Data Science Community Newsletter

DATA SCIENCE NEWS

University Data Science News

The campus foment continues.

We'll start at **Harvard University** where graduate students have **left their classrooms** to picket for higher wages, better protections from harassment and discrimination, and affordable health care. The university — one of the wealthiest in the world with **an endowment of $40 billion** — has offered to raise graduate students' wages to $15 an hour. The Chair of the Department of Government issued **a memo** to get faculty aligned with the administration, noting that "faculty are management" and that the Department "will inform the University administration of those TFs in courses in the Department of Government who have reported they are striking, or who have not responded to our request for this information." Keep in mind that $15/hour is the mandated minimum wage in some states. All Harvard graduate students have bachelors' degrees and were/are considered the most promising applicants to the admissions committees in their departments which would seem to value their work at more than minimum wage.

The University of California system saw **picket lines at 10 of its campuses**, populated by members of its largest union: **American Federation of State, County and Municipal Employees (AFSCME)** in the sixth walkout in the past 3 years. The workers are protesting the slow creep of contract workers, who are paid less and receive no benefits. As universities come under intense pressure to maintain high levels of service with lower flows of state subsidies, there is a logic to their decision to replace unionized workers (and tenured faculty) with contract workers. But it's hard to argue that cuts should come at the expense of staffers' health insurance.

Syracuse University students are not picketing, but they are **staging marches, sit-ins, and calling for the resignation** of Chancellor **Kent Syverud**, Department of Public Safety Chief **Bobby Maldonado**, DPS Associate Chief **John Sardino** and **Dolan Evanovich**, senior vice president for enrollment and student experience. This follows a **series of ongoing racist and anti-Semitic incidents on campus** that **the students feel** have not been adequately addressed by leadership.

Tufts University sent out a memo announcing that it would **remove the Sackler name** from their medical school, their school of graduate biomedical sciences, and from within their medical school programs. They note that due to the high cost of the opioid epidemic in human lives, "it is clear that the Sackler name, with its link to the current health crisis, runs counter to the school's mission." The school also announced it will "adopt stronger screening procedures for donors,...develop more stringent conflict-of-interest policies, and both develop and make publicly available guiding principles

for gift acceptance." Another university in Massachusetts, **MIT** faced similar scrutiny — though no similar promise to make donor screening procedures public — following a revelation that MIT took money from **Jeffrey Epstein**.

"Philanthropy is no license to do bad stuff," **Patricia Illingworth** announces in the title **of a new essay**. Her conclusion: "Nonprofits must exercise due diligence for moral licensing before they accept large donations."

While I realize that it may seem the incidents above have little to do with data science, it is important to remember that we are a community of humans first. On the good days, we also get to be scientists. I encourage everyone to lean on each other for support and understanding and to provide support and understanding for those who could use it.

The Association of Internet Researchers (AoIR) has released an update to their **ethical guidelines for [internet] research**. There is an entire appendix dealing with algorithmic research and its ethical imperatives so don't be fooled by the "internet research" title. These guidelines are worth reading for anyone working with human-derived data. This is one of the most thoughtful, comprehensive groups dealing with questions that impact all data science research: what kind of data can ethically be collected? How do researchers deal with cultural differences in expectations of privacy? What are the researchers responsibilities to ethical norms that may conflict such as the science-based imperative to share data in order to foster scientific reproducibility versus the privacy-imperative to reduce the spread of personally identifiable data?

University of Washington in Seattle and **Washington State University** through UW's **Center for an Informed Public** have **formed a partnership to work on deepfakes research** with $5 million from the **John S. and James L. Knight Foundation**.

The **Howard Hughes Medical Institute** (HHMI) has **opened a national competition to find 20 new Investigators**. These appointments are among the very best research appointments available in the country, coming with 7 year terms and an illustrious cadre of colleagues. Applicants must have PhDs or MDs, a tenure-track position (with or without tenure), and between 5 and 15 years of post-training experience. Applications are due March 18, 2020.

Bruce D. Haynes of the **University of California, Davis** and **Sebastian Benthall** of **NYU Law** argue that in order to avoid algorithmic racial injustice and unfairness tied to race without ignoring the ongoing social significance of race in America, data scientists should **use more biological data — biometrics and genetics — not less**. While I agree that ignoring race is neither a good idea nor possible given the pervasive penetration of racial difference across a huge range of categories of social and biological data, I disagree that unleashing unsupervised machine learning on genetic and biometric data is a fair corrective action.

An international group of fifteen climate science researchers used three decades of fisheries and ocean temperature data to study the impact of rising ocean temperatures on aquatic ecosystems,

focusing on fishes. The group found that there has been relatively little geographic dispersion in fish populations where ocean temperatures have been stable, but fish populations in warming or cooling seas have **shifted their territory**. Some seek new depths, but where depth is unavailable, populations start to move laterally. This presents risk as parts of the ecosystem that used to physically co-occur become separated because they used to be able to find a temperature sweet spot together, but are now moving away from one another. Of note, coastal humans **are also struggling to adapt** to the changing seas.

Unwanted mail — both the physical kind and email — **has a huge carbon footprint**. I will be leading a webinar on January 10th to help everyone use their shiny new rights under the California Consumer Protection Regulation (or their somewhat older rights under GDPR), to reach out and stop the flow of unwanted physical mail and email spam.

There is a new open access publication venue for data science! **The Harvard Data Science Review** is a journal based at **MIT Press** that has two issues so far. We need more open access places to publish interdisciplinary data science so this couldn't come soon enough. :)

Sage OCEAN has a new white paper / ongoing project for taking an inventory of all the useful tools for social scientific data science: ***The ecosystem of technologies for social science research***. If you'd like to get your research tools or software listed, there's a **GitHub repo** — do a pull request.

University of North Carolina is facing a serious shortage of seats in computer science classes. Wait lists for three classes have 100+ students on them. Professors are burnt out taking on more and more students. The department can't find the money or talent to hire their way out of the problem. This may be a story about UNC, but it is typical — the **graphic, created by *The Daily Tar Heel* student newspaper, says it all**.

After reading this article on professors who are finding ways to **avoid flights to conferences**, I want to hear from you. How are you balancing the demands of your research — including conference travel and training models - against what it will take to collectively limit the production of greenhouse gases? What tips can you share? Have any departments done anything interesting in this space?

Company Data Science News

From the ashes of **Toys R Us**, the struggling toy seller that closed all 800 stores last year, two **new 'stores'** may mark the future of a data-saturated product development. The stores in Houston and Paramus, New Jersey, lease space to brands like **LEGO** and **Mattel**. The toymakers give kids a chance to play with actual toys. The stores are loaded with audio recorders and other sensors that capture data for use by the toy designers. All sales that cannot be filled in the store are **fulfilled** by **Target.com** which is undoubtedly using that data to stock its own retail shelves with the right toys. Expect more of this kind of experiential retail surveillance. If it leads to better products, let's be cautious before accusing anyone of spying on kids. It's not spying if it's spatially limited, informed, and used for the purposes to which parents are consenting.

Tim Berners-Lee announced a **Contract for the Web** written by 80 organizations for over a year with the backing of 150. It's a plan to protect the web from eroding from within due to deepfakes, misinformation, toxic bullying, and privacy violations. The backers include **Google**, **Twitter**, **Facebook**, **Microsoft**, the **Electronic Frontier Foundation**. All backers have to adhere to the principles or they will be removed from the list of endorsers, which will put some pressure on companies like Google, which obtained the medical records of 50 million people without their explicit consent. In essence, the Contract requires endorsers to fight for a free and inclusive internet that respects privacy and fosters productive communities.

Government Data Science News

In response to the new rule requiring the **EPA** to reference only studies for which all data is available, the editors of *Nature*, the *Science* family of journals, *PLoS*, *Proceedings of the National Academy of Sciences*, *Cell*, and *The Lancet* have published a response. The authors **argue in paragraph three**: "We support open sharing of research data, but we also recognize the validity of scientific studies that, for confidentiality reasons, cannot indiscriminately share absolutely all data. Datasets featuring personal identifiers—including studies evaluating genomes of thousands of people to characterize medically relevant genetic variants—are but one example. Such data may be critical to developing new drugs or diagnostic tools but cannot be shared openly; even anonymized personal data can be subject to re-identification, and it has been a longstanding practice for agencies and journals to acknowledge the value of data privacy adjustments. The principles of careful data management, as they inform medicine, are just as applicable to data regarding environmental influences on public health. Discounting evidence from the decision-making process on the basis that some data are confidential runs counter to the EPA stated mission 'to reduce environmental risks…based on the best available scientific information.'"

I absolutely support their arguments, though it's rather unprecedented for a group of editors to be making a case to a federal agency. It's hard to tell what kind of impact this could have.

New York Mayor Bill DeBlasio has **created an Algorithms Management and Policy Officer position** to "develop guidelines and best practices to assist city agencies in their use of algorithms." This comes less than a month after the city released the **report of its task force** on the social impact of using algorithms in civic decision making. The task force itself was vocally unhappy with the report at the time, but have endorsed the new position as a necessary step towards the socially beneficial use of AI.

Caroline Haskins of *Buzzfeed* (but writing for *Vice*) is reporting that **over 600 police departments in the US** have partnered with **Ring**, an **Amazon** company that makes home security systems, including motion-activated cameras designed for exterior use. Ring has been advising police that they "do not need a warrant in order to request camera footage from residents." Since the cameras can be pointed at the public sidewalk and roadways in front of people's homes, basically anyone could be captured on film. Some police officers and departments have wondered about the privacy rights aspects of being able to gain warrantless access to unmarked cameras in public space. Your homework for this week: email your police department and ask if they have a partnership with Ring.

Let me know what you find.

Singapore has released a **national AI strategy** and created a new national office for AI. The projects are mostly not new, but the national coordination is a marvel to me, compared to the US.

Extra Extra

I am a terrible human being for admitting this, but one of the greatest sources of relief upon leaving academia is no longer having to writing letters of recommendation. The pressure to write the best letter in order to ensure my students' futures in competition with more senior professors who undoubtedly had better name recognition was maximally anxiety-inducing. If two of my students were both applying for the same job/fellowship: Oy! Then there's the part about getting requests for letters 48 hours or less before they're due and the hoops required to actually submit the things. Turns out, **I am not alone in disliking the process**.

Are you a Simons Fellow or a frequenter of the Flatiron Institute? Curious how the very private **Jim Simons** made his fortune with data science? **Barry Ritholtz interviewed** **Gregory Zuckerman** about his new book *"The Man Who Solved the Market: How Jim Simons Launched the Quant Revolution."*
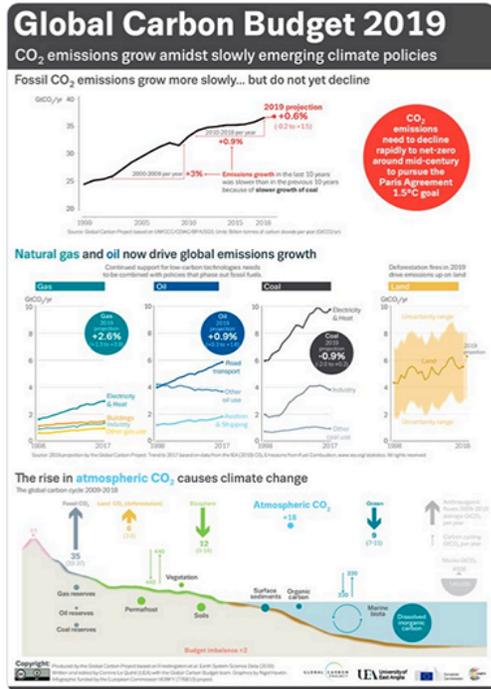
Data Visualization of the Week

Twitter, Eric Holthaus from December 03, 2019

Eric Holthaus @EricHolthaus

Just in: Global carbon emissions hit a new all-time high in 2019, up 0.6% from last year.

This news is shockingly important and heartbreakingly serious. Not only are we entirely failing to reduce emissions, we are making the climate emergency worse at an increasingly fast rate.

4:30 PM · Dec 3, 2019 · Twitter for Android

688 Retweets   643 Likes

Tweet of the Week

Twitter, Jonathan Merritt from December 04, 2019

Lynna Odel and Bonni Stachowiak liked

Jonathan Merritt @JonathanMerritt

Can we eliminate either "effect" or "affect?" Doesn't matter which. We only need one honestly, and I'm tired of making a mistake.

7:48 PM · Dec 4, 2019 · Twitter for iPhone

38 Retweets   530 Likes

EVENTS

Observable Community Meetup

San Francisco, CA December 10, starting at 5:30 p.m., 215 2nd St. [rsvp required]

Book Talk | The Code: Silicon Valley and the Remaking of America

New York, NY December 11, starting at 6 p.m. "**NYU's Institute for Public Knowledge** invites you to join for a book talk for *The Code: Silicon Valley and the Remaking of America*, featuring the author **Margaret O'Mara** in conversation with **Meredith Broussard**." [rsvp required]

NYC UX + Data Meetup: Finding Inspiration in Historic Data Visualizations

New York, NY December 11, starting at 6:45 p.m., **General Assembly** (902 Broadway, 4th Floor). Speaker: **Jason Forrest**. [$$]

Berkeley Initiative for Transparency in the Social Sciences (BITSS) Annual Meeting

Berkeley, CA December 13, starting at 9:30 a.m., **Haas School of Business**. "The event brings together stakeholders from academia, scholarly publishing, and policy to share knowledge and discuss the evolving movement toward research transparency and efforts to strengthen the standards of openness and integrity in the social sciences." [free, rsvp required]

Nicholas A. Christakis, "Social Network Interventions"

Boston, MA December 16, starting at 4 p.m. "**The Boston Children's Hospital Computational Health Informatics Program** (CHIP) invites you to our Landmark Ideas Series Lecture!"

CONNECT: The Symposium of Wu Tsai Neuro and ChEM-H

Stanford, CA February 11, starting at 8:30 a.m., **Stanford University**, Li Ka Shing Center and Paul Berg Hall. "**Stanford ChEM-H** and **Wu Tsai Neurosciences Institute** will host a joint scientific symposium on the occasion of the official dedication of the new ChEM-H and Neurosciences Buildings." [registration required]

The 5th Annual Scaled Machine Learning Conference

Mountain View, CA February 26, 2020, starting at 9 a.m. "The creators of TensorFlow, Kubernetes, Apache Spark, Keras, Horovod, Allen AI, Apache Arrow, MLPerf, OpenAI, Matroid, and others will lead discussions about running and scaling machine learning algorithms on a variety of computing platforms." [$$$]

Introducing Customize, the Food Personalization Summit

New York, NY February 27. "A one-day executive summit that will explore the impact of personalization across the food system." [$$$]

PyCon US

Pittsburgh, PA April 15-23, 2020. Registration is now open. [$$$]

Southern Data Science Conference

Atlanta, GA April 16-17, 2020. "If you are a data science enthusiast who is interested in learning from the experts and top researchers about the cutting-edge technologies and best practices in data science, don't hesitate to register." [$$$]

Artificial Intelligence and Games

Copenhagen, Denmark June 22-26, 2020. "The summer school is dedicated to the uses of artificial intelligence (AI) techniques in and for games. After introductory lectures that explain the background and key techniques in AI and games, the school will introduce participants the uses of AI for playing games, for generating content for games, and for modeling players." [save the date]

Who We Are - Visualizing NYC by the Numbers

New York, NY Through September 20, 2020 at **Museum of the City of New York**. "In anticipation of the 2020 census, Who We Are: Visualizing NYC by the Numbers showcases work not just by data analysts and demographers, but also by cutting-edge contemporary artists and designers who use these tools to enliven and humanize statistics and to shed new light on how we understand our urban

environment and ourselves." [$$]

DEADLINES
Contests/Award
HHMI Investigator Program Opens National Competition
"The **Howard Hughes Medical Institute** (HHMI) is seeking to appoint approximately 20 new Investigators through a national open competition. This effort will expand HHMI's community of nearly 300 Investigators, who perform basic biological research across the nation." Deadline to apply is March 18, 2020.

Conferences
Inaugural O'Reilly Infrastructure & Ops Conference - Call for speakers
Santa Clara, CA June 15-18, 2020. "Be part of the inaugural **O'Reilly Infrastructure & Ops Conference** by leading a talk about your greatest systems success or failure. If 50 minutes isn't enough time, dive deeper by teaching a three-hour tutorial on your favorite topic." Deadline for proposals is December 10.

The First International Workshop on Literature-Based Discovery - Call for Papers
Singapore May 11-14. "Literature-based discovery (LBD) addresses a key knowledge discovery challenge of the Big Data era: How to generate novel and actionable knowledge from vast, diverse, and seemingly disconnected fragments of information. Incorporating state-of-the-art data science, natural language processing, and network science techniques." Deadline for submissions is January 24, 2020.

Education Opportunities
Facebook AI Residency Program
"The AI Residency program will pair you with an AI Researcher and Engineer who will both guide your project. With the team, you will pick a research problem of mutual interest and then devise new deep learning techniques to solve it. We also encourage collaborations beyond the assigned mentors." Deadline to apply is January 31, 2020.

Open Call for Fellowship Applications - Berkman Klein Center for Internet & Society at Harvard University
"This opportunity is for those who wish to spend 2020-2021 in residence in Cambridge, MA as part of the Center's vibrant community of research and practice, and who seek to engage in collaborative, cross-disciplinary, and cross-sectoral exploration of some of the Internet's most important and compelling issues." Deadline for applications is January 31, 2020.

Now Open: Applications for CCST's Science & Technology Policy Fellowship in Sacramento
The **California Council on Science and Technology** "is seeking PhD scientists and engineers for a year of public service and government leadership training in the California State Legislature AND California state agencies." Deadline to apply is March 1, 2020.

Studies/Surveys
U.S. Citizenship and Immigration Services Fee Schedule and Changes to Certain Other Immigration Benefit Request Requirements
"The Department of Homeland Security (DHS) proposes to adjust certain immigration and

naturalization benefit request fees charged by U.S. Citizenship and Immigration Services (USCIS). USCIS conducted a comprehensive biennial fee review and determined that current fees do not recover the full costs of providing adjudication and naturalization services. DHS proposes to adjust USCIS fees by a weighted average increase of 21 percent, add new fees for certain benefit requests, establish multiple fees for petitions for nonimmigrant workers, and limit the number of beneficiaries on certain forms to ensure that USCIS has the resources it needs to provide adequate service to applicants and petitioners." Deadline to provide comments is December 16.

The Carpentries Long-Term Impact Survey is Open

"We are officially opening our Long-Term Impact Survey and would like anyone who has taken a Data, Software, or Library Carpentry workshop between April 2018 and December 2019 to share their story by completing the survey." Deadline to participate is December 20.

TOOLS & RESOURCES

Princeton IoT Inspector

Danny Huang et al. from April 07, 2019

Automatically discovers IoT devices and analyzes their network traffic

Helps you identify security and privacy issues with graphs and tables

Want to Truly Master Scikit-Learn? 2 Essential Tips from Core Developer Himself

YouTube, Towards Data Science from November 23, 2019

Towards Data Science "interviews Andreas Muller from Scikit-learn who shares 3 practical machine learning techniques all data scientists needs to know." [video, 24:43]

Alibaba makes its artificial intelligence platform open source

Data Economy, Antony Savvas from November 28, 2019

"The move to **GitHub** will make lives easier for big data specialists in the machine learning sector, and no doubt provide some useful ecosystem links to the Chinese giant."

"Alibaba Cloud has made the core codes of its Alink data processing platform open-source to help widen the development opportunities for artificial intelligence and machine learning."

Django 3.0 release notes

Django from December 01, 2019

"These release notes cover the new features, as well as some backwards incompatible changes you'll want to be aware of when upgrading from Django 2.2 or earlier."

Amazon CodeGuru (Preview)

Amazon Web Services from December 03, 2019

"**Amazon** CodeGuru is a machine learning service for automated code reviews and application performance recommendations. It helps you find the most expensive lines of code that hurt application performance and keep you up all night troubleshooting, then gives you specific recommendations to fix or improve your code."

SFU global collaboration creates world's first open-source database of natural microbial products

Simon Fraser University, SFU News from December 05, 2019

"**Simon Fraser University** associate professor **Roger Linington** and a team of international collaborators have created the Natural Products Atlas, which holds information on nearly 25,000 natural compounds and serves as a knowledge base and repository for the global scientific

community."

Open-Sourcing Metaflow, a Human-Centric Framework for Data Science

Medium, Netflix Tech Blog; David Berg, Ravi Kiran Chirravuri, Romain Cledat, Savin Goyal, Ferras Hamad, Ville Tuulos from December 03, 2019

"Data scientists want to retain their freedom to use arbitrary, idiomatic Python code to express their business logic — like they would do in a Jupyter notebook. However, they don't want to spend too much time thinking about object hierarchies, packaging issues, or dealing with obscure APIs unrelated to their work. The infrastructure should allow them to exercise their freedom as data scientists but it should provide enough guardrails and scaffolding, so they don't have to worry about software architecture too much."

Better wildfire and smoke predictions with new vegetation database

University of Washington, UW News from December 04, 2019

"Researchers from the **University of Washington** and **Michigan Technological University** have created the first comprehensive database of all the wildfire fuels that have been measured across North America. Called the North American Wildland Fuel Database, the tool incorporates the best available measurements of vegetation in specific locations, and allows fire managers to see where information about fuels is missing altogether."

CAREERS
Tenured and tenure track faculty positions
Professor of Data Science and Mathematics
New York University, Center for Data Science and Courant Institute of Mathematical Sciences; New York, NY

Open Rank Joint Faculty Position
New York University, Center for Data Science and Center for Neural Science; New York, NY

Assistant / Associate Professor - Biostatistics
New York University, Center for Data Science and Department of Biostatistics; New York, NY

Tenure-Track
University of Pittsburgh, The School of Computing and Information (SCI); Pittsburgh, PA

Tenure Track Software Engineering & Data Science Faculty (Open-Rank)
University of St. Thomas, School of Engineering; St. Paul, MN

Full-time, non-tenured academic positions
Data Scientist, Media Manipulation
Harvard University, Kennedy School of Government; Cambridge, MA

Senior Software Developer
Indiana University, Cyberinfrastructure for Network Science Center; Bloomington, IN

Research Scientist (Open Rank) - Foundational; Domain Applications; Open source Software Development
Columbia University, Data Science Institute; New York, NY

Research Scientist (Open Rank) - Research Collaboration; Software Engineering
Columbia University, Data Science Institute; New York, NY

Reference & Instruction Librarian

University of Maryland-Baltimore County; Baltimore, MD

Appointment Stream, Teaching Faculty Positions

University of Pittsburgh, The School of Computing and Information (SCI); Pittsburgh, PA

Postdocs

HEP/Astro Experiment

Brandeis University, School of Physics; Waltham, MA

Post-Doctoral Research Fellow/Scientist

Columbia University, Data Science Institute; New York, NY

Postdoctoral Scholar

University of Washington, Center for an Informed Public (CIP); Seattle, WA

Theoretical Computer Science Postdocs

University of Toronto, Department of Computer Science; Toronto, ON, Canada

Post Doctoral Fellow - Harden

University of Texas, Population Research Center; Austin, TX

Moore-Sloan Faculty Fellow.

New York University, Center for Data Science; New York, NY

Postdoctoral Scholar - Economic Modeling

University of California-Santa Barbara, Environmental Market Solutions Lab; Santa Barbara, CA

ETH- FDS Postdoctoral Fellow

ETH Zurich, ETH Foundations of Data Science; Zurich, Switzerland

Postdoctoral positions

Max Planck Institute for Social Anthropology; Halle, Germany

Post Doctoral Fellow in Data Science and Population Informatic

Texas A&M University, Population Informatics Lab; College Station, TX

Full-time positions outside academia

Matplotlib Research Software Engineering Fellow

NumFOCUS; Austin, TX

Deputy Director of Instructor Training

The Carpentries; Oakland, CA

Research Software Engineer: Infrastructure and Development Operations

EcoHealth Alliance; New York, NY

Senior Machine Learning Engineer, NLP

Samsung NEXT; New York, NY

Vice President for Scholarly Programs

National Humanities Center; Research Triangle Park, NC

Bayesian Statistician, Research & Development

Houston Astros; Houston, TX

Data Scientist

American Civil Liberties Union, National Office; New York, NY

Associate Director

ASAPbio; Greater Boston, MA

Internships and other temporary positions

2020 Summer Football Data Master's Intern

    National Football League; New York, NY

MSc Project: Understanding Arctic Coastal Ecosystems

    Carleton University, Department of Biology; Ottawa, ON, Canada

Google Research Intern, 2020

    Google; Research Lab locations across the U.S.

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.

To send us an announcement for the newsletter, please email laura.noren@nyu.edu and brad.stenger@gmail.com. We retain curatorial discretion.

Data Science Community Newsletter Issue 186.