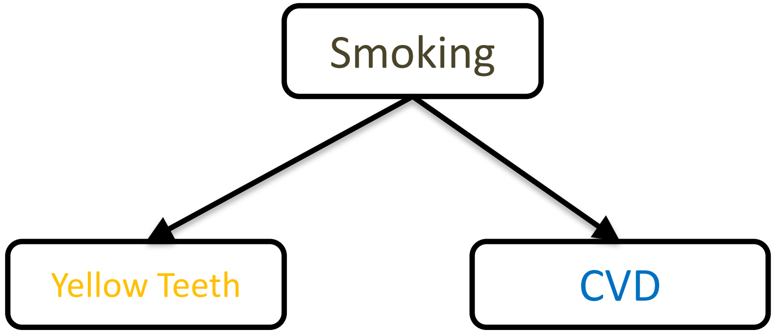
Probabilistic Graphical Models

Structure Learning – Pt2

Causality

Reminder:What if you do not know the graph

Graph G captures the qualitative relations



JPD J encodes the quantitative probabilistic properties

		CVD		
Yellow Teeth	Smoking	Y	N	
Y	Y	0.17	0.06	0.13
N	Y	0.06	0.02	0.08
Y	N	0.02	0.06	0.08
N	N	0.15	0.46	0.61
		0.4	0.6	1

Markov Condition (MC):

Every variable is **independent** of its non-descendants in the graph given its parents.

Faithfulness

Faithfulness Condition:

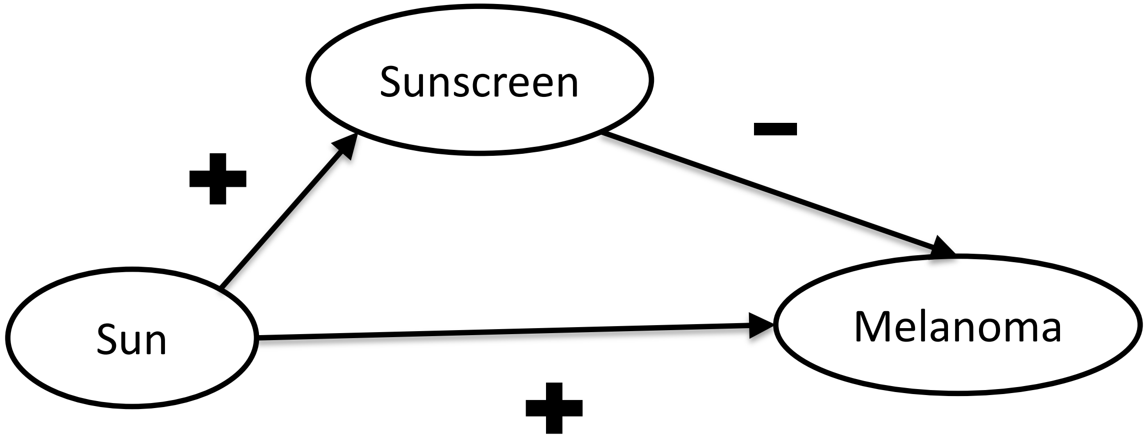
Independences stem **only** from the structure, **not the parameterization** of the distribution.

We say that the graph and the distribution are **faithful to each other**.

MC $\quad DSep(A, B|\mathbf{Z})$ in $G \Rightarrow A \perp\!\!\!\perp B|\mathbf{Z}$ in J

MC+FAITHFULNESS $\quad DSep(A, B|\mathbf{Z})$ in $G \Leftrightarrow A \perp\!\!\!\perp B|\mathbf{Z}$ in J

Faithfulness



The parameters do not cancel each other out!

Hypothesis Testing

- Identify the research question
- Writing the statistical hypotheses in terms of parameters of interest.
- Collect data and calculate a statistic
- Find the distribution of the statistic under the null hypothesis
- Find the p-value (probability that the result we got or a more extreme one happens just by chance given that the null hypothesis is true).
- Decide if the p-value is small or large
- Reject if p-value is lower than the significance threshold a.

Testing (In)Dependencies

Hypothesis Testing

- Identify the research question Is smoking independent from CVD?

- Writing the statistical hypotheses in terms of parameters of interest.
  P(smoking, CVD) = P(smoking)P(CVD)
- Collect data and calculate a statistic

- Find the distribution of the statistic under the null hypothesis

- Find the p-value (probability that the result we got or a more extreme one happens just by chance given that the null hypothesis is true).

- Decide if the p-value is small or large

- Reject if p-value is lower than the significance threshold a.

Example: Independence

- You have a population of 520 people
  - 160/520 smoke.
  - 210/520 have CVD.

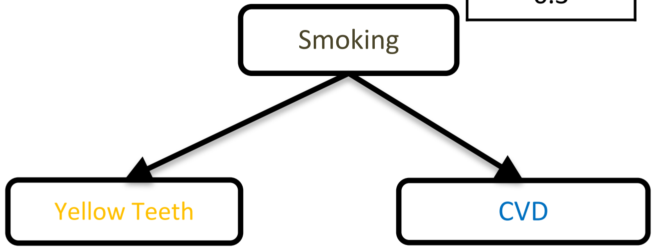
		CVD: Y	CVD: N	Total
Smoking	Y	120	40	160
	N	90	270	360
	Total	210	310	520

Contingency table

Reverse-engineering the graph

What you want

P(Smoking)
0.3



Smoking	P(YT)
Y	0.85
N	0.10

Smoking	P(CVD)
Y	0.75
N	0.15

Can we find the graph where the only d-separation is CVD and Yellow teeth given smoking?

What you have

You can use tests of conditional independence to identify the set of conditional independencies:

Here you only have one independence:

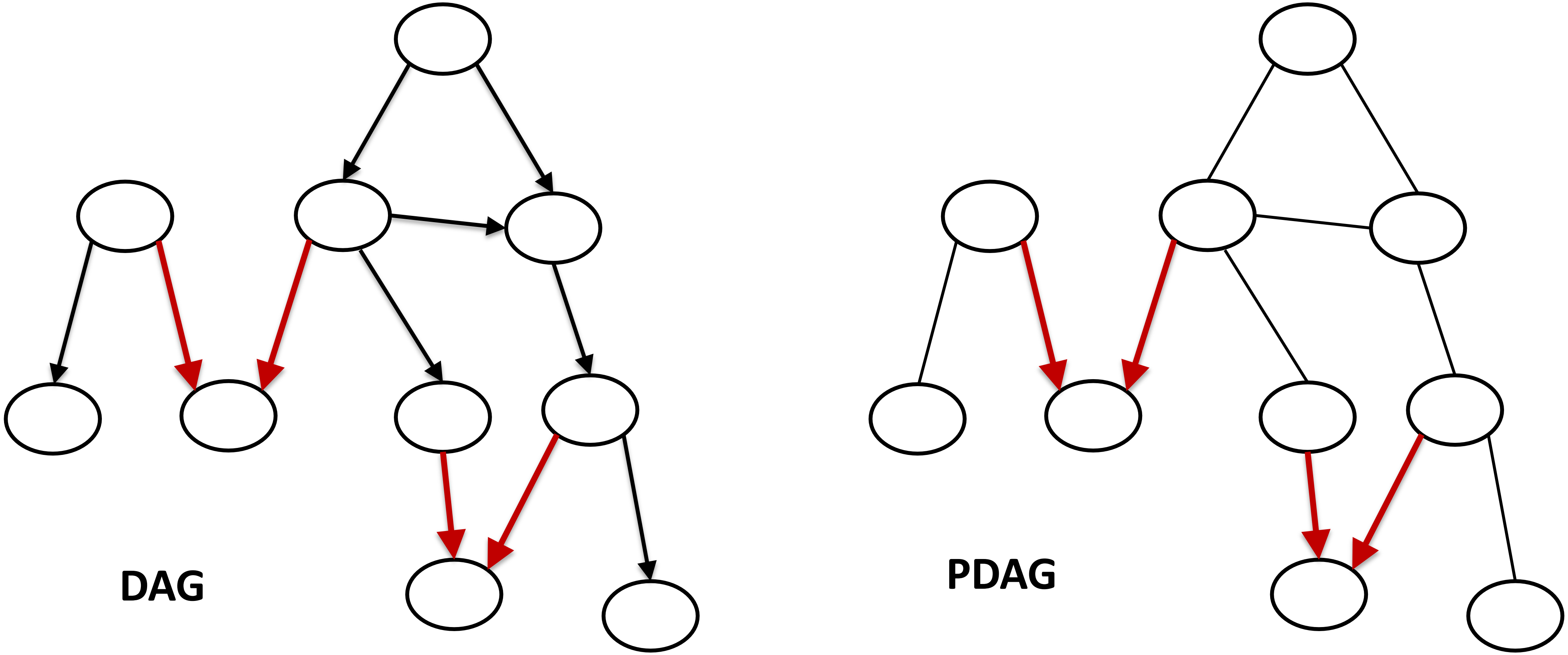
CVD $\perp\!\!\!\perp$ Yellow Teeth | Smoking

And the rest are dependencies:

Smoking $\not\perp\!\!\!\perp$ Yellow Teeth | $\emptyset$
Smoking $\not\perp\!\!\!\perp$ Yellow Teeth | CVD

Smoking $\not\perp\!\!\!\perp$ CVD | $\emptyset$
Smoking $\not\perp\!\!\!\perp$ CVD | Yellow Teeth

Pattern DAGs



- You can still "read" all conditional independencies entailed by the Markov Condition in the graph using d-separation.

The PC algorithm

Search strategy:

- Identify the skeleton of your PDAG:
  - Begin with the full graph.
  - For k=0:number of variables -2
  - Using heuristic 3
    - For each pair of adjacent variables X, Y,
      - look within Adjacencies(X)\Y or Adjacencies(Y)\X for a set of k observed variables **Z** such that $X \perp\!\!\!\perp Y | \mathbf{Z}$.
      - If you succeed, remove X-Y.
- Orient all invariant edges of the Markov Equivalence class
  - Apply R0
  - While no more rules are applicable, apply R1-R3

Rules R0-R3 are complete (Meek, 1995)

PC algorithm

Introduced by Peter Spirtes and Clark Glymour in 1993.

One of the first algorithms to perform causal discovery from cross-sectional data.

Uses a complete set of orientation rules and therefore identifies the PDAG that faithfully represents the conditional independencies it identifies.

- The PDAG is maximally informative, in the sense that every un-oriented edge has different orientations in different DAGs in the Markov Equivalence class.

Most current constraint-based algorithms are extensions/improvements of the PC algorithm.

PC Algorithm - Complexity

Suppose that the maximum number of parents for any variable in the graph is k.
Then the worst-case number of tests of conditional independence performed by PC is:

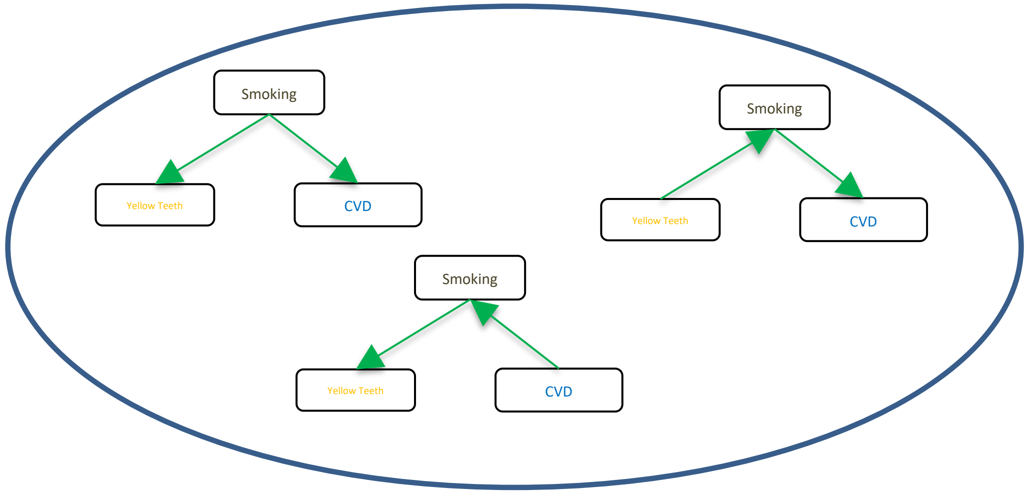
$$2\binom{n}{2}\sum_{i=0}^{k}\binom{n-1}{i}$$

which is bounded by

$$\frac{n^2(n-1)^{k-1}}{(k-1)!}$$

i.e., polynomial to the number of variables, exponential to the maximum number of parents.

Learning causal networks as a model selection problem



Sample (Person)	Smoking	CVD	Yellow Teeth
1	Yes	Yes	No
2	No	No	No
3	Yes	Yes	Yes
4	No	No	Yes
5	Yes	No	No
6	No	Yes	Yes
⋮			
52	No	Yes	No

Identify all DAGs that maximize the posterior probability of the graph given the data: P(G|D) (or some other data-fitting criterion in general)

Posterior probability of the graph

Probability of the data given the graph

Prior probability of the graph

$$P(G|D) = \frac{P(D|G) \times P(G)}{P(D)} = \frac{\boxed{P(D|G)} \times \boxed{P(G)}}{\boxed{P(D)}}$$

Normalization constant

Posterior probability of the graph

Probability of the data given the graph

Prior probability of the graph

$$P(G|D) = \frac{P(D|G) \times P(G)}{P(D)} = \frac{P(D|G) \times P(G)}{P(D)}$$

You can ignore it since it does not depend on the graph structure.

Scoring function

Find G: $argmax_G \, P(D|G) \times P(G)$

Scoring function

Uniform/based on prior knowledge/favoring sparsity

Find G: $argmax_G \, P(D|G) \times P(G)$

Scoring function

Uniform/based on prior knowledge/favoring sparsity

Find G: $argmax_G\, P(D|G) \times P(G)$

Average over all possible parameters (of the joint probability distribution).

$$\int_{\boldsymbol{\theta}} P(D|G,\theta)P(\theta)d\boldsymbol{\theta}$$

Scoring function

Uniform/based on prior knowledge/favoring sparsity

Find G: $argmax_G \, P(D|G) \times P(G)$

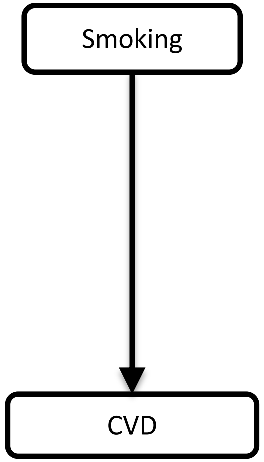
Average over all possible parameters (of the joint probability distribution).

$$\int_{\boldsymbol{\theta}} P(D|G,\theta)P(\theta)d\boldsymbol{\theta} \quad = \int_{\boldsymbol{\theta}} P\left(D\middle|\, \boldsymbol{\theta}_{\boldsymbol{x}|\boldsymbol{pa}(\boldsymbol{x})}\right) f(\boldsymbol{\theta})d\boldsymbol{\theta}$$

The parameterization depends on the graphical structure.

Scoring function

$$P(D|G) = \int_{\boldsymbol{\theta}} P\left(D \middle| \boldsymbol{\theta}_{\boldsymbol{x}|\boldsymbol{pa}(\boldsymbol{x})}\right) f(\boldsymbol{\theta}) d\boldsymbol{\theta} =$$



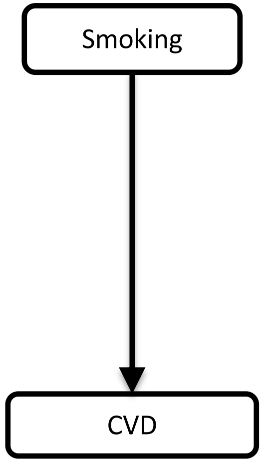
	P(Smoking)
Yes	θ_s
No	$1 - \theta_s$

	P(CVD)	
Smoking	Yes	No
Yes	$\theta_{C\|S}$	$1 - \theta_{C\|S}$
No	$\theta_{C\|NS}$	$1 - \theta_{C\|NS}$

Scoring function

$$P(D|G) = \int_{\boldsymbol{\theta}} P\left(D \middle| \boldsymbol{\theta}_{\boldsymbol{x}|\boldsymbol{pa}(\boldsymbol{x})}\right) f(\boldsymbol{\theta}) d\boldsymbol{\theta} =$$

$$\prod_{x} \int_{\theta_{x|pa(x)}} P\left(D \middle| \theta_{x|pa(x)}\right) f\left(\theta_{x|pa(x)}\right) d\theta_{x|pa(x)}$$



	P(Smoking)
Yes	θ_s
No	$1 - \theta_s$

	P(CVD)	
Smoking	Yes	No
Yes	$\theta_{C|S}$	$1 - \theta_{C|S}$
No	$\theta_{C|NS}$	$1 - \theta_{C|NS}$

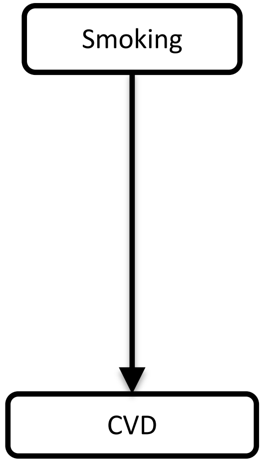
- Score is decomposable:
- It is a product of terms involving only a variable and its parents.

Scoring function

$$P(D|G) = \int_{\boldsymbol{\theta}} P\left(D\middle|G, \boldsymbol{\theta}_{\boldsymbol{x}|\boldsymbol{pa}(\boldsymbol{x})}\right) f(\boldsymbol{\theta})d\boldsymbol{\theta} =$$

$$\prod_{x} \int_{\theta_{x|pa(x)}} P\left(D\middle|G, \theta_{x|pa(x)}\right) f\left(\theta_{x|pa(x)}\right)d\theta_{x|pa(x)}$$

$$\int_{\theta_s} P(D|\,\theta_s)\ f(\theta_s)d\theta_s \int_{\theta_{c|ns}} P\left(D\middle|\,\theta_{c|s}\right) f\left(\theta_{c|s}\right)d\theta_{c|s} \int_{\theta_{c|ns}} P\left(D\middle|\,\theta_{c|ns}\right) f\left(\theta_{c|ns}\right)d\theta_{c|ns}$$



	P(Smoking)
Yes	θ_s
No	$1 - \theta_s$

	P(CVD)	
Smoking	Yes	No
Yes	$\theta_{C|S}$	$1 - \theta_{C|S}$
No	$\theta_{C|NS}$	$1 - \theta_{C|NS}$

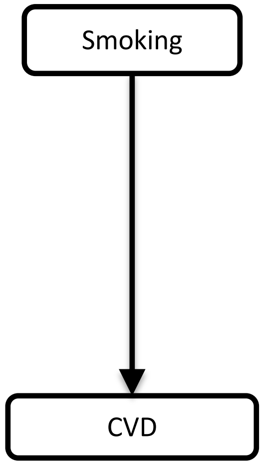
Scoring function

$$P(D|G) = \int_{\boldsymbol{\theta}} P\left(D\middle|G,\boldsymbol{\theta}_{\boldsymbol{x}|\boldsymbol{pa}(\boldsymbol{x})}\right) f(\boldsymbol{\theta})d\boldsymbol{\theta} =$$

$$\prod_{x} \int_{\theta_{x|pa(x)}} P\left(D\middle|G,\theta_{x|pa(x)}\right) f\left(\theta_{x|pa(x)}\right)d\theta_{x|pa(x)}$$

$$\int_{\theta_s} P(D|\,\theta_s)\; f(\theta_s)d\theta_s \int_{\theta_{c|ns}} P\left(D\middle|\,\theta_{c|s}\right)\; f\left(\theta_{c|s}\right)d\theta_{c|s} \int_{\theta_{c|ns}} P\left(D\middle|\,\theta_{c|ns}\right)\; f\left(\theta_{c|ns}\right)d\theta_{c|ns}$$

This score is a marginal likelihood, and can be computed in closed form for some families of distributions that have conjugate priors



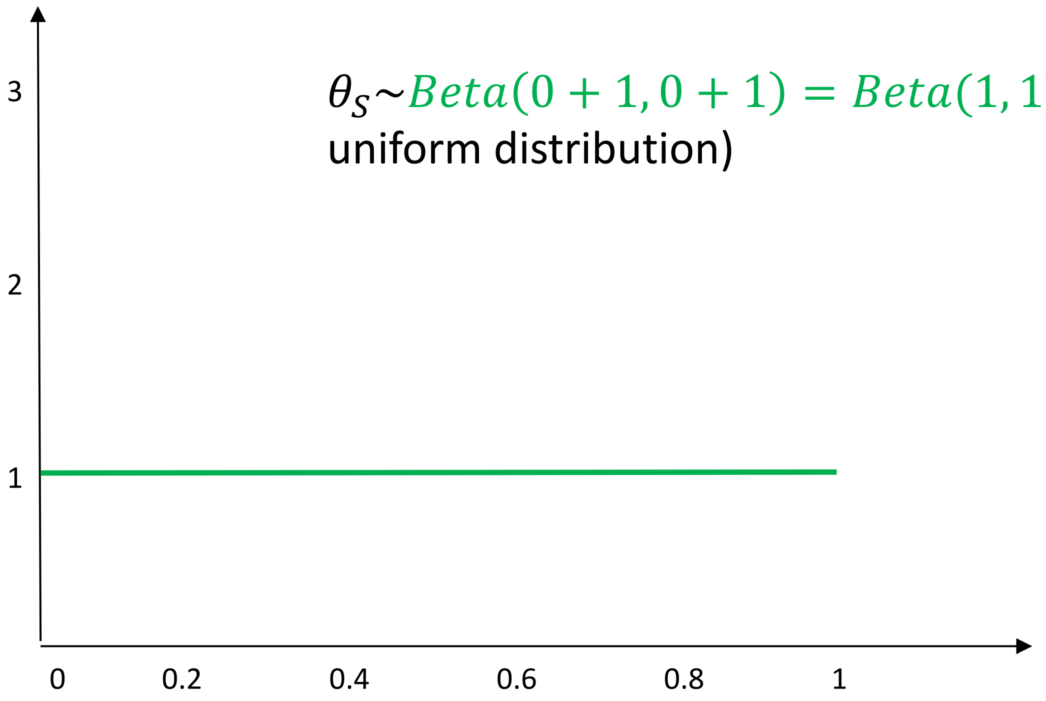
	P(Smoking)
Yes	θ_s
No	$1-\theta_s$

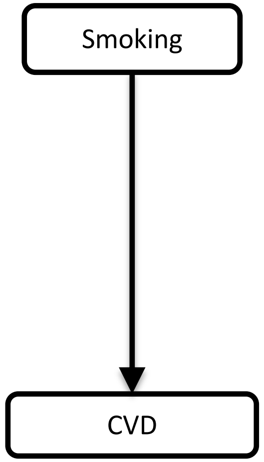
	P(CVD)	
Smoking	Yes	No
Yes	$\theta_{C|S}$	$1-\ \theta_{C|S}$
No	$\theta_{C|NS}$	$1-\theta_{C|NS}$

Scoring function

You have observed 0 smokers and 0 non smokers. (Prior)

$\theta_S \sim Beta(0+1, 0+1) = Beta(1,1)$ (the uniform distribution)





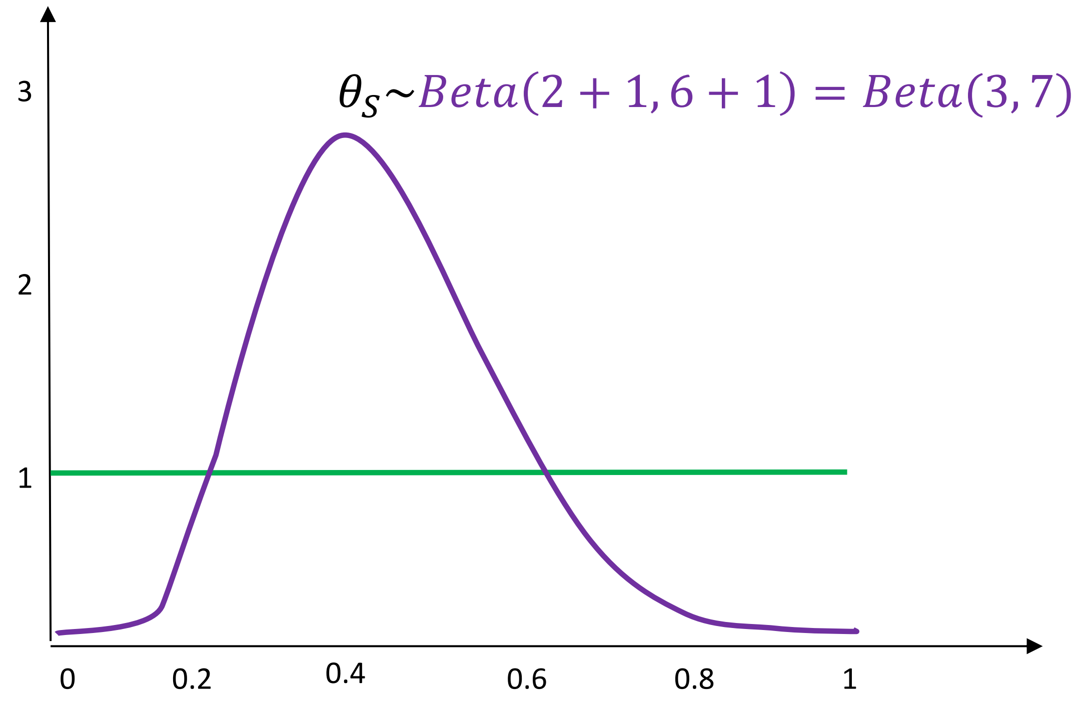
	P(Smoking)
Yes	θ_s
No	$1-\theta_s$

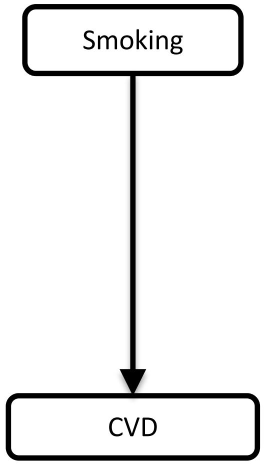
	P(CVD)	
Smoking	Yes	No
Yes	$\theta_{C\|S}$	$1-\ \theta_{C\|S}$
No	$\theta_{C\|NS}$	$1-\theta_{C\|NS}$

Reminder: Bayesian Statistics.

Scoring function

You then observe 2 smokers and 6 non-smokers. Bayesian Update :





	P(Smoking)
Yes	θ_s
No	$1-\theta_s$

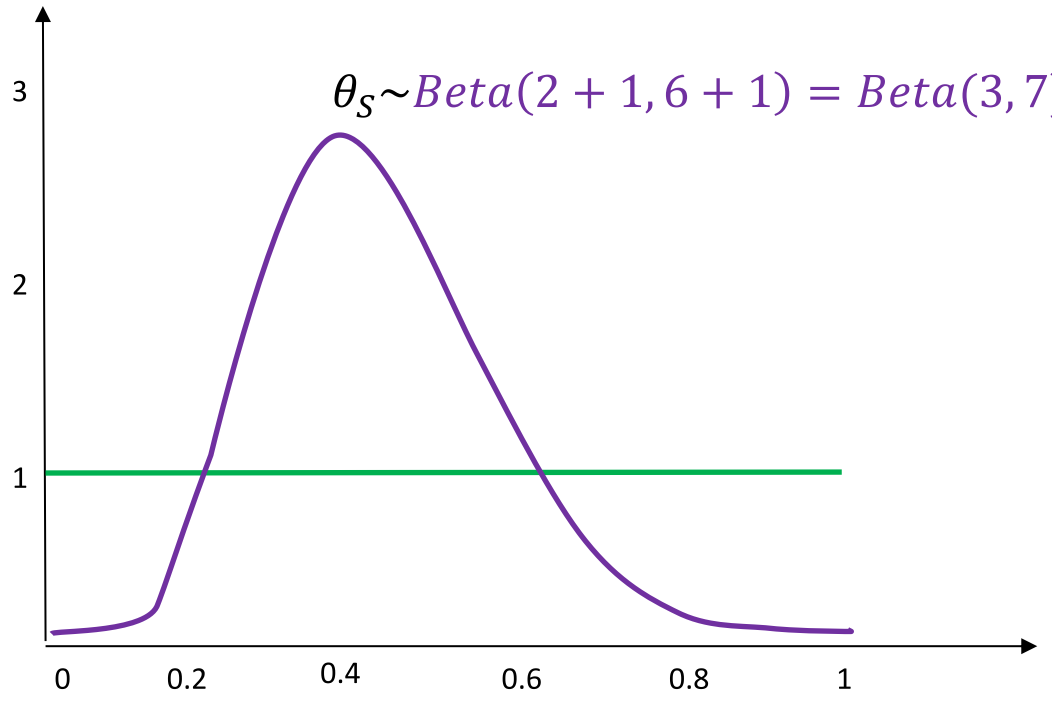
	P(CVD)	
Smoking	Yes	No
Yes	$\theta_{C\|S}$	$1-\theta_{C\|S}$
No	$\theta_{C\|NS}$	$1-\theta_{C\|NS}$

Reminder: Bayesian Statistics.

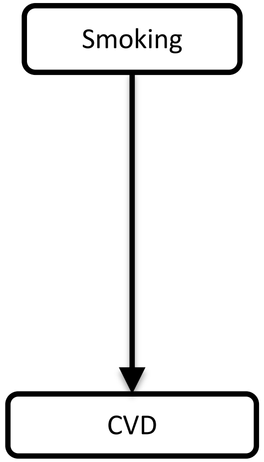
Scoring function

You then observe 2 smokers and 6 non-smokers. Bayesian Update:

$\theta_S \sim Beta(2+1, 6+1) = Beta(3,7)$



You now believe that the proportion of smokers to non smokers is close to 3:7



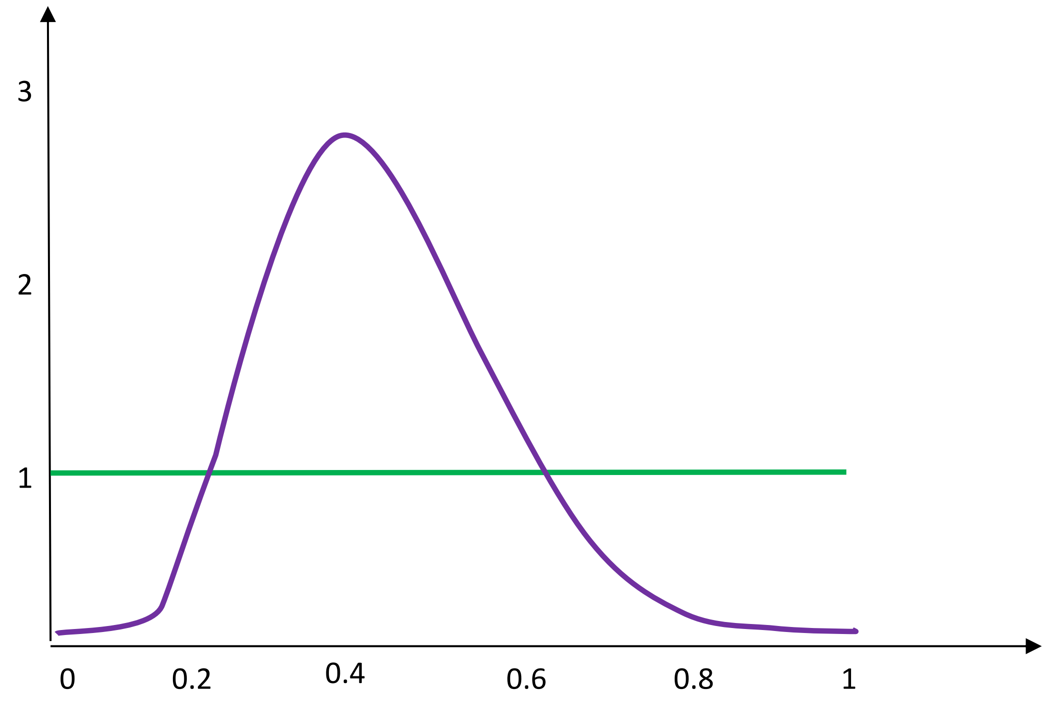
	P(Smoking)
Yes	θ_s
No	$1-\theta_s$

	P(CVD)	
Smoking	Yes	No
Yes	$\theta_{C\|S}$	$1-\theta_{C\|S}$
No	$\theta_{C\|NS}$	$1-\theta_{C\|NS}$

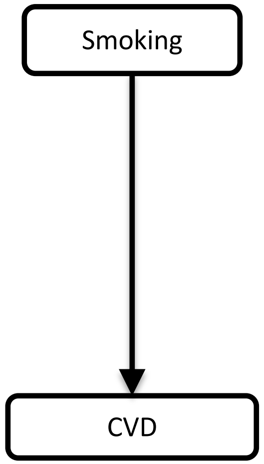
Bayesian Statistics.

Scoring function

You then observe 2 smokers and 6 non-smokers. Posterior:



You now believe that the proportion of smokers to non smokers is close to 3:7



	P(Smoking)
Yes	θ_s
No	$1 - \theta_s$

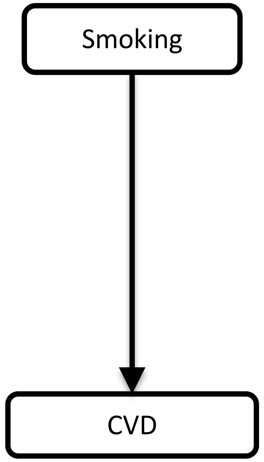
	P(CVD)	
Smoking	Yes	No
Yes	$\theta_{C\|S}$	$1 - \theta_{C\|S}$
No	$\theta_{C\|NS}$	$1 - \theta_{C\|NS}$

Bayesian Statistics.

Scoring function

$$\int_{\theta_S} P(D|\theta_s)\ f(\theta_s)d\theta_s = \int_{\theta_S} \prod_i (X_i|\theta_s)\, f(\theta_s)d\theta_s =$$

$$\frac{\Gamma(2)\Gamma(6)}{\Gamma(8)} = 0.0238$$



	P(Smoking)
Yes	θ_s
No	$1 - \theta_s$

	P(CVD)	
Smoking	Yes	No
Yes	$\theta_{C\|S}$	$1 - \theta_{C\|S}$
No	$\theta_{C\|NS}$	$1 - \theta_{C\|NS}$

Computed in closed form!

Example Search Strategy (Greedy Search)

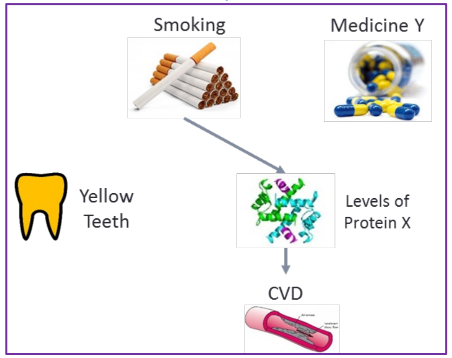
Initialize G as the empty/full/random graph and score.

Score all networks that can be produced by G with a single change: adding/removing/reversing an edge, ensuring G remains a DAG (no cycles).

Keep the change that resulted in the highest-scoring network.

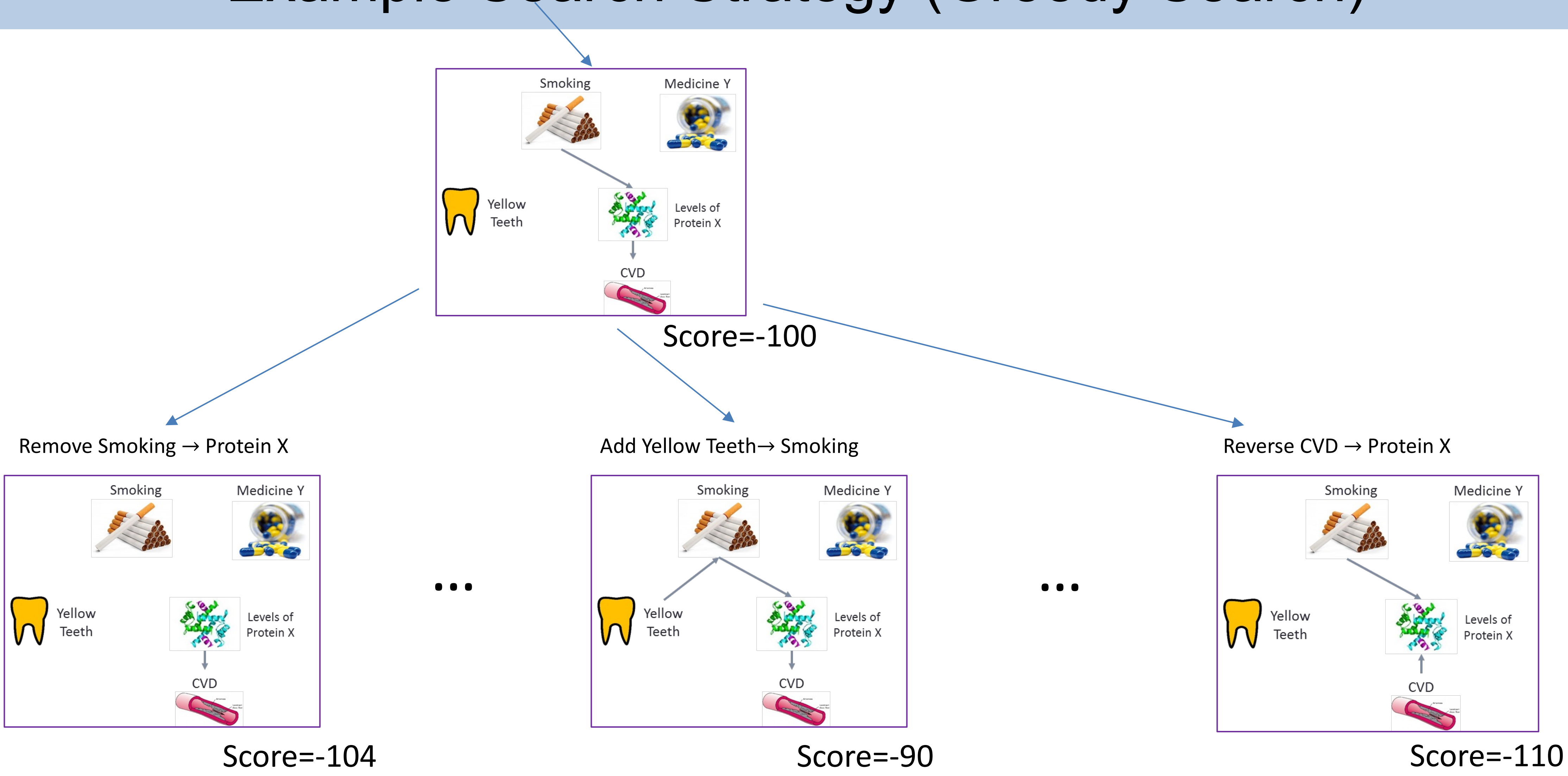
Until no single action improves the score.

Example Search Strategy (Greedy Search)

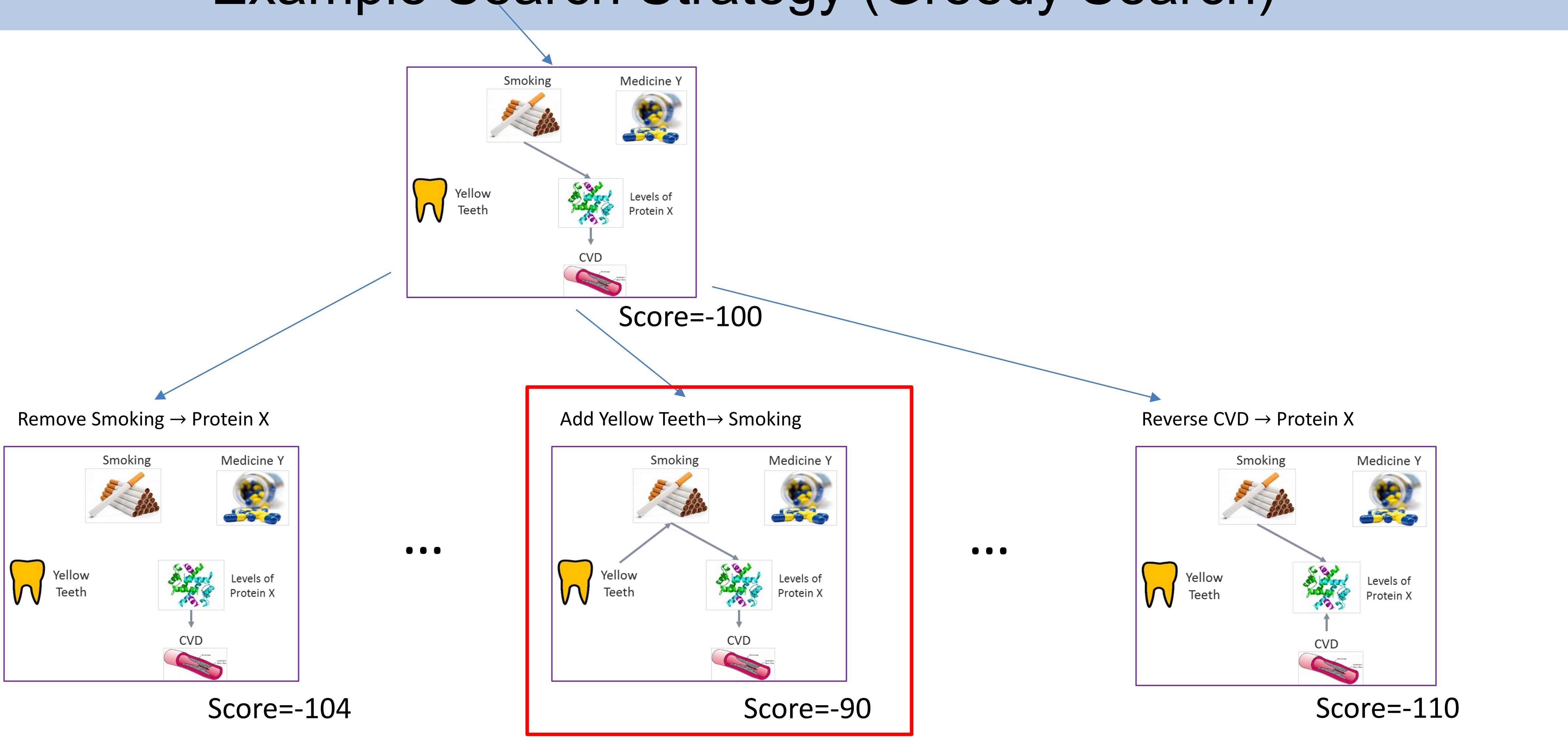


Score=-100

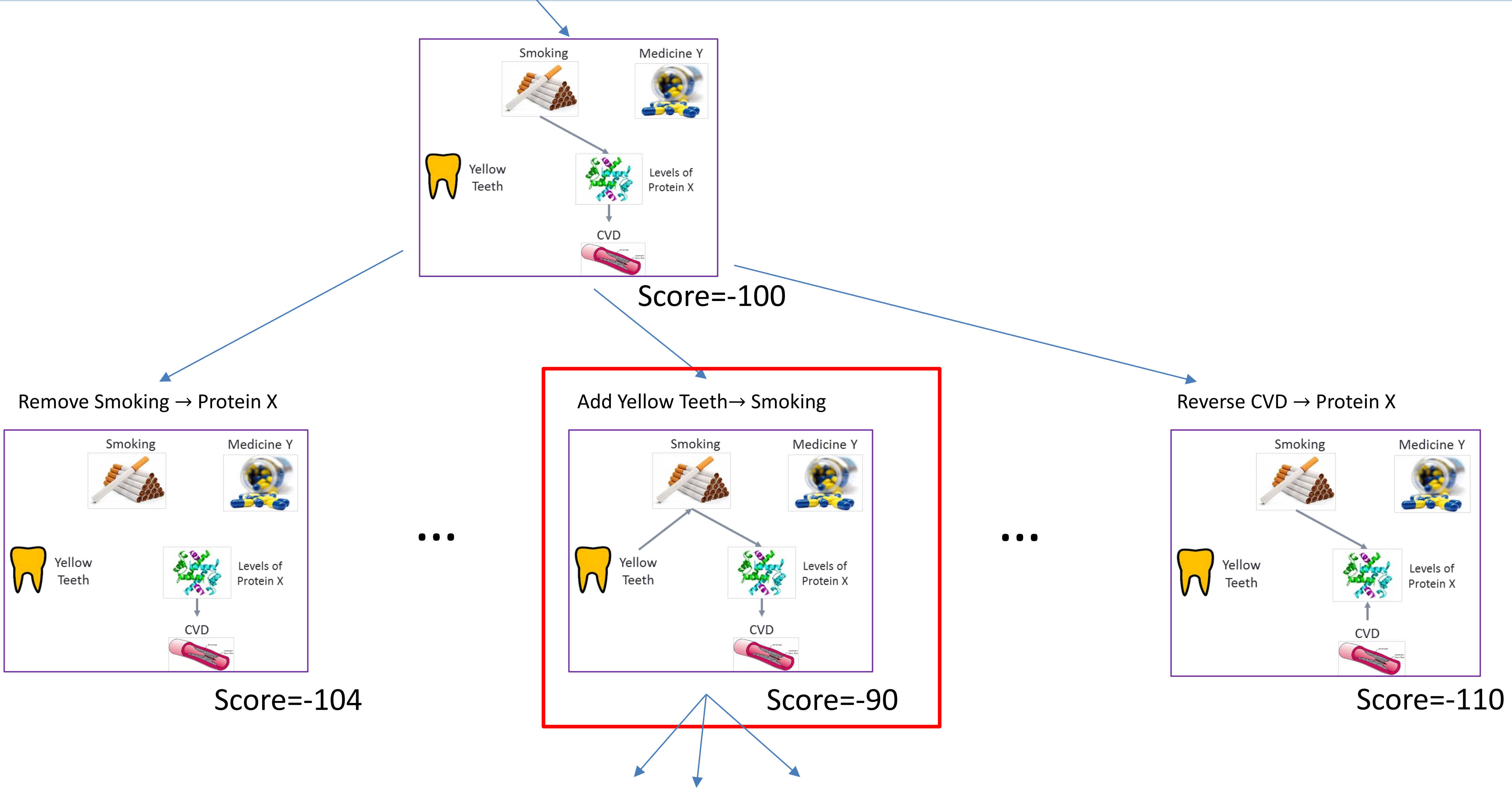
Example Search Strategy (Greedy Search)



Example Search Strategy (Greedy Search)



Example Search Strategy (Greedy Search)



Search-and-Score CBN learning

Other search strategies are possible.

- e.g. BFS, DFS, Genetic algorithms, TABU search.

You can search in the space of PDAGs.

- e.g. GES algorithm, (Chickering, 1996)

You may get stuck in local minima.

- Avoid by random restarts, simulated annealing, stochastic greedy search.

Exact methods exist for actually scoring all possible networks (e.g. Koivisto and Sood, 2004)

- Using dynamic programming & bounded number of parents per variable.
- $O(n2^n)$ space + time complexity, not possible for more than ~20-40 variables.

Comparison

Constraint-Based

Easier to extend to different types of data (e.g., censored).

Easier to extend to networks with latent variables (next time).

More efficient in learning the skeleton of the network.

Search-and-score

Robust to small samples.

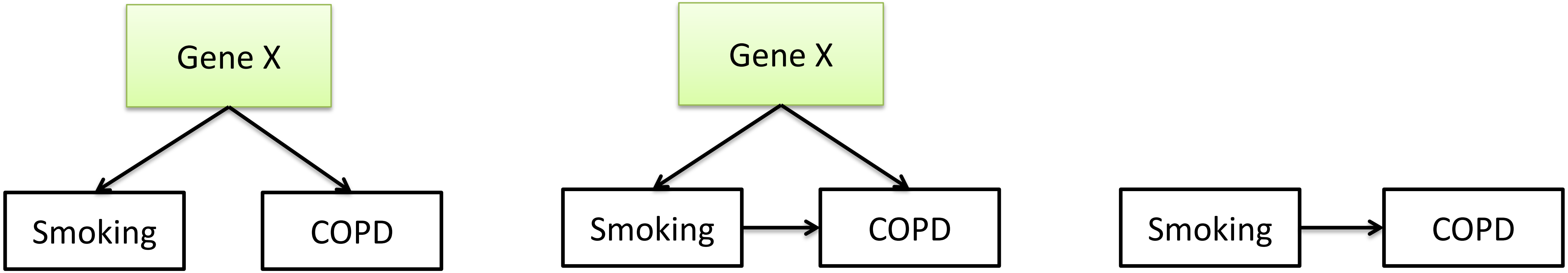
Easier to incorporate priors on the networks.

Better in identifying the edge orientations.

Exact methods also exist, limited to ~20-40 variables.

Modelling causality

Association is not Causality



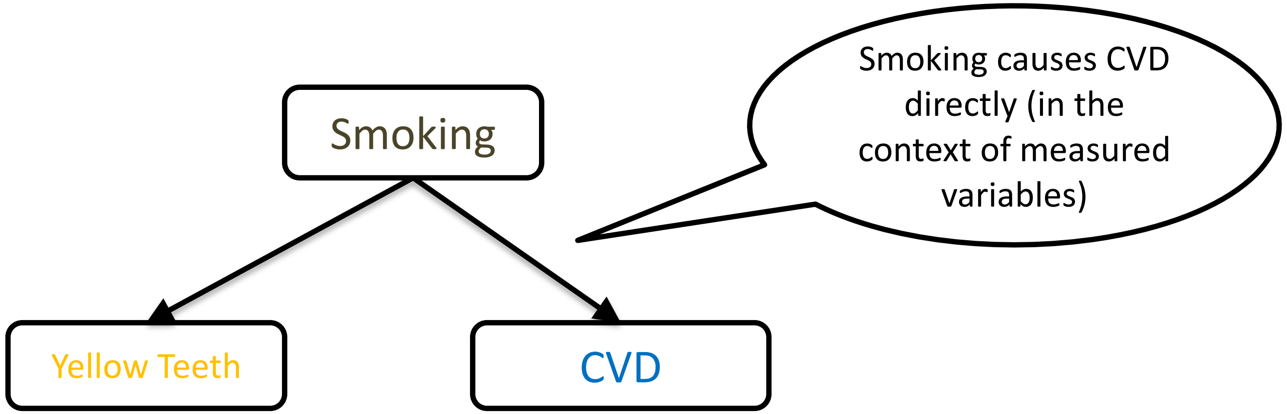
Three models, all imply Smoking and COPD are dependent

$P(COPD|Smoking) \neq P(COPD)$

In model 1, changing smoking habits does not affect the probability of getting COPD

$P(COPD|do(Smoking)) = P(COPD)$

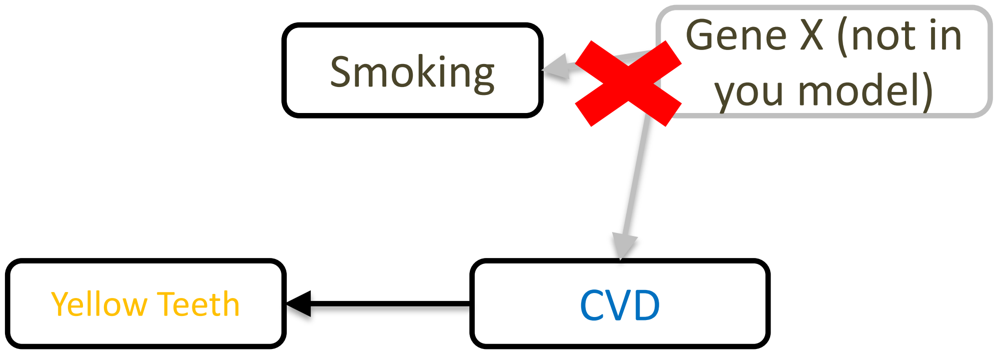
Modeling causality



Recipe for creating a causal graph:

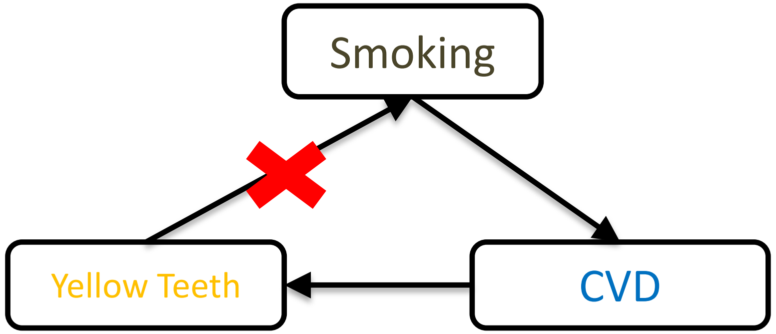
- Model variables as graph nodes.
- Add directed edges corresponding to <u>direct causation</u>.

Modeling Causality



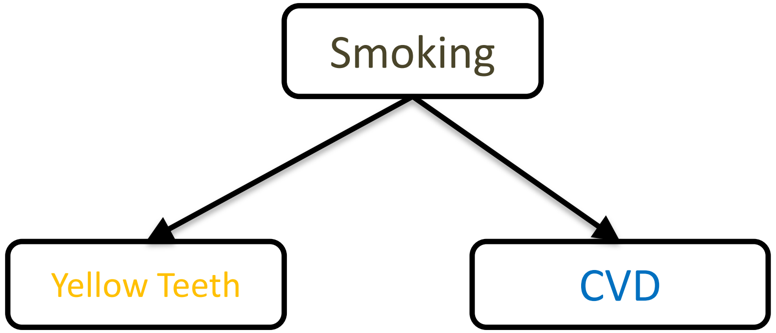
For this lecture:

- No hidden common causes (causal sufficiency).
- No causal feedback.
- Causal structure is described by a Directed Acyclic Graph (DAG).



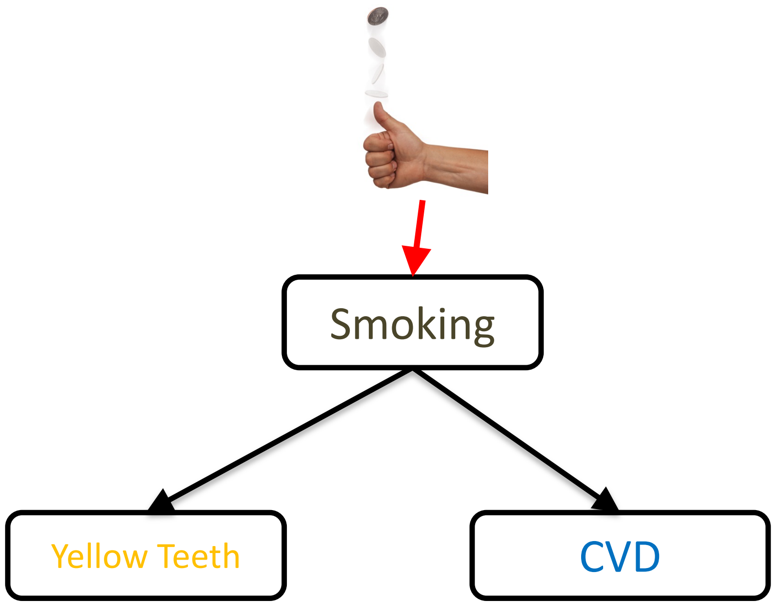
Not allowed (yet)

Modeling Interventions



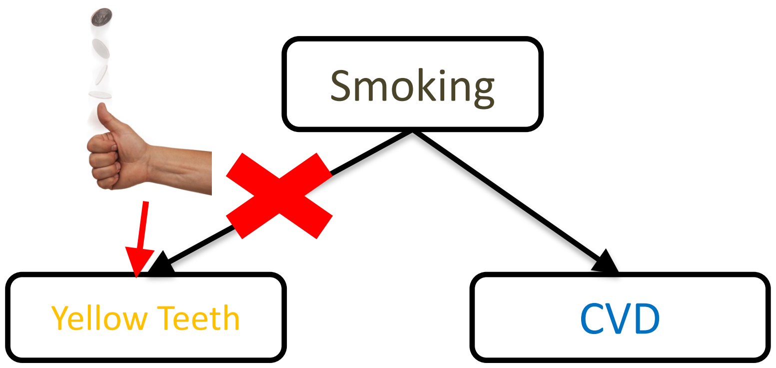
- Ideal Interventions: You completely set the value/distribution of a variable.
  - e.g. assign to treatment/placebo group
  - The type of intervention you would typically like to do, not always possible.
- fat-hand interventions affect more than one variable at a time.
  - sometimes a result of bad experimental design.

Modeling Interventions



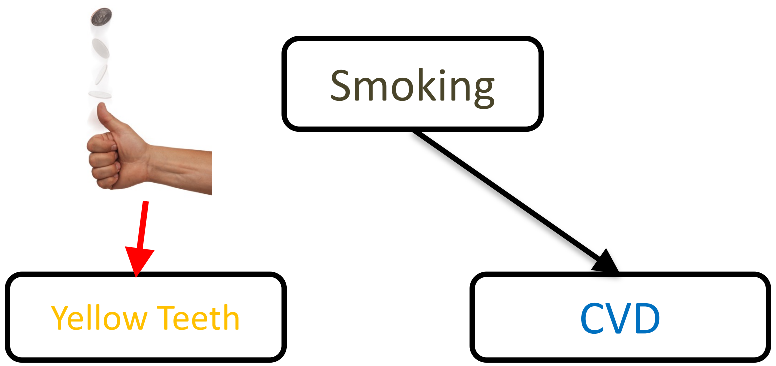
- Intervening on the cause:
  - You force half your sample to smoke, ban the rest from smoking.
  - More smokers than non-smokers have yellow teeth.

Modeling Interventions



- Intervening on the effect:
  - You stain half your sample's teeth yellow, you whiten the teeth of the rest.
  - Smokers do not have yellow teeth more than non-smokers.

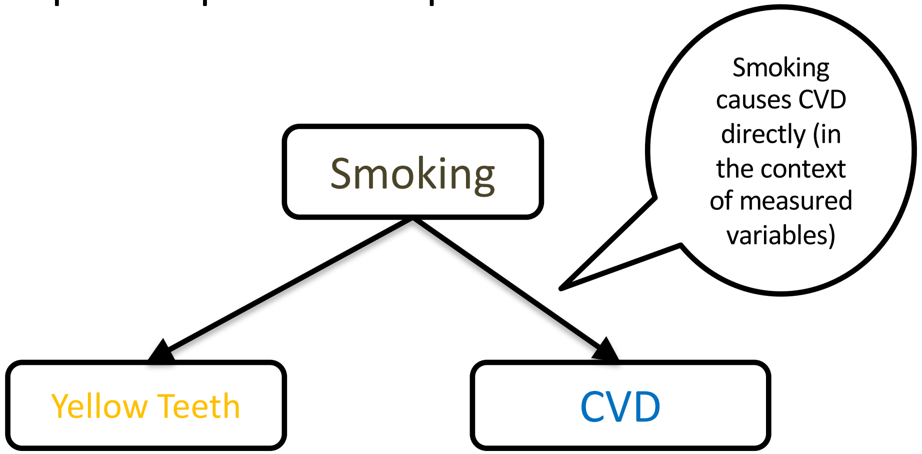
Modeling Interventions



- Graph Surgery/do operator removes all edges that are incoming to the manipulated variable.
- Causal relationships are now described by the **manipulated graph.**

Modeling probabilistic causality

Graph G captures the qualitative causal relations



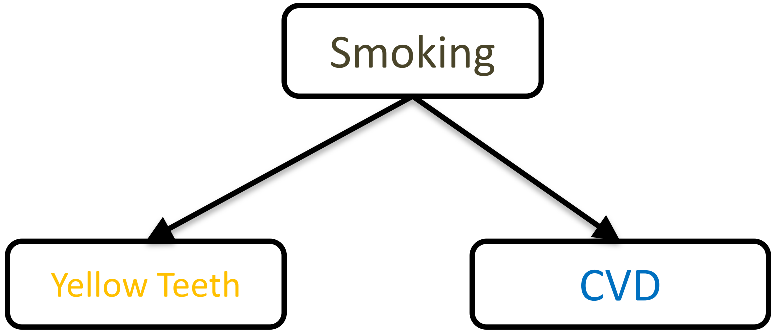
What connects the two?

JPD J encodes the quantitative probabilistic properties

		CVD		
Yellow Teeth	Smoking	Y	N	
Y	Y	0.17	0.06	0.13
N	Y	0.06	0.02	0.08
Y	N	0.02	0.06	0.08
N	N	0.15	0.46	0.61
		0.4	0.6	1

Causal Markov Condition

Graph G captures the qualitative causal relations



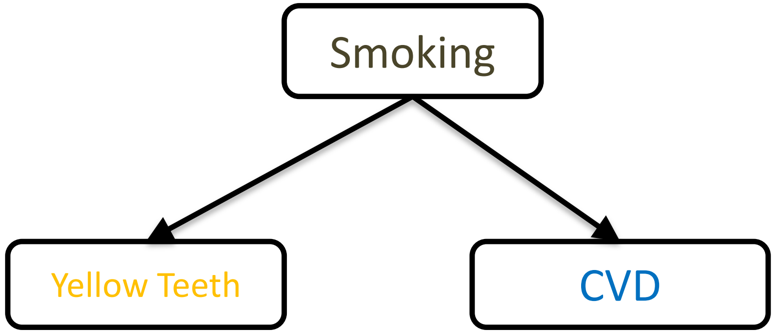
JPD J encodes the quantitative probabilistic properties

		CVD		
Yellow Teeth	Smoking	Y	N	
Y	Y	0.17	0.06	0.13
N	Y	0.06	0.02	0.08
Y	N	0.02	0.06	0.08
N	N	0.15	0.46	0.61
		0.4	0.6	1

Causal Markov Condition (CMC):
Every variable is **independent** of its **non-effects** (non-descendants in the graph) given its **direct causes** (parents).

Causal Markov Condition

Graph G captures the qualitative causal relations



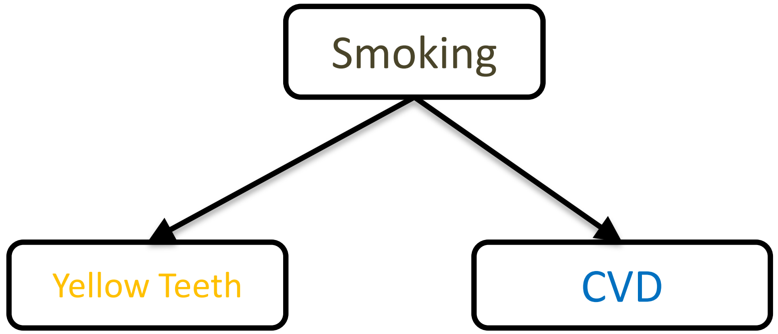
JPD J encodes the quantitative probabilistic properties

		CVD		
Yellow Teeth	Smoking	Y	N	
Y	Y	0.17	0.06	0.13
N	Y	0.06	0.02	0.08
Y	N	0.02	0.06	0.08
N	N	0.15	0.46	0.61
		0.4	0.6	1

Causal Markov Condition (CMC):
Every variable is **independent** of its **non-effects** (non-descendants in the graph) given its **direct causes** (parents).

Learning the value of **intermediate** and **common** causes renders variables **independent.**

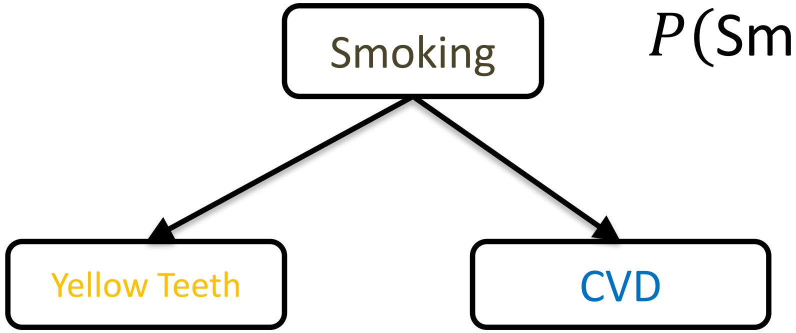
Factorization with the CMC



P(Smoking, Yellow Teeth, CVD) =
P(Smoking) ×
P(Yellow Teeth | Smoking) ×
P(CVD | Smoking)

In general: $$P(\mathrm{V}) = \prod_i P(V_i | \text{Parents of } V_i \text{ in the graph})$$

Causal Bayesian Network



$P(\text{Smoking}) = 0.3$

$P(\text{Yellow Teeth}|\text{Smoking}) = 0.85$

$P(\text{Yellow Teeth}|\neg\text{Smoking}) = 0.1$

$P(\text{CVD}|\text{Smoking}) = 0.75$

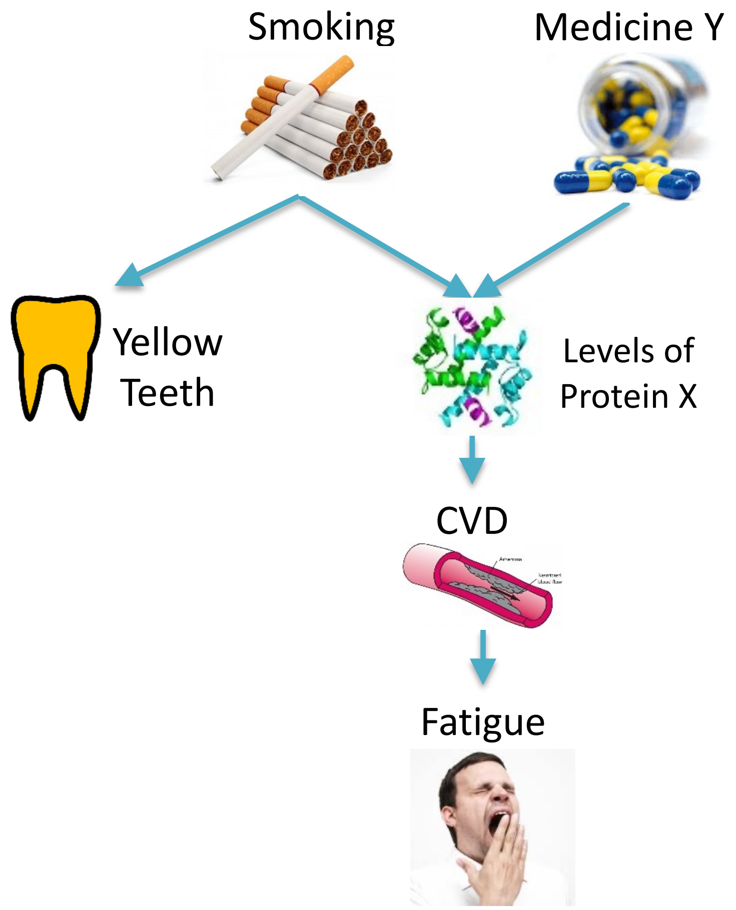
$P(\text{CVD}|\neg\text{Smoking}) = 0.15$

Causal DAG and conditional probability tables define a Causal Bayesian Network

Things you can do with a Causal Bayesian Network

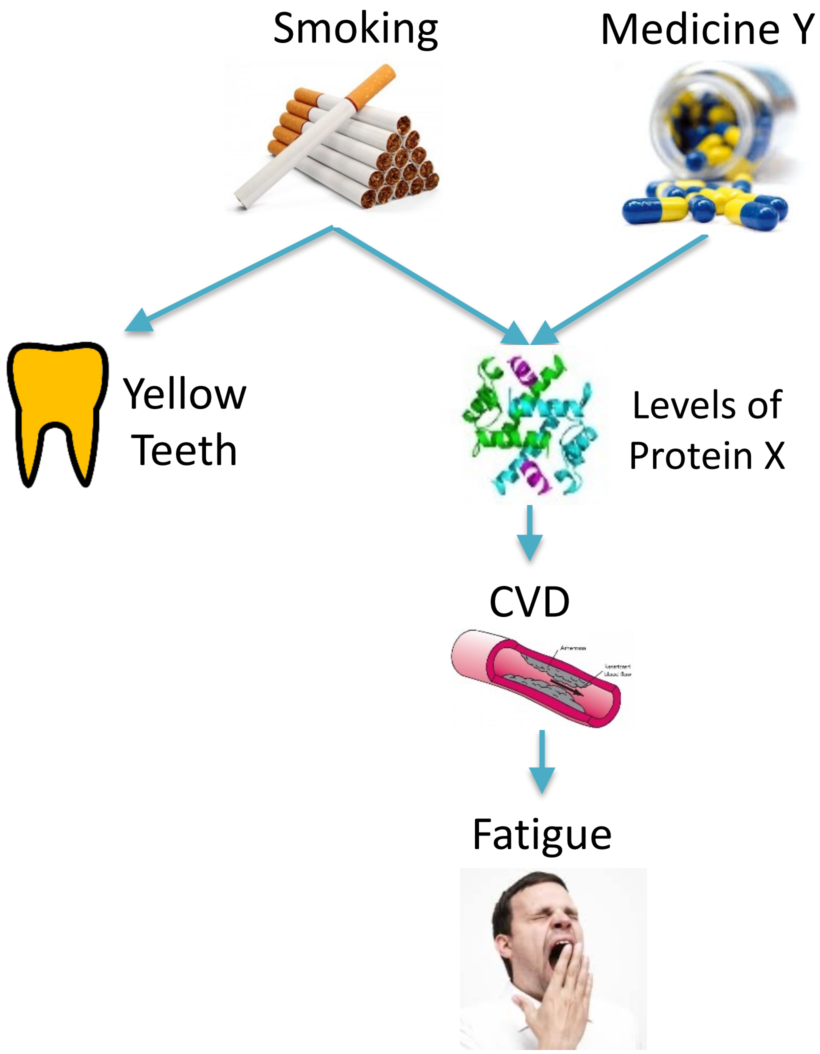
1. Factorize the joint probability distribution.

2. Answer questions like:
   1. Is Smoking independent from Fatigue given Levels of Protein X?
      - Smoking ⫫ Fatigue | Levels of Protein X?
   2. What is the probability of getting CVD if I have high levels of Protein X?
      - P(CVD | Levels of Protein X=high) = ?
   3. Will I reduce the probability of getting CVD if I design a drug that lowers the levels of protein X?
      - P(CVD | do(Levels of Protein X=low))?



CMC and conditional independencies

Every variable is independent of its **non-effects** given its **direct causes.**



The d-separation criterion

Algorithm to determine all independencies entailed by the Causal Markov Condition.

Paths in the graph represent information flow (or lack thereof)

The d-separation criterion

Open (d-connecting) paths :
A path is d-connecting given **Z** iff
every collider on the path is in **Z** or has a descendant in **Z**
AND
every non-collider on the path is not in **Z.**

Otherwise, the path is blocked (d-separating).

The same path can be
d-connecting given $\boldsymbol{Z_1}$,
d-separating given $\boldsymbol{Z_2}$

The d-separation criterion

Algorithm to determine all independencies that are entailed by the CMC.

Conditional independencies in the joint distribution can be decided based on the absence of open paths in the graph:

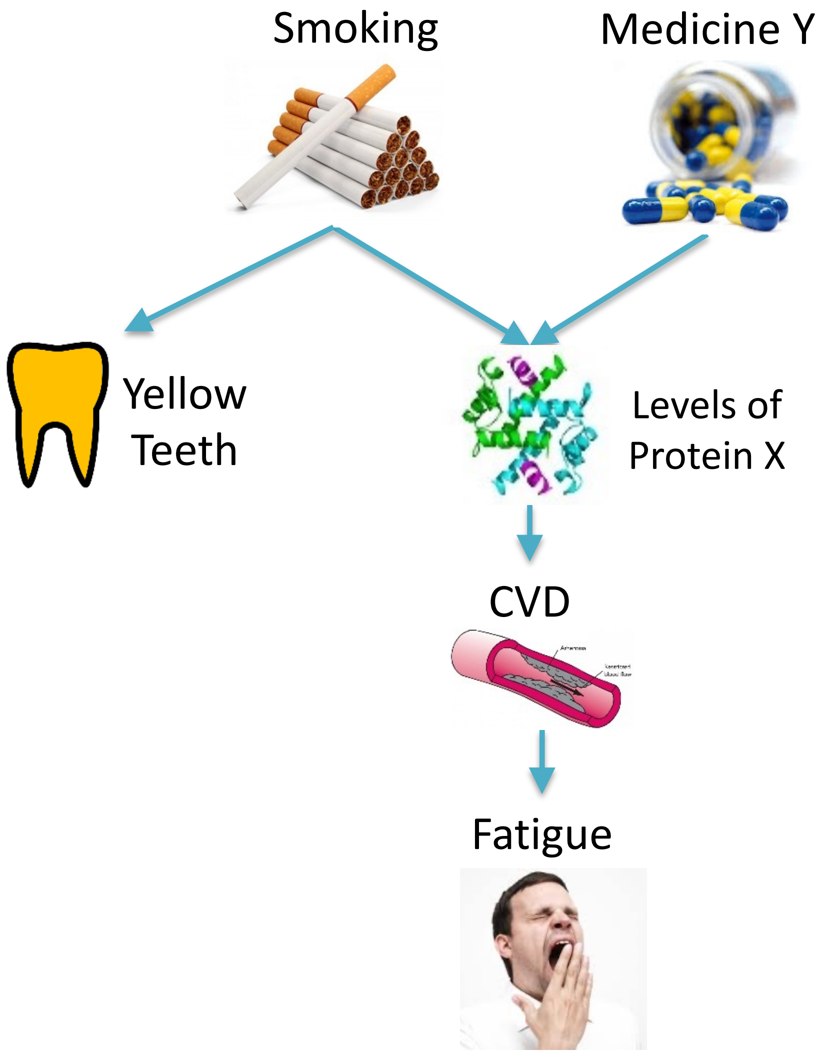
Open paths are called d-connecting paths (given a set of variables).
If no open path exists, the endpoints are d-separated (given the set of variables).
Otherwise, the endpoints are d-connected (given the set of variables)

Notation: $dsep(A, B|\mathbf{Z})$: A and B are d-separated given $\mathbf{Z}$.
$dcon(A, B|\mathbf{Z})$: A and B are d-connected given $\mathbf{Z}$.

The d-separation criterion

To find if $dsep(X, Y|\mathbf{Z})$ in the graph:

1. Find the paths from X to Y (ignoring orientations).
2. If there exists no open path given **Z**, then $dsep(X, Y|\mathbf{Z})$.

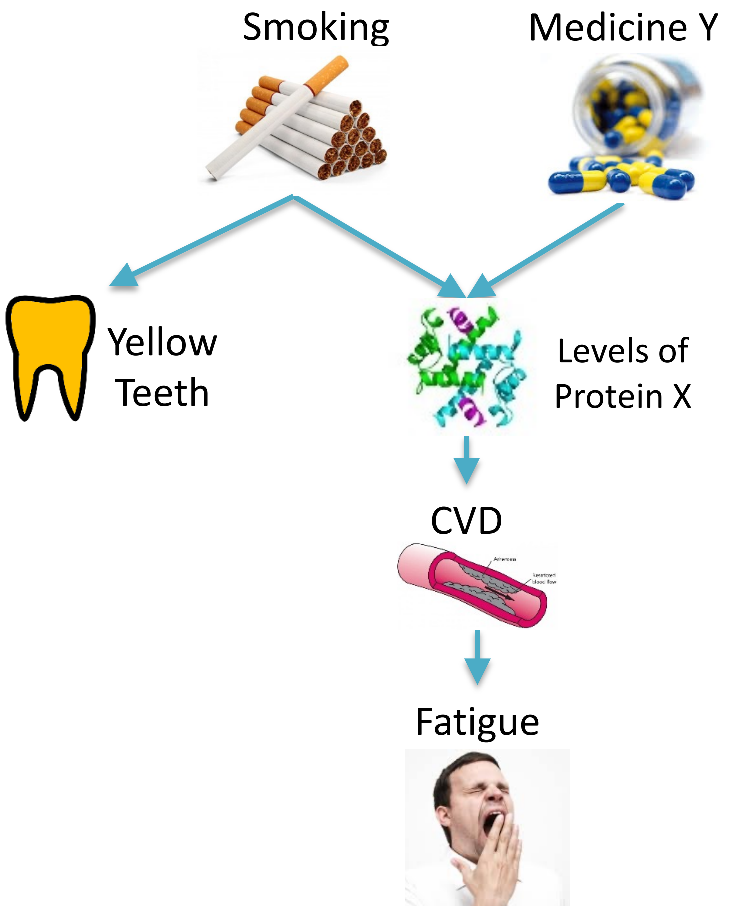


In CBNs:
$dsep(X, Y|\mathbf{Z})$ in $G \Rightarrow X \perp\!\!\!\perp Y|\mathbf{Z}$ in P

Things you can do with a Causal Bayesian Network

1. Factorize the joint probability distribution.

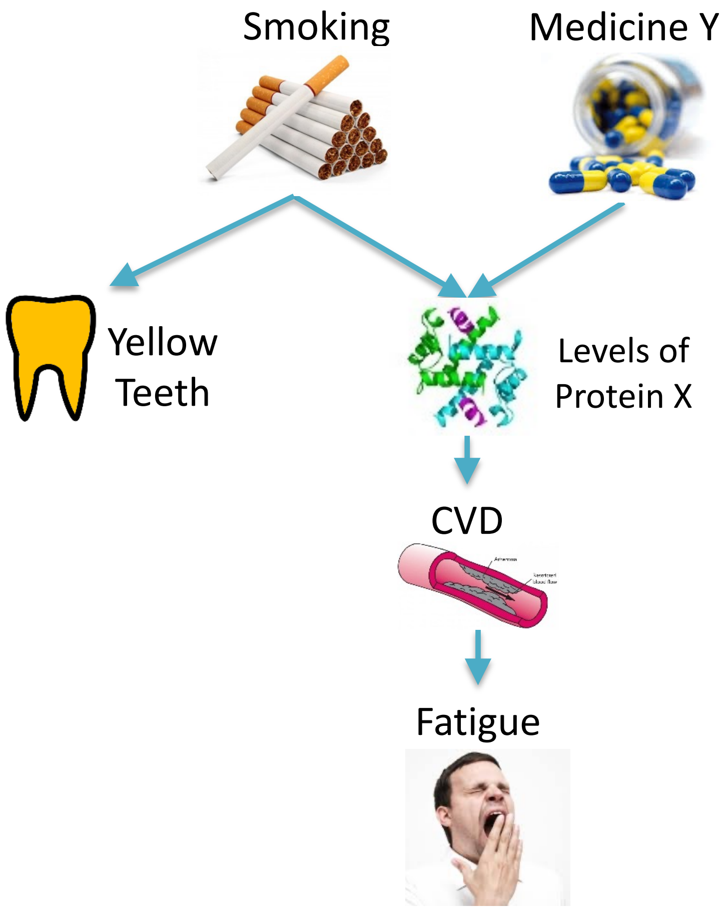
2. Answer questions like:
   1. Is Smoking independent from Fatigue given Levels of Protein X?
      - Smoking $\perp\!\!\!\perp$ Fatigue | Levels of Protein X?
   2. What is the probability of getting CVD if I have high levels of Protein X?
      - P(CVD| Levels of Protein X=high) = ?
   3. Will I reduce the probability of getting CVD if I design a drug that lowers the levels of protein X?
      - P(CVD|do(Levels of Protein X=low))?



Probabilistic Inference: Easy

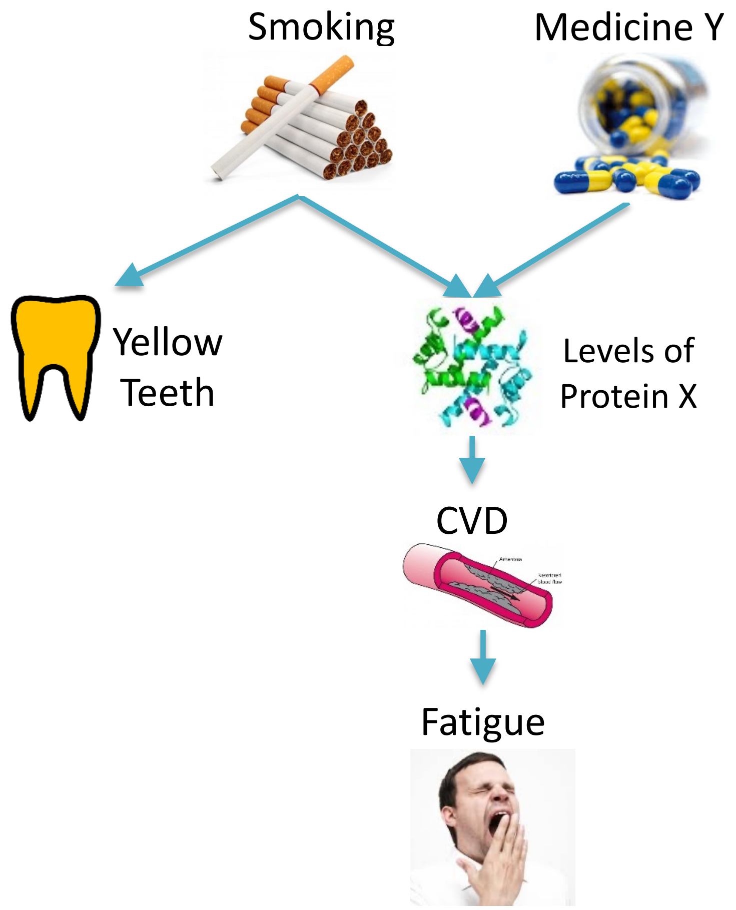
You measure all covariates for a patient.
(smoking, medicine y, yellow teeth, protein x)
What is the probability they have CVD?

P(CVD| Levels of Protein X=high, Smoking=yes, Medicine Y = no, Yellow Teeth = yes) = ?



Probabilistic Inference: Easy

You measure all covariates for a patient.
(smoking, medicine y, yellow teeth, protein x)
What is the probability they have CVD?

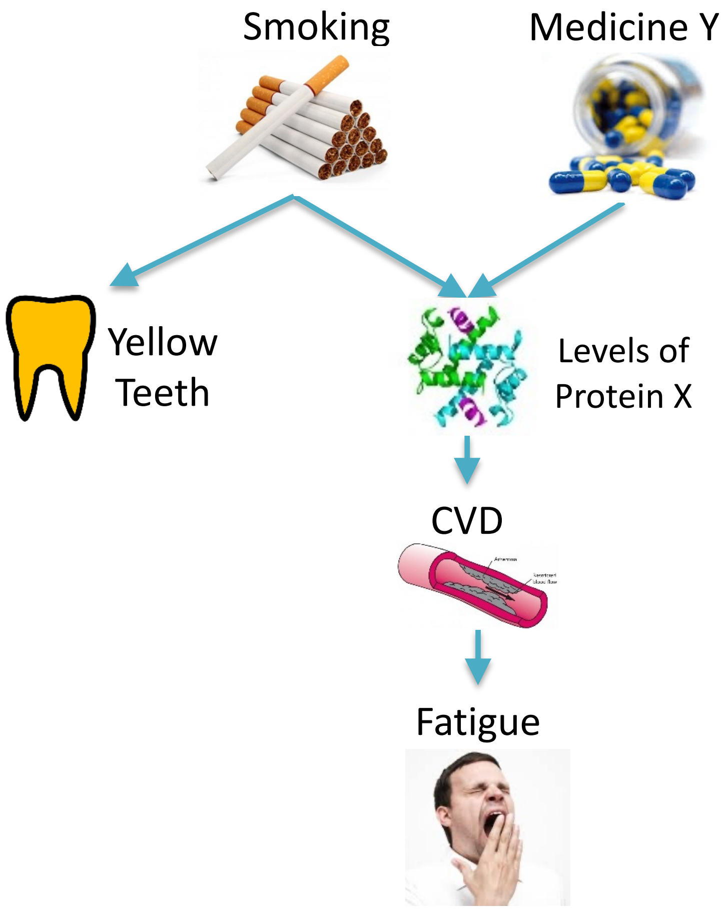


P(CVD| Levels of Protein X=high, Smoking=yes, Medicine Y = no, Yellow Teeth = yes) =

P(CVD|Levels of Protein X)

Probabilistic Inference: Easy

You measure all covariates for a patient.
(smoking, medicine y, yellow teeth, protein x)
What is the probability they have CVD?



P(CVD| Levels of Protein X=high, Smoking=yes, Medicine Y = no, Yellow Teeth = yes) =

P(CVD|Levels of Protein X)

Probabilistic Inference: hard

In general, probabilistic inference is NP-hard.

Exact algorithms can have better average-case performance, particularly for distributions where the integrals can be computed in closed form.

E.g., junction tree, belief propagation

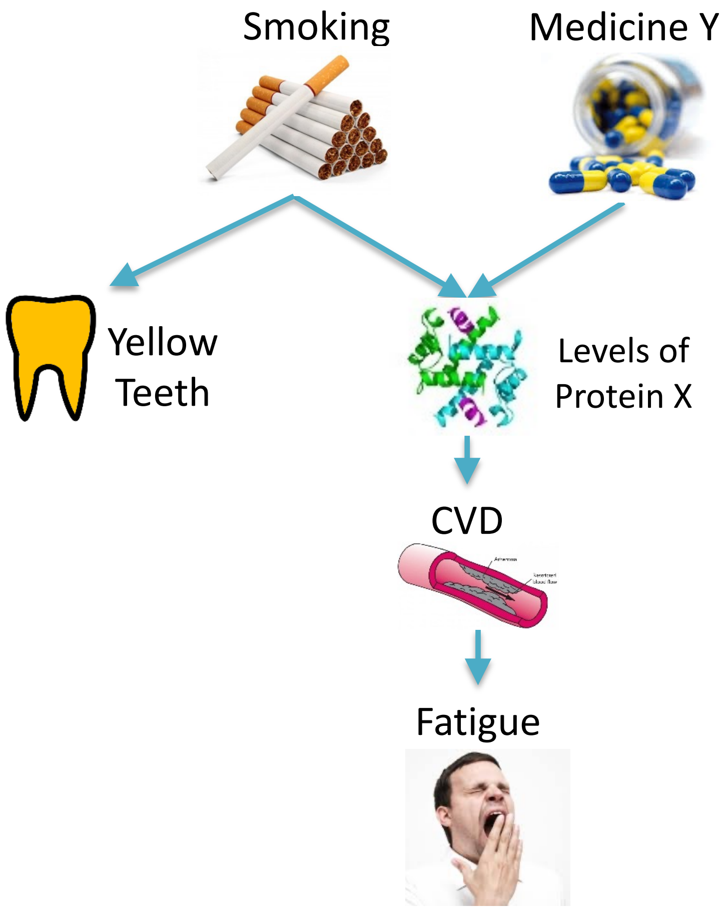
Otherwise, approximate inference using Sampling/MCMC

Probabilistic Inference: Hard

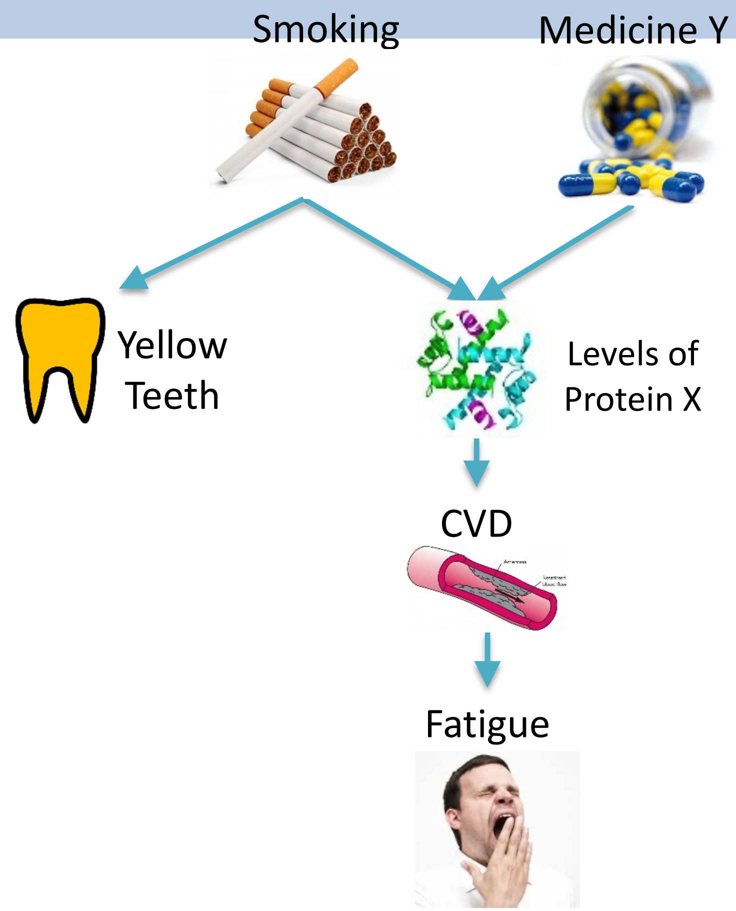
You do not have measurements for protein X, you only know that a patient smokes and does not take medicine Y.

What is the probability they have CVD?

P(CVD| Smoking=yes, Medicine Y = no) =



Probabilistic Inference: Hard



You do not have measurements for protein X, you only know that a patient smokes and does not take medicine Y.

What is the probability they have CVD?

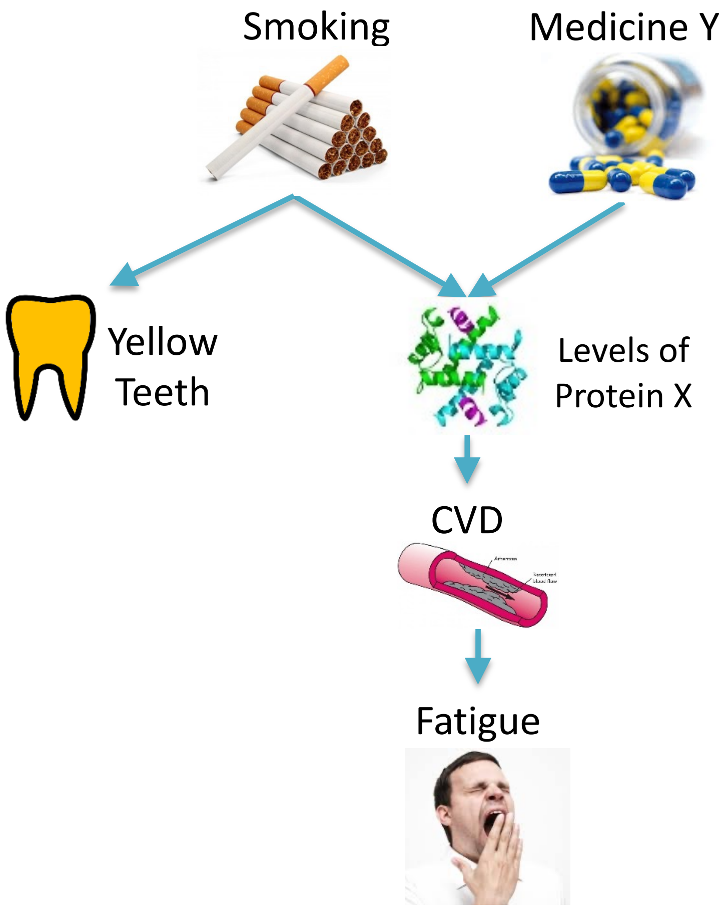
P(CVD| Smoking=yes, Medicine Y = no) =

$\sum_{protein\ X}$ P(CVD|Smoking=yes, Medicine Y=no, Protein X)P(ProteinX|Smoking= Yes, Medicine Y=no) =
$\sum_{protein\ X}$ P(CVD|Protein X)P(ProteinX|Smoking= Yes, Medicine Y=no) =

Things you can do with a Causal Bayesian Network

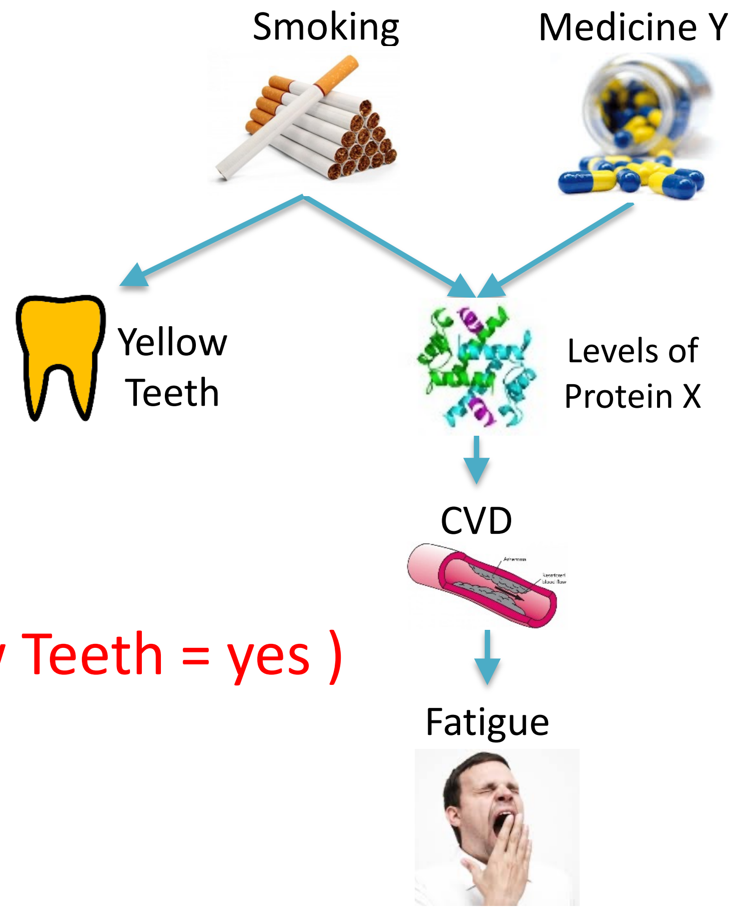
1. Factorize the joint probability distribution.

2. Answer questions like:
   1. Is Smoking independent from Fatigue given Levels of Protein X?
      - Smoking ⫫ Fatigue | Levels of Protein X?
   2. What is the probability of getting CVD if I have high levels of Protein X?
      - P(CVD | Levels of Protein X=high) = ?
   3. Will I reduce the probability of getting CVD if I design a drug that lowers the levels of protein X?
      - P(CVD | do(Levels of Protein X=low))?



Causal Inference

You measure some covariates for a patient. (medicine y, yellow teeth)
What is the probability they will get CVD if you make them quit smoking??

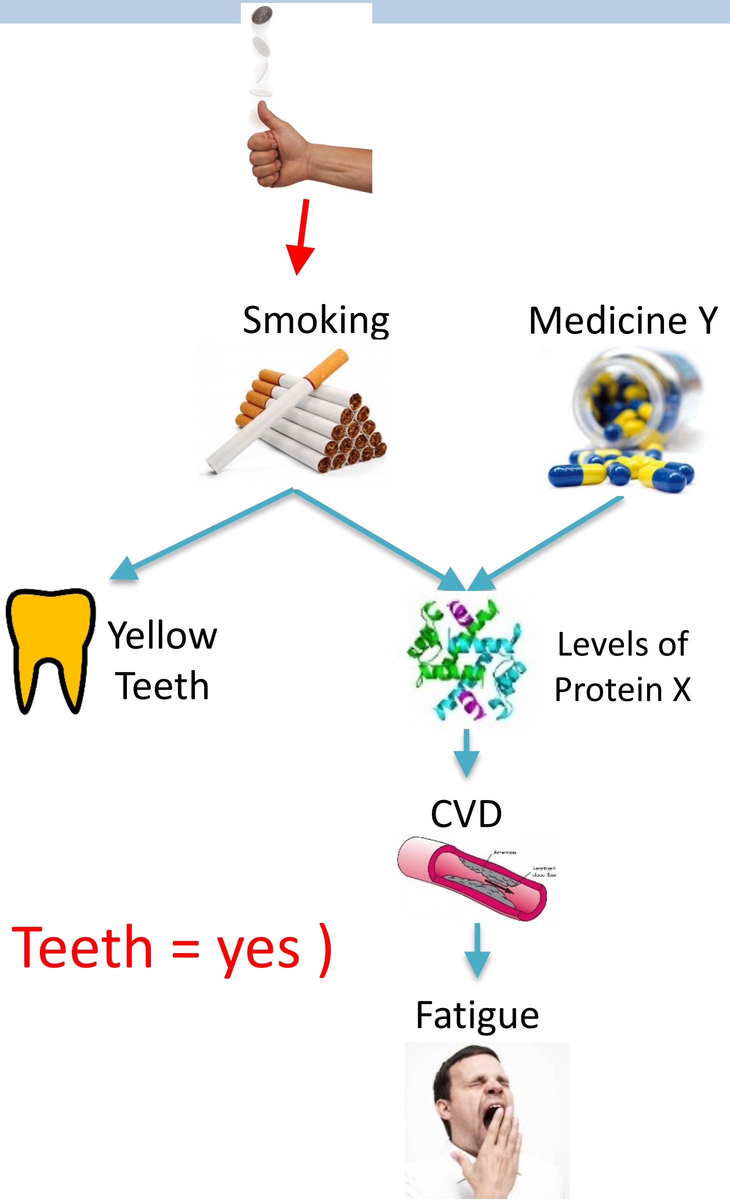


P(CVD| do(Smoking=no), Medicine Y = no, Yellow Teeth = yes)
= ?

Causal Inference

You measure some covariates for a patient.
(medicine y, yellow teeth)
What is the probability they will get CVD if you make them quit smoking??

If you measure all covariates, you can do inference on the manipulated graph



P(CVD| do(Smoking=no), Medicine Y = no, Yellow Teeth = yes) = ?

The do-calculus

Rule 1: Insertion/deletion of observations

$$P(Y|do(X),Z,W) = P(Y|do(X),W) \text{ if } \mathrm{d}sep(Y,Z|X,W) \text{ in } G_{\overline{X}}$$

Rule 2: Action/observation exchange

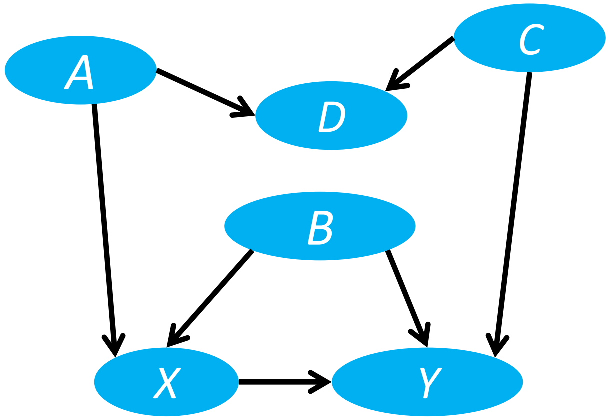
$$P(Y|do(X),do(Z),W) = P(Y|do(X),Z,W) \text{ if } dsep(Y,Z|X,W) \text{ in } G_{\overline{X}\underline{Z}}$$

Rule 3: Insertion/deletion of actions

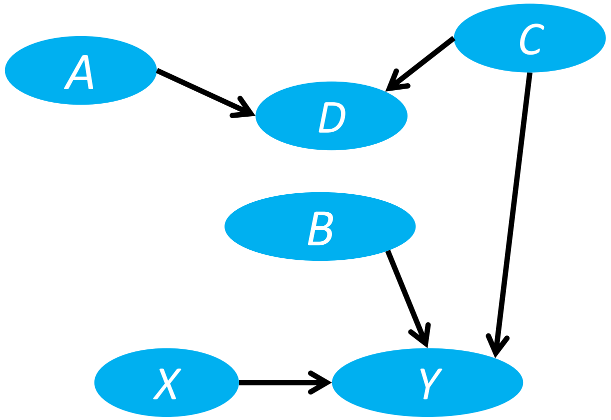
$$P(Y|do(X),do(Z),W) = P(Y|do(X),W) \text{ if } \mathrm{dsep}(Y,Z|X,W) \text{ in } G_{\overline{XZ(W)}}$$

where $Z(W)$ is the set of Z-nodes that are not ancestors of any W-nodes in $G_{\overline{X}}$

Notation used in the do-calculus

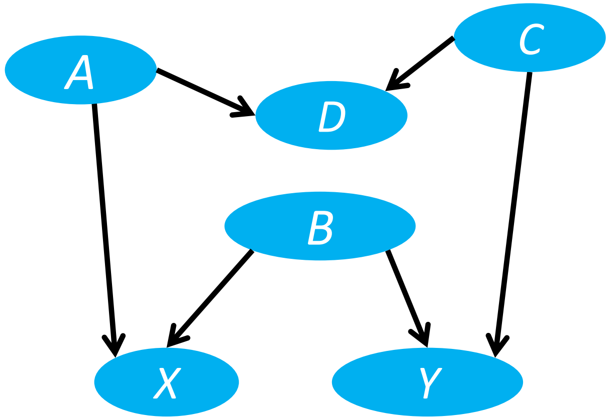


G



$G_{\overline{X}}$

Remove all edges into X (manipulated graph)



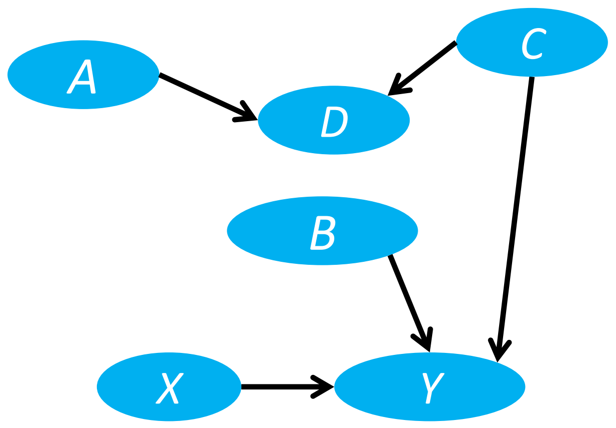
$G_{\underline{X}}$

Remove all edges out of X

Rule 1: Insert/Delete Observation

Rule 1: Insertion/deletion of observations

$P(Y|do(X),Z,W) = P(Y|do(X),W)$ if $\mathrm{d}sep(Y,Z|X,W)$ in $G_{\overline{X}}$



Independence in $G_{\overline{X}}$:

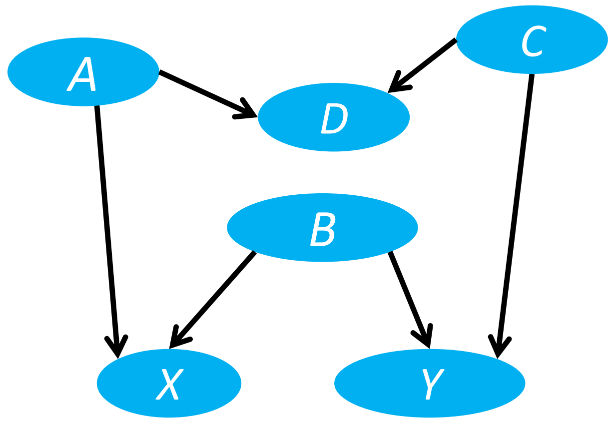
If Z is independent of Y given W in $G_{\overline{X}}$, you can remove Z from

$P(Y|do(X),A) = P(Y|do(X))$

Rule 2: Action/Observation exchange

Rule 2: Action/observation exchange

$P(Y|do(X),do(Z),W) = P(Y|do(X),Z,W)$ if $dsep(Y,Z|X,W)$ in $G_{\overline{X}\underline{Z}}$



Blocking backdoor paths

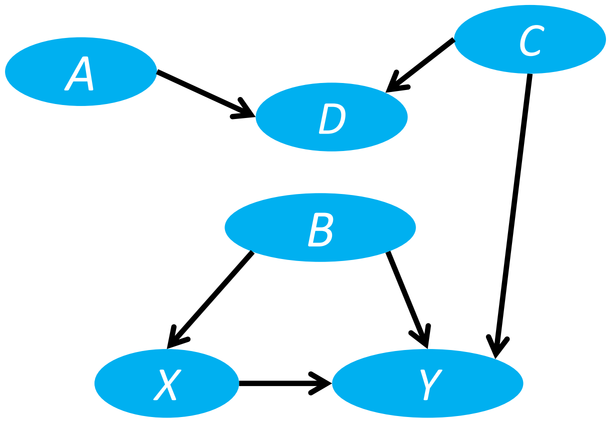
If all non causal paths between X and Y are blocked, observing is the same as acting

$P(Y|do(X),B) = P(Y|X,B)$

Rule 3: Insert/Delete Action

Rule 3: Insertion/deletion of actions

$P(Y|do(X), do(Z), W) = P(Y|do(X), W)$ if $\text{dsep}(Y, Z|X, W)$ in $G_{\overline{XZ(W)}}$ where $Z(W)$ is the set of Z-nodes that are not ancestors of any W-nodes in $G_{\overline{X}}$



If there is not path from Z to Y, you can remove $do(Z)$

$P(D|do(X)) = P(D)$

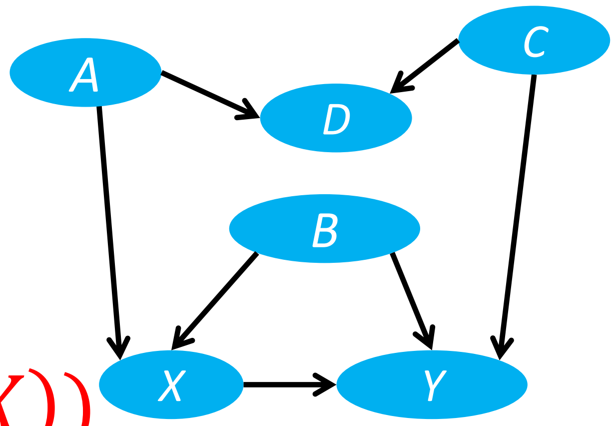
Do-calculus

An algorithm for converting “do”-probabilities to “see”-probabilities
You know the graph, and you have an estimate of the observational probability distribution, and you want to answer: what is $P(Y|do(X))$?

You can use the rules of do-calculus to get an answer.

Do-calculus / Example 1

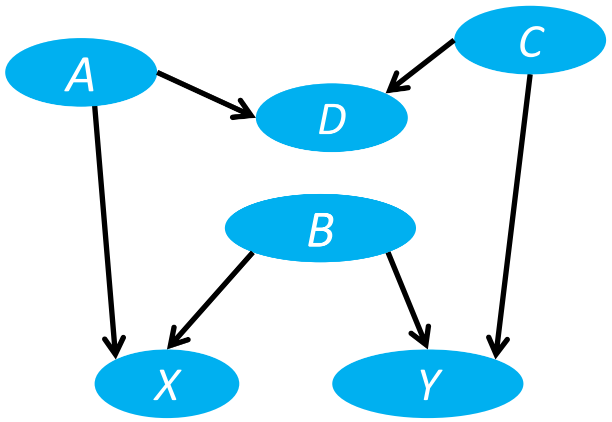
What is $P(Y|do(X))$?



$$P(Y|do(X)) = \sum_B P(Y|do(X), B)P(B|do(X))$$

Do-calculus / Example 1

Rule 2:
$P(Y|do(X), do(Z), W) = P(Y|do(X), Z, W)$
if $(Y \perp\!\!\!\perp Z|X, W)$ in $G_{\overline{X}\underline{Z}}$



What is $P(Y|do(X))$?

$$P(Y|do(X))$$
$$= \sum_B P(Y|do(X), B) P(B|do(X))$$

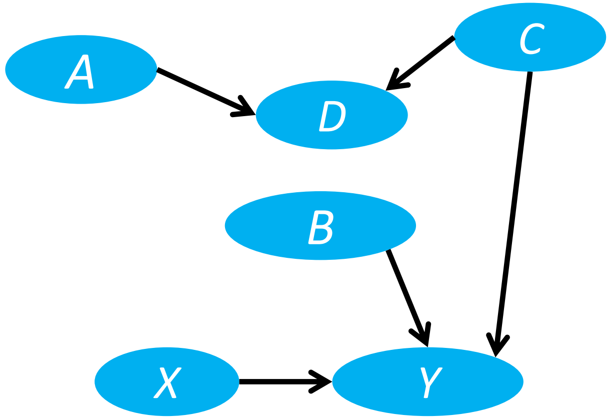
Rule 2 with
$Z \leftarrow X$
$X \leftarrow \emptyset$
$W \leftarrow B$

$$\sum_B P(Y|X, B) P(B|do(X))$$

Do-calculus / Example 1

Rule 3:

$P(Y|do(X), do(Z), W) = P(Y|do(X), W)$ if $(Y \perp\!\!\!\perp Z|X, W)$ in $G_{\overline{XZ(W)}}$



What is $P(Y|do(X))$?

$$P(Y|do(X)) = \sum_B P(Y|do(X), B) P(B|do(X))$$

$$= \sum_B P(Y|X, B) P(B|do(X))$$

$$= \sum_B P(Y|X, B) P(B)$$

Rule 3 with

$Z \leftarrow X$

$X \leftarrow \emptyset$

$W \leftarrow \emptyset$

$Y \leftarrow B$

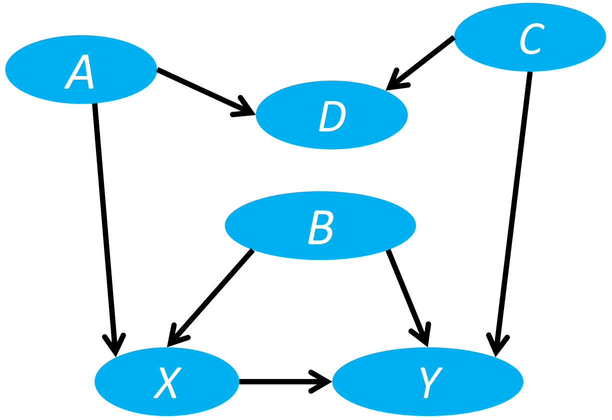
Backdoor criterion: estimate the average treatment effect

What is $P(Y|do(X))$?

$$P(Y|do(X)) = \sum_B P(Y|do(X), B)P(B|do(X))$$

$$= \sum_B P(Y|X, B)P(B|do(X))$$

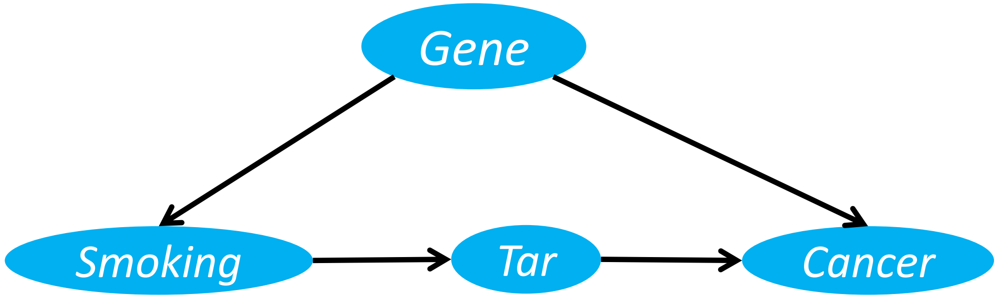
$$= \sum_B P(Y|X, B)P(B)$$



Find a set of pre-treatment covariates that block all backdoor paths, and "adjust" for their influence

Do-calculus / Example 2

What is $P(C|do(S))$?



$P(C|do(S)) = \sum_t P(C|do(S),t)P(t|do(S))$ [Probability axioms]

$= \sum_t P(C|do(S),do(t))P(t|do(S))$ [Rule 2: exchange t/do(t)]

$= \sum_t P(C|do(S),do(t))P(t|S)$ [Rule 2: exchange do(S)/S]

$= \sum_t P(C|\,do(t))P(t|S)$ [Rule 3: Remove do(S)]

$= \sum_{s'} \sum_t P(C|\,do(t),s')P(s'|do(t))P(t|S)$ [Probability axioms]

$= \sum_{s'} \sum_t P(C|\,t,s')P(s'|do(t))P(t|S)$ [Rule 2: Exchange t/do(t)]

$= \sum_{s'} \sum_t P(C|\,t,s')P(s')P(t|S)$ [Rule 3: Remove do(t)]

Do-calculus

Allows us to get post-intervention probabilities from pre-intervention probabilities

Complete for identification of post-intervention probabilities:

- If we can identify a post-intervention probability from the pre-intervention probability, we can do this using some combination of do-calculus rules+ the axioms of probability.

Things you can do with a Causal Bayesian Network

1. Factorize the joint probability distribution.

2. Answer questions like:
   1. What is the probability of getting CVD if I have high levels of Protein X?
      - P(CVD| Levels of Protein X=high) = ?

   2. Is Smoking independent from Fatigue given Levels of Protein X?
      - Smoking ⫫ Fatigue|Levels of Protein X?

   3. Will I reduce the probability of getting CVD if I design a drug that lowers the levels of protein X?

      - P(CVD|do(Levels of Protein X=low))?