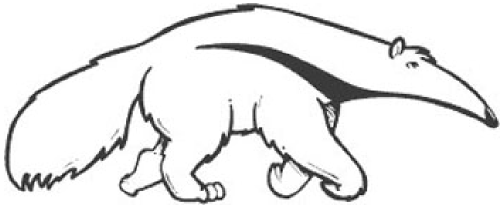
Probability:
Reasoning Under Uncertainty

CS171, Fall Quarter, 2019
Introduction to Artificial Intelligence
Prof. Richard Lathrop



Read Beforehand: R&N 13

Outline

- Representing uncertainty is useful in knowledge bases
  - Probability provides a coherent framework for uncertainty
- Review of basic concepts in probability
  - Emphasis on conditional probability & conditional independence
- Full joint distributions are intractable to work with
  - Conditional independence assumptions allow much simpler models
- Bayesian networks (next lecture)
  - A useful type of structured probability distribution
  - Exploit structure for parsimony, computational efficiency
- **Rational** agents **cannot** violate probability theory

Uncertainty

Let action *At* = leave for airport *t* minutes before flight
Will *At* get me there on time?

Problems:

1. partial observability (road state, etc.)
2. multi-agent problem (other drivers' plans)
3. noisy sensors (uncertain traffic reports)
4. uncertainty in action outcomes (flat tire, etc.)
5. immense complexity of modeling and predicting traffic

Hence a purely logical approach either

1. risks falsehood: "*A25* will get me there on time", or
2. leads to conclusions that are too weak for decision making:

"*A25* will get me there on time if there's no accident on the bridge and it doesn't rain and my tires remain intact, etc., etc."

"*A1440* should get me there on time but I'd have to stay overnight in the airport."

Uncertainty in the world

- Uncertainty due to
  - Randomness
  - Overwhelming complexity
  - Lack of knowledge
  - …
- Probability gives
  - natural way to describe our assumptions
  - rules for how to combine information
- Subjective probability
  - Relate to agent's own state of knowledge: $P(A25|\text{no accidents}) = 0.05$
  - Not assertions about the world; indicate degrees of belief
  - Change with new evidence: $P(A25 \mid \text{no accidents, 5am}) = 0.20$

Propositional Logic and Probability

- Their ontological commitments are the same
  - The world is a set of facts that do or do not hold

Ontology is the philosophical study of the nature of being, becoming, existence, or reality; **what exists in the world?**

- Their epistemological commitments differ
  - **Logic agent** believes true, false, or no opinion
  - **Probabilistic agent** has a numerical degree of belief between 0 (false) and 1 (true)

Epistemology is the philosophical study of the nature and scope of knowledge; **how, and in what way, do we know about the world?**

Making decisions under uncertainty

- Suppose I believe the following:
  - P(A25 gets me there on time | …) = 0.04
  - P(A90 gets me there on time | …) = 0.70
  - P(A120 gets me there on time | …) = 0.95
  - P(A1440 gets me there on time | …) = 0.9999

- Which action to choose?
- Depends on my preferences for missing flight vs. time spent waiting, etc.
  - Utility theory is used to represent and infer preferences
  - Decision theory= probability theory + utility theory

- Expected utility of action a in state s

$$= \sum_{\text{outcome in Results}(s,a)} P(\text{outcome}) * \text{Utility}(\text{outcome})$$

- A rational agent acts to maximize expected utility

Example: Airport

- Suppose I believe the following:
  - P(A25 gets me there on time | …) = 0.04
  - P(A90 gets me there on time | …) = 0.70
  - P(A120 gets me there on time | …) = 0.95
  - P(A1440 gets me there on time | …) = 0.9999
  - Utility(on time) = $1,000
  - Utility(not on time) = −$10,000
- Expected utility of action a in state s

$$= \sum_{\text{outcome in Results}(s,a)} P(\text{outcome}) * \text{Utility}(\text{outcome})$$

E(Utility(A25)) = 0.04*$1,000 + 0.96*(−$10,000) = −$9,560
E(Utility(A90)) = 0.7*$1,000 + 0.3*(−$10,000) = −$2,300
E(Utility(A120)) = 0.95*$1,000 + 0.05*(−$10,000) = $450
E(Utility(A1440)) = 0.9999*$1,000 + 0.0001*(−$10,000) = $998.90

- Have not yet accounted for disutility of staying overnight at the airport, etc.

Random variables

- **Random Variable**:
  - Basic element of probability assertions
  - Similar to CSP variable, but values reflect probabilities not constraints.
    - Variable: A
    - Domain: $\{a_1, a_2, a_3\}$ <-- events / outcomes

- Types of Random Variables:
  - Boolean random variables : *{ true, false }*
    - e.g., *Cavity (= do I have a cavity?)*
  - Discrete random variables : one value from a set of values
    - e.g., *Weather is one of {sunny, rainy, cloudy ,snow}*
  - Continuous random variables : a value from within constraints
    - *e.g., Current temperature is bounded by (10°, 200°)*

- Domain values must be exhaustive and mutually exclusive:
  - One of the values must always be the case (**Exhaustive**)
  - Two of the values cannot both be the case (**Mutually Exclusive**)

Random variables

- **Example**: Coin flip
  - Variable = R, the result of the coin flip
  - Domain = {heads, tails, edge} <-- must be exhaustive
  - P(R = heads) = 0.4999 }
  - P(R = tails) = 0.4999 } <-- must be exclusive
  - P(R = edge) = 0.0002 }

- Shorthand is often used for simplicity:
  - <u>Upper-case letters for variables, lower-case letters for values.</u>
  - E.g., P(A) ≡ <P(A=a1), P(A=a2), …, P(A=an)> for all n values in Domain(A)
    - Note: P(A) is a <u>vector</u> giving the probability that A takes on each of its n values in Domain (A)
  - E.g., P(a) ≡ P(A = a)
    P(a|b) ≡ P(A = a | B = b)
    P(a, b) ≡ P(A = a ∧ B = b)

- Two kinds of probability propositions:
  - **Elementary propositions** are an assignment of a value to a random variable:
    - e.g., *Weather = sunny;* e.g., *Cavity = false* (abbreviated as ¬*cavity*)
  - **Complex propositions** are formed from elementary propositions and standard logical connectives :
    - e.g., *Cavity = false* ∨ *Weather = sunny*

Probability

- P(a) is the probability of proposition "a"
  - E.g., P(it will rain in London tomorrow)
  - The proposition "a" is actually true or false in the real world
  - P(a) is our <u>degree of belief</u> that proposition "a" is true in the real world
  - P(a) = "prior" or marginal or unconditional probability
  - Assumes no other information is available

- Axioms of probability:
  - 0 <= P(a) <= 1
  - P(NOT(a)) = 1 – P(a)
  - P(true) = 1
  - P(false) = 0
  - P(a OR b) = P(a) + P(b) – P(a AND b)

- Any agent that holds degrees of beliefs that contradict these axioms will act sub-optimally in some cases
  - e.g., de Finetti (R&N pp. 489-490) proved that there will be some combination of bets that forces such an unhappy agent to lose money every time.

- **Rational agents <u>cannot</u> violate probability theory.**

Interpretations of probability

- **Relative Frequency**: *Usually taught in school*
  - P(*a*) represents the frequency that event *a* will happen in repeated trials.
  - Requires event *a* to have happened enough times for data to be collected.

- **Degree of Belief**: *A more general view of probability*
  - P(*a*) represents an agent's degree of belief that event *a* is true.
  - Can predict probabilities of events that occur rarely or have not yet occurred.
  - Does not require new or different rules, just a different interpretation.

- Examples:
  - a = "life exists on another planet"
    - What is P(a)? We all will assign different probabilities
  - a = "California will secede from the US"
    - What is P(a)?
  - a = "over 50% of the students in this class will get A's"
    - What is P(a)?

Concepts of probability

- **Unconditional Probability**
  - **P(a)**, the probability of "a" being true, or **P(a=True)**
  - Does not depend on anything else to be true (**unconditional**)
  - Represents the probability prior to further information that may adjust it (**prior**)
  - Also sometimes "**marginal**" probability (vs. joint probability)

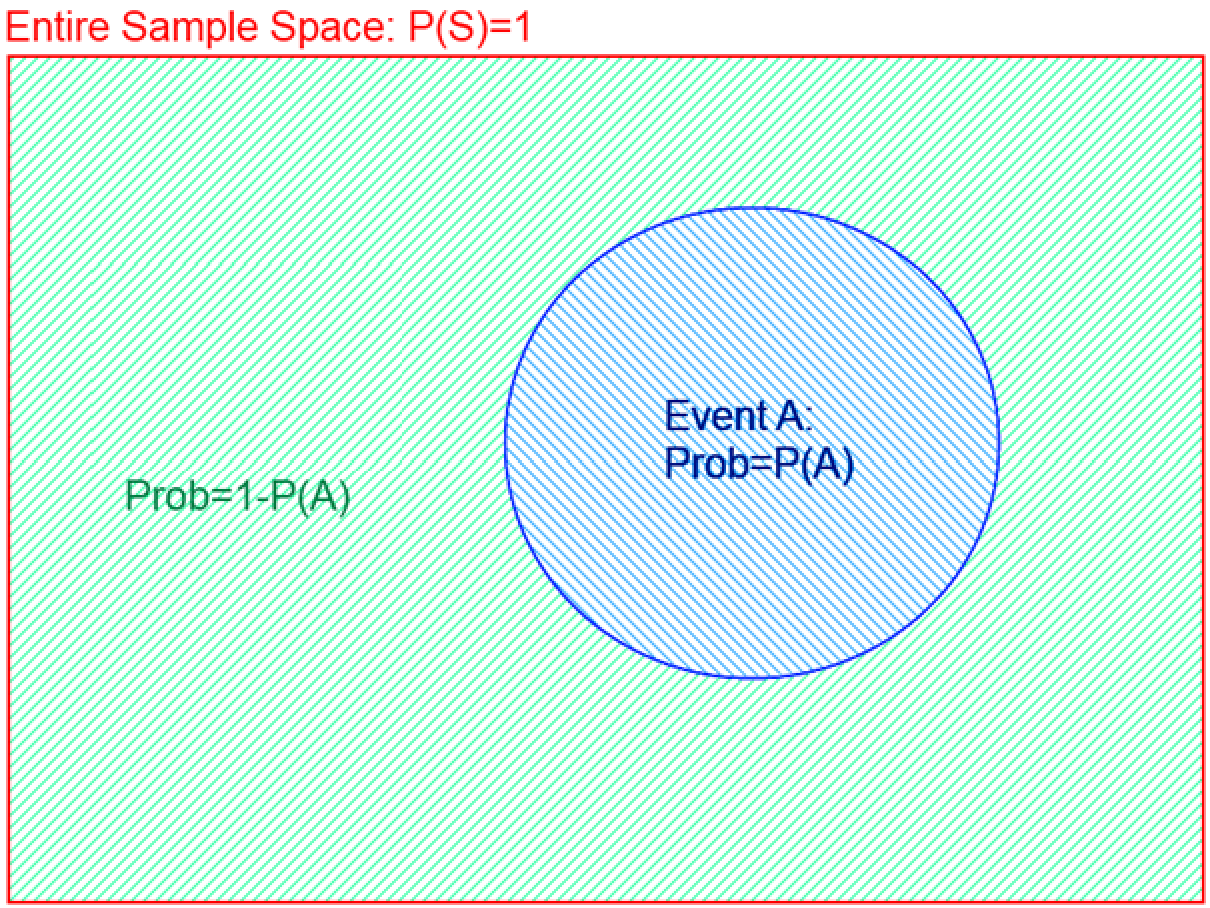
- **Conditional Probability**
  - **P(a|b)**, the probability of "a" being true, given that "b" is true
  - Relies on "b" = true (**conditional**)
  - Represents the prior probability adjusted based upon new information "b" (**posterior**)
  - Can be generalized to more than 2 random variables:
    - e.g. P(a|b, c, d)

We often use comma to abbreviate AND.

- **Joint Probability**
  - **P(a, b) = P(a ∧ b)**, the probability of "a" and "b" both being true
  - Can be generalized to more than 2 random variables:
    - e.g. P(a, b, c, d)

Probability Space

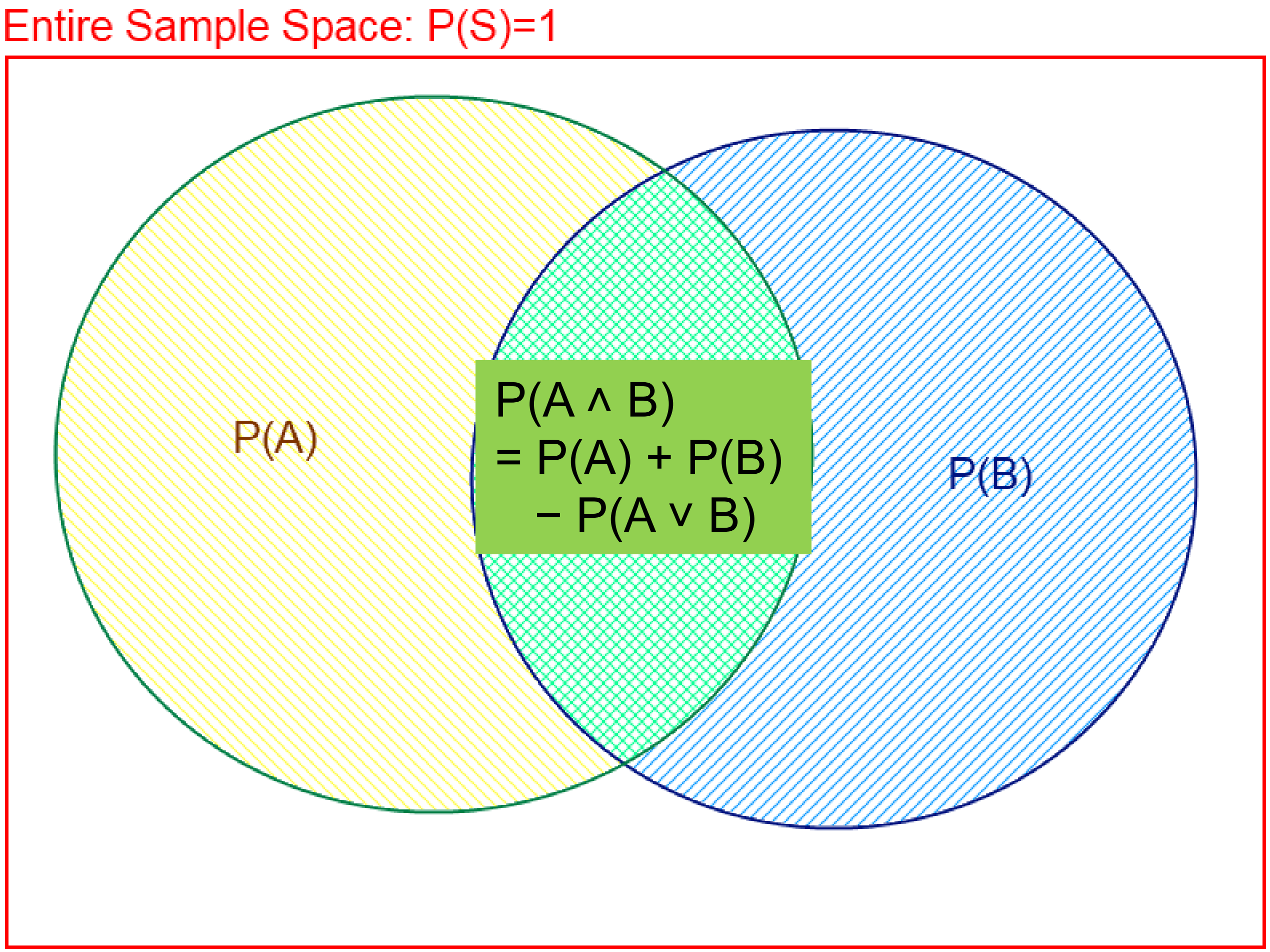
$$P(A) + P(\neg A) = 1$$



Area = Probability of Event

AND Probability

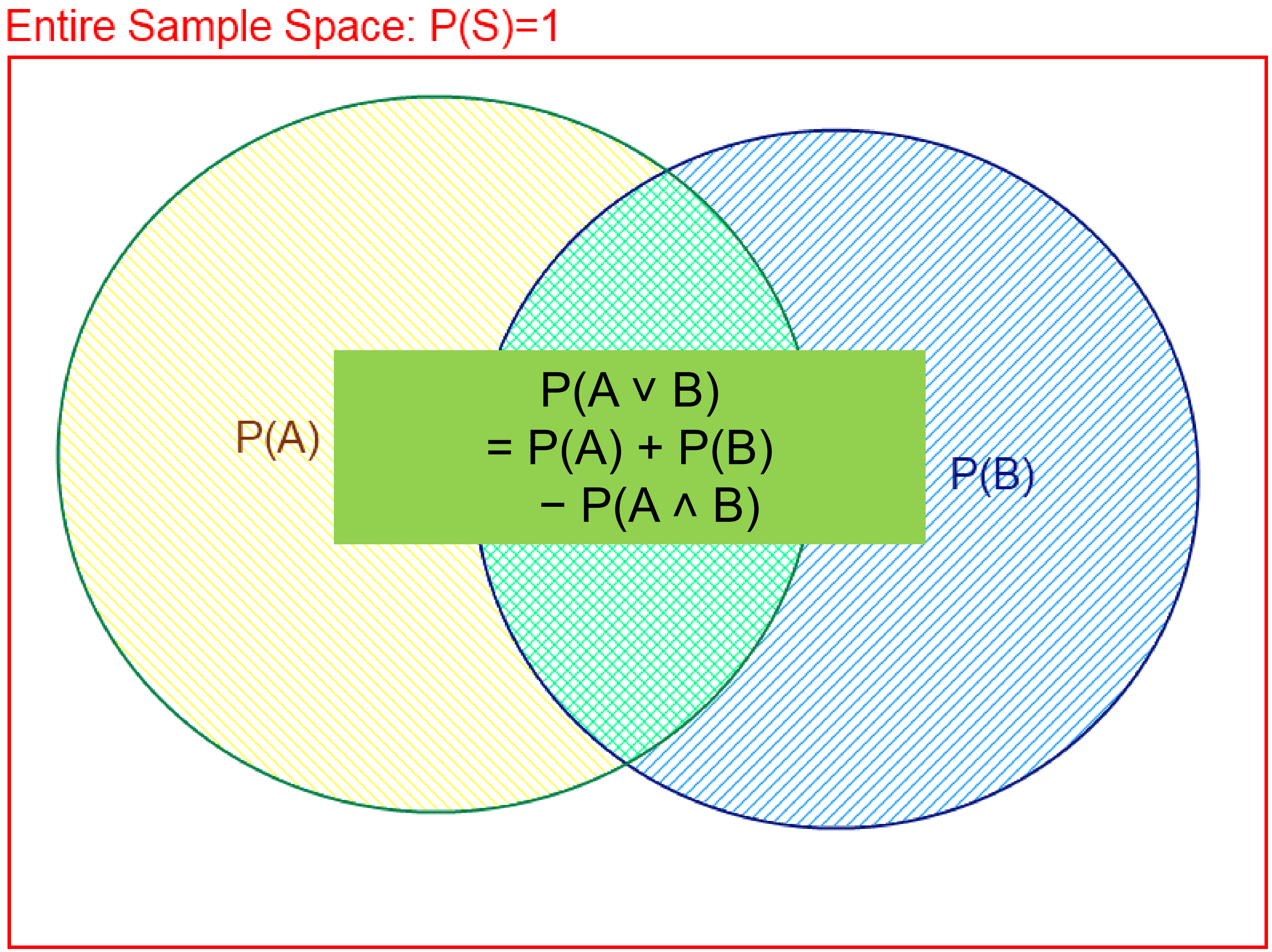
$$P(A, B) = P(A \wedge B) = P(A) + P(B) - P(A \vee B)$$



Area = Probability of Event

OR Probability

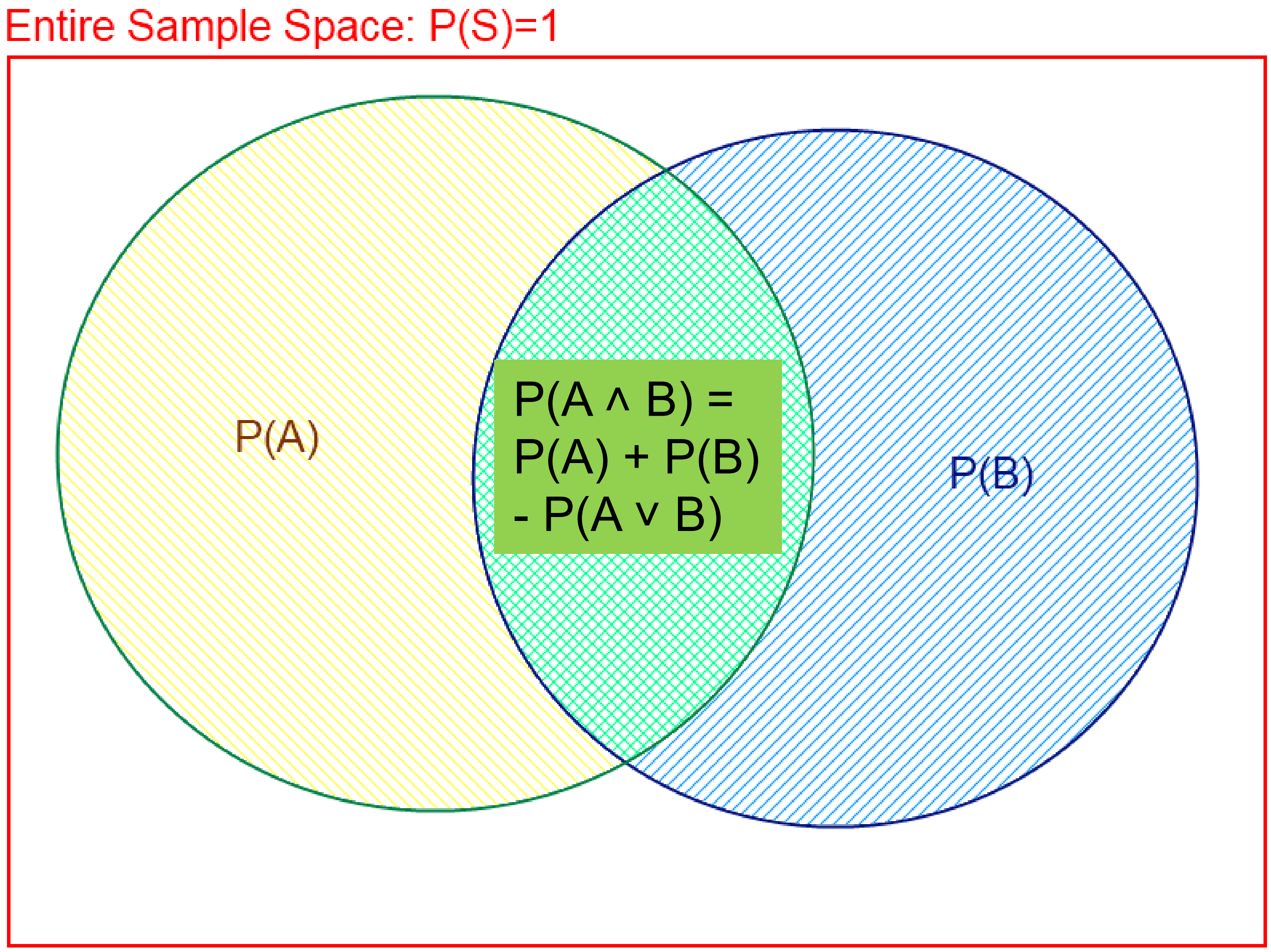
$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$



Area = Probability of Event

Conditional Probability

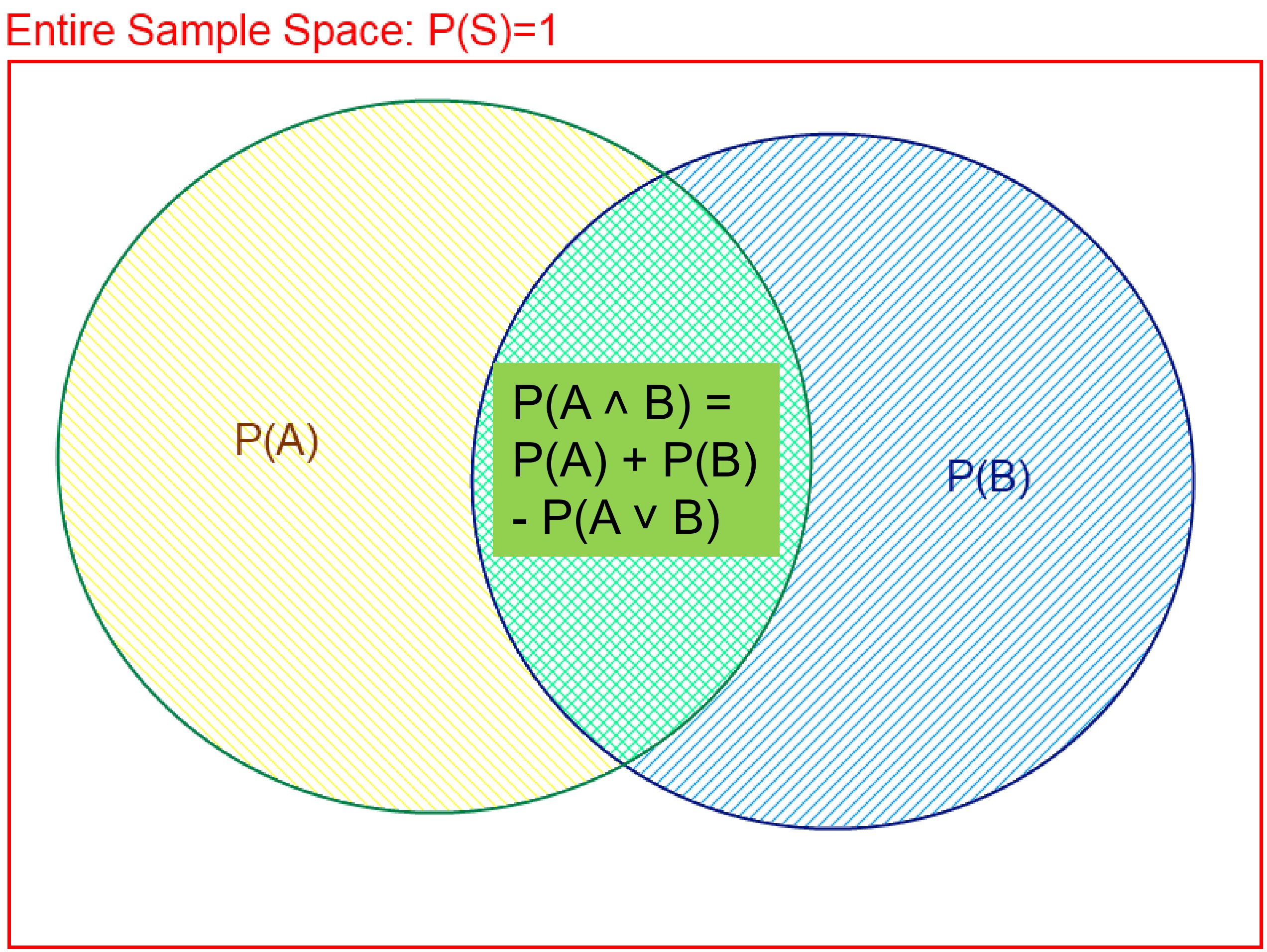
$$P(A \mid B) = P(A, B) / P(B) = P(A \wedge B) / P(B)$$



Area = Probability of Event

Product Rule

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



Area = Probability of Event

Using the Product Rule

- **Applies to any number of variables:**
  - P(a, b, c) = P(a, b|c) P(c) = P(a|b, c) P(b, c)
  - P(a, b, c|d, e) = P(a|b, c, d, e) P(b, c|d, e)

- **Factoring**: (AKA **Chain Rule** for probabilities)
  - By the product rule, we can always write:

    P(a, b, c, … y, z) = P(a | b, c, … y, z) P(b, c, … y, z)

  - Repeating this idea, we can completely factor P(a, b, …, z):

    P(a, b, c, … y, z)

    = P(a | b, c, … y, z) P(b | c, ... y, z) P(c| ... y, z) ... P(y|z)P(z)

  - These relationships hold for any ordering of the variables

We often use comma to abbreviate AND.

Examples of complete Factoring Using the Product Rule (can use any variable ordering)

- $P(a, b) = P(a|b)(b)$

- $P(a, b, c) = P(a|b, c)P(b, c)$

  $= P(a|b, c)P(b|c)P(c)$ <= complete factoring

- $P(a, b, c, d) = P(a|b, c, d)P(b, c, d)$

  $= P(a|b, c, d)P(b|c, d)P(c, d)$

  $= P(a|b, c, d)P(b|c, d)P(c| d)P(d)$ <= complete factoring

- $P(a, b, c, d, e) = P(a|b, c, d, e)P(b, c, d, e)$

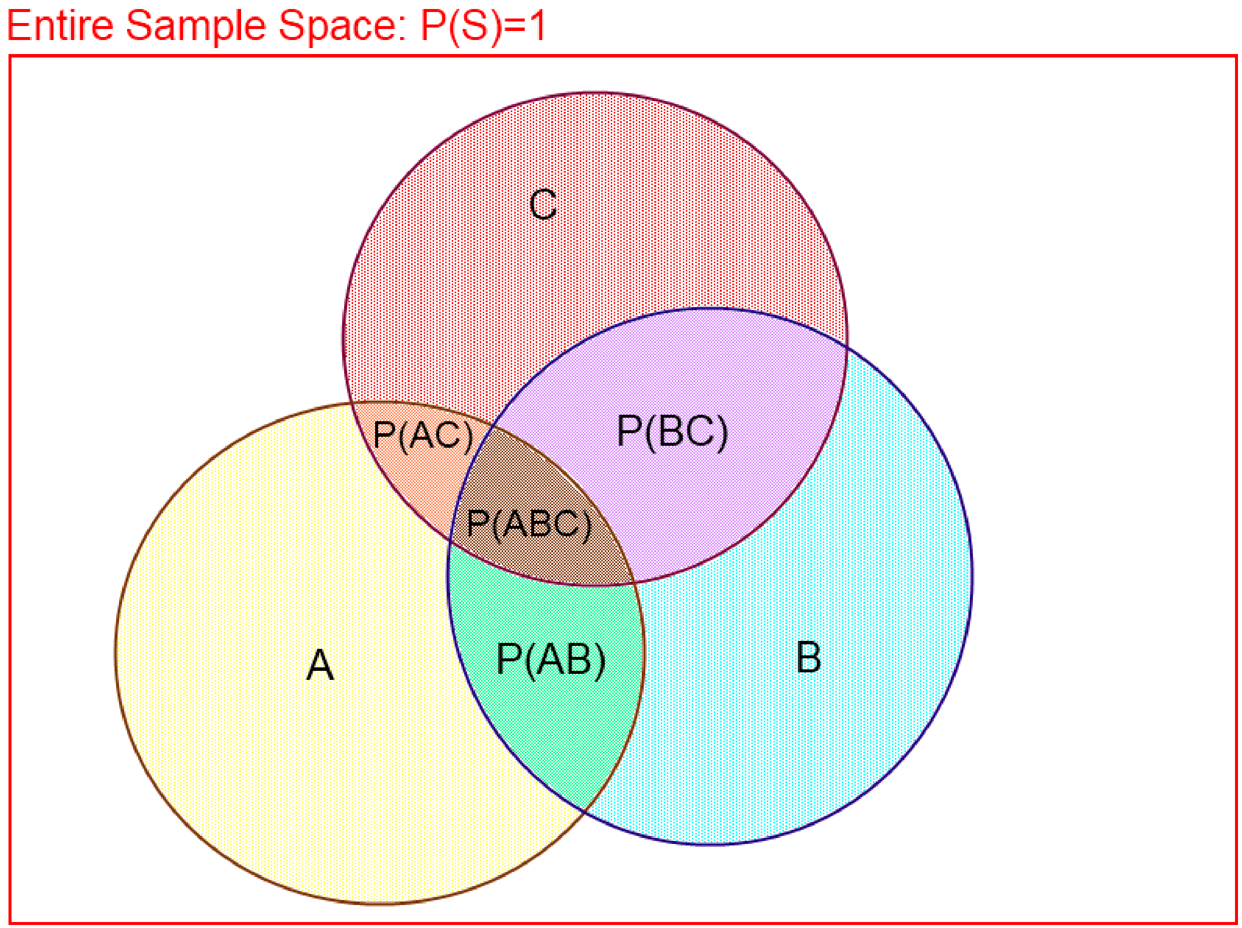
  $= P(a|b, c, d, e)P(b|c, d, e)P(c, d, e)$

  $= P(a|b, c, d, e)P(b|c, d, e)P(c| d, e)P(d, e)$

  $= P(a|b, c, d, e)P(b|c, d, e)P(c| d, e)P(d|e)P(e)$ <= complete

Sum Rule

$$P(A) = \Sigma_{B,C} P(A,B,C) = \Sigma_{b \in B, c \in C} P(A,b,c)$$



Area = Probability of Event

Using the Sum Rule

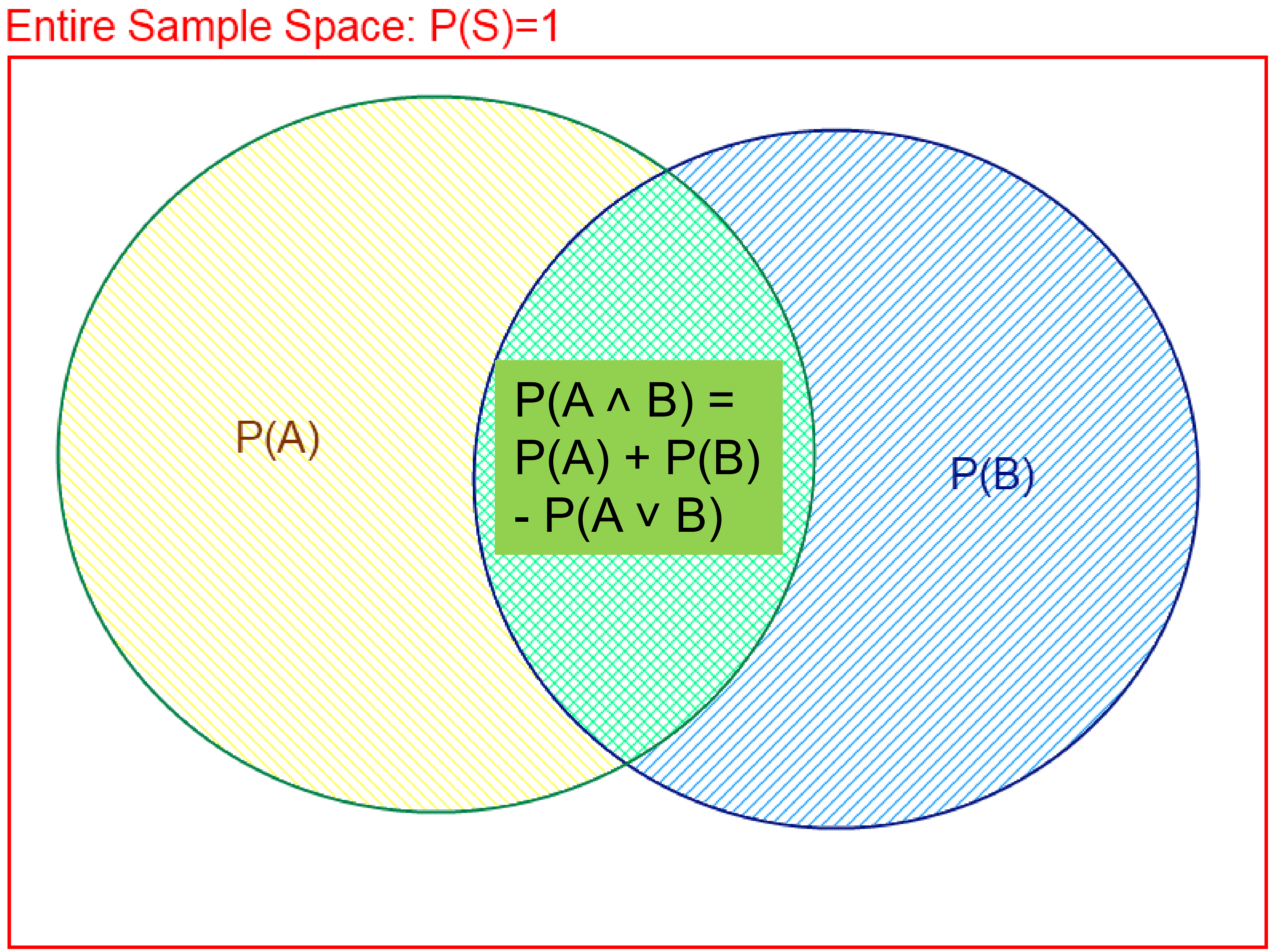
- We can marginalize variables out of any joint distribution by simply summing over that variable:
  - $P(b) = \Sigma_a \Sigma_c \Sigma_d P(a, b, c, d)$
  - $P(a, d) = \Sigma_b \Sigma_c P(a, b, c, d)$

We often use comma to abbreviate AND.

- **For Example**: Determine probability of catching a fish
  - Given a set of probabilities *P(CatchFish, Day, Lake)*
  - Where:
    - *CatchFish =* *{true, false}*
    - *Day =* *{mon, tues, wed, thurs, fri, sat, sun}*
    - *Lake =* *{blue lake, ralph lake, crystal lake}*
  - Need to find P(CatchFish = True):
    - $P(CatchFish = true) = \Sigma_{day} \Sigma_{lake} P(CatchFish = true, day, lake)$

Bayes' Rule

P(B|A) = P(A|B) P(B) / P(A)



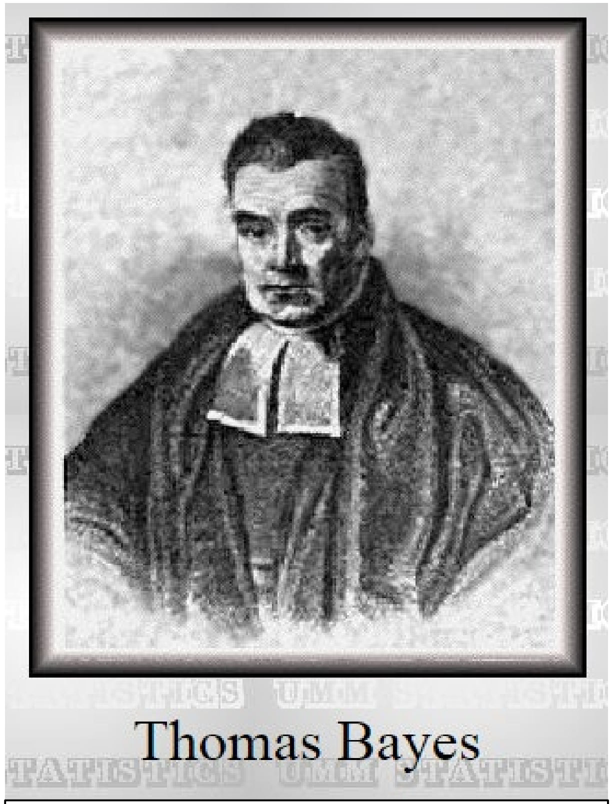
Area = Probability of Event

Derivation of Bayes' Rule

- **Start from Product Rule:**
  - $P(a, b) = P(a|b)\, P(b) = P(b|a)\, P(a)$
- **Isolate Equality on Right Side:**
  - $P(a|b)\, P(b) = P(b|a)\, P(a)$
- **Divide through by P(b):**
  - $P(a|b) = P(b|a)\, P(a) / P(b)$ <-- Bayes' Rule
- **"Bayes' rule underlies most modern approaches to uncertain reasoning in AI systems."** — R&N p. 9

Who’s Bayes?

- Reverend Thomas Bayes (c. 1701 – 1761) was an English minister and mathematician. **His ideas have created much controversy and debate among statisticians**….
- The paper that describes Bayes’ Theorem (or Bayes’ Rule) was discovered in his office after his death. Allegedly, he was trying to prove the existence of God by mathematics; though this is not certain and other motives also are alleged. His paper was sent to the Royal Society with a note, “Some of your members may be interested in this.” It was published by, and read to, the Royal Society. **Nowadays, it has given rise to an immense body of statistical and probabilistic work.**



Portrait purportedly of Bayes used in a 1936 book, but it is doubtful the portrait is actually of him. No earlier claimed portrait survives.

Summary of probability rules

- **Product Rule**: (aka Chain Rule)
  - **P(a, b) = P(a|b) P(b) = P(b|a) P(a)** Probability of "a" and "b" occurring is the same as probability of "a" occurring given "b" is true, times the probability of "b" occurring.
  - e.g., *P(rain, cloudy) = P(rain | cloudy) * P(cloudy)*
  - **P(a, b, c, …, y, z) = P(a|b, c, …, y, z) P(b|c, …, y, z) … P(y|z)P(z)**

- **Sum Rule**: (aka Law of Total Probability)
  - **P(a) = Σ_b P(a, b) = Σ_b P(a|b) P(b),** where B is any random variable
  - Probability of "a" occurring is the same as the sum of all joint probabilities including the event, provided the joint probabilities represent all possible events.
  - Can be used to "marginalize" out other variables from probabilities, resulting in prior probabilities also being called marginal probabilities.
    - e.g., *P(rain) = $\Sigma_{Windspeed}$ P(rain, Windspeed)*
      where Windspeed = {0-10mph, 10-20mph, 20-30mph, etc.}

- **Bayes' Rule**:
  - **P(b|a) = P(a|b) P(b) / P(a)**
  - Acquired from rearranging the product rule.
  - Allows conversion between conditionals, from P(b|a) to P(a|b).
    - e.g., b = disease, a = symptoms
      More natural to encode knowledge as P(a|b) than as P(b|a).

Full Joint Distribution

- We can fully specify a probability space by a full joint distribution:

  - A full joint distribution contains a probability for every possible combination of variable values. This requires:

    Π_{vars} ($\mathbf{n}_{var}$) probabilities

    where $\mathbf{n}_{var}$ is the number of values in the domain of variable **var**

  - E.g. P(A, B, C), where A,B,C have 4 values each; Full joint distribution specified by 4^3 values = 64 values

T	D	C	P(T,D,C)
0	0	0	0.576
0	0	1	0.008
0	1	0	0.144
0	1	1	0.072
1	0	0	0.064
1	0	1	0.012
1	1	0	0.016
1	1	1	0.108

  - For **n** variables each with **m** values, requires $\mathbf{m}^{\mathbf{n}}$ probabilities
  - E.g., a realistic problem of 100 Boolean variables requires > 10^{30} probabilities (intractable)

- Using a full joint distribution, we can use the product rule, sum rule, and Bayes' rule to create any combination of joint, marginal, and conditional probabilities.

Marginal Probability

- Can fully specify a probability space by constructing a full joint distribution
- Example: dentist
  - T: have a toothache
  - D: dental probe catches
  - C: have a cavity
- Joint distribution
  - Assigns each event (T=t, D=d, C=c) a probability
  - Probabilities sum to 1.0
- Law of total probability:

$$p(C=1) = \sum_{t,d} P(T=t, D=d, C=1)$$

= 0.008 + 0.072 + 0.012 + 0.108 = 0.20

  - *Some* value of (T,D) must occur; values are disjoint
  - "Marginal probability" of C; "marginalize" or "sum over" T,D
  - Early actuaries wrote row & column totals in their probability table margins

T	D	C	P(T,D,C)
0	0	0	0.576
0	0	1	0.008
0	1	0	0.144
0	1	1	0.072
1	0	0	0.064
1	0	1	0.012
1	1	0	0.016
1	1	1	0.108

The effect of evidence

Example from Russell & Norvig

- Example: dentist
  - T: have a toothache
  - D: dental probe catches
  - C: have a cavity

T	D	C	P(T,D,C)
0	0	0	0.576
0	0	1	0.008
0	1	0	0.144
0	1	1	0.072
1	0	0	0.064
1	0	1	0.012
1	1	0	0.016
1	1	1	0.108

- Recall p(C=1) = 0.20
- Suppose we observe D=0, T=0?

$$p(C=1|D=0,T=0) = \frac{p(C=1,D=0,T=0)}{p(D=0,T=0)}$$

$$= \frac{0.008}{0.576 + 0.008} = 0.012$$

Called *posterior probabilities* or *conditional probabilities*

- Observe D=1, T=1?

$$p(C=1 \mid D=1, T=1) = \frac{0.108}{0.016 + 0.108} = 0.871$$

T	D	C	P(T,D,C)
0	0	0	0.576
0	0	1	0.008
0	1	0	0.144
0	1	1	0.072
1	0	0	0.064
1	0	1	0.012
1	1	0	0.016
1	1	1	0.108

The effect of evidence

- Example: dentist
  - T: have a toothache
  - D: dental probe catches
  - C: have a cavity

T	D	C	P(T,D,C)
0	0	0	0.576
0	0	1	0.008
0	1	0	0.144
0	1	1	0.072
1	0	0	0.064
1	0	1	0.012
1	1	0	0.016
1	1	1	0.108

- Combining these rules:

$$p(C = 1 | T = 1) = \frac{p(C = 1, T = 1)}{p(T = 1)}$$

$$= \frac{0.012 + 0.108}{0.064 + 0.012 + 0.016 + 0.108} = 0.60$$

$$p(T = 1) = 0.20$$

Called the *probability of evidence*

Computing posteriors

- Sometimes it is easiest to normalize last

$$p(C|T=1) = \frac{1}{p(T=1)}\, p(C, T=1) \propto p(C, T=1) = \sum_d p(C, d, T=1)$$

T	D	C	P(T,D,C)
0	0	0	0.576
0	0	1	0.008
0	1	0	0.144
0	1	1	0.072
1	0	0	0.064
1	0	1	0.012
1	1	0	0.016
1	1	1	0.108

Assign T=1

D	C	F(D,C)
0	0	0.064
0	1	0.012
1	0	0.016
1	1	0.108

Sum over D

C	G(C)
0	0.08
1	0.120

Normalize

C	P(C|T=1)
0	0.40
1	0.60

- The normalizing constant α is used to abbreviate normalization

$$p(C|T=1) = \alpha\, \Sigma_d\, p(C, d, T=1) = \Sigma_d\, p(C, d, T=1) \,/\, p(T=1)$$

Independence

- X, Y independent:
  - p(X=x,Y=y) = p(X=x) p(Y=y)   for all x,y
  - Shorthand: p(X,Y) = P(X) P(Y)
  - Equivalent: p(X|Y) = p(X)   or p(Y|X) = p(Y)   (if p(Y), p(X) > 0)
  - Intuition: knowing X has no information about Y (or vice versa)

Independent probability distributions:
P(A,B,C) = P(A) * P(B) * P(C)

A	P(A)
0	0.4
1	0.6

B	P(B)
0	0.7
1	0.3

C	P(C)
0	0.1
1	0.9

Joint:

A	B	C	P(A,B,C)
0	0	0	.4 * .7 * .1 = .028
0	0	1	.4 * .7 * .9 = .252
0	1	0	.4 * .3 * .1 = .012
0	1	1	.4 * .3 * .9 = .108
1	0	0	.6 * .7 * .1 = .042
1	0	1	.6 * .7 * .9 = .378
1	1	0	.6 * .3 * .1 = .018
1	1	1	.6 * .3 * .9 = .162

We may omit leading zeroes to save space and effort.

This property can **greatly** reduce representation size!

Note: it is hard to “read” independence from the joint distribution.
We can “test” for it, but to do so requires a number of tests equal to the size of the joint distribution.

Conditional Independence

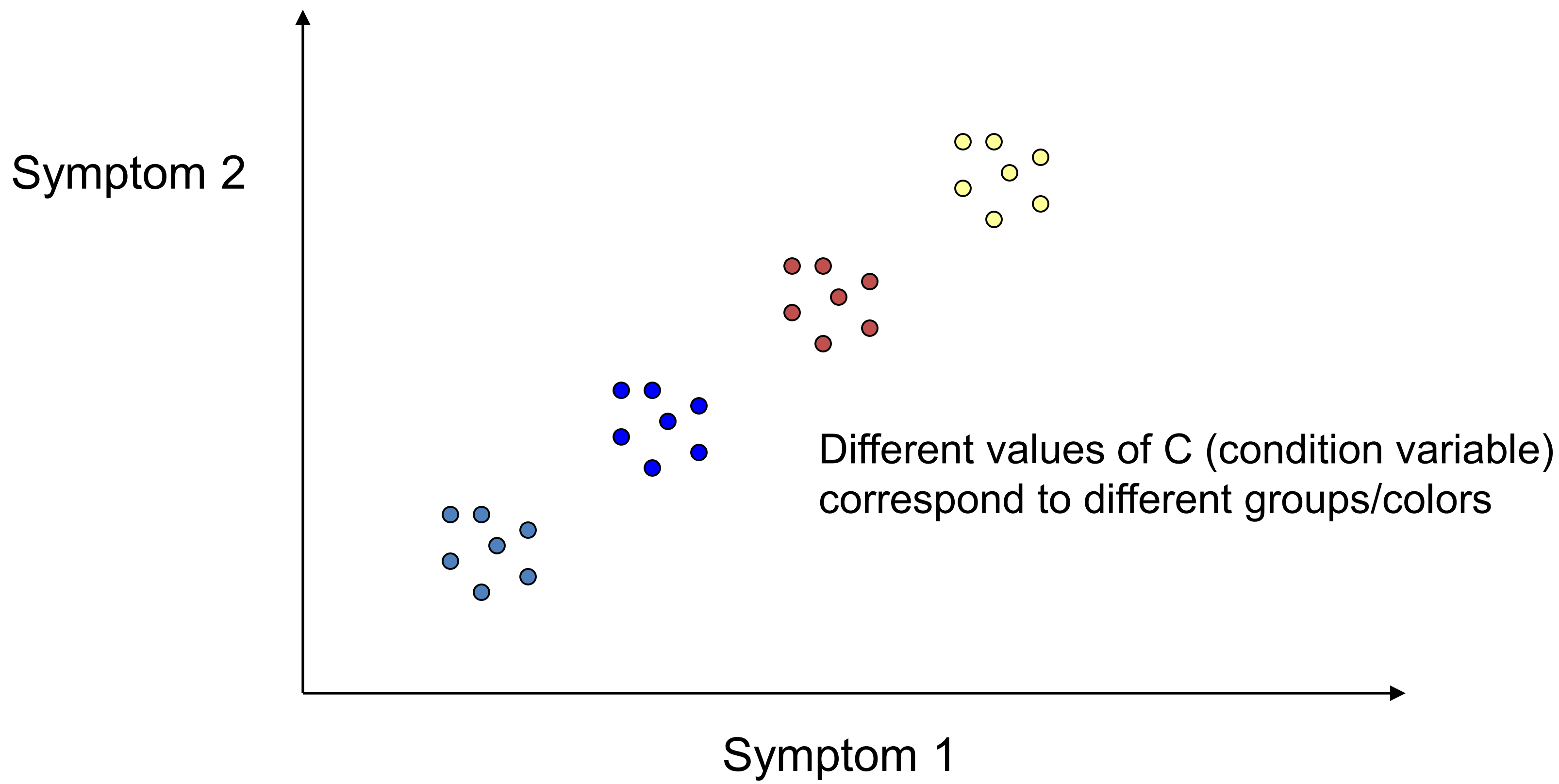
- X, Y independent given Z
  - $p(X=x,Y=y|Z=z) = p(X=x|Z=z)\ p(Y=y|Z=z)$ for all x,y,z
  - Equivalent: $p(X|Y,Z) = p(X|Z)$ or $p(Y|X,Z) = p(Y|Z)$ (if all > 0)
  - Intuition: X has no additional info about Y beyond Z's

- Example

X = height p(height|reading, age) = p(height|age)
Y = reading ability p(reading|height, age) = p(reading|age)
Z = age

Height and reading ability are dependent (not independent), but are conditionally independent given age

Conditional Independence



Symptom 1 and symptom 2 are <u>conditionally independent, given group.</u>

But clearly, symptom 1 and 2 are <u>marginally dependent, unconditionally.</u>

Conditional Independence Example:

- X, Y independent given Z
  - $p(X=x,Y=y|Z=z) = p(X=x|Z=z)\, p(Y=y|Z=z)$ for all x,y,z
- A box contains two coins: one regular coin, P(heads) = .5, and one fake two-headed coin, P(heads)=1. I choose a coin at random and toss it twice. Define the following events.
  - A = First coin toss results in heads
  - B= Second coin toss results in heads
  - C= Coin 1 (regular) has been selected
- $P(A \wedge B) = 5/8 \neq P(A)\, P(B) = 9/16$, so *A* and *B* are not independent
  - Event A makes it more likely I selected the two-headed coin, which makes Event B more likely. <u>Knowing Event A gives information about Event B.</u>
- $P(A \wedge B | C) = 1/4 = P(A|C)\, P(B|C)$, so *A* and *B* are independent given C
  - <u>Given C, knowing Event A gives **no** information about Event B.</u>

Conditional Independence Example:

- X, Y independent given Z
  - $p(X=x,Y=y|Z=z) = p(X=x|Z=z)\, p(Y=y|Z=z)$ for all x,y,z

- Consider two brothers John and Joseph, both having a genetic disease. These two events are dependent as they are brothers.

- However, given the condition that Joseph is an adopted son of the family makes the events conditionally independent.

Conditional Independence Example:

- X, Y independent given Z
  - $p(X=x,Y=y|Z=z) = p(X=x|Z=z)\ p(Y=y|Z=z)$ for all x,y,z

- Rain causes both increased umbrella usage and worsened road conditions. These events are not independent because seeing lots of umbrellas makes worsened road conditions more likely.

- However, given the condition that it is raining makes the events conditionally independent. Once you know it is raining, seeing umbrellas tells you nothing more about road conditions.

Conditional Independence

T	D	C	P(T,D,C)
0	0	0	0.576
0	0	1	0.008
0	1	0	0.144
0	1	1	0.072
1	0	0	0.064
1	0	1	0.012
1	1	0	0.016
1	1	1	0.108

- X, Y independent given Z
  - p(X=x,Y=y|Z=z) = p(X=x|Z=z) p(Y=y|Z=z) for all x,y,z
  - Equivalent: p(X|Y,Z) = p(X|Z) or p(Y|X,Z) = p(Y|Z)
  - Intuition: X has no additional info about Y beyond Z's
- Example: Dentist
  - P(T,D|C) = P(T|C) * P(D|C)

Conditionally independent distributions:

T	C	P(T|C)
0	0	.9
0	1	.4
1	0	.1
1	1	.6

D	C	P(D|C)
0	0	.8
0	1	.1
1	0	.2
1	1	.9

Joint:

Conditional probabilities:

T	D	C	P(T,D|C)
0	0	0	.9 * .8 = .72
0	0	1	.4 * .1 = .04
0	1	0	.9 * .2 = .18
0	1	1	.4 * .9 = .36
1	0	0	.1 * .8 = .08
1	0	1	.6 * .1 = .06
1	1	0	.1 * .2 = .02
1	1	1	.6 * .9 = .54

Again, hard to "read" from the joint probabilities; only from the conditional probabilities.

Like independence, can **greatly** reduce representation size!

We may omit leading zeroes to save space and effort.

Conditional Independence

- Formal Definition:
  - 2 random variables A and B are conditionally independent given C iff:

    P(a, b|c) = P(a|c) P(b|c),   for all values a, b, c

- Informal Definition:
  - 2 random variables A and B are conditionally independent given C iff:

    P(a|b, c) = P(a|c)   OR   P(b|a, c) = P(b|c),   for all values a, b, c

  - P(a|b, c) = P(a|c) tells us that learning about b, given that we already know c, provides no change in our probability for a, and thus b contains no information about a beyond what c provides.

- Naïve Bayes Model:
  - Often a single variable can directly influence a number of other variables, all of which are conditionally independent, given the single variable.
  - E.g., k different symptom variables X_1, X_2, … X_k, and C = disease, reducing to:

    $$P(C \mid X_1, X_2, \ldots. X_K) = \alpha\, P(C) \prod P(X_i \mid C)$$

    ↑ The normalizing constant α is used to abbreviate normalization

Full Joint vs Conditional Independence

- Example : 4 Binary Random Variable (A,B,C,D)
  - Full Joint Probability Table
    - 1 Table with 16 rows
  - Conditional Independence
    - P(A,B,C,D) = P(A) P(B|A) P(C |A, B) P(D| A,B,C) (no saving yet..)
    - if… P(D|A, B) = P(C |A), P(D|A,B,C) = P(D|A) [Naïve Bayes Model]
      - P(A,B,C,D) = P(A) P(B|A) P(C |A) P(D|A)
      - 4 Tables. With at most 4 rows
- If we had N Binary Random Variables
  - Full Joint Probability Table
    - 1 Table with 2^(N) Rows; N = 100, 2^100 ~= 10^30
  - Naïve Bayes Model (Conditional Independence)
    - N tables with at most 4 rows!

Conclusions…

- Representing uncertainty is useful in knowledge bases.
- Probability provides a framework for managing uncertainty.
- Using a full joint distribution and probability rules, we can derive any probability relationship in a probability space.
- Number of required probabilities can be reduced through independence and conditional independence relationships
- Probabilities allow us to make better decisions by using decision theory and expected utilities.
- **Rational agents cannot violate probability theory.**