Data Science Community Newsletter

Data Science Community Newsletter features journalism, research papers and tools/software for September 6, 2020

Please let us (Micaela Parker, Steve Van Tuyl, Laura Noren, Brad Stenger) 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 .

Dear Reader

On behalf of the Alfred P. Sloan Foundation, ADSA is conducting a short survey of universities to explore access to and benefits of staff data scientists. Our goal is to help universities make informed decisions about investing in data science capacity. Our aggregated findings will be made publicly available. For the purposes of the survey, we are defining staff data scientists as non-faculty academic staff who use computational, statistical, and/or mathematical methods and tools to help extract knowledge from data. The survey is meant for researchers at any level and in any field who already work with staff data scientists or are interested in establishing a connection to them.

SURVEY LINK

The survey is completely voluntary and should take 5-10 minutes to complete. The data that respondents enter will be collected by Abt Associates, a policy research and evaluation firm. Abt will analyze the data and develop a public report. Please consider participating and/or distributing to your colleagues.
Thank you!

Academic Data Science News

The National Science Foundation took an important step toward implementing a national plan for artificial intelligence research, creating five national AI Institutes, as well as helping the U.S. Department of

Agriculture to establish two more food-related AI Institutes. All of the Institutes have $20 million in funding earmarked for their first five years. The list (with the academic homebase for the Institute):

- NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (University of Oklahoma) press release
- NSF AI Institute for Foundations of Machine Learning (University of Texas, Austin) press release
- NSF AI Institute for Student-AI Teaming (University of Colorado, Boulder) press release
- NSF AI Institute for Molecular Discovery, Synthetic Strategy, and Manufacturing (University of Illinois at Urbana-Champaign) press release
- NSF AI Institute for Artificial Intelligence and Fundamental Interactions (Massachusetts Institute of Technology) press release
- USDA-NIFA AI Institute for Next Generation Food Systems (University of California, Davis) press release
- USDA-NIFA AI Institute for Future Agricultural Resilience, Management, and Sustainability (University of Illinois at Urbana-Champaign) press release

NSF will have more national AI Institutes, and will continue to partner with federal government agencies, specifically mentioning Department of Transportation and Department of Homeland Security in the press release. According to the White House Office of Science and Technology Policy press release announcing the AI Institutes, NSF has $300 million more to spend and has a 12 month timetable to announce the remaining Institutes.

Historical context (and a handy USA map) for the NSF AI Institutes comes not from OSTP but from the Computing Community Consortium. CCC's 2019 Artificial Intelligence Roadmap report included a recommendation for National AI research centers as part of a thrust to establish robust U.S. research infrastructure. You can sort of see the two other CCC AI Roadmap thrusts, workforce development and long-term administrative planning, in these NSF AI Institutes. That's encouraging. Given the choice, it's better to not charge forward blindly into the future.

The NSF did not forget about Data Science while binging on AI grantmaking. The University of Washington is going to lead a multi-university research center, the Institute for Foundations of Data Science, backed by $12.5 million in NSF funding. Collaborators on this center are University of Wisconsin-Madison, University of California-Santa Cruz and University of Chicago. A second $12.5 million research center, the Foundations of Data Science Institute will bring together researchers from

the University of California-Berkeley and Massachusetts Institute of Technology, plus Boston University, Northeastern University, Harvard University, Howard University and Bryn Mawr College. The centers and grants are under NSF's "Transdisciplinary Research in Principles of Data Science" (TRIPODS). It's a lot going on. The NSF press release explains what's happening better than I can.

Open Platforms and Knowledge Networks are another central infrastructure recommendation that appears in the CCC AI Roadmap. The new NSF Institutes are "highly collaborative" but not explicitly "Open." Fortunately capital-O Open research is top of mind in at least a few places. Phil Bourne, Dean of the School of Data Science at University of Virginia, wants badly to see more Open Science publishing, in large part to enable and improve scientific content mining. Bard College professor Kristin Lane pre-printed an essay on the challenges of doing Open Science at small liberal arts colleges, with co-authors from Haverford, Washington & Lee University and Lewis & Clark College, also h/t Brian Nosek. Lonni Besancon, from Monash University in Australia, together with a global collection of life scientists, authored a pre-print highlighting the health benefits of making pandemic-related research Open.

There are a few more major infrastructure-related grants to report. University of California-Berkeley received $10 million from NSF and Simons Foundation to investigate theoretical underpinnings of deep learning. The University of Toronto took in C$9.5 million from Canadian government sources, earmarked for 33 technology research projects across the school's 3 campus locations. NSF will also be giving $10 million for a cloud computing testbed called Chameleon, a group led by University of Chicago researchers. Cloudbank, another NSF-funded public cloud computing project based at University of California-San Diego's San Diego Supercomputer Center, became operational, one year after its $5 million award.

Seeing money flow into so much useful infrastructure does a lot to set up a better post-pandemic research environment. Confidence is and should be high. If only the talk about Open Science had dollar signs attached. It would go one step further to putting the collective U.S. AI research enterprise on solid footing.

Northwestern University announced a new MBAi joint degree program between Kellogg School of Business and McCormick School of Engineering.

University of Texas-El Paso will be offering Doctoral degrees in Data Science.

The University of Manitoba will have an undergraduate major concentration in Data Science.

Purdue University will offer a Graduate Certificate in the school's new online Applied Data Analytics

program.

Trustees at The Ohio State University approved the school's new Translation Data Analytics Masters degree program last November, and it is already enrolling students. The program is designed for working professionals and emphasizes experiential learning. More at https://tdai.osu.edu/mtda/ [h/t Cathie Smith].

Editor's Picks

Sharon Sputz at Columbia University's Data Science Institute asked DSCN, "Anyone know how to get a GPT-3 API key" from OpenAI? Good question. The short answer is, "Buy one. There's a license you can purchase." And it's expensive. But the long answer has to do with the cost of AI research and what it means to monetize an API. Let me mention that this long answer was helped by two recent articles, one by Dave Gershgorn in the Medium online magazine OneZero, and another by Ben Dickson writing at The Next Web.

AI research is prohibitively expensive. It's a double hammer. Talent costs money and so does all of the computing that is required. One important aspect of an API is how it can bring talent to the organization. If the compute overhead for an API is low, you can offer your APIs for free, and justify any expense as community building that helps with hiring. That's why public APIs at places like The New York Times don't cost anything. GPT-3 isn't public, but an elite talent who is plugged into the OpenAI community should be able to get an API key. The path into the OpenAI community is gated however, meaning OpenAI will let you know when they want you in their community and not the other way around. The exception, of course, is if you're wanting and authorized to spend whatever it costs per month for the API access.

Ultimately there will be a reckoning of haves and have-nots when it comes to superior algorithms and APIs. Open Science and Open Platforms are good ways to close the gap, part way but not all the way. For organizations, like Columbia DSI, where spending decisions are deeply considered, the opportunity is to vote with your wallets, especially if it's a portfolio approach that mixes Open and private resources in an organization's stakeholding.

Amy Bruckman's essay in Communications of the ACM is a super-useful explainer on computing ethics, and more generally, on all new technology ethics. Her ideas transfer easily, to classrooms, and to individuals' motivations and decisions. Let me summarize Dr. Bruckman's argument into two fundamental

points, as I understood them. First, ethics are a conversation, and second, ethics in practice require follow through. The ethics conversation is often difficult, but frequently crucial. If the conversation is something that leads to a call to action, then the follow through is the action. Follow through comes in different forms. Examples are personal work with positive impact, or collective action like Google employees' opposition to Project Maven, or public policy in government. Basically, you don't get the follow through unless you have the conversation. So have the ethics conversation; it's a need-to, not a want-to.

Despite the challenges, ethics conversations are out there, often with a call to action. Just remember how important it is to follow through. Some examples:

- The term 'ethical AI' is finally starting to mean something (VentureBeat, Carly Kind)
- Algorithms promised efficiency. But they've worsened inequality (CNN Business, Zamira Rahim)
- AI has a high IQ but no emotional intelligence, and that comes with a cost (BBC Science Focus Magazine, Rana el Kaliouby)
- Laptop and Chromebook shortage hurts lower-income students (Vox, Recode, Sara Morrison)
- Online School Is Harder Thanks to Unequal Internet Access (The Atlantic, Olga Khazan)
- Call for Comment: How Can Hospitals Improve the Health of Their Communities? (Johns Hopkins Center for Health Equity)

The COVID virus continues to ravage the United States, university campuses included. Applied science, including applied data science, often needs quality leadership to succeed. No surprise then that failings in our pandemic — on campus, with testing, in government, in research — can be traced to situations where leadership has been lacking. Alternatively, there are examples like these American grad students where promising new leaders are emerging.

Research News

There's a roadmap for Computational Social Science, jointly written by many of the voices you would want for such a thing, but it's behind the Science journal paywall. So far the best summary that we have seen is this tweet by Alondra Nelson.

Robots!

- Dusty Robotics CEO Tessa Lau Discusses Robotics Start-Ups and Autonomous Robots for Construction (Robotics Business Review, Joanne Pransky)

- Michigan announces plans for autonomous vehicle corridor (Smart Cities World)
- The robotics revolution is here, and it's changing how we live (National Geographic, David Berreby)

Outer Space! (Stratosphere included)

- What Does the Future of Astronomy Hold? We'll Find Out Soon (Discover Magazine, Yvette Cendes)
- Fifty new planets confirmed in machine learning first (University of Warwick, News & Events)
- Technical Document: Impact of Satellite Constellations on Optical Astronomy and Recommendations Toward Mitigations (NSF, NOIR Lab)
- NOAA and World View partner on stratospheric composition research (TechCrunch, Darrell Etherington)

Inventions!

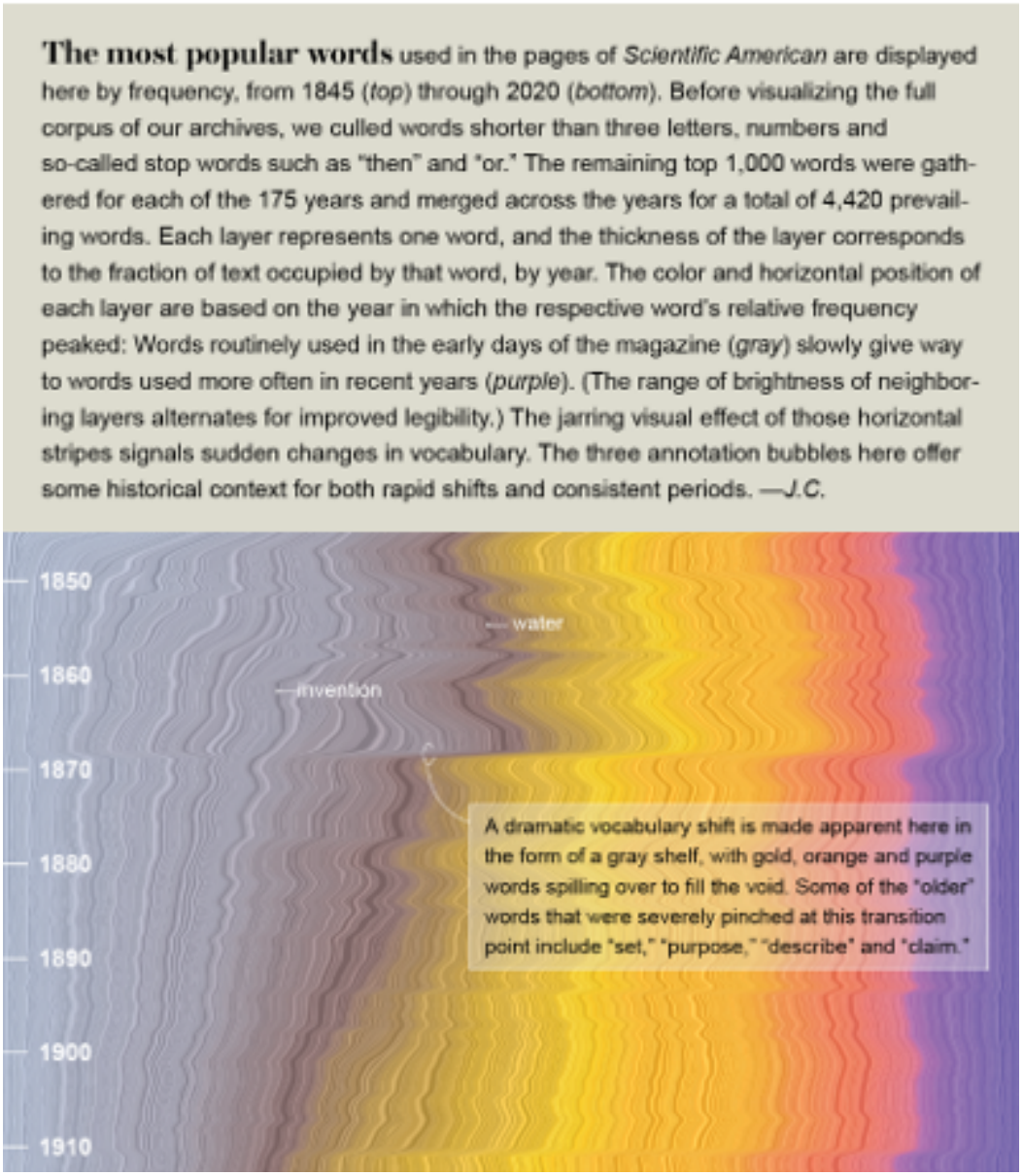
- UMass Amherst Scientists Invent New Sensing Eye Mask (University of Massachusetts Amherst, News & Media Relations)
- Researchers Can Duplicate Keys from the Sounds They Make in Locks (Jason Kottke)
- xTech Brain Operant Learning Technology – xTechBOLT – prize competition (U.S. Department of Defense)

Tweet of the Week

Twitter, Katy Huff from September 01, 2020

Data Visualization of the Week

Scientific American; Stefaner Moritz, Lorraine Daston, Jen Christiansen from August 18, 2020



See the all of the *The Language of Science* visualizations and analysis at Scientific American.

Events

New Jersey Institute of Technology 2020 Fall Seminar Series

Online September 9, starting at 4 p.m. Eastern time. Speaker: **Yifan Hu** from **Yahoo Research Labs**.

[registration required]

Computing Researchers Respond to COVID-19: National Health Symposium- Operationalizing AI in Health

Online September 14-15. "The symposium will explore artificial intelligence (AI) and the continuum between essential research and development through their translation from innovation into operational impact. Participants will learn about real-world AI applications for healthcare, harnessing AI technologies to accelerate advances while doing no harm, and ensuring the safety and security of healthcare while realizing AI's full potential." [registration required]

CSAIL co-launches virtual summit on AI in healthcare

Online October 1-2. "The aim of the summit is to boost effective collaboration among leading AI academics, healthcare experts and business leaders to support innovation in healthcare." [registration required]

The 2020 ADSA Annual Meeting is FREE and virtual!

Online October 14-16. "We will bring together data science methodologists and domain researchers from all disciplines and career stages to share breakthroughs and new approaches in data science research and education, with a strong emphasis on responsible data science. We are encouraging new, untested ideas to promote brainstorming for innovation and promote collaborative feedback and engaging discussions." [registration required]

Tools & Resources

Announcing the new Jupyter Book.

Jupyter Blog, Chris Holdgraf from August 12, 2020

"Jupyter Book is an open source project for building beautiful, publication-quality books, websites, and documents from source material that contains computational content. With this post, we're happy to announce that Jupyter Book has been re-written from the ground up, making it easier to install, faster to use, and able to create more complex publishing content in your books. It is now supported by the Executable Book Project, an open community that builds open source tools for interactive and executable documents in the Jupyter ecosystem and beyond."

13 major climate change reports released so far in 2020

Yale Climate Connections, Michael Svoboda from August 19, 2020

"In this edition of our bookshelf feature, **Yale Climate Connections** highlights a baker's dozen of these

reports, selected to reflect the broad range of concerns that intersect with climate change, including water, national security, media, health, food, finance, energy, and climate and environmental justice."

A community-maintained standard library of population genetic models

eLIfe, Jeffrey Adrion et al. from June 23, 2020

"The explosion in population genomic data demands ever more complex modes of analysis, and increasingly, these analyses depend on sophisticated simulations. Recent advances in population genetic simulation have made it possible to simulate large and complex models, but specifying such models for a particular simulation engine remains a difficult and error-prone task. Computational genetics researchers currently re-implement simulation models independently, leading to inconsistency and duplication of effort. This situation presents a major barrier to empirical researchers seeking to use simulations for power analyses of upcoming studies or sanity checks on existing genomic data. Population genetics, as a field, also lacks standard benchmarks by which new tools for inference might be measured. Here, we describe a new resource, stdpopsim, that attempts to rectify this situation. Stdpopsim is a community-driven open source project, which provides easy access to a growing catalog of published simulation models from a range of organisms and supports multiple simulation engine backends. This resource is available as a well-documented python library with a simple command-line interface."

Announcement: Access to the COVID-19 Data Analytics Platform Now Open

National Institutes of Health (NIH), National Center for Advancing Translational Sciences (NCATS) from September 02, 2020

"Researchers studying COVID-19 now are able to access an innovative new analytics platform that contains clinical data from the electronic health records of people who were tested for the novel coronavirus or who have had related symptoms. Part of the **NCATS National COVID Cohort Collaborative (N3C) Data Enclave**, the centralized and secure data platform features powerful analytics capabilities for online discovery, visualization and collaboration."