Lecture 1

Course introduction

COURSE MATH 637: Mathematical Techniques in Data Science

WEBSITE https://jacobcalvert.com/teaching/math637fall2026/

INSTRUCTOR Jacob Calvert (calvert@udel.edu)

OFFICE HOURS MW 2:45pm–3:45pm, Ewing Hall 509

Plan for today

• Introductions, questionnaire	10 min
• Course overview	30 min
• Data science life cycle	10 min
• Plan for next class, questions	5 min

About me

Photo: Doug Merriam

Academic:

- Developing probability theory to answer basic questions about nature
- PhD from Berkeley Stats

Industry:

- Nine years at clinical data science startups
- Mentored dozens of data scientists at all levels, from variety of STEM backgrounds

Other interests: Movies, books, tennis, coffee

Questionnaire

Plan for today

- Introductions, questionnaire 10 min
- Course overview 30 min
- Data science life cycle 10 min
- Plan for next class, questions 5 min

What is data science?

> **Yu and Barter (2024):** Data science is not simply a subfield of statistics or computer science. Instead, it is the integration of statistical and computational thinking into real-world domain problems in science, technology, and beyond.

What is data science?

> **Yu and Barter (2024):** Data science is not simply a subfield of statistics or computer science. Instead, it is the integration of statistical and computational thinking into real-world domain problems in science, technology, and beyond.

Other answers include:

What is data science?

> **Yu and Barter (2024):** Data science is not simply a subfield of statistics or computer science. Instead, it is the integration of statistical and computational thinking into real-world domain problems in science, technology, and beyond.

Other answers include:

- Data science is a rebranding of statistics.

What is data science?

> **Yu and Barter (2024):** Data science is not simply a subfield of statistics or computer science. Instead, it is the integration of statistical and computational thinking into real-world domain problems in science, technology, and beyond.

Other answers include:

- Data science is a rebranding of statistics.
- Data science is statistics on a Mac. (Taken from @bigdataborat.)

What is data science?

> **Yu and Barter (2024):** Data science is not simply a subfield of statistics or computer science. Instead, it is the integration of statistical and computational thinking into real-world domain problems in science, technology, and beyond.

Other answers include:

- Data science is a rebranding of statistics.
- Data science is statistics on a Mac. (Taken from @bigdataborat.)
- Data science is what data scientists do.

Overview

This course will introduce fundamental techniques for data exploration, prediction, model selection, and interpretation.

> **Main objective:** Develop a mathematical understanding of data science techniques—their assumptions, properties, and limitations—alongside their practical implementation and use.

Beyond lectures:

- Several meetings dedicated to solving problems and programming
- Final projects will evaluate course material and the critical thinking skills at the core of effective data science

Representative list of topics

Dimension reduction	Singular value decomposition, principal component analysis, autoencoders
Density estimation	Gaussian mixture models, variational autoencoders
Regression	Least squares, MLE, MAP, regularization, Bayesian linear regression
Classification	k-Nearest neighbors, logistic regression, decision trees, random forests, support vector machines
Neural networks	Multilayer perceptron, backpropagation, stochastic gradient descent
Model assessment	Cross-validation, calibration
Frontier topics	Conformal prediction, explainability, tabular foundation models

Representative list of topics: "pillars"

Dimension reduction	Singular value decomposition, principal component analysis, autoencoders
Density estimation	Gaussian mixture models, variational autoencoders
Regression	Least squares, MLE, MAP, regularization, Bayesian linear regression
Classification	k-Nearest neighbors, logistic regression, decision trees, random forests, support vector machines
Neural networks	Multilayer perceptron, backpropagation, stochastic gradient descent
Model assessment	Cross-validation, calibration
Frontier topics	Conformal prediction, explainability, tabular foundation models

Four types of techniques, briefly

Dimension reduction
Given a set of p-dimensional data points, find a mapping of the data to a k-dimensional space ($k < p$) that roughly preserves its "structure."
(E.g., for visualization when $k \leq 3$.)

Density estimation
Estimate the common distribution from which a set of data points were sampled.
(E.g., to generate new samples.)

Regression
Given pairs of examples $\boldsymbol{x}_n$ and *continuous* labels y_n, learn a function $f(\boldsymbol{x}) \approx y$.
(E.g., for prediction.)

Classification
Like regression, but with *discrete* labels.

Tentative schedule

Topic:	# Meetings:
• Intro	2
• Dimension reduction, density estimation basics	6
• Regression basics	6
• Classification basics	4
• Advanced techniques for each “pillar”	14
• Frontier topics	5
• Project work, presentations	4
	—
	= 41

Prerequisites

- Probability theory and basic statistics at the level of MATH 350
- Multivariable calculus at the level of MATH 243
- Linear algebra at the level of MATH 349
- Basic programming skills, ideally in Python

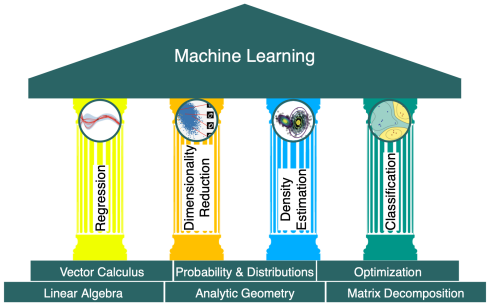


Fig. 1.1 of Deisenroth, Faisal, and Ong (2020).

Assessment

Component:	Share of grade:
• In-class quizzes	15%
• Homework (theory and programming)	25%
• Midterm (**friendly!**)	30%
• Final project report	15%
• Final project presentation, Q&A	15%

> **Note:** We'll use the default Canvas scale: A $\geq 94\%$, A– $\geq 90\%$, B+ $\geq 87\%$, B $\geq 84\%$, B– $\geq 80\%$, C+ $\geq 77\%$, etc. Any changes will be to your benefit.

Quizzes and homework

Quizzes: short, given regularly, on recent lectures and assignments.

- No make-ups. Instead, three lowest scores will be dropped.

Homework: ≈ 10 assignments over the semester.

- Two 72-hour extensions, but tell me before the deadline passes.
- The two lowest scores will be dropped.

> **Note:** We will regularly use our meetings to solve problems and program together. Homework extends that work, and quiz and midterm questions draw on both.

Final project

- Individually or in small groups; we will settle on appropriate scope
- Written report and a presentation with Q&A
- Brief proposal and progress check come earlier in the term

Key idea: Assessed on how well you justify your choices at each step, especially in the Q&A.

Use of AI: Only as a *programming* aid, and only on homework and the project. Not for theory problems, prose, your reasoning, or modeling judgment. Declare any use.

Rationale for assessment

- Data scientists continue to be accountable for modeling decisions and conclusions drawn from analysis

Rationale for assessment

- Data scientists continue to be accountable for modeling decisions and conclusions drawn from analysis
- To get (and keep) a data science career, you must be able to defend your judgment calls, because there is often no “right” answer

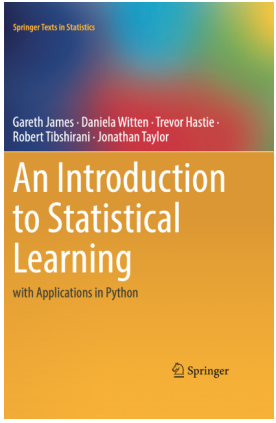
Rationale for assessment

- Data scientists continue to be accountable for modeling decisions and conclusions drawn from analysis
- To get (and keep) a data science career, you must be able to defend your judgment calls, because there is often no “right” answer
- The assessments in the course are intended to reflect this reality:
  AI can help you implement your data science process, but you are still accountable for it!

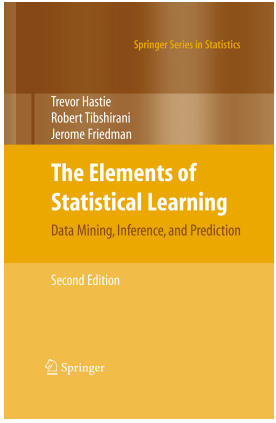
Learning resources

- We’ll draw on many books and online resources.
- All the books we will refer to are freely accessible.

Core textbooks

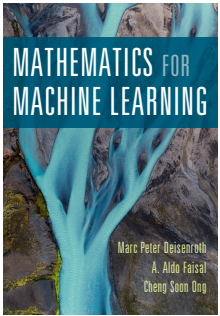


https://www.statlearning.com

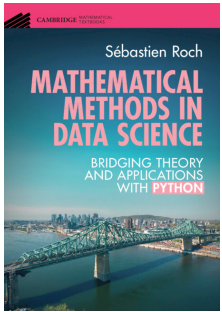


https://hastie.su.domains/ElemStatLearn/

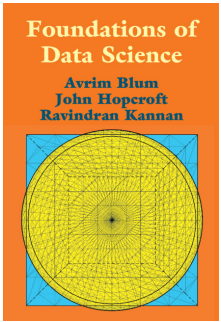
We'll also draw on ...



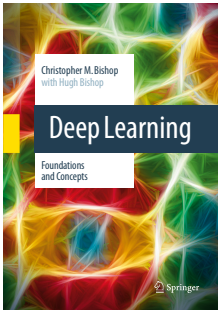
https://mml-book.github.io/



https://mmids-textbook.github.io/

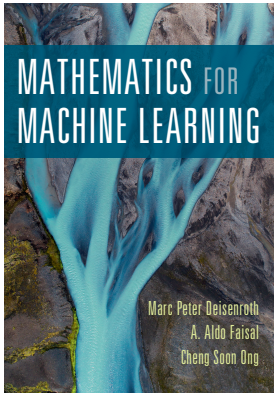


https://home.ttic.edu/~avrim/book.pdf

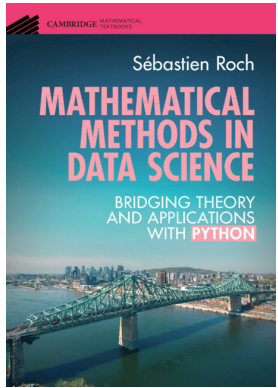


https://www.bishopbook.com/

“I need a math refresher.”

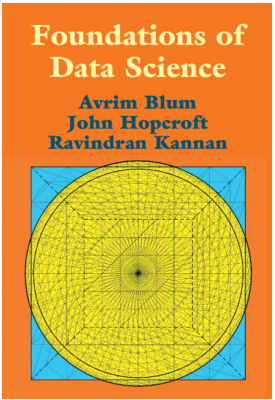


https://mml-book.github.io/



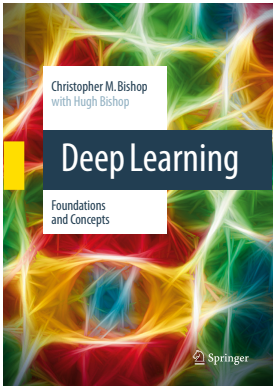
https://mmids-textbook.github.io/

“Less data science, please.”



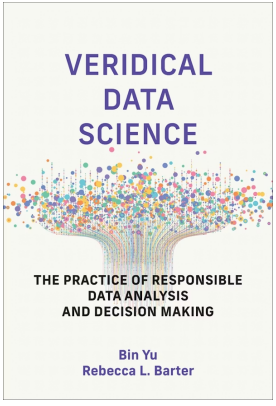
https://home.ttic.edu/~avrim/book.pdf

“I wish this were (more of) an ML course!”



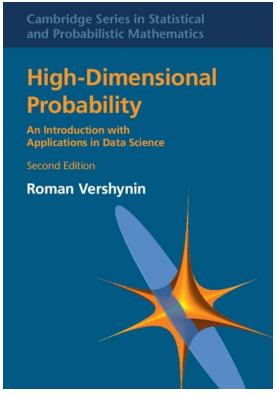
https://www.bishopbook.com/

Bonus: "Give me a practical DS philosophy."



https://vdsbook.com/

Bonus: A companion for probabilists



https://webapps.math.uci.edu/~rvershyn/papers/HDP-book/HDP-2.pdf

Plan for today

- Introductions, questionnaire 10 min
- Course overview 30 min
- Data science life cycle 10 min
- Plan for next class, questions 5 min

Data science life cycle

Fig. 2.1 of VDS Pt. I.

Data science life cycle: example of sepsis prediction

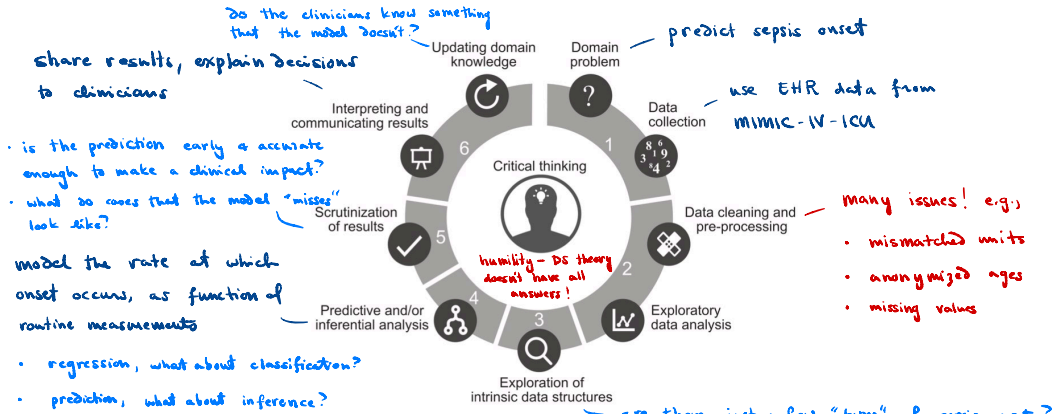


Fig. 2.1 of VDS Pt. I.

Data science life cycle: course coverage

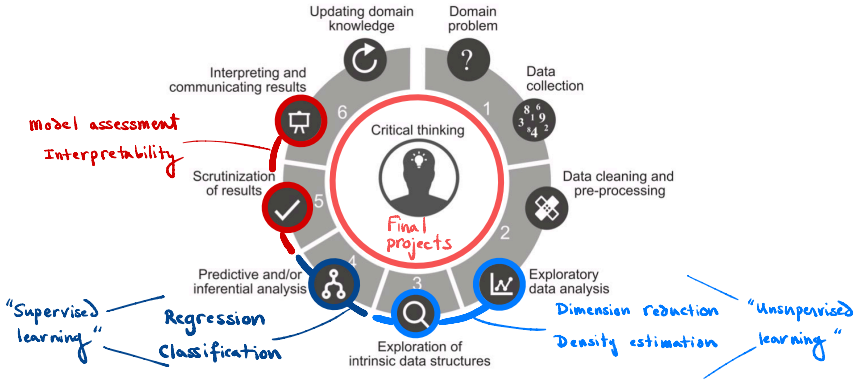


Fig. 2.1 of VDS Pt. I.

Plan for today

• Introductions, questionnaire	10 min
• Course overview	30 min
• Data science life cycle	10 min
• Plan for next class, questions	5 min

Plan for next class

Depending on your questionnaire responses, we will:

- Have a first look at Python in Google Colab
- Review some important facts from linear algebra
- Prepare for lecture 3 on “principal component analysis”

Questions?

- **The best way to reach me is through Canvas message**
- If you email me about the course, use ”MATH 637” in the subject line