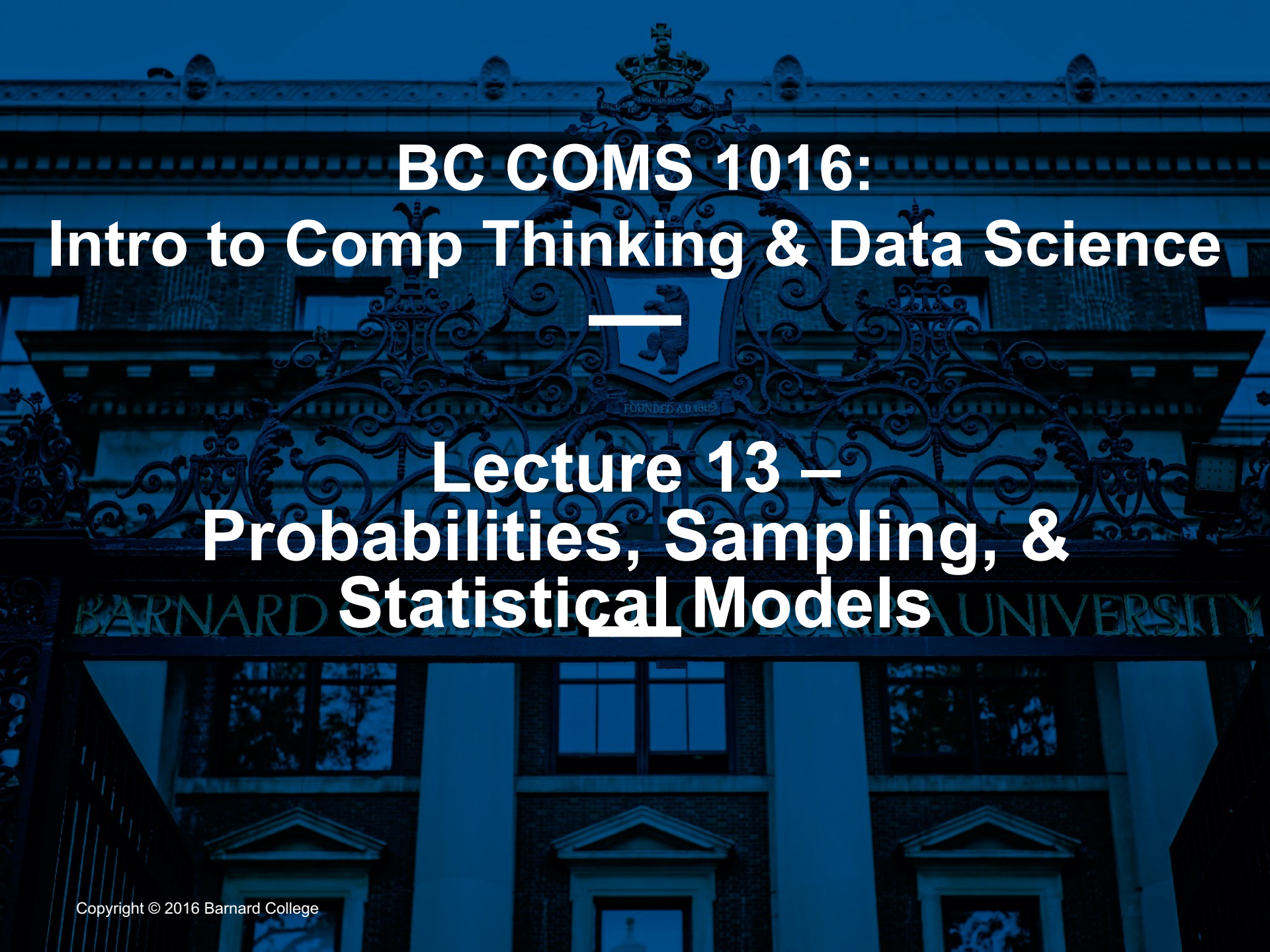


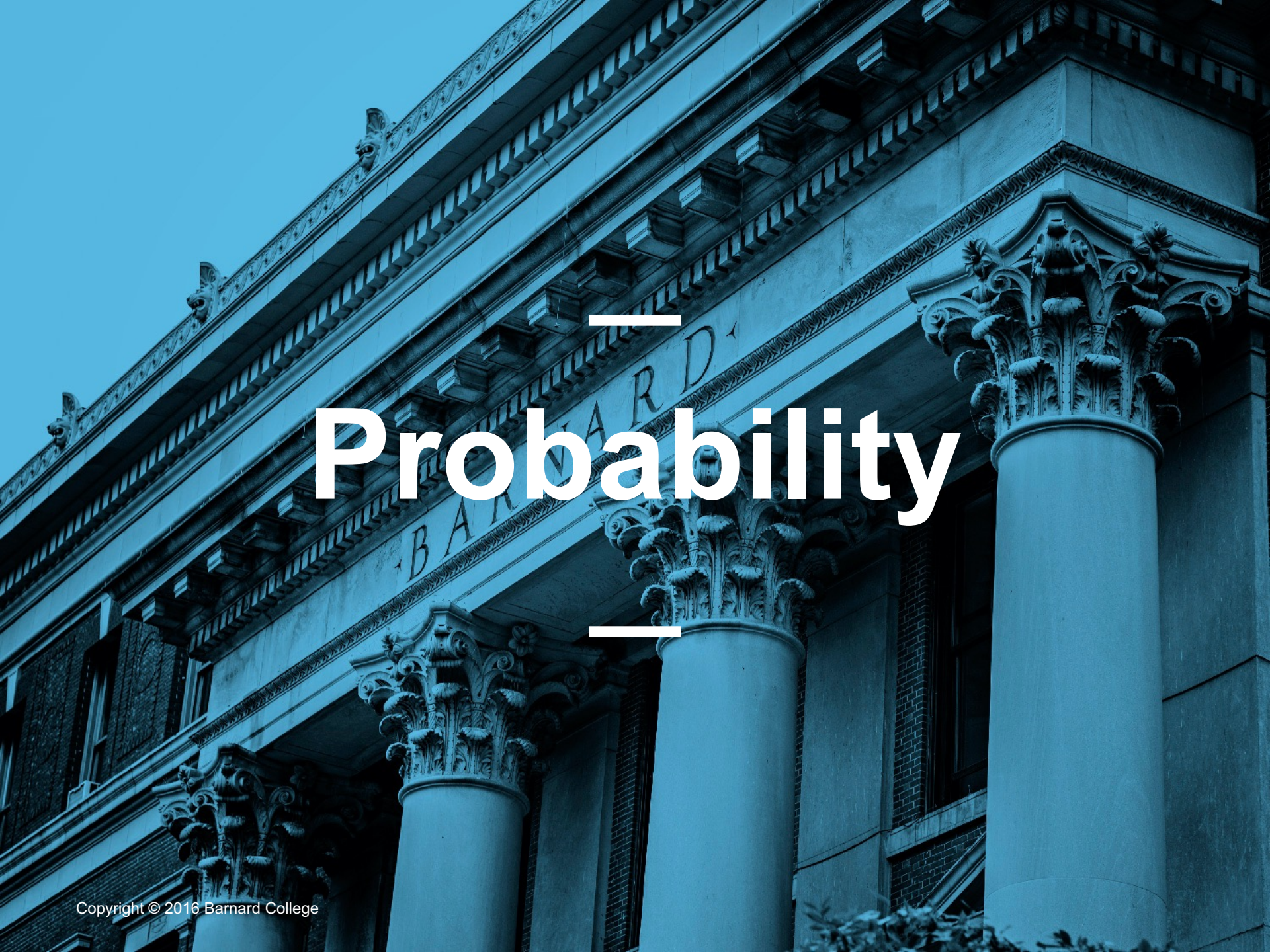
BC COMS 1016: Intro to Comp Thinking & Data Science

—

Lecture 13 – Probabilities, Sampling, & Statistical Models

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



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**

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

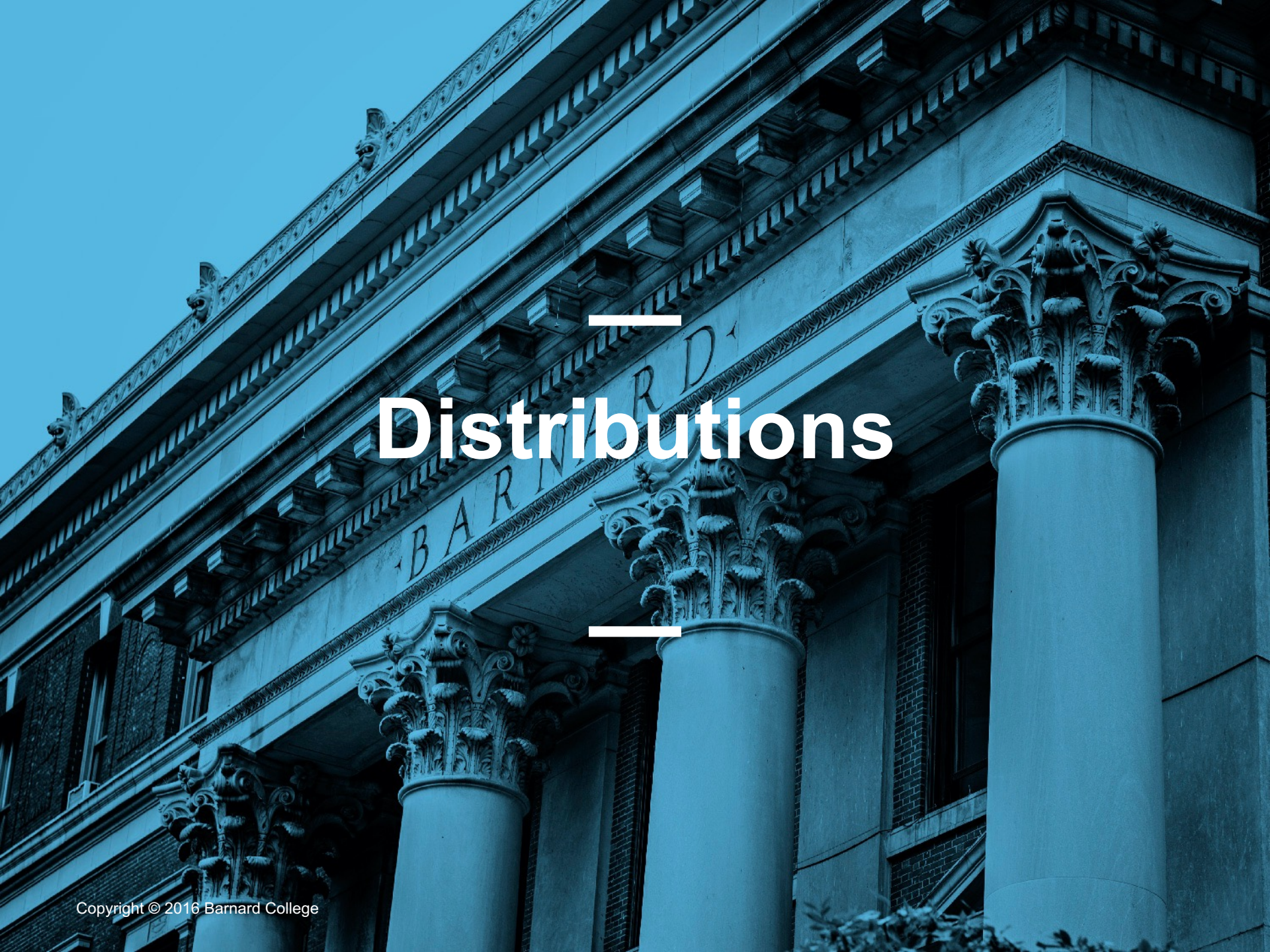
B) P(neither Harmon or Shaibel are in our sample)
= (48/50 * 47/49) = 0.9208

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

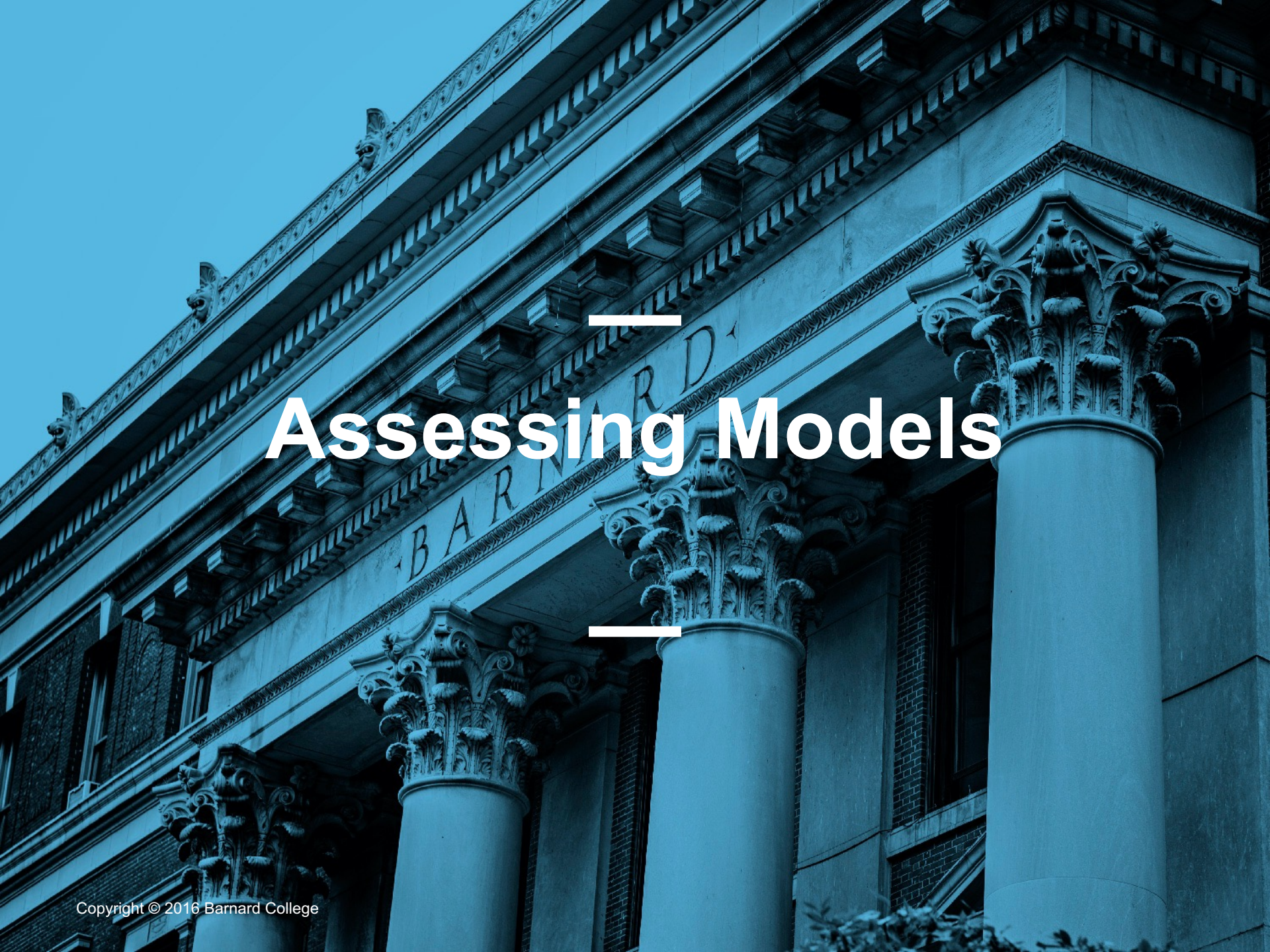
Probability distribution of a statistic

- 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