NYU Data Science Community Newsletter features journalism, research papers, events, tools/software, and jobs for August 19, 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.

Infrastructures for data sharing: Astro, exome, ocean…and kaggle?

The infrastructure for organizing data has hit a few milestones this summer.

Within data science for science, astrophysics is often held as a model case for how other sciences can do data-driven and computational research. How many disciplines are now copying the concept behind arxiv.org for preprint handling? We can thank physics for that bright idea.

So what does data storage and data sharing look like in astrophysics? **NASA** recently launced a new agency-wide data portal. NASA and ESA together maintain portals for astrophysics and space telescope data, but the National Radio Astronomy Observatory, the National Optical Astronomy Observatory, and the European Southern Observatory also maintain data portals. Then **Harvard** has a catalogue of derived data products and a website that links to specific instruments and they data they have gathered. In other words, there is no single data sharing platform even for a single sub-discipline with no human subjects issues.

Genetics - a field with strong patient-protection requirements that can hamper data sharing - can now lean on **MIT's Broad Institute** for protein-genomic data. The Exome Aggregation Consortium (ExAC) database for the human exome (protein-coding region) allows clinical geneticists to compare individual patient data to patterns in the general public. For a taste of what ExAC can do, check out **Monkol Lek** et al. who explain how the 60,000 human exomes in ExAC can be used for efficient "filtering of candidate disease-causing variants, and for the discovery of human 'knockout' variants in protein-coding genes."

Elsewhere in patient data sharing, the debate from last week's newsletter in which the *New England Journal of Medicine* opposed the larger body of medical journal publishers' (ICMJE) recommendation that study data need to be released within 6 months of publication continues. This week in *PLOS Medicine*, **Jean-Paul Chretien** et al. make a case for sharing patient data "for all health research", not just outbreaks like Ebola and Zika, to improve rapid response to public health crises.

In underwater sonar and fisheries ecology, the Water Column Sonar Data Archive is overseen by **NOAA** and collaborators, **National Marine Fisheries Service** and the **University of Colorado**, to collect and share "27 TB of raw sonar data" useful for studying fisheries and other coastal ocean ecology questions.

Amidst all of this field-specific data-sharing work, I found Kaggle's announcement that they want to be the "the home of open data" and the singular site for open data sharing rather ambitious.

As an organizational sociologist, it seems more likely that each professional sub-field, both in academia and beyond, will maintain their own platforms for data sharing so that they can be responsive to the particular sociotechnical needs of their own community. But this siloing does present challenges for cross-disciplinary work and adherence to best-practices in software development and data management.

Expect to see more here about platforms for data sharing.
- Laura

Data Science News

Commonwealth Awards $5 Million to UMass Amherst to Support New Data Science Collaborative

UMass Amherst from August 17, 2016
"The Baker-Polito administration announced a $5 million grant to the **University of Massachusetts Amherst** to establish the UMass Amherst Data Science/Cybersecurity Research and Education Collaborative, a public-private partnership designed to accelerate data science innovation in the Pioneer Valley region of Western Massachusetts." [video, 1:37]

Not such good news in Ohio:
- **University of Akron** puts the brakes on data science center (August 17, cleveland.com)

Yoshua Bengio's answer to What are the pros and cons of Generative Adversarial Networks vs Variational Autoencoders?

Quora, Yoshua Bengio from August 16, 2016
"An advantage for VAEs (Variational AutoEncoders) is that there is a clear and recognized way to evaluate the quality of the model (log-likelihood, either estimated by importance sampling or lower-bounded). Right now it’s not clear how to compare two GANs (Generative Adversarial Networks) or compare a GAN and other generative models except by visualizing samples."

"A disadvantage of VAEs is that, because of the injected noise and imperfect reconstruction, and with the standard decoder (with factorized output distribution), the generated samples are much more blurred than those coming from GANs."

More experts making sense of things:
- Answer to How does Keras compare to other Deep Learning frameworks like Tensor Flow, Theano, or Torch? (August 15, Quora, **Francois Chollet**)
- Introducing Variational Autoencoders (in Prose and Code) (August 12, Fast Forward Labs Blog, **Miriam Shiffman**)
- Ian Goodfellow on Quora (August 12, Quora, **Ian Goodfellow**)
- The Hardest Part About Microservices: Your Data (July 14, Christian Posta, Software Blog)

Ranking Relevance in Yahoo Search

KDD 2016, KDD Topics from August 15, 2016
"In this paper, we give an overview of the solutions for relevance in the **Yahoo** search engine. We introduce three key techniques for base relevance – ranking functions, semantic matching features and query rewriting." [Yahoo search engineers are giving away the merchandise before their store closes; download pdf for full text.]

Astronomers Are On A Celestial Treasure Hunt. The Prize? Planet Nine

NPR, Weekend Edition Saturday from August 13, 2016
"Nobody's actually seen the new planet. The reason astronomers think it's out there is the strange behavior of some smallish objects in the Kuiper Belt, a collection of celestial objects orbiting in the outer reaches of the solar system."

Data Science Challenges

KDnuggets, Neil Lawrence from August 17, 2016
"This post is thoughts for a talk given at the **UN Global Pulse** lab in Kampala as part of the second Data Science in Africa Workshop at the UN Global Pulse Lab in Kampala, Uganda. It covers challenges in data science."

"Data is a pervasive phenomenon. It affects all aspects of our activities. This diffusiveness is both a challenge and an opportunity. A challenge, because our expertise is spread thinly: like raisins in a fruitcake, or nuggets in a gold mine. It is an opportunity, because if we can resolve the challenges of difussion we can foster a multi-faceted benefits across the entire University."

The Perks and Perils of Interdisciplinary Research

The Social Science Research Council from August 16, 2016
"In this essay, **Erin Leahey** discusses how interdisciplinary research affects academic careers, the visibility of research, and scholarly productivity. She also reports on an ongoing project that explores the ways in which universities support interdisciplinary work."

More on practicing interdisciplinary research:

- [1608.03251] Choosing Collaboration Partners. How Scientific Success in Physics Depends on Network Positions (August 10, arXiv, Physics > Physics and Society; **Raphael H. Heiberger**, **Oliver J. Wieczorek**)

10 Rules Of Thumb For Hockey Analysts

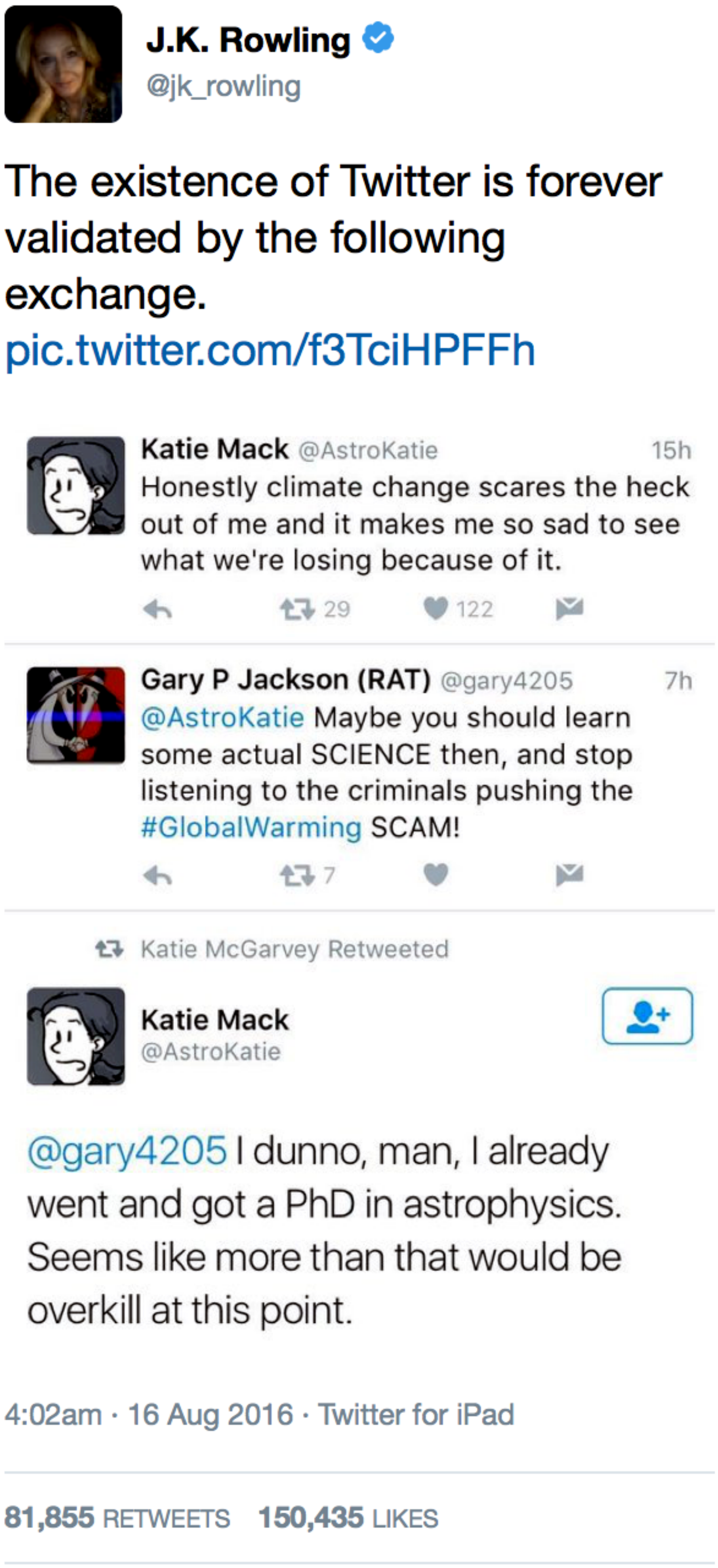
Hockey Graphs, Jack Han from August 16, 2016
"1. The point of hockey is to create goal differential. The point of hockey analysis is to find ways to improve it."

Also, upcoming Hockey Analysis Conferences:

- **RIT** Hockey Analytics Conference (on Saturday, September 10)
- **Babson College** Hockey Analytics Conference "Analytics on Ice: The Long Change" (on Saturday, October 1)

Tweet of the Week

Twitter, J.K. Rowling from August 16, 2016



J.K. Rowling ✓
@jk_rowling

The existence of Twitter is forever validated by the following exchange. pic.twitter.com/f3TciHPFFh

Katie Mack @AstroKatie 15h
Honestly climate change scares the heck out of me and it makes me so sad to see what we're losing because of it.

↩ 29 ♥ 122

Gary P Jackson (RAT) @gary4205 7h
@AstroKatie Maybe you should learn some actual SCIENCE then, and stop listening to the criminals pushing the #GlobalWarming SCAM!

↩ 7 ♥

Katie McGarvey Retweeted

Katie Mack
@AstroKatie

@gary4205 I dunno, man, I already went and got a PhD in astrophysics. Seems like more than that would be overkill at this point.

4:02am · 16 Aug 2016 · Twitter for iPad

81,855 RETWEETS **150,435** LIKES

Events

Challenges and Advances on Big Data in Neuroimaging

"This conference is to bring together statisticians, clinicians, data scientists and graduate students in neuroimaging to exchange ideas of the current challenges and developments in statistical methods and applications in the research of big neuroimaging data."

Cleveland, OH Thursday-Friday, August 25-26. [$$$]

Astro Hack Week 2016

"AstroHackWeek is, in part, a summer school. The mornings will offer lectures and exercises covering essential skills for working effectively with large astronomical datasets. Past years have seen topics such as machine learning, Bayesian inference, frequentist statistics, databases, numerical Python, and visualization."

Berkeley, CA Monday-Friday, August 29-September 2, at the **University of California-Berkeley**. Registration is open.

International Data Week, September 11-17

Denver, CO "The International Data Week will bring together data scientists, researchers, industry leaders, entrepreneurs, policy makers and data stewards to explore how best to exploit the data revolution to improve our knowledge and benefit society through data-driven research and innovation." [$$$]

AoIR 2016 - BrowseSessions

Berlin, Germany "The theme for this year’s conference is Internet Rules!"

Wednesday-Saturday, October 5-8, hosted by **Alexander von Humboldt Institute for Internet and Society** and the **Hans Bredow Institute for Media Research**. [$$$]

Machine Learning Unconference

San Francisco, CA "Please join us [Open AI] for our first Machine Learning Unconference, an experimental gathering driven by its participants rather than an organizing committee."

Friday-Saturday, October 7-8, at OpenAI.

Women in Statistics and Data Science Conference

"The **American Statistical Association** invites you to join us for the 2016 Women in Statistics and Data Science Conference—the only one for the field tailored specifically for women!"

Charlotte, NC Thursday-Saturday, October 20-22. [$$$]

Deadlines

Reproducibility case studies

deadline: Journal/Publishing

Justin Kitzes (jkitzes@berkeley.edu) is seeking case studies that provide concrete workflow examples that embody reproducible research. **Berkeley Institute for Data Science** (BIDS) plans to publish the case studies and credit authors accordingly while also collecting the studies in a GitHub repository. The procedural and formatting instructions for the case studies are rather specific, and project reviewers will appreciate your attention to detail.

More upcoming reproducibility workshops:

- Fall 2016 Center for Open Science Workshop (October 06 at University of Wisconsin-Madison)
- 1-day Reproducibility Conference Coming to Columbia University December 2016! (December 09 at Columbia University)

Computational Approaches to Social Modeling workshop (ChASM 2016)

deadline: Conference

Bellevue, WA The workshop will precede SocInfo 2016 on Monday, November 14, 2016.

Deadline for workshop papers submission is Saturday, August 27.

2016 AGU Data Visualization Storytelling Competition: Requirements, Criteria, and Award Information

deadline: Contest/Award

"The competition is open to all students (2 and 4 year undergraduate and graduate students) who are U.S. citizens. Individual submissions and team submissions (up to three people) will be accepted. All teams must identify a project lead. The project lead is responsible for submitting the application. Submissions must address one of the three themes described in the evaluation criteria below."

Deadline to apply is Thursday, September 15.

NASA's Space Robotics Challenge: The Tasks, the Prizes, and How to Participate

deadline: Contest/Award

"Before the SRC itself, there’s a qualification round. A simulated R5 will have to identify a pattern of colored lights blinking on a panel, push a button, and walk through a doorway without falling down. The 20 top scoring teams (based on speed and accuracy) will each get US $15,000 and move on to the virtual competition."

Deadline to register will remain open until September 16.

Computational Health

deadline: Conference

Perth, Australia "We invite research contributions for the 26th World Wide Web Conference Computational Health Track."

Deadline for abstract submission is Wednesday, October 19.

CDS News

Dissertation Defense -- Yoni Halpern (Monday, August 22nd, 10am)

David Sontag from August 18, 2016
Ph.D. student **Yoni Halpern's** thesis defense, next Monday Aug. 22nd at 10am at 719 Broadway (intersection with Washington Place), 12th floor, Room 1221.

Yoni's thesis is at the intersection of machine learning and health care, and he introduces several new tools for data science that should be of broad interest to the **NYU** and Data Science communities.

NSF Award Search: Award#1637108 - RIDIR: Collaborative Research: Computational and Historical Resources on Nations and Organizations for the Social Sciences (CHRONOS)

National Science Foundation from August 11, 2016
"The project will demonstrate how computational techniques can aid both qualitative and quantitative social science research on a range of areas of major public interest, expanding knowledge about terrorism, intelligence, international trade and aid."

Pablo Barberá Wins the 2016 Franklin L. Burdette/Pi Sigma Alpha Award

Political Science Now from August 18, 2016
Examining individuals using **Twitter** accounts in Spain, Germany, and the U.S., Barberá constructs a dynamic measure of the political ideology of Twitter users based on who they follow. His analysis shows that users joined by weak ties are incidentally exposed to diverse political opinions and become more moderate over time.

Tools & Resources

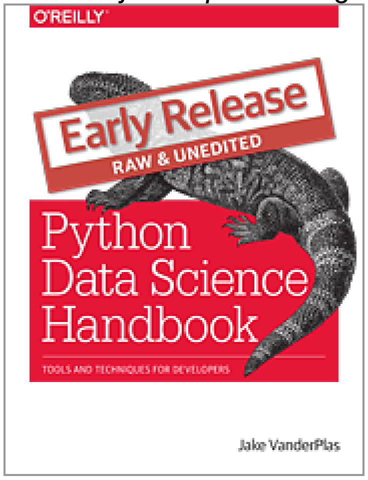
Scalable data science with R

O'Reilly Media, Federico Castanedo from August 10, 2016
"In the particular case of R, data size problems usually arise when the input data do not fit in the RAM of the machine and when data analysis takes a long time because parallelism does not happen automatically. Without making the data smaller (through sampling, for example) this problem can be solved in two different ways:"

- "Scaling-out vertically", using a machine with more available RAM.
- "Scaling-out horizontally"

Supplemental materials for my OReilly project, the Python Data Science Handbook

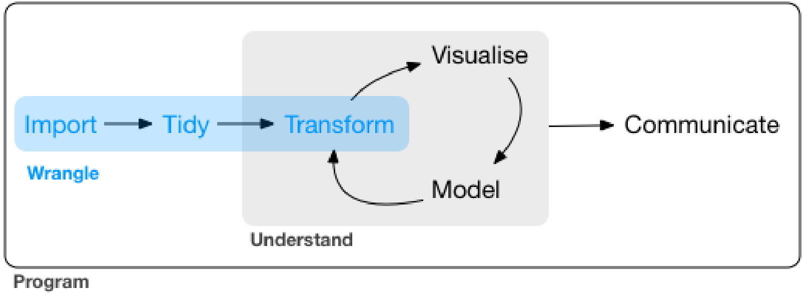
GitHub - jakevdp from August 11, 2016


O'REILLY
Early Release
RAW & UNEDITED
Python Data Science Handbook
TOOLS AND TECHNIQUES FOR DEVELOPERS
Jake VanderPlas

"See also the free companion project, A Whirlwind Tour of Python."

R for Data Science - Wrangling section

R for Data Science book, Hadley Wickham from August 17, 2016



"In this part of the book, you’ll learn about data wrangling, the art of getting your data into R in a useful form for visualisation and modelling."

Colorgorical

Connor Gramazio from August 18, 2016

"To generate a palette with n colors, just enter the number of colors you want and click Generate. Bigger palettes will take longer than smaller palettes to make. Results will automatically appear when ready."

"For greater detail, please consult our paper or the source code."

Bit By Bit: Social Research in the Digital Age

Matthew Salganick from August 18, 2016

"The central premise of this book is that the digital age creates new opportunities for social research. Researchers can now observe behavior, ask questions, run experiments, and collaborate in ways that were simply impossible in the quite recent past. Along with these new opportunities also come new risks; researchers can now harm people in ways that were impossible in the quite recent past. The source of these opportunities and risks is the transition from the analog age to the digital age. This transition did not happen all at once—like a light-switch turning on—and, in fact, the transition is not yet complete. But, by this point we've seen enough to know that something big is happening."

Building a Data Pipeline with Airflow

Mark Litwintschik from August 01, 2016
"When I first began using Airflow I was relieved to see that at it's core is a plain and simple Flask project. I was able to read through it's Python codebase in a morning and have confidence that I could work my way through it's architecture."

"In this blog post I'll setup a data pipeline that takes currency exchange rates, stores them in PostgreSQL and then caches the latest exchange rates in Redis."

Careers

Tenured and tenure track faculty positions

Assistant Professor - Social Impact of Science, Medicine, and Technology

University of California-San Diego from August 01, 2016
San Diego, CA "These positions are part of a bold multi-discipline, multi-year initiative in practical ethics that spans the entire University. In this initiative, traditional normative ethical concerns are embedded in cross-disciplinary research areas such as the humanities, social sciences, physical sciences, biological sciences, geosciences, engineering, business, and medicine."

Assistant / Associate Professor of Research Ethics, KU Medical Center

University of Kansas from August 18, 2016
Kansas City, MO "The Department of History and Philosophy of Medicine is recruiting to its faculty a mid-level scholar in the history of science and medicine, medical ethics, science studies, or the philosophy or social sciences of science and medicine."

Employment | Linguistics | New York University

New York University from August 18, 2016
New York, NY "The Department of Linguistics at New York University invites applications for a tenure-track position at the assistant or associate level in computational phonology, beginning September 1, 2017."

Department of Communication, Cornell University

Cornell University from August 18, 2016
Ithaca, NY "The Department of Communication at Cornell University is searching for a 9-month, tenure-track faculty member in Communication and Social Behavior at the Assistant Professor level."

Full-time, non-tenured academic positions

Scientific Application Developer, Physics

Princeton University from August 16, 2016
Princeton, NJ "Princeton University is seeking one or more Scientific Application Developers/Software Engineers to work with the High Energy Experiment group in the Physics Department. High Energy Physics (HEP) focuses on understanding the elementary particles that are the fundamental constituents of matter and their interactions."

Senior Informatics Researcher - Renaissance Computing Institute

University of North Carolina-Chapel Hill from August 19, 2016
Chapel Hill, NC "This position provides leadership to RENCI research projects in the areas of computational, informatics, and domain expertise to support genomics and genetics research."

Full-time positions outside academia

Data Scientist, ONE

ONE from August 02, 2016
London, England "ONE is a global campaign and advocacy organization co-founded by Bono and backed by more than seven million people from around the world" ... "The Data Scientist will be responsible for advising ONE on the potential of data to provide new insights into ONE's work and mission using advanced data mining, statistical analysis, machine-learning and visualization techniques."

Postdocs

Genomics Postdocs-Computational Biology Department - Carnegie Mellon University

Carnegie Mellon University from August 18, 2016
Pittsburgh, PA "Applications are invited for postdoc positions in computational genomics in Jian Ma's lab at the Computational Biology Department in the School of Computer Science at Carnegie Mellon University."

Postdoc position available immediately

t+z statistics, Columbia University from August 17, 2016
New York, NY "A full-time postdoctoral position is available beginning immediately in the research group of **Professor Tian Zheng** [Columbia University, Department of Statistics] working on analysis of large spatiotemporal data sets, in close cooperation with our collaborators in neural imaging."