



## ADSA Data Science Community Newsletter

**Data Science Community Newsletter** features journalism, research papers and tools/software for March 16, 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](#).

### **From the desk of Laura Norén**

**Federated learning** is super popular in industry. Why? It distributes computational load to user's devices and offers decent protection for user privacy. Federated learning has started showing up in academic research, mostly in medical imaging applications, and in masters level curricula. We put together a short overview to get our readers up to speed if they aren't totally sure what federated learning is, what it's good for, or whether the claims about privacy protection are acceptable at face value.

More at <https://academicdatascience.org/newsletter/from-the-desk>.

### **Academic Data Science News**

**Allison Morgan** and **Aaron Clauset** from **University of Colorado** recently found that, among academics, women publish less frequently than men and that the [demands of parenthood is a cause](#). The finding backs up a **National Academies of Sciences, Engineering, and Medicine** [report on how women academics are faring](#) and caps an unseemly trend where women in academia are discouraged at work in their chosen profession. **Dani Bassett** from **University of Pennsylvania** led a group that found significant [under-citation of female authors in neuroscience journal papers](#). A recent **National Bureau of Economics Research** [paper](#) by **Pascaline Dupas**, **Alicia Sasser Modestino**, **Muriel Niederle**, **Justin Wolfers** (and 97 research economists known collectively as the **Seminar Dynamics Collective**) identified prevailing anti-social uncollegial behaviors toward women at economics seminars (despite [issues with methodology](#)). Notice that these issues are not problems with the diversity employment pipeline. The academics being discouraged made it through the talent funnel, only to experience impediments that male peers do not face. Problems still exist in the pipeline. Sports analytics, a research profession tied to academic research, was recently [called out](#) as "overwhelmingly white and male." Enough authorities are aware. **UNESCO** spent time [discussing unequal hiring](#) at its recent global dialogue on bias in AI. The EU commissioner for research and innovation announced the [Women TechEU initiative](#). In British Columbia, **Athena Pathways** is a consortium helping [advance the careers of hundreds of women in data science](#). The professional hockey team in Pittsburgh [recently hired Katerina Wu](#) as a data scientist reporting to **Sam Ventura**, the Penguins' director of hockey operations and a former **Carnegie Mellon** statistics professor. Feel good about the progress, but realize that what lasts is when professionals thrive in their occupations, not when they are hired into new jobs.

**Phil Bourne**, Stephenson Dean of the **School of Data Science at University of Virginia**, [authored](#) a one-year status report of UVA SDS. In-state rival **Virginia Tech** [broke ground](#) on its new 115,000 square-foot Data and Decision Sciences Building.

The **School of Computer Science at Carnegie Mellon** named **Kit Needham** as its [first-ever Assistant Dean for Entrepreneurship Initiatives](#). Needham has been helping entrepreneurs on campus since 2008. Student entrepreneurship services are not new but they are focusing more on data science. **Northeastern University** is [partnering](#) with **Techstars** on a startup incubator at the school's **Roux Institute** in Portland, Maine. **Princeton** recently [concluded its first Startup Bootcamp](#),

also in partnership with **TechStars**, held during the January wintersession. The **University of Wisconsin-Madison** [recently opened](#) The Entrepreneurship Science Lab within the school's interdisciplinary **Institute for Discovery**.

Sometimes the path to maximum impact isn't a for-profit. **Project Jupyter** founders **Fernando Perez** and **Brian Granger** met as **University of Colorado** grad students. The real work began years later after Perez had created the original IPython Notebooks (2001) and Granger wrote code for IPython Notebook server (2011). The code sprint that took IPython to Project Jupyter happened in 2014. Journalist **Matt Asay** recently wrote [the complete history](#) at **Techrepublic**.

**Andy Peters** at the **Atlanta Journal-Constitution** [documents](#) the current push by big West Coast tech companies to establish offices in Georgia. The reason: Atlanta's black tech talent.

California state college students interact [regularly with AI chatbots](#), according to the **Los Angeles Times**. Engagement rates are high, reportedly above 90 percent. The systems with names like Billy Chat, Cougar Bot and Csunny help students to keep on track toward graduation, but they also absorb enormous amounts of pandemic-related student angst.

**Florida State University** [launched](#) an Interdisciplinary Data Science Master's Degree program (IDS). The program is exclusively on-campus in Tallahassee and first students will begin during Fall 2021.

**Petros Koumoutsakos**, who joined the **Harvard** faculty this semester, [assumed the directorship](#) of **Harvard's Institute for Applied Computational Science**.

The **College of Engineering and Computer Science at Florida Atlantic University** [established](#) the **Center for Connected Autonomy and Artificial Intelligence** (CCA-AI). The Center is included in FAU's \$1 million grant from National Science Foundation to establish graduate degree programs in artificial intelligence for high-achieving, low-income students.

**Brookdale Community College** in New Jersey [has added](#) a Data Science Option to the school's Associate Degree program in Mathematics/Science.

**Vanderbilt University's Data Science Institute** has produced what can only be called a "sizzle reel" starring **Jesse Spencer-Smith**, the Institute's Chief Data Scientist.

## SPONSORED CONTENT



**BIDS Cross Domain Initiatives (XDs)** connect data science community members from all STEAM fields who use graphs and network analysis, image analysis, and text analysis for applications across domains. Please join us for **BIDS GraphXD 2021** on **March 23**, and **BIDS Image XD 2021** on **May 17-18**, when distinguished researchers and educators from academia and industry will present new analytical methods, tools, and discoveries. To get involved in the BIDS XD community online, [sign up for the BIDS XD Community on Discord](#).

## Editor's Picks

The **National Security Commission on Artificial Intelligence** has issued its [final report](#). The Commission, chaired by **Eric Schmidt**, calls on the U.S. government to address [the threat of China](#) becoming the “AI superpower” by more than doubling our AI research and development spending to \$32 billion a year by 2026. (China has [already surpassed](#) the U.S. in journal citations.) It also suggests establishing a new body to help **President Biden** guide AI policies, relaxing immigration laws for highly skilled AI researchers, [creating a new university](#) to train digitally skilled civil servants, and accelerating the adoption of new technologies by U.S. intelligence agencies. The report – stating that the U.S. is “not prepared to defend or compete in the A.I. era” – also calls on the U.S. to reject calls for a global ban on highly controversial AI-powered

autonomous weapons, saying that China and Russia are unlikely to keep to any treaty they sign.

The pandemic stay-at-home life has surfaced enormous public interest into food, and into food technology. Recent developments include being able to take the [guesswork out of ripening cheese](#) and using AI to [analyze flavors in coffee](#). **Columbia University** postdoc **Emily Spratt** created [algorithms to investigate Alain Passard's](#) haute cuisine plate presentations. Somewhat lower brow, fast-food chains are testing the use of AI to speed up the ordering process ([McDonalds](#), [Lee's Famous Recipe Chicken](#)). Finally – and for some, most importantly – creamy or crunchy? Find out what [your PB preference says about you](#) in survey results.

It may come as no surprise to learn that when tweets about COVID-19 are analyzed using machine learning, results show social media having significant [influence on people's health beliefs](#). Perhaps more surprising is work showing [the effects of coordinated information campaigns](#) that rely heavily on networked participation – particularly how these groups discuss, interpret, or conduct their own data analysis in order to come to entirely different conclusions. Using the rhetoric and tools of mainstream data science, anti-mask data visualizations directly contradict those made by newspapers and public health organizations, and it can often be difficult to reconcile these discussions around the data. Fighting data with data – while not new – is a particularly corrosive – and in the case of COVID-19 downright deadly – aspect of our divisive culture wars. New developments may soon be able to offer [lab-free COVID testing](#), with results delivered on your cell phone in under an hour, and an [at-home COVID-19 testing system](#) is being evaluated by **NIH**. But if something like half of the U.S. population don't wear masks in public, ease of testing (and of getting vaccinated) don't seem to be the robust opponents needed to combat social media-fed disinformation.

The term "hyperscalers" refers to the global cloud-computing providers. It's become a helpful one-word bundle for all of the usual suspects among huge Internet software and data corporations. [This article](#), for example, *China's Hyperscalers Strive to Keep Pace in Open Source*. With that here's a quick hyperscalers update. **Google** remains [mired](#) in its self-created AI Ethics [quagmire](#). **Microsoft** is pushing for [decentralized user IDs](#) via blockchain. Also, **Jesus Rodriguez** [analyzed](#) the company's new Azure initiatives. **Amazon** partnered with **Code.org** on [pre-college computer science](#)

[curriculum](#). **Apple** is making a [stutter detection add-on](#) for Siri. **MIT Technology Review** [profiles Joaquin Quiñonero Candela](#), Facebook's Director of AI, digging into **Facebook's Responsible AI group** and its not-so-responsible mission.

## Research News

**Stanford University's Institute for Human-Centered Artificial Intelligence** published its [annual AI Index Report](#). Among the many interesting details, the report says that 65 percent of graduating PhDs are taking industry jobs (2019 data), up from 44 percent two years earlier.

Two new sources of artificial intelligence case studies are out. An [ongoing series of AI case studies](#) from **MIT Stephen A. Schwarzman College of Computing** are "rigorously" peer-reviewed. The first four are out, available for free with open access. A [stand-alone case study](#) authored by **Brian Green (Santa Clara University)** and **Daniel Lim (World Economic Forum)** puts **Microsoft's** "journey to operationalize its responsible innovation practices" under a microscope. Case-based learning, it seems, is finding its place in data science education.

A new **University of Vermont** [study](#) reveals [connections between climate changes and reduced diet diversity](#) and food security for children under 5, with data gathered from across 19 countries in Africa, Asia and Latin America. Using hierarchical modeling, the study synthesizes health and climate data, revealing climate change impacts on children's health to be at least equal to those incurred from variability in wealth, sanitation and education. These effects are one of the drivers behind the [push to prepare food systems for climate change](#). As part of this push, **MIT** is working on [analytic tools](#) to detect biotic and abiotic stress, monitor plant hormonal signaling, and characterize soil, phytobiome, and crop health in a non- or minimally invasive manner. And while (mostly) unrelated to climate change, another threat to agriculture is revealed through research showing that [at least one third of the topsoil in the Midwest is gone](#), mainly due to centuries of plowing. (This research is based on aerial observations, which can also fuel the data gathered to fight wildfires.)

Ongoing degradation to the environment is causing a [rapid decline in the number of](#)

butterflies in the Western U.S., and it is now suspected that the ubiquity of microplastic particles around the globe – being found on the High Alps, in the Arctic and Antarctica – is a major driver of snow and ice melt, perhaps rivalling that of black carbon particles. And in a weird, M. C. Escher-ish kind of development, climate change itself is now driving up insurance costs for solar installations – which of course are an effective way to lower climate change effects.

Trustworthy artificial intelligence is important for environmental research applications (e.g. weather, climate, coastal oceanography) to influence public policy. If the AI applications don't get used, risk communication specialists want to know why. Additional work is being done on analyzing metrics to quantify connectivity of networks and monitor changes over time in an ecoscape (landscape or seascape). (Especially helpful for species conservation – see butterflies above.)

Social Determinants of Health is a phrase that has gained traction in the past few months. The **University of California-San Francisco** has formalized the idea in a new data platform with "150 social determinants of health variables" sourced from electronic health records. There's plenty of low hanging fruit for SDoH research without the UCSF platform. Racial bias and knee pain and less accurate pulse oximetry readings for darker skin are two examples. As more and more useful health sensors come from wearable and at-a-distance devices, expect continued progress into understanding the nexus of social factors and personal health. And as a recent **National Academies' report** investigating Americans' working-age mortality shows, there are significant future public policy implications.

The **MIT Press** launched **Direct to Open (D2O)**, an effort to move scholarly books away from a market-based single-purchase business model to a collaborative, library-supported open access model. The goal, according to MIT Press publisher **Amy Brand**, is to "increase the impact and accessibility of high-quality scholarship."

**Raspberry Pi**, makers of small, inexpensive, yet powerful microcontroller hardware, has an in-house chip development team that is promising built-in machine learning for their next-iteration products. The current iteration, Pico, costs four dollars and shows promise as an ML platform. In its short time available the Pico has sold 250,000 units, with 750,000 units currently on back order.



Researchers at the **Institute for Systems Biology** in Seattle have discovered [how microorganisms evolve cooperative behaviors](#) in a tour de force of computational biology. Taking advantage of model organisms for syntropic interactions (one organism supplies the other with a nutrient in exchange for a reward) researchers sequenced hundreds of samples and investigated mutations that accumulated in the organisms over the course of 1,000 generations. Cooperation, they found, is the result of rare mutations that manage to persist at very low frequency within a larger population. These same patterns occur independently in similar genes across independently evolving populations. Parallel evolution. Amazing.

## Tweet of the Week

Twitter, Talia Ringer from Mar 9, 2021



## Data Visualization of the Week

Twitter, Brendan Dolan-Gavitt from Mar 4, 2021







## Events

See the [ADSA Events Page](#) for more details and more opportunities.

### Join us next week on March 17th for our Diversity, Equity & Inclusion SIG!

**Online** March 17, starting at 10 a.m. Pacific. "We will discuss the **ARC Network** - a STEM Equity Brain Trust and their mission to achieve gender equity for faculty in higher education STEM disciplines." [registration required]

### AI's Impact on Teaching

**Online** March 17 starting at 2 p.m. Eastern. "A panel of leading thinkers and practitioners in the use of AI will join **The Chronicle of Higher Education** for a discussion of AI and machine learning in higher ed's future." [rsvp required]

### Social Media Summit @ MIT

**Online** April 22, starting at 9:15 a.m. Eastern. [free, registration required]

## Deadlines

### Contests/Award

#### [Inaugural Patrick J. McGovern Tech for Humanity Prize](#)

"Anyone in any field pursuing creative and meaningful projects that examine, interpret, develop, or apply technology to shape a better future for humanity may apply." Deadline for applications is March 28.

### Education Opportunities

#### [ASA Pride Scholarship](#)

The scholarship "was established to raise awareness for and support the success of LGBTQ+ statisticians and data scientists and allies. The scholarship will celebrate their diverse backgrounds and showcase the invaluable skills and perspectives these individuals bring to the ASA, statistics, and data science." Deadline for nominations is March 31.

#### [Allen Institute launches fellowship for data scientists to study the brain](#)

"The **Shanahan Foundation** Fellowship at the Interface of Data and Neuroscience, a 3-year fellowship program co-hosted by the **Allen Institute** and the **University of Washington**, is now accepting applications" Deadline for applications is April 15.

#### [University of Michigan Data Science Fellow Program](#)

"This two-year program provides outstanding young researchers with intensive data science experience as they ready themselves for independent research and faculty positions." Deadline for applications is April 23.

## Tools & Resources

### Designing Data Beyond the Dashboard

*Salesforce Design, Jason Winters from Feb 22, 2021*

"The best analytics apps are seamlessly embedded where your users already do their daily work and push the right insights at the right time. This model offers access to relevant data in context, enabling users to take action without leaving their workflow."

### Understanding Deep Learning (Still) Requires Rethinking Generalization

*Communications of the ACM; Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, Oriol Vinyals from Mar 1, 2021*

"Despite their massive size, successful deep artificial neural networks can exhibit a remarkably small gap between training and test performance. Conventional wisdom attributes small generalization error either to properties of the model family or to the regularization techniques used during training.

Through extensive systematic experiments, we show how these traditional approaches fail to explain why large neural networks generalize well in practice."

**Who is ML interpretability meant for, and how can we design methods + interfaces that are truly understandable and relevant for them? Our CHI paper proposes a framework for thinking about these stakeholders beyond categorizations like "domain expert"**

*Twitter, Harini Suresh from Mar 3, 2021*

"For now, find it here! <http://vis.csail.mit.edu/pubs/beyond-expertise-roles.pdf>"

## 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](https://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](#).