NYU Data Science Community features journalism, research papers, events, tools/software, and jobs for May 20, 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

Google's CEO sums up his AI vision: "Hi. How can I help?"

CNET from May 18, 2016
Sundar Pichai, who takes the I/O stage for the first time since becoming CEO in October, says he's on a "journey" from mobile to AI. Good thing he has an assistant.

Also, from Google I/O:
- "Aw, so cute!": Allo helps you respond to shared photos (May 18, Google Research Blog, Ariel Fuxman)
- Google's Making Its Own Chips Now. Time for Intel to Freak Out (May 19, WIRED, Business)
- Android N, Daydream VR, Google Home and more: Everything announced at Google I/O 2016 (May 18, CNET)

Machine learning, A.I to follow on the priority list for businesses: SAP

ZDNet from May 17, 2016
SAP believes the next technology adoption phase for businesses will be around how they can use intelligent applications to assist them with their operations.

Also, in enterprise software:
- Salesforce CEO: I see an 'AI-first world' (May 18, Business Insider)

How Big Data and Data Analysis Are Changing Our Understanding of Cities

CityLab, Richard Florida from May 18, 2016
There's been no shortage of hype about the relationship between cities and data, especially so-called big data. For large numbers of tech companies, cities, and even a growing number of urbanists, data promises to solve all manner of urban problems, from predictive policing to improving traffic flow to promoting energy efficiency.

An even bigger potential role for new kinds of data lies in helping researchers and policy-makers better understand how cities and neighborhoods grow and evolve—but only if done right.

Also: in interdisciplinary urban data research:
- The fourth urban revolution (May 15, Pittsburgh Post-Gazette, Opinion)
- Use our infographics to explore the rise of the urban planet (May 19, Science, Latest News)

Social-sciences preprint server snapped up by publishing giant Elsevier

Nature News & Comment from May 17, 2016
After trying without success more than a decade ago to set up preprint servers — where academics share their papers before peer review — science-publishing giant **Elsevier** is now buying one. It is paying an undisclosed sum for the Social Science Research Network (SSRN), one of the world's most popular repositories of research in economics, law and the social sciences.

Also, in preprints:
- It's the Data, Stupid: What Elsevier's purchase of SSRN also means (May 18, Savage Minds blog, Christopher Kelty)
- Elsevier Acquires SSRN (May 17, The Scholarly Kitchen blog, Roger C. Schonfeld)
- Preprints for the life sciences (May 20, Science; Jeremy Berg et al.)

Better models for brain disease

Proceedings of the National Academy of Sciences; Helen Shen from May 17, 2016
Predicting outcomes and, crucially, developing psychiatric drugs has proven exceedingly difficult in recent decades. Inadequate animal models have been a major stumbling block, researchers say. ... "It's a very exciting time," says **Guoping Feng**, a neuroscientist at the **Massachusetts Institute of Technology** (MIT) in Cambridge, Massachusetts. "Between the technology development and the genetic findings, this is the first time that we've been able to begin digging deep into the causes and neurobiology of these disorders."

[1605.04462] Natural Language Processing for Mental Health: Large Scale Discourse Analysis of Counseling Conversations

arXiv, Computer Science > Computation and Language; Tim Althoff, Kevin Clark, Jure Leskovec from May 14, 2016
Mental illness is one of the most pressing public health issues of our time. While counseling and psychotherapy can be effective treatments, our knowledge about how to conduct successful counseling conversations has been limited due to lack of large-scale data with labeled outcomes of the conversations. In this paper, we present a large-scale, quantitative study on the discourse of text-message-based counseling conversations.

Imker to Lead Illinois Efforts in Multi-Institution Data Curation Network Funded by Sloan Foundation

University of Illinois Library from May 12, 2016
Heidi Imker, Director of the Research Data Service and Associate Professor at the **University of Illinois** Library, will lead local participation in a new Data Curation Network funded by the **Sloan Foundation**. The grant will enable six institutions, including Illinois, to pilot a "network of expertise" model for data curation services.

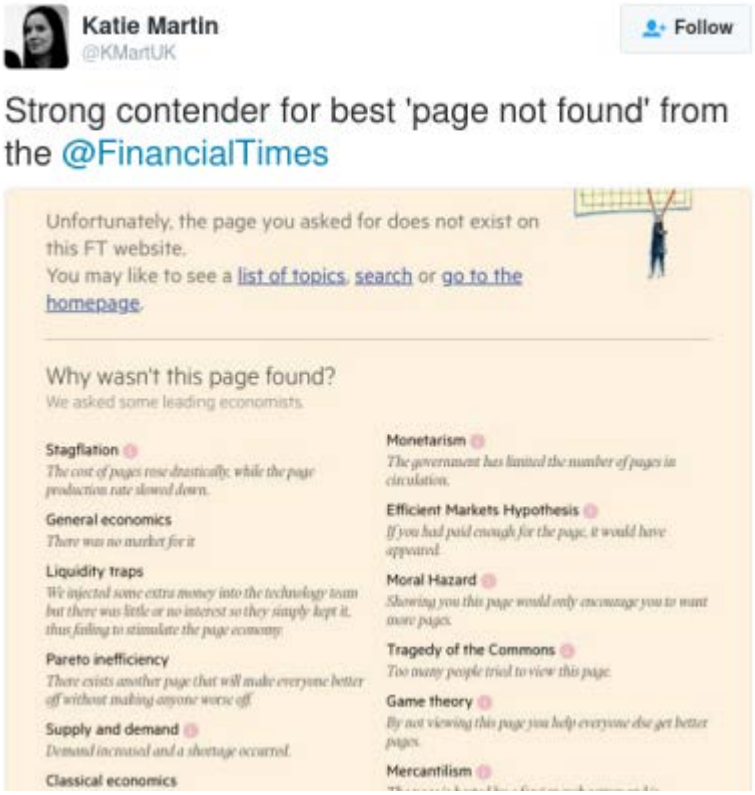
Bossy girls, Parser McParseface, and why deep learning is not just another fad

Pete Warden's blog from May 15, 2016
Deep learning is different, and I believe this fervently because I've seen the approach deliver record-beating results in practical applications across an amazing variety of different problems. That's why TensorFlow is so important to me personally. ... It's

also why I was over the moon to see another **Google** research team release Parsey McParseface!

Tweet of the Week: Strong contender for best 'page not found' from the @FinancialTimes

Twitter, Katie Martin from May 07, 2016



Events

Program - NLP+CSS: Workshops on Natural Language Processing and Computational Social Science

Language is perhaps the most salient outcome of complex social processes. We do not expect teenagers to speak like senior citizens, and we recognize the mutual dependency between language and social factors. Although this interdependence is at the core of models in both natural language processing (NLP) and (computational) social sciences (CSS), these two fields still exist largely in parallel, holding back research insights and potential applications.

This workshops aims to advance the joint computational analysis of social sciences and language by explicitly connecting social scientists, network scientists, NLP researchers, and industry partners.

Hannover, Germany May 22, 2016, Hannover, Germany. This is one of two companion workshops on NLP+CSS in 2016.

Artificial Intelligence: Law and Policy

The **University of Washington School of Law** is delighted to announce a public workshop on the law and policy of artificial intelligence, co-hosted by the **White House** and UW's Tech Policy Lab. The event places leading artificial intelligence experts from academia and industry in conversation with government officials interested in developing a wise and effective policy framework for this increasingly important technology. The event is free and open to the public but requires registration.

Seattle, WA Tuesday, May 24, at the University of Washington School of Law

May 25, Clinical Machine Learning Talk - Marzyeh Ghassemi

Marzyeh Ghassemi is a PhD student in the Clinical Decision Making Group (MEDG) in MIT's Computer Science and Artificial Intelligence Lab (**CSAIL**) supervised by Prof. Peter Szolovits. Her research uses machine learning techniques and statistical modeling to predict and stratify relevant human risks.

New York, NY Wednesday, May 25, at 11 a.m. in the 715 Broadway 12th floor large conference room (intersection of Broadway and Washington Place)

CUSP Research Seminar Series | June 15

Please join us for a research seminar with **Jeff Jonas**, **IBM** Fellow and Chief Scientist of Context Computing.

Brooklyn, NY on Wednesday, June 15, at 11 a.m., Jacobs Seminar Room at the Center for Urban Science and Progress (1 Metrotech, 19th Floor)

Information+ Conference Tickets

The inaugural Information+ conference will bring together researchers and practitioners in information design and information visualization to discuss common questions and challenges in these rapidly changing fields.

Vancouver, British Columbia Thursday, June 16, at Emily Carr University of Art + Design.

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

Any data management system needs to work together with people, whose needs determine the goals for the system, and who must provide the input and who need to work effectively with the output. Data management systems will work much better when they take account of the cognitive and physiological characteristics of the people involved. Recent technology trends (such as touch screens, motion detection, and voice recognition) are widening the possibilities for users to interact with systems, and many information-provision industries are shifting to personalized processing to better target their services to the users' wishes. 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. ... HILDA intends to be a forum where people from varied communities engage with one another's ideas. We are keen to have submissions that present initial ideas and visions, just as much as reports on early results, or reflections on completed projects.

San Francisco, CA Sunday, June 26, 2016, Co-located with SIGMOD 2016

SESYNC to Co-Sponsor Workshop on Social Network Analysis

The **National Socio-Environmental Synthesis Center** (SESYNC) is pleased to co-sponsor the following summer workshop: *Social Network Analysis: An Introduction with an Emphasis on Application in R*.

Annapolis, MD Monday-Friday, June 27–July 1, 2016 at the National Socio-Environmental Synthesis Center (SESYNC); Instructor: Dr. Lorien Jasny, University of Exeter and former SESYNC postdoctoral fellow.

Deadlines

KDD 2016 Workshop on Large Scale Sports Analytics

For the 3rd successive year, we will be running the KDD workshop on Large-Scale Sports Analytics. The objective of this workshop is to bring together researchers and analysts from academia and industry who work in sports analytics, data mining and machine learning. We hope to enable meaningful discussions about state-of-the-art in sports analytics research, and how it might be improved upon.

San Francisco, CA Sunday, August 14. Workshop precedes KDD 2016

Deadline for submissions is Friday, May 27.

O'Reilly Artificial Intelligence Conference, September 26 - 27

The O'Reilly Artificial Intelligence Conference Call for Speakers is open.

New York, NY on Monday-Tuesday, September 26-27.

Deadline to apply to speak is Monday, June 6.

Text as Data 2016 | Seventh Annual New Directions in Analyzing Text as Data

We invite you to submit a paper to be presented, or to submit your name as discussant or as an attendee, at the seventh annual research conference on “New Directions in Analyzing Text as Data” that will be held at **Northeastern University** on October 14-15, 2016. This two-day conference draws together scholars from many different universities and disciplines to discuss developments in text as data research.

Boston, MA Friday-Saturday, October 14-15 at Northeastern University.

Deadline for abstract submissions is Monday, June 20.

D4GX 2016: Program Committee Announced, Call for Papers Open

D4GX is **Bloomberg's** Data for Good Exchange event.

Our “call for papers” is open and we are encouraging data scientists, academics and industry experts to share their success stories, challenges and visions for future applications of data science that can help solve problems for the social good.

New York, NY on Sunday, September 25, at Bloomberg Global Headquarters (731 Lexington Ave)

Deadline for paper abstract submissions is Friday, July 1.

OPODIS 2016 - The 20th International Conference on Principles of Distributed Systems

OPODIS is an open forum for the exchange of state-of-the-art knowledge on distributed computing and distributed computer systems. All aspects of distributed systems are within the scope of OPODIS.

Madrid, Spain Tuesday-Friday, December 13-16.

Deadline for submissions is Monday, August 22.

CDS News

When to Trust Robots with Decisions, and When Not To

Harvard Business Review, Vasant Dhar from May 17, 2016
Smarter and more adaptive machines are rapidly becoming as much a part of our lives as the internet, and more of our decisions are being handed over to intelligent algorithms that learn from ever-increasing volumes and varieties of data.

As these “robots” become a bigger part of our lives, we don’t have any framework for evaluating which decisions we should be comfortable delegating to algorithms and which ones humans should retain. That’s surprising, given the high stakes involved.

I propose a risk-oriented framework for deciding when and how to allocate decision problems between humans and machine-based decision makers.

Also, by Professor Dhar:

- How can we control intelligent systems no one fully understands? (May 16, TechCrunch)

Tools & Resources

The end of CartoCSS

Mapbox, Tom MacWright from May 13, 2016

... Vector maps have been one of the largest leaps in Mapbox's history. We built a map renderer from the ground up. We created the industry standard for distributing vector map tiles, the Mapbox Vector Tile spec, which has also been adopted by other companies, including **Esri**. We keep improving every detail of the system, from size efficiency to text rendering and animations, and building an ecosystem around Mapbox GL JS, the Mapbox iOS SDK, and the Mapbox Android SDK.

As part of this transition, we created a new language: the Mapbox GL Style Specification. It is the realization of the lessons we learned creating CartoCSS, and it is the next step in cartographic styling. Like the Mapbox Vector Tile specification, it's an open standard and we're seeing other organizations starting to adopt it.

Deep Learning Software

NVIDIA Developer from May 01, 2016
Deep learning algorithms use large amounts of data and the computational power of the GPU to learn information directly from data such as images, signals, and text. **NVIDIA** DIGITS offers an interactive workflow-based solution for image classification. Deep learning frameworks offer more flexibility with designing and training custom deep neural networks and provide interfaces to common programming language. The NVIDIA Deep Learning SDK offers powerful tools and libraries for the development of deep learning frameworks such as Caffe, CNTK, TensorFlow, Theano, and Torch.

Apache Spark 2.0: Introduction to Structured Streaming

O'Reilly Media, Strata + Hadoop World from May 16, 2016
Michael Armbrust and **Tathagata Das** explain updates to Spark version 2.0, demonstrating how stream processing is now more accessible with Spark SQL and DataFrame APIs. [video, 52:38]

RoogleVision released - a Package for Image Recognition

Florian Teschner from May 16, 2016
First to the naming; it basically is an arbitrary condensation of "R + **Google** Cloud Vision API". I wonder why google chooses to mix google with vision. In my opinion it sounds pretty much like "to goggle with vision", which makes limited sense. For the functionality; the package enables convenient Image Recognition, Object Detection, and OCR using the Google's Cloud Vision API.

Careers

Postdoctoral Scholar - Research Associate

University of Southern California, Information Sciences Institute from May 13, 2016
Marina Del Rey, CA Information Sciences Institute (ISI), part of the Viterbi School of Engineering at USC, seeks applicants for two Postdoctoral Researcher positions with a focus on modeling and forecasting with multivariate and heterogeneous time series data. The positions are for one year, with possibility of a renewal for two more years.

The successful candidate will work on a federally sponsored project together with a team of experts in machine learning, computer science, and social sciences. The

central objective is to develop an algorithmic framework for analyzing multivariate time-series data, and generating forecasts about various events of interests from two broad categories: Socio-political and societal events (elections, epidemics, etc.); and cybersecurity events (large scale security breaches, new viral malware, etc.).

Networking Tips for Younger PhD Students

Jean Yang, updated sporadically at best blog from May 13, 2016
This post was a collaboration with **Nadia Polikarpova** and **Shachar Itzhaky**, done while we were supposed to be collaborating on other things.

A younger student in the group where I did my PhD is going to his first conference next week and my advisor sent him my way for advice. Nadia, Shachar, and I had already been discussing research (and attending a BBQ) for hours at this point, so we welcomed the opportunity to discuss something else. Here's what we came up with.

What Data Scientist Shortage? Get Serious and Get Talent

DataInformed, Thomas Davenport from May 17, 2016
I have concluded that if you are serious about analytics people, there are ways to ensure that they are available within your organization when they are needed.

The two main approaches to finding and hiring data science talent, of course, are hiring people in the external labor market, and training or retraining people yourself. The former approach is much more common, though I'm not sure it should be.