Berkeley's Undergraduate Data Science Curriculum: Year 1 Pedagogical Overview

Cathryn Carson, History
David Culler, Electrical Engineering and Computer Sciences
Bob Jacobsen, Dean of Undergraduate Studies

University of California, Berkeley
November 2016

Table of Contents

1. Executive summary

Berkeley's Data Science education program was created to make it possible for every undergraduate at Berkeley to engage capably and critically with data, and building on this foundation to provide pathways to further progress in many fields of study. The Data Science program aims at a comprehensive curriculum built from the entry level upward to meet students' needs for data fluency, to serve a broad range of programs of study that now or in the future will integrate data science capacities, and to provide the depth of understanding required for an eventual major. In the program's first year (2015-16, 12 new courses) over 550 students participated; in its second year (2016-17) over 1000 are anticipated to take part.

The curriculum has been launched with strong entry-level offerings that provide conceptual understanding and hands-on experience with modern computationally-based statistical techniques applied to real-world data. These courses are constructed in a modular fashion that provides the benefits of a common platform together with ways to customize offerings to particular areas and needs. The starting point is a lower-division class on Foundations of Data Science (Data 8) that is linked with a growing family of connector courses that tailor the program to a range of domains.

While rigorous and responsive to the latest intellectual developments, the introductory courses make data science accessible to any Berkeley student, tapping into their diverse interests in nuanced questions that can be answered with analyses on data and relying only on university-entry mathematics requirements and no prior computing experience. The intellectual content of the Foundations class, its conception of integrating computational, inferential, and critical thinking, and the overall program architecture of Foundations plus connectors have been described as profoundly transformative by Berkeley's peers.

Students in these classes are diverse, enter the program with a broad range of previous experience with computing and statistics, and are engaged in a wide set of majors (more than 50 to date). Just over 50% of students taking the Foundations course in Fall 2016 are female, and underrepresented students enroll at a level comparable to the campus at large. Social and ethical implications of data science are consciously addressed and are proving to be engaging to students. Student appreciation is high for both Foundations of Data Science and connectors, an incredibly active student community is supporting the program, and the desire for follow-on courses is strong.

There has been broad, continuing outreach to faculty in many departments to assess how a modularly constructed Data Science curriculum can integrate with their students' needs, both at the entry level and in follow-on formats. Faculty will offer three new advanced Data Science courses in Spring 2017 (Data 100 [CS/Stat C100], Stat 28, and Stat 140), as well as incorporating data science into a growing number of courses and programs of study. Teams of faculty are now forming to develop proposals for major and minor programs of study in Data Science, which can be expected to be available to students once they are approved by the Academic Senate.

This report provides a pedagogically oriented overview of the program to date. It gives context for the Berkeley Data Science curriculum in terms of student needs, experiences, and course-taking patterns, so these can be integrated into the university's ongoing planning. It addresses key lessons drawn from the first year of experience in the Data Science education program, notably:

- The pedagogical success of the entry-level offerings (Data 8 and connectors)
- The value of a broad-ranging, modular, and still integrated program
- The significant effort required make the program connect across the campus
- The extraordinary opportunity of designing for diversity and inclusion in student populations, interests, and support mechanisms.

Taking this first-year start-up as establishing proof of concept for a significant part of the student body, the curriculum can be refined and extended to larger populations of students, its effects can be followed into their ensuing course-taking, and its integration with programs of study across campus can be moved forward, assuming adequate institutional structure and support are provided.

2. Background

Berkeley faculty across many disciplines have collaboratively created a model for a comprehensive undergraduate data science curriculum. Starting from the blueprint in the January 2015 report of the Data Sciences Education Rapid Action Team (DSERAT), the curriculum is built around a modular core-and-connections structure that can serve as a platform on which many academic programs can build.

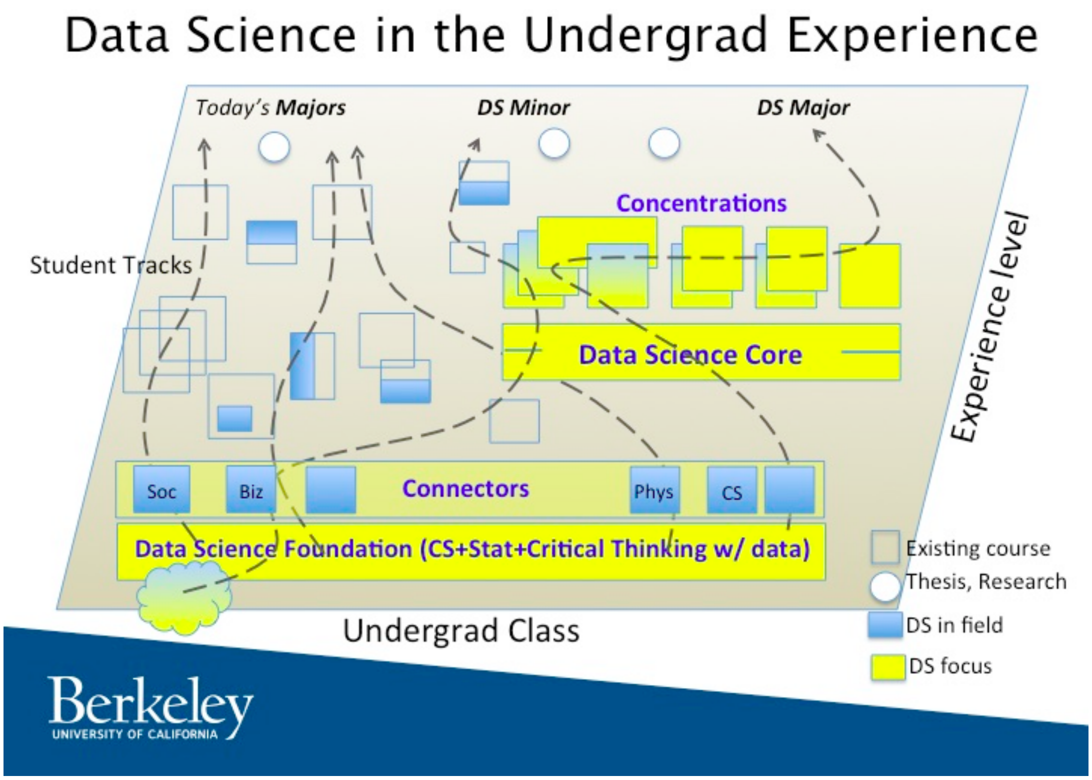
The Data Science curriculum was launched at the entry level in 2015-16 with an innovative introductory course and a suite of connector courses that relate to students' areas of interest, now ranging from neuroscience to civil engineering to demography to ethics. The entry-level courses are designed to provide the base for later classes in a broad range of departments that will be able to leverage and extend what students have learned. The upper tiers of the program are now being developed and will provide additional depth and connect across the campus with major and integrated minor offerings. As previewed in the DSERAT report, the program engages with societal and ethical issues around data science not only in course content, but also throughout the program design, incorporating best practices around diversity, equity, and inclusion so that the curriculum is welcoming to students of many backgrounds and interests.

The curriculum that is now being created aims at a comprehensively integrated program. It responds to the experience of faculty of the transformation of their own fields of research and teaching by the cross-cutting possibilities of data science, and to fast-growing student demand for courses in computing, inference, and hands-on work with real data, as reflected in very large numbers of students enrolling in preexisting courses covering parts of this material in separated fashion. The curriculum aims to integrate a full appreciation of the lifecycle of working with data with the computational and mathematical knowledge that underlies it. It follows a modular design that allows it both to leverage common teaching of exceptional quality and shared infrastructure in a highly cost-effective manner, and to create tailored offerings designed and "owned" by departments. In staying strongly coupled to student interests and diverse programs' needs, it must operate flexibly and responsively even as it scales up fast.

Through its start-up phase the data science curriculum has been very lightly staffed (~1.0 staff FTE through 2015-16 across several units, providing both technical infrastructure and programmatic support) and temporarily shepherded by the L&S Dean of Undergraduate Studies. In addition to the dean's resources and a start-up allocation of programmatic, TAS, and capital renovation funding, it has drawn heavily on individual faculty investment, staff commitment, and provision of additional resources by multiple departments and support units. Part of the program has been driven ahead by strong collaboration between EECS and Statistics; the other, more distributed part has been managed collectively, with many additional participants involved.

Program leadership has been provided by the Dean of Undergraduate Studies and two other faculty members involved in the leadership of the Data Science initiative. With the expiration of the DSPI in summer 2016, the program continues to operate on an provisional footing with interim faculty leadership, with TAS and programmatic resources allocated or carried forward from 2015-16, and with gifts (in dollars and in kind, amounting so far to roughly half the scale of campus investment) secured by program leadership and development staff in multiple units. The program's integration into departmental TAS processes is starting. A proposal for its institutional regularization in relation to a new decanal division for computation and data is now being considered by campus.

Curriculum schematic



An overview is available at http://data.berkeley.edu/data-science-education-program.

3. Year 1, 2015-16 (with preview of Fall 2016)

3.1. Foundations of Data Science (Data 8)

Pedagogical approach

Berkeley's data science education program starts at the introductory level, with a 4-unit foundational course, Foundations of Data Science, familiarly Data 8 (CS C8 / Info C8 / Statistics C8) that teaches core computational and inferential concepts while enabling students to work constructively with real data. The course was developed in spring and summer 2015 by a coalition of faculty across multiple disciplines and taught in two offerings in 2015-16 (a pilot in fall 2015, followed by the first regular offering in spring 2016). It has so far been collaboratively taught by Distinguished Teaching Award recipient Professor Ani Adhikari (Statistics) and Professor John DeNero (EECS), recipient of the Diane McEntyre Award for Excellence in Teaching. Other faculty have expressed interest in teaching it in the future.

The Foundations course is built on three interrelated perspectives: inferential thinking, computational thinking, and critical engagement with questions of real-world relevance. As Prof. Adhikari describes the intent of the course, "All students should have access to a course that develops data literacy, so that they can use modern data analysis as an approach to any problem or investigation that they encounter in any discipline." At the same time, Data 8 students develop a strong conceptual understanding of the mathematical structures underlying statistical thinking and learn ways of thought to work effectively with a modern programming language (Python) with modern data analysis frameworks, starting on the first day of class and continuing through each homework, project, and lab. In addition to teaching critical computing concepts, programming skills, and statistical inference, the course is based on hands-on analysis of a variety of real-world datasets, including economic and spatial data, and it delves into social and legal issues surrounding data analysis, including issues of privacy and data ownership.

From a data analysis perspective, students understand:

- Visualization for understanding and communication (graphs, histograms, bars, scatters, maps)
- Distributions and random sampling (with and without replacement)
- Properties of several statistics (median, mean, max, total variation distance)
- Testing statistical hypotheses
- Estimation, prediction, and assessing predictions and models
- Regression and correlation
- Clustering and classification

- Comparison, causality, and decisions

In the process students gain a solid understanding of **classical statistical concepts**:

- Probability theory, e.g., complements and multiplication rule, birthday surprise, permutations
- Distributions of data (categorical and numerical) and of probabilities
- Empirical distributions
- Law of averages, Central Limit Theorem
- Sampling variability, standard errors of estimates
- *p*-values and error probabilities in tests of hypotheses
- Bootstrap, permutation tests, null hypothesis
- Bayes' Rule and the probability of false positives

Much of this they discover computationally and then codify symbolically. In that process, they master **computational concepts**:

- Data types and data structures (tuples, lists, arrays, tables)
- Representation, operators, interpretation
- Sequencing, conditionals, iteration, comprehensions
- Use and definition of functional abstractions
- Data parallel programming techniques
- Higher-order functions
- Database operations (select, filter, join)
- Repetition, convergences, searching, sorting
- Testing, debugging, exceptions
- Objects and modules

Data 8 pedagogy is centered on a single powerful data structure that is as natural to use as a spreadsheet, but allowing students to grow from simple manipulation and visualization through to sophisticated statistical techniques used widely today in industry and research. Rather than a traditional exposure to programming that is focused on learning syntax, dealing with files and tools, and working idealized problems, students learn how to construct sound analysis processes in a computational document, starting from acquiring data and proceeding through a series of steps in a modern programming language, to arrive at a meaningful observation.

The class syllabus is available at data8.org; the online textbook authored by the instructors is available at inferentialthinking.com.

In Data 8 there is a natural back-and-forth of the mathematical concepts, the computational processes, and the experience in applying them to data. For example, students sample from the null hypothesis repeatedly through computational simulations to form and assess distributions, rather than relying on asymptotic properties. (All of the statistical operators are available as primitives in the computing environment, along with numerical operators, database operators, and visualization methods.)

The manner in which Data 8 introduces students to modern computational techniques and statistical techniques in a completely integrated fashion has been called "revolutionary" by our peers.[1] It amounts to an intellectually rigorous consolidation of the growing integration of statistics and computing in data science. The course deliberately fuses these fields and brings them down to their shared foundations. This pedagogical strategy lets Data 8 deliver deep understanding of fundamentals in a way that connects naturally with a broad range of applications, cementing conceptual understanding through direct experience with data analysis in a computational setting. Colleagues at leading universities (e.g., Harvard and Yale) have sought out our help in adapting their own teaching of data science to the Berkeley paradigm.

Pedagogical construction

As a computer science offering, the Foundations course differs from many others. "This course is really focused on how we study the world through the lens of computing," Prof. DeNero observes. "Even students with no background in computer science are able to do this. They go through the whole experience themselves—working with the data, asking and answering interesting questions."

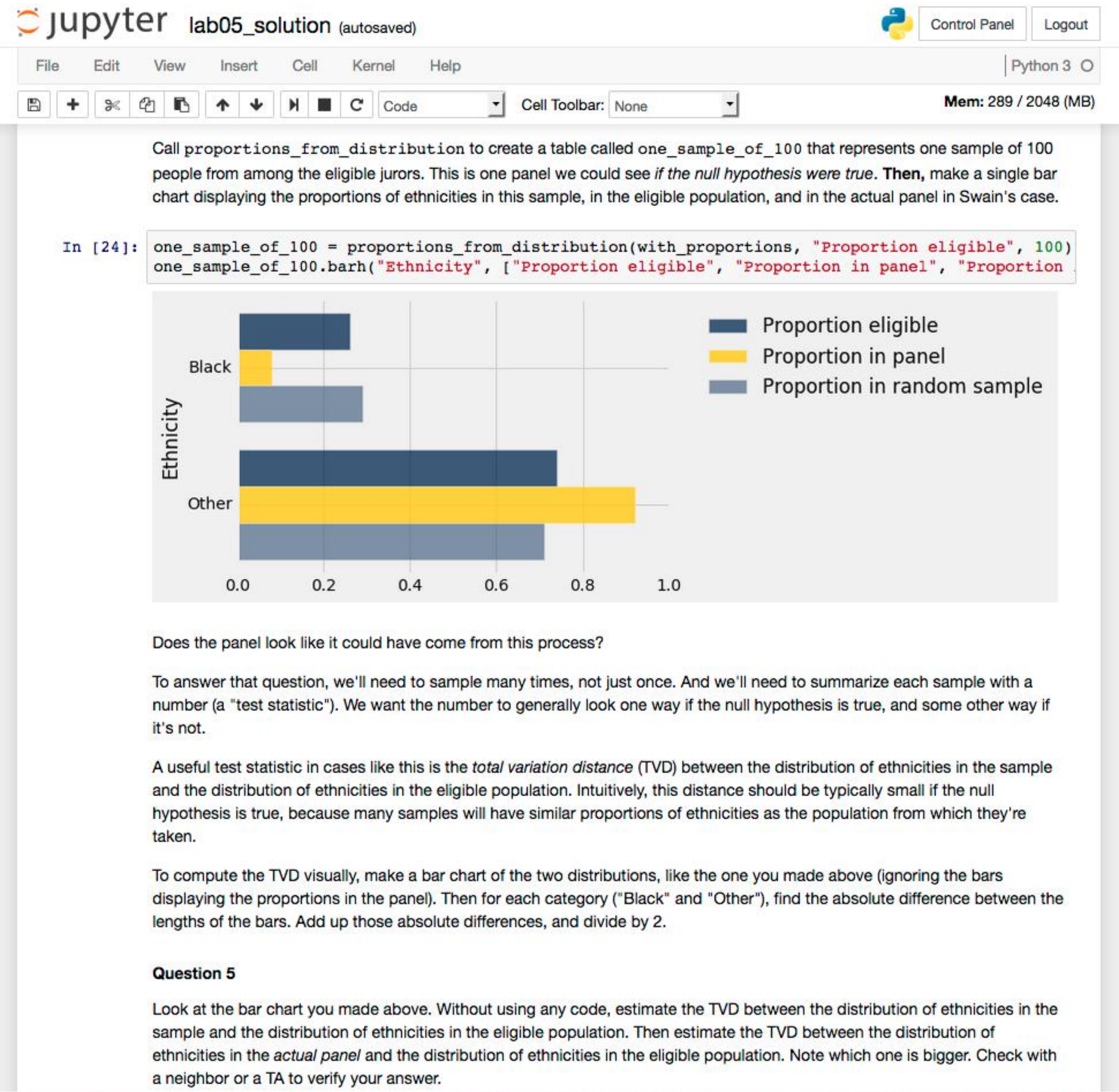
The Foundations course is lab-driven and project-driven to enable students with little or no programming background to become proficient with data analysis. Each Data 8 student enrolls in a lab section of 30 students, meeting weekly for 2 hours. Students undertake 3 two-week projects:

- Water use in California: mapping the water districts and overlaying IRS taxable income data by zip code
- Murder rates and the death penalty: nonparametric inference; the importance of visualization
- Classification of song lyrics: hip-hop or country?

Students learn to program and work with data in Jupyter notebooks, which enable browser-based computation in the cloud. The platform allows students to develop hands-on experience and intuition within a computational document in step-by-step, narrative form (and avoid installing software locally on their computers). The cloud platform supports computing examples in lecture, lab activities, homework and project assignments, and the students' own explorations. Because it is integral to the course experience, the course platform needs to be robust, stable, and easy to use, ideally more than it is now.

Fig. 1. Example of a Jupyter notebook (lab, Week 5, Fall 2016)

[1] Report of the External Review Committee of the Berkeley Department of Statistics, November 2015, p. 3.



Call `proportions_from_distribution` to create a table called `one_sample_of_100` that represents one sample of 100 people from among the eligible jurors. This is one panel we could see *if the null hypothesis were true*. **Then,** make a single bar chart displaying the proportions of ethnicities in this sample, in the eligible population, and in the actual panel in Swain's case.

```
In [24]: one_sample_of_100 = proportions_from_distribution(with_proportions, "Proportion eligible", 100)
         one_sample_of_100.barh("Ethnicity", ["Proportion eligible", "Proportion in panel", "Proportion
```

Does the panel look like it could have come from this process?

To answer that question, we'll need to sample many times, not just once. And we'll need to summarize each sample with a number (a "test statistic"). We want the number to generally look one way if the null hypothesis is true, and some other way if it's not.

A useful test statistic in cases like this is the *total variation distance* (TVD) between the distribution of ethnicities in the sample and the distribution of ethnicities in the eligible population. Intuitively, this distance should be typically small if the null hypothesis is true, because many samples will have similar proportions of ethnicities as the population from which they're taken.

To compute the TVD visually, make a bar chart of the two distributions, like the one you made above (ignoring the bars displaying the proportions in the panel). Then for each category ("Black" and "Other"), find the absolute difference between the lengths of the bars. Add up those absolute differences, and divide by 2.

Question 5

Look at the bar chart you made above. Without using any code, estimate the TVD between the distribution of ethnicities in the sample and the distribution of ethnicities in the eligible population. Then estimate the TVD between the distribution of ethnicities in the *actual panel* and the distribution of ethnicities in the eligible population. Note which one is bigger. Check with a neighbor or a TA to verify your answer.

Course construction

Enrollments (census data, Cal Answers)

Semester	Course offering	Enrollment
Fall 2015	CS 94 / Stat 94 pilot	109
Spring 2016	CS C8 / Info C8 / Stat C8	447
Fall 2016 (mid-sem)	CS C8 / Info C8 / Stat C8	509

After the Fall 2015 pilot, Data 8 expanded to the largest available classroom in its first regular offering in Spring 2016. In Fall 2016, the Foundations course reached its shortterm capacity limit (~500 seats/semester). It will be offered at the same scale in Spring 2017. The large pilot size and unusually fast growth for a new course match expectations derived from enrollments for introductory computing and statistics courses.

Prerequisites and requirements. Data 8 is designed for freshmen and sophomores of all intended majors. As it develops conceptual understanding from the ground up, it has no prerequisites beyond high-school algebra. It satisfies several requirements, including the statistics requirement in the overwhelming majority of the 20+ majors requiring statistics (see full list at data.berkeley.edu/requirements). In some cases (including Economics and the undergraduate Business major), programs have chosen to require the combination of Data 8 plus Statistics 88 (the Statistics connector course). Twenty majors have modified their programs to allow their students to take advantage of these courses in fulfilling existing requirements; several others (such as Engineering Mathematics and Statistics) are considering it for their requirements, having moved to bring it onto the program list of electives. Data 8 has been approved by the Letters & Science Executive Committee to satisfy the L&S Quantitative Reasoning requirement.

Course sequences building on Data 8. Data 8 will be able to serve as the main entry point (prerequisite) for other courses, including those in the planned Data Science major and minor programs. Design and proposal of courses for those programs is now timely, starting with the proposed Data 100 class envisioned in the DSERAT report. Other courses can also be developed with Data 8 as prerequisite, such as Stat 28 and Stat 140. (More details on Data 100, Stat 28, and Stat 140 are given later in this report.)

Relation to other course offerings

Data 8 provides a new entry point into lower-division statistics, alongside Stat 2, Stat 20, Stat 21, and courses in other departments. The Statistics Department has strongly recommended Data 8 to all majors that have relied on Stat 2, noting that for many students Data 8 is a better option. The department has likewise encouraged Data 8 plus Stat 88 as an alternative to Stat 20 for majors that require statistics based on calculus. For orientation, combined enrollment in Stat 2, Stat 20, and Stat 21 has been largely stable in recent years at roughly 2,000 students annually. (Significant numbers of Berkeley students choose to take introductory statistics at community college; these students are not included in campus enrollment counts.) Data 8 at its first-year level is roughly half the scale of Berkeley's other introductory statistics offerings combined.

Data 8 likewise provides a new, large-capacity point of entry for introductory computing, alongside CS 10 (Beauty and Joy of Computing), CS 61A (Structure and Interpretation of Computer Programs, the prerequisite for more advanced CS classes and required for CS and EECS majors), as well as courses in other programs, notably E7 (Introduction to Computer Programming for Scientists and Engineers). Among introductory CS offerings, Data 8 is most directly targeted at providing key capacities for other majors. For orientation, CS 61A has experienced a steep growth curve, leading to a Fall 2016 enrollment of 1,553 students in a single course offering. While the hypothesis has been that Data 8 will attract significant student interest that is more aligned with data science than with computer science per se, and while Data 8 enrollment has in its first year grown to a significant fraction of CS 61A enrollment already, it will be for the next years to show what its long-term effects on overall CS demand at Berkeley will be.

Effective use of resources. Foundations of Data Science serves students more efficiently than Berkeley's previous separate course offerings. The course is built on shared pedagogy, combining the most broadly needed content from distinct classes in Statistics and CS into 4 units of a single course. It conveys key skills in computing and statistics in an outward-looking way designed to be useful to students from a wide range of majors. When Data 8 is combined with connector course options for customization, the program offers a common platform for entry-level teaching that gives the benefit of shared delivery without being one-size-fits-all. Data 8 is a highly efficient way to provide high-value instruction in both statistics and computing at low cost per student credit hour (SCH).

Student demographics and experience

Gender and URM status (Fall 2016)

- **Gender parity is conspicuous**. Just over half (50.2%) of Fall 2016 Data 8 students identify as female.
- **Underrepresented students are significant**. 21.8% of students consider themselves members of underrepresented minorities at Berkeley; 11.9% answer "I don't know."
- **The demographics of Data 8 students are nationally distinctive.** A diverse Berkeley pipeline offers significant promise for shaping the profession of data science.

Year of studies

- **Freshmen and sophomores** are strongly represented at 53.8%.
- **A roughly equal fraction are more advanced students.** A number of graduate students are taking Data 8. (Some faculty are auditing it as well.)
- **Multiple pathways bring students to Data 8**. 18% of students have done one or more years of undergraduate study outside of Berkeley (e.g., community college).

Programming and statistics background

- **Students without previous programming are effectively served**. More than 50% start Data 8 with limited programming (self-described as "none," "terrible," or "bad").
- **Students who have previously taken statistics** find that the Data 8 approach is significantly different and not repetitive, at the same time leading to significantly greater comprehension (as judged by instructors on the basis of course performance).
- **Students with a wide range of previous preparation in statistics and computing still find new material to learn**. Instructors provide guidance and additional exercises

for students who come in with more background.

Majors

- **Students come from a broad range of undergraduate units**, including
    - College of Letters & Science: Divisions of Arts & Humanities, Biological Sciences, Mathematical & Physical Sciences, Social Sciences, Undergraduate Studies (all 5 divisions)
    - College of Engineering, College of Natural Resources, College of Environmental Design, College of Chemistry
    - Haas School of Business, School of Public Health, School of Social Welfare
- **56 majors and intended majors** are represented in fall 2016.
- **The largest majors represented** have existing computing or statistics requirements (computer science, economics, psychology, statistics, business administration, and cognitive science). Other majors among the top 20 include public health, molecular and cell biology, environmental economics and policy, political science, mathematics, and media studies.
- **More than 80 combinations of majors** (double and triple majors) are included, pointing to the multidisciplinary interests of Data 8 students.

The class serves a broad population and does not track students into particular areas of study. Instead it provides a common foundation on which other programs can build, with customization provided by departmentally-designed connectors.

In addition to students in technical majors, Prof. DeNero observes, “We’ve had strong engagement from students in literature, history, ethnic studies, areas that aren’t traditionally seen as related to Data Science—since today, studying any of these fields will also involve computing with data.” Prof. Adhikari adds, “Students from non-CS and Stats majors have skills that are very important—they ask different questions of the data. What I learned is that our students had an ability to generalize, and they were able to ask the broader questions in a way that they don’t in a regular introductory stats class.”

Student learning is significant. Students’ answers to conceptual and analytical questions have impressed instructors and observers in class Q&A, in lab settings, and on exams. Early reports from instructors of subsequent courses (e.g., Stat 134, Concepts of Probability) suggest that Data 8 students can perform at a high level in classes requiring mastery of statistical knowledge.

Integration with other programs of study will take additional thought and attention as Data 8 scales. Statistics is being taught in a new way in Data 8, and instructors of follow-on classes in other programs of study should be engaged around their expectations of student preparation and learning. In some cases integration is simple, as in the Department of Economics, which has determined that Data 8 plus Stat 88 meets its needs. In other cases, as more students from different majors take Data 8, more pedagogical dovetailing will be required, as in programs of study that draw heavily on familiar statistical methodology in their “methods” courses (for instance, in the social, behavioral, and environmental sciences; Psychology and Public Health are two current examples). In addition to discussion among instructors, mechanisms such as connectors and “translation” processes may be helpful. It will also be important to work through modes of integration and sequencing with those programs of study that draw on computing in their requirements, as for statistical computing, simulation, modeling, etc.

Support systems: Students are provided with a strong support network beyond the core staff of faculty, GSIs, and Undergraduate GSIs, including access to lab assistants, supplemental office hours, tutoring by members of student groups, and, as of Fall 2016, a new Data Scholars program for students from underrepresented groups (see below).

Student community: The peer community around Data 8 extends into supporting the next offerings of the course. The passage from Data 8 student to tutor to lab assistant to UGSI is strongly mentored by course instructors and is modeled on the pipeline approach used in EECS to scale large CS classes (mostly 8-hour-a-week appointments to fit with students’ demanding programs). There is substantial engagement by previous students in developing the Data 8 technical infrastructure.

Students have made a video about their experience:
https://www.youtube.com/watch?v=D5W7Zu15WjA
Student interest in Data 8 seems to be as much a viral phenomenon (word of mouth and social media) as the outcome of official circulation through formal university channels.

Student survey responses (Spring 2016)

- **85%** of students said they were happy or very happy about their decision to take the course (4 or 5 on a 5-point scale, instructor survey)
- **77%** of students said they learned a lot in terms of skills and ideas by taking the course (4 or 5 on a 5-point scale, instructor survey)
- **84%** of students formally enrolled in the CS offering of the class, when asked how worthwhile this course was compared to others they’d taken, rated the course either a 6 or 7 on a 7-point scale (Eta Kappa Nu student survey)

Some student reactions

- “One of the things I most enjoy about data science is the diversity—my classmates range from English majors to bio majors to computer science majors —all looking at data from our different perspectives.”
- “This class puts theory into practice. I was able to use data to tell powerful visual stories about the struggles I experienced growing up in southeast LA.”
- “Out of all the classes I’ve taken, this class gave me the most practical knowledge. I’m applying it in my internship at Google already.”

3.2. Connector courses

Connector courses are a key part of the entry-level program, giving substance to the idea that data science is anchored in ground-level engagement with many domains. In the modular architecture of the program, connectors give departments full latitude to decide how best to leverage Data 8: what to fill out, how to supplement, and where to add in alternate perspectives and needs.

Pedagogically, connectors offer a ready way to start bringing data science approaches directly into the experience of students in many majors, setting them up to use these skills later in their programs (and ideally advancing that capacity in the faculty skill set as well). Connectors have been offered so far by instructors in a wide range of fields of study: from statistics to history to cognitive science, from demography to civil engineering to literature, from ecology to computer science to social networks, from ethics to genomics to geospatial analysis. (A full list of connectors is in the appendix.)

The program architecture leverages student learning and infrastructure from Data 8. This approach lets connector students and instructors begin immediately with questions and data drawn from their own fields. Connectors are designed to be taken at the same time as, or after, the Foundations course. A philosophy of symmetry governs the relationship, as students in the connectors bring their experiences into Data 8 class meetings, and as case studies, datasets, and questions that are surfaced in the connectors feed back into Data 8 course design.

Connectors form a spectrum from application of Data 8 techniques in varied fields of study to expanding technical depth in aspects of the course. Devising the right way to offer a connector targeting students in a particular program of study involves thoughtful integration with the rest of that concentration. Here the leadership is intrinsically in the hands of the faculty in the area of concentration, with support from the data science program.

An exciting development is that as faculty take specialized areas of study down to the freshman level, the effort reveals commonalities with faculty in other programs; so some connectors are now being provided jointly or in rotation. In each department a connector may reveal other topics that would be well served by the connector course instrument. Thus the fabric of connector faculty relationships seems to weave in both dimensions.

The pedagogical design of an entry-level curriculum with integrated connectors is seen outside of Berkeley as no less innovative than the Foundations class itself. Harvard's Derek Bok Center for Teaching and Learning devoted a full-day symposium to the connector course

[2] Report of the External Review Committee of the Berkeley Department of Statistics, November 2015, p. 3.

model in Spring 2015. The “combined package,” in the words of another group of peers, “will soon become a model for the rest of the world.”[2]

Course offerings

Roughly 50% of Data 8 students have chosen to take a connector so far. An overview of enrollments is given in the table below; more details are given in an appendix.

Semester	Connectors		Enrollment	
	# of courses	# new	# of students	% of Data 8
Fall 2015	6	6	59	54%
Spring 2016	11	6	217	49%
Fall 2016 (mid-sem)	10	5	277	54%

Course construction

- Connectors are mostly numbered 88, though this is not a required designation.
- Multiple departments have had connector courses approved by COCI. An “incubator” function is provided by L&S 88 for first-time offerings.
- Connectors (e.g., Statistics 88) can be made part of a set of options or a required sequence in one or more programs of study.
- Connectors can offer a more focused or smaller learning setting for students. In the initial stage, some pilot connectors have been small as the student pool grows and instructors gain experience. It should be anticipated that some connectors will remain small to medium size, while others will need to become quite large.
- Connectors can have prerequisites (e.g., calculus) as appropriate.
- So far, connectors have been offered as 2-unit courses. Observation suggests that connectors of 3 (or 4) units may also be important (see below).

Connector offerings

- Connectors can be coordinated with the Data 8 syllabus in a variety of ways, as makes sense for different fields of study.
- Connector instructors typically draw on assistance from previous Data 8 students in designing exercises and supporting lab instruction.
- Connector instructors have been drawn so far from ladder faculty, visitors, lecturers, and postdoctoral fellows.
- Connector instructors have access to the standardized computing environment, lab space, and student support available to the Foundations class.
- The program offers several modes of support to pilot a connector. Some seed funding is available to stand up a course before it is regularized as part of regular departmental

curriculum planning and TAS budgeting.

Course experience

Many students are excited about connectors, some of them exceptionally so, making strong statements that they find it the most inspiring part of the program.[3] Some students come back for multiple classes in a broad range of subjects. They see applications and connections as an essential element in a data science curriculum and connector classes as a way to explore new areas. Combined with the high level of enrollment, qualitative feedback has been a significant confirmation as the program has gotten off the ground.

There are open questions about how to appropriately shape connectors a) for this student audience, b) in connection with the Foundations course. The present cohort of connector instructors have been working their way through pedagogical questions about course content and approach, including:

- creating course goals and exercises appropriate for entry-level students
- providing assistance with programming challenges
- deciding how to align with material covered in the Foundations course
- managing domain-specific customization of Data 8's pedagogical approach

Unit value of connectors is a key open question. While there continues to be a lot of enthusiasm for 2-unit connectors, some student and faculty input suggests that they can be over-packed with content and take preparation time out of proportion to their unit value. Some departments also report that 2-unit courses fit awkwardly into faculty teaching expectations or do not integrate with breadth requirements. For some programs of study, it will be valuable to try out 3-unit (or possibly 4-unit) connectors (in the form of new classes or redesign of current classes) and see how they can be integrated into student pathways.

Faculty engagement is critical. Because the Foundations course is a novel way to approach teaching data analysis, many connector instructors look for additional preparation to get up to speed. Some sit in on Data 8 or take up self-study of Data 8 materials. As described in the appendix, the faculty short course on data science pedagogy and practice, which offered a 30-hour program of instruction and lab work in early summer 2016, was broadly welcomed and highly valued by participants.

Coordination of connectors takes significant work. Even with standardization on a common platform, support for connector offerings takes a lot of coordination. Compared to the rest of the entry-level data science curriculum, it is considerably more staff-

[3] Because classes are sponsored by multiple departments, it is not simple to collect and standardize course evaluations. The assessments given here come from qualitative interviews and surveys done by the program and most connector instructors in Spring 2016.

intensive, faculty-intensive, and resource-intensive and requires more institutional support than it now has.

Comments from students:

- “I wasn't exactly sure what to expect going in, but I was surprised by the depth of research we examined in this field (digital humanities) and the knowledge of the instructor.”
- “I did not expect to be working on data at nearly the same level that the original researchers would examine their data, which was very exciting. This connector pushed me to learn how to code in a very real sense.”
- “This class turned out to be more technical than I had expected, but it is also something I invite and am pleased with because it's nice to apply statistical methods to real world examples.”
- “I am looking into data science as a full major and potentially minoring in CS. I am definitely taking more cs courses now because of this class.”
- “I can now more comfortably handle big data while working in my environmental economic and GIS research. This is quintessential to the research I plan on doing in the future.”
- “This course went beyond my expectations. I learned things in this course that I would never have gotten the chance to cover elsewhere. The course was fascinating and raised new ideas for me that influenced other areas of my work as well.”

Comments from connector instructors

- “My favorite moments came from the students. I had a mix of about 50% students majoring in Math/Stats/CS and 50% in environmental science/anthropology/geology or other domain area; and I enjoyed seeing each side realize the role of the other.” — Carl Boettiger, Professor, Environmental Science, Policy, and Management, instructor for Data Sciences in Ecology and the Environment in Spring 2016.
- “The primary value added of the connectors lies in the training to communicate and share new, complex ideas in a parsimonious way with a heterogeneous audience. The first-year college experience benefits a lot from an exposure to data science. It benefits even more when that is combined with a seminar experience in which students learn to use tools in plain language to explore a problem.” — Ryan Edwards, Visiting Associate Professor, instructor for Health, Human Behavior and Data in Spring 2016.
- “It has been great engaging Berkeley undergrads and encouraging them to think about ethics, values, and what is fundamentally *human* about data. Pushing students to broaden their ideas around data, where it comes from, and how it makes a difference—sometimes good, sometimes bad—in the world is key to cultivating an ethical and humane future for data science.” — Anna Hoffman, School of Information, instructor for Data and Ethics in Spring and Fall 2016.

4. Additional aspects

4.1. Pedagogical infrastructure and resources

The Jupyter notebook environment is critical to the entry-level data science curriculum. Generous assistance from the Berkeley-based Jupyter team (partially housed in BIDS) has been essential to the program. In moving quickly to scale, the JupyterHub infrastructure that supports the curriculum has attracted significant interest outside of Berkeley and has immediately pressed up against the limits of the technology now available.

The Tables abstraction (http://data8.org/datascience/tables.html) developed specifically for Data 8 provides a simple, pedagogically centered “dataframe” abstraction that allows students to transition easily from spreadsheets to a full programming environment. It integrates fully with the Jupyter environment and provides a stepping stone to complex dataframe environments, such as R and Pandas.

The cloud infrastructure that supports the program is being actively developed by a cooperative team of students, staff, faculty, and volunteers across multiple units, in the open source community, and among industry collaborators. At this pilot stage, industry contributions equivalent to several hundred thousand dollars have been essential to allowing the program to grow.

Approaches used in the data science instructional labs may be adapted for computation-intensive teaching in other areas of campus. The program’s cloud-based infrastructure allows instructional labs to be built without desktop computers and associated overhead. To provide computing capacity for students without laptops, a semester-long laptop loan program has been piloted with the Library with generous donations from supporters.

Inexpensive renovations in three instructional lab spaces in Summer 2016 have provided full-time space for Data 8 and connector labs and office hours at the current level of enrollment. Layout is 30 seats each, clustered into groups around shared table space. In addition to dedicated instructional labs, students have created spaces for collaborative work by making extensive use of overflow space in BIDS, the Library Data Lab, and D-Lab. If other space can be provided for office hours, each instructional lab can support 375 students enrolled in Data 8 and a proportional number in connectors spaced over the week. This space allocation will need to be increased no later than 2017-18 for the curriculum to grow.

4.2. Short course for faculty on Data Science pedagogy

Faculty have asked for support learning Data 8 material and incorporating it into their teaching and practice. Beyond encouraging them to audit Data 8 (as several have each semester), the Data Science program has experimented with several mechanisms to satisfy faculty requests .

In June 2016, 35 faculty and instructors from a broad range of disciplines devoted a week of summer to a 30-hour course on Pedagogy and Practice of Data Science. (An additional 35 registered but could not be served at this time due to capacity and scheduling constraints.) Participants included faculty from a range of departments across campus, including American Cultures, Demography, Economics, Haas School of Business, History, Linguistics, Math, Near Eastern Studies, Neuroscience, Optometry, Physics, Political Science, Rhetoric, and Sociology.

The faculty short course, co-taught by Foundations of Data Science instructors Adhikari and DeNero, covered the key teaching methodologies for Berkeley's data science education program and its new way of thinking statistically, and gave participants hands-on experience programming in Python using the Jupyter notebook environment. The course also offered panel discussions with connector faculty and a group of Data 8 and connector students. Participant satisfaction in a follow-up survey was high. The material in the short course will be offered to faculty again as soon as planning and teaching capacity allows.

Selected quotes from survey respondents:

- "I loved the integration of the statistical concepts and the computing; seeing really is believing, and I think this will motivate students to study statistics in a more rigorous way as well (especially those who are more reluctant to engage the math)."
- "I definitely want to work on developing a 'module' for my Ethnic Studies courses to both animate the existing course material from another dimension, and also to bring into our students' lives a taste of other methodological techniques which can complement our field of study."
- "Seeing how you guys 'un-AP' the students was great. I found the way in which the early lectures simultaneously motivated fundamental basic stats concepts and learning to write code illuminating. As a relatively new member of the Berkeley community, I found the interactions with instructors from *all* over campus incredibly beneficial."
- "I learned a lot both from the class itself and the other faculty in the room. We as a campus should do more of this. How many other initiatives are out there that we could all build from?"

The short course has made it possible for additional faculty to take up data science in their teaching. Several participants in the short course have signed on to offer connector courses in 2016-17.

The notebook approach to teaching data science has gotten uptake, including among instructors for whom data science is new. Along with bringing the approach into connectors, faculty have partnered with students to develop notebook-based modules for existing courses from entry level to more advanced. As an example, a Rhetoric R1A instructor in Fall 2016 developed a three-class module on polling data in the context of the current election. The instructor commented, "Quite seriously, today was great. ... this is a true immersion experience. The python notebooks looked spectacular, and I think 80% of the class will be talking about this at their dorms tonight." She continued, "It's been a bit of trailblazing adventure here and there, but … the students started buzzing when they recognized their own neighborhood data in the tables. That was exciting, and satisfying, for all of us." Her final conclusion: "I'm stunned by how pedagogically sound the overall approach is."

4.3. Data Scholars program

To serve underrepresented students in data science, the Data Science program partnered with student organizers and D-Lab (the Social Sciences Data Laboratory) in Fall 2016 to initiate a Data Scholars program, building on the models of the Biology Scholars and CS Scholars programs. Data Scholars is a student community that provides a welcoming environment and support resources to underrepresented students taking Foundations of Data Science. The program offers drop-in office hours and collaboration opportunities to students completing Data 8 homeworks and labs.

Diversity and inclusion are core values of the Data Science program. Data Scholars provides a set of deliberate mechanisms to realize those values. Professor Ani Adhikari, who has provides faculty guidance, observes that diversity is a critical element of the course culture and especially of its pipeline of student staff. "Their presence will matter in the future," she says. "Who gets to decide what data will be studied, and what kinds of questions are asked of the data? This group has to be as broad as possible, to bring in a range of perspectives." In the view of the lead student organizer, Luis Macias, "The whole goal of this course is to make data science as accessible as possible to students of many majors and backgrounds. Data Scholars is an extension of that."

4.4. Student engagement

A vigorous and engaged student community has rapidly grown up around the Data Science education program, with a strong social media presence and a partial home base in BIDS. Undergraduate student teams support key parts of the program's work (curriculum support, front- and back-end infrastructure, outreach, diversity and inclusion, mapping, and research collaboration), and student staff in Data 8 and the connectors are actively engaged in assisting instructors and teaching. Four independent student groups offer tutoring and follow-on opportunities for data science projects and careers.

5. Pedagogical reflections and lessons learned

Foundations of Data Science (Data 8)

- **Data 8 is intellectually coherent, broadly accessible, highly valued by students, and demographically sound.** It has rapidly attracted a large entry-level audience, while advanced students draw value as well. Early indications are it succeeds in teaching important material in an innovative way.
- **Students being served by Data 8 reach across the spectrum of majors.** They are concentrated so far in particular majors, particularly in programs with existing Statistics and CS requirements. It will be important to reach out to a broader range of programs, on both ends of the spectrum of quantitative preparation.
- **As enrollment expands, it may be necessary to offer additional sections of Data 8.** The best split may be by previous programming rather than area of study.
- **The course technical platform is at the cutting edge.** It should be stabilized and packaged for export.

Connectors

- **Roughly half of Data 8 students take connectors,** often reporting significant enthusiasm for the connector concept and offerings.
- **Connectors can be successfully implemented in a broad range of fields**, with different expectations on student preparation. Uptake in the sciences and engineering will depend on prerequisites and integration with or reworking of existing programs of study. Uptake in fields with “methods” courses (as in the social sciences) will require coordination with departmental offerings.
- **Connectors may increasingly need to go above 2 units.**
- **Engaging ladder faculty as connector instructors is important**, particularly for propagating the program into follow-on courses in application domains.
- **Connector design takes significant preparation and effort.** The Data Science program will need to continue investing in mechanisms like course development support, student assistants, and the faculty short course.

Program architecture

- **The design of the entry-level program around Data 8 is efficient and modular.** The combination of large lectures and smaller labs with supplemental support scales effectively, drawing on a pipeline of undergraduate course staff.
- **Coordination of connectors and integration with existing majors is challenging, resource-intensive, and a significant amount of work.** The connector concept and overall program architecture as envisioned cannot succeed without that additional effort.

- **The construct of a shared platform plus customization** through integration with other programs of study is a key architectural element for the design of the program and an eventual data science major and minors.

Student experience and curricular implications

- **Students value the breadth of the data science program,** including the broad-ranging examples and the spectrum of majors in the student population.
- **Diversity and inclusion demand dedicated attention** to make all students feel welcome. Along with course culture, practices, and forms of support, this involves looking out for underrepresented populations and transfer students. Programs such as Summer Bridge will be valuable to build.
- **Students want to see the program continue into the upper division.** Large enrollments in subsequent courses can be assumed. If a major and minors are offered, they will likely be of significant size (several hundred students per year).
- **Movement of students into data science courses will affect other programs** in ways yet to be foreseen. Along with CS and Statistics enrollments, there will likely be effects on other quantitative, data-analytic majors.
- **Many students will take data science to satisfy major or college requirements** (statistics, quantitative reasoning). Where those requirements exist, imposing new ones may not be in students' interest. Where they do not exist but arguably should, data science should be considered a primary option.
- **Judging by enrollment patterns, many students have essentially imposed a computing requirement on themselves**. Data science is a good way to meet it for many of them.
- **Computing provisioning for students needs careful thought.** This is an institutional gap at Berkeley at present. Students often bring their own computing resources, but the unprovisioned minority is a real concern. The cloud provides scalable computing, but its costs are in cycles and people (administrative resources and development) rather than hardware. Its cost is not significantly less expensive than the comparable budget for textbooks.

6. Next steps

6.1. New courses in preparation for Spring 2017

New connectors are being developed by faculty in several departments (e.g., in Legal Studies, MCB, Mathematics, and Geography). The data science program is assisting faculty with developing materials for integration into existing courses (e.g., in Near Eastern Studies) and is providing assistance with other new courses (e.g., in computational offerings within the Digital Humanities program).

3 new courses are being offered in Spring 2017 that take Data 8 as prerequisite:

Principles and Techniques of Data Science (CS C100 / Stat C100): Data 100 is an intermediate-level gateway to upper-division courses (e.g., within the anticipated Data Science major and minor programs) that will cover both the foundations and the applications of data science. It has been collaboratively designed by a team of faculty to follow Data 8 and approved by COCI. Students explore the data science lifecycle, including question formulation, data collection and cleaning, exploratory data analysis and visualization, statistical inference and prediction, and decision-making. This class focuses on quantitative critical thinking and key principles and techniques needed to carry out this cycle. These include languages for transforming, querying and analyzing data; algorithms for machine learning methods including regression, classification and clustering; principles behind creating informative data visualizations; statistical concepts of measurement error and prediction; and techniques for scalable data processing. Beyond Data 8, prerequisites are additional depth in computation (programming and abstractions) and mathematics (linear algebra and calculus).

Statistical Methods for Data Science (Stat 28): Stat 28 is a new lower-division course approved by COCI to serve students in many disciplines who have taken Data 8 and want to learn more advanced techniques without the additional mathematics called on in upper-division statistics. Topics include group comparisons and ANOVA, standard parametric statistical models, multivariate data visualization, multiple linear regression and classification, classification and regression trees and random forests. An important focus of the course is on statistical computing and reproducible statistical analysis. Students are introduced to the widely used statistical language R and obtain hands-on experience in implementing a range of commonly used statistical methods on real-world datasets. Data 8 is the only prerequisite.

Probability for Data Science (Stat 140): This new course introduces students to probability theory using both mathematics and computation, the two main tools of the subject. It has been approved by COCI. The contents have been selected to be useful for data science, and include discrete and continuous families of distributions, bounds and approximations, dependence, conditioning, Bayes methods, random permutations, convergence, Markov chains and reversibility, maximum likelihood, and least squares prediction. Labs will cover a variety of topics including matches in random sampling, distance between distributions, Page rank, and Markov Chain Monte Carlo methods. The prerequisites are Data 8 and one year of calculus. Data 8 gives students a practical understanding of randomness and sampling variability. Stat 140 will capitalize on this, abstraction and computation complementing each other throughout. Students will develop multiple approaches to problem solving, understand the difference between theory and simulation, and appreciate the power of both. It can be used in place of Stat 134 for entry into the Stat major and as prereq for Stat 150-level courses that currently have Stat 134 as a prereq. Beyond Data 8, the prerequisite is a year of calculus.

6.2. Proposals for major and minor programs in Data Science

An overwhelming majority of Data 8 students, when asked in a survey at the end of the Spring 2016 Foundations course, said that they would be interested in either a Data Science major or minor, even though these degree programs do not yet exist.[4] Potential designs for major and minor programs in are being actively explored by cross-disciplinary groups of faculty that are now forming, with strong interest from students who have already taken part in the program. Depending on how other campus processes unfold, the Senate can expect to see proposals from the faculty later in 2016-17.

Starting approaches for possible major and minor programs are found in the DSERAT report. The DSERAT blueprint is being adapted for this year's design discussions, as faculty can now draw on experience with the student population in the entry-level courses and increasingly with the segment moving on to more advanced classes. Along with existing courses that can be drawn into the programs, new courses will likely need to be developed. Themes that can be expected to carry forward in major and minor design are

- Strong conceptual, mathematical, and computational foundations, as would fit with the kind of program (major or minor) and students' expected trajectories
- Integration of students' course-taking with an area of external concentration (such as an application domain)
- Direct attention to human, societal, and organizational facets of data science

[4] On a scale of 1 to 5, more than 50% of Spring 2016 Data 8 students surveyed (418 responses) indicated an interest level of 4 or 5 in a data science major. More than 75% indicated that level of interest in a minor. Among the 193 first-year students, 86% wanted to major or minor.

If solutions that support it are available from campus, these will be broad interdisciplinary major and minor programs rather than narrow disciplinary ones. That will take working through how these programs of study can interact with and leverage course offerings available through other majors. As the data science major and minor programs are designed and students begin engaging with them, other programs may find they can work together with them, so that students find it possible to take (for instance) double-majors in data science and a broad range of other subjects. Along with EECS and Statistics, it is likely that other departments will find they want to internally adjust their curriculum and offerings. For instance, the Math department has a key role in providing course content for likely data science majors, and their engagement in the conversations so far has been powerfully constructive.

Campus-level commitment will make it possible to have thoughtful coordination with a broad range of departments and with campus-level resource allocation processes (TAS, etc.). At this point is possible to give only a rough sense of the likely size of these programs (potentially among the largest on campus), and only general speculations about other follow-on impacts (such as effects on other programs of study). Here planning in the abstract will be less useful than careful tracking, communication, adaptive development, and collective problem-solving at the campus level.

6.3. Other pedagogical developments

Other elements that call for pedagogical development are

- A Summer Bridge offering (anticipated for Summer 2017)
- Additional integration with the programs of the Student Learning Center and other modes of student support
- Support for other programs of study that want to integrate data science approaches by adding new elements or new courses
- Integration with undergraduate research opportunities
- Integration with graduate education

6.4. Prospects

The pace and shape of these efforts will depend on the course of campus-level discussions about organizational institutionalization and fundraising. Efforts so far have been interim and provisional. They have borne more by good-faith commitment from faculty and staff than a solid organizational footing.

More narrow programs can be built on the basis of existing collaboration between EECS and Statistics. The distributed, campus-crossing part of the program cannot be adequately handled in that way. The latter needs dedicated resources, fundraising, and organizational structure that need to stay directly coordinated with these departments. In its present form it is highly at risk.

Appendix 1. Hallmarks of Berkeley's approach

Starting early to build a data science mindset
Many universities have data science programs focused solely on masters or PhD programs and research. A distinctive aspect of Berkeley's program is that it begins early, with entry-level students, in order to introduce data science as core component of a liberal education. Here, data science is approached not as a specialized field for advanced practitioners, but as a way for any student to approach problems across all disciplines.

No prerequisites required
No prerequisites are required to take the introductory Foundations of Data Science course, which covers many concepts that would normally be introduced in upper-division Statistics or Computer Science courses. The course was designed so that students can interact with real data and address questions that would normally take several semesters of coursework to reach.

Integrative, cross-disciplinary collaboration
The data science program builds on decades of collaboration at Berkeley between Statistics, Computer Science, and other disciplines to create courses that add up to something greater than the sum of their parts. The Foundations course was developed collaboratively by a cross-disciplinary group of faculty inspired by the perspectives it could offer opened on their home areas, and connector courses are offered in coordination with a range of disciplines, from Cognitive Science to History to Civil Engineering.

A grassroots, faculty-driven effort
Berkeley's Data Science program has been built from the ground up by the faculty with a collective spirit that could not be achieved by other means. The Foundations course has been designed and taught by a group of faculty from multiple departments coming together. The summer Data Science short course drew participants from a range of disciplines, further embedding data science into instructional programs across campus.

Responsive to students
The Data Science program responds to the direction of student course-taking, patterns around majors and minors, and students' own hopes for their futures. The diversity of student pathways into and through data science is integral to the program's design, with a broadly accessible entry point and modular construction. Students have been part of building and projecting the program for their peers.

Committed to diversity and inclusion
Berkeley's program invites a diverse student body to see itself as part of data science. The structure of the program, the community it fosters, and the pedagogical approach of its entry level courses communicate Berkeley's commitments to inclusion and equity. A year into the program, Data 8 has attracted a student population that is demographically representative of the campus. As Data 8 students are recruited into its teaching staff, they provide role models for their peers. Berkeley's investment in a diverse data science program has potential to change the world at large.

An open-source approach
The Foundations of Data Science textbook and many course materials are freely available online, as is the datascience library developed for Data 8 pedagogy in Python.

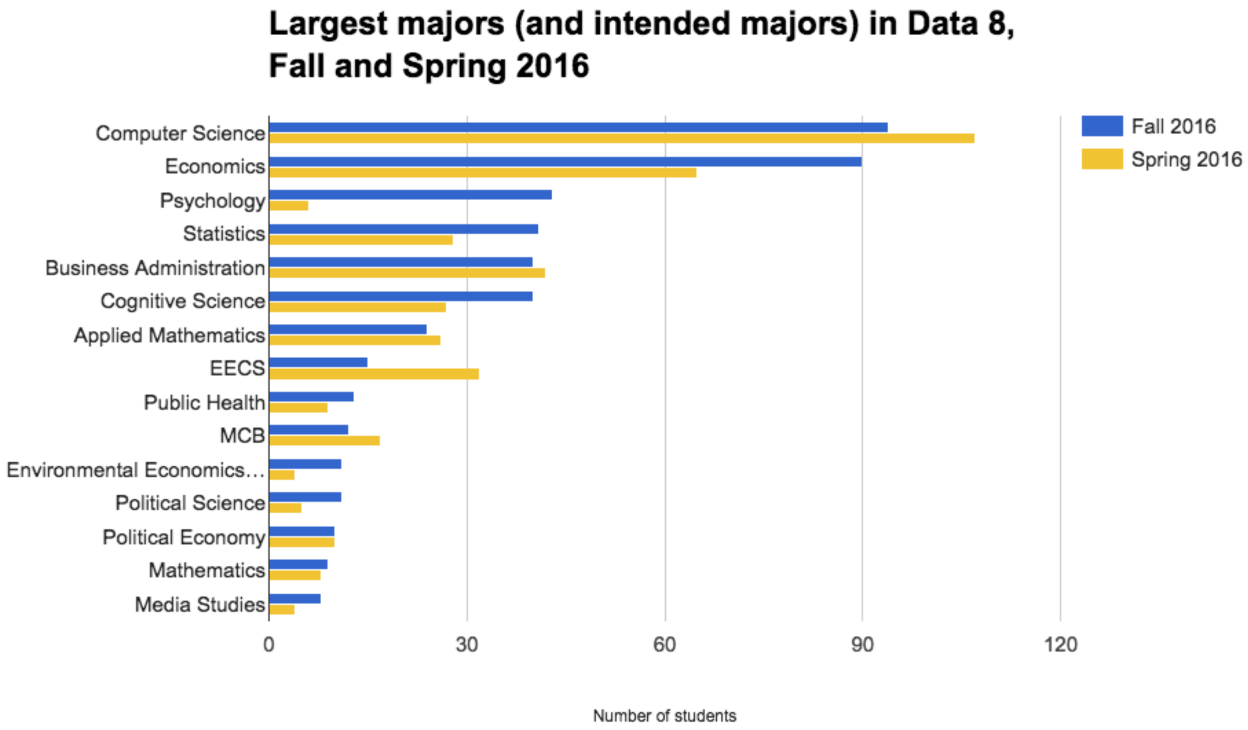
An education program inextricably linked to research
At a campus renowned for its research strength, it is unsurprising that the data science education program is closely tied to the campus' research efforts. Students, at both the undergraduate and graduate levels, are playing important roles with faculty research and through unique platforms such as the Berkeley Institute for Data Science (BIDS). Astronomy professor Joshua Bloom underlines the pivotal role of data science education in preparing the next generation of researchers: "For us in a research institution, we are domain scientists, and we see data science as a prerequisite for our students to do cutting-edge research."

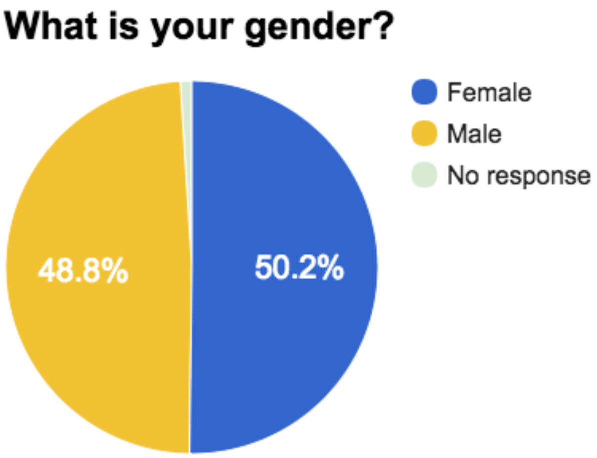
First day of class, Foundations of Data Science, January 2016

Appendix 2: Demographic information

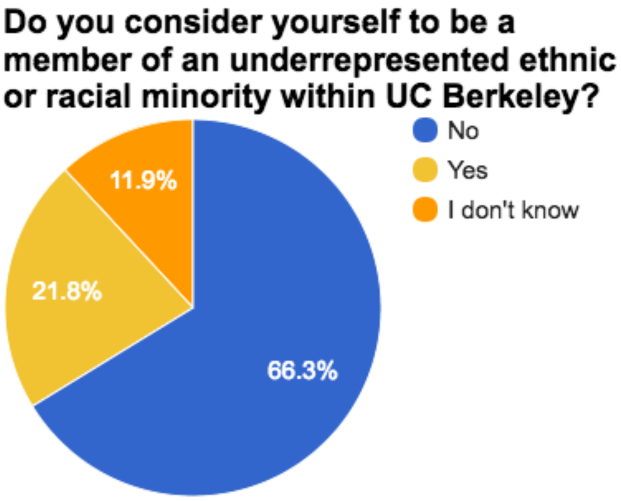
Data 8 Student Survey, Fall 2016
Responses: N=504

Total number of majors (and intended majors) in Data 8, as self-identified: 56

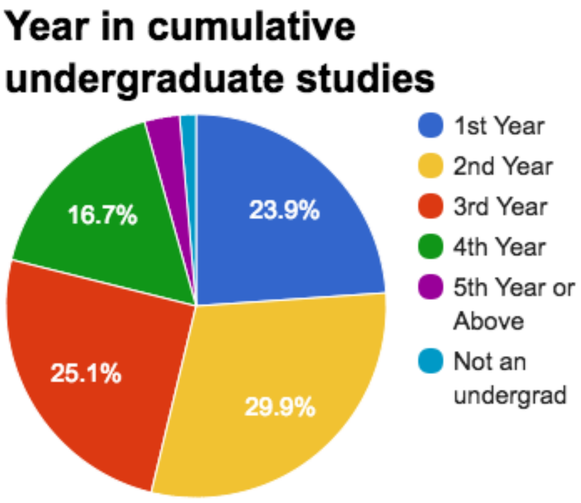




What is your gender?

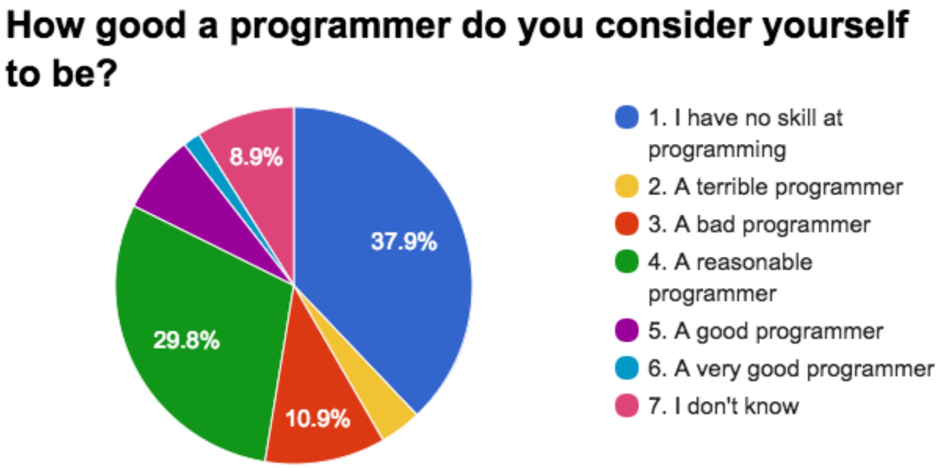


Do you consider yourself to be a member of an underrepresented ethnic or racial minority within UC Berkeley?

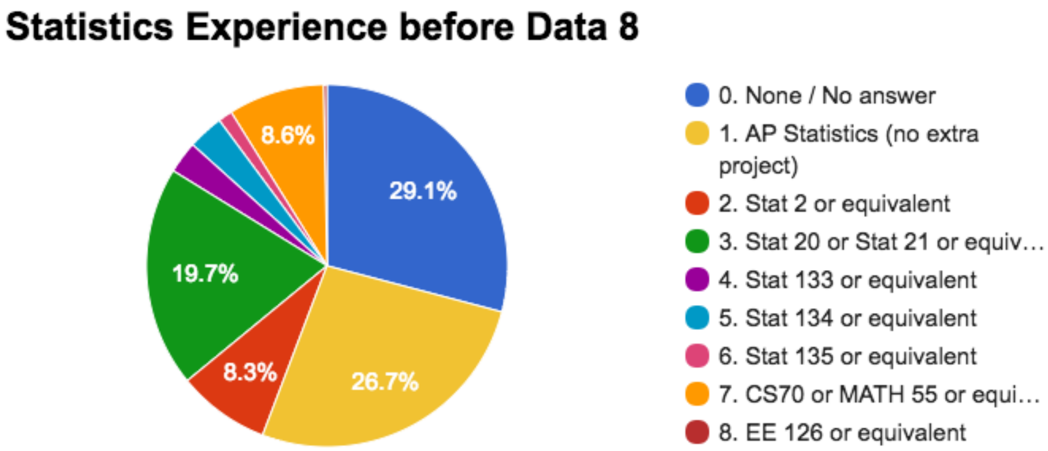


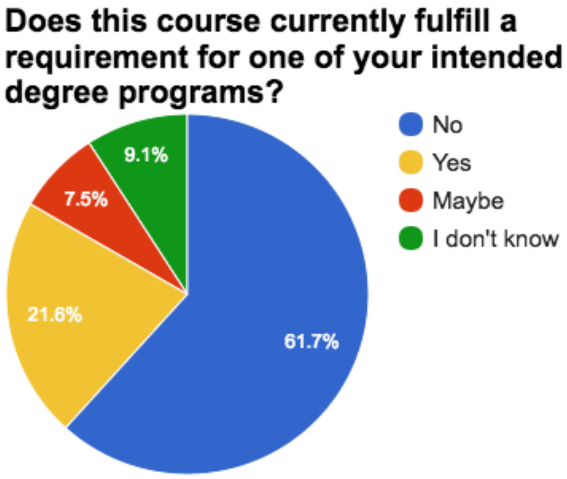
Year in cumulative undergraduate studies

18% have done 1 or more years of undergraduate studies outside of Berkeley

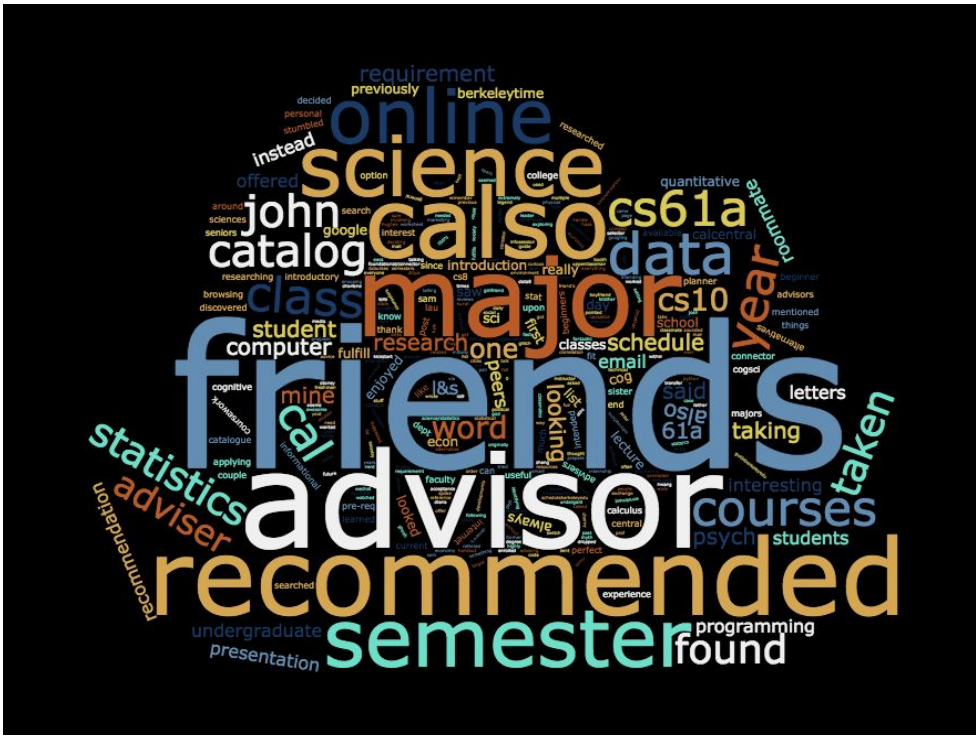


More than 50% of class have limited programming experience (categories 1-3)

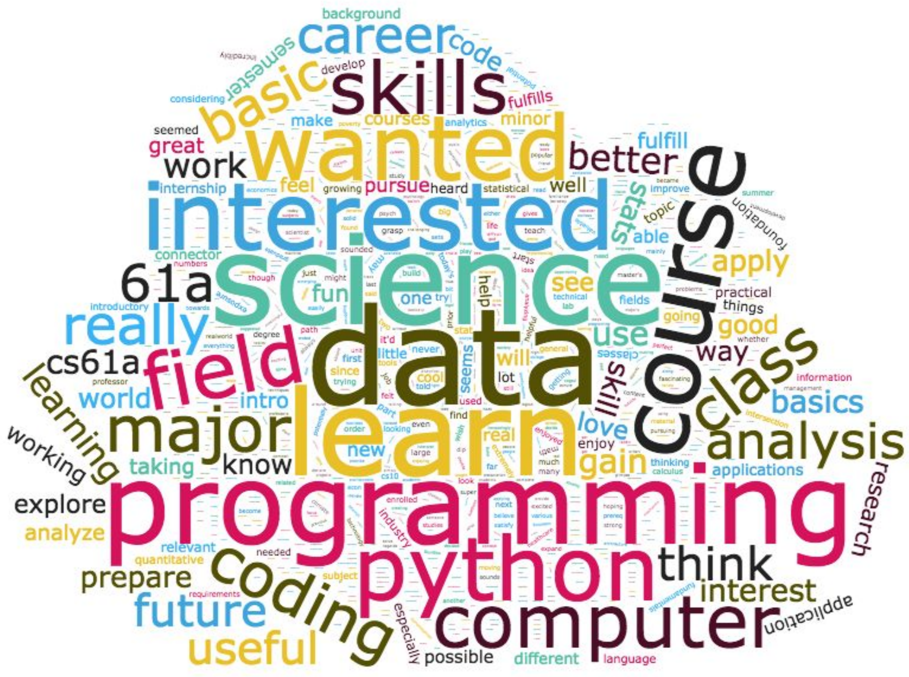




How did you hear about the class?



Why did you take the course?



Total number of majors (and intended majors) in Data 8, as self-identified (Fall 2016)

By size	Count	Alphabetically	Count
Computer Science	94	Anthropology	1
Economics	90	Applied Mathematics	24
TBD	45	Applied Physics	1
Psychology	43	Architecture	3
Statistics	41	Art History	1
Business Administration	40	Biology	1
Cognitive Science	40	Business Administration	40
Applied Mathematics	24	Chemical Biology	1
EECS	15	Chemical Engineering	2
Public Health	13	Chemistry	2
MCB	12	Civil Engineering	3
Environmental Economics and Policy	11	Cognitive Science	40
Political Science	11	Computer Science	94
Political Economy	10	Conservation and Resource Studies	3
Mathematics	9	Dance and Performance Studies	1
Media Studies	8	Development Studies	1
History	4	Economics	90
ISF	4	EECS	15
Physics	4	Engineering Mathematics and Statistics	2
Civil Engineering	3	Engineering Physics	1
Conservation and Resource Studies	3	English	3
English	3	Environmental Economics and Policy	11
Integrative Biology	3	Environmental Science	2
Mechanical Engineering	3	Film	1
ORMS	3	Geography	1
Society and Environment	3	History	4
Sociology	3	IEOR	1
Legal Studies	3	Integrative Biology	3
Philosophy	3	ISF	4
Architecture	3	Legal Studies	3
Chemical Engineering	2	Linguistics	2

Chemistry	2	Materials Science and Engineering	2
Engineering Mathematics and Statistics	2	Mathematics	9
Environmental Science	2	MCB	12
Linguistics	2	Mechanical Engineering	3
Materials Science and Engineering	2	Media Studies	8
Molecular Environmental Biology	2	Microbial Biology	1
Social Welfare	2	Molecular Environmental Biology	2
Urban Studies	2	Music	1
Anthropology	1	Nutritional Science	1
Applied Physics	1	ORMS	3
Art History	1	Peace and Conflict Studies	1
Biology	1	Philosophy	3
Chemical Biology	1	Physics	4
Dance and Performance Studies	1	Political Economy	10
Development Studies	1	Political Science	11
Engineering Physics	1	Psychology	43
Film	1	Public Health	13
Geography	1	Rhetoric	1
IEOR	1	Social Welfare	2
Microbial Biology	1	Society and Environment	3
Music	1	Sociology	3
Nutritional Science	1	Statistics	41
Peace and Conflict Studies	1	Sustainable Environmental Design	1
Rhetoric	1	TBD	45
Sustainable Environmental Design	1	Urban Studies	2

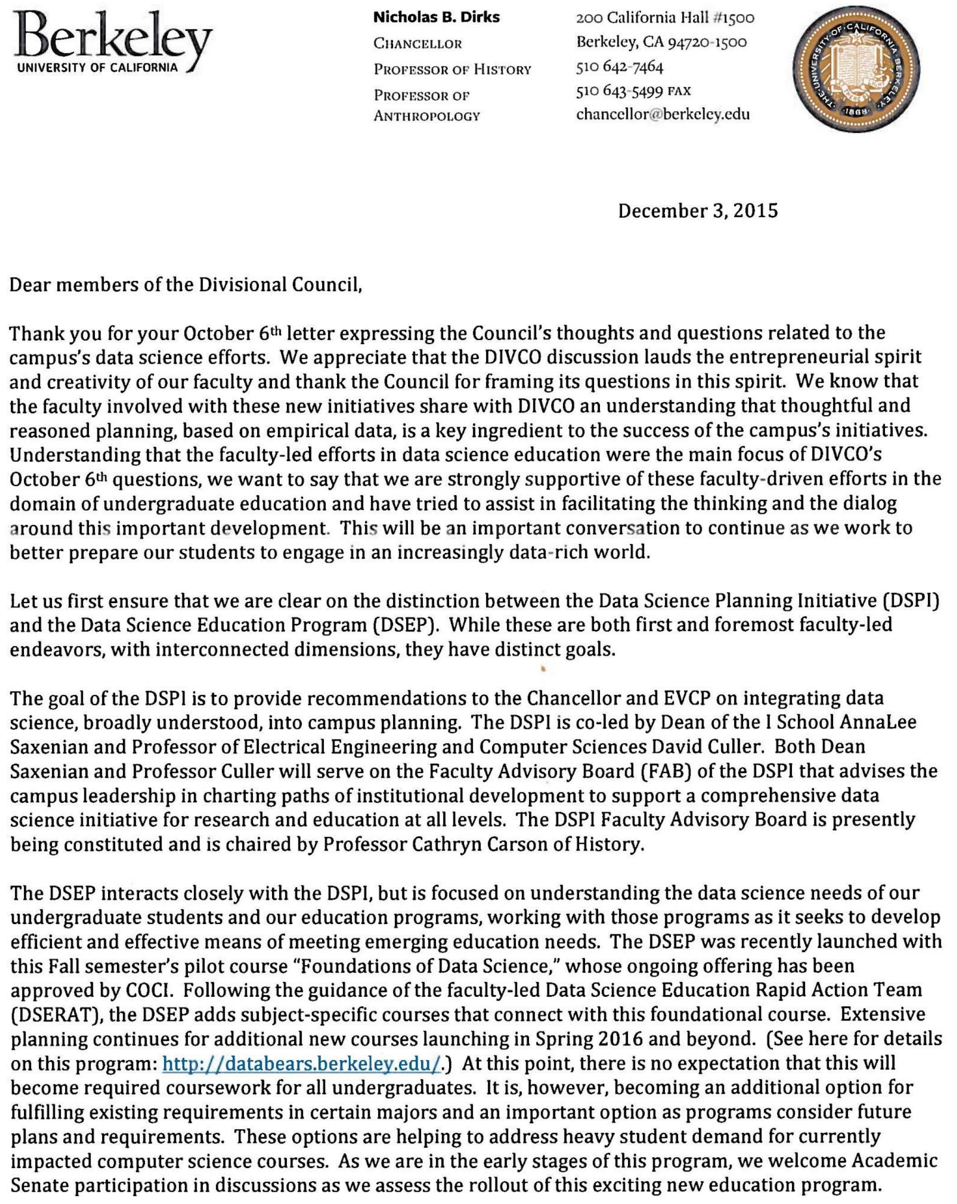
Appendix 3: Connector enrollments

Connectors and enrollments (census data, Cal Answers)

Course	Title	Enrollment		
		Fa 2015	Sp 2016	Fa 2016
CEE 88	Data Science for Smart Cities		30	38
CS 88	Computational Structures in Data Science		21	36
Cog Sci 88	Data Science and the Mind	14	23	21
ESPM 88A	Geospatial Data Explorations	6	9	
ESPM 88B	Data Sciences in Ecology and the Environment		22	
Hist 88	How Does History Count?	5	3	
Info 88A	Data and Ethics		17	37
L&S 39	Race, Policing, and Data Science	6		
L&S 88-1	Health, Human Behavior, and Data	6	14	
L&S 88-2	Literature and Data		6	
L&S 88-1	Child Development Around the World			4
L&S 88-3	Genomics and Data Science			15
L&S 88-4	Social Networks			24
L&S 88-5	Data Science for Cognitive Neuroscience			12
L&S 88-6	Data Science, Demography, and Immigration			10
Stat 88	Probability and Mathematical Statistics in DS	22	54	80
Stat 89A	Introduction to Matrices and Graphs in DS		18	
Enrollment total (17 connectors)		59	217	277
Enrollment in Data 8		109	447	509
Connectors in relation to Data 8 enrollment		0.54	0.49	0.54

(Fall 2015 connectors were taught as L&S 39s.)

Appendix 4. Letter to the Divisional Council of the Academic Senate, December 2015



Berkeley
UNIVERSITY OF CALIFORNIA

Nicholas B. Dirks
Chancellor
Professor of History
Professor of Anthropology

200 California Hall #1500
Berkeley, CA 94720-1500
510 642-7464
510 643-5499 fax
chancellor@berkeley.edu

December 3, 2015

Dear members of the Divisional Council,

Thank you for your October 6th letter expressing the Council's thoughts and questions related to the campus's data science efforts. We appreciate that the DIVCO discussion lauds the entrepreneurial spirit and creativity of our faculty and thank the Council for framing its questions in this spirit. We know that the faculty involved with these new initiatives share with DIVCO an understanding that thoughtful and reasoned planning, based on empirical data, is a key ingredient to the success of the campus's initiatives. Understanding that the faculty-led efforts in data science education were the main focus of DIVCO's October 6th questions, we want to say that we are strongly supportive of these faculty-driven efforts in the domain of undergraduate education and have tried to assist in facilitating the thinking and the dialog around this important development. This will be an important conversation to continue as we work to better prepare our students to engage in an increasingly data-rich world.

Let us first ensure that we are clear on the distinction between the Data Science Planning Initiative (DSPI) and the Data Science Education Program (DSEP). While these are both first and foremost faculty-led endeavors, with interconnected dimensions, they have distinct goals.

The goal of the DSPI is to provide recommendations to the Chancellor and EVCP on integrating data science, broadly understood, into campus planning. The DSPI is co-led by Dean of the I School AnnaLee Saxenian and Professor of Electrical Engineering and Computer Sciences David Culler. Both Dean Saxenian and Professor Culler will serve on the Faculty Advisory Board (FAB) of the DSPI that advises the campus leadership in charting paths of institutional development to support a comprehensive data science initiative for research and education at all levels. The DSPI Faculty Advisory Board is presently being constituted and is chaired by Professor Cathryn Carson of History.

The DSEP interacts closely with the DSPI, but is focused on understanding the data science needs of our undergraduate students and our education programs, working with those programs as it seeks to develop efficient and effective means of meeting emerging education needs. The DSEP was recently launched with this Fall semester's pilot course "Foundations of Data Science," whose ongoing offering has been approved by COCI. Following the guidance of the faculty-led Data Science Education Rapid Action Team (DSERAT), the DSEP adds subject-specific courses that connect with this foundational course. Extensive planning continues for additional new courses launching in Spring 2016 and beyond. (See here for details on this program: http://databears.berkeley.edu/.) At this point, there is no expectation that this will become required coursework for all undergraduates. It is, however, becoming an additional option for fulfilling existing requirements in certain majors and an important option as programs consider future plans and requirements. These options are helping to address heavy student demand for currently impacted computer science courses. As we are in the early stages of this program, we welcome Academic Senate participation in discussions as we assess the rollout of this exciting new education program.

Because the DSPI and the DSEP are both broad-based efforts that do not sit squarely within any particular academic unit, limited campus seed funding, provided primarily to the L&S Dean of Undergraduate Studies, has assisted the activities of the DSPI and the DSEP. This initial funding will support these activities through Fall 2016, by which time the DSPI and DSEP teams with support from the administration will develop a funding plan based strongly on philanthropy and external funding sources that can support the continuation and expansion of these efforts. The DSEP is off to an excellent start, and as the level of student demand becomes increasingly clear, we will (as we do for all course offerings) assess the extent to which TAS support through normal campus processes would be appropriate. We will monitor the program to ensure expenses do not outpace available funds, and decisions on the creation of new courses, the appointment of instructors and GSIs, and the outfitting of classrooms will utilize campus processes.

We agree that it is important to ensure coordination between the DSEP and a wide range of campus departments, as well as with the Undergraduate Initiative (UGI). We can assure you that the DSEP and UGI are indeed in dialog, in part to identify and serve appropriate overlapping goals. Bob Jacobsen, Interim Dean of Undergraduate Studies in L&S and member of the UGI Steering Committee, has responsibility for the planning and growth of the DSEP and reports in this matter to the Provost. The faculty-led and faculty-driven DSEP has been consulting with many dozens of faculty in departments across campus. We are particularly excited to see how multiple units and disciplines are working closely together to develop the program in ways that are mutually beneficial and leverage each other's strengths and capacities. This process has also given rise to important conversations within and across units about how to best serve student needs and enable important research directions, looking toward the future. We expect that these exchanges will naturally inform the conversations to be undertaken by the DSPI, on whose Faculty Advisory Board Dean Jacobsen will serve as well. Under the leadership of the program and dean, the DSEP will continue to assess the student experience of data science on the campus at large and in the courses it offers. Early student responses are very promising.

We expect that the program's growth will be shaped by all the usual elements of the Academic Senate processes. The DSEP's leadership have begun its exchanges with the Senate by engaging COCI as its first step, and, as faculty who are deeply invested in mounting a strong set of course offerings for Berkeley's students, they have aimed to be responsive to COCI's questions. We know that they are committed to continuing their engagement with relevant Senate committees.

We thank the Council for its interest and look forward to continuing to discuss these new and evolving initiatives around data science.

With warm regards,

Nick

Nicholas B. Dirks
Chancellor

Claude

Claude M. Steele
Executive Vice Chancellor & Provost