Data Sciences @ Berkeley
The Undergraduate Experience

Sketch 1.2
By the Data Sciences Education Rapid Action Team
Version: 1/19/2015

Outline

1. Executive summary

Providing access for our students to the fundamental structure, principles, and ramifications of data-analytic thinking is a key educational desideratum for an increasingly data-rich world. This document offers a concrete proposal for a curriculum that will make cutting-edge, critical engagement with data an integral feature of a new liberal arts education and a core interdisciplinary capacity shared by all Berkeley undergraduates. Critical thinking with data is an area of key importance to many fields of study, central to job opportunities in many industries, and integral to personal and professional decision-making. Tackling this challenge for our students with an ambition distinctive to Berkeley means drawing on our multiple strengths, comprehensive excellence, and campus-wide scale. It represents an opportunity to markedly improve our students' experience by aligning the desire to acquire data science capacities with the structure of major programs and with course innovation. Taking off from signals of student interest – large inflows of students into computing and statistics classes, to start with – it creates pathways through the curriculum that open doors for our students into data science in the diverse ways that respond to their future trajectories, current needs, and varied backgrounds.

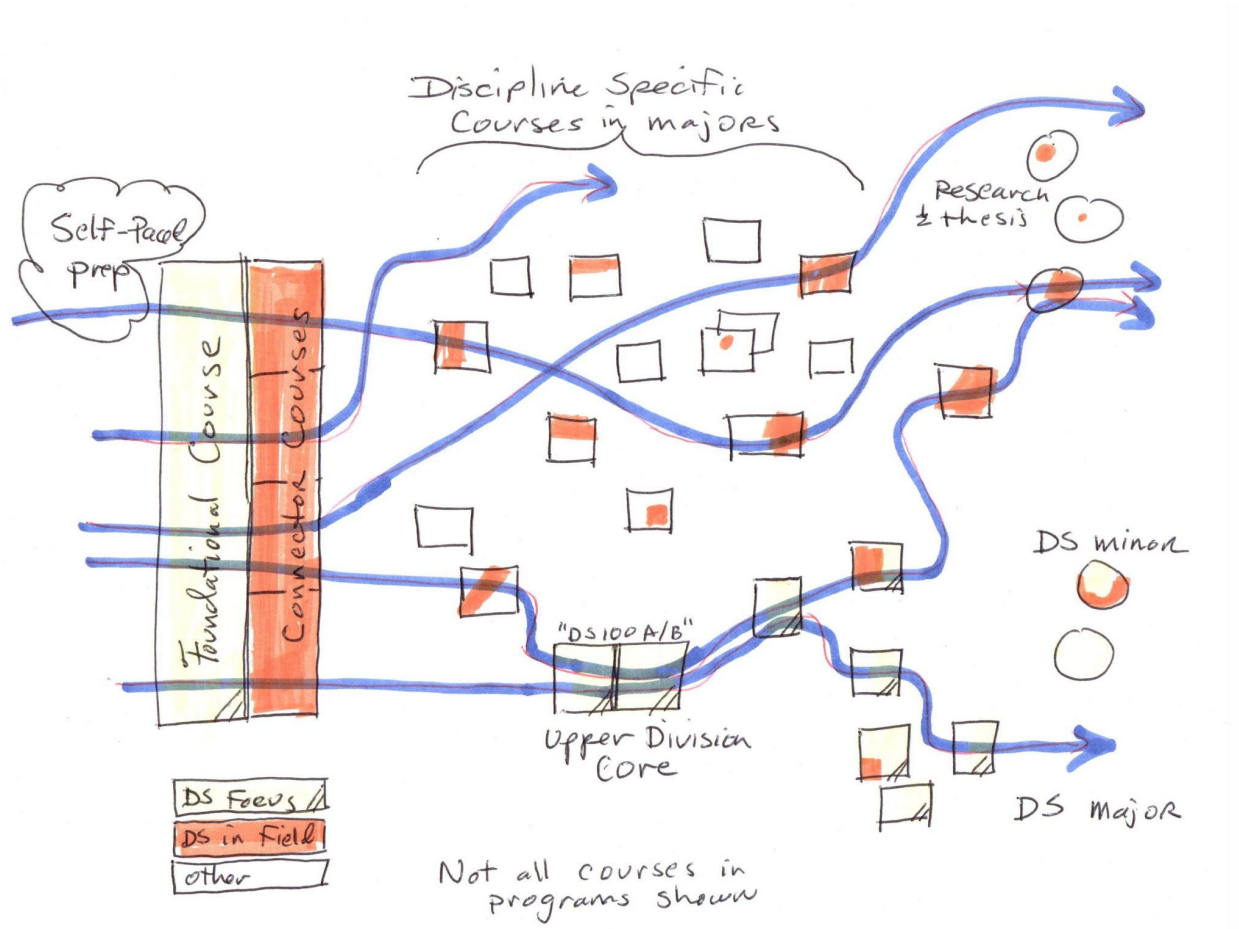
Based on our conversations with faculty across the Berkeley campus, we formulate a multi-tiered structure for a comprehensive data science education program that displays both breadth and depth.[1] It is anchored in a new foundational offering, or suite of courses, intended to scale to the entire freshman class. The foundational offering will present key elements of introductory computational and inferential thinking in an integrated fashion, cementing conceptual understanding through direct experience with data. Closely allied with this is a suite of "connector" courses, rigorously engaging many disciplinary areas by means of focused projects and framing a critical understanding of the social and ethical context of data and analysis, while tailoring material to the diversity of student backgrounds and interests. Built on this foundational offering and connectors are multiple opportunities to advance teaching and learning campus-wide. Departments and programs in many areas will have the chance to evolve their curricula in ways that support disciplinary learning, critical thinking about data, and undergraduate research.

Likewise building on this foundation, intensive treatment of data science is rooted in a novel one-year upper-division core, which provides a gateway into both a broadly useful minor, which can serve students across campus, and an inventively designed major with deep coverage. These actions, taken jointly, will bring our deep campus expertise in data science research and professional training to serve our undergraduates. They will establish Berkeley as the leader in a national landscape of institutions that are invested in data science, none of them offering the breadth of vision and ambition that our proposal reveals.

[1] A schematic of the proposed curriculum is included at the end of this summary. A fuller discussion is found in Section 5, Student experience in the future, and Section 6, Proposed curriculum structure.

From our examination, the key challenges in instituting such a program also represent important opportunities. Many majors have accumulated requirements over time and feel highly constrained. Accommodating and utilizing the growth of data science capacities in the student body will involve thoughtful re-examination. Creating a network of connections between the core and the extensions, as well as between the minor and host majors, will require faculty engagement. That engagement is precisely what will propel our university into a 21st-century posture in its engagement with data across the curriculum. Indeed, the courses that the foundational offering and the corresponding upper-division data science core displace are rooted in seminal texts of decades past; a new seminal body of teaching materials will likely need to come out of Berkeley. Creating this new structure will require new resources to be invested in multiple areas, from infrastructure to curriculum design to academic staff and faculty FTE. The huge growth in demand for courses that only partially address our students' data science needs have already introduced severe imbalances; the systematic process of planning and resourcing can rationalize priorities.

This document seeks to provide a framework for a broader discussion among Berkeley's faculty and campus leadership. With the unique depth and breadth of Berkeley faculty in data science, development of a freshman foundational offering could begin immediately. Initial pilot offerings could be ready as early as 2015-16 and should be on the scale of a couple hundred students with a handful of early-adopter connectors. An upper-division core course could be piloted shortly after. A concerted 3- to 4-year growth and learning process would be required to refine the foundational offerings to address the full complement of student abilities, backgrounds, and interests. A similar timeframe would be needed to build out follow-on opportunities in domains across campus, as well as flesh out key upper-division courses for the minor and major. With agile planning and deliberate speed, Berkeley can stake a claim to leadership in transforming this future-facing aspect of undergraduate education with signature effect.



Student pathways through courses touched by the proposed data sciences (DS) curriculum.

- Time proceeds from left to right.
- The vertical axis spans the breadth of student interests and majors.
- Courses are schematically represented by boxes; research opportunities, by circles.
- Student pathways are marked in blue.
- Course content focused on DS appears in yellow with a hatched corner.
- Course content applying DS to application fields appears in orange.

2. Task force operation

This document is the work of a task force operating through the fall of 2014. In response to signals of student demand and to high levels of faculty interest, the Data Sciences Education Rapid Action Team was formed mid-summer by the Chancellor and Provost. The team was asked to move quickly and report on a short timeframe. Members are:

Cathryn Carson (co-chair), History	clcarson@berkeley.edu
Bob Jacobsen (co-chair), Physics and Interim Dean, L&S Undergraduate Studies	jacobsen@berkeley.edu
David Culler, EECS	culler@berkeley.edu
Michael Franklin, EECS	franklin@cs.berkeley.edu
Michael Jordan, EECS and Statistics	jordan@eecs.berkeley.edu
AnnaLee Saxenian, Dean, School of Information	anno@ischool.berkeley.edu
Jasjeet Sekhon, Political Science and Statistics	sekhon@berkeley.edu
Bin Yu, Statistics and EECS	binyu@stat.berkeley.edu

In weekly meetings since early August, we have focused on three things: understanding the Berkeley landscape, learning about initiatives at other universities, and developing a curriculum proposal. In this first round, we have been lucky enough to exploit insights from several faculty and instructors already teaching this material. We have talked with multiple deans and a subset of department/program chairs and staff, with campus leadership, Research IT (Office of the CIO), ETS, the Library, D-Lab, and BIDS. We are especially glad to have had analytic assistance from the Office of Planning and Analysis and L&S Deans Office staff.

2.1. Invitation

In line with the Chancellor and Provost's request, our proposal has been developed on a fast timescale. Our data collection, consultation, and design process is continuing; we are taking an experimental and iterative approach. This document is explicitly an appeal for feedback and input, especially helping us identify errors we have made or issues we have missed. We warmly invite comments to any member of the task force.

We have tried putting the big picture in the main text and details in appendices. We would be glad for responses to both.

3. Framing the initiative

3.1. Vision

Ever-growing demands on our capacity to reason reliably, intelligently, and creatively from data will change how our students tune their academic trajectories, carry out their careers, and live in their world. In the context of thorough-going transformations in the ways our society engages with data, all educated individuals should be able to interpret statements of inference drawn from empirical data, such as those that now appear routinely in the news or in statements about financial or medical matters, and also to utilize statistical reasoning and to be able to acquire, manipulate, and process appropriate empirical data in their decision-making. To do these well requires certain computational abilities, as well as enough experience and understanding to distinguish causation from coincidence and avoid inferential and cognitive biases. It also requires understanding of how data is collected, processed, and classified in order to think critically about the social and ethical implications of data. In the context of a campus-wide undergraduate curriculum that stretches across diverse fields of study and extends from entry level through engagement in research and capstone experiences, we owe it to our students to provide opportunities to exercise these abilities at stages from literacy to competency to mastery in preparation for the widest variety of life paths.

The initiative that our task force is proposing comes amid the convergence of computation, massive new data streams, and sophisticated strategies of inference that are changing the face of contemporary life. The “data sciences” – a toolkit of rigorous and imaginative approaches to working with data from diverse new sources and at all scales – are emerging as instruments of the future for tackling a wide range of problems of intellectual, personal, and societal import. They give us new ways of grasping patterns, collectivities, and systematic effects that remain invisible to us without statistical and computational tools; of understanding the linkages from data to knowledge to decision-making under conditions of uncertainty; of exploiting domain-specific computational possibilities fluidly and reliably and seeking cross-fertilization across them; and of critically engaging the constructive and creative possibilities opened up by data collection and computation, as well as their challenging ethical and social entanglements.

3.2. Rationale

Taking the amorphous term “data science” to point to something uniting computation, statistics, information management, and application-domain engagement with real-life data, we are persuaded that the interdisciplinary phenomenon behind it is significant and real. More than that, we see the inviting possibility of embracing a reinvention of statistical education in the era of pervasive computation. With today’s affordances we can now can give our students the chance to learn by hands-on manipulation, centered on projects using real data to integrate the teaching of computational and inferential thinking. Compared to past approaches that veered

between abstract formalism and push-button techniques, a unified approach grounded in real-world relevance promises to make the teaching of data analysis actually "stick." Moreover, existing curricula have fallen into a consumer/producer model of computational and inferential ideas, whereby statisticians, applied mathematicians, and computer scientists produce the techniques and algorithms used by others. A better model recognizes that real-world problem domains generate new challenges that are often best recognized by researchers who are steeped in a domain. Similarly, producers and consumers of data were historically split but are now sometimes integrated – often in fractured and challenging ways. Finally, real-world relevance raises a wide range of new issues that we have not confronted in the past and demands the contributions of social scientists, humanists, and other scholars, ranging from research design and communicating results to complex ethical challenges associated with data collection and analysis. Thus, not only is real-world relevance likely to enliven the teaching of data analysis, it reflects the underlying dynamic of this domain.

The payoffs for our students are targeted to their intended paths through Berkeley and beyond. While such diversity is essential and built into this proposal, we simultaneously have the opportunity to do something distinctively integrated at the campus scale. The domains that will be informed by the data sciences in the future extend beyond academic research and data analytics in industry, reaching into a wide range of real-world careers. Every major on campus harbors students who will critically engage with data as producers AND consumers in individual decision-making, civic settings, and their professional lives. We can create pathways for them out into a world that is being transformed by the possibilities of data science.

Note the counter: good CS students enrolling in 218, 290 …

The demand for this curriculum is already making itself felt. *Already the bulk of our undergraduates are taking entry-level courses in computation and (separately) statistical reasoning*, with massive waves of enrollment reaching into the upper division. Iconically, CS 61A, the introductory computer science course *for majors*, is expected to serve 2,500 students this year. Combining CS 61A enrollments with other introductory computing offerings brings the number to nearly 5,000 undergraduates this year. This growth at Berkeley matches a trend seen nationwide. The bulk of students enrolling in introductory computing courses, moreover, apparently do not intend to specialize in computer science (in either of Berkeley's two CS majors in the College of Engineering within EECS or in the College of Letters & Science), with evidence of particular growth in the social and physical sciences. In statistics, introductory courses inside and outside the Statistics department serve more than 3,500 students and have been growing as well. Over the past 5 years, the number of statistics majors has grown from 80 to 400; L&S CS majors from 140 to over 700; and EECS majors from 900 to over 1200, with a shift in balance from roughly equal in EE and CS to 3/4ths CS. Remarkably, too, the number of students across campus investing in double majors has tripled in the last five years, the largest numbers of them showing up in statistics, computer science (L&S), economics, applied mathematics, and EECS. In 2013-14, *the majority* of undergraduates in statistics left Berkeley with two (or more) majors.

Thus there is increasing student interest in fields related to "data science," even if students can meet it on our campus only by mechanically assembling a series of technical offerings. While it

will take careful empirical study to pin down the drivers, our initial inquiries into the conjunction of student interest and societal transformations suggests that our curriculum is ripe to be rethought and reconfigured. We find it inviting to do so, moreover, in the context of serving Berkeley's full undergraduate student population. We ask the campus to envision providing this experience to those students who have not always felt fully welcomed into the data sciences until now, addressing challenges around preparation, intellectual or disciplinary orientation, and social processes of inclusion and exclusion.

3.3. Principles

Our envisioned curriculum assumes that every individual who studies at Berkeley should be prepared to understand and develop points of view based on the analysis of data as well as evaluate arguments made by others. Our students should learn to think about the personal, social, and scientific contexts in which data are gathered, and they should be able to think through the philosophical, moral and ethical, political, legal, and economic consequences of tackling individual and societal problems with this set of tools. They should learn how to ask, "Would it be possible to answer that question with the appropriate data?" and to think clearly about how to obtain the appropriate data. Our students should learn about the computational and inferential underpinnings of data analysis and should learn about these underpinnings in an integrated manner, in the context of problems that matter in the real world. In settings and at levels appropriate to their own trajectories, they should be empowered to conduct their own analysis of data and to think creatively about how to work with data, as well as how to communicate the findings of complex data analytics to non-specialists.

We can offer these opportunities to our students with Berkeley's signature across-the-board quality and rigorously critical engagement. Core to our proposal is the determination that our students grapple with the limitations, inherent inconsistencies, and missing elements in data, the effects these have on the inherent quality of decision-making, the methods for assessing the risks of those decisions given incomplete data and understanding, and the alternative methodologies that may be brought in as complements. In this process we underline the critical role of visualization and other approaches that communicate data and analysis to assist in exploration, understanding, and decision-making. We are framing data science as a process of discovery, interpretation, analysis, and extraction to underwrite choices, interpretations, and decisions in individual, organizational, and societal settings. The data revolution raises many social, political and ethical issues, and the broad student population should be taught to reason critically about these issues.

Our proposal thus grows from the following principles:

- The University should make it possible for all undergraduates to gain experience with data science and encourage students to take fullest advantage. The desire to acquire these capacities should not dictate a student's major. Appropriate offerings should be available for students with the backgrounds and interests typical of a wide spectrum of majors.

- Foundational course offerings that develop these computational, statistical, and critical abilities should be available in a form that scales to the size of the entire undergraduate student body and is accessible across that diverse body at the lower division level. These offerings should provide a strong foundation in data literacy/numeracy and open the door to further learning on that foundation. This foundational data science content should be part of the core experience of undergraduate education at Berkeley.
- A wide spectrum of courses should be able to take advantage of students' foundational computational, statistical, and critical capabilities in addressing data-oriented aspects of diverse subject matter. For example, advanced courses in many fields besides computing and statistics may bring in new, data-oriented modules or topics, and they should be encouraged to do so.
- Additional courses that address data science concepts in depth, up to and including the ability to carry out a minor, should be available to students in many fields and accessible within their major requirements.
- A data science major should be developed as a cohesive major in its own right. This can and should be done with the broad understanding of data science that underwrites the rest of this initiative. The design of the major must be collaborative and adaptive as the demands of the field and the needs of students evolve.

Together these principles guide the formation of an overall data science education plan that supports the emerging needs of our students in graduate programs, professional careers, and leadership roles. Ultimately such a plan would need to be resourced, coordinated, and executed by the faculty.

3.4. The challenge and the call

This curricular change would have implications for many departments, programs, and majors. Beyond sheer challenges of scale, there are intellectual and practical considerations growing from Berkeley's differentiated scene. For this curriculum to be integrated with students' needs in their majors, it will only succeed if we have detailed discussions, program-by-program, of existing pathways, prerequisites, requirements, and follow-on courses. In some areas our students already have packed-full schedules. Meeting their needs requires thoughtful exploration of the ways in which existing offerings may not serve them optimally. In other areas our students need entirely new offerings, additional support, and new infrastructure. In all cases, building a curriculum for our students integrally involves the judgment and coordinated engagement of domain-area faculty and the Academic Senate.

The process calls for a data-driven approach that proceeds iteratively and experimentally, gathers quantitative and qualitative data on the student experience, and tracks the results of innovations. It requires deliberate attention to challenges around equity and best practices around inclusion, in ways that are critical to the success of the enterprise. It needs to be done in a thoughtful process of piloting and expanding offerings in a fashion that delivers high quality to our students, devising both an educational model and a campus infrastructure that works at

scale. Finally, a data sciences curriculum requires significant investment of new resources, as we detail below.

And yet by tackling the challenge at scale, Berkeley can do something that no other university has imagined. That is to integrate the data sciences as a core component of liberal education. As other schools rush to create narrower data science programs, Berkeley has shown the intellectual ambition to define data science capaciously by engaging faculty members campus-wide. Capitalizing on our deep strength in data science *and* our broad-spectrum excellence, this program can be a Berkeley signature in conception as well as a defining feature of our undergraduate education. Berkeley has a leading role to play in data science because of our faculty strength, our exceptional graduate programs, and our professional degree offerings. We owe it to our undergraduates to extend the continuum of student experience to them as well.

4. Student experience now

Many students entering Berkeley today were born after the rise of the web, having grown up in a world of continuous communication and interaction without bounds. With much of the world's knowledge collected and indexed for search at the slightest inclination, they often acquire understanding by "questioning down" – searching on a thread and gleaning underlying concepts by traversing links and searching additional fragments – rather than building up. Our students are subjects of continuous data analytics, with services delivering them targeted news, advertisements, social connections, and more based on observations of their actions. They are aware that their experience is modulated in this way, just as they have been exposed to the promises opened up by the possibilities of political campaign analytics or personalized medicine and to the excitement of new techniques put to work in their areas of study. It would be no surprise that they should see becoming an educated person to include gaining the ability to garner their own data, perform their own analytics and interpretation, and think critically about the social and technical contexts within which those actions take place.

Yet the current data science experience at Berkeley – gaining computational and statistical capabilities and applying them meaningfully to real data, with an understanding of the challenges, pitfalls, and associated issues – is chaotic and anachronistic, as it is at most other institutions. Students are gaining basic skills in large number through courses that are not designed for their array of needs and that pay almost no attention to the potential integration of computational and inferential thinking, and only sporadically to designing intelligent research questions, collecting data, understanding it in context, and communicating its import.

As filled out in Appendix 1, large cohorts of Berkeley students are in fact enrolling in one of several courses covering introductory computing (nearly 5,000 this year) and introductory statistics (over 3,500 this year, plus significant numbers taking statistics at community college).[2] The landscape of course-taking is complicated by the existence of multiple versions of introductory offerings inside and outside the core departments.[3] From the diversity of offerings we can conclude both the widespread need for this material and some sense that courses initially designed by departments to their own expectations are not necessarily serving broader campus needs.

Once admitted, Berkeley students face an immediate schism between the desire to gain basic data and computational capabilities on the one hand, and the requirements of potential majors

[2] The existence in several colleges (including L&S) of Quantitative Reasoning requirements that can be satisfied by these courses is worth noting, although it is by no means a dominant driver of student enrollments. That is true even in statistics, where most students who meet QR by course-taking end up satisfying the requirement.

[3] For instance, E 7, Math 128, and Stat 133 are computing courses, while Math 10A/B (developed for biology majors and now serving psychology students) and Public Health 142 provide introductory statistics content. Appendix 1 gives more details.

of study on the other. The vast majority of undergraduates taking introductory courses in both computer science and statistics are Letters & Science undeclared. This reflects both the size of College and its commitment to a liberal arts education in which students gain breadth and critical thinking before declaring a major. However, only a handful of majors recognize either computer science or statistics courses as part of the major, other than as an elective or, in some cases, as a prerequisite.[4] Computer Science does not even accept Statistics courses as fulfilling its own statistics requirement. Many science and engineering majors view their requirements as completely filling their students' schedules, so there would be no room for a computing or statistics course. Evidently, students take such courses anyways before declaring the major. And although many of our students work hard to gain this capability, our faculty rarely utilize it or cannot count on it, and major programs have not adapted to it. Despite the growth in introductory computing enrollments, only a relatively small number of courses outside of Computer Science have an introduction to computing as a prerequisite.[5] Courses in many disciplines have introductory statistics as a prerequisite, but faculty report that relying on students to have learned statistical thinking in those settings to be able to apply it immediately in domains is often a pedagogical mis-step. It is also worth observing that as many students as there are who take introductory statistics or computing at Berkeley, there are also large numbers who do not, and their distribution across areas of study and interest is not remotely uniform. The real or perceived inaccessibility of these courses, and the lack of a common baseline that could build on them, undermines the ability of Berkeley faculty to set higher instructional goals in their own domains. In particular, the current Quantitative Reasoning requirement (which is centered in L&S and used by some other parts of campus) does not go very far toward meeting this need.[6]

Moving further, to garner more than the most basic skills in this area requires taking several advanced courses that contain a lot of additional material. Tellingly, an analysis done within Computer Science concluded that the material most important to a data science program is contained in small sections of several of the current courses. To get these with the current courses involves essentially completing the major, and some fraction of students choose to do just that. The current over-enrollment in these courses, as in Statistics, further forces that choice, since without declaring the major, getting a seat is these courses in unlikely.

We see other important needs falling through the cracks. Undergraduate courses addressing critical questions about data science in societal context are essentially absent from the curriculum, leaving students to jerry-rig an interpretive frame. For guidance in working hands-on

---

[4] Economics and Applied Mathematics allow either computer science or statistics courses to form an external cluster, but they cannot be combined. Statistics allows computer science to be an application cluster.

[5] Courses outside of EECS that build on Berkeley's introductory computing offerings can be found, to our knowledge, in Bioengineering, Chemical and Biomolecular Engineering, Civil and Environmental Engineering, Mechanical Engineering, and Cognitive Science.

[6] Most L&S undergraduates who do not pass out of QR at entry, either by standardized test scores or by adequate performance on a mathematics exam, satisfy the requirement by taking Stat 2 or, in lesser numbers, Math 23. Almost none take a CS course.

with data, our undergraduates have access to a relatively small number of upper-division courses in the application domains. To support their own learning, they often have to seek out extra-curricular resources such as the Library Data Lab (heavily used by Economics students) or cannibalize training offerings targeted at graduate students and faculty in the Social Sciences Data Laboratory (D-Lab). The computing infrastructure for classes requiring instructional work with datasets (compute resources, software or virtual environments, staffing, space) is also highly constrained. In the absence of a campus-wide instructional computing baseline, Berkeley's infrastructure is de facto limited to only some parts of campus, and some departments have given up on the possibility of teaching courses that require significant projects of data analysis. This infrastructure is a prerequisite for the broad-based data science education we envision.

Finally, dedicated data science offerings at Berkeley are sparse, even as they are in high demand. With the exception of the runaway attraction of Stat 133 (Concepts in Computing with Data, expected to serve 700 students this year), two recently introduced advanced machine learning courses in Computer Science and in Statistics, and an experimental data science course in CS, the deep inter-relationship of statistics and computer science is missing from the curriculum. Even more glaring is the absence of understanding how to apply data science methods to real-world issues specific to the domain areas and the messy, noisy data that go along with those. The application domains are disadvantaged, and space for reflecting on data collection and on contextual or societal issues is basically absent.

Overall, rather than pathways through the curriculum, our students encounter somewhat fragmented courses with pieces of relevant material scattered throughout. Few majors accommodate, much less place explicit value on such capabilities, while more advanced courses in those majors that could benefit can rarely take advantage of the abilities that some students gain. There are some bright spots of integration with domain-area problems, but also many dark patches, and there is a real gap in providing critical thinking skills with and about data. These consequences extend beyond the undergraduate curriculum. Many graduate programs recognize the importance of data science abilities in important research areas but must employ work-arounds to fill the voids in current undergraduate offerings, while our students' opportunities to meet the huge professional demand in industry and other real-world settings go unmet.

Appendix 1 - Current student experience

5. Student experience in the future

We foresee a very different student experience: one that would better prepare our graduates for their professional careers, for graduate studies, and for functioning as an informed person in a data-rich world. The posture of the University that is conveyed in the admissions process should be that data competency is important, it is a part of everyday life, it involves certain computational, statistical, and mathematical abilities, but it is it not “just for nerds” any more than reading is “just for bookworms” or understanding American institutions is “just for policy wonks.” It involves design, presentation and communication; it involves thinking critically about what and how; it involves judgment and ethics. It should be clear that access to this basic educational element is available to all students, regardless of choice of college, eventual selection of major, or pre-college experience.

As such, course offerings need to be created that span the diversity of abilities and backgrounds of entering students and that are tailored towards making the study relevant in their lives. A set of expectations should be put in place, and pervasively communicated, that this material is valuable and appropriate for all students, including those who have previously felt socially or educationally excluded. Major programs need to be examined and adjusted to ensure that there is opportunity for a student to gain such competence. In many cases this will require only including the foundational data science offerings in the quantitative breadth requirements of the program. For a few specialties it will require re-examination of the introductory courses for majors, which may cover aspects of computing or statistics in an isolated manner.

Gaining this competency should, with time, enrich the experience in a variety of courses that are not about data-related skills, but which, increasingly, touch in some manner the data itself. For this to happen, faculty need to see themselves as more than consumers of data products, and certainly not just as objects of analytics (while recognizing that we all are), but potentially as producers of data, data analysis, and presentation. We may find that it is important to encourage faculty to take part in the foundational courses that form part of the modern student experience, but not of their own.

It should be clear these data abilities are not just a modern vocational fix; they have depth. Not only are they widely relevant, but further studies can lead students from competency to effectiveness, expertise, and judgment in all spheres of a data-rich world. This may involve follow-on courses relevant to particular domains, the opportunity to pursue a minor, or even to gain true expertise by making data science itself one’s major. These additional pathways will fill what is today a substantial gap in the preparation for graduate study in a number of fields whose frontiers are becoming increasingly data-centric.

5.1. Foundational course

Delivering this experience begins with the creation of a new course, or suite of courses, on modern data science. This is not just a matter of repackaging material from existing courses or breaking them into modules to pick from. Its development does begin by drawing upon elements in the introductory computer science courses, CS 10 and CS 61A – not just programming methods, but concepts of abstraction, representation, and algorithm – and equally upon concepts in the introductory statistics courses, Stat 2, 20, and 21, not just formulas, but concepts of summarization, uncertainty, and inference. These would be truly integrated. One learns basic programming techniques by computing descriptive statistics of real data and learns basic concepts of statistics by implementing them and seeing what understanding they convey about matters of relevance in our students' lives. The symbolic formulation appears hand-in-hand with its algorithmic implementation and understanding of its application. In this setting, concepts of probability, estimation, and inference will be taught in a computationally grounded fashion that represents the modern practice of statistics, rather than the "hand calculator" orientation of the seminal texts, which have pervaded statistics educational practice for decades. A question as basic as "Are these two collections of numbers significantly different?" where, say, one of them represents observations of a test group and the other of a control group, is better answered by computing properties of permuted samples[7] of both than by comparing classic statistical summaries, and it is an interesting computational experience to do so.

Distributions are not just idealizations of what the data might look like, but real things that are computed and visualized, and that can be related to models and to concepts of uncertainty. The development of such statistical techniques motivates learning more sophisticated computing concepts, such as permutation, sorting, data structures, and higher order functions. Inference, hypothesis testing, and significance are grounded in experience with data.

Such a pedagogic approach should not be viewed as "going soft on the math." It recognizes that conceptual understanding can be developed, perhaps even *better* developed, through direct experience and computational actions performed with one's own hands, rather than through symbolic manipulation. The symbolic formulation takes on new meaning as a concise representation of both the algorithmic process and the empirical understanding. This, in part, recognizes that students today are more familiar with computational manipulation of representations of the real world than they are with symbolic idealizations of it. But, more so, it

[7] In this technique, you repeatedly compare new samples made from random mixtures of elements from the test and control groups. If those two groups are really different, the mixtures will lie somewhere in between them. If the differences between the test and control groups are just due to random fluctuations, then there will be other randomly-selected samples with similar differences. This is an example of modern inference techniques that rely on the availability of lots of computation, and therefore can handle complex data without having to boil it down to just a couple of representative summary numbers.

allows exposure to computational forms of analysis that are hard to describe symbolically, such as resampling, regression, approximation, and model formation, which are what they will actually use. And even more so, it provides a path to learning these concepts in a setting where it is natural to think about where the data comes from, what it represents and what it does not, what the analyses mean, and how to relate this understanding to the deluge of data and analytics they encounter every day.

Indeed, in formulating this course, we believe it will be necessary to draw in some topics from courses that today follow the isolated statistical and computational introductions, such as certain important data structures and algorithms, the use of databases, and visualization techniques. But by tackling a carefully selected set of computational and statistical topics in this integrated, hands-on manner with data, it will be possible to actually learn more in less time. More of what is learned will be useful as well as more likely to "stick." The experience of Stat 133 (Concepts in Computing with Data), in particular, shows huge promise for the lower division. There are no computing or statistics prerequisites for this course, and the ballooning enrollments and high levels of student engagement underscore our confidence in this approach. We have also taken significant lessons from the recent launching of Math 10, which includes some elements of probability and statistics in a two-course sequence intended for majors in the life sciences.

We therefore propose to create a **foundational course providing a one-semester experience for all students**. It will be structured around **a common 4-unit "core" course**, to be taken at the same time as one of **multiple, notionally 2-unit "connector" courses** that run in parallel with and "plug in" to the foundational core. The core course is a large, single course providing a shared experience to all students, to be taught by a dedicated team through a combination of in-person, on-line and group study methods, as we discuss in more detail below. The connector courses then bring in faculty from every part of campus to develop focused content and exercises relevant to broad areas of student interest. Students entering Berkeley as freshmen would take the core and the connector together during either the Fall or Spring semester of their freshman year.

The introduction to formal data science concepts should not be isolated from their application. In both the core course and the connectors the principles of data science will be conveyed with use of real-world content. Major segments of the core course can be placed in context through structuring the material around perhaps four vignettes which sequentially develop the themes discussed above at increasing levels of depth. The sequence of vignettes could be, for example:

- *"Is There Really a Difference?"* – an exploration of how to tell whether two sets of numbers (data) are the same or different. This vignette introduces programming and statistical skills via basic quantitative and visualization techniques. From the beginning, it demonstrates and requires use of appropriate best practices for both individual and project-based computing.
- *"The Power and Peril of Pictures"* – a mix of visualization and inference methods, aimed at showing how hard it is to do a really good job of understanding data. This will involve

the students in analyzing and presenting imperfect, real-world data through their own code and reasoning, and studying ways that can be done well and go wrong.
- *"Whose Data is This Anyway?"* – exploration of privacy and anonymization in the context of data analysis. This includes meaty issues, such as how balance privacy guarantees with the ability to do high-quality inference. Genomic data might be a typical case.
- *"How Should I Act?"* – an introduction to decision-making, both at the individual and social level. Ethical issues can be explored. Some probability theory will be introduced, along with issues around data validity and reliability.

The connector courses break up the single common core cohort for more customized instruction without forcing premature choices on students to specialize. There might be one connector for students interested in a group of social sciences centered on large population data (economics, political economy, political science) and another centered on smaller cohorts (psychology, linguistics, cognitive science). Engineering fields may elect to use such connectors to further develop the understanding of numerical methods while applying them to varied domain-specific problems. Computer Science and Statistics may elect to further develop the algorithms and methods around the core. The set of connectors must be carefully coordinated so that choosing a specific one early would not close doors to future study in other areas; in some cases, a student might want to return to take another connector.

Eventually we foresee 15-20 connector options developed by units across campus to provide
- a range of disciplinary options through choice of data sources and questions, and
- a range of depth, perhaps through having H ("Honors") versions of certain connector courses for students who can and want to study in more detail.

This would result in connector classes of 100-150 students, serving a total 3,000 students each semester.

Together, these courses would take 6 units, to be compared to the 30 units/year nominal for freshmen. That is a bit more than the 4-unit courses now typically taken for the L&S Quantitative Reasoning requirement, which this option would satisfy. It is comparable, on the other hand, to the typical load (one or two 4-unit courses) for the L&S Reading & Composition requirement. Having students take it uniformly early would allow adaptations of other lower-division courses that could benefit from using it as a prerequisite.

CS and Statistics lower-division offerings have demonstrated that these general topics can be taught in large scale to the majority of Berkeley students. We believe that the large core / focused connector approach will improve the quality of the teaching, while at the same time allowing us to reach all of the entering class. We can also see that scaling requires paying attention to the full course ecosystem, including the availability and preparation of teaching assistants as well as instructors for the core and connector courses. For teaching assistants we can see experimenting with an approach that CS has piloted to serve its enrollment growth, namely, cultivating a vertical learning community that prepares previous students in a course to serve as undergraduate TAs when graduate students are not available in numbers sufficient to fill out the full staff. Rather than a technical fix, this approach should be approached as an

opportunity to expand the undergraduate experience to encompass community-building and engagement with other students seeking to learn.

In every aspect of course preparation, from core materials to connector design to the preparation of instructional staff, the foundational course has to communicate respect for the diversity of student experience. That respect goes beyond the simple variety of students' interests and possible majors. For this curriculum to be worthy of Berkeley's ambitions, it needs to address dynamics of inclusion around data science approaches and fields, which have historically been seen by some of our students as excluding them for reasons that include race and gender. Proven approaches to creating inclusive, diverse classrooms need to be built into the course materials and all its pedagogical encounters. Only with focused attention to equity and inclusion will this curriculum deliver its intended effect.

There are several other important issues that the course raises, and we preliminarily reflect on them in the appendix. We would particularly welcome discussions with faculty and advising staff around serving students coming in with different backgrounds, addressing issues of equity and inclusion head-on, making the foundational course content available for junior transfers, providing additional support options for students who need it, and integrating with students' programs of study and schedules.

That said, we feel that ultimately the questions raised by the foundational offering are best resolved through experience developing and deploying the new courses, rather than by opining about it now. The core will need to be developed through pilots on the scale of a few hundred initially and grown. This will give an opportunity to understand how it scales in various dimensions and how various constituencies receive it. Involving faculty broadly in the iterative refinement process over, say, two years will provide an opportunity for engagement that will be essential for developing the connectors. Departments will need time to evolve their course offerings and majors to take advantage of, and to accommodate, the course and the new level of student data competency. Temporary adapter courses are likely to exist for some period. Modern pedagogic practices that provide modularity and specialization through various use of on-line mechanisms can potentially be brought to bear. This might yield multiple threads through a common syllabus, some going into greater depth into certain topics, or it might yield multiple tracks, or possibly a network of sub-semester modules. The agile development methodology that is an important aspect of modern data analytics is well understood (outside academia) and can well be applied to the course itself.

Appendix 2 - The foundational course

5.2. “Follow-on” classes out in the departments

With the foundational offering in place, students should be able to build on it in a wide range of programs. These opportunities will take different forms for students with different intentions, and the possibilities here are diversely inflected across the campus.

We can see that this effort will be more straightforward in some domains than in others, governed in part by sequences and pathways through existing major curricula and by constraints such as accreditation requirements. For instance, in the life sciences, the desire to give students better access to both computing and statistical skills has been a long-standing topic of conversation. The introduction of Math 10 has begun addressing at least part of this question. The shape of Math 10, and the possible connections to more advanced courses in the varied biology programs across campus, will be a key question for DS curriculum design.

Throughout, the goal of “follow-on” courses is *enabling*: enabling students to go further, enabling departments and programs to add options. The possibilities can only be explored in thoughtful, coordinated conversation with faculty from many departments; our suggestions are meant to spur creative thinking rather than bound and define.

Students can expand competency in **working with data** in ways suited to their own majors and interests. We see this as unfolding at multiple levels:

- **Lower- and upper-division courses centering on methods of data analysis** can take the foundational course as a baseline. Questions of domain-specific methods can be tackled in a critical fashion with less review of basic statistics. Students coming in with foundational hands-on competency will be freed up to develop it into expertise relevant to their own course of study. Starting from students’ experience engaging computationally with real data (and the infrastructure built up to support it), these courses can also continue the strategy of project-based work.
- **Existing upper-division offerings** can add in new components (assignments, projects, course segments) that go deeper in working with data. Assuming adequate infrastructure is provided, we see major opportunities here for nearly all domain areas. Instructors will be able to assume the core concepts and practices, rather than re-cover them too quickly and too schematically as just a set of tools. This could potentially balance out certain methodological divides, in that it will be safer to assume that all students will have the capacity to work rigorously with multiple methods. These courses can also create domain-specific opportunities to address how instrumentation and data collection, classification, and organization shape analysis.
- **New upper-division offerings** can move our curricula further into engagement in the data sciences in their disciplinary domains. Some departments and colleges have begun investing in faculty appointments that cross over to data science domains – in joint appointments with Statistics, for instance, leading to cross-listed courses – and

encouraging faculty to offer courses on computational methods, working with geospatial data, or digital humanities. We can see that process accelerating in the context of the larger DS curriculum, making space for curricular innovation in many domains.

Students can also go deeper in **critical exploration of data** in a wide range of disciplines. We find some of these opportunities particularly inviting, as they leverage Berkeley's broad-spectrum excellence and intellectual ambition. The opportunities created by mixing students from different majors in these courses can have signal impact in broadening their undergraduate experience, addressing questions such as:

- epistemology and meta-reflections on reasoning with data
- data in the creative curriculum, arts, and new media
- reproducibility of data, data interoperability, data ownership and governance
- ethical issues of data collection and use; privacy issues with inference about humans
- uses of data and data-driven argumentation in social, organizational, and political context
- data in the context of methodological pluralism, asking good research questions and fitting methods to ends

Finally, the combination of early lower-division experience followed by in-depth methods classes will give a much larger number of students a strong preparation for **undergraduate research**. This can increase both the depth and the breadth of undergraduate research engagement. Because all students in a given major are being engaged, in many disciplines this curriculum will also make it easier to integrate research-based experiences into the major.

Making these changes requires infrastructure, incentives, and coordination. We discuss that below.

5.3. “Follow-on” classes in computer science, statistics, and possibly engineering and math

Creating data science foundational offerings will likely have consequences for curricula in Statistics and Computer Science. Beyond relieving enrollment pressure on particular lower division courses, they may end up being part of the ongoing renovation of curricula for majors. In Mathematics, there may be other implications as students seek out different kinds and combinations of follow-on competence (e.g., Math 54, linear algebra and differential equations, and Math 55, discrete mathematics). These effects should be thoughtfully tracked through student enrollment data and discussions with faculty, students, and advising staff.

In addition to connectors for the foundational course, it seems likely that Engineering and Applied Math may want to develop a different kind of follow-on to provide some of the computational science and numerical methods present in E 7 and Math 128A respectively. The computational introduction may allow these to be streamlined. Currently these are taught in a specific environment, Matlab, which is utilized by various follow-on courses and texts in the field. While students move much more easily between computational environments today than in decades past and environments are largely converging, SciPy and IPython being one such example, the coverage of numerical methods, convergence, and error propagation in the simulation of physical phenomena has unique aspects. At the same time, Monte Carlo techniques and other kinds of stochastic methods are of growing importance.

We expect as well that the creation of upper-division data science offerings (described in the next sections) may very well impinge on curricula for majors in Statistics, CS, Applied Math, IEOR, and possibly other domains. It would be wise to bring faculty from those departments into the conversation around data science upper-division planning.

5.4. DS advanced offerings, the core, the minor, and the major

We recognize further that data science has a coherent intellectual core that, while broad and evolving, centers on data as first-class concept, viewed broadly as a basis for analytical insight and inference, not merely a discrete object on which computation is performed. Data traverses the full range from observational concreteness to high-level scientific abstraction. It serves as a singular bridge between fields. It is powerful, for example, for students to reflect on data architecture, data models, and the ramifications of conceiving metadata as data, and to recognize that while it is created in one context with particular provenance, data can take on a life of its own with the potential to transit across multiple uses and settings. In particular, uncertainty pervades all settings in which data is used as partial evidence to infer underlying truths, to form predictions and to make decisions. The need to make honest assessments of uncertainty, to gather observations so as to reduce uncertainty, and to communicate uncertainty are part-and-parcel of the modern scientific method, particularly in the context of collaborative, cross-disciplinary initiatives. As data becomes a valued resource across all domains, it is essential to work with the potential legal, ethical, and privacy risks associated with its collection, categorization, and analysis – particularly in contexts involving personal, behavioral, and social data – and to think critically and rigorously about how it is made actionable in societal and organizational settings.

In all these ways, data science is emerging out of existing disciplines in response to the challenges posed by the unruly, large-scale forms of data now at our disposal. Yet it holds the promise of being far more than an opportunistic collection of practices. Data science is in the process of becoming, we believe, an intellectual formation with its own foundations and habits of mind.

The Rapid Action Team is persuaded that Berkeley should scope, shape, and define this science in ways that make it available to our undergraduate students. New career paths grounded in data science have emerged in industry, civil society, and academic research, and we wish to empower our students to pursue such careers. By making this effort, Berkeley can chart an intellectually coherent trajectory through a potentially dizzying set of transformations, ones that elsewhere are largely being tackled by patching together course offerings from whatever disciplines locally have seized the high ground. The emerging field of data science presents a challenge for institutions that have built undergraduate offerings within curricular divisions that have been in place for decades. We think Berkeley can claim a unique leadership role in advanced data science offerings no less than our broad-based foundational course.

We envision an upper-division program (classes supporting stand-alone course-taking, a minor, and a major) that develop distinctive core competencies, each of which links to existing academic communities. Our experience in the field leads us to identify five core competencies:

- Statistical foundations: algorithmic and mathematical foundations of inference and decision-making as a process, including sampling, optimization, simulation and design, together with procedures for representing, propagating, and controlling uncertainty
- Scalable computing: study of trade-offs and constraints involving temporal factors, complexity, storage, and transport in a technological and architectural setting for the processes of inference and decision-making with data of all sorts
- Knowledge representation and management: formulation of data models, indexing, schema, lineage, provenance, dynamical models, spatial models and causal representations, reproducibility, curation
- Utilization: organization, visualization, and communication of data and the outcomes of the inferential process for decision making in a context
- Critical thinking: social, ethical, personal, organizational, and institutional implications of data and of the inferential process, including the importance of context, origins, and impacts

We envision an integrated initial development of these competencies into a two-semester upper-division core sequence, with more advanced courses providing deeper, focused treatment. Course lists associated with each of the core DS competencies provide both the structure and specifics of the minor and major programs. The courses in these categories will be drawn both from existing upper-division courses and from new courses created expressly for the DS major. We anticipate that these courses will be developed over time by faculty in Computer Science, Statistics, I-School, IEOR and other units with an interest in data sciences education, including some courses deliberately constructed to provide grounding in more than one competency.

In many cases these will be existing courses that become adopted into the DS program, which may cause them to evolve. In other cases, new courses will be developed, to achieve either broader applicability or greater depth. For example, a new Statistical Machine Learning course might replace the current overlapping offerings, CS 189 and Stat 154; moreover, some of the material in the current instantiations of these courses can be drawn in the core course, allowing the machine learning course to focus on more advanced material that is closer to graduate curricula and to real-world practice.

A curriculum in an evolving field such as data science must be adaptive. It must also be an intellectual crossroads. Taking seriously the bridging nature of the concepts of data and inference, upper-division DS offerings need to be thoughtfully connected to the rest of our students' experiences, building a minor that integrates with other programs and a major that can be completed simultaneously with other programs for students who choose that route. Part of our task is to build structures of collegial relations and organizational forms that support connective, adaptive curriculum design.

The DS program will thus be defined and offered in cooperation with units across campus, and might include fields such as mathematics, genetics, econometrics, demography, cognitive science, biostatistics, earth and planetary sciences, epidemiology, digital humanities, materials

science and engineering, and so forth. There are many upper-division courses on campus where computational and inferential ideas are already present and which would be natural choices for incorporation or evolution. We believe that the lists can and should remain quite open-ended as more disciplines embrace data-oriented reasoning and units will increasingly elect to offer courses that can provide concentration as they incorporate these modes of research and teaching.

Finally, we expect that a range of upper-division courses in DS will be highly attractive to PhD students in multiple disciplines. That need must be accommodated in planning, though graduate training is not the focus of our report.

5.4.1 The upper-division core

Central to the systematic development of a data science education program is the creation of an **2-semester upper-division core sequence** (let's call it DS 100A/B) that anchors both the minor and the major program and can also serve students as a stand-alone offering. It extends the concept of integrating statistical and computational concepts, skills, and practices in the framework of working with real data. The course content meets student needs by unifying material currently spread among multiple courses. It will be necessary to identify what pieces get extracted from those courses, how they get adapted, and how they get put together to form a whole.[8]

A guiding principle for the design of upper-division DS core is to develop a solid understanding the "data sciences life cycle," consisting coarsely of research design → collection → preparation → analysis → utilization. In practice, this life cycle is not linear, but highly iterative, and it has several distinct dimensions. From a data management perspective, it involves acquisition, organization, storage, retrieval, and curation of data. From a data analytics perspective, in involves filtering, fusing, model fitting, model selection, prediction, causal inference, and diagnostics. From a data engineering perspective, it involves the tools and techniques for constructing robust data pipelines, parallelization, testing, performance, reproducibility, and transparency. From a decision-making perspective, it involves visualization, presentation, and interaction with stakeholders. While many today view data primarily from one of these perspectives, the life cycle proceeds along all these dimensions in an interrelated fashion. Moreover, each stage of the life cycle along any of these dimensions presents social, ethical, political, organizational, and privacy implications.

Understanding the life cycle of data places an emphasis on the "use of" data science techniques in the DS core. In DS 100A/B, students would become effective at *using* important data structures and inference procedures as embodied in modern data analytics toolsets and

[8] We see important elements in Stat 135 (concepts in statistics), Stat 133 (computing with data), CS 61A/B (programming and data structures), CS 70 (discrete math and probability), CS 169 (software engineering), CS 186 (databases), CS 194 (data visualization), Math 128A (numerical analysis), and other courses in these and additional departments.

technology, and this user's perspective will help inform their understanding of the underlying foundational principles. A key element of their practice is "real"; students will be exposed to the realities of real world data that is inherently "dirty." They will directly address issues of cleaning, selection, decoration, integration, and the many aspects of curation. They will grapple with integrating a meticulously collected source with other sources. They will learn to think about issues of bias, variance, confidence intervals, and diagnostics. They will experience the framing of data analysis, including the context of the acquisition and processing of data and communication of the results of the analysis.

Because the nature of data science is interdisciplinary and collaborative, the new course sequence would likely be taught in such a way as to teach students teamwork through group projects, give them experience with written reports and class presentations, and provide training in interpersonal, communication (written and spoken), and leadership skills. Ideally, such groups will bring together diverse perspectives. We believe that ethical issues and social context can and should be "baked into" DS 100A/B in the same way that they are integral to the foundational offering.

5.4.2. Minor

Students who want to pair an existing major course of study with additional depth and sophistication in data science should be able to by taking a minor that consists roughly of taking the DS upper-division core, additional competency-focused courses, and a DS application experience course in or related to their major area.

The DS minor will provide both a core of common material and a set of structured opportunities to pursue domain-specific questions aligned with the distinct areas in which data science can be engaged. Along with skills, tools, and concepts, the minor should give students opportunities to learn critical thinking capacities around asking good questions of data, understanding the limitations of what can be learned from the data on hand and what additional data may be needed to answer the question of interest, and working through the ethical consequences of acquiring and analyzing different sorts of data. Collaborative skills are important and should be built into the minor program. The mixing of students from different areas within the DS minor will expose students to a range of different perspectives on the role of data science and its many different modes of application.

Although there is much work to be done to settle the structure of a minor, we present a tentative plan here. The lower-division foundational course (DS10) described in Section 5.1 and a corresponding 2-unit lower-division cross-listed connector course (DS C10x) would be prerequisites for the minor. These would be followed by the upper-division foundational course (DS 100A/B.) Students would also take at least two courses covering at least two distinct competencies from the list above. Finally, students in the DS minor will be required to take an application course that integrates data sciences with applications in their major field. Ideally, this course will be developed and offered by the faculty in their major field or by clusters of faculty from allied departments engaged with similar kinds of data. It is intended to give students direct

access to the diverse data science issues that come up in the context of applied practice, including both technical issues (research design, methods, and tools) and social, practical, and normative ones (data collection standards, privacy, ethical challenges, data-driven policymaking).

5.4.3. Major

We believe that Berkeley should offer a major in data science. As we have discussed, new career paths in data science have opened to our students that call for knowledge that we can only awkwardly deliver to them now. These trajectories can lead them directly into careers in industry and commerce; they also provide compelling preparation for doctoral programs in a wide range of increasingly data-driven fields. Moreover, we believe that data science has a coherent intellectual core that is best captured in the context of an undergraduate major, and a depth that demands additional coursework to achieve grounded understanding and fluent praxis. While the focus in the minor is the home discipline, with data science as a supporting discipline, in the major the focus is reversed. The need for grounding in a real-world scientific or social domain remains; the focus cannot be entirely on computer science and statistics because it is in the context of emerging real-world data analysis problems that the need for an integrated approach is most apparent. Data science majors need to understand how data analysis ideas connect to underlying concepts in a domain, and how domain-specific context conditions their overall approach to problem solving.

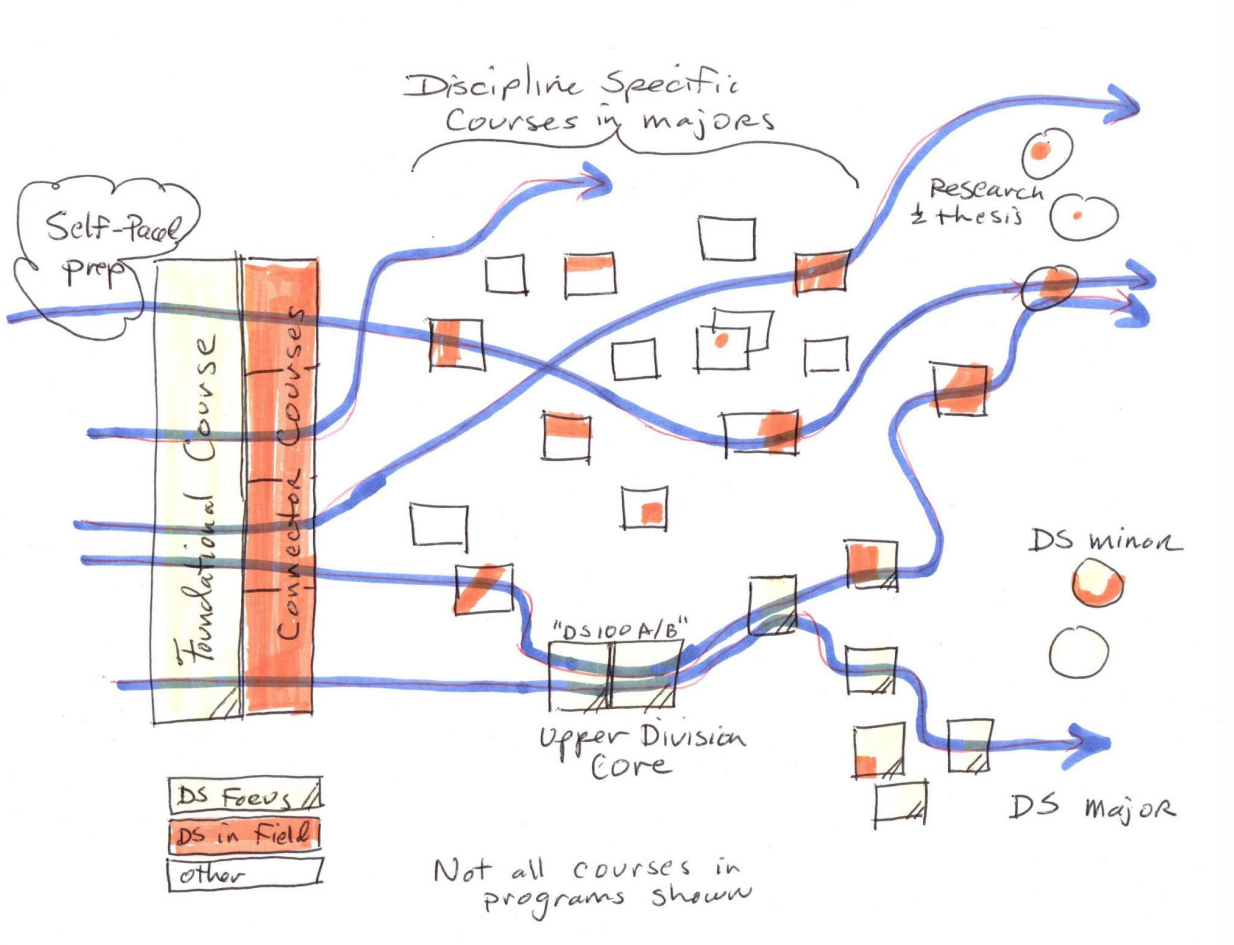
Although this report is not the place to lay out a definitive plan for the major, we believe that a broad outline can reasonably be supplied. The proposed DS major will build on the foundational course (DS 10) described in section 5.1. We expect that all prospective majors will take a corresponding 2-unit lower-division cross-listed connector course (DS C10x), ideally simultaneously. We also recommend that students anticipating a major in data science take one or more of the additional lower-division follow-on data science offerings that may be developed in a range of departments and/or in statistics or computer science, and we expect that there will be a need for additional requirements or prerequisites in computing, mathematics, and statistics.

All majors will be required to take the two-semester upper-division core course (DS 100A/B) described in section 5.4.1. This course, like the lower-division core, will blend computational and inferential thinking, but at a more advanced level, and establish a direct understanding of the data lifecycle. The course will assume more advanced programming and mathematical skills than can be assumed for lower division students.

Beyond these four courses (DS 10, DS C10x, DS 100A, DS 100B), we foresee the major consisting of upper-division offerings amounting to seven or eight courses . We envision this program as structured around the five areas of core competency identified above as central to data science (statistical foundations, scalable computing, knowledge representation, utilization, and critical thinking). In a sketch, a valid program would include, for each area of competency, at least one course with primary coverage of that area. In addition, a valid program would have a concentration consisting of an area that is covered by at least three courses. If there proves to

be significant interest in building data-science-related courses in other majors across campus, a concentration might also consist in part of an application area to a distinct problem domain.

6. Proposed curriculum structure



All students encounter Data Sciences early, through the combined foundational offering. Examples, from the top down:

- Some students take a straightforward path, building on the foundational course in one or more discipline-related courses, and use the information in research or capstone experiences.
- Some might take a more complex path, picking one connector early but later moving through classes into another area.
- There are multiple paths through the connector and lower-division classes into different majors.
- Students may choose to add a data sciences minor to their discipline-specific major. That minor may include some classes that mix discipline-specific and DS core content.
- Finally, some students will proceed to the data sciences major. That path goes through mostly DS course content, with some discipline-specific content in the original connector and perhaps some mixed courses to broaden the application examples.

7. What would this change?

What would this change for students?

- Currently some of our students never get an exposure to this area, and never develop any of the basic skills. This proposal changes that situation, because it provides students with basic capabilities and understandings for later life.
- By building this into the lower-division (ideally first-year) curriculum, students will rapidly understand what it means to study this domain, and whether they really want to. That, combined with a well-articulated breadth of offerings from further study in the context of individual fields through the minor and major, and some well-focused advising, helps them plan their careers at Berkeley much more intentionally.
- On the downside, this can be overwhelming: Another piece in a full program, when the student just wants to study their field! We're talking about integrating data sciences into many of those fields, and more will probably take it up with time, but students don't necessarily see this. So we need to work on communicating with students and with helping them become more thoughtful about their educations. (This is true for every initiative, though, and it's something that the Chancellor's Undergraduate Initiative must deal with.)

What would this change for faculty?

- For uninterested ones, not much. Many faculty will see little change in their courses or research.
- Many faculty will see small changes: Students who want to do research with them will have a little more capability in this area. Students may frame research questions for class papers and projects differently, etc.
- Some faculty will choose to make significant, desirable changes: They'll be able to raise the level of their courses that overlap with this area. They can take advantage of skills and ideas taught in the foundational course, spending less time on introductory material in their course and therefore getting to the interesting part earlier. There will be resources (development grants, faculty Chairs) available to make this possible, so it's not just another unfunded mandate; that's definitely a change for the better.
- A few faculty will want to make significant changes in the introductory math, statistics, and computer science classes they teach. Math 10, which includes a statistics component, might shift to be best taken alongside the foundational course, or perhaps use it as a prerequisite, for example.
- Some of the people currently advanced classes teaching in math, statistics and CS may see smaller classes (which is good) of students who are more focused on the specific statistics/CS content (which is very good).

What would this change for departments, Divisions, and Colleges?

- There will be some interest in revisiting lower- and upper-division curricula as new courses based on data sciences capabilities are created. This kind of change can be disconcerting for faculty.
- More resources (faculty positions, staff, etc) will be available for efforts in this direction. Departments will have to be careful to make sure that existing people get an opportunity to do exciting things in this area, not just giving them to the new people.
- Departments teaching large amounts of lower-division introductory topics in this area will have a tectonic (slow, but with occasional earthquakes) adaption as we develop and teach the foundational course, minor, and major, and students start to vote with their feet.
- Eventually, a new Data Sciences department or interdisciplinary unit may emerge. It's impossible to predict how that will go. If it were to be created, it would require significant work and commitment.

8. How could we implement this?

There are two primary things needed to make this happen:

- **Resources** – accounting for people with time available, classrooms, IT infrastructure, staffing, expense funds, etc.
- **Roles & responsibilities** – an effective way to organize the effort, keep campus-wide engagement and the benefits that come with it, get to shared understandings or standards of what courses like this should provide.

As the plan is refined, it will become increasingly concrete. In the interest of soliciting reactions, we offer the following skeleton.

8.1. Resources

The goal of the **foundational course** is to serve 6,000 students per year. We have existence proofs from Statistics and Computer Science that this can be done via very large classes for about 50% of those students. We're proposing to take over parts of that model and adapt other parts for the curriculum we're envisioning.

- A small cadre of faculty and lecturers are needed to deliver the core course.
- At scale, about 15-20 connector courses would have to be staffed.
- GSIs and UGSIs would be needed in suitable numbers, which will require careful consideration of both funding and student cadre size.
- Course development and training time is needed for the entire teaching group, specifically including the GSIs and UGSIs.

Just as important, infrastructure is needed to make this a quality Berkeley common experience.

- Classroom space for the core and connector classes, both large sections and project/lab space. The redevelopment of Moffitt undergraduate library provides a unique opportunity to address at least some of this in a central place that's already strongly associated with undergraduates.
- IT infrastructure for teaching and project work (which may take the form of hardware, virtual environments, staff support, and other resources). This is potentially a significant investment, given the lack of campus-wide infrastructure currently in place.

Developing **follow-on courses in departments** requires attention to capacity and incentives. This needs to be done via several approaches to work for both existing and new faculty. It's very important that there be a balance of extra support and incentives for existing faculty with the possibility of new faculty. We cannot achieve our goals on the backs of our current people, nor can we only give the opportunities in this area to new people.

- Course development grants would provide existing faculty time and assistance to adapt their existing courses. As the Presidential Chairs program has demonstrated, a few of these per year, over many years, would enable adoption across numerous departments.

- A group of Faculty Chairs, along the model of the Hewlett Chairs, would allow faculty to make a five-year commitment to larger projects, such as development of multiple new courses or restructuring majors.
- For departments where new research directions align with the undergraduate educational direction, fractional or possibly full faculty FTE positions should be made available. If a department wants to assign 0.5 FTE for a new faculty position that links research and teaching in this area, the other 0.5 FTE could come from a centrally-held pool. Strong commitments to teaching would be necessary, of course.

Additional infrastructure may also be needed for this component, along with that needed for the foundational course. The infrastructure may need to be regionalized and include support for access to datasets as well as classic IT infrastructure.

In order to make definite statements about the **minor and major**, we will need to develop the proposed curriculum more fully and carefully work through how it draws from and interacts with teaching in existing courses and majors. The process will be interdisciplinary, very iterative, and driven in large part by the faculty developing the courses that make up the balance of the minor and major programs. We expect that faculty strength will need to be added. An adequate level of staffing and advising will need to be tuned as we gather signals of student interest.

Across the initiative we may need to build in infrastructure and resources for students to access course material in non-central locations, extra-curricular tutoring and support systems (currently crucial for students taking computer science and statistics), training for TAs and faculty, and other elements that support student learning. Appointing new faculty requires attending to all the ordinary components of recruitment, including start-up packages and space.

8.2. Roles and responsibilities

We are describing an educational initiative that is developed, delivered, and engaged across the entire campus. We discuss below how to get started, but eventually we must develop a structure to govern that effectively.

The foundational course needs coordination of the (small) team that develops and delivers the core component, and the (many) instructors across campus who develop and deliver the numerous connector courses. This is in some ways similar to the American Cultures courses, and we suggest a similar approach: A committee of faculty from across campus with expertise who handle coordination and evaluation.

The development of courses across the departments, including courses that will make up the eventual minor and major, will go through the usual departmental and Senate processes. A small amount of coordination is needed, and can be ensured either through augmentation of the committee for the foundational course, or (more likely, due to workload) through a parallel committee. Note that faculty FTE allocation includes research priorities, start-up funding, and other issues that far exceed our charge, which extends to the undergraduate education aspects

of Data Sciences. Those must eventually be coordinated at a higher level. We do suggest, though, that the Administration pay consistent attention to the organizational issues involved in the development of the Data Sciences minor and major.

9. Next steps

Here we are reporting on work to date. Our effort is not complete. We intend to continue solicit feedback from faculty and students, gather additional data for analysis, and do additional outreach over the next months.

There will need to be a campus-level decision on whether to pursue this initiative. That will be tied to other decisions around research directions, funding and fundraising priorities, and other considerations beyond the scope of our charge.

In the meantime, we believe that there's enough faculty and interest already in this area to start development of the foundational offering as the key next step.

We propose that a pilot version of foundational course be developed along with perhaps two to three initial connector classes from different parts of campus. The diverse connectors will help ensure that the core course sticks to its mission of being common across campus, while at the same time testing what it takes to get student and faculty involvement. Having several classes of 100-200 students each semester may also offload some of the still-growing pressure on the lower-division CS and statistics classes. Once the content and scaling issues have been settled, ramping up to the entire entering class requires creating more connector classes to have sufficient range, and more instructional resources to offer enough sections. Providing resources that allow a (small group of) department(s) to create and test-run a connector course would greatly help, as would resources to help with initial deployment, but eventually these large-enrollment classes should be funded like any other departmental responsibility.

Ideally, such a pilot could be delivered for the first time in Fall 2015, be repeated in an second-round version in Spring 2016 while additional connectors are developed, and then start to ramp up toward its final size in 2016-17.

Time is of the essence if we are to have a pilot foundational course offered in the Fall. We would need to organize and resource the effort promptly, so that the Senate's Committee on Courses of Instruction can consider and approve the course this coming Spring, and incoming freshman students can select it during orientation in June 2015. Course schedules for Fall 2015 have already been set, but the demand for introductory statistics and CS is so high that we don't expect any difficulty filling a Fall pilot.

There is a lot of faculty interest in developing the curriculum for the upper-division data science core, major, and minor, specifically through adaptation of existing courses and development of new ones. Because building those is a long process, it's important to start. The immediate action needed is to create a group of interested faculty, coordinated with but broader than our task force, to concretely work on these. Initial resources for this are small, just a bit of support

and encouragement. By next academic year (2015-16), it will be important to have a few course development grants and some release time available to make real progress in this area.

10. Conclusion

At the Chancellor and Provost's request, we have examined the status and prospects for data sciences undergraduate education at Berkeley. We have talked to faculty across the campus, examined existing courses and student programs, and considered Berkeley's environment.

We believe that now is an opportune time for Berkeley to increase its data science education for undergraduates, specifically through developing and deploying a balanced program:

- A foundational lower-division course accessible to everyone, with broad applicability
- A suite of courses that advance the use of data sciences within the broad swath of disciplines available to Berkeley undergraduates
- A high-quality minor for students who wish to couple data sciences with their major discipline
- A data sciences major for those students who want this to be their primary area of study

Together, this will better prepare the full range of Berkeley graduates to engage with this growing area during the rest of their studies, through their career, and in their entire life. Although our remit has been limited to undergraduate education, we believe that an initiative in this area would also enhance other work in research and graduate education across the campus. Because all professionals – not just those in the academy – are grappling with the proliferation of data and the need to make sense of it, we believe this initiative will provide ample opportunities to build bridges to outside organizations and individuals across many sectors and domains. We expect that those connections will benefit Berkeley in manifold ways.

We have laid out concrete next steps that can be taken to develop this direction through initial development of courses. We encourage the campus to embark on those, because we'll learn a lot and we're sure there will be broad interest in the courses. In some sense, those next small steps are inevitable.

Beyond that, however, there is a strategic decision needed: Shall we give this effort the priority and raise the resources needed to do it right? Making that decision is the most important next step. We recommend that the campus choose to do so.

Appendices

Appendix 1 - Current student experience

To provide a sense of scale of student commitment, in 2014-15 nearly 5,000 undergraduates are expected to take some form of an introductory computing course. For roughly half of those students, their choice will be the extremely demanding course intended for majors, CS 61A. At the same time, over 3,500 students are expected to take an introductory statistics course inside or outside the Statistics department (Stat 2 or Stat 20). Moreover, according to DARS data, large numbers of our undergraduates take statistics off-campus, for instance, at community college. Presently one-third of Berkeley undergraduates taking an introductory course in the Statistics department enroll in Stat 20, the offering that serves future majors in statistics (and other fields with significant demand for statistical reasoning). As indication of the scale of change, in 2007-08 nearly 1,600 students took an introductory computing course, one-third of them the course for majors, and just under 3,000 students took introductory statistics, fewer than one-fifth of them in Stat 20. These enrollments should be understood against a backdrop of roughly 7,500 bachelors degrees awarded each year.

Computing

A distinct, large cohort is the College of Engineering, which has a computing requirement across the college. Here we see a different schism. Outside EECS, only Industrial Engineering and Operations Research (IEOR) and Bioengineering accept a CS course as fulfilling the computing requirement. The remaining majors in the College of Engineering require E 7, which provides an introduction to computing in the context of numerical methods (although students may petition to substitute CS 61A). E 7 serves approximately 800 students a year, declining in recent years below 700. Only IEOR recognizes Statistics as contributing to its major requirements.

During the past five years a gentler CS introduction, CS 10 (The Beauty and Joy of Computing), has been available, which now serves 650 students per year, about a third of whom go on to take CS 61A. Other majors have substantially grown or recently begun introductory computing offerings as part of their upper division. In Statistics, Stat 133 (Concepts in Computing with Data), which historically served about 150 students a year, increased to over 250 three years ago and may be twice that size this year. In Mathematics, Math 128A (Numerical Analysis) grew steadily from under 200 to nearly 350 and may exceed 500 this year. Other small offerings collectively serve another 50 or so students.

Statistics

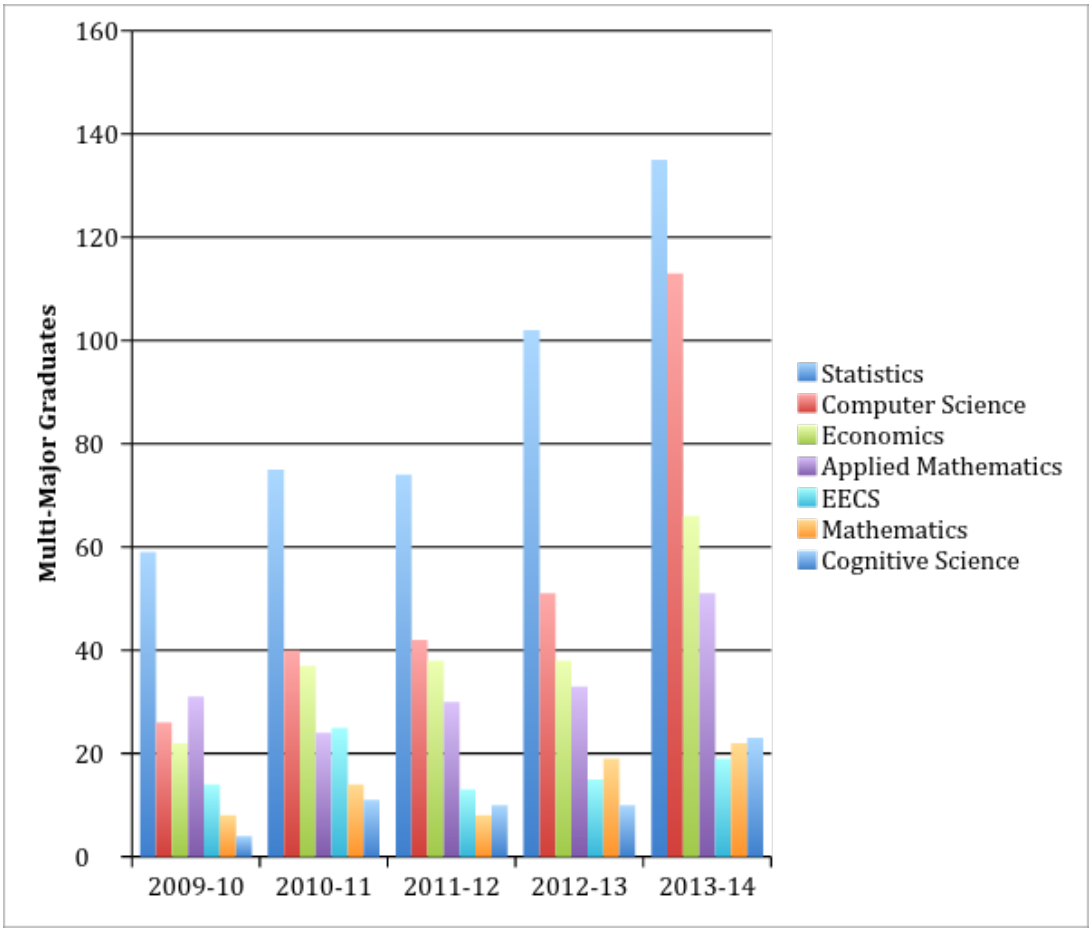
Introductory statistics has been a stable collection of major and service courses, with Stat 20 intended first of all for future Statistics majors, Stat 2 for non-majors, and Stat 21 for Business majors. Collectively, enrollments have been relatively uniform around 2,800 students per year;

however, there has been a substantial shift to Stat 2 (historically less that 20% to now nearly 40% of total enrollments). Other departments offer introductory statistics as well. The most significant of these is Public Health 142, historically below 300 annual enrollments, but recently grown to nearly 500 (serving graduate students as well as undergraduates). Also, the new Math 10A/B sequence includes some statistics and is now serving 300 students per year.

Patterns of majors

Students have responded to the mismatch between the desire to gain computational and statistical skills and the various major requirements in two interesting ways. First, we see a huge growth in students who after taking these introductory offerings go on to declare the major. As indicated above, over the past 5 years, the number of statistics majors has grown from 80 to 400; L&S CS majors from 140 to nearly 700; and EECS majors from 900 to over 1200, with a shift in balance from roughly equal in EE and CS to 3/4ths CS.

Secondly, we see a huge growth in double and triple majors, from 50 such students graduating in 2008 to over 250 in 2014, with Statistics and Computer Science being the two most popular. In 2014, 135 of the 234 Statistics graduates were double majors and 113 of 402 CS graduates were as well. Economics, Applied Math, and EECS follow these at substantially lower levels. Economics, Applied Math, Business Administration, and Math are the most popular combination with Statistics, in order of popularity, while Applied Math, Cognitive Science, Math, Economics, and Business Administration combine with Computer Science. Surely, these patterns in both single and double majors are also a reflection of macroeconomics factors, job prospects, and family pressure.



Current openings for data science in the undergraduate curriculum

Statistics would be a natural choice for current students interested in Data Science because it includes an “applied cluster” of three courses which could be drawn from Computer Science or several other areas, only one lower division course, and a flexible upper division.

Computer Science is not particularly well structured to accommodate data science interests, as it requires a demanding lower-division program of five courses and, although flexible, a collection of specialized upper-division courses. As noted earlier, an analysis carried out within Computer Science concluded that the material most important to a data science program is contained in small sections of several of the current courses. To get these with the current courses involves essentially completing the major, and many students do just that.

The third natural point of Data Science concentration would be the I-School, but it currently does not offer an undergraduate program. It does offer an on-line professional masters in information and data science. As this was created as an on-line degree, a set of courses were created around it in the on-line format. The College of Engineering offers an MEng through EECS with a concentration in Data Science and Systems, which has largely been assembled out of pre-existing course offerings, plus a capstone project and leadership courses. Statistics offers an

MA program in Statistics generally, which does permit a Computer Science course to taken as an elective and includes a capstone with “a substantial data analysis project.”

Graduate implications

Shortcomings at the undergraduate level are reflected among the current graduate student population. Several programs, including astronomy, materials science, neuroscience, nuclear engineering, and the social sciences, hold boot camps and workshops to provide students with the computational and inferential skills needed to carry out their research. The massive uptake of training offerings since the 2013 opening of D-Lab extends significantly beyond the social sciences to every school, college, and division on campus. In some cases, these interstitial offerings essentially cover undergraduate material. In other cases, they demand practical skills and tools that are not taught in departmental curricula in Computer Science or Statistics with a heavy theoretical orientation, whose students rely on self-teaching or informal learning to pick up many practical skills. However, it is not material from any particular course, but important snippets that would appear in a number of different courses. This recognized shortcoming in the backgrounds of our graduate students reflects the lack of adequate undergraduate data science offerings globally, not just at Berkeley.

Appendix 2 - The foundational course

One obvious question is whether the core is a single massive course or a suite of variations on the theme. The sheer scale requires multiple offerings regardless. The diversity of backgrounds demands that students will need to acquire the material differently. Obviously, the task force discussed the many alternatives and trade-offs at great length. In such an integrated course, a student may excel in certain aspects and be at sea in others. For example, a student with a great deal of programming experience may find the statistical concepts daunting and the artistic aspects of presentation incomprehensible, or vice versa. However, we come down for designing the foundational offerings around shared material with adequate local adaptations through the connector strategy. Beyond the commitment to statistical and computing competence in the course's instructors and the strong desire not to "track" students into presumed majors in their freshman year, we see real payoffs in offering a common student experience that will be distinctive to Berkeley to boot.

Experience shows that not all Berkeley students take quantitative courses on campus, and that some struggle with the ones that they do take. We intend that there be core and connector courses at a level that all entering students can access. Particularly once the foundational course has ramped up to the level of reaching all Berkeley students, it will be necessary to have connector classes that meet all parts of the entering cohort at an appropriate level. There may be, say, a need for a 3-unit connector that provides additional time and experience for students with limited high-school preparation, or self-paced preparatory material.

A small number of students might arrive at Berkeley with both the breadth and the depth of knowledge that it would make sense to skip this foundational course. There won't be many, because the conceptual content of the course far exceeds the Computer Science and Statistics AP classes as they are normally taught in high school. For the few that do have that, we will provide an option to demonstrate that and move past this foundational course. Many more are likely to almost have that, in the sense of some grounding in computing and/or statistics; those are best served by taking the foundational course with a more advanced connector. Creating those connectors should be a priority.

Our principles address all students, including those who transfer to Berkeley. It's likely that aspects of this foundational course will make their way into community college curricula, but that's a slow process, and many Berkeley students now come through other routes. We must assume that transfer students will want and need access to this material. Therefore, we propose that Summer Sessions provide a version of the core plus a small number of connectors during the summer before transfer students' first semester at Berkeley. Growing numbers of transfer students are arriving early and could take it, particularly since 6 units qualifies the student for financial aid. Second, propose that some connectors be created for use with the Fall core course that take advantage of the characteristics of the more developed and focused transfer population.