Scientists’ ugly fights about data sharing

As a sociologist, it is not at all surprising that sharing is hard. The technology for sharing is often difficult. Researchers found that even when astronomers shared their data and software, link rot and other maladies of technological decay left 44% of these good intentions practically meaningless ten years on. NSF now requires data management plans, which should help, data hoarding lingers in the culture of some disciplines.

To me, the data sharing fight is a major professional shift about how to do science that has been provoked by the increasingly computational nature of scientific practice. Data are digital, many methods are code-based, and it is more possible to duplicate and share digital data than, say, specialized cell lines or large, expensive instruments. The problem, then, is that the act of sharing these intermediate products - code and data - runs counter to other tenets in the culture of science. In the fading ownership model, which dovetails nicely with a broader capitalist ethos, data belong to the scientist or team who gathered them. The paradigm gathering steam is moving towards an “open science” or research trust model of data in which data belongs to science, and should be stored collectively in a trust.

The fight between The New England Journal of Medicine and most other top researchers and publishers in medicine made it appear that NEJM cares most about protecting scientists’ professional careers by allowing them to hoard their own data. It is a relief to see Nature’s editorial board call for the designers of algorithms to make public the source of the data sets they use to train and feed them”. Nature also celebrated changes to the U.S. Department of Health and Human Services policy on disappointing clinical trials: those failures must be disclosed, even if they don’t help specific scientists’ careers. If only this applied to all research, not just clinical trials.

One major concern about sharing human-generated data is that privacy protections have to be phenomenal. Gary King, a social scientist and statistician at Harvard, released a Private data sharing interface (PSI), “to enable researchers in the social sciences and other fields to share and explore privacy-sensitive datasets with the strong privacy protections of differential privacy”. While we’re talking about practical tools and policies for data sharing, here is a frameworks for citation practices for citing software, software packages, and datasets. These types of tools are often presented as a sociotechnical response to the professionalization gap. If researchers are cited for their software and data similar to the way they receive citations for publication, perhaps their careers can be boosted, not scooped, by sharing their data, software, and algorithms.

The juiciest long read of the week is from statistician Andrew Gelman who filets psychologist Susan Fiske for writing about, “self-appointed data police” who crash careers by publicly exposing errors or catching researchers in games of p-hacking for professional gain.
Data Science News

Illumina, Global Leader of DNA Sequencing, Dominates! the market

*Fast Company* Christina Farr from September 19, 2016

Illumina sells sequencers to hospitals and labs and consumer-style tests to individuals. "With spin-off investments Grail and Helix, and a new software-savvy CEO, Illumina is poised to make DNA an even bigger part of your life." But what is at stake when a private company is this dominant in a market with huge implications for public good?

Non-profit models for large-scale health science:

- Allen Institute publishes highest resolution map of the entire human brain to date (September 15, *Allen Institute for Artificial Intelligence*)
- A Diversity of Genomes - New DNA from understudied groups reveals modern genetic variation, ancient population shifts (September 21, *Harvard Medical School*)

In text-as-data stack this week from Harvard: Image-to-Markup Generation


"Building on recent advances in image caption generation and optical character recognition (OCR), we present a general-purpose, deep learning-based system to decompile an image into presentational markup." ... "Our model does not require any knowledge of the underlying markup language, and is simply trained end-to-end on real-world example data."

Also in Language & Text:

- Computers Can Sense Sarcasm? Yeah, Right (August 26, Scientific American, Jesse Emspak)
- Google acquires natural language understanding startup Api.ai (September 19, VentureBeat, Jordan Novet)
- Etsy buys Blackbird Technologies to bring AI to its search (September 19, TechCrunch, Ingrid Lunden)

Amazon Says It Puts Customers First. But Its Pricing Algorithm Doesn’t.

*ProPublica; Julia Angwin and Surya Mattu* from September 20, 2016

"Amazon bills itself as “Earth’s most customer-centric company.” Yet its algorithm is hiding the best deal from many customers."

Show and Tell: image captioning open sourced in TensorFlow

*Google Research Blog, Chris Shallue* from September 22, 2016

Open source release "contains significant improvements to the computer vision component of the captioning system, is much faster to train, and produces more detailed and accurate descriptions" Git repo at the bottom of the blog linked here.
A Blockchain is a digital platform that hosts a digital ledger of transactions and shares it among a distributed network of computers. The cryptography technology allows each participant on the network to manipulate the ledger in a secure way without the need for a central authority.

More Blockchain:
- Is an Editable Blockchain the Future of Finance? (September 20, MIT Technology Review, Jamie Condliffe)
- Blockchain beyond Bitcoin (September 21, Raconteur Media, UK)

In data for good this week: Google Maps teaches algorithm to identify city trees
- Futurity, Robert Perkins-USC from September 19, 2016
  "A new method uses data from satellite and street-level images, such as the ones you can see in Google Maps, to automatically create an inventory of street trees that cities may use to better manage urban forests."

More Data for Good:
Matchmaking to Find Homes for Refugees (September 19, Bloomberg View, Alex Teytelboym & Scott Duke Kominers)
P2P insurance firm Lemonade launches out of stealth, powered by chatbots, morals, and big bucks (September 21, VentureBeat, Paul Sawers)
Computational Sustainability Virtual Seminar - Measuring progress towards sustainable development goals with machine learning (takes place on September 27, CompSustNet)
Data for Good Exchange 2016 event (takes place on September 25, Bloomberg L.P.)

Tweet of the Week
Twitter, Stef Lewandowski from September 22, 2016
Stef Lewandowski
@stef

You are in a startup. All around is a burning runway. There are exits to the North and East. You have a bootstrap. There is a VC here.

> |

1:29pm · 22 Sep 2016 · TweetDeck

26 RETWEETS  59 LIKES

Peter MacRobert @PeteM
@stef ask the VC for money

Stef Lewandowski @stef
@PeteM the VC shakes his head: "You never ask a VC for money. Always ask for advice". The runway burns you to a crisp. You are in a startup.
Events

The 2016 ICPSR Data Fair!
   Online Monday-Thursday, 26-29 September 2016. [free]

Columbia Data Science Student Challenge
   New York, NY Friday-Saturday, 30 September-1 October 2016, at Davis Auditorium (4th Fl Schapiro Center). [free]

Data Science in Agriculture Summit
   Chicago, IL 10 October 2016, precedes the NSF Midwest Big Data Hub All Hands Meeting. [free]

Lucene/Solr Revolution 2016
   Boston, MA Tuesday-Friday, October 11-14. [$$$$]

All-Hands Meeting – Midwest Big Data Hub
   Chicago, IL Tuesday-Wednesday, 11-12 October 2016 at Big Ten Conference Center in Rosemont, near O'Hare Airport. [free]

MinneMUDAC: Dive Into Water Data event
   Eden Prairie, MN Saturday, 5 November 2016 at Optum (13625 Technology Drive). [free]

Partnering For Cures medical research conference
   New York, NY Sunday-Tuesday, 13-15 November 2016 [$$$$]

Deadlines

Contest/Award

MIT Sloan Sports Analytics Conference Research Paper Competition
   Deadline for abstract submission is Monday, 28 September 2016.

Conference

Social Network Analysis and Computational Social Science Track, 26th World Wide Web Conference
   Perth, Australia Deadline for abstracts' submissions is Wednesday, 17 October 2016.

NIPS Symposium - Machine Learning and the Law
   Barcelona, Spain Symposium is Thursday, 8 December 2016, at NIPS 2016. Deadline for papers submissions is Thursday, 3 November 2016.

More NIPS 2016 Workshops (submission deadline, workshop date):
   Recurrent Neural Networks Symposium (15 October 2016, 8 December 2016)
   Future of Interactive Learning Machines Workshop (14 October, 9 or 10 December)
   NIPS Workshop on learning intuitive physics (14 October, 9 December)
   Workshop on Machine Learning for Health (28 October, 9 December)
   Approximate Inference (1 November, 9-10 December)
   Interpretable ML for Complex Systems (20 October, 9 December)
   Neural Abstract Machines & Program Induction (14 October, 10 December)
   End-to-end Learning for Speech and Audio Processing Workshop (20 October, 10
PyAstro17 applications are open!

Leiden, The Netherlands The workshop will be held on May 8-12, 2017 at the Lorentz Center. Deadline for applications to participate in the Python in Astronomy 2017 workshop is Friday, 9 December 2016.

Big data conference: Strata + Hadoop World, March 13 - 16, 2017
San Jose, CA The deadline for speaker proposals is Friday, 30 September 2016.

Education Opportunity
Global Fellowship in Human Rights
Students identify human rights organizations with which to work and propose their own summer projects and/or internships. The organizations should have the capacity to host students and incorporate them into the substantive aspects of their human rights work in meaningful ways. Deadline for applications is Saturday, 30 October 2016.

Journal/Publishing
CFP: “Countercultures of Data” – Special Issue of Philosophy & Technology
Deadline for paper submissions is 24 October 2016.

RFP
Russell Sage Foundation Issues RFP for Computational Social Science Projects
Through the initiative, grants of up to $150,000 will be awarded in support of innovative social science research that brings new data and methods to bear on questions of interest in the foundation’s core programs in Behavioral Economics; the Future of Work; Race, Ethnicity, and Immigration; and Social Inequality. Deadline for Letters of Inquiry is Wednesday, November 30.

NYU Center for Data Science News
GoViral Study
Chunara Lab, NYU Tandon School of Engineering from September 22, 2016
Help an NYU data and public health researcher Rumi Chunara from home while learning what makes you sick. [video, 2:00]

[1609.03126] Energy-based Generative Adversarial Network
arXiv, Computer Science > Learning; Junbo Zhao, Michael Mathieu, Yann LeCun from September 14, 2016
We introduce the "Energy-based Generative Adversarial Network" model (EBGAN) which views the discriminator as an energy function that associates low energies with the regions near the data manifold and higher energies with other regions. Similar to the probabilistic GANs, a generator is trained to produce contrastive samples with minimal energies, while the discriminator is trained to assign high energies to these generated samples.

Upcoming No-Cost Data Science Events in NYC
• The Robotic Scientist: Automating discovery, from cognitive robotics to computational biology on September 27
  Brooklyn, NY 3:30 p.m., Tandon School of Engineering, MakerSpace Lecture Hall (6
• **Human-oriented Technology in Data-Intensive Startups** on September 28
  New York, NY Our kick-off event features short talks and a panel with three experts at cutting-edge data intensive startups. -- 7 p.m., Pulitzer Hall, Columbia University.

• **NYU Center for Data Science Master's Program Information Webinars** on September 28
  Online To learn more about the MSDS program, please attend our webinar for an overview of the program and application process, 9:30 a.m.

• **Moore-Sloan Data Science Lunch Seminar Series** on September 28
  New York, NY Michael Z. Gill (The Governance Lab, NYU) Lunch provided. -- Wednesday, September 21, from 12:30 - 1:30 p.m. at the NYU Center for Data Science, 60 5th Avenue, 7th Floor.

• **Clinical Trial Transparency and Reproducibility Discussion Panel and Workshop at NYU** on September 29
  New York, NY; NYU, Center for Open Science, and AllTrials USA at 1:30 pm at Langone Medical Center, Skirball 4th Floor Seminar Room

• **NYU to Host Public Debate: “Do Replication Projects Cast Doubt on Many Published Studies in Psychology?”** on September 29
  New York, NY NYU’s Center for Mind, Brain and Consciousness will host the debate. Starts at 5 p.m. in the Kimmel Center for University Life.

• **Text as Data Speaker Series** on September 29
  New York, NY Regina Barzilay (MIT) — How Can NLP Help Cure Cancer? — from 4-5:30 p.m. in room 217, 19 West 4th St.

• **OCR Friday with Blacki Migliozzi** on September 30
  Brooklyn, NY Our studio is excited to kick off our OCR Friday fall series with Blacki Migliozzi! Blacki is a data journalist at Bloomberg News, where he develops interactive infographics for data-driven stories. Starts at 6:30 p.m.

• **Ethics of Artificial Intelligence** on October 14
  New York, NY Friday-Saturday, 14-15 October 2016, at NYU.

• **Gravitational Waves, Black Holes and Einstein's Legacy** on October 14
  New York, NY NYU is honored to have Rainer Weiss, Emeritus Professor of Physics at MIT and co-founder of the Laser Interferometer Gravitational-wave Observatory (LIGO) give a public lecture, part of its prestigious Klosk lecture series. Starts at 11 a.m.

• **2nd Annual NYC Taxi and Limousine Commission Hackathon Kickoff** on October 15
  At General Assembly (10 E 21st St, 2nd Floor).

• **1-day Reproducibility Conference Coming to Columbia University December 2016!** on December 09
  New York, NY Columbia University and other New York City research institutions are hosting a one-day symposium to showcase a robust discussion of reproducibility and research integrity among leading experts, high-profile journal editors, funders and researchers.

• **StanCon** on January 21
  New York, NY Saturday, 21 January 2017, starting at 9 a.m., Davis Auditorium, Columbia University

**Tools & Resources**

**High Performance Compute for the JVM - A Prerequisite for DL4J (Part 1)**

*Skymind engineering blog, Susan Eraly and Vyacheslav Kokorin* from September 12, 2016

"To make the benefits of hardware acceleration available to the JVM we have to deal with the
complexity and challenges of bridging Java and native C++.

**Reinforcement Learning for Torch: Introducing torch-twrl**
*Twitter blog, Kory Mathewson* from September 16, 2016
"To develop effective algorithms, rapid iteration and testing is important, torch-twrl aims to make implementing and innovating fast and easy."

**CodeBuff smart formatter**
*GitHub - antlr* from September 16, 2016
"This repository is a step towards a universal code formatter that uses machine learning to look for patterns in a corpus and format code using those patterns."

**Copybara: A tool for transforming and moving code between repositories.**
*GitHub - google* from September 16, 2016
"The most common use case involves repetitive movement of code from one repository to another. Copybara can also be used for moving code once to a new repository."

**A gentle introduction to random forests using R**
*Kailash Awati, Eight to Late* from September 20, 2016
"I describe random forests...and then do a demo using the R randomForest library."

**ZaliQL: A SQL-Based Framework for Drawing Causal Inference from Big Data**
*arXiv, Computer Science > Databases; Babak Salimi, Dan Suciu* from September 13, 2016
"We describe a suite of techniques for expressing causal inference tasks from observational data in SQL....we introduce several optimization techniques that significantly increase speed."

**Visdown - Markdown for Visualisation**
*Amit Kapoor* from September 20, 2016
Make visualisations using only markdown

**Architecting smart applications on top of Apache Spark**
*Medium, Stepan Pushkarev* from September 21, 2016
"a set of architectural guidelines for building a platform to expose analytics as services for enterprise solutions." Includes a review of open source components.

**Global Historical Daily Weather Data in BigQuery**
*Google Cloud Platform Blog, Lak Lakshmanan* from September 22, 2016
"From 80,000+ stations in 180 countries, spans decades and has been quality-checked to ensure that it's temporally and spatially consistent."

**Show and Tell: image captioning open sourced in TensorFlow**
*Google Research Blog, Chris Shallue* from September 22, 2016
Open source release "contains significant improvements to the computer vision component of the captioning system, is much faster to train, and produces more detailed and accurate descriptions" Git repo at the bottom of the blog linked here.

**Careers**
Tenured and tenure track faculty positions
Assistant Professor - Analytics and Information Management
Leeds School of Business, the University of Colorado at Boulder; Boulder, CO

Assistant Professor - Higher Order Cognition
Department of Cognitive Sciences, University of California, Irvine; Irvine, CA

Assistant Professor – Business Intelligence and Predictive Data Science
University of Toronto Mississauga; Mississauga, Ontario, Canada

Assistant/Associate (4 Positions) - Computer Science & Complex Systems
Vermont Complex Systems Center, University of Vermont; Burlington, VT

Assistant Professor - Political Methodology
Department of Political Science, Northwestern University; Evanston, IL

Associate or Full Professor - Communication
University of Washington; Seattle, WA

Professor (Open Rank) – Data Science
Worcester Polytechnic Institute; Worcester, MA

Assistant Professor in Digital Humanities
Penn Arts & Sciences, University of Pennsylvania; Philadelphia, PA

Professor of English - Renaissance Drama
Carnegie Mellon University; Pittsburgh, PA

Full-time, non-tenured academic positions
Research Analyst (accountability)
Data & Society, New York University; New York, NY

Academic Coordinator, Data Science Initiative
University of California-Davis; Davis, CA

Data Visualization Specialist
Georgia Tech Library; Atlanta, GA

Data Acquisition Manager
Institute for the Interdisciplinary Study of Decision Making, New York University; New York, NY

Postdocs
Postdoctoral Scholar - Research Associate in Machine Learning
University of Southern California, Information Sciences Institute; Marina Del Rey, CA
Postdoctoral Fellow: Six Degrees of Francis Bacon
Department of English, Carnegie Mellon University; Pittsburgh, PA

Postdoctoral Fellowship
National Socio-Environmental Synthesis Center (SESYNC); Annapolis, MD
Full-time positions outside academia
Data Scientist, Media Analytics team
Fusion; New York, NY

Developers and data scientists!
U.S. Department of Commerce; Greater Washington DC

Chief Data Officer
City of Santa Monica; Santa Monica, CA

Scientist - Social Dynamics Research
Nokia Bell Labs; Cambridge, England

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 070.