School of Computer Science
Carnegie Mellon

10-601B Introduction to Machine Learning

Directed Graphical Models (aka. Bayesian Networks)

Readings:
Bishop 8.1 and 8.2.2
Mitchell 6.11
Murphy 10

Matt Gormley
Lecture 21
November 9, 2016

Reminders

- Homework 6
  - due Mon., Nov. 21
- Final Exam
  - in-class Wed., Dec. 7

Outline

- **Motivation**
  - Structured Prediction
- **Background**
  - Conditional Independence
  - Chain Rule of Probability
- **Directed Graphical Models**
  - Bayesian Network definition
  - Qualitative Specification
  - Quantitative Specification
  - Familiar Models as Bayes Nets
  - Example: The Monty Hall Problem
- **Conditional Independence in Bayes Nets**
  - Three case studies
  - D-separation
  - Markov blanket

MOTIVATION

Structured Prediction

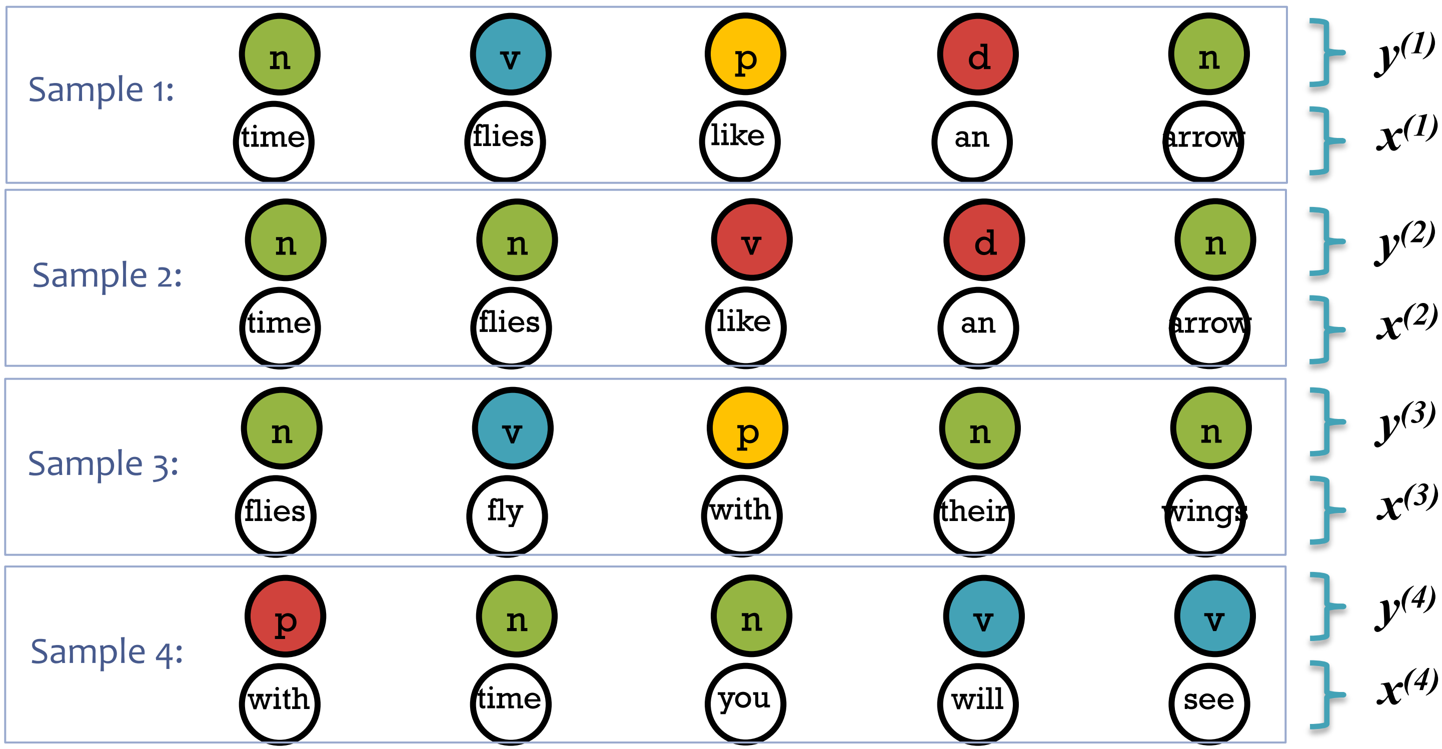
- Most of the models we've seen so far were for **classification**
  - Given observations: $\boldsymbol{x} = (x_1, x_2, \ldots, x_K)$
  - Predict a (binary) **label:** y
- Many real-world problems require **structured prediction**
  - Given observations: $\boldsymbol{x} = (x_1, x_2, \ldots, x_K)$
  - Predict a **structure:** $\boldsymbol{y} = (y_1, y_2, \ldots, y_J)$
- Some *classification* problems benefit from **latent structure**

Structured Prediction Examples

- **Examples of structured prediction**
  - Part-of-speech (POS) tagging
  - Handwriting recognition
  - Speech recognition
  - Word alignment
  - Congressional voting
- **Examples of latent structure**
  - Object recognition

Dataset for Supervised Part-of-Speech (POS) Tagging

Data: $\mathcal{D} = \{\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)}\}_{n=1}^{N}$



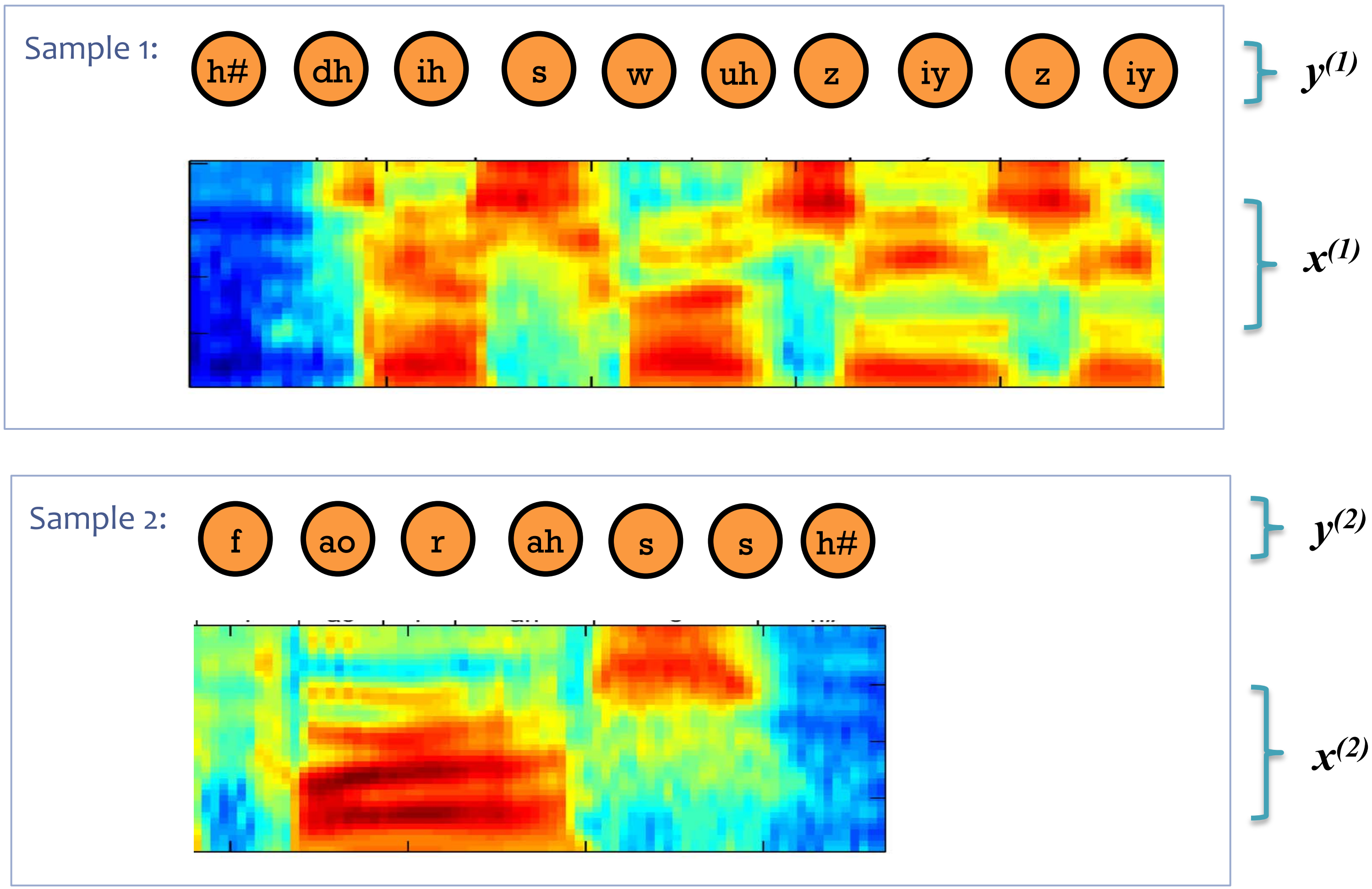
Dataset for Supervised Handwriting Recognition

Data: $\mathcal{D} = \{\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)}\}_{n=1}^{N}$

Figures from (Chatzis & Demiris, 2013)

Dataset for Supervised Phoneme (Speech) Recognition

Data: $\mathcal{D} = \{\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)}\}_{n=1}^{N}$

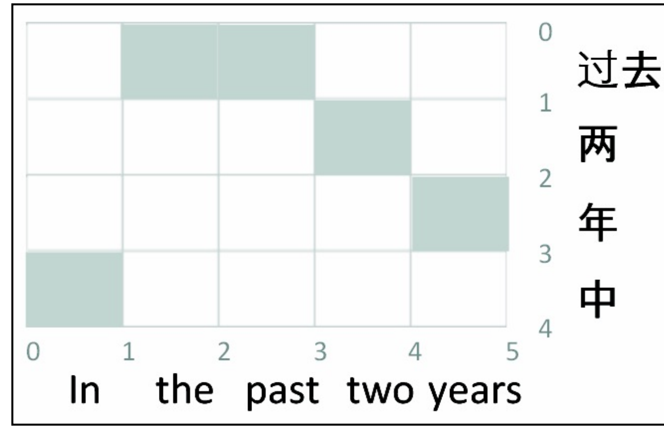


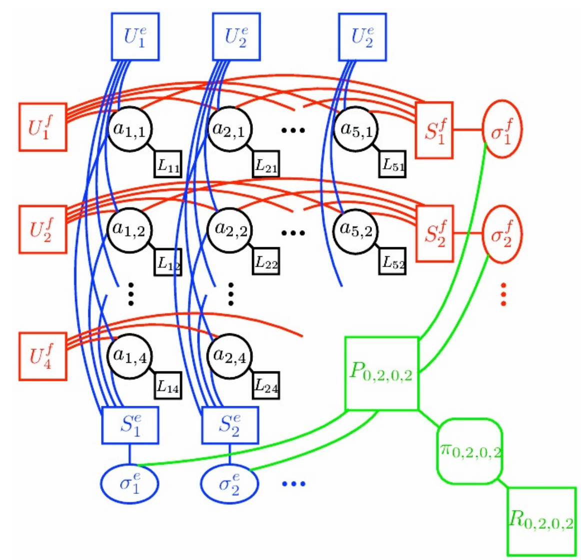
Figures from (Jansen & Niyogi, 2013)

Application:

Word Alignment / Phrase Extraction

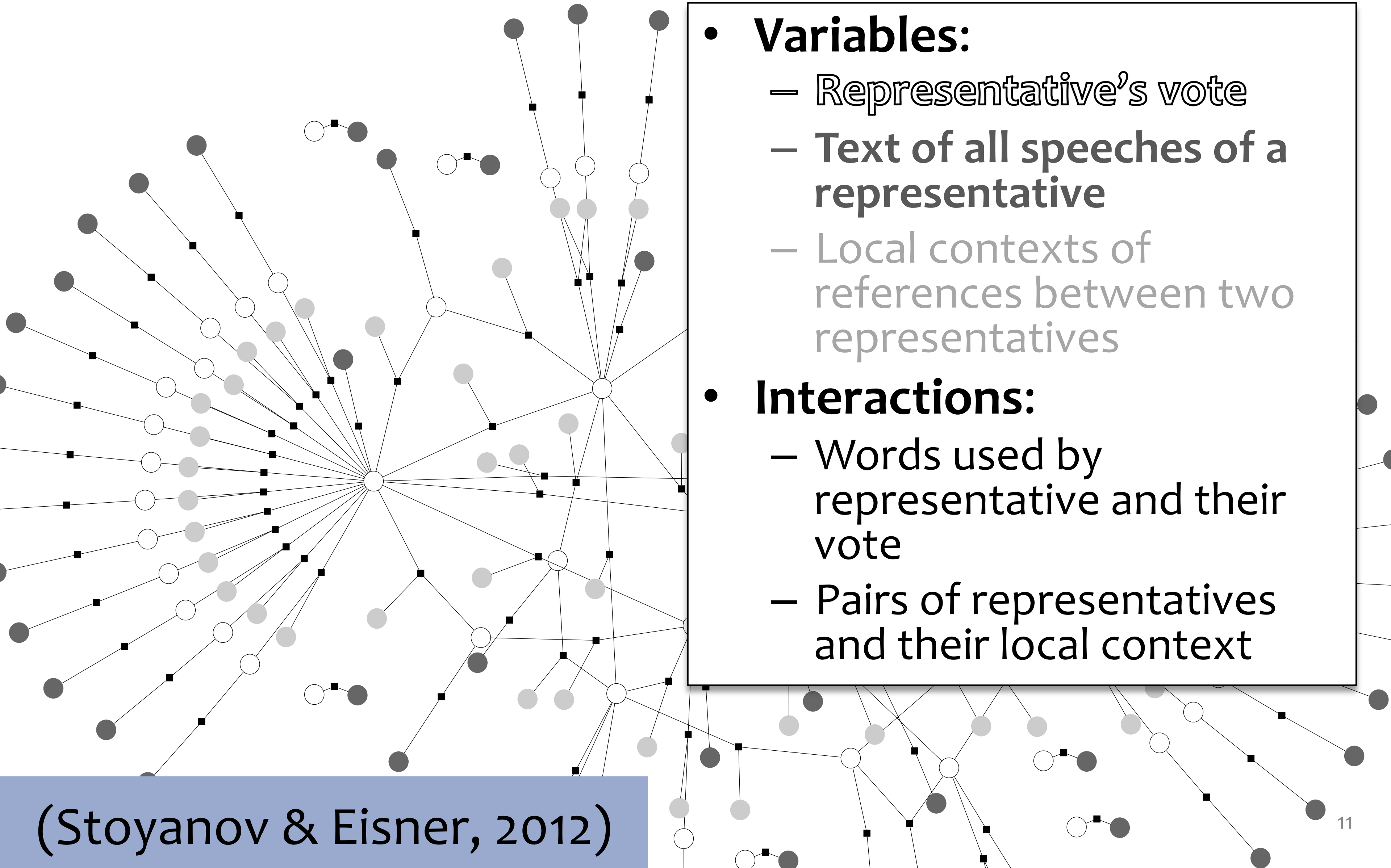
- **Variables (boolean):**
  - For each (Chinese phrase, English phrase) pair, are they linked?
- **Interactions:**
  - Word fertilities
  - Few “jumps” (discontinuities)
  - Syntactic reorderings
  - “ITG contraint” on alignment
  - Phrases are disjoint (?)





(Burkett & Klein, 2012)

Congressional Voting



- **Variables:**
  - Representative's vote
  - **Text of all speeches of a representative**
  - Local contexts of references between two representatives
- **Interactions:**
  - Words used by representative and their vote
  - Pairs of representatives and their local context

(Stoyanov & Eisner, 2012)

Structured Prediction Examples

- **Examples of structured prediction**
  - Part-of-speech (POS) tagging
  - Handwriting recognition
  - Speech recognition
  - Word alignment
  - Congressional voting
- **Examples of latent structure**
  - Object recognition

Case Study: Object Recognition

Data consists of images $\boldsymbol{x}$ and labels y.

$x^{(1)}$

pigeon $y^{(1)}$

$x^{(2)}$

rhinoceros $y^{(2)}$

$x^{(3)}$

leopard $y^{(3)}$

$x^{(4)}$

llama $y^{(4)}$

Case Study: Object Recognition

Data consists of images x and labels y.

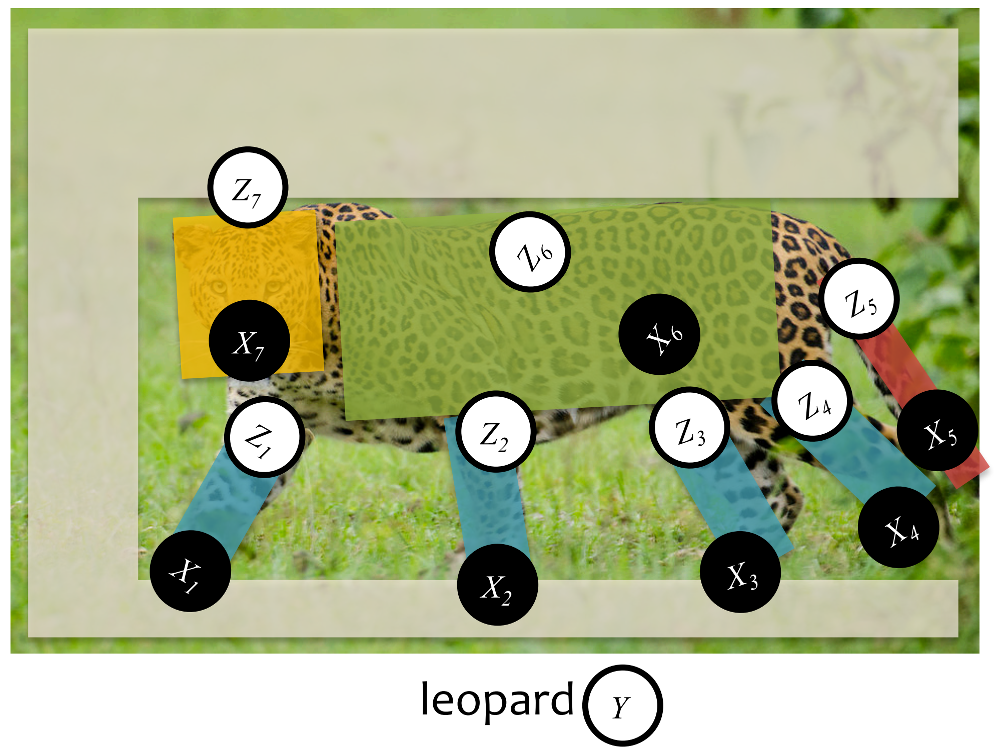
- Preprocess data into "patches"
- Posit a latent labeling z describing the object's parts (e.g. head, leg, tail, torso, grass)
- Define graphical model with these latent variables in mind
- z is not observed at train or test time

leopard

Case Study: Object Recognition

Data consists of images $\boldsymbol{x}$ and labels y.

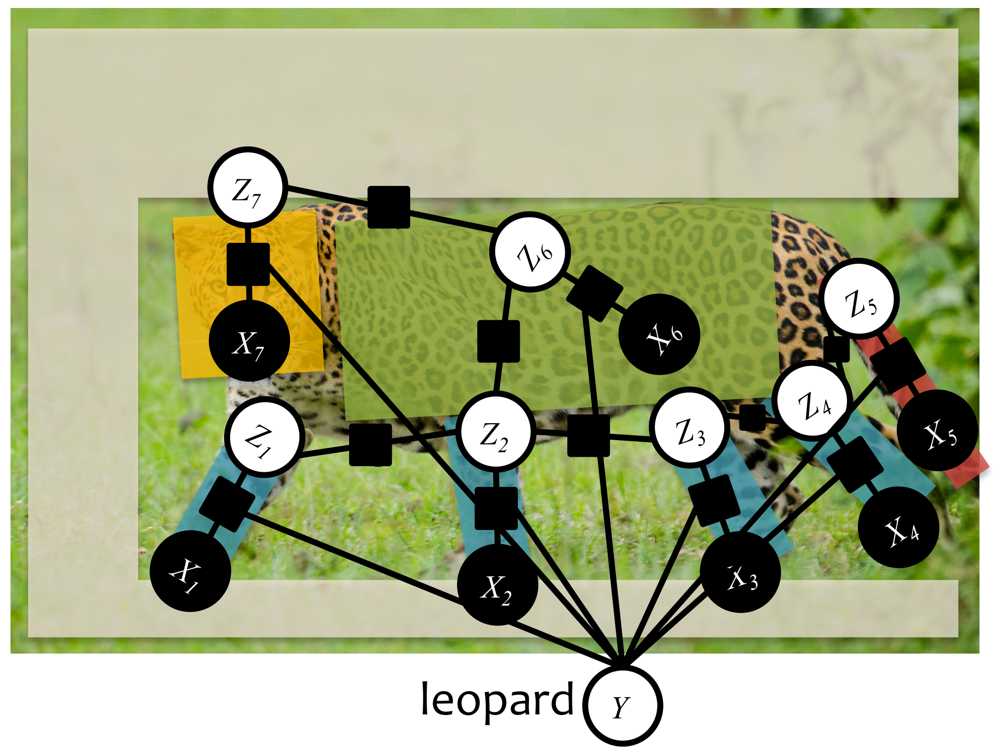
- Preprocess data into "patches"
- Posit a latent labeling z describing the object's parts (e.g. head, leg, tail, torso, grass)
- Define graphical model with these latent variables in mind
- z is not observed at train or test time



Case Study: Object Recognition

Data consists of images $\boldsymbol{x}$ and labels y.

- Preprocess data into "patches"
- Posit a latent labeling z describing the object's parts (e.g. head, leg, tail, torso, grass)
- Define graphical model with these latent variables in mind
- z is not observed at train or test time



Structured Prediction

Preview of challenges to come…

- Consider the task of finding the **most probable assignment** to the output

Classification

$$\hat{y} = \underset{y}{\operatorname{argmax}}\, p(y|\mathbf{x})$$

where $y \in \{+1, -1\}$

Structured Prediction

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}}\, p(\mathbf{y}|\mathbf{x})$$

where $\mathbf{y} \in \mathcal{Y}$

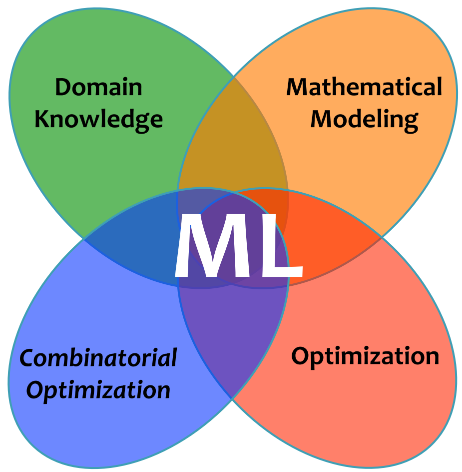
and $|\mathcal{Y}|$ is very large

Machine Learning

The **data** inspires the structures we want to predict

Our **model** defines a score for each structure

It also tells us what to optimize



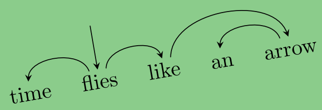
Inference finds {best structure, marginals, partition function} for a new observation

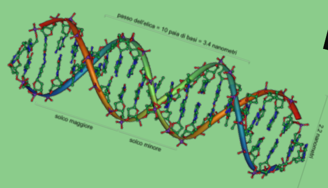
Learning tunes the parameters of the model

(**Inference** is usually called as a subroutine in learning)

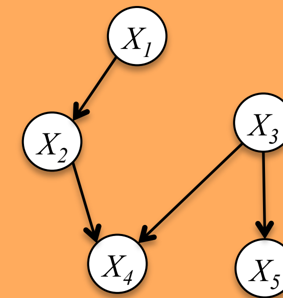
Machine Learning

Data

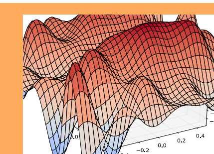




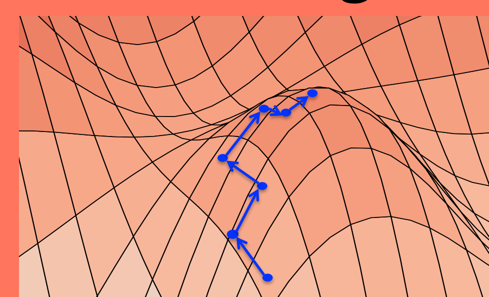
Model



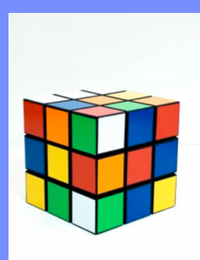
Objective



Learning



Inference



(**Inference** is usually called as a subroutine in learning)

BACKGROUND

Background: Chain Rule of Probability

For random variables A and B:

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

For random variables X_1, X_2, X_3, X_4:

$$\begin{aligned} P(X_1, X_2, X_3, X_4) = & P(X_1|X_2, X_3, X_4) \\ & P(X_2|X_3, X_4) \\ & P(X_3|X_4) \\ & P(X_4) \end{aligned}$$

Background: Conditional Independence

Random variables A and B are conditionally independent given C if:

$$P(A, B|C) = P(A|C)P(B|C) \qquad (1)$$

or equivalently:

$$P(A|B, C) = P(A|C) \qquad (2)$$

We write this as:

$$A \perp\!\!\!\perp B|C$$

Later we will also write: $I\langle A, \{C\}, B\rangle$

Bayesian Networks

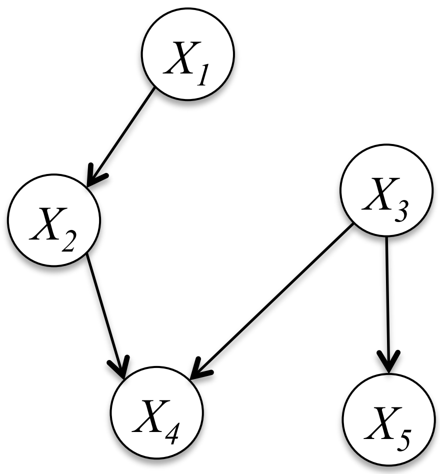
DIRECTED GRAPHICAL MODELS

Whiteboard

Writing Joint Distributions

- Strawman: Giant Table
- Alternate #1: Rewrite using chain rule
- Alternate #2: Assume full independence
- Alternate #3: Drop variables from RHS of conditionals

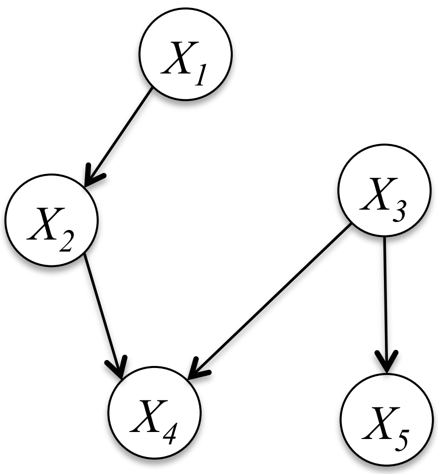
Bayesian Network



Definition:

$$P(X_1 \ldots X_n) = \prod_{i=1}^{n} P(X_i \mid parents(X_i))$$

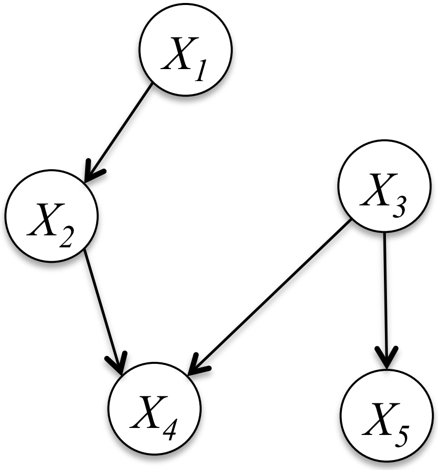
Bayesian Network



Definition:

$$p(X_1, X_2, X_3, X_4, X_5) =$$
$$p(X_5|X_3)p(X_4|X_2, X_3)$$
$$p(X_3)p(X_2|X_1)p(X_1)$$

Bayesian Network



Definition:

$$p(X_1, X_2, X_3, X_4, X_5) = p(X_5|X_3)p(X_4|X_2, X_3) p(X_3)p(X_2|X_1)p(X_1)$$

- A Bayesian Network is a **directed graphical model**
- It consists of a graph **G** and the conditional probabilities **P**
- These two parts full specify the distribution:
  - Qualitative Specification: **G**
  - Quantitative Specification: **P**

Qualitative Specification

- Where does the qualitative specification come from?
  - Prior knowledge of causal relationships
  - Prior knowledge of modular relationships
  - Assessment from experts
  - Learning from data
  - We simply link a certain architecture (e.g. a layered graph)
  - …

Whiteboard

If time…

- Example: 2016 Presidential Election

Towards quantitative specification of probability distribution

- Separation properties in the graph imply independence properties about the associated variables
- For the graph to be useful, any conditional independence properties we can derive from the graph should hold for the probability distribution that the graph represents

- **The Equivalence Theorem**

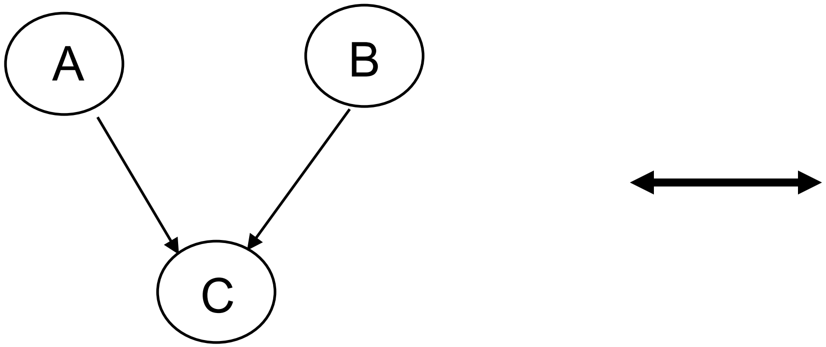
  For a graph G,

  Let D_1 denote the family of all distributions that satisfy I(G),

  Let D_2 denote the family of all distributions that factor according to G,

  Then $D_1 \equiv D_2$.

Quantitative Specification



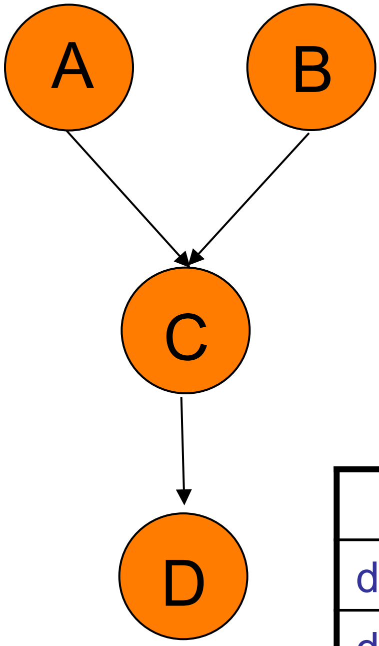
$p(A,B,C) =$

Conditional probability tables (CPTs)

a^0	0.75
a^1	0.25

b^0	0.33
b^1	0.67

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



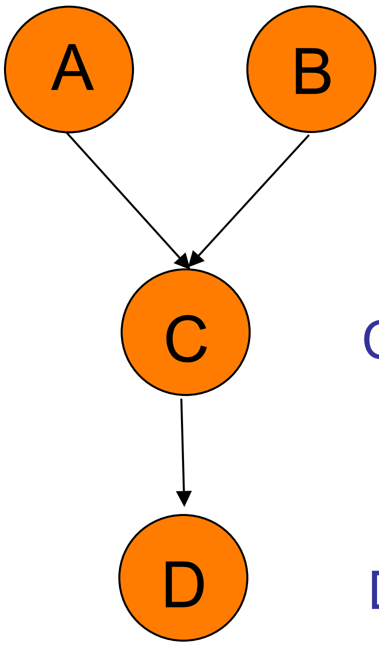
	a^0b^0	a^0b^1	a^1b^0	a^1b^1
c^0	0.45	1	0.9	0.7
c^1	0.55	0	0.1	0.3

	c^0	c^1
d^0	0.3	0.5
d^1	07	0.5

Conditional probability density func. (CPDs)

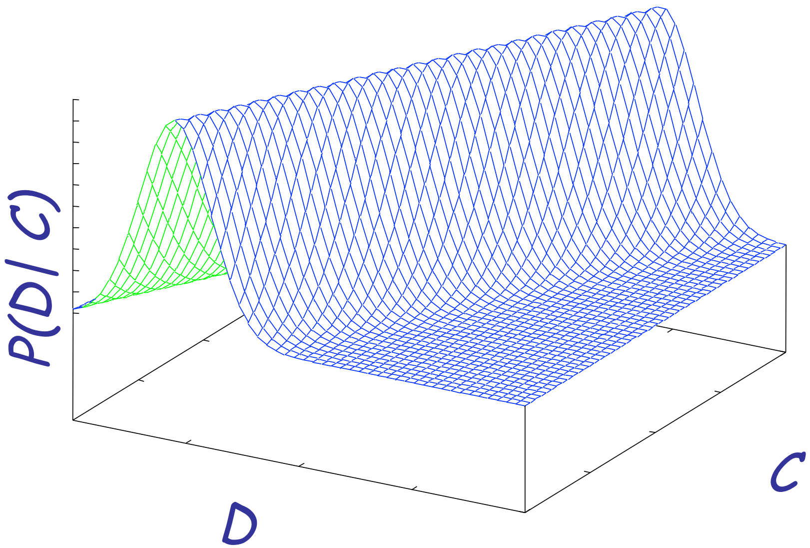
$A \sim N(\mu_a, \Sigma_a)$ $B \sim N(\mu_b, \Sigma_b)$

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

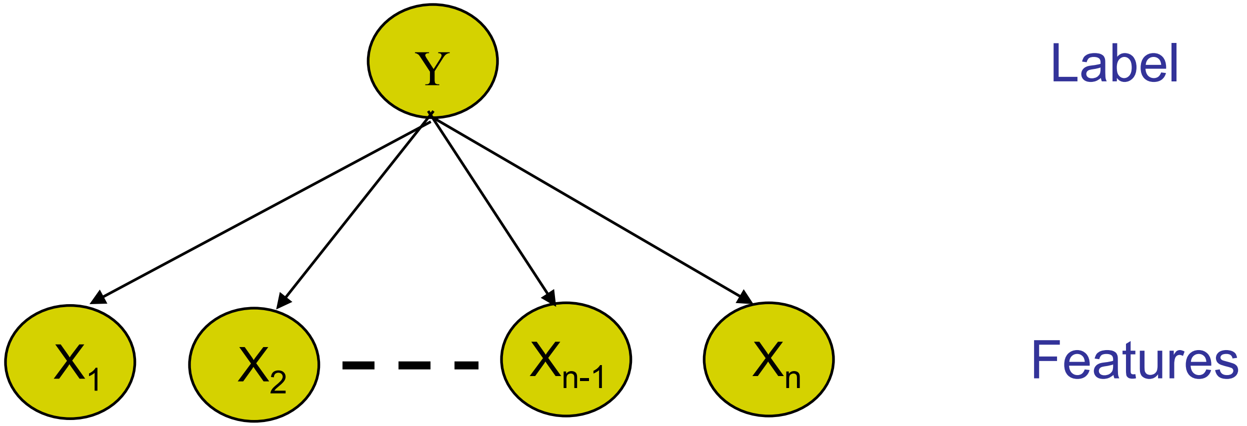


$C \sim N(A+B, \Sigma_c)$

$D \sim N(\mu_a + C, \Sigma_a)$



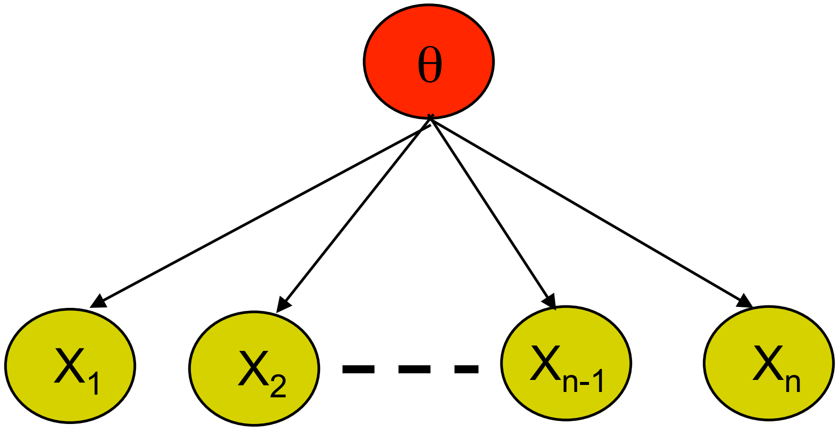
Conditional Independencies



What is this model

1. When Y is observed?
2. When Y is unobserved?

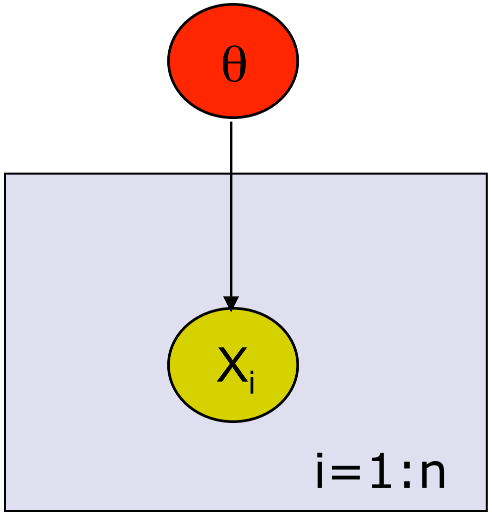
Conditionally Independent Observations



Model parameters

Data = $\{y_1, \ldots y_n\}$

"Plate" Notation



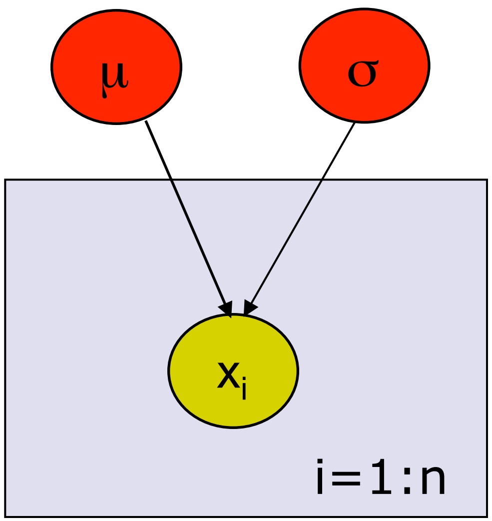
Model parameters

Data = $\{x_1, \ldots x_n\}$

Plate = rectangle in graphical model

variables within a plate are replicated
in a conditionally independent manner

Example: Gaussian Model



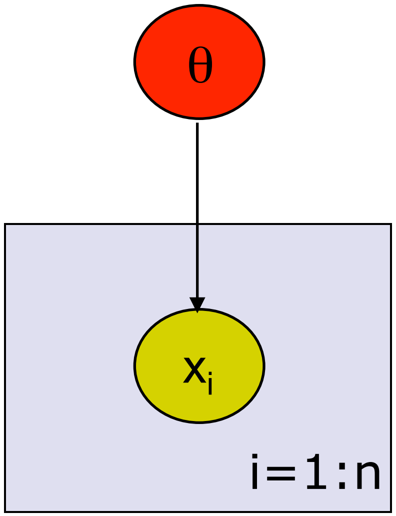
Generative model:

$$p(x_1,\ldots x_n \mid \mu, \sigma) = \prod p(x_i \mid \mu, \sigma)$$
$$= p(\text{data} \mid \text{parameters})$$
$$= p(D \mid \theta)$$

where $\theta = \{\mu, \sigma\}$

- Likelihood $= p(\text{data} \mid \text{parameters})$
  $= p(D \mid \theta)$
  $= L(\theta)$
- Likelihood tells us how likely the observed data are conditioned on a particular setting of the parameters
  - Often easier to work with $\log L(\theta)$

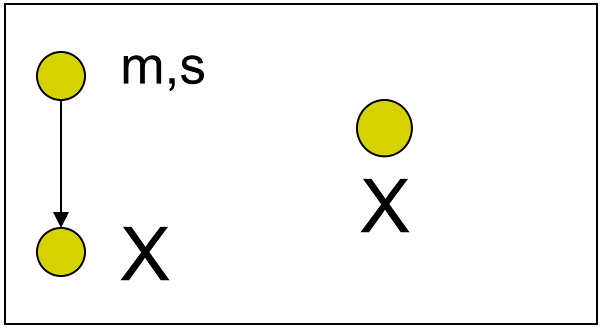
Bayesian models



More examples

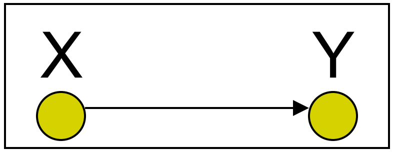
Density estimation

Parametric and nonparametric methods



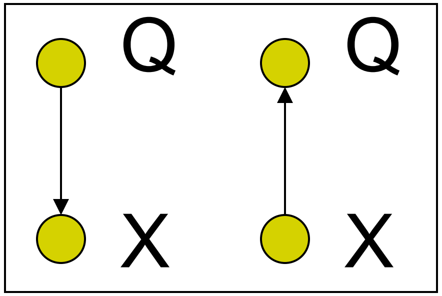
Regression

Linear, conditional mixture, nonparametric



Classification

Generative and discriminative approach



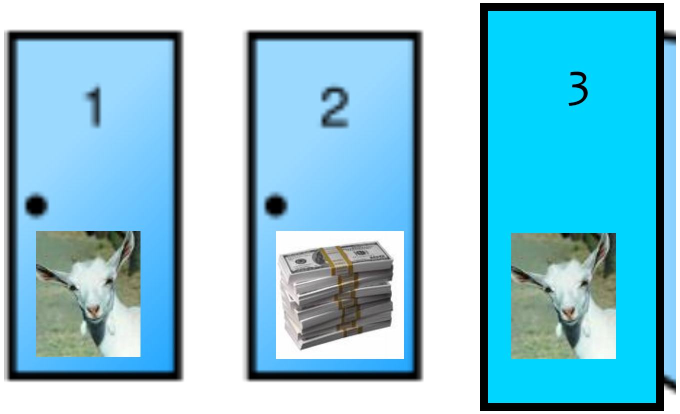
EXAMPLE:
THE MONTY HALL PROBLEM

Extra slides from last semester

Slide from William Cohen

The (highly practical) Monty Hall problem

- You're in a game show. Behind one door is a prize. Behind the others, goats.
- You pick one of three doors, say #1
- The host, Monty Hall, opens one door, revealing… a goat!



You now can either

- stick with your guess
- always change doors
- flip a coin and pick a new door randomly according to the coin

Extra slides from last semester

Slide from William Cohen

The (highly practical) Monty Hall problem

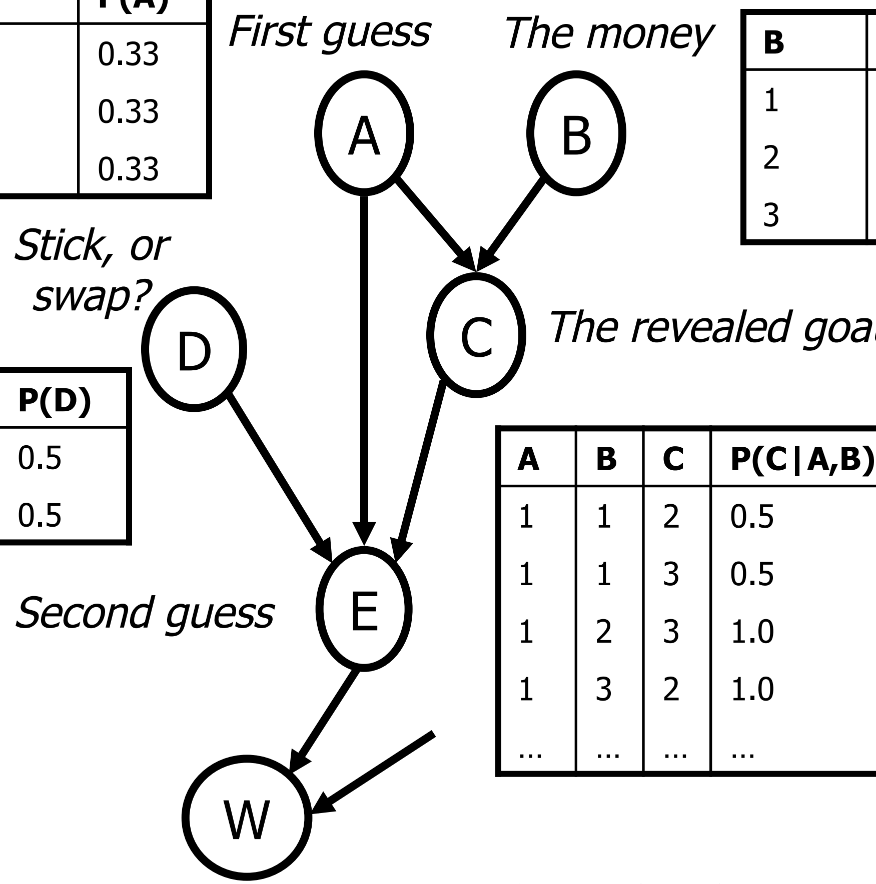
Extra slides from last semester

- You're in a game show. Behind one door is a prize. Behind the others, goats.
- You pick one of three doors, say #1
- The host, Monty Hall, opens one door, revealing… a goat!
- You now can either stick with your guess or change doors

A	P(A)
1	0.33
2	0.33
3	0.33

B	P(B)
1	0.33
2	0.33
3	0.33

D	P(D)
Stick	0.5
Swap	0.5



A	B	C	P(C|A,B)
1	1	2	0.5
1	1	3	0.5
1	2	3	1.0
1	3	2	1.0
…	…	…	…

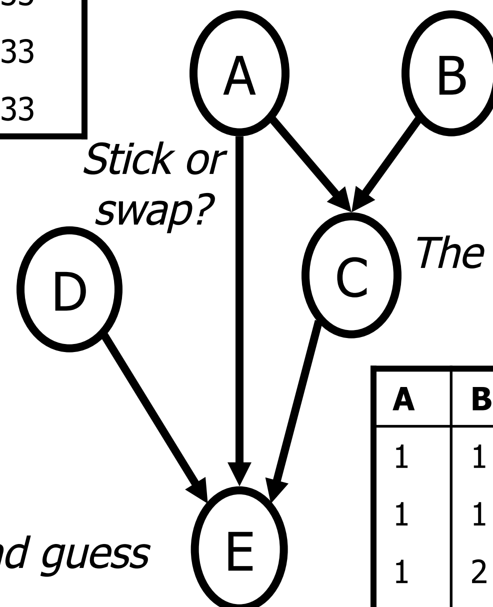
$$P(C=c \mid A=a, B=b) = \begin{cases} 1.0 & \text{if } (a \neq b) \wedge (c \notin \{a,b\}) \\ 0.5 & \text{if } (a = b) \wedge (c \notin \{a,b\}) \\ 0 & \text{otherwise} \end{cases}$$

Slide from William Cohen

The (highly practical) Monty Hall problem

Extra slides from last semester

A	P(A)
1	0.33
2	0.33
3	0.33



B	P(B)
1	0.33
2	0.33
3	0.33

$$P(E=e \mid A,C,D) = \begin{cases} 1.0 & \text{if } (e=a) \wedge (d=stick) \\ 1.0 & \text{if } (e \notin \{a,c\}) \wedge (d=swap) \\ 0 & \text{otherwise} \end{cases}$$

A	B	C	P(C|A,B)
1	1	2	0.5
1	1	3	0.5
1	2	3	1.0
1	3	2	1.0
…	…	…	…

If you stick: you win if your first guess was right.

If you swap: you win if your first guess was wrong.

A	C	D	P(E|A,C,D)
…	…	…	…

$$P(C=c \mid A=a, B=b) = \begin{cases} 1.0 & \text{if } (a \neq b) \wedge (c \notin \{a,b\}) \\ 0.5 & \text{if } (a = b) \wedge (c \notin \{a,b\}) \\ 0 & \text{otherwise} \end{cases}$$

Slide from William Cohen

The (highly practical) Monty Hall problem

We could construct the joint and compute P(E=B|D=swap)

…again by the chain rule:

P(A,B,C,D,E) =

P(E|A,C,D) *

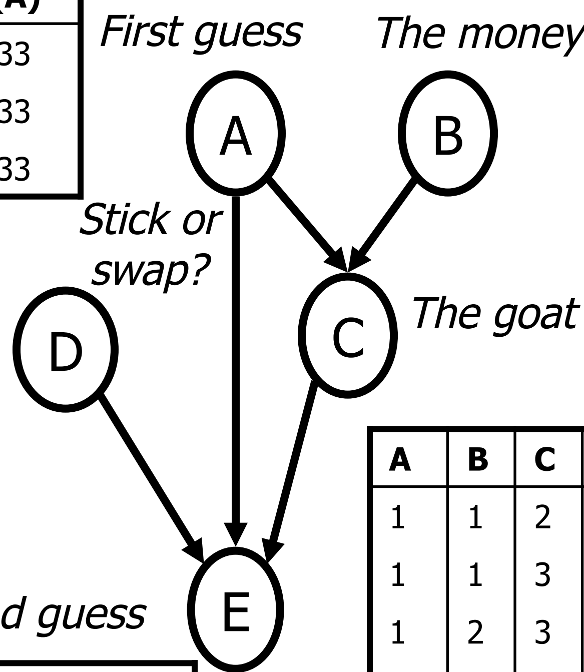
P(D) *

P(C | A,B) *

P(B) *

P(A)

A	P(A)
1	0.33
2	0.33
3	0.33



B	P(B)
1	0.33
2	0.33
3	0.33

A	B	C	P(C|A,B)
1	1	2	0.5
1	1	3	0.5
1	2	3	1.0
1	3	2	1.0
…	…	…	…

A	C	D	P(E|A,C,D)
…	…	…	…

Extra slides from last semester

Slide from William Cohen

The (highly practical) Monty Hall problem

We could construct the joint and compute P(E=B|D=swap)

...again by the chain rule:

P(A,B,C,D,E) =

P(E | A,~~B~~,C,D) *

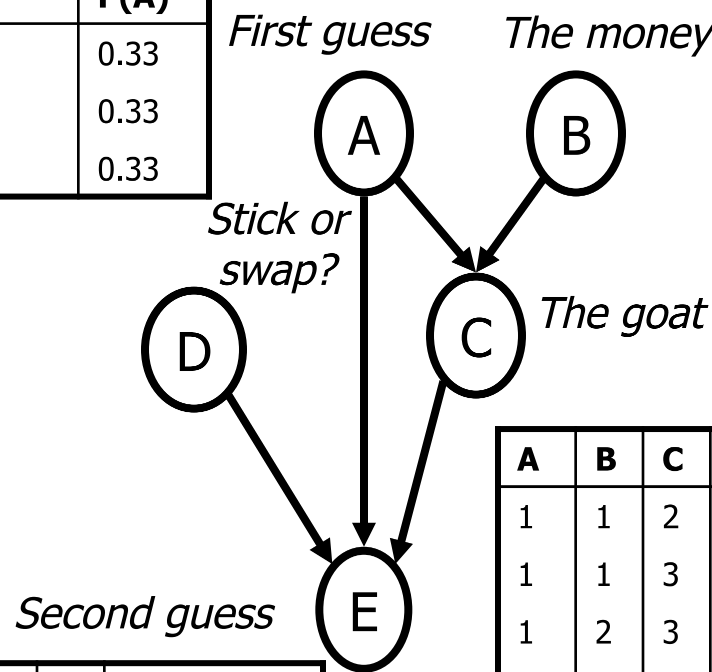
P(D | ~~A,B,C~~) *

P(C | A,B) *

P(B | ~~A~~) *

P(A)

A	P(A)
1	0.33
2	0.33
3	0.33



B	P(B)
1	0.33
2	0.33
3	0.33

A	B	C	P(C|A,B)
1	1	2	0.5
1	1	3	0.5
1	2	3	1.0
1	3	2	1.0
…	…	…	…

A	C	D	P(E|A,C,D)
…	…	…	…

Extra slides from last semester

Slide from William Cohen

The (highly practical) Monty Hall problem

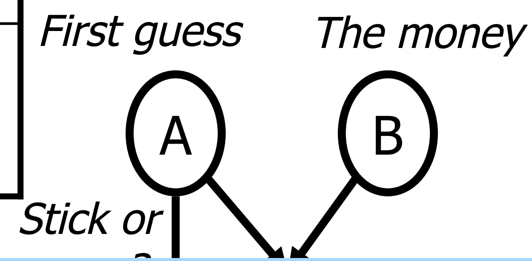
The joint table has…?

3*3*3*2*3 = 162 rows

The *conditional probability tables* (CPTs) shown have … ?

3 + 3 + 3*3*3 + 2*3*3
= 51 rows < 162 rows

A	P(A)
1	0.33
2	0.33
3	0.33



B	P(B)
1	0.33
2	0.33
3	0.33

Seco

A	C	D
…	…	…

Big questions:

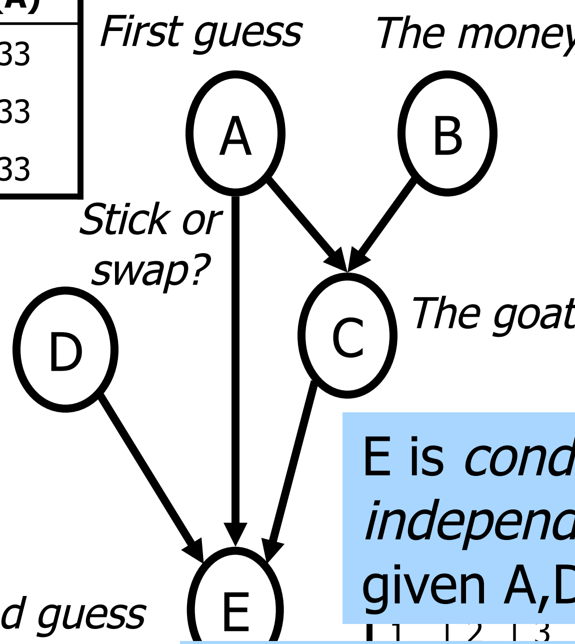
- *why* are the CPTs smaller?
- how *much smaller* are the CPTs than the joint?
- can we compute the answers to queries like P(E=B|d) *without* building the joint probability tables, just using the CPTs?

Extra slides from last semester

The (highly practical) Monty Hall problem

Why is the CPTs representation smaller? Follow the money! (B)

A	P(A)
1	0.33
2	0.33
3	0.33



B	P(B)
1	0.33
2	0.33
3	0.33

$$P(E=e \mid A,C,D) = \begin{cases} 1.0 & \text{if } (e=a) \wedge (d=stick) \\ 1.0 & \text{if } (e \notin \{a,c\}) \wedge (d=swap) \\ 0 & \text{otherwise} \end{cases}$$

E is *conditionally independent* of B given A,D,C

$$\forall a,b,c,d,e$$
$$P(E=e \mid A=a, C=c, D=d)$$
$$= P(E=e \mid A=a, B=b, C=b, D=d)$$

P(E|A,C,D)				
	1	2	3	1.0
…	…	…	…	…

$$E \perp B \mid A,C,D$$

$$I < E,\{A,C,D\},B >$$

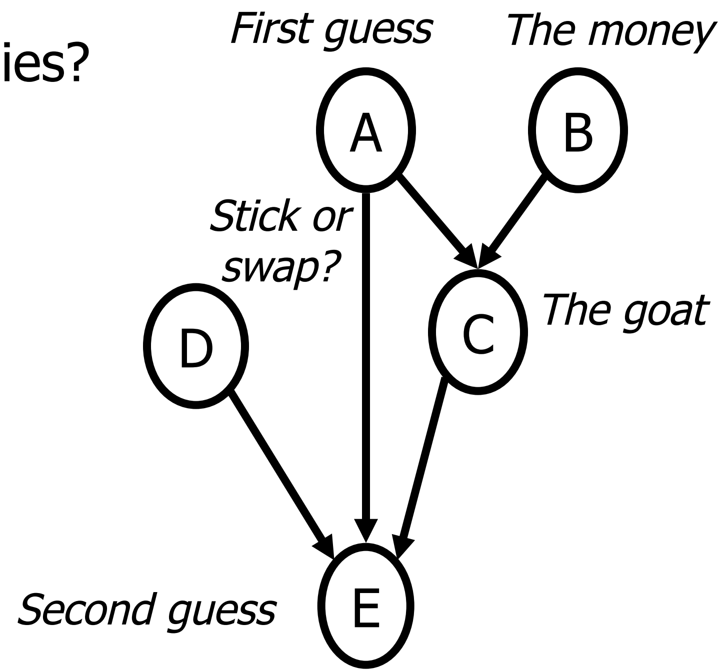
Extra slides from last semester

Slide from William Cohen

The (highly practical) Monty Hall problem

What are the conditional indepencies?

- I<A, {B}, C> ?
- I<A, {C}, B> ?
- I<E, {A,C}, B> ?
- I<D, {E}, B> ?
- …



Extra slides from last semester

Slide from William Cohen

GRAPHICAL MODELS: DETERMINING CONDITIONAL INDEPENDENCIES

Slide from William Cohen

What Independencies does a Bayes Net Model?

- In order for a Bayesian network to model a probability distribution, the following must be true:

  Each variable is conditionally independent of all its non-descendants in the graph given the value of all its parents.

- This follows from

$$P(X_1 \ldots X_n) = \prod_{i=1}^{n} P(X_i \mid parents(X_i))$$

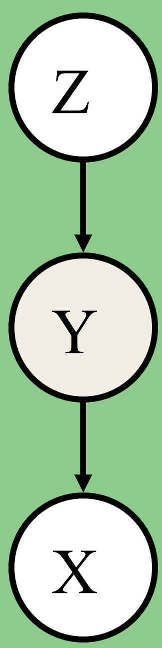
$$= \prod_{i=1}^{n} P(X_i \mid X_1 \ldots X_{i-1})$$

- But what else does it imply?

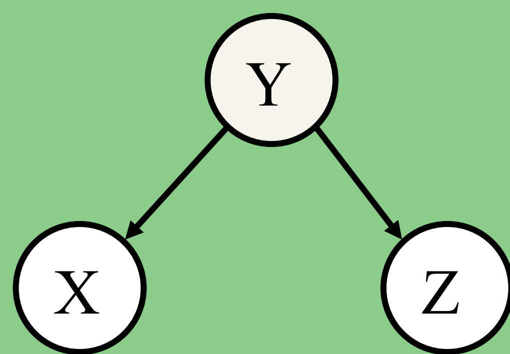
What Independencies does a Bayes Net Model?

Three cases of interest…

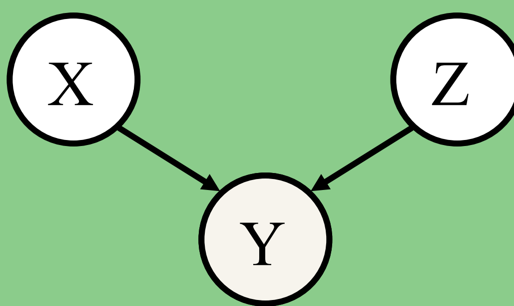
Cascade



Common Parent



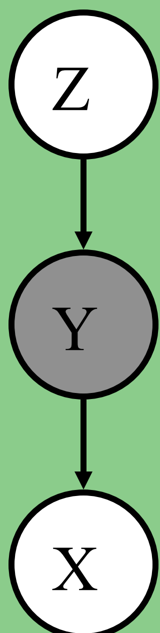
V-Structure



What Independencies does a Bayes Net Model?

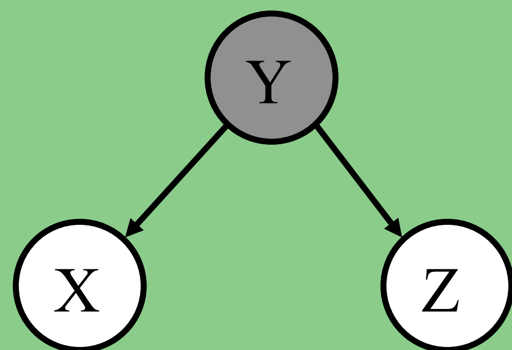
Three cases of interest…

Cascade



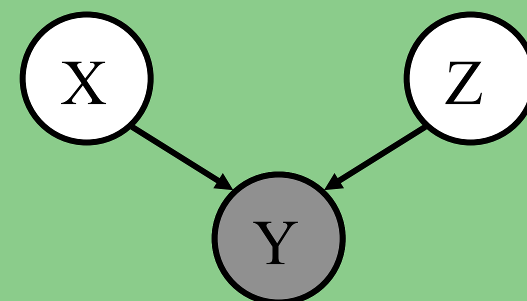
$$X \perp\!\!\!\perp Z \mid Y$$

Common Parent



$$X \perp\!\!\!\perp Z \mid Y$$

V-Structure



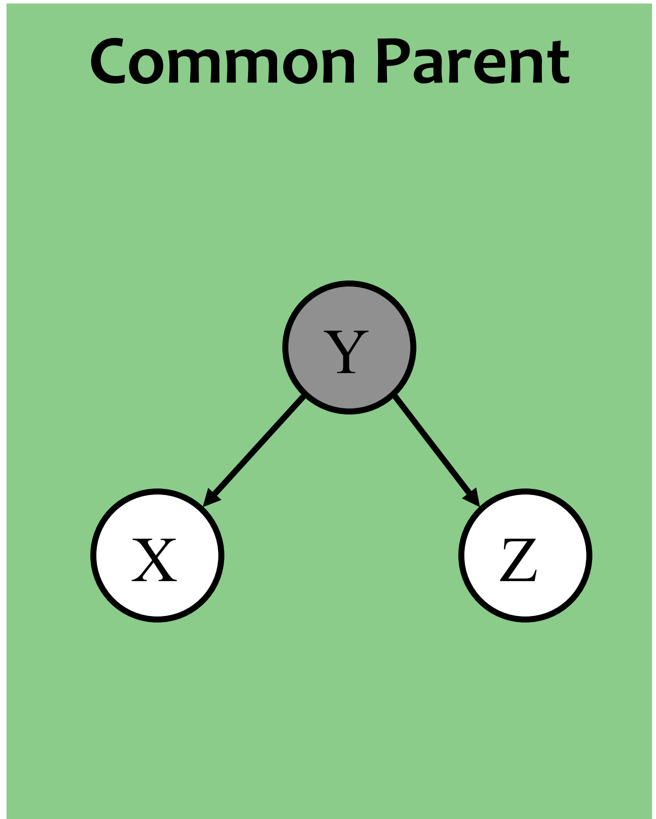
$$X \not\perp\!\!\!\perp Z \mid Y$$

Knowing Y
decouples X and Z

Knowing Y
couples X and Z

Whiteboard

Proof of conditional independence

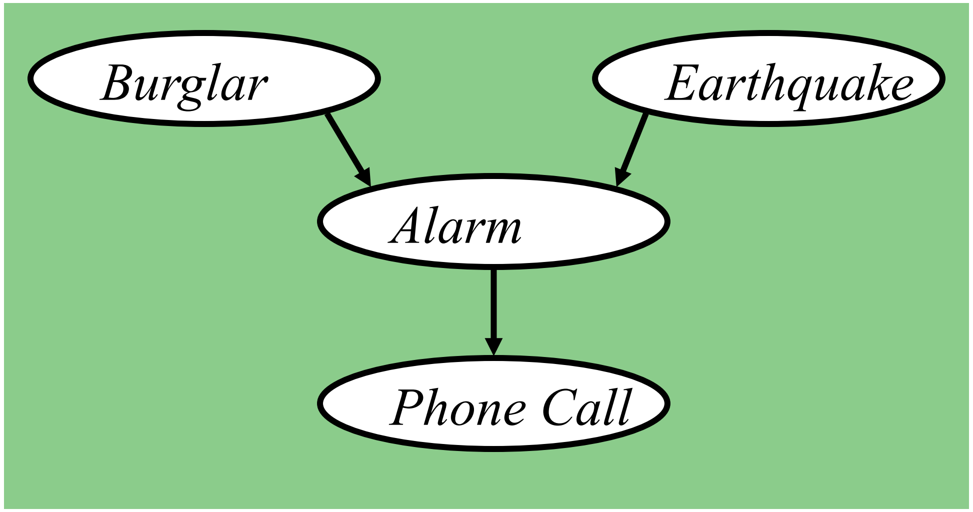


$$X \perp\!\!\!\perp Z \mid Y$$

(The other two cases can easily be shown just as easily.)

The "Burglar Alarm" example

- Your house has a twitchy burglar alarm that is also sometimes triggered by earthquakes.
- Earth arguably doesn't care whether your house is currently being burgled
- While you are on vacation, one of your neighbors calls and tells you your home's burglar alarm is ringing. Uh oh!



Quiz: True or False?

$$Burglar \perp\!\!\!\perp Earthquake \mid PhoneCall$$

D-Separation (Definition #1)

- Fortunately, there is a relatively simple algorithm for determining whether two variables in a Bayesian network are conditionally independent: *d-separation*.

- **Definition**: variables *X* and *Z* are *d-separated* (conditionally independent) given a set of evidence variables *E* iff every undirected path from *X* to *Z* is "blocked", where a path is "blocked" iff one or more of the following conditions is true: ...

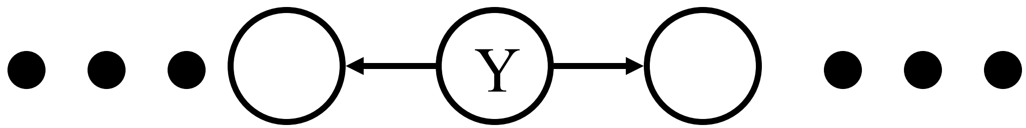
ie. X and Z are dependent iff there exists an unblocked path

Slide from William Cohen

D-Separation (Definition #1)

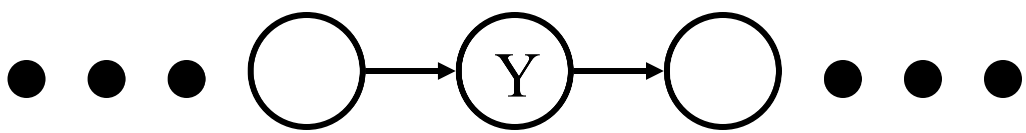
A path is "blocked" when...

- There exists a variable Y on the path such that
  - it **is** in the evidence set *E*
  - the arcs putting *Y* in the path are "tail-to-tail"



unknown "common causes" of X and Z impose dependency

- Or, there exists a variable *Y* on the path such that
  - it **is** in the evidence set *E*
  - the arcs putting *Y* in the path are "tail-to-head"



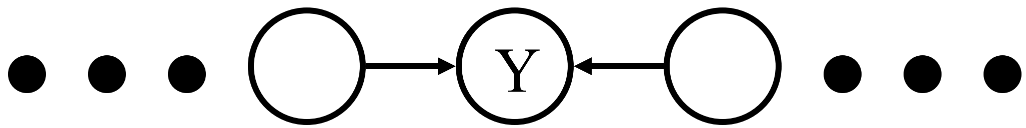
unknown "causal chains" connecting X an Z impose dependency

- Or, ...

D-Separation (Definition #1)

A path is "blocked" when...

- … Or, there exists a variable *V* on the path such that
  - it **is NOT** in the evidence set *E*
  - **neither are any of its descendants**
  - the arcs putting *Y* on the path are "head-to-head"



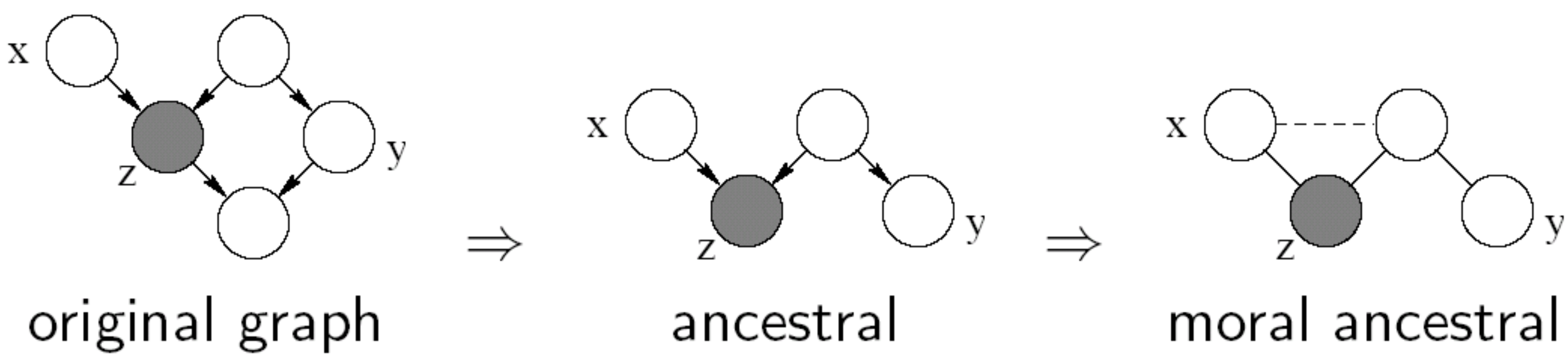
<u>Known</u> "common symptoms" of X and Z impose dependencies… X may "explain away" Z

D-Separation (Definition #2)

- D-separation criterion for Bayesian networks (D for Directed edges):

  Definition: variables X and Y are *D-separated* (conditionally independent) given Z if they are separated in the *moralized* *ancestral* graph

- Example:



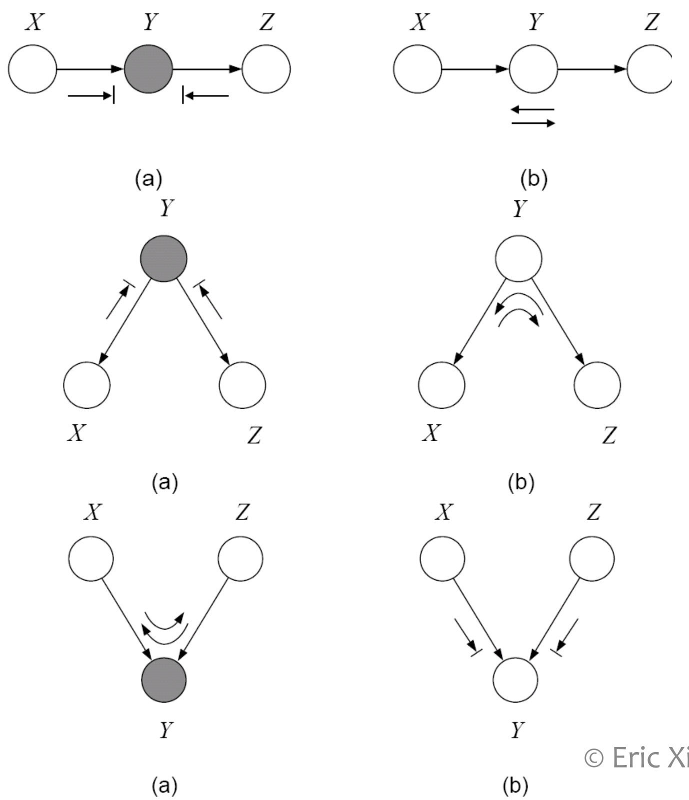
D-Separation

- Theorem [Verma & Pearl, 1998]:
  – If a set of evidence variables *E* d-separates *X* and *Z* in a Bayesian network's graph, then *I*<*X*, *E*, *Z*>.
- *d*-separation can be computed in linear time using a depth-first-search-like algorithm.
- Be careful: d-separation finds what *must* be conditionally independent
  – "Might": Variables may actually be independent when they're not d-separated, depending on the actual probabilities involved

Slide from William Cohen

"Bayes-ball" and D-Separation

- X is **d-separated** (directed-separated) from Z given Y if we can't send a ball from any node in X to any node in Z using the "*Bayes-ball*" algorithm illustrated bellow (and plus some boundary conditions):



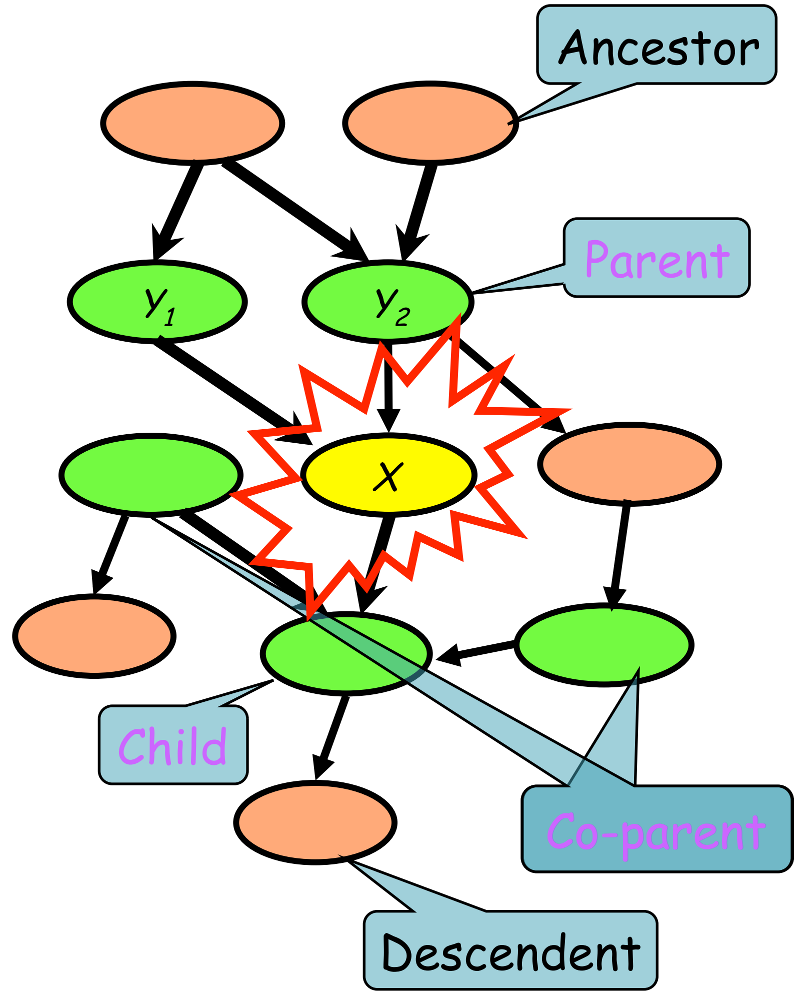
- Defn: $\mathcal{I}(\mathcal{G})$=all independence properties that correspond to d-separation:

$$\mathrm{I}(G) = \left\{ X \perp Z \middle| Y : \mathrm{dsep}_G(X; Z|Y) \right\}$$

- D-separation is sound and complete

Markov Blanket

A node is conditionally independent of every other node in the network outside its Markov blanket



Summary: Bayesian Networks

Structure: *DAG*

- Meaning: a node is conditionally independent of every other node in the network outside its Markov blanket
- Local conditional distributions (CPD) and the DAG completely determine the joint dist.
- Give causality relationships, and facilitate a generative process