NYU Data Science Community news articles, blog posts and research papers for March 4, 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.

Inside the Artificial Intelligence Revolution: Pt. 1

Rolling Stone, Jeff Goodell from February 29, 2016
Welcome to robot nursery school," Pieter Abbeel says as he opens the door to the Robot Learning Lab on the seventh floor of a sleek new building on the northern edge of the UC-Berkeley campus. The lab is chaotic: bikes leaning against the wall, a dozen or so grad students in disorganized cubicles, whiteboards covered with indecipherable equations. Abbeel, 38, is a thin, wiry guy, dressed in jeans and a stretched-out T-shirt. He moved to the U.S. from Belgium in 2000 to get a Ph.D. in computer science at Stanford and is now one of the world's foremost experts in understanding the challenge of teaching robots to think intelligently. But first, he has to teach them to "think" at all. "That's why we call this nursery school," he jokes. [video autoplays, pre-roll + 1:31]

Also:

- The Promise of Artificial Intelligence Unfolds in Small Steps (The New York Times, February 28)
- What counts as artificially intelligent? AI and deep learning, explained (The Verge, February 29)
- AlphaGo and AI Progress (Miles Brundage, February 27)

Precision Medicine at one year: A soaring White House summit and the potholes ahead

HealthNewsReview.org from February 29, 2016
Last week there was a big shindig at the White House reviewing progress from the first year of the million-person Precision Medicine Initiative (PMI). ... This is an exciting scientific undertaking — one that merits the attention these outlets have devoted to it. But the coverage sounded mostly like cheerleading, and none of these stories included a skeptical word about the many challenges ahead and how they could thwart the initiative’s lofty objectives. I watched the webcast of the event with a critical eye and took notes as I was watching. Here are a six of the things that I thought journalists should have been thinking about and writing about as they covered the event.

Also:

- Precision medicine for understudied populations (TEDMED blog, Roxanne Dameshjou, March 2)
- Vanderbilt, Google's Verily to Launch Precision Medicine Initiative Cohort (GEN

News Highlights, February 25)
- Obama Is Using The Bully Pulpit To Set Patient Data Free (Forbes, Matthew Herper, February 25)
- President Weighs In on Data From Genes (The New York Times, February 25)
- Biden's cancer moonshot effort looks at Utah database (STAT, February 26)

The Mirage of a Citizen Data Scientist

KDnuggets, Gregory Piatetsky from March 02, 2016
... I dislike "citizen data scientist" for two reasons.

First, the word "citizen" has very misleading connotations, especially given the heated debate now in the US regarding immigration.

Having organized and attended many US conferences on data mining and data science, I observed that the majority of researchers and attendees are actually immigrants or visitors. Whether they have become US citizens like me, or are permanent residents, or visitors, has very little relation to the quality of their work.

Second, the term "citizen" Data Scientist implies that people without much training can do the work of a Data Scientist.

Start-up makes sense of transit data

Los Angeles Times from February 28, 2016
On the Internet, traffic is easily tracked. Google and Facebook have algorithms that know what users are searching for. Online retailers can monitor what shoppers are buying. Newspapers can see in real-time how many readers are viewing an article.

The same hasn't been true for traffic in the physical world, where data-gathering has been decidedly more low-tech.

Research Partners Invited!

UC-Berkeley, ENVIRONMENT and SOCIETY: DATA SCIENCES for the 21st CENTURY (DS421) from February 29, 2016
Global environmental change poses critical environmental and societal needs, and the next generation of students are part of the future solutions. This National Science Foundation Research Traineeship (NRT) in Data Science for the 21st Century prepares graduate students at the University of California Berkeley with the skills and knowledge needed to evaluate how rapid environmental change impacts human and natural systems and to develop and evaluate data-driven solutions in public policy, resource management, and environmental design that will mitigate negative effects on human well-being and the natural world. Trainees will research topics such as management of water resources, regional land use, and responses of agricultural systems to economic and climate change, and develop skills in data visualization, informatics, software development, and science communication.

In a final semester innovative team-based problem-solving course, trainees will collaborate with an external partner organization to tackle a challenge in global environmental change that includes a significant problem in data analysis and interpretation of impacts and solutions.

Everpub: reusable research, 21st century style

Tim Head, Konrad Hinsen, C. Titus Brown, Kyle Cranmer from February 28, 2016
The open source/open science community is rapidly converging around a set of technologies that will enable highly reproducible and reusable computer-aided research. These technologies include environments to encode and encapsulate dependencies, cloud compute to execute workflows, collaboration technologies that enable remixing, and text formats that enable comparison and merging.

We believe that the time is right to develop a vertical spike through the problem space, with tools to go from an empty directory to a fully rendered paper with an associated workflow that can be executed, reviewed, and remixed. We will explore a specific vertical integration of the existing tools in a focused way, find points of general technical agreement, and map areas where further work is needed. In the process, we will provide a technical basis for demos and extension. Engagement with a broad community, open discussion, and community brainstorming will build consensus about "solved" problems as well as discovering the hard knots of disagreement. Finally, open community building around this problem will inevitably yield serendipitous long-term interactions.

Renter Beware: Study Finds Craigslist Catches Barely Half of Scam Rental Listings

NYU, Tandon School of Engineering from March 01, 2016
Apartment hunters in big cities know the drill: They spot a listing for a well-priced, attractive place and make an inquiry, only to be met with demands for an instant credit check or an upfront fee to access the full listing. Savvier home hunters spot these scams immediately, but others fall through the cracks, making popular rental listing sites like Craigslist a highly lucrative spot for fraud.

A new study by researchers at the New York University Tandon School of Engineering finds that Craigslist fails to identify more than half of scam rental listings on the site’s pages and that suspicious postings often linger for as long as 20 hours before being removed—more than enough time to snare victims, especially in competitive housing markets.

The research team was led by Damon McCoy, an assistant professor of computer science and engineering, along with Elaine Shi, an assistant professor of computer science at Cornell University, and Youngsam Park, a doctoral student at the University of Maryland.

Content ID and the Rise of the Machines

Electronic Frontier Foundation from February 26, 2016
In 2007, Google built Content ID, a technology that lets rightsholders submit large databases of video and audio fingerprints and have YouTube continually scan new uploads for potential matches to those fingerprints. Since then, a handful of other user-generated content platforms have implemented copyright bots of their own that scan uploads for potential matches.

Platforms have no obligation to seek out and block infringing content, and such an obligation would entrench existing providers by stifling new platforms before they could achieve popularity. But if a large platform decides it's in its interests to evaluate every item of user material for potential infringement, the process probably must be automated—at least in part. The problem comes when humans fall out of the picture. Machines are good at many things—making the final determination on your rights isn't one of them.

Events

Symposium Examines Technology, Privacy, and the Future of Education

The Technology, Privacy, and the Future of Education symposium, hosted by the Department of Media, Culture, and Communication at NYU Steinhardt, brings together educational specialists, journalists, and academics to open a dialogue around the pedagogical, legal, and ethical repercussions of the use of new technologies in educational environments.

The symposium will take place on Friday, March 4 from 2-6 p.m. at 239 Greene Street, Floor 8.

Computer Science Colloquium: New machine learning for ubiquitous genomics and beyond

JOB TALK — James Zou, Microsoft Research New England and MIT

Large-population human datasets are being generated that can transform science and medicine. New machine learning techniques are necessary to unlock this data resource and enable discoveries. I will first survey recent advances in human population genomics, and describe new computational techniques we have developed to connect genetic and epigenetic variations to human diseases. These methods required significant innovations in latent-variable models, non-convex optimization and histogram estimation. I collaborated closely with the largest genomics consortia to apply these approaches—which are scalable and have strong mathematical guarantees—to systematically estimate the effects of mutations and to identify disease biomarkers.

Monday, March 7, at Warren Weaver Hall 1302. Refreshments at 11:15 a.m.

Presentation at 11:30 a.m.

Dean for Science Lecture - Peter Dayan

IISDM is pleased to announce that Dr. Peter Dayan, Professor of Computational Neuroscience and Director of the Gatsby Computational Neuroscience Unit at University College London, will be the 2016 speaker for the annual New York University Dean for Science Lecture in Neuroeconomics. Professor Dayan is a preeminent researcher in computational neuroscience with a primary focus on the application of theoretical computational and mathematical methods for understanding neural systems. We are looking forward to hearing about his groundbreaking work and its impact on multiple disciplines.

Monday, March 7, at 5 p.m., NYU Rosenthal Pavilion 60 Washington Square South, 10th floor

Data Science Showcase Panel

The Data Science Showcase will start with a talk by Zaid Harchaoui, on the history of AI research and its public perception, followed by a panel discussion on the future of AI with Ernest Davis, Vasant Dhar, Yann LeCun, and Gary Marcus.

Wednesday, March 9, from 4:30-7 p.m. at Kaufman Management Center, Stern School of Business, Rm KMC 5-50

Computer Science Colloquium: Automated Discovery and Learning of Complex Movement Behaviours

JOB TALK — Igor Mordatch, University of California at Berkeley

In order to create truly autonomous physical robots, understand the underlying principles behind human movement, or tell narratives in animated films and interactive games, it is necessary to synthesize movement behaviours with the same wide variety, richness and complexity observed in humans and other animals. Moreover, these behaviours should be discovered automatically from only a few core principles, and not be a result of extensive manual engineering or a mimicking of demonstrations. In this talk at the intersection of robotics, computer graphics and biomechanics, I will show work on novel trajectory and policy optimization methods that give rise to a range of behaviours such getting up, climbing, moving objects, hand manipulation, acrobatics, and various cooperative actions involving multiple characters all in a single system.

Wednesday, March 9, at Warren Weaver Hall 1302. Refreshments at 11:15 a.m. Presentation at 11:30 a.m.

Moore-Sloan Data Science Lunch Seminar Series

Wednesday, March 9 — Uri Shalit, NYU, Computer Science

This seminar will take place 1/2 hour earlier than usual, from 12:00-1:30
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 10, will be Laura Nelson (Northwestern, Kellogg)

Tyranny of the Algorithm? Predictive Analytics & Human Rights

Bernstein Institute for Human Rights Annual Conference

Tyranny of the Algorithm? Predictive Analytics & Human Rights

Monday-Tuesday, March 21-22, at NYU

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 opportunities:

- Moore-Sloan Fellows (NYU)
- Moore-Sloan Fellows (UW-Seattle)
- Moore-Sloan Fellows (UC-Berkeley)
- Data & Society post-doctoral fellowships

NYU Digital Humanities Project Showcase

We are pleased to announce an NYU Digital Humanities Project Showcase to be held on Friday April 29th at NYU's Center for the Humanities (5th floor: 20, Cooper Square). This event provides a forum for faculty, staff, and students to learn about each other's work, create connections, and start new conversations. Open to an audience from both inside and outside the university, the event will feature the work of NYU's vibrant and diverse DH community.

Members of NYU interested in sharing a DH project should fill out the application form at http://goo.gl/forms/ZJPqGGoUW7.

Deadline for applications is Monday, March 14.

Call for Participation - DML2016

Join us October 5-7, 2016 for the 7th annual Digital Media and Learning Conference. This international gathering brings together a vibrant and diverse community of innovators, thinkers, and progressive educators to delve into leading-edge topics in digital media and learning. We build connections across research, design, and practice in the service of progressive, equitable, and youth-centered approaches to learning with technology.

Deadline to apply to participate is Monday, April 4.

Neurohackweek -- September 5-9, 2016

Summer school for neuroimaging and data science ... Neurohackweek is a 5-day hands-on workshop in neuroimaging and data science, held at the University of Washington eScience Institute. Participants will learn about technologies used to analyze human neuroscience data, and to make analysis and results shareable and reproducible.

Deadline to apply: Monday, April 18

Tools & Resources

TensorFlow for Poets

Pete Warden's blog from February 28, 2016
I feel very lucky to be a part of building TensorFlow, because it's a great opportunity to bring the power of deep learning to a mass audience. I look around and see so many applications that could benefit from the technology by understanding the images, speech, or text their users enter. The frustrating part is that deep learning is still seen as a very hard topic for product engineers to grasp. That's true at the cutting edge of research, but otherwise it's mostly a holdover from the early days. There's

already a lot of great documentation on the TensorFlow site, but to demonstrate how easy it can be for general software engineers to pick up I'm going to present a walk-through that takes you from a clean OS X laptop all the way to classifying your own categories of images. You'll find written instructions in this post, along with a screencast showing exactly what I'm doing.

How to Code and Understand DeepMind's Neural Stack Machine

Andrew Trask, i am trask blog from February 25, 2016
Summary: I learn best with toy code that I can play with. This tutorial teaches DeepMind's Neural Stack machine via a very simple toy example, a short python implementation. I will also explain my thought process along the way for reading and implementing research papers from scratch, which I hope you will find useful.

Four pitfalls of hill climbing

Chris Said, The File Drawer blog from February 28, 2016
One of the great developments in product design has been the adoption of A/B testing. Instead of just guessing what is best for your customers, you can offer a product variant to a subset of customers and measure how well it works. While undeniably useful, A/B testing is sometimes said to encourage too much "hill climbing", an incremental and short-sighted style of product development that emphasizes easy and immediate wins.

Discussion around hill climbing can sometimes get a bit vague, so I thought I would make some animations that describe four distinct pitfalls that can emerge from an overreliance on hill climbing.