Causal Inference:
prediction, explanation, and intervention

Lecture 4: Graphical models and Bayesian Networks

Samantha Kleinberg
samantha.kleinberg@stevens.edu

Main topics today

- Graphical models intro
  - Representation
  - Inference
  - Learning
- What makes a graphical model causal?
- How to use graphical models to answer causal questions
  - Predicting effects of actions
  - Counterfactual queries

Why graphical models?

Smoking	Chronic Bronchitis	Lung Cancer	Fatigue	Mass on X-ray	P(S,CB,LC,F,M)
F	F	F	F	F	0.2
F	F	F	F	T	0.1
	…				…
T	T	T	T	T	0.4

n binary variables = 2^n entries

Some notation

- $X_1, X_2 \ldots X_n$ are a set of random variables
  - X_1 = variable, x=particular value of X_1
  - For coin, X={heads, tails}

$$\sum_x P(x) = \sum_{x \in \{heads, tails\}} P(x)$$

- $P(X_1, X_2 \ldots X_n)$ is set of joint probability distributions over all assignments of variables
  - If each variable binary, need 2^n parameters
  - Gives set of equations

Some more notation

Binary variables X,Y

$$P(x) = P(x \mid y)P(y) + P(x \mid \neg y)P(\neg y)$$

General case

$$P(x) = \sum_{Y} P(x \mid y)P(y)$$

Chain rule

Last week: $P(A \wedge B)=P(A|B)P(B)$

Multiple variables:

$$
\begin{aligned}
P(X,Y,Z) &= P(X|Y,Z)P(Y,Z) \\
&= P(X|Y,Z)P(Y|Z)P(Z)
\end{aligned}
$$

Chain rule (general)

$$P(X_1 \ldots X_n) = P(X_1 \mid X_2 \ldots X_n) P(X_2 \mid X_3 \ldots X_n) \ldots P(X_n)$$

$$= \prod_{i=1}^{n} P(X_i \mid X_{i+1} .. X_n)$$

Key observation: independence

- If variables independent, fewer parameters needed
- When $X_1 \ldots X_n$ are all independent

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

Conditional independence

Yellowed fingers and lung cancer are dependent:

$$P(Y,L)>P(Y)P(L)$$

But they are independent conditioned on smoking (S)

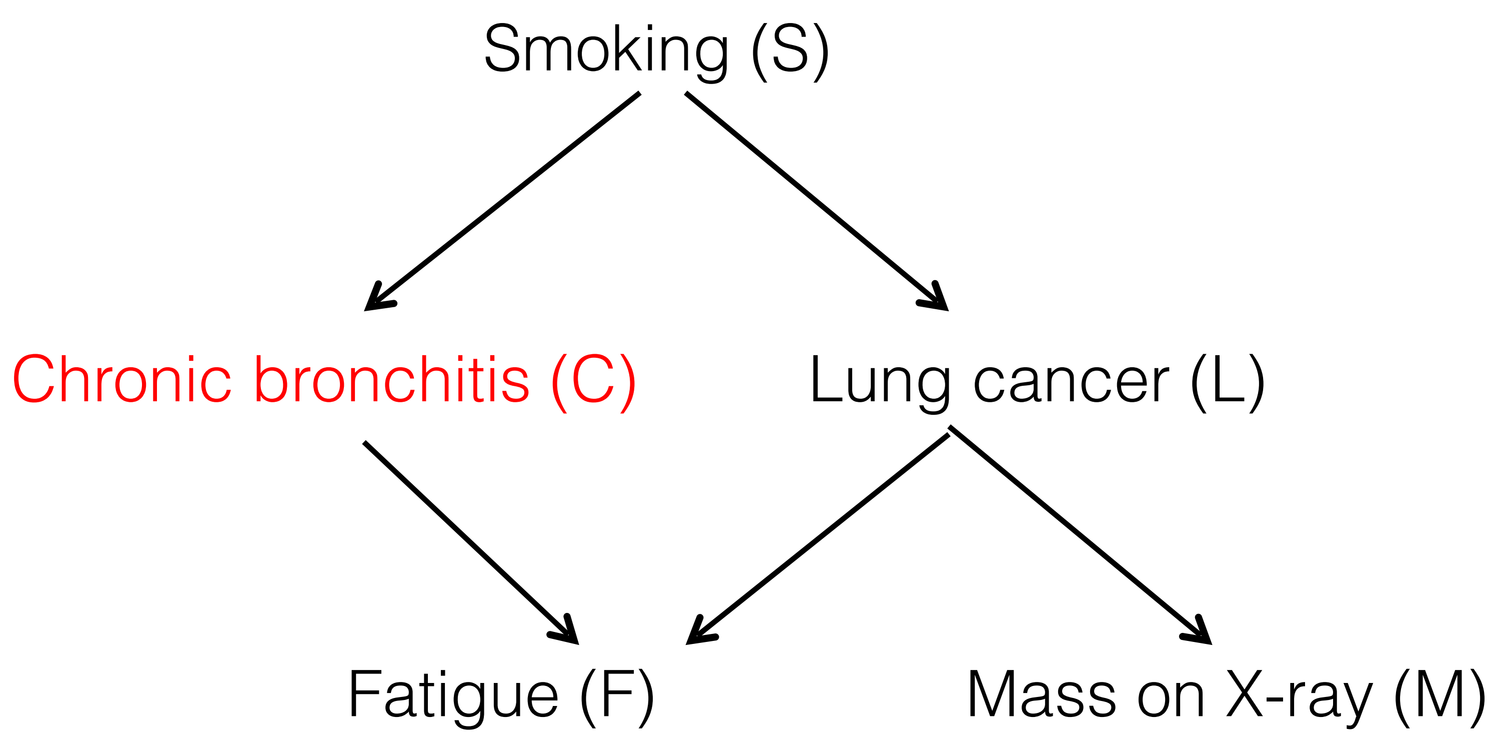
$$P(S,Y,L)=P(Y|L,S)P(L|S)P(S)$$
$$=P(Y|S)P(L|S)P(S)$$

Graphical model

- Describes independencies in a set of variables

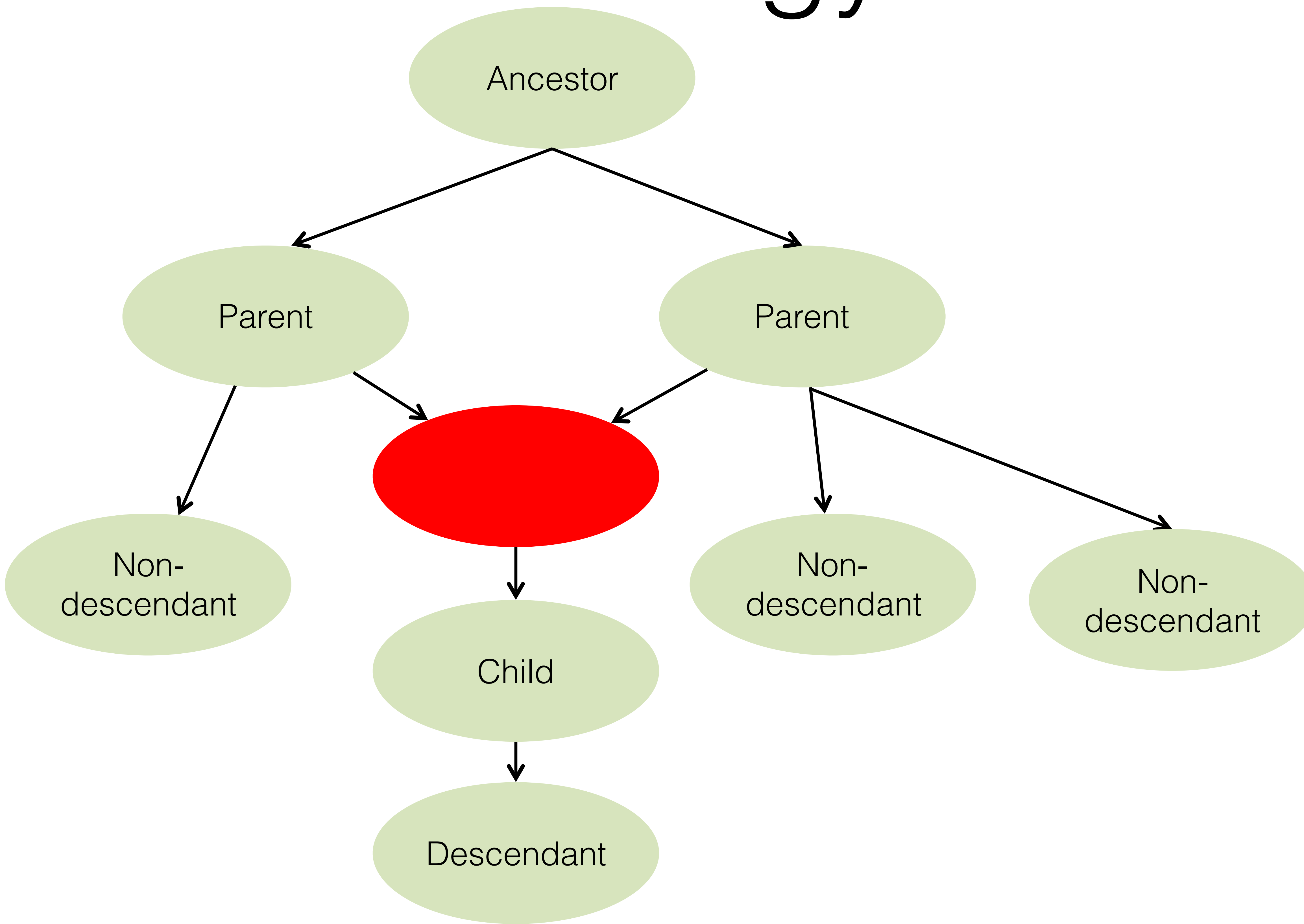
- Directed
  - Bayesian network
  - HMM (next week)
- Undirected

Graphical model and independencies



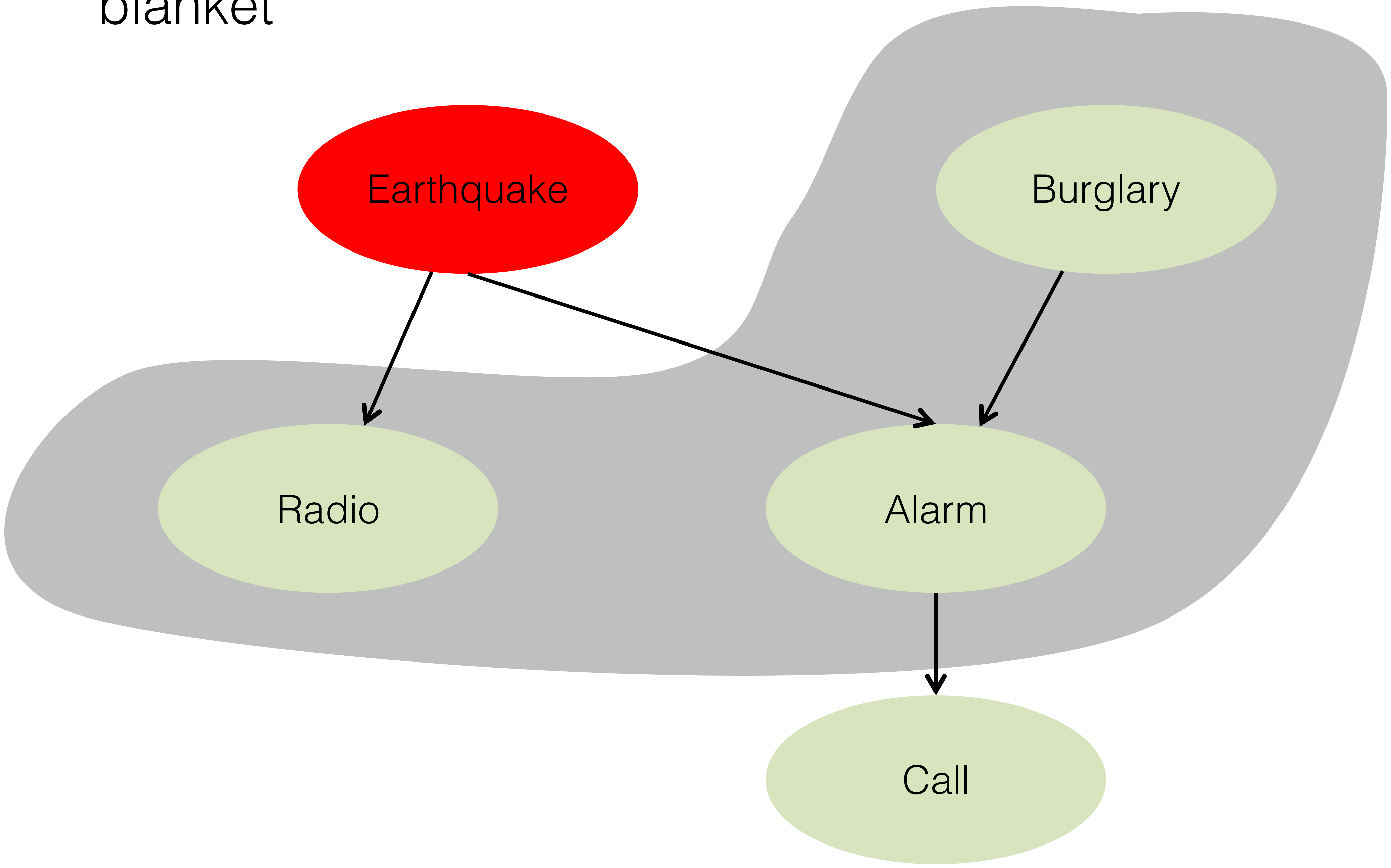
A node is independent of its non-descendants given its parents

Terminology



Markov blanket

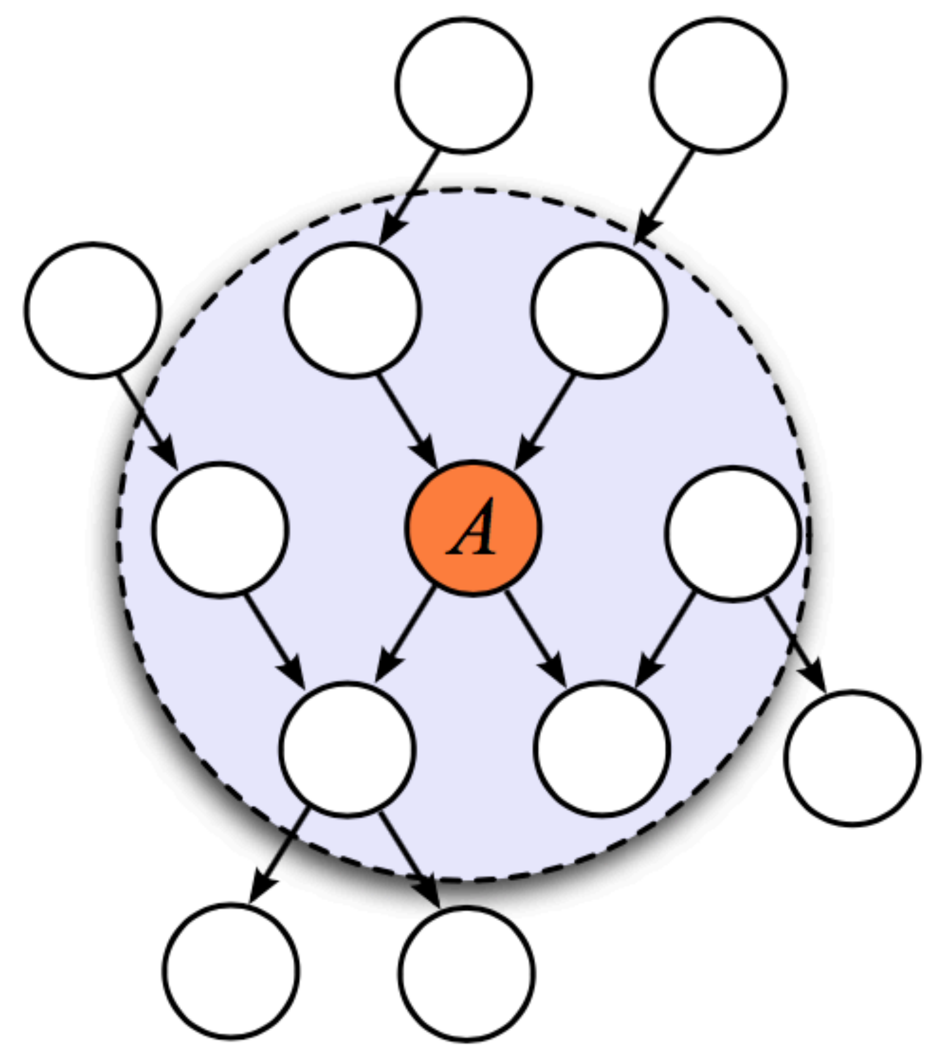
- Markov blanket: parents, children, children's parents
- Node independent of all others conditioned on Markov blanket



d-separation and Markov blanket

- Markov blanket: set of nodes that separate a node from all others

- d-separation: Method for determining whether a pair of nodes (or sets of nodes) are independent conditioned on another set

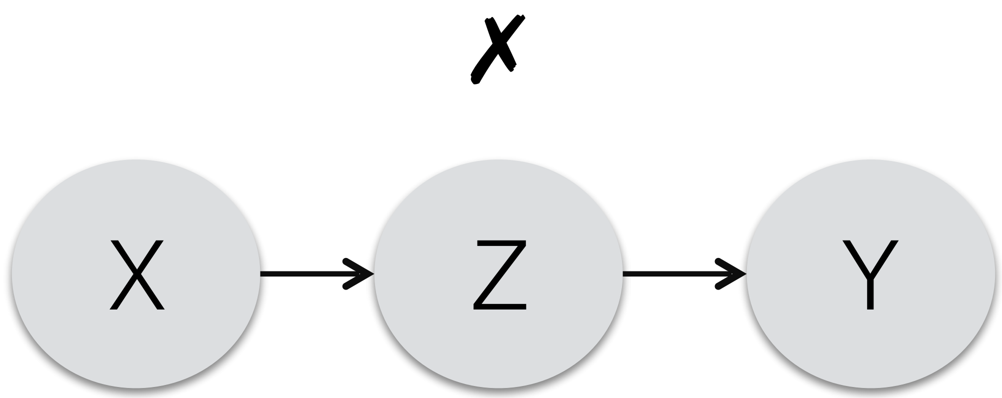


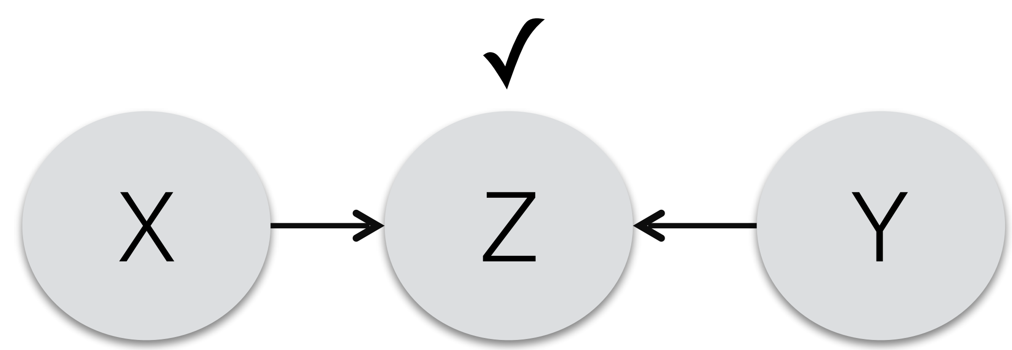
d-separation

- Equivalent statements, for sets of nodes X, Y, Z in graph G:
  - X and Y are d-separated by Z (Z can be node or set of nodes) in G
  - X and Y are conditionally independent given Z
  - Z blocks all paths between X and Y

Definition: d-separation

- Node v is a collider if two arrowheads meet at v





- X and Y are d-connected by Z in graph G iff
  - Exists an undirected path between a vertex in X and vertex in Y s.t. for every collider C on the path, C or descendant of C is in Z and no non-collider on path is in Z
- X and Y are d-separated by Z in G iff they are not d-connected by Z in G

Example 1

- $X \rightarrow Y \rightarrow Z$

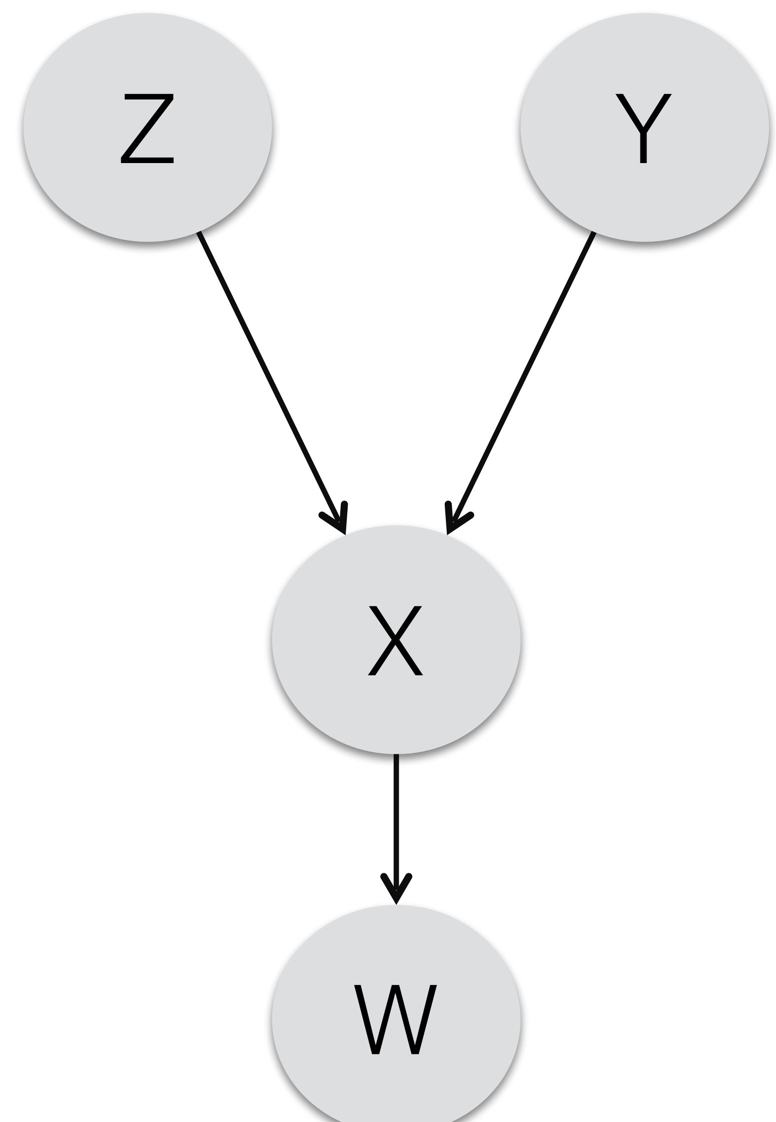
- $X \leftarrow Y \rightarrow Z$

- In both cases, X,Z d-separated by Y: no colliders on path from X to Z, and X and Z not d-connected by Y

Example 2

- Are Y,Z d-separated by W?

- No, d-connected by X and W is descendent



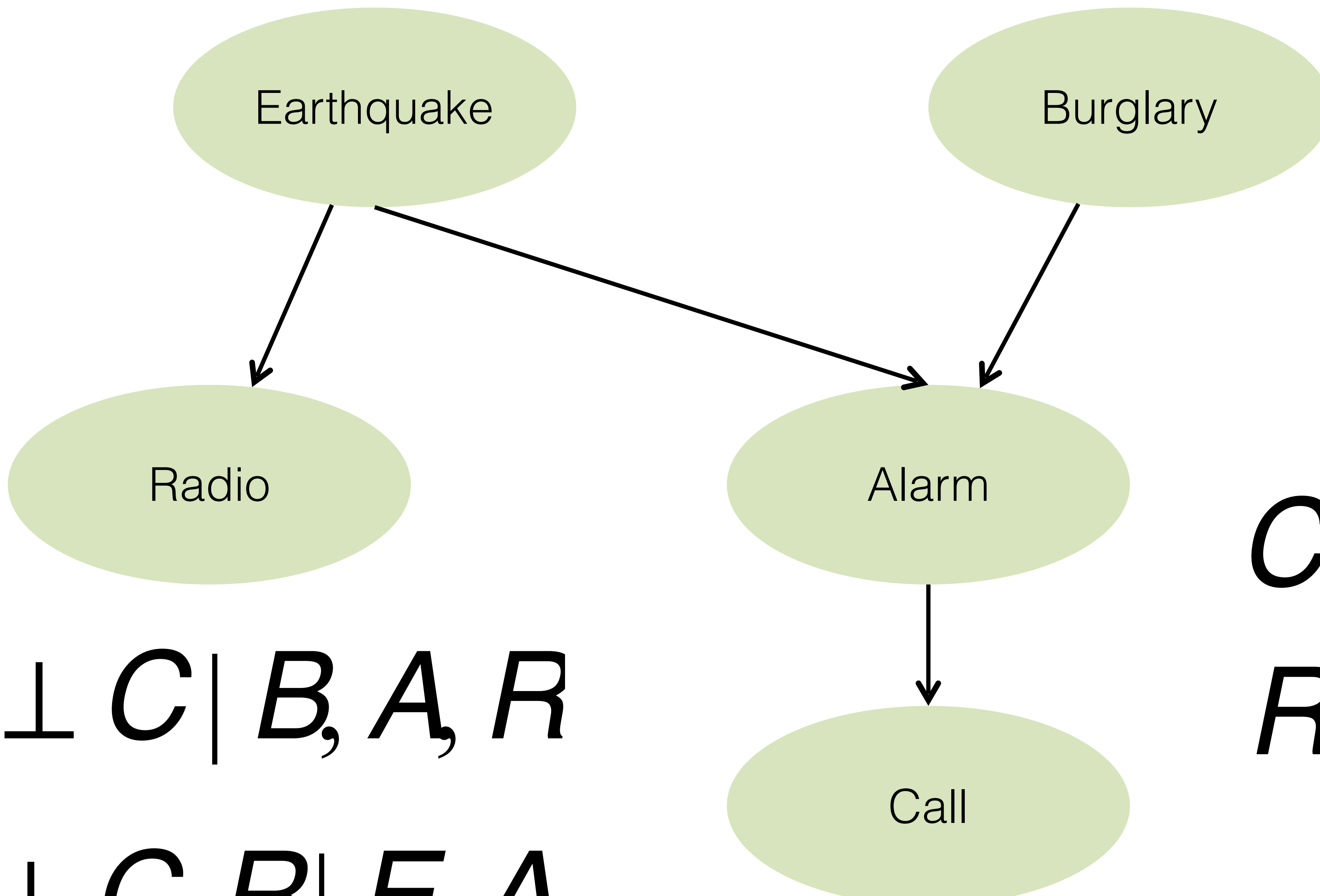
DAG and joint probability

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

Factorize:

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

Conditional independence from graph



$$E \perp B$$

$$A \perp R \mid E, B$$

$$B \perp E, R$$

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

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

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

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

Components of a Bayesian network

1. Directed acyclic graph

2. Conditional probability distributions

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

Conditional probability distribution

- In this lecture: discrete (usually binary valued) variables

- Conditional probability tables

X	Y	P(Z=T)	P(Z=F)
T	T	0.5	0.5
T	F	0.2	0.8
F	T	0.1	0.9
F	F	0.3	0.7

Simple method for building a BN

For $X_1, X_2, \ldots X_{n,}$:

-Add X_i to graph

-Add edges from $pa(X_i)$ where

$$P(X_i|X_1 \ldots X_{i-1}) = P(X_i|pa(X_i))$$

-Add conditional probability $P(X_i|pa(X_i))$

Pearl, J (1988) Probabilistic Reasoning in Intelligent Systems

Example (1)

- Coin A, Coin B

$$A \perp B$$

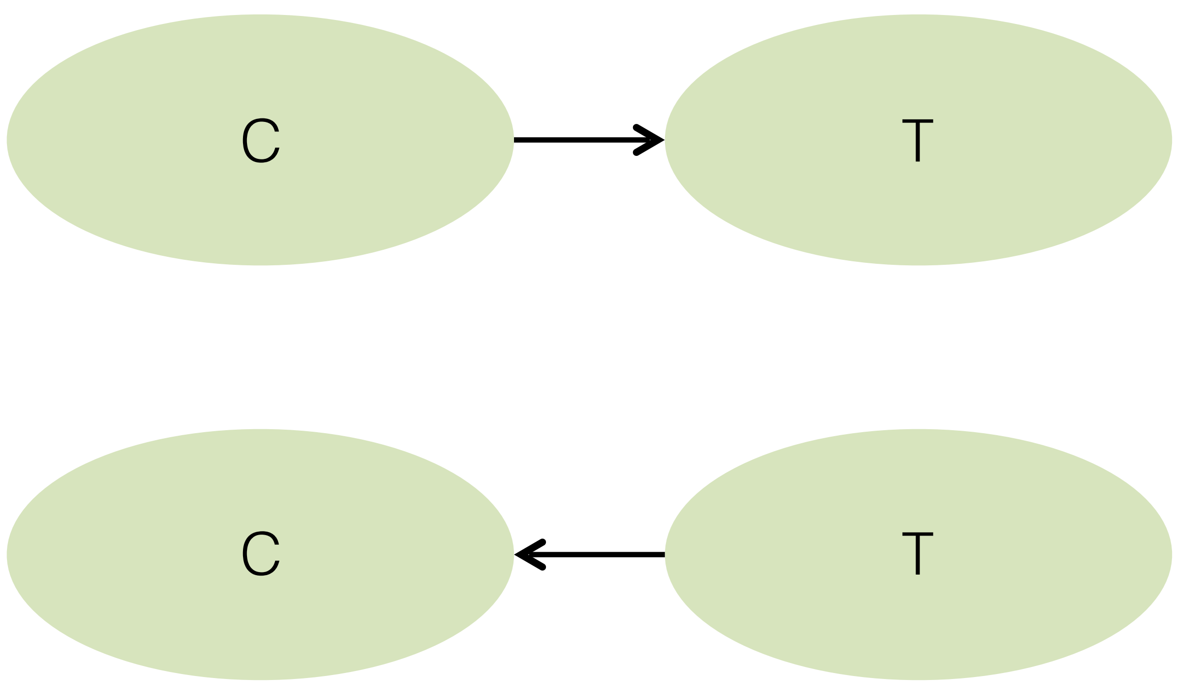
- What's Bayesian network?

Example (2)

- Cavity, Toothache

$$C \not\perp T$$

- Possible BNs?

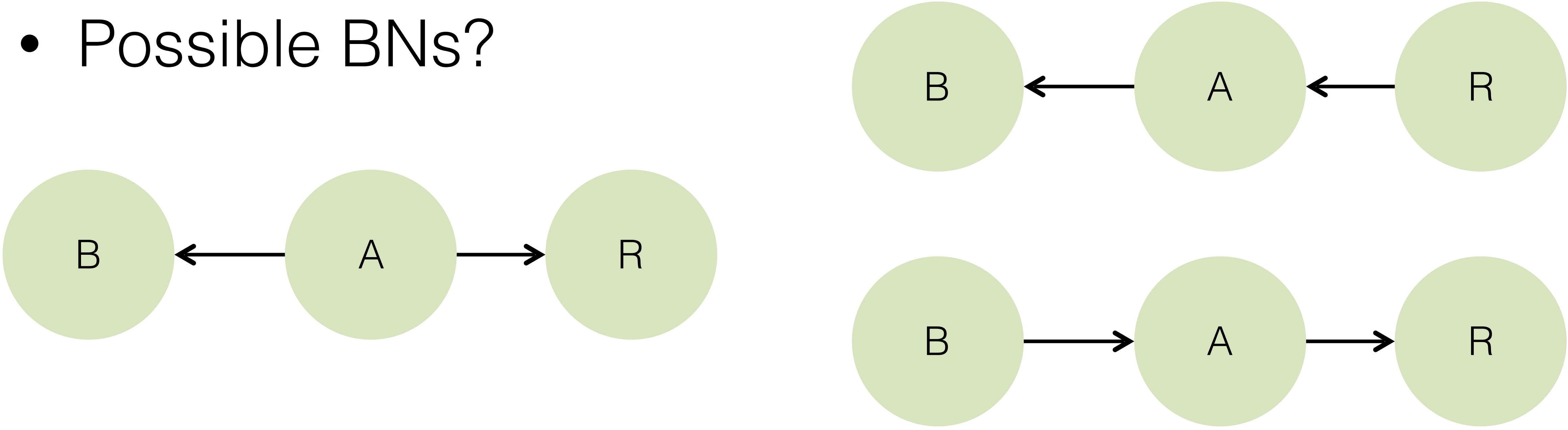


Example (3)

- Barometer, Rain, Air Pressure

$$B \perp R | A$$

- Possible BNs?

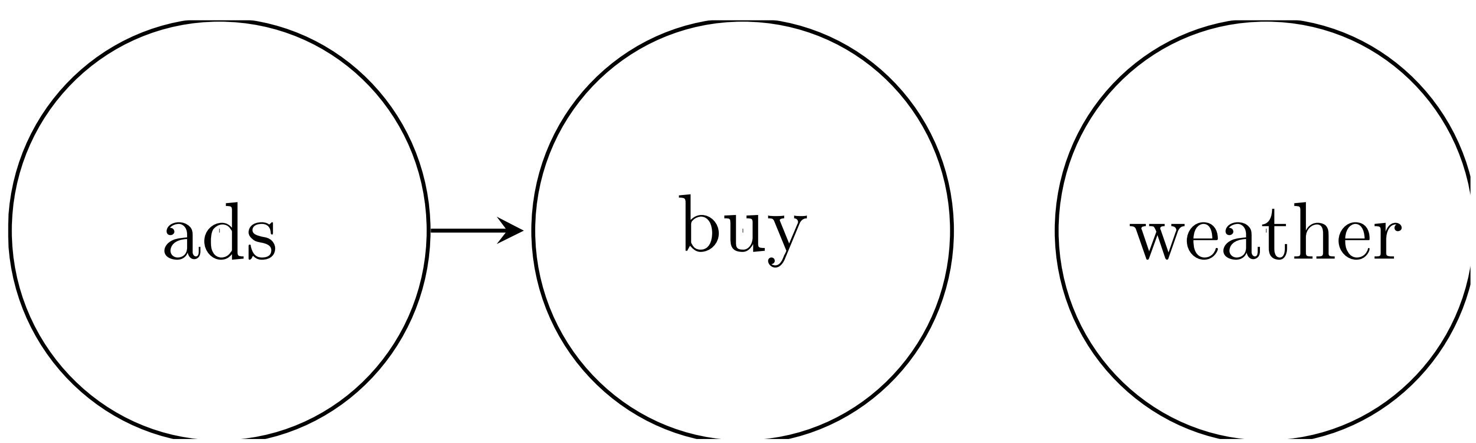


Notes on creating a BN

- Multiple BNs consistent with same relationships
- Order of operations can matter

Causal graph

- Arrows denote direct causes
  - Edge from X to Y means X causes Y
- DAG



What makes a BN causal?

- Causal Markov condition
- Faithfulness
- Causal sufficiency

- + a few other assumptions, e.g. variables “correctly” specified

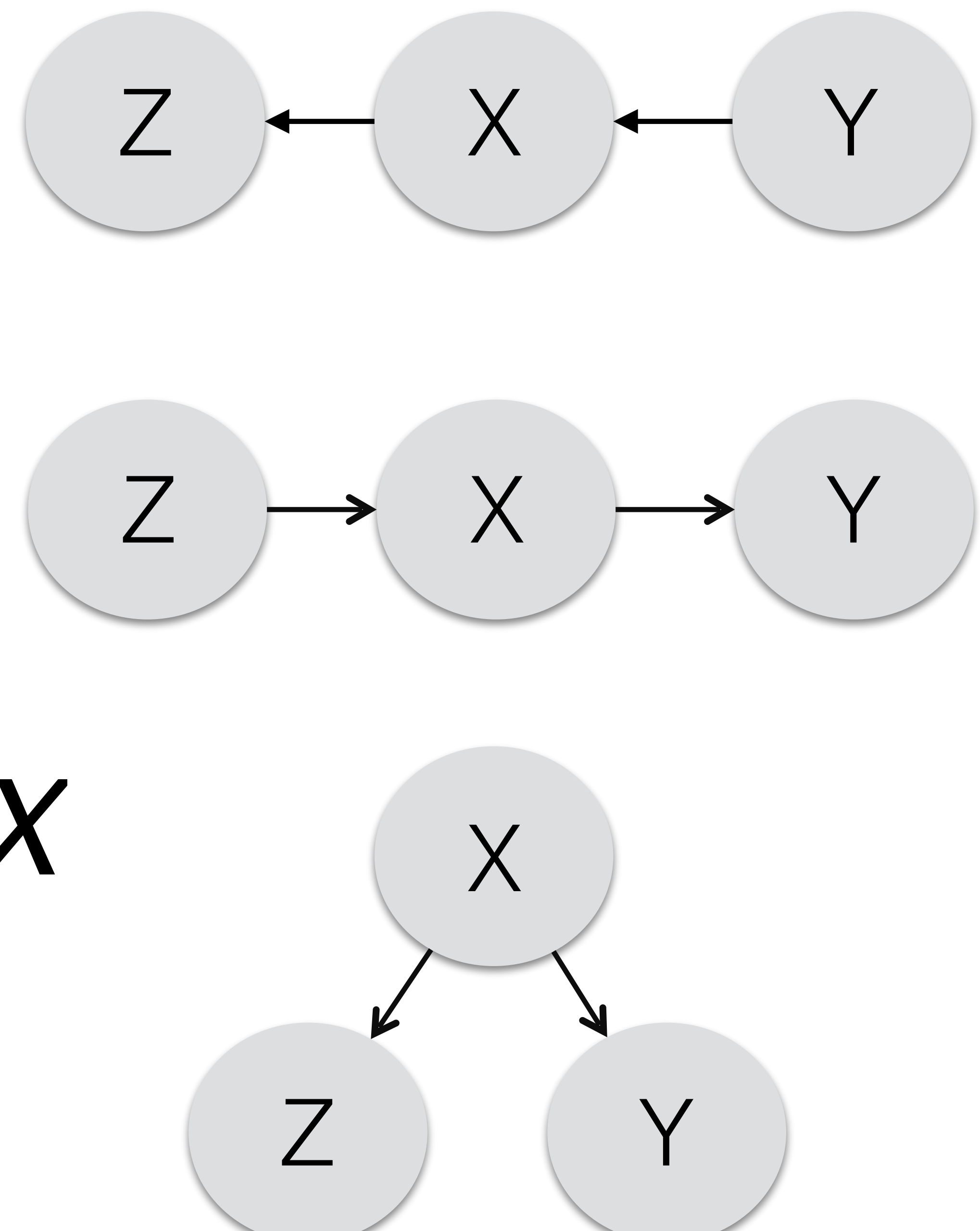
Causal Markov condition (CMC)

Node in the graph is independent of all of its non-descendants (direct and indirect effects) given its direct causes

But…

- Independence implies many networks
- Network implies1 set of independence relationships

$$Y \perp Z \mid X$$



…Also

$$P(C_1 = H \wedge C_2 = H) > P(C_1 = H)P(C_2 = H)$$

$$5/10 > 6/10 * 6/10$$

$$C_1 \not\perp C_2$$

Coin 1	Coin 2	# obs.
H	H	5
T	T	3
H	T	1
T	H	1

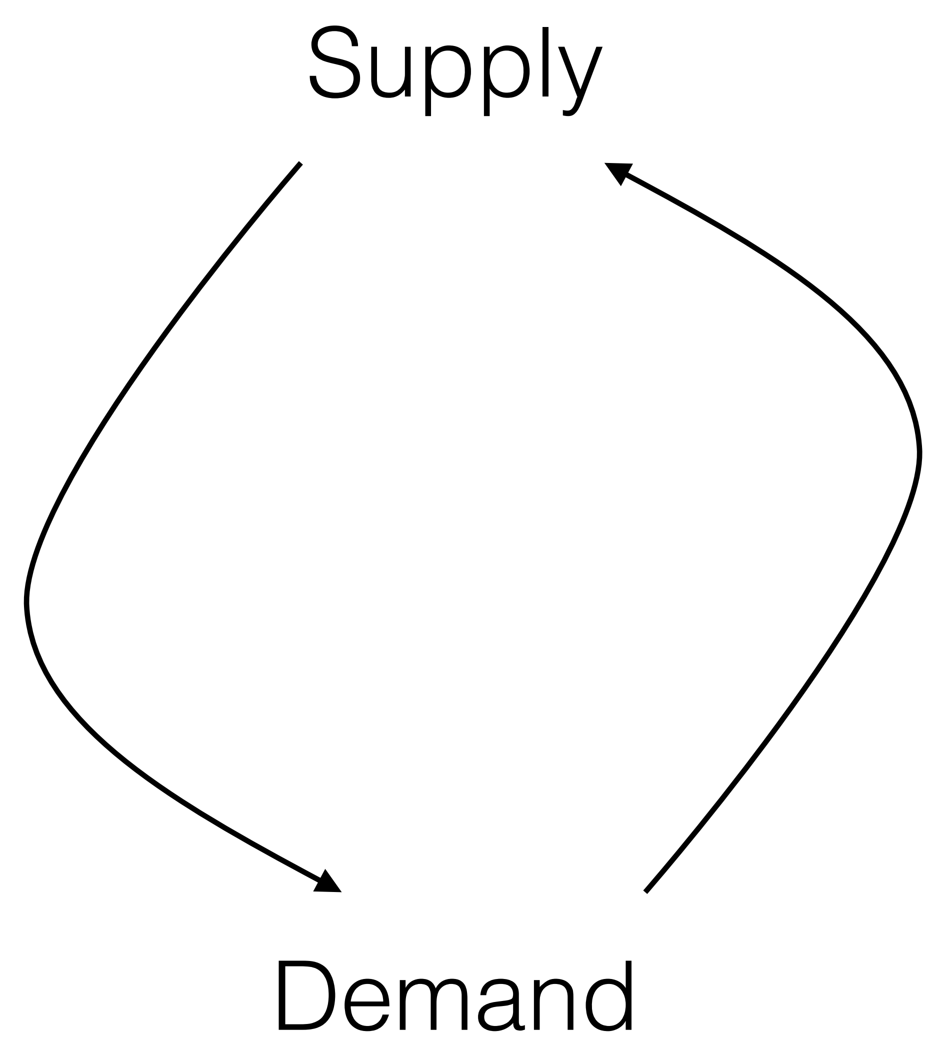
CMC and screening off

- Recall Common Cause Principle (CCP)
    - If $P(X \wedge Y) > P(X)P(Y)$ then either X causes Y (or vice versa) or they have a common cause

- Now: if $P(X \wedge Y) > P(X)P(Y)$ and they have a common cause C, it means X ind Y | C

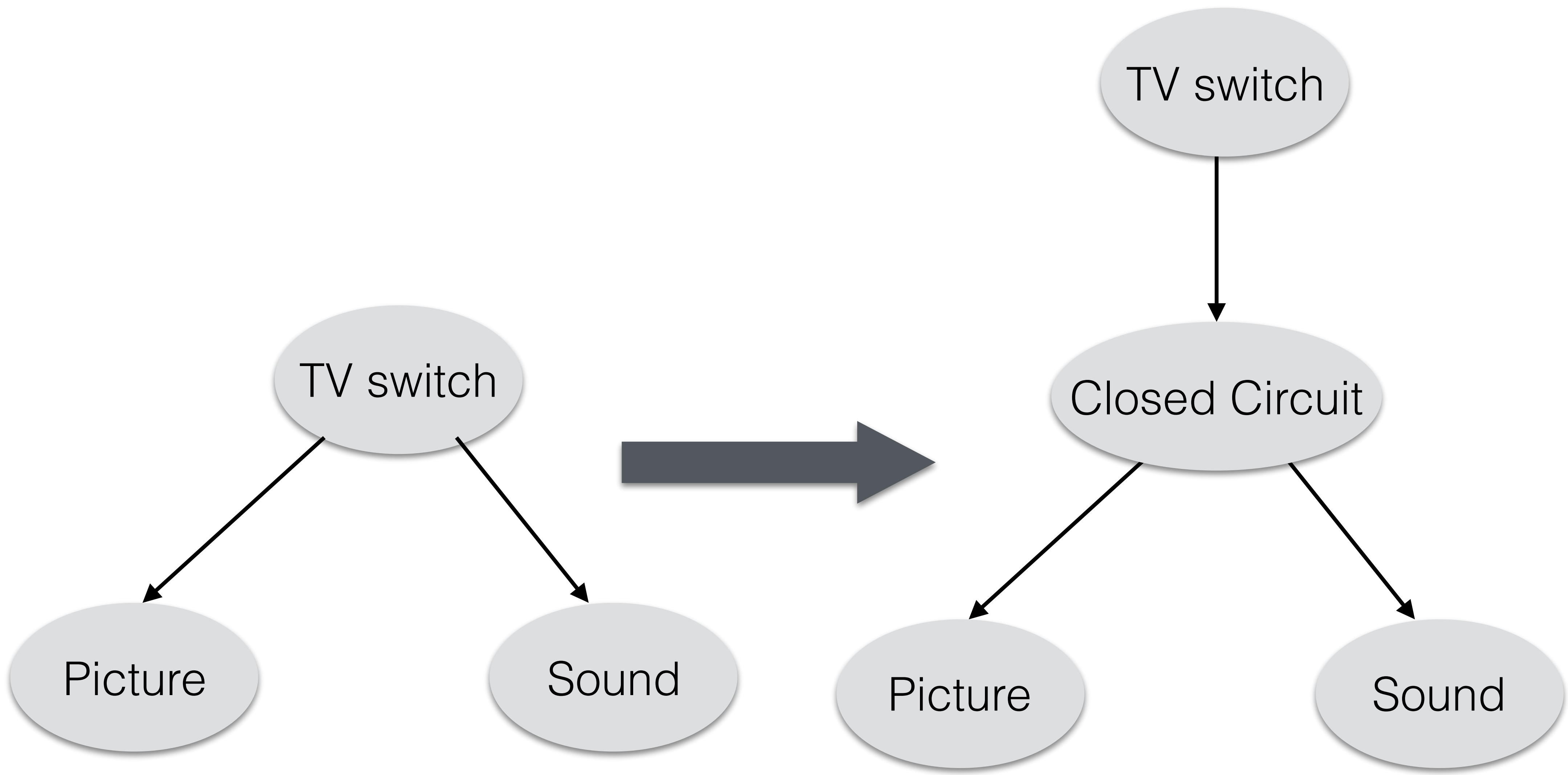
- Note that CCP seeks single common cause. CMC allows for sets of nodes.

Problems: feedback



Problems: Indeterminism

- $P(\text{picture}|\text{switch}) < P(\text{picture}|\text{switch},\text{sound})$



Spirtes, P., Glymour, C., & Scheines, R. (2000). Causation, prediction, and search. MIT Press

Completeness of graph

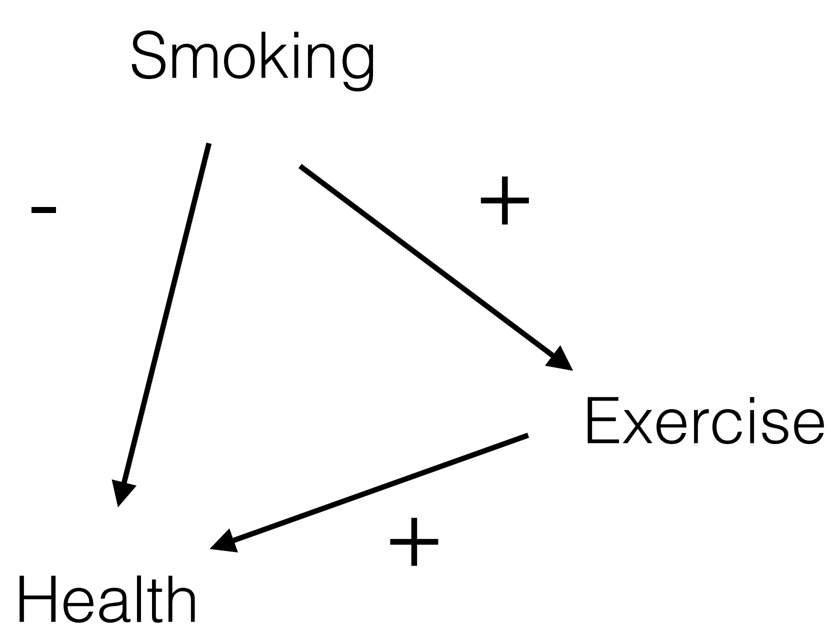
- Complete: all common causes included, all causal relations among variables included

- Incomplete: not all intermediate factors necessarily included

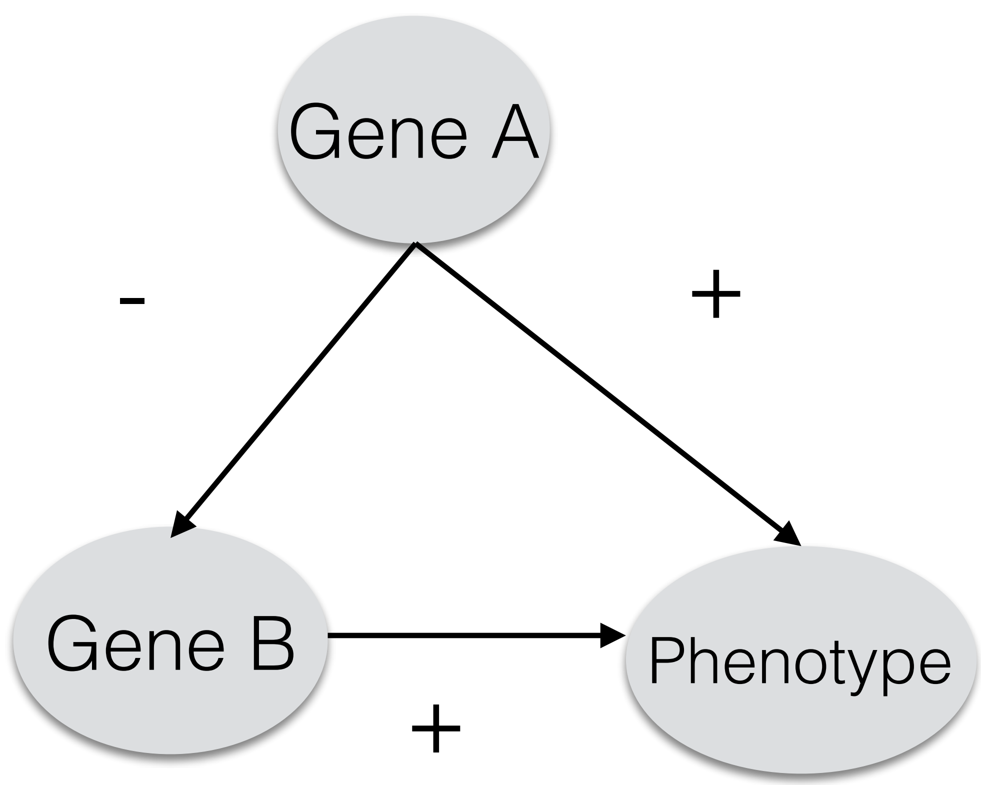
Faithfulness

- Exactly the dependencies in the underlying structure hold in the data
  - i.e. Independence relations not from chance but from structure
  - No canceling out

Example

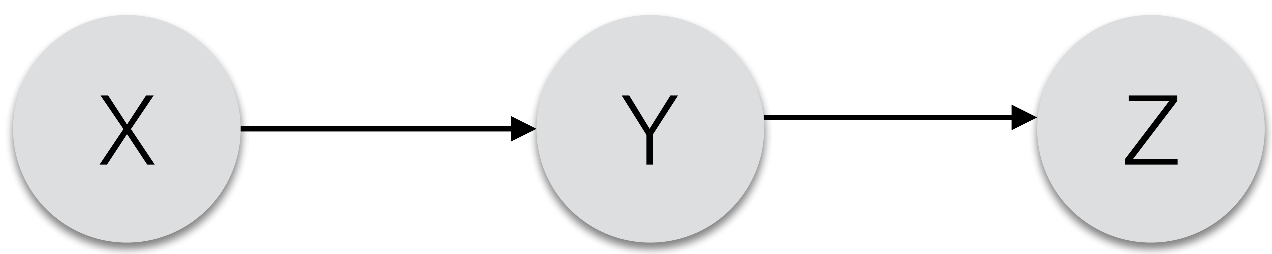


Another example

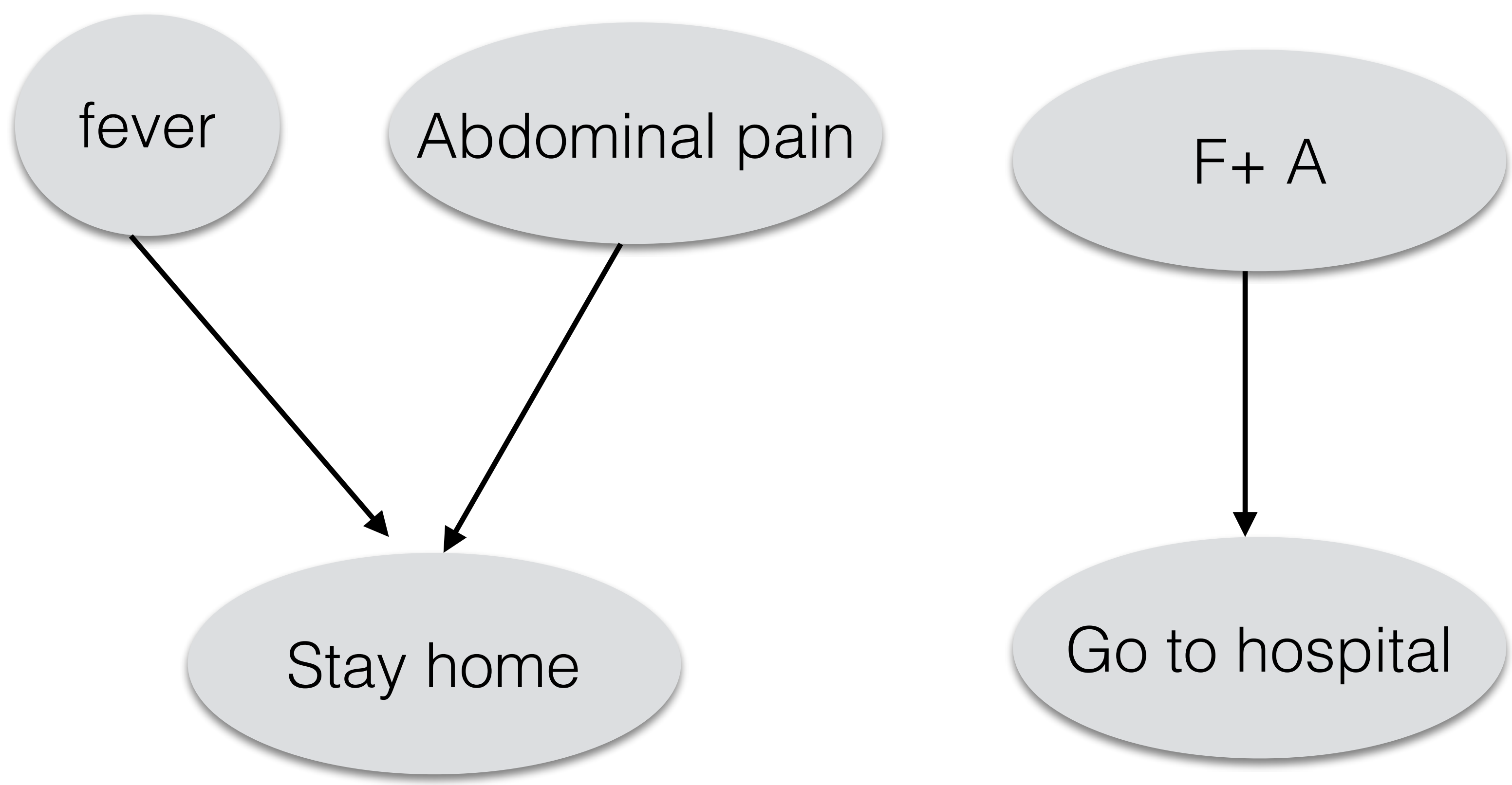


A final example (deterministic chain)

$$X \perp Z \mid Y$$



Selection bias



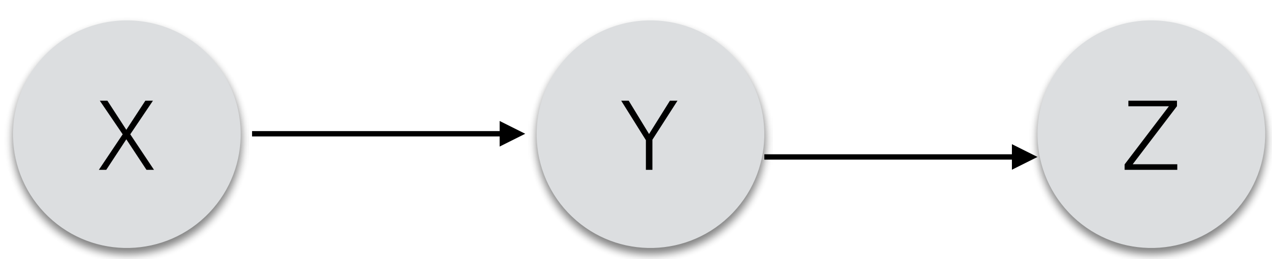
Cooper, G. F. (1999). An overview of the representation and discovery of causal relationships using bayesian networks. In C. Glymour & G. F. Cooper (Eds.), Computation, causation, and discovery. AAAI Press and MIT Press

Recap of problems for faithfulness

- Only true in large sample limit
- Simpson's paradox
- Selection bias
- Statistical tests

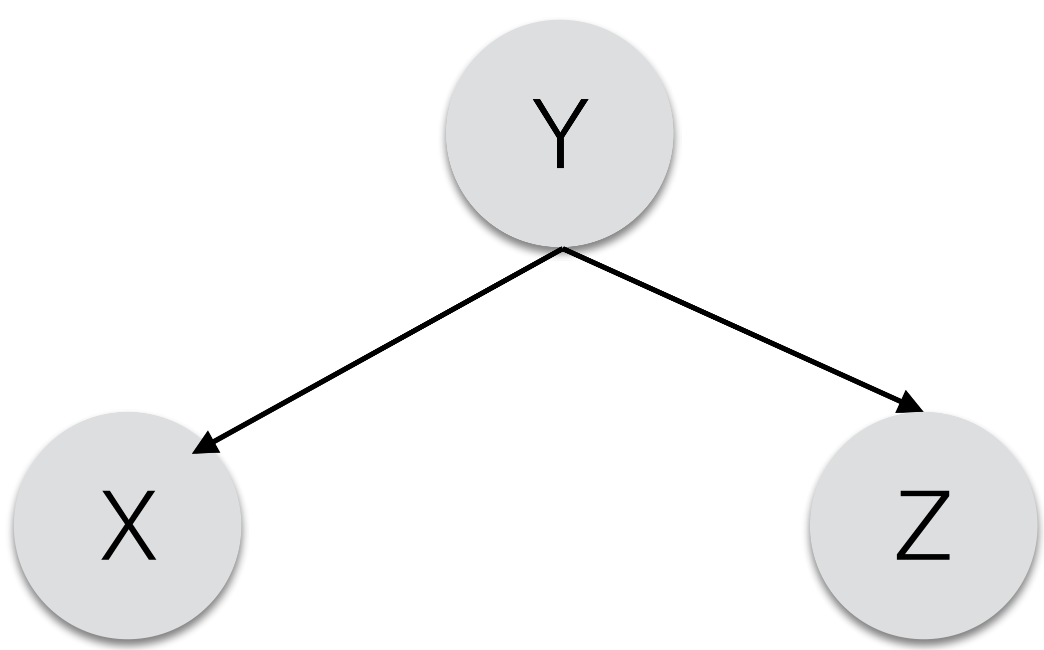
- CMC: population produced by structure has these independencies

- Faithfulness: population has only these independencies Why do we need both?



Causal sufficiency

- All common causes of pairs of variables measured
- Not sufficient if Y not measured



Completeness vs. sufficiency

- Completeness: common causes are included in causal graph
- Sufficiency: all common causes have been measured

In absence of sufficiency…

- Can still learn something
  - Some relationships may appear in all graphs
  - Can find set of all graphs representing independence relations, with nodes for possible hidden variables
- Timing information helps

Recap of causal inference with BN

- What makes a Bayesian network causal?
    - The assumptions: CMC, sufficiency, faithfulness

- Assumptions+Data—> Independencies —> Causal BN(s) —> effects of interventions

Uses for BNs

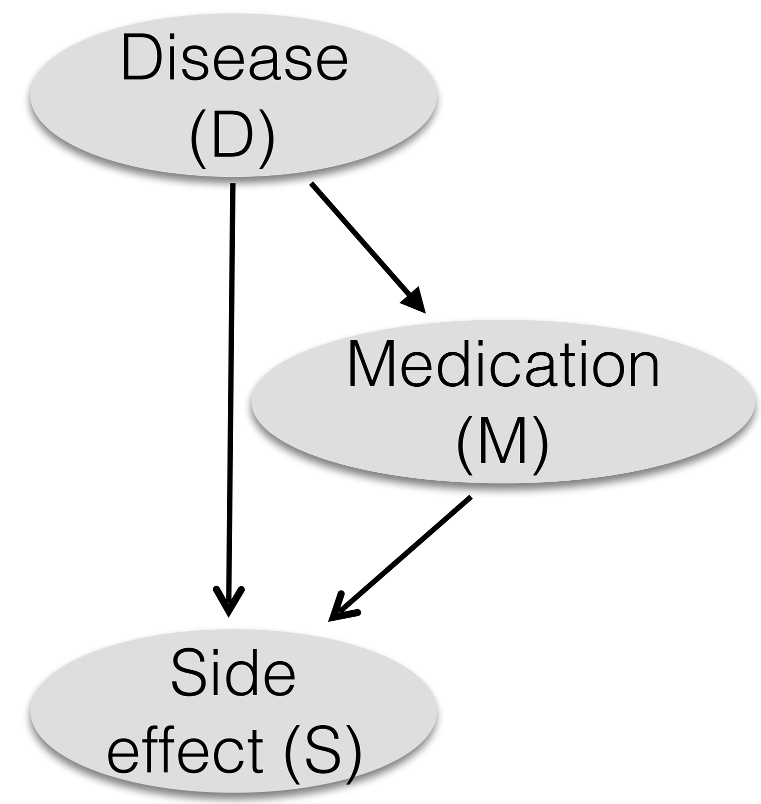
- Actions
    - What happens if we do X?
- Counterfactuals
    - What if things happened differently?
- Explanations
    - Why did X happen?

Counterfactuals reminder

- If I had not gone running, I would not have gotten a sunburn
- If the patient had taken the drug, she would have recovered
- Had I bought shares of Apple stock in 2004, I would have made a large profit

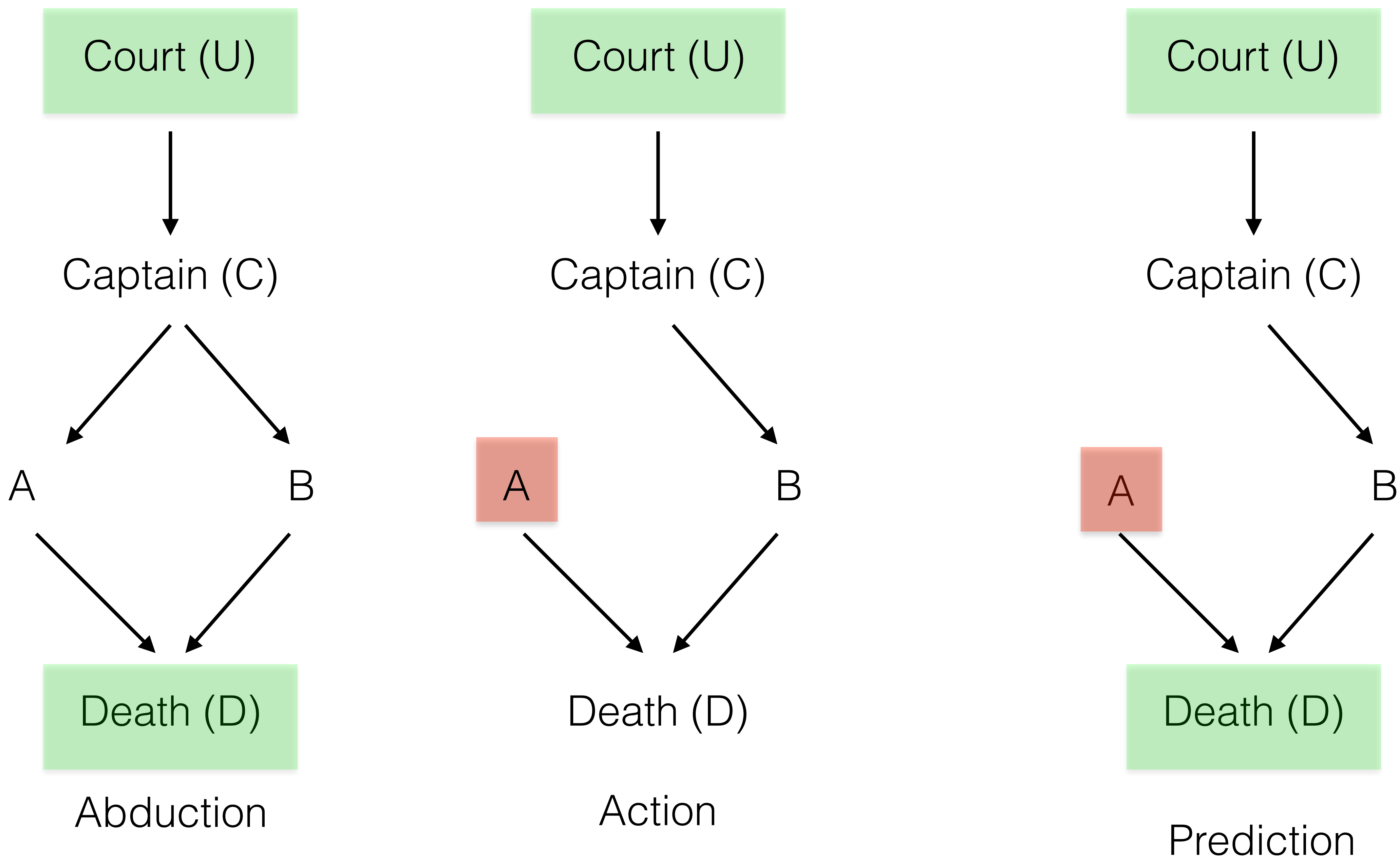
Pearl on Counterfactuals

- Like do(), except backward looking and changing value of variable
- Three steps
  - Abduction: use evidence to interpret past
  - Action: change to hypothetical values
  - Prediction: see consequences of actions



If D, then D would still be true if A were false

$D \rightarrow D_{\neg A}$



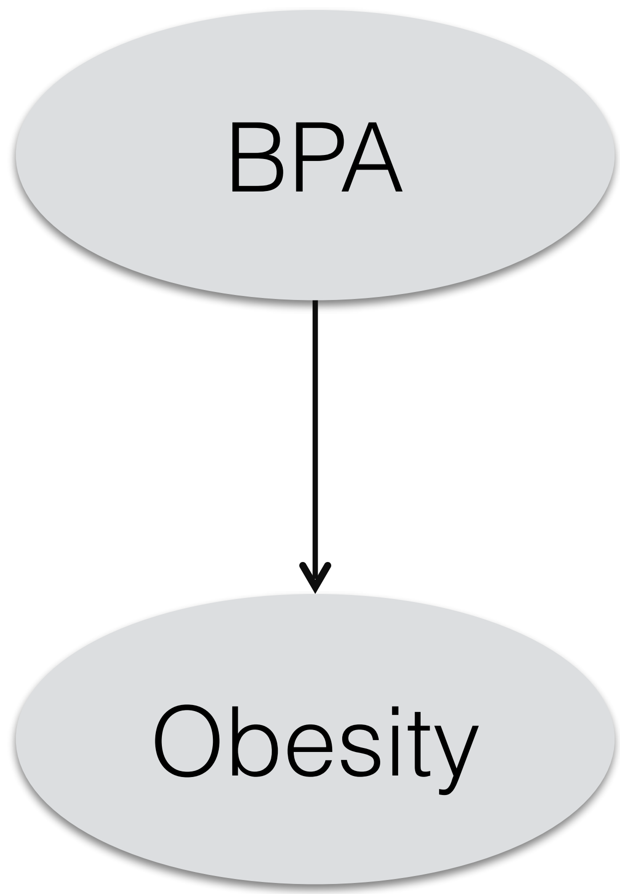
Pearl, J. (2000). Causality: Models, reasoning, and inference. Cambridge University Press

JOURNAL CLUB

Back to BNs

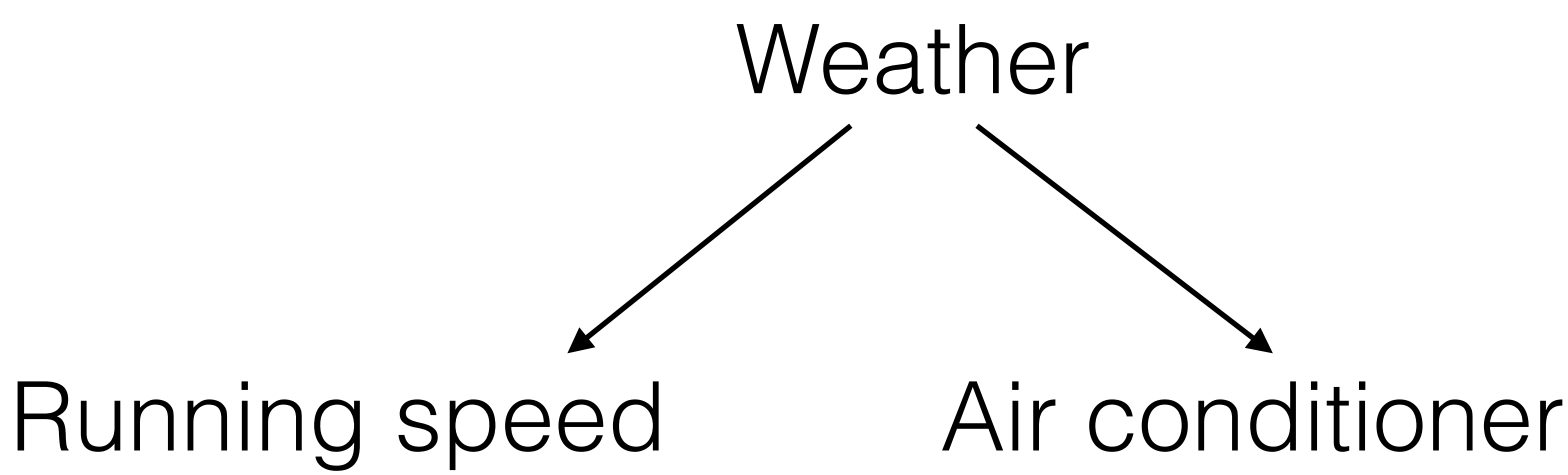
- Interventions
- Structure + parameter learning

Manipulability

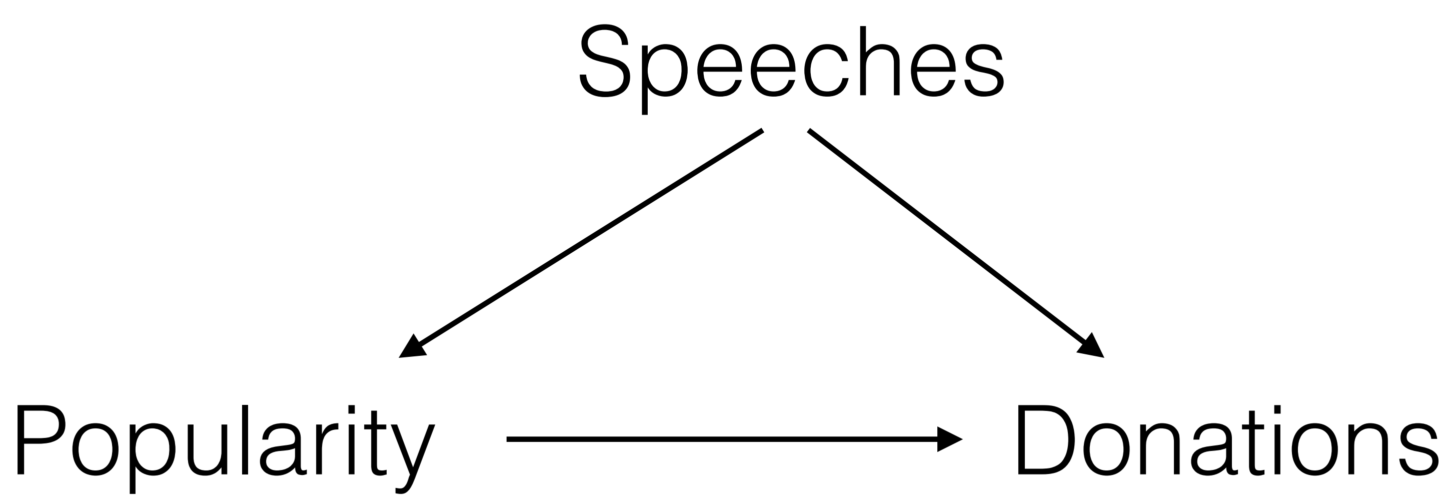


Ideal manipulations

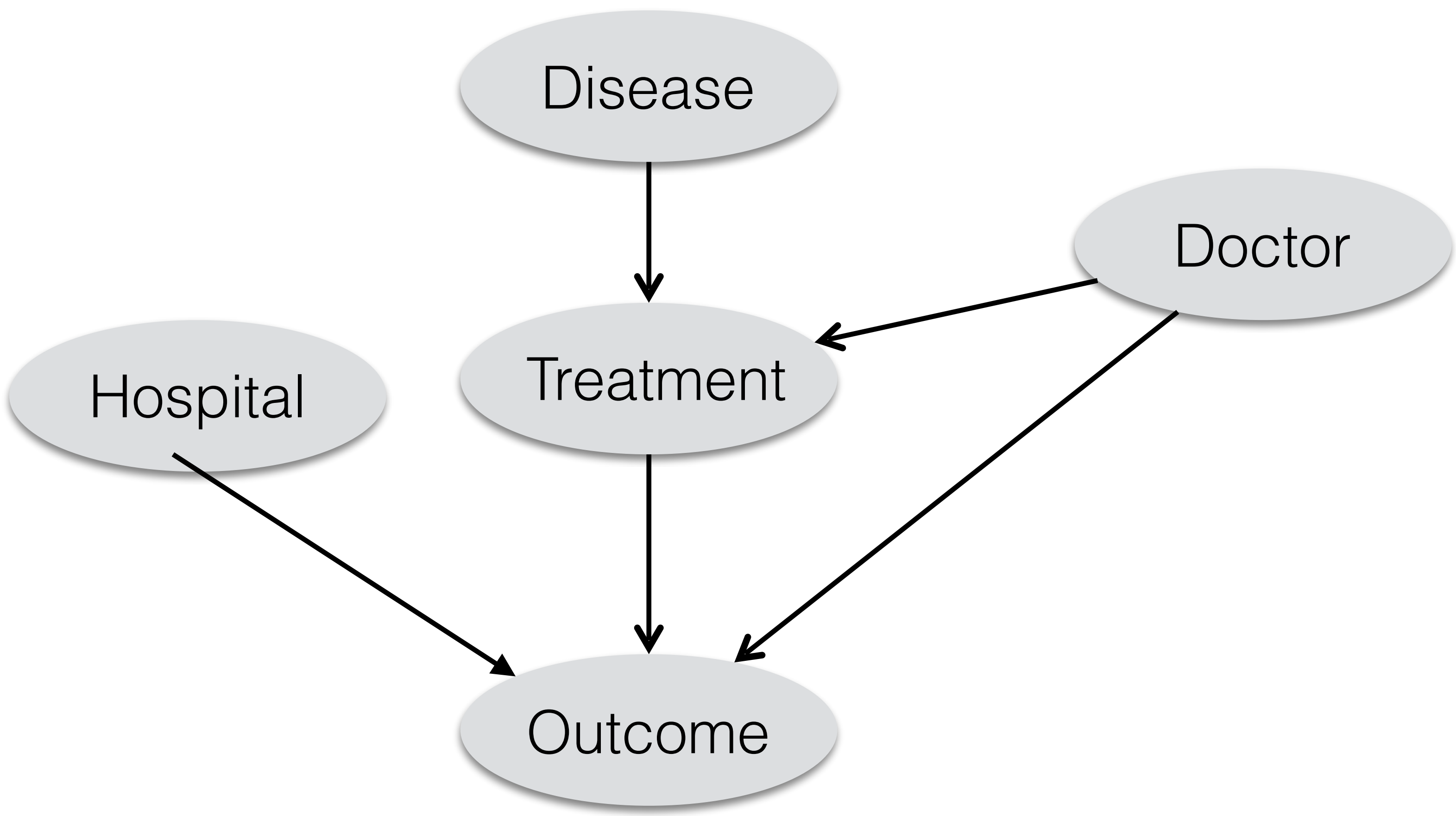
Definition: change in value of a variable that does not introduce any other changes (except those produced by the change in variable)

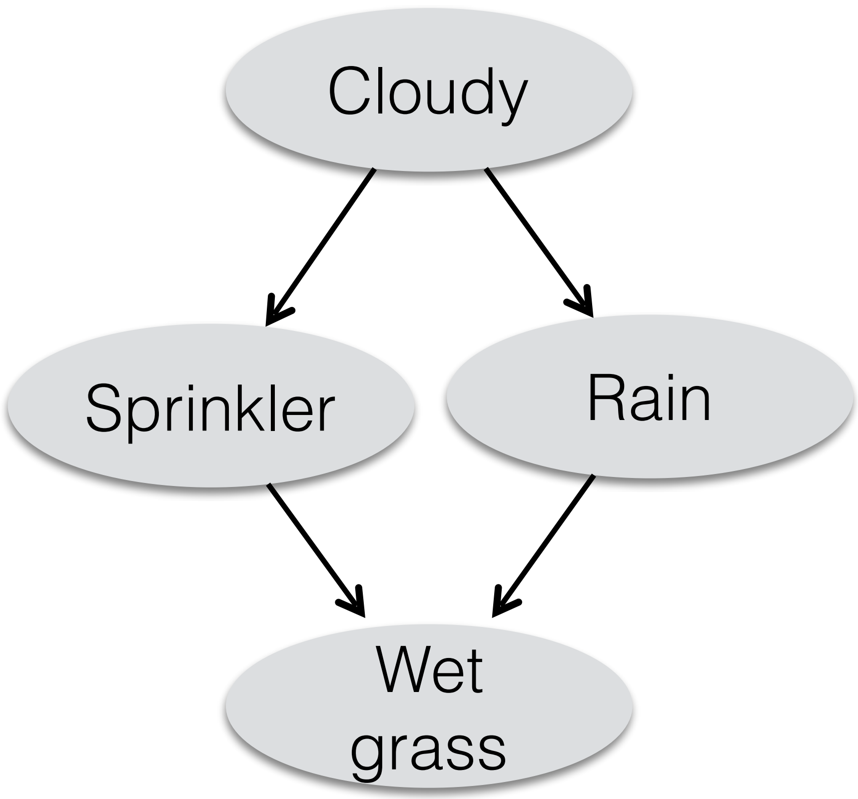


Testing popularity, how do we manipulate it’s value?



Seeing versus doing

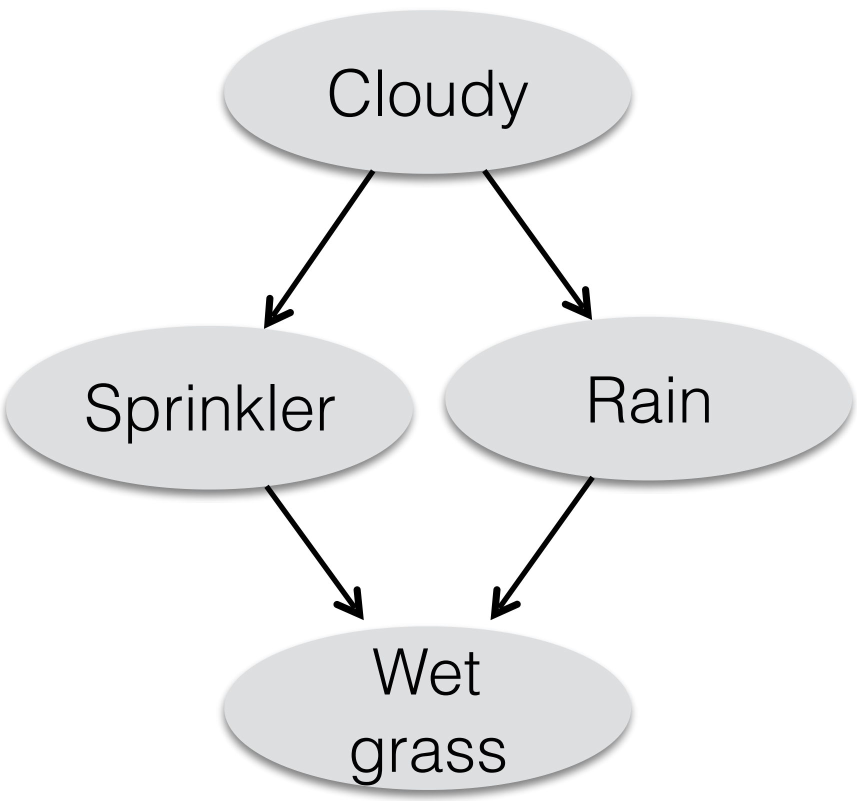




What's P(C) if I turn the sprinkler on?

Is this the same as P(C|S=T)?

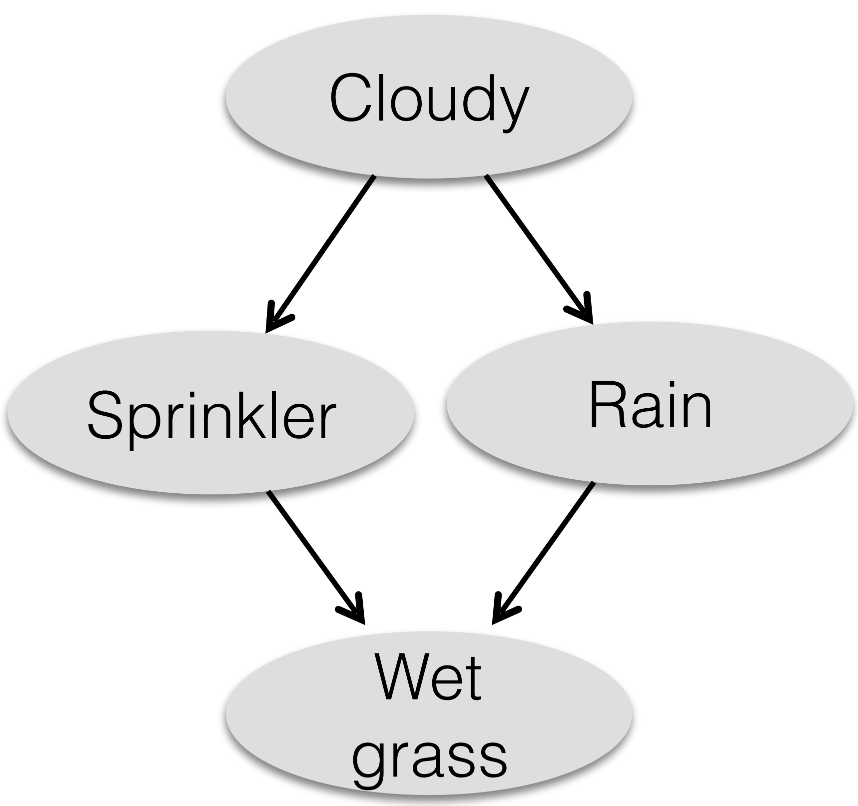
Intervention and joint probability



- != just incorporating evidence
  - Evidence: set value of observed values
  - Intervention: set value by forcing variable to take value independent of its parents' values

If turn on sprinkler, the fact that it's on no longer gives info about C

Intervention and joint probability



$$P(C,S,W,R) = \sum_{C,S,W,R} P(c)P(s \mid c)P(r \mid c)P(w \mid s,r)$$

$$P(C,W,R \mid do(s)) = \sum_{C,W,R} P(c)\underline{P(s)}P(r \mid c)P(w \mid s,r)$$

do() operator

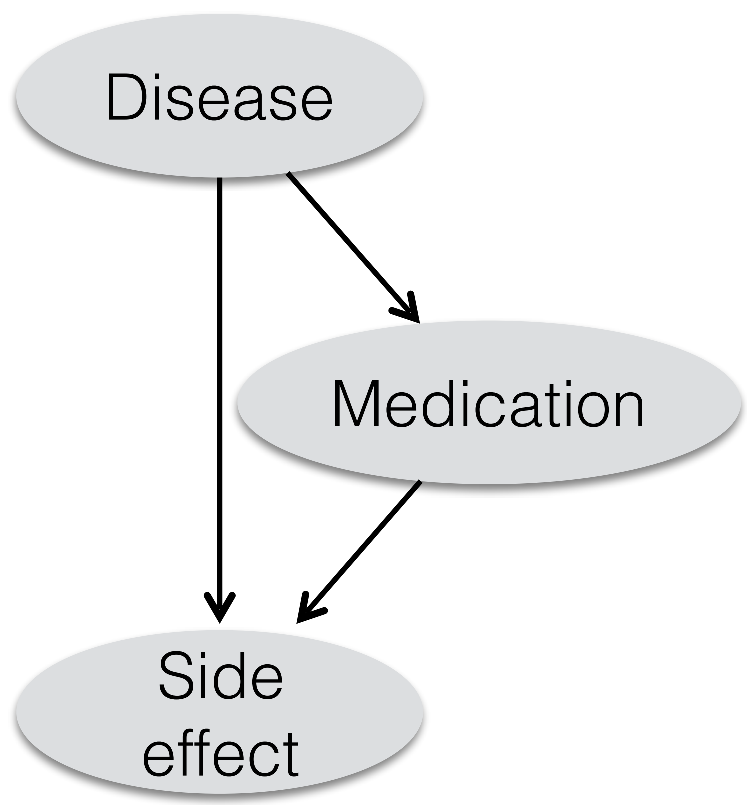
- Model can help us determine the effect of interventions

- P(X=x|Y=y) != P(X=x|set Y=y)

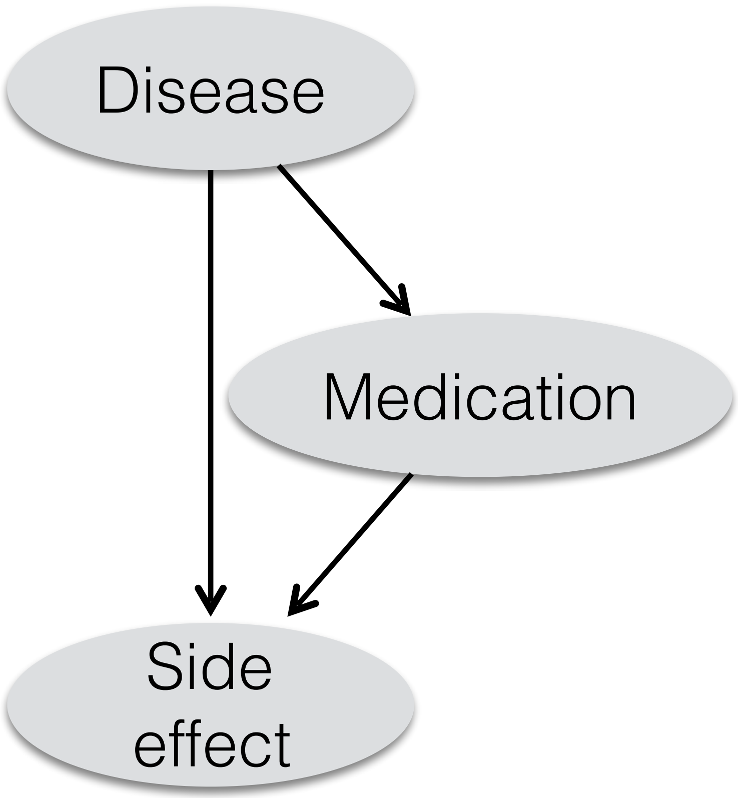
- Big assumption: can set variable T/F!

Example

P(S|do(M))



Example



$$P(s \mid \mathrm{do}(m)) = P(s \mid \hat{m}))$$

$$= \sum_d P(s, d, \hat{m}) / P(\hat{m}) = \sum_d P(s \mid d, \hat{m}) P(d \mid \hat{m}) P(\hat{m}) / P(\hat{m})$$

BUT! $P(d \mid \hat{m}) = P(d)$

$$P(m) / P(m) = 1$$

SO

$$= \sum_d P(s \mid d, \hat{m}) P(d)$$

Summary of do-calculus

- Insertion/deletion of observations
- Action/Observation exchange
- Insertion/deletion of actions

- In general, may have unobserved/hidden variables

Some caveats

- Time
- Modularity
- Possibility of intervening
- Efficacy

Learning

- Structure
- Parameters

Structure learning approaches

- Search and score
  - Define scoring function for how well structure fits data
  - Search over set of graphs, maximizing scoring function

- Notes
  - Computational complexity

Structure learning approaches

- Constraint based
  - Use repeated conditional independence tests
  - E.g. Connect all nodes with undirected edges, repeatedly do conditional independence tests to remove/orient edges
- Notes
  - An incorrect edge removal/orientation affects later tests

Search and score

- Define scoring function
  - Bayesian information criterion (BIC)
    - Aims to find most compact graph fitting data
    - Note minimality condition
    - $\log P(D|G) \approx \log P(D|G,\theta_G) - \log N/2 * \mathrm{Dim}(G)$

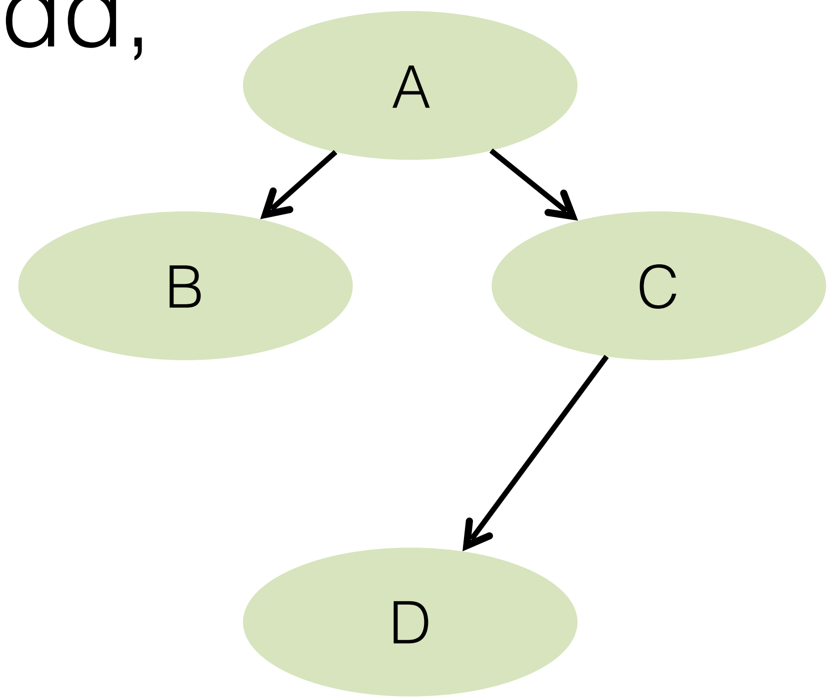
    Dim=#parameters, m=data size

- How to search over set of graphs?

- Overfitting

Heuristics

- Can't search over all graphs
- But
  - Can explore nearby graphs: add, remove, reorient edges



Greedy hillclimbing

- Apply all possible alterations, pick highest scoring change, iterate with new graph until no changes improve score
- Can get stuck in local maxima
  - But can randomly restart
  - Simulated annealing

Constraint-based learning

- Find conditional independencies
  - Note: have to decide when to accept/reject independence
  - Add edges (or remove) according to these
- Get CPT after

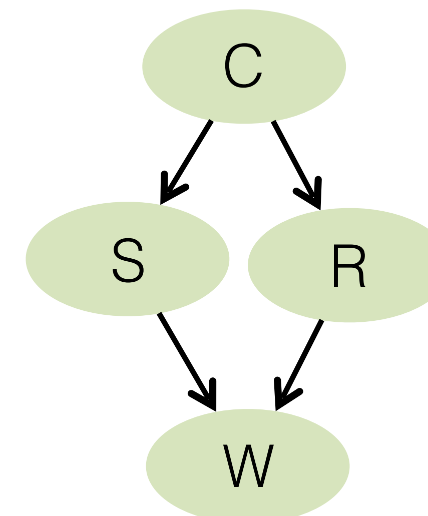
Things to beware of with inference

- Sample size
- Missing data (not just variables)
- Multiple testing (and FDR)
- What structures DAG can/cannot represent (e.g. time series and feedback)
- Variable representation

The good news

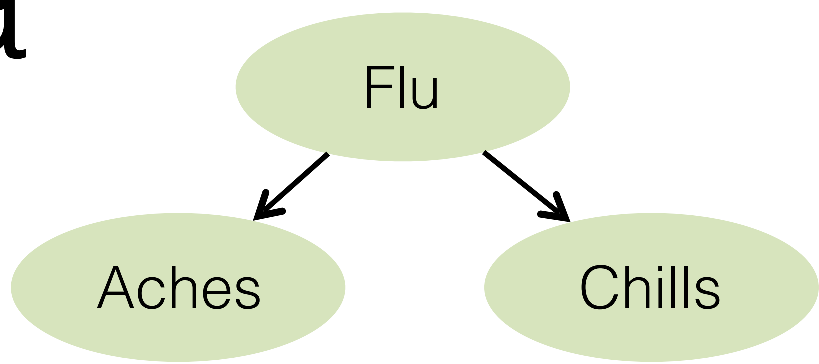
- Can add time
- Can experiment
- Methods for testing assumptions

Parameter learning



- Have structure, what's CPT for each node?
- Two cases
  - Complete data (no missing variables or values)
  - Incomplete data

Parameter learning: complete data

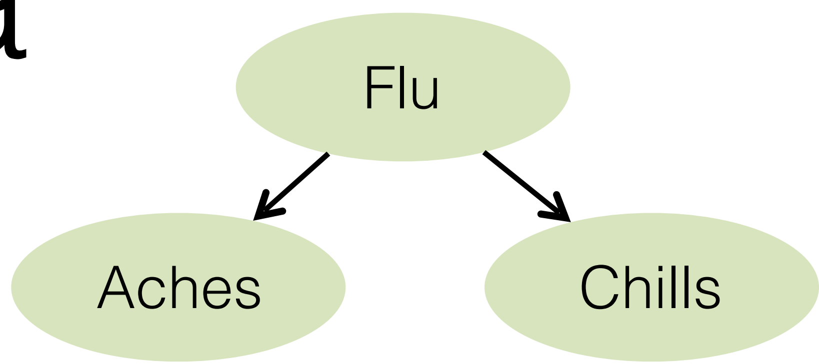


A	C	F
T	T	T
T	F	T
F	T	T
F	F	F
T	F	F
F	F	F
T	F	T

A	C	F	P
T	T	T	1/7
T	T	F	0/7
T	F	T	2/7
F	T	T	1/7
F	T	F	0/7
F	F	T	0/7
T	F	F	1/7
F	F	F	2/7

- Three parameters
  - P(F), P(A|F) , P(C|F)

Parameter learning: complete data

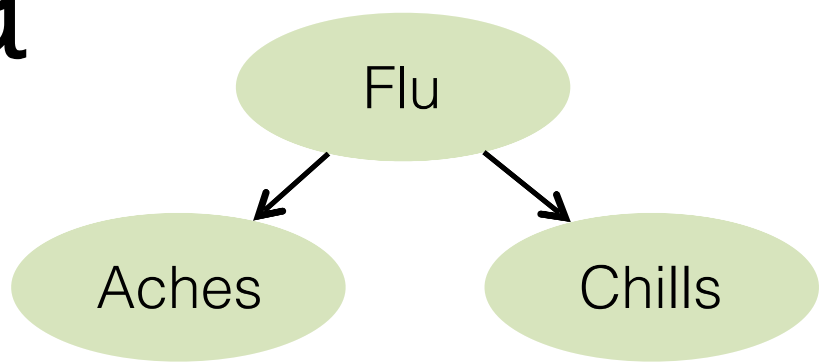


A	C	F	P
T	T	T	1/7
T	T	F	0/7
T	F	T	2/7
F	T	T	1/7
F	T	F	0/7
F	F	T	0/7
T	F	F	1/7
F	F	F	2/7

P(F=T)	P(F=F)
4/7	3/7

- Three parameters
  - P(F), P(A|F) , P(C|F)

Parameter learning: complete data



A	C	F	P
T	T	T	1/7
T	T	F	0/7
T	F	T	2/7
F	T	T	1/7
F	T	F	0/7
F	F	T	0/7
T	F	F	1/7
F	F	F	2/7

P(F=T)	P(F=F)
4/7	3/7

- Three parameters
  - P(F), P(A|F) , P(C|F)

F	P(A=T)	P(A=F)
T	(3/7)/(4/7) =3/4	(1/7)/(4/7) =1/4
F	(1/7)/(3/7) =1/3	(2/7)/(3/7) =2/3

Software

Overview: http://www.cs.ubc.ca/~murphyk/Software/bnsoft.html

Inference recap

	BN	DBN	Granger	Temporal logic
Results	Graph			
Time	No			
Data	C/D/M			
Cycles	No			
Latent vars.	Yes			
Prediction	Yes			
Token cause	Counterfactual - based			

Further reading

- Graphical models and causality
  - Spirtes, P., Glymour, C., & Scheines, R. (2000). Causation, prediction, and search. MIT Press
  - Pearl, J. (2000/2009). Causality: Models, reasoning, and inference. Cambridge University Press.
- Actual cause
  - Pearl's book
  - Halpern, J. Y., & Pearl, J. (2005). Causes and explanations: A structural-model approach. Part I: Causes. The British Journal for the Philosophy of Science, 56(4), 843-887.

For next week

- How can we find how long it takes for smoking to cause lung cancer?
- When to buy/sell a stock after you hear some news?