NYU Data Science Community Newsletter features journalism, research papers, events, tools/software, and jobs for July 15, 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

SciPy 2016: Scientific Computing with Python Conference - YouTube

YouTube, Enthought from July 15, 2016
SciPy 2016, the fifteenth annual Scientific Computing with Python conference, was held July 11-17, 2016 in Austin, Texas. This video playlist includes talks by **Hanna Wallach**, **Brian Granger**, **Stuart Geiger**, **Jess Hamrick**, **Sebastian Benthall**, and more. [58 videos playlist]

Also from SciPy:
- scipy_2016_notes (July 15, github.com/chendaniely)

Microsoft launches data science degree to plug the skills gap, more courses could follow

VentureBeat, Paul Sawers from July 14, 2016
Microsoft has unveiled a new degree program as it looks to address the "significant skills gap" that exists in the field of data science.

Announced at its Worldwide Partner Conference yesterday, the Microsoft Professional Degree (MPD) program is touted as "the first program of its kind to offer employer-endorsed, university-caliber curriculum for professionals at any stage of their career." The course will be offered through **Edx.org**.

On #AINow: Beyond Transparency, what is design and ethics in algorithms and artificial intelligence?

Medium, Caroline Sinders from July 14, 2016
Are we giving AI engineers and creators the right tools to be ethical and create ethical products, algorithms, and software?

More from recent conferences:
- useR 2016, A Love Story (July 12, Medium, Moore Data, Chris Mentzel)
- Dispatch: The White House's and NYU's Artificial Intelligence Workshop #AINow (July 12, LinkedIn, Khurram Nasir Gore)
- Fast Forward Labs: What We Liked at AINow (July 08, Fast Forward Labs Blog)
- Machined Learnings: ICML 2016 Thoughts (July 04, Paul Mineiro, Machined Learnings blog)
- Microsoft Research at IJCAI 2016: Developing technologies that allow people and machines to collaborate (July 08, Microsoft Research, Eric Horvitz)

GovLab's 8 takeaways for handling data responsibly

Technical.ly Brooklyn from July 14, 2016
The devil's in the data. A new report from **NYU Tandon's Governance Lab** attempts to define guidelines for its deployment.

How technology disrupted the truth

The Guardian, Media from July 12, 2016
Social media has swallowed the news – threatening the funding of public-interest reporting and ushering in an era when everyone has their own facts. But the consequences go far beyond journalism.

Also in data journalism:
- 77 | Polygraph and The Journalist Engineer Matt Daniels (July 01, Data Stories; Enrico Bertini, Moritz Stefaner and guest, Matt Daniels)
- R in the data journalism workflow at FiveThirtyEight (July 12, Nathan Yau, Flowing Data blog, and Andrew Flowers, Five Thirty Eight)
- Robot Sports Journalism: Is This The End Or A Fresh Start? (July 14, Vocativ, Joe Lemire)
- [1607.03057] Learning from the News: Predicting Entity Popularity on Twitter (July 11, arXiv, Computer Science > Social and Information Networks; Pedro Saleiro, Carlos Soares)
- What will the Internet of Things do to journalism? (July 14, Columbia University, Tow Center for Digital Journalism, Francesco Marconi)

Developing SocArXiv — a new open archive of the social sciences to challenge the outdated journal system.

London School of Economics, Impact of Social Sciences blog from July 11, 2016
Philip Cohen argues a cultural shift is taking place in the social sciences. He introduces SocArxiv, a fast, free, open paper server to encourage wider open scholarship in the social sciences.

Another new Open publishing platform:
- The ReScience Journal (July 14, Tiziano Zito)

Biodiversity falls below ‘safe levels’ globally

University College London, UCL News from July 14, 2016
Levels of global biodiversity loss may negatively impact on ecosystem function and the sustainability of human societies, according to UCL-led research. biodiversity hotspots

“This is the first time we’ve quantified the effect of habitat loss on biodiversity globally in such detail and we’ve found that across most of the world biodiversity loss is no longer within the safe limit suggested by ecologists” explained lead researcher, **Dr Tim Newbold**.

Rice wins interdisciplinary 'big data' grant

Rice University News & Media from July 12, 2016
The three-year program will serve as a point of contact for six graduate students, two

postdoctoral researchers and several undergraduates as they pursue statistics and computer science projects in the **Rice** research groups to which they're assigned.

More university data science:

- U-M launches two specializations for new generation of data scientists (July 11, University of Michigan, The University Record)
- NYU Steinhardt and StartEd Launch NY Edtech Accelerator and Incubator (July 06, NYU Steinhardt)

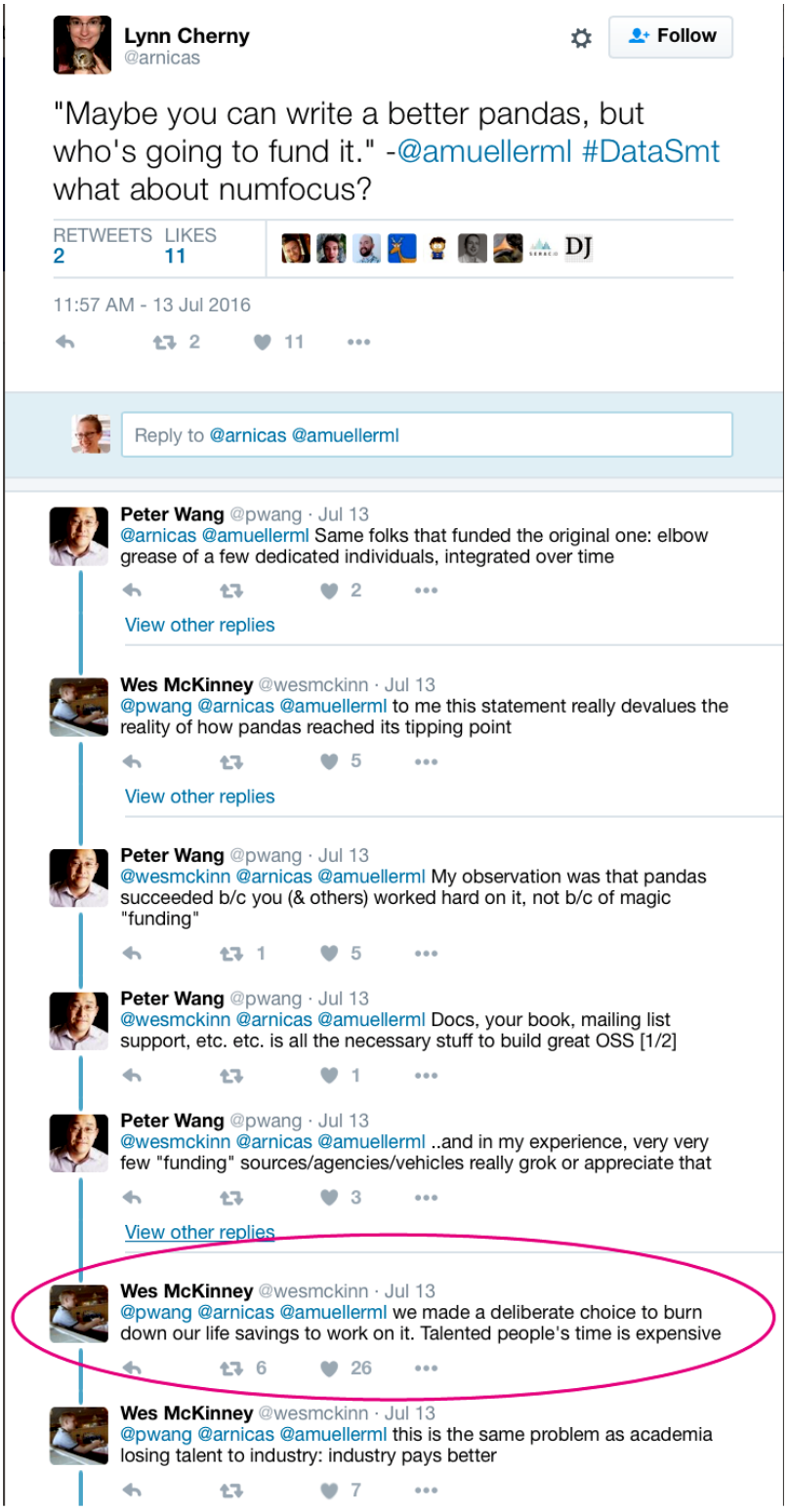
[1607.03320] What Happens After You Both Swipe Right: A Statistical Description of Mobile Dating Communications

arXiv, Computer Science > Social and Information; Jennie Zhang, Taha Yasseri from July 12, 2016

This paper looks at one of these sets of data: metadata of approximately two million conversations, containing 19 million messages, exchanged between 400,000 heterosexual users on an MDA. Through computational analysis methods, this study offers the very first large scale quantitative depiction of mobile dating as a whole. We report on differences in how heterosexual male and female users communicate with each other on MDAs, differences in behaviors of dyads of varying degrees of social separation, and factors leading to "success"-operationalized by the exchange of phone numbers between a match.

Tweet of the Week

Twitter; Lynn Cherney, Peter Wang and Wes McKinney from July 13, 2016



Events

Artificial Intelligence And The Law

Join the **Wikimedia Foundation** for a discussion on the intersection of the law and emerging technologies, such as driverless cars, web crawlers, and lethal autonomous weapons. Panelists will explore the legal challenges presented by these technologies in the areas of international law, employment, intellectual property, and tort liability.

San Francisco, CA Tuesday, July 19 at Wikimedia Foundation (149 New Montgomery Street)

Seattle: Women in Data Science: Analyzing the Stories - July 2016

Don't miss this special event as we look into the future of data science, understand the questions to ask, and learn the mistakes startups make when managing their data. The panelists include **Elaine Werffeli** of Ecuiti, **Claire Jaja** of Atlas Informatics, and **Alice Zheng** formally of Dato & Microsoft.

Seattle, WA Wednesday, July 20, at Code Fellows (2901 3rd Ave Suite 300), starting at 6:30 p.m.

Science of Music Hackathon

Presented in collaboration with HAMR: Hacking Audio and Music and the 17th International Society for Music Information Retrieval Conference: Science of Music Hackathon!

New York, NY Friday-Saturday, August 5-6 at 45 W 18th St, 3rd Floor

12th International Workshop on Mining and Learning with Graphs (MLG 2016)

San Francisco, CA Held in conjunction with KDD'16 on Sunday, August 14.

O'Reilly Artificial Intelligence Conference

Discover the real-world opportunities of applied artificial intelligence

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

Deadlines

Call For Papers – SocInfo'16

deadline: Conference

Bellevue, WA The International Conference on Social Informatics (SocInfo16) is an interdisciplinary venue that brings together researchers from the computational and social sciences to help fill the gap between the two communities.

The deadline for full paper submissions is Wednesday, July 20.

NYU Stern – Master of Science in Business Analytics

deadline: Education Opportunity

The MS in Business Analytics Program is a one-year, part-time program divided into

five on-site class sessions (modules). Two of the five modules occur outside of **NYU Stern** in global rotating locations, allowing you to expand your international network of valuable peers and contacts.

Deadline to apply for the program beginning in May, 2017 is August 1.

Request For Proposals | California Initiative to Advance Precision Medicine

deadline: RFP

We are pleased to announce the release of this Request for Proposals (RFP). This RFP will help serve as a means to identify approximately six proof-of-principle Demonstration Projects to advance precision medicine in California.

Deadline for concept proposals is Monday, August 8.

EMNLP NLP+CSS Doctoral Consortium CFP

deadline: Conference

This doctoral consortium aims to bring together students and faculty mentors across NLP and the social sciences, to encourage interdisciplinary collaboration and cross-pollination.

Austin, TX The November 6 consortium event is part of a workshop at EMNLP, one of the top conferences in natural language processing

Deadline for submissions is Friday, August 12.

Call for Papers - The Conference on Digital Experimentation @ MIT

deadline: Conference

Cambridge, MA The purpose of the Conference on Digital Experimentation at **MIT** (CODE Conference) is to bring together leading researchers conducting and analyzing large scale randomized experiments in digitally mediated social and economic environments, in various scientific disciplines including economics, computer science and sociology, in order to lay the foundation for ongoing relationships and to build a lasting multidisciplinary research community.

The deadline for paper submissions is Friday, August 12.

TalkingData Launches Data Science Competition Featuring $25,000 Prize Pool

deadline: Contest/Award

TalkingData, the leading Big Data and analytics company in China, is launching its “Global Data Science Competition” beginning today, July 11, 2016 and ending September 5, 2016. The event will being held on **Kaggle** and is working in partnership with **Turi**, the leading machine learning platform.

Deadline to participate in the Kaggle competition is Monday, September 5.

Tools & Resources

Release of IPython 5.0

Project Jupyter from July 07, 2016
We are pleased to announce the release of IPython 5.0 LTS (or Long Term Support). IPython is the Python kernel for Jupyter and the interactive Python shell; it provides a rich set of features for fluid interactive computation in Python at the terminal, in the Jupyter Notebook and across all other clients that support the Jupyter architecture.

Project Malmo, which lets researchers use Minecraft for AI research, makes public debut

The Official Microsoft Blog from July 07, 2016
Microsoft has made Project Malmo, a platform that uses the world of Minecraft as a testing ground for advanced artificial intelligence research, available for novice to experienced programmers on GitHub via an open-source license.

- github.com/Microsoft/malmo (July 11, GitHub - Microsoft)

Altair - Declarative statistical visualization library for Python

GitHub - ellisonbg from July 11, 2016
Altair is developed by **Brian Granger** and **Jake Vanderplas** in close collaboration with the **UW Interactive Data Lab**.

With Altair, you can spend more time understanding your data and its meaning. Altair's API is simple, friendly and consistent and built on top of the powerful Vega-Lite JSON specification. This elegant simplicity produces beautiful and effective visualizations with a minimal amount of code.

Computer-Assisted Keyword and Document Set Discovery from Unstructured Text

Gary King from July 04, 2016
We develop a computer-assisted (as opposed to fully automated) statistical approach that suggests keywords from available text without needing structured data as inputs.

Introducing the free Microsoft R Client

Microsoft, Revolutions from July 11, 2016
Over the years, we've shared several posts on using the ScaleR package to import, process, visualize and analyze large data sets with R. Until now, you needed to have access to a Microsoft R Server license to take advantage of the package. Now, you can use all of the capabilities of ScaleR free of charge with Microsoft R Client for Windows, which is available for download now.

A Future for R: A Comprehensive Overview

Henrik Bengtsson from June 25, 2016
The purpose of the future package is to provide a very simple and uniform way of evaluating R expressions asynchronously using various resources available to the

user.

In programming, a future is an abstraction for a value that may be available at some point in the future. The state of a future can either be unresolved or resolved. As soon as it is resolved, the value is available instantaneously. If the value is queried while the future is still unresolved, the current process is blocked until the future is resolved. It is possible to check whether a future is resolved or not without blocking. Exactly how and when futures are resolved depends on what strategy is used to evaluate them.

New Connected Vehicle Data Environments from the Following Projects of the Dynamic Mobility Application (DMA) Program are Now Available in the Research Data Exchange

Computing Community Consortium, CCC Blog from July 13, 2016
The Research Data Exchange (RDE) is a web-based data resource provided by the **USDOT Intelligent Transportation Systems (ITS) Program**. It collects, manages, and provides access to archived and real-time multi-source and multi-modal data to support the development and testing of ITS applications.

A Tale of Three Apache Spark APIs: RDDs, DataFrames, and Datasets

Databricks Blog, Jules Damji from July 14, 2016
In this blog, I explore three sets of APIs—RDDs, DataFrames, and Datasets—available in a pre-release preview of Apache Spark 2.0; why and when you should use each set; outline their performance and optimization benefits; and enumerate scenarios when to use DataFrames and Datasets instead of RDDs. Mostly, I will focus on DataFrames and Datasets, because in Apache Spark 2.0, these two APIs are unified.

Our primary motivation behind this unification is our quest to simplify Spark by limiting the number of concepts that you have to learn and by offering ways to process structured data.

Python 3 for Scientists

Open Astronomy from July 14, 2016
The primary aim of this page is to share information about useful new Python 3 features that may be useful to scientists for everyday work, as well as information about things you can do right now to prepare for the Python 3 transition, and how to try Python 3 (without necessarily switching over completely).

Careers

Managing Director Data Science Campus

UK Statistics Authority, The Office for National Statistics from July 12, 2016
Newport, South Wales We have been given a unique opportunity to build something rather special.

Civic Analytics Postgraduate Fellowship Program

NYU Center for Urban Science and Progress from July 13, 2016
New York, NY The Civic Analytics Postgraduate Fellowship Program invites three post-master's graduates from diverse disciplines for a full-time, nine-month urban data science program from September to June in residence at the **New York University Center for Urban Science and Progress** (CUSP).

Lead Product Designer at Fieldbook

Fieldbook from July 14, 2016
San Mateo, CA At **Fieldbook**, we're pursuing an ambitious mission to create the world's best tool for working with structured data. As our design lead, your challenge is to bring the power of relational data modeling to non-expert users.

Duke Population Research Institute - DATABASE ANALYST II

Duke Population Research Institute from July 08, 2016
Durham, NC The Duke Population Research Institute (DuPRI) is seeking a highly motivated Computational Social Sciences Consultant to work closely with DuPRI faculty and researchers in the social and biological sciences at Duke University.