NYU Data Science Community features journalism, research papers, events, tools/software, and jobs for April 8, 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

First, design for data sharing

Nature Biotechnology; John Wilbanks & Stephen H Friend from April 07, 2016
To upend current barriers to sharing clinical data and insights, we need a framework that not only accounts for choices made by trial participants but also qualifies researchers wishing to access and analyze the data.

Amid Public Feuds, A Venerated Medical Journal Finds Itself Under Attack

ProPublica, The Boston Globe from April 05, 2016
A widely derided editorial, a controversial series of articles and delayed corrections have prompted critics to question the direction of the *New England Journal of Medicine*.

Moderating Harassment in Twitter with Blockbots

Berkeley Institute for Data Science, R. Stuart Geiger from April 06, 2016
I've been working on a research project about counter-harassment projects in **Twitter**, where I've been focusing on blockbots (or bot-based collective blocklists) in Twitter. Blockbots are a different way of responding to online harassment, representing a more decentralized alternative to the standard practice of moderation—typically, a site's staff has to go through their own process to definitively decide what accounts should be suspended from the entire site. I'm excited to announce that my first paper on this topic will soon be published in *Information, Communication, and Society*.

Community-level data science and its spheres of influence: beyond novelty squared

UW eScience Institute, Brittany Fiore-Gartland and Anissa Tanweer from April 04, 2016
Data science has many characterizations, but in academia it is often talked about as pushing the limits of both methodological and domain science, what **Josh Bloom**, a Professor of Astronomy at U.C. Berkeley, has referred to as "novelty squared". Bloom sees this as the "great challenge of modern interdisciplinary scientific collaboration". The idealized characterization of data science in academia is also represented in the idea of shifting from the traditional T-shaped scientists, who have deep expertise in a single domain, to (Pi)-shaped scientists with deep expertise in both a domain and methodological science (as coined by **Alex Szalay** and discussed here and here. As (Pi)-shaped data scientists, they are primed to innovate in multiple disciplinary trajectories. Bloom and others have argued that these characterizations of novelty squared and (Pi)-shaped scientists represent the "unicorn" of data science.

Fed's Kashkari: Risks still lurk in banking system

CNBC from April 05, 2016
Post-financial crisis changes have made the financial system more stable, but policymakers need to acknowledge that large banks still pose systemic risk, Minneapolis Fed President **Neel Kashkari** contended Tuesday.

Kashkari, 42, started at the bank earlier this year and quickly made further regulation a top priority. On Monday, he hosted a symposium in Minneapolis called "Ending Too Big to Fail," where experts floated ideas about how best to make the financial system safer.

New undergraduate majors and minors to debut in fall 2016

MIT News from April 08, 2016
New major bachelor of science programs include Management, Business Analytics, Finance, and Mathematical Economics (pending approval at the April MIT faculty meeting).

New minor programs include Computer Science, offered by the Department of Electrical Engineering and Computer Science; Design, offered by the Department of Architecture and the School of Architecture and Planning (in collaboration with departments across MIT); Entrepreneurship and Innovation, jointly offered by the School of Engineering and MIT Sloan School of Management; Statistics and Data Science, offered by the Institute for Data, Systems, and Society (IDSS); and Management, Business Analytics, and Finance, offered by MIT Sloan.

Recognize the value of social science

Nature News & Comment, Andrew Webster from April 05, 2016
If the science community is serious about integrating social science into its thinking and operations — and statements by everyone from *Nature* and the UK government to **Paul Nurse**, former president of the Royal Society, indicate that it is — then we social scientists must do more to make this happen.

Our input is necessary because, too often, the reach and influence of research is discovered only with hindsight. Lessons are 'learned' only after the social implications of new domains of science and technology have provoked controversy or challenged existing norms.

Also, Restructuring the Social Sciences: Reflections from Harvard's Institute for Quantitative Social Science (March 21, *Political Science Now*, **Gary King**)

What social media data could tell us about the future

news @ Northeastern from April 07, 2016
Northeastern's **Alessandro Vespignani**, Sternberg Family Distinguished University Professor of physics, computer science, and health sciences, has teamed up with an interdisciplinary group of scientists to develop an innovative method to map how tweets about large-scale social events spread. Using massive **Twitter** datasets and sophisticated quantitative measures, it tracks how information about political protests, large business acquisitions, and other "collective phenomena" gather momentum,

peak, and fall over time, from city to city, and where the impetus comes from for that trajectory.

Self-regulatory information sharing in participatory social sensing

EPJ Data Science journal; Evangelos Pournaras et al. from April 01, 2016
Participation in social sensing applications is challenged by privacy threats. Large-scale access to citizens' data allow surveillance and discriminatory actions that may result in segregation phenomena in society. On the contrary are the benefits of accurate computing analytics required for more informed decision-making, more effective policies and regulation of techno-socio-economic systems supported by 'Internet-of Things' technologies. In contrast to earlier work that either focuses on privacy protection or Big Data analytics, this paper proposes a self-regulatory information sharing system that bridges this gap. This is achieved by modeling information sharing as a supply-demand system run by computational markets. On the supply side lie the citizens that make incentivized but self-determined decisions about the level of information they share. On the demand side stand data aggregators that provide rewards to citizens to receive the required data for accurate analytics. The system is empirically evaluated with two real-world datasets from two application domains: (i) Smart Grids and (ii) mobile phone sensing. Experimental results quantify trade-offs between privacy-preservation, accuracy of analytics and costs from the provided rewards under different experimental settings. Findings show a higher privacy-preservation that depends on the number of participating citizens and the type of data summarized.

[1603.09326] Estimating Treatment Effects using Multiple Surrogates: The Role of the Surrogate Score and the Surrogate Index

arXiv, Statistics > Methodology; Susan Athey et al. from March 30, 2016
Estimating the long-term effects of treatments is of interest in many fields. A common challenge in estimating such treatment effects is that long-term outcomes are unobserved in the time frame needed to make policy decisions. One approach to overcome this missing data problem is to analyze treatments effects on an intermediate outcome, often called a statistical surrogate, if it satisfies the condition that treatment and outcome are independent conditional on the statistical surrogate. The validity of the surrogacy condition is often controversial. Here we exploit that fact that in modern datasets, researchers often observe a large number, possibly hundreds or thousands, of intermediate outcomes, thought to lie on or close to the causal chain between the treatment and the long-term outcome of interest. Even if none of the individual proxies satisfies the statistical surrogacy criterion by itself, using multiple proxies can be useful in causal inference. We focus primarily on a setting with two samples, an experimental sample containing data about the treatment indicator and the surrogates and an observational sample containing information about the surrogates and the primary outcome.

Events

DS Industry Speaker Series — April 12, Dave Thomas, Kx Labs — The Future of Database - Revolution or Evolution?

In the last decade years there have been significant improvements in computer processors, storage and networking. ... In this talk we briefly explain the essential features, similarities and differences between these different database perspectives. We examine evolution of Kx Kdb+ given the data challenges of our customers and the market. We then discuss the impact of modern hardware on database architecture and data languages.

Tuesday, April 12, starting at 12:30 p.m. at the **Center for Data Science**, located at 726 Broadway, 7th Floor

NYU Computer Science Department Colloquium — Marginalization is not Marginal: Non-Convex, Bayesian-Inspired Algorithms for Sparse and Low-Rank Estimation

Speaker: **David Wipf**, **Microsoft Research Beijing**

Many practical applications of sparsity and low rank matrices do not benefit from this luxury; rather, because of intrinsic correlations in the signal dictionary (or related structure in rank minimization problems), convex algorithms must be used in regimes where theoretical support no longer holds. Moreover, in some situations it has been shown that convex relaxations are in fact provably bad. Consequently, non-convex algorithms, while perhaps theoretically less accommodating, may nonetheless produce superior results. Here we will examine non-convex estimation algorithms, many of which originate from Bayesian machine learning ideas, that thrive in environments where more popular convex alternatives fail. In all cases, theoretical model justification will be provided independent of any assumed prior distributions.

Wednesday, April 13, starting at 11:30 a.m., in Warren Weaver Hall 1302

Moore-Sloan Data Science Lunch Seminar Series

Wednesday, April 13 — **Andy Guess** from **NYU Social Media and Political Participation (SMaPP) Lab**

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.

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 14, will be **Sven-Oliver Proksch** (**McGill**).

2016 Atlantic Causal Inference Conference

The Atlantic Causal Inference Conference is a gathering of statisticians, biostatisticians, epidemiologists, economists, social science and policy researchers to discuss methodologic issues with drawing causal inferences from experimental and non-experimental data. The inaugural meeting was held in 2005 with a small group of researchers at **Columbia University** and has since grown into an annual event with over 100 attendees.

Thursday-Friday, May 26-27, at **New York University**, Kimmel Center

Gaia Sprints — A project to support exploitation of the Gaia First Data Release.

The idea behind the Sprints is to bring together people who have an interest in timely exploitation of the Gaia First Data Release. These are not traditional scientific meetings; they are intended to facilitate completion of first scientific papers. The Sprints will be structured to support collaborative refinement and execution of (fairly) mature scientific ideas. It is hoped that new partnerships will form and lead to co-authored publications for the scientific literature ready or near-ready by the end of each Sprint. (Advance registration required.)

Monday-Friday, October 17-21, at the **Simons Center for Computational Astrophysics**, 160 Fifth Avenue, 7th Floor, New York, NY

Deadlines

Deep Learning Summer School 2016

Deep neural networks that learn to represent data in multiple layers of increasing abstraction have dramatically improved the state-of-the-art for speech recognition, object recognition, object detection, predicting the activity of drug molecules, and many other tasks. Deep learning discovers intricate structure in large datasets by building distributed representations, either via supervised, unsupervised or reinforcement learning.

This summer schools is aimed at graduate students and industrial engineers and researchers who already have some basic knowledge of machine learning (and possibly but not necessarily of deep learning) and wish to learn more about this rapidly growing field of research. This year's edition of the summer school is organized by **Aaron Courville** and **Yoshua Bengio**.

Deadline for applications is Monday, April 11.

HILDA 2016: Workshop on Human-In-the-Loop Data Analytics

HILDA is a new workshop that will allow researchers and practitioners to exchange ideas and results relating to how data management can be done with awareness of the people who form part of the processes. A sample of topics that is are in the spirit of this workshop includes, but is not limited to: novel query interfaces, interactive query refinement, data exploration and analysis, data visualization, human-assisted data integration and cleaning, perception-aware data processing, database systems

designed for highly interactive use cases, empirical studies of database use, and crowd-powered data infrastructure. (Co-located with **SIGMOD 2016** in San Francisco, CA.)

Deadline for submissions is Monday, April 18.

ICML 2016 Workshops: #Data4Good: Machine Learning in Social Good Applications

The goal of this workshop is to bring together experts from different fields of machine learning, statistics, data science, social sciences and social activism to explore the opportunities for machine learning in applications with social impact. The workshop will consist of: 1) invited presentations from the leading practitioners in the field, and 2) a series of 20 minute presentations on research that fits the theme of machine learning for social good; broadly construed, this could be machine learning related social good applications, or new machine learning methods or theory of particular interest for social good applications.

Deadline for submissions is Sunday, May 1.

Also, 2016 Workshop on Human Interpretability in Machine Learning at ICML 2016 (same May 1 deadline)

SBP-BRiMS 2016 Grand Data Challenge

Fundamental research problems exist in how to fuse data, how to identify the relevant portions of the data, how assess change in the data, how to sample the data, and how to visualize the data. These issues must be met to advance social theorizing and improve policy analysis. This year's SBP-BRiMS challenge problem invites you to take part in addressing one or more of these challenges.

Using at least one of four political event datasets (GDELT, KEDS, ICEWS, Phoenix) and one other data set (which may be a second one of these event datasets, or any other relevant dataset), this year's challenge problem asks participants to address any issue of interest to you or your team that involves events and their distribution over time or space. All entries must have both a strong social theory, political theory or policy perspective and a strong methodology perspective.

Deadline for abstract submissions is Sunday, May 1.

CDS News

[1604.00289] Building Machines That Learn and Think Like People

arXiv, Computer Science > Artificial Intelligence; Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, Samuel J. Gershman from April 01, 2016
Recent progress in artificial intelligence (AI) has renewed interest in building systems that learn and think like people. Many advances have come from using deep neural networks trained end-to-end in tasks such as object recognition, video games, and board games, achieving performance that equals or even beats humans in some

respects. Despite their biological inspiration and performance achievements, these systems differ from human intelligence in crucial ways. We review progress in cognitive science suggesting that truly human-like learning and thinking machines will have to reach beyond current engineering trends in both what they learn, and how they learn it. Specifically, we argue that these machines should (a) build causal models of the world that support explanation and understanding, rather than merely solving pattern recognition problems; (b) ground learning in intuitive theories of physics and psychology, to support and enrich the knowledge that is learned; and (c) harness compositionality and learning-to-learn to rapidly acquire and generalize knowledge to new tasks and situations. We suggest concrete challenges and promising routes towards these goals that can combine the strengths of recent neural network advances with more structured cognitive models.

When Social Movements Reach Their Tipping Point, and Why Scientific Collaboration is Important

Annenberg School for Communication, Bruno Goncalves from April 05, 2016
Unlike top-down actions like a corporate announcement or a governmental decree that then have rippling effects on the public, events of collective effervescence start small — a tight circle of activists, a maverick trendsetter — and grow organically until they suddenly seem to explode. Protesters pour into the streets. Rumors become national headlines. Americans under 25 suddenly all wear high-waisted jeans.

Finding a way to identify that tipping point where a movement explodes was one of the aims of the study. [video, 1:21]

Tools & Resources

pomegranate — pomegranate 0.4.0 documentation

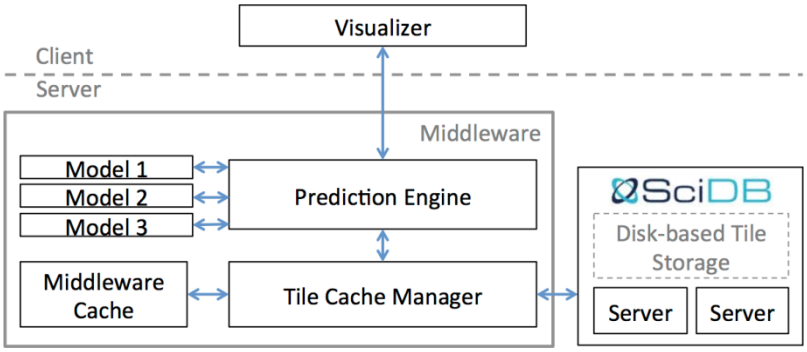
Jacob Schreiber from April 04, 2016
pomegranate implements fast, efficient, and extremely flexible probabilistic modelling for Python. It grew out of the YAHMM package where many of the components of hidden Markov models could be rearranged to form other probabilistic models, such as general mixture models and markov chains. pomegranate is flexible enough to allow nesting of these components to form models such as general mixture model hidden Markov models (GMM-HMMs) or Naive Bayes comparing a hidden Markov model to a Markov chain.

Deep Learning for Chatbots, Part 1 – Introduction – WildML

Denny Britz, WildML blog from April 06, 2016
In this series I want to go over some of the Deep Learning techniques that are used to build conversational agents, starting off by explaining where we are right now, what's possible, and what will stay nearly impossible for at least a little while. This post will serve as an introduction, and we'll get into the implementation details in upcoming posts.

ForeCache: Raising the Bar in Big Data Visual Exploration

MIT CSAIL, Intel Science & Technology Center for Big Data from April 06, 2016



To push beyond the limitations of current DBMSs and support interactivity, we developed the ForeCache visual exploration system (see Figure 1). ForeCache uses a client-server architecture: The user interacts with a visualization interface running on the client machine (i.e., the user's laptop), and the client retrieves the corresponding data by issuing requests to a DBMS running on a remote server. For its extensive support for scientific analysis operations, we use the array-based DBMS SciDB as our back-end. To further boost back-end performance, ForeCache includes a server-side middleware layer inserted in front of the DBMS, which pre-fetches data into a main memory cache in anticipation of the user's future interactions.

api-packages.Rmd

GitHub - hadley/httr (Hadley Wickham) from April 07, 2016
So you want to write an R client for a web API? This document walks through the key issues involved in writing API wrappers in R. If you're new to working with web APIs, you may want to start by reading "An introduction to APIs" by zapier.

dbpatterns - create, share, explore database patterns

Fatih Erikli from April 01, 2013
Dbpatterns is a service that allows you to create, share, explore database models on the web. GitHub: https://github.com/fatiherikli/dbpatterns.

NASA, Japan Make ASTER Earth Data Available At No Cost

NASA from April 01, 2016
Beginning today, all Earth imagery from a prolific Japanese remote sensing instrument operating aboard NASA's Terra spacecraft since late 1999 is now available to users everywhere at no cost.

The public will have unlimited access to the complete 16-plus-year database for **Japan's Ministry of Economy, Trade and Industry (METI)** Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) instrument, which images Earth to map and monitor the changing surface of our planet. ASTER's database currently consists of more than 2.95 million individual scenes.

Careers

Million-dollar babies -- As Silicon Valley fights for talent, universities struggle to hold on to their stars

The Economist from April 02, 2016
... In the past universities employed the world's best AI experts. Now tech firms are plundering departments of robotics and machine learning (where computers learn from data themselves) for the highest-flying faculty and students, luring them with big salaries similar to those fetched by professional athletes.

Last year **Uber**, a taxi-hailing firm, recruited 40 of the 140 staff of the **National Robotics Engineering Centre at Carnegie Mellon University**, and set up a unit to work on self-driving cars. That drew headlines because Uber had earlier promised to fund research at the centre before deciding instead to peel off its staff. Other firms seek talent more quietly but just as doggedly.

Also, Hot Commodity (March 28, *MIT Technology Review* feature on **Andrej Karpathy**)

Short Term Research Assistant — The Governance Lab (GovLab) at NYU

NYU, The Government Lab from April 04, 2016
The **Governance Lab (GovLab) at NYU**, in collaboration with a small research team at the **Columbia Business School**, seeks a motivated and engaged Research Assistant interested in learning more about the topics of entrepreneurship, open data, and social networks. The project s/he will be working on seeks to understand to what extent the free release of U.S. government data impacts entrepreneurship. The assistant will work closely with a Professor and postdoc at Columbia Business School on managing and launching a survey to various startups, along with other research tasks; and with the Chief Research and Development Officer and project manager at the GovLab on survey outreach and implementation.

The ideal candidate will be familiar with navigating NYU's library systems and online databases, be capable of formulating broad questions and seeking answers from diverse sources, and exhibit an outstanding attention to detail and organization. S/he should be comfortable working with spreadsheets and managing data. Experience conducting research and/or coordinating research projects is a plus

Seeking Full-stack Engineer / Data Scientist, Open Syllabus Project

The Open Syllabus Project from April 04, 2016
The Open Syllabus Project is an academic data-mining project at **Columbia** and **Stanford** that's extracting structured information from a corpus of 1M+ college course syllabi. What's actually being taught in college classrooms? How has this changed over time? What can we learn about the organization of the modern university from large-scale trends in the texts that are being assigned? How can insights from these data be applied to curriculum development, education policy, and lifelong learning?

We're looking for someone who has experience with large-scale data analysis, natural language processing, web archiving, and web application development to help us grow OSP into a comprehensive, feature-rich authority about teaching trends in higher education.

Increase Your Earnings by Freelancing

Ellevest, Melissa Cullens from March 14, 2016
Figuring out the value of what you do can at first seem like a challenge. You don't want to just spout out a number that "sounds good," or "seems right." Here's a general equation for calculating your hourly rate, with some real numbers as examples:

- **$3000** — Monthly personal expenses, including household bills
- **$1200** — Monthly overhead and operating costs, including internet, a new laptop every two years, printing, software, phone, etc.
- About **$459 — $5500** into an IRA, annually, broken out by month
- About **$4659** — Total monthly costs
- That means billing **$55,899.96** annually, and we haven't gotten to taxes yet. Remember, you have to set aside 30% for Uncle Sam.
- About **$55,900** — Annual income needed to cover basic and personal expenses, overhead, and IRA contributions
- About **$72,670** — Plus 30% for taxes
- About **$87,204** — Plus 20% for an investment account [optional]
- Which means you need to bill about **$7,267** each month. If you were working 40 hours a week, 52 weeks a year, you'd charge **$42** an hour. But I'm guessing the whole reason you went freelance in the first place is to have flexible work hours and vacation time. If you want to take three weeks off for vacation each year, and budget the standard five sick days, your hourly rate goes up to **$45** or **$46.**

Also, With 'Gigs' Instead of Jobs, Workers Bear New Burdens (March 31, *The New York Times, The Upshot* blog, Neil Irwin)

The next hot job in Silicon Valley is for poets

The Washington Post from April 07, 2016
Until recently, **Robyn Ewing** was a writer in Hollywood, developing TV scripts and pitching pilots to film studios.

Now she's applying her creative talents toward building the personality of a different type of character — a virtual assistant, animated by artifical intelligence, that interacts with sick patients.

Fast Forward Labs — Client-focused Data Scientist

Fast Forward Labs from April 06, 2016
We're looking for someone to fill a unique role -- it's a mix between data science, project management, and client relations. This person will focus on maintaining our clients' strong satisfaction with the Fast Forward Labs R&D and technical advising product, while also contributing to the development of future products.

Fast Forward Labs is a research company (headquartered in the Lower East Side) that helps organizations recognize and develop new product and business opportunities through emerging machine intelligence technologies. We offer a research subscription service and advisory services to small and large companies in a wide set of industries.

Postdoctoral Fellow in Center for Complex Networks and Systems

IU Bloomington School of Informatics and Computing from April 07, 2016

The **Center for Complex Networks and Systems Research (CNetS.indiana.edu)** has one open postdoctoral position to study critical processes in networks of networks. The appointment starts in June 2016 for one year and is renewable for other two years, subject to funding and performance. The salary is competitive and benefits are generous.

The postdoc will join a dynamic and interdisciplinary team that includes computer, physical, and cognitive scientists. The postdoc will work with Prof. **Filippo Radicchi**.

For best consideration completed applications must be received by May 15.