San Diego State University Logo

San Diego State University Logo CS577 Principles and Techniques of Data Science

Fall 2024

# COURSE INFORMATION

Class Days: W

Class Times: 7:00 pm -- 9:40 pm

Class Location: PSFA 350

Instructor: Sean Kang

Phone:

Email: skang7@sdsu.edu

Office: online only

Office Hours (and by appointment):

M – noon-2pm

W - noon-2pm & short time before or after classCourse Overview

This course covers the basic principles and techniques of data science. These include languages for transforming, querying and analyzing data; algorithms for machine learning methods including regression, classification, and clustering; principles behind creating informative data visualizations; statistical concepts of measurement error and prediction; and techniques for scalable data processing. This class explores the key areas of data science including question formation, data collection, data cleaning, visualization, statistical inference, predictive modeling, and decision making.

The goals of this course are:

* Enable students to start careers as data scientists by providing experience working with real-world data, tools, and techniques
* Empower students to apply computational and inferential thinking to tackle real-world problems.

**Course Catalog Description**

Data classification, cleaning, common representation and operations, dimensionality reduction, life cycle, regression, statistical inference, and visualization.

Effective Fall 2021, students who register for face-to-face classes are expected to attend as indicated in the course schedule. Faculty teaching face-to-face courses will not be required to create a new, alternative on-line class as an accommodation for any student.

Students with medical conditions that would present a COVID-related risk in a face-to-face instructional setting should contact the Student Ability Success Center (<https://sdsu.edu/sasc>) to begin the process of getting support. Students who do not adhere to the [Covid19 Student Policies](http://csrr.sdsu.edu/) or the directives of their faculty will be directed to leave the classroom and will be referred to the Center for Student Rights and Responsibilities.

Do not come to campus if you do not feel well. Remain home and monitor your symptoms and seek medical attention as needed.

**Student Learning Outcomes**

* Acquire and clean data to prepare it for analysis.

In assignments you will find real datasets and clean the data before analyzing it. On exams you will be asked the proper way to clean different types of data.

* Apply machine learning and statistical methods on data to make predictions and draw conclusions.

In assignments you will apply various machine learning and statistical methods on data to draw conclusions and make predictions. In exams you will be asked what methods to apply in specific cases and justify your selection.

* Utilize visualizations to aid in data analysis.

You will be using visualizations to help analysis data in labs and assignments. In exams you will be asked to explain different visualizations including their strengths and limitations.

* Apply computational and inferential thinking to tackle real-world problems.

In assignments you will need to apply computational and inferential thinking to analysis real-world problems/issue. For example, given detailed election results can you determine if there was systematic cheating or voter fraud.

* Evaluate other's data analysis/studies for validity and limitations.

During class we will look at existing studies/analysis. You will be asked to point out issues with the studies and suggest better ways of conducting the study/analysis. On exams you will be given small case study and will have to comment on the study methods.

* Use industry standard languages, notebooks, and tools to perform data analysis.

Assignments will require you to use industry standard languages and tools.

**Real Life Relevance**

Data Science is a growing field that is being used in many areas that affect people’s daily lives. Companies are increasing their use of Data Science. This course will train students to use Data Science to investigate business, scientific and social issues, and problem by collecting and analyzing data.

Relation to other Courses:

CS 200 Introduction to Data Science and Python covers some of the same material but this course covers the material in much more depth. If this courses material seems too advanced, try taking CS 200. CS649 Big Data covers some of the same topics but in the context of large amounts of data. You should take this course before taking CS 629. CS 566 Machine Learning covers the machine learning algorithms covered in this course in greater detail. It also covers more machine learning algorithms than this course.

# Enrollment Information

Prerequisites

CS 210 (formerly numbered CS 310). Math 254, Stat 250, Working knowledge of Python.

Adding/Dropping Procedures

Students will be allowed to add or drop the course in the first two weeks of the semester.

# Course Materials

| Materials | Required or optional | Where and how it can be obtained |
| --- | --- | --- |
| Principles and Techniques of Data Science, Sam Lau, Joey Gonzalez, and Deb Nolan | Optional | https://www.textbook.ds100.org |
| An Introduction to Statistical Learning, Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani | Optional | http://www-bcf.usc.edu/~gareth/ISL/ |
| Data Science from Scratch, Joel Grus, O'Reilly Media, April 2015 | Optional | SDSU Bookstore |

Student will need access to a computer running Chrome, Safari, or Edge. Students are expected to be able to run Python 3.x on their local machines.

# Course Structure and Conduct

This course is taught as a traditional lecture course. The lecture will meet once a week. You will have assignments and exams in this course. Unless indicated explicitly on an assignment you are to work by yourself. Students will do a semester project. Course material will be available online through Blackboard (or Canvas).

The course will use SDSU's learning management system (blackboard or Canvas depending on what is selected). If outside of class you have questions about the course, course content, assignments etc. use Blackboard/Canvas to ask your questions. Frequently other students will be interested in the answer to your questions. If you have questions of personal nature email the instructor directly or talk to the instructor in person.

Estimated time commitment.

As 3-unit lecture course students will spend 3 hours in lecture and between 9-12 hours outside of class per week.

How to participate in the course

This is not a do-it-yourself, work-at-your-own-speed course. The material in this course is technical and will build on each other. If you fall behind it will be difficult to catch up. So, it is important that you do the readings on time, before attending class.

Let me know how you're doing. Send me an email if you're facing difficulties and together let's see what we can work out!

Interacting with me

Questions about the course and course material should be asked either in class, in office hours or on canvas discussions. If you have question of person nature either send email or come to office hours to discuss them.

# Course Assessment and Grading

**Assignments**

In 4-6 assignments, students need to select the appropriate techniques to analysis a set of data. Students will spend time over multiple days to complete an assignment.

**Project**

A project is completely open ended. Students need to find the dataset(s) you need to investigate an issue. What can be learned from the dataset? What are the issues with the dataset? What techniques should they use? It is expected that the quality of the project report is high enough that one could submit it to a data science competition like Kaggle or used by San Diego City to justify some action. Work on the project is expected to take at least a month.

Grading Policy

Your grade in the course will be determined as follows:

Exams

Mid-term exam, and one Final

Grading Guideline

Mid-term exam 25%

Final exam 30%

Assignments 30%

Semester project 15%

Late submission policy: 10% score deduction per week.

Wrong format policy: 10% score deduction – if the submission is in the wrong format.

Excused Absence Make-up Policies. This will be handled on a case-by-case base.

No late submissions will be accepted after Dec 18, 2024.

Schedule

There will be exams in week 9, and a final on the date scheduled by the university. Students need to submit project proposals week 5 of the semester.

As a 3-unit lecture course students will spend 3 hours in lecture and between 9-12 hours outside of class per week.

Schedule of Topics

| week | Topic |
| --- | --- |
| 1  Aug 26 | Lecture 1 Over  Course introduction,  Data science life cycles |
| 2  Sept 2 | Lecture 2, 3  Assignment 1 announced  Data Sampling and Probability  Pandas  Holiday only on Monday – does not affect class |
| 3  Sept 9 | Lecture 4  More Pandas  Question and Answer |
| 4  Sept 16 | Lecture 5, 6  Assignment 1 Due  Assignment 2 Announced  More Panda  Data Cleaning, EDA |
| 5  Sept 23 | Lecture 7, 8  Data Cleaning, EDA  Anomaly Detection  Review Assignment 1 |
| 6  Sept 30 | Lecture 9, 10  Data Anomaly  Data Warehousing |
| 7  Oct 7 | Lecture 11, 12  Assignment 2 Due  Assignment 3 announced.  Visualization  Regular Expressions  **Oct 9 – Not a school holiday** |
| 8  Oct 14 | Lecture 13, 14  SQL  Modeling  Review Assignment 2 |
| 9  Oct 21 | Midterm – in class |
| 10  Oct 28 | Lecture 15 Simple Linear Regression  Lecture 16  Gradient Descent  Assignment 3 due  Assignment 4 announced  **Semester Project Snapshot Draft Due** |
| 11  Nov 4 | Lecture 17  Lecture 18  Feature Engineering  Overfitting  Review midterm and assignment 3 |

|  |  |
| --- | --- |
| 12  Nov 11 | Lecture 19  Lecture 20,  Assignment 4 due  Cross Validation  Regularization |
| 13  Nov 18 | Lecture 21  Lecture 22  Review assignment 4  Multiple Linear Regression  Logistic Regression |
| 14  Nov 25 | Holiday week |
| 15  Dec 2 | Lecture 25,26  Decision Trees  Review |
| 16  Dec 9 | Review  **Final Report due – Dec 8** |
| Dec 12-18 | Final week |

# Academic Honesty

The University adheres to a strict [policy regarding cheating and plagiarism](http://go.sdsu.edu/student_affairs/srr/cheating-plagiarism.aspx). These activities will not be tolerated in this class. Become familiar with the policy and what constitutes plagiarism. Any cheating or plagiarism will result in failing this class and a disciplinary review by the University. These actions may lead to probation, suspension, or expulsion.

Examples of Plagiarism include but are not limited to:

* Using sources verbatim or paraphrasing without giving proper attribution (this can include phrases, sentences, paragraphs and/or pages of work)
* Copying and pasting work from an online or offline source directly and calling it your own
* Using information you find from an online or offline source without giving the author credit
* Replacing words or phrases from another source and inserting your own words or phrases
* Submitting a piece of work you did for one class to another class

# Turnitin

Students agree that by taking this course all required papers may be subject to submission for textual similarity review to [Turnitin.com](http://turnitin.com/) for the detection of plagiarism. All submitted papers will be included as source documents in the Turnitin.com reference database solely for the purpose of detecting plagiarism of such papers. You may submit your papers in such a way that no identifying information about you is included. Another option is that you may request, in writing, that your papers not be submitted to www.turnitin.com. However, if you choose this option you will be required to provide documentation to substantiate that the papers are your original work and do not include any plagiarized material.

# Technical Support for Blackboard

Student support for Blackboard is provided by the Library Computing Hub, located on the 2nd floor of Love Library. They can be reached at 619-594-3189 or hub@mail.sdsu.edu

# Netiquette

Netiquette is online etiquette. It is important that all participants in courses be aware of proper online behavior and respect one another.

Use appropriate language for an educational environment:

* Use complete sentences.
* Use proper spelling and grammar.
* Avoid slang and uncommon abbreviations.
* Do not use obscene or threatening language.

Remember that the University values diversity and encourages discourse. Be respectful of differences while engaging in online discussions. Find SDSU’s netiquette guidelines at this link: <http://its.sdsu.edu/learning-management-system/student-netiquette>. For more information about Netiquette, see [The Core Rules for Netiquette](http://www.albion.com/netiquette/corerules.html) by Virginia Shea.

# Students with Disabilities

If you are a student with a disability and believe you will need accommodations for this class, it is your responsibility to contact Student Ability Success Center at (619) 594-6473. To avoid any delay in the receipt of your accommodations, you should contact the Student Ability Success Center as soon as possible. Please note that accommodations are not retroactive, and that I cannot provide accommodations based upon disability until I have received an accommodation letter from Student Ability Success Center. Your cooperation is appreciated.

**Starting Fall 2019 all accommodated exams will be booked through SASC Connect, an online portal for Student Ability Success Center. Paper booking forms will not be accepted.**

# Student Services:

A complete list of all academic support services is available on the [Academic Success](http://go.sdsu.edu/student_affairs/academic_success.aspx) section of the [SDSU Student Affairs](http://go.sdsu.edu/student_affairs) website.

For help with improving your writing ability, the staff at the SDSU [Writing Center](http://writingcenter.sdsu.edu/) is available in person and online.

[Counseling and Psychological Services](http://go.sdsu.edu/student_affairs/cps/Default.aspx) offers confidential counseling services by licensed psychologists, counselors, and social workers. More info can be found at their website or by contacting (619) 594-5220. You can also Live Chat with a counselor <http://go.sdsu.edu/student_affairs/cps/therapist-consultation.aspx> between 4:00pm and 10:00pm, or call San Diego Access and Crisis 24-hour Hotline at (888) 724-7240.

# Student Privacy and Intellectual Property

Students maintain intellectual property rights to work products they create as part of this course unless they are formally notified otherwise.

Blackboard Grade Center is used to assure privacy of student grades and feedback on individual assignments unless students have granted written waivers.

Students will be notified at the time of an assignment if copies of student work will be retained beyond the end of the semester or used as examples for future students or the wider public.

# Copyright Policy

SDSU respects the intellectual property of others and we ask our faculty & students to do the same.

It is best to assume that any material (e.g., graphic, html coding, text, video, or sound) on the Web is copyrighted unless specific permission is given to copy it under a [Creative Commons License](http://creativecommons.org). More information about the use of copy written material in education as part of the [TEACH Act](https://www.provost.ncsu.edu/copyright/toolkit/) and [Copyright Fair Use Guidelines](http://fairuse.stanford.edu/). Whenever possible, you should attribute the original author of any work used under these provisions.