Data Science Community Newsletter

Data Science Community Newsletter features journalism, research papers and tools/software for November 6, 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 .

Errata

Last issue we highlighted but forgot to link to an essay at Wired.com by Casey Fiesler and Natalie Garrett. Please re-visit their essay, Ethical Tech Starts With Addressing Ethical Debt, now properly linked. Sincere apologies.

And anyone who has spent much time in New Haven knows that Southern Connecticut State University is west of Yale University, not east.

Academic Data Science News

University silos no longer provide an adequate understanding of the pandemic in a college community. A pre-print, not peer reviewed study in La Crosse, Wisconsin by Kabara Cancer Research Institute, showed increases in COVID cases among the community's senior living residents have been linked genetically to student cases at the University of Wisconsin-La Crosse branch. In Vermont, a/k/a Fort Vermont for its success at keeping coronavirus out, St. Michael's College has had an outbreak that has been linked to a series of spread events at a community hockey rink 40 miles away. The chains of coronavirus transmission are person-to-person, and the extent is reflected in its R0, the basic reproduction number. Another study at the University of California-Irvine, found that stress and depression stemming from personal and emotional loss during the pandemic was pervasive, and the severity was frequently the result of exposure to mass media coverage of COVID-19. Health and well-being are both products of the staggering complexity that is human interaction, but there's progress being made. C. Brandon Ogbunu, writing at Wired.com, says to be optimistic. Network science, he feels, is coming of age as an essential method for grasping phenomena, like this pandemic, that are complicated and interconnected.

Cornell University has been held up by Bloomberg News (link) and Inside Higher Ed (link) as a model students-in-attendance campus. Infection test positivity rates have been as low at 0.006%. Bloomberg's writer, Emma Court, a Cornell alum, provided some backstory, explaining the pivotal role Provost Michael Kotlikoff played; as a veterinary scientist with in-depth understanding of infectious disease, he oversaw creation of the school's COVID testing lab, based in Cornell's Veterinary College. The University of Notre Dame brought its on-campus testing lab online in late-September. This was before the school's President, John Jenkins, contracted COVID, testing positive days after attending a crowded, mostly maskless White House event. By mid-October new cases were again spiking at the South Bend school, and Notre Dame announced that the maximum size of any on-campus gathering would be reduced from 20 to 10 people. (MIT research now shows that limiting gatherings to 10 or fewer people reduces super-spreader events.)

The cause of the outbreak: football. Attendance at the Saturday night football game between Notre Dame and Florida State did not play a role, but tailgating and watch party gatherings did. Football has been implicated in coronavirus surges in Texas college towns like Lubbock, home to Texas Tech University. Washington State University researchers have mathematical models showing in-person sports events increase COVID-19 cases by 25% in their best-case scenarios. Big football schools in the Southeast and Midwest rank high on the list of universities with the most COVID cases. Cornell and Notre Dame had both experienced outbreaks in August tied to athlete populations, but were able to regain control. The Stanford Daily student newspaper reported that athletes account for more than half of Stanford University's COVID cases. In the article, professor Yvonne Maldonado points to the types of gatherings and social networks among and around athletic teams as sources of additional risk faced by students who play inter-collegiate sports. Mitigating the risk has created a double standard, seen at

the University of Michigan which put the general student population into lockdown while the school's football team went ahead with its games against University of Minnesota and Michigan State University.

It's a mistake to point a giant foam finger and say "Sports Bad!" when there's an urgent need for nuance in the discussion. Stanford University health economist Maria Polyakova makes the point that the pandemic's negative health impacts have a complicated inter-relationship with the pandemic's negative economic impacts. Every state, she found, experienced economic damage. "Health damages early on, however, were highly geographically concentrated," she wrote. Maldonado, in the Stanford Daily article, sees human concentration levels affecting campus athletes disproportionately, where they have “more exposure to one another, to other athletes, if they are at practice or training.” Optimal policy is bound to hurt some more than others, but it's a mistake to see pandemic-related sacrifice as a hard yes/no issue, whether that's locking down businesses or canceling sports. There's been impressive scientific progress to counter COVID, so much human immuno-chemistry and basic biochemistry and AI-assisted therapies, but the social science lags.

Tech, and more importantly the future of tech, will be more inclusive and diverse. Facebook AI announced that it is collaborating with U.S. universities that serve significant populations of Black and Latino students to co-teach and fund graduate-level online deep learning courses, a program that piloted with Georgia Tech in Spring 2020. Microsoft has several programs in place to make AI systems more inclusive of people with disabilities. Netflix and 2U are partnering with Norfolk State University to launch online bootcamps meant to increase exposure to the tech industry. National Science Foundation awarded Clemson University a grant that will measure the degree to which students feel empowered to enact change through data science. COVID data collection has highlighted the need for indigenous data control. And the pandemic has also exposed the "homework gap" – yet another indicator of the digital divide in the U.S.

There is growing awareness of AI’s potential for social good and for data science to be interdisciplinary. Progress has meant a step back for every two steps forward. There are grave ethical concerns as AI takes an ever larger decision-making role in an ever-increasing number of industries. There is inherent racism in our funding mechanisms and bias in our rankings, and the acknowledgement that systemic racism shaped the ecosystems of our cities. Maybe it's one step forward and one step back. Regardless, keep at it.

Science has an in-depth article (paywall) on academic job prospects for U.S. science Ph.Ds, reporting that its careers site shows 70% fewer jobs than one year ago. There is also a movement afoot to make "innovation and entrepreneurship" a fourth prong – joining teaching, research and service as crucial metrics for faculty promotion and tenure – but the proposal is not being supported by the American Association of University Professors. MIT and the school's new Schwarzman College of Computing will be supporting 50 new faculty positions. 25 of them will be in the college and 25 will be inter-disciplinary and housed in other departments. Also, the University of Oxford is establishing the endowed DeepMind professorship, and two University of Pennsylvania alumni have pooled their resources to establish a Presidential Professorship in data science that will be housed in the School of Arts & Sciences.

The University of Arizona will use a major gift from the Lundin Family to establish an interdisciplinary school of mining and mineral resources. Missouri Science and Technology University received a $300 million gift from June and Fred Kummer. It will be used to create a new school of innovation and entrepreneurship in addition to other research programs. The Walther Cancer Foundation is making an $11 million investment in the bioinformatics collaboration between Purdue University and Indiana University. The Institute for Complex Adaptive Matter at University of California-Davis (ICAM), received a $1 million grant from the Gordon and Betty Moore Foundation to fund international science exchanges, part of Moore Foundation's $185 million Emergent Phenomena in Quantum Systems program.

Arizona State University is set to offer a new online data science degree starting in Fall 2021. Yale School of Public Health has a new concentration in climate change and health. University of Texas-Arlington now offers a bachelor's degree in business analytics. Santa Fe Institute will use a National Endowment for the Humanities grant to establish a training institute for humanities scholars seeking computational and quantitative skills.

The University of Amsterdam is establishing its research Data Science Centre. Other new data science research centers coming online in the U.S. include: the Partnership to Advance Throughput Computing (PATh) project at University of Wisconsin-Madison, the Center of Excellence in Artificial Intelligence and Machine Learning at Howard University, and the 5G Innovation Hub at University of Illinois Research Park.

Northeastern University named Usama Fayyad to lead the school's Institute of Experimental Artificial Intelligence. Casey Greene will be the Director of a new Center for Health Artificial Intelligence at the University of Colorado School of Medicine. Harvard University is hiring data visualization artist technologists Fernanda Viegas and Martin Wattenberg. Both will hold endowed professorships in the Computer Science Department. Wattenberg joins in Fall 2021 and Viegas will follow in Spring 2022.

Editor's Picks

The sense of humor is making a comeback, and not a moment too soon. Falaah Arif Khan and Zachary Lipton debuted their Superheroes of Machine Learning comic: Volume 1 Machine Learning Yearning. Khan has also collaborated with NYU Center for Data Science researcher Julia Stoyanovich on educational materials to teach ethics and data responsibility. AI-generated jokes is real research and, when discussed expertly, can be insightful and kind of funny. And check out Scientific study on procrastination delayed in The Cavalier Daily student newspaper by University of Virginia freshman satirist Camila Cohen Suárez.

New research using massive data sets is addressing many facets of climate change. A new daily global temperature data set, produced and validated by researchers at the University of Minnesota and University of California-Santa Barbara, could prove valuable in studying human health impacts from heat waves, risks to agriculture, droughts, potential crop failures, and food insecurity, as NOAA reports that September 2020 was the warmest September on record and University of Massachusetts-Amherst researchers warn of the irreversible effect of record warm waters on the environment. Fortunately, we can at least better predict sea level rise with Rutgers' new model. Researchers at Tufts University are using data to gain greater understanding of corals and how to protect them, and University of South Florida is partnering with NOAA to map the ocean, while up on the sea surface, Princeton is building robotic floats to monitor ocean health and a major Arctic ocean research expedition comes to a close. But while we are also reminded that we can't outrun climate change, University of California-Santa Barbara reports that healthcare can actually be a climate solution. And even though we all know climate and weather are two different things, we still need to know if we will need to take an umbrella – making us very glad that National Science Foundation is backing using AI to forecast the weather.

The National Institutes of Health announced a broad, new policy on data management and sharing where the data management plans basically describe how grantees will share their data. The U.S. Food and Drug Administration (FDA) and the White House Office of Management and Budget (OMB) are also introducing new technological workflows, (FDA link, OMB link). The Brookings Institution has free advice on new technological workflows to the U.S. Congress: Invest in Open-Source Software. And, no surprise, the U.S. Department of Defense has fast-track funding for AI technology prototypes.

High-level failures at the U.S. Centers for Disease Control were a subject of long form journalism by excellent science writers, James Bandler and his team at ProPublica (link), and Charles Piller at Science (link).

Overseas, in the probably-should-have-happened-already department, the UK government added statistical and machine learning expertise to the country's COVID-19 response by partnering with The Alan Turing Institute and the Royal Statistical Society. And British Parliament is attempting to draft a first set of rules for the European Union on Artificial Intelligence. But the forward-looking Australian government is close to funding a Facebook-alternative social network for its citizens.

There is a national election underway and it's our pleasure to give you something else to read about. We will mention however that the State of California passed Proposition 24, the California Privacy Rights Act (CRPA). CRPA augments and tightens loopholes in the California Consumer Privacy Act, which is one of the strongest consumer data control and privacy laws in the U.S.

Research News

An enterprise software developer in Ireland, NearForm, has become the go-to resource for contract tracing apps. Early on there was no norm for whether contract tracing apps would be mandatory or voluntary – a huge problem when it comes to creating the requirements list for a user interface designer. NearForm's COVID Tracker product reached 35% voluntary adoption rates in Ireland, and is good evidence that they have something usable, and four U.S. states (Pennsylvania, Delaware, New Jersey, New York) agree. NearForm is working with European Union authorities on much needed improvement to the app's cross-border interoperability. Contract tracing apps work, as the medical journal The Lancet points out, but the biggest successes have come in South Korea, China and Singapore, where adoption was mandatory. A new paper by Eszter Hargittai, Elissa M. Redmiles, Jessica Vitak and Michael Zimmer found that experienced Internet users were more willing to adopt contract tracing apps (except when the app was not insurance-provided), suggesting that these users had sufficient know how to address any privacy concerns on their own.

The pandemic has enabled governments to curtail internet freedom. Digital rights watchdog organization Freedom House reports that contract tracing has contributed to government surveillance in Russia and India. Universities have been tracking students' data on remote learning platforms in the UK. Reports of student surveillance at U.S. universities show that it is present and not necessarily pervasive, but there is a growing outcry, first at University of Illinois and more recently at Miami University in Ohio, against Proctorio, an online test proctoring product.

Usable, data-backed software is crucial to getting the pandemic under control. The team at Johns Hopkins University has done superior work with its global coronavirus counting dashboard. The threat of a COVID plus flu "twindemic" has prompted physicians and coders to collaborate on new symptom checking tools that help to differentiate between the illnesses.

Trust, crucial to usability, is shaping up as the weak link for data applications in medicine. Calls for explainability, transparency and reproducibility are gaining urgency, and while valuable, they aren't necessary for trust. Look at how widely recommendation systems are used for digital content and commerce. Marketing professors Chiara Longoni (Boston University) and Luca Cian (University of Virginia) call it "word-of-machine effect" when people prefer AI recommenders to human recommenders. The University of Pennsylvania Health System, Penn Nudge Unit used machine learning predictions for short-term mortality to identify high risk patients and then to initiate conversations with them around end-of-life goals before it's too late. Here, a machine recommendation and a behavioral nudge significantly increased quality of care. There's a long and growing list of AI-assisted clinical diagnostics (Alzheimer's disease at University of Southern California, Cardiac CTs at Rensselaer Polytechnic Institute, Osteoarthritis at NIH National Institute on Aging) which, hopefully, can also incorporate their own behavior-aware innovations.

Data Visualization of the Week

arXiv, IEEE VIS 2020 Best Paper Award; Alex Kale, Matthew Kay, Jessica Hullman from October 30, 2020

Visual Reasoning Strategies for Effect Size Judgments and Decisions

Alex Kale, Matthew Kay, and Jessica Hullman

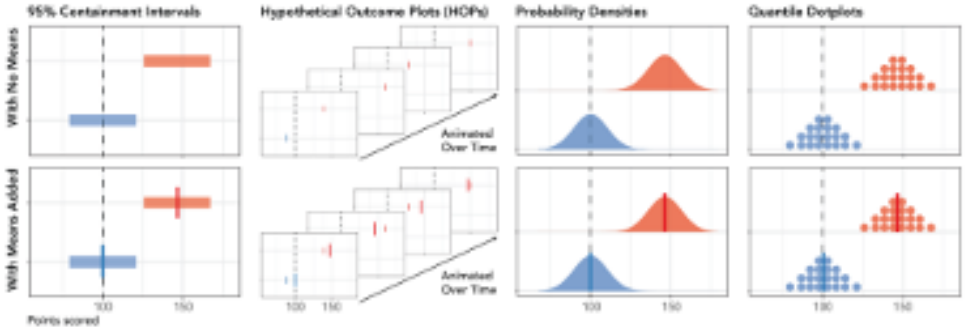

Fig. 1: Visualization designs evaluated in our experiment.

Abstract— Uncertainty visualizations often emphasize point estimates to support magnitude estimates or decisions through visual comparison. However, when design choices emphasize means, users may overlook uncertainty information and misinterpret visual distance as a proxy for effect size. We present findings from a mixed design experiment on Mechanical Turk which tests eight uncertainty visualization designs: 95% containment intervals, hypothetical outcome plots, densities, and quantile dotplots, each with and without means added. We find that adding means to uncertainty visualizations has small biasing effects on both magnitude estimation and decision-making, consistent with discounting uncertainty. We also see that visualization designs that support the least biased effect size estimation do not support the best decision-making, suggesting that a chart user's sense of effect size may not necessarily be identical when they use the same information for different tasks. In a qualitative analysis of users' strategy descriptions, we find that many users switch strategies and do not employ an optimal strategy when one exists. Uncertainty visualizations which are optimally designed in theory may not be the most effective in practice because of the ways that users satisfice with heuristics, suggesting opportunities to better understand visualization effectiveness by modeling sets of potential strategies.

Index Terms—Uncertainty visualization, graphical perception, data cognition

Tweet of the Week

Twitter, Ian Bogost from October 19, 2020

Events

University of Michigan Data Science Annual Symposium 2020
Online November 10-11. [registration required]

TMLS Annual Conference & Expo 2020
Online November 16-19. "The Toronto Machine Learning Summit (TMLS) is a community with over 9,000 active members that works to promote and encourage the adoption of successful machine learning initiatives within Canada and abroad." [$$]

COMPLEX NETWORKS 2020
Online December 1-3. The conference "aims at bringing together researchers from different scientific communities working on areas related to complex networks." [$$$]

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.

Introducing Outlier — A Conference Hosted by the Data Visualization Society
Online February 4-5, 2021. "I'm happy to announce that we will be hosting our first conference! Outlier will be hosted virtually on February 4th and 5th. The events committee and I have been doing a lot of thinking around what our ideal virtual conference would look like and we're super excited to bring this vision to fruition!" [save the date]

Save the Date & Celebrate Int'l Women's Day with WiDS!
Online March 8, 2021. "The Women in Data Science (WiDS) Worldwide conference is a technical conference featuring outstanding women in data science and related fields such as artificial intelligence across a wide range of domains. Join us as we follow the sun around the world to bring you thought leaders from academia, industry, non-profits, and government." [save the date]

A global collaboration to move artificial intelligence principles to practice

"On May 6–7, 2021, MIT will host — most likely online — the first AI Policy Forum Summit, a two-day collaborative gathering to discuss the progress of the task forces towards equipping high-level decision-makers with a deeper understanding of the tools at their disposal — and trade-offs to be made — to produce better public policy around AI, and better AI systems with concern for public policy. Then, in fall 2021, a follow-on event at MIT will bring together leaders from across sectors and countries and, built atop the leading research from the task forces, the forum will provide a focal point for work to move from AI principles to AI practice, and serve as a springboard to global efforts to design the future of AI." [save the date]

Tools & Resources

AI Engineers Need to Think Beyond Engineering

Harvard Business Review, Donald Martin, Jr. and Andrew Moore from October 28, 2020

"Computer scientists need to do more to understand and account for the underlying societal contexts in which these technologies are developed and deployed."

"Here at Google, we started to lay the foundations for what this approach might look like. In a recent paper co-written by DeepMind, Google AI, and our Trust & Safety team, we argue that considering these societal contexts requires embracing the fact that they are dynamic, complex, non-linear, adaptive systems governed by hard-to-see feedback mechanisms. We all participate in these systems, but no individual person or algorithm can see them in their entirety."

A Taxonomy of Training Data: Disentangling the Mismatched Rights, Remedies, and Rationales for Restricting Machine Learning

Info Justice, Benjamin Sobel from October 26, 2020

"The chapter taxonomizes different applications of machine learning according to the qualities of their training data. Four categories emerge: (1) public-domain training data, (2) licensed training data, (3) market-encroaching uses of copyrighted training data, and (4) non-market-encroaching uses of copyrighted training data."

Behavioral Testing of NLP models with CheckList

Amit Chaudhary from October 07, 2020

"In this post, I will explain the overall concept of CheckList and the various components that it proposes for evaluating NLP models."

The best Low-Code Machine Learning Libraries in Python

Medium, Spatial Data Science, Abdishakur from October 30, 2020

"Low-code/No-code platforms and libraries enable users to run machine learning models easily by providing a ready-to-use code and functions. You can access these functions either through a web interface or writing minimal code."

"While no-code platforms are the easiest way to train a Machine Learning model through drag and drop interface, they lack flexibility."

"On the other hand, the low-code ML is the sweet spot and middle ground. They offer both flexibility and easy to use code. You still have to write some code, but that is bare minimum compared to other typical machine learning libraries."