Data Science Community Newsletter

Data Science Community Newsletter features journalism, research papers and tools/software for November 23, 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 .

Academic Data Science News

New and much needed approaches for teaching data science include putting a Practicum course at the center of a data science program, intentionally organized as a hybrid between an educational classroom and an industry-like environment, as **Boston University** has done for the past 5 years in their MS in Statistical Practice (MSSP) program. **Harvard's Access to Justice Lab** is experimenting with using actuarial risk assessment mechanisms, one of many evidence-based practices now used in legal spheres, while a **Cornell Law School** professor teamed up with **Cornell** computer science researchers to develop a model language for real estate conveyances using programming language as a structure. And researchers at **Caltech** are using machine learning to solve partial differential equations. PDEs are the cornerstone of engineering education, a fulcrum for linking natural phenomena to mathematical, techncial problem solving, something that an AI-solver for PDEs won't change. What will change is the psychic torture students incur using approximation methods to solve PDEs. Good-bye Finite Elements. You will not be missed.

The Internet of Things will be here sooner, rather than later, according to **Pete Warden**, blogger and lead of the **TensorFlow Mobile/Embedded team at Google**. "What I see is that there is a lot of latent demand for technology that I believe will become feasible over the next few years," he recently wrote, "and the scale of that demand is so large that it will lead to a massive increase in the number of embedded devices shipped." Warden forsees a low-power, low-price threshhold that, once broken, will open the floodgates. MCUNet is an example of the disruption Warden expects: faster and more accurate machine intelligence at the Internet's edge, with lower cost and power requirements. **Song Han's** group at **MIT Department of Electrical Engineering and Computer Science** used two "tiny deep learning" components to invent MCUNet. TinyEngine is a sort-of operating system that directs resource management. TinyNAS is a neural architecture search

algorithm that feeds compact neural nets matched to TinyEngine's available microcontrollers. “Everything put together is just one megabyte of flash," says Han. Industry standards for IoT devices are, wisely, treading wwater, as technologies like MCUNet come online. For example, vendors recently forked a new standalone working group, **Project Connected Home over IP** (ProjectCHIP), that will enable IoT devices in homes, offices and commercial spaces to communicate.

It takes at least 140 scientists from all corners of the globe to get your animal database published in either **Science** or **Nature**. The small sample (5) constituting that sweeping generalization about databases is: bird genomes, mammal genomes, Arctic animals in motion, bees and earthworms. Each corpus is a monumental effort and a big data window into global change. These are efforts that started as a good idea and snowballed. These datasets will aid animal conservation and increase our understanding of evolution. They also reflect the enormous diversity, but also some of the bias in contemporary science – northern species, nearer to richer nations, predominate. There are exactly 33 types of earthworm in the UK, but there are far, far more unknown types of earthworm in the tropics. "Every time they dig a new hole, they find a new species," says **Helen Phillips**, a soil ecologist at the German Centre for Integrative Biodiversity Research. The authors from the **Zoonomia Project**, the mammals dataset, "found that that species with less genetic diversity have higher extinction rates." And they know that 47 mammals have a "high likelihood of being reservoirs or intermediate hosts for the SARS-CoV-2 virus." This is our world at its closest to ground truth.

Google released the freely available Objectron Dataset, an effort to provide a benchmark for 3-dimensional object detection, trying to do in 3-D what Imagenet does in 2-D. Objects in Objectron are non-human, which should help it to avoid bias issues that are becoming a source of distress with Imagenet, a library of 2-D Internet photos frequently used to train facial recognition. Researchers from **Carnegie Mellon** and **George Washington University** are "quantifying biased associations between representations of social concepts (e.g., race and gender) and attributes in images" in Imagenet, potentially introducing bias from source data into models.

One more dataset, a "developmental atlas" of gene expression in fruit fly neurons, was published in **Nature** by **NYU** researchers led by **Claude Desplan**. Like these other giant datasets it was an immense undertaking and it highlights the incredible diversity found in nature; blah, blah, blah. Noteworthy here is how the atlas was created, using single-cell mRNA sequencing, a method that pulls meaningful data from small bits of genetic code inside cells.

Repeat the sequencing for hundreds of thousands of cells and machine learning can differentiate and spell out the function of those cells.

Controversies abound in the wider world of higher education. The latest chapter in a long-running court case saw a federal appeals court rule that **Harvard** does not intentionally discriminate against prospective Asian American students, paving the way for a **Supreme Court** challenge. While the high court has upheld the use of race as a factor in college admissions multiple times, that could change under the court's new conservative majority.

Facebook is once again trying to hinder independent research when it demanded that **NYU** researchers stop using the web plugin Ad Observer, which copies political ads that users see on Facebook and puts them in the **Ad Observatory** public database with the aim of increasing transparency on ad targeting. Ad Observer is still up and running, and uncovering targeted ads posted by the "fringe conspiracy QAnon movement." The found ads, according to **Jeremy Merrill** at **The Markup**, surface methods for sidestepping Facebook oversight of misleading, potentially incendiary content.

Students learning remotely are being subjected to absurd levels of exam monitoring surveillance – while paying full tuition, we should note, as well as in some cases using software that fails to recognize students with darker skin tones. Students are required to do things such as banging their foam earplugs on their desktop on-camera to prove they are not earbuds. Monitoring software such as Respondus is facing a growing backlash. In all of these cases – stay tuned.

The Board of Regents for the **University of California System** approved construction for a 415,000 square foot Data Hub building at the **University of California-Berkeley**. The building will house the school's inter-disciplinary **Division of Computing, Data Science and Society**.

University of Washington now offers a minor in data science. The program is for non-STEM students and was developed primarily for students in the arts, social sciences and humanities.

University of Arkansas's College of Engineering announced a new multidisciplinary data science scholarship program funded by **National Science Foundation** to enhance and increase the graduation rate of underrepresented STEM undergraduate students interested in careers in data science.

The **University of Cambridge** has officially launched its new **Centre for AI in**

Medicine, funded by **AstraZeneca** and **GlaxoSmithKline** and focused on developing AI and ML technologies aiming to transform clinical trials, personalized medicine and biomedical discovery.

Georgia State University Library is launching the Public Interest Data Literacy (PIDLit) initiative to expand programs promoting data literacy and career preparedness with a focus on reaching first-year students and underrepresented groups.

New York University's Masters in Data Science will now offer an Industry Concentration, targeted to respond to the needs of companies and includes a required internship and practical training.

University of Maryland has a grant from the **U.S. Department of Agriculture National Institute of Food and Agriculture** (USDA-NIFA) to develop a next-generation food safety risk assessment model by combining emerging techniques in both food safety and machine learning.

The **Cascadia Data Alliance** has announced new awards to fund medical research at the **Fred Hutchison Cancer Research Center**, **University of Washington eScience Institute**, **BC Cancer**, **University of British Columbia Data Science Institute**, and the **Knight Cancer Institute at Oregon Health and Science University**.

Editor's Picks

History can often go unnoticed amid technology's shiny newness, but the stories are nonetheless incredible. **Marcel Salathe** created a modern disease spread model for a 1630-31 plague outbreak in Venice after digitizing historical records. And the **Carrier** record label recently released a 1984 concert performance by multi-instrumentalist **George Lewis**, an important digital music pioneer.

A survey by the **Financial Times** finds growing resentment among under-30s, who's economic well-being has been affected disproportionately by the pandemic.

Camera traps on steroids! Canadian conservation biologists have high hopes for the algorithms to recognize, then follow and track, individual grizzly bears. It's a strategy that has increased public concern for orcas, but bears lack the whales' distinctive body markings. Researchers find that facial features are the best grizzly auto-indicators, even when bears get super-fat before hibernation. Bio-logging is a similar but more invasive technology that attaches the camera + sensor packages to the animals. **Osaka University** researchers use AI to detect seabirds' behaviors using low-cost sensors, which in turn activates and

deactivates the lightweight but high-resource cameras. Seabird privacy is not a consideration. Agri-scientists have been using computer vision for years to identify traits that generate superior yields. Canadian and American soybean scientists have developed an effective analysis pipeline and are getting very good at linking genetics to root system architectures. They can then optimize RSAs for whatever growing conditions. Below-ground traits, no surprise, are more interesting than those superficial, easy-to-spot above-ground traits.

University of Minnesota researchers published research on using new methods to assess how biodiversity loss impacts forest ecosystems by determining how sunlight reflects off the surface of the forest canopy using spectral images taken from an airplane. **L'Oreal** is using its store-locator algorithm to direct people to make-up recycling centers. And the "Right to Repair" bill passed by a wide margin in **Massachusetts**, making a bold move against our throwaway culture.

OTOH, the carbon footprint of AI and cloud computing is becoming an increasingly urgent concern, says **Sarah DeWeerdt** in **Anthropocene** (link) and **Jacob Dykes** in **Geographical Magazine** (link). Artificial lighting is known to be extremely disruptive to the natural world, say **University of Exeter** (UK) researchers. To top it all off, "Do Americans really care about climate change?," asks **Joshua Eferighe**. (Spoiler alert: The majority do not think it will impact them directly. So much for trusting science.)

There's a push-me-pull-you happening in the privacy sphere as **Tim Berners-Lee** and **MIT** kick off four pilots of the Solid data-privacy technology, the **Canadian Privacy Commissioner** (btw, when will the US have a Privacy Commissioner? and "never" is not an acceptable answer) has recommendations for regulating AI. Meanwhile users are becoming increasingly wary of how their payment app data is used and stored, how big data companies are accessing their personal data, and how AI is influencing our buying habits.

Research News

COVID vaccines are a 21st Century technology transfer success story, so far. Vaccines made by a **BioNTech**-**Pfizer** collaboration and by **Moderna** are on track for **FDA** emergency approvals and public inoculations. What does this miracle of science tell us about the role of data science in contemporary technology transfer? **STAT** produced an excellent longform origin story on mRNA vaccines for COVID (link) and the European research magazine, **Horizon**, has an excellent mRNA vaccine "5 Things" explainer (link). The first challenge one faces in making an mRNA vaccine is somehow

avoiding the body's immune response when these artificial compounds enter cells. Finding the perfect modified nucleoside (mRNA building blocks) to make fake mRNAs work like real mRNAs is data intensive, difficult work. BioNTech-Pfizer have accumulated experienced working on mRNA vaccines against flu, which has helped despite a lack of flu vaccine success. Alternatively, Moderna was given billions of **U.S. Operation Warp Speed** dollars to fund their work. The second challenge then becomes the human testing necessary to show clinical effectiveness. Both vaccines require two doses, in order to tamp down any immune response that might occur. BioNTech-Pfizer completed testing first, claiming an astonishing 95% effectiveness, because it requires 3 weeks in between doses, whereas Moderna requires 4 weeks in between doses. Pfizer has applied for FDA emergency approval of its vaccine.

Once a vaccine gets all the way to widespread commercialization, there's another, more epidemiological data science problem: Who gets vaccinated? BioNTech-Pfizer have been as transparent as possible throughout their development process, a strategy that they hope will counter any skepticism that the public expresses toward their vaccine. **Pew Research** figures suggest that only 51% of Americans would "definitely" or "probably" go for COVID vaccination. If only half of America gets vaccinated then it helps to give shots to the right, not wrong, half. **Christopher Cox**, writing at **WIRED**, says it's better if we inoculate the most social before the most at-risk. There's good research out of **Stanford** and **Northwestern** that cross references metro mobility data and COVID case data to create rate of infection models. The study led by **Jure Lescovec** used data collected between March and May and identified hot spots for infection in urban areas during that time window. Maybe vaccination policy isn't a Who question, but a Where question.

A COVID vaccine doesn't do anything to repair the psychological damage caused by the pandemic. Ample evidence is showing that our current times are tough. The original population-scale optimism-pessimism gauge, the Hedonometer developed by **Peter Dodds** and **Chris Danforth**, now at the **University of Vermont Complex Systems Center**, has been around since 2008, using a randomly selected ten percent of daily Twitter tweets to measure happiness. Dodds recently spoke with **Alex Goldman** on the popular **Reply All** podcast. Apparently the worst day in Hedonometer history was this past May 31, a Sunday in a weekend full of George Floyd protests. **University of Washington** professor **Tim Althoff** and Ph.D student **Jina Suh**, with **Ryen White** and **Eric Horvitz** from **Microsoft Research**, used search data to gauge fundamental human needs (based on Maslow's hierarchical framework), again at population-scale. Aspirational needs, like goal-setting, declined significantly, they found. These population-scale studies, and another study from **MIT** based

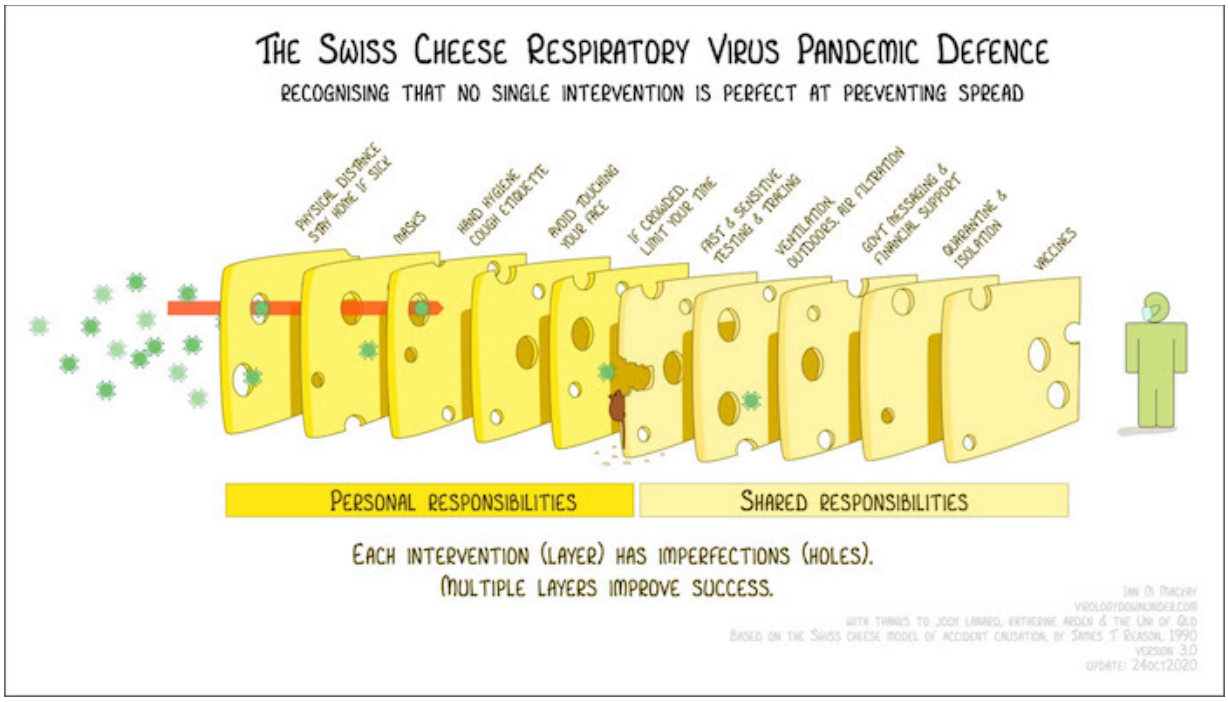
on a gigantic **Reddit** dataset accumulated between January and April 2020, are evidence of significant drops in our collective well-being. If you believe that a social media datum is a thinking, breathing human, and you realize that "expressions of basic needs like health and financial security rose dramatically during the pandemic," as Althoff told DSCN by email, then policymakers can't stay deaf to this information.

Tweet of the Week

Twitter, Zachary Levenson from November 19, 2020

Data Visualization of the Week

Jason Kottke, Ian Mackay from November 16, 2020



Deadlines

Contests/Award

NFL Big Data Bowl 2021 - Help evaluate defensive performance on passing plays

"This competition uses NFL's Next Gen Stats data, which includes the position and speed of every player on the field during each play. You'll employ player tracking data for all drop-back pass plays from the 2018 regular season. The goal of submissions is to identify unique and impactful approaches to measure defensive performance on these plays." Deadline for submissions is January 7, 2021.

Conferences

Networks 2021: A Joint Sunbelt and NetSci Conference

Washington, DC July 6-11, 2021. "We expect this to be the largest networks conference ever held. It will combine the annual meeting of the **International Network for Social Network Analysis** (Sunbelt XLI), and the annual meeting of the **Network Science Society** (NetSci 2021)." Deadline for abstracts submissions is January 24, 2021.

Education Opportunities

Rising Stars in Data Science

Online January 11-12, 2021. "The Rising Stars in Data Science workshop is a new initiative from the Center for Data and Computing (CDAC) at the University of Chicago, focusing on celebrating and fast tracking the careers of exceptional data scientists at a critical inflection point in their career: the transition from PhD to postdoctoral scholar, research scientist, or tenure track position."

Deadline for applications is November 23.

Stanford Population Health Summer Research Program

"To provide training and experience in population health research for college students who are from underrepresented and historically excluded groups in the health sciences." Deadline for applications is January 15, 2021.

Studies/Surveys

Observable Community Survey

"We expect this survey to take no more than 5 minutes to complete. We really appreciate your feedback!"

Tools & Resources

The Future of Distributed Machine Learning

Coiled, Andreas Müller & Hugo Bowne-Anderson from November 03, 2020

"We recently chatted with Andy Müller, core developer of scikit-learn and Principal Research Software Development Engineer at Microsoft. Andy is one of the most influential minds in data science with a CV to match. He shares his thoughts on distributed machine learning with open-source tools like Dask-ML as well as proprietary tools from the big cloud providers."

big news: we are starting a non-profit! It is called 2i2c, which stands for "The International Interactive Computing Collaboration".

Twitter, Chris Holdraf from November 10, 2020

"2i2c has a few core goals:"

- "Manage interactive computing infrastructure for research and education"
- "Develop and improve tools in interactive computing for these use-cases"
- "Support open source tools and communities that underlie this infrastructure"

First release of the Array API Standard

Consortium for Python Data API Standards from November 10, 2020

"The main goal of this standard: make it easier to switch from one array library to another one, or to support multiple array libraries as compute backends in downstream packages. We'd also like to emphasize that if some functionality is not present in the API standard, that does not mean it's unimportant, or that we're asking existing array libraries to deprecate it. Instead it simply means that that functionality at present isn't supported - likely due to it not being present in all or most current array libraries, or not being used widely enough to have been included so far. The use cases section of the standard may provide more insight into important goals."

Data literacy training: What you need to know

The Enterprisers Project, Piyanka Jain from November 12, 2020

"Successful data literacy training programs are never one-size-fits-all. Consider this expert advice to avoid common mistakes and design a data skills plan that works."

Subscribe to the New NSF CISE Newsletter

Computing Community Consortium, CCC Blog from November 10, 2020
"The National Science Foundation (NSF) Computer and Information Science and Engineering (CISE) directorate just announced a new newsletter that will share 'periodic updates about CISE and NSF broadly, including up-to-date information about [their] newest programs and activities.'"

Click here to receive the Data Science Community Newsletter and/or to have us follow your twitter feed so that our data science twitter bot can easily grab links from your tweets.

Data Science Community Newsletter Issue 206