NYU Data Science Community Newsletter features journalism, research papers, events, tools/software, and jobs for August 5, 2016

Please let us (Laura Noren, Brad Stenger) know if you have something to add to next week's newsletter. We are grateful for the generous financial support from the Moore-Sloan Data Science Environment and to NYU's Center for Data Science.

NEJM, ICMJE go head-to-head over clinical trial data-sharing rules

The New England Journal of Medicine (NEJM) and the International Committee of Medical Journal Editors (ICMJE) are having a heated disagreement about the best practices for sharing randomized clinical trial (RCT) data. Last January, the ICMJE proposed to

"require authors to share with others the deidentified individual-patient data (IPD) underlying the results presented in the article (including tables, figures, and appendices or supplementary material) no later than 6 months after publication."
Commenters from the disciplines represented by the journals frequently pointed out that the six month embargo is too long; data should be available at the time of publication to increase the likelihood that confirmatory studies would arrive quickly for the sake of patients.

But…

Just this week, NEJM responded and so did Senator Elizabeth Warren, JD. NEJM opposed the six month embargo. They instead suggest data should be shared,

"a minimum of 2 years after the first publication of the results and an additional 6 months for every year required to complete the study, up to a maximum of 5 years"
According to NEJM, this would mean a data sharing embargo of 2.5 years for a typical small trial and 5 years for a large trial. They also recommended that,
"persons who were not involved in an investigator-initiated trial but want access to the data should financially compensate the original investigators for their efforts and investments in the trial and the costs of making the data available."
Commenter **Roger Bumgarner** growled that
"as long as medical investigators continue to prioritize publication rights over what is best for the patients, progress will be slowed"
and scientist **Bjorn Brembs** asked if NEJM is
"proposing that dying patients should wait a few years such that medical researchers can climb the career ladder?"
Fighting words are also cascading across twitter today.

Getting back to the issue of financial incentives, but taking the side of open data advocates, **Senator Warren** wrote,

"widespread practices of data sharing can also help to address concerns about conflicts of interest that may arise when clinical trials are funded by industry sponsors that stand to profit from favorable research results. By making trial results available for independent scrutiny by outside reviewers, data sharing makes it less likely that trial sponsors can buy the analysis and results they want."
She also noted that standardizing the rules for data sharing would increase uniformity of data-sharing expectations across funding agencies.

University of Pennsylvania data scientist **Daniel Himmelstein**, creator of a public biology data-sharing portal Hetionet, would have preferred more legal clarity before recently launching his platform. While lawsuits against academics for sharing information are rare, they do happen (e.g. Aaron Swartz). Lack of legal and normative clarity has a chilling effect because, as Himmelstein found,

"most academics aren't in the position to risk setting off a legal battle"
We are, however, willing to take a public stance in comments, tweets, and GitHub releases, which is a good start.

Data Science News

How a happy moment for neuroscience is a sad moment for science

Medium, Mark Humphries from August 01, 2016
The **Allen Institute for Brain Science** released a landmark set of data in June. Entitled the “Allen Brain Observatory”, it contains a vast array of recordings from the bit of cortex that deals with vision, while the eyes attached to that bit of cortex were looking at patterns. Not too exciting, you say. In some respects you’d be right: some mouse brain cells became active when shown some frankly boring pictures. Experimental neuroscience is eternally lucky that mice have a very high boredom threshold.

The release of this data took a privately funded institute. It could not have come from a publicly-funded scientist. It is a striking case-study in how modern science is worryingly broken, because it prioritises private achievement over the public good.

[1607.08237] The population of long-period transiting exoplanets

arXiv, Astrophysics > Earth and Planetary Astrophysics; Daniel Foreman-Mackey, Timothy D. Morton, David W. Hogg, Eric Agol, Bernhard Schölkopf from July 27, 2016
The Kepler Mission has discovered thousands of exoplanets and revolutionized our understanding of their population. This large, homogeneous catalog of discoveries has enabled rigorous studies of the occurrence rate of exoplanets and planetary systems as a function of their physical properties. ... we perform a fully automated search for long-period exoplanets with only one or two transits in the archival Kepler light curves. When applied to the 40,000 brightest Sun-like target stars, this search produces 16 long-period exoplanet candidates. Of these candidates, 6 are novel discoveries and 5 are in systems with inner short-period transiting planets.

More exoplanets:
- Astronomers have released a list of list of the 20 most “Earth-like” planets (August 05, Wired UK)
- What Does the Universe Do When We're Not Looking? (July 19, Universe Today, **Fraser Cain**)

Northeastern University's College of Computer & Info. Sci. launches 6 new combined majors programs

Northeastern University from July 29, 2016
Six new undergraduate combined majors were announced.
- CS and Criminal Justice
- CS and English
- CS and History
- CS and Philosophy
- CS and Sociology
- CS and Design

Also in new data science courses at Boston universities:
- MIT's new online course addresses data science (MIT News, 4 August 2016)

[1607.07403] On the Ubiquity of Web Tracking: Insights from a Billion-Page Web Crawl

arXiv, Computer Science > Social and Information Networks; Sebastian Schelter, Jérôme Kunegis from July 25, 2016
We perform a large-scale analysis of third-party trackers on the World Wide Web from more than 3.5 billion web pages of the CommonCrawl 2012 corpus. We extract a dataset containing more than 140 million third-party embeddings in over 41 million domains. To the best of our knowledge, this constitutes the largest web tracking

dataset collected so far ... we confirm that trackers are widespread (as expected), and that a small number of trackers dominates the web (**Google**, **Facebook** and **Twitter**). In particular, the three tracking domains with the highest PageRank are all owned by Google. The only exception to this pattern are a few countries such as China and Russia.

Also in tracking + data:

- Universities are tracking their students. Is it clever or creepy? (August 03, The Guardian, **Chris Jutting**)
- Publishers' Dilemma: Judge A Book By Its Data Or Trust The Editor's Gut? (August 02, NPR, All Tech Considered)

Epigenetics and aging

Science Advances; Sangita Pal and Jessica K. Tyler from July 29, 2016
Over the past decade, a growing number of studies have revealed that progressive changes to epigenetic information accompany aging in both dividing and nondividing cells. ... Several important conclusions emerge from these studies: rather than being genetically predetermined, our life span is largely epigenetically determined; diet and other environmental influences can influence our life span by changing the epigenetic information; and inhibitors of epigenetic enzymes can influence life span of model organisms. These new findings provide better understanding of the mechanisms involved in aging.

This AI Will Craft Tweets That You'll Never Know Are Spam

MIT Technology Review, Tom Simonite from August 04, 2016
Industry researchers trained machine-learning software to write tweets like a human to reply to some people using the hashtag #Pokemon, in a demonstration of how advances in software that understands language could be used to trick people online. Roughly a third of people targeted by the software clicked on a benign link sent along by the software to test how convincing it was.

More AI headlines:

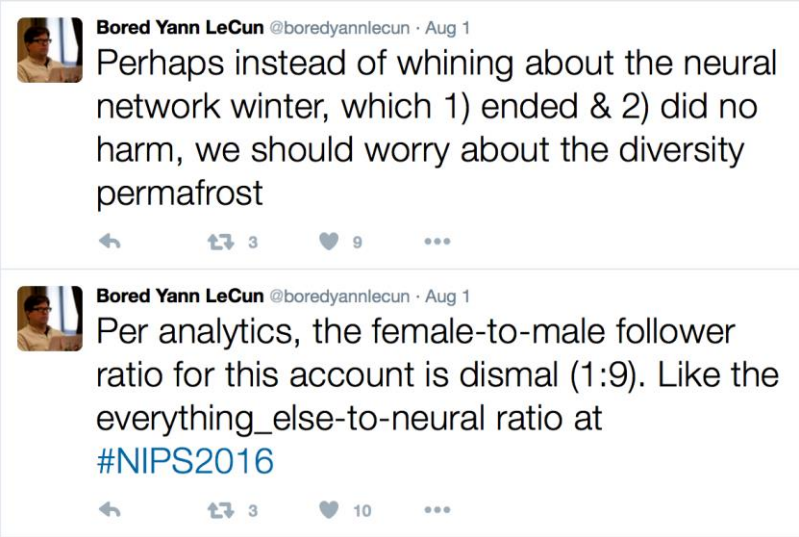
- OpenAI Is Calling for Techie Cops to Battle Code Gone Rogue (August 02, WIRED, Business)
- Machines v. hackers: Cybersecurity's artificial intelligence future (July 25, Christian Science Monitor, CSMonitor.com)
- Make Algorithms Accountable (August 01, The New York Times, Opinion, **Julia Angwin**)
- How To Fool AI Into Seeing Something That Isn't There (July 29, WIRED, Security)

Moore Diversity

Medium, Moore Data, Carly Strasser from August 03, 2016
Diversity has been on my mind a lot lately. It's a vague phrase (like "sustainability" or "governance" or "best practices")—but it's one that I keep finding myself chatting about with our grantees, fellow conference attendees, and other **Moore Foundation** personnel. I won't attempt to cover the gamut of diversity-related topics in this post, however I want to get a few of my more recent thoughts out; this blog post will focus on (primarily gender) diversity at conferences.

Tweet of the Week

Twitter, Bored Yann LeCun from August 01, 2016



Events

HAMR | ISMIR 2016

HAMR@ISMIR 2016 will provide a space for individuals from various institutions, backgrounds, and experience levels to test out novel ideas as opposed to finishing a polished project and paper.

New York, NY Friday, August 5 at **Spotify NYC** (45 W 18th St., 3rd Floor) starting at 7 p.m. [Tonight!]

12th International Workshop on Mining and Learning with Graphs (MLG 2016)

San Francisco, CA Held in conjunction with KDD'16 on Sunday, August 14. [$$$]

Data Science for Social Good Conference

This conference will highlight the successes, opportunities, and challenges faced by the growing Data Science for Social Good community by bringing key members from each community (academia, governments, non-profits, foundations, social enterprises, and corporations) together to share best practices, learn from each other, and generate new collaboration opportunities.

Chicago, IL Wednesday-Thursday, August 24-25, at the University of Chicago Gleacher Center. [$$]

South Hub Sponsors a Fall Workshop Series

The South Hub is sponsoring several workshops this fall, and registration for these events is now open. Participants come from academic research institutions across the 16 states that comprise the **South Big Data Innovation Hub** and industrial partners across the country.

Atlanta, GA The first workshop, “Data Infrastructure for Materials and Advanced Manufacturing Workshop” will be held at Georgia Tech on August 25. [travel awards available]

Open Data Maker's Hackathon

During the GODAN Summit 2016 happening September 15-16, 2016, **GODAN** [Global Open Data for Agriculture & Nutrition] is hosting the Open Data Maker’s Hackathon, bringing together software and agricultural innovators to create the beginnings of practical solutions that allow for better utilizing, collecting, and/or making more accessible open data to improve our food system. In an effort to support young, next generation innovators, this hackathon is intended for current university students and/or entrepreneurs 26 years old and younger.

New York, NY Thursday-Friday, September 15-16, at the New York Hilton Midtown.

Plotcon 2016 - Speakers and topics in R

New York, NY The world’s most visionary conference for data visualization in scientific computing, finance, business, and journalism. Tuesday-Friday, November 15-18, at 55 Broadway. [$$$]

Deadlines

Call for Papers - The Conference on Digital Experimentation @ MIT

deadline: Conference

Cambridge, MA The purpose of the Conference on Digital Experimentation at **MIT** (CODE Conference) is to bring together leading researchers conducting and analyzing large scale randomized experiments in digitally mediated social and economic environments, in various scientific disciplines including economics, computer science and sociology, in order to lay the foundation for ongoing relationships and to build a lasting multidisciplinary research community.

The deadline for paper submissions is Friday, August 12.

International Prize in Statistics: Submitting Nominations

deadline: Contest/Award

The biannual International Prize in Statistics is stewarded and managed by a foundation comprising representatives of the five major statistical organizations working cooperatively to develop this prestigious award: the **American Statistical Association**, **Institute of Mathematical Statistics**, **International Biometric Society**, **International Statistical Institute**, and **Royal Statistical Society**.

Mirroring the successful approach employed by other prestigious scientific prizes, the International Prize in Statistics recognizes an individual statistician or team of statisticians (groups of individuals working on similar ideas as teams of individuals or organizations) for “a single work or body of work.”

Deadline for nominations in Monday, August 15.

NYC Digital Humanities - Third Annual Graduate Student Project Award

deadline: Education Opportunity

We are pleased to announce our third annual cross-institutional NYCDH digital

humanities graduate student project award. We invite all graduate students attending an institution in New York City and the metropolitan area to apply.

Deadline to apply is Monday, August 15.

The Lightning Challenge: Doctoral Students

deadline: Contest/Award

Applicants to participate in The Lightning Challenge: **NYU** Doctoral Students are asked to make and upload a 3 minute video of themselves talking about their dissertation research. This video and the accompanying application will be evaluated by the selection committee, who will choose Lightning Challenge participants from those who apply.

Applications are due by Monday, August 15, 2016.

Workshop on Data and Algorithmic Transparency

deadline: Conference

New York, NY The Workshop on Data and Algorithmic Transparency (DAT'16) is being organized as a forum for academics, industry practicioners, regulators, and policy makers to come together and discuss issues related to increasing role that "big data" algorithms play in our society. It will be at Columbia University on Saturday, November 19.

Deadline for submissions is Friday, September 9.

CDS News

AI NexusLab

NYU Tandon School of Engineering, NYU Future Labs from July 28, 2016
AI NexusLab is a four-month program run by the **NYU Future Labs** to support AI companies' going from ideation and MVP and product-market fit. The AI NexusLab will recruit the top AI startups from across the world to come to NYC for a four-month program. Companies will receive $100k to join the lab, and will gain access to two full-time technical experts, a network of mentors including NYU AI faculty experts, abundant resources, and a rigorous program to guide startups to market entry.

Should Children be Taught Computer Science?

NYU Center for Data Science from July 29, 2016
As digital technologies become increasingly entrenched into our world, the conversation surrounding when and how to teach computer science is becoming increasingly important. Should computer science courses be requisite the way that math courses are? And if so, at what age would we start educating kids about computer science, and in what capacity?

To find out more about the current conversation surrounding computer science in childhood education, we spoke with two professionals in the field.

Tools & Resources

JSM2016slides: Links to slides for talks at the 2016 Joint Statistical Meetings in Chicago

GitHub - kbroman from August 01, 2016
Pull requests welcome! Or add an issue, or tweet @kwbroman.

Probabilistic data structures in Python

O'Reilly Media, Paco Nathan from August 01, 2016
This tutorial is intended for a Python programmer who has some background working with big data, who now needs to learn how to apply probabilistic data structures for analytics with large-scale data and streaming applications, and especially for use cases that require both. [video, 27:31]

A Survey of Deep Learning Techniques Applied to Trading

Greg Harris from May 30, 2016
Deep learning has been getting a lot of attention lately with breakthroughs in image classification and speech recognition. However, its application to finance doesn't yet seem to be commonplace. This survey covers what I've found so far that is relevant to systematic trading.

Library for fast text representation and classification.

GitHub - facebookresearch from August 04, 2016
fastText is a library for efficient learning of word representations and sentence classification. Requires Python 2.6 or newer.

data.world

data.world from July 11, 2016
At least 18M open datasets exist today. Only 2.4M websites existed at **Google's** launch in 1998. Open data holds great promise for society, and we think its impact should grow as quickly as its volume. We've built a platform where the world's problem solvers can find and use a vast array of high-quality open data.

Careers

Specialist, SBC LTER Information Manager - University of California-Santa Barbara

University of California-Santa Barbara from August 01, 2016
Santa Barbara, CA The **UCSB Marine Science Institute** seeks a Scientific Information Manager to support and promote the scientific mission of the Santa Barbara Coastal Long-term Ecological Research (SBC LTER) project, part of an NSF-sponsored program engaged in interdisciplinary research on population, community and ecosystem level processes in the context of environmental change.

John Derby Evans Professorships in Media Technology

University of Michigan School of Information from August 01, 2016
Ann Arbor, MI The **School of Information** and the **Department of Communication Studies in the College of Literature, Science, and the Arts** at the **University of Michigan** invite applications for two tenured faculty positions at the Associate Professor rank focusing on the social implications of digital media. ... These positions together constitute a cross-disciplinary cluster hire in the area of "digital futures."

Penn State: Director for the Institute for CyberScience

KDnuggets from July 28, 2016

State College, PA The successful candidate will be an accomplished scholar with the vision and leadership to direct the **Institute for CyberScience** and its advanced computing resources. The ICS is one of five university-wide research institutes that are centrally positioned within the Office of the Vice President for Research to accelerate discovery and advance interdisciplinary, collaborative team science, across the University.

CSHL Faculty Positions

Cold Spring Harbor Laboratory from August 05, 2016
Cold Spring Harbor, NY Cold Spring Harbor Laboratory (CSHL) is searching at the Assistant Professor level for highly talented individuals to join its **Simons Center for Quantitative Biology**.

We're Hiring a new Dash Service Manager!

California Digital Library from August 03, 2016
Oakland, CA California Digital Library is recruiting for a new UC3 Service Manager. This position will oversee the product management and outreach activities for the Dash project and service, as well as offer research data management and digital preservation consulting for the UC community.