Advanced Data Science

Topic 11b – Part 4

1. What We'll Cover

This topic will introduce…

- What is data science.
- Key concepts – the scientific method.
- Useful terminology.

- Important tools - Statistics.
- Data collection & Experiment Design.

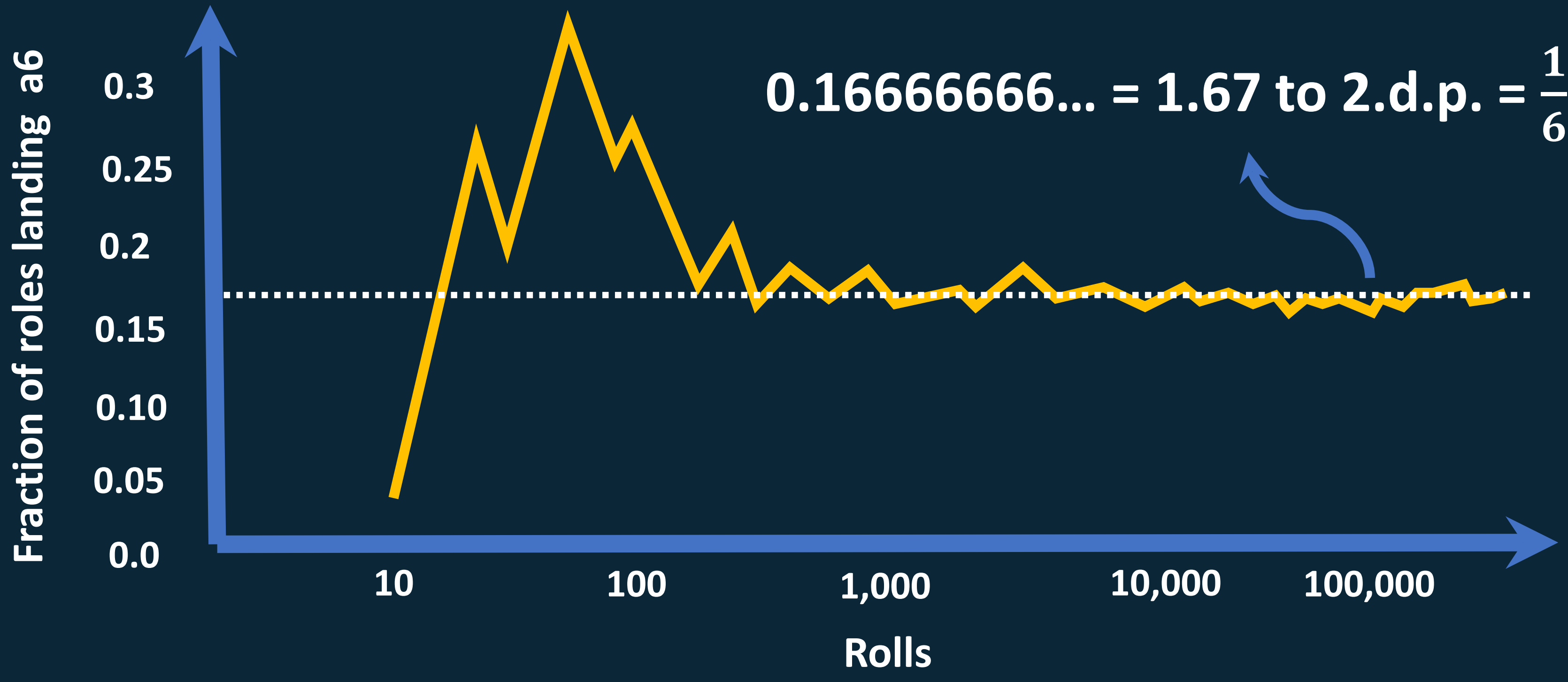
- Probability basics.
- Data distributions.
- Hypothesis testing.

} Part 4

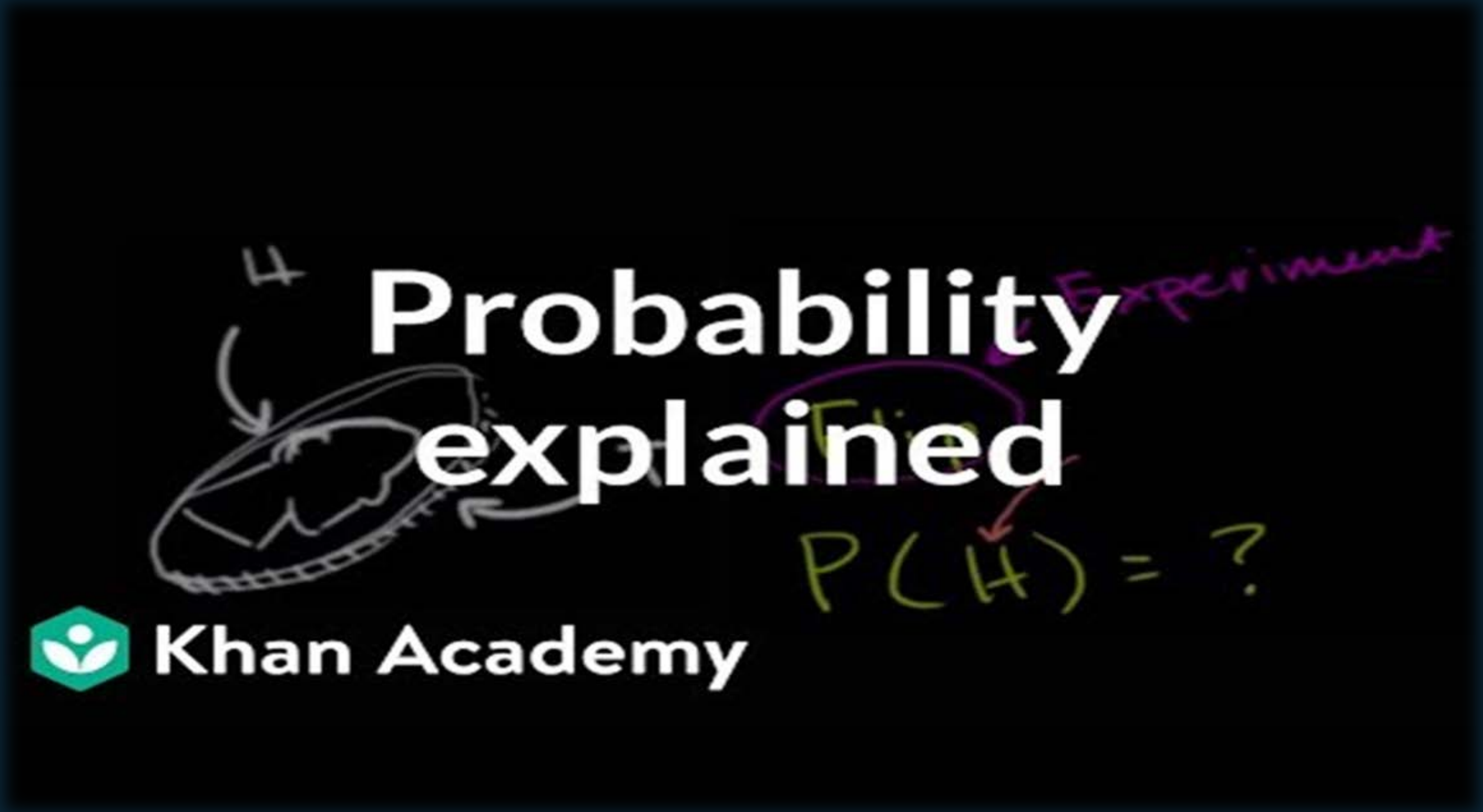
The aim: to help you understand what it means to be a data scientist and to get you familiar with data science tools.

2. Probability

- Studying probability helps us understand randomness in data.
- We often think about randomness in terms of random variables.
- To understand probability we first think a little about it's nature.
- Just because something is probable, does not mean it will happen….



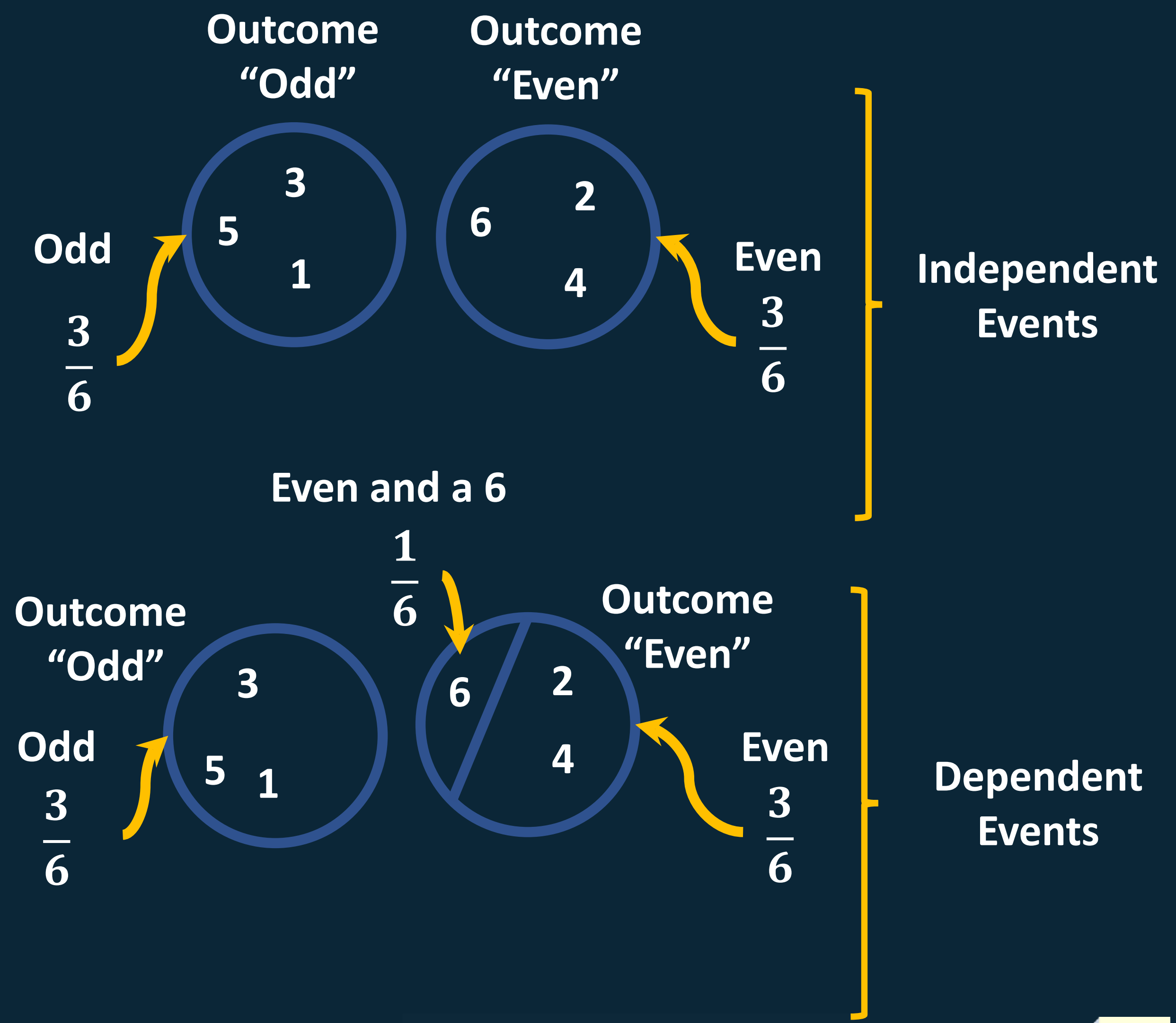
3. Probability Basics



Credit: Khan Academy

4. Disjoint or Mutually Exclusive

- During our first die experiment, we considered a specific type of probabilistic event.
- Each trial outcome was an **independent event**.
- Another way to say this - each event is **mutually exclusive** which means **disjoint** – they cannot happen a the same time.
- For instance I cannot get an odd and an even number during a dice roll – there is only one die and therefore 1 disjoint outcome.
- The outcome “roll a six and get an even number” are not disjoint events – these can both happen with a roll of the dice.



5. Probability Notation

- We express probabilities using notation.
- At first this notation can appear confusing. I promise it is relatively straightforward - notation is just representing numbers that you can do basic math with.
- An upper case P can be read as "the Probability".
- Usually probability refers to some outcome or event. To show this, you may use an index such as i.
- We can see that probabilities can be described as fractions, or as decimal numbers.
- Sometimes we may see probabilities written like $P(Event)$. Here "Event" is a placeholder for any event we can think off.
- I'll use this notation from now on, as it's easier for teaching.

P means the probability.

↓

P_i means the probability of event i occurring.

↓

If $P_{i=6}$ is probability of rolling a 6 then,

$$P_6 = \frac{1}{6} = 0.166666 \ldots$$

↓

$P(\ Event)$ means the probability of "Event" occurring.

↓

Teaching Notation → $P(\ Roll\ a\ 6) = \frac{1}{6} = 0.16666 \ldots$

6. Probability Notation – Not (¬)

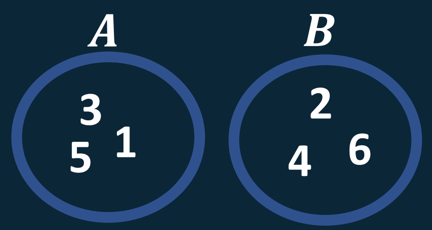
- We may also use variables to represent probabilities or events – makes the notation easier to interpret.
- For example, we could use the letter A to represent "Roll a odd number".
- Using variables makes things more concise.
- There are operators you should know about too.
- First there is the "not" operator represented by the symbol: ¬.
- Not is used to negate the probability of an event happening.
- The sum of probabilities for all events should always add up to 1.

$$P(Roll\ a\ odd\ number) = \frac{3}{6} = 0.5 \quad \text{vs.} \quad P(A) = \frac{3}{6} = 0.5$$

If A = prob. of rolling an odd number

and B = prob. of rolling an even number

Probability of A $= \frac{3}{6} = \frac{1}{2} = 0.5 = 50\%$



Then $\neg A$ = probability of not rolling an odd number

$$P(\neg A) = \frac{3}{6} = 0.5$$

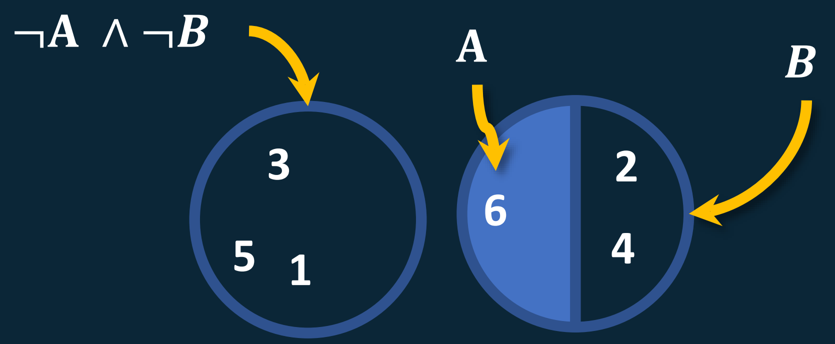
It follows that $\neg A = B$ and thus $\neg B = A$

7. Probability Notation – AND & OR

- For dependent events, we can describe the probability of all outcomes occurring, or just one of many occurring.
- We can do this using notation for logical AND (∧). & logical OR (∨).
- For example, we can define the probability of rolling a 6 and an even number, or the probability of rolling a number divisible by 2 or 3.

If A = probability of rolling a 6 and

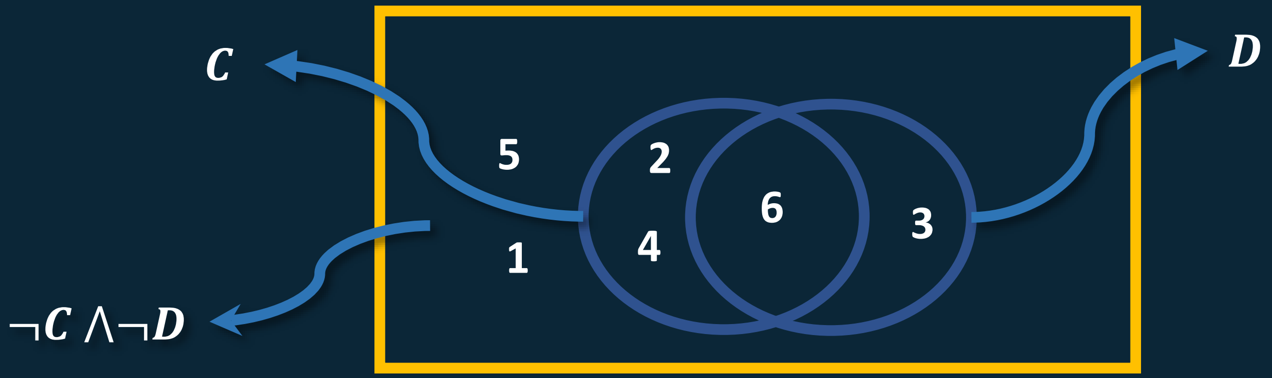
B = probability of rolling an even number



$$P(A \wedge B) = \frac{1}{6}$$

If C = probability of rolling a number divisible by 2

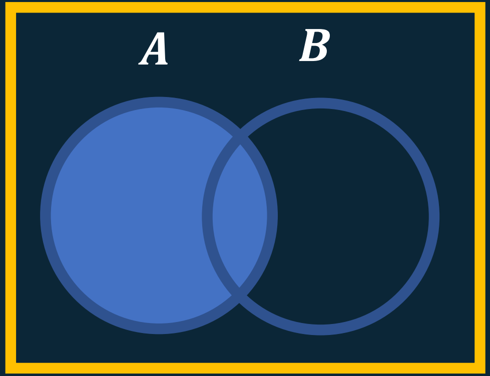
D = probability of rolling a number divisible by 3



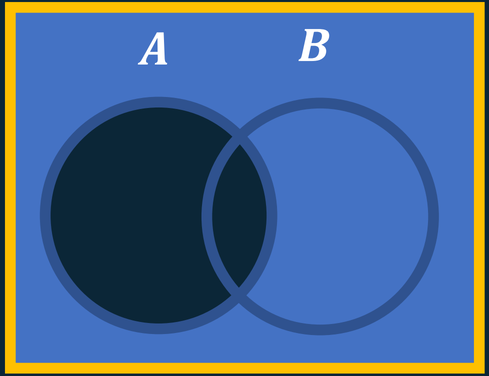
$$P(C \vee D) = \frac{4}{6}$$

8. Notation Recap – Dependent Events

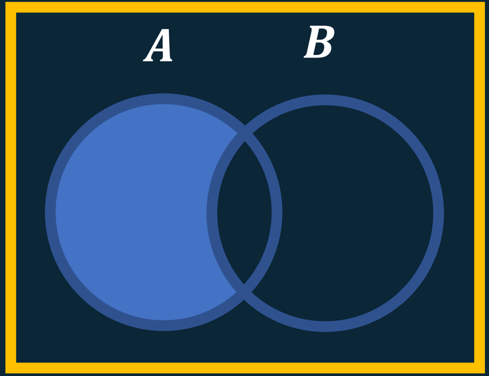
$P(A)$



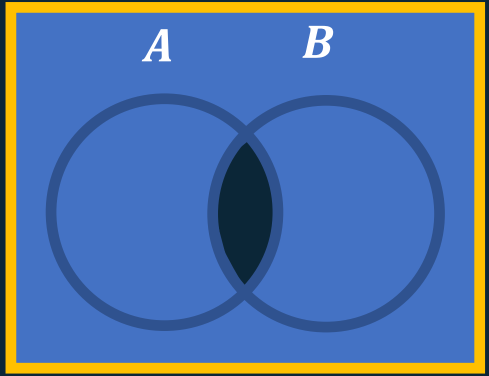
$P(\neg A)$



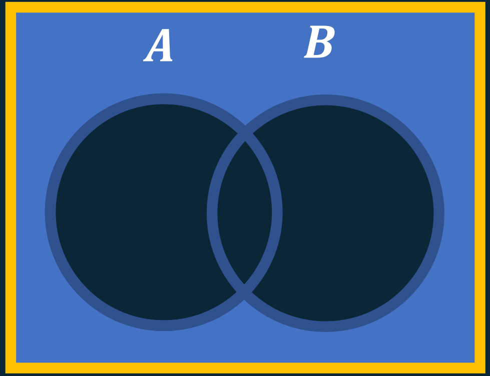
$P(A \wedge \neg B)$



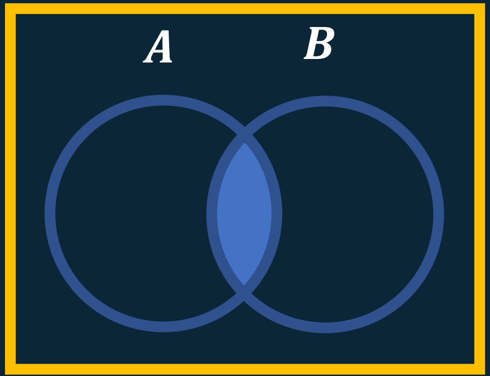
$P(\neg A \vee \neg B)$



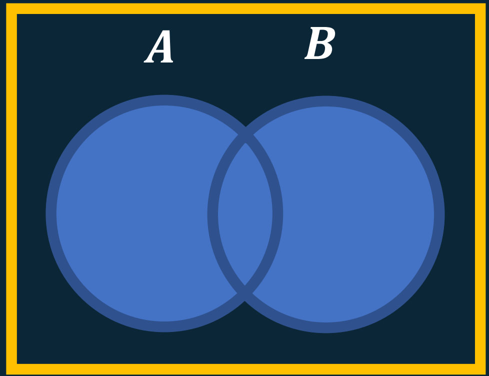
$P(\neg A \wedge \neg B)$



$P(A \wedge B)$

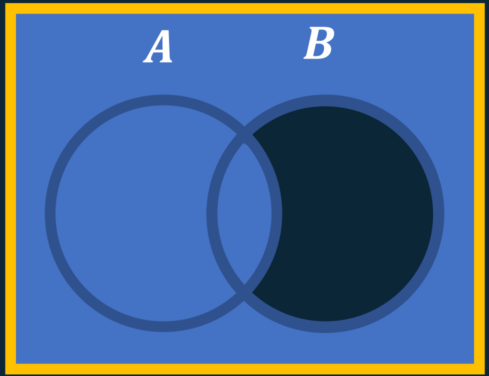


$P(A \vee B)$

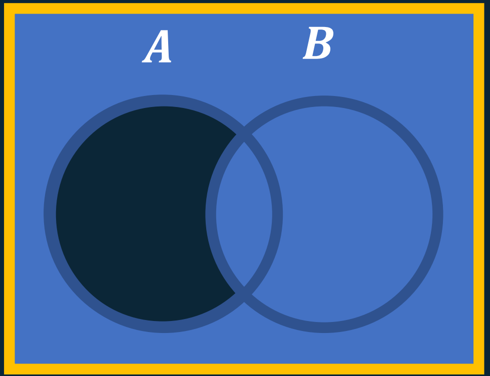


Shaded regions represent events that occur – events here are not independent!

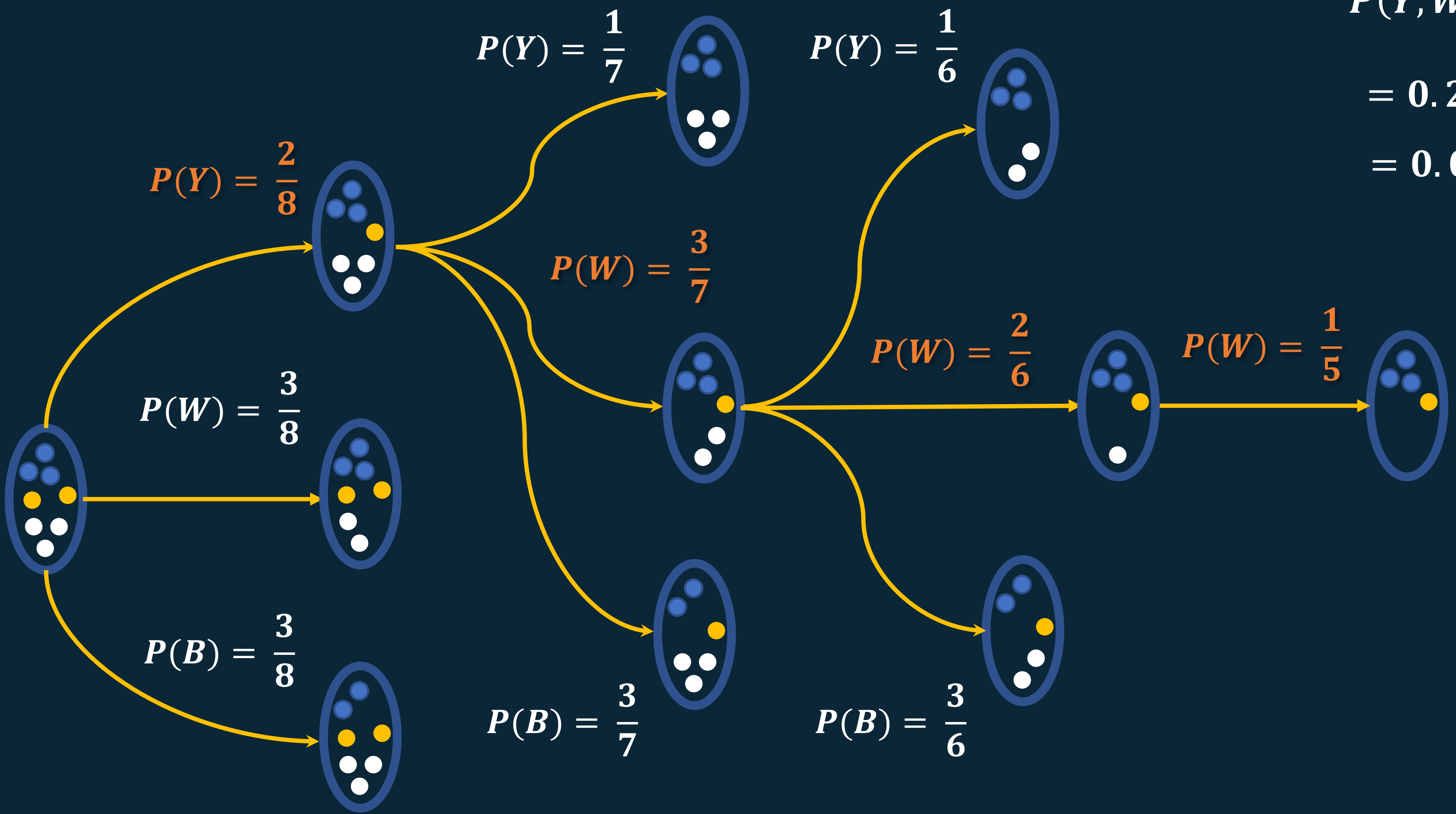
$P(A \vee \neg B)$



$P(\neg A \vee B)$



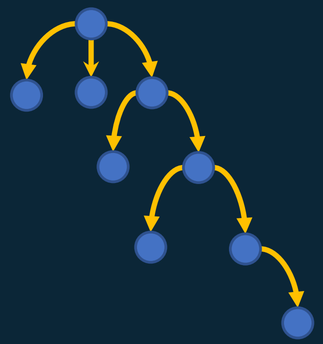
9. Tree diagrams



$$P(Y, W, W, W) = \frac{2}{8} \times \frac{3}{7} \times \frac{2}{6} \times \frac{1}{5}$$

$$= 0.25 \times 0.4286 \times 0.333\dot{3} \times 0.2$$

$$= 0.007142857 = \sim 0.714\%$$



$P(Y)$ = Prob. Picking Yellow

$P(B)$ = Prob. Picking Blue

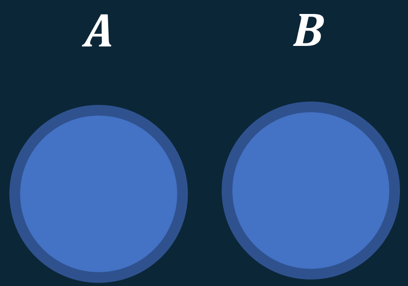
$P(W)$ = Prob. Picking White

10. Probability Rules

Addition rule - disjoint outcomes

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

$$P(A_1 \lor A_n) = P(A_1) + \ldots + P(A_n)$$



If A = Even die roll and
B = odd die roll

$$P(A \lor B) = 1.0$$

Multiplication rule - disjoint outcomes

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

$$P(A_1 \land A_n) = P(A_1) \times \ldots \times P(A_n)$$

If A = Even die on roll 1 and
B = odd die roll on roll 2

$P(A) = 0.5$ $P(B) = 0.5$

$$P(A \land B) = 0.5 \times 0.5$$
$$= 0.25$$

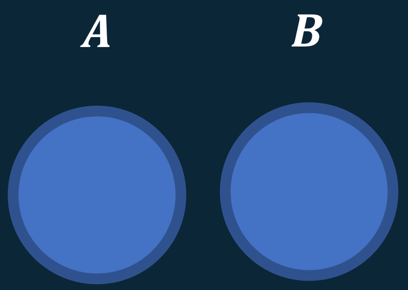
Die 1	Die 2	Prob.
Even	Even	$\frac{1}{4} = 0.25$
Even	Odd	$\frac{1}{4} = 0.25$
Odd	Even	$\frac{1}{4} = 0.25$
Odd	Odd	$\frac{1}{4} = 0.25$
Total		1.0

11. Probability Rules

Addition rule - any outcomes

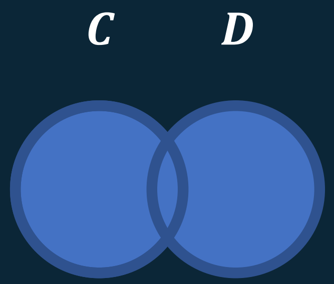
$$P(A \lor B) = P(A) + P(B) - P(A \land B)$$

Probability that at least A or B occurs.



If A = Even die roll and B = odd die roll

$$P(A \lor B) = 1.0$$

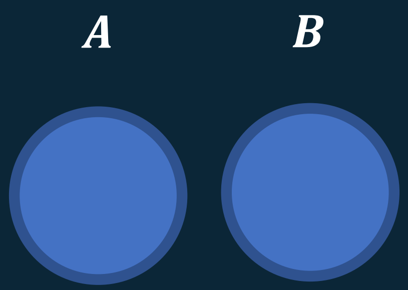


If C = Even die roll or D = Rolling a 6

$$P(C \lor D) = 0.5 + 0.166 \cdot -(0.166 \cdot) = 0.5$$

Complement rule

$$P(A) + P(\neg A) = 1$$

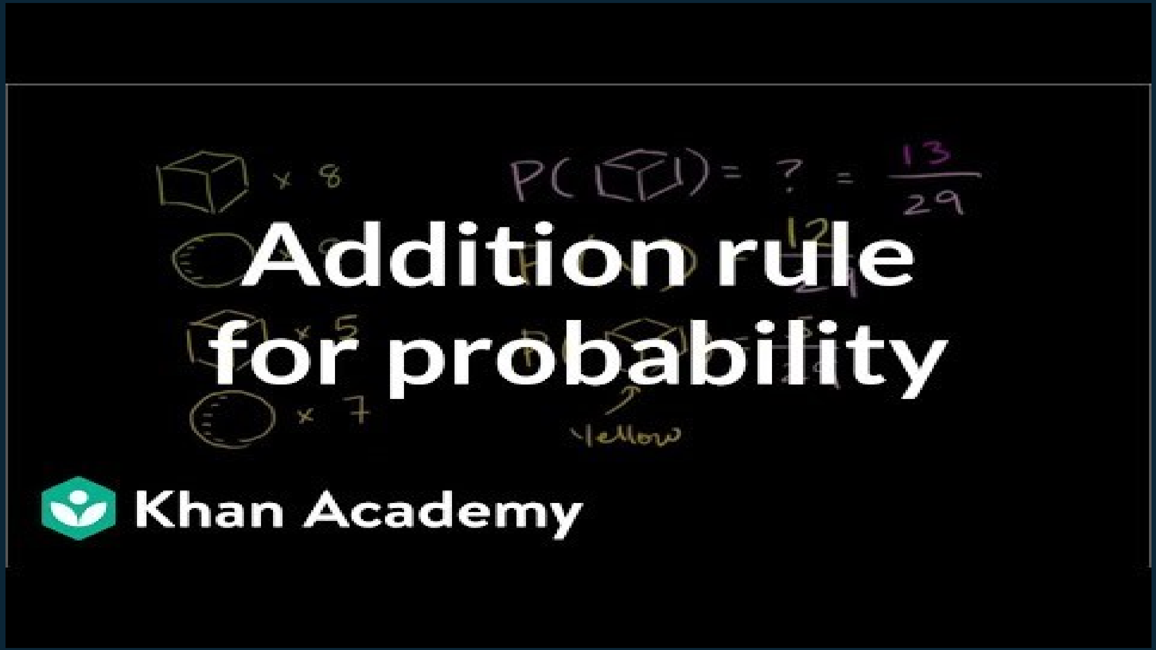


If A = Even die roll + $\neg A$ = odd die roll

$$P(A) + P(\neg A) = 0.5 + 0.5 = 1$$

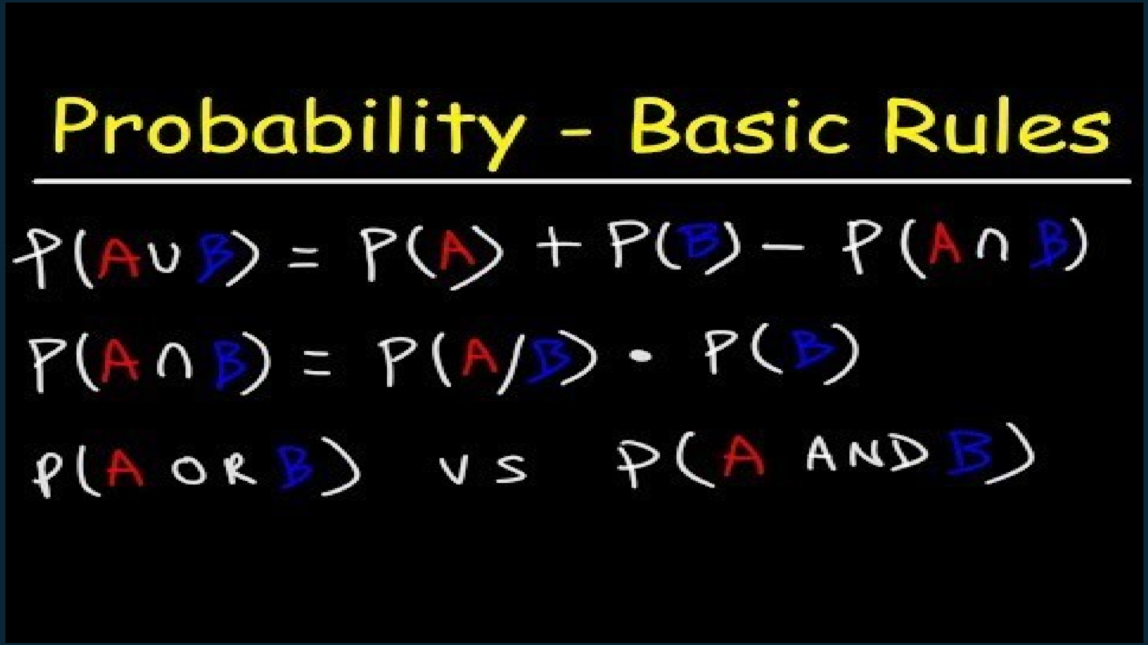
12. Probability Rules

Addition rule - disjoint outcomes



Credit: Khan Academy

Multiplication rule - disjoint outcomes



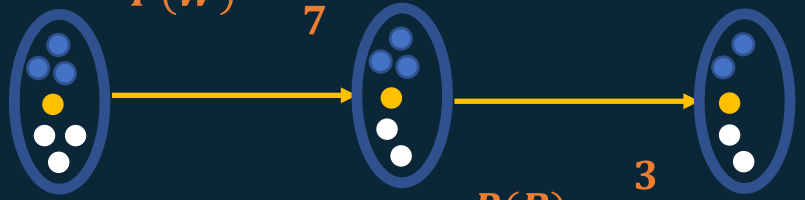
Credit: The Organic Chemistry Tutor

13. Conditional Probability

A given that *B* happened → $P(A|\,B) = \dfrac{P(A\ \wedge B)}{P(B)}$

Read "|" symbol as meaning "given".

$P(W) = \dfrac{3}{7}$



$P(B) = \dfrac{3}{6}$

$$P(W, B) = \frac{3}{7} \times \frac{3}{6}$$
$$= 0.21428571428$$
$$= \sim 21.4\%$$

Probability of sequence

vs.

Probability of Blue given that white picked.

$$P(B|\,W) = \frac{P(B) \times P(W)}{P(W)} = \frac{\frac{3}{6} \times \frac{3}{7}}{\frac{3}{7}} = \frac{0.5 \times 0.42857142857}{0.42857142857} = \frac{0.214285714285}{0.42857142857} = 0.5$$

Edge Hill University

</TECHUP_WOMEN>

14. Practical Example

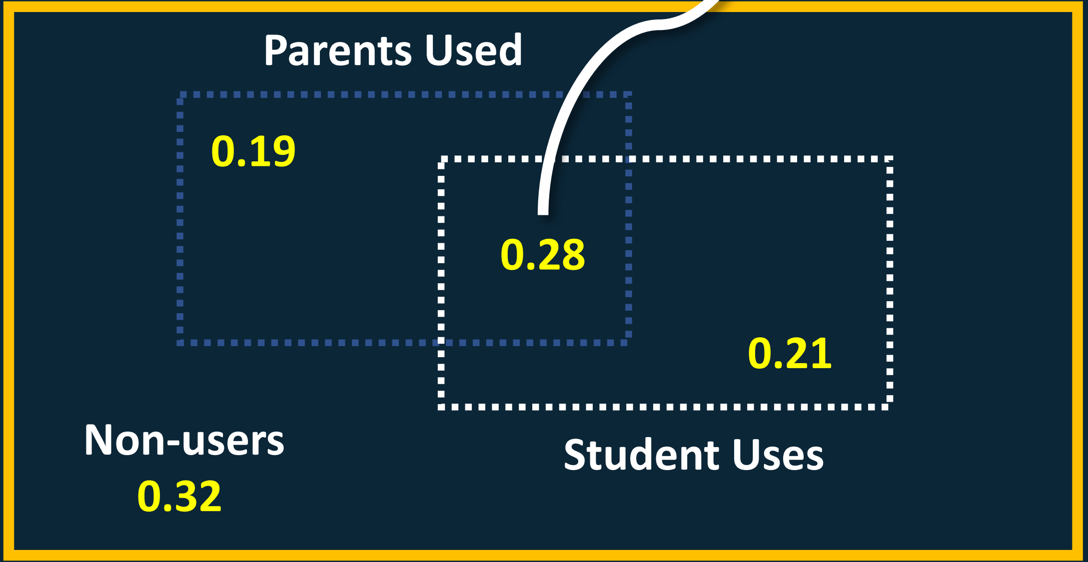
- Suppose we are government data scientists, put in charge of guiding policy based on data.
- We're given data describing drug use by students and their parents.
- We form a contingency table using the data.
- It shows the number of parents who used drugs in the past, and students actively using drugs now.

$$\frac{125}{445} = 0.28089887\ldots = \sim 0.28 = \sim 28\%$$

Parents

Student		Used	Not	Total
	Uses	125	94	219
	Not	85	141	226
	Total	210	235	445

Contingency Table



Probabilities sum to 1

15. Practical Example

- A student who doesn't use drugs is chosen at random. What is the chance that at least one of her parents used drug in the past?
- We can use the conditional probability rule – or look a the contingency table.

A = Parent Used

B = Student Uses

		Parents		
		Used	Not	Total
Student	Uses	125	94	219
	Not	85	141	226
	Total	210	235	445

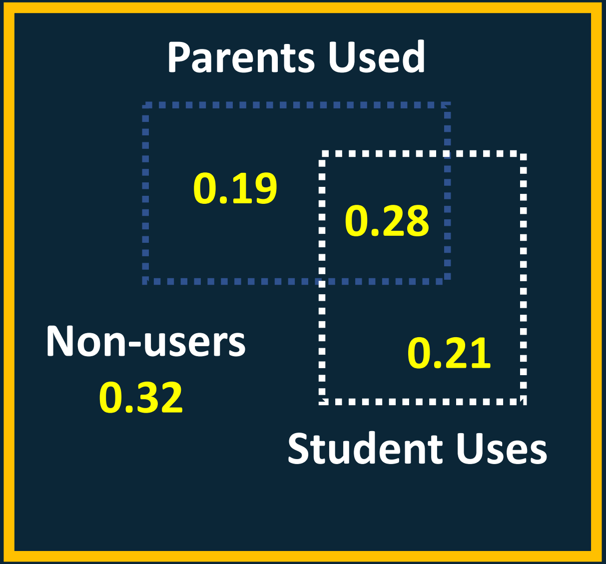
Contingency Table

$$P(A \mid \neg B) = \frac{85}{226} = 0.376 = 37.6\%$$

$$P(B) = \frac{219}{445} = 0.492 = 49.2\%$$

$$P(B \wedge \neg A) = \frac{94}{445} = 0.21 = 21\%$$

$$P(B \mid A) = \frac{125}{210} = 0.6 = 60\%$$



16. Rules Summary

Addition rule - disjoint outcomes

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

Multiplication rule - disjoint outcomes

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

Addition rule - any outcomes

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$

Complement rule

$$P(A) + P(\neg A) = 1$$

Conditional Probability rule

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

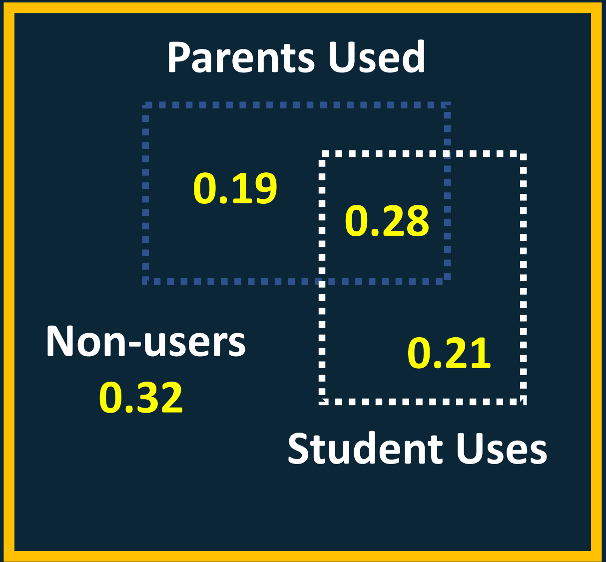
A = Parent Used
B = Student Uses

Lot's of questions we can now ask and answer using probability rules!

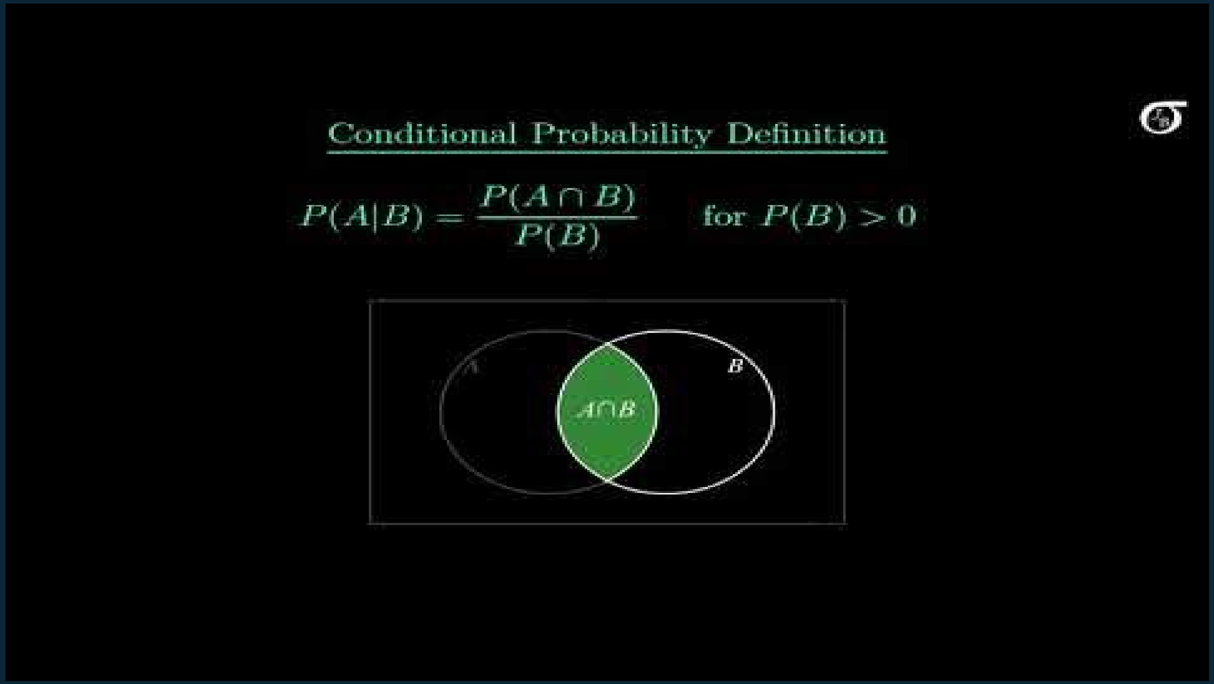
Parents

Student		Used	Not	Total
	Uses	125	94	219
	Not	85	141	226
	Total	210	235	445

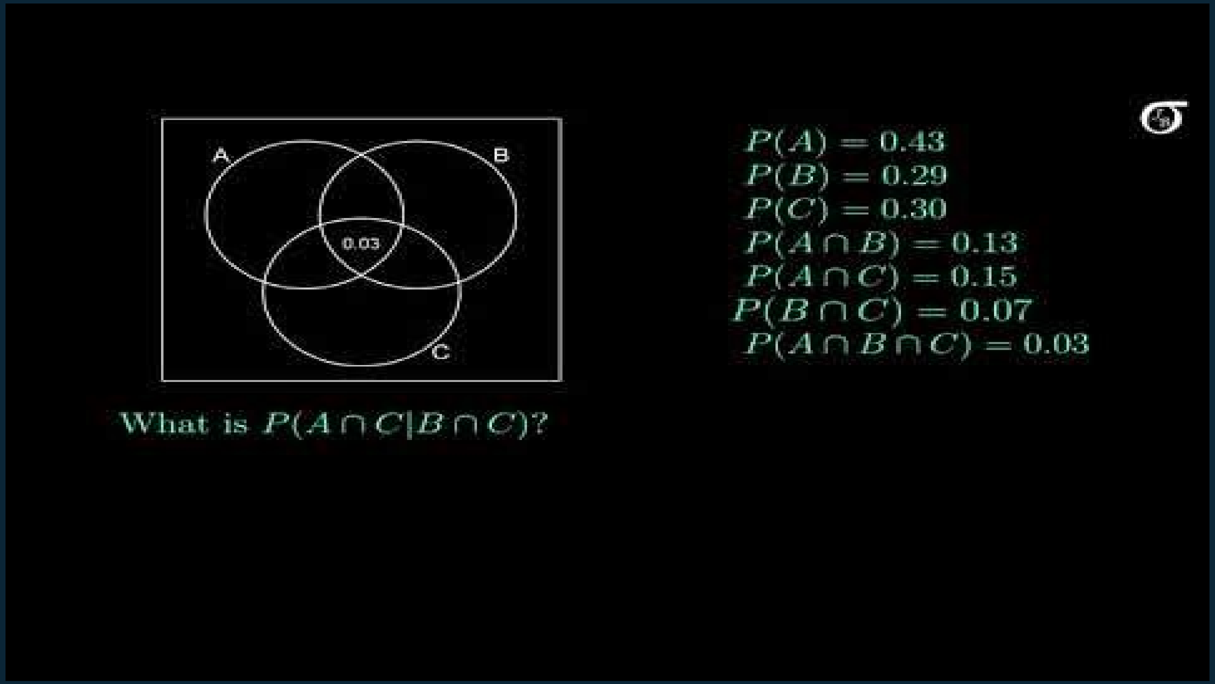
Contingency Table



17. Conditional Probability - Review



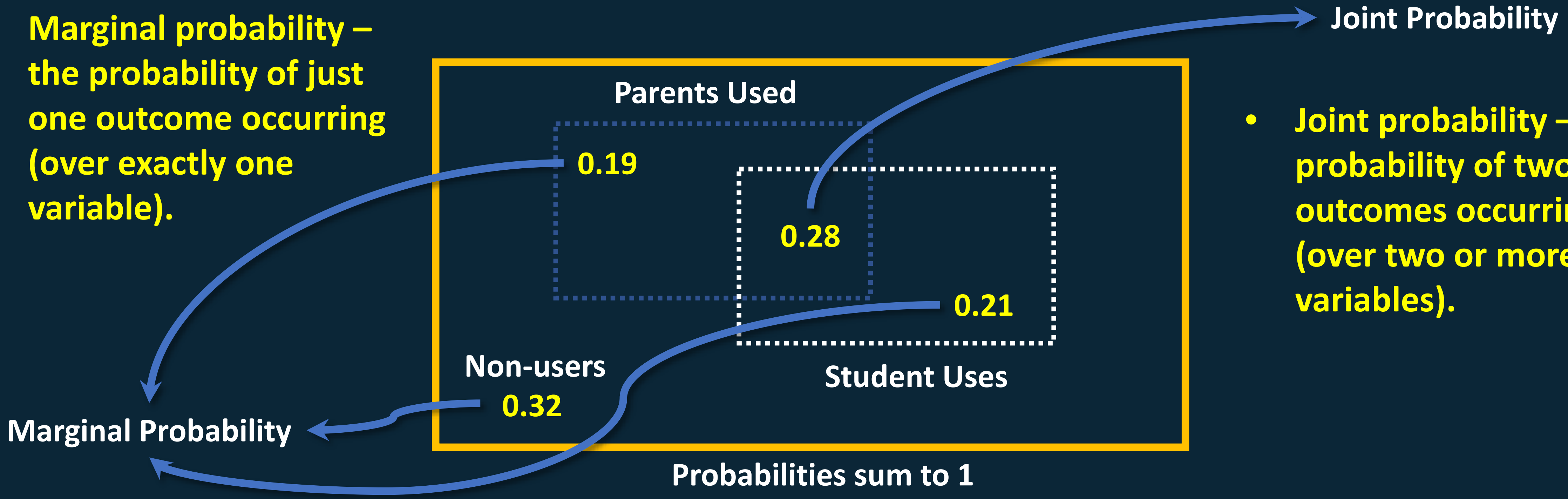
Credit: jbstatistics



Credit: jbstatistics

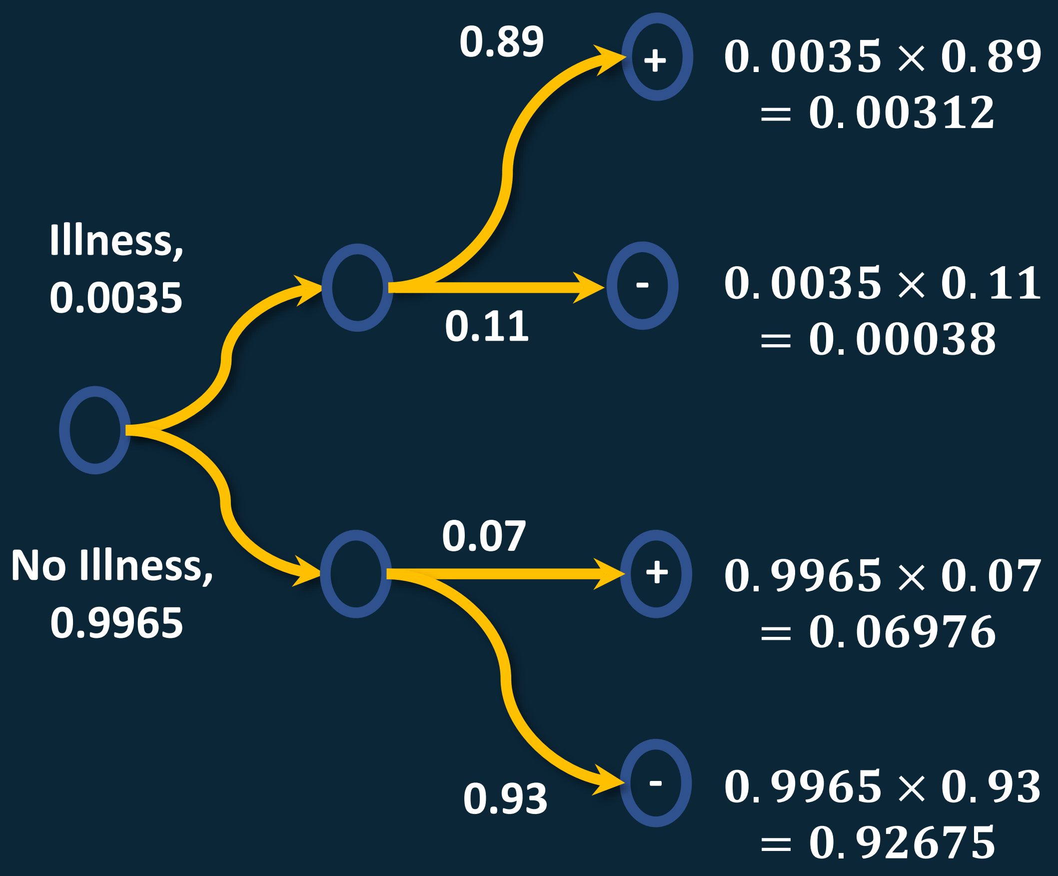
18. Marginal vs Joint

- Marginal probability – the probability of just one outcome occurring (over exactly one variable).



- Joint probability – the probability of two outcomes occurring (over two or more variables).

19. Inverting Probabilities



- Sometimes we have data describing a specific situation that doesn't quite suit our purposes.
- Suppose patients are being studied during a medical trial.
- We see that the disease is very rare – there is an ~99.7% chance that an arbitrary patient doesn't have the disease.
- The aim of the trial is to determine how effective a procedure is for detecting the disease.
  - Ill patients - 89% chance they test +.
  - Ill patients - 11% chance they test -.
  - Healthy patients - 7% chance they test +.
  - Healthy patients - 93% chance they test -.
- Likelihood of a person testing positive or negative:
  - P(ill patient testing +) = 0.312%.
  - P(ill patient testing -) = 0.038%.
  - P(healthy patient testing +) = 6.976%.
  - P(healthy patient testing -) = 92.675%.

20. Inverting Probabilities

$P(ill \mid tests\ postive)$?

- Sounds ok so far – but what if I want to ask a different question. I want to ask – what is the probability that a patient testing positive is ill?
- We can invert the probability tree to ask this question. All the data is there, except it is very difficult to do – it's get complicated.
- The probability of a patient testing positive, being ill, is actually just 4% - so how did I compute it?
- I used our final rule from probability theory to compute the likelihood – this rule is called Bayes Theorem.

21. Bayes Theorem

- This simple theorem let's us compute conditional probabilities in an easy way.
- It's very powerful – it actually forms the basis for a number of machine learning based systems.
- What does it say?

The probability of A given that B has happened, is equal to the probability of B given that A has happened, multiplied by the probability of A, all divided by the probability of B.

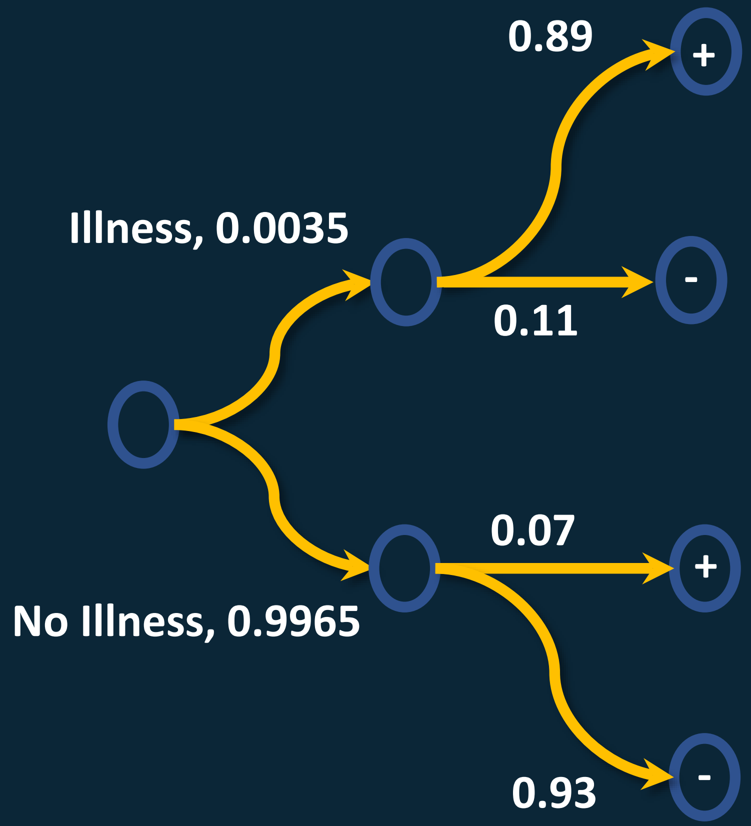
- It is unlikely that this will make sense just by looking at it.
- We can expand it so it makes more sense, but it's still a lot to digest.
- So let's use it to answer the question we posed on the last slide.

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

Expands to…

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

22. Bayes Theorem



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

Dividing term

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

Multiplicative term

Test:

$If\ P(B \wedge A) = 0.5$

AND $if\ P(A) = 0.8$

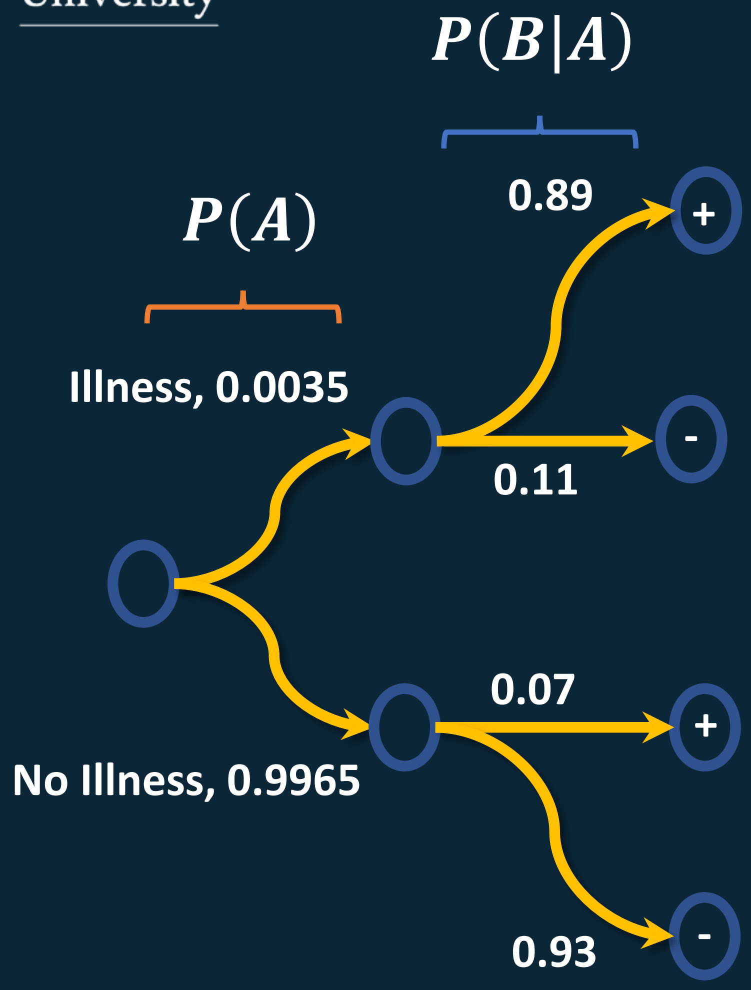
Then $P(B|A) = 0.625$

So $P(B|A) \times P(A) = 0.5$

Cancels to:

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

23. Bayes Theorem



A = Patient has illness
B = Positive test

$$P(B \wedge A) = 0.89 \times 0.0035 = 0.00312 = 0.312\%$$

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

Hint: $\frac{numerator}{denominator}$

$$P(B) = P(B \wedge A) + P(B \wedge \neg A)$$

+ test and Ill

+ test and not Ill

$$P(B) = \underbrace{P(A)} \times \underbrace{P(B|A)} + P(\neg A) \times P(B|\neg A)$$

$$= 0.00035 \times 0.89 + 0.9965 \times 0.07 = 0.07288 = 7.288\%$$

24. Bayes Theorem

A = Patient has illness

B = Positive test

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

$$P(B) = 0.07288$$

$$P(B \land A) = 0.00312$$

$$P(A|B) = \frac{0.00312}{0.07288} \approx 0.0428 \approx 4.28\%$$

25. Bayes Theorem

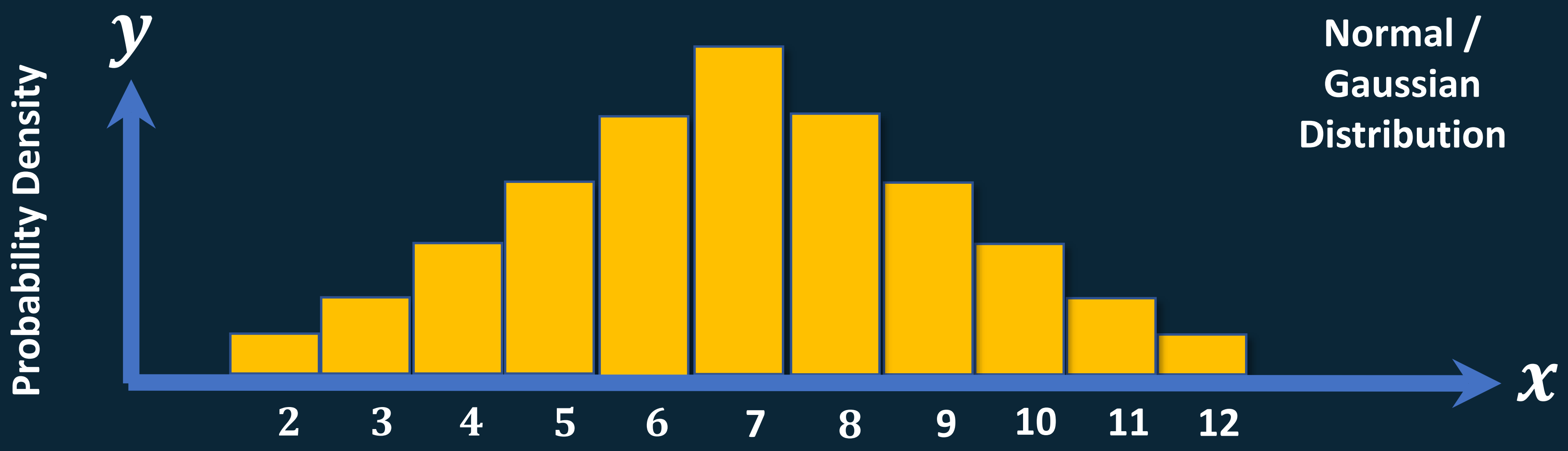
$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

Credit: Wireless Philosophy

26. Probability/ Data distributions

Ways to get a sum of 7:

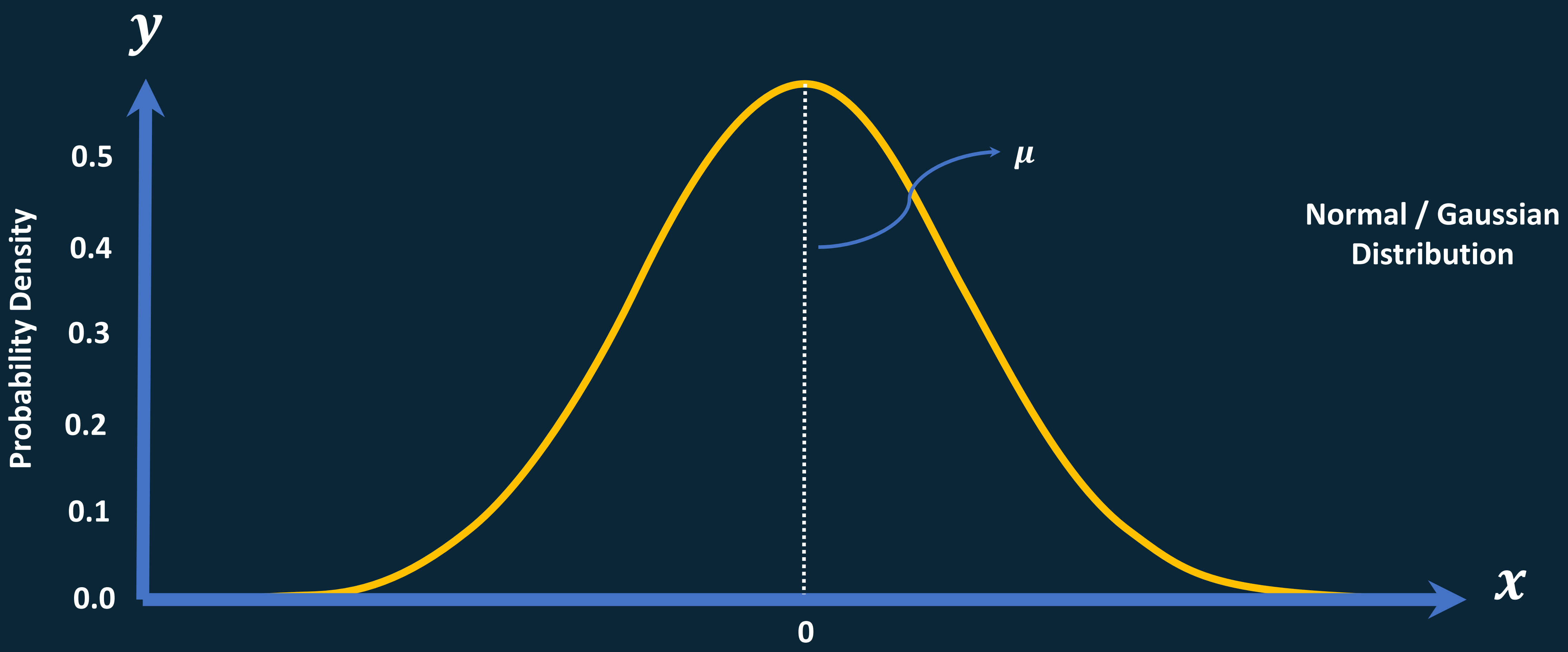
6 + 1
1 + 6
3 + 4
4 + 3
5 + 2
2 + 5



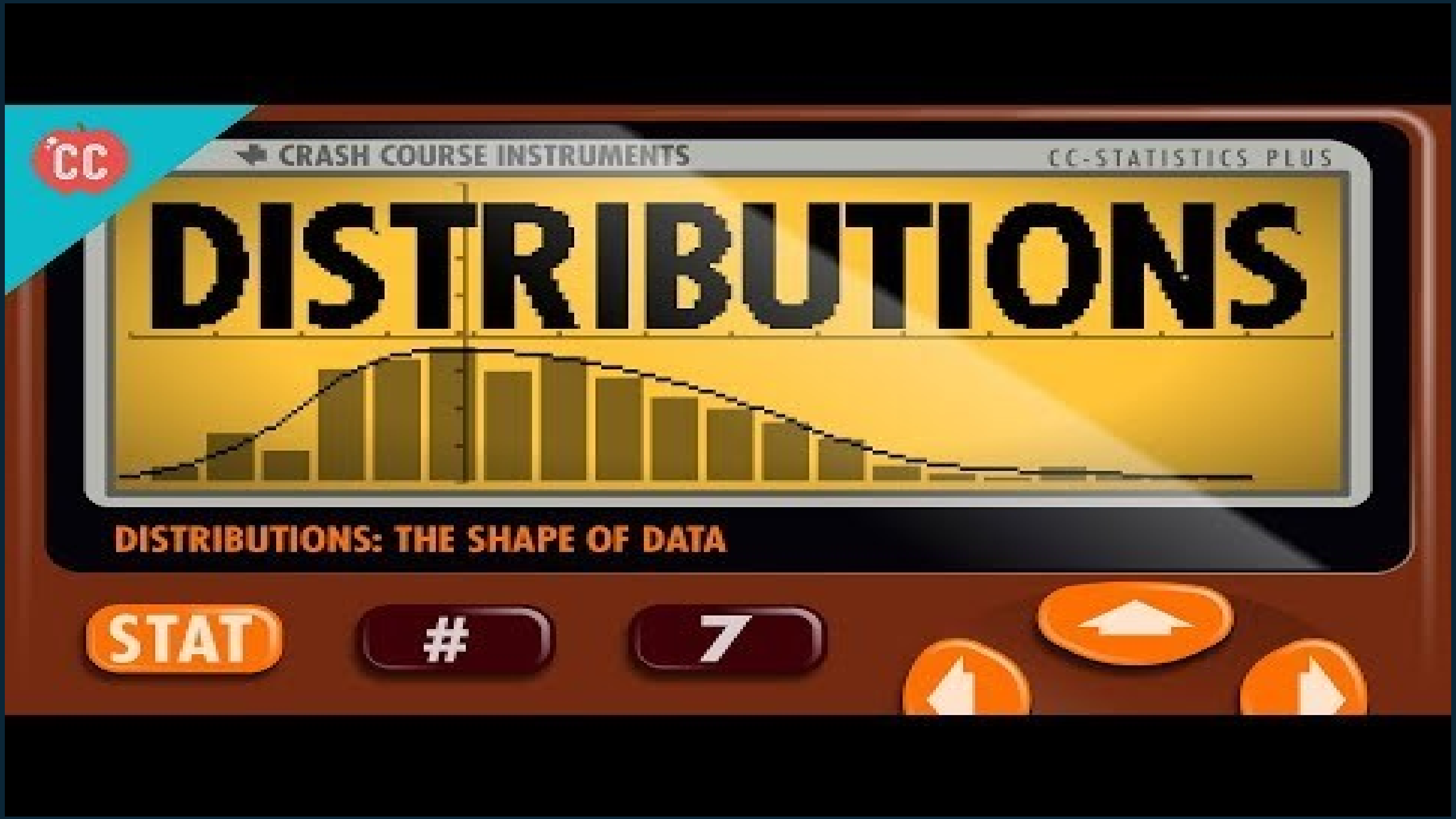
Normal / Gaussian Distribution

Sum	2	3	4	5	6	7	8	9	10	11	12
Prob.	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

27. Continuous distributions



28. Continuous distributions



Credit: CrashCourse

Edge Hill University

29. Bayes Theorem

</TECHUP_WOMEN>

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

Credit: CrashCourse

30. Checkpoint

We've reached another checkpoint. Let's recap what we've introduced so far.

- The nature of probability.
- Different types of probabilistic event.
- Probability notation.
- How to define events and express them happening independently or together.
- Tree-diagrams.
- The rules of probability.
- Conditional probability.
- Bayes Theorem.
- Data distributions.

This puts you in a great place to tackle our next topic – hypothesis testing.