Harvard Data Science Review • Issue 3.2, Spring 2021

Data Science and Computing at UC Berkeley

Jennifer Chayes¹

¹Division of Computing, Data Science, and Society, University of California Berkeley, Berkeley, California, United States of America

Published on: Apr 30, 2021

DOI: <https://doi.org/10.1162/99608f92.12c8533a>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

ABSTRACT

Data science and computing are becoming central to education, scholarship, and societal impact; indeed, they are being woven into the fabric of our society. I discuss how UC Berkeley is organizing its educational, research, and institutional structures to respond to the opportunities and challenges this creates.

Keywords: data science, computing, statistics, information, multi-disciplinarity

1. Introduction

Data science and computing are becoming lenses through which we view and interpret the world, and a framework through which we interact with and alter the world. They are central to data-driven advances in many fields, from STEM to the social sciences to the professional disciplines. At the same time, computing has been leading the development of the social and economic platforms that are rapidly becoming the fabric of our society, for better and for worse, both locally and globally.

Against this backdrop, universities are grappling with how to organize data science and computing efforts across their campuses. There are educational mandates: Driven by tremendous demand from students and employers, there is an almost insatiable demand for data science and computing degrees. There are scholarly mandates: Advances in data science and computing are driving new approaches to other fields, as other fields are challenging the foundations of data science and computing. There must be deep and reciprocal relationships between data science and computing, and essentially all the other disciplines on campus. There are societal mandates: As algorithmically-driven data platforms become the foundation for much of our economic, social, and political interaction, as well as our educational, public health, social welfare, and criminal justice systems, scholars in data science and computing must be in deep conversation with the rest of the university and with society more broadly. We must approach computing and data science with human perspective and purpose. It is time to unify computing and data science with its social mission and responsibility.

This article will focus on the way one institution—UC Berkeley—is responding to these demands. This article is very much from a UC Berkeley perspective, and yet it highlights many of the opportunities and challenges faced by other institutions. About a year ago, I joined UC Berkeley as the inaugural Associate Provost of the Division of Computing, Data Science, and Society (CDSS), and Dean of the School of Information. CDSS is in the process (a long process at the University of California) of becoming a new College, on par with the other Colleges at UC Berkeley—Letters and Sciences, Engineering, and several others. It contains pieces from both Letters and Sciences and Engineering, as well as a professional school.

CDSS contains the Department of Statistics, the Department of Electrical Engineering and Computer Sciences (EECS) (shared equally with the College of Engineering), and the School of Information. CDSS also contains

other entities. The Center for Computational Biology brings together about 50 faculty from across campus (Statistics, EECS, Bioengineering, Biology, the School of Public Health, and more)—most of the faculty 0% time, but four faculty 50%—and it offers a Ph.D. in Computational Biology. The Berkeley Institute for Data Science (BIDS) was one of the three original Moore-Sloan-funded Data Sciences Institutes, along with those at NYU and the University of Washington. The Data Science Education Program offers the Data Science major, a unique joint major of Statistics and EECS, with much of the curriculum developed by faculty in other units, from Human Context and Ethics to a host of domain disciplines. It is the fastest-growing major in the history of UC Berkeley. This is only our third graduating class, and yet 700 UC Berkeley seniors will graduate in the Data Science major in 2021.

In this piece, I talk both about the potential for data science and computing and about how we build educational, research, and institutional structures to support our aspirations. This is not meant to be exhaustive or definitive; it is a snapshot in time of my view of the potential of data science and computing, rooted in my decades of experience building multidisciplinary research groups and labs, deeply informed by my recent experiences here at UC Berkeley, and inspired by the hopes of what we and other institutions can build.

2. What is Data Science?

You will probably get as many definitions of ‘data science’ as you have data scientists. There is already a compelling body of literature on this question, some of the most thoughtful in the pages of the *Harvard Data Science Review* (HDSR). I highly recommend the collection of articles from HDSR on [“Perspectives on Data and Data Science.”](#) Jeannette Wing’s (2020) already classic [“Ten Research Areas in Data Science”](#) begins by framing the yet-unanswered question of whether data science is a discipline, and then provides a scope of the ‘field’ in terms of ten major areas of study. In the same issue, Xuming He and Xihong Lin’s (2020) piece, [“Challenges and Opportunities in Statistics and Data Science: Ten Research Areas,”](#) starts from more of a statistics (and biostatistics) framing, and follows the evolution of data science, posing its own ten challenges, several of which overlap with Wing’s areas of study. Both pieces elicited many thought-provoking pieces by discussants; see the perspective page cited above.

I start from the position that data science is the bringing together of the core disciplines of computing, statistics, and information with a broad spectrum of other disciplines. On one level, data science is a collaboration between the core and other fields to ask the right questions, collect and analyze data to address those questions, interpret it, draw inferences, and propose theories and interventions—often with the goal of scientific or societal impact. On another level, it is the setting of frameworks, questions, and aspirations from a broad array of other disciplines. These challenge the core disciplines and support a much more powerful intellectual agenda.

How can data science do this? Because data science properly sits at the nexus of STEM, and what I have begun to call the human-centered disciplines (HCD), which includes social sciences, arts, and humanities, as well as

the professional disciplines of public health, policy, social welfare, education, law, and medicine. As such, it has potential to be a game-changer for academia and the world.

Multiculturalism is inherent to data science. We should not simply learn the languages of other disciplines, but also understand what questions those disciplines ask, and why they are important.

This multiculturalism begins with the constituent disciplines of data science: statistics, computing, and information. Indeed, many universities have struggled with how (or even whether) to bring these three disciplines together. In many cases, data science has grown out of statistics, with bridges to computer science forming later, but grounded in statistics. Achieving data science as a true partnership of statistics and computing, and uniting it deeply with the study of information, has not been done that often. One thriving example that comes to mind is Cornell University, where such a unit was formed about 15 years ago. We here at UC Berkeley are doing something similar at our base, but also with more explicit connections to other disciplines.

It is only when computing and statistics researchers have an appreciation of the very different cultures, questions, and approaches of other fields that deep multidisciplinary data science will occur. This is true even in interactions with STEM fields, where there is often an assumption of similar culture with statistics and computing. Computer scientists and statisticians typically have very different world views from those in the physical, biological, medical, and environmental sciences, and engineering; different questions are asked, and different impacts are sought. Computing and statistics in deep interaction with the biological, physical, medical, and environmental sciences, and engineering, is already enabling a revolution in those fields; it is just the beginning.

Data science informed by the arts and humanities is the framework we must strive to build, because it will be fairer and more beautiful, and will inspire us to do deeper and more important work. Data science ambitions become both more grounded, and more elevated in purpose, through deep interaction with the professional disciplines—public policy, health, education, social welfare, law, and medicine.

Computing and data science must incorporate a keen understanding of the ethical issues inherent in data sets and algorithms. We must formulate explicit methods to audit, as well as mitigate, the biases and limitations of our data sets and algorithms. We must also be adept at thinking through the human, political, and social justice implications of the data sets and algorithms. Machine learning informed by ethical, legal, historical, and moral scholarship, and grounded by interacting and iterating with people in the real world, can produce decisions that are less biased than human judgments; it can transform the way we approach societal decision making. Privacy considerations must be front and center in both the collection and the use of data sets. In certain domains, such as personalized medicine and criminal justice, it is essential that the predictions of data science be humanly interpretable. Again, there is much recent work on the Ethics of Data Science; many are collected in the HDSR collection on [“AI and Responsible Data Science.”](#) The authors of these works are motivated not only by

scientific concerns, but also by societal and even deeply personal ones; for the latter, I strongly recommend Michael Jordan's (2020) article, ["Artificial Intelligence – The Revolution Hasn't Happened Yet."](#)

Computing and data science provide a framework for approaching our most pressing societal problems: climate change, personalized medicine, health, the future of work, child welfare, elder care, social justice, and more. Even questions that at first seem rather technical, e.g., climate change, have human factors and contexts which must be addressed by proposed solutions. We need psychologists, historians, economists, political scientists, and others, along with climatologists, earth science researchers, environmental researchers, material scientists, machine learning researchers, statisticians, and engineers, to address a problem like climate change, which is as much an issue of individual and political will for social change as it is a scientific and technological issue. As algorithms become the medium for much of our interaction with our public systems, we must work with our colleagues in education, social welfare, public health, public policy, and criminal justice to make sure our constituents are well-served.

Especially at this critical time in our history, racial justice must be at the forefront of education, scholarship, and social impact. The existing data science and computing platforms, which have become the fabric of many of our social systems, can tend to exacerbate rather than mitigate bias. Often this is unintentional, as in the case of image processing software which does not work well on black faces, or software to determine whether to grant loans, which substitutes proxies for protected attributes like race. Increasingly, it is also intentional as those who want to exacerbate racial injustice usurp the power of social platforms. It is time to bring the disciplines of computing, statistics, and information together with the broader university community to build these systems in ways that will help to overcome systemic racism and discrimination rather than to increase it. It is our responsibility to bring all voices to the table at the earliest stages of data science and computing research to ensure that our future reflects our values.

3. Combining Computing and Data Science

Why combine the core of computing (which includes all of computer science, AI, and decision science) with data science? There are deep intellectual and societal reasons. The intellectual arc of computing has taken it far beyond its early framing as the mathematics, science, and engineering of computation, and has brought it much closer to STEM and to a host of humanist, social, and professional disciplines. These interactions challenge the core of computing, and these challenges are the impetus for breakthroughs in the core.

For a historical perspective of the intellectual arc of computing, I turn to a framing of that history by Turing Laureate Butler Lampson. Lampson was one of the central participants in the computing revolution that occurred at Xerox PARC in the 1970s and has continued to be a guiding light for the field. He says (Lampson, 2014) that there have been three ages of computing—the Age of Simulation, the Age of Communication, and the Age of Interaction with the Physical World. I often extend the third to be simply the Age of Interaction, both with the physical world and with public systems.

The Age of Simulation had its heyday in the 1950s-1970s, though there of course continues to be deep wonderful work advancing both the core of simulation, and its applications. It is interesting to note that computer science was typically a field in Letters and Sciences during the Age of Simulation; it was much closer to mathematics than to engineering. At the same time, the physical aspects of computing and electrical engineering were typically done in Colleges of Engineering.

The second age—the Age of Communication—where computing became the medium of our communication, had its roots (though no one realized it as such) in the Arpanet. Then in the 1970s, people like Bob Taylor, Butler Lampson, and others at Xerox PARC began to wire together the computers of the day—which everyone viewed as large calculators or simulators—to usher in the Communication Age. At most universities, computer science moved to Colleges of Engineering during this period.

Finally, sometime in the last decade, we have entered the Age of Interaction, with computing as the medium for our interactions both with the physical world (Internet of Things, e.g., sensors, self-driving cars, etc.) and with the public service world (public health, education, social work, law, government). What do I mean by interaction with the public service world? Increasingly, decisions about such things as what public schools you will attend, what social and public health services you will receive, whether you are considered a potential suspect in a crime, and, if arrested, how your bail will be set, are made algorithmically, using training data that is often skewed and algorithms which can intensify the bias in the data.

At the beginning of 2021, we can look back at the best and the worst of the Age of Communication. The white-collar economy in developed nations was able to carry on in the horrifying shadow of COVID-19, with computing as our means of communication. We bought our groceries and other goods via the Internet. Some of our children received parts of their education (albeit far from perfectly). We saw doctors, kept in touch with friends and family through social media, and even said goodbye to loved ones through computing-mediated means. We also saw the worst, much of it delivered through social media, from increased rates of teen suicide brought on by bullying, to destabilization of democracy due to misinformation. As I write this, the world is still reeling from the events of January 6, 2021, which have their roots in social media platforms. Computing as our means of communication had grown without the benefit of deep dialogue with the human-centered disciplines. We anticipated neither the unintentional effects of amplification of bias in standard machine learning algorithms, e.g., racial bias in facial recognition (Buolamwini and Gebru, 2018), nor the intentional subversion of the power of computing as a means of communication. These are both very difficult to fix ‘after the fact,’ though many of us are working to find normative, regulatory, and technical solutions.

Now we are on the threshold of the Age of Interaction. We cannot begin to imagine the promise of computing (with statistics and data science) as our means of interaction. At its best, it will free us to accomplish so much more, to provide care and enable experience at a new level, to distribute the gains of society more equitably, to tackle climate change, and to finally realize the promise of personalized medicine. I shudder to imagine its worst—both from unintentional amplification of bias and from intentional subversion of our powerful means

of interaction. We are already beginning to see the effects of unintentional amplification of bias, e.g., in our medical systems in the United States, where a black person with the same insurance as a white person often gets worse care due to bias amplification in algorithms (Obermeyer et al., 2019). What about subversion of the power of these systems? Physical systems, like self-driving cars or security systems—pieces of the massive Internet of Things (IoT)—could be subverted by those with nefarious goals. Subversion of the means of public systems are even more frightening since they are not as easily detectable: what if the same players who used social media to destabilize democracy on January 6, 2021, gain control of algorithms which increasingly run our public systems—education, social welfare, public health, criminal justice—to intentionally amplify racism in these systems? Today, we can spoof a self-driving car into not stopping at a stop sign by altering the stop sign in ways undetectable to humans. Could someone alter the data in a public system, again in ways undetectable to humans, and intentionally limit the opportunities for some group of people? It is time to pull statistics and computing—the core elements of data science—out from the exclusive realm of STEM and put them in their proper place at the nexus of STEM with the human-centered disciplines.

With a human-centered approach to data science and computing, we will be able to better frame the perspective and purpose, and to anticipate the problems, in this new medium. But we will also need tremendous innovation in the core of computing and statistics to move the field dramatically forward, to address the most difficult technical challenges and threats, and to realize the greatest opportunities.

4. Educational Opportunities and Responsibilities

At UC Berkeley, the creation of our new Division of Computing, Data Science, and Society was initially driven by students who wanted to be educated in the emerging field of data science. A few years ago, UC Berkeley developed an incredible curriculum for our Data Science Education Program; see the article in this issue by Ani Adhikari, John DeNero, and Michael Jordan. It was the product of thousands of hours of our deeply committed faculty and students—computer scientists, statisticians, information scientists, historians, ethicists, legal scholars, geographers, ecologists, artists, and many more. The major has four components: computing, statistics, human context and ethics, and a disciplinary focus. It is built on doing—not just studying—data science, with Jupyter notebooks integrated into the earliest stages of the program. The data science major is the fastest growing major in the history of Berkeley; only three years into the major, we will be graduating over 700 seniors this spring. Data science is also important for a general education. Over 6,000 UC Berkeley students take a data science course each year, including over 3,000 annually who take our rigorous introductory course, Data 8. Our data science education curriculum is now being used at numerous universities throughout the world. Indeed, hundreds of colleges and universities participate each year in our annual Data Science Education Workshop. Our curriculum is used widely by both public and private colleges and universities, and by community colleges.

Who are the UC Berkeley Data Science students? There are some who knew they wanted to do computer science or statistics already in high school. But the majority are those who came to UC Berkeley to do

something else, often in the arts, humanities, sciences, or the professional domains, but also in the STEM fields. They learn about data science through our intro course, where they realize that they have both an aptitude and an affinity for a data-driven approach to whatever their field of interest. They are disciplinarily and potentially demographically more diverse than the standard computing or statistics students. Many of them will take their data science education into fields well beyond science and technology, becoming the leaders in those fields. Whether they are passionate about climate change or biomedicine or social justice, they will find that a human-centered data-driven approach will help them to achieve their aspirations. We are proud to nurture this new generation of leaders. Data science also provides a path into technology for those who discovered this calling after high school; hence it is also a means for democratization of the tech industry. For students of all backgrounds and aspirations, data science can be their vehicle for social mobility and societal impact.

Just as UC Berkeley has led in creating a data science curriculum, the openness of our students and faculty to new disciplines that serve the needs of society is enabling further opportunities. Eric Schmidt has asserted that, for our government to function properly, the US needs 500,000 “public interest technologists”—people with training in computing and data science who enter public service at the Federal or State level. Our data science major provides a way to create generations of public interest technologists, and our Data Science Education Workshop provides a way to scale this to hundreds of other colleges and universities.

Data science literacy is becoming increasingly important as a component of general education (see the wonderful collection of HDSR pieces on [“Data Science Education”](#)), and in particular the article by Alan Garber (2019), [“Data Science Education: What the Educated Citizen Needs to Know.”](#) Society is constantly being challenged by intentionally or unintentionally misleading information and prediction. Our defense against this misinformation is a solid education enabling us to think critically and confidently about data. Education provides a framework for thinking through the social and political implications of data sets and algorithms, including an understanding of the incentives implicit in the data and algorithms, and how better to align those incentives with our values and goals. There needs to be a fabric of human context and ethics in data science education. General education on the role of data and algorithms in society is as important as specialized education in applications of data science.

The opportunities go well beyond undergraduate education. We have begun to talk with many other schools and colleges across the UC Berkeley campus about the potential to offer master’s degrees in ‘data science + X,’ including biology, social science, urban planning, business, law, public policy, social welfare, education, public health, journalism, and many others. Some of these will be in-person while others will be online. Finally, we intend to develop joint Ph.D. programs, e.g., we are currently developing a new Ph.D. curriculum jointly with UCSF in Computational Precision Health. Many new Ph.D. programs will also be started with other departments and schools at UC Berkeley. Scholars from other universities have asked me whether we plan to offer a Ph.D. in Data Science. As of this writing, we are not yet decided. Before we offer a Ph.D., we have to answer the question of whether data science is a discipline, (see, e.g., Wing, 2020), and if so, what defines it.

5. The Intellectual and Societal Mandate

Our aspirations for intellectual and societal impact include both the core of CDSS and beyond.

5.1. The Core

The core of CDSS comprises three academic departments and schools—Statistics, Electrical Engineering and Computer Sciences (EECS) (shared jointly with the College of Engineering), and the School of Information—as well as other research and educational entities like the Berkeley Institute for Data Science (BIDS), the National Science Foundation West Big Data Innovation Hub (with management shared by the University of Washington and UC San Diego), and the D-Lab, an organization on the Berkeley campus which helps to train social scientists and humanists in basic data science.

Statisticians have been working on applications in deep collaboration with disciplinarians in other fields for a century or more: they are the original data scientists, and they remain central today. They have worked for decades with the ‘big data’ of the day, but also with ‘smaller data’ from which they are nevertheless able to draw statistically impactful conclusions. Much of the relevant human data from public systems to biomedicine is not ‘big data’ in the current conventional sense, but rather high-dimensional sparse data, requiring fundamental advances in statistics and statistical machine learning.

Computing is another central component, complementary to statistics in many ways. Clearly, machine learning (ML) and artificial intelligence—including vision and image analysis, natural language processing, robotics—are essential to data science. But so too are computing systems, database systems, programming languages, algorithm design, theory, privacy, security, and more which underpin so much of data science.

Finally, the School of Information, which first emerged at Berkeley (as at other institutions) as the evolution of the Library School, is focused on the convergence of information, humans, and technology. This includes information of all types, both qualitative and quantitative, and how that information enables us to address issues from media manipulation to global poverty. Incentive considerations are fundamental to the research in the School of Information. The school in many senses incorporates both the core and connections beyond the core; it provides a model for how computing and data science might interact with the non-STEM disciplines.

In the context of the core, we have amazing research agendas and aspirations, of which I will mention just a few. In statistics, we are increasingly doing proper causal inference, enabling accurate prediction, and proposing scientifically grounded policy interventions. We can now do causal inference on questions for which it would until recently have been difficult, impractical, or immoral to do A/B tests (i.e., different interventions for different groups) by identifying random variation in the data in ways that are likely uncorrelated with what we are trying to measure.

In computing, we are embarking on what Butler Lampson calls the Age of Interaction with the physical world. How do we create systems which interact productively with us without knowing our preferences? Indeed, although we often now use fixed objective functions (e.g., reward functions in reinforcement learning), true preferences differ not only among different people, but within a single person in different contexts. How does the mere existence of the interaction change the context? What happens when we extend this to society—to networks of agents, each with their own context-dependent preferences and incentives? We must develop new frameworks in the core of computing to enable these interactions.

The School of Information has deep and compelling research programs across two areas which overlap deeply with computing, but do so with more economic, policy, legal, and humanist perspectives: digital security, safety and trust; and evolving human-data interfaces. Fairness, accountability, and security of AI is a central element of the School of Information, and indeed of all three of our core disciplines.

5.2. The Breadth and the Promise

Since arriving at UC Berkeley a year ago, I have had the thrill of speaking with faculty in our Divisions of Social Sciences, Biology, Mathematics and Physical Sciences, Arts and Humanities, Colleges of Engineering, Natural Resources, Chemistry, and Environmental Design, Schools of Business, Law, Social Welfare, Education, Public Health, Public Policy, Journalism, and Optometry, as well as the Medical School at UC San Francisco, and Lawrence Berkeley National Laboratory. In every case, we have found tremendous potential for joint initiatives and research and educational agendas with computing, statistics, and information. Let me discuss several nascent efforts to address some of the most urgent challenges facing society.

Biomedicine and Health: With UCSF, and the Berkeley Division of Biology, the Colleges of Chemistry and Engineering, the Schools of Public Health and Optometry, we are asking many questions, e.g.:

- How can we integrate heterogeneous data—from the genome, clinical tests, imaging, health care notes, demographic information—to provide truly personalized medicine?
- How do we decipher the roles of the immunome and the microbiome, two new sources of data that, together with the genome, will be to the next quarter century what the genome has been to the last? From an individual's immunome, can we read the cancers their body is fighting (e.g., ovarian or pancreatic cancer), and intervene before they are detectable by other means? Can we also read the autoimmune diseases individuals are incubating, before certain immune cells destroy the islets of Langerhans to produce type 1 diabetes, or destroy the myelin sheaths to produce multiple sclerosis? Can we anticipate and avoid the next potential pandemic?
- How do we use machine learning to accelerate drug design and to optimize the synthesis and manufacture of those drugs?
- How do we develop platforms for public health which take the best data and models we have, and provide a front end to enable interventions by public health workers who are likely not trained in data science?

- How do we get beyond the limitations of the health care data sets we have collected, which tend to be dominated by Caucasian men? How do we design health care algorithms which decrease bias in the data rather than exacerbating it?

More broadly, we are committed to creating computing and statistical platforms for integrating our broad knowledge and methodology in biomedicine and health, to allow us to make the best possible decisions on scales from the individual to societies. We are in the process of hiring joint faculty and designing a joint Ph.D. in Computational Precision Health with UCSF to address some of the clinical aspects of these questions.

Climate and Sustainability: We are working with faculty in the Berkeley Colleges of Natural Resources, Chemistry, Engineering, and Environmental Design, the Divisions of Math and Physical Sciences, and Social Sciences, the Schools of Business, Law, Public Policy, and Public Health, as well as Lawrence Berkeley National Laboratory. The scope of work in climate and sustainability at Berkeley is amazing. For us, the most pressing question is how we integrate this work to be more than the sum of its parts—a vehicle both for addressing climate change and for providing policy makers with sound methods to choose between different possible interventions.

We have data-driven technical solutions coming out of collaborations with the Colleges of Natural Resources, Engineering, and Chemistry, as well as departments like Earth and Planetary Sciences. Fantastic materials are being developed, in collaborations between material scientists in our College of Chemistry and machine learning experts in EECS, to sequester carbon in novel ways, thereby mitigating climate change. We are devising hybrid approaches to climate prediction, integrating well-established climate modeling with newer approaches of machine learning; both elements are essential. And much, much more.

We also have scholars considering the human and geopolitical dimensions, especially in the developing world, which is more deeply affected by climate change. Consider, for example, large-scale migration caused by climate-induced destruction of peoples' homelands, and the resulting geopolitical destabilization. In reference to this, my colleague Max Auffhammer said, "Bangladesh is where climate change shakes hands with nuclear war." UC Berkeley is currently hiring five faculty in Environmental Justice. We also need to address adaptation in the developed world. We must try to prevent but also react effectively to horrific fire seasons like those of Australia or the Western US in 2020, or hurricanes that constantly batter the Southeastern US, and occasionally reach the Northeastern US, as occurred with Hurricane Sandy in 2012.

What about the economic incentives, and the drivers of policy? Even if we have wonderful technical solutions, what incentives will cause us to move from one form of energy to another? What are the financial and societal implications of choosing the social cost of carbon as our metric for regulation? We have the top groups in the world in environmental and energy economics, business, law, and policy, each producing data-driven insights and solutions.

What are the security issues created by climate change? How will mitigation and adaptation choices made by one geopolitical entity impact the climate of another? How likely is ‘climate terrorism,’ and how would we attempt to prevent it? If we cannot prevent it, what is our strategy to deal with it?

Then, of course, there is the factor of human behavior. While we still have uncertainty in the physical models of climate, it is dwarfed by uncertainty due to human behavior. How can we begin to understand behavior, especially in an arena like climate change where there are unclear incentives for individual action? What is necessary to drive collective action? How do we make the long-term effects of climate change visceral to individuals today?

Our goal is to build a platform that integrates the best models and the most extensive data from STEM and HCD (the human-centered disciplines)—data from physical sensors, libraries of possible materials, economic indicators, government and international agency data, behavioral data—and use machine learning and statistics (particularly causal inferences) to propose the best interventions given limited resources.

Human Welfare and Social Justice: We are working with faculty in the Berkeley Schools of Social Welfare, Education, Public Health, Law, and Public Policy, and the Division of Social Sciences.

When I arrived in Berkeley a year ago, I wanted us to build our program in Fairness, Accountability, and Interpretability in AI and ML, bringing together ethicists, legal scholars, information scientists, social scientists, statisticians, and computer scientists. This is an incredibly important new field in which we have top scholars in all three of our core disciplines, as well as many other disciplines across campus.

Talking with my colleagues across campus, I soon realized we could also do more. Among our alumni/ae are K-12 teachers, public social workers, public health workers, public defenders, and government officials in the state of California and beyond. How can we work with our alumni/ae on the ground to frame the problems, understand in advance the unintended consequences of naïve questions asked or interventions adopted, decide which data to collect or at least how to understand the limitations of the data we have, do proper causal inference, and suggest ethically, socially, and scientifically grounded interventions?

Again, can we create platforms that take in this diverse data, do proper causal inference, and enable us to propose the best interventions for policy makers? Could the front end of this platform be appropriately tailored to people in the field, e.g., taking in health data on individuals in Memphis, TN, or Oakland, CA (with appropriately different interfaces for the two locations), and providing public health workers with a simple interface to suggest interventions for individuals? Could the data we collect, with proper security, be used to help increase the quality of interventions proposed to federal or state officials who are deciding on broad policy interventions?

Could we create platforms to address the power inequities between prosecutors and defenders in the realm of criminal justice? According to Rebecca Wexler, Assistant Professor of Law at UC Berkeley, there are some

people on death row whose convictions were based on the output of proprietary algorithms which cannot be queried, and which were used beyond their tested domain of validity. How do we take in the relevant data and recommend appropriate challenges to enable public defenders to get the right to query these algorithms? Rediet Abebe, Assistant Professor of EECS, is working with Rebecca Wexler to challenge the use of this software beyond its scope of validation. And having done this, how do we scale it so that the results of this work are available to public defenders throughout the country, even if they have no training in data science?

How do we sort through heterogeneous data to find exonerating evidence? The Berkeley Institute for Data Science, an institute within the Division of Computing, Data Science, and Society, is working with the co-founder of the Innocence Project, probing data recently released by New York City and the state of California to find evidence of police misconduct to use as exonerating evidence for people on death row; the evidence includes audio, video, handwriting on scraps of paper, spreadsheets without proper annotation and more. The heterogeneity of the data requires advances in core areas of Computing and Statistics. Moreover, we are working to make sure that the tools we create are usable by criminal defense attorneys and advocates who have absolutely no training in computing or data science. Joe Hellerstein and Sarah Chasins of EECS, and Aditya Parameswaram of the School of Information and EECS, are developing new methodologies in AI, databases, human-computer interaction, and visualization to enable this.

As we enter the Age of Interaction, with computing mediating our interactions with our public systems, we must ensure that we have protocols for collecting diverse data, methods for auditing that diversity, algorithms which mitigate rather than exacerbate bias, explanations for the outcomes of those algorithms, and methods for auditing those outcomes. We must also anticipate how these systems can be appropriated by individual criminals to state actors, and we must build in safeguards and auditing systems to avoid this. We must weave privacy, security, and trust into the fabric of these systems.

Other Areas: The areas highlighted above are far from exhaustive. Below I briefly mention a few additional multidisciplinary intellectual agendas.

- The intersection of computing and data science with economics and business poses fascinating new questions. New methodologies at the boundary of microeconomics and machine learning have enabled prediction with synthetic experiments, revolutionizing the way decisions are made on the back ends of many platforms in the tech industry—from setting the price of products, to allocation of human and capital resources. Michael Jordan speaks eloquently on the need to consider interacting networks of humans and AI; he speaks of the need for market-aware machine learning and machine-learning aware markets. This new field embraces algorithmic game theory, which has been at the intersection of computing and economics for a couple of decades, the new area at the boundary of microeconomics and ML, and the science of networks which bring together many agents from individuals to corporations.
- The intersection of computing and data science with the fundamental sciences is broad, deep, and long-standing. Astronomers have been dealing with data for centuries, and indeed developed some early work in

statistics; that interaction between astronomy and data science continues to this day. There are now attempts to ‘learn’ astrophysics models through data. Physics and computing have deep and reciprocal interaction. Methods from statistical physics are being used to try to understand the behavior of deep learning and other learning systems. Quantum computing, which over the last two decades has grown from a nascent area to an important frontier in both physics and computing, is yielding deep insights both into the physical systems and into the nature of computation. Earth and planetary sciences were already becoming data sciences before the term ‘data science’ was coined. Many of the questions we consider here have huge ramifications for climate and sustainability—e.g., climate modeling, etc.—while others focus on geological questions, and more.

- The intersection of computing and data science with the arts, humanities, media, and the interpretive social sciences is relatively new. There are obvious connections between ethics, sociology, and questions of fairness in AI, and there is much work to be done here. The way computing is remaking the fabric of our communication and interaction goes far beyond questions of fairness. Deep anthropological studies are being done which inform the more quantitative social analyses. Computing as a medium and muse is being found throughout the arts. Here at UC Berkeley, we have a group of faculty members from arts, humanities, and interpretive social science who call themselves Human Technology Futures; they include but go well beyond science and technology studies, embracing a broad array of frames and expertise. We can also look at data storytelling, or more specifically, data journalism. How do we tell stories with data? How do we contextualize the data? How do we learn to tell stories for the heart and the mind?

We have only just begun.

6. The Institutional Structure

CDSS includes three core Departments and Schools—the Department of EECS (shared equally with the College of Engineering), the Statistics Department, and the School of Information. In addition, it contains the Center for Computational Biology, the Berkeley Institute for Data Science, and the Data Science Education Program.

I have been asked repeatedly how we at CDSS plan to ‘share EECS equally’ with the College of Engineering. There are not that many large EECS departments left in the United States. MIT comes to mind as another example, and indeed MIT is ‘sharing EECS equally’ between their new Schwarzman College of Computing and their School of Engineering in much the same way that UC Berkeley plans to do. At both institutions, the more CS-oriented faculty have matters handled by the computing unit, while the more EE-oriented faculty have matters handled by the Engineering unit. At both universities, most EECS faculty feel strong intellectual connections across the broad scope of the department, and see their aspirations represented by both computing and engineering. At other universities, large EECS departments split long ago into Computer Science and Electrical Engineering (or Electrical and Computer Engineering) Departments. But even in these cases, it is sometimes not clear whether one or both departments want to be either in a computing unit or an engineering

unit or both. I imagine that, in a decade or so, there will be a book or two written on how various universities dealt with the growth of computing and data science as disciplines that go beyond engineering, while maintaining much overlap with it. Right now, the devil of co-governance is in the details.

I have also been asked why statistics and EECS have been brought together, specifically was that a ‘top-down’ or ‘bottom-up’ decision? For anyone who knows UC Berkeley, the answer is obvious—structural changes occur at UC Berkeley only if they are bottom-up. Statistics and EECS have seen their intellectual arcs bend towards each other, especially in the domains of AI, ML, causal inference, and information. There are many joint faculty members in these two departments at UC Berkeley. The two faculties took it upon themselves to collectively build the data science major, jointly with faculty from other disciplines. Working together to create the curriculum helped to refine the already developing joint intellectual vision, and to strengthen the belief that these communities must come together to educate the next generation of leaders with this vision.

The School of Information at UC Berkeley shares faculty with EECS, and, for over five years, has been offering a very popular professional degree in Information and Data Science. The School of Information’s focus on the intersection of humans, technology and information is core to the vision of CDSS.

Going beyond our core units, what institutional structures can we build to support the multidisciplinary research and educational objectives of CDSS? How do we welcome faculty and students from across campus to explore various levels of engagement with data science and computing? How do we ensure that we do not simply create more structure, but instead support fundamentally new intellectual and social agendas where the whole is greater than the sum of the parts? How do we create the institutions and platforms that incent and enable collective action on the world’s most urgent challenges?

Our institutional structure is just coming into focus over the next year or two, as we embark on the multi-stage process of becoming a college at UC Berkeley. Undoubtedly, what I say here will evolve in response to feedback from the UC Berkeley and broader UC Academic Senates; I hesitate to present even preliminary thoughts. On the other hand, each week I am approached by leaders at other universities, contemplating or just embarking on data science and computing schools or colleges. Here I present to you our current thinking, subject to consultation, modification, evolution, and approval.

In addition to our core departments and schools, our cross-divisional data science major, and our existing institutes and labs, will create new institutional structures to support the new intellectual and social agendas we aspire to build. In CDSS, we are creating a new multi-culture, which starts from shared values and vision. Central to this culture is the value of inclusivity—demographically, disciplinarily, and in terms of roles. Data science welcomes a much broader group of students than computing or statistics, students who were not given the pre-college opportunities to develop expertise or even to imagine themselves in these roles. We also welcome a much broader group of disciplines and disciplinary interests, among our faculty as well as our students. And we welcome our staff as equal partners with our faculty and students.

There is often an uneasy tension between STEM and the human-centered disciplines. We must create a culture of curiosity and mutual respect, or we will not have deep collaboration nor achieve our goals. We must have an impact on the world much more broadly than our individual intellectual agendas can support; this requires that we embrace the expertise and contributions of our staff as well as of our faculty. We will need staff leaders as well as faculty leaders to build the kinds of university-government-industry-philanthropic partnerships that will support us in moving the dial on biomedicine and health, climate and sustainability, and human welfare and social justice.

We value creativity, experimentation, and risk taking. Universities can sometimes be conservative. We need to empower our students and young faculty to cross conventional academic boundaries and take risk in the pursuit of their vision and their aspiration for societal impact. A career without failure is an abridged career. We need to de-risk the path for students and young faculty, so that they learn to be sufficiently expansive in their aspirations. These will be the generations that create the new fields at the boundary of computing and data science with other disciplines.

Building a new multi-culture requires leaders who exemplify the strands of those cultures; it is essential that we have joint faculty between our core and the other divisions, colleges, and schools on the Berkeley campus. Sometimes those joint faculty will have part of their appointments in one of our core departments or schools. But sometimes they will not fit naturally into EECS, Statistics, or the School of Information, or they might want to thoroughly devote themselves to creating a new discipline. We are therefore building a new core structure—not simply a department, nor an institute nor a center—which we currently call the Data Commons, to reside within CDSS. The Data Commons will support collaboration and incubate nascent fields.

The Commons will support research, educational, and faculty endeavors that cannot be supported fully by our existing academic units. It will be the home (or the home away from home) of groups of faculty, staff, and students who focus on specific multidisciplinary agendas and it will have new multidisciplinary undergraduate courses and graduate degrees. Many faculty members from across campus will have various levels of engagement with the Commons, including:

1. **Research:** There will be various multidisciplinary research units in the Commons, to support collaboration between our core departments and schools, and other academic units. Many of these will be of limited duration (say, 3-5 years), and will bring in the funding necessary to support these endeavors. Some of these units will reinvent themselves periodically (say every 5 years) with many of the same faculty but new research agendas. There will be shared graduate students and postdoctoral fellows whose intellectual agendas are supported by these groups. There may also be institutes which host seminars, workshops, and other focused activities, with visitors as well as Berkeley students, postdoctoral fellows, and faculty. Visionary staff will help us to build research, government, and industrial partnerships for these units.
2. **Curriculum Development and Teaching:** Many faculty members from across campus develop curricula for and teach in our data science education program for undergraduates; many of these courses are cross

listed with departments and schools across campus. There is also interest in creating new majors which span computing and data science and other disciplines on campus. We expect to have a good number of new master's programs which will be joint with other entities on campus. Finally, there will be the creation of new multidisciplinary Ph.D. programs. Faculty from other departments can engage with this teaching and curriculum development through temporary affiliations with CDSS, as specified in MOUs with their home departments.

3. **Faculty-Holding Groups:** The primary purpose of these will be to nurture nascent fields. Since the holding of faculty appointments is permanent, these will be in areas that have very clear intellectual agendas involving computing, data science, and information in collaboration and conversation with other academic disciplines across campus. Some of the faculty in these groups may have a positive fraction of their appointments here; others will simply be affiliates (0%-time appointments). In addition to their research agendas, these faculty will create new educational programs—undergraduate, master's, and Ph.D. Some of these groups may eventually become departments or schools, while others may maintain their identities within the Commons.

How do we evaluate scholarship in the Data Commons, particularly for faculty hiring and advancement? How do we evaluate the quality of our educational programs and make sure that our students are well-served? I have thought about these types of issues for my entire career, indeed since I received my Ph.D. in mathematical physics from Princeton in 1983. I have always straddled fields, and mentored literally hundreds of multidisciplinary graduate students, post-docs, researchers, and faculty over the years. Moreover, I co-founded and led several multidisciplinary research labs at Microsoft for 23 years before moving to UC Berkeley last year.

The Data Commons will provide a guide which clearly articulates the characteristics of excellent multidisciplinary scholarship. We will judge our faculty in hiring and promotion based on this guide, which will lay out in general terms what are our criteria for excellence in multidisciplinary work. Within any specific multidisciplinary unit, the relevant hiring or promotion committee—which will usually include faculty from other departments or schools in addition to those in the unit — will clearly explain how the general criteria will be interpreted in the context of that specific multidisciplinary endeavor. We will rely on these criteria when we hire faculty and when we request reference letters for hiring or advancement. We are acutely aware that different fields have different publication venues and conventions. We will make sure to identify these in advance, and to support our faculty in advancing their research agendas across these different venues and conventions.

The educational programs will similarly have general criteria for excellence in multidisciplinary education, which will be adapted to the evaluation of specific programs. We will make sure that there are clear career paths for these multidisciplinary students. We will also determine appropriate levels of guidance and support for both students and young faculty in these emerging areas.

The type of criteria we set up for evaluation of multidisciplinary work will also be useful in the context of evaluating faculty who are joint between one of our core departments or schools and other academic units on campus.

Figure 1. Aerial view of the planned new home of CDSS in the northwest part of the UC Berkeley campus. Images courtesy of Weiss/Manfredi and Gensler architectural firms. Note: Design is provisional and in development.

Figure 2. View of gathering spaces for the campus community.

7. The Physical Structure

It is not just institutional structure that determines the success of an endeavor like the Division of Computing, Data Science, and Society. Our physical space must provide fertile ground for the new multidisciplinary

cultures we aspire to create. There are some buildings that enable deep collaboration, e.g., the home of the Isaac Newton Mathematics Institute at Cambridge University, and others which inhibit it. At UC Berkeley we are lucky to have the opportunity to design a new home for CDSS.

A portion of our new building will welcome the entire university community, with lovely outdoor space, a large cafe, a gallery, institutes, lecture halls, and shared educational endeavors. It will be a nexus for STEM and the human-centered disciplines, a place where undergrads, grad students, faculty, and staff from across campus can discover data science and make it their own.

A portion of the building will be research focused, and home to faculty, staff, postdoctoral fellows, and grad students. For this, we envision less individual space and more space that support mixing, coming together, the serendipity of interaction over time, leading to deep collaborations, fostering the development of new fields, and addressing society's most urgent needs. Our design incorporates the lessons learned in new academic buildings as well as in the technology industry. We are planning a building which comprises a series of neighborhoods on different scales, including convening spaces of small and larger groups of faculty members with associated grad students, post-docs, and staff. The boundaries between neighborhoods will be porous. Think of neighborhood piazzas in Rome. I usually have my morning espresso in the piazza nearest my hotel, but I often wander to other squares, each with its own special character; there are more intimate squares and there are grander squares like the Piazza Navona. New intellectual and social agendas will be born in these squares, agendas that will give us purpose and make our hearts sing.

Figure 3. View of unstructured interaction space for research.

Figure 4. View of event space on the roof, looking towards the UC Berkeley Campanile.

Acknowledgments

So many people have contributed to my understanding of data science and computing over the years. I’d particularly like to thank a few of my UC Berkeley colleagues: my three dear friends who brought me to UC Berkeley—Michael Jordan, Scott Shenker, Bin Yu; my new colleague, Ion Stoica, who has deeply supported our efforts; the founders of the original Berkeley Data Science Division, who have been trusted advisors before and during my time here—Cathryn Carson and David Culler; the innovative leaders of our institutes: Saul Perlmutter and David Mongeau, David Harding and Claudia von Vacano; our phenomenally talented and dedicated Associate Deans in CDSS—Hany Farid, Deb Nolan, Oliver O’Reilly, Nathan Sayre, Steve Weber, and Kathy Yelick; the rest of the CDSS leadership team—Meredith Lee, Cindy LuBien, Rebecca Miller, and Rebecca Ulrich; and of course my husband, collaborator, and constant companion on our shared intellectual and institutional journey during last quarter of a century—Christian Borgs. I am also forever grateful to our visionary lead donor, who is deeply committed to the future of UC and the students of California, and who made an unprecedented gift towards our new building.

Disclosure Statement

Jennifer Chayes has no financial or non-financial disclosures to share for this article.

References

AI and responsible data science. (2020) Harvard Data Science Review. <https://hdsr.mitpress.mit.edu/ai-and-responsible-data-science>

Data science education. (2020). Harvard Data Science Review. <https://hdsr.mitpress.mit.edu/data-science-education>

Perspectives on data and data science. (2020). Harvard Data Science Review. <https://hdsr.mitpress.mit.edu/perspectives-on-data-and-data-science>

Buolamwini, J., & T. Gebru. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, 81, 1–15. <https://www.media.mit.edu/publications/gender-shades-intersectional-accuracy-disparities-in-commercial-gender-classification/>

Garber, A. M. (2019). Data science: What the educated citizen needs to know. *Harvard Data Science Review*, 1(1). <https://doi.org/10.1162/99608f92.88ba42cb>

He, X., & X. Lin. (2020). Challenges and opportunities in statistics and data science: Ten research areas. *Harvard Data Science Review*, 2(3). <https://doi.org/10.1162/99608f92.95388fcb>

Jordan, M. I. (2019) Artificial intelligence—The revolution hasn’t happened yet. *Harvard Data Science Review*, 1(1). <https://doi.org/10.1162/99608f92.f06c6e61>

Lampson, B. (2014). Lampsonfest Dinner, Cambridge, MA.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366(6464), 447–453. <https://doi.org/10.1126/science.aax2342>

Wing, J. M. (2020). Ten research challenge areas in data science. *Harvard Data Science Review*, 2(3). <https://doi.org/10.1162/99608f92.c6577b1f>

©2021 Jennifer Chayes. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](#), except where otherwise indicated with respect to particular material included in the article.