NYU Data Science Community Newsletter features journalism, research papers, events, tools/software, and jobs for August 7, 2015.

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

NYU Data Science Community Newsletter Issue 013.

---

**Data Science News**

**Welcoming BIDS 2015 Data Science Fellows**

*Berkeley Institute for Data Science* from July 24, 2015

We are thrilled to introduce our 2015 cohort of data science fellows! With diverse research backgrounds and experiences, the new Fellows will compliment BIDS founding group in driving data science innovations and enhancing collaborations across UC Berkeley and beyond.

**UW’s “Advanced Data Science” Ph.D. option launches!**

*UW CSE News* from August 04, 2015

The [University of Washington](https://www.washington.edu)’s Graduate School has approved the creation of a Ph.D. option in “Advanced Data Science” — an initiative of UW’s $2.8 million [National Science Foundation IGERT](https://www.nsf.gov) (Integrative Graduate Education and Research) award in data science, led by UW CSE’s [Magda Balazinska](https://www.washington.edu). The goal of the option is not to educate all students in the foundations of data science, but rather to provide advanced education to the students who will push the state-of-the-art in data science methods in their domain — to educate the next generation of thought leaders who will both build and apply new methods of data science.

**The Internet of Things and the Future of Farming**

*The New York Times*, Bits blog from August 03, 2015

The Internet of Things — the vision of a world brimming with communicating sensors and digital smarts — occupies the peak of [Gartner](https://www.gartner.com)’s most recent “hype cycle.” And a report released two months ago by the [McKinsey Global Institute](https://www.mckinsey.com) laid out the potential multitrillion-dollar payoff from the emerging technology.

At a two-day workshop last week in San Jose, Calif., hosted by the [National Science Foundation](https://www.nsf.gov) and the [National Consortium for Data Science](https://www.nationalconsortium.org), a few dozen academics, corporate technologists and government officials met and mostly wrestled with the thorny technical and policy issues that must be addressed if the potential of the Internet of Things is to be realized. They were working to come up with a research agenda to make practical progress on challenges like security, privacy and standards. A glimpse of the looming security concerns came two weeks ago, when [Fiat Chrysler](https://www.fiatusa.com) recalled 1.4 million vehicles after two researchers hacked into a Jeep Cherokee and showed they could remotely control its engine, brakes and steering.
Making data count

Scientific Data from August 04, 2015

Science is built on a foundation of data. The production and publication of that data should be recognized as valuable scholarship, but data lacks an essential prerequisite for modern-day scholarly recognition—accepted metrics of significance. The scientific community has traditionally estimated the impact of a journal article by counting the number of subsequent references to it; more recently, a suite of web-based alternative metrics (‘altmetrics’) offer faster assessment and the chance capture other kinds of impact. Data can be fit into these article-centered assessment systems by proxy, via data descriptor articles in journals like Earth Systems Science Data or Scientific Data. Another approach is to apply familiar metrics directly to datasets published in online databases or repositories. Complicating matters, the same metric may mean different things with respect to articles versus datasets. A researcher can read an article online closely without downloading the PDF version, but if they view a dataset landing page without downloading the data, their level of engagement is almost certainly low. A better understanding of how to measure data impact is critical if we are to reward data creators and incentivize data publication.

Could venture capitalists use data analytics to improve success?

Fortune, Tech from August 05, 2015

Ask anyone in venture capital about their business model and they will probably tell you it’s all about the “hits.” In the VC world, a hit is a startup that makes it big, returning many multiples of a venture fund’s initial investment. Hits are great for everyone—investors, entrepreneurs, job seekers—but the problem is they don’t happen very often. William Hambrecht, a legendary venture capitalist who made early investments in Apple, Genentech, and Google, says the odds of a big hit are about one in 10. “A few others will work out, and you’re going to lose in a lot,” he says.

But what if venture capital could boost its odds to 50-50, or even two out of three? With $48 billion in VC investment in 2014, such an improvement would prevent huge amounts of money from being lost on startups that never had much of a chance of surviving the harsh competitive environment. The challenge is to identify those likely laggards well before the market rejects their idea and, perhaps more importantly, to see the big hits before anyone else. Venture capital has long relied on subjective, intuitive methods of assessing startups, but that’s changing as more firms are bringing data science and consistency into their decision-making.

The Tao of open science for ecology

ESA Online Journals, Ecosphere from July 23, 2015

The field of ecology is poised to take advantage of emerging technologies that facilitate the gathering, analyzing, and sharing of data, methods, and results. The concept of transparency at all stages of the research process, coupled with free and open access to data, code, and papers, constitutes “open science.” Despite the many benefits of an open approach to science, a number of barriers to entry exist that may prevent researchers from embracing openness in their own work. Here we describe several key shifts in mindset that underpin the transition to more open science. These shifts in mindset include thinking about data stewardship rather than data ownership, embracing transparency throughout the data life-cycle and project duration, and accepting critique in public. Though foreign and perhaps frightening at first, these changes in thinking stand to benefit the field of ecology by fostering collegiality and broadening access to data and findings. We present an overview of tools and best practices that can enable these shifts in mindset at each stage of the research process, including tools to support data management planning and reproducible analyses, strategies for soliciting constructive feedback throughout the research process, and methods of broadening access to final research products.
Creating Made-Up Data From Real Census Information Could Protect Individual Privacy
*The Atlantic* from July 30, 2015
Synthetic datasets allow researchers to study social systems without compromising individual identities—but how reliable is the information they’re using?

Data Management Threshold Concepts
*Data Ab Initio* from August 04, 2015
We’ve been going through the new ACRL “Framework for Information Literacy for Higher Education” recently at work. This document discusses ways to teach students how to search and understand information resources, framing critical skills as “threshold concepts”. While the Framework itself is interesting, I’m really intrigued by the idea of a threshold concept and wonder if there are any threshold concepts for data management.

For those unfamiliar with the term, a “threshold concept” is an idea that, once understood, completely reframes the way you view a topic. It’s like seeing a hidden image in that it’s very difficult to un-see the image afterward. Threshold concepts are so fundamental to understanding that it’s actually necessary to understand the concept in order to progress in the field.

Data scientists to CEOs: You can’t handle the truth
*Venture Beat* from August 01, 2015
Too many big data initiatives fail because companies, top to bottom, aren’t committed to the truth in analytics. Let me explain.

In January 2015, the Economist Intelligence Unit (EIU) and Teradata (full disclosure: also my employer) released the results of a major study aimed at identifying how businesses that are successful at being data-driven differ from those that are not.

Among its many findings, there were some particularly troubling, “code red” results that revealed CEOs seem to have a rosier view of a company’s analytics efforts than directors, managers, analysts, and data scientists.

**MuseumofModernArt/collection · GitHub**
*GitHub, MuseumofModernArt* from July 22, 2015
*The Museum of Modern Art (MoMA)* acquired its first artworks in 1929, the year it was established. Today, the Museum’s evolving collection contains almost 200,000 works from around the world spanning the last 150 years. The collection includes an ever-expanding range of visual expression, including painting, sculpture, printmaking, drawing, photography, architecture, design, film, and media and performance art. … At this time, the data is available in CSV format, encoded in UTF-8. While UTF-8 is the standard for multilingual character encodings, it is not correctly interpreted by Excel on a Mac. Users of Excel on a Mac can convert the UTF-8 to UTF-16 so the file can be imported correctly.

**Dat Goes Beta**
*U.S. Open Data* from July 29, 2015
After a long year of alpha testing, which started on August 19th, we are excited to announce our launch of a new, even-more-stable phase of dat. Beta starts now.
Let us know what you’re working on and how dat might work for you. We’ll even come to your lab to help you get up and running, implement features, and fix bugs in real time. That’s first-class service!

Events
Program | KDD 2015, 10-13 August 2015, Sydney
KDD 2015 is a premier conference that brings together researchers and practitioners from data mining, knowledge discovery, data analytics, and big data. KDD 2015 will be the first Australian edition of KDD, and is its second time in the Asia Pacific region.

Monday-Thursday, August 10-13, in Sydney, Australia

Dining on Data: The Future of Cooking, Eating & Running a Food Biz
This month, the Food+Tech Meetup will hear from five companies that are using data to improve everything from cooking to restaurant concept and menu development. Join us for an inside look at these companies’ business models, the innovative ways they use data and their lessons learned.

Tuesday, August 11, at 7 p.m., Redwood Studios, 55 9th Street, Brooklyn

This American Life’s Audio Hackathon
We are inviting developers, coders, designers, producers, sound designers...anyone who has skills and ideas to offer to join us. We will group you into teams, and after introductory talks by audio professionals, you will participate in a two day hackathon. Afterward, teams will present their creations to our speakers and to employees from our partners, who will evaluate and critique their work.

Saturday-Sunday, September 19-20, at ThoughtWorks, 99 Madison Ave

Registration Deadline: Saturday, August 15

PyGotham 2015 Tickets, New York
PyGotham is an eclectic Py-centric conference covering many topics. There's a diverse speaker list, and some things which will be quite different.

Saturday-Sunday, August 15-16, at AMA New York Executive Conference Center, 1601 Broadway, Times Square

Deadlines
Fellows | Mozilla Science Lab
Are you a champion of open science and open data? Mozilla is seeking researchers eager to advance openness in science and data within their institutions.

The Mozilla Fellowships for Science present a unique opportunity for researchers who want to influence the future of open science and data sharing within their communities.

Application Deadline: Friday, August 14
Esri Human Health and Climate Change App Challenge
The White House just released more than 100 health-related open data resources to keep people more informed. To reinforce our commitment and support President Obama's Climate Data Initiative, Esri announces the Esri Human Health and Climate Change App Challenge.

Submission Deadline: Friday, August 14

WORKSHOP ON INFORMATION IN NETWORKS (WIN)
WIN is a Social Networks Summit intended to foster collaboration and to build community. The increasing availability of massive networked data is revolutionizing the scientific study of a variety of phenomena in fields as diverse as Computer Science, Economics, Physics and Sociology. Yet, while many important advances have taken place in these different communities, the dialog between researchers across disciplines is only beginning. The purpose of WIN is to bring together leading researchers studying ‘information in networks’ – its distribution, its diffusion, its value, and its influence on social and economic outcomes – in order to lay the foundation for ongoing relationships and to build a lasting multidisciplinary research community.

Friday-Saturday, October 2-3, at Stern School of Business, NYU

Abstract Submission Deadline: Friday, August 21

ACM SIGSPATIAL 2015, November 3-6 in Seattle
The ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems 2015 (ACM SIGSPATIAL 2015) is the twenty-third event in a series of symposia and workshops that began in 1993 with the aim of bringing together researchers, developers, users, and practitioners in relation to novel systems based on geo-spatial data and knowledge, and fostering interdisciplinary discussions and research in all aspects of geographic information systems.

Workshops include UrbanGIS 2015 and Geo Privacy, and have paper submission deadlines that begin in mid-August.

CDS News
CDS Professor of the Year and YP Mobile Labs Chief Scientist David Rosenberg Pushes the Boundaries of Machine Learning And Predictive Analytics
NYU Center for Data Science from August 03, 2015
Recently named NYU Center for Data Science Professor of the Year, CDS Adjunct Associate Professor David Rosenberg is currently Chief Scientist of YP Mobile Labs at YP and a former Chief Scientist at Sense Networks (acquired by YP in 2014). Dr. Rosenberg specializes in machine learning, artificial intelligence, predictive analytics and statistical modeling.

Dr. Rosenberg earned his Bachelor of Science degree in mathematics from Yale University, his Master of Science degree in applied math (computer science focus) from Harvard University, and his Ph.D. in statistics from the University of California, Berkeley. During this year’s spring semester, he taught the CDS course “Machine Learning and Computational Statistics,” which he will again offer in Spring 2016.
Click here to receive the NYU Data Science Community Newsletter 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@nyu.edu by 9 pm Eastern time on Thursday evenings for inclusion in Friday’s newsletter. We retain curatorial discretion.

NYU Data Science Community Newsletter Issue 013.