NYU Data Science Community Newsletter features journalism, research papers, events, tools/software, and jobs for July 10, 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.

Data Science News
CMU experiment aspires to make Pittsburgh world’s first ‘smart city’
Pittsburgh Post-Gazette from July 09, 2015
Pittsburgh residents who dream of the day that the region’s bridges and roads can talk to their Internet-connected cars to plot the best pathway around potholes should be willing guinea pigs in Carnegie Mellon University’s latest experiment.

After claiming the title of the world’s first university with an integrated computer network in the 1980s, CMU is now gunning for the designation as frontier for the Internet of Things — a term for products with Internet connectivity built into them — by turning its Oakland campus into a living lab.

And it plans to make all of Pittsburgh part of the experiment.

We need a measured approach to metrics
Nature News & Comment from July 08, 2015
Metrics evoke a mixed reaction from the research community. A commitment to using data and evidence to inform decisions makes many of us sympathetic to, even enthusiastic about, the prospect of granular, real-time analysis of our own activities. If scientists cannot take full advantage of the possibilities of big data, then who can?

Yet we only have to look at the blunt use of metrics such as journal impact factors, h-indices and grant-income targets to be reminded of the pitfalls. Some of the most precious qualities of academic culture resist simple quantification, and individual indicators can struggle to do justice to the richness and plurality of our research. Too often, poorly designed evaluation criteria are distorting behaviour and determining careers. At their worst, metrics can contribute to what Rowan Williams, the former Archbishop of Canterbury, UK, calls a “new barbarity” in our universities. Metrics hold real power: they are constitutive of values, identities and livelihoods.

Study Suggests Google’s Ad-Targeting System May Discriminate
MIT Technology Review from July 06, 2015
Researchers say Google’s ad-targeting system sometimes makes troubling decisions based on data about gender and other personal characteristics.

Big Data: Astronomical or Genomical?
PLOS Biology from July 07, 2015
Genomics is a Big Data science and is going to get much bigger, very soon, but it is not known whether the needs of genomics will exceed other Big Data domains. Projecting to the year 2025, we compared genomics with three other major generators of Big Data:
astronomy, YouTube, and Twitter. Our estimates show that genomics is a “four-headed beast”—it is either on par with or the most demanding of the domains analyzed here in terms of data acquisition, storage, distribution, and analysis. We discuss aspects of new technologies that will need to be developed to rise up and meet the computational challenges that genomics poses for the near future. Now is the time for concerted, community-wide planning for the “genomical” challenges of the next decade.

Using math to route ambulances
ORiginals Season 1, Episode 1 - Prof. Laura McLay - YouTube
YouTube, ORiginals Research from July 06, 2015
Prof. Laura McLay, at the University of Wisconsin Madison, discusses her research using mathematics to route ambulances. [video autoplays, 5:48]

CSAIL report: Giving government special access to data poses major security risks
MIT News from July 07, 2015
In recent months, government officials in the United States, the United Kingdom, and other countries have made repeated calls for law-enforcement agencies to be able to access, upon due authorization, encrypted data to help them solve crimes.

Beyond the ethical and political implications of such an approach, though, is a more practical question: If we want to maintain the security of user information, is this sort of access even technically possible?

That was the impetus for a report — titled Keys under doormats: Mandating insecurity by requiring government access to all data and communications — published today by security experts from MIT’s Computer Science and Artificial Intelligence Lab (CSAIL), alongside other leading researchers from the U.S. and the U.K.

YouTube, ACM from July 06, 2015
Michael Stonebraker has made fundamental contributions to database systems, which are one of the critical applications of computers today and contain much of the world's important data. [video autoplays, 1:15:48]

Seven Things Cell [Phone] Data Shows About Life In Yemen
Foreign Affairs from July 06, 2015
Advanced information technologies have revolutionized the way the world works and how people conceptualize it. Massive troves of information, known as “big data,” are aggregated and shared on a daily basis, recording an array of human behaviors and interactions at an unprecedented level of granularity. One form of such data is call data records, which, while preserving the anonymity of subscribers and the privacy of content, allow researchers to track the volume of traffic, timing, and location of calls. In combination with increasingly powerful computers, such data have shed light on important questions in the developed world on topics including marketing, health care, urban planning, and environmental policy.

Call data can also help us understand violent places in the developing world that are largely inaccessible. At a time when Yemen remains highly volatile, for example, anonymous Yemeni cell phone metadata from 2010 to 2013 that include over ten million users and several hundred million calls vividly capture patterns of Yemeni daily life, as well as celebrations, religious practices, involvement in politics, and reactions to violence.

$6M for UC Berkeley and Cal Poly to expand and enhance open-source software for
scientific computing and data science

Gordon and Betty Moore Foundation from July 07, 2015

Three foundations pledged $6M over the next three years to Project Jupyter, an open-source software project that supports scientific computing and data science across a wide range of programming languages via a large, public, open and inclusive community.

Fernando Perez of University of California, Berkeley and Lawrence Berkeley National Laboratory and Brian Granger of California Polytechnic University, San Luis Obispo will lead the project at their institutions. Perez and Granger’s efforts with Project Jupyter are the result of their work developing IPython, a popular user interface for interactive computing across multiple programming languages.

With this award from the Leona M. and Harry B. Helmsley Charitable Trust, Alfred P. Sloan Foundation, and Gordon and Betty Moore Foundation, these researchers will expand and improve the capabilities of the Jupyter Notebook, a web-based platform that allows scientists, researchers and educators to combine live code, equations, narrative text and rich media into a single, interactive document.

Coders nudge Port Authority toward open data

Capital New York from July 06, 2015

The chief technology officer of the Port Authority hopes to make more of the agency’s data available to outside developers, in part inspired by a recent PATH train-focused project at a hackathon and a data science school.

Three students at the New York City Data Science Academy analyzed PATH turnstile data to produce a visualization showing the circumstances leading to morning crowding on the trains from New Jersey to New York City.

Events

PyData Conference in Seattle

Here in the Azure Machine Learning team we live and breathe two scripting languages: R and Python. We’re working on exciting products that will advance the state of the art in how data scientists employ these languages and their stacks.

Both these ecosystems are fueled by the open source community, of course, and we are looking for ways to give back. For instance, we were recently the Keystone sponsor at PyCon 2015 and have contributed to the Jupyter/IPython project prior to that. Today, I’m very happy to announce that we will be sponsoring and hosting the first PyData conference here on the Microsoft campus – and you are invited!

Friday-Sunday, July 24-26, at Microsoft Campus in Redmond, WA

Leveraging Big Data and Predictive Knowledge to Fight Disease

Drug development is entering an era of precision medicine that is centered on the analysis of massive amounts of data. The ability to integrate, interrogate, model and interpret biological, chemical, pharmacological, genomic and clinical data holistically is key to making more effective and truly personalized medicines to fight disease. Researchers are using innovative technologies and computational techniques to develop predictive knowledge for the identification of promising new treatments, new therapeutic uses for existing molecules, patients who are good candidates for particular clinical trials or treatment protocols, and
population signals of adverse drug reactions. This symposium explores the many uses of big data and predictive knowledge to guide drug development and clinical trials.

Tuesday, July 28, at 8:30 a.m., The New York Academy of Sciences

**Deadlines**

**Big Data & Society: Call for Proposals: Special theme on "Critical Data Studies"**

Critical Data Studies (CDS) is a growing field of research that focuses on the unique theoretical, ethical, and epistemological challenges posed by “Big Data.” Rather than treat Big Data as a scientifically empirical, and therefore largely neutral phenomena, CDS advocates the view that data should be seen as always-already constituted within wider data assemblages. Assemblages is a concept that helps capture the multitude of ways that already-composed data structures inflect and interact with society, its organization and functioning, and the resulting impact on individuals' daily lives. CDS questions the many assumptions about data that permeate contemporary literature on information and society by locating instances where data may be naively taken to denote objective and transparent informational entities.

Proposal Deadline: July 10, 2015 ... Proposals of 1000 words are invited for consideration and inclusion in the Special Theme for an Original Research Article, Commentary, or essay in the Early Career Research Forum section.

**NYC Media Lab | NYC Media Lab & Bloomberg Data for Good Exchange (DGX) Prize & Impact Grant Program**

Calling all students, faculty and university researchers! Are you currently working on data-driven research relating to education, government innovation, public health or climate change? Are you currently collaborating with a nonprofit / NGO on data analysis for a civic cause? Do you believe in the power of data to connect ideas for social impact?

Deadline for Submissions: Saturday, August 1

**NYCDH Second Graduate Student Project Award**

The NYCDH Steering Group are pleased to announce our second annual cross-institutional NYCDH digital humanities graduate student project award. We invite all graduate students in the city of New York to apply by August 1, 2015. First prize winner(s) will receive a cash prize of $1000. Two runner up positions will receive $500 each. All three winning proposals will have the opportunity to receive support from one or more of the many centers affiliated with NYCDH. Winners will also receive exposure on our site and through our social media outlets.

For more information about this award and how to apply: [http://nycdh.org/nycdh-sgpa-2015/](http://nycdh.org/nycdh-sgpa-2015/)

Application Deadline: Saturday, August 1

**Cognitive Computing Challenge**

... We want a solution that can not only learn how and what data to extract from data sources like spreadsheets, word processor files, and computer generated PDF files but also learn how to map the data found in these documents to specified target fields in a database. ... This is a qualifying problem which requires you to process MLS (Multiple Listing Service) data from 300 MLS training records into a database. You will be provided with a document summarizing the correct format of each field to be extracted.
Submission Deadline is Monday, January 11, 2016. Submissions began on June 8.

**CDS News**

Data Stories #56: Amanda Cox on Working With R, NYT Projects, Favorite Data Twitter, Datastori.es from June 26, 2015
Stop whatever it is that you're doing and listen to the latest @datastories, with @amandacox
http://t.co/YEg9wTRrSK

— Alberto Cairo (@albertocairo) June 26, 2015

Data Stories is a bi-weekly podcast on data visualization with Enrico Bertini (NYU Polytechnic) and Moritz Stefaner.

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 009.