NYU Data Science Community Newsletter features journalism, research papers, events, tools/software, and jobs for 14 August 2015.

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 NYU’s Center for Data Science.

NYU Data Science Community Newsletter Issue 014.

Data Science News

Data Carpentry Grant from the Moore Foundation
Software Carpentry, Moore Foundation from August 11, 2015
We are extremely pleased to announce that Data Carpentry has received $750,000 in funding from the Gordon and Betty Moore Foundation.

Statistical technique helps cancer researchers understand tumor makeup, personalize care
A new statistical method for analyzing next-generation sequencing (NGS) data that helps researchers study the genome of various organisms such as human tumors and could help bring about personalized cancer treatments was presented today at a session of the 2015 Joint Statistical Meetings (JSM 2015) in Seattle.

Yuan Ji, director of the Program for Computational Genomics and Medicine Research Institute at NorthShore University HealthSystem and associate professor of biostatistics at The University of Chicago, described the new technique--called Bayesian feature allocation models--during a presentation titled "Bayesian Models for Heterogeneity in Human Cancers."

40 Data Mavericks Under 40
import.io blog from August 06, 2015
We get to meet lots of exciting data innovators at Extract, so we decided to use our knowledge to create the first Data Mavericks list. A lot of thought went in to choosing people who are truly making a difference in the data scene.

It was no easy task narrowing our list down, but 40 under 40 seemed like a good place to start. Here they are. In alphabetical order, please enjoy our list of the Top 40 Data Mavericks under 40.

Multilevel Models and Political Advertising
**Bad Hessian, Adam Slez** from August 11, 2015

The graph above recently appeared as part of Scott Walker’s Twitter feed. Presumably, the idea is to suggest that under Walker’s leadership, Wisconsin has done better than the country as a whole when it comes to unemployment, though an alternative version of the ad makes it somewhat more personal, using the same basic figures to suggest that Walker—a Republican presidential candidate—is outperforming sitting Democratic president Barack Obama. In these ads, the Walker campaign repeatedly highlights the fact that the unemployment rate in Wisconsin is lower than the national average. Note, however, that the unemployment rate in Wisconsin was already lower than the national average when Walker took office. In other words, Walker inherited a good labor market. If we want to measure Walker’s effect on the Wisconsin economy, we need to look at changes in the unemployment rate over time.

I decided to throw some data at the problem. To put Wisconsin’s performance in perspective, I estimated the effect of time on the monthly seasonally-adjusted unemployment rate in each state. Covering the period from January 1, 2011 to June 1, 2015, the data came from the Bureau of Labor Statistics—the same source used by Walker and company.

**Q&A with Jonathan Pillow on dissecting the brain using math and neuroscience**

**Princeton University**, *News at Princeton* from August 10, 2015

The brain is the ultimate big-data problem. Its billions of neurons give rise to numerous abilities, such as making decisions, interpreting color and even recognizing your best friend. **Jonathan Pillow**, a Princeton University assistant professor of psychology and the Princeton Neuroscience Institute, aims to understand the brain by using math and statistics to make sense of the reams of information collected by brain-imaging studies. Recently arrived from the University of Texas-Austin, Pillow, who also is affiliated with the University’s Center for Statistics and Machine Learning, sat down to talk about how he got into neuroscience, his approach to teaching, and his latest research published earlier this month in the journal Science.

**Mike Jordan and BDAS in Science**
UC Berkeley, AMPLab from July 31, 2015
The BDAS Stack made its debut in the prestigious journal Science earlier this month. A Review article by our own Mike Jordan and CMU’s Tom Mitchell entitled “Machine Learning: Trends, Perspectives and Prospects” appeared in the July 17, 2015 issue (subscription required), and includes the following diagram (which should be familiar to the AMPLab:faithful):

How’s the weather? Using artificial intelligence for better answers
Microsoft Research, Next at Microsoft blog from August 10, 2015
Here are two things we know about the weather: We have tons of data about it, and we care very deeply about it.

It’s those two factors that led Microsoft researchers Ashish Kapoor and Eric Horvitz to turn to the latest advances in artificial intelligence to try to find a better answer to that age-old question: What’s the weather forecast?

Turning your R (or Python) models into APIs
**Jo-fai Chow**, *Blend it like a Bayesian!* blog from August 09, 2015
More and more real-world systems are relying on data science and analytical models to deliver sophisticated functionality or improved user experiences. For example, Microsoft combined the power of advanced predictive models and web services to develop the real-time voice translation feature in Skype. Facebook and Google continuously improve their deep learning models for better face recognition features in their photo service.

Some have characterised this trend as a shift from Software-as-a-Service (SaaS) to an era of Models-as-a-Service (MaaS). These models are often written in statistical programming languages (e.g., R, Python), which are especially well suited to analytical tasks.

With analytical models playing an increasingly important role in real-world systems, and with more models being developed in R and Python, we need powerful ways of turning these models into APIs for others to consume.

**Building the Next New York Times Recommendation Engine**
The New York Times, Open blog from August 11, 2015
The New York Times publishes over 300 articles, blog posts and interactive stories a day.

Refining the path our readers take through this content — personalizing the placement of articles on our apps and website — can help readers find information relevant to them, such as the right news at the right times, personalized supplements to major events and stories in their preferred multimedia format.

In this post, I'll discuss our recent work revamping The New York Times's article recommendation algorithm, which currently serves behind the Recommended for You section of NYTimes.com.

**Data Science for Social Good: Improving Service to Homeless Familes**
UW eScience Institute from August 05, 2015
We are living in an age where data plays a role in almost everything we do. While the power of big data is already being harnessed in science and technology, the question remains how to bring the same disruptive impact to public policy and social good. Companies like Microsoft, Google and Facebook are using massive amounts of user data to create products that engage and entertain users (and to find the most effective advertisement to display to them). In a variety of scientific fields, new measurement devices are producing larger and larger quantities of data about everything from remote galaxies to our own DNA, accelerating our progress towards a better understanding of the universe. Some have even gone so far as to say that data is "unreasonably effective." But how does one use data to promote social good? How does one harness the lessons learned in analyzing data from the internet, or data from scientific measurements, to address a social problem as challenging and complex as family homelessness?

This summer, the University of Washington's eScience Institute is hosting the first installment of a Data Science For Social Good program to address this question.

**Mapping NYC Taxi Data**
**Daniel Forsyth** from August 07, 2015

Earlier this week the **New York City Taxi & Limousine Commission** officially released yellow and green taxi trip record data for all of 2014 and up to June of 2015. This includes millions of records that include pick-up and drop-off dates and times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. The data, which was previously only available through submission of a formal Freedom of Information Law (FOIL) request, is available in CSV format as well as from Google's BigQuery tools [1].

After seeing the aforementioned comment and its accompanying visualization I wanted to have a go at replicating it in python. The first step was getting the data into pandas. In this situation it was much easier to query the data from GBQ than to download the individual CSV files. (Note that you must have the google-api-python-client installed and be logged into GBQ and create a new project for this to work properly). Pandas has an included pandas.io.gbq module that allows you to parse the results of a GBQ query into a dataframe very easily. The following query creates a dataframe that includes the latitude and longitude of all pickup locations in 2015, this ends up being around 750,000 records.

**Events**

**This American Life's Audio Hackathon**

We are inviting developers, coders, designers, producers, sound designers...anyone who has skills and ideas to offer to join us. We will group you into teams, and after introductory talks by audio professionals, you will participate in a two day hackathon. Afterward, teams will present their creations to our speakers and to employees from our partners, who will evaluate and critique their work.

Saturday-Sunday, September 19-20, at ThoughtWorks, 99 Madison Ave

Registration Deadline: Saturday, August 15

**NYC Media Lab 2015 Annual Summit**

NYC Media Lab’s Annual Summit is a snapshot of the best thinking, projects, and talent in digital media from universities in NYC and beyond.

This is an opportunity for media executives, technologists, and decision makers to see more than 100 university prototypes and demonstrations that explore interesting technologies and applications related to the future of media. Attendees will be also invited to roll up their sleeves during a series of interactive workshops led by NYC faculty.

Friday, September 25, at 9 a.m., NYU Skirball Center for the Performing Arts

**New England Symposium on Statistics in Sports**
The 2015 New England Symposium on Statistics in Sports will be a meeting of statisticians and quantitative analysts connected with sports teams, sports media, and universities to discuss common problems of interest in statistical modeling and analysis of sports data. The symposium format will be a mixture of invited talks, a poster session, and a panel discussion.

Saturday, September 26, at Harvard University Science Center

**Deadlines**

**The Workshop on Information in Networks, Call for Submissions**

We are excited to announce the 6th annual WIN Workshop, 2015, taking place October 2-3 in New York City. The last five years have seen the development of an intimate and influential community, around topics which at their core involve ‘information in networks’—its distribution, its diffusion, its value, and its influence on social and economic outcomes. Scholars from fields as diverse as computer science, economics, physics, political science and sociology have come together to lay the foundation for ongoing relationships and to build a multidisciplinary research community. The sixth year of WIN will bring innovative new content and an even more vibrant discussion to the forum. More information: http://www.winworkshop.net/ and http://winworkshop.net/win2015/WINCALL2015.pdf

Deadline for Abstract Submissions: Friday, August 21

**UNICEF and Bloomberg Announce Data Science Researcher-in-Residence Program**

The researcher-in-residence, supported by Bloomberg, will help leverage knowledge transfer and capacity building on Data Science for humanitarian development at a global level while helping to design new data science methods or applying existing methods to help solve pressing problems that face large humanitarian organizations such as UNICEF.

Candidates interested in applying for this role can download and submit an application from http://www.bloomberg.com/company/d4gx/#unicef.

Deadline for applications is 21 August 2015.

**ACM SIGSPATIAL 2015, November 3-6 in Seattle**

The ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems 2015 (ACM SIGSPATIAL 2015) is the twenty-third event in a series of symposia and workshops that began in 1993 with the aim of bringing together researchers, developers, users, and practitioners in relation to novel systems based on geo-spatial data and knowledge, and fostering interdisciplinary discussions and research in all aspects of geographic information systems.

Workshops include UrbanGIS 2015 and Geo Privacy, and have paper submission deadlines that begin in mid-August.

**Cognitive Computing Challenge**
... We want a solution that can not only learn how and what data to extract from data sources like spreadsheets, word processor files, and computer generated PDF files but also learn how to map the data found in these documents to specified target fields in a database. ... This is a qualifying problem which requires you to process MLS (Multiple Listing Service) data from 300 MLS training records into a database. You will be provided with a document summarizing the correct format of each field to be extracted.

Submission Deadline is Monday, January 11, 2016. Submissions began on June 8.

**CDS News**

**Introduction to Machine Learning with Python: Sarah Guido, Andreas Mueller:**

*9781449369415: Amazon.com: Books*

Amazon.com Books, O'Reilly Publishing from August 10, 2015

Many Python developers are curious about what machine learning is and how it can be concretely applied to solve issues faced in businesses handling medium to large amount of data. Machine Learning with Python teaches you the basics of machine learning and provides a thorough hands-on understanding of the subject.

You’ll learn important machine learning concepts and algorithms, when to use them, and how to use them. The book will cover a machine learning workflow: data preprocessing and working with data, training algorithms, evaluating results, and implementing those algorithms into a production-level system.

Available now for pre-sale. Publication: January 2016.

*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 014.**