Data Science Community Newsletter features journalism, research papers and tools/software for June 1, 2021.

Please let us (Micaela Parker, Steve Van Tuyl, Catherine Cramer, 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

"There's been a rapidly growing interest in computational social science," said **Duncan Watts** at a panel organized by **University of Massachusetts**. He discussed three challenges to the field: research design, data collection and rethinking the model of social science to a "solution-oriented science." Researchers at **Duke University** are looking at how the Duke community coped with the pandemic and Northeastern University researchers are looking at factors for achieving collective intelligence in human groups – or not.

Fortunately collective intelligence is overflowing on social media. And now that sarcastic tone can be detected thanks to a new sentiment analysis tool developed by researchers at the **Complex Adaptive Systems Laboratory at the University of Central Florida** that uses deep learning to analyze social media text for sarcasm. What if you are being sarcastic with your Significant Other? Wearables are being developed that assess physiology, vocal characteristics and language to predict an upcoming conflict about 80 percent of the time. **Adela Timmons**, psychology professor at **Florida International University**, says "If we [psychologists] could be in the context in which the problem is occurring, in the optimal moment… we could potentially do more and be more effective."

Too much positivity in online reviews is the focus of research at **University of Massachusetts Boston**, where researchers have developed a computational linguistics tool called the Evaluative Lexicon, which measures the emotion in people's text, found to be a better predictor of the success of what's being rated than the rating itself. "The basic idea is that mass-scale emotion is predictive of future mass-scale behavior," says researcher **Matthew Rocklage**. Probably a more serious concern (said without sarcasm) is the research from **University of California-San Diego** showing that "published papers in top psychology, economics, and general interest journals that fail to replicate are cited more than those that replicate."

Sheldon Jacobson, a computer and data science professor at **University of Illinois**, has a recent opinion piece in the **Indianapolis Star** (in addition to his regular essays for **The Hill**). Public data literacy

(or lack of it) has limited effective public health policy at the end of our pandemic, and according to Jacobson, data scientists should work on cultivating wider appreciation for evidence-based decisions by using more context, targeting narrower audiences and leaving out the brain-melting torrents of rapid-fire numbers, facts and figures. Still, shifting the public demand curve towards more and better science reporting will not necessarily decrease the supply of weak, wrong or bad science that flows from traditional science publishing into mass media. **Jevin West** and **Carl Bergstrom**, instructors of the popular "Calling BS" course at the **University of Washington**, told **Vice** magazine that the "same attention economy that drives likes on tweets" also drives which research gets published in the first place, often at the cost of nuance in everyone's understanding of science.

Data and computing in K-12 education is a priority in many places. The **State of New Jersey** gave funds to **Rutgers University**, **Fairleigh Dickinson University** and **Kean University** to create computer science learning hubs to train K-12 educators. The **Mississippi state legislature** is considering bills to institute a new high school computer science curriculum. **Michigan Department of Education** has its C2 Pipeline program for Detroit high school students, and will be running summer workshops through **Wayne State University** with technology help from **Amesite**, makers of an online learning platform. The national organization **AI4K12**, led by **Christine Gardner-McCune** (**University of Florida**), **David Touretzky** (**Carnegie Mellon University**) and **Deborah Seehorn** (**Computer Science Teachers Association**), is a primary resource advancing these state initiatives. The group keeps an up-to-date national list of states' K-12 initiatives. **Code.org**, a Seattle non-profit, is creating a high school course similar to the AP Computer Science A course, but with an inclusive focus. The course is set to pilot in **Georgia**, **Oklahoma**, **New Mexico** and **Ohio**.

Still more momentum for youth AI skill-building comes from **NVIDIA**, which recently undertook a three-year education partnership with the **Boys & Girls Clubs of Western Pennsylvania** that builds on a prior **Pennsylvania** state grant to advance STEM education. **Cynthia Brazeal** and **Hal Abelson** from **MIT Media Lab** have also established a societal AI-literacy initiative called RAISE (Responsible AI for Social Empowerment and Education) with a significant preK-12 education component. The **U.S. Naval Surface Warfare Center, Crane Division** (NSWC Crane) and **Indiana University** are collaborating on a $1.7 million pilot program to enhance artificial intelligence education programming for rural students. The program for Indiana middle school students kicks off this summer.

Rajendra Raj from **Rochester Institute of Technology** and **Amruth Kumar** from **Ramapo College** will be leading a joint task force to revise undergraduate computer science curricula. Participants come from the **IEEE Computer Society**, the **Association for Computing Machinery** (ACM) and, for the first time, the **Association for the Advancement of Artificial Intelligence** (AAAI). The 19 member task force represents universities located in the U.S., Brazil, Canada, Egypt, India and Ireland. AI skills and improving diversity are focus areas. Expect the new guidelines in 2024.

The **University of California-Davis Bourns College of Engineering** will offer the first Master of Science degree in robotics available through the **University of California System**. **Amit Roy Chowdhury**, professor of electrical and computer engineering, is program chair.

The **School of Civil and Environmental Engineering at Cornell** now has a "smart cities concentration" available to undergraduate students beginning in the fall 2021 semester. CEE Director **Linda Nozick** expects the five-course concentration to eventually become a minor degree program.

Emory University will establish the Brain Health Personalized Medicine Institute, headed by neurologist **Allan Levey**. The institute will integrate data from **Emory Healthcare** into Emory's clinical and research neuroscience programs.

University of Virginia will offer a professional three-course certificate in Data Science for Business Strategy through **UVA Darden Executive Education & Lifelong Learning**. The courses are offered both on-line and in-person.

Oral Roberts University in Tulsa, OK, has new masters degrees in Computer Science and in Data Science. The **University of North Carolina-Greensboro** is starting a doctoral program in computer science.

The Board of Trustees at **Ferris State University** in Michigan approved a new bachelors degree in Artificial Intelligence. The Board also approved construction plans for a $29.5 million building, the Center for Virtual Learning, that will house the school's new Artificial Intelligence program.

The **Machine Learning Center at Georgia Tech** graduated its first three Ph.D. students: **Cyrille Combettes**, **Harsh Shrivistava** and **Haomin Jiang**. **Georgia Tech** also announced a masters degree in Urban Analytics administered by the **Georgia Tech's School of City and Regional Planning** and leveraging the institute's programs in Industrial Engineering and Computing.

The **University of North Carolina-Chapel Hill** will have an undergraduate data science minor launching fall 2021. The school's pan-University data science initiative is "soon-to-launch."

For the past eight years the **University of Pittsburgh** has themed its academic class year around an important topic of universal interest across the Pitt community, a "Year of"-something. 2021-22 will be the Year of Data and Society. **Eleanor Mattern**, a teaching assistant professor in **Pitt's School of Computing and Information**, is campaign chair.

Johns Hopkins University launched the Pandemic Data Initiative to follow through on data and communication deficiencies that **Hopkins' Coronavirus Resource Center** (CRC) researchers uncovered as the COVID-19 outbreak blanketed the U.S.

Richard King Mellon Foundation approved a $150 million grant to **Carnegie Mellon University**. $75 million makes it possible for CMU to go forward with a planned $210 million state-of-the-art science building on Forbes Avenue, adjacent to the campus' **Carnegie Museums**. The second $75 million is seed funding for a new robotics research center and an institute for advanced materials that will be co-located at Hazelwood Green, a former steel mill site near campus.

University of Texas-San Antonio received a $1 million gift from Frost Bank that will create an endowment to support graduate fellowships and undergraduate research activities at the **UTSA School of Data Science**. UTSA recently named **David Mongeau** to be the school's founding director. Mongeau had been the executive director of **University of California-Berkeley's Institute for Data Science**.

A five-year agreement between the **University of Maryland**, **University of Maryland-Baltimore County** (UMBC) and the **U.S. Army Research Laboratory** (ARL) will focus on robotics and artificial intelligence, and could bring up to $68 million in research funding. **Derek Paley**, director of the **Maryland Robotics Center**, is the lead researcher.

Howard University and **Amazon Web Services** have a joint initiative to increase the cloud services training and resources available to students and educators. Aided by AWS experts, "Howard University is creating a new master's degree program focused on data science and will integrate cloud concepts into other areas of study, including business, computer science, and engineering."

Creative Destruction Lab (CDL) is a nonprofit tech accelerator based at **University of Toronto's Rotman School of Management** and it will launch its third U.S. outpost, **CDL-Seattle**, at the **University of Washington Foster School of Business** in partnership with **Microsoft**, the **UW College of Engineering**, the **Allen School of Computer Science & Engineering**, and **UW CoMotion**. The initial focus for CDL-Seattle is computational health. **University of California-Berkeley** also has a new nonprofit startup incubator, the **Bakar BioEnginuity Hub**, located in what had been the school's art museum. There will be space for as many as 80 startups. Climate change and human health are its initial focus, according to general manager **Gino Segre**.

BlackBerry has expanded its existing joint innovation program with the **University of Waterloo**. The program is part of Waterloo's newly established **Gateway for Enterprises to Discover Innovation**, a group tasked with streamlining corporate engagement. BlackBerry also has an extensive ongoing collaboration with **Carleton University** in Ottawa.

Editor's Picks

As Nobel prizewinner **Daniel Kahneman** recently told **The Guardian**, "When linear people are faced with exponential change, they're not going to be able to adapt to that very easily." So what happens as we adapt to AI's persistence in our lives and work? **NIST** has published a draft document Artificial Intelligence and User Trust which

looks at whether human trust in AI is measurable, proposing nine factors that contribute to a person's potential trust in an AI system. "It's how the user thinks and feels about the system and perceives the risks involved in using it," says co-author **Brian Stanton**. (The document is open for comment until July 30.)

Cold Spring Harbor Laboratory researchers are looking at a weird artifact of AI machine learning tools: that the researchers who design these tools do not know what rules the machines make for themselves. A new method, Global Importance Analysis, tests what rules machines generate to make predictions about patterns in RNA strands – making them "show their work." The method is now available for anyone to use. **Stanford University** researchers use biomedical examples to illustrate how algorithmic bias and disparity can arise due to inadequate outcome choice, data collection, and model evaluation, calling for ensuring that biomedical AI benefits diverse populations.

The one-year anniversary of the murder of **George Floyd** is also the one-year anniversary of several tech company promises to fight racism through changes to their facial recognition policies. **IBM** announced it was getting out of the facial recognition business entirely, and **Amazon** has just announced it will stop selling face recognition technology to police — indefinitely, which the **ACLU** calls "a major win for racial justice, curbing police powers, and privacy." (Amazon, however, continues to sell its Ring home video doorbell, which provides video footage to over 1800 police departments across the country, and in an OpEd in **The Guardian** is called "the largest civilian surveillance network the U.S. has ever seen.")

Castro Community Benefit District (in San Francisco's Castro District) is considering putting in "public safety cameras" in an effort to reduce crime. Executive Director **Andrea Aiello** says, "[The video] gets uploaded into a portal and then we give law enforcement access to the portal and they can use the video." But as **Lee Hepner** with the **Harvey Milk LGBTQ Democratic Club** points out, "If surveillance cameras prevented crime in San Francisco or in the Castro, then the hundreds of individually-owned cameras that are already in the Castro would be doing that job. The only thing this does is provide a centralized surveillance network that we know will be used to target immigrants, poor people, Black and brown people, queer people in a beacon of free expression which is the Castro neighborhood."

Airports are implementing live tracking like systems sold by **LiveReach Media** (LRM), a "comprehensive motion analytics and digital out-of-home marketing platform" and a majority of school-based apps are sending student data to 3rd-party companies – companies like **Facebook** and **Google** – without the schools' or parents' consent. We might be making progress on facial recognition but the rise of surveillance in general seems relentless. Most disturbing of all might be what **Senator Ron Wyden** calls "warrantless surveillance of Americans" by the **U.S. Department of Defense**, referring to the purchase of data related to internet browsing history and location data. Wyden is currently proposing a new piece of legislation called *The Fourth Amendment Is Not For Sale Act* which would force some agencies to obtain a warrant for location and other data and would ban the use of the much-hated **Clearview AI**. Both of the main avenues of law enforcement surveillance – public facial recognition technology and the sale of location data – need to be controlled, tightly.

A review study led by **Stanford Medical School** epidemiologists **Steven Goodman** and **Tahima Nasserie** showed that hospitalized Covid patients exhibited an enormous range of symptoms, 84 different symptoms by their count. Over seventy percent of patients in the study had at least one symptom that persisted sixty days or longer. Long Covid is shaping up to become an immense public health burden.

Mini-organs invented by stem cell biologists like **Shuibing Chen** at **Weill Cornell Medicine** are a major technical advance helping researchers to understand Covid. The extensive new, global collaboration networks are a major process advance and will have lasting impact, according to the genome scientist and entrepreneur **William Haseltine**. A perspective article by **Enrique Valera**, **Rashid Bashir** and colleagues from **University of Illinois** says that new point of care (POC) diagnostics will enable rapid population-scale for infectious disease testing that can function at industrial-scale, personal-scale and any regional- or community-scale in between. A look-ahead investigation into socio-cultural drivers of Covid infection rates by anthropologist **R. Alexander Bentley** (**University of Tennessee**) found that nations with greater openness and tolerance, and nations with low trust in institutions tended to experience the more severe outbreaks. Reporter **Helen Branswell** of **STAT** also wrote a reminder that pandemics do eventually subside naturally, as the co-evolution of infectious organisms and human immune systems modulate, but don't necessarily quell, the severity of the disease that caused outbreaks.

Looking back at Covid, the **University of Pennsylvania** gave the **New England Journal of Medicine** a thorough write-up of the school's Covid surveillance processess and systems. **University of Cambridge** applied mathematician **Michael Roberts** also looked back, measured the contribution of artificial intelligence to Covid diagnosis and found it disappointing. **University of Wisconsin** mathematician **Jordan Ellenberg** calls out health data analysts who opted to fit curves as their main Covid prevalence predictor, without regard for disease biology, cultural norms or government rulemaking. "Real modeling is always a dance between predictable dynamics and our unpredictable responses," Ellenberg wrote in **Slate**.

Research News

Apple announced updates to a variety of its assistive tools. The company used ML and computer vision to produce VoiceOver, a screenreader which makes apps usable for those with low or no vision. Apple plans to work directly with developers to make apps accessible from the start; recently the company added image description to the service along with Markup, allowing users to add their own descriptions to images. Other needs are addressed with AssistiveTouch when using Apple Watch, eye-tracking when using iPads, and additional iPhone support for hearing aids. Its introduction of new iPhone background sounds to help us “stay calm” will come in handy when faced with sticker shock at Apple prices.

Twitter went public with its work on improving its photo cropping algorithm using a saliency algorithm and testing it for gender and race-based bias. One of their conclusions is “not everything on Twitter is a good candidate for an algorithm, and in this case, how to crop an image is a decision best made by people.” Their “Responsible ML” initiative seems to indicate that not using ML is the most responsible choice. Tech companies in general can now opt-in to the Race Equality Index, which is an attempt to evaluate companies’ Diversity, Equity and Inclusion (DEI) reports by comparing them to each other, with a goal of "creating a definitive benchmark for equity in tech."

Stanford Health Care started using **Qualtrics** for their post-care patient surveys to make sure they get "representative feedback from the diverse community" that Stanford treats. **Stanford Medicine** established a Commission on Justice and Equity and its Health Equity Committee is gathering

“information on race/ethnicity, sexual orientation/gender identity and socioeconomic status, to understand where disparities exist so it can ensure equitable access to care for all” and the Stanford School of Medicine is undergoing a Curriculum Review. All of these developments are in reaction to the **American Medical Association** declaring racism to be a public health threat. Understanding financial inequality on a global scale has proved challenging, as data is most lacking from the parts of the world experiencing the highest rates of poverty. New research using remotely sensed nighttime light data as a proxy for inequality may help fill this gap. While researchers found “a significant relationship between the resulting light-based inequality indicator and existing estimates of net income inequality,“ they also state that, "While it is widely acknowledged that inequality has profound effects on the functioning of societies, data remain frustratingly scarce." There’s work to be done.

NASA is getting ready to launch the James Webb Space Telescope (JWST) which will be able to work in the infrared spectrum in order to probe the atmospheres of thousands of newly discovered exoplanets for molecules such as carbon dioxide, water, methane, and others that could suggest the presence of life. The development of the JWST proved to be the most complex and difficult project in NASA’s history, taking 25 years and costing $9 billion. Fortunately for NASA, the agency can depend on thousands of amateur astronomers to work on two other complex projects — TESS and Planet Patrol, which need humans to manually confirm the identification of exoplanets out of millions of potential targets. "Citizen scientists have already discovered half of the known comets," says NASA astronomer and citizen science officer **Marc Kuchner**, "They’re making discoveries left and right."

Large-scale data capture from distributed commodity sensors in cell phones has become a creative exercise for scientists. **NASA** earth scientist **Sten Odenwald** used cell phone magnetometers to investigate space weather and Sun-related geomagnetic storms. A **Google** prototype (not **FDA**-approved) matches skin diseases to high-quality cell phone pictures of all kinds of skin types. **MIT** researchers used global cell phone location data to test whether city dwellers ever travel any distance further than what's necessary. Guess what? They don't! "This kind of intuitive notion had never been empirically tested. When we did it we found an incredibly regular and robust law — which we have called the visitation law,” said **Carlo Ratti**, director of **MIT's Senseable City Lab**, which led the research project. "Uh, duh," said everybody else.

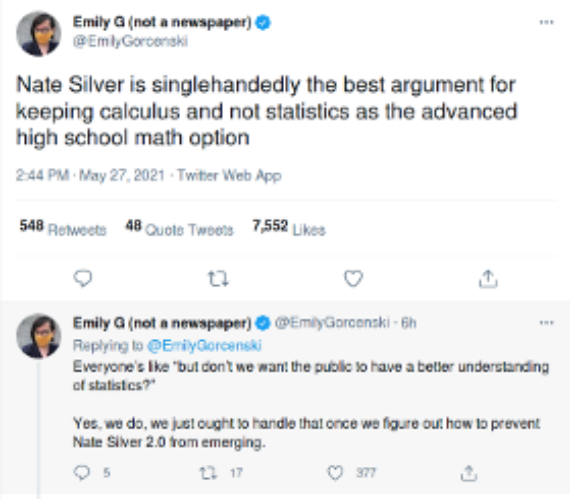
Jon Kleen at the **University of California-San Francisco** used a dime-sized, flexible, high-density electrode grid invented by **Razi Hacque** at **Lawrence Livermore National Laboratory** to record brain signals in and out of the hippocampus, the brain's memory center. Kleen found brain waves at different frequencies and coming from different directions (likely carrying different information) arriving at the hippocampus during a task. Future progress into how the hippcampus integrates information will come as researchers develop smaller, denser electrode arrays. Biomedical engineers at **Duke University** are using machine learning to automate the active brain imaging that's done with two-photon calcium imaging. Here neurons appear as flashes of light in video frames, and will overlap when thousands of neurons are in the image frame. Deep learning has improved to where images can be segmented (circling the light flashes) in real-time instead of the hours or days required before.

Commercial integrated photonic accelerators designed to speed neural network calculations are starting to hit the market with new and developing products from companies like **Lightmatter** (Boston), **Luminous.co** (Mountain View, CA), **LightIntelligence** (Boston), **LightOn** (Paris) and **Optalysys** (Wakefield, England). Input data for a neural net is supplied electronically and then converted to optical signals which are easier to process in parallel (separate the beam) and at extraordinary speed (the rate of photon detection) before the output product is converted back to electronics. **Jeff Hecht** wrote an excellent near-term roadmap at **Laser**

Focus World. Neuromorphic computing, another burgeoning next-gen data processing architecture where memory and processing are co-located, is the subject of its own joint academic-industry roadmap described by **Nicole Hemsoth** at **The Next Platform**. According to this roadmap initial applications of neuromorphic computing will occur first at the edge (robotics, embedded systems, wearable technology) rather than in datacenters. Biological or DNA-computing does not have the speed to be a viable data processor, but researchers at **Boise State** borrowed the idea of a Lite-Brite to encode DNA with pegboard-like fluorescent patterns for rapid microscope inspection. The prototype's data density was 330 gigabits per centimeter and given that data stored in DNA lasts for centuries the DNA Lite-Brite could be a viable archival storage medium.

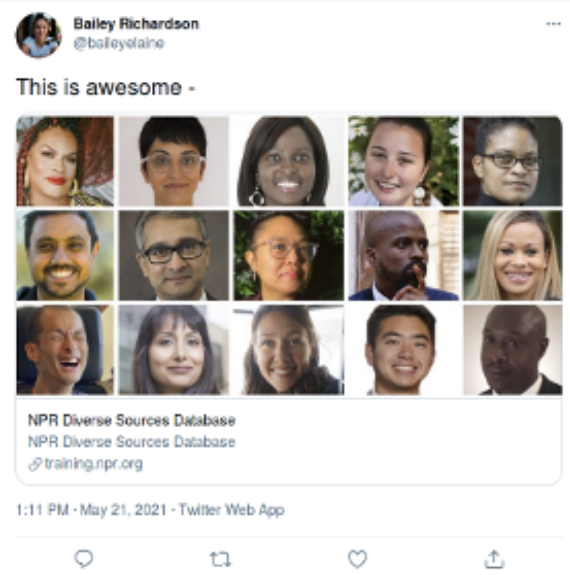
Tweet of the Week

Twitter, @EmilyGorcenski from May 27, 2021



Data Visualization of the Week

Twitter, Bailey Richardson from May 21, 2021



Events

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

Value of Science: Data, Products & Use
Online June 23-24 "Join the **National Center for Science and Engineering Statistics** and the **Coleridge Initiative** for this two-day conference to advance understanding of the value of data by showcasing new data, products and use resulting from recent NCSES investments." [registration required]

Sustainable AI - Conference
Online June 15-17. Organized by the **Aimee van Wynsberghe** (CCSRE) from the **University of Bonn**, supported by **Alexander von Humboldt Foundation**, and the **Institute for Science and Ethics** (HAI). [registration required]

Data Science Connect K-12
Online June 17 starting at 10:30 a.m. Eastern. [free, registration required]

Deadlines

Contests/Award
The TensorFlow Microcontroller Challenge
Deadline for submissions is June 19.

RFPs
The Call for Proposals for the @ MWBigDataHub Community Development and Engagement program is open!
"Details: http://midwestbigdatahub.org/cde/" Deadline for research proposals is July 5.

Studies/Surveys
Let's talk about the 2021 #CSforALLSummit...
"As we plan for this year's 2021 #CSforALL Summit, we need your feedback. Please take this survey to help us evaluate the type of interest for an in-person event vs hybrid and/or virtual CSforALL Summit"

Tools & Resources

As part of the #ResponsibleCS Challenge, 22 schools and universities have come together to create a playbook to help students bring ethics to the design of technology products.
Twitter, Mozilla from May 12, 2021

"Learn more and contribute to the playbook"

Google's New Wave of Machine Learning Capabilities
Jesus Rodriguez, TheSequence newsletter from May 23, 2021

"The AI announcements at **Google I/O** were highly diverse. One of the most interesting was the unveiling of Vertex AI, a new managed cloud service to accelerate the deployment and maintenance of ML models. Vertex AI is a particularly interesting release and it shows significant overlap with Google Cloud ML. Another announcement that captured the headlines was LaMDA, a new conversational model that can

produce more natural dialogs by better understanding its context. Google also announced a new generation of Tensor Processing Units (TPUs) chips to power AI workloads as well as new AI-native personalization capabilities to its Firebase mobile backed platform."

BaseTen
Product Hunt from May 19, 2021

"The fastest way to build apps powered by machine learning. Deploy models with a few lines of code, serve APIs without infrastructure or framework nightmares, and build stateful, interactive user interfaces to power real, functional applications"

Careers

See the ADSA Jobs Page for 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. The newsletter is written and the content is compiled by the Academic Data Science Alliance. Our archive of newsletters is at cds.nyu.edu/newsletter and is the process of transitioning to another, permanent location. Our mailing address is 1037 NE 65th St #316; Seattle, WA 98115.