Data Science Community Newsletter

22 December 2019 **Data Science Community Newsletter Issue 188**

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.

Dear readers, we have almost come to the end of another year. This is the final regular newsletter for 2019. Next week we'll have a special issue that looks back on what we've covered, how we've grown, and tries to contextualize the topics that came up over and over again.

But first, thank you for being part of the data science community, for doing science thoughtfully, and for bringing your whole selves to the table to advocate for ethical, inclusive practices. We are absolutely impressed at the advances, the thoughtfulness, and the fortitude to carry on in this emerging field. Our appreciation runs deep for the past year of great work that you have undertaken and seen through.

Here's to doing our collective best to building a beneficient data science community,

-Brad and Laura

DATA SCIENCE NEWS

University Data Science News

Several pieces of open access news this week:

Joelle Pineau of **Facebook Artificial Intelligence Research** in Montreal and **McGill University** is pushing to make sure **AI research avoids the reproducibility crises** other disciplines have confronted. This means code sharing and reproducibility tests as well as a reporting checklist specific to machine learning.

Sixty professional societies including the **American Dental Association**, the **American Chemical Society**, the **American Sociological Association** have co-signed **a letter** to the Trump administration supporting a 12-month embargo on open access to research published using tax dollars. They argued that shortening the embargo to make research available more quickly would, "undermine cooperative efforts to address these bigger, higher priorities, and risks the continued international leadership for the U.S. scientific enterprise." Many scholars have expressed **consternation** at the letter and favor more immediate open access.

The **U.S. Congress** renewed and strengthened funding for the **Open Textbook Pilot Grant Program with $7 million** for 2020. It's a drop in the bucket when spread across the entire country, but a drop in the right end of the bucket. Making textbooks free for students to use requires fairly small upfront allotment for savings that impact thousands of students.

In a **thematically similar, but unrelated report**, the **Scholarly Publishing and Academic Resources Coalition** (SPARC), outlines the infrastructural and procurement process changes that

need to happen in order for academic institutions to take meaningful guardianship over their data. The concern is that third party vendors are rushing into the space before guidelines about what kind of data can be passed to third parties, whether it's long-term cheaper to build data infrastructures in-house, what the data retention policies ought to be, and whether there's a role for open access to some kinds of typical academic data. It's a thoughtful piece of work that I hope many data librarians, data management offices, and admissions offices will read (admissions is already invested in vendor relationships involving sensitive data and impactful algorithms).

Let's not forget Open Source software. **NCEAS** researcher and **Openscapes** founder, **Julia Stewart Lowndes** reminds us in her *Scientific American* **blog essay** that "Open Software means Kinder Science."

A new report by **RWTH Aachen University** suggests that **5G is going to put an alarming energy demand on data centers**. In just five years, the energy burden associated with the switch to 5G could equal the annual energy consumption of three mid-sized German cities: Cologne, Dusseldorf, and Dortmund.

University of Illinois at Chicago pulled in a **$1.5 million grant** from the **NSF** to start a Foundations of Data Science Institute. It will focus on "the representation and structure of data, machine learning and complexity, and robustness and privacy."

University of California, Santa Cruz established a **masters degree in natural language processing**. It will be based at the school's Silicon Valley campus.

Australian Computer Society has a **new Artificial Intelligence Hub** in Melbourne.

Washington University in St. Louis is starting a PhD in a new **Division of Computational and Data Sciences**. Their data science center focuses on methods, psych, brain science, political science, social work, and public health.

Notre Dame University **received $25 million to consolidate and expand on all their existing efforts in data science** with the Lucy Family Institute for Data & Society.

Science declared the **first image of a black hole** to be the biggest scientific breakthrough of 2019.

Canadian physicist **Robert Walkow** and doctoral candidate **Roshan Achal** and colleagues presented on a **major breakthrough in the materials engineering aspect of chip hardware** that could lead to the production of computers that are 100 times more energy efficient. This cannot commercialize soon enough - the carbon footprint of computing is already quite large and growing rapidly under pressure from 5G networks and training machine learning models.

The **American Community Survey**, the most important national survey besides the Census, came out with a new release last week. The **main takeaway for higher education** is that there are

250,000 fewer people enrolled in courses than there were just one year ago. The decline is linked to high tuition, fewer high school graduates, and the strength of the economy - a strong economy tends to correlate with lower enrollments. This means (more) financial strife for tuition-driven schools.

Company Data Science News

Apple has a team developing a **satellite network to provide wireless telephony to mobile phones**, bypassing typical networks. Similar projects have been attempted by companies like **Facebook**, **Iridium** (now part of **Garmin**) and **Globalstar**, but the route to profitability was never clear. Apple is generally successful at turning a profit, so this time, this project, might be different.

Facebook is **providing the funding** for a free online course designed for journalists called **"Identifying and Tackling Manipulated Media."** In other words: how to fight misinformation and deepfakes. Even though it's designed for journalists, I think everyone could benefit from learning how to identify photos or videos that have been altered.

The **tech fail of the week** comes not from one of the Big Four, but from **Knightscope**, a company that makes a robotic security cop that glides around Huntington Park, a neighborhood in Southeast Los Angeles. When a fight broke out, a bystander pushed the button on the robot to initiate a call to 911, but it told her to "stand aside" and continued on its preprogrammed route. The fight continued and eventually a woman was removed on stretcher with a head laceration. The company explained that the 911 feature isn't active. How any reasonable person could figure out that the buttons were inoperable continues to escape me.

Seventy-five percent of the Big Four (**Google**, **Amazon**, **Apple**) are **supporting a common standard for smart home applications** called Connected Home over IP (CHIP). IoT industry watcher **Stacy Higginbotham**, who has "long blamed the sad state of the smart home on a lack of a standard," is optimistic, "It will be good for both consumers and developers."

Google has **fired another worker** who had made a pop-up (shown only to other Googlers) that explained their labor rights. It said nothing more than what companies are legally obligated to post about labor rights, usually they do it on a big laminated poster near the coffee maker. She did it by using Google tools to create a popup, as one would expect an employee to do, but was subsequently told her use of the tools was inappropriate. Googlers commenting on Twitter seemed to side with the fired worker, at least in the sense that they didn't understand the company's explanation of its own acceptable use policy.

It's been a stellar December for New York City NLP startup **Hugging Face**. The company's NeurIPS poster describing **DistilBERT**, a lightweight BERT for edge devices, was a hit. The company followed up with a **$15 million dollar funding round**.

Government Data Science News

The **US Census Bureau** has **established a Trust and Safety team** to prevent the spread of misinformation and disinformation about the US Census. **Facebook** has **agreed to combat the spread of misinformation** about the Census on its platform.

For the first time in more than two decades, **the US federal government will make research funding available to study gun violence**. **Centers for Disease Control and Prevention** and the **National Institutes of Health** have $25m available, split evenly, to fund the systematic study of gun violence. (How many people had to die before this funding became available?)

NOAA has turned an experiment in **providing data about clouds, in the cloud**, into a durable ongoing project that makes climate data available at no cost to the public. The three main cloud providers, **AWS**, **Google Cloud**, and **Microsoft Azure**, are all participating partners. It's extremely important to make our best climate data available for free, to everyone. The climate crisis is the biggest crisis our planet has ever faced and sometimes it seems like only scientists are rising to confront the challenge.

Extra Extra

This is the article you should open in a tab and come back to in the quiet days between Christmas and New Years. It's absolutely the one to read as America heads into an election year.

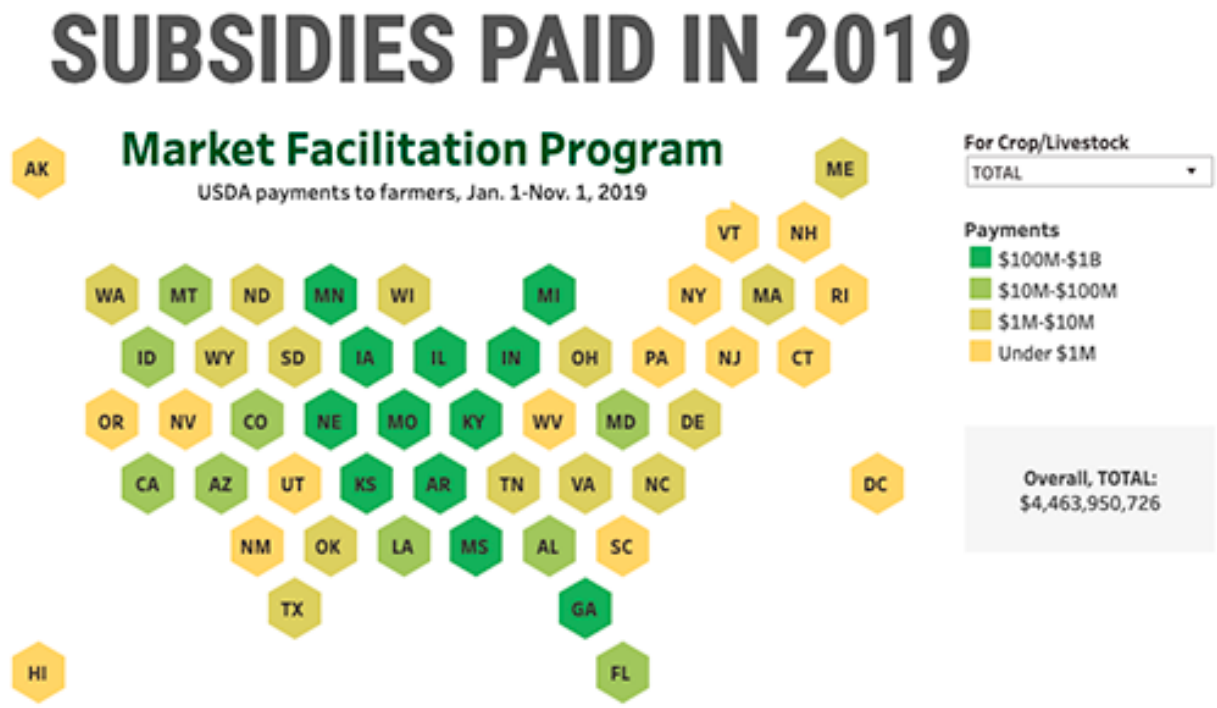
This essay by **David Karpf**, **George Washington University** Associate Professor of Media and Public Affairs, is perhaps the most important, thoughtful, relevant piece I have read about the intersection of data science, misinformation, and American political culture. He reminds us to be skeptical about the impact of misinformation campaigns, for which we do *not* have good evidence. His analysis is smart, but the real takeaway is for us, the class of people who might contribute to the hype cycle around data science. Given that it's not entirely clear if misinformation changes the way people vote or participate in political discourse, it's important to consider the second-order impacts of people like us, people who know better than to contribute to hype cycles in the absence of rigorous evidence, treating misinformation as a powerful scourge on democracy. Karpf reminds us that Americans have never been a particularly informed *vox populous*, but the way we spin our democratic myth rests on a belief that we are — or at least *were* — well-informed participants in the democratic process. If we, the people trained well enough to evaluate the first order impacts and identify second order consequences fail to do so, then powerful forces may shift the myth of American democratic participation in ways that become self-fulfilling. The non-contribution to election discourse undermines the very notion that everyday voters matter without shying away from the fact that everyday voters in America have always been uninformed.

Tweet of the Week

Twitter, NewStatistics from December 20, 2019

Data Visualization of the Week

Twitter, APM Research Lab from December 16, 2019



EVENTS

BostonCHI - Jared Spool 2020: The Four horsemen of the Quantitative UX Metrics

Burlington, MA Jan 16, 2020, starting at 6:30 p.m., **Microsoft** (5 Wayside Road). [registration required]

Announcing the 3rd Annual WiDS Datathon

Online "This year's datathon will begin in January 2020 and last until February 24, 2020. Winners will be announced at the WiDS Conference at **Stanford University** and via livestream. Sign up now to participate, and we will send you the link when the competition begins!" [save the date]

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." [$$$]

2020 FCSM Research and Policy Conference

Washington, DC April 14-16, 2020, Washington DC Convention Center. "The 2020 **Federal Committee on Statistical Methodology** Research and Policy Conference will focus on the **Federal Statistical System's** role in equipping agencies and the public to leverage data resources for evidence-based policymaking." [$$$]

Tutorials - ACL 2020

Seattle, WA July 5, 2020. [save the date]

DEADLINES

Contests/Award

Nominations Open for 2020 SAGE-CASBS Award

The award "recognizes outstanding achievement in the behavioral and social sciences that advance our understanding of pressing social issues. The award underscores the role of the social and behavioral sciences in enriching and enhancing public policy and good governance." Deadline for nominations is March 16, 2020.

Conferences

Call for submissions for CarpentryCon 2020 – and for your t-shirt designs

Madison, WI June 29-July 1, 2020, at **University of Wisconsin**. "Programme submissions are currently being requested for **CarpentryCon**, which will be held in Wisconsin from 29 June to 1 July 2020. Talented (and otherwise!) designers are also encouraged to submit their design ideas for the conference t-shirt."

Call For System Demonstrations - ACL 2020

Seattle, WA July 6-8. "Submissions may range from early research prototypes to mature production-ready systems. Of particular interest are publicly available open-source or open-access systems. We additionally strongly encourage demonstrations of industrial systems that are technologically innovative given the current state of the art of theory and applied research in computational linguistics." Deadline for submissions is January 31, 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

Now accepting fellowship applications: Solutions Reporting on Health Interventions (Deadline: Jan. 19)

"Over the next year, [**Solutions Journalism**] will support 10 journalists in reporting on solutions from around the world that could help U.S. communities improve their health and well-being." Deadline to apply is January 19, 2020.

TOOLS & RESOURCES

Zwicky Transient Facility - Public Data Release 2

Zwicky Transient Facility from December 11, 2019

"The **Zwicky Transient Facility** (ZTF) and **IPAC at the California Institute of Technology** announce the second ZTF Public Data Release. ZTF is an optical time-domain survey covering the northern sky visible from **Palomar Observatory**. This release builds upon the first data release to include products from (i) an additional 6 months of survey operations from the public portion of the survey, giving a total observation span of March 2018-June 2019, and (ii) data acquired under private survey time during the first ~3.4 months of the survey, spanning March 2018-June 2018."

Caselaw Access Project

Harvard Library Innovation Project from October 31, 2018

"The Caselaw Access Project (CAP) expands public access to U.S. law. Our goal is to make all

published U.S. court decisions freely available to the public online, in a consistent format, digitized from the collection of the **Harvard Law Library**."

FastMRI initiative releases neuroimaging data set

Facebook Artificial Intelligence; Tullie Murrell, Anuroop Sriram, Nafissa Yakubova, Larry Zitnick from December 12, 2019

"FastMRI, a joint research collaboration between **Facebook AI** and **NYU Langone Health** to use AI to speed up magnetic resonance imaging (MRI) scans, is announcing a new open source data set from NYU Langone Health, along with baseline models and a newly expanded research paper to help the AI research community accelerate and broaden research in this area."

Create agents that monitor and act on your behalf. Your agents are standing by!

GitHub - huginn from December 08, 2019

"Huginn is a system for building agents that perform automated tasks for you online. They can read the web, watch for events, and take actions on your behalf. Huginn's Agents create and consume events, propagating them along a directed graph."

APIs' reach expanding beyond developers, survey shows

ZDNet, Joe McKendrick from December 12, 2019

"Survey of 10,000 managers and professionals finds more non-developers working with APIs. Small teams, internal APIs are more the rule."

How To Run Docker Images and Make Your Own! Docker for Data Science/Python

Medium, Aaron Abrahamson from December 19, 2019

"This is a followup to my previous post, What is Docker, and why is it useful for data science? In this post I am going to show you how to setup Docker on your machine, and create a Docker Image from a Dockerfile, and then how to get a Docker Container running PostgreSQL up and running."

Google Cloud Platform (GCP) Security Best Practices

Assured, Patrik from December 19, 2019

"The intention of this blog post is to make a walk-through of a couple of GCP's features and give security recommendations and advice on how to configure your GCP environments."

CAREERS

Tenured and tenure track faculty positions

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

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

Postdocs

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

Postdoctoral research position on survey research with us at Columbia School of Social Work

Columbia University, Columbia Population Research Center; New York, NY

Postdoctoral Researcher

University of Oxford, Oxford Internet Institute; Oxford, England

Postdoctoral Research Fellow

North Carolina State University, Department of Computer Science, Innovative Educational Computing laboratory; Raleigh, NC

Postdoctoral Researcher in High Resolution Spectroscopy

University of Michigan, Astronomy; Ann Arbor, MI

Postdoctoral Researcher in Exoplanets and their Host Stars

University of Michigan, Astronomy; Ann Arbor, MI

Postdoctoral Fellow in Nature & Human Health

University of Vermont, Gund Institute for the Environment; Burlington, VA

Full-time positions outside academia

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

Research Associate, Social and Demographic Trends

Pew Research Center; Washington, DC

SQL Developer and Stats Analyst

National Hockey League; New York, NY

Statistician

U.S. Department of Homeland Security; Washington, DC

Internships and other temporary positions

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

2020 Data Science Internship, Informatics and Analytics

Dana-Farber Cancer Institute; Boston, MA

Summer Research Program

New York University, Center for Urban Science + Progress; Brooklyn, NY

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 188.