Data Science Community Newsletter

22 November 2019 **Data Science Community Newsletter Issue 184**

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

Professor **Paul Dourish** of **University of California-Irvine** is a pre-eminent science and technology studies scholar who argues that the field of user experience design and human-computer interaction needs to **return to its roots**. He notes that the field's mission used to be to "nurture and sustain human dignity and flourishing" but that it has become more about delighting users individually, at times adopting "cavalier attitudes toward data privacy" and generally "reconfigur[ing] privacy and security as problems for individuals". He argues that centering *design* rather than centering humans, that focusing on user "delight" rather than broader moral commitments is the result of field drift. He challenges his field to "recover what has been lost". He makes a compelling set of points.

Joshua Grubbs at **Bowling Green State University** uses a philosophical concept — **moral grandstanding** — to explain much of the communication that happens on social media. "Moral grandstanding can take many forms, the researchers write, 'in a quest to impress peers, someone may trump up moral charges, pile on in cases of public shaming, announce that anyone who disagrees with them is obviously wrong, or exaggerate emotional displays in taking ideological positions.' They may also ramp up the situation, seeking to out-do others in their outrage." This perfectly describes what makes me so uncomfortable on twitter — the moral outrage is such a demanding communication style that often seems to make vanishingly little difference to the underlying concern. Readers, if you ever think I'm doing some moral grandstanding, please let me know.

The **National Center for Supercomputing Applications** (NCSA) at the **University of Illinois at Urbana-Champaign** has **opened a new Center for Artificial Intelligence Innovation** that will amplify and continue a lot of good work that is already happening in deep learning, machine learning, and AI at UIUC. I should note that UIUC has gotten less than the requisite amount of attention it deserves in this newsletter. It already had the **moderately large Illinois Data Science Initiative** so the NCSA is not their first foray into data science. And the **Discovery Partners Institute** program, a Chicago beachhead for innovation and entrepreneurship, expands the impact footprint for UIUC technology.

MIT, **USC**, **Duke University** and **The Cleveland Clinic** **each received $260 million** — the largest ever for both USC and the Cleveland Clinic — from the sale of the **Lord Corporation** to **Parker-Hannifin** for $3.68 billion. The unusual windfall results from Thomas Lord's 1982 decision to set aside

ownership shares in Lord Corp. to each of the four institutions. The Lord Corp. made products that "dampened noise and vibration, specialty adhesives" and other materials used in the automotive and aerospace industries. May every institution find a pot of gold at the end of a generous inventor's rainbow.

Julia Lane of **NYU**, **Paco Nathan** of **O'Reilly**, and **Ian Mulvany** of **SAGE Ocean** hosted **the Rich Context workshop** in DC last week. They brought together academia, industry, and top government officials to think about how we can all get access to richly contextualized datasets that make our research easier, more meaningful, and more impactful. Hopefully, there will be more from this group soon.

University of Pennsylvania received a $25 million donation from **Harlan M. Stone** **for a new Data Science Building**.

Columbia University, the **Flatiron Institute** (underwritten by **Simons Foundation**) in New York City and the **Max Planck Society** in Germany **will form** "the Center for Nonequilibrium Quantum Phenomena. The center aims to understand, control and manipulate the uniquely useful properties of quantum materials." They will work on quantum computing, cryptography, and "technologies not yet imaginable". Yes, that is a direct quote. I guess it's fairly common to prance around spouting hyperbolic prognostications about the future.

Company Data Science News

Thomson Reuters and **Lexis-Nexis** are **facing pressure** from lawyers, legal scholars, and law students who are angry that those companies' contracts with the **Department of Homeland Security** via **ICE** and **Palantir** are **enabling the deportation of immigrants** that the lawyers and legal scholars are working to protect. **Sarah Lamdan** of **CUNY School of Law** is quoted in *The Intercept* saying, "[l]awyers are funding the companies that are building ICE's surveillance system, which totally works against their clients...They're paying collectively millions of dollars (est. $54 million) to Thomson Reuters and Lexis every year, and then those companies are putting it into R&D, where they are creating products for ICE and law enforcement." If you are a lawyer, legal scholar, or law school student wishing to join the resistance against Lexis Nexis and Thomson Reuters on this issue, there is a petition **here**.

Harry Shum head of **Microsoft AI and Research**, **announced that he would be leaving the company** after an illustrious career. It sounds like he might end up in academia, but he made no specific announcement about what he would be doing next. Microsoft CTO **Kevin Scott** assumes Shum's position as head of Microsoft AI and Research effective immediately.

Microsoft is **opening a new hub for AI and data science work** in Louisville, Kentucky. They are starting quite modestly, by hiring four fellows to join executive **Ben Reno-Weber**. This micro-hub for AI is an interesting approach. Opening an outpost in a co-working space populated mostly by fellows is the corporate equivalent of a pop-up shop — it can be set up and torn down almost overnight.

Over the past couple weeks I wrote about how **Google's** purchase of **FitBit** and its **partnership with Ascension hospitals** increased the tech giant's access to precious health data that had previously been scant in the Google family of products. This week Google has **announced** it is moving into banking. A checking account service called **Cache** (funny funny) will be available soon. Google is on an apparently unstoppable quest that will result in a corporate entity holding the most complete full-person data available in human history.

Apple ran a **study** using the Apple watch to monitor heartbeats for possible episodes of atrial fibrillation. This monitoring mechanism seemed to work, providing a model for future health wearables projects.

Apple is calling for patients to participate in **three different longitudinal health studies**. One focusing on cardiovascular fitness requires participants to have an Apple watch. So that's totally a representative sample! There's another study on women's health that involves a monthly survey. A third study looks at hearing and I have no idea what kind of hardware is involved. While Apple is generally quite privacy protecting, I'm not entirely sure I understand their sampling procedure.

Fortune magazine wrote up **an annual survey of CIO's which** found that machine learning is becoming more widely used in business and that there's no sign of deceleration, though actual use is still low. "Almost 70% of CIOs say that AL or machine learning are used on only 1% to 20% of their companies' tech projects." Consider the strength of your job security, readers.

Government Data Science News

Senators Amy Klobuchar (D-MN) and **Lisa Murkowski (R-Alaska)** are looking at amending HIPAA so that **it covers health data collected from wearables**, which are currently out-of-scope. HIPAA coverage is determined largely by which entity gathered the data. If it's from a health care provider then it's covered. If it's not coming from a health care provider, then it's out of scope, no matter how similar or sensitive it may be. Expanding the entities that are considered HIPAA covered entities to include tech companies would usher in a host of changes that could go beyond wearable devices to include other kinds of health-related data.

The **US Patent and Trade Organization** is **accepting comments** on **the status of AI generated content**. Can AI be trained on copyrighted content (with or without paying royalties)? "Should a work produced by an AI algorithm or process, without the involvement of a natural person contributing expression to the resulting work, qualify as a work of authorship protectable under U.S. copyright law? Why or why not?" There are 18 total questions. The one I'm most interested in is: "How, if at all, does AI impact the need to protect databases and data sets? Are existing laws adequate to protect such data?" I'm not sure if current laws are strong enough to protect datasets, but I have no idea how to hold an AI accountable for hoarding data, using data it shouldn't have, or making biased decisions drawn from using problematic data. The ruling cannot come too soon. The first musical album recorded by an AI was released by **Holly Herndon**. The **composition details are all in the training process**, which took months.

Israel has a **severe shortage of educational capacity** to meet the demand for training in data science and AI. Israel has a thriving tech scene, but not enough university professors. And short, targeted coursework has failed to keep actual or would-be employees on top of their games.

NASA's Parker Solar Probe opened a **new trove** of data from our sun gathered on two fly-bys.

ARPA-E announced $15 million in funding for **23 projects** for machine learning and AI into energy technology.

The **National Science Foundation** is also **soliciting comments**. Writing in NSF-ese, the federal agency wants comments about "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". I'm sure some of you want to contribute, especially those who are planning to get consistent funding from NSF.

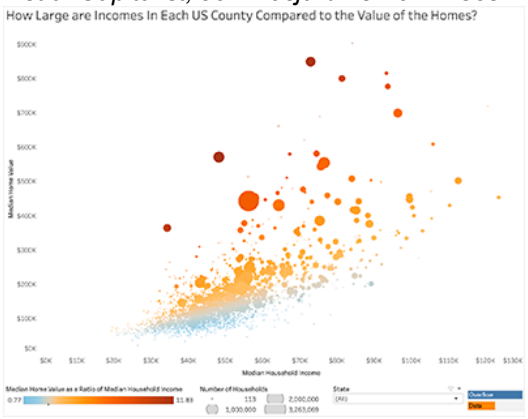
Extra Extra

This is an excellent **list of do's and don'ts for people who write about AI**, especially journalists. Here's my top two tips:

1. Don’t: imply autonomy where there is none
2. Don't: cite opinions of famous smart people who don’t work on AI

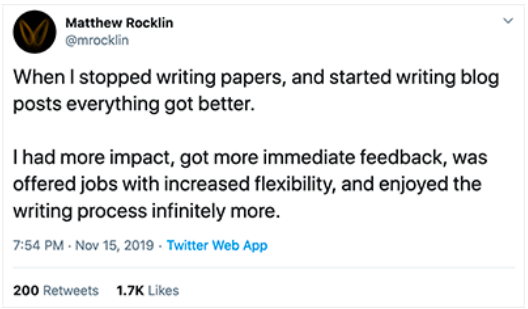
Data Visualization of the Week [Interactive]

Visual Capitalist, Jeff Desjardins from December 27, 2016



Tweet of the Week

Twitter, Matthew Rocklin from November 15, 2019



EVENTS

D2K Distinguished Lecture Series: Hadley Wickham; 2019 COPSS Presidents' Award Winner

Houston, TX November 22, starting at 3 p.m., **Rice University**. [free]

SoCal All Hands useR Meeting

Burbank, CA November 23, starting at 10 a.m. "At the end of 2019, let's get together and have an entire day of R fun at **Warner Bros** in Burbank. 2019 has been a great year for SoCal R users. Several R Users Groups and R-Ladies either started or restarted to serve all communities in Southern California." [free, registration required]

Women in Machine Learning Workshop at NeurIPS 2019

Vancouver, BC, Canada December 9, starting at 8 a.m. "This technical research workshop gives female faculty, research scientists, and graduate students in the machine learning community an opportunity to meet, network and exchange ideas, participate in career-focused panel discussions with senior women in industry and academia and learn from each other." [free, registration 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." [$$$]

Conference to Focus on AI in Arkansas

North Little Rock, AR "The sixth annual **Arkansas Bioinformatics Consortium** conference is set for Feb. 10-11 at the Wyndham Riverfront in North Little Rock. Its theme is 'Artificial Intelligence in Arkansas.'"

Women in Data Science (WiDS) 2020 at Stanford University

Stanford, CA March 2, 2020. "This one-day technical conference features amazing thought leaders in data science from academia, industry, non-profits, and government. Topics presented will cover a wide range of domains from data ethics and privacy, healthcare, data visualization, and more." [$$$]

DEADLINES

Contests/Award

Rainforest XPrize

"The winning team will survey the most biodiversity in at least three stories of a rainforest (emergent, canopy, understory, and forest floor) in 8 hours and use that data to produce the greatest number of

new insights after 48 hours." Deadline to submit materials for Guidelines Feedback is December 22.

Conferences

Call for Participation - C+J 2020

"Save the Date! C+J 2020 will be held at **Northeastern University** in Boston on March 20-21, 2020. Submission deadline Dec. 13th, 2019."

NYC Open Data Week

New York, NY February 28-March 7. "Open Data Week is a festival celebrating **NYC Open Data** as a free public resource about NYC — guided by the philosophy that open data is for all New Yorkers." Deadline for proposal submissions is December 13.

Education Opportunities

CDS Moore-Sloan Faculty Fellow

"The **Center for Data Science at New York University** invites applications for its CDS Moore-Sloan Faculty Fellow positions. Building on the successes with the Moore-Sloan Fellows program, CDS has created a Faculty Fellow program to continue to develop outstanding researchers in Data Science." Deadline to apply is December 23.

Princeton Center for Information Technology Policy: Call for Visitors 2020-21

"CITP seeks applicants for various visiting positions each year. Visitors are expected to live in or near Princeton and to be in residence at CITP on a daily basis. They will conduct research and participate actively in CITP's programs." Deadline for applications is December 31.

Applications for the 2020 Data Science for Social Good Fellowship at Carnegie Mellon University are now open - Apply to be a fellow, mentor, project manager,

"The Data Science for Social Good Fellowship is a full-time summer program to train aspiring data scientists to work on machine learning, data science, and AI projects with social impact in a fair and equitable manner." Deadline for applications is January 31, 2020.

Studies/Surveys

Take the 2019 HackerRank Developer Skills Survey

"What's the best place for developers to learn new skills and what new skills are they learning? How are engineering leaders hiring the developers they need?"

"These are some of the questions we want to learn more about in our survey and the insights we want to share with you."

RFP

The power of open data to transform and engage communities: a call for ideas

"**Knight Foundation** is issuing an open call for ideas that advance the concept of open data and civic engagement to encourage a new set of transformative approaches for using, understanding and taking action with public data. Selected recipients can earn a share of up to $1 million in funding for their ideas and projects." Deadline for proposals is December 13.

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

Practical Compositional Fairness: Understanding Fairness in Multi-Task ML Systems

DeepAI, Xuezhi Wang, et al. from November 05, 2019

"Most literature in fairness has focused on improving fairness with respect to one single model or one single objective. However, real-world machine learning systems are usually composed of many different components. Unfortunately, recent research has shown that even if each component is “fair,” the overall system can still be “unfair”. In this paper, we focus on how well fairness composes over multiple components in real systems."

How to tune a Decision Tree?

Towards Data Science, Mukesh Mithrakumar Mukesh Mithrakumar from November 11, 2019

"How do the hyperparameters for a decision tree affect your model and how do you choose which ones to tune?"

DataOps is more than just DevOps for data

SD Times, Christina Cardoza from November 13, 2019

"Development, testing, security and operations have all been transformed to keep up with the pace of software today — but one piece is still missing. Data is now becoming a roadblock to Agile and DevOps initiatives."

Kaolin: A PyTorch Library for Accelerating 3D Deep Learning Research

arXiv, Computer Science > Computer Vision and Pattern Recognition, Krishna Murthy et al. from November 13, 2019

"We present Kaolin, a PyTorch library aiming to accelerate 3D deep learning research. Kaolin provides efficient implementations of differentiable 3D modules for use in deep learning systems. With functionality to load and preprocess several popular 3D datasets, and native functions to manipulate meshes, pointclouds, signed distance functions, and voxel grids, Kaolin mitigates the need to write wasteful boilerplate code. Kaolin packages together several differentiable graphics modules including rendering, lighting, shading, and view warping."

How to Lock Down Your Health and Fitness Data

WIRED, Security, David Nield from November 17, 2019

"Whether you're a Fitbit user worried about Google's recent $2.1 billion purchase of the company or just generally privacy conscious, you should pay attention to where your health and fitness data goes and who has access. It's among the most sensitive data you have." ... "It shouldn't take long, and it follows the same principles as any other data privacy audit: Check which data is being collected, which parts of it are public, and how many of your apps can access it."

Gen: a general-purpose probabilistic programming system with programmable inference

built on Julia .ical Feedback

The Julia Language, Marco Cusumano-Towner from July 25, 2019

"This talk introduces a new flexible and extensible probabilistic programming system called Gen, that is built on top of Julia. Gen's extensible set of modeling DSLs can express probabilistic models that combine Bayesian networks, black box simulators, deep learning, structure learning, and Bayesian nonparametrics; and Gen's inference library supports custom algorithms that combine Markov chain Monte Carlo, particle filtering, variational inference, and numerical optimization." [**GitHub source**]

CAREERS

Tenured and tenure track faculty positions

Assistant, Associate, Full Professor

University of Washington, Information School, Center for an Informed Public; Seattle, WA

Open Positions in CS (2)

University of California-Santa Barbara, Department of Computer Science; Santa Barbara, CA

Sociotechnical Assistant/Associate/Full Professor (Open Rank)

University of Maryland, Department of Criminology and Criminal Justice and the College of Information Studies; College Park, MD

Assistant Professor - Data Science Ethics

University of California-San Diego, Halıcıoğlu Data Science Institute; La Jolla, CA

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

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

Research Scientist/Engineer 3

University of Washington, Data Intensive Research in Astrophysics and Cosmology (DIRAC) Institute; Seattle, WA

Life Science Research Professional 2

Stanford University, Natural Capital Project; Palo Alto, CA

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

Postdocs

Data Science and Applied AI Postdoctoral Scholars Program

University of Chicago, Center for Data and Computing and the Center for Applied AI; Chicago, IL

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

Full-time positions outside academia

Computer Scientist – Machine Learning

Argonne National Laboratory, Leadership Computing Facility; Lemont, IL

Data Storyteller

The Pudding, Polygraph; New York, NY

Data Engineer

Zillow Group, StreetEasy; New York, NY

Internships and other temporary positions

Research Assistant, Social Instabilities in Labor Futures

Data & Society Research Institute; New York, NY

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

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 184.