Data Science Community Newsletter

Data Science Community Newsletter features journalism, research papers and tools/software for October 9, 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

On the one hand, the majority of college students who responded to a new poll believe colleges could have handled their return to campus much better, and in a myriad of ways. On the other, these same challenges are inspiring innovation in the form of testing wastewater at the University of Arizona, creating wearables to stop the COVID spread at the University of Florida, a COVID symptom-monitoring facemask at the University of Rhode Island, launching a new app at UC Berkeley, and attempts to keep students from infecting locals at UT Austin, Texas Tech and the Texas A&M campuses. Emory and Georgia Tech together lead an National Institutes of Health program to verify COVID-19 diagnostic tests. And the fact that the pandemic has finally pushed universities into fully embracing online learning – not just talking about it – is seen as a long-overdue plus.

National Science Foundation put out an impressive group of programs that intend to diversify data science. Rochester Institute of Technology announces a data science course for non-computing majors. Florida Atlantic University will recruit talented undergraduates with low-income backgrounds and then financially support the ones who enter FAU's new program as juniors. The support goes for three years, leading to Bachelor's and Master's degrees in AI, autonomous systems or machine learning. And Yale will create a brand new program to train "Research Computing Facilitators" (RCFs), staff trained to match university research projects to advanced computing resources at small- and medium-size universities, schools like Southern Connecticut State University, which is 3 miles east of Yale but light-years away in terms of the technical capability to apply data science.

New Mexico State University announced that it will improve community-wide Internet access to high performance computing resources, initially partnering with New Mexico Highland University. And the state of South Dakota is on its way to being able to offer its first doctorate in computer science, a joint program involving South Dakota State University at Brookings and Dakota State University at Madison.

The Alfond Foundation is set to give a whopping $500 million in grants to higher education and research institutes in the state of Maine. Major beneficiaries are the Roux Institute, a Northeastern University outpost in Portland that focuses on graduate education and research in AI and computational biology, and the University of Maine System, which got $75 million to establish its College of Engineering, Computing and Information Science.

The Temerty Foundation granted C$250 million to the University of Toronto Faculty of Medicine. Part of the gift will be used to establish a new Centre for AI Research and Education in Medicine.

National Science Foundation (NSF) has awarded Duke University engineers a $3M AI grant for training the next generation in the new convergent field of materials and computer science research. Duke engineers will also be collaborating with the University of Pittsburgh and UPMC on a $1 million grant to preserve patient data privacy within big data healthcare research.

The Geospatial Institute at St. Louis University (GeoSLU) received a $5 million grant from U.S. National Geospatial-Intelligence Agency (NGA) to train intelligence officers. NGA is currently replacing its Saint Louis hub with a new campus, and has long collaborated with the University.

Jeannette Wing, writing in the Harvard Data Science Review, authored her "Ten Research Challenge Areas in Data Science." It's a good list, prefaced by three meta-questions related to the larger question of "Is Data Science a Discipline?" There's no good way to preview the pre-questions or the list, so let's skip

to her conclusion, "What will data science be in 10 or 50 years? The answer to this question is in the hands of the next-generation researchers and educators. To advance and study data science will take a commitment to learn the vocabulary, methods, and tools from multiple, traditionally siloed disciplines."

Big questions drive the field forward but universities have their own ways of organizing the people who do data science. Dave Hunter, a Penn State statistics professor, has spearheaded the school's informal, grassroots community-building around inter-disciplinary data science. At the University of Chicago, the Computer Science department chair Mike Franklin and the statistics deparment chair Dan Nicolae have designed and proposed a new undergraduate data science major. (UC established a program leading to a data science minor in 2019.) These informal and formal structures are subject to the invisible hand of economics. According to Wired, University of Toronto grad student Mohamed Abdalla found "more than half of tenure-track AI faculty at four prominent universities who disclose their funding sources have received some sort of backing from Big Tech."

Results from a survey by University of Michigan Institute for Research on Innovation & Science (IRIS) estimates that active university research capacity has dropped by 30 percent as a result of the pandemic, an estimate derived from research vendor spending at 10 research universities.

Researchers led by a team at Oregon State University will establish a Center of Excellence in Genomic Science, to be housed at OSU's new Translational and Integrative Sciences Center in the College of Agricultural Science. NIH has granted $10 million for the Genomic Science center that also includes scientists from Lawrence Berkeley National Laboratory, Johns Hopkins University, Jackson Laboratory, Queen Mary University of London and the European Bioinformatics Institute.

The National Academies of Sciences, Engineering, and Medicine recently announced a new committee being formed to advance a systems approach to studying the Earth. In keeping with this approach, Colorado State University announced EarthWorks, a new partnership with the National Center for Atmospheric Research (NCAR) to create a high-resolution version of the Community Earth System Model (CESM), which is an open-source model used by many researchers to improve understanding of the complex interplay of atmospheric, oceanic, land surface and sea ice processes that comprise the Earth system. As NSF rolls out its Convergence Accelerator programs, interdisciplinary projects are emerging such as the one led by the University of Arizona and Princeton to model the nation's groundwater.

Canada's Alberta Machine Intelligence Institute (Amii) will get C$9 million from regional government to apply machine learning to greenhouse gas emissions. Alas, critics point out that the money is one-third of what was originally promised, only to be cut from budget plans.

Editor's Picks

Orwell didn't predict this. Big Brother government technologies haven't scaled well. The Los Angeles Police Department has used facial recognition software 30,000 times in the last decade, a large and unpleasant number, but not quite big data. Whereas UK police officers have been told not to download and use the National Health System (NHS) contract tracing app, keeping them out of the fray of an app that has been downloaded 12 million times. From immigration to elections to public health, government big data solutions are in the works but not really working.

There's reason for optimism however. Beta testing for instance, like in Amsterdam (Netherlands), where city government is developing an Algorithm Register to allow the public to get acquainted with the city's algorithmic systems and give feedback. The state of California published a "data strategy" document to guide analysts and developers through the state's myriad data sets and streams. U.S. local governments are starting to put the Apple-Google platform for COVID contact tracing to work after a lengthy gestation period. (See the New York/New Jersey example, and the North Carolina example.) And New York University researcher Julia Lane argues in a just-published book for rethinking how the U.S. measures its society. Collaborative solutions that are coherent and not rushed, and which build on solid technical principles, have the best chance to work.

While it's plenty obvious that women are equal to men in computing skills, scratch the surface and you will find women feeling less confident in those skills, according to researchers from Villanova.

Medical AI tools sometimes reflect a geographic bias, calling for larger and more diverse datasets, say researchers at the Stanford Institute for Human-Centered AI. Algorithmic bias is more than geographic. It can be pervasive. Both Twitter and Zoom have been called on the carpet recently – in both cases for cropping out all or part of Black users faces. And sometimes it's institutional. An NYU Public Safety Lab report reveals a pattern of race-based NYPD misconduct.

Casey Fiesler and Natalie Garrett in Wired are calling for an analysis of what they call “ethical debt” – fixing issues related to bias and inequality BEFORE the newest tool gains widespread use, instead of after, as with “zoombombing” and the aforementioned cropping problem. Love a smart debt metaphor. See also, Research Debt, and the godfather, Technical Debt.

Research News

According to a new paper published in the New England Journal of Medicine there have been crucial data collection and management problems in the testing criteria (too few tests taken), in the test itself (too expensive, too slow for test results) and in the population-wide evaluation of results (too imprecise for population-level disease surveillance). The paper authors feel we need a new, different type of testing regime. COVID-19 testing should be inexpensive, fast, easy and, as a result, frequent and universal. There's an emerging consensus that new rapid tests can be a paradigm shift in COVID testing.

Independent of testing regimes, poor data handling has plagued efforts to test broadly. Some of the most significant problems:
- government bureaucracy
- academic bureaucracy
- algorithmic bias
- systemic undercounting
- and in England, Excel spreadsheets

Producing models out of all the COVID-related data is a major challenge. And social media echo chambers have formed around "COVID-19 elites" as groups of individuals look for information about the pandemic.

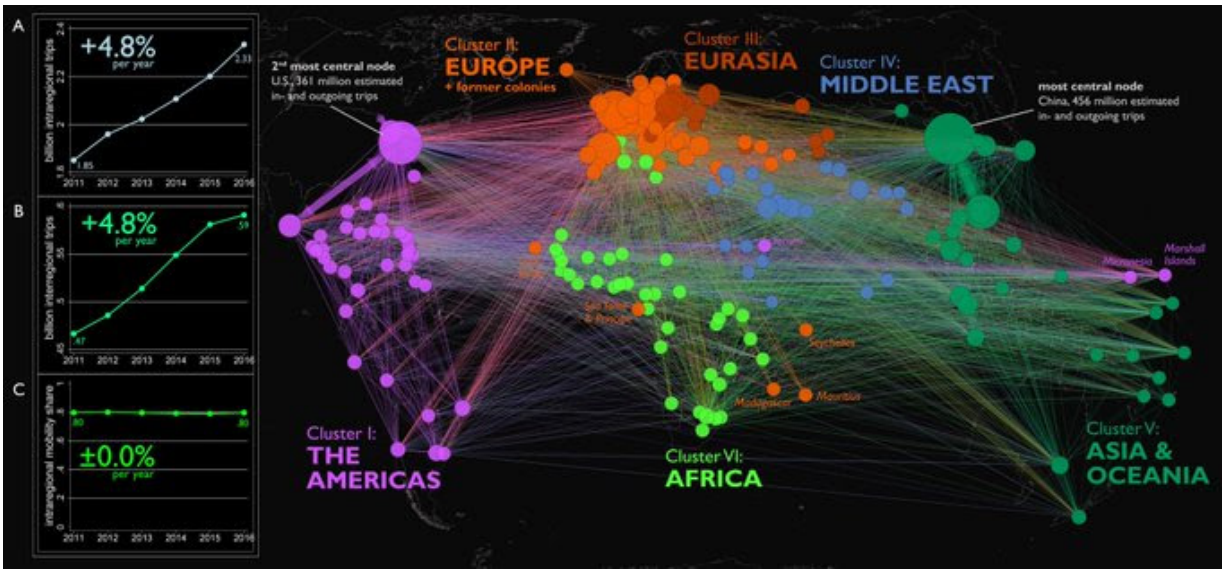
Los Angeles has reduced inequity in the city's COVID testing and treatment, crediting an extensive data dashboard for the success. Allison Crawford, a clinical psychiatrist and University of Toronto professor, goes further. She authored and recommends her Digital Health Equity Framework to identify the factors that lead to inequity in virtual health services that have become crucial elements of COVID prevention and care. Tools for applying the Framework are in development.

"The brain-computer interface is coming and we are so not ready for it" is the title of a recent longform article in Bulletin of the Atomic Scientists. It's an excellent status report on what's been a bellwether technology for applied data science. Other medical technology domains have been subjects for similar status reports and for important technical advances. Wearable technology, ambient intelligence and mHealth apps are revolutionizing in-home and in-hospital care, as well as cardiology. Emerging zoonotic disease prediction is a subject of intense investigation at Georgetown University, and a major investment by the U.S. National Institute for Allergy and Infectious Diseases. Regenerative medicine is advancing with machine learning recipes for bioscaffolds at Rice University, and with pattern matching among transplanted engineered tissues at Carnegie Mellon. Are we so not ready? It may not matter.

Data Visualization of the Week

Twitter, Emanuel Deutschmann from September 28, 2020

"80 percent of all human movements between countries occur within world regions."



Tweet of the Week
Twitter, Bryn Stole from September 30, 2020

Bryn Stole
@brynstole

A bias of ~The Media~ that never gets talked about is the strong bias toward coherence.

Make a reporter sit through a bizarre, rambling, nonsensical speech and they'll be inclined to summarize it in some way that makes actual sense and grab the single sensical quote.

12:09 PM · Sep 30, 2020 · Twitter for iPhone

1.9K Retweets **215** Quote Tweets **8.1K** Likes

Events

'Information and Uncertainty in Data Science' Discussion Forum
Online October 9, starting at 4 p.m. Pacific time. Speaker: Gerald Friedland, Adjunct Assistant Professor, EECS, UC Berkeley. [registration required]

JupyterCon 2020 General Sessions (Times in UTC) - In Progress Schedule
Online October 12-16. "We developed a vision for JupyterCon Online as a learning platform, unconstrained by synchronous schedules or geographical location, coalescing a multitude of mini-events

and rad new content, learning experiences, and online social interactions." [$$$]

Purdue University's Big Data, Safe Food conference goes forward virtually

Online October 12-15. “The main question the conference will focus on is how can as academia, industry and government entities collaborate to address and prevent foodborne diseases using data collected from farm to fork. Food has a long and complex supply chain, and we need to discuss how big data can influence food safety, along with who manages and shares these data points.” [registration required]

United Nations World Data Forum

Online October 19-21. "Brings together data and statistical experts and users from governments, civil society, the private sector, donor and philanthropic bodies, international and regional agencies, the geospatial community, the media, academia and professional bodies. 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]

Bias^2 Seminar: Ziad Obermeyer, UC Berkeley

Online October 22, starting at 1:30 p.m. Eastern time. "Title: Algorithms that reinforce racial biases, and algorithms that fight them" [registration required]

Vis In Practice Workshop

Online October 26, starting at 10 a.m. Mountain time. "VisInPractice provides an opportunity for practitioners and researchers to share experiences, insights, and ideas in applying visualization and visual analytics to real use cases." [save the date]

Join us for Data for Black Lives III

Cambridge, MA December 11-13 at MIT Media Lab. "Returning to Cambridge will make it possible for us to accomplish all of our goals: expand our programs, continue to build our network, and host the best conference yet." [save the date]

Tools & Resources

graph-tool: Efficent network analysis with python

Tiago P. Peixoto from September 17, 2020

"Graph-tool is an efficient Python module for manipulation and statistical analysis of graphs (a.k.a. networks). Contrary to most other python modules with similar functionality, the core data structures and algorithms are implemented in C++, making extensive use of template metaprogramming, based heavily on the Boost Graph Library."

How randomized response can help collect sensitive information responsibly

Google, PAIR Explorables, Adam Pierce and Ellen Jiang from September 17, 2020

"Giant datasets are revealing new patterns in cancer, income inequality and other important areas. However, the widespread availability of fast computers that can cross reference public data is making it harder to collect private information without inadvertently violating people's privacy. Modern randomization techniques can help preserve anonymity."

Data Cleaning IS Analysis, Not Grunt Work

Counting Stuff newsletter, Randy Au from September 15, 2020

"The act of cleaning data imposes values/judgments/interpretations upon data intended to allow downstream analysis algorithms to function and give results. That’s exactly the same as doing data analysis. In fact, “cleaning” is just a spectrum of reusable data transformations on the path towards doing a full data analysis."

On the #COVID19 transparency front: JNJ has announced that they will share the Clinical Study Report and clinical trial participant data from their vaccine trial with researchers through @Yale Open Data Access (YODA) Project.

Twitter, Harlan Krumholz from September 24, 2020

"To provide context for the JNJ vaccine data sharing, @Yale YODA Project provides access to data to researchers w/ credible scientific proposal, agreement to report results, & commitment to protect participant privacy. Access decision is independent. https://yoda.yale.edu"

In my first-ever tweet, I am thrilled to share NL4DV, a new toolkit that helps people prototype natural language (NL) interfaces for data visualization.

Twitter, Arpit Narechania from October 01, 2020

"Given a dataset and an NL query, NL4DV recommends a list of Vega-Lite specifications corresponding to the query. This allows specifying visualizations using NL in environments like @ProjectJupyter and developing applications such as an NL-based @vega_vis generator/editor"

At Princeton CITP, we were concerned by media reports that political candidates use psychological tricks in their emails to get supporters to donate. So we collected 250,000 emails from 3,000 senders from the

2020 U.S. election cycle.
Twitter, Arvind Narayanan from October 05, 2020
"Here's what we found. https://electionemails2020.org"

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 205