ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

Μηχανική Μάθηση

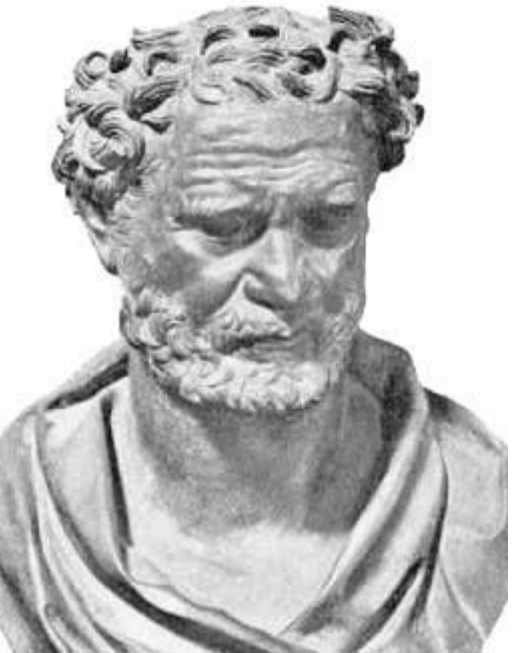
Ενότητα 11: Introduction to **causal discovery**: A **Bayesian Networks** approach

Ιωάννης Τσαμαρδίνος
Τμήμα Επιστήμης Υπολογιστών

Pt.1 - Introduction

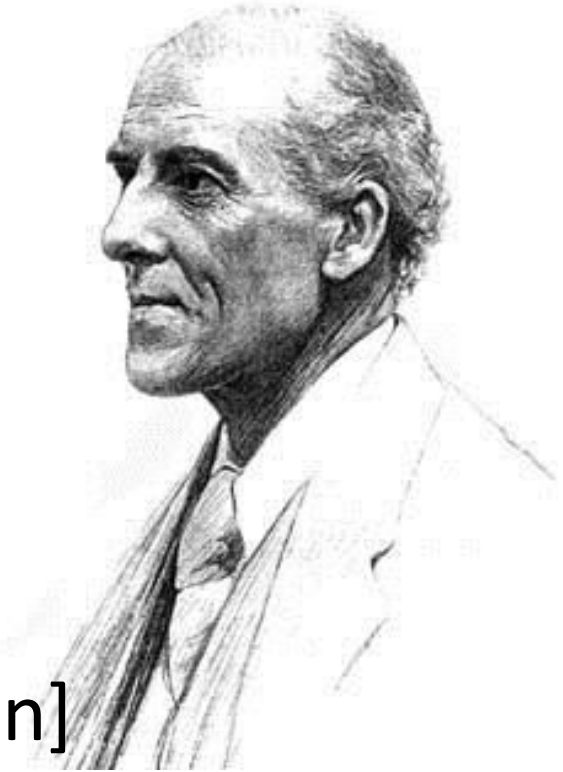
Democritus said that he would rather discover a single cause than be the king of Persia

Δημόκριτος έλεγε βούλεσθαι μάλλον μία ευρείν αιτιολογίαν ή την Περσών βασιλείαν εαυτού γενέσθαι



“Beyond such discarded fundamentals as ‘matter’ and ‘force’ lies still another fetish amidst the inscrutable arcana of modern science, namely the category of cause and effect”

[K. Pearson]



What is causality?

- What do you understand when I say :
  - smoking causes lung cancer?

What is (probabilistic) causality?

- What do you understand when I say :
  - smoking causes lung cancer?

If you decide to start smoking, you will increase your chances of getting lung cancer (vs. deciding not to start)

Why do we need causality?

"Decreasing interest rates causes a decrease in unemployment rate"

Everything else being equal, lower interest rates to decrease unemployment

"High expression of gene *X* is increasing the chances of metastasis"

Design a drug to block expression of gene *X* to lower the chances of metastasis

Association?

- We want to know what causes lung cancer
- We have documented the **smoking** habits of 10000 people, whether they get **lung cancer** by 60yrs and whether they have **yellow** stains on their **teeth**

		Lung cancer	
Smoking	Yellow Teeth	Yes	No
Yes	Yes	100	400
Yes	No	100	400
No	Yes	1	450
No	No	9	8540

Data from 10000 people*

*fictional

Association

- *X* and *Y* are associated
  - Observing the value of *X* may change the distribution of the (observed) values of *Y*
  - Knowledge of *X* provides information for *Y*
  - *X* is predictive for *Y*
  - and vice versa

  - Makes no claims about the distribution of *Y,* if instead of observing, we intervene on the values of *X*

Sampling

- How do you chose 10000 people?
  - Randomly selected people from the general public
  - Identically and independently distributed (i.i.d)
    - Identically: sampled from the same population (distribution)
    - Independently: previous samples do not affect what the future samples will be
  - Other sampling schemes may affect the measured associations (e.g., case-control studies, experimental studies, selection bias)

Measuring Association

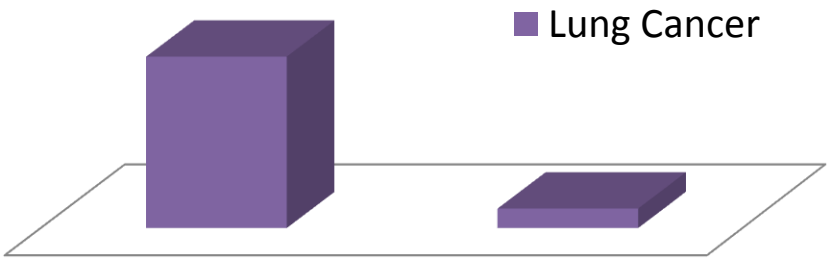
- Mutual Information (MI) (information theory)
- Association / correlation / effect size (statistic)
  - Pearson (linear) correlation, Spearman correlation for continuous variables
  - Cramér's V for nominal variables
  - Many other measures
- Mutual Information
  - General measure that assumes knowledge of the distribution
  - Specific parametric choices may make it equivalent to statistical approaches

Determining Dependency

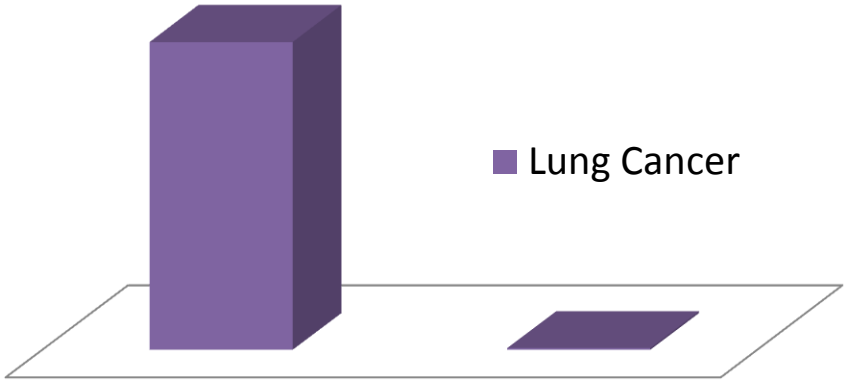
- MI > 0, Association > 0 $\Leftrightarrow$ Dependency
- Perform a hypothesis testing on the null assumption that “Association = 0” and obtain a p-value
  - Statistical approaches explicitly address finite sample estimation problems
- Threshold Mutual Information
  - Threshold interpretation depends on sample size

Association is NOT Causality

Yellow Teeth vs Lung Cancer



Smoking vs Lung Cancer



		Lung cancer	
Smoking	Yellow Teeth	Yes	No
Yes	Yes	100	400
Yes	No	100	400
No	Yes	1	450
No	No	9	8540

Data from 10000 people

Association is NOT Causality

- Yellow teeth and lung cancer are associated
- Can I bleach my teeth and reduce the probability of getting lung cancer?
- Is Smoking really causing Lung Cancer?

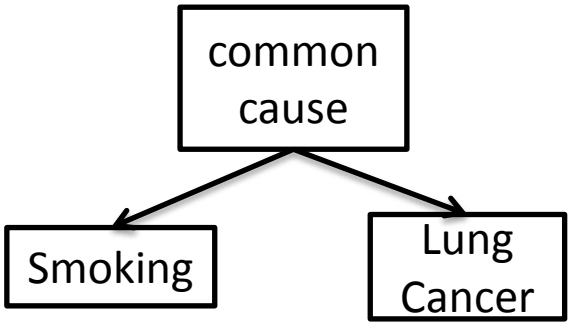
BUT

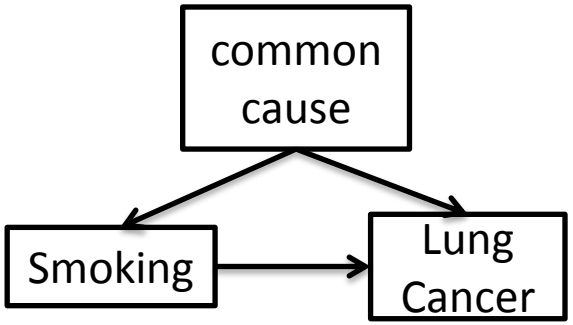
"If A and B are correlated, A causes B OR B causes A OR they share a latent common cause"

[Hans Reichenbach]

Is Smoking Causing Lung Cancer?

All possible models*





*assuming:

1. Smoking precedes Lung Cancer
2. No feedback cycles
3. Several hidden common causes can be modeled by a single hidden common cause

A way to learn causality

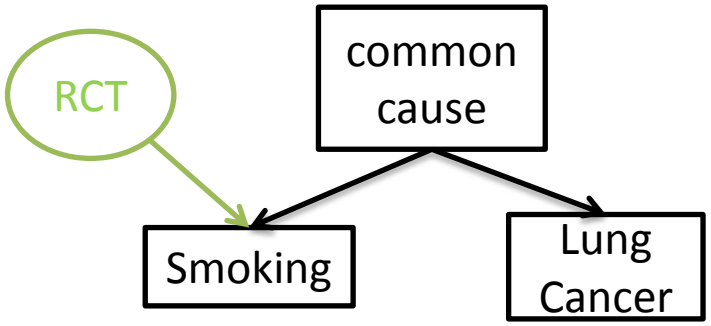
1. Take 200 people
2. Randomly split them in control and treatment groups
3. Force control group to smoke, force treatment group not to smoke
4. Wait until they are 60 years old
5. Measure correlation

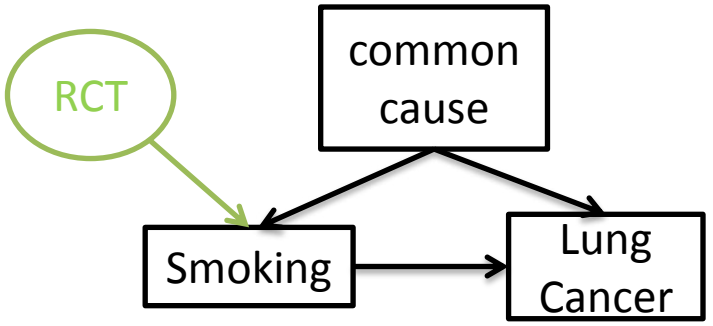
[Randomized Control Trial]

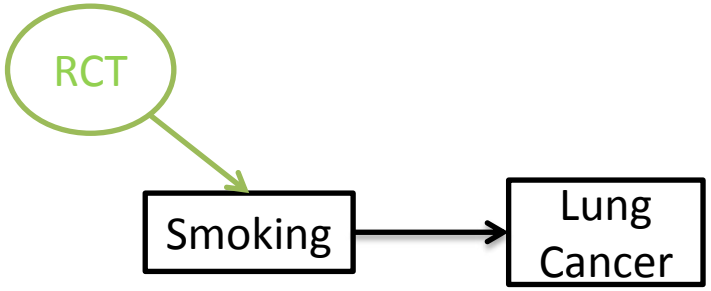
[Sir Ronald Fisher]

Manipulation

All possible models*

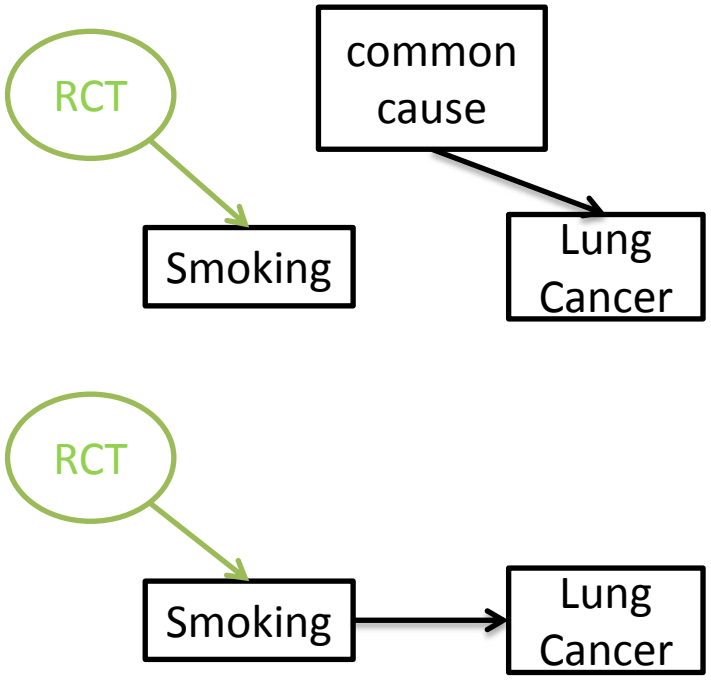


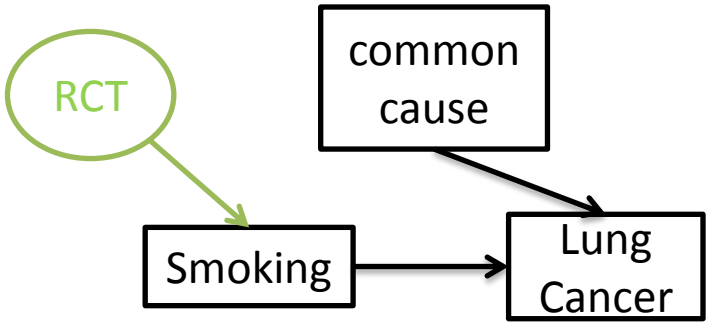




Manipulation removes other causes

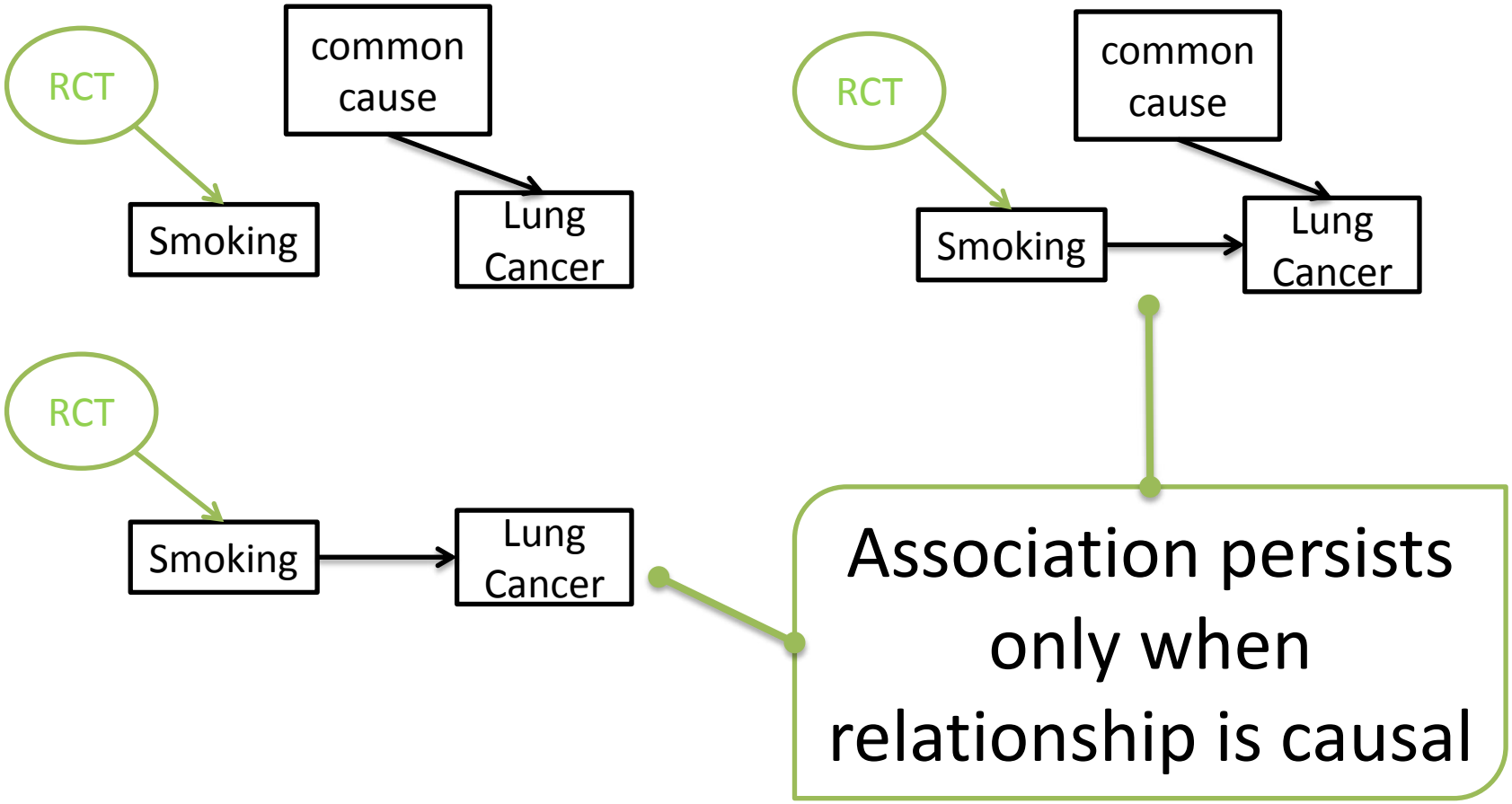
All possible models*





Manipulation removes other causes

All possible models*



RCTs are hard

- Can we learn anything from observational data?

RCTs are hard

- Can we learn anything from observational data?

"If A and B are correlated, A causes B OR B causes A OR they share a latent common cause"

[Hans Reichenbach]

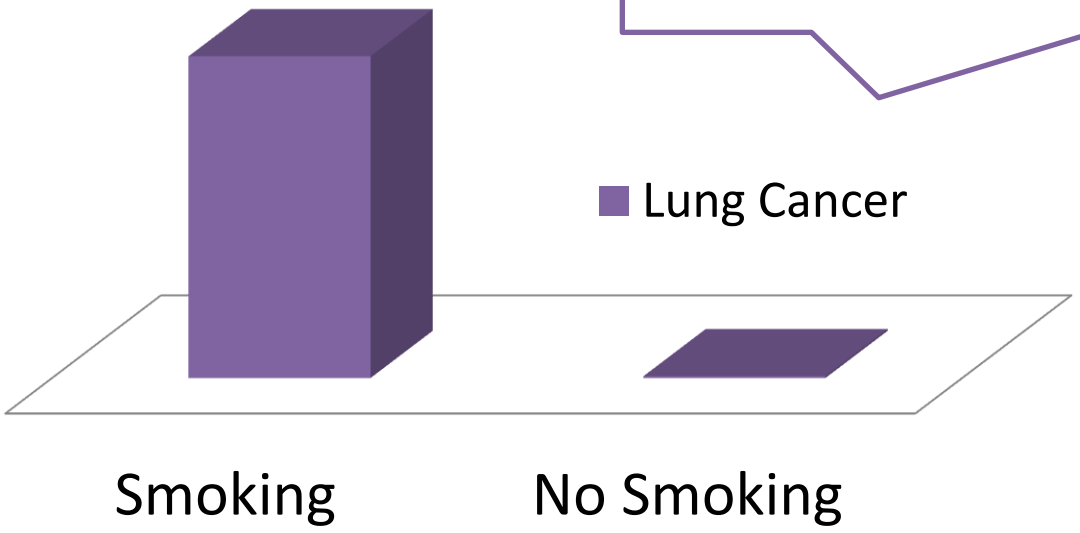
(Un)Conditional (In)dependence

- Lung Cancer in the general population: 2.1%

(Un)Conditional (In)dependence

- Lung Cancer among people who don't smoke: 0.1%
- Lung Cancer among people who smoke: 20 %

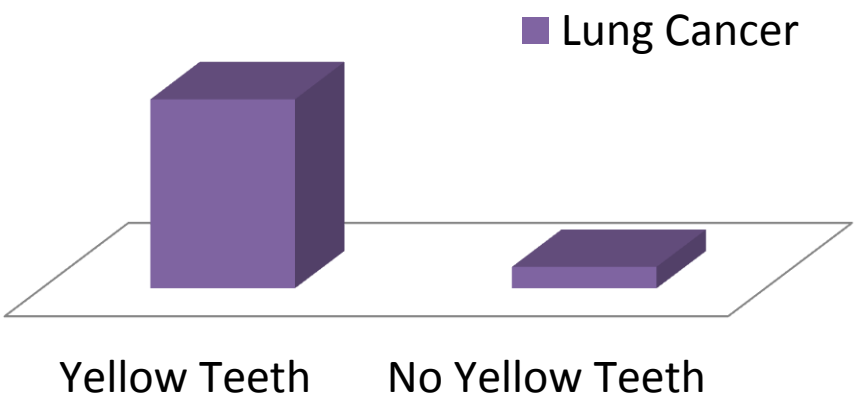
$Dep(Smoking, Lung\ Cancer | \varnothing)$



(Un)Conditional (In)dependence

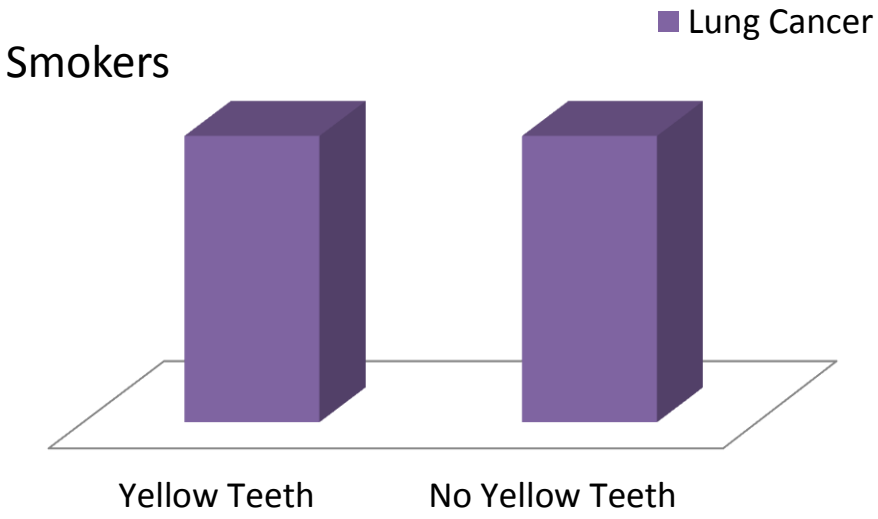
- Lung Cancer among people with no yellow teeth: 1.1%
- Lung Cancer among people who have yellow teeth: 10.6%

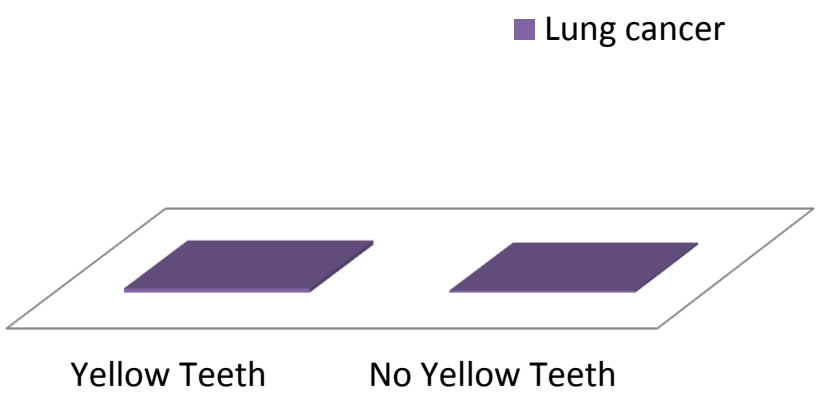
$$Dep(\text{Yellow Teeth}, \text{Lung Cancer} \mid \varnothing)$$



Conditional (In)dependence

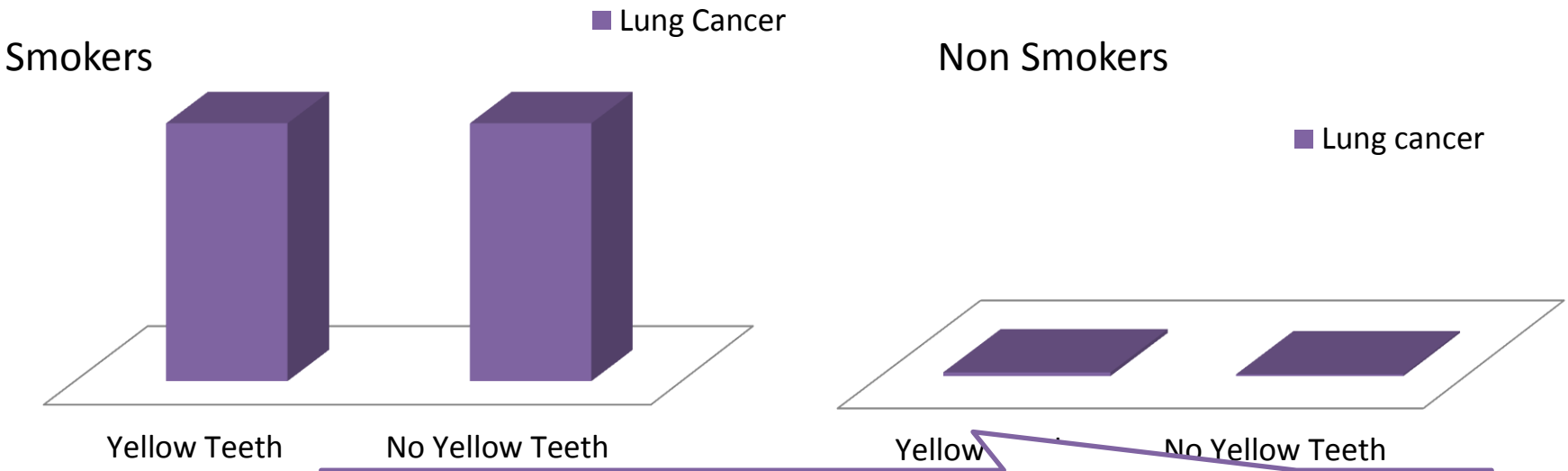
- Lung Cancer among people who smoke AND have yellow teeth: 20%
- Lung Cancer among people who smoke AND have no yellow teeth: 20%
- Lung Cancer among people who don't smoke AND have yellow teeth: 0.01%
- Lung Cancer among people who don't smoke AND have no yellow teeth: 0.01%





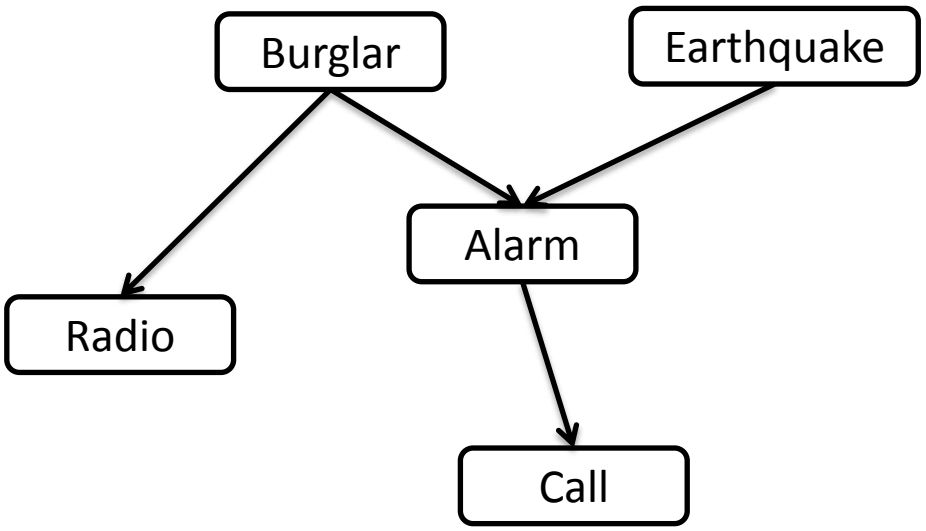
Conditional (In)dependence

- Lung Cancer among people who smoke AND have yellow teeth: 20%
- Lung Cancer among people who smoke AND have no yellow teeth: 20%
- Lung Cancer among people who don't smoke AND have yellow teeth: 0.01%
- Lung Cancer among people who don't smoke AND have no yellow teeth: 0.01%



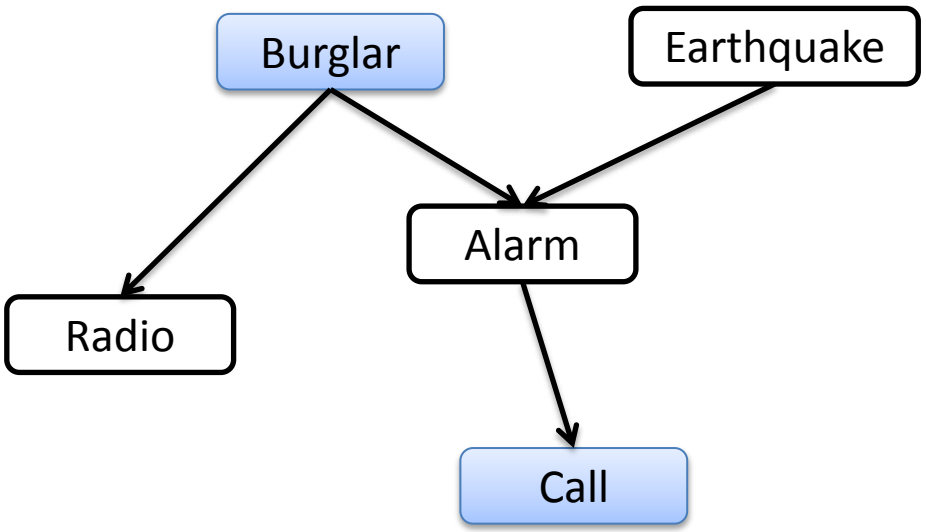
Ind(Lung Cancer, Yellow Teeth| Smoking)

Conditioning and Causality



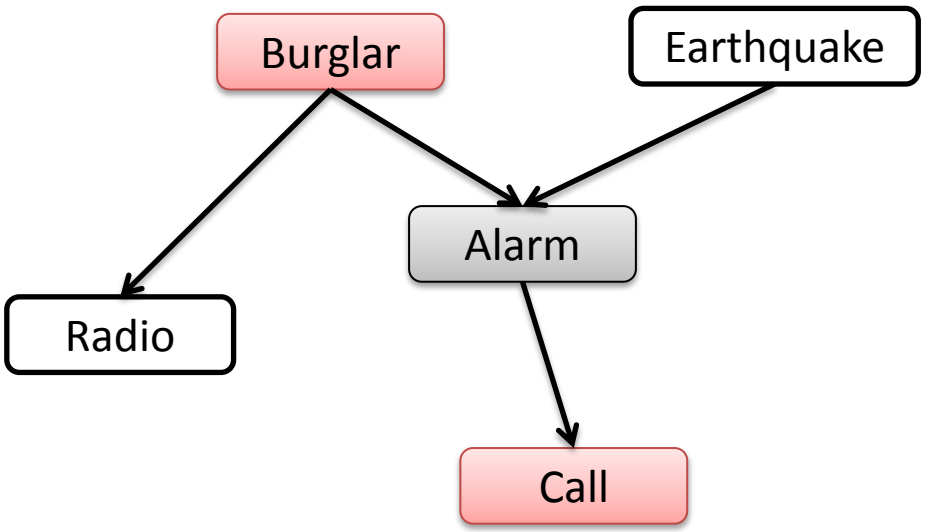
[example by Judea Pearl]

Conditioning and Causality



Dep(Burglar, Call|∅)

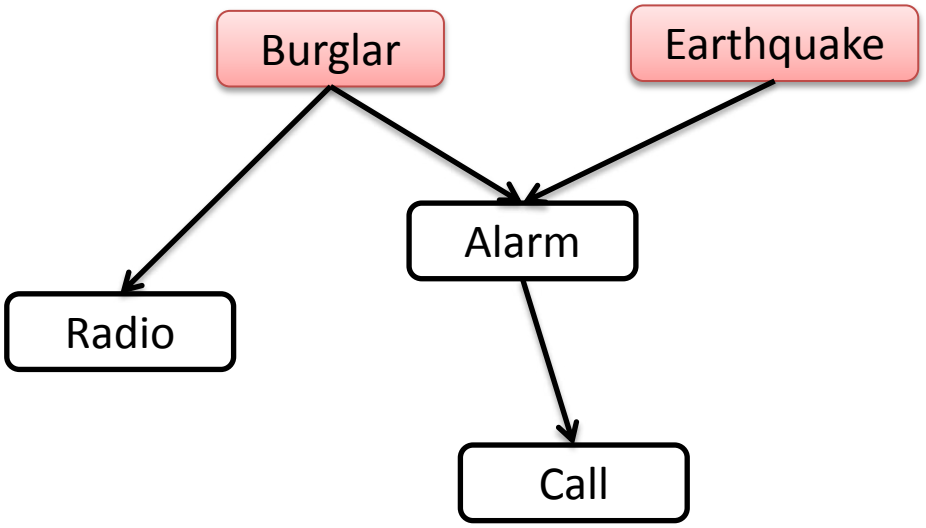
Conditioning and Causality



Learning the value of intermediate and common causes renders variables independent

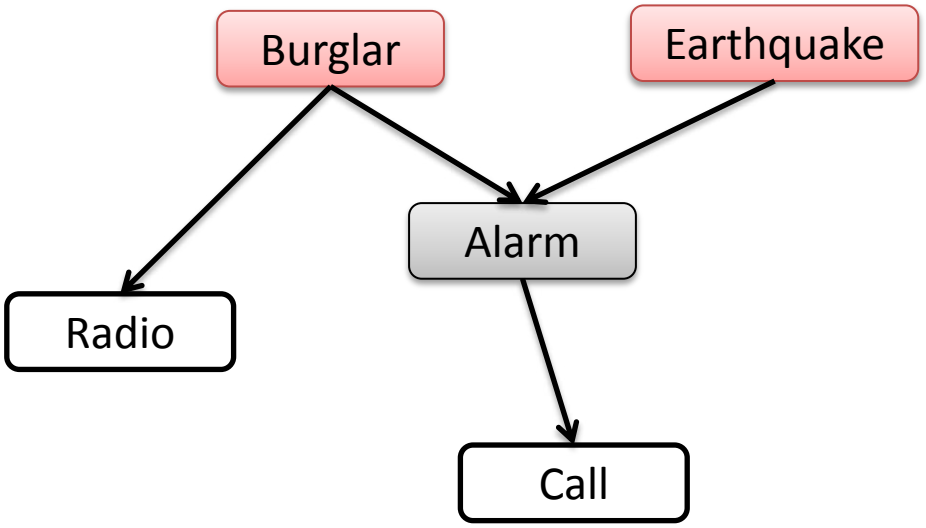
Ind (Burglar, Call|Alarm)

Conditioning and Causality



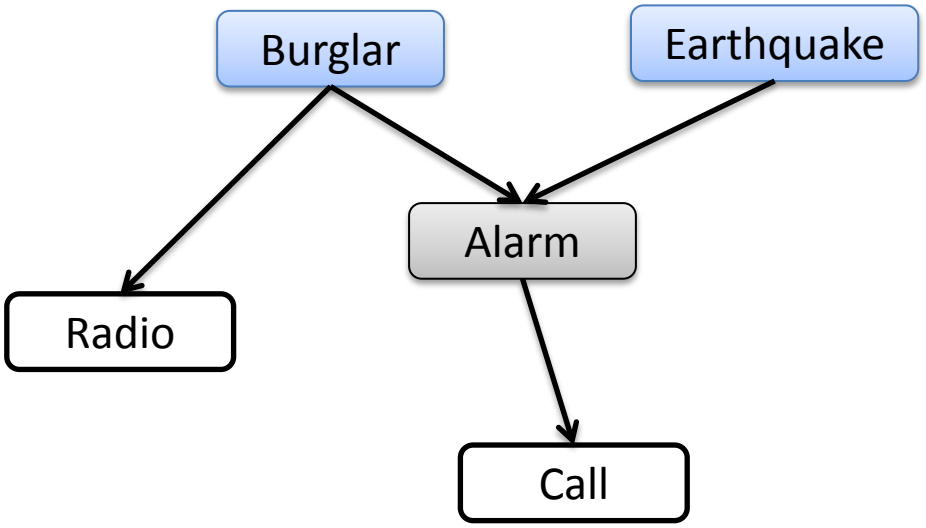
Ind (Burglar, Earthquake|∅)

Conditioning and Causality



Ind (Burglar, Earthquake|∅)

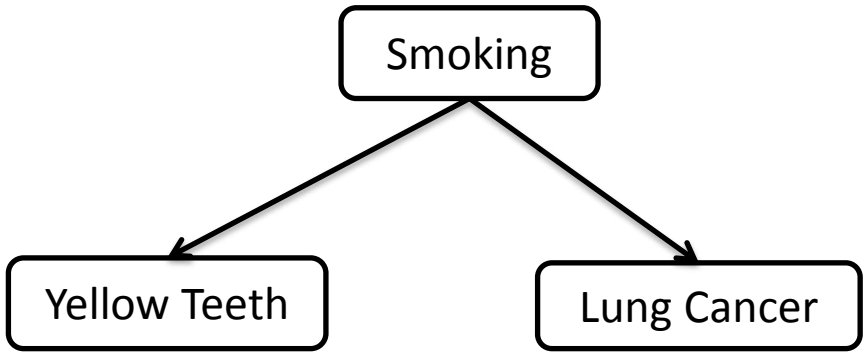
Conditioning and Causality



Learning the value of common effects renders variables dependent

Dep (Burglar, Earthquake| Alarm)

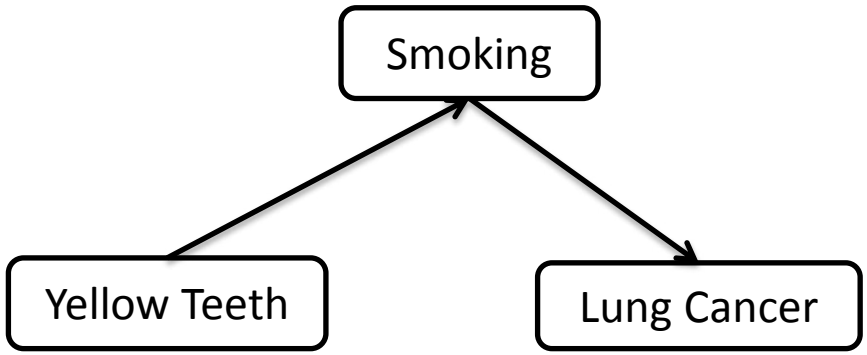
Observing a causal model



What would you observe?

- Dep(Lung Cancer, Yellow Teeth|∅)
- Dep(Smoking, Lung Cancer |∅)
- Dep(Lung Cancer, Yellow Teeth |∅)
- Ind(Lung Cancer, Yellow Teeth| Smoking)

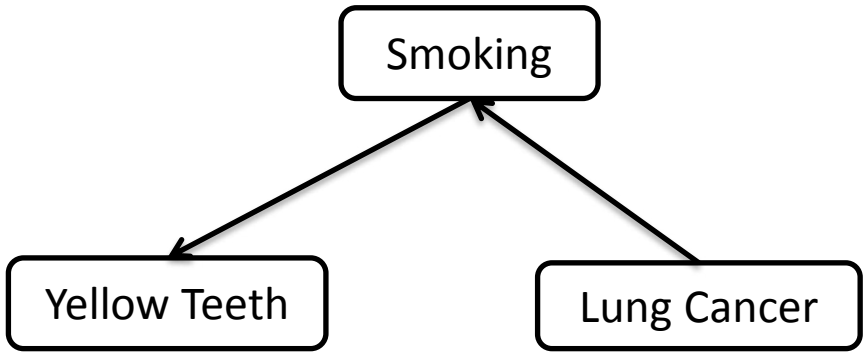
Observing a causal model



What would you observe?

- Dep(Lung Cancer, Yellow Teeth|∅)
- Dep(Smoking, Lung Cancer |∅)
- Dep(Lung Cancer, Yellow Teeth |∅)
- Ind(Lung Cancer, Yellow Teeth| Smoking)

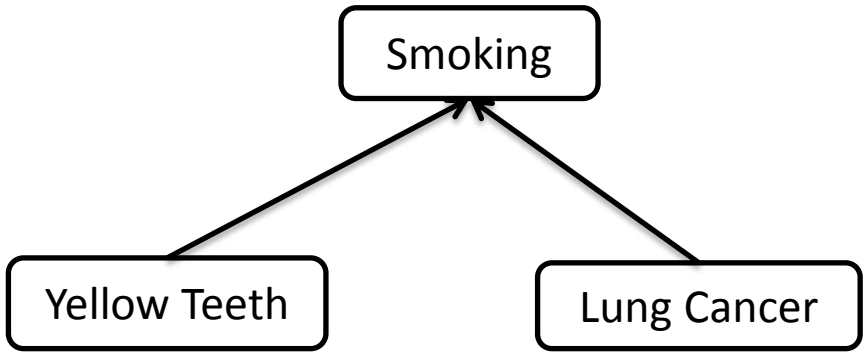
Observing a causal model



What would you observe?

- Dep(Lung Cancer, Yellow Teeth|∅)
- Dep(Smoking, Lung Cancer |∅)
- Dep(Lung Cancer, Yellow Teeth |∅)
- Ind(Lung Cancer, Yellow Teeth| Smoking)

Observing a causal model

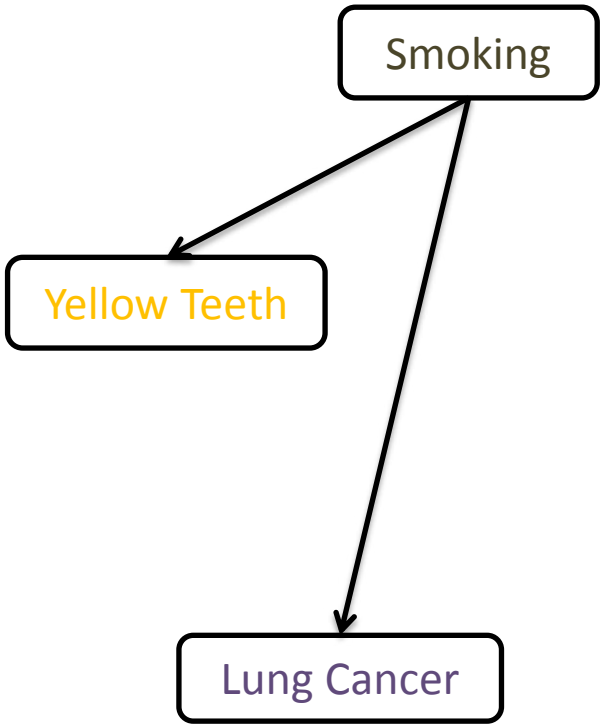


What would you observe?

- Dep(Lung Cancer, Yellow Teeth|∅)
- Dep(Smoking, Lung Cancer |∅)
- Ind(Lung Cancer, Yellow Teeth |∅)
- Dep(Lung Cancer, Yellow Teeth| Smoking)

Causal Bayesian Networks*

Graph $\mathcal{G}$



JPD J

		Lung Cancer	
Smoking	Yellow Teeth	Yes	No
Yes	Yes	0,01	0,04
Yes	No	0,01	0,04
No	Yes	0,000045	0,044955
No	No	0,000855	0,854145

Assumptions about the nature of causality connect the graph $\mathcal{G}$ with the observed distribution J and allow reasoning

*almost there

Conclusions, Pt 1

- Association measures the information for *Y* from **observing** the values of *X*
- Causality dictates what will happen if someone **intervenes on** (sets) the values of *X*
- Means to measure association and determine dependency
- Conditional Dependency
- Major means to determine causality is the RCT
- A causal structure implies certain expectations about properties of the observed distribution (e.g., conditional dependencies and independencies)

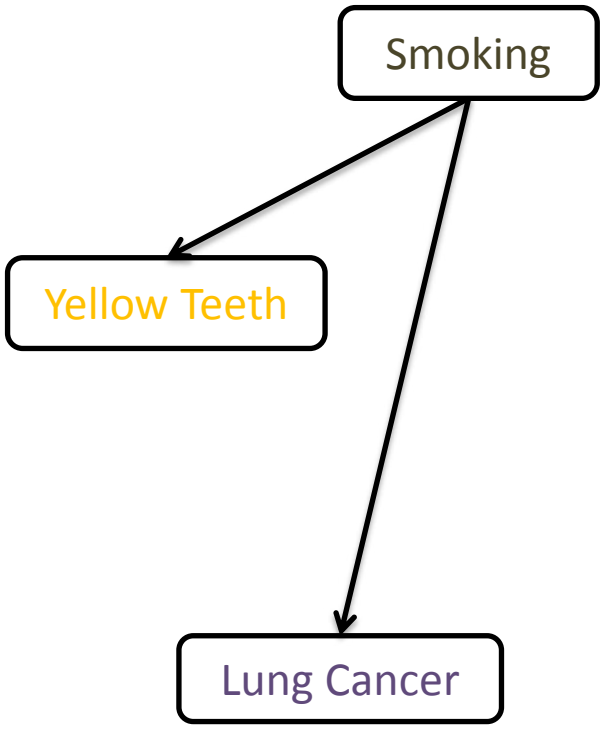
Bibliography

- Pearson, K., 1911. *The Grammar of Science*, London: Black Publishers.
- Fisher, R.A., 1935. *The Design of Experiments*, New York: Hafner Publishing.
- Reichenbach, H., 1956. *The Direction of Time*, Berkeley: University of California Press.
- Pearl, J., 2000. *Causality: Models, Reasoning, and Inference*, Cambridge University Press.

Pt.2 – (Causal) Bayesian Networks

Causal Bayesian Networks*

Graph $\mathcal{G}$



JPD J

		Lung Cancer	
Smoking	Yellow Teeth	Yes	No
Yes	Yes	0,01	0,04
Yes	No	0,01	0,04
No	Yes	0,000045	0,044955
No	No	0,000855	0,854145

*almost there

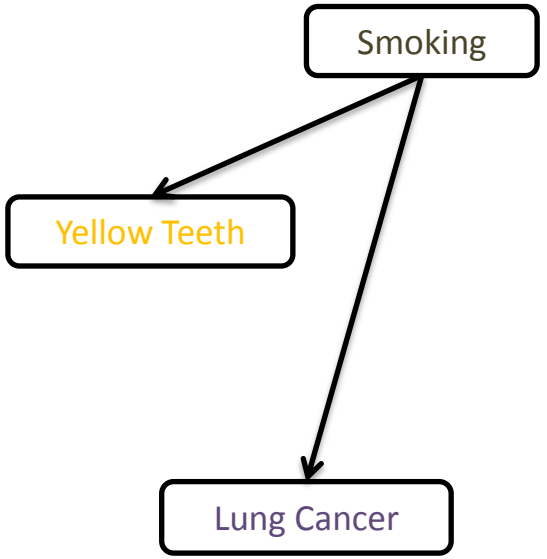
Causal Bayesian Networks*

Connecting the graph and the joint probability distribution:

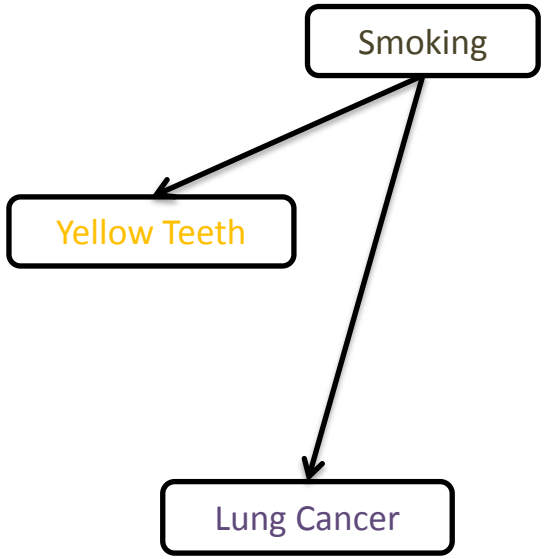
- Learning the value of intermediate and common causes renders variables independent

- Stems from our intuition about the nature of causality

		Lung Cancer	
Smoking	Yellow Teeth	Yes	No
Yes	Yes	0,01	0,04
Yes	No	0,01	0,04
No	Yes	0,000045	0,044955
No	No	0,000855	0,854145



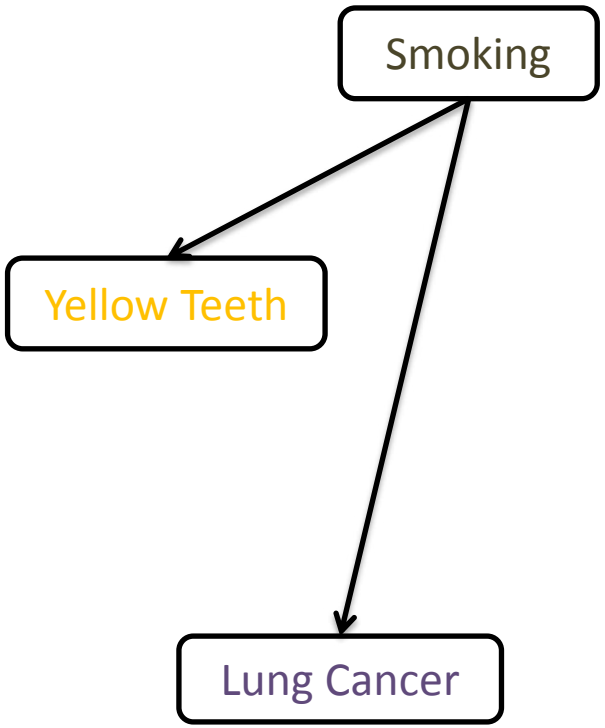
Causal Markov Condition (CMC)



		Lung Cancer	
Smoking	Yellow Teeth	Yes	No
Yes	Yes	0,01	0,04
Yes	No	0,01	0,04
No	Yes	0,000045	0,044955
No	No	0,000855	0,854145

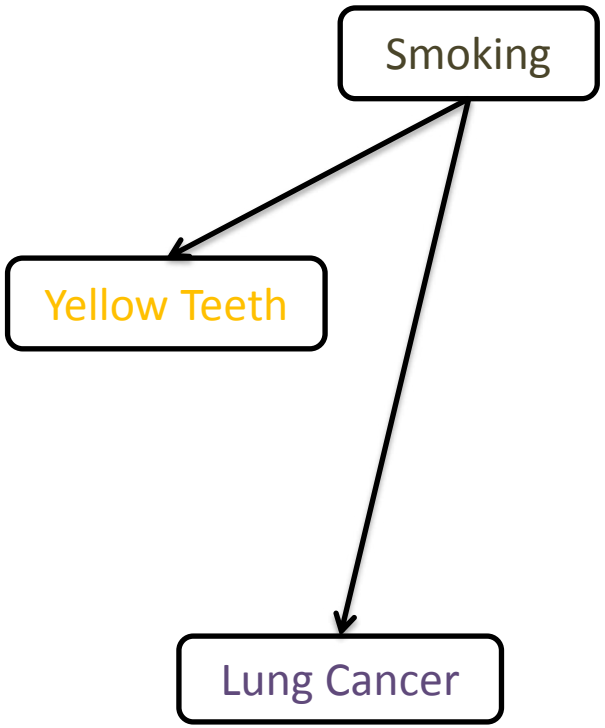
Every variable is independent of its non-effects (descendants in the graph) given its direct causes (parents)

Causal Markov Condition



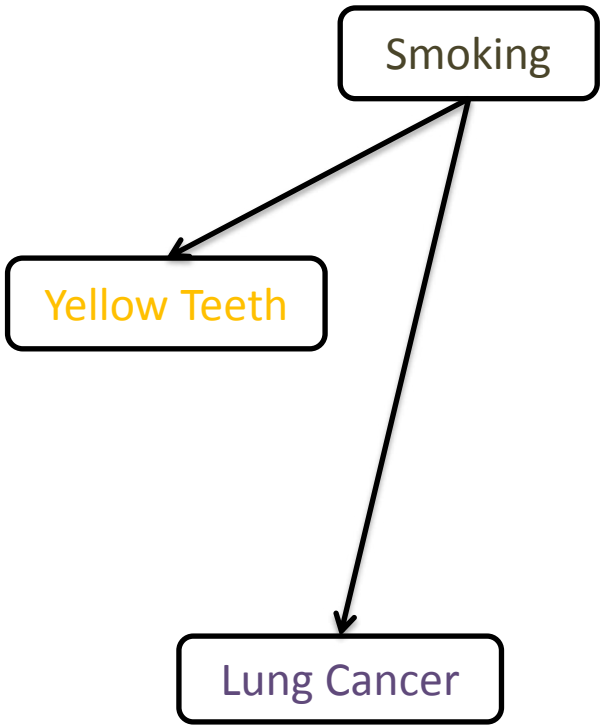
P(Yellow Teeth, Smoking, Lung Cancer)

Causal Markov Condition



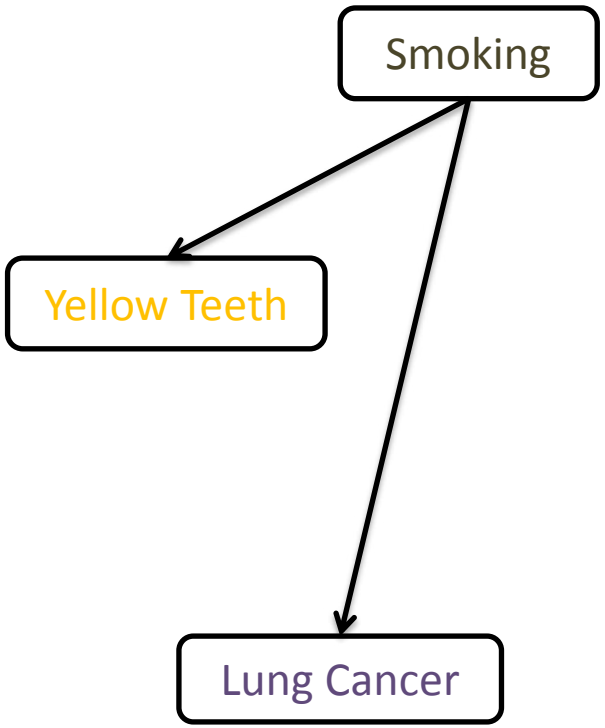
P(Yellow Teeth, Smoking, Lung Cancer) =
P(Smoking) ×
P(Yellow Teeth | Smoking) ×
P(Lung Cancer | Smoking, Yellow Teeth)

Causal Markov Condition



P(Yellow Teeth, Smoking, Lung Cancer) =
P(Smoking) ×
P(Yellow Teeth| Smoking) ×
P(Lung Cancer| Smoking, ~~Yellow Teeth~~)

Causal Markov Condition

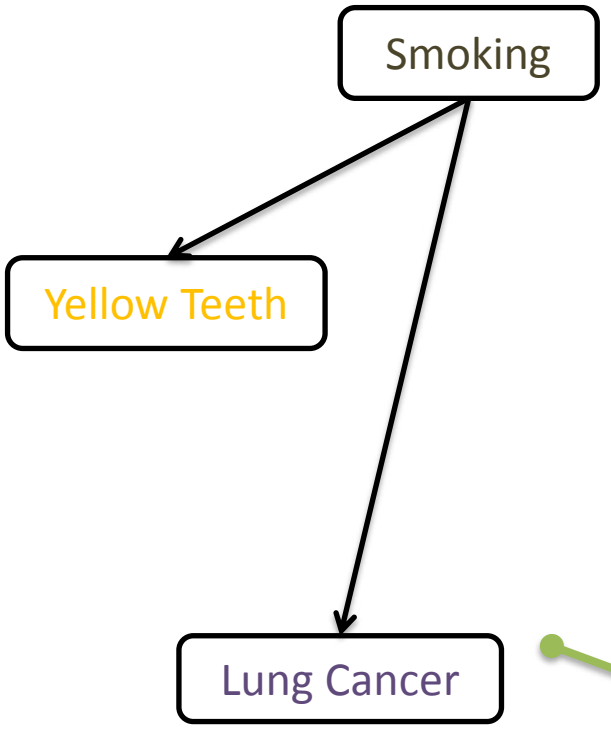


P(Yellow Teeth, Smoking, Lung Cancer) =
P(Smoking) ×
P(Yellow Teeth| Smoking) ×
P(Lung Cancer| Smoking)

$$P(\mathbf{V}) = \prod P(Vi|\ Pa(Vi))$$

Causal Bayesian Networks

Graph $\mathcal{G}$

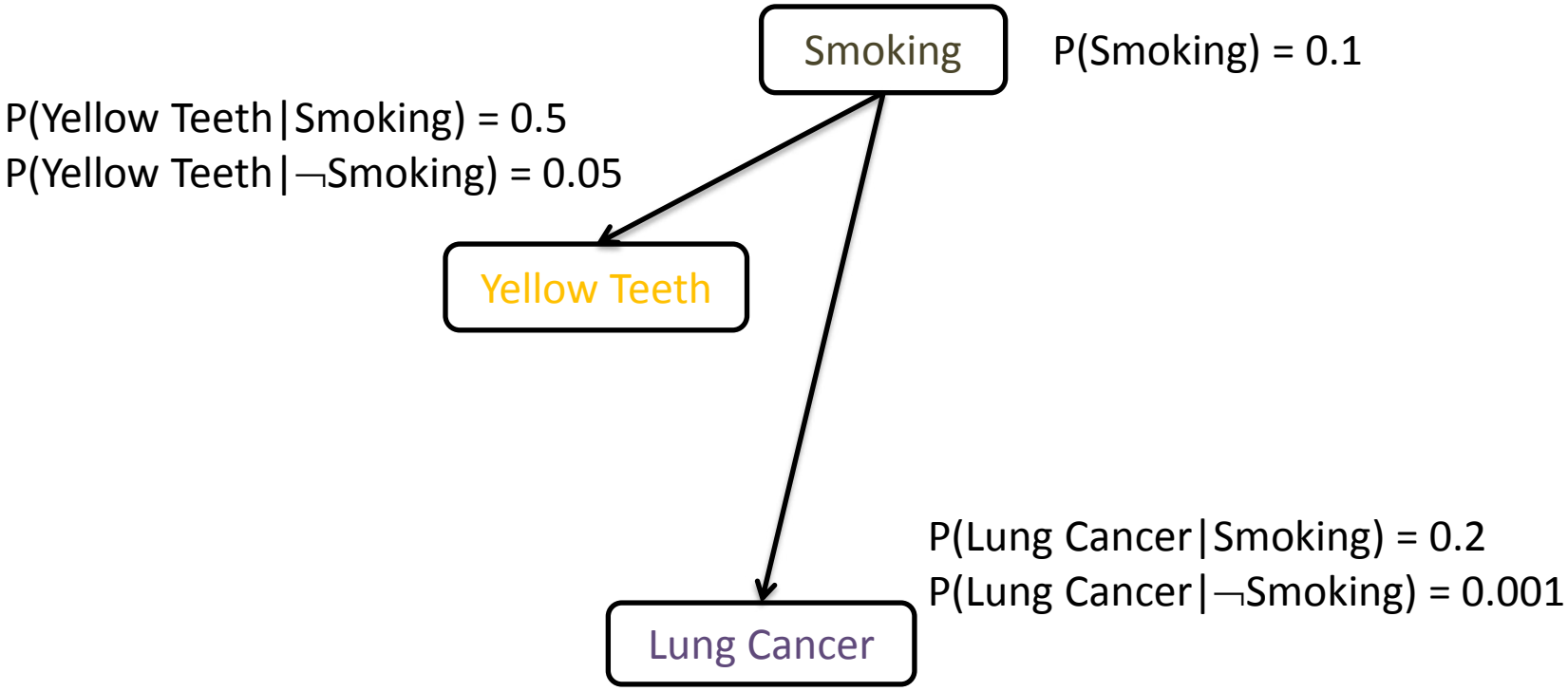


JPD J

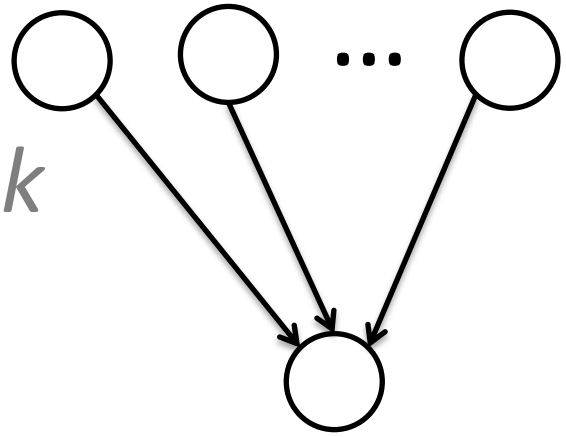
		Lung Cancer	
Smoking	Yellow Teeth	Yes	No
Yes	Yes	0,01	0,04
Yes	No	0,01	0,04
No	Yes	0,000045	0,044955
No	No	0,000855	0,854145

Causal Markov Condition

Factorization with the CMC



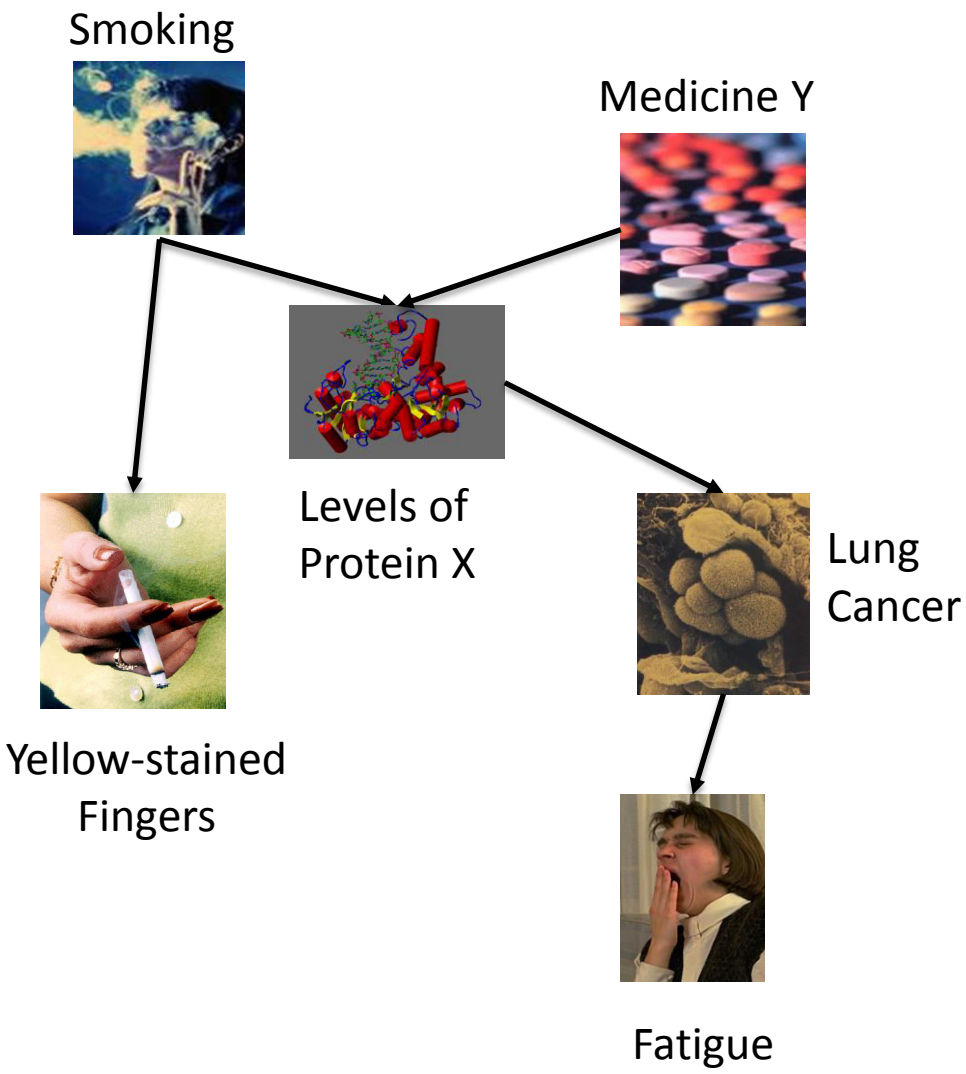
Factorization with the CMC



- Assume *n* binary variables, at most *k* parents each.
- Using P(**V**): $2^n - 1$ parameters
- Using $\prod$ P(Vi| Pa(Vi)) **: n*** $2^k - 1$ parameters

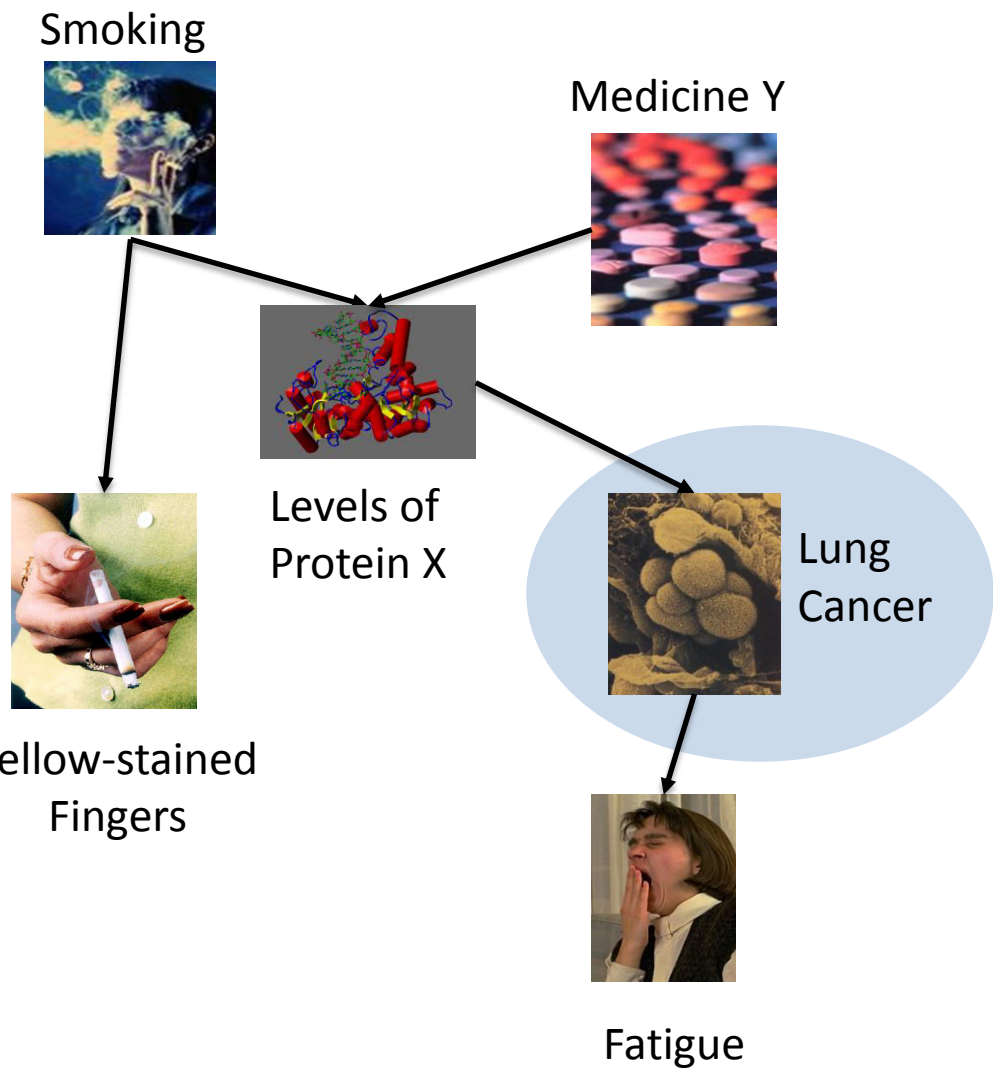
Reasoning with the CMC

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



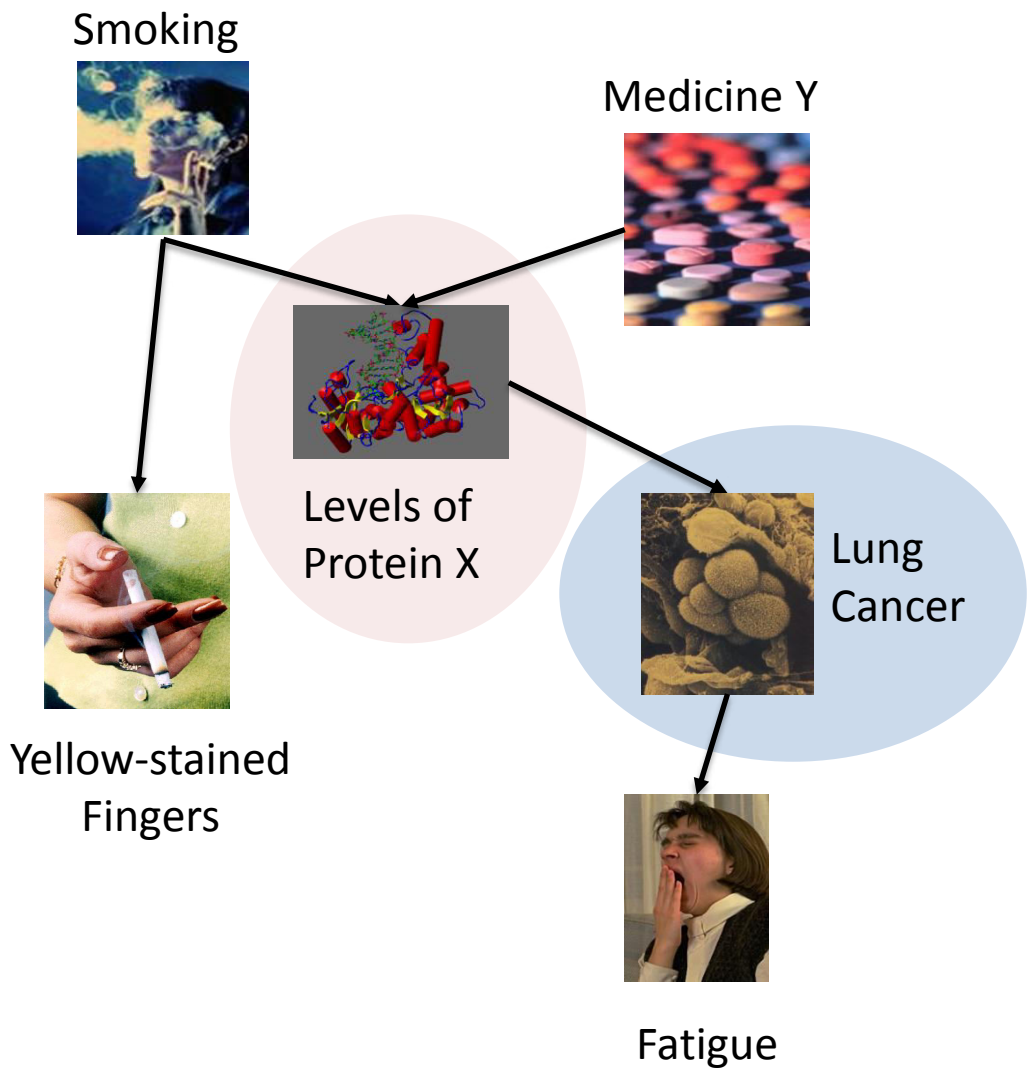
Reasoning with the CMC

Lung Cancer is independent of its non-effects given its direct causes.



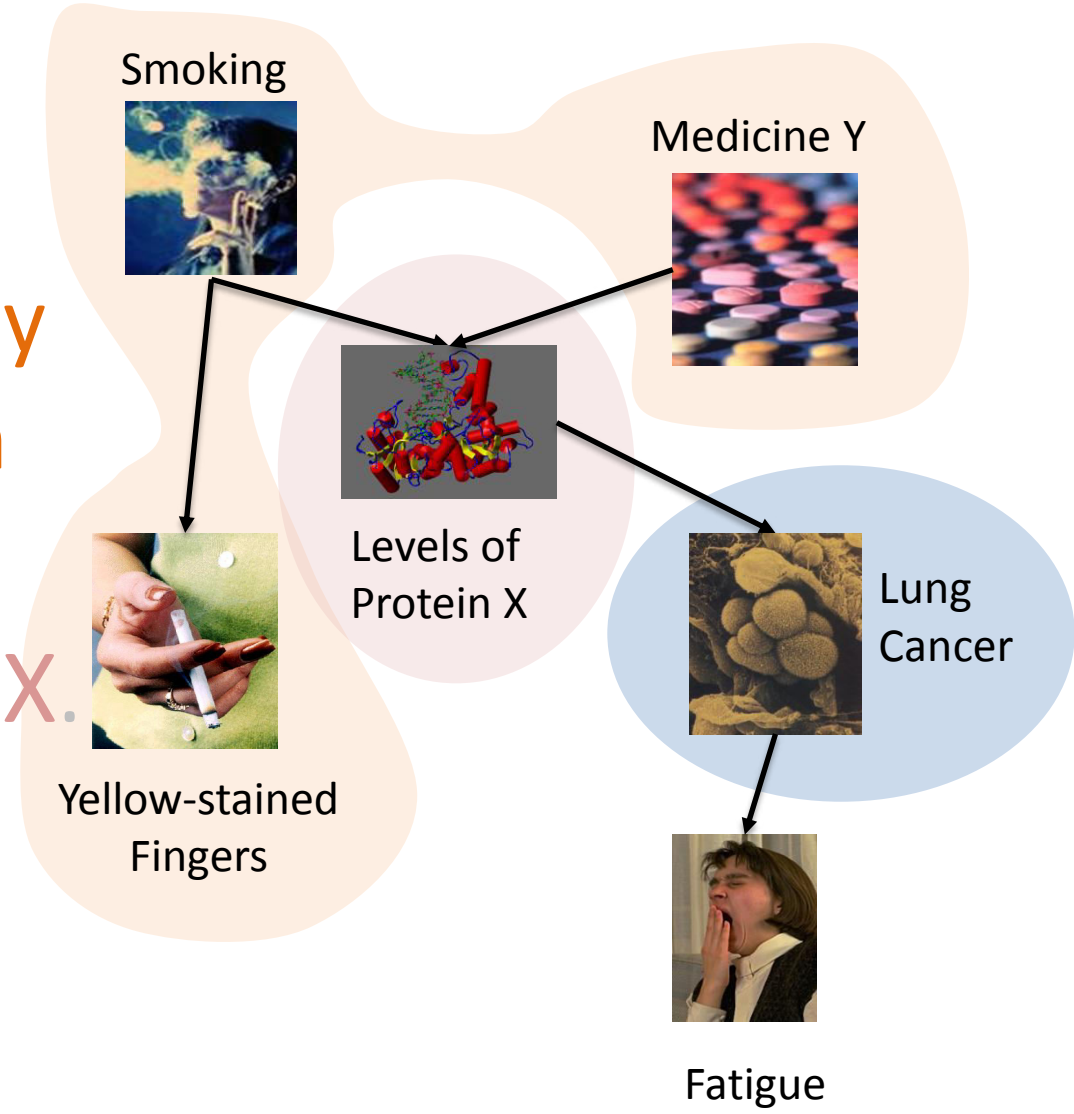
Reasoning with the CMC

Lung Cancer is independent of its non-effects given its Levels of Protein X



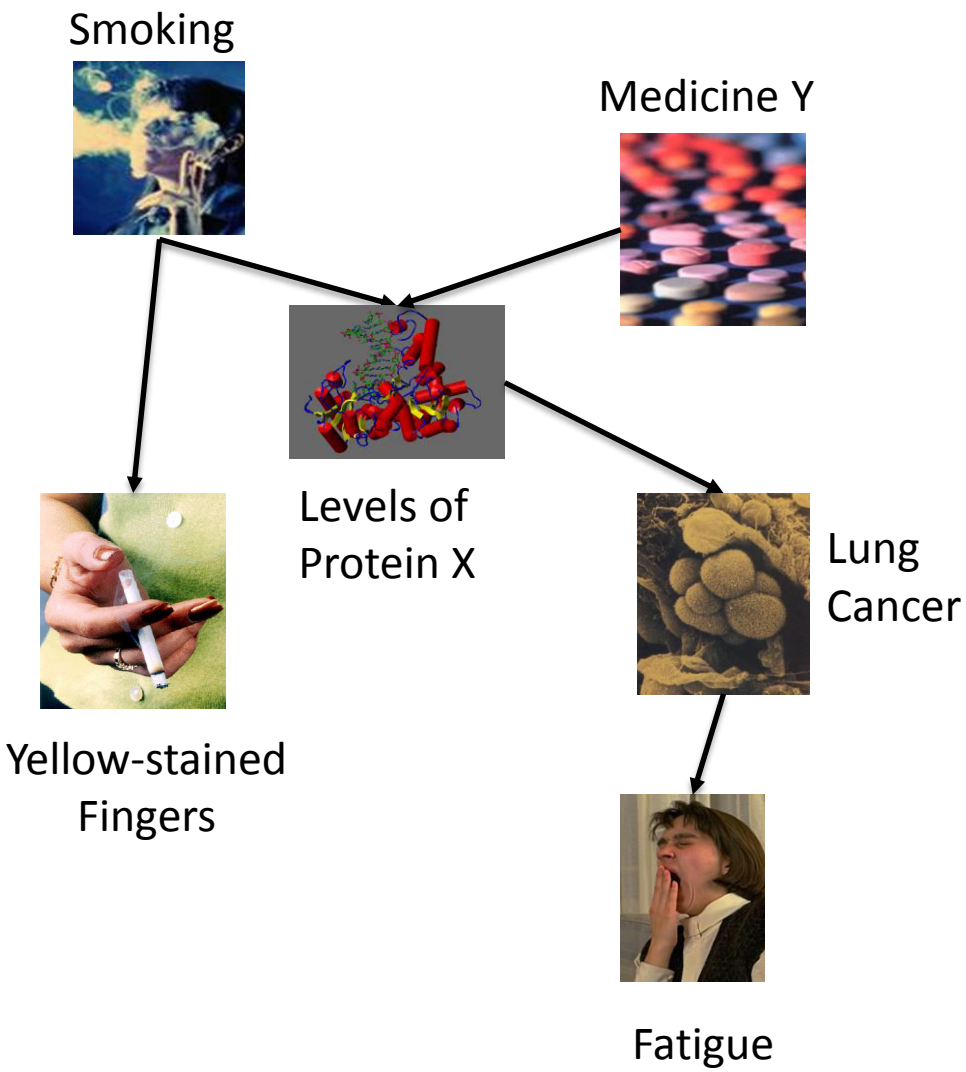
Reasoning with the CMC

Lung Cancer is independent of any variable other than Fatigue given the Levels of Protein X.



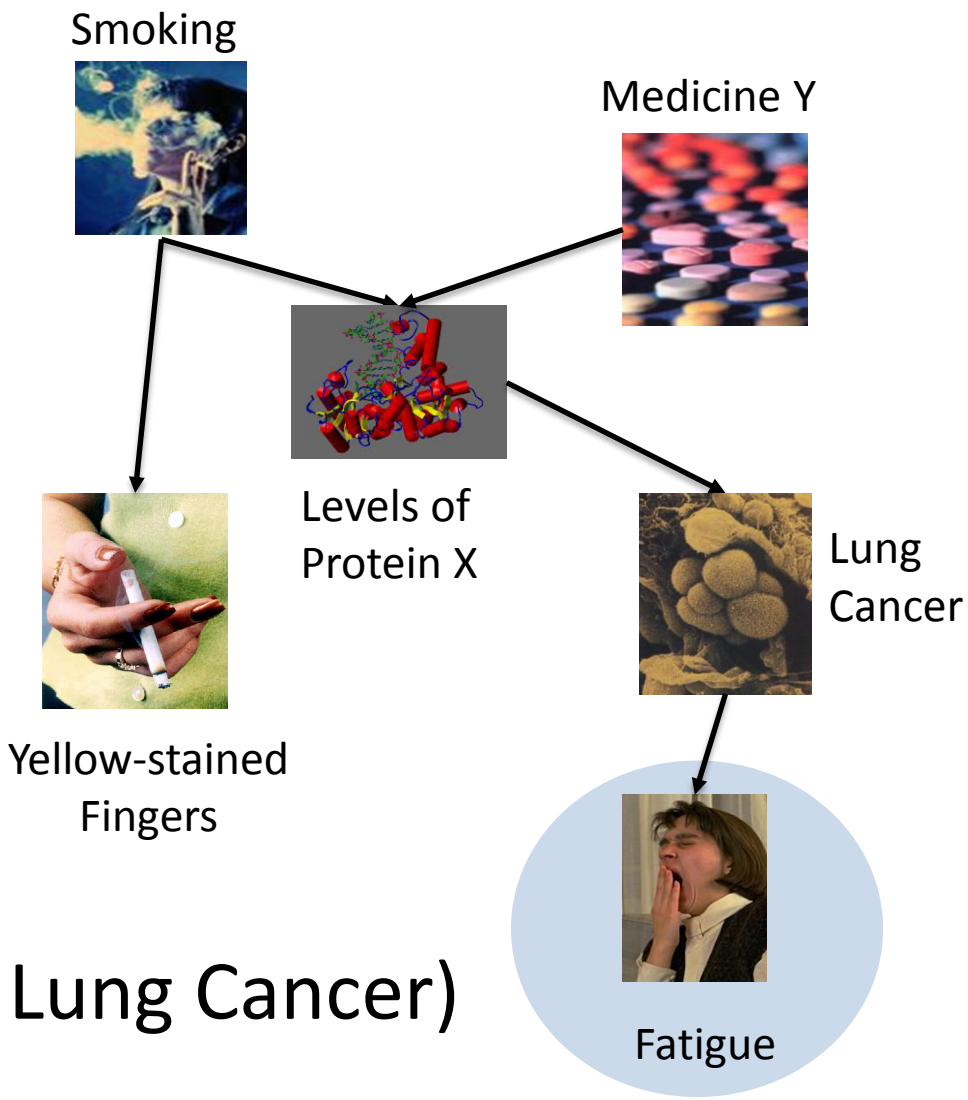
Entailed Independencies

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



Entailed Independencies

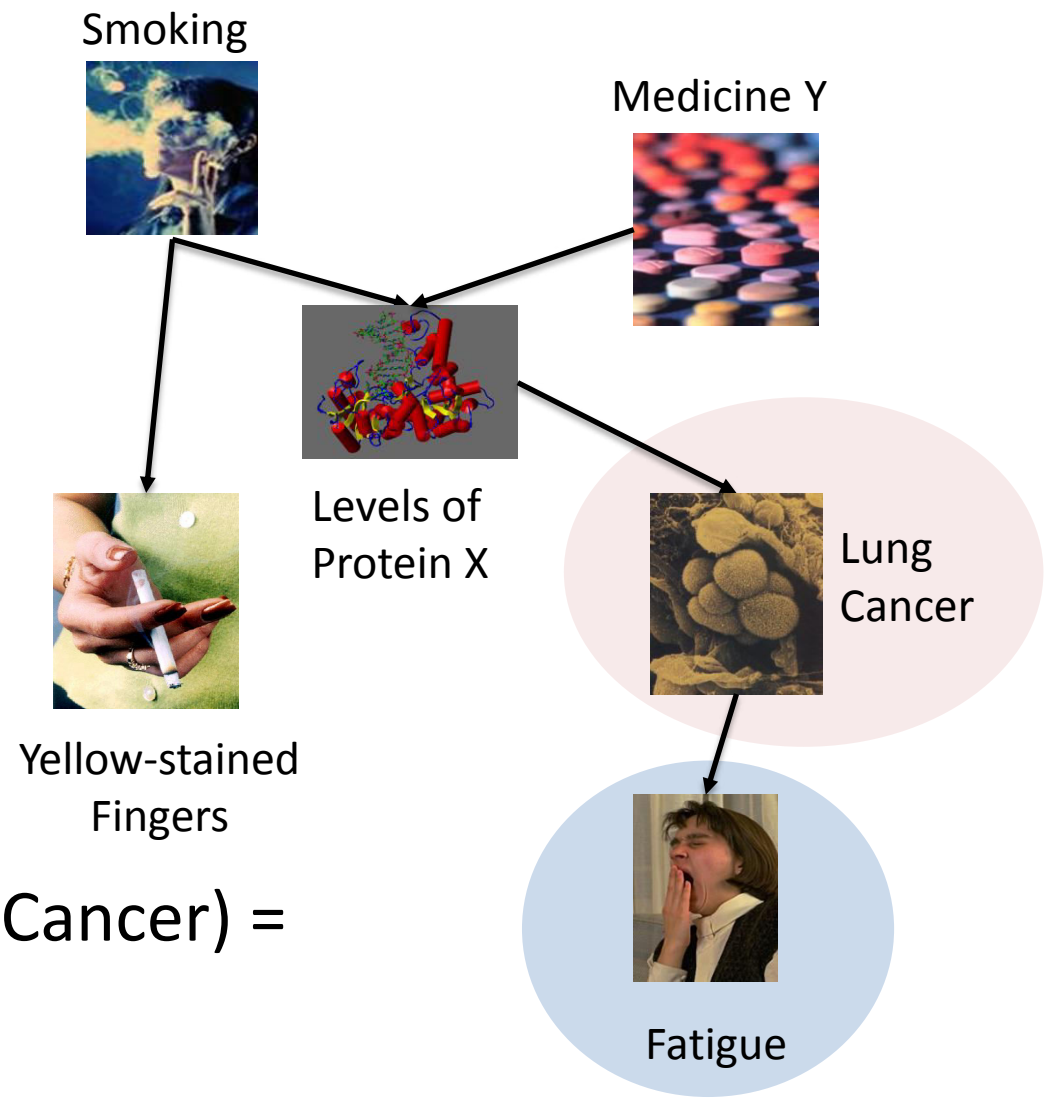
Fatigue is independent of its non-effects given its direct causes.



Ind(Fatigue, Smoking| Lung Cancer)

Entailed Independencies

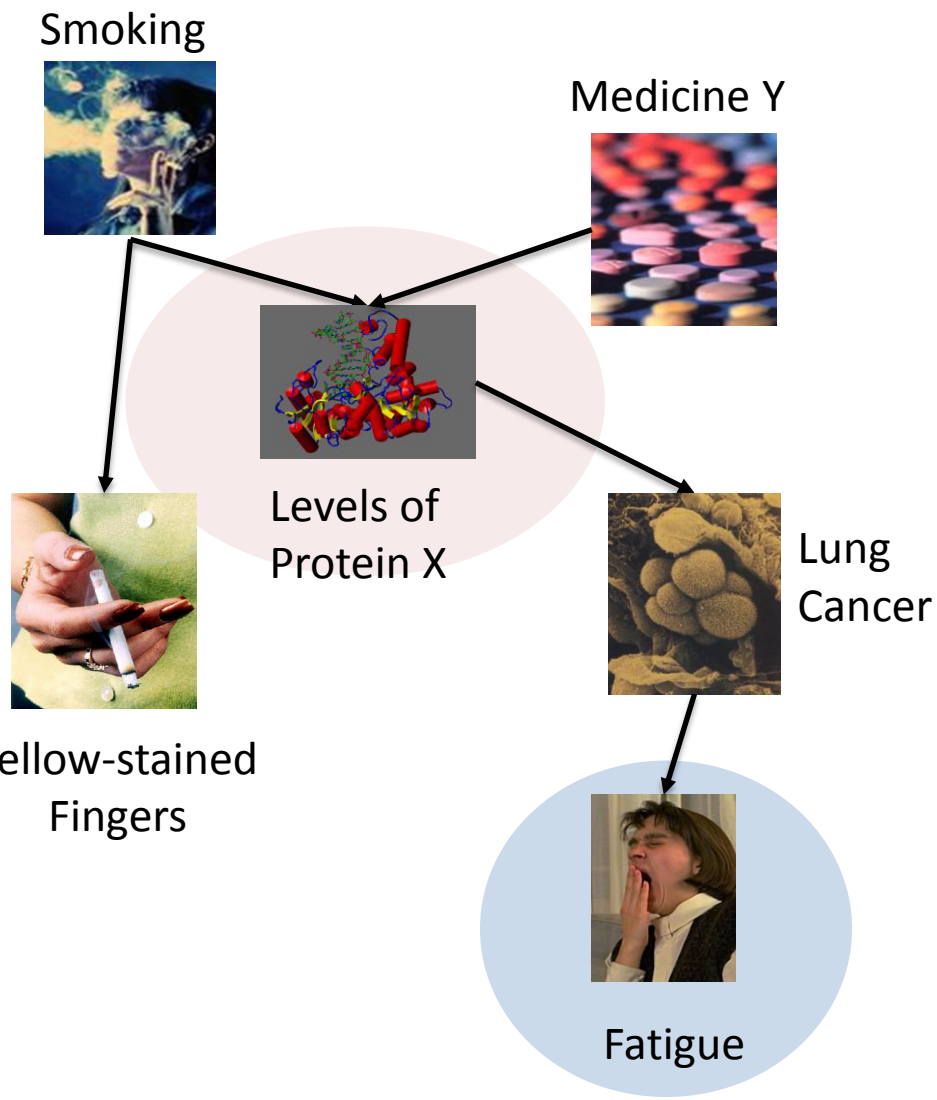
Fatigue is independent of Smoking given Lung Cancer



P(Fatigue| Smoking, Lung Cancer) = P(Fatigue| Lung Cancer)

Entailed Independencies

How about P(Fatigue| Smoking, Levels of Protein X)?



Entailed Independencies

P(Fatigue | Smoking, Levels of Protein X) =

$$P(F \mid S, X, LC=yes) \times P(LC= Yes \mid S, X) +$$
$$P(F \mid S, X, LC=no) \times P(LC = no \mid S, X) =$$

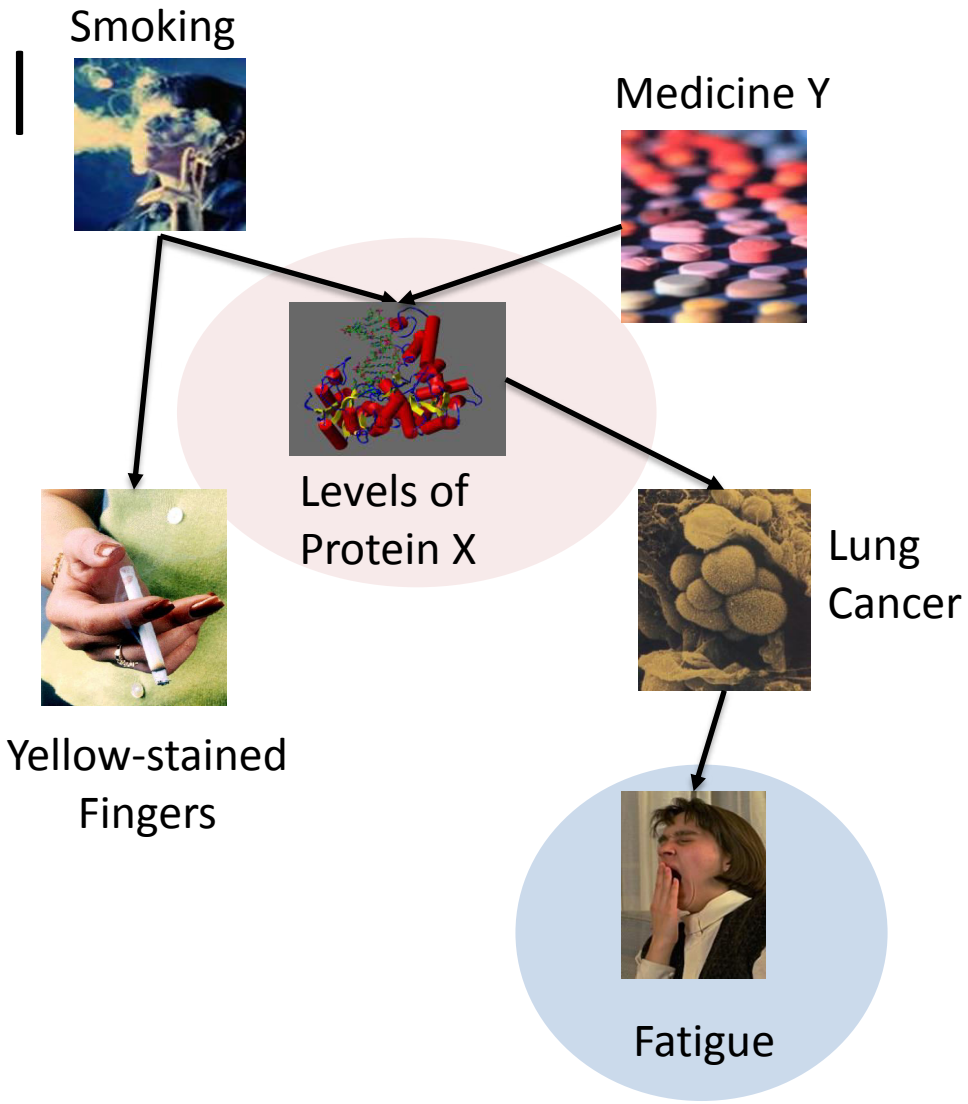
$$P(F \mid X, LC=yes) \times P(LC = Yes \mid X) +$$
$$P(F \mid X, LC=no) \times P(LC= no \mid X) =$$

P(Fatigue|Levels of Protein X)

Entailed Independencies

Ind(Fatigue, Smoking| Levels of Protein X)

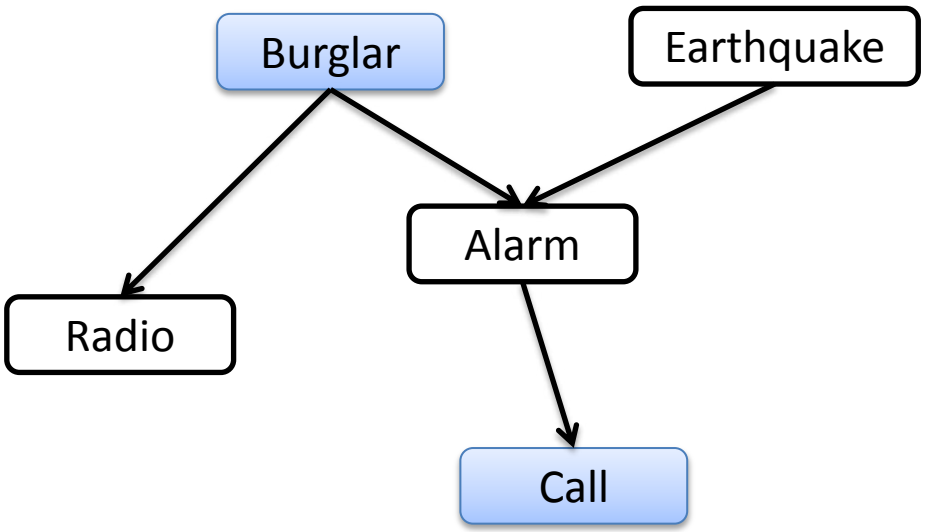
- What other independencies are *entailed* by the CMC?
- Do we have to do the math?



The d-separation criterion

- An algorithm to determine independencies that hold in a CBN.
- Let’s try to understand the intuition

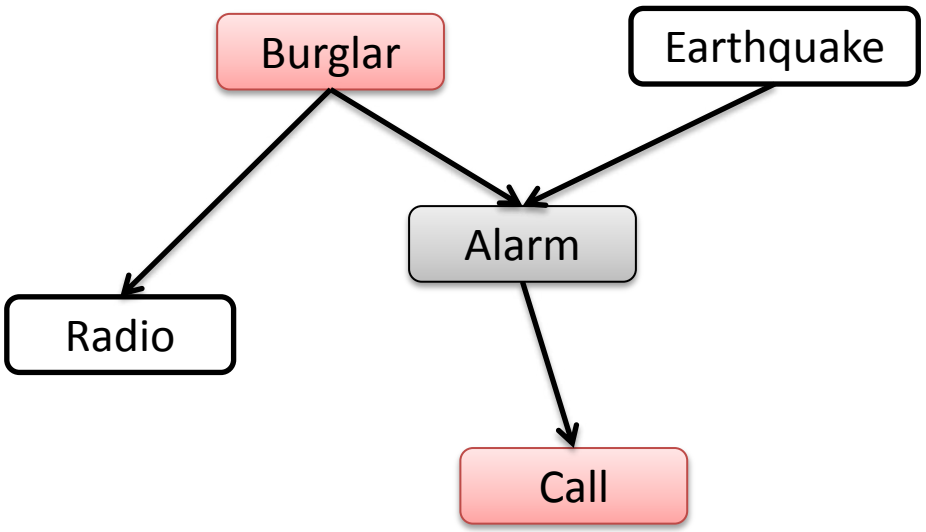
Open Paths



A causal path or a common cause is an open path (it allows information to flow)

Dep(Burglar, Call|∅)

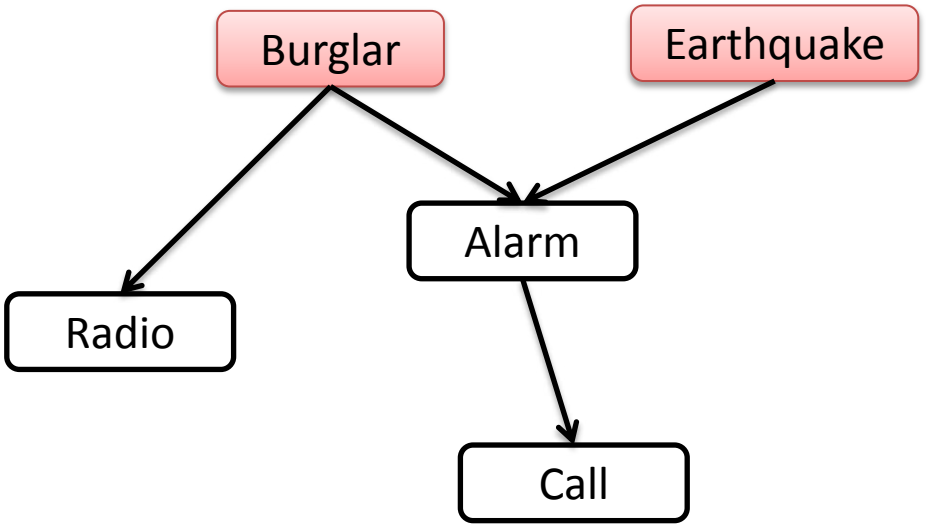
Blocked Paths



Conditioning on **intermediate** and **common** causes **blocks** the path

Ind (Burglar, Call | Alarm)

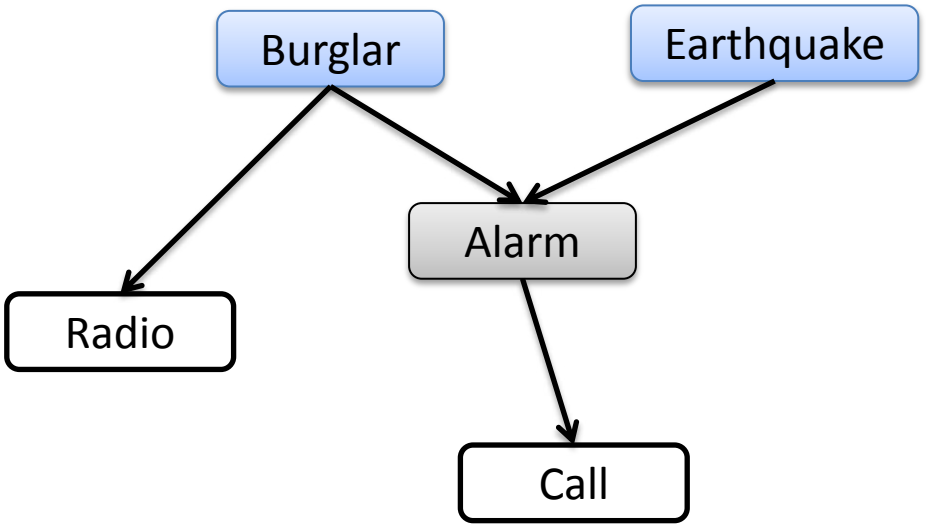
Blocked Paths



A path that goes through a common effect is a **blocked** path (no information flows).

Ind (Burglar, Earthquake|∅)

Open Paths



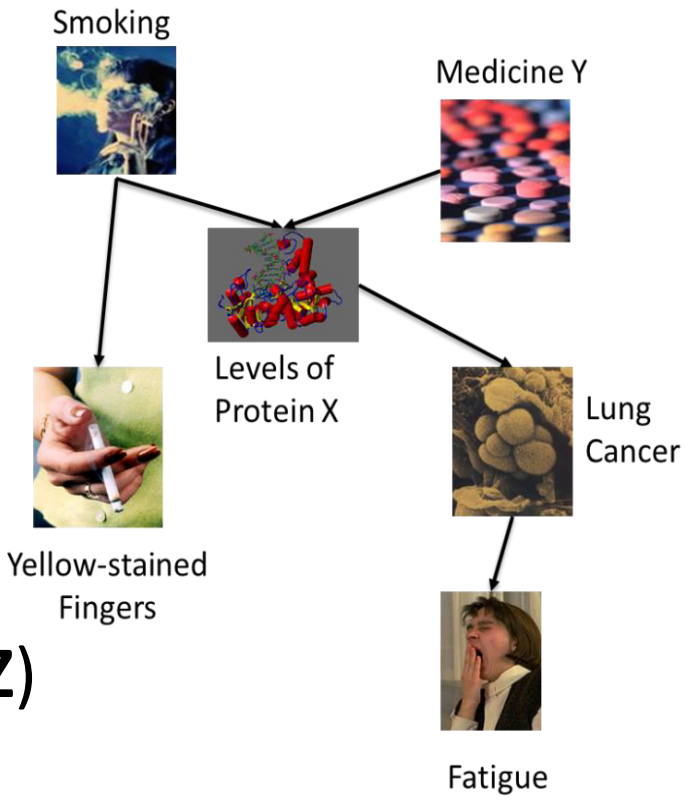
However, conditioning on common effects **opens** the path (information flows through the path.)

Dep (Burglar, Earthquake| Alarm)

The d-separation criterion

- You want to know if Ind(A, B|**Z**) is entailed by the CMC in a CBN

1. Find the paths from A to B, regardless orientation
2. If there exists no open path * conditioned on Z, then Ind(A, B|**Z**)

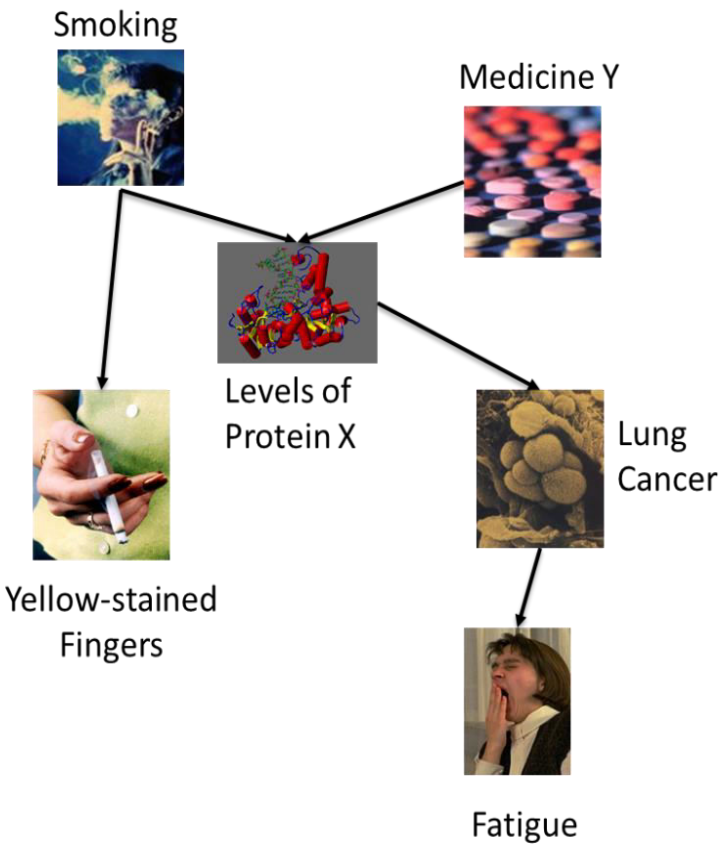


*symb. DSep(A, B|**Z**)

The d-separation criterion

- You want to know if Ind(A, B|**Z**) is entailed by the CMC in a CBN

1. Find the paths from A to B, regardless orientation
2. If there exists no open path conditioned on **Z** Ind(A, B|**Z**)
3. Else?**



symb. DConn(A, B|Z**)

Faithfulness

In all CBNs

$DSep(A, B|\mathbf{Z}) \Rightarrow Ind(A, B|\mathbf{Z})$

In faithful CBNs

$DSep(A, B|\mathbf{Z}) \Leftrightarrow Ind(A, B|\mathbf{Z})$

Faithfulness

- What does it mean really?
- The causal structure fully determines the independencies; independencies are not accidental
- Infinitesimal perturbations of the probabilities will not change the independencies (stability)

No independencies due to the particular parameters of the conditional probability tables (e.g., associations from different paths cancelling out)

Faithfulness

- Is it realistic?
- Assume you are given a graph and you select the parameters of the conditional probability tables randomly following a Dirichlet distribution
- The probability you get a non-faithful BN are zero (Lebesque measure is zero)
- Helpful to devise efficient asymptotically correct methods

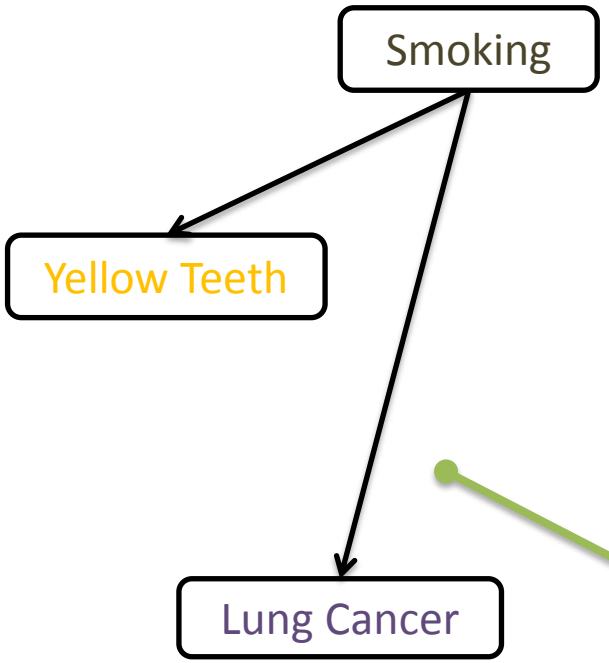
[Meek, C, UAI 1995]

Faithfulness

- Is it realistic?
- Too low associations:
  - For finite sample, they are not detectable and may lead to non-faithfulness (for all practical purposes)
- Too high correlations (determinism or close-to-determinism)
  - May lead to non-faithfulness
- Natural selection may be biasing towards creating non-faithful distributions in systems in nature (e.g., cells)!

Causal Bayesian Networks

Graph $\mathcal{G}$

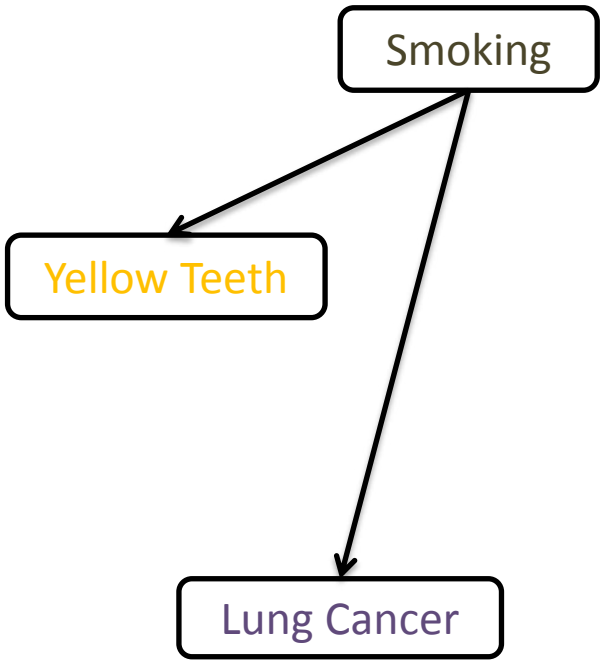


JPD J

		Lung Cancer	
Smoking	Yellow Teeth	Yes	No
Yes	Yes	0,01	0,04
Yes	No	0,01	0,04
No	Yes	0,000045	0,044955
No	No	0,000855	0,854145

1. Causal Markov Condition
2. Faithfulness

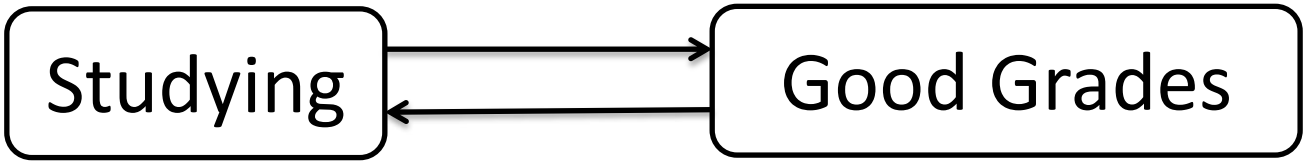
Causal Markov Condition (CMC)



		Lung Cancer	
Smoking	Yellow Teeth	Yes	No
Yes	Yes	0,01	0,04
Yes	No	0,01	0,04
No	Yes	0,000045	0,044955
No	No	0,000855	0,854145

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

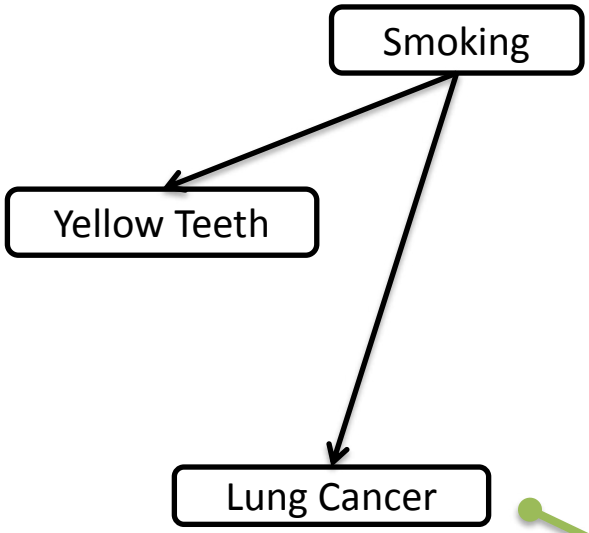
Causal Markov Condition (CMC)



- Studying causes Good Grades causes more studying (at a later time!)…
- Hard to define without explicitly representing time
- If all relations are linear, we can assume we sample from the distribution of the equilibrium of the system when external factors are kept constant
  - Path-diagrams (Structural Equation Models with no measurement model part) allow such feedback loops
- If there is feedback and relations are not linear, there may be chaos, literally and metaphorically

Causal Bayesian Networks*

Directed Acyclic Graph $\mathcal{G}$

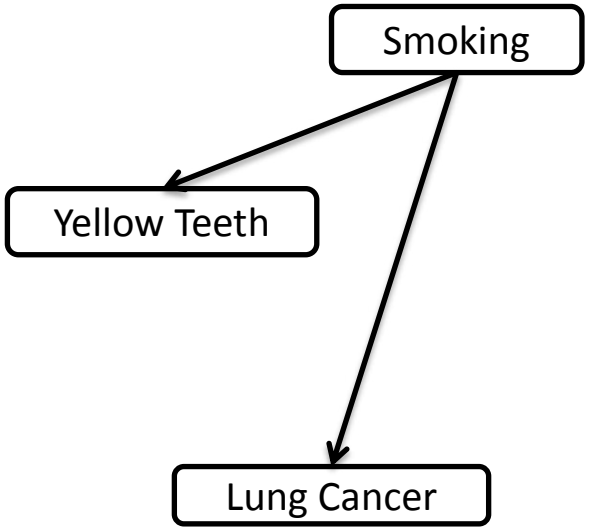


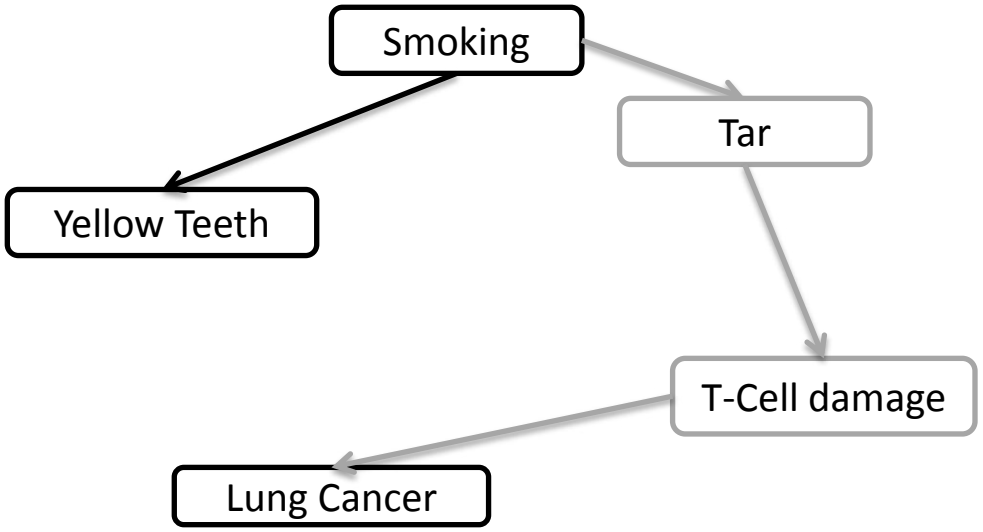
JPD J

		Lung Cancer	
Smoking	Yellow Teeth	Yes	No
Yes	Yes	0,01	0,04
Yes	No	0,01	0,04
No	Yes	0,0005	0,0450
No	No	0,0004	0,8541

1. Causal Markov Condition
2. Faithfulness
3. Acyclicity

Causal Interpretation



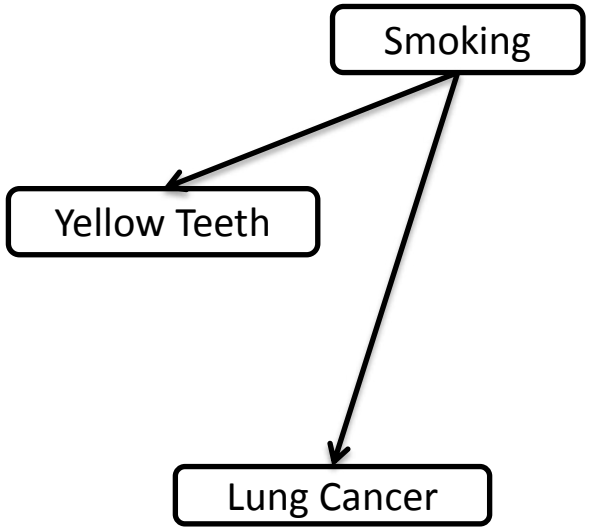


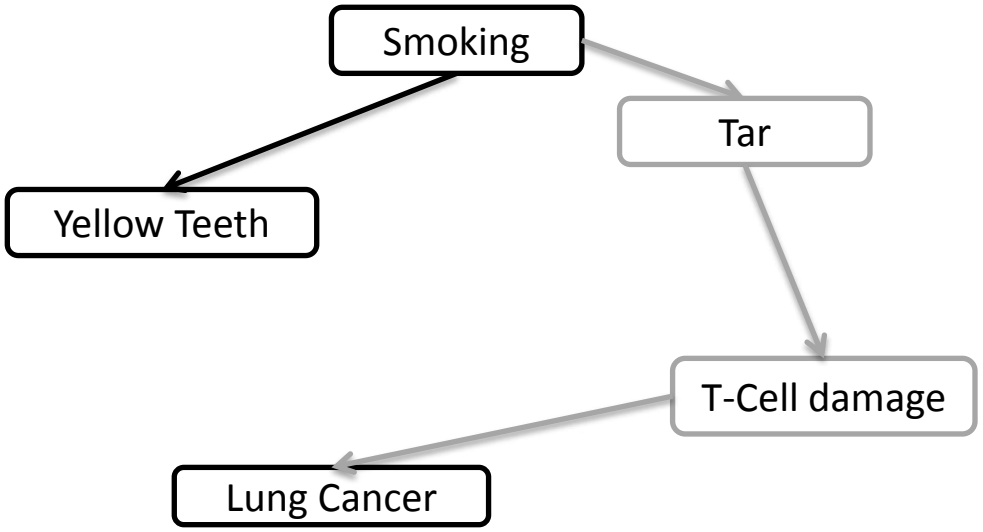
Smoking causes Lung Cancer

Smoking causes Tar Increase* causes T-Cell Damage* causes Lung Cancer

*unobserved

Causal Interpretation





causal relation and conditional independencies are modeled correctly

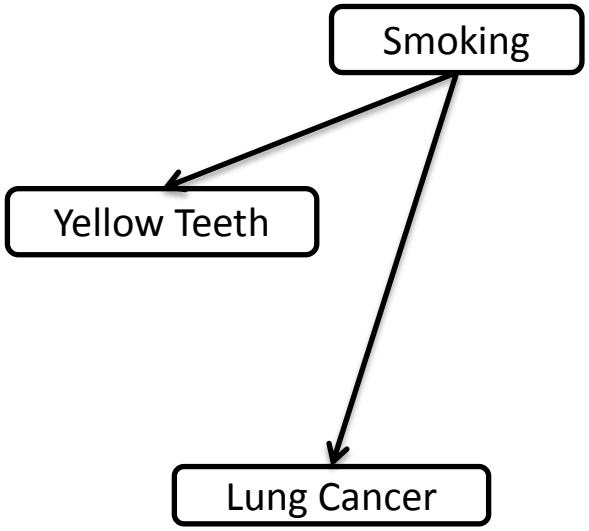
Smoking causes Lu

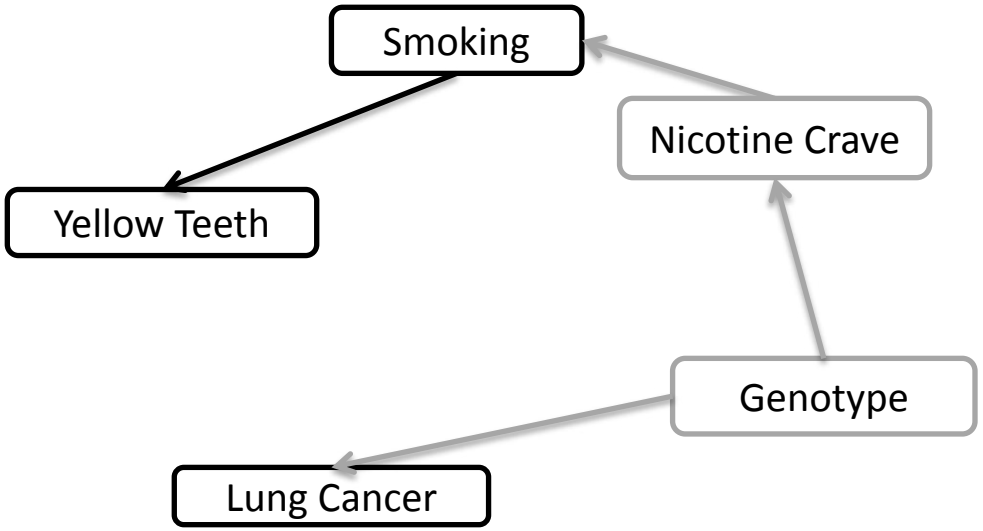
Increase*

ge* causes

*unobserved

Causal Interpretation





Smoking causes Lung Cancer

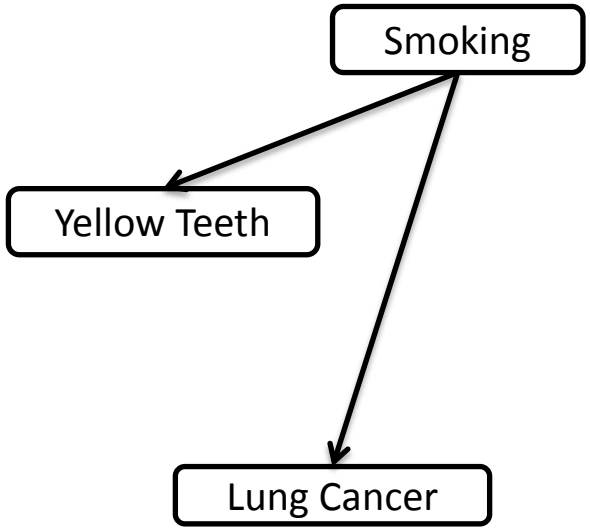
Genotype* causes Nicotine Crave*

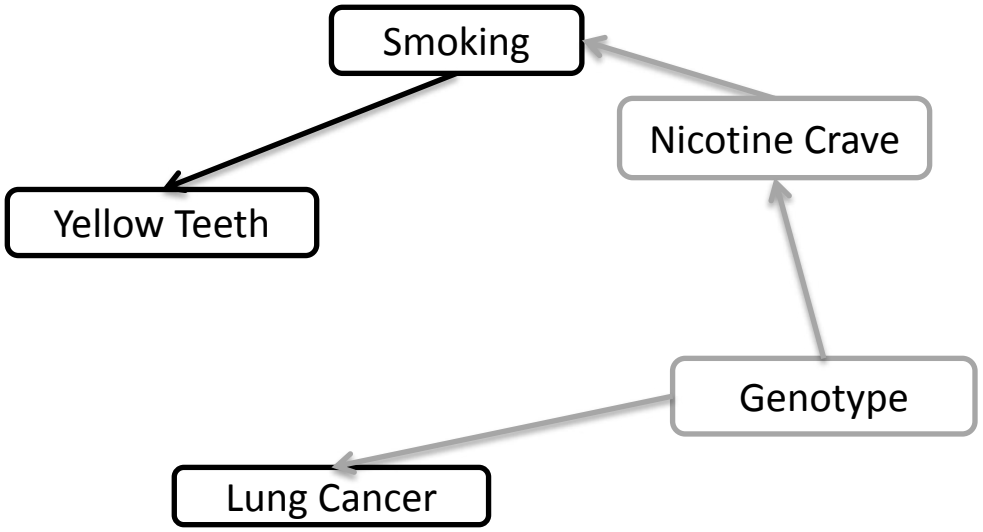
causes Smoking, Genotype*

causes Lung Cancer

*unobserved

Causal Interpretation





Smoking causes Lu

conditional independencies are modeled correctly, but **not causal relation**

Nicotine Crave*

notype*

*unobserved

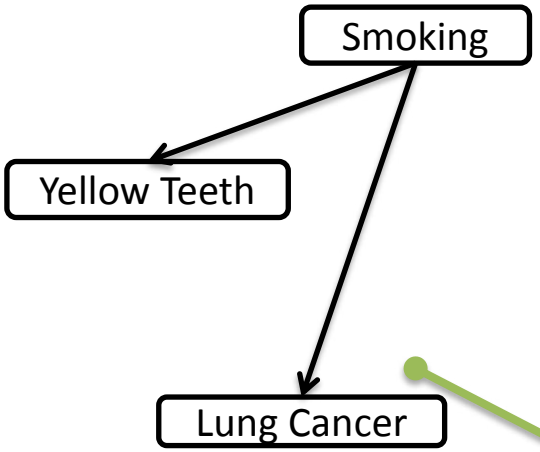
Causal Sufficiency

- Assume what is called Causal Sufficiency
  - No pair of variables has a latent (unobserved) common cause

- This is a pretty strong assumption
- Hidden confounders a major reason why some people debate we absolutely need experiments

Causal Bayesian Networks

DAG $\mathcal{G}$



JPD J

Smoking	Yellow Teeth	Lung Cancer: Yes	Lung Cancer: No
Yes	Yes	0,01	0,04
Yes	No	0,01	0,04
No	Yes	0,0005	0,0450
No	No	0,0004	0,8541

1. Causal Markov Condition
2. Faithfulness
3. Acyclicity
4. Causal Sufficiency

Bayesian Networks

I don’t like all these assumptions

I kind of liked reducing the parameters of the distribution

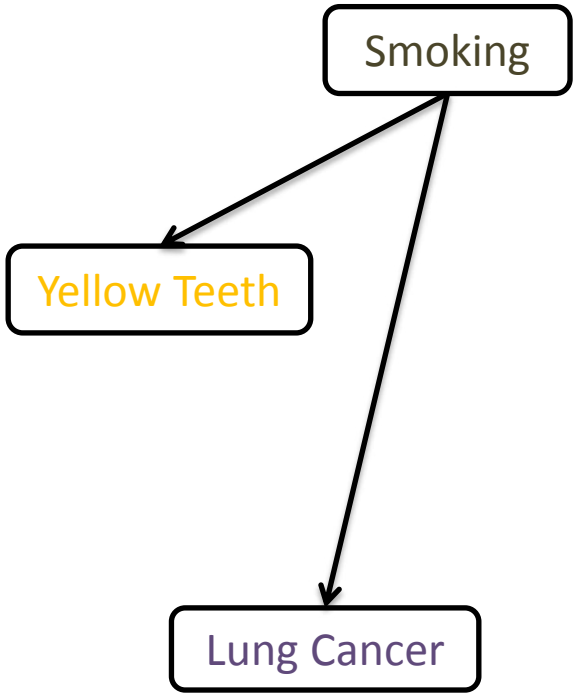
Bayesian Networks

I don’t like all these assumptions

I kind of liked reducing the parameters of the distribution

Drop the Causal part!

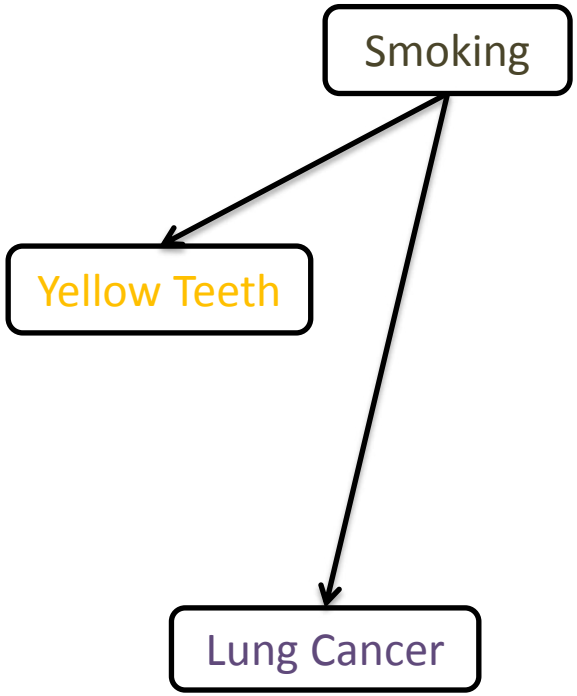
Causal Markov Condition (CMC)



		Lung Cancer	
Smoking	Yellow Teeth	Yes	No
Yes	Yes	0,01	0,04
Yes	No	0,01	0,04
No	Yes	0,000045	0,044955
No	No	0,000855	0,854145

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

Markov Condition (MC)



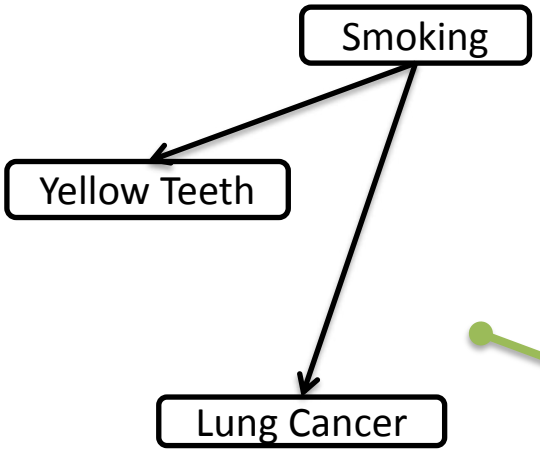
		Lung Cancer	
Smoking	Yellow Teeth	Yes	No
Yes	Yes	0,01	0,04
Yes	No	0,01	0,04
No	Yes	0,000045	0,044955
No	No	0,000855	0,854145

Every variable is independent of its non-descendants given its parents.

(can always be made to hold by adding more edges)

Bayesian Networks

DAG $\mathcal{G}$

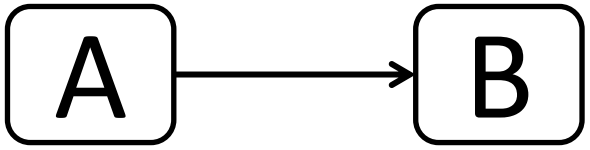


JPD J

Smoking	Yellow Teeth	Lung Cancer: Yes	Lung Cancer: No
Yes	Yes	0,01	0,04
Yes	No	0,01	0,04
No	Yes	0,0005	0,0450
No	No	0,0004	0,8541

1. Markov Condition
2. Faithfulness
3. Acyclicity

Semantics



Causal Bayesian Networks:

A causes B

Bayesian Networks:

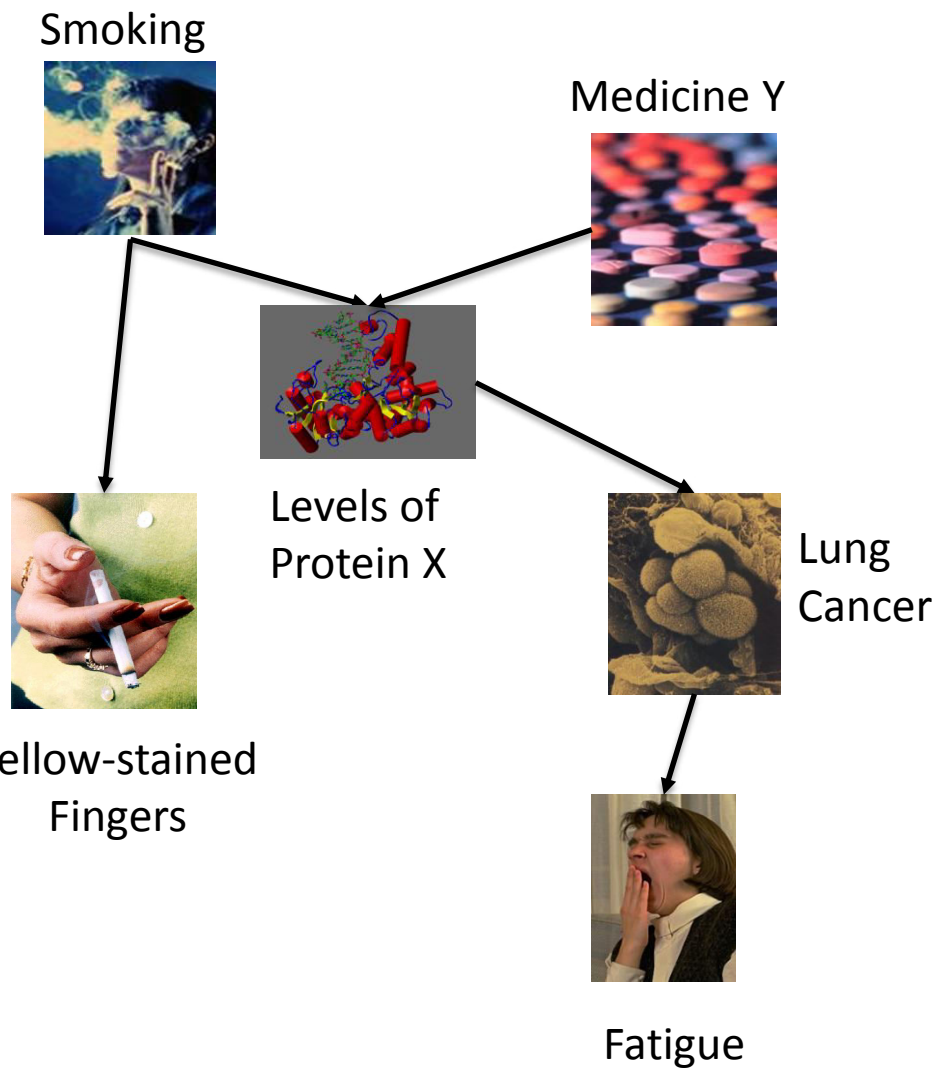
A gives unique information for B

Using a Bayesian Network

1. Factorize the jpd

2. Answer questions like:
   1. P(Lung Cancer| Levels of Protein X) = ?

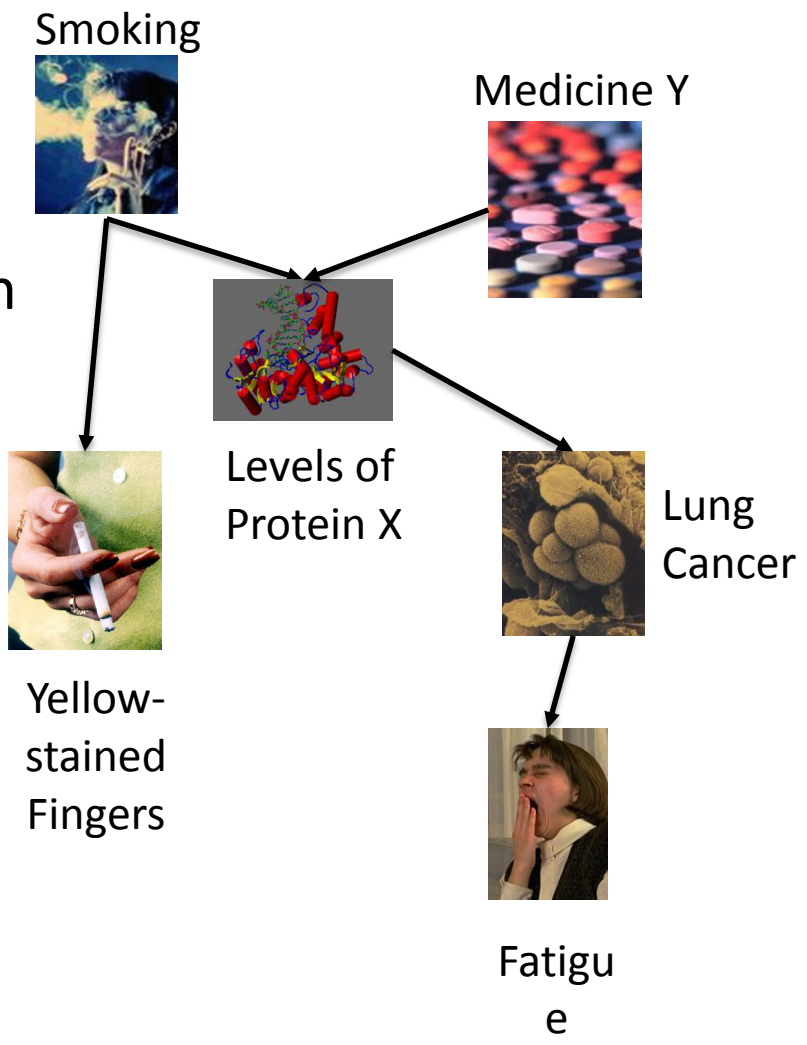
   2. Ind(Smoking, Fatigue| Levels of Protein X)?



Using a **Causal** Bayesian Network

1. Factorize the jpd

2. Answer questions like:
   1. P(Lung Cancer| Levels of Protein X) = ?
   2. Ind(Smoking, Fatigue| Levels of Protein X)?
   3. What will happen if I design a drug that blocks the function of protein X (**predict effect of interventions**)?



A Closer Look at the Assumptions

	Causal	Non-Causal
(Causal) Markov Condition	Connects structure with effect of interventions	Can always be made to hold by adding edges
Faithfulness	Required for (relatively) efficient learning Facilitates characterization of equivalent networks	Required for (relatively) efficient learning Facilitates characterization of equivalent networks
Acyclicity	Causal-feedback loops create problems with the semantics and reasoning	Acyclic graphs can still encode all distributions (not restrictive)
Causal Sufficiency	Required to causally interpret an edge	Not required

Conclusions, Pt.2

- Causal Bayesian Networks quantitatively represent the probabilistic causal relations among a set of variables
- Dependencies and independencies can be read-off the graph using the *d*-separation
- Typical assumptions:
  - Causal Markov Condition
  - Faithfulness
  - No feedback loops (acyclicity)
  - Causal Sufficiency
- Bayesian Networks:
  - inspired by causality but drop the causal claims
  - Thousands of applications
  - Fewer restrictive assumptions
  - Appropriate when probabilistic reasoning is the goal, not causal reasoning (predict effect of interventions)

Bibliography

- Spirtes, P., Glymour, C. & Scheines, R., 2001. *Causation, Prediction, and Search* Second Edi., The MIT Press.
- Neapolitan, R., 2003. *Learning Bayesian Networks*, Prentice Hall
- Meek, C., 1995. Strong completeness and faithfulness in Bayesian networks. In *UAI 1995*

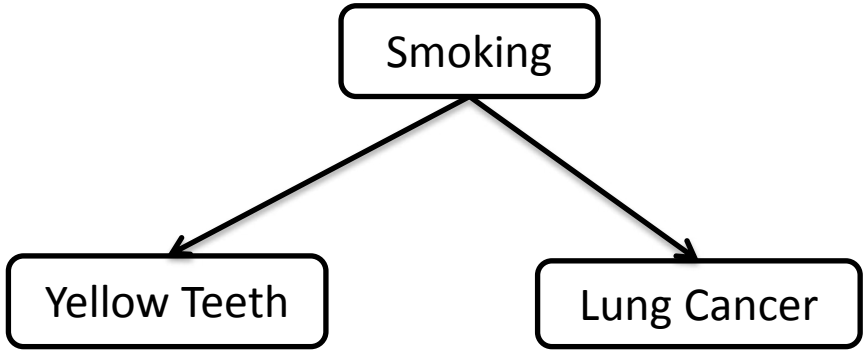
Pt. 3 - Learning Bayesian Networks

Often the case

Raw data

Subject #	Smoking	Yellowed Fingers	Levels of Protein X	Fatigue	Medicine Y	Lung Cancer
1	0	0	0.23	1	1	0
2	1	1	0.1	1	0	0
3	1	1	0.7	0	0	0
4	0	1	0.92	1	0	0
5	0	0	1.2	0	1	0
6	1	1	1.4	0	1	0
7	1	1	5.4	1	1	1
8	0	1	0.89	1	0	0
9	0	0	0.7	1	1	1
10	1	1	0.56	0	1	0
11	1	1	0.16	0	1	0
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
10000	1	1	3.2	1	1	1

Observing a causal model



What would you observe?

- Dep(Lung Cancer, Yellow Teeth|∅)
- Dep(Smoking, Lung Cancer |∅)
- Dep(Lung Cancer, Yellow Teeth |∅)
- Ind(Lung Cancer, Yellow Teeth| Smoking)

Raw data

Subject #	Smoking	Yellow Teeth	Lung Cancer
1	0	0	0
2	1	1	0
3	1	1	0
4	0	1	0
5	0	0	0
6	1	1	0
7	1	1	1
8	0	1	0
9	0	0	0
10	1	1	0
11	1	1	0
.	.	.	.
.	.	.	.
.	.	.	.
10000	1	1	1

Learning the network

Constraint-Based Approach	Score-Based (Bayesian)
Test conditional independencies in data and find a faithful DAG that encodes them (with d-separations)	Find the DAG with the maximum a posteriori probability given the data

Learning the network

Constraint-Based Approach

- Easier to extend to different types of data (e.g., survival)
- Easier to extend to networks with latent variables (MAGs)
- Easier to turn to local (learn parts of the network)

Score-Based

- Robust to small samples
- Incorporates priors on the networks
- Better in identifying the edge orientations (personal experience)

Learning the network

Constraint-Based Approach

- SGS [Spirtes, Glymour, & Scheines 2000]
- PC [Spirtes, Glymour, & Scheines 2000]
- TPDA [Cheng et al., 1997]
- CPC [Ramsey et al, 2006]

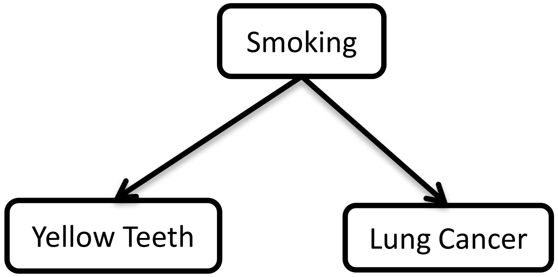
Hybrid

- MMHC [Tsamardinoas et al. 2006]
- CB [Provan et al. 1995]
- BENEDICT [Provan and de Campos 2001]
- ECOS [Kaname et al. 2010]

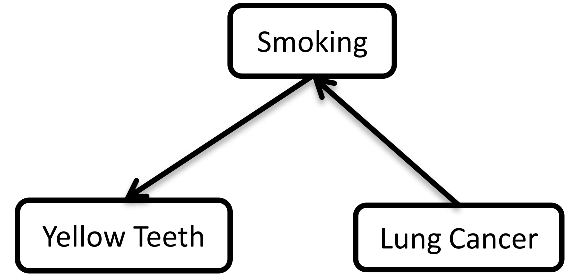
Bayesian Approach

- K2 [Cooper and Herskowitz 1992]
- GBPS [Spirtes and Meek 1995]
- GES [Chickering and Meek 2002]
- Sparse Candidate [Friedman et al. 1999]
- Optimal Reinsertion [Moore and Wong 2003]
- Rec [Xie, X, Geng, Zhi, JMLR 2008]
- Exact Algorithms [Koivisto et al., 2004] , [Koivisto, 2006] , [Silander & Myllymaki, 2006]

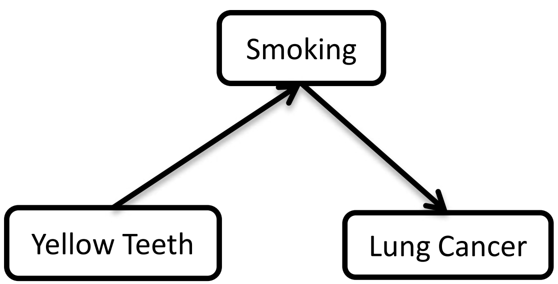
Learning the Network



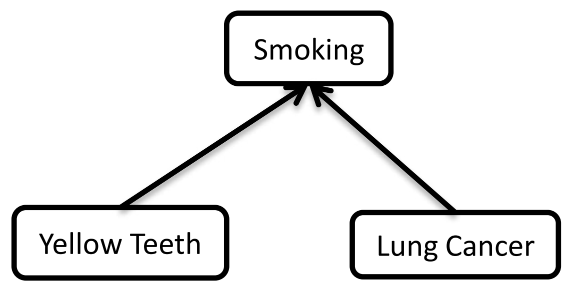
- Dep(Lung Cancer, Yellow Teeth |∅)
- Dep(Smoking, Lung Cancer |∅)
- Dep(Lung Cancer, Yellow Teeth |∅)
- Ind(Lung Cancer, Yellow Teeth| Smoking)



- Dep(Lung Cancer, Yellow Teeth |∅)
- Dep(Smoking, Lung Cancer |∅)
- Dep(Lung Cancer, Yellow Teeth |∅)
- Ind(Lung Cancer, Yellow Teeth| Smoking)

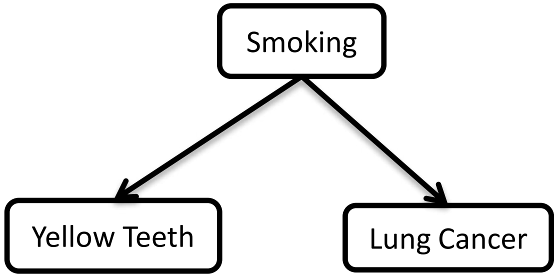


- Dep(Lung Cancer, Yellow Teeth |∅)
- Dep(Smoking, Lung Cancer |∅)
- Dep(Lung Cancer, Yellow Teeth |∅)
- Ind(Lung Cancer, Yellow Teeth| Smoking)

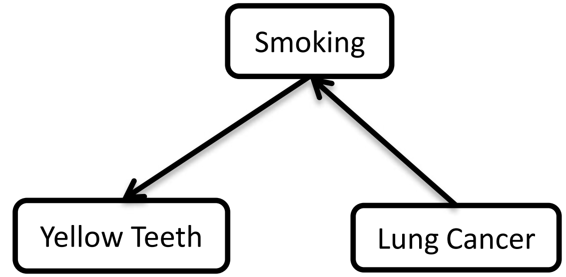


- Dep(Lung Cancer, Yellow Teeth |∅)
- Dep(Smoking, Lung Cancer |∅)
- Ind(Lung Cancer, Yellow Teeth |∅)
- Dep(Lung Cancer, Yellow Teeth| Smoking)

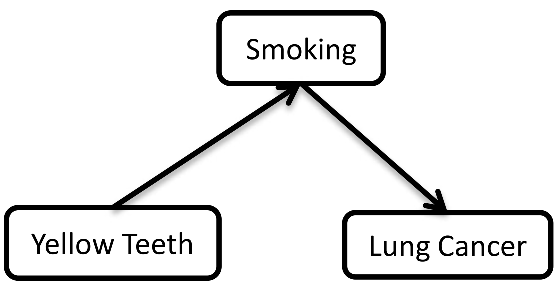
Learning the Network



- Dep(Lung Cancer, Yellow Teeth |∅)
- Dep(Smoking, Lung Cancer |∅)
- Dep(Lung Cancer, Yellow Teeth |∅)
- Ind(Lung Cancer, Yellow Teeth| Smoking)

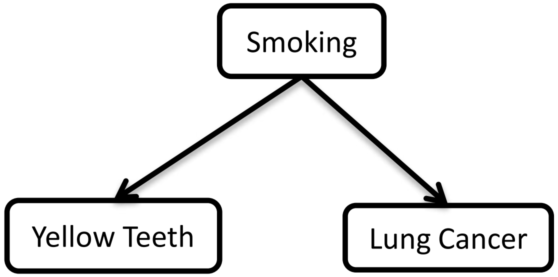


- Dep(Lung Cancer, Yellow Teeth |∅)
- Dep(Smoking, Lung Cancer |∅)
- Dep(Lung Cancer, Yellow Teeth |∅)
- Ind(Lung Cancer, Yellow Teeth| Smoking)

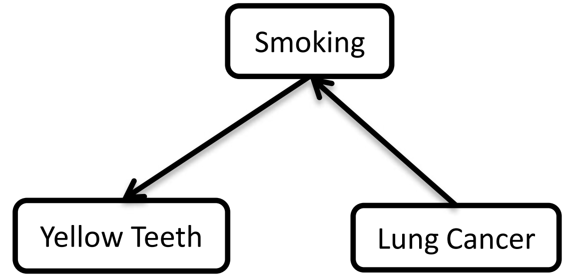


- Dep(Lung Cancer, Yellow Teeth |∅)
- Dep(Smoking, Lung Cancer |∅)
- Dep(Lung Cancer, Yellow Teeth |∅)
- Ind(Lung Cancer, Yellow Teeth| Smoking)

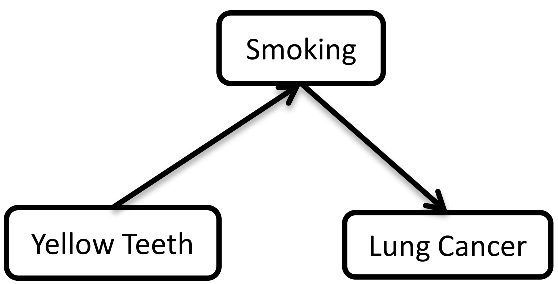
Markov Equivalence



- Dep(Lung Cancer, Yellow Teeth |∅)
- Dep(Smoking, Lung Cancer |∅)
- Dep(Lung Cancer, Yellow Teeth |∅)
- Ind(Lung Cancer, Yellow Teeth| Smoking)



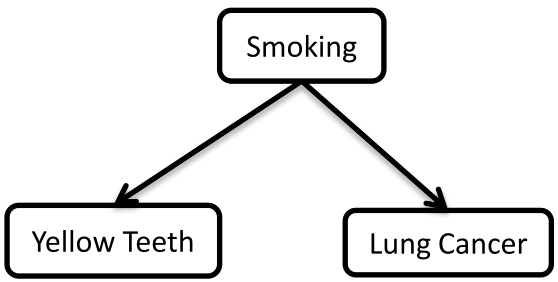
- Dep(Lung Cancer, Yellow Teeth |∅)
- Dep(Smoking, Lung Cancer |∅)
- Dep(Lung Cancer, Yellow Teeth |∅)
- Ind(Lung Cancer, Yellow Teeth| Smoking)



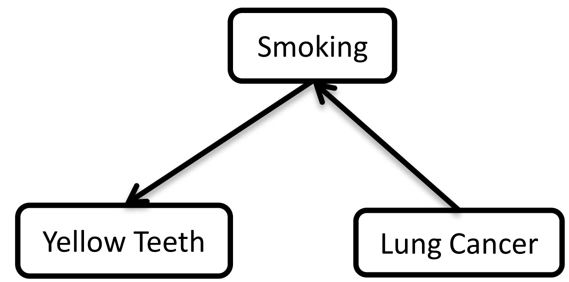
- Dep(Lung Cancer, Yellow Teeth |∅)
- Dep(Smoking, Lung Cancer |∅)
- Dep(Lung Cancer, Yellow Teeth |∅)
- Ind(Lung Cancer, Yellow Teeth| Smoking)

Two networks are Markov Equivalent if and only they entail (by the Markov Condition) the same set of independencies

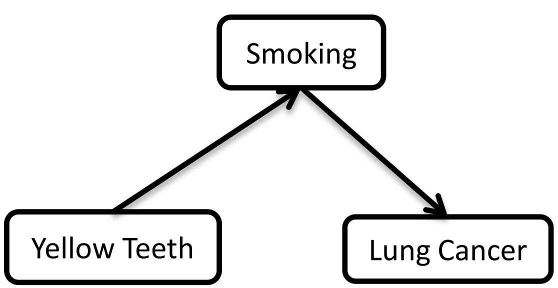
Markov Equivalence



- Dep(Lung Cancer, Yellow Teeth $|\varnothing$)
- Dep(Smoking, Lung Cancer $|\varnothing$)
- Dep(Lung Cancer, Yellow Teeth $|\varnothing$)
- Ind(Lung Cancer, Yellow Teeth| Smoking)



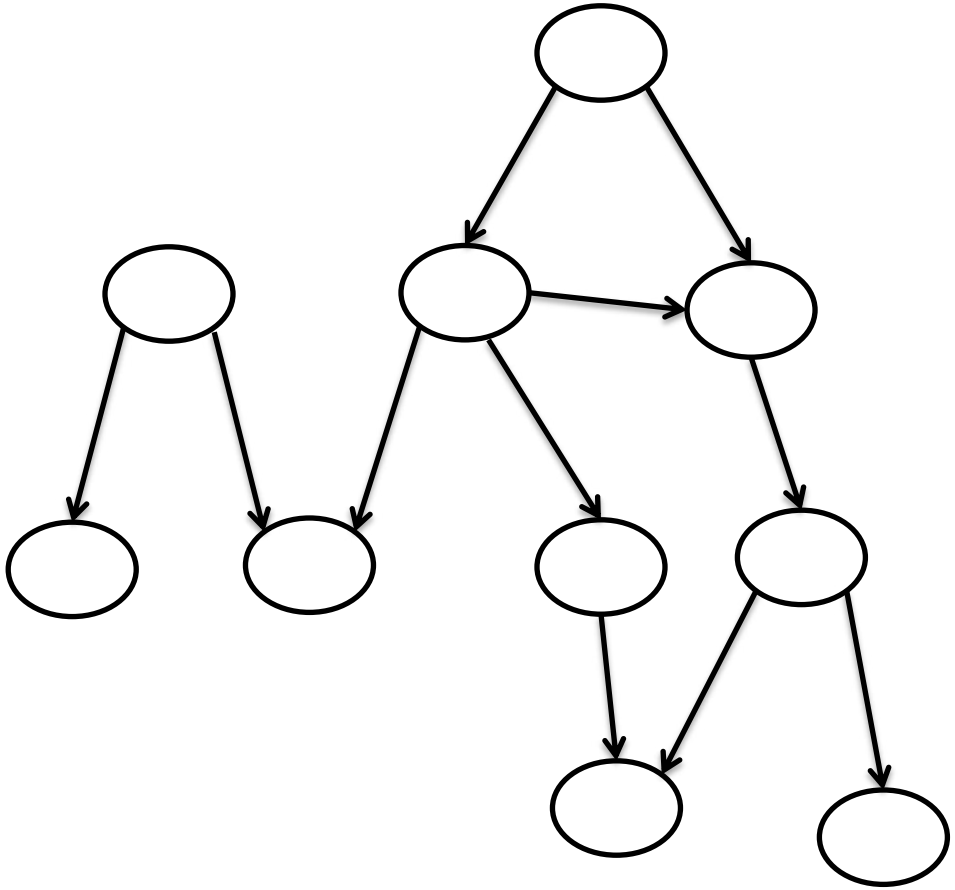
- Dep(Lung Cancer, Yellow Teeth $|\varnothing$)
- Dep(Smoking, Lung Cancer $|\varnothing$)
- Dep(Lung Cancer, Yellow Teeth $|\varnothing$)
- Ind(Lung Cancer, Yellow Teeth| Smoking)



- Dep(Lung Cancer, Yellow Teeth $|\varnothing$)
- Dep(Smoking, Lung Cancer $|\varnothing$)
- Dep(Lung Cancer, Yellow Teeth $|\varnothing$)
- Ind(Lung Cancer, Yellow Teeth| Smoking)

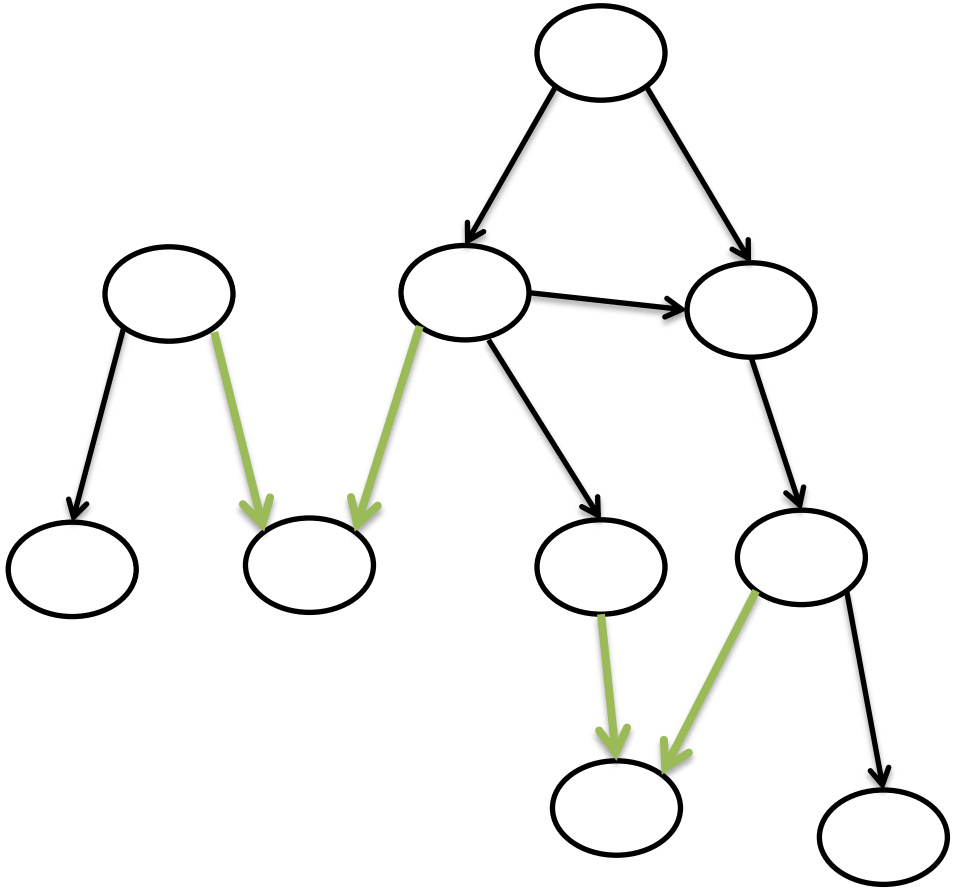
Two networks are Markov Equivalent if and only they have the same edges and the same v-structures

Markov Equivalence



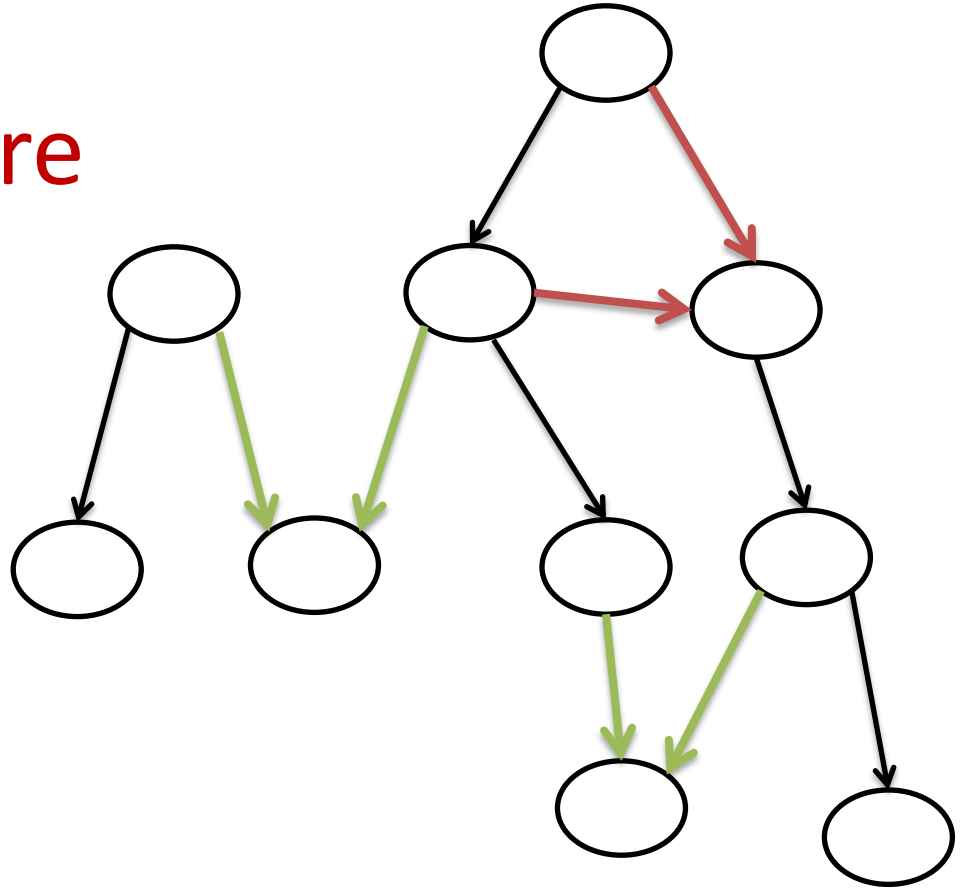
Markov Equivalence

- v-structures



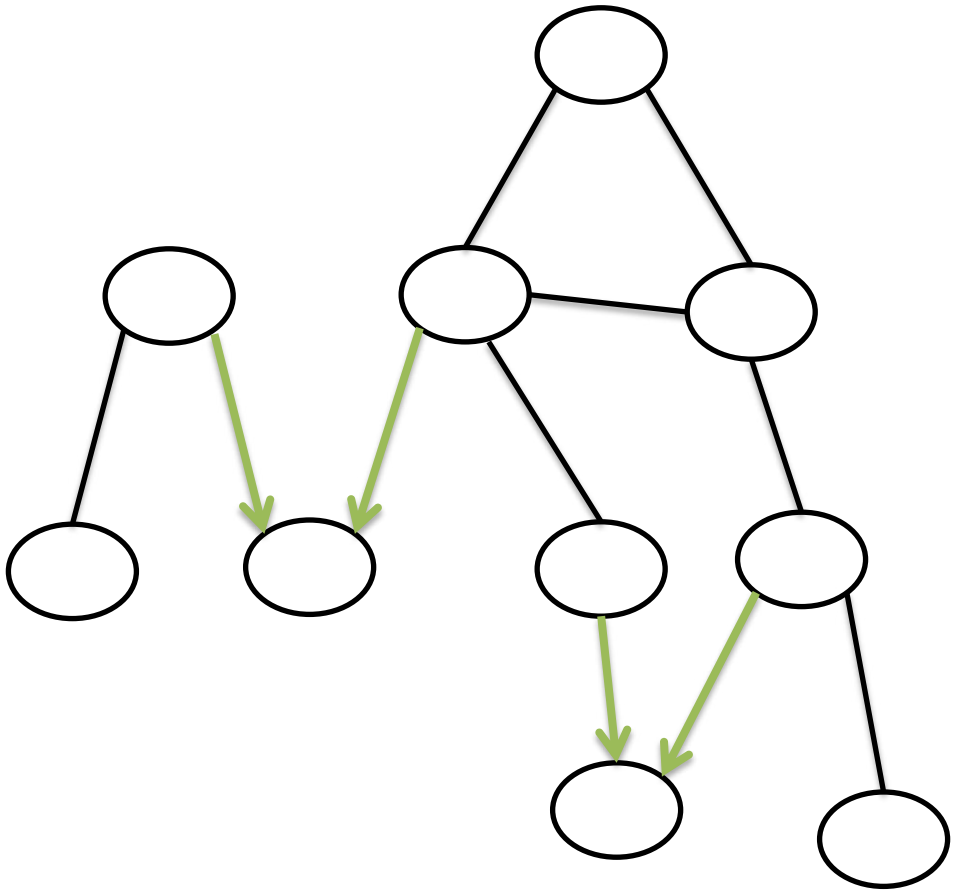
Markov Equivalence

- v-structures
- not a v-structure

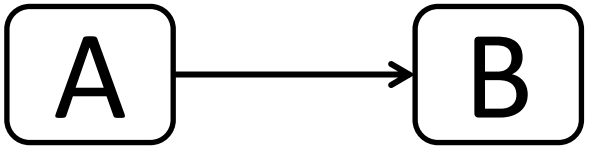


Pattern DAG (PDAG)

- Represents a class of Markov Equivalent DAGs
- Has the same edges as every other DAG in the class
- Has only orientations (arrows) shared by all the DAGs in the class



Semantics in PDAGs



A causes *B*

Either *A* causes *B*
or vice versa
Both cases fit the
data equally well

The PC Algorithm

- Learning the skeleton:
  - Iff there exists no set of variables **Z** s. t.

$$\text{Ind}(A, B \mid \mathbf{Z}^*) \quad A - B \text{ in } \mathcal{G}'$$

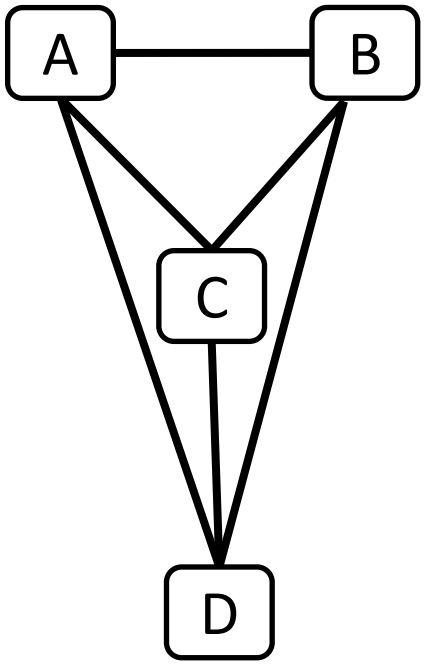
- Learning v-structures:
  - If $A - C - B$ and $\text{Ind}(A, B \mid \mathbf{Z})$, $C \cap \mathbf{Z} = \emptyset$,

$$A \rightarrow C \leftarrow B \text{ in } \mathcal{G}$$

- Perform all other orientations entailed by acyclicity and the set of v-structures found

*separating set

The PC Algorithm – an example



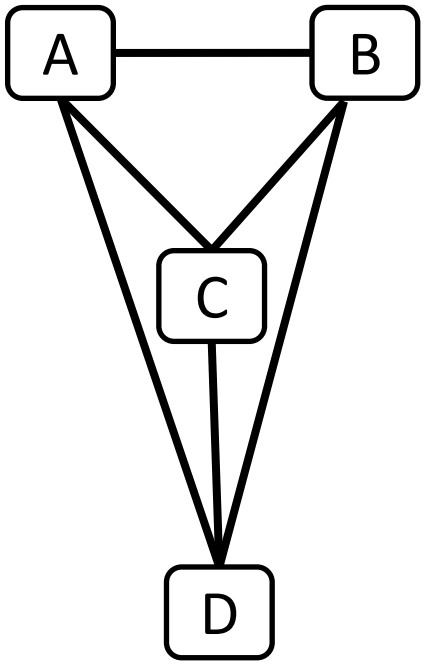
Ind (A, B | ∅)

Ind (A, D | C)

Ind (B, D | C)

Begin with the full graph

The PC Algorithm – an example



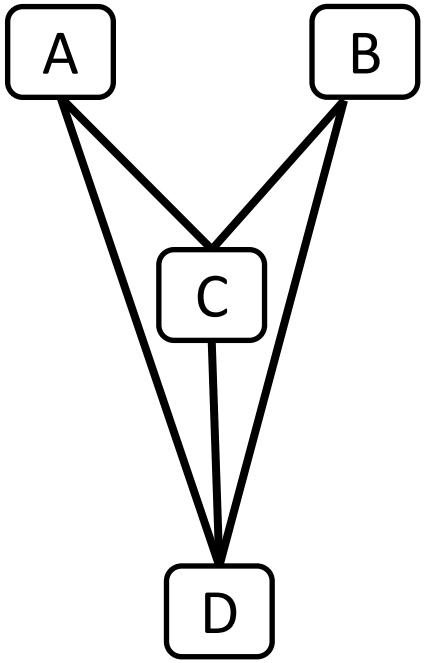
Ind (A, B | ∅)

Ind (A, D | C)

Ind (B, D | C)

A and B do not share an edge

The PC Algorithm – an example



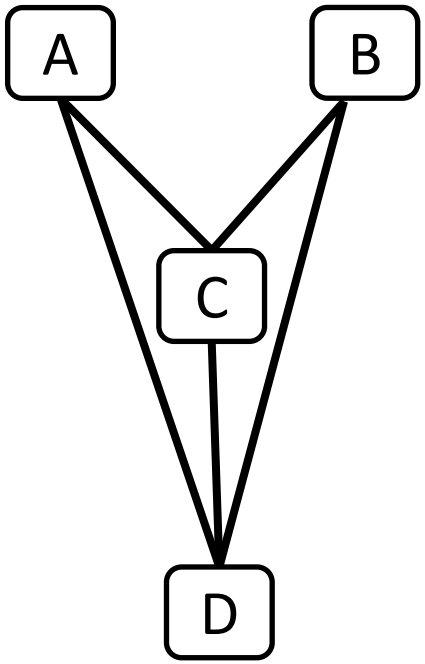
Ind (A, B | ∅)

Ind (A, D | C)

Ind (B, D | C)

Remove A—B

The PC Algorithm – an example



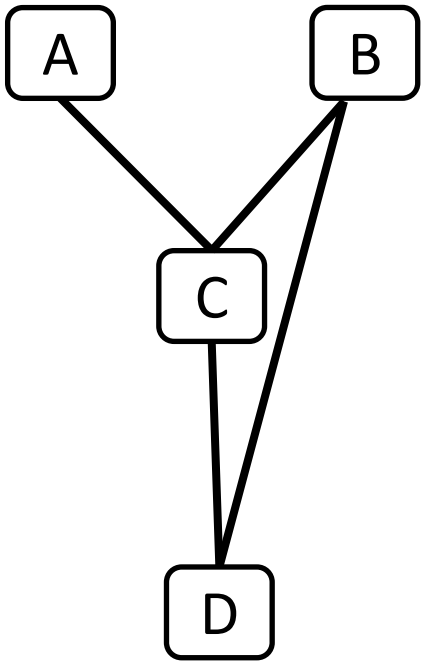
Ind (A, B | ∅)

Ind (A, D | C)

Ind (B, D | C)

A and D do not share an edge

The PC Algorithm – an example



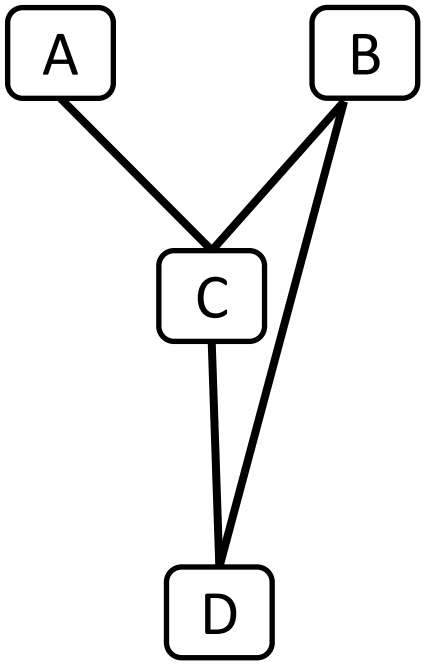
Ind (A, B | ∅)

Ind (A, D | C)

Ind (B, D | C)

Remove A—D

The PC Algorithm – an example



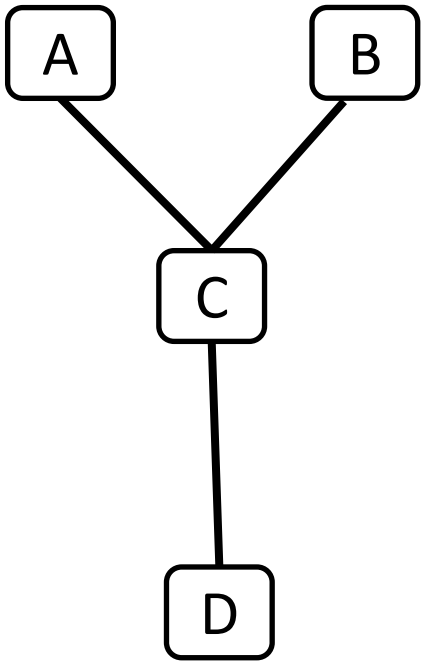
Ind (A, B | ∅)

Ind (A, D | C)

Ind (B, D | C)

B and D do not share an edge

The PC Algorithm – an example



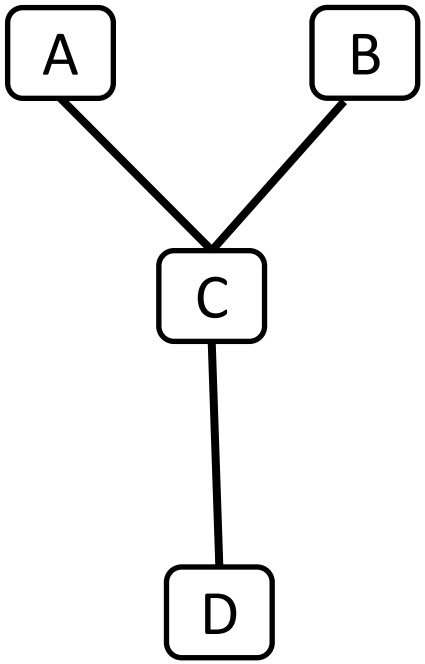
Ind (A, B | ∅)

Ind (A, D | C)

Ind (B, D | C)

Remove B—D

The PC Algorithm – an example



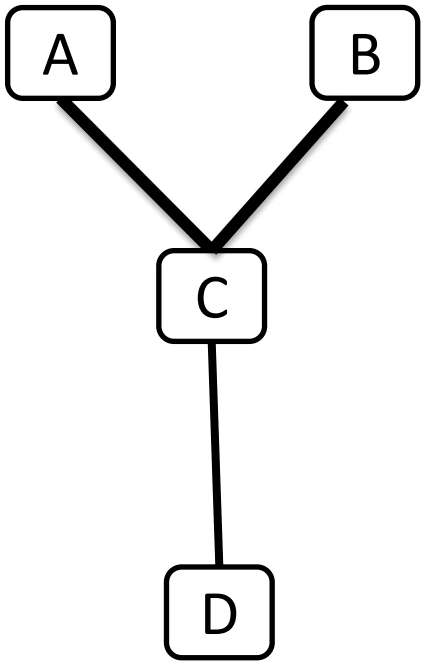
Ind (A, B | ∅)

Ind (A, D | C)

Ind (B, D | C)

Skeleton identification phase is complete

The PC Algorithm – an example



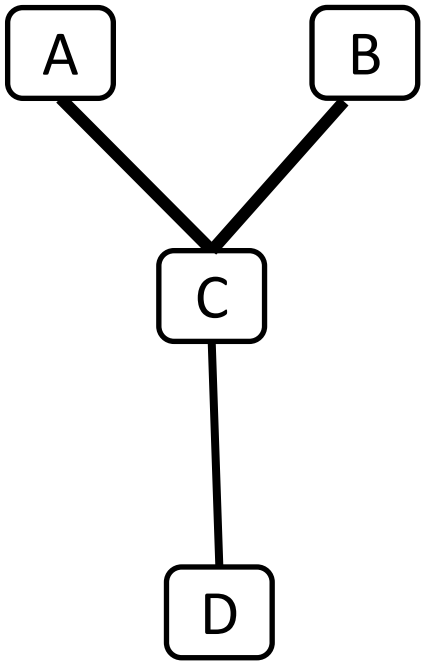
Ind (A, B | ∅)

Ind (A, D | C)

Ind (B, D | C)

A—C—B is a possible v-structure

The PC Algorithm – an example



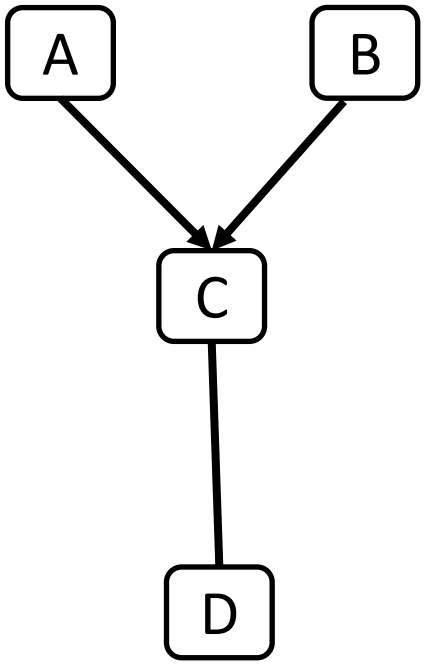
Ind (A, B | ∅)

Ind (A, D | C)

Ind (B, D | C)

A—C—B is a v-structure

The PC Algorithm – an example



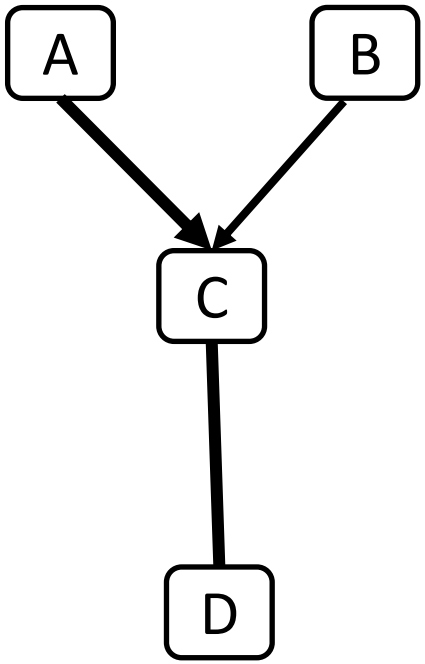
Ind (A, B | ∅)

Ind (A, D | C)

Ind (B, D | C)

Orient A→C←B

The PC Algorithm – an example



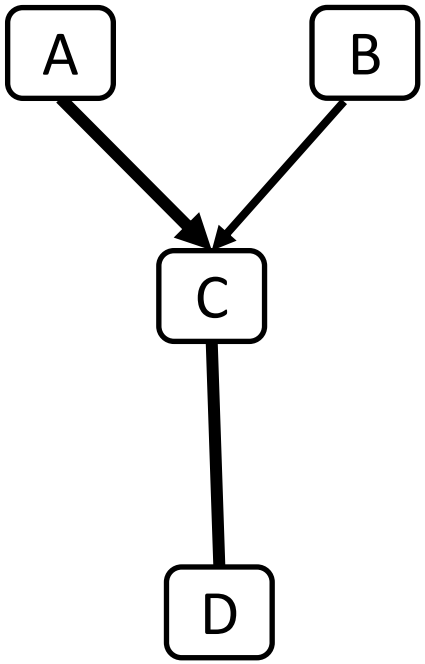
Ind (A, B | ∅)

Ind (A, D | C)

Ind (B, D | C)

A—C—D is a possible v-structure

The PC Algorithm – an example



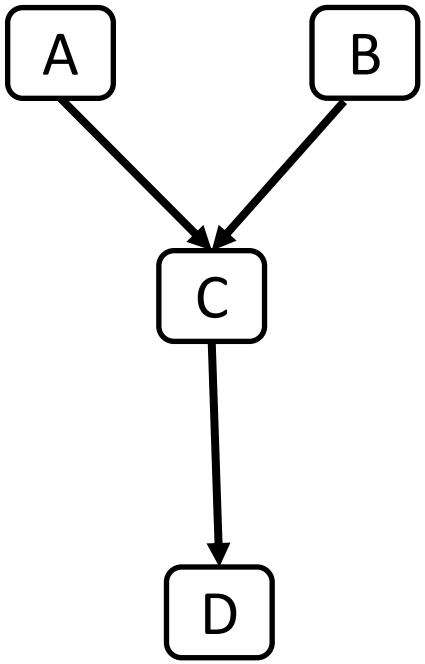
Ind (A, B | ∅)

Ind (A, D | C)

Ind (B, D | C)

A—C—D is NOT a v-structure

The PC Algorithm – an example



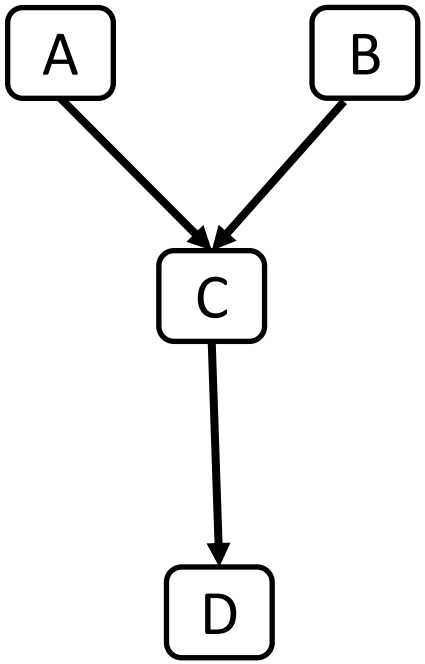
Ind (A, B | ∅)

Ind (A, D | C)

Ind (B, D | C)

Orient C→D

The PC Algorithm – an example



Ind (A, B | ∅)

Ind (A, D | C)

Ind (B, D | C)

Final Output

Considerations

- Search strategy for independencies / structure affects learning quality and efficiency
- Determining dependencies or independencies introduces errors; errors propagate
- For a causal interpretation
  - Check the sensitivity / confidence of a feature ($A \rightarrow B$) [Tsamardinos, Brown 2008, Friedman]
  - Convert the Causal Bayesian Network to a PDAG

Conclusions, Pt.3

- Algorithms for learning Bayesian Networks (global) or parts of networks (local) : decent learning accuracy and scalability (thousands of variables)
- NP-Hard problem [Chickering et al., 1996]
- Constraint-Based
  - Check for dependencies / independencies
- Score-based
  - Maximize fitting to the data (score)
  - Greedy searches approximate but scalable
  - Exact
- Still intense research area

Bibliography

- Cheng, J., Bell, D. & Liu, W., 1997. Learning Bayesian Networks from Data: An Efficient Approach Based on Information Theory. In *CIKM 1997*.
- Ramsey, J., Spirtes, P. & Zhang, J., 2006. Adjacency Faithfulness and Conservative Causal Inference. In *UAI 2006*
- Tsamardinos, I., Brown & Constantin, A., 2006. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1), pp.31-78.
- Provan, G. & Singh, M., 1995. Learning Bayesian Networks Using Feature Selection. In *AI&STATS 1995*
- Acid, S. & de Campos, L.M., 1996. Benedict: an algorithm for learning probabilistic Bayesian networks. In *IPMU 1996*Kaneme,
- Kojima, K. et al., 2010. Optimal search on clustered structural constraint for learning bayesian network structure. *Journal of Machine Learning Research*, 11, pp.285-310.
- Xie, X. & Geng, Z., 2008. A Recursive Method for Structural Learning of Directed Acyclic Graphs. *Journal of Machine Learning Research*, 9, pp.459-483.
- Moore, A. & Wong, W.K., 2003. Optimal reinsertion: A new search operator for accelerated and more accurate Bayesian network structure learning. In *ICML 2003*
- Koivisto, M. & Sood, K., 2004. Exact Bayesian Structure Discovery in Bayesian Networks D. M. Chickering, ed. *Journal of Machine Learning Research*, 5, pp.549-573.

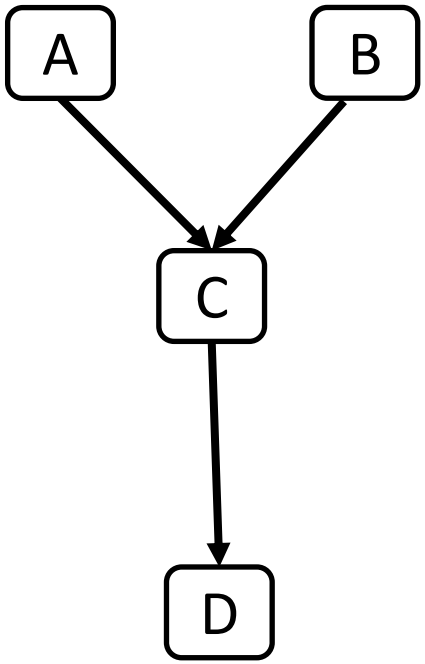
Bibliography

- Cooper, G.F. & Herskovits, E., 1992. A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning*, 09(4), pp.309-347.
- Spirtes, P. & Meek, C., 1995. Learning Bayesian networks with discrete variables from data. In *KDD 1995*
- Friedman, N., Nachman, I. & Pe'er, D., 1999. Learning Bayesian Network Structure from Massive Datasets: The Sparse Candidate'' Algorithm. In *UAI 1999*
- Chickering, D.M. & Meek, C., 2002. Finding Optimal Bayesian Networks. In *UAI 2002*
- Silander, T. & Myllymaki, P., 2006. A simple approach for finding the globally optimal Bayesian network structure. In *Proceedings of Uncertainty in Artificial Intelligence*.
- Koivisto, M., 2006. Advances in exact Bayesian structure discovery in Bayesian networks. In *UAI 2006*
- Chickering, D.M., Geiger, D. & Heckerman, D., 1994. Learning Bayesian Networks is NP-Hard. In *Learning from Data: Artificial Intelligence and Statistics V*.

Causal Discovery

Looking deeper into the assumptions and potential pitfalls

The PC Algorithm – an example



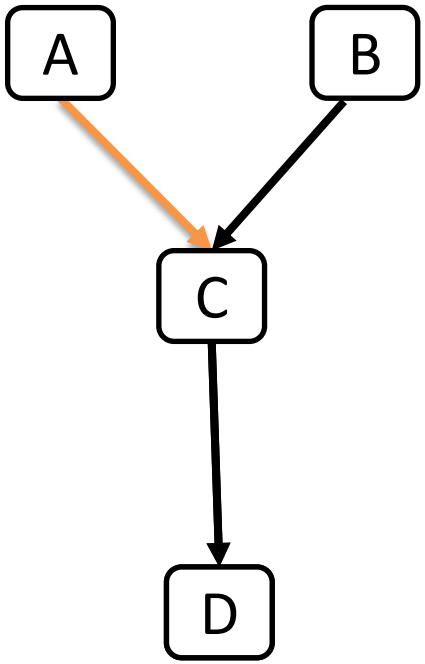
Ind (A, B | ∅)

Ind (A, D | C)

Ind (B, D | C)

Final Output

Causal Interpretation



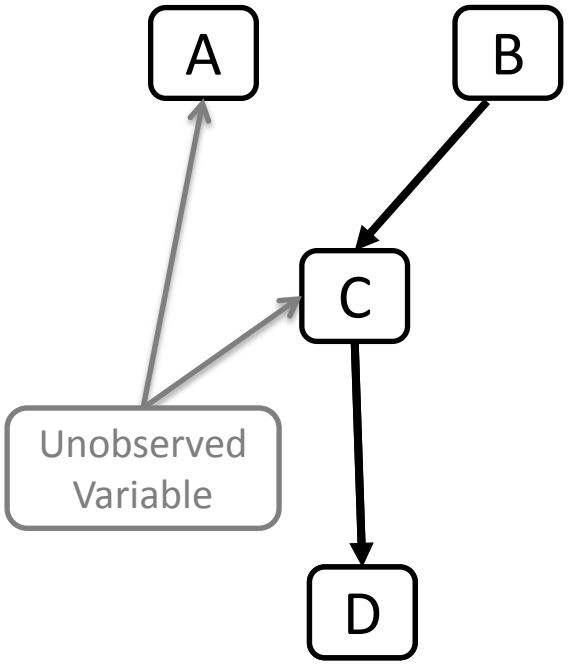
Ind (A, B | ∅)

Ind (A, D | C)

Ind (B, D | C)

- Is A → C causal?

Causal Interpretation



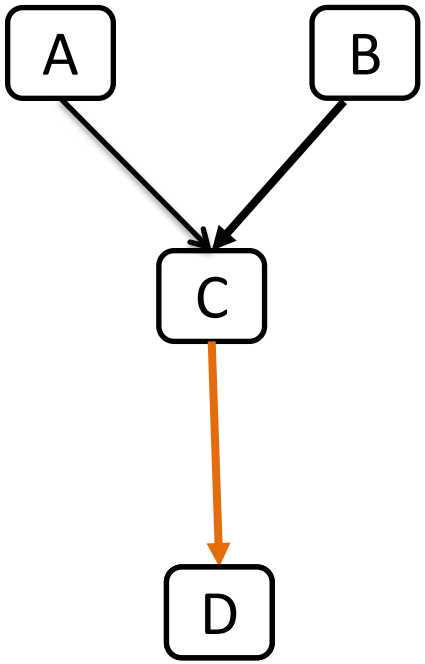
Ind (A, B | ∅)

Ind (A, D | C)

Ind (B, D | C)

- Is A → C causal?
- What would we observe?

Causal Interpretation



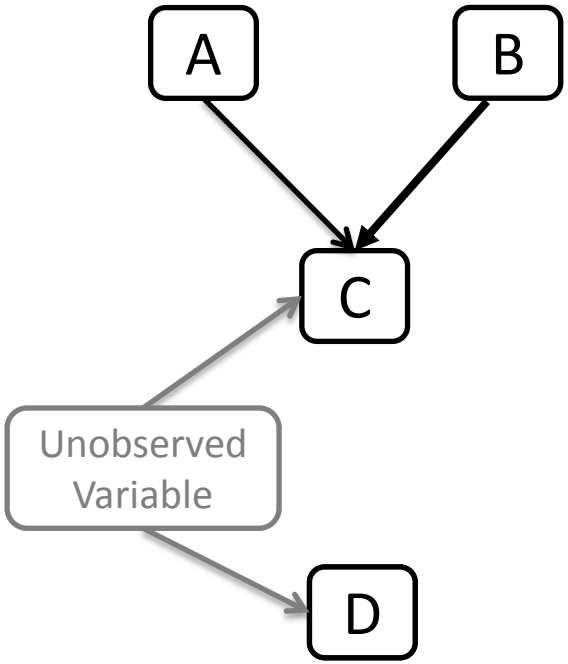
Ind (A, B | ∅)

Ind (A, D | C)

Ind (B, D | C)

- Is C → D causal?
- What would we observe?

Causal Interpretation



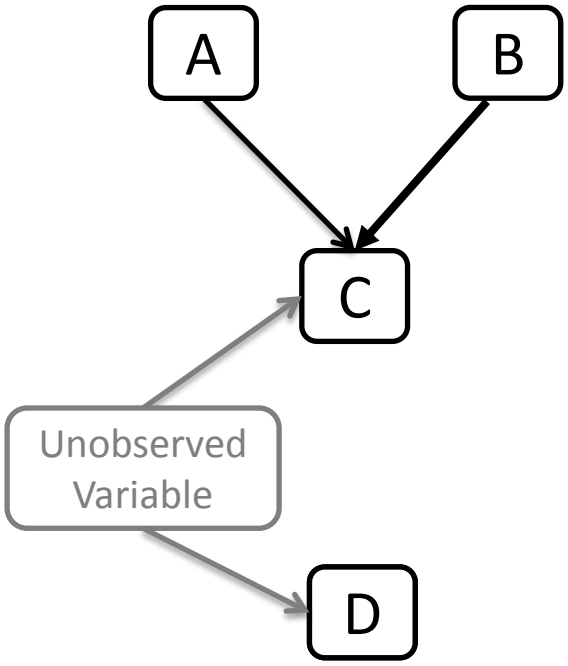
Ind (A, B | ∅)

Ind (A, D | C)

Ind (B, D | C)

- Is C → D causal?
- What would we observe?

Causal Interpretation



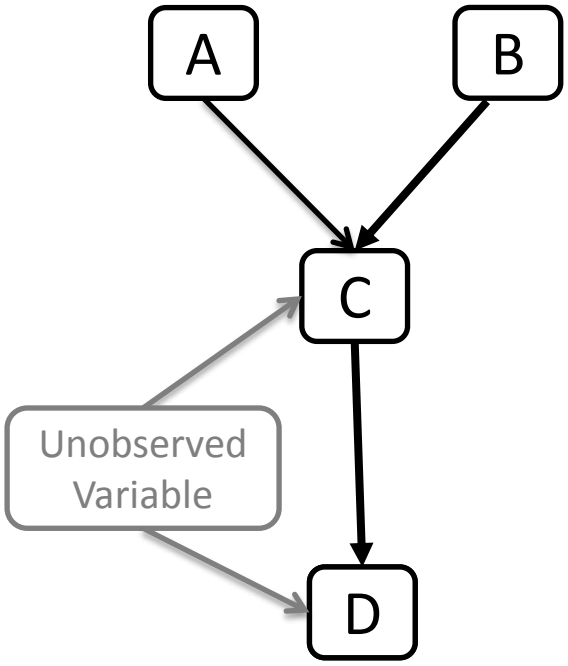
Ind (A, B | ∅)

Ind (A, D | C)

Ind (B, D | C)

- Is C → D causal?
- What would we observe?

Causal Interpretation



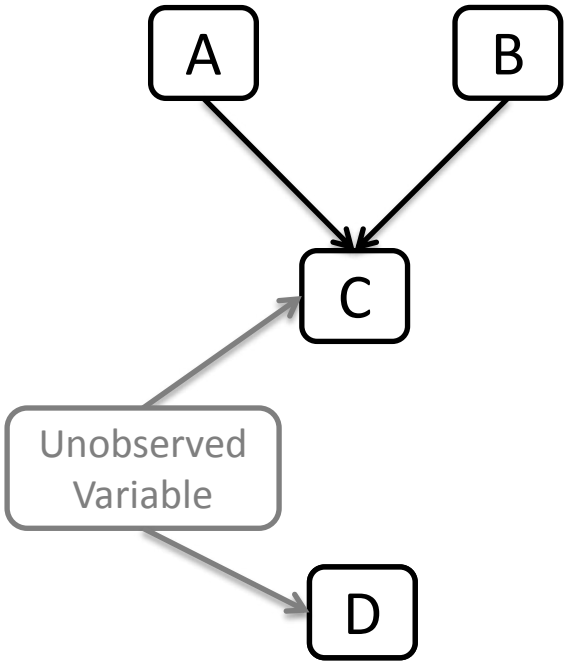
Ind (A, B | ∅)

Ind (A, D | C)

Ind (B, D | C)

- Is C → D causal?
- What would we observe?

Causal Interpretation



Ind (A, B | ∅)

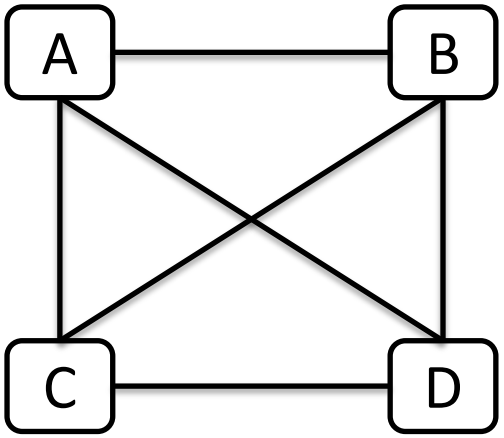
Ind (A, D | C)

Ind (B, D | C)

What went wrong?

Violation of the Causal Sufficiency Assumption

Latent Common Causes

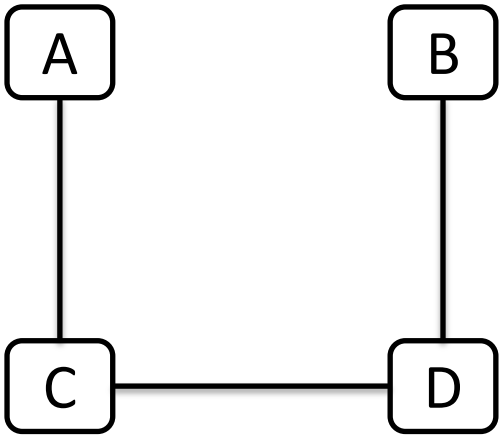


Ind (A, B | $\varnothing$)

Ind (A, D | $\varnothing$)

Ind (B,C | $\varnothing$)

Latent Common Causes



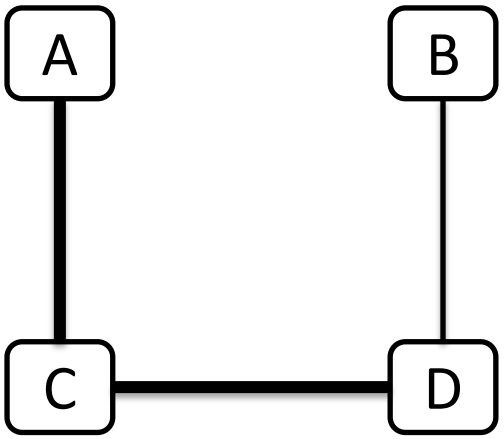
Ind (A, B | ∅)

Ind (A, D | ∅)

Ind (B, C | ∅)

Skeleton Identification

Latent Common Causes



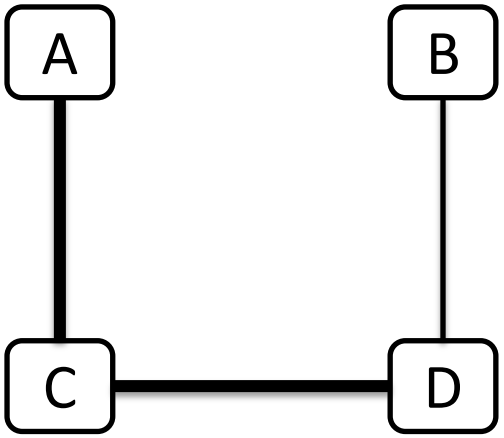
Ind (A, B | ∅)

Ind (A, D | ∅)

Ind (B, C | ∅)

Skeleton Identification

Latent Common Causes



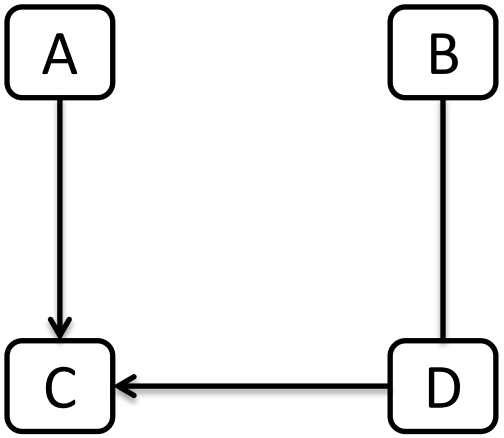
Ind (A, B | ∅)

Ind (A, D | ∅)

Ind (B, C | ∅)

A-C-D is a v-structure

Latent Common Causes



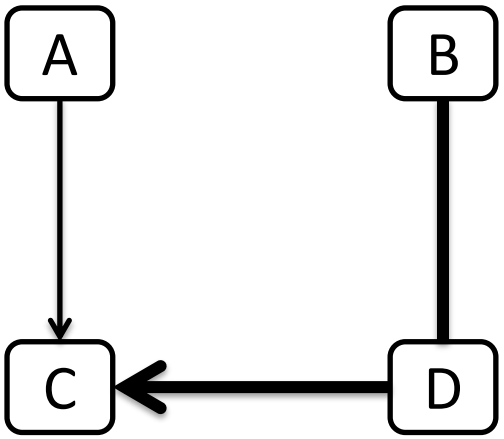
Ind (A, B | ∅)

Ind (A, D | ∅)

Ind (B, C | ∅)

A-C-D is a v-structure

Latent Common Causes



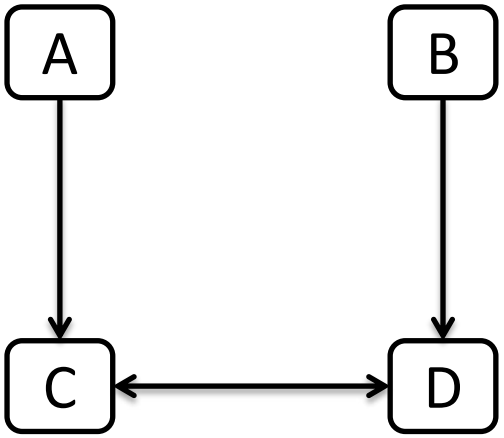
Ind (A, B | ∅)

Ind (A, D | ∅)

Ind (B, C | ∅)

B-D-C is a v-structure

Latent Common Causes



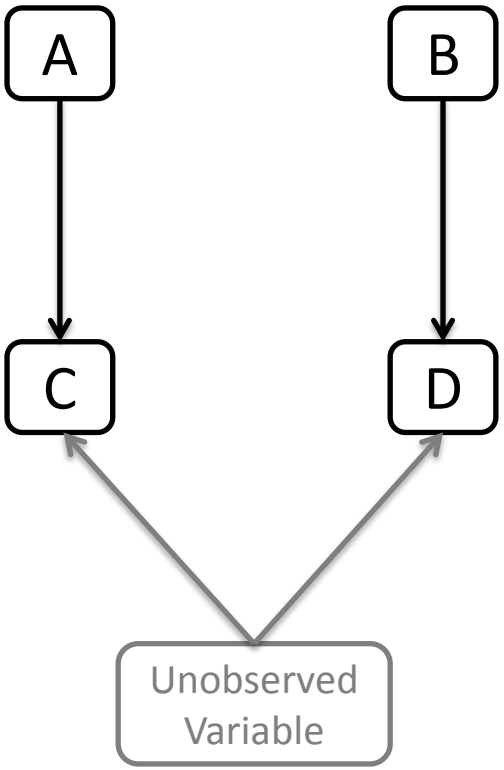
Ind (A, B | ∅)

Ind (A, D | ∅)

Ind (B, C | ∅)

B-D-C is a v-structure

Latent Common Causes



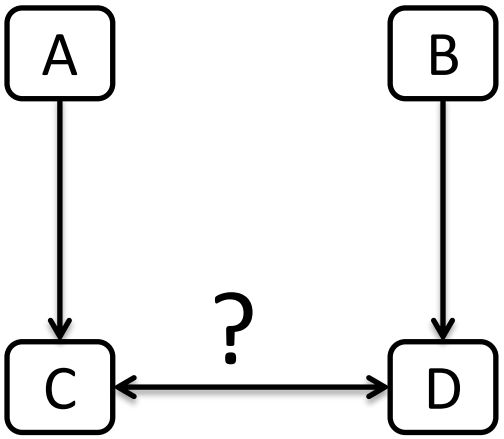
Ind (A, B | ∅)

Ind (A, D | ∅)

Ind (B, C | ∅)

Orientations imply there is a latent common cause

Latent Common Causes



Ind (A, B | ∅)

Ind (A, D | ∅)

Ind (B, C | ∅)

This is not a BN

What went wrong?

Violation of the Causal Sufficiency Assumption

Latent Common Causes

- We can sometimes detect the violation of causal sufficiency

- If we're not sure that causal sufficiency holds, we can't be sure an edge A→ B is causal

Violation of the Causal Sufficiency Assumption

Maximal Ancestral Graphs

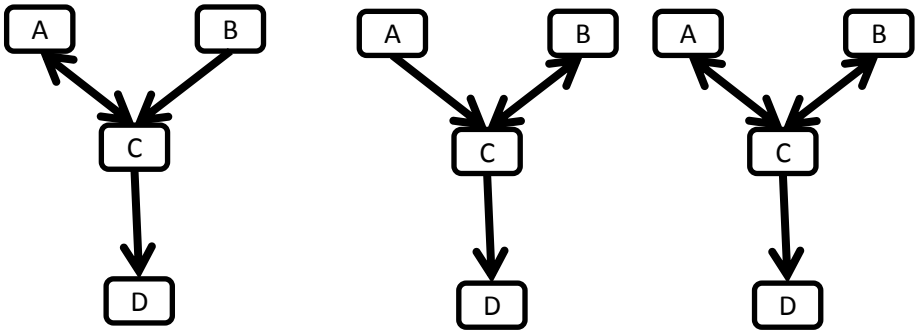
- Two kind of edges (Mixed Graphs)
  - Directed (A → B: A is an ancestor of B)
  - Bi-directed (A ↔ B : A is not an ancestor of B, B is not an ancestor of A)
  - Missing edge means no direct causation (the reverse does not hold though)
- No Directed/ Almost Directed Cycles Allowed

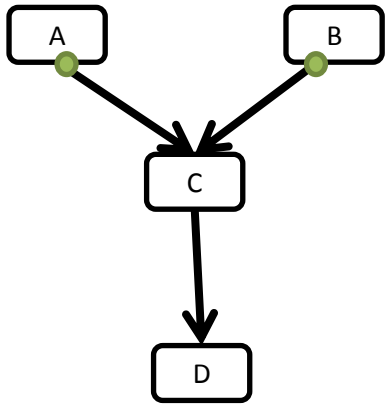
MAGs

- Straight-forward edge interpretation:
  - Edges declare ancestry or non ancestry
  - Missing edge declare no direct causality
- If a distribution is faithful to a MAG (DAG), any marginal of the distribution is faithful to a MAG
- **M-separation** criterion captures independencies
- **Markov Equivalent** MAGS are statistically indistinguishable

Partially oriented Ancestral Graph

- Represents a class of Markov Equivalent MAGs
- Has the same edges as every other MAG in the class
- Has only orientations (arrows an tails) shared by all the MAGs in the class ; uncertainties are denoted with circles
- FCI algorithm can identify the PAG





Selection Bias

- You want to test whether genotype X causes Lung Cancer
- You take 100 Lung Cancer patients (cases)
- For every Lung Cancer patient you add to the survey a non-Lung Cancer patient (controls)

Selection Bias

- You want to test whether genotype X causes Lung Cancer
- You take 100 Lung Cancer patients (cases)
- For every Lung Cancer patient you add to the survey a non-Lung Cancer patient (controls) from the same hospital

Dep (Genotype X, Lung Cancer| ∅)

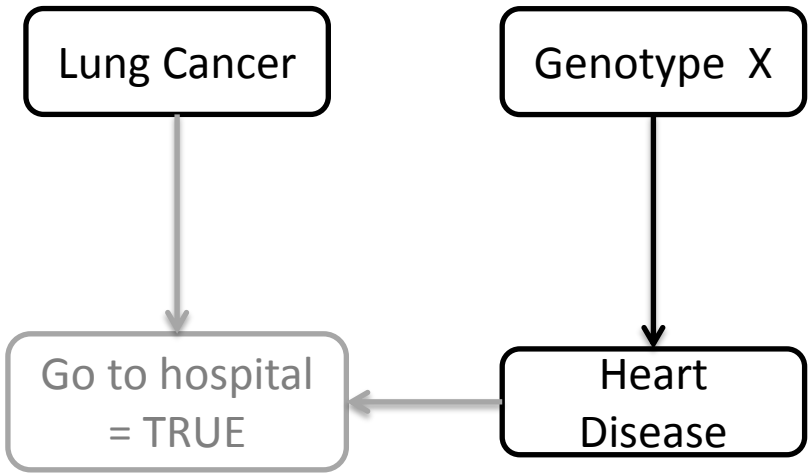
Selection Bias

- You want to test whether genotype X causes Lung Cancer
- You take 100 Lung Cancer patients (cases)
- For every Lung Cancer patient you add to the survey a non-Lung Cancer patient (controls) from the same hospital

Dep (Genotype X, Lung Cancer| ∅)

Genotype X causes Lung Cancer

Spurious associations due to experimental design

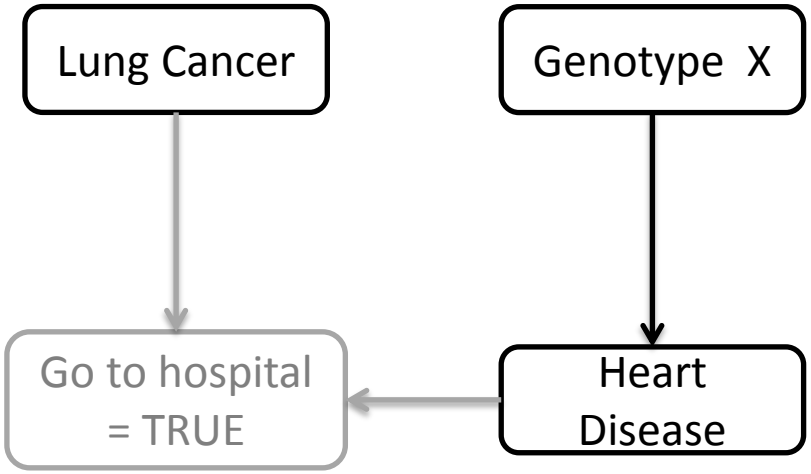


- You want to test whether genotype X causes Lung Cancer
- You take 100 Lung Cancer patients (cases)
- For every Lung Cancer patient you add to the survey a non-Lung Cancer patient (controls) from the same hospital

Dep (Genotype X, Lung Cancer| ∅)

Genotype X causes Lung Cancer

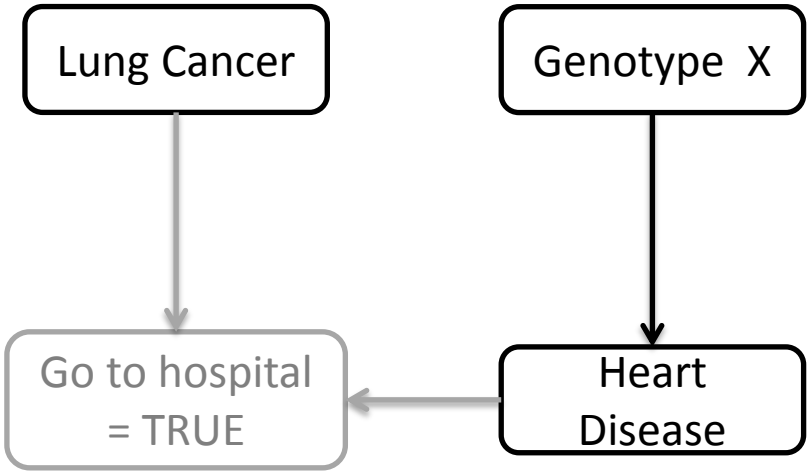
Selection Bias



What went wrong?

Violation of the Causal Markov Condition

Selection Bias



MAGs can also handle selection bias, BUT the edges have different semantics

What went wrong?

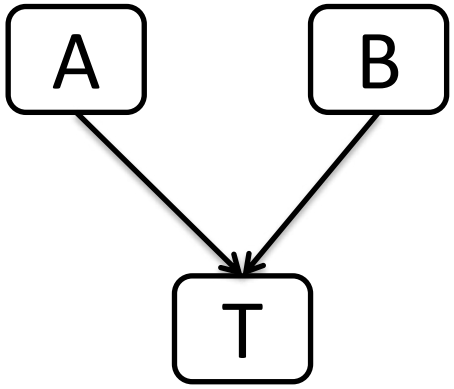
Violation of the Causal Markov Condition

Collinearity and Determinism

Disappearing Associations

Measurement Error

Observing an XOR function



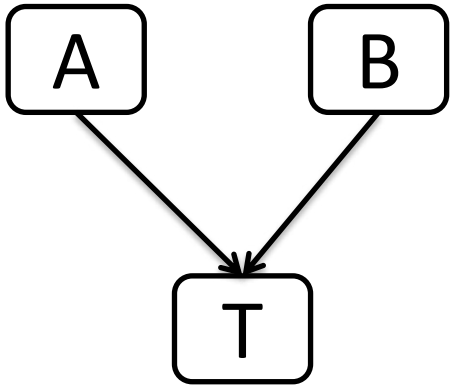
T =A XOR B

		T	
A	B	0	1
0	0	0.25	0
0	1	0	0.25
1	0	0	0.25
1	1	0.25	0

$P(A \cap T) = 0.25 = P(A)*P(T)$
$P(A \cap \neg T) = 0.25 = P(A)*P(T)$
$P(\neg A \cap T) = 0.25 = P(\neg A)*P(T)$
$P(\neg A \cap \neg T) = 0.25 = P(\neg A)*P(\neg T)$

$P(A \cap T) = P(A)*P(T)$

Observing an XOR function



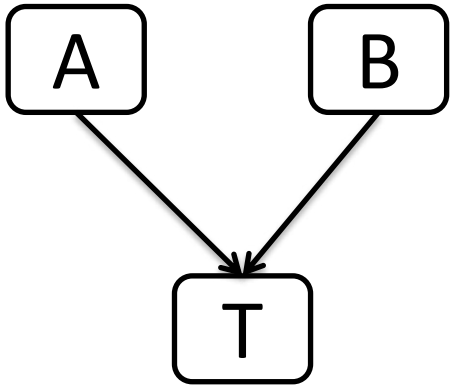
T =A XOR B

		T	
A	B	0	1
0	0	0.25	0
0	1	0	0.25
1	0	0	0.25
1	1	0.25	0

$P(A \cap T) = 0.25 = P(A)*P(T)$
$P(A \cap \neg T) = 0.25 = P(A)*P(T)$
$P(\neg A \cap T) = 0.25 = P(\neg A)*P(T)$
$P(\neg A \cap \neg T) = 0.25 = P(\neg A)*P(\neg T)$

$P(A \cap T) = P(A)*P(T)$

Observing an XOR function



T =A XOR B

		T	
A	B	0	1
0	0	0.25	0
0	1	0	0.25
1	0	0	0.25
1	1	0.25	0

Dep(A, T|∅)

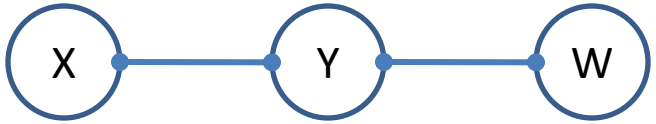
Ind(A, T|∅)

$P(A \cap T) = 0.25 = P(A)*P(T)$
$P(A \cap \neg T) = 0.25 = P(A)*P(T)$
$P(\neg A \cap T) = 0.25 = P(\neg A)*P(T)$
$P(\neg A \cap \neg T) = 0.25 = P(\neg A)*P(\neg T)$

$P(A \cap T) = P(A)*P(T)$

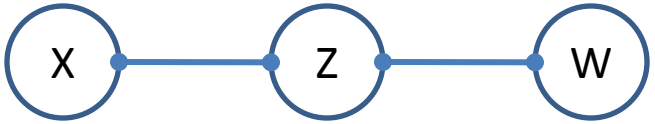
Example

Dataset measuring *X, Y, W*
PAG G_1



$J = \{ \langle X, W \mid Y \rangle \}$

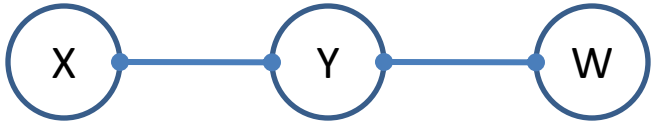
Dataset measuring *X, Z, W*
PAG G_2

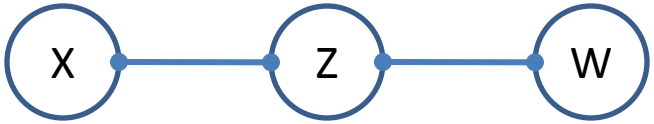


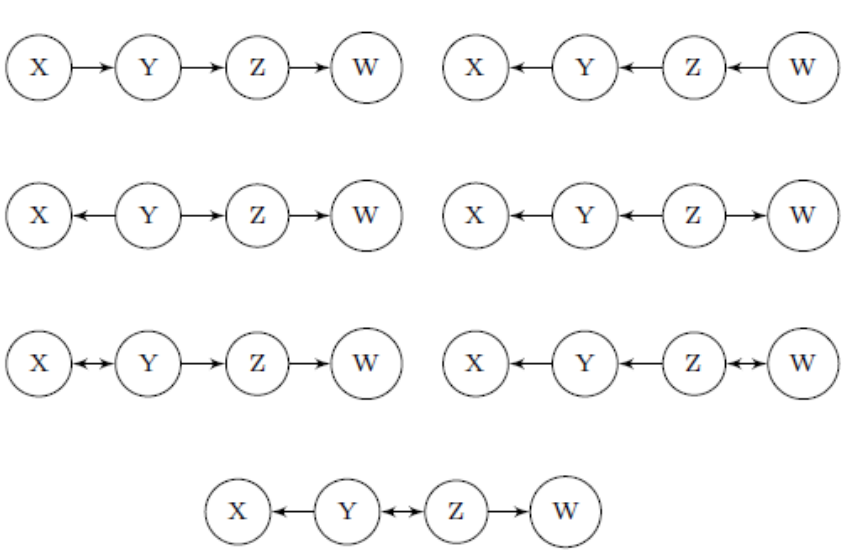
$J = \{\langle X, W \mid Z \rangle \}$

- Find a single causal model on {*X, Y, Z, W*} that is consistent with both independence models

Solutions to the Example

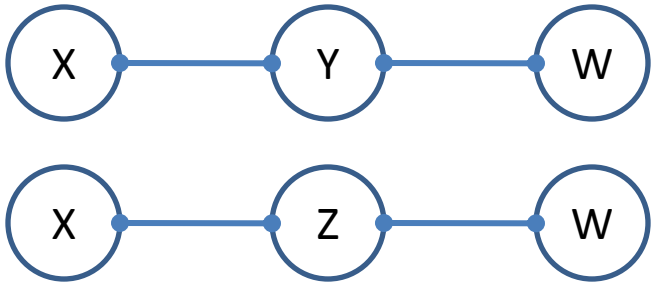




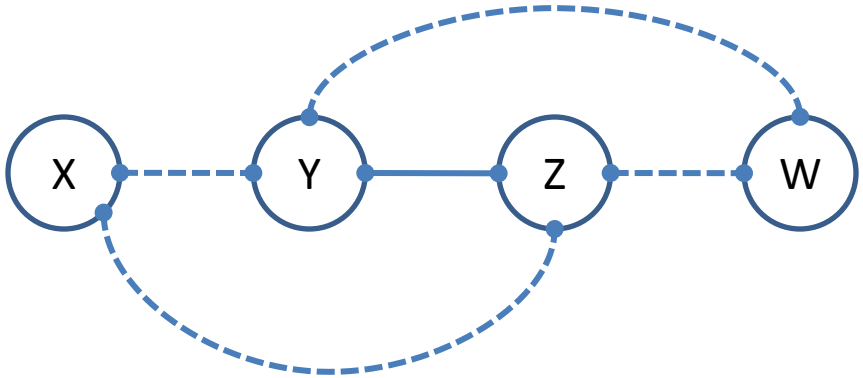


Proof-of-Concept INCA Predictions

If



then



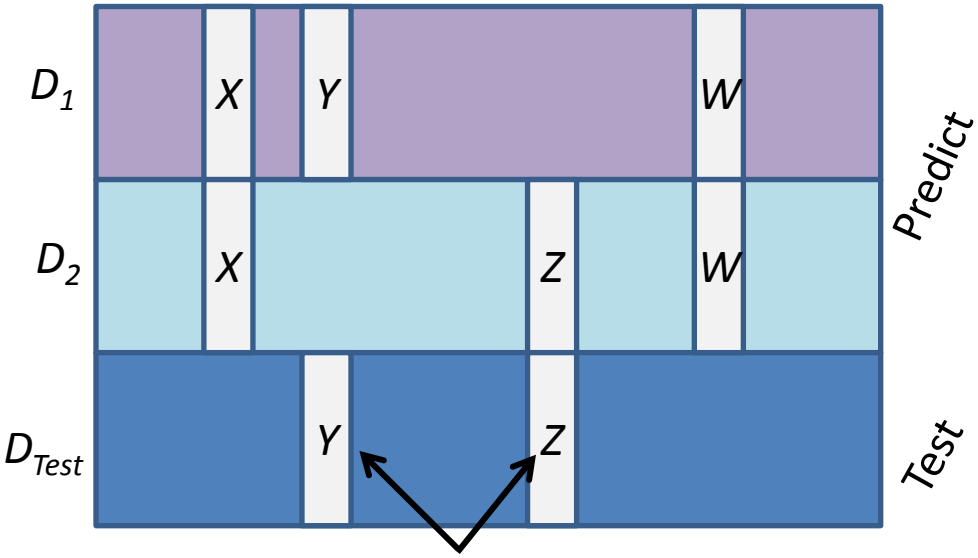
- If
  - $J_1 = \{ \langle X, W \mid Y \rangle \}$ in dataset D_1
  - $J_2 = \{ \langle X, W \mid Z \rangle \}$ in dataset D_2
- Then
  - Predict *Y and Z are associated* $\neg(Y \perp Z)$

Making It Work on Real Data

1.Original Dataset

2.Split to D_1 , D_2 and D_{test} containing different samples

3.Find *X, Y, W in* D_1 and *X, Z, W,* in D_2 such that the *FTR* applies



Association?

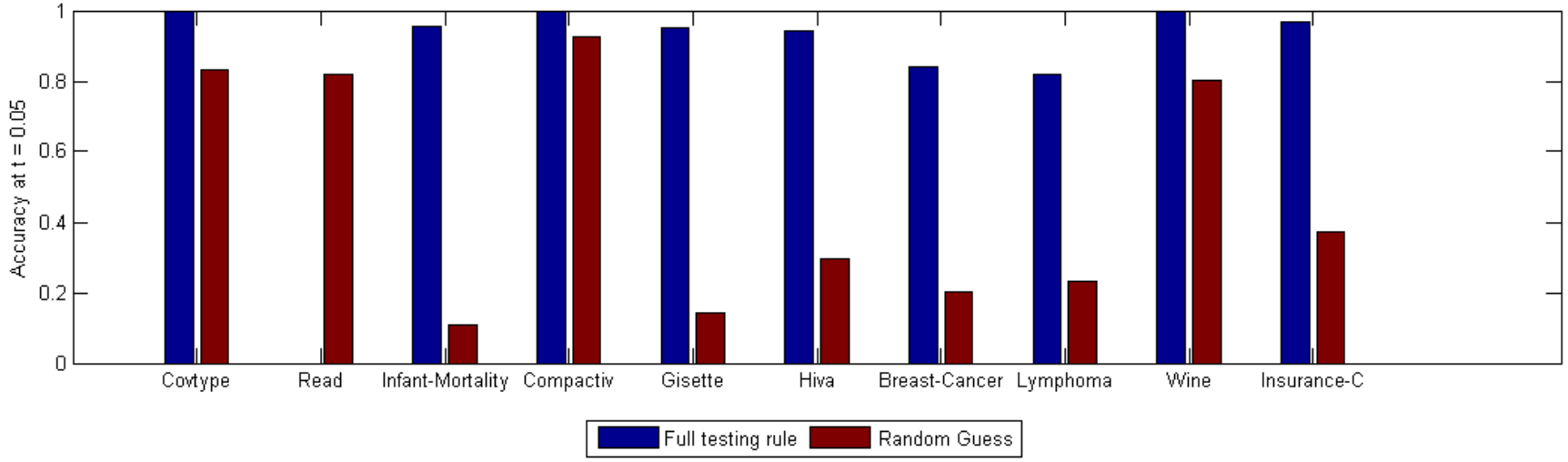
Datasets Employed

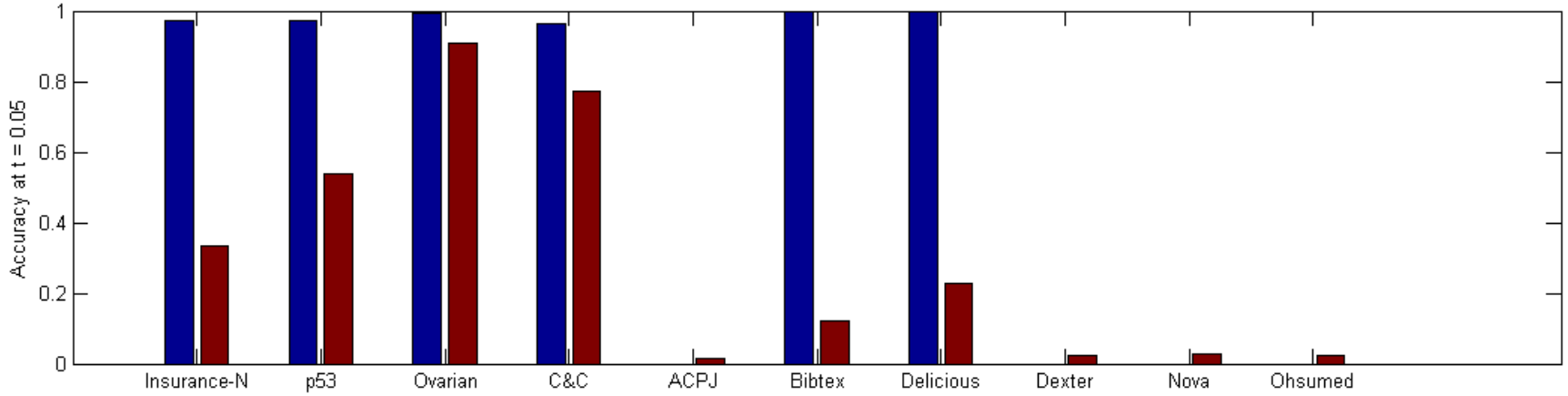
Name	Reference	# istances	# vars	Group Size	Vars type	Scient. domain
Covtype	Blackard and Dean (1999)	581012	55	55	N/O	Agricultural
Read	Guvenir and Uysal (2000)	681	26	26	N/C/O	Business
Infant–mortality	Aliferis et al. (2010)	5337	83	83	N	Clinical study
Compactiv	Alcalá-Fdez et al. (2009)	8192	22	22	C	Computer science
Gisette	Aliferis et al. (2010)	7000	5000	50	C	Digit recognition
Hiva	Guyon et al. (2006)	4229	1617	50	N	Drug discovering
Breast–Cancer	Aliferis et al. (2010)	286	17816	50	C	Gene expression
Lymphoma	Aliferis et al. (2010)	237	7399	50	C	Gene expression
Wine	Cortez et al. (2009)	4898	12	12	C	Industrial
Insurance–C	Elkan (2001)	9000	84	84	N/O	Insurance
Insurance–N	Elkan (2001)	9000	86	86	N/O	Insurance
p53	Danziger et al. (2009)	16772	5408	50	C	Protein activity
Ovarian	Aliferis et al. (2010)	216	2190	50	C	Proteomics
C&C	Frank and Asuncion (2010)	1994	128	128	C	Social science
ACPJ	Aliferis et al. (2010)	15779	28228	50	C	Text mining
Bibtex	Tsoumakas et al. (2010)	7395	1995	50	N	Text mining
Delicious	Tsoumakas et al. (2010)	16105	1483	50	N	Text mining
Dexter	Aliferis et al. (2010)	600	11035	50	N	Text mining
Nova	Aliferis et al. (2010)	1929	12709	50	N	Text mining
Ohsumed	Aliferis et al. (2010)	5000	14373	50	C	Text mining

Performance Metric

- Ground truth is unknown
- Accuracy
  - The percentage of p-values < 0.05
- May include false positives and exclude false negatives in the calculation

Predicting Dependencies: FTR vs Random





Number of Predictions

Dataset	Full testing rule	Min. testing rule	Transit. rule
Covtype	222	42978	58518
Read	0	27	5025
Infant-Mortality	22	9870	8663
Compactiv	135	763	4134
Gisette	423	74223	261095
Hiva	554	167799	305901
Breast-Cancer	1833	597652	1688313
Lymphoma	7712	1182824	1112963
Wine	4	84	466
Insurance–C	1839	57718	60695
Insurance–N	226	34344	56615
p53	46647	2469957	2516926
Ovarian	539165	2107067	2248459
C&C	99241	536785	351557
ACPJ	0	547	31556
Bibtex	1	15975	85640
Delicious	856	78502	187456
Dexter	0	8	752
Nova	0	676	17807
Ohsumed	0	157	10139

Bibliography

- Richardson, T. & Spirtes, P., 2002. Ancestral Graph Markov Models. *The Annals of Statistics*, 30(4), pp.962-1030.
- Zhang, J., 2008. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17), pp.1873-1896.

Success Stories

Feature Selection from a Bayesian Network Perspective

- Find the set of variables that are (collectively/multivariately) the most predictive of your target
  - Shouldn't they be "related" to the target? How exactly?
  - How do you measure how predictive they are?
  - How do you find them?

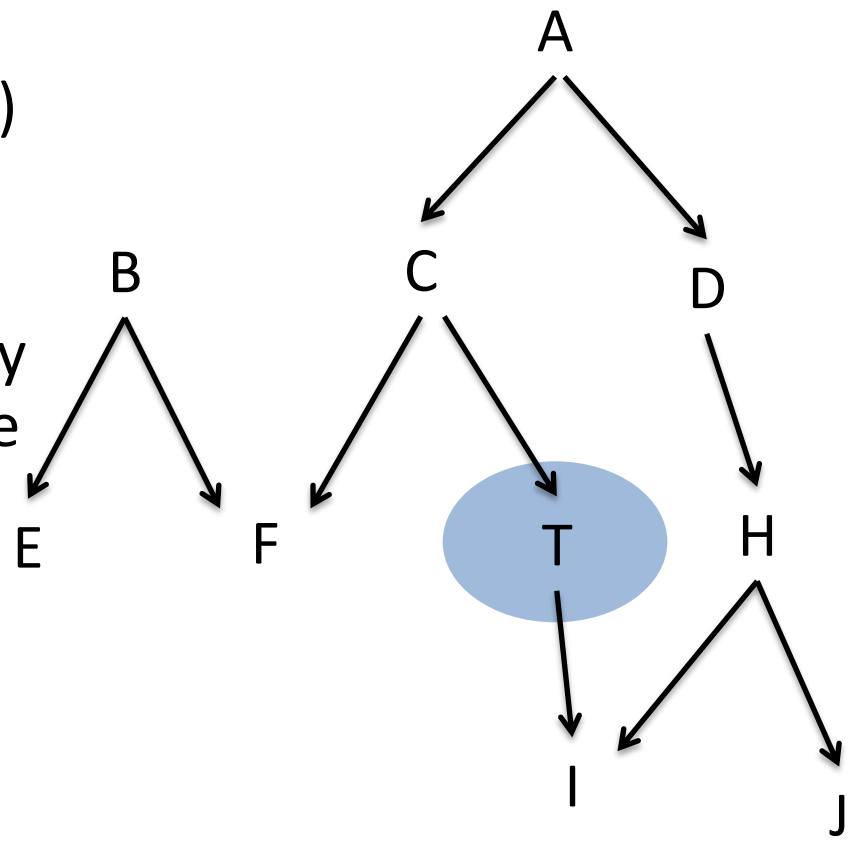
Feature Selection As a Solution to High-Dimensional Analysis

- Reduce the number of required observed quantities (variables/features) to build a predictive/diagnostic model

- Definition: Select the variable subset of minimal size with the maximal predictive or diagnostic, classification power for target variable T

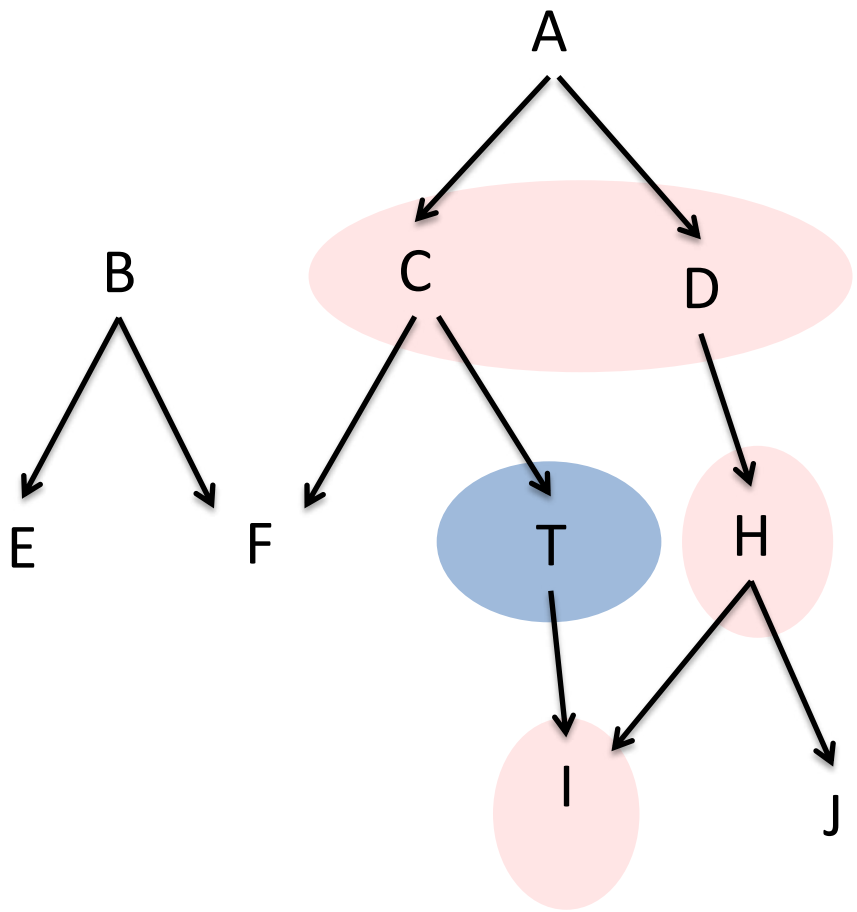
Causality and Feature Selection

- "Relevant" variables provide information for *T:* Dep(*X ; T* | ∅)
- "Redundant" variables are "relevant" but don't provide any additional information given the selected variables
- Crude definition of "relevancy" and "redundancy" but often used



Causality and Feature Selection

- Markov Blanket is unique in Faithful networks
- Relevant: any variable *X* with a directed path to or from *T*
- Redundant: a relevant variable that is made d-separated from *T* given the selected variables
- Non relevant variables maybe in the Markov Blanket (required for optimal selection), e.g., *H*
- What is redundant depends on the selected variables



Causality and Feature Selection

- The smallest subset with the optimal predictive power is the set of
  - Parents (direct causes)
  - Children (direct effects)
  - Spouses (direct causes of the direct effects)

  of the target variable (to predict) in the network that fits the data

- This set is called the Markov Blanket of the target
- The Markov Blanket is unique in Faithful distributions

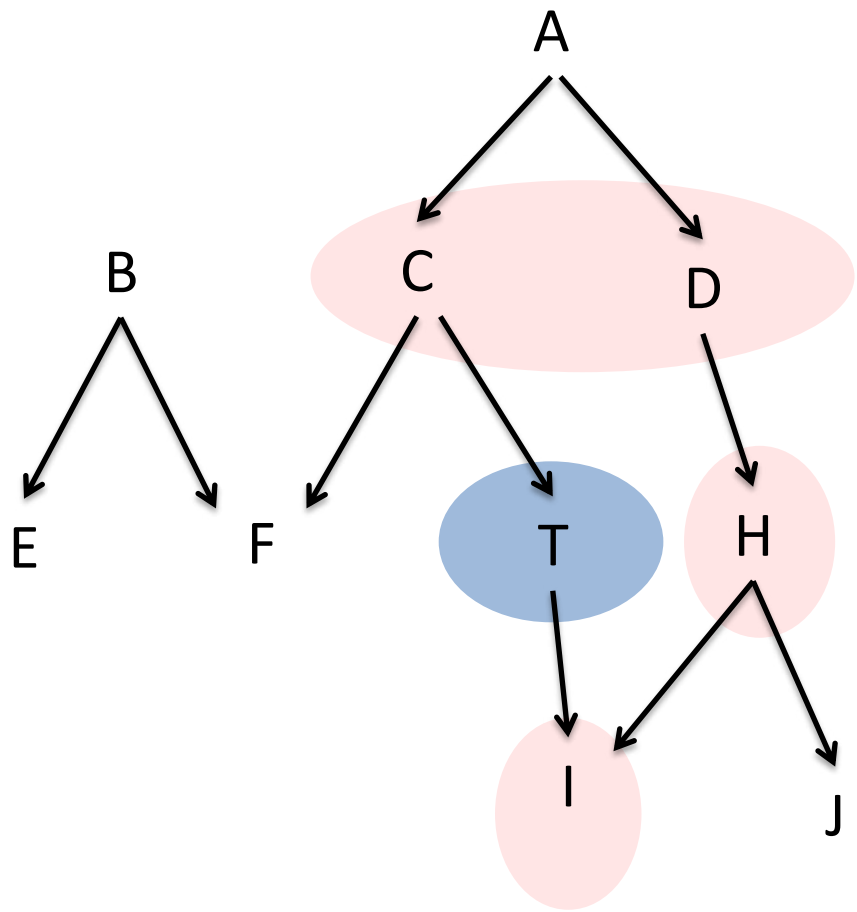
- Connections among Bayesian Networks, Markov Blanket, Feature Selection, Relevant Variable, and more
  - [Tsamardinos, Aliferis, AI&Stats 2003]

Causality and Feature Selection

- The Markov Blanket of *T* is:
  - Parents (direct causes)
  - Children (direct effects)
  - Spouses (direct causes of the direct effects)

  in the causal network

- Knowing the values of the Markov Blanket variables renders knowledge of the values of all other variables superfluous



Causal-Based Feature Selection

- Identify the Markov Blanket of the target using methods based on causal theories
- Use the Markov Blanket variables to build the final predictive or diagnostic models
- Design efficient and accurate algorithms that identify the MB without having to learn the whole network
  - Max-Min Markov Blanket, [Tsamardinos, Aliferis, Statnikov, KDD 2003]
  - HITON [Aliferis, Tsamardinos, Statnikov, AMIA 2003]
  - General framework and extended evaluation
    - [Aliferis, Statnikov, Tsamardinos, et. al. JMLR 2010]

Method	*Reference*
No feature selection	
RFE (recursive feature elimination SVM-based method)	(Guyon et al., 2002)
UAF-KruskalWallis-SVM (univariate ranking by Kruskal-Wallis statistic and feature selection with SVM backward wrapper)	(Statnikov et al., 2005a; Hollander and Wolfe, 1999)
UAF-Signal2Noise-SVM (univariate ranking by signal-to-noise statistic and feature selection with SVM backward wrapper)	(Guyon et al., 2006b; Statnikov et al., 2005a; Furey et al., 2000)
UAF-Neal-SVM (univariate ranking by Radford Neal's statistic and feature selection with SVM backward wrapper)	Chapter 10 in (Guyon et al., 2006a)
Random Forest Variable Selection (RFVS)	(Diaz-Uriarte and Alvarez de Andres, 2006; Breiman, 2001)
LARS-Elastic Net (LARS-EN)	(Zou and Hastie, 2005)
RELIEF (with backward wrapping by SVM)	(Kononenko, 1994; Kira and Rendell, 1992)
L0-norm	(Weston et al., 2003)
Forward Stepwise Selection	(Caruana and Freitag, 1994)
Koller-Sahami (with backward wrapping by SVM)	(Koller and Sahami, 1996)
IAMB	(Tsamardinos and Aliferis, 2003; Tsamardinos et al., 2003a)
K2MB	(Cooper et al., 1997; Cooper and Herskovits, 1992)
BLCD-MB	(Mani and Cooper, 2004)
FAST-IAMB	(Yaramakala and Margaritis, 2005)
HITON-PC (semi-interleaved)	Novel algorithm

Dataset name	*Domain*	*Num. variables*	*Num. samples*	*Target*	*Reference*
Infant_ Mortality	Clinical	86	5,337	Died within the first year	(Mani and Cooper, 1999)
Ohsumed	Text	14,373	5,000	Relevant to neonatal diseases	(Joachims, 2002)
ACPJ_ Etiology	Text	28,228	15,779	Relevant to etiology	(Aphinyanaphongs et al., 2006)
Lymphoma	Gene expression	7,399	227	3-year survival: dead vs. alive	(Rosenwald et al., 2002)
Gisette	Digit recognition	5,000	7,000	Separate 4 from 9	*NIPS 2003 Feature Selection Challenge* (Guyon et al., 2006a)
Dexter	Text	19,999	600	Relevant to corporate acquisitions	*NIPS 2003 Feature Selection Challenge* (Guyon et al., 2006a)
Sylva	Ecology	216	14,394	Ponderosa pine vs. everything else	*WCCI 2006 Performance Prediction Challenge*

Dataset name	*Domain*	*Num. variables*	*Num. samples*	*Target*	*Reference*
Ovarian_ Cancer	Proteomics	2,190	216	Cancer vs. normals	(Conrads et al., 2004)
Thrombin	Drug discovery	139,351	2,543	Binding to thrombin	KDD Cup 2001
Breast_ Cancer	Gene expression	17,816	286	Estrogen-receptor positive (ER+) vs. ER-	(Wang et al., 2005)
Hiva	Drug discovery	1,617	4,229	Activity to AIDS HIV infection	WCCI 2006 Performance Prediction Challenge
Nova	Text	16,969	1,929	Separate politics from religion topics	WCCI 2006 Performance Prediction Challenge
Bankruptcy	Financial	147	7,063	Personal bankruptcy	(Foster and Stine, 2004)

Results Overview

- Compactness
  - HITON-PC wins (statistically significantly) 20 cases
  - HITON-PC ties (non statistically significant result) 16 cases
  - HITON-PC loses (statistically significantly) 6 cases
    - with significance loss of predictive power of the other methods
- Predictive Power
  - HITON-PC wins (statistically significantly) 9 cases
  - HITON-PC ties (non statistically significant result) 33 cases
  - HITON-PC loses (statistically significantly) 1 case
- Time Efficiency
  - Thrombin dataset with> 100,000 features HITON-PC requires 10 to 52 minutes single-CPU time and less than 3 hours when parameters are automatically optimized by cross-validation

Extensions to Survival Analysis

- Causal-based variable selection extended for survival data, where censorship of patients is possible
- Compared against most other methods in the field (filtering, forward selection, Bayesian variable selection, etc.)
- Each algorithm coupled with several regressors
- Statistically significantly the best performing algorithm against all algorithms that reduce the model to less than 20 variables
- [Lagani, Tsamardinos, Bioinformatics (2010)]
- Extensions to other types of data (temporal) are under investigation

Predicting causal effects in large-scale systems from observational data

What will happen if you knock down Gene X?

	Gene A	Gene B	…	Gene X
1	0.1	0.5		1.2
2	0.56	2.32		0.7
…				
n	7	0.4		2.4

[Maathius et al., Nature Methods, 2010]

Predicting causal effects in large-scale systems from observational data

What will happen if you knock down Gene X?

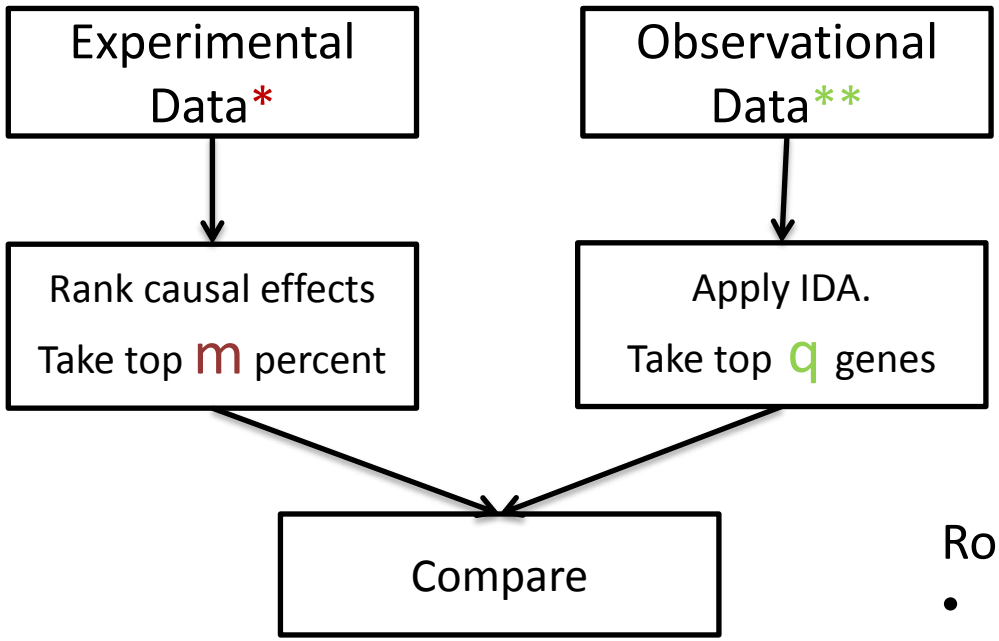
	Gene A	Gene B	…	Gene X
1	0.1	0.5		1.2
2	0.56	2.32		0.7
…				
n	7	0.4		2.4

Intervention calculus when **D**AG is **A**bsent

1. Learn equivalence class of DAGs
2. For every DAG G in the equivalence find $Pa_X(G)$
3. Causal effect C_G of X on V in DAG G is
   1. 0, if $V \in Pa_X(G)$
   2. coefficient of X in $V \sim X + Pa_X(G)$, otherwise
4. Causal effect C of X on V is the minimum of all C_G

Predicting causal effects in large-scale systems from observational data

IDA Evaluation

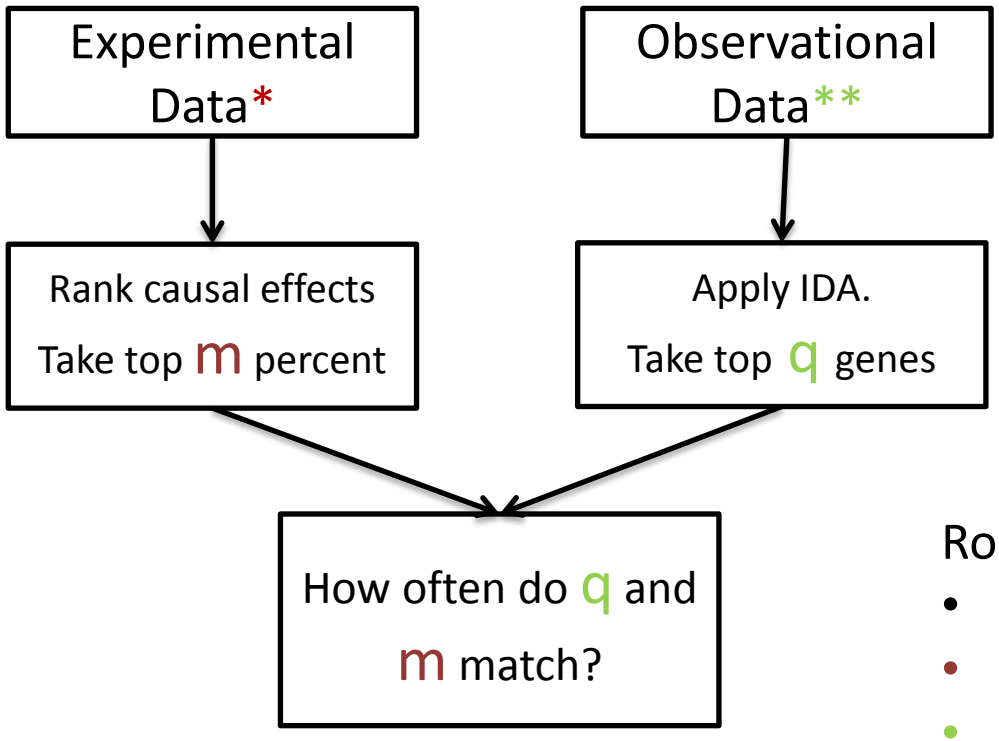


Rosetta Compendium data:

- 5,361 genes
- 234 single-gene deletion mutants*
- 63 wild-type measurements**

Predicting causal effects in large-scale systems from observational data

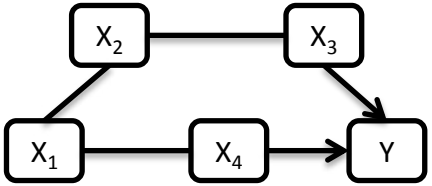
IDA Evaluation



Rosetta Compendium data:

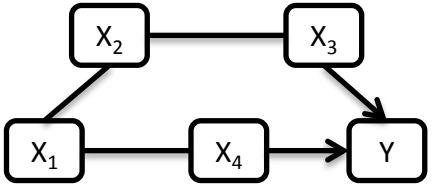
- 5,361 genes
- 234 single-gene deletion mutants*
- 63 wild-type measurements**

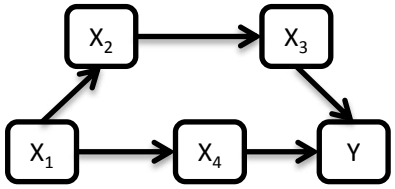
Predicting causal effects in large-scale systems from observational data



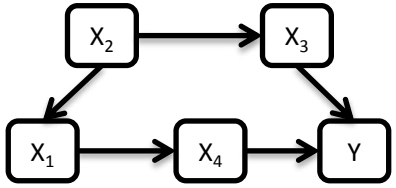
[Example from Maathuis et al, 2009]

Predicting causal effects in large-scale systems from observational data

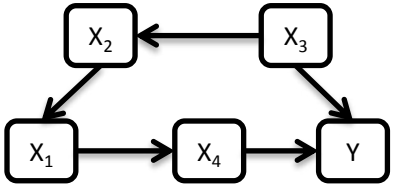




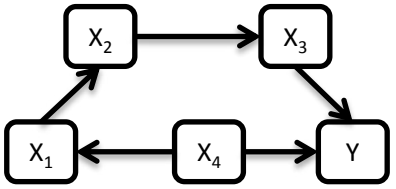
$Pa_1(G_1)= \{\varnothing\}$



$Pa_1(G_2)= \{X_2\}$

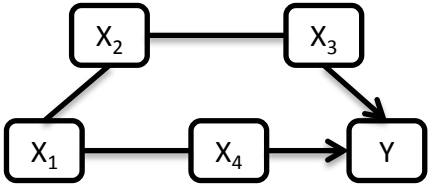


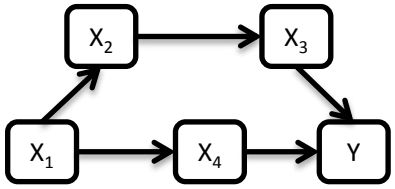
$Pa_1(G_3)= \{X_2\}$



$Pa_1(G_4)= \{X_2\}$

Predicting causal effects in large-scale systems from observational data



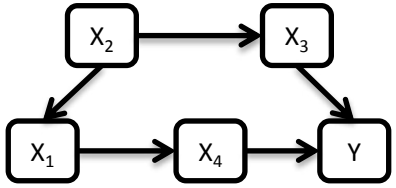


$Pa_1(G_1) = \{\varnothing\}$

Regress Y on X_1

↓

C_1

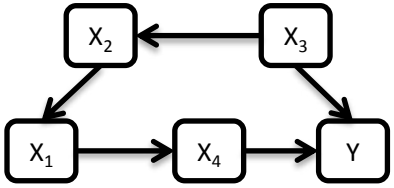


$Pa_1(G_2) = \{X_2\}$

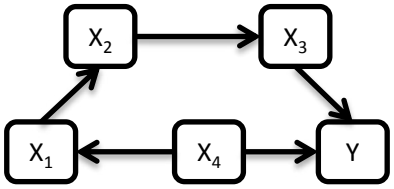
Regress Y on X_1 and X_2

↙ ↘

C_2 C_3



$Pa_1(G_3) = \{X_2\}$



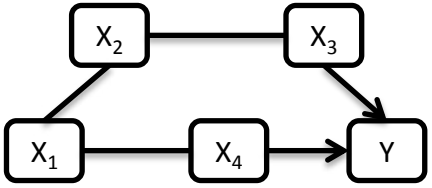
$Pa_1(G_4) = \{X_2\}$

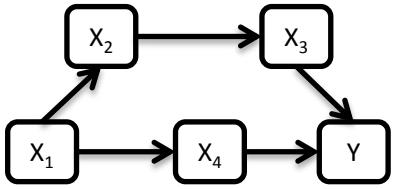
Regress Y on X_1 and X_4

↓

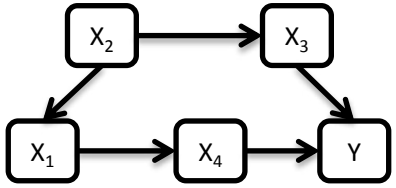
C_4

Predicting causal effects in large-scale systems from observational data

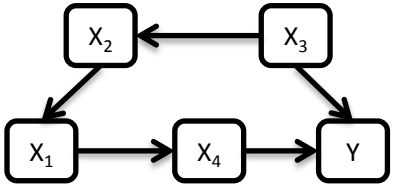




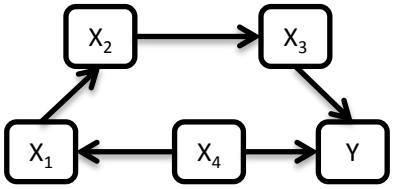
$Pa_1(G_1) = \{\varnothing\}$



$Pa_1(G_2) = \{X_2\}$



$Pa_1(G_3) = \{X_2\}$



$Pa_1(G_4) = \{X_2\}$

Regress Y on X_1

↓

C_1

Regress Y on X_1 and X_2

↙ ↘

C_2 C_3

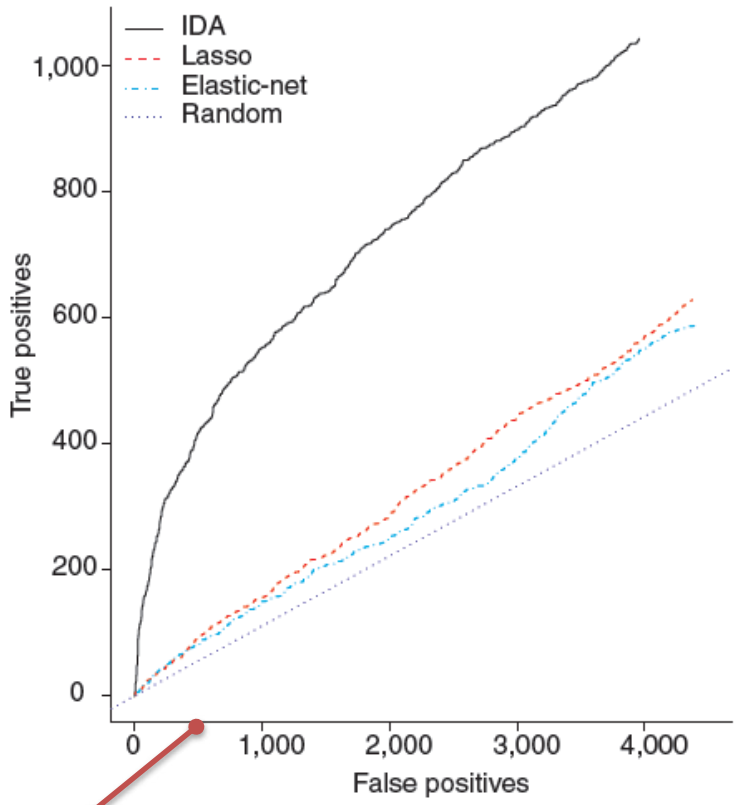
Regress Y on X_1 and X_4

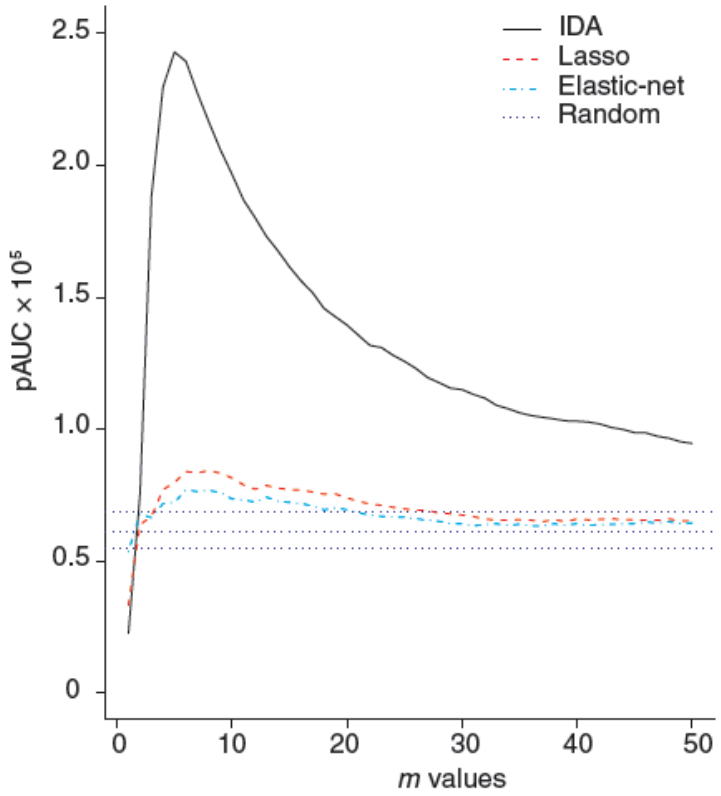
↓

C_4

$C = \min \{ C_1, C_2, C_3, C_4 \}$

Predicting causal effects in large-scale systems from observational data





- m=10%

An integrative genomics approach to infer causal associations between gene expression and disease

Causal Inference with the use of genetic variation

- DNA variation $\longrightarrow$ gene expression $\longleftrightarrow$ phenotype.
- DNA variation used to identify susceptible loci for phenotypic traits (QTLs).

[Schadt et al., Nature Genetics, 2005]

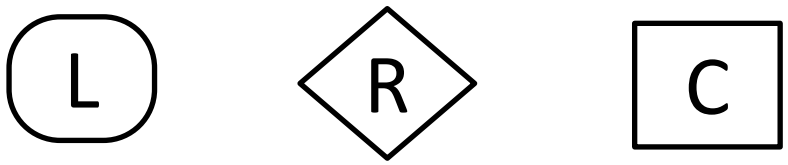
An integrative genomics approach to infer causal associations between gene expression and disease

Causal Inference with the use of genetic variation

- DNA variation ⟶ gene expression ⟷ phenotype.
- DNA variation used to identify regions in DNA susceptible for phenotypic traits (QTLs).

Use gene expression as a phenotypic trait caused by genetic variation (identify eQTLs).

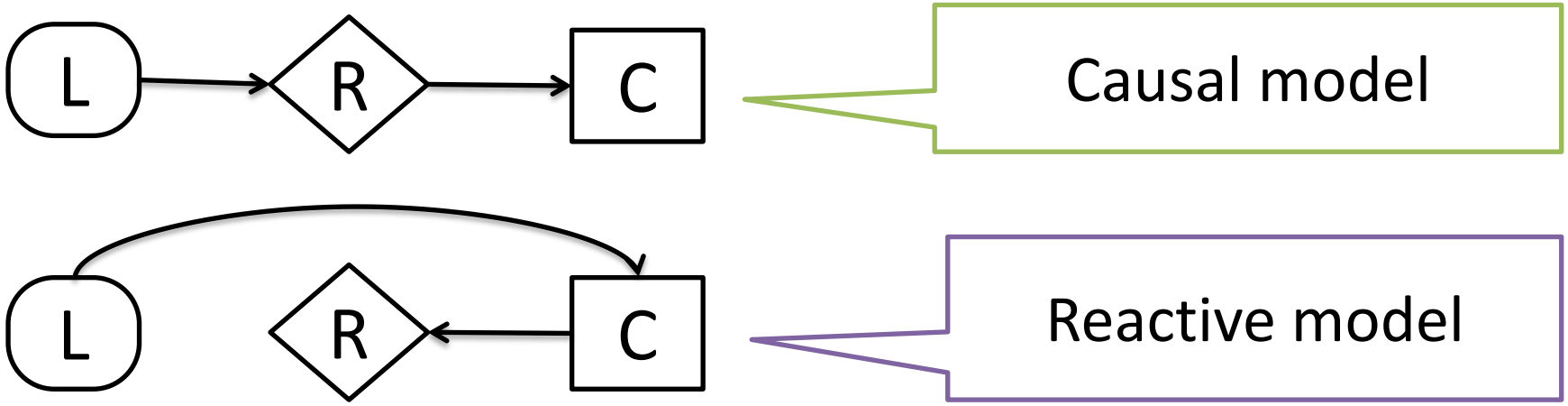
An integrative genomics approach to infer causal associations between gene expression and disease



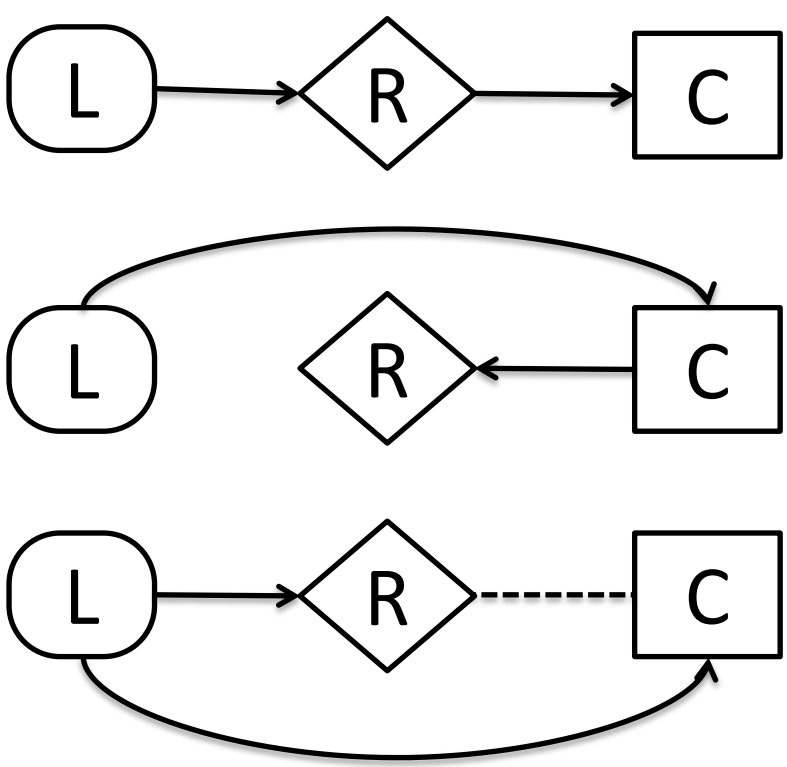
The expression of a gene (R) and a complex trait (C) are correlated with a common QTL (L).

An integrative genomics approach to infer causal associations between gene expression and disease

An integrative genomics approach to infer causal associations between gene expression and disease



An integrative genomics approach to infer causal associations between gene expression and disease

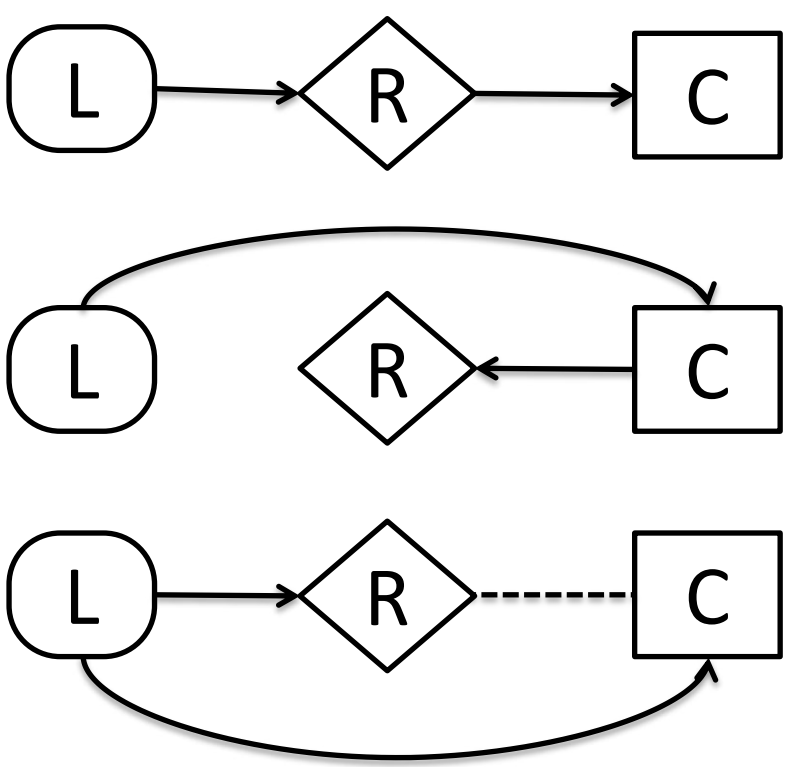


Causal model

Reactive model

Independent model

An integrative genomics approach to infer causal associations between gene expression and disease



$P(L) \times P(R|L) \times P(C|R)$

$P(L) \times P(C|L) \times P(R|C)$

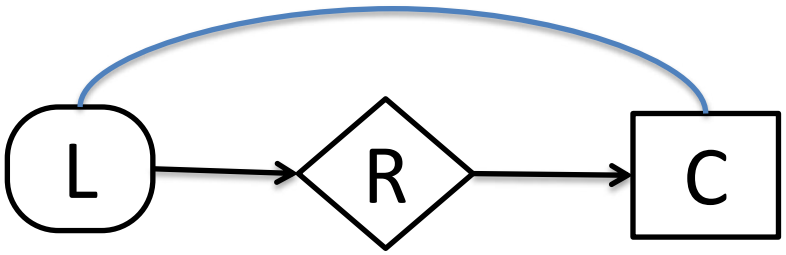
$P(L) \times P(R|L) \times P(C|R, L)$

An integrative genomics approach to infer causal associations between gene expression and disease

- 111 mice from segregated population
- Expression of 23,574 genes (R)
- Genotyped at 139 microsatellite markers. (L)
- Omental Fat Pad Mass trait (C)

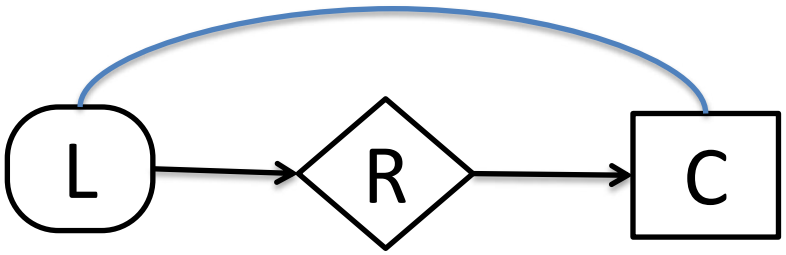
An integrative genomics approach to infer causal associations between gene expression and disease

1. Identify loci susceptible for disease



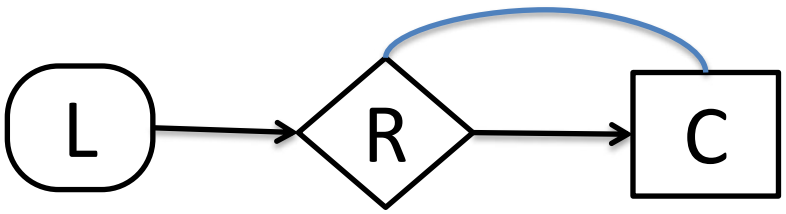
An integrative genomics approach to infer causal associations between gene expression and disease

1. Identify loci susceptible for disease
    - 4 QTLs



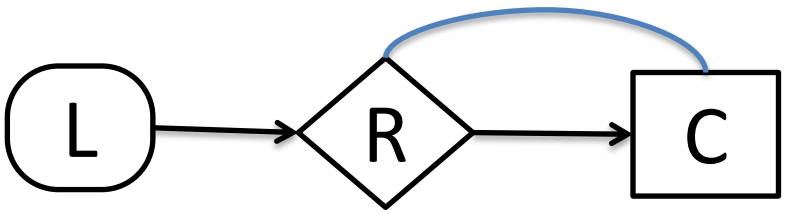
An integrative genomics approach to infer causal associations between gene expression and disease

1. Identify loci susceptible for disease
    - 4 QTLs
2. Identify gene expression traits correlated with the disease



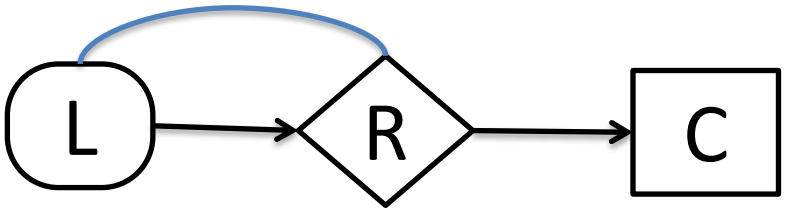
An integrative genomics approach to infer causal associations between gene expression and disease

1. Identify loci susceptible for disease
   - 4 QTLs
2. Identify gene expression traits correlated with the disease
   - 440 genes



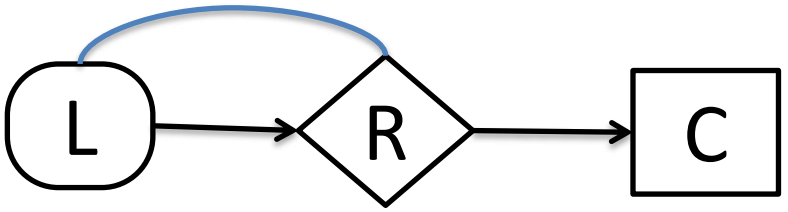
An integrative genomics approach to infer causal associations between gene expression and disease

1. Identify loci susceptible for disease
   - 4 QTLs
2. Identify gene expression traits correlated with the disease
   - 440 genes
3. Identify genes with eQTLs that coincide with the QTLs



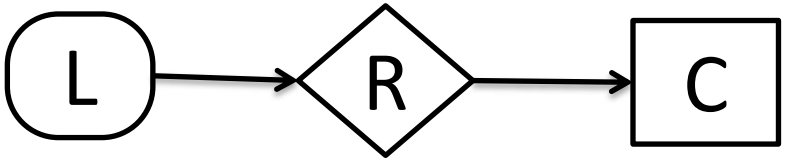
An integrative genomics approach to infer causal associations between gene expression and disease

1. Identify loci susceptible for disease
   - 4 QTLs
2. Identify gene expression traits correlated with the disease
   - 440 genes
3. Identify genes with eQTLs that coincide with the QTLs
   - 113 genes, 267 eQTLs



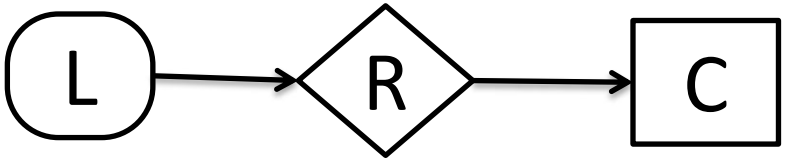
An integrative genomics approach to infer causal associations between gene expression and disease

1. Identify loci susceptible for disease
   - 4 QTLs
2. Identify gene expression traits correlated with the disease
   - 440 genes
3. Identify genes with eQTLs that coincide with the QTLs
   - 113 genes, 267 eQTLs
4. Identify genes that support causal models



An integrative genomics approach to infer causal associations between gene expression and disease

1. Identify loci susceptible for disease
   - 4 QTLs
2. Identify gene expression traits correlated with the disease
   - 440 genes
3. Identify genes with eQTLs that coincide with the QTLs
   - 113 genes, 267 eQTLs
4. Identify genes that support causal models
5. Rank genes by causal effect



An integrative genomics approach to infer causal associations between gene expression and disease

1. Identify loci susceptible for disease
   - 4 QTLs
2. Identify gene expression traits correlated with the disease
   - 440 genes
3. Identify genes with eQTLs that coincide with the QTLs
   - 113 genes, 267 eQTLs
4. Identify genes that support causal models
5. Rank genes by causal effect

Top-ranked genes are the strongest causal candidates

An integrative genomics approach to infer causal associations between gene expression and disease

1. Identify loci susceptible for disease
   - 4 QTLs
2. Identify gene expression traits correlated with the disease
   - 440 genes
3. Identify genes with eQTLs that coincide with the QTLs
   - 113 genes, 267 eQTLs
4. Identify genes that support causal models
5. Rank genes by causal effect

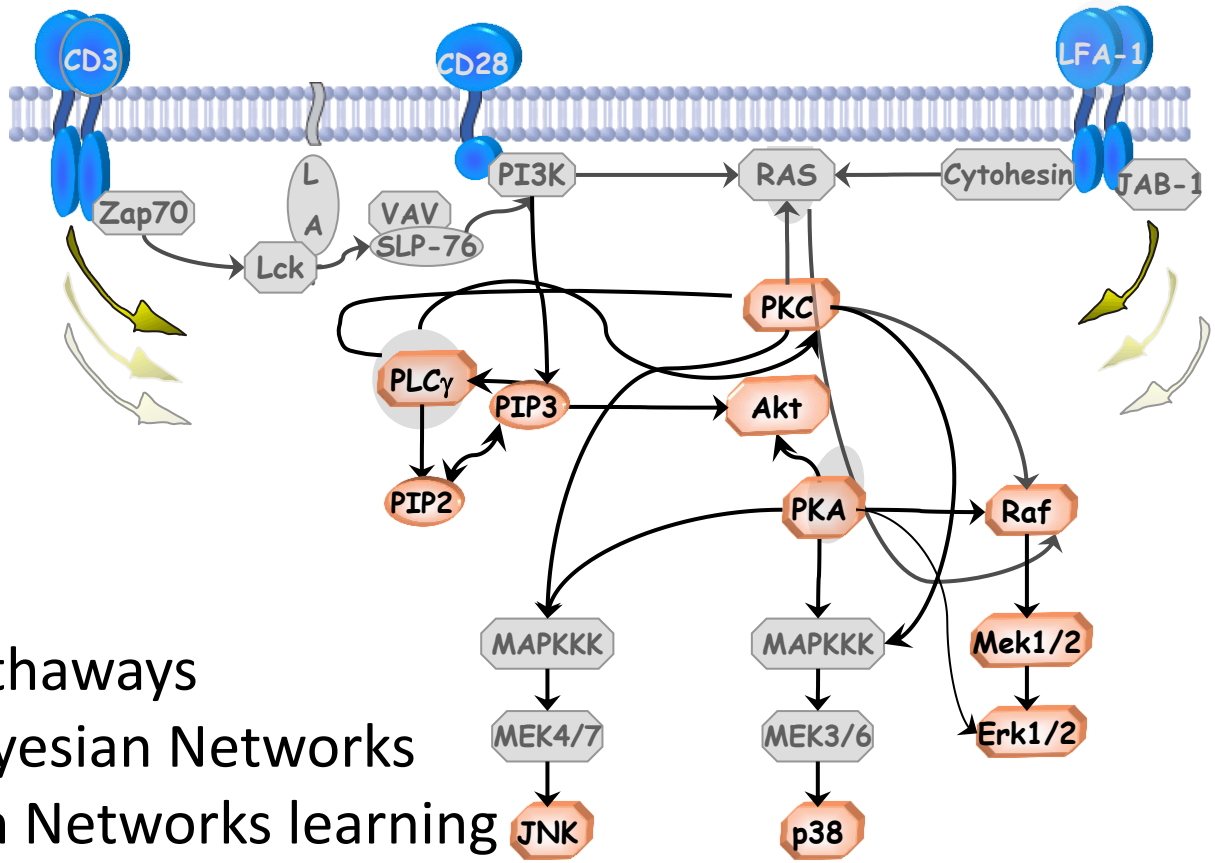
4 top-ranked genes were experimentally validated

An integrative genomics approach to infer causal associations between gene expression and disease

1. Identify loci susceptible for disease
   - 4 QTLs
2. Identify gene expression traits correlated with the disease
   - 440 genes
3. Identify genes with eQTLs that coincide with the QTLs
   - 113 genes, 267 eQTLs
4. Identify genes that support causal models
5. Rank genes by causal effect

One of them ranked 152 out of the 440 based on mere correlation

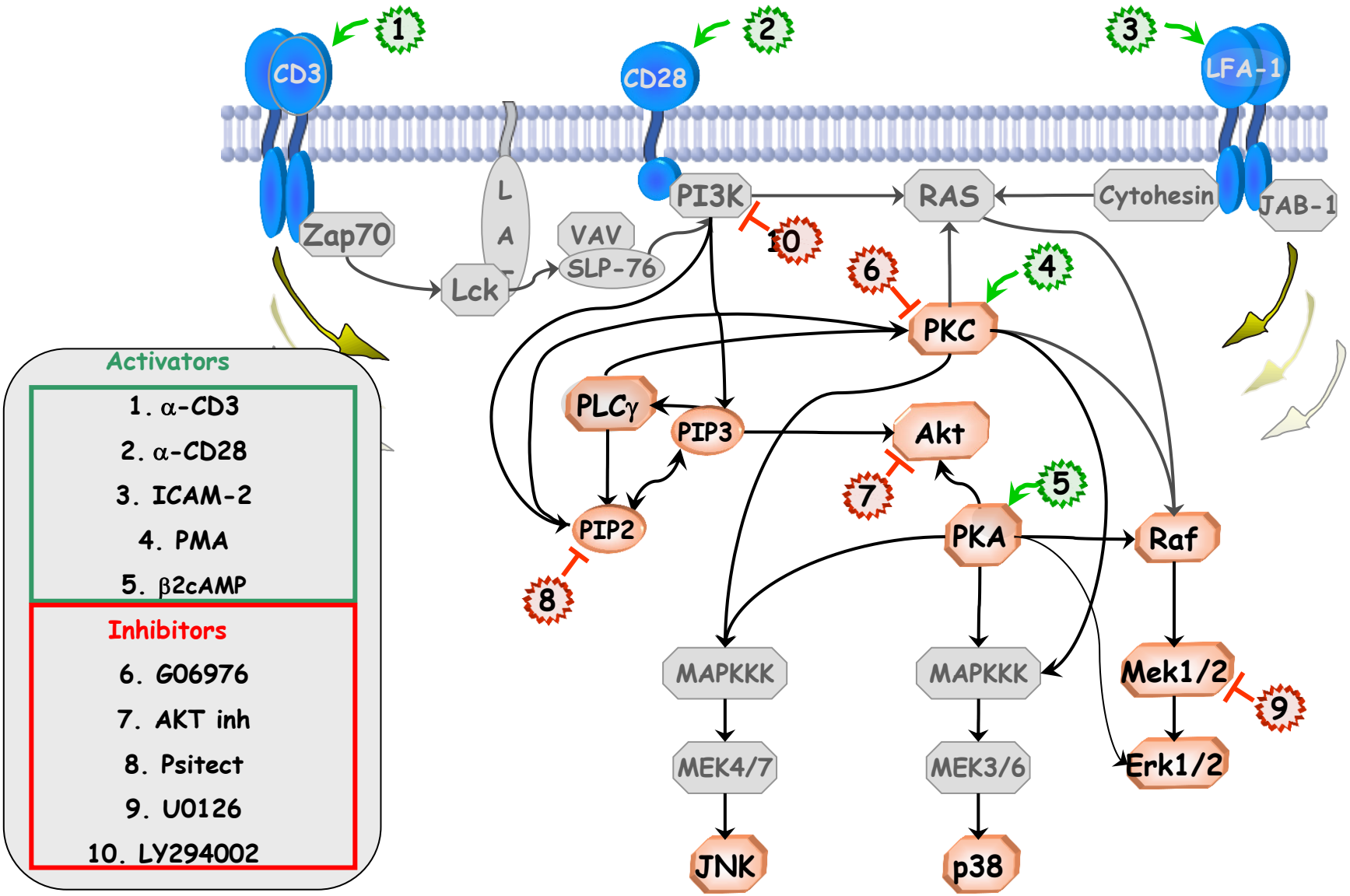
Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data



- Protein Signaling Pathaways resemble Causal Bayesian Networks
- Use Causal Bayesian Networks learning to reconstruct a Protein Signaling pathway

[K. Sachs, et al. ***Science*** **,** (2005)]

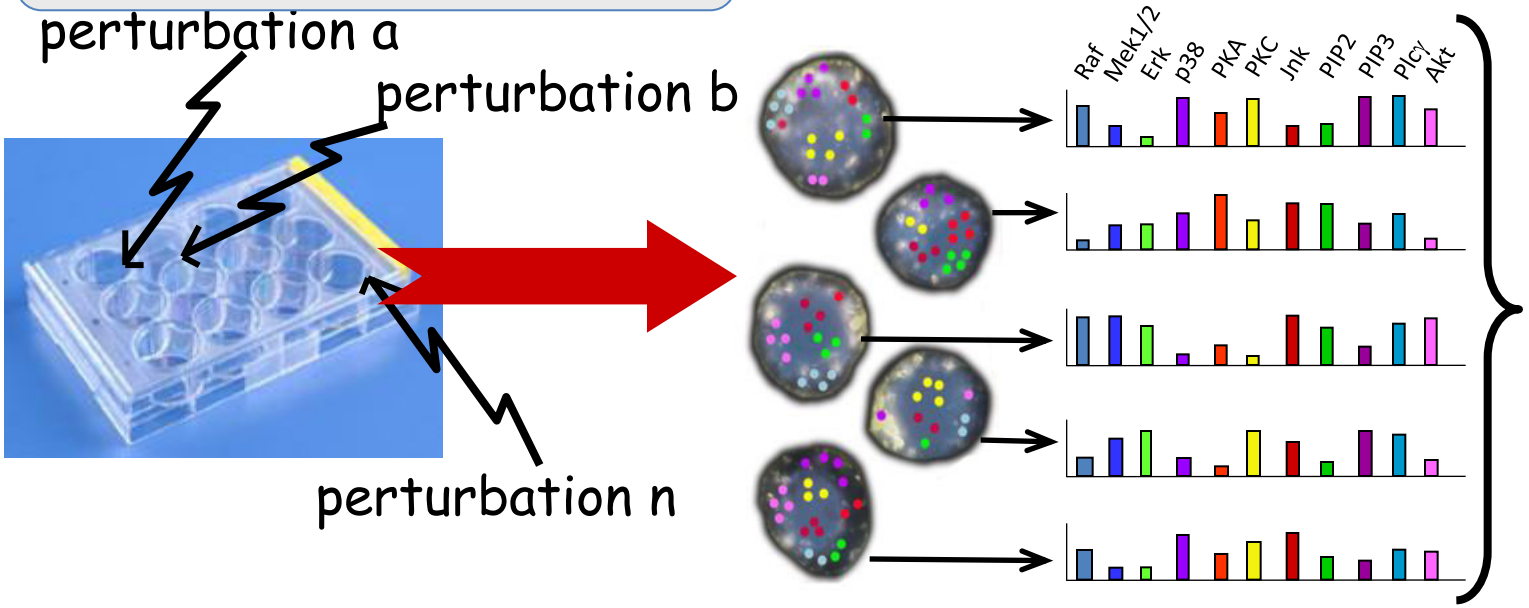
Stimulations and perturbations



T-Lymphocyte Data

Conditions (multi-well format)

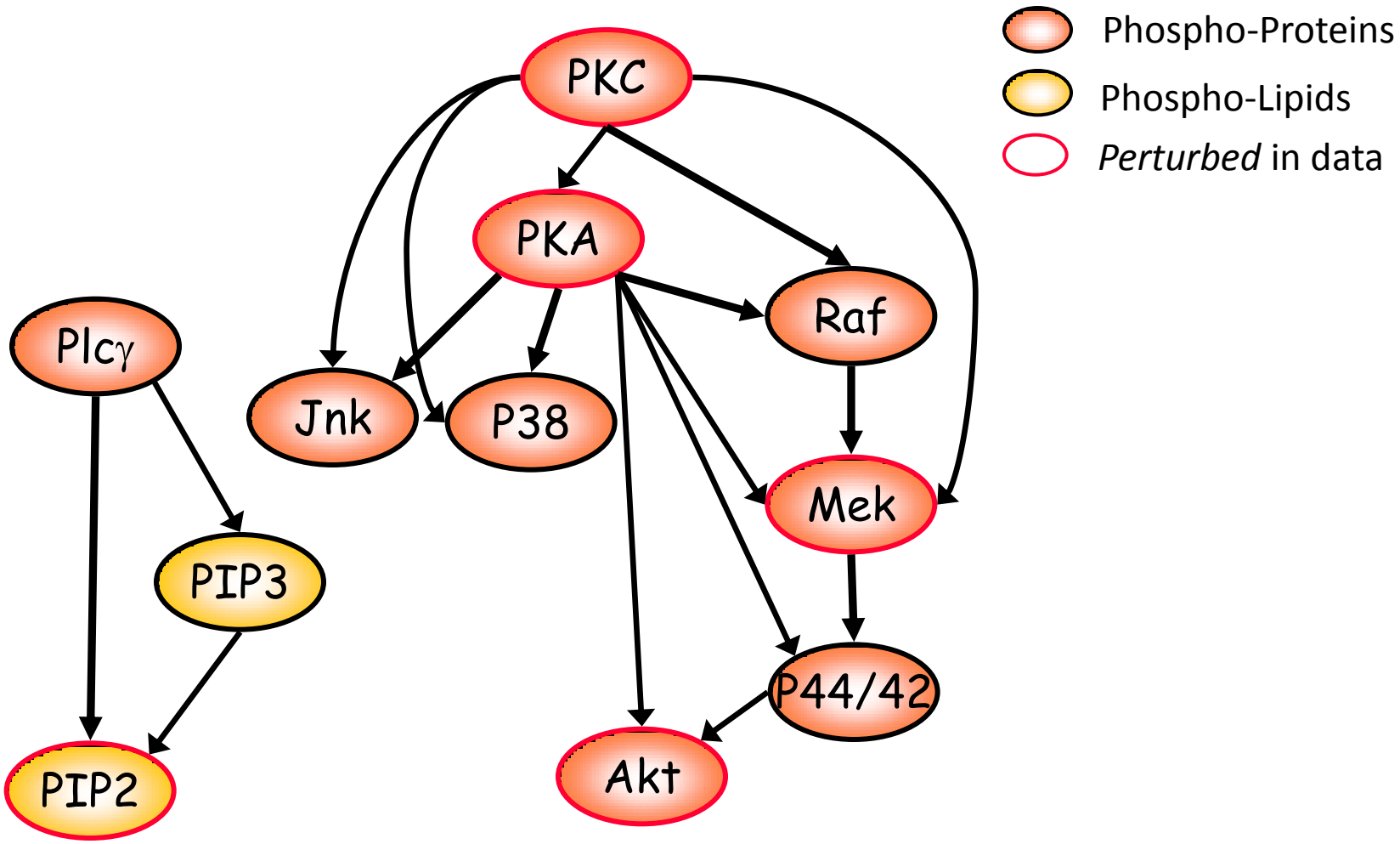
12 Color Flow Cytometry



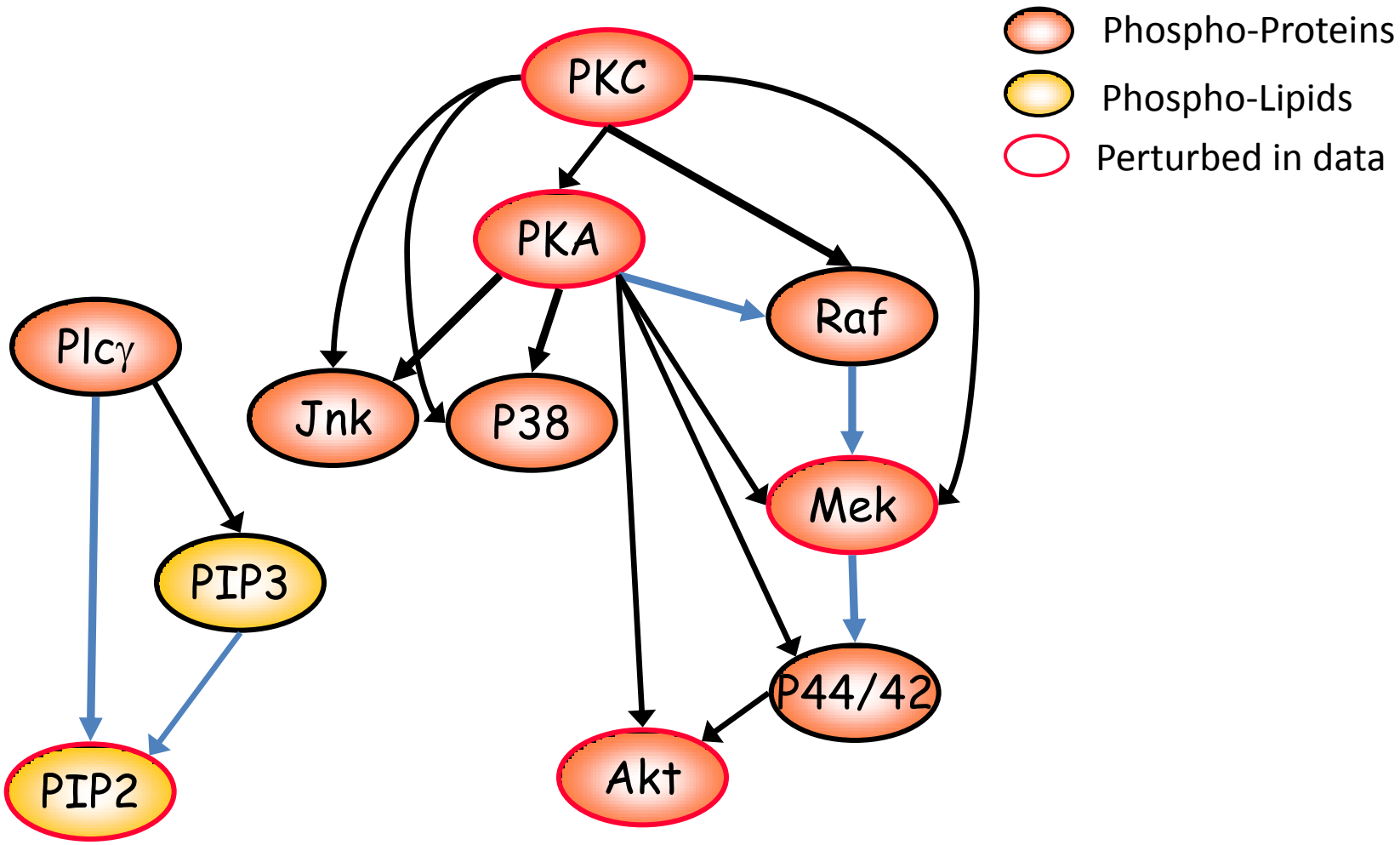
- Primary human T-Cells
- 9 conditions
  - (6 **Specific** interventions)

- 9 phosphoproteins, 2 phospolipids
- 600 cells per condition
  - 5400 data-points

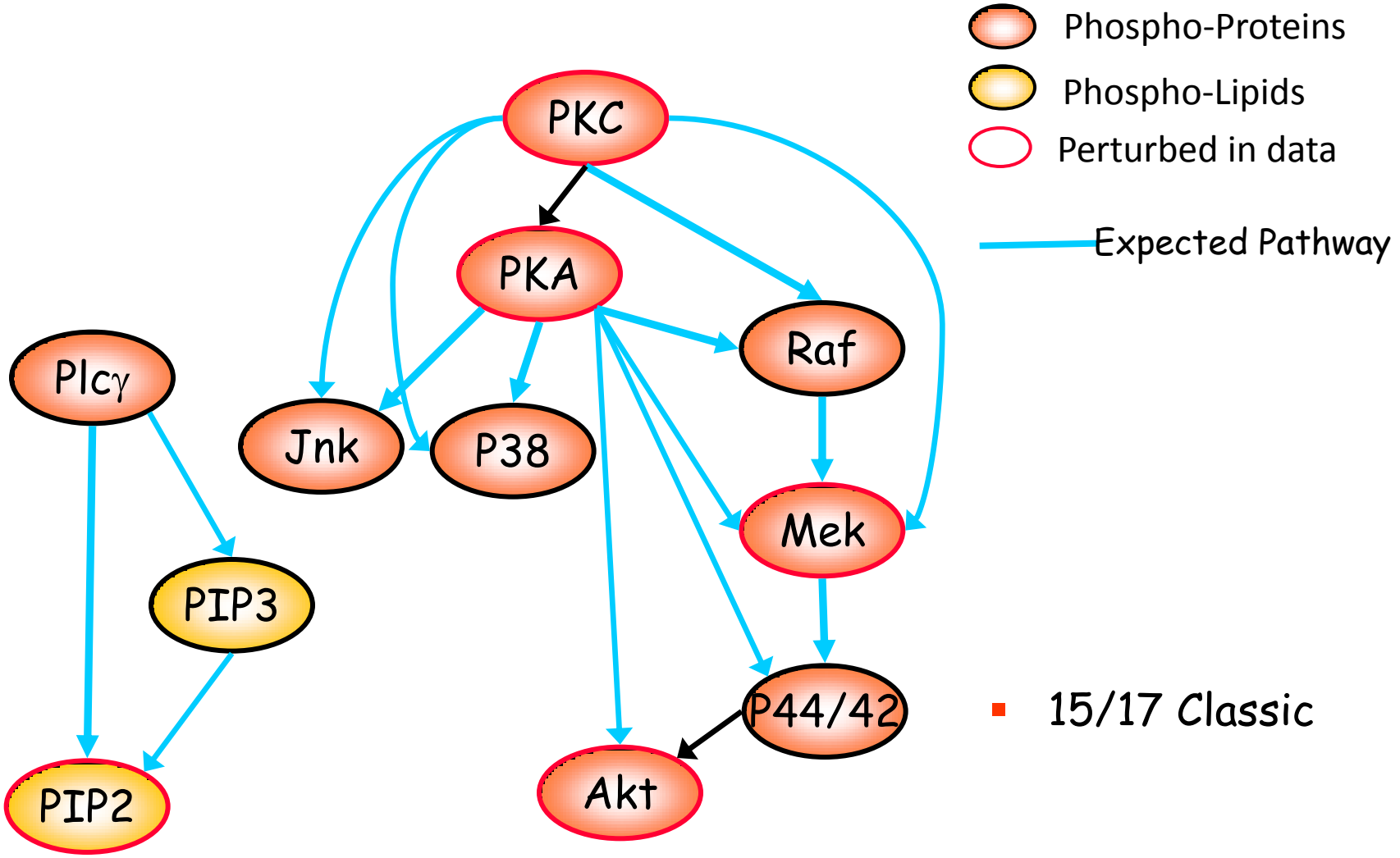
Inferred Network



How well did we do?

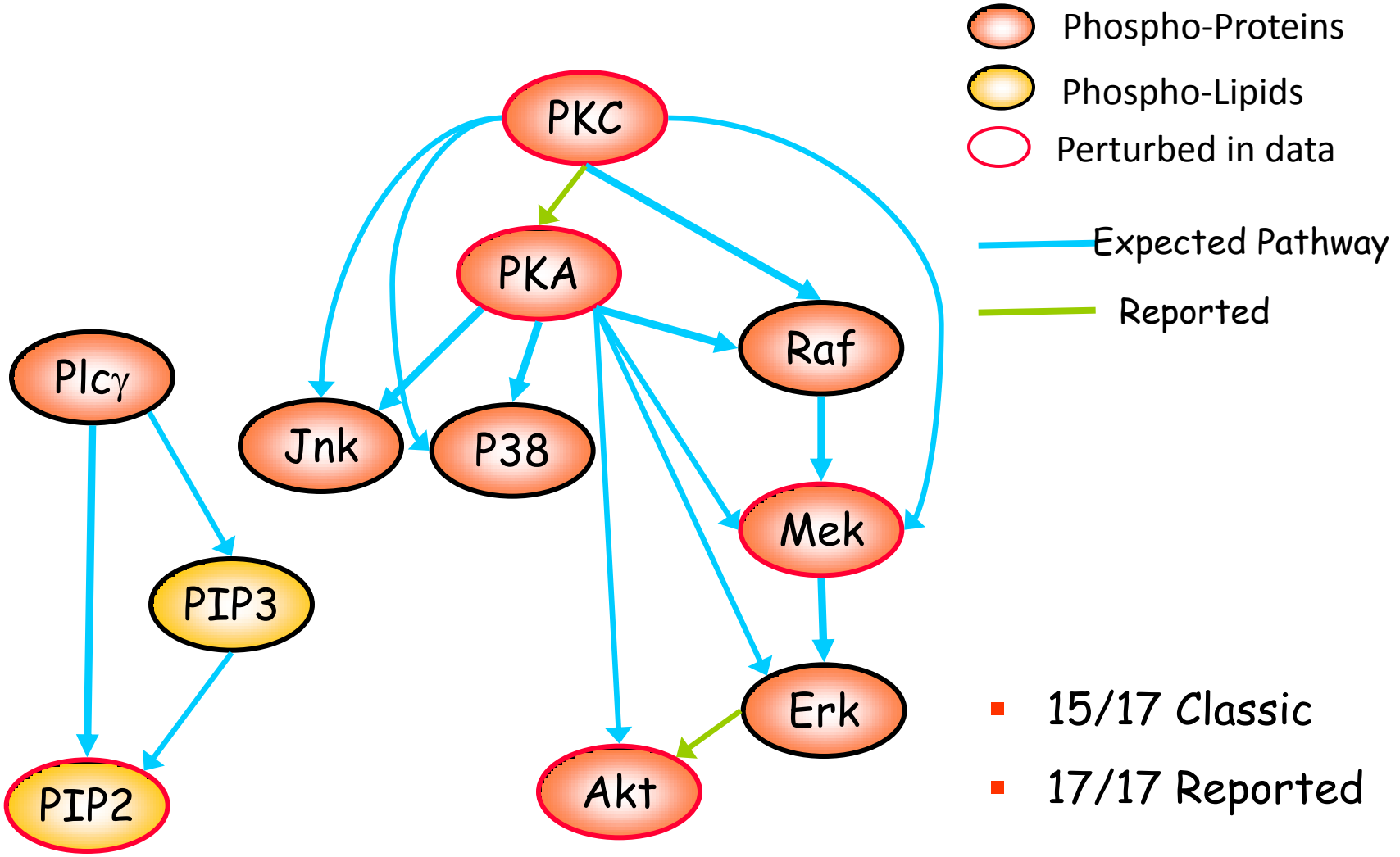


How well did we do?

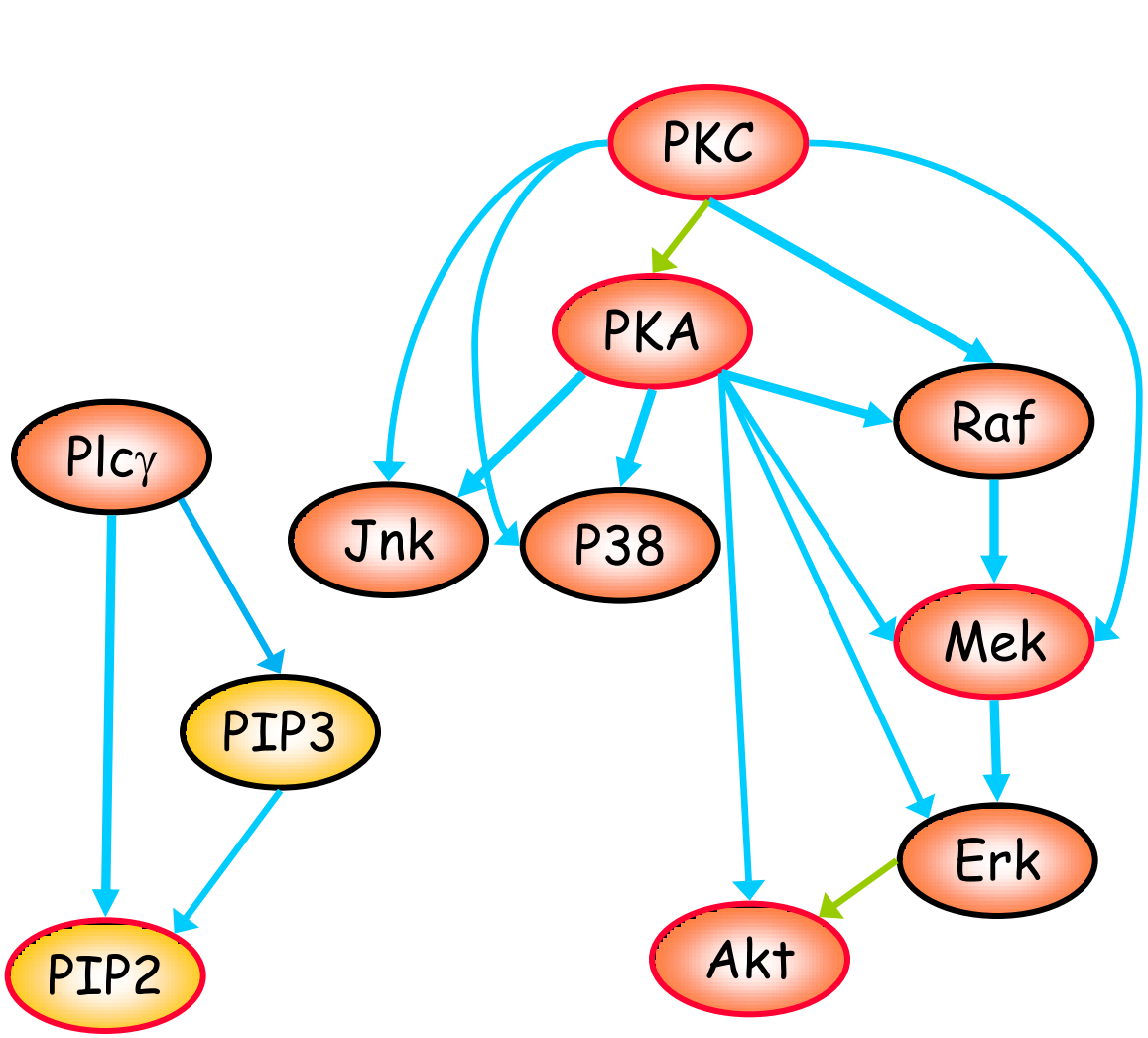


- 15/17 Classic

How well did we do?



How well did we do?



Phospho-Proteins

Phospho-Lipids

Perturbed in data

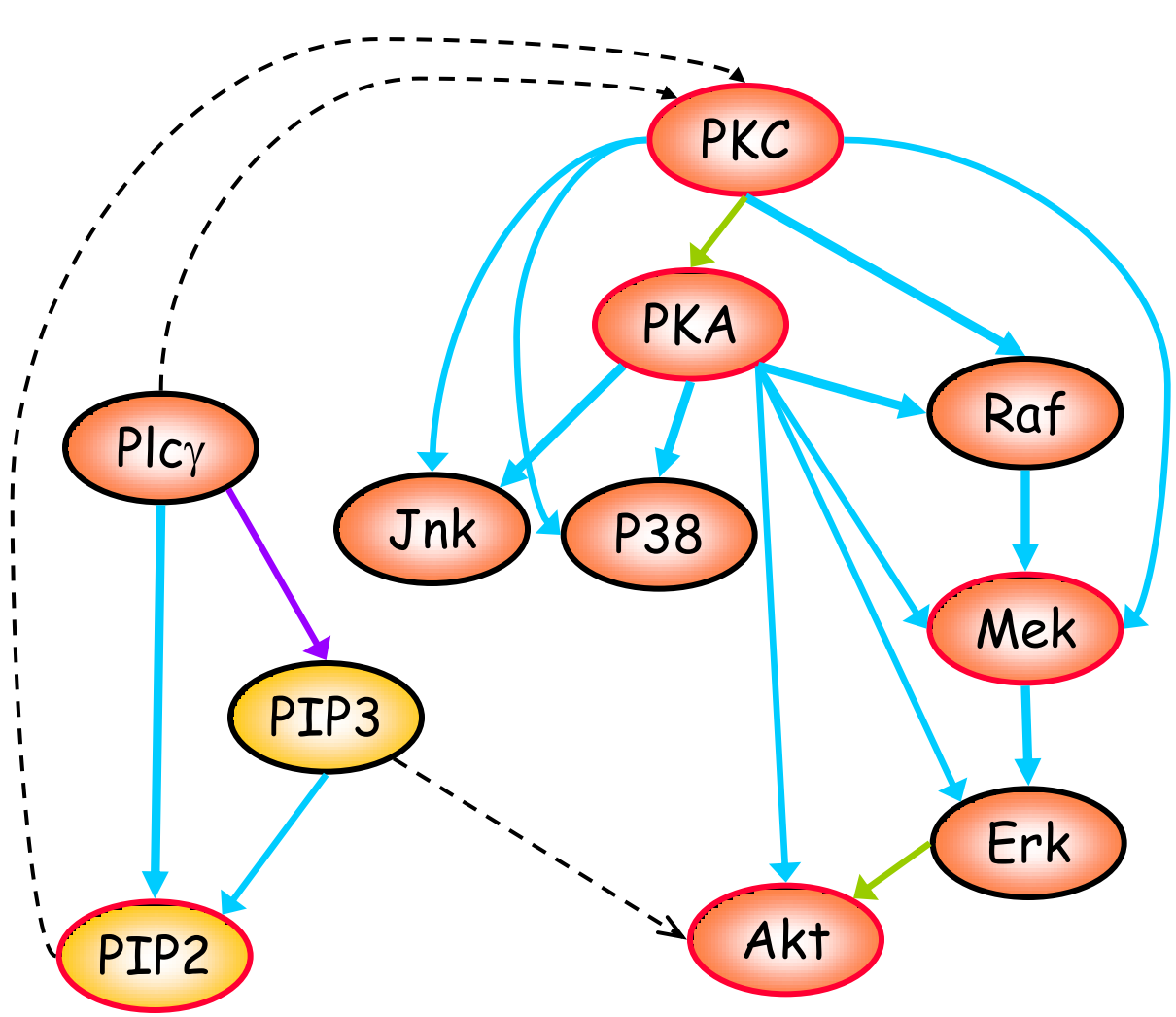
Expected Pathway

Reported

Reversed

- 15/17 Classic
- 17/17 Reported

How well did we do?

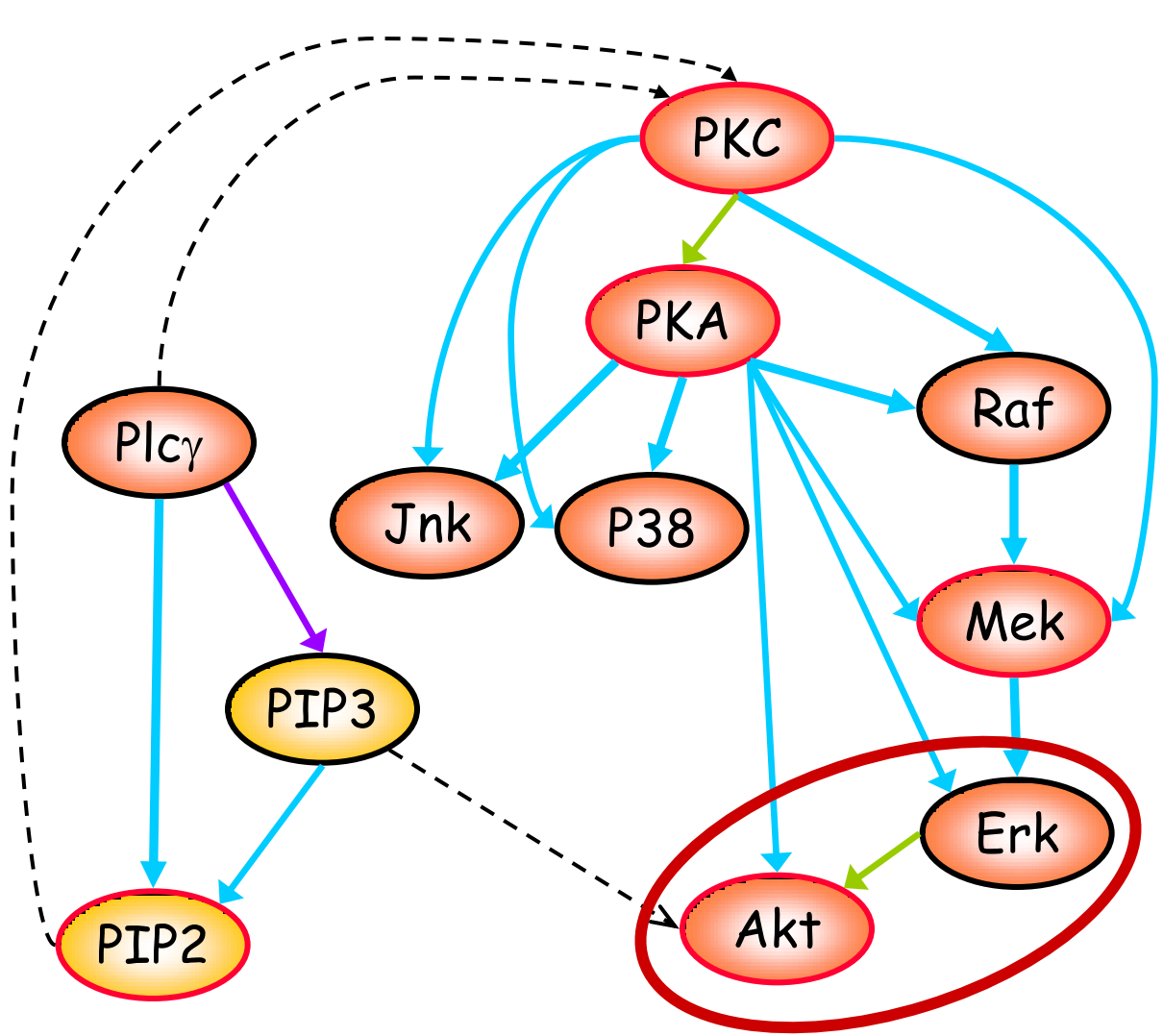


Phospho-Proteins
Phospho-Lipids
Perturbed in data

Expected Pathway
Reported
Reversed
Missed

- 15/17 Classic
- 17/17 Reported
- 3 Missed

How well did we do?



Phospho-Proteins
Phospho-Lipids
Perturbed in data

Expected Pathway
Reported
Reversed
Missed

- 15/17 Classic
- 17/17 Reported
- 3 Missed

Bibliography

- Tsamardinos, I. & Aliferis, C.F., 2003. Towards Principled Feature Selection: Relevancy, Filters and Wrappers. In *AI&STATS 2003*
- Aliferis, C.F., Tsamardinos, I. & Statnikov, A., 2003. HITON: a novel Markov Blanket algorithm for optimal variable selection. In *AMIA 2003*
- Tsamardinos, I., Aliferis, C.F. & Statnikov, A., 2003. Time and sample efficient discovery of Markov blankets and direct causal relations. In *KDD 03*
- Lagani, V. & Tsamardinos, I., 2010. Structure-based variable selection for survival data. *Bioinformatics*, 26(15), pp.1887-1894.
- Aliferis, C.F., et al., 2010. Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part I : Algorithms and Empirical Evaluation. *Journal of Machine Learning Research*, 11, pp.171-234.
- Aliferis, C.F. et al., 2010. Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part II : Analysis and Extensions. *Journal of Machine Learning Research*, 11, pp.235-284
- Maathuis, M.H. et al., 2010. Predicting causal effects in large-scale systems from observational data. *Nature Methods*, 7(4), pp.247-248.
- Maathuis, M.H., Kalisch, M. & Bühlmann, P., 2008. Estimating high-dimensional intervention effects from observational data. *Annals of Statistics*, 37(6A), pp.3133-3164.
- Schadt, E.E. et al., 2005. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics*, 37(7), pp.710-717.
- Sachs, K. et al., 2005. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721), pp.523-529.

Τέλος Ενότητας

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα στο Πανεπιστήμιο Κρήτης**» έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Σημειώματα

Σημείωμα αδειοδότησης (1)

- Το παρόν υλικό διατίθεται με τους όρους της άδειας χρήσης Creative Commons Αναφορά, Μη Εμπορική Χρήση, Όχι Παράγωγο Έργο 4.0 [1] ή μεταγενέστερη, Διεθνής Έκδοση. Εξαιρούνται τα αυτοτελή έργα τρίτων π.χ. φωτογραφίες, διαγράμματα κ.λ.π., τα οποία εμπεριέχονται σε αυτό και τα οποία αναφέρονται μαζί με τους όρους χρήσης τους στο «Σημείωμα Χρήσης Έργων Τρίτων».

[1] http://creativecommons.org/licenses/by-nc-nd/4.0/

Σημείωμα αδειοδότησης (2)

- Ως **Μη Εμπορική** ορίζεται η χρήση:
  - που δεν περιλαμβάνει άμεσο ή έμμεσο οικονομικό όφελος από την χρήση του έργου, για το διανομέα του έργου και αδειοδόχο
  - που δεν περιλαμβάνει οικονομική συναλλαγή ως προϋπόθεση για τη χρήση ή πρόσβαση στο έργο
  - που δεν προσπορίζει στο διανομέα του έργου και αδειοδόχο έμμεσο οικονομικό όφελος (π.χ. διαφημίσεις) από την προβολή του έργου σε διαδικτυακό τόπο

- Ο δικαιούχος μπορεί να παρέχει στον αδειοδόχο ξεχωριστή άδεια να χρησιμοποιεί το έργο για εμπορική χρήση, εφόσον αυτό του ζητηθεί.
-

Σημείωμα Αναφοράς

«Μηχανική Μάθηση. Introduction to causal discovery».
Έκδοση: 1.0. Ηράκλειο 2015. Διαθέσιμο από τη δικτυακή διεύθυνση:
https://opencourses.uoc.gr/courses/course/view.php?id=362.

Διατήρηση Σημειωμάτων

Οποιαδήποτε αναπαραγωγή ή διασκευή του υλικού θα πρέπει να συμπεριλαμβάνει:

- το Σημείωμα Αναφοράς
- το Σημείωμα Αδειοδότησης
- τη δήλωση Διατήρησης Σημειωμάτων
- το Σημείωμα Χρήσης Έργων Τρίτων (εφόσον υπάρχει)

μαζί με τους συνοδευόμενους υπερσυνδέσμους.