NYU Data Science Community Newsletter features journalism, research papers, events, tools/software, and jobs for August 26, 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 for Social Good: The must-have apprenticeship

I spent the last week hearing about Data Science for Social Good (DSSG) summer fellowships currently wrapping up at the University of Chicago, the University of Washington in Seattle, and Georgia Tech in Atlanta. There are also non-summer programs at Syracuse City Hall in Central New York State, and at IBM, north of New York City in Westchester County (now hiring).

Get one of these fellowships. Seriously.

One of the most compelling ways to dive into data science is to work on projects that impact thousands, maybe even millions of people. Projects that utilize city or regional data to develop data-driven solutions are sociotechnically challenging, real-world intervention points that offer excellent fast-paced apprenticeships with actual data and partners with political capital at stake. Pressure cooker? Yeah, you could say that. Learning how to function in these collaborations gives the fellows a huge advantage. Their peers in tech companies or banks may be working primarily with other data scientists, never having to translate or justify their attentions to stakeholders who don't fully share the social good mission or (more likely) have infrequently gone beyond Excel. (Remember: Excel is bad for science.)

Plan to apply for one of these fellowship programs next summer. The experience is unbeatable and the alumni network alone is worth it.

DSSG network for current city agents and academics

Not a student anymore? That's cool, too.

There are networks of city agencies and academic labs built for sharing insights, software, and workflows from the DSSG projects. At the city level, there's a monthly conference call on which twenty Chief Data Officers and Chief Information Officers compare notes to see what they can share with each other (but this is only for city level CIOs and CDOs). At the academic level, the MetroLab Network based at Carnegie Mellon, has 34 city-university pairs sharing best practices and projects. MetroLab Network is a great first-stop for academics looking to get into DSSG. WhatWorksCities, funded by Bloomberg Philanthropies, does something similar for cities wishing to compare data-driven opportunities for improvement.

Looking for funding right now? Check out **Stanford's** Good Data Grants with the **Bill and Melinda Gates Foundation**.

What about the projects?

What do projects and partnerships look like?

University of Chicago's DSSG projects:

- Smarter crowdsourcing for crisis maps with Ushahidi
- Predictive Enforcement of Pollution and Hazardous Waste Violations in New York State

with the New York State Department of Environmental Conservation
- Optimizing Waste Collection from Portable Sanitation in Kenya with Sanergy

University of Washington's DSSG projects**:**

- CrowdSensing the Mexican Census to Detect Poverty with Early Identification of Unsafe Foods using Amazon reviews
- Dashboard for Exploring Electronic Fare Cards, known to Seattlites as ORCA cards (more)
- OpenSidewalks for Accessibility

Georgia Tech's DSSG projects**:**

- Firebird: Predicting Fire Risk and Prioritizing Fire Inspections in Atlanta with the **Atlanta Fire & Rescue Department**

Employers, hire these fellows

I hope future employers step up to fund these programs as apprenticeships and view these programs as meaningful differentiators of work ethic and character the way they currently see **Teach for America** or serving in **the Peace Corps**. Full disclosure: none of my students are current DSSG fellows.

Organizational challenges remain

DSSG organizers **Rayid Ghani** (U of C), **Lauren Haynes** (U of C), and **Sarah Stone** (UW-Seattle) agree that the apprenticeship model is great for training, but limited when it comes to implementation. In Chicago, "a new report suggests that a data-driven tool meant to reduce gun violence was ignored by police and, in a few cases, may have been misused". As an organizational sociologist, I would have been shocked if city agencies changed their practices following summer fellow presentations with no unintended side effects. Some agencies have taken action rapidly, but only with extremely narrow problems, excellent data and modeling, and ongoing collaboration. Like plants, children, and websites, any good data science project needs care and maintenance to continue to function. The implementation shortcomings point to a gap that needs to be filled and takes nothing away from the value of the DSSG apprenticeship model.

The value of purpose

What's more, most fellows were relieved to be working towards a 'good' deeper than selling ad space online or even fighting credit card fraud. As former Silicon Valley tech entrepreneur turned Chief Data Officer for the White House's Office of Science and Technology Policy **DJ Patil** declared, "having a purpose bigger than yourself, bigger than your company, is a great motivation to get out of bed every morning."

-Laura Norén, from the field

Data Science News

Spreadsheet software defaults damage science

Genome Biology; Mark Ziemann, Yotam Eren and Assam El-Osta from August 23, 2016

As I tweeted earlier this week, Excel is bad for science. Libre Office and most spreadsheet applications are no better. The issue this time? "Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions." [full text]

A new research trend? DARPA wants to fund XAI (Explainable Artificial Intelligence)

McGill University, Newsroom from August 18, 2016
It's nice to see that more transparency, less black-boxing is in store: **DARPA's** technology could be "making life or death decisions". New funding is available for "research in Explainable AI, that is, AI that will output an answer and the criteria that led the system to its decision."

Programmable network routers

MIT News from August 23, 2016
Hardware is an occasionally overlooked area in data science, possibly due to our disciplinary distance from electrical engineering. Programmable routers allow for algorithm updates which are typically "hardwired into the routers' circuitry...[so] if someone develops a better algorithm, network operators have to wait for a new generation of hardware before they can take advantage of it."

CrowdAI Builds Smarter Image Recognition

Y Combinator, The Macro blog from August 19, 2016
"CrowdAI is building smarter image recognition. They are currently working with satellite, drone and self-driving car companies to provide them with scalable image recognition." ... "We sat down with **Devaki Raj**, **Pablo Garcia**, and **Nic Borensztein** to talk about what they're building."

More Image Recognition:

- Full Resolution Image Compression with Recurrent Neural Networks (August 18, arXiv, Computer Science > Computer Vision and Pattern Recognition; **George Toderici** et al.)
- Segmenting and refining images with SharpMask (August 25, Facebook Code, Engineering Blog; **Piotr Dolla**)
- Science AMA Series: I'm **Abe Davis**, last week my research video about "Interactive Dynamic Video" (IDV) hit the front-page of Reddit (August 23, reddit.com/r/science)

Bloomberg Media Names Global Head of Data Science

FishbowlNY, Chris O'Shea from August 24, 2016
"**Bloomberg Media** has named **Michelle Lynn** global head of data science and insights" ... "Lynn comes to Bloomberg from **Dentsu Aegis Network**, where she served as chief insights officer."

More in Data Science in Business:

- **Apple** Acquires Personal Health Data Startup **Gliimpse** (August 22, Fast Company, **Christina Farr** and **Mark Sullivan**)
- An Exclusive Look at How AI and Machine Learning Work at **Apple** (August 24, Medium, Backchannel, **Steven Levy**)
- Digital Feeding Frenzy Erupts: Internet of Things, Analytics Drive M&A Activity To Record Levels (August 18, Forbes, **Joe McKendrick**)
- **WhatsApp** to Share User Data With **Facebook** (August 25, Wall Street Journal, **Deepa Seetharaman** and **Brian R. Fitzgerald**)

Tweet of the Week

Twitter, Bogdan Botezatu from August 25, 2016

Events

@Scale 2016 lineup announced!

San Jose, CA "Engineers from leading Silicon Valley Internet technology companies and more will be discussing their newest solutions for addressing engineering challenges and building for scale." -- Wednesday, August 31 [$$$]

Kaizen Data Conference

San Francisco, CA "Kaizen Data is an applied data science conference focused on data analytics, processing, management, visualization and machine learning." -- Friday-Saturday, September 16-17, at the **Galvanize** SOMA campus. [$$$]

Clinical Trial Transparency and Reproducibility Discussion Panel and Workshop at NYU

New York, NY "Please join us for a free afternoon of clinical research transparency and reproducibility discussion and learning co-hosted by **New York University, Center for Open Science**, and **AllTrials USA** (part of Sense About Science USA)." -- Thursday, September 29, beginning at 1:30 p.m. at NYU Langone Medical Center, Skirball 4th Floor Seminar Room

OpenTrials launch date + Hack Day

Berlin, Germany "OpenTrials will officially launch its beta on Monday 10th October 2016 at the World Health Summit in Berlin. After months of work behind-the-scenes meeting, planning, and developing, we're all really excited about demoing OpenTrials to the world and announcing how to access and use the site!" ... "If that wasn't enough, we also have a confirmed date and location for the OpenTrials Hack Day – it will take place on Saturday 8th October at the German office of **Wikimedia** in Berlin."

PAPIs '16 — PAPIs.io — Where makers of Predictive Applications & APIs meet

Boston "PAPIs '16 will be the 3rd International Conference on Predictive Applications and APIs" -- Monday-Wednesday, October 10-12. [$$]

Geohackweek

Seattle, WA "Geohackweek is a 5-day workshop to be held at the **University of Washington eScience Institute**. Participants will learn about open source technologies used to analyze geospatial datasets. -- Monday-Friday, November 14-18

Deadlines

Take O'Reilly's 2017 Data Science Salary Survey

deadline: Survey

As a data professional, you are invited to share your valuable insights. Help us gain insight into the demographics, work environments, tools, and compensation of practitioners in our growing field. All responses are reported in aggregate to assure your anonymity.

Nominate for the Congressional Innovation Fellowship

deadline: Career Opportunity

Nominating a friend, family member or colleague will show him or her that you think they have what it takes to help bring our government into the 21st Century. Nominees will also gain access to exclusive events and trainings with some of the country's top

technology leaders.

Deadline for nominations is Thursday, September 1. Deadline to apply is Friday, September 30.

Data by the People, for the People: Join the White House Open Data Innovation Summit

deadline: Conference

Washington DC "The **White House** will showcase recent and future open data and My Data achievements at the September 28th Open Data Innovation Summit with Solutions Showcase!"

Deadline to be considered as a speaker or Solutions Showcase exhibitor is Thursday, September 1.

Seeking Public Input on the HHS Open Government Plan for 2016–2018

deadline: Survey

"Every two years, we've worked across all corners of **HHS** to coordinate our strategies for making government more open. Earlier this summer, we called out for your ideas on getting our plan started. Today, we're back in the village square to engage you once again and invite you into our open government plan."

The deadline to contribute comments is Friday, September 9.

EMNLP 2016 - Joint Call for Student Scholarship Applications and Student Volunteers

deadline: Conference

Austin, TX The 2016 Conference on Empirical Methods in Natural Language Processing will be held on Tuesday-Friday, November 1-4. Applicants for either the Student Volunteer Program or the Student Scholarship Program must be full-time students.

Deadline to apply is Tuesday, September 13, 2016

Web, Social Media, and Cellphone Data for Demographic Research

deadline: Conference

Bellevue, WA "There is unfortunately very limited communication between population researchers and data scientists. This workshop is intended to foster communication and exchange between the two communities." -- Workshop precedes Socinfo 2016 on Monday, November 14.

Deadline for submissions is Friday, September 30.

CDS News

Big Data is a junkyard.

Medium Bruno Goncalves from August 23, 2016
Our Moore-Sloan Fellow **Bruno Goncalves** reminds us of the composite nature of individuals when drawn from multi-platform social media data: "Every user shows only a piece of himself by using [particular platforms], but by carefully analyzing large amounts of users one might get a fuller picture of human behavior".

Tools & Resources

*** This week with 100% TensorFlow ***

Overview — opveclib

Hewlett Packard from August 18, 2016
"The Operator Vectorization Library, or OVL, is a python library for defining high performance custom operators for the TensorFlow platform. OVL enables TensorFlow users to easily write, test, and use custom operators in pure python without sacrificing performance. This circumvents the productivity bottleneck of implementing, building, and linking custom C++ and CUDA operators or propagating them through the Eigen code base."

RNNs in Tensorflow, a Practical Guide and Undocumented Features – WildML

Denny Britz, WildML blog from August 21, 2016
"Using an RNN should be as easy as calling a function, right? Unfortunately that's not quite the case. In this post I want to go over some of the best practices for working with RNNs in Tensorflow, especially the functionality that isn't well documented on the official site."

Deep Deterministic Policy Gradients in TensorFlow

Patrick Emami from August 21, 2016
"Deep Reinforcement Learning has recently gained a lot of traction in the machine learning community due to the significant amount of progress that has been made in the past few years. Traditionally, reinforcement learning algorithms were constrained to tiny, discretized grid worlds, which seriously inhibited them from gaining credibility as being viable machine learning tools."

Text summarization with TensorFlow

Google Research Blog, Peter Liu from August 24, 2016
"We're open-sourcing TensorFlow model code for the task of generating news headlines on Annotated English Gigaword, a dataset often used in summarization research. We also specify the hyper-parameters in the documentation that achieve better than published state-of-the-art on the most commonly used metric as of the time of writing."

New TensorFlow Code for Text Summarization

Fast Forward Labs Blog from August 25, 2016
"TensorFlow code works well on relatively short input data ... but struggles to achieve strong results on longer, more complicated text. We faced similar challenges when we built Brief (our summarization prototype) and decided to opt for extractive summaries to provide meaningful results on long-form articles like those in the *New Yorker."*

TensorFlow in a Nutshell? — Part One: Basics

Medium, Camron Godbout from August 22, 2016
"The fast and easy guide to the most popular Deep Learning framework in the world."

Careers

Tenured and tenure track faculty positions

Assistant Professor or Associate Professor, Department of Communication
University of California-Davis; Davis, CA

Assistant Professor (multiple openings), Social Impact of Science, Medicine, and Technology
University of California-San Diego; San Diego, CA

Assistant / Associate Professor of Research Ethics, Kansas Univ. Medical Center
University of Kansas; Kansas City, MO

Assistant or Associate Professor, Dept. of Linguistics
New York University; New York, NY

Assistant Professor; Department of Communication
Cornell University; Ithaca, NY

Assistant Professor, Management and Organizations, Kellogg School of Management
Northwestern University; Evanston, IL

Associate Professor (2 openings) Population Health and Labor Demography
Max Planck Institute for Demographic Research; Rostock, Germany

Assistant Professor (4 openings), Department of Politics
New York University; New York, NY

Full-time, non-tenured academic positions

Scientific Application Developer (1+ positions), Physics
Princeton University; Princeton, NJ

Senior Informatics Researcher - Renaissance Computing Institute
University of North Carolina-Chapel Hill; Chapel Hill, NC

Postdocs

Strategic Data Project Data Fellow, Center for Education Policy Research
Society for Research on Educational Effectiveness; Cambridge, MA

Full-time positions outside academia

Sports Intelligence Programme Manager
UK's High Performance Sport System; London, England

Data Scientist, Marketplace Belonging
Airbnb; San Francisco, CA

Data Analyst or Director
Turnaround for Children, New York, NY

Software Curator for Systems and Environments
Rhizome; New York, NY

Junior Developer
StatDNA; Seattle, WA

Senior Grant Program Specialist, National digital platform portfolio development
Office of Library Services, Washington, DC

OPT OUT: If you do not want to receive this newsletter, please email brad.stenger@nyu.edu with the word 'unsubscribe' in the subject line.
OPT IN: Feel free to forward the Data Science newsletter to colleagues. They can sign up for the newsletter using this web form.