CS 581

Advanced Artificial Intelligence

February 14, 2024

Announcements / Reminders

- **Please follow the Week 06 To Do List instructions (if you haven't already)**
- **Written Assignment #02: due on Monday 02/19 at 11:59 PM CST**
- **Programming Assignment #01: due on Sunday 03/03 at 11:59 PM CST**

- **Midterm Exam: 02/21/2024**
  - **Section 02 – Make arrangements with Mr. Charles Scott**
  - **WE WILL HAVE OUR EXAM IN A DIFFERENT ROOM (WH113)**

Plan for Today

- **Probability refresher**

Random Experiment

- The agent needs reason in an uncertain world
- Uncertainty can be due to
  - Noisy sensors (e.g., temperature, GPS, camera, etc.)
  - Imperfect data (e.g., low resolution image)
  - Missing data (e.g., lab tests)
  - Imperfect knowledge (e.g., medical diagnosis)
  - Exceptions (e.g., all birds fly except ostriches, penguins, birds with injured wings, dead birds, …)
  - Changing data (e.g., flu seasons, traffic conditions, etc.)
  - …
- The agent still must act (e.g., step on the breaks, diagnose a patient, order a lab test, …)

Probability In AI: Selected Application

- Classification
  - Naïve Bayes, logistic regression, neural networks
  - Maximum likelihood estimation, Bayesian estimation, gradient optimization, backpropagation
- Decision making
  - Episodic decision making, Markov decision processes, multi-armed bandits
  - Value of information, Bellman equations, value iteration, policy iteration, UCB1, ε-greedy
- Reinforcement learning
  - Prediction, control, Monte-Carlo methods, temporal difference learning, Sarsa, Q-learning

Random Experiment

Random Experiment is a process by which we observe something uncertain.

An outcome is a result of a random experiment.

The set of all possible outcomes is called the sample space S (frequently labeled Ω).

Outcomes / Sample Space / Event

Outcome: A **result of a random experiment**

Sample space $\mathbb{S}$: The **set of all possible outcomes**

Event: A **subset** of the **sample space $\mathbb{S}$**

Events: Union and Intersection

If A and B are events, then

$$A \cap B$$

and

$$A \cup B$$

are also events

$\cup$ **- union ("or")**

$\cap$ **- intersection ("and")**

Events: Union and Intersection

If A and B are events, then

$$A \cap B$$

and

$$A \cup B$$

are also events

Simple event: An event that cannot be decomposed

Events: Union and Intersection

We observe that event

$$A \cap B$$

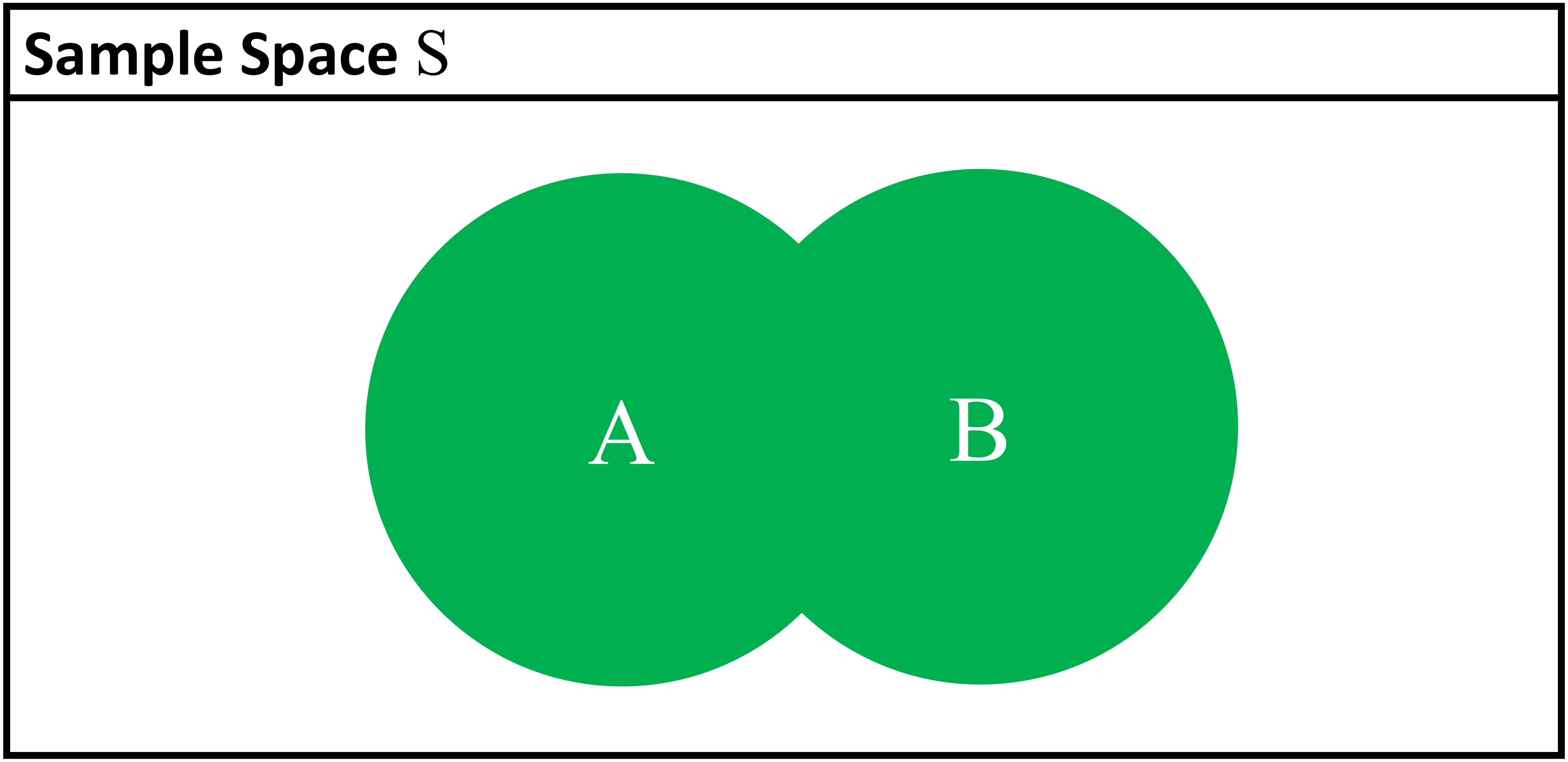
occurs if event A and event B occur

We observe that event

$$A \cup B$$

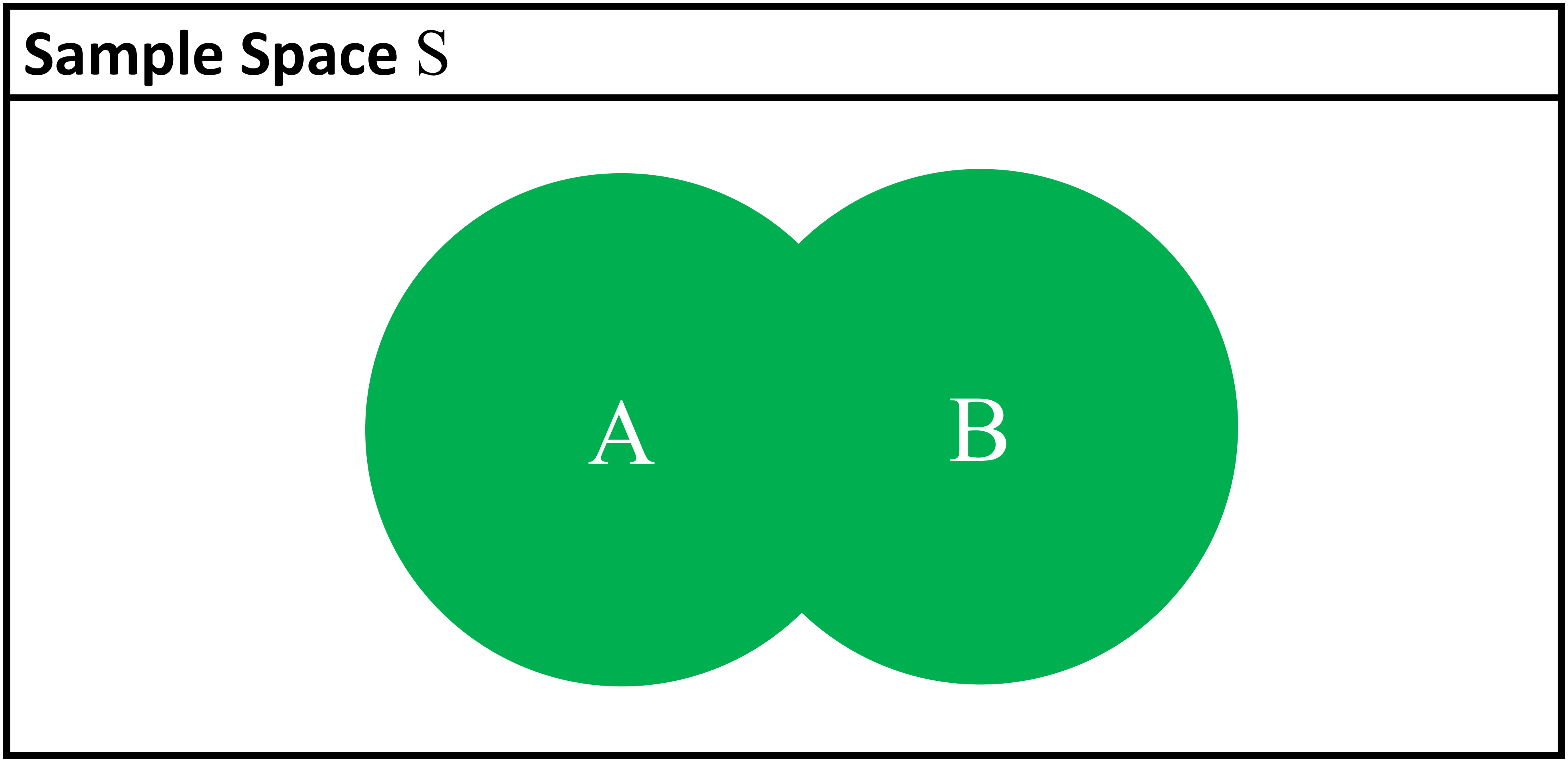
occurs if either event A or event B occur

Event (Set) Union: Venn Diagram



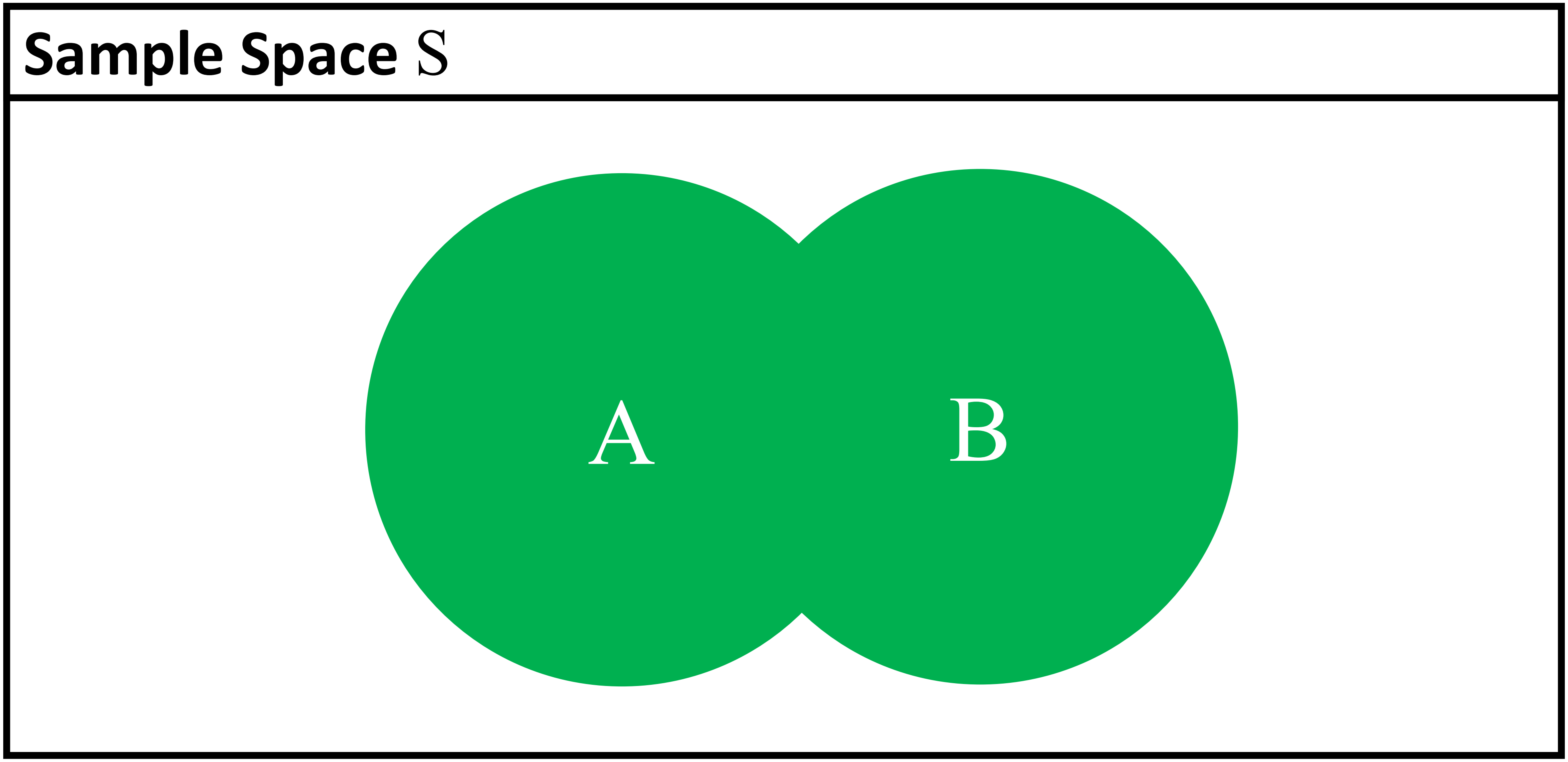
$A \cup B$

Event (Set) Union: Venn Diagram



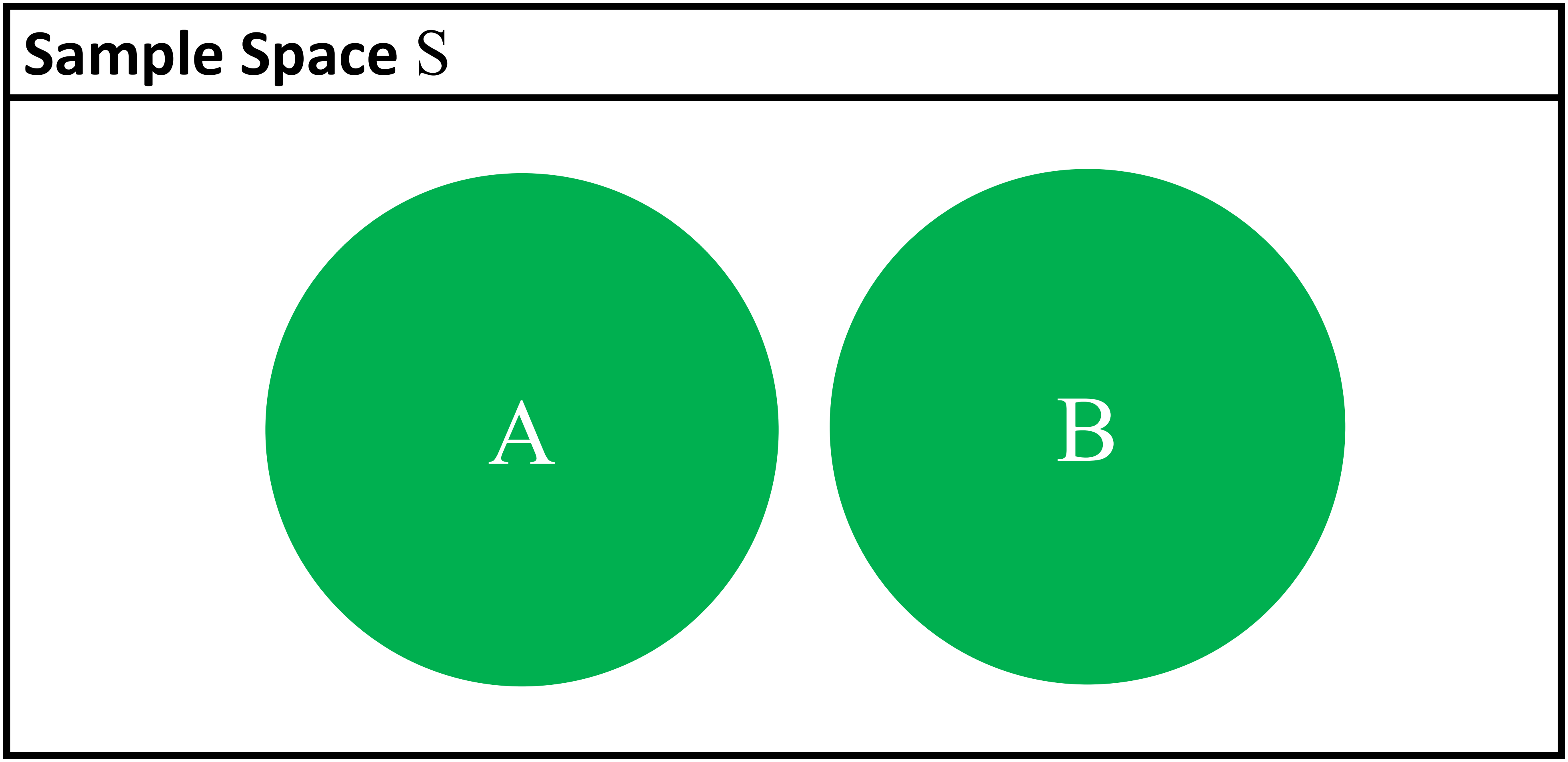
$$A \cup B \neq \varnothing$$

Event (Set) Union: Probability



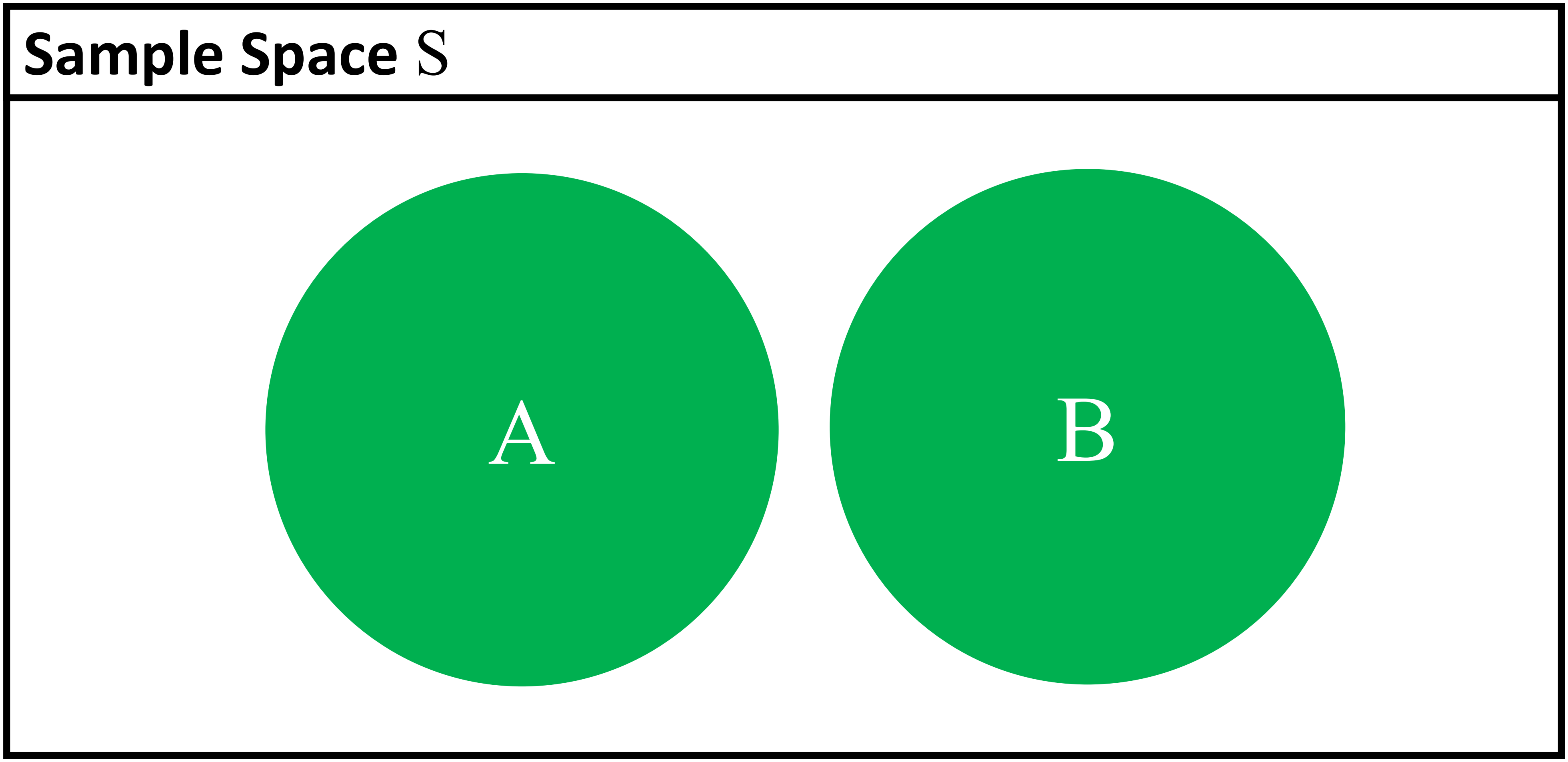
Events A and B are overlapping

Event (Set) Union: Venn Diagram



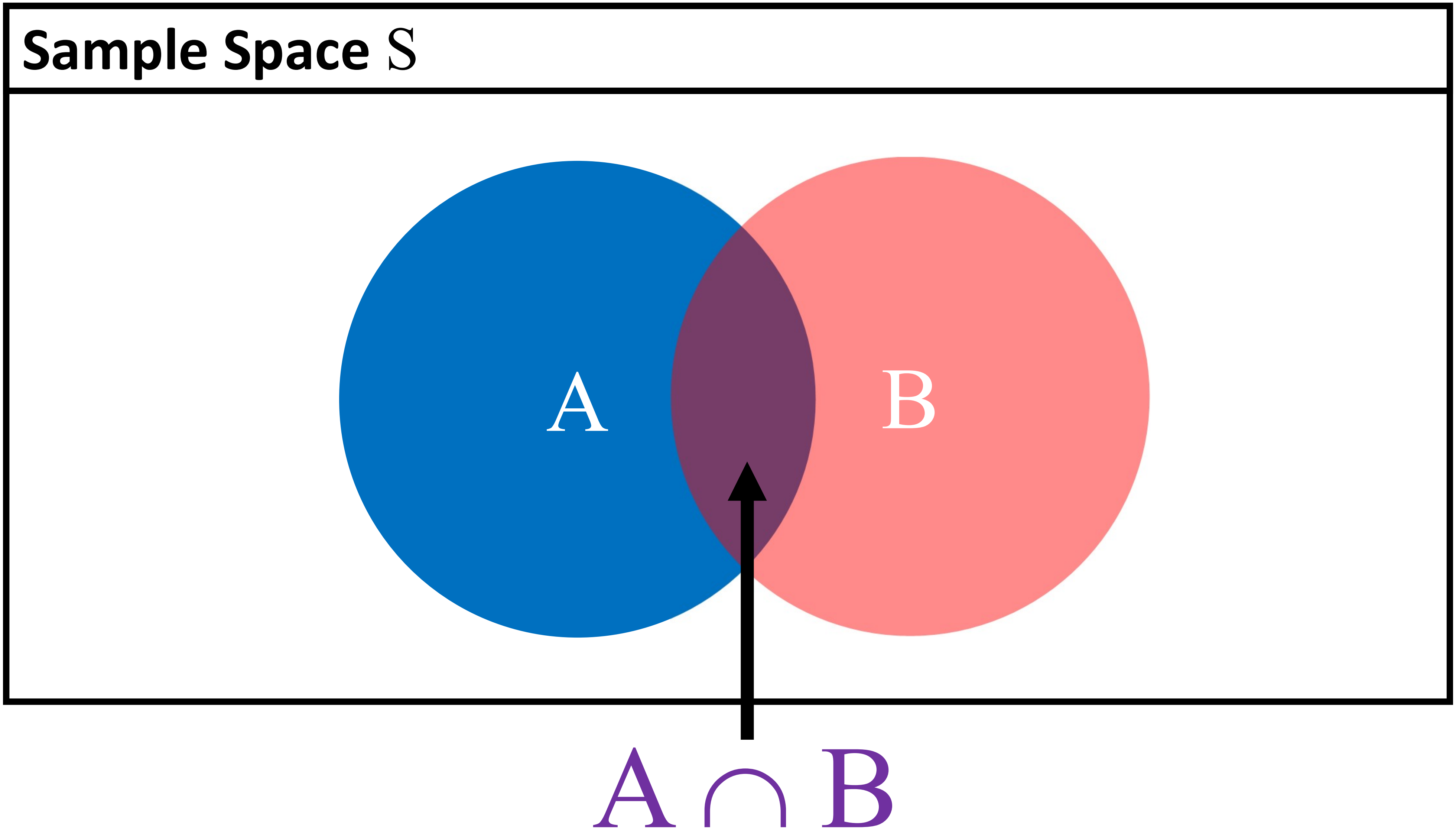
$$A \cup B \neq \varnothing$$

Event (Set) Union: Probability

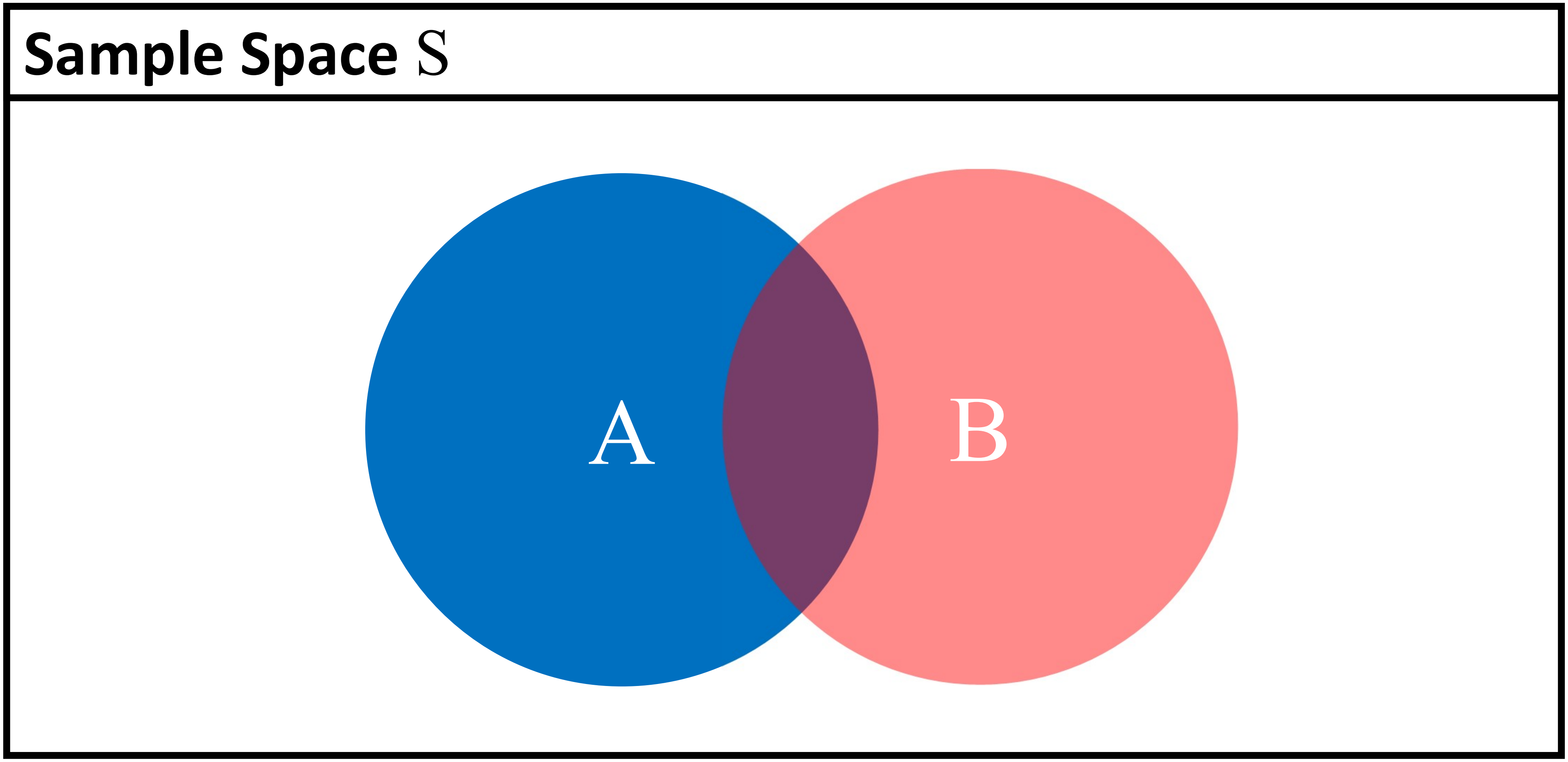


Events A and B are disjoint / mutually exclusive

Event (Set) Intersection: Venn Diagram

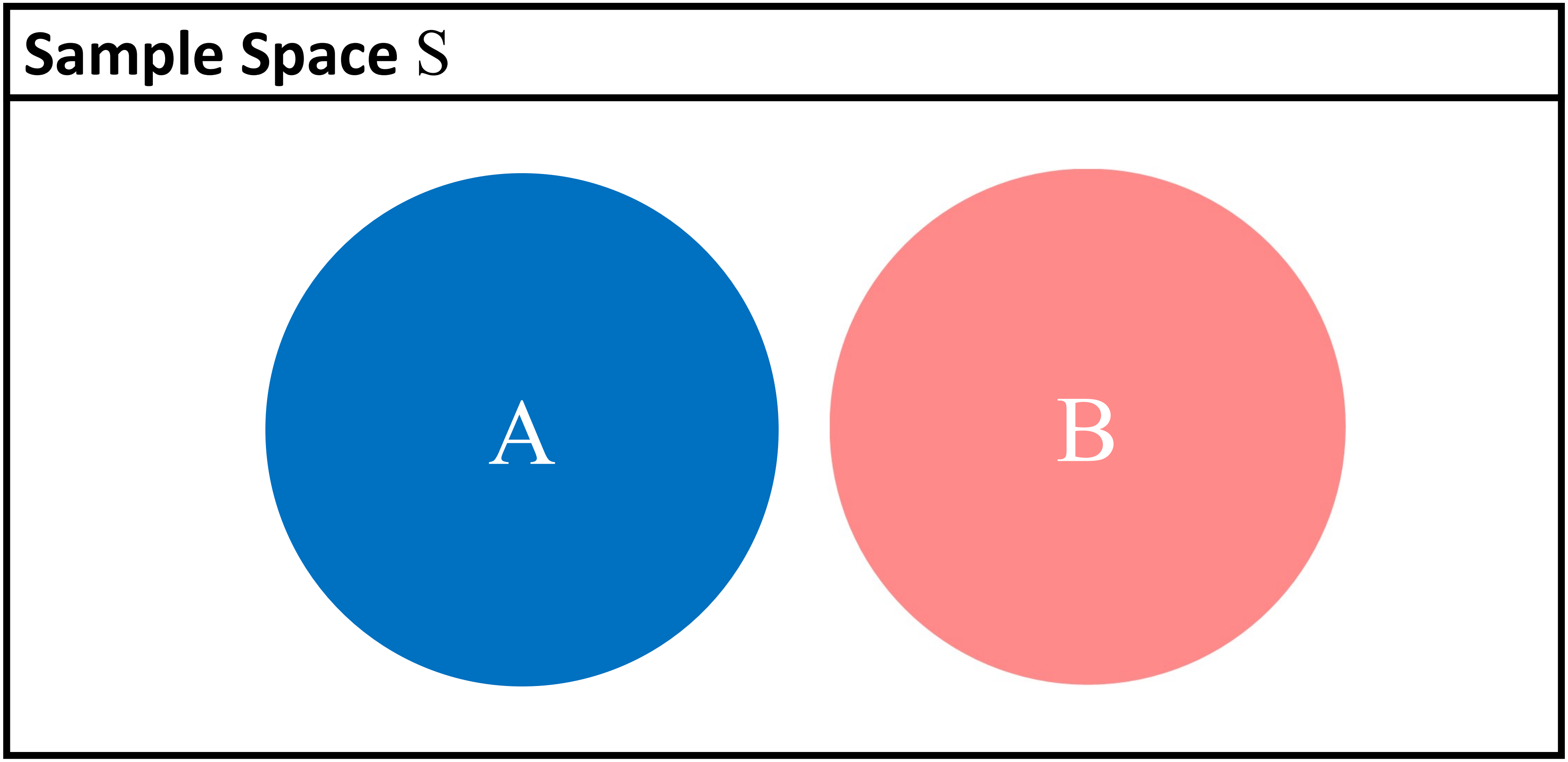


Event (Set) Intersection: Venn Diagram



$$A \cap B \neq \emptyset$$

Event (Set) Intersection: Venn Diagram

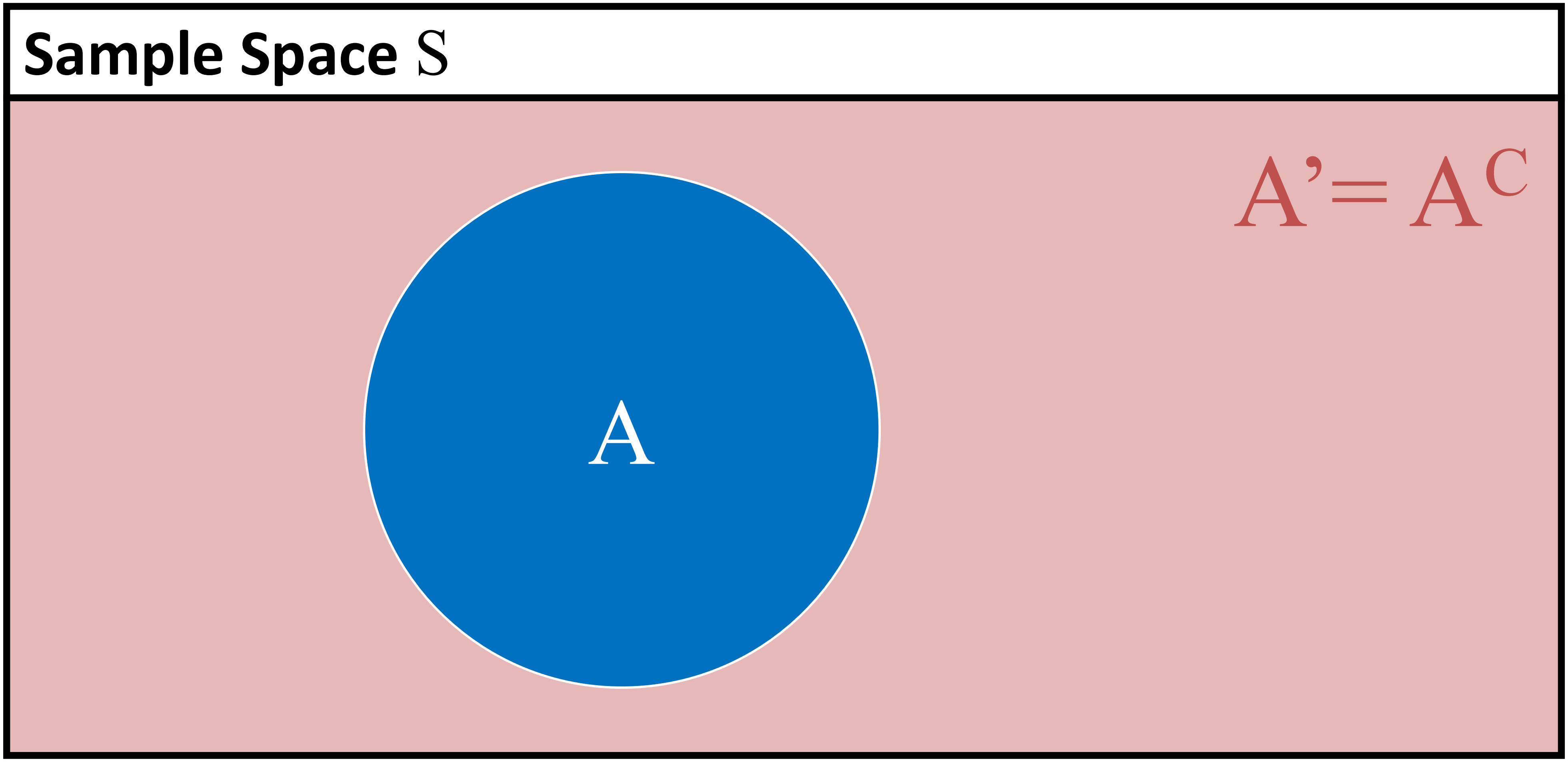


$$A \cap B = \varnothing$$

Events: Complementary Event

The complement of any event A is the event A' ("not A"), i.e. the event that A does not occur.

Complement: Venn Diagram

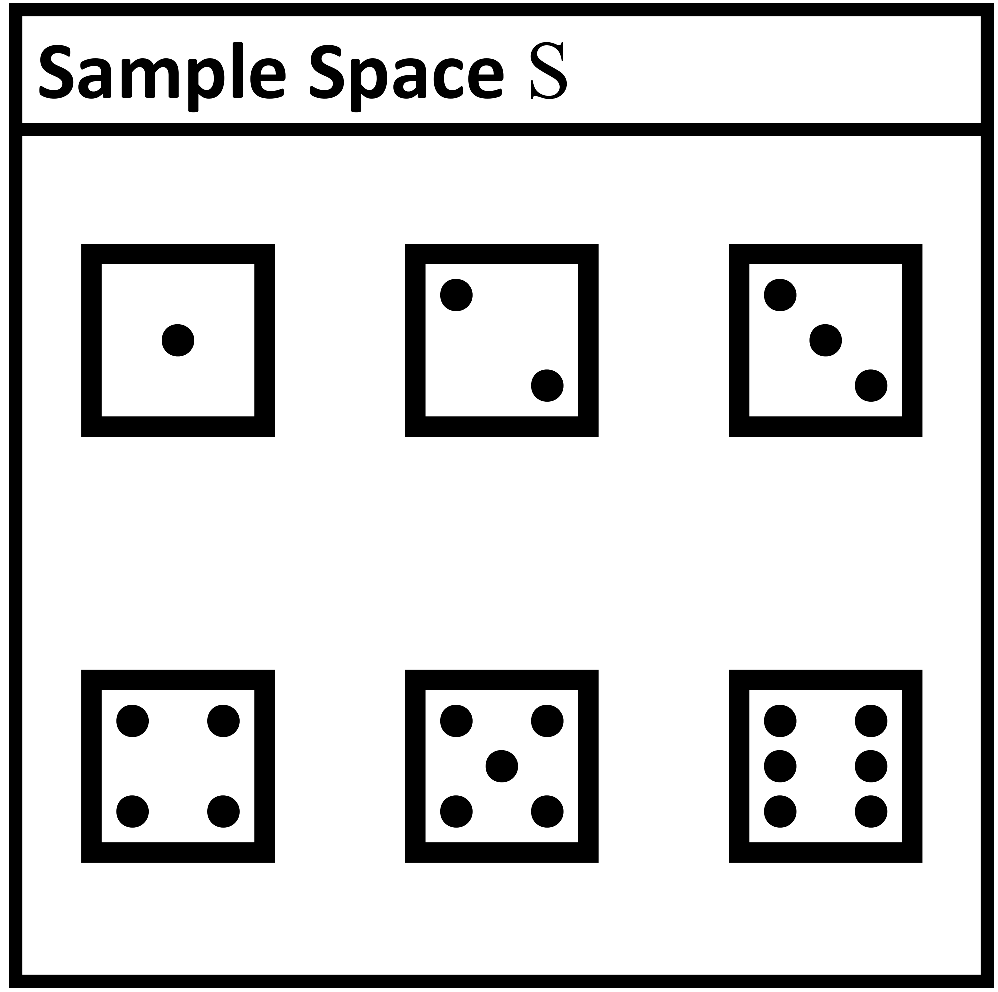


$$A \cup A' = A \cup A^C = S$$

Single Die Roll: Sample Space S

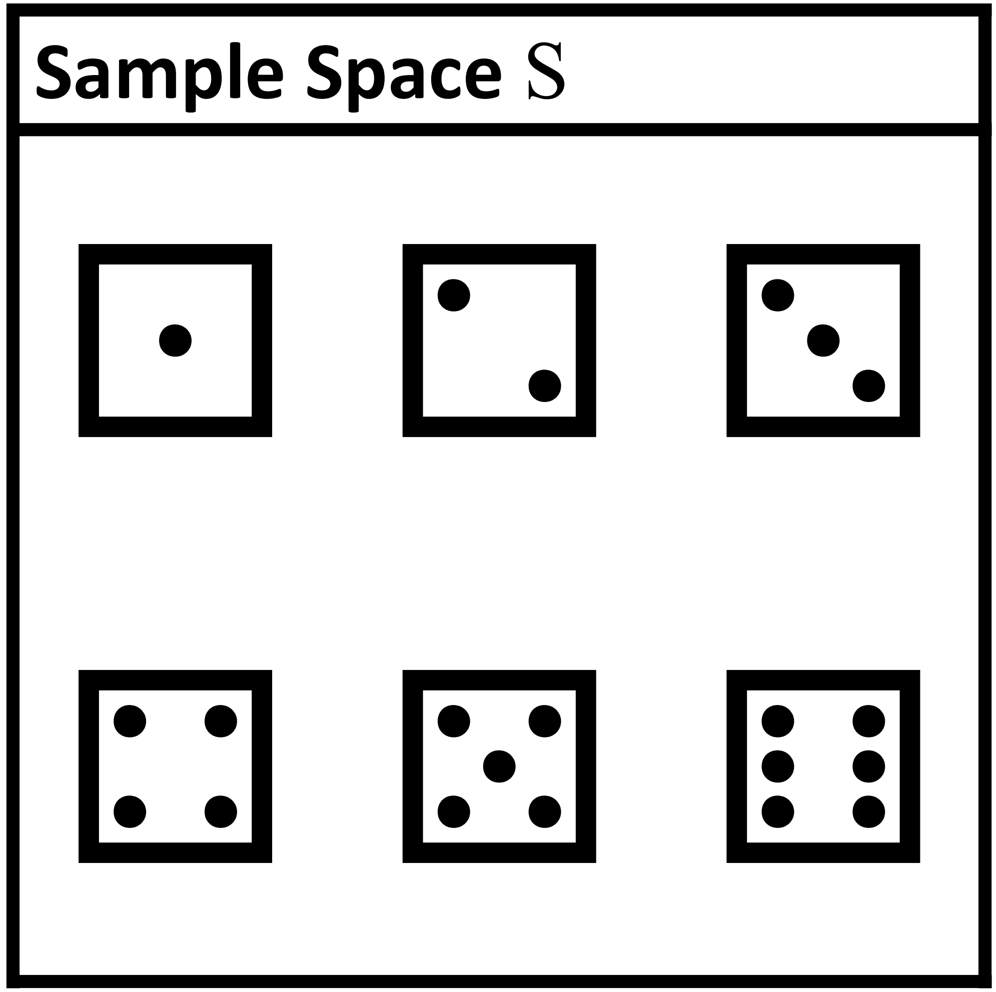
The set of all simple events of an experiment is called the sample space S.

S = {⚀, ⚁, ⚂, ⚃, ⚄, ⚅}



Single Die Roll: Sample Space S

The set of all **simple events** of an experiment is called the **sample space** S.



$$S = \{⚀, ⚁, ⚂, ⚃, ⚄, ⚅\}$$

Sample space S size (set cardinality): $|S| = 6$

Coin Toss: Sample Space S

The set of all simple events of an experiment is called the sample space S.

$\mathrm{S} = \{$  ,  $\}$

Sample Space S

Coin Toss: Sample Space S

The set of all **simple events** of an experiment is called the **sample space** S.

$\mathrm{S} = \{$  ,  $\}$

Sample space S size (set cardinality): $|\mathrm{S}| = 2$

Sample Space S

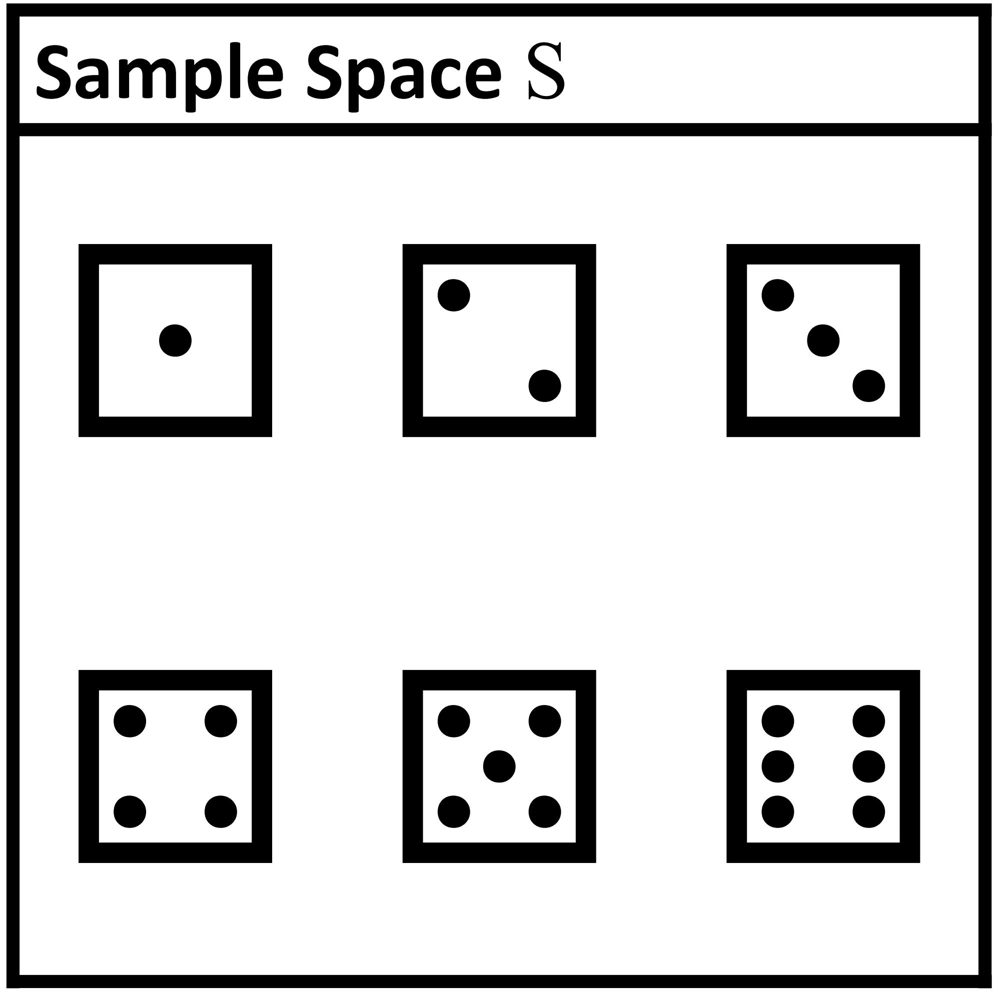
Example: Single Fair Die Roll

Random experiment:
rolling a single fair die

Outcome: result of rolling a single fair die

Sample space S:

S = {⚀, ⚁, ⚂, ⚃, ⚄, ⚅}



Example: Single Fair Coin Toss

Random experiment:

tossing a fair coin

Outcome: result of tossing a fair coin

Sample space S:

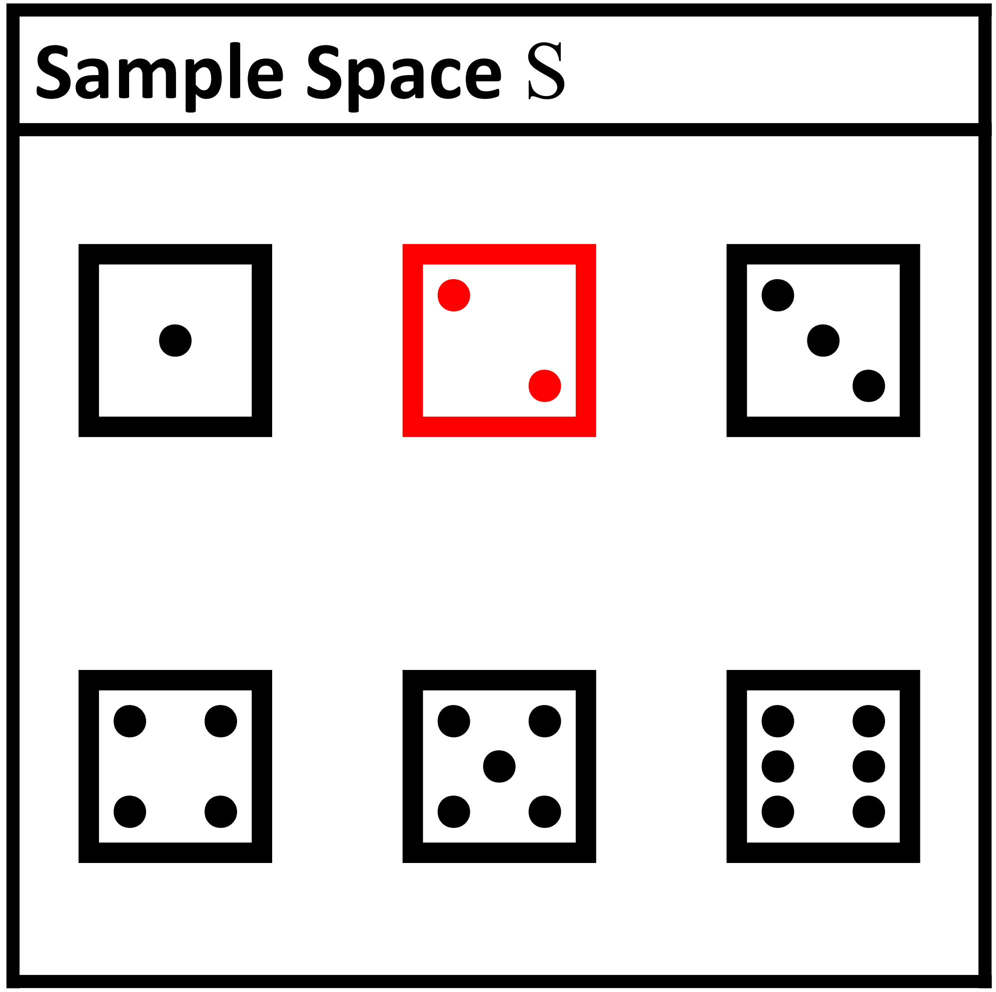
S = { ,  }

Sample Space S

Single Die Roll: Events

Event F: "two" rolled

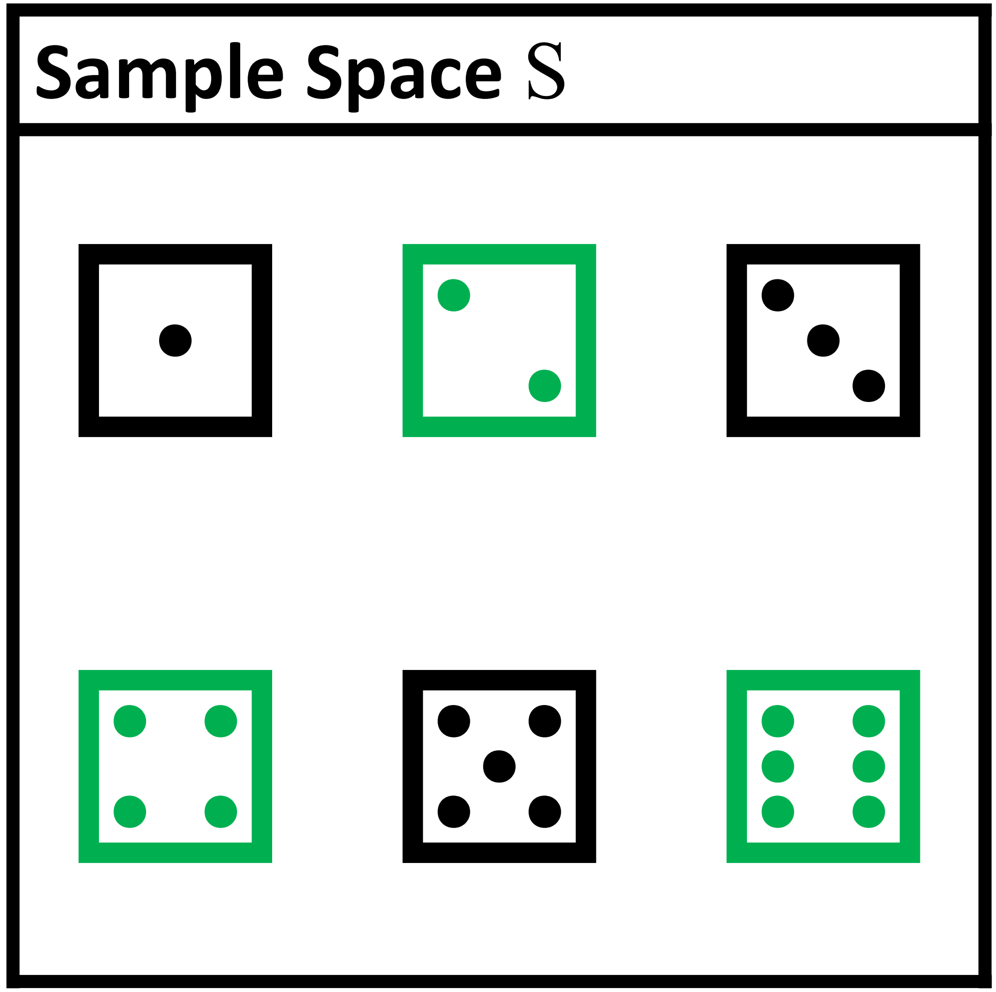
$\mathrm{F} = \{$  $\}$



Single Die Roll: Events

Event G: even number rolled

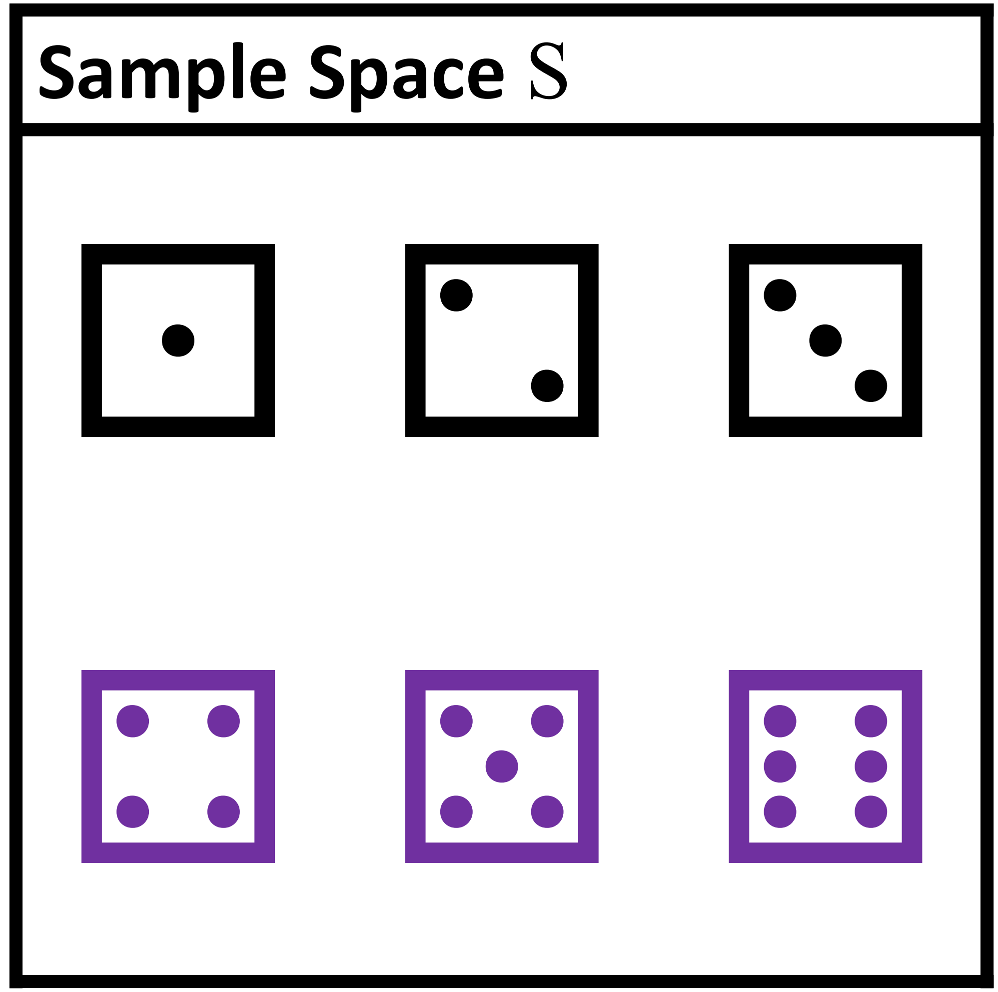
$G = \{$⚁, ⚃, ⚅$\}$



Single Die Roll: Events

Event H: number greater than 3 rolled

$$\mathrm{G} = \{⚃, ⚄, ⚅\}$$



Probability (Measure)

Probability (measure)

$$\mathrm{P(A)}$$

is a value assigned to an event A.

Probability $\mathrm{P(A)}$ is a value between 0 and 1 (inclusive) that shows how likely event A is.

Probability Theory

The main subject of probability theory is to develop tools and techniques to calculate probabilities of different events.

Axioms of Probability

Axiom 1:

For any event A**,** $P(A) \geq 0$

Axiom 2:

Probability of the sample space S **is** $P(S) = 1$

Axiom 3:

If A_1**,** A_2**, … are disjoint events, then**

$$P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \ldots$$

Single Die Roll: Probabilities

Consider events:

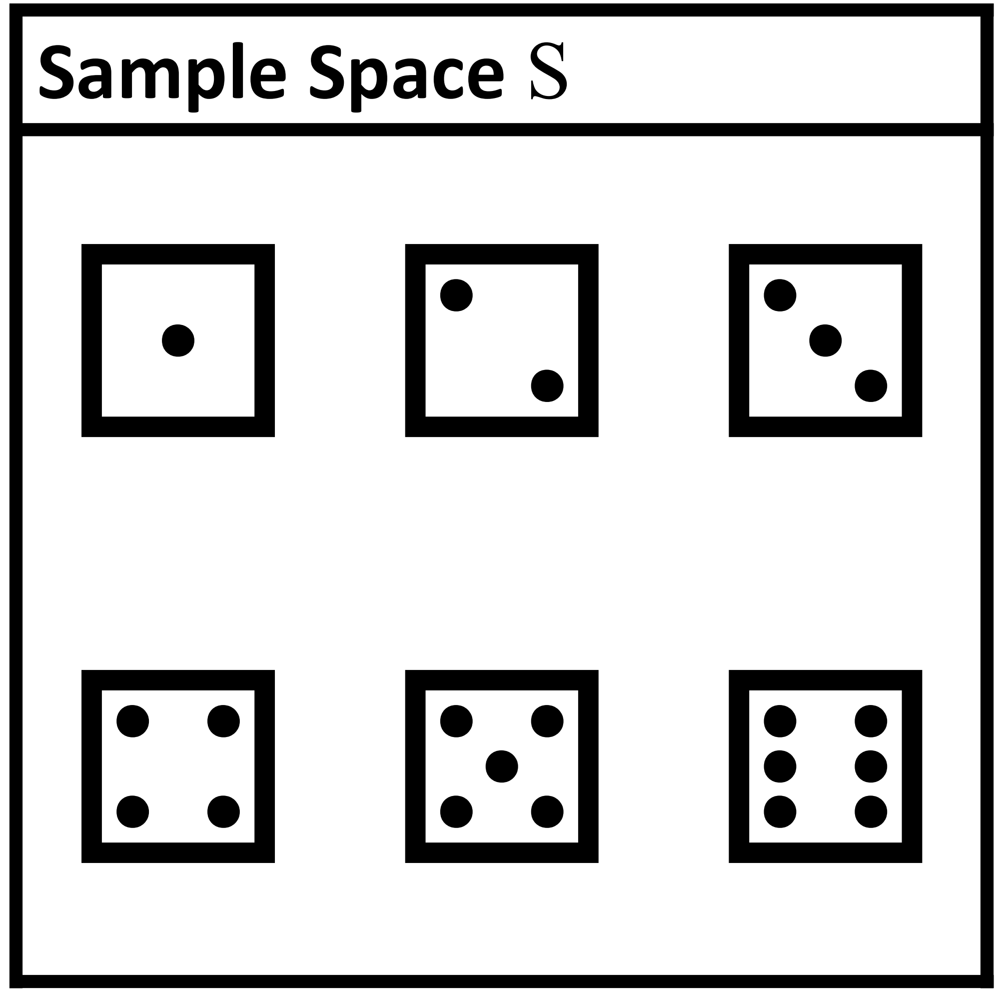
$A = \{$⚂$\}$

$B = \{$⚅$\}$

Probabilities:

$$P(A) = |A|/|S| = 1/6$$

$$P(B) = |B|/|S| = 1/6$$

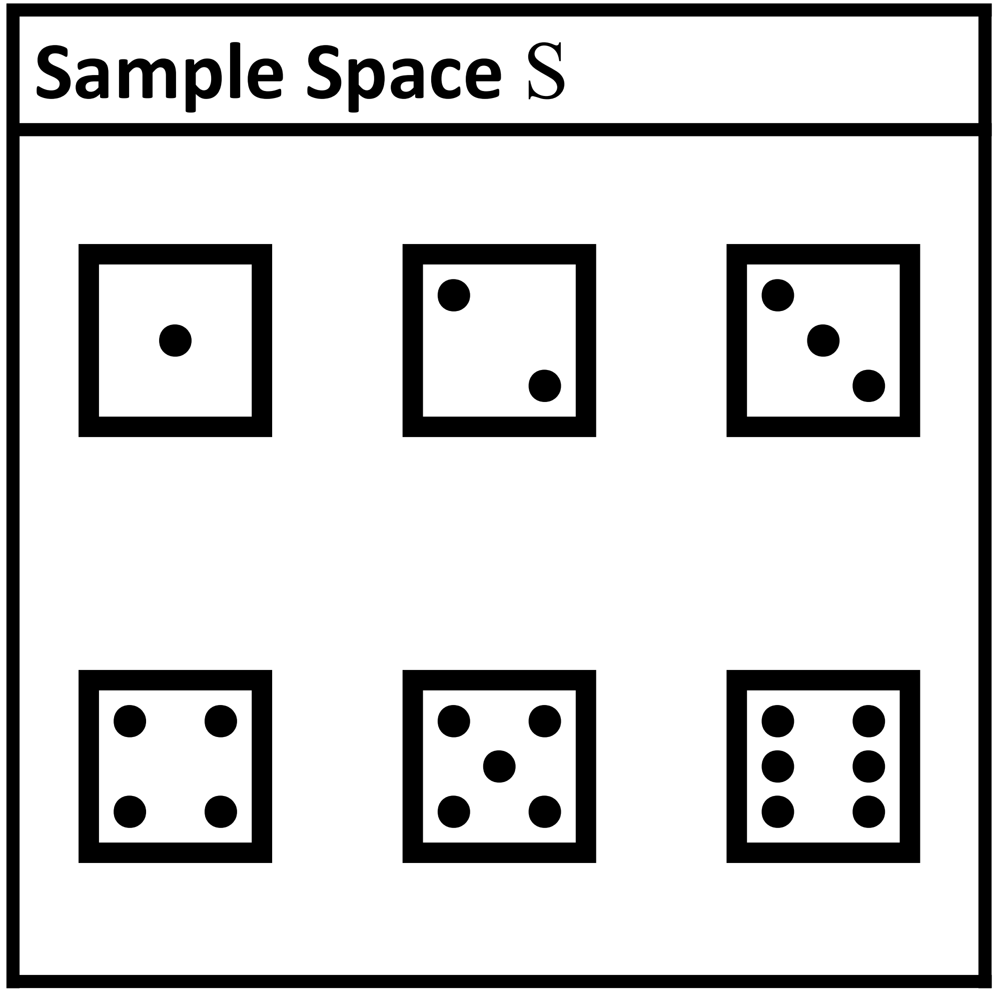


Single Die Roll: Probabilities

Consider events:

$$C = A \cup B = \{⚂\} \cup \{⚅\}$$

$$D = A \cap B = \{⚂\} \cap \{⚅\}$$



Probabilities:

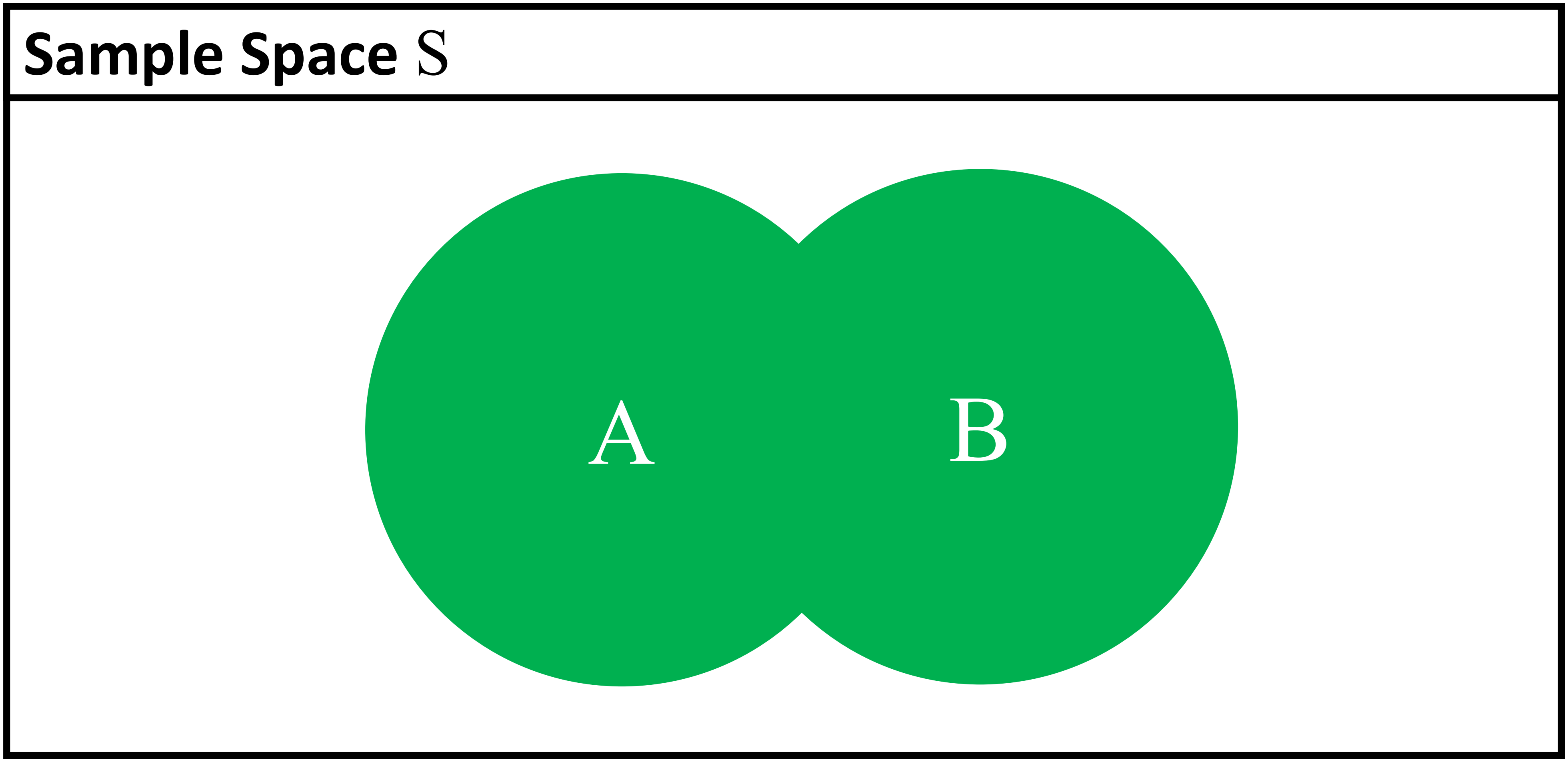
$$P(C) = P(A \cup B) =$$

$$= P(A) + P(B) = 2/6$$

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

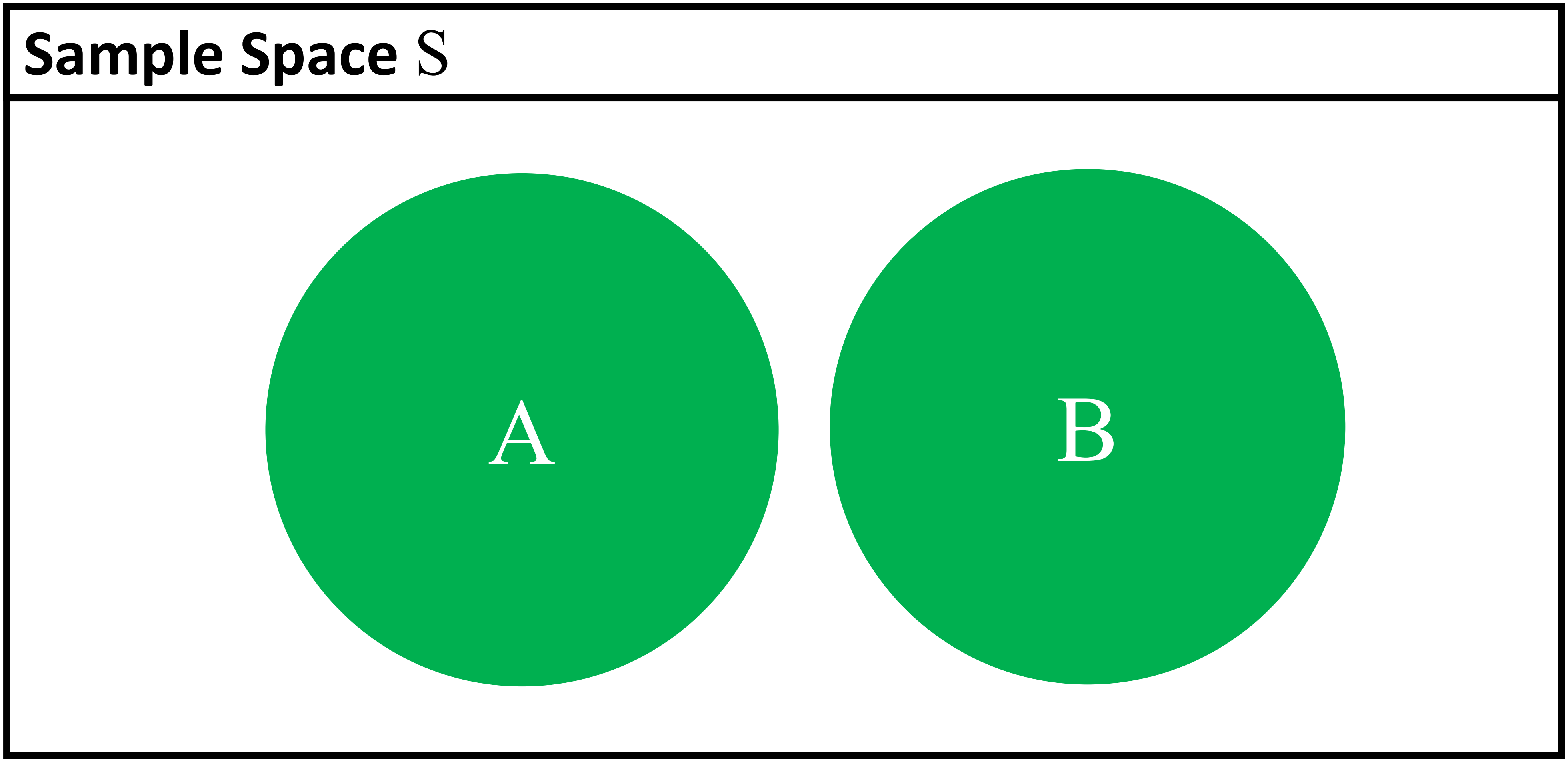
$$= |A \cap B|/|S| = 0/6$$

Event (Set) Union: Probability



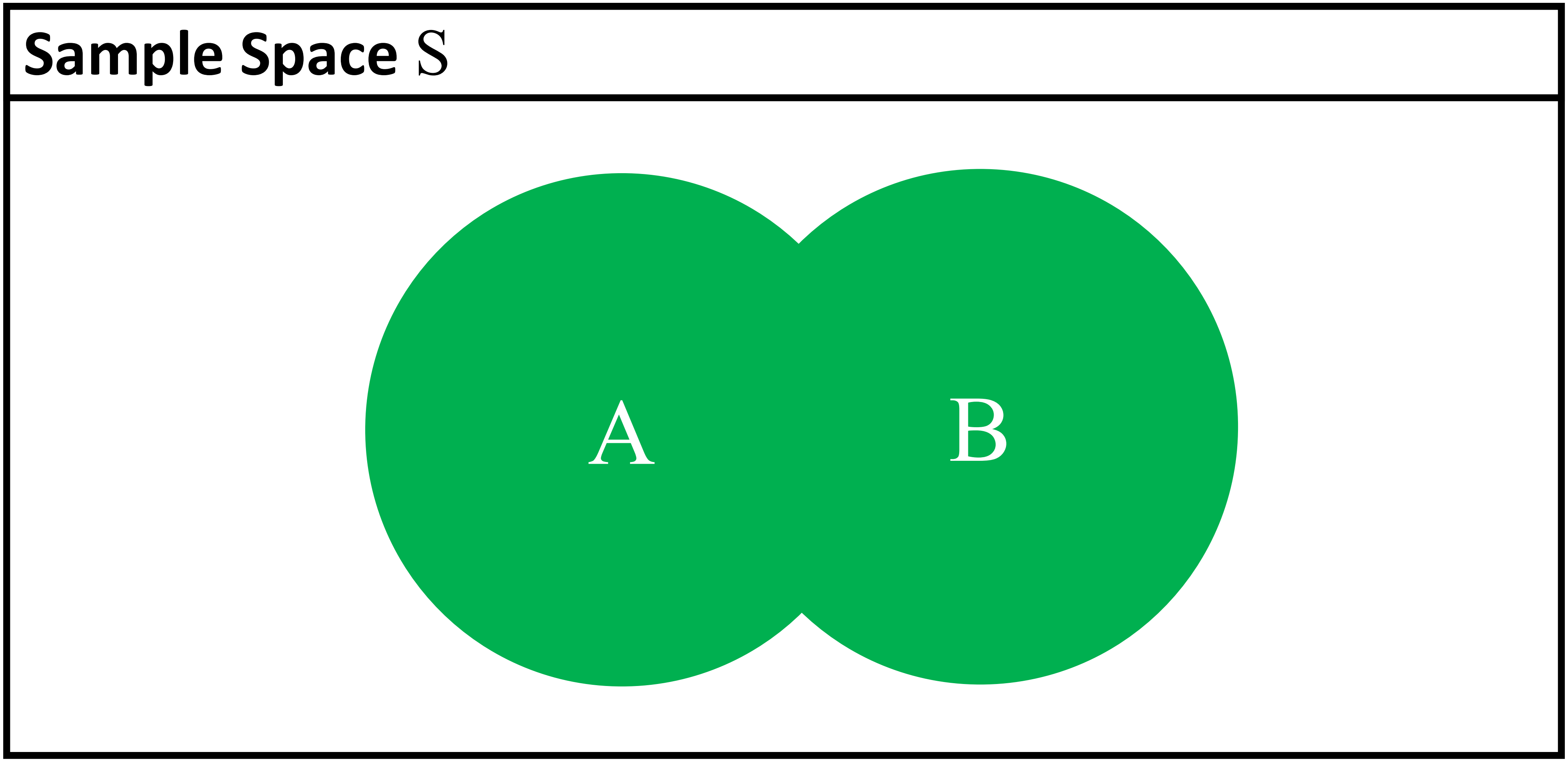
$$P(A \cup B) = P(A \text{ or } B)$$

Event (Set) Union: Probability



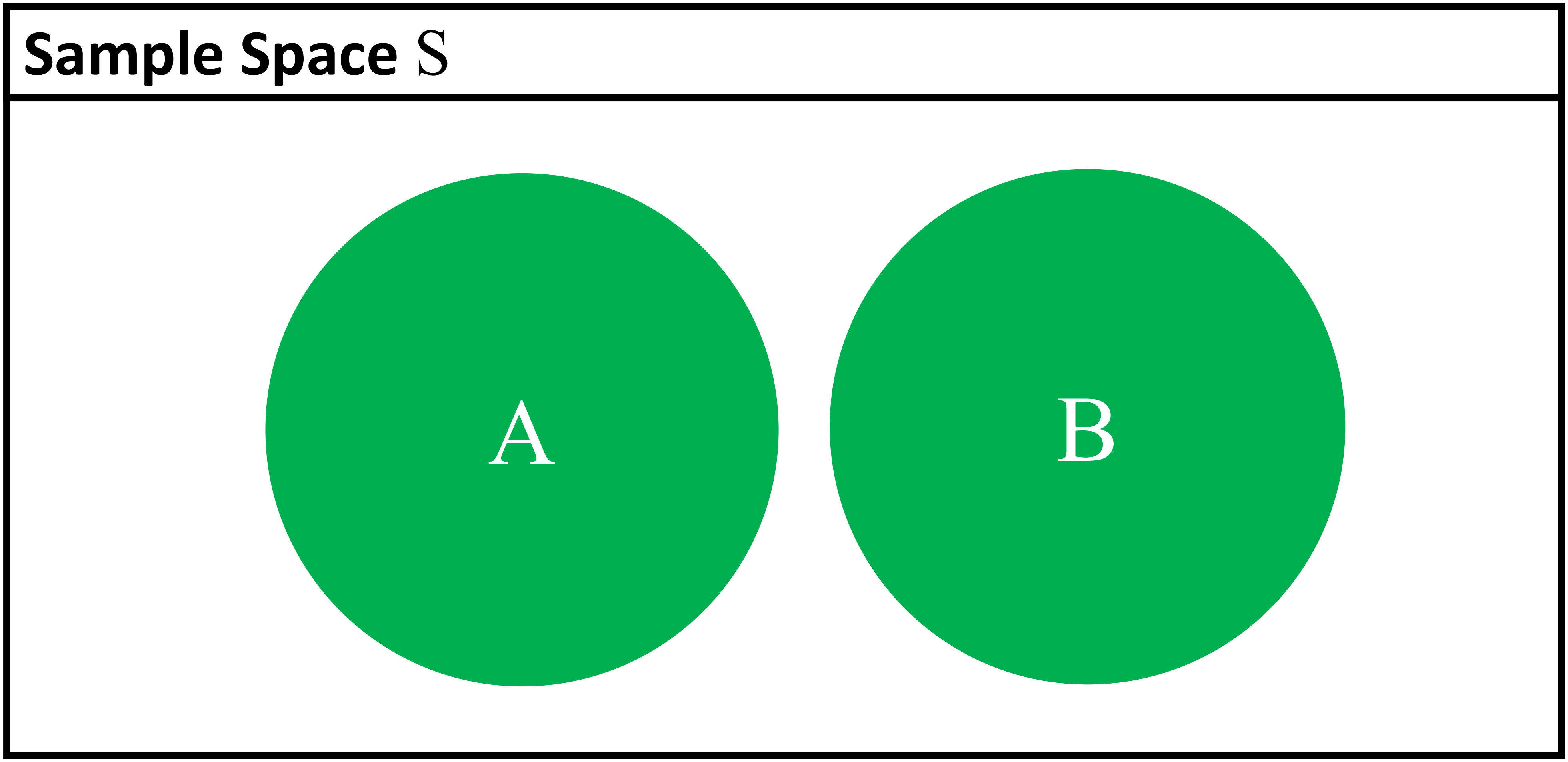
$$P(A \cup B) = P(A \text{ or } B)$$

Event (Set) Union: Probability



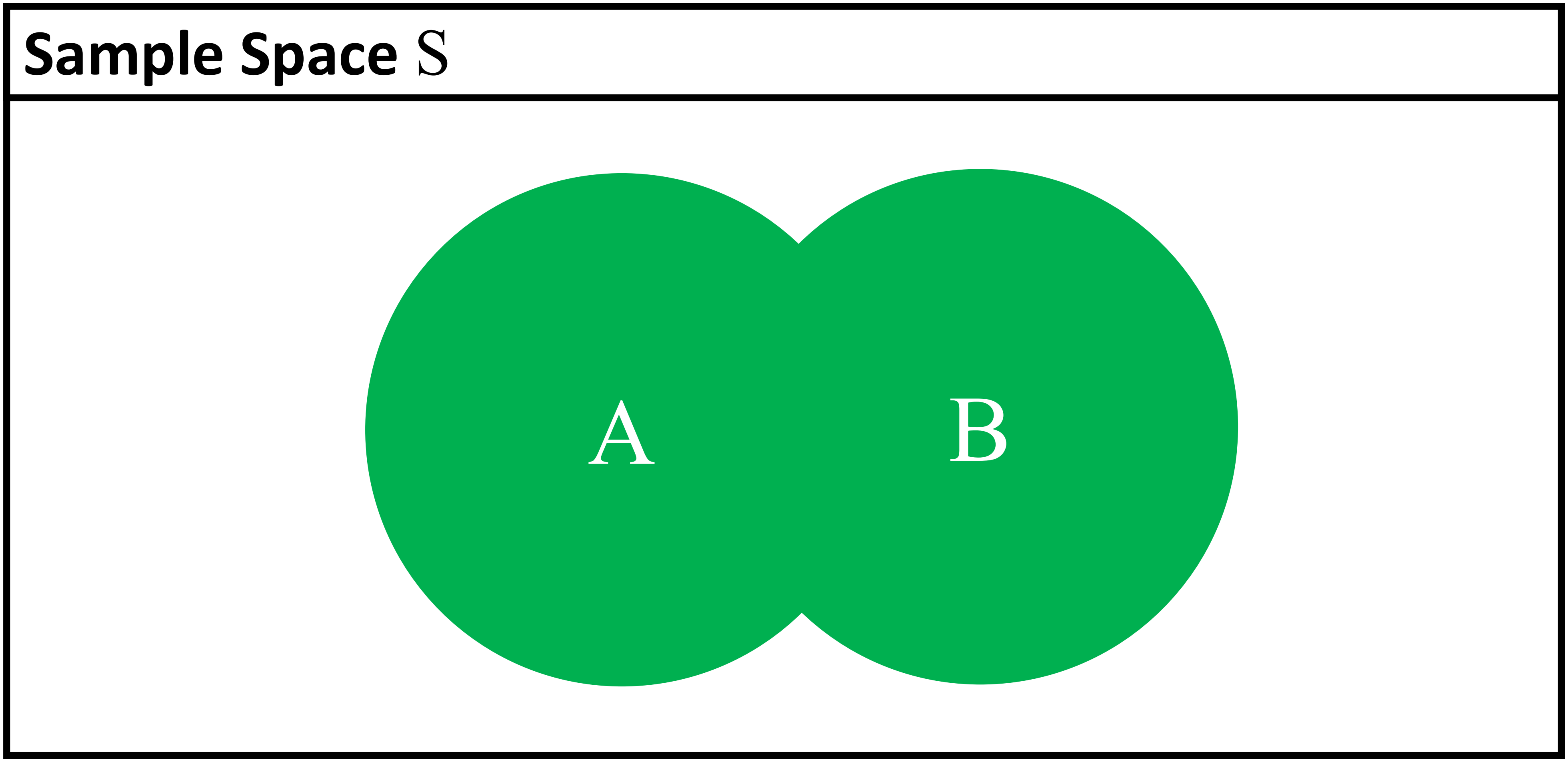
Events A and B are overlapping

Event (Set) Union: Probability



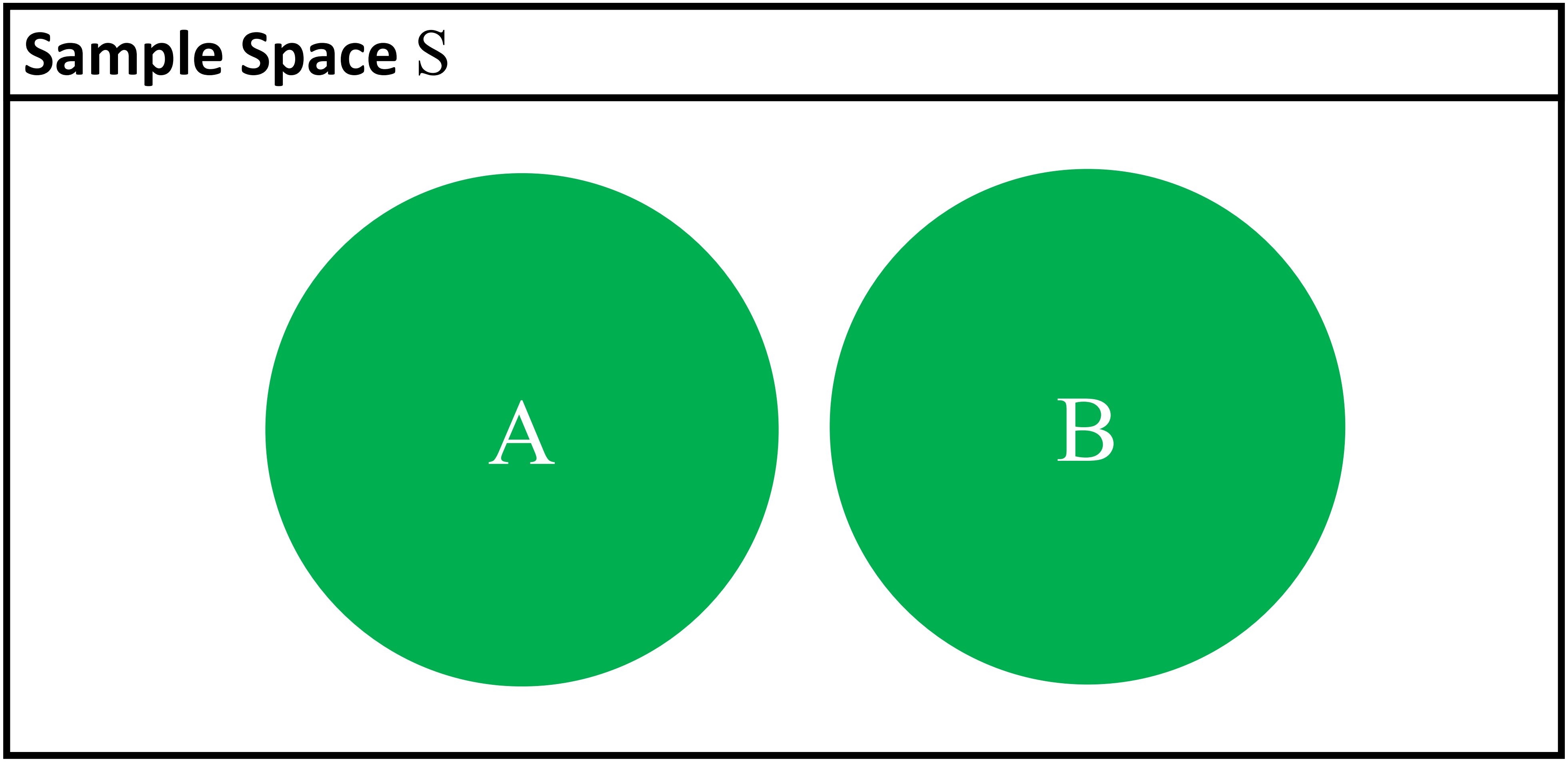
Events A and B are disjoint / mutually exclusive

Event (Set) Union: Probability



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

Event (Set) Union: Probability



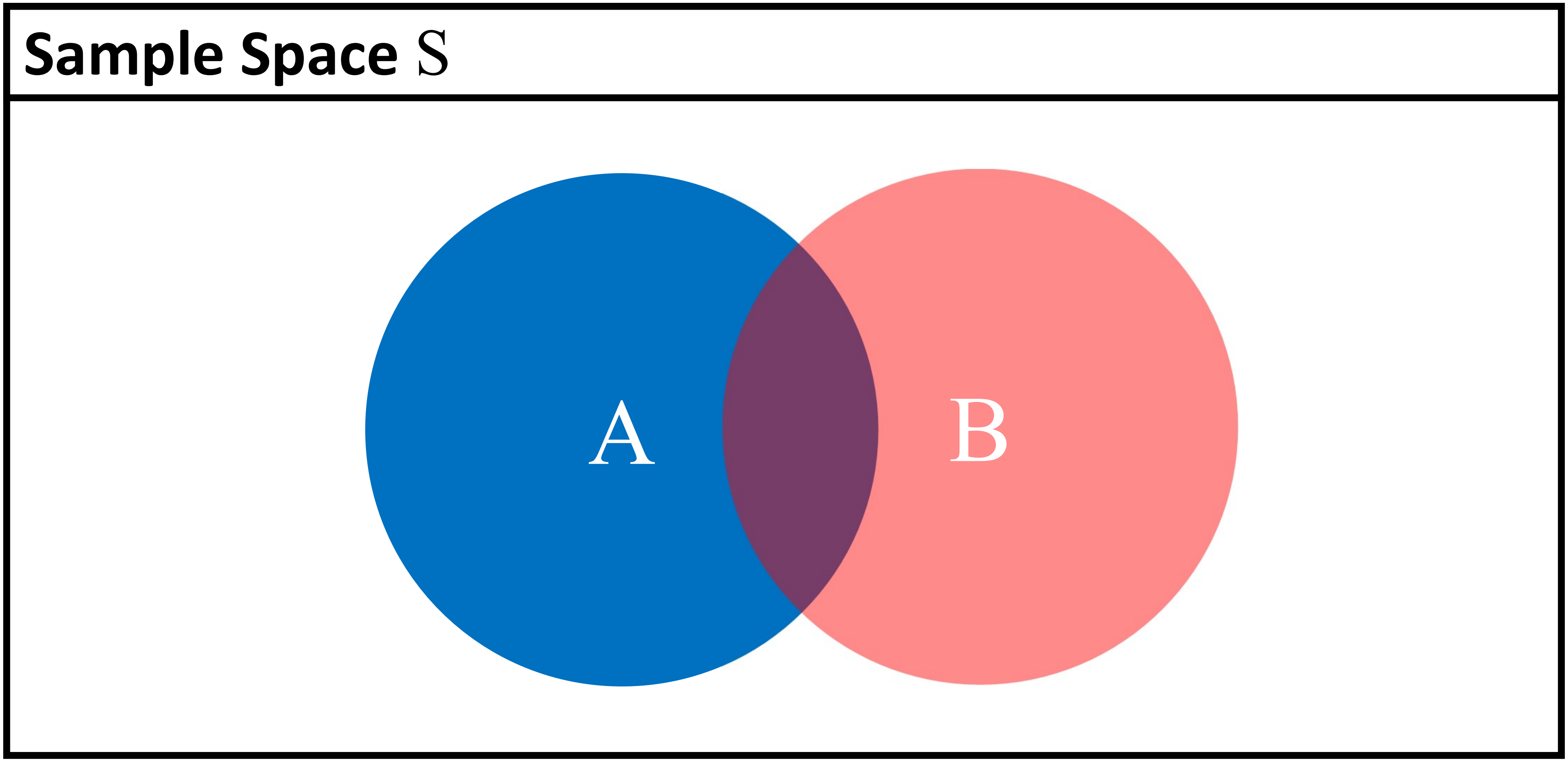
$$P(A \cup B) = P(A) + P(B)$$

Joint Probability

The probability of event A and event B occurring (or more than two events). It is the probability of the intersection of two or more events.

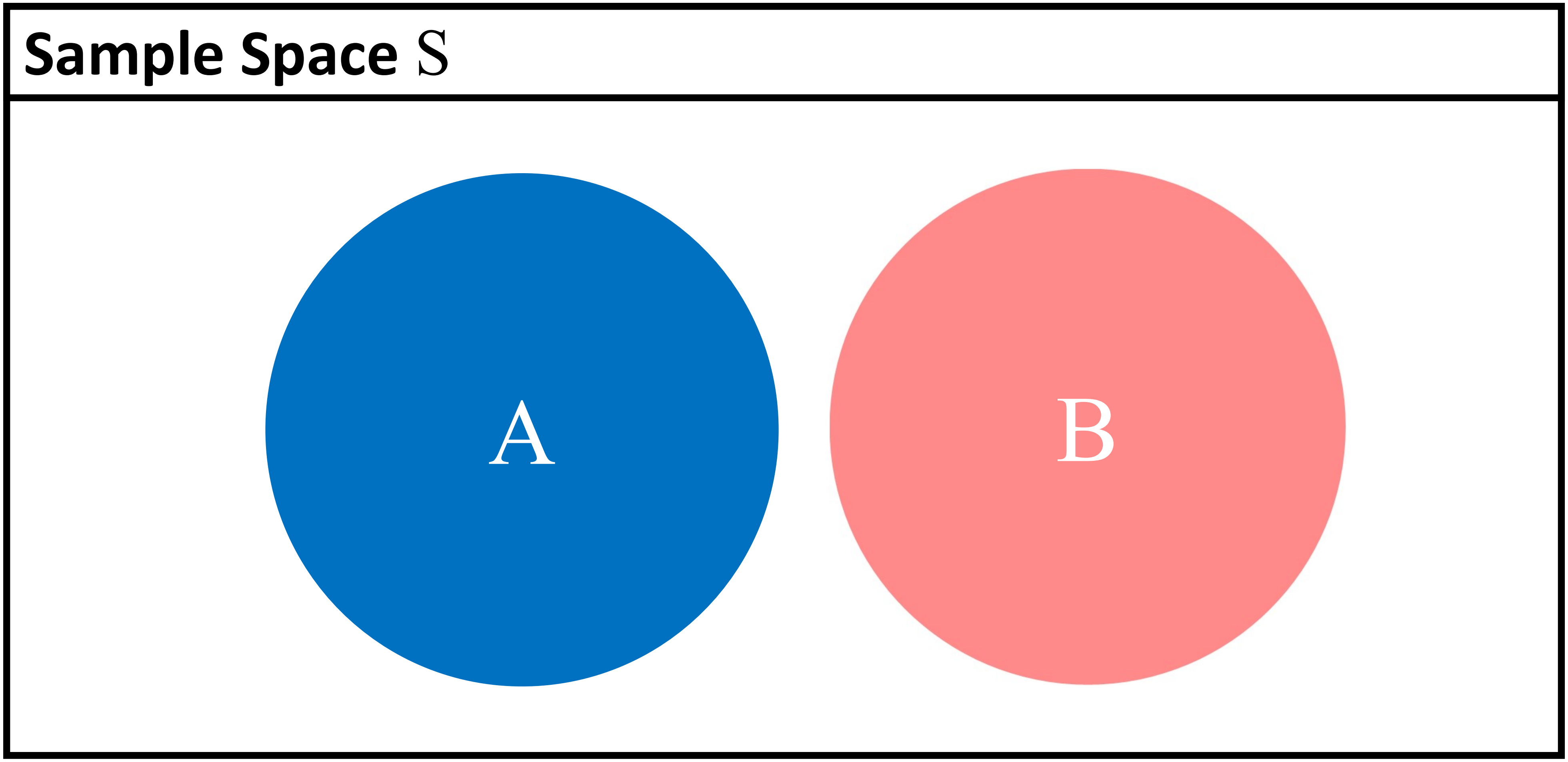
$$\mathrm{P}(\mathrm{A} \cap \mathrm{B}) = \mathrm{P}(\mathrm{A}, \mathrm{B}) = P(A \text{ and } B) = P(A \wedge B)$$

Event (Set) Intersection: Probability



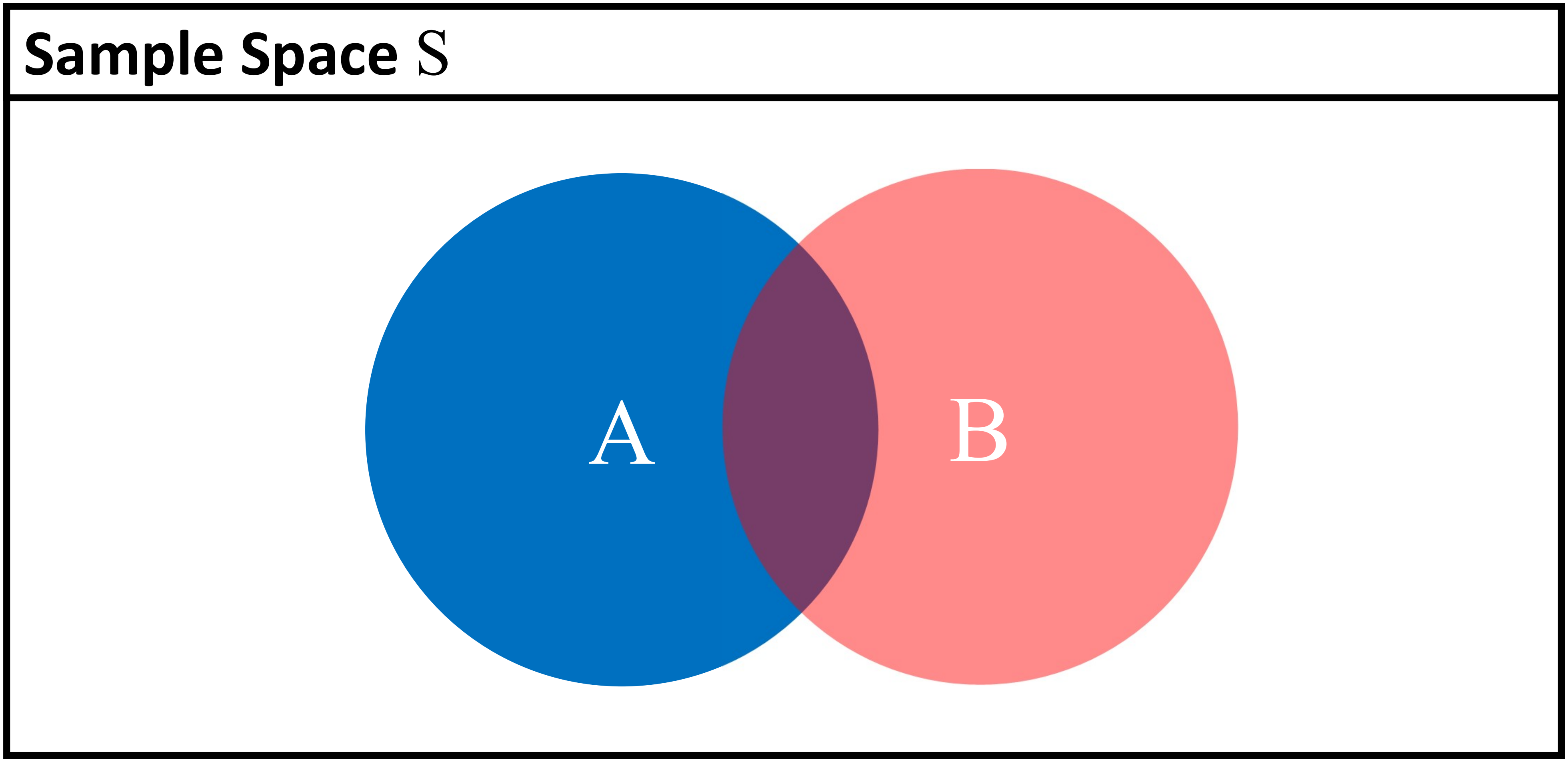
$$P(A \cap B) = P(A \text{ and } B) = P(A, B) = P(A \wedge B)$$

Event (Set) Intersection: Probability



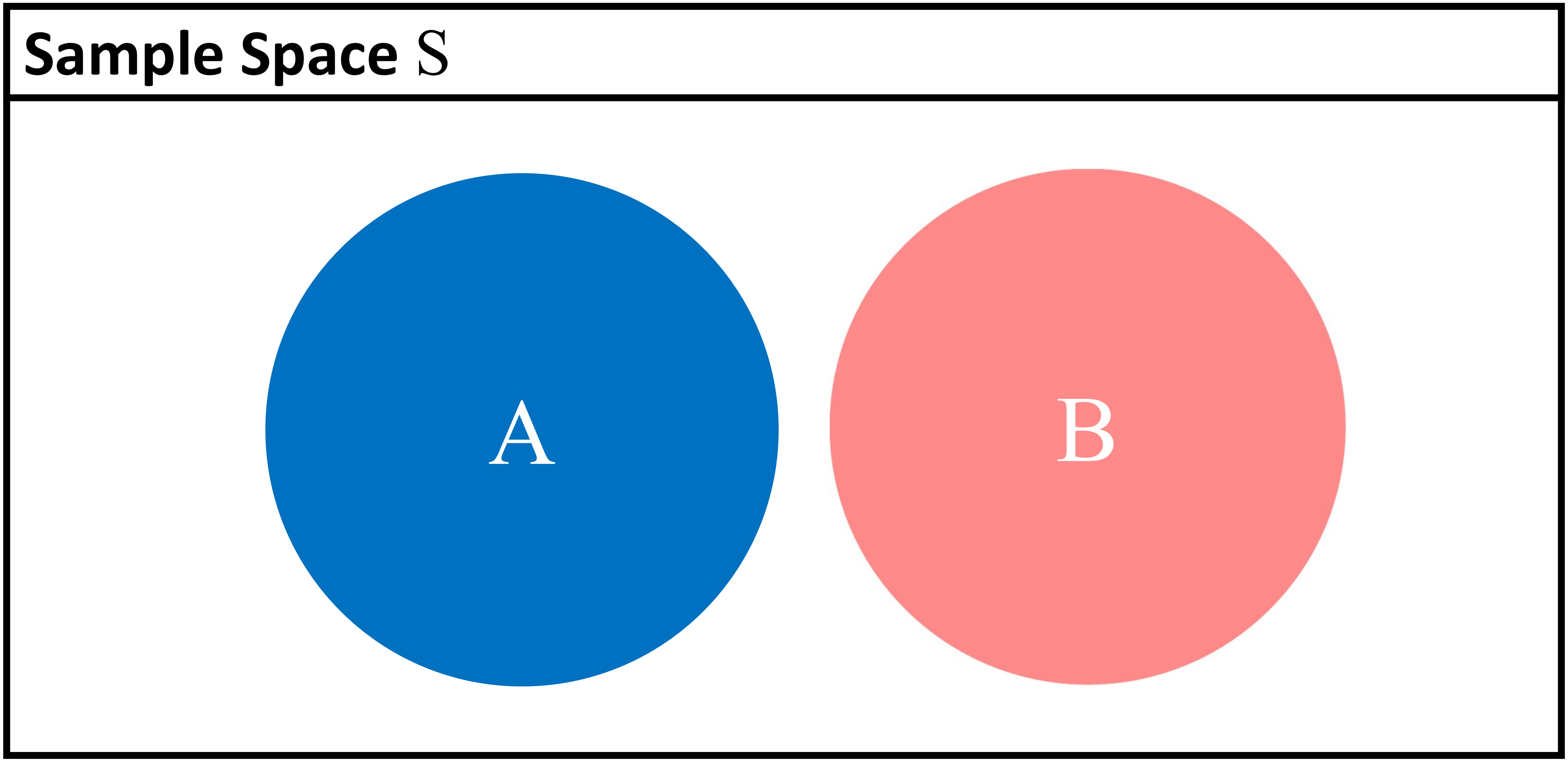
$$P(A \cap B) = P(A \text{ and } B) = P(A, B) = P(A \wedge B)$$

Event (Set) Intersection: Probability



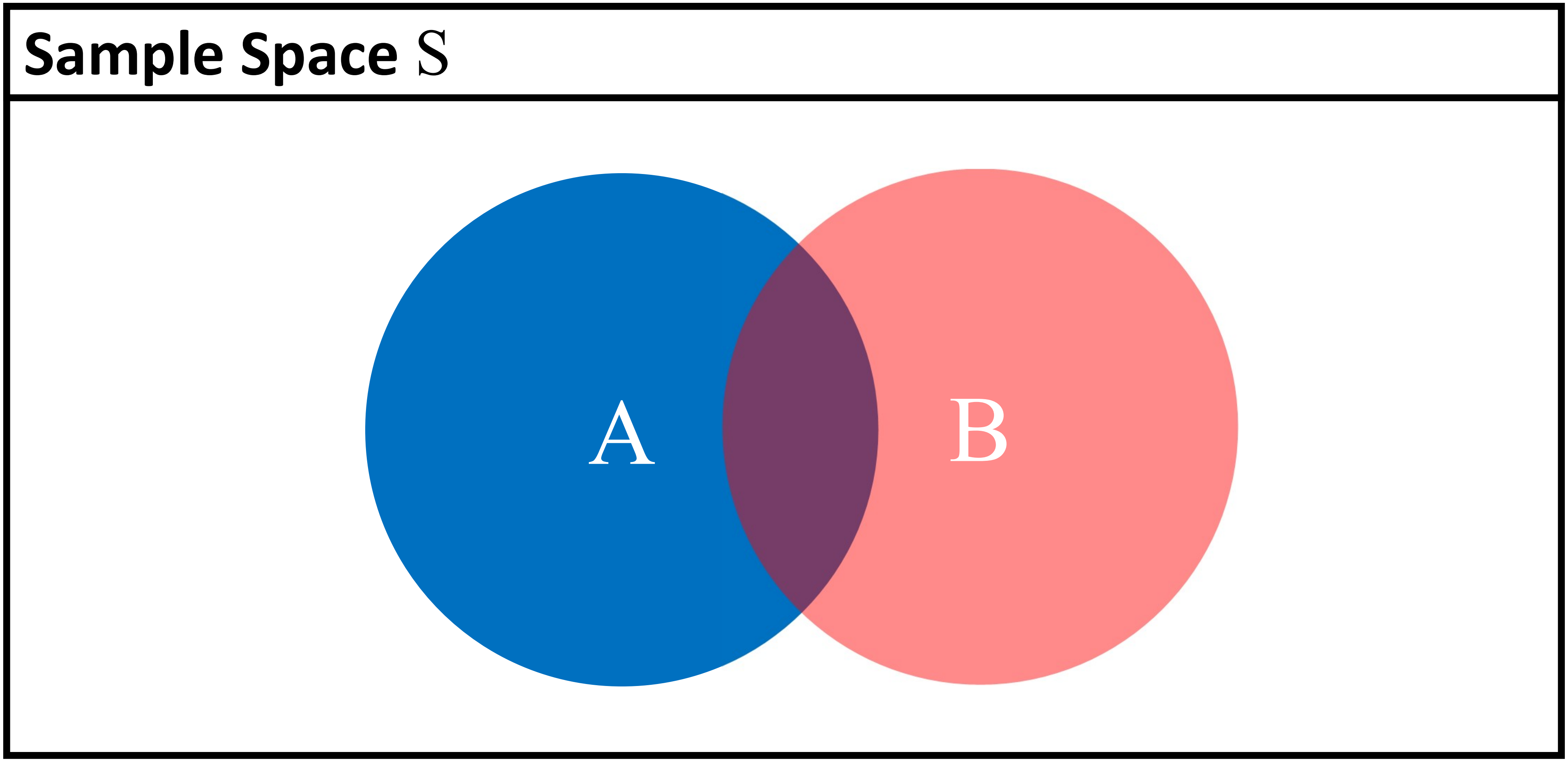
Events A and B are overlapping

Event (Set) Intersection: Probability



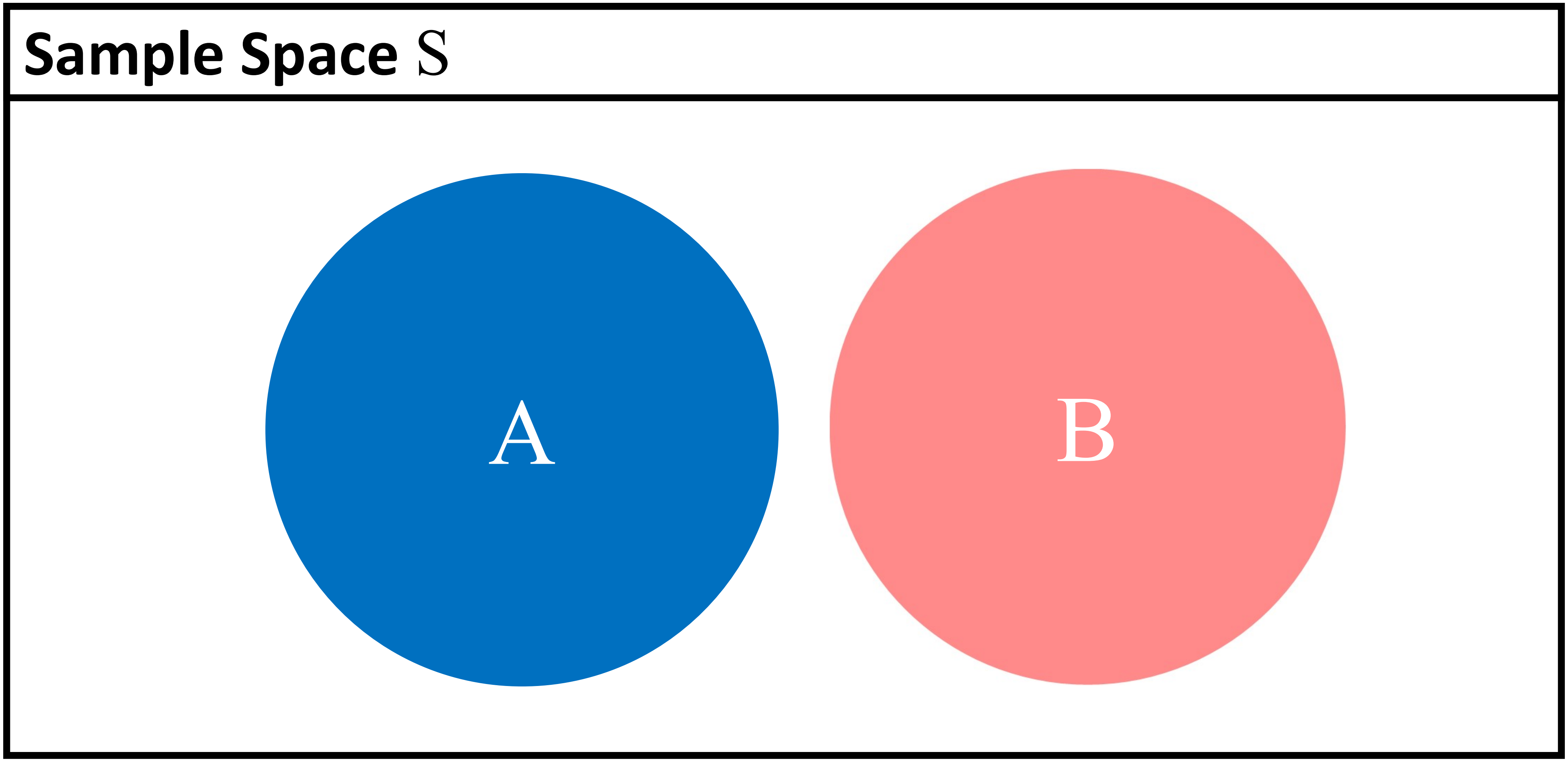
Events A and B are disjoint / mutually exclusive

Event (Set) Intersection: Probability



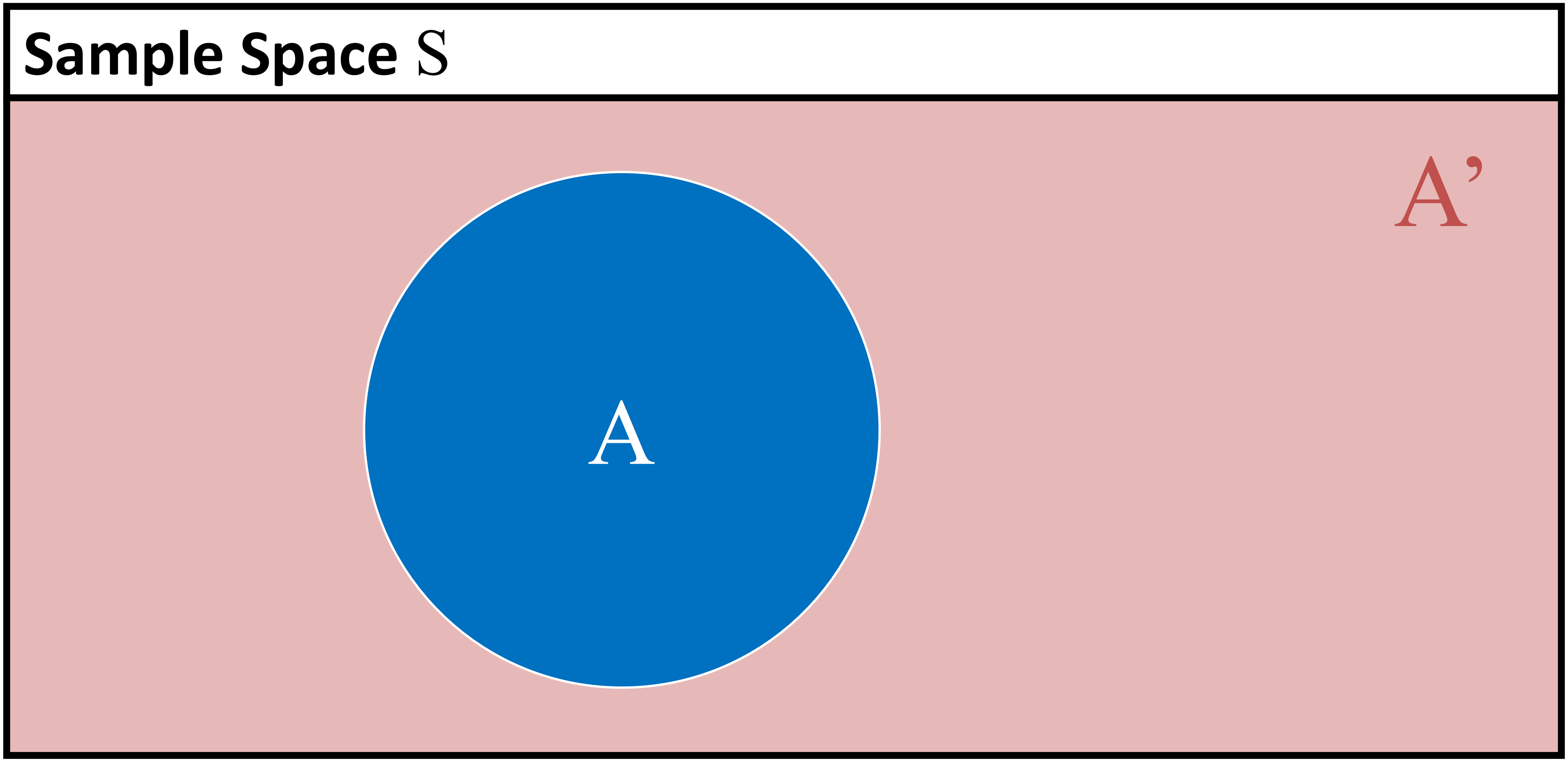
$$P(A \cap B) = P(A, B) > 0$$

Event (Set) Intersection: Probability



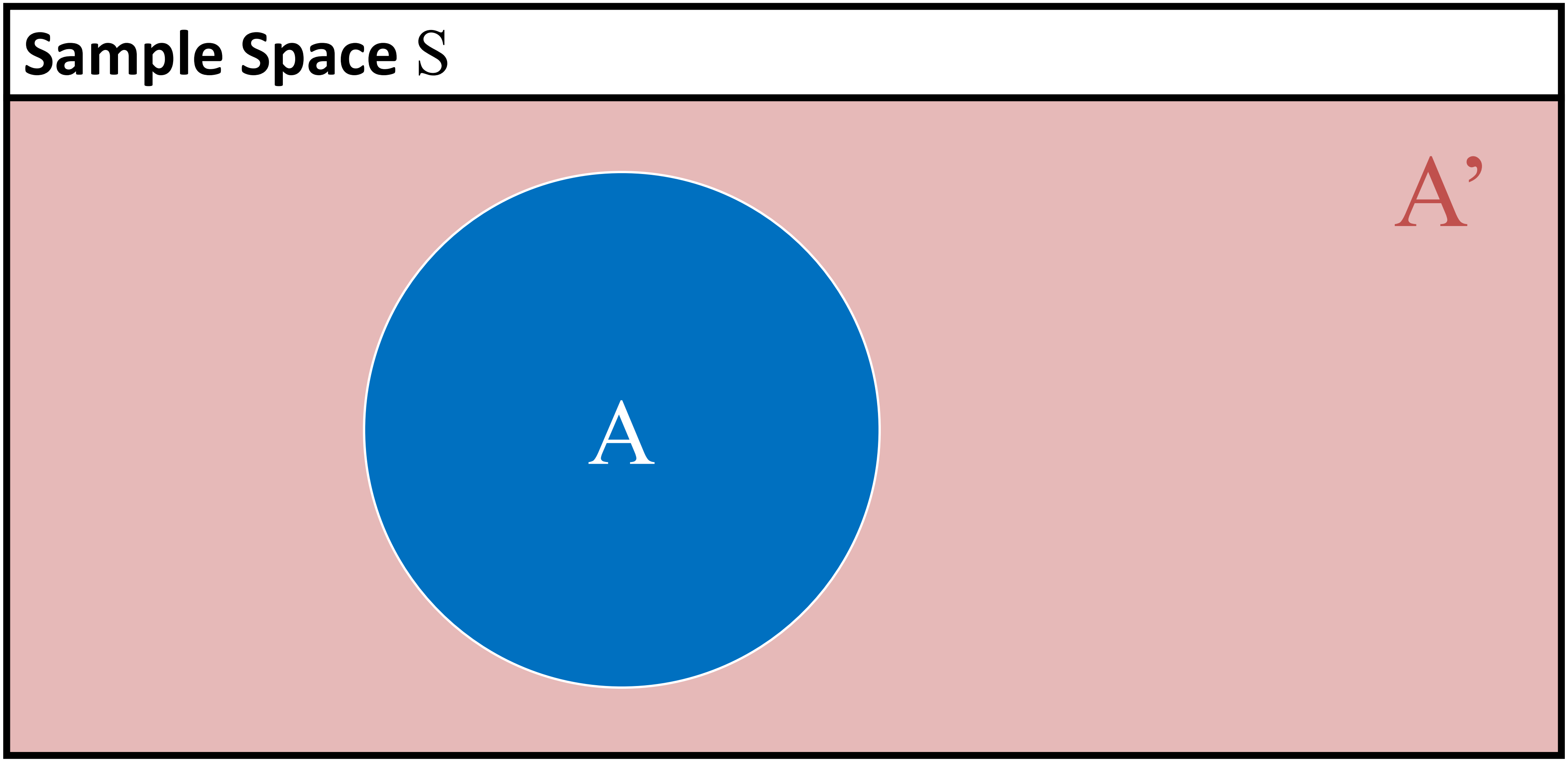
$$P(A \cap B) = P(A, B) = P(\varnothing) = 0$$

Complementary Events: Probability



$$P(A \cup A') = P(S) = 1$$

Complementary Events: Probability



$$P(A) = 1 - P(A')$$

Probability Theory and Propositions

Assume that A and B are sentences in propositional logic.

- $\mathrm{P}(\mathsf{T}) = 1$
- $\mathrm{P}(\perp) = 0$
- $\mathrm{P}(\mathrm{A} \vee \mathrm{B}) = \mathrm{P}(\mathrm{A}) + \mathrm{P}(\mathrm{B})$ **if** $\neg(\mathrm{A} \wedge \mathrm{B})$ **is a tautology**
- $\mathrm{P}(\mathrm{A}) + \mathrm{P}(\neg\mathrm{A}) = 1$
- $\mathrm{P}(\mathrm{A}) = \mathrm{P}(\mathrm{B})$ **if** $(\mathrm{A} \Leftrightarrow \mathrm{B})$ **is a tautology (logical equivalence)**
- $0 \leq \mathrm{P}(\mathrm{A})$ **for any sentence** A

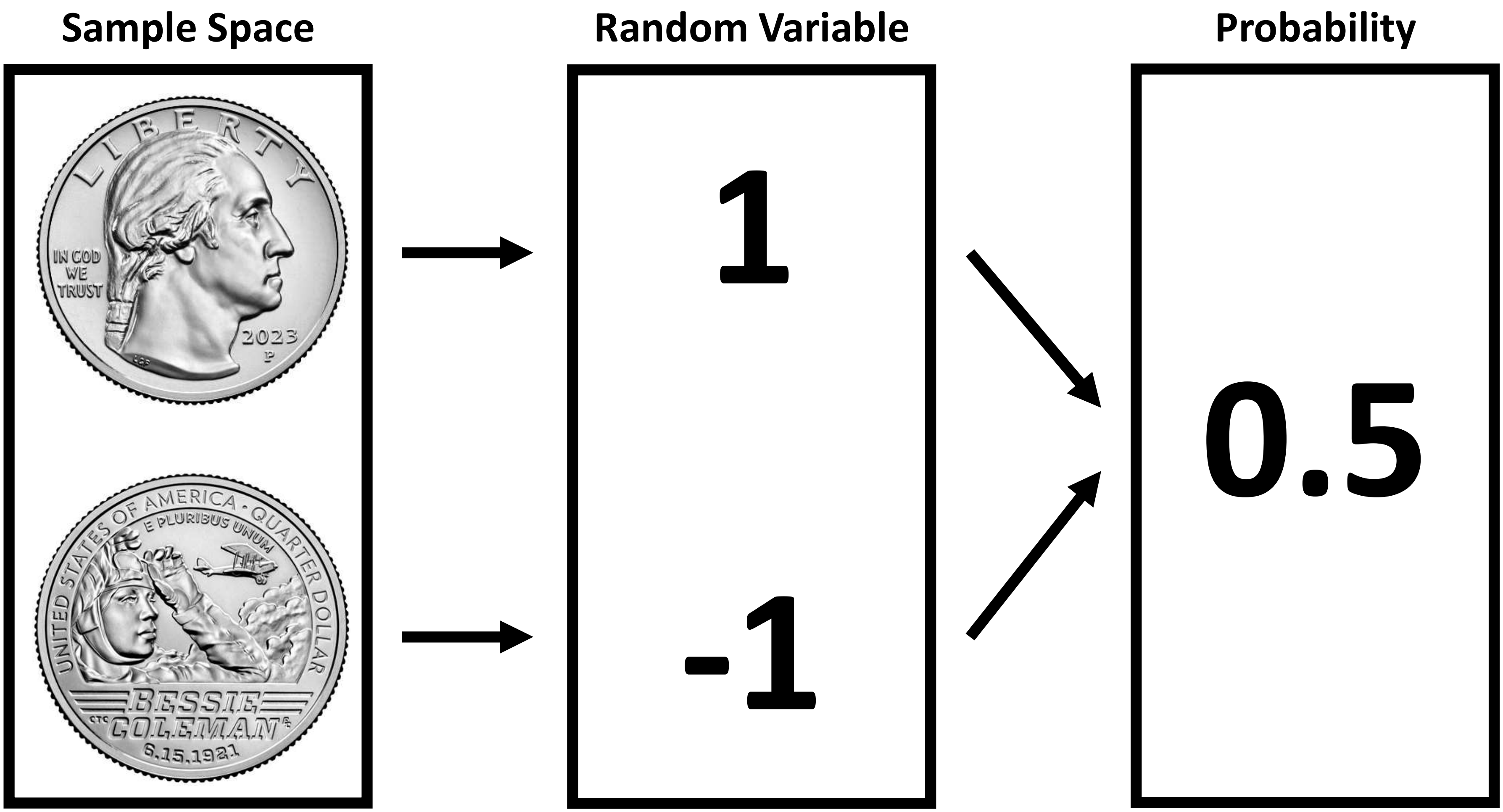
Random Variable

A Random Variable is a mathematical formalization of a quantity or object which depends on random events

A Random Variable X is a function mapping events/outcomes from the sample space S to a measurable space (such as $\mathbb{R}$) :

$$\mathrm{X}: \mathrm{S} \rightarrow \mathbb{R}$$

Random Variable

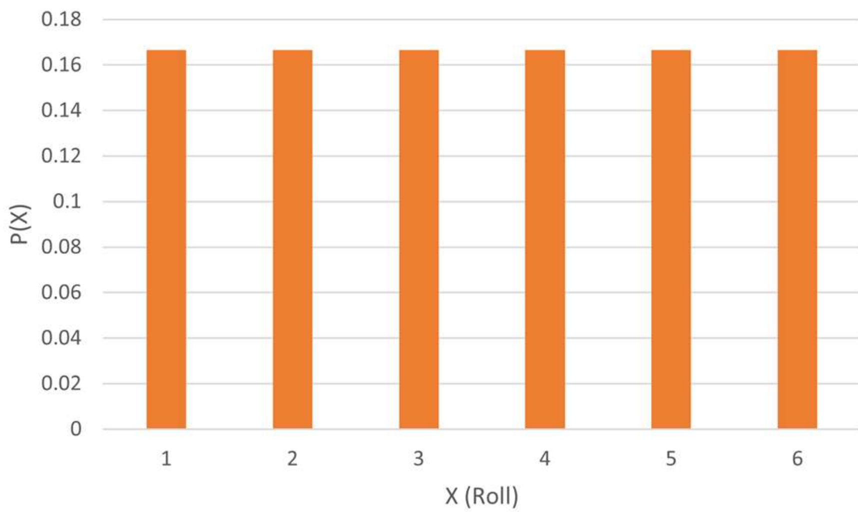


Random Variable Distribution

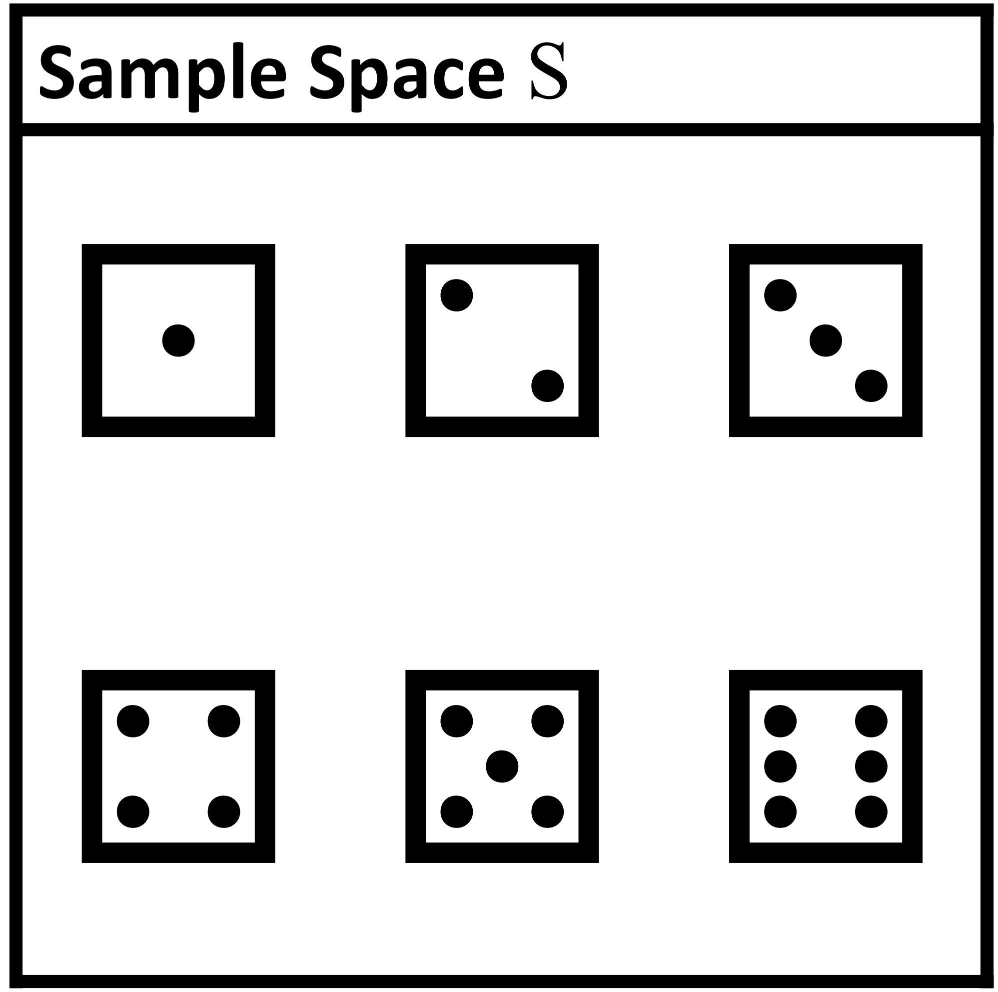
The probability distribution for a discrete random variable X can be perceived as a frequency distributions.

It is a graph, table or formula that gives the possible values of X and the probability $P(X)$ associated with each value of X.

Single Die Roll: Distribution



X	P (X)
⚀	1/6
⚁	1/6
⚂	1/6
⚃	1/6
⚄	1/6
⚅	1/6



Random Variable: Typical Notation

- Capital: X: a variable
- Lowercase: x: a particular value of X
- $Val(X)$: the set of values X can take
- Bold Capital: $\mathbf{X}$: a set of variables
- Bold lowercase: $\mathbf{x}$: an assignment to all variables in $\mathbf{X}$
- $P(X = x)$ will be shortened as $P(x)$
- $P(X = x \cap Y = y)$ will be shortened as $P(x, y)$
- $\mathbf{P}(X)$: probability distribution for X

Random Variable: Typical Notation

- Pick variables of interest/relevance
  - Medical diagnosis
    - Age, gender, weight, temperature, …
  - Loan application
    - Income, savings, payment history, …
  - other
- Every variable has a domain
  - Binary (e.g., True/False)
  - Categorical (e.g., Red/Green/Blue)
  - Real-valued (e.g., 97.8)
- Possible world
  - An assignment to all variables of interest

Joint Probability

The probability of event A and event B occurring (or more than two events). It is the probability of the intersection of two or more events.

$$\mathbf{P}(\mathrm{A} \cap \mathrm{B}) = \mathbf{P}(\mathrm{A}, \mathrm{B}) = \mathbf{P}(\mathrm{A \text{ and } B}) = \mathbf{P}(\mathrm{A} \wedge \mathrm{B})$$

For example (specific probability shown):

$$\mathrm{P}(\text{pressure} = 90, \text{temperature} = 100, \text{volume} = 6) = 0.1$$

For any random variables: $f_1, f_2, \ldots, f_n$ **:**

$$\boldsymbol{P}(f_1, f_2, \ldots, f_n)$$

Probability Model

A fully specified probability model associates a numerical probability $\mathrm{P}(\omega)$ with each possible world (assume there is a finite number of such worlds):

$$0 \leq \mathrm{P}(\omega) \leq 1, \text{for every} \sum_{\omega \in S} \mathrm{P}(\omega) = 1$$

Joint Probability

The probability of event A and event B occurring (or more than two events). It is the probability of the intersection of two or more events.

$$\mathbf{P}(\mathrm{A} \cap \mathrm{B}) = \mathbf{P}(\mathrm{A}, \mathrm{B}) = \mathbf{P}(\mathrm{A \text{ and } B}) = \mathbf{P}(\mathrm{A} \wedge \mathrm{B})$$

For example (specific probability shown):

ONE POSSIBLE "WORLD":

$$\mathrm{P}(\text{pressure} = 90, \text{temperature} = 100, \text{volume} = 6) = 0.1$$

For any random variables: $f_1, f_2, \ldots, f_n$ **:**

$$\boldsymbol{P}(f_1, f_2, \ldots, f_n)$$

Random Variables, Events, Logic

An **event** is the set of possible worlds where a given predicate is true

- Roll two dice
  - The possible worlds are (1,1), (1,2), …, (6,6); 36 possible worlds
  - Predicate = two dice sum to 10
  - Event = {(4,6), (5,5), (6,4)}
- Toothache and cavity
  - Four possible worlds: $(t, c), (t, \sim c), (\sim t, c), (\sim t, \sim c)$
  - Some worlds are more likely than others
  - Predicate can be anything about these variables: $t \wedge c, t, t \vee \sim c,$

Complex Joint Probability Distribution

Consider a complex joint probability distribution involving N random variables $f_1, f_2, f_3, \dots, f_{N-1}, f_N$. [values can be OTHER than true/false and non-binary]

2^N Possible Worlds (Models)	f_1	f_2	f_3	…	f_{N-1}	f_N	Joint Probability	2^N values
	true	true	true	…	true	true	0.0011	
	true	true	true	…	true	false	0.0451	
	true	true	false	…	false	true	0.1011	
	…	…	…	…	…	…	…	
	false	false	true	…	true	false	0.0909	
	false	false	true	…	false	true	0.0651	
	false	false	false	…	false	false	0.2021	

(Columns f_1 … f_N: N Random Variables)

Frequentist versus Causal Perspective

- **Frequentist view:**

  Probability represents long-run frequencies of repeatable events.

- **Causal perspective:**

  Probability is a measure of belief.

Prior (Unconditional) Probabilities

Degree of belief that some event A is occurred *in the absence of any other related information* is called unconditional or prior probability (or "prior" for short) $P(A)$.

Examples:

$$P(\text{isRaining})$$

$$P(\text{dieRoll} = 5)$$

$$P(\text{CourseFinalGrade} = \text{'A'})$$

$$P(\text{toothache})$$

Conditioning

Conditioning is a process of revising beliefs based on new evidence e:

- **start by taking all background information (prior probabilities) into account**
- **if new evidence e is acquired, a conditional probability of some proposition A given evidence e can be calculated (posterior probability): $P(A \mid e)$**

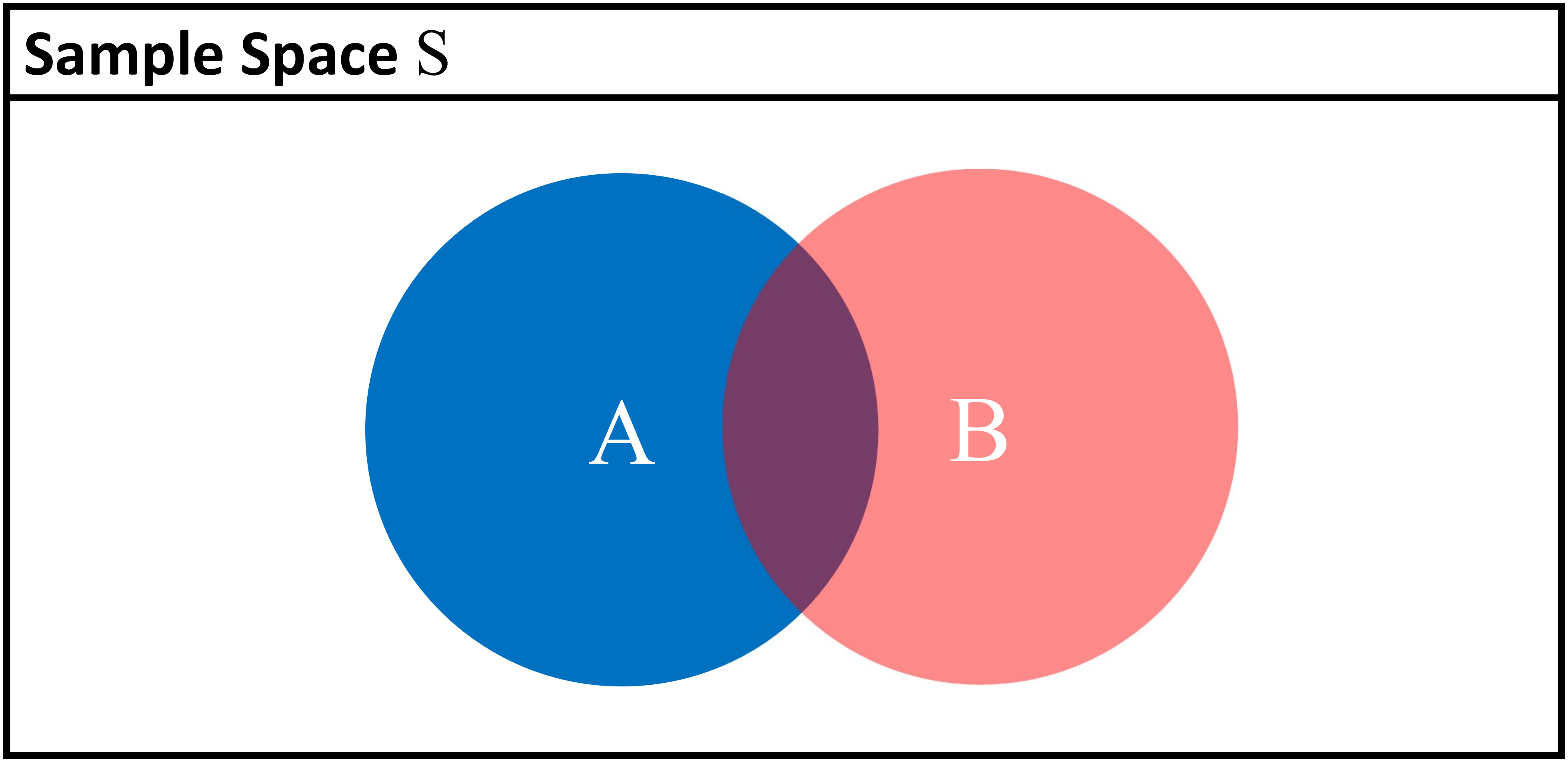
Conditional Probability

If A and B are two events in sample space S, then conditional probability of A given B is defined as:

$$P(A \text{ given } B) = P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

where: $P(B) > 0$

Conditional Probability: Venn Diagram



$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A, B)}{P(B)} = \frac{P(A \wedge B)}{P(B)}$$

Conditional Probability

If A and B are two events in sample space S, then conditional probability of A given B is defined as:

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

where: $P(B) > 0$ ← [Otherwise B is impossible]

Conditional Probability: Notation

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

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$$

$$P(A \mid B) = \frac{P(A, B)}{P(B)}$$

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

Conditional Probability

If A **and** $\mathrm{evidence}$ **are two events in sample space** S**, then conditional probability of** A **given** $\mathrm{evidence}$ **is defined as:**

$$P(A \mid \mathrm{evidence}) = \frac{P(A \cap \mathrm{evidence})}{P(\mathrm{evidence})}$$

where: $P(\mathrm{evidence}) > 0$

Posterior (Conditional) Probabilities

Typically, there is going to be some information, called evidence e, that affects our degree of belief about some event A being occurring. This allows us to also consider conditional or posterior probability (or "posterior" for short) $P(A \mid e)$.

Examples ($P(A \text{ given } e)$):

$$P(\text{isRaining} \mid \text{cloudy})$$

$$P(\text{CourseFinalGrade} = \text{'A'} \mid \text{CoursePA1Score} > 80)$$

$$P(\text{cavity} \mid \text{toothache})$$

Evidence e

Evidence e rules out possible worlds incompatible with e.

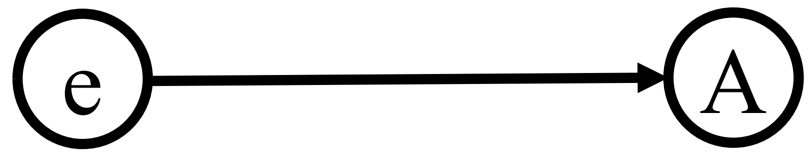
Prior vs. Posterior Probabilities

Prior Probability

$$P(A)$$

BTW: it is also $P(A \mid T)$

Posterior Probability



$$P(A \mid e)$$

Conditional Probability: Notation

$$P(A \mid \text{evidence}) = \frac{P(A \cap \text{evidence})}{P(\text{evidence})}$$

$$P(A \mid \text{evidence}) = \frac{P(A \text{ and evidence})}{P(\text{evidence})}$$

$$P(A \mid \text{evidence}) = \frac{P(A, \text{evidence})}{P(\text{evidence})}$$

$$P(A \mid \text{evidence}) = \frac{P(A \wedge \text{evidence})}{P(\text{evidence})}$$

Conditional Probability: Notation

$$\mathrm{P}(A, B, C, D \mid E, F, G) = \frac{\mathrm{P}(A, B, C, D, E, F, G)}{\mathrm{P}(E, F, G)}$$

Axioms of Conditional Probability

Axiom 1:

For any event A**,** $P(A \mid B) \geq 0$

Axiom 2:

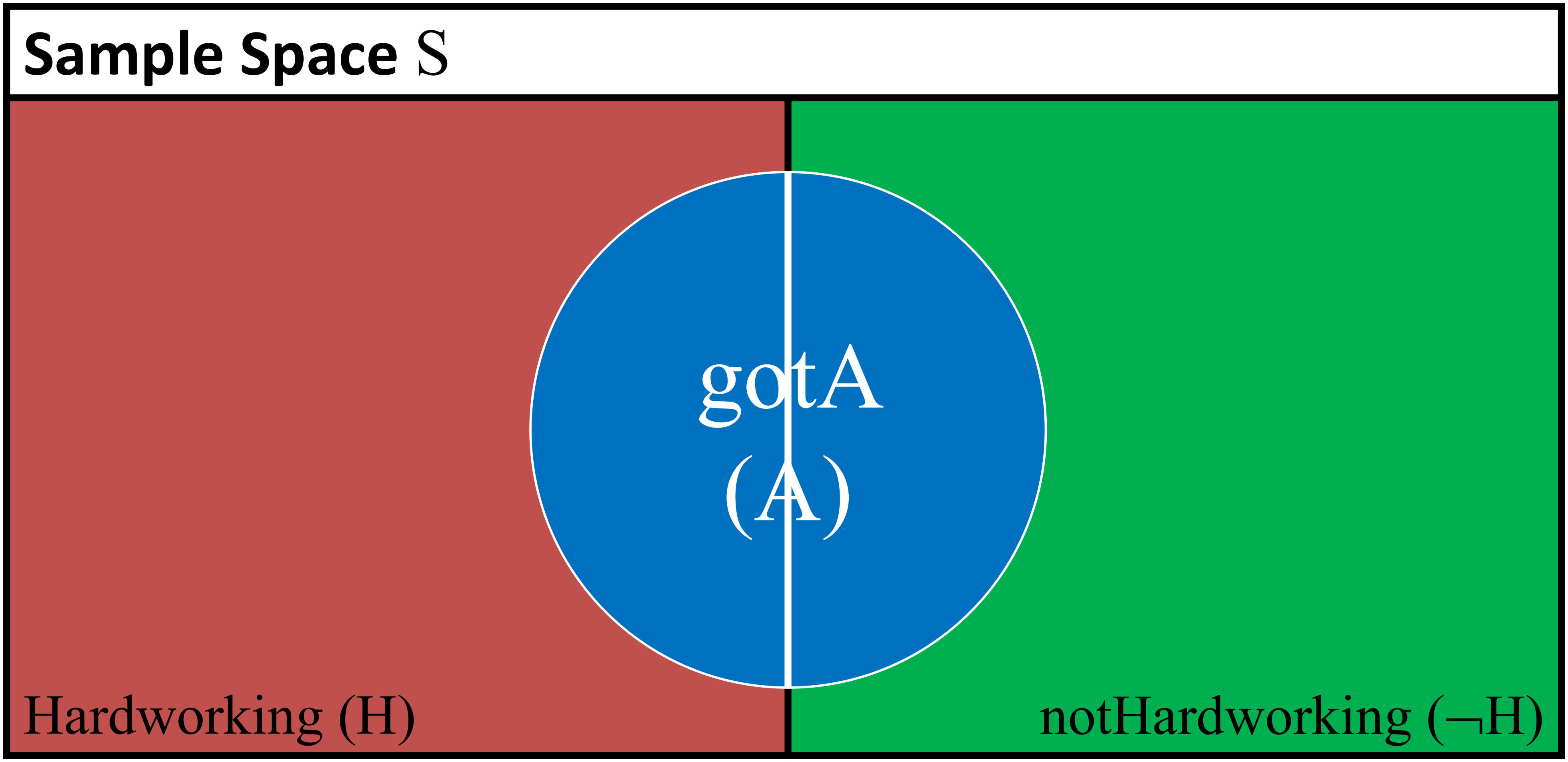
Conditional probability of B **given** B **is** $P(B \mid B) = 1$

Axiom 3:

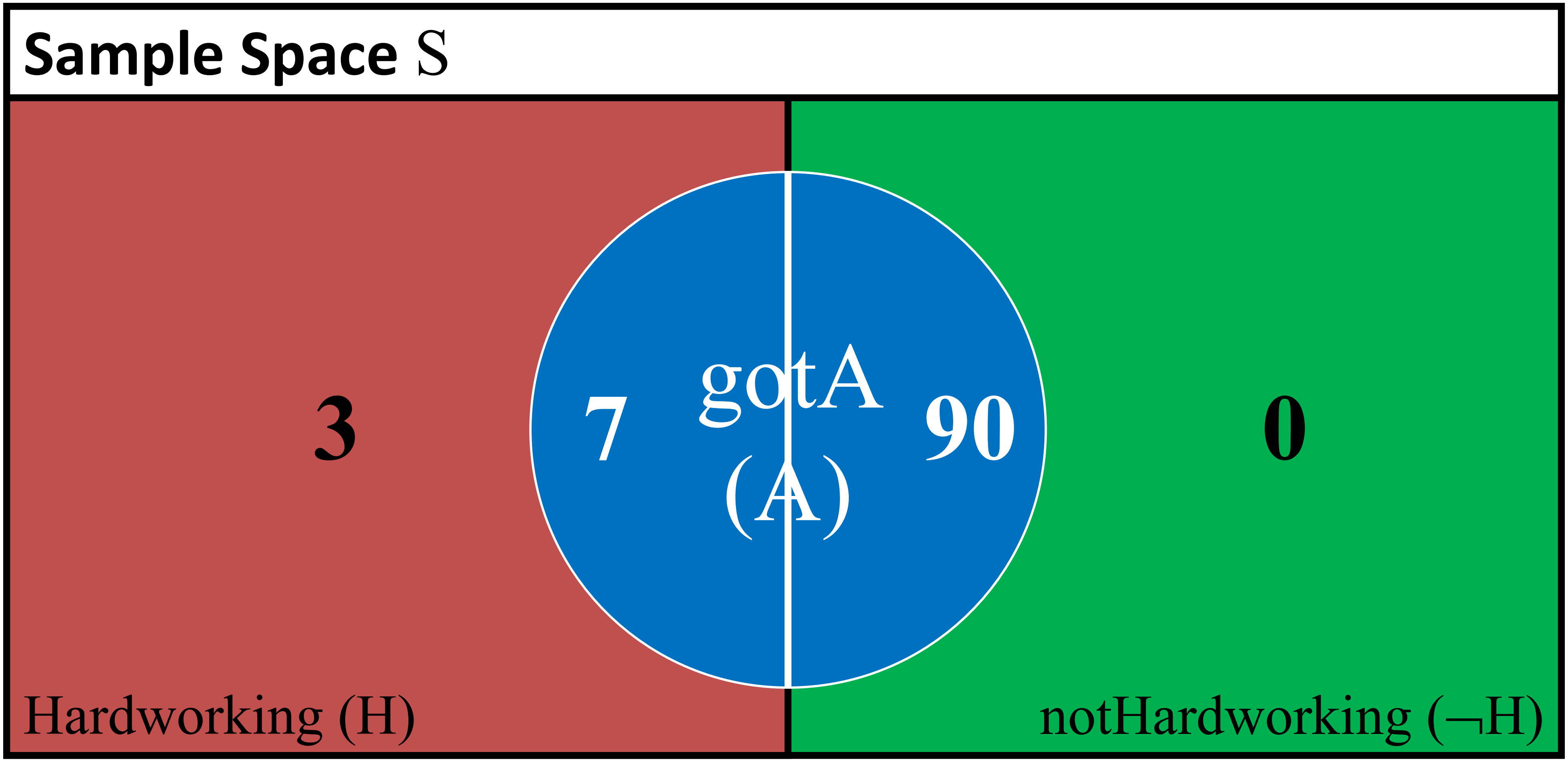
If A_1**,** A_2**, … are disjoint events, then**

$$P(A_1 \cup A_2 \cup \ldots \mid B) = P(A_1 \mid B) + P(A_2 \mid B) + \ldots$$

Conditional Probability: Visualization

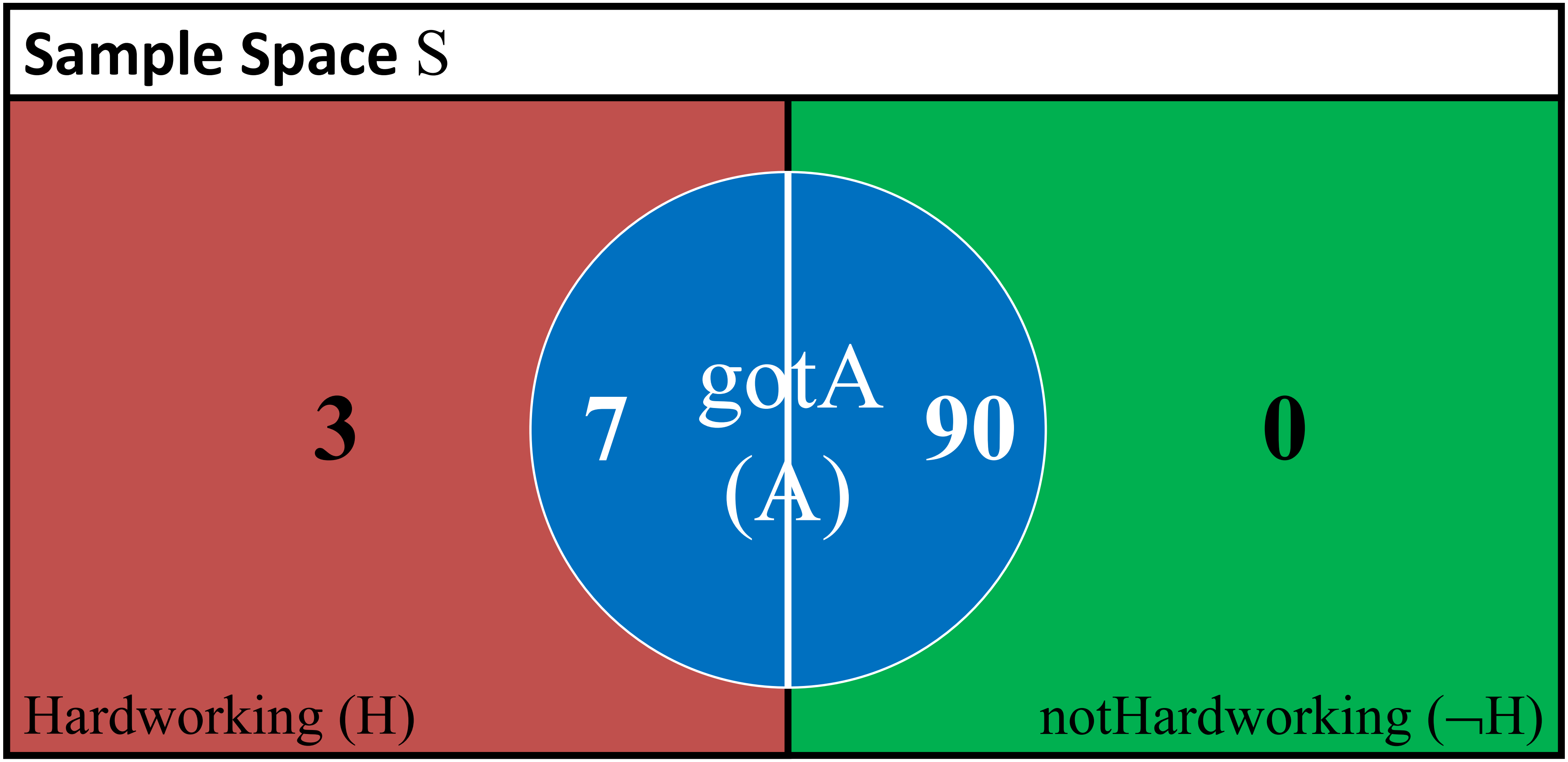


Conditional Probability: Visualization



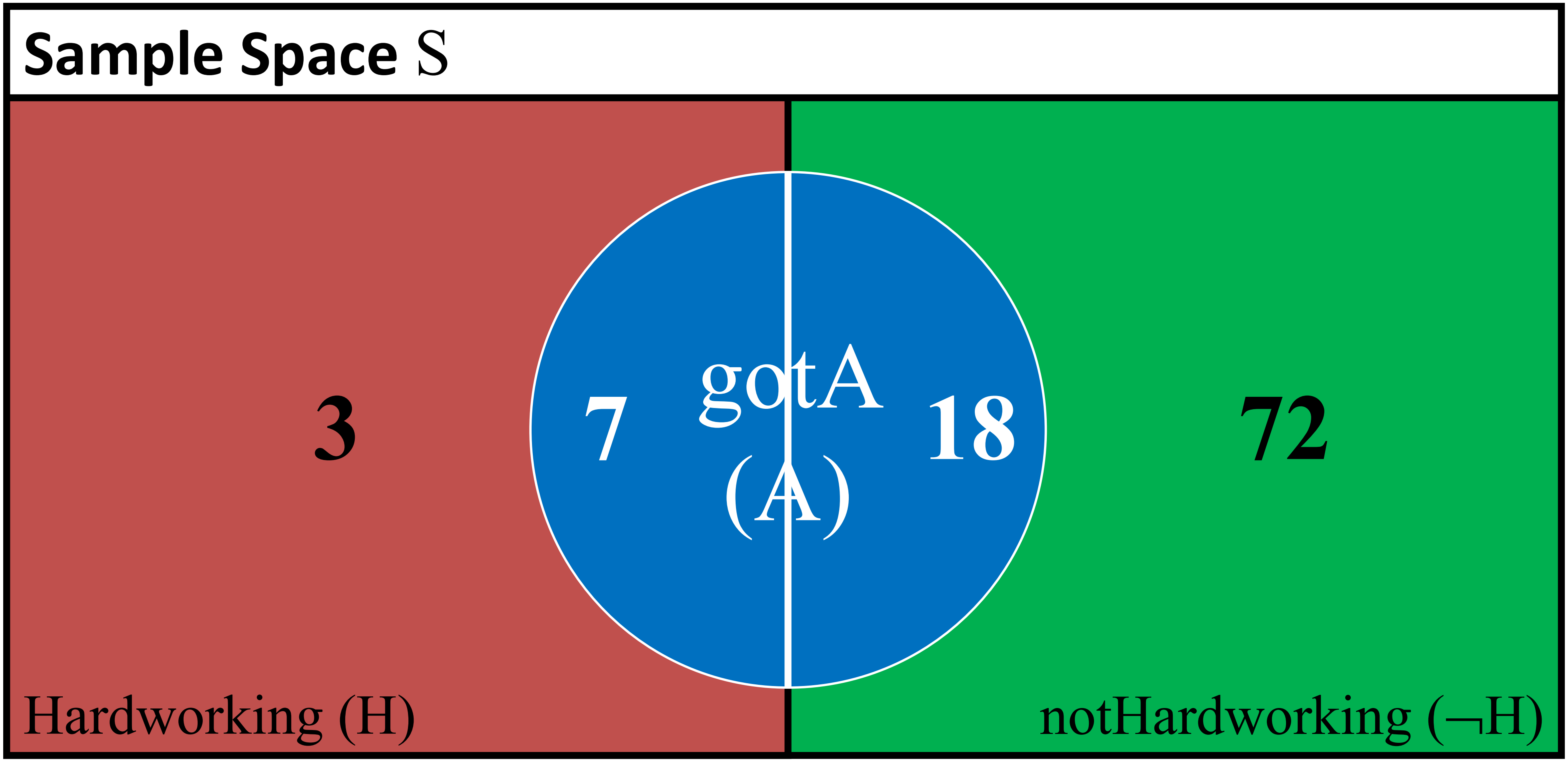
$$P(H \mid A) = ?$$

Conditional Probability: Visualization



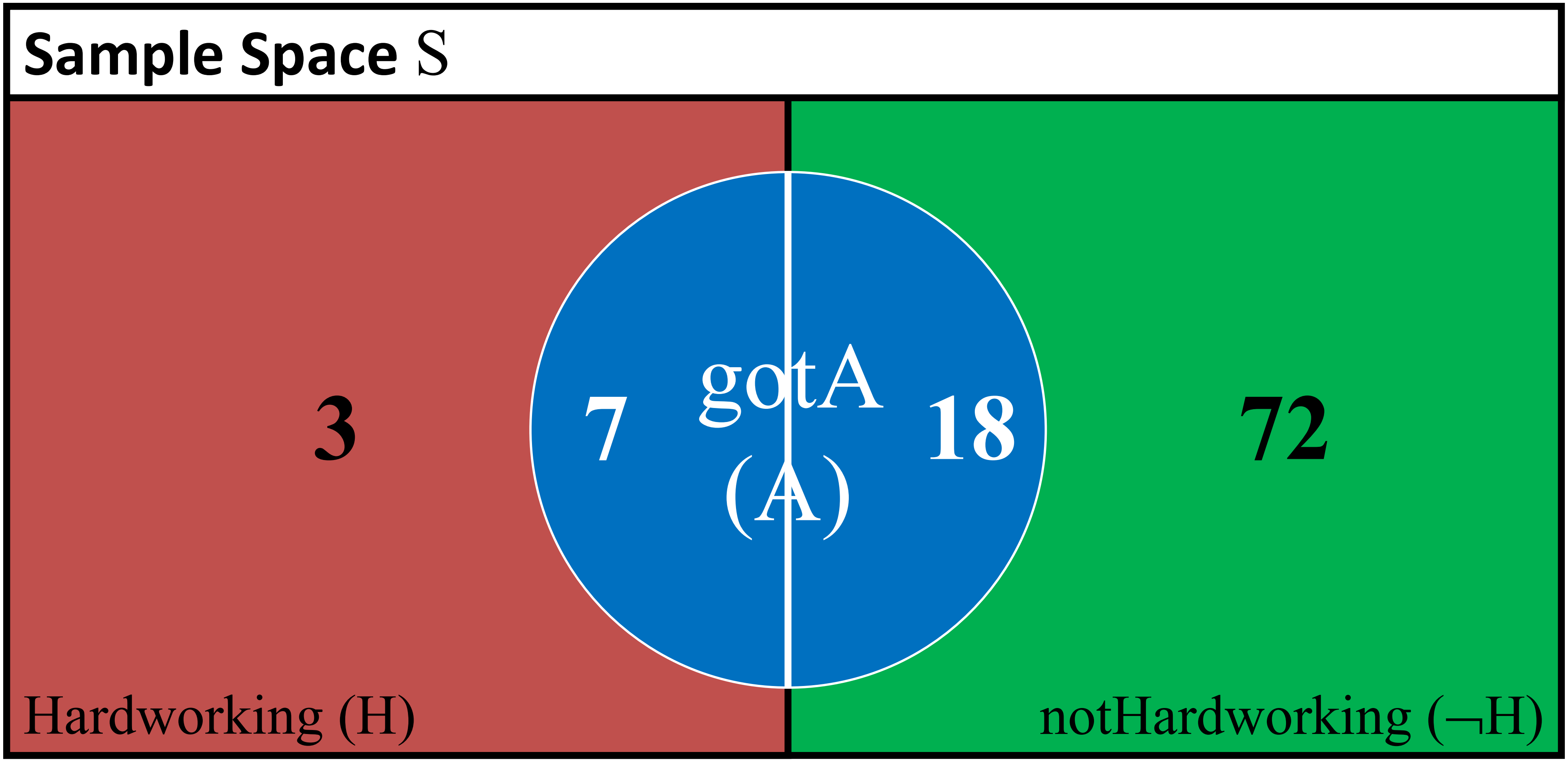
$$P(H \mid A) = \frac{P(H \cap A)}{P(A)} = \frac{7/100}{97/100} = \frac{7}{97}$$

Conditional Probability: Visualization



$$P(H \mid A) = ?$$

Conditional Probability: Visualization



$$P(H \mid A) = \frac{P(H \cap A)}{P(A)} = \frac{7/100}{25/100} = \frac{7}{25}$$

Chain Rule

Conditional probabilities can be used to decompose joint probabilities using the chain rule. For any random variables $f_1, f_2, \ldots, f_n$ and values $x_1, x_2, \ldots, x_n$:

$$P(f_1 = x_1, f_2 = x_2, \ldots, f_n = x_n) =$$

$$P(f_1 = x_1) *$$

$$P(f_2 \mid f_1 = x_1) *$$

$$P(f_3 \mid f_1 = x_1, f_2 = x_2) *$$

$$\ldots$$

$$P(f_n = x_n \mid f_1 = x_1, \ldots, f_{n-1} = x_{n-1}) =$$

$$= \prod_{i=1}^{n} P(f_i = x_i \mid f_1 = x_1, \ldots, f_{i-1} = x_{i-1})$$

Independence

Two events are independent if one does not convey any information about the other.

Two events A and B are independent if:

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

Independence

Two events A and B are independent if:

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

So (from conditional probability formula):

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

Disjointment vs. Independence

Concept	Meaning	Formulas
Disjoint	Events A and B cannot occur at the same time	$A \cap B = \varnothing$ $P(A \cup B) = P(A) + P(B)$
Independent	Event A does not give any information about event B	$P(A \mid B) = P(A)$ $P(B \mid A) = P(B)$ $P(A \cap B) = P(A) * P(B)$

Independence

If two events A and B are independent:

- **events A and B' are independent**
- **events A' and B are independent**
- **events A' and B' are independent**

Independence

If A_1, A_2, … , A_N are independent events:

$$P(A_1 \cup A_2 \cup \ldots \cup A_N) =$$

$$= 1 - (1\text{-}P(A_1)) * (1\text{-}P(A_1)) * \ldots * (1\text{-}P(A_N))$$

Conditional Independence

Random variable X is conditionally independent of random variable Y given Z if for all $x \in Dx$, for all $y \in Dy$, and for all $z \in Dz$, such that

$$P(Y = y \wedge Z = z) > 0 \textbf{ and } P(Y = y' \wedge Z = z) > 0$$

$$P(X = x \mid Y = y \wedge Z = z) = P(X = x \mid Y = y' \wedge Z = z)$$

In other words, given a value of Z, knowing Y's value DOES NOT affect your belief in value of X.

Conditional Independence

The following four statements are equivalent as long as conditional probabilities:

1. X **is conditionally independent of** Y **given** Z
2. Y **is conditionally independent of** X **given** Z
3. $P(X \mid Y, Z) = P(X \mid Z)$
4. $P(X, Y \mid Z) = P(X \mid Z) * P(Y \mid Z)$

Conditional Independence

Consider three random variables: P(owerful), H(appy), R(ich)

with domains:

$$D_P = \{\text{powerful, powerless}\},\ D_H = \{\text{happy, unhappy}\},\ D_R = \{\text{rich, poor}\}$$

Now, when:

$$P(H = \text{happy}, R = \text{rich}) > 0 \textbf{ and } P(H = \text{unhappy}, R = \text{rich}) > 0$$

and:

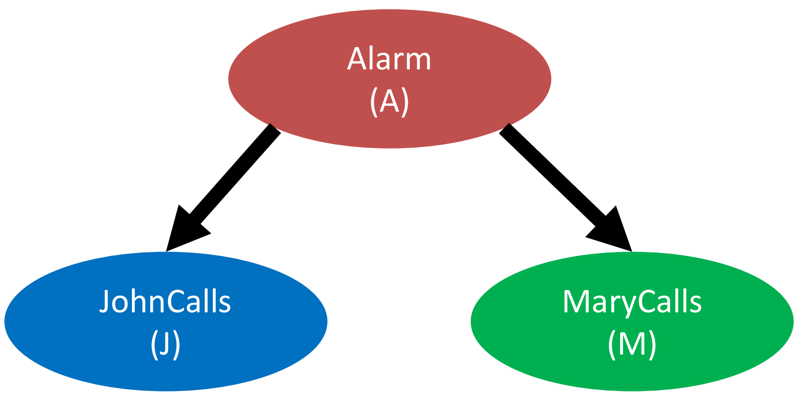
$$P(P = \text{powerful} \mid H = \text{happy}, R = \text{rich}) = P(P = \text{powerful} \mid H = \text{unhappy}, R = \text{rich})$$

In other words, given a value of R, knowing Y's value DOES NOT affect your belief in the value of X.

"Being un/happy does not make you less powerful, if you are rich."

More On Conditional Independence

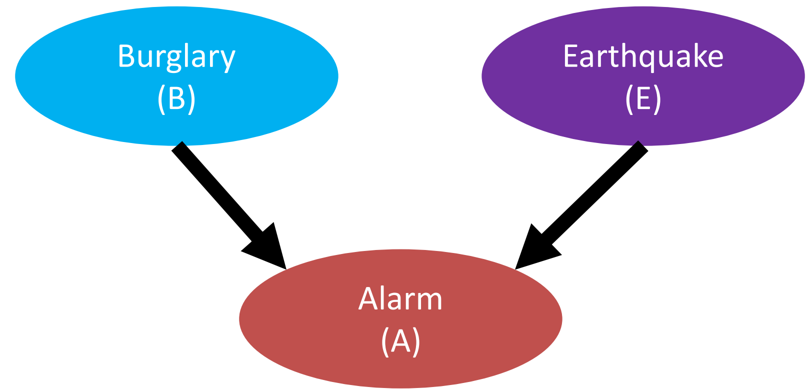
Common Cause:



JohnCalls and MaryCalls
are NOT independent

JohnCalls and MaryCalls are CONDITIONALLY independent given Alarm

Common Effect:



Burglary and Earthquake
are independent

Burglary and Earthquake are NOT CONDITIONALLY independent given Alarm

Bayes' Rule

$$P(A \mid B) = \frac{P(B \mid A) * P(A)}{P(B)}$$

Bayes' Rule

$$P(A \mid B) = \frac{P(B \mid A) * P(A)}{P(B)}$$

Bayes' Rule

$$P(cause \mid effect) = \frac{P(effect \mid cause) * P(cause)}{P(effect)}$$

Bayes' Rule

$P(cause \mid effect)$ **diagnostic direction relation**

$$P(cause \mid effect) = \frac{P(effect \mid cause) * P(cause)}{P(effect)}$$

$P(effect \mid cause)$ **causal direction relation**

Bayes' Rule

$P(disease \mid symptoms)$ **diagnostic direction relation**

$$P(disease \mid symptoms) = \frac{P(symptoms \mid disease) * P(disease)}{P(symptoms)}$$

$P(symptoms \mid disease)$ **causal direction relation**

Bayes' Rule

Why is this useful?

- **Because in practice it is easier to get probabilities for** P(effect|cause) **and** P(cause) **than for** P(cause|effect)

$$P(disease \mid symptoms) = \frac{P(symptoms \mid disease) * P(disease)}{P(symptoms)}$$

  - **It is easier to know what symptoms diseases cause. It is harder to diagnose a disease given symptoms**

Bayes' Rule

$$P(cause \mid effect) = \frac{P(effect \mid cause) * P(cause)}{P(effect)}$$

Problem: a single card is drawn from a standard deck of cards. What is the probability that we drew a queen if we know that a face card (J, Q, K) was drawn?

$$P(queen \mid face) = \frac{P(face \mid queen) * P(queen)}{P(face)}$$

$$P(queen \mid face) = \frac{1 * 4 / 52}{12 / 52} = \frac{1}{3}$$

Bayes' Rule

$$P(cause \mid effect) = \frac{P(effect \mid cause) * P(cause)}{P(effect)}$$

Problem: Calculate probability that a patient has meningitis if a patient has stiff neck. Meningitis is a cause of neck stiffness in 70% of cases, probability of having meningitis is 1/50000. Stiff neck happens to 1% of patients.

$$P(m \mid s) = \frac{P(s \mid m) * P(m)}{P(s)}$$

$$P(m \mid s) = \frac{0.7 * 1/50000}{0.01} = 0.0014$$

Bayes' Rule: Another Interpretation

Another way to think about Baye's rule: it allows us to update the hypothesis H in light of some new data/evidence e.

$$P(H \mid e) = \frac{P(e \mid H) * P(H)}{P(e)}$$

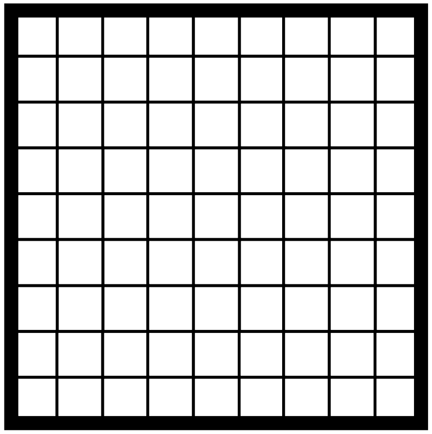
$$P(Hypothesis \mid evidence) = \frac{P(evidence \mid Hypothesis) * P(Hypothesis)}{P(evidence)}$$

where:

- $P(H)$ **- probability of the** Hypothesis H **being true BEFORE we see new data/**evidence e **(<u>prior</u> probability)**
- $P(H \mid e)$ **- probability of the** Hypothesis H **being true AFTER we see new data/**evidence e **(<u>posterior</u> probability)**
- $P(e \mid H)$ **- probability of new data/**evidence e **being true under the** Hypothesis H **(<u>likelihood</u>)**
- $P(e)$ **- probability of new data/**evidence e **being true under ANY hypothesis (<u>normalizing constant</u>)**

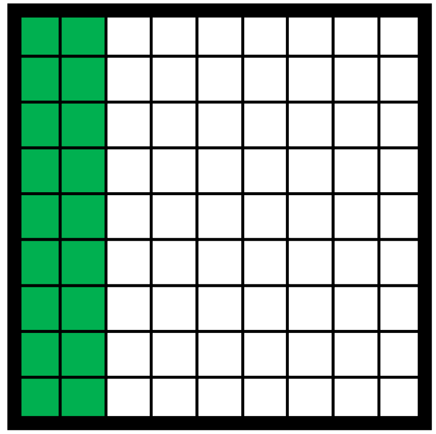
Bayes' Rule: Visual Interpretation

All possible cases



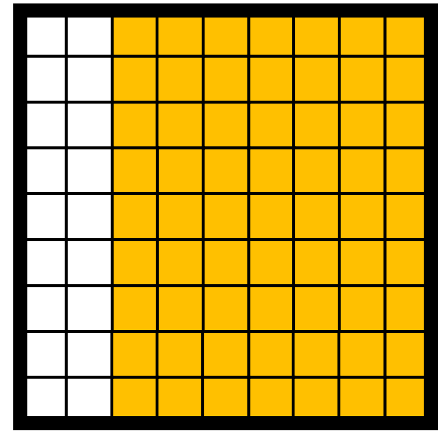
Cases where Hypothesis H **is true**

P(H)



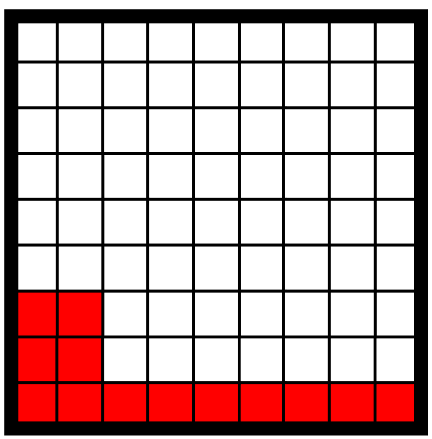
Cases where Hypothesis H **is false**

P(¬H)

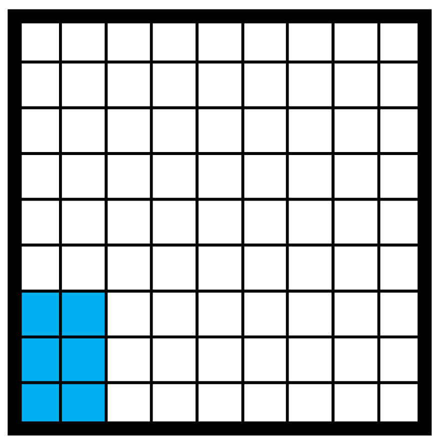


Cases where evidence e **is true**

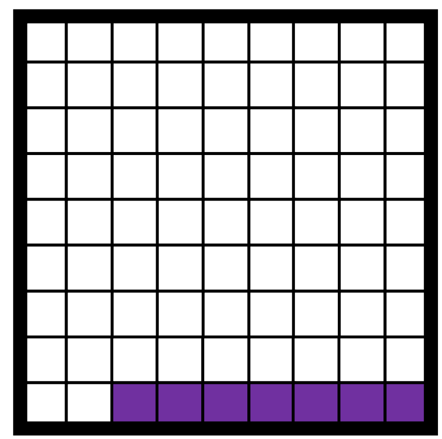
P(e)



Cases where evidence e **is true given** Hypothesis H **true** P(e | H)



Cases where evidence e **is true given** Hypothesis H **false** P(e | ¬H)



Bayes' Rule: Visual Interpretation

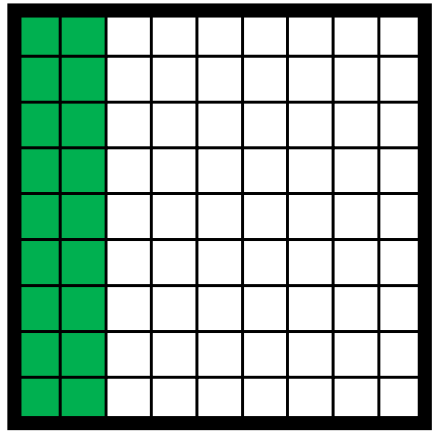
Bayes' Rule:

$$P(H \mid e) = \frac{P(e \mid H) * P(H)}{P(e)}$$

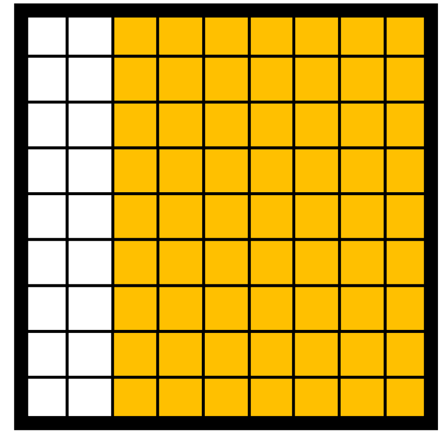
$$P(H \mid e) = \frac{P(e \mid H) * P(H)}{P(e)}$$

$$P(H \mid e) = \frac{P(e \mid H) * P(H)}{P(H) * P(e \mid H) + P(\neg H) * P(e \mid \neg H)}$$

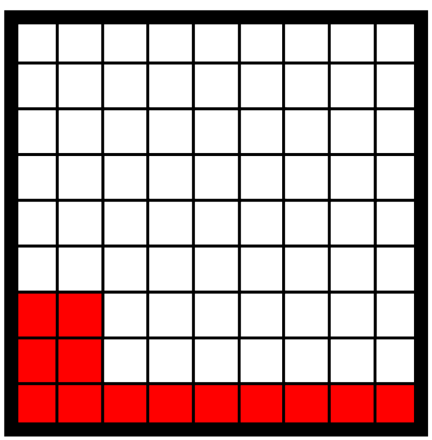
Cases where Hypothesis H **is true**
P(H)



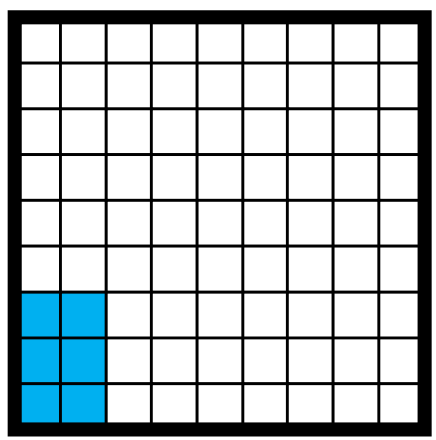
Cases where Hypothesis H **is false**
P(¬H)



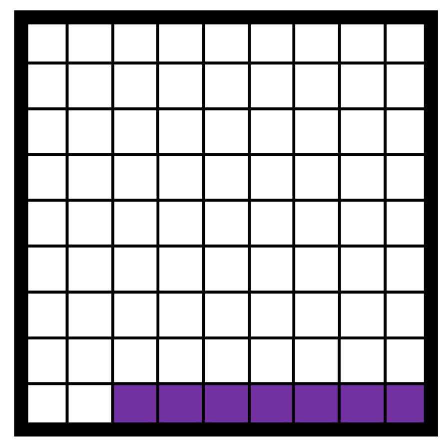
Cases where evidence e **is true**
P(e)



Cases where evidence e **is true given** Hypothesis H **true** P(e | H)



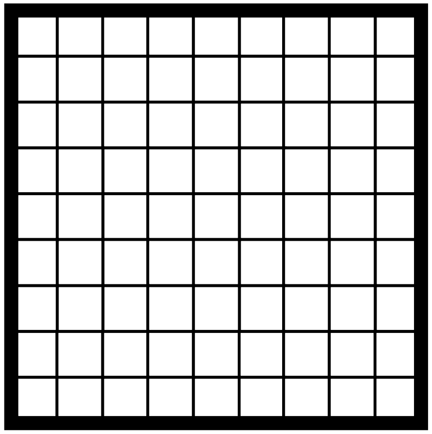
Cases where evidence e **is true given** Hypothesis H **false** P(e | ¬H)



Bayes' Rule: Visual Interpretation

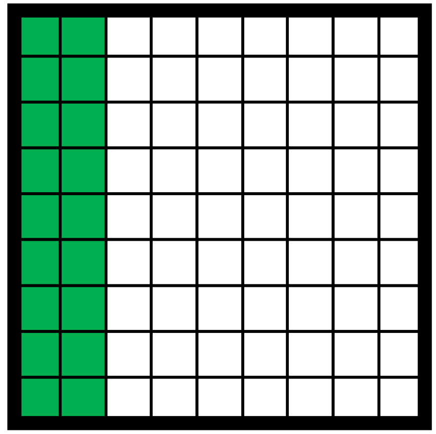
All Students

Hypothesis H: **graduate student**



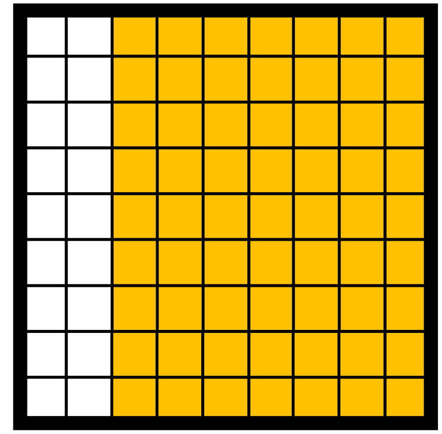
Cases where Hypothesis H **is true**

$P(H) = P(grad = true)$



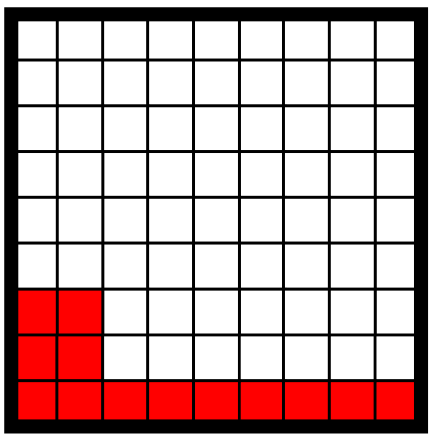
Cases where Hypothesis H **is false**

$P(\neg H) = P(grad = false)$



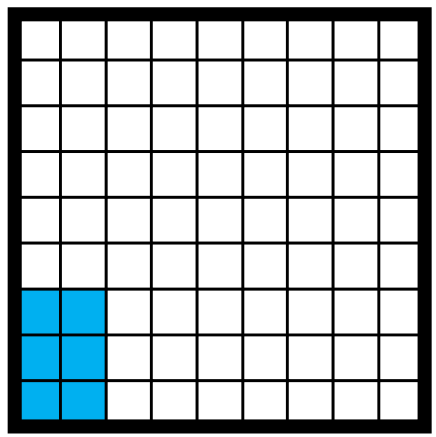
Cases where evidence e **is true**

$P(e) = P(female = true)$



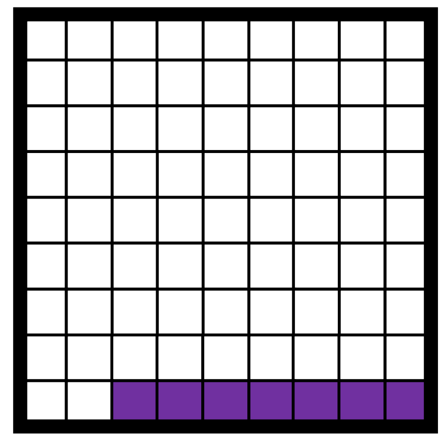
Cases where e **true given** H **true**

$P(e \mid H) = P(female = true \mid grad = true)$



Cases where e **true given** H **false**

$P(e \mid \neg H) = P(female = true \mid grad = false)$



Bayes' Rule: Visual Interpretation

Given (<u>made up</u> roster data):

%of G students: P(H)

%of UG students: P(¬H)

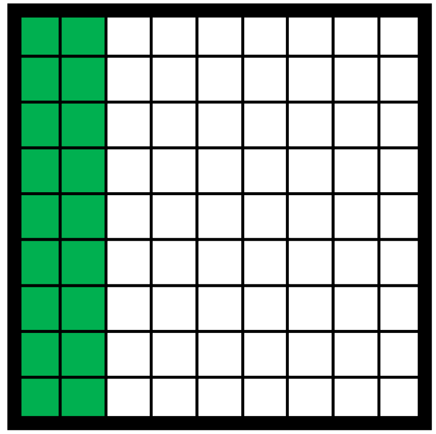
%of female students: P(e)

%of female G students: P(e | H)

%of female UG students: P(e | ¬H)

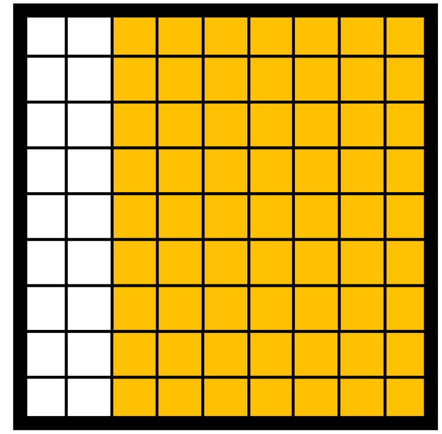
Cases where Hypothesis H **is true**

$P(H) = 18 / 81$



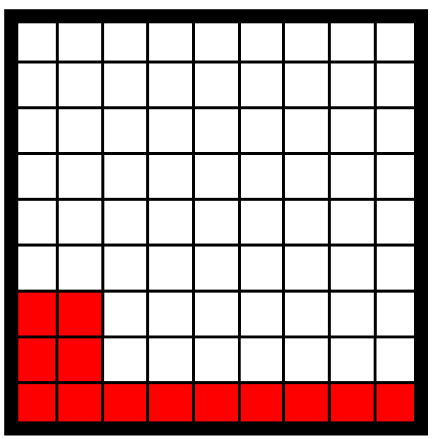
Cases where Hypothesis H **is false**

$P(\neg H) = 63 / 81$



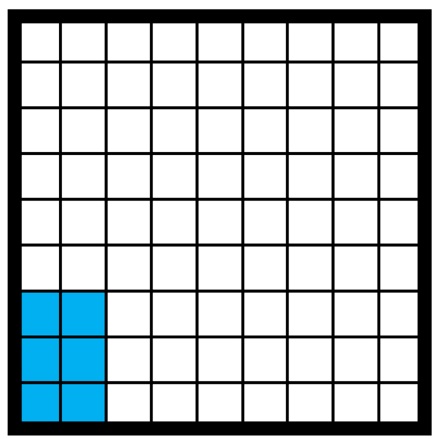
Cases where evidence e **is true**

$P(e) = 13 / 81$



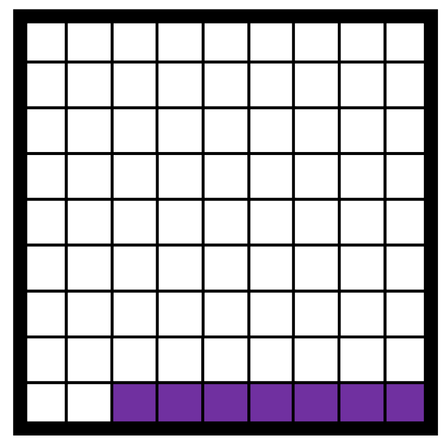
Cases where e **true given** H **true**

$P(e \mid H) = 6 / 18$



Cases where e **true given** H **false**

$P(e \mid \neg H) = 7 / 63$



Bayes' Rule: Visual Interpretation

Bayes' Rule:

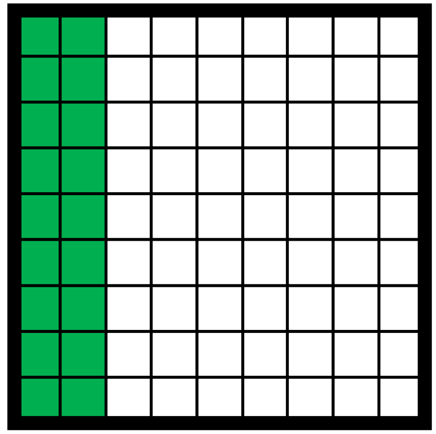
$$P(H \mid e) = \frac{P(e \mid H) * P(H)}{P(e)}$$

$$P(H \mid e) = \frac{P(e \mid H) * P(H)}{P(e)}$$

$$P(H \mid e) = \frac{P(e \mid H) * P(H)}{P(H) * P(e \mid H) + P(\neg H) * P(e \mid \neg H)}$$

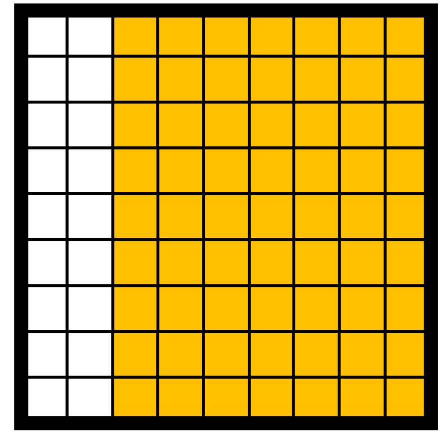
Cases where Hypothesis H **is true**

P(H) = 18 / 81



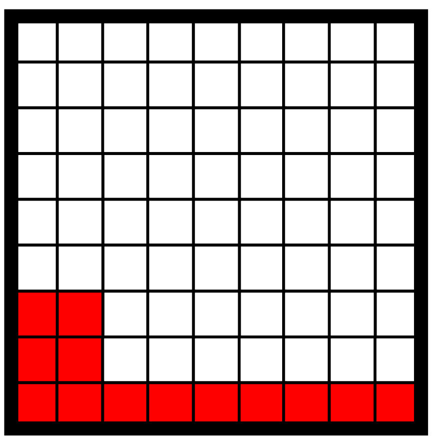
Cases where Hypothesis H **is false**

P(¬H) = 63 / 81



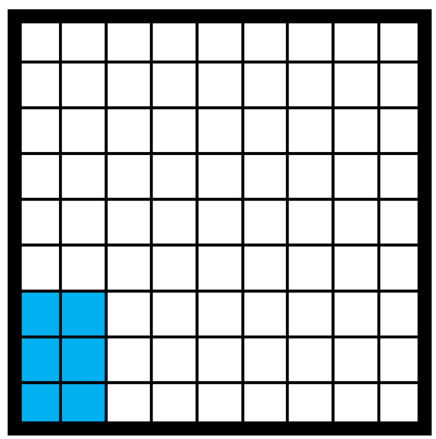
Cases where evidence e **is true**

P(e) = 13 / 81



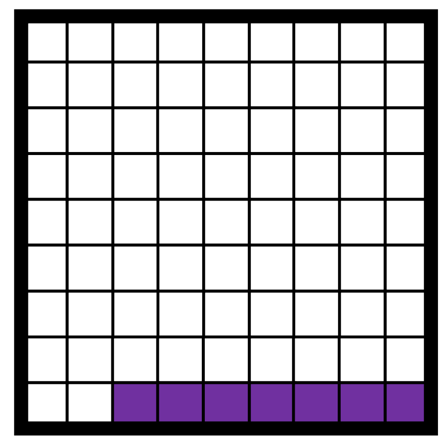
Cases where e **true given** H **true**

P(e | H) = 6 / 18



Cases where e **true given** H **false**

P(e | ¬H) = 7 / 63



Bayes' Rule: Visual Interpretation

Bayes' Rule:

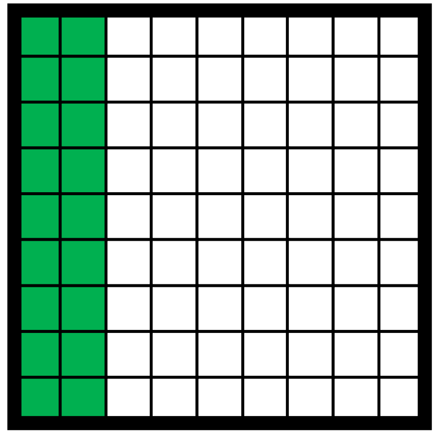
$$P(H \mid e) = \frac{P(e \mid H) * P(H)}{P(e)}$$

$$P(H \mid e) = \frac{6/18 * 18/81}{13/81}$$

$$P(H \mid e) = \frac{6/18 * 18/81}{18/81 * 6/18 + 63/81 * 7/63}$$

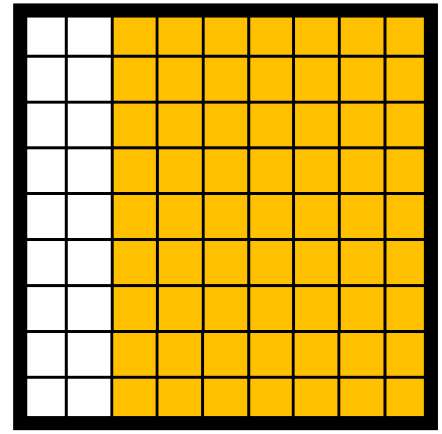
Cases where Hypothesis H **is true**

P(H) = 18 / 81



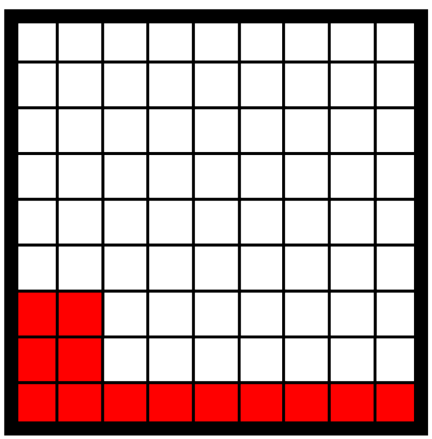
Cases where Hypothesis H **is false**

P(¬H) = 63 / 81



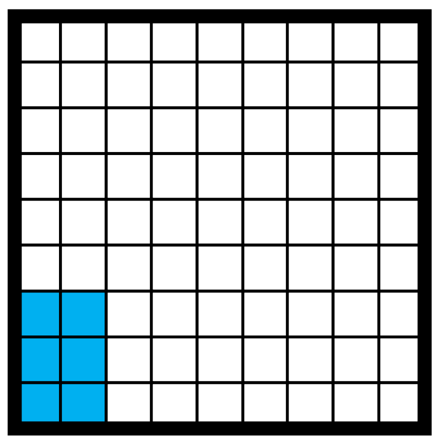
Cases where evidence e **is true**

P(e) = 13 / 81



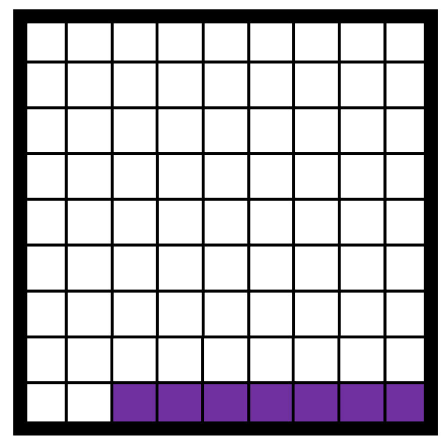
Cases where e **true given** H **true**

P(e | H) = 6 / 18



Cases where e **true given** H **false**

P(e | ¬H) = 7 / 63



Bayes' Rule: Visual Interpretation

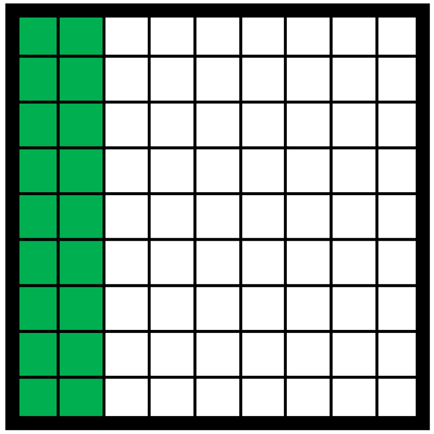
Bayes' Rule:

$$P(H \mid e) = \frac{P(e \mid H) * P(H)}{P(e)}$$

$$P(H \mid e) \approx 0.462$$

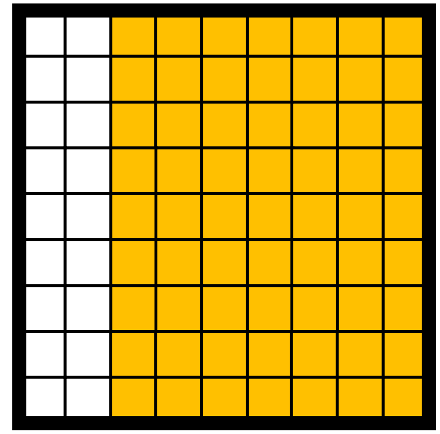
Cases where Hypothesis H **is true**

P(H) = 18 / 81



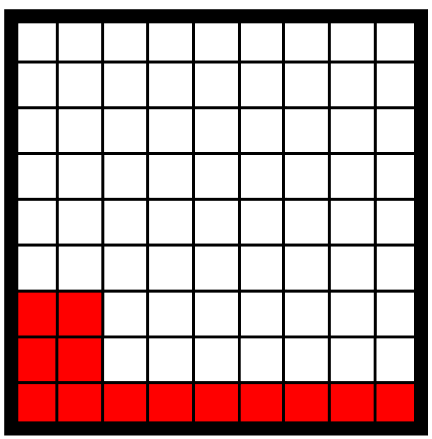
Cases where Hypothesis H **is false**

P(¬H) = 63 / 81



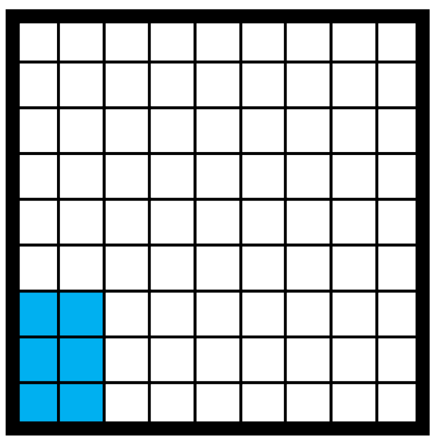
Cases where evidence e **is true**

P(e) = 13 / 81



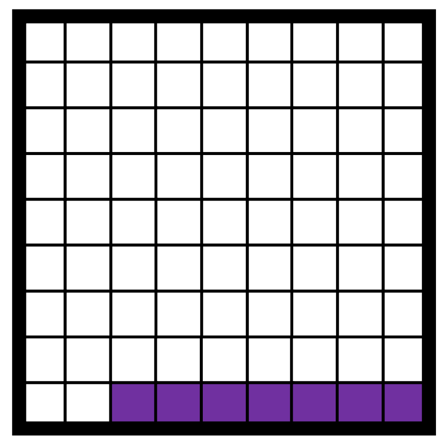
Cases where e **true given** H **true**

P(e | H) = 6 / 18



Cases where e **true given** H **false**

P(e | ¬H) = 7 / 63



Bayes' Rule: Visual Interpretation

Prior probability:

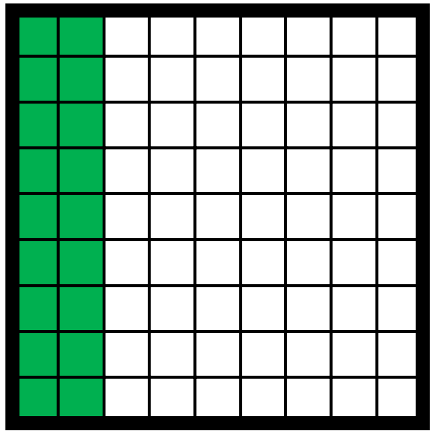
$P(H) = 18 / 81 \approx 0.222$

Posterior probability:

$P(H \mid e) \approx 0.462$

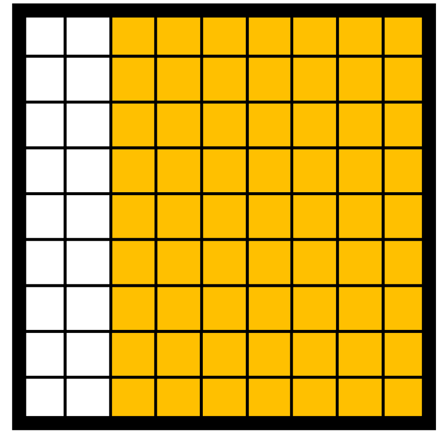
Cases where Hypothesis H is true

P(H) = 18 / 81



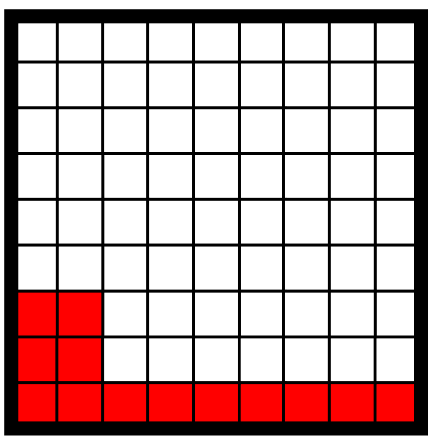
Cases where Hypothesis H is false

P(¬H) = 63 / 81



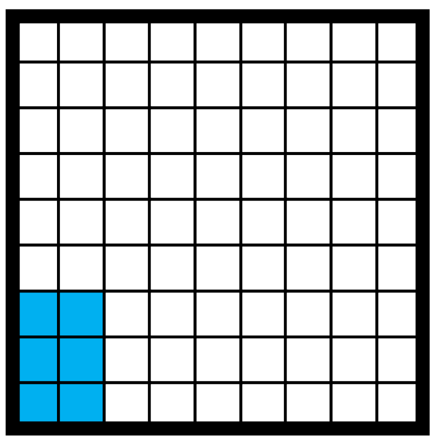
Cases where evidence e is true

P(e) = 13 / 81



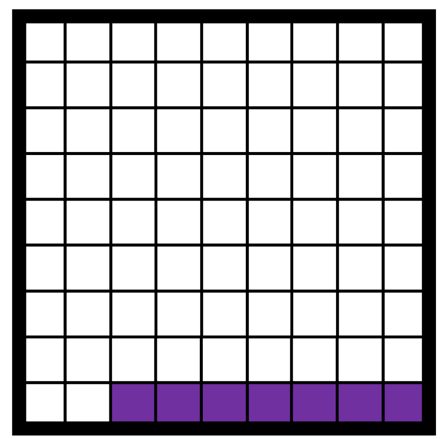
Cases where e true given H true

P(e | H) = 6 / 18



Cases where e true given H false

P(e | ¬H) = 7 / 63



Bayes' Rule: Belief/Probability Update

A student approaches the podium. Without looking I create a hypothesis H**:**

this is a grad student ($\mathrm{grad} = \mathrm{true}$)

My <u>belief</u> in H **being true is based on prior probability:**

$$P(H) = 18\,/\,81 \approx 0.222$$

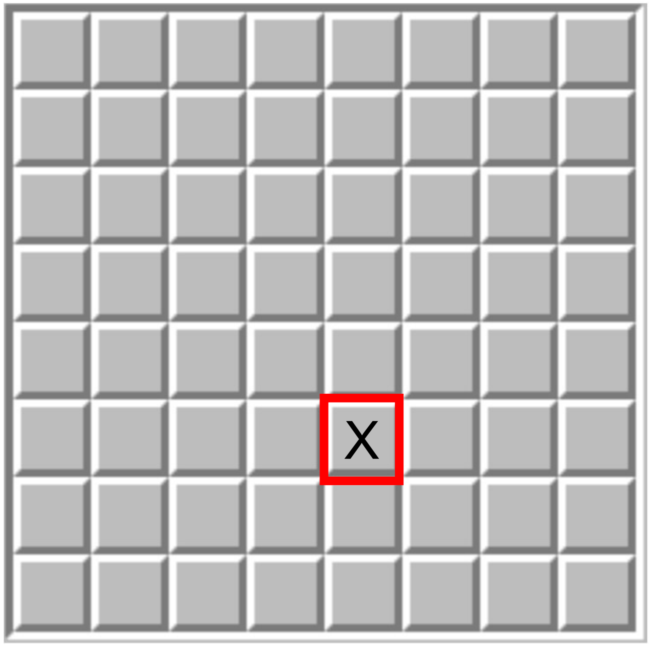
I look up and see a female student, which is <u>new data / evidence</u> e ($\mathrm{female} = \mathrm{true}$)**. Bayes' Rule helps me update my <u>belief</u> in** H **being true with <u>posterior</u> probability:**

$$P(H \mid e) = \frac{6\,/\,18 * 18\,/\,81}{18\,/\,81 * 6\,/\,18 + 63\,/\,81 * 7\,/\,63} \approx 0.462$$

Playing Minesweeper with Bayes' Rule

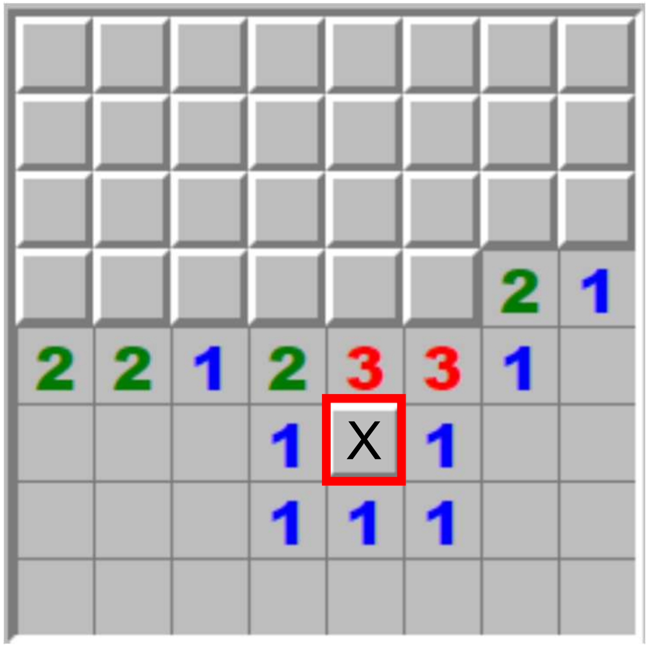
Prior probability / belief:

$$P(X = mine) = 0.5$$



Posterior probability / belief:

$$P(X = mine \mid \boldsymbol{evidence}) = 1.0$$



Marginal Probability

Marginal probability: the probability of an event occurring $P(A)$.

It may be thought of as an unconditional probability.

It is not conditioned on another event.

Full Joint Probability Distribution

H: grad	e: female	P(H, e) = P(H ∧ e): P(grad ∧ female)	Conditional probabilities
true	true	$P(H \mid e)*P(e) \approx 0.074$	$P(H \mid e) = \frac{P(e \mid H) * P(H)}{P(e)} = \frac{6/18 * 18/81}{13/81} \approx 0.462$
true	false	$P(H \mid \neg e)*P(\neg e) \approx 0.148$	$P(H \mid \neg e) = \frac{P(\neg e \mid H) * P(H)}{P(\neg e)} = \frac{12/18 * 18/81}{68/81} \approx 0.176$
false	true	$P(\neg H \mid e)*P(e) \approx 0.086$	$P(\neg H \mid e) = \frac{P(e \mid \neg H) * P(\neg H)}{P(e)} = \frac{7/63 * 63/81}{13/81} \approx 0.538$
false	false	$P(\neg H \mid \neg e)*P(\neg e) \approx 0.691$	$P(\neg H \mid \neg e) = \frac{P(\neg e \mid \neg H) * P(\neg H)}{P(\neg e)} = \frac{56/63 * 63/81}{68/81} \approx 0.824$
		SUM = 1	

Joint probabilities calculated using the Product Rule:

$$P(A \wedge B) = P(A \mid B) * P(B)$$

Conditional probabilities calculated using Bayes' Rule:

$$P(A \mid B) = \frac{P(B \mid A) * P(A)}{P(B)}$$

Joint Probability Distribution

H: grad	e: female	P(H, e) = P(H ∧ e): P(grad ∧ female)
true	true	$P(grad = true \wedge female = true) = P(H, e) = P(H \wedge e) = P(H \mid e) * P(e) \approx 0.074$
true	false	$P(grad = true \wedge female = false) = P(H, \neg e) = P(H \mid \neg e) * P(\neg e) \approx 0.148$
false	true	$P(grad = false \wedge female = true) = P(\neg H, e) = P(\neg H \mid e) * P(e) \approx 0.086$
false	false	$P(grad = false \wedge female = false) = P(\neg H, \neg e) = P(\neg H \mid \neg e) * P(\; e) \approx 0.691$
		SUM = 1

Joint probabilities calculated using the Product Rule:

$$P(A \wedge B) = P(A \mid B) * P(B)$$

Conditional probabilities calculated using Bayes' Rule:

$$P(A \mid B) = \frac{P(B \mid A) * P(A)}{P(B)}$$

Joint Probability Distribution

H: grad	e: female	$P(H, e) = P(H \wedge e)$: $P(grad \wedge female)$
true	true	**0.074**
true	false	**0.148**
false	true	**0.086**
false	false	**0.691**
		SUM = 1

If we know the joint probability distribution, we can infer:

- **marginal probabilities** $P(H)$, $P(\neg H)$, $P(e)$, **and** $P(\neg e)$
- **conditional probabilities** $P(H \mid e)$, $P(H \mid \neg e)$, $P(\neg H \mid e)$, **and** $P(\neg H \mid \neg e)$

Joint Probability: Marginalization

H: grad	e: female	P(H, e) = P(H ∧ e): P(grad ∧ female)
true	true	**0.074**
true	false	**0.148**
false	true	**0.086**
false	false	**0.691**
		SUM = 1

Probability P(H)**:**

$$P(H) = P(grad = true) = 0.074 + 0.148 \approx 18 / 81$$

Probability P(H)**: "sum of all probabilities where** H **true"**

Joint Probability: Marginalization

H: grad	e: female	$P(H, e) = P(H \wedge e)$: $P(grad \wedge female)$
true	true	**0.074**
true	false	**0.148**
false	true	**0.086**
false	false	**0.691**
		SUM = 1

Probability $P(e)$**:**

$$P(e) = P(female = true) = 0.074 + 0.086 \approx 13 / 81$$

Probability $P(e)$**: "sum of all probabilities where** e **true"**

Joint Probability Distribution

H: grad	e: female	$P(H, e) = P(H \wedge e)$: $P(grad \wedge female)$
true	true	**0.074**
true	false	**0.148**
false	true	**0.086**
false	false	**0.691**
		SUM = 1

Joint probabilities calculated using the Product Rule:

$$P(A \wedge B) = P(A \mid B) * P(B)$$

Conditional probabilities calculated using Bayes' Rule:

$$P(A \mid B) = \frac{P(B \mid A) * P(A)}{P(B)}$$

Joint Probability: Conditionals

H: grad	e: female	P(H, e) = P(H ∧ e): P(grad ∧ female)
true	true	**0.074**
true	false	**0.148**
false	true	**0.086**
false	false	**0.691**
		SUM = 1

From product rule:

$$P(H \wedge e) = P(H \mid e) * P(e)$$

we can derive:

$$P(H \mid e) = \frac{P(H \wedge e)}{P(e)}$$

Joint Probability: Conditionals

H: grad	e: female	P(H, e) = P(H ∧ e): P(grad ∧ female)
true	true	0.074
true	false	0.148
false	true	0.086
false	false	0.691
		SUM = 1

From product rule:

$$P(H \wedge e) = P(H \mid e) * P(e)$$

we can derive:

$$P(H \mid e) = \frac{P(H \wedge e)}{P(e)} = \frac{0.074}{0.074 + 0.086} \approx 0.462$$

Full Joint Probability Distribution

H: grad	e: female	P(H, e) = P(H ∧ e): P(grad ∧ female)
true	true	**0.074**
true	false	**0.148**
false	true	**0.086**
false	false	**0.691**
		SUM = 1

Joint probabilities calculated using the Product Rule:

$$P(A \wedge B) = P(A \mid B) * P(B)$$

Conditional probabilities calculated using Bayes' Rule:

$$P(A \mid B) = \frac{P(B \mid A) * P(A)}{P(B)}$$

Number of Parameters

- **Assuming everything is binary**
- $P(V_1)$ **requires**
  - **1 independent parameter**
- $P(V_1, V_2, \ldots, V_n)$ **requires**
  - **2^n-1 independent parameters**
- $P(V_1|V_2)$ **requires**
  - **2 independent parameters**
- $P(V_1, V_2, \ldots, V_n \mid V_{n+1}, V_{n+2}, \ldots, V_{n+m})$ **requires**
  - **$2^m \times (2^n\text{-}1)$ independent parameters**

Continuous RV

We define **probability density function**, p(x), a non-negative integrable function, such that

$$P(X \leq a) = \int_{-\infty}^{a} p(x)dx$$

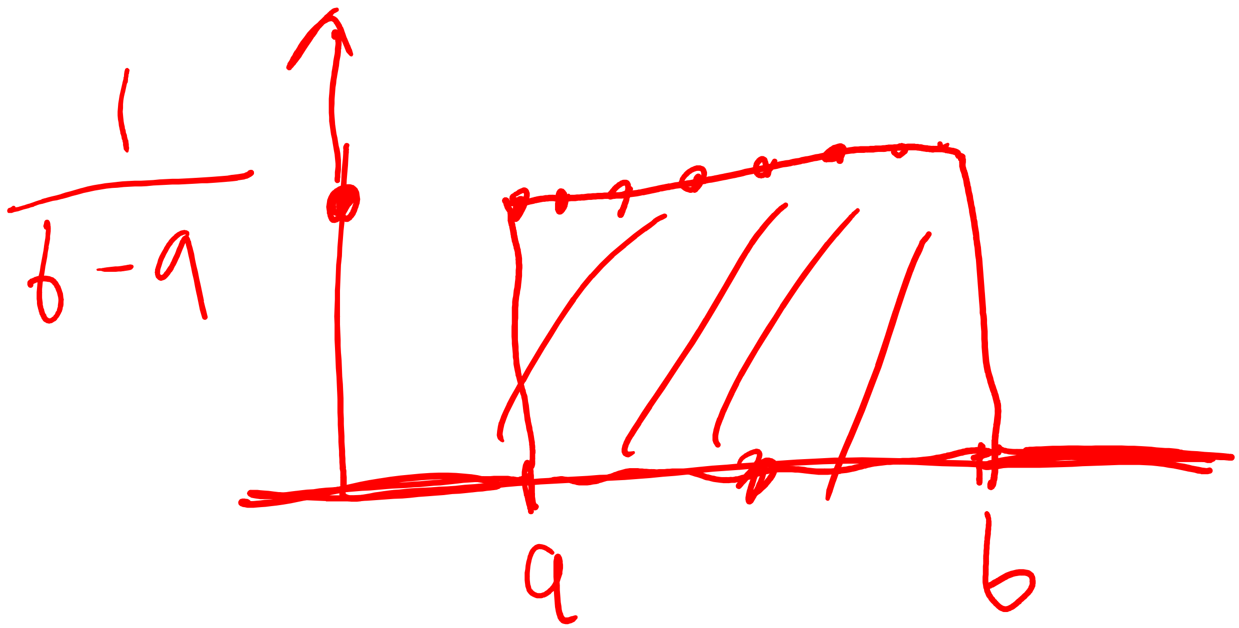
$$P(a \leq X \leq b) = \int_{a}^{b} p(x)dx$$

$$\int_{Val(X)} p(x)dx = 1$$

Uniform Distribution

A variable X has a **uniform distribution** over $[a,b]$ if it has the PDF

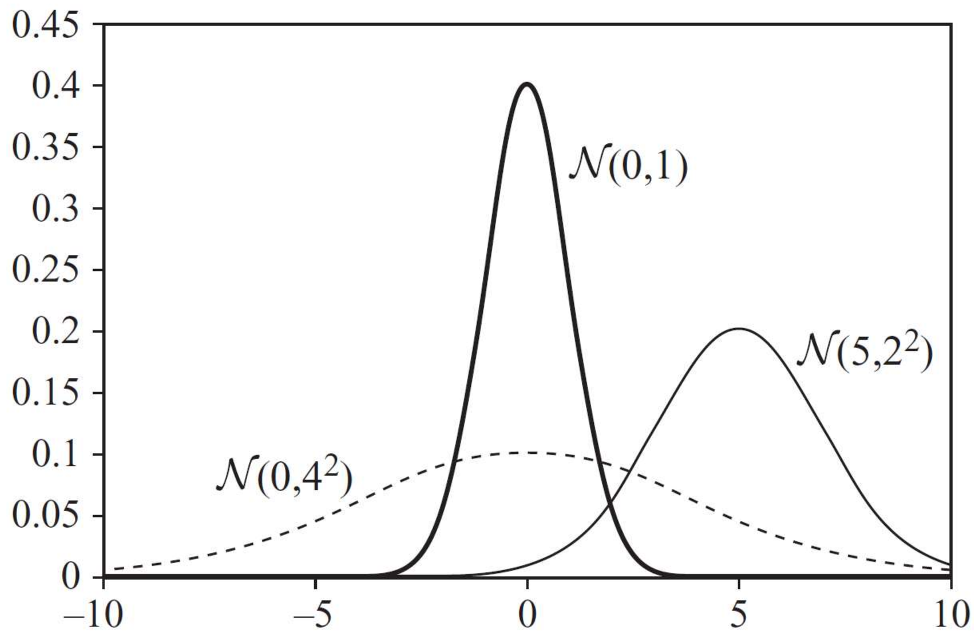
$$p(x) = \begin{cases} \dfrac{1}{b-a} & a \le x \le b \\ 0 & otherwise \end{cases}$$



Gaussian Distribution

A variable X has a Gaussian distribution with mean μ and variance σ^2, if it has the PDF

$$p(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Expectation

$$E_P[X] = \sum_x xP(x)$$

$$E_P[X] = \int_x xp(x)dx$$

$$E_P[aX + b] = aE_P[X] + b$$

$$E_P[X + Y] = E_P[X] + E_P[Y]$$

$$E_P[X \mid y] = \sum_x xP(x \mid y)$$

Variance

$$Var_P[X] = E_P\left[\left(X - E_P[X]\right)^2\right]$$

$$Var_P[X] = E_P\left[X^2\right] - \left(E_P[X]\right)^2$$

$$Var_P[aX + b] = a^2 Var_P[X]$$