**NYU Data Science Community** features journalism, research papers, events, tools/software, and jobs for February 26, 2016.

Please let us *(Laura Noren, Brad Stenger)* know if you have something to add to the newsletter. We are grateful for the generous financial support from the Moore-Sloan Data Science Environment and to NYU's Center for Data Science.  

**NYU Data Science Community Newsletter Issue 040.**

---

**Data Science News**  
**How to Think About Bots**  
VICE, Motherboard from February 23, 2016  
*A botifesto*

We live in a world of bots. Generally speaking, these sets of algorithms are responsible for so much on the backend of the internet, from making Google searches possible to filling up your spam folder. But an emergent kind of bot, capable of interacting with humans and acting on their behalf, is playing a more active role in our everyday lives.

*Also, more Bots essays*  
*Data & Society:Points blog*:  
Bots: A definition and some historical threads *(Allison Parrish, February 24)*  
Our friends, the bots? *(Alexis Lloyd, February 25)*  
What is the Value of a Bot? *(danah boyd, February 26)*

---

**Beware: Employers Mine Your Personal Data to Predict Health**  
*Data Science Association, Michael Walker* from February 18, 2016  
... There is a good chance your employer and health insurer (both partnered with big government) is currently or soon about to mine data about where and how you shop, where you eat out, whether and when you vote, and the prescription drugs you use - to predict your state of health, forecast future health problems and recommend preventive treatments.

---

**Berkeley Lab, UC Berkeley Scientists to Participate in New NASA Space Telescope Project**  
*Berkeley Institute for Data Science* from February 19, 2016  
**BIDS** senior fellow and faculty director **Saul Perlmutter** will lead a team of 29 scientists from 15 institutions for the Department of Energy's space telescope project. The team will "explore mysteries of dark energy, hunt for distant planets, [and] retrace universe's history."

---

**Analysing the Social Networks of 19th-20th Century Literature**  
**Derek Greene** from February 21, 2016  
The “Nation, Genre and Gender” research project at [University College Dublin] is currently creating a large digital corpus of Irish and English novels from the period 1800-1922. Our
Now, the company has made it available on GitHub for everyone to use.

able to perform 'deep learning' on the vast ocean of data kept in
Yahoo! Inc., is getting into the artificial intelligence (AI) game with the release of new
Yahoo just made deep learning easier with CaffeOnSpark
source .tex for the tutorial file EPV_demo.pdf can be built from EPV_demo.Rnw using
Predicting NBA Possession Outcomes.”

GitHub
struggled to work well together due to both high-level architectural and low-level technological
While the
tool of choice for a wide variety of data engineering and data science needs. Hugely
rapidly creating visualizations for analysis and presentation. With
Today we are excited to announce the official 1.0 release of
huge numbers of friends millions or tens of millions of followers in some cases.

The friendship paradox is straightforward to explain. It comes about because of the skewed
and scalable performance demanded by the world's largest enterprises.

8.
environments.
open data as a tool for civic and social impact!

Monday, February 29, Warren Weaver Hall 1302, 11:30 a.m.
In order to gain insight into the reasons for such success, in this talk we will start by studying
information from images, speech and text, with excellent statistical generalization. These are
Berkeley
NYU Computer Science Department Colloquium: Convolutional Networks
GE's Big Bet on Data and Analytics
has explored the influence of college on preparing students with the capacity, desire and
... where do innovators come from? And how do they acquire their skills?
Kensho
Nadler
Ars Technica,
We're starting in the UK, where the National Health Service is hugely important to our team.

Researchers create super-efficient Wi-Fi
Ars Technica, University of Washington from February 23, 2016
A team of computer scientists and electrical engineers from the University of Washington has
developed an extremely power-efficient version of Wi-Fi wireless networking technology that
consumes 10,000 times less power than the current Wi-Fi components, allowing Wi-Fi
networking to be built into a much wider range of devices. The team will present a paper
(PDF) with the results of their research into what they have dubbed Passive Wi-Fi at the
upcoming USENIX Symposium on Networked Systems Design and Implementation in March.

I'm Excited! A Post Pre-Print-Posting-Powwow Post
Michael Eisen from February 18, 2016
I just got back from attending a meeting organized by a new group called ASAPbio whose
mission is to promote the use of pre-prints in biology. ... By the end of the meeting's 24 hours
it seemed like nearly everyone in attendance was sold on the idea that biomedical researchers
should all post pre-prints of their work, and had already turned their attention to questions
about how to do it. And there was a surprisingly little resistance to the idea that post-
publication review of papers initially posted as pre-prints could, at least in principle, fulfill the
functions that pre-publication review currently carries out. That's not to say there weren't
concerns and even some objections – there were, as I will discuss below. But these were all
dealt with to varying degrees, and there seemed to be a general attitude these concerns can

10 Breakthrough Technologies 2016
MIT Technology Review from February 23, 2016
Which of today's emerging technologies have a chance at solving a big problem and opening
up new opportunities? Here are our picks. The 10 on this list all had an impressive milestone
in the past year or are on the verge of one. These are technologies you need to know about
right now.

DeepMind Health
Google DeepMind from February 24, 2016
We founded DeepMind to solve intelligence and use it to make the world a better place by
developing technologies that help address some of society's toughest challenges. It was clear
to us that we should focus on healthcare because it's an area where we believe we can make
a real difference to people's lives across the world.

We're starting in the UK, where the National Health Service is hugely important to our team.
be addressed, and did not constitute reasons not to proceed.

**The Robots Are Coming for Wall Street**  
When Daniel Nadler woke on Nov. 6, he had just enough time to pour himself a glass of orange juice and open his laptop before the Bureau of Labor Statistics released its monthly employment report at 8:30 a.m. He sat at the kitchen table in his one-bedroom apartment in Chelsea, nervously refreshing his web browser — Command-R, Command-R, Command-R — as the software of his company, Kensho, scraped the data from the bureau’s website. Within two minutes, an automated Kensho analysis popped up on his screen: a brief overview, followed by 13 exhibits predicting the performance of investments based on their past response to similar employment reports.

Nadler couldn’t have double-checked all this analysis if he wanted to. It was based on thousands of numbers drawn from dozens of databases. He just wanted to make sure that Kensho had pulled the right number — the overall growth in American payrolls — from the employment report. It was the least he could do, given that within minutes, at 8:35 a.m., Kensho’s analysis would be made available to employees at Goldman Sachs.

**Straight A students may not be the best innovators**  
*The Conversation*, Matthew Mayhew from February 19, 2016  
... where do innovators come from? And how do they acquire their skills?

One place – perhaps among the best – is college. Over the past seven years, my research has explored the influence of college on preparing students with the capacity, desire and intention to innovate.

In this time we’ve learned that many academic and social experiences matter quite a bit; grades, however, do not matter as much.

**GE’s Big Bet on Data and Analytics**  
*MIT Sloan Management Review* from February 18, 2016  
GE has bet big on the Industrial Internet — the convergence of industrial machines, data, and the Internet. The company is putting sensors on gas turbines, jet engines, and other machines; connecting them to the cloud; and analyzing the resulting flow of data. The goal: identify ways to improve machine productivity and reliability. This MIT Sloan Management Review case study looks at how this traditional manufacturer is remaking itself into a modern digital business.

**Events**

**NYU Computer Science Department Colloquium: Convolutional Networks against the Curse of Dimensionality** — Joan Bruna, University of California at Berkeley  
Convolutional Neural Networks (CNN) are a powerful class of non-linear representations...
Convolutional Neural Networks (CNNs) are a special class of non-linear representations that have shown through numerous supervised learning tasks their ability to extract rich information from images, speech and text, with excellent statistical generalization. These are examples of truly high-dimensional signals, in which classical statistical models suffer from the so-called curse of dimensionality, referring to their inability to generalize well unless provided with exponentially large amounts of training data.

In order to gain insight into the reasons for such success, in this talk we will start by studying statistical models defined from wavelet scattering networks, a class of CNNs where the convolutional filter banks are given by complex, multi-resolution wavelet families.

Monday, February 29, Warren Weaver Hall 1302, 11:30 a.m.

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 3, will be Slav Petrov (Google NY) – Towards Universal Syntactic Processing of Natural Language.

Storytelling with Open Data: A Tool for Civic Impact
Join WagnerTech and Bridge: Students for Social Innovation for a panel discussion on using open data as a tool for civic and social impact!

Friday, March 4, at The Puck Building, The Rudin Family Forum for Civic Dialogue, 2nd Fl. (295 Lafayette Street), starting at 5 p.m.

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.

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.

**Edge Tools in a Digital World**

**The Social Science Research Council**, in partnership with the **New York Public Library**, is launching a three-part series of events titled, “Edge Tools in a Digital World” featuring **John Seely Brown**, **Yochai Benkler**, and **David Krakauer**, among other amazing contributors in the digital and media worlds. With a focus on complexity and design, these conversations seek to underscore creative approaches that might advance understandings and guide action in this increasingly stressed planet.

Tuesday, March 8, in the NYPL’s Trustee’s Room, Bryant Park, starting at 5 p.m. To RSVP, please contact **Kate Grantz** at grantz@ssrc.org.

**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.

**Deadlines**

**Announcing the Google Internet of Things (IoT) Technology Research Award Pilot**

While there has been significant progress in [IoT], there remain significant challenges in terms of (1) interoperability and a standardized modular systems architecture, (2) privacy, security and user safety, as well as (3) how users interact with, manage and control an ensemble of devices in this connected environment.

It is in this context that we are happy to invite university researchers to participate in the **Internet of Things (IoT) Technology Research Award Pilot**. This pilot provides selected researchers in-kind gifts of Google IoT related technologies (listed below), with the goal of fostering collaboration with the academic community on small-scale (~4-8 week) experiments, discovering what they can do with our software and devices.

Deadline for proposals is Monday, February 29, 2016.

**Call for WiML 2016 Organizers**

We are seeking organizers for the 11th Workshop for Women in Machine Learning (WiML). The workshop will once again be co-located with the Neural Information Processing Systems (NIPS) conference and will be held in December 2016 in Barcelona, Spain.

Deadline to respond is Monday, February 29, 2016.
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/ΖJPqGG0UW7.

Deadline for applications is Monday, March 14, 2016.

ACL 2016

For the first time, the annual meeting of the Association for Computational Linguistics (ACL) takes place in Germany. ACL 2016 will be held at the Humboldt University, in the heart of Berlin, August 7-12, 2016.

As in previous years, the program of the conference includes poster sessions, tutorials, workshops, and demonstrations in addition to the main conference. ACL is the premier conference of the field of computational linguistics, covering a broad spectrum of diverse research areas that are concerned with computational approaches to natural language.

Deadline for long-paper submissions is Friday, March 18, 2016.

CDS News

White House Honors Psychology's Gureckis with Presidential Early Career Award for Scientists and Engineers

NYU News from February 22, 2016

New York University’s Todd Gureckis, an associate professor in the Department of Psychology, has been awarded a Presidential Early Career Award for Scientists and Engineers (PECASE). The awards, announced by the White House, are the highest honor bestowed by the U.S. government on science and engineering professionals in the early stages of their independent research careers.

First Evidence for the Happiness Paradox—that Your Friends Are Happier than You Are


Johan Bollen at Indiana University in Bloomington and a few pals [including NYU CDS’ Bruno Goncalves] have found the first evidence of a happiness paradox on Twitter. They say that it is good evidence that social network use can affect the well-being of a significant proportion of the planet’s population.
The friendship paradox is straightforward to explain. It comes about because of the skewed way people collect friends on online social networks such as Twitter and Facebook. Most people have a small number of friends—a few dozen or so. But a tiny fraction of people have huge numbers of friends millions or tens of millions of followers in some cases.

**Tools & Resources**

**Introducing Vega-Lite**

*Medium, UW Interactive Data Lab* from February 23, 2016

Today we are excited to announce the official 1.0 release of **Vega-Lite**, a high-level format for rapidly creating visualizations for analysis and presentation. With **Vega-Lite**, one can concisely describe a visualization as a set of encodings that map from data fields to the properties of graphical marks, using a JSON format. Vega-Lite also supports data transformations such as aggregation, binning, filtering, and sorting, along with visual transformations including stacked layouts and faceting into small multiples.

**Python and Apache Hadoop: A State of the Union**

*Cloudera VISION blog* from February 17, 2016

Over the last five years, the rapid growth of Python’s open source data tools have made it a tool of choice for a wide variety of data engineering and data science needs. Huge success stories that we now take for granted, such as Jupyter, Pandas, and scikit-learn, were comparatively nascent efforts only a few years ago. Today, data teams worldwide love Python for its accessibility, developer productivity, robust community, and “batteries-included” open source libraries.

During the same time period, the Apache Hadoop ecosystem has risen to the challenge of collecting, storing, and analyzing accelerating volumes of data with the robustness, security, and scalable performance demanded by the world’s largest enterprises.

While the Python and Hadoop ecosystems have each flourished in their own right, they have struggled to work well together due to both high-level architectural and low-level technological challenges.

**Demo of NBA Expected Possession Value model**

*GitHub/dcervone* from December 26, 2015

This repository contains data and code offering a demo of the NBA Expected Possession Value model presented in the paper "A Multiresolution Stochastic Process Model for Predicting NBA Possession Outcomes."

The main document that introduces and illustrates the code/data is EPV_demo.pdf. The source .tex for the tutorial file EPV_demo.pdf can be built from EPV_demo.Rnw using RStudio.

**Yahoo just made deep learning easier with CaffeOnSpark**

*SiliconANGLE* from February 26, 2016

Yahoo has now made its popular data science tool Caffe available for use with Apache Spark. …
Yahoo! Inc., is getting into the artificial intelligence (AI) game with the release of new internally-built software under an open-source license. Called CaffeOnSpark, the software is able to perform ‘deep learning’ on the vast ocean of data kept in Yahoo’s Hadoop file system. Now, the company has made it available on GitHub for everyone to use.

Click here to receive the NYU Data Science Community Newsletter OR to have us follow your twitter feed so that our data science twitter bot can easily grab links from your tweets.

To send us an announcement for the newsletter, please email laura.noren@nyu.edu and brad.stenger@nyu.edu by 9 pm Eastern Time on Thursday evenings for inclusion in Friday’s newsletter. We retain curatorial discretion.

NYU Data Science Community Newsletter Issue 040.