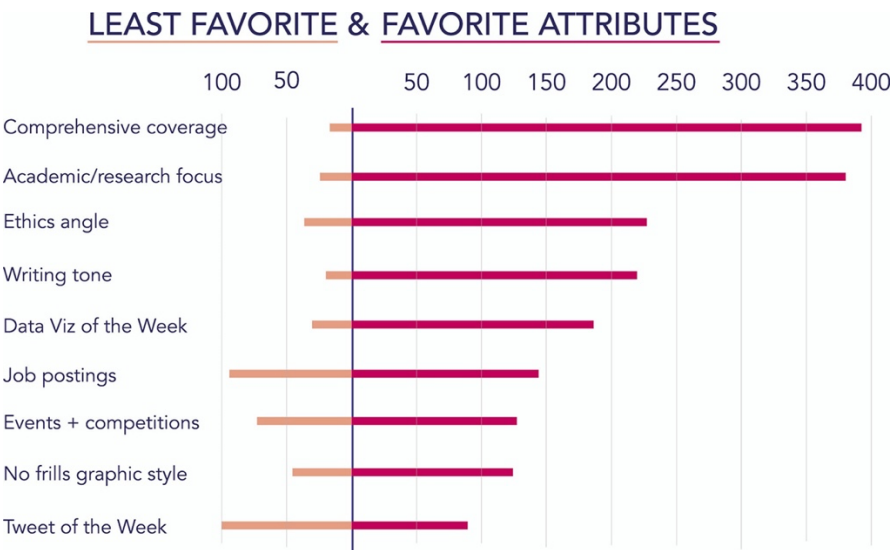
ADSA Data Science Community Newsletter

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

What do DSCN readers want?



You like our comprehensive coverage. "Tweet of the Week" not so much. Thank you readers! Your input is an enormous help. Read more about our survey results at the Academic Data Science Alliance From the Desk blog.

Academic Data Science News

The **U.S. Senate** passed the US Innovation and Competition Act (S.1260), which includes a funding increase for NSF of $81 billion over the next 5 years, including $29 billion for a **Directorate for Technology and Innovation**. The new directorate would support projects in 10 focus areas, including materials science and engineering, quantum computing, and artificial intelligence and machine learning. The act includes $10 billion to create regional technology hubs across the country. “If we’re going to make this big investment in big technologies, we want to ensure there’s geographic diversity and success,” says **Deborah Altenburg** from the **Association of Public and Land-grant Universities**. “As a nation, it doesn’t serve us well to only have the coasts participating in these efforts.” The House has its own proposal, the **National Science Foundation** for the Future Act (H.R.2225), offering a budget increase of $1 billion each year for the next 5 years. “The House bill is more in line with what I’d consider the traditional activities

of NSF," says **Tim Clancy**, president of **Arch Street**, a Washington DC consulting firm. The two bills differ in their philosophies of how the NSF should fund various activities: while the House bill relies on traditional peer review, the Senate's pushes for alternative ways to fund outside peer review, akin to the **Defense Advanced Research Projects Agency** (DARPA). **The White House** has requested $10.2 billion for the NSF, including $865 million for a Directorate for Technology, Innovation, and Partnerships intended to "serve as a cross-cutting platform that leverages, energizes, and rapidly brings to the market and to society the innovations that result from all of NSF's investments."

NSF and the **White House Office of Science and Technology Policy** (OSTP) announced the formation of the **National Artificial Intelligence Research Resource Task Force**. As directed by Congress in the "National AI Initiative Act of 2020," the task force will serve as a federal advisory committee, developing an implementation roadmap for the National AI Research Resource, a shared research infrastructure providing AI researchers and students with access to computational resources, high-quality data, educational tools and user support. "America's economic prosperity hinges on foundational investments in our technological leadership," said newly-installed Science Advisor to the President and OSTP Director **Eric Lander**. "The National AI Research Resource will expand access to the resources and tools that fuel AI research and development, opening opportunities for bright minds from across America to pursue the next breakthroughs in science and technology." In addition to assessing how the government will monitor and regulate foreign influence, Lander is particularly focused on addressing the next pandemic, building on the promise of a new agency within the **National Institutes of Health** (NIH). Lander is the first OSTP Director to be a Cabinet-level appointee, a strong signal of support by the Biden White House.

Michael Littman, data science professor at **Brown University**, called out his research colleagues involved in collusion rings. This was in a Viewpoint article published in the *Communications of the ACM*. The issue

resonated. **Jacob Buckman's** response essay found its way to reddit/r/MachineLearning. Littman writes that collusion threatens integrity. Buckman followed, saying that integrity is shades of gray, and more "blatant academic fraud" will help to clarify the state of the computer science enterprise. Magazine opinion columns, blog posts, **Reddit** discussion, this newsletter – these are all checks on integrity. And they are all relatively new entries into our common discourse, making it hard to project forward what, if any impact these remarks will have. Integrity checks won't make data science into an oasis of fairness in an unfair world, but it might open eyes to the erosion of trust that can dismantle collaboration among highly inter-disciplinary researchers in computer and data science.

Preserving integrity is a process. **Elisabeth Bik** is a research integrity consultant who scrutinizes research papers and publicly flags the concerns she finds in published scientific works. *Chemistry World* recently spotlighted Bik's process in an article on paper mills and fraud in science publishing. Bik has been sued by **Didier Raoult**, a prominent infectious disease researcher whose pre-print suggested that the anti-malarial drug hydroxychloroquine could treat COVID-19. Bik flagged problems with the same Raoult paper after his pre-print was rushed to publication by the *International Journal of Antimicrobial Agents*. This was in March 2020. Bik has received widespread support, according to journalist **Holly Else**, writing in *Nature*, who also points out the "chilling effect" that could occur if the legal action succeeds.

Tech journalist **Nancy Scola** has an essay on the limits fact-checking has on social media at her **Substack** newsletter, *Slow Build*. It's an impossible task for small groups of truth seekers to address incorrect details at the scale of social media, and incentives favor the continuation of what's false, not the insertion of what's correct. Accountability, as carried out by social media companies, has become theater, something "that takes the pressure off them to responsibly operate their networks." The stakes are high. The work is difficult. Maintain integrity. Preserve trust.

Iowa Board of Regents approved a two-year masters degree program in Artificial Intelligence at **Iowa State**

University. Plans are to enroll five students this Fall and eventually grow to 70-80 students.

The **University of Wisconsin School of Medicine and Public Health** launched the **UW Center for Health Disparities Research**. Co-directors **Amy Kind**, **Barbara Bendlin** and **Andrea Gilmore-Bykovskyi** will lead a $28.5 million grant-funded initiative called "The Neighborhood Study" that will collect Alzheimer's Disease research data and examine how social determinants of health impact brain health.

The **Cornell Lab of Ornithology** received a $24 million gift to establish the **K. Lisa Yang Center for Conservation Bioacoustics**. The center was originally founded 30 years ago to study communication among whales and elephants.

The **University of Illinois' Granger College of Engineering** and **IBM** have agreed to a ten-year, $200 million funding plan to jointly establish the **Discovery Accelerator Institute**. The new Institute will include new facility construction for research in computing and quantum technologies.

The **University of California-Berkeley** secured $75 million across 3 major gifts to fund construction of The Gateway, the future home for **Berkeley's Division of Computing, Data Science and Society** (CDSS). Berkeley plans to have space for more than 1,600 students, faculty and staff at The Gateway upon completion in 2025.

Big, shiny new buildings have also begun construction at **Miami University** in Oxford OH, and at the **University of Southern California** in Los Angeles. Miami's Richard M. McVey Data Science Building will provide 87,000 square-feet of inter-disciplinary lab, office and instructional space. USC's Dr. Allen and Charlotte Ginsburg Human-Centered Computation Building will be a 116,000 square-foot "living lab" housing research centers in artificial intelligence, machine learning and robotics.

The **Silver School of Social Work at New York University** received $16 million from **Dr. Constance Silver** and **Martin Silver** for the express purpose of bringing data science to social work. The gift will found the new **Center on Data Science and Social Equity** at NYU Silver.

Three **Stanford University** groups – the **Precourt Institute for Energy**, the **Stanford Institute for Human-Centered Artificial Intelligence** and the **Bits & Watts Initiative** – will fund two new projects at the nexus of artificial intelligence and energy systems. One project led by **Ines Azevedo**, **Sally Benson**, **Adam Brandt**, **Ram Rajagopal** and **John Weyant** will focus on pathways toward zero emissions and decarbonization. The other project led by Rajagopal, Azevedo, **Arun Majumdar** and **Andrew Ng** will build tools for new electricity infrastructure.

Microsoft chief executive officer **Satya Nadella** and his wife **Anu Nadella** donated $2 million to the **University of Wisconsin-Milwaukee**. The university will put the money into a new Fund for Diversity in Tech Education. Satya Nadella completed his masters degree in Computer Science at UWM in 1990.

The **University of Delaware** launched a new master's program in biopharmaceutical sciences with support from **AstraZeneca**, **Bristol Myers Squibb** and **Merck & Co**. The 15-month program will emphasize analytics and data science, and will include an internship that runs during the entire term of study.

University of Miami (the other one, in Florida) joined the **National Science Foundation-sponsored Center for Accelerated Real-Time Analytics** (CARTA). The CARTA consortium is led by **Yelena Yesha** at the **University of Maryland, Baltimore County**, and counts **Rutgers University**, **North Carolina State University** and **Tel Aviv University** in Israel as other members.

DataRobot, an enterprise AI multi-national based in Boston, has partnered with **West Virginia University** to establish a "creative office" near campus in Morgantown. DataRobot expects to grow the outpost quickly, claiming to have taken on 500+ new hires in the past six months.

Editor's Picks

According to **University of Minnesota** professor **Steven Ruggles** the **U.S. Census Bureau** plans to "replace the American Community Survey (ACS) microdata with 'fully synthetic' data over the next 3 years. The problematic move comes as the

Research News

The high seas – defined as ocean areas more than 200 nautical miles offshore – are beyond the jurisdiction of any nation, making them popular amongst pirates (Argh!) but they are also home to the largest reservoirs of ocean biodiversity on the planet. A team of researchers led by **University of California-Santa Barbara** have sufficient data and research to begin mapping and modeling hot spots of biodiversity on the high seas to support the development of a new inter-governmental treaty to allow nations to establish comprehensive, cross-sector marine protected areas (MPAs) on the high seas. The researchers used a "systematic conservation prioritization software" called prioritizr R that identifies potential areas that meet conservation objectives while allowing for commercial fishing activities. "We synthesized insight from over 20 billion data points about ocean wildlife — and about how people were using the ocean—to try to find high seas biodiversity hot spots deserving of protection," said **Doug McCauley**, an assistant professor of ecology, evolution, and marine biology at UCSB, who led the research team.

Declining fish biodiversity can affect human nutrition, according to a computer modeling study led by **Cornell** and **Columbia University** researchers. The findings apply to fish biodiversity worldwide, as more than 2 billion people depend on fish as their primary source of animal-derived nutrients. "As you lose biodiversity, you have these tradeoffs that play out in terms of the aggregate quantity of nutrients," said author **Sebastian Heilpern**, postdoc in the **Cornell Department of Natural Resources and the Environment**, adding the system becomes "more risky to further shocks." Practical steps could include establishing and enforcing "no-take zones" – areas set aside by the government where natural resources can't be extracted. Scientists at **Duke University Marine Lab** and the **Wildlife Conservation Society** (WCS) used a deep-learning algorithm to analyze more than 10,000 drone images of mixed colonies of seabirds in Argentina's Malvinas/Falkland Islands. The results were within 5% of human counts about 90% of the time. "Using drone surveys and deep learning gives us an alternative that is remarkably accurate, less disruptive and significantly easier. One person, or a small team, can do it," said **Madeline C. Hayes**, who led the study.

Bringing a level of sustainability to oceanographic exploration is the goal of **Seatrec**, which manufactures profiling floats — robotic data-gathering devices that monitor the ocean's physical, chemical and geological characteristics. Seatrec's innovation is to make the floats rechargable through the use of the ocean's thermal energy. Seatrec founder **Yi Chao** says this technology enables floats to last indefinitely, and sample more frequently with more sensors. Such capability allows

for more comprehensive data gathering, which can be deployed to better predict hurricanes, for example.

Lots of ocean reporting occurred on and around June 8 — World Ocean Day — and **National Geographic** won the coverage, announcing that its world maps would now have 20 percent more ocean (by name). National Geographic added a fifth major saltwater basin, the Southern Ocean, which surrounds Antarctica and joins the original four: Atlantic, Pacific, Indian and Arctic.

Puneet Batra leads machine learning at **Broad Institute** in Cambridge MA. (**Eric Lander's** old stomping ground.) He recently outlined how the **Eric and Wendy Schmidt Center** at the Institute, endowed with $300 million from the Schmidt and Broad foundations, intends to approach machine learning, saying that "Machine learning needs to move from predictive accuracy to causal modeling, from 'what?' to why?'" Biological machine learning, according to Batra, seeks to understand natural laws, not just solve problems. He expects the approach to lead to advances in computing and, at the same time, improve patient care.

The **National Institutes of Health** (NIH) granted $1.2 million to **Rice University** computer scientist **Lydia Kavraki** to investigate "protein-ligand interactions in cancer research." The project fills a knowledge gap in fundamental structural biology and builds tools managing the enormous combinatorials of proteins, ligands and small molecules with anti-tumor activity. Kavraki invented probabilstic roadmap algorithms for robots' motion planning, a helpful starting point for understanding how biomolecules' structure and motion are intimately related to their function.

Big science meets data science meets life science can also be brute force systems biology wrangling at larger and larger scales. **Naftali Kaminski**, pulmonologist at **Yale School of Medicine**, analyzed 15,000 endothelial cells (oxygen exchange cells) from 73 donors to create a cellular blueprint of what healthy lungs should do and look like. **Jeff Lichtman's** lab at **Harvard** – with help from **Google's Connectomics group** – created and released the "HO1" dataset, a 1.4 petabyte rendering of a small fragment of human cerebral cortex. The data covers "roughly one cubic millimeter of brain tissue, and includes tens of thousands of reconstructed neurons, millions of neuron fragments, 130 million annotated synapses, 104 proofread cells, and many additional subcellular annotations and structures" and can be viewed in web browsers with its Neuroglancer interface. British synthetic biologist **Jason Chin** used CRISPR to replace 18,000 sections of E-coli genome (4 million base pairs) to invent the Syn61 strain of the model organism. Syn61 can test multiple amino acids against a protein simultaneously, boosting productivity from what is typically a one-at-a-time cellular bioengineering practice.

Sometimes big-data-life science means NLP-ing the medical literature corpus, which **Nigam Shah** and **Jason Fries** from **Stanford Institute for Human-Centered Artificial Intelligence** did to invent Trove, an open-source framework that uses "'weak supervision' to automatically classify entities in clinical text using publicly available ontologies (databases of biomedical information) and expert-generated rules." The framework augments clinician's notes and performs on par with manual annotations done by doctors, adding a knowledge management capability that helped to clarify dynamic symptoms in infected population, something they tested during early-Covid.

While farmers have always genetically engineered crops through seed selection, AI and genetic engineering are producing new commercial crop strains that can better withstand climate change effects. CRISPR is now being used to teach plants to absorb nitrogen from the air, which would reduce the need for chemical fertilizer, and bacteria can be introduced to allow plants to gain more energy and water-use efficiency through photosynthesis. **Cornell** engineers and plant scientists recently invented nanoscale sensors that measures the water in leaves of still-growing plants. Conservation-minded farmers and the researchers and NGOs interested in supporting them can help make the financial case for environmentally-friendly practices through the use of a new guide published by the **Environmental Defense Fund**. And the job skills needed to be a successful farmer are changing swiftly, according to farm-tech entrepreneur **Raviv Itzhaky**. Prominent among them is acquiring data analysis skills, as well as skills related to human-robot interaction (HRI), as harvesting becomes increasing mechanized. Robotics that use computer vision for weeding can reduce pesticide use by 90 percent. The knock-on effect of the inevitable massive reduction in the need for migrant workers will however affect humans, not robots.

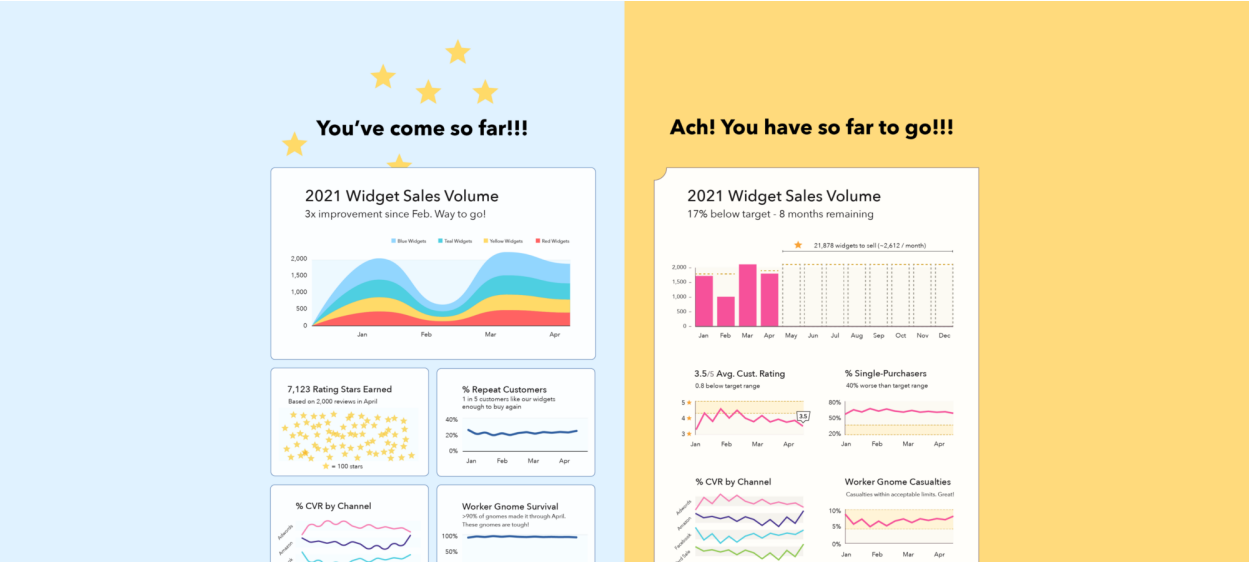
PiAutoStage could make geoscience research more accessible. It's a research tool that automatically take pictures of entire thin sections (slivers of geological material) and stitches them into digital panoramic microscope images that can be analyzed anywhere, and the tool can be 3-D printed for $200, so that anyone can make and use it. (It consists of an open-source mechanism that moves the sample around the microscope, attached to a high-resolution integrated camera and inexpensive Raspberry Pi computer.) PiAutoStage could also provide instructors a new, affordable resource to digitize and use specimens and literary materials they already have in their curriculum. A new paper shows how subterranean microbes can trap globally significant amounts of carbon, which eventually will be released into the atmosphere, representing an overlooked factor in efforts to balance Earth's deep carbon cycle. These massive networks of microbes are found in subduction zones, such as the area around **Costa Rica**, the site of the study. The researchers also

found evidence for a second group of microbes that live off the organic leftovers of the carbon-sequestering bacteria. "There's a whole world happening underneath Costa Rica," says **Karen Lloyd**, co-author and microbiologist at the **University of Tennessee, Knoxville**. The researchers suspect similar activity is taking place in other subduction zones all over the world.

Machine learning is being used to improve weather forecasting with research being done at **Stony Brook University** focused on a phenomenon known as the Madden-Julian oscillation (MJO), a belt of thunderstorms that moves slowly eastward toward the central Pacific Ocean every 40 to 50 days. While MJO has been used for three-to-four-week weather forecasting, computer modeling has previously not been able to simulate all aspects of MJO and therefore extended-range forecasts has had a larger margin of error. The Stony Brook team combined weather forecast models and observations with a machine learning process (a Deep Learning bias correction using all of the data) to forecast the MJO. Forecast errors were reduced by 80 to 90 percent, according to **Hyemi Kim** in the **School of Marine and Atmospheric Sciences** (SoMAS).

Data Visualization of the Week

Nightingale, "Dashboard Psychology: Effective Feedback in Data Design" by Eli Holder, June 14, 2021



Nightingale, the Journal of the Data Visualization Society, is now free to readers. DVotW is part of an excellent how-to article by data designer, **Eli Holder**.

Events

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

Data Science Coast-to-Coast Seminar

Online June 16, starting at 12 p.m. Pacific. Organized by the **Academic Data Science Alliance**. [please sign up in advance]

SciPy 2021 | Attend

Online July 12-18. [$$$]

Academic Data Science Alliance, 2021 ADSA Annual Meeting

Savannah, Georgia, and Online November 10-12. The theme of the 2021 ADSA Annual Meeting is the "Transfer Power of Data Science." [save the date]

Deadlines

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

Student Perceptions of Data Science Survey

"The #DataScience for #SocialJustice project created a survey to advance knowledge around data science exposure within the context of social issues and its impact on students, especially those underrepresented in STEM fields."

Conferences

VAHC 2021 (12th workshop on Visual Analytics in Healthcare)

Online October 24 or 25. VAHC 2021 will be held in conjuction with the **IEEE VIS 2021** conference. Deadline for submissions is July 19.

Tools & Resources

BERT for Humanists

GitHub, Melanie Walsh from May 15, 2021

"What impact might BERT-like models have in the field of the digital humanities? What impact might digital humanists have on our understanding and application of BERT-like methods?"

"The **BERT for Humanists** project is developing resources to help answer these questions and enable DH scholars to explore how BERT-like models can be used in their research and teaching."

Twitter on Elastic + Neural Nets

Twitter, The Institute for Ethical AI & Machine Learning, and Twitter Engineering Blog, Rakshit Wadhwa and Ryan Turner from April 30, 2021

"A fascinating post from **Twitter Engineering** around how they developed a recommendation engine to annotate legacy datasets to a new standardized taxonomy."

Careers

See the ADSA Jobs Page for more opportunities.