NYU Data Science Community features journalism, research papers, events, tools/software, and jobs for April 29, 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

Government to establish council for data science ethics

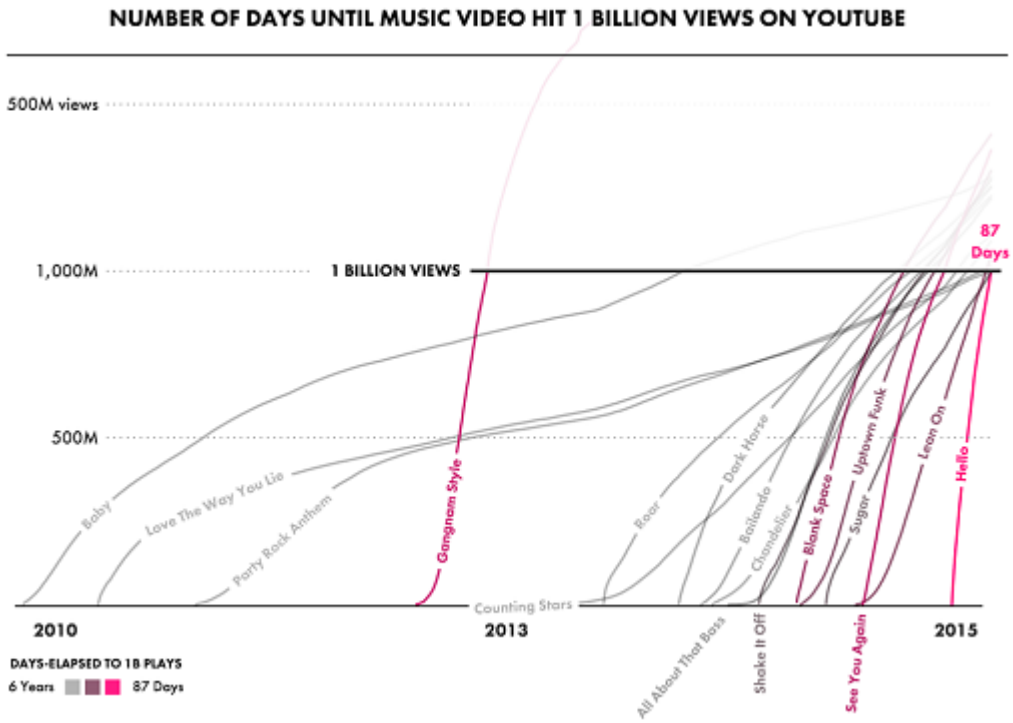
PHG Foundation from April 27, 2016
The **UK Government** is to set up a Council of Data Science Ethics. The development comes in response to recommendations from the Science and Technology Committee's report The big data dilemma published earlier this year.

The Council of Data Science Ethics will be established within the Alan Turing Institute as a means of "addressing the growing legal and ethical challenges associated with balancing privacy, anonymisation, security and public benefit".

Also:
- Case Study: The Ethics of Using Hacked Data: Patreon's Data Hack and Academic Data Standards (April 4, **Council for Big Data, Ethics, and Society**; **Nathaniel Poor** & **Roei Davidson**)

When Music Becomes Popular, Faster

Polygraph, YouTube Music from March 25, 2016



Here's a fun thing: there are 17 music videos that have hit 1 billion views, EVER. And

15 of the 17 songs crossed 1 billion views in the last year.

Also:
- Acoustic Voxels: Computational Optimization of Modular Acoustic Filters (April 27, CreativeAI, **Columbia University Computer Science**)
- **New York University's Music Experience Design Lab** Teams Up with Soundtrap Online Music Recording Studio (April 26, Business Wire, MathScienceMusic.org)

Science AMA Series: I'm George Church, professor at Harvard and MIT, founder of PersonalGenomes.org.

reddit.com/r/science from April 18, 2016
Hi Reddit! I'm **George Church** and my lab is developing technologies for genome sequencing, gene editing, and DNA nanotechnology (bio).

One area that has attracted a lot of attention recently is the CRISPR technology for editing the genetic information in living cells, a sort of 'nano-surgery' that can be used to treat genetic disease at the root cause.

Movidius Unveils Artificial Intelligence on a Stick

PC Magazine from April 28, 2016
Hot on the heels of a new artificial intelligence-capable thermal imaging camera, chip maker Movidius has yet another AI implementation up its sleeve: a USB stick that can allow pretty much any Linux computer to handle advanced neural networks, one of the building blocks of AI. The San Mateo, Calif.-based company announced the device, called the Fathom Neural Compute Stick, today along with its Fathom deep-learning software framework. Together, the two products will enable device manufacturers to move AI processing from the cloud to native deployment in end-user devices.

Stealing Google's Coding Practices for Academia

Dave Andersen, Dave's Data blog from April 27, 2016
I'm spending the year in Google's Visiting Faculty program. ... One of my explicit goals was to steal ideas from Google that I could feed back into teaching and mentoring.

Exponential growth of R's open source community threatens commercial competitors

TechRepublic, Matt Asay from April 21, 2016
With more than 2 million users and developers, how can proprietary vendors stand against the R programming language and software environment's open source community?

Google to encourage employee startups

San Jose Mercury News from April 24, 2016
Add founding a new company to the list of things **Google** employees can do without leaving the office. The Internet search giant, which already famously offers on-site perks such as all-you-can-eat snacks, massages and fitness centers, is reportedly launching a new startup incubator that lets entrepreneurially minded employees pursue their dreams

without leaving the mother ship. Dubbed "Area 120," the incubator will be based in one of Google's new San Francisco buildings, according to tech news site The Information, which cited anonymous "people familiar with the project."

Also:
- Race For AI: Google, **Facebook**, **Amazon**, **Apple** Grab Artificial Intelligence Startups (April 10, CB Insights)
- Google believes its superior AI will be the key to its future (April 21, The Verge)

Interacting with Robots? Mais Oui!

NYU Tandon School of Engineering from April 19, 2016
The French-American Doctoral Exchange Seminar (FADEx), a program organized by the **Office for Science & Technology at the Embassy of France** in the United States, is aimed at encouraging collaboration among American and French doctoral students with similar research interests and forging what could turn into long-lasting scientific partnerships.

The theme of FADEx 2016 is cyber-physical systems, which integrate computation, networking, and physical processes and are used in the fields of transportation, manufacturing, healthcare, and more.

Also, in University data science:
- Advancing ingenuity (April 27, **Harvard John A. Paulson School of Engineering and Applied Sciences**)
- Is Data Science a Liberal Art? (April 25, SmartData Collective, **George Mount**)
- Data Science Across Disciplines and the program's Final Poster Session (April 26, **University of Illinois**, University Library)

Events

MLTalks Series: Helen Margetts in Conversation with Ethan Zuckerman

How does the changing use of social media affect politics? In her recent book, *Political Turbulence*, **Helen Margetts** and colleagues **Peter John**, **Scott Hale** and **Taha Yasseri** show how social media are now inextricably intertwined with the political behavior of ordinary citizens, and exert an unruly influence on the political world. ... In this talk, Professor Margetts will discuss the implications of these findings both for political science research and the future of the modern state.

Boston, MA. Tuesday, May 3, starting at 2 p.m., **MIT Media Lab**, 3rd Floor Atrium

BIDS Spring 2016 Data Science Faire

This year's Data Science Faire will feature a wide variety of poster/demo exhibits from students and researchers on campus as well as a series of data science–related lightning talks from BIDS fellows. The event will culminate with keynote address from **Lucas Merrill Brown**, Data Scientist and Digital Expert at the **US Digital Service at the White House**.

Berkeley, CA. Tuesday, May 3, starting at 1:30 p.m., 190 Doe Library

Moore-Sloan Data Science Lunch Seminar Series

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. This week we hear from **Daphna Harel** of PRIISM at **NYU Steinhardt**.

New York, NY. Wednesday, May 4, from 12:30 - 1:30 p.m. at the **NYU Center for Data Science**, 726 Broadway, 7th floor.

Workshop on Networks

The workshop will cover Networks, Random Graphs and Statistics with a quality speaker list.

New York, NY. Wednesday-Friday, May 4-6, at **Columbia University** and at the **Union Theological Seminary** near Columbia.

Text as Data Speaker Series

The NYU 'Text-as-Data' speaker 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. This week we hear from **Mark Dredze** (**Johns Hopkins University Bloomberg School of Public Health**) on *Topic Models for Identifying Public Health Trends*.

New York, NY. Thursday, May 5, starting at 4 p.m. in room 217, 19 West 4th St (unless otherwise noted).

Health Datapalooza

Health Datapalooza is a national conference focused on liberating health data, and bringing together the companies, startups, academics, government agencies, and individuals with the newest and most innovative and effective uses of health data to improve patient outcomes.

Washington DC. Sunday-Wednesday, May 8-11

Workshop III: Cultural Patterns: Multiscale Data-driven Models (Schedule)

The proliferation of cultural data has given data-driven approaches a significant edge in modeling various cultural phenomena. This workshop focuses on such approaches that make use of mathematical tools in machine learning, data mining, network science, and computational social science. We are particularly interested in presenting methods, both normative and descriptive, that offer a gestalt or structure-first approach to culture analysis and that provide a multi-layered summarization of these phenomena suitable for exploration at multiple scales. These models are applied to various datasets such as social and information networks, social media, narrative and story detection in texts, group dynamics or behavior, and collaboration and competition leading to emergent

behavior.

This workshop will include a poster session; a request for posters will be sent to registered participants in advance of the workshop.

Los Angeles, CA. Monday-Thursday, May 9-13, at **UCLA IPAM**

Deadlines

2016 Workshop on Human Interpretability in Machine Learning

The 2016 Workshop on Human Interpretability in Machine Learning (WHI 2016), held in conjunction with ICML 2016, will bring together researchers who study the interpretability of predictive models, develop interpretable machine learning algorithms, and develop methodology to interpret black-box machine learning models (e.g., post-hoc interpretations).

New York, NY. Deadline for submissions is Sunday, May 1.

Call for Software Carpentry Foundation Subcommittees and Task Forces

The 2016 Steering Committee would like to encourage and invite the community to propose new initiatives in the form of subcommittees (a standing group for ongoing activities) and task forces (an ad hoc group focused on a finite task). Read on to learn more about existing initiatives and for information about how to propose a new initiatives that will shape our community.

BITSS and BIDS Collaboration: Call for Reproducible Workflows

BITSS and the Reproducibility Working Group at the **Berkeley Institute for Data Science** are collaborating on an edited volume of reproducible workflows in the social sciences, and we are looking for submissions.

Also, see Call for Reproducibility Workflows by **Cyrus Dioun** and **Garret Christensen** at the *Bad Hessian* blog.

BITSS: Research Transparency 2-day workshop

There is growing interest in research transparency and reproducibility across the social sciences. This workshop is a crash course on the problems of publication bias, inability to replicate research, and specification searching (or p-hacking, among other names) that have heretofore caused researchers problems. We will cover recent methodological progress in this area, including study registration, pre-analysis plans, disclosure standards, and open sharing of data and materials, drawing on experiences in economics, political science, and psychology, as well as other social sciences.

Ann Arbor, MI. Tuesday-Wednesday, July 5-6, at the **University of Michigan**. Deadline to apply is Sunday, May 15.

Call for Abstracts - The 2016 Conference on Complexity Systems

The Conference on Complex Systems (CCS) has become a major venue for the Complex Systems community since 2003. After last year success in USA, we are now back in Europe. AMSTERDAM CCS 2016, will be the major international conference and event for complex systems and interdisciplinary science.

Amsterdam, The Netherlands. Deadline for abstracts' submissions is Sunday, May 15.

Data Science Game

Data Science Game is a French association run by volunteer data scientists and students, supported by **Paris-Saclay University**. Each year, we organize an international data science competition for students.

Paris, France. Deadline to register is Tuesday, May 31.

Complex Networks 2016

The International Workshop on Complex Networks and their Applications aims at bringing together researchers from different scientific communities working on areas related to complex networks.

Two types of contributions are welcome: theoretical developments arising from practical problems, and case studies where methodologies are applied. Both contributions are aimed at stimulating the interaction between theoreticians and practitioners.

University of Milan, Milan, Italy. Deadline for submissions is Monday, September 5.

Tools & Resources

15 Must Read Books for Entrepreneurs in Data Science

AnalyticsVidhya from April 25, 2016
... The books listed below gives immense knowledge and motivation in technology arena. Reading these books will give you the chance to live many different entrepreneurial lives. Take them one by one. Don’t get overwhelmed. I’ve displayed a mix of technical and motivational books for entrepreneurs in data science.

Sequence-to-sequence model with LSTM encoder/decoders and attention

GitHub - harvardnlp from April 24, 2016
Torch implementation of a standard sequence-to-sequence model with attention where the encoder-decoder are LSTMs. Also has the option to use characters (instead of input word embeddings) by running a convolutional neural network followed by a highway network over character embeddings to use as inputs.

Keras as a simplified interface to TensorFlow: tutorial

The Keras Blog from April 24, 2016
If TensorFlow is your primary framework, and you are looking for a simple & high-level model definition interface to make your life easier, this tutorial is for you.

Keras layers and models are fully compatible with pure-TensorFlow tensors, and as a result, Keras makes a great model definition add-on for TensorFlow, and can even be used alongside other TensorFlow libraries. Let's see how.

Druid Query Optimization with FIFO: Lessons from Our 5000-Core Cluster

Metamarkets, Charles Allen from April 25, 2016
A large strength of using Druid as a data store and aggregation engine is its ability to horizontally scale. Whenever more data is in the system, or whenever faster compute times are desired, it is simply a matter of throwing more hardware at the problem, and Druid auto-detects, and auto-balances its workloads. At **Metamarkets** we are currently ingesting over 3M events/ second (replicated) into our Druid cluster and have multiple hundreds of historical nodes serving this data across multiple tiers. ... the balancing algorithms in Druid are not perfect, which means segments will not be perfectly balanced across the cluster, but will usually be “good enough” for most use cases. As such, many clusters will end up with some degree of over-committing cores to number of data segments that are needing to be scanned. This leads to an interesting aspect of Druid’s processing queue. The release of Druid 0.9.0 adds the feature flag druid.processing.fifo. Let’s take a look at where this flag comes from and how it should be used.

Vega-Lite

University of Washington Interactive Data Lab from April 12, 2016
Vega-Lite is a high-level visualization grammar. It provides a concise JSON syntax for supporting rapid generation of visualizations to support analysis. Vega-Lite specifications can be compiled to Vega specifications.

How Kalman Filters Work, Part 1

An Uncommon Lab, Tucker McClure from April 26, 2016
... When performed as part of an algorithm, this type of thing is called recursive state estimation. Unfortunately, only a small fraction of mechanical engineers, electrical engineers, and data scientists receive any formal education on the subject, and even fewer develop an intuitive understanding for the process or have any knowledge about practical implementation. While there are very many good books on the math behind it and the details of how to apply it to certain specific problems, this article will take a different approach. We’ll focus on developing:

- an intuition for recursive state estimation,
- a broad knowledge of the strongest and most general types,
- and a good idea of the implementation details.

Careers

Review and Edit, Writing Workshops Planned for JSM

Amstat News from April 01, 2016
Two workshops are planned for JSM 2016 in Chicago that are designed for new researchers to develop capabilities for research publication. The Writing Workshop for New Researchers continues the series of workshops in which each participant receives

individual mentoring by an experienced journal editor. Additionally, the new Review and Edit Workshop will be offered by a team of executive editors of leading journals to junior researchers as they take up new responsibilities for reviewing technical articles and joining the ranks of associate editors. **Chicago, IL.**

Detectica - Jobs

Detectica from April 24, 2016
Detectica employs state-of-the-art machine learning to build trader surveillance systems used to identify malicious and collusive trading activity. We analyze the trading patterns and the electronic communications to recognize early and conclusively cases of fraud. We are founded by and comprised of some of the best minds in data science and machine learning. **New York, NY.**

Jobs · Hammer Lab -- Operations Lead

Icahn Institute at Mount Sinai, Hammer Lab from April 25, 2016
Hammer Lab seeks an operations lead in NYC to facilitate our work in cancer immunotherapy. An ideal candidate will have an interest in and some knowledge of biomedicine, software development, and data analysis, education in business or operations, and work experience in both engineering/product development and sales/marketing/business development. If you're organized, outgoing, and love to write, you'll enjoy the work and be a critical contributor to the success of the lab. **New York, NY.**

Working at Twitter - Data Scientist (Content Insights)

Twitter from April 26, 2016
Join the Pulse team and work with a diverse set of partners in the company including product and media partners, engineers and data scientists. Your work will influence how we measure content about people, events, and topics in our platforms. In every decision that you influence, you will see our improve and be more valuable to **Twitter** users. **San Francisco, CA.**

MSR's Social Media Collective is looking for a 2015-16 Research Assistant

Microsoft Research, Social Media Collective from April 27, 2021
Starting in July 2016 the **MSR Social Media Collective** will welcome a new RA to work directly with **Nancy Baym**, **Kate Crawford**, **Tarleton Gillespie**, and **Mary L. Gray**. An appropriate candidate will be a self-starter who is passionate and knowledgeable about the social and cultural implications of technology. Strong skills in writing, organization and academic research are essential, as are time-management and multi-tasking. Minimal qualifications are a BA or equivalent degree in a humanities or social science discipline and some qualitative research training. A Masters degree is preferred. **Cambridge, MA.**

Library Research Data Specialist

Caltech Library from April 29, 2016
The **Caltech Library** is seeking a Research Data Specialist to work with the Caltech community to develop tools and services in support of the research data curation

lifecycle, from creation to publication, dissemination and reuse. Ideal candidates will have strong computational skills, a deep appreciation for and knowledge of data management best practices, and a passion for working with individuals and teams to define and develop this new programmatic direction for the Caltech Library. **Pasadena, CA.**