Calendar News Contact Us

eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

Search ...

Community-level data science and its spheres of influence: beyond novelty squared

Share

Brittany Fiore-Gartland and Anissa Tanweer

Data science has many characterizations, but in academia it is often talked about as pushing the limits of both methodological and domain science, what Josh Bloom, a Professor of Astronomy at U.C. Berkeley, has referred to as “**novelty squared**”. Bloom sees this as the “great challenge of modern interdisciplinary scientific collaboration”. The idealized characterization of data science in academia is also represented in the idea of shifting from the traditional T-shaped scientists, who have deep expertise in a single domain, to Π (*Pi*)-shaped scientists with deep expertise in both a domain and methodological science (as coined by **Alex Szalay** and discussed **here** and **here**. As Π-shaped data scientists, they are primed to innovate in multiple disciplinary trajectories. Bloom and others have argued that these characterizations of novelty squared and Π-shaped scientists represent the “**unicorn**” of data science.

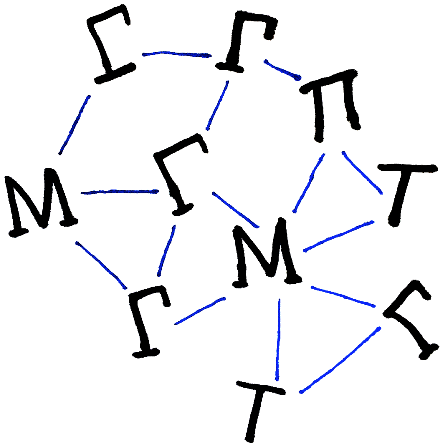
The mythological, elusive nature of the ideal data scientist was supported through our ethnographic fieldwork. In search of these so-called “data scientists” across the data science environment, we could find scarcely anyone who consistently identified as a “data scientist.” We found more often that “data scientist” was a partial and relational identity, one layer of many. We encountered lots of data science, emerging across a network of interactions across a multitude of differently shaped scientists, from T to Γ (*Gamma*) to Π to even M (*Mu*) -shaped! Individuals mostly did not see themselves as particularly Π-shaped, a category many reserved for only a very select few that had accomplished an exceptional level of expertise and contribution to multiple fields. As one of us has

Search ...

Recent Posts

- Panama Papers Leak
- Community-level data science and its spheres of influence: beyond novelty squared
- Mayor Murray Signs Historic Open Data Executive Order
- A New Innovation Model for the 21st Century
- Call for

discussed **elsewhere**, perhaps a more realistic characterization of data science includes a network of differently shaped scientists practicing data science on a community level.



T, Π, Γ, M network. A network view of community-level data science reveals a mix of T-shaped, Γ-shaped, M-shaped, and Π-shaped scientists practicing data science.

Home About Us Research Education Get Involved

Incubator (DSI) program at University of Washington's eScience Institute so that we could observe a cross-section of data science collaborations from across the campus. DSI is a program that aims to bring together data scientists and domain scientists to work on focused, intensive, collaborative projects over the course of an academic quarter. The program solicits proposals from scientists across campus for projects that require data science expertise. The accepted applicants are then paired with one or two data science liaisons at the eScience Institute.

As ethnographers we were participant observers in the DSI and found that it represents a promising model for cultivating what we term community-level data science. Instead of breeding unicorns that embodied the singular, ideal "data scientist", we found that the DSI fosters data science practice on the community level such that work is distributed across domain and methods-based approaches and occurs at the intersections of these different

Proposals – 2016 Data Science for Social Good

Archives

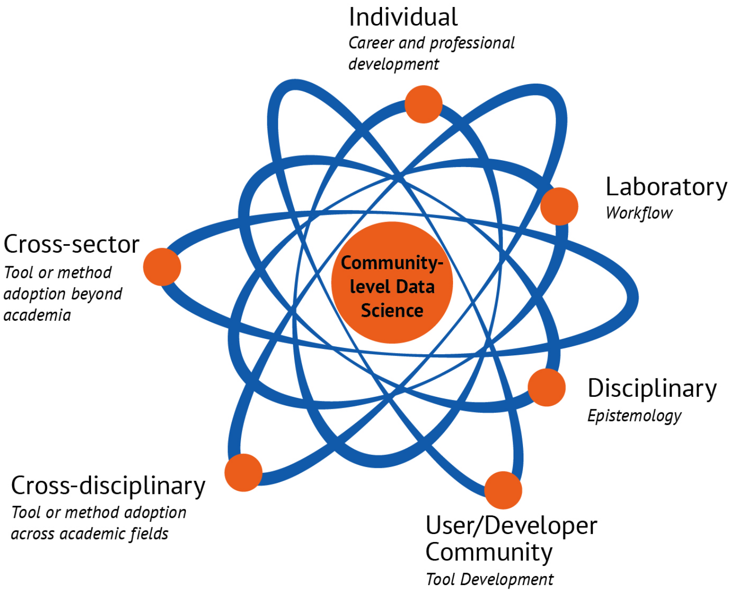
- April 2016
- March 2016
- February 2016
- January 2016
- December 2015
- November 2015
- October 2015
- September 2015
- August 2015
- June 2015
- May 2015
- April 2015
- March 2015
- February 2015
- October 2014

forms of expertise. Beyond acknowledging the DSI's role in advancing their particular research projects, participants articulated an awareness that they were engaged the process of learning what it means to be part of a data science community. In other words, participants recognized data science as a cultural practice.

As such, the novelty-squared model falls short of sufficiently characterizing the DSI interactions. Our respondents didn't talk about the value of data science solely, or even first and foremost, in terms of simultaneously pushing the known limits of domain knowledge and computational methods. Rather, they imagined the "novel" in data science as its ability to influence a range of day-to-day work practices and to expand an array of scientific and professional possibilities.

This program provides a window into a range of data science collaborations that have significant value without all being characterized by a novelty squared approach. How do we make sense of this range of interactions and understand the different motivations and values that are imagined by participants? Based on our ethnographic observations of this program and interviews with all participants we categorized the different kinds of value we heard participants imagine and ascribe to data science into six spheres of influence.

Spheres of Influence


Individual
Career and professional development
Laboratory
Workflow
Cross-sector
Tool or method adoption beyond academia
Community-level Data Science
Disciplinary
Epistemology
Cross-disciplinary
Tool or method adoption across academic fields
User/Developer Community
Tool Development

License

Individual

Some participants told us that data science helped distinguish them from other scholars in their fields by equipping them with uncommon skills or enabling them to generate novel data sets. For example, one participant told us:

> *"[When] you are on the job market you present kind of one paper that you are going to be a solo author on and it's becoming less and less common to use these kind of prepackaged data sets.* ***It's starting to be that you need to bring new data to the table, to answer any questions, to get a good job.****"*

Data science in this case is imagined to be influencing the sphere of the individual by providing researchers with skills, tools, and data that can help them advance their career and professional goals.

Laboratory

Other participants have come to the data science incubator hoping to gain new skills and learn new tools or systems they can take back to their labs or collaborative research groups in order to optimize their workflow. For one participant, it was about using the tools and techniques learned in the incubator for automating an existing workflow that involved processing files manually. For another participant it was about bringing a more seamless and transparent workflow to their lab using Git. In the words of this participant:

> ***What I'd like to do is build a better infrastructure.*** *Programming is on Git, so I can look at [lab member's] programs. I can download them. I can check them. [....] "Do I trust this code? Do I not trust this code? How much is this a bunch of hacks? How much is this tidy code?" [....] I knew I wanted to start using Git. I did not use Git, so then moving [the] lab over and doing it, [when you] don't even know how to use it yourself, it's just never ever going to work."*

What was important about this person's data science incubator experience was the acquisition of new skills and tools that would allow her to streamline and optimize work routines across a team of faculty and student researchers. For a number of participants, it is not necessarily about scientific advancement in the short term; rather it is about investing in scientific advancement and productivity in the long term. This perspective foregrounds the potential for community-level data science to influence the sphere of the laboratory.

Disciplinary

Sometimes, participants saw data science as providing novel ways of developing theory within their discipline. One participant told us:

> *"[Traditional theorists in my field] say 'This is my regression model. It's these x's that I guessed, basically, are important…Then I'm going to use that model to predict y.' There's never a process of kicking out any regressors that are potentially unimportant. We consider that to be a very old-timey way of doing it.* ***That's the way we've been doing it since before we had these machine learning techniques that could tell us that stuff.****"*

This person sees data science as providing a new way of knowing, or introducing epistemological novelty to his particular domain. So in this case, we can consider data science to be influencing the disciplinary sphere.

User/Developer Community

Another area of novelty people sometimes identified was their contribution to the development of data science tools. One participant described her experience this way:

> *"… I was one of the only people doing these calculations and doing a lot of numerical stuff through [the newly developed cloud-based data management system]. I think a lot of the other queries were looking at text. It was kind of a different focus.* ***I was finding some really good bugs for them.****"*

Because the contributions she made could potentially affect anyone who is using or developing the new tool, in this case we've characterized community-level data science as influencing the sphere of the user/developer community.

Cross-disciplinary

Sometimes researchers were developing novel approaches to examining an object of study that is common across a number of fields. This was the case with one participant who was developing tools and methods to analyze the entire corpus of text on the Internet Archive:

> *"At this conference we went to, there was a group of people that were trying to do link analysis [on the Internet Archive] …. I've been looking at the text. I told them … I'd send them all of the stuff that I've done, and they were really excited about that because they want to start looking at the text, but* ***they hadn't considered doing that yet because it's 90 terabytes of text. How do you look at that?!****"*

In cases like this, the methods and tools being developed could be applied to similar data generated and analyzed across different academic domains. Therefore, we can think of community-level data science as potentially influencing the cross-disciplinary sphere.

Cross-sector

At times, participants in our study talked about how novel data science tools and methods they were developing could be applied to questions beyond academia. This idea was expressed by a participant who thought his algorithms for analyzing time series might have commercial applications with other types of data:

> *"I would love to, for example, start a company up, that tries to model all these time domain streams coming from wearables or the internet of things. Where these things are producing these data streams and people aren't necessarily listening. But if they were –* ***if you had the right software listening to these streams – you could actually understand what's happening.****"*

Because this person envisions taking software that was developed for academic research and adapting it for use in a commercial setting, we can say that the imagined influence of data science is taking place in a cross-sector sphere that spans academic and commercial sectors, as well as public and private sectors.

Toward a community-level data science

As these examples show, we found that the wide ranging imaginations for the value of data science were often not at the level of “novelty squared.” To be sure, many of the DSI projects would have been intractable without the incubator; posing research questions the participants couldn’t or wouldn’t otherwise ask. In some cases this novel science required a specific input and contributions from the data science liaisons and in other cases the novel science was supported through the dedicated time for elbow-to-elbow learning of tools and techniques from the data science canon. But the incubator participants often characterized what were major breakthroughs for them as trivial tasks for their data science liaisons. In other words, we didn’t see much of the idealized imagination of a data science that simultaneously pushes the frontiers of domain knowledge and computational methods.

When the projects did seem to be pursuing novelty squared, some participants were ambivalent about that dual trajectory and identified a tension between the mandate to develop novel computational tools and techniques on the one hand, and the mandate of doing what was best for their science on the other hand. But that’s a topic for another installment.

Nonetheless, as we’ve shown, for these participants, the data science work accomplished in the DSI was imagined to be novel and transformational within a range of spheres. Many of the researchers we observed saw value in becoming acculturated to the data science community. They talked about how learning the languages, tools, processes, and norms of data science would advance their careers, allow them to keep pace with the trends in their fields, optimize their work routines, alter the repertoires of their disciplines and subdisciplines, and connect them to the array of expertise and resources their research required. In addition to incubating science projects, the DSI incubates researchers within a data science community in which they are immersed in the cultural language of data science, learning how to think like a data scientist and how to be part of a data science community. As such, if we want to understand the evolution of data science in academia, we need to recognize that it is not solely about developing the capacity of individual researchers to employ novel methods, but also about forging a community-level data science that can reshape the cultural contours of the academy.

Recent Posts

Panama Papers Leak

April 6th, 2016

Mayor Murray Signs Historic Open Data Executive Order

March 4th, 2016

A New Innovat Model for the 2 Century

March 4th, 2016

W UNIVERSITY of WASHINGTON