The Institute for Quantitative Social Science at Harvard University

The Institute for Quantitative Social Science

The Big Picture: A New Collaborative Landscape for Social Science

For Harvard to remain at the cutting edge of social science research, we must recognize, and continue to prepare for, an underappreciated but historic change beginning to sweep the field: enormous quantities of highly informative data are inundating almost every area we study. In the last half-century, the information base of social science research has primarily come from three sources: survey research, end of period government statistics, and one-off studies of

particular people, places, or events. In the next half-century, these sources will still be used and improved, but the number and diversity of other sources of information are increasing exponentially, and are already many orders of magnitude more informative than ever before. We attach a recent two-page article in *Science* that summarizes some of these remarkable developments (Gary King. "Ensuring the Data Rich Future of the Social Sciences." *Science* 331 (2011): 719-721, copy at http://j.mp/mw64M8).

The impact of quantitative social science on the outside world in the last decade has been unprecedented and is growing fast. In fact, what areas of university research come anywhere near the impact quantitative social science (and related technologies, methodologies, and data) have had on the world? These approaches have remade most Fortune 500 companies; established whole new industries; changed much of the rest of the business world; led to the largest increase in the expressive capacity of the human race in history; and reinvented medicine, friendship networks, political campaigns, public health, legal analysis, policing, economics, sports, public policy, and program evaluation, among many others. The social sciences are getting to the point where enough information, infrastructure, methods, and theories may finally be available to understand and ameliorate some of the most important, but previously intractable, problems that affect human societies. In the last few months, even the media has begun to recognize these developments under the banners of "big data," "data science," and "data analytics."

The consequences of these amazing developments for the day-to-day lives of social science faculty and students are substantial and growing. Whereas social scientists once worked on their own, alone in their offices, many are now working on much more highly collaborative, interdisciplinary, lab-style research. The knowledge and skills necessary to access and use these new data sources do not exist within any one of the traditionally defined social science disciplines. Through collaboration across fields, however, we can begin to address the interdisciplinary substantive knowledge needed, along with the engineering, computational, ethical, and informatics challenges before us.

A promising side effect of this change in research style is that the most significant division within the field, that between quantitative and qualitative researchers, is showing signs of breaking down. Social scientists from both traditions are working together now more than ever before, because many of the new data sources meaningfully represent the focus and interests of both groups. The information collected by qualitative researchers in field notes, video, audio, unstructured text from archival collections, and many other sources is now being recognized as valuable and actionable data sources, for which new quantitative approaches can be applied. At the same time, quantitative researchers are realizing that their approaches can be viewed or adapted to assist, rather than replace, the deep knowledge of qualitative researchers, and they have taken up the challenge of providing added value to these new richer types of data. The divergent interests of the two camps also converge at the need for tools to cope with, organize, preserve, and share this onslaught of data, the search for new understandings of where meaning exists in the world and how it can be represented systematically, and the rise of inherently collaborative projects where each researcher brings his or her own knowledge and skills to attack common goals. We aim to lead and help nurture

this important development since it has the potential to greatly strengthen the research output of social science as a whole.

A central goal of IQSS is to keep Harvard among the leaders of the social sciences internationally by anticipating and responding to the unprecedented challenges these new data sources pose for our field. Outfitting every social science faculty member with his or her own lab—on the scale of those in, say, chemistry or biology (where startup funds for a single senior faculty might include several million dollars of startup money, 3,500 square feet of lab space, and a dozen employees)—is obviously unrealistic in the near future, but we also aim to make it unnecessary; our goal is not to replicate the physical and natural science model within the social sciences. Instead, our goal is a far more efficient approach that involves building *common infrastructure* to solve problems across the labs and research programs of the diverse interests of Harvard's faculty and students. This approach is in line with recent research, which shows that the most common way for younger scientists to collaborate is through the sharing of resources, such as data, tools, and practices.[1] Our common infrastructure takes many forms, with new forms arising frequently. As technology changes, we adapt IQSS so we can seek out new opportunities to make a difference. In the last few years, we have built large open source computer programs, started new seminar series, run international conferences, brought together scholars from disciplines who have rarely collaborated before, taken over and built quasi-research projects that make the Harvard administration more efficient, started new programs, closed down completed programs, spawned commercial and nonprofit startups, educated students and faculty in new technologies, data, methods, and theories, and led many other activities.

We are especially gratified by the number of other universities who are creating centers they describe as emulating IQSS. For the last several years, we have been contacted by representatives of more than more than 20 other schools seeking advice on how to get an IQSS-like center up and running. Many have followed up with personal visits to Harvard; we've even developed a standard protocol for these visitors of which groups they need to meet, what categories of advice we offer, and the different ways we can collaborate and support each other going forward.

The Beginning: Administrative Unification

Once we received support from the administration, we began to build IQSS by first administratively unifying several distinct entities. With one of the most highly decentralized governance structures there is, this was a major political achievement for Harvard, and ultimately was to the benefit of everyone involved. These units included the Harvard-MIT Data Center (HMDC, a venerable Harvard institution providing fee-based services to MIT, with roots here dating to the 1960s), the Center for Basic Research in the Social Sciences (a new center, now defunct, which has become the administrative core of IQSS), the Murray Archive (the endowment of which funds the Dataverse Network® and other archiving

[1] Aghakhani et al., "Emerging collaborative services in Personal Digital Libraries," Computer Sciences and Convergence Information Technology (ICCIT), 2010 5th International Conference on Computer Sciences and Convergence Information Technology, vol., no., pp.487-491, 2010).

operations), and the Program on Political Economy. At the same time, we also helped to establish, under IQSS, the Center for Geographic Analysis.

Progress in Building a University-Wide Social Science Community through Research Technology and Support

In the five years since IQSS was founded, we built a thriving research community around a constantly changing and improving set of seminars, conferences, professional services, research projects, and community-building activities. IQSS has implemented a programmatic model that brings together researchers from across the disciplines for intellectual collaboration and debate, and also provides these researchers with tools and services that make them more efficient.

Our overall operational plan: researchers are first attracted to IQSS because of specific services we provide that make their research better or more efficient—seminars, conferences, grant administration, financial and HR management, strategic planning advice, high performance computing, technology consulting, original software, etc. With our assistance, they get their work done better and faster. While they are receiving these services, they cross paths with other researchers often from apparently distant areas, find collaborative opportunities, and as a result make substantial contributions to building our research community. Individual scholars do not always know their important contribution to the collective, but we are all much better as a result. The result is that the community here is flourishing and now filled with social scientists from disciplines representing the different departments and schools at Harvard and well beyond. Faculty and students come to IQSS because they can do their research more efficiently here; they stay for the community and for contacts with other researchers.

Below is a snapshot of a few of the more recent successes we have had:

- We have built a **strong and diverse community**, with over 275 active affiliates including faculty (nearly 90), students (over 100), visitors (more than 40), and staff from around the University. (These are real numbers; unlike other centers, we do not let researchers apply to be affiliates; we list them if they actively and substantially participate in the life of the Institute, and our staff happens to notice them. We remove them if they do not participate; thousands more at Harvard not on this list, and numerous others outside of Harvard, benefit from our services.)
- We have **expanded the tools and resources offered** and made them available to all users, regardless of school, and where appropriate outside of Harvard. Our computing cluster, the Research Computing Environment (RCE), for example, now serves 463 active researchers who hail from 8 different Harvard schools.
- We have **developed several extremely popular new products for researchers, including the OpenScholar® platform,** which how more than 1,000 individual sites set up by faculty and students. This platform has been chosen to serve as the base for the new Central and FAS long-term administrative website strategy for academic and administrative departments. Scholars all over Harvard are adopting this at a rapidly growing rate.

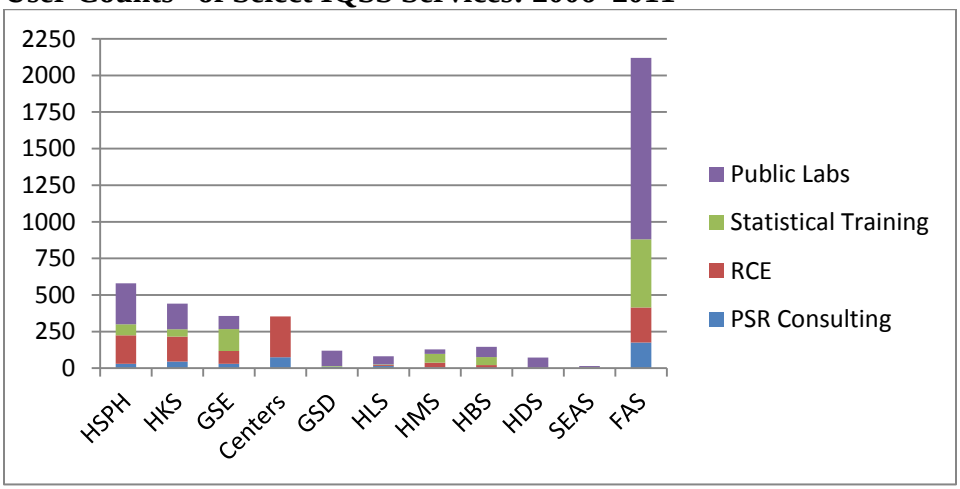
- We have digitized the previously at-risk Murray Research Archive collection, and we have **created a new digital data repository, the IQSS Dataverse Network® (DVN)**, that has expanded the Murray's holdings from its original 300 studies to nearly 40,000 studies with over 650,000 files—now the largest collection of social science data in the world. The Dataverse Network software we developed is open source and has received contributions and been installed by numerous universities around the world. It has also been formally adopted by the Harvard University Library, which will take an active role in operations, promotion, and extensions to the sciences and humanities.
- We **serve as the official home for all Harvard Business School sponsored awards**, managing 6 grants for 5 PIs, which will bring in approximately $1.8M in incremental research funding that would have otherwise not come to Harvard.[2] We also help make the sponsored research application process seamless for numerous social scientists in FAS and other schools who do not have access to high quality support. Many new collaborations have taken place because of the new interactions among researchers this activity has created.
- We work hard to arrange **collaboration with groups around campus,** including both large organizations, like HUIT (Harvard University Information Technology) and the Harvard University Library, and also more discrete entities, such as the Center for Population and Development Studies and the Harvard-Smithsonian Center for Astrophysics. These are just a few of the many collaborations we have established.
- We helped to **initiate, and now serve as the administrative home** for, a variety of specific initiatives, such as the Center for Geographic Analysis, Education Innovation Laboratory (EdLabs), and the Robert Wood Johnson's Scholars in Health Policy program.
- We have **helped faculty launch new research programs benefiting everyone** under the IQSS umbrella, including the Program on Survey Research, the Program on Text Analysis, Qualitative Social Science@Harvard, the Program on Experience Based Learning in the Social Sciences (PEBLSS), among others
- We **sponsor conferences, workshops, and training events** that involve more than 750 of our faculty and students annually. Our events range from graduate student-only luncheons to large-scale conferences that include participants from around the world.
- We **provide faculty with undergraduate research support**, each year offering nearly 45 undergrads and 20 faculty members the opportunity to work together and learn from each other. The connections among these undergraduates, and between the undergraduates and faculty, are clearly as important as specific purposes for which the funding is granted.
- We **offer more than $50,000 annually for travel and research grants to over 30 graduate students** from the FAS, HKS, and HSPH.

[2] The Harvard Business School does not have sponsored research capabilities on staff. Before IQSS made its services available, faculty interested in sponsored funding either funneled their proposals through a third party or most often opted not to apply.

Cross-School Impact: Expanding Our Boundaries

IQSS and social science often cross school boundaries. A quick look at four of the IQSS services that have been expanded across Harvard (Public Computing Labs, Statistical Training, Research Computing Environment, and PSR Consulting) shows that the usage rate is nearly equally divided between FAS and non-FAS schools. Over the past five years, in these categories, IQSS has supported 2,474 FAS and research center users (or 56%) and 1,941 users (44%) from other Harvard schools.

User Counts* of Select IQSS Services: 2006–2011



*Users include account holders for Public Labs and RCE, individuals attending Statistical Training classes, and individuals asking for consulting services with PSR.

Formal Mission: Enhancing Social Science Research

IQSS is a University-wide institute located physically and administratively within FAS. We facilitate the creation, dissemination, and preservation of scientific knowledge about human society, its problems, and their solutions. We support these aims through our highly collaborative environment and the scalable infrastructure we are continually building. IQSS also offers many related services to support faculty, students, staff, and programs across a wide variety of disciplines in all of Harvard's schools. Our scientific mission is: (1) to create, and make widely accessible, statistical, computational, and analytical tools for the social sciences; and (2) to use these tools for understanding and solving major problems that affect society and the well-being of human populations. The organizational mission is: (1) to foster interdisciplinary, often large-scale, and highly collaborative projects that cannot readily be accomplished in isolation within the traditional setting of individual departments; and (2) to build a scientific culture where faculty, students, and staff work side by side, not only to solve their own problems in their own disciplines, but also to seek out problems in unrelated or applied areas amenable to the same approach.[3]

[3] We use the term "social science" to refer to areas of scholarship dedicated to understanding, or improving the well-being of, human populations. "Quantitative" refers to statistical, computational, or mathematical methods. Social scientists typically conduct quantitative analyses using data observed at the level of the person or groups

IQSS is an unusual *hybrid organization*, both a research center and an integral part of the Harvard administration. We direct and facilitate large, interdisciplinary research projects ourselves; build infrastructure that facilitates research of students and faculty; incubate, administer and host research groups and technology centers; and administer professional staff and IT tools that increase the productivity of many others around the University. We also combine roles, most obviously when we take routine activities of the administration, turn those into research projects, automate tasks, and greatly extend the impact, efficiency, creativity, and productivity of the effort.

We specialize in *infrastructure that scales*, so that spending a dollar on IQSS positively impacts more faculty and students than would normally be the case. We are able to do this by our focus on research computing infrastructure that is naturally amenable to use by large numbers; by our day-to-day emphasis on creating synergies among the different parts of the Institute, with the help of faculty from all over the University who interact here; and by marshaling the efforts of the open source communities in contributing software and other assistance from inside and outside of Harvard.

The Harvard-MIT Data Center (HMDC) is a part of IQSS, but we have kept its identity separate for reasons related to our contract in aiding MIT, among others. HMDC is responsible for providing technical infrastructure, tools, and support both for our own research and development as well as that of social science researchers across the University. HMDC specializes in research computing technology and is always looking to harness the newest resources for the scientific community. For one recent example, we are teaming up with the CIO and the SEAS Director of Academic Computing to design and build a new tool that will enable automated spillover from each of our internal Harvard clusters to commercial clusters, such as Amazon's elastic computing. This should save a great deal of money in direct funds for hardware, software, personnel, electricity for cooling, and physical resources. The HMDC team is also continuing to shed commodity services as our technology evolves from research projects to commodity items. We are shifting functions—including email and mailing list hosting, departmental network storage and standard website hosting—to the central HUIT team. This process allows HMDC to stay at the cutting edge of research and development; to help faculty and students analyze data in new resources, classes, and customized support; and to build new products in new areas.

---

of persons, such as countries or areas. The term is most commonly applied to empirical and quantitative areas within academic disciplines in the Faculty of Arts and Sciences, such as Sociology, Political Science (called "Government" at Harvard), Economics, Psychology, and Anthropology. The term is also used for quantitative analyses of public policy at the Kennedy School and educational research within the Graduate School of Education. Social science is called other things in other areas but the category is much wider than the term. It includes what Law School faculty call "empirical research," and many aspects of research at the Medical and Business schools. It also includes a large fraction of faculty from the School of Public Health, although they have different names for these activities such as epidemiology, demography, and outcomes research. IQSS also has rich connections with the School of Engineering and Applied Sciences, the Harvard Initiative for Global Health, and the Broad Institute (along with many more substantively oriented centers), including joint grant proposals and collaborative research.

Goals and Priorities: Building Communities through Needed Tools, Services, and Resources

Our main goal is to continue building multi- and inter-disciplinary communities with common research interests, which benefit from using the same services and tools, sharing resources and having a place to interact. We accomplish this with the combination of:

- Administrative and Technology Services – Providing sponsored research administration, research technology consulting, user support, computer labs, training, events organization, etc.
- Product Development – Building statistical tools and large software applications that improve the research cycle and increase scholarly recognition.
- Programs and Centers – Being home of a number of affiliated research centers and programs that foster new social science research and collaborations.

Our institute could be defined as either a center that focuses on social science research and extends its methodologies and research products to other disciplines, or a center that focuses on multi-disciplinary computational and data science using social science expertise. In fact, IQSS is both. These two synergetic approaches result into a broad set of goals and future for the Institute.

We have prioritized our short-term goals (1-3 years) based on high demand research needs and opportunities to collaborate with other groups at Harvard. When we find another group that shares a common interest for which we already have a service or a product, we build a collaboration to continue enhancing that service or product. When we determine a research need that can be fulfilled with a shareable, reusable and/or scalable solution, we develop that solution. And when we identify a service or product already fully supported by another Harvard-wide group, we transfer it or merge it in order to reduce cost and have room to focus on unique projects.

Our short-term goals include:

- Build a collaborative data storage solution based on the Dataverse Network and transfer its operational responsibility to the Harvard Library and HUIT, allowing us to focus on developing the next generation tools for data management.
- Continue to grow our preservation efforts, both in data collection and in scholarly research into archival technologies, so that the many new and innovative forms of data can be collected, preserved, and shared.
- Partner with HUIT to create a new web site building platform based on the OpenScholar product and expand the support from faculty to departments, while we continue working on new tools for scholarly recognition.
- Establish a permanent research technology consulting services team to support the increasing community of data-driven social science researchers. We are actively working on connecting our three research consultants (focusing on different types of statistical and programming support) to the three social science preceptors (in mathematics and statistics, geographic information systems, and survey research), our desktop support group, and several other noncredit teaching efforts we host or foster.

- Migrate commodity technology services to HUIT (email, web site hosting, network files) so that we can continue to develop new services.
- Develop an automatic and transparent mechanism on our servers to access computing cycles from other available clusters, internal and external to Harvard.
- Release Consilience, a new text clustering software tool.
- Improve post-award reporting services for faculty.
- Grow graduate and undergraduate programs through greater enrollment and participation in events.
- Decrease the number of physical servers maintained locally and move increasingly to hosting IQSS services on cloud resources.
- By necessity, we have developed an expertise in security (of data and servers), and of privacy. We pushed a standardized five-level protocol for different types of data security, and the University had formally adopted and promulgated it. Going forward, we will continue to serve as a resource for data security and privacy solutions for social sciences.

Identifying long term goals in a field changing so fast is difficult, but our general plans are as follows:

- Expand IQSS's network of partners, within and beyond Harvard, to allow us to continue focusing on innovation. Traversing the political obstacles to collaboration often requires substantial effort, and it is not always fruitful, but when it works the result can be enormously beneficial to all involved.
- Serve as one of the University's primary centers for statistical and other software R&D to spawn new tools and products that improve research, scholarly work, and administration.
- Provide software development services by managing tournaments for creating research tools/problem-solving using development services such as TopCoder.
- Focus on the initial research and development for a tool or a problem, and build reusable and scalable solutions that can be later offered Harvard-wide.
- Expand the community of researchers beyond the social sciences by building a consortium of all the institutes at Harvard that focus on computational science and data science. For example, the Institute for Applied Computational Science (IACS), the Institute for Theory and Computation, statistics groups, etc.
- Be a home for scholars from diverse disciplines interested in business intelligence, data engineering, data science; provide workshops, internships, place for sabbaticals, etc.
- Solve technology problems related to exponential growth in data availability (search, sort, storage, analysis, etc.)
- Help facilitate social science beyond Harvard, by encouraging the open source communities forming around our products and contributing to others, by helping those establishing IQSS-like institutes elsewhere, and building collaborations with related projects elsewhere.

- Continue to influence federal funding priorities to increase their support of social science data and research

In order to complete our projects and accomplish our goals, it is critical for us to retain and motivate our staff. In a fast-changing and dynamic field, we need to keep our organization ready to adapt quickly, and to continue to improve our staff's working space, recognition and compensation. To expand or start new research projects, we plan to increase grant support for internal technology research and development (software and infrastructure) when possible.

Activities: Our Services, Products and Affiliated Programs

Administrative Services

IQSS serves as the administrative home for a variety of PIs, programs and even a University-wide center. In the past four years, the number of groups for which we provide services has grown 250%, from 7 to 19. Some of this growth has been deliberate as we build the community; some has been at the request of the FAS or Central Administration. The administrative and technical incubation we provide has led to a number of important advances.

Programmatically, a critical goal of our administrative services is to build a community of scholars. We find that these administrative services are a tremendous help in that regard: quite frequently, scholars come for the services and stay for the interactions with other scholars. We have shown this model to work many times in different areas and different ways, but it is almost always successful at creating more synergistic value beyond the services as listed. We are thus far more open to building administrative services upon request, as long as the management overhead can be handled relatively easily.

On funding, IQSS has directly and indirectly brought in a large amount of grant support, much of which would not have come to Harvard otherwise. We have also saved the faculty a great deal of money, by the services we have developed. To take one example: to hire someone to build a unique web site for a faculty member can cost $5,000 to $25,000 per site; the more than 1,000 diverse-looking and operating OpenScholar sites, supported by a single installation of our software, has conservatively saved the University millions of dollars. We have received some donations and gifts, but we would like the opportunity to take development efforts to the next level. The pitch, in terms of the impact of quantitative social science on the world, is ready, and we have made good connections with some highly capable donors. With the University's permission, we hope to be able to secure some larger gifts. IQSS has been one of the few research centers specifically mentioned in the Dean's capital planning proposal, and so we are optimistic. While we continue to pursue various methods of developing new revenue sources, it is important to remember that IQSS was not intended to become self-sustainable. IQSS is a service center, a research center, a builder and operator of University infrastructure, and a part of the University administration.

Sponsored Research Administration. IQSS provides grant support to researchers in the social sciences. We provide both pre- and post-award support to a growing number of faculty from across the University. Our services include helping prepare submission materials and navigating them through the Office of Sponsored Programs process; processing and tracking expenses during a project; hiring and managing research-related staff; ensuring post-award reporting; etc. We also connect researchers to the technologies required to move their projects forward, providing consulting assistance, collaborating on new tools, and introducing faculty to other resources from across the University and beyond.

We now manage 37 active grants with 65 unique accounts (excluding EdLabs). An additional 25-35 proposals are submitted annually. Our grant support is non-exclusive and extends beyond the FAS to include faculty in HBS, HMS, SPH and GSE. This expansion of grant support not only follows the Provost's goal of increasing sponsored research across the schools, but it also brings in significant overhead and research dollars. In FY11, IQSS generated $297K in overhead from non-FAS PIs, and that figure is expected to jump nearly 70% to $500K in FY12. Indirect cost recovery goes to the FAS and not to IQSS. With the exception of CGA, which pays a modest 15% of the Director of Finance's salary regardless of activity volume, our standard policy is not to charge PIs for our services. This model has been reinforced by both FAS and the Center when asking us to take on additional groups, such as EdLabs and the RWJ program. In those cases where a PI's portfolio might be significant, we ask that s/he use grant funding to support a local administrator to reduce the burden on our team.

Total sponsored research funding managed by IQSS for faculty across Harvard (but excluding EdLabs) was $4.7M in FY11 and is projected to be $5.3M in FY12. Much of this funding, including the FAS grants, would have been lost to the University and the PIs without IQSS support due to unavailable pre- and post-award support.

Below are examples of our successful administrative support:

- We helped Sendhil Mullainathan build a fabulous research group (*research42*), and then spawn a separate, non-profit entity.
- Roland Fryer's *Education Innovation Laboratory (EdLabs)* was built under the administrative oversight of IQSS, and has now raised enough funds and developed enough experience that it may soon become its own FAS entity.
- The *Center for Geographic Analysis* has been able to focus on critical software like AfricaMap while IQSS manages its HR and finances.
- The *Robert Wood Johnson Scholars in Health Policy Research* program received a third tranche of funding for another 3-year period.
- The *Qualitative Social Science @ Harvard* project, partially funded by IQSS, has received funding two years running for J-term classes managed by our events coordinator.
- At the request of the HBS administration, IQSS serves as *the sponsored research home for HBS* researchers. Faculty, including Michael Porter, Lee Fleming, and Karim Lakhani, have applied for and received significant federal grants with IQSS assistance.

Technology Services

General Technology Infrastructure

HMDC serves as the technology infrastructure arm of IQSS, providing tools, resources, and support for analyzing data. This includes maintenance and development of the social science computing cluster; management of specialized, research-oriented computer labs for the social sciences; research and development on, and the incubation of, new general technology; training classes for specific, social science software technologies; the creation of specific research-related tools and applications; and dedicated research technology consultants who provide one-on-one and project specific support for research computing. Many of these specific activities are discussed below, but the actual infrastructure that supports all of them should not be forgotten.

Note: Services that began as IQSS research or as quasi-research, infrastructure-building projects, that now can be obtained as commodities (such as web, file, email and listserv hosting) are being gradually migrated over to the new HUIT organization; where there is overlap with research activities (file hosting, for example, related to the computing cluster or the Dataverse Network archive), the responsibility will remain with IQSS.

Key facts and figures about HMDC infrastructure:

- Leveraging technological advances to take advantage of increased density has allowed for a reduction in total servers while growing computing cores.
    - 128 total servers, down from a high of 199
    - 1,288 total computing cores, up from 510 in 1997
- Virtualization has enabled multiple services to be run on a smaller number of servers.
    - 174 operating system instances currently being run
    - Usable disk capacity of 251 TB
- 209 TB of network-attached storage
- 42 TB of local server storage
- Active research data (12 TB) includes data for 355 projects
- Archive data (30 TB)
- Personal and departmental data (3 TB) includes management of 1,392 separate file shares, including 1,242 home directories and 88 group shares (to be partially migrated to HUIT)
- 324 TB of data back-ups stored on tape. Real-time mirrors of data and failover servers for key applications are kept in a disaster-recovery site at 60 Oxford Street.

- 57 websites hosted (to be partially migrated to HUIT)
- 1,936 active email accounts with 8 separate domains (migration to HUIT in process)
- 94 active e-mail mailing lists (migrated to HUIT)
- 122 users of RT (user incident management system)

Research Technology Consulting

Starting in late 2011, IQSS established a new Research Technology Consulting program by hiring a team of Ph.D. consultants from a variety of disciplines who provide training and expertise in both quantitative and qualitative analyses. The Research Technology Consulting services include the following:

- Data analysis support and programming services
- Research project planning and guidance selecting appropriate technology for research projects
- Facilitating appropriate organization, storage and sharing of data
- Training on the use of both established software packages and emerging tools

As of March 2011, the Research Technology Consulting team has helped roughly 100 different research projects, offering as many as 60 hours of assistance to a single project. The requests have come from faculty, post-docs, students, staff, and visiting scholars from seven schools, including numerous FAS departments, programs, and centers. These requests have needed assistance in such diverse areas as research design, data security, data inquiry, programmatic scripting, text scraping, database design, statistical support, and using software packages such as R, Stata, SAS, Matlab, atlas.ti, Dedoose, and Python.

Technical Training Classes

We offer half-day statistical software training courses on R, Stata, SAS, numeric data resources, and IQSS/HMDC services more generally. Over the past two years, we have added three new courses on R (R and Statistics, R Programming, and R Graphics) and three new monthly courses on the Dataverse Network, OpenScholar, and RCE. The Stata training includes introductory courses on basic Stata use, data management, regression analysis, and graphics. The R training includes introductory courses on basic R use, statistical analysis, programming, and graphics. In 2006-2007, there were just two courses (Introduction to Stata and Introduction to Numeric Data Resources) as compared to today, when we offer 13 different courses. IQSS also offers a subset of these courses to HMS/HSPH personnel in the Longwood area.

The majority of attendees are graduate students, post-docs, and visiting professors. Attendance has doubled from 130 attendees in FY08 to 265 in FY11. Attendees come from FAS, GSD, HBS, HKS, HMS, and HSPH.

User Support Services

The IQSS User Services team is responsible for providing technical support to more than 550 customers (who use roughly 700 devices) in the CGIS complex and associated wood frame buildings. The team interacts directly with faculty, staff, students, and visitors who have questions, problems, or specific requests regarding technology. The team regularly provides advice to end-users on new technology including mobile devices, printers, software, laptops, and desktop computers. Their service has been characterized as immediate and vital; they are widely praised by users who readily admit that they would be lost without the team's knowledge and responsiveness.

In the past year, the User Services team has responded to 2,644 user support tickets as well as triaging all 4,455 tickets that came into IQSS. They review each ticket and send it to the proper group within IQSS (OpenScholar, Dataverse, RCE, etc.) with response time measured in minutes vs. hours in other support models. The User Services team has expanded their scope and technical depth to inject their friendly, responsive style to support not only desktop services but also front-end research computing system administration.

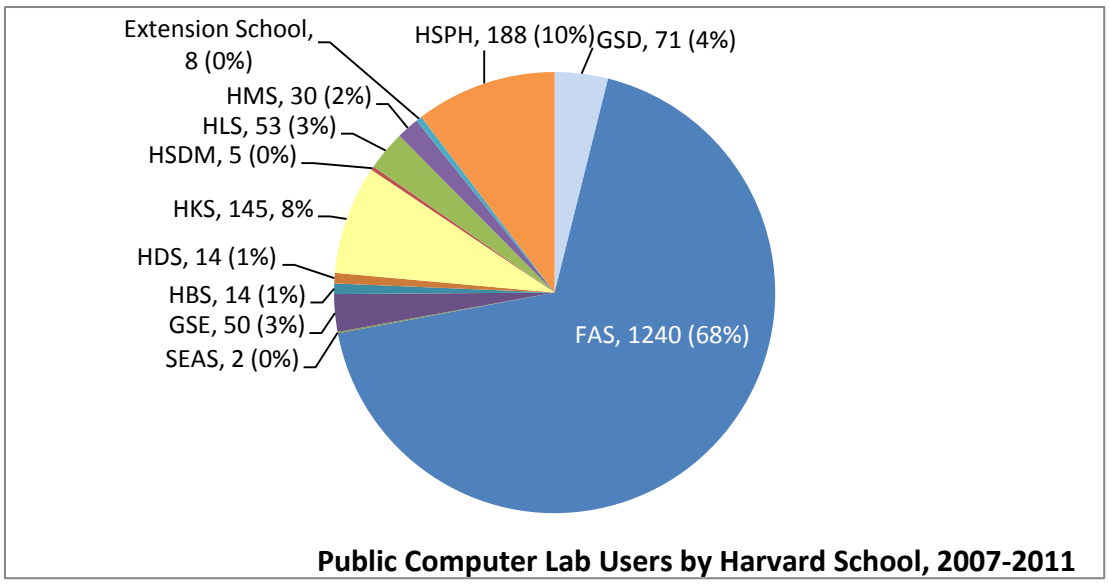
It is important to note that the User Services team has been critical in building connections with the new HUIT team. In particular, they recently helped the HUIT IRIS team migrate hundreds of CGIS user email accounts from IQSS servers to HUIT servers, shepherding users before and after the migration. When senior management began thinking of ways to better integrate the IQSS and HUIT, we discovered that the group had already taken the initiative and forged many of the links we were considering. The team keeps itself up-to-date with the wide range of services offered throughout IQSS and HUIT, and is able to serve as a liaison and route the requests to the right expert very quickly.

The positive feedback that the User Services team receives is remarkable. Of the feedback tickets sent in over the past year, 97% were to thank the team for their work. For example, one administrative staff member wrote, "I have friends (non-HU) who tell tales of cringing when opening tickets and interacting with their (non-HU) IT - this is never the case at the WCFIA. I think i (sic) can safely speak for a majority of my colleagues when i (sic) say that HMDC help is a treasured organization upon which we can always rely." Another commented on how the technician was "fabulous and gave blood, sweat, and tears to help me with my ancient PALM." They routinely receive praise such as, "Quick support"; "Very professional, efficient, helpful and friendly"; "Always a great job"; "Incredibly responsive, diligent and efficient"; "Quick and right on the money"; "Always very nice and very useful. I appreciate their patience and kindness." Just as important, the team teaches while they problem-solve. As one user noted, the team is "always very willing to help and offer advice," while another echoed that he "learned a few new tricks." Anyone who has ever worked in a service industry will tell you not to expect much in the way of real thanks—our users go out of their way to praise the

team because they give that much attention to customer service. As one person put it, "HMDC IT staff responses are fast, right, and friendly." The overall satisfaction with our team is considerably higher than any related team on campus.

Public Computer Labs

We offer public labs in the CGIS buildings, comprising 43 work spaces, including 2 large public labs and a training lab on the concourse level of CGIS Knafel. Our workstations offer the most current versions of all commonly used statistical programs as well as many more specialized tools for the social sciences. Printers and scanners are also available for use. During 2007–2011, 1,993 new visitor accounts were created. The lab is always open and available for use by all Harvard personnel. MIT researchers are also permitted to use the lab as part of the HMDC arrangement.



The labs are used most often by graduate students (many of whom are listed here as FAS, but also come from joint programs with other schools) and the specific breakdown of accounts, by personnel type, is shown below for 2007–2011:

- Graduate Students, 50.9% (996 accounts)
- Undergraduate Students, 30.2% (592 accounts)
- Visitor, 11.0% (215 accounts)
- Staff, 4.8% (93 accounts)
- Faculty, 2.0% (40 accounts)
- Fellow/Post-Doc, 1.1% (21)

Products

OpenScholar

The OpenScholar project started in response to a growing demand by faculty and students at IQSS for individual and personalized websites that allow for quick posting and easy discoverability of publications and related research work. OpenScholar rapidly evolved into a web site building tool for faculty and research groups across Harvard and beyond. Today, the full-featured, open-source website creation tool boasts scalable infrastructure that is easy to administer, with a single installation capable of hosting thousands of personal or project sites. Separate installations of OpenScholar now exist at more than *50* other universities from Princeton and the University of Chicago to Fitchburg State in Massachusetts. We benefit by making the source code available outside of Harvard because others contribute code, suggestions, and bug fixes, while Harvard and IQSS are recognized as the thought and software leader.

OpenScholar has also changed from a research project to a central part of the long-term FAS and Central Administration web strategy. This year, we started a collaboration with HUIT and HPAC to use OpenScholar as the main web site building platform for Harvard sites. As part of this collaboration, OpenScholar will be expanded to support academic and administrative department web sites. Also, HUIT will offer support to departments to help them through their web site building process.

Key facts and figures about OpenScholar:

- 2008 – Development begins
- 2009 – Beta release of scholar.harvard.edu
- 2010 – Beta release of projects.iq.harvard.edu
- 2011 – Production release of both scholar and projects tools

	Scholar sites*	***Project sites***	***Faculty sites***
2010	175	56	30
2011	816	127	~300
2012 (up to April 3)	1012	205	~350

**Includes faculty, graduate student, post-doc and staff sites.*

- More than 2,000 support tickets handled
- Approximately 15,000 publications uploaded
- More than 50 installations in other universities/institutions
- Monthly training sessions offered since summer 2011

Dataverse Network

The Dataverse Network (DVN) began in 2006 (building on previous, related projects of ours), as a research project to help scholars from Harvard and MIT share social science datasets. Through its partnerships with similar groups including ICPSR at Michigan, the Odum Institute at UNC, and many other institutions, *DVN now supports access to more social*

science data than any other repository in the world. An open-source project, with many contributors, DVN enables data publishing, sharing and archiving while also promoting formal data citation. In close coordination with the Harvard University Library and FAS Research Administration Services, DVN also now serves as a standard solution for researchers required to show long-term accessibility to research project data for federal funding and/or journal publication.

The DVN team collaborates with groups both inside and outside of Harvard and across a wide range of disciplines. We have been awarded a Library Labs grant this year to build a query tool within DVN that will allow data owners and repository administrators to learn more about the usage of individual data sets. We worked with both the Center for Geographic Analysis and the Rappaport Institute to develop a Boston Data Portal. The Harvard-Smithsonian Center for Astrophysics has embraced DVN as its data management solution. The Institute for Neurodegenerative Disease at Massachusetts General Hospital is working with DVN to define metadata extensions supporting health and biomedical data; and Stanford's Center for Poverty and Inequality recently collaborated with us on data visualization enhancements.

This year, we also started a collaboration with the Harvard Library to make available the Dataverse to all University disciplines. By this summer, the Library (together with HUIT hosting services) will provide to the Harvard community the IQSS Dataverse Network for social science data and the Harvard-Smithsonian Center Dataverse Network for astronomy data. As part of this collaboration, the Library, together with IQSS, will offer new services to help scholars with data preservation, data management and legal issues, and teach them how to use the Dataverse Network.

The Murray Research Archive is also part of the Dataverse Network project. A large fraction of the data from the original archive is accessible through the Dataverse Network, and the current archive has expanded to include all the social science data deposited in the IQSS Dataverse Network.

Key facts and figures about the Dataverse Network:

- 2006 – Development begins
- 2007 – Beta release
- 2008 – First production release

	2008	2009	2010	2011	2012 (up to Arpil 3)
Dataverses* (total)	225	510	707	999	1172
Dataverses* (Harvard)	28	40	58	78	88
Dataverses* (external)	140	178	222	280	338
Dataverses (in	57	292	427	641	746

progress)					
Studies	10,676	16,895	24,388	41,005	52,271
Total Files Available	197,000	250,000	379,000	671,000	711,781

*A "dataverse" is a virtual archive, a collection of datasets or studies, owned, managed, or curated by a researcher or research group and housed in the Dataverse Network.

Note: The focus of DVN has historically been on quantitative data in the social sciences. Of the 700 faculty at FAS, about 300 are in the social sciences. Of these, we about half-regularly use quantitative data in their research. We are working with groups outside IQSS to expand the Dataverse Network not only to provide more added value to qualitative data, but also to support data outside social science. In addition, traditionally qualitative researchers now have some of the largest archiving needs, since video, audio, and photographs take large amounts of space; we are working with some of these groups as well, and are building out this capacity with the Harvard University Libraries.

- More than 665,000 files accessible via IQSS Dataverse Network
- More than 533,000 data downloads
- Average number of daily accesses: 1,254 (388 visits/day)
- Nearly 1,400 support tickets in total (351 for CY11 through October)
- More than 5,000 downloads of the entire Dataverse Network software package
- At least 15 installations in use at other universities around the world

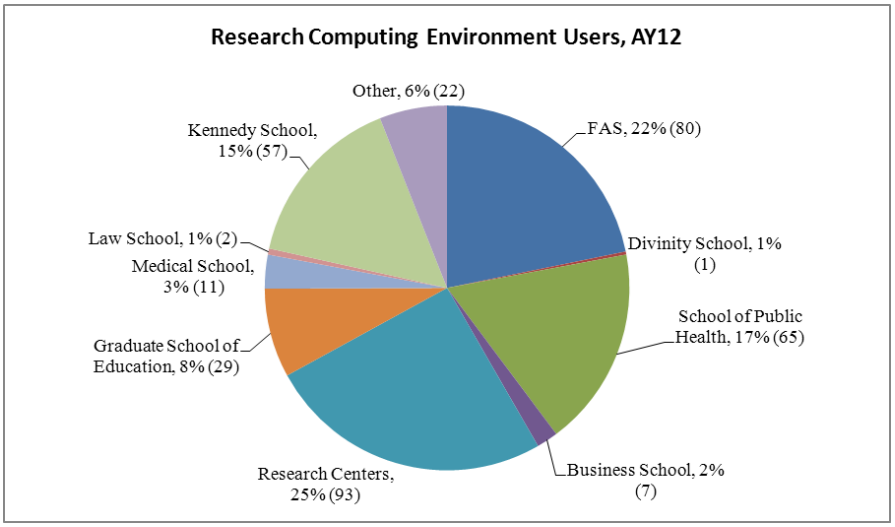
Research Computing Environment

The Research Computing Environment (RCE) has evolved over the last five years from a loose band of independent servers dedicated to individual professors for their own personal computing work, to a shared pool of scalable resources that is accessible to all and that offers an easy-to-use, persistent interface and a tiered model of use. Today's RCE provides a remote connection environment which enables access from anywhere via the internet and allows a session to persist over time. It also offers a graphical desktop environment with application menus and windowing features. And, of course, it includes statistical software run on powerful servers capable of handling large data analysis jobs. With the Provost's support, we have been able to offer the RCE tool to researchers across the University.

Due to space constraints, we have had a flat number of servers over the past five years. We have still been able to satisfy increasing compute capacity and data storage needs, however, by leveraging technological advances for increased density and increased computing cores. Now at our maximum cooling capacity and unable to upgrade the air conditioning due to the FAS financial situation, we have been looking for other alternatives to meet ever growing needs. Recently we developed a collaboration with University CTO Jim Waldo and SEAS Director of Academic Computing Robert Parrott to develop the next generation cloud platform for extending local compute clusters by accessing resources from other sources. This will likely be another open source tool that will be available across the University and will allow groups, in addition to ours, to expand their computing cycles at any given moment without purchasing additional hardware.

Key facts and figures about the Research Computing Environment:

- More than 636 computing cores
- Over 450 active users including people from 8 different Harvard schools



- On average, 50% of the cluster in use at any given moment
- Includes an entirely separate, smaller cluster of large machines with a small number of cores that individual users can reserve for different types of dedicated computing needs

Zelig

Our primary statistical programming product is *Zelig: Everyone's Statistical Software*, which began in 2000 as a research concept led by graduate students and faculty affiliated with IQSS, and is now has hundreds of thousands of users around the world. As with OpenScholar and the Dataverse Network, Zelig is an open source project; Harvard and IQSS benefit as many from inside and outside the University contribute code to the project and recognize our creative leadership role. Zelig is a package for R, an open source programming approach widely used in the statistical community across numerous fields. It standardizes how to run statistical models in order to facilitate their usage; it is built on and facilitates Harvard faculty research into ontologies of statistical modeling. Zelig makes it far easier to run cutting edge statistical methods without having to understand the cacophony of different programming styles, documentation standards, computer program syntax, and substantive examples in existing R packages. Initially, Zelig contained only 4 statistical models; it now has nearly 100. As the project moved to IQSS in 2008, development became more productive and robust with better documentation. The team continues to improve documentation and add models to the package.

Key facts and figures about Zelig:

- 2008 – Initial package released with 56 models
- 2009 – Integrated with other software, including the Dataverse Network
- 2011 – Developed Zelig 4, with a new API added to facilitate the development of new models
- Often named among the Top 10 of the more than 3,500 most downloaded packages from CRAN (the archive for all R packages)
- Quarterly training sessions offered since 2010

Consilience

Consilience is a work in progress and will offer a new way to discover and organize huge volumes of unstructured text; we think of it as computer-assisted conceptualization from unstructured text. Word processors provide computer-assisted writing—they do not write for you, but they make writing easier and more efficient. Our Consilience tool is being designed to offer computer-assisted reading. Whereas no machine could read a page of text better than a human being, zooming in to extract meaning, humans are not good at zooming out and understanding a large number of documents all at once. This tool is aimed at providing this zoom in/zoom out capability. The target audience includes almost all of those at Harvard who write literature reviews, review field notes, try to understand academic literatures, delve into government regulations, comprehend social media posts, or "read" any volume or source of unstructured text.

Simultaneously applying to a set of documents all known clustering algorithms and, due to a mathematical advance, all algorithms that *could be* invented, a researcher will be able to navigate the universe of possible clusterings and also drill down to each individual document. This project was originally begun as part of the Program on Quantitative Methods and joined forces with the Product Development team in 2010. We were recently awarded a Library Labs grant to build a proof of concept design and evaluate whether or not it will be useful to archivists at the Harvard University Library to have an ingest front-end that allows them to use this tool to understand, organize and label email archives and similar digital content worth preserving. At the end of 2011, we had an internal alpha release, and we plan to have a beta release in 2012.

Programs and Community

IQSS is the home of several faculty-led programs. IQSS develops and makes available a suite of programs to faculty, students, and staff across the University. Short descriptions of these programs are detailed below:

Program on Quantitative Methods

The flagship of IQSS's research program, the Program on Quantitative Methods (PQM) encompasses a wide variety of projects that actively develop innovative methods and software

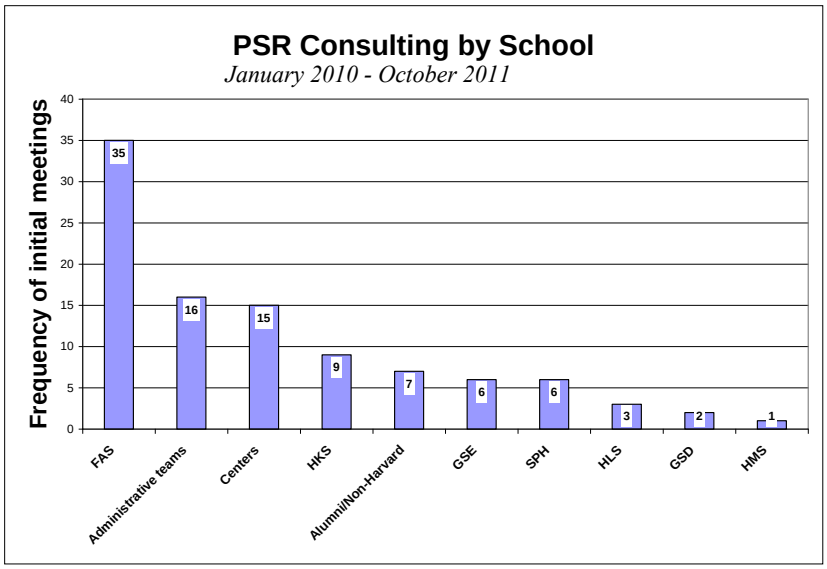
as a way of building and unifying the methodological subfields now existing separately within most social science disciplines. PQM sponsors a weekly, for-credit workshop (Applied Statistics, GOV 3009), during which participants discuss statistical innovations and applications. It is billed as a "tour of Harvard's statistical innovations and applications with weekly stops in different disciplines" and is regularly attended by a dozen faculty and 40-50 students throughout the University. PQM maintains the Social Science Statistics blog, which is moderated by graduate students and enables a free exchange of information pertaining to statistics, formalizing and recording some of the numerous hallway conversations underway. PQM also serves as the home for new software tools, such as the program *rbuild* which enables collaboration on the development of R packages hosted within the IQSS Research Computing Environment. These tools are all open-source and designed to scale for maximum scientific benefit. Many of the tools that started in PQM, including Zelig and Consilience, have moved into full-scale production with the IQSS team, while continuing to be worked on as research projects.

Program on Survey Research

The Program on Survey Research (PSR) provides expertise and resources on survey methodology to enhance the quality of research and teaching across all schools and units of Harvard. PSR is poised to capitalize on tremendous opportunities to help the social sciences. Survey research accounts for fully half of all quantitative research on human subjects in academia, but with massive increases in non-response and cell phone usage, its scientific foundations are crumbling. PSR is searching for solutions to this problem. At the same time, web surveys pose an enormous opportunity because they are far less expensive to use (the marginal cost of an additional respondent is almost nil), but they also present a serious challenge, since random selection on the internet is impossible. Thus far, we only have partial answers to these issues, but we are taking a leadership role in maintaining continued progress. PSR sponsors conferences and workshops that bring together survey scholars and practitioners in the media, government and commercial sectors. It also offers a variety of resources for students and faculty interested in scientific survey methods, including courses, short workshops, and one-on-one advising. PSR's services are open to all faculty, students, staff, and affiliates of the University.

Key facts and figures about PSR:

- Over 100 initial one-on-one consultations from more than 50 different people looking for help on survey-related projects in the past 22 months
- Reach includes FAS, HKS, SPH, GSE, HBS, GSD, HLS, HMS, and various administrative groups on campus (Institutional Research, Office of Undergraduate Education, Harvard College Library, etc.…)
- Annual conferences on survey experiments since 2008
- Holds University license for Key Survey, a web-based survey tool



Program on Text Research

The statistical analysis of text is a mainstay of almost all academic disciplines and the explosion of the web guarantees that its role will only increase in importance. The Program on Text Research (PTR) is seeking to become a world leader in the creation, preservation, and dissemination of statistical text analysis knowledge. Led by *Professor Arthur Spirling* with strong support from *Professor Peter Bol*, *Stuart Shieber*, and others, PTR seeks to inform and influence (and be informed and influenced by) scholars working with text who are essentially unaware—or perhaps wary—of the use of statistical approaches. The focus here is on qualitative work in fields such as history, English, and modern languages, anthropology, sociology, religious, and cultural studies. PTR is committed to building up a repository of plain text data for textual analysis and hopes to augment its teaching and consulting services. Each year, PTR hosts a conference with participants hailing from a multitude of disciplines and departments.

Program for Experience Based Learning in the Social Sciences (PEBLSS)

The Program for Experience Based Learning in the Social Sciences (PEBLSS) provides a modern support infrastructure to Harvard programs and faculty that create experiential (or "active") learning opportunities for Harvard students. Led by *Professor Dustin Tingley*, PEBLSS coordinates common interests in experiential learning and reduces common costs of effectively using and evaluating this type of teaching tool. PEBLSS is designed to enhance existing Harvard institutions by providing scalable services predicated on modern social science principles and to create new opportunities for experiential learning where appropriate. The program will provide a common infrastructure and service base that speaks to the shared needs of an experience based curriculum.

Data Privacy Lab (DPL)

The Data Privacy Lab (DPL) is dedicated to creating technologies and related policies with provable guarantees of privacy protection while allowing society to collect and share private (or sensitive) information for many worthy purposes. Under the direction of *Professor Latanya Sweeney*, the DTL accomplishes this by partnering with institutions, agencies, and corporations facing real-world privacy concerns.

Work in the DTL consists of two competing teams. The first team's focus is on developing ways to learn sensitive information from disparate and seemingly innocent information. Members of this team are called "data detectives" and their results are termed *semantic learning algorithms.* Being good at learning sensitive information from data allows the DTL to better understand what is needed to be "data protectors," which is the name given to members of the second team. By constructing solutions that can provably control what can be learned, the DTL provides effective *privacy technology* for real-world problems.

Global History of Elections Program (GHEP)

With recent advances in information technology—GIS mapping technology, digitized historical archives, digitized parliamentary transcripts, and online historical census reports etc.—it is now possible for scholars to undertake historical quantitative research on elections going back to the 19th century. The Global History of Elections Program (GHEP), led by *Professor Daniel Ziblatt*, aims to explore this new area of research by linking and synthesizing two centuries of national census data from as many countries as possible with election, roll call and petition data going back to 1800.

NASA Tournament Lab (NTL)

Under the leadership of *Professor Karim R. Lakhani* and *Professor Kevin Boudreau*, the NASA Tounament Lab (NTL) designs and fields competitions to create the best computer code for NASA systems, an approach often referred to as "crowd sourcing." With the creation of the NTL, Lakhani, Boudreau, and others are able to conduct research into the optimal design parameters for innovation competitions of this type, facilitating the use of these tournaments within the public and private sectors. The NTL also supports contests for Health and Human Services.

Roybal Center for the Study of Social Networks and Well Being

The Roybal Center for the Study of Social Networks and Well Being at Harvard investigates in detail the effects and the interplay of social networks and geographic area on health and healthcare. Headed by *Professor Nicholas Christakis*, the Roybal Center's mission is to promote research that explores how social networks affect the health and well being of Americans across the lifespan, with the goal of developing and implementing practical methods that will help to improve the health and well being of older people.

Student Programs

Undergraduate Research Scholars Program

By taking part in the Undergraduate Research Scholars (URS) program, students and faculty engage in a unique collaborative experience that fosters teamwork and a sense of academic community. At the same time that faculty members receive support for their research, students

become an integral part of an academic team and gain first-hand experience in on-going, faculty-led research projects in a diverse array of fields. Our Undergraduate Research Scholars program is led by sociology *Professor Filiz Garip*, who ensures that students work side-by-side with faculty and contribute to the discovery and development of new tools, methods and knowledge.

We currently have 22 faculty members from 10 departments and 3 schools involved in the URS program, and 44 undergraduate research scholars.

Graduate Student Program
IQSS graduate students are nominated by current faculty affiliates. Affiliate status is offered only to those who regularly participate in the life of the Institute and make contributions to the IQSS community (such has running seminars, teaching short courses, mentoring others, etc.). Graduate students are eligible for: participation in monthly luncheons, regular workshops and conferences; conference and research grants; publication on our website and white paper series; office space; logistical and financial support for workshops; and more.

We currently have 111 active graduate student affiliates from 5 schools, including FAS (79%), SPH (10%), HKS (5%), GSE (5%), and HBS (1%).

Seminars, Workshops, Conferences
We regularly bring members of the community together through seminars, workshops, conferences, and training sessions. These activities vary in size and shape in order to best accommodate the subject matter and goals of the individual event, but are nearly always open to researchers from across the University, and often beyond.

Our flagship event, sponsored by the Program on Quantitative Methods (see *Section E. Program on Quantitative Methods*, above) is the weekly *Applied Statistics Workshop.* This workshop offers faculty and graduate students a chance to see statistical innovations applied and developed in active research across a broad range of disciplines. Attendance each week includes a dozen faculty members and 40-50 graduate students. The workshop is listed as a course in the Department of Government, the Sociology Department, and HSPH; speakers over the past three years have included members of the government, economics, psychology, sociology, and statistics departments, as well as presenters from five different Harvard schools (business, education, law, medicine and public health) and multiple visiting faculty. The course is open to students from all disciplines. Other classes offered through IQSS include the Graduate Student Workshop in Political Economy, the annual Math Prefresher, and the Political Psychology and Behavior Workshop.

We also run regular conferences and seminars each year. Many of these larger-scale events have been funded in part through generous gifts from Eric C. Mindich and have included conferences such as Data Citations, Political Economy and Public Law, New Directions nText Analysis, Survey Equality, and others. Our seminars include the Positive Political Economy Seminar, the Faculty Discussion Group on Political Economy, and the Colloquium

on Complexity and Social Networks. We also provide logistical and sometimes financial support for a variety of seminars run by students (Graduate Student Workshop in Political Economy, Graduate Methods and Models, etc.), the Data Privacy Lab, Qualitative Social Science@Harvard, and more.

In total, we run more than 20 events each year (recurring and non-recurring), with close to 500 participants hailing from the FAS (roughly one-third), other Harvard schools (one-third), and outside Harvard (one-third). We also co-sponsor and help fund another half dozen events.