



Data Science Community Newsletter

16 December 2019

Data Science Community Newsletter Issue 187

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.

DATA SCIENCE NEWS

University Data Science News

In a surprising new study, researchers found that, "20 percent of 205,011 men in a large genetic database called the **UK Biobank** have **lost Y chromosomes from some detectable proportion of their blood**. By age 70, 43.6 percent of men had the same issue." In addition to suggesting some risks associated with older men fathering children, there's widespread interest in understanding how the loss of the Y chromosome may be linked to testicular, prostate, kidney cancer and glioma. The UK Biobank is **shifting medical genetic research towards bigger sample sizes that yield more robust findings**, though we should still keep in mind that the samples are non-representative (94% white, largely middle class or upper middle class participants).

Sarah E. Hill, professor of evolutionary psychology and psychology at **Texas Christian University** notes that almost all medical research is performed on male animals and then on human men, leaving the medical community severely **under-informed about the way women may respond to treatment**. She asks, "If the top research journals will publish your research without including female participants, would you go through all of the trouble when it might ultimately shoot you in the foot?" Readers, I know many of you (most of you?) care about accuracy and are upset that the accuracy of medical treatment may be sub-optimal for half the population. In her piece, she explains why women's hormonal cycles make them more complicated research subjects and how this can double or triple the cost and time it takes to include women. In the context of pressure to publish, publish, publish, and no pressure to produce gender-parity in medical research, it is obvious how we ended up here but less obvious how to reset and recover.

On the other side of medical research, the authorship side, **senior women scholars are 40% less likely to write an invited commentary article** in a medical journal than their colleagues who are men, holding years of experience and quality/quantity of articles published constant.

The **AI Now Institute** released its **2019 report**, focusing on about a dozen areas of concern, everything from AI and the climate crisis to neighborhood surveillance tools to the impact of algorithmic speed on workers to "the tough road toward sociotechnical perspectives." As one of the only major public facing AI think tanks in the US, their report is worth a skim for anyone working in AI.

Harvard University is in the midst of a strike on the part of its graduate students who serve as teaching fellows (TFs). Harvard's **Provost Alan Garber** has **provided the following guidance to departments sending spring semester offer letters** to TFs: "This offer is conditioned on your

commitment to being able to start on X date. Please indicate in writing whether you can commit to starting on that date.' If the candidate says that they can commit to that date, then we can assume that they are not honoring the strike and will report to work." This is an age-old tactic of union busting. H/t [Jake Anbinder's twitter](#).

A lawsuit filed by plaintiffs in the **Compton Unified School District** claims that the **University of California** is **discriminating against disabled, low-income, multilingual, and underrepresented minority applicants** by requiring the SAT or ACT for admissions. The UC system has been studying the use of standardized tests for months and will release its report on schedule next spring. **The College Board**, owner of the SAT, is facing an existential threat with this suit, stating that they feel it contains "a number of false assertions."

Predatory publishers are the street criminals of the academic circuit, publishing studies that may be patently false and taking fees from authors who may not realize these journals are utter rubbish. A new panel of 43 scholars has taken the first step towards pulling these weeds by **defining** them. A predatory journal is one that publishes, "false or misleading information, deviat[es] from best editorial and publication practices, lack[s] transparency, and uses aggressive, indiscriminate solicitation."

Renata Rawlings-Goss Executive Director of the **NSF South Big Data Hub** is running a 5-week networking and career-boosting course: **The Data Career Academy** starting in January. It's limited to ten seats so Dr. Rawlings-Goss is running '**discovery calls**' with interested parties to ensure matching expectations.

Case Western Reserve University has launched a **new Computer & Data Sciences Department** with \$5 million from alumnus **Kevin Kranzusch**. They are naming a chair after Kranzusch and have opened a **hiring portal** to receive applications.

Purdue University trustees **approved the funds** to build a new \$40 million, 86,000 square foot building to house the school's growing data science program. Big 10 cohort, **University of Illinois**, **will renovate two historic buildings** for data science, Illini Hall and Altgeld Hall, rather than demolish and replace them with new construction.

In Canada, the **University of Waterloo** has launched **The Quantum Alliance**, an industry-academia partnership.

Happy holidays: **Harvard's Case Law Project** is making **360 years of case law in 6.7 million legal cases** available to the public via API or large zip files. There's a powerful search tool and an historical trend visualizer, too.

New research lends more evidence to the fact that **opening grocery stores in "food deserts" does not lead to healthier eating** by community members. When it comes to fresh fruits and vegetables, it's not a field of dreams situation.

American and European cell and molecular biologists had their annual meeting during which they [discussed the future of open science in their field](#). Open access publishing, opening scientific findings to anyone who wants to see them, including the names of reviewers in peer-review, and, presumably sharing data (though that was not mentioned in the write-up) are seen as important characteristics of the future of the field. Revising funding and other incentives structures to align with these expectations is an ongoing challenge. Notably, one commenter from the article indicated support for the concept of open science, but noted that open access publishing fees are so high that they are fully half a year's grant in some countries.

Meanwhile, **The Wharton School** discovered that there is money in sports — it's "a multibillion-dollar international industry" — and [launched the Wharton Sports Analytics and Business Initiative](#).

Company Data Science News

Ring cameras, now owned by **Amazon**, have been hacked repeatedly. It happened again last week when [a hacker used the Ring camera](#) that a family had installed in their three daughters' bedroom four days earlier. He also used the speaker to talk to them. No report on what he said, but please keep in mind that it's a bad idea to install these cameras in bedrooms or any other rooms where you or your family members change their clothes or undertake other naked activities. Maybe just don't put the cameras anywhere in your house.

And...you may want to [think twice about having listening devices](#) like **Alexa**, **Google Home**, **Apple's Siri** and **Facebook's Portal** in your house. In order to train the AI transcription programs, they capture snips of audio which have included all manner of private utterances (a young man pondering rape, people reciting their addresses and phone numbers, a man trying to order a sex toy, other men trying to hit on Alexa in crude ways, etc). If you have a listening device in your home it is unreasonable to expect you'll always be on your best, most above-board behavior. Amazon devices - by far the most popular - come with terms of service that allow the company to record and store your data indefinitely. If you value your privacy, you may want to ditch the home listening devices and re-learn how to play music, turn on the lights, and type text into search fields and text messages without robotic assistance.

The DNA-database company that led to the prosecution of the Golden State Killer, **GEDMatch** was [acquired last week](#) by a forensic genetics firm that typically processes crime scene DNA, **Verogen**. In the immediate aftermath of the Golden State Killer discovery, GEDMatch was used in about [70 other cases](#), but the firm faced an outcry from its users and changed its terms of service. After the revision, only DNA from people who explicitly opted into participating in law enforcement investigations could be used to investigate crimes. It remains to be seen exactly how the new ownership will make use of the data, but it brings up good questions about what happens to data during an acquisition. The data goes to the new owners. New terms of service and privacy policies come into play. This is one of the sleeper ethical questions that has arisen in the data saturated age: because data is durable and can be easily copied, it will outlast the contexts in which it was created and the purposes it was meant to serve. Collectively, we need to deal with this ethical dilemma of the out-of-context survival of complete, sensitive data records.

A study of **LinkedIn** job post data found that **AI Specialists, Data Scientists, and Data Protection Officers are the most sought** after positions in 14 of the 17 countries surveyed.

Plus.ai, a self-driving trucking company, **sent 40,000 pounds of Land O'Lakes** butter across the country without mishap last week. The two human drivers who accompanied the shipment were likely bored most of the time. The system uses cameras, radar, and lidar to keep trucks from running into anything or running off the road altogether. Land O' Lakes notes that Q4 is peak-demand butter season. Don't I know it. :)

The **Future of Privacy Forum** announced the winners of its 10th annual Privacy Papers for Policy Makers Awards. All of the papers were written in 2019 and are fully linked **in the article**. All are also broadly legible to people who aren't privacy specialists and touch on topics from **user-interface design** to GDPR to algorithmic fairness/auditing.

NeurIPS 2019 also announced its **winning papers**.

Goldman Sachs discovered the **power-broker status of APIs in 2017**: "The way we see it [the future] going is to APIs." Arguably, they were a little late to the game, but it is notable when a giant investment group pivots towards selling data.

Verizon acquired **Yahoo** which decided to kill off Yahoo Groups earlier this fall. Archivists have been trying to preserve the content in the Groups, but **Verizon blocked their efforts**. This is a lesson in why it can be better not to use platforms owned by giant tech companies who can capriciously pull the plug whenever it pleases them.

Government Data Science News

The **U.S. Census Bureau** has faced one hurdle after another in the run-up to the 2020 decennial census. A **new Reuters report** faults external contractors **Pegasystems** and cybersecurity contractor **T-Rex** for failing penetration tests and denial of service attacks in 2018. Those vulnerabilities have not been sufficiently addressed, according to insiders, leaving some of the world's most valuable ground truth data at risk of falsification. The Census hired Pega and T-Rex because it is moving to an online format for the first time in 2020, but insiders note that given the immense cost and ongoing security and functionality concerns, the Census may have been better off if they'd kept the project in-house.

San Diego County will **halt the use of facial recognition on January 1, 2020**. This is remarkable because San Diego County had one of the largest, longest running implementations used by more than 30 agencies. San Diego County had been sharing the data with ICE, but that was ended in October. The reason the entire program is going into a freeze state is new legislation in California that bans the use of police body cameras starting January 1st.

Brookline — a city adjacent to Boston — has become the **fourth U.S. city to ban the use of facial recognition technology** by civic officers. This means no buying it, no using it, no making contracts

with companies that use it.

Meanwhile, police in **Moscow**, Russia (not Idaho), have been **taking bribes to allow civilians to access live feeds and check the previous five days of footage from 175,000 surveillance cameras** in that city. Some of them paid enough in bribe money to be able to use the facial recognition feature. So much for privacy!

Germany has been running a process to develop a **National Research Data Infrastructure** for over a year now. We haven't said peep about this €85 million per annum plan (h/t to **Dan Katz**). The goal is to provide "science-driving data services to research communities" and it will tackle issues like establishing cross-disciplinary meta-data standards, interoperable data management measures and collaboration with researchers outside of Germany. The program will eventually be run out of locations at 30 consortia partners. This kind of national, cross-disciplinary effort to store and distribute all sorts of types of data is wise — if governments don't do it, scientists end up using commercial and open source tools that may not be tailored for scientists or for public access.

Finland provided free basic **online training in AI literacy to 1% of its population** (55,000 people) in only a few months. At the moment, Finland holds the rotating EU presidency and would like to extend its AI educational success across the EU, training more than 50 million EU citizens in **basic AI literacy**. I'm not seeing that kind of state or federal AI leadership in the US.

As more of us are convinced we hit **peak San Francisco** at some point in the past decade, new tech hubs are continuing to develop. There are the regular names: Seattle, New York, Austin, Boston, and DC. But there's **also LA**. Sure, it lacks public transportation, but it leads in beautiful sunny days. On the other hand, a mainstay of LA's think tank scene, the **Broad Center**, has announced it is **moving cross country**, to the **Yale University School of Management**.

The **U.S. Veterans Administration** (VA) has **announced** a new **National Artificial Intelligence Institute** to leverage innovative tools and be a centralizing location for academic, government, nonprofit, and industry partners working on AI approaches to health care.

Extra Extra

There is a new **Feminist Data Manifesto**, a product of an August 2019 **University of Michigan** workshop. Per the preamble, "It refuses harmful data regimes and commits to new data futures". The manifesto was largely written by academics, though they refer to the academy as "ethically compromised", a power hierarchy in need of "subversion, undermining, opening, making possible". It's certainly worth reading in full, but what I like is that at the same time they push against today's dominant data ethos (and creatively think about future realms of data dominance), they end on the kind of contingent optimism I associate with my favorite feminist manifestos. They write, "Data can be a check-in, a story, an experience or set of experiences, and a resource to begin and continue dialogue. It can - and should always - resist reduction. Data is a thing, a process, and a relationship we make and put to use. *We can make it and use it differently.*"

There is also a new [PBS Frontline documentary on AI](#). It clocks in at just under 2 hours running time, 28 minutes shorter than *Star Wars: The Rise of Skywalker*, which opens this week.

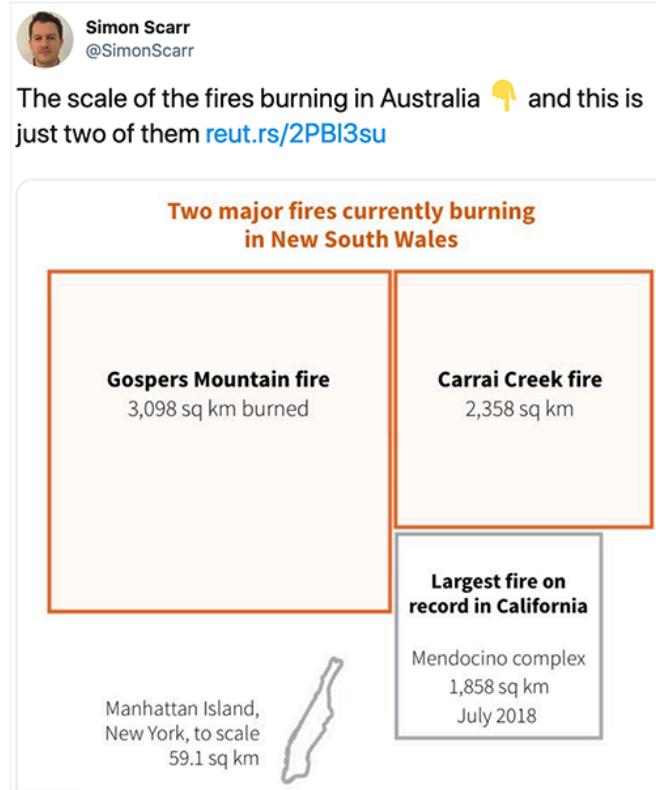
Tweet of the Week

Twitter, Isa Watson from December 09, 2019



Data Visualization of the Week

Twitter, Simon Scarr from December 11, 2019



EVENTS

The Misinformation Virus

Online December 17, starting at 20:00 GMT, BBC 4. "**Angela Saini** investigates the lethal spread of alternative facts and discovers that the very architecture of the web amplifies dangerous pseudoscience online." [audio, 37:00]

Databite No. 127: Jasmine McNealy, An Ecological Approach to Data Governance

New York, NY January 8, 2020, starting at 5 p.m. "Dr. McNealy argues that we require an ecological approach for understanding this era of emergent technology and data — both for creating adequate policy, and for protecting the vulnerable." [rsvp required]

Data Day Texas

Austin, TX January 25, 2020. "Originally launched in January 2011 as one of the first NoSQL / Big Data conferences, Data Day Texas each year highlights the latest tools, techniques, and projects in the data space, bringing speakers and attendees from around the world to enjoy the hospitality that is uniquely Austin." [\$\$\$]

Applied Machine Learning Days at EPFL

Lausanne, Switzerland January 25-29. "Five days of 30 hands-on sessions and 29 tracks on machine learning and artificial intelligence with top speakers from around the world." [\$\$\$\$]

CBJ Hockey Analytics Conference

Columbus, OH February 7-8. "A two-day conference that will showcase state-of-the-art research happening in the hockey analytics community and explore the use of data-driven analysis being done by industry professionals today." [\$\$]

O'Reilly Strata Data & AI Conference

San Jose, CA March 15-18, 2020. "Data feeds AI; AI makes sense of data. So it also made sense to combine the O'Reilly Strata Data and AI Conferences—covering two of the most pressing technological trends of the decade—and giving you access to the full breadth of both programs." [\$\$\$ \$]

DEADLINES

Contests/Award

#TFWorld TF 2.0 Challenge hackathon

"Use our official release of TensorFlow 2.0 to do something nifty: build a model, a mobile or web application, an art installation, or something else entirely! The sky is the limit - and creativity is encouraged. We're excited to see what you build!" Deadline for submissions is December 31.

5th annual NFL 1st and Future competition - Innovations to Advance Athlete Health and Safety Competition

"Submissions for innovations that could improve player health and safety, including but not limited to: protective equipment, medical devices, sensors and training devices. Up to four start-ups will be selected as finalists and will have the chance to present their innovations on stage in Miami." Deadline for entries is January 2, 2020.

CONFERENCES

3rd annual Symposium on Applications of Contextual Integrity

Chicago, IL September 21-22, 2020, at **University of Chicago**. "The aim of the symposium is to foster interaction among diverse communities of research and practice using contextual integrity to reason about privacy, and to design and evaluate, craft regulation, and generate formal logics for privacy." Deadline for submissions is June 22, 2020.

4th Workshop on Immersive Analytics: Envisioning Future Productivity for Immersive Analytics

Honolulu, HI April 25/26, 2020, part of **CHI 2020**. "This workshop will aim to identify the key productivity challenges for data-centric, immersive systems." Deadline for submissions is February 11, 2020.

HILDA 2020 Workshop on Human-In-the-Loop Data Analytics

Portland, OR June 9, 2020, co-located with **SIGMOD 2020**. "HILDA brings together researchers and practitioners to exchange ideas and results on human-data interaction. It explores how data management and analysis can be made more effective when taking into account the people who design and build these processes as well as those who are impacted by their results." Deadline for submissions is March 23, 2020.

Education Opportunities

Foreign Affairs Information Technology Fellowship

"The U.S. Department of State is seeking to attract top tech talent to the Foreign Service that reflects

the diversity of the United States. Women, members of minority groups underrepresented in the Foreign Service, and students with financial need, are encouraged to apply to this challenging and rewarding program." Deadline for applications is February 14, 2020.

Studies/Surveys

U.S. Citizenship and Immigration Services Fee Schedule and Changes to Certain Other Immigration Benefit Request Requirements

"The **Department of Homeland Security** (DHS) proposes to adjust certain immigration and naturalization benefit request fees charged by **U.S. Citizenship and Immigration Services** (USCIS). USCIS conducted a comprehensive biennial fee review and determined that current fees do not recover the full costs of providing adjudication and naturalization services. DHS proposes to adjust USCIS fees by a weighted average increase of 21 percent, add new fees for certain benefit requests, establish multiple fees for petitions for nonimmigrant workers, and limit the number of beneficiaries on certain forms to ensure that USCIS has the resources it needs to provide adequate service to applicants and petitioners." Deadline to provide comments is December 16.

RFP

About Kaggle's Open Data Research Grant

"Researchers are an important part of **Kaggle's** community and we want to expand our support for them through the first ever Kaggle Open Data Research Grant. Graduate students, PhD candidates, research scientists, post-doctoral fellows, and faculty at accredited universities are all invited to apply." Deadline for applications is January 9, 2020.

TOOLS & RESOURCES

OAI-PMH Service Updates

DataCite, Richard Hallett from November 25, 2019

"Our OAI-PMH service is one of the common ways we offer to harvest our public metadata, and we are launching a new version this Wednesday. This technology refresh allows us to continue supporting the OAI-PMH service. For the most part, there is no functional change, we adhere to the OAI-PMH standards and have attempted to keep the service as backward compatible as possible. The main change in the new service is that it uses our REST API instead of directly integrating with our Solr search index. The REST API uses our newer Elasticsearch-based search index under the hood, and with OAI-PMH being the last service still depending on Solr, this allows us to retire our Solr search index, completing the transition to Elasticsearch that started in 2018."

Introducing "The Loop": A Foundation in Listening

Stack Overflow, Sara Chipp and Juan Garza from November 25, 2019

"TLDR; We're going to be sharing our product development process with you, from feedback loops to timelines. We'll be doing so through our new series – The Loop. You can give us your thoughts on what you'd like to see us do by filling out this survey: Through the Loop. We'll also be releasing Moderator Training and some new feedback mechanisms to help us form decisions as we grow."

Altair: Declarative Visualization in Python — Altair 4.0.0 documentation

Jake VanderPlas from December 12, 2019

"Altair is a declarative statistical visualization library for Python, based on Vega and Vega-Lite, and the source is available on GitHub."

Analysis of Text-Analysis Syllabi: Building a Text-Analysis Syllabus Using Scaling

PS: Political Science & Politics journal; Nadjim Fréchet, Justin Savoie, Yannick Dufresne from November 29, 2019

"Text analysis is taught in most major universities; many have entire courses dedicated to the topic. This article offers a systematic review of 45 syllabi of text-analysis courses around the world. From these syllabi, we extracted data that allowed us to rank canonical sources and discuss the variety of software used in teaching. Furthermore, we argue that our empirical method for building a text-analysis syllabus could easily be extended to syllabi for other courses. For instance, scholars can use our technique to introduce their graduate students to the field of systematic reviews while improving the quality of their syllabi."

Google AI Blog: Fairness Indicators: Scalable Infrastructure for Fair ML Systems

Google AI Blog, Catherina Xu and Tulsee Doshi from December 11, 2019

"We recently released a beta version of Fairness Indicators, a suite of tools that enable regular computation and visualization of fairness metrics for binary and multi-class classification, helping teams take a first step towards identifying unjust impacts. Fairness Indicators can be used to generate metrics for transparency reporting, such as those used for model cards, to help developers make better decisions about how to deploy models responsibly."

CAREERS

Tenured and tenure track faculty positions

Tenure-Track

University of Pittsburgh, The School of Computing and Information (SCI); Pittsburgh, PA

Tenure Track Software Engineering & Data Science Faculty (Open-Rank)

University of St. Thomas, School of Engineering; St. Paul, MN

Chair and Professor with Tenure, Department of Computer and Data Sciences

Case Western Reserve University, Case School of Engineering; Cleveland, OH

Full-time, non-tenured academic positions

Reference & Instruction Librarian

University of Maryland-Baltimore County; Baltimore, MD

Appointment Stream, Teaching Faculty Positions

University of Pittsburgh, The School of Computing and Information (SCI); Pittsburgh, PA

Research Faculty Positions in Statistical Sciences (SDAD) - Biocomplexity

University of Virginia, Biocomplexity Center; Charlottesville, VA

Program Director, Social Sciences

University of Maryland, University College; Largo, MD

Assistant Research Professor

Pennsylvania State University, Social Science Research Institute; University Park, PA

Visiting Research Scientist

University of Illinois, School of Information Sciences; Champaign, IL

Senior Research Associate, ROSI, TM (Return on Sustainability Investment)

New York University, Leonard N. Stern School of Business: Center for Sustainable Business; New York, NY

Executive Director, Center for Information, Technology, and Public Life

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

Postdocs

Moore-Sloan Faculty Fellow.

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

Postdoctoral Scholar - Economic Modeling

University of California-Santa Barbara, Environmental Market Solutions Lab; Santa Barbara, CA

ETH- FDS Postdoctoral Fellow

ETH Zurich, ETH Foundations of Data Science; Zurich, Switzerland

Postdoctoral positions

Max Planck Institute for Social Anthropology; Halle, Germany

Post Doctoral Fellow in Data Science and Population Informatic

Texas A&M University, Population Informatics Lab; College Station, TX

Postdoctoral Research Associate (SDAD) - Biocomplexity

University of Virginia, Biocomplexity Center; Charlottesville, VA

Postdoctoral Fellowship Announcement

Stanford University, Meta-Research Innovation Center at Stanford (METRICS); Palo Alto, CA

Postdoctoral Positions

Johns Hopkins University, Department of Psychological & Brain Sciences; Baltimore, MD

Postdoctoral Fellowship in Medical Humanities

Rice University, Rice Academy Postdoctoral Fellowship; Houston, TX

Full-time positions outside academia

Research Software Engineer: Infrastructure and Development Operations

EcoHealth Alliance; New York, NY

Senior Machine Learning Engineer, NLP

Samsung NEXT; New York, NY

Vice President for Scholarly Programs

National Humanities Center; Research Triangle Park, NC

Bayesian Statistician, Research & Development

Houston Astros; Houston, TX

Data Scientist

American Civil Liberties Union, National Office; New York, NY

Associate Director

ASAPbio; Greater Boston, MA

Chief Privacy Officer

City & County of San Francisco; San Francisco, CA

Principal Researcher

Microsoft Research NYC; New York, NY

Director of Membership & Communications

MetroLab Network; Washington, DC

Computer Scientist

U.S. Department of Defense, Office of the Secretary of Defense; Arlington, VA

Backend Developer (full-time)

International Consortium of Investigative Journalists; Paris, France

Internships and other temporary positions**Google Research Intern, 2020**

Google; Research Lab locations across the U.S.

Data Science for the Public Good 2020 Recruitment

University of Virginia, Biocomplexity Institute and Initiative

Netflix Research Summer Internship

Netflix; Los Gatos, CA

Research Intern - Fairness, Accountability, Transparency, and Ethics in AI (FATE)

Microsoft Research NYC; New York, NY

Data Wrangling Project Contractors

Analytics Institute; Remote

Research Intern

Microsoft, Adaptive Systems and Interaction Group; Redmond, WA

GRADUATE STUDENT ASSISTANT

California State Water Resources Control Board; Sacramento, CA

Research Assistant to Global AI Narratives (Part Time, Fixed Term)

University of Cambridge, Leverhulme Centre for the Future of Intelligence (CFI); Cambridge, England

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 187.