RANDOMNESS IN DATA

ST101 – DR. ARIC LABARR

PROBABILITY (AND RISK)

RANDOMNESS IN DATA

CHANCE

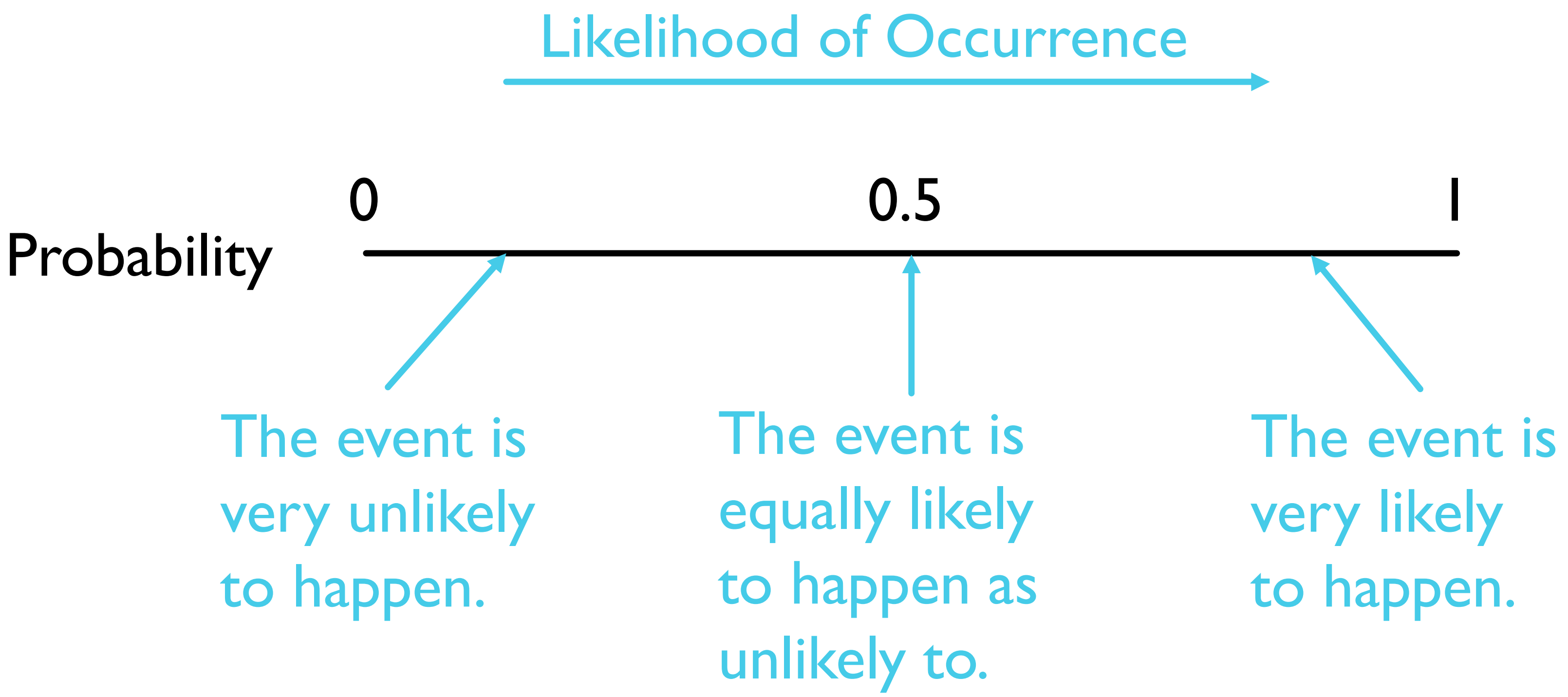
- **Random** – an outcome is random if we know the particular outcomes that something could have but are unsure of which of those outcomes is about to happen.
- People throughout history have tried to measure patterns in randomness and answer the question, "what would happen if we did this many times?"

CHANCE

- Try flipping a coin.
- Each flip is completely random → you are unsure of the specific outcome.
- If your coin is fair (evenly weighted), then in many flips you should get approximately 50% heads and 50% tails.

PROBABILITY

- The **probability** that an event happens is a numerical measure of the likelihood of that event's occurrence.



PROBABILITY

- The **probability** that an event happens is a numerical measure of the likelihood of that event's occurrence.
  - Probabilities are numbers between 0 and 1.
  - Percentages are numbers between 0 and 100.

- **Sample space**: the collection of **all** possible outcomes in a random process.
  - Sum of all probabilities for an experiment must sum to 1.

EVENTS

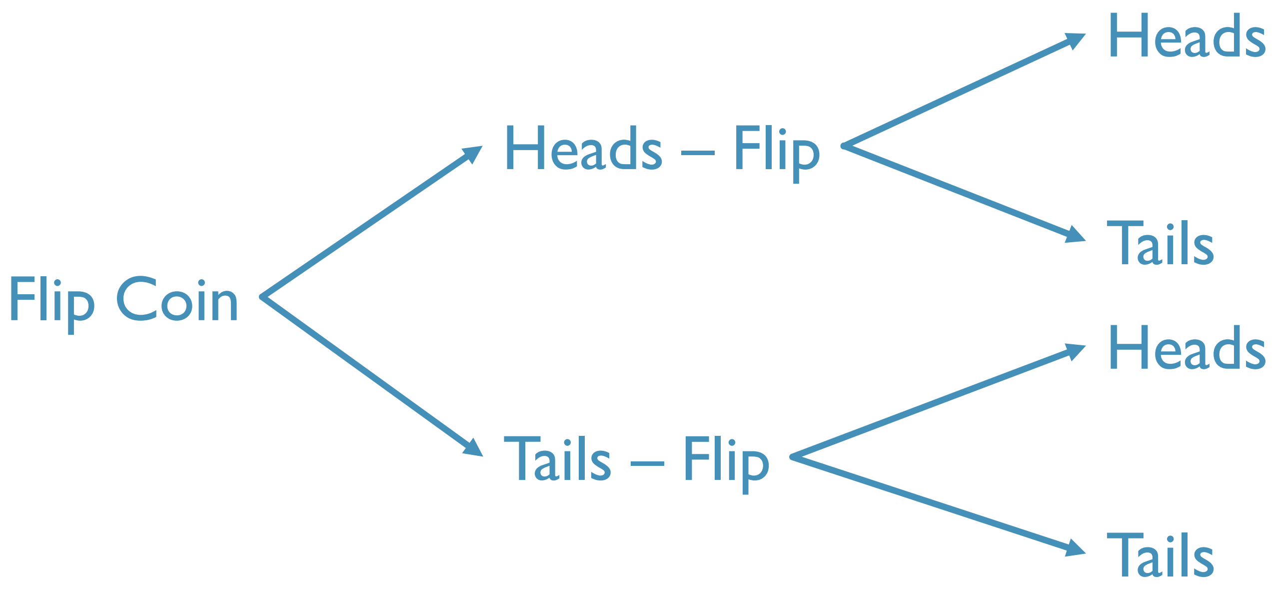
- An **event** is a collection of one or more outcomes from a process whose result cannot be predicted with certainty.
- The probability of an event *A* is denoted, *P(A)*.
- Examples:
  - When flipping a fair dice, what is the probability of it landing on heads?
  - When rolling a fair dice, what is the probability of rolling a 6?

MULTI-STEP RANDOM PROCESSES

- If a random process consists of a sequence of k steps in which there are a certain number of outcomes per step, then the total number of outcomes is the multiplication of these.
- For example:
  - Step 1 has n_1 outcomes
  - Step 2 has n_2 outcomes
  - Step 3 has n_3 outcomes
  - Step 4 has n_4 outcomes
  - The total number of outcomes is $n_1 \times n_2 \times n_3 \times n_4$

TREE DIAGRAMS

- Multi-step random processes can be visualized easily with tree diagrams.
- For example, you have a 2-step random process where you flip a coin twice:



ASSIGNING PROBABILITIES

- Probabilities of an event occurring must be between 0 and 1.
- The sum of the probabilities of **all** events in an experiment must equal 1.
- There are three typical methods for assigning probabilities to events:
  1. Classical Method
  2. Relative Frequency Method
  3. Subjective Method

CLASSICAL METHOD

- The classical method of assigning probabilities assumes that all events have **equally likely outcomes**.
- If an experiment has n possible outcomes, then each outcome gets a probability of $1/n$.
- For example, rolling a die has the following sample space:

$$S = \{1,2,3,4,5,6\}$$

- Each with the probability of $1/6$

RELATIVE FREQUENCY METHOD

- The relative frequency method of assigning probabilities assigns probabilities based on **experimentation or historical data**.
- For example, you don't believe that I have a fairly weighted dice so you ask me to roll it 100 times and get the following:

Value of the Roll	Frequency	Experimental Probability
1	10	0.10
2	25	0.25
3	42	0.42
4	7	0.07
5	10	0.10
6	6	0.06

SUBJECTIVE METHOD

- Circumstances might change rapidly in the events you are trying to build probabilities for, so things shouldn't be based solely on historical data.
- Use both a combination of historical data values as well **experience and intuition** about how likely an event will be to occur.
- Best probability estimates are typically a combination of subjective and classical/relative frequency methods.

SUMMARY

- An outcome is random if we know the particular outcomes that something could have but are unsure of which of those outcomes is about to happen.
- The probability that an event happens is a numerical measure of the likelihood of that event's occurrence.
  - Classical Method
  - Relative Frequency Method
  - Subjective Method
- An event is a collection of one or more outcomes from a process whose result cannot be predicted with certainty.

LAW OF LARGE NUMBERS

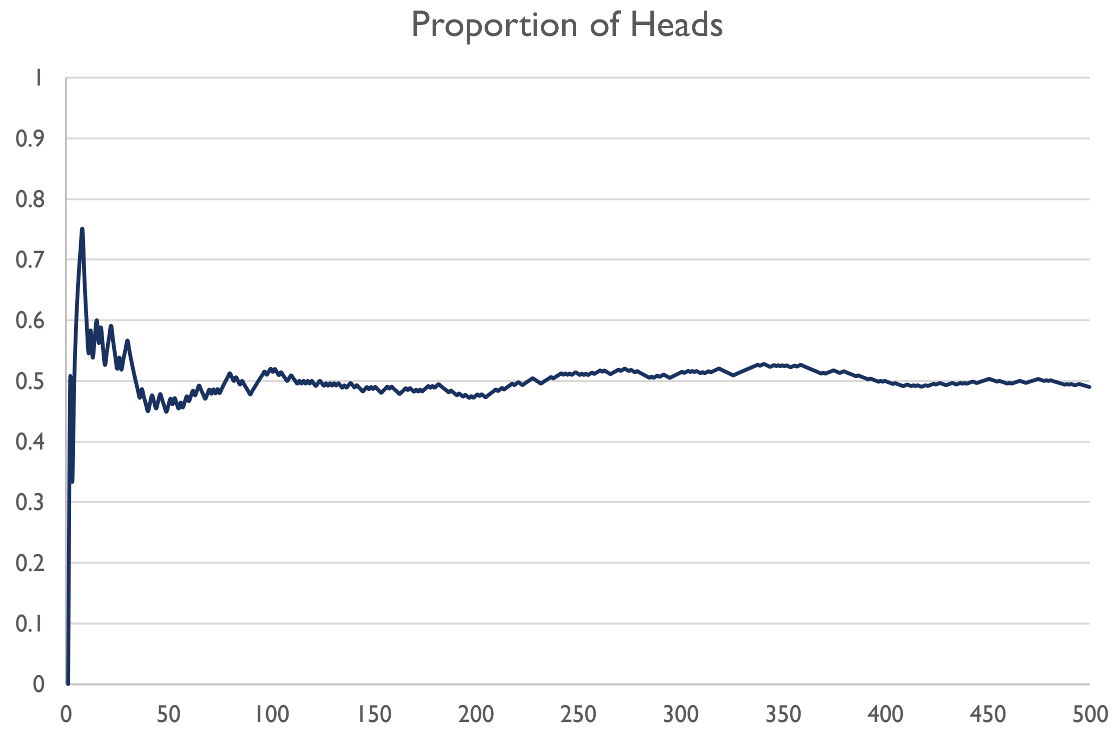
RANDOMNESS IN DATA

CHANCE

- Try flipping a coin.
- Each flip is completely random → you are unsure of the specific outcome.
- If your coin is fair (evenly weighted), then in many flips you should get approximately 50% heads and 50% tails.
- Chance behavior is unpredictable in the short run, but predictable in the long run.

EXAMPLE – TOSS A COIN

- Toss a coin 500 times and record the proportion of heads as you go.
- Early on, the proportion of heads can vary drastically.
- In the long run, it goes to what we expect.

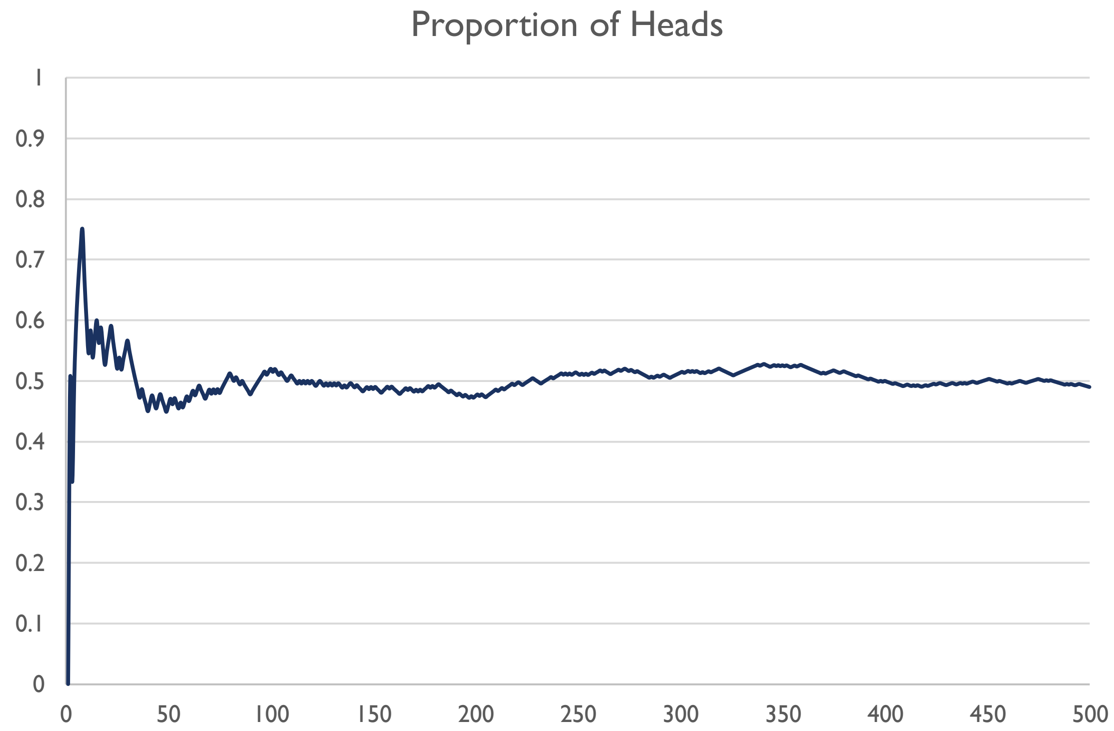


LAW OF LARGE NUMBERS

- If we flip coin enough times, the overall proportion of times it lands on heads (or tails) gets closer to 50%.
- Reasonable to assume that it is a fair coin – half of the time it lands on heads.
- The **law of large numbers** states that as the number of independent trials increases, *in the long run* the proportion for a certain event gets closer and closer to a single value (the probability of the event).

LAW OF LARGE NUMBERS

- The **law of large numbers** states that as the number of independent trials increases, *in the long run* the proportion for a certain event gets closer and closer to a single value (the probability of the event).



MYTH OF SHORT RUN PREDICTABILITY

- Chance behavior is unpredictable in the short run, but predictable in the long run.
- This is counter-intuitive to most people.

- Example – Which of the following outcomes of flipping a fair coin 4 times is more probable?

H, T, T, H   H, H, H, H

MYTH OF SHORT RUN PREDICTABILITY

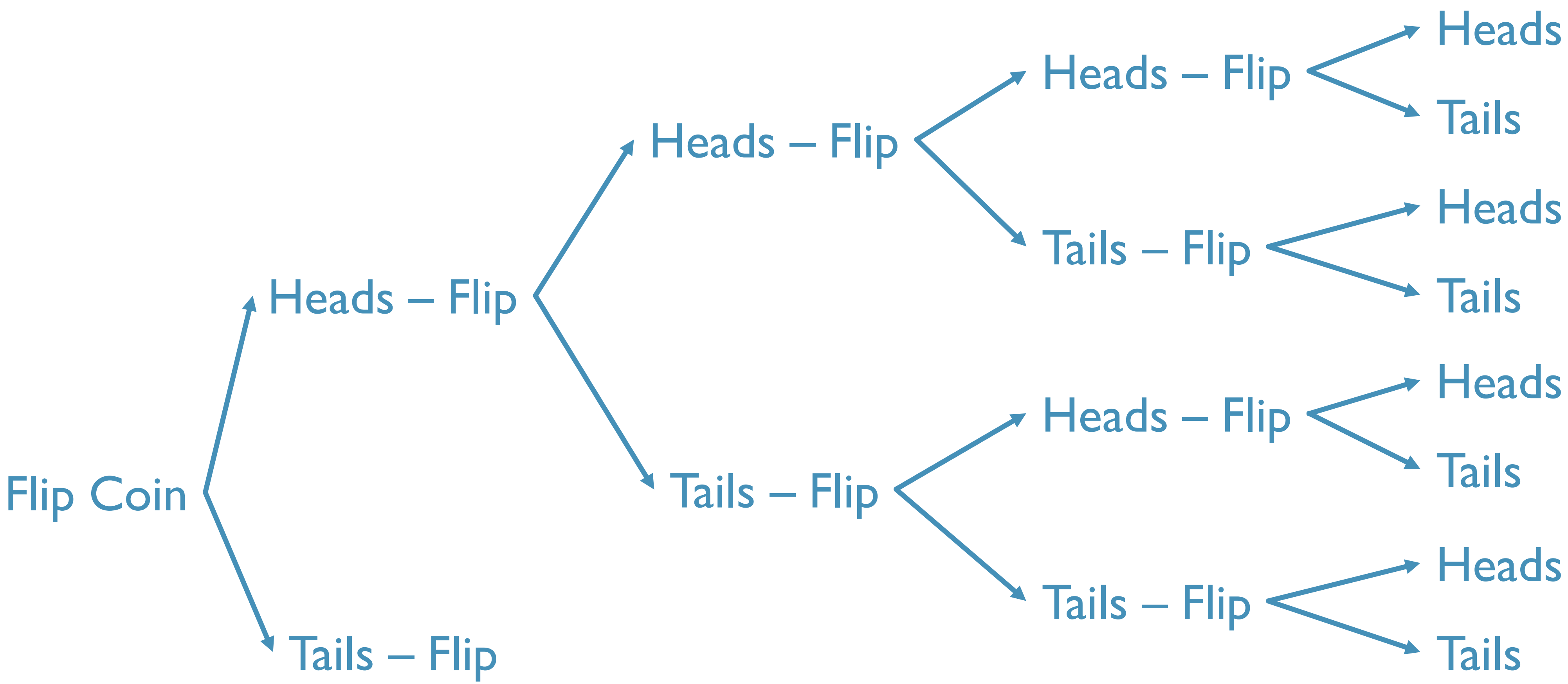
- Chance behavior is unpredictable in the short run, but predictable in the long run.
- This is counter-intuitive to most people.

- Example – Which of the following outcomes of flipping a fair coin 4 times is more probable?

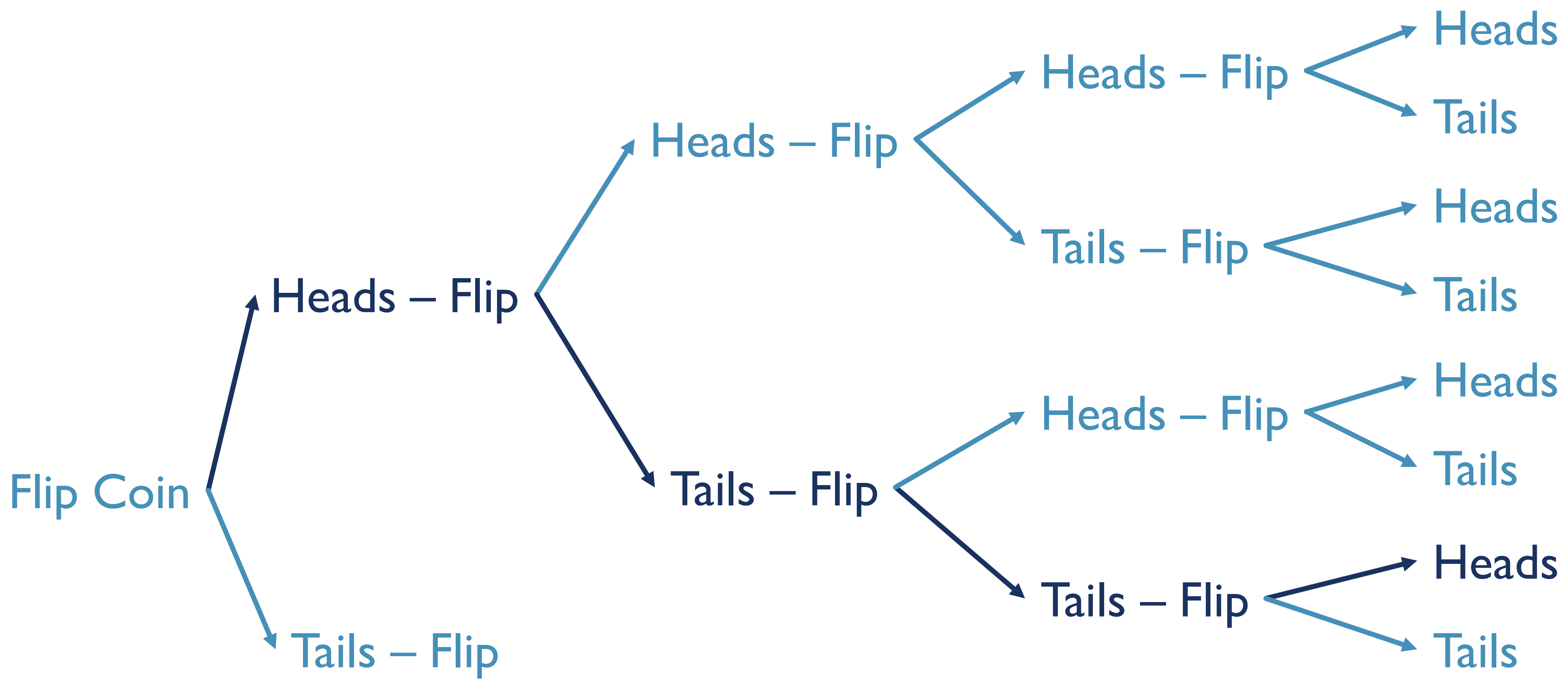
H, T, T, H   H, H, H, H

SAME!

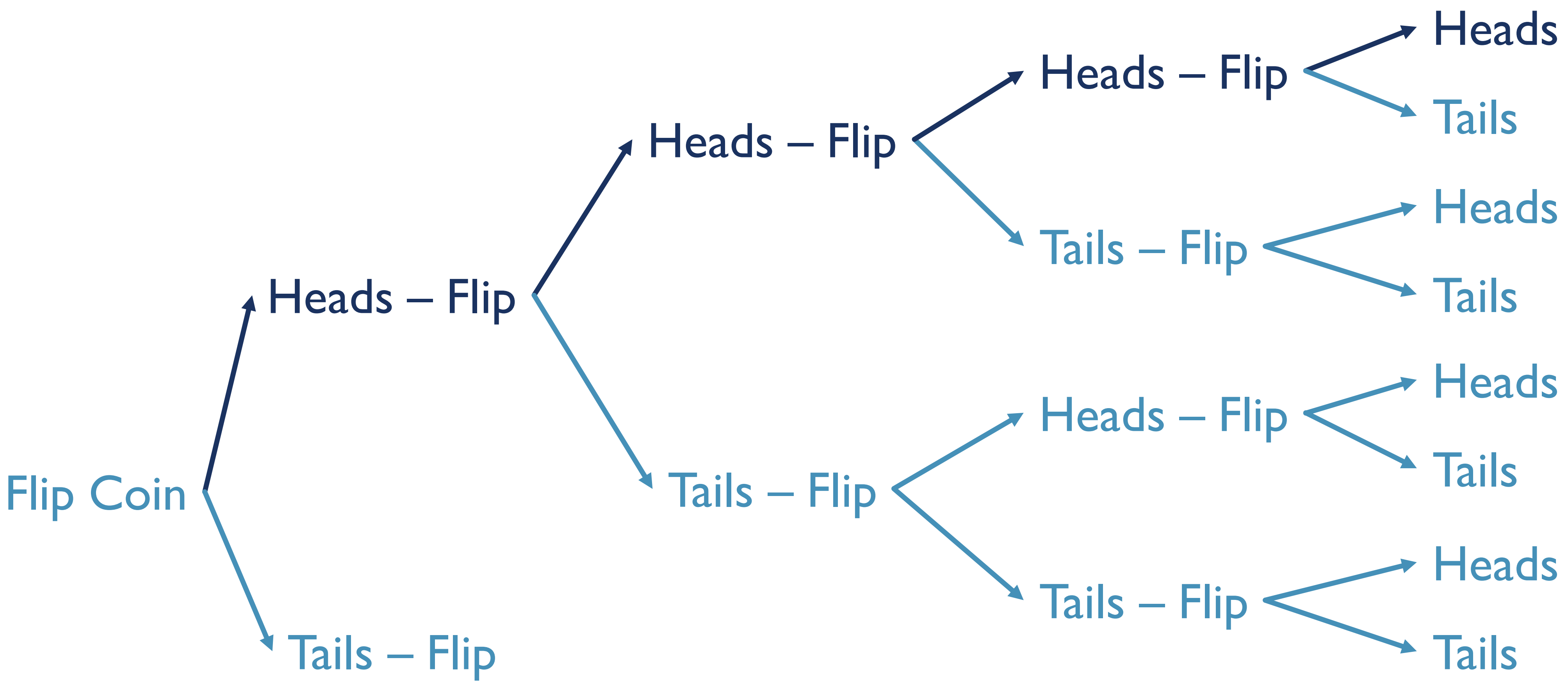
MYTH OF SHORT RUN PREDICTABILITY



MYTH OF SHORT RUN PREDICTABILITY



MYTH OF SHORT RUN PREDICTABILITY



MYTH OF THE "HOT HAND"

- Chance behavior is unpredictable in the short run, but predictable in the long run.
- This is counter-intuitive to most people.

- Example – A player in basketball makes 4 straight shots as compared to makes 1, misses 2, then makes 1. Same idea as the coins!

H, T, T, H　　　　H, H, H, H

MYTH OF THE "LAW OF AVERAGES"

- Chance behavior is unpredictable in the short run, but predictable in the long run.
- This is counter-intuitive to most people.

- Example – A couple has four kids that are all boys. They think their 5th kid must "be due" to have a girl. Same idea as the coins! Still 50/50.

MYTH OF THE "LAW OF AVERAGES"

- Chance behavior is unpredictable in the short run, but predictable in the long run.
- This is counter-intuitive to most people.
- The "law of averages" doesn't exist → things even out in the long run, but chances of outcomes do not change.

SUMMARY

- Chance behavior is unpredictable in the short run, but predictable in the long run.
- The law of large numbers states that as the number of independent trials increases, *in the long run* the proportion for a certain event gets closer and closer to a single value (the probability of the event).
- The "law of averages" doesn't exist → things even out in the long run, but chances of outcomes do not change.

BASIC PROBABILITY RULES

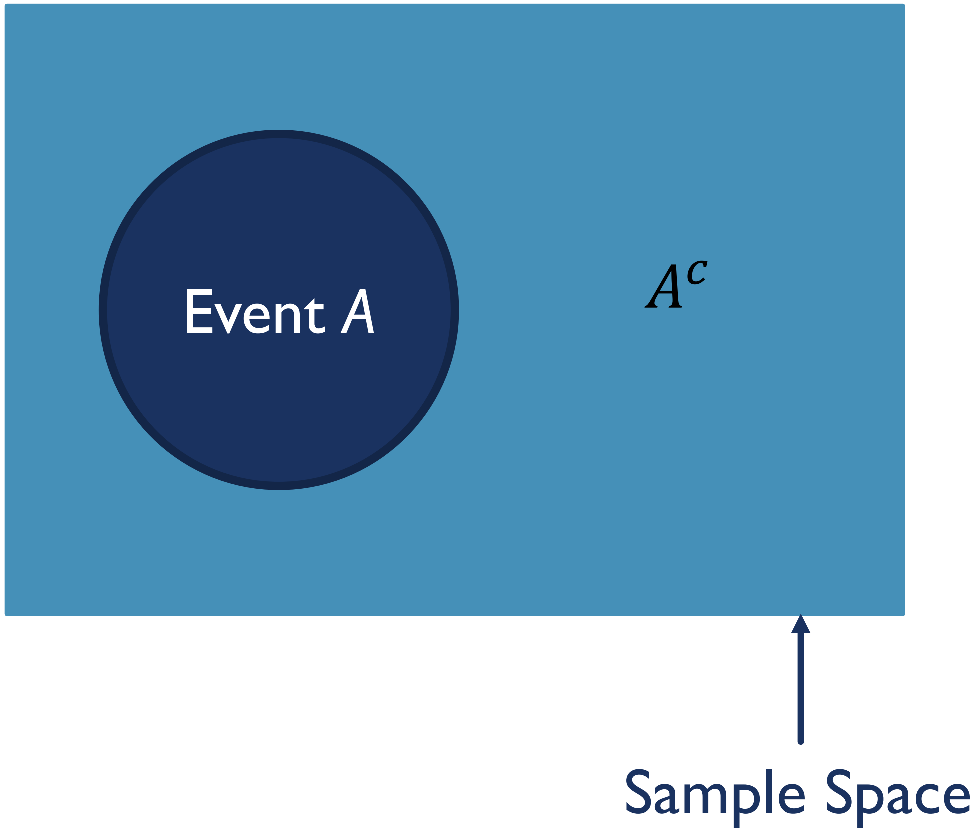
RANDOMNESS IN DATA

BASIC RELATIONSHIPS

- You do not always know all of the sample point probabilities in an event.
- However, there are some basic probability relationships that still can be used to calculate the probability of an event occurring:
  - Complement of an Event
  - Union of Two Events
  - Intersection of Two Events
  - Mutually Exclusive Events

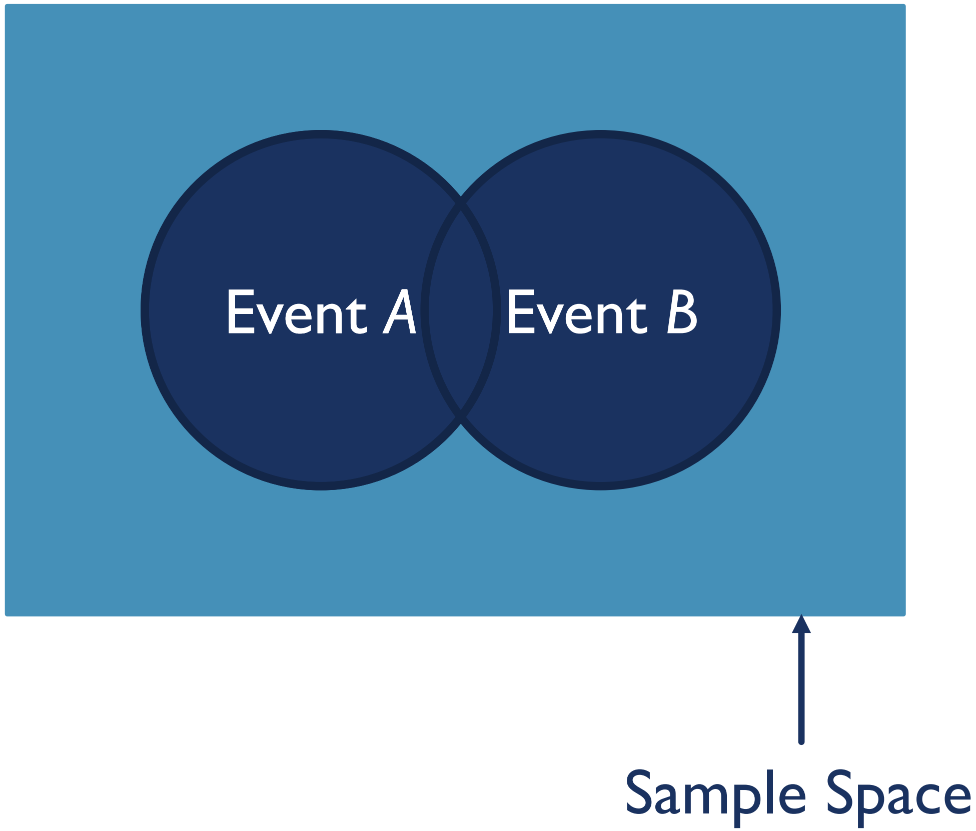
COMPLEMENT OF AN EVENT

- The **complement** of an event A is defined to be the event consisting of all sample points that are not in A.
- The complement of A is denoted by either A^c or $\bar{A}$.



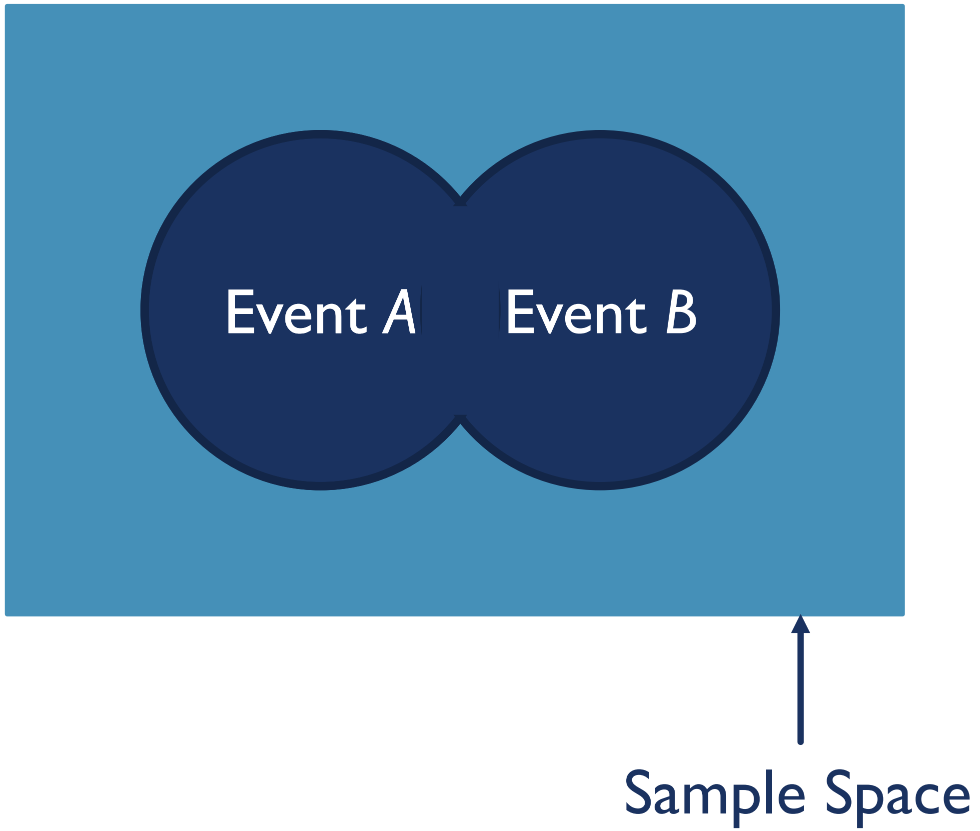
UNION OF TWO EVENTS

- The **union** of an event A and an event B is the event containing all sample points that are in A or B or both.
- The union of A and B is denoted by $A \cup B$.



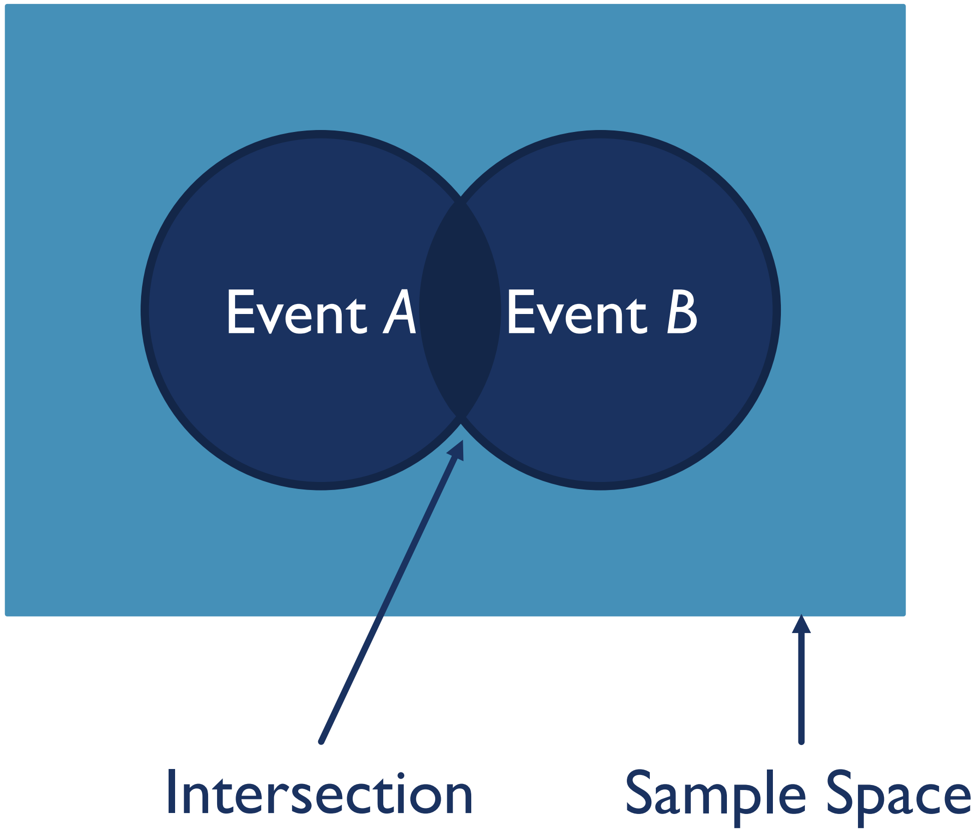
UNION OF TWO EVENTS

- The **union** of an event A and an event B is the event containing all sample points that are in A or B or both.
- The union of A and B is denoted by $A \cup B$.



INTERSECTION OF TWO EVENTS

- The **intersection** of an event *A* and an event *B* is the event containing all sample points that are in **both** *A* and *B*.
- The intersection of *A* and *B* is denoted by $A \cap B$.



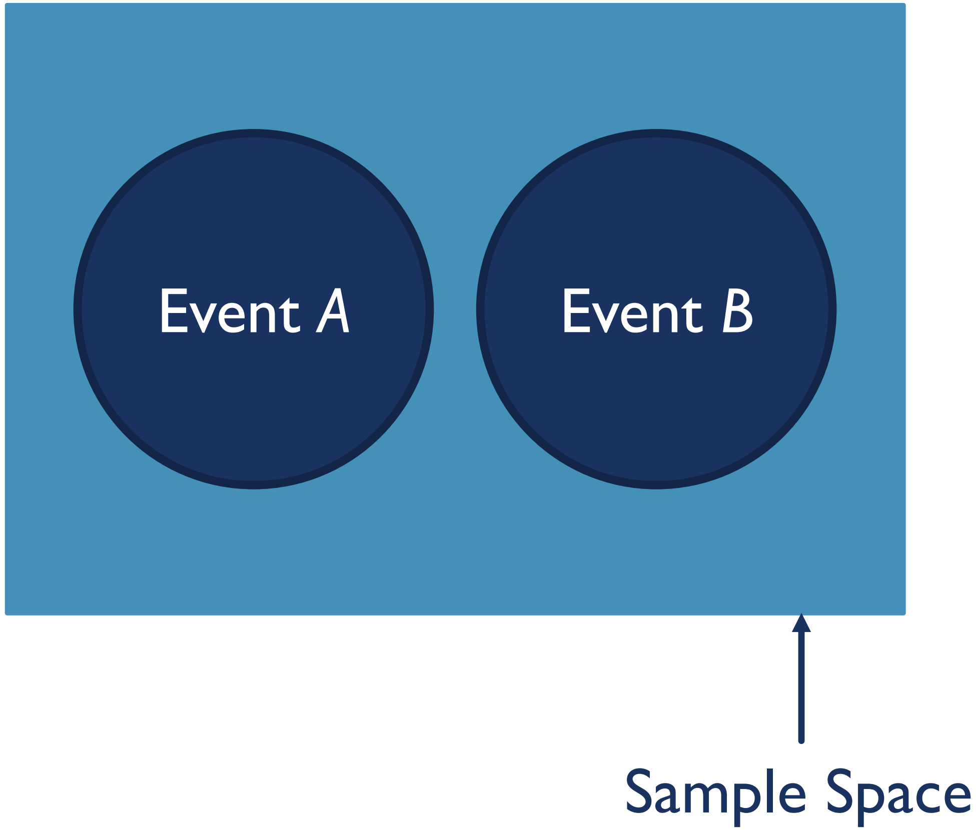
ADDITION LAW

- The **addition law** provides a way to compute the union of events *A* and *B*:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

MUTUALLY EXCLUSIVE EVENTS

- Two events are **mutually exclusive** if the events have no sample points in common – do not intersect.
- This also means that the events cannot both occur. If one event occurs, the other cannot.



ADDITION LAW – **MUTUALLY EXCLUSIVE EVENTS**

- The **addition law** provides a way to compute the union of events *A* and *B*:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

If two events are mutually exclusive they do not intersect.

$$P(A \cup B) = P(A) + P(B)$$

EXAMPLE

Weather	Spring	Summer	Fall	Winter	Total
Clear of Cloudy	626,986	799,443	519,487	312,036	**2,257,952**
Misty	288,096	250,679	302,510	155,573	**996,858**
Rain or Snow	3,507	11,007	19,616	3,739	**37,869**
Total	**918,589**	**1,061,129**	**841,613**	**471,348**	**3,292,679**

EXAMPLE

- What is the probability a random customer uses the bike service in Fall **and** it was raining or snowing?

EXAMPLE

Weather	Spring	Summer	Fall	Winter	Total
Clear of Cloudy	626,986	799,443	519,487	312,036	**2,257,952**
Misty	288,096	250,679	302,510	155,573	**996,858**
Rain or Snow	3,507	11,007	19,616	3,739	**37,869**
Total	**918,589**	**1,061,129**	**841,613**	**471,348**	**3,292,679**

EXAMPLE

- What is the probability a random customer uses the bike service in Fall **and** it was raining or snowing?

$$\frac{19{,}616}{3{,}292{,}679} = 0.006 = 0.6\%$$

EXAMPLE

- What is the probability a random customer uses the bike service in Fall **or** it was raining or snowing?

EXAMPLE

Weather	Spring	Summer	Fall	Winter	Total
Clear of Cloudy	626,986	799,443	519,487	312,036	**2,257,952**
Misty	288,096	250,679	302,510	155,573	**996,858**
Rain or Snow	3,507	11,007	19,616	3,739	**37,869**
Total	**918,589**	**1,061,129**	**841,613**	**471,348**	**3,292,679**

EXAMPLE

- What is the probability a random customer uses the bike service in Fall **or** it was raining or snowing?

$$\frac{37{,}869}{3{,}292{,}679} + \frac{841{,}613}{3{,}292{,}679} - \frac{19{,}616}{3{,}292{,}679} = \frac{859{,}866}{3{,}292{,}679} = 0.261 = 26.1\%$$

SUMMARY

- The complement of an event *A*, (A^c), is defined to be the event consisting of all sample points that are not in *A*.
- The union of an event *A* and an event *B*, $(A \cup B)$, is the event containing all sample points that are in *A* or *B* or both.
  - Computed with Addition Law.
- The intersection of an event *A* and an event *B*, $(A \cap B)$, is the event containing all sample points that are in both *A* and *B*.
- Two events are mutually exclusive if the events have no sample points in common – do not intersect.

CONDITIONAL PROBABILITIES

RANDOMNESS IN DATA

CONDITIONAL PROBABILITIES

- The probability of an event given that another event has occurred is called a **conditional** (or **joint**) **probability**.
- The conditional probability of *A* given that *B* has already occurred is denoted by $P(A|B)$.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

MULTIPLICATION LAW

- The **multiplication law** provides a way to compute the probability of the intersection of two events as long as you know the conditional probabilities:

$$P(A \cap B) = P(A|B) \times P(B)$$

OR

$$P(A \cap B) = P(B|A) \times P(A)$$

INDEPENDENT EVENTS

- If the probability of an event A is not changed by the existence of event B, then the two events are called **independent**.

$$P(A|B) = P(A) \quad \text{OR} \quad P(B|A) = P(B)$$

MULTIPLICATION LAW – INDEPENDENT EVENTS

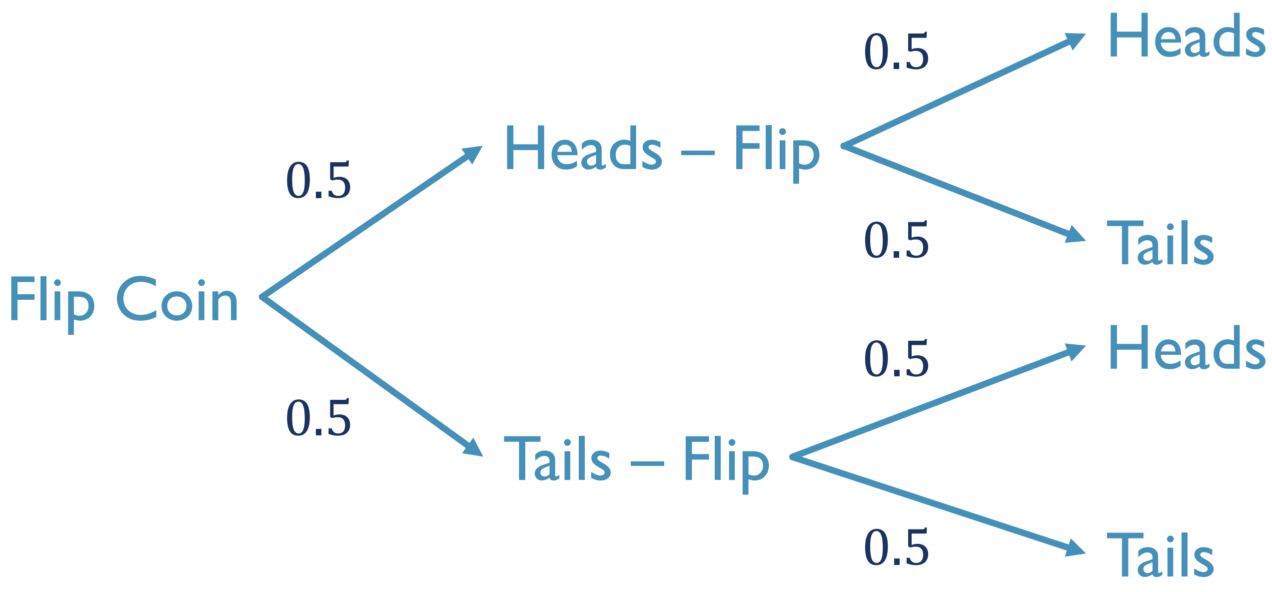
- The **multiplication law** provides a way to compute the probability of the intersection of two events as long as you know the conditional probabilities:

$$P(A \cap B) = P(A) \times P(B)$$

ONLY IF EVENTS ARE INDEPENDENT

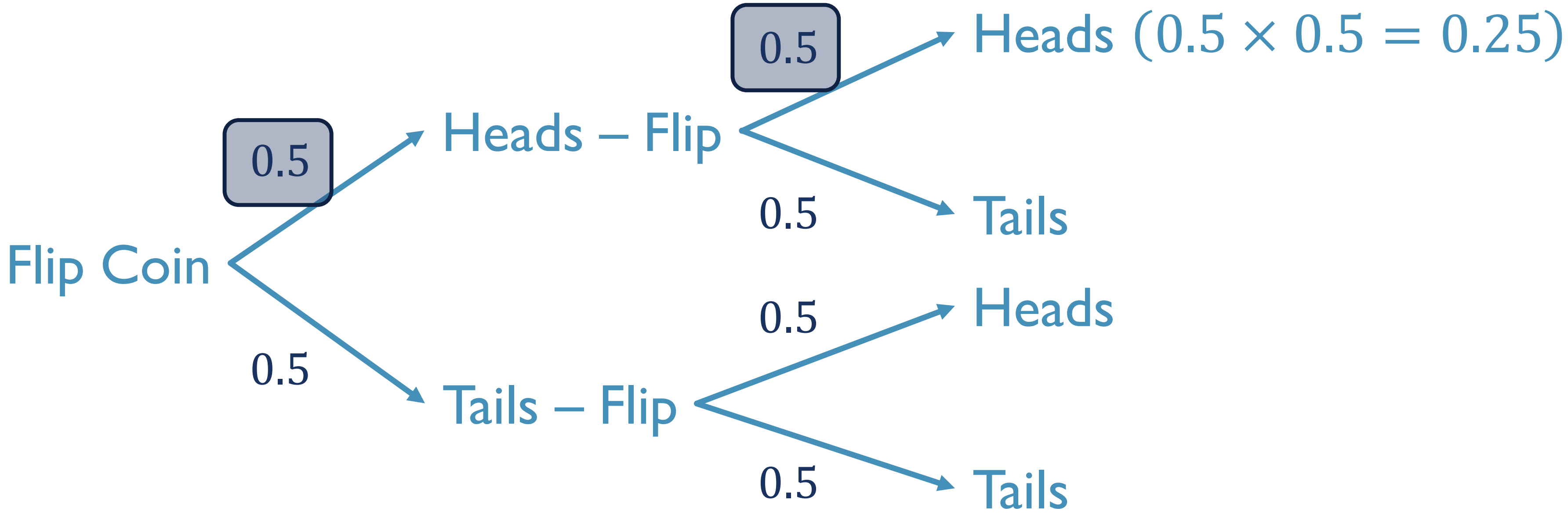
TREE DIAGRAMS

- Tree diagrams can help calculate probabilities in a series of independent events.
- For example, you have a 2-step random process where you flip a coin twice (independent flips):



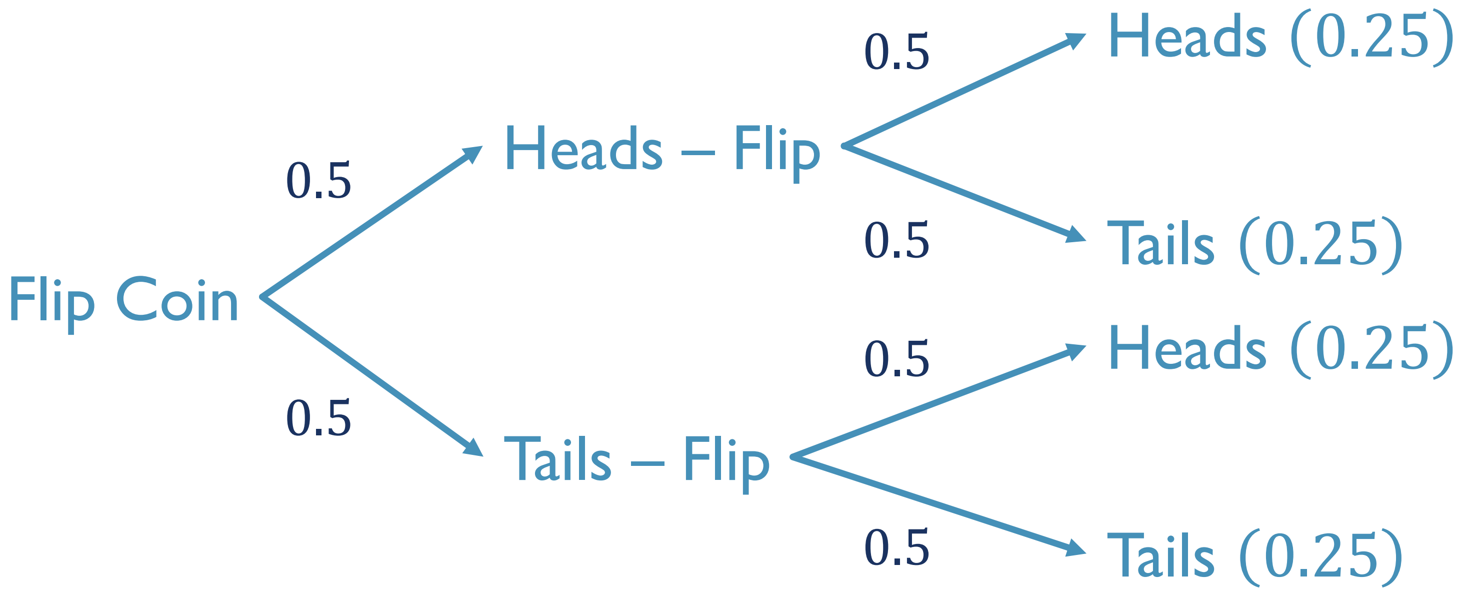
TREE DIAGRAMS

- Tree diagrams can help calculate probabilities in a series of independent events.
- For example, you have a 2-step random process where you flip a coin twice (independent flips):



TREE DIAGRAMS

- Tree diagrams can help calculate probabilities in a series of independent events.
- For example, you have a 2-step random process where you flip a coin twice (independent flips):



EXAMPLE

Weather	Spring	Summer	Fall	Winter	Total
Clear of Cloudy	626,986	799,443	519,487	312,036	**2,257,952**
Misty	288,096	250,679	302,510	155,573	**996,858**
Rain or Snow	3,507	11,007	19,616	3,739	**37,869**
Total	**918,589**	**1,061,129**	**841,613**	**471,348**	**3,292,679**

EXAMPLE

- What is the probability a random customer uses the bike service in on a misty day **given** it is winter?

EXAMPLE

Weather	Spring	Summer	Fall	Winter	Total
Clear of Cloudy	626,986	799,443	519,487	312,036	**2,257,952**
Misty	288,096	250,679	302,510	155,573	**996,858**
Rain or Snow	3,507	11,007	19,616	3,739	**37,869**
Total	**918,589**	**1,061,129**	**841,613**	**471,348**	**3,292,679**

EXAMPLE

- What is the probability a random customer uses the bike service in on a misty day **given** it is winter?

$$\frac{155{,}573}{471{,}348} = 0.33 = 33\%$$

EXAMPLE

- What is the probability a random customer uses the bike service in on a misty day **given** it is winter?

$$\frac{155{,}573}{471{,}348} = 0.33 = 33\%$$

OR

$$\frac{P(A \cap B)}{P(B)} = \frac{\left(\frac{155{,}573}{3{,}292{,}679}\right)}{\left(\frac{471{,}348}{3{,}292{,}679}\right)} = \left(\frac{155{,}573}{3{,}292{,}679}\right) \times \left(\frac{3{,}292{,}679}{471{,}348}\right) = 0.33$$

SUMMARY

- The probability of an event given that another event has occurred is called a conditional (or joint) probability.
- The multiplication law provides a way to compute the probability of the intersection of two events as long as you know the conditional probabilities.
- If the probability of an event *A* is not changed by the existence of event *B*, then the two events are called independent.