ADSA Data Science Community Newsletter

Data Science Community Newsletter features journalism, research papers and tools/software for July 29, 2021.

Please let us (Micaela Parker, Steve Van Tuyl, Catherine Cramer, 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.

CORRECTION

Last week, we incorrectly stated that the University of Wyoming's faculty Senate submitted plans for a major restructuring of the university. The plan was submitted by the school's administration. We also mistakenly had "Final Draft" in the email subject line.

PREDICTING PROTEIN STRUCTURE

European Molecular Biology Laboratory (EMBL) and **DeepMind** **launched** the **AlphaFold Protein Structure Database**, a public dataset with 350,000 protein structure predictions. The AlphaFold algorithms for those predictions **was made public** a few days earlier in a *Nature* paper that included its open source code. The AlphaFold algorithms' announcement was **a few days after** the **University of Washington** published RoseTTAFold, a near-equivalent structure prediction algorithm, in *Science* with freely available code.

This is just the start. EMBL and DeepMind expect the database to predict structures for almost every sequenced protein known to science, more than 100 million structures. **David Baker**, leader of the UW project, **told** *Science* journalist **Robert Service** that he sees the field moving forward rapidly, away from static structures and toward dynamic protein-protein interactions.

An **editorial** in *Nature* compared the breakthrough to the Human Genome Project, where competition and open methodology led to a litany of valuable tools and insights. A better comparison might be the Periodic Table. Like **Dmitri Mendeleev's** **original "periodic" scribble from 1869** the AlphaFold and RoseTTAFold predictions **make room for discoveries** not yet made, like the "dark proteome" of protein-generating sequences that should, but don't yet appear in **Protein Data Bank** training data.

The current periodic table has rows that transpose Mendeleev's original columns, and it needed the "left-step" insight based on quantum-mechanical electron configurations provided by physicist **Charles Janet** in 1928. Designers would eventually drop the f-block elements of alkali

metals and alkali earth metals below the rest of the table, to fit better on a single page. An elegant expression of natural law, like the periodic table, **takes time and enormous effort** to do the work and to achieve consensus. We have completed one step. In time we'll understand life science like we understand materials science. The investigation won't end but the clarity in how, when and why things happen will become far more refined.

RIGHT TO AVOID RECOGNITION *OR* FAWKES, FACIAL RECOGNITION STINKS

Emily Wenger and **Shawn Shan** developed and open-sourced an algorithm called **Fawkes** that proves it is possible to build a "poison model" to fool facial recognition tech by applying a "facial mask" to photos of faces, undermining the ability of facial recognition models to correctly match photos to names. Wearing masks is another way to thwart facial recognition in public spaces — a highly open source solution.

The right to avoid facial recognition technologies used by civic agencies is limited to **thirteen cities** mostly on the east and west coasts and **Minneapolis**. Commercially, it's a mixed bag. **Target**, headquartered in Minneapolis, doesn't use it or have plans to use it, but **Macy's does**.

At the U.S. federal level, **20 agencies** are using facial recognition, but many major agencies like the **Department of Defense**, the **General Accounting Office (GAO)** (see **new video on AI accountability guidelines**) and **NIST** are considering limits or ethical guidelines for AI that could impact the use of facial recognition.

The objections to facial recognition are of two basic types. The first objection is that facial recognition is inaccurate, particularly for darker skinned people, leaving innocent brown-skinned people more likely to be pegged as troublemakers. Earlier this week, a Black teenager was denied access to a skating rink after being **misidentified** as someone who had caused trouble at the rink earlier. The facial recognition

system over-fitted to her eyeglasses, which were substantially similar to the glasses worn by the troublemaker. For this type of objection, the obvious solution is to improve accuracy. Better accuracy does NOT address the second class of objections to facial recognition, which is that people should know and be able to object to the increasingly personal, pervasive, and persistent surveillance, especially when the technology is used on behalf of those with power (governments, large corporations) while negative consequences are usually born by those without much power (individuals, small businesses, activist groups). For type 2 objections, any decent solution would have to move toward leveling fundamental power imbalances between the surveilled and the surveillers.

These types of objections to algorithmic surveillance also surface with other types of personal, pervasive, persistent technology like ad networks that sit on mobile phone apps tracking people's behavior and geolocation. A Catholic Substack publication **outed a priest as gay** after the Substack app and an ad network they use tied him to **Grindr** and yielded enough location data to track him to gay bars. He was **forced to resign**. The growing outrage about ad networks' disregard for data guardianship principles has amplified.

Where I have heard less objection — possibly due to lack of awareness — is to voice recognition, even though it is another biometric identifier frequently used in AI applications. Human voices produce unique audio wave forms, making them good identifiers. As a tech ethicist, I recommend curtailing the use of identifiers people cannot change like their faces, voices, fingerprints, DNA, the way they walk, their birthdate, and their past addresses. Only **New York** and **Illinois** have specific legislation around the use of biometric data, but **New York's law** only applies to commercial entities and allows biometric data collection if there's a notice posted in plain sight. Not a particularly good way of rebalancing power dynamics.

And let's not ignore surveillance of other species! Two new preprints show that animal DNA can be **sucked out of the air** and used to tell which animals have been nearby. This is both not the same as facial recognition (the techniques are about identifying animal species, not individual animals) but would trigger strident objections were it more sensitive and applied to humans. If facial recognition is banned but it becomes possible to suck DNA out of the air, we'll need broader regulations, not tech-specific bans and moratoria.

THE POWER GRID

The **National Energy Market** in **Australia**, like most grids, requires forecasts from electricity

generators every 5 minutes to ensure current production will meet demand five minutes from now. Using AI, researchers at **Monash University** **improved these predictions**, which are a challenge for solar farms with simpler weather models where a solitary cloud floating across the sun can cause a sudden, temporary dip in production. Monash researchers developed a machine learning approach with real-time sensor data from the farms. “Wind turbines already measure wind speed, wind direction, and power generation,” said Christoph Bergmeir, lead researcher on the **ARENA project**. Improving predictions by using available sensor data may not directly lead to peak shaving, but it does make it easier to integrate renewables into grids effectively, which distrubutes energy generation and decreases carbon emissions.

In the U.S. the **National Renewable Energy Laboratory** (NREL) is **working with** the **California Energy Commission** to pinpoint the number of electric vehicle chargers needed to meet the state’s goals of putting at least 5 million light-duty zero-emission vehicles on California roads and reducing greenhouse gas emissions to 40% below 1990 levels by 2030. NREL’s analysis found that “up to 1.2 million chargers — beyond those located at single-family homes — may be necessary,” (California has 70,000 chargers now.) Eventually the demand for out-of-home EV charging will be far less than today's demand for gas stations (and far more integrated into parking lots), which could reduce the number of EV chargers required.

Buildings, which consume 75% of electricity in the U.S., are the subject of another NREL **collaboration** with the **Department of Energy's Lawrence Berkeley National Laboratory**. "We don’t hear more about the role of our buildings as a significant resource for the clean energy transition is because it’s been challenging to quantify that resource at a large scale," said **Jared Langevin**, lead author of the **study**. By operating higher-performing equipment and shifting the time when its usage takes place, the authors found demand management technologies could avoid the need for up to one-third of coal- or gas-fired power generation drawn by buildings. This would mean at least half of all such power plants slated to be brought online between now and 2050 would be unnecessary. According to NREL's **Achilles Karagiozis**, director of **NREL’s Building Technologies and Science Center**, buildings are "a significant source of flexibility for grid operators, primarily in dialing down electricity demand during times when it would normally be at its peak," like on hot summer days. DOE recently released a **National Roadmap for Grid-Interactive Efficient Buildings**, which provides recommendations for how to triple the efficiency and flexibility of the buildings sector by 2030.

The **University of Minnesota** is offering seed grants to municipalities, community buildings, and recommendations to large corporations to support adding lithium ion battery storage capacity alongside renewable energy generators instead of sending excess directly to the grid at the time of generation. (See **slide presentations**). The batteries facilitate daily peak shaving (energy consumption is usually highest in the afternoon at which time energy companies often raise rates) and give facilities the ability to store energy for overnight use when their renewable generation is limited. Adding batteries to renewable generators empowers end users — from corporations to small cities – and incentivize greater renewable use. Negotiating with power companies remains a challenge. Power companies like battery storage when they recoup the savings, but can be ornery when end users add batteries to reduce their own bills.

A bit of government intervention with regulated utilities can help these research projects turn into

scalable business models. Expect **California** to lead the way in the U.S.

GOVERNMENT DOES THINGS, TOO

California's legislature unanimously approved "a plan to build a statewide, open-access fiber network". The middle-mile project doesn't extend all the way to residences, but it does bring fiber-speed internet closer to rural and other underserved areas. At those points, any ISP will be granted "non-discriminatory access" to build last mile connections to end users. This should actually advance progress towards 'broadband for all' in the U.S.

FOLLOW THE MONEY

$220,000,000 **Joe and Clara Tsai Foundation** -> Wu Tsai Human Performance Alliance (currently a consortium of **Stanford University**, **Boston Children's Hospital** (a **Harvard Medical School** affiliate), **University of California San Diego**, **University of Kansas**, **University of Oregon**, and the **Salk Institute for Biological Studies** for the "study of peak performance in athletes"

C$180,000,000 **Canada Pension Plan Investment Board**, **Softbank** and other investors -> **Deep Genomics**, an AI-based drug development startup founded by **Brendan Frey** who studied under **Geoff Hinton** and also founded the **Vector Institute for Artificial Intelligence**.

$35,000,000 **State of Maine** -> **UMaine system** to "develop a workforce for Maine"

$24,000,000 **Dorilton Ventures** and other VCs -> **Julia Computing** in a funding round

$7,700,000 **U.S. Department of Energy** -> **University of Washington**, **University of Michigan**, **Texas A&M University**, **University of Colorado**, **University of California-Irvine**, **University of Illinois**, **Northern Arizona University**, **Pennsylvania State University**, **Columbia University**, **University of Wisconsin**, **University of California-San Diego** to fund eleven studies that will increase understanding of Earth system predictability and improve DoE's climate model

$5,000,000 **National Science Foundation** -> **San Diego Supercomputing Center** for a National Research Platform (NRP)

$3,500,000 **National Institutes of Health's Eunice Kennedy Shriver National Institute of Child Health and Human Development** -> **Brown University, Population Studies and Training Center** focuses on: migration and urbanization, development and the environment, family health, reproduction health, HIV and AIDS, and health disparities

$2,300,000 **National Institutes of Diabetes and Digestive and Kidney Diseases**, part of the **National Institutes of Health** -> **Wayne State University School of Medicine** biochemistry professor **Kezhong Zhang** will study disruptions in human circadian rhythm and their affect on metabolic and cardiovascular disease

$2,250,000 **Toyota** -> **University of Kentucky College of Engineering**, **Bluegrass Community and Technical College** where students can now earn associates and bachelors degrees in Engineering Technology.

$1,500,000 **The Gordon and Betty Moore Foundation** -> **UC Berkeley** "to outfit spotter planes with improved infrared detectors to learn more about how fires spread"

$1,500,000 **U.S. Department of Defense** -> **Washington State University** for Northwest Virtual

Institute for Cybersecurity Education and Research

$300,000 **National Science Foundation** -> **University of South Florida**, **The AI Education Project** to study the public's understanding of artificial intelligence technologies

$300,000 **The Andrew W. Mellon Foundation** -> **Johns Hopkins University** for "Black Beyond Data: Computational Humanities and Social Sciences Laboratory for Black Digital Humanities"

+$534 **California Board of Regents** -> undergraduate students for a tuition and fee hike hitting Fall 2022

NEW PROGRAMS

Institute for Rebooting Social Media **Harvard University's Berkman Klein Center for Internet & Society** ... The Institute will be led by **Jonathan Zittrain** and **James Mickens**.

The Alliance, UK (official name TBD) **BCS**, **Royal Statistical Society**, **Alan Turing Institute** have formed an alliance to develop the "standards of professional competence and behaviour expected of people who work with data which impacts lives and livelihoods"

Diagnostic Accelerator **Harvard University + Boston's Brigham and Women Hospital** to "compress the time frame for introducing diagnostic technologies" with new diagnostic assays

Max Welling and his new molecular simulation lab **Microsoft Research Amsterdam** See: podcast transcript

AI Imaging Certificate **Radiological Society of North America** "to deliver a pathway for radiologists, including those who don't consider themselves technologically savvy, to understand and learn how to apply AI to their practices."

USF Institute for Microbiomes **University of South Florida Health** "to advance the research and development of innovative treatments and other microbiome-based solutions"

Data Labs: Roadmap to Recovery **National Governors Association** and the Beeck Center for Social Impact and Innovation at **Georgetown University** 9-month program "for state leaders looking to launch a data-informed project"

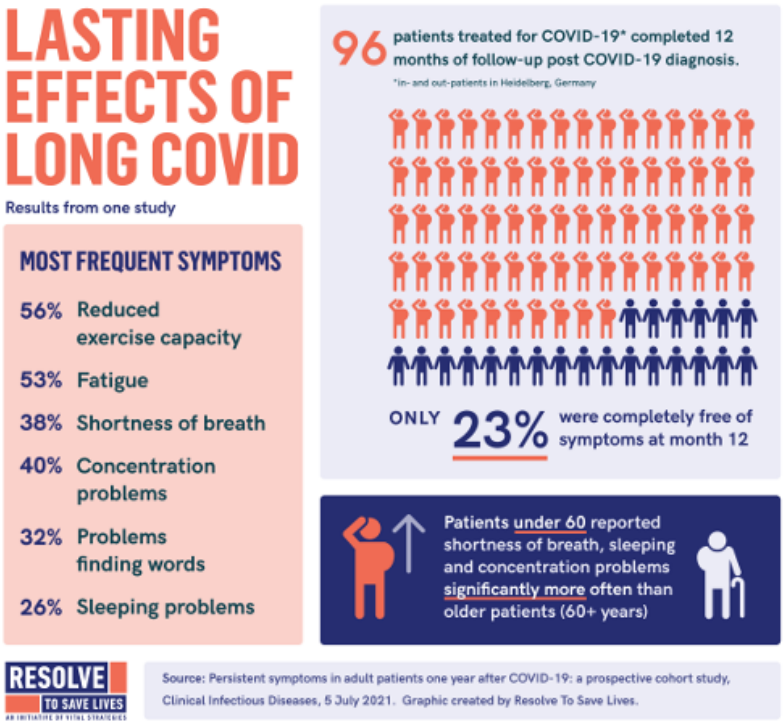
Computer Science Masters Degree **Angelo State University**

Data Science Major **Washington University in St Louis** offered as a bachelor of science or a bachelor of arts

Data Science Minor **Pepperdine University**

DATA VISUALIZATION OF THE WEEK

Medium, Dr. Tom Frieden and Resolve To Save Lives from July 9, 2021



Events

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

Australia's National Science Quiz

Online August 19, starting at 6 p.m. AEST. "Two teams will answer a series of thought-provoking questions that delve into the many branches of science using humour and some good scientific and mathematical reasoning."

United Nations World Data Forum

Bern, Switzerland, and Online October 3-6. "Data experts and users gather to spur data innovation, mobilize high-level political and financial support for data, and build a pathway to better data for sustainable development." [registration required]

Deadlines

Contests/Award

Strengthening Democracy Challenge

"The challenge is a joint project between academics and practitioners to crowdsource and identify short, scalable interventions to reduce anti-democratic attitudes, support for partisan violence, and/or partisan animosity among Americans." Deadline for submissions is October 1.

RFPs

Knight announces open call to fund new research into combatting disinformation in communities of color

"Seeking proposals for research that can lead to effective interventions that help fight disinformation

campaigns targeting communities of color." Support up to $175,000. Deadline for submissions is September 15.

Conferences

The big one is back: PyData Global 2021 is coming Oct 28-30. We can't wait to see and chat with world's most interesting innovators once again.

Online "Do you have ideas to share? Submit a proposal for your presentation." Deadline for submissions is August 15.

Tools & Resources

IBM, MIT and Harvard release DARPA "Common Sense AI" dataset at ICML 2021

IBM Research from July 19, 2021

"The research has been done with our colleagues at **MIT** and **Harvard University** to accelerate the development of AI that exhibits common sense. These tools rely on testing techniques that psychologists use to study the behavior of infants."

NASA Expands Access to Planet Data to All US Federal Civilian Agencies

Planet.com, Tanya Harrison from July 22, 2021

"**NASA** has expanded our contract with their Commercial SmallSat Data Acquisition (CSDA) Program to provide access to **PlanetScope** imagery for scientific research use for all U.S. Federal Civilian researchers and **National Science Foundation** funded researchers, including their contractors and grantees."

Careers

See the ADSA Jobs Page for more opportunities.

Executive Director of the NextGen Data Science and Analytics Innovation Center, University of Missouri System; Kansas City, MO.

Assistant Professor of Sociology, University of Arizona; Tucson, AZ.

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. The newsletter is written and the content is compiled by the Academic Data Science Alliance. Our archive of newsletters is at cds.nyu.edu/newsletter and is the process of transitioning to another, permanent location. Our mailing address is 1037 NE 65th St #316; Seattle, WA 98115.