Data Science Academic Programs Committee Report

March, 2016

Charge and Recommendations

The 2013 report of the "Big Data Implementation Committee"[1] proposed the creation of a Data Science Institute, later implemented as the Data Science Initiative (DSI). As interest and engagement has grown in the DSI, and world-wide interest in Data Science increases, the campus began to consider whether this new field "might benefit from an independent academic home at UC Davis." This committee was charged to "*think futuristically about possible structures, and to consider these issues from the perspective of research, education, faculty, graduate and undergraduate student programs. How an academic unit could foster interdisciplinary research services, outreach and collaboration is another consideration*." We summarize the key recommendations below.

Graduate Education. The committee unanimously recommends that a new graduate group be formed, with its initial priority to immediately define new Master's degree programs and ultimately a new PhD program. We note that there are existing graduate programs in Data Science-related areas, for example, the Medical School operates the Health Informatics graduate group. While we anticipate that Statistics and Computer Science would participate in the new graduate group, they could also create a separate graduate group to more computationally-focused and not cross-disciplinary.

Undergraduate Education. The committee recommends (with 14 of 16 members in favor and 2 abstentions) the creation of a new major that includes the curricular elements discussed in Part 1 below. As part of our deliberations, we considered a number of options for housing a new major, including the creation of a new department, a new FTE-holding program, and possibly expanding the portfolio of existing departments or some other academic structure. Most of the committee felt that locating a new major within existing departments would ultimately constrain the major's curriculum, given the committee's desire to implement an undergraduate major that fully reflected Data Science's integrative and interdisciplinary nature. At the end of deliberations, the majority of the committee recommended forming a new Data Science FTE-holding program or department. It is important to note that, even with a new program or department, close collaboration with existing departments will be critically important.

The following summary discusses salient points from committee deliberations, lays out the advantages and disadvantages of different types of academic programs, defines the necessary faculty expertise to offer the new Data Science major, and loosely discusses potential budget impact.

Part 1. Overview

Definition of Data Science

The Committee began its discussions by coalescing around defining Data Science as comprising (adapted from Donoho[2], Cleveland, Chambers, and Tukey):

- Framing Data-Driven Questions

[1] The full report can be found here **http://tinyurl.com/qyhe7tj**

[2] Particularly https://dl.dropboxusercontent.com/u/23421017/50YearsDataScience.pdf

- Data Acquisition
- Data Exploration and Preparation
- Data Structure and Transformation
- Data Description and Modeling
- Computing with Data
- Statistical and Machine Learning
- Insights and Interpretation
- Data Visualization, Presentation, and Communication
- Data governance, provenance, ethics, security & privacy
- Study of Data Science and its process and pipeline

We note that while there are existing courses for many of these topics, others will require new courses to be defined, and some existing courses may need revising to meet the needs of a Data Science-oriented program.

Core Curriculum of a Data Science Program

The Committee also unanimously agreed that the conceptual framework for a new undergraduate curriculum would include core offerings in each of the 6 areas shown in Figure 1. The Committee strongly believes that Data Science is an integrative and interdisciplinary subject, so students should be educated in the essentials of multiple fields and develop the skills requisite for a synthesis of those fields. We also want students to have more than superficial knowledge of a domain area and to have experience in applying Data Science to that domain. Students will learn to find, frame and solve problems with data and computing through one or more capstone projects, similar to the Design Course in many engineering programs.

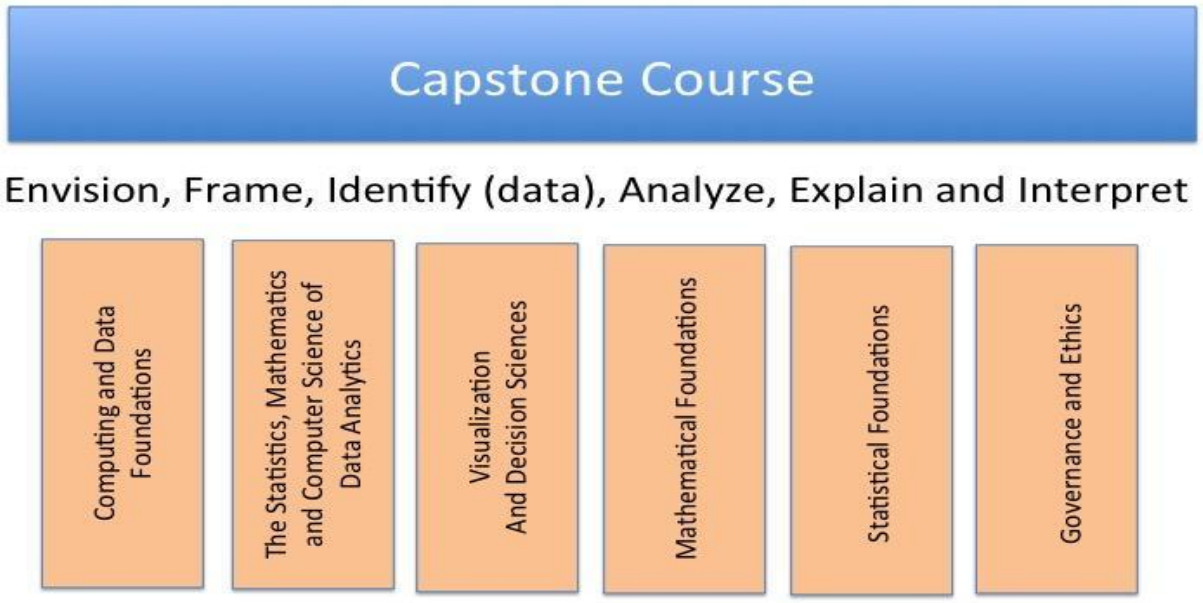

Figure 1. Overview of the core areas of the curriculum

Why Data Science depends on but is not exclusive to Statistics and Computer Science

Data Science relies extensively on concepts and methods from mathematics, statistics and computer science. However, the majority of the committee felt strongly that Data Science extends beyond these aspects, focusing on the entire process and pipeline of data-driven discovery which necessarily means including data acquisition, management, governance, privacy and ethics; data visualization; statistical and machine learning methods; optimization and algebra; workflows; high-performance computing for data analysis; and framing questions and interpreting insights. No single existing discipline provides expertise in all of these aspects, and most existing disciplines would be reluctant to expand their undergraduate curriculum to include courses covering these additional topics.

It is, however, essential that the statistics department retain its deep connection with Data Science as a core discipline. Originally part of the mathematics department, the UC Davis statistics department is a relatively recent creation – beginning as the Intercollege Division of Statistics in 1979, jointly managed by the Colleges of Letters & Science, and Ag & Environmental Science and the Schools of Medicine and Vet Med, and becoming a department in the Math & Physical Sciences Division only in 2000. The statistical end of Data Science is perhaps the most important current research frontier in statistics, and over the past two years the department has added several faculty in the area. The interests of these new assistant professors, as well as those of the department as a whole, need to be considered in the creation of a new undergraduate program and its academic home.

To a lesser extent, a similar situation exists in computer science, which has also hired new assistant professors in aspects of Data Science and plans to recruit more, and has a number of tenured faculty in relevant areas such as data mining, machine learning and visualization. As with statistics, the strength of the Computer Science program going forward depends on developing strength in the computational side of Data Science, and in applications such as computer vision, natural language processing, image processing, social computing and bioinformatics.

Part 2: Possible Academic Structures

New Department or "Program"

The committee recommends the creation of a new academic unit for the Data Science academic degree programs, but notes that there are two options for the unit, each with different strengths and weaknesses.

The first option is to create a new department and the second option is to create an FTE-holding "program"[3]. A program is not a department, but can hold FTE and deliver undergraduate majors. We refer to the convention of establishing such programs in Letters & Science (particularly HArCS and DSS)[4] for supporting new and smaller fields. Programs often have affiliated faculty from other units, so a program structure might be the most expedient way to leverage the data scientists already at Davis in various departments, advance an interdisciplinary research agenda, and build a strong graduate group.

[3] Calling these smaller FTE-holding units "programs" is very confusing, given our frequent reference to academic "programs" of education such as undergraduate major and minor degrees, master's and PhD degrees, etc. We found no good way around this given that administrative "program" units are named as such.

[4] These are listed on the L&S website http://ls.ucdavis.edu/our-college/harcs/depts-programs-harcs/ and http://ls.ucdavis.edu/our-college/soc-sci/depts-programs-dss/

On the other hand, the committee was unclear how such a program might be recognized outside the University and whether the program structure would be strong enough to recruit superior faculty in this highly competitive new field.

The "core" faculty of a newly created academic unit would initially need at least 6 FTE, with additional senior faculty carrying joint appointments in the new unit and their current department. Joint appointments will be a necessary and desirable design feature for either the new department or new program option, so the terms and expectations of these appointments will need to be carefully documented for evaluations and regular review (for example, appointments could be for 5 years with renewal requiring formal approval).

The main advantage of creating a new unit (whether a department or program) lies in the flexibility to specifically design it to meet the needs of a new, highly integrative, interdisciplinary, and evolving STEM discipline. An additional advantage is that a new unit would be best positioned to support, and to collaborate with, Data Science-related programs in specific disciplines and departments, e.g., Business Analytics or Health Care Analytics and Environmental or Precision Agriculture Analytics, because collaborative opportunities via joint appointments or new co-taught courses in these domain-specific programs would challenge existing departments. The committee anticipates that new faculty hires would be trained in a Data Science-related area (Figure 1), but also include significant experience with applied area within other disciplines. A final advantage is that a new unit, which draws expertise from several related disciplines, would be the best at supporting a truly ambitious agenda for innovation and leadership in Data Science, and would be the most likely to survive the test of several decades of activity in Data Science.

The two obvious disadvantages to this model are: 1) new faculty hires will be required and existing faculty will be needed for joint appointments, particularly in the early years, to help establish the department or program and deliver curriculum, and 2) it incurs the overhead of a supporting new academic unit.

With either a new department or program it would be important to avoid unnecessary duplication. It is the Committee's expectation that classes from both statistics and computer science will be necessary. We also see the potential for a more specialized Data Science, computational statistics or programming, or similar programs in statistics, computer science, and possibly other fields, leveraging the same core courses offered by each of these departments.

Existing Department(s)

The Committee also considered the possibility of expanding existing departments in order to offer new Data Science academic programs. The natural departments for this include statistics, mathematics, and computer science. In fact, the committee discussed a proposal provided by the statistics department to house a new undergraduate degree program entirely in that department and to offer a major collaboratively with computer science. A large majority of the committee was not supportive of that approach for several reasons, but in particular the proposal lacked key educational areas (e.g., data governance) and did not fully integrate non-statistical elements (e.g., data acquisition, domain application) of the Data Science curriculum. These are features of the new data science major that the majority of the Committee agrees are essential to the new major.

The curriculum must support an inclusive definition of Data Science, which necessarily includes the full range of topics noted in Figure 1. Because of this, the new major's scope would go beyond those offered by existing departments. A coalition of several departments may have the resources for one or two of the academic programs we envision, and could serve as a stepping stone, but it is unlikely to support all of the features of the undergraduate major that we feel are necessary. Finally, creating new degree programs by simply recycling existing courses from different departments fundamentally fails to recognize Data Science as a new and different intellectual topic, fails to distinguish our programs nationally, and does not leverage new opportunities (spanning research, training and funding).

Part 3: Committee Recommendations and Conclusions

The committee has two main recommendations: the **creation of new academic programs**, including a new undergraduate major, as described in Part 1; the **creation of a new FTE-holding program or department**, as discussed in Part 2.

A new committee should be formed to design the detailed curriculum for the new undergraduate major. We suggest that, given the novelty of such a degree program, an **external advisory group** would be helpful in establishing the new curriculum; this external advisory group should comprise both acknowledged academics leaders in the field (e.g. David Donoho from the Stanford statistics department whose recent publication on Data Science was highly influential to the committee) as well as industry leaders whose workforce we will be developing.

Although the committee took a broad look at the impact of alternative academic structures, we elected to leave the detailed design of a new department or program to a future effort. More discussion is needed with existing departments that will be key contributors to any new Data Science degree programs. While a HIP proposal has been submitted that, if successful, could create new faculty positions necessary for a new major, a **joint appointment framework** that incentivizes faculty to successfully participate in the establishment of this new discipline remains to be developed.

The committee discussed, but did not agree on a college to house the new academic unit, given that several colleges and schools have a strong stake and want a role in its creation and governance. We note that there have been mechanisms in the past to solve this, such as the original formation of statistics as a "cross-college division" managed by a working group of Deans from L&S, CAES, SOM and SOVM, and the more recent Marine Sciences Program. So it is quite possible that an inter-college department or program might be preferable to one housed within a single college.

Finally, in any scenario, the **budget model needs to be reviewed** to provide incentives for current departments and their faculty to participate in offerings delivered through the new program, while also allowing the new department or program to grow and attract excellent new faculty. The committee recommends taking a hard look at how revenues and expenditures would be handled such that there would be sharing of credit for majors that are supported by multiple departments.

In closing, the committee recognizes that an opportunity exists to distinguish UC Davis in Data Science with truly innovative and interdisciplinary educational offerings. We also recognize that capturing this opportunity will require the university to move carefully and with sensitivity but also with alacrity, if we hope to capture the opportunity to provide leading educational programs in this new field. One of our

members reminded us that there are real and growing consequences to our existing programs of *not* moving quickly to begin to offer these new academic programs.

Appendix I -- Data Science Academic Programs Committee Membership

MacKenzie Smith, co-chair (University Library)[1,2,3]
Deb Niemeier, co-chair (faculty; Civil Engineering; School of Education)
Nina Amenta (chair, Computer Science)
Hemant Bhargava (faculty; GSM)[1]
Ian Davidson (faculty; Computer Science)
Joe Dumit (faculty; STS)
Jonathan Eisen (faculty; Genome Center/CBS/EVE/SOM/MMI)[2]
Michael Friedlander (faculty, Math)
Ben Houlton (faculty; LAWR/CAES)[2,3,4]
Yufang Jin (faculty; LAWR/CAES)
Louise Kellogg (faculty; EPS)[1,3]
Thomas Lee (chair; Statistics)
Prasant Mohapatra (faculty; Computer Science)[1]
Tom Nesbitt (faculty, AVC SOM)
Wolfgang Polonik (faculty; Statistics)[2]
Randy Siverson (emeritus, Political Science; Associate Dean, L&S)
Duncan Temple Lang (faculty, Statistics; Director, DSI)[1]

[1] member, DSI Advisory Group
[2] member, Big Data Implementation Committee
[3] member, DSI Implementation Committee
[4] on sabbatical as on January 1, replaced by Yufang Jin

Appendix II – Potential Academic Programs

The McKinsey Global Institute report from 2011 states "By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions." Recent market research evaluating demand for a Business Analytics master's program adds "approximately 2.9 million jobs nationally would benefit from being filled with someone who has a master's degree in business analytics; that figure is expected to grow to 3.4 million by 2025." Analysis of job postings on Web sites such as LinkedIn or Monster.com confirms the demand, especially with an average of between 30%–50% salary increase over traditional positions such as "statistician" and "data analyst".

The statistics major in most universities (including UC Davis) has doubled every 2 to 3 years over the last decade. In this year's applications to graduate programs in computer science at UC Davis, 700 of the 1200 stated their primary interest was in Data Science and/or machine learning. Enrollment in existing Data Science-related courses on the Davis campus also clearly illustrates the demand. And both faculty and students (at all levels and from many fields) are requesting new courses and mentorship to cover many aspects of Data Science across the broad range of opportunities UC Davis supports.

Undergraduate major

A Data Science major provides a broad theoretical and applied training in the entire process and pipeline of Data Science, including statistical and machine learning; data visualization; data acquisition, cleaning and exploratory data analysis; computational reasoning; data technologies; data governance (ethics, privacy, confidentiality) and framing questions and interpreting results to gain insights.

Undergraduate minors

A Data Science minor will expose students from different majors to Data Science concepts relevant to their major and useful when they graduate. This credential improves their hiring prospects, but also allows the best students to pioneer data-driven discovery in their primary field in graduate school and beyond.

A data studies minor is already being designed with a primary audience of social science and humanities students. The focus is on leveraging the different training to ask the right questions, and interpret the results in a richer context. The program also provides the students with sufficient concepts, skills and vocabulary to work with data and with data scientists.

A Data Science track within the major in statistics has also been designed. The implementation of this plan has been delayed due to the long delay in getting approval of key (newly introduced) courses and the ongoing undergraduate program review.

Different departments/programs may decide to create separate programs entitled Data Science minor in x, e.g., economics, environmental science, business. These would rely on courses offered as part of the new Data Science major, and combine these with data-driven courses in the specific field.

Graduate Programs (certificates, Master's, PhD)

There is demand for a master's degree in Data Science, illustrated by the very strong job market. Similar to an undergraduate minor, students will benefit from a graduate certificate in Data Science. Many PhD students in different fields require Data Science training for their own research. Formalizing this as a

certificate gives them some credentials in the field, and also will help develop the hybrid researchers we will need in the future to evangelize data-driven research in these other fields.
Research in Data Science (e.g., infrastructure, workflows) does not have a traditional home in statistics, computer science and mathematics. A PhD program is needed to develop the next generation of researchers in Data Science, complementing those in these other fundamental fields.

Support for other degree programs (e.g. health informatics, biostatistics, business analytics)

Many programs will now, and in the future, add Data Science components to existing courses and programs. Similar to statistics, mathematics and computer science, there will be large demand for Data Science service courses. Some of these will be core components of other majors and graduate programs.