

COMS BC1016

Introduction to Computational Thinking and Data Science

Lecture 11: Models

BARNARD COLLEGE OF COLUMBIA UNIVERSITY

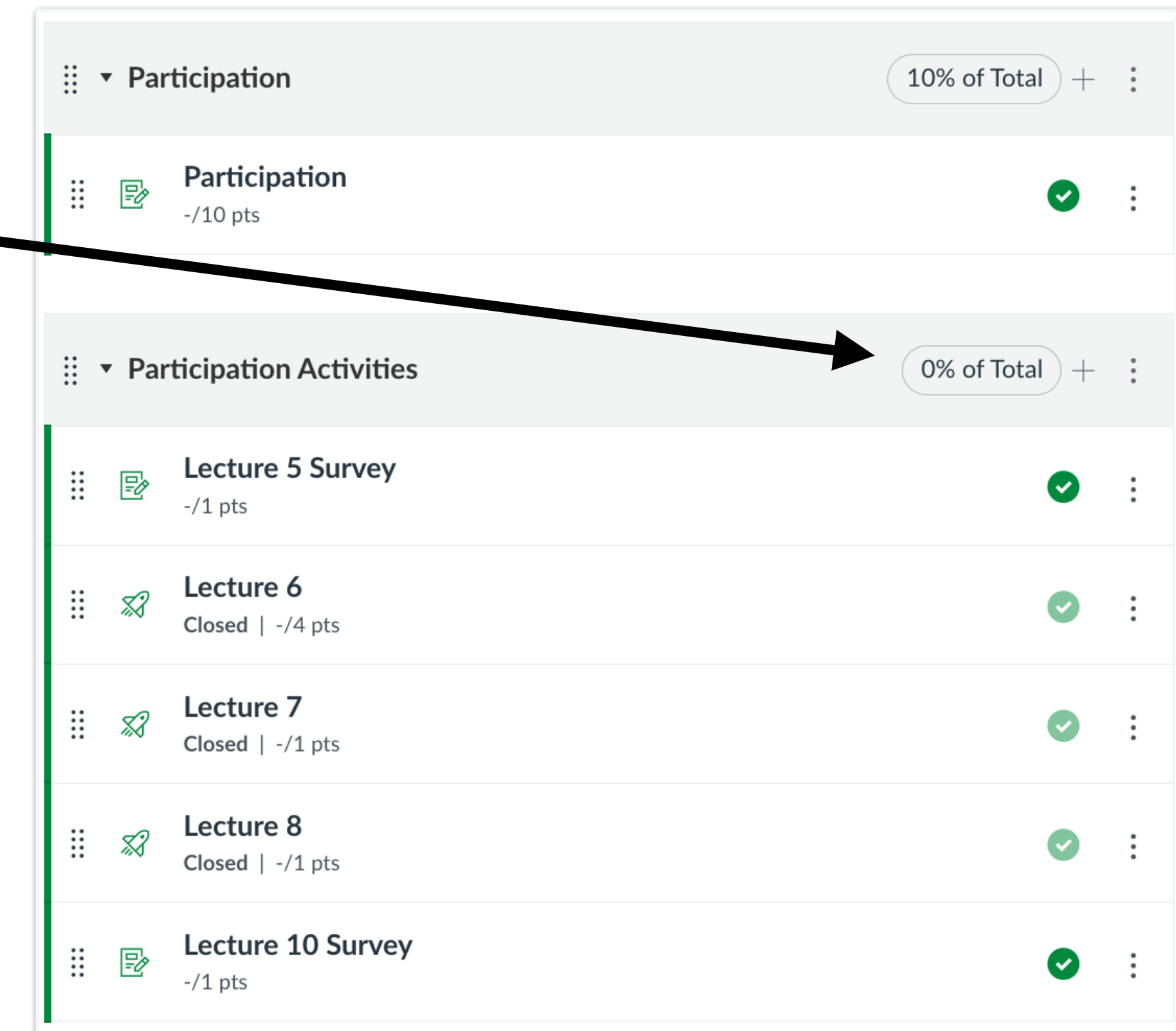
Sep 30, 2025

HW Reminders

- The format of Multiple Choice Questions is the same as HW 2
- Your answer should be put as **a number or an array of numbers** corresponding to the statement(s)
- Do not put the string / array of strings!
- Unless otherwise specified (i.e., a question specifically tells you so), you should be using what we're learning in class for the homeworks
- Use of external libraries/techniques may result in lost points or having to explain your answers for credit

A Note on the Participation Grade

- This is *not* the fraction of in-class surveys/activities you complete
 - These are more for me to gauge who's around, do some simple checks of understanding, and incorporate class data into our demos
 - Updated Courseworks to make it more clear
- Participation grade is influenced by how much you *actively participate*
 - Asking/answering questions in class
 - Office hours (myself or TAs)
 - EdStem
- Hopefully you look like what's on the photo roster because I'm using that to remember people



Review: Simulations

Monty Hall Problem

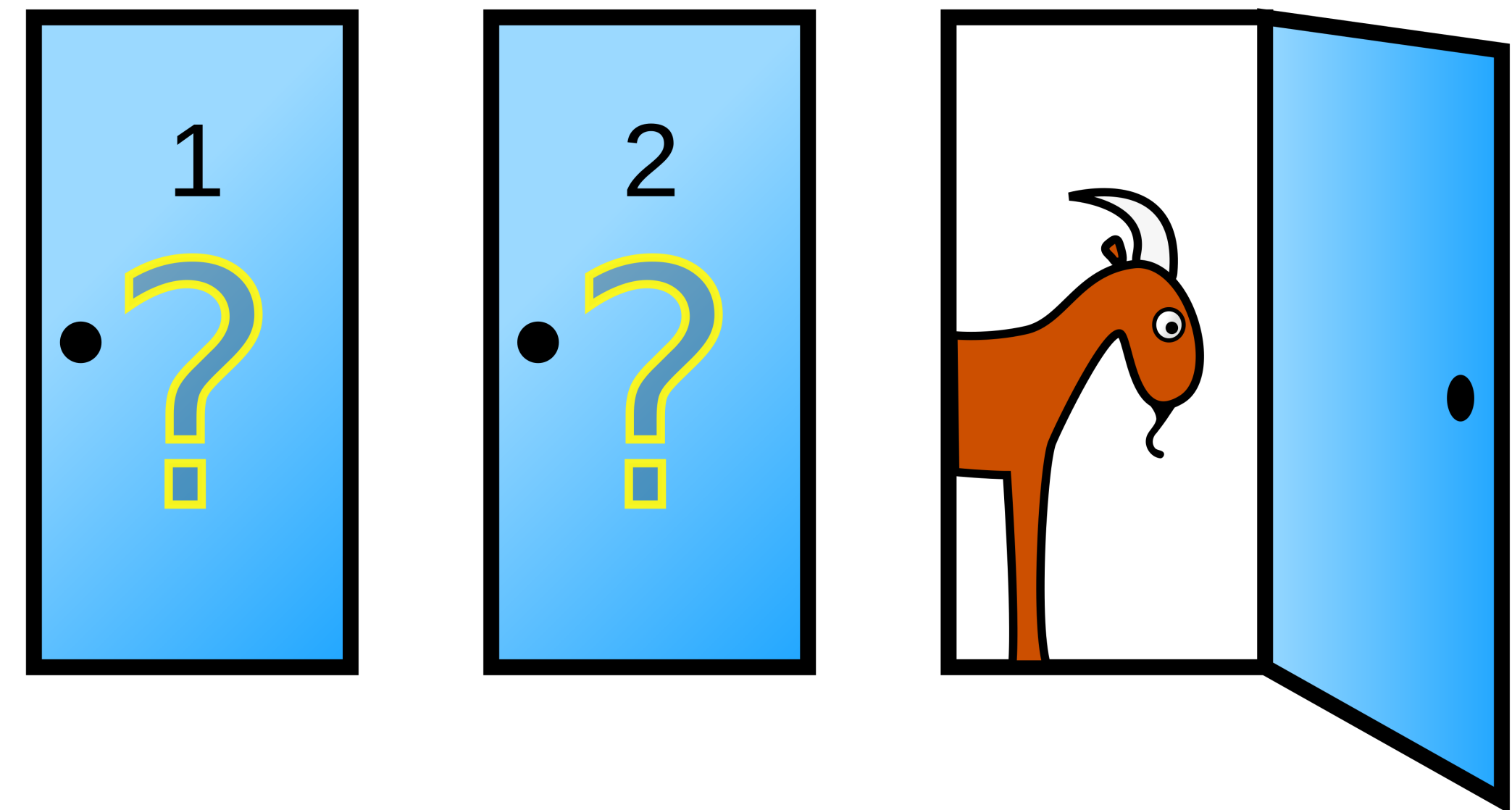
“Suppose you're on a game show, and you're given the choice of three doors:

Behind one door is a car; behind the others, goats.

You pick a door [...] and the host [...] opens another door [...] which has a goat.

He then says to you, "Do you want to [switch]?"

Is it to your advantage to switch your choice?"



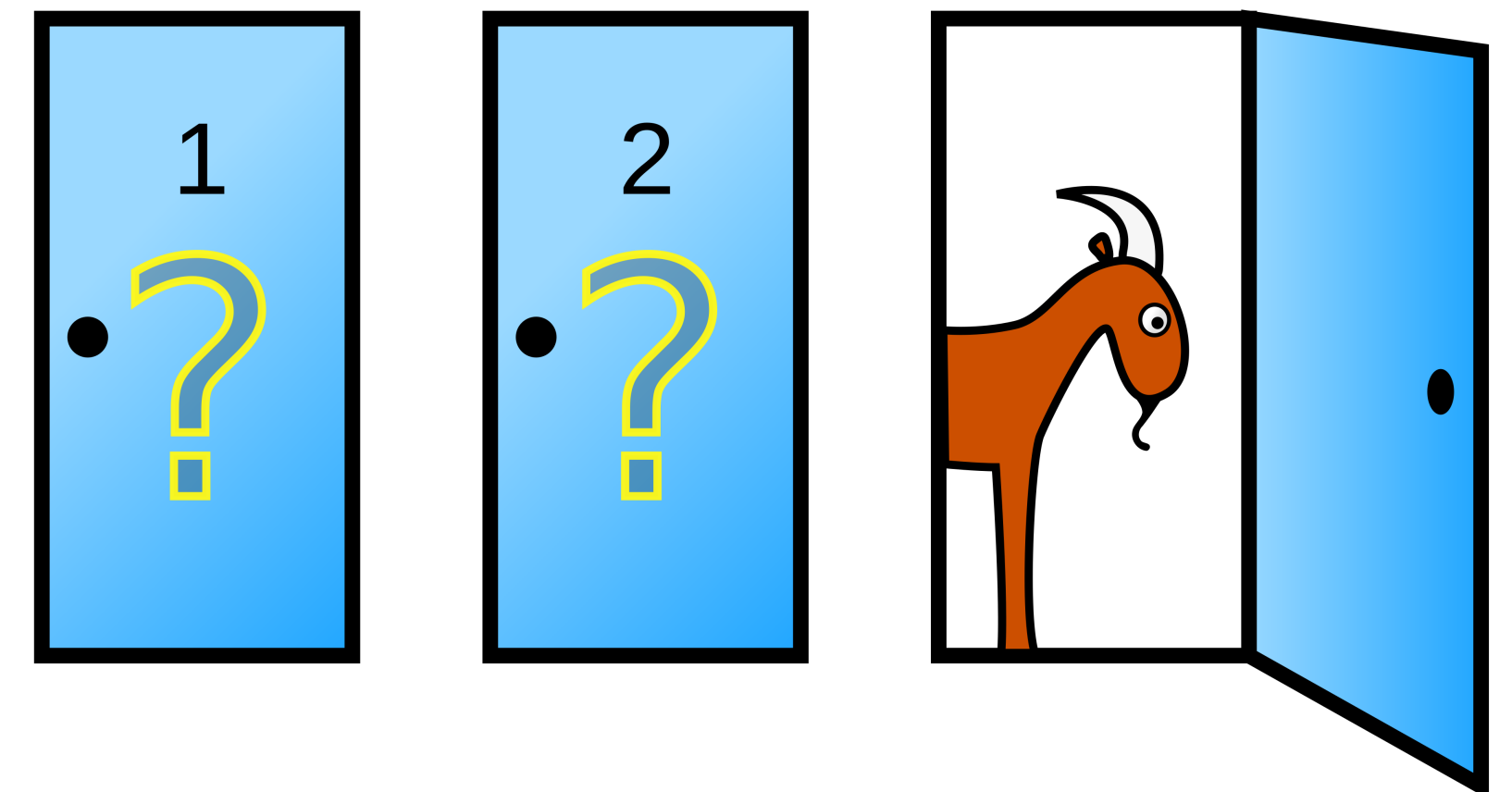
By Cepheus - Own work, Public Domain, <https://commons.wikimedia.org/w/index.php?curid=1234194>

Simulating the Monty Hall Problem

Question: Is it to our advantage if we switch doors?

- Assume the cars and goats are assigned to doors randomly and our initial door is chosen randomly

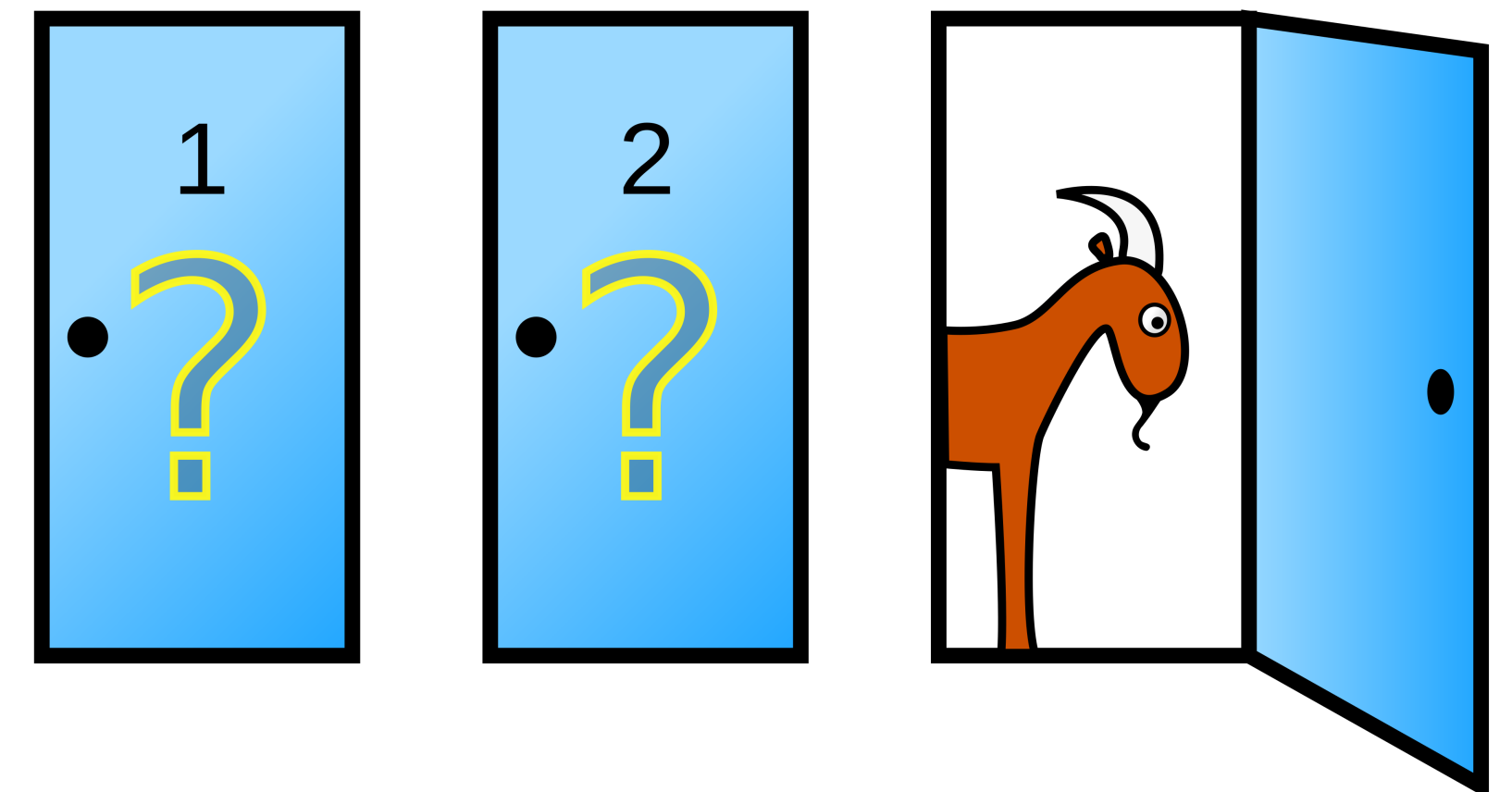
What will help us answer our question?



Process for Simulations

1. Figure out what you want to simulate
2. Write a function whose output is the outcome of a single simulation
3. Repeat the simulation for some number of iterations
4. Add results array to a table so you can plot the results

We started this last class.
Let's rewind and talk about
what we were trying to do



Last Time: Distributions

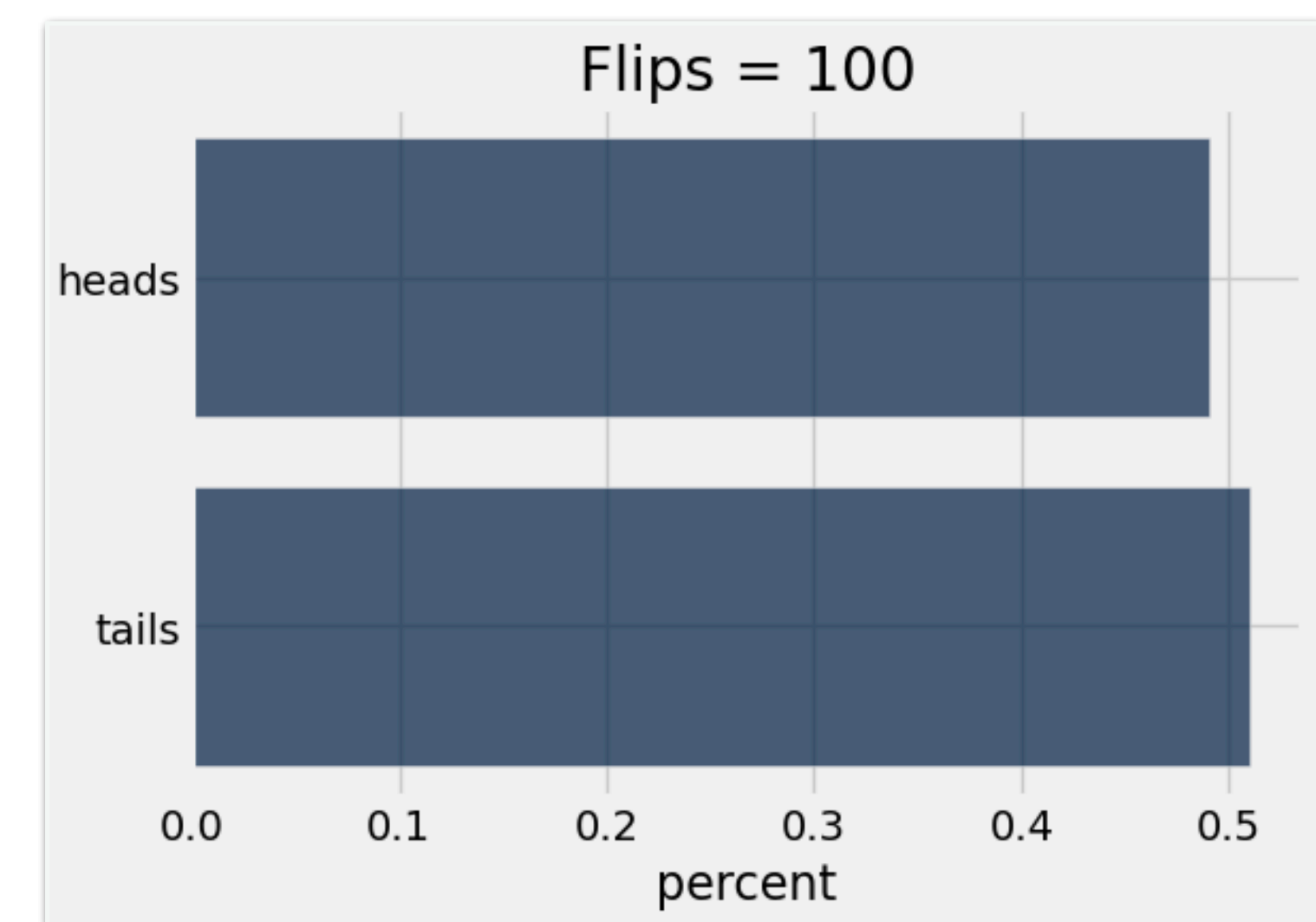
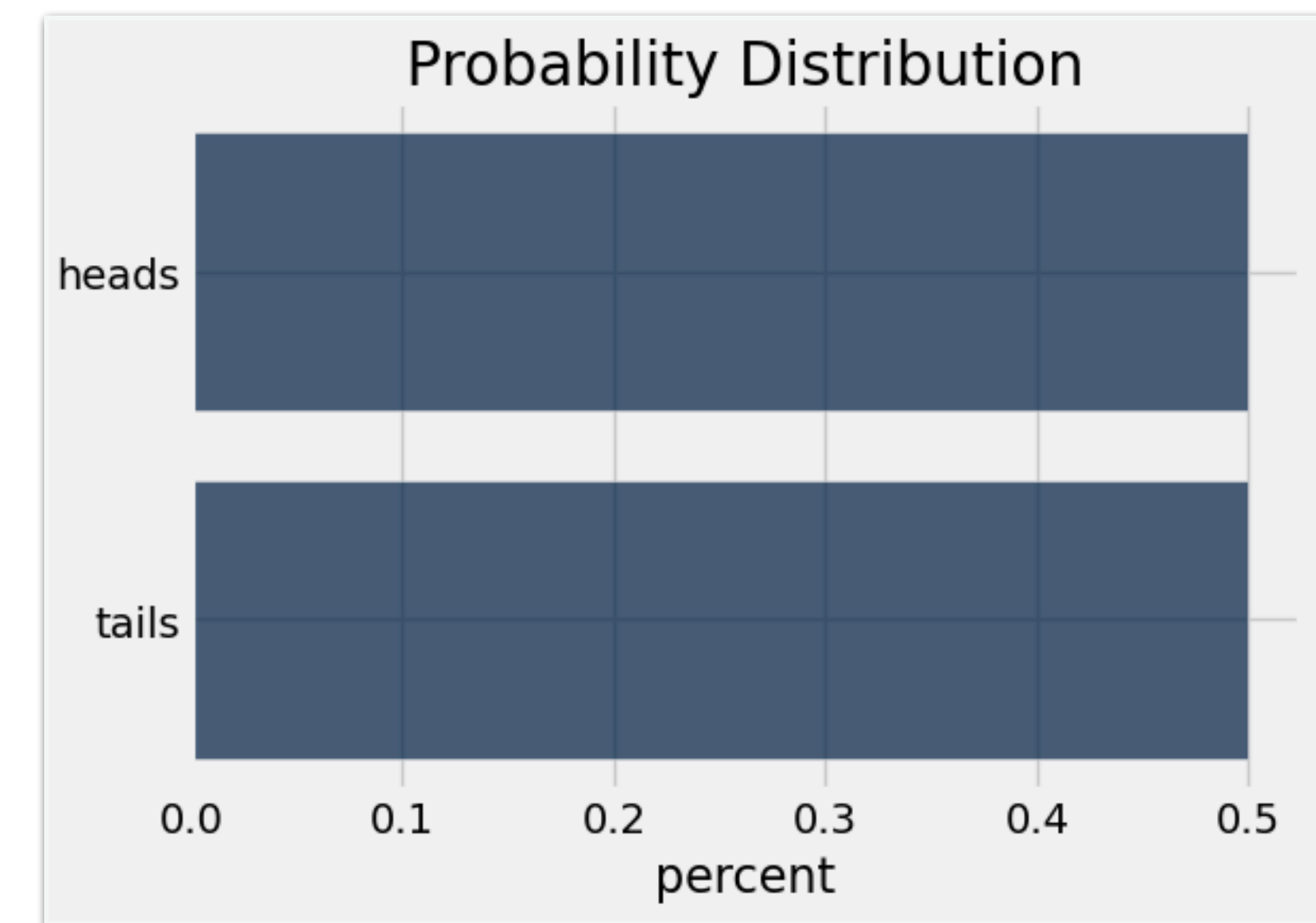
Distributions

Recall categorical distributions (how often each unique value appears)

- If we have a population, we can get a list of unique values and how often they appear
- If we take a sample, the list of unique values may change based on the sample
- There can be differences in what's in the population vs what we see based on taking samples

Distributions

- **Probability Distribution:** All possible values & the probability of each value
 - Example: 50% heads, 50% tails
- **Empirical Distribution:** The observed values and proportion of times it took each value
 - Example: I flipped 100 coins and N of them were heads and rest were tails



Law of Averages / Large Numbers

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

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

Statistical Inference

Statistical Inference: drawing conclusions based on data in random samples

- Create an **estimate** of an **unknown value** using sample data and statistics
- Inference occurs from not being able to know an entire population
 - Estimates change based on the sample you draw
 - Statistics help you measure how much you expect those differences to vary

Important Terminology

Parameter: Number associated with the population

- Example: average, max, min, mean

Statistic: A number calculated from the sample

A statistic can be used as an **estimate** of a parameter

- Example: sample mean, sample max, sample min

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 the statistic and all corresponding probabilities
- Can be hard to calculate
 - Either have to do the math
 - Or have to generate all possible samples and calculate statistic based on each sample

Empirical Distribution of a Statistic

- Empirical distribution of a statistic:
 - Based on simulated values of the statistic
 - Consists of all the observed values of a statistic and the proportion of times each value appeared
- Good approximation to the probability distribution of the statistic if the number of repetitions in the simulation is large

Demo: Probability Distributions and Empirical Distributions

Assessing Models

Models

A **model** is a set of assumptions about data

- In data science, many models involve assumptions about the processes that involve randomness
- Question: Does the model fit the data?

Assessing Models

- Suppose we have a statistical model that describes how data should behave
 - If we simulate data according to the assumptions of the model, we can learn what the model predicts
- We can then compare the predictions to the data that were observed (irl data)
- If the data and the model's predictions are not consistent, that is evidence against the model.

1960s Supreme Court Case: *Swain v Alabama*

Amendment VI of the U.S. Constitution:

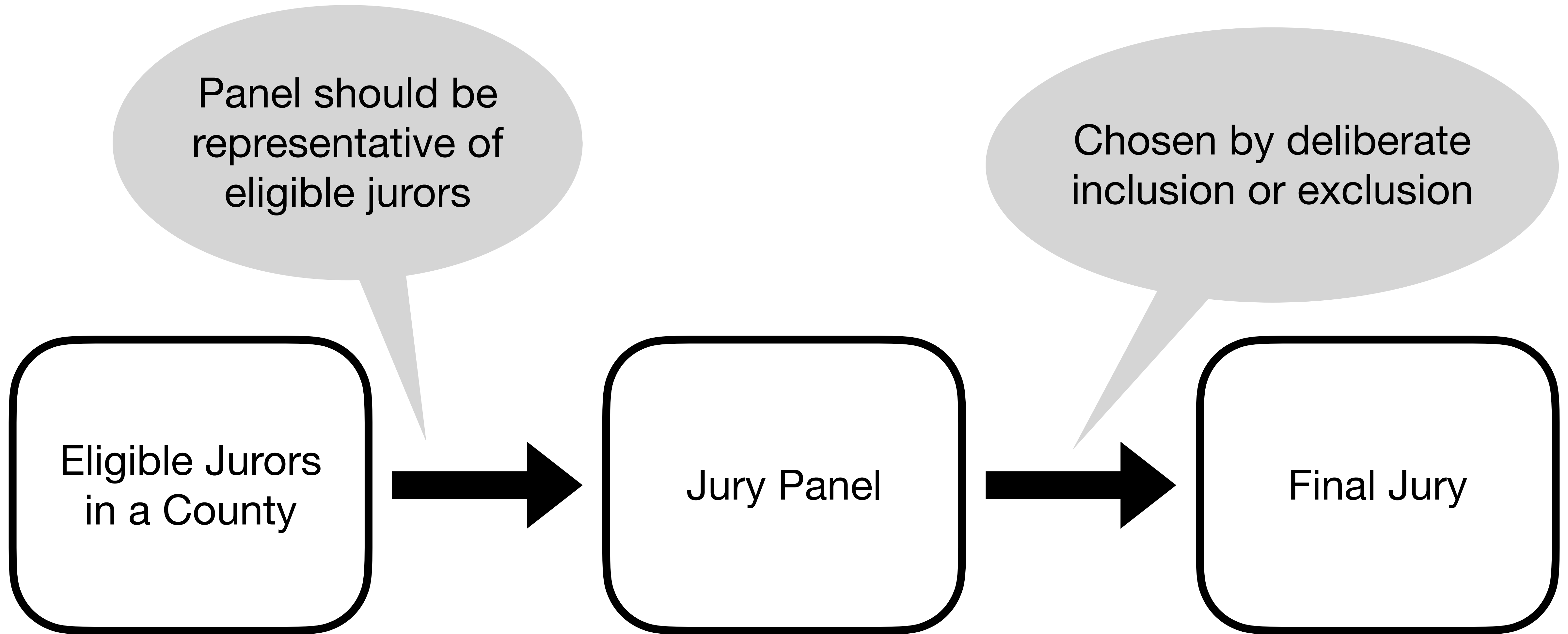
“In all criminal prosecutions, the accused shall enjoy the right to a speedy and public trial, by an impartial jury of the State and district wherein the crime shall have been committed”

- An *impartial* jury should be selected from a panel that is representative from the population of the relevant region

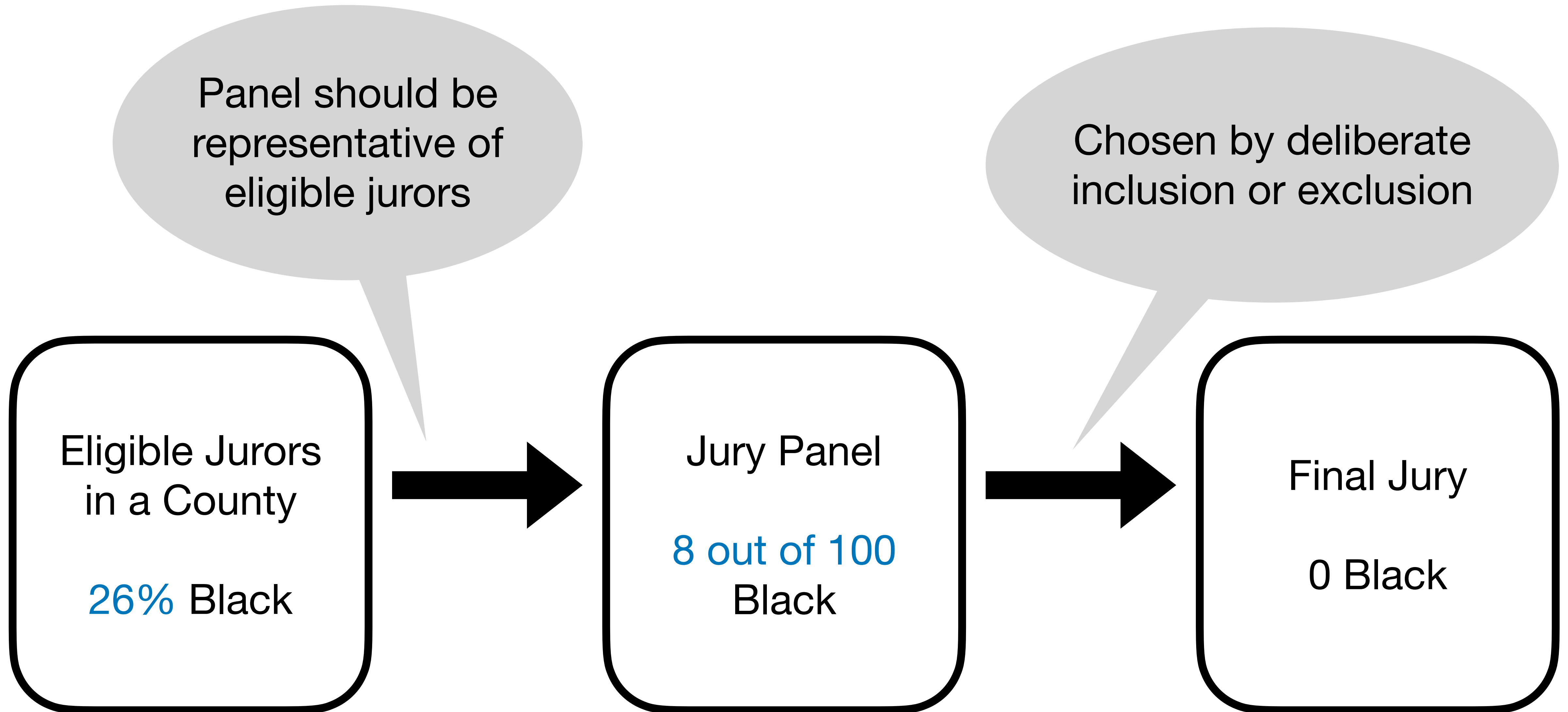
Robert Swain was a Black man indicted and convicted by an all white jury in Talladega County, Alabama

- He appealed to the Supreme Court due to the lack of representation on juries in Talladega County

Jury Selection



Swain vs. Alabama Jury



The Appeal

- Swain's argument: Unfair jury selection

Panel should be
representative of
eligible jurors

Eligible Jurors
in a County

26% Black



Jury Panel

8 out of 100
Black

Supreme Court Ruling, 1965

- Swain's argument: Unfair jury selection
- Supreme Court denied the appeal

Panel should be
representative of
eligible jurors

Eligible Jurors
in a County

26% Black



Jury Panel

8 out of 100
Black

Supreme Court wrote:
“The overall percentage
disparity has been small”

Our Question

- **Court's view:** 8/100 is less than 26%, but not different enough to show Black panelists were systematically excluded
- **Question:** Would an 8% black jury be a realistic outcome if jury selection were truly unbiased?

Sampling from a Categorical Distribution

- Sample at random from a categorical distribution
 - `sample_proportions(sample_size, pop_distribution)`
- `pop_distribution` is a list or array that adds up to 1
- Samples at random from the population
 - Returns an array containing the empirical distribution of the categories in the sample

Steps in Assessing a Model

1. Choose a **statistic** that will help you decide whether the data supports the **model** or an **alternative view** of the world
2. **Simulate the statistic** under the assumptions of the model
3. Draw a histogram of the simulated values
 - This is the model's prediction for how the statistic should come out
4. **Compute the statistic** from the sample in the study
 - If the two are not consistent: evidence against the model
 - If the two are consistent: data supports the model *so far*

Assessing *Swain v Alabama*

1. Choose a statistic that will help you decide whether the data supports the model or an alternative view of the world
2. Simulate the statistic under the assumptions of the model
3. Draw a histogram of the simulated values
4. Compute the statistic from the sample in the study

Assessing *Swain v Alabama*

1. Choose a **statistic** that will help you decide whether the data supports the **model** or an **alternative view** of the world
2. Simulate the statistic under the assumptions of the model
3. Draw a histogram of the simulated values
4. Compute the statistic from the sample in the study

Model:

Panelists were selected at random.
The number of Black panelists is by *chance*

Alternative View:

Too few Black panelists for it to have been a truly random sample

Statistic: Number of Black panelists

Demo:

Swain vs. Alabama

Our Conclusion

- Percent of Black panelists (8%) was **highly unlikely** under random sampling
 - It's *possible*, but extremely unlikely
 - Suggests the assumptions about the model are wrong
- We observed **statistical bias**, when differences between the parameters and the statistics are systematically in one direction
- The verdict was eventually overruled in *Baton v. Kentucky* (1986)
 - "The dismissal of jurors without stating a valid cause for doing so—may not be used to exclude jurors based solely on their race."

Next Class

- Today
 - Simulations
 - Distributions
 - Models
- Monday
 - More on Models
 - Empirical Simulations