UNIVERSITY OF SAN FRANCISCO | College of Arts and Sciences

MS in Data Science

Table of Contents

UNIVERSITY OF SAN FRANCISCO | College of Arts and Sciences

MS in Data Science

Name of the Program

MS in Data Science

Degree Type

Graduate

Contact Information

Diane Woodbridge
Program Academic Director/Associate Professor
MS in Data Science
dwoodbridge@usfca.edu

Mission Statement

The mission of our program is to produce graduates who possess a theoretical and practical understanding of many classical and modern statistical modeling and machine learning techniques; who use contemporary programming languages and technologies to scrape, clean, organize, query, summarize, visualize, and model large volumes and varieties of data; and who use their knowledge and skills to successfully solve real-world data-driven business problems and to communicate those solutions effectively.

Notes: The MSDS program's mission statements were ratified by its faculty during a January 2016 vote over email

Program Learning Outcome

Our program learning outcomes (PLOs) were changed in December of 2018 and ratified by a vote of the faculty to the following:

PLO 1. Possess a theoretical understanding of classical statistical models (e.g., generalized linear models, linear time series models, etc.), as well as the ability to apply those models effectively.

PLO 2. Possess a theoretical understanding of machine learning techniques (e.g., random forests, neural networks, naive Bayes, k-means, etc.), as well as the ability to apply those techniques effectively.

PLO 3. Effectively use modern programming languages (e.g., R, Python, SQL, etc.) and technologies (AWS, Hive, Spark, Hadoop, etc.) to scrape, clean, organize, query, summarize, visualize, and model large volumes and varieties of data.

PLO 4. Prepare for careers as data scientists by solving real-world, data-driven, business problems with other data scientists, and understand the social, ethical, legal, and policy issues that increasingly challenge and confront data scientists.

PLO 5. Develop professional communication skills (e.g., presentations, interviews, email etiquette, etc.), and begin integrating with the Bay Area data science community.

Current Curricular Map

There has been a few changes in our curriculum.

1. Swapping the sequences of MSDS 626 and MSDS 694 - MSDS 626 was offered in intersession, and MSDS 697 was offered in Spring 2. Since there are no dependencies between the two courses, the changes did not affect the rearrangement or curriculum updates of any other courses.
2. Introducing MSDS 699: Machine Learning Lab - We newly designed and introduced the course as machine learning skill sets and practices are most wanted for students' careers.
3. Introducing MSDS 689: Data Structures and Algorithms - To help students understand fundamental computer science topics and pass technical interviews and coding challenges, we introduced this course to cover core concepts for writing efficient programs.

Program Phase	Bootcamp				Fall Module One					Fall Module Two						Intersession	Spring Module One					Spring Module Two					Summer 2			
	MSDS 501: Computation for Analytics	**MSDS 502**: Review of Linear Algebra	**MSDS 504**: Review of Probability and Statistics	**MSDS 593**: EDA and Visualization	**MSAN 691**: Relational Databases	**MSDS 601**: Linear Regression Analysis	**MSDS 692**: Data Acquisition	**MSDS 610**: Communications for Analytics	**MSDS 640**: Seminar Series I	**MSDS 621**: Introduction to Machine Learning	**MSDS 699**: Machine Learning Laboratory	**MSDS 604**: Time Series Analysis	**MSDS 694**: Distributed Computing	**MSDS 605**: Practicum I	**MSDS 641**: Seminar Series II	**MSDS 629**: Experiments in Data Science	**MSDS 697**: Distributed Data Systems	**MSDS 630**: Advanced Machine Learning	**MSDS 689**: Data Structures and Algorithms	**MSDS 625**: Practicum II	**MSDS 642**: Seminar Series III	**MSDS 633**: Ethics in Data Science	**MSDS 626**: Case Studies in Data Science	**MSDS 603**: Product Analytics	**MSDS 627**: Practicum III	**MSDS 643**: Seminar Series IV	**MSDS 631**: Special Topics in Analytics	**MSDS 627**: Practicum IV	**MSDS 644**: Seminar Series V	**10 required hours of interview training**
Program Phase	**Bootcamp**				**Fall Module One**					**Fall Module Two**						**Intersession**	**Spring Module One**					**Spring Module Two**					**Summer 2**			
Units	1	1	1	1	1	2	2	1	0	2	1	2	1	1	0	2	2	2	1	2	0	1	2	2	2	0	2	1	0	0
Program Learning Outcome 1. Possess a theoretical understanding of classical statistical models (e.g., generalized linear models, linear time series models, etc.), as well as the ability to apply those models effectively.			I	D		I						D		I		M				D			D		M		M	M		M
Program Learning Outcome 2. Possess a theoretical understanding of machine learning techniques (e.g., random forests, neural networks, naive Bayes, k-means, etc.), as well as the ability to apply those techniques effectively of data.						I				D	I	I		I			D	M			D		I	D	D	D		M	M		
Program Learning Outcome 3. Effectively use modern programming languages (e.g., R, Python, SQL, etc.) and technologies (AWS, Hive, Spark, Hadoop, etc.) to scrape, clean, organize, query, summarize, visualize, and model large volumes and varieties of data.	I	I		I	D		D			D		I	M	I			D	M	D	D				D	M	D		M			
Program Learning Outcome 4. Prepare for careers as data scientists by solving real-world, data-driven, business problems with other data scientists, and understand the social, ethical, legal, and policy issues that increasingly challenge and confront data scientists.				D	I		I	I						D		D					D		M	D	D	M			M		M
Program Learning Outcome 5. Develop professional communication skills (e.g., presentations, interviews, email etiquette, etc.), and begin integrating with the Bay Area data science community.								I	I					I	D						D	M	I		D	M	M		M	M	D

Assessment Schedule between APRs

We are assessing 5 PLOs throughout the 3-year plan.

Year	Assessment Schedule
Year 1	Assessing PLO 1 based on statistics courses including MSDS 502, 504, 602, 604, and 623
Year 2	Assessing PLO 2, 3 and 4 based on computational courses including MSDS 621, 630, 689, 691, 692, 694, 697, and 699
Year 3	Assessing PLO 5 based on communication and business courses including MSDS 610 and 603

This report presents the assessment of PLO 5. The last assessment was done in 2018-2019.

Description of the Assessment Methodology

Four faculty (Cody Caroll, Robert Clements, Sundardas Dorai-Raj, and Uri Schonfeld) in MSDS contributed representative questions from one or more written exams taken from three data science courses and practicum related to PLO5.

These learning outcomes are loosely described as professional communication and behavioral skills.

Here are the number of questions broken down by program learning objectives and courses:

Module	**Course**	**PLO5**
Fall 1	MSDS 610: Communications for Analytics	5
Fall 2	MSDS 605: Practicum I	-
Spring 1	MSDS 625: Practicum II	-
Spring 2	MSDS 603: Data Science Entrepreneurship	2
	MSDS 633: Ethics in Data Science	9
	MSDS 627: Practicum III	-
Summer 1	MSDS 632: Practicum IV	1
	Total	

Note:

The practicum courses (MSDS 605, 625, 627, and 632) were evaluated through a survey conducted by the practicum company and faculty mentors at the end of the year.

The assessment report does not include the Seminar Series (MSDS 640, 641, 642, and 644) as it is a 0-unit credit P/F course, and not everyone is required to participate in the Q&A sessions.

All exam questions used for this assessment report can be found in the Appendix section.

Rubrics

We decided on a scoring mechanism that could be used across all courses and all questions on those exams used in this report:

Score	Description
3	Student has mastered material necessary to answer a specific question
2	Student did not give a perfect answer to the question but had a solid grasp on the concepts
1	Student did not achieve a minimum level of competency for a specific question
0	Student did not answer

For the purposes of this assessment report, faculty returned to their exams, sometimes months after initially grading the students, to select and score assessment questions as 3, 2, or 1.

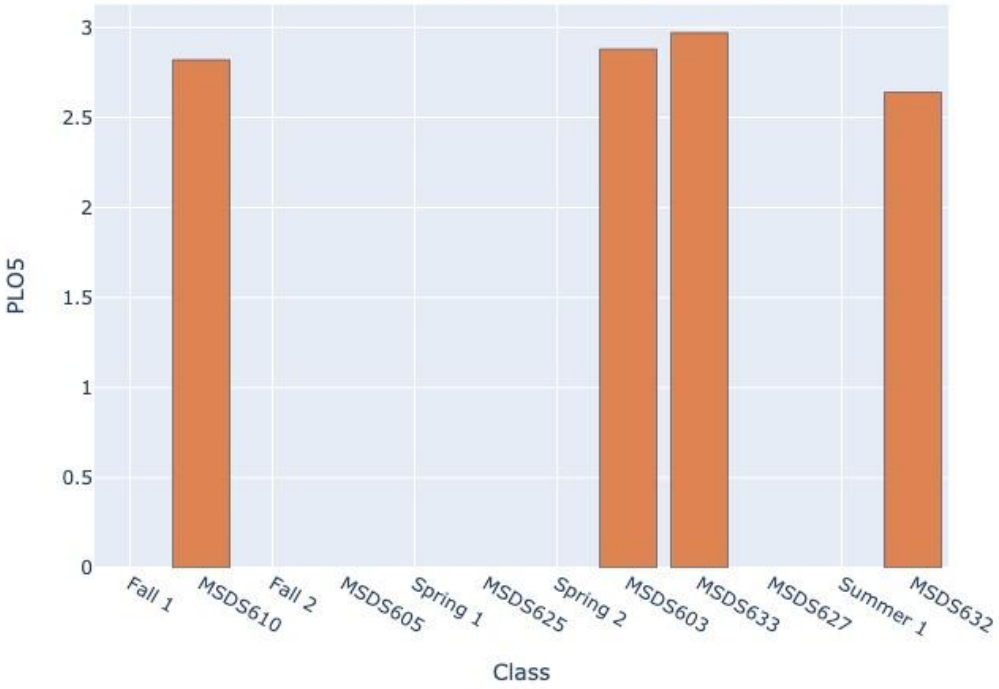
Description of Results

The averages of PLO5 for all seven courses ranged from 2.64 to 2.97, where each module had 2.82, 2.93, and 2.64, respectively, showing that students achieved close to “mastery levels.”

Note that four practicum courses (MSDS 605: Practicum I, MSDS 625: Practicum II, MSDS 627: Practicum III, and MSDS 632: Practicum IV) were evaluated at the end of the year using the survey of practicum company mentors. While this represents students’ professionalism and capability to complete tasks at work, each mentor has different expectations and standards and might not best represent outcomes with objective standards. Based on our experience, requesting and receiving feedback for every module is not possible, as mentors often ignore messages. While we check in with companies on a need basis, we only request final evaluations at the end of the year with a request to submit a practicum interest form for the next academic year.

PLO5. Develop professional communication skills (e.g., presentations, interviews, email etiquette, etc.), and begin integrating with the Bay Area data science community.

To evaluate PLO5, we selected assignments that involve communication through presentations, interviews, code reviews, written reports, social media posts (LinkedIn), and surveys from mentors at the practicum companies. Since many courses and the program maintain high standards of professionalism, students consistently received high scores and demonstrated continuous improvement throughout the year.

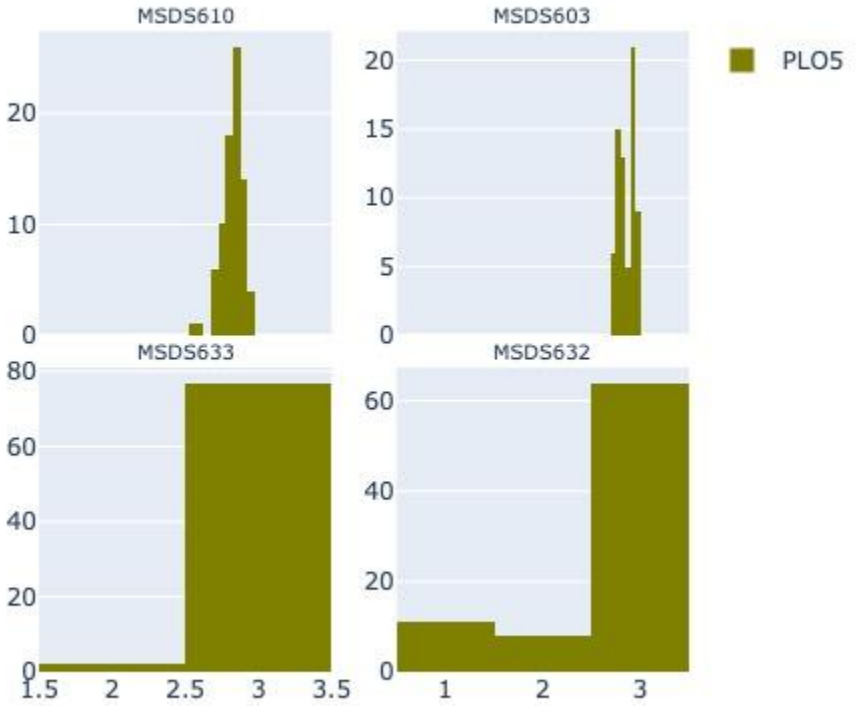


Description of Sharing and Responses

In most classes, students achieved between a solid to a mastery level and showed improvement over the year. We also investigated the distribution of grades per student for PLO5. In most classes, the grades were normally or close to normally distributed.

For practicum (MSDS632), some students received 1, which is “Student did not achieve a minimum level of competency for a specific question”. We assume that this is related to the timeline of the survey, which was towards the end of the program when the students were actively looking for a job. Due to the mass layoffs in the tech industry, our students had a very stressful last module to look for a job. Unfortunately, this led some of the international students to engage less with their practicum and caused dissatisfaction on the company side with students' performance. In addition, many practicum companies this year have chosen to be fully remote or hybrid, and the communication between students and companies often got delayed and affected the students’ engagement.

To improve professionalism and meet the expectations of the practicum companies, the program now encourages companies to provide in-person options and requires students who work for fully remote companies to be in person at least once a week on campus under the supervision of practicum supervisors.



Level	Percentage of Students (2019)
Complete Mastery of the outcome	8.7%
Complete Mastery of the outcome	20.3%
Mastered some parts of the outcome	66%
Did not master the outcome at the level intended	5%

In 2019, the program offered only two courses (MSDS610 and MSDS603) meeting PLO5 However, we were able to include an additional course (MSDS633) that aligns with PLO5. In addition, our practicum faculty director (Robert Clements) designed a survey and received feedback from mentors at the practicum company. Moreover, when comparing the 2019 assessment results, we observed a significant improvement in students' communication and professionalism.

According to Fortune, even though the number of job postings in the tech industry has reached a record low this year for the third consecutive year, MSDS managed to achieve an 80% employment rate by the end of October (Note: Our students were awarded their degrees in August.). The median base salary was $132K, with a maximum base salary of $189.5K. Since data science interviews involve multiple stages, including behavioral and technical rounds that require strong communication skills, the high employment rate directly reflects the outcome of PLO5. The faculty and staff are dedicated to assisting alumni in their job search by providing resume reviews, mock interviews, and career workshops.

Discussion

The cohort assessed this year in the MSDS program largely achieved a high standard, with solid to mastery grades. This success can be attributed to the efforts of our faculty in providing high-quality education and adapting the curriculum to equip our students with the skills that the industry demands.

The evaluation of PLO5 was conducted by four faculty members from courses that emphasize communication and professionalism. Each instructor was asked to compile the questions and anonymize the students' answers in a shared folder for review. The final report was also shared with all other faculty members in the program to ensure an accurate assessment of learning objectives and outcomes. Going forward, the program has decided to incorporate a set of questions/assignments in each course to assess student learning.

The feedback from AY 2019 reviews mentioned that the Program Learning Outcomes (PLOs) lack the use of active verbs, which makes them difficult to assess. Unfortunately, due to the loss of many faculty members, we have not had the opportunity to update our PLOs. However, we are planning to address this issue in AY 2023 by restructuring the course subjects and reviewing the PLOs.

Furthermore, we will explore better metrics and methods for assessing students, including the development of new rubrics that can provide a more detailed evaluation of student learning.

During our curriculum meeting, we will strategize ways to balance the number of questions in each PLO to promote a more balanced learning experience and evaluate student progress. We may also consider consulting with FDCD, the Center for Teaching Excellence (CTE), and Educational Technology Services (ETS) to align the PLOs with each course.

While we have a better faculty coverage this year, we still have a shortage of tenured faculty members, with only two currently in such positions. This places additional burdens on junior faculty members who have to take on administrative duties. Additionally, the program faces resource constraints, especially in terms of space for our students to study and work on their remote practicum projects.

While the overall salary in the tech industry increases, the faculty salary has not been increased accordingly. We believe that faculty salaries should be higher than the median base salary of our new graduates to ensure retention. - Many faculty members in the program left the industry for their passion, but this affects their quality of life and the lives of their families which can affect retention.

Unlike faculty members on the main campus who have access to the gym and reasonably priced parking, our faculty and students do not have these amenities. This can pose challenges for recruitment and retention.

Appendix

MSDS 603: Data Science Entrepreneurship

The assessment was done through the following assignments and milestones.

- Milestone #3/ Homework #3 - Customer interview reports
- Milestone #5/ Homework #5 - In-class presentations

MSDS 605: Practicum I, MSDS 625: Practicum II, MSDS 627: Practicum III, and MSDS 632: Practicum IV

The assessment was done based on the survey answered by practicum company mentors.

- Survey Question: Did the student exceed, meet, or barely meet expectations in regard to professionalism, communication, and work effort?

MSDS 610: Communications for Analytics

The assessment was done through the following assignments.

- Creating a Professional LinkedIn
- Oral Presentations
- Code Demo
- Professionalism during the class
- Final Project - Explain a technical DS topic & give presentation

MSDS 633: Ethics in Data Science

- Case Study and written report

  In 2016, two Danish social science researchers used data scraping software developed by a third collaborator to amass and analyze a trove of public user data from approximately 68,000 user profiles on the online dating website OkCupid. The purported aim of the study was to analyze “the relationship of cognitive ability to religious beliefs and political interest/participation” among the users of the site.

  However, when the researchers published their study in the open access online journal Open Differential Psychology, they included their entire dataset, without use of any deanonymizing or other privacy-preserving techniques to obscure the sensitive data. Even though the real names and photographs of the site’s users were not included in the dataset, the publication of usernames, bios, age, gender, sexual orientation, religion, personality traits, interests, and answers to popular dating survey questions was immediately recognized by other researchers as an acute privacy threat, since this sort of data is easily re-identifiable when combined with other publically available datasets.

  That is, the real-world identities of many of the users, even when not reflected in their chosen usernames, could easily be uncovered and relinked to the highly sensitive data in their profiles, using commonly available re-identification techniques. The responses to the survey questions were especially sensitive, since they often included information about users’ sexual habits and desires, history of relationship fidelity and drug use, political views, and other extremely personal information. Notably, this information was public only to others logged onto the site as a user who had answered the same survey questions; that is, users expected that the only people who could see their answers would be other users of OkCupid seeking a relationship. The researchers, of

course, had logged on to the site and answered the survey questions for an entirely different purpose—to gain access to the answers that thousands of others had given.

When immediately challenged upon release of the data and asked via social media if they had made any efforts to anonymize the dataset prior to publication, the lead study author Emil Kirkegaard responded on Twitter as follows: “No. Data is already public.” In follow-up media interviews later, he said: “We thought this was an obvious case of public data scraping so that it would not be a legal problem.” [1] When asked if the site had given permission, Kirkegaard replied by tweeting “Don’t know, don’t ask. :)” [2] A spokesperson for OkCupid, which the researchers had not asked for permission to scrape the site using automated software, later stated that the researchers had violated their Terms of Service and had been sent a take-down notice instructing them to remove the public dataset. The researchers eventually complied, but not before the dataset had already been accessible for two days.

Critics of the researchers argued that even if the information had been legally obtained, it was also a flagrant ethical violation of many professional norms of research ethics (including informed consent from data subjects, who never gave permission for their profiles to be used or published by the researchers). Aarhus University, where the lead researcher was a student, distanced itself from the study saying that it was an independent activity of the student and not funded by Aarhus, and that “We are sure that [Kirkegaard] has not learned his methods and ethical standards of research at our university, and he is clearly not representative of the about 38,000 students at AU.” The authors did appear to anticipate that their actions might be ethically controversial. In the draft paper, which was later removed from publication, the authors wrote that “Some may object to the ethics of gathering and releasing this data… However, all the data found in the dataset are or were already publicly available, so releasing this dataset merely presents it in a more useful form." [3]

Question 4.1:

What specific, significant harms to members of the public did the researchers’ actions risk? List as many types of harm as you can think of.

Question 4.2:

How should those potential harms have been evaluated alongside the prospective benefits of the research claimed by the study’s authors? Could the benefits hoped for by the authors have been significant enough to justify the risks of harm you identified above in 4.1?

Question 4.3:

List the various stakeholders involved in the OkCupid case, and for each type of stakeholder you listed, identify what was at stake for them in this episode. Be sure your list is as complete as you can make it, including all possible affected stakeholders.

Question 4.4:

The researchers’ actions potentially affected tens of thousands of people. Would the members of the public whose data were exposed by the researchers be justified in feeling abused, violated, or otherwise unethically treated by the study’s authors, even though they have never had a personal interaction with the authors? If those feelings are justified, does this show that the study’s authors had an ethical obligation to those members of the public that they failed to respect?

Question 4.5:

The lead author repeatedly defended the study on the grounds that the data was technically public (since it was made accessible by the data subjects to other OkCupid users). The author's implication here is that no individual OkCupid user could have reasonably objected to their data being viewed by any other individual OkCupid user, so, the authors might argue, how could they reasonably object to what the authors did with it? How would you evaluate that argument? Does it make an ethical difference that the authors accessed the data in a very different way, to a far greater extent, with highly specialized tools, and for a very different purpose than an 'ordinary' OkCupid user?

Question 4.6:

The authors clearly did anticipate some criticism of their conduct as unethical, and indeed, they received an overwhelming amount of public criticism, quickly and widely. How meaningful is that public criticism? To what extent are big data practitioners answerable to the public for their conduct, or can data practitioners justifiably ignore the public's critical response to what they do? Explain your answer.

Question 4.7:

As a follow-up to Question 4.5, how meaningful is it that much of the criticism of the researchers' conduct came from a range of well-established data professionals and researchers, including members of professional societies for social science research, the profession to which the study's authors presumably aspired? How should a data practitioner want to be judged by his or her peers or prospective professional colleagues? Should the evaluation of our conduct by our professional peers and colleagues hold special sway over us, and if so, why?

Question 4.8:

A Danish programmer, Oliver Nordbjerg, specifically designed the data scraping software for the study, though he was not a co-author of the study himself. What ethical obligations did he have in the case? Should he have agreed to design a tool for this study? To what extent, if any, does he share in the ethical responsibility for any harm to the public that resulted?

Question 4.9:

How do you think the OkCupid study likely impacted the reputations and professional prospects of the researchers and of the designer of the scraping software?