NYU Data Science Community Newsletter features journalism, research papers, events, tools/software, and jobs for June 10, 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

Open Access Policy for #MooreData

Medium, Moore Data, Carly Strasser from June 07, 2016
Open access to research articles has been in the news quite a bit lately (see the SciHub controversy, the preprints in biology discussion, and the European Union’s recent announcement). The Data-Driven Discovery team at the Moore Foundation has also been discussing open access, particularly as it relates to the publications generated by our #MooreData researchers. Our grantee population is fairly progressive when it comes to open science, and many of the outputs that they generate are already publicly available (including proposals, software, workflows, and publications). It is therefore easy for us to imagine that they would embrace a policy that mandates open access for research articles that they produce. That said, we are always open to discussions!

Also, in open access and science publishing:

- Taking up TOP (June 03, Science, Editorial; Marcia McNutt)
- Now, a better way to find and reward open access (June 05, Impactstory blog)
- R Passes SAS in Scholarly Use (finally) (June 08, r4stats.com)
- Bias against Novelty in Science: A Cautionary Tale (June 01, National Bureau of Economic Research)
- Predicting the Impact of Scientific Concepts Using Full Text Features (June 06, Allen Institute, Semantic Scholar; Kathleen Mckeown et al.)

Former NASA chief unveils $100 million neural chip maker KnuEdge

VentureBeat, Dean Takahashi from June 06, 2016
It’s not all that easy to call KnuEdge a startup. Created a decade ago by **Daniel Goldin**, the former head of the National Aeronautics and Space Administration, KnuEdge is only now coming out of stealth mode. It has already raised $100 million in funding to build a “neural chip” that Goldin says will make data centers more efficient in a hyperscale age.

Researchers Uncover a Flaw in Europe’s Tough Privacy Rules

The New York Times from June 03, 2016
Europe likes to think it leads the world in protecting people’s privacy, and that is particularly true for the region’s so-called right to be forgotten. That legal right allows people connected to the Continent to ask search engines like **Google** to remove links about themselves from online search results under certain conditions.

Yet that right — one of the world’s most widespread efforts to protect people’s privacy online — may not be as effective as many European policy makers think, according to new research by computer scientists based, in part, at **New York University**.

DH at Berkeley Awards More than $200K in Grants

University of California-Berkeley, Digital Humanities from June 07, 2016
The grants will promote collaborative research and the development of new DH courses. 14 research teams will be funded for projects that range from database development to algorithmic analytical tools.

[1606.00776] Multiresolution Recurrent Neural Networks: An Application to Dialogue Response Generation

arXiv, Computer Science > Computation and Language; Iulian Vlad Serban, Tim Klinger, Gerald Tesauro, Kartik Talamadupula, Bowen Zhou, Yoshua Bengio, Aaron Courville from June 02, 2016
We introduce the multiresolution recurrent neural network, which extends the sequence-to-sequence framework to model natural language generation as two parallel discrete stochastic processes: a sequence of high-level coarse tokens, and a sequence of natural language tokens. There are many ways to estimate or learn the high-level coarse tokens, but we argue that a simple extraction procedure is sufficient to capture a wealth of high-level discourse semantics.

Also, in text analysis:

- Introducing DeepText: Facebook's text understanding engine (June 01, Facebook Code, Engineering Blog)
- [1606.00372] Conversational Contextual Cues: The Case of Personalization and History for Response Ranking (June 01, arXiv, Computer Science > Computation and Language; Rami Al-Rfou, Marc Pickett, Javier Snaider, Yun-hsuan Sung, Brian Strope, Ray Kurzweil)
- Predicting the Impact of Scientific Concepts Using Full Text Features (June 06, Allen Institute, Semantic Scholar; Kathleen Mckeown et al.)

Computer Vision Research: The deep "depression"

LinkedIn, Nikos Paragios from June 05, 2016
... almost all the community now seems to target the development of more complex pipelines (that most likely cannot be reproduced based on the elements presented in the paper) which in most of the cases have almost no theoretical reasoning behind that can add 0,1% of performance on a given benchmark. Is this the objective of academic research? Putting in place highly complex engineering models that simply explore computing power and massive annotated data?

Opinion: Big data biomedicine offers big higher education opportunities

Proceedings of the National Academy of Sciences; John Darrell Van Horn from June 07, 2016
I like to tell my students a story about a time back in the “olden days” of the early 1990s when I was a newly minted postdoctoral fellow at the **National Institutes of Health**. Our laboratory conducted brain imaging studies using positron emission tomographic imaging and would routinely obtain 6–12 functional image volumes per subject in our experiments. The director of our neuroimaging laboratory asked me to

explore the purchase of a new computer hard drive capable of storing our growing collection of brain imaging data files. At that time, we had already accumulated quite a number of subjects and wanted to easily access their data for a variety of different statistical analyses—and we expected to obtain much more data.

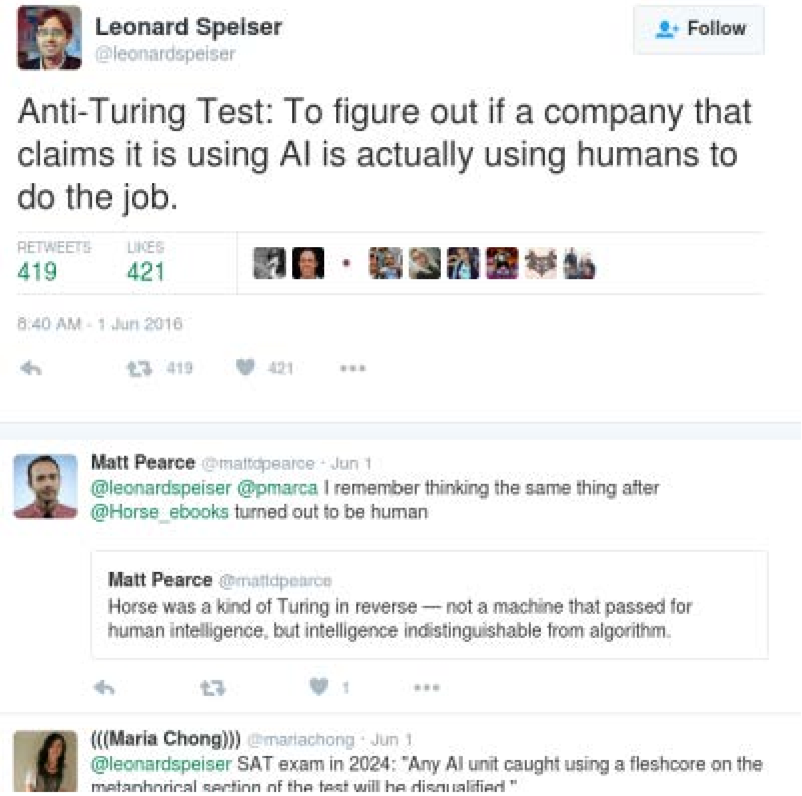
So I searched, and I compared and contrasted, rating drive quality, price, and capacity. With our not-so-expansive $4,000 budget, I eventually decided on the drive that would surely satisfy our needs. It was among the largest self-contained external hard drives of its kind at that time. Its massive 4 gigabyte capacity seemed infinite. People would come to visit the laboratory to just gaze upon it.

Reflection on the Data Science Profession

YouTube, Work-Bench from June 03, 2016
Delivered by **Drew Conway** (CEO, Alluvium) at the 2016 New York R Conference on April 8th and 9th at Work-Bench. Conway reflects on how New York City became a hub for data science and innovation. [video autoplays, 19:32]

Tweet of the Week

Twitter, Leonard Speiser from June 09, 2016



Leonard Speiser
@leonardspeiser

Anti-Turing Test: To figure out if a company that claims it is using AI is actually using humans to do the job.

RETWEETS 419 LIKES 421

8:40 AM - 1 Jun 2016

Matt Pearce @mattdpearce · Jun 1
@leonardspeiser @pmarca I remember thinking the same thing after @Horse_ebooks turned out to be human

Matt Pearce @mattdpearce
Horse was a kind of Turing in reverse — not a machine that passed for human intelligence, but intelligence indistinguishable from algorithm.

(((Maria Chong))) @mariachong · Jun 1
@leonardspeiser SAT exam in 2024: "Any AI unit caught using a fleshcore on the metaphorical section of the test will be disqualified."

Events

Data Science Summit 2016

Industry leaders predict that most applications will be powered by machine learning within the next 3-5 years. To keep pace with this changing landscape, business and academic leaders are developing new tools and techniques that maximize the ability to create and leverage powerful machine learning. We have brought together these innovators for the 5th annual Data Science Summit in San Francisco.

Moore-Sloan Data Science Environment participants: **Andreas Mueller** (NYU), **Magdalena Balazinska** (UW-Seattle), **Emily Fox** (UW-Seattle), **Bill Howe** (UW-Seattle)

San Francisco, CA Tuesday-Wednesday, July 12-13 [$$$]

The Rise of Machine Learning

Welcome to the Silicon Valley Robotics Influencer series, hosted at **HAX Accelerator** and curated by **Cory Kidd** of Catalia Health, **Tim Smith** of Element Public Relations and **Andra Keay** of Silicon Valley Robotics. This topic looks at the state of the machine learning with special guests: **Joshua Bloom** (Wise.io), **Sarah Osentoski** (Mayfield Robotics), **Carol Reiley** (Drive.ai), **Quentin Hardy** (New York Times - Moderator).

San Francisco, CA Tuesday, June 14, starting at 6 p.m., HAX Accelerator (479 Jessie Street) [$$]

Software Carpentry: University of Washington - Seattle

Software Carpentry's mission is to help scientists and engineers get more research done in less time and with less pain by teaching them basic lab skills for scientific computing. This hands-on workshop will cover basic concepts and tools, including program design, version control, data management, and task automation.

Seattle, WA Tuesday-Wednesday, June 14-15, at WRF Data Science Studio, Physics/Astronomy Tower (6th Floor). [$]

CUSP Research Seminar Series | June 15

Join **NYU Center for Urban Science and Progress** for a research seminar with **Jeff Jonas**, **IBM** Fellow and Chief Scientist of Context Computing.

Brooklyn, NY on Wednesday, June 15, at 11 a.m., Jacobs Seminar Room at the Center for Urban Science and Progress (1 Metrotech, 19th Floor)

WiML ICML Luncheon 2016

Underrepresented minorities and undergraduates interested in machine learning research are Encouraged to expect. The luncheon is co-located with ICML.

New York, NY Saturday, June 21, at 11 Times Square, Microsoft building (6th floor) startiing at 12 noon

HILDA 2016: Workshop on Human-In-the-Loop Data Analytics

Any data management system needs to work together with people, whose needs determine the goals for the system, and who must provide the input and who need to work effectively with the output. Data management systems will work much better when they take account of the cognitive and physiological characteristics of the people involved. Recent technology trends (such as touch screens, motion detection, and voice recognition) are widening the possibilities for users to interact with systems, and many information-provision industries are shifting to personalized processing to better target their services to the users' wishes. HILDA is a new workshop that will allow researchers and practitioners to exchange ideas and results relating to how data management can be done with awareness of the people who form part of the processes. ... HILDA intends to be a forum where people from varied communities engage with one another's ideas. We are keen to have submissions that present initial ideas and visions, just as much as reports on early results, or reflections on completed projects.

San Francisco, CA Sunday, June 26, 2016, Co-located with SIGMOD 2016 [$$$]

SafArtInt 2016 - WORKSHOP ON SAFETY AND CONTROL FOR ARTIFICIAL INTELLIGENCE

The Public Workshop on Safety and Control for Artificial Intelligence (SAF|ART|INT) is a jointly-sponsored event of the **White House Office of Science and Technology Policy** (OSTP) and **Carnegie Mellon University**. The workshop will explore the potential future of AI and AI applications, the emerging technical means for constructing safe and secure systems, how safety might be assured, and how we can make progress on the challenges of safety and control for AI.

Pittsburgh, PA Tuesday, June 28, at Carnegie Mellon University

Deadlines

Workshop on Algorithms for Modern Massive Data Sets.

Registration fees are waived for students (non postdoc) with an approved poster presentation.

Berkeley, CA Tuesday-Friday, June 21-24 in Stanley Hall.

Deadline for submissions is Sunday, June 12.

Stanford Medicine X and Symplur announce an Everyone Included™ social media research challenge

Stanford Medicine X and **Symplur** are pleased to announce a joint initiative designed to spark scholarly research activity in healthcare social media. The Stanford Medicine X | Symplur Everyone Included™ Research Challenge seeks to encourage all health care stakeholders to collaborate on health care social media research.

Deadline for submissions is Friday, June 17.

Computational Social Science Workshop

The aim of this satellite is to address the question of ICT-mediated social phenomena emerging over multiple scales, ranging from the interactions of individuals to the emergence of self-organized global movements. We would like to gather researchers from different disciplines and methodological backgrounds to form a forum to discuss ideas, research questions, recent results, and future challenges in this emerging area of research and public interest.

Amsterdam, The Netherlands Wednesday, September 21, co-located with 2016 Conference on Complex Systems.

Deadline for abstract submissions is Monday, June 20.

Call For Workshops – SocInfo'16

The SocInfo 2016 Committee invites proposals for Workshops Day at the Eighth International Conference on Social Informatics (SocInfo 2016).

Seattle, WA The Workshops Day will be held on Monday, 14 November.

Deadline for workshop submissions is Friday, July 1.

Mozilla Fellowships for Science

We're looking for researchers with a passion for open source and data sharing, already working to shift research practice to be more collaborative, iterative and open. Fellows will spend 10 months starting September 2016 as community catalysts at their institutions, mentoring the next generation of open data practitioners and researchers and building lasting change in the global open science community.

Deadline for applications is Saturday, July 16.

MacArthur Foundation is offering $100 million to a group that identifies a social problem and can solve it

The **MacArthur Foundation**, which has doled out billions of dollars in “genius grants” and to nonprofit organizations working on major social challenges, now wants to give away $100 million to solve a societal problem that might not be on its radar.

Deadline for submissions is Monday, October 3.

CDS News

2016 NYU Data Science Seed Grant Awards

NYU Center for Data Science from June 09, 2016

The Seed Grant Selection Committee is thrilled to announce that, after a rigorous evaluation process involving multiple referee reports, the following grant proposals were chosen for funding by the **Moore-Sloan Data Science Environment**.

- **Brian Parker** and **Christine Vogel**: “Statistics Meets Transcriptomics: Time-Series

Responses of Post-Transcriptional Regulation By Families of Conserved RNA Structures”

- **Ralph Grishman** and **Alastair Smith**: “Health and Death of Political Leaders”
- **Jonathan Winawer** and **Heiko Müller**: “The Standard Cortical Observer”
- **Florian Knoll** and **Carlos Fernandez-Granda**: “Estimation of Multiple Tissue Compartments from Magnetic-Resonance-Fingerprinting Data”
- **Preeti Raghavan** and **Aaditya Rangan**: “Determining Treatment Algorithms for Patient Subgroups in Stroke Rehabilitation”
- **Thomas Kirchner** and **Kyunghyun Cho**: “Image-based Community Asset and Risk Factor Surveillance System using Deep Learning”
- **Nathaniel Beck** and **David Sontag**: “Applying Machine Learning Methods to Integrated Time Series”

Tools & Resources

Make your own tagging system from scratch

Kequc, Nathan Lunde-Berry from June 03, 2016
Build a tagging tool from scratch rather than using one that is pre-made. You get more control, there isn't unused code, and you learn something too. In this article I will go through the process of building a tagging system from scratch.

idbr: access the US Census Bureau International Data Base in R

AriLamstein.com, Kyle Walker from June 06, 2016
The **US Census Bureau’s** International Data Base (IDB) is one of the best resources on the web for obtaining both historical and future projections of international demographic indicators. I’ve long used the IDB in my teaching, generally using its web interface to download data extracts. However, the Census Bureau also makes the IDB accessible via its API, which makes it much more convenient for programmers to access the data. Earlier this year, I wrote the R package idbr (https://github.com/walkerke/idbr) to help R programmers use the IDB in their projects.

ParaText: CSV parsing at 2.5 GB per second

Wise.io from June 07, 2016
Despite extensive use of distributed databases and filesystems in data-driven workflows, there remains a persistent need to rapidly read text files on single machines. Surprisingly, most modern text file readers fail to take advantage of multi-core architectures, leaving much of the I/O bandwidth unused on high performance storage systems. Introduced here, ParaText, reads text files in parallel on a single multi-core machine to consume more of that bandwidth. The alpha release includes a parallel Comma Separated Values (CSV) reader with Python bindings. ... In our tests, ParaText can load a CSV file from a cold disk at a rate of 2.5 GB/second and 4.2 GB/second out-of-core from a warm disk. ParaText can parse and perform out-of-core computations on a 5 TB CSV file in under 30 minutes.

Release TensorFlow v0.9.0 RC0 · tensorflow/tensorflow · GitHub

GitHub - tensorflow from June 06, 2016

Major Features and Improvements

- Python 3.5 support and binaries
- Added iOS support
- Added support for processing on GPUs on MacOS
- And more

All-in-one Docker image for Deep Learning

GitHub - saiprashanths from June 07, 2016
Here are Dockerfiles to get you up and running with a fully functional deep learning machine. It contains all the popular deep learning frameworks with CPU and GPU support (CUDA and cuDNN included). The CPU version should work on Linux, Windows and OS X. The GPU version will, however, only work on Linux machines.

10 Useful Python Data Visualization Libraries for Any Discipline

Mode Blog, Melissa Bierly from June 08, 2016
Today, we're giving an overview of 10 interdisciplinary Python data visualization libraries, from the well-known to the obscure. We've noted the ones you can take for a spin without the hassle of running Python locally, using Mode Python Notebooks.

Open Sourcing Photon ML

LinkedIn Engineering, Paul Oglivie from June 07, 2016
Machine learning is a key component of **LinkedIn's** relevance-driven products. We use machine learning to train the ranking algorithms for our feed, advertising, recommender systems (such as People You May Know), email optimization, search engines, and more. For an in-depth example, check out these posts (part one and two) on how LinkedIn applies machine learning for ranking the feed.

These algorithms play an important role in determining user experience for content-rich websites, so it's critical that we provide our engineers with easy-to-use machine learning tools that create high-quality models that are fast and scale to large datasets. To meet these needs, we have developed Photon ML, a machine learning library for Apache Spark.

Careers

Urban Institute - Junior Data Visualization Developer

Urban Institute from June 06, 2016
Washington, DC The Institute is seeking a junior data visualization developer to join our vibrant and growing Visual Communications team. To build on our continued success with powerful digital content, we're looking for a talented and multifaceted visual thinker. This is a unique opportunity to deliver Urban's vital research to an expanding audience and to explore the possibilities of telling stories with data. The developer will work closely with research center directors and staff, executive office staff, and the digital communications team to translate research needs into user-focused data visualizations, interactive tools, and applications for Urban Institute websites.

PhD Studentship, OpenSystems

University of Barcelona from June 06, 2016
Barcelona, Spain We offer an opportunity to join OpenSystems group and undertake research in Human Behavior using participatory practices and running physically embedded collective experiments. Research will specifically focus on human mobility (GPS tracking) and/or human decision making (digital interfaces monitoring human actions in for instance social dilemmas) with experiments situated in real-world but controlled scenarios.

Postdoctoral Research Associate - Politics

Princeton University from June 09, 2016
Princeton, NJ The Departments of Politics and Sociology at Princeton University seek applicants for two positions at the rank of postdoctoral research associate, more senior research associate, or associate professional specialist. The primary responsibilities of the positions are to design and teach a series of computing workshops for data analysis in the social sciences.

Image assay development postdoc job description

Broad Institute of Harvard and MIT from June 09, 2016
Cambridge, MA This is an opportunity to make important contributions through many different individual projects as well as by supporting thousands of researchers around the world who are accomplishing great things with our lab’s open-source software, CellProfiler.

The Art of Pivoting

Medium, Boris Adryan from June 04, 2016
The Art of Pivoting, or less pretentious, how I changed from being a frustrated life science academic to using my skills as well-paid consultant for industrial engineering problems.