Integrative Graduate Education and Research
Traineeship Program in Data Science

I. Summary

Increasingly, professional and research practices, perhaps simply “creative” practices, depend on data and data processing, on drawing conclusions from or in some way adapting to complex quantitative observations of the physical world. And yet, no single discipline can claim ownership of the entire “pipeline” from data collection technologies, to transmission and storage, to visual analysis, modeling and decision-making. Within UCLA, researchers in many departments are faced with a shifting terrain of evidence-based or data-based methodologies. The faculty members listed in Section IV provide a sense of the diversity of fields involved. In response, many departments are developing hybrid training programs that prepare students both in their primary disciplines as well as in “analytics” or “informatics.” Even within traditionally quantitatively oriented areas like computer science, engineering, mathematics and statistics, there is a growing recognition that students need a broader, interdisciplinary view of data, models, computation and inference. Rather than separately augment each program, the proposed IGERT will cull expertise from across campus, engaging different disciplines and exposing students to a larger science of data. Through collaborative courses, workshops and unique summer research experiences, our curriculum will be organized around the “flow” of data from collection outward, and will probe the complex feedback loops that exist between data methods and technologies developed in different disciplines that tap into this flow. In short, we propose to train a new kind of professional, and, ultimately, a new degree through a full-fledged inter-departmental program in data science.

II. Motivation

Almost every aspect of our lives is “rendered” in data. New data collection technologies have made it easy to record continuous, high-resolution measurements of our physical environment (weather patterns, seismic events, the human genome). We are also constantly monitoring our movements through and interactions with our physical surroundings (automobile and air traffic, large-scale land use, advanced manufacturing facilities). In computer-mediated settings, our activities either depend crucially on or consist entirely of complex digital data (networked games, peer-to-peer technologies, Web site and Internet usage).

As a result, the flow of data has become an important force in contemporary life.

Last quarter the Statistics Department began a two-year project, a graduate seminar, to track the flow of data through specific sites in the greater Los Angeles metropolitan area and to study how data collection, processing and analysis shape our experience of our local community and its institutions.

- Through interviews with doctors, nurses and administrators at the UCLA Medical Center, we examined the adoption of electronic medical records, both in terms of the complications related to database coordination and end-user visualization and analysis requirements, as well as the radical transformation that such a system represents for the delivery of healthcare.

- At the Los Angeles Unified School District, we found that federally mandated student assessments have spawned new flows of data, new periodic assessments, which are beginning to restrict curriculum design and are fueling new discussions of accountability.

- In a guest lecture at UCLA last quarter, the Chief Scientist at Disney, a civil engineer by training, examined both the variety of data and the patterns of data exchange that support construction projects in the park, and presented his ideas for creating a new organizational chart (a new reporting structure for his staff) to streamline data flow.

- The data from a network of 15,000 sensors embedded in the 1,700 miles of freeways that cut across Los Angeles are used by Caltrans to help characterize and ultimately regulate the flow of automobiles throughout the system via adaptive ramp metering (controlling access) and targeted text messages
(large displays reporting incident alerts and travel time estimates); the interplay between autonomous control and centrally-directed responses is the subject of ongoing experimentation.

- At the Frank O. Gehry Partners, students learned that architectural (digital) models are becoming powerful carriers of information, and that this concentration of data and the format in which it is expressed, are restoring the architect’s role as a “master builder,” the primary coordinator for more and more of the process leading from design to construction.

- In addition to these sites, we have started to discuss the role of data and data processing at the National Parks Service in the Santa Monica Mountains, focusing particular attention on so-called adaptive environmental management strategies; we have scheduled interviews of the Chairs of the Los Angeles Democratic and Republican Committees, and plan to examine the effects of recent state legislation mandating an “accessible voter verified paper audit trail” for all electronic voting systems as of June 1, 2006; and, finally, we will consider “points of entry” by examining flows of data that regulate the movement of people and cargo at Los Angeles International Airport and at the Port of Los Angeles.

While our project has only just begun, a number of themes have emerged related to the collection, transmission (dissemination), analysis and “preservation” of data. At the source of the flow, what are the technical and social implications of data collection, literally the translation of objects and phenomena in the physical world into bits? Is there, or should there be, a centralized authority that is charged with data collection and dissemination or are data produced and organized in a more organic fashion? What technologies might promote the sharing of data in “safe” ways? What are the competing models (deterministic mathematical, probabilistic, or data-based representations) for objects and phenomena and how do they organize or “expose” the information they carry? If the flow consists of large quantities of complex, dynamic data, how do “subscribers” understand or express patterns or regularities, and how do these forms of expression affect their view of the phenomena being described? How are decisions made or, abstractly, how are optimization problems solved, when the underlying data are noisy or uncertain? Who has access to the data or views of the data, and at what resolution; and what is the role of legislation in setting these limits? How are data or derived analyses presented to the general public for decision making? Addressing these questions in a meaningful way obviously requires an interdisciplinary approach.

In addition to the sites listed above, our IGERT draws inspiration from the interdisciplinary research projects taking place at the Center for Embedded Networked Sensing. Within the CENS, teams of environmental scientists, engineers, computer scientists and statisticians are working together to design the next generation of earth observing systems; the interactions between faculty, graduate and undergraduate students from these different disciplines have spawned genuinely new research questions in each of the represented fields. In some cases, science questions have led to the development of new engineering systems; while in others the opposite is true, with new observational technologies suggesting unimagined science experiments. While the complexity of data coming from these systems often requires new analysis techniques; we have also had statistical considerations of “what to measure and when” lead to new control or actuation algorithms for the observing platforms. While students in CENS graduate with a diverse skill set, the proposed IGERT formalizes this experience, greatly expanding the curriculum thematically and providing much deeper basic training in data science.

III. Organization and Budget

The proposed IGERT will create unique research and traineeship opportunities for students and faculty from both disciplines that have traditionally contributed to the flow of data (specifically, statistics, biostatistics, mathematics, computer science and engineering), as well as disciplines that are now asked to make greater use of data and computational methodologies. The involvement of departments like information studies will add a “critical” component to the curriculum, challenging participants to question the form and function of technologies at each point in the flow. For the students and faculty from fields like sociology, anthropology, architecture and history, we are not only providing end-to-end training in data science, but we are also creating a unique opportunity for new collaborations to form. This application-
focus will undoubtedly produce new research in each of the fields involved. As a result, we feel that this kind of program is a perfect match for the IGERT CFP.

The program is intended to catalyze a cultural change in graduate education, for students, faculty, and institutions, by establishing innovative new models for graduate education and training in a fertile environment for collaborative research that transcends traditional disciplinary boundaries.” (IGERT CFP, p. 3)

The proposed IGERT will target students in the first or second years of their doctoral programs and will provide fellowships for two years. The NSF budget cap on IGERTS could fund at most 20 UCLA graduate students each year; and, in crafting a formal budget our emphasis will be on maximizing the number of graduate fellowships. We will fund curriculum development, seminars and other activities with matching grants and department releases. We anticipate a fairly even split in fellowships allocated to primarily data-centered disciplines like mathematics, statistics and computer science, and to more applied (from the perspective of data science) disciplines like geography, anthropology and sociology. Through the IGERT itself we will offer a series of primarily team-taught graduate seminars in different aspects of data flow: Sampling and data collection could be taught by a biostatistician, an engineer and a sociologist; database technologies could draw from both information studies and computer science; visualization might involve a geographer (GIS), a statistician and a designer; and inference could combine perspectives from a numerical analyst and a decision scientist. The interplay between these points of view will provide students with powerful new ways of thinking about and reasoning with data. In addition to these seminar courses, students will be offered a curriculum of (possibly specialized) classes from departments like statistics, computer science, engineering, and information studies. These would provide basic introductory material on concepts related to data science.

If this (pre)pre-proposal is accepted, we will form a working group including the faculty listed in Section IV to identify Co-PI's. Hansen will serve as PI and, together with this working group, will organize the writing of a formal proposal, taking input from the different centers, institutes and departments involved. In terms of institutional support, the IGERT will need a meeting space on campus, with flexible workspace for up to 20 students. While we anticipate that most graduate student participants in the program will have offices in their home departments, we need to provide common space for IGERT participants to encourage collaboration. We will secure matching funds through grants or private contributions to provide collaboration tools and computing equipment for this workspace.

We will leverage the existing resources of IDRE, IPAM and other relevant institutes on campus to create a series of lectures and workshops related to data science. These are aimed at the larger UCLA community and will highlight the importance of data science in research and professional practices. We will also secure matching funds to host visitors and guest lecturers. We have already started to identify industry support for this effort (Google), and hope to offer unique summer research opportunities for students in the program. Our brief experience with the graduate seminar described at the beginning of this (pre)pre-proposal has put us in contact with groups in the greater Los Angeles metropolitan area that have an interest in hiring students with the training we are proposing. Representatives from these and other local groups will help form an advisory board for the IGERT that could be called on for internships, lectures and even the identification of private funding sources.

Finally, we expect that participants in the proposed IGERT will help (re)design undergraduate courses and summer research programs engaging topics related to data science. After initial discussions with engineering, for example, we anticipate introducing ideas from statistical sampling into systems design courses. We expect that this kind of crossbreeding will take place across many departments once the relevant faculty members have started collaborating. In addition, summer research opportunities could be coordinated with new and existing REUs and we are considering a mentorship program pairing IGERT graduate participants with undergraduates. One possible realization of this would involve community-based projects, suggested perhaps by local advisory board members or other City, County or State organizations. At this point, we leave the precise details of our undergraduate offering to the working group and the PI's.
IV. Institutional Support

The Data Science IGERT depends on participation from faculty and departments directly involved in the collection, storage, transmission and analysis of data, as well as disciplines that draw on data as part of their research or creative practice. We have only just begun to assess interest across campus, and the following faculty, departments and institutes have agreed to be involved in the formulation of the IGERT’s mission and its organization.

UCLA Faculty

- Jeff Brantingham (Anthropology)
- Dick Berk (Statistics)
- Andrea Bertozzi (Mathematics)
- Christine Borgman (Information Studies)
- Jason Cong (Chair, Computer Science)
- Dana Cuff (Architecture)
- Adnan Darwiche (Computer Science)
- Jan DeLeeuw (Chair, Statistics)
- Deborah Estrin (CENS, Computer Science)
- Thomas Gillespie (Geography)
- Anne Gilliland-Swateland (Chair, Information Studies)
- Mark Green (Mathematics)
- Alex Hall (Atmospheric Sciences)
- Mark Hansen (Statistics)
- Katherine Hayles (English)
- Kayo Ide (Atmospheric Sciences)
- William Kaiser (Electrical Engineering)
- Jerry Kang (Law)
- Alan Laub (Electrical Engineering, Mathematics)
- Casey Reas (Design)
- Jan Reiff (History)
- Neal Richman (Urban Planning)
- Phil Rundel (Ecology and Evolutionary Biology)
- Chiara Sabatti (Statistics)
- Ali Sayed (Chair, Electrical Engineering)
- Michael Shin (Geography)
- Mani Srivastava (Electrical Engineering)
- Victoria Vesna (Design)

UCLA Departments

- Statistics
- Information Studies
- Biostatistics
- Computer Science
- Electrical Engineering

UCLA Institutes

- Institute for Digital Research and Education
- Institute for Pure and Applied Mathematics
- Center for Information as Evidence
- Center for Embedded Networked Sensing
- Center for Statistical Computing
V. Letters of support

We augment the list of names in Section IV with a few of the enthusiastic emails and letters we have received supporting the proposed IGERT for “Data Science.” In addition to a strong commitment on the part of the CENS, the heads of both IDRE and IPAM feel this proposal is “exceptionally innovative” (Green) and are excited to participate.

From: "Laub, Alan" <laub@ats.ucla.edu>
Date: March 7, 2006 6:19:17 PM PST
To: "Mark Hansen" <cocteau@stat.ucla.edu>
Subject: IGERT in Data Science

Hi Mark,

I've read through your IGERT preproposal and I'm very excited. As you know, I've just finished submitting a proposal to the DOE on a UCLA SciDAC Institute for High-Performance Parallel Scientific Computing. We discuss funding for about half the number of graduate students as you have, and you're right: ultimately, it's all about the data, both data in and data out. We will, of course, be interested especially in parallel data movement but it will all be relevant to what you are doing. And even if we don't get funded, the same remarks apply! So count us in, and count us fascinated in effecting the same cultural change in graduate education.

With best regards,

Alan
Director, IDRE

Alan J. Laub
310-825-4245
laub@ucla.edu
Mark

The IGERT proposal is exciting. There is an increasing awareness in statistics that the core of the discipline is data science, and not modeling or inference. Moreover the data science perspective is a great way to get into different cooperative projects with many different disciplines.

In fact, it seems interesting to use data science as a possible organizing principle to construct one or more GE courses, or even as a center of the UCLA Quantitative Requirement.

For all these reasons, the IGERT proposal is a bold and original step of great importance to statistics. Let me know any way I can help.

-- Jan
Professor Mark Hansen
Department of Statistics
UCLA

March 7, 2008

Dear Mark,

I am writing to support the Data Science IGERT proposal that you are in the process of submitting. I consider this proposal to be exceptionally innovative. I believe that you have the skills of a program to train a new type of professional, one that is going to be needed in lots of fields as a specialist who understands the issues involved in collecting, analyzing, storing and retrieving the mix of data appropriate to that field. It would be extremely valuable to IPAM to have such an entity at UCLA. I also think it would give a very interesting direction to the Statistics Department that would allow it to develop in a unique way and offer a program that no other place will be offering.

The contacts and expertise that you bring to the table constitute a one-of-a-kind resource, starting with your joint appointment in Statistics and Design, and your amazing range of interests. It is hard to imagine anyone else on campus who could pull this off, but your background is perfect for it.

It is quite unusual to see a proposal that is science-based and yet involves the North Campus in such interesting ways. I regard this as an especially interesting feature of the proposal.

I anticipate that if this proposal is successful, there will be numerous ways in which IPAM can be of use. I would foresee many opportunities for joint workshops, and indeed this would result in immeasurably strengthening IPAM's connections with the Statistics community.

Best regards,

Mark L. Green, Director
Institute for Pure and Applied Mathematics