NYU Data Science Community news articles, blog posts and research papers for March 11, 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.

Terra Bella and Planet Labs's Most Consequential Year Yet

The Atlantic, Robinson Meyer from March 09, 2016
... This is the home of Terra Bella—the satellite company, formerly known as Skybox, that Google purchased for $500 million in June 2014. In the next 18 months, it plans to put more than a dozen new satellites into orbit. This will increase its imagery "refresh rate"—that is, how often any one spot on Earth is photographed—from one new image every three days to four to five new images per day.

Terra Bella is part of a larger group of satellite companies that promise to transform the way we see Earth. Planet Labs is another: An independent startup based in San Francisco, it estimates that in the next 12 months, it will have more than 100 satellites beaming imagery down to Earth. That will give it an almost-daily imagery refresh rate.

The ASA's statement on p-values: context, process, and purpose

American Statistical Association, The American Statistician from March 07, 2016
The American Statistical Association (ASA) has released a "Statement on Statistical Significance and P-Values" with six principles underlying the proper use and interpretation of the p-value. The ASA releases this guidance on p-values to improve the conduct and interpretation of quantitative science and inform the growing emphasis on reproducibility of science research. The statement also notes that the increased quantification of scientific research and a proliferation of large, complex data sets has expanded the scope for statistics and the importance of appropriately chosen techniques, properly conducted analyses, and correct interpretation.

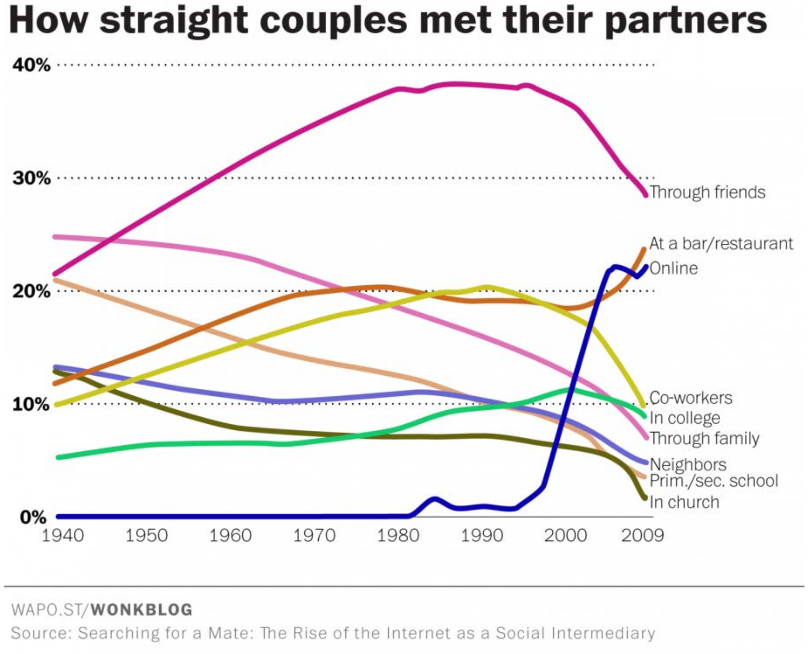
More:

- The problems with p-values are not just with p-values: My comments on the recent ASA statement (Andrew Gelman, March 7)
- Statisticians Found One Thing They Can Agree On: It's Time To Stop Misusing P-Values (FiveThirtyEight, March 7)
- The American Statistical Association statement on p-values (Psychonomic Society, Richard Morey, March 7)

There are only three ways to meet anyone anymore

The Washington Post from March 08, 2016
Last month, the BBC explained how love has changed over the years. "It was easier in the olden days. Future spouses could be found living around the corner. Or at least in your part of town," the piece said, directing attention to a series of charts. There was

one about how close to one another people who ended up together used to live (the answer is very close). Another about how the average age at which people get married has evolved (it has, as you probably know, been creeping upwards for some time).


How straight couples met their partners

WAPO.ST/WONKBLOG
Source: Searching for a Mate: The Rise of the Internet as a Social Intermediary

Announcing the 2016 Google PhD Fellows for North America, Europe and the Middle East

Google Research Blog, Michael Rennaker from March 10, 2016
Google created the PhD Fellowship program in 2009 to recognize and support outstanding graduate students doing exceptional research in Computer Science and related disciplines. Now in its eighth year, our fellowship program has supported hundreds of future faculty, industry researchers, innovators and entrepreneurs.

Reflecting our continuing commitment to supporting and building relationships with the academic community, we are excited to announce the 39 recipients from North America, Europe and the Middle East. We offer our sincere congratulations to Google's 2016 Class of PhD Fellows.

The Artist Archives Project - David Wojnarowicz

NYU Center for the Humanities from March 09, 2016
The Artist Archives Project develops information resources for the display and conservation of contemporary art. The initiative responds to a growing need for museum and archive professionals to work with artists in documenting their production methods, and building knowledge for future treatment and re-activation of their work. The first undertaking of the Artist Archives Project is to create an information resource devoted to the multidisciplinary artist/activist David Wojnarowicz. The resource will contain technical and historical information about the artist's films, video, photography, paintings, drawings, and performance work. The artist's personal archive

in the Fales Downtown Collection at NYU serves as a primary source for research.

The project takes a digital humanities approach by assigning equal emphasis to the content and the database / web portal design.

Pushback -- The 2016 campaign is putting the most influential political-science book in recent memory to a stiff test

The Economist from March 05, 2016
Of all the theories to explain the unexpected success of Donald Trump's presidential campaign, this, surely, is the most novel. Forget about a disaffected working class buffeted by globalisation and automation, pent-up racial resentments finding an outlet or the advent of the 24-hour news cycle. No: in the assessment of Daniel Drezner, a professor of international politics at Tufts University, it's the political scientists who are to blame.

Also:

- How politicians should use Twitter bots. (Slate, Tim Hwang and Samuel Woolley, March 8)

The Code That Runs Our Lives

YouTube, The Agenda with Steve Paikin from March 03, 2016
From searching on Google to real-time translation, millions of people use deep learning every day, mostly without knowing it. It's a form of artificial intelligence designed to mimic the human brain. Geoffrey Hinton is a professor in the department of computer science at the University of Toronto. His work on deep learning has been snapped up by Google and is now being used to power its search engine. He joins The Agenda to discuss deep learning and the future of artificial intelligence. [video autoplays, 15:00]

Google's AI machine v world champion of 'Go': everything you need to know

The Guardian from March 09, 2016
The Google DeepMind challenge match will pit the world's top player of the ancient Chinese board game against the world's most sophisticated Artificial Intelligence programme. Here is everything you need to know about this clash between advanced technology and old-fashioned human wit.

Also:

- All Matches - Google DeepMind Challenge Match: Lee Sedol vs AlphaGo (YouTube, Deep Mind)
- Google's AI Wins Pivotal Second Game in Match With Go Grandmaster (WIRED Business, March 10)

Comprehensive, Open Cancer Data Repository to Tap Cognitive Insights from Watson

Scientific Computing, New York Genome Center from March 02, 2016

At the White House Precision Medicine Initiative Summit on February 25, 2016, the New York Genome Center and IBM announced that they are collaborating to create a comprehensive and open repository of genetic data to accelerate cancer research and scale access to precision medicine using cognitive insights from IBM Watson. Analyzing this data alongside the medical community's growing knowledge about cancer could help accelerate the ability of doctors to deliver personalized treatment to individual patients.

IBM and New York Genome Center are working together to build the capacity to house the contributed data, train Watson's cognitive computing capabilities for genomic analysis and enable the Center's member institutions and other research collaborators to sequence and analyze DNA and RNA from patients' tumors.

Accelerating Discovery with New Tools and Methods for Next Generation Social Science

DARPA from March 04, 2016
To begin to assess the research opportunities provided by today's web-connected world and advanced technologies, DARPA today launched its Next Generation Social Science (NGS2) program. The program aims to build and evaluate new methods and tools to advance rigorous, reproducible social science studies at scales necessary to develop and validate causal models of human social behaviors. The program will draw upon and build across a wide array of disciplines—including social sciences like sociology, economics, political science, anthropology, and psychology, as well as information and computer sciences, physics, biology and math.

Events

JupyterDays Boston 2016

We're excited to announce our next event! Join us for JupyterDays Boston in Cambridge, MA on March 17-18, 2016. The event will be hosted at the Harvard Law School, Wasserstein Hall, Milstein East A & B, and organized by O'Reilly Media, Harvard-Smithsonian Center for Astrophysics Library and the Harvard Law School Library. This event is being co-organized with core Jupyter/IPython project contributors, some of whom will be present.

Thursday-Friday, March 17-18, at Harvard

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

Political Analytics 2016

Political Analytics is a one-day conference at Harvard University featuring top minds from media, politics, and academics. We are starting an exciting conversation about the growing role of data and analytics in determining the winners and losers in politics. Our goal is to promote new methods, technology, and discussions for the improved analysis of politics.

The conference is open to the public and will feature a variety of discussions and dynamic presentations by leaders in the field.

Friday, April 1, at Harvard University

Machine Learning in Finance Workshop 2016

The Data Science Institute at Columbia University and Bloomberg LP are pleased to announce a workshop on "Machine Learning in Finance". The workshop will be held at Columbia University under the auspices of the Financial and Business Analytics Center, one of the constituent centers in the DSI, and the Center for Financial Engineering.

Friday, April 1, at Columbia University

3rd Annual Big Data in Biology Summer School.

The 2016 Big Data in Biology Summer School offers eleven intensive courses that span general programming, high throughput DNA and RNA sequencing analysis, proteomics, and computational modeling. These courses provides a unique hands-on opportunity to acquire valuable skills directly from experts in the field. Each course will meet for three hours a day for four days (either in the morning or in the afternoon) for a total of twelve hours.

Also: 9th Annual Summer Statistics Institute — The 2016 Summer Statistics Institute offer 26 courses that span introductory statistics, statistical software, statistical methods and statistics applications. Each course will meet for four half-days, either mornings or afternoons, for a total of twelve hours.

Both training events take place at University of Texas at Austin on May 23-26.

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.

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.

IEEE VIS 2016

All conferences at IEEE VIS allow both single-blind (not anonymized) as well as double-blind (anonymized) submissions. Double-blind submissions are allowed for those authors who want to submit their work anonymously. Therefore, those authors should NOT include their name or institution on the cover page of the initial submission, and should make an effort to ensure that there is no revealing information in the text (such as obvious citations to authors' previous work, or making acknowledgments to colleagues of long standing). Authors should also avoid posting their submitted manuscript on the web until the final notification date. To reiterate, the choice of complete anonymity (i.e., single or double-blind) is optional. Authors can reveal their names and affiliations in the first round of the review cycle if they choose not to anonymize their work.

Deadline for abstract submission is Monday, March 21. Full papers are due Thursday, March 31.

Moore-Sloan Seed Grants

The Moore-Sloan Data Science Environment at NYU is announcing a unique funding opportunity, open to all NYU faculty, that aims to bring together data scientists and domain scientists to foster collaborations and generate new ideas. For details, please see the Seed Grant Announcement. If you are interested in applying for the Seed Grant, you must first submit an online letter of interest by March 25, 2016. You must then attend the "open dating session" to pitch your ideas to the other attendees. The open dating session is scheduled from 4pm to 6pm on April 4, 2016 and the location is to be announced. The PIs must submit a formal proposal (1-2 pages) within two weeks of the "open dating" session. The proposals will be reviewed by the MSDSE Methods Working Group based on the project impact, innovation and scientific merits of the proposal. The results will be announced by May 13, 2016.

Deadline for submitting this online letter of interest is Friday, March 25.

CDS News

CDS Faculty Interview: Arthur Spirling

NYU Center for Data Science from March 10, 2016
Communications between embassies, government entities, and diplomats take the form of classified diplomatic cables. In 2010, over 150,000 of these cables were released by Wikileaks, a nonprofit organization that publishes classified government documents. The effect of the leak was twofold; not only did previously secret information become readily available, but now, the general population could glimpse into the inter-workings of diplomacy.

Last year, Arthur Spirling, an Associate Professor of Politics and Data Science at New York University, co-authored a paper titled, “Dimensions of Diplomacy,” regarding his research on these Wikileaks cables. We got the chance to ask him a few questions about his research, his findings, and the nature of governmental secrecy.

Organising Astro Hack Week Part 1: How to organise a hack week

Daniela Huppenkothen, Daniela's blog from March 06, 2016
This will be a series of posts about organizing Astro Hack Week 2015. What is Astro Hack Week, you might rightfully ask? Let’s start with that, before we go into the organization in detail.

It’s a five-day workshop that’s part academic summer school—with tutorials and lectures on cutting-edge data analysis topics and methods—and part hackathon, with free time for teams to self-organize and hack (work quickly, but productively) on their own data analysis problems. ... But this post is not about all the fun stuff that happens at a hack week. This post is about the basic organisation of Astro Hack Week, and will be followed by followed by more specific posts about different aspects of it.

Tools & Resources

Applying Fair Use - Copyright - Research Guides at New York University

New York University from February 26, 2016
In order to balance the interests of the creators of copyrighted works with the public's ability to benefit from those works, copyright law includes the exemption of Fair Use.

Fair use allows limited use of copyrighted material without permission for purposes such as criticism, parody, news reporting, research and scholarship, and teaching.

A book for all: Data Management for Researchers by Briney

Christie Bahlai, Practical Data Management for Bug Counters blog from March 07, 2016
If you’re a data management enthusiast like me (yes, we exist, and there’s actually a bunch of us), you’ve probably head about Kristin Briney’s Book, *Data Management for Researchers*. I received a copy for review a few months ago, and have been taking my time to savor it. But if you’ve heard of this book, chances are that although you’ll certainly find aspects of it useful, you’re probably the metaphorical choir that we, the data managers, are preaching to. You might even argue that there are lots of data management resources out there- why a book? But Briney does something unique here, and I have been enthusiastic to recommend it to everyone around me.

leaf at f0b11961b5a0649544a1101b960c670a0bebf57c: The Hacker's Machine Learning Engine

GitHub, autumnai from March 07, 2016
Leaf is a Machine Intelligence Framework engineered by software developers, not scientists. It was inspired by the brilliant people behind TensorFlow, Torch, Caffe, Rust and numerous research papers and brings modularity, performance and portability to deep learning. Leaf is lean and tries to introduce minimal technical debt to your stack.

Leaf is a few months old, but thanks to its architecture and Rust already one of the fastest Machine Intelligence Frameworks in the world.

How to use The Guardian's API to download article data for content analysis (in Python 3.x)

GitHub, dannguye from March 08, 2016

The Guardian offers an API as deep and robust as the New York Times Article API when it comes to content analysis.

The Guardian's API offers more than "1.7 million pieces of content", with published items as far back as 1999. You can register as a developer here, which gets you 5,000 API hits a day and an API key.

Machine Learning Meets Economics

MLDB.ai, Nicolas Kruchten from January 27, 2016
The business world is full of streams of items that need to be filtered or evaluated: parts on an assembly line, resumés in an application pile, emails in a delivery queue, transactions awaiting processing. Machine learning techniques are increasingly being used to make such processes more efficient: image processing to flag bad parts, text analysis to surface good candidates, spam filtering to sort email, fraud detection to lower transaction costs etc.

In this article, I show how you can take business factors into account when using machine learning to solve these kinds of problems with binary classifiers. Specifically, I show how the concept of expected utility from the field of economics maps onto the Receiver Operating Characteristic (ROC) space often used by machine learning practitioners to compare and evaluate models for binary classification. I begin with a parable illustrating the dangers of not taking such factors into account.

Quick Intro to NMF (the Method and the R Package)

Norm Matloff, Mad (Data) Scientist blog from March 05, 2016
Nonnegative matrix factorization (NMF) is a popular tool in many applications, such as image and text recognition. If you’ve ever wanted to learn a little bit about NMF, you can do so right here, in this blog post, which will summarize the (slightly) longer presentation here. The R package NMF will be used as illustration.