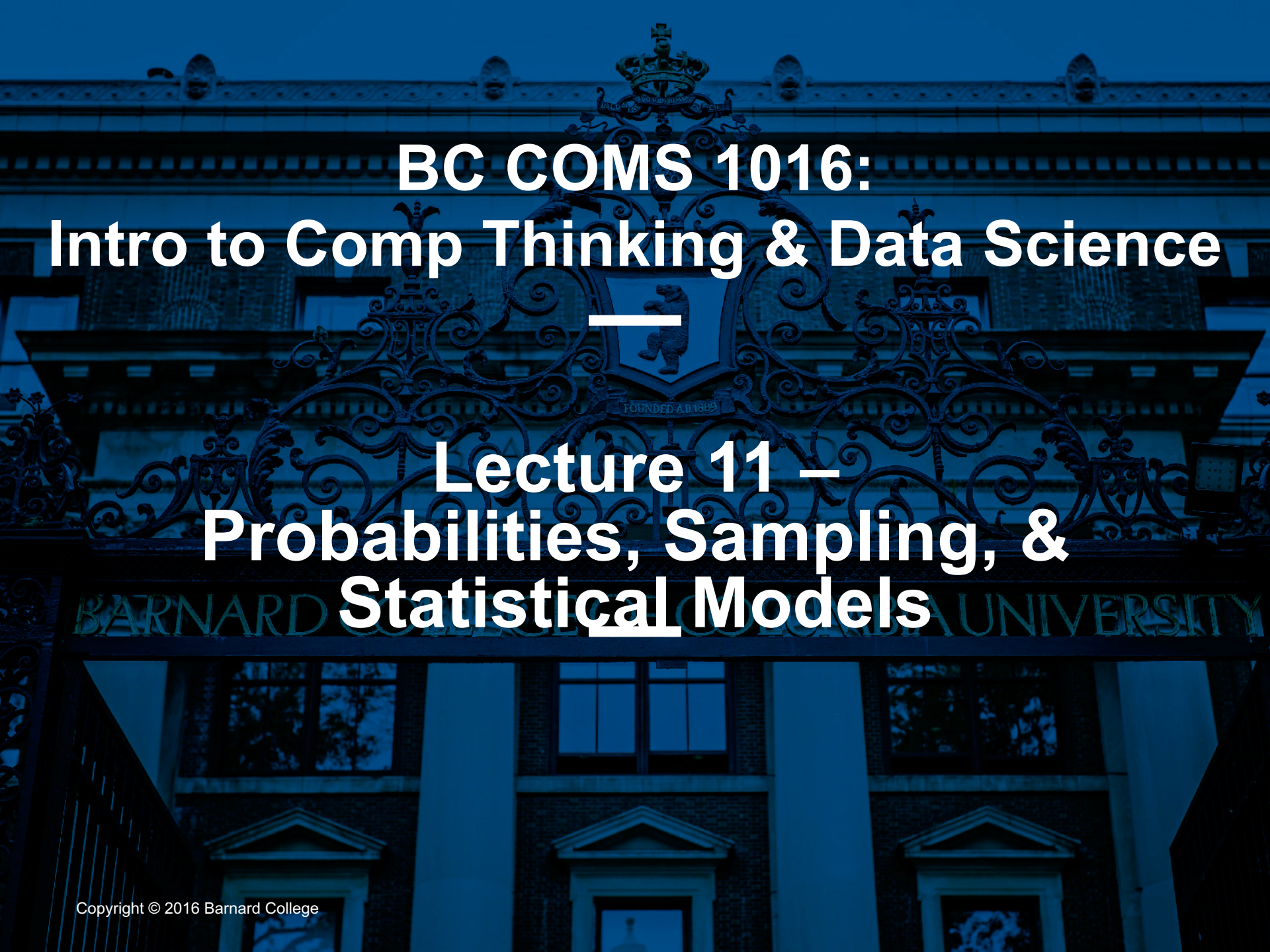


BC COMS 1016: Intro to Comp Thinking & Data Science

—

Lecture 11 – Probabilities, Sampling, & Statistical Models

Announcements

- HW04 - Applying Functions and Iteration
  - Due Tonight (11/16)
- Lab 05 - Assessing Models: Examining the Therapeutic Touch
  - Postponed – no lab due Wednesday 11/18
- HW05 - Probability, Simulation, Estimation, and Assessing Models
  - Due Thursday 11/19
- Checkpoint/Project 1:
  - Due Wednesday 11/18
- Checkpoint/Project 2 (midterm):
  - Released Thursday 11/19, due Tuesday 11/24

Homework 04 – common issues

- What does a row in a Table represent?
  - *Each individual in our table*
- What does each column in a Table represent?
  - *The attributes*
  - *Attribute for a specific individual*
- How do we find how many individual's in a Table?
  - .num_rows
- How do we find how many attributes in a Table?
  - .num_columns
- len(Table) will give us the number of columns, not the number of rows
- Question 1.3: the function expects a row as input and not an index (integer) of a row

On apologizing

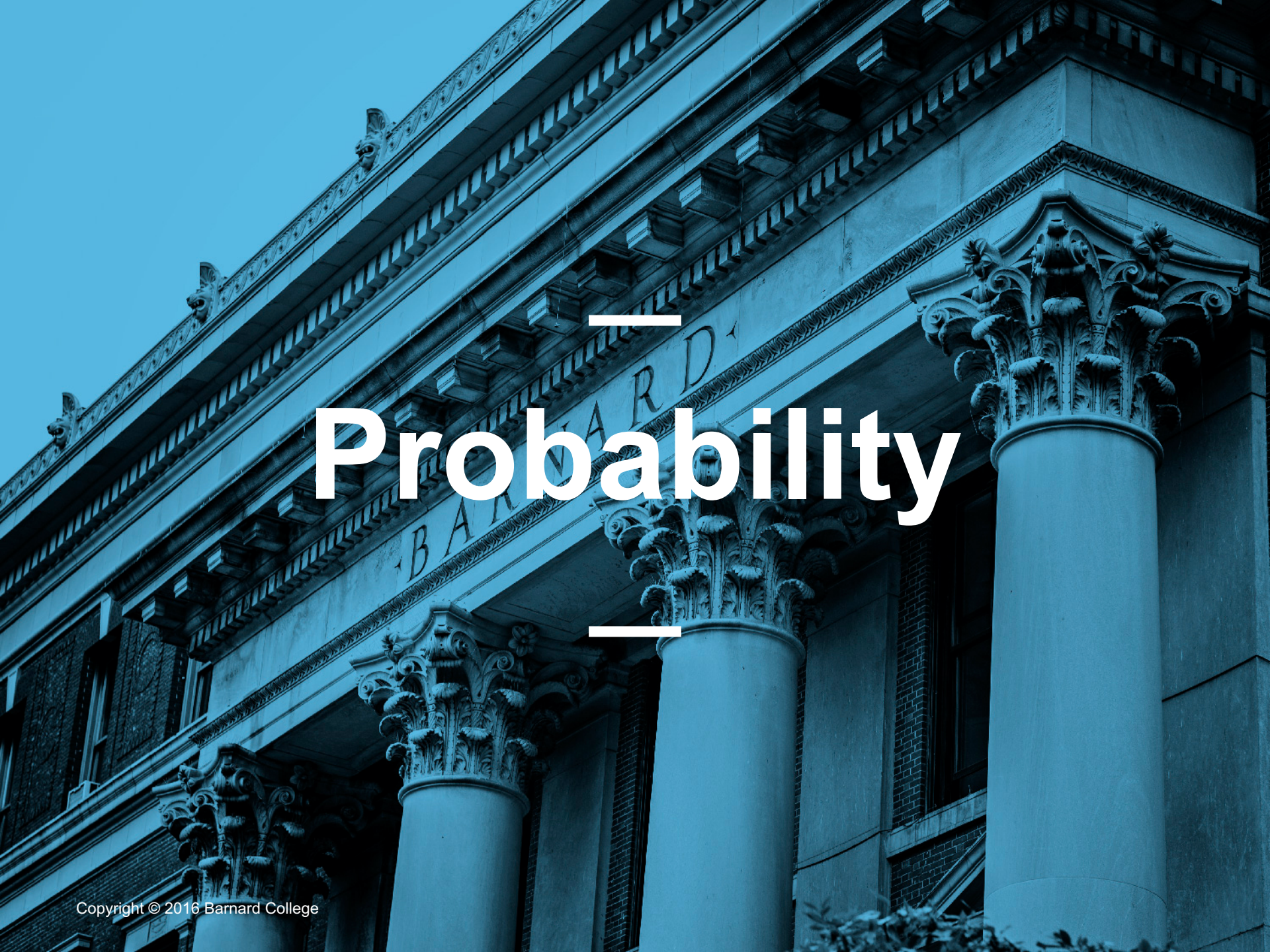
"Sorry for taking up so much of your time, I am new to coding and data science so have quite a few questions"

Do not apologize for:

- Asking questions
- Taking advantage of the resources we offer
  - Especially if this material is new to you

In general:

- Don't apologize for "taking up space in the room"
- Own it!



Probability

Basics

- **Lowest value**: 0
  - Chance of event that is impossible
- **Highest value**: 1 (or 100%)
  - Chance of event that is certain

- If an event has chance 70%, then the chance that it doesn't happen is:
  - 100% - 70% = 30%
  - 1 – 0.7 = 0.3
  - We call this the **Complement**

Equally Likely Outcomes

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

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

A Question

- I have 3 cards: **ace of hearts**, **king of diamonds**, and **queen of spades**
- I shuffle them and draw two cards at *random without replacement.*
- What is the chance that I get the Queen followed by the King?

Approach 1: Enumerate all outcomes

- What is the chance that I get the Queen followed by the King?
    1. Queen, King
    2. Queen, Ace
    3. Ace, King
    4. Ace, Queen
    5. King, Queen
    6. King, Ace

Approach 1: Enumerate all outcomes

- What is the chance that I get the Queen followed by the King?
  1. Queen, King
  2. Queen, Ace
  3. Ace, King
  4. Ace, Queen
  5. King, Queen
  6. King, Ace

Approach 1: Enumerate all outcomes

- What is the chance that I get the Queen followed by the King?

  1. Queen, King
  2. Queen, Ace
  3. Ace, King
  4. Ace, Queen
  5. King, Queen
  6. King, Ace

- Answer: 1/ 6

Approach 2: Probabilities of the sequences

- What is the chance that I get the Queen followed by the King?

- What's the probability I first draw Queen and what's the probability I then draw King

Approach 2: Probabilities of the sequences

- Step 1:
  - Draw Queen from {Ace, King, Queen}
  - What's the probability of drawing Queen? **1/3**
- Step 2:
  - Draw King from {King, Ace}
  - What's the probability of drawing King? **1/2**
- Combining them:
  - What's 1/2 of 1/3? **1/6**

Multiplication Rule

Chance that two events *A* and *B* both happen

= P(*A* happens) x P(*B* happens given that *A* has happened)

- The answer is *less than or equal* to each of the two chances being multiplied
- The more conditions you have to satisfy, the less likely you are to satisfy them all

Addition Rule

If event *A* can happen in *exactly one* of two ways, then

$$P(A) = P(\text{first way}) + P(\text{second way})$$

- The answer is *greater than or equal* to the chance of each individual way

Complement: At Least One Head

- What the probability that I flip coins and I get at least one head?
- In 3 tosses:
  - Any outcome *except* TTT (tails, tails, tails)
  - P(TTT) = (1/2) x (1/2) x (1/2) = 1/8
  - P(at least one head) = 1 – P(TTT) = 1 – (1/8) = 87.5%
- In 10 tosses:
  - 1 – (1/2)**10 ≅ 99.9%

Probability & Sampling

Discussion Question

A population has 50 people, including Harmon and Shaibel. We sample two people at random without replacement.

A) P(both Harmon and Shaibel are in our sample)

B) P(neither Harmon or Shaibel are in our sample)

Discussion Question

A population has 50 people, including Harmon and Shaibel. We sample two people at random without replacement.

A) P(both Harmon and Shaibel in our sample)
= P(first Harmon, second Shaibel) + P(first Shaibel, second Harmon)
= (1/50 * 1/49) + (1/50 * 1/49). = 0.0008

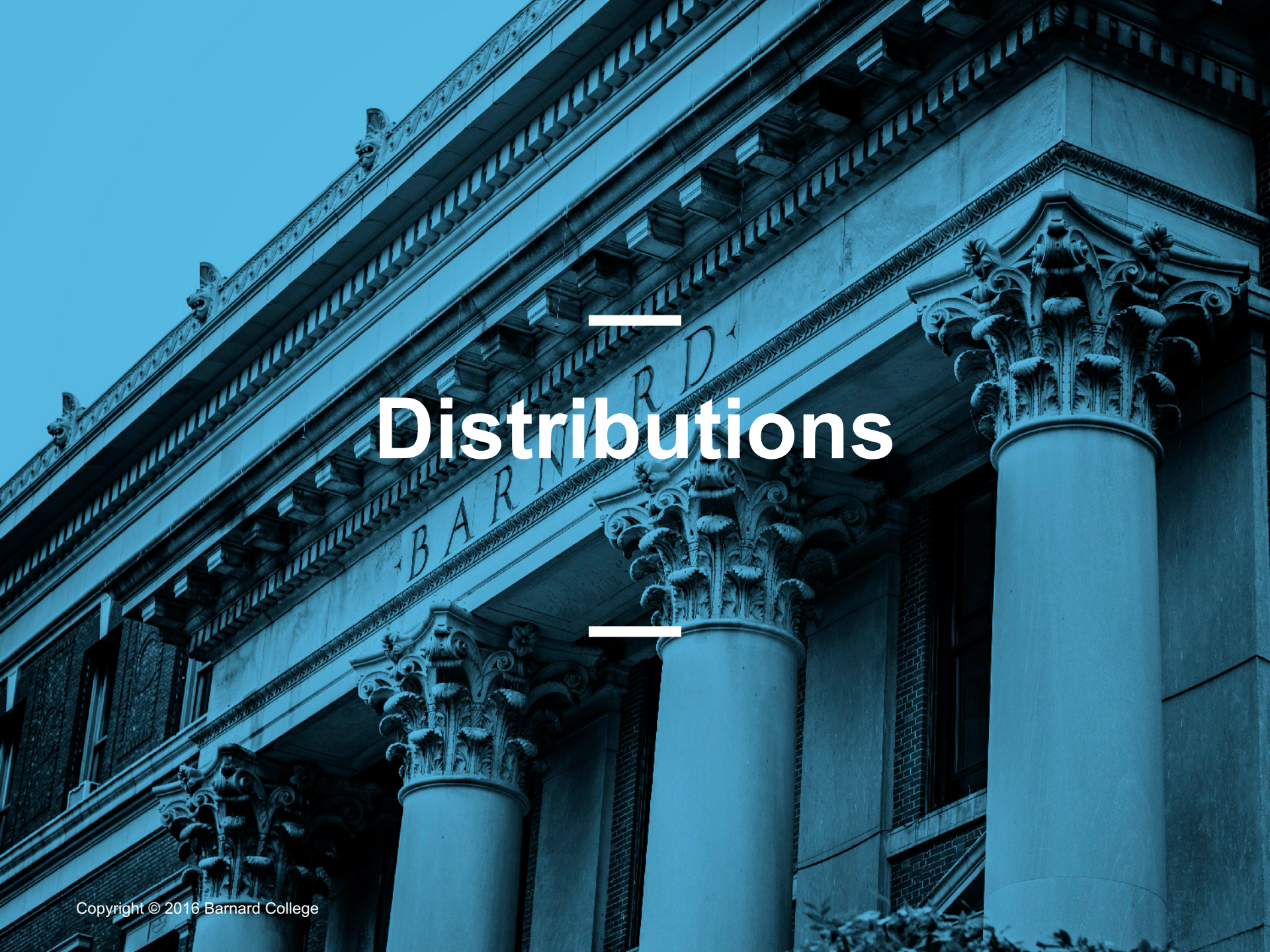
B) P(neither Harmon or Shaibel are in our sample)
= (48/50 * 47/49) = 0.9208

Random Samples

- Deterministic sample:
  - Sampling doesn't involve chance
- Random sample:
  - Before the sample is drawn, you have to know selection probability for each group in the population
  - Note: not every group has to have an equal chance of being drawn
- Uniform Random Sample:
  - Each individual has an equal chance of being selected

Sample of Convenience

- Example: sample consists of whoever walks by
- Doesn't guarantee a "random" sample
- A sample is random if before we sample we have an idea of:
  - the population we are sampling from
  - the chance of selection for each group in our population



Distributions

Probability Distribution

- Random quantity with various possible values
- “Probability Distribution”:
  - All the possible values of a quantity
  - The probability of each of the values
- Computing the probability distribution:
  - Math
  - Simulation …. often easier

Empirical Distribution

- "Empirical" – based on observations
- Observations can be a repeated experiment
- "Empirical Distribution":
  - All observed values
  - The proportion of times each value appears

Large Random Samples

Law of Averages / Law of Large Numbers

If a chance experiment is repeated many times, independently and under the same conditions, then the proportion of times that event occurs gets closer to the theoretical probability of the event

As you increase the number of rolls of a die, the proportion of times you see the face with 5 dots gets closer to 1/6

Empirical Distribution of a Sample

If the sample size is large,

then the empirical distribution of a uniform random sample

resembles the distribution of the population,

with high probability

A Statistic

Inference

- **Statistical Inference:**
  - Making conclusions based on data in random samples

- **Example:**
  - Use the data to guess the value of an unknown number

fixed

Depends on the random sample

  - Create an **estimate** of an unknown quantity

Terminology

- **Parameter**
  - Numerical quantity associated with the population
- **Statistic**
  - A number calculated from the sample
- A statistic can be used as an **estimator** of a parameter

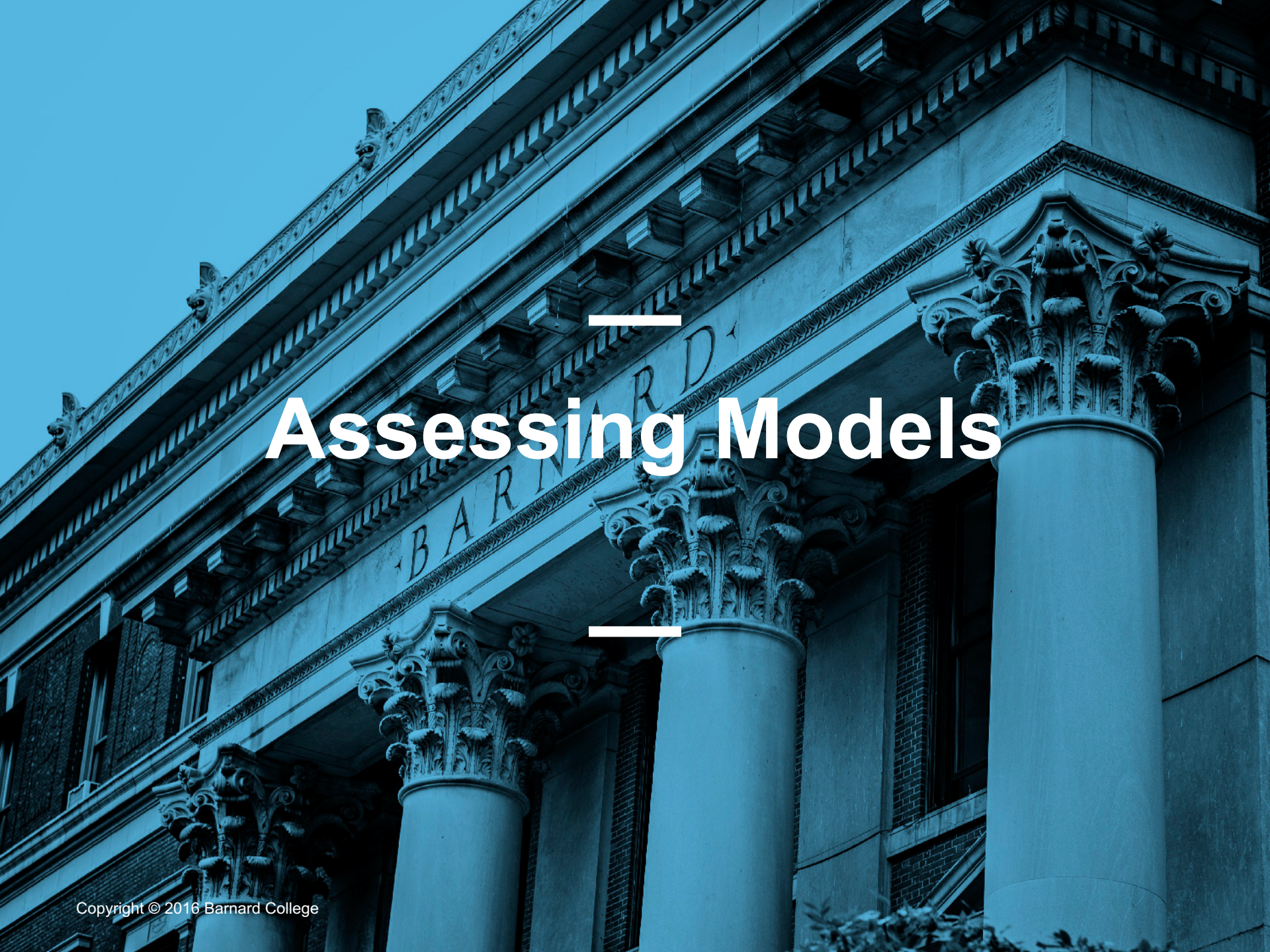
Probability distribution of a statistic

- Values of a statistic vary because random samples vary
- "Sampling distribution" or "probability distribution" of the statistic:
  - All possible values of a statistic
  - and all corresponding probabilities
- Can be hard to calculate:
  - Either have to do math
  - Or generate all possible samples and calculate the statistic based on the each sample

Empirical Distribution of a Statistic

- Based on simulated values of a statistic
- Consists of all observed values of the statistic,
- and the proportion of times each value appeared

- Good approximation to the probability distribution of a statistic
  - If the number of repetitions in the simulation is large



Assessing Models

Models

- A model is a set of assumptions about the data
- In data science, many models involve assumptions about processes that involve randomness:
  - "Change models"
- **Key question:** does the model fit the data?

Approach to Assessing Models

- If we can simulate data according to the assumptions of the model, we can learn what the model predicts
- We can compare the model's predictions to the observed data
- If the data and the model's predictions are not consistent, that is evidence against the model