A Proposal for a

MASTER OF SCIENCE DEGREE IN DATA SCIENCE

UNIVERSITY OF CALIFORNIA
UC RIVERSIDE

Marlan and Rosemary Bourns College of Engineering

University of California – Riverside

Riverside, CA 92521

Submitted by

Vassilis Tsotras on behalf of the MS Data Science Program Committee
Professor, Department of Computer Science and Engineering
Director, Data Science Center
Program Director, Data Science Undergraduate Program

M.S. Data Science Approvals

Approvals	Dates

Contact Information:

For any questions, please contact:

Vassilis J. Tsotras

Department of Computer Science and Engineering
Bourns College of Engineering
University of California Riverside, CA 92521 USA

Phone: 951-827-2888
Email: tsotras@cs.ucr.edu

Table of Contents

EXECUTIVE SUMMARY 5

SECTION 1: INTRODUCTION 6

SECTION 2: PROGRAM 13

SECTION 3: PROJECTED NEEDS 17

SECTION 4: PROGRAM FACULTY AND STAFF 19

SECTION 5: COURSES 20

SECTION 6: RESOURCE REQUIREMENTS 25

SECTION 7: GRADUATE STUDENT SUPPORT 26

SECTION 8: GOVERNANCE 27

SECTION 9: SENATE REGULATION CHANGES 28

APPENDIX A: PROGRAM BYLAWS 30

APPENDIX B: NEW COURSE SYLLABI 34

APPENDIX C: LETTERS OF SUPPORT 44

EXECUTIVE SUMMARY

This document is a proposal for a Master of Science (M.S.) degree in Data Science (DS), which will be jointly managed by the departments of Computer Science and Engineering (CSE) and Electrical and Computer Engineering (ECE). Degree requirements and administration of the program are described in this document.

Data has become ubiquitous in everyday life, impacting every profession, including manufacturing, logistics, health care, public safety, and the military. Data also permeates all aspects of science, engineering, and other academic disciplines. As a result, the field of Data Science has emerged as a new academic discipline: the study of data itself. Data Science deals with obtaining insight and information from the analysis of large collections of data. The proposed MS in DS is a comprehensive program studying how data can be collected, transformed, analyzed, and used to solve problems across many application areas.

At UCR, relevant courses related to data management, data mining, information retrieval, big data, machine learning, and artificial intelligence have been offered in the Computer Science & Engineering and the Electrical & Computer Engineering Departments. These courses are regularly offered and are very popular. However, our current MS curricula in BCOE do not permit students to obtain a focused mastery of Data Science.

The proposed program will allow students with an undergraduate degree from a quantitative field, some experience in algorithms and software engineering, and an exposure to introductory statistics to enroll in a masters-level program in Data Science that will grant them a broad understanding of the subject, while focusing on parts of it for a deeper understanding depending upon their interests.

The new program will rely on existing faculty and will be built mostly on existing courses (only three new courses will be added) within the two departments. It will leverage upon existing facilities in the two departments. Future course offerings will also be through CSE and ECE and the program faculty will be from these departments.

SECTION 1: INTRODUCTION

1.1 Program Objectives

The objective of the MS in Data Science program is to provide training in various aspects of the data lifecycle. Students will gain exposure to data collection, data cleaning, data integration, data management, and data visualization, as well as the theories and techniques necessary for data analysis from data mining, machine learning, information retrieval, and artificial intelligence.

The program aims to admit students from various backgrounds with undergraduate training in quantitative fields (e.g., engineering, physics, math, statistics). We expect that applicants will have some experience in programming, software engineering, and algorithms, and some exposure in probability/statistics. The committee overseeing the formation of the program has considered this aspect very carefully and designed a program that provides both breadth and depth. Two new courses were designed with this purpose in mind: They introduce students from different backgrounds to the basic tools and theory in the Data Science field. Students will then be exposed to the breadth of the area through a set of core courses. They will also be able to focus on various aspects of data science and gain in-depth knowledge through specific electives. At the end, students will complete a capstone project (new course) where they will combine technical, analytic, and interpretive skills to design and execute a large-scale data science project that has a focus on real-world applications.

It is also possible to accept students whose undergraduate education did not include the expected experience in programming, software engineering etc. Examples are students whose undergraduate degrees are in chemistry, biology, economics or sociology. Such students may still be admitted to the program with the stipulation that they complete missing courses at the undergraduate level at UCR. The CSE and ECE departments are working on a “bridge” program that could be used as a first step by students who need instruction in undergraduate fundamentals, such as programming, algorithms, and data structures, prior to entering graduate programs in CSE, ECE, or Data Science. Through the bridge program, students without the appropriate background can still finish their MS degree in Data Science within 2 years. We expect the bridge program to increase the reach of this Data Science MS program in the near future.

1.2 Historical Development of Data Science and Departmental Strengths

We live in a world where data is being generated continuously by scientific experiments, digital processes, sensors, social media, mobile devices, etc. The term “big data” refers to data that is arriving from multiple sources at an alarming volume, velocity, and variety. Data Science is a new field that deals with the management of and extraction of knowledge from big data. As a scientific field, Data Science affects research in many domains, including biological sciences, physical sciences, social sciences, and humanities. The importance of Data Science is evident by various related UC-wide initiatives. As an example, UCB has recently created a separate Data Science Division (https://data.berkeley.edu/).

The White House “*Big Data Research and Development Initiative*” committed $200 million to “extract knowledge and insights from large and complex collections of digital data, accelerate the pace of discovery in science and engineering, strengthen our national security, and transform teaching and learning.” NIH launched the *Big Data To Knowledge (BD2K)* initiative “to enable biomedical research as a digital research enterprise, to facilitate discovery and support new knowledge.” *Harnessing the Data Revolution* is part of NSF’s 10 Big Ideas. In particular, “Engaging NSF's research community in the pursuit of fundamental research in data science and engineering, the development of a cohesive, federated, national-scale approach to research data infrastructure, and the development of a 21st-century data-capable workforce.” Other funding agencies (DARPA, IARPA, etc.) have similar research initiatives.

In addition to research, Data Science heavily influences economics and business. Data has become ubiquitous in everyday life: It impacts every profession, from entry-level office workers to CEOs, from team coaches to general managers, from accountants to CFOs. Businesses now have data available to them at a scale that is historically unprecedented; harnessing this data for insight on what customers want provides them with a competitive advantage. Traditional companies (Ford, Walmart, General Electric, etc.) today pride themselves as being transformed to big-data businesses.

Fueled by the explosion of data, Data Science jobs have proliferated and the demand for data scientists is extremely high; moreover, this demand is expected to be strong for years to come. A 2016 McKinsey report forecasted a shortfall of roughly 250,000 data scientists by 2024. Data scientists are the no. 1 most promising job in America for 2019, according to a report from LinkedIn. Similarly, according to Glassdoor, a recruiting site, Data Scientist has been the best job in the US (2015-2019) with around 113K median base salary. Three-fifths of the data science and analytics jobs are in the finance and insurance, professional services, and information technology sectors, but the manufacturing, health care, and retail sectors also are hiring significant numbers of data scientists. We thus expect that the new program will be in high demand among students and will serve the UCR community well.

As another indication of the interest in Data Science, we have experienced high demand among graduate students for related courses (Data Mining, Machine Learning, AI, Big Data, etc.) For example, around 70-100 students attended “CS235: Data Mining”, in its last few offerings; similarly “CS236: Database Management”, “CS 226: Big Data Management” and “CS229: Machine Learning” have enrollments around 50-60 students. We believe many of these students would prefer a degree more concentrated on these particular topics, particularly one with a coordinated project to provide hands-on experience. Thus, we feel this proposed MS program will better serve many of our current students.

Preparing the workforce in Data Science is also important for the local community. Here in the Inland Empire, for example, the Naval Surface Warfare Center (NSWC) in Corona has launched the Universal Hub for Big Data, a project to collect and share Navy data, which will require a qualified workforce. Our ability to keep high-tech employers like this in the region depends on our ability to supply professionals capable of satisfying their technical needs. NSWC has recently contacted BCOE expressing strong interest in the proposed MS in DS program.

Finally, the MS in Data Science will be of interest as a career next path, to the UCR students graduating from the newly approved BS in Data Science program. We expect that some of these students will continue to pursue a PhD degree in CSE or ECE. Further, a BS+MS will be a possibility in the future.

We thus believe that the MS in Data Science program will be instrumental in educating the future Data Scientists by building their expertise from solid core knowledge, covering the essentials in managing and analyzing data, as well as covering the applications of Data Science in real life problems.

The CSE and ECE Departments have many faculty that perform research related to Data Science. The existing strength was instrumental in creating recently the Data Science Center. Moreover, through a Data Science Cluster, three more faculty members were hired (Papalexakis, Eldawy, Oymak). Section 4 discusses the initial program faculty (currently 10 CSE and 5 ECE members). There are existing strong research groups working on Big Data, Database Management, Data Mining, Artificial Intelligence, Deep Learning, Time Series, and Vision. Related research is published in the top conferences and journals, and is consistently funded by various grants from NSF, Army, Navy, DARPA and other funding agencies. Graduates from these groups are very much sought after from the industry (including Google, Amazon, LinkedIn, Microsoft, Facebook etc.)

1.3 Enrollment Projections

We believe that the new MS in Data Science will help to increase the overall graduate enrollment in BCOE, which is also a college aim. In Fall 2019, the three MS programs offered by CSE, ECE or both had the following enrollment: CSE 168, ECE 37 and CEN 43 MS students. As of 7/31/20, the number of Fall 2020 SIRs were: CSE 85, ECE 50, CEN 46. We believe that the new MS in Data Science will be at least as popular as the ECE and CEN MS programs.

We thus aim to start with 20 students in the first year of the program and reach a steady state of 50-60 students within 5 years. This would be achieved without hurting enrollment in the CSE, CEN and ECE MS programs (or the new MS in Robotics), since the MS in Data Science offers a different career path than these other MS programs.

Further, we expect that many of these students will stay on for PhDs in CSE or ECE, thus allowing us to select PhD students who have already been at UCR.

1.4 Relation to Other Programs in UCR and the UC System

We note that the MS in Data Science will be a state-supported program focused on students that are interested in the on-campus experience. It is thus different from the self-supported BCOE MSOL program that offers a Data Science specialization (among others).

UCR has recently approved an M.S. in Business Analytics from the School of Business (in collaboration with the Statistics department). This degree is different from our proposed program as it focuses on non-technical aspects of data management and analysis while we are looking at the computational side of data. UCR’s MS in Business Analytics is more equivalent to UCI’s MS in Business Analytics.

Within the UC system, UC San Diego has a (self-supported) Masters of Advanced Studies program in Data Science and Engineering that runs over Fridays/Saturdays; this program is offered through their Engineering school. Similarly, UCLA offers through the Engineering School, an on-line Master of Science in Engineering With Certificate of Specialization in Data Science Engineering. UC Berkeley has an on-line M.S. in Information and Data Science that is offered through their School of Information. They further provide the “5th Year Master of Information and Data Science” program, open to Berkeley undergraduate students as a path to earning a professional master’s degree in one additional calendar year. UC Berkeley also has a (self-supported) MS in Engineering program through their Electrical Engineering and Computer Science Department, that offers a concentration in Data Science and Systems. UC Irvine offers a M.S. in Business Analytics offered by the School of Business and has a more business rather than a technical focus.

MS in Data Science is offered by many top universities that have strong research in the area. Examples include the Master in Computational Data Science at CMU, the Masters in Data Science at NYU and the Masters in Data Science at Columbia University.

1.5 Contributions to Diversity

Because of its ubiquitousness and inherent interdisciplinarity, Data Science has an enormous, and still largely untapped, potential for increasing diversity in computing. In Fall 2016, women accounted for only 13.8% of the Computer Science and Electrical Engineering undergraduate enrollment (including all majors offered by the departments). Because our program draws on undergraduates from a more diverse set of majors, we expect to have a more balanced set of applicants. Further, due to Data Science’s relationship to a large variety of application areas, we expect this major to appeal to a broader set of students (including more women) than a traditional Computer Science or Electrical Engineering degree. The recent surge of workshops and conferences that promote diversity in Data Science and related fields, with prominent examples including “WiML” (Women in Machine Learning; https://wimlworkshop.org/), “WiDS” (Women in Data Science; https://www.widsconference.org/), and “BPDM” (Broadening Participation in Data Mining; https://www.facebook.com/BPDMProgram) is strong empirical evidence for the validity of our premise. Similarly there are initiatives like CAWIT (Center for Advancing Women in Technology; https://www.cawit.org/)

whose aim is to increase the participation of women in computing and information technology, by developing new interdisciplinary computing degree programs that educate more women innovators for the Digital Age. CAWIT has recently supported our undergraduate Data Science major with a grant to enable us start and advertise the program.

UCR is an accredited Hispanic Serving Institution (OPEID 00131600), with approximately 35% Hispanic enrollment. This provides an ideal environment for recruitment of underrepresented graduate students. We expect most of the MS in DS students to come from STEM undergraduate programs. Having an MS in DS program will allow us to attract a large number of these students by providing a focus area, thus enriching the diversity of our graduate student pool. The fast-growing nature of the Data Science field is a great motivating factor for these students to complete an MS before entering the workforce. We also expect that some of the MS students will stay on for a PhD, thus enhancing diversity in the associated PhD programs too. The oversight committee of the MS in DS program will organize an open feedback session at the end of each academic year in order to obtain qualitative feedback from students and instructors. In addition, the committee will perform quantitative diversity assessment through anonymous student survey and evaluation, in collaboration with the two participating departments.

1.6 Comments from other UC programs

We identified four Data Science related MS programs in other UCs (all self-supported). We sent copies of this proposal to the chairs of these programs, with a cover letter using the sample provided by the Senate instructions. These programs were (including the date of the letter): (1) UCB: Master of Information and Data Science (9/4/2020), (2) UCB: Master of Engineering in EECS with concentration on Data Science and Systems (9/4/2020), (3) UCSD: Master of Advanced Studies in Data Science and Engineering (9/1/2020), and (4) UCLA: Master of Science in Engineering With Certificate of Specialization in Data Science Engineering (9/1/2020). However, no comments on the proposed program have been received, with three of the programs not responding and the other one stating that this was out of their domain.

1.7 Administration of the Program

The program will be led by a Program Director, assisted by an Associate Director. While the Director will focus on the overall program and coordination among the departments, the Associate Director will serve the role of Graduate Advisor taking care of all graduate student advising issues within the program. A staff member will help the faculty Directors in administering the program. The program faculty will consist of Senate faculty in related research areas from the two departments (see list of the initial program faculty in Section IV). In the interest of efficient administration, a core group of faculty will be appointed to oversee the program and coordinate efforts with the two departments. This Oversight Committee will consist of 5 faculty from the two departments (three from CSE and two from ECE), including the Director and Associate Director.

This initial proposal was created by the following group of faculty:

Samet Oymak (ECE)
Vagelis Papalexakis (CSE)
Mariam Salloum (CSE)
Christian Shelton (CSE)
Vassilis Tsotras (CSE) - Committee Chair

1.8 Evaluation of the Program

As is the norm for all graduate programs at the UCR campus, the program will follow the Senate-mandated review (once every six or seven years). Beginning with the second year, the Program Committee will initiate an internal review of the M.S. Data Science Program.

SECTION 2: PROGRAM

Below we describe the undergraduate admission requirements, the program of study and provide a sample time plan.

2.1 Admission Requirements

All applicants to this program must have completed a Bachelor's degree or its approved equivalent from an accredited institution and to have attained undergraduate record that satisfies the standards established by the Graduate Division and University Graduate Council. Students need experience in a quantitative field with experience in programming, software engineering, algorithms, and background in statistics. Competence in these areas is defined by the following UCR undergraduate courses (or equivalents):

- CS 141 - Intermediate Data Structures and Algorithms
- CS 100 - Software Construction
- MATH 010A - Multivariable Calculus
- MATH 031 - Linear Algebra
- A course covering foundations of probability and statistics (such as STAT 155 - Probability and Statistics for Science and Engineering, or, EE 114 - Probability, Random Variables, and Random Processes in Electrical Engineering)

Applicants who fail to meet this criterion may sometimes be admitted with course deficiencies, provided they take remedial steps to cover the deficiencies. A student who is deficient in a competency area may be asked to complete the corresponding UCR course with a letter grade of at least B, or to pass a challenge examination based on that course's final exam with a grade of at least B. All such remedial work cannot be counted towards the MS degree requirements and should be completed within the first year of graduate study, and in all cases the deficiency(s) must be corrected BEFORE a student can enroll in any graduate course from the same specialty area. The details will be decided by the Graduate Advisor of the program in consultation with the student. The CSE and ECE departments are working currently on a 'bridge' program that can be used as a first step by students who lack basic undergraduate background in programming, algorithms and data structures.

All applicants must submit scores from the Graduate Record Exam, General Test (GRE). Relevant GRE subject tests may be beneficial to the candidate's application, but are not required. Applicants whose first language is not English are required to submit acceptable scores from the TEST of English as a Foreign Language (TOEFL) or the International English Language Testing System (IELTS) unless they have a degree from an institution where English is the exclusive language of instruction. Additionally each applicant must submit letters of recommendation, as per the admission requirements. All other application requirements are specified in the graduate application.

2.2 Data Science MS Program

The MS Data Science program will be built using existing courses, with three new courses[1] (CS/EE 251A: Data Analytics and Exploration, CS/EE 251B: Fundamentals of Data Science and CS/EE 279: Capstone Project in Data Science). The MS in Data Science requires the completion of 49 units of coursework, including a capstone project. There are no thesis or comprehensive exam options; i.e. it falls in the category of a Master's II (with capstone).

Units are divided among core courses (6 courses, for a total of 24 units), elective courses (5 courses, for a total of 20 units), a professional development course (1 unit) and the capstone course. All students must complete the same core courses. Elective courses are selected by the student within a list of possible courses, and students can petition to select a course not on the list.

Core courses:

1. CS/EE 251A: Data Analytics and Exploration (new course)
2. CS/EE 251B: Fundamentals of Data Science (new course)
3. CS 224: Fundamentals of Machine Learning
4. CS 226: Big Data Management
5. CS 235: Data Mining Techniques
6. CS 236: Database Management

Elective courses:

The five electives can be selected from the following two lists of elective courses; at least three of the courses must be from list A. The description of all the elective courses is available later in the proposal. Students may petition for other elective courses; such

[1] All three courses have been approved by both the CSE (10/7/20) and ECE (10/14/20) Departments and are in the process of senate approval.

petitions require approval of the program graduate advisor.

Elective List A:

1. CS 205: Artificial Intelligence
2. CS 225: Spatial Computing
3. CS 227: Probabilistic Models for Artificial Intelligence
4. CS/EE 228: Introduction to Deep Learning
5. CS 229: Machine Learning
6. CS 242: Information Retrieval and Web Search
7. CS/EE 248: Optimization for Machine Learning
8. EE 231: Convex Optimization in Engineering Applications
9. EE 240: Pattern Recognition
10. EE 244: Computational Learning

Elective List B:

1. CS 210. Scientific Computing
2. CS 211. High Performance Computing
3. CS/EE 217: GPU Architecture and Parallel Programming
4. CS 234: Computational Methods for Biomolecular Data
5. EE 241: Advanced Digital Image Processing
6. EE 243: Computer Vision
7. EE 250: Information Theory

Capstone Experience: Students must complete a capstone course CS/EE 279: Capstone Project in Data Science (new course), under the guidance of the capstone instructor member. The description of the capstone course appears in Section 5.

Professional Development Requirement: Students will satisfy the professional development requirement by enrolling in one of the following courses: one quarter of CS 287 (Colloquium in Computer Science), or GDIV 403 (Research and Scholarship Ethics), or at least one unit of CS 298I (Individual Internship).

2.3 Other Requirements

There are no field or qualifying examinations. There is no thesis/dissertation or final examination. There are no special requirements over and above the Graduate Division minimum requirements.

2.4 Sample Program

Below we provide a sample program. Assuming that a student has no deficiencies and is full-time, the normative time from matriculation to degree is 4 quarters. Using the currently planned bridge program in CSE, it is expected that students with deficiencies can still graduate within 2 years.

	Fall	**Winter**	**Spring**
Year 1	CS/EE 251A CS/EE 251B CS 236	CS 226 CS 224 Elective	CS235 Elective Elective
Year 2	Elective Elective CS/EE 279		

SECTION 3: PROJECTED NEEDS

3.1 Student Demand and Opportunities

We expect a large demand for the new program. The numbers of students in related programs, like the BS and MS programs in CSE, ECE and CEN continue to increase. The proportion of domestic students in related MS programs is around 17% in CSE, 20% in CEN and 33% in ECE; we expect that for the MS in DS the percentage of domestics will be closer to the ECE example. This is because Data Science seems to be popular with domestic students. Moreover, there are a number of students enrolled in the new B.S. program in Data Science and this program will draw from students who complete the DS B.S. program (the BS in DS inaugural class of Fall 2020 is 100% domestic students). Many students in the various Undergraduate Professional Societies have expressed interest in a Data Science graduate program. While we have most of the courses, the structures of the existing programs do not allow them to take the proper set of courses required for specialized training in Data Science This demand is directly related to opportunities for students after graduation.

3.2 Opportunities for Placement of Graduates

Based on our experience from graduate students working in relevant areas (Databases, Data Mining, Artificial Intelligence, Machine Learning etc.) there is currently very high demand from industry. Moreover, as mentioned in the Introduction, according to Glassdoor, a recruiting site, Data Scientist has been the best job in the US (2015-2019).

3.3 Importance to the Discipline

As a scientific field, Data Science affects research in many domains, including biological sciences, physical sciences, social sciences, and humanities. In addition to the newly approved Data Science undergraduate program, the proposed MS in Data Science allows students to concentrate further on this important field.

3.4 Meeting the needs of Society

Data is an important societal asset. By training more students in Data Science we also create more “citizen scientists”. According to CitizenScience.gov (the official government website dedicated to Citizen Science), a citizen scientist “...participates voluntarily in the scientific process, addressing real-world problems in ways that may include formulating research questions, conducting scientific experiments, collecting and analyzing data, interpreting results, making new discoveries, developing technologies and applications, and solving complex problems". Such involvement can engage the American public in addressing societal needs and accelerating science, technology, and innovation.

3.5 Relation to Research and Faculty Interests

A critical mass of our faculty are engaged in research and teaching across the full range of areas relevant to the proposed MS program. This is also evident from the fact that almost all the courses for the program already exist at UCR. These areas are already of high interest to faculty. Moreover, faculty is well funded in these research areas.

3.6 Program Differentiation

A current list of Data Science related programs in California appears at http://datascience.community/colleges. The majority of these programs are on-line, or focus in Business Analytics. Some are MS programs that provide concentrations in Data Science. Below we describe how the proposed MS in Data Science differs.

The MS in Data Science is a state-supported program. It is thus different from UCR's MSOL program in Data Science as well other similar online programs in other UC campuses. It is also different from MS in Business Analytics or MS in Information Management programs as such programs focus on non-technical aspects of data management and analysis while we are looking at the computational side of data. UCR's MS in Data Science is also different from UC Berkeley's (self-supported) MS in Engineering with Concentration in Data Science and Systems. To get that concentration, students need to take 4 technical courses from a list of approved EECS courses and a capstone project. The UCR MS in Data Science is not a concentration but the whole focus of the MS degree though a well designed curriculum that offers many opportunities to students to train in Data Science related coursework. Our program is also different from the Masters of Advanced Studies program in Data Science and Engineering from UC San Diego (also self-supported), which runs Fridays/Saturdays and focuses on mid-career professionals. It is also different from UCLA's Master of Science in Engineering With Certificate of Specialization in Data Science Engineering, which is similar to our MSOL program.

The MS in Data Science at UCR will be the first state-supported MS program in this important subject.

Among private California Institutions there are two related programs: (1) the MS in Information Systems & Technology with concentration in Data Science, offered by Claremont Graduate University, and (2) the Master of Science in Computer Science (Data Science) offered by USC. We believe that we offer a very competitive program from a public institution that is actually named MS in Data Science.

SECTION 4: PROGRAM FACULTY AND STAFF

The list of the Program Faculty (with a link to their publications) appears below:

CSE

Ahmed Eldawy (Assistant Professor; PhD; https://dblp.uni-trier.de/pers/hd/e/Eldawy:Ahmed)
Vagelis Hristidis (Professor; PhD; https://dblp.uni-trier.de/pers/hd/h/Hristidis:Vagelis)
Eamonn Keogh (Professor; PhD; https://dblp.uni-trier.de/pers/hd/k/Keogh:Eamonn_J=)
Paea LePendu (Assistant Teaching Professor; PhD; https://dblp.uni-trier.de/pers/hd/l/LePendu:Paea)
Amr Magdy (Assistant Professor; PhD; https://dblp.uni-trier.de/pers/m/Magdy_0001:Amr.html)
Evangelos Papalexakis (Assistant Professor; PhD;
https://dblp.uni-trier.de/pers/hd/p/Papalexakis:Evangelos_E=)
C.V. Ravishankar (Professor; PhD; https://dblp.uni-trier.de/pers/hd/r/Ravishankar:Chinya_V=)
Mariam Salloum (Assistant Teaching Professor; PhD; https://dblp.uni-trier.de/pers/hd/s/Salloum:Mariam)
Christian Shelton (Professor; PhD; https://dblp.uni-trier.de/pers/hd/s/Shelton:Christian_R=)
Vassilis Tsotras (Professor; PhD; https://dblp.uni-trier.de/search?q=tsotras)

ECE

Salman Asif (Assistant Professor; PhD; https://dblp.uni-trier.de/pers/hd/a/Asif:Muhammad_Salman)
Bir Bhanu (Professor; PhD; https://dblp.uni-trier.de/pers/hd/b/Bhanu:Bir)
Samet Oymak (Assistant Professor; PhD; https://dblp.uni-trier.de/pers/hd/o/Oymak:Samet)
Amit Roy-Chowdhury (Professor; PhD; https://dblp.uni-trier.de/pers/hd/r/Roy=Chowdhury:Amit_K=)
Nanpeng Yu (Assistant Professor; PhD; https://dblp.uni-trier.de/pers/hd/y/Yu:Nanpeng)

STAFF

One FTE for administrative support, primarily for graduate student admissions, enrollment and advising. Initial support may be less than 1 FTE, ramping up as the program matures.

TEACHING RESOURCES

The new program is based on existing courses from CSE and ECE. The three new courses will be cross-listed between the two departments which will share responsibilities in teaching them. In the Appendix we include letters of support from the two department chairs that also discuss the sharing of the teaching.

SECTION 5: COURSES

Core Courses

New Courses Developed for the Program

CS/EE 251A. *Data Analytics and Exploration* (4) Lecture, 3 hours; outside research, 3 hours. Prerequisite(s): CS141, CS100, Stat 155 or EE114 or equivalent. This course covers important algorithms relevant to the lifetime of data from data collection and cleaning to integration, data mining and analytics. Topics include: sketch algorithms for computing statistics on data streams; mining social graphs, including community detection and graph partitioning; Data Science lifecycle and techniques on data cleaning, data integration, Exploratory Data Analysis, and visualization.

CS/EE 251B. *Fundamentals of Data Science* (4) Lecture, 3 hours; outside research, 3 hours. Prerequisite(s): Math 010A, Math 031 or EE 020, CS100, Stat 155 or EE114 or equivalent. Explores theoretical tools in data science and their applications in data science. The course introduces and motivates statistical and computational viewpoints on data analysis. Topics include the manipulation of data as vectors, drawing inferences from data as distributions, and quantifying data uncertainty for data analysis. The course will also include in-class and homework exercises on practical applications of these theoretical data science tools.

CS/EE 279. *Capstone Project in Data Science* (4) Lecture, 1 hour; outside research, extra readings, 9 hours. Prerequisite(s): Enrollment in Master in Data Science. Co-requisites: CS/EE 251A, CS/EE 251B, CS224, CS226, CS235, CS236. Covers combining technical, analytic, and interpretive skills to design and execute a large-scale data science capstone project that has a focus on real-world applications. Provides an opportunity to integrate all of the core skills and concepts learned throughout the program and prepares students for long-term professional success in the field. Emphasizes collaboration and communication in both written and oral form.

Existing Core Courses

CS 224: *Fundamentals of Machine Learning.* (4) Lecture, 3 hours; outside research, 3 hours. Prerequisite(s): CS 100, STAT 155, MATH 31 . A study of generative and discriminative approaches to machine learning. Topics include probabilistic model fitting, gradient-based loss optimization, regularization, hyper-parameters, and generalization. Includes experience with data science programming environments, data from practice, and performance metrics.

CS 226. *Big-Data Management* (4) Lecture, 3 hours; term paper, 3 hours.

Prerequisite(s): CS 166. Introduction to the architecture and design of big data management systems. Covers the design of distributed file systems and high throughput databases. Description of popular programming paradigms for big data including MapReduce and Resilient Distributed Datasets. Includes a course project with hands-one experience on open-source big data systems. May be taken Satisfactory (S) or No Credit (NC) with consent of instructor and graduate advisor.

CS 235. *Data Mining Techniques* (4) Lecture, 3 hours; term paper, 1.5 hours; project, 1.5 hours per week. Prerequisite(s): CS 141, CS 166; CS 170 is recommended. CS 235 online section; enrollment in the online Master of Science in Engineering program. Provides students with a broad background in the design and use of data mining algorithms and tools. Includes clustering, classification, association rules mining, time series clustering, and Web mining. May be taken Satisfactory (S) or No Credit (NC) with consent of instructor and graduate advisor.

CS 236. *Database Management Systems* (4) Lecture, 3 hours; outside research, 3 hours. Prerequisite(s): CS 141; CS 153 or equivalent; CS 166; or consent of instructor. Covers principles of file systems; architecture of database management systems; data models; and relational databases. Also examines logical and physical design of databases; hardware and software implementation of database systems; and distributed databases (e.g., query processing, concurrences, recovery). May be taken Satisfactory (S) or No Credit (NC) with consent of instructor and graduate advisor.

Electives

CS 205. *Artificial Intelligence* (4) Lecture, 3 hours; discussion, 1 hour. Prerequisite(s): CS 170 or equivalent. Examines knowledge representation and automated reasoning and their use in capturing common sense and expert knowledge. Also addresses predicate and nonmonotonic logics; resolution and term rewriting; reasoning under uncertainty; theorem provers; planning systems; and belief networks. Includes special topics in natural language processing, perception, logic programming, expert systems, and deductive databases. May be taken Satisfactory (S) or No Credit (NC) with consent of instructor and graduate advisor.

CS 210. *Scientific Computing* (4) Lecture, 4 hours. Prerequisite(s): CS 012, MATH 010A; MATH 031 or equivalent; or consent of instructor. Utilizes scientific computing in a specific computer science research area. Provides a foundation for pursuit of further studies of special topics in scientific computing. May be taken Satisfactory (S) or No Credit (NC) with consent of instructor and graduate advisor.

CS 211. *High Performance Computing* (4) Lecture, 3 hours; research, 3 hours. Prerequisite(s): CS 161 or consent of instructor. Introduces performance optimization for sequential computer programs. Covers high performance computing on multicore shared memory computers and on distributed memory computing clusters. Also covers high performance scientific libraries and computing application development using pthreads, OpenMP, and Message Passing Interface (MPI) parallel file systems. May be taken Satisfactory (S) or No Credit (NC) with consent of instructor and graduate advisor.

CS/EE 217. *GPU Architecture and Parallel Programming* (4) Lecture, 3 hours; consultation, 1 hour. Prerequisite(s): CS 160 with a grade of “C-” or better or consent of instructor. Introduces the popular CUDA based parallel programming environments based on Nvidia GPUs. Covers the basic CUDA memory/threading models. Also covers the common data-parallel programming patterns needed to develop a high-performance parallel computing applications. Examines computational thinking; a broader range of parallel execution models; and parallel programming principles. May be taken Satisfactory (S) or No Credit (NC) with consent of instructor and graduate advisor.

CS 225. *Spatial Computing* (4) Lecture, 3 hours; individualized study, 3 hours. Prerequisite(s): graduate standing; or consent of instructor. Introduction to the spatial computing technologies and techniques. Covers the fundamentals, the present, and the emerging use cases of spatial data analysis systems. Topics include spatial data modelling, spatial relationships, storage, indexing, query processing, and recent trends in the field. Includes a research-oriented project and hands-on experience on spatial technologies. May be taken Satisfactory (S) or No Credit (NC) with consent of instructor and graduate advisor.

CS 227. *Probabilistic Models for Artificial Intelligence* (4) Lecture, 3 hours; written work, 3 hours. Prerequisite(s): CS 141, STAT 155. Covers methods for representing and reasoning about probability distributions in complex domains. Focuses on graphical models and their extensions such as Bayesian networks, Markov networks, hidden Markov models, and dynamic Bayesian networks. Topics include algorithms for probabilistic inference, learning models from data, and decision making. May be taken Satisfactory (S) or No Credit (NC) with consent of instructor and graduate advisor.

CS/EE 228: *Intro to Deep Learning.* (4) Lecture, 3 hours; written work, 3 hours. Prerequisite(s): CS 225 or EE 236 or EE 231 or EE 244 or CS 171 or EE 142 or consent of the instructor. Explores fundamentals of deep neural networks and their applications in various machine learning tasks. Includes the fundamentals of perception, approximation, neural network architectures, loss functions, and generalization. Addresses optimization methods including backpropagation, automatic differentiation, and regularization. Covers non-standard problems including autoencoders, weak supervision and probabilistic models. Presents applications in machine learning/computer vision.

CS 229. *Machine Learning* (4) Lecture, 3 hours; outside research, 3 hours. Prerequisite(s): CS 100, STAT 155. CS 229 online section; enrollment in the Online Master-in-Science in Engineering program. A study of supervised machine learning that emphasizes discriminative methods. Covers the areas of regression and classification. Topics include linear methods, instance-based learning, neural networks, kernel machines, and additive models. May be taken Satisfactory (S) or No Credit (NC) with consent of instructor and graduate advisor.

CS 234. *Computational Methods For Biomolecular Data* (4) Lecture, 3 hours; research, 3 hours. Prerequisite(s): CS 111; CS 141 or CS 218; STAT 155 or STAT 160A. A study of computational and statistical methods aimed at automatically analyzing, clustering, and classifying biomolecular data. Includes combinatorial algorithms for pattern discovery; hidden Markov models for sequence analysis; analysis of expression data; and prediction of the three-dimensional structure of RNA and proteins. May be taken Satisfactory (S) or No Credit (NC) with consent of instructor and graduate advisor.

CS 242. *Information Retrieval and Web Search* (4) Lecture, 3 hours; term paper, 1.5 hours; project, 1.5 hours per week. Prerequisite(s): CS 141, CS 166. CS 242 online section; enrollment in the online Master of Science in Engineering program. Introduces Information Retrieval (IR) principles and techniques for indexing and searching document collections with special emphasis on Web search. Includes text processing, ranking algorithms, search in social networks, search evaluation, and search engines scalability. May be taken Satisfactory (S) or No Credit (NC) with consent of instructor and graduate advisor.

CS/EE 248: *Optimization for Machine Learning.* (4) Lecture, 3 hours; research, 3 hours. Prerequisite(s): CS 229 or EE 231 or EE 244 or consent of the instructor. Explores efficient optimization algorithms for machine learning. Emphasizes fundamental principles, provable guarantees, and contemporary results. Includes fundamentals of optimization (first-order methods, stochastic algorithms, accelerated schemes, non-convex optimization, regularization, and black-box optimization). Also covers connections to statistical learning (empirical risk minimization, finite-sample guarantees, and high-dimensional problems).

EE 231. *Convex Optimization in Engineering Applications* (4) Lecture, 3 hours; term paper, 3 hours. Prerequisite(s): EE 230. Covers recognizing and solving convex optimization problems that arise in engineering applications. Explores convex sets, functions, and optimization problems. Includes basics of convex analysis, least-squares, linear and quadratic programs, semidefinite programming, minimax, and other problems. Addresses optimality conditions, duality theory, theorems of alternative and applications, interior-point methods, and applications in engineering.

EE 240. *Pattern Recognition* (4) Lecture, 3 hours; research, 3 hours. Prerequisite(s): EE 141 or consent of instructor. EE 240 online section; enrollment in the Online Master-in-Science in Engineering program . Covers basics of pattern recognition techniques. Topics include hypothesis testing, parametric classifiers, parameter estimation, nonparametric density estimation, nonparametric classifiers, feature selection, discriminant analysis, and clustering.

EE 241. *Advanced Digital Image Processing* (4) Lecture, 3 hours; outside research, 3 hours. Prerequisite(s): EE 152 or consent of instructor. Covers advanced topics in digital image processing. Examines image sampling and quantization, image transforms, stochastic image models, image filtering and restoration, and image data compression

EE 243. *Advanced Computer Vision* (4) Lecture, 3 hours; outside research, 3 hours. Prerequisite(s): EE 146 or consent of instructor. EE 243 online section; enrollment in the Online Master-in Science in Engineering program. A study of three-dimensional computer vision. Topics include projective geometry, modeling and calibrating cameras, representing geometric primitives and their uncertainty, stereo vision, motion analysis and tracking, interpolating and approximating three-dimensional data, and recognition of two-dimensional and three- dimensional objects.

EE 244. *Computational Learning* (4) Lecture, 3 hours; research, 3 hours. Prerequisite(s): graduate standing or consent of instructor. Explores fundamental computational learning techniques. Topics include elements of learning systems, inductive learning, analytic learning, case-based learning, genetic learning, connectionist learning, reinforcement learning and integrated learning techniques, and comparison of learning paradigms and applications.

EE 250. *Information Theory* (4) Lecture, 3 hours; extra reading, 3 hours. Prerequisite(s): EE 215. An overview of fundamental limitations imposed on communication systems. Topics include Shannon’s information measures, weak and strong typicality, lossless data compression, source and channel models and Shannon’s coding theorems, channel capacity and the rate-distortion function, Gaussian sources and channels, and limits of communication between multiple terminals.

SECTION 6: RESOURCE REQUIREMENTS

All the technical resources required by the M.S. Data Science program are already available in and for the two participating departments including computing facilities, library resources, teaching laboratories and research facilities. The only additional resources would be office space and one FTE for administrative support (initial support may be less than 1 FTE, ramping up as the program matures).

SECTION 7: GRADUATE STUDENT SUPPORT

MS Data Science students are expected to be self-supported. However, GSR and Teaching Assistantships may be available on a case-by-case basis.

SECTION 8: GOVERNANCE

The Program Faculty will consist of Senate faculty in related research areas to Data Science, drawn from the two departments. Program Faculty members shall support the program through instruction of courses, supervision of students, activity in Data Science research, or program administration. All Program Faculty are eligible to vote on matters related to the MS in Data Science Program. All changes to the MS in Data Science Program or curriculum must be approved by a majority of the Program Faculty.

The program will be led by a Program Director, assisted by an Associate Director. The Director is appointed by the Dean of BCOE with consultation from the Program Faculty. The Program Director will rotate among the 2 departments. While the Director will focus on the overall program and coordination among the departments, the Associate Director will serve the role of Graduate Advisor taking care of all graduate student advising issues within the program. A staff will help the faculty Directors in administering the program. The staff will report to the Director and the Director will report to the Dean of BCoE.

A core group of the program faculty (including the Director and Associate Director) will be appointed to form the Oversight Committee, whose task is to oversee the program and coordinate efforts with the departments. The committee will consist of three faculty from CSE and two faculty from ECE.

SECTION 9: SENATE REGULATION CHANGES

No changes in Senate Regulations at the Divisional level or in the Assembly of the Academic Senate will be required.

APPENDIX A: PROGRAM BYLAWS

MS in Data Science Program Bylaws
Creation Date: June 2, 2020
Approval Date:

I. Objective
   A. The MS in Data Science is housed in the Bourns College of Engineering (BCOE), and is a joint program between the departments of Computer Science and Engineering (CSE) and Electrical and Computer Engineering (ECE).
   B. The objective of the MS in Data Science is to provide training in various aspects of Data Science. Students graduating from the program will gain exposure to the foundational principles underlying the full data lifecycle, from storage to management to analysis.

II. Membership
   A. The faculty associated with the program, called the Program Faculty, is drawn from UCR Senate faculty in related research areas from the CSE and ECE departments.
   B. Program Faculty members shall support the program through instruction of courses, supervision of students, activity in Data Science research, or program administration.
   C. All Program Faculty are eligible to vote on matters related to the MS in Data Science Program.
   D. All changes to the MS in Data Science Program or curriculum must be approved by a majority of the Program Faculty.
   E. UCR Senate faculty outside of CSE and ECE whose research or teaching activities align with the mission of the MS in Data Science are eligible to be Cooperating Faculty in the program. Cooperating Faculty do not have a vote in the program, but are eligible to participate in meetings of the Program Faculty.
   F. Membership Changes
      1. Nominations of prospective members to the Program Faculty or Cooperating Faculty may be made by any faculty member in CSE or ECE.
      2. New Program Faculty or Cooperating Faculty shall be appointed by a majority vote of the Program Faculty, based on a review of the nomination and the recommendation of the Oversight Committee, defined in III.A below.

3. Members of the Program Faculty may terminate their association with the MS in Data Science Program after so informing the Program Director in writing.
4. Participation as Program Faculty or Cooperating Faculty shall be reviewed every three years to ensure that all members are meeting their obligations to the MS in Data Science Program.

III. Administration

A. A core group of faculty, called the MS in Data Science Program Oversight Committee, shall oversee the program and coordinate efforts with the departments.

B. Composition

1. The Program Oversight Committee is chaired by the Director, or by the Associate Director in the Director's absence.
2. The Program Oversight Committee consists of five (5) members (including the Director and Associate Director), all of whom are members of the Program Faculty.
3. Three (3) faculty from CSE and two (2) faculty from ECE departments shall be on the Oversight Committee. Faculty with joint appointments in multiple departments shall specify the one department they represent.

C. Duties

1. The duties of the Director include
   - a. providing overall academic and administrative leadership for the program,
   - b. overseeing the development and implementation of program policies,
   - c. representing the interests of the program to the College, the Campus and University administrators,
   - d. calling and chairing meetings of the program,
   - e. managing the program's budgets,
   - f. ensuring the accuracy of publications related to the program including web pages and catalog copy, and
   - g. coordinating the program's teaching needs with the teaching assignments of the constituent departments.
2. The duties of the Associate director include
   - a. serving as the Graduate Advisor for the MS in Data Science program,
   - b. coordinating administration with the Office of Graduate Studies,
   - c. submitting course change or approval forms, and

d. assisting the Director as needed.

D. Appointments

1. The Dean of BCOE appoints the Director with consultation from the Program Faculty, in a manner consistent with the appointment of other program directors and department chairs. The Director reports to the BCOE Dean.
2. It is expected that Directors should alternate between the two departments. Any exception will require a majority vote of the Oversight Committee.
3. Director appointments are for three (3) years, except when circumstances require otherwise.
4. Members of the Oversight Committee, other than the Director, are nominated and elected by the Program Faculty, in accordance with the provisions of bylaw III.B above.
5. The Associate Director will be appointed by the Director from the membership of the Oversight Committee.

IV. Meetings

A. The Program Faculty

1. The Program Faculty will meet as necessary, but at least once a year.
2. Three or more faculty from the Program Faculty can call a meeting.

B. The Program Oversight Committee

1. The Program Oversight Committee will meet at least once per academic term, on a schedule set by the Director.
2. Three or more faculty from the Program Oversight Committee can call a meeting.

C. Members will be notified of meetings at least a week in advance.

D. A quorum for meetings of the Program Faculty consist of 50% of the Program Faculty.

E. A quorum for meetings of the Program Oversight Committee consist of 4 members of the Program Oversight Committee.

APPENDIX B: NEW COURSE SYLLABI

CS/EE 251A : Data Analytics and Exploration

Spring 2021

Instructor: Mariam Salloum / Vagelis Papalexakis
Contact Info: msalloum@cs.ucr.edu / epapalex@cs.ucr.edu

Credits / Type

4.0 Units
Lecture: 3 hours
Research (outside): 3 hours

Description:

This course covers important algorithms relevant to the lifetime of data from data collection and cleaning to integration, data mining and analytics. Topics include: sketch algorithms for computing statistics on data streams; mining social graphs, including community detection and graph partitioning; Data Science lifecycle and techniques on data cleaning, data integration, Exploratory Data Analysis, and visualization.

Prerequisite(s): CS141, CS100, Stat 155 or EE114 or equivalent.

Relevant Textbooks

- (abbreviated MMD) Mining of Massive Datasets by Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman
- (abbreviated EDA) Experimental Design and Analysis by Howard J. Seltman. 2018
- Selected papers (See assigned readings in the schedule)

Grading:

- Homework (x5) 35% (assignments include both a written and programming component)
- Midterm (x2) 40%
- Final Project 25%

Tentative Schedule

Week	Lecture Topics	Readings (Book/Papers)
1	Probability Review. Distributions (single and multi-dimensional), Central-limit Theorem, Expectation, Mean, Variance, Moments,	https://www.stat.cmu.edu/~hseltman/309/Book/chapter3.pdf (EDA book Ch.3)
2	Mining Data Streams: sampling, filtering (e.g. bloom filters), sketch algorithms	http://infolab.stanford.edu/~ullman/mmds/ch4.pdf (MMDS book Ch. 4)

3	Sketch Algorithms Cont. (Count-Min and Heavy Hitters)	Notes from http://theory.stanford.edu/~tim/s15/l/l2.pdf
4	Mining Social Graphs: Intro to Social Networks as Graphs, Clustering (distance measures, Betweenness and Girvan-Newman alg.)	http://infolab.stanford.edu/~ullman/mmds/ch10n.pdf (MMDS book Ch. 10)
5	Mining Social Graphs Cont.: Community detection and Graph Partitioning (Finding Clique, Bipartite Graphs, Partitioning)	NA
6	Data Science lifecycle & Exploratory Data Analysis & Ethics of Big Data	http://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf (EDA book Ch. 4-7) Voosen, P., Big-Data Scientists Face Ethical Challenges After Facebook Study. The Chronicle of Higher Education. Retrieved from https://www.chronicle.com/article/Big-Data-Scientists-Face/150871
7	Data Visualization (including topics such as dimensionality reduction, tSNE)	(t-SNE) https://lvdmaaten.github.io/publications/papers/JMLR_2008.pdf (UMAP) https://arxiv.org/pdf/1802.03426.pdf
8	Data cleaning techniques	http://dc-pubs.dbs.uni-leipzig.de/files/Rahm2000DataCleaningProblemsand.pdf
9	Data Integration methods & ETL (schema matching, record-linkage, data fusion)	Big Data Integration by Xin Luna Dong and Divesh Srivastava
10	Data Integration methods cont.	NA

CS/EE 251B: Fundamentals of Data Science

Spring 2021

Instructor: Samet Oymak, Christian Shelton
Contact Info: oymak@ece.ucr.edu, cshelton@cs.ucr.edu

Credits and type

4.0 Units
Lecture: 3 hours
Research (outside): 3 hours

Course Information

A. Course Description

Explores theoretical tools in data science and their applications in data data science. Introduces and motivates statistical and computational viewpoints on data analysis. Topics include the manipulation of data as vectors, drawing inferences from data as distributions, and quantifying data uncertainty for data analysis. Also includes in-class and homework exercises on practical applications of these theoretical data science tools.

B. Prerequisite(s) Math 010A, Math 031 or EE020, CS100, Stat 155 or EE114 or equivalent, or permission by instructor

Syllabus

Week 1

Data as a vector I: motivation for linear algebra in data science, norms of vectors and matrices, eigenvalues and eigenvectors, fundamental subspaces

Week 2

Data as a vector II: Hermitian and positive semidefinite matrices, singular values, QR decomposition, principal component analysis (PCA), low-rank approximation

Week 3

Data analysis with linear algebra: least-squares, pseudo-inverse, condition number, ridge regression, in-class exercise on MNIST dataset and PCA

Week 4

Data as a distribution I: motivation for statistics and probability in data science, the randomness in data, random variable, conditional probability, expectation, variance, moments

Week 5

Data as a distribution II: covariance matrices, correlation coefficient, data normalization, multivariate Gaussians, law of large numbers, in-class exercise on analyzing covariance matrices on the Adult dataset

Week 6
Inference with data: Parameter estimation, unbiased estimator, bias-variance decomposition, maximum likelihood estimator (MLE), maximum a posteriori estimation (MAP), log likelihood

Week 7
Applications of Estimation: in-class exercises on MLE in clinical data, minimum mean-square error (MMSE), prediction with least-squares, coefficient of determination, in-class exercise on MMSE in time series prediction

Week 8
Quantifying uncertainty with data: hypothesis testing, confidence intervals, p-value, Student's t-test, bootstrapping, in-class exercise on hypothesis testing on the movie ratings

Week 9
Optimization with data: the role of data in modern optimization problems, loss functions, convexity, gradient, in-class exercise on gradient descent and least-squares on the Adult dataset

Week 10
Overflow: Finish the material from earlier weeks or practice for the final exam.

Textbooks and Related Materials
Recommended sources:
1. Textbook for linear algebra topics: Gilbert Strang, “Linear Algebra and Learning from Data”, Wellesley-Cambridge Press, 2019.
2. Textbook for statistical topics: Larry Wasserman, “All of Statistics: A Concise Course in Statistical Inference”, Springer, 2013. [available as a free pdf from https://link.springer.com/content/pdf/10.1007%2F978-0-387-21736-9.pdf]

Grading TBD
Participation 5%
HWs 40% (mix of coding projects and problem solving on paper)
Midterm 25%
Final 30%

CS/EE 279 : Capstone Project in Data Science

Fall 2022

Instructor: Mariam Salloum
Contact Information: msalloum@cs.ucr.edu

Credits/Type
4.0 Units
Lecture: 3 hours
Research (outside): 3 hour

Short Description (<= 50 words)

Covers combining technical, analytic, and interpretive skills to design and execute a large-scale data science capstone project that has a focus on real-world applications. Provides an opportunity to integrate all of the core skills and concepts learned throughout the program and prepares students for long-term professional success in the field. Emphasizes collaboration and communication in both written and oral form.

Prerequisites: Enrollment in Master in Data Science.
Co-requisites: CS/EE 251A, CS/EE 251B, CS224, CS226, CS235, CS236.

Course Objectives

At the end of this course, students will be able to demonstrate their knowledge, skills and abilities to develop and execute a data science project using real-world data and effectively communicate their results to a technical and non-technical audience.

Students will be able to:

- Formulate a research question, problem or hypothesis that can be answered or tested using real-world data;
- Collect and manage data to devise solutions to their research question, problem or hypothesis;
- Select, apply and evaluate models, tools and methods to address their research question, problem or hypothesis. This includes building an end-to-end analysis pipeline covering data sourcing, cleaning/preparation, integration and transformation, and visualization;
- Interpret and assess their results and evaluate the limitations of their findings;
- Prepare a professional report of their work and effectively communicate their findings to a technical and non-technical audience.

Grading

Students will work on a quarter-long project in teams of 2-3 students. The grading rubric is focused on group and individual project representations, project report, and a final web-based deliverable. In addition to these assignments, students are evaluated based on their participation in class discussions, and by their group-mates based on contributions to the group.

- 5% - Class participation (class discussions) and weekly meetings with course instructor
- 50% - Project Deliverables
  - Proposal (due Week 2) - Project proposal
  - Phase 1 (due Week 4) - Code and status report
  - Phase 2 (due Week 7) - Code and status report
  - Phase 3 (due Finals Week) - Code, and final report
- 35% - In-class presentations - instructor evaluation and peer feedback on presentations
- 10% - Web-based final deliverable

Readings

There is no textbook for this course. Readings are drawn from various relevant books, articles and academic papers that are available online.

Schedule

Week 1 - Introduction

Topics

- Reviewing the data science life-cycle
- Case studies of organizations using “big data” effectively
- Project and group selection

Required Readings

- Steve LaValle, Eric Lesser, Rebecca Shockley, Michael S. Hopkins and Nina Kruschwitz, Big data, analytics, and the path from insights to value. MIT Sloan Management Review (2010). https://sloanreview.mit.edu/projects/analytics-the-new-path-to-value/
- Robert, Christian. 2020. “The 9 Pitfalls of Data Science.” *Chance* 33. [UCR Library]
- (optional) Press. G. (2013). A very short history of data science. Forbes. Retrieved from https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#7cfe337655cf

Week 2 - Data Science Applications

Topics

- Data anonymity
- Selected readings from DS applications, focused on areas such as Social Media Analysis, Social and Information Networks, Healthcare and Medicine

Required Readings

- Anna Monreale, Gennady Andrienko, Natalia Andrienko, Fosca Giannotti, Dino Pedreschi, Salvatore Rinzivillo , Stefan Wrobel. Movement Data Anonymity through Generalization, Transactions on Data Privacy, 2010. http://www.tdp.cat/issues/tdp.a045a10.pdf
- Selected readings from papers highlighted in https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9912/9874
- (optional) Voosen, P., Big-Data Scientists Face Ethical Challenges After Facebook Study. The Chronicle of Higher Education. Retrieved from https://www.chronicle.com/article/Big-Data-Scientists-Face/150871

Week 3 - Data Science Ethics

Topics

- Skills for collecting, storing, sharing and analyzing data derived from human subjects including data used in algorithms and examining ethical implications.

Required Readings

- 1 - Data Skeptic
  - O'Neil, Cathy. (2013) On being a data skeptic. p. 1-19. Sebastopol, CA: O'Reilly Media. http://www.oreilly.com/data/free/files/being-a-data-skeptic.pdf boyd, danah. (2017)
  - Ethical side effects of the publish or perish system: p-hacking and small sample size Gelman, Andrew and Loken, Eric. (2014) The statistical crisis in science. American Scientist. Accessed online: https://www.americanscientist.org/issues/id.16259,y.2014,no.6,content.true,page.1,css.pri nt/issue.aspx
- 2- Data Sharing / Ethics
  - Libby Bishop (2017). Big data and data sharing: Ethical issues. UK Data Service, UK Data Archive. Accessed online: https://bigdata.ukdataservice.ac.uk/media/604711/big-data-and-data-sharing_ethicalissues.pdf
- 3- Building fair systems/ Ethics
  - Toward accountability: Data, Fairness, Algorithms, Consequences. Data and Society: Points. [blog post] Accessed online: https://points.datasociety.net/toward-accountability-6096e38878f0

    - ○ Crawford, Kate. (2013, April 1) The hidden biases in big data. Harvard Business Review. https://hbr.org/2013/04/the-hidden-biases-in-big-data Lerman, Jonas. (2013, September)
    - ○ Big data and its exclusions. Stanford Law Review. Accessed online: https://www.stanfordlawreview.org/online/privacy-and-big-data-big-data-and-its-exclusion

Week 4 - Group Presentations I

Topics

- Group Presentations

Required Readings

- NA

Week 5 - Communication and Storytelling

Topics

- Power of storytelling and narrative
- Tactics for presenting and sharing information

Required Readings

- Selected readings from Interactive storytelling : 7th International Conference on Interactive Digital Storytelling, ICIDS 2014, Singapore, Singapore, November 3-6, 2014 : proceedings

Week 6 - Data Visualizations

Topics

- How people and organizations process information and make decisions
- Use of data visualization for communication

Required Readings

- Selected readings from 2019 IEEE Visualization in Data Science (VDS) IEEE Visualization in Data Science (Conference) (2019 : Vancouver, B.C.)
- Gavett, G. (2014) How data visualization answered of retail's most vexing questions. Harvard Business Review. Retreived from https://hbr.org/2014/05/how-data-visualization-answered-one-of-retails-most-vexing-questions

Week 7 - Group Presentations II

Topics

- Group Presentations

Required Readings

- NA

Week 8 - New trends / topics in Data Science

Topics

- Highlight current research work in data science

Required Readings

- Selected readings from KDD, ICML, VLDB, IEEE Big Data

Week 9 - *Guest presentations*

Topics

- Guest speaker will discuss their experience in industry

Required Readings

- NA

Week 10 - Final Group Presentations and Deliverables

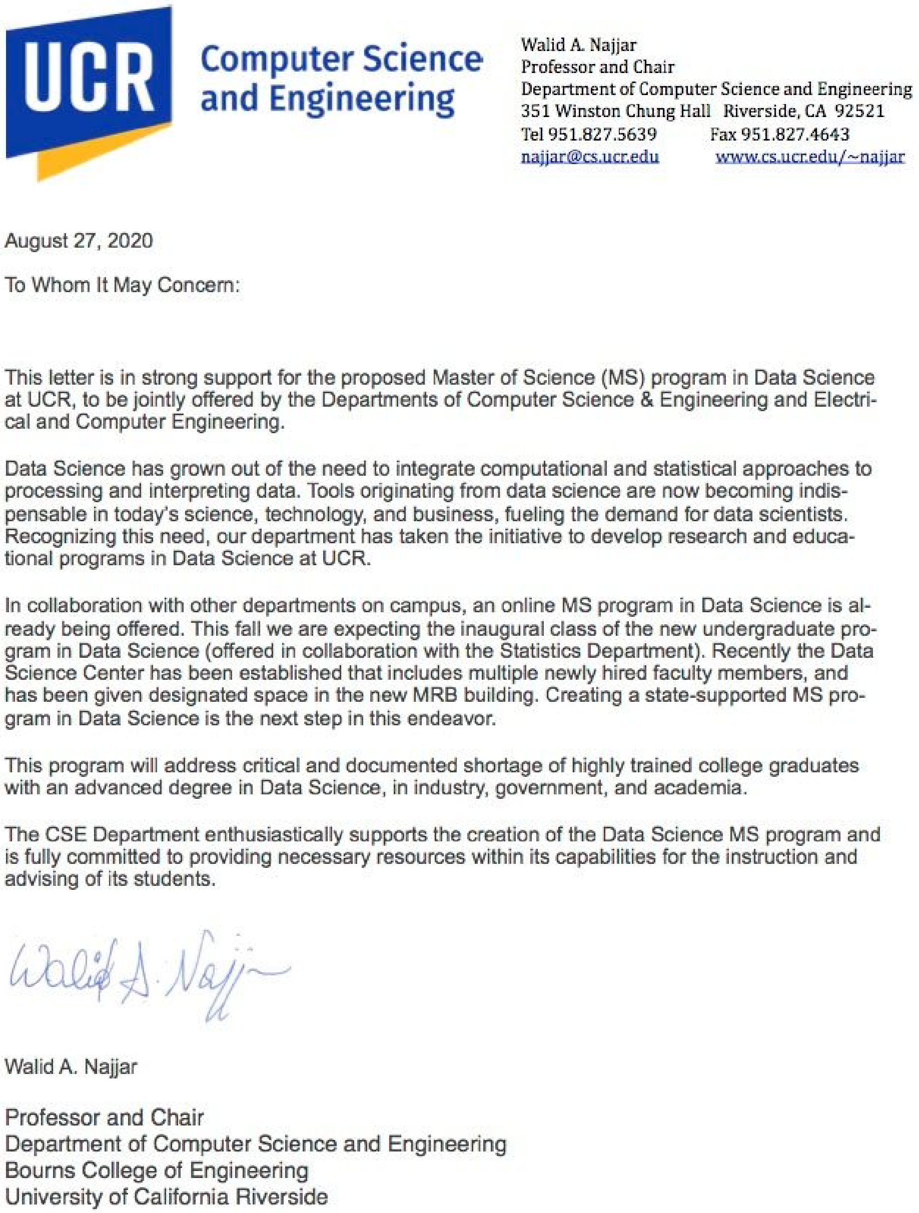
Topics

- Prepare for final in-class group presentations
- Deliver final presentations and submit project deliverables

Required Readings

- None

APPENDIX C: LETTERS OF SUPPORT



UCR Computer Science and Engineering

Walid A. Najjar
Professor and Chair
Department of Computer Science and Engineering
351 Winston Chung Hall Riverside, CA 92521
Tel 951.827.5639 Fax 951.827.4643
najjar@cs.ucr.edu www.cs.ucr.edu/~najjar

August 27, 2020

To Whom It May Concern:

This letter is in strong support for the proposed Master of Science (MS) program in Data Science at UCR, to be jointly offered by the Departments of Computer Science & Engineering and Electrical and Computer Engineering.

Data Science has grown out of the need to integrate computational and statistical approaches to processing and interpreting data. Tools originating from data science are now becoming indispensable in today's science, technology, and business, fueling the demand for data scientists. Recognizing this need, our department has taken the initiative to develop research and educational programs in Data Science at UCR.

In collaboration with other departments on campus, an online MS program in Data Science is already being offered. This fall we are expecting the inaugural class of the new undergraduate program in Data Science (offered in collaboration with the Statistics Department). Recently the Data Science Center has been established that includes multiple newly hired faculty members, and has been given designated space in the new MRB building. Creating a state-supported MS program in Data Science is the next step in this endeavor.

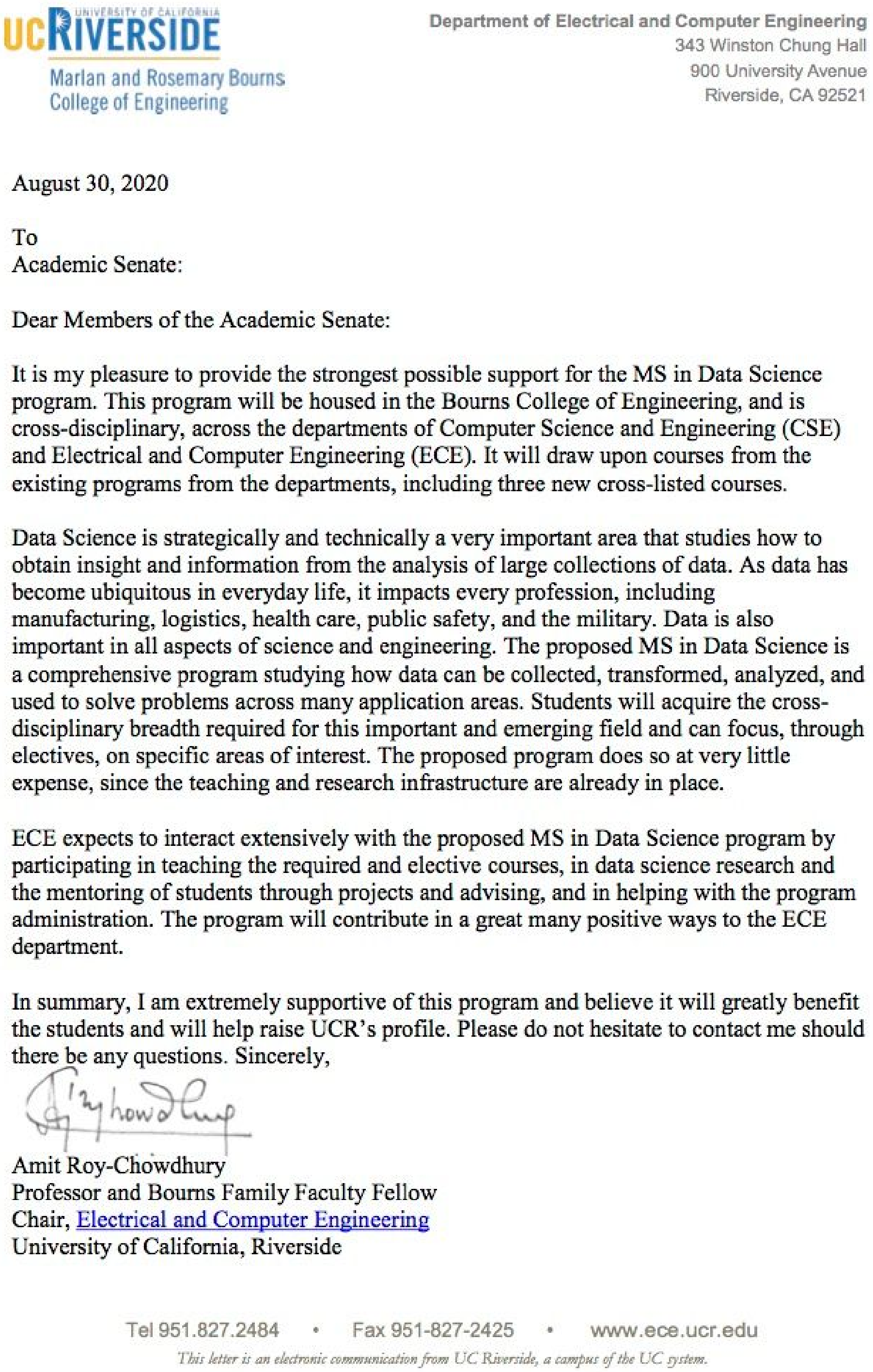
This program will address critical and documented shortage of highly trained college graduates with an advanced degree in Data Science, in industry, government, and academia.

The CSE Department enthusiastically supports the creation of the Data Science MS program and is fully committed to providing necessary resources within its capabilities for the instruction and advising of its students.

Walid A. Najj

Walid A. Najjar

Professor and Chair
Department of Computer Science and Engineering
Bourns College of Engineering
University of California Riverside



UNIVERSITY OF CALIFORNIA
UCRIVERSIDE
Marlan and Rosemary Bourns
College of Engineering

Department of Electrical and Computer Engineering
343 Winston Chung Hall
900 University Avenue
Riverside, CA 92521

August 30, 2020

To
Academic Senate:

Dear Members of the Academic Senate:

It is my pleasure to provide the strongest possible support for the MS in Data Science program. This program will be housed in the Bourns College of Engineering, and is cross-disciplinary, across the departments of Computer Science and Engineering (CSE) and Electrical and Computer Engineering (ECE). It will draw upon courses from the existing programs from the departments, including three new cross-listed courses.

Data Science is strategically and technically a very important area that studies how to obtain insight and information from the analysis of large collections of data. As data has become ubiquitous in everyday life, it impacts every profession, including manufacturing, logistics, health care, public safety, and the military. Data is also important in all aspects of science and engineering. The proposed MS in Data Science is a comprehensive program studying how data can be collected, transformed, analyzed, and used to solve problems across many application areas. Students will acquire the cross-disciplinary breadth required for this important and emerging field and can focus, through electives, on specific areas of interest. The proposed program does so at very little expense, since the teaching and research infrastructure are already in place.

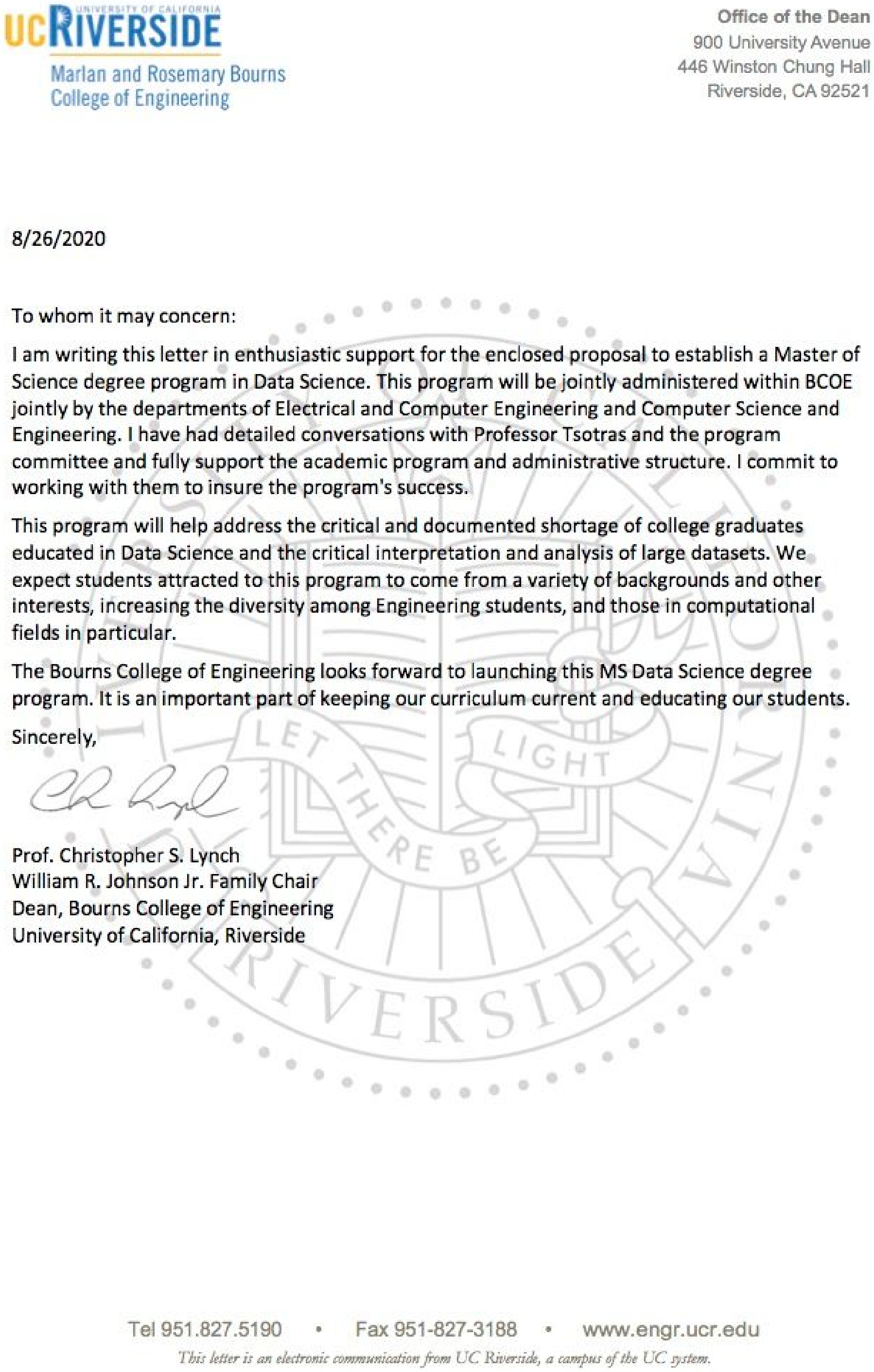
ECE expects to interact extensively with the proposed MS in Data Science program by participating in teaching the required and elective courses, in data science research and the mentoring of students through projects and advising, and in helping with the program administration. The program will contribute in a great many positive ways to the ECE department.

In summary, I am extremely supportive of this program and believe it will greatly benefit the students and will help raise UCR's profile. Please do not hesitate to contact me should there be any questions. Sincerely,

Amit Roy-Chowdhury
Professor and Bourns Family Faculty Fellow
Chair, Electrical and Computer Engineering
University of California, Riverside

Tel 951.827.2484 • Fax 951-827-2425 • www.ece.ucr.edu

This letter is an electronic communication from UC Riverside, a campus of the UC system.



UNIVERSITY OF CALIFORNIA
UCRIVERSIDE
Marlan and Rosemary Bourns
College of Engineering

Office of the Dean
900 University Avenue
446 Winston Chung Hall
Riverside, CA 92521

8/26/2020

To whom it may concern:

I am writing this letter in enthusiastic support for the enclosed proposal to establish a Master of Science degree program in Data Science. This program will be jointly administered within BCOE jointly by the departments of Electrical and Computer Engineering and Computer Science and Engineering. I have had detailed conversations with Professor Tsotras and the program committee and fully support the academic program and administrative structure. I commit to working with them to insure the program's success.

This program will help address the critical and documented shortage of college graduates educated in Data Science and the critical interpretation and analysis of large datasets. We expect students attracted to this program to come from a variety of backgrounds and other interests, increasing the diversity among Engineering students, and those in computational fields in particular.

The Bourns College of Engineering looks forward to launching this MS Data Science degree program. It is an important part of keeping our curriculum current and educating our students.

Sincerely,

Prof. Christopher S. Lynch
William R. Johnson Jr. Family Chair
Dean, Bourns College of Engineering
University of California, Riverside

Tel 951.827.5190 • Fax 951-827-3188 • www.engr.ucr.edu

This letter is an electronic communication from UC Riverside, a campus of the UC system.