

Intro and First Day Stuff

Lecture 1 - CMSE 381

Prof. Lianzhang Bao

Michigan State University

::

Dept of Computational Mathematics, Science & Engineering

Mon, Jan 12, 2026

People in this lecture



Dr. Bao (he/him)
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/2387930>

The screenshot shows the D2L course homepage for SS26-CMSE-381-001. The top navigation bar includes links for Course Home, Content, Course Tools, Assessments, Communication, Help, Course Admin, and More. The main title "SS26-CMSE-381-001 - Fundamentals of Data Science Methods" is displayed on a banner with a green and blue abstract background. Below the banner, the "Announcements" section shows a message: "There are no announcements to display. [Create an announcement](#)". The "Need Help?" section provides contact information for the MSU IT Service Desk, including local and toll-free phone numbers.

SS26-CMSE-381-001 - Fundamentals of Data Science Methods

Announcements

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

Need Help?

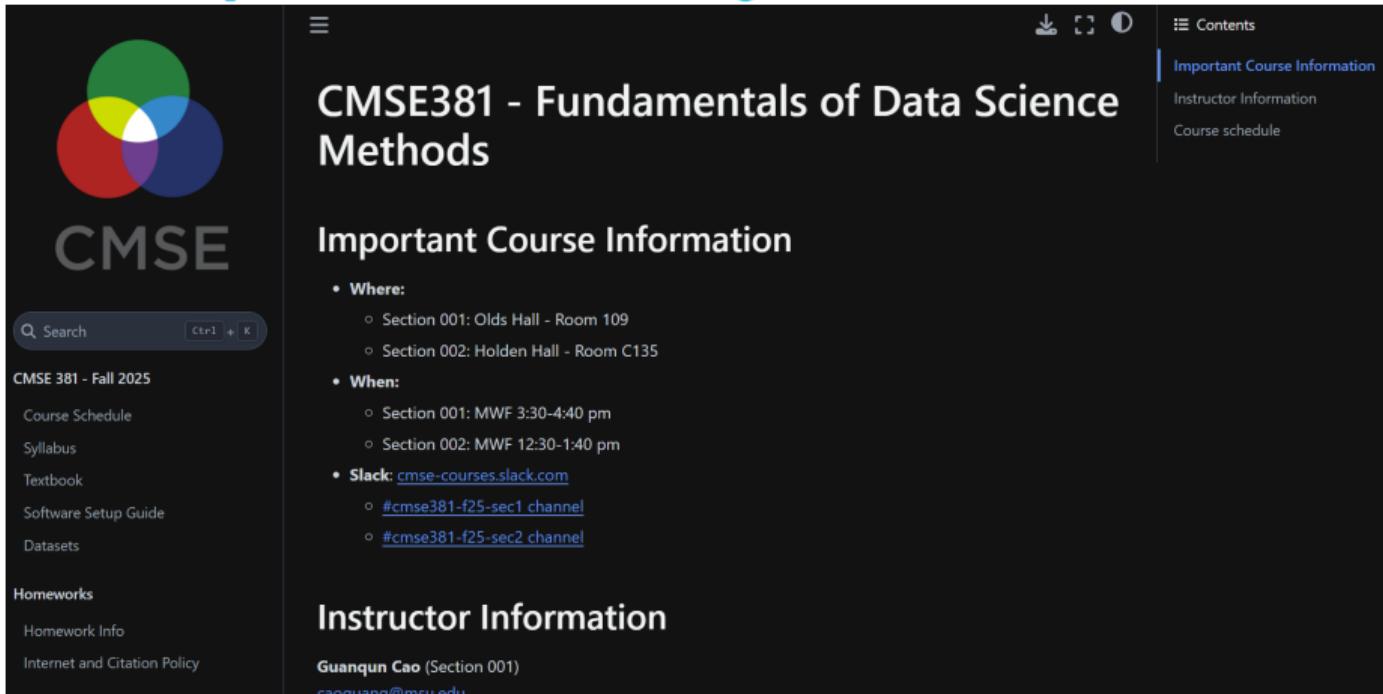
MSU IT Service Desk:
Local: (517) 432-6200
Toll Free: (844) 678-6200

Course Website and where to find slides and jupyter notebooks

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

—or—

<https://msu-cmse-courses.github.io/CMSE381-S26/>



The screenshot shows the CMSE381 course website with a dark theme. The header features a logo with three overlapping circles (red, green, blue) and the text 'CMSE'. A search bar and a 'Ctrl + K' keyboard shortcut are visible. The main content area displays the course title 'CMSE381 - Fundamentals of Data Science Methods' and a 'Important Course Information' section. This section includes a bulleted list: '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), and 'Slack' (cmse-courses.slack.com, #cmse381-f25-sec1 channel, #cmse381-f25-sec2 channel). Below this is an 'Instructor Information' section for Guanqun Cao (Section 001, gcauqun@msu.edu). The sidebar on the right contains links for 'Contents', 'Important Course Information', 'Instructor Information', and 'Course schedule'.

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: cmse-courses.slack.com
 - [#cmse381-f25-sec1 channel](#)
 - [#cmse381-f25-sec2 channel](#)

Instructor Information

Guanqun Cao (Section 001)
gcauqun@msu.edu

CMSE 381 - Fall 2025

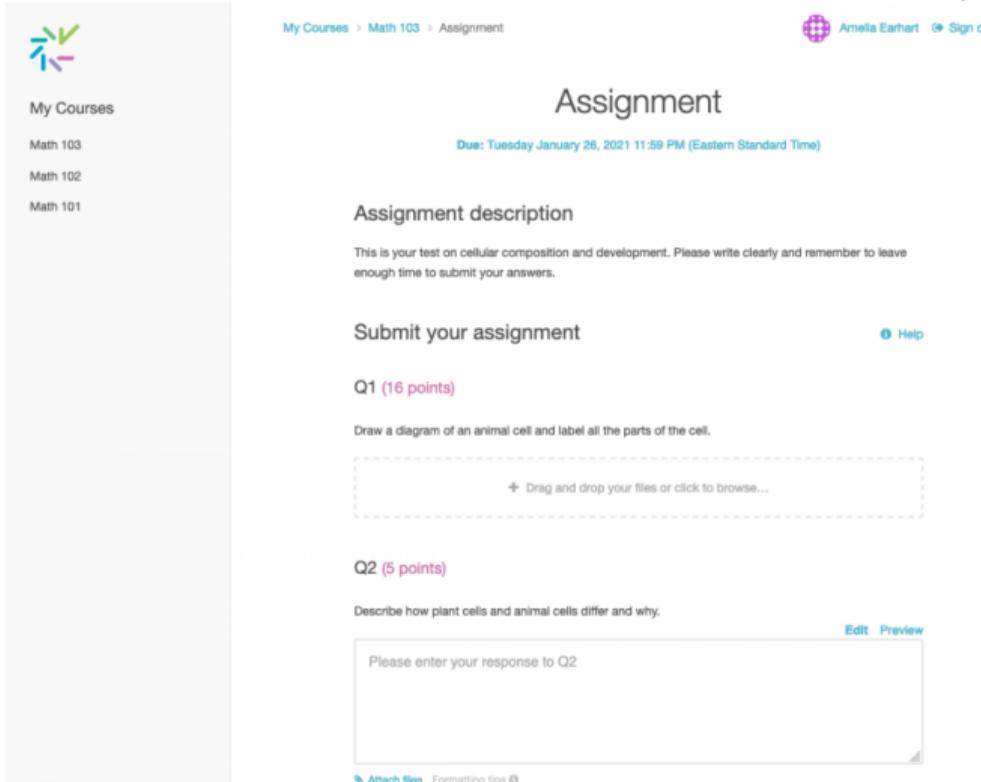
- Course Schedule
- Syllabus
- Textbook
- Software Setup Guide
- Datasets

Homeworks

- Homework Info
- Internet and Citation Policy

Crowdmark and where we grade your quizzes/midterms

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



The screenshot shows the Crowdmark assignment interface. At the top, there is a navigation bar with 'My Courses' (Math 103), 'Assignment', and a user profile for 'Amelia Earhart' with a 'Sign out' option. The main title is 'Assignment' with a due date of 'Tuesday January 26, 2021 11:59 PM (Eastern Standard Time)'. The 'Assignment description' section contains the text: 'This is your test on cellular composition and development. Please write clearly and remember to leave enough time to submit your answers.' Below this, the 'Submit your assignment' section has a 'Help' link. The first question, 'Q1 (16 points)', asks for a diagram of an animal cell and provides a file upload area with the placeholder 'Drag and drop your files or click to browse...'. The second question, 'Q2 (5 points)', asks for a description of how plant cells and animal cells differ and why, with a text input field containing 'Please enter your response to Q2'. There are 'Edit' and 'Preview' buttons above the text input field. At the bottom, there are links for 'Attach files' and 'Formatting tips'.

Office hours

The screenshot shows a dual-pane interface. On the left is the CMSE 381 Spring 2025 course schedule, featuring a CMSE logo, a search bar, and sections for Syllabus, Textbook, Datasets, Homeworks, Homework Info, and Internet and Citation Policy. On the right is a Google Calendar titled "Google calendar for office hours" for January 2025. The calendar lists specific office hours for Dr. Bao and Dr. Zhang on various dates, with Dr. Bao's office hours marked on Tuesdays and Dr. Zhang's on Wednesdays.

SUN 29	MON 30	TUE 31	WED Jan 1	THU 2	FRI 3	SAT 4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
	• 12:30pm CMSE3 • 3:30pm CMSE3i		• 12:30pm CMSE3 • 3:30pm CMSE3i		• 12:30pm CMSE3 • 3:30pm CMSE3i	
19	20	21	22	23	24	25
		• 9am Dr. Bao off	• 9am Dr. Bao off • 10am Dr. Zhang 2 more		• 12:30pm CMSE3 • 3:30pm CMSE3i	
26	27	28	29	30	31	Feb 1
	• 10am Dr. Zhang • 12:30pm CMSE3 • 3:30pm CMSE3i	• 9am Dr. Bao off	• 9am Dr. Bao off • 10am Dr. Zhang 2 more		• 12:30pm CMSE3 • 3:30pm CMSE3i	

CMSE381-S2025
Events shown in time zone: (GMT-05:00) Eastern Time - New York
Google Calendar

Dr. Bao

Time: Tues 9 am - 11 pm
(Starting 1/20)

In-person (EB 2507L)

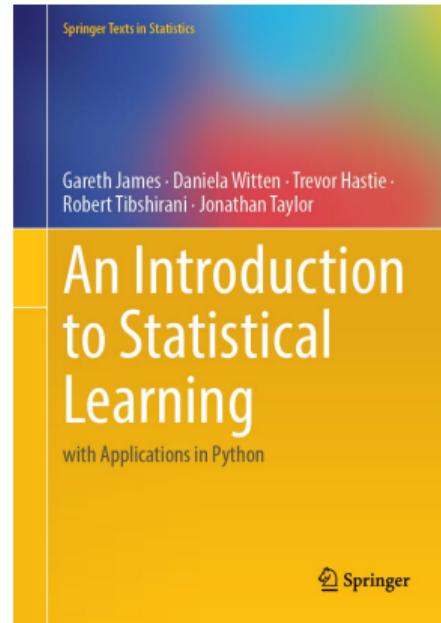
Siyu Guo

Time: Wed 2-3pm
In-person (EB 2504)

Textbook

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-S26/Course_Info/Schedule.html

Lec #	Date	Topic	Reading	HW	Pop Quizzes	Notes
1	M 1/12	Intro / Python Review	1			
2	W 1/14	What is statistical learning	2.1		Q1	
3	F 1/16	Assessing Model Accuracy	2.2.1, 2.2.2			
4	M 1/19	MLK - No Class				
5	W 1/21	Linear Regression	3.1			
6	F 1/23	More Linear Regression	3.1	HW #1 Due Sun 1/25		
7	M 1/26	Multi-linear Regression	3.2			
8	W 1/28	Probably More Linear Regression	3.3			
9	F 1/30	Last of the Linear Regression				
10	M 2/2	Intro to classification, Bayes classifier, KNN classifier	2.2.3	HW #2 Due Sun 2/1		
11	W 2/4	Logistic Regression	4.1, 4.2, 4.3.1-3			
12	F 2/6	Multiple Logistic Regression / Multinomial Logistic Regression	4.3.4-5	HW #3 Due Sun 2/8		
	M 2/9	Project Day & Review				
	W 2/11	Midterm #1				
12	F 2/13	Leave one out CV	5.1.1, 5.1.2			

12	F	2/13	Leave one out CV	5.1.1, 5.1.2			
13	M	2/16	k-fold CV	5.1.3			
14	W	2/18	More k-fold CV	5.1.4-5			
15	F	2/20	k-fold CV for classification	5.1.5			
16	M	2/23	Subset selection	6.1			
17	W	2/25	Shrinkage: Ridge	6.2.1			
18	F	2/27	Shrinkage: Lasso	6.2.2	HW #4 Due Sun 3/1		
	M	3/2	Spring Break				
	W	3/4	Spring Break				
	F	3/6	Spring Break				
19	M	3/9	PCA	6.3			
20	W	3/11	PCR	6.3			
	F	3/13	Review				
	M	3/16	Midterm #2				
21	W	3/18	Polynomial & Step Functions	7.1-7.2			
22	F	3/20	Step Functions; Basis functions; Spline Splines	7.2-7.4			
23	M	3/23	Regression Splines	7.4			

23	M	3/23	Regression Splines	7.4			
24	W	3/25	Decision Trees	8.1	HW #6 Due Wed 3/25	Q7	
25	F	3/27	Random Forests	8.2.1, 8.2.2	HW #7 Due Sun 3/29	Q8	
26	M	3/30	Maximal Margin Classifier	9.1			
27	W	4/1	SVC	9.2			
28	F	4/3	SVM	9.3, 9.4	HW #8 Due Sun 4/5	Q9	
29	M	4/6	Single Layer NN	10.1			
30	W	4/8	Multi Layer NN	10.2			
31	F	4/10	CNN	10.3	HW #9 Due Sun 4/12	Q10	
32	M	4/13	Unsupervised learning / clustering	12.1, 12.4			
33	W	4/15	Virtual: Project Office Hours				
	F	4/17	Review				
	M	4/20	Midterm #3				
	W	4/22					Project Due
	F	4/24					
					No final exam		

Grade distribution

<i>Estimated Points</i>	
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

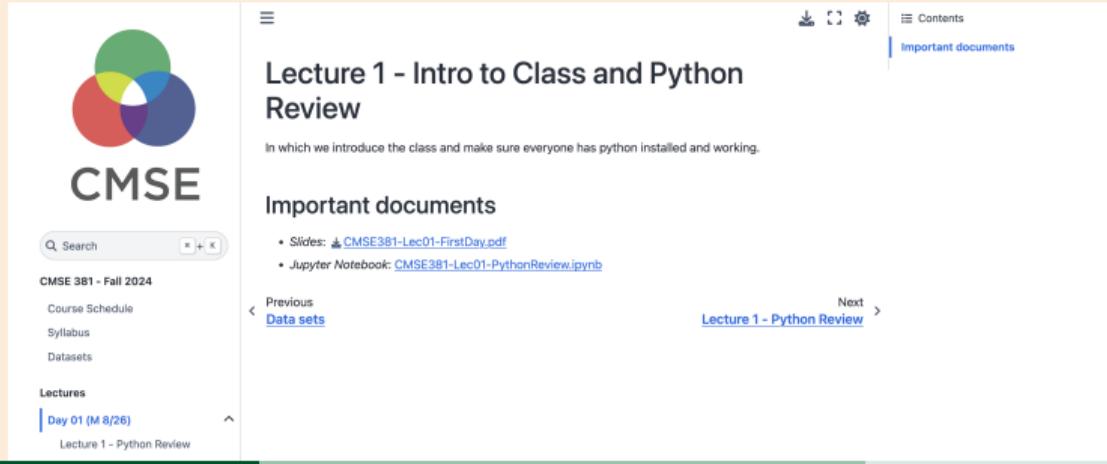
- Get in a group of about 4.
- Open this google doc:
tinyurl.com/CMSE381-S26-genAI
- In your group, brainstorm cases where someone might use generative AI in the context of our class.
- Once you have added a few, start adding arguments for or against whether we should allow the use of that context in 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) or (msu-cmse-courses.github.io/CMSE381-S26/) and download the jupyter notebook for the Python Review Lab.
- Get started!



The screenshot shows the CMSE 381 class website. The header features the CMSE logo (three overlapping circles in green, red, and blue) and the text 'CMSE'. The main content area is titled 'Lecture 1 - Intro to Class and Python Review'. Below the title, a sub-header reads: 'In which we introduce the class and make sure everyone has python installed and working.' A section titled 'Important documents' lists 'Slides: CMSE381-Lec01-FirstDay.pdf' and 'Jupyter Notebook: CMSE381-Lec01-PythonReview.ipynb'. Navigation links include 'Previous Data sets' and 'Next Lecture 1 - Python Review'. The sidebar on the left includes links for 'Course Schedule', 'Syllabus', 'Datasets', 'Lectures' (with 'Day 01 (M 8/26)' selected), and 'Lecture 1 - Python Review'.

Next time

- Weds: What is statistical learning?
(Reading 2.1)
- First HW Due Sunday, 1/25
- Quiz sometime **this** week
- Office hours:
 - ▶ Most up-to-date on the website
 - ▶ Starting next week

Lec #	Date	Topic	Reading	HW	Pop Quizzes	Notes
1	M 1/12	Intro / Python Review	1			
2	W 1/14	What is statistical learning	2.1		Q1	
3	F 1/16	Assessing Model Accuracy	2.2.1, 2.2.2			
	M 1/19	MLK - No Class				
4	W 1/21	Linear Regression	3.1		Q2	
5	F 1/23	More Linear Regression	3.1			
6	M 1/26	Multi-linear Regression	3.2		HW #1 Due Sun 1/25	
7	W 1/28	Probably More Linear Regression	3.3		Q3	
8	F 1/30	Last of the Linear Regression				
9	M 2/2	Intro to classification, Bayes classifier, KNN classifier	2.2.3		HW #2 Due Sun 2/1	
10	W 2/4	Logistic Regression	4.1, 4.2, 4.3.1-3		Q4	
11	F 2/6	Multiple Logistic Regression / Multinomial Logistic Regression	4.3.4-5	HW #3 Due Sun 2/18		