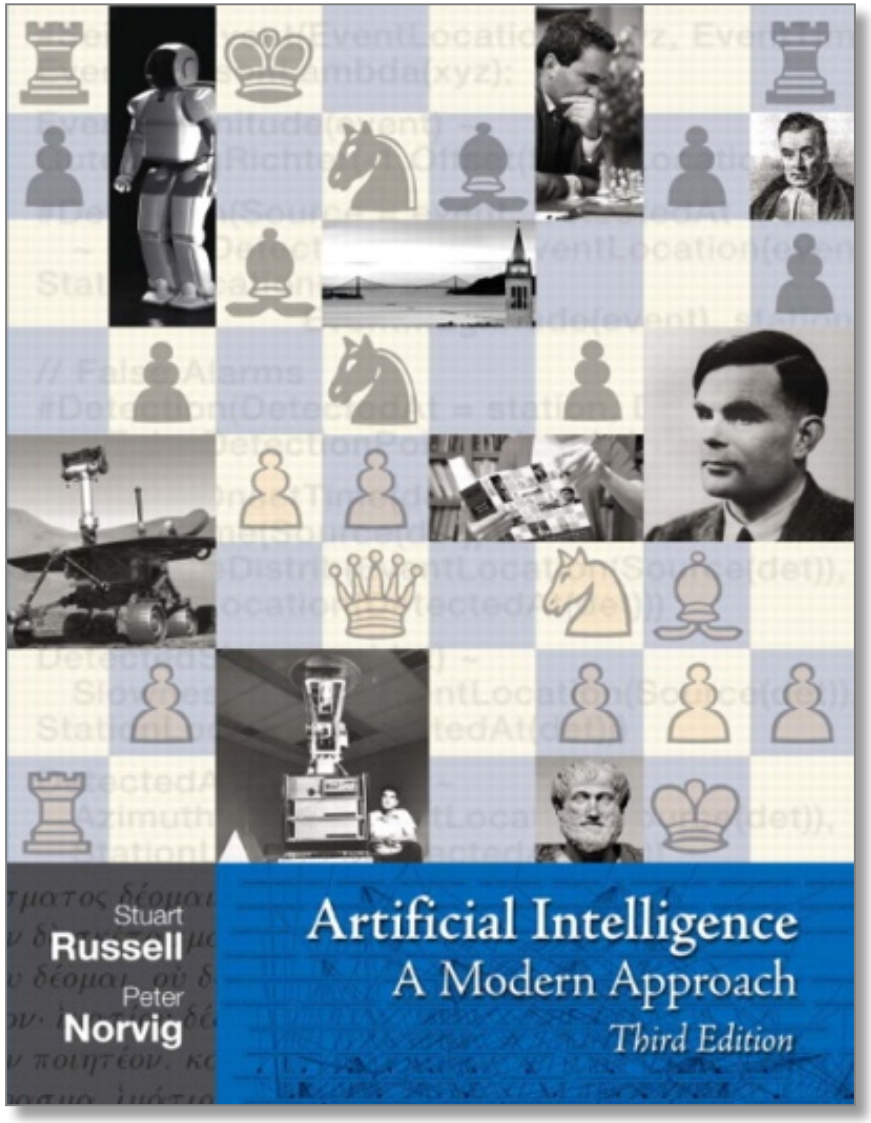
Probabilities and Markov Models

Read AIMA
Chapter 15 “Probabilistic Reasoning Over time”
(15.1-15.5)



Review: Uncertainty

- General situation:
  - **Observed variables (evidence)**: Agent knows certain things about the state of the world (e.g., sensor readings or symptoms)
  - **Unobserved variables (states)**: Agent needs to reason about other aspects (e.g. where an object is or what disease is present)
  - **Model**: Agent knows something about how the known variables relate to the unknown variables
- Probabilistic reasoning gives us a framework for managing our beliefs and knowledge

Review: Random Variables

- A random variable is some aspect of the world about which we (may) have uncertainty
  - R = Is it raining?
  - U = Is the director carrying an umbrella?
- We denote random variables with capital letters

Review: Probability Distributions

- Unobserved random variables have distributions

$P(T)$

T	P
hot	0.5
cold	0.5

$P(W)$

W	P
sun	0.6
rain	0.1
fog	0.3

Shorthand notation:

$$P(hot) = P(T = hot),$$
$$P(cold) = P(T = cold),$$
$$P(rain) = P(W = rain),$$
$$\dots$$

OK *if* all domain entries are unique

- A distribution is a TABLE of probabilities of values
- A probability (lower case value) is a single number

$$P(W = rain) = 0.1$$

- Must have: $\forall x \;\; P(X = x) \geq 0$ and $\sum_x P(X = x) = 1$

Review: Joint Distributions

- A *joint distribution* over a set of random variables: $X_1, X_2, \ldots X_n$ specifies a real number for each assignment (or *outcome*):

$$P(X_1 = x_1, X_2 = x_2, \ldots X_n = x_n)$$

$$P(x_1, x_2, \ldots x_n)$$

  - Must obey: $P(x_1, x_2, \ldots x_n) \geq 0$

$$\sum_{(x_1, x_2, \ldots x_n)} P(x_1, x_2, \ldots x_n) = 1$$

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

- Size of distribution if n variables with domain sizes d?
  - For all but the smallest distributions, impractical to write out!

Review: Probabilistic Models

- A probabilistic model is a joint distribution over a set of random variables

- Probabilistic models:
  - (Random) variables with domains
  - Assignments are called *outcomes*
  - Joint distributions: say whether assignments (outcomes) are likely
  - *Normalized:* sum to 1.0
  - Ideally: only certain variables directly interact

Distribution over T,W

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

Review: Events

- An *event* is a set E of outcomes

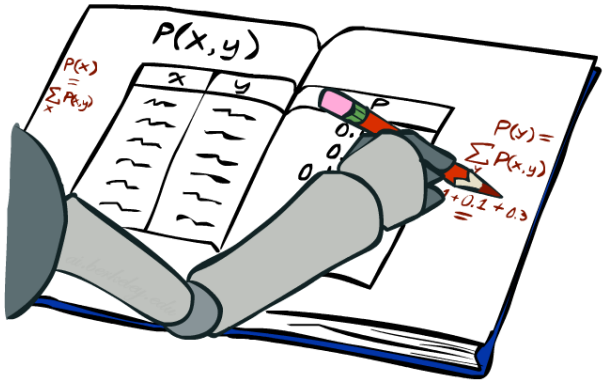
$$P(E) = \sum_{(x_1 \ldots x_n) \in E} P(x_1 \ldots x_n)$$

- From a joint distribution, we can calculate the probability of any event
  - Probability that it's hot AND sunny?
  - Probability that it's hot?
  - Probability that it's hot OR sunny?
- Typically, the events we care about are *partial assignments*, like P(T=hot)

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

Review: Marginal Distributions



- Marginal distributions are sub-tables which eliminate variables
- Marginalization (summing out): Combine collapsed rows by adding

$P(T,W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(t) = \sum_s P(t,s)$$

$P(T)$

T	P
hot	0.5
cold	0.5

$$P(s) = \sum_t P(t,s)$$

$P(W)$

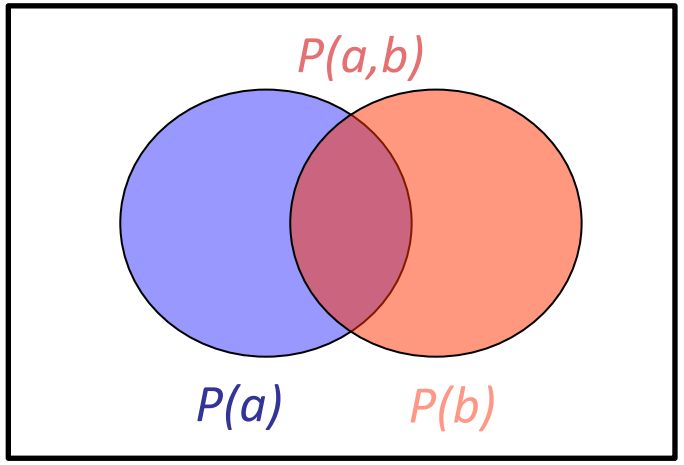
W	P
sun	0.6
rain	0.4

$$P(X_1 = x_1) = \sum_{x_2} P(X_1 = x_1, X_2 = x_2)$$

Review: Conditional Probabilities

- A simple relation between joint and conditional probabilities
  - In fact, this is taken as the *definition* of a conditional probability

$$P(a|b) = \frac{P(a,b)}{P(b)}$$



$P(T,W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(W = s|T = c) = \frac{P(W = s, T = c)}{P(T = c)} = \frac{0.2}{0.5} = 0.4$$

$$= P(W = s, T = c) + P(W = r, T = c)$$
$$= 0.2 + 0.3 = 0.5$$

Review: Conditional Distributions

- Conditional distributions are probability distributions over some variables given fixed values of others

Conditional Distributions

$P(W|T)$

$P(W|T = hot)$

W	P
sun	0.8
rain	0.2

$P(W|T = cold)$

W	P
sun	0.4
rain	0.6

Joint Distribution

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

Review: Probabilistic Inference

- Probabilistic inference: compute a desired probability from other known probabilities (e.g. conditional from joint)

- We generally compute conditional probabilities
  - P(on time | no reported accidents) = 0.90
  - These represent the agent's *beliefs* given the evidence

- Probabilities change with new evidence:
  - P(on time | no accidents, 5 a.m.) = 0.95
  - P(on time | no accidents, 5 a.m., raining) = 0.80
  - Observing new evidence causes *beliefs to be updated*

Review: Inference by Enumeration

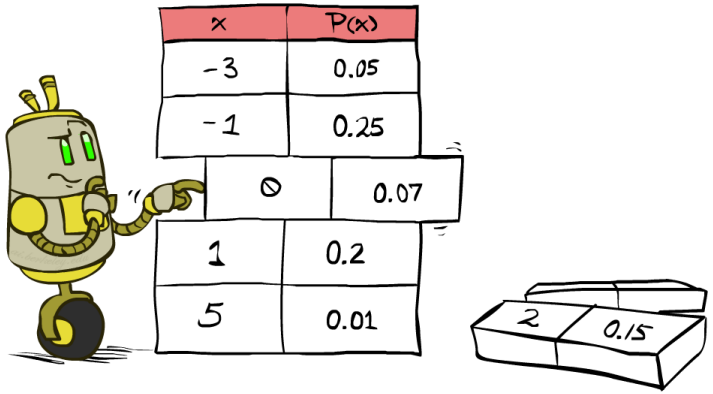
- General case:
  - Evidence variables: $E_1 \dots E_k = e_1 \dots e_k$
  - Query* variable: Q
  - Hidden variables: $H_1 \dots H_r$

  $\left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} X_1, X_2, \dots X_n$ *All variables*

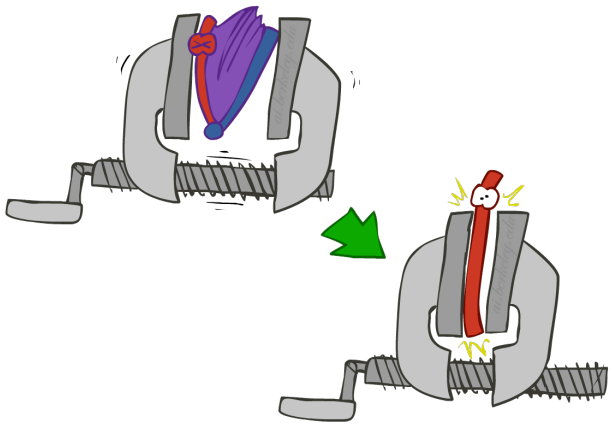
- We want: $P(Q|e_1 \dots e_k)$

** Works fine with multiple query variables, too*

- Step 1: Select the entries consistent with the evidence



- Step 2: Sum out H to get joint of Query and evidence



$$P(Q, e_1 \dots e_k) = \sum_{h_1 \dots h_r} P(\underbrace{Q, h_1 \dots h_r, e_1 \dots e_k}_{X_1, X_2, \dots X_n})$$

- Step 3: Normalize

$$\times \frac{1}{Z}$$

$$Z = \sum_q P(Q, e_1 \cdots e_k)$$

$$P(Q|e_1 \cdots e_k) = \frac{1}{Z} P(Q, e_1 \cdots e_k)$$

Review: The Product Rule

$$P(y)P(x|y) = P(x,y)$$

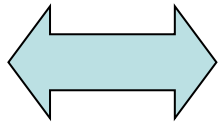
- Example:

$P(W)$

R	P
sun	0.8
rain	0.2

$P(D|W)$

D	W	P
wet	sun	0.1
dry	sun	0.9
wet	rain	0.7
dry	rain	0.3



$P(D,W)$

D	W	P
wet	sun	
dry	sun	
wet	rain	
dry	rain	

Review: The Chain Rule

- More generally, can always write any joint distribution as an incremental product of conditional distributions

$$P(x_1, x_2, x_3) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2)$$

$$P(x_1, x_2, \dots x_n) = \prod_i P(x_i|x_1 \dots x_{i-1})$$

- Why is this always true?

Review: Bayes' Rule

- Two ways to factor a joint distribution over two variables:

$$P(x, y) = P(x|y)P(y) = P(y|x)P(x)$$

- Dividing, we get:

$$P(x|y) = \frac{P(y|x)}{P(y)}P(x)$$

- Why is this at all helpful?
  - Lets us build one conditional from its reverse
  - Often one conditional is tricky but the other one is simple
  - Foundation of many systems we'll see later (e.g. ASR, MT)

- In the running for most important AI equation!

Review: Inference with Bayes' Rule

- Example: Diagnostic probability from causal probability:

$$P(\text{cause}|\text{effect}) = \frac{P(\text{effect}|\text{cause})P(\text{cause})}{P(\text{effect})}$$

- Example:
  - M: meningitis, S: stiff neck

$$\left.\begin{aligned} P(+m) &= 0.0001 \\ P(+s|+m) &= 0.8 \\ P(+s|-m) &= 0.01 \end{aligned}\right\}$$ Example givens

$$P(+m|+s) = \frac{P(+s|+m)P(+m)}{P(+s)} = \frac{P(+s|+m)P(+m)}{P(+s|+m)P(+m) + P(+s|-m)P(-m)} = \frac{0.8 \times 0.0001}{0.8 \times 0.0001 + 0.01 \times 0.999}$$

  - Note: posterior probability of meningitis still very small
  - Note: you should still get stiff necks checked out! Why?

Example: Independence?

$P(T)$

T	P
hot	0.5
cold	0.5

$P_1(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$P(W)$

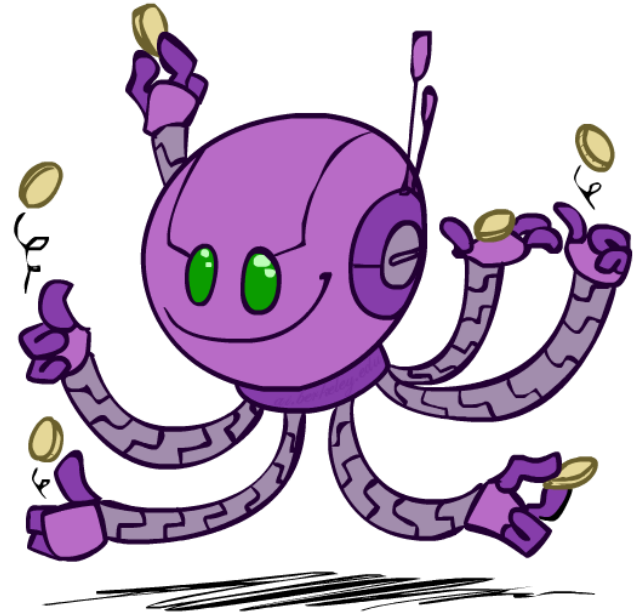
W	P
sun	0.6
rain	0.4

$P_2(T, W) = P(T)P(W)$

T	W	P
hot	sun	0.3
hot	rain	0.2
cold	sun	0.3
cold	rain	0.2

Example: Independence

- N fair, independent coin flips:



$P(X_1)$

H	0.5
T	0.5

$P(X_2)$

H	0.5
T	0.5

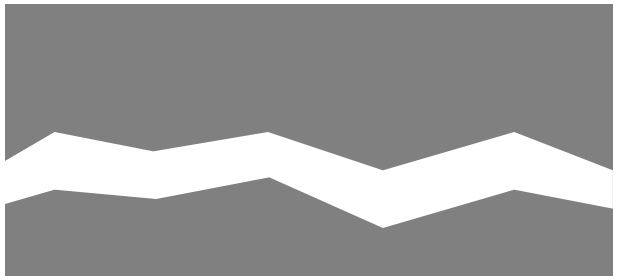
$\ldots$

$P(X_n)$

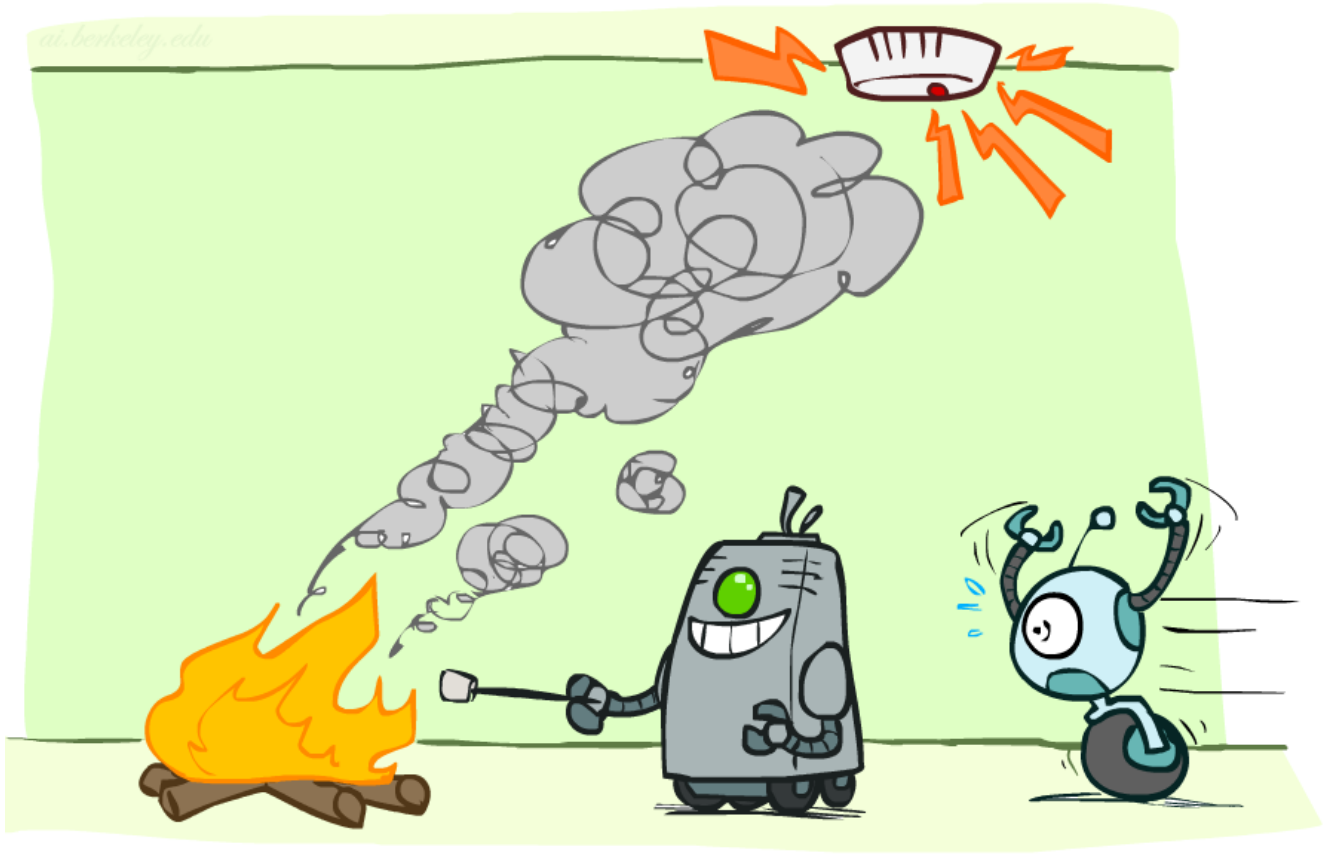
H	0.5
T	0.5

$P(X_1, X_2, \ldots X_n)$

2^n

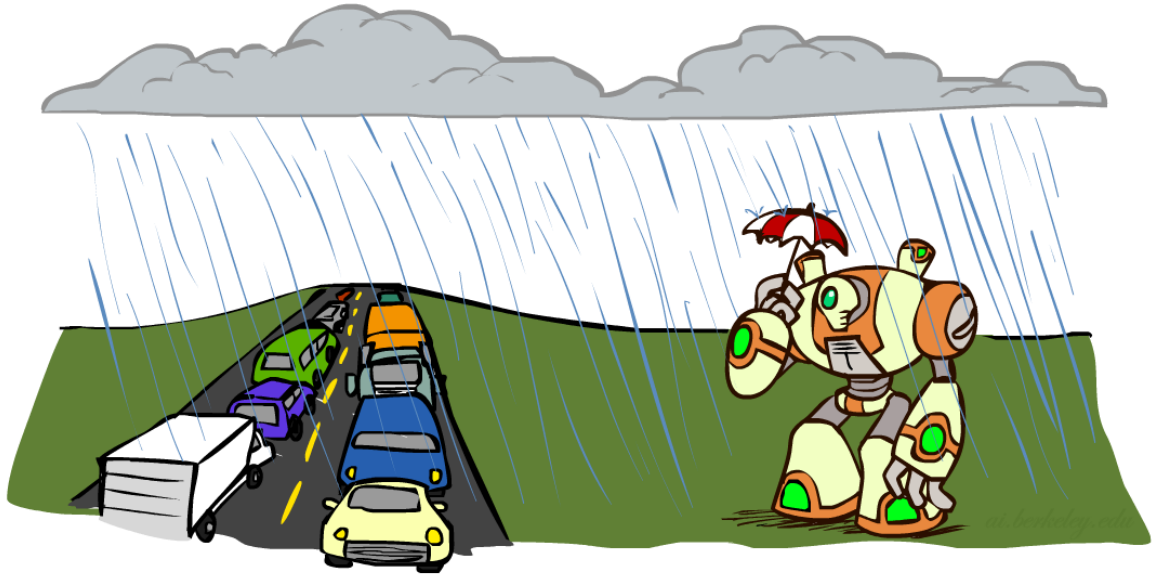


Conditional Independence



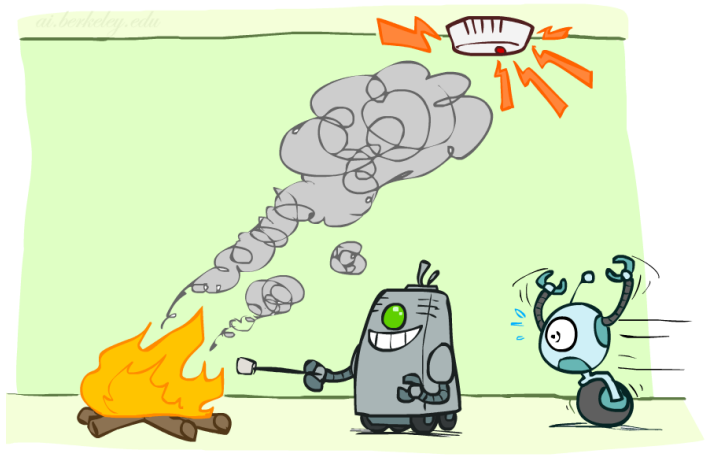
Conditional Independence

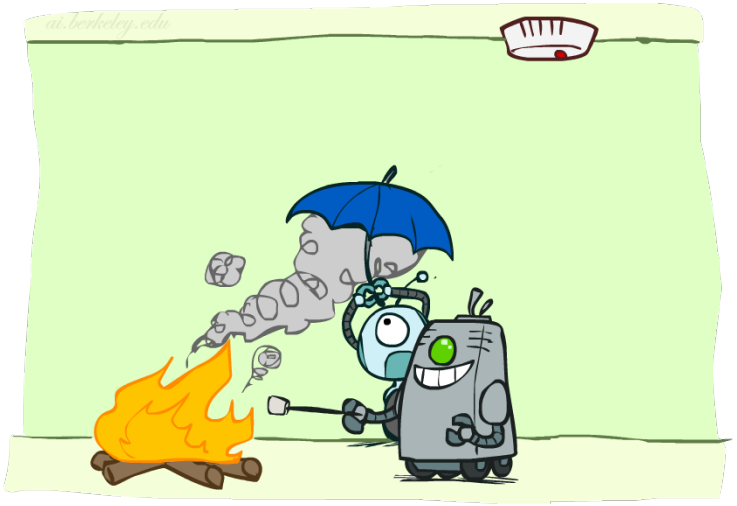
- What about this domain:
  - Traffic
  - Umbrella
  - Raining



Conditional Independence

- What about this domain:
  - Fire
  - Smoke
  - Alarm





Probability Recap

- Conditional probability $$P(x|y) = \frac{P(x,y)}{P(y)}$$
- Product rule $$P(x,y) = P(x|y)P(y)$$
- Chain rule $$\begin{aligned} P(X_1, X_2, \ldots X_n) &= P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)\ldots \\ &= \prod_{i=1}^{n} P(X_i|X_1, \ldots, X_{i-1}) \end{aligned}$$
- X, Y independent if and only if: $\forall x, y : P(x,y) = P(x)P(y)$
- X and Y are conditionally independent given Z if and only if: $X \perp\!\!\!\perp Y|Z$

$$\forall x, y, z : P(x,y|z) = P(x|z)P(y|z)$$

Independence

- Two variables are *independent* in a joint distribution if:

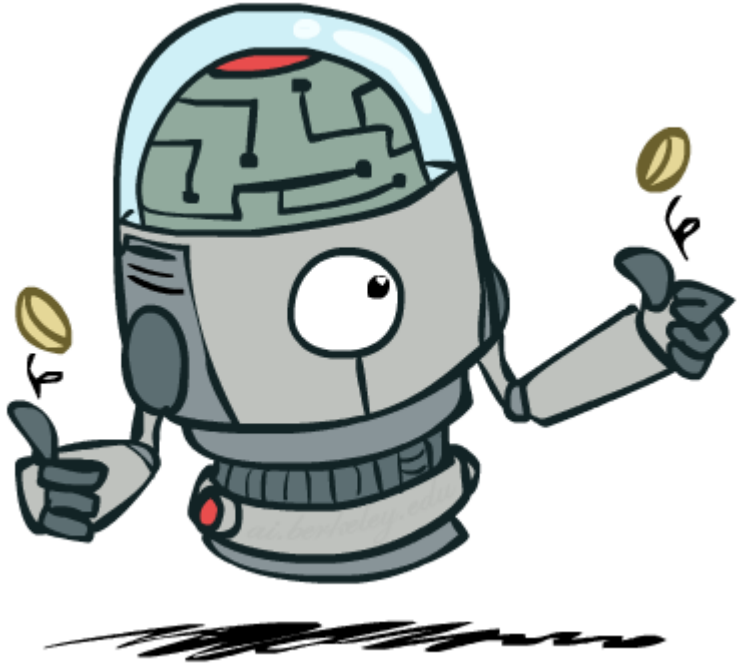
$$P(X,Y) = P(X)P(Y)$$

$$X \perp\!\!\!\perp Y$$

$$\forall x, y \; P(x,y) = P(x)P(y)$$

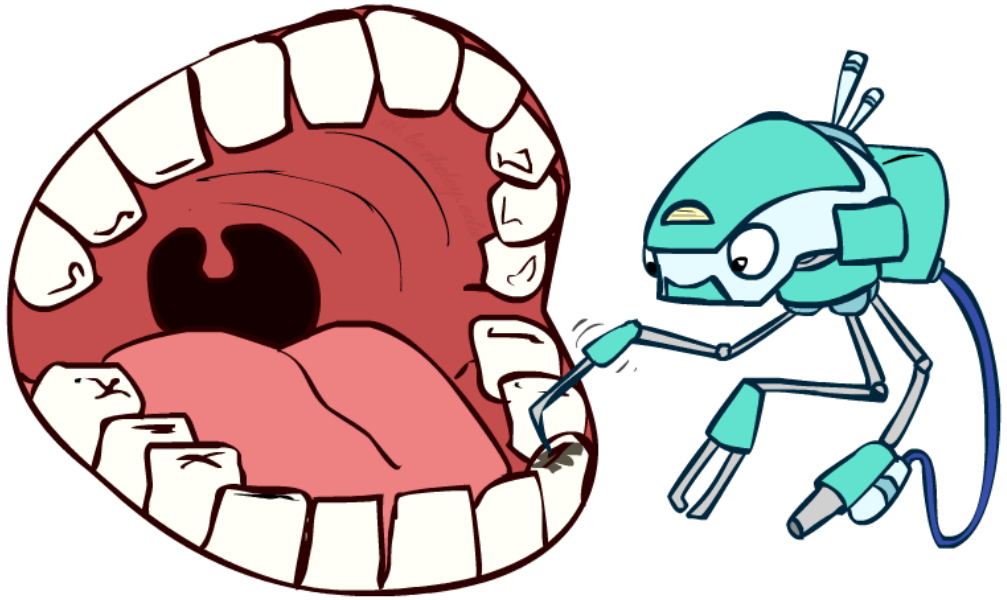
  - Says the joint distribution *factors* into a product of two simple ones
  - Usually variables aren't independent!

- Can use independence as a *modeling assumption*
  - Independence can be a simplifying assumption
  - *Empirical* joint distributions: at best "close" to independent
  - What could we assume for {Weather, Traffic, Cavity}?



Conditional Independence

- P(Toothache, Cavity, Catch)
- If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:
  - P(+catch | +toothache, +cavity) = P(+catch | +cavity)
- The same independence holds if I don’t have a cavity:
  - P(+catch | +toothache, -cavity) = P(+catch| -cavity)
- Catch is *conditionally independent* of Toothache given Cavity:
  - P(Catch | Toothache, Cavity) = P(Catch | Cavity)
- Equivalent statements:
  - P(Toothache | Catch , Cavity) = P(Toothache | Cavity)
  - P(Toothache, Catch | Cavity) = P(Toothache | Cavity) P(Catch | Cavity)
  - One can be derived from the other



Conditional Independence

- Unconditional (absolute) independence very rare (why?)
- *Conditional independence* is our most basic and robust form of knowledge about uncertain environments.
- X is conditionally independent of Y given Z

$$X \perp\!\!\!\perp Y | Z$$

if and only if:

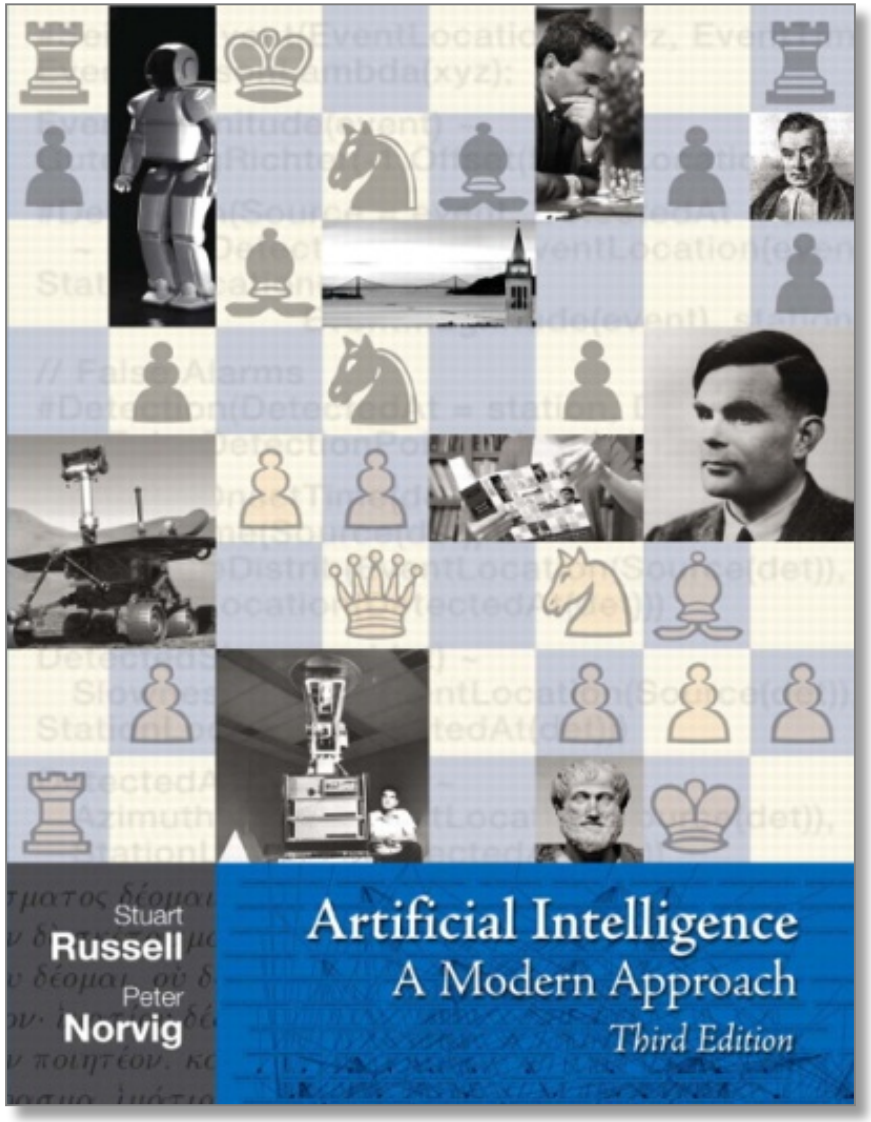
$$\forall x, y, z : P(x, y|z) = P(x|z)P(y|z)$$

or, equivalently, if and only if

$$\forall x, y, z : P(x|z, y) = P(x|z)$$

Probabilistic Reasoning Over Time

Read AIMA
Chapter 15 (15.1-15.5)



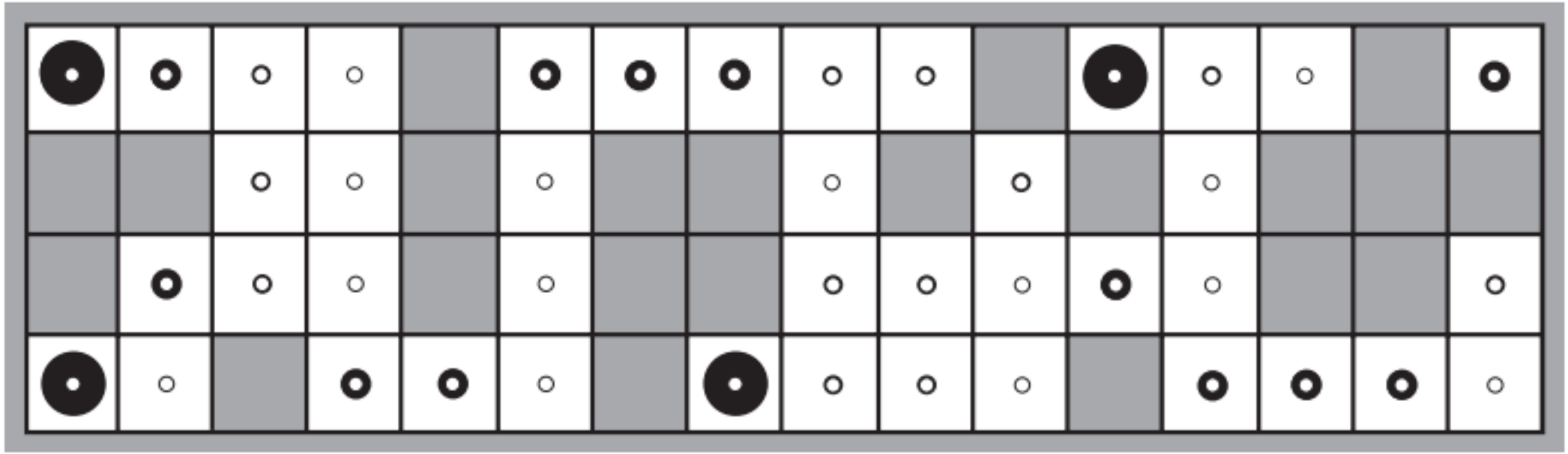
These slides are courtesy of Dan Klein and Pieter Abbeel for UC Berkeley's Intro to AI course

Partially Observable Environments

- So far, we have considered **full-observable environments** like chess, or gridworld
- In **partially-observable environments** we don't see everything
- In this case an agent needs to maintain a **belief state** that represents represents which states are currently possible

Robot Localization

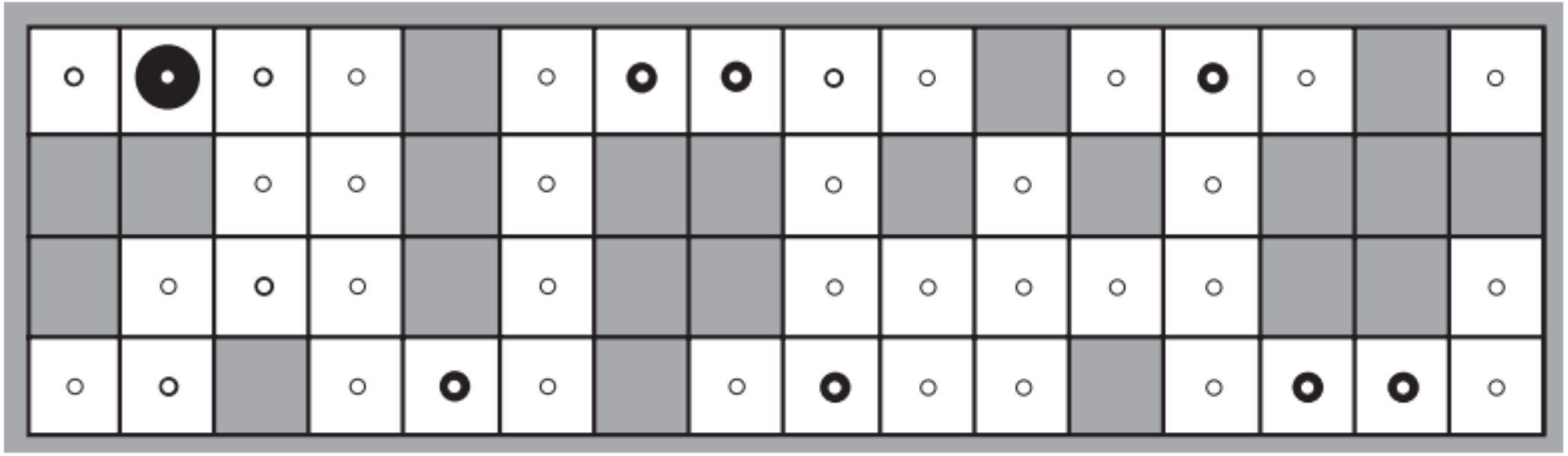
Given a map and sensor data, the robot updates its beliefs about where it is.



Sensor data with one observation saying there are obstacles to the North, South and West.

Robot Localization

Given a map and sensor data, the robot updates its beliefs about where it is.



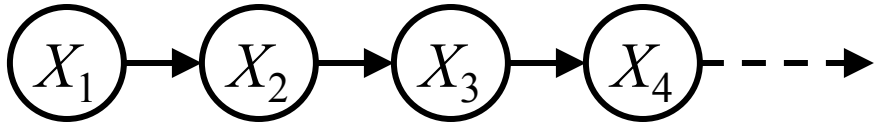
Sensor data with one observation saying there are obstacles to the North, South and West. The robot moves, and then takes another observation with obstacles to the North and South, and updates its belief state.

Reasoning over Time or Space

- Often, we want to reason about a sequence of observations
  - Robot localization
  - Speech recognition
  - Medical monitoring
- Need to introduce time (or space) into our models

Markov Models

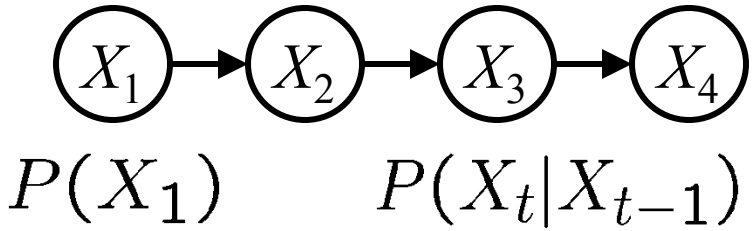
- Value of X at a given time is called the state



$$P(X_1) \qquad P(X_t|X_{t-1})$$

- Parameters: called transition probabilities or dynamics, specify how the state evolves over time (also, initial state probabilities)
- Stationarity assumption: transition probabilities the same at all times
- Same as MDP transition model, but no choice of action

Joint Distribution of a Markov Model



- Joint distribution:

$$P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2|X_1)P(X_3|X_2)P(X_4|X_3)$$

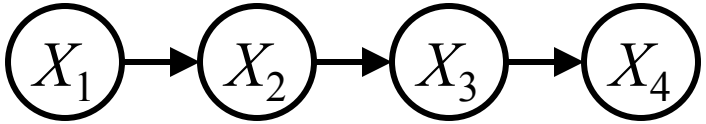
- More generally:

$$P(X_1, X_2, \ldots, X_T) = P(X_1)P(X_2|X_1)P(X_3|X_2)\ldots P(X_T|X_{T-1})$$

$$= P(X_1)\prod_{t=2}^{T} P(X_t|X_{t-1})$$

- Questions to be resolved:
  - Does this indeed define a joint distribution?
  - Can every joint distribution be factored this way, or are we making some assumptions about the joint distribution by using this factorization?

Chain Rule and Markov Models



- From the chain rule, every joint distribution over X_1, X_2, X_3, X_4 can be written as:

$$P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)P(X_4|X_1, X_2, X_3)$$

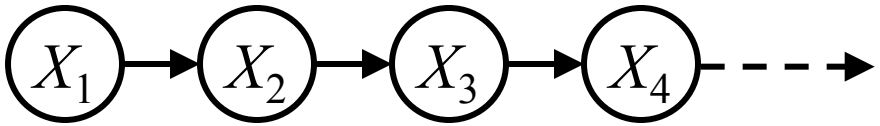
- Assuming that

$$X_3 \perp\!\!\!\perp X_1 \mid X_2 \quad \text{and} \quad X_4 \perp\!\!\!\perp X_1, X_2 \mid X_3$$

results in the expression posited on the previous slide:

$$P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2|X_1)P(X_3|X_2)P(X_4|X_3)$$

Chain Rule and Markov Models



- From the chain rule, every joint distribution over $X_1, X_2, \ldots, X_T$ can be written as:

$$P(X_1, X_2, \ldots, X_T) = P(X_1) \prod_{t=2}^{T} P(X_t | X_1, X_2, \ldots, X_{t-1})$$

- Assuming that for all *t*:

$$X_t \perp\!\!\!\perp X_1, \ldots, X_{t-2} \mid X_{t-1}$$

gives us the expression posited on the earlier slide:

$$P(X_1, X_2, \ldots, X_T) = P(X_1) \prod_{t=2}^{T} P(X_t | X_{t-1})$$

Markov Models Recap

- Explicit assumption for all t : $X_t \perp\!\!\!\perp X_1, \ldots, X_{t-2} \mid X_{t-1}$
- Consequence, joint distribution can be written as:

$$P(X_1, X_2, \ldots, X_T) = P(X_1)P(X_2|X_1)P(X_3|X_2)\ldots P(X_T|X_{T-1})$$

$$= P(X_1)\prod_{t=2}^{T} P(X_t|X_{t-1})$$

- Implied conditional independencies: (try to prove this!)
  - Past variables independent of future variables given the present

  i.e., if $t_1 < t_2 < t_3$ or $t_1 > t_2 > t_3$ then: $X_{t_1} \perp\!\!\!\perp X_{t_3} \mid X_{t_2}$
- Additional explicit assumption: $P(X_t \mid X_{t-1})$ is the same for all t

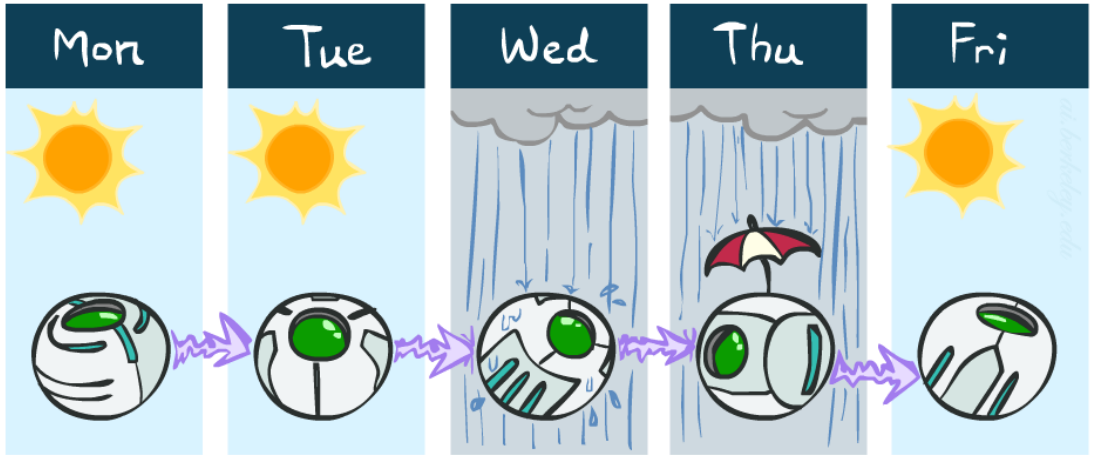
Conditional Independence

- Basic conditional independence:
  - Past and future independent of the present
  - Each time step only depends on the previous
  - This is called the (first order) Markov property

- Note that the chain is just a (growable) BN
  - We can always use generic BN reasoning on it if we truncate the chain at a fixed length

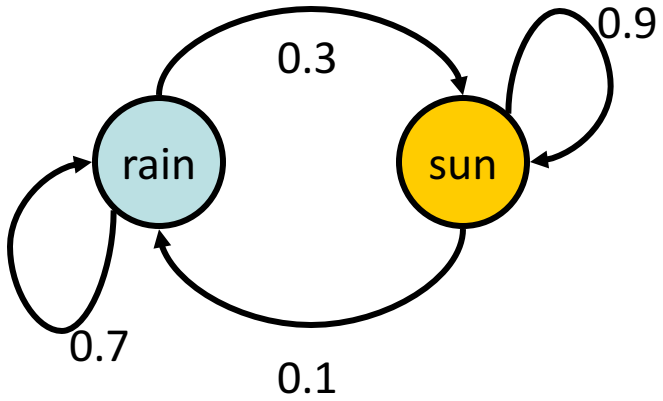
Example Markov Chain: Weather

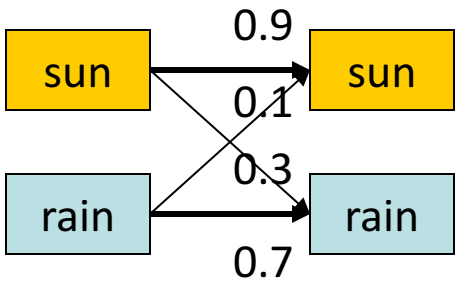
- States: X = {rain, sun}
- Initial distribution: 1.0 sun
- CPT $P(X_t \mid X_{t-1})$:



X_{t-1}	X_t	$P(X_t \mid X_{t-1})$
sun	sun	0.9
sun	rain	0.1
rain	sun	0.3
rain	rain	0.7

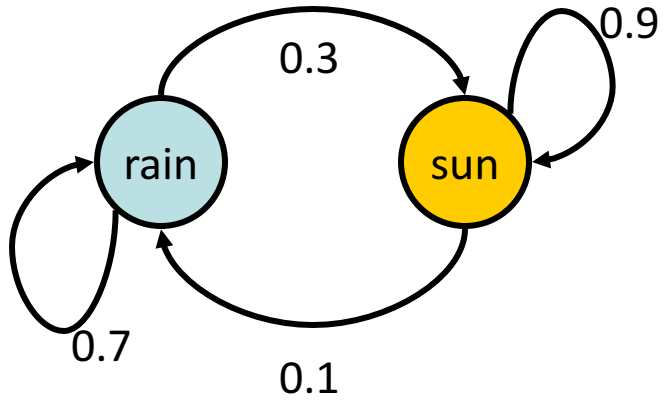
Two new ways of representing the same CPT





Example Markov Chain: Weather

- Initial distribution: 1.0 sun



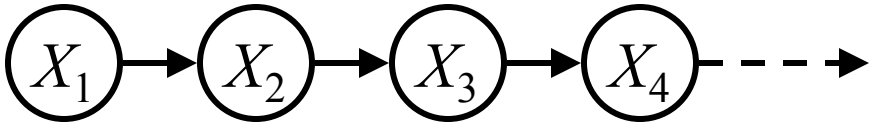
- What is the probability distribution after one step?

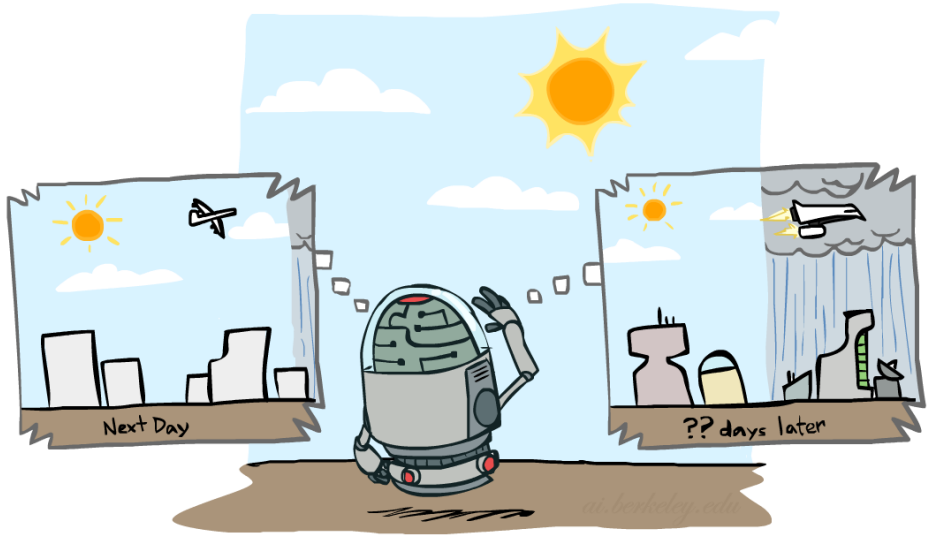
$$P(X_2 = \text{sun}) = P(X_2 = \text{sun}|X_1 = \text{sun})P(X_1 = \text{sun}) + P(X_2 = \text{sun}|X_1 = \text{rain})P(X_1 = \text{rain})$$

$$0.9 \cdot 1.0 + 0.3 \cdot 0.0 = 0.9$$

Mini-Forward Algorithm

- Question: What's P(X) on some day t?





$$P(x_1) = \text{known}$$

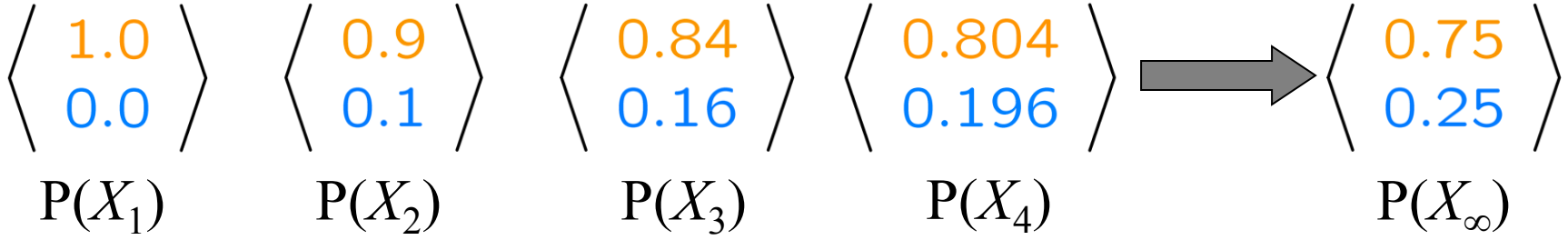
$$P(x_t) = \sum_{x_{t-1}} P(x_{t-1}, x_t)$$

$$= \sum_{x_{t-1}} P(x_t \mid x_{t-1}) P(x_{t-1})$$

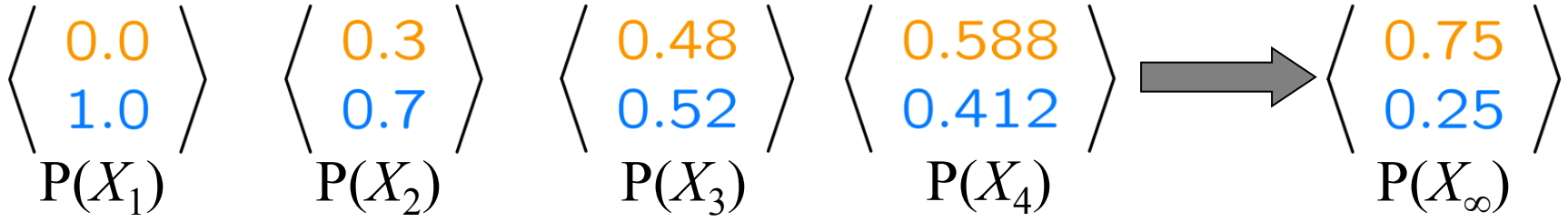
Forward simulation

Example Run of Mini-Forward Algorithm

- From initial observation of sun



- From initial observation of rain



- From yet another initial distribution $P(X_1)$:

[Demo: L13D1,2,3]

Video of Demo Ghostbusters Basic Dynamics

Video of Demo Ghostbusters Circular Dynamics

Video of Demo Ghostbusters Whirlpool Dynamics

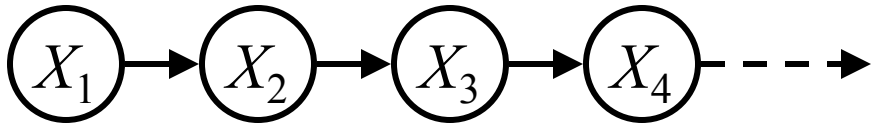
Stationary Distributions

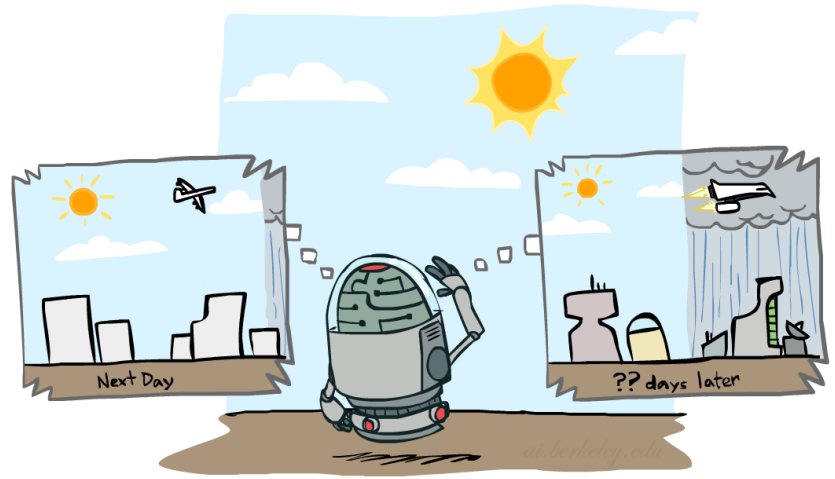
- For most chains:
  - Influence of the initial distribution gets less and less over time.
  - The distribution we end up in is independent of the initial distribution
- Stationary distribution:
  - The distribution we end up with is called the stationary distribution P_∞ of the chain
  - It satisfies

$$P_\infty(X) = P_{\infty+1}(X) = \sum_x P(X|x)P_\infty(x)$$

Example: Stationary Distributions

- Question: What's P(X) at time t = infinity?





$$P_\infty(sun) = P(sun|sun)P_\infty(sun) + P(sun|rain)P_\infty(rain)$$
$$P_\infty(rain) = P(rain|sun)P_\infty(sun) + P(rain|rain)P_\infty(rain)$$

$$P_\infty(sun) = 0.9P_\infty(sun) + 0.3P_\infty(rain)$$
$$P_\infty(rain) = 0.1P_\infty(sun) + 0.7P_\infty(rain)$$

$$P_\infty(sun) = 3P_\infty(rain)$$
$$P_\infty(rain) = 1/3P_\infty(sun)$$

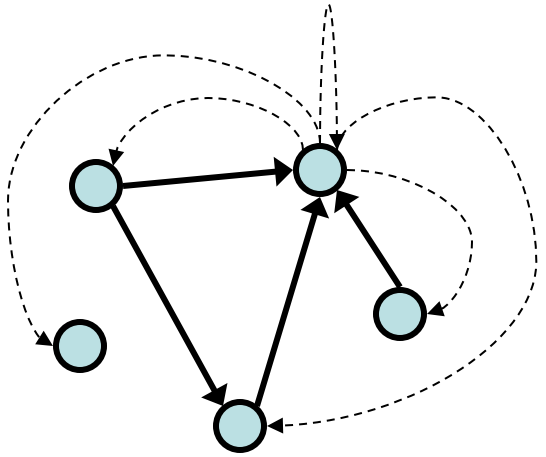
Also: $P_\infty(sun) + P_\infty(rain) = 1$

$$P_\infty(sun) = 3/4$$
$$P_\infty(rain) = 1/4$$

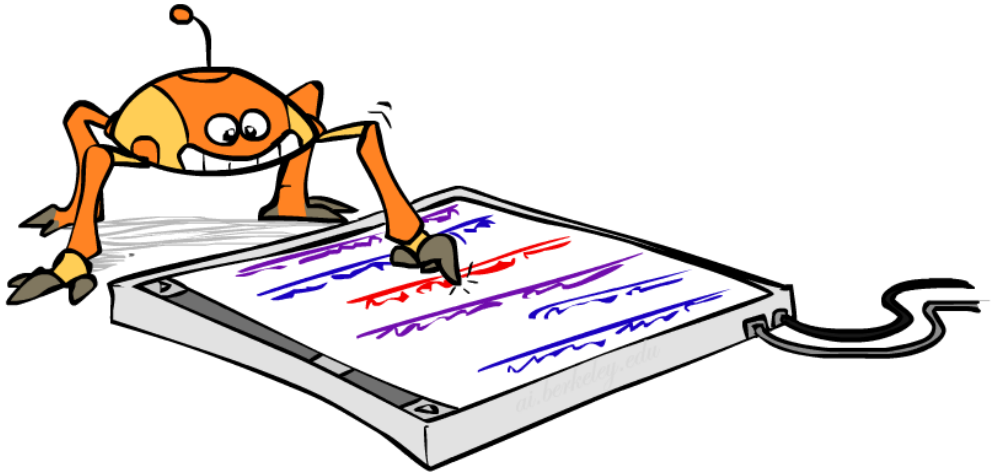
X_{t-1}	X_t	$P(X_t \mid X_{t-1})$
sun	sun	0.9
sun	rain	0.1
rain	sun	0.3
rain	rain	0.7

Application of Stationary Distribution: Web Link Analysis

- PageRank over a web graph
  - Each web page is a state
  - Initial distribution: uniform over pages
  - Transitions:
    - With prob. c, uniform jump to a random page (dotted lines, not all shown)
    - With prob. 1-c, follow a random outlink (solid lines)



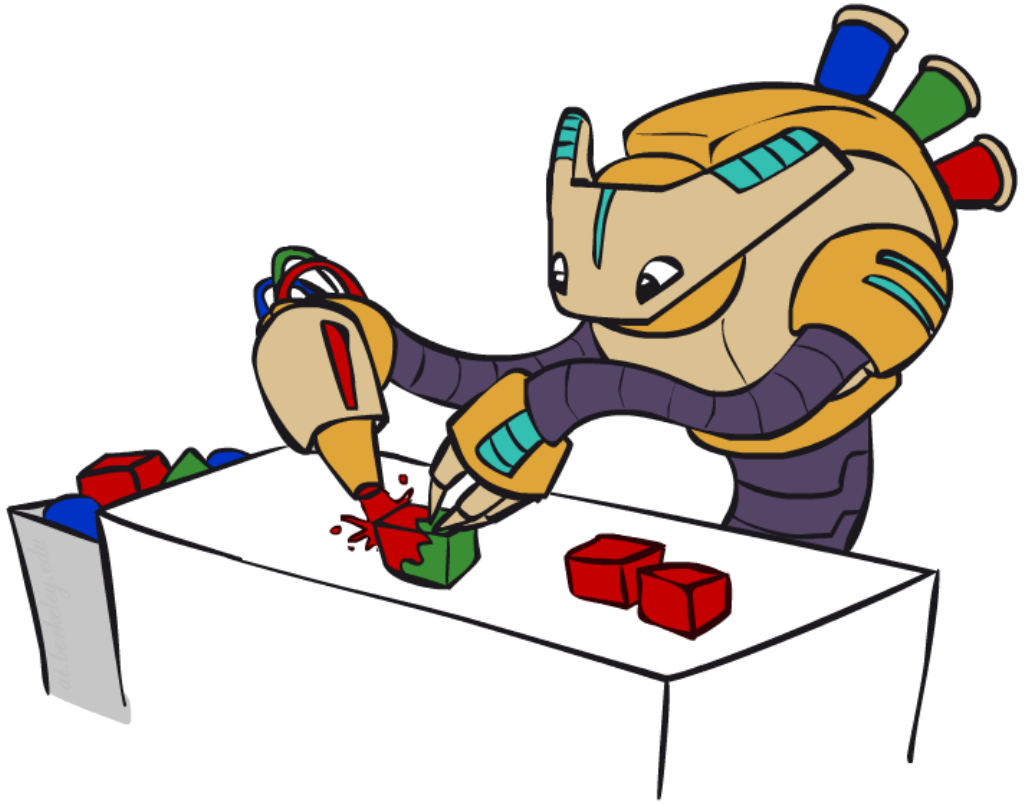
- Stationary distribution
  - Will spend more time on highly reachable pages
  - E.g. many ways to get to the Acrobat Reader download page
  - Somewhat robust to link spam
  - Google 1.0 returned the set of pages containing all your keywords in decreasing rank, now all search engines use link analysis along with many other factors (rank actually getting less important over time)



Application of Stationary Distributions: Gibbs Sampling*

- Each joint instantiation over all hidden and query variables is a state: $\{X_1, \ldots, X_n\} = H \cup Q$
- Transitions:
  - With probability 1/n resample variable X_j according to

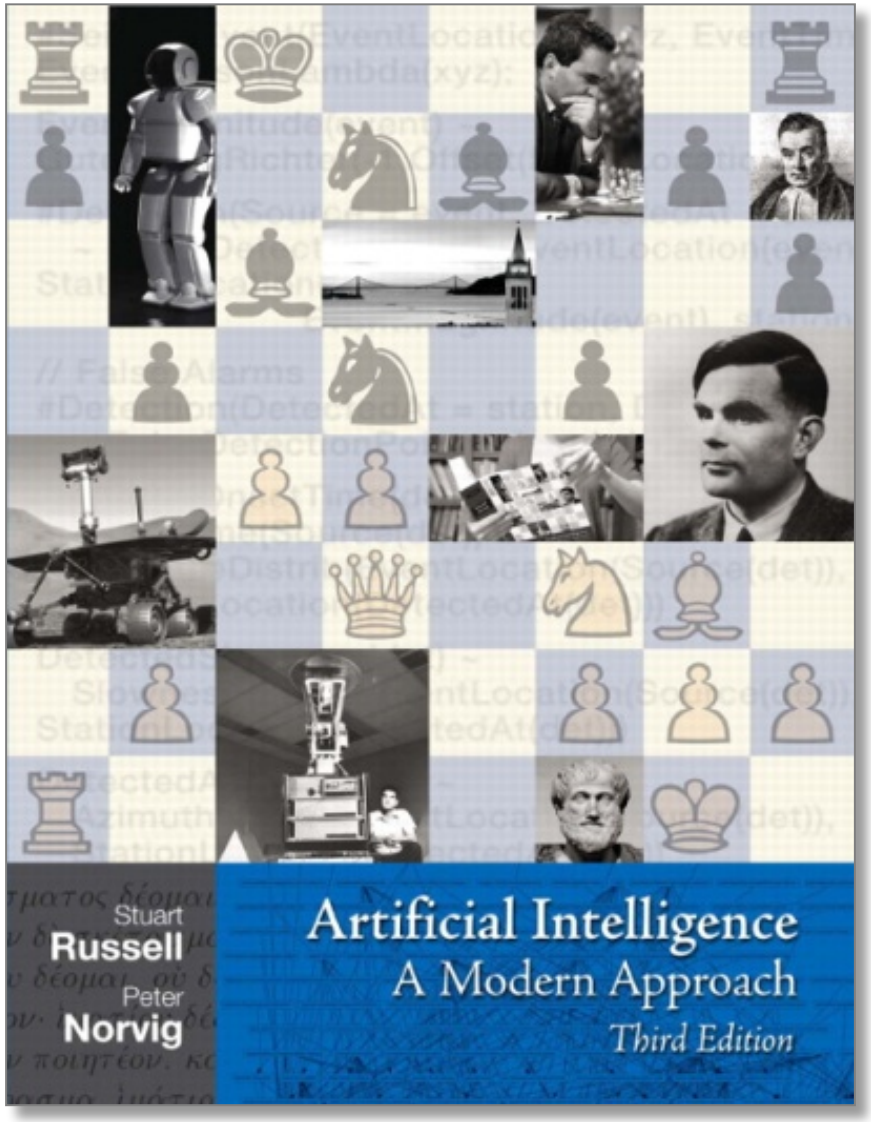
    $P(X_j \mid x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n, e_1, \ldots, e_m)$
- Stationary distribution:
  - Conditional distribution $P(X_1, X_2, \ldots, X_n \mid e_1, \ldots, e_m)$
  - Means that when running Gibbs sampling long enough we get a sample from the desired distribution
  - Requires some proof to show this is true!



Next Up: Markov Models for NLP

Probabilities and Markov Models

Read AIMA
Chapter 15 “Probabilistic Reasoning Over time”
(15.1-15.5)



Review: Uncertainty

- General situation:
  - **Observed variables (evidence)**: Agent knows certain things about the state of the world (e.g., sensor readings or symptoms)
  - **Unobserved variables (states)**: Agent needs to reason about other aspects (e.g. where an object is or what disease is present)
  - **Model**: Agent knows something about how the known variables relate to the unknown variables
- Probabilistic reasoning gives us a framework for managing our beliefs and knowledge

Review: Random Variables

- A random variable is some aspect of the world about which we (may) have uncertainty
  - R = Is it raining?
  - U = Is the director carrying an umbrella?
- We denote random variables with capital letters
- Like variables in a CSP, random variables have domains
  - R in {true, false} (often write as {+r, -r})
  - T in {hot, cold}
  - D in $[0, \infty)$
  - L in possible locations, maybe {(0,0), (0,1), …}

Review: Probability Distributions

- Unobserved random variables have distributions

$P(T)$

T	P
hot	0.5
cold	0.5

$P(W)$

W	P
sun	0.6
rain	0.1
fog	0.3

- A distribution is a TABLE of probabilities of values
- A probability (lower case value) is a single number

Shorthand notation:

$$P(hot) = P(T = hot),$$
$$P(cold) = P(T = cold),$$
$$P(rain) = P(W = rain),$$
$$\dots$$

OK *if* all domain entries are unique

- Must have:

$$P(W = rain) = 0.1 \quad \text{and}$$

$$\forall x \;\; P(X = x) \geq 0 \qquad \sum_x P(X = x) = 1$$

Review: Joint Distributions

- A *joint distribution* over a set of random variables: $X_1, X_2, \ldots X_n$ specifies a real number for each assignment (or *outcome*):

$$P(X_1 = x_1, X_2 = x_2, \ldots X_n = x_n)$$

$$P(x_1, x_2, \ldots x_n)$$

  - Must obey:

$$P(x_1, x_2, \ldots x_n) \geq 0$$

$$\sum_{(x_1, x_2, \ldots x_n)} P(x_1, x_2, \ldots x_n) = 1$$

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

- Size of distribution if n variables with domain sizes d?
  - For all but the smallest distributions, impractical to write out!

Review: Probabilistic Models

- A probabilistic model is a joint distribution over a set of random variables

- Probabilistic models:
  - (Random) variables with domains
  - Assignments are called *outcomes*
  - Joint distributions: say whether assignments (outcomes) are likely
  - *Normalized:* sum to 1.0
  - Ideally: only certain variables directly interact

Distribution over T,W

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

Review: Events

- An *event* is a set E of outcomes

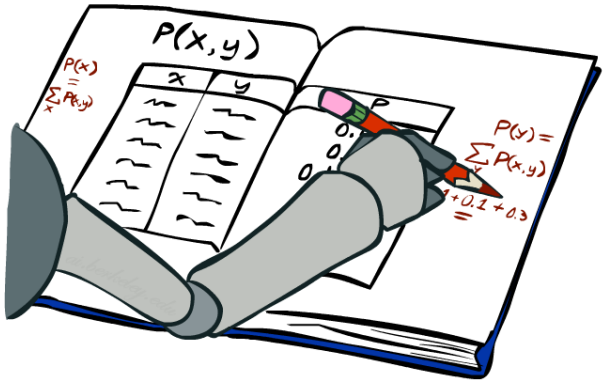
$$P(E) = \sum_{(x_1 \ldots x_n) \in E} P(x_1 \ldots x_n)$$

- From a joint distribution, we can calculate the probability of any event
  - Probability that it's hot AND sunny?
  - Probability that it's hot?
  - Probability that it's hot OR sunny?
- Typically, the events we care about are *partial assignments*, like P(T=hot)

$$P(T, W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

Review: Marginal Distributions



- Marginal distributions are sub-tables which eliminate variables
- Marginalization (summing out): Combine collapsed rows by adding

$P(T,W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(t) = \sum_s P(t,s)$$

$P(T)$

T	P
hot	0.5
cold	0.5

$$P(s) = \sum_t P(t,s)$$

$P(W)$

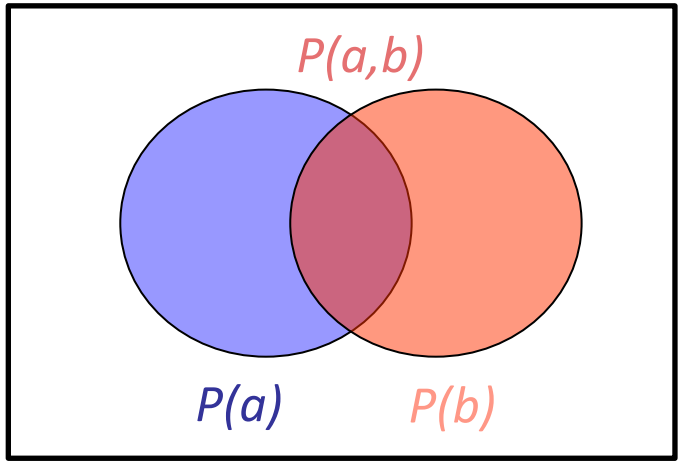
W	P
sun	0.6
rain	0.4

$$P(X_1 = x_1) = \sum_{x_2} P(X_1 = x_1, X_2 = x_2)$$

Review: Conditional Probabilities

- A simple relation between joint and conditional probabilities
  - In fact, this is taken as the *definition* of a conditional probability

$$P(a|b) = \frac{P(a,b)}{P(b)}$$



$P(T,W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(W = s|T = c) = \frac{P(W = s, T = c)}{P(T = c)} = \frac{0.2}{0.5} = 0.4$$

$$= P(W = s, T = c) + P(W = r, T = c)$$
$$= 0.2 + 0.3 = 0.5$$

Review: Conditional Distributions

- Conditional distributions are probability distributions over some variables given fixed values of others

Conditional Distributions

$P(W|T)$

$P(W|T = hot)$

W	P
sun	0.8
rain	0.2

$P(W|T = cold)$

W	P
sun	0.4
rain	0.6

Joint Distribution

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

Review: Probabilistic Inference

- Probabilistic inference: compute a desired probability from other known probabilities (e.g. conditional from joint)

- We generally compute conditional probabilities
  - P(on time | no reported accidents) = 0.90
  - These represent the agent's *beliefs* given the evidence

- Probabilities change with new evidence:
  - P(on time | no accidents, 5 a.m.) = 0.95
  - P(on time | no accidents, 5 a.m., raining) = 0.80
  - Observing new evidence causes *beliefs to be updated*

Review: Inference by Enumeration

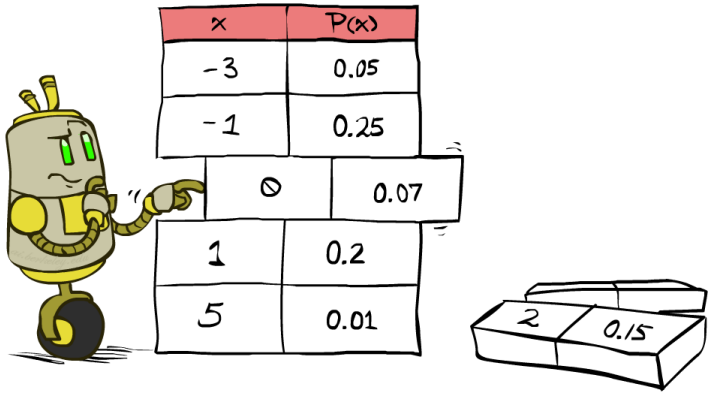
- General case:
  - Evidence variables: $E_1 \dots E_k = e_1 \dots e_k$
  - Query* variable: Q
  - Hidden variables: $H_1 \dots H_r$

  $\left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} X_1, X_2, \dots X_n$ *All variables*

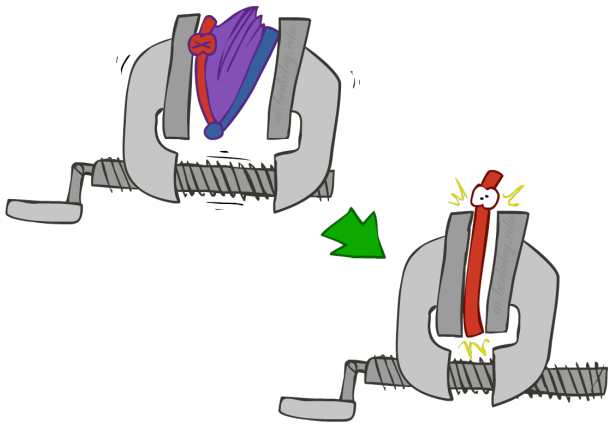
- We want: $P(Q|e_1 \dots e_k)$

** Works fine with multiple query variables, too*

- Step 1: Select the entries consistent with the evidence



- Step 2: Sum out H to get joint of Query and evidence



$$P(Q, e_1 \dots e_k) = \sum_{h_1 \dots h_r} P(\underbrace{Q, h_1 \dots h_r, e_1 \dots e_k}_{X_1, X_2, \dots X_n})$$

- Step 3: Normalize

$$\times \frac{1}{Z}$$

$$Z = \sum_q P(Q, e_1 \cdots e_k)$$

$$P(Q|e_1 \cdots e_k) = \frac{1}{Z} P(Q, e_1 \cdots e_k)$$

Review: The Product Rule

$$P(y)P(x|y) = P(x,y)$$

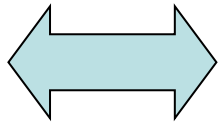
- Example:

$P(W)$

R	P
sun	0.8
rain	0.2

$P(D|W)$

D	W	P
wet	sun	0.1
dry	sun	0.9
wet	rain	0.7
dry	rain	0.3



$P(D,W)$

D	W	P
wet	sun	
dry	sun	
wet	rain	
dry	rain	

Review: The Chain Rule

- More generally, can always write any joint distribution as an incremental product of conditional distributions

$$P(x_1, x_2, x_3) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2)$$

$$P(x_1, x_2, \dots x_n) = \prod_i P(x_i|x_1 \dots x_{i-1})$$

- Why is this always true?

Review: Bayes' Rule

- Two ways to factor a joint distribution over two variables:

$$P(x, y) = P(x|y)P(y) = P(y|x)P(x)$$

- Dividing, we get:

$$P(x|y) = \frac{P(y|x)}{P(y)}P(x)$$

- Why is this at all helpful?
  - Lets us build one conditional from its reverse
  - Often one conditional is tricky but the other one is simple
  - Foundation of many systems we'll see later (e.g. ASR, MT)

- In the running for most important AI equation!

Review: Inference with Bayes' Rule

- Example: Diagnostic probability from causal probability:

- Example: $$P(\text{cause}|\text{effect}) = \frac{P(\text{effect}|\text{cause})P(\text{cause})}{P(\text{effect})}$$
  - M: meningitis, S: stiff neck

$$\left.\begin{aligned} P(+m) &= 0.0001 \\ P(+s|+m) &= 0.8 \\ P(+s|-m) &= 0.01 \end{aligned}\right\}$$ Example givens

$$P(+m|+s) = \frac{P(+s|+m)P(+m)}{P(+s)} = \frac{P(+s|+m)P(+m)}{P(+s|+m)P(+m) + P(+s|-m)P(-m)} = \frac{0.8 \times 0.0001}{0.8 \times 0.0001 + 0.01 \times 0.999}$$

  - Note: posterior probability of meningitis still very small
  - Note: you should still get stiff necks checked out! Why?

Example: Independence?

$P(T)$

T	P
hot	0.5
cold	0.5

$P_1(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$P(W)$

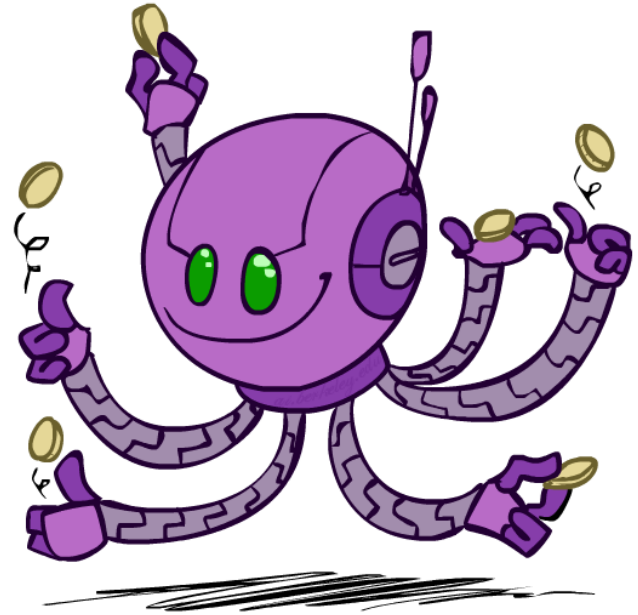
W	P
sun	0.6
rain	0.4

$P_2(T, W) = P(T)P(W)$

T	W	P
hot	sun	0.3
hot	rain	0.2
cold	sun	0.3
cold	rain	0.2

Example: Independence

- N fair, independent coin flips:



$P(X_1)$

H	0.5
T	0.5

$P(X_2)$

H	0.5
T	0.5

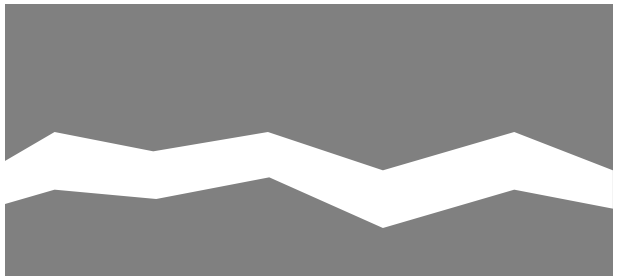
$\ldots$

$P(X_n)$

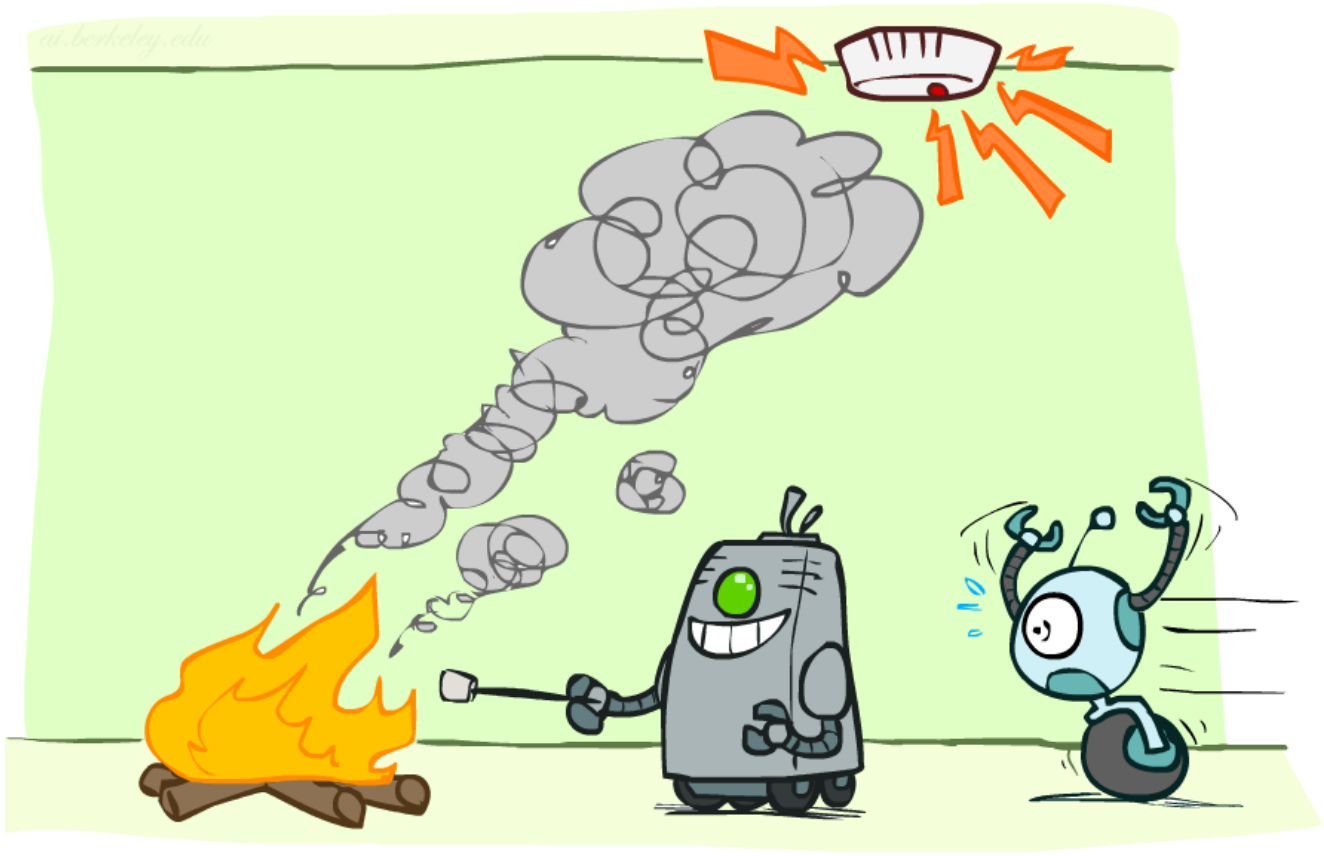
H	0.5
T	0.5

$P(X_1, X_2, \ldots X_n)$

2^n

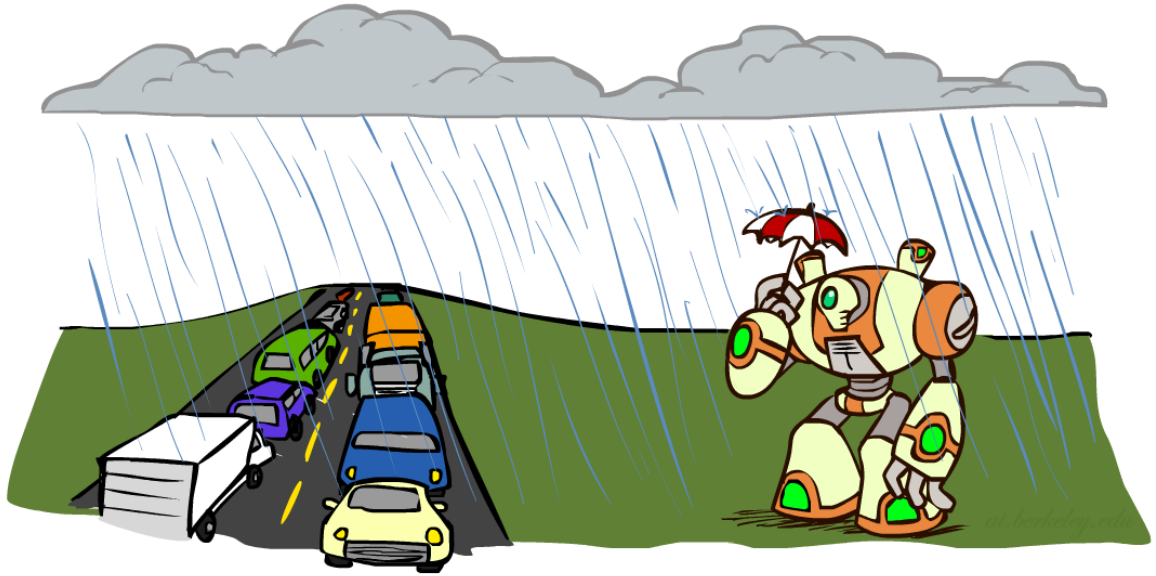


Conditional Independence



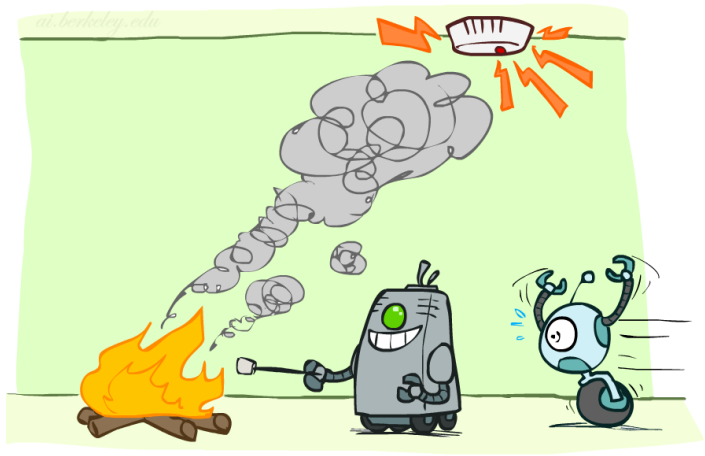
Conditional Independence

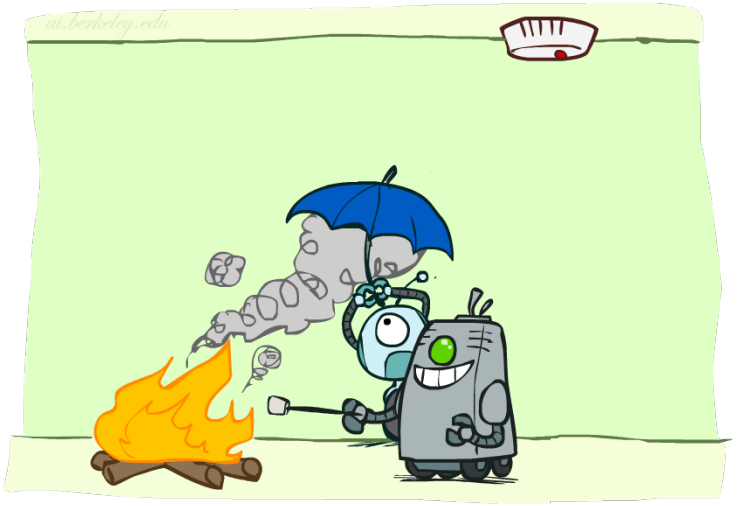
- What about this domain:
  - Traffic
  - Umbrella
  - Raining



Conditional Independence

- What about this domain:
  - Fire
  - Smoke
  - Alarm





Probability Recap

- Conditional probability $$P(x|y) = \frac{P(x,y)}{P(y)}$$
- Product rule $$P(x,y) = P(x|y)P(y)$$
- Chain rule $$\begin{aligned} P(X_1, X_2, \ldots X_n) &= P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)\ldots \\ &= \prod_{i=1}^{n} P(X_i|X_1, \ldots, X_{i-1}) \end{aligned}$$
- X, Y independent if and only if: $\forall x, y : P(x,y) = P(x)P(y)$
- X and Y are conditionally independent given Z if and only if: $X \perp\!\!\!\perp Y|Z$

$$\forall x, y, z : P(x,y|z) = P(x|z)P(y|z)$$

Independence

- Two variables are *independent* in a joint distribution if:

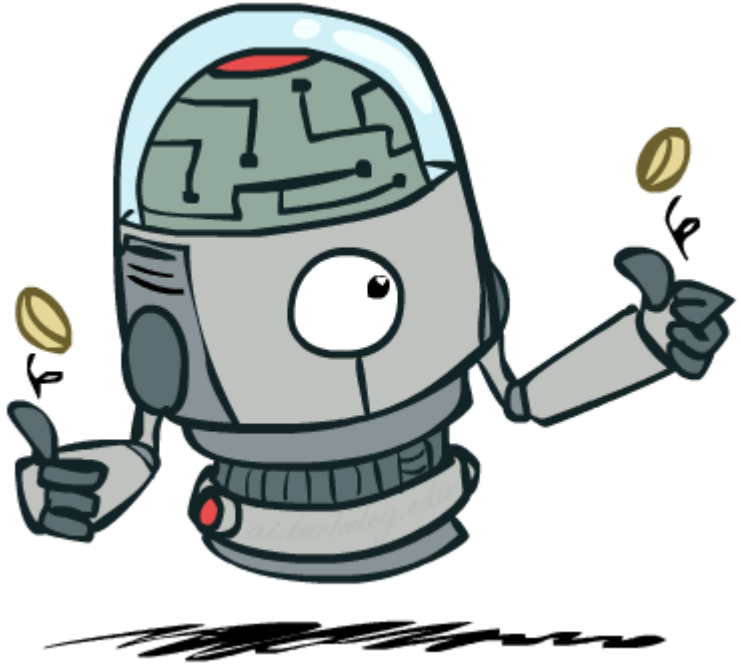
$$P(X,Y) = P(X)P(Y)$$

$$X \perp\!\!\!\perp Y$$

$$\forall x, y \; P(x,y) = P(x)P(y)$$

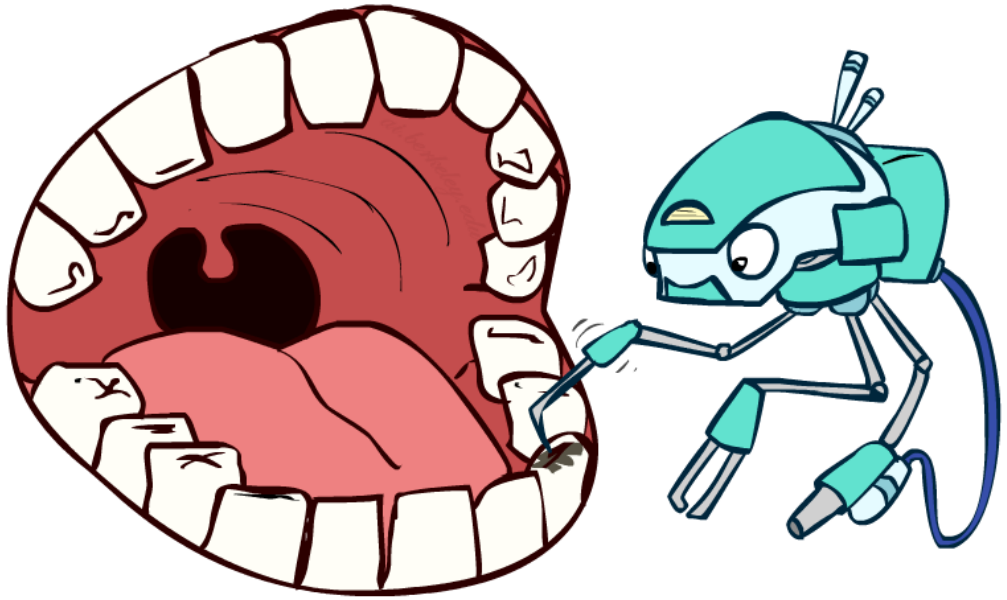
  - Says the joint distribution *factors* into a product of two simple ones
  - Usually variables aren't independent!

- Can use independence as a *modeling assumption*
  - Independence can be a simplifying assumption
  - *Empirical* joint distributions: at best "close" to independent
  - What could we assume for {Weather, Traffic, Cavity}?



Conditional Independence

- P(Toothache, Cavity, Catch)
- If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:
  - P(+catch | +toothache, +cavity) = P(+catch | +cavity)
- The same independence holds if I don’t have a cavity:
  - P(+catch | +toothache, -cavity) = P(+catch| -cavity)
- Catch is *conditionally independent* of Toothache given Cavity:
  - P(Catch | Toothache, Cavity) = P(Catch | Cavity)
- Equivalent statements:
  - P(Toothache | Catch , Cavity) = P(Toothache | Cavity)
  - P(Toothache, Catch | Cavity) = P(Toothache | Cavity) P(Catch | Cavity)
  - One can be derived from the other



Conditional Independence

- Unconditional (absolute) independence very rare (why?)
- *Conditional independence* is our most basic and robust form of knowledge about uncertain environments.
- X is conditionally independent of Y given Z $X \perp\!\!\!\perp Y | Z$

  if and only if:

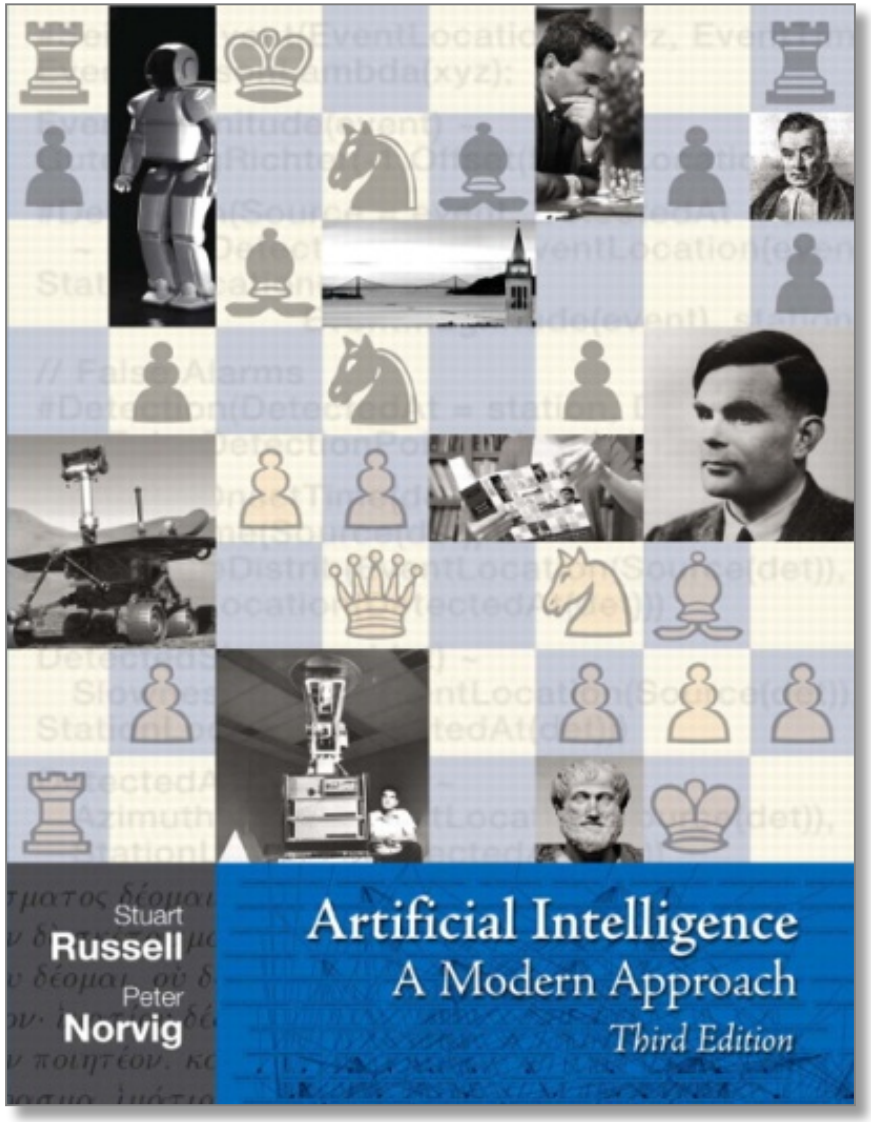
$$\forall x, y, z : P(x, y|z) = P(x|z)P(y|z)$$

or, equivalently, if and only if

$$\forall x, y, z : P(x|z, y) = P(x|z)$$

Probabilistic Reasoning Over Time

Read AIMA
Chapter 15 (15.1-15.5)



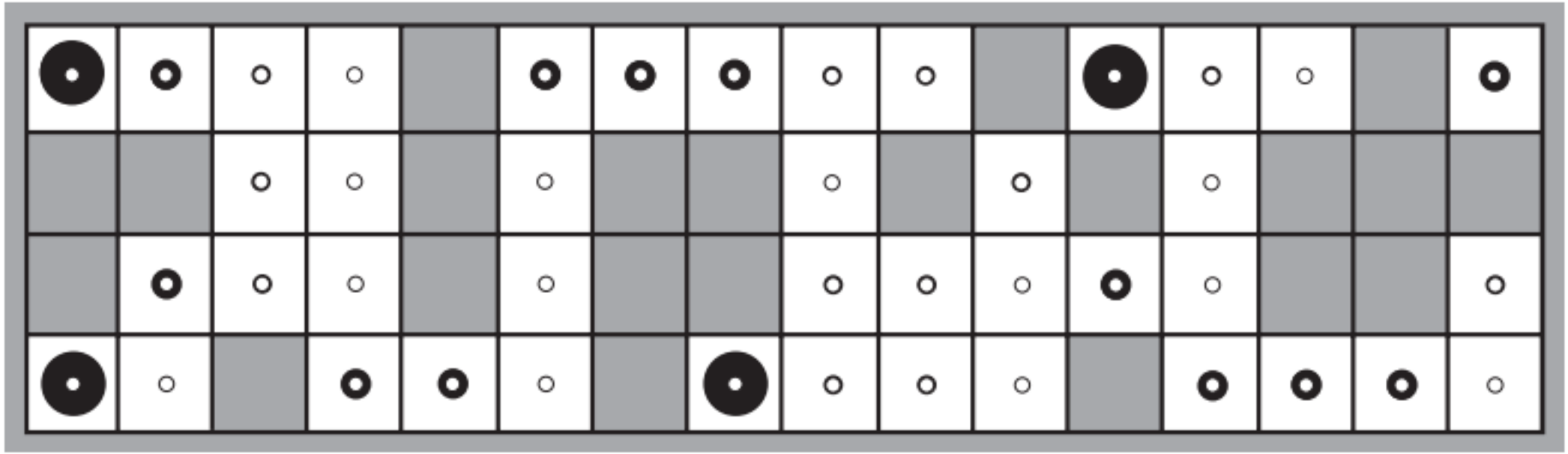
Some of these slides are courtesy of Dan Klein and Pieter Abbeel for UC Berkeley's Intro to AI course

Partially Observable Environments

- So far, we have considered **full-observable environments** like chess, or grid world
- In **partially-observable environments** we don't see everything
- In this case an agent needs to maintain a **belief state** that represents represents which states are currently possible

Robot Localization

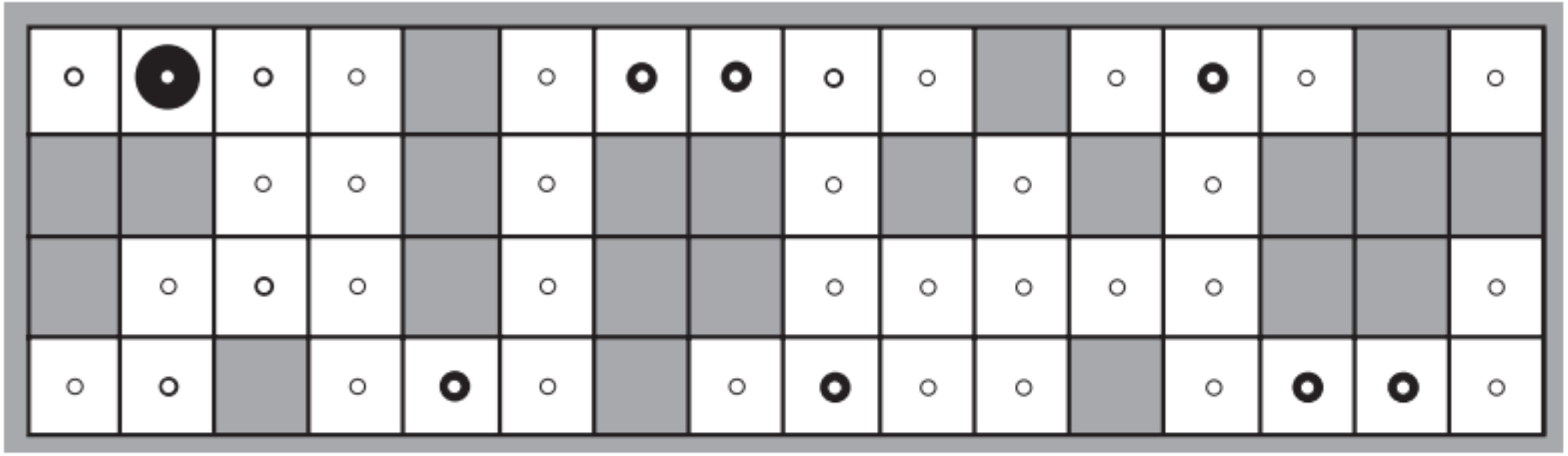
Given a map and sensor data, the robot updates its beliefs about where it is.



Sensor data with one observation saying there are obstacles to the North, South and West.

Robot Localization

Given a map and sensor data, the robot updates its beliefs about where it is.



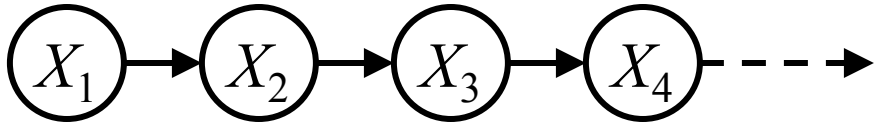
Sensor data with one observation saying there are obstacles to the North, South and West. The robot moves, and then takes another observation with obstacles to the North and South, and updates its belief state.

Reasoning over Time or Space

- Often, we want to reason about a sequence of observations
  - Robot localization
  - Speech recognition
  - Medical monitoring
- Need to introduce time (or space) into our models

Markov Models

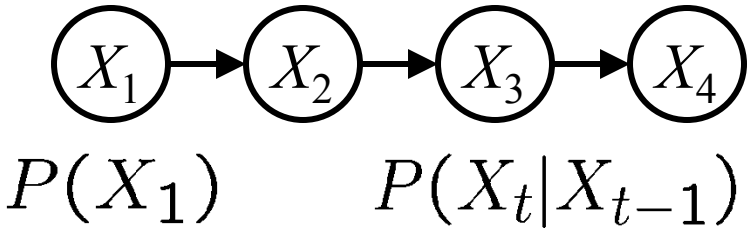
- Value of X at a given time is called the state



$$P(X_1) \qquad P(X_t|X_{t-1})$$

- Parameters: called transition probabilities or dynamics, specify how the state evolves over time (also, initial state probabilities)
- Stationary assumption: transition probabilities the same at all times
- Same as MDP transition model, but no choice of action

Joint Distribution of a Markov Model



- Joint distribution:

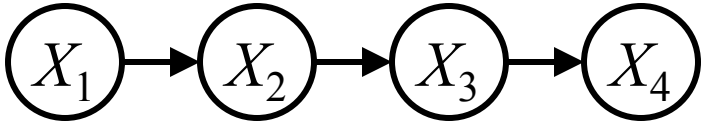
$$P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2|X_1)P(X_3|X_2)P(X_4|X_3)$$

- More generally:

$$P(X_1, X_2, \ldots, X_T) = P(X_1)P(X_2|X_1)P(X_3|X_2)\ldots P(X_T|X_{T-1})$$

$$= P(X_1)\prod_{t=2}^{T} P(X_t|X_{t-1})$$

Chain Rule and Markov Models



- From the chain rule, every joint distribution over X_1, X_2, X_3, X_4 can be written as:

$$P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)P(X_4|X_1, X_2, X_3)$$

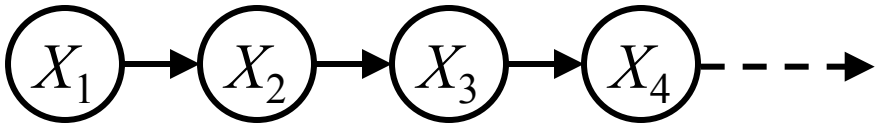
- Assuming that

$$X_3 \perp\!\!\!\perp X_1 \mid X_2 \quad \text{and} \quad X_4 \perp\!\!\!\perp X_1, X_2 \mid X_3$$

results in the expression posited on the previous slide:

$$P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2|X_1)P(X_3|X_2)P(X_4|X_3)$$

Chain Rule and Markov Models



- From the chain rule, every joint distribution over $X_1, X_2, \ldots, X_T$ can be written as:

$$P(X_1, X_2, \ldots, X_T) = P(X_1) \prod_{t=2}^{T} P(X_t | X_1, X_2, \ldots, X_{t-1})$$

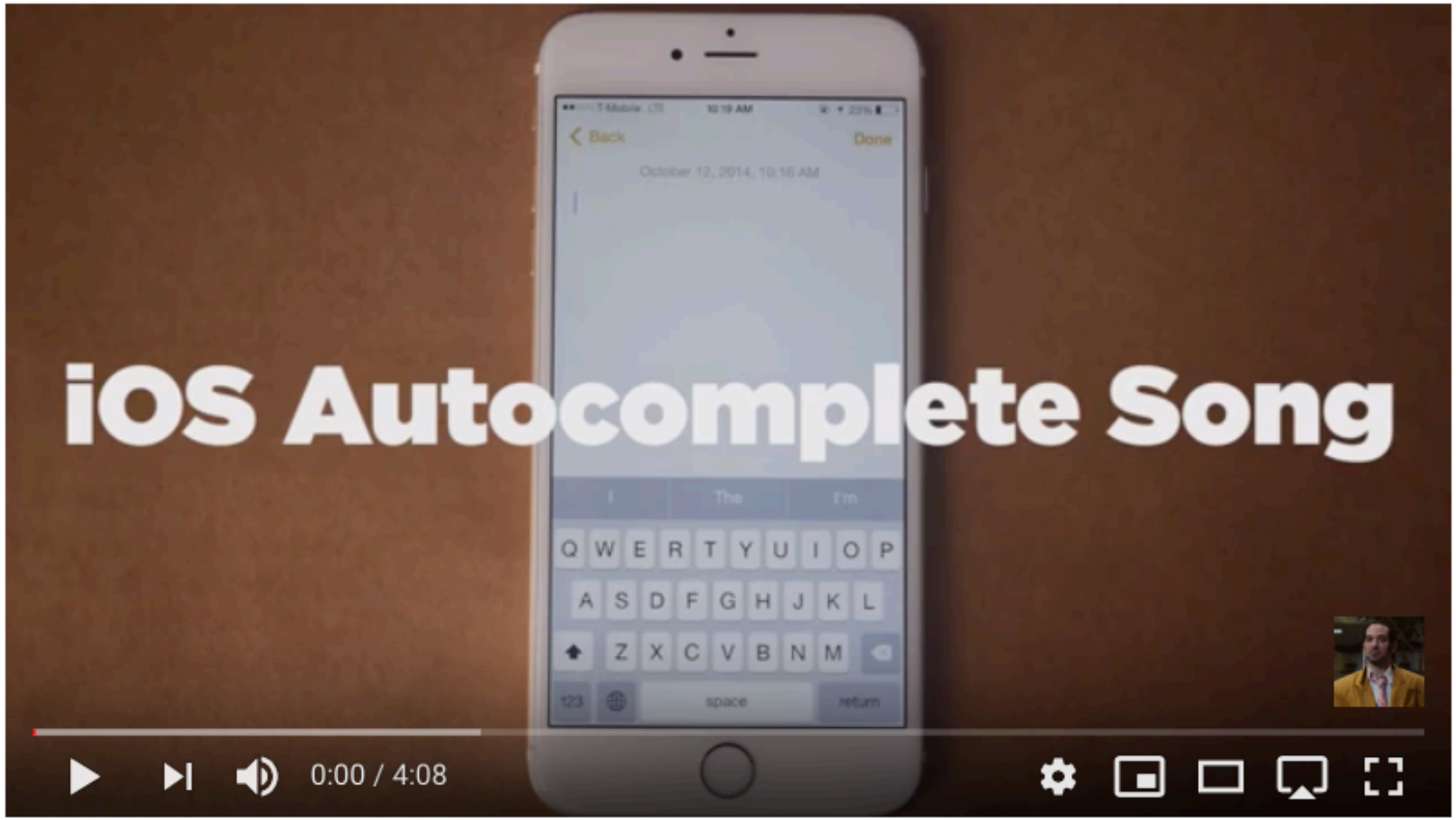
- Assuming that for all *t*:

$$X_t \perp\!\!\!\perp X_1, \ldots, X_{t-2} \mid X_{t-1}$$

gives us the expression posited on the earlier slide:

$$P(X_1, X_2, \ldots, X_T) = P(X_1) \prod_{t=2}^{T} P(X_t | X_{t-1})$$

Markov Models for Natural Language Processing



♫ iOS Autocomplete Song | Song A Day #2110

https://www.youtube.com/watch?v=M8MJFrdfGe0

Probabilistic Language Models

- One goal: assign a probability to a sentence
  - Autocomplete for texting
  - Machine Translation
  - Spelling Correction
  - Speech Recognition
- Other Natural Language Generation tasks: summarization, question-answering, dialog systems

Probabilistic Language Modeling

- Goal: compute the probability of a sentence or sequence of words
- Related task: probability of an upcoming word
- A model that computes either of these
- Better: **the grammar** But **language model** or **LM** is standard

Probabilistic Language Modeling

- Goal: compute the probability of a sentence or sequence of words

$$P(W) = P(w_1,w_2,w_3,w_4,w_5 \ldots w_n)$$

- Related task: probability of an upcoming word

$$P(w_5 | w_1,w_2,w_3,w_4)$$

- A model that computes either of these

$P(W)$ or $P(w_n | w_1,w_2 \ldots w_{n-1})$ is called a **language model**.

- Better: **the grammar** But **language model** or **LM** is standard

How to compute P(W)

- How to compute this joint probability:
  - P(the, underdog, Philadelphia, Eagles, won)
- Intuition: let's rely on the Chain Rule of Probability

The Chain Rule

The Chain Rule

- Recall the definition of conditional probabilities

  p(B|A) = P(A,B)/P(A) Rewriting: **P(A,B) = P(A)P(B|A)**

- More variables:

  $P(A,B,C,D) = P(A)P(B|A)P(C|A,B)P(D|A,B,C)$

- The Chain Rule in General

$$P(x_1,x_2,x_3,\ldots,x_n) = P(x_1)P(x_2|x_1)P(x_3|x_1,x_2)\ldots P(x_n|x_1,\ldots,x_{n-1})$$

The Chain Rule applied to compute joint probability of words in sentence

The Chain Rule applied to compute joint probability of words in sentence

$$P(w_1 w_2 \ldots w_n) = \prod_i P(w_i \mid w_1 w_2 \ldots w_{i-1})$$

P("the underdog Philadelphia Eagles won") =

P(the) × P(underdog|the) × P(Philadelphia|the underdog)

× P(Eagles|the underdog Philadelphia)

× P(won|the underdog Philadelphia Eagles)

How to estimate these probabilities

- Could we just count and divide?

How to estimate these probabilities

- Could we just count and divide? Maximum likelihood estimation (MLE)

$$P(\text{won} \mid \text{the underdog team}) = \frac{\text{Count(the underdog team won)}}{\text{Count(the underdog team)}}$$

- Why doesn't this work?

Simplifying Assumption = Markov Assumption

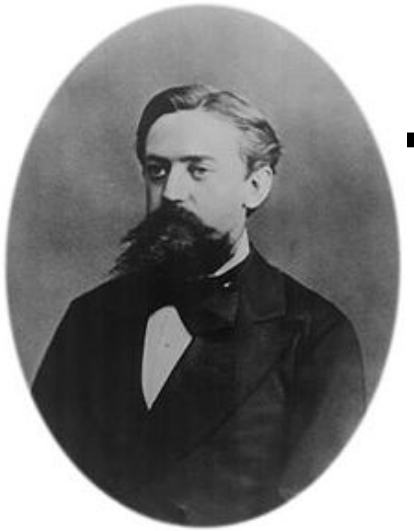
Simplifying Assumption = Markov Assumption

- $P(won|the\ underdog\ team) \approx P(won|team)$
- Only depends on the previous k words, not the whole context
- $\approx P(won|underdog\ team)$
- $\approx P(w_i|w_{i-2}\ w_{i-1})$
- $P(w_1w_2w_3w_4...w_n) \approx \prod_i^n P(w_i|w_{i-k} \dots w_{i-1})$
- K is the number of context words that we take into account

How much history should we use?

unigram	**no history**	$\prod_{i}^{n} \mathrm{p}(w_i)$	$p(w_i) = \frac{count(w_i)}{all\ words}$
bigram	1 word as history	$\prod_{i}^{n} \mathrm{p}(w_i \mid w_{i-1})$	$p(w_i \mid w_{i-1}) = \frac{count(w_{i-1} w_i)}{count(w_{i-1})}$
trigram	2 words as history	$\prod_{i}^{n} \mathrm{p}(w_i \mid w_{i-2} w_{i-1})$	$p(w_i \mid w_{i-2} w_{i-1}) = \frac{count(w_{i-2} w_{i-1} w_i)}{count(w_{i-2} w_{i-1})}$
4-gram	3 words as history	$\prod_{i}^{n} \mathrm{p}(w_i \mid w_{i-3} w_{i-2} w_{i-1})$	$p(w_i \mid w_{i-3} w_{i-2} w_{i-1}) = \frac{count(w_{i-3} w_{i-2} w_{i-1} w_i)}{count(w_{i-3} w_{i-3} w_{i-1})}$

Historical Notes



1913	**Andrei Markov counts 20k letters in *Eugene Onegin***
1948	Claude Shannon uses n-grams to approximate English
1956	Noam Chomsky decries finite-state Markov Models
1980s	Fred Jelinek at IBM TJ Watson uses n-grams for ASR, think about 2 other ideas for models: (1) MT, (2) stock market prediction
1993	Jelinek at team develops statistical machine translation $argmax_e p(e\|f) = p(e)\, p(f\|e)$
	Jelinek left IBM to found CLSP at JHU Peter Brown and Robert Mercer move to Renaissance Technology

Simplest case: Unigram model

$$P(w_1 w_2 \ldots w_n) \approx \prod_i P(w_i)$$

Some automatically generated sentences from a unigram model

```
fifth an of futures the an incorporated a a the
inflation most dollars quarter in is mass

thrift did eighty said hard 'm july bullish

that or limited the
```

Bigram model

- Condition on the previous word:

$$P(w_i \mid w_1 w_2 \ldots w_{i-1}) \approx P(w_i \mid w_{i-1})$$

```
texaco rose one in this issue is pursuing growth in a boiler
house said mr. gurria mexico 's motion control proposal

without permission from five hundred fifty five yen

outside new car parking lot of the agreement reached

this would be a record november
```

N-gram models

- We can extend to trigrams, 4-grams, 5-grams
- In general this is an insufficient model of language
  - because language has **long-distance dependencies**:

"The computer(s) which I had just put into the machine room on the fifth floor is (are) crashing."

- But we can often get away with N-gram models

Estimating N-gram Probabilities

LANGUAGE MODELING

Estimating bigram probabilities

- The Maximum Likelihood Estimate

$$P(w_i \mid w_{i-1}) = \frac{count(w_{i-1}, w_i)}{count(w_{i-1})}$$

$$P(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})}$$

An example

$$P(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})}$$

<s> I am Sam </s>

<s> Sam I am </s>

<s> I do not like green eggs and ham </s>

$$P(\texttt{I} \mid \texttt{<s>}) = \tfrac{2}{3} = .67 \qquad P(\texttt{Sam} \mid \texttt{<s>}) = \tfrac{1}{3} = .33 \qquad P(\texttt{am} \mid \texttt{I}) = \tfrac{2}{3} = .67$$

$$P(\texttt{</s>} \mid \texttt{Sam}) = \tfrac{1}{2} = 0.5 \qquad P(\texttt{Sam} \mid \texttt{am}) = \tfrac{1}{2} = .5 \qquad P(\texttt{do} \mid \texttt{I}) = \tfrac{1}{3} = .33$$

Problems for MLE

- Zeros

Train	Test
denied the allegations	denied the memo
denied the reports	
denied the claims	
denied the requests	

- P(memo|denied the) = 0
- And we also assign 0 probability to all sentences containing it!

Problems for MLE

- Out of vocabulary items (OOV)
- <unk> to deal with OOVs
- Fixed lexicon L of size V
- Normalize training data by replacing any word not in L with <unk>

- Avoid zeros with smoothing

Practical Issues

- We do everything in log space
  - Avoid underflow
  - (also adding is faster than multiplying)

$$\log(p_1 \times p_2 \times p_3 \times p_4) = \log p_1 + \log p_2 + \log p_3 + \log p_4$$

Language Modeling Toolkits

- SRILM
  - http://www.speech.sri.com/projects/srilm/
- KenLM
  - https://kheafield.com/code/kenlm/

Google N-Gram Release, August 2006

AUG
3

All Our N-gram are Belong to You

Posted by Alex Franz and Thorsten Brants, Google Machine Translation Team

Here at Google Research we have been using word n-gram models for a variety of R&D projects,

...

That's why we decided to share this enormous dataset with everyone. We processed 1,024,908,267,229 words of running text and are publishing the counts for all 1,176,470,663 five-word sequences that appear at least 40 times. There are 13,588,391 unique words, after discarding words that appear less than 200 times.

Google N-Gram Release

- serve as the incoming 92
- serve as the incubator 99
- serve as the independent 794
- serve as the index 223
- serve as the indication 72
- serve as the indicator 120
- serve as the indicators 45
- serve as the indispensable 111
- serve as the indispensible 40
- serve as the individual 234

http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html