A PROPOSAL

FOR A

PROGRAM OF GRADUATE STUDY

IN

DATA SCIENCE AND ENGINEERING

LEADING

TO A

MASTER OF ADVANCED STUDY DEGREE

Revised January 6, 2014
A Proposal for a Master of Advanced Study in Data Science and Engineering.
Department of Computer Science & Engineering in Collaboration with the San Diego Supercomputer Center, UC San Diego

Executive Summary

The Department of Computer Science & Engineering (CSE) in collaboration with the San Diego Supercomputer Center, proposes a new master’s degree program in “Data Science and Engineering” (DSE) leading to a Master of Advanced Study (MAS). Data Science refers to large, diverse, complex, longitudinal, and/or distributed data sets generated from instruments, sensors, Internet transactions, email, video, click streams, and/or all other digital sources available today and in the future.

An increasing fraction of human and non-human activity around the world is digitized and shared, resulting in a tsunami of data covering all aspects of life. Managed well, this data can be used to unlock new sources of economic value, provide fresh insights into science and hold governments to account.

Processing big data also creates new challenges. Despite the abundance of tools to capture, process and share all this information-sensors, computers, mobile phones and the like-it already exceeds the available storage space. Moreover, protecting privacy is becoming harder as increasing amounts of information on individuals is collected and shared between a huge variety of services, from cell phones to social networks to medical institutions.

A new kind of professional has emerged, the data scientist, who combines the skills of software programmer, database manager, statistician and storyteller/artist to create mathematical models of the data, identify trends and deviations, and present them in effective visual ways that can be understood and appreciated by others. The term "data scientist" is to the point. The scientist's goal is to understand systems by creating simplified models that capture its essence. The data scientist aims to model systems which manifest themselves through data that is often much larger than the capacity of typical machines. The system can be biological (genomic and clinical data from malignant tumors of all cancer patients), physical (plate-tectonics or a bridge) or human communication (social networks, real-time data generated by mobile devices), but the goal is the same - simplified models that can be used to predict or identify major events.

The program is designed for working professionals with a broad background in education and/or training in related areas of computer science, or other engineering or mathematics and with substantial experience in data analysis. It consists of seven required courses (three foundational courses and four core courses), two electives chosen from six course options, and a two-quarter capstone project course, for a total of 38 units. The program is designed to be completed over a two-year period. The admission selection and program administration will benefit from the process used by current MAS programs offered at UCSD under the supervision of the Office of Graduate Studies.
I. Table of Contents

II. INTRODUCTION ................................................................................................................................................. 4

   A. GOALS AND OBJECTIVES OF THE PROGRAM ........................................................................................................ 4
   B. HISTORICAL DEVELOPMENT OF THE FIELD ........................................................................................................ 4
   C. TIMETABLE FOR DEVELOPMENT OF THE PROGRAM .............................................................................................. 5
   D. RELATIONSHIPS OF THE PROGRAM TO EXISTING PROGRAMS ........................................................................... 5
   E. INTERRELATIONSHIP OF THE PROGRAM WITH OTHER UC INSTITUTIONS ................................................................. 5
   F. DEPARTMENT OR GROUP WHICH WILL ADMINISTER THE PROGRAM ........................................................................ 6

III. PROGRAM .......................................................................................................................................................... 6

   A. REQUIREMENTS .................................................................................................................................................. 6
   B. FOREIGN LANGUAGE ........................................................................................................................................... 6
   C. PROGRAM OF STUDY ........................................................................................................................................ 6
   D. CAPSTONE PROJECT, EVALUATION, AND EXAMINATIONS ...................................................................................... 10
   E. QUALIFYING EXAMINATIONS .............................................................................................................................. 11
   F. THESIS AND/OR DISSERTATION .......................................................................................................................... 11
   G. FINAL EXAMINATION ........................................................................................................................................ 11
   H. SPECIAL REQUIREMENTS OVER AND ABOVE GRADUATE DIVISION MINIMUM REQUIREMENTS ......................... 11
   I. RELATIONSHIP OF MASTER’S AND DOCTOR’S PROGRAMS ................................................................................... 11
   J. SPECIAL PREPARATION FOR CAREERS IN TEACHING .......................................................................................... 11
   K. SAMPLE PROGRAM ........................................................................................................................................... 11
   L. NORMATIVE TIME FROM MATRICULATION TO DEGREE .......................................................................................... 12

IV. PROJECTED NEED ............................................................................................................................................ 12

   A. STUDENT DEMAND FOR THE PROGRAM ............................................................................................................. 12
   B. OPPORTUNITIES FOR PLACEMENT OF GRADUATES ............................................................................................ 13

V. IMPORTANCE TO THE DISCIPLINE .................................................................................................................. 13

VI. WAYS IN WHICH THE PROGRAM WILL MEET THE NEEDS OF SOCIETY ............................................................ 14

VII. RELATIONSHIP OF THE PROGRAM TO RESEARCH & INTERESTS OF THE FACULTY ............................................... 14

VIII. PROGRAM DIFFERENTIATION .......................................................................................................................... 14

IX. FACULTY .......................................................................................................................................................... 14

X. COURSES .......................................................................................................................................................... 15

   A. FOUNDATIONAL COURSES .................................................................................................................................... 16
   B. CORE COURSES .................................................................................................................................................... 16
   C. ELECTIVE COURSES .......................................................................................................................................... 17
   D. CAPSTONE COURSE .......................................................................................................................................... 18

XI. RESOURCE REQUIREMENTS ............................................................................................................................ 18

XII. GRADUATE STUDENT SUPPORT ........................................................................................................................ 20

XIII. CHANGES IN SENATE REGULATIONS ................................................................................................................ 20

XIV. PROGRAM GOVERNANCE AND ADMINISTRATION .......................................................................................... 20

XV. APPENDICES .................................................................................................................................................... 22

   A. BIOGRAPHY OF PARTICIPATING FACULTY ........................................................................................................... 22
   B. LETTERS OF SUPPORT ......................................................................................................................................... 45
   C. INFORMATION REQUIRED BY CPEC .................................................................................................................. 56
   D. COURSE APPROVAL FORMS .................................................................................................................................. 59
   E. CATALOG COPY ..................................................................................................................................................... 73
   F. INSTITUTIONS WITH DATA ANALYTICS PROGRAMS ............................................................................................ 76
   G. EXTERNAL REVIEWS AND RESPONSE ................................................................................................................ 79
II. Introduction

A. Goals and Objectives of the Program

The proposed program is offered by the Department of Computer Science and Engineering (CSE) in collaboration with the San Diego Supercomputer Center (SDSC). Its goal is to provide an opportunity for training in the emerging area of Data Science and Engineering (DSE) to working professionals drawn from government and industry. This goal is supported by a program with specific objectives to teach theoretical and practical aspects of data modeling, data analysis, data query and curation, machine learning, programming and benchmarking, and to offer these instructions in a format that is accessible to the designated student demographics through day-long sessions on Fridays and weekends. It is an objective of the program to offer instruction that is supported by substantial hands-on experience in the practice of DSE topical areas.

B. Historical Development of the Field

When the Sloan Digital Sky Survey started work in 2000, its telescope in New Mexico collected more data in its first few weeks than had been amassed in the entire history of astronomy. Now, a decade later, its archive contains a whopping 140 terabytes of information. A successor, the Large Synoptic Survey Telescope, due to come on stream in Chile in 2016, will acquire that quantity of data every five days.

Such astronomical amounts of information can be found closer to Earth too. Wal-Mart, a retail giant, handles more than one million customer transactions every hour, feeding databases estimated at more than 2.5 petabytes (PB) -the equivalent of 167 times the books in America's Library of Congress. Facebook, the social-networking website, is home to 40 billion photos. Decoding the human genome involves analyzing three billion base pairs--which took 10 years the first time it was done, in 2003, but can now be achieved in one week.

All these examples point to the same phenomenon: an increasing fraction of human and non-human activity around the world is digitized and shared, resulting in a tsunami of data covering all aspects of life. Managed well, this data can be used to unlock new sources of economic value, provide fresh insights into science and hold governments to account. But they are also creating a host of new problems. Despite the abundance of tools to capture, process and share all this information (sensors, computers, mobile phones and the like) it already exceeds the available storage space.

A new kind of professional has emerged, the data scientist, who combines the skills of software programmer, database manager, statistician and storyteller/artist to create statistical models of the data, identify trends and deviations and present them in effective visual ways that can be understood and appreciated by others. The term "data scientist" is to the point. The scientist's goal is to understand systems by creating simplified models that capture its essence. The data scientist aims to model systems that manifest themselves through big data. The system can be biological (genomic and clinical data from malignant tumors of all cancer patients), physical (plate-tectonics or a bridge) or human communication (social networks, real-time data generated by mobile devices), but the goal is the same - simplified models that can be used to predict or identify major events.
While the organization and analysis of data as outlined above seems daunting and insurmountable, this can indeed be overcome by the recent developments. Advances in methods and tools that help us organize data (Hadoop, Big Table), search and query data (SQL), reduce complexity (dimensionality reduction) and learn from data (machine learning) have enabled us to mount an attack that organizes and makes sense of the volume of data. It is, therefore a timely goal to put together a graduate program that teaches students important concepts and skills that go beyond practical issues in various domains where Data Science is practiced.

C. **Timetable for Development of the Program**

We anticipate that this program will begin in the Fall of 2014. Initial enrollment is estimated at 22-26 students with approximately 25-30 new students per year at steady-state, for a total of 50-60 enrolled students at any given time.

<table>
<thead>
<tr>
<th>Event</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Industry Visits and Needs Assessment</td>
<td>Summer 2011</td>
</tr>
<tr>
<td>Proposal Exploration Team Formed</td>
<td>Fall 2012</td>
</tr>
<tr>
<td>Proposal Submitted to UCSD Graduate Council</td>
<td>April 2013</td>
</tr>
<tr>
<td>Revised Proposal Submitted</td>
<td>Fall 2013</td>
</tr>
<tr>
<td>UCSD Proposal Submission to CCGA</td>
<td>Fall 2013</td>
</tr>
<tr>
<td>CCGA Approval</td>
<td>February 2014</td>
</tr>
<tr>
<td>UCOP Approval</td>
<td>March 2014</td>
</tr>
<tr>
<td>Marketing and Information Sessions</td>
<td>April 2014</td>
</tr>
<tr>
<td>Applications</td>
<td>April-June 2014</td>
</tr>
<tr>
<td>Admission of First Cohort</td>
<td>Late Spring-Summer 2014</td>
</tr>
<tr>
<td>Orientation</td>
<td>Summer 2014</td>
</tr>
<tr>
<td>Program Offered and Courses Begin</td>
<td>Fall 2014</td>
</tr>
</tbody>
</table>

D. **Relationships of the Program to Existing Programs**

The area of Data Science has emerged as one of the most important areas of study, and CSE and SDSC are among the leaders in this field. We understand that related efforts are in progress at other campuses given the currency of the topic and its need. At the time of this writing, we believe that the only masters level program in analytics currently offered in California is through the College of Arts and Sciences, University of San Francisco. In addition, only two programs (Stanford and UCSD) offer certificates in Data Mining. Within the UC system, UC Berkeley recently launched a 27-unit Masters in Information and Data Science (MIDS). However, it is entirely on-line, except for one 4-5 day immersion session on the campus.

The proposed program builds upon a number of courses offered in the Department of Computer Science at UCSD. There is also a strong undergraduate program in a related area of databases and machine learning that offers access to preparatory courses for student who may need these. Nationwide, there are several institutions offering master degree programs in analytics, and these are primarily offered by business schools and directed towards that audience.

E. **Interrelationship of the Program with Other UC Institutions**

There is no interrelationship with other UC Institutions.
**F. Department or Group Which Will Administer the Program**

The program will be offered by the Department of Computer Science and Engineering (CSE) in close collaboration with the San Diego Supercomputer Center (SDSC). The pedagogical direction of the program, admission decisions, outreach to industry, financial decision making, and strategy for student advising, will be guided by a steering committee that will include the Chair of CSE, Director of SDSC, Associate Dean for MAS programs, and director and co-director of the program. In addition, academic program review and student advising will be the responsibility of the CSE department.

Based on our experience with the management of other MAS programs, we anticipate a continuous evaluation process that includes not only traditional evaluation measures for a graduate program such as student feedback, external review and graduate council oversight, but also input from the major companies and industry sectors from which we draw students. These evaluation components and measures taken in response will be part of our official review record for evaluation by the Graduate Council at its periodic program reviews.

As per Senate requirements, the program will be reviewed on an eight year cycle.

**III. Program**

**A. Requirements**

The MAS in Data Science and Engineering is directed to students with backgrounds in engineering, physical sciences, social sciences and business, but with a strong background in mathematics (linear algebra, calculus, probability and statistics), programming and algorithmic thinking, problem solving and independent thinking, and knowledge of programming languages. A student will be expected to have a working knowledge of cloud infrastructure and some experience with analytic methods and tools. Because prospective students are expected to be full-time working professionals, the lecture-style courses will be offered in a format and at a time convenient for working professionals (this could be in the evenings or on alternating Friday-Saturdays).

Admissions requirements are equivalent to those for the Master of Science with one exception: two years of work experience may be substituted for GRE exam scores.

- Bachelor’s degree in science or engineering, with priority for computer science, electrical engineering, statistics, economics, physics, chemistry, and bio-informatics
- Undergraduate GPA of at least 3.0 on 4.0 scale
- Two years of work experience in statistical data analysis
- Work experience with data management, data organization, and data analysis is considered a plus
- Three letters of recommendation, one of which should be from the applicant’s current employer. For those who are unemployed, the current employer requirement may be waived.
- TOEFL or TSE scores (international applicants only)

**B. Foreign Language**

Foreign language is not required for this degree.

**C. Program of Study**

Unit Requirements
The MAS in Data Science and Engineering program proposes a 38-unit/10-course degree that is to be taken in consecutive quarters (excluding summers) over a period of two years. The curriculum consists of six 4-unit required courses (three foundational and three core courses), one 2-unit core course (a case study seminar), two 4-unit elective courses and a 4-unit capstone team-based project course that spans two quarters. The capstone will require a combination of in-class, laboratory, and off-campus work. See Section I.K for a sample two-year course program and sample elective pairings by emphasis.

Data Science is an area that combines computer science, statistics and data analysis. In our experience, the people that succeed in this area belong to roughly three groups:

1. People with a degree in computer science, typically working in the Information Technology (IT) department of an organization.
2. People with a degree in statistics, typically working in the Business Intelligence (BI) department of a company.
3. People with degree in a physical science (physics, chemistry, meteorology, biology etc.). Data analysis is typically a central part of these people’s experience and passion.

The goal of our program is to deepen the students understanding in their original field and to provide them with the language and concepts to be able to effectively communicate with people with a different background.

First Year: Foundational and Core Courses

The first year of the MAS students complete six of the seven required courses. The goal of the first year is bring all students to a level of competency in the foundational and core subject areas of the degree and to prepare them for collaboration in the second year. They earn 22 of their required 38 units of the degree.

In fall quarter, students complete one of the core courses, a 2-unit seminar, [DSE290] Case Studies in Data Science. The case studies expose students to the needs and uses of different technologies and their roles in model building. The subjects introduced are covered in detail in the rest of the curriculum, with in-depth training that utilizes these concepts in the capstone project.

As industry practitioners, students will come into the program with differing strengths and knowledge. The three required foundational courses offered in fall and winter quarters are designed to bring all students up to an equal knowledge base in programming, statistics, and databases. The foundational courses are:

1. [DSE200] Python for Data Analysis. In order to enable students to collaborate with other students, they need a common programming language. We chose python because it is a relatively easy language to learn, because of the rich open-source libraries for data analysis and visualization.
2. [DSE210] Introduction to Probability and Statistics. Statistics is the foundation of data analysis. Computer science students typically have weak backgrounds in statistics. This course is modeled after [CSE103] with the same name and is designed to familiarize students with basic concepts in probability and statistics.
3. [DSE201] Data Management Systems: Big data is organized in big databases. Knowledge of SQL is therefore a requirement for DSE. Students with a strong foundation in one of these areas can waive the need to take the course by passing an exam, and may then complete additional electives to fulfill the units requirement.
In Spring quarter of the first year, students complete two more core courses:

1. [DSE230] Data Analysis using Hadoop and Spark. This course is central to the whole program. It is here that the students are introduced to very large and massively parallel data storage systems and to the map-reduce framework for parallel computation. The most popular framework for big data analysis.

2. [DSE220] Machine Learning. This course extends the basic concepts introduced in [DSE210] and familiarizes the students with machine learning algorithms for classification, regression, clustering and semi-supervised learning.

Second Year: Core and Elective Courses and Capstone Project

During the second year, students complete their last required core course, two electives, and the capstone team project. They earn the remaining 16 units for the degree.

In fall quarter, they take the core course [DSE203] Data integration and ETL. The course is designed to provide students with the fundamentals of data integration.

Elective courses specialize in subareas, and there are some dependencies between the courses (see diagram and list of courses below). The students will request their desired electives at the end of the first academic year. These requests will be based on the student’s interests and input from the faculty advisors about the training needed for teams to successfully complete their capstone projects. An elective course might not be offered if the number of students requesting it is too small (the threshold is expected to be around 10).

The final part of the MAS degree is a two-quarter capstone team project completed in winter and spring of the second year (2 units quarter, 4 units total; in progress (IP) grade in winter with final grade in spring). The capstone project requires that students apply the course material to a realistic engineering project of potential interest and value to a company, industry, or government entity.

Each capstone project team will consist of 3-4 students. The students should complement each other, encompassing at least two of the three backgrounds described above. The groups will form at the end of the first year and will begin to define their project with a faculty advisor. Based on this definition the advisor will determine which elective courses have to be taken by at least one person in the group. The capstone will require students to access the facilities in the San Diego Supercomputer Center that consists of one or more clusters, each with a capacity of 1 Petabyte (PB).

Course Dependencies

The following graph depicts prerequisite dependencies between the required foundational and core courses, and electives.
List of Courses

**Foundation Courses**
- DSE 200 Python for Data Analysis
- DSE 201 Data Management Systems
- DSE 210 Statistics and Probability Using Python

**Core Courses**
- DSE 203 Data Integration & ETL
- DSE 220 Machine Learning
- DSE 230 Data Analysis Using Hadoop, and Spark
- DSE 290 Case Studies in Data Science

**Elective Courses**
Elective courses offered each year will depend on the interests of students and training needs of the capstone. A course may not be offered if a threshold of 10 students is not reached.
- DSE 221 Data Analysis Using R
- DSE 232 Performance Measurement
- DSE 240 Online Analytics Applications
- DSE 241 Data Visualization
- DSE 250 Beyond Relational Data Models
- DSE 251 Managing Large-Scale Graph Data

**Capstone Course**
- DSE 260 Data Science Design
Schedule and Mode of Delivery
The program will be delivered at a time convenient for working professionals and may be offered in a concentrated timeframe, such as alternating Fridays and Saturdays each quarter, with each course conforming to equivalent hours in an academic quarter.

The department has extensive experience with existing MAS programs both in terms of the use of a learning management system (LMS) and digital media by students. MAS programs must accommodate the needs of professional, working students by providing access to course materials and communication tools outside of normal 9-5 work hours. Students are expected to complete work when they are not in class, and the LMS serves as a central location for access to homework, grades, digitally recorded class sessions, discussion boards etc., and other communication tools.

All courses in the program are set up in the LMS with a course webpage that includes basic information such as syllabi, homework, contact information, digitally recorded lectures, discussion boards, etc. How each instructor chooses to use the tools varies depending on the preference of the instructor. Some require students to answer questions posed and monitor students’ online responses; others simply give the students the option of participating in the use of forums, etc.

Our experience with existing MAS courses is that student attendance is nearly 100% despite the fact that roll is not taken. Students find in-class structure and participation beneficial; in addition there are also opportunities for interaction and networking around the class schedule.

While working professionals acquire much knowledge in the course of their work life, some may have been away from the classroom for some time and may need additional review of the material presented in class. In addition, we recognize that on occasion students will have to miss class due to illness or some other unavoidable circumstance. Since the program provides information in a very intensive format, missing a class can have very serious consequences if a student were unable to make up the work. For all these reasons, we digitally record video and sound for all MAS classes, and our students find review of these recordings extremely valuable.

D. Capstone Project, Evaluation, and Examinations
There are no field examinations. However, the MAS DSE program will require a 4-unit capstone team-based project course that runs as an in-progress course in winter and spring quarters of the second year. The final grade is assigned at the end of spring quarter.

The project topic will be determined with the assistance of the faculty lead in the program. If students are unable to select a topic of relevant interest, possible topics for consideration will be provided for them. The intent is that the capstone project will be a 3-4 person team design project conducted with people who work in the same company. If this is not possible, the capstone project may be an individual effort supervised by the lead faculty member. Each team will have a lead faculty member, and the same faculty member may lead more than one team.

The main objective of this project will be for the small teams to demonstrate their critical thinking, organization, and design skills in attacking a problem within the DSE field. At the end of the quarter, a final report must be handed in and a short technical oral presentation must accompany the deliverable. In addition, there might be a demonstration of the working prototype. The presentation and report will be reviewed by both the instructor and the other class participants in a peer-to-peer review process.
Performance on the capstone project will be used as the primary measure of the student’s learning in the MAS program. Students will be evaluated based on their grasp of the course material, ability to apply the course material to the capstone project and individual contribution to the overall success of the team project. Satisfactory completion of the capstone project is required for award of the MAS degree.

**E. Qualifying Examinations**
No Qualifying Examination is required.

**F. Thesis and/or Dissertation**
No thesis or dissertation is required.

**G. Final Examination**
No final examination is required. Evaluation of the completed capstone project will serve as the final examination for the program.

**H. Special Requirements Over and Above Graduate Division Minimum Requirements**
There are no special requirements over and above Graduate Division requirements.

**I. Relationship of Master’s and Doctor’s Programs**
This program will result in a supplemental Master’s degree with no relationship to a doctoral program.

**J. Special Preparation for Careers in Teaching**
Teaching is not a component of this degree.

**K. Sample Program**
There are 10 courses totaling 38 units (including the 4-unit capstone project) required for completion of this program. Each 4-unit quarter-length class is equivalent to 30 lecture-hours. The lecture-style classes and case study seminar (DSE 290) may be conducted from 8:00 AM to 5:00 PM alternating Fridays and Saturdays. All courses will fall within the start and end dates of each quarter. The courses in the DSE program are interrelated so that the student gains increasing expertise in the two key domain areas, with the capstone project synthesizing the learning at the end of the process. For a description of which electives will be offered each year, see Section C, Second Year section.

<table>
<thead>
<tr>
<th>Sample Two-Year Plan (Instructors for Cohort 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Year</strong></td>
</tr>
<tr>
<td>---------</td>
</tr>
</tbody>
</table>
| Year 1  | DSE 200-Freund  
          DSE 290-Baru | DSE 201-Papakonstantinou  
          DSE 210-Dasgupta | DSE 220-Balac  
          DSE 230-Freund |
| Year 2  | DSE 203-A. Gupta  
          Elective | DSE 260 (In Progress)-Papakonstantinou  
          Elective | DSE 260 (Completion)-Papakonstantinou |
### Sample Elective Pairings by Subareas

<table>
<thead>
<tr>
<th>Subarea</th>
<th>Fall</th>
<th>Winter</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Analysis</td>
<td>DSE 232</td>
<td>DSE 221</td>
</tr>
<tr>
<td>2. Databases A</td>
<td>DSE 240</td>
<td>DSE 241</td>
</tr>
<tr>
<td>3. Databases B</td>
<td>DSE 250</td>
<td>DSE 251</td>
</tr>
<tr>
<td>4. Database Mix</td>
<td>DSE 240</td>
<td>DSE 250</td>
</tr>
<tr>
<td>5. DB-An Mix A</td>
<td>DSE 240</td>
<td>DSE 232</td>
</tr>
<tr>
<td>6. DB-An Mix B</td>
<td>DSE 240</td>
<td>DSE 221</td>
</tr>
<tr>
<td>7. DB-An Mix C</td>
<td>DSE 250</td>
<td>DSE 232</td>
</tr>
<tr>
<td>8. DB-An Mix D</td>
<td>DSE 250</td>
<td>DSE 221</td>
</tr>
<tr>
<td>Possible Courses Taught</td>
<td>DSE 232, 240, 250</td>
<td>DSE 221, 232, 241, 250, 251</td>
</tr>
</tbody>
</table>

### L. Normative Time From Matriculation to Degree.

The normative time from matriculation to degree is two years. This takes into account the target student population of experienced working professionals, as well as the course sequence as outlined in the previous section.

### IV. Projected Need

#### A. Student Demand for the Program

The proposed program is designed for experienced working professionals who want to augment their education in engineering. Industry need and student demand for this program is very strong, as is discussed in the following paragraphs.

Roughly speaking, business employees that work in analytics have two types of background. The first are people with background in business, statistics and modeling. The position of such people is often called Business Intelligence (BI) or Analytics. The second are people with backgrounds in computer science and computer programming. These people often work in the Information Technology (IT) department. With the advent of analytics in general and Data Science in particular, there is a growing need for people that have a solid understanding of both BI and IT. The goal of our program is to educate people with such background by emphasizing both areas and showing their inter-relations.

The concept for the DSE program emerged from discussions with local industry representatives of our Departmental Advisory Councils and the Corporate Affiliates Program. We expect that there will be a robust demand for this program, due to the broad application of the topic of study and large availability of data intensive companies in the San Diego area.

The clearest market analysis in support of the opportunity in ‘Big Data’ and long-term trends in this domain comes from a recent McKinsey Global Institute report that identified ‘Game Changer’ opportunities for US growth ([http://www.mckinsey.com/insights/americas/us_game_changers](http://www.mckinsey.com/insights/americas/us_game_changers)). Among the five such opportunities (energy, trade, infrastructure and talent), Big Data was identified as accounting for .8-1.7% of the GDP of 2020. The May 2011 McKinsey Global Institute report, Big Data: The next frontier for innovation, competition and productivity ([http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation](http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation)), predicted that the need of data scientist will exceed their production by 140,000-190,000 by the year 2018. McKinsey projected a
shortfall of 1.5 million additional managers and analysts in the U.S. who can “ask the right questions and consume the results of the analysis of big data effectively.”

In a separate ‘bottoms-up’ study, the EMC Corporation, a publicly traded company with 60,000 employees, recently interviewed 497 data scientist and business intelligence professionals from around the world. The results of their study on the need for Data Scientist pointed to some interesting trends in the computing industry. About two-thirds of the individuals polled believe the demand for data scientists will outpace supply in the next five years with nearly 30% coming from professionals in disciplines other than computer science. The study also cited the lack of training and resources as the biggest obstacles to data science in organizations. The EMC study also found that data science professionals were over 2.5 times more likely to have a master’s degree and over 9 times more likely to have a PhD than business intelligence professionals. These observations directly support the case for the need for rigorous scientific training for the professionals moving into the data science field.

The San Diego Economic Development Corporation (EDC) tracks San Diego’s standing among the 25 most populous U.S. metropolitan areas. In an analysis of talent, the region ranked 2nd (51%) in the percent of college graduates with science or engineering majors and 12th (34%) in the percent population with bachelor’s degrees or higher. The San Diego Software Industry Council, the region’s premier professional association for software technology, data analytics and cloud-based application communities, hosted a Data Summit for the last three years with over 300 attendees each year and a speakers list of the foremost thinkers and analysts. In a presentation to the previous San Diego mayor, a group of business, academic, and government leaders shared that over the last 20 years, this region has become home to the nation’s pre-eminent analytics community with the right mix of large, established companies alongside a constellation of over 40 smaller, niche-focused companies that are pushing back the boundaries of analytics and data science applications largely inspired by an unmatched academic resource in and an experienced professional services network to support the community’s growth. We expect to have students in the MAS-DSE drawn from all of these communities. This forms an extraordinarily large pool of students for years to come.

B. Opportunities for Placement of Graduates

The intention of this degree is for the graduates’ development and (possibly) promotion within existing job positions. As a program intended for working engineers, many students will receive support from their companies and will be employed during the program. As a result, the placement of graduates is not a primary concern to this program except that it may improve prospects for advancement and future employment opportunities. For students who are unemployed or interested in alternate job prospects, the Jacobs School Engineering Student Services and the campus Career Service Center will provide assistance in terms of résumé advising, internship opportunities, and jobs listing. The support letters by industry (appended) indicate a strong demand for such training for their employees in this area.

V. Importance to the Discipline

Today, US industry and government agencies recognize that our business and scientific communities are undergoing a profound transformation with the use of large-scale, diverse, and high-resolution data sets that allow for data-intensive decision-making, including clinical decision making, at a level never before imagined. New statistical and mathematical algorithms, prediction techniques, and modeling methods, as well as multidisciplinary approaches to data collection, data analysis and new technologies for sharing data and information are enabling a paradigm shift in government, the marketplace and in research. Advances in machine learning, data mining, and visualization are enabling new ways of extracting useful information in a timely fashion from massive data sets, which complement and extend existing methods
of hypothesis testing and statistical inference. As a result, a number of companies and agencies are
developing big data strategies to align with their missions. This proposal for the graduate program in
Data Science and Engineering is multidisciplinary and focuses on methodologies and tools to develop
next generation systems. It fills an important need that exists today. The proposed MAS program is
concerned with educating students at the intersection of multi-disciplinary engineering and science
disciplines. Because the field is still maturing, this program will provide students with the opportunity
for a state-of-the art interdisciplinary engineering education as well as a rich area for significant and
heretofore largely unexplored challenges.

VI. Ways in Which the Program Will Meet the Needs of Society
The proposed graduate program directly addresses the needs of society by training students in an
emerging field of study that can be immediately applied to interests in the public and private sectors.
Data Science and Engineering is uniquely a value-added endeavor that makes use of existing data for
new markets and science.

VII. Relationship of the Program to Research & Interests of the Faculty
As demonstrated in the faculty biographies in the appendices, the program directly aligns with the
research interests of faculty members participating in the program.

VIII. Program Differentiation
The proposed program does not duplicate any existing program nation-wide. The proposed MAS
program will not interfere with existing departmental programs or any other existing interdisciplinary
programs. First and fundamentally, the MAS program is targeting a student population that is not
currently served by any existing programs. Rather than traditional full-time students who can only take
coursework during the day, these MAS courses will be scheduled when the target audience of working
professionals is available. Second, in contrast to more purely academic pursuits in traditional computer
science programs, the proposed MAS curriculum builds on fundamentals in computer science and
extends tools into a new realm of study that has direct application for the working professional.

Several Institutions (see list in Appendix F) have established masters program in analytics over the past
few years, but only a small fraction are part-time with fewer still on the West Coast. The programs are
extremely varied, with most of the programs geared towards students of business and statistics. Two of
the programs that appear to be relatively similar to ours are the NYU program and the Northwestern
program. The unique strength of the DSE program at UCSD is its emphasis on the database side. Most
programs go to Hadoop but no further. Our proposed classes include graph-based data taught by Alin
Deutsch, a class on online analytics taught by Yannis Papakonstantinou, and also classes on column-
based databases. In addition, a distinctive feature of our program is its collaboration with the San Diego
Supercomputer Center and the ability of the students not only to engage with the technical staff at SDSC
but also to participate in a capstone project that utilizes the unique facilities of the San Diego
Supercomputer Center.

IX. Faculty
The program incorporates the expertise of ladder-rank faculty members who are interested in
participating in the program and technical experts from the San Diego Supercomputer Center. The
capstone course will be managed by Drs. Yoav Freund and Yannis Papakonstantinou, Professors in the
Department of Computer Science, and Chaitanya Baru from the San Diego Supercomputer Center.
The faculty who are engaged in this program are at the forefront of advancing the core scientific and technological means of managing, analyzing, visualizing, and extracting useful information from large, diverse, distributed and heterogeneous data sets. Appendix A includes vitae of participating faculty.

The proposed MAS curriculum is set up so that all participating faculty members teach as off-load instructors at off-hours; this will not impact the ability of faculty to ensure quality delivery of all other departmental programs. All of the faculty members listed for particular courses in the Courses section below have made a commitment to teach in this program (Profs. Elkan, Freund, Dasgupta, Deutch, and Papakonstaninou). As the program evolves, we expect the engagement of additional faculty members from the machine learning, computer vision and database groups. The proposed MAS DSE, then, will draw primarily upon our faculty in the following areas:

- **Machine Learning:** Yoav Freund, Sanjoy Dasgupta, Gary Cottrell, Charles Elkan, Lawrence Saul, and Kamalika Chaudhuri
- **Databases:** Yannis Papakonstantniu, Alin Deutsch, and Victor Vianu
- **Algorithms, Vision, and Bioinformatics:** Mohan Paturi, Russell Impagliazzo, Andrew Kahng, and Vineet Bafna

The department has invested heavily into growing the area of Machine Learning, doubling in size from three to six faculty members over the past six years. Given the number one priority of "Data Sciences" over the past two years and two separate slots available to the faculty (one in intersection with computer systems discussed below), this growth is likely to continue.

With additional recruiting expected across the areas of Big Data, Systems, Bioinformatics and Vision (the department is currently conducting search for five slots across these four areas), we anticipate even greater capacity for teaching our current courses and any anticipated additions to our curriculum. Further, the MAS DSE represents a commitment of 10-12 courses per year -- which could be taught on-load or off-load -- in partnership with San Diego Super Computer Center. SDSC already has identified a core group of six researchers (Chaitan Baru, Amarnath Gupta, Natasha Balac, Paul Rodriguez, Amit Chourasia, and Richard Sinkovits) who have committed to teaching DSE courses. The department expects to carry a 6-9 course workload related to the DSE program, which will be taught entirely off-load for at least the first two years.

Ladder-rank faculty engaged in the program will normally teach one course per year and never more than two courses per year. In addition, the department will be hiring additional faculty in this important area and our partnership with SDSC brings a wealth of expertise and resources that are outstanding and unique, expanding the list of qualified instructors available to teach in the program. Faculty participation in the program is voluntary and those faculty members listed are committed to this teaching opportunity. Growth in the program will bring welcome opportunities for SDSC and CSE professional researchers to participate in the teaching program as well.

Finally, we anticipate that some revenue generated from this proposed program may actually enhance existing graduate programs through hiring of additional Teaching Assistants or even Lecturers to reduce student-to-assistant or student-to-instructor ratios, and/or offer additional courses. The DSE program is expected to provide additional resources and collaborations for our students and faculty.

**X. Courses**

As noted above in VIII. Faculty and on the course approval forms (see Appendix D), a large number CSE and SDSC personnel are able to teach DSE courses, although below we have named only one or two per course as an indication of those particularly interested in teaching it. In some cases the faculty may team teach, but as a general rule one instructor will teach the course in a given quarter.
A. *Foundational Courses*

[DSE 200] Python for Data Analysis (CSE: Yoav Freund, SDSC: Robert Sinkovits)
The goal of this course is to bring students with diverse background and experience to a common level of competency in programming in the context of complex and noisy data. Solid competency in Python programming provides its owner with autonomy and independence in their work. Introduction to object oriented programming using python. Regular expressions. Numpy and Numerical Processing. Ipython and Plotting. Data analysis using PANDAS. Webpage scraping using scrapy. The twitter API. NLTK.

This course will provide an introduction to the management of structured data beginning with an introduction to database models including relational, hierarchical, and network approaches. It will also cover topics in database system implementation including query languages and system architectures; parallel, column-oriented, and array-based database systems; advanced SQL features including user-defined functions (UDFs), triggers, statistical functions; and support for spatial data.


B. *Core Courses*

[DSE 203] Data Integration & ETL (SDSC: Amarnath Gupta)
The course is designed to provide students with the fundamentals of data integration and includes: schema mapping and matching, entity disambiguation, ontology development and management, data provenance, and crowd sourcing and machine learning as strategies for integration. The course will also require hands-on projects in which students will work on a data integration problem requiring integration of two or more datasets taken from an application domain of their choice (e.g. geospatial data, healthcare, financial applications, bioinformatics, etc).

This course provides a broad introduction to the practical side of machine-learning and data analysis. The topics covered in this class include topics in supervised learning, such as k-nearest neighbor classifiers, decision trees, boosting and perceptrons, and topics in unsupervised learning, such as k-means, PCA and Gaussian mixture models.

Map-reduce, streaming analysis, and external memory algorithms and their implementation using the Hadoop and its eco-system: HBase, Hive, Pig and Spark. The class will include assignment of analyzing large existing databases.
[DSE 290] Case Studies in Data Science (Various CSE and SDSC)
Case studies discussed by speakers from industry, government and academia expose students to the needs and uses of different technologies and their roles in model building.

C. Elective Courses

[DSE 221] Data Analysis Using R (SDSC: Natasha Balac)
R, an open source software project with an extensive library of freely available packages and the capability to apply most modern statistical methods, has emerged as a leading statistical computing environment. This course will focus on providing fundamental compute skills necessary for effective data analysis and machine learning tasks by applying modern statistical methods implemented in R. The course covers practical issues in statistical computing including data preparation, manipulation, analysis and the generation of analytical, predictive and graphical results. Topics in statistical data analysis, machine learning and graphics applications will be provided along with practical working examples. Machine learning topics are introduced as needed when addressing real world data mining case studies.

[DSE 232] Performance Measurement (SDSC: Chaitanya Baru)
This course will introduce practical and pragmatic considerations related to the performance of big data solution approaches, cover the fundamentals of computer performance measurement, especially as applied to database and big data systems, and provide an understanding of the primary determinants of performance in big data systems. The course will cover: tools and techniques for performance monitoring and performance tuning; the role of benchmarking and how to interpret benchmark results in context; how to read query plans and perform optimizations and database tuning; behavior of schedulers and governors, including systems like YARN, Mesos, Fair Scheduler, FIFO, HPC schedulers; performance characteristics of data analytics operations; practical limits to performance scaling; and recent results in big data performance and ongoing work in performance optimizations.

[DSE 240] Online Analytics Applications (CSE: Yannis Papakonstantinou, SDSC: Chaitanya Baru)
The course will cover the functionality of online analytics applications from the business analyst point of view; basics of application and data infrastructure architecture; data organizations for systematic data precomputation. The use of data warehouses, data cubes (with emphasis on ROLAP organizations) and materialized views; Fast Data use cases and relevant technologies: combining transactional and analytical databases; incremental maintenance of precomputed views; use of novel database systems: parallel, column & Hadoop/mapreduce-enhanced; visual interfaces & dashboards for custom analytics applications. Application development technologies and methodologies with emphasis on Model-View frameworks facilitating reflection of state on view. Overview of web-based visualization libraries; custom web-based visualizations (D3); achieving online performance with approximations.

[DSE 241] Data Visualization (SDSC: Amit Chourasia)
The goal for the course is to use visualization as a tool to explore trends, relationships, confirm hypothesis, communicate findings and gain insight about data. This course will focus on teaching students the principles and techniques for creating visual representation from raw data. The course exercises will be based on publicly available datasets and utilize freely available tools like
D3.JS and VisIt. The course will be modeled similar to Stanford’s visualization CS448 course and will include an introduction to visualization, vis foundation review, color, interaction, dashboards and Heat Maps, introduction to D3.Js. high dimensional data, network data, geographic data, text data, scientific visualization: isosurface, volume rendering, and introduction to VisIt.

[DSE 250] Beyond Relational Data Models (CSE: Alin Deutsch)
The course covers data models, query languages and models of computation beyond those employed in relational databases. It addresses new developments that have gained attention with the advent of the Web 2.0 and Big Data revolutions. The topics are presented in a unifying framework and include: key-value pairs as data model, as used in Google's Big Table; Object-Oriented Data Model, with its practical support in relational databases via the Object-Relational Mapping (involves ODMG standards ODL and OQL, and recent systems such as Ruby on Rails); semi-structured databases (data organized as graph with labels on nodes and edges), query languages based on reachability constraints between nodes: conjunctive regular path queries); XML databases, as special case of semi-structured databases in which the graph is a tree (this involves associated standards such as XML Schema, XPath and XQuery); RDF databases (with associated OWL and SPARQL standard).

[DSE 251] Managing Large-Scale Graph Data (SDSC: Amarnath Gupta)
Large-scale graphs appear in many diverse applications including the World Wide Web, Social Networks, Human Communication (e.g., phone call graphs, email graphs), Professional Networks (who knows/follows whom), Biological Networks, and Linked Data Graphs. The goal of this course is twofold (a) get students acquainted with data management issues related to graphs, including storage, indexing, querying, and computing with large graph data, and (b) give them a hands-on experience with Neo4j and Gremlin. Prerequisite: Successful completion of DSE250 or written permission of the instructor. The lecture portion of the course will cover: Basic Principles of Graph Data Management, Storage Techniques for Large-Scale Graphs, Indexing Graphs, Query Processing for Graphs, Computing Graph Functions, Special Processing Techniques for Citation Networks, Social Networks, Biological Networks

D. Capstone Course

[DSE 260] Data Science Design Capstone Project (CSE: Yoav Freund, Yannis Papakonstantinou / various faculty, SDSC: Chaitanya Baru/various staff)
A team design project in the final two quarters of the program culminates in a final report and an oral presentation of the capstone project. In addition, there might be a demonstration of the working prototype. The project will start by identifying a domain of interest and the available data sources that will be used to study the domain. From this starting point there will be two parallel and interdependent lines of work: data extraction, Transformation and Loading (ETL), and statistical analysis and model building. The ultimate goal will be to present a processing pipeline which transforms the raw data into more usable forms and models which separates between the predictable and the unpredictable aspects of the underlying system.

XI. Resource Requirements
No new campus resources will be required to support this program as it is completely self-supporting. Based on the experience of administering other self-supporting MAS programs in the Jacobs School,
enrollment is estimated to be approximately 22-26 students the first year and increase to a steady-state of 25-30 students per year for a total of 50-60 enrolled students at any given time.

Our experience with three different MAS programs recently launched indicates a fundamental viability and sustainability of the proposed MAS program that will benefit by the experience and existing infrastructure.

Nationally, there are several self-supporting master’s programs currently being offered at private and public institutions. Stanford’s EE MS degree (15 courses) charges $58,950.1 USC’s online M.S. in Computer Engineering (9 courses) charges $36,720.2 UCLA’s online MS in engineering program (36 units) charges $30,000 per student.3 The UCSD SOM’s MAS in Leadership of Healthcare Organizations (42 units) charges $28,500;4 its MAS in Clinical Research (36 units) charges $28,000.5 The MAS in International Studies offered by International Relations and Pacific Studies (48 units) charges $36,000/student. Students in the proposed UCSD MAS in Structural Health Monitoring (36 units) will be charged $33,000. The above numbers are just program fees, not including health insurance fees ($594/quarter) and registration fees ($324/quarter). As shown in the program’s financial forecast summarized in Table 1 below, the program is viable. The following paragraph presents a summary of the financial analysis.

Because the faculty instructors are already teaching full load, we assume 10-12 instructors of courses (including the case study seminar and the capstone project course) are paid off load at 1.5 months of their salary but averaged at $16,000 per 4-unit course in the MAS program. This average, slightly higher than 1.5 months of a Full Professor, Step II ($14,250), is substantially higher than the cap of $10,000 that Summer Session pays its instructors. In addition, a TA (50% appointment, or 20 hours/week) is assigned for 20-30 students in the class. This is substantially more generous than the current TA allocation in the department (50% TA for 60 students in graduate CSE courses, for example) and is required in order to provide these MAS students the additional high-quality support that they will need to re-enter the academic environment after their hiatus. A 50% TA appointment includes a salary of $5,885/quarter, a fee payment of $5,216.50, and the possibly a non-resident supplemental tuition payment of $5034/quarter, depending on residency status of the TA. There will be one full-time technical support person to manage the cluster and to aid with technical issues such as software installation and configuration, load management and throughput maximization. The IT cost is estimated to be $90,000 and includes technical support as well as a number of non-technical services such as access to the program’s learning management system, related software and hardware support and for videotaping the lectures. Hardware/equipment is projects at $5000 per 12 students. We will reserve 10% of the fees generated for financial aid, $5,000 to $10,000 each, for those who are unemployed or whose tuition fees are not covered by their employers; financial awards will be made beginning with the second cohort, after a reserve has been established. The Jacobs School Dean’s Office charges 8% for administrating the program (staff support, room reservation, receiving payments, troubleshooting student issues, recruiting students, etc.). Budget calculations include the 4% UCSD fee and the 1.5% UCOP assessment fee, which will start with Cohort 3, and the Office of Graduate Studies assessment.

1 http://scpd.stanford.edu/becomeAStudent/tuitionAndFees.jsp
2 http://mapp.usc.edu/mastersprograms/tuitionandfees/tuitionfees.html
3 http://msengrol.seas.ucla.edu/
4 http://lhco.ucsd.edu/prospectivestudents/fees.cfm
5 http://clre.ucsd.edu/prospectivestudents/fees.cfm
Table 1: Summary of Projected Budget

<table>
<thead>
<tr>
<th>Description</th>
<th>Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td># Students in cohort</td>
<td>24</td>
</tr>
<tr>
<td># Courses (7 required, capstone, 2-4 electives)</td>
<td>11</td>
</tr>
<tr>
<td>Instructors’ stipend @ $16,000</td>
<td>$176,000</td>
</tr>
<tr>
<td>TA salary+fees+tuition (50% TA @ 20-30 stu/course)</td>
<td>$177,491</td>
</tr>
<tr>
<td>IT (video, website)</td>
<td>$90,000</td>
</tr>
<tr>
<td>Hardware/equipment</td>
<td>$10,000</td>
</tr>
<tr>
<td>Financial Aid @ 10% of fees collected</td>
<td>$79,200</td>
</tr>
<tr>
<td>Admin fee @ 8% of fees collected</td>
<td>$63,360</td>
</tr>
<tr>
<td><strong>Total Expenses</strong></td>
<td><strong>$596,051</strong></td>
</tr>
<tr>
<td>UCSD 4% assessment</td>
<td>$23,842</td>
</tr>
<tr>
<td>UCOP 1.5% assessment (beginning 3rd yr, based on expenditures 2 years prior)</td>
<td>$5,841</td>
</tr>
<tr>
<td>Office of Graduate Studies assessment ($4,000 + $220/stu)</td>
<td>$9,280</td>
</tr>
<tr>
<td><strong>Total Assessments</strong></td>
<td><strong>$38,963</strong></td>
</tr>
<tr>
<td>Student fees ($33,000/stu)</td>
<td>$792,000</td>
</tr>
<tr>
<td><strong>Net Income to Department</strong></td>
<td><strong>$156,986</strong></td>
</tr>
</tbody>
</table>

XII. Graduate Student Support
Those who are unemployed or whose tuition fees are not covered by their employers are eligible for financial aid of $5,000 to $10,000, based on merit and financial need.

XIII. Changes in Senate regulations
No changes to Senate regulations are proposed.

XIV. Program Governance and Administration
The program is interdisciplinary in nature and involves two units associated with the Jacobs School of Engineering. Initial responsibility for administration will reside in the Department of Computer Science and Engineering which will be the home department for the program. Curriculum oversight of the degree program, as well as responsibility for the academic advising of students, selection of faculty, access to student and faculty services, and program evaluation will be centered in the Department of Computer Science. However, all faculty participating in the program will have equal access and opportunity to recommend modifications and changes to the curriculum and program.

Ultimate authority and responsibility for the program rests with the program directors who are members of the Academic Senate. A steering committee that consists of the MAS program directors, the department chair and the Associate Dean for MAS programs will provide advice to the directors on:

- Student recruitment
- Academic standards and other requirements for continuing students enrollment in the program
- Curriculum oversight
- Advice concerning students who fail to fulfill requirements of the program
- Funds management oversight
- Continuous evaluation guidance
- Events and program-wide activities, discussions and retreat, general oversight etc.
Students will be mentored by faculty involved in the program. Admission into the MAS DSE program will be limited to the number of students who can be supervised by the available and interested faculty. A committee of three will manage student admissions. Two members will be professors from CSE (initially Yoav Freund and Yannis Papakonstantinou) and one member from SDSC (initially Chaitanya Baru). Acceptance into the program will require a majority vote.

Academic advising will be handled through the Department of Computer Science and Engineering, providing support for programmatic advising in terms of courses, degree requirements, University requirements, academic and programmatic policies and procedures, etc. At the same time, we want to ensure that the MAS students are integrated with the rest of the engineering student body and that they benefit from close relationships with the entire Jacobs School student infrastructure. This aspect will be managed by the MAS program office and will operate outside of standard business.

A student services administrator will be designated to serve as a liaison between the University and the MAS students to ensure the unique needs of working adult students are met and to provide guidance to students regarding University procedures, especially those specific to MAS programs. These student service functions include:

- Organizing prospective student application packets, supporting the Admissions Committee, and helping to determine appropriate fees
- Assisting students with enrollment, enforcing deadlines, and ascertaining that students receive ancillary services such as parking stickers, ID cards, and computer registration
- Checking fee waivers and exemption contracts, providing loan packets, explaining financial obligations to the students
- Conveying student comments concerning administration to faculty course planners and making sure that student needs are met
- Ascertaining that students complete courses needed for the degree, assisting them with career transitions or placement interests
- Assistance with visa, housing, insurance and other arrangements for students from outside the San Diego region

Financial management involves setting program fees, planning a budget, monitoring actual vs. projected income, monitoring expenditures, reporting to campus and OP budget and audit offices, and negotiating fee discounts.

Faculty services are an additional element of program administration. Instructor services will be handled directly by the Jacobs School in conjunction with the departmental MSOs and student affairs directors, funded by MAS fees. This function involves reserving and obtaining access to suitable classrooms, scheduling course meetings, ordering textbooks and obtaining audiovisual materials, administering class evaluations and grade reports, preparing occasional instructors for their teaching assignments and negotiating salaries with them.

All program promotion will be done with the close collaboration of the Jacob School’s Corporate Affiliates Program.
XV. Appendices

A. Biography of Participating Faculty
NATASHA BALAC
San Diego Supercomputer Center
University of California, San Diego
nbalac@sdsc.ucsd.edu

A. Education
Vanderbilt University, Nashville, TN Ph.D. 2002, Computer Science
Vanderbilt University, Nashville, TN M.S. 1999, Computer Science
MTSU, Murfreesboro, TN. B.S. 1997, Computer Science

B. Positions and Honors
5/12-present: Director, Predictive Analytics Center of Excellence, San Diego Supercomputer Center (SDSC), University of California San Diego (UCSD), San Diego, CA
10/04 – 5/12: Research Scientist IV Manager, San Diego Supercomputer Center (SDSC)
10/03 – 10/04: Associate Research Scientist III, SDSC, UCSD, San Diego, CA
1/02 – 10/03: Data Mining Consultant, ProStaff, Inc. San Diego, CA
8/00 – 6/01: Software Engineer, RF Microsystems, Nashville, TN
5/00 – 8/00: Software Engineer, RTS Wright Industries Inc., Nashville, TN
5/99 – 8/01: Research Assistant, Vanderbilt University, Nashville, TN

Teaching Positions
- 2003 – present: Instructor, University of California San Diego Extension
  Designed, created and teaching Data Mining certificate program courses.
- 1994-1997: Instructor and assistant for 100 and 300 level computer science courses, Middle Tennessee State University

Honors
- Upsilon Phi Epsilon, Computer Science Honor Society
- Outstanding Young Americans
- Tau Beta Pi, National Honor Society
- Phi Kappa Phi, National Honor Society

Professional Societies
- American Association for Artificial Intelligence
- Association for Computing Machinery, ACMKDD
- Institute of Electrical and Electronic Engineers

C. Selected peer-reviewed publications


Natasha Balac and Daniel M. Gaines. ERA: Knowledge Acquisition in Robotic Domains. ICAPS 2003.

a. Professional Preparation
Indian Institute of Technology, Madras Electronics Engineering B.Tech, 1979
University of Florida Electrical Engineering M.E., 1983
(Outstanding Master’s Thesis Award, Dept. of Electrical Engineering)
University of Florida Electrical Engineering Ph.D., 1985

b. Academic and Professional Appointments
2013 – Associate Director, Data Initiatives, San Diego Supercomputer Center, UCSD.
2011 – Director, Center for Large-scale Data Systems research (CLDS), SDSC.
2007 – Distinguished Scientist, San Diego Supercomputer Center (SDSC), University of California, San Diego.
2004 – 2007 Division Director, Science R&D, SDSC, UC San Diego.
2001 – 2004 Co-Director, Data and Knowledge Systems Program, SDSC, UC San Diego.
1992 – 1996 Advisory Programmer, Database Technology Institute, IBM Almaden Research Center, San Jose, CA, and IBM Toronto Laboratory, North York, Ontario.
1985 – 1992 Assistant Professor, CSE Division, Dept. of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI.
1979 – 1981 Programmer Analyst, Management Services Division, Telco (Tata Motors), Pune, India.

c. Publications

Other Publications


d. Synergistic Activities
1. Program Chair, 4th Workshop on Big Data Benchmarking, IEEE BigData Conference, October 2013, San Jose, CA (with Shane Canon, LBL; Ian Foster, U.Chicago/ANL).
2. Program Chair, Workshop on Big Data in Science, IEEE BigData Conference, October 2013, San Jose, CA (with Tilmann Rabl, U.Toronto; Milind Bhandarkar, Pivotal).
3. Chair, Program / Organizing Committee, Workshops on Big Data Benchmarking. First Workshop: May 8-9, 2012, San Jose, CA; Second Workshop: December 17-18, 2012, Pune, India; Third Workshop: July 16-17, 2013, Xi’an, China.
4. Member, Organizing Committee, 7th Extremely Large Databases Conference, September 9-12, 2013, Palo Alto, CA.

e. Collaborators and Other Affiliations
a. Collaborators and Co-editors
Mark Adams, JCVI; Ramon Arrowsmith, Arizona State; Karen Basen-Enquist, MD Anderson Cancer Center, UT (MDACC); Michael Carey, UC Irvine; Christopher Crosby, UNAVCO; Wendy Demark-Wahnefried, UAB; Deborah Estrin, UCLA; Ian Foster, U.Chicago/Argonne National Labs; Rusty Gage, Salk Institute; Ann Gates, UTEP; Sara Graves, Univ. of Alabama-Huntsville (UAH); Tom Horan, Claremont Graduate University; Randy Keller, U.Oklahoma; Tanu Malik, U.Chicago; Charles Meertens, UNAVCO; Don Middleton, NCAR; Susan Peterson, MDACC; Rahul Ramachandran, UAH; Richard Scheuerman, JCVI; Craig Venter, JCVI.

b. Graduate and Postgraduate Advisors
Ph.D./Master’s Thesis Advisor: Prof. Stanley Y. W. Su, Dept. of Computer Science and Engineering, University of Florida, Gainesville, FL.

c. Thesis Advisor and Postgraduate-Scholar Sponsor
Thesis Advisor
- Prof. Ophir Freider, Robert L. McDevitt, K.S.G., K.C.H.S. and Catherine H. McDevitt L.C.H.S. Chair in Computer Science and Information Processing. Chair, Department of Computer Science, Georgetown University, Washington DC.
- Dr. Piyush Goel, Co-founder, everypath.com, Santa Clara, CA.
- Dr. Sriram Padmanabhan, Distinguished Engineer and Chief Architect for Information Server, IBM Silicon Valley Labs, Santa Teresa, CA.

Postgraduate-Scholar Sponsor
- Dr. Tezeswi Tadepalli, Research Associate, University of Mississippi.
- Dr. Kai Lin, Research Programmer, San Diego Supercomputer Center.
- Dr. Bertram Ludaescher, Professor, Computer Science Dept., UC Davis.
AMIT CHOURASIA

San Diego Supercomputer Center
University of California, San Diego
9500 Gilman Dr., MC 0505
La Jolla, CA - 92093

Telephone: (858) 822-3656
Fax: (858) 534-8303
Email: amit@sdsc.edu

A. Professional Preparation
M.S. in Computer Graphics Technology  2001-2003 Purdue University, West Lafayette, USA
Baccalaureate in Architecture (Hons) 1996-2001 Indian Institute of Technology, Kharagpur, India

B. Appointments
2009 – Present Sr. Visualization Scientist San Diego Supercomputer Center, UCSD
2003 – 2009 Visualization Scientist San Diego Supercomputer Center, UCSD

C. Publications (selected from over 40 publications)
Five most relevant publications

Five other relevant publications

D. Synergistic Activities
• **XSEDE & TeraGrid Participation**
  Lead Visualization Services Group at SDSC
  Provide visualization expertise through XSEDE Advanced User Support Services
• **Conference Organization**
  Visualization Showcase Chair XSEDE 2012
  Publicity Chair VisWeek 2012
  Chair IEEE Visualization Contest (2009-2010), Co-Chair IEEE Visualization Contest 2008
• **Program Committee Member**
  IEEE Visualization Conference
  EuroVis Conference
  International Symposium on Visual Computing
  Volume Graphics Technical Committee

E. Collaborators
**Collaborators**
David Camp (Lawrence Berkeley National Laboratory and Univ. Of California, Davis)
Hank Childs (Lawrence Berkeley National Laboratory and Univ. Of California, Davis)
Yifeng Cui (SDSC/ Univ. of California San Diego)
Steve Day (San Diego State University)
Geoffrey Ely (SIO/ Univ. of California San Diego)
Christoph Garth (Univ of California, Davis)
Thomas Jordan (University of Southern California)
Kenneth Joy (Univ of California, Davis)
Homayoun Karimabadi (Univ. of California San Diego)
Philip Maechling (University of Southern California)
Amitava Majumdar (SDSC/Univ. of California San Diego)
Bernard Minister (IGPP Univ. of California San Diego)
Mike Norman (SDSC/ Univ. of California San Diego)
Kim Olsen (San Diego State University)
Emily Roxworthy (Univ. of California San Diego)
Jamison H. Steidl (Univ. of California Santa Barbara)
P. K. Yeung (Georgia Tech Univ.)

**Graduate Advisors**
Mark Bannatyne (IUPUI), Gary Bertoline (Purdue), Clark Cory (Purdue), Terry Burton (Purdue)

**Graduate Students**
Emmett Mcquinn, Patrick Yau
High-dimensional statistics, clustering, algorithms for finding underlying patterns in high-dimensional data, machine learning

Professor Sanjoy Dasgupta develops algorithms for the statistical analysis of high-dimensional data. Such data is now widespread, in domains ranging from environmental modeling to genomics to web search. The geometry of high-dimensional spaces presents unusual challenges; many traditional statistical procedures were developed with one- or two dimensional data in mind and do not scale well to this modern context. Some of them are very inefficient; others give poor results because of counter-intuitive effects in high dimension. Dasgupta has developed the first provably correct, efficient algorithms for a variety of canonical statistical tasks, especially related to clustering (grouping) data. He is one of the few machine learning researchers whose work combines algorithmic theory with geometry and mathematical statistics. He adds a strong theoretical focus to UCSD's CSE artificial intelligence and bioinformatics groups.

Capsule Biography

Prior to joining the UCSD Jacobs School in 2002, Sanjoy Dasgupta was a senior member of the technical staff at AT&T Labs-Research, where his work focused on algorithms for data mining, with applications to speech recognition and to the analysis of business data. Prof Dasgupta received a Ph.D. in Computer Science in 2000 from UC Berkeley and a B.A. in Computer Science from Harvard in 1993. He is a member of the editorial boards of the Journal of Machine Learning Research, the Journal of Artificial Intelligence.

Research


Semistructured and XML data, data security, adaptive distributed query design, and the design and optimization of query languages. Professor Deutsch's research centers around the XML (Extended Markup Language) and the efficient and effective integration of heterogeneous data from multiple sources. XML is widely accepted as the standard for data exchange between businesses on the Internet. However, most corporations publish only selected portions of their proprietary business data as XML documents, and even then only virtually, by exposing an interface against which XML queries can be formulated. This interface allows the user to query the data without having to know where exactly it is physically stored, or under what form. In order to be answered, user queries must be "reformulated" as queries on the actual proprietary data. The MARS (Mixed and Redundant Storage) System solves the reformulation problem in a very general setting that allows mixed (XML and relational) storage for the proprietary data and takes advantage of redundancies in the storage (materialized views, indexes and caches) that can enhance performance. The underlying reformulation algorithm is guaranteed to find a reformulation if one exists. Moreover, this algorithm achieves optimal complexity bounds under reasonable restrictions on the expressive power of user queries.

Capsule Bio:
Alin Deutsch is interested in XML databases and XML query languages. He received his Ph.D. in computer science from the University of Pennsylvania in 2002. He received his M.S. in computer science from the Technical University of Darmstadt (Germany) in 1995 and his B.S. in computer engineering from the Bucharest Polytechnic University (Romania) in 1993 - graduating with honors from each institution. During his doctoral studies at the University of Pennsylvania, Deutsch worked as an intern at the AT&T Research Labs and Texas Instruments. As a teaching assistant at the University of Pennsylvania, he taught Java programming techniques and a tutorial on XML technologies. Due to his extensive experience, he is able to provide an effective mix of high-level technical detail and real-world interpretations - enabling him to interact with a wide variety of audiences. Deutsch's entrepreneurial skills have led to a patent in 2000 based on query optimization.

Selected Publications:
Google Scholar Publications
Web Page:
http://www.db.ucsd.edu/People/alin/
CHARLES ELKAN

Professor, Computer Science and Engineering
University of California, San Diego
Email: celkan@ucsd.edu
Office Phone: 858-534-8897

Automated reasoning, artificial intelligence, machine learning, database systems, expert systems, computational biology, and data mining. Professor Elkan has expanded his early interest in fuzzy logic to writing learning, search, and reasoning algorithms. He was an early pioneer in Web applications of artificial intelligence, notably in the area of information retrieval (e.g. how search engines process very general or very specific queries to yield the best matches possible). Elkan has developed algorithms for reasoning about database queries and updates, and methods of formalizing commonsense knowledge about causation. He also co-authored a powerful Web engine for comparing proteins and DNA--allowing biologists to detect shared features and evolutionary relationships among the flood of protein and DNA sequence data produced by the Human Genome Project. Elkan is currently co-directing the Knowledge and Data Engineering (KDE) effort within Cal-(IT)2--a broad effort in database and data mining research to support applications involving massive data sets (initially from medical imaging and environmental sensor networks). He has also taken his work in AI and applied it to stock-market trading strategies. Elkan can also talk on how scientists and engineers communicate their ideas to the media and the public, having authored the CSE department's "Notes on Giving a Research Talk."

Capsule Bio:
Charles Elkan joined the UCSD faculty in 1990, after earning his Ph.D. that same year in computer science at Cornell University. He did his undergraduate degree at Cambridge University. Elkan was a postdoctoral fellow at the University of Toronto. In 1998-99, he was a visiting Associate Professor in computer science at Harvard University. While at Harvard, he was Senior Scientist at the software firm Knowledge Stream Partners. Elkan has consulted for Hewlett-Packard, SAIC, Sony, IBM, and Alcoa. He has won numerous best-paper awards, including first-place at the CoIL Challenge 2000 data mining competition. Elkan is the co-founder of UCSD's Artificial Intelligence Laboratory.

Selected Publications:
Google Scholar Publications

Institute Affiliation:
California Institute for Telecommunications and Information Technology

Web Page:
http://www-cse.ucsd.edu/users/elkan/
YOAV FREUND

Education

Hebrew University, Jerusalem  Math and Physics  B.Sc., 1982
Hebrew University, Jerusalem  Computer Science  M.Sc. 1989
University of California, Santa Cruz  Computer Science  Ph.D., 1993

Appointments

2005-present  Professor, Department of Computer Science and Engineering
               University of California, San Diego

2003-2005  Senior Research Scientist in the Center for Computational Learning Systems.
            Columbia University

2001-2003  Senior Research Scientist.
            Banter Inc.

1993-2001  Senior Technical Staff member, Machine Learning and AI group
            Bell Laboratories, later AT&T Labs—Research

1986-1989  Head of software development team for a real-time embedded system.
            TAAS - Israel Armament Development Industries.

1982-1986  Member of technical staff. Image processing and Pattern Recognition Dept.
            RAFAEL - Israeli Defense Research and Development

Five Products relevant to this proposal

Valmianski I, Shih AY, Driscoll JD, Matthews DW, Freund Y and Kleinfeld D. Automatic identification of

Mayank Kabra, (Advisor Yoav Freund) Automated cancer detection and drug discovery: two biomed- ical

Arvey A, Hermann A, Hsia CC, Ie E, Freund Y and McGinnis W. Minimizing off-target signals in RNA
fluorescent in situ hybridization. Nucleic Acids Res. 2010 Jun;38(10)

Paré, A, Lemons, D, Kosman, D, Beaver, W, Freund, Y and McGinnis W. Transcriptional analysis of the
Hox gene Scr at single molecule resolution yields evidence for transcriptional bursting during

Roy Liu, Yoav Freund and Glen Spraggon, Image-based crystal detection: a machine-learning ap-

Other representative products


JBoost: A software package for Boosting. (http://jboost.sourceforge.net/)


Synergistic activities
Action editor, Journal of Machine Learning Research
Work with Dr Stephen Baird, Surgical Pathologist, on computer analysis of pathology images.
Work with Professors McGinnis and Bier, Biologists, on analysis of FISH images of drosophila embryos.
Work with Professors Nguyen and Rao on the smart media lab in CALIT2 / UCSD.
Work with the Image Analysis group in the Broad Institute and the Ruvkun and Altshuler labs in MGH.

Collaborators and advisors

Collaborators (last 4 years)
UCSD: Sanjoy Dasgupta, William McGinnis, Other: Kimmen Sjolander (Berkeley), Glen Spraggon (Novartis Research Foundation).


Graduated Students
German Creamer (2005) – Now a Professor in Stevens Technical Institute
Eugene Ie (2007) – Now works for Pinterest
Evan Ettinger (2010) – Now works for Google
Mayank Kabra (2011) – Now doing a post-doc in Janelia Farm
William Beaver(2012)
AMARNATH GUPTA

San Diego Supercomputer Center
University of California, San Diego
9500 Gilman Drive, MC 0505
La Jolla, CA 92093-0505

Telephone: 858 822-0994
Fax: 858 822-0906
E-mail: gupta@sdsc.edu

Professional Preparation
Indian Institute of Technology, Kharagpur, India
B. Tech. Mechanical Engineering 1984

University of Texas, Arlington
M.S. Biomedical Engineering 1987

Jadavpur University, India
Ph.D. (Engineering) Computer Science 1994

Appointments
2007 Full Research Scientist, San Diego Supercomputer Center
2001 Associate Research Scientist, San Diego Supercomputer Center
1998 Assistant Research Scientist, San Diego Supercomputer Center
1998 Visiting Research Scientist, San Diego Supercomputer Center
1995 Scientist, Virage, Inc., San Mateo, CA
1995 Visiting Researcher, Visual Computing Laboratory, UC San Diego
1994 Sr. IT Engineer, CMC Limited, India
1988 Computer Engineer, Indian Statistical Institute, Calcutta, India

Related Products


Other Significant Products


Synergistic Activities
- A domain-specific ontology based search system – developed as a part of the Neuroscience Information Framework for NIH
- A search system for finding paths and subgraphs in RDF data – being developed in the context of LAMHDI, an ongoing project for NIH
- A data integration system for heterogeneous events – collaborative work with Ramesh Jain, UCI

Collaborators
Chaitan Baru (SDSC), Mark Ellisman (UCSD), Bertram Ludäscher (SDSC), Michael Kifer (SUNY at Stony Brook, NY), Maryann E. Martone (UCSD), Richard Marciano (SDSC), Reagan Moore (SDSC), Yannis Papakonstantinou (UCSD), Victor Vianu (UCSD), Ilya Zaslavsky (SDSC)

Graduate Students and Postdoctoral Researchers
Shabbar Tambawalla, UC San Diego, now in the industry
Paul Nguyen, UC San Diego, now in the industry
Shailendra Bhonsle, UC San Diego, now in the industry
Setareh Rafatirad, UC Irvine, George Mason University
Li Chen, UC San Diego, IBM
Sandeep Gupta, UC San Diego, now in the industry
Hitesh Sabnani, UC San Diego, now in the industry
Mingyan Gao, UC Irvine, Google

Graduate and Postgraduate Advisors
Aditya Bagchi, Indian Statistical Institute, Calcutta, India
George V. Kondraske, Dept. of Electrical Engineering, University of Texas at Arlington
Embedded systems and mobile computing, including the integration of software and hardware to make computers more portable and energy efficient ('power-aware' computing). Professor Gupta is a pioneer in "codesign" of hardware and software for embedded microsystems. He works on new architectures for mobile devices that take into account their constraints: battery life, a small footprint, less memory, and so on. The research goal is to create system architectures that allow mobile computers to manage power more efficiently. Gupta is also an expert on system modeling and design tools. He teaches courses in computer-aided design (CAD) for digital circuits and systems, and his research extends to algorithms for automated design of very large-scale integrated (VLSI) circuits. Gupta is also an expert on adaptive computing architectures that permit, to a greater degree, built-in flexibility for better performance, fault tolerance etc. One current project focuses on how to design systems that can "learn" from the way they are being used, to allow the system to make the most efficient use (of power, for instance).

Capsule Bio:
Rajesh Gupta is QUALCOMM professor in Computer Science & Engineering at UC San Diego, California. He received his B. Tech. in Electrical Engineering from IIT Kanpur in 1984, MS in EECS from UC Berkeley in 1986 and a Ph. D. in Electrical Engineering from Stanford in 1994. Earlier he worked at Intel Corporation, Santa Clara and on the Computer Science faculty at University of Illinois, Urbana-Champaign and UC Irvine. His current research is focused on energy efficient and mobile computing issues in embedded systems. In recent years, Gupta and his students have received a best paper award at IEEE/ACM DCOSS’08 and a best demonstration award at IEEE/ACM IPSN/SPOTS’05. Gupta is a recipient of the Chancellor's Fellow at UC Irvine, UCI Chancellor's Award for excellence in undergraduate research, NSF CAREER Award and achievement awards at Intel. He has served as EIC of IEEE Design & Test of Computers and chair of the steering committee of IEEE Transactions on Mobile Computing. Gupta is a Fellow of the IEEE.

Selected Publications:
Google Scholar Publications

Institute Affiliation:
California Institute for Telecommunications and Information Technology

Web Page:
http://www.cse.ucsd.edu/~gupta
a. Professional Preparation

b. Appointments
2010-present, Director, San Diego Supercomputer Center
2009-2010, Interim Director, San Diego Supercomputer Center
2008-2009, Chief Scientific Officer, San Diego Supercomputer Center
2007-present, Distinguished Professor, UCSD, La Jolla, CA
2000-2007, Professor of Physics, UCSD, La Jolla, CA
1992-2000, Professor of Astronomy, UIUC, Champaign-Urbana, IL
1991-1992, Associate Professor of Astronomy, UIUC, Urbana, IL
1989-2000, Senior Research Scientist, NCSA, Urbana, IL
1986-1989, Research Scientist, NCSA, Urbana, IL
1984-1986, Staff Member, X-Division, LANL, Los Alamos, NM
1980-1984, Staff Member, MPI for Astrophysics, Garching, Germany
1975-1980, Student Employee, B-Division, LLNL, Livermore, CA

c. Products: 5 Most Relevant to Current Project
5. The Enzo Project, http://enzo-project.org. Enzo is a community-developed adaptive mesh refinement simulation code, designed for rich, multi-physics hydrodynamic astrophysical calculations.

Other Products


d. Synergistic Activities
Norman directs the San Diego Supercomputer Center, a leading center for cyberinfrastructure research and development. He also directs the Laboratory for Computational Astrophysics at UCSD which develops the ZEUS, ENZO, and CELLO community codes.
Norman is the principal investigator of the $20M Gordon data-intensive supercomputer project funded by the NSF.
Norman has also been involved in many public outreach projects through videos and films featuring astrophysical simulations, including IMAX films, Planetarium Sky shows, and Nova and Discovery Channel programs.

e. Collaborators and Other Affiliations
a. Collaborators and Co-Editors
Tom Abel (Stanford), Greg Bryan (Columbia), Jack Burns (CU), Eric Hallman (CU), Fabian Heitsch (Munich), Lars Hernquist (Harvard), David Kirkman (UCSD), Alexei Kritsuk (UCSD), Chris Loken (CITA), Mordecai-Mac Low (AMNH), Patrick Motl (LSU), Kentaro Nagamine (UCSD), Ake Nordlund (Copenhagen), Paolo Padoan (UCSD), Alex Razoumov (ORNL), Yoel Rephaeli (Tel Aviv), Meir Shimon (Tel Aviv), Volker Springel (MPA), David Tytler (UCSD), Arthur Wolfe (UCSD), Jason Prochaska (UCSC)
b. Graduate and Postdoctoral Advisors
James R. Wilson (LLNL), Karl-Heinz A. Winkler (LANL)
c. Thesis Advisor and Postgraduate-Scholar Sponsor
Tom Abel (Stanford), Peter Anninos (LLNL), Dinshaw Balsara (Notre Dame), James Bordner (UCSD), Greg Bryan (Columbia), David Clarke (St. Mary's), David Collins (LANL), Greg Daues (NCSA), Robert Fiedler (Cray), John Hayes (LLNL), Tridivesh Jena (Zementis), Pakshing Li (Berkeley), Byung-Il Jun (LLNL), Wen-Ching Lin (Zementis), Henry Neeman (U. Oklahoma), Brian O'Shea (MSU), Alexei Razoumov (St. Marys), James Stone (Princeton), Pascalis Paschos (Northwestern), Daniel Reynolds (SMU), Stephen Skory (CU Boulder), Doug Swesty (SUNY), Dan Whalen (CMU), Hao Xu (LANL), Miguel-Angel Yanez (UCSD), Yu Zhang (Prowess Systems)

Norman has supervised 16 graduate students and 15 postdocs with some overlap.
YANNIS PAPAKONSTANTINOY

Professor
Computer Science and Engineering Department
University of California, San Diego

Office: 3248 Computer Science & Engineering Building
Phone: (858) 8221612
Fax: (858) 5340729
Email: yannis@cs.ucsd.edu

Short Bio
Yannis Papakonstantinou is a Professor of Computer Science and Engineering at the University of California, San Diego. His research is in the intersection of data management technologies and the web, where he has published over eighty research articles. He has given multiple tutorials and invited talks, has served on journal editorial boards and has chaired and participated in program committees for many international conferences and workshops. Yannis enjoys to commercialize his research and to inform his research accordingly. He was the CEO and Chief Scientist of Enosys Software, which built and commercialized an early XML-based Enterprise Information Integration platform. Enosys Software was acquired in 2003 by BEA Systems. His lab’s FORWARD platform (for the rapid development of data-driven Ajax applications) is now in use by many commercial applications. He is involved in data analytics in the pharmaceutical industry and is in the technical advisory board of Brightscope Inc. He is the inventor of seven patents. Yannis holds a Diploma of Electrical Engineering from the National Technical University of Athens, MS and Ph.D. in Computer Science from Stanford University (1997) and an NSF CAREER award for his work on data integration.

Current Research
- Rapid WYSIWYG development of web applications, supported by NSF 0917379 and a Google Research award
- Database-driven Ajax applications, supported by NSF 1018961
- Parallel platform for semi-structured data search and analysis, supported by NSF 0910820
- Inconsistency Resolution in Online Databases, supported by NSF 1117527
- Reengineering Database Systems for Fast SSDs, supported by NSF 1219125
- Low latency browser-based web computation, supported by NSF 1219263
- E-platforms for Patient Empowerment and Population Health Improvement, supported by NSF 1237174
Teaching

- CSE135, Web Application Development (senior undergraduate and MS class, current): The course covers application server essentials, database design and SQL programming with emphasis on the special requirements of web applications, authentication/authorization techniques, the Model-View-Controller methodology and Javascript/Ajax programming. The main part of the course and the class project uses the Java stack. The latter part uses Ruby-on-Rails to teach Object Relational Mapping and Ajax using partial rendering. HTML5 is also discussed.


- CSE87, Computing and the Universe (freshman seminar, current): Imagine a powerful computer that behaves as a human. Is it conscious? Imagine a computer simulating a universe. Could it be our universe? If so, is God ... a programmer?! Ranging from Plato, Descartes and classic sci-fi movies to readings on the singularity, we ponder how computing may shape the future and humanity's approach to big questions.

- CSE21 Mathematics for Algorithms and Systems Analysis
- CSE132B, Database Applications: How to design and build complex applications that use a relational database system
- CSE232B, Database Systems: Advanced Topics & Implementation
PAUL F. RODRIGUEZ, Ph.D.

San Diego Super Comptuer Center
University of California, San Diego
La Jolla, CA. 92037
phone: (858) 534-8326
email: p4rodriguez@ucsd.edu

EDUCATION/POSITIONS

Research Programmer Analyst, University of California, San Diego, 2008-
Associate Specialist, University of California, Irvine, 2004-2008
Postdoctoral Fellow, University of California, Los Angeles, 2003-2004
Postdoctoral Fellow, University of Virginia, Charlottesville, Virginia, 1999-2002
Ph.D. Cognitive Science, University of California, San Diego, 1999
M.S. Computer Science, George Mason University, Fairfax, Virginia, 1993
B.S. Mathematics/Computer Science, University of California, Los Angeles, 1985

SELECT PUBLICATIONS


MAS Data Science and Engineering, Revised January 2014

43


**COURSEWORK**

Advanced Statistical Methods for fMRI Analysis, University of California, Irvine

A broad and technical survey course of experimental design, linear, nonlinear, Bayesian, and multivariate methods for statistical inferencing, connectivity analysis, exploratory analysis, and pattern recognition for fMRI.

**INVITED TALKS**


Mechanisms of Brain and Mind II: Linguistic and non-linguistic communication mechanisms in the brain. Tateshina, Japan. Co-organized by RIKEN Brain Science Center and ATR.

B. Letters of Support
April 15, 2013

Interim Dean Juan Ladner
Jacobs School of Engineering
UC San Diego
9500 Gilman Drive MC 0403
La Jolla, CA 92030-0403

Dear Dean Ladner,

I am writing in support of the proposal by the Department of Computer Science and Engineering in conjunction with the San Diego Supercomputer Center to establish a new program leading to a Master of Advanced Study in Big Data Science. The development of this program could not come at a better time for our industry.

Cubic Transportation Systems is a leading integrator and developer of electronic transit fare collection systems, as well as a provider of other services for intelligent travel solutions. In California, Cubic developed and currently services such systems for the San Francisco Bay Area (MTC, BAART, SFMTA, etc.), the Los Angeles Area (LA Metro, Foothill Transit, LA DOT, etc.) and the San Diego Area (SANDAG, MTS, NCTD).

One of the major benefits of electronic transport payment solutions is the data generated by these systems. There is notable demand from our customers for solutions that help them better understand their data to improve planning for infrastructure and optimize operational processes. Cubic envisions a future where travelers will be given intelligent travel advice based on advance data analytics.

As a consequence, there is an urgent and growing need for engineers who can work effectively in a multidisciplinary environment who are literate in the science of big data and the importance of storing, managing, retrieving and presenting this data to decision makers. Our engineers need continued training in the advanced skills and techniques required for this rapidly changing new field. The proposed program addresses the development of the essential skills required for engineers working in Data Science systems in my company.
Interim Dean Juan Lasheras  
Jacobs School of Engineering  
UC San Diego  
9500 Gilman Drive MC 0403  
La Jolla, CA 92037-0403

Dear Dean Lasheras,

I am writing in support of the proposal by the Department of Computer Science and Engineering in conjunction with the San Diego Supercomputer Center to establish a new program leading to a Master of Advanced Study in Big Data Science. The development of this program could not come at a better time for our industry.

There is an urgent and growing need for engineers who can work effectively in a multi-disciplinary environment who are literate in the science of big data and the importance of storing, managing, retrieving and presenting this data to decision makers. The proposed program addresses the development of essential skills required for engineers working in Data Science systems in my company.

Opera Solutions prides on maintaining its technical leadership and competitive position in our industry and one of the most important components of these efforts is the continuous education of our workforce. The MAS in Big Data Science program proposed would help us not only address the shortage of talent in this new field but contribute to the available pool of talent for this expanding industry. We were pleased to be able to provide ideas and inputs to the curriculum and applaud the interdisciplinary nature of this program and its relevance to our future skill needs.

Big Data analytics and management is a core technological need for our company and the proposed program will be most attractive for recruiting employees. I look forward to the participation of our employees in this valuable program as I believe it will give us a competitive advantage in this rapidly developing technology sector.

Sincerely,

Joseph Milana, PhD  
Chief Scientist, Global Head of Analytics.
Interim Dean Juan Lasheras  
Jacobs School of Engineering  
UC San Diego  
9500 Gilman Drive MC 0403  
La Jolla, CA 92037-0403  

Dear Dean Lasheras,

I am writing in support of the proposal by the Department of Computer Science and Engineering in conjunction with the San Diego Supercomputer Center to establish a new program leading to a Master of Advanced Study in Big Data Science. The development of this program could not come at a better time for our industry.

Named to the Computerworld Top 100 "Best Places to Work IT", the Kaiser Permanente Information Technology (IT) organization is the team that powers the health and IT leadership of Kaiser Permanente. Partnering with KP business units, regions and physician groups through portfolios of business, IT provides the information technology that is combined with KP's integrated health model to provide real-time and personalized health care that is smart, networked, collaborative, affordable, and preventive.

In the next few years, Kaiser Permanente will see dramatic growth in the quality and quantity of IT-enabled services the organization delivers to members. With health care reform, we will also likely see many more members. To serve the organization's growing needs, KP has increased the capacity of our data centers, which are the foundation of all the services IT provides to clinicians, members, and employees.

There is a growing need for engineers who can work effectively in a multi-disciplinary environment who are literate in the science of big data and the importance of storing, managing, retrieving and presenting this data to decision makers. Our engineers need continued training in the advanced skills and techniques required for this rapidly changing new field. The proposed program addresses the development of the essential skills required for engineers working in Data Science systems in my company.

Kaiser Permanente strives to maintain its technical leadership and competitive position in our industry and one of the most important components of these efforts is the continuous education of our workforce. The MAS in Big Data Science program proposed would help us not only address the shortage of talent in this new field but contribute to the available pool of talent for this expanding industry. We were pleased to be able to provide ideas and inputs to the curriculum and applaud the interdisciplinary nature of this program and its relevance to our future skill needs.

Big Data analytics and management is a core technological need for our company and the proposed program will be most attractive for our employees seeking to advance their career and our company.

Sincerely,

Scott McGuirk  
Director, National University Relations Recruitment  
Kaiser Permanente
April 17, 2013

Interim Dean Juan Lasheras
Jacobs School of Engineering
University of California, San Diego
9500 Gilman Drive MC 0403
La Jolla, CA 92093-0403

Dear Dean Lasheras,

I am writing in support of the proposal by the Department of Computer Science & Engineering and the San Diego Supercomputer Center to establish a new program leading to a Master of Advanced Study degree in Big Data Science at UCSD. The development of this program will be of clear benefit to the financial services industry.

Encore's industry leadership position in the areas of analytics and information technology depends critically on investment in education and training. The proposed MAS in Big Data Science program will address the shortage of talent in this growing field and contribute to the available pool of talent for our company and the financial services industry. We were pleased to be able to provide guidance about the curriculum, and applaud the interdisciplinary nature of this program and its relevance to our future business needs.

There is an urgent and growing need for data analysts, scientists, and engineers capable of high quality, multidisciplinary work based on Big Data scientific tools and technologies. Encore's analytic teams require leading edge training in the advanced skills and techniques required for these important computational tasks. The proposed program addresses the development of these essential skills, required for quantitative professionals working on complex, Big Data problems at my company.

The ability to make decisions through the application of Big Data analytics is a clear and growing need for our company. I look forward to the participation of our employees in this valuable program and believe it will significantly enhance our competitive strengths.

Sincerely,

Christopher Trepel, Ph.D.
Senior Vice President and Chief Scientific Officer
Interim Dean Juan Lasheras
Jacobs School of Engineering
UC San Diego
9500 Gilman Drive MC 0403
La Jolla, CA 92030-0403

Dear Dean Lasheras,

I am writing in support of the proposal by the Department of Computer Science and Engineering in conjunction with the San Diego Supercomputer Center to establish a new program leading to a Master of Advanced Study in Big Data Science. The development of this program could not come at a better time for our industry.

There is an urgent and growing need for engineers who can work effectively in a multi-disciplinary environment who are literate in the science of big data and the importance of storing, managing, retrieving and presenting this data to decision makers. Our engineers need continued training in the advanced skills and techniques required for this rapidly changing new field. The proposed program addresses the development of the essential skills required for engineers working in Data Science systems in my company.

ProvideCommerce strives to maintain its technical leadership and competitive position in our industry and one of the most important components of these efforts is the continuous education of our workforce. The MAS in Big Data Science program proposed would help us not only address the shortage of talent in this new field but contribute to the available pool of talent for this expanding industry. We were pleased to be able to provide ideas and inputs to the curriculum and applaud the interdisciplinary nature of this program and its relevance to our future skill needs.

Big Data analytics and managements is a core technological need for our company and the proposed program will be most attractive for our employees seeking to advance their career and our company. I look forward to the participation of our employees in this valuable program as I believe it will give us a competitive advantage in this rapidly developing technology sector.

Sincerely,

Hani Yassin
Vice President, IT Business Services
Interim Dean Juan Lasheras  
Jacobs School of Engineering  
UC San Diego  
9500 Gilman Drive MC 0403  
La Jolla, CA 92093-0403

Dear Dean Lasheras,

I am writing in support of the proposal by the Department of Computer Science and Engineering in conjunction with the San Diego Supercomputer Center to establish a new program leading to a Master of Advanced Study in Big Data Science. I have been active in this field for the past 20 years and it continues to grow in importance.

This field requires people with solid technical skills in what are currently cross discipline areas, including mathematics, statistics, computer science, physics, neuroscience and behavioral psychology. The program being constructed by UCSD is well designed to meet the goals of these needs.

ID Analytics continue to hire people with graduate degrees in technical fields who have these skills. We have done well in hiring several people out of various UCSD programs, and this proposed program will produce more candidates that will be attractive to us. More importantly, I can attest to the explosive nature of this field, and companies will likely be facing a shortage of people with this background over the next decade.

I can give you a personal story to underscore this. In the past year both of my children have entered into this fast growing field. My daughter with an undergrad degree in math is now a data scientist, and my son, with a new PhD in high energy physics, just got hired to be a data modeler. Their careers paths are very positive because of this fast growing field.

This UCSD Masters program is well designed and will produce people that are very attractive to hiring companies in this expanding field.

Sincerely,

[Signature]

Stephen Coggeshall  
Chief Technology Officer, ID Analytics
April 5, 2013

Interim Dean Juan Lasheras  
Jacobs School of Engineering  
UC San Diego  
9500 Gilman Drive MC 0403  
La Jolla, CA 92037-0403

Dear Dean Lasheras,

I am writing in support of the proposal by the Department of Computer Science and Engineering in conjunction with the San Diego Supercomputer Center to establish a new program leading to a Master of Advanced Study in Big Data Science. The development of this program could not come at a better time for our industry.

There is an urgent and growing need for engineers who can work effectively in a multi-disciplinary environment who are literate in the science of big data and the importance of storing, managing, retrieving and presenting this data to decision makers. Our engineers need continued training in the advanced skills and techniques required for this dynamic new field. The proposed program addresses the development of essential skills for engineers working in Data Science systems in my company.

San Diego Gas and Electric strives to maintain its technical leadership and competitive position in our industry. One of the most important components of these efforts is the continuous education of our workforce. The MAS in Big Data Science program proposed would help us not only address the shortage of talent in this new field but contribute to the available pool of talent for this expanding industry. We were pleased to be able to provide ideas and inputs to the curriculum and applaud the interdisciplinary nature of this program and its relevance to our future skill needs.

Big Data analytics and management is a core technological need for our company and the proposed program will be most attractive for our employees seeking to advance their career and our company. I look forward to the participation of our employees in this valuable program as I believe it will give us a competitive advantage in this rapidly developing technology sector.

Sincerely,

Chris Baker
Qualcomm, Inc.

5775 Morehouse Drive
San Diego, CA 92121
United States

Interim Dean Juan Lasheras
Jacobs School of Engineering
UC San Diego
9500 Gilman Drive MC 0403
La Jolla, CA 92037-0403

Dear Dean Lasheras,

I am writing in support of the proposal by the Department of Computer Science and Engineering in conjunction with the San Diego Supercomputer Center to establish a new program leading to a Master of Advanced Study in Big Data Science. The development of this program could not come at a better time for our industry.

There is an urgent and growing need for engineers who can work effectively in a multi-disciplinary environment who are literate in the science of big data and the importance of storing, managing, retrieving and presenting this data to decision makers. Our engineers need continued training in the advanced skills and techniques required for this rapidly changing new field. The proposed program addresses the development of the essential skills required for engineers working in Data Science systems in my company.

Qualcomm strives to maintain its technical leadership and competitive position in our industry and one of the most important components of these efforts is the continuous education of our workforce. The MAS in Big Data Science program proposed would help us not only address the shortage of talent in this new field but contribute to the available pool of talent for this expanding industry. We were pleased to be able to provide ideas and inputs to the curriculum and applaud the interdisciplinary nature of this program and its relevance to our future skill needs.

Big Data analytics and management is a core technological need for our company and the proposed program will be most attractive for our employees seeking to advance their career and our company. I look forward to the participation of our employees in this valuable program as I believe it will give us a competitive advantage in this rapidly developing technology sector.

Sincerely,

Erik Miller
Staffing Manager
UCSD Lead Recruiter
Interim Dean Juan Lasheras  
Jacobs School of Engineering  
UC San Diego  
9500 Gilman Drive MC 0403  
La Jolla, CA 92030-0403

Dear Dean Lasheras,

I am writing in support of the proposal by the Department of Computer Science and Engineering in conjunction with the San Diego Supercomputer Center to establish a new program leading to a Master of Advanced Study in Big Data Science. The development of this program could not come at a better time for our industry.

There is an urgent and growing need for engineers who can work effectively in a multi-disciplinary environment who are literate in the science of big data and the importance of storing, managing, retrieving and presenting this data to decision makers. They need continued training in the advanced skills and techniques required for this rapidly changing new field. The proposed program addresses the development of the essential skills required for engineers working in Data Science systems in my company.

As the world's number one independent provider of data integration software, Informatica strives to maintain its technical leadership and competitive position in our industry. The MAS in Big Data Science program proposed would help us not only address the shortage of talent in this new field but contribute to the available pool of talent for this expanding industry. We were pleased to be able to provide ideas and inputs to the curriculum and applaud the interdisciplinary nature of this program and its relevance to our future skill needs.
March 29, 2013

Interim Dean Juan Lasheras  
Jacobs School of Engineering  
UC San Diego  
9500 Gilman Drive MC 0403  
La Jolla, CA 92093-0403  

Dear Dean Lasheras,

I am writing in support of the proposal by the Department of Computer Science and Engineering in conjunction with the San Diego Supercomputer Center to establish a new program leading to a Master of Advanced Study in Big Data Science. The development of this program could not come at a better time for our industry and for Mitchell International.

There is an urgent and growing need for engineers who can work effectively in a multi-disciplinary environment who are literate in the science of big data and the importance of storing, managing, retrieving and presenting this data to decision makers. Our engineers need continued training in the advanced skills and techniques required for this rapidly changing new field. The proposed program addresses the development of the essential skills required for engineers working in Data Science systems in my company.

Mitchell strives to maintain its technical leadership and competitive position in our industry and one of the most important components of these efforts is the continuous education of our workforce. The MAS in Big Data Science program proposed would help us not only address the shortage of talent in this new field but contribute to the available pool of talent for this expanding industry. We were pleased to be able to provide ideas and inputs to the curriculum and applaud the interdisciplinary nature of this program and its relevance to our future skill needs.

Big Data analytics and management is a core technological need for our company and the proposed program will be most attractive for our employees seeking to advance their career and our company. I look forward to the participation of our employees in this valuable program as I believe it will give us a competitive advantage in this rapidly developing technology sector. Furthermore, I look forward to hiring graduates from this program in order to optimize our development work in the area of Big Data.

Sincerely,

Jack Farnan  
Senior Vice President, Human Resources
**C. Information Required by CPEC**

1. **Name of Program**: Jacobs School of Engineering, Department of Computer Science & Engineering, Master of Advance Study in Data Science and Engineering

2. **Campus**: University of California, San Diego

3. **Degree/Certificate**: Master of Advance Study (MAS) in Data Science and Engineering

4. **CIP Classification** (to be completed by Office of the President):

5. **Date to be started**: Fall 2014

6. **If modification of existing program, identify that program and explain changes**.

7. **Purpose (academic or professional training) and distinctive features (how does this program differ from others, if any, offered in California?)**: The goal is to provide an opportunity for training in the emerging area of Data Science to working professionals drawn from government and industry. Within the UC system, this would be a first program of this kind (a recently launched Masters in Information and Data Science is entirely on-line). However we understand that related efforts are in progress at other campuses given the currency of the topic and its need. At the time of this writing, we believe that the only master’s level program in analytics currently offered in California is through the College of Arts and Sciences, University of San Francisco. In addition, only two programs (Stanford and UCSD) offer certificates in Data Mining.

8. **Type(s) of students to be served**: Prospective students for this program include mid-career professional engineers from businesses and government agencies.

9. **If program is not in current campus academic plan give reason for proposing program now**: N/A

10. **If program requires approval of a licensure board, what is the status of such approval?** No licensure board approval is required.

11. **Please list special features of the program (credit for experience, internships, lab requirements, unit requirements, etc.)**

    The MAS degree brings together the expertise of faculty in the department of Computer Science and Engineering with the talent of the San Diego Supercomputer Center to address an important need of industry and government. The program is geared toward working professionals and prepares them for leadership in the growing field of Data Science. The program employs a capstone project that will emphasize team engineering in topics of relevance to the students’ employment.

12. **List all new courses required**:

    Department, Course Number, Title, Hours/Week Lecture Lab.

    Department of Computer Science and Engineering; all classes offered on an alternating Friday and Saturday schedule during Fall, Winter and Spring Quarters.

    **Foundation and Case Study Courses**
    - DSE 200 Python for Data Analysis
    - DSE 201 Data Management Systems
    - DSE 210 Statistics and Probability Using Python
Core Courses
- DSE 203 Data Integration & ETL
- DSE 220 Machine Learning
- DSE 230 Data Analysis Using Hadoop, and Spark
- BSD 290 Case Studies in Data Science

Elective Courses (Subset of 2-4 offered each year)
- DSE 221 Data Analysis Using R
- DSE 232 Performance Measurement
- DSE 240 Online Analytics Applications
- DSE 241 Data Visualization
- DSE 250 Beyond Relational Data Models
- DSE 251 Managing Large-Scale Graph Data

Capstone Course
- DSE 260 Data Science Design

13. List all other required courses:
Department, Course Number, Title, Hours/Week Lecture Lab.
See above.

14. List UC campuses and other California institutions, public or private, which now offer or plan to offer this program or closely related programs: University of San Francisco offers a full time M.S. in Analytics, requiring 35 credit hours, at a cost of approximately $39,550. The program began in 2012. In 2013, UC Berkeley launched an on-line Masters of Information and Data Science program which requires 27 units, at a cost of $2,222/unit.

15. List any related program offered by the proposing institution and explain relationship.
Currently only a certificate program in data mining is offered through University Extension. In addition, an undergraduate program in a related area of databases and machine learning is offered by the Department of Computer Science and Engineering at UC San Diego.

16. Summarize employment prospects for graduates of the proposed program. Give results of job market survey if such have been made.
The primary purpose of this degree is not to improve prospects for employment since the enrollees already possess basic relevant knowledge. That said however, the degree will no doubt improve the employment opportunities for the students who complete the program. The MAS degree will help students be more productive, effective and achieve greater influence and leadership within their current organization.

The McKinsey Global Institute’s June 2011 report, “Big data: The next frontier for innovation, competition, and productivity”, projects the U.S. will need 140,000-190,000 predictive analysts and 1.5 million managers/analysts by 2018. While a local market survey has not been conducted, faculty teaching in the program have been asked by industry and government entities to provide supplemental short courses to their employees in this area. Further we have been received 10 (with more expected) letters of support from local companies expressing interest sending employees to participate in the program.

17. Give estimated enrollment for the first 5 years and state basis for estimate.
Enrollment for the first year of the program is estimated to be ~24 professional engineers. This estimate is based on the lengthy review process required to seek approval for a new degree program and our ability to market a new program within a very short time period. As the program becomes known in the general community, enrollment is expected to grow to a total enrollment of 50-60 students at steady state.

18. Give estimates of the additional cost of the program by year for 5 years in each of the following categories: FTE Faculty, Library Acquisitions, Computing, Other Facilities, Equipment. Provide brief explanation of any of the
costs where necessary.
The program is completely self-funding. No need for additional faculty FTE is anticipated since resources generated from the program will pay directly for faculty teaching in the program. No additional library acquisitions, computing, facilities or equipment is anticipated that is not already accounted for in the program fee structure.

19. How and by what agencies will the program be evaluated?
The MAS degree program will be evaluated every eight years by the Academic Senate at UCSD. In addition, surveys will be conducted of students participating in the program in terms of the quality of teaching and program delivered. Surveys by graduates of the program are also anticipated.
D. Course Approval Forms

UNIVERSITY OF CALIFORNIA, SAN DIEGO
REQUEST FOR COURSE APPROVAL

- New Course
- Reinstatement
- Deletion
- Renumbering: old number
- Summer Session Only
- Change in Course

Nature of change:

Effective Qt. Yr.
Fall 2014

<table>
<thead>
<tr>
<th>Subject &amp; Num.</th>
<th>Units</th>
<th>Title: Python for Data Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSE 200</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hours Per Week Expected of Student</th>
<th>Lec</th>
<th>Sem</th>
<th>Dis</th>
<th>Lab</th>
<th>Studio</th>
<th>Tut</th>
<th>Practicum</th>
<th>PE Act</th>
<th>Med Clerk</th>
<th>Outside Prep</th>
<th>Other(describe)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

If the course has multiple discussion or other sections, how should the grade reports be printed (check one)?
- Single List of all Students
- By Dis Section
- By Lab Section
- By Studio Section
- By Tut Section

- Grading: Undergraduate
- Standard Grading (letter or P/NP)
- P/NP Only
- Grading: Graduate and SOM
- Standard Option (Graduate)
- S/U Permitted (Graduate)
- S/U Only
- IPL/FSOM Core only

May be taken for credit [1] time(s). If more than once, justify:
- Final Exam Given
- If not, explain:

COURSE DESCRIPTION (In concise catalog description style, 40 word limit)

Prerequisites: MAS Data Science and Engineering major

ENFORCEMENT (List prerequisites and other restrictions to be enforced by computer (see instructions).
Prerequisites that must be completed: Enrolled in MAS Data Science and Engineering major
Prerequisites that may be concurrent:
Corequisites (must be concurrent):
Other restrictions

Special course characteristics. Check all boxes that apply and see instructions for required explanations:
- Use of animals
- Use of computer resources
- IP Grading
- Cross listed with
- Conjoined with

Instructor and title: Dr. Yoav Freund, Professor, CSE; Dr. Robert Sinkovits, SDSC

JUSTIFICATION:
This course has been created as part of the curriculum for the Masters in Advanced Sciences in Data Science and Engineering.

Department Chair          date

Registrar                 date

APPROVALS-GRADUATE COURSE | APPROVALS-UNDERGRADUATE COURSE

- Dean, School of Medicine date
- Dean of Graduate Studies date
- Council of Provosts date
- CEP Subcommittee on Courses date

Extent of approval: □ Indefinite □ Summer Only Expires at the end of Quarter, 19

PO 2873 (REV. 7/80) Approved for Use by CEP on November 17, 1994
## UNIVERSITY OF CALIFORNIA, SAN DIEGO
### REQUEST FOR COURSE APPROVAL

<table>
<thead>
<tr>
<th>Subject &amp; Num.</th>
<th>Units</th>
<th>Title: Data Management Systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSE 201</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

### Hours Per Week Expected of Student
- Lec: 3
- Sem: 0
- Dis: 0
- Lab: 0
- Studio: 0
- Tut: 0
- Practicum: 0
- PB Act: 0
- Med Clerk: 0
- Outside Prep: 0
- Other (describe): 0

### If the course has multiple discussion or other sections, how should the grade reports be printed (check one)?
- [ ] Single List of all Students
- [X] By Dis Section
- [ ] By Lab Section
- [ ] By Studio Section
- [ ] By Tut Section

### Grading-Undergraduate
- [ ] Standard Grading (letter or P/NP)
- [X] P/NP Only

### Grading-Graduate and SOM
- [ ] Standard Option (Graduate)
- [ ] S/U Permitted (Graduate)
- [X] S/U Only
- [ ] HEP/FYSOM Core only

### May be taken for credit [1] times(s). If more than once, justify:
- [X] Final Exam Given. If not, explain:

### COURSE DESCRIPTION (In concise catalog description style, 40 word limit)

Introduction to structured data and database systems. Covers topics in database system implementation including query languages and system architectures; parallel, column-oriented, and array-based database systems; advanced SQL features including user-defined functions (UDFs), triggers, statistical functions; support for spatial data.

### Enforced Prerequisites:

Enrollment in MAE Data Science and Engineering major, DSE 200

### Enforcement (List prerequisites and other restrictions to be enforced by computer (see instructions). Prerequisites that must be completed: Enrollment in MAE Data Science and Engineering major

### Prerequisites that may be concurrent:

Corerequisites (must be concurrent): Other restrictions:

### Special course characteristics. Check all boxes that apply and see instructions for required explanations.
- [X] Use of animals
- [X] Use of computer resources
- [X] IP Grading
- [ ] Cross listed with
- [ ] Conjoined with

### Instructor and title:

Dr. Yannis Papakonstantinou, Professor, CSE; Dr. Chaitanya Baru, SDSC

### JUSTIFICATION:

This course has been created as part of the curriculum for the Masters in Advanced Sciences in Data Science and Engineering.

### Department Chair date Registrar date

### APPROVALS-GRADUATE COURSE

- Dean, School of Medicine date
- Dean of Graduate Studies date
- Graduate Council date

### APPROVALS-UNDERGRADUATE COURSE

- Council of Provosts date
- CEP Subcommittee on Courses date

### Extent of approval: [ ] Indefinite [X] Summer Only Expires at the end of Quarter, 19

---

FO 2075 (REV. 7/90) Approved for Use by CEP on November 17, 1994

---

MAS Data Science and Engineering, Revised January 2014
# UNIVERSITY OF CALIFORNIA, SAN DIEGO

## REQUEST FOR COURSE APPROVAL

<table>
<thead>
<tr>
<th>Subject &amp; Num.:</th>
<th>DSE 203</th>
<th>Units:</th>
<th>4</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Title:</th>
<th>Data Integration and ETL</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Hours Per Week</th>
<th>Lec</th>
<th>Sen</th>
<th>Dis</th>
<th>Lab</th>
<th>Studio</th>
<th>Tut</th>
<th>Practicum</th>
<th>PB Act</th>
<th>Med Clerk</th>
<th>Outside Prep</th>
<th>Other (describe)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9</td>
<td></td>
</tr>
</tbody>
</table>

If the course has multiple discussion or other sections, how should the grade reports be printed (check one)?

- [ ] Single List of all Students
- [ ] By Dis Section
- [ ] By Lab Section
- [ ] By Studio Section
- [ ] By Tut Section
- [ ] By PB Act
- [ ] By Med Clerk
- [ ] By Outside Prep
- [ ] Other (describe)

Grading-Undergraduate:  
- [ ] Standard Grading (letter or P/NP)
- [ ] P/NP Only

Grading-Graduate:  
- [ ] Standard Option (Graduate)
- [ ] S/U Permitted (Graduate)
- [ ] S/U Only
- [ ] HUP/F (SOM Core only)

May be taken for credit [1] times. If more than once, justify:

- [ ] No Final Exam Given. If not, explain:

Project

COURSE DESCRIPTION (In concise catalog description style; 40 word limit)

The course is designed to provide students with the fundamentals of data integration: schema mapping and matching, entity disambiguation, ontology development and management, data provenance, crowd sourcing. Project required: Data integration problem requiring integration of two or more data sets taken from an application domain of the student's choice.

Prerequisites:  
- Enrollment in MAS Data Science and Engineering major; DSE 200 and DSE 201

ENFORCEMENT: List prerequisites and other restrictions to be enforced by computer (see instructions).

Prerequisites that must be completed:  
- Enrollment in MAS Data Science and Engineering major; DSE 200 and DSE 201

Prerequisites that may be concurrent:

Corequisites (must be concurrent):

Other restrictions:

Special course characteristics:  
- [ ] Use of animals
- [ ] Use of computer resources
- [ ] IP Grading
- [ ] Cross listed with
- [ ] Conjoined with

Instructor and title:  
- CSE: Staff, SDSC: Dr. Anamnath Gupta

JUSTIFICATION:

This course has been created as part of the curriculum for the Masters in Advanced Sciences in Data Science and Engineering.

\[\text{Signature}\]

Department Chair  

Registrar  

APPROVALS (GRADUATE COURSE)  

Dean, School of Medicine  

Dean of Graduate Studies  

Graduate Council  

Council of Provosts  

CEP Subcommittee on Courses  

Extent of approval:  
- [ ] Indefinite
- [ ] Summer Only

Expires at the end of Quarter, 19

FO 2075 (REV. 7/90) Approved for Use by CEP on November 17, 1994
### UNIVERSITY OF CALIFORNIA, SAN DIEGO

#### REQUEST FOR COURSE APPROVAL

<table>
<thead>
<tr>
<th>Subject &amp; Num.: DSE 210</th>
<th>Units: 4</th>
<th>Title: Probability and Statistics Using Python</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hours Per Week Expected of Student</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lab</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Sem</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lab</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Studio</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tut</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Practicum</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PE Act</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Med Clerk</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outside Prep</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Other (describe)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

If the course has multiple discussion or other sections, how should the grade reports be printed (check one)?

- [X] Single List of all Students
- [ ] By Dis Section
- [ ] By Lab Section
- [ ] By Studio Section
- [ ] By Tut Section

Grading: Undergraduate

- [ ] Standard Grading (letter or P/NP)
- [ ] P/NP Only

Grading: Graduate and SOM

- [X] Standard Option (Graduate)
- [ ] S/U Permitted (Graduate)
- [ ] S/U Only
- [ ] H/P/FR(SOM Core only)

May be taken for credit [1] time(s). If more than once, justify:

- [X] Final Exam Given. If not, explain:

**COURSE DESCRIPTION** (In concise catalog description style, 40 word limit)


**Prerequisites:** Enrollment in MAS Data Science and Engineering major; DSE 200

**ENFORCEMENT** (List prerequisites and other restrictions to be enforced by computer (see instructions).

Prerequisites that must be completed: **Enrollment in MAS Data Science and Engineering major; DSE 200**

Prerequisites that may be concurrent:

Corequisites (must be concurrent):

Other restrictions:

Special course characteristics. Check all boxes that apply and see instructions for required explanations.

- [ ] Use of animals
- [ ] Use of computer resources
- [ ] IP Grading
- [ ] Cross listed with
- [ ] Conjoined with

**Instructor and title:** Dr. Sanjoy Dasgupta, Professor, CSE; Dr. Natasha Balac and Dr. Paul Rodriguez, SDSC

**JUSTIFICATION:**

This course has been created as part of the curriculum for the Masters in Advanced Sciences in Data Science and Engineering.

**Signature**

Department Chair date Registrar date

---

### APPROVALS-GRADUATE COURSE

<table>
<thead>
<tr>
<th>Dean, School of Medicine</th>
<th>date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dean of Graduate Studies</td>
<td>date</td>
</tr>
</tbody>
</table>

### APPROVALS-UNDERGRADUATE COURSE

<table>
<thead>
<tr>
<th>Graduate Council</th>
<th>date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Council of Provosts</td>
<td>date</td>
</tr>
</tbody>
</table>

**Extent of approval:**

- [ ] Indefinite
- [ ] Summer Only

Expires at the end of Quarter, 19
UNIVERSITY OF CALIFORNIA, SAN DIEGO
REQUEST FOR COURSE APPROVAL

Subject & Num.: **DSE 220**  Units: 4  Title: **Machine Learning**

<table>
<thead>
<tr>
<th>Hours Per Week</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Lesson</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seminar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discussion</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lab</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Studio</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tutor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Practicum</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PE Act</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Med Clerk</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outside Prep</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9</td>
</tr>
<tr>
<td>Other (describe)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

If the course has multiple discussion or other sections, how should the grade reports be printed (check one)?
- Single List of all Students
- By Dis Section
- By Lab Section
- By Studio Section
- By Tut Section

Grading-Undergraduate
- Standard Grading (letter or P/NP)
- P/NP Only

Grading-Graduate and SOM
- Standard Option (Graduate)
- S/U Permitted (Graduate)
- S/U Only
- H/P/F/SOM Core only

May be taken for credit [1] time(s). If more than once, justify:
- Final Exam Given
- If not, explain:

COURSE DESCRIPTION (In concise catalog description style, 40 word limit)
A broad introduction to machine-learning. Topics include supervised learning: k-nearest neighbor classifiers, decision trees, boosting and perceptrons. Unsupervised learning, k-means, PCA, Gaussian mixture models, Random Projection Trees.

Prerequisites: Enrollment in MAS Data Science and Engineering major; DSE 210

ENFORCEMENT (List prerequisites and other restrictions to be enforced by computer (see instructions).
Prerequisites that must be completed: Enrollment in MAS Data Science and Engineering major, DSE 210
Prerequisites that may be concurrent:
Corequisites (must be concurrent):
Other restrictions:

Special course characteristics. Check all boxes that apply and see instructions for required explanations.
- Use of animals
- Use of computer resources
- IP Grading
- Credit listed with
- Conjoined with

Instructor and title: Dr. Charles Elkan, Professor, CSE; Dr. Natasha Balci, SDSC

JUSTIFICATION:
This course has been created as part of the curriculum for the Masters in Advanced Sciences in Data Science and Engineering.

Rojer Singh

Department Chair                  date                  Registrar                  date

APPROVALS-GRADUATE COURSE        APPROVALS-UNDERGRADUATE COURSE

Dean, School of Medicine                  date  

Dean of Graduate Studies                  date

Graduate Council                  date

Council of Provosts                   date

CEP Subcommittee on Courses                  date

Extent of approval:  
- Indefinite
- Summer Only

Expires at the end of Quarter, 19
UNIVERSITY OF CALIFORNIA, SAN DIEGO
REQUEST FOR COURSE APPROVAL

Subject & Num.: DSE 221  Units: 4  Title: Data Analysis Using R

<table>
<thead>
<tr>
<th>Lec</th>
<th>Sem</th>
<th>Dis</th>
<th>Lab</th>
<th>Studio</th>
<th>Tut</th>
<th>Practicum</th>
<th>PE Act</th>
<th>Med Clerk</th>
<th>Outside Prep</th>
<th>Other (describe)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9</td>
<td></td>
</tr>
</tbody>
</table>

If the course has multiple discussion or other sections, how should the grade reports be printed (check one)?
- ☐ Single List of all Students
- ☐ By Dis Section
- ☐ By Lab Section
- ☐ By Studio Section
- ☐ By Tut Section

Grading-Undergraduate: ☐ Standard Grading (letter or P/NP)  ☐ P/NP Only
Grading-Graduate and SOM: ☐ Standard Option (Graduate)  ☐ S/U Permitted (Graduate)  ☐ S/U Only  ☐ H/P/F/FSOM Core only

May be taken for credit [1] time(s). If more than once, justify:
- ☐ Final Exam Given.  If not, explain:

COURSE DESCRIPTION (In concise catalog description style, 40 word limit)
Fundamental computer skills necessary for effective data analysis and machine learning tasks, applying modern statistical methods implemented in R. Practical issues in statistical computing, including data preparation, manipulation, analysis and the generation of analytical, predictive and graphical results.

Prerequisites: Enrollment in MAS Data Science and Engineering major; DSE 210

ENFORCEMENT (List prerequisites and other restrictions to be enforced by computer see instructions):
Prerequisites that must be completed: Enrollment in MAS Data Science and Engineering major; DSE 210
Prerequisites that may be concurrent:
Corquisites (must be concurrent):
Other restrictions:
- ☐ Use of animals  ☐ Use of computer resources  ☐ IP Grading  ☐ Cross listed with  ☐ Conjoined with

Instructor and title: STAFF, CSE; Dr. Natasha Balac, SDSC

JUSTIFICATION:
This course has been created as part of the curriculum for the Masters in Advanced Sciences in Data Science and Engineering.

[Signature]

Department Chair  date  Registrar  date

<table>
<thead>
<tr>
<th>APPROVALS-GRADUATE COURSE</th>
<th>APPROVALS-UNDERGRADUATE COURSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dean, School of Medicine</td>
<td>date</td>
</tr>
<tr>
<td>date</td>
<td></td>
</tr>
<tr>
<td>Deon of Graduate Studies</td>
<td>date</td>
</tr>
<tr>
<td>date</td>
<td>Council of Provosts</td>
</tr>
<tr>
<td>date</td>
<td></td>
</tr>
<tr>
<td>Graduate Council</td>
<td>date</td>
</tr>
<tr>
<td>date</td>
<td>CEP Subcommittee on Courses</td>
</tr>
<tr>
<td>date</td>
<td></td>
</tr>
<tr>
<td>Extent of approval: ☐ Indefinite  ☐ Summer Only</td>
<td>Expires at the end of Quarter, 19</td>
</tr>
</tbody>
</table>

FO 2073 (REV. 7/90) Approved for Use by CEP on November 17, 1994
### UNIVERSITY OF CALIFORNIA, SAN DIEGO

**REQUEST FOR COURSE APPROVAL**

<table>
<thead>
<tr>
<th>Subject &amp; Num.</th>
<th>DSE 230</th>
<th>Units</th>
<th>4</th>
<th>Title: <strong>Data Analysis Using Hadoop and Spark</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Lec</td>
<td>3</td>
<td>Sen</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dis</td>
<td></td>
<td>Lab</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tut</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Practicum</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PB Act</td>
<td></td>
<td>Med Clerk</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outside Prep</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>(describe)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Effective Qtr Yr:</strong></td>
<td>Fall 2014</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**hours Per Week Expected of Student**

<table>
<thead>
<tr>
<th>Grade</th>
<th>Undergraduate</th>
<th>Standard Grading (Letter or P/NP)</th>
<th>P/NP Only</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade</td>
<td>Graduate and SOM</td>
<td>Standard Option (Graduate)</td>
<td>S/U Permitted (Graduate)</td>
</tr>
</tbody>
</table>

**If the course has multiple discussion or other sections, how should the grade reports be printed (check one)?**

| Single List of all Students | By Dis Section | By Lab Section | By Studio Section | By Tut Section |

**Grading-Undergraduate**

**May be taken for credit [1] time(s). If more than once, justify:**

**Final Exam Given:**

**If not, explain:**

**COURSE DESCRIPTION** (In concise catalog description style; 40 word limit)

- Map-reduce, streaming analysis, and external memory algorithms and their implementation using the Hadoop and its ecosystem: HBase, Hive, Pig and Spark. The class will include assignment of analyzing large existing databases.

**Prerequisites:**

- MAS Data Science and Engineering major, DSE 201 and DSE 210

**ENFORCEMENT** (List prerequisites and other restrictions to be enforced by computer (see instructions)).

**Enrolled in MAS Data Science and Engineering major, DSE 201 and DSE 210**

**Prerequisites that may be concurrent:**

**Corequisites (must be concurrent):**

**Other restrictions:**

**Special course characteristics. Check all boxes that apply and see instructions for required explanations.**

- Use of animals
- Use of computer resources
- IP Grading
- Cross listed with
- Conjoined with

**Instructor and title:** Dr. Youn Freund, Professor, CSE; Dr. Paul Rodriguez, SDSC

**JUSTIFICATION:**

This course has been created as part of the curriculum for the Masters in Advanced Sciences in Data Science and Engineering.

**Department Chair**

**Registrar**

**APPROVALS-GRADUATE COURSE**

<table>
<thead>
<tr>
<th>Dean, School of Medicine</th>
<th>date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dean of Graduate Studies</td>
<td>date</td>
</tr>
<tr>
<td>Graduate Council</td>
<td>date</td>
</tr>
</tbody>
</table>

**APPROVALS-UNDERGRADUATE COURSE**

<table>
<thead>
<tr>
<th>Council of Provosts</th>
<th>date</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEP Subcommittee on Courses</td>
<td>date</td>
</tr>
</tbody>
</table>

**Extent of approval:**

- Indefinite
- Summer Only

Expires at the end of Quarter, 19

**FO 2075 (REV. 7/90) Approved for Use by CEP on November 17, 1994**

---

**MAS Data Science and Engineering, Revised January 2014**

65
UNIVERSITY OF CALIFORNIA, SAN DIEGO
REQUEST FOR COURSE APPROVAL

Subject & Num.: DSE 232  Units: 4  Title: Performance Measurement

If the course has multiple discussion or other sections, how should the grade reports be printed (check one)?
☐ Single List of all Students  ☐ By Dis Section  ☐ By Lab Section  ☐ By Studio Section  ☐ By Tut Section

Grading-Undergraduate  ☐ Standard Grading (letter or P/NP)  ☐ P/NP Only
Grading-Graduate and SOM  ☐ Standard Option (Graduate)  ☐ S/U Permitted (Graduate)  ☐ S/U Only  ☐ HP/F (SOM Core only)

May be taken for credit [1] time(s). If more than once, justify:
☐ Final Exam Given.  If not, explain:

COURSE DESCRIPTION (in concise catalog description style, 40 word limit)
Introduction to topics in measurement and prediction of performance of big data systems. Fundamentals of performance measurement as applied to database and big data systems; coverage of the primary determinants of performance in big data systems, including cost.

Prerequisites: Enrollment in MAS Data Science and Engineering major; DSE 230

ENFORCEMENT (List prerequisites and other restrictions to be enforced by computer (see instructions).
Prerequisites that must be completed: Enrollment in MAS Data Science and Engineering major, DSE 230
Prerequisites that may be concurrent:
Corequisites (must be concurrent):
Other restrictions:

Special course characteristics. Check all boxes that apply and see instructions for required explanations.
☐ Use of animals  Use of computer resources  ☐ IP Grading  ☐ Cross listed with  ☐ Conjoined with

Instructor and title: STAFF, CSE; Dr. Chaitanya Baru, SDSC

JUSTIFICATION:
This course has been created as part of the curriculum for the Masters in Advanced Sciences in Data Science and Engineering.

Robert

Department Chair  date  Registrar  date

APPROVALS-GRADUATE COURSE  APPROVALS-UNDERGRADUATE COURSE

Dean, School of Medicine  date

Dean of Graduate Studies  date  Council of Provosts  date

Graduate Council  date  CEP Subcommittee on Courses  date

Extent of approval: ☐ Indefinite  ☐ Summer Only  Expires at the end of  Quarter, 19

FO 2075 (REV. 7/90) Approved for Use by CEP on November 17, 1994
UNIVERSITY OF CALIFORNIA, SAN DIEGO
REQUEST FOR COURSE APPROVAL

- New Course
- Reinstatement
- Deletion
- Renumbering: old number
- Summer Session Only
- Change In Course
- Nature of change:
- Effective Qt Yr: Fall 2014

### Subject & Num.: DSE 240  |  Units: 4  |  Title: **Online Analytics Applications**

<table>
<thead>
<tr>
<th>Hours Per Week</th>
<th>Lec</th>
<th>Sem</th>
<th>Dis</th>
<th>Lab</th>
<th>Studio</th>
<th>Tut</th>
<th>Practicum</th>
<th>Pr Essential</th>
<th>PE Act</th>
<th>Med Clerk</th>
<th>Outside Prep</th>
<th>Other(describe)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected of Student</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

If the course has multiple discussion or other sections, how should the grade reports be printed (check one)?
- [ ] Single List of all Students
- [ ] By Dis Section
- [ ] By Lab Section
- [ ] By Studio Section
- [ ] By Tut Section

- [ ] Grading-Undergraduate
- [ ] Standard Grading (letter or P/NP)
- [ ] P/NP Only

- [ ] Grading-Graduate and SOM
- [ ] Standard Option (Graduate)
- [ ] S/U Permitted (Graduate)
- [ ] S/U Only
- [ ] HEP/F(SOM Core only)

May be taken for credit [1] time(s). If more than once, justify:
- [ ] Final Exam Given
  - If not, explain:

COURSE DESCRIPTION (In concise catalog description style, 40 word limit)


Prerequisites: Enrollment in MAS Data Science and Engineering major; DSE 201 and DSE 230

ENFORCEMENT (List prerequisites and other restrictions to be enforced by computer (see instructions))

Prerequisites that must be completed: Enrollment in MAS Data Science and Engineering major; DSE 201 and DSE 230

Prerequisites that may be concurrent:
- [ ] Corequisites
  - [ ] Corequisites (must be concurrent):
  - [ ] Other restrictions:

Special course characteristics. Check all boxes that apply and see instructions for required explanations.
- [ ] Use of animals
- [ ] Use of computer resources
- [ ] IP Grading
- [ ] Cross listed with
- [ ] Conjoined with

Instructor and title: GSE: Yannis Papakonstantinou, SDSC: Chaitanya Baru

JUSTIFICATION:

This course has been created as part of the curriculum for the Masters in Advanced Sciences in Data Science and Engineering.

*Signed*

Department Chair  date  Registrar  date

<table>
<thead>
<tr>
<th>APPROVALS-GRADUATE COURSE</th>
<th>APPROVALS-UNDERGRADUATE COURSE</th>
</tr>
</thead>
</table>

- Dean, School of Medicine  date
- [ ] Deans of Graduate Studies  Council of Provosts  date
- [ ] Graduate Council  date  CEP Subcommittee on Courses  date

Extent of approval:
- [ ] Indefinite
- [ ] Summer Only

Expires at the end of Quarter, 19

FO 2075 (REV. 7/90) Approved for Use by CEP on November 17, 1994
UNIVERSITY OF CALIFORNIA, SAN DIEGO
REQUEST FOR COURSE APPROVAL

Subject & Num.: DSE 241  Units: 4  Title: Data Visualization

<table>
<thead>
<tr>
<th>Lec</th>
<th>Sem</th>
<th>Dis</th>
<th>Lab</th>
<th>Studio</th>
<th>Tut</th>
<th>Practical</th>
<th>PB Act</th>
<th>Med Clerk</th>
<th>Outside Prep</th>
<th>Other(describe)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

If the course has multiple discussion or other sections, how should the grade reports be printed (check one)?
- [ ] Single List of all Students
- [ ] By Dis Section
- [ ] By Lab Section
- [ ] By Studio Section
- [ ] By Tut Section

Grading-Undergraduate
- [ ] Standard Grading (letter or P/NP)
- [ ] P/NP Only

Grading-Graduate and SOM
- [ ] Standard Option (Graduate)
- [ ] S/U Permitted (Graduate)
- [ ] S/U Only
- [ ] PUB/PYSOM Core only

May be taken for credit [1] times(s). If more than once, justify:
- [X] Final Exam Given. If not, explain

COURSE DESCRIPTION (In concise catalog description style, 40 word limit)
The use of visualization as a tool to explore trends, relationships, confirm hypothesis, communicate findings and gain insight about data. Principles and techniques for creating visual representation from raw data. Exercises will be based on publicly available datasets and utilize freely available tools like D3.js and VisIt.

Prerequisites: Enrollment in MAS Data Science and Engineering major, DSE 240

Enrollment in MAS Data Science and Engineering major, DSE 240

Pre-requisites must be completed:

Enrollment in MAS Data Science and Engineering major, DSE 240

Pre-requisites that may be concurrent

Corequisites (must be concurrent):

Other restrictions:

Special course characteristics. Check all boxes that apply and see instructions for required explanations.
- [ ] Use of animals
- [ ] Use of computer resources
- [ ] IP Grading
- [ ] Cross listed with
- [ ] Conjoined with

Instructor and Title: STAFF, CSE. Dr. Amit Chourasia, SDSC

JUSTIFICATION:
This course has been created as part of the curriculum for the Masters in Advanced Sciences in Data Science and Engineering.

[Signature]

Department Chair  date  Registrar  date

APPROVALS-GRADUATE COURSE  APPROVALS-UNDERGRADUATE COURSE

Dean, School of Medicine  date

Dean of Graduate Studies  date  Council of Provosts  date

Graduate Council  date  CEP Subcommittee on Courses  date

Extent of approval:  [ ] Indefinite  [ ] Summer Only  Expires at the end of Quarter, 19

FO 2075 (REV. 7/90) Approved for Use by CEP on November 17, 1994
UNIVERSITY OF CALIFORNIA, SAN DIEGO
REQUEST FOR COURSE APPROVAL

- New Course    - Reinstatement    - Deletion    - Renumbering: old number _______ Summer Session Only
- Change in Course    - Nature of change:

<table>
<thead>
<tr>
<th>Subject &amp; Num.: DSE 250</th>
<th>Units: 4</th>
<th>Title: Beyond Relational Data Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lec 3</td>
<td>Sem</td>
<td>Dis</td>
</tr>
<tr>
<td>Hours Per Week Expected of Student:</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- If the course has multiple discussion or other sections, how should the grade reports be printed (check one)?
  - [ ] Single List of all Students  - [ ] By Dis Section  - [ ] By Lab Section  - [ ] By Studio Section  - [ ] By Tut Section

- Grading-Undergraduate:  - [ ] Standard Grading (letter or P/NP)  - [ ] P/NP Only
- Grading-Graduate and SOM:  - [ ] Standard Option (Graduate)  - [ ] S/U Permitted (Graduate)  - [ ] S/U Only  - [ ] Elective (SOM Core only)

May be taken for credit [1] times(s). If more than once, justify:

- [X] Final Exam Given. If not, explain:

COURSE DESCRIPTION (In concise catalog description style, 40 word limit)

Data models, query languages and models of computation beyond those employed in relational databases. New developments that have gained attention since the advent of the Web 2.0 and Big Data revolutions. Topics are presented in a unifying framework.

Prerequisites: Enrollment in MAS Data Science and Engineering major; DSE 201

ENFORCEMENT (List prerequisites and other restrictions to be enforced by computer (see instructions)).
Prerequisites that must be completed: Enrollment in MAS Data Science and Engineering major, DSE 201
Prerequisites that may be concurrent:
Corequisites (must be concurrent):
Other restrictions:

- [ ] Use of animals  - [ ] Use of computer resources  - [ ] IP Grading  - [ ] Cross listed with  - [ ] Conjoined with

Instructor and title. Dr. Alin Deutsch, Professor, CSE

JUSTIFICATION:

This course has been created as part of the curriculum for the Masters in Advanced Sciences in Data Science and Engineering.

[Signature]

Department Chair date Registrar date

APPROVALS-GRADUATE COURSE

- [ ] Dean, School of Medicine date
- [ ] Dean of Graduate Studies date
- [ ] Graduate Council date

APPROVALS-UNDERGRADUATE COURSE

- [ ] Council of Provosts date
- [ ] CEP Subcommittee on Courses date

Extent of approval:  - [ ] Indefinite  - [ ] Summer Only  Expires at the end of Quarter, 19

FO 2075 (REV. 7/90) Approved for Use by CEP on November 17, 1994
UNIVERSITY OF CALIFORNIA, SAN DIEGO
REQUEST FOR COURSE APPROVAL

Subject & Num.: DSE 251  Units: 4  Title: Managing Large-Scale Graph Data

<table>
<thead>
<tr>
<th>Hours Per Week</th>
<th>Lec</th>
<th>Sem</th>
<th>Dis</th>
<th>Lab</th>
<th>Studio</th>
<th>Tut</th>
<th>Practicum</th>
<th>PB Act</th>
<th>Med Clerk</th>
<th>Outside Prep</th>
<th>Other(describe)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

If the course has multiple discussion or other sections, how should the grade reports be printed (check one)?
- Single List of all Students
- By Dis Section
- By Lab Section
- By Studio Section
- By Tut Section

Grading (Undergraduate) □ Standard Grading (letter or P/NP) □ P/NP Only

Grading (Graduate and SOM) □ Standard Option (Graduate) □ S/U Permitted (Graduate) □ S/U Only □ HP/FSOM Core only

May be taken for credit [1] time(s). If more than once, justify.

[No] Final Exam Given. If not, explain. Project as described below.

COURSE DESCRIPTION (In concise catalog description style, 40 word limit)
The goal of this course is twofold: acquaint students with data management issues related to graphs (including storage, indexing, querying, and computing with large graph data) and give them a hands-on experience with Neo4j and Gremlin. Project: Students complete a fully hands-on project whereby they build a new application or a new extension to the software system described above.

Prerequisites: Enrollment in MAS Data Science and Engineering major; DSE 250

ENFORCEMENT (List prerequisites and other restrictions to be enforced by computer (see instructions)).

Prerequisites that must be completed: Enrollment in MAS Data Science and Engineering major, DSE 250

Prerequisites that may be concurrent:

Corequisites (must be concurrent):

Other restrictions:

Special course characteristics. Check all boxes that apply and see instructions for required explanations.
- Use of animals
- Use of computer resources
- IP Grading
- Cross listed with
- Conjoined with

Instructor and title: STAFF, CSE; Dr. Amarnath Gupta, SDSC

JUSTIFICATION:
This course has been created as part of the curriculum for the Masters in Advanced Sciences in Data Science and Engineering.

Department Chair

Registrar

APPROVALS-GRADUATE COURSE

Dean, School of Medicine

Date

Dean of Graduate Studies

Date

Graduate Council

Date

APPROVALS-UNDERGRADUATE COURSE

Council of Provosts

Date

CEP Subcommittee on Courses

Date

Extent of approval: □ Indefinite □ Summer Only Expires at the end of Quarter, 19

FO 2075 (REV. 7/90) Approved for Use by CEP on November 17, 1994
### UNIVERSITY OF CALIFORNIA, SAN DIEGO
#### REQUEST FOR COURSE APPROVAL

- **New Course**
- **Reinstatement**
- **Deletion**
- **Renumbering:** old number
- **Summer Session Only**
- **Change in Course**
- **Nature of change:**
- **Effective Qtr Yr:**

<table>
<thead>
<tr>
<th>Subject &amp; Num.</th>
<th>Units</th>
<th>Title: Data Science Capstone Design Project</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSE 260</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hours Per Week</th>
<th>Expected of Student</th>
<th>Lect</th>
<th>Sem</th>
<th>Dis</th>
<th>Lab</th>
<th>Studio</th>
<th>Tut</th>
<th>Practicum</th>
<th>PE Act</th>
<th>Med Clerk</th>
<th>Outside Prep</th>
<th>Other (describe)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

If the course has multiple discussion or other sections, how should the grade reports be printed (check one)?
- ☑ Single List of all Students
- ☐ By Dis Section
- ☐ By Lab Section
- ☐ By Studio Section
- ☐ By Tut Section

<table>
<thead>
<tr>
<th>Grading-Undergraduate</th>
<th>☐ Standard Grading (letter or P/NP)</th>
<th>☐ P/NP Only</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grading-Graduate and SOM</td>
<td>☑ Standard Option (Graduate)</td>
<td>☐ S/U Permitted (Graduate)</td>
</tr>
<tr>
<td></td>
<td>☐ S/U Only ☐ P/F/SOM Core only</td>
<td></td>
</tr>
</tbody>
</table>

May be taken for credit [ ] time(s). If more than once, justify:
( ) Final Exam Given. If not, explain: Project.

**COURSE DESCRIPTION** (In concise catalog description style, 40 word limit)
A team design project in the final two quarters of the MAS Data Science and Engineering program culminates in the Spring with a final report and an oral presentation of the capstone project. In addition, there might be a demonstration of the working prototype.

**Prerequisites:**

**ENFORCEMENT** (List prerequisites and other restrictions to be enforced by computer (use instructions):)
Prerequisites that must be completed: DSE 200, 201, 210, 203, 220, 230
Prerequisites that may be concurrent: Corequisites (must be concurrent):
Other restrictions:

Special course characteristics. Check all boxes that apply and see instructions for required explanations:
- ☑ Use of animals
- ☑ Use of computer resources
- ☐ IP Grading
- ☐ Cross listed with
- ☐ Conjoined with

Instructor and title: Professors Yoav Freund, Yannis Papakonstantinou, Professors, CSE; SDSC; Chaitanya Baru

**JUSTIFICATION:**
This course has been created as part of the curriculum for the Masters in Advanced Sciences in Data Science and Engineering.

---

Department Chair: [Signature]

Registrar: [Signature]

**APPROVALS-GRADUATE COURSE**

Dean, School of Medicine: [Signature] [Date]

Dean of Graduate Studies: [Signature] [Date]

Graduate Council: [Signature] [Date]

Council of Provosts: [Signature] [Date]

CEP Subcommittee on Courses: [Signature] [Date]

Extent of approval: ☑ Indefinite ☐ Summer Only Expires at the end of Quarter, 19

---

PO 2073 (REV. 7/86) Approved for Use by CEP on November 17, 1994
UNIVERSITY OF CALIFORNIA, SAN DIEGO
REQUEST FOR COURSE APPROVAL

- New Course
- Reinstatement
- Deletion
- Remumbering: old number
- Summer Session Only
- Change in Course

Nature of change:
Effective Qtr Yr: Fall 2014

<table>
<thead>
<tr>
<th>Subject &amp; Num.</th>
<th>DSE 290</th>
<th>Units: 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title:</td>
<td>Case Studies in Data Science</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hours Per Week</th>
<th>Lec</th>
<th>Sem</th>
<th>Dis</th>
<th>Lab</th>
<th>Studio</th>
<th>Tut</th>
<th>Practicum</th>
<th>FE Act</th>
<th>Med Clerk</th>
<th>Outside Prep</th>
<th>Other(describe)</th>
</tr>
</thead>
</table>

If the course has multiple discussion sections, how should the grades be reported? (Check one):
- [ ] By Dis Section
- [ ] By Lab Section
- [ ] By Studio Section
- [ ] By Tut Section

Grading-Undergraduate:
- [ ] Standard Grading (letter or P/NP)
- [ ] P/NP Only

Grading-Graduate and SOM:
- [ ] Standard Option (Graduate)
- [ ] S/U Permitted (Graduate)
- [ ] S/U Only
- [ ] HP/F(SOM Core only)

May be taken for credit: [ ] time(s). If more than one, justify:
- [ ] Final Exam Given. If not, explain:

COURSE DESCRIPTION (in concise catalog description style, 40 word limit!)
Case studies discussed by speakers from industry, government and academia expose students to the needs and uses of different technologies and their roles in model building.

Prequisites: Enrollement in MAS Data Science and Engineering major.

ENFORCEMENT (List prerequisites and other restrictions to be enforced by computer (see instructions).

Prerequisites that must be completed: Enrollment in MAS Data Science and Engineering.

Prerequisites that may be concurrent:
Corequisites (must be concurrent):
Other restrictions:

Special course characteristics: Check all boxes that apply and see instructions for required explanations.
- Use of animals
- Use of computer resources
- Computer grade
- Cross listed with
- Conjoined with

Instructor and title: CSE and SDSC Various Faculty

JUSTIFICATION:
This course has been created as part of the curriculum for the Masters in Advanced Sciences in Data Science and Engineering.

[Signature]
Department Chair date Registrar date

APPROVALS-GRADUATE COURSE

Dean, School of Medicine date

Dean of Graduate Studies date

Graduate Council date

APPROVALS-UNDERGRADUATE COURSE

Council of Provosts date

CEP Subcommittee on Courses date

Extent of approval: [ ] Indefinite [ ] Summer Only Expires at the end of Quarter, 19
E. Catalog Copy

Master of Advanced Studies

The Department of Computer Science and Engineering offers the master of advanced studies (MAS) degree in Data Science and Engineering. The degree requires thirty-eight units of work including a capstone team project. For full-time students with an adequate background in engineering and science, all requirements can be completed within six quarters of part-time study.

Final Project Capstone Requirement, No Thesis

In the Data Science and Engineering program, an “alternative plan” requirement is satisfied by a four-unit capstone project requirement.

Required Courses

Students entering the MAS program for a degree in Data Science and Engineering will undertake courses in programming, analysis, and applications management and visualization. This program requires three foundational courses, four core courses, and two electives totaling thirty-four units, plus a capstone team project course of four units, for a total of thirty-eight units.

Foundational Courses

- DSE 200 Python for Data Analysis
- DSE 201 SQL Database Management Systems
- DSE 210 Statistics and Probability Using Python

Core Courses

- DSE 203 Data Integration & ETL
- DSE 220 Machine Learning
- DSE 230 Data Analysis Using Hadoop, and Spark
- DSE 290 Case Studies in Data Science

Elective Courses

A subset of these courses will be offered each year.

- DSE 221 Data Analysis Using R
- DSE 232 Performance Measurement
- DSE 240 Online Analytics Applications
- DSE 241 Data Visualization
- DSE 250 Beyond Relational Data Models
- DSE 251 Managing Large-Scale Graph Data

Capstone Course

- DSE 260 Data Science Capstone Design Project

Capstone Project

Students are required to execute a team project with an adviser while enrolled in DSE 260. The project will require a combination of in-class, laboratory, and off-campus work. It culminates with a final report and an oral presentation of the capstone project. In addition, there might be a demonstration of the working prototype.
Courses

DSE 200. Python for Data Analysis (4)
Prerequisites: MAS Data Science and Engineering major

DSE 201. Data Management Systems (4)
Introduction to structured data and database systems. Covers topics in database system implementation including query languages and system architectures; parallel, column-oriented, and array-based database systems; advanced SQL features including user-defined functions (UDFs), triggers, statistical functions; support for spatial data.
Prerequisites: MAS Data Science and Engineering major; DSE 200

DSE 203. Data Integration & ETL (4)
The course is designed to provide students with the fundamentals of data integration: schema mapping and matching, entity disambiguation, ontology development and management, data provenance, crowd sourcing. Project required: Data integration problem requiring integration of two or more data sets taken from an application domain of the student’s choice.
Prerequisites: MAS Data Science and Engineering major; DSE 201

DSE 210. Statistics and Probability Using Python (4)
Prerequisites: MAS Data Science and Engineering major; DSE 200

DSE 220. Machine Learning (4)
A broad introduction to machine-learning. Topics include supervised learning: k-nearest neighbor classifiers, decision trees, boosting and perceptrons. Unsupervised learning, k-means, PCA, Gaussian mixture models, Random Projection Trees.
Prerequisites: MAS Data Science and Engineering major; DSE 210

DSE 221. Data Analysis Using R (4)
Fundamental compute skills necessary for effective data analysis and machine learning tasks, applying modern statistical methods implemented in R. Practical issues in statistical computing, including data preparation, manipulation, analysis and the generation of analytical, predictive and graphical results.
Prerequisites: MAS Data Science and Engineering major; DSE 220

DSE 230. Data Analysis Using Hadoop, and Spark (4)
Map-reduce, streaming analysis, and external memory algorithms and their implementation using the Hadoop and it’s eco-system: HBase, Hive, Pig and Spark. The class will include assignment of analyzing large existing databases.
Prerequisites: MAS Data Science and Engineering, DSE 201 and DSE 210
DSE 232. Performance Measurement (4)
Introduction to topics in measurement and prediction of performance of big data systems. Fundamentals of performance measurement as applied to database and big data systems; coverage of the primary determinants of performance in big data systems, including cost.
Prerequisites: MAS Data Science and Engineering major; DSE 230

DSE 240. Online Analytics Applications (4)
Prerequisites: MAS Data Science and Engineering major; DSE 201

DSE 241. Data Visualization (4)
The use of visualization as a tool to explore trends, relationships, confirm hypothesis, communicate findings and gain insight about data. Principles and techniques for creating visual representation from raw data. Exercises will be based on publicly available datasets and utilize freely available tools like D3.JS and VisIt.
Prerequisites: Enrollment in MAS Big Data major, DSE 240

DSE 250. Beyond Relational Data Models (4)
Data models, query languages and models of computation beyond those employed in relational databases. New developments that have gained attention since the advent of the Web 2.0 and Big Data revolutions. Topics are presented in a unifying framework.
Prerequisites: MAS Data Science and Engineering major; DSE 201 and DSE 230

DSE 251. Managing Large-Scale Graph Data (4)
The goal of this course is twofold: acquaint students with data management issues related to graphs (including storage, indexing, querying, and computing with large graph data) and give them a hands-on experience with Neo4j and Gremlin. Project: Students complete a fully hands-on project whereby they build a new application or a new extension to the software system described above.
Prerequisites: MAS Data Science and Engineering major; DSE 250

DSE 260. Data Science Capstone Design Project (4)
A team design project in the final two quarters of the MAS Data Science and Engineering program culminates in the Spring quarter with a final report and an oral presentation of the capstone project. In addition, there might be a demonstration of the working prototype.

DSE 290. Case Studies in Data Science (2)
Case studies discussed by speakers from industry, government and academia expose students to the needs and uses of different technologies and their roles in model building.
F. Institutions with Data Analytics Programs

The list of selected programs below was taken from the Institute for Advanced Analytics webpage: http://analytics.ncsu.edu/?page_id=4184

*Programs offering a Master of Science in Analytics and similar degrees (Program delivery format: FT=full-time; PT=part-time; O=online. Tuition: r=resident. Tuition estimates may vary at a given institution based on the delivery format. Because tuition typically rises over time, multi-year programs may cost more than the estimate, which is based on 2012-13 data.)*

Institute for Advanced Analytics, North Carolina State University - M.S. in Analytics
FT, 10 months, 30 Credit hrs; established in 2007
Estimated Cost $21,000(r), $36,300

School of Engineering, Northwestern University - M.S. in Analytics
FT – 15 months, 11 courses; established in 2012
Estimated Cost: $58,000

School of Business, Louisiana State University at Baton Rouge - M.S. in Analytics
FT – 12 months, 36 Credit Hrs; established in 2011
Estimated Cost: $8,500(r), $22,000

College of Arts and Sciences, University of San Francisco - M.S. in Analytics
FT – 11 months, 35 credit hrs, established in 2012
Estimated Cost $39,550

Discovery Informatics Institute, Rutgers University - M.B.S. in Analytics
FT/PT, 12-21 months, 43 Credit hrs; established in 2012
Cost: $38,000(r), $62,000

College of Business, University of Tennessee at Knoxville - M.S. in Business Analytics
FT, 17 months, 38-39 credit hrs; established in 2010
Estimated Cost $14,200(r); $40,000

Stern School of Business, New York University - M.S. in Business Analytics
PT, 12 months, 14 courses; established in 2013
Estimated Cost: $63,500

Graduate School of Business, Fordham University - M.S. in Business Analytics
FT, 12 months, 30 credit hrs; established in 2012
Estimated Cost: $34,740

School of Business, George Washington University - M.S. in Business Analytics
FT/PT, 10-24 months, 33 Credit hrs; established in 2013
Estimated Cost $46,860

College of Business, Drexel University - M.S. in Business Analytics
PT/O, 20+months, 45 Credit hrs; established in 2012
Estimated Cost: $45,000

**School of Business, Michigan State University - M.S. in Business Analytics**
FT, 12 months, 30 credit hrs; established in 2013
Estimated Cost $36,000(r), $39,000

**School of Business, Arizona State University - M.S. in Business Analytics**
FT, 9 months, 30 credit hrs; established in 2013
Estimated Cost: $29,600(r), $44,100

**Institute for Applied Computational Science, Harvard University - M.S. in Computational Science and Engineering**
FT, 10 months, 8 courses; established in 2013
Estimated Cost: $40,400

**Center for Data Science, New York University - M.S. in Data Science**
FT, 15-21 months, 36 credit hrs; established in 2013
Estimated Cost $49,800

**School of Continuing Studies, Northwestern University - M.S. in Predictive Analytics**
PT/O, 20+months, 11 courses; established in 2011
Estimated Cost: $40,700

**School of Computing, DePaul University - M.S. in Predictive Analytics**
PT/O, 20+months, 52 Credit hrs; established in 2010
Estimated Cost: $37,000

**School of Business, Saint Joseph’s University - M.S. in Business Intelligence**
FT/PT/O, 15-24 months, 30 credit hrs; established in 2008
Estimated Cost: $26,760

**School of Technology Management, Stevens Institute of Technology - M.S. in Business Intelligence and Analytics**
FT/PT, 20+months, 36 Credit hrs; established in 2012
Estimated Cost $46,000

**College of Humanities and Sciences, Southern Methodist University - M.S. in Applied Statistics and Data Analytics**
FT, 18-24 months, 36 Credit hrs; established in 2013
Estimated Cost: $65,600

**School of Business, University of Connecticut - M.S. in Business Analytics and Project Management**
FT/PT, 20+months, 33 Credit hrs; established in 2012
Estimated Cost: $22,500
School of Business, University of Texas at Austin - **M.S. in Information, Risk and Operations Management**
FT, 10 months, 36 Credit hrs; established in 2013
Estimated Cost: $32,000(r), $38,000

School of Business, University of Maryland, College Park - **M.S. in Business: Marketing Analytics**
FT, 9 months, 30 Credit hrs; established in 2013
Estimated Cost: $45,000(r), $55,000
G. External Reviews and Response

(1) Reviewers

In an effort to facilitate review of our proposed program, we invited appropriate faculty members from other universities and as well as external professionals who are leaders in the field to comment on our proposal as outlined in CCGA’s Guidelines for an Expedited Review for Self-Supporting Programs. These reviews are enclosed and responses to reviewer comments have been incorporated into the current proposal. Reviewers were selected to provide a cross-section of industry experts in the field as well as other academicians knowledgeable about data science. These include national leaders of Big Data initiatives such as NSF Expeditions for sister campuses of Berkeley, Santa Barbara and Irvine.

Divyakant Agrawal, Professor of Computer Science, UC Santa Barbara
Dr. Agrawal’s area of expertise is database systems, distributed computing, data warehousing, and large-scale information systems. In 2007 Dr. Agrawal served as VP of Data Solutions and Advertising Systems at ASK.com. and was the Chief Architect for building the next-generation Business Intelligence and Data Warehousing system at ASK.com. In addition, he developed revenue-sensitive products at ASK.com by applying data-mining and machine-learning technologies over ASK.com’s historical data.

Michael J. Franklin, Thomas M. Siebel Professor of Computer Science, UC Berkeley
Dr. Franklin's area of expertise is large-scale data management infrastructure and applications (“Big Data”) primarily in Database (DB) and Operating Systems and Networking Technology (OSNT) areas. Director of the Algorithms, Machines and People Lab (AMPLab) - an industry and government-supported collaboration specializing in data management, cloud computing, statistical machine learning and topics necessary for making sense of vast amounts of varied and unruly data. AMPLab received an NSF CISE "Expeditions in Computing" award, as part of the White House Big Data Research initiative in 2012.

Waqar Hasan, CEO and Founder InsightOne
Dr. Hasan has 20 years of experience in building high-energy organizations and delivering leading edge software or service products. He served as Vice President of Data Systems at Yahoo! where, he led and managed Yahoo!’s data platform that served all lines of business including Audience properties and Search and Brand advertising. With data from 500 million consumers, these systems are among the largest in the world, affecting nearly all of Yahoo!’s revenue streams. Dr. Hasan founded DB Wizards (acquired by Adeosoft), ran the optimizer team at Informix (now IBM), worked on data research at HP Labs and IBM Almaden, and has consulted for companies such as Microsoft, Visa, Reed-Elsevier and Deloitte. He received his Ph.D. in Computer Science from Stanford University and his B.S. in Computer Science from the Indian Institute of Technology, Kanpur.

Edward Lazowska holds the Bill and Melinda Gates Chair in Computer Science & Engineering at the University of Washington
Dr. Lazowska's research and teaching concern the design, implementation, and analysis of high-performance computing and communication systems. For the first ten years of his career, Prof. Lazowska's principal focus was computer system performance: the development of effective performance evaluation techniques, and the use of these techniques to gain insight about significant computer systems and computer system design issues. He then turned his attention to the design and implementation of distributed and parallel computer systems - work that yielded a number of widely-embraced approaches to kernel and system design in areas such as thread management, high-performance local and remote communication, load sharing, cluster computing, and the effective use of
the underlying architecture by the operating system. Current research includes information technology to support sustainable rural development, data architecture for the Ocean Observatories Initiative, control theory applied to computer system management, and evolving a broad research agenda in Network Science & Engineering. Dr. Lazowska is a Member of the National Academy of Engineering, a Fellow of the American Academy of Arts & Sciences, a Member of the Washington State Academy of Sciences, a Fellow of the Association for Computing Machinery, a Fellow of the Institute of Electrical and Electronics Engineers, and a Fellow of the American Association for the Advancement of Science.

Sharad Mehrotra, Professor of Computer Science, Bren School of Information Sciences, UC Irvine
Dr. Mehrotra's area of expertise is next generation database management systems that provide natural and efficient support for complex multidimensional data sets. Multidimensional data sets abound in numerous application domains in which database technology is currently being deployed (e.g. medical information systems require databases to provide native support for X-rays, volumetric MRI scans, and time varying volumetric information).

(2) Response and Reviewer's Comments

Title
Two reviewers (Franklin and Lazowska) suggested a change in title that reflects a focus on data science. We agree and have retitled the proposal “Data Science and Engineering”, accordingly.

Viability
We were heartened by the nearly unanimous comments concerning the relevance and viability of the program and resounding support for the need and development of a program in data science. Two reviewers (Mehtrotra and Hasan) indicated that the scale up will occur faster than projected. While we took a very conservative approach in our initial proposal, we have revised the numbers to reflect a more realistic initial enrollment of 24 students as recommended. Scaling up is a good problem to have, and we are heartened by support that the program will have large demand. Rapid growth will become a resource management issue, but resources will be available to meet the needs to scale the program.

We were interested to learn that UC Berkeley recently launched a Masters of Information and Data Science. However, their program is entirely on-line and only 27 units.

Curriculum
One reviewer (Agrawal) recommends thinking beyond a "self supporting" program and establishing a regular MS program. We agree with Dr. Agrawal’s vision, with the caveat that we plan for the MAS to serve as a precursor to the development of an MS program in this area. While Dr. Agrawal is not a fan of MAS programs in general, we recognize that the MAS program, with a focus on industry needs, has an audience with a focus different from traditional MS programs. To that end, it has value in and of itself as has been shown with our existing MAS programs. Our intent is to expand the intellectual core of the MAS and as we gain experience with the MAS. This intellectual core grows naturally into an MS concentration which we plan to offer our students in the future. And, as suggested by reviewer Mehtrotra, it “will evolve, going forward .....in security/privacy, NLP methods, vision processing”.

Also, as suggested by Franklin and Agrawal, we expect the courses and programs to attract a broader constituency, which will benefit our existing students as has occurred with the experiential lab infrastructure created for our Wireless Embedded Systems (WES) MAS program -- hardware, software,
laboratory manuals, etc. As with our WES program, the offerings in Data Science and Engineering will certainly support of our future MS specialization.

Several reviewers commented on the courses and curriculum proposed, recommending the addition of courses or particular focus areas (statistics, visualization etc). We agree, these areas provide fundamental building blocks, and they are covered either in the curriculum proposed or in the elective depth sequences. The intent of the program is to provide data scientist with the ability to go deeper in a particular area of interest through the use of elective courses. That said, we added a required case studies seminar to the curriculum, as part of the core course requirements. It provides students with exposure to the needs and uses of different technologies and their roles in model building. The subjects introduced in the case studies course are covered in detail in the rest of the curriculum, with in-depth training that utilizes these concepts in the capstone projects, where students use the skills learned in application to real-world problems.

Reviewer Agrawal commented that the program will produce data generalists (not sought after by big technology companies and research laboratories) not data scientists. This may be an issue of naming, but we unapologetically support the notion of well-trained data generalists. We do not, however, concur with his conclusion concerning their utility. Our intent is to start on a practical side, providing students with fundamental skills and knowledge that build the student’s capabilities as data generalists. We added a case study seminar in the first quarter to expose students to practitioners from industry, government and academia and to give them an insight into solving big data problems. The capstone project will give our students the opportunity to work on a large and complex data science problem and apply the knowledge and tools that they have learned.

Dr. Agrawal also described the core curriculum as upper-division undergraduate courses and introductory level or programming tools. We recognize that as industry practitioners, students will come into the program with differing strengths and knowledge. As such, we have designed the foundational requirements to bring all students up to an equal knowledge base in programming, statistics and databases. The courses are mezzanine courses that straddle upper-division undergraduate and graduate level courses. However, the core and elective depth sequence courses are squarely at the advanced graduate level with increasing levels of sophistication and complexity. We also recognize that students may have extensive experience and knowledge in a particular foundational area; we may allow students to test out of these courses and instead complete their programs with the required units using depth sequences provided through the proposed electives.

Reviewer Lazowska recommends that statistics, machine learning, scalable computing, data management and data visualization should be required. While we support the notion that these elements form the basis of data science, we disagree that all must be required equally. In the case of visualization, for example, we recognize the importance of the area, but at this point don’t consider visualization to be an absolute requirement in the way that programming, statistics or databases are. We have two data visualization courses as electives to give students the opportunity to learn this important area. We also recognize that the program and curriculum will evolve as the program and field matures. At some point we may choose to modify the curriculum; of course any change in the program will only be done with Senate approval.

Reviewer Franklin suggests consideration of using the analytics workflow/lifecycle as a unifying concept for the curriculum, providing a holistic view of data science early in the curriculum. The required Database Systems course in the program covers the lifecycle of data warehousing and analytics. In
particular, it covers data integration architectures, warehousing and creation of added value views and the development of analytics applications on them. We have changed the course name to Data Management Systems to make apparent that the course covers the lifecycle of such applications.

Reviewer Franklin also asked for clarification on the distinction between foundational and core courses and suggested that the capstone may not be appropriate for students with substantial work experience. First, foundational courses are courses required as background with core courses serving as courses central to the degree and from which students gain the knowledge needed to successfully complete their capstone projects. The capstone projects are based in problems/opportunities proposed within the industry/organizations from which the students come, making the issues and problems at hand more relevant particularly to the student with substantial work experience. We view the capstone as an opportunity for students to solve real problems that they are experiencing in their work situation. Thus, as suggested by reviewer Hasan, the capstone projects may be most effective when done in conjunction with industry partners.

Faculty
A number of reviewers commented on UCSD’s undergraduate workload. UCSD’s Academic Senate last year granted the CSE department approval to limit undergraduate enrollment. As such, undergraduate enrollment in the department will level off over the next year.

All the instructors on the proposal from CSE department are indeed ladder rank faculty and the program has broad support within CSE faculty. However, we agree with the reviewers that more ladder-rank faculty are needed for long-term sustenance of this program. As we are actively recruiting additional faculty members, we are joined by a uniquely strong group of researchers from SDSC who have stepped forward to teach and shape the proposed program. Indeed, the reviewers acknowledge the strength of the proposal due to this direct participation from SDSC. We have thought long and hard about delivering a quality program that will enhance opportunities for intellectual growth and development of our existing students and faculty, and we believe we have done so with the proposed program.

Dr. Franklin commented that Yoav Freund is in charge of the lion’s share of courses. Dr. Freund is indeed identified on many of the courses as he is both faculty director and lead for the program as conceptualized. However, this does not necessarily require that he be the instructor for all courses. As faculty are recruited and where appropriate, other faculty and research staff will be available to teach in the program.

Reviewer Lazowska mentioned that the proposal appeared “Thin in terms of statistics faculty”. Professor Freund has done research and published in statistics and in information theory (see below). In addition, we are reaching out to faculty in the mathematics department to determine if there is interest in participating in the program.


Budget

Dr. Agrawal mentions that there is not a line-item in the budget for staff support. This is due to the fact that staff are shared with existing MAS programs and the cost is included in the central administrative fee identified at 8%.

Drs. Franklin and Mehrotra commented on the level of remuneration proposed, suggesting that a larger financial incentive may be needed for faculty to teach in program. We recognize that market salaries in this area may require remuneration beyond the $16K/course (which in the budget is an average cost). However, faculty remuneration may be as high as $20K/course.

As mentioned by reviewers, we may have been too conservative in our original budget and have included a more realistic estimate for enrollment at 24, as suggested by Dr. Mehrotra. Dr. Mehrotra also suggested that some costs are likely to be reduced in the future. We agree, and fortunately will have the funds for rapid scale up, should that be needed.

Applicants

Dr. Hasan recommended replacing the requirement of “2 years of work experience in statistical data analysis” with “2 years work experience in statistical data analysis or business intelligence or data management.” We understand Dr. Hasan’s point but remain committed to focusing on applicants with a background in engineering, statistics or science as we don’t expect a significant number of students to come into our program with a management background.

Dr. Franklin commented that a background in math may be overly restrictive. He suggested we consider administering an entrance exam, or conditional acceptance with remedial courses prior to start of program, should applicants not possess the appropriate formal mathematical background. We agree that a formal mathematical education should not be required and as suggested plan to maintain our high standards of admission through the use of an entrance exam to ensure that students have the appropriate mathematical background.
REVIEW OF FOR COORDINATING COMMITTEE FOR GRADUATE AFFAIRS
UC ACADEMIC COUNCIL

Subject: Proposal Review – Master of Advanced Study in Big Data Science
Reviewer: Divyakant Agrawal, Professor of Computer Science, UC Santa Barbara
Area of Expertise: Database systems, distributed computing, data warehousing, and large-scale
information systems. In 2007 Dr. Agrawal served as VP of Data Solutions and Advertising
Systems at ASK.com. and was the Chief Architect for building the next-generation Business
Intelligence and Data Warehousing system at ASK.com. In addition, he developed revenue-
sensitive products at ASK.com by applying data-mining and machine-learning technologies
over ASK.com’s historical data.
Return by: October 7, 2013 (if unable to do so please contact lindy)
Return to: Lindy Nagata, lnagata@ucsd.edu; 858-822-2457 (as pdf or as image file)

Topics for Consideration in Review:

1. Relevance, importance and viability of the proposed program
The emergence of Big Data is indeed has emerged as an important scientific and
technological challenge in multiple domains ranging from scientific investigations to
industrial settings. At present these activities (i.e., scientific research and advanced
technology development) are occurring in a rather fragmented and ad hoc manner. Much
of the information in popular press and technology media comes as a result of advances
that are being made in industrial settings. In general, industrial companies especially in the
Internet context are successful in doing so because they have access to domain data (a
critical ingredient) and have large amount of resources (they can throw large amount of
compute and storage resources).

As a researcher who has focused on data problems both from management and analysis
point-of-view for the past two decades, I think the current industrial paradigm of throwing
more resources to solve “big data” problems is both irresponsible and unsustainable. It is
therefore extremely important to train the next generation of young engineers and
scientists who will have the in-depth knowledge and understanding of the complexities and
nuances of big data problems. In particular, building models leveraging advanced statistical
and mathematical techniques combined with scalable database and machine learning
technologies is the only way out of the “big data” conundrum. Throwing more data and
computers at “big data” is not going to magically provide us the answers to the questions
that we desperately seek.

Thus, the evolution of “big data science” as an advanced scientific and engineering
discipline is relevant, important, and viable. I am pleased that University of California a San
Diego has decided to lead this effort.
Based on the current version of the proposal, I note that the proponents of this proposal are unaware that a similar program (Master of Information and Data Science administered by UC Berkeley’s School of Information) was approved by the University of California in 2012-13.

2. **Quality and academic rigor**
   
   I have carefully studied the academic program that comprises six core courses, two electives, and a Capstone project. Although I am extremely supportive of the effort given my personal bias, my concern is that the academic curriculum as proposed is not going to result in a cohort of students who will be sought after by the big technology companies and research laboratories at the national level. The program as it stands will produce “Data Generalists” not “Data Scientists”. The rationale for my evaluation is that most of these courses are amalgamation of courses from four different areas: computer programming, probability and statistics, databases, and machine learning. Furthermore, computer programming (Python and SQL) and statistics course as described are at best at the level of upper division undergraduate course (which corresponds to 50% of the core curriculum). The remaining three courses seem to be at an introductory level (machine learning) or are exposing students to programming tools for big data (Hadoop and ETL tools).

   What is missing is, for example, a course that exposes students to “Model Building” by using data from different domains (e.g., highly structured click-stream data to semi-structured data to linked data to text data).

   A more compelling curriculum structure would be to create a sequence of 3 critical courses Data Science I, II, and III – where students are exposed to all four (actually six: Statistics, Programming, Databases, Systems, Data Mining, and Machine Learning) areas at increasing levels of sophistication and complexity. With this modification and ensuring that the intake of students is high-quality (mathematically and scientifically mature), UCSD will have a compelling program.

3. **Adequacy of the size and expertise of faculty to administer**

   The quality of both CSE and SDSC faculty is very strong. This is an excellent team and the combination of a strong computer science department coupled with a large infrastructure facility (with access to both data and computation) is likely to make UCSD a “Data Science” destination if done right. I would urge the team to think beyond the idea of “self-supporting” program for generating short-term revenues but to structure this program as pre-cursor to an advanced research and education destination for data science.

4. **Adequacy of the facilities and budgets**

   The proposal plans to use the existing facilities available from CSE and SDSC.

   With regard to the budget, the proposal mentions the need for staff support but I do not see it as line-item in the detailed budget.
5. **Applicant pool and placement prospects for the graduates**

The proposal will be structured as a self-supporting program. With that the proposal is targeting working professionals from San Diego metropolitan area. I have no data to ascertain the quality of this pool. However, the quality of this applicant pool could be significantly improved if this is structured as a regular graduate academic program.

6. **General comments and/or aspects of the program that could be strengthened**

I understand the desire on the part of administering units (Deans and Department Chairs) to use self-supporting programs as a way to augment operational budgets that have declined substantially due to the budgetary situation in the University of California. But in the long term, it would be much better to structure this program as a regular academic program — since Big Data and Data Science will be an important and relevant discipline of increasing importance.
Subject: Proposal Review – Master of Advanced Study in Big Data Science
Reviewer: Michael J. Franklin, Thomas M. Siebel Professor of Computer Science, UC Berkeley
Area of Expertise: Large-scale data management infrastructure and applications ("Big Data") primarily in Database (DB) and Operating Systems and Networking Technology (OSNT) areas. Director of the Algorithms, Machines and People Lab (AMPlab) - an industry and government-supported collaboration specializing in data management, cloud computing, statistical machine learning and topics necessary for making sense of vast amounts of varied and unruly data. AMPlab received an NSF CISE "Expeditions in Computing" award, as part of the White House Big Data Research Initiative in 2012.

Return by: October 11, 2013 (if unable to do so please contact Lindy)
Return to: Lindy Nagata, lnagata@ucsd.edu; 858-822-2457 (as pdf or as image file)

Big Data science

Topics for Consideration in Review:

This is a very timely and very exciting program that I expect will be of great interest to practitioners in the San Diego area and beyond. The combination of strengths of the UCSD CS Department and those of the San Diego Supercomputer Center make for a very interesting program and one that has significant intellectual advantages over many programs being developed elsewhere to address this topic.

Below I respond to the specific topics suggested for my consideration.

1. Relevance, importance and viability of the proposed program

The program scores very high in terms of relevance and importance. As the proposal document clearly points out, Data Science and Big Data (more about the name “Big Data Science” later – I am not a fan) are areas of tremendous interest across virtually all industries, scientific domains and government agencies. The emerging understanding of the skill set required of Data Scientists indicates that such skills cross boundaries and are not well supported by existing programs in established domains such as Computer
Science, Statistics or Computer Engineering. Thus, new programs combining aspects of these areas are needed, and the proposed program does a good job of integrating these different perspectives.

The proposal states that the "concept for the BDS program emerged from discussions with local industry representatives of our Departmental Advisory Councils and the Corporate Affiliates Program". This participation clearly shows, and the result is a proposed program that should be of great interest to practitioners and their employers.

In terms of "viability", I think there is good reason to believe that from an intellectual standpoint, the area of Data Science as a new concentration will remain important and likely gain importance in the foreseeable future. The viability of this particular program of course, also depends on additional issues such as institutional and faculty buy-in, as discussed below.

All in all, I give the program high marks for relevance and timeliness. It fills a real and growing need in the industrial community that the program is intended to serve. I also believe that there could be benefits for the more traditional audience of Computer Science students at the graduate and undergraduate levels, and I expect that there will be significant demand from those students for access to these courses or others like them.

2. Quality and academic rigor

Given that the program is clearly designed for working professionals, it would appear that the program would be of sufficient quality and rigor that one would expect from a UCSD professional degree. The course sequence seems well thought out, combining mathematical and statistical foundations, data management foundations and advanced topics, as well as practical training on modern tools and techniques.

A capstone project is a useful and important component of the program, integrating and apply the diverse topics presented in the regular course work. I am a bit concerned that such a capstone may not be appropriate for students with substantial work experience, who have already done such
things, but the concept of hand-forming teams based on different skills may help mitigate this by enabling such students to take on roles outside of their existing experience.

The suite of proposed courses is quite large, although I am concerned about whether or not a sufficient level of interest and effort from ladder rank faculty can be sustained, given the large number of courses that are being proposed.

I personally did not quite see a large distinction between the three “foundational” courses and the three “core” courses and I don’t fully understand why that distinction is being made, although I do not believe that this has a significant impact on the quality or rigor of them program.

There is one important topic that I did not see addressed in the proposal, namely, in many emerging discussions of Data Science, attention is paid to the “analytics lifecycle and workflow” – that is, the set of steps and processes that one must go through to achieve a successful Data Science project or product. While the pieces of this workflow are covered through the various courses, I would think that there would be significant benefit in introducing the more holistic view of Data Science early in the curriculum, where it could be used as a unifying structure for the remainder of the courses.

3. **Adequacy of the size and expertise of faculty to administer**

If I have any concerns about the proposal, it is around the ability to really staff all of these courses on a regular basis, with sufficient participation from ladder rank and equivalent faculty. There are a lot of courses here, and I find it hard to believe that the existing faculty are not already feeling overburdened given the recent surge in demand for computer science education at the undergraduate and graduate level.

The proposal states that the named faculty have all made a commitment to do this, and if true, then there is certainly more than enough expertise on the faculty to get it done. Still, the biggest risk I see here is that the
program competes for the time and mindshare of already over-extended faculty. And while the proposal states that more faculty will be hired, those faculty will also have primary responsibilities for the existing undergraduate and graduate programs.

One can't help but notice that in the proposal as it stands, many courses do not have an affiliated CSE faculty member, and that for those that do, Prof. Yoav Freund is in charge of the lion's share of them. This limited coverage is a concern, and perhaps an indication that there may not be sufficient buy-in from CSE faculty.

The proposal states that faculty will teach "off-load courses at off-hours". Again, I worry that busy faculty will not be able to do this. And while there are some financial incentives in place, these do not seem to be sufficiently high given the other demands on faculty time.

4. Adequacy of the facilities and budgets
I am not an administrator so I don't have a strong view on the budget issues. As stated above, the stipend for faculty seems low if the goal is to get ladder rank CSE faculty to be involved.

In terms of facilities I expect that they will be more than adequate for this program, and in fact, given access to the computational and data resources at SDSC, these resources will be a key distinguishing feature of the program compared to similar programs elsewhere.

5. Applicant pool and placement prospects for the graduates

As stated above, I expect that there will be strong interest in this program from the very beginning. While the proposal focuses on regional interest, I expect that there could be significant national and international interest in this program as well.

One thing in the proposal that surprised me was the requirement for a "strong background in mathematics (linear algebra, calculus, probability and statistics)",. I suspect that significant interest in the program could come
from those in industry who have limited formal training in these areas but who have relevant job experience, or who want to move into Data Science as a new field. It seems overly restrictive to require this expertise. An alternative could be to have an entrance exam as part of the application project, and for those candidates who are otherwise qualified, but who may be lacking the formal math and stats background, they could be offered a conditional admittance with a requirement to take and pass a remedial course the summer before the start of their program. Such a course could even be offered online.

In terms of the placement prospects of graduates – I believe that they will be in great demand, and that this program can be a real career booster and accelerator for a wide range of students. Companies will be extremely interested, and perhaps could even be tapped as a source of additional financial support.

6. General comments and/or aspects of the program that could be strengthened

In the sections above, I made a number of comments and suggestions for strengthening the program. I do have one further major issue, however, the name of the program “Big Data Science” seems problematic to me for a number of reasons. In the near term, it just doesn’t sound particularly professional or rigorous. In the longer term, I worry that the term “Big Data” is already having a bit of a negative reaction due to an overload of industry hype, and the fact that that most interesting Data Science problems don’t have all that much data behind them (at least not in terms of multiple Terabytes or Petabytes). This backlash is only going to get worse. I would suggest dropping the “Big” from the title, and focusing on “Data Science”, which I believe, has real currency as an emerging discipline and skill set.

Of my other comments above, I believe the key one from a curriculum point of view is to consider using the analytics workflow/lifecycle as a unifying concept. From the organizational point of view, it is really a question of how to get already overloaded faculty to buy into this new program.
In summary, I think this is a very exciting program that meets a real need in industry today. I believe that it will serve as a model for similar programs around the country and around the world, and I wish you success in this new endeavour.
REVIEW OF FOR COORDINATING COMMITTEE FOR GRADUATE AFFAIRS
UC ACADEMIC COUNCIL
Waqar Hasan

Topics for Consideration in Review:

1. **Relevance, importance and viability of the proposed program**

   Big Data has attracted tremendous investment by the major corporations, venture capitalists and governments and at the same time the area faces a shortage of qualified personnel.

   This is about the most relevant program that could be created by a CS department. I believe you will see applicants exceed program capacity significantly and graduates to be in high demand for at least the next decade.

2. **Quality and academic rigor**

   The program is well designed and does a good job of providing courses that cover the depth and breadth of theoretical and hand-on knowledge required by a data scientist.

3. **Adequacy of the size and expertise of faculty to administer**

   The faculty is well qualified and there is sufficient breadth of faculty expertise to offer a high quality program.

4. **Adequacy of the facilities and budgets**

   I do not see a problem. I do note that my background is in industry rather than academics and therefore the judgment of people more qualified than me on this matter should take precedence.

5. **Applicant pool and placement prospects for the graduates**

   I would recommend replacing the requirement of “2 years work experience in statistical data analysis” with “2 years work experience in statistical data analysis or business intelligence or data management”.

   There is a tremendous shortage of data scientists in industry and in my view the largest pool of applicants will be people working in business intelligence or data management. The numbers of people who have experience in statistical data analysis (statisticians) is relatively low.

   The placement prospects for graduates are exceptionally good. There is a tremendous demand for skilled data scientists that will continue for at least a decade.

6. **General comments and/or aspects of the program that could be strengthened**
(a) I believe the size of the program is quite modest compared to the demand and consideration might be given to a faster pace of scale-up.

(b) Given the relevance of the program to needs in industry, it will be worth considering whether the Capstone projects can be run as projects with industry partners.
October 29, 2013

Dear Dean Pisano

This letter is in response to your request for my assistance in evaluating the UCSD proposal on creating a new MAS program on Big Data Sciences. I would begin by commending UCSD Computer Science Engineering Department and San Diego Super Computing Center for proposing what I consider to be a very important and timely program that will have far reaching consequences.

We are living in the time of data revolution where machines, sensors, a variety of data capture devices has enabled us to collect and monitor every aspect of our lives whether they be personal experiences, health, social interactions, or our interactions with the engineered and physical systems. The ability to automatically and seamlessly monitor social as well as physical worlds at various spatial and temporal granularities has created unprecedented opportunities leading to major data-centric innovations, new opportunities, new efficiencies, and new industries. Companies such as Google and Yahoo! have used such data to provide improved search, better personalized experiences of individuals on the internet, designed novel ways to monetize and funding the new ideas through placement of advertisement. Moving beyond such internet companies, other organizations such as the health care providers such as UCI medical center have developed tools based on social media data which provides highly granular public opinion feedback about their doctors and services. Product companies such as Apple also track such feedback when they launch new product lines or new models. Data centric analysis for business decision making is by no means a new concept — the whole database warehousing industry which supports business intelligence has over the years built substantial tools that empower retail giants like Walmart or large telecommunication companies such as ATT to identify customer needs, plan inventories, organize showrooms, manage consumers and resources. With the advent of cloud computing and availability of granular data through variety of sensors (including cell phones) and social media has created an opportunity and the need for such powerful data analytics tools long used by fortune 500 companies to be universally available for all to use and benefit from — whether they be new startups, small organizations or individuals.

The proposal makes a compelling case based on hard numbers and facts that such a trend towards data-centric computing will not just continue, but will significantly increase over years to come. All this creates a unique need to create a work force capable tackling big data challenges. The new work force must possess unique multidisciplinary abilities to collect, process, analyze, aggregate, mine, visualize large amounts of data in the context of new applications and opportunities. The UCSD proposal on data sciences addresses exactly this need to train students in variety of areas in computer science and engineering, mathematics, and arts focusing on data management technologies, information extraction and integration, scalable approaches to machine learning, graph analysis, data mining, new models of cluster computing (such as Hadoop) and system design. While one can argue to an extent that existing CS curriculum in various schools (UCSD included) already offers many courses that cover all the above concepts required for big data science, it is only through a dedicated program on data science such as the one proposed, that students can gain experience with variety of design constraints and interrelationships amongst different component technologies that are required.

It is for these reasons I am very enthusiastically supportive of the proposed UCSD effort. Let me further address the questions that you asked in the introduction.
REVIEW OF FOR COORDINATING COMMITTEE FOR GRADUATE AFFAIRS
UC ACADEMIC COUNCIL

Subject: Proposal Review – Master of Advanced Study in Big Data Science
Reviewer: Edward Lazowska, Bill and Melinda Gates Chair in Computer Science & Engineering,
University of Washington
Area of Expertise: Computer systems: modeling and analysis, design and implementation, distributed
and parallel systems. Data-intensive science.
Return by: October 7, 2013 (if unable to do so please contact Lindy)
Return to: Lindy Nagata, lnagata@ucsd.edu; 858-822-2457 (as pdf or as image file)

Topics for Consideration in Review:

1. Relevance, importance and viability of the proposed program

2. Quality and academic rigor

3. Adequacy of the size and expertise of faculty to administer

4. Adequacy of the facilities and budgets

5. Applicant pool and placement prospects for the graduates

6. General comments and/or aspects of the program that could be strengthened

Many major universities either have created (or are in the process of creating) course sequences,
Masters programs, and interdisciplinary graduate programs in "data science."

In the past year, workforce projections for "data scientists" by relatively authoritative organizations
have been breathtaking. It is Masters programs that will meet this need. One would expect very large
demand for such a program in the San Diego area. UCSD, because of its strength in Computer Science &
Engineering and in the San Diego Supercomputer Center, is well positioned to offer a flagship program.

The proposal is strong. I recommend wholeheartedly that it be supported. I offer a small number of
comments for the leaders of the effort to consider:
1. I would get rid of "Big" in the title. It's passé. It's also misleading: "the 3 V's" are equally challenging — "volume," "velocity," and "variety."

2. I would make visualization a required course. The foundations of data science are statistics, machine learning, scalable computing, data management, and data visualization. All of these should be required.

3. While UCSD looks sort of thin in terms of Statistics faculty (Stanford and the University of Washington each have two entire departments — Statistics and Biostatistics; it appears that at UCSD, Statistics is buried in the Mathematics department), it nonetheless seemed odd that no one from Statistics was involved.

4. I know UCSD CSE very well. They are an excellent program that has made truly remarkable progress in the past decade, but they are now buried by their undergraduate teaching load, and their relatively new building is stuffed to the gills. As Dean, you need to confront both of these issues, as priorities. I would not use these issues as excuses to block or delay this excellent proposal; rather, I would use the need for this excellent program as yet another reason to do something about CSE's undergraduate enrollments and its facilities.
Relevance, importance and viability of the proposed program: The program addresses a very important and growing need of educating/reeducating our work force to become proficient in data sciences. These are skills that are important today but will likely be a requirement in the future. While UCSD is taking the lead in creating such a program, I anticipate that such programs will be replicated/refined and offered in variety of schools in the future since, as the proposal highlights, the need for data scientists is expected to grow at a very fast pace and no single school or program can meet the emerging needs.

Quality and Academic Rigor: The proposal have done an excellent job of including what I would consider to be all the core elements – basics of statistics, basics of data management, introduction to modern architectures, cluster computing, graph analysis, statistical analysis packages, etc. – of such a program. As is always the case, additional/new needs will emerge when the program is launched (possibly based on student feedback, industrial input, and the experience gained in running the program). The proposal includes mechanisms necessary for such future adaptations based on emerging needs (e.g., through new elective courses and capstone projects). So my strong recommendation is the program be approved without any delay. Once the program is launched it will naturally evolve in new directions and increase its coverage. Going forward, I would advise the faculty in-charge to additionally consider including electives in areas such as security/privacy, NLP methods, vision processing to provide additional opportunities to students to focus the program towards their interests. UCSD has all of the skill sets for such courses already in their faculty.

 Adequacy of the size and expertise of faculty to administer: The UCSD faculty driving the program are amongst the leaders in their respective disciplines. For instance, Prof. Yannis Papakonstantinou and Alin Dutech (whose research expertise overlaps with me) are world recognized experts in the areas of data integration, semi-structured data, and analytics. Prof. Victor Vianu is perhaps the leading authority in theoretical principles underlying database systems who has recently been focusing on verification of data driven systems. The machine learning and statistics team is equally strong with Prof. Yoav Freund, Sanjoy Dasgupta, and Serge Belongie whose work is at the intersection of learning and computer vision. Finally, Dr. Anamath Gupta and Chaitan Baru from SDSC bring in the strength in large data management, modeling, graph analysis, and enterprise data management. These are only a handful of faculty involved. Basically, both the core team of drivers of this proposal as well as other faculty involved are at the forefront of research leading to big data systems. I do not see any shortcoming of the faculty as such in getting the program launched. I am sure additional faculty with complementary skills will be added to the program as new needs arise and the scope of the program increases gets refined.

 Adequacy of the facilities and budgets: The proposal does budget analysis from both a conservative estimate with only 12 students and a more realistic analysis with about 24 students. In both cases, the budget suggests that the program is self-sustaining. In my opinion baseline with 12 students is too conservative and quite likely the program will be able to attract 20-30 students in its first year itself, if advertised properly. So I do not see any major issues with the program being able to raise its own funds. In terms of budget allocation, the TA support (one TA for every 15-30 student) seems reasonable. This is more generous that the TA support for regular courses in UCSD but is the proposal articulates such additional TA support is necessary to ensure high quality of the program -- this argument seems fully justified. The faculty compensation of $16K/4 unit course seems perhaps a bit on the low side to me if the courses are taught as extra load outside the regular teaching. The point that this is more generous compared to summer courses is well taken. However, to any understanding, the summer courses are taught largely by instructors who are at a substantially lower base salary scale compared to the regular tenure track faculty. It would appear to me that the program should consider (if it is possible) to increase faculty remuneration to attract the "star" UCSD faculty (who are listed in the proposal) to actually participate in the program by teaching courses instead of potentially offloading the teaching to institutional staff hired to teach for this program. I do think that the program will be able to easily sustain higher faculty pay.

In terms of facilities, the budget includes a $90K/ year of IT cost which includes technical support as well as web site maintenance etc. While this expense seems reasonable in the initial years, there might be opportunities to save some money in later years when the program is well established. In terms of the hardware budget, $5K for hardware/computing is included for every 12 students. Typically, one would expect that in a program such as this, students should have access to large machines running hadoop, pregel etc. UCSD (specially SDSC) maintains large computing and storage systems which should be available to students enrolled in the program (for instance, for their capstone project), hence there should be no need for an elaborate dedicated hardware budget for this program. Thus a seemingly low budget of $5K seems reasonable.
Applicant pool and placement prospects for the graduates: The program is focused on attracting students who are already employed in the industry who are increasingly finding the need to deal with big data in their daily jobs. These include individuals who are working engineers and data analysts. As a result, the proposal does not focus on placement. The intent is to leverage existing university placement resources to support students in the program who are looking for such help. To me, this is not just sufficient but prudent since there is not much sense in replicating services that are available to students already through other means.

In summary, I enthusiastically support the proposed creation of the MAS program in Big data science. The proposal is very well thought out and addresses an emerging need. The program has a truly distinguished set of associated faculty who are leading experts in various areas of relevance to the program. The budget seems reasonable from the perspective of creating a self-sustaining program. Indeed, I welcome the creation of such a program at UCSD and will urge my university UCI to consider doing something similar to address the emerging needs of future workforce.

Please do not hesitate to contact me if I can be of further assistance.

Prof. Sharad Mehrotra
Director, Center for Emergency Response Technologies
Vice Chair, of Graduate Studies
Department of Computer Science
University of California, Irvine
CA 92617
December 20, 2013

Fr: Amarnath Gupta, Ph. D.
   Research Scientist,
   Director, Advanced Query Processing Lab

To: Whom it may concern

Re: Commitment to teach in the MAS program

I am happy for the opportunity to take part in the planned Master Of Advanced Studies in Data Science and Engineering. I am hereby committing to teach the classes assigned to me each year for at least three years.

Sincerely,

Amarnath Gupta
To Whom It May Concern:

I am happy for the opportunity to take part in the planned Master Of Advanced Studies in Data Science and Engineering. I am hereby committing to teach two classes per year for this program.

Sincerely,

Chaitan Baru
Director, Institute for Data Science and Engineering
Associate Director, Data Initiatives
San Diego Supercomputer Center

January 2, 2014
Jan 2, 2014

To Whom It May Concern:

I welcome the opportunity and commit to teaching classes for the planned Master Of Advanced Studies in Data Science and Engineering.

Sincerely,

[Signature]

Amit Chourasia
Senior Visualization Scientist
San Diego Supercomputer Center
University of California, San Diego
Phone: +1-858-822-3656
December 20, 2013

To whom it may concern:

I have carefully examined the proposal for a new Master of Advanced Studies degree program in Data Science. It is an exciting and timely endeavor, and I would be pleased to be involved with it. I can commit to teaching the DS 210 course once a year, off-load.

Sincerely,

Sanjoy Dasgupta
Professor
Department of Computer Science and Engineering
University of California, San Diego
letter of interest and intent to teach in the MAS program

Charles Elkan <elkan@cs.ucsd.edu>  Sat, Dec 21, 2013 at 10:58 AM
Reply-To: elkan@ucsd.edu
To: Rajesh Gupta <gupta@cs.ucsd.edu>
Cc: Yoav Freund (eng.ucsd) <yfreund@ucsd.edu>, Sanjoy Dasgupta <dasgupta@eng.ucsd.edu>, Yannis Papkekonstantinou <yannis@cs.ucsd.edu>, Lynne Keith-McMullin <lmcmullin@eng.ucsd.edu>, Lindy Nagata <lnagata@eng.ucsd.edu>

Dear Rajesh: This message is to confirm my commitment to teach DSE 220 (Machine Learning) in Winter 2015 as part of the MAS in DSE program. I understand that teaching will be be in Friday/Saturday blocks, and that it will be on an off-load basis. Please note that it is possible that I will be on a leave of absence away from UCSD in 2014/15, in which case unfortunately I will not be able to teach DSE 220.

Best wishes, Charles

[Quoted text hidden]
December 20, 2013

To Whom It May Concern:

I am happy for the opportunity to take part in the planned Master Of Advanced studies in Data Science and Engineering. I am hereby committing to teach two classes each year for at least five years.

Sincerely,

Yoav Freund
Professor, Computer Science and Engineering
University of California, San Diego
December 20, 2013

To Whom It May Concern:

I am happy for the opportunity to take part in the planned Master Of Advanced studies in Data Science and Engineering. I am hereby committing to teach two classes each year for at least five years.

Sincerely,

Yannis Papakonstantinou
Professor, Computer Science and Engineering
University of California, San Diego
December 20, 2013

To Whom It May Concern:

I am excited about the opportunity to take part in the planned Master Of Advanced studies in Data Science and Engineering. We at SDSC have found that it is an increasingly important topic with an increasing audience. I am hereby committing to teach at least one class each year for at least five years.

Sincerely,

Paul Rodriguez, PhD
San Diego Super Computer Center
University of California, San Diego
Date: 12/21/13  
Subject: Master of Advanced studies in Data Science and Engineering

To Whom It May Concern:

I am happy for the opportunity to take part in the planned Master Of Advanced studies in Data Science and Engineering. I am hereby committing to teach at least one class each year for at least five years.

Sincerely,

Robert Sinkovits  
Director, Scientific Applications Group  
San Diego Supercomputer Center