ADSA Data Science Community Newsletter

Data Science Community Newsletter features journalism, research papers and tools/software for December 31, 2021.

Please let us (Micaela Parker, Catherine Cramer, Brad Stenger, Laura Norén) know if you have something to add to next week's newsletter. We are grateful for the generous financial support from the Academic Data Science Alliance.

DEMAND FOR COMPUTER SCIENCE FACULTY INCREASING; DATA SCIENCE/AI LEAD DEMAND

Continuing a trend that was briefly interrupted by the pandemic, students deciding to major in computer science (CS) and to take computer science courses continued to push class sizes up. **Cal Poly San Luis Obispo** was able to enroll just **200 out of the 6000 students** who applied to the CS major. It also fueled demand for additional faculty lines, according to a **study** by **Craig Willis** of **Worcester Polytechnic Institute**:

“We analyzed ads from 400 institutions seeking to fill hundreds of tenure-track faculty positions in Computer Science," Willis wrote on the **Computing Research Association** website. "This number

is a 70% increase from last year at this time (mid-November) and is a comparable number to the 394 institutions searching for 2020. The number of tenure-track positions sought is doubled from last year and up 6% from two years ago indicating a recovery in demand after a one-year drop due to the pandemic. The number of BS/BA institutions seeking faculty is at an eight-year high with top PhD and private PhD institutions at eight-year highs in the number of positions being sought."

Data science and AI were the most sought-after specialties, "aggregating the Data Science, AI/DM/ML and Databases clusters again resulted in close to a third of all hires sought." Though not addressed in the Willis study, at the DSCN we saw more postings joint-hires in data science. These are tenure-track lines in which the faculty hire will have ~50% of a tenure line held by the data science institute/center and the other half held in conjunction with a domain-specific department such as political science, physics, sociology, biology, etc. We expect these joint-hires to slowly become more prevalent, especially after the first cohort successfully navigates the process of becoming tenured. It's trickier with two departments involved.

AI IN POLICING — CONTESTED TERRAIN

The past year has seen a major reversal in one social trend: crime had been dropping for years but skyrocketed across the US in 2021. **Carjackings**, **homicides**, **robberies** — all up. This introduced incredible social pressure on local mayors to do something, raising the pressure on questions about the role of AI in policing. Using predictive analytics to guide enforcement is a practice dating back to the 1990s that was supposed to stretch small police budgets by getting the greatest attention where crime was highest. This led to over policing of minority communities and an increased arrest rate of black and brown people, especially for drug related crimes. Police and court system budgets continue to under deliver on safety, especially the protection of minority communities, leading cities like Minneapolis to put the status quo on the ballot this year. Going forward, the city will be determining just how to provide for community safety and could consider using more technology, though not facial recognition, which was already banned.

At the very top, the **U.S. Department of Defense** introduced a new position – **Chief Digital and Artificial Intelligence Officer**. The DoD also released new **guidelines** for the responsible use of artificial intelligence. These moves set a certain tone across the military that is largely tangential to domestic policing.

The **City of Chicago** found **problems** with the popular, expensive, AI-run ShotSpotter program. It over-polices minority communities and has led to the false arrests and imprisonment of black men. The AI here is not the flashy headline grabbing type — no facial recognition or risk profiling. It's used to predict whether a bang was a gunshot or a similar-sounding but innocuous bang from a car backfiring, fireworks, etc. Of course, getting it wrong can send police out to the neighborhoods where the microphones are (guess where they are usually installed?), and can then increase the arrest rate in those areas, even without an increased prevalence of gun fire.

As we enter the local budgeting season, expect to see more calls for the use of tech-assisted policing tools to combat increases in crime. Be ever so wary of the total design and implementation of these systems. In the 40+ year history of tech-enabled policing, very few technologies have reduced racial bias and many have increased it.

Featured Jobs

See the ADSA Jobs Page for more opportunities.

THE UNIVERSITY OF CHICAGO

Postdoctoral Scholar University of Chicago - Data Science Institute. Chicago, IL.
Preceptor in Data Science, University of Chicago - Data Science Institute. Chicago, IL.

MAJOR DEVELOPMENTS IN APPLIED AI: PROTEIN FOLDING AND NATURAL LANGUAGE PROCESSING

In terms of the biggest research developments over the past year in applied artificial intelligence, we feel that the major advances in proteomics — specifically protein folding — and the continued advances in natural language processing are the two stand-outs. We covered the parallel course between **DeepMind** (**AlphaFold 2**) and a group of academics led by a team at the **University of Washington** as they developed models to predict how a protein will fold based on DNA. Both teams produced excellent models and both have shared those models publicly. This is a huge win for science and will lead to more cheap and quick drug discovery cycles which should ultimately reduce human suffering. Bravo to everyone who worked on and continues to work on these projects.

Our other selection for biggest development in AI is also shared across teams working on the same project: natural language processing. In this case, **Google**, **Microsoft + NVidia**, **Alibaba**, and **DeepMind** **all** produced new models in 2021. Notably, it can be difficult for academic research labs to compete directly in this space, but there are many academic collaborations under the surface. Google released a **trillion parameter NLP model** to kick off the year. Microsoft and NVidia jointly released **Megatron** in October. DeepMind got into the game with a smaller, but in some ways more accurate model called **Gopher**. The vast amount of enthusiasm, energy, and resources dedicated to the problem of natural language processing is not without ethical considerations. See the post at the DeepMind link above for more there.

What is likely to be unlocked by at-scale deployment of these models are things like: better real-time content moderation to scrape some of the toxic goo off the internet before anyone sees it; the ability to translate accurately between languages so we can all access more information generated by people who don't talk the way we do; the ability to have machines read and summarize text we don't have time to read; the ability for people with sight or hearing problems to participate fully. There will be mistakes and there will be bias in these models. The teams working on them seem reasonably perceptive to breakage and funny model behavior at the margins, so we are confident that net benefits will be widespread and that languages, people with certain accents and/or speech impediments, and people who use a lot of unique slang/phraseology will eventually be included, even if they aren't initially seeing all their needs met.

Major kudos to the entire field of Natural Language Processing for your achievements.

WEARABLE AND QUANTIFIED

We process thousands of sources to bring you DSCN: **Twitter** streams, rss feeds from news and blogs, Google news alerts. Our choice for under-the-radar "feed of the year" in 2021 is the ACS Sensors Twitter account (**https://twitter.com/ACS_Sensors**). *ACS Sensors* is a relatively new journal (founded 2016) with a mission to cover "all aspects of chemical and biological sensing" in its pages, and fortunately for us, with its Twitter feed. According to journal editor-in-chief **J. Justin Gooding**, the pandemic **has raised awareness** of sensing technology as a crucial aspect for rapid diagnosis and for consumer eHealth applications. And sensor data capture for personal health has investors'

attention. **Northwestern University** used $75 million in state grants and alumni gifts to **establish a startup incubator** for digital sensing materials, supporting **John Rogers** and **Mark Hersam**, two faculty entrepreneurs, on day one. Rogers has pioneered development of stretchable silicon for wearable sensors, and one of his companies, **Rheos**, just received **a $2.2 million seed investment** for wearable sensors that monitor cerebral spinal fluid buildup. Hersam has synthesized **bi-layer borophene**, a wonder nano-material with properties similar to graphene but made from boron instead of carbon, with breakthrough potential for energy and sensor applications.

Michael Snyder, the **Stanford University** computational biologist, closed out a productive 2021 with a paper **describing an algorithm** which "reads heart rate as a proxy for physiological or mental stress," potentially indicating illness days before symptoms appear. The algorithm was unable to differentiate between stressors and requires individual contextualization — one person gets stress-alerted for running a marathon, another for traveling across country, another for a drunken bender. Still, according to Snyder, the algorithm was able to detect 80% of Covid-19 cases before or at onset of symptoms. Snyder and a long list of collaborators also made headlines during calendar 2021 for important big data studies of **the gut microbiome** and **wearable glucose monitors**, for **prediction models of personal blood tests**, for a **large-scale study of college athletes** and for open-sourcing a **Personal Health Dashboard** platform that collects and stores individuals' wearables and genomics data for analysis. Snyder compares what he's doing to **a jigsaw puzzle**, "You really don't stand a very good chance from 5 pieces if it's a 1,000-piece puzzle. When you put all 1,000 pieces there, you have a pretty good idea what that picture is." Help Snyder hone his algorithms and grow his data sets by enrolling in a study at **https://innovations.stanford.edu/wearables**.

The consumer experience of eHealth wearable technology and big data fits into two design patterns: a project manager for accomplishing wellness and fitness goals, and a recommender system that diverts your time and attention from work, video games and The Mandelorian. **University of Pennsylvania** social scientist **Katy Milkman** and 29 behavioral science colleagues **undertook a "megastudy"** with 61,293 members of **24 Hour Fitness**, a national chain of work out centers. Milkman **told** **CNN**, "If people are hoping to boost their physical activity or change their health behaviors, there are very low-cost behavioral insights that can be built into programs to help them achieve greater success," like the 53 different tactics for improving individual gym attendance that the study tested. Even though investors are funding research and startups geared towards wearables and wellness, the business models are not rock solid. Who pays for preventive health improvement in the general population? Is it government, employers, insurers or Jane and Joe Sixpack (they want it all — beers and abs!)? And can the success of individuals on these platforms scale to groups? Does a winning soccer team or a more productive satellite office justify thousands of dollars in cloud-based software as a service? No one is telling these companies that consumers' data cannot become alternative revenue streams. Because every consumer wearable derives most of its utility from the progress dashboards on its app those consumers face some onerous work, if not outright lock-in, if they desire a switch between brands. Textile manufacturers are **embroidering bio-sensors into workout benches**, clearing the way for the 2022 "super-megastudy" that answers everything.

tibi iorga is a UK-based med-tech product manager. She can see a distant future where **users control their data and are accountable to themselves**, rather than ceding choice architectures to auto-recommenders and accountability to health or training project plans. But for now, she writes, you can buy in or you can choose to be a couch potato.

Featured Events

See the ADSA Events Page for more details and more opportunities.

Draper Data Science Business Plan Competition. Open to students anywhere in the world. Info Session: 18 Jan 2022. Submission deadline: 31 Jan 2022. Event: April 13-15, 2022. 1st prize — $50,000 investment.

Call for Nominations: Future Leaders Summit 2022
Attendance fully funded.
Dates: Nominations due 21 Jan 2022; Event 6-7 April 2022.
Eligibility: US-based grad students and postdocs with research on responsible data science and AI.

RESEARCH ACCESS TO BIG TECH'S DATA

About a decade ago, it was common for academic researchers to officially or unofficially use data from companies like **Google**, **Twitter**, **eBay**, **Flickr**, **Foursquare** and **Facebook** for research purposes. The **Cambridge Analytica** case, ensuing fines, and discursive retribution from the academic community marked a clear turning point in relationships that had already begun to become more restrictive. This year, Facebook launched its second post-Cambridge Analytica program for data sharing with researchers. But! Facebook was also found to ignore or downplay research undertaken by its own research team (See **Frances Haugen's testimony**) and shut down an ongoing political **Ad Observer** research project by **Laura Edelson** at **NYU**. In this case, users had explicitly consented to the terms of the project, so typical objections to passive research do not apply.

These stories are part of a larger unfolding story that is of great interest to DSCN readers. What kind of research can be done from within tech companies? Will it make a difference to users? If readers stay in academia and want to study the impact of large platforms on things like adolescents and eating disorders, school shootings, political polarization, trafficking in protected species, the influence of online ad-spend, and so many other valid research questions — how should they expect to proceed? What's the best way to set up their careers? Their specific projects?

As it stands, there are still APIs available for research use (often with a fee):

+ **Twitter** has **an API for academic research**. Gold Star partner.
+ **LinkedIn API**
+ **Yelp API**
+ **Google** has **240 developer APIs** including several for **YouTube**
+ **Flickr API** (**remember** Flickr?)

But companies like Foursquare and Facebook that used to make their data more widely available have become less accessible to researchers, often to protect users' privacy. It is clear that these privacy concerns are valid in some, but not all, cases. Generally speaking, it would be nice if there were a way to obtain user consent for research use of various types of user-generated data designed and operated at

scale by an independent organization like the **Social Science Research Council**. I hope to see this kind of project take off in 2022 because partnerships with specific companies tend to break down when those companies confront the **FTC** and/or the relevant **Data Protection Office** in the EU. It is extremely difficult to be upset with a company that is attempting to comply with regulators. It would be nice if regulators envisaged data sharing not strictly as a risk that needs to be mitigated, but as a nuanced terrain where data sharing needs to be allowed along a clear, brightly lit path that prioritizes not just the rights of individual users but also the overall social benefit that accrues from carefully designed, publicly available research.

Our DSCN readers care about this topic quite a bit. The most clicked on **coverage** last year was about Facebook opening access to its **new platform** for research data sharing. While the new Facebook + academia program was uncomfortably hush-hush in terms of which researchers are part of the program and what kind of research they are doing, allowing a little breathing room may not be such a bad thing. Dousing criticism on this fledgling V2 program may shut it down completely, which would be a net negative. No matter how you feel about Facebook, many people use it to socialize. Many small businesses rely on it. Keeping some kind of door cracked to share data — as limited as it may be — is better than bricking over the door altogether.

REGULATORS PUT ANTI-MONOPOLY PROTECTIONS AT THE TOP OF THEIR 2022 AGENDA

The **FTC**, the **UK Competition and Markets Authority**, and the **EU Competition Authority** have put anti-monopoly rule making at the top of their agenda for 2022. The UK already blocked Facebook's proposed acquisition of **Giphy** in late 2021.

We are not yet sure how this may impact data science and AI, we know that regulators are concerned with the concentration of power they see across the tech sector. Keep in mind that regulations revolving around user privacy had the consequence of making it harder for academics to use data stored on big tech platforms. Even though no regulators are aiming at academia and non-profit research institutes, regulator's decisions could have implications for academic data science research.

NEW PROGRAMS, FOLLOW THE MONEY

Click through to access **a structured spreadsheet** of New Programs and money moving around in academic data science.

DATA VISUALIZATION OF THE YEAR

Flowing Data, Nathan Yau from December 9, 2021

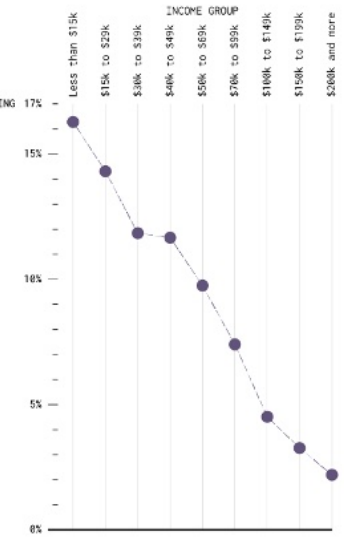
Nathan Yau of *Flowing Data* has consistently created quietly thought-provoking visualizations that push into the deeper context of issues in the news. **This one** looks at the proportion of income that goes to different types of spending, by income level. There is far more to the visualization than we were able to include. Go forth and immerse yourself. At least it isn't news about covid or climate change.

Lower Income Spending

These are the categories that lower income groups tend to spend more on, percentage-wise, than higher income groups. Note the downward trends.

HOUSING
Rented dwellings

Lower income groups spend a larger share of their income on rented dwellings, whereas higher income groups tend to spend more on owned dwellings.

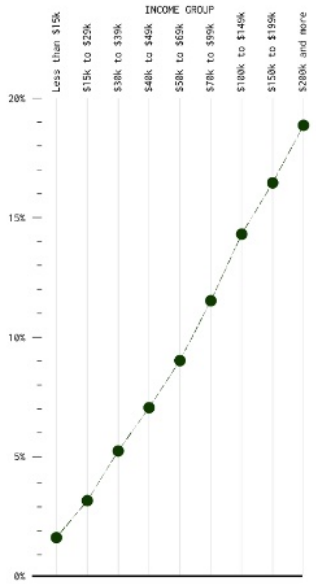


Higher Income Spending

These are the categories that higher income groups tend to spend more on, percentage-wise, than lower income groups. Note the upward trends.

PERSONAL INSURANCE AND PENSIONS

When you make more money, you have more to save, both by share of spending and total spending.



Deadlines

Contests/Award

[Google Open Source Expert Prize]

"Submit your public notebooks using **Google Open Source** frameworks and / or in-depth discussion posts to be considered for a $1000 monthly award!" First submission deadline is January 24, 2022

Tools & Resources

Databases in 2021: A Year in Review

Ottertune blog, Andy Pavlo from December 28, 2021

"It was a wild year for the database industry, with newcomers overtaking the old guard, vendors fighting over benchmark numbers, and eye-popping funding rounds. We also had to say goodbye to some of our database friends through acquisitions, bankruptcies, or retractions."

About Us: The Data Science Community Newsletter was founded in 2015 in the Moore-Sloan Data Science Environment at NYU's Center for Data Science. We continue to be supported by the Gordon and Betty Moore Foundation and the Alfred P. Sloan Foundation through the Academic Data Science Alliance. Our archive of newsletters is at https://academicdatascience.org/resources/newsletter. Our mailing address is 1037 NE 65th St #316; Seattle, WA 98115.