Data-Intensive Studies Center (DISC) at Tufts University

Proposal Informed by the Work of the Planning Committee[1]

Draft: October 5, 2016

Executive Summary

We propose the development of a major new Tufts University Center dedicated to the integration, acquisition, and application of data-intensive research, scholarship, and education. As data acquisition, storage and analysis begins to influence every area of research scholarship and education, from the STEM and biomedical fields to the social sciences to the arts and humanities, the Data-Intensive Studies Center (DISC) will advance our ability to benefit from data resources being collected all around us. From the open source data requirements attached to federally-funded research findings, to the migration of health information to electronic records, to the internets of data and things, our world and the research we do is often limited by our ability to derive meaning from large sets of data.

Extracting meaning from complex data requires sophisticated algorithms and statistical tools to identify subtle patterns and correlations within and between datasets. The greatest challenges and opportunities are associated with the huge expansion of "Big Data" in recent years. These data are made possible by advances in the rapidly evolving, interdisciplinary, application-focused disciplines of Statistics, Data Science and Computational Science. The mining of these data for scholarship, discovery, innovation and education, in fields as diverse as genetics and social policy, will henceforth be referred to as *Applied Data Science*.

While Applied Data Science requires computational and storage infrastructure, it is important to keep in mind that the most important resource required for it to flourish is *people*. Applied Data Science is naturally interdisciplinary, bringing together scholars and researchers from a broad array of domain areas to collaborate with computer scientists, mathematicians, statisticians, machine learning experts and others, in a process of discovery. This broad collaboration must also be supported by education, training and service in cutting-edge statistical, data-scientific and computational-scientific theories and approaches.

The initiatives that we propose in this document are designed to create this vibrant community of scholars, researchers and students, so that Tufts University will proactively seize this opportunity and become a key player in what we feel certain will be the most important transformation of the academic endeavor to take place in our lifetimes.

To incubate and sustain Applied Data Science culture DISC will:

---

[1] Planning Committee Members: Bruce Boghosian (chair); Sarah Booth; Misha Kilmer; Matthias Scheutz; Gavin Schnitzler; Giovanni Widmer. The Planning Committee was supported by Augusta Rohrbach, Celia Campbell, Miriam McLean, Jo Wellins , and Kirby Johnson.

- Bring stakeholders together
- Use existing Tufts' projects as prototypes and foundations
- Build strengths through iterative processes
- Provide tools, training, and techniques to transfer skills across campuses
- Facilitate the creation and external funding of novel, interdisciplinary collaborations
- Create promoters to demonstrate and inspire

• Mission

A 2005 United States Presidential Report on Computational Science stated that "traditional disciplinary boundaries within academia and Federal R&D agencies severely inhibit the development of effective research and education in computational science, and that the paucity of incentives for longer-term multidisciplinary, multi-agency, or multi-sector efforts stifles structural innovation." The report's recommendations for academia included the following:

- *Universities must significantly change their organizational structures to promote and reward collaborative research.*

- *Universities must implement new multidisciplinary structures to provide rigorous, multifaceted educational preparation for the growing ranks of computational scientists that the Nation will need to remain at the forefront of scientific discovery.*

The mission of the DISC is to address these recommendations by creating an environment at Tufts that allows faculty, staff, fellows, and students to conduct transformational interdisciplinary research involving statistics, Data Science and the Computational Sciences, while also providing service and training in Applied Data Science and Applied Computational Science for all of Tufts University. Because it will be a matrix connecting all of Tufts University's schools and partner institutions, the Center director will report directly to the Provost's Office.

• Rationale

Why a Data-Intensive Studies Center?

A key rationale for the creation of the DISC as a center is to bring stakeholders together, working across existing departments, schools, and campuses in matrix fashion, to facilitate cutting-edge research and to prepare Tufts graduates for a future in which data analytics is increasingly important.

Research:
The dramatic, ongoing expansion of "big data" provides huge opportunities for Tufts researchers in essentially every field. Indeed, in our recent survey, 233 respondents indicated they were interested in using one or more types of "big data" in their research (82% of respondents). To seize these opportunities requires the active help of people with expertise in computational and statistical approaches to data science, both to conceptualize possibilities and to help advance

inherently multidisciplinary research projects. The need for this help can be seen in our survey results, where 179 respondents (69%) said they would like to consult with Tufts faculty with data analytics expertise, and 51 respondents said they had projects in mind that they have been unable to submit for funding due to a lack of data analytics expertise. By working across schools and campuses, the DISC faculty will be able to facilitate data-intensive research in every area of academic endeavor, through collaborations, training programs, grant writing support and active consultations, available to all Tufts researchers as well as Tufts partner institutions.

Teaching:
Knowing how to compress, search, manipulate and analyze large data sets is becoming increasingly important in the workplace. A recent McKinsey report concluded that "...by 2018 the United States will experience a shortage of 190,000 skilled data scientists, and 1.5 million managers and analysts capable of reaping actionable insights from the big data deluge." Moreover, a new federal National Strategic Computing Initiative (NSCI) has "thrust computational science and advanced computing into the spotlight." [2] Preparing Tufts graduates for these challenges will require a coordinated effort to create new courses and programs in computational and data-intensive research and data science. Just as the opportunities of data-intensive research apply to almost every academic field, these courses and programs must be accessible to students in every school and on every campus. DISC will be ideally situated to be the hub for building these new programs, bringing together all stakeholders to plan courses, degree programs, majors and minors with maximal accessibility and impact. DISC will work with the Provost's Office and Schools to secure resources to hire teaching faculty with the necessary expertise, allowing new courses or programs to be jump started.

In addition to formal courses and degree programs, there is a great need for practical training in research techniques needed to analyze diverse large data sets. Similarly, there is a need for training in design and use of high performance computing techniques and software. Patterning off the success of recent hands-on bioinformatics modules funded by a Tufts Innovates grant, DISC faculty will teach continuously evolving training modules in cutting edge data-intensive research approaches in critical research areas, open to students, postdocs and faculty across all campuses. DISC will also partner with Research Technology/TTS and the libraries to ensure that training courses in the fundamentals of statistical packages and modern programming languages and high performance computing are available to all who need them.

Cultures of Computational and Data Science at Tufts:
The Center will serve as a nexus for students, post-docs, faculty, and staff, whose research interests involve both the development of new methods in these disciplines as well as the application of these techniques in the context of specific research domains. By hosting web resources and forums, mini-symposia, seminars, journal clubs and research solution panels teleconferenced between campuses, and partnering with other institutions to maximize resources and audiences in shared activities, the Center will grow a much-needed culture of computational, data and statistical sciences. This will allow researchers to identify and leverage expertise outside of their local departments and collectively pursue visionary interdisciplinary research challenges, beyond the expertise of individual faculty members, building institutional knowledge so that

[2] Miriam Quintal, New Initiative Focuses on Computational Science and Advanced Computing, SIAM News, March 2016

specialized expertise is not lost when students or post-docs move on. Center researchers, faculty, and staff will be data experts and advocates—able to teach cutting edge techniques, disseminate best practices, and help faculty and students realize the potential for data analytics and computational science in their research and careers.

A Tufts-wide resource:
By creating an umbrella structure that reports to the Provost's Office, but at the same time operates both across and within existing schools, the DISC will use the intimate domain knowledge of departments and schools to focus on the overarching challenges of Applied Data Science and Applied Computational Science. DISC's ability to potentially bridge all areas of academic study will foster growth based on Tufts-specific challenges and values.

Guided by research-driven priorities and strengths, DISC will grow at just the right rate for the community it seeks to serve, support, and develop (see Figure 1). Drawing on the success of Tufts Collaborates and the CTSI, the Center will be able to bring together existing faculty expertise to fulfill its mission. Initially building on existing faculty expertise, DISC will grow through flexible joint appointment mechanisms. It will also be able to target new hires to identify and fill critical gaps in the overall Statistics / Data Science / Computational Science expertise portfolio of Tufts as a whole.

• Proposed outcomes

Support for data-intensive research at Tufts:
Center faculty, trainees and staff spread across all the campuses, will actively provide training and help individual researchers and their laboratories identify opportunities, and overcome currently daunting start-up barriers, to apply the power of Data Science and Computational Science approaches to their research. Many of these consultations are expected to create novel, interdisciplinary, collaborative projects that will be highly competitive for outside funding. The Center would also be in a strong position to partner with the Tufts CTSI which is required to offer researchers a user-friendly data management system, and encourage compatibility of their research systems with broadly accepted content and technical standards including those adopted by the Department of Health and Human Services for use in U.S. health care and public health operations. In addition, the Center would help

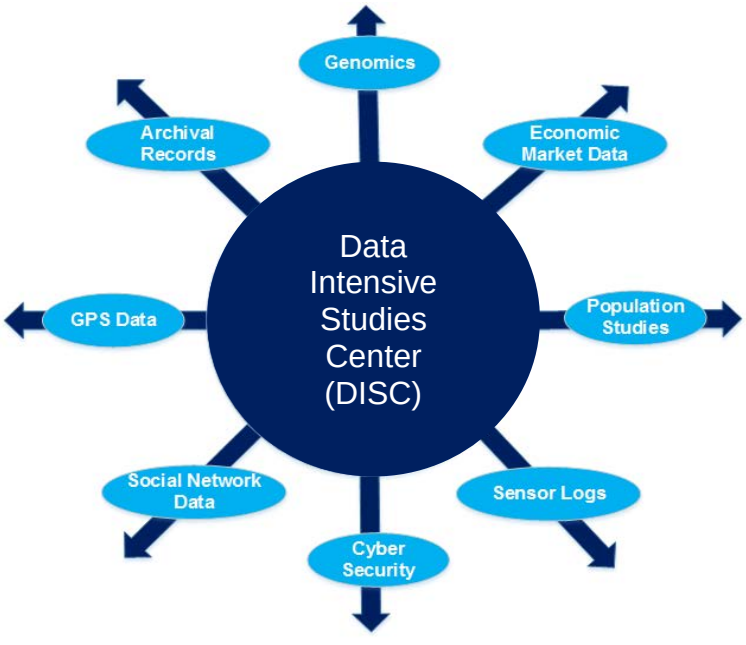

Figure 1: A schematic depicting how DISC will serve as a central nexus for expertise to advance multidisciplinary research in a broad array of fields.

identify opportunities for relevant multi-laboratory, cross-disciplinary research programs, inter-institutional collaborations and provide assistance in assembling program project or consortium grants.

Creation of data science courses and programs:
Center faculty will work with relevant departments to develop and implement curricula in statistical, computational and applied data sciences that will be available to students and researchers on every Tufts campus. Center facilitated courses will provide the practical expertise needed to advance research, and will support new degree programs and concentrations, within existing Schools, that prepare Tufts graduates for leadership roles in a future full of "big data." The Center will also collaborate with existing schools and partner organizations to develop professional masters/certificate programs that will both help address the national need for data scientists, and also provide revenue to support other Center activities. Finally, the proactive, cross-disciplinary and cross-campus nature of a Center is expected to greatly inspire alumni and other donors who would like to see Tufts seize the enormous potential of the ongoing Data Science revolution.

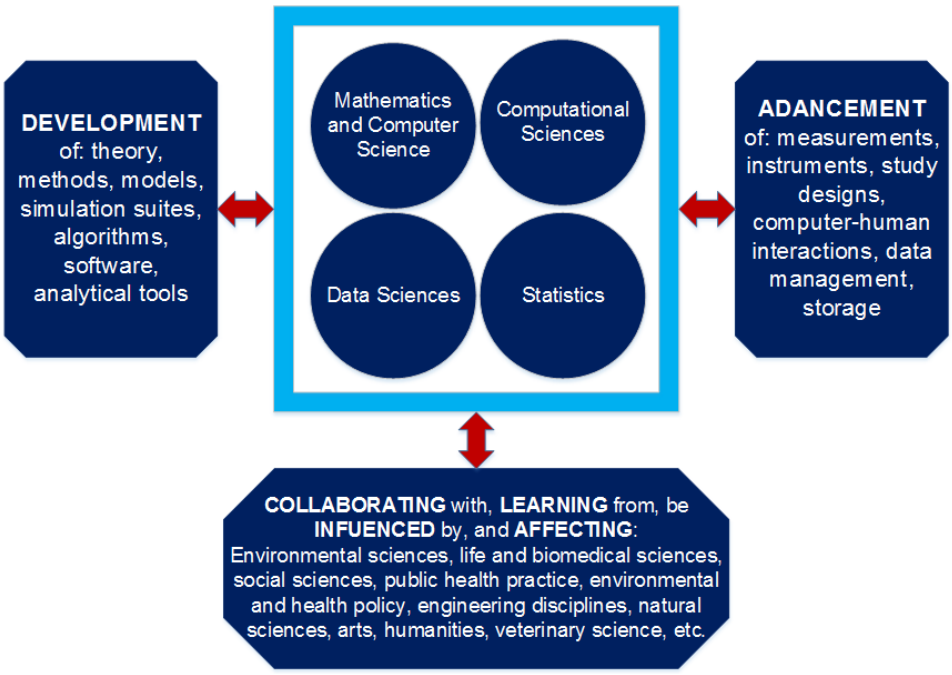

Figure 2. Interrelationship among disciplines, research fields, and research and training areas as DISC cultivates a data science culture in the research space. The iterative flow deliberately sequences divergent and convergent steps, responding to Tufts specific priorities, values, and constraints.

Illustrations of just a few of the areas in which the Center could benefit research and pedagogy in of a broad range of Tufts Schools and departments is shown in Figures 1 & 2. From the analysis of large clinical data sets, to web scraping, text mining, sentiment and content analysis, and large-scale simulation, Tufts researchers will benefit from the culture of computational and data science fostered at DISC. Rather than outsourcing through collaborations, DISC will bring data-intensive study to bear on the interdisciplinary culture that defines Tufts' research community.

• Center programming:

A brief listing of the activities of the Center is given below, followed by a discussion of existing

strengths and initiatives and their relation to Center activities.

Key aspects of Center programming:

1) Research initiated by center faculty.
2) Research consultations with other Tufts faculty (to conceptualize and advance multidisciplinary data-intensive projects): Center faculty
3) Grant application assistance (to identify funding sources & opportunities, strengthen data science aspects of proposals, and coordinate multi-investigator proposals): Center director, faculty & grants officer.
4) Intensive collaborations (often arising from consultations that have led to funded projects): Center faculty.
5) Seed grant program (to allow nascent multidisciplinary data-intensive collaborations to generate preliminary data needed for outside funding): Center faculty & staff.
6) External funding for the Center mission.
7) Essential didactic courses (in collaboration with relevant schools and departments … our survey identified particular needs for courses in upper level statistics, optimization methods, machine learning and bioinformatics): Center faculty in partnership with school/department faculty.
8) Hands-on practical courses (in collaboration with relevant schools and/or independently, to teach cutting edge data analytic approaches needed to advance research - such as next generation sequencing data analysis - available to all Tufts researchers): Center faculty, plus students & post-docs as TAs.
9) New degree programs, majors & minors (help design & implement, in collaboration with all interested schools & departments, reducing redundancy & enhancing accessibility to all Tufts students): Center director & faculty, in partnership with other schools & departments.
10) Data Science master's, certificate and/or continuing education programs (in collaboration with relevant schools): Center faculty.
11) Infrastructure improvements (in collaboration with Research Technology /TTS & others: necessary software or computer upgrades, seamless videoconferencing for DISC-supported cross-campus courses, seminars & meetings): Center director & staff.
12) Web presence (forums for discussion, faculty and student research/interest listings - connecting expertise with needs, how-to's and links to web resources): Center faculty & staff.
13) Outreach (mini-symposia, sponsored lectures, journal clubs, research solution panels, often teleconferenced to be available on all campuses): Center faculty & staff.
14) Evaluation (periodic evaluation of the center's success relative to previously-established benchmarks, and identification of evolving needs/opportunities, using these assessments to make necessary adjustments to center priorities or structures).

DISC will build from existing strengths, collaborating closely with established institutes, schools and programs, and either bolstering or integrating newer initiatives. Existing strengths and resources at Tufts include: (1) strong statistical expertise for clinical studies (CTSI) and for epidemiology (InForMID), (2) a strong and diverse Mathematics and Computer Science faculty, (3) the powerful Tufts HPC Cluster and other research computing resources provided by Research Technology/TTS, (4) bioinformatics expertise at Tufts Genomics Core Facility, (5) several organizations and initiatives, by faculty and students in and across all Tufts campuses.

Among the initiatives already up and running within the School of Arts and Sciences, are the Perseus Digital Library Project,[3] the Perseids project[4], and the Bodin Project;[5] these will be able to expand in partnership with, or find a home in, DISC. Other projects from the Arts and Humanities are likely to follow. For instance, big data is having a growing impact on musicology. Music theorists are using large data sets to understand historical change in musical repertoires and to explore the roles of conventional structures in harmony and form. Large databases of medieval music make it possible to understand transmission of melodies and quotation that otherwise would take years to uncover. Last year Tufts hosted Alessandro Bratus, a visiting scholar who created a 5 million Euro EU grant using digital humanities approaches to tracking social and creative networks of Italian film and song. DISC would not only make projects like this possible, but likely.

Within STEM sciences, DISC will provide a locus of support for initiatives and programs that are integral to its mission and the intellectual life of scholars and researchers at Tufts. For instance, DISC could facilitate the success of the Cyber Security and Policy program—a collaborative effort that brings together strengths from 3 schools (A&S; Engineering; and Fletcher), by providing needed faculty expertise or by supporting existing faculty to devote their time to build and sustain the program. Similarly, DISC would be able to build from the researcher-supported Computational Biology Initiative (with members from Cummings, TUSM, the Friedman School of Nutrition, HNRCA, Tufts Genomics and A&S)[6], to support data analytics approaches in biomedical research and education. As DISC develops, it will work closely with other schools to strengthen and modernize existing programs or to launch new programs or concentrations. For instance, the Center would help establish needed Bioinformatics courses for students at TUSM, Cummings and Fletcher schools, and A&S, and might partner with these schools to launch a Bioinformatics degree program. The Center could also partner with Mathematics, Computer Science and other relevant programs to develop professional masters and certificate programs, as well as continuing education programs, strengthening a culture of computational and data science at Tufts and helping to fill a national need for data science literacy, while generating tuition support to fund Center activities.

---

[3] Since 1985, the Perseus Digital Library Project has explored what happens when libraries move online. Perseus is a practical experiment exploring the possibilities and challenges of digital collections in a networked world.

[4] The Perseids platform allows students and scholars to collaborate on digital editions and annotations in an integrated online environment. After successfully completing a first phase of development, Perseids is now poised to support the elaboration of a new undergraduate curriculum in the Humanities in the US and abroad. The cornerstone of this initiative is the integration of teaching and research in the Humanities classroom, emulating the way students participate in research teams in the natural sciences.

[5] Using the Perseids platform, the Bodin Project will produce an electronic variorum edition of Jean Bodin's *Les six livres de la république.*

[6] http://sites.tufts.edu/cbi/about-us/

Together, these programs are as essential for the future of a liberal arts education as they are integral to 21st century research approaches. In each case, DISC will work to help support these new programs, both directly, by providing faculty with needed expertise or by supporting the efforts of existing faculty, and, indirectly, by creating a vibrant culture of data science across Tufts. In this way, DISC will provide schools the needed intellectual space to expand interdisciplinary, collaborative programs, helping prepare and shape active citizens and productive members of 21st century workforce with 21st century tools.[7]

By virtue of its cross-campus nature, DISC will also be able **to reduce duplication of effort** by fostering communication and coordination between, and/or by partnering with, schools considering new programs related to computational and data science. It will also be able to identify and fill gaps not covered by individual programs. For example, building on the existing portfolio of activities, the Center and its affiliates will teach continuously evolving, short, hands-on courses, designed to provide functional, essential competencies in the most relevant and powerful big data techniques in a range of fields (available to students, staff and faculty on all campuses). In addition, the Center's partnership with Research Technology/TTS and the libraries will ensure that tutoring, workshops and courses are available to allow students, at any level and across all campuses, to rapidly become proficient on computing platforms, statistical packages and programming approaches needed for their degree programs or research. The Center will also be able to provide the support needed for small programs (which have barely enough faculty time to cover required courses for those programs' students), to open its doors to qualified students from other schools and programs at Tufts. It will also be able to fund some infrastructure improvements to make cross-campus learning at Tufts much more powerful and user-friendly.

The second, equally important, core function of the Center will be to support and nurture cross disciplinary research using applied data science approaches. In addition to the research training made possible by the hands on and didactic courses, described above, the Center will more directly facilitate faculty research projects, through active consultations--to brainstorm possible opportunities and to provide the expertise needed to jump start projects. It will also sponsor a vigorous peer-reviewed seed grant program. The Center will also provide active support for the development of small and large cross-disciplinary grant proposals, with the expectation that the small investment in consultative support will allow the development of powerful, well-funded collaborative projects. The Center will also partner with Research Technology/TTS and research core facilities such as Tufts Genomics, to ensure that cutting edge hardware and software resources for data-intensive or compute-intensive applications in every field, as well as the resources to run them, are installed, well documented and accessible. In summary, by providing a centralized forum for data science across Tufts University, the DISC will catalyze innovations and insights in all areas of academic endeavor in ways that are not currently possible at Tufts.

• **Competitive landscape**:

The collaborative space DISC will provide is also a central feature of its charter and will make it

[7] https://tinyurl.com/insidehighered-liberarts

a unique resource in the region. While other universities can, and undoubtedly will, develop their own strategies for data/computational science, most of them revolve around the STEM fields. DISC has the opportunity—supported by a successful track record—to create a collaborative space sustained by a culture empowered by data science but not solely devoted to it. In other words, DISC will be STEM-rich, but not STEM-centric. The DISC initiative capitalizes on Tufts' distinctive position in the institutional landscape as a research-intensive university with a strong commitment to the liberal arts. Thus, DISC will attract a particular kind of workforce—one dedicated to taking on the grand challenges but with a perspective that is balanced between the sciences and the liberal arts.

We have excellence in the classical STEM fields, and we additionally have the Fletcher School (international cyber security), the Museum School (digitized art), the Cummings School (veterinary clinical informatics and translational medicine), the Medical Center (next-generation genome sequencing), the CTSI (medical records and health policy) and the Perseus Digital Library, to name a few footholds Tufts has in the data-verse. Perseus is a Mellon-funded project, and its offspring characterizes Tufts' strength in the Digital Humanities. Since its inception in 1985, Perseus has spawned a nested set of projects that, as part of an integrated effort, would distinguish DISC from other local and regional efforts to serve an expanding arena of data-rich scholarly research that is not limited to or determined by the STEM fields. Part of what sets DISC apart from local comparisons, such as Northeastern's NuLab or Boston University's Center for Computational Science Center,[8] is the way in which Tufts' exceptionally collaborative environment brings together strengths from across its campuses, schools, and centers, rather than isolating them. Thus, our argument for centralizing data and computational resources is more than a matter of conservation of effort. Locating inquiries in the fields centrally will both concentrate and distribute knowledge through the research matrix, fostering the kind of collaborative investigations that have become a signature of Tufts' institutional profile.[9] DISC researchers will find the relatively low barriers to collaboration at Tufts shrink even further as space and time for the exchange of ideas are built into the very design of the enterprise.

• **Risks**:

The greatest risk of all is inaction. But in order to act, and act strategically, we must develop a shared vision. DISC's success depends on deep collaboration among and between departments, schools, and campuses. This collaboration is needed to allow DISC to tap the expertise of existing faculty, and for DISC to recruit outstanding new faculty through joint hire arrangements. Collaboration between DISC and existing departments is also needed for DISC's teaching mission. For instance, emerging plans for program development in departments and schools forecast the value of DISC for Tufts. **It will be critical for DISC and these departments to**

[8] http://bu-ccs.com/

[9] Some examples: http://www.northeastern.edu/nulab/
http://www.bu.edu/datascience/
http://www.bc.edu/research/research-day
https://www.brown.edu/research/projects/digital-society/
https://www.ccv.brown.edu/research/vignettes

coordinate closely in developing programs relevant to the DISC mission, to avoid duplication of effort, to allow DISC resources to strengthen these efforts, and to ensure that courses essential for interdisciplinary data-intensive research be available to Tufts faculty and their research teams on all Tufts and Tufts affiliated campuses.

Safeguarding against these potential risks is the fact that the initial DISC director and core faculty will represent many schools and departments. They will, thus, already have close ties to their home departments, which will greatly facilitate the negotiation of mutually beneficial agreements on these issues. Moreover, since DISC's core mission is to benefit all schools and campuses by increasing Tufts-wide capabilities in fundamental and applied data science, the Center will seek to have all its interactions be clearly beneficial for the schools and departments it partners with. This will be made possible by tapping Center-specific resources, such as external center grants, donations and endowment, and tuition from new programs in which the Center plays a predominant role. Accordingly, despite these potential risks it is anticipated that, with stable leadership to ensure steady development and the collaborative support of the entire university, DISC can create a vital intellectual center capable of empowering the Tufts research and teaching enterprise with all that data science promises.