Yale University

Report of the University-Wide Data-Intensive Social Science Committee (DISSC)

Table of Contents

Executive Summary

In late 2016, President Peter Salovey identified data-intensive social science as a top academic priority for Yale, emphasizing the application of empirical methods and data to public policy issues and matters of social concern. "A great university should be engaging in the great debates of its era, and our students – the leaders of tomorrow – should participate. But that engagement must be grounded in evidence-based inquiry and rigorous analysis of facts."

The President's vision for social science presents both an exciting opportunity and a timely challenge. Social science research can not only help us understand, describe, and measure social behavior. It can also promote human welfare by providing the tools to develop and evaluate strategies for tackling the most pressing problems facing people both in the U.S. and around the globe, including income inequality and social mobility; democracy and civic engagement; migration, demographic change, and political identity; trade, innovation, and jobs; criminal justice; taxing and spending; education and early childhood development, and the environment and climate change.

Social science has long illuminated important issues and is exceptionally well-positioned to offer new insights into the role of individuals, groups, institutions, and markets in social life. The availability of massive amounts of social and behavioral data, rapidly increasing computational power, and potent new methods for analyzing all sorts of data are transforming how many social scientists do their work. Digitalization of very large databases, the data-streams produced by social media and as a by-product of commercial transactions, and vast archives of text and images give researchers the capacity to achieve remarkable precision and texture in the description and analysis of the patterns of social behavior. Computational power and analytical methods permit the design and launching of research programs that would have seemed like science fiction just a few years ago. A study of legislative polarization that might have spent years assessing the speeches from a single session of Congress can now instead analyze the millions of pages of text comprising the entire Congressional Record in seconds. Millions of employment and earning records can be merged with other administrative data bases to uncover hidden patterns linking early childhood environment to an individual's lifetime opportunities. Insurance claims databases can be mined to uncover how medical billing rules lead to unexpected and potentially ruinous medical expenses. Data from satellite images can be used to describe the height and density patterns of city growth across the world and assess how variation in global urbanization patterns will affect carbon emissions. Social scientists are tackling large and important questions relevant to social problem-solving and finding new answers that are improving our understanding of the world and informing the design of more effective and equitable solutions to our problems.

Yale's social science research community is among the strongest in the world, and our faculty are already using data in imaginative and impactful ways. Making the investments to allow faculty to take full intellectual advantage of rapidly developing analytical methods and providing the infrastructure for innovative data use (including legal support for data use agreements and protections to ensure that data is held securely and shared according to the terms of any agreements) will ensure that our social scientists remain at—and push forward—the research frontier. And Yale's students should be prepared to engage the world with confidence and positive impact by learning the methods and principles of quantitative analysis, how to evaluate and apply empirical evidence wisely, and how to participate in scientific discovery.

To chart a path forward, the University-Wide Committee on Data-Intensive Social Science (DISSC) was charged to recommend the key priorities in data-intensive social science for the next decade in the areas of research infrastructure, teaching, and organizational structures and behaviors. We were asked to gather input from faculty across the University, take inventory of our current resources and strengths that could support data-intensive, policy-relevant social science, and benchmark against peers to understand how other universities are responding to similar challenges and opportunities.

We devoted significant time to gathering information and conducting our deliberations. We solicited input from the community broadly, at the school, departmental, and individual faculty levels through emails and an online survey. We also conducted one-on-one and group interviews with faculty, instructors, department chairs, DUSs, and DGSs. During the 2018-19 academic year, we organized ten focus area group meetings in which we invited colleagues from different departments and schools to guides us to the research frontiers of their respective areas of specialty. Focus areas included education, finance, international development economics, health, environment, urbanization, criminal justice, and methodology. In this format, we met with over fifty faculty from nine schools. The Committee gathered information on peer institutions through emails, phone interviews, and site visits.

The Committee used two criteria to evaluate the recommendations: impact on Yale and feasibility for Yale. This provided a framework for considering how to prioritize the recommendations. We also attempted to estimate the annual cost of each recommendation if implemented so that we could prioritize and make trade-offs among options.

Based upon this process, we offer the following recommendations for research infrastructure, teaching, and organizational structures and behaviors.

Research Infrastructure

Yale should stand among the top universities in the world in developing the methods of data-intensive social science and in the application of these methods to advance basic knowledge and address important policy issues.

Goals:

I. ***As rapid technological progress expands data availability, the set of analytical tools, and computational power, researchers should have state-of-the-art physical and human infrastructure to reach and extend the research frontiers of their disciplines.*** The technological revolution characterized by a proliferation of data sources, increased computing power, and new analytical techniques is changing how we do social science research. This shift requires a reorientation of how we think of the resources that support this research. We believe that social science increasingly requires facilities akin to the core facilities in common use in the natural sciences. Often these facilities hinge on specialized and highly skilled staff rather than physical infrastructure, so it is imperative that the University create ways to attract, build, and sustain these people.

II. ***The University should build the social science research community and foster coordination across departments and schools.*** Yale can harness the quality and breadth of expertise in data-intensive social science by supporting the identification and realization of research affinities across departments and schools. Expertise should not remain siloed but should be visible and accessible across our campus. The social sciences represent an outstanding opportunity to build connections because the social sciences are not only in the Faculty of Arts and Sciences but also are distributed across many Yale schools, including SOM, the Law School, the School of Public Health, FES, and the Medical School.

III. ***Basic services supporting data-intensive social science should be selectively improved.*** The Committee identified selective areas in which basic services could be strengthened. Areas such as IT support, statistical consulting and training for both researchers and students, and access to the Institutional Review Board (IRB) for research involving human subjects could benefit from additional attention to the needs of social science researchers. These services are used by a large community, so even modest improvements will yield significant benefits.

Recommendation:

Create a Data-Intensive Social Science Center

We recommend the creation of a data-intensive social science center at Yale, anchored by a core facility for sensitive and restricted-use data. The facility would provide the infrastructure to support the secure acquisition and use of new data resources. The center would also promote exposure to and skill development in the new computational and statistical methods that are being developed in statistics, computer science, and application fields. The center would become a crossroads for data-intensive social science researchers from across the University, creating University-wide awareness of the events, information, and resources available for data-intensive social science, and building community for researchers across the University with expertise and interests in the different methods and applications of data-intensive social science.

To accomplish this mission, the center would serve six primary functions:

1. **Facilitate acquisition, computing and storage** of sensitive and restricted-use data by building a secure data facility with a team of expert staff
2. **Provide research consulting and data science services** with project planning, programming support, and guidance to other University resources
3. **Offer significant seed grants** to stimulate cross-departmental and cross-school connections and research collaborations
4. **Provide outreach and a web portal** to introduce social science researchers to Yale's data science resources
5. **Build community** and stimulate intellectual exchange and collaboration around the data, methods, and innovative research designs and applications of data-intensive social science through workshops, conferences, visiting scholars, and research fellows
6. **Host IT support** specialized to data- and computational-intensive social science

Each of these functions is described in greater detail in the full report below.

Teaching

All Yale College students should develop the habits of mind that will enable them to identify the strengths and weaknesses in empirical evidence, ask probing questions about empirical claims, and use quantitative evidence wisely in forming opinions and making decisions. All Yale College students who seek to achieve mastery of quantitative methods should have a clear path to reaching high levels of expertise. Our educational objectives should evolve as new skills and tools are developed and as research designs for data-intensive social science improve.

Goals:

I. **Offer basic courses for key ideas and methods of data-intensive research and analysis.** Yale College should offer accessible but rigorous courses that cover the fundamental insights about research design, reasoning about quantitative evidence, and using quantitative evidence in belief formation and decision-making. Every Yale College student should have an opportunity to experience the excitement of empirical investigation and discovery in content areas of academic interest to them. Students should have multiple exposures to the basics and multiple opportunities for developing and applying their data analysis skills. We should promote a culture of rigorous and wise engagement with empirical claims.

II. **Establish paths to advanced achievement**. These pathways should have multiple entry points and include a pathway for students who begin their education at Yale College without substantial prior exposure to statistics, programming, or advanced math.

III. **Strengthen communication, coordination, innovation, and assessment.** There should be coordination among instructors across campus and students should know what is being offered. Yale College should periodically update its goals for teaching and learning and measure progress towards them. Undergraduate students should also be able to self-evaluate their levels of understanding, mastery and progress. There should also be mechanisms in place to learn about what is being done well at other universities.

Recommendations:

1. **Enhance basic data-intensive course offerings in Yale College**

   To offer accessible but rigorous courses that cover the fundamental insights about research design, reasoning about quantitative evidence, and using quantitative evidence in belief formation and decision-making, the Committee recommends enhancing basic data-intensive course offerings in Yale College by:

   a. Establishing data-intensive course sections across the Yale College curriculum to provide students with the opportunity for multiple exposures to data analysis with applications they care about
   b. Creating interdisciplinary, “signature” lecture courses in data-intensive social science to attract general background undergraduate students and provide an initial exposure to

rigorous empirical social science inquiry, principles of research design, evaluation of evidence quality, and decision-making using data

c. Expanding the YData[1] course to include additional connector seminars with a broad reach across humanities, social science, and science

2. Create a pathway to advanced achievement in quantitative social science in Yale College

The Committee recommends establishing a pathway to advanced achievement in quantitative social science for undergraduate first years and sophomores. This would not be a major; however, it would bring together a cohort of first and second years who would take a collection of courses designed to provide rigorous and accelerated preparation for future quantitative social science majors. The collection of courses for this proposed program would be worked out by a faculty committee in consultation with DUSs and DGSs and would be evaluated though standard channels. The program would provide a broad but rigorous introduction to the tools and applications in the quantitative social sciences, position undergraduate students to design and execute outstanding senior projects and prepare students for graduate school or work as an RA with professors on research frontier projects.

3. Expand the pre-doctoral program currently housed in the Tobin Center for Economic Policy

The Tobin Center's Economics Pre-Doctoral Fellows Program supports policy-relevant economics research by providing a high-quality education and training experience for individuals with bachelor's or master's degrees who are considering pursuing a Ph.D. in economics or a closely related discipline. The fellows work for one to two years as full-time research assistants for faculty mentors along with receiving additional training. The Committee recommends building this program to provide opportunities for fellows to work with faculty in data-intensive social science disciplines outside of economics (e.g., psychology, political science, linguistics, etc.). A major goal of this expansion is to provide additional resources to help promote diversity and inclusion in data-intensive social science.

4. Establish annual meeting of intro stats and research design instructors and relevant Yale College DUSs to improve communication, coordination, innovation, and assessment

The Committee recommends that Yale College organize an annual meeting of instructors of introductory statistics and research design courses and relevant Yale College DUSs. This meeting would serve as a forum for departments and instructors to better coordinate concepts and techniques taught in the classroom, identify gaps in level or material covered, share best practices, and brainstorm innovative and collaborative teaching ideas. This group would also

[1] YData (S&DS 123b, aka YaleData) was introduced in Spring 2019. YData is modelled after UC Berkeley's popular Data 8 course and is designed to be a highly interdisciplinary, unintimidating, introductory data science course for students with little or no background in statistics, math, or computer science and who aren't necessarily interested in pursuing these fields as their majors. The course teaches traditional statistics and data science concepts using Python. In its first year, YData offered three half-credit connector seminars for students to take concurrently with the main course on diverse topics in applied data science, including political campaigns, exoplanet astronomy, and text processing. These connector seminars for YData help reinforce the main ideas, themes, and techniques learned in the core class by giving students a chance to apply them with relevant datasets in an area of substantive interest. The primary goal of the class is to teach students "the data science way of thinking" and prepare them to critically analyze the information they come across day-to-day in the news, scientific studies, and elsewhere.

review course offerings and check that online course listings and descriptions are accurate. A work product of this annual meeting could be an updated guide contrasting the various Yale College introductory courses. This guide could be shared with the DUS of each Faculty of Arts and Science department and then disseminated to advisers to share with their undergraduate students.

5. **Consider appointing a committee to evaluate the Yale College quantitative reasoning (QR) requirement**

   The Yale College Dean should consider appointing a committee to evaluate how the QR requirement is functioning and to consider either changing the requirement or adjusting its implementation given the increased prominence of computation and data-intensive analysis. The committee should also explore how peer institutions approach requirements around baseline level of exposure to data analysis and statistical reasoning.

Each of these recommendations is described in greater detail below.

Recommendations on Organizational Structures and Behaviors

In the Charge for the University-Wide Committee on Data-Intensive Social Science, Provost Ben Polak asked the Committee to “make suggestions about organizational structures and behaviors that could support data-intensive social science at Yale, [particularly] recommendations about mechanisms for better coordinating across Yale to improve efficiency, innovation, and impact, and mechanisms for rapidly learning relevant developments and innovations occurring at other universities.”

In addition to the research and teaching recommendations, above, the Committee has two recommendations specific to organizational structures and behaviors.

Recommendations:

1. **Establish a University-wide committee to share information about data-intensive social science**

   Yale should consider establishing a twice annual meeting of social science center directors, department chairs, and other academic leaders who are most involved with data-intensive social science research. This would bring together approximately 15-20 people who have significant oversight responsibility in this area. These meetings would be for sharing information, coordinating plans, and providing advice to university administration and service units. We recommend that for at least one of these meetings each year, the key service providers such as ITS, YCRC, OSP, and Yale Library are included. This meeting could be used to discuss faculty needs and for faculty to provide advice and reactions to service unit plans.

2. **Learn about developments at other universities and industry-leading organizations**

   We propose that Yale's Office of Institutional Research and Strategic Analysis (OIR/SA) produce a periodic memo on "Innovations and Lessons," perhaps annually, on major developments supporting data-intensive social science research and teaching at industry-leading organizations and peer universities. The memo would describe important infrastructure investments, important changes in data policies and data availability, key programs being started, and programs being discontinued. This memo should be sent to university deans, center directors, department chairs, and shared with the faculty. Based on this research, OIR/SA might propose one or more faculty site visits each year to places that seem especially innovative or cost-effective.

Relation to the University Science Strategy Priorities

In 2017, the University Science Strategy Committee (USSC) was formed to make recommendations for Yale's scientific research investments over the coming decades. The USSC identified five areas of science for strategic investment. For three of these areas, Integrative Data Science and its Mathematical Foundations, Neuroscience, and Environmental and Evolutionary Science, the tools, empirical insights, and theoretical models of the social sciences are important resources for intellectual progress. DISSC Committee members believe that it would be valuable to invest, where relevant, in social science areas adjacent to these science priorities to draw on the relevant methods and theoretical perspectives of the social sciences. This suggestion applies to all three of the priorities mentioned, but we focus on the priority which most clearly intersects with the DISSC's charge, the USSC's Data Science recommendation.

A broad approach to data science would position the University to excel at basic research that advances the methods of data science, applied research that uses the tools of data science to advance disciplinary and interdisciplinary research programs, and applied research on the societal consequences of the revolution in computation and related technology. From the standpoint of the social sciences, we might organize these ideas into two themes: accelerating social science research using data science and understanding the impact of computation on society.

Regarding the first of these themes, advances in data sources (e.g., digital media trails, administrative data, transactions data, text and image archives, sensitive and restricted-use data sets, remote sensing data, location tracking), computational power, and analytical methods (including advances in machine learning, natural language processing, and image processing) are transforming how we study traditional questions at the heart of the social sciences. The USSC report contains an excellent discussion of the ways that new data sources and analytical techniques may propel research in the social sciences.

The proposed Center for Data-Intensive Social Science is designed to support efforts to take full advantage of these growing research opportunities and the proposed center fits in well with the emerging university organizational schema of forming centers that are devoted to advancing research at the intersection of data science and some large segment of the intellectual landscape. It would complement the USSC's proposed Institute for Integrative Data Science and its Mathematical Foundations and other existing centers and institutes such as the Center for Biomedical Data Science, Quantitative Biology Institute (QBio), and the Digital Humanities Lab, while addressing the unique needs of data-intensive social science.

Regarding the second of these themes, how rapid technological change is affecting society, the USSC report notes that “the world is currently undergoing a data revolution comparable to the industrial revolution in its potential impact…Not a single aspect of society today will be left untouched by the data revolution.” Although there is currently no coordinated University-wide initiative in place to understand how advances in computation are transforming society, some significant initial efforts to address the social and individual impact of the technological revolution are underway at Yale. There are substantial efforts to build research capacity on the human impact of the computational revolution at many peer institutions, including Stanford, MIT, and Berkeley.

Assessing how Yale might engage with the social impact of computation is a matter of concern to all disciplines and schools and therefore both beyond the scope of the DISSC’s charge and the limited range of expertise of DISSC’s members. That said, DISSC believes that Yale will not remain a center for innovation and excellence in data-intensive and policy relevant social science research if it fails to play a significant if not leading role in engaging with the technology-led transformations and opportunities that characterize our era. If this direction is of interest, we recommend as a next step for the Provost to appoint a working group to explore how the University can lead in this area.

Acknowledgements

The Committee is grateful to the many members of the Yale community who assisted in this process. We thank those who prepared departmental self-reflections, participated in focus groups for research and teaching, and provided input to the Committee in response to our requests. We are particularly grateful to Limor Peer, Jill Parchuk, Melanie Maksin, and Zack Cooper for their contributions to the Committee's work. We also thank Peter Schiffer for his active engagement throughout the process.

Committee Members

Committee Chair:

Alan Gerber, FAS Dean of Social Science, Charles C. & Dorathea S. Dilley Professor of Political Science, Professor in the ISPS, of Economics, and of Public Health (FAS)

Committee Faculty Members:

Judith Chevalier, William S. Beinecke Professor of Finance and Economics (SOM/FAS)

Marvin Chun, Dean of Yale College, Richard M. Colgate Professor of Psychology and Professor of Neuroscience (FAS/YSM)

Pinelopi Goldberg, Elihu Professor of Economics (FAS/SOM)

Harlan Krumholz, Harold H. Hines Jr. Professor of Medicine (Cardiology) and Professor in the Institution for Social and Policy Studies, of Investigative Medicine and of Public Health (Health Policy) (YSM/YSPH)

John Lafferty, John C. Malone Professor of Statistics & Data Science (FAS)

Yair Listokin, Shibley Family Fund Professor of Law and Professor of Management (YLS/SOM)

Andrew Metrick, Janet L. Yellen Professor of Finance and Management (SOM)

A. David Paltiel, Professor of Public Health (Health Policy), Professor of Management, and Professor in the Institution for Social and Policy Studies (YSPH/SOM)

Karen Seto, Frederick C. Hixon Professor of Geography and Urbanization Science (FES)

Ebonya Washington, Samuel C. Park Jr. Professor of Economics (FAS/SOM)

Committee Staff Members:

Tim Pavlis, Associate Vice President for Strategy and Academic Business Operations

Naureen Rashid, Special Adviser for Strategy and Academic Business Operations

Introduction

In November 2016, President Salovey identified data-intensive social science as a top academic priority for Yale, emphasizing the application of empirical social science to public policy issues. To identify major ideas for strategic new investment in data-intensive social science, Provost Polak organized the University-Wide Committee on Data-Intensive Social Science (DISSC) (see Committee charge in the Appendix). Our Committee was convened in January 2018 and met regularly through May 2019.

We were charged to engage broadly with the University community to identify ideas in which additional investments would have a maximum impact on the overall quality of teaching and research in data-intensive social science at Yale. We were asked to develop a prioritized list of ideas that could be accomplished with no additional resources as well as with an additional two to four million dollars in annual expenditures. Each idea was assessed in terms of impact, resources required (funding, space, faculty, etc.), and feasibility.

The Committee was moderate sized to keep the discussions manageable and efficient. Given the University-wide scope of its charge, the small committee size required that the Committee's members were not fully representative of the breadth of social science at Yale. At the initial charging meeting, the Provost asked members not to consider themselves as members representing their own department or school, but rather to take a "university-wide view" of data-intensive social science at Yale and to conduct their deliberations accordingly. The ideas under our consideration had to be big enough and broad enough to garner support from the entire Committee. We were encouraged to think beyond the boundaries that may exist between departments and between schools. The outcome of this process is a set of recommendations for significant new investment in data-intensive social science.

We lay out specific recommendations for strengthening research, teaching, and our organizational structures. Our goal is to take advantage of existing strengths at the University and to make key strategic investments to move Yale to the forefront of research and teaching in data-intensive social science. We believe that data-intensive social science at Yale is often constrained, not by the quality or quantity of good ideas, but by our structural support and the ability of our community to organize itself around those ideas. Our recommendations for organizational structures are meant to ensure that mechanisms are in place for better coordination across the University to improve efficiency, innovation, and impact and for learning relevant developments and innovations occurring at other universities so that we can fully engage with the ever-evolving research and teaching frontier of data-intensive social science.

This report begins with a summary of the deliberation process we followed to reach our recommendations and is followed by vision and goals and specific recommendations for investment in research infrastructure, teaching, and organizational structures and behaviors. The recommendations in this report represent the consensus opinion of the Committee, and we hope the University will find them helpful in implementing a strategic plan for data-intensive social science in the coming decade.

Respectfully,

University-Wide Committee on Data-Intensive Social Science

Deliberation Process

Given the scope of our charge, the Committee undertook an extensive process for collecting and considering input from multiple stakeholders across the University. We viewed our role as being a conduit to collect and organize the many excellent ideas that are emerging across the University.

We broadly solicited input from the community, at the school, departmental and individual faculty levels. The committee conducted one-on-one and group interviews with faculty, instructors, and department chairs. The Committee solicited comments on the main final teaching recommendations from FAS social science chairs and DUSs. We also asked groups of faculty in FAS social science departments and relevant professional schools sets of questions via email. The questions included various groupings of the following:

1. What changes or initiatives would you propose to enhance research and teaching in data-intensive social science at Yale over the next decade? Could you please write just a few sentences about at least one idea that would not require a large expenditure and a few sentences about one initiative that might cost a lot of money? Small ideas are fine, but also please try to think of big things that would in your view make a big difference.
2. Could you give us a general sense of what sorts of sensitive data you handle in your research, how the data is stored, and how it is made secure? Are there services that Yale could provide to improve things in this domain? Do you know of any colleagues who use a lot of sensitive data whose perspectives we should be sure to obtain?
3. What concepts and techniques related to the collection, use, analysis, interpretation, and communication of data should every student graduating from Yale College be familiar with? (We use familiarity to mean not minimal acquaintance but a level of understanding that implies a reasonable degree of sensitivity, maturity, and sophistication)
4. What habits of mind should we instill in Yale College students to help them to think critically about the results and reporting of scientific, medical and other studies, and how do we accomplish this?
5. What role can data and evidence play in public policy, professional life, and personal life decisions?
6. What should Yale College students learn about how human rationality, irrationality, and bias affect the evaluation of empirical claims and individual and collective decision-making?

A Qualtrics survey was also sent out to faculty asking for input on technology, programs, and services; teaching and curriculum; what other institutions are doing; structural/organizational changes; major issues; and major strengths. Faculty were able to attach more detailed proposals to the survey in addition to their text comments.

During the 2018-19 academic year, the Committee organized ten focus area group meetings in which we invited colleagues from different departments and schools to guide us to the infrastructure and support needs at the research frontiers of their respective areas of specialty. Focus areas included education, finance, international development economics, health, environment, urbanization, criminal justice, and methodology. In this format, we met with over fifty faculty from nine schools.

Throughout our input-gathering process we also engaged with faculty at peer institutions who were involved with successful initiatives related to data-intensive social science. The Committee gathered information on peer institutions through online research, email, phone interviews, and site visits.

The Committee carefully considered the criteria to evaluate the ideas we received. We established two overarching criteria, Impact on Yale and Feasibility for Yale, which provided a framework for considering how to prioritize the recommendations. Based on this framework, the Committee discussed each idea, and each Committee member independently scored each idea. The scores were tabulated, and ideas were prioritized based on the highest cumulative scores.

In parallel with this process, we estimated the resources needed for each of the ideas and calculated their annual cost if fully implemented. Key cost drivers include the cost of staff, students, faculty, equipment and the capital and operating expenses needed for space (new or renovated, purchased or leased). These estimates were considered relative to the cost targets that were provided to the Committee in our charging instructions.

Full Report

Research Infrastructure

Overview

The Committee's recommendations focus on cross-cutting initiatives that provide the foundation for scholars to do excellent work in data-intensive, policy-relevant social science and to be highly competitive in seeking external support. Our recommendations build on the substantial University resources already devoted to supporting data-intensive social science research, which include the new Tobin Center for Economic Policy, Institution for Social Policy Studies, MacMillan Center, Economic Growth Center, and the Cowles Foundation. Yale has significant infrastructure to support data-intensive research, including the census data center (FSRC), the Yale Center for Research Computing (YCRC), Statlab, and the Yale library system's collection of data librarians and other specialists.

The Committee sought ways to provide an environment for faculty to do outstanding research and for faculty-led initiatives to take shape and flourish. Yale faculty will advance research and change the world in ways that are impossible to predict. Yale should produce a low friction environment for research and interaction among our scholars. This will foster creativity and productivity and help us to attract and retain the best scholars in the world.

The faculty input and Committee discussions reflect an understanding that the revolution in computational power and data availability, along with advances in data analysis techniques and the development of software that implements new analytical methods, has created exceptional opportunities for rapid advancement in knowledge across many domains. As new data sources become available and new analytical methods are introduced, it is a challenge for scholars to stay at the research frontiers of their disciplines. Our recommendations attempt to respond to these emerging changes in the research environment.

We begin this section of the report by stating our vision and goals for research at Yale that are based upon President Salovey's academic priority for enhancing data-intensive, policy-relevant social science. We then discuss our specific recommendations, explaining how each supports our goals. Finally, the section ends with other ideas and observations from our Committee's input-gathering process and deliberations. Although these ideas were not in our final list of prioritized recommendations, they came up frequently in our discussions, and we believe that there is value in considering them for future implementation.

Vision and Goals

President Salovey has identified data-intensive social science as an academic priority. Yale should stand among the top universities in the world in developing the methods of data-intensive social science and in the application of these methods to advance basic knowledge and address important policy issues.

Goals:

I. ***As rapid technological progress expands data availability, the set of analytical tools, and computational power, researchers should have state-of-the-art physical and human infrastructure to reach and extend the research frontiers of their disciplines.*** The technological revolution characterized by a proliferation of data sources, increased computing power, and new analytical techniques is changing how we do social science research. This shift requires a reorientation of how we think of the resources that support this research. We believe that social science increasingly requires facilities akin to the core facilities in common use in the natural sciences. Often these facilities hinge on specialized and highly skilled staff rather than physical infrastructure, so it is imperative that the University create ways to attract, build, and sustain these people.

II. ***The University should build the social science research community and foster coordination across departments and schools.*** Yale can harness the quality and breadth of expertise in data-intensive social science by supporting the identification and realization of research affinities across departments and schools. Expertise should not remain siloed but should be visible and accessible across our campus. The social sciences represent an outstanding opportunity to build connections because the social sciences are not only in the Faculty of Arts and Sciences but also are distributed across many Yale schools, including SOM, the Law School, Public Health, FES, and the Medical School.

III. ***Basic services supporting data-intensive social science should be selectively improved.*** The Committee identified selective areas in which basic services could be strengthened. Areas such as IT support, statistical consulting and training for both researchers and students, and access to the Institutional Review Board for research involving human subjects could benefit from additional attention to the needs of social science researchers. These services are used by a large community of scholars, so even modest improvements will yield significant benefits.

The Committee's recommendations and observations that follow support these goals. As this is a dynamic field, we also emphasize the importance of effective faculty input and governance to ensure that our resources evolve with the times. We anticipate that new needs will arise, and some will diminish, and these are changes that the Committee cannot yet anticipate. Hence, we see the need for University-wide governance and a regular process that will evolve these recommendations over time. Further, as the scale and complexity of social science research increases we anticipate more opportunities for collaboration across institutions to pool resources; faculty leadership will be required to identify and realize these possibilities.

Recommendations

Recommendation: Establish a Data-Intensive Social Science Center

Increasingly, the data science frontier in the social sciences involves researchers working with large administrative data sets, data with use restrictions, and data for which privacy and security concerns are crucial. Managing the challenges posed by these research opportunities requires special infrastructure and specialized and highly skilled staff that go beyond the expertise of an individual researcher. These dynamics are similar to those in the sciences, where institutions provide shared "core" facilities such as microscopes or gene sequencers for use by many researchers. These cores amortize the cost of the equipment, provide for special expertise in the staff, and can create opportunities for research collaboration among faculty using similar equipment. We believe it is time for leading universities to provide such facilities for social science research.

Therefore, we recommend the creation of a data-intensive social science center at Yale, anchored by a core facility for sensitive and restricted-use data. The primary mission of the faculty-directed center would be to provide the infrastructure to support the acquisition, security, and use of the new data resources that are transforming social science research. The center would also promote exposure to and skill development in the new computational and statistical methods that are being developed in statistics, computer science, and application fields. The center would become a crossroads for data-intensive social science researchers from across the University, creating University-wide awareness of the events, information, and resources available for data-intensive social science, and building a community for researchers across the University with expertise and interests in the different methods and applications of data-intensive social science.

To accomplish this mission, the center would serve six primary functions:

1. **Facilitate acquisition, computing and storage** of sensitive and restricted-use data by building a secure data facility with a team of expert staff
2. **Provide research consulting and data science services** with project planning, programming support, and guidance to other University resources
3. **Offer significant seed grants** to stimulate cross-department and cross-school connections and research collaborations
4. **Provide outreach and a web portal** to introduce social science researchers to Yale's data science resources
5. **Build community** and stimulate intellectual exchange and collaboration around the data, methods, and innovative research designs and applications of data-intensive social science through workshops, conferences, visiting scholars, and research fellows
6. **Host IT support** specialized to data-intensive and computational social science

The center will require space, modest staff, and computing resources.

1. Secure data facility

The need to facilitate use of sensitive and restricted-use data is at once the most critical as well as the most fluid. Cutting edge social science research increasingly centers on pioneering analysis of novel data sets or novel uses of restricted administrative data sets. The most talented and ambitious early career social scientists increasingly distinguish themselves through analysis of data sets that they are the first to obtain (or assemble) and analyze. Other scholars will exploit new opportunities made available by combining and analyzing massive data sets. These scholars would benefit from expedited access to data that is otherwise difficult to obtain. Researchers report varying requirements from data providers for how data should be handled but common issues with the challenges of how Yale's current infrastructure is set up to support their needs. This includes not just secure computing infrastructure, but also an acute need to more efficiently process data use agreements (DUAs), including a realistic calibration of the risk the University must assume to access cutting-edge data.

We propose that Yale build a secure data facility to reduce the barriers separating researchers from the data they need to advance knowledge and address important policy issues. We identified several critical functions that the secure data facility should perform:

a. Building and maintaining infrastructure: Design, develop, maintain, and continuously improve the infrastructure necessary (both physical infrastructure and specialized staff) to facilitate data-intensive social science research
b. Supporting individual researchers and research initiatives: Provide technical and legal expertise to support researchers seeking to conduct data-intensive research, including acquisition, management, security, access, and analytical support
c. Exploring opportunities to pool and share data assets already at Yale: discover and maintain broad knowledge of the full landscape of data-intensive social science work at Yale; identify and facilitate opportunities for sharing, collaboration, and savings across research assets, including the possibility of constructing data enclaves with broad but secure access across the Yale research community
d. Improving access to external data sources and partnerships: Identify existing data enclaves and intermediaries beyond Yale and seek access based on data availability and faculty interest; seek out partners beyond Yale for data-sharing collaborations, and perhaps take the lead in forming such data consortia
e. Provide a seamless connection between Yale's services to data-intensive social science researchers and other shared services on campus and off campus, including YCRC, the new Office of Corporate Strategy and Engagement, and the University Library

Based on extensive and repeated input from faculty in addition to a scan of how other universities and social science research centers address needs in this area, we provide an order-of-magnitude estimate that assumes we need approximately three staff people, some computing infrastructure, as well as physical space to house secure research "cold rooms."

Investing in specialized and highly skilled staff is critical. The three-person staff team would be responsible for addressing the complicated legal, security, and technology issues associated with acquiring and sharing complex data and for ensuring that the data is documented, maintained, and structured in a manner that supports an outstanding user experience for the researcher. We believe there is need for (1) a Chief Data Counsel, who is a DUA specialist with experience relevant for data-intensive social science, to maintain critical external relationships with data partners, negotiate DUAs, oversee DUA counsel, support University-wide efforts to minimize risk and increase efficiency, and collaborate closely with the Senior Data Engineer; (2) a Senior Data Engineer to coordinate technical elements of data acquisition, set up and manage the overall security and wellness of the data, work with staff at YCRC and ITS on technical solutions, maintain data quality tests and monitoring routines, and serve as a technical adviser on University-wide data security strategy to minimize risk and increase efficiency; and (3) a Research Consultant to facilitate use of available data sets and support researchers with data discovery, onboarding, access, and analysis in a secure research environment. The team should be guided by a faculty director, ideally the director of the proposed data-intensive social science center, who would share oversight responsibility with a steering committee representing the university data-intensive social science community.

We envision this team being co-located and working closely with social science researchers seeking to acquire and use large administrative data sets and sensitive and restricted use data. Frequent contact with Yale social scientists will reduce bottlenecks in communication, and exposure to Yale social scientists and visiting scholars will support the staffs' continuous learning of practices, data sets, and research interests across Yale and at other universities and centers.

2. Research consulting and data science services

Students as well as faculty seek varying levels of support in research design and consulting on analytical tools. Some of these services are offered on campus today – most notably the Statlab under the auspices of the library and research computing expertise in YCRC. The faculty director of the proposed center, who would be in continuous contact with social science researchers, could help direct researchers to existing service units and would be well positioned to advise those who run the existing service units on how to match services to researchers' needs. The center would be a crossroads for data-intensive social science, and so it would be natural for the Statlab to provide regular basic on-site consulting services tailored to the focus of the new center. Some faculty will need higher-level programming expertise to assist with research beyond the scope or expertise of consultation at the StatLab. Higher-level assistance in planning and programming to support data-intensive research could be provided by the Research Consultant on the center's secure data team. YCRC is also exploring solutions such as certifying external vendors who can provide fixed-duration programming support. The center should have resources to support such efforts and stay abreast of demand to determine if other support modalities for programming would be appropriate, up to and including building some additional in-house capacity. The center could serve as an incubator for expanded data science services.

3. Seed grants to build connections across Yale's departments and schools

Innovative research often stems from people from different disciplines and backgrounds coming together to attack a common problem. To encourage such innovative work, we propose offering seed grants to catalyze conversations among researchers across departments and schools and help get innovative cross-department and cross-school projects off the ground. We envision the typical seed grant being sufficient to hire a post-doc for a year, which is often enough to turn an idea into meaningful progress on a joint project.

In addition to the innovative research stimulated by the seed grants, the seed grants are aimed at producing an important public good, a stronger communication network among Yale researchers. The seed grants would encourage the investment of time and energy in research conversations among potential collaborators from different parts of the campus. These cross-department and cross-school conversations and collaborations will in turn produce a stronger network among data-intensive social scientists at Yale, leading to a more efficient flow of expertise and information across Yale's departments and schools. A strong network is especially valuable during this period of rapid methodological innovation because it will help to ensure that all of Yale's research communities benefit from timely exposure to new data sets, analytical techniques, research designs, and applications.

We recommend the center administer a program to provide meaningful internal grants to support data-intensive projects, awarded through an Request for Proposal (RFP) process. Projects might request funding for shared personnel, data acquisition, hardware, or other project related expenses. To build community and connections across Yale, we would very strongly encourage proposals that span departments and schools. To conserve University resources for important uses for which other sources of funds are not easily available, proposals in research areas with significant sources of external funding should be encouraged to describe a path to application for such funds. Seed grants may yield significant returns given growing funding for some areas of data-intensive social science research, including work at the intersection of artificial intelligence and social science.[2]

4. Web Portal and Outreach to Yale researchers

A common theme from our focus groups with Yale faculty was that they are unaware of many of the resources already available at the University and do not know how to access these resources. Therefore, outreach will be an important task for the center. Not only can the center provide information on the services it offers, but it could also become a repository for information that directs Yale researchers to other resources across campus. A web portal as well as a well-networked executive director and staff can provide this function. Additionally, a University-wide mailing list of researchers interested in data-intensive social science could be curated and used to share information on upcoming events and relevant grants, highlight new and innovative research, etc.

5. Workshops, conferences, visiting scholars, and fellows

The faculty input and Committee discussions reflect an understanding that the revolution in computational power and data availability, along with advances in data analysis techniques and the development of software that implements new analytical methods, has created exceptional opportunities for rapid advancement in knowledge across many domains. It is essential to Yale's position as a research leader that methodological innovation and expertise developed outside of Yale is made available to Yale researchers as it is being developed and applied and that this knowledge is shared rapidly among Yale's researchers. Multidisciplinary workshops, self-organized working groups (like those

[2] There appears to be some room for expanding Yale's social scientists' participation in the grant economy. Relative to peers, Yale is attracting fewer outside research dollars in the social sciences. According to NSF's Fiscal Year 2017 Higher Education Research and Development Survey, total R&D expenditures in the social sciences in 2017 were $2.55B (15% increase from 2014). Among the institutions surveyed in 2017, Yale ranked 74th, with $9.05M in R&D expenditures in all social sciences.

formed around the Berkeley D-Lab[3]), and conferences at the intersection of data science and social science would bring together researchers from across and outside the University who are interested in learning and sharing new data-intensive social science applications and methods and forming research collaborations. Inviting visiting scholars from universities and industry on campus would also build collaborations through an exchange of ideas. Graduate students from Yale and elsewhere might join the center as research fellows, attending events, presenting their research, and teaching workshops on new techniques that they are using in their work.

6. Host Specialized IT support

Faculty, especially those whose research involves significant data work, have sometimes found it challenging to work with Yale ITS support, citing long wait times and lack of specialized support. At some peer institutions, departments and units with similar IT needs are grouped together into service clusters, and IT personnel who service the clusters are hired according to the specialized needs of the clusters' researchers.

We understand that Yale's ITS is considering some ideas for revising IT service provision, including organizing workers into service teams for units with similar IT needs. This suggestion seems very promising. The high correlation among the needs of people working in the same area, given use of similar software, hardware, etc., would lead social science-specific IT specialists to get used to recurrent issues and be able to solve them more efficiently. If this direction is pursued, the proposed data-intensive social science center could serve as the base for an IT team dedicated to serving a cluster of departments and units with relatively heavy data and computation needs. Whatever adjustments in service provision are contemplated, we recommend that ITS work closely with the relevant departments or units to ensure that there is complete clarity on the skills and qualifications needed for faculty IT support and to involve faculty closely in the search process for any new staff.

Other Observations

The center represents the single highest-impact investment that we believe Yale can make to improve the data-intensive social sciences. During our deliberation the Committee logged some other observations about the field at Yale that we capture below for further reflection by the community.

Observations on current services at Yale today

Researchers praised the quality of the IRB staff, but in nearly all focus groups faculty raised concern about major research delays due to how long the review process takes. The IRB staff is currently stretched and will be stretched further as research in the data-intensive social sciences grows. Expanding the IRB's capacity and having designated staff members liaise with social science researchers through information sessions and regular office hours would be helpful, as would updating the web interface with simple instructions and easy-to-use templates.

[3] Examples of Berkeley D-Lab working groups: Computational Text Analysis Working Group, GeoMatters Working Group, Securing Research Data, and Social Media Research Working Group

The StatLab is the unit of the Yale University Library that provides basic statistical consulting and workshops on methods and software. The StatLab provides workshops on statistics and data gathering methods and analysis using software as well as approximately 400 hours annually of one-on-one consulting, 2/3 of which is provided to graduate students. It is headquartered at the Center for Science and Social Science Information, a library in the concourse level of the Kline Tower (KT). In contrast to the arrangement common at peer institutions, Yale's social science statistical support services are not co-located with either a major social science department or the central campus library. It is possible that the StatLab's location has influenced its activity level. The usage pattern for StatLab shows heavy use by the School of Forestry & Environmental Studies, which is located yards from KT, but less use by FAS undergraduates compared to use of similar services at peer institutions, such as Princeton's Data and Statistical Services (DSS).

As part of the Committee input process, we met with library staff who oversee the StatLab. We discussed several ideas for how to better align StatLab services with the growing needs of social science faculty, students, and researchers. Ideas include providing more statistical and data services in closer proximity to many social science departments, such as through significant expansion of service provision in the StatLab's Rosenkranz Hall satellite facility, a site in close proximity to large numbers of social science researchers and students. Following a model developed at Princeton, the StatLab could also provide enhanced remote consulting via email or self-help by building or linking to carefully curated online resources that thoroughly address specific commonly asked questions. To confirm that services are meeting the needs of social science researchers and to enhance accountability, the StatLab could begin to use a standardized and consistent method of collecting and reporting service usage and user-satisfaction with services and programming. Finally, to improve the alignment of StatLab services with both teaching and research, StatLab could form a steering committee of faculty that meets regularly to provide advice and oversight. Although it would be costly in faculty effort, the steering committee could consider whether basic statistical consulting should have a faculty director. Further, it would be worth seriously considering whether rather than expanding service provision in Rosenkrantz the Statlab should instead be moved to the proposed Data-Intensive Social Science Center, as there would be clear advantages in proximity and faculty oversight to moving it to the new center.

The University spends substantial amounts on the support services most relevant for data-intensive social science research. These research support units include the YCRC, IRB, Statlab, ITS, and OSP's grants and contracts staff. However, there do not seem to be sufficient systems in place to measure service quality and provide regular and impactful feedback to service providers on faculty and student experiences with these services. Our Committee's input process gave faculty an opportunity to voice praise and complaints and to offer suggestions. In the absence of this forum, these positive and negative comments would have otherwise gone unheard or would have been shared privately. As the University seeks to offer new services and improve existing ones, it will be beneficial for all service units to institutionalize the practice of measuring user satisfaction and sharing information about user satisfaction with the community. This will facilitate identification of specific issues and focus attention on the management of the critical outcomes, spot trends in service quality over time, and stimulate valuable discussions of researcher satisfaction with the services.

Observations on Yale versus peers

Yale appears to have fewer ladder faculty specializing in data-intensive, policy-oriented social science research than peer universities. In the Yale Economics department, for instance, there is currently strong faculty representation in some subfields, such as micro-economic theory, econometrics, economic development and trade. However, a large amount of data-intensive, policy-oriented social science is done by economists specializing in applied microeconomic analysis, especially those working in industrial organization, labor economics, and public finance. Joe Altonji, Professor of Economics at Yale, provided the Committee with a memo tallying the number of faculty members in these three fields in the economics departments, business schools, and policy schools at Yale and six peer institutions. According to Professor Altonji's assessment, Yale has similar strength to key peers in industrial organization, but in public finance and labor economics Yale has fewer faculty.[4] Another major center of data-intensive, policy-oriented work is the sociology subfield of stratification and inequality. According to a recent Yale sociology department self-study and analysis by the chair of sociology, Yale has fewer faculty in this area than peer institutions. The chair of political science reports a similar pattern of relative underrepresentation versus peer institutions in more quantitative research, especially among senior faculty. The Committee views the relative lack of faculty in data-intensive policy oriented social science to be an important barrier to Yale's research strength in this area and an issue worth further consideration.

[4] This exercise requires some judgment calls about assigning faculty to sub-fields and should be viewed as only a rough estimate of faculty strength in data-intensive policy-oriented microeconomic research. According to Altonji's April 2019 tally, Yale has 13 faculty in the sub-fields of public finance and labor economics versus an average of 27.5 at the most comparable peer institutions (Harvard (35), Chicago (38), Princeton (17), and Stanford (20)). MIT, which has very few undergraduate economic majors, has 13 faculty in public finance and labor economics, while Berkeley has 26. Yale's faculty numbers in Industrial Organization were counted as similar to the comparison set (Harvard (10), Chicago (9), Princeton (3), Stanford (7.5), MIT (8), Berkeley (11)).

Teaching

Overview

The Committee's charge states that: "A great university should be engaging in the great debates of its era, and our students—the leaders of tomorrow—should participate. But that engagement must be grounded in evidence-based inquiry and rigorous analysis of facts."

The Yale College graduate is frequently confronted with empirical claims as an organizational leader, thought leader, and community member. Yale College graduates will make the greatest contribution to society if they develop the habits of mind to ask probing questions about empirical claims, know the strengths and weaknesses of different common research designs, and understand how to incorporate evidence to form judgements and make decisions.[5]

Distinctive challenges to inference and prediction in the social sciences have led to methods of quantitative and theoretical analysis designed to address these challenges. One key feature of social science problems is that researchers must often use non-experimental data to determine cause and effect relationships. Measuring cause and effect in observational data is extremely challenging. There are sometimes difficult measurement issues, such as the imprecision and bias that arise from self-reported behaviors and attitudes. Observed patterns of behavior are frequently the product of choices made by individuals or the strategic decisions of organizations. Because these actions often reflect unobservable differences across the actors and these actions are also the result of other factors that are hard to observe or adequately measure and model, it is difficult to disentangle the observed correlations between actions and outcomes from the true causal effects of actions on outcomes. Experimental research can sometimes be used to measure causal relationships in social science, but in contrast to the physical sciences, how people respond to the "treatment" will typically depend in important ways on history, experimental conditions, or subtle differences in the study populations. Relationships found in experiments may therefore be highly context dependent. The analytical techniques and the empirical intuitions that are needed to account for these problems and the skills in identifying opportunities to design research that overcome these challenges differ from the experimental design and measurement skills developed in the physical and biological sciences. It is therefore appropriate to think of research design and data analysis for the social sciences as a related but separate and distinct domain of knowledge that is not acquired as a by-product of the study of mathematics or of experimental research in the physical and biological sciences.

Recent developments in technology and analytical techniques have made learning data-intensive methods a matter of special excitement and urgency. The revolution in computational power, data types and data availability, and analytical techniques have combined to create unprecedented opportunities

[5] The Committee was directed to identify what every Yale student should know about engaging in evidence-based inquiry and rigorous analysis of facts and how they might learn these things. This naturally focused our attention on the foundational knowledge that every college graduate should master. Although we gathered some input about the professional and other schools, we lacked the expertise and scholarly authority to productively engage in assessment and recommendations for these schools and therefore our fact-finding and recommendations are restricted to undergraduate education and social science graduate education. That said, the input we received suggests that many of the lessons we have distilled about foundational knowledge apply to students across the University.

for using data to understand individuals, social groups, and institutions. Researchers have increasing access to new types of data, including large scale administrative records, commercial transactions, text and image data and archives, location tracking, satellite data, and social media digital trails. These novel data, combined with new techniques in machine learning, text and image processing, and other methods, will produce new insights into how cities grow and the environmental consequences of different patterns of growth, how consumers make consumption and savings decisions, how social media changes political communication, where people spend their time and with whom they interact, how treatment effects in large scale experiments vary with subject characteristics and context, and how health and intergenerational class mobility is related to geography.

Data-intensive social science research offers a rare combination of practical application and intellectual excitement. It can illuminate complex challenges like poverty, health care, and climate change, and provide valuable insights into the consequences of different solutions. There is a growing demand for evidence-based policy knowledge not only in government, but also in the advocacy and NGO communities. At the same time, engaging in data-intensive social science offers remarkable opportunities for intellectual and personal growth. A structured empirical investigation is a confrontation between researchers' beliefs about how the world works and external reality. Whether a particular set of beliefs are confirmed or refuted is often less important than the ongoing process of learning and discovery. Data-intensive social science research thus imparts both confidence and a sense of humility. Researchers experience the possibility of intellectual progress and also gain a deeper understanding of the limits to our collective knowledge.

Yale College students vary in their level of interest and engagement with questions of empirical analysis and research design. The goal of the Committee’s recommendations is to provide undergraduate students with a range of accessible, engaging, and rigorous course and research opportunities to develop fundamental knowledge and skills in the data-intensive social sciences. For students who just want the basics, there should be engaging and intellectually rigorous courses that cover the fundamentals, such as YData and “signature” lecture courses. For students who want more exposure to data-intensive methods, there should be multiple opportunities for reinforcement of the key techniques and concepts, including applications of data analysis and research design in the student’s area of academic interest, for example, through a data-intensive section of a departmental course of interest. Students who seek to develop deep substantive knowledge of the domain areas of social science and to master the traditional literatures in these areas should also have many opportunities to develop data analysis skills and a chance to use these skills to study questions they care about. These students can complete further courses in statistics & data science to earn a certificate in data science or enroll in the pathway to advanced achievement in quantitative social science. Interested students can also gain exposure to research design, methods, and computing through numerous research assistantship opportunities on campus. For students who want to achieve advanced levels of mastery, there should be clear paths to achieving the highest level of proficiency. Yale College should provide training that is as rigorous and engaging as any available in the world for undergraduate students who are intensely interested in these topics.

We begin with a brief summary of our key findings on teaching and learning at Yale College. We then provide recommendations for enhancing the teaching and learning of the methods and applications of data-intensive social science at Yale College.

Summary of Key Findings

The Committee investigated key concepts and techniques related to data that faculty believe all undergraduate students should be familiar with, the landscape of introductory statistics and research design courses currently offered in Yale College, undergraduate student course taking patterns, fulfillment of the Quantitative Reasoning requirement, and learning opportunities outside the classroom. These were the key findings:

1. There is rough consensus among faculty on what concepts every Yale College student ought to know. These include central concepts in statistics and research design such as principles of probability, regression analysis, statistical significance, measurement and sampling error, causal inference, modeling, and common data collection issues. In addition, faculty highlighted the importance of numeracy, sophisticated quantitative reasoning, and intelligent evidence assessment.
2. There are numerous introductory statistics and research design courses in Yale College covering overlapping material and there are no structures in place to ensure that there is coordination among instructors or identification of gaps in level or material covered.
3. Course taking patterns suggest that many undergraduate students may be getting only minimal exposure to the key concepts and methods of empirical inquiry. After reviewing course patterns for "first courses" in statistics and research design (which are prerequisites for more advanced courses), it appears that 28% of Yale undergraduates take no courses that are centrally focused on data analysis techniques or statistical methods and only 24% take more than one such course.[6] Most Yale College social science students take the minimum number of statistics courses required for their major (typically 1 but sometimes 0).
4. The Yale College quantitative reasoning requirement does not appear to be leading most students to do substantial data-intensive coursework. The Committee looked at students in the Yale College Class of 2018 who took no more than two QR courses during their time at Yale, i.e. met the minimum QR requirement. 43% of these students took no stats courses to fulfill their QR requirement. Based on these students' course-taking trends, the QR requirement is most commonly fulfilled by Econ 115: Introductory Microeconomics and is often fulfilled by non-statistical math courses.
5. There are excellent learning opportunities in data-intensive research outside the classroom, such as: the Tobin Undergraduate Research Assistantship program and the Herb Scarf Summer Research program in the FAS economics department, the Dahl Scholars program and the Director's Fellows program in Yale's Institution for Social and

[6] The Committee analyzed course taking patterns for students in Yale College Class of 2018. There may be some statistics and research design courses missed by this analysis and there may be some courses with substantial statistics components that are missed as well.

Policy Studies (ISPS), and numerous research assistant opportunities in the psychology department.

The Appendix contains a more detailed description of these findings.

Vision and Goals

All Yale College students should develop the habits of mind that will enable them to identify the strengths and weaknesses of empirical evidence, ask probing questions about empirical claims, and use quantitative evidence wisely in forming opinions and making decisions. All Yale College students who seek to achieve mastery of quantitative methods should have a clear path to reaching high levels of expertise. Our educational objectives should evolve as new skills and tools are developed and as research designs for data-intensive social science improve.

Goals:

I. **Offer basic courses for key ideas and methods of data-intensive research and analysis.** Yale College should offer accessible but rigorous courses that cover the fundamental insights about research design, reasoning about quantitative evidence, and using quantitative evidence in belief formation and decision-making. Every Yale College student should have an opportunity to experience the excitement of empirical investigation and discovery in content areas of academic interest to them. Students should have multiple exposures to the basics and multiple opportunities for developing and applying their data analysis skills. We should promote a culture of rigorous and wise engagement with empirical claims.

II. **Establish paths to advanced achievement**. These pathways should have multiple entry points and include a pathway for students who begin their education at Yale College without substantial prior exposure to statistics, programming, or advanced math.

III. **Strengthen communication, coordination, innovation, and assessment.** There should be coordination among instructors across campus, and students should know what is being offered. Yale College should periodically update its goals for teaching and learning and measure progress towards them. Undergraduate students should also be able to self-evaluate their levels of understanding and mastery and their progress. There should be mechanisms in place to learn about what is being done well at other universities.

The Committee's recommendations that follow are designed to support these goals. The faculty input process, focus groups, examination of existing courses and course taking patterns, peer benchmarking, and Committee discussions produced several promising ideas for enhancing the teaching of data-intensive social science in Yale College. There are suggestions for strengthening each of the levels from basic to advanced undergraduate training. The order of the recommendations reflects the Committee's priorities, taking into account impact, cost and ability to implement. The Committee recognizes that each of these recommendations will require evaluation through standard channels (e.g., faculty committees, consultation with DUSs and DGSs).

Recommendations

Recommendation 1: Enhance basic data-intensive course offerings in Yale College

Promoting the first goal of offering basic courses for key ideas and methods of data-intensive research and analysis, the committee proposes three recommendations: (1) establish data-intensive course sections, (2) create "signature" lecture courses in data-intensive social science, and (3) expand the YData course. These recommendations aim to offer students multiple, accessible, but rigorous, exposures to the fundamental insights about research design, quantitative reasoning, and belief formation and decision-making using quantitative evidence.

a. Establish data-intensive course sections

Students benefit from receiving direct support when learning how to work with and analyze data. This includes setting up computers and software, writing code, loading data, and step-by-step assistance with problem sets and problem solving. There was significant enthusiasm on the Committee to adapt the successful Yale College model of writing-intensive course sections for data-intensive course sections, which would require designated teaching fellows who teach fewer students, undergo additional training, and provide intensive feedback on undergraduate written work. Data-intensive sections would provide Yale College students with the opportunity for multiple exposures to data analysis with applications they care about. Yale College should consult with the S&DS department and explore the possibility of counting data-intensive sections as half credits towards the data science certificate. Implementing this recommendation would require faculty leadership and incremental teaching fellows (TFs). Yale College could start by piloting this program through the creation of five to ten sections to assess student demand and perfect the model.

b. Create "signature" lecture courses in data-intensive social science

There are currently few large, popular, "signature courses" in research design, the application of data-intensive methods to social problems, or the use of data in forming opinions and making decisions in Yale College. Through our focus groups with faculty, the Committee gauged high energy around developing "signature" lecture courses that would attract a sizable number of general background undergraduate students and would provide a first or second exposure to rigorous empirical social science inquiry, principles of research design, evaluation of evidence quality, and decision-making using data and probability. These courses would be interdisciplinary, and potentially team taught, general courses and not part of the standard course sequences in departments. Implementation would rely on faculty willingness to step up to design and teach these courses.[7]

[7] The single example at Yale we are aware of is Professor Woo-kyoung Ahn's course, PSYC 179 "Thinking," which is offered this fall semester and covers material related to these themes. The course provides "a survey of

The Appendix includes further examples of popular “signature” courses at peer institutions, such as Raj Chetty’s “Using Big Data to Solve Economic and Social Problems” at Harvard and Carl Bergstrom’s and Jevin West’s “Calling Bullshit: Data Reasoning in a Digital World” at University of Washington. These types of courses can teach important material in an accessible way and, when a sufficient number of students share the course experience, can also spark community, conversation, and culture around a discipline across the entire university.

c. Expand YData course

Launched in 2019, YData is designed to be a highly interdisciplinary, unintimidating, introductory data science course for students with little or no background in statistics, math, or computer science and who are not necessarily interested in pursuing these fields as their majors. The Committee recommends expanding YData to include additional connector seminars with a broad reach across humanities, social science, and science. Ideally, this course would become an interdisciplinary data science hub at Yale, attracting students from all academic disciplines. New connector seminars could leverage Yale’s existing strengths in areas such as law, English, history, economics, psychology, and medicine. The primary goal of the class is to teach students “the data science way of thinking” and prepare them to critically analyze the information they come across day-to-day in the news, scientific studies, and elsewhere.

Recommendation 2: Create a pathway to advanced achievement in quantitative social science in Yale College

Much of the practice of social science, whether in research, policy, or the private sector has rapidly changed in the past decades and now rests on a set of mathematical and quantitative tools and theoretical ideas that require significant time and dedication to master. There are some programs in place to provide the technical background for undergraduate study in quantitative social science, such as the economics and mathematics major. However, there is currently a lack of a broader and more intensive option that spans across the social science disciplines.

The Committee recommends establishing a pathway to advanced achievement in quantitative social science for undergraduate first years and sophomores. This would not be a major; however, it would bring together a cohort of first and second years who would take a collection of courses designed to provide rigorous preparation for future quantitative social science majors. There is good reason to expect student demand. Undergraduate social science enrollments are increasing, especially in quantitative areas, including global affairs, statistics and data science, and computer science. The last new development in the quantitative social sciences at Yale was the S&DS major, which increased the number of department majors from three to over thirty. There is also the precedent of the successful Northwestern Mathematical Methods in the Social Sciences program (more details on this program are

psychological studies on thinking and reasoning, with discussion of ways to improve thinking skills. Topics include judgements and decision-making, causal learning, logical reasoning, problem solving, creativity, intelligence, moral, reasoning, and language and thought." The lecture is offered this year for the first time and the course enrollment is over 400 undergraduate students, which suggests a huge latent demand for survey courses on these and related topics.

in the Appendix). Feedback from faculty and chairs suggests there is likely to be high undergraduate student demand for a challenging program in this area. This program could also be used as a recruiting instrument to attract highly motivated students interested in quantitative social sciences to Yale College.

The details and collection of courses for this proposed program would be worked out by a faculty committee in consultation with DUSs and DGSs and would be evaluated though standard channels. We strongly recommend that any program have entry points in both the first and sophomore year to allow undergraduate students to use their first year to deepen their preparation for the program. There is significant faculty enthusiasm for teaching highly motivated students, and we believe that constructing a robust program would likely require the university to develop only a few new courses to supplement existing ones. One proposed model would be: students take six courses, two per semester for three semesters, to acquire foundational knowledge in programming, computational methods, and statistics, along with intermediate level exposure to mathematical models applied across the social sciences. The program would be built around three themes: (1) mathematical and computational foundations, (2) empirical analysis and research design, and (3) mathematical representation of human behavior. The program would provide students a broad but very rigorous introduction to the tools and applications in quantitative social science early in their time at Yale College, position students to design and execute outstanding senior projects, and prepare students for graduate school or work as an RA with professors on research frontier projects. It would also provide excellent background training in analytical methods that can be applied in government, non-profit, finance, and industry jobs as well.

Recommendation 3: Expand the pre-doctoral program currently housed in the Tobin Center for Economic Policy

“According to a recent National Science Foundation survey of earned doctorates, less than 5 percent of PhD recipients in economics or related fields are underrepresented minorities.”[8] Yale currently has two pre-doctoral programs in the social sciences that strongly encourage applications from underrepresented groups: (1) the Tobin pre-doctoral program and (2) ESI-PREP. The Tobin pre-doctoral program is a one to two-year program aimed to provide education and training to individuals who are planning to apply to a PhD in Economics, or a closely related discipline. Pre-docs work as full-time RAs with faculty members who are primarily doing economic research, enroll in one course per semester, and attend weekly professional development and research seminars. ESI-PREP is a one-year program open to recent college graduates in all divisions (i.e., humanities, social sciences, sciences) with a strong desire to pursue a PhD. In the 2017-18 cohort, 3 out of 10 post-bacs were placed in a social science department. These programs have begun to address the need to diversify graduate programs; however, there is still additional unmet need.

The Committee recommends building on Yale’s existing Tobin pre-doctoral program to increase the number of slots dedicated to data-intensive social science outside of economics (e.g., psychology, political science, sociology, etc.). A major goal of this expansion is to provide additional resources to promote diversity and inclusion and to help address the underrepresentation of certain groups in data-intensive social science more broadly. We envision adding two additional pre-doctoral slots per cohort,

[8] https://gsas.harvard.edu/diversity/research-scholar-initiative

so four additional pre-doctoral fellows in any given year for a two-year program. Tobin pre-doctoral fellows receive compensation and half of that compensation is paid by Tobin while the other half is paid by faculty. The Committee proposes a similar setup for the additional pre-doctoral hires – the University would fund half of their compensation while faculty would continue to fund the other half. For the initial pilot, the University may consider collaborating with existing centers to partially fund the additional slots. Given that the initial infrastructure and leadership is already in place for the Tobin program, we believe building and expanding on it would be a relatively smooth and easily implementable process.

Recommendation 4: Establish annual meeting of intro stats and research design instructors and relevant Yale College DUSs to improve communication, coordination, innovation, and assessment

There are numerous Yale College introductory statistics and research design courses that tend to cover overlapping statistical concepts and techniques and differ primarily based on disciplinary focus (e.g., health, economics, politics). There are currently no structures in place to ensure that there is coordination among instructors. The Committee recommends that Yale College organize an annual meeting of relevant Yale College DUSs and instructors of introductory statistics and research design courses. This meeting would serve as a forum for departments and instructors to better coordinate concepts and techniques taught in the classroom, identify gaps in level or material covered, share best practices, and brainstorm innovative and collaborative teaching ideas. This group would also review course offerings and check that online course listings and descriptions are accurate. A work product of this annual meeting could be an updated annual guide contrasting the various introductory courses. This guide could be shared with the DUS of each FAS department and then disseminated to advisers to share with their undergraduate students.

More generally, it would be valuable for DUSs (and perhaps department chairs) to discuss course offerings to identify any statistics, data science, and research design courses that offer similar material in multiple departments, determine whether such multiple offerings stem from real differences in course objectives or a failure to coordinate, and make minor adjustments that would permit teaching a single course. Further, DUSs could discuss unmet student needs and explore whether departments could cooperate and offer new courses jointly to address student needs efficiently.

Recommendation 5: Consider appointing a committee to evaluate the Yale College quantitative reasoning (QR) requirement

Beginning with the Class of 2009, Yale College students have been required to complete two courses in each of three disciplinary areas (humanities, natural science, and social science) and fulfill skills requirements in foreign language, writing (two courses), and quantitative reasoning (two courses).[9] Students fulfill the QR requirement with a range of courses, with many being not data-intensive in nature. After discussions with faculty from across the University, it seems crucial for all Yale College

[9] https://science.yalecollege.yale.edu/academics/faculty-resources/qr-courses-without-prerequisite/qr-courses

graduates to leave with the basic statistical and research skills and habits of mind to ask probing questions about empirical claims, understand the strengths and weaknesses of different research designs, evaluate the strength of evidence, and understand how to incorporate evidence to form judgements and make decisions.

The Yale College Dean should consider appointing a committee to evaluate how the QR requirement is functioning and to consider either changing the requirement or adjusting its implementation given the increased prominence of computation and data-intensive analysis. The committee should also explore how peer institutions approach requirements around baseline level of exposure to data analysis and statistical reasoning. For example, Harvard recently implemented a quantitative reasoning with data (QRD) requirement, replacing its past empirical and mathematical reasoning requirement. The purpose of the new QRD requirement is to ensure undergraduates "reach a level of quantitative facility involving mathematical, statistical, and computational methods that will enable them to think critically about data as it is employed in fields of inquiry across the FAS."[10] For more detail on this requirement and similar requirements at peer institutions, please refer to the Appendix.

Other Observations

The surveyed faculty identified key skills and concepts that every Yale College student should know, but it is not clear how to determine whether undergraduate students develop these skills. Beyond placement exams, Yale College does not administer tests outside of courses to measure skill attainment. One way to measure progress toward our teaching goals is to develop some tools for self-assessment. As this could both measure attainment and help guide a student's course selection, it would be useful for such tools to be available to students well before graduation. One Committee member stated that perhaps an undergraduate student entering their senior year might take a self-assessment and conclude that they ought to build up their quantitative reasoning skills before graduation.

Based on suggestions from social scientists drawn from every department in the FAS social science division, the Committee notes that there is interest in further exploration of establishing an interdisciplinary program in data science and computational social science for PhD students at Yale. The program would provide deeper training in data science and computation to interested social science PhD students and would lead to a graduate certificate in data science. The program would be cross-disciplinary and would create synergies by bringing together students from across the social sciences while still training these students on how to incorporate the analysis into their own discipline-specific research. Yale could be exceptionally well suited for such a program, with its strengths in statistics, computer science, data analysis and econometrics, as well as its traditional strengths in economic, political, network and social science. Existing centers such as the Cowles Foundation, ISPS, the Economic Growth Center, and the Yale Institute for Network Studies would create research opportunities and an enabling research environment for students and scholars.

[10] https://www.harvardmagazine.com/2019/04/harvard-course-requirements-quantitative-reasoning

Organizational Structures and Behaviors

Overview

In the Charge for the University-Wide Committee on Data-Intensive Social Science, Provost Ben Polak asked the Committee to "make suggestions about organizational structures and behaviors that could support data-intensive social science at Yale, [particularly] recommendations about mechanisms for better coordinating across Yale to improve efficiency, innovation, and impact, and mechanisms for rapidly learning relevant developments and innovations occurring at other universities."

Several of the Committee's recommendations in this report aim to promote Yale's efficient use of resources and spur innovation. In addition to the research and teaching recommendations, above, the Committee has two recommendations specific to organizational structures and behaviors.

Recommendations

Recommendation 1: Establish a University-wide committee to share information about data-intensive social science

The anticipated infrastructure needs of data-intensive social science suggest that the benefits to interdisciplinary and cross-school collaboration will continue to grow. Therefore, it is imperative that the University establish mechanisms to facilitate communication and coordination across the University's various centers, departments, and schools dealing with data-intensive social science.

Yale should consider establishing a twice annual meeting of social science center directors, department chairs, and other academic leaders who are most involved with data-intensive social science research. This would bring together approximately 15-20 people who have significant oversight responsibility in this area. These meetings would be for sharing information, coordinating plans, and providing advice to university administration and service units. We recommend that for at least one of these meetings each year, the key service providers such as ITS, YCRC, OSP, Yale Library, are included. This meeting could be used to discuss faculty and for faculty to provide advice and reactions to service unit plans.

Recommendation 2: Learn about developments at other universities and industry-leading organizations

The revolution in computational power and data availability, along with advances in data analysis techniques and the development of software that implements new analytical methods, has created exceptional opportunities for rapid advance in knowledge across many domains. Given the dynamic and rapidly growing nature of this area, it is impossible to predict what specific topics will be the most exciting for future direction. For Yale faculty and students to surpass the research frontiers of their disciplines, it is important that the University stay up to date with innovative developments in data-intensive social science among industry-leading organizations and peer universities. Learning from others' mistakes and successes will help Yale implement new policies and programs to better promote research and teaching in data-intensive social science.

We propose that Yale's Office of Institutional Research and Strategic Analysis (OIR/SA) produce a periodic memo on "Innovations and Lessons," perhaps annually, on major developments improving data-intensive social science research and teaching at industry-leading organizations and universities. The memo would describe important infrastructure investments, important changes in data policies and data availability, and key programs being started or discontinued at peer institutions. This memo would be sent to Deans, Center Directors, and social science Chairs, and also shared with faculty so that innovative developments regarding research support and teaching in data-intensive social science are common knowledge among interested members of our community. Based on this research, OIR/SA might propose one or more faculty site visits each year to places that seem especially innovative or cost-effective.

Relation to the University Science Strategy Priorities

In 2017, the University Science Strategy Committee (USSC) was formed to make recommendations for Yale's scientific research investments over the coming decades. The USSC identified five areas of science for strategic investment. For three of these areas, Integrative Data Science and its Mathematical Foundations, Neuroscience, and Environmental and Evolutionary Science, the tools, empirical insights, and theoretical models of the social sciences are important resources for intellectual progress. DISSC Committee members believe that it would be valuable to invest, where relevant, in social science areas adjacent to these science priorities to enhance Yale's leadership in these areas to draw on the relevant methods and theoretical perspectives of the social sciences. This suggestion applies to all three of the priorities mentioned, but we will focus our discussion on the priority which most clearly intersects with the DISSC's charge, the USSC's Data Science recommendation.

No discipline has exclusive claim to data or data science, so we emphasize the value to Yale of intellectual interchange among the broadest possible community of scholars. The breadth of data science suggests that each community will have its own specific needs and priorities, and we recognize that the needs and contributions of the social sciences are not identical with the needs and contributions of the natural sciences and engineering and applied sciences. Hence, DISSC believes we must invest in priorities specific to the social sciences, while social scientists also participate in the cross-discipline data science efforts described by the USSC.

A broad approach to data science would position the University to excel at basic research that advances the methods of data science, applied research that uses the tools of data science to advance disciplinary and interdisciplinary research programs, and applied research on the societal consequences of the revolution in computation and related technology. From the standpoint of the social sciences, we might organize into three categories the ways that our set of opportunities and challenges are changing due to technological developments that have followed from advances in analytics and computation:

- ***accelerating social science research*** using data science,
- ***understanding and managing the social impact*** of data science and technological change (this category centers on the societal changes occurring now), and
- ***reimagining core human activities*** through data science and related technologies (this category centers on what new things are made possible as the result of advances in data science).

For purposes of this discussion, we will group these latter two categories under ***computation and society***.

Yale's social sciences are well-positioned to make important contributions in each of these areas. Yale's Faculty of Arts and Sciences (FAS) is home to some of the world's most creative and productive social scientists and includes departments of economics, political science and psychology that rank among the very top departments in the world. The FAS also includes a rapidly expanding department of statistics and data science, and excellent departments of sociology, linguistics, and anthropology, as well as a top-ranked department of history. Yale has relevant excellence in its professional schools, including the nation's top-ranked law school, and an outstanding business school, school of medicine, school of public health, school of nursing, and school of forestry & environmental studies. The new Jackson school of

Global Affairs will provide additional research strength. This extraordinary collective strength produces a concentration of scholars that can reasonably aspire to a position of excellence that provides Yale with one of the best social science faculties in the world. Maintaining and enhancing this excellence will require us to engage with the research opportunities presented by data availability and technological innovation.

Theme 1: Accelerate social science research using data science

Advances in data sources (e.g. digital media trails, administrative data, transactions data, text and image archives, sensitive and restricted use data sets, remote sensing data, location tracking), computational power, and analytical methods (including advances in machine learning, natural language processing, and image processing) are transforming how we study traditional questions at the heart of the social sciences. The USSC report contains an excellent discussion of the ways that new data sources may propel research in the social sciences. Faculty and students who seek to reach and advance the research frontiers in their respective disciplines using these rapidly developing tools and resources will require support to enable them to work on novel problems and applications. This research will in turn produce new analytical paradigms and spur the development of new data science tools for further social science applications.

The proposed Center for Data-Intensive Social Science is designed to support efforts to take full advantage of these research opportunities and fits in well with the emerging university organizational schema of forming centers that are devoted to advancing research at the intersection of data science and some large segment of the intellectual landscape. It would complement the USSC's proposed Institute for Integrative Data Science and its Mathematical Foundations and other existing centers and institutes such as the Center for Biomedical Data Science, Quantitative Biology Institute (QBio), and the Digital Humanities Lab, while addressing the unique needs of data-intensive social science.

Theme 2: Computation and society

The USSC report notes that "the world is currently undergoing a data revolution comparable to the industrial revolution in its potential impact…Not a single aspect of society today will be left untouched by the data revolution." We agree and highlight the reference to impact on society. The social sciences engage centrally with these issues, and DISSC recognizes at least two vectors along which the social sciences can collaborate on an endeavor complementary with the USSC data science priority.

The first vector is understanding and managing the societal impact of technology. Data science and related technological developments in computing and artificial intelligence are transforming the human environment and society. We are experiencing rapid change in communications, politics, work, and markets, with enormous consequences for psychological well-being, economic performance, social equality, identity, and governance. These are among the most pressing issues of the day, and engagement with these issues is both an intellectual challenge and, at the university level, a social responsibility.

Although there is no coordinated University-wide initiative in place at Yale, some initial efforts to address the social and individual impact of the technological revolution are already underway here. For example, substantial projects at the law school such as the Information Society Project (ISP), which supports a community of interdisciplinary scholars exploring issues at the intersection of law, technology, and society, and the Social Media Governance Initiative (SMGI), which explores social media companies' responsibilities in maintaining conditions and values that are necessary for democracy (e.g., civil discourse, respect for others, and health community) are engaged in community building and research that is shaping how we think about the impact of advances in technology on society. In addition, researchers at the school of management have recently formed the Thurman Arnold Project to study the competitive features of digital platforms and the consequences and regulation of concentration in the tech industry. Faculty from across the FAS have begun an initiative on computation and society to promote research and discussion at the intersection of technology and society. An initial informational meeting of this group, led by Elisa Celis and Nisheth Vishnoi, drew approximately 25 faculty members from across the University. The potential interest at Yale in this broad area is further suggested by the reception of the Spring 2019 Workshop on AI, Ethics, and Society, which was organized by Nisheeth Vishnoi (Computer Science), Jack Balkin (YLS), Elisa Celis (S&DS), and Zoltan Szabo (Philosophy), and gathered some thirty faculty from four schools (FAS, SOM, YSM, and YLS) and 15 departments for panels on AI's Impact on Society, AI and Morality, AI and the Legal Sphere, and AI, Ethics & Society at Yale.

The second vector is reimagining society using the tools of the technological revolution. The disjunction between the gradual evolution of institutions and practices, and the rapid change in technical feasibility, opens a space for radical reinvention of key social activities. Rethinking of status quo institutions and practices is possible in the wake of the technological revolution. Consider the enduring features of human society, such as educating the next generation, caring for the sick and the poor, organizing work and exchange, forming teams, choosing leaders, making decisions about the community, and sharing ideas. The ways that we perform these core activities have been built up slowly, over many generations, in response to changing values, local experimentation and accumulating experience, and technical constraints. We are now experiencing rapid, perhaps unprecedented, technological advances which dramatically relax the technical constraints on what is possible. The dimensions of this epochal change in technical possibilities are broad-reaching. There is an ongoing revolution in data production, communication, computing, analytical tools, and automation of human capabilities. Point-to-point, group, and mass communications across vast distances that would have been slow and expensive if not impossible are now free and instantaneous. Tasks that might have required large armies of workers can now be automated, and text and images that would have required thousands of years to view and understand can now be processed in an instant. Data that would have been infeasible to collect and store can now be easily obtained and stored at minimal cost. Patterns that would have been impossible to detect can now be discovered.

There are substantial efforts to build research capacity on the human impact of the computational revolution at many peer institutions, including Stanford, MIT, and Berkeley. In the Appendix, we describe these initiatives. Efforts by Yale to advance a University-wide initiative on computation and society would require planning to identify Yale's priorities and comparative advantage.

We underscore again that these are suggestions for further consideration and development as one possible elaboration of the USSC priorities, rather than DISSC's recommendations. Assessing how Yale might engage with the social impact of computation is a matter of concern to all disciplines and schools and therefore beyond both the scope of DISSC's charge and the range of expertise of the DISSC's members. That said, we believe that Yale will not remain a center for innovation and excellence in data-intensive and policy relevant social science research if it fails to play a significant if not leading role in engaging with the technology-led transformations and opportunities that characterize our era. If this direction is of interest, we recommend as a next step for the Provost to appoint a working group to explore how the University can lead in this area.

Appendix

1. Committee Charge

Charge for University-Wide Committee on Data-Intensive Social Science

From Provost Ben Polak
December 7, 2017

President Salovey has identified data-intensive social science as a top academic priority for Yale (*see* University Priorities and Academic Investments). This committee will play a key role in investigating and guiding to this priority.

Social science at Yale is quite strong, thanks to investments we have already made in people, programs, and facilities. It takes place across the campus: in departments in the FAS, in the Law School, the School of Management, parts of Forestry, Public Health, and elsewhere.

The application of data to public policy questions – to the great issues of the day – is an area that spans schools and departments, and that would complement our existing strengths. A great university should be engaging in the great debates of its era, and our students—the leaders of tomorrow—should participate. But that engagement must be grounded in evidence-based inquiry and rigorous analysis of facts.

The first task of this committee is to gather input from faculty across the university and to take inventory of our current resources and strengths that could support data-intensive, policy-relevant social science. The committee should also look outside of Yale to understand how other universities are responding to similar challenges and opportunities. The committee should examine the potential for progress in this area for Yale, and what progress might mean.

Then, keeping in mind our missions of teaching and research, I ask that the committee

1. Establish the key priorities in data-intensive social science for the next decade. These could include common courses or particular sets of skills for students, shared resources (analogous to science cores), or other initiatives. Please assess each idea in terms of impact, resources required (funding, space, faculty, etc.), feasibility, and whether Yale has a comparative advantage.
2. Develop prioritized lists of ideas that could be accomplished at current levels of resources, as well as those that would be possible with an additional $2m or $4m in annual expenditures.
3. Make suggestions about organizational structures and behaviors that could support data-intensive social science at Yale. In particular, I would appreciate recommendations about mechanisms for better coordinating across Yale to improve efficiency, innovation, and impact, and mechanisms for rapidly learning relevant developments and innovations occurring at other universities.

4. Consider what we should we expect our students to know - or at least to have the opportunity to learn – in this area. What are we in fact teaching them, and how are we teaching it?

Finally, I ask each member of the committee not to think of themselves as representing their particular subject area, school, or department but instead as representing Yale, to take a long-range and University-wide view. I appreciate the creativity, wisdom, and institutional citizenship that this will require, and I thank each of you in advance. I look forward to working with you - and learning from you – on this important undertaking.

2. Research Infrastructure

Peer Examples for Research Infrastructure Recommendations and Observations

1. Center for Data-Intensive Social Science

One example to consider is Harvard's Institute for Quantitative Social Science (IQSS). IQSS was founded in 2005 and is now the university's largest social science research center. It is led by an enthusiastic faculty director who has a vision of IQSS building cutting edge social science infrastructure, fostering an interdisciplinary community of social scientists, and facilitating research to solve some of the greatest problems affecting society. The Institute provides a broad array of service, including: collaborative spaces, seed grants, consulting, workshops, core tech support, research computing environment, and more. These are services that we are considering for our proposed Center for Data-Intensive Social Science. IQSS additionally offers data science services (e.g., research project planning, software training, statistics, visualization, secure storage), data curation services, a data repository, and specialized consulting. An example of an excellent unit within IQSS is Harvard's Center for Geographic Analysis (CGA), which consists of four full-time GIS experts who have either an M.A. or PhD background. The CGA is managed by a separate faculty director and provides specialized research services for faculty and students in addition to introductory workshops. The unit charges a subsidized hourly rate for its services. These additional services may be areas the University wants to consider investing in down the road if there is sufficient need and demand.

2. Seed grants

Columbia's Data Science Institute has the Seed Funds Program that funds up to five projects, up to $100,000 annually, for a maximum of two years. The purpose of the program is to encourage and support novel proposals at the intersection of data science and other domains that bring together researchers from different disciplines across campus. The Institute seeks proposals that can be developed and ideally submitted to government, industry, or foundations for external funding in the future.

Human-Centered Artificial Intelligence (HAI) at Stanford awards up to 25 grants of up to $75,000 each. The seed funding program is in its second year. The purpose of the grants is to support innovative, ambitious, and interdisciplinary research in Human-Centered Artificial Intelligence. HAI encourages proposals involving collaborations of faculty and students across different departments and/or schools. Stanford also has an Environmental Ventures Fund through its Woods Institute for the Environment that provides seed grants, from $5,000 up to $200,000 over two years, for interdisciplinary research projects related to the environment and sustainability among faculty who have not previously worked together.

Additionally, Stanford Bio-X has a successful Interdisciplinary Initiatives Seed Grants Program (IIP), which awards two-year seed grants of $200,000 per project to collaborative projects in areas related to bioengineering, biosciences, and biomedicine. Since the program's inception in 2000, Stanford has awarded seed grants to 212 interdisciplinary projects involving over 360 faculty from five Stanford schools and dozens of departments. The seed grants awarded since 2000 have resulted in over $270 million in external funding awarded to the university, a tenfold return.

Stanford's Social X-Change Accelerator is another program making significant investments to generate and scale up collaborative research, in partnership with the public, private, and social sectors, addressing concrete social problems. The level of support is beyond the "seed grant" level. Its mission is "to push the frontiers of social science and to craft solutions and policies for challenging societal issues such as economic opportunity, polarization, and ineffective institutions."[11] Social X-Change "is envisioned as a platform that will: (a) connect talented researchers and practitioners around concrete problems, (b) house, resource, and support these partnerships, (c) support the collection of evidence to identify potential interventions, (d) support the launch and evaluation of new interventions, and (e) scale promising solutions. These partnerships will span substantive areas including poverty and inequality, health, education, criminal justice, the environment, economic opportunity, and governance, polarization, and civic engagement, with local, national, and global reach. These partnerships will take the form of 'impact labs' and [the goal] is to support up to 20 impact labs working on different issues with five-year, multi-million-dollar commitments."[12]

Penn State offers several types of funding to support faculty research. Its review process is evaluative (seek excellent proposals) and developmental (support faculty to initiate research programs that can attract external funding). Approximately 40% of level 1 and level 2 external grant proposals submitted receive funding. A detailed description of the different funds available are listed below:

- Level 1
    - $500-$5,000 for 6-12 month period
    - Criteria for review: interdisciplinary collaboration, well-articulated plan of activities, team of investigators who range in seniority and experience, clear contribution to social science
- Level 1 RDC Funding (for research that will be conducted in Penn State's Census Research Data Center)
    - $500-$10,000 for 6-12 month period
- Level 2
    - $5,000-$20,000 for 12-24 month period; can cover salary replacement cost of $7,500 for one course buyout

[11] https://ourvision.stanford.edu/vision-initiatives/research
[12] https://www.timeshighereducation.com/unijobs/en-us/listing/174361/administrative-associate-/

- SSRI co-funded faculty: Co-funded faculty members may be junior or senior faculty members; all have demonstrated research expertise in strategic areas identified by SSRI; once SSRI has selected an area/areas of strategic research activity in which to hire, an announcement is developed and sent to relevant colleges asking department heads to develop proposals to create a position in their department in the identified area(s) of interest.; proposals reviewed and rank-ordered by SSRI Steering Committee, who then submit them to the SSRI Directors. The Directors then consult with the SSRI Advisory Committee (comprised of the Vice President for Research, and the Deans from the Colleges of Agricultural Sciences, Education, Health and Human Development, The Liberal Arts, and the Vice President for Research in the College of Medicine). Final decisions are then made; once a proposal is approved, SSRI will provide up to $3000 to departments to assist in recruitment efforts. SSRI then commits to paying up to 50 percent of the start-up costs as well as up to 50 percent of the salary (renewable after a successful SSRI review every five years)
- SSRI Faculty Fellows Program: (1) Mentored Fellowships provide funding for a faculty member for up to two course releases during an academic year (up to $7,500 per course or the equivalent for those who do not have resident instruction responsibilities) for study and training in new research areas with the guidance and support of a mentor or mentor team. The mentor/mentor team will also receive up to $1,000 in summer supplement; (2) Collaborative Fellowships provide funding for a new team of faculty members for up to three course releases (up to $7,500 per course, and no more than one release per faculty member on the team) to develop a novel, interdisciplinary project
- Commonwealth Campuses Research Collaboration Development Program
    - Faculty can submit research proposals that will require access to Penn State's shared facilities
    - In 2018, 20 awards of up to $10,000
- Consortium to Combat Substance Abuse
    - Community Fellows Program: provide funding for tenure track faculty members for up to two course releases across one or two academic years (up to $7,500 per course release); also eligible to apply for Community Collaboration funds (up to $5,000 for community activities; faculty teams of two can divide the course release funds between them
    - Seed funding: at least $100,000 of funding is available through this seed grant solicitation via the SSRI's Level 1 and Level 2 mechanisms
    - Strategic hires: 12 new tenure-track faculty members over next 4 years (national search began in Fall 2018), search through departments (in 2018, had 5 dept. searches: bio-behavioral health, human development & family studies, neural & behavioral sciences, psych, sociology & criminology)
- Frances Keesler Graham Early Career Professorship: provides supplemental funding to social and behavioral science faculty members at Penn State who are working in the interdisciplinary field of developmental neuroscience; award will rotate every three years, providing seed money (approximately $20,000 per year) for innovative research projects and programs

3. Reorganized University IT support

Economics is a data-intensive field and useful to illustrate the IT approach of peer institutions. There are several distinctive characteristics of the services provided at MIT Economics, which was cited as a place with superior services. The department of about 40 faculty has 3 full-time IT professionals who assist with both desktop and HPC: (1) Andrew Dormer, Sr. IT Operations Manager with a Graduate Certificate in Information Security, a B.S. in Information Science, and over 13 years of IT work experience; (2) Mark Leary, IT Manager with a B.S. in Computer Science and over 17 years of IT work experience; and (3) Carl Anderson, System Analyst. In addition to providing desktop support, these individuals maintain the computing infrastructure behind the scenes, including email, web and database services, and research computing.

Harvard's Economics department has a dedicated IT desktop support staff person who is part of the Social Science Division Administrative Service Group. Stanford's Economics department until recently had an in-house desktop support staffer, but the department switched and now participates in IT desktop support team coverage (a team supports a cluster of departments). Princeton has a dedicated IT support person with an office in the Economics department.

4. Observation regarding Statlab

There may be lessons to learn from the specific practices at other schools. There are some features of Princeton's Data and Statistical Services (DSS) that are worth careful consideration. The Princeton consulting model is structured as follows. The largest volume of consulting occurs in a group setting. Students fill a room (there is capacity for 20-25 students at a time and it is common that all spots are taken) and work at computers. They get advice from consultants who circulate around the room. Students talk with consultants, try to implement the advice, and ask more questions if (as) they arise. There is also an email consulting service which guides researchers to the extensive library of on-line responses that DSS staff have prepared for frequently asked questions. For more complicated issues, there are office-hour consulting sessions. In addition to graduate student consultants, Princeton DSS is staffed by a full-time consultant, who has a PhD in political science and over twenty-five years of work and research experience in the public, private, and academic sectors.

5. Measuring User Satisfaction

One example to consider is the feedback process Harvard's Institute for Quantitative Social Science (IQSS) uses for its desktop support services. IQSS uses a Request Tracker ticketing system, similar to Yale ITS, through which clients receive a notification email when a ticket has been created and resolved. Within the resolution notification email, users have the option to share their experience. Here is a sample email generated by the ticketing system:

"According to our records, your request regarding …. has been resolved. If you have any further questions or concerns, please respond to this message.

HOW DID WE DO?

1. If you are completely satisfied with the resolution of this issue, and how HMDC staff handled it, read no further and take no action.
2. If issues remain that we can help with, please respond to this message with details and we will continue to work on it.

3. If this issue has been resolved, and you would like to comment on the resolution or your interaction with our staff, HMDC management would appreciate you clicking the link below to answer a few questions.

If you can’t view the link above, here are the questions….

HMDC Trouble Ticket Feedback

Please Note: If you want your feedback to remain anonymous: select 'anonymous@nowhere.com”

1-Please provide your email address: From the drop down menu client has the option to choose (their own email address or anonymous@)

2-How satisfied were you with the resolution of your issue? From the drop down menu client has the option to choose (Completely Satisfied, Satisfied, Unsatisfied, Completely Unsatisfied)

3-How satisfied were you with your interaction with HMDC staff? From the drop down menu client has the option to choose (Completely Satisfied, Satisfied, Unsatisfied, Completely Unsatisfied)

4-If you have suggestions for how we might improve our service, please describe these here. Any details you could provide would be appreciated. (Just a blank field)”

3. Teaching

Key Findings

1. There is a rough consensus among faculty on what concepts every Yale College student ought to know.

DUSs and instructors of data-intensive courses were asked, “What concepts and techniques related to the collection, use, analysis, and interpretation of data should every student graduating from Yale College be familiar with? (We use familiarity to mean not minimal acquaintance but a level of understanding that implies a reasonable degree of sensitivity, maturity, and sophistication)”. Faculty identified the central concepts in statistics and research design including: probability, regression analysis, statistical significance, measurement and sampling error, causal inference, modeling, and common data collection issues (e.g., bias, missing data). We also conducted focus groups with faculty who taught data analysis and research design courses. Faculty focus groups highlighted skills beyond the central technical concepts in statistics and research design and emphasized the need for courses that produced numeracy, sophisticated quantitative reasoning, and intelligent evidence assessment. At the very least, Yale College students should be equipped with the skills necessary to critically evaluate the meaning and validity of statistics and empirical findings presented in their daily lives and in the news. Table 1 summarizes the specific feedback on this set of concepts and techniques.

Table 1

What concepts and techniques related to the collection, use, analysis, and interpretation of data should every student graduating from Yale College be familiar with?	
Concept/Technique	**Detail**
Understanding different types of biases in data	Students should be able to evaluate the trustworthiness of sources. Students should understand common data collection issues, such as bias induced by survey question wording, subjective coding, non-random sample selection, and missing data.
Grasping key concepts in statistics and probability	Students should have a strong grasp on key concepts such as mean, median, variance, standard deviation, distributions, basic probability, and Bayes' Rule. Students should understand how to interpret results of things like medical tests in light of prior frequencies.
Gaining comfort with core methods in statistical analysis	Students should be comfortable with different types of regressions (e.g., linear, logistic, etc.) and when to use each, understanding the difference between dependent and independent variables, interpreting regression coefficients, etc. Students should also be familiar with various methods of statistical hypothesis testing (e.g., t-test, non-parametric tests) and have a numerical and intuitive understanding of concepts tied to statistical significance (e.g., p-values, confidence intervals, uncertainty). It is essential that students develop the ability to correctly interpret statistical tests and critically evaluate statistical claims. They should also understand the limitations of significance testing. Our graduates should be able to intelligently process/consume this material and communicate it to an audience possessing various levels of technical familiarity.
Designing and conducting research	It is key that students are able to translate a policy or theoretical question into an informative empirical inquiry in a systematic way. Students should be able to formulate a hypothesis or research question, identify an appropriate source

	of data that would test the hypothesis, extract relevant measures from the data, analyze them appropriately, and finally write up the results in a clear and compelling way.
Sampling	All graduates of Yale College should understand the difference between completely describing a population versus using a sample. They should also understand the consequences of samples, including uncertainty and selection issues.
Differentiating between causation and correlation	Students should understand the distinction between a causal relationship and correlation. They should understand the difference between observational and experimental work and resist causal interpretations of descriptive quantities.
Visualizing data	All students should be familiar with basic but effective data visualization techniques. They should be able to communicate their own results through graphs and charts, accurately interpret published visualizations, and be wary of common pitfalls.
Programming	Learning to code is critical for data literacy. Therefore, every student should have basic competence in computer programming.
Assessing counterfactuals	Students should have an understanding of the roles and differences between data description and assessment of counterfactual quantities. The latter includes so-called "causal effects" but is not limited to this. All students should understand that the latter (1) is required for using data to answer any question beyond description--beyond "what happened" or "what do we see" -- and therefore necessary for answering most empirical questions in social sciences; and (2) requires an abstract framework (a model) within which one can define the counterfactual -- e.g., what would have happened had the treated group not received treatment. All students should develop familiarity with a variety of abstract frameworks/models and with application of appropriate methods that allow identification and estimation of the quantities of interest in such models. The most appropriate models and statistical tools will naturally vary by field.
Learning about Randomized Control Trials	Students should learn about the Randomized Controlled Trial as a procedure for “creating”

	data that are “designed”. Learning about this tool provides an opportunity for students to draw out the following points that are critical for data literacy: (1) descriptive questions are different from causal questions; (2) the definition of causation requires counterfactuals, (3) summaries of data (estimates) are different from the things we want to know (estimands), (4) uncertainty attends to all estimates; (5) larger sample sizes mean less uncertainty. In a course that teaches about RCTs, students can also encounter these fundamental concepts: how to summarize a distribution (mean, median, mode, standard deviation, variance), null hypothesis significance testing, and correlation doesn't imply causation.

2. There are numerous Yale College introductory statistics and research design courses covering overlapping material and there are no structures in place to ensure that there is coordination among instructors or identification of gaps in level or material covered.

Table 2

Course	Description
1. S&DS 100 Introductory Statistics	An introduction to statistical reasoning. Topics include numerical and graphical summaries of data, data acquisition and experimental design, probability, hypothesis testing, confidence intervals, correlation and regression. Application of statistical concepts to data; analysis of real-world problems.
2. S&DS 101-105 Intro Stats	Each of these courses, led by an expert from the field of study (life science, political science, social science, medicine), introduces statistical reasoning and emphasizes how statistics is applied to the particular discipline. Topics include numerical and graphical summaries of data, data acquisition and experimental design, probability, hypothesis testing, confidence intervals, correlation and regression. Students will learn to apply statistical concepts to data using Minitab and reach conclusions about real-world problems.
3. ECON 131 Econometrics & Data Analysis*	Basic probability theory and statistics, distribution theory, estimation and inference, bivariate regression, introduction to multivariate regression, introduction to statistical computing.

4. ECON 135 Intro Probability & Stats	Foundations of mathematical statistics: probability theory, distribution theory, parameter estimation, hypothesis testing, regression, and computer programming. Recommended for students considering graduate study in economics.
5. GLBL 121 Applied Quant Analysis	Mathematical fundamentals that underlie analytical approaches in public policy and the social sciences. Statistical approaches include descriptive statistics, principles of sampling, hypothesis tests, simple linear regression, multiple regression, and models for analyzing categorical outcomes.
6. PSYC 200 Statistics	Measures of central tendency, variability, association, and the application of probability concepts in determining the significance of research findings.
7. SOCY 162 Methods in Quant Sociology	Introduction to methods in quantitative sociological research. Topics include: data description; graphical approaches; elementary probability theory; bivariate and multivariate linear regression; regression diagnostics. Students use Stata for hands-on data analysis.

* As of academic year 2018-19, ECON 131 is no longer offered. It has been replaced with ECON 117: Introduction to Data Analysis and Econometrics.

The Yale College introductory courses listed in Table 2 cover the statistics and probability foundations. There is a limit to how much material can be covered in a single semester. These courses do not focus on research design strategies beyond basic experimental design or regression for observational data or on issues like publication bias and cognitive bias in incorporating information and in decision-making. Based on the course syllabi, these introductory courses tend to cover similar statistical concepts and techniques and differ primarily based on disciplinary focus (e.g., health, economics, politics).

There are currently no structures in place to ensure that there is coordination among instructors or identification of gaps in level or material covered. There is no social science divisional curriculum process or FAS-wide process that would promote Yale College courses that teach general principles using examples from across the varied social science disciplines or beyond these disciplines.

3. Course taking patterns suggest that many undergraduate students may be getting only minimal exposure to the key concepts and methods of empirical inquiry.

Developing the habits of mind that support vigorous and probing engagement with empirical claims requires multiple exposures to the analytical tools and extensive practice in their application. Although it is hard to know with certainty since some techniques and ideas will be part of courses that are focused on subject area information, after reviewing course patterns for “first courses” in statistics and research design (which are prerequisites for more advanced courses) it appears that 28% of Yale College undergraduates take no courses that are centrally focused on data analysis techniques or statistical methods and only 24% take more than one such course.[13] Most social science undergraduate students take the minimum number of statistics courses required for their major (typically 1 or 0). See Appendix Figures 1, 2, and 3.

The most distinctive analytical challenge in social science research is isolating causal effects from non-experimental data. The basic courses in statistics and probability in Yale College do appear to cover most or all of the basic statistics concepts. However, they do not provide comprehensive coverage of the challenges of research design for measuring causal effects in social science settings. Further, these courses do not focus on the practical issues that arise in interpreting empirical literatures, such as the problems with publication bias. These courses do not focus on the characteristic errors in how people incorporate new information into prior beliefs or the use of empirical evidence in decision-making. It is unclear how many students are getting exposure to the range of descriptive statistical analysis that is now possible due to developments such as the advances in manipulation of administrative data sets and the increasing use of digital media trails to document and analyze human behavior.

It is possible that examining enrollment in introductory courses misses important channels for building data analysis skills and if we had a fuller picture this would alter our assessment. However, there are other indications that undergraduate students are not attaining the level of proficiency that would be ideal. Faculty voiced concern, based on their experience and interactions with undergraduate students, about whether students are internalizing fundamental concepts of statistics and research design. In a graduation survey, Yale College Class of 2018 students in the humanities and social sciences reported less development of quantitative skills than what was reported by students at peer institutions. See Appendix Figure 4. The most direct method for determining what students know would be to conduct a detailed survey of Yale College seniors. However, Yale College does not measure what students have learned since enrollment or what the students have mastered after two years at Yale or when they graduate, and there are no tools provided to students to encourage self-assessment of their knowledge level.

4. The Yale College quantitative reasoning requirement does not appear to be leading most students to do substantial data-intensive coursework.

The Yale College quantitative reasoning (QR) requirement requires that students take two courses with a quantitative reasoning designation. The Committee looked at students in the Yale College Class of 2018 who took no more than two QR courses during their time at Yale, i.e. met the minimum QR requirement. This population comprised 22% of the 1,298 student sample. 43% of these students took no stats courses to fulfill their QR requirement. Based on these students' course-taking trends, the QR requirement is most commonly fulfilled by Econ 115: Introductory Microeconomics and is often fulfilled by non-statistical math courses (e.g., Math 190: Fractal Geometry, Math 101: Geometry of Nature, Math 107: Math in the Real World, Math 112 and Math 115: Single Variable Calculus), as can be seen in Appendix Figure 5. Two thirds of total QR course enrollments among these students are in non-statistical courses. These courses do not provide significant coverage of social science research design, data analysis, or use of data in forming judgments and making decisions. Further, taking a single course

[13] The Committee analyzed course taking patterns for students in Yale College Class of 2018. There may be some statistics and research design courses missed by this analysis.

covering the basics of data analysis and statistic to meet a requirement is unlikely to leave a lasting impression on the typical student.

There is a distinction between data-intensive, empirical social science and general quantitative skills. Yale College science students are repeatedly exposed to quantitative material, such as math, mathematic models, and experimental data. However, social science data analysis poses a distinctive set of challenges. One key feature of social science problems is that in many cases researchers must use non-experimental data to determine cause and effect relationships. Another challenge is that human beings adapt and reactions often depend on history and context. Relationships found in a particular study may be highly context dependent. Social scientists are often faced with challenges such as issues with measurement, discrepancies between reported behavior and actual behavior, and problems of endogenous choice. The analytical techniques and the empirical intuitions that are needed to account for these problems are not identical to the experimental design and measurement skills developed in physical and biological sciences. It is common for extremely intelligent and highly educated people in positions of authority who have not developed the habits of mind associated with social science data analysis to mistake correlations between human behavior (or experiences) and "outcomes" for causal relationships.

5. There are excellent learning opportunities in data-intensive research outside the classroom.

There are research assistant opportunities for undergraduate students to apply the concepts and techniques they learn in the classroom to data-intensive research projects. The FAS economics department provides extensive opportunities for undergraduate research. The Tobin Undergraduate Research Assistantship program provides undergraduate students with exposure to conducting research in economics by working with a professor for ~10 hours per week for one to two semesters. In the 2018-19 academic year, 70 students participated as Tobin Research Assistants, working with 34 different professors. Some project topics were Rural-Urban Wage Gaps, Voting Rights of Native Americans, and the Impact of LGBTQ Discrimination on Health Disparities. Another program is the Herb Scarf Summer Research program, through which students get directly involved in the ongoing research of professors by working with them in New Haven for ~160 hours over the summer. For Summer 2019, 27 students were selected to work with 13 different professors. Projects included Spatial Policies and Economic Growth; Human Capital, Migration, and the Returns to Schooling; and more.

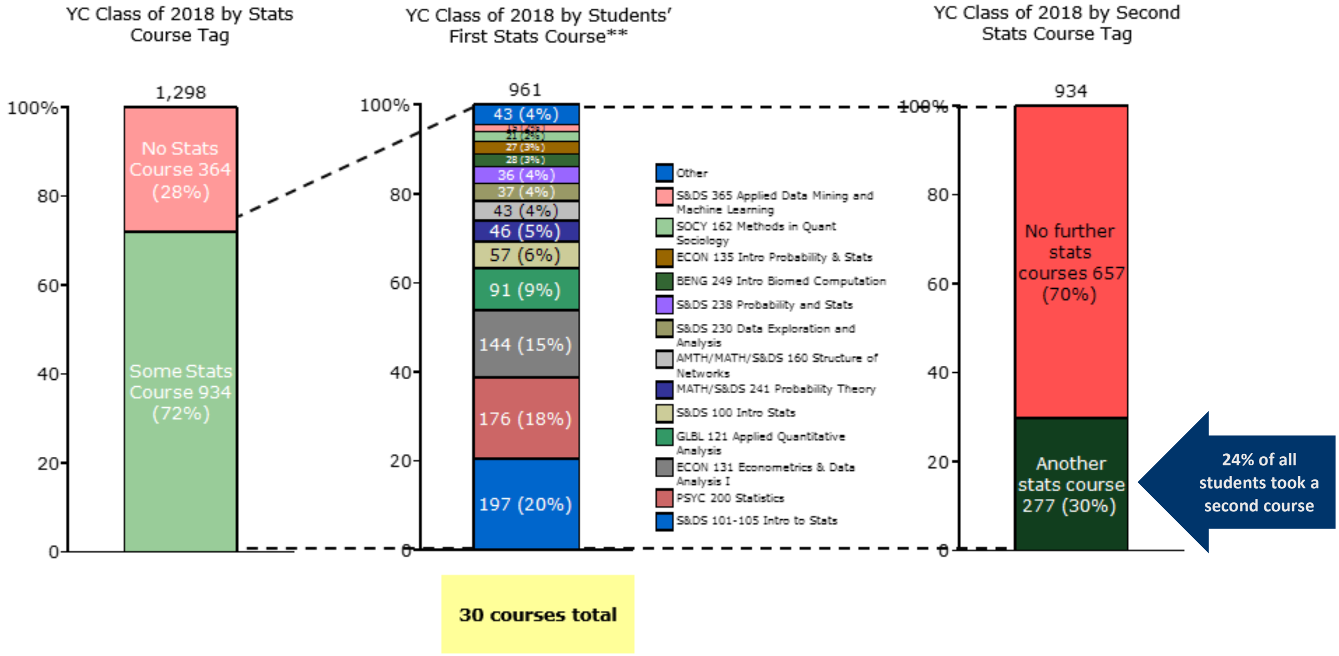
Yale Institution for Social and Policy Studies (ISPS) offers two fellowship opportunities for undergraduates. The Dahl Scholars program provides students with opportunities to engage in policy-oriented academic research with leading scholars through a year-long research and mentorship fellowship. Dahl Scholars work approximately 8-10 hours a week as research assistants, attend and prepare for 4-5 peer review meetings, and complete independent public policy research projects by the end of their program. Dahl Scholars typically plan on attending graduate school or joining organizations conducting rigorous policy research (e.g., think tanks, NGOs). ISPS also sponsors the Director's Fellows program, which provides undergraduates with sophisticated policy training and work experiences to bridge the gap between theory and practice in U.S. domestic policymaking. Over the course of a year, Director's Fellows attend biweekly policy seminars with leading researchers, government officials, and policy experts; participate in a policy internship; and complete a series of op-eds and policy briefs.

Fellows graduate prepared to contribute to high level domestic policy discussions through continued academic research and writing or working directly in policy.

There are also many research assistant opportunities in the psychology department, specifically in neuro, cognitive, developmental, social, and clinical psychology. Research assistants in psychology gain exposure to working in lab settings, conducting experiments, and collecting and analyzing quantitative and qualitative data. Some examples of projects include identifying the role of the gut microbiome in complex behaviors relevant to compulsive features of addiction, OCD, or binge eating; understanding how men and women transition into their parenting role and how this may be affected by psychopathology; and uncovering how children and adults think and reason about social groups and intergroup experiences. Advanced statistical methods include machine learning or structural equation modeling.

Figures

Figure 1: Yale College Class of 2018 Stats* Course Taking Trends



Note: *Stats courses include: all Yale College courses in STAT/S&DS department plus BENG 249 Intro Biomedical Computing, ECON 131, ECON 132, ECON 135, ECON 136, ECON 420 Applied Microeconometrics, ECON 481 Empirical Microeconomics, GLBL 121 Applied Quant Analysis, SOCY 162 Methods in Quantitative Sociology, PSYC 200 Statistics

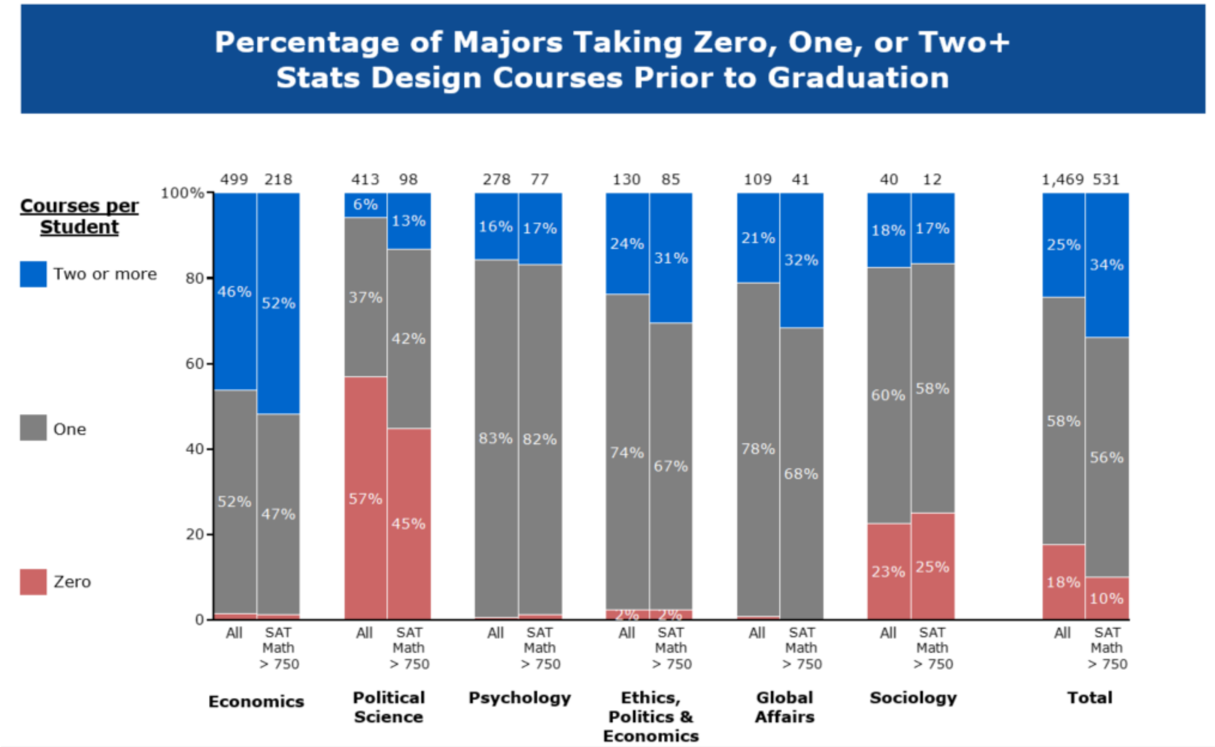
**If a student took multiple stats courses in his first semester taking any stats course, all of those courses would get counted in this graph

Figure 2: Statistics Course Requirements by Major for Yale College Class of 2018 and Prior

Major	Number of Stats Courses Required
Economics	1
Ethics, Politics, and Economics	1
Global Affairs	1
Political Science	0
Psychology	1
Sociology	0*

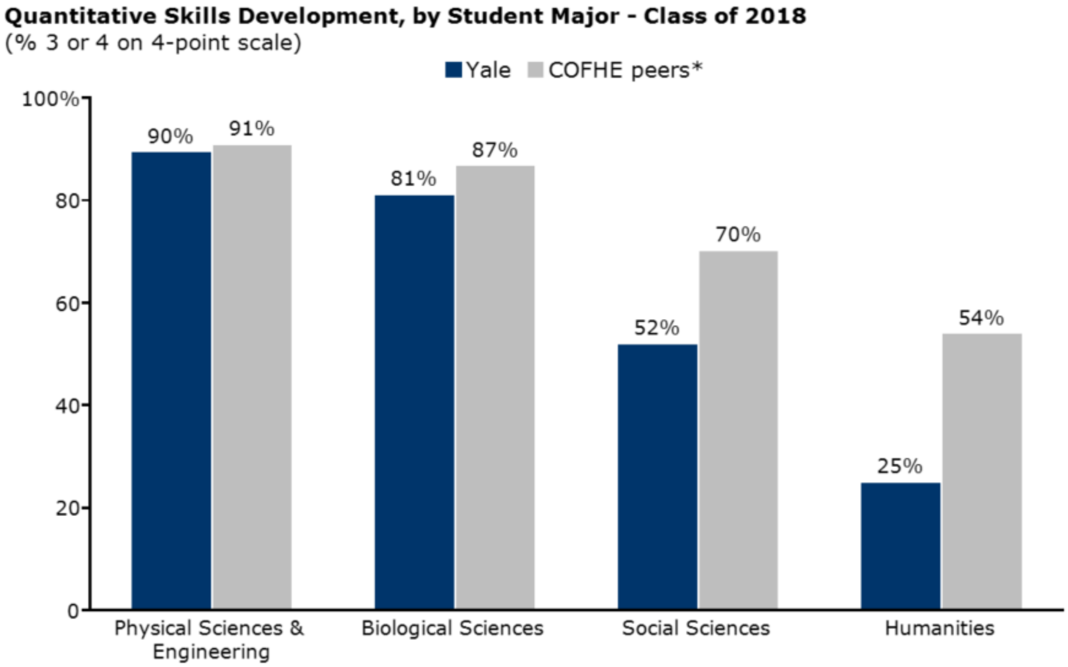
Note: * Not required for standard major; for concentrations in markets and society or health and society one stats course is required
Source: Yale College Programs of Study

Figure 3: Yale College Classes of 2014-17 Stats Course Taking by Major



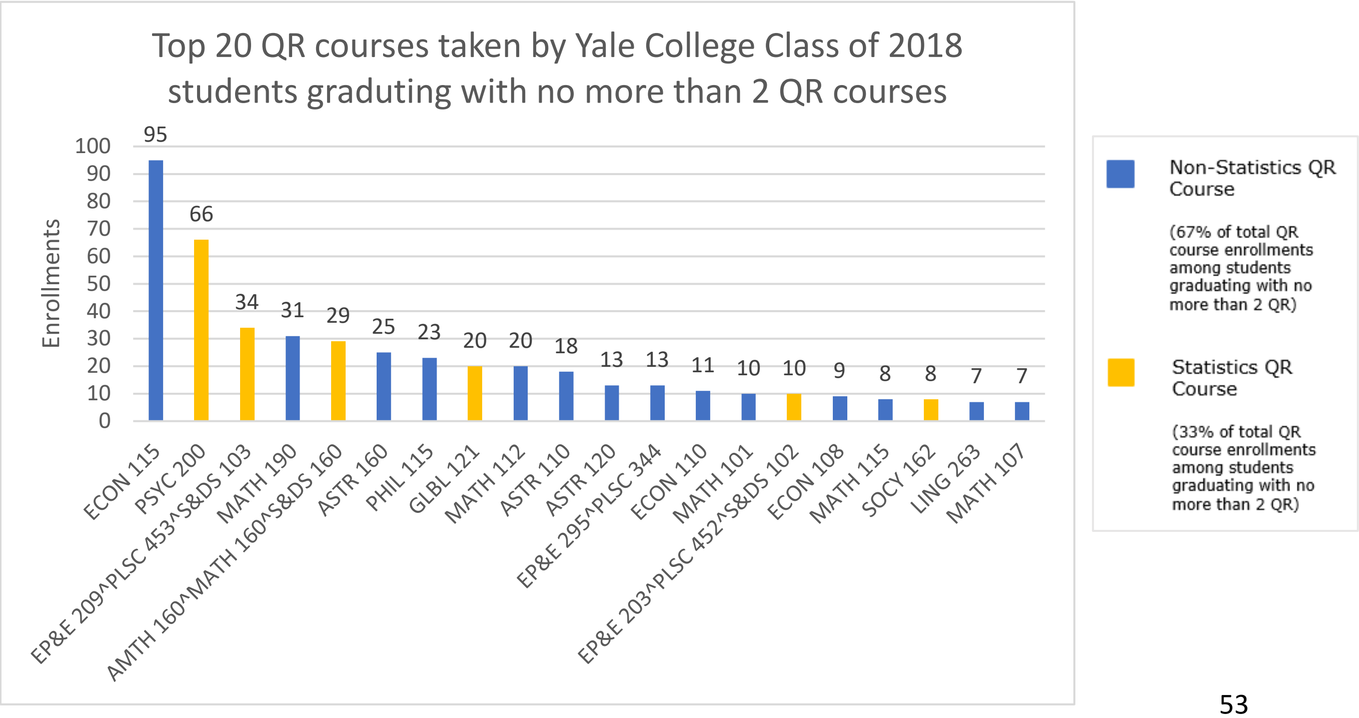
Note: Includes Yale College graduates in Classes of 2014-17, single majors only (excludes ~15% with two majors)
Source: Office of Institutional Research

Figure 4: Yale College Class of 2018 Reporting Quantitative Skills Development by Major Division



Note: * COFHE is an organization of top private US institutions (n=39); survey question asked, "To what extent has your experience at Yale contributed to your knowledge, skills, and personal development in using quantitative tools (e.g., statistics, graphs)?"; 1-4 response scale, with 1 as "very little or none", 2 as "some", 3 as "quite a bit", and 4 as "very much"
Source: Office of Institutional Research

Figure 5: Fulfilling the Yale College Quantitative Reasoning Requirement



Raj Chetty's "Using Big Data to Solve Economic and Social Problems" at Harvard and Carl Bergstrom's and Jevin West's "Calling Bullshit: Data Reasoning in a Digital World" at University of Washington are great examples of "signature" courses that teach important material in an accessible way and, when a sufficient number of students share the course experience, can also help to shape campus culture.

Raj Chetty's course was the third-largest course this spring (2019) at Harvard with over 400 students. The course requires no previous coursework in economics and teaches students how to understand and develop policy proposals for key social and economic problems of our time (e.g., equality of opportunity, education, health care, climate change, and crime) using "big data". The course covers statistical methods and data analysis techniques, including regression analysis, causal inference, quasi-experimental methods, and machine learning. The course also invites leading practitioners to discuss how they use big data in real-world applications.

Carl Bergstrom's and Jevin West's course is aimed at teaching students how to understand and evaluate research and recognize and call out misleading and fake empirical claims. The course covers concepts such as causality and correlation, Bayes' Rule and conditional probabilities, data visualization, machine learning, and more. Students see these concepts applied in real world case studies, such as food stamp fraud, traffic improvements, and the gender gap in 100-meter dash times. The course syllabus includes engaging and interdisciplinary readings, for example "The Will Rogers Phenomenon – Stage Migration and New Diagnostic Techniques as a Source of Misleading Statistics for Survival in Cancer," "The Parable of the Google Flu: Traps in Big Data Analysis," and "Rumor Cascades" in social media.

Another interesting example of an interdisciplinary and accessible "signature" course is the "Sense & Sensibility & Science" course co-taught at Berkeley by Nobel Laureate Saul Perlmutter (Physics), John Campbell (Philosophy), and Robert MacCoun (Public Policy/Law). This course successfully incorporates real-world examples and hands-on small group exercises into each lecture to make the material more engaging. Although this course focuses more on science literacy, the fundamental concepts and skills the course teaches (e.g., uncertainty, causal reasoning, sanity checks, biases) by bringing together natural science, social science, and humanities is something that could be tailored for social science literacy as well.

Stanford's CS 106 "Programming Methodology" is the largest introductory programming course and one of the largest courses at the university. The course teaches Java and emphasizes modern software engineering principles (e.g., object-oriented design, decomposition, encapsulation, abstraction, and testing). The course is "explicitly designed to appeal to humanists and social scientists as well as hard-core techies [and]…most Programming Methodology graduates end up majoring outside of the School of Engineering."[14] Stanford identified computer science as a mission-critical area and conducted a nationwide search for the right instructor to teach this introductory course in a way that was intellectual, not just mechanical. CS 106 is an example of how a well-taught, accessible, and engaging introductory course can spark community, conversation, and culture around a discipline across the entire university.

[14] https://see.stanford.edu/Course/CS106A

2. YData course

Yale’s YData course was modelled after UC Berkeley’s popular Data 8 course, titled “Foundations of Data Science,” which was introduced in Fall 2015. The course is designed for students who have not previously taken any statistics or computer science courses and covers core concepts of inference and computing, while providing students with opportunities to work with real data in weekly labs. The course also offers small-group tutoring sessions for students who may need additional support. Since 2015, UC Berkeley has offered 25 connector courses to supplement the core Data 8 course. Topics for connector courses have included: immigration, crime and punishment, ethics, ecology and the environment, history, literature, child development, economic development, and more. Data 8 has been incredibly popular at UC Berkeley. The University aims to offer it every semester to meet student demand. In Spring 2018, the University had to limit the course to ~1,000 students but hopes to offer more seats in future semesters.

3. Pathway to advanced achievement in quantitative social science

Although we envision a program more computational and empirical, it is useful to consider the Northwestern program called Mathematical Methods in the Social Sciences (MMSS). Each year approximately 30 first-year and 30 sophomores are admitted to this 2-year MMSS program. The program consists of 12 quarters of course work (approximately 8 semester courses) in mathematics, statistics, economic theory, and mathematical models of psychology, political science, and economics, plus a senior project. The program, started in 1978, appears both intellectually rigorous and quite popular, and it boasts an impressive record of post-graduate placement.

4. Pre-doctoral programs

Harvard’s Graduate School of Arts and Sciences Research Scholar Initiative in Economics (RSI) is a pre-doctoral program that provides mentored research and training for individuals interested in pursuing a PhD in economics or related fields. The program strongly encourages applications from underrepresented minorities and aims to enhance the competitiveness of individuals’ applications to top PhD programs. RSI admits three to four Economics scholars each year, and each scholar spends one to two years in the program, working as a research assistant for an assigned faculty member and taking graduate-level courses. Scholars have access to Harvard University resources (e.g., libraries, tutoring, professional development seminars) and receive a monthly stipend, health insurance, GRE preparation, a one-time relocation fee, and tuition for one to two classes per semester. Economics scholars also have the opportunity to join weekly workshops organized by Harvard’s Department of Economics and participate in group social activities. RSI alumni have gone on to PhD programs at top universities including Harvard, University of Chicago, NYU, UC Berkeley, and University of Pennsylvania. RSI in Economics is funded by the Alfred P. Sloan Foundation.

Columbia has a “Bridge to the PhD Program in STEM,” which is designed to increase the participation of students from underrepresented groups in PhD programs in STEM disciplines, including economics. Individuals accepted to the program are hired as full-time research assistants for up to two years and conduct research under the mentorship of faculty, post-doctoral researchers, and graduate students.

The Economics Bridge participants receive an annual salary of $50,123, an annual stipend of $2,000 for professional and educational expenses, GRE test preparation, and University benefits (e.g., health, retirement, etc.). Participants also enroll in one to two courses per semester related to their future field of study, attend monthly one-on-one check-in meetings with the Program's Director to evaluate their progress, and attend program-sponsored professional development workshops. The Bridge Program strongly encourages applicants from historically underrepresented groups.

Stanford has two pre-doctoral programs, one affiliated with its Graduate School of Business (GSB) and another affiliated with Stanford's Institute for Economic Policy Research (SIEPR). Both are geared towards individuals who want to gain training and research experience before applying to top PhD programs in business, economics, public policy, or related fields. The Stanford GSB Research Fellows Program is a two-year fellowship, and fellows have the opportunity to work closely with faculty in their field of interest on empirical research papers; take doctoral-level courses in business, economics, statistics, math or related fields; and regularly attend field seminars. Fellows receive exposure to the various fields of study at Stanford GSB, including Accounting, Economics, Finance, Marketing, Operations, Information & Technology, Organizational Behavior, and Political Economics. The program emphasizes diversity and strongly encourages applications from women and underrepresented groups. The SIEPR Predoctoral Research Fellowship is a full-time one to two-year program. There are twenty slots at any given time (i.e., hire ~10-12 each year) and almost all are funded by foundations or other sources. The program receives ~650 applicants for the 10-12 slots each year. Incoming fellows have a strong quantitative background, an interest in learning cutting-edge research methods, and experience programming in Stata, R, or other statistical packages. Fellows spend a significant portion of their time working on two empirical research projects with faculty members assigned based on project preferences and can take graduate-level courses at Stanford, up to one course per quarter. Fellows receive an annual stipend of $49,600, health insurance, and tuition for courses. They also receive mentorship from the faculty members they have been paired with.

5. Quantitative Reasoning (QR) requirement

Starting in Fall 2019, Harvard implemented a quantitative reasoning with data (QRD) requirement, replacing its past empirical and mathematical reasoning requirement. The purpose of the new QRD requirement is to ensure undergraduates "reach a level of quantitative facility involving mathematical, statistical, and computational methods that will enable them to think critically about data as it is employed in fields of inquiry across the FAS." QRD courses are "purpose-built, thematic courses deliberately outside professors' disciplinary teaching, with a recognizable, distinctive pedagogy and explicit connection to the wider world, with assignments that are outward-facing and perhaps even directly engaged with real-world challenges."[15]

Stanford has a comprehensive skills requirement for its undergraduate students. Of particular interest is Stanford's Applied Quantitative Reasoning (AQR) requirement. AQR courses teach inferential and inductive reasoning and provide opportunities for students to actively apply these methods of reasoning through direct manipulation of data, models, software, or other quantitative tools. Examples of

[15] https://www.harvardmagazine.com/2019/04/harvard-course-requirements-quantitative-reasoning

courses that fulfill the AQR requirement include: “Bio 141: Biostatistics”, “CS 102: Big Data: Tools and Techniques, Discoveries and Pitfalls”, “CSRE 184E: Race, Gender, and Literary Digital Humanities.” The AQR requirement is an applied complement to the Formal Reasoning (FR) requirement which focuses on logical and deductive reasoning and can be fulfilled with more theoretical courses in mathematics and computer science.

Another notable example is University of Chicago, which requires its students to fulfill a general education requirement by taking fifteen courses spread over seven areas of study, one of which is social science. The University offers five, three-course social science sequences in the general education program for students to choose. Two of these five sequences, “Social Science Inquiry” and “Mind,” have significant data-intensive components. In “Social Science Inquiry,” students are introduced to social science research tools and learn how to collect data, conduct experiments, and make causal inferences from statistics. Students also gain hands-on experience working with large data sets and conducting their own substantial research projects. This three-course sequence emphasizes students developing “a critical perspective on many perennial social questions and, ultimately, acquiring ‘quantitative literacy’, essential skills in an increasingly data-driven world.” In “Mind,” students develop essential habits of mind such as critically evaluating the legitimacy of social scientific questions, empirical evidence, and data used to test hypotheses and to support causal claims. [16]

4. Relation to the University Science Strategy Priorities

Peer Examples for Computation and Society

1. Stanford Institute for Human-Centered Artificial Intelligence (HAI)[17]

Stanford HAI’s mission is to “advance AI research, education, policy, and practice to improve the human condition…and to become an interdisciplinary, global hub for AI thinkers, learners, researchers, developers, builders and users from academia, government and industry, as well as leaders and policymakers who want to understand and leverage AI’s impact and potential.” HAI emerged from Stanford’s long-range planning process, which began in 2017 and solicited ideas to promote creativity and innovation and accelerate solutions to improve society. The Institute will be housed in a new 200,000 square foot building, along with the new Data Science Institute, at the heart of campus. Stanford is reportedly looking to raise more than $1 billion for HAI, the same as the funding target MIT has set for its own interdisciplinary AI institute, the Schwarzman College of Computing. Money raised will go to research grants, academic gatherings, buying data processing power, and attracting back some talent that has left academia for industry jobs in recent years

HAI has three research areas of focus: (1) human impact, (2) augment human capabilities, and (3) intelligence. The Institute awards 25 seed grants of up to $75,000 each year to spur novel research and in its first two grant cycles, it has already committed funding to fifty research projects. The Institute also offers human-centered AI courses on campus and online (e.g., “AI in Real Life Seminar Series”, “The

[16] U Chicago College Catalog, http://collegecatalog.uchicago.edu/archives/2014-2015/thecollege/socialsciences/
[17] https://hai.stanford.edu/

Politics of Algorithms", "Regulating AI") and hosts events such as monthly community building receptions (e.g., AI & Education, AI & Civic Architecture), seminars (e.g., AI & Accessibility: Ethical Considerations), an annual AI and Human Rights symposium, and an annual conference on AI Ethics, Policy, and Governance.

HAI plans to hire at least twenty new faculty, including ten junior fellows, from across fields. Current affiliated faculty and staff include:

- Co-directors John Etchemendy, Professor of Philosophy, and Fei-Fei Li, Professor of Computer Science (respectively, the former provost and former director of Stanford AI Lab)
- 6 faculty associate directors
- 12 faculty member design team
- 10 staff members – 1 HAI deputy director, 1 director of research, 1 director of administration, 3 administrative associates, 1 research and financial analyst, 1 events planner 1 HR and Operations administrator, 1 faculty affairs and fellowship coordinator
- 23-member advisory council from government, industry, and academia (e.g., former Google CEO Eric Schmidt, LinkedIn co-founder Reid Hoffman, director of Microsoft Research Labs Eric Horvitz)
- 19 distinguished fellows from industry and academia
- ~150 participating faculty from all seven schools at the university

2. MIT Schwarzman College of Computing[18]

The mission of MIT's Schwarzman College of Computing is "to accelerate pioneering research and innovation in computing, build a profound awareness of the ethical implications and societal impact, and above all, educate leaders for the algorithmic future." The creation of the College was motivated by increasing student interest in computer science, the burgeoning opportunities for research to benefit from advanced computational knowledge and capabilities, and the increasing role of computing and AI in every facet of our lives, including education, the environment, ethics, finance, health, policy, security, transportation, and the global economy. The College will be housed in a new building centrally located on campus, scheduled to be completed in 2022. MIT has made a $1 billion commitment "address the global opportunities and challenges presented by the prevalence of computing and the rise of artificial intelligence," and has received a $350 million foundational gift from Stephen Schwarzman, and as of October 18, 2018, has raised an additional $300 million in support.

The College aims to respond to increased student demand for computing curricula by expanding course offerings and programs to educate students in every discipline to be "bilingual" and empower students to become "creative computational thinkers and doers with the cultural, ethical, and historical consciousness to use technology for the common good." The College will also implement new programs exploring the intersection of ethics/societal impact and computing – undergraduate research opportunities, graduate fellowships in ethics and AI, support for interdisciplinary faculty collaboration, and programs to attract distinguished individuals from other universities, government, industry, and journalism.

[18] https://computing.mit.edu/

MIT plans to create 50 new faculty positions, with 25 to be appointed to advance computing in the new College, and 25 to be appointed jointly in the College and departments across MIT. The addition of faculty will also naturally lead to growth of graduate students and post docs. Current affiliated faculty and staff include:

- Inaugural dean, Dan Huttenlocher
- Provost's task force composed of MIT faculty, students, and staff to focus on (1) organizational structure, (2) faculty appointments, (3) academic degrees, (4) social implications and responsibilities of computing, (5) computing infrastructure; task force submitted preliminary reports on June 5th which will go through a community comment period
- The department of Electrical Engineering and Computer Science (EECS), the Computer Science Artificial Intelligence Lab (CSAIL), the Institute for Data, Systems, and Society (IDSS), the MIT Quest for Intelligence, the MIT-IBM Watson AI Lab, and the Center for Computational Engineering will all become part of the MIT Schwarzman College of Computing

3. UC Berkeley Inclusive Intelligence Initiative[19]

In its Spring 2018 Strategic Planning report, UC Berkeley proposed an Inclusive Intelligence Initiative to develop "a new approach to artificial intelligence and data science that is: 1) inclusive of individuals from all backgrounds to benefit the greater good; 2) inclusive of a broad community of scholars from engineering, business, the arts, humanities, psychology, neuroscience, sociology, politics, philosophy, history, and other disciplines; 3) inclusive of developing a broad array of related approaches and technologies such as data science, artificial intelligence, robotics, sensing, machine learning, etc.; and 4) inclusive of both human and artificial intelligence and the way they interact, complement, and enhance each other." The Initiative aims to "promote both continued technological innovation and a broad investigation of the societal and ethical implications of artificial intelligence, robotics, and data sciences." As the Initiative is currently in the planning phase, ideas for implementation include:

- "Call out interdisciplinary intersections in which faculty across campus are prepared to launch collaborative research projects, so they can make concrete headway on problems such as algorithmic fairness and interpretability, new processes for political governance, the future of work, and data-intensive solutions to societal problems"
- "Invest in foundational areas across the disciplines where the human futures of technology will be shaped, from philosophy, history, social theory, and the arts to a new human-centered engineering discipline for data- and learning-focused fields"
- "Create campus-crossing educational programs for undergraduates and graduates to gain grounding across the human and societal challenges of technology, and for mid-career professionals to reflect on their lived experience and return with new ideas"
- "Draw together the campus's nascent student programs in data, computing, ML, and AI for social impact and underwrite their growth with high-level campus support"
- "Create the world's leading public forum on ethical paths into a technological future, making the campus an international magnet for faculty and students and a showpiece for the state"

[19] https://strategicplan.berkeley.edu/wp-content/uploads/2018/12/Appendix_A1-Inclusive-Intelligence.pdf

4. NYU – AI Now[20],

The AI Now Institute at NYU was founded in 2017 by Kate Crawford and Meredith Whittaker and is an interdisciplinary research center dedicated to understanding the social implications of artificial intelligence. Its four main research areas include: (1) rights & liberties, (2) labor & automation, bias & inclusion, (4) safety & critical infrastructure. The Institute organizes an annual symposium and regular workshops (e.g., “Machine Learning, Inequality and Bias Roundtable”, “Convening on Immigration, Data, and Automation in the Trump Era”). AI Now receives funding and support from Luminate, MacArthur Foundation, Microsoft Research, Google, Ford Foundation, The Ethics & Governance of AI Initiative, Deep Mind Ethics & Society.

The Institute is in the process of searching for an executive director. Current affiliated faculty and staff include: 2 co-directors; director of strategy & operations; director of policy research; 2 program associates; 2 technology fellows; 5 postdoc researchers; 5 affiliates (visiting professor, research fellow, artist fellow, writer in residence); 9 NYU area leads; 10-member advisory board from industry, government, and academia.

[20] https://ainowinstitute.org/