Data Science Community Newsletter

13 March 2020 **Data Science Community Newsletter Issue 193**

Please let us (Laura Norén, Brad Stenger) know if you have something to add to the newsletter. We are grateful for financial support from the Academic Data Science Alliance.

Once again this week, COVID *is* the news. On March 11th and 12th scores of universities sent emails to students announcing that spring semester courses would be taught via **Zoom** (or one of its competitors). They were told not to come back if they were on spring break or to be out of the dorms in a week or less. Eight states, plus LA Unified and San Diego County, announced earlier today that K-12 schools would be canceled.

As a sociologist, I started collecting these institutional responses in a **spreadsheet**. I would like your help, readers. Can you please jump into that spreadsheet and enter data about your universities' responses to COVID? There are 100 universities so far, mostly in the US, but also in Canada, the UK, Germany, Switzerland, and one each from Italy and Slovakia. First, check to see if your university is already there. If not, add it and populate all the way across. Leave comments in cell A1 if you've got suggestions for improvement.

If you'd like tips on how to go to all-virtual teaching, two **Stanford** employees put together a guide for **"Teaching Effectively During Times of Disruption."** There's also a **new "gone virtual" Slack channel**, set up by **Zach Lieberman**, for doing the whole academic life from home.

Though the American federal government is doing a piss poor job of managing COVID testing, American institutions of higher education appear to be ahead of their European counterparts when it comes to getting students out of dorms quickly and decisively. Yes, this has caused hardship for students. Yes, some of them will remain in dorms because they have nowhere safe to go. And, yes, this mass evacuation will save lives.

-Laura and Brad

DATA SCIENCE NEWS

University Data Science News

MIT is using machine learning to **discover new antibiotics** and it's turning up some unexpected — but promising — results.

Student loan debt in the U.S. **has grown to $1.56 trillion** shared by 45 million Americans, largely those who acquired the debt after 2003. (Coincidentally, the Trump Administration just announced a **$1.5 trillion in loans available** to shore up short term money markets in the banking industry.) The topic of how to address this situation has come up on the campaign trail; **Bernie Sanders** has proposed simply eliminating it altogether. There's been less discussion of the **Trump administration** plan. Relevant context: Obama introduced income-based repayment plans that allow people to pay no more than 10% of their income towards their loans for 20 years. If they still have a balance at that point, it is forgiven. About 80% ($167 billion) of the loan debt currently in that program was taken out to obtain graduate degrees. **The Trump administration plan** raises the percentage borrowers would repay from 10% to 12.5%; shorten the payment term for undergraduates to 15 years, but *lengthen* the repayment term to 30 years for graduate students. Given that many people do not complete graduate degrees until their late 20s or early 30s, that would put many advanced degree holders into repayment for their entire working careers. It's very different from the Sanders plan and could have a major impact on US students' participation rates in Masters and PhD programs.

The Trump administration, led by Senior Adviser **Ivanka Trump**, is also **pushing** to allow Pell Grants to be used to cover certificate programs, many of which have been shown to have low placement rates (~50%) into jobs that pay $25,000 per year or less. Previous administrations have attempted to limit Pell grants so that they could only be used for degree programs that led to high rates of degree completion, job placement, and above-the-poverty line salaries.

Oregon State University **received an $8.7 million** **DARPA** grant to develop a deep learning algorithm for robots, trained in part on videos of toddlers provided by **Karen Adolph**, an **NYU** behavioral psychologist. **University of Utah** roboticist **Tucker Hermans** is also a named collaborator on the grant.

Cornell University librarians passed a **resolution addressing the creeping use of student data**, including data that can "lead to detailed portraits of individuals' lives, thoughts, values, health, and personal, intellectual, creative, and intellectual interests, even in cases where such data has been 'de-identified' or 'anonymized'." One of their recommendations is to ensure that students must actively opt-in to any location tracking and any other projects that, "collect personal behavioral data." This would make it difficult to use typical cybersecurity software, which doesn't necessarily do location tracking, but it could be described as collecting "personal behavioral data" for activities such as logins. Logins aren't likely to be particularly privacy sensitive on their own, but they are critical for providing legitimate cybersecurity.

Carnegie Mellon University's flu forecasting team has some good news: **flu season peaked two weeks ago**. Though the team would typically disband by the end of May, it will continue its work as COVID cases increase across the US.

Professor **Barend Mons** of the **Leiden University Medical Center** in Belgium argues that data stewardship — making data FAIR (findable, accessible, interoperable and reusable) — doesn't happen by itself and can't be squeezed into the already overflowing interstitial moments of academic life. Instead, Mons **calls for setting aside 5% of research budgets** to pay for (human) data stewards to achieve the FAIR standards, manage data, and maintain ethical standards in sharing, storage, and reuse. Though this would add to the university's overhead, it should free researchers time to focus on research rather than data stewardship. It's quite difficult to argue with the notion that data stewardship is trivial or that knowing there's an expert taking care of it would free up time for researchers to work in other areas. Worth thinking about further.

Illinois Institute of Technology is **launching a new School of Computing** to house the Institute's computer science, data science, artificial intelligence, applied mathematics, cybersecurity, and information technology and management programs. The effort was supported by Chicagoan **Chris Gladwin**, a billionaire serial tech entrepreneur.

Harry Shum, former head of **Microsoft** AI and Research, has **returned to academia**. Professor Shum will now be the head of the computer vision and computer graphics PhD program at **Tsinghua University**.

Spleeter, **a free, open-source AI tool** accepts a recorded piece of music and outputs each instrument or instrument group, plus vocals into four tracks. This takes remixing to a whole new level of ease and creativity, though keep in mind that copyright still applies. (I'm not a lawyer, I don't know exactly how copyright works on individual instrumental tracks.)

Company Data Science News

Twitter rewrote its developer policy to **explicitly allow academic research via its API and the academic use of data coming from it**. Have I mentioned how much I like Twitter as a company lately? No? I really like Twitter as a company.

As clinical device makers and hospitals start incorporating AI algorithms into their products and processes, there are **open questions about who bears responsibility** when they lead to medical mistakes, which is inevitable. Medicine is difficult and the stakes are high. Algorithms change over time, making it hard to use legal mechanisms designed for more linear design and development processes. Plus, "random errors, in which artificial intelligence misses an obvious abnormality for inexplicable reasons, are uncommon and difficult to predict. But statistically speaking, their occurrence is certain." It's not as simple as overriding an AI, either, because there can be liability associated with deliberately ignoring a suggested course of action, too. Obtaining additional scans and diagnostics is one way doctors can gain more clarity, but that also **raises costs** for hospitals and patients. These are the types of extremely challenging conversations that arise around shared cognition. We're just getting started.

OpenBiome a company selling stool specimens, has been implicated in the **deaths of two recipients and the hospitalization of four others**. Three of the donors produced stool containing illness-causing bacteria, according to the **Food and Drug Administration**. This seems like a field that is begging for better data science. Transferring that many different types of bacteria into a person with so many other bacteria is going to result in all sorts of unpredictable outcomes. Why isn't there more data science? [insert joke about the quantity and quality of data janitorial labor here]

American, **Delta**, and **United Airlines** **slashed the number of flights on their schedules**. This was before **Trump** announced a travel ban between the U.S. and many E.U. countries for those who aren't American citizens or green card holders. Couple that with all the cars that aren't commuting to work and school every day and maybe we'll hit those Paris Accord emissions targets after all? Can we learn how to connect meaningfully while remaining local?

Two companies are using mobile phone location data in privacy-infringing ways. Reston-based **Babel Fish** and **Google's Sensorvault** are both used by government officials to identify specific users who have been in given locations. Sensorvault **works best on Android**. Since Babel Fish culls location data from popular mobile apps, it's more likely to be effective on either Android phones or **Apple** iPhones, though details are skimpy. **Google** is not selling Sensorvault as a service, but law enforcement can submit search requests during investigations, which they apparently do about 180 times per week. Babel Fish *does* sell its data to agencies such as **Customs and Border Patrol**, which has **aroused the attention** of Senator **Ron Wyden** (D-OR) who's concerned that warrantless use of personal location data may be a 4th amendment violation.

Honeywell has been developing **the world's most powerful quantum computer** over the past several years, recently adding partnerships with **Microsoft** and **JP Morgan Chase**, who is developing algorithms for it. The machine has a quantum Volume of 64, a metric specific to quantum computers.

Honeywell also introduced a machine learning approach to building climate and HVAC controls, **Honeywell Forge Energy Optimization**. Implementation can lead to a 10% reduction in energy for climate control without impacting user comfort. If the system can balance temperatures in an office so that there are no longer hot spots and cold spots, I will nominate it for a Nobel.

Government Data Science News

The Pentagon will **reconsider the JEDI cloud computing award** process that gave the $10 billion contract to **Microsoft**. **Amazon**, a competitor for the contract, claims that the award went to Microsoft due to meddling by **President Trump**. It is known that Trump is not a fan of Amazon CEO **Jeff Bezos**.

The Department of Health and Human Services released guidance on how patient data should be stored and processed so that it can **be accessed via API**. This is a fascinating experiment in data-as-a-service and I'll expect many third party apps to populate this space, giving patients an array of opportunities to layer their everyday health data with their official medical data. One excellent example of a use-case for patient data is what the **Boston University's Physical Therapy Center** where they are **integrating data** such as number of visits, range of motion, type of therapy, and even copay amounts (higher copays may lead to fewer visits and poorer outcomes).

California, **Washington**, **New York**, and **Massachusetts** account for 90% of the jobs requiring AI skills in the U.S. However, the **ZipRecruiter** study **reported by VentureBeat**, noted that housing costs in these four states are so high that a crop of secondary states are gaining quickly, including **Colorado**, **Utah**, **Virginia**, **Texas**, and **Arizona**.

Abigail Keller at the thinktank **American Enterprise Institute** is **using Twitter to provide a running tally of available COVID diagnostic tests** in the U.S. There's also **COVID+ case data by**

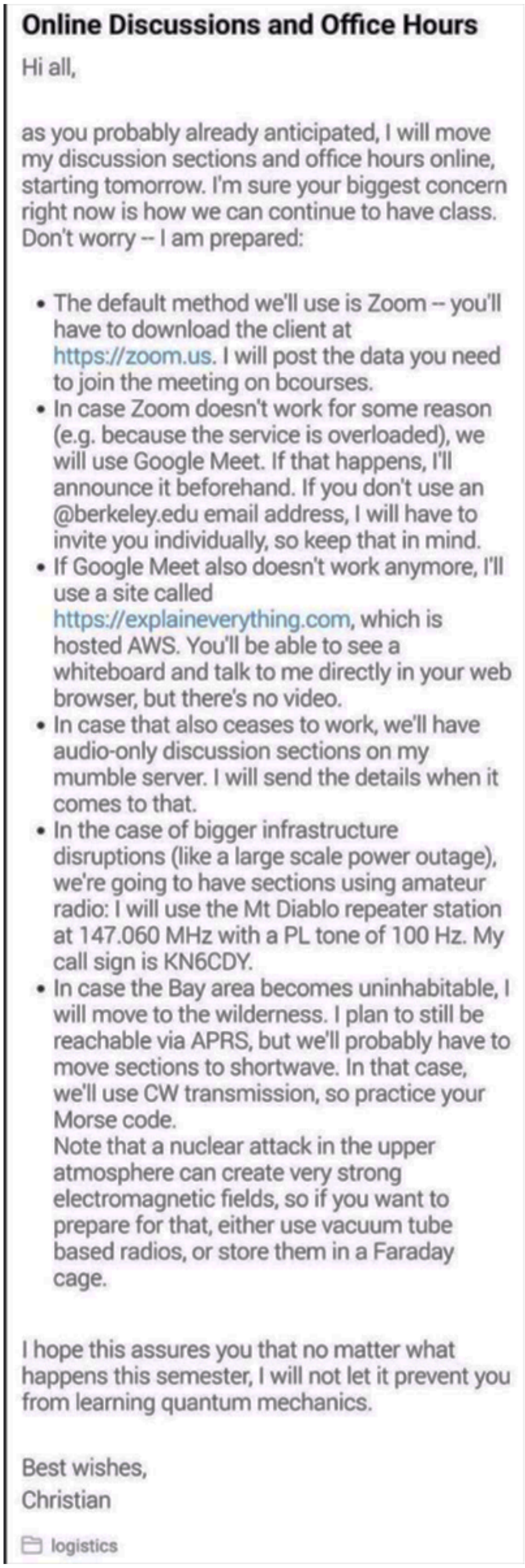
state, updated daily, where possible.

Extra Extra

Two identical twin sisters both got breast cancer. One, a professor in the U.K., received care through the **National Health Service**. The other, a Federal government worker in the U.S., received care through her employer-sponsored plan. Both are in remission, but **the U.S.-based twin spent $17,500**, plus tens of hours of time negotiating with insurance companies.

Tweet of the Week

Twitter, Jenna Franke from March 10, 2020



Online Discussions and Office Hours

Hi all,

as you probably already anticipated, I will move my discussion sections and office hours online, starting tomorrow. I'm sure your biggest concern right now is how we can continue to have class. Don't worry -- I am prepared:

- The default method we'll use is Zoom -- you'll have to download the client at https://zoom.us. I will post the data you need to join the meeting on bcourses.
- In case Zoom doesn't work for some reason (e.g. because the service is overloaded), we will use Google Meet. If that happens, I'll announce it beforehand. If you don't use an @berkeley.edu email address, I will have to invite you individually, so keep that in mind.
- If Google Meet also doesn't work anymore, I'll use a site called https://explaineverything.com, which is hosted AWS. You'll be able to see a whiteboard and talk to me directly in your web browser, but there's no video.
- In case that also ceases to work, we'll have audio-only discussion sections on my mumble server. I will send the details when it comes to that.
- In the case of bigger infrastructure disruptions (like a large scale power outage), we're going to have sections using amateur radio: I will use the Mt Diablo repeater station at 147.060 MHz with a PL tone of 100 Hz. My call sign is KN6CDY.
- In case the Bay area becomes uninhabitable, I will move to the wilderness. I plan to still be reachable via APRS, but we'll probably have to move sections to shortwave. In that case, we'll use CW transmission, so practice your Morse code.

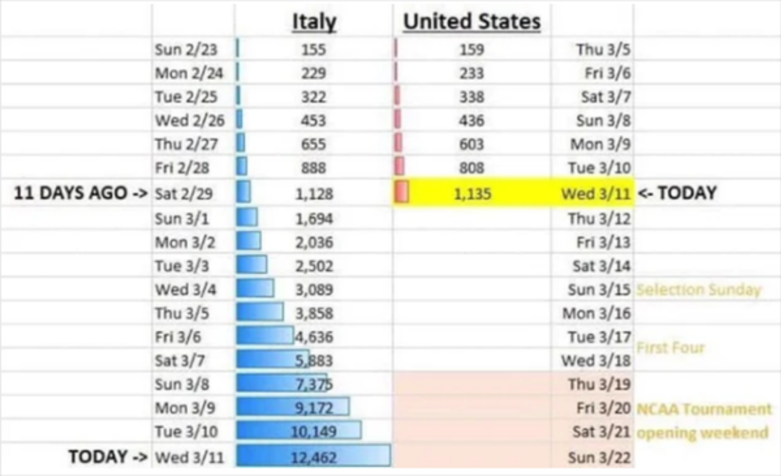
  Note that a nuclear attack in the upper atmosphere can create very strong electromagnetic fields, so if you want to prepare for that, either use vacuum tube based radios, or store them in a Faraday cage.

I hope this assures you that no matter what happens this semester, I will not let it prevent you from learning quantum mechanics.

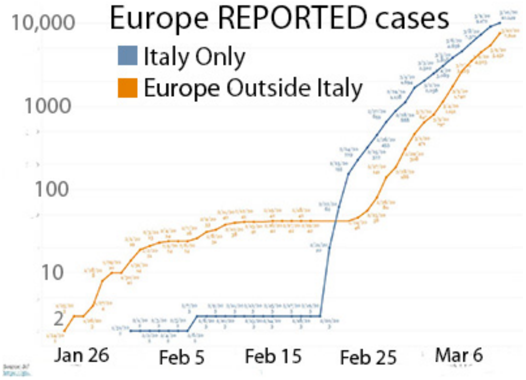
Best wishes,
Christian

Data Visualization of the Week

reddit/r/Dataisbeautiful from March 13, 2020



		Italy	United States		
	Sun 2/23	155	159	Thu 3/5	
	Mon 2/24	229	233	Fri 3/6	
	Tue 2/25	322	338	Sat 3/7	
	Wed 2/26	453	436	Sun 3/8	
	Thu 2/27	655	603	Mon 3/9	
	Fri 2/28	888	808	Tue 3/10	
11 DAYS AGO ->	Sat 2/29	1,128	1,135	Wed 3/11	<- TODAY
	Sun 3/1	1,694		Thu 3/12	
	Mon 3/2	2,036		Fri 3/13	
	Tue 3/3	2,502		Sat 3/14	
	Wed 3/4	3,089		Sun 3/15	Selection Sunday
	Thu 3/5	3,858		Mon 3/16	
	Fri 3/6	4,636		Tue 3/17	First Four
	Sat 3/7	5,883		Wed 3/18	
	Sun 3/8	7,375		Thu 3/19	
	Mon 3/9	9,172		Fri 3/20	NCAA Tournament
	Tue 3/10	10,149		Sat 3/21	opening weekend
TODAY ->	Wed 3/11	12,462		Sun 3/22	



EVENTS

COVID-19 and AI: A Virtual Conference

Online April 1. "COVID-19 and AI: A Virtual Conference will address a developing public health crisis. Sponsored by the Stanford Institute for Human-Centered Artificial Intelligence (HAI), the event will convene experts from Stanford and beyond to advance the understanding of the virus and its impact on society. It will be livestreamed to engage the broad research community, government and international organizations, and civil society." [Event details and agenda will be posted soon.]

Academic Data Science Alliance, Data Science Education Special Interest Group Call

Online April 2, starting at 10 a.m. Pacific. "Is your institution considering starting a data science degree program? Come hear three ADSA community members discuss how their organization decided to create data science degree programs and how these programs took the shape they did. We'll be hearing from **Sarah Stone** (**University of Washington**), **Ajay Anand** (**University of Rochester**), and **HV Jagadish** (**University of Michigan**)." [registration required]

Does Science Self-Correct? What We've Learned At Retraction Watch

Remote TBD April 24, starting at 3 p.m., NYU (Pless Hall, 82 Washington Square East). "The Center of Health and Rehabilitation Research is very pleased to announce our Spring Speaker Event for 2020 featuring Dr. Ivan Oransky, MD." [free, registration required]

Registration Open for 2nd Edition of Machine Learning School in The Netherlands: July 7-10, 2020

Breukelen, Netherlands July 7-10 at Nyenrode Business School. "It's the ideal learning setting for professionals that wish to solve real-world problems by applying Machine Learning in a hands-on manner. This includes analysts, business leaders, industry practitioners, and anyone looking to boost their team's productivity by leveraging the power of automated data-driven decision making." [$$$]

DEADLINES

Conferences

SCF 2020 - ACM Symposium on Computational Fabrication

Boston, MA June 28-30 at **Boston University**. The symposium "will gather experts and enthusiasts from many areas of academia and industry, in order to explore the use of computational tools for the creation of physical things. SCF provides a venue for participants to discuss cutting-edge results, cross-pollinate ideas, and strengthen interdisciplinary connections and collaborations." Deadline for paper submissions is April 10.

Call for Paper: International Workshop on Federated Learning at IJCAI 2020

Yokohama, Japan July 13 (tentative). "In order to explore how the AI research community can adapt to this new regulatory reality, we organize this one-day workshop in conjunction with the **29th International Joint Conference on Artificial Intelligence**." Deadline for submissions is April 26.

RecSys 2020 – Call for Contributions

Rio de Janeiro, Brazil September 22-26. The conference "strongly encourages the submission of algorithmic papers that repeat and analyze prior work." ... "Submissions regarding replicability or reproducibility papers are welcome in all areas related to recommender systems (see the main track Call for Papers for a list of topics)." Deadline for submissions is May 4.

NeurIPS 2020 Call for Papers

Vancouver, BC, Canada December 6-12. Deadline for abstracts submissions is May 5.

FOCI '20 Preliminary Call for Papers

Boston, MA "The 10th USENIX Workshop on Free and Open Communications on the Internet (FOCI '20) will take place August 11, 2020, and will be co-located with the **29th USENIX Security Symposium**." Deadline for paper submissions is May 21.

TOOLS & RESOURCES

Keyboard shortcuts for Gmail

Google, Gmail Help from March 02, 2020

You can use keyboard shortcuts to navigate your inbox and messages, format text, and complete actions like archiving and deleting.

ACM publishes new journal on the Internet of Things

EurekAlert! Science News, Association for Computing Machinery from March 05, 2020

The **Association for Computing Machinery** "published the inaugural issue of *ACM Transactions on Internet of Things* (TIOT). The new ACM journal features novel research contributions and experience reports in several research domains whose synergy and interrelations enable the Internet of Things vision. TIOT focuses on system designs, end-to-end architectures, and enabling technologies, and publishes results and insights corroborated by a strong experimental component."

Writing a collaboration policy

Flatiron Center for Computational Astrophysics, Astro Data Group, Ruth Angus from February 28, 2020

"Inspired by a conversation with **Andy Casey** who we visited at **Monash University** in Melbourne last week, I decided to write a Collaboration Policy, and share my feelings on the topic at our weekly group meeting."

"At Monash, Andy showed me the Research Expectations page of his website, which is a document outlining the expectations he has for his students, and what their expectations might be for him. I love this idea and would like to create my own research expectations guidelines for any students working with me, current or future. As I started to write it however, I realized I wanted to write a zeroth-order expectations document first. Something a little more general, targeted at everyone I work with: a collaboration policy."

How to Build High-Performing Engineering Teams

HackerRank Blog, Vivek Ravisankar and Mike Tria from March 06, 2020

"This post discusses how to assess developer candidates and build high-performing engineering teams."

Fixing academic posters: the #BetterPoster approach

astrobites blog, Oliver Hall from February 28, 2020

Mike Morrison, a psychology PhD student, wants to make poster sessions more efficient. In his opinion, the ‘cardinal sin’ of posters is that they often require somebody to read them for 10 minutes straight, in a time-pressured environment. He identified 3 things that posters should embody, and breaks them down as follows:

Posters should maximise the amount of insight for people attending a poster session.

Detailed information should still be present, but not the core of the poster.

These design goals should be easy to achieve and accessible to new and old scientists.

National Wilderness Areas

ArcGIS Hub from March 09, 2020

"A map service depicting parcels of **Forest Service** land congressionally designated as wilderness such as National Wilderness Areas. This map service provides display, identification, and analysis tools for determining current boundary information for Forest Service managers, GIS Specialists, and others."

CAREERS

Tenured and tenure track faculty positions

Director of the Ethical Tech Law and Policy Clinic

Duke University, School of Law; Durham, NC

Associate Professor/Professor/Full Professor Research Position in Data Management and Data Wrangling

Hasselt University, Data Science Institute; Hasselt, Belgium

Full-time, non-tenured academic positions

Application for Data Scientist - Schmidt DataX Project

Princeton University, Center for Information Technology Policy; Princeton, NJ

Research Assistant in the Computational Approaches to Modeling Language (CAMeL) Lab - Division of Science

New York University, NYU Abu Dhabi; Abu Dhabi, United Arab Emirates

Computing Sciences Researcher I

University of Arizona, Agricultural Data Science group; Tucson, AZ

R&D Software Engineer I

University of Arizona, Agricultural Data Science group; Tucson, AZ

OHI/O Program Coordinator

The Ohio State University, Computer Science and Engineering Department; Columbus, OH

Research Director

University of Chicago, Urban Crime Labs; Chicago, IL

Program Analyst, Ocean Team

University of California-Berkeley, Haas School of Business; David and Lucile Packard Foundation; Berkeley, CA

Visiting Assistant Professor in Learning Engineering

Boston College, Lynch School of Education; Chestnut Hill, MA

Postdocs

Postdoc in Music Cultures & AI

KTH Royal Institute of Technology, Department of Intelligent Systems; Stockholm, Sweden

Postdoctoral Researcher, Machine Learning and Visualization

Apple, Machine Learning and AI; Pittsburgh, PA

Knight Postdoctoral Research Fellow - "algorithms, misinformation, polarization, propaganda, political institutions, and platform companies"

University of North Carolina, School of Information and Library Science; Chapel Hill, NC

Hewlett Postdoctoral Research Fellow - "misinformation, partisanship, polarization, propaganda, political institutions, and journalism"

University of North Carolina, School of Information and Library Science; Chapel Hill, NC

AI/Machine Learning/Encryption Postdoctoral Research Associate (CS Position)

University of North Carolina, School of Information and Library Science; Chapel Hill, NC

AI/Machine Learning/Encryption Postdoctoral Research Associate (Social Science PhD)

University of North Carolina, School of Information and Library Science; Chapel Hill, NC

Postdoctoral Associate - Open Science

Carnegie Mellon University, University Libraries; Pittsburgh, PA

Postdocs (2) Computational Sociology - Science of Science

University of Copenhagen, Department of Sociology; Copenhagen, Denmark

Full-time positions outside academia

Data Science Assistant

Pew Research Center, Data Labs; Washington, DC

Deputy Chief Technology Officer for Policy & Future Planning

New York City, Mayor’s Office of the Chief Technology Officer; New York, NY

Data Engineer - Informatics Workflows

Sage Bionetworks; Seattle, WA

Principal Software Automation Engineer, Machine Learning Toolchain

iRobot; Bedford, MA

Research Scientist, HCI Research - AI/ML

Apple, Machine Learning & AI; Seattle, WA

Sr. Grants Management Specialist (Senior Programs Officer)

National Foundation on the Arts and the Humanities, Institute of Museum and Library Services; Washington, DC

Natural Resources Specialist (Monitoring Data Coordinator)

Department of the Interior, Bureau of Land Management; Lakewood, CO

Director of Technology Projects

Electronic Frontier Foundation; San Francisco, CA

Diversity and Inclusion Research Fellow

Partnership on AI; San Francisco, CA

Research and Analytics Director, Chatbots

National Domestic Workers Alliance, NDWA Labs; New York, NY, or Remote

Data Science Manager

Center for Policing Equity; Washington, DC

Research Analyst (Health Policy)

The Urban Institute; Washington, DC

Senior Researcher

Girl Scouts; New York, NY

Research Analyst, Africa Growth Initiative

The Brookings Institution, Global Economy and Development program; Washington, DC

Data Analysis Team Manager

Space Telescope Science Institute; Baltimore, MD

Senior Data Visualization Editor

McKInsey, Global Editorial Services; Atlanta, Boston, New York City or Waltham, MA

Research Engineer, AllenNLP

The Allen Institute for Artificial Intelligence; Seattle, WA

Policy Fellow or Policy Counsel (Youth & Education)

Future of Privacy Forum; Washington, DC

Quantitative Researcher & Developer

Hudson and Thames Quantitative Research; New York, NY

Data Scientist

Observable; San Francisco, CA

Machine Learning Engineer - High-Performance Deep Learning and Neural Computing

Target Data Sciences; Sunnyvale, CA

Internships and other temporary positions

Seasonal field research assistants, Great Basin

University of California-Davis, Earth Research Institute; Nevada and eastern California

Project Manager for Data Science for Social Good

Alan Turing Institute, University of Warwick; Coventry, England

Microsoft Research Data Science Summer School

Microsoft Research; New York, NY

the Congressional innovation Scholars program

TechCongress; Washington, DC

Basketball Analytics Summer Intern

Los Angeles Lakers; El Segundo, CA

Adelie Penguin Population Ecology Internship

Point Blue Conservation Science; Ross Island, Antarctica

part-time grad student work

New York University, Marron Institute; New York, NY, or Berlin, Germany

Data Science Intern - Merch Product Development

Stitch Fix; San Francisco, CA

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 193.