CS 540 Introduction to AI Basic Probability and Statistics

Daifeng Wang

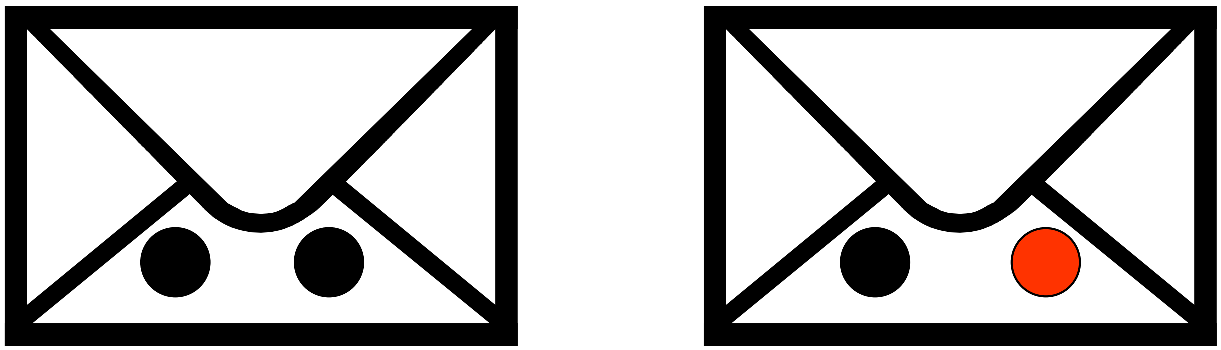
daifeng.wang@wisc.edu

University of Wisconsin, Madison

Based on slides from Xiaojin Zhu (http://pages.cs.wisc.edu/~jerryzhu/cs540.html), modified by Daifeng Wang

Reasoning with Uncertainty

- There are two identical-looking envelopes
  - one has a red ball (worth $100) and a black ball
  - one has two black balls. Black balls worth nothing



- You randomly grabbed an envelope, randomly took out one ball – it's black.
- At this point you're given the option to switch the envelope. To switch or not to switch?

Outline

- Probability
  - random variable
  - Axioms of probability
  - Conditional probability
  - Probabilistic inference: Bayes rule
  - Independence
  - Conditional independence

Uncertainty

- Randomness
  - Is our world random?
- Uncertainty
  - Ignorance (practical and theoretical)
    - Will my coin flip end in head?
    - Will bird flu strike tomorrow?
- Probability is the language of uncertainty
  - Central pillar of modern day artificial intelligence

Sample space

- A space of events that we assign probabilities to
- Events can be binary, multi-valued, or continuous
- Events are mutually exclusive
- Examples
  - Coin flip: {head, tail}
  - Die roll: {1,2,3,4,5,6}
  - English words: a dictionary
  - Temperature tomorrow: R_+ (kelvin)

Random variable

- A variable, x, whose domain is the sample space, and whose value is somewhat uncertain
- Examples:
  - x = coin flip outcome
  - x = first word in tomorrow’s headline news
  - x = tomorrow’s temperature
- Kind of like x = rand()

Probability for discrete events

- Probability P(x=a) is the fraction of times x takes value a
- Often we write it as P(a)
- There are other definitions of probability, and philosophical debates… but we'll not go there
- Examples
  - P(head)=P(tail)=0.5 fair coin
  - P(head)=0.51, P(tail)=0.49 slightly biased coin
  - P(head)=1, P(tail)=0 Jerry's coin
  - P(first word = "the" when flipping to a random page in R&N)=?
- Demo: http://www.bookofodds.com/

Probability table

- Weather

Sunny	Cloudy	Rainy
200/365	100/365	65/365

- P(Weather = sunny) = P(sunny) = 200/365
- P(Weather) = {200/365, 100/365, 65/365}
- For now we'll be satisfied with obtaining the probabilities by counting frequency from data…

Probability for discrete events

- Probability for more complex events A
  - P(A="head or tail")=? fair coin
  - P(A="even number")=? fair 6-sided die
  - P(A="two dice rolls sum to 2")=?

Probability for discrete events

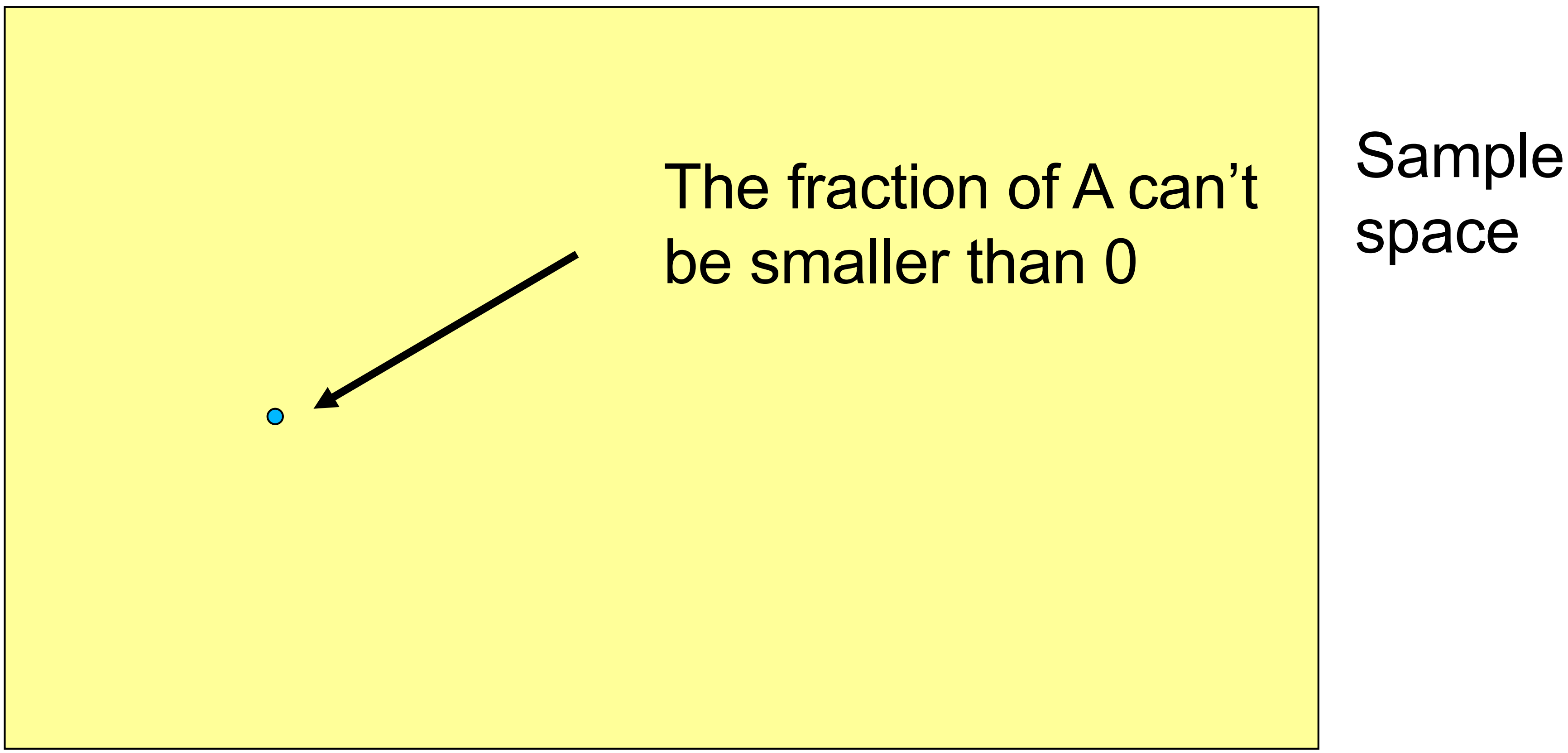
- Probability for more complex events A
  - P(A="head or tail")=0.5 + 0.5 = 1 fair coin
  - P(A="even number")=1/6 + 1/6 + 1/6 = 0.5 fair 6-sided die
  - P(A="two dice rolls sum to 2")=1/6 * 1/6 = 1/36

The axioms of probability

- $P(A) \in [0,1]$
- $P(true)=1$, $P(false)=0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

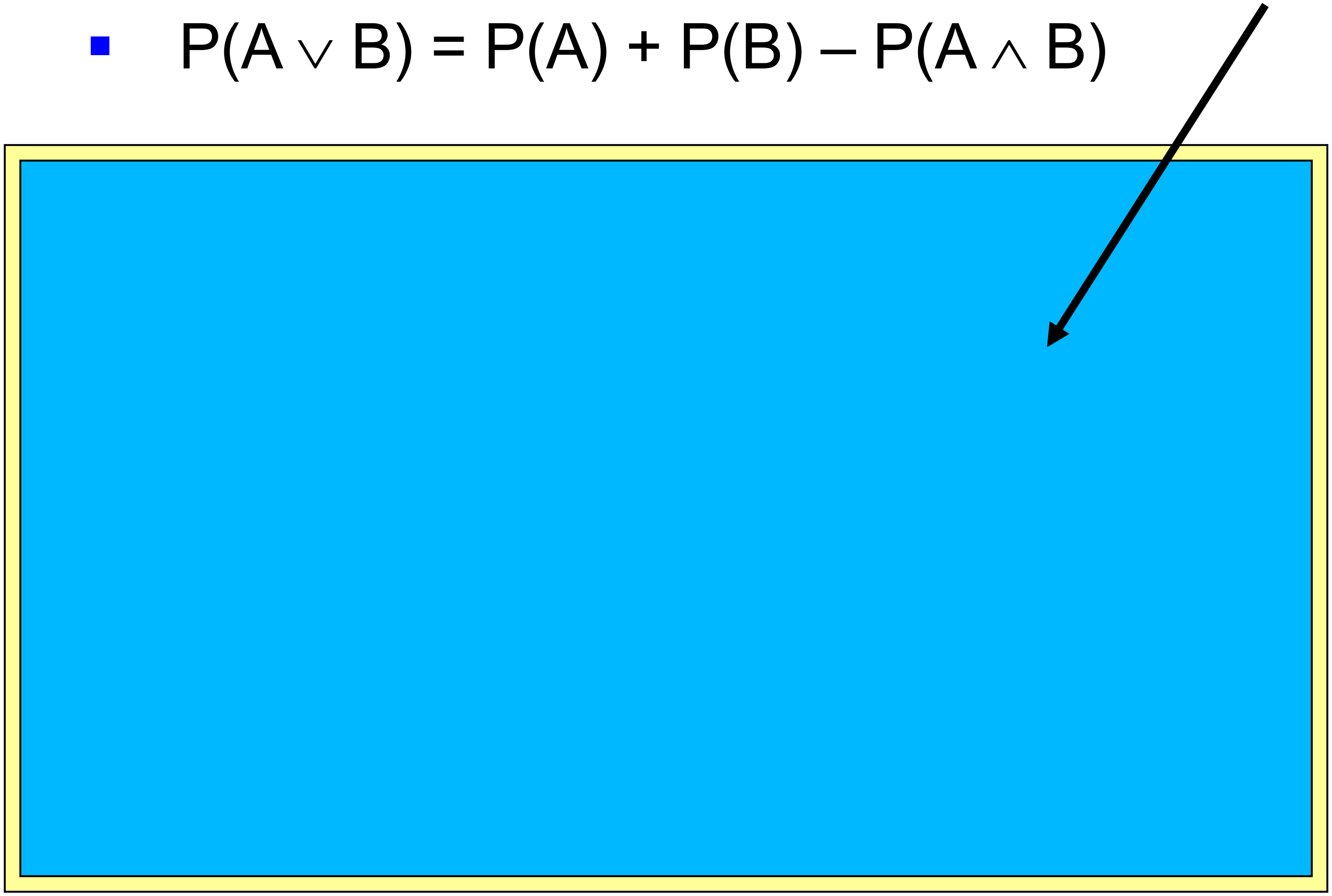
The axioms of probability

- $P(A) \in [0,1]$
- P(true)=1, P(false)=0
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$



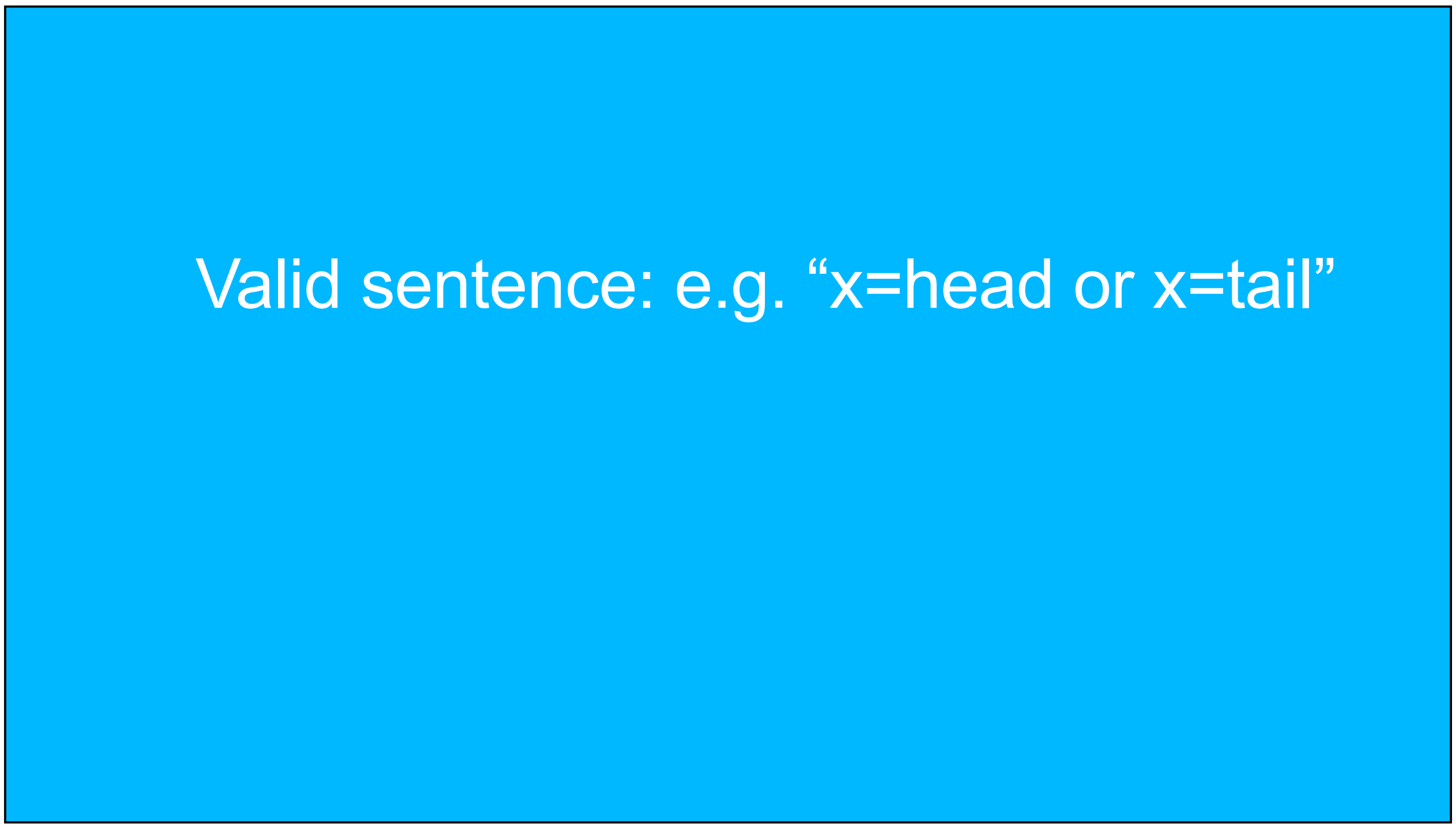
The axioms of probability

- P(A) $\in$ [0,1]
- P(true)=1, P(false)=0
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$



The axioms of probability

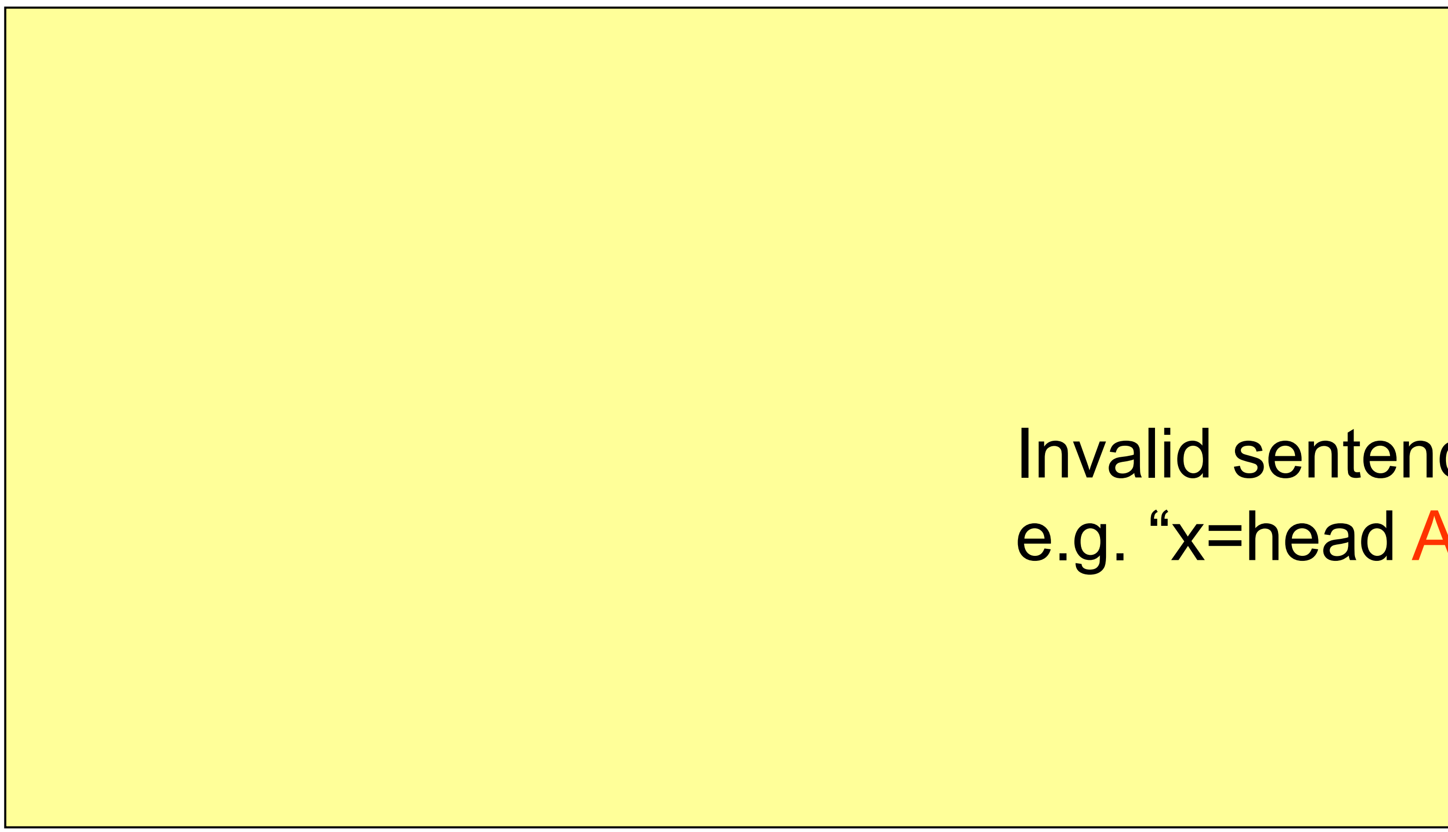
- $P(A) \in [0,1]$
- $P(true)=1$, $P(false)=0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$



Sample space

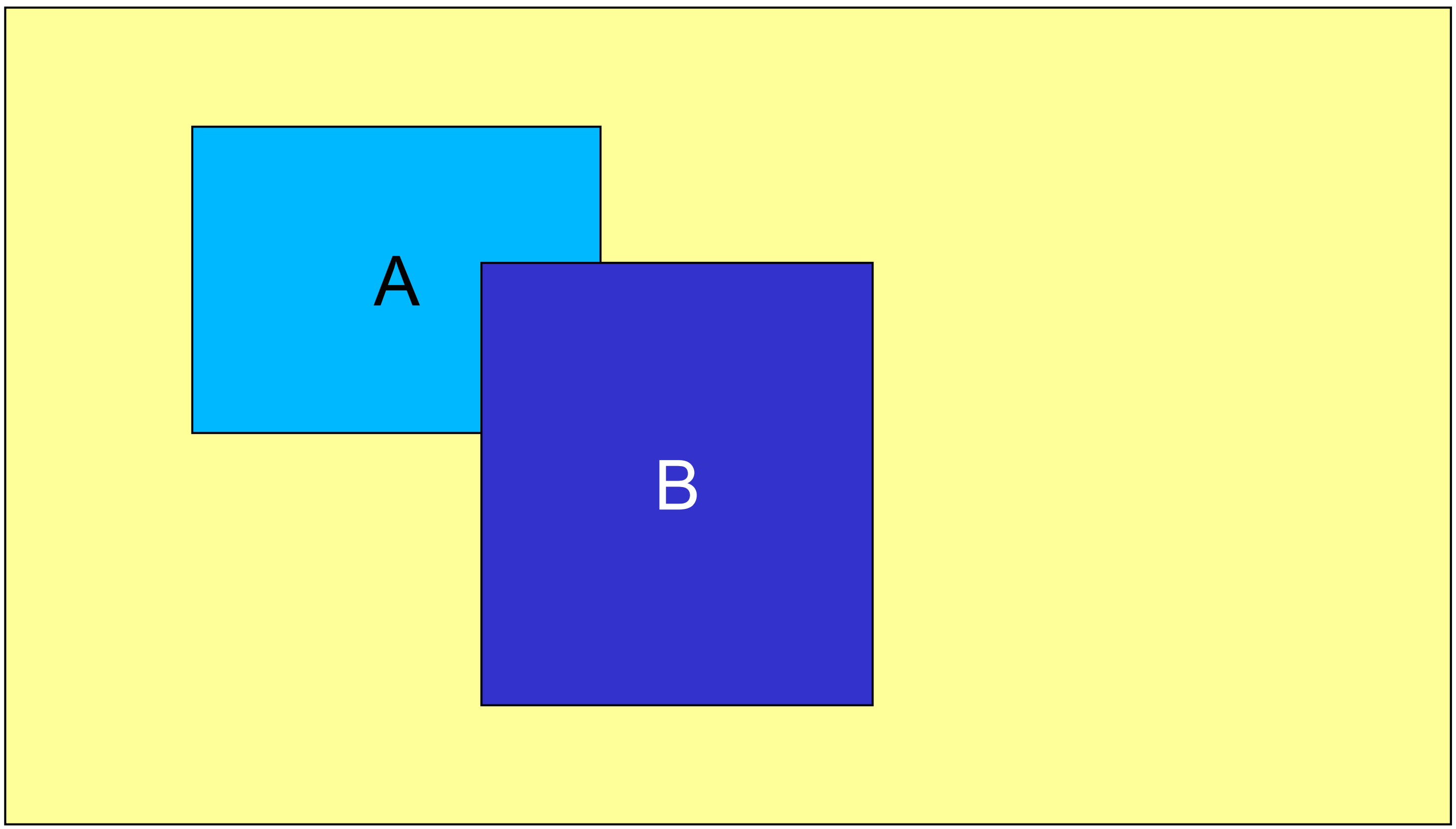
The axioms of probability

- $P(A) \in [0,1]$
- $P(true)=1$, $P(false)=0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$



The axioms of probability

- $P(A) \in [0,1]$
- $P(true)=1$, $P(false)=0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$



Sample space

Some theorems derived from the axioms

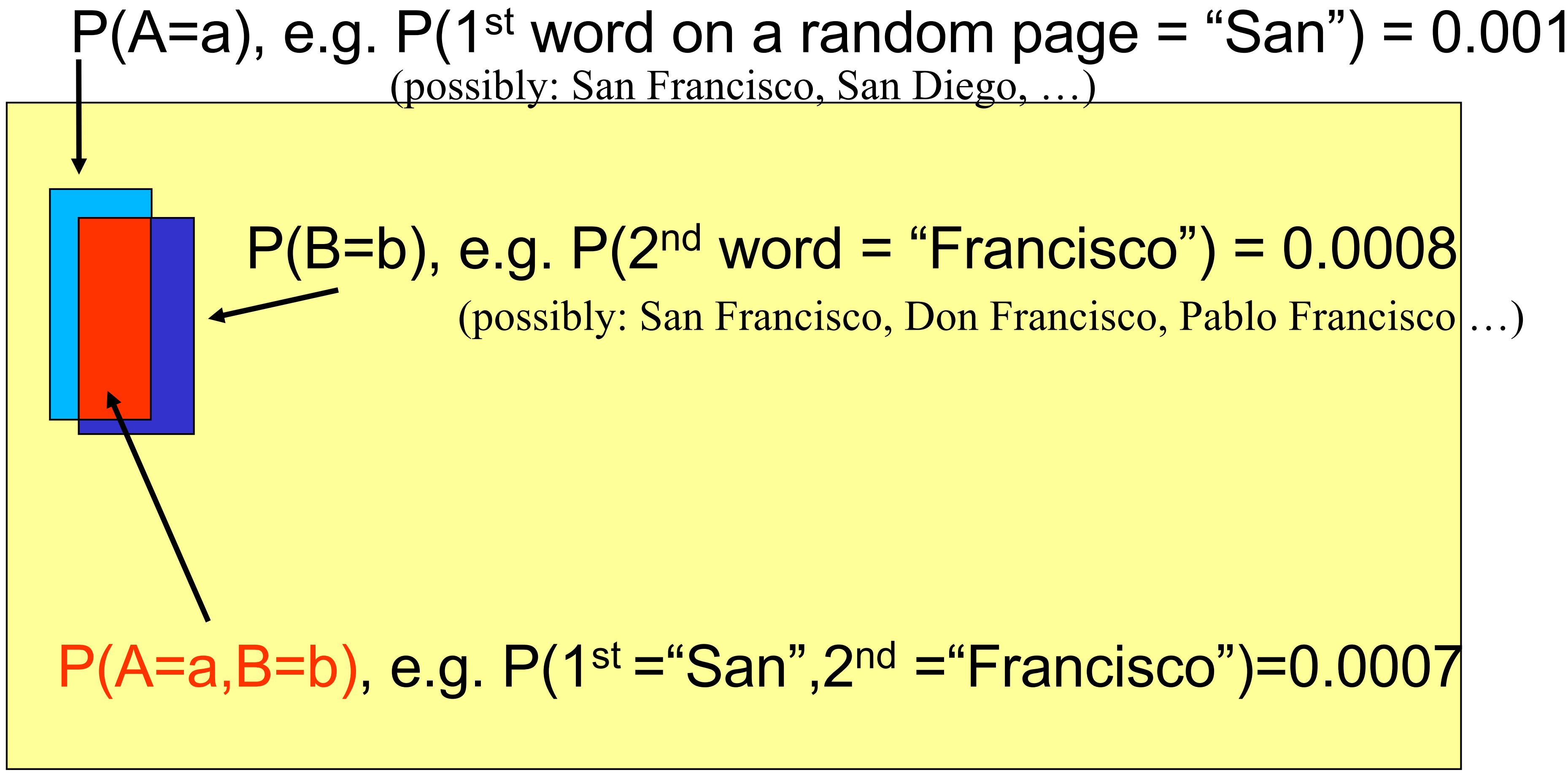
- $P(\neg A) = 1 - P(A)$ picture?
- If A can take k different values $a_1 \ldots a_k$:

  $$P(A=a_1) + \ldots P(A=a_k) = 1$$

- $P(B) = P(B \wedge \neg A) + P(B \wedge A)$, if A is a binary event
- $P(B) = \sum_{i=1\ldots k} P(B \wedge A=a_i)$, if A can take *k* values

Joint probability

- The joint probability P(A=a, B=b) is a shorthand for P(A=a ∧ B=b), the probability of both A=a and B=b happen



Joint probability table

weather

temp		Sunny	Cloudy	Rainy
	hot	150/365	40/365	5/365
	cold	50/365	60/365	60/365

- P(temp=hot, weather=rainy) = P(hot, rainy) = 5/365

- The full joint probability table between N variables, each taking k values, has k^N entries (that's a lot!)

Marginal probability

- Sum over other variables

weather

temp		Sunny	Cloudy	Rainy
	hot	150/365	40/365	5/365
	cold	50/365	60/365	60/365
	Σ	200/365	100/365	65/365

P(Weather)={200/365, 100/365, 65/365}

- The name comes from the old days when the sums are written on the margin of a page

Marginal probability

- Sum over other variables

weather

temp		Sunny	Cloudy	Rainy	Σ
	hot	150/365	40/365	5/365	195/365
	cold	50/365	60/365	60/365	170/365

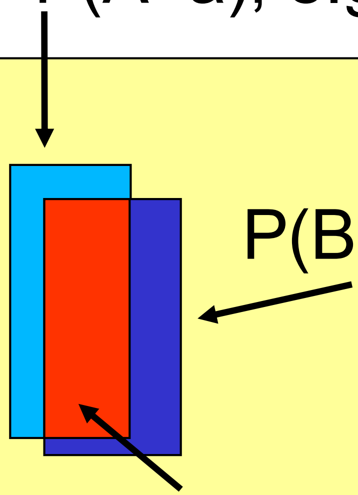
P(temp)={195/365, 170/365}

- This is nothing but $P(B) = \sum_{i=1\ldots k} P(B \wedge A=a_i)$, if A can take *k* values

Conditional probability

- The conditional probability P(A=a | B=b) is the fraction of times A=a, within the region that B=b

P(A=a), e.g. P(1st word on a random page = "San") = 0.001



P(B=b), e.g. P(2nd word = "Francisco") = 0.0008

P(A=a | B=b), e.g. P(1st="San" | 2nd ="Francisco")=**0.875**

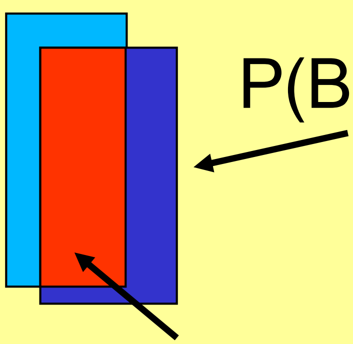
(possibly: San, Don, Pablo …)

Although "San" is rare and "Francisco" is rare, given "Francisco" then "San" is quite likely!

Conditional probability

- P(San | Francisco)
  $= \#(1^{st}=S \text{ and } 2^{nd}=F) / \#(2^{nd}=F)$
  $= P(San \wedge Francisco) / P(Francisco)$
  $= 0.0007 / 0.0008$
  $= 0.875$

P(S)=0.001
P(F)=0.0008
P(S,F)=0.0007



P(B=b), e.g. P(2^{nd} word = "Francisco") = 0.0008

P(A=a | B=b), e.g. P(1^{st}="San" | 2^{nd} ="Francisco")=**0.875**

(possibly: San, Don, Pablo …)

Conditional probability

- In general, the conditional probability is

$$P(A=a \mid B) = \frac{P(A=a, B)}{P(B)} = \frac{P(A=a, B)}{\sum_{\text{all } a_i} P(A=a_i, B)}$$

- We can have everything conditioned on some other events C, to get a conditional version of conditional probability

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

'|' has low precedence.
This should read P(A | (B,C))

The chain rule

- From the definition of conditional probability we have the chain rule

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

- It works the other way around

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

- It works with more than 2 events too

$P(A_1, A_2, \ldots, A_n) =$

$P(A_1) * P(A_2 \mid A_1) * P(A_3 \mid A_1, A_2) * \ldots * P(A_n \mid A_1, A_2 \ldots A_{n-1})$

Reasoning

How do we use probabilities in AI?

- You wake up with a headache (D'oh!).
- Do you have the flu?
- H = headache, F = flu

Logical Inference: if (H) then F. (but the world is often not this clear cut)

Statistical Inference: compute the probability of a query given (conditioned on) evidence, i.e. P(F|H)

[Example from Andrew Moore]

Inference with Bayes' rule: Example 1

Inference: compute the probability of a query given evidence
(H = headache, F = flu)

You know that

- P(H) = 0.1 "one in ten people has headache"
- P(F) = 0.01 "one in 100 people has flu"
- P(H|F) = 0.9 "90% of people who have flu have headache"

- How likely do you have the flu?
  - 0.9?
  - 0.01?
  - …?

[Example from Andrew Moore]

Inference with Bayes' rule

Bayes rule

Essay Towards Solving a Problem
in the Doctrine of Chances (1764)

$$p(F \mid H) = \frac{P(F,H)}{P(H)} = \frac{P(H \mid F)P(F)}{P(H)}$$

- P(H) = 0.1 "one in ten people has headache"
- P(F) = 0.01 "one in 100 people has flu"
- P(H|F) = 0.9 "90% of people who have flu have headache"

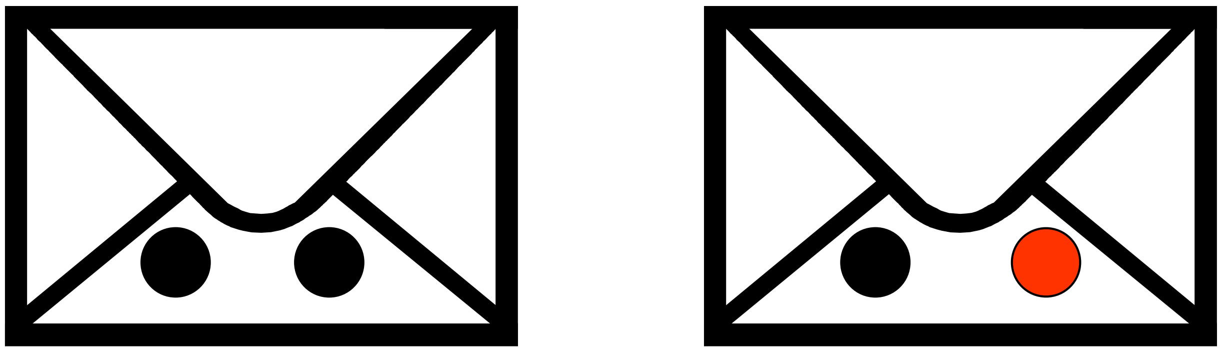
- P(F|H) = 0.9 * 0.01 / 0.1 = 0.09
- So there's a 9% chance you have flu – much less than 90%
- But it's higher than P(F)=1%, since you have the headache

Inference with Bayes' rule

- P(A|B) = P(B|A)P(A) / P(B)   Bayes' rule
- Why do we make things this complicated?
  - Often P(B|A), P(A), P(B) are easier to get
  - Some names:
    - **Prior P(A)**: probability before any evidence
    - **Likelihood P(B|A)**: assuming A, how likely is the evidence
    - **Posterior P(A|B)**: conditional prob. after knowing evidence
    - **Inference**: deriving unknown probability from known ones
- In general, if we have the full joint probability table, we can simply do P(A|B)=P(A, B) / P(B) – more on this later…

Inference with Bayes' rule: Example 2

- In a bag there are two envelopes
  - one has a red ball (worth $100) and a black ball
  - one has two black balls. Black balls worth nothing



- You randomly grabbed an envelope, randomly took out one ball – it's black.
- At this point you're given the option to switch the envelope. To switch or not to switch?

Inference with Bayes' rule: Example 2

- E: envelope, 1=(R,B), 2=(B,B)
- B: the event of drawing a black ball
- $P(E|B) = P(B|E)*P(E) / P(B)$
- We want to compare $P(E=1|B)$ vs. $P(E=2|B)$
- $P(B|E=1) = 0.5$, $P(B|E=2) = 1$
- $P(E=1)=P(E=2)=0.5$
- $P(B)=3/4$ (it in fact doesn't matter for the comparison)
- $P(E=1|B)=1/3$, $P(E=2|B)=2/3$
- After seeing a black ball, the posterior probability of this envelope being 1 (thus worth $100) is smaller than it being 2
- Thus you should switch

Independence

- Two events A, B are independent, if (the following are equivalent)
  - P(A, B) = P(A) * P(B)
  - P(A | B) = P(A)
  - P(B | A) = P(B)
- For a 4-sided die, let
  - A=outcome is small
  - B=outcome is even
  - Are A and B independent?
- How about a 6-sided die?

Independence

- Independence is a domain knowledge
- If A, B are independent, the joint probability table between A, B is simple:
  - it has k^2 cells, but only 2k-2 parameters. This is good news – more on this later…
- Example: P(burglary)=0.001, P(earthquake)=0.002. Let's say they are independent. The full joint probability table=?

Independence misused

A famous statistician would never travel by airplane, because he had studied air travel and estimated that the probability of there being a bomb on any given flight was one in a million, and he was not prepared to accept these odds.

One day, a colleague met him at a conference far from home. "How did you get here, by train?"

"No, I flew"

"What about the possibility of a bomb?"

"Well, I began thinking that if the odds of one bomb are 1:million, then the odds of two bombs are (1/1,000,000) x (1/1,000,000). This is a very, very small probability, which I can accept. So now I bring my own bomb along!"

An innocent old math joke

Conditional independence

- Random variables can be dependent, but conditionally independent
- Your house has an alarm
  - Neighbor John will call when he hears the alarm
  - Neighbor Mary will call when she hears the alarm
  - Assume John and Mary don't talk to each other
- JohnCall independent of MaryCall?
  - No – If John called, likely the alarm went off, which increases the probability of Mary calling
  - $P(MaryCall \mid JohnCall) \neq P(MaryCall)$

Conditional independence

- If we know the status of the alarm, JohnCall won't affect Mary at all

  P(MaryCall | Alarm, JohnCall) = P(MaryCall | Alarm)

- We say JohnCall and MaryCall are conditionally independent, given Alarm
- In general A, B are conditionally independent given C
  - if P(A | B, C) = P(A | C), or
  - P(B | A, C) = P(B | C), or
  - P(A, B | C) = P(A | C) * P(B | C)