NYU Data Science Community news articles, blog posts and research papers for March 18, 2016

Each week we curate and share news articles, blog posts, research papers and events that we think will be of interest to our community of data scientists. Let us (Laura Noren, Brad Stenger) know how we can make this information better. Thank you to NYU and the Moore-Sloan Data Science Environment.

Rapid assessment of disaster damage using social media activity

Science Advances; Yury Kryvasheyeu, Haohui Chen, Nick Obradovich, Esteban Moro, Pascal Van Hentenryck, James Fowler and Manuel Cebrian from March 11, 2016

Could social media data aid in disaster response and damage assessment? Countries face both an increasing frequency and an increasing intensity of natural disasters resulting from climate change. During such events, citizens turn to social media platforms for disaster-related communication and information. Social media improves situational awareness, facilitates dissemination of emergency information, enables early warning systems, and helps coordinate relief efforts. In addition, the spatiotemporal distribution of disaster-related messages helps with the real-time monitoring and assessment of the disaster itself. We present a multiscale analysis of Twitter activity before, during, and after Hurricane Sandy. We examine the online response of 50 metropolitan areas of the United States and find a strong relationship between proximity to Sandy's path and hurricane-related social media activity.

NYU Wireless gives away millimeter wave simulator for 5G

RCR Wireless News from March 15, 2016

Looking to help speed the development of 5G technology related to millimeter wave spectrum, NYU Wireless opened up a channel simulator tool based on the group's research and experiments. NYU Wireless is a Brooklyn, New York,-based research center focused on next-generation wireless networks and devices; Ted Rappaport serves as the founding director of the institution.

As "5G" technology and standards continue to evolve, millimeter wave is being looked at to support more efficient and high-capacity data transmission. Specifically, the new software supports channel simulation in frequencies ranging from 28 GHz to 73 GHz, with French operator Orange already using the technology.

Would You Give All Your Personal Data to Science?

New York Magazine, Science of Us blog from March 10, 2016

... In 2017, researchers at New York University will begin to assemble a similarly massive database with the Kavli HUMAN Project, an audaciously ambitious study that will track the biology and behavior of 10,000 New Yorkers for the next 20 years. Their humble goal: to gather enough data over time to learn "everything there is to know about a group of people," said Paul Glimcher, a neuroscientist and economist, who is the director of the project. The project will begin by recruiting

2,500 New York volunteers, from across all five boroughs, whose entire households will also need to agree to the terms of the research project. For the next two decades, practically everything that happens in their lives will turn to data, which will be made accessible by researchers in a wide variety of fields, such as medicine, psychology, sociology, economics, and public policy.

More Data Collection & Security:

- Mobile data sharing in emergencies – consent, care and control (Linnet Taylor, March 15)
- NYU professor questions big data collection (The Brown Daily Herald, March 16)
- Health Apps Routinely Sell User's Data With Little Notice (Bloomberg BNA, March 9)
- Court rulings threaten to upset defences against data breach claims (Financial Times, March 16)
- Ebola: A Big Data Disaster (The Centre for Internet and Society, Sean McDonald, March 1)

Can you put a price on nature? A Californian nonprofit thinks it can

The Guardian, Guardian Sustainable Business from March 13, 2016

... A California nonprofit called The Earth Genome thinks it can change that, working with the world's second largest chemical company, Dow, to prove the power of its data-crunching tool.

"Just because scientists make valuations, doesn't make that [calculation] valuable," says Glen Low, a former consultant who co-founded The Earth Genome with Steve McCormick, former president of the Gordon and Betty Moore Foundation and The Nature Conservancy, in 2014. "They publish information that says, well this is what this part of nature is worth. But you have to ask: did anyone use that information to make a better decision which led to better conservation of that resource or to better outcomes for the corporation?"

More Environmental Data Science:

- Researchers propose satellite mission to improve understanding of global vegetation change (University of Minnesota, March 7)
- An API for the World's Weather & Climate Data (Planet OS, March 14)
- New climate science innovation center opens downtown (WLOS, Asheville NC, March 11)
- Can a long-dead reverend help save Amazonia's freshwater dolphins? (Science, ScienceInsider, March 11)
- Seafood CSI (Hakai Magazine, March 8)
- A Computer With a Great Eye Is About to Transform Botany (WIRED Design, March 17)

CSCW 2016: Beyond the Belmont Principles

LinkedIn, SlideShare, Jessica Vitak from March 01, 2016
Pervasive information streams that document people and their routines have been a boon to social computing research. But the ethics of collecting and analyzing available—but potentially sensitive—online data present challenges to researchers. In response to increasing public and scholarly debate over the ethics of online data research, this paper analyzes the current state of practice among researchers using online data. Qualitative and quantitative responses from a survey of 263 online data researchers document beliefs and practices around which social computing researchers are converging, as well as areas of ongoing disagreement. The survey also reveals that these disagreements are not correlated with disciplinary, methodological, or workplace affiliations. The paper concludes by reflecting on changing ethical practices in the digital age, and discusses a set of emergent best practices for ethical social computing research.

Mercedes Boots Robots From the Production Line

Bloomberg Business from February 25, 2016
Mercedes-Benz offers the S-Class sedan with a growing array of options such as carbon-fiber trim, heated and cooled cupholders and four types of caps for the tire valves, and the carmaker's robots can't keep up.

With customization key to wooing modern consumers, the flexibility and dexterity of human workers is reclaiming space on Mercedes's assembly lines. That bucks a trend that has given machines the upper hand over manpower since legendary U.S. railroad worker John Henry died trying to best a motorized hammer more than a century ago.

"Robots can't deal with the degree of individualization and the many variants that we have today," Markus Schaefer, the German automaker's head of production, said.

New Canada Excellence Research Chair

Government of Canada from March 15, 2016
The Minister of Science, Kirsty Duncan, today announced Dr. Erik Snowberg as the Canada Excellence Research Chair (CERC) in Data-Intensive Methods in Economics at The University of British Columbia (UBC). Dr. Snowberg comes to Vancouver from the California Institute of Technology. He will become Canada's first CERC in the social sciences and the country's 26th CERC overall.

Dr. Snowberg aims to help transform UBC into the world leader in data-driven political economy, a field that focuses on the relationship between politics and the economy. He and his research team are developing new ways to analyze big data — massive datasets of information produced and stored by companies, governments and private citizens. Dr. Snowberg's research will lead to new analytical tools for governments to use when making policy decisions.

Events

Tyranny of the Algorithm? Predictive Analytics & Human Rights

Bernstein Institute for Human Rights Annual Conference

The 2016 annual conference will leverage the interdisciplinary strengths of the Robert L. Bernstein Institute to consider the human rights implications of the varied uses of predictive analytics by state actors. As a core part of this endeavor, the conference will examine—and seek to advance—the capacity of human rights practitioners to access, evaluate, and challenge risk assessments made through predictive analytics by governments worldwide.

Monday-Tuesday, March 21-22, at NYU

Moore-Sloan Data Science Lunch Seminar Series

Wednesday, March 23 — Michael Blanton, NYU, Physics

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, March 24, will be Cristian Danescu-Niculescu-Mizi (Cornell).

Advanced Git and Github

Modern research involving data analysis increasingly uses programming to increase efficiency and allow for more effective use of data.As code becomes a more and more essential part of research activities, we need to treat it with the same care that we treat other research products. The first step towards more maintainable software development and data analysis is using version control on all research and analysis code. Git is a popular tool for tracking individual and collaborative development of code.

This workshop takes a look at advanced usage and collaboration using Git and

GitHub, including: the concept of branches, and how to manipulate them with merge and rebases, forks and pull requests, and we'll even rewrite history using rebase, and possible workflows.

Tuesday, March 29, at 4 p.m., Bobst Library, Rm. 619

More NYU Libraries Training: Data Visualization Clinic No.2 on Friday, March 25.

Skills-Building Workshop: Tableau Data Visualization

The Center for Human Rights and Global Justice and Tandon School of Engineering are pleased to host Dash Davidson, a Data Analyst at Tableau Software, who will demonstrate Tableau's powerful suite of data visualization tools.

Monday, April 4, starting at 6:30 p.m. in Vanderbilt Hall, Smart Classroom 214

Deadlines

CDS Data Science Fellow

Data Science Fellows will be expected to work on research at the boundaries between datascience methods and another field of scholarly activity (domain science, humanities, ethics). They will lead independent, original research programs with impact in one or more scholarly domains and in one or more methodological domains (computer science, statistics, and applied mathematics). They are also encouraged to develop collaborations with partners in other universities and in industry.

Fellowship applicants should send a curriculum vitae, list of publications, and brief statement of research interests (no longer than 4 pages) to ds-jobs-group@nyu.edu, and also arrange to have three letters of recommendation sent as soon as possible. Applications are still being accepted.

More post-doctoral and fellowship opportunities:

- Moore-Sloan Fellows (NYU)
- Moore-Sloan Fellows (UW-Seattle)
- Moore-Sloan Fellows (UC-Berkeley)
- Open Web Fellows Program (Mozilla, Ford Foundation)
- Post-doc in Computer Science at University of Michigan (Grant Schoenebeck)
- The MIT School of Humanities, Arts, and Social Sciences offers a two-year Postdoctoral Fellowship in Digital Humanities.

Women in Statistics and Data Science Conference, Speed Abstract Submission

We are calling for speed session abstracts from senior, mid-level, and junior stars representing the industrial, academic, and government communities who would like

to present their life's work or share their perspectives on the role of women in today's statistics and data science fields.

Deadline to submit abstracts is Thursday, March 24.

Google Summer of Code

Spend your summer break writing code for an open source software project!

Applications open March 14, 2016 at 15:00 (EDT)

Deadline for application submission is Friday, March 25.

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.

Tools & Resources

Decibel: Dataset Branching for Collaborative Data Management

ISTC Big Data from March 16, 2016

The methods many data scientists currently use to coordinate operations are often ad hoc and rely on making full, redundant copies of an entire dataset in their individual workspaces. This not only wastes storage, but also woefully restricts collaborations: users cannot easily share patches or modifications to datasets, users cannot easily track which versions of a dataset were used for certain experiments, and there is no easy way to determine who is using particular versions of a dataset. ... To remedy this problem, a team of ISTC researchers has introduced DataHub, a hosted, collaborative and on-line analytics platform that allows users to easily upload,

modify, query, and share datasets.

Also, similarly: Databrary helps researchers collaborate by sharing video and data (Penn State University, Daily Collegian, March 16)

Celebrating Figaro 4.0: What is Structured Factored Inference?

Avi Pfeffer, Practical Probabilistic Programming blog from March 16, 2016
Figaro 4.0 has just been released, available from http://www.cra.com/figaro. The headline new feature is called “Structured Factored Inference”, or SFI. So I’d like to take this opportunity to explain what SFI is about and why we’re pursuing it.

State of the Art JavaScript in 2016

Medium, JavaScript and Opinions, Francois Ward from February 28, 2016
... the good news is the ecosystem is starting to slow down. Projects are merging. Best practices are starting to become clear. People are building on top of existing stuff instead of building new frameworks.

As a starting point, here’s my personal picks for most pieces of a modern web application. Some choices are likely controversial and I will only give basic reasoning behind each choices. Keep in mind they’re mostly my opinion based on what I’m seeing in the community and personal experiences. Your mileage may vary.

Introduction to Scikit Flow

Yuan Tang, Yuan's Blog from March 14, 2016
Scikit Flow is a simplified interface for TensorFlow, to get people started on predictive analytics and data mining. It helps smooth the transition from the Scikit-learn world of one-liner machine learning into the more open world of building different shapes of ML models. You can start by using fit/predict and slide into TensorFlow APIs as you are getting comfortable. It’s Scikit-learn compatible so you can also benefit from Scikit-learn features like GridSearch and Pipeline.

Introducing Kafka Streams: Stream Processing Made Simple

confluent, Jay Kreps from March 10, 2016
I’m really excited to announce a preview of a new feature in Apache Kafka called Kafka Streams. Kafka Streams is a Java library for building distributed stream processing apps using Apache Kafka. It will be part of the upcoming Kafka 0.10 release and we’ve made a preview version available to make it easy to try out now. The Kafka Streams source code is available under the Apache Kafka project.

What’s New — pandas 0.18.0 documentation

pandas from March 13, 2016
This is a major release from 0.17.1 and includes a small number of API changes,

several new features, enhancements, and performance improvements along with a large number of bug fixes. We recommend that all users upgrade to this version.