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

(f)utility of preprint publishing?

This year is the 25th anniversary year of arxiv.org, the scrappy physics archive that has proven the utility and importance of webbased preprints for science communication. Why do so many great ideas start in physics? And why are scientists, funder, and publishers still so hung up about the role of preprints in the publishing process two and a half decades later? Earlier this summer, forprofit publishing **Elsevier** bought the Social Science Research Network, upsetting sociologists and inspiring them to start SocArxiv. It remains to be seen how publication and download patterns will change in the social sciences, if at all.

Richard Horton, EditorinChief of *The Lancet* counts preprint archives as part of The crisis in scientific publishing arguing that, "volume is not value....the defining quality of publishing is judgment." Prepublication does skip a number of traditional publishing steps like peer review and journalled editing, but does this leave the entire process devoid of judgment or does it simply relocate, reassign, and restructure how and by whom judgments are made? Focusing on the promise of open source technology rewritten copyright policies, **Carly Strasser** (Data-Driven Discovery, Moore Foundation), writes about the advantages of a coming preprint deluge to adapt scienceinaction and combine information on the fly. She argues that this is, "a rare moment in history [when] as a community, we can decide to make decisions about the technology and service ecosystem that can help to usher in a new era of research communication," and offers insightful pointers on how to build a sociotechnical publishing system that works for science and scientists.

In biomedicine, there is a related debate about preregistering experimental designs, a step that precedes publishing preprints. **Michael Frank** (Developmental Psychology, Stanford), argues that experiments should be preregistered and offers three key insights and four justifications for weaving preregistration into research workflows. Meanwhile, a team at **Pitt** has released a search engine for biomedical studies as a onestop infoshop for arXiv Quantitative Biology, bioRxiv, F1000Research, and PeerJ Preprints though not ASAPbio.

We will continue to cover news in preprints across the disciplines.

Edu + RegTech, FinTech, Risk Management = masters degree program$

Elsewhere at the intersection of data science, academia, and corporate information peddlers, **Thomson Reuters** and **KPMG** are joining forces with the **Data Science Institute at Imperial College London** to motivate "faster innovation in FinTech and RegTech industries". Competing for the attention and tuition dollars of future regulators and central bankers, **Yale University** announced a oneyear masters in systemic risk management with coursework taught by **Timothy Geithner** (but no data scientists).

Scroll down to Tools & Resources if you want to try using Zipline: A trading library for Python.

Data Science News

Follow-up of Kepler data yields more than 100 confirmed exoplanets

University of California-Santa Cruz, Newscenter from July 18, 2016
International team reports the biggest haul of new worlds yet uncovered by **NASA's** K2 mission, including many worlds that could potentially support life.

Model-based projections of Zika virus infections in childbearing women in the Americas

Nature Microbiology, GitHub - TAlexPerkins from July 25, 2016
The epidemic trajectory of this viral infection poses a significant concern for the nearly 15 million children born in the Americas each year. Ascertaining the portion of this population that is truly at risk is an important priority. Our results suggest that 1.65 (1.45–2.06) million childbearing women and 93.4 (81.6–117.1) million people in total could become infected before the first wave of the epidemic concludes. Based on current estimates of rates of adverse fetal outcomes among infected women, these results suggest that tens of thousands of pregnancies could be negatively impacted by the first wave of the epidemic. [full text + data]

Also in the American Zika epidemic:
- Local Transmission of Zika Likely Occurring in Florida, 4 cases confirmed (July 29, Medscape, **Robert Lowes**)

Microsoft Can't Shield User Data From Government, U.S. Says

Bloomberg, Kartikay Mehrotra from July 22, 2016
The U.S. says there's no legal basis for the government to be required to tell **Microsoft Corp.** customers when it intercepts their e-mail.

The software giant's lawsuit alleging that customers have a constitutional right to know if the government has searched or seized their property should be thrown out, the government said in a court filing. The U.S. said federal law allows it to obtain electronic communications without a warrant or without disclosure of a specific warrant if it would endanger an individual or an investigation.

Also in data security:
- As biometric scanning use grows, so does security risk (July 24, NBC News, **Chiara Sottile**)
- How the Chinese government fabricates social media posts for strategic distraction, not engaged argument (July 27, Working Paper, **Gary King**, **Jennifer Pan**, and **Margaret [Molly] Roberts**)

USC crafts tech system using mobile apps, AI to expand care

Health Data Management from July 25, 2016
Using mobile apps, "virtual doctors," data collection and analysis systems, world-class diagnostic and wearable sensors coupled with experiential design and engaging, expert patient health information, the VCC delivers wireless, on-demand access to **Keck Medicine of USC** experts while doctors go beyond telemedicine models for remote management and care of patients regardless of location.

More on health, technology & data:

- Inside Genomics Pioneer Craig Venter's Latest Production (July 25, MIT Technology Review, Business Report on Precision Medicine)
- Better Screening Using Big Data (July 05, Journal of Oncology Practice, **Debra Patt**)
- Academic Medical Orgs Leap into Precision Medicine Initiative (July 27, HealthIT Analytics)
- The Genomics Inflection Point: Implications for Healthcare (July 25, Rock Health; **Lauren Devos**, **Teresa Wang**, **Sandya Iyer**)

GECCO 2016 | Best Paper Nominations

Genetic and Evolutionary Computation Conference from July 20, 2016
The Genetic and Evolutionary Computation Conference (GECCO 2016) will present the latest high-quality results in genetic and evolutionary computation.

Predictive Models on Random Data

Data Skeptic podcast from July 23, 2016
This week is an insightful discussion with **Claudia Perlich** about some situations in machine learning where models can be built, perhaps by well-intentioned practitioners, to appear to be highly predictive despite being trained on random data. Our discussion covers some novel observations about ROC and AUC, as well as an informative discussion of leakage. [audio, 36:31]

Using Linked Census, Survey, and Administrative Data to Assess Longer-Term Effects of Policy: Proceedings of a Workshop—in Brief

The National Academies Press from July 24, 2016
The American Opportunity Study (AOS) is envisioned to create an intergenerational panel—using existing data at the person level—to study both social and economic mobility and the effectiveness of programs and policies that affect that mobility. ... [This workshop report], held on May 9, 2016, in Washington, D.C., was to more fully explore the value and potential uses of the AOS throughout a broad range of social science research.

More Census stuff:

- Using Census Bureau Data Made Easier: New Statistical Testing Tool Answers the Question "Is This Comparison Statistically Significant?" (July 21, U.S. Census Bureau, Random Samplings blog)
- New American Community Survey Tables and Products (July 21, United States Census Bureau press release)

Welcoming BIDS 2016 Data Science Fellows

Berkeley Institute for Data Science from July 25, 2016
We are thrilled to introduce our **Berkeley BIDS** 2016 cohort of data science fellows!

- **Rebecca Barter**, Statistics
- **Orianna DeMasi**, Electrical Engineering & Computer Science
- **Chris Holdgraf**, Neuroscience

- **Dmitriy Morozov**, Computational Research, Lawrence Berkeley National Lab
- **Laura Nelson**, Digital Humanities & Sociology
- **Alexandra Paxton**, Cognitive and Brain Science
- **Lauren Ponisio**, Environmental Science, Policy, and Management
- **Nelle Varoquaux**, Statistics

Yann LeCun - Session on Jul 28, 2016 - Quora

Quora from July 28, 2016
The first of many questions asked/answered:
What are some recent and potentially upcoming breakthroughs in deep learning? ...
The most important one, in my opinion, is adversarial training (also called GAN for Generative Adversarial Networks). This is an idea that was originally proposed by **Ian Goodfellow** when he was a student with **Yoshua Bengio** at the University of Montreal.

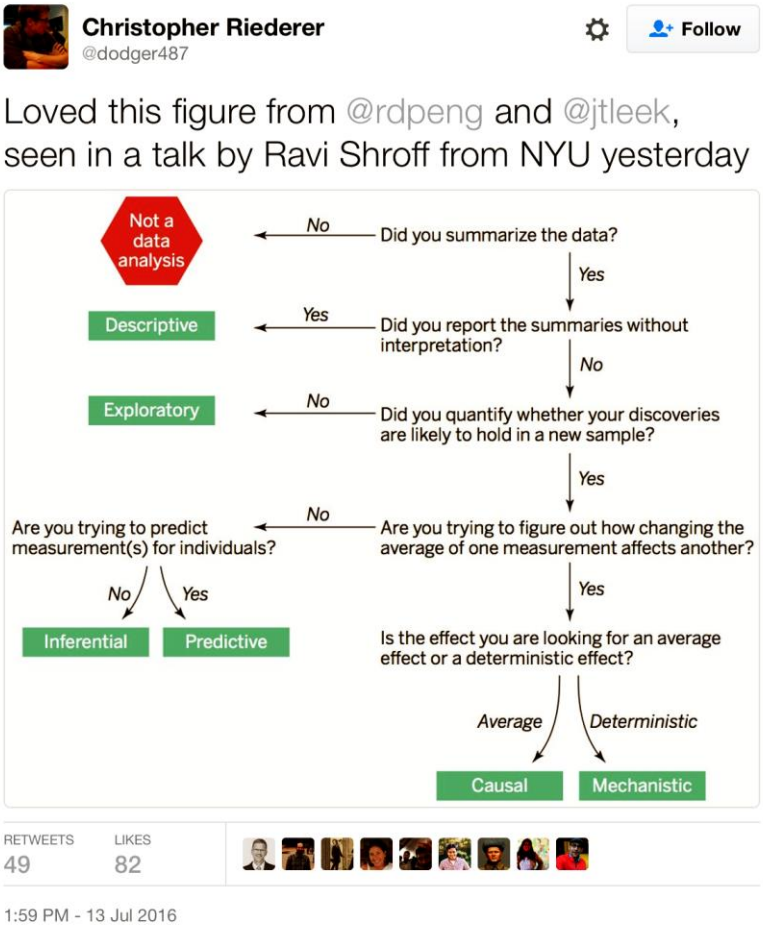
Athletes, coaches trying to find balance between analytics and 'gut feeling'

The Seattle Times from July 24, 2016
The new sports battleground is no longer about the value of a stats approach vs. a traditional one. Most teams by now realize that blending the two offers a better shot at winning. The bigger challenge is how to get humans to catch up to the numbers.

Also, in Sports + Data:
- Rise of Data Analytics in Football: The rise and rise of Leicester City (July 22, Outside of the Boot, **Jack Coles**)
- How USA Cycling is Using Data to Prepare for Rio (July 26, TrainingPeaks, YouTube)
- Putting it all together: A hockey systems, stats, tools, and talent evaluation primer (July 24, Blue Seat Blogs, **Dave Shapiro**)
- Who Do you Want Throwing your Darts – A Monkey or Eric Bristow? (July 7, Leaders Performance Institute, **Scott Drawer**)

Tweet of the Week

Twitter, Christopher Reiderer from July 13, 2016



Events

JupyterDay Atlanta 2016

Lead organizers **Tony Fast** and **Nick Bollweg** have a great event planned that is committed to education, open source, and community, which is organized by the local business, academic, and open source community.

Atlanta, GA Saturday, August 13, at Georgia Tech Research Institute (GTRI) Conference Center. [$$]

Artificial Intelligence Hackathon Event

Come one, come all to NYC's first Artificial Intelligence Hackathon, hosted in collaboration with **Clarifai**!

New York, NY Saturday, August 13 at General Assembly (902 Broadway, 4th Floor).

Data Science for Social Good Conference

This conference will highlight the successes, opportunities, and challenges faced by

the growing Data Science for Social Good community by bringing key members from each community (academia, governments, non-profits, foundations, social enterprises, and corporations) together to share best practices, learn from each other, and generate new collaboration opportunities.

Chicago, IL Wednesday-Thursday, August 24-25, at the University of Chicago Gleacher Center. [$$]

Master R Developer Workshop

This class will be a good fit for you if you have some experience programming in R already. You should have written a number of functions, and be comfortable with R's basic data structures (vectors, matrices, arrays, lists, and data frames). You will find the course particularly useful if you're an experienced R user looking to take the next step, or if you're moving to R from other programming languages and you want to quickly get up to speed with R's unique features.

New York, NY Monday-Tuesday, September 12-13, at AMA Conference Center (1601 Broadway) [$$$$]

Deadlines

Call for Participation -- Visualization in Data Science workshop

deadline: Conference

Baltimore, MD The workshop [preceding IEEE VIS 2016 on Monday, October 26] will feature a series of keynote presentations by leading data scientists covering visualization in data science. In addition, there will be presentations by speakers selected from submitted abstracts.

Deadline for submissions is Friday, August 5.

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.

CFP ICDM-2016 Workshop on Data Mining Systems and their Applications on the Cloud: CLOUDMINE

deadline: Conference

Barcelona, Spain The CLOUDMINE workshop aims to bring together researchers and practitioners working on cloud based data mining systems and applications. Part of ICDM 2016.

Deadline for workshop papers is Friday, August 12.

Open Cities Summit - Medialab-Prado Madrid

deadline: Conference

Madrid, Spain On October 5th 2016, one day before the International Open Data Conference in Madrid (Spain), the city council of Madrid in alliance with other international institutions will be hosting the Open Cities Summit at **Medialab Prado**.

The deadline for proposals is Monday, August 15.

The Urban Accelerator by MINI and HAX Futures | URBAN-X

deadline: Contest/Award

Our Mission: To catalyze, educate, invest in, and advocate for startups who are shaping the future of cities through technology.

Application deadline for URBAN-X 02 is Tuesday, September 6.

Workshop on Data and Algorithmic Transparency

deadline: Conference

New York, NY The Workshop on Data and Algorithmic Transparency (DAT'16) is being organized as a forum for academics, industry practicioners, regulators, and policy makers to come together and discuss issues related to increasing role that "big data" algorithms play in our society. It will be at Columbia University on Saturday, November 19.

Deadline for submissions is Friday, September 9.

Tools & Resources

Time Series Prediction With Deep Learning in Keras

Medium, IIOT from July 24, 2016

Time Series prediction is a difficult problem both to frame and to address with machine learning. In this post you will discover how to develop neural network models for time series prediction in Python using the Keras deep learning library.

How to Quantize Neural Networks with TensorFlow

TensorFlow from July 24, 2016

The computation demands of training grow with the number of researchers, but the cycles needed for inference expand in proportion to users. That means pure inference efficiency has become a burning issue for a lot of teams. That is where quantization comes in. It's an umbrella term that covers a lot of different techniques to store numbers and perform calculations on them in more compact formats than 32-bit floating point.

Introduction to Zipline: A Trading Library for Python

Quant Insti from July 18, 2016

Zipline is a Python library for trading applications that powers the Quantopian service. It is an event-driven system that supports both backtesting and live-trading. We demo how to install Zipline and how to implement Moving Average Crossover strategy, and calculate P&L, Portfolio value, etc.

Python 3 Readiness - Python 3 support table for most popular Python packages

Nar Chhantyal from July 28, 2016
This site shows Python 3 support for 360 most downloaded packages on PyPI

- 339 Green packages support Python 3
- 21 White packages don't support Python 3 yet.

Spark Release 2.0.0 | Apache Spark

Apache Software Foundation from July 27, 2016
Apache Spark 2.0.0 is the first release on the 2.x line. The major updates are API usability, SQL 2003 support, performance improvements, structured streaming, R UDF support, as well as operational improvements. In addition, this release includes over 2500 patches from over 300 contributors.

Careers

Department of Communication - Quantitative Models of Human Communication

University of California-Davis from July 26, 2016
Davis, CA Assistant Professor (Tenure Track) or Associate Professor, Quantitative Models of Human Communication. For this position, we seek a scholar with research interests focused on quantitative model building in communication.

Smith College: Massachusetts: Program in Statistical and Data Sciences - Assistant Professor of Statistical and Data Sciences

Interfolio, Smith College from July 27, 2016
Northampton, MA The successful candidate will be prepared to teach statistics at all levels, advise and mentor students, and must provide evidence of excellence in teaching and of an active research program.

IBM Social Good Fellow

IBM from July 27, 2016
Yorktown Heights, NY The IBM Social Good Fellowship is an opportunity for graduate students and postdoctoral scholars to develop their skills and develop data science solutions that benefit humanity

160725 Data Scientist - Enveritas.pdf

Enveritas from July 27, 2016
New York, NY Enveritas is a startup nonprofit that offers a new model of verifying sustainability. It was founded in 2016 by two experts in the coffee sector and has secured funding for platform development, testing and scaling. We are recruiting a global team of highly-motivated professionals for roles in software engineering, data

science, operations (on the ground in Africa, Asia and Latin America), and senior management.

Assistant Professor – Critical Computation and New Media - University of Toronto

University of Toronto from July 26, 2016
Toronto, ON, Canada We seek candidates with research expertise in Critical Computation and New Media, focusing on epistemological considerations surrounding computational intelligence and new media, the politics embedded in algorithms and code, media history, limits and/or comparative technologies of computation, and software studies related to new media.

Postdoctoral Associate, Statistical Modeling For Evaluating Human Motion

MIT, Stirling Research Group from July 29, 2016
Cambridge, MA POSTDOCTORAL ASSOCIATE, Aeronautics and Astronautics, to join the Statistical Modeling for Evaluating Human Motion project. The Stirling Research Group has an opening in the domain of statistical modeling for a collaborative project between **MIT** and the **University of Michigan**, sponsored by the **U.S. Army**.