



# Probability Basics

These slides were assembled by Eric Eaton, with grateful acknowledgement of the many others who made their course materials freely available online. Feel free to reuse or adapt these slides for your own academic purposes, provided that you include proper attribution. Please send comments and corrections to Eric.

# Probabilities

- 30 years of AI research danced around the fact that the world was inherently uncertain
- Bayesian Inference:
  - Use probability theory and information about independence
  - Reason diagnostically (from evidence (effects) to conclusions (causes))...
  - ...or causally (from causes to effects)
- Probabilistic reasoning generally gives probabilistic results

# Discrete Random Variables

- Let  $A$  denote a random variable
  - $A$  represents an event that can take on certain values
  - Each value has an associated probability
- Examples of binary random variables:
  - $A$  = I have a headache
  - $A$  = Sally will be the US president in 2020
- $P(A)$  is “the fraction of possible worlds in which  $A$  is true” (Frequentist)
  - ... or  $P(A)$  is a measure of our uncertainty (Bayesian)
  - We could spend hours on the philosophy of this, but we won't

# Visualizing $A$

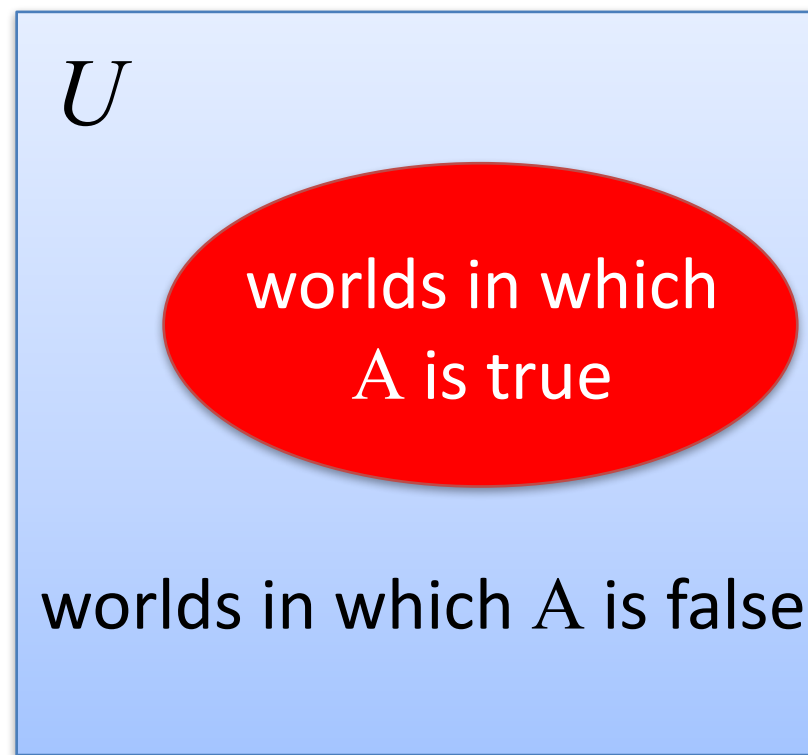
- Universe  $U$  is the sample space of all possible worlds
  - Its area is 1
  - $P(U) = 1$

- $P(A) = \text{area of red oval}$

- Therefore:

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

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



# Axioms of Probability

Kolmogorov showed that three simple axioms lead to the rules of probability theory

- de Finetti, Cox, and Carnap have also provided compelling arguments for these axioms

1. All probabilities are between 0 and 1:

$$0 \leq P(A) \leq 1$$

2. Valid propositions (tautologies) have probability 1, and unsatisfiable propositions have probability 0:

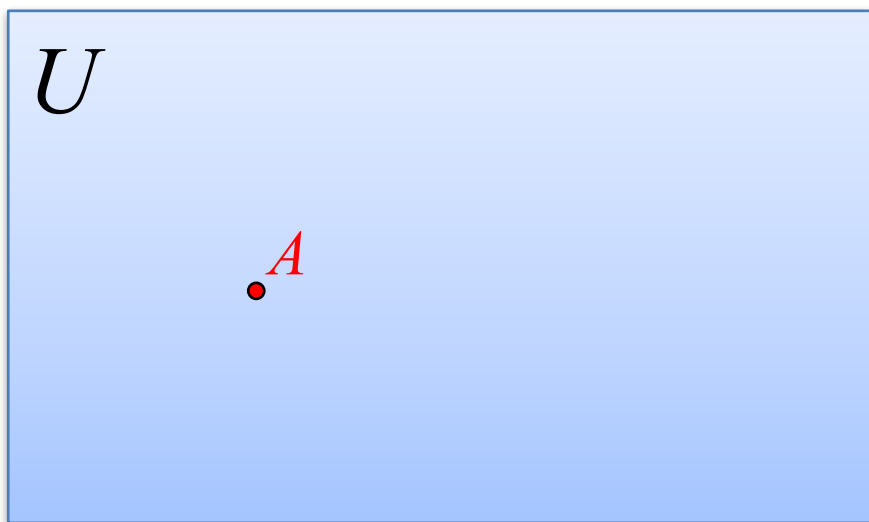
$$P(\text{true}) = 1 ; \quad P(\text{false}) = 0$$

3. The probability of a disjunction is given by:

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

# Interpreting the Axioms

- $0 \leq P(A) \leq 1$
- $P(\text{true}) = 1$
- $P(\text{false}) = 0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

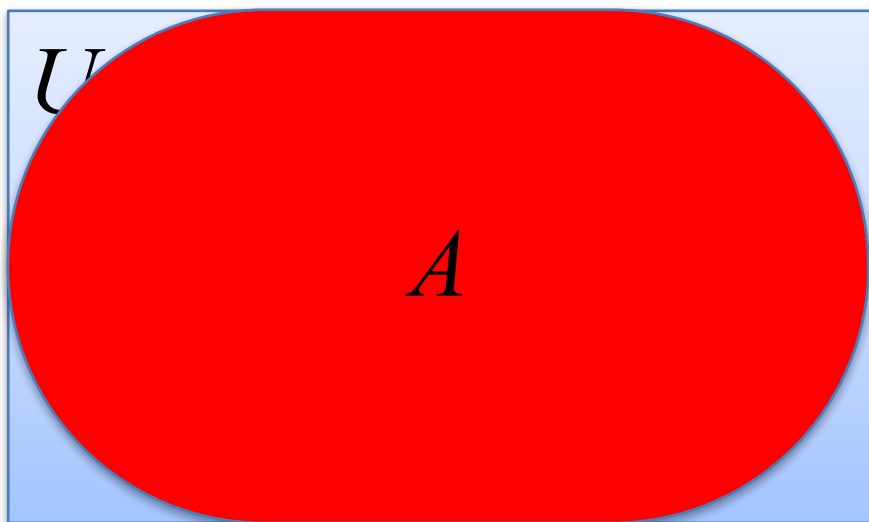


The area of  $A$  can't get any smaller than 0

A zero area would mean no world could ever have  $A$  true

# Interpreting the Axioms

- $0 \leq P(A) \leq 1$
- $P(\text{true}) = 1$
- $P(\text{false}) = 0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

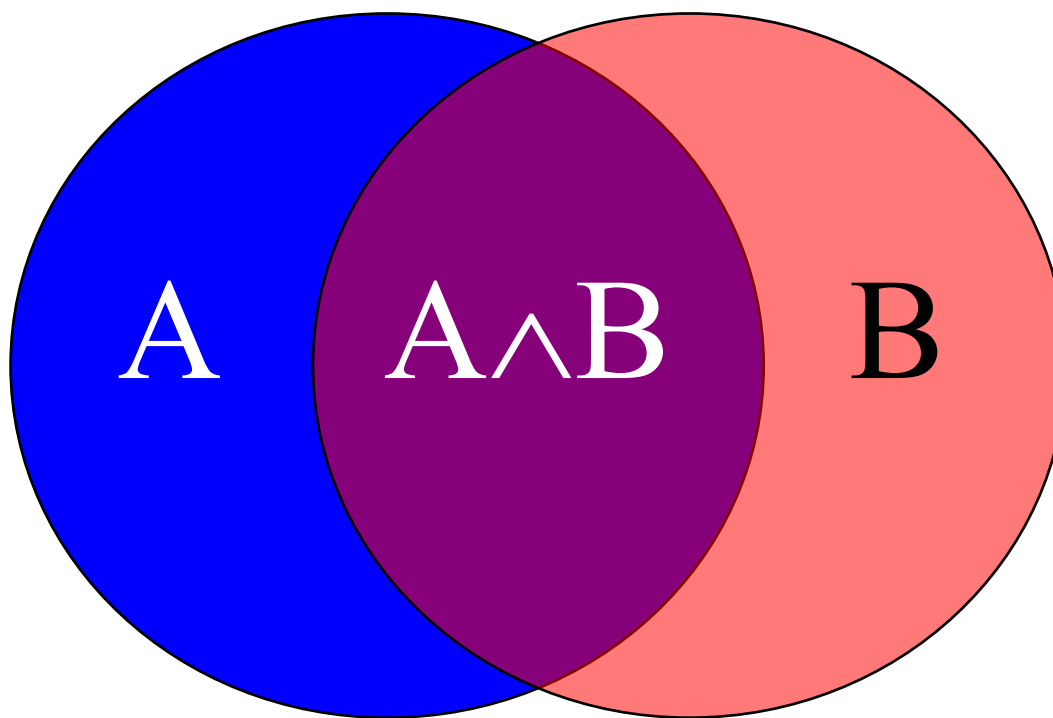


The area of  $A$  can't get any bigger than 1

An area of 1 would mean  $A$  is true in all possible worlds

# Interpreting the Axioms

- $0 \leq P(A) \leq 1$
- $P(\text{true}) = 1$
- $P(\text{false}) = 0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$





# These Axioms are Not to be Trifled With

- There have been attempts to develop different methodologies for uncertainty:
  - Fuzzy Logic
  - Three-valued logic
  - Dempster-Shafer
  - Non-monotonic reasoning
- But the axioms of probability are the only system with this property:

If you gamble using them you can't be exploited by an opponent using some other system [di Finetti, 1931]

# An Important Theorem

$$0 \leq P(A) \leq 1$$

$$P(\text{true}) = 1; \quad P(\text{false}) = 0$$

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

From these we can prove:

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

Proof: Let  $B = \neg A$ . Then, we have

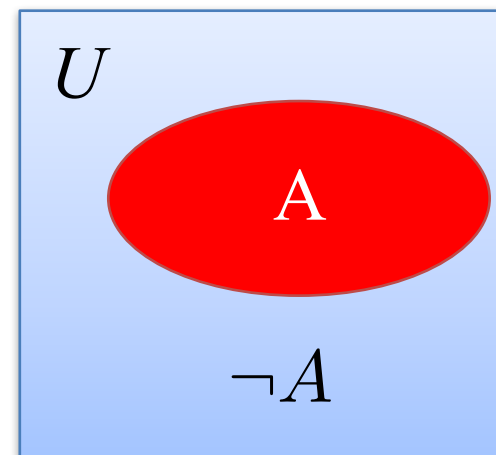
$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$

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

$$P(\text{true}) = P(A) + P(\neg A) - P(\text{false})$$

$$1 = P(A) + P(\neg A) - 0$$

$$P(\neg A) = 1 - P(A) \quad \square$$



# Another Important Theorem

$$0 \leq P(A) \leq 1$$

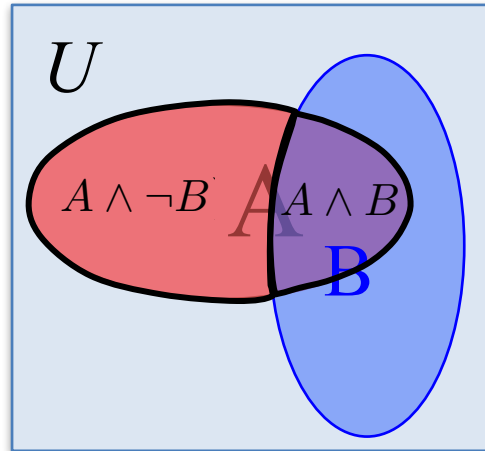
$$P(\text{True}) = 1; \quad P(\text{False}) = 0$$

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

From these we can prove:

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

How?



# Multi-valued Random Variables

- Suppose  $A$  can take on more than 2 values
- $A$  is a *random variable with arity  $k$*  if it can take on exactly one value out of  $\{v_1, v_2, \dots, v_k\}$
- Thus...

$$P(A = v_i \wedge A = v_j) = 0 \quad \text{if } i \neq j$$

$$P(A = v_1 \vee A = v_2 \vee \dots \vee A = v_k) = 1$$

$$1 = \sum_{i=1}^k P(A = v_i)$$

# Multi-valued Random Variables

- We can also show that:

$$P(B) = P(B \wedge [A = v_1 \vee A = v_2 \vee \dots \vee A = v_k])$$

$$P(B) = \sum_{i=1}^k P(B \wedge A = v_i)$$

- This is called **marginalization** over  $A$

# Prior and Joint Probabilities

- **Prior probability**: degree of belief without any other evidence
- **Joint probability**: matrix of combined probabilities of a set of variables

Russell & Norvig's Alarm Domain: (boolean RVs)

- A world has a specific instantiation of variables:  
(alarm  $\wedge$  burglary  $\wedge$   $\neg$ earthquake)
- The joint probability is given by:

$P(\text{Alarm}, \text{Burglary}) =$

	alarm	$\neg$ alarm
burglary	0.09	0.01
$\neg$ burglary	0.1	0.8

Prior probability  
of burglary:

$P(\text{Burglary}) = 0.1$

by marginalization  
over Alarm

# The Joint Distribution

*e.g., Boolean variables  $A, B, C$*

Recipe for making a joint  
distribution of  $d$  variables:

# The Joint Distribution

*e.g., Boolean variables  $A, B, C$*

Recipe for making a joint distribution of  $d$  variables:

1. Make a truth table listing all combinations of values of your variables (if there are  $d$  Boolean variables then the table will have  $2^d$  rows).

<b>A</b>	<b>B</b>	<b>C</b>
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1



# The Joint Distribution

*e.g., Boolean variables  $A, B, C$*

Recipe for making a joint distribution of  $d$  variables:

1. Make a truth table listing all combinations of values of your variables (if there are  $d$  Boolean variables then the table will have  $2^d$  rows).
1. For each combination of values, say how probable it is.

<b>A</b>	<b>B</b>	<b>C</b>	<b>Prob</b>
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10

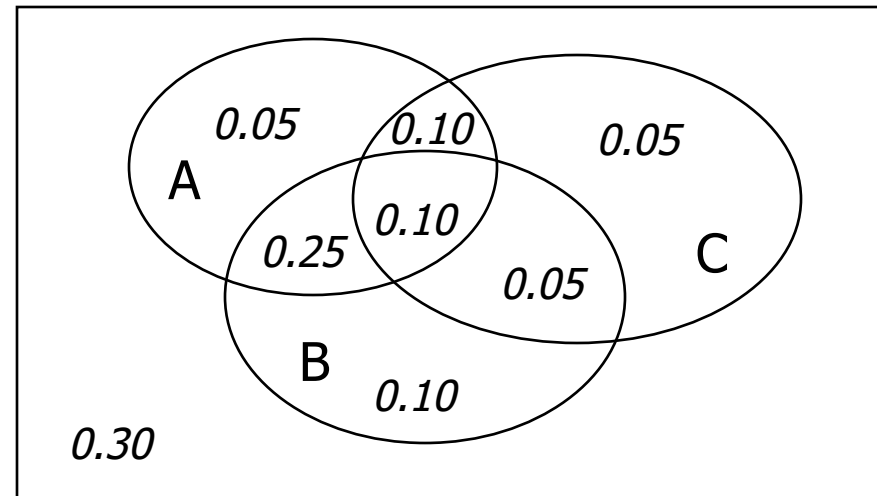
# The Joint Distribution

*e.g., Boolean variables  $A, B, C$*

Recipe for making a joint distribution of  $d$  variables:

1. Make a truth table listing all combinations of values of your variables (if there are  $d$  Boolean variables then the table will have  $2^d$  rows).
1. For each combination of values, say how probable it is.
2. If you subscribe to the axioms of probability, those numbers must sum to 1.

<b>A</b>	<b>B</b>	<b>C</b>	<b>Prob</b>
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10



# Inferring Prior Probabilities from the Joint

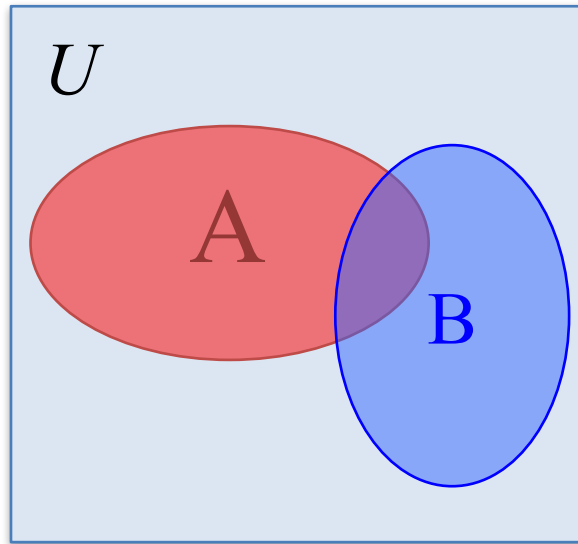
	alarm		¬alarm	
	earthquake	¬earthquake	earthquake	¬earthquake
burglary	0.01	0.08	0.001	0.009
¬burglary	0.01	0.09	0.01	0.79

$$\begin{aligned}P(\text{alarm}) &= \sum_{b,e} P(\text{alarm} \wedge \text{Burglary} = b \wedge \text{Earthquake} = e) \\&= 0.01 + 0.08 + 0.01 + 0.09 = 0.19\end{aligned}$$

$$\begin{aligned}P(\text{burglary}) &= \sum_{a,e} P(\text{Alarm} = a \wedge \text{burglary} \wedge \text{Earthquake} = e) \\&= 0.01 + 0.08 + 0.001 + 0.009 = 0.1\end{aligned}$$

# Conditional Probability

- $P(A | B)$  = Probability that  $A$  is true given  $B$  is true



What if we already know that  $B$  is true?

That knowledge changes the probability of  $A$

- Because we know we're in a world where  $B$  is true

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

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

# Example: Conditional Probabilities

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

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

P(Alarm, Burglary) =

	alarm	$\neg$ alarm
burglary	0.09	0.01
$\neg$ burglary	0.1	0.8

$$\begin{aligned} P(\text{burglary} \mid \text{alarm}) &= P(\text{burglary} \wedge \text{alarm}) / P(\text{alarm}) \\ &= 0.09 / 0.19 = 0.47 \end{aligned}$$

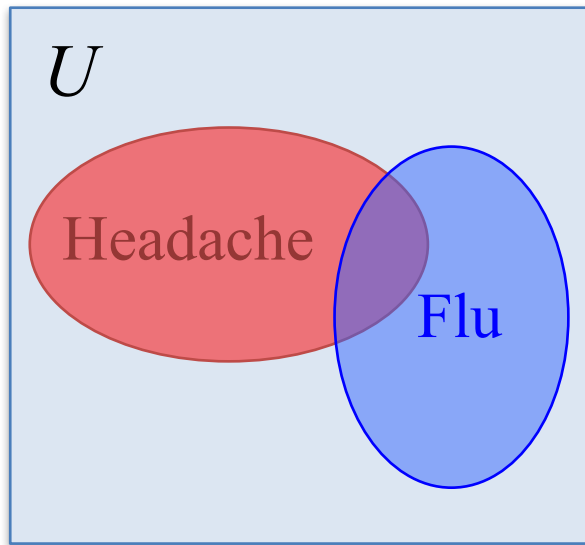
$$\begin{aligned} P(\text{alarm} \mid \text{burglary}) &= P(\text{burglary} \wedge \text{alarm}) / P(\text{burglary}) \\ &= 0.09 / 0.1 = 0.9 \end{aligned}$$

$$\begin{aligned} P(\text{burglary} \wedge \text{alarm}) &= P(\text{burglary} \mid \text{alarm}) P(\text{alarm}) \\ &= 0.47 * 0.19 = 0.09 \end{aligned}$$

# Example: Inference from Conditional Probability

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

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



$$P(\text{headache}) = 1/10$$

$$P(\text{flu}) = 1/40$$

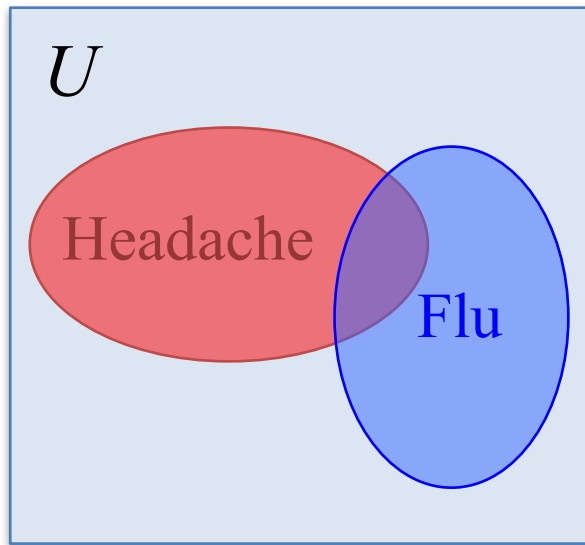
$$P(\text{headache} \mid \text{flu}) = 1/2$$

“Headaches are rare and flu is rarer, but if you’re coming down with the flu, then there’s a 50-50 chance you’ll have a headache.”

# Example: Inference from Conditional Probability

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

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



$$P(\text{headache}) = 1/10$$

$$P(\text{flu}) = 1/40$$

$$P(\text{headache} \mid \text{flu}) = 1/2$$

One day you wake up with a headache. You think: “Drat! 50% of flus are associated with headaches so I must have a 50-50 chance of coming down with flu.”

Is this reasoning good?

# Example: Inference from Conditional Probability

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

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

$$P(\text{headache}) = 1/10$$

$$P(\text{flu}) = 1/40$$

$$P(\text{headache} \mid \text{flu}) = 1/2$$

Want to solve for:

$$P(\text{headache} \wedge \text{flu}) = ?$$

$$P(\text{flu} \mid \text{headache}) = ?$$

$$\begin{aligned} P(\text{headache} \wedge \text{flu}) &= P(\text{headache} \mid \text{flu}) \times P(\text{flu}) \\ &= 1/2 \times 1/40 = 0.0125 \end{aligned}$$

$$\begin{aligned} P(\text{flu} \mid \text{headache}) &= P(\text{headache} \wedge \text{flu}) / P(\text{headache}) \\ &= 0.0125 / 0.1 = 0.125 \end{aligned}$$



# Bayes' Rule

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

- Exactly the process we just used
- The most important formula in probabilistic machine learning

(Super Easy) Derivation:

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

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

these are the same

Just set equal...

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

and solve...



**Bayes, Thomas (1763)** An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, **53:370-418**

# Bayes' Rule

- Allows us to reason from **evidence** to **hypotheses**
- Another way of thinking about Bayes' rule:

$$P(\text{hypothesis} \mid \text{evidence}) = \frac{P(\text{evidence} \mid \text{hypothesis}) \times P(\text{hypothesis})}{P(\text{evidence})}$$

In the flu example:

$$P(\text{headache}) = 1/10$$

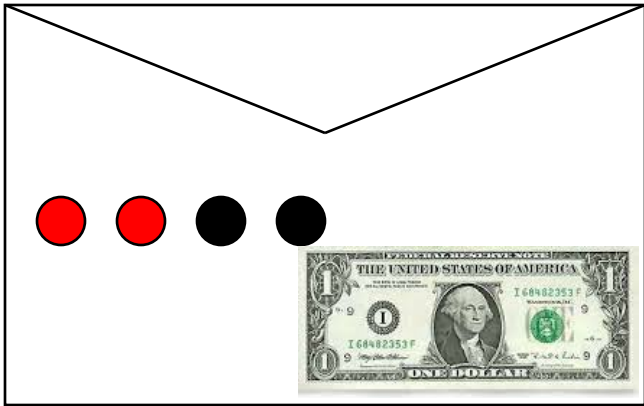
$$P(\text{flu}) = 1/40$$

$$P(\text{headache} \mid \text{flu}) = 1/2$$

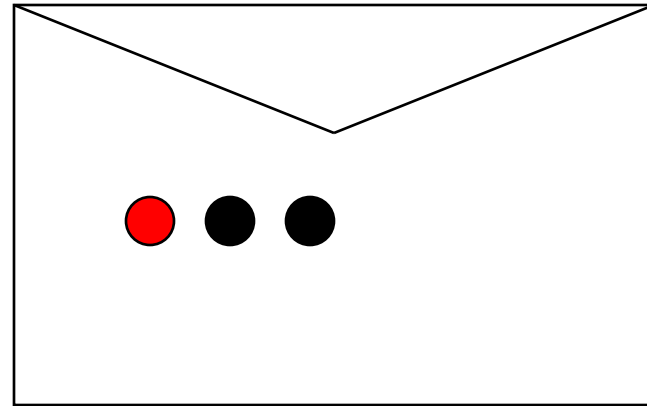
Given evidence of headache, what is  $P(\text{flu} \mid \text{headache})$  ?

Solve via Bayes rule!

# Using Bayes Rule to Gamble



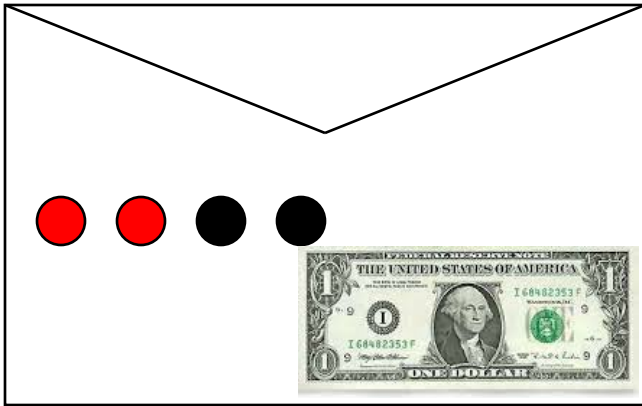
The “Win” envelope has a dollar and four beads in it



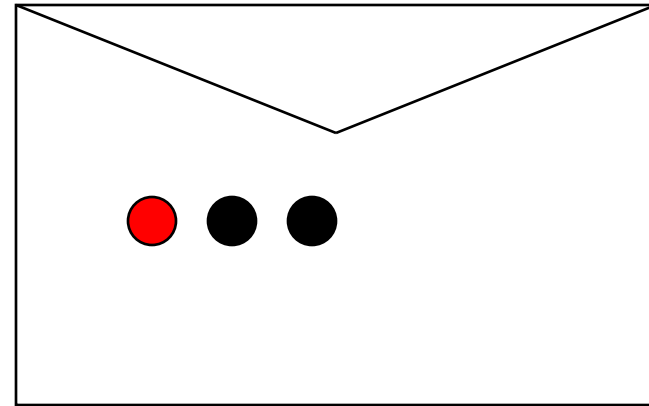
The “Lose” envelope has three beads and no money

**Trivial question:** Someone draws an envelope at random and offers to sell it to you.  
How much should you pay?

# Using Bayes Rule to Gamble



The “Win” envelope has a dollar and four beads in it



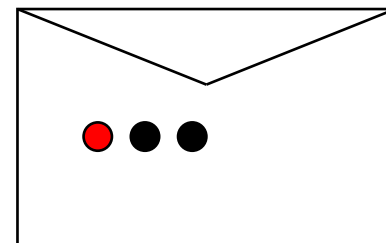
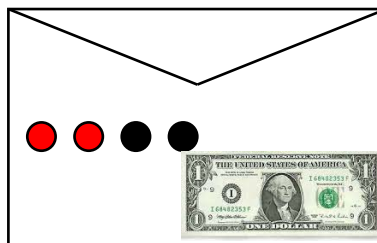
The “Lose” envelope has three beads and no money

**Interesting question:** Before deciding, you are allowed to see one bead drawn from the envelope.

Suppose it's black: How much should you pay?

Suppose it's red: How much should you pay?

# Calculation...



Suppose it's black: How much should you pay?

$$P(b \mid \text{win}) = 1/2$$

$$P(b \mid \text{lose}) = 2/3$$

$$P(\text{win}) = 1/2$$

$$\begin{aligned} P(\text{win} \mid b) &= \alpha P(b \mid \text{win}) P(\text{win}) \\ &= \alpha 1/2 \times 1/2 = 0.25\alpha \end{aligned}$$

$$\begin{aligned} P(\text{lose} \mid b) &= \alpha P(b \mid \text{lose}) P(\text{lose}) \\ &= \alpha 2/3 \times 1/2 = 0.3333\alpha \end{aligned}$$

$$1 = P(\text{win} \mid b) + P(\text{lose} \mid b) = 0.25\alpha + 0.3333\alpha \Rightarrow \alpha = 1.714$$

$$P(\text{win} \mid b) = 0.4286$$

$$P(\text{lose} \mid b) = 0.5714$$

# Independence

- When two sets of propositions do not affect each others' probabilities, we call them **independent**
- Formal definition:

$$\begin{aligned} A \perp\!\!\!\perp B &\Leftrightarrow P(A \wedge B) = P(A) \times P(B) \\ &\Leftrightarrow P(A \mid B) = P(A) \end{aligned}$$

For example, {moon-phase, light-level} might be independent of {burglary, alarm, earthquake}

- Then again, maybe not: Burglars might be more likely to burglarize houses when there's a new moon (and hence little light)
- But if we know the light level, the moon phase doesn't affect whether we are burglarized

# Exercise: Independence

<b>P(smart <math>\wedge</math> study <math>\wedge</math> prep)</b>	<b>smart</b>		<b><math>\neg</math>smart</b>	
	<b>study</b>	<b><math>\neg</math>study</b>	<b>study</b>	<b><math>\neg</math>study</b>
<b>prepared</b>	0.432	0.16	0.084	0.008
<b><math>\neg</math>prepared</b>	0.048	0.16	0.036	0.072

Is *smart* independent of *study*?

Is *prepared* independent of *study*?

# Exercise: Independence

$P(\text{smart} \wedge \text{study} \wedge \text{prep})$	smart		$\neg\text{smart}$	
	study	$\neg\text{study}$	study	$\neg\text{study}$
prepared	0.432	0.16	0.084	0.008
$\neg\text{prepared}$	0.048	0.16	0.036	0.072

Is *smart* independent of *study*?

$$P(\text{study} \wedge \text{smart}) = 0.432 + 0.048 = 0.48$$

$$P(\text{study}) = 0.432 + 0.048 + 0.084 + 0.036 = 0.6$$

$$P(\text{smart}) = 0.432 + 0.048 + 0.16 + 0.16 = 0.8$$

$$P(\text{study}) \times P(\text{smart}) = 0.6 \times 0.8 = 0.48$$

So yes!

Is *prepared* independent of *study*?



# Conditional Independence

- Absolute independence of  $A$  and  $B$ :

$$A \perp\!\!\!\perp B \iff P(A \wedge B) = P(A) \times P(B)$$

$$\iff P(A \mid B) = P(A)$$

**Conditional independence** of  $A$  and  $B$  given  $C$

$$A \perp\!\!\!\perp B \mid C \iff P(A \wedge B \mid C) = P(A \mid C) \times P(B \mid C)$$

- e.g., Moon-Phase and Burglary are **conditionally independent given** Light-Level
- This lets us decompose the joint distribution:
$$P(A \wedge B \wedge C) = P(A \mid C) \times P(B \mid C) \times P(C)$$
  - Conditional independence is weaker than absolute independence, but still useful in decomposing the full joint

# Take Home Exercise:

## Conditional independence

$P(\text{smart} \wedge \text{study} \wedge \text{prep})$	smart		$\neg\text{smart}$	
	study	$\neg\text{study}$	study	$\neg\text{study}$
prepared	0.432	0.16	0.084	0.008
$\neg\text{prepared}$	0.048	0.16	0.036	0.072

Is *smart* conditionally independent of *prepared*, given *study*?

Is *study* conditionally independent of *prepared*, given *smart*?