Looking Back, Looking Ahead

Dennis Sun
Stanford University

DATASCI 112

March 13, 2026

1. A Look Back
2. A Look Ahead
3. The Data Science Majors and Minor
4. Remaining Tasks

1. A Look Back
2. A Look Ahead
3. The Data Science Majors and Minor
4. Remaining Tasks

Goal of this Class

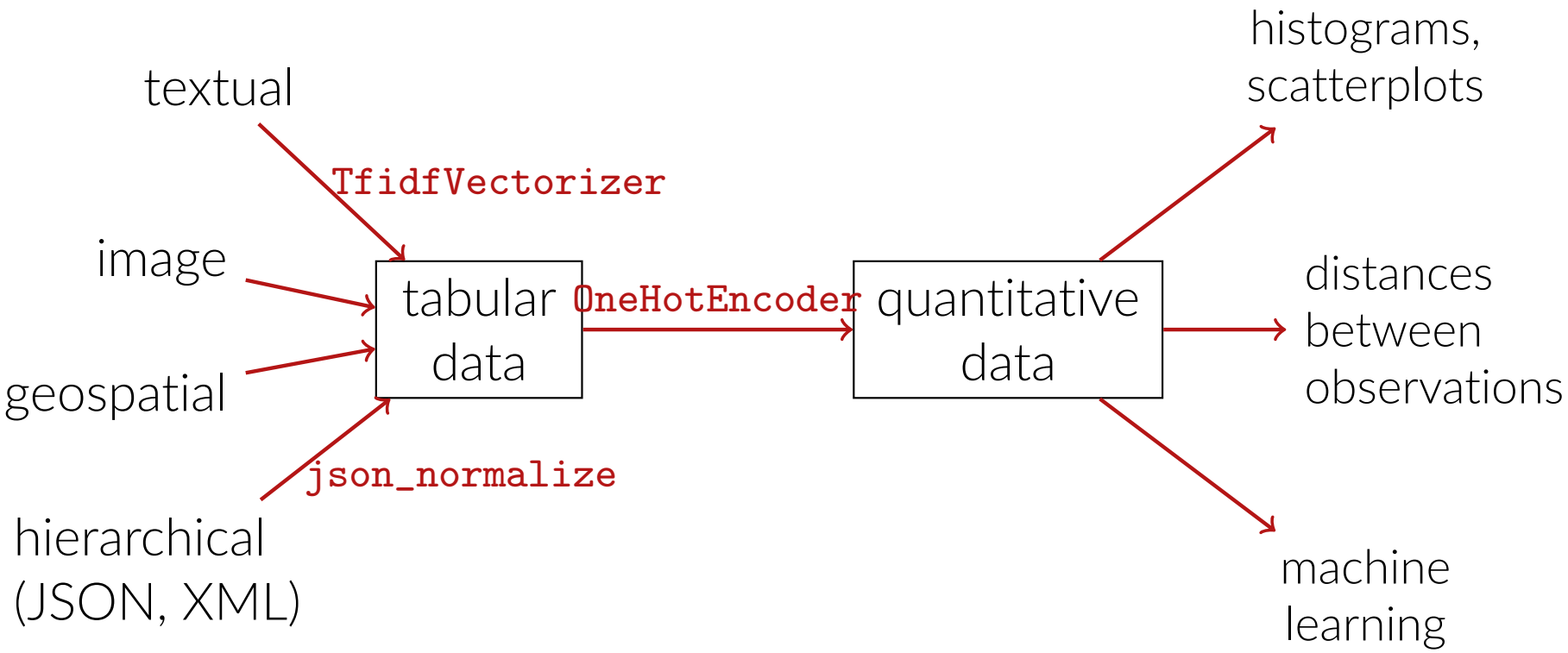
The main goal was to equip you with a set of data science tools so that you can start solving problems with data.

The lessons focused mostly on the tools, but I hope

- the guest lectures
- your final project

give you a glimpse of how these tools fit into the larger data science process.

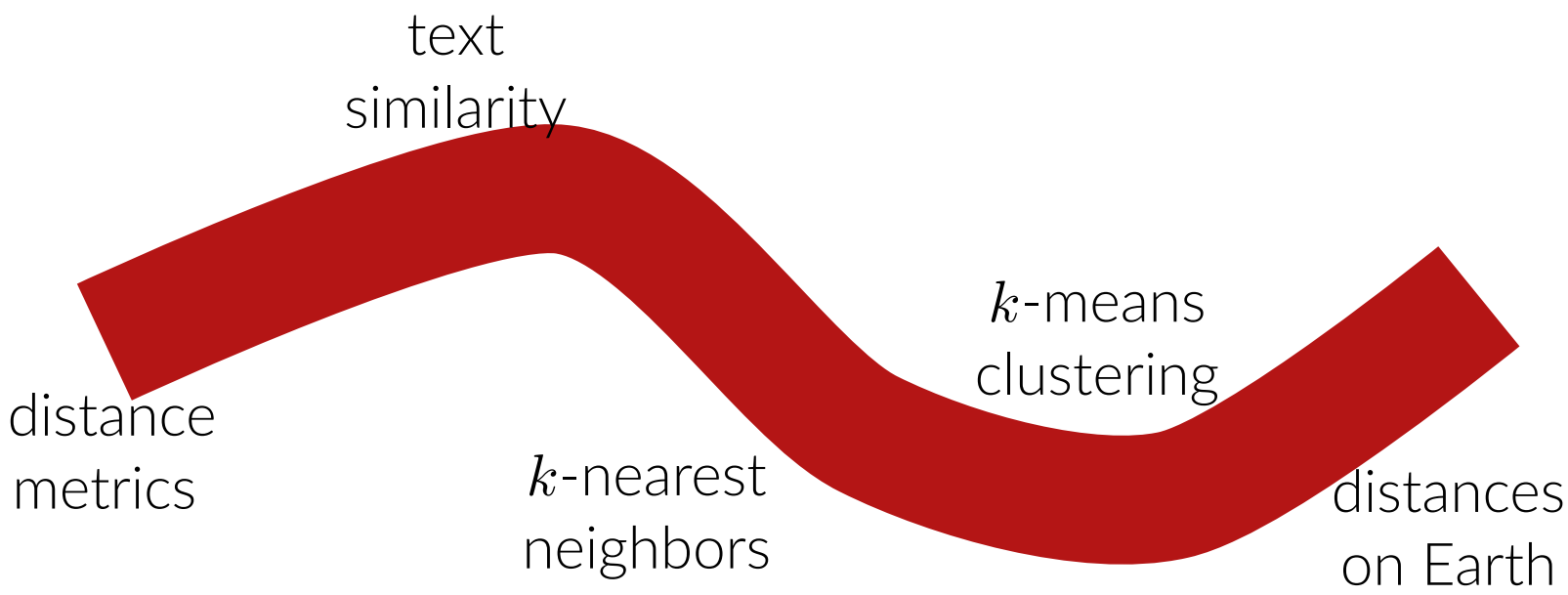
Themes of this Class



1. Many data science tools are designed for tabular data, specifically quantitative data.
2. But data comes in many shapes and sizes, and we have to convert them to tabular form to use these tools.

Themes of this Class

There have been a few strands that have run continuously throughout this course.



Along the way, we refined our understanding of distance metrics:

- cosine distance for textual data
- Haversine distance for geospatial data

Course Evaluations

Course evaluations are a good opportunity to reflect on what you learned!

One of the questions in the Stanford course evaluations is:
"How well did you achieve the learning goals of this course?"

I'm interested in your answer to this question! But what exactly were the learning goals?

- Syllabus:

  ## Learning Objectives

  - Acquire and process tabular, textual, hierarchical, and geospatial data.
  - Uncover patterns by summarizing and visualizing data.
  - Apply machine learning to answer real-world prediction problems.

- Course Description:

  DATASCI 112: Principles of Data Science

  A hands-on introduction to the methods of data science. Strategies for analyzing and visualizing tabular data, including common patterns and pitfalls. Data acquisition through web scraping and REST APIs. Core principles of machine learning: supervised vs. unsupervised learning, training vs. test error, hyperparameter tuning, and ensemble methods. Introduction to data of different shapes and sizes, including text, image, and geospatial data. The focus is on intuition and implementation, rather than theory and math. Implementation is in Python and Jupyter notebooks, using libraries such as pandas and scikit-learn. Course culminates in a final project where students apply the methods to a data science problem of their choice.

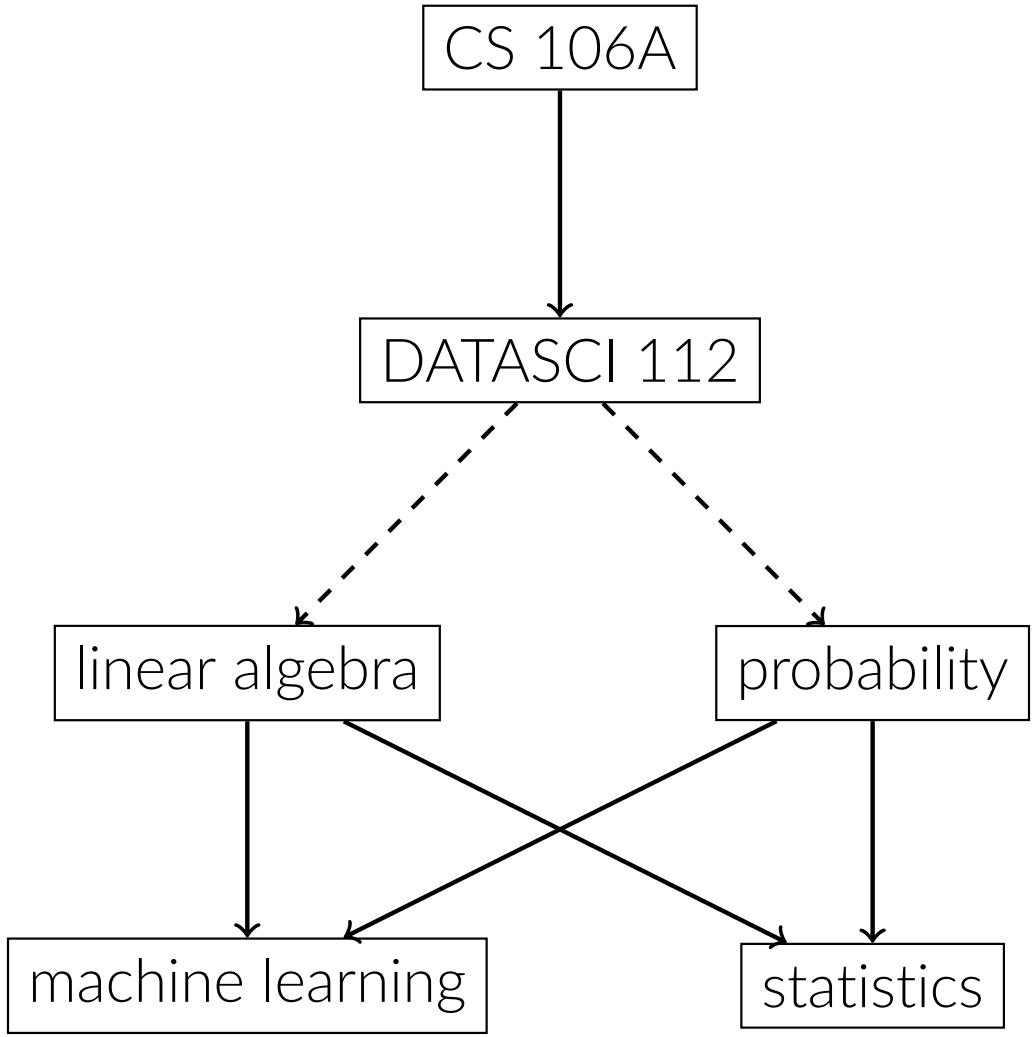
1 A Look Back

2 A Look Ahead

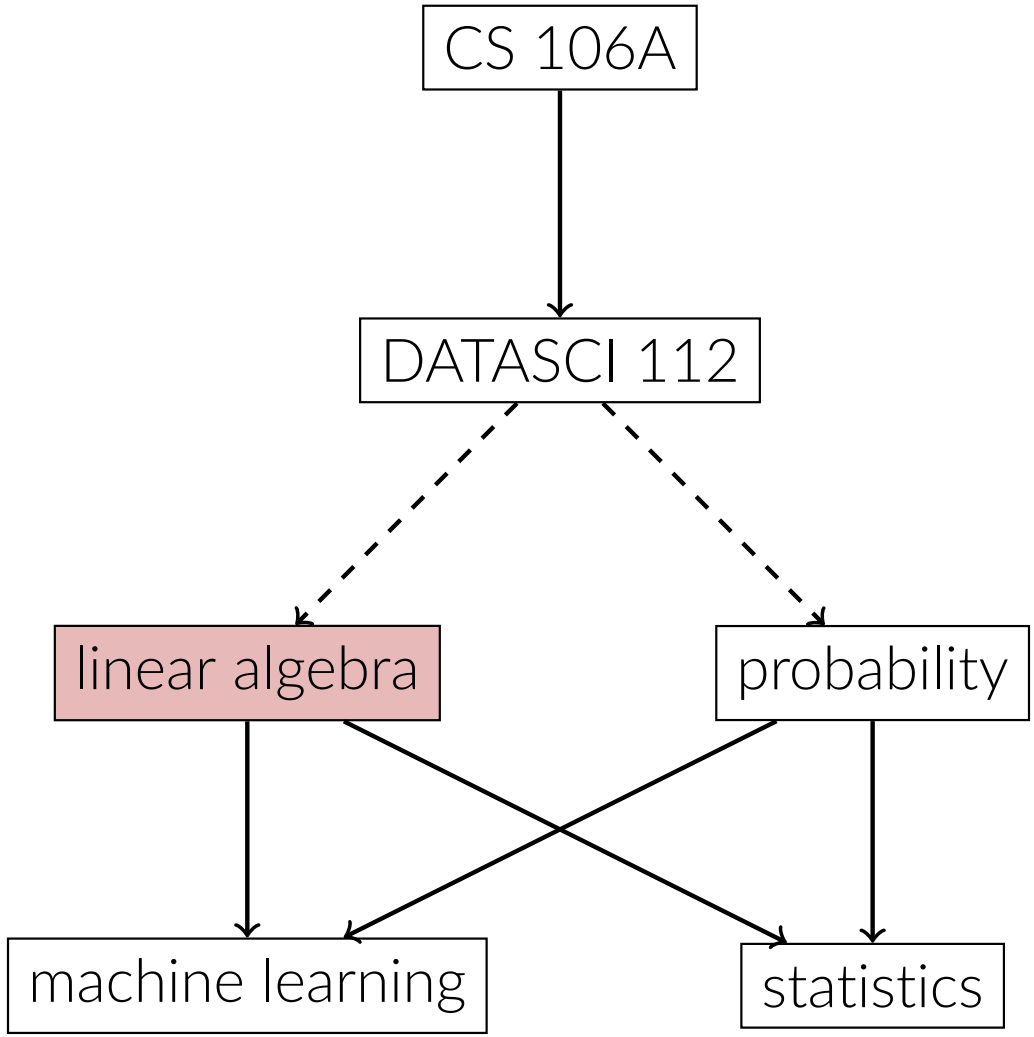
3 The Data Science Majors and Minor

4 Remaining Tasks

Road to Data Science



Road to Data Science



Linear Algebra

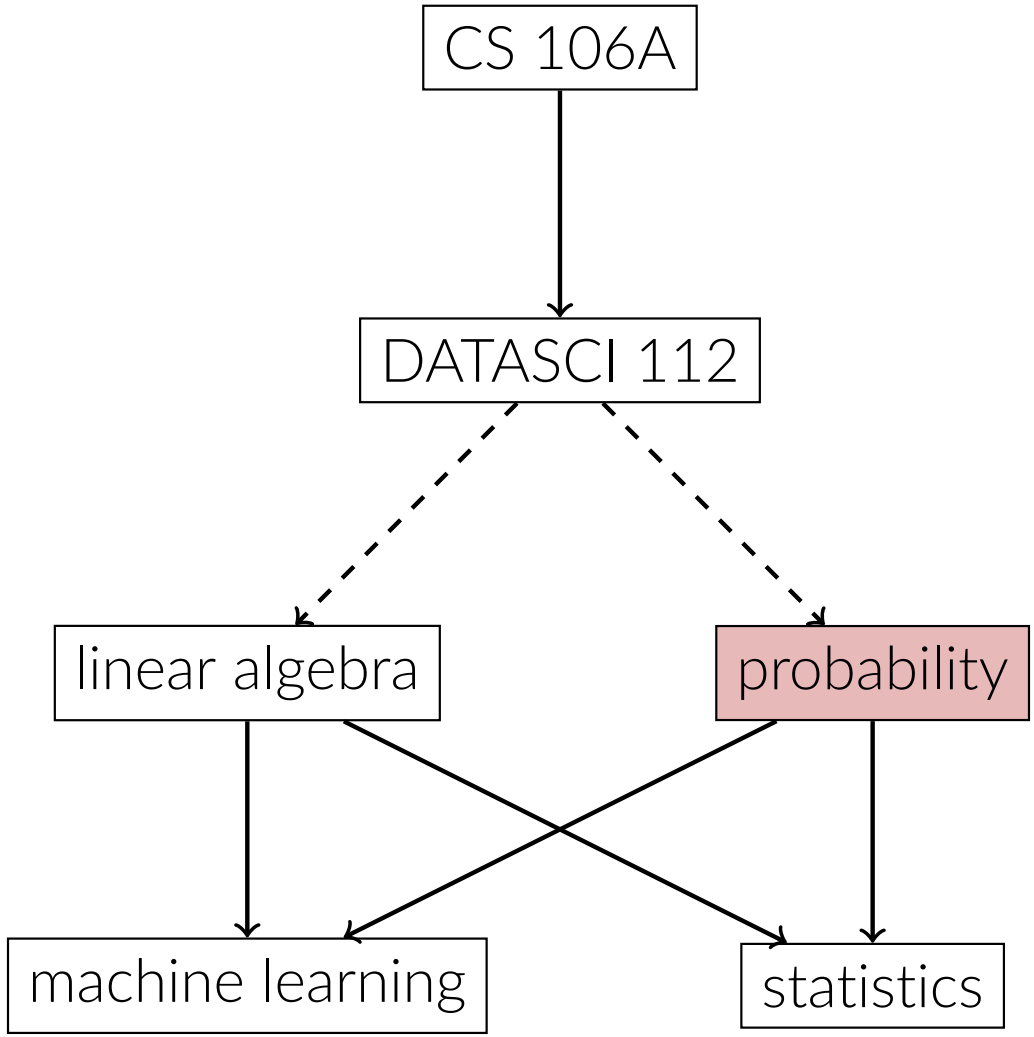
A `DataFrame` is a matrix X and a `Series` is a vector $\mathbf{y}$.

Linear algebra is the study of matrices and vectors.
It is the math underlying many concepts in DATASCI 112, from distance metrics to model fitting.

Classes:

- MATH 51 is a first introduction. Take it as soon as possible!
- MATH 104 and ENGR 108 delve deeper into linear algebra.

Road to Data Science



Probability

Many data science questions can be answered with joint, marginal, and conditional probabilities.

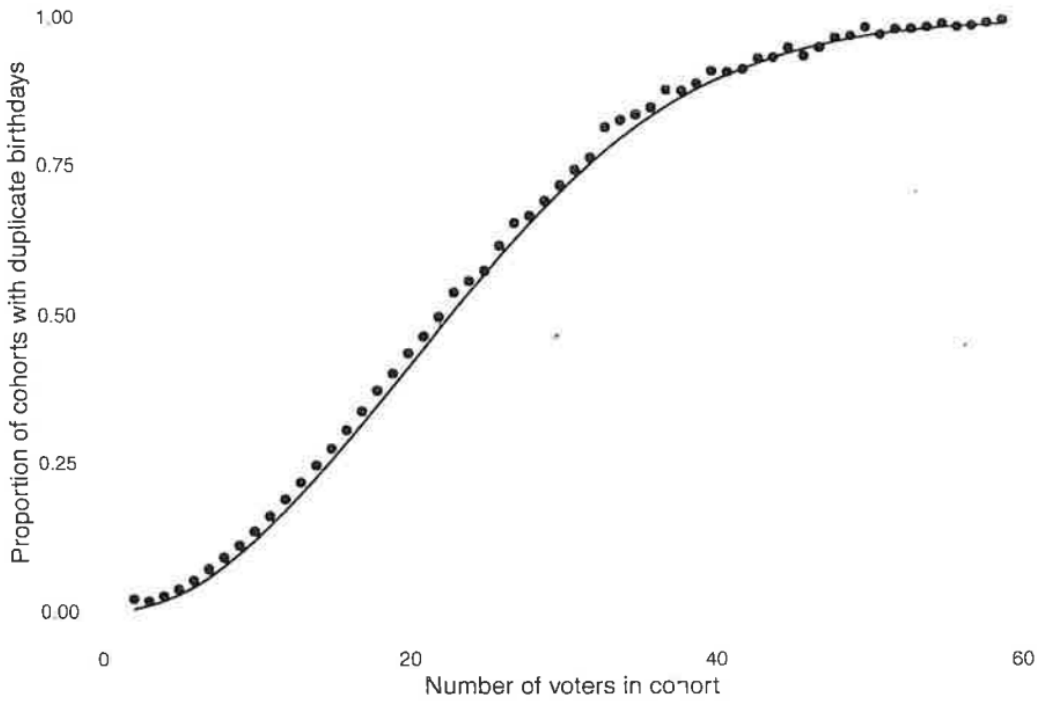
Take STATS 117!

- It will focus on the foundations of probability, including puzzles, paradoxes, and interview questions.
- Prerequisite is single-variable calculus.
- Offered Spring, Summer, and Autumn quarters.
- Replaces CS 109 for Computer Science majors if you eventually take machine learning.

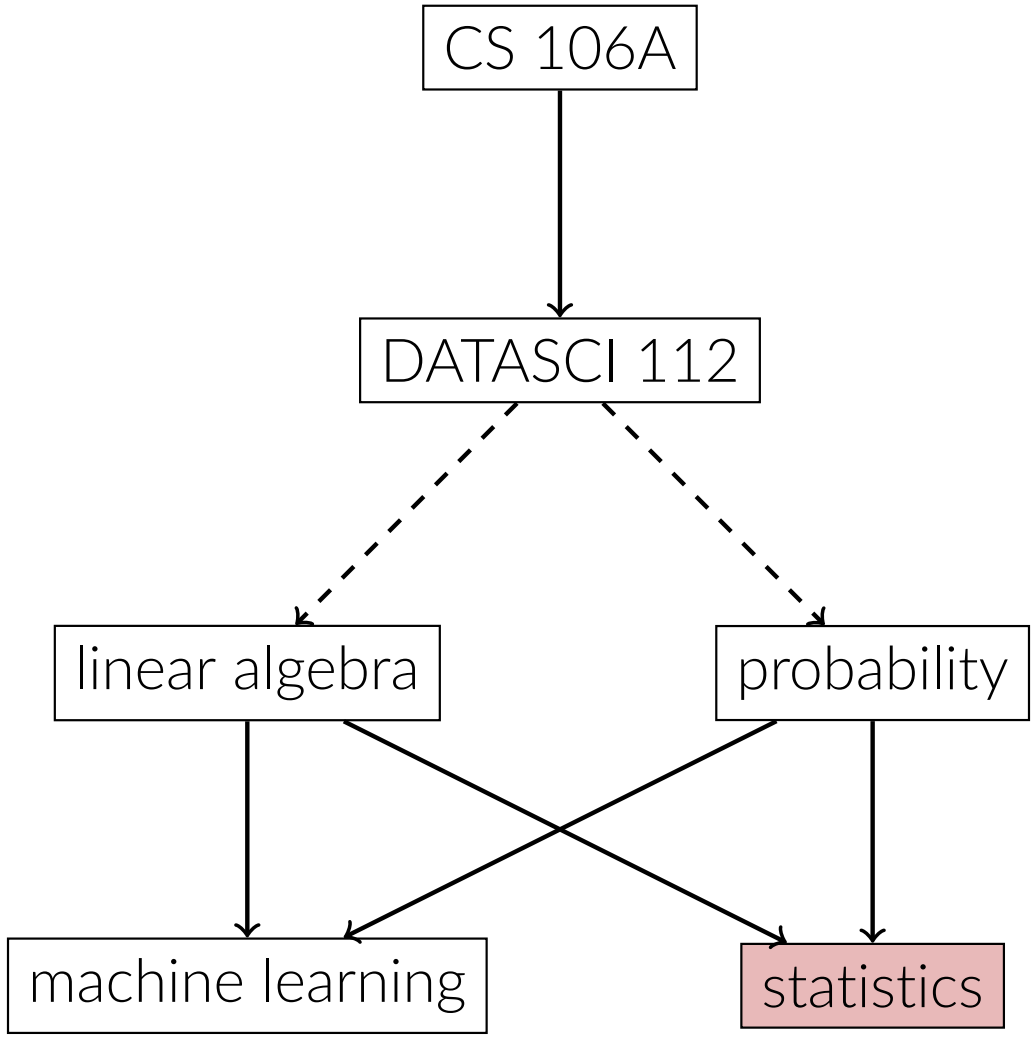
Example: What is the probability that (at least) two people share a birthday in this room?

The Birthday Problem Has Data Science Applications!

- Look at voter records for people with the same name born in the same year (i.e., “Steve Miller”s born in 1984).
- Consider these as people in one room and see if anyone shares a birthday.
- Aggregate over all names and birth years.



Road to Data Science



Statistics

Statistics is the science of uncertainty in data.

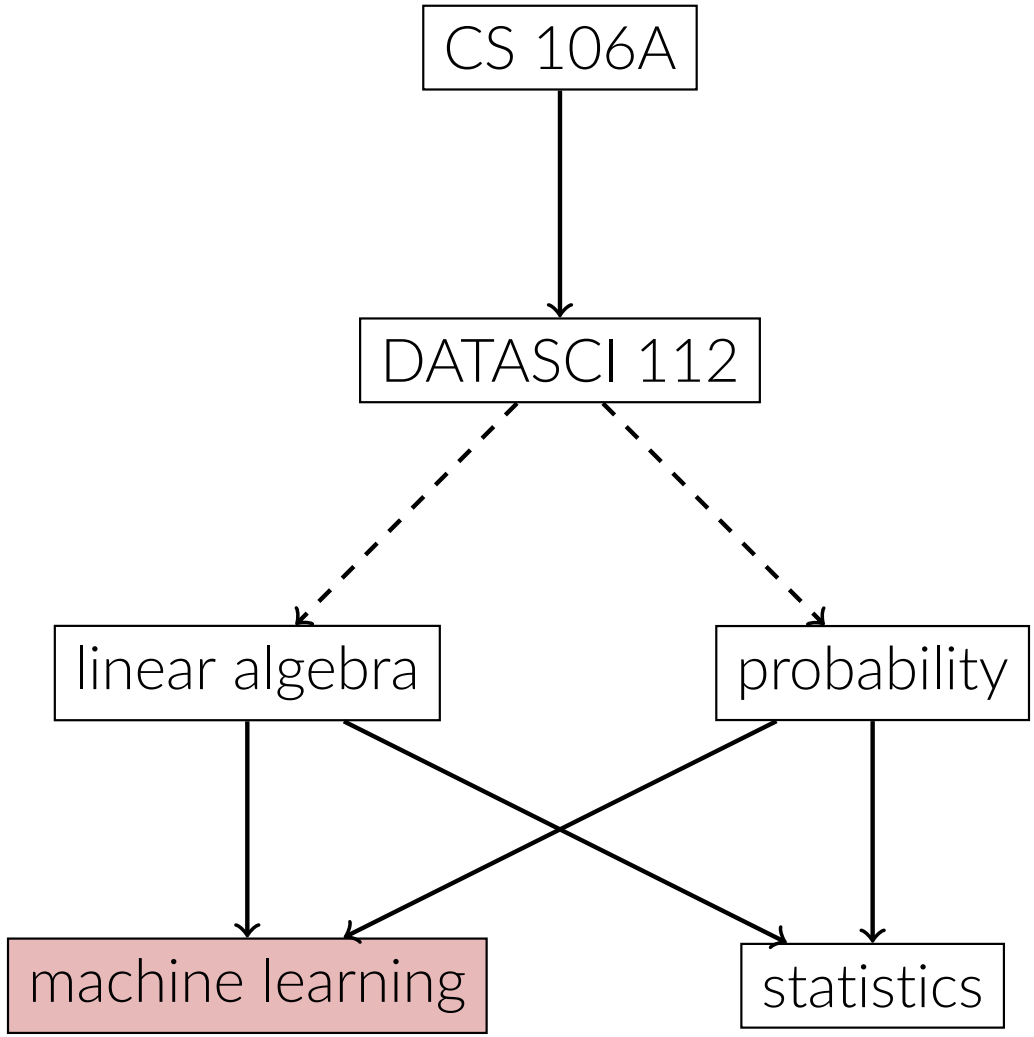
By quantifying uncertainty, statistics prevents us from seeing patterns in data when there are none.

- It uses linear algebra to represent the data.
- It uses probability to model the uncertainty.

Classes:

- **Intro Stats:** STATS 60, STATS 110, STATS 141, CME 106, ECON 102A
- **Statistical Methods:** STATS 191 (without linear algebra), STATS 203 (with linear algebra)
- **Theoretical Stats:** STATS 118 $\rightarrow$ STATS 200

Road to Data Science



Machine Learning

We have covered quite a bit of machine learning in DATASCI 112.

- In fact, I would argue that we've covered the most important things a practitioner should know.
- Very few data scientists need to do machine learning that can't be done in Scikit-Learn.

In an advanced machine learning class (CS 229, STATS 202), you'd learn:

- more algorithms (decision trees, neural networks, etc.)
- the math behind them (loss functions, optimization)
- how to implement them from scratch

In my opinion, only a handful of researchers need to know details like these, and many practitioners screw up more basic things:

- not splitting the data into a training set and a validation set
- not scaling the data

1 A Look Back

2 A Look Ahead

3 The Data Science Majors and Minor

4 Remaining Tasks

Majors in Data Science

	B.S.	B.A. Social Systems	B.A. Artistic and Cultural Analysis
Core	• Math • Computer Science • Statistics • Optimization	• Math • Computer Science • Statistics • Optimization • Social Systems (econ, sociology, poli sci, psych)	• Math • Computer Science • Statistics • Optimization • Humanities (English, history, art, music)
Courses Next Quar-ter	• STATS 117 • MATH 51 or 104	• STATS 117 • DATASCI 154: Data Science for Social Impact	• STATS 117 • DATASCI 156: Thinking and Making with Data

Minor in Data Science

- Linear Algebra: MATH 51 or ENGR 108
- Computer Science: CS 106A (or B)
- Data Science: this class!
- Probability
- Statistics
- Data Science Methodology (elective)

Minor in Data Science

- ~~Linear Algebra: MATH 51 or ENGR 108~~
- ~~Computer Science: CS 106A (or B)~~
- ~~Data Science: this class!~~
- Probability
- Statistics
- Data Science Methodology (elective)

Most of you only need 3 courses to complete the minor.

Recommended Courses Next Quarter:

- Probability: STATS 117
- Data Science Methodology: DATASCI 154 or 156

Opportunities in Statistics

- Minor in Statistics
- M.S. in Statistics (co-term)

1 A Look Back

2 A Look Ahead

3 The Data Science Majors and Minor

4 Remaining Tasks

Remaining Tasks

- Lab 6B (optional, only if you want to replace your lowest lab part)
- Final Project
- Course Evaluations

Thanks for a great quarter!

We're excited to see the projects at the poster sessions next week!