Introduction to causal discovery: A Bayesian Networks approach

Ioannis Tsamardinos[1, 2]

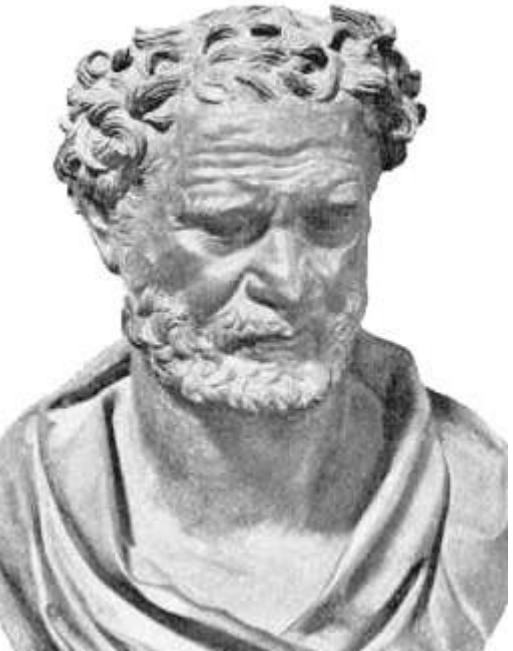
Sofia Triantafillou[1, 2]

Vincenzo Lagani[1]

[1]Bioinformatics Laboratory,
Institute of Computer Science,
Foundation for Research and Tech., Hellas

[2]Computer Science Department
University of Crete

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



“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?

- *A* causes *B*:
  - A causally affects B
  - Probabilistically
  - Intervening onto values of *A* will affect the distribution of *B*
  - in some appropriate context

Statistical Association (Unconditional Dependency)

- Dep(*X, Y* | ∅)

- *X* and *Y* are associated
  - Observing the value of *X* may change the conditional distribution of the (observed) values of *Y: P(Y | X) ≠ P(Y)*
  - Knowledge of *X* provides information for *Y*
  - Observed *X* is predictive for observed *Y* and vice versa
  - Knowing *X* changes our beliefs for the distribution of *Y*

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

- Several means for measuring it

Association is NOT Causation

- 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?

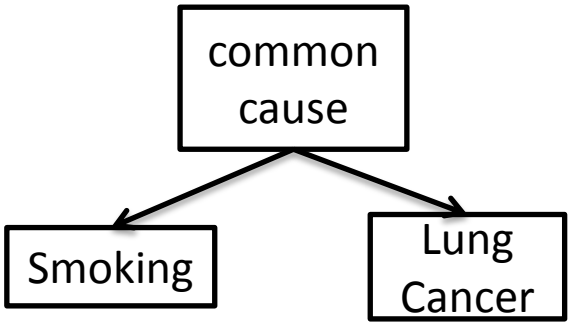
BUT

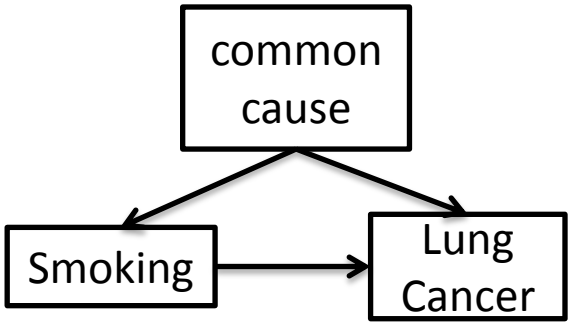
"If A and B are correlated, A causes B OR B causes A OR they share a latent common cause"

[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

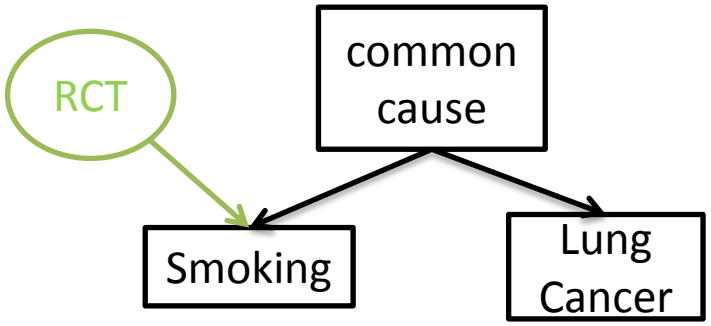
1. Take 200 people
2. Randomly split them in control and treatment groups
3. Force control group to smoke, force treatment group not to smoke
4. Wait until they are 60 years old
5. Measure correlation

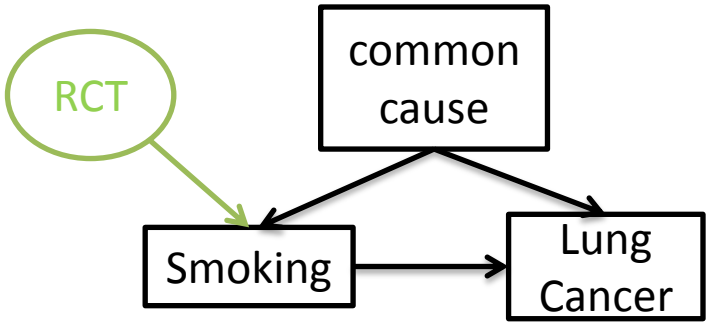
[Randomized Control Trial]

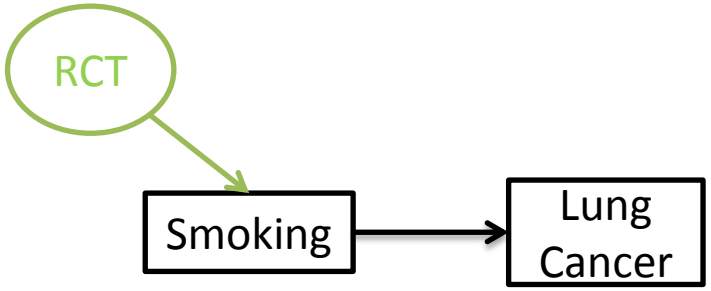
[Sir Ronald Fisher]

Manipulation

All possible models*

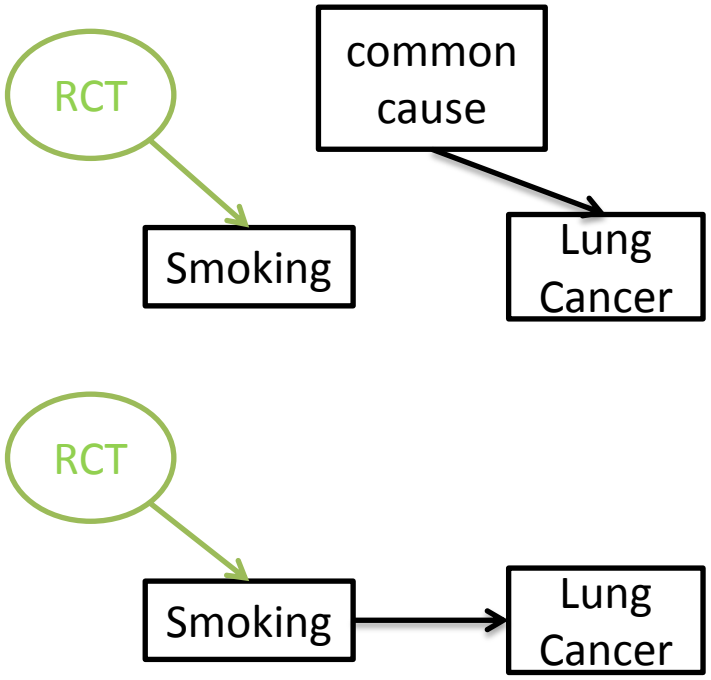


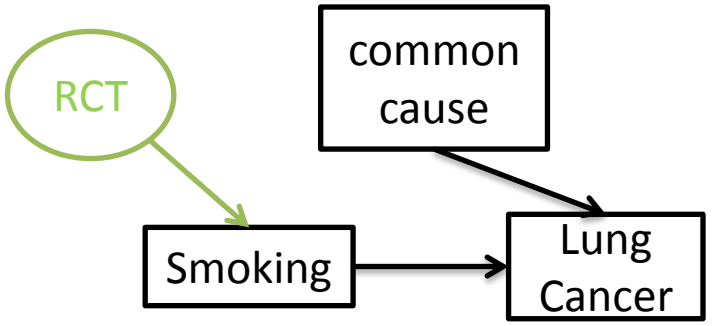




Manipulation removes other causes

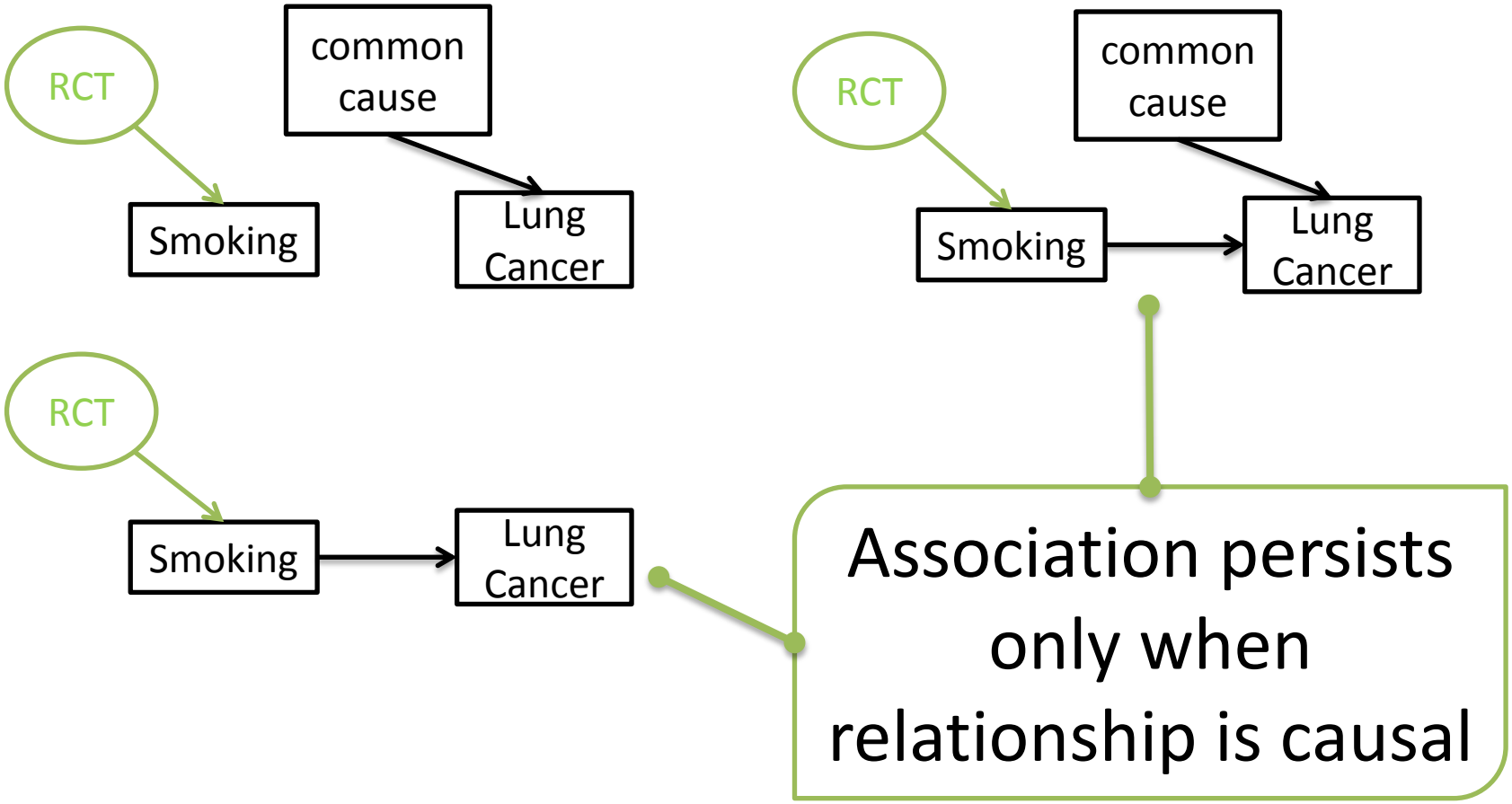
All possible models*





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]

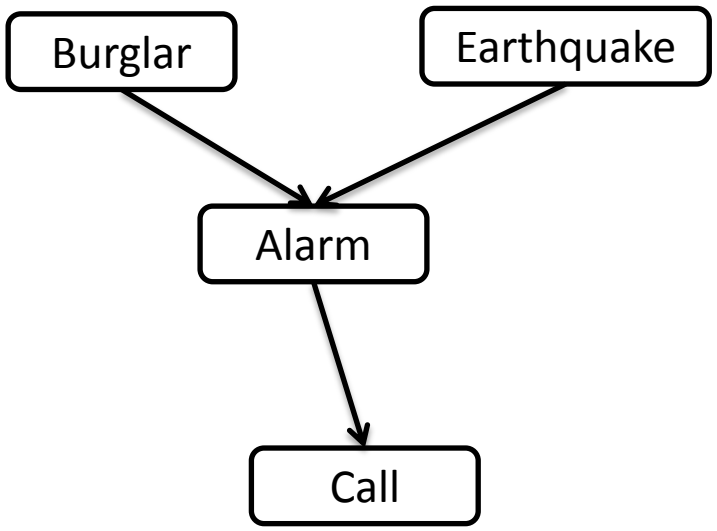
Conditional Association (Conditional Dependency)

- Dep(*X, Y* | **Z**)

- *X* and *Y* are associated conditioned on ***Z***
  - For some values of ***Z*** (some context)
  - Knowledge of *X* still provides information for *Y*
  - Observed *X* is still predictive for observed *Y* and vice versa

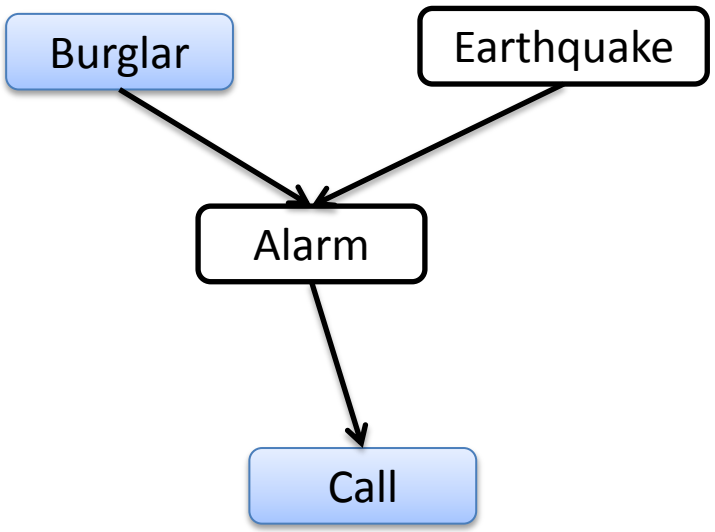
- Statistically estimable

Conditioning and Causality



[example by Judea Pearl]

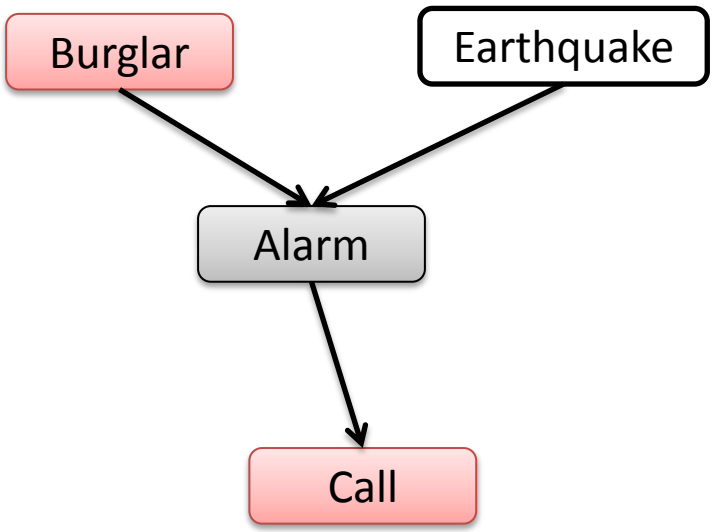
Conditioning and Causality



Dep(Burglar, Call|∅)

Burglar provides information for Call

Conditioning and Causality

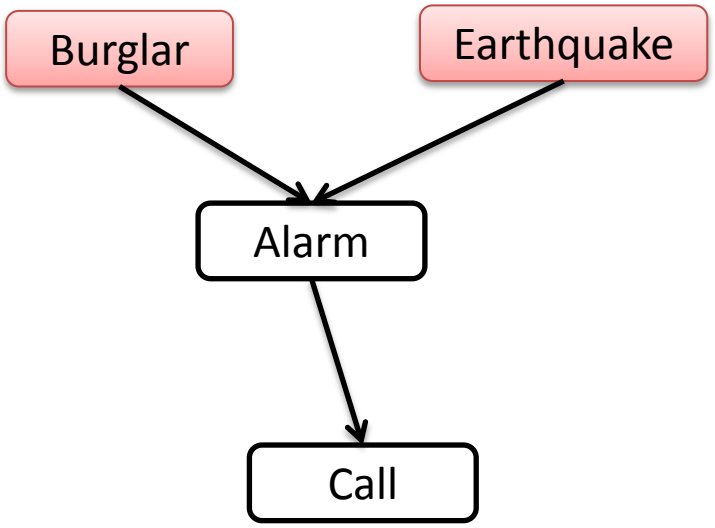


Learning the value of intermediate and common causes renders variables independent

Ind (Burglar, Call| Alarm)

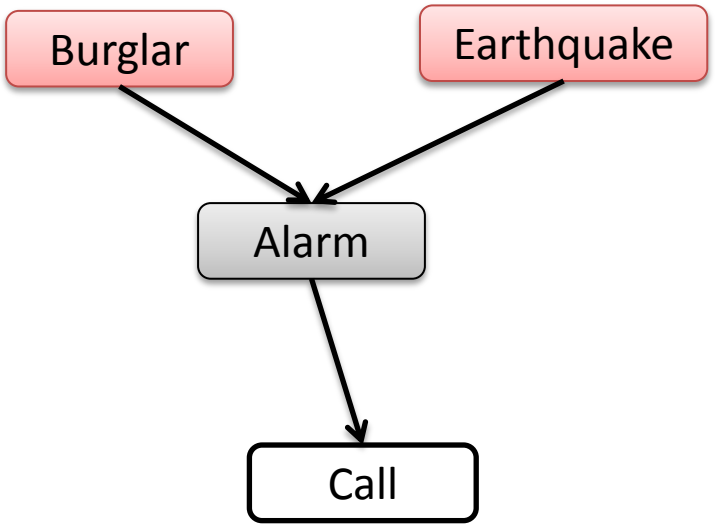
Burglar provides no information for Call once Alarm is known

Conditioning and Causality



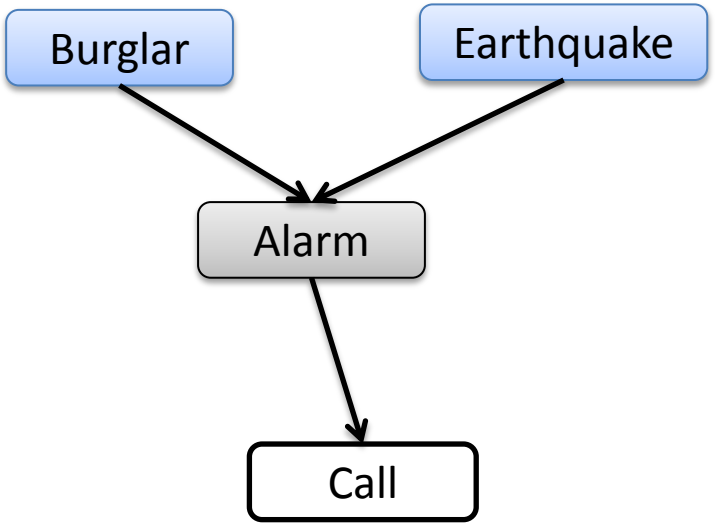
Ind (Burglar, Earthquake|∅)

Conditioning and Causality



Ind (Burglar, Earthquake|∅)

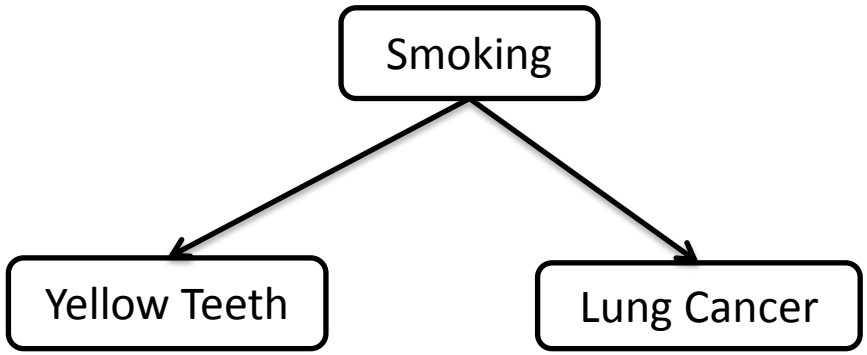
Conditioning and Causality



Learning the value of common effects renders variables dependent

Dep (Burglar, Earthquake| Alarm)

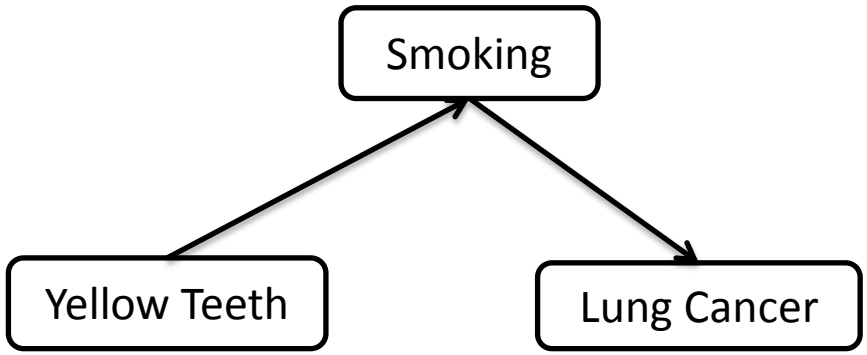
Observing data from a causal model



What would you observe?

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

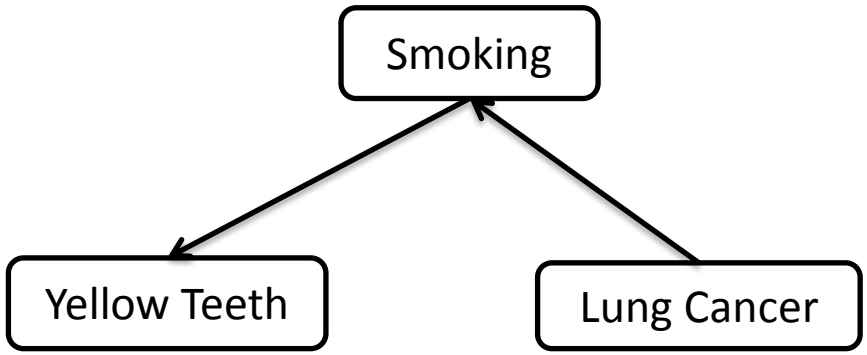
Observing data from a causal model



What would you observe?

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

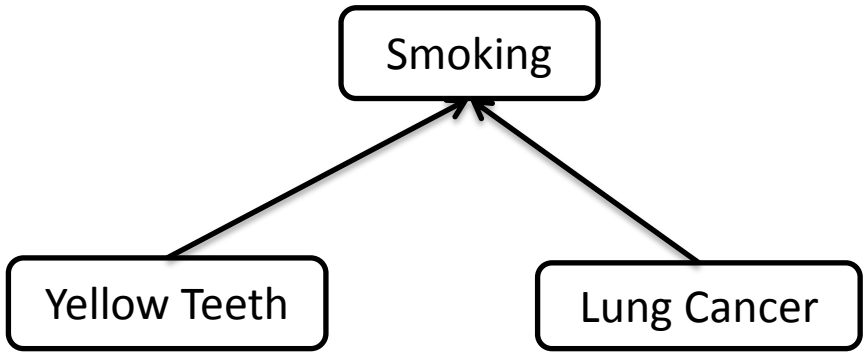
Observing data from a causal model



What would you observe?

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

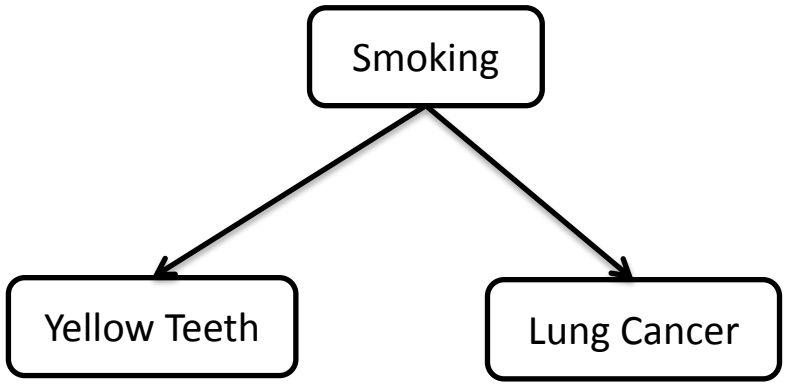
Observing data from a causal model

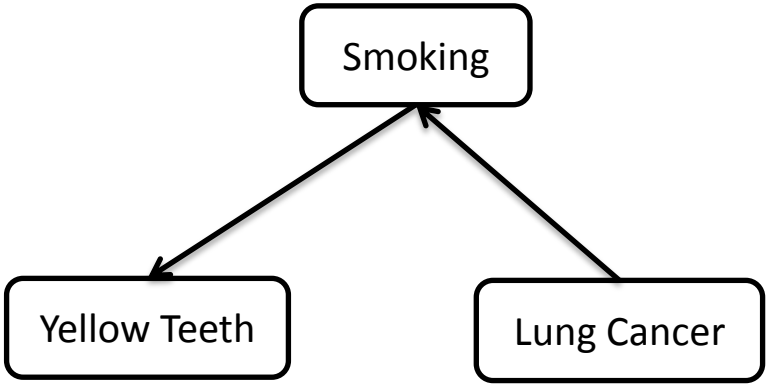


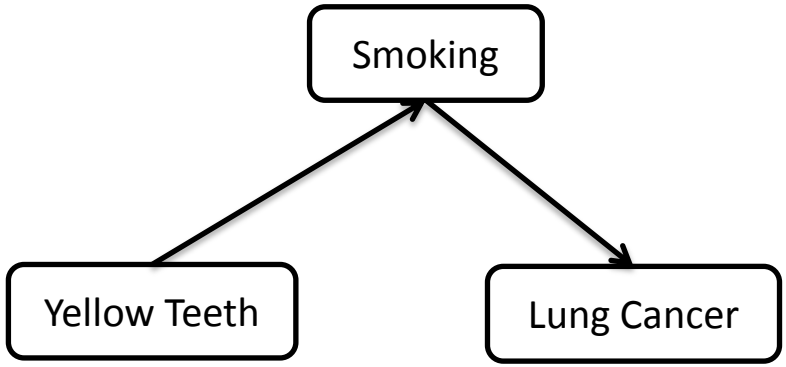
What would you observe?

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

Markov Equivalent Networks



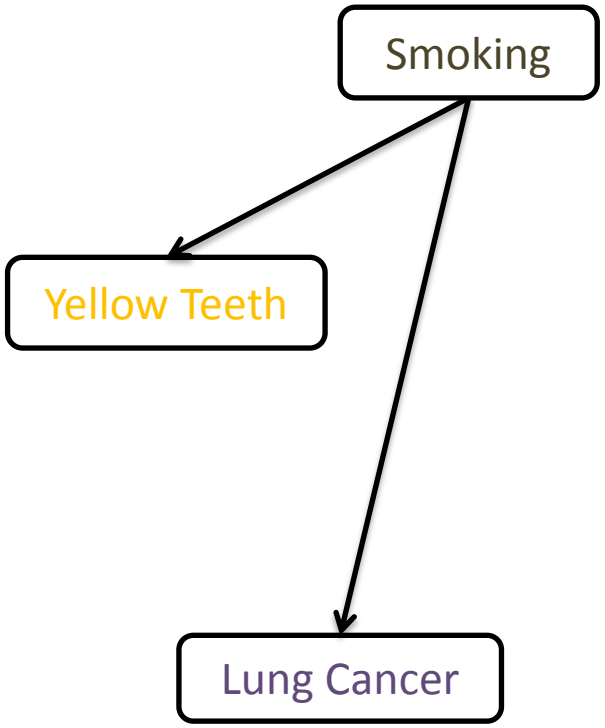




- Same conditional Independencies
- Same skeleton
- Same v-structures (subgraphs $X \rightarrow Y \leftarrow Z$ no X-Z)

Causal Bayesian Networks*

Graph $\mathcal{G}$



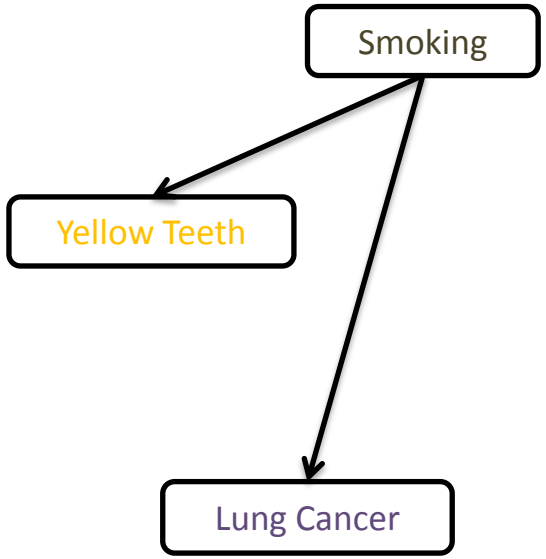
JPD $\mathcal{J}$

		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 $\mathcal{J}$ and allow reasoning

*almost there

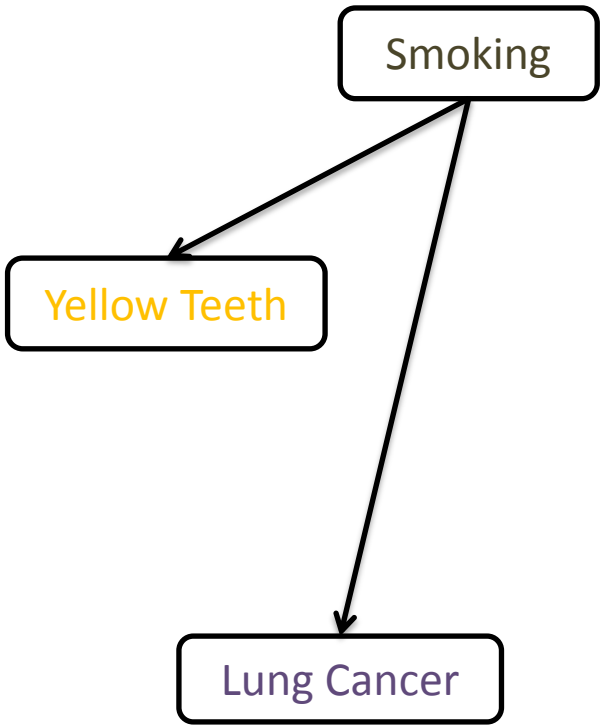
Causal Markov Condition (CMC)



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

Causal Markov Condition

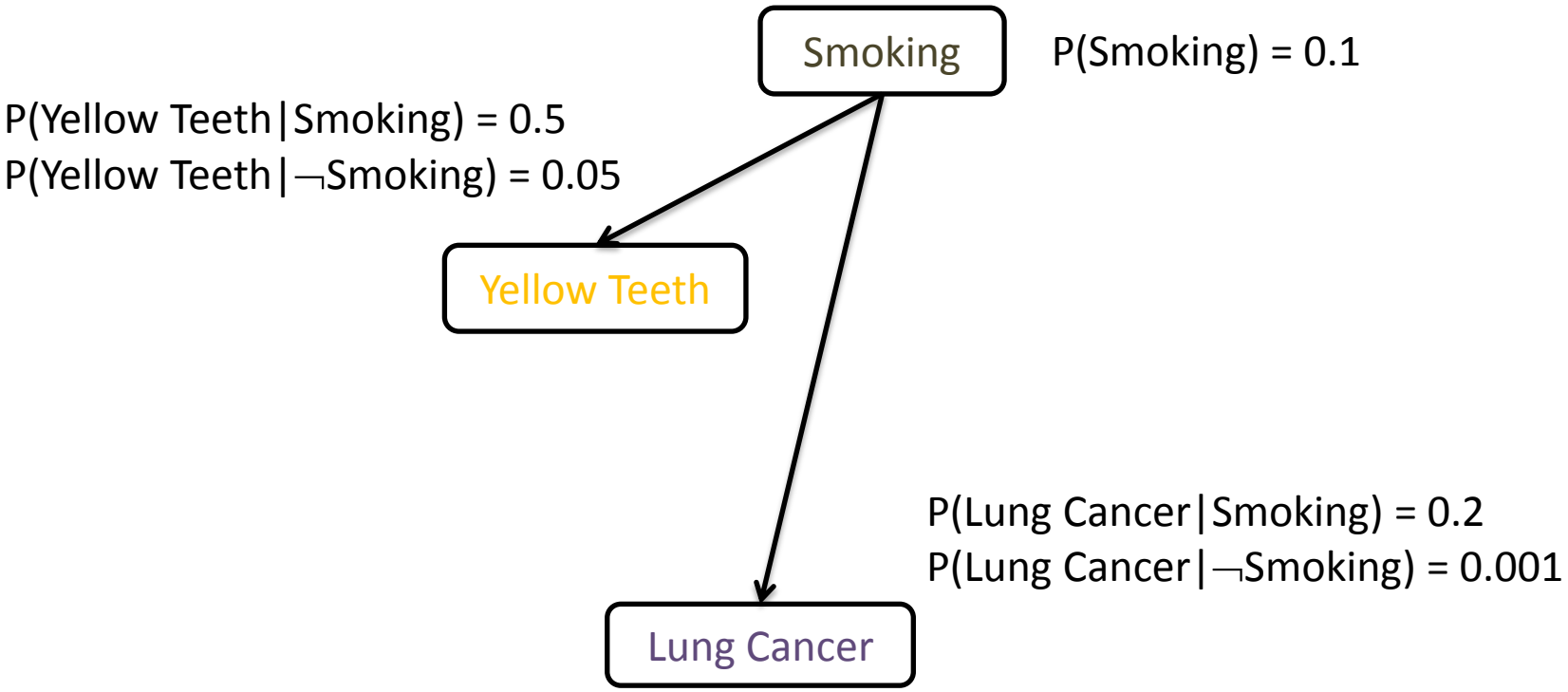


P(Yellow Teeth, Smoking, Lung Cancer) =
P(Smoking) ×
P(Yellow Teeth| Smoking) ×
P(Lung Cancer| Smoking)

$$P(\mathbf{V}) = \prod P(V_i \mid Pa(V_i))$$

Factorization with the CMC

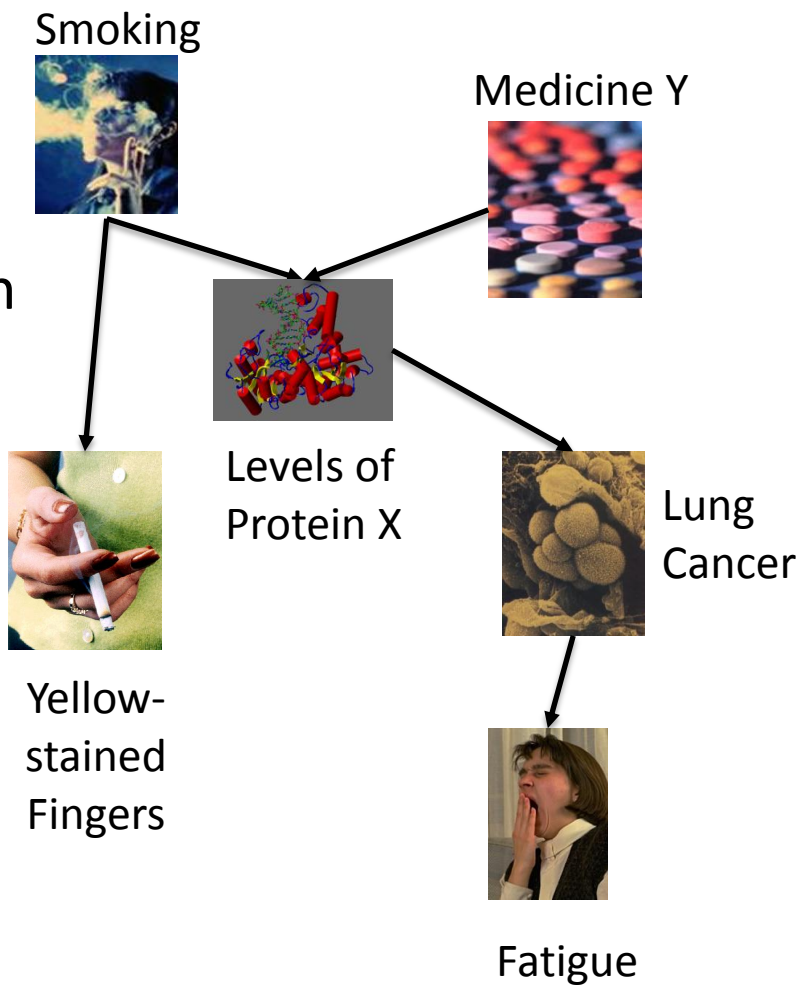
P(Yellow Teeth, Smoking, Lung Cancer) =
P(Smoking) $\times$
P(Yellow Teeth| Smoking) $\times$
P(Lung Cancer| Smoking)



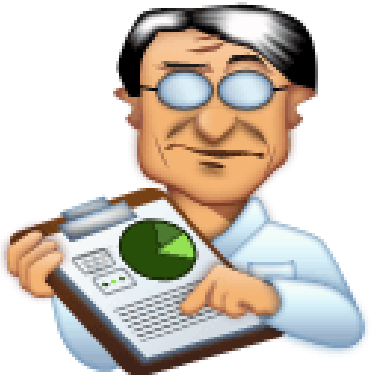
Using a **Causal** Bayesian Network

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**)?



Bayesian Networks



I don’t like all these assumptions

I kind of liked reducing the parameters of the distribution

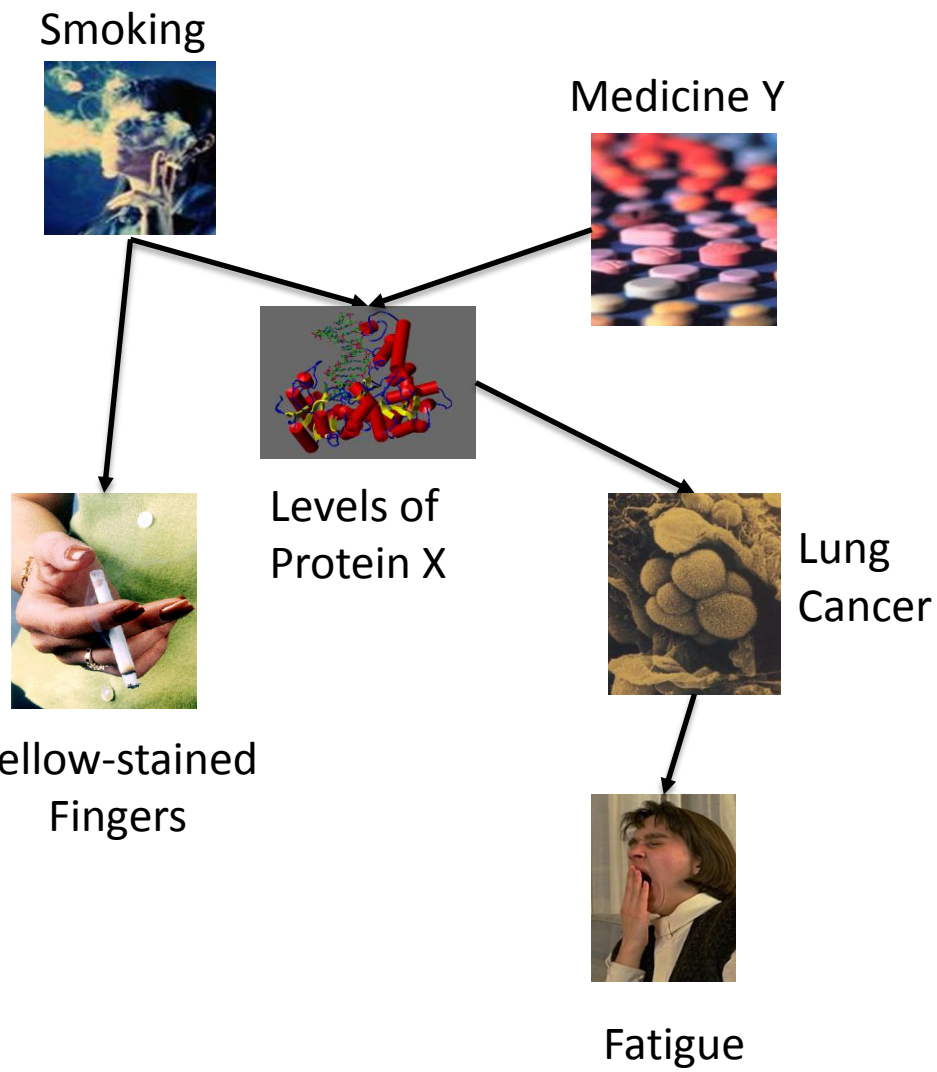
Drop the Causal part!

Using a Bayesian Network

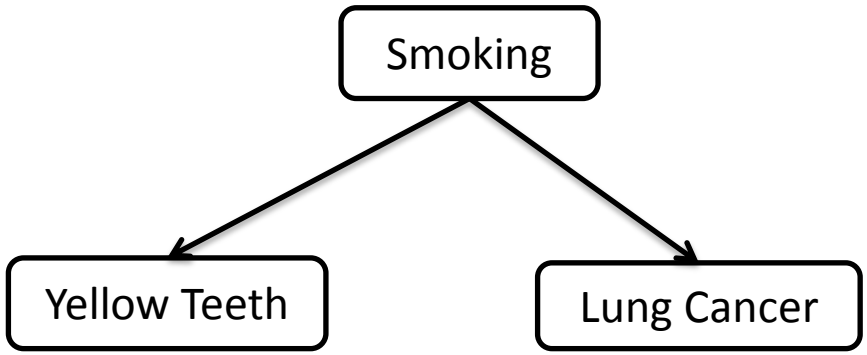
1. Factorize the jpd

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

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



Observing a causal model



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

What would you observe?

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

Learning Set of Equivalent Networks

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

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 [Tsamardinos 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]

Many more!!!

Assumptions*

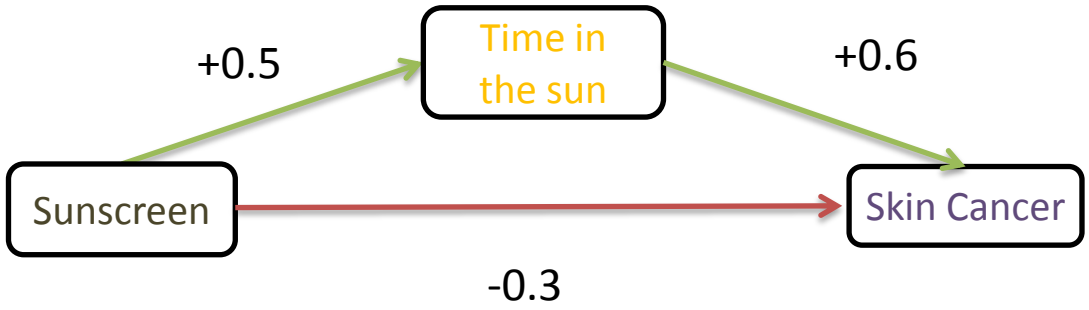
- Tests of Conditional Independences / Scoring methods may not be appropriate for the type of data at hand

- Faithfulness
- No feedback cycles
- No determinism
- No latent variables
- No measurement error
- No averaging effects

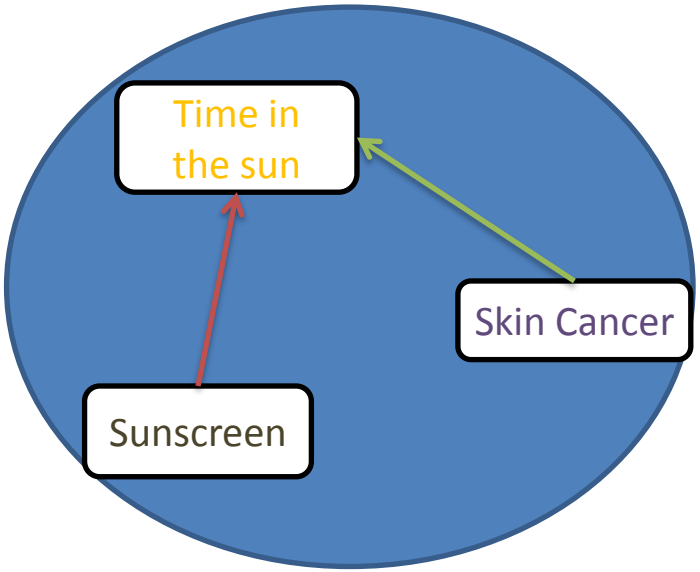
*aka why the algorithms may not work

Faithfulness

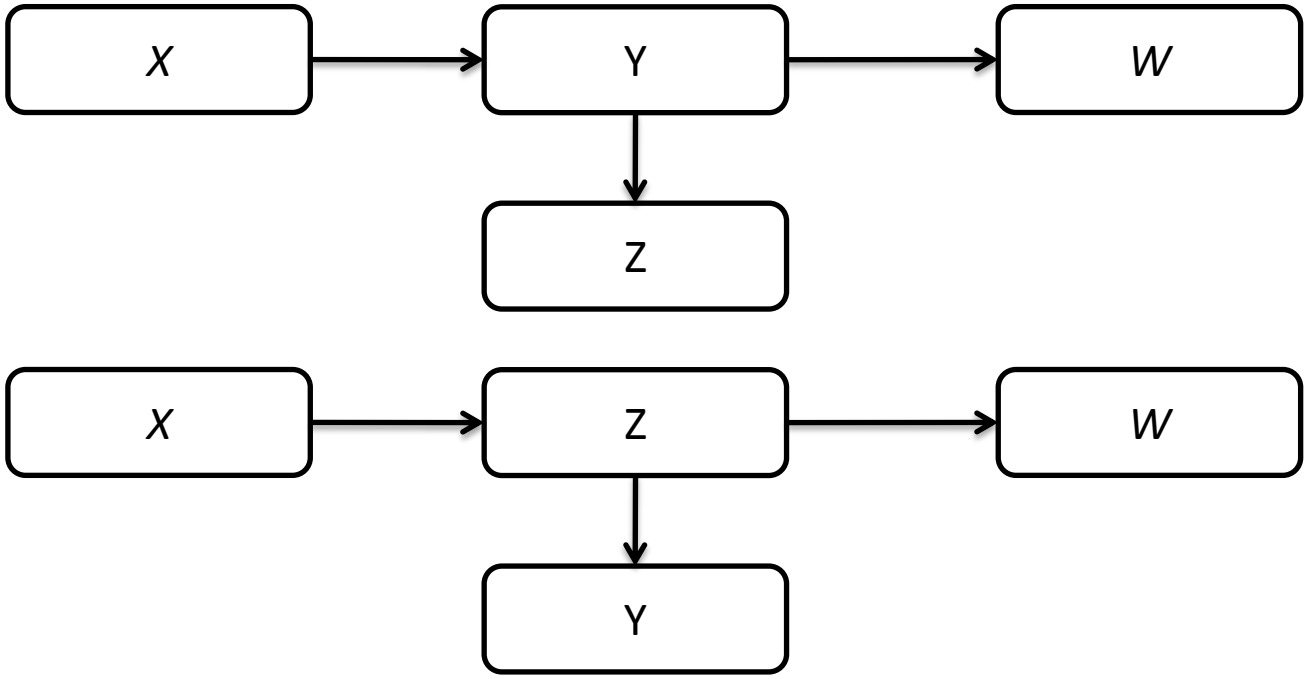
- ALL conditional (in)dependencies stem from the CMC



- Markov Condition does not imply:
  - Ind(Skin Cancer, Sunscreen)
- Unfaithful if:
  - Ind(Skin Cancer, Sunscreen)

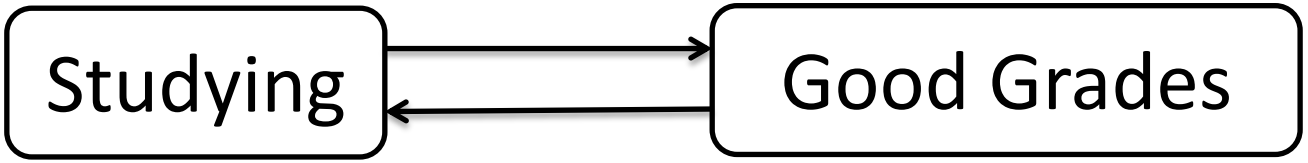


Collinearity and Determinism



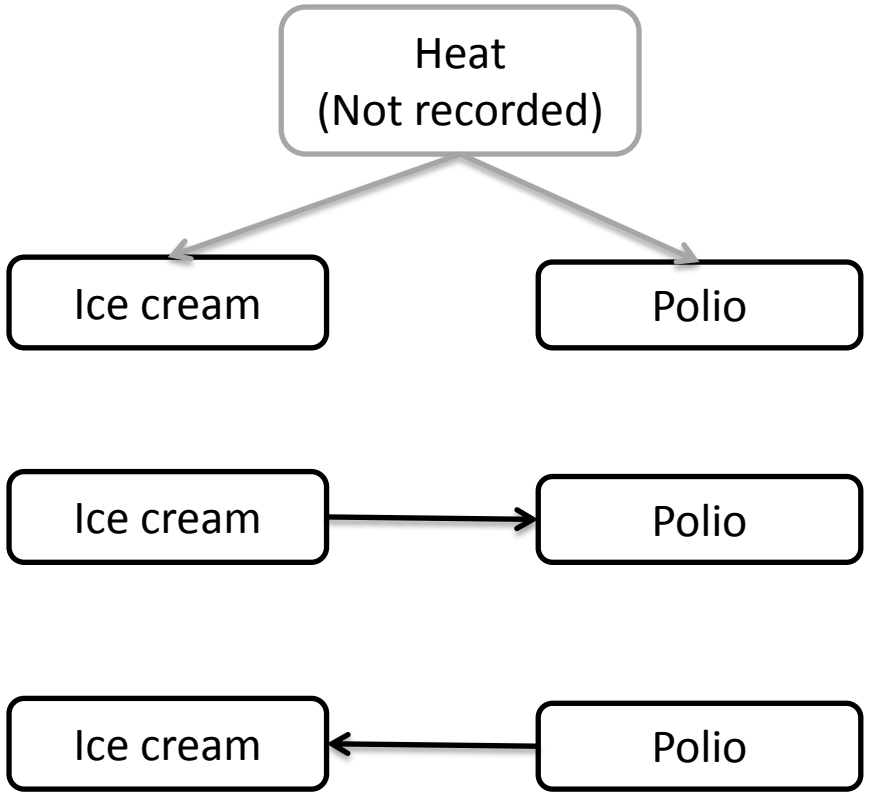
- Assume *Y* and *Z* are information equivalent (e.g., one-to-one deterministic relation)
- Cannot distinguish the two graphs
- A specific type of violation of Faithfulness

No Feedback Cycles



- 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 (mathematically) and metaphorically

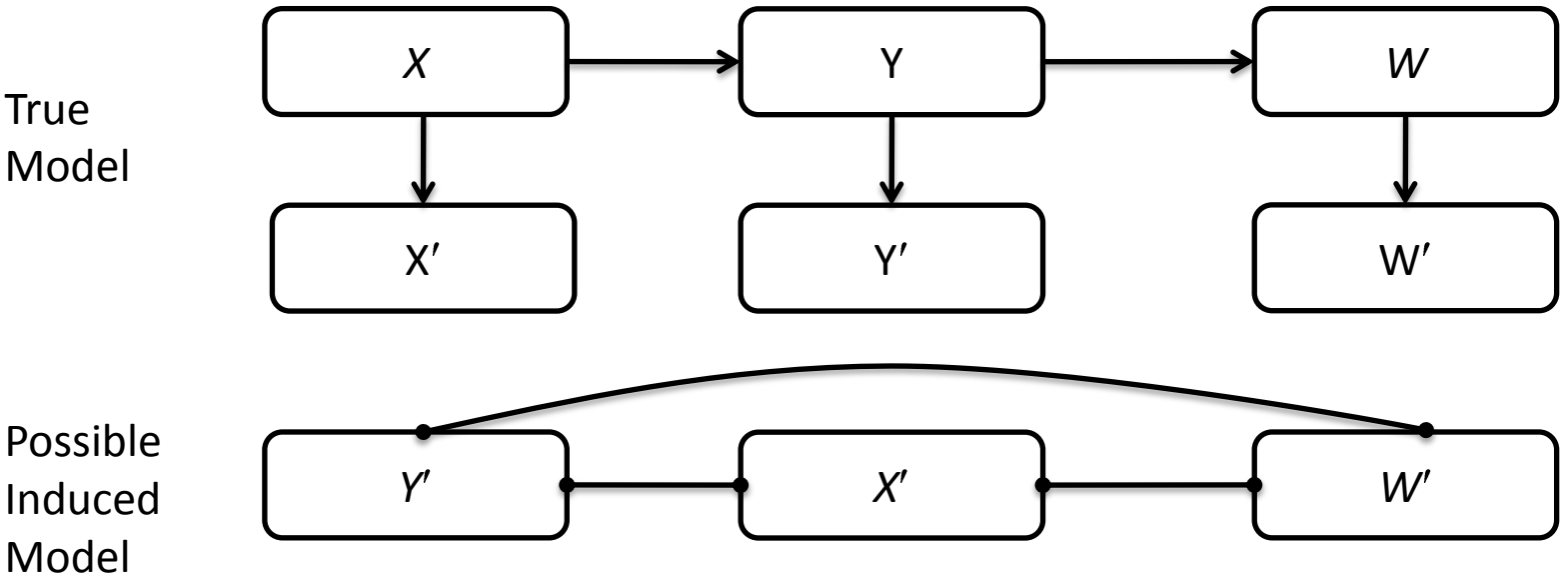
No Latent Confounders



- Dep (Ice Cream, Polio)
- No CAUSAL Bayesian Network on the modeled variables ONLY captures causal relations correctly

- Both Bayesian Networks capture associations correctly (not always the case)

Effects of Measurement Error



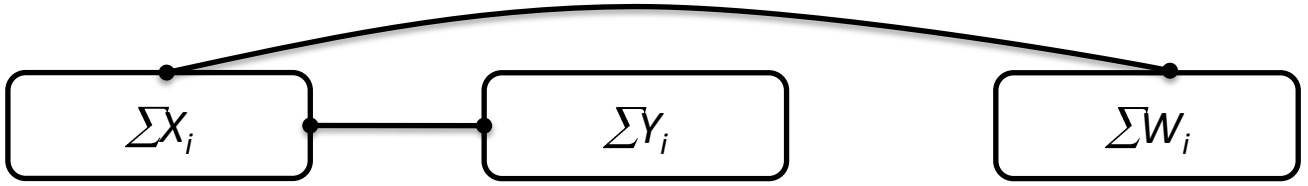
- *X, Y, Z* the actual physical quantities
- *X′, Y′, Z′* the measured quantities (+ noise)
- If *Y′* is measured with more error than *X′* then Dep(*X′;W′ | Y′*)

Effects of Averaging

- Almost all omics technologies measure average quantities over millions of cells
- The quantities in the models though refer to single-cells

True Model

Possibly Induced Model



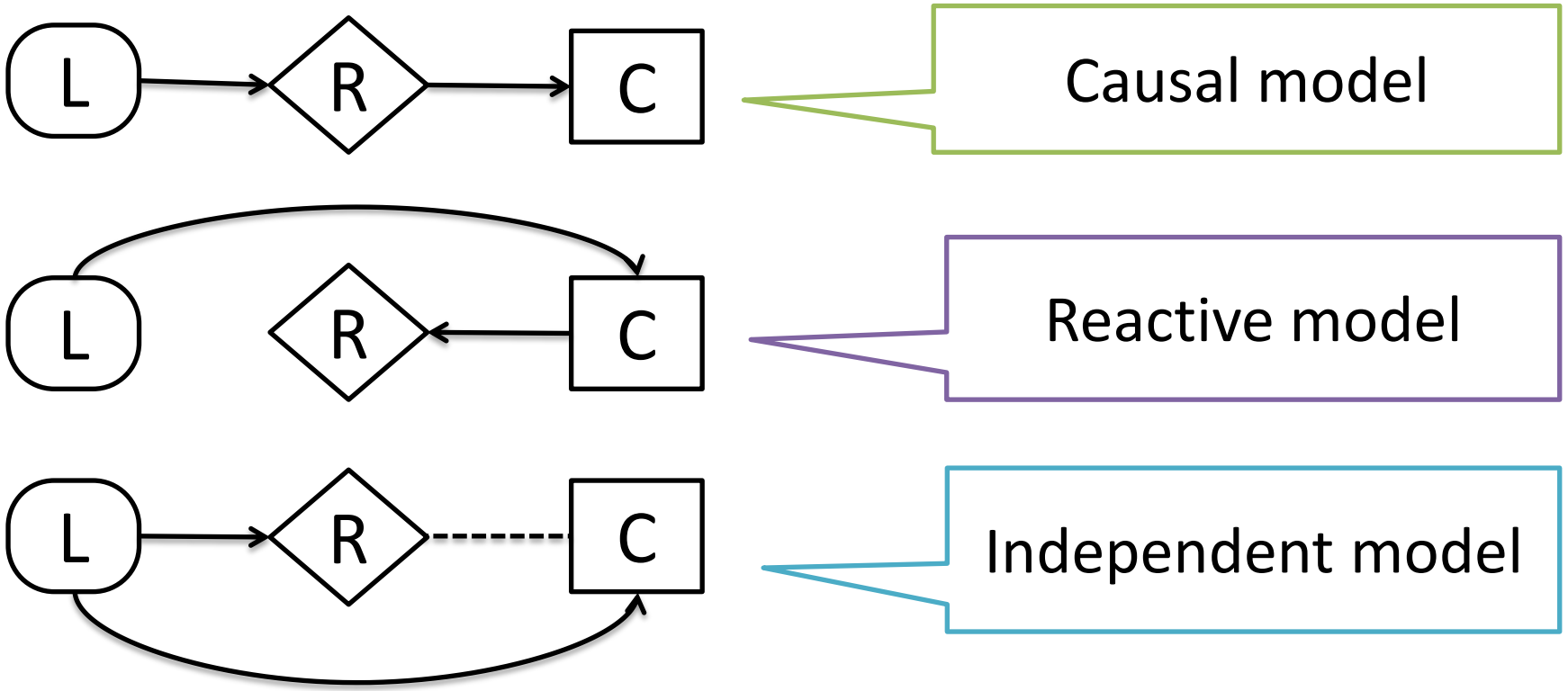
Success Stories

- Identify genes that cause a phenotype.
  - [Schadt et al., ***Nature Genetics***, 2005]
- Reconstruct causal pathways.
  - [K. Sachs, et al. ***Science*** , (2005)]
- Identify causal effects.
  - [Maathius et al., ***Nature Methods***, 2010]
- Predict association among variables never measured together.
  - [Tsamardinos et al., ***JMLR***, 2012]
- Select features that are most predictive of a target variable.
  - [Aliferis et. al., ***JMLR***, 2010]

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

L –Locus of DNA variation
R – gene expression
C- Phenotype (Omental Fat Pad Mass trait)

Biological knowledge: Nothing causally affects L



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

1. Identify loci susceptible for causing the 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

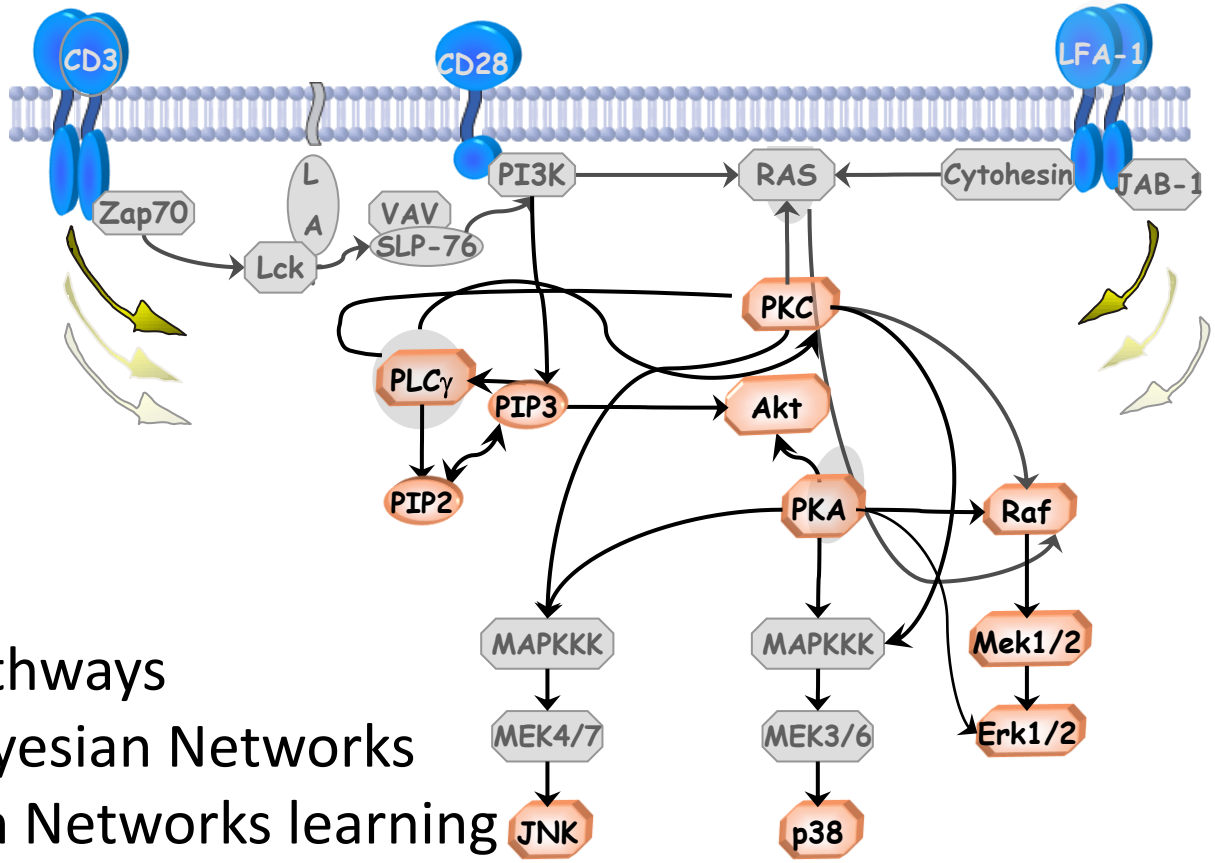
One of them ranked 152 out of the 440 based on mere correlation

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

1. Identify loci susceptible for causing the 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 of the top ranked genes where experimentally validated as causal

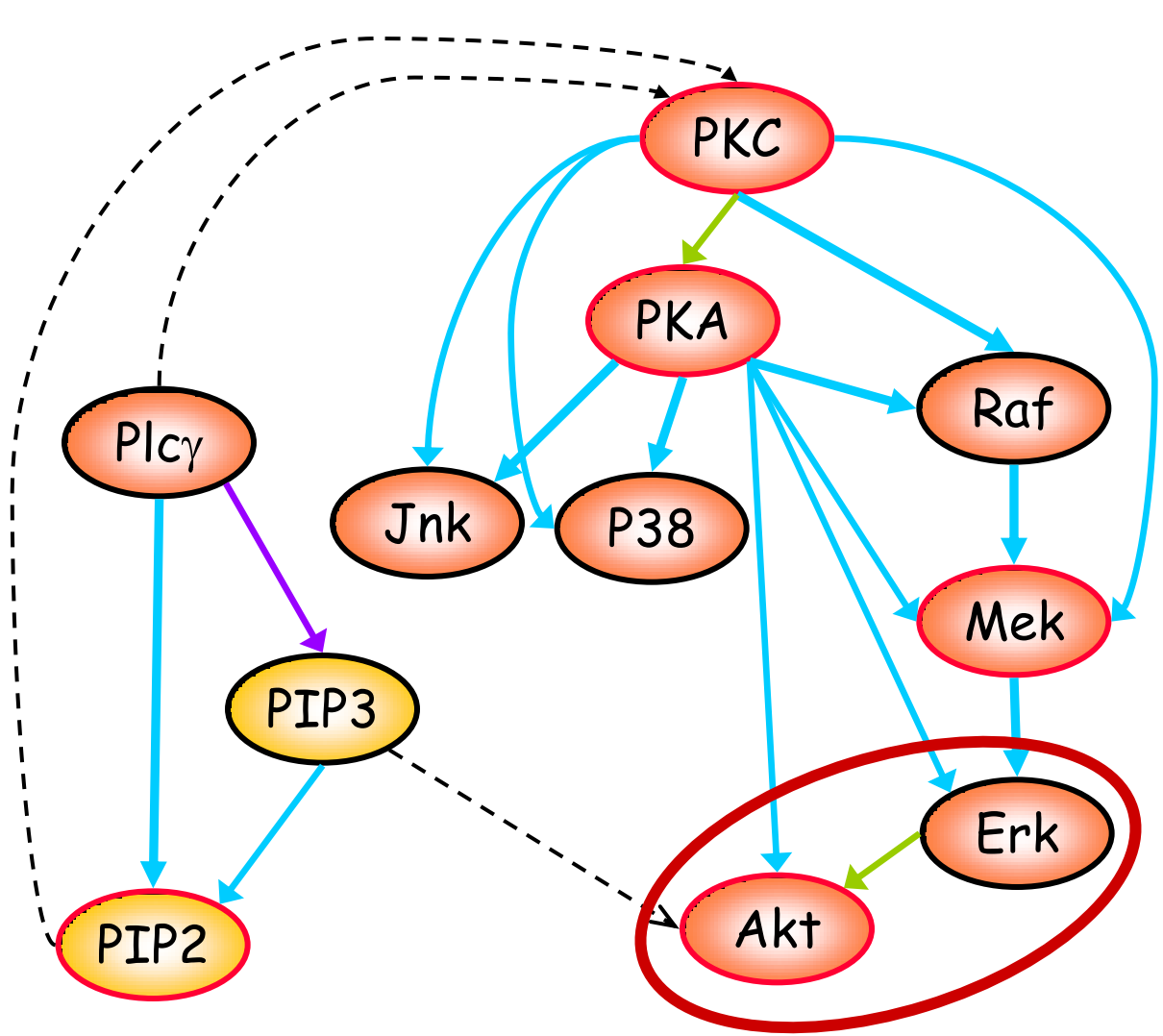
Causal Protein-Signaling Networks Derived from Multi-parameter Single-Cell Data



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

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

Reconstructed vs. Actual Network



Phospho-Proteins
Phospho-Lipids
Perturbed in data

Expected Pathway
Reported
Reversed
Missed

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

Predicting causal effects in large-scale systems from observational data

What will happen if you knock down Gene X?

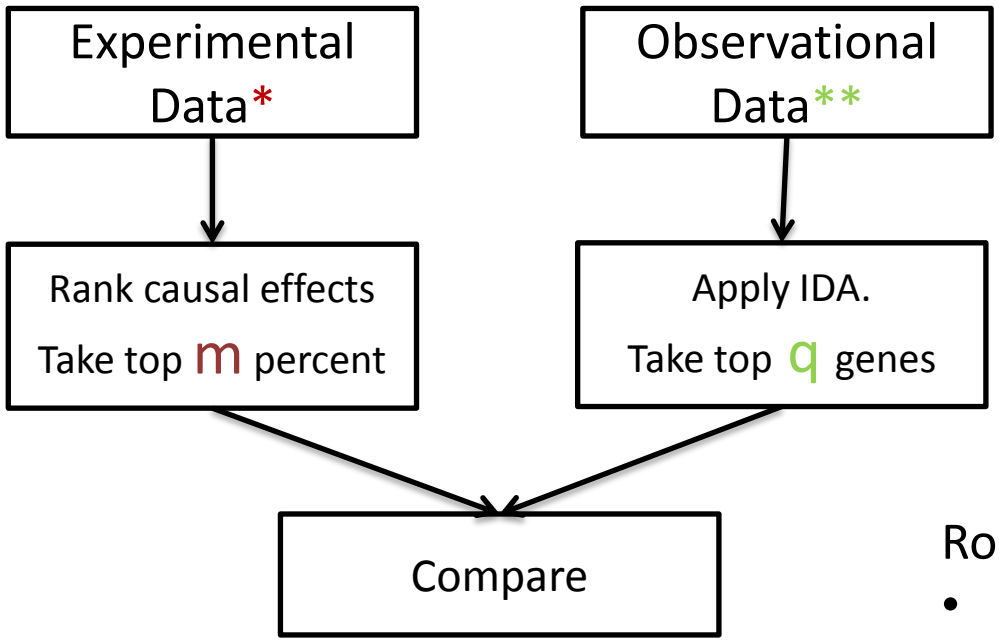
	Gene A	Gene B	…	Gene X
1	0.1	0.5		1.2
2	0.56	2.32		0.7
…				
n	7	0.4		2.4

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

1. For every DAG G faithful to the data
2. 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
3. Causal effect C of X on V is the minimum of all C_G

Predicting causal effects in large-scale systems from observational data

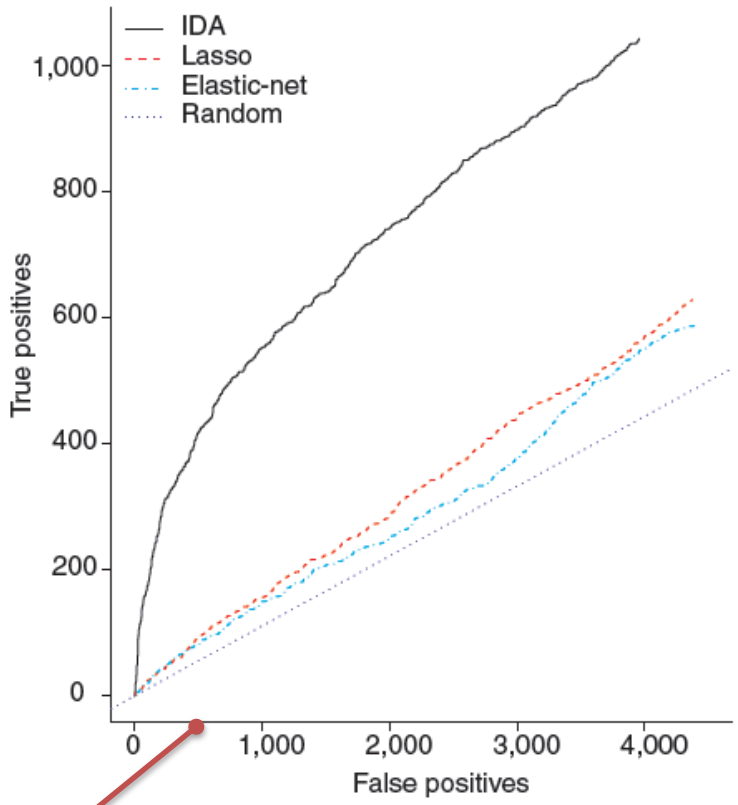
IDA Evaluation

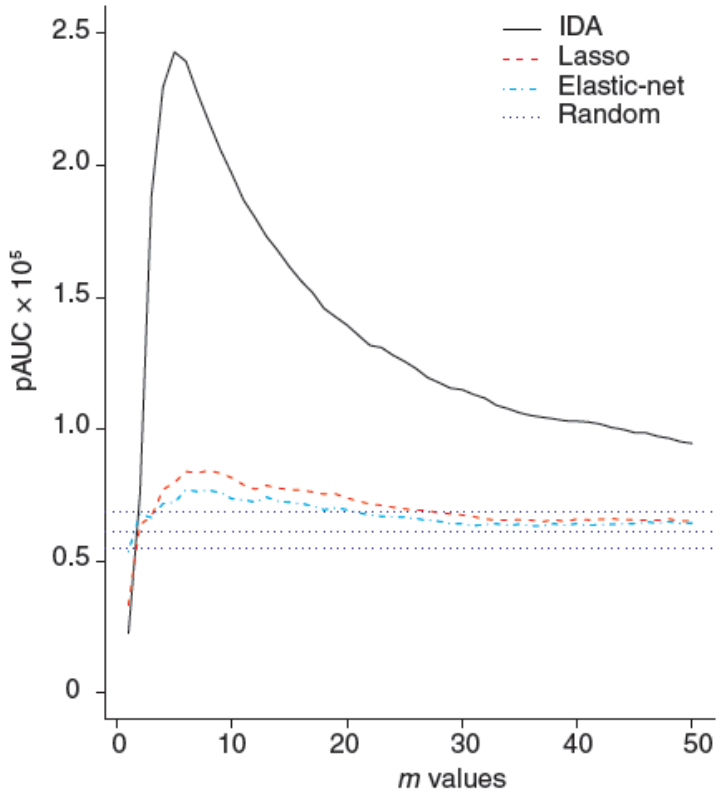


Rosetta Compendium data:

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

Predicting causal effects in large-scale systems from observational data



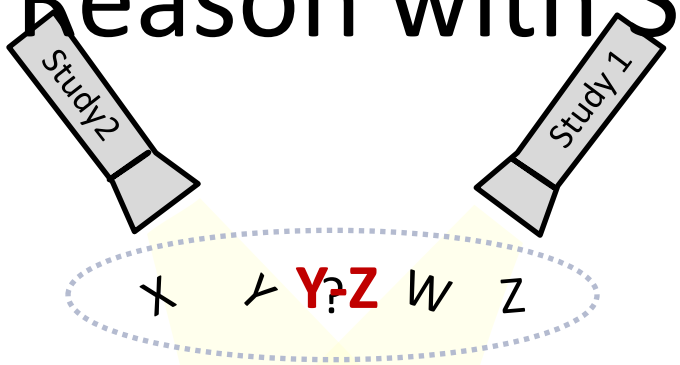


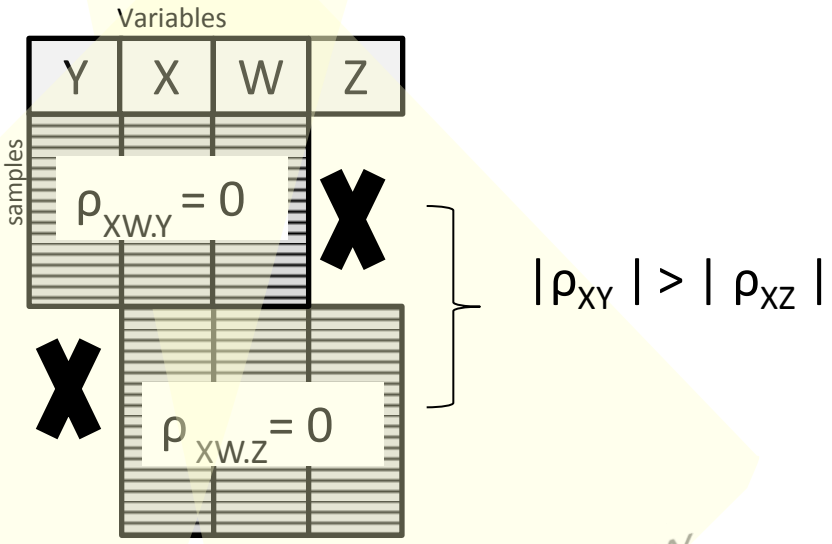
- m=10%

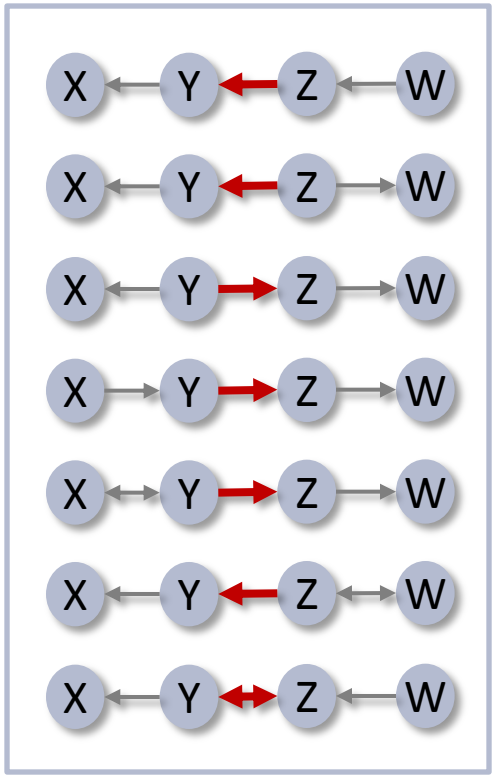
Integrative Causal Analysis

- Make inferences from multiple heterogeneous datasets
  - That measure quantities under different experimental conditions
  - Measure different (overlapping) sets of quantities
  - In the context of prior knowledge
- General Idea:
  - Find all CAUSAL models that simultaneously fit all datasets and are consistent with prior knowledge
  - Reason with the set of all such models

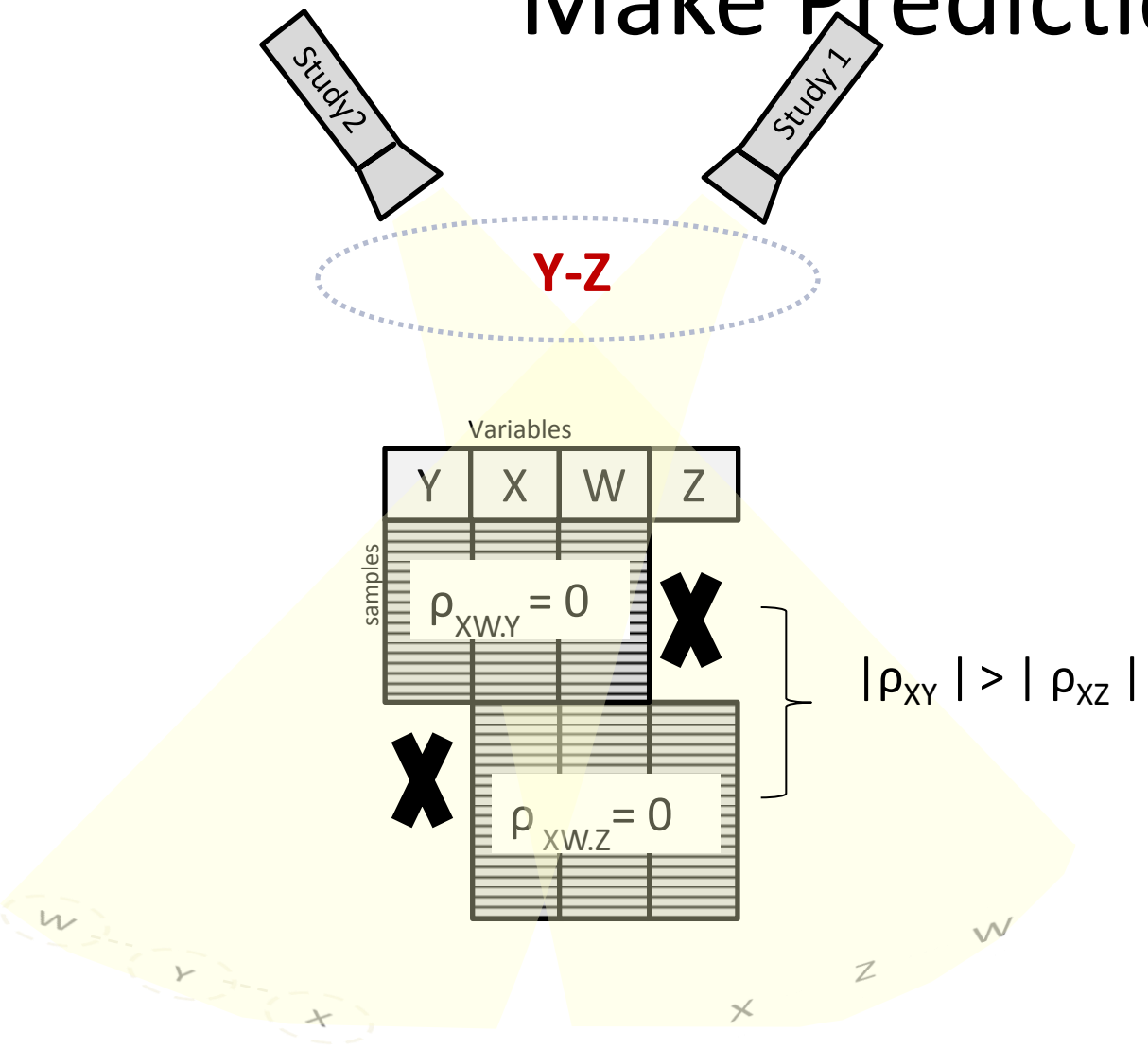
Reason with Set of Solutions

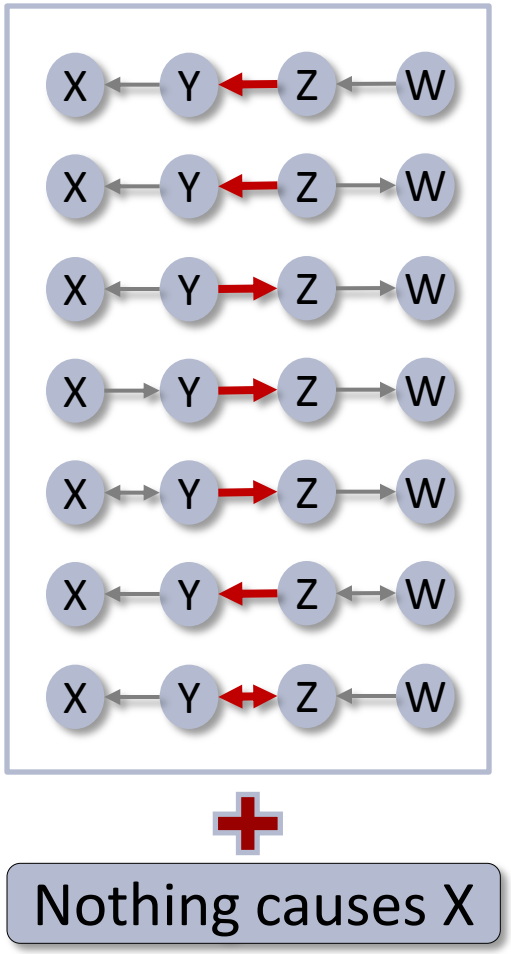






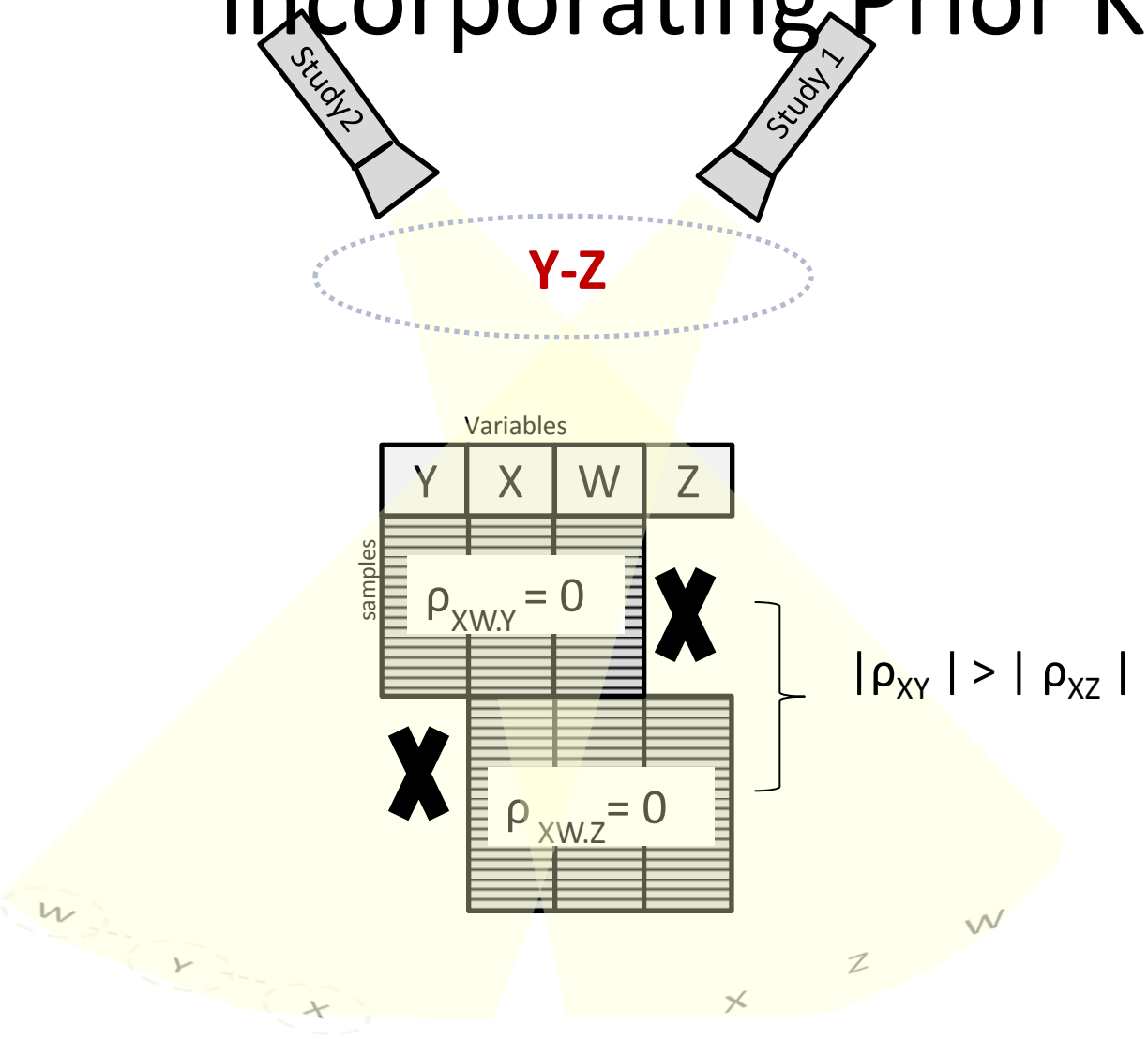
Make Predictions

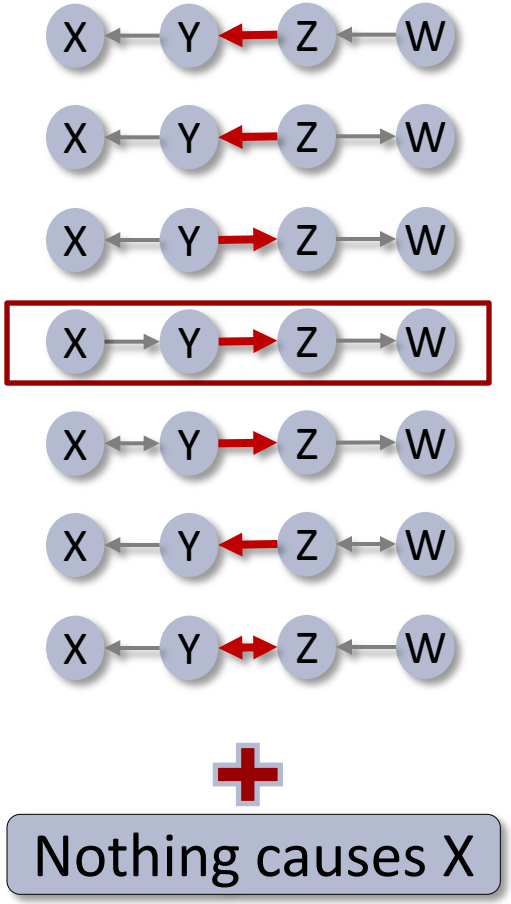




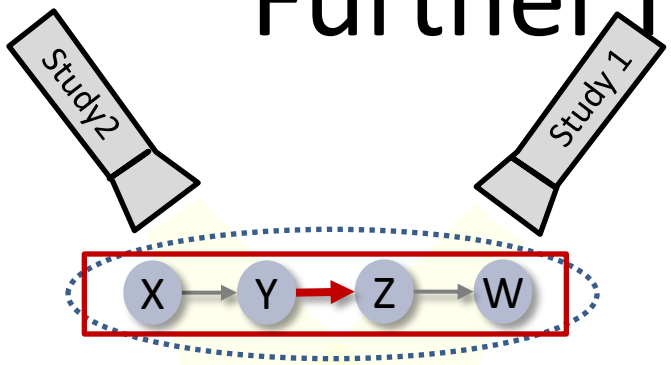
Nothing causes X

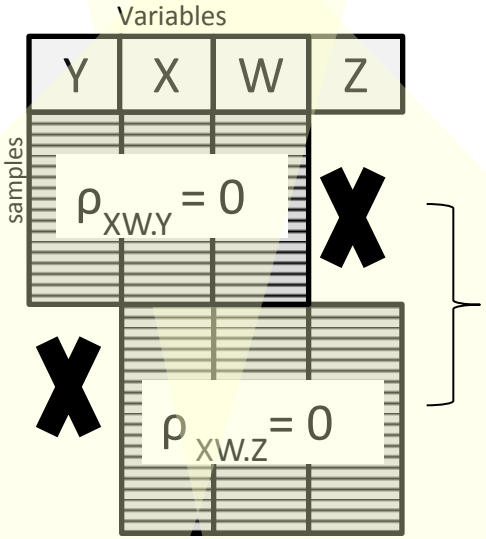
Incorporating Prior Knowledge



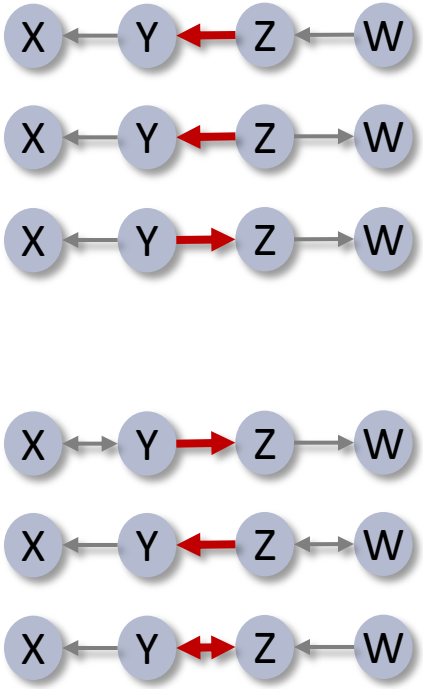


Further Inductions





$|\rho_{XY}| > |\rho_{XZ}|$



Nothing causes X

Changing ***Y*** will have an effect on ***Z***

Proof-of-concept Results

I Tsamardinos, S Triantafillou and V Lagani,
Towards Integrative Causal Analysis of Heterogeneous Datasets and Studies,
***Journal of Machine Learning Research*, to appear**

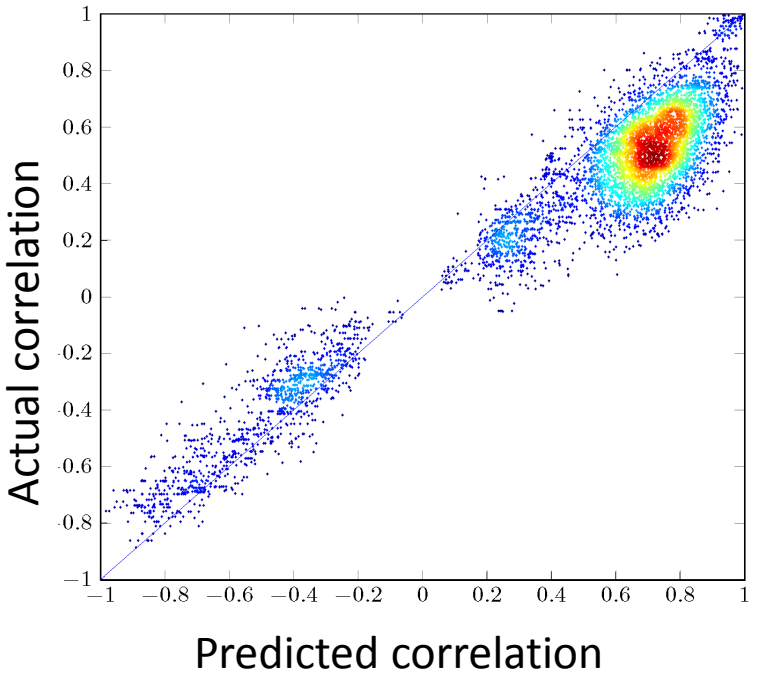
20 datasets

- **Biological, Financial, Text, Medical, Social**

698897 predictions

98% accuracy

- **vs. 16% for random guessing**



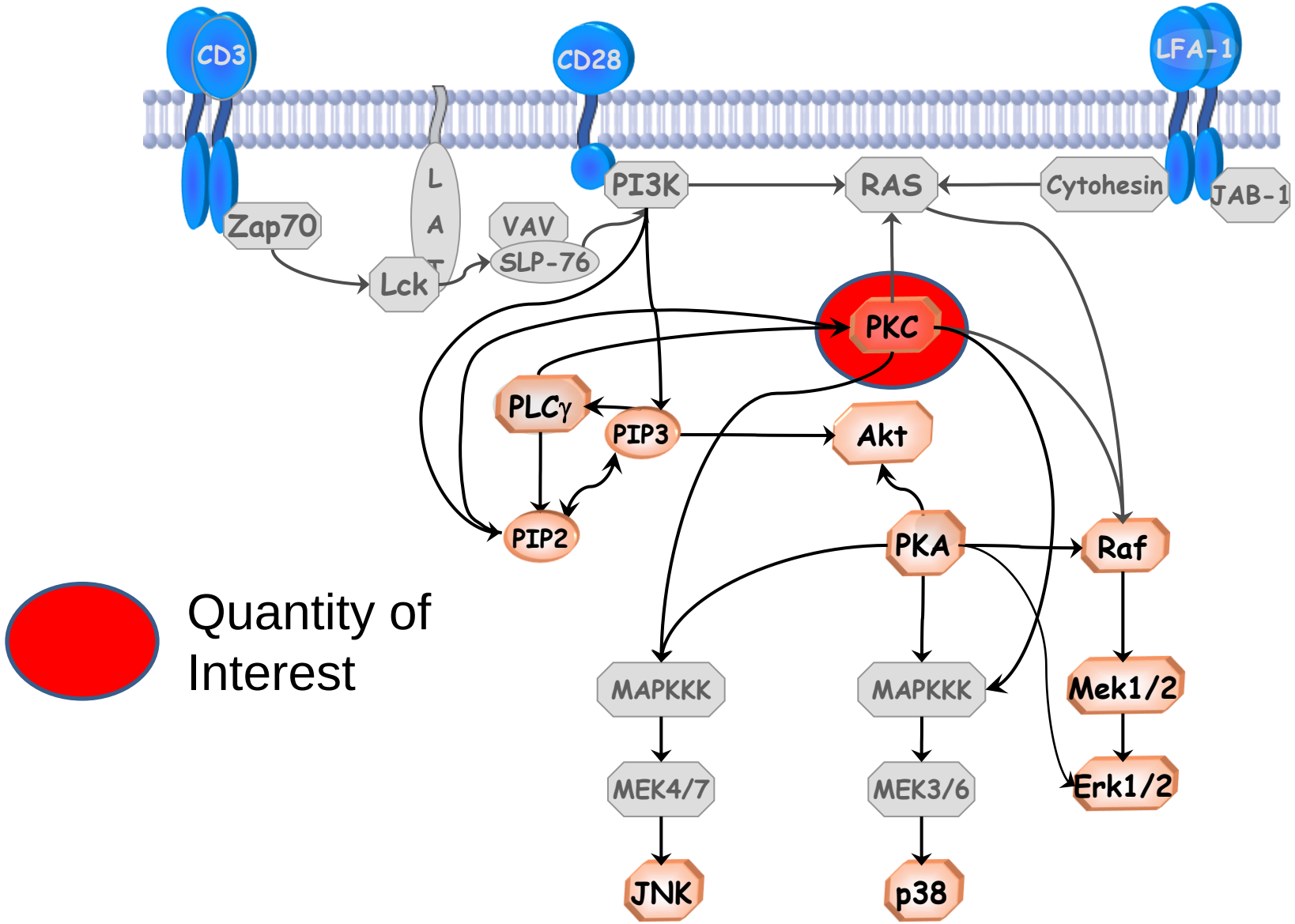
0.79 R^2 between predicted and sample correlation

Causality and Feature Selection

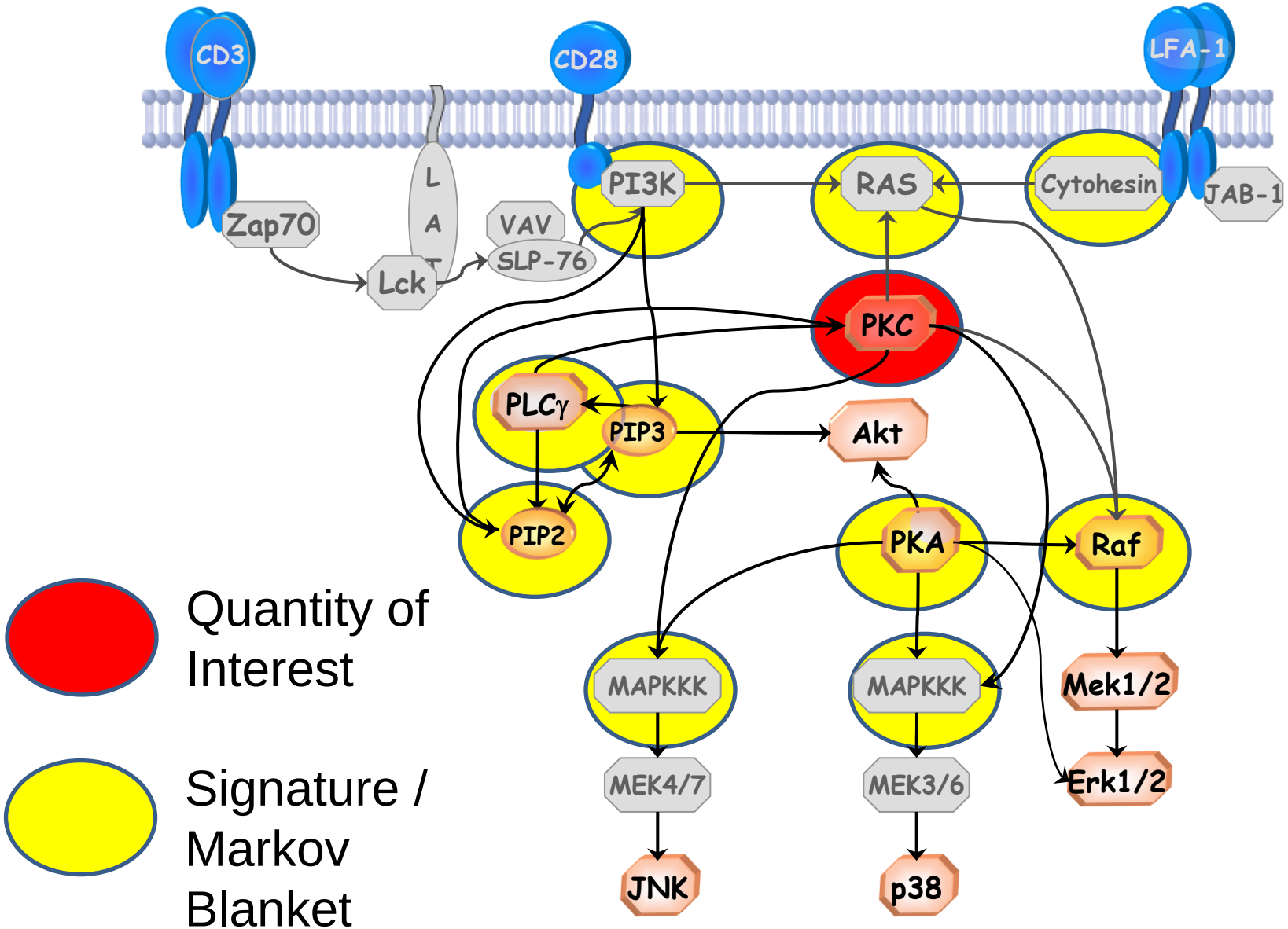
- Question: Find a minimal set of molecular quantities that collectively carries all the information for optimal prediction / diagnosis (target variable) (**Molecular Signature**)
  - **Minimal**: throw away irrelevant or superfluous features
  - **Collectively**: May need to consider interactions
  - **Optimal**: Requires constructing a classification / regression model and estimating its performance
- Answer*:It is the **direct causes**, the **direct effects**, and the **direct causes of the direct effects** of the target variable in the BN (called the Markov Blanket in this context)

* Adopting all causal assumptions

Markov Blanket



Markov Blanket



Markov Blanket Algorithms

- Efficient and accurate algorithms applicable to datasets with hundreds of thousands of variables
  - Max-Min Markov Blanket, [Tsamardinos, Aliferis, Statnikov, KDD 2003]
  - HITON [Aliferis, Tsamardinos, Statnikov, AMIA 2003]
  - [Aliferis, Statnikov, Tsamardinos, et. al. JMLR 2010]
- State-of-the-art in variable selection

Objective

- Identifying a set of transcripts able to predict IKAROS gene expression.
- The selected set should be:
  - Maximally informative: able to predict IKAROS expression with optimal accuracy
  - Minimal: containing no redundant or uninformative transcripts

Data

- Genome-wide transciptome data from HapMap individual of European descent [Montgomery et al., 2010]
  - Lymphoblast cells
  - 60 distinct individuals
  - Approximately. ~140K transcripts

- RKPM values freely available from ArrayExpress
  - www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-197

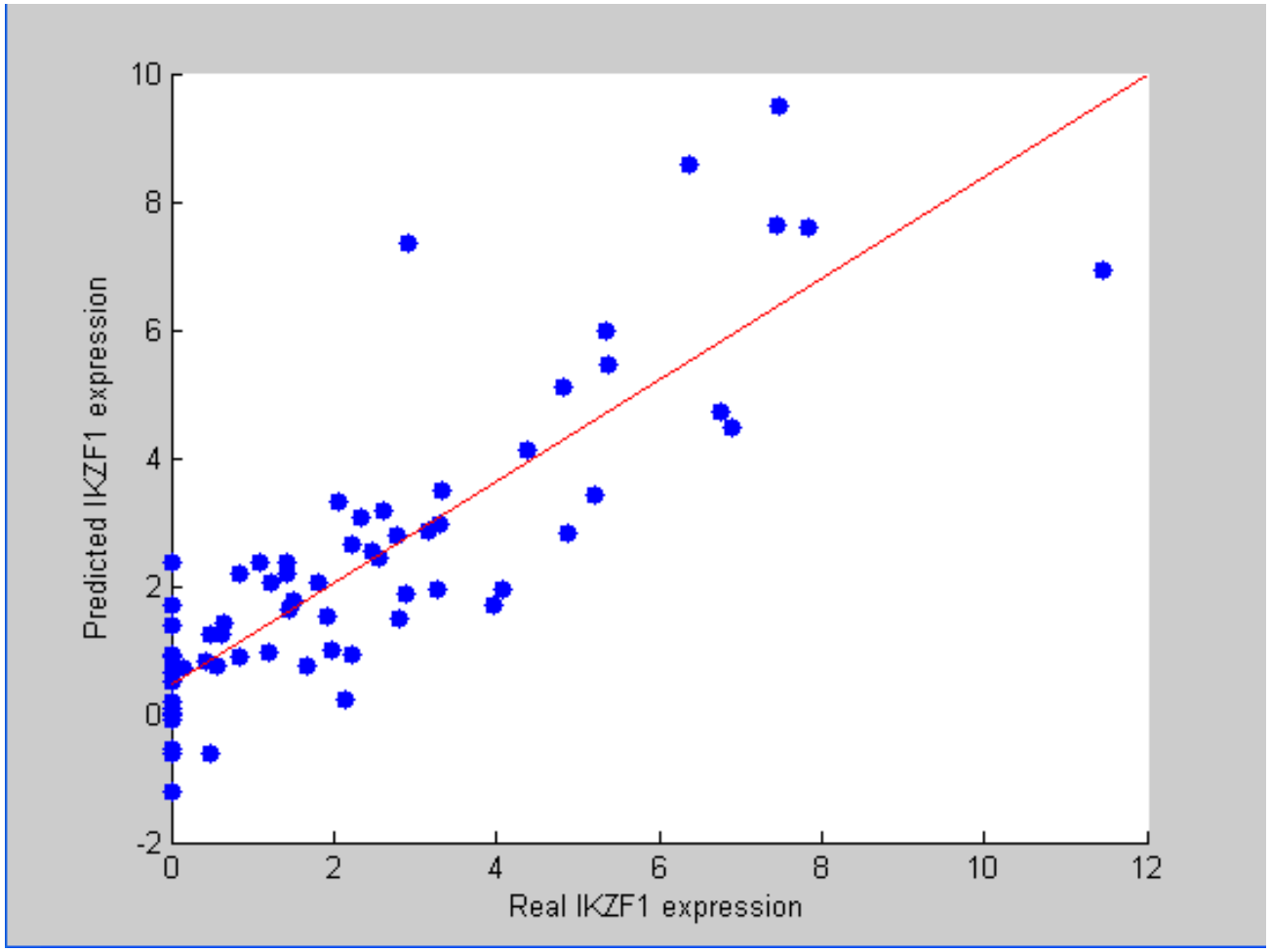
Methods

- Constraint-based, local learning feature selection method for identifying multiple signatures
  - [Tsamardinos, Lagani and Pappas, 2012]
- Support Vector Machine (SVM) for providing testable predictions
  - [Chang and Lin, 2011]
- Nested cross validation procedure for:
  - setting algorithms' parameters
  - providing unbiased performance estimations
  - [Statnikov, Aliferis, Tsamardinos, et al., 2005]

Results

- 22 different signatures found to be equally maximally predictive
  - Mean Absolute Error: 1.93
  - R^2: 0.7159
  - Correlation of predictions and true expressions:
    - 0.8461 (p-value < 0.0001)
- Example signature:
  - ENST00000246549, ENST00000545189, ENST00000265495, ENST00000398483, ENST00000496570
- Corresponding to genes:
  - FFAR2, ZNF426, ELF2, MRPL48, DNMT3A

Predicted vs. Observed IKZF1 values



Beyond This Tutorial

Textbooks:

- Pearl, J. *Causality: models, reasoning and inference* (Cambridge University Press: 2000).
- Spirtes, P., Glymour, C. & Scheines, R. *Causation, Prediction, and Search*. (The MIT Press: 2001).
- Neapolitan, R. *Learning Bayesian Networks*. (Prentice Hall: 2003).

Beyond This Tutorial

Different principles for discovering causality

- Shimizu, S., Hoyer, P.O., Hyvärinen, A. & Kerminen, A. A Linear Non-Gaussian Acyclic Model for Causal Discovery. *Journal of Machine Learning Research* **7**, 2003-2030 (2006).

Hoyer, P., Janzing, D., Mooij, J., Peters, J. & Schölkopf, B. Nonlinear causal discovery with additive noise models. *Neural Information Processing Systems (NIPS)* **21**, 689-696 (2009).

Causality with Feedback cycles

- Hyttinen, A., Eberhardt, F., Hoyer, P.O., Learning Linear Cyclic Causal Models with Latent Variables. *Journal of Machine Learning Research*, 13(Nov):3387-3439, 2012.

Causality with Latent Variables

- Richardson, T. & Spirtes, P. Ancestral Graph Markov Models. *The Annals of Statistics* **30**, 962-1030 (2002).
- Leray, P., Meganck, S., Maes, S. & Manderick, B. Causal graphical models with latent variables: Learning and inference. *Innovations in Bayesian Networks* **156**, 219-249 (2008).

Conclusions

- Causal Discovery is possible from observational data or by limited experiments
- Beware of violations assumptions and equivalences
- Causality provides a formal language for conceptualizing data analysis problems
- Necessary to predict the effect of interventions
- Deep connections to Feature Selection
- Allows integrative analysis in novel ways
- Advanced theory and algorithms exist for different sets of (less restrictive) assumptions
- Way to go still, particularly in disseminating to non-experts

References

- S. B. Montgomery et al. Transcriptome genetics using second generation sequencing in a Caucasian population. Nature 464, 773-777 (1 April 2010)
- I. Tsamardinos, V. Lagani and D. Pappas. Discovering multiple, equivalent biomarker signatures. In Proceeding of the 7th conference of the Hellenic Society for Computational Biology & Bioinformatics (HSCBB) 2012
- C. Chang and C. Lin. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011
- A. Statnikov, C.F. Aliferis, I. Tsamardinos, D. Hardin, and S.Levy. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics, 21:631-643, 2005.