ADSA
Data Science
Community
Newsletter

Data Science Community Newsletter features journalism, research papers and tools/software for September 8, 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.

COVID & THE ACADEMY

Children represented **26.8% of weekly reported COVID cases** in the week from 8/26-9/2 in the United States, according to the **American Academy of Pediatrics**. Many schools hadn't started back yet during that window; the case count among children is expected to increase dramatically in the next weeks and months. Alabama pediatrician **David Kimberlin** noted that, "having schools wide open with no masking, especially with our low vaccination rates, is going to be a disaster" on the ***In the Bubble* podcast**.

In **Georgia**, classes have resumed in the state university system without mask or vaccination mandates. Faculty like **Matthew Boedy**, who described his teaching duties as **"an emotional hellscape"** are **planning a statewide protest** on September 13th. Nearly 300 students and alumni of **University of North Carolina-Chapel Hill** have signed **an open letter** "Enough is Enough" calling for a vaccine mandate, frequent COVID-19 testing, and more testing centers on campus. Hundreds of residence hall advisors at **Stanford University**, which is still a week away from starting classes, **have gone on an indefinite strike**. The catalyst for the strike, according to **The Stanford Daily** student newspaper, was a student staff member who tested positive for COVID-19 after attending an indoor training session with other RAs, prompting requests for virtual training.

At least four faculty members at universities around the country **have resigned** upon being denied requests to teach remotely.

Some of last year's high school seniors are worried that the cheating they did to get by during the pandemic will make it harder for them to succeed as incoming freshmen: "cheating has made me unconfident" and being expected to do the work is **"a frightening thought."**

Meanwhile, researchers at **The Ohio State University** are **dusting dorms and other campus buildings for COVID-19 RNA**, which can survive in detectable amounts for up to a month. The dust is not infectious, but it can demonstrate where COVID-19+ people have been.

It is important to have humor at times such as these. Enter: **McSweeney's on COVID in the classroom**.

RESEARCH UPDATES

A research team led by **Isao Shitanda** at **Tokyo University of Science** has **installed sensors powered by urine** in adult diapers to monitor blood glucose levels, alleviating the need for blood draws.

University of Illinois Urbana-Champaign graduate students **Jiayang (Kevin) Xie** and **Parthiban Prakash** with postdocs **John Ferguson**, **Samuel Fernandes** and **Charles Pignon** and Professor **Andrew Leakey** used an approach that combined plant genetic data with imaging data of plant stomata to **predict which which plant specimens were likely to be more drought tolerant**.

A new paper from the **Molecular Information Systems Lab** at the **University of Washington** describes a **new class of reporter proteins** that can be read by commercially available nanopore sensors. “We’re essentially making it possible for these cells to ‘talk’ to computers about what’s happening in their surroundings at a new level of detail, scale and efficiency that will enable deeper analysis than what we could do before," said grad student researcher **Nicolas Cardozo**. The new proteins have magnetically charged tails that get them sucked into the nanopore sensors where machine learning helps classify them.

A new machine-learning based approach for predicting the edge of the Arctic ice sheet (or lack thereof – welcome to an ice free Arctic future) **outperforms the existing model** with substantially lower compute requirements. The IceNet model was developed by a team of 15 British scientists, 1 German, and 1 American. Predicting where ice will be in a couple months can help prevent crashes between ships and whales, disturbances of walrus populations that have to haul themselves onto land when there's no ice but are easily, often fatally, spooked by humans, and can help predict extreme cold weather events during North American winters.

Monya Baker has tools for reproducible experimental protocols (mostly life science) in her *Science* **feature article**. And there's a new set of Gold, Silver, Bronze reproducibility standards for machine learning papers **proposed** in *Nature Methods*. The standards would require data, code, models, programming environments, and one-click re-run ability for Gold. Defensibly ambitious.

Featured Jobs

See the ADSA Jobs Page for more opportunities.

Data Scientist at Acorn AI

Location: New York, Boston

INTERNATIONAL STUDENTS IN COMPUTER SCIENCE

The **National Foundation for American Policy** has a new study about international students in CS grad programs. Though it's not surprising there are more international students coming to CS programs over the past two decades – that time period overlaps with large overall growth in the field – it was surprising to see how dominant international representation is at the graduate student level. The trend in the past couple decades (1998-2019) had graduate enrollments in computer and information sciences up "with students from the U.S. increasing their numbers by 91% and international students increasing their enrollments by 310%."

The numbers:

International Student Enrollment in Grad programs:
+ 72% of computer and information science
+ 74% of electrical engineering
+ 71% of industrial and manufacturing engineering

I'm guessing some of this has to do with U.S. born CS people going directly into high-paying industry careers straight out of undergrad, while international students may need to "prove" themselves with credentialing first by getting undergrad degrees in their home countries to prove themselves to U.S. institutions of higher edu, then getting credentials in the U.S. to prove themselves to U.S. employers while in a visa cycle (F-1 to OPT) that eases their way into U.S. jobs.

However, that long arc increase masks a recent decline in international enrollment starting in about 2017 that was likely accelerated by Trump administration policies which made it difficult for international students to matriculate.

Any decrease in CS enrollments will have a **negative impact on the U.S. tech sector**.

For more, see "**US Power in International Higher Education**" edited by **Jenny J. Lee**

PHILIP GLASS ON AI MUSIC

Composer **Philip Glass** listened to AI generated music that had been trained on his corpus of work. **He didn't love it** – not a big surprise – explaining that "what's wrong with it is no one's listening to it....This is the difference between art and, let's say, a bunch of ideas that don't have an emotional direction to them." The AI produced the musical equivalent of the "fancy babble" that natural language generators have a tendency to create. The (musical) phrases may be locally intact, even interesting, but there's no narrative structure to the whole.

Ludwig van Beethoven, dead since 1827, will have no comment when **Beethoven Orchestra Bonn** **premieres** his unfinished 10th Symphony on October 9. The work to complete the symphony was led by data scientist **Matthias Roder** and musicologist **Christine Siegert** and funded by **Deutsche Telekom**.

Picking the narrative theme back up, **Mark Reidl**, a leading natural language processing researcher who has worked with AI in attempts to generate fictional narratives, **explains**: "there is no guarantee that the neural network will generate a text that is coherent or drives to a particular point or goal. Furthermore, as the story gets longer, the more of the earlier context is forgotten (either because it falls outside of a window of allowable history or because neural attention mechanisms prefer recency). This makes neural language model based story generation systems 'fancy babblers' – the stories tend to have a stream-of-consciousness feel to them."

Still, creative work can meaningfully incorporate AI as a tool. Artist **Bas Uterwijk** works closely with a facial imaging AI to create photo-realistic depictions of ancient people (e.g. **Nefertiti**) and fictional people (e.g. **Shakespeare's** Juliet) based on depictions of them in sculpture and portraiture. Definitely worth a scroll through his **photo book of historic notables** which includes **Jesus**, **Tutankhamun**, and **Aphrodite**.

UNIVERSITY OF CALIFORNIA-BERKELEY ORDERED TO HALT ENROLLMENT INCREASE

UC Berkeley has been **ordered to stop plans** that include increasing enrollment by 33.7% (11,285 students) between 2005 and 2020 by an **Alameda County** judge. The university must freeze its enrollment at 2020-21 levels until it has re-worked two environmental reviews pertinent to new construction on campus. A spokesperson for Berkeley is "optimistic" that this ruling will have no impact on future enrollment. The suit was brought by **Save Berkeley's Neighborhoods** who were initially joined by the **City of Berkeley**, which later settled with UC Berkeley and discontinued their participation. This is an important case for town/gown power struggles everywhere, but especially in California, where the legal precedent set in this case will be most meaningful.

REGULATION – UPDPA, BIASED BANKING, HEALTH APPS

City and state governments in the U.S. pursue data breach, data privacy, and AI regulations using state- and city-specific regulations. This presents a **complex regulatory environment** and has given rise to a new **Uniform Personal Data Protection Act (UPDPA)** – a piece of uniform law written by the **Uniform Law Commission**. States are encouraged to introduce the UDPDA in their next legislative sessions to promote a uniform legal environment. The proposed law is notably weaker than the existing California Consumer Protection Act / California Privacy Rights and Enforcement Act as well as the Virginia Consumer Data Protection Act and the proposed Washington Privacy Act that has also been introduced in Minnesota. In the UPDPA, there is **no right for consumers to remove their data** and broad protections for pseudonymized data, which concerns data protection experts who are concerned about the ease of re-identifying supposedly pseudonymized data. Local governments have also introduced moratoria on the use of facial recognition and sought to limit the use of other biometric data.

Data privacy protections would do little to advance social justice in consumer banking and, in fact, have been used to deny researchers access to data that could help discover *how* racism has been **embedded in mortgage underwriting**. A recent investigative journalism project by **Emmanuel Martinez** and **Lauren Kirchner** of **The Markup** had access to 17 factors used in over 2 million conventional mortgage applications. They found that compared to financially similar white applicants, lenders were:

+ 40% more likely to turn down Latinx applicants
+ 50% more likely to deny Asian/Pacific Islander applicants >br/> + 70% more likely to deny Native American applicants
+ 80% more likely to reject Black applicants

The director of PR for the **American Bankers Association** responded that the data used had "limitations" but declined to discuss the nature of the limitations or provide any additional data or evidence that would refute the findings.

Another techno-solutionist area where regulators are playing catch up is direct-to-consumer health apps, of which there are now 318,000 with 200 new apps popping up on a daily basis. The **U.S. Food and Drug Administration** regulates apps deemed to pose a "moderate to high risk to user safety", but the majority are left to the app stores for review. **Andrea Coravos** et al., proposed a **"nutrition label"** approach to standardize the way users inform themselves about health apps. **Germany** **adopted** an approach that allows state-based insurance to cover the cost of apps that have proven their therapeutic effectiveness to regulators. This being the U.S., there are also experts **calling for the market to sort it out** because regulation would cost too much and be "difficult."

SPONSORED CONTENT

Become a Founding Member of ADSA!
There's still time left, join by October 1st for special institutional member benefits
https://academicdatascience.org/get-involved/adsa-membership
ADSA

If your institution pays membership dues by October 1st, 2021, we will list you as a Founding Member organization on our website and in our Annual Member Book. Your institution will also receive 1 FREE Content Box in the DSCN. And we will give you a podium shout-out at the 2021 Leadership Summit and Annual Meeting! Click to Learn More

U.S. ASTRONOMY'S BIG TELESCOPES FACE PRECARITY

U.S.-based astronomers conduct a survey of priorities in their field once a decade that they use to guide funding, allowing the field to embark on long term, expensive projects like giant telescopes. The latest decadal survey is nearing release and could impact the fate of two telescope projects – the Giant Magellan Telescope (GMT) in Chile and the Thirty Meter Telescope (TMT) in Hawaii. Native Hawaiians object to having a large research telescope built on their sacred land, resulting in lawsuits for the TMT. The GMT is proposed for Chile and would join the Large Synoptic Survey Telescope (LSST) and Extremely Large Telescope (ELT), backed by Europeans, which has led some to wonder if there *must* be a new U.S.-backed telescope in Chile. It's not clear that the 2020 decadal survey will rank either the GMT, the TMT, or both as top funding priorities. Both projects **face an uncomfortable level of uncertainty**

AUSTRALIAN RESEARCH COUNCIL REJECTS APPLICATIONS THAT CITE PREPRINTS

In a bizarre, punitive, anti-science policy change, the **Australian Research Council** has **invalidated any grant application that cited a preprint** during its most recent funding round. At least 32 researchers had their applications rejected for citing preprints. The Australian physics and astronomy community was hit particularly hard. Many members have signed an **open letter** strongly opposing the policy change and accusing the ARC of promoting "academic misconduct."

FOLLOW THE MONEY

$40,929,700,000 **Harvard University's** current endowment, the largest in the world. See **top 30**.

$2,000,000,000 **National Security Administration** -> **Hewlett-Packard** for a 10-year contract to **improve data management, machine learning, and high performance computing** for the US's "relentless surveillance" agency.

$1,346,995,500 **German Ministry for Economic Affairs and Energy** -> **Tesla** for a **manufacturing facility** to make batteries and Tesla Model Y.

$50,000,000 **National Science Foundation** -> Quantum Leap Challenge Institute **will have** two lead schools – **University of Maryland, College Park** for Robust Quantum Simulation and **University of Chicago** for Quantum Sensing in Biophysics and Bioengineering, each of which will be joined by other institutions.

$20,000,000 **Chris Bickell** -> **University of Pittsburgh** football program. Following the terms of the agreement, the head coach will henceforth be referred to as the **[Insert Coach Name Here] Chris Bickell '97 Head Football Coach**.

$13,700,000 **Department of Energy** Advanced Scientific Computing Research program -> separate projects at **Oak Ridge National Laboratory**, **Texas State University**, **University of California-San Diego**, **Fermi National Accelerator Laboratory** **and others** to "develop new mathematical and computer-science techniques to shrink these data sets by removing trivial or repetitive data while preserving the important scientific information that can lead to discovery."

$100,000,000 **Germany** -> **World Health Organization Hub for Pandemic and Epidemic Intelligence** for a new, **still-forming project** to "to bring together, in real time, information on emerging public health crises."

$7,000,000 **National Institute of Environmental Health Sciences** 8-year grant -> **Robyn Tanguay** at **Oregon State University** to **study** the impact of "10,000 chemicals commonly found in food additives, medicines, consumer products and industrial chemicals" on zebrafish.

$3,500,000 **Chan Zuckerberg Initiative** -> **American Dental Association Science & Research Institute** to **create** "a cell atlas of the nose, mouth, and airways from birth through adolescence."

$2,000,000 **National Science Foundation** Emerging Frontiers in Research and Innovation program -> **Hongfei Lin** and a team from **Washington State University** and **University of Washington** to develop "catalytic processes to **improve plastics recycling** and make it cost effective."

$1,500,000 **National Science Foundation** Harnessing the data Revolution program -> **Penn State University** for **data science** curriculum development.

$1,000,000 **National Science Foundation** -> **University of Hawaii at Manoa** to "provide 230 elementary educators professional development in **how to promote CS and valued culture-based outcomes**."

$750,000 **Department of Energy** -> **Penn State University** for **data science** curriculum development.

$710,000 **National Science Foundation** -> **Rutgers University** (lead institute) with **Indiana University**, **Temple University** and **New York Institute of Technology**. "to build a nationwide, community-based **mobile edge sensing and computing infrastructure**"

$398,288 **National Science Foundation** Ethical and Responsible Research Program -> **Jason Borenstein**, **Ellen Zegura**, and **Charles Isbell** at **Georgia Tech** to "enable groups historically underrepresented in computing to" **directly influence** computing ethics curriculum.

$? **Mozilla**, **Omidyar Network**, **Schmidt Futures**, and **Craig Newmark Philanthropies** -> **Casey Fiesler** for a **spreadsheet** with **over 300** tech ethics syllabi.

NEW PROGRAMS

Cannabis Minor **City University of New York, Medgar Evers College** "science and business faculty collaborated" on the new minor. Data science is not the only hot new field.

Institute for Data Science **University of Mississippi** launched Spring 2020, no degree programs.

Google Intrinsic **Google** launched a robots-as-a-service platform.

LSST Interdisciplinary Network for Collaboration and Computing **University of Washington** and **Carnegie Mellon University** which is an entity set up to enable data storage and sharing from the Legacy Survey of Space and Time (LSST) telescope in northern Chile.

Lab to Life (L2L) **Purdue** an "innovation community" set up to enable invention during "the evolution from 5G to 6G" with an "open, neutral-host technology infrastructure" on a 400 acre campus.

Data Journalism Masters Degree **University of Maryland** either all online, all in-person, or a mix.

Davis Institute for Artificial Intelligence **Colby College** with **Amanda Stent** recently hired to lead.

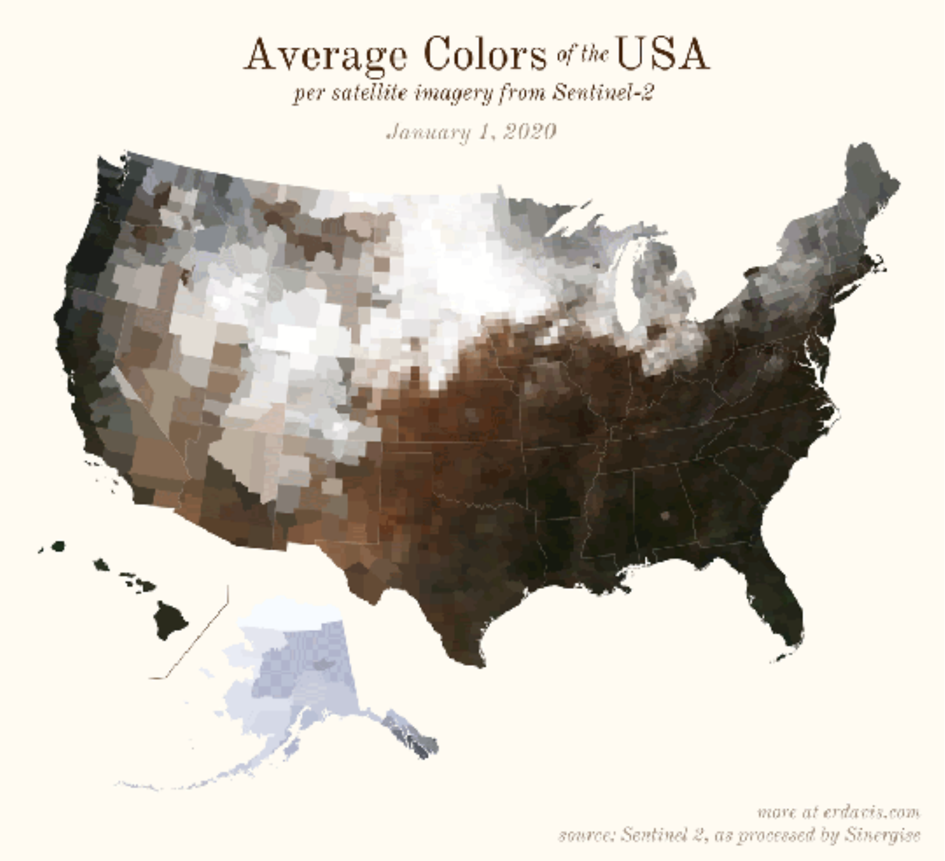
Data Sciences Institute **University of Toronto** with $10 million in grant funding available to University of Toronto faculty.

School of Mathematical and Data Sciences **West Virginia University** including a new Data Science BS degree.

Data-Enabled Computational Engineering and Science MS **Brown University** accepting applications now **for fall 2022** start.

DATA VISUALIZATION OF THE WEEK

Average Seasonal Colors of the USA in Erin Davis' Data Stuff blog from September 7, 2021


Average Colors of the USA
per satellite imagery from Sentinel-2
January 1, 2020
more at erdavis.com
source: Sentinel 2, as processed by Sinergise

Deadlines

Contests/Award

Model the Future to Win a Scholarship

"The **Actuarial Foundation** invites high-school students to conduct an actuarial research project in which they make recommendations to companies, organizations, government agencies, or other groups based on their mathematical models, real-world data analysis, and risk management." Deadline for submissions is mid-November.

Education Opportunities

Become a UW Data Science Postdoctoral Fellow

"Receive associated benefits like research funding and community engagement opportunities such as weekly activities and annual data science conferences. Incoming UW postdocs should apply by Sept 15th."

Tools & Resources

I am pleased to announce that the camera ready version of my new textbook, "Probabilistic Machine Learning: An Introduction", is finally available from http://probml.ai.

Twitter, Kevin Patrick Murphy from August 30, 2021

"Hardcopies will be available from **MIT Press** in Feb 2022."

Research Spotlight: CDS and Grid AI researchers propose new method of self-supervised learning
Medium, NYU Center for Data Science from August 27, 2021
"In the realm of self-supervised learning, state-of-the-art algorithms can broadly be divided into two groups: contrastive approaches and non-contrastive approaches. But what if there could be a third, unique approach? This is the main question that was answered in a recent publication authored by CDS PhD Student **William Falcon**, along with researchers at **Grid AI** and CDS faculty **Kyunghyun Cho**. The paper, titled 'AAVAE: Augmentation-Augmented Variational Autoencoders' introduces the titular augmentation augmented variational autoencoders."
Computer Scientist Explains Machine Learning in 5 Levels of Difficulty
YouTube, WIRED from August 18, 2021
"**WIRED** has challenged computer scientist and **Hidden Door** cofounder and CEO **Hilary Mason** to explain machine learning to 5 different people; a child, teen, a college student, a grad student and an expert." [video, 26:08]

Events

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

Lisa Nakamura’s work is and has always been Star-struck and I am super proud that she is the next speaker in MSR’s (free, public, recorded, awesome) Race and Tech lecture series.
Online September 22, starting at 10 a.m. Pacific.

Introducing the Bay Area Open Science Group
Online September 28, starting at 2 p.m. Pacific, with "**Biftu Mengesha** from **UCSF** discussing Innovating Education in Reproductive Health."

Melissa Dell - LayoutParser: A Unified Toolkit for Deep Learning-Based Document Image Analysis
Online October 4, starting at 12 p.m. Pacific. **Stanford Digital Economy Lab** Seminar Series: "**Melissa Dell** of **Harvard University** will join S-DEL Director **Erik Brynjolfsson** to discuss of LayoutParser, an open-source library for streamlining the usage of DL in DIA research and applications." [registration required]

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://academicdatascienceorg/resources/newsletter. Our mailing address is 1037 NE 65th St #316; Seattle, WA 98115.