Data Science Community Newsletter

25 November 2019 **Data Science Community Newsletter Issue 185**

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 Moore-Sloan Data Science Environment.

DATA SCIENCE NEWS

University Data Science News

Over one-third of graduate students sought help for anxiety or depression during graduate school, according to *Nature's* **recent survey** of 6300 graduate students around the world. A **similar survey** of 50,000 students in the UK found that 86% report marked levels of anxiety.

In the UK, it appears that problems that start in graduate school, continue deep into the career cycle of academics. Lecturers and professional staff across the UK began an 8-day long strike. The notoriously low-paid faculty and staff are **asking for better wages**, pay equity, improved working conditions and pensions. *The Guardian* has **more**.

NYU Law students **are organizing in support** of **Robert Jackson Jr** a professor on leave who is currently serving on the **Securities and Exchange Commission**. The students support his tenure at the SEC and do not want him to step down, as he is planning to do, fearing that other SEC commissioners will push harmful policies forward.

Stanford's Human-Centered AI group announced its **first major corporate partner** last week: **IBM**. The three major initiatives are: trusted artificial intelligence (e.g. typical AI ethics topics like fairness, robustness, explainability, and transparency), natural language processing, and neuro-symbolic computation. This appears to be a mutually beneficial partnership. The financial terms weren't revealed; I'm referring to the mutual intellectual benefits.

Duke University engineers have **developed a smart microscope** that can adjust its lighting conditions so that they are optimal for making specific diagnoses, including malarial diagnoses.

Yale University is going to **move a bunch of the data science 'usual suspects' disciplines into the same building**. The departments of Astronomy, Mathematics, and Statistics and Data Science, as well as parts of the Department of Physics will be stacked on top of one another in the Kline Tower. There will be a new institute, as yet unnamed, but likely to mimic the centers and institutes for data science we see elsewhere in academia.

University of Delaware is **leaning into FinTech**, planning to move part of its research effort into a new building in Delaware Technology Park, home of Delaware Biotechnology Institute and down the street from Newark High School. It will cost $38m, which is an indication of the strength of their interest (i.e., STRONG).

I almost always ignore conference summaries because they are boring. Summarizing well is harder than it looks, apparently. But this week this **IEEE Vis Conference** summary by **Michael Correll** is **included** because it is packed with easily digestible, salient take-aways. Priming participants with stories about a data visualization they're about to see is incredibly powerful; basically, the viewers will see what they've been told to see, and not much else. Also, it's important but rare to visualize uncertainty — error bars, box plots, halos — the data viz community may need to take a stand and adopt more complex visualizations that include uncertainty.

Speaking of data visualization, you may want to **check out** this year's winners of the *Information is Beautiful Awards*. The visualization about what the Swiss worry about tells you really everything you need to know about that country's attitude towards immigration and financial security.

Company Data Science News

Liquidata is a company that makes an **open source version-controlled git sql database product** called Dolt. In other words, it's the love child of git + sql, that allows datasets and databases to be forked, diffed, etc all quite seamlessly. I had been looking for something like this and thought I'd share in the hopes of solving a problem for others.

Google hires great people. I believe them to be consistently smart, broad-minded, creative, hard working, and cool under pressure, based on my personal experience with current and former Googlers. The company culture was generally open and fairly altruistic with lots of internal forums full of high quality advice on a surprising range of topics. When people who are smart, creative, hard working, broad minded, and good at thinking collectively organize themselves, they are formidable. Google's management is now running into sustained pushback from its talented, well-organized employees. Four employees who have been actively organizing to challenge company efforts **have been terminated in the past 10 days**, sparking accusations of union busting. Earlier this month, the company **trimmed its anyone-can-ask-anything egalitarian all-hands meetings** by prohibiting questions about company culture and moving from bi-weekly to monthly frequency. Googlers on **Twitter** confirmed that **employee organizing efforts** are global. Employment law, however, is local so it will be interesting to see how this plays out. At the moment, American Googlers appear to be most likely to be terminated or "re-orged" onto less favorable teams.

GenapSys is a biological hardware maker that produces an attractively priced gene sequencing tool. How attractive is the price? It's $10,000, which is about 1/100th the cost of existing systems. That's so low it could revolutionize the field and super charge certain applications of computational genetics. The company **closed a Series C round** of $90 million this week.

The New Yorker **profiles Silicon Valley renegade venture capitalist**, **Roger McNamee**. Actor/ comedian **Sacha Baron Cohen** recently joined McNamee by also **dunking on** **Facebook**.

Government Data Science News

The **National Institutes of Health** has **released** a **draft policy** outlining the agency's data sharing expectations. All researchers, not just those receiving $500,000 or more, will be expected to follow

privacy protocols for protecting subjects and share their data with the broader research community. Comments will be accepted through January 10, 2020. The agency expects to publish its final policy directive later in 2020.

Also at the **NIH**: A scandal involving **potentially unintentional fraud related to a loan repayment program** (LRP) unfolds. The program is designed "to keep promising young biomedical scientists in academic research by helping repay school loans that can run up to hundreds of thousands of dollars." The controversy is whether or not those enrolled in the program can also take money from big pharmaceutical companies for consulting, speaking engagements, or research partnerships. The rules have shifted around over the years - allowing some Pharma work at times and then banning it. One hundred eighty-two people have been identified as double-dipping by *Science*, though the cases have not be fully investigated yet.

Transparency is an NIH priority. **Noni Byrnes**, has been Director of **NIH Center for Scientific Research** since February. **She is profiled** by **Andrea Widener** in *Chemical & Engineering News*. “When you don’t know what’s going on, you’re much more likely to assume the worst,” she tells C&EN. “Any attempt we have to make the process transparent, to bring more people into our system, is going to help that.”

Elsewhere, the **Department of Energy's National Renewable Energy Lab announced a partnership** with **Hewlett Packard** to improve operational efficiencies at data centers.

It's not news that the state of **California** needs money. It is news when the **California Department of Motor Vehicles sells personal information** for revenue.

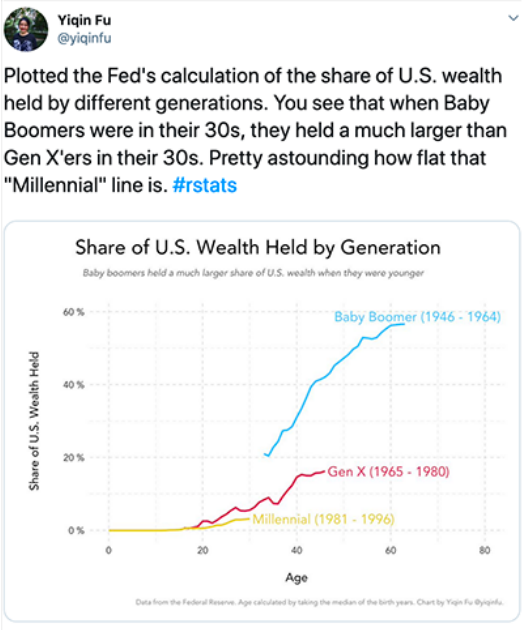
Extra Extra

I got excited about the idea of offsetting the carbon footprint of model-based testing and parameter tuning. When I **looked into carbon offsets**, I unsurprisingly discovered that some are openly fraudulent and many others aren't really offsets because the trees would have been planted anyway. I may, instead, look into contributions to the **Trust for Public Land** (they buy up land so that it cannot be developed, thereby preserving entire ecosystems).

This story — **The most useful app is Find My Friends** — is one that I am grateful for this Thanksgiving. It's about why people choose to use *Find My Friends* the app for stalking their friends (with consent), the anti-feminist history of geography, and several poignantly intimate uses of GPS.

Tweet of the Week

Twitter, Yiquin Fu from November 25, 2019



Data Visualization of the Week

Twitter, Arthur Charpentier from November 25, 2019

EVENTS

ML@GT Presents AI at Facebook Scale with Facebook VP of AI Jerome Pesenti

Atlanta, GA December 2 at 10:30 a.m., **Georgia Tech** Marcus Nanotechnology Building. [free]

TextXD: Text Analysis Across Domains

Berkeley, CA December 3-6. "The premier natural language processing conference at the **University of California, Berkeley**." [registration required]

Climate Resilient Smart Cities: Human-Technology Integration

Washington, DC December 3 starting at 9 a.m., National Academies Keck Building (500 5th St., NW). [registration required]

Challenges for Deep Reinforcement Learning in Complex Environments

Brooklyn, NY December 6, starting at 11 a.m. Speaker: **Raia Hadsell**, Head of Robotics Research at **DeepMind**. [free]

ACM FAT* - 2020 Registration

Barcelona, Spain January 27-30, 2020. "A computer science conference with a cross-disciplinary focus that brings together researchers and practitioners interested in fairness, accountability, and transparency in socio-technical systems." [$$$]

PyCon US

Pittsburgh, PA April 15-23, 2020. Registration is now open. [$$$]

Artificial Intelligence and Games

Copenhagen, Denmark June 22-26, 2020. "The summer school is dedicated to the uses of artificial intelligence (AI) techniques in and for games. After introductory lectures that explain the background and key techniques in AI and games, the school will introduce participants the uses of AI for playing games, for generating content for games, and for modeling players." [save the date]

Who We Are - Visualizing NYC by the Numbers

New York, NY Through September 20, 2020 at **Museum of the City of New York**. "In anticipation of the 2020 census, Who We Are: Visualizing NYC by the Numbers showcases work not just by data analysts and demographers, but also by cutting-edge contemporary artists and designers who use these tools to enliven and humanize statistics and to shed new light on how we understand our urban environment and ourselves." [$$]

DEADLINES

Contests/Award

NFL Big Data Bowl

"In this competition, you will develop a model to predict how many yards a team will gain on given rushing plays as they happen. You'll be provided game, play, and player-level data, including the position and speed of players as provided in the **NFL's** Next Gen Stats data. And the best part - you can see how your model performs from your living room, as the leaderboard will be updated week after week on the current season's game data as it plays out." Deadline for final submissions is November 27.

Education Opportunities

Google PhD Fellowship Program

The fellowships "directly support graduate students as they pursue their PhD, as well as connect them to a **Google** research mentor." Deadline to apply is November 30.

Santa Fe Institute - Research Experiences for Undergraduates

"A ten-week residential research opportunity in which students develop innovative research projects in collaboration with an **SFI** mentor." Deadline for applications is January 13, 2020.

High School Summer Internship Program (HS-SIP) 2019

"High School SIP (HS-SIP) provides an opportunity to spend a summer working at the **NIH** side-by-side with some of the leading scientists in the world, in an environment devoted exclusively to biomedical research." Deadline to apply is February 1, 2020.

Studies/Surveys

NIH's DRAFT Data Management and Sharing Policy: We Need to Hear From You!

"**NIH** has released for public comment in the Federal Register a Draft NIH Policy for Data Management and Sharing along with supplement draft guidance." ... "To facilitate public comments, NIH has established a web-portal where folks can easily and securely provide their feedback." Deadline for feedback submissions is January 10, 2020.

RFP

Dear Colleague Letter: Request for Information on Data-Focused Cyberinfrastructure Needed to Support Future Data-Intensive Science and Engineering Research

"This Request for Information (RFI) invites the community to provide input to **NSF** on specific data-intensive S&E research questions and challenges and the essential data-related CI services and capabilities needed to publish, discover, transport, manage and process data in secure, performant and scalable ways to enable that data-intensive research. Recognizing that data-oriented CI and services exist in many S&E disciplinary domains, NSF is particularly interested in understanding how broader cross-disciplinary and domain-agnostic solutions can be devised and implemented, along with the structural, functional and performance characteristics such cross-disciplinary solutions must possess." Deadline for submissions is December 16.

TOOLS & RESOURCES

Census Differential Privacy Exploration

Caliper Corporation from November 14, 2019

"To help people assess some of the implications and unintended consequences of Differential Privacy, **Caliper®** is providing several maps for public inspection. This map, created with Maptitude®, shows the change in population for every Congressional district after applying Differential Privacy."

List of Machine Learning and Deep Learning conferences in 2019 / 2020

Tryo Labs from November 25, 2019

"This list provides an overview with upcoming ML conferences and should help you decide which one to attend, sponsor or submit talks to."

Data Stories 150 | Highlights from IEEE VIS'19 with Tamara Munzner and Robert Kosara

Enrico Bertini and Moritz Stefaner from November 20, 2019

We have **Tamara Munzner** from the **University of British Columbia**, Vancouver, and **Robert Kosara** from **Tableau Research** on the show to go through some of our personal highlights from the **IEEE Visualization Conference 2019**." [audio, 1:02:01]

Release Notes (1.0.2) · Visual Coding - Neuropixels dataset

GitHub - AllenInstitute from October 14, 2019

"The 1.0.2 release brings support for our new Visual Coding - Neuropixels dataset. This is a large-scale survey of mouse subcortical and visual cortical regions using high-density Neuropixels probes."

Introducing ksqlDB

confluent from November 20, 2019

"KSQL as it has existed thus far has been about continuously transforming streams of data. It allows you to take existing Apache Kafka® topics and filter, process, and react to them to create new derived topics. These topics can represent pure events or updates to some keyed table that is being

materialized off the stream. For example, I could join together many sources of data I have about my customers to create a continually updated “unified customer profile.” KSQL continually processes the stream of incoming events and updates those materializations. Until now, KSQL has only had support for this kind of continuous, streaming query against its tables of data."

CAREERS

Tenured and tenure track faculty positions

Assistant, Associate, Full Professor

University of Washington, Information School; Seattle, WA

Machine Learning Algorithms and Applications

Michigan State University, Department of Computational Mathematics, Science and Engineering; Lansing, MI

Tenure-Track/Tenured Positions In Computer Science

Illinois Institute of Technology; Chicago, IL

Assistant Professors (Tenure Track) of Computer Science (Data Science)

ETH Zurich, Department of Computer Science; Zurich, Switzerland

Assistant Professor Sociology/IACS

Stony Brook University, Sociology Department; Stony Brook, NY

Tenure Track Faculty Position In Computer Science and Data Science

Vanderbilt University, Department of Electrical Engineering and Computer Science; Nashville, TN

Assistant Professor of Computer Science (Tenure Track)

University of Vermont, Department of Computer Science; Burlington, VT

Assistant/Associate Professor-Data Science

University of Massachusetts Amherst, College of Information and Computer Sciences; Amherst, MA

Full-time, non-tenured academic positions

LSST Alert Production Software Developer

Large Synoptic Survey Telescope (LSST) and University of Washington, Data Intensive Research in Astrophysics and Cosmology (DIRAC) Institute; Seattle, WA

CIG Research Scientist

University of Washington, EarthLab, Climate Impacts Group; Seattle, WA

group leader position

Institut du Cerveau et de la Moelle épinière (Brain and Spine Institute); Paris, France

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

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

Data Scientist, Media Manipulation

Harvard University, Kennedy School of Government; Cambridge, MA

Senior Software Developer

Indiana University, Cyberinfrastructure for Network Science Center; Bloomington, IN

Postdocs

Post Doc Researcher- Fairness, Accountability, Transparency, and Ethics (FATE) group

Microsoft Research; New York, NY

LSE Fellows in Computational Social Science (2)

London School of Economics, Department of Methodology; London, England

Postdoctoral Scholar

University of California-Berkeley, School of Information; Berkeley, CA

Postdoctoral Research Assistant

Queen Mary University of London, School of Mathematical Sciences; London, England

Janelia Theory Fellow Program

Howard Hughes Medical Institute, Janelia Research Campus; Ashburn, VA

HEP/Astro Experiment

Brandeis University, School of Physics; Waltham, MA

Post-Doctoral Research Fellow/Scientist

Columbia University, Data Science Institute; New York, NY

Full-time positions outside academia

Data Storyteller

The Pudding, Polygraph; New York, NY

Data Engineer

Zillow Group, StreetEasy; New York, NY

Internships and other temporary positions

Data Journalism Intern

Associated Press; Washington, DC

Data Journalist Intern

Two Sigma; New York, NY

Research Intern - Information and Data Sciences

Microsoft Research AI, Information and Data Sciences (IDEAS) group; Redmond, WA

2020 Summer Football Data Master's Intern

National Football League; New York, NY

MSc Project: Understanding Arctic Coastal Ecosystems

Carleton University, Department of Biology; Ottawa, ON, Canada

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 185.