10-601 Introduction to Machine Learning

Machine Learning Department
School of Computer Science
Carnegie Mellon University

Bayesian Networks (Part II)

Graphical Model Readings:
Murphy 10 – 10.2.1
Bishop 8.1, 8.2.2
HTF --
Mitchell 6.11

HMM Readings:
Murphy 10.2.2 – 10.2.3
Bishop 13.1 – 13.2
HTF --
Mitchell –

Matt Gormley
Lecture 23
April 12, 2017

Reminders

- Peer Tutoring
- Homework 7: Deep Learning
  - Release: Wed, Apr. 05
  - Part I due Wed, Apr. 12
  - Part II due Mon, Apr. 17

Start Early

BAYESIAN NETWORKS

Bayes Nets Outline

- **Motivation**
  - Structured Prediction
- **Background**
  - Conditional Independence
  - Chain Rule of Probability
- **Directed Graphical Models**
  - Writing Joint Distributions
  - Definition: Bayesian Network
  - Qualitative Specification
  - Quantitative Specification
  - Familiar Models as Bayes Nets
- **Conditional Independence in Bayes Nets**
  - Three case studies
  - D-separation
  - Markov blanket
- **Learning**
  - Fully Observed Bayes Net
  - (Partially Observed Bayes Net)
- **Inference**
  - Background: Marginal Probability
  - Sampling directly from the joint distribution
  - Gibbs Sampling

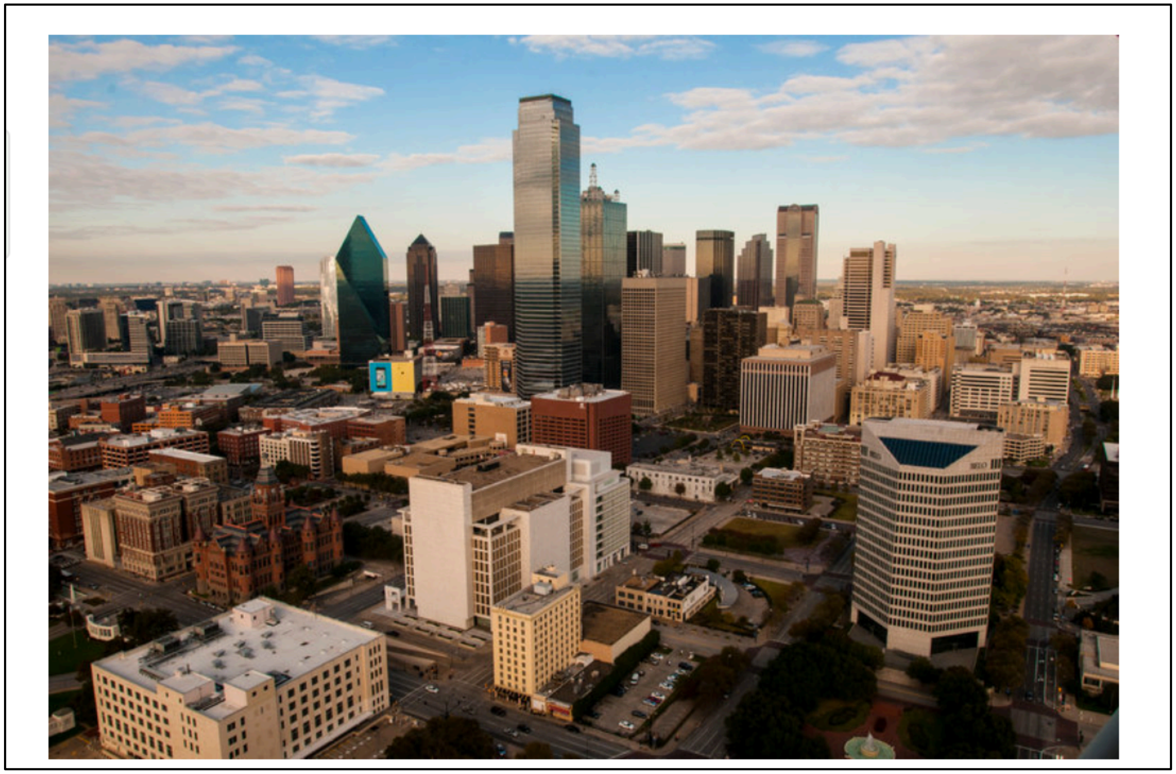
Last Lecture

This Lecture

Bayesian Networks

DIRECTED GRAPHICAL MODELS

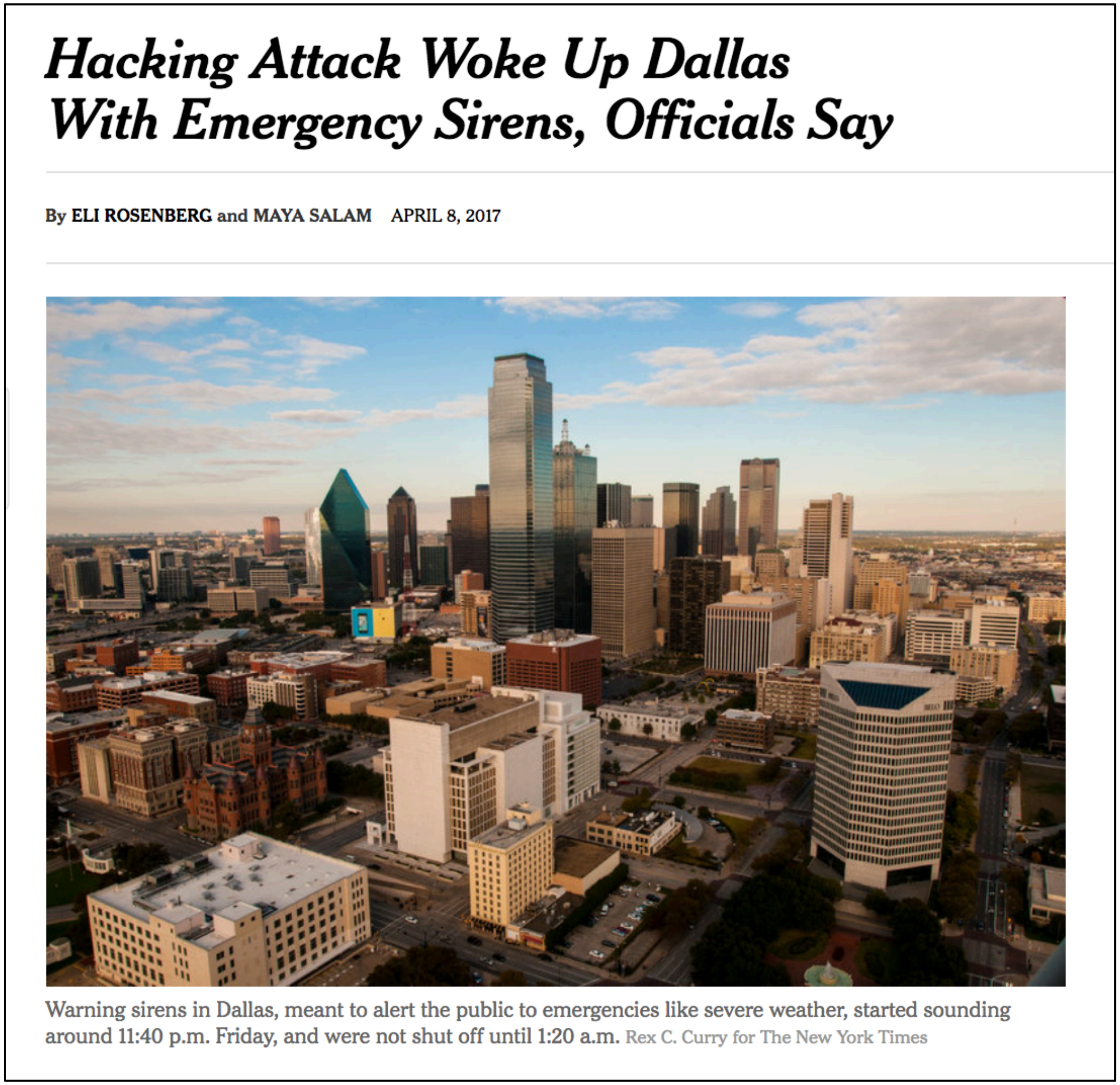
Example: Tornado Alarms



1. Imagine that you work at the 911 call center in Dallas
2. You receive six calls informing you that the Emergency Weather Sirens are going off
3. What do you conclude?

Figure from https://www.nytimes.com/2017/04/08/us/dallas-emergency-sirens-hacking.html

Example: Tornado Alarms



Hacking Attack Woke Up Dallas With Emergency Sirens, Officials Say

By ELI ROSENBERG and MAYA SALAM APRIL 8, 2017

Warning sirens in Dallas, meant to alert the public to emergencies like severe weather, started sounding around 11:40 p.m. Friday, and were not shut off until 1:20 a.m. Rex C. Curry for The New York Times

1. Imagine that you work at the 911 call center in Dallas
2. You receive six calls informing you that the Emergency Weather Sirens are going off
3. What do you conclude?

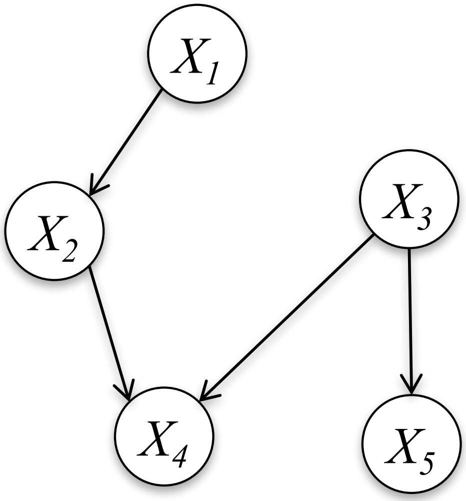
Figure from https://www.nytimes.com/2017/04/08/us/dallas-emergency-sirens-hacking.html

Directed Graphical Models (Bayes Nets)

Whiteboard

- – Example: Tornado Alarms
- – Writing Joint Distributions
    - Idea #1: Giant Table
    - Idea #2: Rewrite using chain rule
    - Idea #3: Assume full independence
    - Idea #4: Drop variables from RHS of conditionals
- – Definition: Bayesian Network

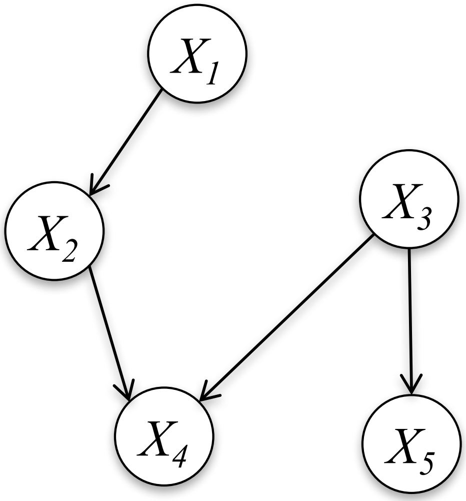
Bayesian Network



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

Bayesian Network

Definition:



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

- 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 (i.e. structure learning)
  - We simply link a certain architecture (e.g. a layered graph)
  - …

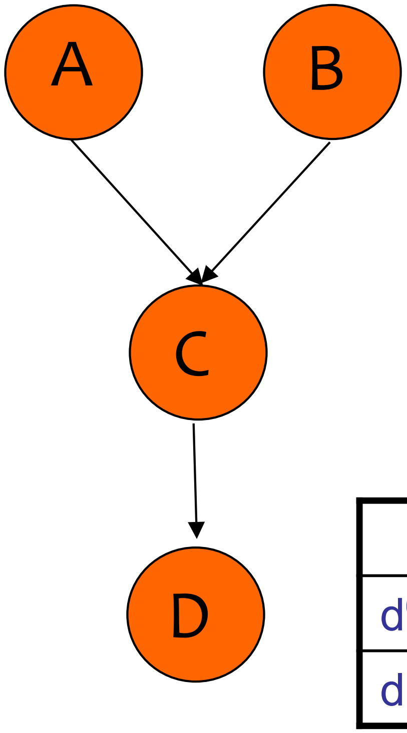
Quantitative Specification

Example: Conditional probability tables (CPTs)
for discrete random variables

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

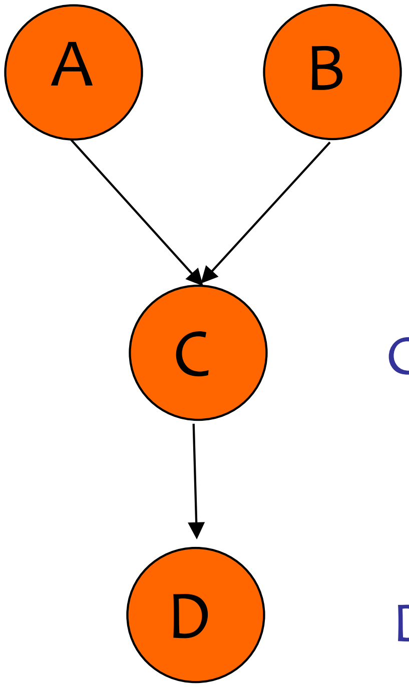
	c^0	c^1
d^0	0.3	0.5
d^1	07	0.5

Quantitative Specification

Example: Conditional probability density functions (CPDs) for continuous random variables

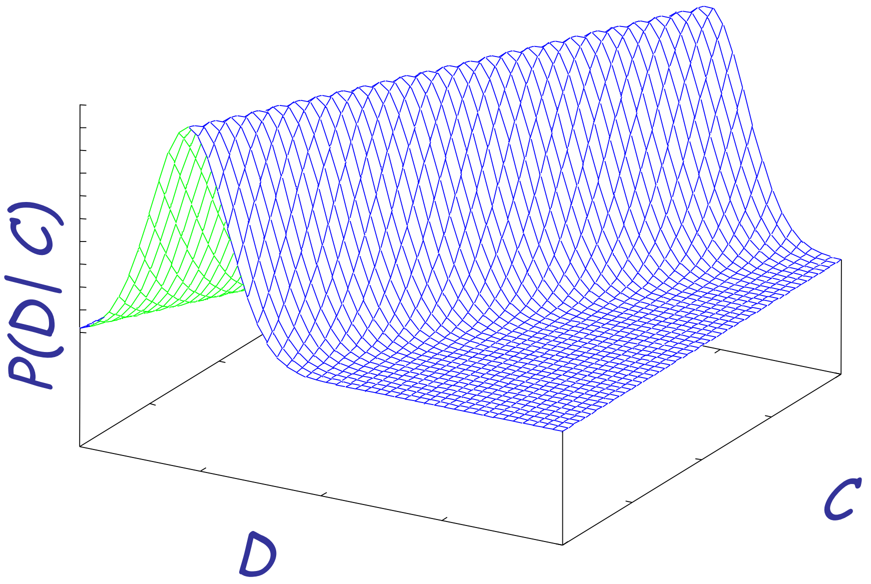
$A \sim N(\mu_a, \Sigma_a)$ $B \sim N(\mu_b, \Sigma_b)$

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



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

$D \sim N(\mu_d + C, \Sigma_d)$



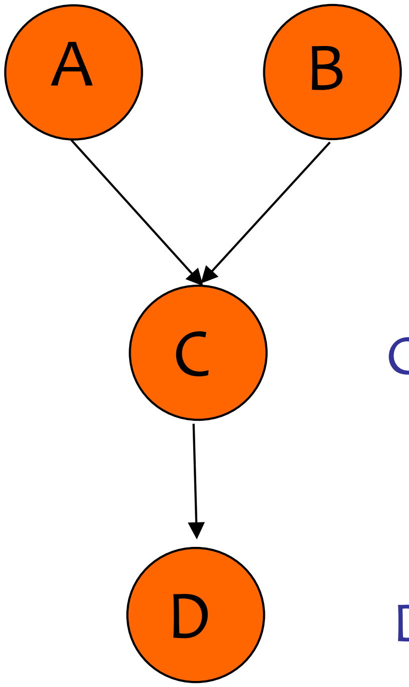
Quantitative Specification

Example: Combination of CPTs and CPDs for a mix of discrete and continuous variables

a^0	0.75
a^1	0.25

b^0	0.33
b^1	0.67

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



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

$D \sim N(\mu_d + C, \Sigma_d)$

Directed Graphical Models (Bayes Nets)

Whiteboard

- Observed Variables in Graphical Model
- Familiar Models as Bayes Nets
  - Bernoulli Naïve Bayes
  - Gaussian Naïve Bayes
  - Gaussian Mixture Model (GMM)
  - Gaussian Discriminant Analysis
  - Logistic Regression
  - Linear Regression
  - 1D Gaussian

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

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

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

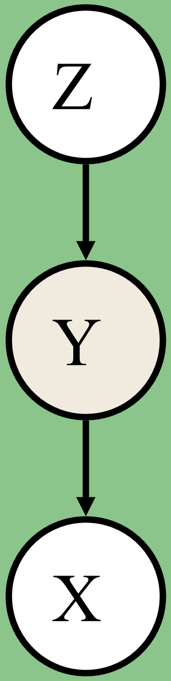
- But what else does it imply?

Slide from William Cohen

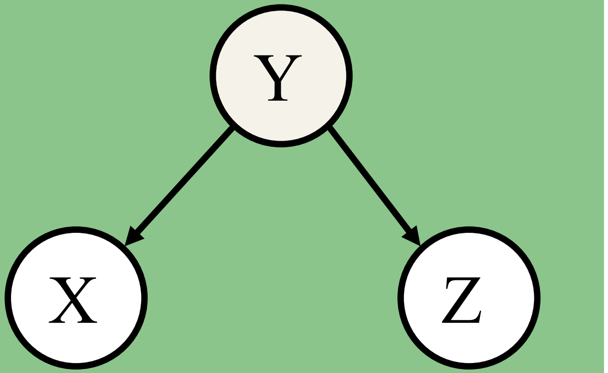
What Independencies does a Bayes Net Model?

Three cases of interest…

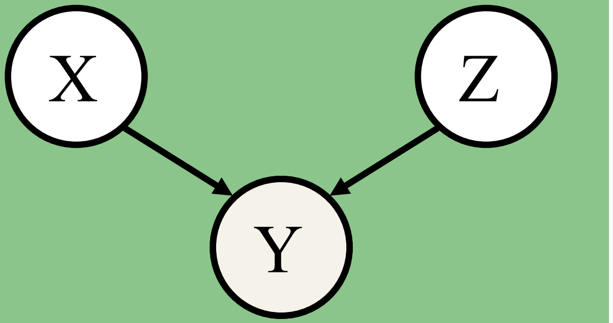
Cascade



Common Parent



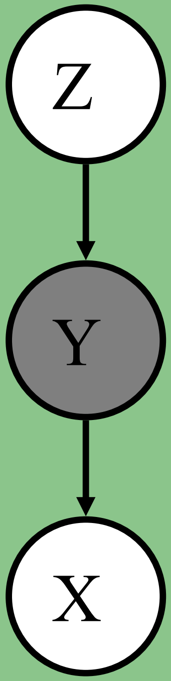
V-Structure



What Independencies does a Bayes Net Model?

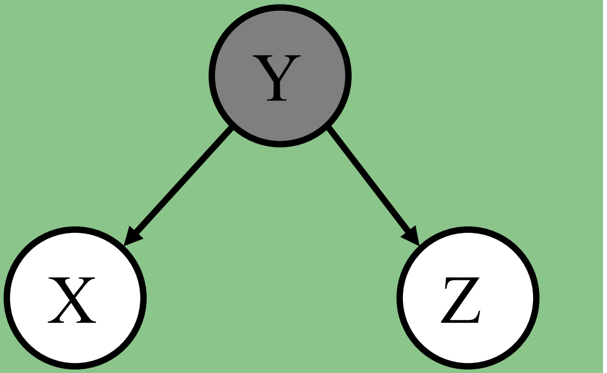
Three cases of interest…

Cascade



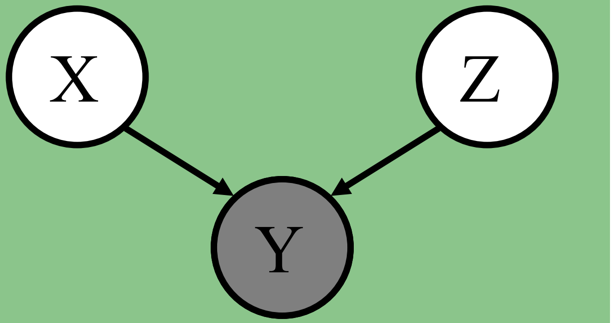
$$X \perp\!\!\!\perp Z \mid Y$$

Common Parent



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

V-Structure



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

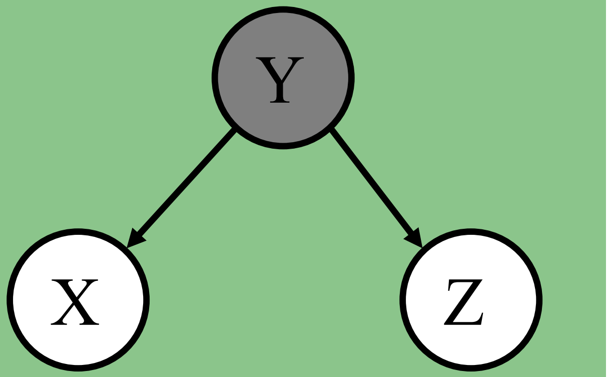
Knowing Y
decouples X and Z

Knowing Y
couples X and Z

Whiteboard

Proof of conditional independence

Common Parent

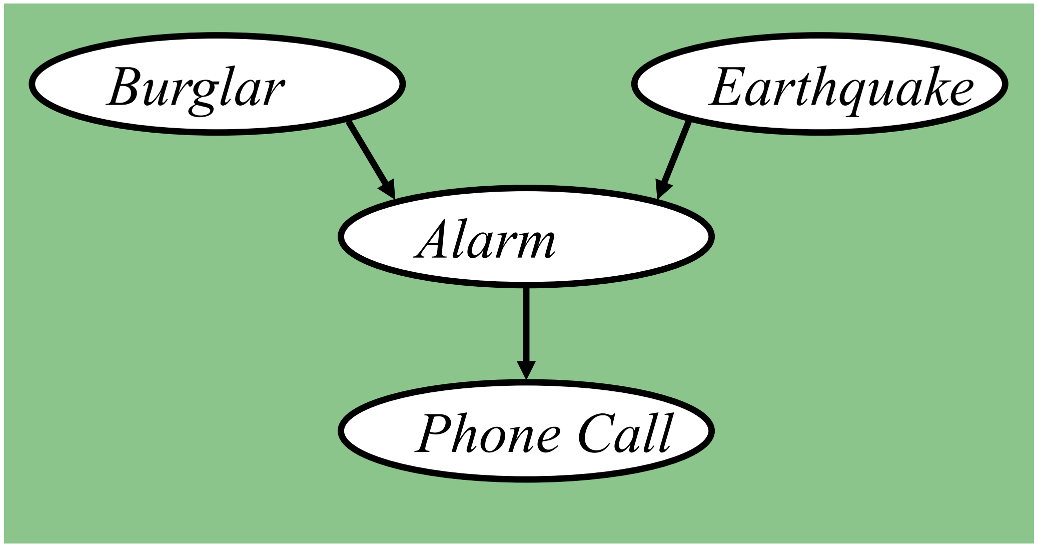


$$X \perp\!\!\!\perp Z \mid Y$$

(The other two cases can 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!

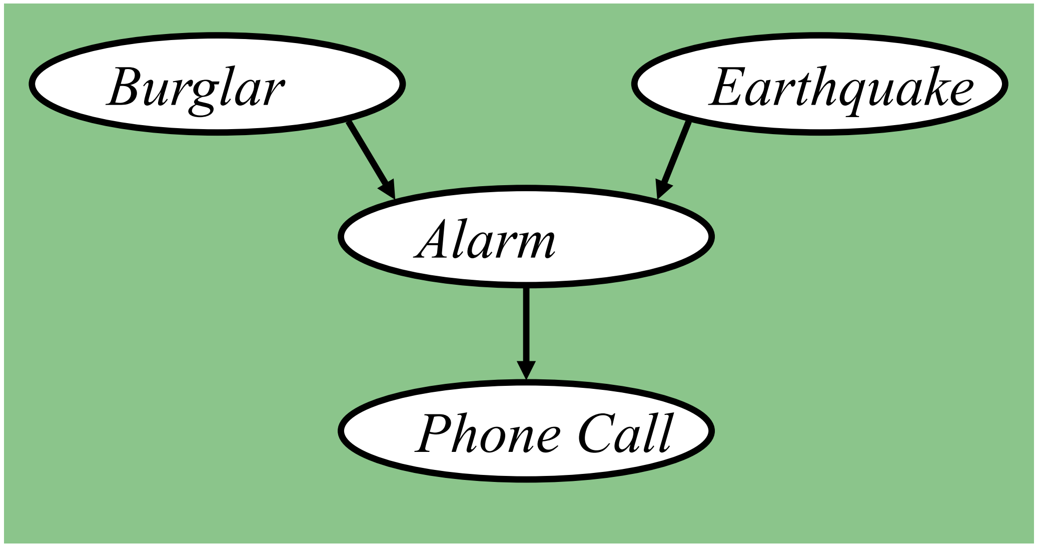


Quiz: True or False?

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

Slide from William Cohen

The “Burglar Alarm” example



- But now suppose you learn that there was a medium-sized earthquake in your neighborhood. Oh, whew! Probably not a burglar after all.
- Earthquake “explains away” the hypothetical burglar.
- But then it must **not** be the case that

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

even though

$$Burglar \perp\!\!\!\perp Earthquake$$

Slide from William Cohen

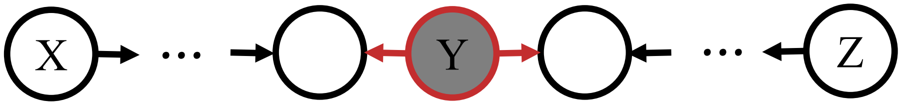
D-Separation

If variables X and Z are **d-separated** given a **set** of variables E
Then X and Z are **conditionally independent** given the **set** E

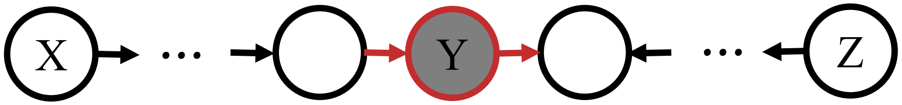
<u>**Definition #1:**</u>
Variables X and Z are **d-separated** given a **set** of evidence variables E iff every path from X to Z is "blocked".

A path is "blocked" whenever:

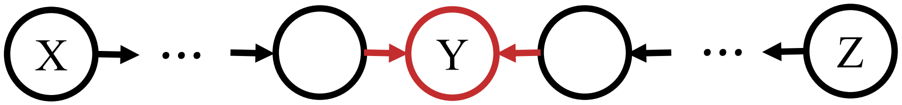
1. ∃ Y on path s.t. Y ∈ E and Y is a "common parent"



2. ∃ Y on path s.t. Y ∈ E and Y is in a "cascade"



3. ∃ Y on path s.t. {Y, descendants(Y)} ∉ E and Y is in a "v-structure"



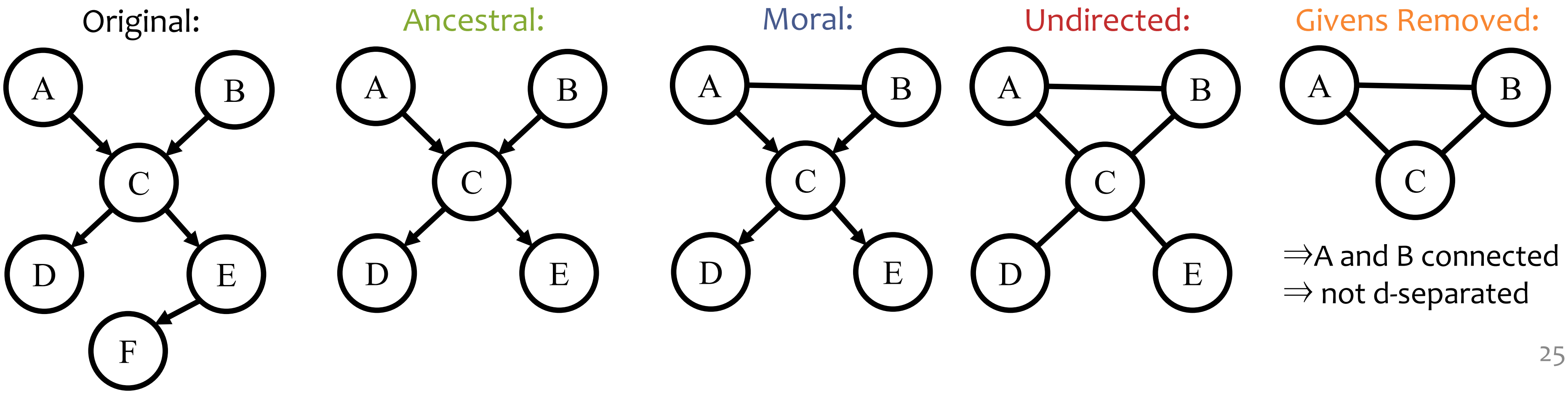
D-Separation

If variables X and Z are **d-separated** given a **set** of variables E
Then X and Z are **conditionally independent** given the **set** E

Definition #2:
Variables X and Z are **d-separated** given a **set** of evidence variables E iff there does **not** exist a path in the **undirected** **ancestral** **moral** graph **with E removed**.

1. **Ancestral graph**: keep only X, Z, E and their ancestors
2. **Moral graph**: add undirected edge between all pairs of each node's parents
3. **Undirected graph**: convert all directed edges to undirected
4. Givens Removed: delete any nodes in E

Example Query: $A \perp\!\!\!\perp B \mid \{D, E\}$



Original:

Ancestral:

Moral:

Undirected:

Givens Removed:

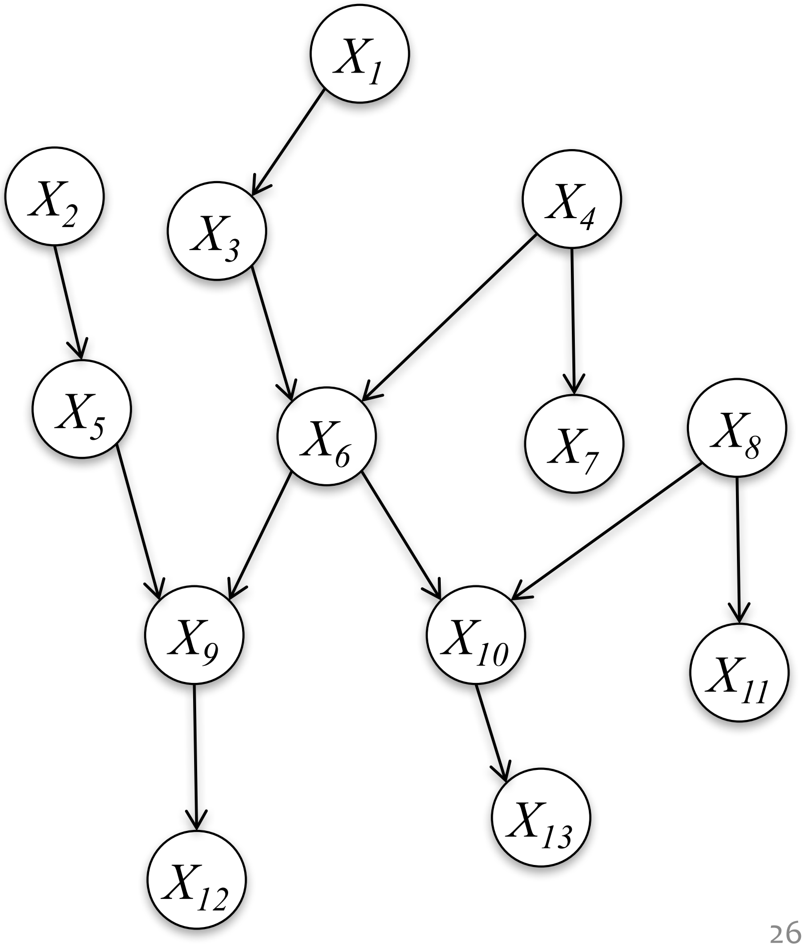
⇒A and B connected
⇒ not d-separated

Markov Blanket

Def: the **co-parents** of a node are the parents of its children

Def: the **Markov Blanket** of a node is the set containing the node's parents, children, and co-parents.

Thm: a node is **conditionally independent** of every other node in the graph given its **Markov blanket**



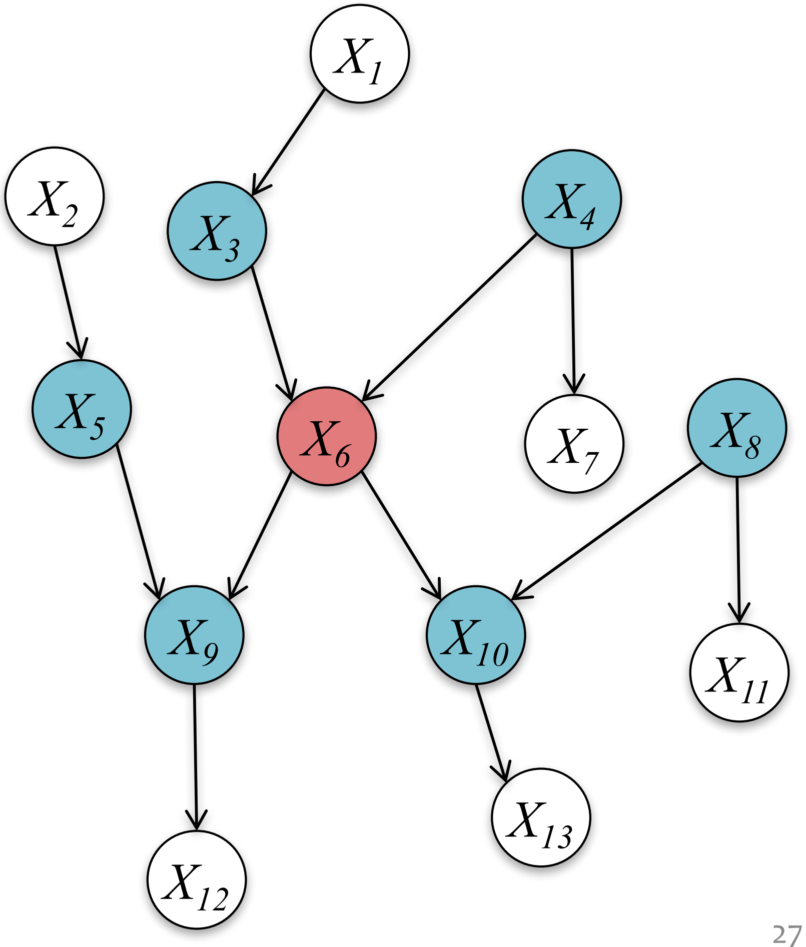
Markov Blanket

Def: the **co-parents** of a node are the parents of its children

Def: the **Markov Blanket** of a node is the set containing the node's parents, children, and co-parents.

Thm: a node is **conditionally independent** of every other node in the graph given its **Markov blanket**

Example: The Markov Blanket of X_6 is $\{X_3, X_4, X_5, X_8, X_9, X_{10}\}$



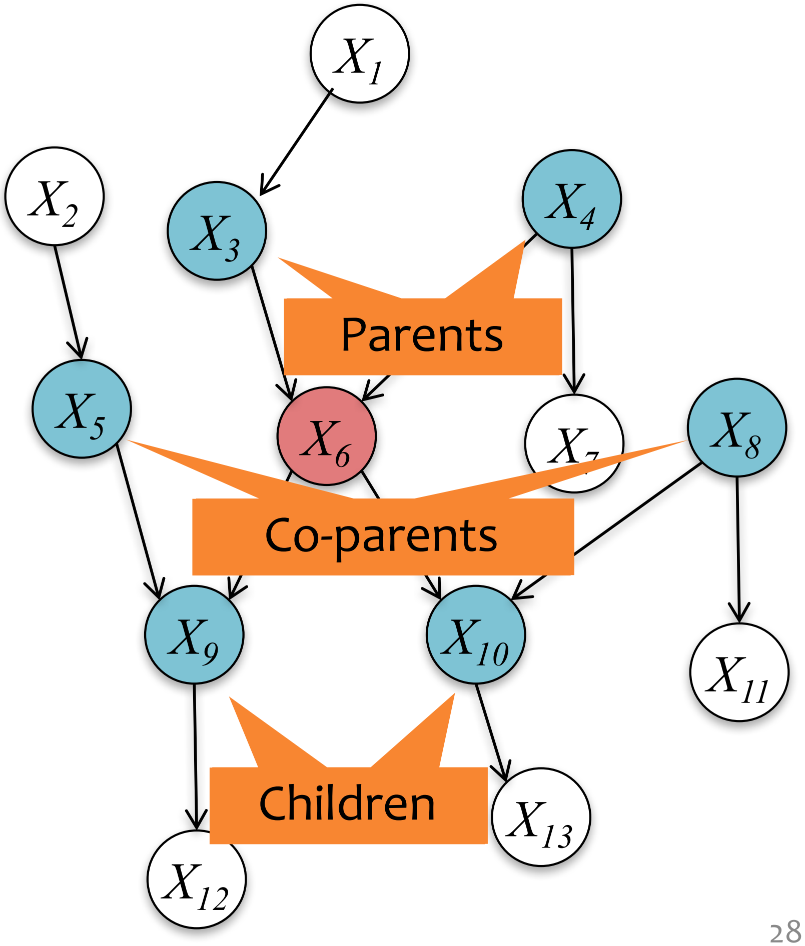
Markov Blanket

Def: the **co-parents** of a node are the parents of its children

Def: the **Markov Blanket** of a node is the set containing the node's parents, children, and co-parents.

Thm: a node is **conditionally independent** of every other node in the graph given its **Markov blanket**

Example: The Markov Blanket of X_6 is $\{X_3, X_4, X_5, X_8, X_9, X_{10}\}$



SUPERVISED LEARNING FOR BAYES NETS

Machine Learning

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

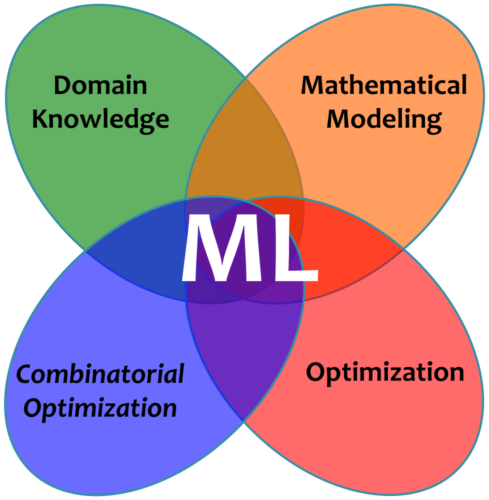
Our **model** defines a score for each structure

It also tells us what to optimize

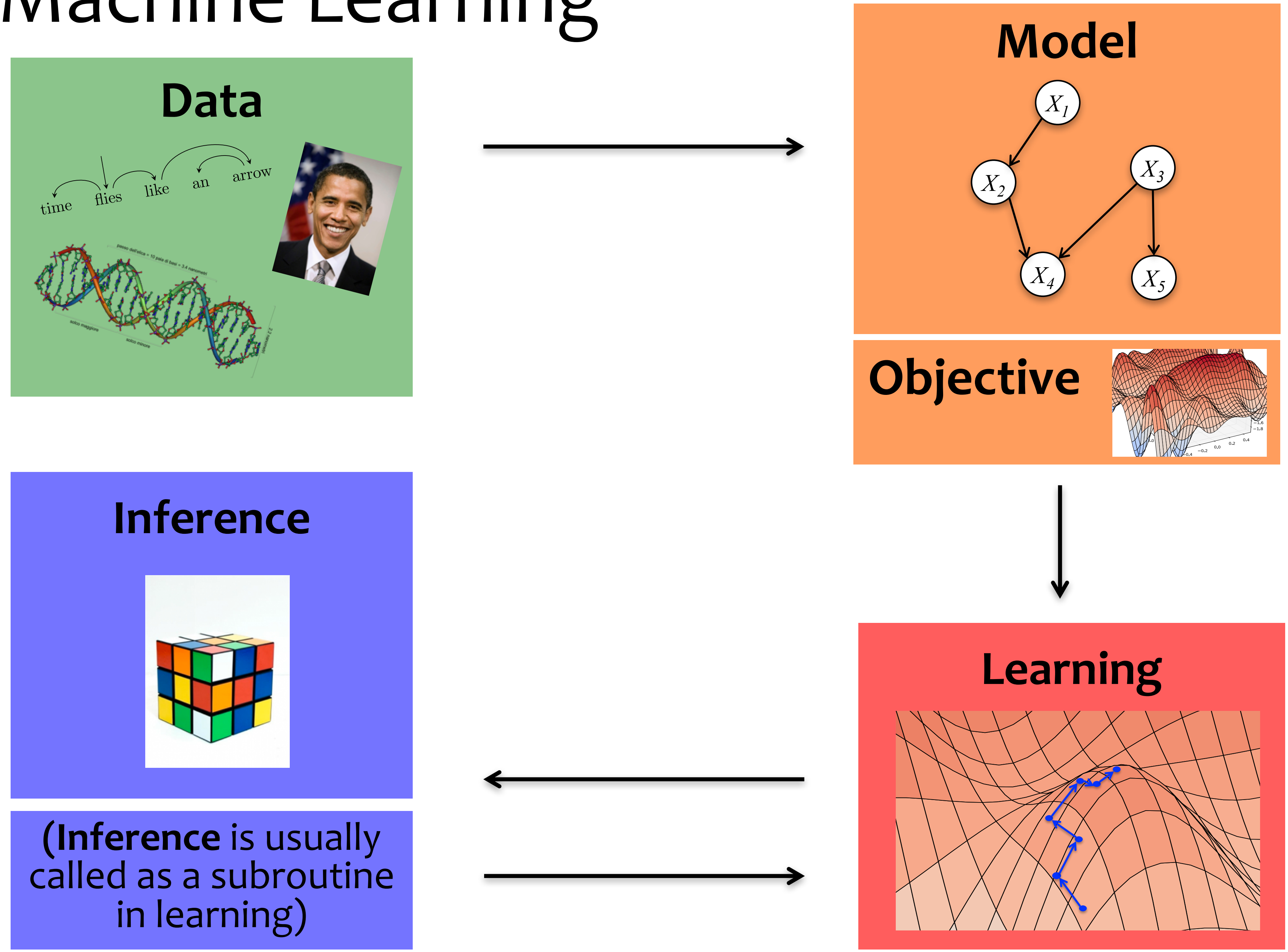
Learning tunes the parameters of the model

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

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



Machine Learning



Data

Model

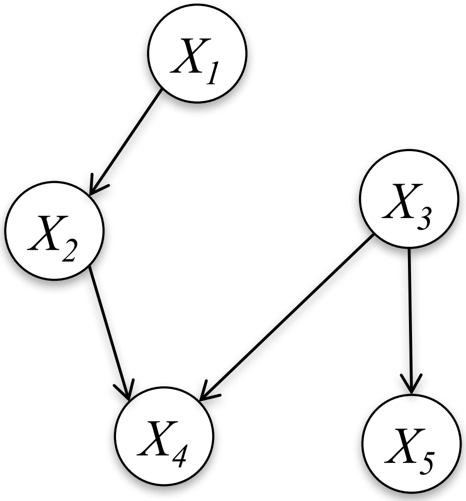
Objective

Learning

Inference

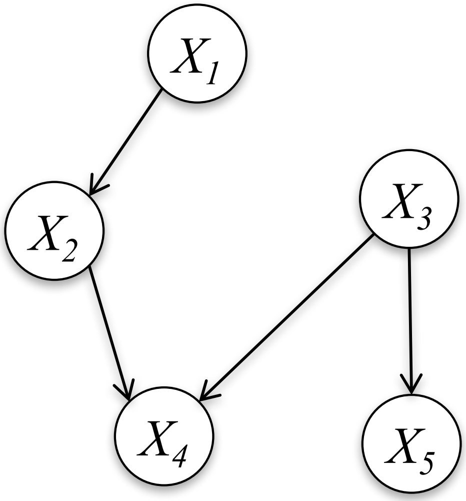
(**Inference** is usually called as a subroutine in learning)

Learning Fully Observed BNs



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

Learning Fully Observed BNs

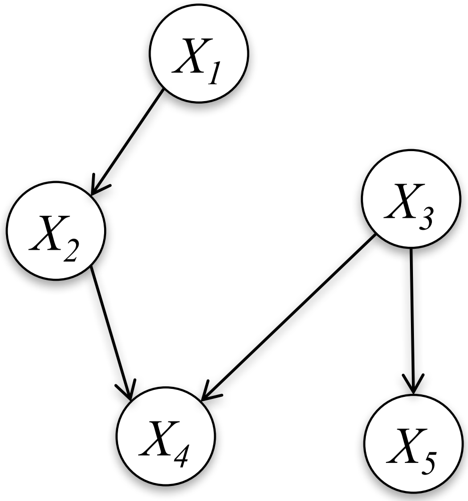


$$p(X_1, X_2, X_3, X_4, X_5) =$$

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

$$p(X_3)p(X_2|X_1)p(X_1)$$

Learning Fully Observed BNs



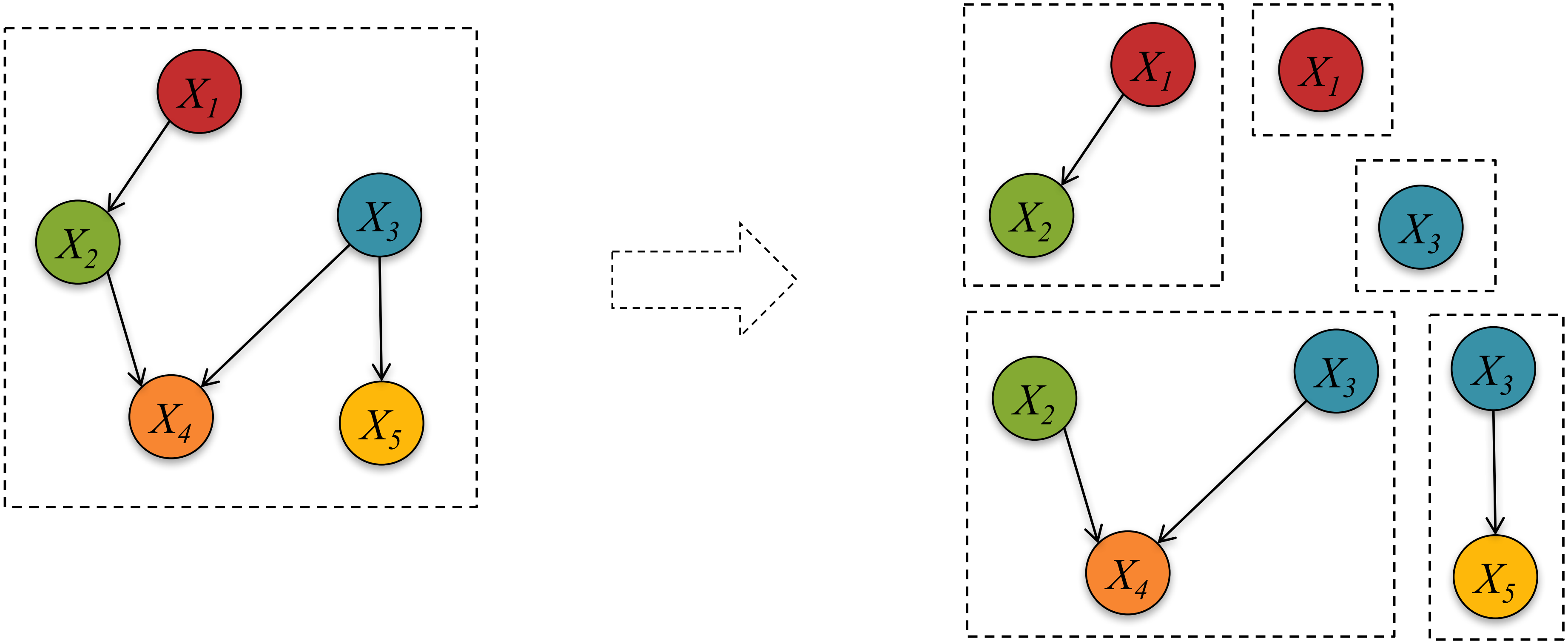
$$p(X_1, X_2, X_3, X_4, X_5) =$$
$$p(X_5|X_3)p(X_4|X_2, X_3)$$
$$p(X_3)p(X_2|X_1)p(X_1)$$

How do we learn these conditional and marginal distributions for a Bayes Net?

Learning Fully Observed BNs

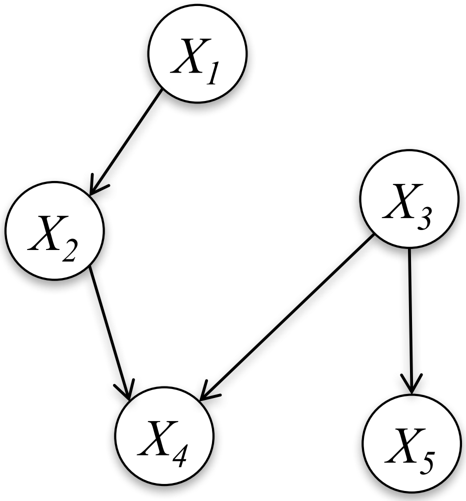
Learning this fully observed Bayesian Network is **equivalent** to learning five (small / simple) independent networks from the same data

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



Learning Fully Observed BNs

How do we **learn** these conditional and marginal distributions for a Bayes Net?



$$\begin{aligned}\boldsymbol{\theta}^* &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log p(X_1, X_2, X_3, X_4, X_5)\\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log p(X_5|X_3, \theta_5) + \log p(X_4|X_2, X_3, \theta_4)\\ &\qquad + \log p(X_3|\theta_3) + \log p(X_2|X_1, \theta_2)\\ &\qquad + \log p(X_1|\theta_1)\end{aligned}$$

$$\theta_1^* = \underset{\theta_1}{\operatorname{argmax}} \log p(X_1|\theta_1)$$

$$\theta_2^* = \underset{\theta_2}{\operatorname{argmax}} \log p(X_2|X_1, \theta_2)$$

$$\theta_3^* = \underset{\theta_3}{\operatorname{argmax}} \log p(X_3|\theta_3)$$

$$\theta_4^* = \underset{\theta_4}{\operatorname{argmax}} \log p(X_4|X_2, X_3, \theta_4)$$

$$\theta_5^* = \underset{\theta_5}{\operatorname{argmax}} \log p(X_5|X_3, \theta_5)$$

Learning Fully Observed BNs

Whiteboard

- Example: Learning for Tornado Alarms