NYU Data Science Community features journalism, research papers, events, tools/software, and jobs for April 22, 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.

Data Science News

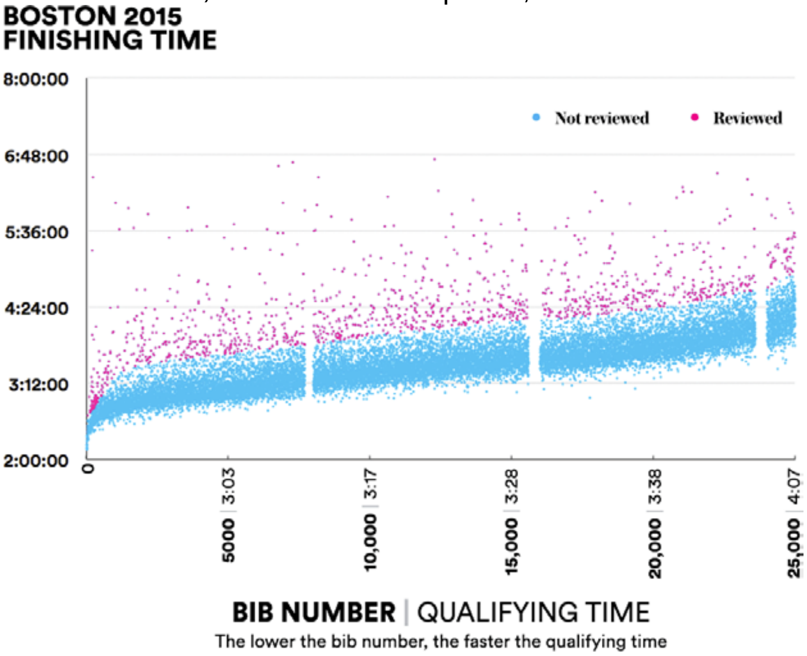
No pressure: NSF test finds eliminating deadlines halves number of grant proposals

Science, ScienceInsider from April 15, 2016
In recent years, the **National Science Foundation** in Arlington, Virginia, has struggled with the logistics of evaluating a rising number of grant proposals that has propelled funding rates to historic lows. Annual or semiannual grant deadlines lead to enormous spikes in submissions, which in turn cause headaches for the program managers who have to organize merit review panels. Now, one piece of the agency has found a potentially powerful new tool to flatten the spikes and cut the number of proposals: It can simply eliminate deadlines.

This week, at an NSF geosciences advisory committee meeting, Assistant Director for Geosciences **Roger Wakimoto** revealed the preliminary results from a pilot program that got rid of grant proposal deadlines in favor of an anytime submission. The numbers were staggering. Across four grant programs, proposals dropped by 59% after deadlines were eliminated.

Dozens Suspected of Cheating to Enter Boston Marathon

Runner's World, Newswire from April 13, 2016



[Derek] Murphy started a blog about race cheating, and last summer with three colleagues began pursuing their ambitious project to figure out how many people got into Boston by breaking race rules. Their first step was to cull the entire list of finishers into a smaller group of likely suspects.

Are Algorithms Ruining How We Discover Music?

FiveThirtyEight from April 21, 2016
On this week's episode of our podcast What's The Point, New York Times jazz and pop critic **Ben Ratliff** discusses his new book *Every Song Ever* and how the everything-at-our-fingertips era is changing the way we listen to music. [audio, 36:04]

Also, this August in NYC:

- 17th International Society for Music Information Retrieval Conference (August 7-11, organized by **NYU** and **Columbia University**)

AllTransit Maps and Visualizes a Nationwide Transit Database, the Most Exhaustive and Accessible Yet

CityLab, Laura Bliss from April 19, 2016
As the social and economic benefits of transit become clearer and clearer, a parade of data-driven maps and websites have tried to evaluate transit access in major American cities: where buses and trains go, who they serve, how effectively, and how often.

Tuesday marks the launch of AllTransit, the most exhaustive and accessible such resource yet. A joint project of the **Center for Neighborhood Technology** and **TransitCenter**, it assembles the largest collection of transit data anywhere—543,000 transit stops, 800 transit agencies, and 15,000 routes nationwide, according to the site. That in itself is a major public service, since agencies aren't (as of yet) required by the DOT to open up their data about connectivity, access, and frequency. AllTransit doesn't offer that data raw (not for free, at least), but it does offer a number of useful ways to explore it.

How big data is helping us understand mental illness

Wired UK from April 19, 2016
Mental health apps have had a tough time of late. Studies from the American Psychiatric Association's Smartphone App Evaluation Task Force and the University of Liverpool have found that, despite a surfeit of mental healthcare apps available online, many lack an "underlying evidence base, a lack of scientific credibility and limited clinical effectiveness".

Not so for Big White Wall. The company, founded by entrepreneur Jen Hyatt in 2007, is a mental healthcare tool that uses data and rigorous clinical governance to provide a service that is both free and effective. The service was recently highlighted as one of the few NHS mandated apps proven to be clinically effective.

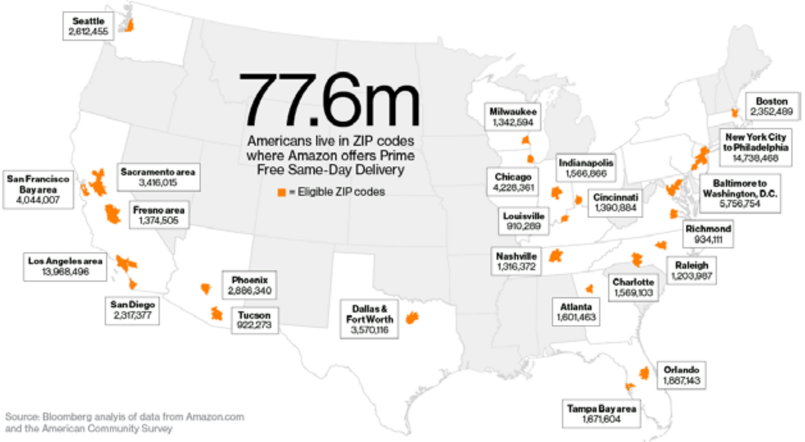
Data Science Research Grants: Announcing Our Third Round of Winners

Bloomberg L.P. from April 19, 2016
The Bloomberg Data Science Research Grant Program aims to support cutting-edge research in the broad field of machine learning, including specific areas such as

natural language processing, information retrieval, machine-translation and deep neural networks. In April 2015 we announced our first round of recipients and in October 2015 we announced our second. Today, we are pleased to announce the winners of our third round of awards.

Amazon Doesn’t Consider the Race of Its Customers. Should It?

Bloomberg from April 21, 2016



Source: Bloomberg analyis of data from Amazon.com and the American Community Survey

As **Amazon** has expanded rapidly to become “the everything store,” it’s offered the promise of an egalitarian shopping experience. On Amazon and other online retailers, a black customer isn’t viewed with suspicion, much less followed around by store security. Most of Amazon’s services are available to almost every address in the U.S. “We don’t know what you look like when you come into our store, which is vastly different than physical retail,” says **Craig Berman**, Amazon’s vice president for global communications. “We are ridiculously prideful about that. We offer every customer the same price. It doesn’t matter where you live.”

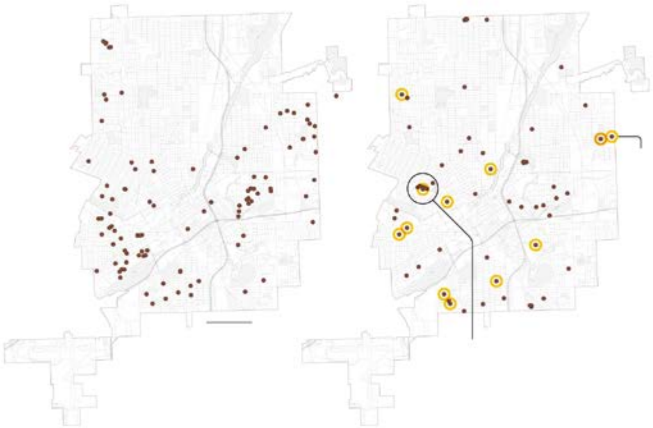
Beyond the Lab: Ethan White

Gordon and Betty Moore Foundation from April 15, 2016

Ethan White, Ph.D., is an investigator in the foundation’s Data-Driven Discovery initiative and directs the Quantitative Macroecology Lab at the **University of Florida**. ... In our first installment of *Beyond the Lab*, Ethan discusses his work in bringing ecology to the data-intensive era.

How Officials Distorted Flint’s Water Testing

The New York Times from April 21, 2016



Local and state officials claimed for months that tests showed that Flint's water had safe levels of lead. But the officials used flawed testing methods, making the levels of lead in the water supply appear far less dangerous than they were.

Three of those officials were charged with crimes on Wednesday, accused of covering up glaring deficiencies in two rounds of lead testing conducted in 2014 and 2015.

Metadata for research data management

OCLC Research, Hanging Together blog from April 18, 2016
That was the topic discussed recently by **OCLC Research** Library Partners Research metadata managers, initiated by **John Riemer** of UCLA. With increasing expectations that research data creation made possible through grant funding will be archived and made available to others, many institutions are becoming aware of the need to collect and curate this new scholarly resource. To maximize the chances that metadata for research data are shareable (that is, sufficiently comparable) and helpful to those considering re-using the data, our communities would benefit from sharing ideas and discussing plans to meet emerging discovery needs.

IRCS will be closing on June 30, 2016, as the SAS and SEAS leadership realigns initiatives with the strategic plans of the two schools.

University of Pennsylvania, Institute for Research in Cognitive Science from April 18, 2016
IRCS was founded just over 25 years ago in 1990, strengthening a tradition of cross-disciplinary collaboration in Cognitive Science at Penn that dates back to the early 1960s and building on a formal program in Cognitive Science that was initiated in 1978 with support from the **Alfred P. Sloan Foundation**.

Also, in the academic research circle of life:

- Final OK for Science and Engineering Complex in Allston (April 14, **Harvard** Gazette)
- Awarded! New centre of excellence and research building for collective behaviour (April 18, **University of Konstanz, Department of Collective Behaviour**)

Events

Moore-Sloan Data Science Lunch Seminar Series

Wednesday, April 27 — **Sunandan Chakraborty** from **NYU Center for Data Science**

Seminar meets from 12:30 - 1:30 p.m.

The Data Science Lunch Seminar Series is an informal weekly gathering of NYU Data Science affiliated persons to discuss data science related topics. Each week there is a 30 minute presentation, over lunch (provided), with additional time for conversation and questions.

2016 NYU Tandon School of Engineering Research Expo

The highly interactive expo will showcase dozens of the most promising and exciting research projects from every academic department, as well as exhibits from the school's Center for K-12 STEM Education, which is dedicated to boosting science, technology, engineering, and math education in New York City's public schools.

Wednesday, April 27, starting at 1 p.m., in Brooklyn.

Text as Data Speaker Series

The NYU 'Text-as-Data' speaker series takes place on Thursdays from 4 – 5:30 pm in room 217, 19 West 4th St (unless otherwise noted). The series provides an opportunity for attendees to see cutting edge text-as-data work from the fields of social science, computer science and other related disciplines.

Thursday, April 28, will be **Molly Roberts** (**UCSD**) on *Summarizing the Patient Record*.

NYU Digital Humanities Project Showcase

We are pleased to announce an NYU Digital Humanities Project Showcase to be held on Friday April 29th at **NYU Center for the Humanities** (5th floor: 20, Cooper Square). This event provides a forum for faculty, staff, and students to learn about each other's work, create connections, and start new conversations. Open to an audience from both inside and outside the university, the event will feature the work of NYU's vibrant and diverse DH community. Presentations will include 10-minute project presentations and two-minute lightning talks, and we will end with a roundtable discussion devoted to identifying priorities for supporting and building the DH community at NYU.

Friday, April 29, at NYU Center for the Humanities (5th floor: 20, Cooper Square)

How Big Data Discriminates

The **NYU Politics Society**, **WagnerTech**, and **SCJR** will host a panel on how the increasing use of data and algorithms in government and public and private organizations can create disparate impact, unintentionally discriminating against underrepresented groups.

Friday, April 29, starting at 5 p.m., the Puck Building, Rudin Auditorium

Edge Tools in a Digital Age

This is the third installment in a series that probes tools for an increasingly complex and connected world.

Original thinkers **John Seely Brown** and **Ann Pendleton-Jullian** will contextualize a series of presentations exploring new methods for listening and understanding, including data mining, visualization, shifting identity frameworks, speculative design, and games. This conversation features game designer **Elan Lee**, **Chris McNaboe** of **The Carter Center**, **Terry Young** of **sparks & honey**, and former Navy SEAL Officer **Coleman Ruiz**.

Monday, May 2, starting at 5 p.m., **New York Public Library**, Stephen A. Schwarzman Building (Fifth Avenue at 42nd St.)

New York University Reproducibility Symposium 2016

Achieving reproducibility in scientific research is a laudable goal, however this has been difficult to achieve. While data and data analysis play a central role in many scientific domains, most papers specify their methods and data only informally and omit important supplemental material. High quality journals have responded to this issue by making reproducibility a requirement for publication. Understanding the challenges to reproducibility and combating them with tools and best practices is therefore of cross-disciplinary relevance.

The **Moore-Sloan Data Science Environment** at NYU is pleased to announce a symposium on reproducibility that will be held on May 3, 2016.

Tuesday, May 3, in Brooklyn at **The Center for Urban Science + Progress** 1 MetroTech Center, 19th Floor (Jacobs Room)

Modern Massive Data Sets (MMDS)

The Workshops on Algorithms for Modern Massive Data Sets (MMDS) address algorithmic and statistical challenges in modern large-scale data analysis. The goals of this series of workshops are to explore novel techniques for modeling and analyzing massive, high-dimensional, and nonlinearly-structured scientific and internet data sets; and to bring together computer scientists, statisticians, mathematicians, and data analysis practitioners to promote the cross-fertilization of ideas.

Tuesday-Friday, June 21-24, at the **University of California-Berkeley**. Deadline for early registration is Sunday, May 1.

Deadlines

SSOE - IEEE SPS Summer School on Signal Processing and Machine Learning for Big Data at University of Pittsburgh

Humans, machines and sensors collectively generate an enormous amount of data on a daily basis. The fact that much of this data is now accessible provides an opportunity to explore, analyze and extract previously unavailable and potentially highly useful information. In many cases, the volume and speed of data generation makes traditional centralized data analysis infeasible. The lack of structure, and the amount of noise and outliers emphasize the need for robust processing across heterogeneous data domains. High dimensionality makes it challenging to visualize and interpret the data. Overall, Big Data analysis presents many challenges and opportunities for current and future signal processing professionals. This Summer School is intended to provide an introduction to the current efforts to explore Big Data from a signal processing perspective. Topics will range from foundations for Big Data analysis and processing (robust statistical methods, sparse representations, numerical linear algebra, machine learning, convergence and complexity analysis) to Big Data applications (social networks, behavior and language analysis, bioinformatics, smart grid, environmental monitoring, and others).

Deadline for registration is Saturday, April 30.

Call for Papers - aaaiaiide16

The 12th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE'16) will be held in October 8-12, 2016 (in Burlingame, California, near SFO Airport). AIIDE'16 is the twelfth annual conference and the fifteenth overall event sponsored by AAAI on the topic of AI and Interactive Entertainment.

Deadline for workshop proposals is May 15. Deadline for conference papers and demo abstracts is May 27.

Submitting Papers, CSCW 2017

This year in particular CSCW would like to invite papers that make a contribution to building CSCW systems including (but not limited to) technical enablers for CSCW applications; methods and techniques for new CSCW services and applications; and evaluation of fully-built CSCW systems and lab and field settings. Authors will be able to direct such submissions to a dedicated subcommittee.

Deadline for submissions is Friday, May 27.

CFP: Studying Social Media and Digital Infrastructures: a workshop-within-a-conference

For fifty years, the Hawaii International Conference on System Sciences (HICSS) has been a home for researchers in the information, computer, and system sciences (http://www.hicss.org/). The 50th anniversary event will be held January 4-7, 2017, at the Hilton Waikoloa Village. With an eye to the exponential growth of digitalization and information networks in all aspects of human activity, HICSS has continued to expand its track on Digital and Social Media (http://www.hicss.org/#!track3/c1xcj).

Deadline for submissions is Wednesday, June 15.

CDS News

The differences between tinkering and research

Julian Togelius from April 18, 2016
Some of us have academic degrees and fancy university jobs, and publish peer-reviewed papers in prestigious journals. Let's call these people researchers. Some (many) others publish bots, hacks, experimental games or apps on blogs, web pages or Twitter while having day jobs that have little to do with their digital creative endeavors. Let's call those people tinkerers.

So what's the difference between researchers and tinkerers?

Data Science Grant Opportunity

NYU Center for Data Science from April 15, 2016
As part of the **Moore-Sloan Data Science Environment's** ongoing attempt to foster interdisciplinary research, they recently announced the NYU Data Science Seed Grant. The grant is open to NYU faculty researchers, and will fund 6-8 proposals of either up to $25,000, or up to $6,000, in addition to providing software engineering assistance from the data science incubator. Neal Beck, an Affiliated Faculty member at CDS and member of Moore-Sloan's Working Methods group, is leading the way for this funding opportunity, and answered a few questions below.

Tools & Resources

Material for Skidmore College MA 276, Statistics in Sports

GitHub - statsbylopez from February 18, 2016
This will be the homepage for code, data, and selected assignments.

Jupyter Dynamic Dashboards from Notebooks

GitHub - jupyter-incubator from April 18, 2016
Extension for Jupyter Notebook that enables the layout and presentation of grid-based dashboards from notebooks.

A (small) introduction to Boosting

Sachin Joglekar's blog from March 06, 2016
Boosting is a machine learning meta-algorithm that aims to iteratively build an ensemble of weak learners, in an attempt to generate a strong overall model.

Lets look at the highlighted parts one-by-one.

Directory of Women in Machine Learning

Women in Machine Learning from April 20, 2016
This is a (necessarily incomplete) list of women active in machine learning maintained by the **Women in Machine Learning** (WiML) organization. (See this link for more information about the annual WiML Workshop event.) If you are organizing an event

related to machine learning, this list serves as a resource for potential speakers.

Hive Plots - Linear Layout for Network Visualization - Visually Interpreting Network Structure and Content Made Possible

Martin Krzywinski from December 09, 2011
The hive plot is a rational visualization method for drawing networks. Nodes are mapped to and positioned on radially distributed linear axes — this mapping is based on network structural properties. Edges are drawn as curved links. Simple and interpretable.

The purpose of the hive plot is to establish a new baseline for visualization of large networks — a method that is both general and tunable and useful as a starting point in visually exploring network structure.

Introducing: Research Stack

Cornell Tech, Open mHealth Lab from April 15, 2016
We introduce to you ResearchStack–the first Android framework for building and designing apps for clinical studies. With funding from the **RWJF**, **Cornell Tech** and **Open mHealth**, and development by **touchlab**, the project kicked off just five months ago to develop a way for developers and researchers with existing iOS apps to easily adapt their apps for Android. There are some 1.4 billion Android users worldwide.

Careers

JSMF - Postdoctoral Fellowship Program

James S. McDonnell Foundation Postdoctoral Fellowship Award in S from June 30, 2016
The Studying Complex Systems program supports scholarship and research directed toward the development of theoretical and mathematical tools contributing to the science of complex, adaptive, nonlinear systems. While the program's emphasis is on the development and application of the theory and tools used in the study of complex research questions and not on particular fields of research per se, JSMF is particularly interested in the continued development of complex systems science, and in projects attempting to apply complex systems approaches to coherently articulated questions. ... The **James S. McDonnell Foundation** Postdoctoral Fellowship Award in Studying Complex Systems provides fellowships to students completing doctoral training in the multidisciplinary field of complex systems science. The JSMF Fellowship is intended to provide students in the final stages of completing a Ph.D. degree more leeway in identifying and securing postdoctoral training opportunities in complex systems research.

Deadline for application and support letters is Thursday, June 30.

Mayor's Office of Operations: Enterprise Data Solutions Team

New York City Mayor's Office from April 18, 2016
Put your data modeling and analytical skills into service. Improve the lives of millions of New Yorkers.

Assessing Demand for PhD Statisticians and Biostatisticians

ASA Community, Steve Pierson from April 14, 2016
We recently received a couple inquiries on the demand for PhD statisticians and biostatisticians. A new piece in Inside Higher Education (IHE), *The Shrinking Ph.D. Job Market*, also addresses the broader market for PhDs and so I thought it would be helpful to share what we know on the statistics and biostatistics PhD market. I think it is true that the IHE piece's summarizing statement—“As number of new Ph.D.s rises, the percentage of people earning a doctorate without a job waiting for them is up”—doesn't hold for statistics and biostatistics.

Postdoc position available: whale biomechanics, energetics, and the consequences of acoustic disturbance

Stanford University, Goldbogen Lab from April 17, 2016
The Goldbogen Lab at the **Hopkins Marine Station** of **Stanford University** invites applications for a postdoctoral position in whale biomechanics and energetics. The start date is flexible. The duration of the fellowship will be at least 1 year with an opportunity to extend up to 3 years with demonstrated performance and funding availability. In addition to broad interests in comparative physiology, the lab seeks candidates with a strong background in math, physics, engineering and computer science. Candidates with a strong background in the programmatic analysis of multi-sensor acoustic tag data, especially DTAGs and the detection of acoustic and kinematic signatures, are encouraged to apply.