ADSA Data Science Community Newsletter

Data Science Community Newsletter features journalism, research papers and tools/software for August 27, 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.

WELCOME BACK TO SCHOOL

As you start back to class, please consider recommending the **Data Science Community Newsletter** to your students and colleagues – especially masters and grad students. It helps them grasp the larger context in which data science operates and includes links to events and jobs that may be great for them.

DATA SCIENCE IN REAL ESTATE

The real estate industry is data-rich with huge social, economic and ecological impact. I talked to three people in the industry to understand which data science methods are most common, which problems "count" as relevant, and how data science findings are presented to consumers. Using data science in real estate can mean developing automated tools to detect misleading input from realtors trying to rent Manhattan apartments to developing flood and fire risk assessments for every property in the US so buyers can make informed choices as climate change reshapes mortgage risk and livability. **Click through** for more *From the Desk of Laura Norén*.

DEPRESSING CASE OF ACADEMIC FRAUD

Joe Simmons, **Uri Simonsoh**, and **Leif Nelson** **writing** for their blog **Datacolada** and working with a team of anonymous researchers uncovered academic fraud in a famous 2012 **paper** by **Shu**, **Mazar**, **Gino**, **Ariely**, and **Bazerman** that found honest behavior can be increased if people are asked to sign a statement of honest intent before providing information. In 2020, the five original authors plus **Kristal** and **Whillans** were **unable to replicate** that finding, which led to a reexamination of the 3 studies that made up the original paper.

It appears highly likely that the data from the third study, provided by Ariely, were fabricated. Three key giveaways: the distribution of miles driven should be a normal curve with a long right tail (some people drive a little, most people drive a moderate amount, others drive a vast amount). But the study distribution looked like a solid block of equally little to moderate driving with nobody going more than 50,000 miles in the study period. Second, when people are asked to self-report large numbers, they are somewhat likely to round to multiples of 100, which they did in the Time 1 data (25% of reported values ended in zero), but not in the Time 2 data (10% of values ended in zero). Third giveaway, half of the data were in Calibri, the other in Cambria. Each of the Calibri data points had a Cambria data point that was a near twin, with what appeared to be a slight, randomly generated perturbation. The Datacolada authors ran 1 million simulations to see if they could replicate the apparent twinning of the Calibri/Cambria pairs, "Under the most generous assumptions imaginable, it didn't happen once. ... These data are not just excessively similar. They are impossibly similar." The datacolada authors cannot assign blame, but they can narrow it down to the fourth author (Dan Ariely), someone in the fourth author's lab, or the insurance company (**The Hartford**) that was involved in gathering the data. The authors further conclude, "scientific fraud is more common than is convenient to believe...eliminating it should be a collective endeavor. Data should be posted." I hope you agree.

[**Reporting** by **Stephanie Lee** of **BuzzFeed** digs into the blame game further.]

UKRI REQUIRES IMMEDIATE OPEN ACCESS; ARXIV TURNS 30

Granting paywall-free open access to research findings has notched another win in the United Kingdom where the tax-payer funded **UK Research and Innovation** has **announced** all recipients of UKRI funding must make their published papers available for free upon publication starting in April 2022. Papers resting on research funded by UKRI must use a Creative Commons copyright license (CC-BY) which legally allows for the liberal distribution of the work. Two other major funders of UK-based research – the **Wellcome Trust** and the **European Research Council** – also mandate open access publication.

arXiv, the original pre-publication "open research sharing platform" founded by **Paul Ginsparg** at **Cornell** **celebrated its 30th anniversary** this August. Ginsparg also **wrote** in *Nature* that open access pre-print servers have done a better job of handling the need to publish COVID studies quickly and accurately than traditional publishers.

COVID ON CAMPUS – WORSE THIS YEAR THAN LAST?

On some campuses already back in session, COVID is already playing out with worse outcomes compared to the same time last year. **University of Wisconsin-Madison**, which does not have a vaccination mandate, set aside empty dorms and used empty hotels last year for quarantine. This year dorms are at full capacity, hotels are full of football fans, and sick students are **lodged in several family housing units in a complex with children who cannot be vaccinated**.
At **University of North Carolina Wilmington**, Professor **Kevin McClure** **tweeted** that they added "another 122 positive cases [yesterday]. over 300 cases total. 52 out of 150 quarantine beds in use. This time last year we added 3 positive cases." It would be unwise to make any broader assumptions about college operations based on two schools. **Dr. Jeremy Farrar**, head of the **Wellcome Trust**, **said** "We've only gotten 18 months into this pandemic, and the pandemic is going faster today than it was in 2020 and much of 2021." ... "We've had 5 pandemic changing variants in the last 6-9 months and that isn't going to slow down." Humility remains a virtue as we decide how to proceed.

REGULATORY DILEMMAS

What happens when big tech companies buy smaller startups? In this case the focus is on the AI space, but the **U.S. Federal Trade Commission** (FTC) is reviewing its role as an anti-trust regulator in the context of large incumbents acquiring startups. Former Chief Competition Economist of the **European Commission**, **Tommaso Valletti**, argues that the five biggest tech companies in the US have acquired 1000 startups in the past 20 years and been blocked 0 times by the FTC. He argues that instead of making regulators prove a merger will be harmful, merging parties should prove their merger won't be harmful. I don't see the legal logic in leaving the burden of proof to the parties who will benefit from the merger. Still, given that startups tend to represent small market share and want to be acquired for the right price, it's not clear that existing regulatory guidance applies or that there is a coherent framework available to identify which acquisitions or acquisition patterns would threaten the fair functioning of the market and/or consumer and community goals. FTC Commissioner **Rebecca Slaughter** is considering new rule-making to address this and other regulatory gaps.

As **Google Health** continues to restructure, shedding its internal leader **David Feinberg** is now the CEO of **Cerner**. Why is health such a hard sector to crack? **IBM Watson** overpromised, then flamed out years ago. **Google** hasn't put forth dramatic wins yet. **Apple** has gotten into consumer health and nibbled on patient care. A range of smaller startups – **Oscar**, **Lemonaid Health**, **Flatiron Health** – have all been incubating their own unique approaches. **STAT** journalist **Casey Ross** discusses the possibility that lax regulation makes it difficult to judge the success of algorithms in clinical settings.

MISINFORMATION SAGAS – FROM CLIMATE CHANGE TO COVID BACK TO CLIMATE CHANGE

We have reported on misinformation as it relates to COVID, but before that, there was a robust misinformation campaign related to climate change. **Carnegie Mellon University** researchers **Aman Tyagi** and **Kathleen Carley** **investigated** the depressing set of conspiracy theories in the climate change misinformation net. They found two that stood out for disbelievers: "the chemtrails theory, which claims that the trails following high-altitude jets are chemical agents being sprayed for nefarious purposes; and the geo engineering theory, which claims that government experiments are causing climate change". There are no identifiable theories or tropes that stand out among believers. When your position rests on scientific evidence, maybe you don't need contrived metaphors and conspiracy theories?

ARE WE DEBATING AI OR THE LIMITS OF CIVIL LIBERTIES?

Prisoners and the people they talk to on the phone are subject to surveillance by a natural-language processing algorithm in New York, Georgia, and Alabama. **Reuters** **reports** that privacy advocates are worried about racial bias. As we discussed last week, that is not the key philosophical critique. It is plausible that the surveillance could be developed in a way that wasn't racially biased. Would that make it acceptable? A key question from legal ethics: Do prisoners and their non-incarcerated conversation partners have a right to avoid constant surveillance? Depending on how/when/where human rights are found to end and prerogatives that prioritize other objectives begin, we may object to persistent surveillance outright.

ARE CANADIAN COMMUNITY COLLEGES TAKING ADVANTAGE OF PUNJABI FARMERS?

"In 2019, 34 percent of the more than 642,000 international students in Canada were from India...[many] from Punjab, and they generally attend small community colleges." *The Walrus* **reports** that a surprising number of Indian farmers are making financially perilous arrangements in order to send their children to Canada, guided by recruiters who imply that permanent residency is just a student visa, community college degree, and postgraduate work permit away.

Careers

See the ADSA Jobs Page for more opportunities.

acornai

Data Scientist at Acorn AI

Location: New York, Boston

CDO TITLE – GOOD ON PAPER, COULD BE AN ORGANIZATIONAL DEATH SENTENCE

Be wary of becoming a Chief Data Officer. Sixty-five percent of "large, data-intensive firms" have one, but the average tenure is only 2-2.5 years, **according** to **Tom Davenport**, **Randy Bean**, and **Josh King.** They note that, "transformational change" is often expected within 18 months, though it is often unclear what is in scope and how objectives ought to be prioritized. Further, "their C-suite compensation levels plus their lack of C-suite history and lack of well-developed C-suite political savvy often makes them targets from day one."

A FINAL PIECE OF BETTER NEWS

During the pandemic, subscribers to the **WHOOP** wearable platform who submitted information about their sleep and exercise habits in 2019 and 2020 were found to both sleep more and exercise more intensely in 2020 during social distancing than in 2019. The authors, **Emily R. Capodilupo** of WHOOP and **Dean J. Miller** of **Central Queensland University** **surmised** that increases in sleep minutes were linked to decreases in "social jet lag" or the loss of sleep due to staying out later on weekends. The study sample should not be generalized to broader populations. WHOOP users are likely to have been more interested in healthy living and the sample was 2/3rds men.

FOLLOW THE MONEY

~$8,600,000,000 **MacKenzie Scott** -> 375 non-profits, generally in unrestricted funds
[**Bloomberg** catalogued ~half of the funds from the secretive "money cannon"]

$1,600,000,000 **Penn Medicine** for **The Pavilion**, a 1.5 million square foot medical center where 10,000 people will work starting later this year.

$1,540,000,000 research funders -> **University of California-San Diego** **"the largest ever"** sponsored research budget in the school's history. [UC San Diego beat UC Berkeley in the research funding race]

$1,050,000,000 research funders -> **University of California-Berkeley** **the first time research funding topped $1 billion at Berkeley**

$250,000,000 **John Deere** -> **Bear Flag Robotics** to **acquire the Silicon Valley startup** "that develops autonomous-driving technology"

$175,000,000 **University of Georgia** -> **Georgia Bulldogs** football program for **"350,000 sq feet of football facilities...since 2015"**

$75,000,000 **National Science Foundation** -> **University of Michigan**, **University of Maryland**, **City University of New York**, **Princeton University**, **University of Southern California**, plus additional partner and affiliate academic institutions throughout the U.S. for five new **NSF Innovation Corps** (I-Corps) Hubs, which will be the backbone for a new nationwide National Innovation Network to **facilitate technology transfer and entrepreneurship**

$60,000,000 **Belmont University** -> **Belmont Data Collaborative** "to build a portfolio of data science programs...both degree programs and continuing education services for working professionals" in Tennessee.

$20,000,000 **UPMC** -> **Novasenta**, a Pittsburgh immunotherapies startup that **uses machine learning to identify drug targets**

$3,800,000 **EduLab Capital Partners** and **Allos Ventures** -> **Codelicious** to "**provide full-year computer science curriculum to K-12 schools**"

$3,000,000 **National Science Foundation** -> **Case Western Reserve University** and **University of Pittsburgh** for the new **Center for Materials Data Science for Reliability and Degradation** (MDS-Rely)

$2,000,000 **National Institute on Aging** -> **Rendever** to **study** "the effects of [Rendever's] virtual reality platform on senior living residents and their families"

$1,970,000 **National Institutes of Health** -> **Shankar Mukherji** of **Washington University** to **understand** "how a cell commits resources to building new parts — and eventually divides into two cells"

$1,300,000 **National Science Foundation** + **National Institutes of Health** -> **Georgia State** for the **study** of "causal learning models in the brain"

$1,000,000 **MassMutual** -> **Boston University Faculty of Computing and Data Sciences**, **spread out over three years**

$1,000,000 **Conoco Phillips** -> **Texas A&M University** to **establish an undergraduate certificate program** **establish an undergraduate certificate program** in Data Analytics for the Petroleum Industry (CERT-DAPI)

$33,140 **Iowa State University** -> consultants at **History Associates Inc.** to **"gather and organize factual evidence"** as the school considers removing the name **Carrie Chapman Catt** from a campus building. Chapman Catt supported women's suffrage and white supremacy. See also: a related **Columbia Missourian** newspaper **piece** on the trouble with naming buildings after people.

NEW PROGRAMS

The **University of Pittsburgh School of Information and Computing** has, after years of planning, **released the course requirements** for students to major in data science.

George Washington University is **adding bachelor of science degrees** in Data Science, Cognitive Science of Language, and Psychological and Brain Sciences to its **Columbian College of Arts and Sciences**.

The **U.S. Centers for Disease Control and Prevention** have enlisted epidemiologists **Marc Lipsitch** and **Rebecca Kahn** from **Harvard T.H. Chan School of Public Health** to establish a new **Center for Forecasting and Outbreak Analytics** at CDC.

Lawrence Berkeley National Laboratory is building the new **Surface Atmosphere Integrated Field Laboratory** (SAIL) outside Crested Butte, Colorado. It is **an effort to improve data capture for high-elevation water cycles** in the Western U.S.

Data Science and Analysis Bachelors **University of Missouri-St. Louis**

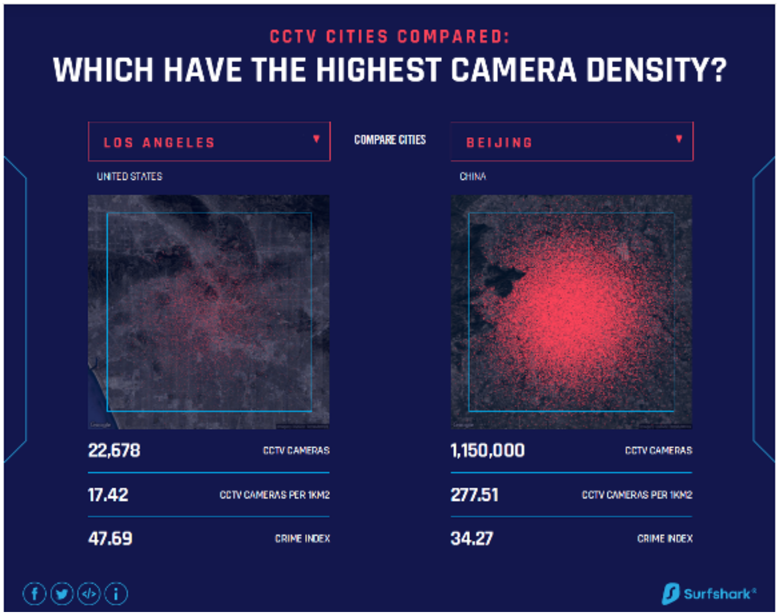
Information Management and Technology Bachelors **Syracuse University** Part-time, fully online or hybrid.

Marquette University **added** an online master of science degree in Health Care Data Analytics. The program is supported by **Everspring**, a Chicago-based education techology services company.

Infosys will build **its first digital development center** in Canada, a 500-person office located outside Toronto, and modeled after the six similar centers that Infosys operates in the U.S.

DATA VISUALIZATION OF THE WEEK

Mapped: The Top Surveillance Cities Worldwide in Visual Capitalist, Avery Koop from January 1, 2021



Deadlines

Contests/Award

[Rethink EMG Challenge]

"The **De Luca Foundation** is offering four awards, in the form of cash and surface Electromyography (sEMG) equipment, for proposals using sEMG to augment current performance assessment methods in professions (PT, OT, Athletic Training, Fitness, Geriatrics, Speech, Telehealth) in which human assessment of movement in health and disease are needed." Deadline for submissions is October 31.

Education Opportunities

"Applications are now open for the Translational Global Infectious Diseases Fellowship at @uvmcomplexity

"We offer early-career scientists a unique opportunity to address pressing modeling challenges in infectious diseases epidemiology!" Review of applications begins on October 1.

Conferences

The deadline to submit papers to the INDIS (Innovating the Network for Data-Intensive Science) workshop has been EXTENDED to September 5!

Online "The 8th annual INDIS Workshop will be held in conjunction with **SC21** on Monday, November 15, 2021 in St. Louis."

Events

See the ADSA Events Page for more details and 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. Our archive of newsletters is at https://academicdatascience.org/resources/newsletter. Our mailing address is 1037 NE 65th St #316; Seattle, WA 98115.