Tufts UNIVERSITY

THE STATE OF DATA SCIENCE AT TUFTS UNIVERSITY

A Community Landscape Analysis

By Joe Hilleary

DISC

D3M@Tufts

Executive Summary

This report summarizes the findings of a series of nearly forty semi-structured interviews with faculty and staff working in data science at Tufts University. Over the course of these interviews, several patterns emerged:

Regarding the current state of research and teaching, ***most data science work can be divided into one of four categories—domain-specific applications, foundational methods, critical data studies, or data infrastructure.***

The first category is by far the most well represented, with faculty from essentially every engineering, natural science, and social science discipline utilizing statistical or machine learning methods within their field. Biomedical applications are especially robust. The second category, which concentrates on foundational advances in data science methods, primarily consists of math and computer science faculty and has more of a focus on machine learning than other techniques. A common complaint was the lack of dedicated statisticians on the faculty and high-level statistical course offerings for students. The other two groups are significantly less well established. With few exceptions, most of the work considering the human processes and impacts around the use of data-driven techniques (critical data studies) takes place in limited modules taught within larger courses. It's also the most likely to be undertaken by those with less established positions such as staff, non-tenured faculty, graduate students, or post-docs. The final category, while relevant and distinct, is represented by only a handful of staff and engineering faculty, who operate the Tufts high performance cluster and think about designing systems to support data science at scale, respectively.

While there is widespread enthusiasm for collaboration and continued development in all of these areas at Tufts, several structural challenges stymie this work. Firstly, ***knowledge of potential collaborators' work and the resources available across campuses is extremely poor.*** *Current modes of communication have been ineffective at surmounting this barrier. These disconnects result in faculty missing out on many of the precise opportunities that they suggest would improve the Tufts data science community, such as symposiums, informal gatherings, and financial support for new initiatives. At the same time, some of these existing resources such as* ***funding and support staff would be insufficient if they were to experience the true level of demand present at the university.***

In addition to improving communication about existing resources, ***there is a need to institutionalize data science work outside of any current department.*** *This would circumvent the present barriers to collaboration and interdisciplinary education created by politics around funding and teaching loads, as well as providing a common home for faculty regardless of their school or disciplinary affiliation. As a visible center for this work, it would provide a single point of contact and/or entry for faculty, staff, and students interested in any aspect of data science. It would also provide the infrastructure to support a shared common core of data science courses. If implemented correctly, this centralized curriculum would allow students in different home disciplines better access to the courses, which presently are scattered across schools and departments, while more equitably meeting departmental needs from a funding and teaching load perspective.*

I. Introduction

The goal of this work was to map the data science community at Tufts across all schools, divisions, and departments and to generate ideas to improve research and teaching in this area. Over the course of the spring and summer of 2023, I conducted interviews with nearly 40 faculty and staff. These interviews broadly focused on three areas:

1. What is the current state of teaching and research in data science at Tufts?
2. To what extent is the data science community interconnected and visible to its members?
3. Where are the gaps in expertise that ought to be filled moving forward?

While the interviews were largely conversational, allowing for organic themes to emerge, each participant was asked the same set of 8 questions (See Appendix A).

i. Defining Data Science

Before delving deeper, it is important to define what is meant by "data science" in the context of this report. The term data science is simultaneously nebulous and pervasive. There is no widely agreed upon definition and many areas that once fell under different banners now use the label of data science. For the purposes of this work, I have deliberately taken a big tent approach, including under data science any teaching or research concerned with the use of data.

ii. Process for Determining Interviewees

To this end, the initial list of interviewees was drawn from the SIS course directory using a keyword search on "data." From that starting point, the list grew through the network effect of asking each interviewee to recommend others to interview. In its current state, this "data science directory" contains more than 200 individuals who are either directly involved in data science work or who've been identified as having valuable perspectives on data science at Tufts.

From this master list, an effort was made to prioritize those most directly involved in data science work, based on the pre-existing knowledge of the D3M and DISC leadership, while ensuring representation of the university as a whole and a wide array of perspectives. Approximately half of the directory received outreach over the course of the interview process. Each potential interviewee was contacted up to three times. If they failed to respond to these emails, no further attempts were made to interview them. Ultimately, interviewees came from more than 25 departments across every school except for Dental Medicine. (See Appendix B for additional figures).

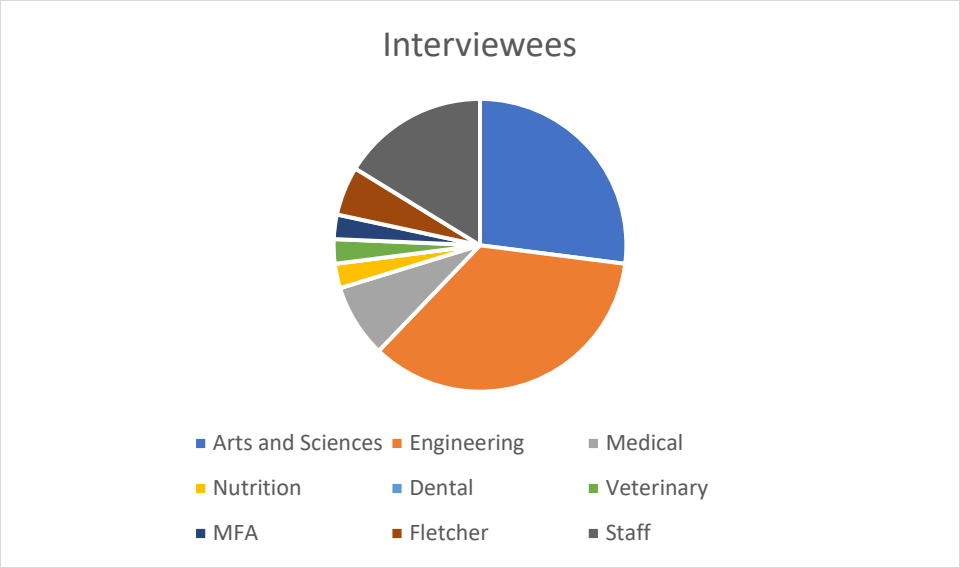

Figure 1. Breakdown of interviewees across schools.

While this work covered a broad cross-section of the data science community, it is by no means exhaustive. I was unable to interview several individuals identified as key figures in the community, and there are certainly areas where work is happening that were not turned up by the process I conducted.

II. Current Research and Teaching

i. Categories of Research and Teaching

The data science community at Tufts is quite sizable. Nevertheless, **nearly all of the teaching and research conducted by the community falls into one of four categories: Domain Specific Applications, Foundational Methods, Critical Data Studies, or Data Infrastructure.**

1. Domain Specific Applications

Most of the data science community at Tufts is focused on the application of data science methods within a particular domain. The domain with the largest representation is biomedical research, which spans not only the Medical School, but also the Engineering School, the School of Arts and Sciences, and the Nutrition School. Within that area, much of the work is on "omics" data. Most of the engineering departments also have faculty working on domain specific data science applications, but the department of Civil and Environmental Engineering appears to have an especially strong focus in this area as a result of faculty research interest. The Gordon Institute, which technically falls under the School of Engineering, also has a few courses on business applications. In the School of Arts and Sciences, social science departments, in particular Economics and Psychology, have faculty focused on data science applications, as do the natural science departments, in particular Physics and Biology. The Department of Urban and Environmental Planning serves as a key center for applied geospatial analysis, while Classics plays a similar role for natural language processing and the digital humanities. The Nutrition School has two core areas of data science applications—molecular nutrition and nutrition epidemiology. The Dental School and Veterinary School are largely focused on professional degree programs, but each has a few faculty who work on data science applications, the latter mostly within epidemiology. Fletcher has a handful of professors that apply data science in their work, particularly geospatial and econometric

methods. Finally, the School of the Museum of Fine Arts (SMFA) has a Digital Media department that may dabble in applied data science and certainly has the opportunity to.

2. Foundational Methods

A subset of the data science community, mostly in the Math and Computer Science departments focuses on the development and teaching of foundational methods. Current research primarily skews towards machine learning, although a few folks across Math, Civil and Environmental Engineering, the Medical School, and the School of Nutrition have expertise in probability and statistics. In many cases, other departments rely on the faculty in this group to provide foundational courses for their students, and many lament the lack of more theoretical statisticians who could teach and research advanced statistical techniques, including time-series analysis, analysis of longitudinal study data, big data analysis, high dimensional data analysis, hazard functions, small-n data analysis, and causal inference.

3. Critical Data Studies

One of the smallest and least formalized groups within the data science community looks at the human processes and impacts of data-driven techniques. Essentially, it is the study of how data science is conducted and its socio-cultural and ethical ramifications. The organization with the most explicit focus on this topic is the undergraduate STS program in the School of Arts and Sciences, but elements of this area are also covered in the Cyber Security and Public Policy program that bridges Fletcher with the Computer Science Department. Beyond a single course on AI ethics in Computer Science, a handful of mostly one-off courses have been taught, typically by impassioned graduate students, post-docs, and junior or non-tenured faculty. Otherwise, it primarily exists at a modular level within much larger courses with applied or technical focuses. Only one faculty member, in the anthropology department, appeared to have this area as the primary focus of their work. In spite of its marginal current presence, a huge proportion of interviewees felt this was actually one of the most important areas for students to develop expertise, with the insight that methods and applications would evolve over time. It was also one of the most common areas highlighted as a gap in faculty expertise.

4. Data Infrastructure

The final group consists of individuals focused on the logistics and systems requirements of working with data at scale. Of the handful of people in this area at Tufts, most are academic support staff, although a few Electrical and Computer Engineering and Computer Science faculty are involved in research in this space as well. Work in this area is motivated by the enormous computational demand and complex logistics of data science projects. On the staff side, this work chiefly consists of maintaining and supporting the use of the Tufts high-performance cluster, while faculty research, and to a lesser extent teaching, tackles how to design systems that facilitate the analysis of and machine learning on massive data sets.

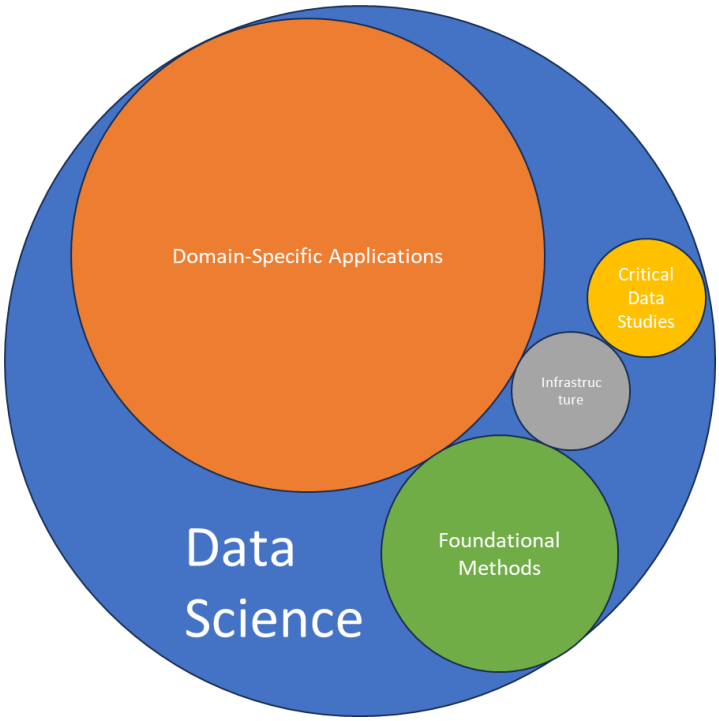

Figure 2. A conceptual illustration of the categories of data science work at Tufts

ii. Core Groups

In addition to, and at times in overlap with, the departments and programs mentioned above, there are a number of other key entities in the Tufts data science landscape.

DISC. The Data Intensive Studies Center was the non-departmental group most frequently mentioned by interviewees. That this work was conducted under the auspices of DISC, may have been a contributing factor in that finding. Nonetheless, the interview process was the first time some faculty had heard of DISC, and many that had, or that had even advised DISC at various points, felt they had little to no idea what DISC was doing. For those familiar with it, DISC is visualized as a resource hub, providing funding, expertise, and coordination to the data science community.

In particular, DISC offers the talents of roughly 6 data scientists who each work on 3-4 longer term research projects in collaboration with faculty. More than 50 faculty have conducted projects with DISC lasting more than a year, but of the faculty interviewed for this report, only a small group felt they were able to take advantage of this resource for reasons that will be outlined in a later section.

TTS. Tufts Technology Services, especially Research Technology, came up almost as frequently as DISC, especially outside the Engineering School. This group is highly utilized as a source of both data science consulting for domain applications and of instructors to teach methods courses or modules. They are also involved in organizing many of the workshops and events that are accessible to students across the university. At the same time, the data science group within TTS was constantly referred to as overworked and at capacity.

TIAI. The Tufts Institute for AI was mentioned a few times as something that existed but whose role was unknown. Only one interviewed faculty member had any direct knowledge of TIAI, saying it essentially consisted of a couple of post-docs using AI to address questions in a few select research areas, but that the hope was to grow it over time.

T-TRIPODS. T-TRIPODS was an NSF-funded center that over the last three years served a vital role in facilitating connections across the data science community. While it did not reach everywhere, it provided the foundation for a core group of data science faculty especially in Computer Science, Math, and Electrical and Computer Engineering, to connect. Faculty that had been involved with or mentioned T-TRIPODS were much more likely to know what was happening in other departments with regard to data science. Many of the recent projects and collaborations in the community can on some level be traced to folks who got to know each other, or got to know each other better, through T-TRIPODS. Had the period of the grant not been disrupted by Covid-19, it might have had an even greater impact.

D3M. Data Driven Decision Making @ Tufts is another fixed-term NSF-funded program, although this one focuses on graduate student training. Within the data science community, it serves as a bridge and a connection builder by drawing together students from across graduate schools, along with, to varying degrees, their advisers. It provides both more casual opportunities for community building along with a structured "Problem Focused Immersion" component that has evolved into multiple ongoing research projects. Although initially the goals of the program further included a creative approach to modular curriculum development, in the eyes of many of those involved in a variety of capacities, that aspect has not yet come to fruition.

CTSI. The Tufts Clinical and Translational Science Institute is *the* key resource for data science at the Medical School and one of the primary reasons for the strength of the biomedical application domain. Within CTSI are the organizations of The Biomedical and Health Data Sciences Collaborative (BHDSC) and The Biostatistics, Epidemiology, and Research Design Center (BERD). These resources collectively house most of the biostatisticians at Tufts. They engage in research and act as a service core to facilitate additional applied research within the biomedical and health domains. They also collaborate with DISC, where their director is a part-time affiliate.

Data Science Program. The School of Engineering offers "data science" concentrations both at the undergraduate and master's level. In practice, there are no dedicated faculty in this program, which largely consists of guided electives in the Computer Science and Math departments. It also focuses heavily on the machine learning side of data science, especially at the graduate level.

Data Analytics Program. The School of Arts and Sciences offers a master's degree in data analytics, which is guided by faculty in the economics department. This program does originate some of its own courses, though these tend to be taught by adjunct faculty and focus on skills for private sector industry.

HNRCA. The Human Nutrition Research Center on Aging is a long-standing USDA-funded research center. While technically its own entity, it shares many faculty with the Nutrition and Medical Schools. Like CTSI, it houses its own service core for biostatistics and data management to facilitate data science applications in the health and biomedical domains.

DIAMONDS. The Directed, Intensive and Mentored Opportunities in Data Science program was a summer research immersion opportunity supported by T-TRIPODS. It served as an entry point for

undergraduate students, especially those from underrepresented backgrounds, into the Tufts data science community.

III. Challenges of Communication and Collaboration

Despite the number of individuals and organizations involved in data science work at Tufts, the community remains highly fragmented. Most interviewees characterized the community as siloed with work happening largely in parallel isolation. While faculty knew or assumed that there were others working on similar problems, most had very little knowledge of the larger community. This was especially true of new staff and junior faculty. A typical faculty member had some knowledge of the work happening in their own department and maybe one or two closely related departments, but nothing beyond that except the general assumption that Computer Science "must teach some courses on machine learning." (See Appendix C for an informal network analysis).

DISC was repeatedly cited as an organization with the potential to serve as a hub of resources and information. At the same time, **most faculty stated they didn't really know what DISC offered, and when they did hear about things like workshops, symposiums, or other talks, it was often too close to the dates the events were occurring for them to schedule time to attend.** This was underscored by the most common suggestions for improving collaboration in the community being resources that DISC already provides, namely seed funding for new interdisciplinary initiatives, regular symposiums where faculty and students can present and learn about each other's work, and a core of data scientists who could assist with methods questions.

i. How It Happens

Instead of formal channels, collaboration and awareness of broader community consistently come about through one of three paths: Graduate Students, Collegial Friendships, or Grant Work.

1. Graduate Students

For professors that advise students, the courses and interactions those students have with other faculty are vital to their understanding of what exists in the broader community. This is in large part because graduate students were seen as having more time to connect than their advisors who had departmental and funding obligations.

2. Collegial Friendships

Faculty also know what their friends are doing. Professors occasionally referenced getting coffee and catching up with social connections as a way to hear about work happening elsewhere in the university. These friends in some cases later became interdisciplinary collaborators. And many of the suggestions for improving collaboration involved providing informal settings for researchers to meet and become friends.

3. Grant Work

Lastly, faculty meet one another when applying for grants. Since going after funding was seen as a high priority task, it was one of the few circumstances in which faculty felt they could justify spending time on outreach to colleagues in other departments. In addition, successful grant applications like T-TRIPODS and D3M were cited as the foundation for circles of collaboration with reverberating effects in the data science community.

ii. What's Needed

Improving collaboration in the data science community comes down in large part to improving communication. Most communication about events happens through email. However, with few exceptions, in order to receive emails about opportunities and resources you need to be on a list-serv for the sponsoring organization, which requires you to know that they exist in the first place. In addition, as mentioned above, these emails often arrive too late for anyone without personalized advanced notice of what is occurring to take part. One solution is to provide earlier notice of upcoming events and to notify through means beyond email, such as widespread posters across the university as a whole (not just specific buildings). Another would be the construction of a shared and well publicized calendar for the data science community.

Faculty and staff also repeatedly mentioned that a landscape map and directory like the ones being created for this project would be incredibly helpful for finding other individuals that might have shared research interests or organizations that could help them with their work. Anecdotally, I was able to facilitate several such connections from having interviewed people who were unaware of one another.

IV. Additional Proposals for Strengthening Data Science at Tufts

In addition to improving communication about existing work and resources, faculty and staff also suggested ways data science at Tufts could be strengthened moving forward. Some of these suggestions were curricular, while others were institutional.

i. Curricular

One theme that arose repeatedly was the tension between domain-oriented teaching and abstract methods-oriented teaching. Across the university there are numerous intro courses to probability or statistical methods taught by different departments. These were seen as vital entry points into data science as students were less likely to initially sign up for introductory theoretical foundations courses than to stumble into data science from an application domain. This approach also facilitated the development of domain expertise, which was widely seen as integral to the effective application of data science methods. At the same time, there was a feeling that students, especially at the graduate level would benefit from a rigorous core of abstract methods courses that would give them a foundation to take back to their domain work. Finally, at the highest level, there was a general recognition that the specific data science methods and tools a student would need to learn to pursue original doctoral research would become extremely dependent on the domain and even the project they were pursuing.

These observations lend themselves to something like an hourglass shaped curriculum for data science. A wide variety of intro courses in applied quantitative methods across the disciplines would serve as feeders into a shared core of service courses that would provide students with a robust theoretical grounding in methods and critical data studies, after which PhD students or those conducting advanced research would be able to further develop their expertise through directed study with experts in their application domains.

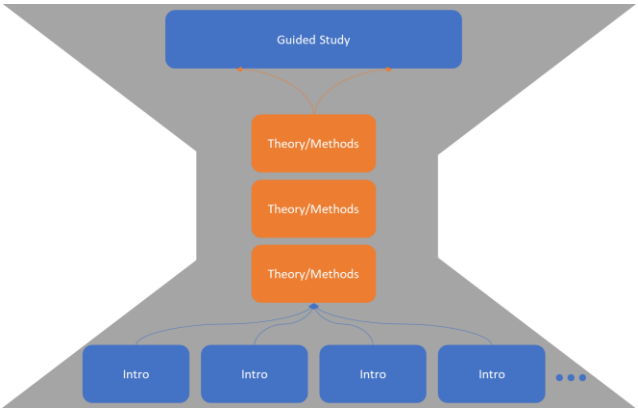

Figure 3. An hourglass curriculum for data science with blue boxes indicating departmental classes, and orange boxes indicating centralized service courses.

Within these courses, many faculty suggested an emphasis on project work. It was felt that in order to gain a deep and intuitive understanding of working with data, there was no substitute for substantial group projects. This approach would also give students the opportunity to bring their application interests into the core courses and learn firsthand about the challenges and benefits of interdisciplinary collaboration.

ii. Institutional

On an institutional level, there was concern that the existing bureaucratic structure of the university precluded taking the most effective approach to data science, which was seen as inherently interdisciplinary. For instance, in the scenario described above, how could shared service courses be taught in a way that was equitable when funding is departmental?

One solution would be to create an extra-disciplinary unit for data science as exists at other universities. This would provide a home for dedicated faculty to work outside the bounds of a traditional department while pursuing methods and critical data studies research and teaching. Other faculty could maintain affiliation to their current departments while being posted part-time to the data science unit. In addition to formally recognizing the reality of the split nature of many professors' work and allowing them to be compensated and incentivized accordingly, housing all data science related courses under one figurative roof, would allow students to find these classes more easily.

Finally, there was a repeated concern among faculty conducting data intensive work that the university was not living up to its obligations regarding access to computing resources. Although faculty had a very positive view of the team operating the high performance cluster, the cluster itself was deemed insufficient to meet the demands of research faculty whose work was primarily digital. Long queues, lack of nodes, and the difficulty of getting projects to run on the cluster led many faculty to feel they were being forced to pay for additional computing resources from cloud providers in order to conduct their research. They were especially frustrated because as they understood it, computation fell under the indirect costs covered by the large portion of their grant money appropriated by the university. Data focused faculty felt they were essentially paying twice for computing resources, while subsidizing more traditional lab-based research. This frustration could be addressed by revisiting the indirect cost rate and what it covers, and/or by significantly scaling up the Tufts cluster.

V. Next Steps

The findings of this report should be considered preliminary and the starting point for further conversation rather than the final word. In particular, faculty, staff, students, and members of the administration need to have (or continue to have) their own discussions about some of the ideas brought forward in this work.

i. Faculty and Staff

While I endeavored to interview a fairly representative set of faculty and staff, by no means were all voices captured in this report. Furthermore, in attempting to summarize general trends, I was unable to give space to every strongly felt minority viewpoint, and I am aware that there remains disagreement even on the proposals I have suggested. Faculty, especially in governing bodies such as the Faculty Senate, should take up this discussion and allow these debates to take place through formal channels.

ii. Students

The opinions and perspectives of students are almost entirely missing from this report. As one of the most important stakeholders, it is vital both that decisionmakers engage with them and solicit their views and that students themselves begin to organize and have discussions around what they would want to see from data science education at Tufts.

iii. Administration

Ultimately, any significant changes to the infrastructure or organization of data science at the university will need to be facilitated by the administration. While there have been discussions and smaller committee meetings about the topics of this report, going forward these conversations should be opened to the broader Tufts community in order to increase transparency and promote consensus on how to move forward.

Appendix A.
Interview Questions

1. How/to what extent is teaching/research about data science a part of your curricula?
2. Is a distinction between “numbers” and “narratives” important for data-driven decision-making?
3. What are the most important concepts you feel students should learn in order to become data science professionals?
4. What courses are you aware of that cover this topic within your department? Other departments?
5. What other faculty, staff, or groups are you aware of at Tufts who do this kind of work?
6. What is your perception of the interconnectedness of data science-related efforts at Tufts?
7. What would “motivate” you and your students to be actively engaged with interdisciplinary data science efforts?
8. If you were to hire three faculty members as a “Data Science Cluster” for Tufts, what type of expertise/skills would you look for?

Appendix B.
Additional Description of Interviewees and Directory Members

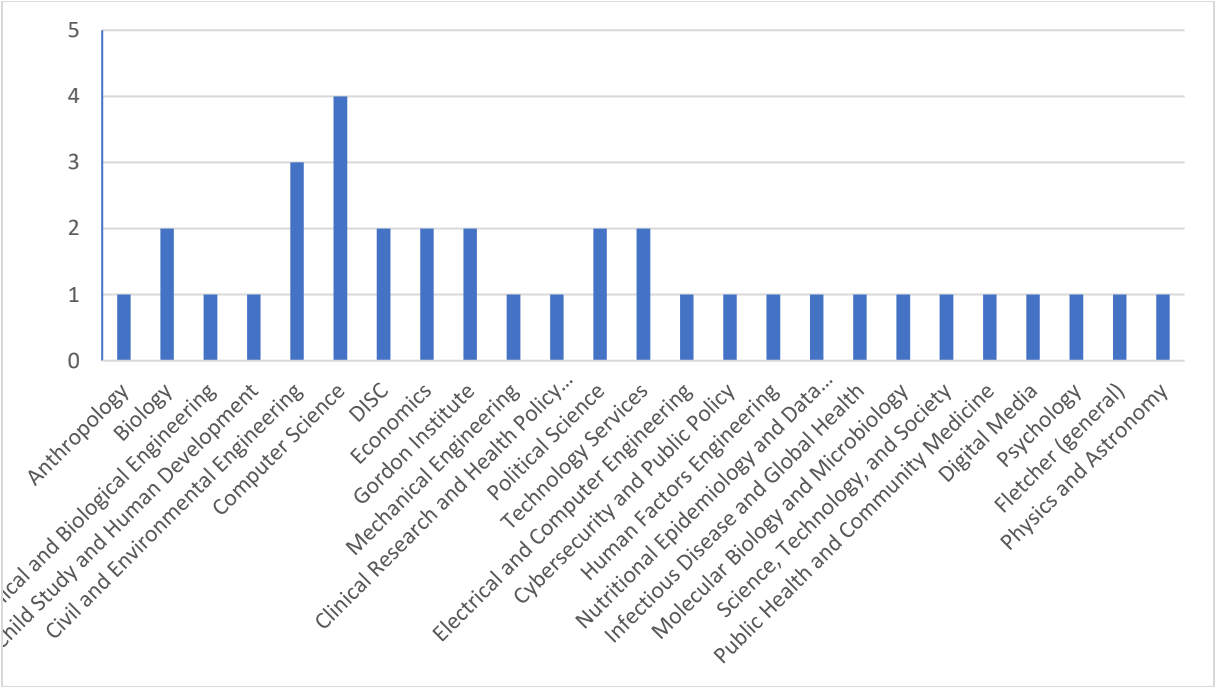

Figure B1. Breakdown of Interviewees by Department

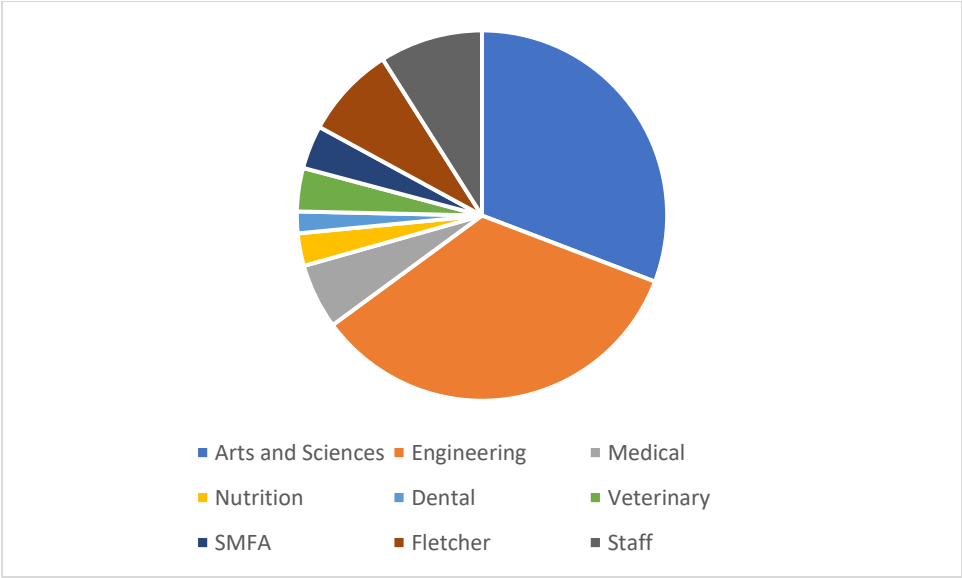

Figure B2. Breakdown of Directory Members by School

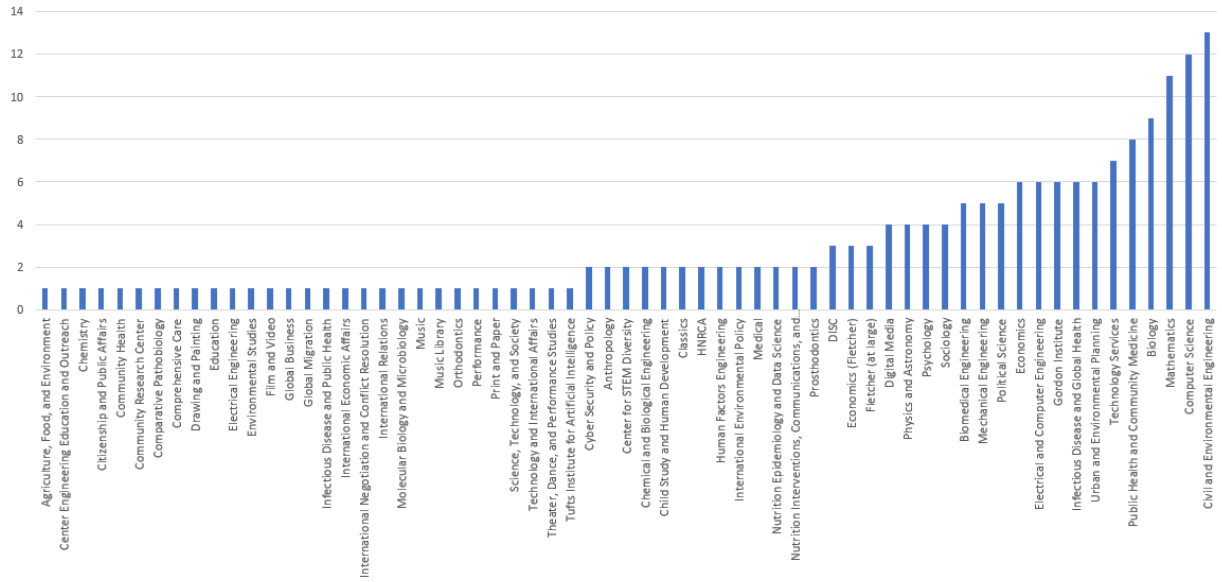

Figure B3. Directory Members by Department

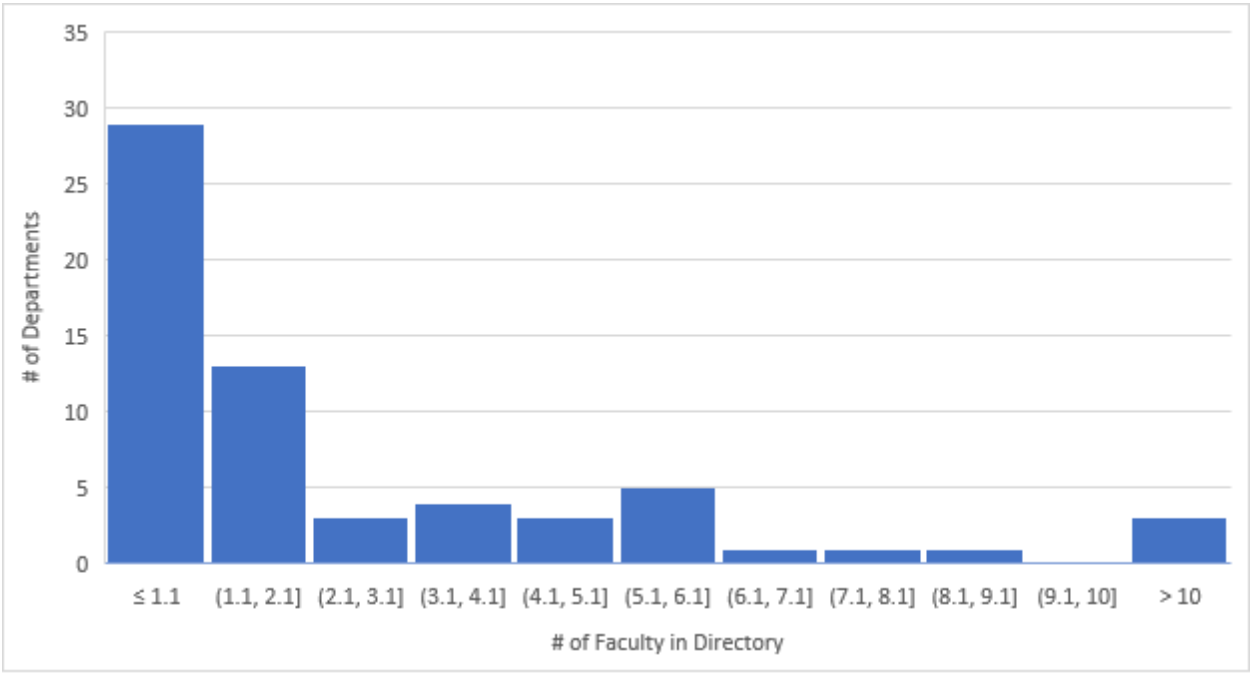

Figure B4. Distribution of Faculty in Directory by Department

Appendix C.
Informal Network Analysis

The network below consists of the interviewed individuals and any other faculty or staff they referenced.

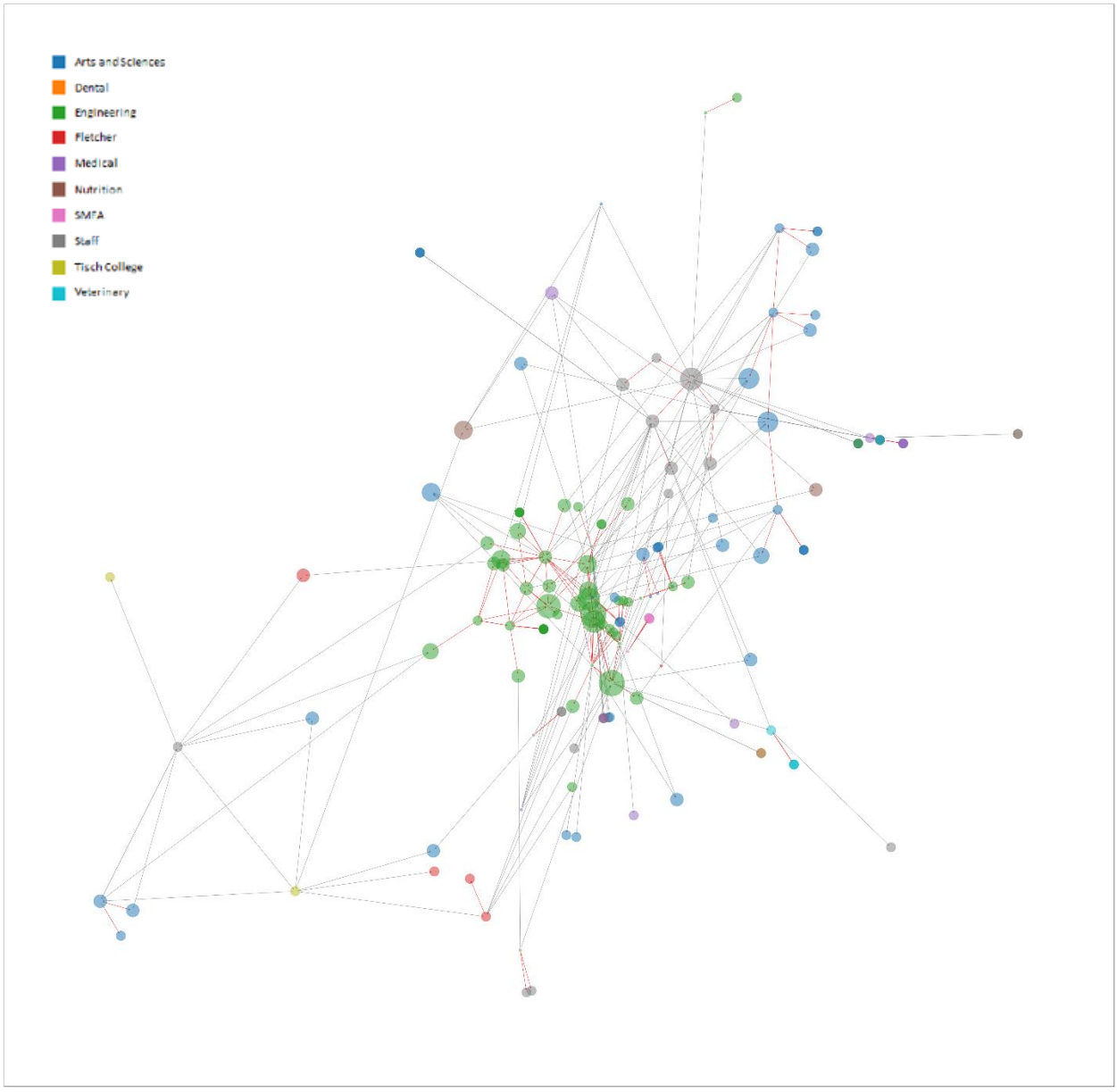

Figure C1. Sample of the Tufts Data Science Community Network

The color of the node indicates the school with which an individual is affiliated, while its size indicates the in-degree of the node (how many people referred to them). Edges are directed, originating from an interviewee and connecting to anyone they referenced during their interview. If the referee was in the same school as the interviewee, the edge is red.

Although this graph is more of an incidental artifact than a core part of the project, it corroborates the narrative of disconnection within the community that emerged from the qualitative aspects of the study. In particular, two useful observations can be drawn from it:

- First, outside of a core of close collaborators in the School of Engineering, the graph is relatively sparse. Out of 159 nodes, more than 100 have an in-degree of 1 or less. A small number of nodes have much higher in-degree values, demonstrating the importance of a few lynch-pin individuals in the data science community. Typically, these key people either lead programs or major grants.

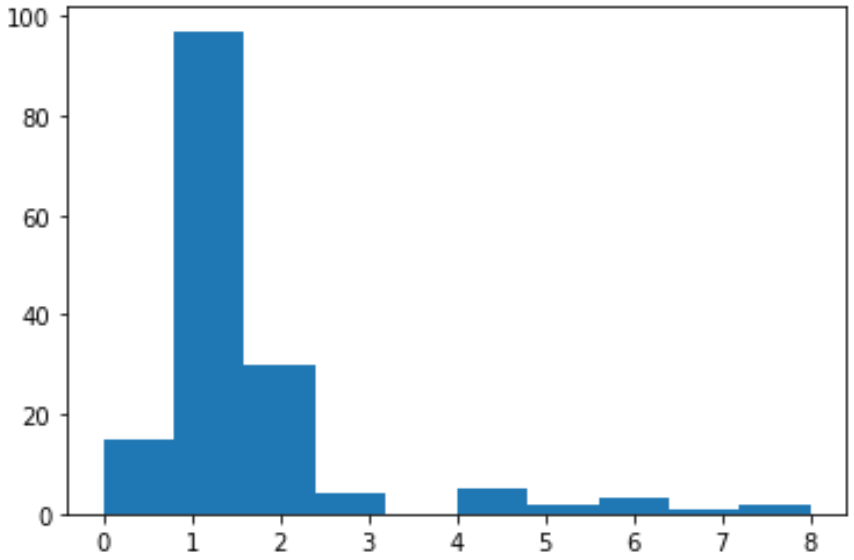

Figure C2. # of Nodes with each In-degree in the Community Graph

- Second, overall just under half of the edges in the graph are internal (between individuals in the same school). In spite of this fact, when looking just at edges originating from interviewees in the School of Engineering or the School of Arts and Sciences, twice as many edges are internal as external. This discrepancy underscores that faculty remain relatively siloed at a school level. Conversely, nearly all edges originating with staff connected to faculty across the schools.

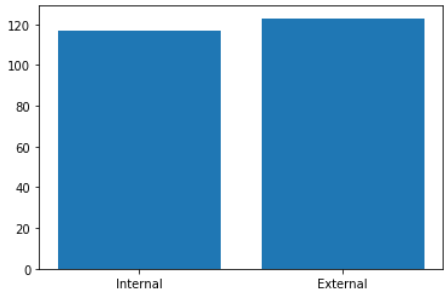

Figure C3. # of Internal vs. External Edges across the Community Graph

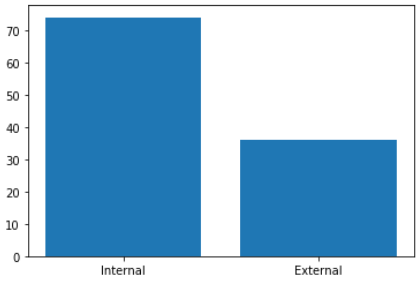

Figure C4. # of Internal vs. External Edges from Nodes in the School of Engineering

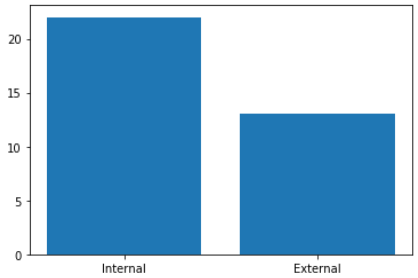

Figure C5. # of Internal vs. External Edges from Nodes in the School of Arts and Sciences

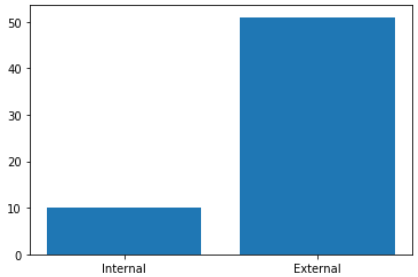

Figure C6. # of Internal vs. External Edges from Staff Nodes

Further insights might be gleaned from constructing and analyzing a similar graph using joint paper and grant data, but that was beyond the scope of this project.