NYU Data Science Community Newsletter features journalism, research papers, events, tools/software, and jobs for June 12, 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 005.

Data Science News

Sidewalk Labs, a Start-Up Created by Google, Has Bold Aims to Improve City Living


Google’s ambitions and investments have increasingly broadened beyond its digital origins in Internet search and online advertising into the arena of physical objects: self-driving cars, Internet-connected eyeglasses, smart thermostats and a biotech venture to develop life-extending treatments.

Now Google is getting into the ultimate manifestation of the messy real world: cities.

The Silicon Valley giant is starting and funding an independent company dedicated to coming up with new technologies to improve urban life.

[1506.03340] Teaching Machines to Read and Comprehend

*arxiv, Computation and Language* from June 10, 2015

Teaching machines to read natural language documents remains an elusive challenge. Machine reading systems can be tested on their ability to answer questions posed on the contents of documents that they have seen, but until now large scale training and test datasets have been missing for this type of evaluation. In this work we define a new methodology that resolves this bottleneck and provides large scale supervised reading comprehension data. This allows us to develop a class of attention based deep neural networks that learn to read real documents and answer complex questions with minimal prior knowledge of language structure.

Another Tottering Step Toward a New Era of Data-Making

*Dart Throwing Chimp* blog from June 09, 2015

Ken Benoit, Drew Conway, Benjamin Lauderdale, Michael Laver, and Slava Mikhaylov have an article forthcoming in the *American Political Science Review* that knocked my socks off when I read it this morning. Here is the abstract from the ungated version I saw:

Empirical social science often relies on data that are not observed in the field, but are transformed into quantitative variables by expert researchers who analyze and interpret qualitative raw sources. While generally considered the most valid way to produce data, this expert-driven process is inherently difficult to replicate or to assess on grounds of reliability. Using crowd-sourcing to distribute text for reading and interpretation by massive numbers of non-experts, we generate results comparable to those using experts to read and interpret the same texts, but do so far more quickly and flexibly. Crucially, the data we collect can be reproduced and extended transparently, making crowd-sourced datasets intrinsically reproducible. This focuses researchers’ attention on the fundamental scientific objective of specifying reliable and replicable methods for collecting the data needed, rather than on the content of any particular dataset. We also show that our approach works straightforwardly with different types of political text, written in different languages. While findings reported here concern text analysis, they have far-reaching implications for expert-generated data in
the social sciences.

Costs Of Slipshod Research Methods May Be In The Billions
NPR, Shots blog from June 09, 2015
Laboratory research seeking new medical treatments and cures is fraught with pitfalls: Researchers can inadvertently use bad ingredients, design the experiment poorly, or conduct inadequate data analysis. Scientists working on ways to reduce these sorts of problems have put a staggering price tag on research that isn't easy to reproduce: $28 billion a year.

That figure, published Tuesday in the journal *PLOS Biology*, represents about half of all the preclinical medical research that's conducted in labs (in contrast to research on human volunteers). And the finding comes with some important caveats.

The $28 billion doesn't just represent out-and-out waste, the team that did the research cautions. It also includes some studies that produced valid results — but that couldn't be repeated by others because of the confusing way the methods were described, or because of other shortcomings. [audio, 3:28]

Please join us in welcoming Data Scientist Bernease Herman to the eScience Institute!
Facebook, UW eScience Institute from June 03, 2015
Bernease was most recently a Software Development Engineer at Amazon where she collaborated with operations research scientists and statisticians to add economic constraints and buying models to Amazon’s Inventory Planning and Control system. Previous to Amazon, Bernease worked on derivatives pricing and predictive modeling at the research arm of Morgan Stanley. Bernease earned her BS in Mathematics and Statistics from the University of Michigan.

Please join us in welcoming Anthony Arendt to the eScience Institute!
Facebook, UW eScience Institute from June 04, 2015
Anthony is a Senior Research Scientist with the Polar Science Center at the Applied Physics Laboratory, where he conducts research on the response of glaciers and ice sheets to changing climate. Anthony joined the eScience Institute in June 2015 and provides expertise in relational databases, geospatial data analytics, and development of lightweight cloud computing solutions for scientific research.

HHS on a mission to liberate health data
GCN from June 05, 2015
... According to Damon Davis, the director for the health data initiative at HHS, HealthData.gov is a catalog of health, social services and research data made available to the public to improve the country’s health.

The relaunch includes user-friendly tools and updates the platform’s underlying technology for more efficient performance, Davis said in a blog, in an effort to nurture more applications, products and services capable of enhancing health care.

The project started with the migration of the catalog content to the DKAN open data and open source platform, which is the same technology used by Data.gov.

Higher-Level Tools for Interactive Data Visualization
ISTC Big Data from June 04, 2015
In popular conversation, discussions of data visualization often involve specialized interactive graphics – typically intended to convey a story – that were hand-crafted by skilled
designers. Over our years working in visualization, a central goal of our group has been to build tools that help designers craft sophisticated graphics, leading us to create tools such as Prefuse, Protovis and D3.js.

In the grand scheme of things, however, finely honed hand-coded visualizations are the exception, not the rule. Instead, the vast majority of the world’s visualizations are produced using end-user software, most notably spreadsheet applications. While valuable, these tools are often created by engineers, who, though well-intentioned, unfortunately lack familiarity with effective visualization design and how to support the iterative process of data analysis. Better tools are needed to help everyone, not just skilled designers, create effective visualizations and better understand their data.

Mortgages Are About Math: Open-Source Loan-Level Analysis of Fannie and Freddie

Todd W. Schneider from June 09, 2015

Fannie Mae and Freddie Mac began reporting loan-level credit performance data in 2013 at the direction of their regulator, the Federal Housing Finance Agency. The stated purpose of releasing the data was to “increase transparency, which helps investors build more accurate credit performance models in support of potential risk-sharing initiatives.”

The so-called government-sponsored enterprises went through a nearly $200 billion government bailout during the financial crisis, motivated in large part by losses on loans that they guaranteed, so I figured there must be something interesting in the loan-level data. I decided to dig in with some geographic analysis, an attempt to identify the loan-level characteristics most predictive of default rates, and more. As part of my efforts, I wrote code to transform the raw data into a more useful PostgreSQL database format, and some R scripts for analysis. The code for processing and analyzing the data is all available on GitHub.

Events

Libraries, Digital Privacy, & Data Literacy | TA3M June Tickets, New York | Eventbrite

Join us for a conversation about the impact of surveillance and data collection on citizens, specifically on disadvantaged communities. Learn more about the privacy and data issues that librarians face in their work and new efforts to empower them to address these issues.

Libraries are among the most trusted institutions in their communities, making librarians uniquely positioned to prepare patrons for the privacy challenges brought about by the pervasiveness of data sharing, profiling, DRM, third-party platforms, and surveillance technologies. Individuals with the greatest digital literacy needs are also the most vulnerable to abuses of personal data, creating an even more urgent need for libraries to address these issues.

Monday, June 15, at 7 p.m., Thoughtworks, 99 Madison Ave

Quantitative Biology: From Molecules to Man

Quantitative Biology: From Molecules to Man will bring together professionals in science, medicine, and engineering to articulate a vision for the future of improving patient health outcomes. Convergence science provides for a data-driven understanding of intricate biological processes across spatial and temporal scales. Achieving breakthroughs in healthcare requires a specific progression of steps from molecular-level experiments to
manipulations and observations in model systems to human-scale investigations, all followed by major epidemiological studies. This one-day meeting will provide a forum for individuals involved in every stage of the process to engage in thought-provoking conversations and to generate actionable ideas for new approaches to finding solutions to some of humanity’s most intractable health challenges.

Thursday, June 18, at New York Academy of Sciences, 250 Greenwich St.

**The NYC Arcade**

The Sheep’s Meow is taking over Playtest Thursday on June 18th to bring you the NYC Arcade. Check out the latest locally-developed games & interactive experiences with the creators!

Want to show a project? Submit your work for the showcase via this form. The full lineup of work will be announced one week before the event.

The mission of The Sheep’s Meow is to support local developers, make game creation inclusive & accessible, and help people get started making their own games.

Thursday, June 18, at 6 p.m., 2 MetroTech Center

**Tutorials Schedule | SciPy 2015**

**New Tutorial Options:** Due to overwhelming demand, our tutorial presenters for the intermediate and advanced Machine Learning with Scikit-Learn tutorials graciously allowed us to add a second run of these sessions on Tuesday, July 7th (in addition to the sessions on Monday, July 6th).

Machine Learning with Scikit-Learn (Intermediate, Monday/Tuesday, 8:00 - noon)
Instructors: Andreas Mueller, Assistant Research Scientist, New York University Center for Data Science, and Kyle Kastner, Graduate Student, Université de Montréal

Machine Learning with Scikit-Learn (Advanced, Monday/Tuesday, 1:00 - 5:00 pm)
Instructors: Andreas Mueller, Assistant Research Scientist, New York University Center for Data Science, and Kyle Kastner, Graduate Student, Université de Montréal

**Software Carpentry tutorial at SciPy 2015**

Software Carpentry’s mission is to help scientists and engineers get more research done in less time and with less pain by teaching them basic lab skills for scientific computing. This hands-on workshop will cover basic concepts and tools, including program design, version control, and task automation. We'll also briefly introduce the basic libraries of scientific Python. Participants will be encouraged to help one another and to apply what they have learned to their own research problems.

Instructors: Azalee Bostroem, Matt Davis, Jess Hamrick, Ted Hart, Katy Huff, Thomas Kluyver, Jens Nielsen, April Wright

Monday-Tuesday, July 6-7, part of the 2015 SciPy Conference in Austin, TX

**Deadlines**

**New England Symposium on Statistics in Sports at Harvard on Saturday, September 26**

The 2015 New England Symposium on Statistics in Sports will be a meeting of statisticians and quantitative analysts connected with sports teams, sports media, and universities to
discuss common problems of interest in statistical modeling and analysis of sports data. The symposium format will be a mixture of invited talks, a poster session, and a panel discussion.

Abstracts Deadline: June 15

Submit Your Demo for the 2015 Annual Summit
Taking place Friday, September 25, NYC Media Lab’s Annual Summit is a crash course in the best thinking, projects, and talent in digital media and communications across the City’s universities. The demo showcase will feature more than 100 digital media projects from students and faculty in NYC and beyond.

This is an opportunity to share your prototypes, research, and startups in a trade show-style environment to a gathering of 800+ executives, technologists, and venture capitalists from NYC Media Lab’s community.

Deadline for first round of submissions: mid-June

Astro Hack Week 2015
ASTRO HACK WEEK IS A WEEK-LONG SUMMER SCHOOL / HACK WEEK / UNCONFERENCE FOCUSED ON ASTROSTATISTICS AND DATA-INTENSIVE ASTRONOMY.

The mornings will be a typical summer school format, with lectures and exercises covering essential skills for working effectively with large astronomical datasets. The afternoons will be entirely unstructured, and offer opportunities for collaborative research, breakout sessions on special topics, and application of the concepts covered during the morning sessions. The vision is to provide a space to encourage learning, research, collaboration, and sharing of expertise, for the benefit of both young and experienced astronomical researchers alike.

Deadline for Expression of Interest: Monday, June 15

AT&T NYU Connect Ability Challenge
The AT&T NYU Connect Ability Challenge is a three-month global software development competition leveraging mobile and wireless technologies to improve the lives of people living with disabilities.

Deadline to submit entries: Wednesday, June 24

CDS News
The Future of Computers Is the Mind of a Toddler
Bloomberg Business from June 09, 2015

... Corporate research labs now rival those in academia in terms of staffing and funding. They have surpassed them in access to proprietary data and computing power to run experiments on. That’s attracting some of the field’s most prominent researchers. LeCun, former director of New York University’s Center for Data Science, joined Facebook in December 2013 to run its AI group. While still teaching a day a week at NYU, he has hired nearly 50 researchers; on June 2, Facebook said it is opening an AI lab in Paris, its third such facility.
Accelerating Data-driven Discovery and Innovation: New York Scientific Data Summit

*Scientific Computing* from June 05, 2015

The first-annual New York Scientific Data Summit (NYSDS) is a no-fee meeting that aims to accelerate data-driven discovery and innovation by bringing together researchers, developers and end-users from academia, industry, utilities and state and federal governments. The Summit is jointly organized by Brookhaven National Laboratory, Stony Brook University and New York University. Sponsors include the New York University Center for Data Science, and the Moore-Sloan Data Science Environment.

The 2015 theme is "Frontiers in Scientific Data." There is no registration fee to attend. Registration includes a sponsored welcome reception, conference dinner and coffee breaks. Note that registration will be limited to 150 participants.

Summit: Sunday-Wednesday, August 2-5, at NYU.

Deadline for Registration and Poster Abstracts: Monday, July 20

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 005.