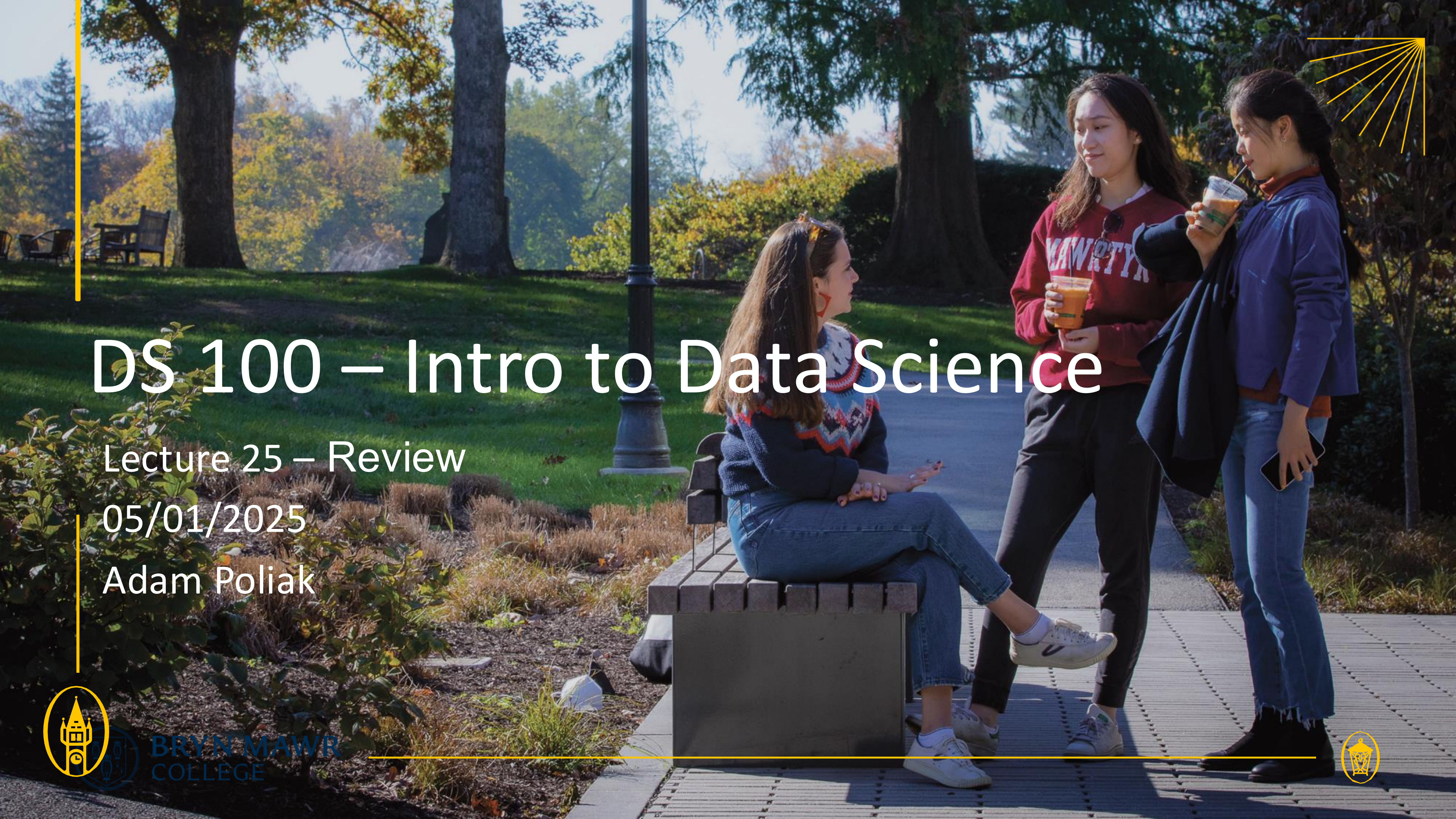


# DS 100 – Intro to Data Science

Lecture 25 – Review

05/01/2025

Adam Poliak

BRYN MAWR COLLEGE

# Announcements

Midsemester feedback form - https://forms.gle/M4jVdGTDgQknWAwo6

HW10 (due Friday May 2nd )

Project 3 (due Friday May 2nd)

Final 05/07 9:30am Old Library 116

# Outline

Review Classification

Evaluation Metrics

Comparing two classifiers

# Classification

# Classifiers

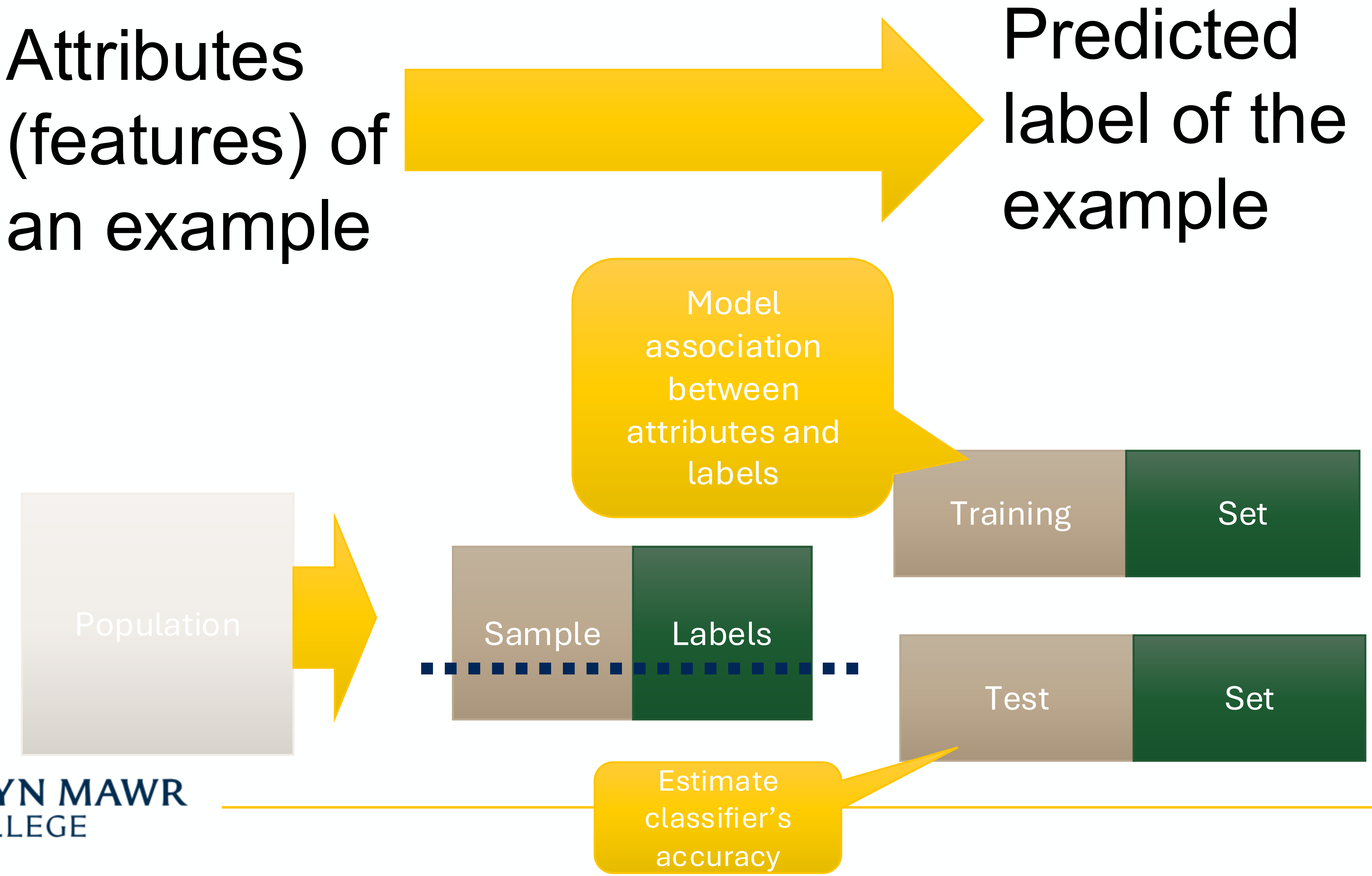
BRYN MAWR COLLEGE

# A Classifier

Attributes (features) of an example

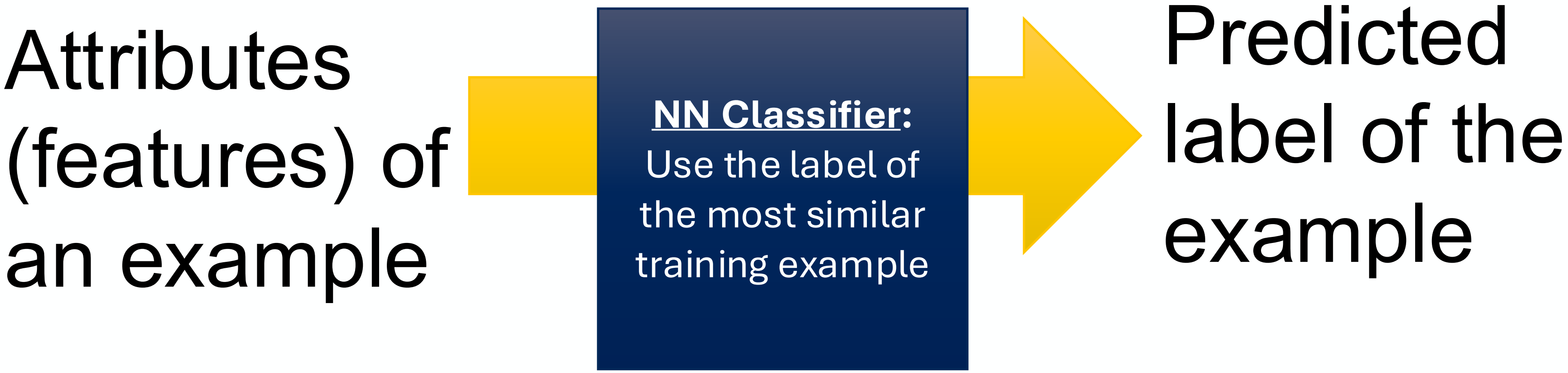

Predicted label of the example

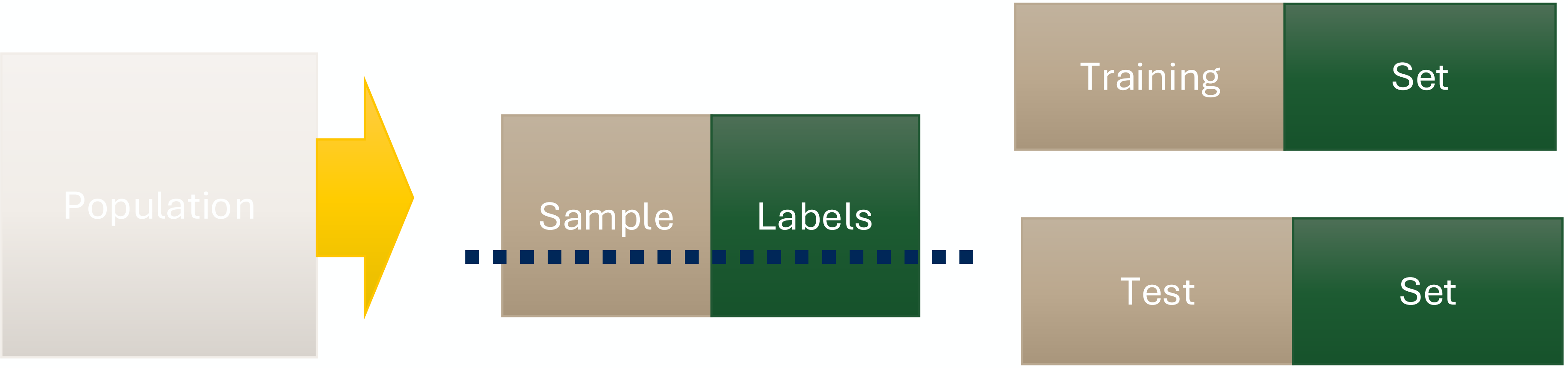
# Training and Evaluating a Classifier



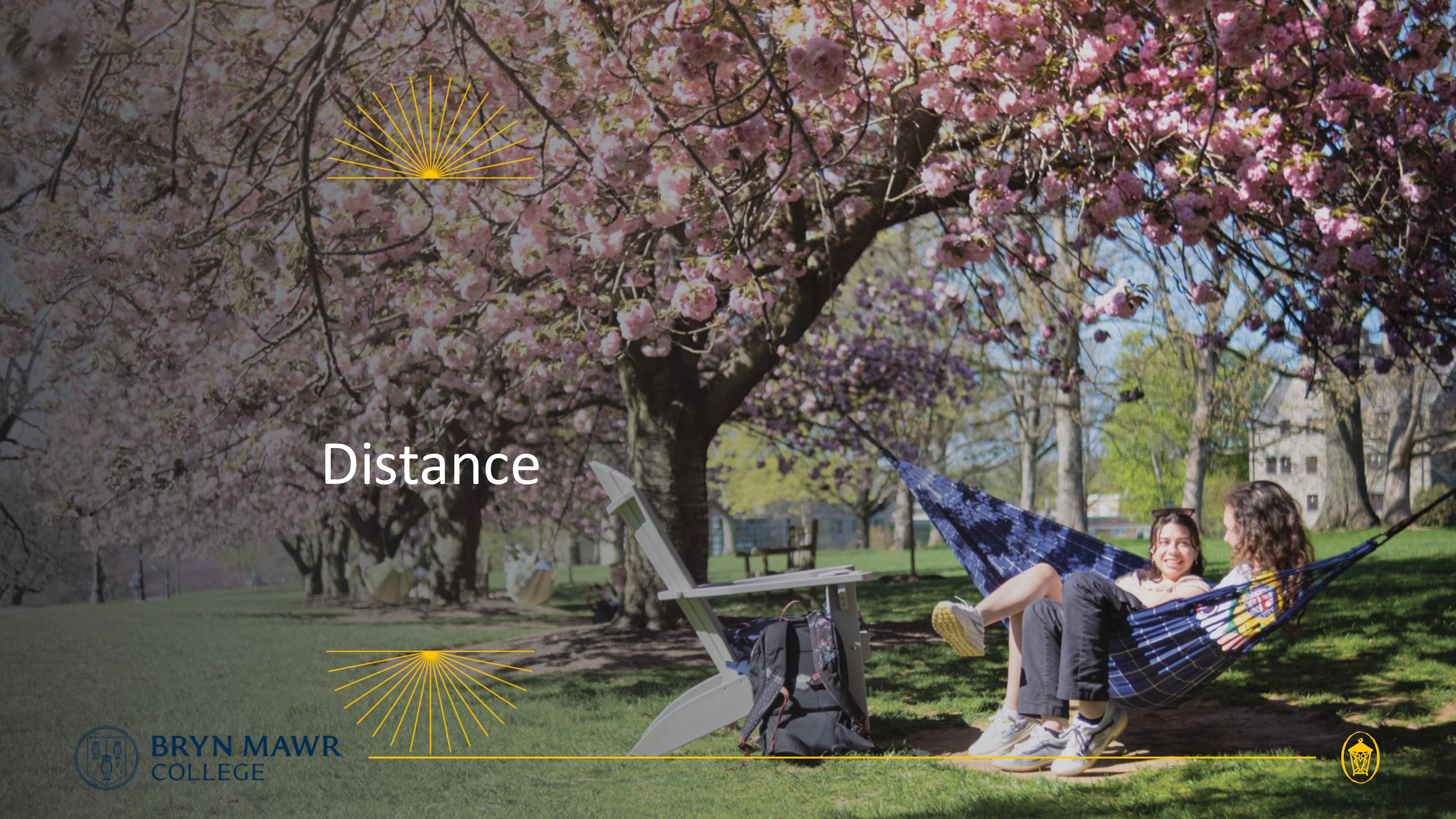
# Nearest Neighbor Classification

# Nearest Neighbor Classifier



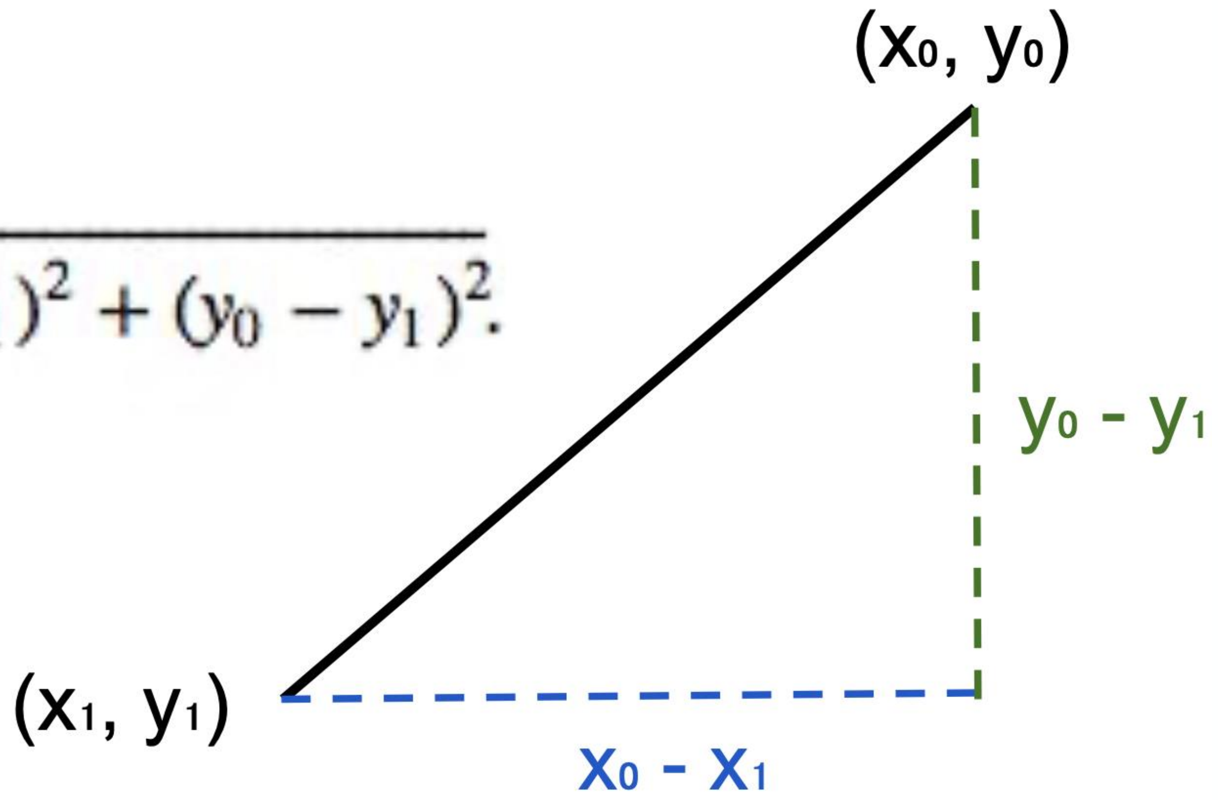



# Distance



# Pythagorea's Formula

$$D = \sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2}.$$



# Course Review
## (in depth)

# Course Outline

- Computation: Python and Tables
- Exploration
  - Discover patterns in data
  - Articulate insights (visualizations)
- Inference
  - Make reliable conclusions about the world
  - Probability & Statistics
- Prediction
  - Informed guesses about unseen data
  - Machine Learning: Regression & Classification

# Computation in Python

Textbook sections

- General features and Table methods: 3.1 - 9.3, 17.3
- sample_proportions: 11.1
- percentile: 13.1
- np.average, np.mean, np.std: 14.1, 14.2
- minimize: 15.4

# Exploring Data

# Describing Data

- Qualitative:
  - Visualizing Distributions: Chapter 7
- Quantitative
  - Center and spread: 14.1-14.3
  - Linear trend and non-linear patterns: 8.1, Chapter 15

# Measures of Center

Median

- 50th percentile, where
- *p*th percentile = smallest value on list that is at least as large as *p*% of the values 13.1
- Median is not affected by outliers

Mean/Average

- Depends on all the values
- smoothing operation
- center of gravity of histogram
  - if histogram is skewed, mean is pulled away from median towards the tail

# Measure of Spread

- Standard deviation (SD) measures roughly how far the data are from their average

- SD = root mean square of deviations from average

Steps: 5 4 3 2 1

## Chebyshev's Bounds

| Range | Proportion |
|---|---|
| average ± 2 SDs | at least 1 - 1/4 (75%) |
| average ± 3 SDs | at least 1 - 1/9 (88.888...%) |
| average ± 4 SDs | at least 1 - 1/16 (93.75%) |
| average ± 5 SDs | at least 1 - 1/25 (96%) |

**True no matter what the distribution looks like**

# Bounds and Normal Approximations

| Percent in Range | All Distributions | Normal Distributions |
|---|---|---|
| Average +- 1 SD | At least 0% | About 68% |
| Average +- 2 SDs | At least 75% | About 95% |
| Average +- 3 SDs | At least 88.888...% | About 99.73% |

# Standard Units Z

"average ± SDs" 14.2

- *z* measures "how many SDs above average"
- Almost all standard units are in the range (-5, 5)
- To convert a value to standard units:

$$z = \frac{\text{value - average}}{\text{SD}}$$

# The Correlation Coefficient *r*

- Measures ***linear*** association
- Based on standard units; pure number with no units
- *r* is not affected by changing units of measurement
- $-1 \leq r \leq 1$
- $r = 0$: No linear association; *uncorrelated*
- *r* is not affected by switching the horizontal and vertical axes
- Be careful before you use it
- 15.1

# Definition of *r*

**Correlation Coefficient** (r) =

average of product of standard(x) and standard(y)

Steps: 4 3 2 1

Measures how clustered the scattered data are around a straight line

**estimate of $y = r \cdot x$**, when both variables are measured in standard units

# Slope and Intercept

$$estimate\ of\ y = slope\ * x + intercept$$

$$\boldsymbol{slope\ of\ the\ regression\ line}$$

$$r\ * \frac{SD\ of\ y}{SD\ of\ x}$$

$$\boldsymbol{intercept\ of\ the\ regression\ line}$$

$$mean(y) - slope\ \times\ mean(x)$$

# Regression Line

- Regression line is the "least squares" line
- Minimizes the *root mean squared error* of prediction, among all possible lines
- No matter what the shape of the scatter plot, there is one best straight line
  - but you shouldn't use it if the scatter isn't linear
- 15.3, 15.4

# Residuals

- Error in regression estimate
- One residual corresponding to each point (*x*, *y*)
- **residual**

  **= observed *y* - regression estimate of *y***

  = vertical difference between point and line
- No matter what the shape of the scatter plot:
  - Residual plot does not show a trend
  - Average of residuals = 0

# Inference

# General Concepts

- Study, experiment, treatment, control, confounding, randomization, causation, association: Chapter 2
- Distribution: 7.1, 7.2
- Sampling, probability sample: 10.0
- Probability distribution, empirical distribution, law of averages: Chapter 10
- Population, sample, parameter, statistic, estimate: 10.1, 10.3
- Model: every null and alternative hypothesis; 16.1

# Goal of Inference

- To make conclusions about unknown features of the population or model, based on assumptions of randomness

# Probability

- Probability theory:
  - Exact calculations
  - Normal approximation for mean of large random sample
  - Accuracy and sample size

## Equally Likely Outcomes

**Assuming** all outcomes are equally likely, the chance of an event A is:

$$P(A) = \frac{\textit{number of outcomes that make A happen}}{\textit{total number of outcomes}}$$

# Large Sample Approximation: CLT

## Central Limit Theorem

If the sample is

- large, and
- drawn at random with replacement,

Then, *regardless of the distribution of the population,*

**the probability distribution of the sample sum (or of the sample mean) is *roughly* bell-shaped**

# Inference: Estimation

# Estimating a Numerical Parameter

- Question: What is the value of the parameter?
- Terms: **predict, estimate, construct a confidence interval, confidence level**
- Answer: Between x and y, with 95% confidence
- Method (13.2, 13.3):
  - **Bootstrap the sample**; compute estimate
  - Repeat; draw empirical histogram of estimates
  - Confidence interval is “middle 95%” of estimates
- Can replace 95% by other confidence level (not 100%)

## Meaning of "95% Confidence"

- You'll never get to know whether or not your constructed interval contains the parameter.
- The confidence is in the process that generates the interval.
- The process generates a good interval (one that contains the parameter) about 95% of the time.
- End of 13.2

## Reasons to use a confidence interval

- To **estimate** a numerical parameter: 13.3
  - Regression **prediction**, if regression model holds: Predict *y* based on a new *x*: 16.3
- To **test** whether or not a numerical parameter is equal to a specified value: 13.4
  - In the regression model, used for testing whether the slope of the true line is 0: 16.2

# Inference: Testing

# Testing Hypotheses

- **Null:** A completely specified chance model, under which you can simulate date.
  - Need to say exactly what is due to chance, and what the hypothesis specifies.
- **Alternative:** The null isn’t true
  - something other than chance is going on; might have a direction
- **Test Statistic:** A statistic that helps decide between the two hypotheses, based on its empirical distribution under the null
- 11.3

# The P-value

- The chance, **under the null hypothesis**, that the test statistic comes out equal to the one in the sample or more in the direction of the alternative
- If this chance is small, then:
  - If the null is true, something very unlikely has happened.
  - Conclude that the data support the alternative hypothesis more than they support the null.
- 11.3

# Error Probability

- Even if the null is true, your random sample might indicate the alternative, just by chance
- The **cutoff** for P is the chance that your test makes the wrong conclusion when the null hypothesis is true
- Using a small cutoff limits the probability of this kind of error
- 11.4

# Testing Data in Two Categories

- **Null:** The sample was drawn at random from a specified distribution.
- Test statistic: Either count/proportion in one category, or distance between count/proportion and what you'd expect under the null; depends on alternative
- Method:
  - Simulation: Generate samples from the distribution specified in the null.
- 11.1 (Swain v. Alabama, Mendel)

# Testing Data in Multiple Categories

- **Null:** The sample was drawn at random from a specified distribution.
- Test statistic: TVD between distribution in sample and distribution specified in the null.
- Method:
  - Simulation: Generate samples from the distribution specified in the null.
- 1.2 (Alameda county juries)

## Comparing Two Numerical Samples

- **Null:** The two samples come from the same underlying distribution in the population.
- Test statistic: difference between sample means (take absolute value depending on alternative)
- **Method for A/B Testing:**
  - Permutation under the null: 12.2 (Deflategate), 12.1 (birth weight etc for smokers/nonsmokers), 12.3 (BTA randomized controlled trial)

# One Numerical Parameter

- **Null:** parameter = a specified value.
- **Alternative:** parameter ≠ value
- Test Statistic: Statistic that estimates the parameter
- Method:
  - Bootstrap: Construct a confidence interval and see if the specified value is in the interval.
- 13.4, 16.2 (slope of true line)

# Causality

- Tests of hypotheses can help decide that a difference is not due to chance
- But they don't say ***why*** there is a difference …
- Unless the data are from an RCT 12.3
  - In that case a difference that's not due to chance can be ascribed to the treatment

# Prediction

# Regression

- Regression model 16.1
- Bootstrap confidence interval for the true slope 16.2
  - Use of this interval to test if the true slope is 0
- Bootstrap prediction interval for *y* at a given value of *x* 16.3

# Classification

- Binary classification based on attributes 17.1
  - *k*-nearest neighbor classifiers
- Training and test sets 17.2
  - Why these are needed
  - How to generate them
- Implementation: 17.4
  - Distance between two points
  - Class of the majority of the *k* nearest neighbors
- Accuracy: Proportion of test set correctly classified 17.5

# Data Science

# Why Data Science

- Unprecedented access to data means that we can make new discoveries and more informed decisions
- Computation is a powerful ally in data processing, visualization, prediction, and statistical inference
- People can agree on evidence and measurement
- Data and computation are everywhere: understanding and interpreting are more important than ever

# Limitations of Data Science

- Evidence and measurements are critical ingredients for good decision-making
  - ...but they're not enough by themselves!
- Data science is a powerful complement to qualitative analysis
  - but it's not a replacement!

# How to Analyze Data

- Begin with a question from some domain, make reasonable assumptions about the data and a choice of methods.
- Visualize, then quantify!
- *Perhaps the most important part*: Interpretation of the results in the language of the domain, without statistical jargon.

# How *Not* to Analyze Data

- Begin with a question from some domain, make reasonable assumptions about the data and a choice of methods.
- Visualize, then quantify!
- *Perhaps the most important part*: Interpretation of the results in the language of the domain, without statistical jargon.

# How to Analyze Data after Data Science 100

- Begin with a question from some domain, make reasonable assumptions about the data and a choice of methods.

- Visualize, then quantify! **Do both using computation**

- *Perhaps the most important part*: Interpretation of the results in the language of the domain, without statistical jargon.

# Data Science 100 – *Analyzing Data with Computation*

- Table manipulation using Python
- Working with whole distributions, not just means
- Decisions based on sampling: assessing models
- Estimation based on resampling
- Understanding sampling variability
- Prediction

# Continuing in data science

Math courses:

- Linear Algebra
- Probability & Statistics (Math H218)

Computer Science Courses

- Data Structures
- Discrete Math
- Algorithms

# Thank you!