

# Intro and First Day Stuff

## Lecture 1 - CMSE 381

Prof. Mengsen Zhang

Michigan State University

::

Dept of Computational Mathematics, Science & Engineering

Mon, Aug 25, 2025

# People in this lecture



**Dr. Zhang** (she/they)  
Assistant Professor, CMSE, MSU



**Siyu Guo** (He/him)  
Graduate Student, CMSE, MSU







# What is this course about?

## Topics:

- Fundamental concepts of data science
- Regression
- Classification
- Dimension reduction
- Resampling methods
- Tree-based methods, etc.

# D2L and where to find grades

<https://d2l.msu.edu/d2l/home/2339736>

🏠 FS25-CMSE-381-002 - Fundamentals of Data Scien...      Mengsen Zhang 

Course Home Content Course Tools ▾ Assessments ▾ Communication ▾ Help Course Admin More ▾

## FS25-CMSE-381-002 - Fundamentals of Data Science Methods

**Announcements ▾**

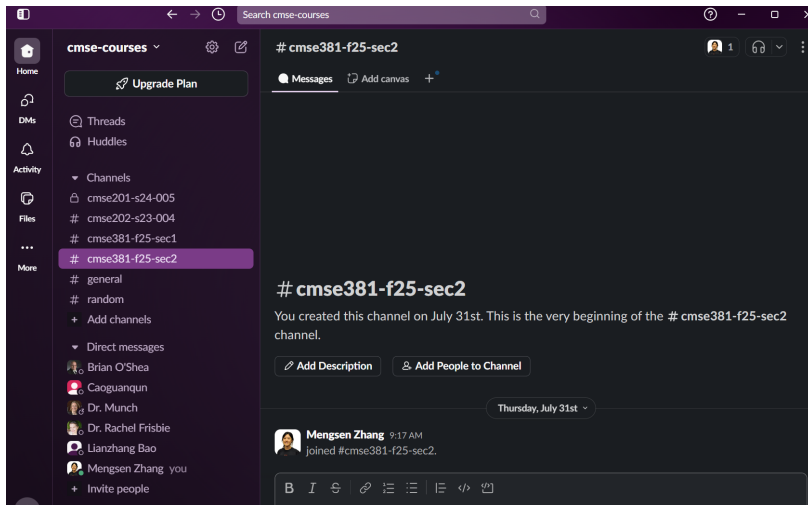
There are no announcements to display. [Create an announcement](#)

**Need Help? ▾**

MSU IT Service Desk:  
Local: (517) 432-6200

# Slack and where to find announcements/ask questions

Join cmse-courses slack: <https://tinyurl.com/cmse-courses-slack-invite>

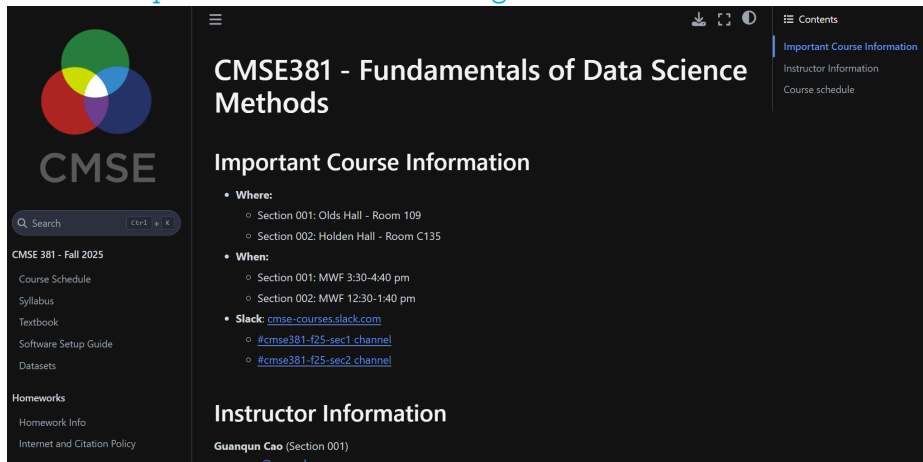


# Course Website and where to find slides and jupyter notebooks

<https://cmse.msu.edu/CMSE381>

—or—

<https://msu-cmse-courses.github.io/CMSE381-F25/>



CMSE

Search  Ctrl + K

CMSE 381 - Fall 2025

- Course Schedule
- Syllabus
- Textbook
- Software Setup Guide
- Datasets

Homeworks

- Homework Info
- Internet and Citation Policy

CMSE381 - Fundamentals of Data Science Methods

Important Course Information

- **Where:**
  - Section 001: Olds Hall - Room 109
  - Section 002: Holden Hall - Room C135
- **When:**
  - Section 001: MWF 3:30-4:40 pm
  - Section 002: MWF 12:30-1:40 pm
- **Slack:** [#cmse381-f25-sec1\\_channel](https://cmse-courses.slack.com)  
[#cmse381-f25-sec2\\_channel](https://cmse-courses.slack.com)

Instructor Information

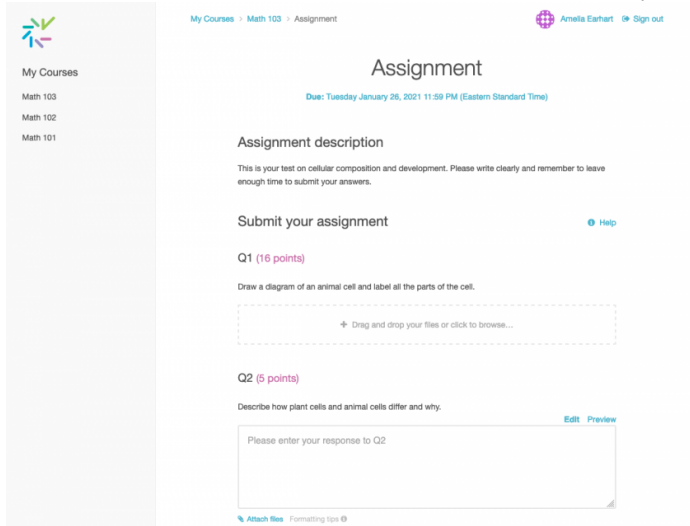
Guanqun Cao (Section 001)  
[caoguanqun@msu.edu](mailto:caoguanqun@msu.edu)

Contents

- Important Course Information
- Instructor Information
- Course schedule

# Crowdmark and where to submit homework

No URL: You will get an automated email from the system (I think.....?)



The screenshot shows the Crowdmark assignment submission page. On the left is a sidebar with the Crowdmark logo and a 'My Courses' section listing 'Math 103', 'Math 102', and 'Math 101'. The main content area has a breadcrumb trail 'My Courses > Math 103 > Assignment' and a user profile for 'Amelia Earhart' with a 'Sign out' link. The title 'Assignment' is centered, with a due date 'Due: Tuesday January 26, 2021 11:59 PM (Eastern Standard Time)'. Below this is the 'Assignment description' section, which states: 'This is your test on cellular composition and development. Please write clearly and remember to leave enough time to submit your answers.' The 'Submit your assignment' section includes a 'Help' link. The first question, 'Q1 (16 points)', asks to 'Draw a diagram of an animal cell and label all the parts of the cell.' It features a dashed box with a plus icon and the text 'Drag and drop your files or click to browse...'. The second question, 'Q2 (5 points)', asks to 'Describe how plant cells and animal cells differ and why.' It includes 'Edit' and 'Preview' links and a text input area with the placeholder 'Please enter your response to Q2'. At the bottom, there are links for 'Attach files' and 'Formatting tips'.

# Office hours

Zoom link: Look up on [calendar on the website](#)

The image shows a screenshot of the CMSE website on the left and a Google Calendar on the right. The website sidebar includes the CMSE logo, a search bar, and links to 'Course Schedule', 'Syllabus', 'Textbook', 'Datasets', 'Homeworks', 'Homework Info', and 'Internet and Citation Policy'. The Google Calendar is titled 'Google calendar for office hours' and shows a monthly view for January 2025. The calendar displays office hours for Dr. Zhang and Dr. Bao. Dr. Zhang's office hours are on Wednesdays from 10 am to 12 pm, starting on January 9th. Dr. Bao's office hours are on Thursdays from 9 am to 10 am, starting on January 10th. The calendar also shows CMSE381-S2025 events.

*Dr. Zhang*

Time: W 10 am - 12 pm  
(Starting 9/3)

Zoom

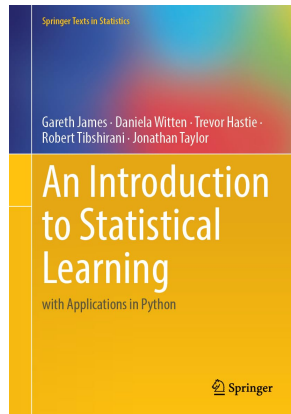
*Siyu Guo*

Time: TBD  
Zoom



**Free download**

<https://www.statlearning.com/>



# Class Structure

- Class is a combination of lecture time, and group work/coding time.
  - ▶ Bring computer every day
  - ▶ Jupyter notebooks
  - ▶ Python
- Once a week, there will be a short check-in quiz. This will be basic content related to lectures since the last class. Possible questions include checking on definitions, or basic understanding of major ideas.
  - ▶ 10 points per quiz
  - ▶ Drop two lowest grades

# Class Structure Pt 2

- Homeworks due once a week, midnight of the day marked in the schedule (mostly Sundays).
  - ▶ 20 points per homework
  - ▶ Drop two lowest grades
  - ▶ Sliding scale:
    - ★ 24 hours late: 5% penalty.
    - ★ 48 hours late: 15% penalty.
    - ★ >48 hours: No late work accepted.
- Three Midterms
  - ▶ See schedule for dates
  - ▶ 100 points each
  - ▶ Not cumulative
- One Project
  - ▶ Analyze dataset using tools in class, submit written report
  - ▶ 100 points
  - ▶ Due at the end of the semester

# Basic Expectations

- attend each class for the full 70 min duration
- take detailed notes on, or beside, the skeleton slides provided.
- complete the jupyter notebook in class.
- read the assigned textbook chapters listed in the course schedule (on course website).
- actively participate in group work and interactive Q&A sessions.
- complete all homework assignments, quizzes, exams, and a semester project.

# Approximate schedule

Up to date version: [https://msu-cmse-courses.github.io/CMSE381-F25/Course\\_Info/Schedule.html](https://msu-cmse-courses.github.io/CMSE381-F25/Course_Info/Schedule.html)

CMSE381\_F2025\_Schedule : Schedule

Lec #	Date	Topic	Reading	HW
1	M 8/25	Intro / Python Review	1	
2	W 8/27	What is statistical learning	2.1	
3	F 8.29	Assessing Model Accuracy	2.2.1, 2.2.2	
4	M 9/1	Labor Day - No Class		
5	W 9/3	Linear Regression	3.1	
6	F 9/5	More Linear Regression	3.1	HW #1 Due Sun 9/7
7	M 9/8	Multi-linear Regression	3.2	
8	W 9/10	Probably More Linear Regression	3.3	
9	F 9/12	Last of the Linear Regression		HW #2 Due Sun 9/14
10	M 9/15	Intro to classification, Bayes classifier, KNN classifier	2.2.3	
11	W 9/17	Logistic Regression	4.1, 4.2, 4.3.1-3	
12	F 9/19	Multiple Logistic Regression / Multinomial Logistic Regression	4.3.4-5	HW #3 Due Sun 9/21
13	M 9/22	Project Day & Review		
14	W 9/24	Midterm #1		
15	F 9/26	Leave one out CV	5.1.1, 5.1.2	

CMSE381\_F2025\_Schedule : Schedule

12	M 9/22	Project Day & Review		
13	W 9/24	Midterm #1		
14	F 9/26	Leave one out CV	5.1.1, 5.1.2	
15	M 9/29	k-fold CV	5.1.3	
16	W 10/1	More k-fold CV	5.1.4-5	
17	F 10/3	k-fold CV for classification	5.1.5	
18	M 10/6	Subset selection	6.1	
19	W 10/8	Shrinkage: Ridge	6.2.1	
20	F 10/10	Shrinkage: Lasso	6.2.2	HW #4 Due Sun 10/12
21	M 10/13	PCA	6.3	
22	W 10/15	PCR	6.3	
23	F 10/17	Review		
24	M 10/20	Fall Break		
25	W 10/22	Midterm #2		
26	F 10/24	Polynomial & Step Functions	7.1-7.2	HW #5 Due Sun 10/28
27	M 10/27	Step Functions; Basis functions; Start Splines	7.2-7.4	
28	W 10/29	Regression Splines	7.4	

29	F 10/17	Review		
30	M 10/20	Fall Break		
31	W 10/22	Midterm #2		
32	F 10/24	Polynomial & Step Functions	7.1-7.2	HW #5 Due Sun 10/28
33	M 10/27	Step Functions; Basis functions; Start Splines	7.2-7.4	
34	W 10/29	Regression Splines	7.4	
35	F 10/31	Decision Trees	8.1	HW #6 Due Sun 11/2
36	M 11/3	Random Forests	8.2.1, 8.2.2	
37	W 11/5	Maximal Margin Classifier	9.1	
38	F 11/7	SVC	9.2	HW #7 Due Sun 11/9
39	M 11/10	SVM	9.3, 9.4	
40	W 11/12	Single Layer NN	10.1	
41	F 11/13	Multi Layer NN	10.2	HW #8 Due Sun 11/16
42	M 11/17	CNN	10.3	
43	W 11/19	Unsupervised learning / clustering	12.1, 12.4	
44	F 11/21	Virtual: Project Office Hours		HW #9 Due Sun 11/23
45	M 11/24	Review		
46	W 11/26	Midterm #3		
47	F 11/28	Thanksgiving		
48	M 12/1	Virtual: Project Office Hours		
49	W 12/3	Virtual: Project Office Hours		
50	F 12/5			Project Due

# Grade distribution

## *Estimated Points*

Homeworks	$(9 \text{ homeworks} - 2 \text{ lowest grades}) \times 20 \text{ points} = 140$
Quizzes	$(10 \text{ Quizzes} - 2 \text{ lowest grades}) \times 10 \text{ points} = 80$
Midterm	$(3 \text{ Midterms}) \times 100 = 300$
Final Project	100
<hr/>	
TOTAL:	620 (Subject to change!)

# Section 1

Intro to class

# What is Statistical Learning?

## Statistical Learning

- Subfield of statistics
- Emphasizes models and their interpretability, precision, and uncertainty

## Machine Learning

- Machine learning has a greater emphasis on large scale applications and prediction accuracy.

*Nowadays....to sound pedantic or techie?*



# Why should you care?

Data is everywhere, getting more complicated and useful. Learning how to analyze data is critical.

- Web data, e-commerce (Amazon, JD, Alibaba)
- Car sales (Tesla, Ford, and GM)
- Sports team (MSU, Lions, etc)
- Politics and government
- Image, videos, text
- even fancier data in biomedicine

# Learning Tools as Black Boxes? Or Math Apocalypse?

- Need to understand the machinery enough to
  - ▶ know what tool to use
  - ▶ know how to interpret output of the tool
- Don't need to rebuild the entire box from scratch

## Example: Email spam

	george	you	your	hp	free	hpl	!	our	re	edu	remove
spam	0.00	2.26	1.38	0.02	0.52	0.01	0.51	0.51	0.13	0.01	0.28
email	1.27	1.27	0.44	0.90	0.07	0.43	0.11	0.18	0.42	0.29	0.01

if (%george < 0.6) & (%you > 1.5)    then spam  
   else email.

if ( $0.2 \cdot \%you - 0.3 \cdot \%george$ ) > 0    then spam  
   else email.

# Supervised learning

- Outcome measurement  $Y$  (also called dependent variable, response, target, label).
- Vector of  $p$  predictor measurements  $X$  (also called inputs, regressors, covariates, features, independent variables).
- In the regression problem,  $Y$  is quantitative (e.g price, blood pressure).
- In the classification problem,  $Y$  takes values in a set of distinct categories (survived/died, cancer class of tissue sample, types of language).

# Unsupervised learning

- No outcome variable, just a set of predictors (features) measured on a set of samples.
- Objective is fuzzier: often explore the intrinsic relation between samples (e.g., clustering) or features (e.g. dimensionality reduction)
- Difficult to know how well you are doing
- Different from supervised learning but can be useful as a pre-processing step for supervised learning.

# Generative AI discussion

Definition via [Wikipedia](#):

*Generative artificial intelligence (AI) is artificial intelligence capable of generating text, images, or other media, using generative models. Generative AI models learn the patterns and structure of their input training data and then generate new data that has similar characteristics.*

Examples:

- ChatGPT
- Bard
- DALL-E

- Get in a group of about 4.
- Open this google doc:  
[tinyurl.com/CMSE381-F25-genAI](https://tinyurl.com/CMSE381-F25-genAI)
- In your group, brainstorm cases where someone might use generative AI in the context of our class.
- Once you have added a few, start adding arguments for or against whether we should allow the use of that context in class.

## Section 2

### Python Review Lab: Pt 1

# Plan for the lab

- Find a group of 4 or so.
- Find the class website ([cmse.msu.edu/CMSE381](https://cmse.msu.edu/CMSE381)) or ([msu-cmse-courses.github.io/CMSE381-F25/](https://msu-cmse-courses.github.io/CMSE381-F25/)) and download the jupyter notebook for the Python Review Lab.
- Get started!

The screenshot displays the CMSE381 course website. On the left is a sidebar with the CMSE logo (a Venn diagram with four overlapping circles in green, red, blue, and purple) and the text 'CMSE'. Below the logo is a search bar and a list of links: 'CMSE 381 - Fall 2024', 'Course Schedule', 'Syllabus', 'Datasets', 'Lectures', and 'Day 01 (M 8/26)'. The main content area is titled 'Lecture 1 - Intro to Class and Python Review' and includes a sub-header 'Important documents' with links to 'CMSE381-Lec01-FirstDay.pdf' and 'CMSE381-Lec01-PythonReview.ipynb'. Navigation links for 'Data sets' and 'Lecture 1 - Python Review' are also visible.



# Next time

- Weds: What is statistical learning? (Reading 2.1)
- No class coming Monday (9/1)
- First HW Due Sunday, 9/7
- Quiz sometime **this** week
- Office hours:
  - ▶ Most up-to-date on the website
  - ▶ Starting next week

CMSE381\_F2025\_Schedule : Schedule

Lec #	Date		Topic	Reading	HW
1	M	8/25	Intro / Python Review	1	
2	W	8/27	What is statistical learning	2.1	
3	F	8.29	Assessing Model Accuracy	2.2.1, 2.2.2	
	M	9/1	Labor Day - No Class		
4	W	9/3	Linear Regression	3.1	
5	F	9/5	More Linear Regression	3.1	HW #1 Due Sun 9/7
6	M	9/8	Multi-linear Regression	3.2	
7	W	9/10	Probably More Linear Regression	3.3	
8	F	9/12	Last of the Linear Regression		HW #2 Due Sun 9/14
9	M	9/15	Intro to classification, Bayes classifier, KNN classifier	2.2.3	
10	W	9/17	Loaistic Rearession	4.1, 4.2, 4.3, 4.4	