Data Science Community Newsletter features journalism, research papers and tools/software for October 7, 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.

ADSA 2021 ANNUAL MEETING CANCELED-ISH
Due to the ongoing pandemic and responses from our community to our travel survey, ADSA has sadly decided to cancel the in-person 2021 Annual Meeting originally scheduled to take place in Savannah, GA this November. Instead, we are excited to bring you virtual breakout sessions and recorded research talks this fall and winter. We are also planning a shorter in-person networking event for March 2022. Keep an eye out here for updates!

APPLYING ETHICAL APPROACHES TO DATA SCIENCE
The AI Now Institute developed "datasheets for datasets" about 5 years ago. Now Karen Boyd has tested the framework in the wild, finding that it does help engineers, data scientists, and programmers move towards more socially responsible project outcomes.

Still, it can be hard to understand how tech ethics is structured within large companies. Recently with Facebook, the public is seeing gaps between what company leaders have previously said in public and what researchers at Facebook have found (some of whom are DSCN readers — hello, readers!). A new partnership between the World Economic Forum, the Markkula Center for Applied Ethics at Santa Clara University, and companies including Microsoft and IBM has produced reports that are extremely useful in this context. They explain — in detail, with charts — how each company structures itself to foster socially beneficial approaches to their products and features. It has been hard to find this level of detail around which roles, committees, and processes are useful within large corporate structures.

DSCN READERS MOVING UP
The 66th person to sign up for the Data Science Community Newsletter back when we started in 2015, Professor Gina Neff, who since left the University of Washington to become Professor of Technology and Society at the University of Oxford has now been appointed to lead the Minderoo Centre for Technology and Democracy at the University of Cambridge. We like to celebrate when our longest-term readers who supported us when we were new and experimental are so well-appointed. :)

People frequently mentioned in the DSCN are also appearing in President Biden’s Council of Advisors on Science and Technology. Representing academic data science and data science research are Saul Perlmutter of University of California-Berkeley; Eric Horvitz, Chief Scientific
Officer at Microsoft; Phil Venables, Chief Information Security Officer of Google Cloud. Lisa Su, CEO of Advanced Micro Devices, and William Dally, chief scientist at NVIDIA, represent the chip industry. All together thirty science and tech notables constitute the most diverse PCAST in U.S. history.

Sadly, DSCN subscriber Yann LeCun reported (see Twitter) that his 2021 NeurIPS submission with co-authors Adrien Bardes and Jean Pounce, "VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning," was rejected.

SPONSORED CONTENT

The $50,000 prize pool Food for Thought competition challenges data scientists to use NLP and machine learning in helping the USDA link food and nutrition datasets. Applications close Sunday, 11/7. Visit our website now to sign up for the informational webinar on 10/12!

MANY AI ROADMAPS; ONLY ONE RETROSPECTIVE

The UK recently released a national roadmap for advancing AI technologies. The UK report positions the country as running in third place behind the U.S. and China and aims to increase educational spending, visa affordances for high skilled AI workers, and "developing a pro-innovation regulatory and governance framework" to accelerate broad adoption of AI across sectors. This signals a possible point of departure from EU-based data protection rule-making which is considered to be quite rigorous.

In Germany, which also sees itself falling behind the U.S. and China on AI adoption and where elections are slated for December, AI is on the campaign agenda. “In 2019, over 50% of AI job vacancies in Germany either could not be filled or were filled later than wished or with less desirable candidates,” and “only around 6% of companies” in the small and medium-sized enterprise category use AI according to a recent German Ministry for Economic Affairs and Energy survey.

Australia also has a new report on its national position with respect to digital technologies that points to the U.S., UK, France, and Canada as leaders and requests additional funding and policy support to increase investment and innovation in areas that include AI as well as 5G, blockchain,
VR/AR, IoT, and quantum computing.

At least with respect to AI, more and more of the world is preparing for launch. It's still not clear that there is an agreed-upon definition of what is and isn't AI. It probably doesn't matter much at this stage, though it will matter later when regulatory efforts catch up to innovation, as is the general pattern.

The U.S., for its part, has announced it will form a committee under the Department of Commerce to investigate the state of the U.S.'s competitive position, scientific ability, and missteps (e.g. criminal justice?) vis a vis AI. Not waiting for the U.S. government, the AI 100 effort that started in 2016 convened at Stanford University and including a range of computer, natural/physical, and social scientists in academia and industry, has released the second report. In terms of benchmarking where the state of the science of AI and its impact on the world stands, this is the best report available. It includes many DSCN readers on its standing committees, ad hoc committees, and author lists.

**LEGISLATING LAW ENFORCEMENT**

The attorney general of Colorado, Phil Weiser, recently completed the state's first civil rights investigation of law enforcement accountability law that was passed during the summer of 2020. The investigation, led by data scientists David Mordecai, Samantha Kappagoda, Michael Kwak, Mihir Gokhale, Noah Mathews, and Peter Horvath, computed chi-square tests for years of interaction data between citizens and police in Aurora, Colorado.

The analytics team found that Aurora police had clear patterns of racially biased policing against Black residents both in terms of likelihood of being detained and in terms of disproportionality of force. The use of unlawful excessive force has perhaps most notably been evidenced in the death of Elijah McClain, a young Black massage therapist and violinist who died when officers and paramedics placed him in a chokehold and injected him with 1.5x the dose of ketamine advised for his weight after receiving a call that a person wearing a ski mask looked "sketchy." There is no evidence that McClain resisted arrest or that he had done anything illegal or threatening.

The Aurora Fire Department's practice of injecting ketamine had been under investigation prior to this report, and an Aurora FD state waiver to administer ketamine outside of a medical setting that had been in place when McClain died has been left in a state of unrenewed moratorium.

Legislation in advance of legal action using data is also critical for determining appropriate police use of digital search warrants and police use of data purchased from third parties. NYU Law School professor Barry Friedman argues that "as a matter of constitutional law" policing agencies should collect no digital data whatsoever without a "sufficient regulatory scheme in place."

Suspicionless searches of consumer databases by law enforcement is a growing problem, according to Jennifer Lynch of the Electronic Frontier Foundation. State legislatures need to strike an equitable balance between public safety and personal privacy.

**IS THERE ENOUGH DATA TO ASSESS MODERNA, J&J BOOSTER SHOTS?**

We were able to look at Israel's all-Pfizer vaccinated population to develop a framework for understanding which sub-groups should receive a booster shot of the Pfizer vaccine. We don't have a similarly large group of well-tracked people either for Moderna or for Johnson & Johnson. We are especially lacking a group of people who received booster doses to prove that a booster of Moderna or J&J will boost the immunogenic response for all-aged recipients, just certain age groups, other subsets, etc. Israel had already been administering third doses, but they only used Pfizer so their data only speaks to Pfizer third doses. Since Pfizer's first and second doses were a bit weaker than Moderna's first and second doses and separated by a different number of weeks, there is
reason to wonder if the necessity and timing of a Moderna third dose and/or the recommended recipient groups should change. Moderna boosters are currently only allowed for those with compromised immune systems and discussions are exploring half-dose boosters. As of yesterday, J&J believes it has enough evidence to recommend a booster for all recipients 18 or older and seeks an audience with regulators at the U.S. Food and Drug Administration.

THE INDUSTRIAL BIO COMPLEX IN YOUR BATHROOM
Vijay Pande, venture capitalist at Andreesen Horowitz and an adjunct professor of structural biology at Stanford, writes (and tweets) that the mobilization of biotech resources for Covid is about to usher an unprecedented era of Industrialized Bio where engineering and AI dramatically increase automation and improve healthcare quality. The trigger: consumer demand. One testbed will be whether we adopt the sensor technologies that are poised to invade our bathrooms. The smart toilet is an obvious starting point. At the other end of the pipe, wastewater testing has proven to be an effective albeit imprecise big-picture Covid data point for communities. And health officials expect more uses for wastewater analysis post-pandemic.

Sweat, saliva and breath sensing are also seeing rapid advances. A new saliva sensor developed by Korean researchers incorporates user-friendly, chewy, chocolatey Tootsie Rolls. Maria Cuartero and Gaston Crespo from KTH Royal Institute of Technology in Stockholm recently published a review of sweat capture and biomarker sensing technologies geared towards athletes in ACS Sensors journal. Jun Ye and Mark Nesbitt at JILA in Colorado have dramatically increased the precision of breath analysis using mid-infrared (mid-IR) cavity-enhanced direct-frequency comb spectroscopy (CE-DFCS). The frequency comb is a laser that produces 10,000 different colors of infrared light. The mirrored cavity amplifies weaker chemical signals and can measure 15,000 chemical absorption optical frequencies in near-real time. The Ye and Nesbitt labs are at work on a compact, possibly bathroom-size version.

FUNDING FOR PARTICLE PHYSICS IS (IRONICALLY) HUGE
LBFN/DUNE, a particle physics project based at the Fermilab in Illinois, is going to cost $3 billion. That's 60% more and 4 years longer than originally projected. The U.S. Department of Energy is committed to press on looking for neutrinos. Just think of how much non-neutrino science could be done with the additional $1.3 billion the project needs just to keep going. Scroll to any of our Follow the Money sections and look at how many projects are winning less than ~$2 million after relentlessly submitting grant applications.

SPONSORED CONTENT
FOLLOW THE MONEY

$480,000,000 Patrick and Shirley Ryan -> Northwestern University to support education and research.

$100,000,000 Penny Pritzker -> Harvard University, Department of Economics. The former U.S. Secretary of Commerce made a massive gift to her alma mater.

$75,000,000 National Science Foundation -> Five new "Harnessing the Data Revolution Institutes" at The Ohio State University, University of Washington, University of Maryland, Baltimore County, Colorado School of Mines, and University of Illinois at Urbana-Champaign.

$73,000,000 US Department of Health and Human Services via the American Rescue Plan - > 10 "historically Black colleges and universities, Hispanic serving institutions, Asian American and Native American Pacific Islander-serving institutions and other minority-serving institutions" to train 4,000 students in health-related IT, including health informatics.

$50,000,000 University of North Texas Health Science Center at Fort Worth -> National Institutes of Health to be the Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity coordinating center.

$36,000,000 Lilly Endowment Inc. -> Central Indiana Corporate Partnership for an initiative called AnalytiXIN with half of the funding to "hire university-level data-science researchers...and the other half...towards the creation of ‘data lakes’" in manufacturing and health-sciences.

$20,000,000 "blockchain pioneer Charles Hoskinson" -> Carnegie Mellon University for the Hoskinson Center for Formal Mathematics.

$17,000,000 National Institute of Child Health and Human Development -> The Ohio State University to establish "the Maternal and Pediatric Precision in Therapeutics Data, Model, Knowledge and Research Coordination Center." Dear academics, names do not need to be this long nor is a long name more descriptive than, for instance, Mother and Child Data Science Center.
$10,800,000 National Institutes of Health -> Brown University's Center for Computational Biology of Human Disease.

$10,000,000 Steve and Alexi Conine ($5m) and Niraj and Jill Shah -> Cornell University, Ann S. Bowers College of Computing and Information Science for a new building. Steve and Niraj founded Wayfair.

$3,250,000 Department of Energy -> University of Chicago and Argonne National Lab to explore using "surrogate models" — simplified models that utilize AI to simulate extremely complex physical systems, including climate and energy systems.

$3,000,000 Google.org -> Blue Conduit and partners to build Open Source tools based on the organization's success at locating underground lead piping in Flint MI.

$2,500,000 National Science Foundation -> University of Michigan — MIDAS, UW, and NYU to establish "a Framework for Integrative Data Equity Systems (FIDES)"

$1,400,000 Zynga's Social Impact Fund -> North Carolina A&T "to elevate Black representation in the gaming industry"

$99,760 Arkansas National Science Foundation Established Program to Stimulate Competitive Research (NSF EPSCoR) -> Arkansas Tech University. Moneys will be used to collect data sets from research faculty for student projects.

$17,700 University of Oregon, Department of Computer Science -> Graduate Employees. The graduate employees petitioned for a larger stipend (between ~$21-24,000 depending on level), but their request was rejected. One of the pre-pandemic activities likely to return: graduate student labor organizing.

**NEW PROGRAMS**

Metropolitan Chicago Data Corps will bring together data scientists from five local universities to offer data science services to agencies and community groups in the Chicago metro area.

Data Science Major Northwestern University, Department of Statistics in response to student demand for the new Data Science minor which went from "76 students in fall 2019 to 349 in fall 2021."

New majors in data science and finance are available to undergraduates at St. Lawrence University.

LawTech Center at University of Virginia to be led by Danielle Citron.

Computational Social Science Lab at UPenn to be led by Duncan Watts.

New York City Pandemic Response Unit will be launched and operated at Columbia University. Wafaa El-Sadr, epidemiologist at the Mailman School of Public Health, will lead PRU. CUNY Graduate School of Public Health and Health Policy was named a key partner in the initiative.
Virginia Tech National Security Institute at Virginia Tech to become "the nation's preeminent academic organization at the nexus of interdisciplinary research, technology, policy, and talent development to advance national security." Notably, "the Department of Defense is Virginia Tech's largest source of federal funding with approximately $50 million in fiscal year 2020."

DATA VISUALIZATION OF THE WEEK
Twitter, Eric Topal, from October 3, 2021

An extraordinary 90 second video of the hotspots and spread throughout the American pandemic, by @BioTurboNick

Click through to view the entire animation (updated weekly) by Nicholas Bauer.
Deadlines

Education Opportunities
We are looking forward to submissions to our special issue on Machine Learning Approaches in Big Data Visualization. This is for IEEE CG&A Magazine. "Submission Deadline: October 29, 2021"

Contests/Award
NFL Big Data Bowl 2022
"Your challenge is to generate actionable, practical, and novel insights from player tracking data that corresponds to special teams play." Deadline for submissions is January 7, 2022.

Tools & Resources
Introducing LingHacks Resources! Wondering where to start learning NLP or do NLP research/activities?
Twitter, LingHacks from September 27, 2021
"Check out our growing collection of resources at http://linghacks.tech/resources. Want to submit a resource for the community? Fill out the form"

GHub: The new place to be for ice sheet scientists
University at Buffalo, News Center from October 5, 2021
"Glaciologists have teamed up with computer scientists to build an online portal for ice sheet science. The new science gateway is called GHub, short for 'Glaciology Hub.'"

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

Milwaukee Data Day
Online October 20, starting at 10 a.m. "Data Day, hosted by Data You Can Use, provides an opportunity for leaders, from neighborhoods to universities, to better understand what data are available, the implications of data on community revitalization, and how to increase connections between research and practice." [$$]

Open to everyone! The first-ever 2021 Microsoft Research Summit
Online "Oct 19 - 21, with over 150 sessions across 16 tracks, provides the global research community with an opportunity learn from experts pushing the frontiers of technology."

Featured Jobs
See the ADSA Jobs Page for more opportunities.

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