Data Science Silicon Valley Pre-Proposal
March 31, 2015

1 Introduction

This document outlines a pre-proposal to develop a world-class Data Science training program in Silicon Valley. At the core of the proposal is a masters program that will be jointly developed by the departments of Computer Science, Applied Mathematics and Statistics, Economics, and Technology Management. This program is an integral part of a more ambitious Data Science Initiative, which includes the development of a Data Science Research Institute focusing on cutting-edge research as well as community outreach and engagement. The research, teaching, and outreach activities that we envision for the Institute are tightly linked and interdependent, and the expectation is that they will all eventually converge into a vibrant institute that spans both the main campus and the Silicon Valley center. A detailed discussion of the Data Science Initiative can be downloaded from `http://datascience.ucsc.edu/proposals/SVDS-Institute.pdf`. Some key features of the program we propose include:

- A study plan which will strongly integrate multi-disciplinary program training created in collaboration with Silicon Valley industry partners.
- A capstone project that has strong links with local industry.
- Strong synergies with existing educational and research programs on campus and the involvement of ladder-rank faculty.

2 Intellectual Rationale

With the advent of the Internet, big companies such as Google, Facebook, and Amazon, as well as small start-ups, are faced with large-scale, heterogeneous, and diverse datasets. Dealing with these complexities requires a multidisciplinary approach. In that regard, data science is emerging as an encompassing discipline that borrows tools from statistics, computer science and econometrics, and whose models and algorithms support the principled and computationally tractable extraction of knowledge from data, in order to facilitate evidence-based decision making. Since data science is fundamental to the success of a number of technology companies, it is not surprising that job demand for data scientists is high, both in the Bay Area and nationally. In fact, a McKinsey Global Institute report estimates that 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."

Although the number of Data Science programs has grown (both nationally and locally) over the last couple of years, there are still opportunities for a well-designed program to grow and thrive. In particular, few existing programs provide a strong link with industry or an integrated curriculum across disciplines. In contrast, the opportunities for creating a truly integrated Data Science curriculum at UC Santa Cruz are clear. Unlike most other universities, the Computer Science, the Applied Mathematics and Statistics departments and the Technology Management departments all belong to the same academic division, facilitating collaborations, planning and resource sharing. Similarly, although the Economics Department belongs to a different division, it is collocated with CS, AMS and TM in the JBE/E2 complex and already has strong ties with them. We strongly believe that the academic program we have envisioned fills a specific niche in the Silicon Valley ecosystem, and a high-quality program in Data Science like the one we propose to develop has the potential to raise UCSC's national and international visibility.

3 Structure of the Program

The Silicon Valley Data Science training program will offer a new Masters in Data Science. The program will require between 40 and 45 credits (8 to 9 five-credit courses) of classroom instruction, along with a

ten-credit, six month internship in which the students will develop a capstone project. The total duration of the program will be between 12 and 15 months. The curriculum is structured as an umbrella program with a common core of basic courses along with a series of distinct tracks that allow students to acquire specialized knowledge in their preferred area of technical expertise.

- Of the 40 to 45 credits of classroom instruction, between 20 and 25 credits will correspond to four to five core courses that will be taken by all students in the program.
- The remaining 20 credits will correspond to four elective courses.
- We expect the six month internships to be with Silicon Valley companies, but some may be with nonprofits. We also expect many of these internships to be paid, but no guarantee will be offered.

To ensure that the program responds to the needs of local companies we will set up an industry advisory board consisting of industry researchers and leaders from Silicon Valley. We have informally approached some industry researchers and they are excited to be part of this initiative. We plan to provide letters of commitment once we move from the pre-proposal to the proposal stage. We also expect that some of the members of this industry advisory board will also be members of the external advisory boards for the Data Science Institute, thus helping integrate the initiatives.

As the program develops, we will regularly invite researchers from industry to teach or co-teach specialized classes as part of the tracks. This will further enhance the desirability of the MS degree, while also leading to a better collaboration with the industry. Again, many researchers have informally shown interest, and we will provide letters of commitment at a later stage.

Although this pre-proposal focuses on an immersive, full-time MS program, it is worthwhile noting that, depending on demand, the program might also offer a subset of courses in the form of an executive certificate, or even offer part-time/executive version of the program for individuals who are employed in the Silicon Valley area but are unwilling or unable to focus on an MS program full time

4 Faculty Involvement

The core courses are expected to cover topics such as probability, statistical inference, machine learning algorithms, data wrangling, visualization, decision theory, high performance computing and ethics, but the exact details of the curriculum will be decided over the next two to three months in consultation with industry partners and all departments involved. The core courses will be, for the most part, created from scratch and tailored to the program. We expect these courses to be multidisciplinary in nature, so most of them will be cross-listed across more than one of the departments involved. The core courses will be taught in Silicon Valley and telecast to campus so that students in the main campus can benefit from the offerings.

We expect that participating departments will offer at least one "default" track consisting of a sequence of courses coherently organized around one particular theme, but an option will also be offered for students to design their own track by mix-and-matching courses. For the most part, we expect these tracks to be based on existing courses already offered to other UCSC graduate students, which will be offered mostly by ladder-rank faculty either at Silicon Valley with telecast to the UCSC main campus, or vice versa. Under either of the two schemes, instructors will be required to teach at least one third of the lectures from the alternate location, and we plan to build financial incentives into our financial model to encourage faculty to do so. The expectation is that this requirement will help bridge the gap between them Santa Cruz campus and the Silicon Valley Center, stimulating a sense of community. We anticipate that a number of ladder-rank faculty who live in the Valley will be keen to teach at least once a year at the Silicon Valley center.

5 Enrollment Goals and Connections to Research

At steady state, we expect that 30 or more students will be enrolled in the masters in Data Science program.

We have a strong faculty, with substantial research agendas centered around data science. This includes not only pioneers, but also a number of young and mid-career faculty in various divisions of the school of

engineering. Furthermore, we recently made a number of strong new hires in the area. This program will leverage these existing strengths on campus. At the same time, we also expect that some of the masters students enrolled in this program will be motivated to switch to a full time PhD program at the main campus.

There is also significant interest in responsible and inclusive data science. This includes diversifying the workforce, and addressing ethical questions. UCSC is a pioneer in both these areas, and we believe that we can offer a unique and new perspective on these issues.

6 Preliminary resource request

The masters program in the Silicon Valley will be offered as a PDST masters, and will be financially viable in steady state. The track system we propose will enable us to utilize a substantial number of existing courses while at the same time providing enough variety to attract students with diverse backgrounds and interests while keeping costs down. The most resource-intensive aspect of our proposal is the capstone project, which will require faculty supervision (in most cases from research-active faculty) in addition to industry participation.

We expect the annual cost of running the program to be roughly $350,000 (see Table 1 below for details on the planned expenses). Hence, we request an allocation of 4 FTEs to the program (one program director and two FTEs to be housed in Computer Science and one FTE to be housed in Applied Mathematics and Statistics). The money behind these FTEs will be used to support the program until we reach steady-state enrollments and a reserve has been established. At that point all expenses would be covered by PDST fees and we will convert these positions into ladder rank faculty (the program director may become a LSOE and stay in charge of the program, while the other three FTEs would become regular research faculty). See the table below for details.

Program director salary (including salary and benefits, will teach 2 courses)	$150,000
Support staff, full time (including salary and benefits)	$60,000
Five TA quarters @ $12,000 each (including stipend, benefits and fees)	$60,000
Seven lecturer quarters @ $10,000 (including stipend and benefits)	$70,000
Marketing	$15,000
Total	$350,000

Table 1: Planned Program Expenses

Developing the Data Science program in Silicon Valley will require minimal additional infrastructure investment. There is no need for costly wet-lab space, or even for computer labs (as students would be expected to own their own laptops and use them to complete assignments). However, there are some key investments that we believe would greatly enhance the chances that the program would be successful, and that would benefit other programs that might run concurrently in the Valley:

- A regular shuttle could be made available for faculty and students who wish to commute to the Silicon Valley center from the main campus.
- Hot desks, short term, and long term offices could be made available to the faculty and students. Many faculty already live over-the-hill and would welcome the opportunity to visit the Institute 1-2 days a week. Similarly, the students of the faculty who are teaching courses in Silicon Valley might be interested in spending a quarter a year there.
- Our goal is to enable people to come for daily, weekly, or monthly visits. Unlike many other disciplines, which require lab setups, access to (non-digital) archives, etc., data scientists are mobile. It is viable for them to work in multiple places, and there are many advantages to building infrastructure that support this.