A Cornell Center for Engaged Data Science

Why Data Science? Research and scholarship in the 21st century are being dramatically transformed by their interaction with data, in a vast array of new forms and at rapidly increasing levels of scale. This is a transformation that is gathering force across essentially all areas of study: it is visible, for example, in the profound effect of high-resolution molecular data and the ever growing genomic data in biology and medicine; in the ways in which digital traces of human behavior have enabled new styles of research in the social sciences; in the opportunities that computational analyses of large text collections have opened for scholarship in the humanities; and in the effect that data-rich approaches have had on central applications in engineering, agriculture, urban infrastructure; on driving breakthroughs in the physical sciences, and in many other areas.

Viewed in this context, the emerging domain of data science is not so much a single free-standing field as an inter-disciplinary effort of unusually broad scope -- impacting, as it does, nearly every area within the university. It is concerned both with the core methods -- computational, mathematical, and statistical -- for working with data, as well as with the integration of these methods in their myriad domains of application.

The universities that will emerge as leaders in data science are those that can address this profound breadth in scope in the context of excellence in the domains of application -- making it possible for the methodological innovations in data science to engage with the full range of applications and problems in the world. Effectively realizing this style of *engaged data science* will be crucial for the next phase of the evolution of data science, and universities that wish to take part in this evolution will require a combination of three crucial ingredients: (i) an institutional culture of cross-disciplinary engagement with broader questions and applications; (ii) a history of world leadership in specific applications of data science across diverse areas, which can serve as a foundation for integration, together with leadership in the core methodological fields that are driving these advances; and (iii) a central coordinating structure that can catalyze integrative activities across its campuses.

Capitalizing on Cornell's Excellence in Data Science Cornell is well-positioned to expand upon and strengthen its role as a leader in data science through its ability to capitalize on the above three ingredients. Cornell's long-standing culture of engagement, inherent in its land-grant mission, and amplified in the past several years through the external engagement mission of Cornell Tech and activities including Engaged Cornell, positions it to explore cross-disciplinary opportunities with substantial external impact. Cornell has also long been a leader in data-intensive applications; indeed, it is recognized as occupying the pre-eminent position in a number of such areas, including open-access scholarly publishing, legal information and public opinion research; citizen science; computational social science; gravitational wave astronomy and cosmology; smart transportation; digital agriculture, plant sciences, evolutionary genomics, and sustainability, including global-scale public health; precision medicine and meta-genomics; informatics-driven personalized health; and the foundations of security and privacy for data systems. The breadth of data-oriented topics in which Cornell leads is perhaps unmatched, and it provides the underpinnings for leadership in data science more broadly.

The ingredient that has arguably, thus far, been missing at Cornell is the third one: a central coordinating structure with resources to create an integrated presence for data science. This is what we propose here -- a *Center for Engaged Data Science* that can build on the extensive area-specific successes in data science at Cornell and serve as a research hub for data science across the campuses. Its goal would be to promote existing and new research activities, as well as to foster and enhance engagements with applications; to coordinate and help develop the necessary educational programs in data science enabling them to span college and campus boundaries; to support the global data resources and archives maintained at Cornell; and to connect with partners outside academia in the realization of new applications.

An Integrated Multi-Campus Approach to Research in Data Science Cornell's leadership in data science leverages scholarship that spans all of its campuses – Cornell Tech, Weill Cornell Medical College and the main campus in Ithaca, as well as the neighboring NYSEAS in Geneva. The proposed Center would have elements that invest in all of them, and be particularly attuned to developing mechanisms for bridging between them. One

distinguishing characteristic of data science research at Cornell is the extent to which it is truly cross-disciplinary, and the key goal of the new Center is to build a platform for strengthening the mechanisms already in place that foster these collaborations. With this in mind, we believe that the largest investment of new resources should be allocated towards the creation of a large pool of Research Assistant Professorships and Junior (Postdoctoral) Fellows in Engaged Data Science. The aim is to create a new critical mass of activity that will provide an exciting platform for research in which data scientists driven by a wide range of application domains work together in a common space that promotes interaction, and enables them to feed off each other's advances. Ideally, we imagine a group of 15 young scholars, 5 per year, each for a 3-year term, that will attract the foremost talent of the generation bridging between doctoral study and regular tenure-track appointments. The Center will provide a nurturing environment for them, equipped with the requisite computational and mentorship infrastructure to catapult their burgeoning research agendas without the need (as they would within a traditional academic appointment) to devote significant effort to funding their research enterprise. These scholars would typically be linked both to one senior faculty member grounded in a particular application domain, and another with particular methodological (computational, mathematical, or statistical) expertise. Further, we expect that many of these will have cross-campus appointments, bridging between them, sometimes 50/50, or in other cases, 80/20. We reason that scholars at this stage of career development are often amenable to such dual-mode appointments given appropriate resources, and their activity will provide inestimable value in linking research activities across the campuses. Technology has improved what can be done remotely, but there still is no substitute for regular physical presence in collaboration.

Another objective of the proposed initiative is to enhance the targeted recruiting of leading scholars to maintain Cornell's preeminence, by setting the agenda for data science in novel domains of application, as well as new methodological advances. Given the impact that data science is having across the full spectrum of academic disciplines, many units are eager to invest in this area. A *Center for Engaged Data Science* will substantially leverage this interest by boosting Cornell's ability to attract the most influential mid-career academics in data science; furthermore, this strategy would be greatly enhanced by creating a resource pool that would provide 5-year startup research funds for two tenured hires per year to be coordinated across the university, particularly for areas of data science research that span multiple colleges. Finally, these should be complemented by newly endowed professorships that serve to energize this data science initiative with senior hires whose interests span multiple units, as well as providing the academic leadership for the Center. Close attention will also be given to opportunities for advancing diversity in hiring for all of these positions. Given the rapid emergence of data science, many of Cornell's peer institutions are competing to match our current leadership, and hence investments in attracting both influential mid-career and senior faculty will be required to sustain this role.

The Center will become the focal point across the campuses for thought leadership and day-to-day, week-by-week research in data science at Cornell. In addition to hosting internationally distinguished experts in regular weekly colloquia and seminars spanning the full breadth of the area, the Center will coordinate a number of research-initiating activities. For example, prior to each semester the Center would assemble a summit of both Cornell and external expertise to jump-start research in a particular emerging domain, bringing together faculty and researchers from all of its campuses in a 2-4-week "boot-camp" atmosphere to commingle experts with the targeted methodological expertise as well as domain experts in the designated area of application. Critical to the success of such an activity is an Innovation Fund for Visiting Scientists that provides funding both for mobility between Cornell campuses as well as accommodating the external expertise. These summits will not only stimulate cutting-edge research, but will strengthen the collaborative bonds between the campuses, and provide significant external visibility for Cornell in data science.

Campus Infrastructure Advancing Data Science Research across the university is being framed by a new data-centric perspective, at scales unimaginable only a few years ago. Methodological advances in data science are creating opportunities by introducing general-purpose tools effective at such scales that impact domains well beyond the scope intended by their creators. This has unleashed a demand for both access to cloud-scale computational resources and methodological expertise. We propose that the most natural approach to meet this demand is to build on the existing campus-wide infrastructure to ensure that innovative research in non-

traditional areas can incorporate massive-scale data-driven approaches. For example, we propose that there be an Infrastructure Fund aimed at seeding computational activity in such emerging areas, complemented by a "consulting-style" outlet for providing support to initiate these activities. The creation of such an entity would not only help to infuse data science within areas for which such expertise is currently "outside their comfort zone", and could coordinate and provide centralized support to enable faculty and researchers to fully leverage existing Cornell facilities with extensive expertise in key elements of data science, such as the Center for Advanced Computing (CAC), the Computational Biology Service Unit (CBSU), and the Cornell Statistical Consulting Unit (CSCU).

One important currency in advancing research in data science is easy access to archival data. Cornell plays a unique leadership role among academic institutions worldwide in heading efforts that innovate in a range of application domains by defining, curating, and promoting accessible corpora of data. A Center for Engaged Data Science would enable these to be linked through a common portal, and help crystalize the notion that Cornell is at the cutting edge in developing and advancing these resources. The Cornell Lab of Ornithology, though its e-Bird project, pioneered the notion of citizen science, and the resulting (and ongoing) data set has been the basis of newfound understanding in bird migration patterns. The Roper Center for Public Opinion Research at Cornell is the largest archive of public opinion data, with surveys dating back to the 1930's. The Legal Information Institute is known internationally as the leading "law-not-com" provider of public legal information, providing both data and software tools to advance open-access research on the full gamut of legal issues. Weill Cornell's department of Healthcare Policy and Research is the home for the New York City Clinical Data Research Network, whose medical histories for as many as 6 million patients enables ground-breaking research in a way that ensures their privacy and security. Cornell's arXiv is the world's premier e-print repository in physics, math, computer science and related disciplines enabling scientists worldwide to share and access research, and also serves as a critical source of data enabling the study of the evolution of science itself. The CAC built and hosts the NANOGrav gravitational wave search archive, PALFA pulsar and Fast Radio Burst search data, and planetary radar observations, comprising half a petabyte and growing, of data made available to the worldwide research community. Viewed collectively, even just these few examples highlight Cornell's role in advancing the curation of data archives to stimulate data-driven research; furthermore, these efforts exemplify Cornell's institutional commitment to data science as a means to guide valid and informed decisions for the public good. Having a common structure with flexible support resources would greatly enhance the visibility of these separate efforts, and would accelerate the process of building further such initiatives within the university.

Enabling Data Science Education Across Cornell Data science is already having a substantial impact on educational programs at Cornell, at all levels and on all campuses, but there is a seemingly insatiable thirst for further opportunities for students to advance their understanding and to seize upon the opportunities that this revolution is creating. The nature of developments in data science require a tight integration between the emerging research and the training of a new generation of researchers; as a result, we believe that it is critical for the proposed Center for Engaged Data Science to play the leading role in helping to shape and coordinate programs, courses, and other educational mechanisms across the full span of the university.

Although there are many pressing needs, the most immediate action would be the creation of a Graduate Field of Data Science, initially as a "minor-only" Field, to help meet the pressing need of offering doctoral students in a broad cross-section of disciplines a means to gain the understanding and computational tools of data science, and integrate large-scale data-driven methods into their own research. This step could then shortly be followed by adding a PhD-granting component, training the next generation of engaged data scientists with an innovative program hand-tailored to meet the needs of this emerging discipline. In a manner similar to the Research Assistant Professorships, we would strive to create a culture in which many of the PhD students are co-advised, with one advisor providing methodological guidance in concert with another providing application domain expertise. Of course, in addition to developing the course infrastructure to support the added demand for appropriate doctoral-level courses in data science, the development of the PhD-granting program will require additional graduate student fellowships and administrative support for implementing this initiative.

For data science at the Master's level, there are already a number of programs either in data science or in heavily overlapping disciplines – for example, there is the MPS in Applied Statistics with the Data Science option offered by the Department of Statistical Science in CIS, an MS in Biostatistics and Data Science in the Department of Health Care Policy & Research at Weill Cornell, and the Data Analytics concentration in the ORIE M.Eng in the College of Engineering; furthermore, there are numerous opportunities for broadening this coverage, such as the new campus-wide Masters in Public Health. One role for the proposed Center will be the coordination of these programs and course offerings as this level, providing a common platform to help distinguish relevant characteristics as they serve differing student populations with varying educational goals.

Perhaps the greatest educational challenge for meeting the expanding needs for exposure to data science comes at the undergraduate level. Basic knowledge in data science will be necessary for students to become informed citizens, much less leading practitioners in their respective fields. It will be important to scale existing courses, and expand offerings to enable the creation of an undergraduate minor in data science in each undergraduate college. A more ambitious goal, and one for which initial steps have already been taken, is the creation of a structure under which a new introductory course, co-taught by several departments, serves as the hub for an array of gateway courses. These gateway courses link the foundational material from the introductory course to a wide span of focused application domains. It is intended that this introductory course will provide an entry point for students across the university who have limited or no prior background in computing and statistics. The course will enable them to become critical thinkers in analyzing real-world data, and introduce them to them to relevant tools in computation and statistical inference; in our data-rich world, it is imperative that we educate our students so that they understand how to validate quantitative statements, as well as to reason about the inherent limitations on making such conclusions reliably. Furthermore, the broader goal is to provide the springboard that connects to the subsequent gateway courses, creating the ability for any student, in any study, to apply a cutting-edge data-centric approach to any discipline. This approach of introducing core material in tandem with a broad range of relevant application domains is known to be effective in attracting a diverse student body for STEM disciplines.

Building a Center for Coordinating Across Campuses We propose that most of the programmatic decisions will be guided by a Cornell-wide rotating executive committee together with an academic director for the Center; there will be ongoing decisions, for example, in recruiting the Research Assistant Professors/Junior Fellows, coordinating among search committees for the mid-career startup funds, and promoting the development of an expanding assortment of gateway courses. The primary function of the Center is to provide a means to focus, coordinate, and build upon a diverse set of initiatives and resources across the campuses in a way that serves to demonstrate Cornell's leading role in this critically important and rapidly evolving discipline. One notable aspect of data science as a discipline is that it is not only rapidly transforming a broad cross-section of research domains, but it is consequently having a similar impact on a broad range of enterprises, both reinvigorating some dormant ones, as well as creating entirely new industries in the process. Consequently, it is extremely important that the Center build strong cooperative relationships with industry, NGO's, and governmental agencies: many of the next generation of important data science problems to solve are likely to emerge from outside of academia; we imagine that the Innovation Fund that promotes both cross-campus and external collaborations will also be used to help Cornell faculty explore industrial collaborations. In light of the important role that industry plays in this domain, we propose that the Center have an energetic, engaged, Industrial Advisory Council to facilitate these interactions.

Cornell's strength in data science is, in large part, due to the institution's special character in promoting and nurturing cross-disciplinary work; there has probably never been another newly emerging discipline based so definitively on the symbiotic relationship between transformative advances in methodological tools and a staggering array of application domains. Within the framework of the "radical collaborations" outlined by the Provost, the discipline of data science has the notable characteristic of interfacing with the full breadth of the other initiatives currently being explored, ranging from genome science to sustainability to the social sciences and the humanities (as well as essentially everything else "in between").

In some ways, Cornell's strength in interdisciplinarity also gives rise to its one weakness in this regard – the fact that there is no centering focus that can help catalyze the full breadth of all of its campuses and serve as the mechanism for coordination and hence allow us to fully capitalize on the inherent potential for Cornell's dominance in this critical emerging discipline. Such a Center requires substantial investment, but due to the excitement over the seemingly limitless nature of the new advances to come, there should be opportunities to use the creation of a Center as a springboard for new endowment. This can be further supported by the potential investment of sponsored research in data science – for example, the NSF has initiated a 2-phase program in TRIPODS, for founding centers for data science research, and Cornell was awarded one of a small number of Phase I exploratory development grants potentially leading to such a more substantial award. Finally, the explosion of demand for educational programs in data science should also help to provide the means for the kinds of support, coordinated by the Center, that would lead to the expansion, development, and infusion of data science opportunities throughout Cornell.

Data science is a disruptive technology that is accelerating and reconfiguring progress in truly every field of study, and having an impact on virtually every aspect of society. The tradition of cross-disciplinary work at Cornell, particularly among the fields both at the core of data science methodologically, and in many of the first key application domains to be influenced by this revolution, has placed Cornell in a position of leadership. The creation of a Center for Engaged Data Science would play a critical role in maintaining and building upon this position, by energizing the recruitment of faculty at the senior and mid-career levels with both targeted additional resources, and by establishing a collaborative environment to house a cadre of junior researchers in data science that span all Cornell campuses and areas of study. The proposed Center would coordinate research activities that foster new collaborations by extending the span of data science to new domains, both by research summits in targeted interdisciplinary areas, as well as with resources to facilitate faculty throughout Cornell to integrate data science methods into their ongoing research. The Center would be a portal for both the infrastructure to support and promote the archiving of data resources at the foundation of many applications, and the computational power and expertise to apply these innovative approaches at scale. The Center would both coordinate and support the development of educational programs in data science at all levels, but most importantly at the doctoral level, where it would be the home for a new graduate field, and at the introductory level, so that a common springboard is available to every undergraduate to study their chosen discipline from a data-centric perspective, including the possibility of minoring in data science. Novel application domains for data science will emerge from industry as well, and the Center will provide a mechanism for encouraging partnerships that capitalize on these new opportunities. Most importantly, such a Center for Engaged Data Science would ensure that Cornell's role in leading the data science revolution worldwide continues for the next generation of advances.

Appendix

Data Science Task Force Charge:

1. What organization, structure, leadership or mechanisms will create a robust and inclusive academic research environment for data science that facilitates recruitment and success of the most promising scholars in the emerging discipline, effectively connects faculty working across the spectrum of discovery through application in data science, enhances the potential for external funding, and provides an effective platform for educational efforts? For example, should we have a Graduate Field of Data Science? Should there be a Data Science Institute? If so, how should it be funded? Should we consider changes in our current department structure?

2. Cornell has a number of units that already engage with and support data science across the university, including the Cornell Center for Advanced Computing (CAC), the Bioinformatics Facility within the Institute of Biotechnology, the Research Data Management Service Group (RDMSG), the Cornell Statistical Consulting Unit (CSCU), the Cornell Institute for Social and Economic Research (CISER), and the Survey Research Institute (SRI), among others. In this context, do we have the right organizational structure and capabilities to achieve the goals set forth above? Are support systems appropriate for the research communities seeking to access data analysis?

3. How can we advance data science educational programs at the undergraduate, masters, and Ph.D. levels in the most effective way?

4. Should Cornell focus on building specific areas of data science in which we can most easily achieve a competitive advantage relative to our peers due to existing strengths at the university? If so, which areas are most promising?

5. Are there ways to organize our recruitment efforts so as to enhance the interactions and collaborations of faculty that are recruited to Cornell in the discovery and application domains?

Data Science Task Force Membership:

NAME	DEPARTMENT
Bean, Rachel	Astronomy
Brachman, Ron	Jacobs Technion-Cornell Institute
Clark, Andrew	Molecular Biology and Genetics
Cornwell, Ben	Sociology
Easley, David	Economics
Enns, Peter	Government
Gore, Michael	Integrative Plant Sciences
Hooker, Giles	Biological Statistics and Computational Biology
Joachims, Thorsten	Computer Science and Information Science
Kleinberg, Jon	Computer Science and Information Science
Liaukonyte, Jura	Applied Economics and Management
Liivak, Oskar	Law
Linster, Christiane	Neurobiology and Behavior
Osofsky, Steven	Population Medicine and Diagnostic Sciences
Pathak, Jyotishman	Weill Department of Medicine
Shmoys, David	Operations Research and Information Engineering
Wagner, Aaron	Electrical and Computer Engineering
Weinstein, Harel	Weill Department of Medicine
Wells, Marty	Biological Statistics and Computational Biology
Wittich, Peter	Physics