Bayes Nets and the Automation of Discovery

Clark Glymour
Carnegie Mellon University
And
Institute for Human and Machine Cognition, Pensacola Florida

Some History

- **Yule, 1890s: Regression model of pauperism**
- **Spearman, 1904-1928: One factor linear latent variable models supported by patterns of quadratic constraints on correlations (tetrad differences)**
- **Wright (1929), Graphical path models of biological mechanisms**
- **Thurstone, 1934: Factor Analysis**
- **Hotelling, 1937: Principal components analysis**
- **Reichenbach, 1956: "Screening off"**
- **Blalock, 1959: Partial correlation graphical model selection**
- **Stroz & Wold (19??): Intervention interpretation for algebraic econometric models**
- **Suppes (197?) Conditional dependence criteria for causality**
- **Rubin (1980s) Counterfactual causal analysis for experimentation**
- **Speed (1980) Markov condition—application to causal social science models.**
- **Granger (1980s) "Causality" for time series**
- **Glymour, Spirtes, Scheines, Kelly, 1987: Algorithmic computation of tetrad constraints for automated model modification**

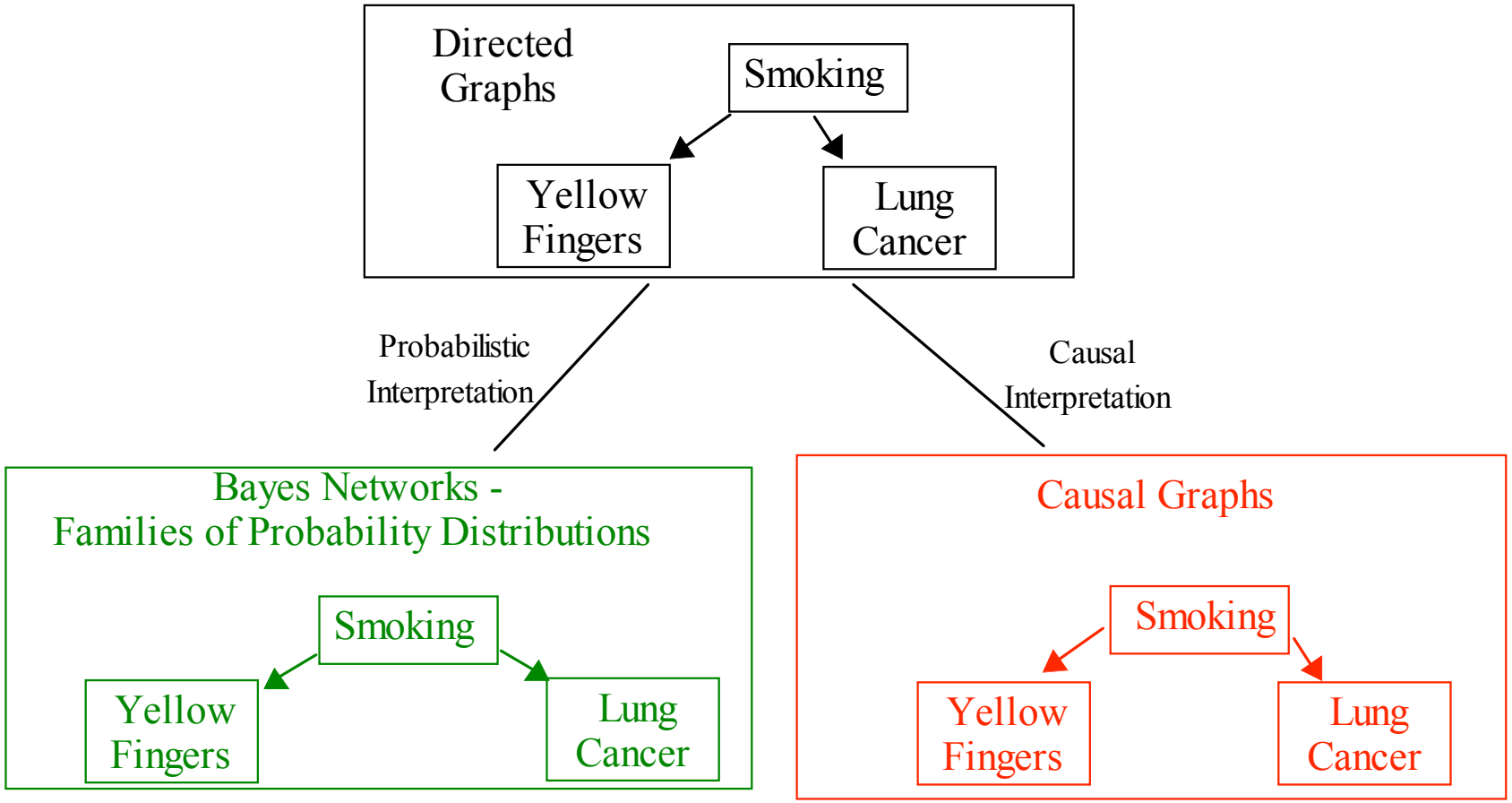
More History

- **Pearl, Geiger & Verma, 1988: D-separation algorithm (no causality); faithfulness assumption; Markov Blanket**
- **Forgotten Dane (1988): Decision procedure for Markov equivalence**
- **Spirtes, Glymour and Scheines, 1993: Intervention algorithms for graphical models with and without latent variables; faithfulness holds measure 1; consistent constraint based search with and without latent variables**
- **Meek, 1996; Chickering, 2002: Consistent Bayesian search without latent variables**
- **Spirtes, 1994: D-separation for linear cyclic models**
- **Richardson, 1996: D-separation equivalence for cyclic models; consistent constraint based search for linear feedback models without latent variables.**
- **Granger, 1998: Linear chain simultaneous causality for time series using tetrad constraints**
- **Bessler, Hoover, Monetta, Spirtes (2001-2004): general simultaneous causality for time series.**
- **Much, much more.**

Applications of Bayes Net Discovery Procedures

- *Pneumonia mortality prediction (Spirtes, et al.)*
- *Photosynthesis mechanisms (Shipley)*
- *Effects of Lead Exposure on IQ (Scheines)*
- *Causes of College Retention (Druzdzel & Glymour)*
- *Forecasting the Corn Export Market (Bessler)*
- *Influence on Biomass of Spartina Grass (Spirtes, et al)*
- *Causes of College Plans (Heckerman)*
- *Remote Satellite Instrument Calibration (Waldemark and Norqvist)*
- *Identification of Mineral Composition from Spectra(Ramsey, et al.)*
- *Climate Teleconnections (Chu, et al.)*
- *Zillions of studies of gene regulatio*
- *Psychology (*Gopnik, Tenenbaum, Danks, Schultz, etc*)*
- *Etc.*

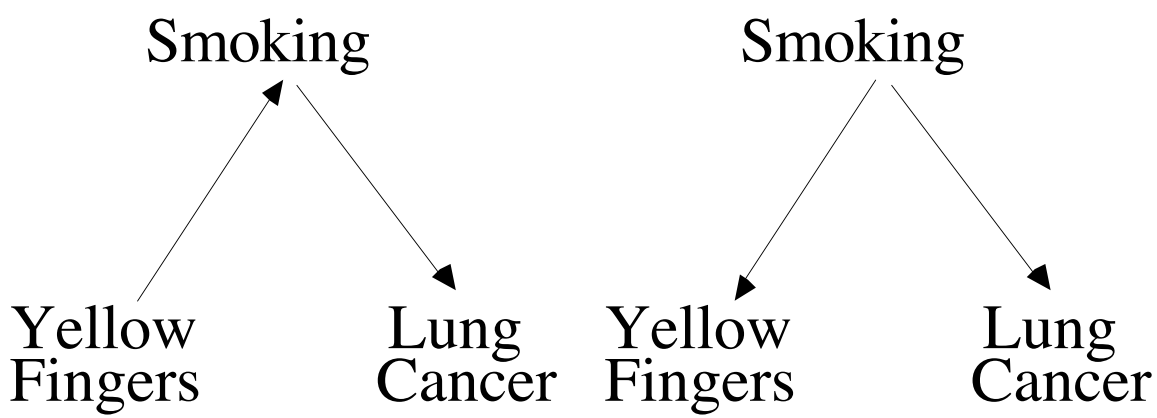
Directed Graphs



Graphs: Causal Interpretation

- *Each edge represents a* direct cause relative to the variables in the graph.

Graph 1 $\neq$ *Graph 2*

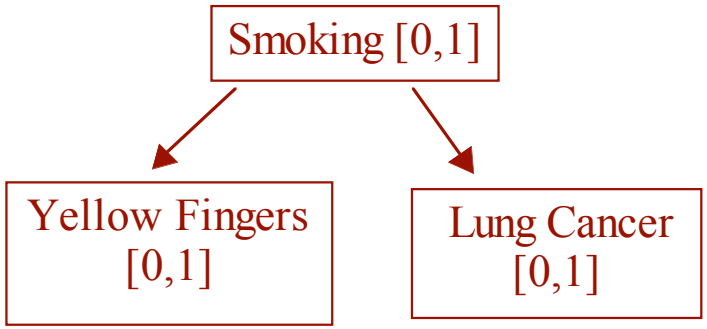


Graphs: Intervention Interpretation

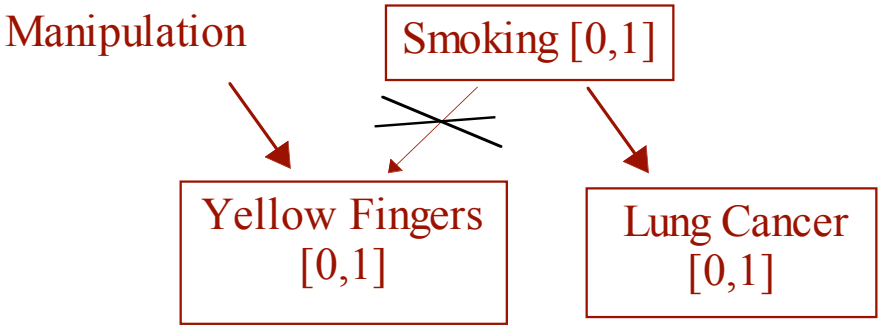
An edge X -> Y indicates that there are values for all other represented variables other than Y such that an intervention that fixes those values and varies X, varies Y.

Causal Graphs: Representing a Manipulation or Intervention

Observed Structure:



Structure upon Manipulating Yellow Fingers:

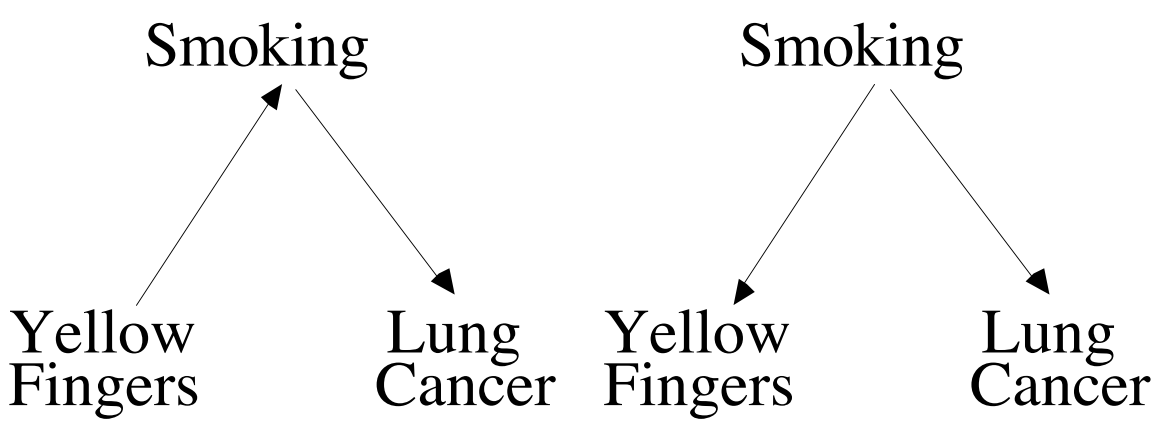


Graphs:
Probabilistic Interpretation

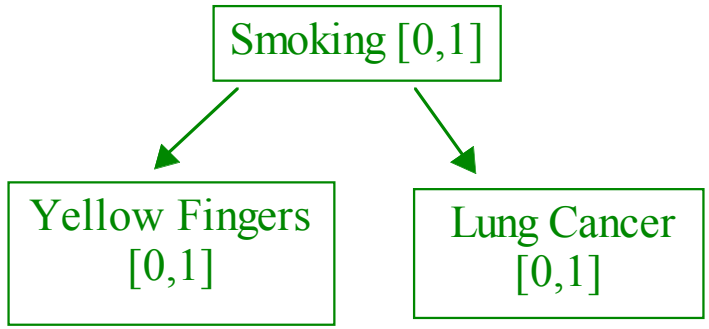
The graph represents a set *of probability distributions sharing the same conditional independence relations*

E.g., in these graphs, Yellow Fingers is independent of Lung Cancer given Smoking.

Graph 1 equivalent to Graph 2



Bayes Networks



The Joint Distribution Factors

According to the Graph,

i.e., for all X in **V**

$P(\mathbf{V}) = \Pi P(X|Parents(X))$

$P(S,Y,F) = P(S)\, P(YF \mid S)\, P(LC \mid S)$

P(S = 0) = .7

P(S = 1) = .3

P(YF = 0 | S = 0) = .99

P(YF = 1 | S = 0) = .01

P(YF = 0 | S = 1) = .20

P(YF = 1 | S = 1) = .80

P(LC = 0 | S = 0) = .95

P(LC = 1 | S = 0) = .05

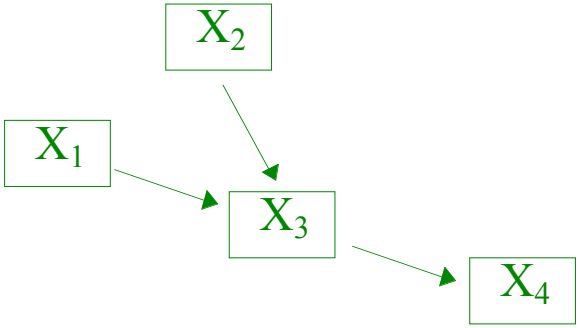
P(LC = 0 | S = 1) = .80

P(LC = 1 | S = 1) = .20

Bayes Networks

<u>Markov Condition</u>:

In a Bayes Network: each variable V is independent (in probability) of its non-descendants, conditional on its parents.

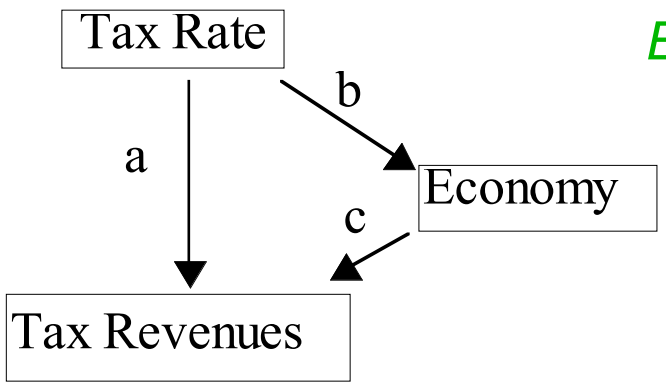


$$X_1 \perp\!\!\!\perp X_2$$

$$X_4 \perp\!\!\!\perp \{ X_1 \; X_2 \} \mid X_3$$

Faithfulness Assumption

Statistical Constraints arise from
Structure, not Coincidence



$$Revenues = aRate + cEconomy + \varepsilon_{Rev.}$$

$$Economy = bRate + \varepsilon_{Econ.}$$

$$a \neq -bc$$

Markov Equivalence Classes

- E.g.

$$X \rightarrow Y \rightarrow Z$$

$$X \leftarrow Y \rightarrow Z$$

$$X \leftarrow Y \leftarrow Z$$

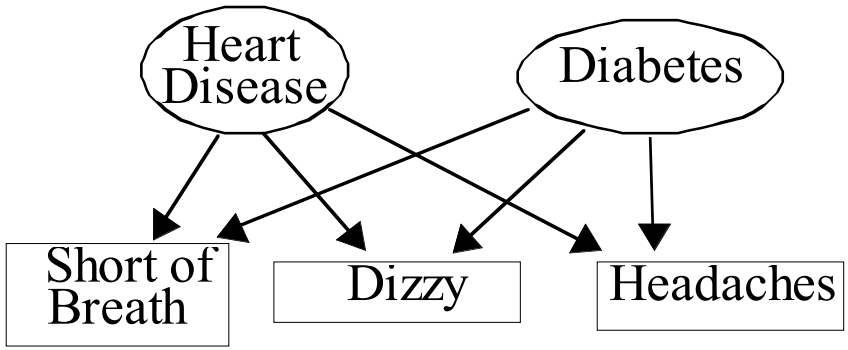
$$X \rightarrow Y \leftarrow Z$$

Why These Assumptions?

- The Markov condition can be derived from a weaker assumption used almost everywhere.
- Inferences to causes (and their absence) in conventional randomized experiments requires special cases of these assumptions.
- Assuming Markov, it has been proved that faithfulness almost always holds.

Bayes Networks: Original Uses

- Updating
  - Classifying
  - Diagnosis



  - E.g., calculate probability of Diabetes from symptoms

Bayes Nets as Causal Models

- Updating or computing conditional probability of possible causes represented in a Bayes net, given observation of an effect is already a simple form of causal inference.
- But we want to discover graphical causal models that correctly predict the effects of interventions on variables.

Discovery

From Experiment: super simplified.

- Wiggle X, if Y wiggles, infer X causes Y

  Invokes the Markov Assumption

- Wiggle X, if Y does not wiggle, infer that X does not cause Y

  Invokes the Faithfulness Assumption

From Observation: super simplified

- Observe that X and Y covary, X precedes Y, infer that X causes Y.

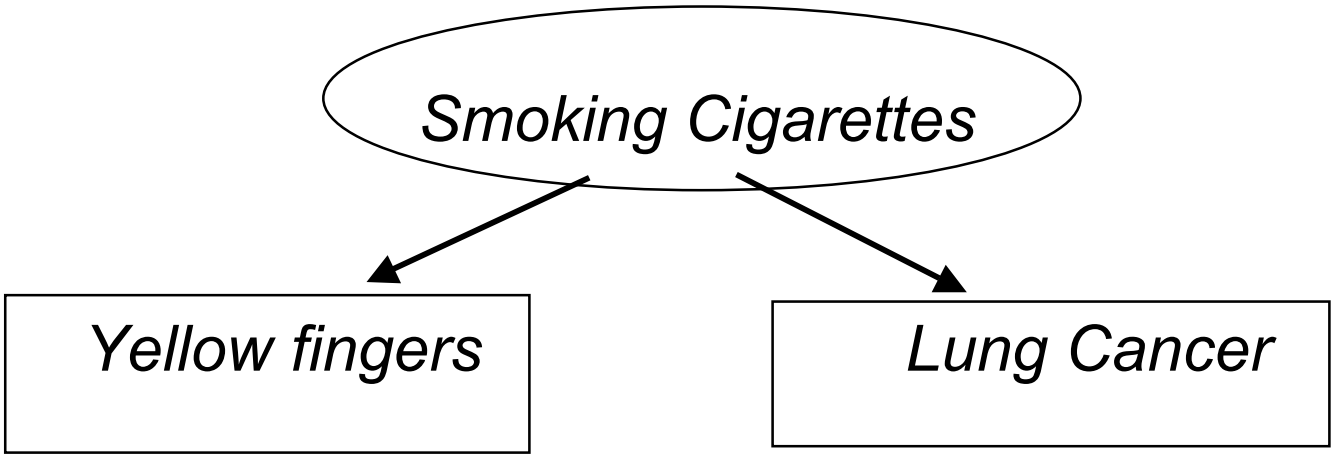
  OOPS!

Search

- Intervention relations don’t track correlation relations in any simple way.
- We can’t infer causal relations from correlations of two variables, but
- We can get causal information from associations of several variables
- Rather than trying to learn everything, we aim to do what proves to be possible: learn features common to all models that will explain the data, consistent with Markov and Faithfulness.

Why Association Isn't Causation

- An unobserved common cause produces the association, or part of it:



Why Association Isn't Causation

- The sample is a mixture of two populations with different probabilities for the associated features.

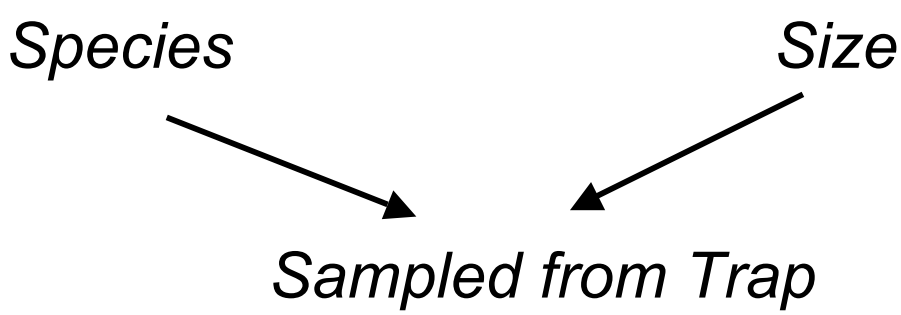
Males	*Females*
Tall, blond	*Short, brunette*

Males + Females

Height ——— *Hair color*

Why Association Isn't Causation

- The Values of the Associated Variables Influence whether a Unit is Sampled
  - Large mesh fish trap with bait only one species likes



What to Do?

- Fake it
  - Build a model and ignore alternatives (standard social science procedure)
  - Don’t distinguish between association and causation (e.g., talk about “risk factors”)
  - Use an inference procedure that is tried but not true (e.g., regression, CART, etc.)

What to Do?

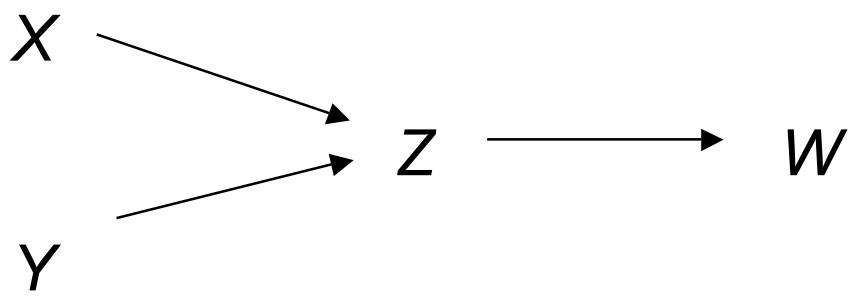
- Make plausible assumptions about the relation between causation and associations
- Characterize the causal information that can be obtained from associations under these assumptions
- Find algorithms that extract the information as efficiently as possible
- Investigate the reliability of the algorithms
- Repeat, with weaker assumptions

How?

- Represent causal structure by directed graphs
- Impose assumptions relating graphs and probability distributions
- Develop search algorithms
  - Based on patterns in the data
  - Based on model scores determined from the data
  - Mixed

Causal Information in Associations

- The members of the Observed Markov Equivalence class of a causal structure may all share some structure. For example:



Every graph in the OME class of this graph contains the path $Z \longrightarrow W$

Computational and Statistical Difficulties

1. If X -> Y in a DAG, then X, Y should be independent conditional on any set of other variables in the graph, but…

We cannot test for conditional independence of X, Y on all subsets of other variables—two many subsets

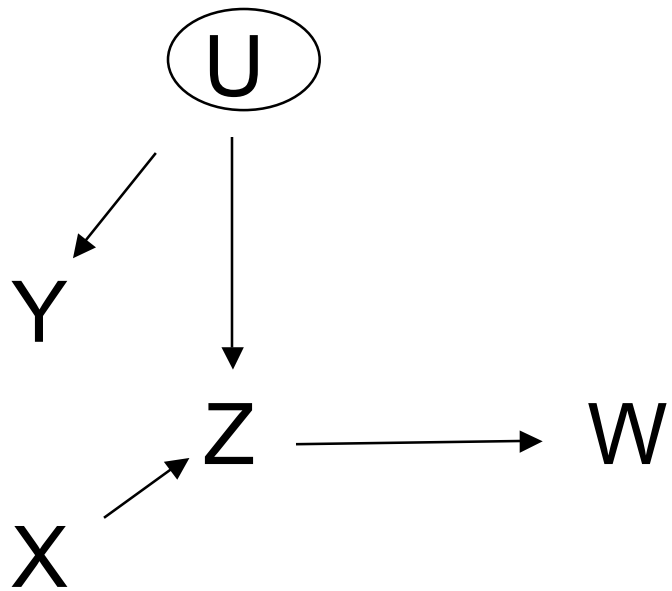
Statistical decisions lose power the more variables are conditioned on.

Consistent search procedures

- PC—tests for conditional independence--consistent when there are no latent confounders—finds some latent confounders.
- FCI—tests for conditional independence consistent when there are latent confounders
- Meek/Chickering—Bayesian search consistent when there are no latent confounders—always bad when there are.
- Washdown/MIMbuild: Clusters measured effects of latent variables and finds latent structure in linear models
- Richardson: Constraint based search for linear feedback models, consistent when there are no latent confounders
- Mixed Ancestral Graph Search: Bayesian search, consistent when there are latent variables, so far only for linear models.
- Danks/Glymour: Consistent constraint based search from databases with no common cases, distinct but not disjoint variable sets, and latent confounders.

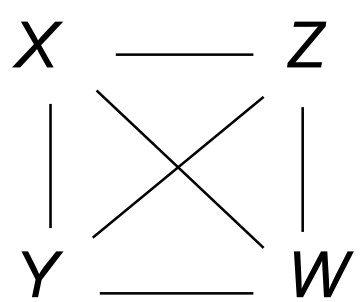
PC Search Illustrated

- True Structure



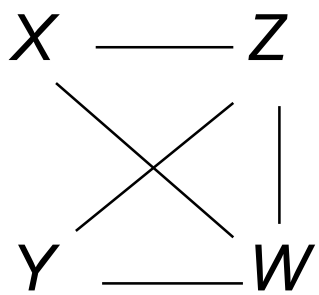
PC Search Illustrated

- Given data for X,Y,Z, W, form the complete undirected graph:



For each pair of variables, test for their independence; remove edges between any independent pair

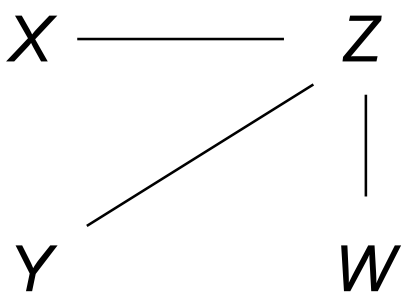
PC Search Illustrated



For each adjacent pair, e.g., X, W, and each third variable adjacent to at least one of them, e.g., Z, test for independence of the pair conditional on the third, e.g. $X \perp\!\!\!\perp W \mid Z$.

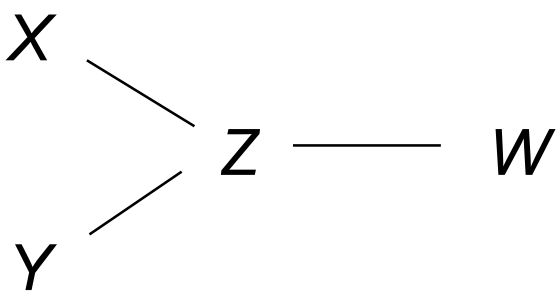
Remove edges between conditionally independent pairs.

PC Search Illustrated



Repeat for each each pair of adjacent variables conditioning on each two variables adjacent to at least one of them; conditioning on each 3 variables, etc., until no further edges are removed.

PC Search Illustrated

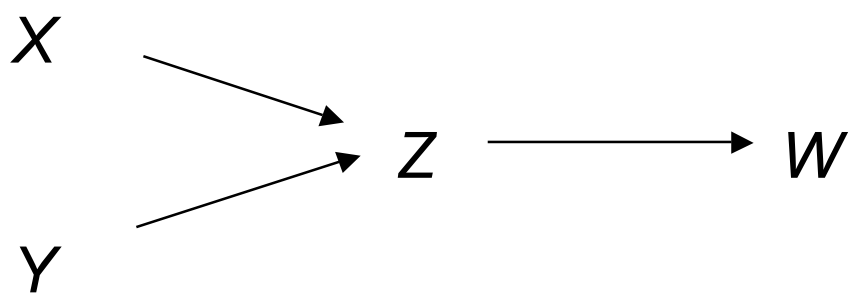


If Z was not conditioned on in removing X - Y edge, orient X - Z and Y- Z into Z (colliders) and mark the X end of X - Z and the Y end of Y - Z.

If Z is a collider, orient Z - W away from Z.

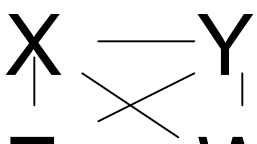
PC Search Illustrated

- Final Result is the structure common to all graphs in the OME class of the true structure:



What about Unobserved Confounders?

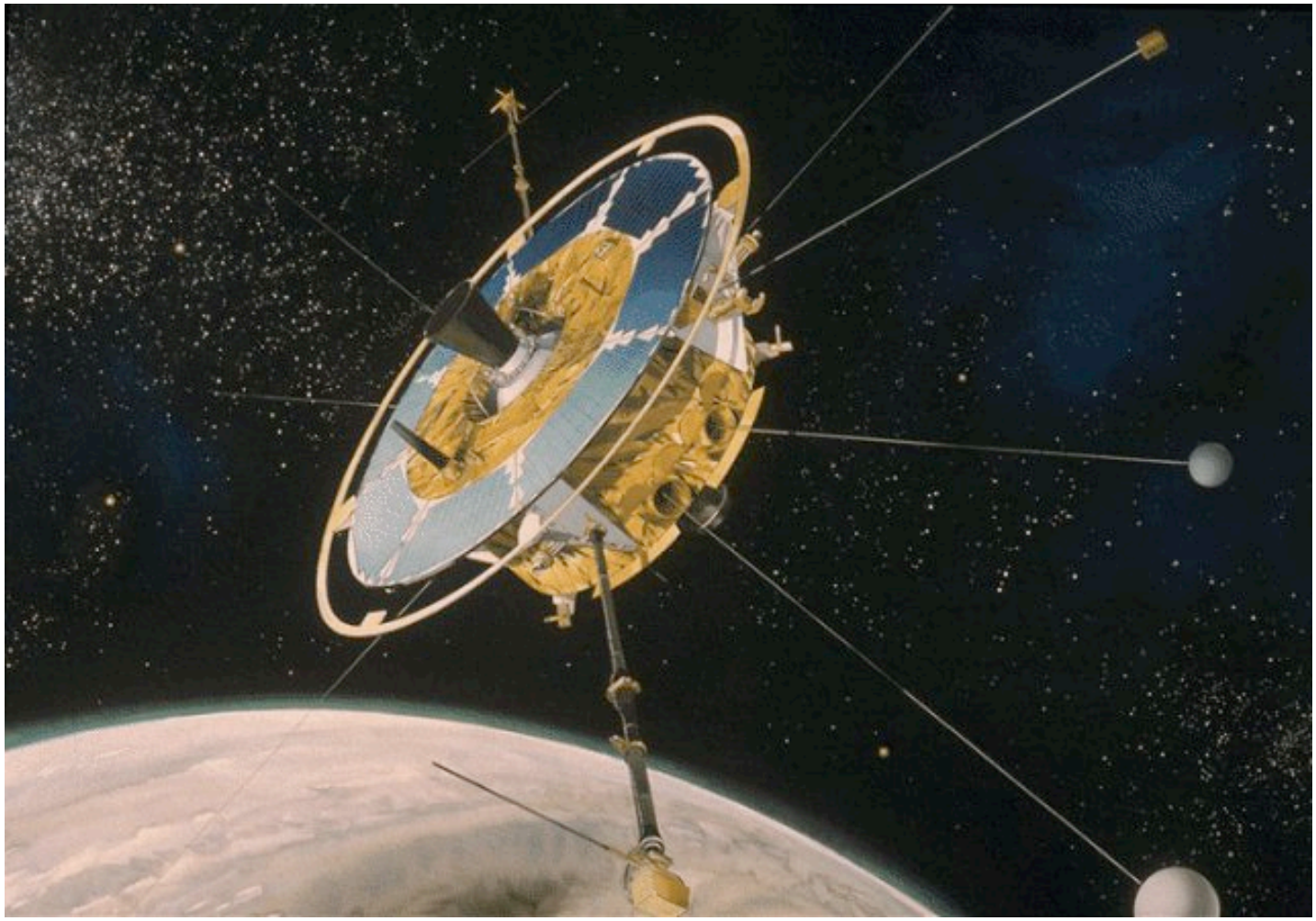
$X \rightarrow Y \leftarrow U \rightarrow Z \leftarrow W$



X ind of Z, W; W ind of X, Y

X—>Y<— >Z<— W Y not conditioned on in removing X, Z edge; Z not condtioned on in removing Y, W edge, hence

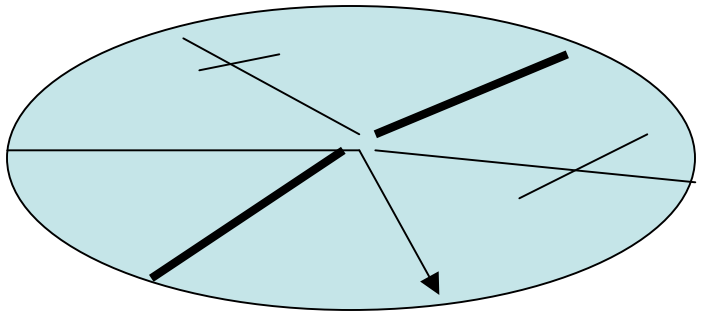
The Swedish Freja Satellite



Miscalibrations

- Miscalibrated mass spectrometer designed to detect ion bursts.
- Bayes net procedures used to identify the errors
- Recalibration reduces estimated errors by half.

Light ions



Heavy ions

Spartina in the Cape Fear Estuary

What FactorsDirectly Influence Spartina Growth in the Cape Fear Estuary?

pH, salinity, sodium, phosphorus, magnesium, ammonia, zinc, potassium…, what?

14 variables for 45 samples of Spartina from Cape Fear Estuary.

Biologist concluded salinity must be a factor.

Bayes net analysis says only pH directly affects Spartina biomass

Biologist's subsequent greenhouse experiment says: if pH is controlled for, variations in salinity do not affect growth; but if salinity is controlled for, variations in pH do affect growth.

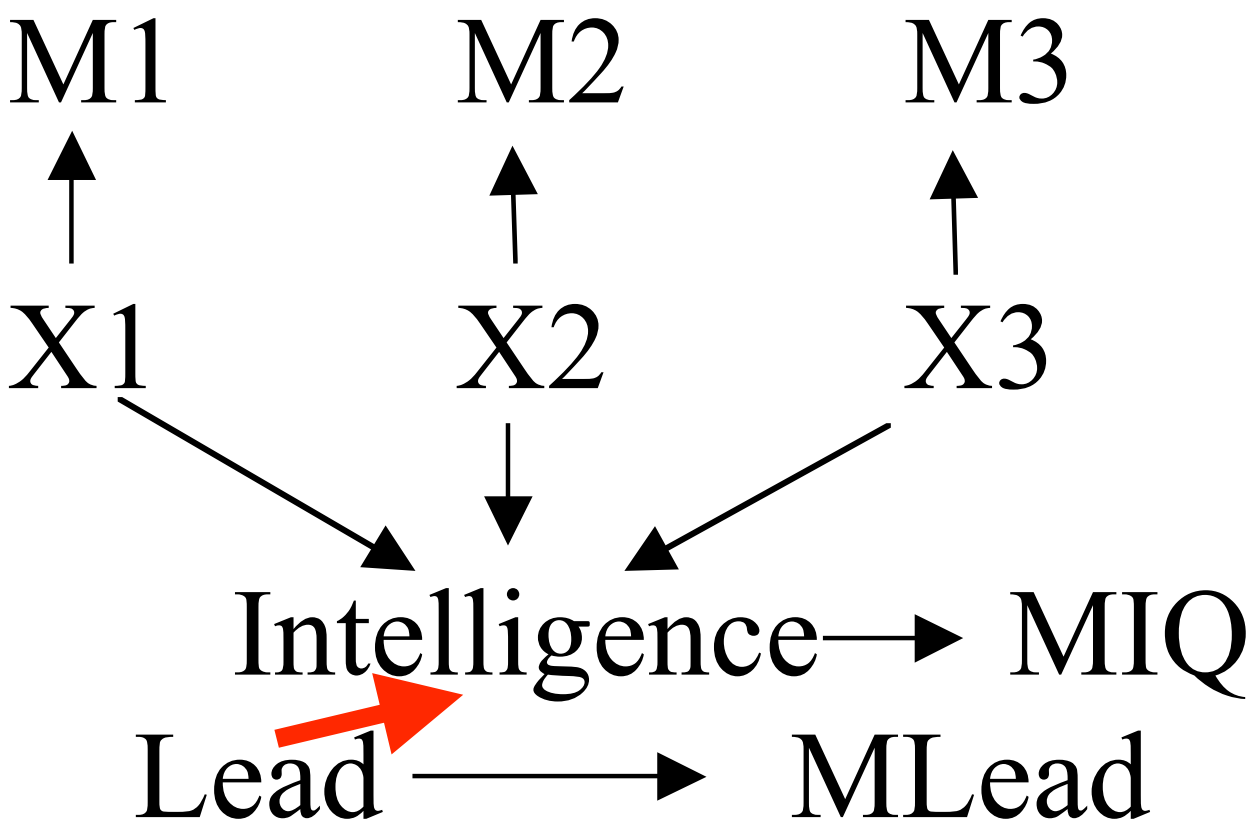
The Influence of Lead Exposure on Children's IQ

- Needleman: ANOVA, many variables, small negative effect of lead exposure
- NIH statisticians: Needleman must redo with stepwise regression
- Needleman: Stepwise regression, 6 significant regressors, small (but bigger) effect of lead exposure
- CMU econometricians: Measurement error—influence of lead exposure cannot be bounded away from zero.

Influence of Low Level Lead Exposure on Children's IQ

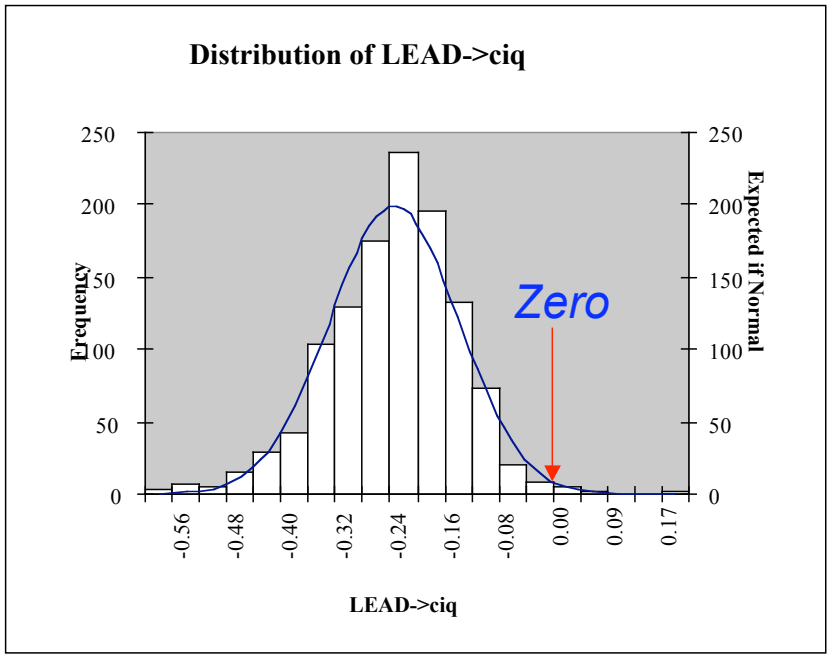
- Scheines: TETRAD program; 3 of the 5 covariates have NO correlation with IQ.
- Scheines: CMU econometricians must give their prior distribution for measurement error.
- Scheines: Bayesian estimation (with Gibbs' sampling for posterior distribution) of effect of low level lead exposure on IQ using;
  - TETRAD selected variables
  - Econometricians' measurement error model
  - Econometricians' priors

The CMU Economists' Model with TETRAD Covariates



Result: Lead is Twice as Bad as Needleman Thought

Marginal Posterior

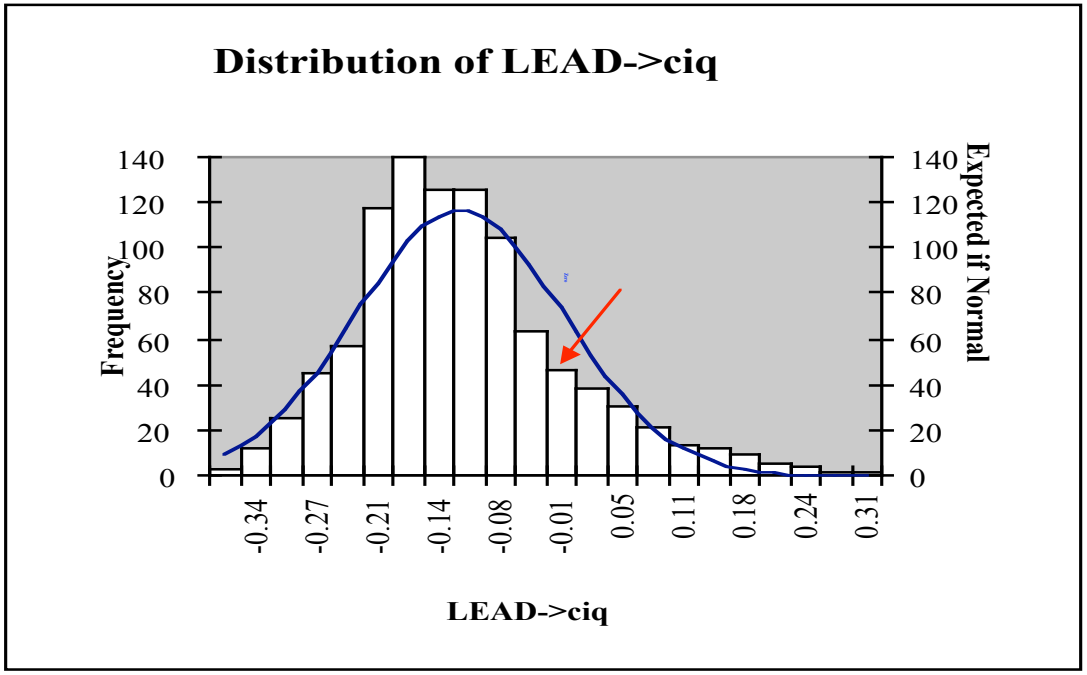


Robust over similar priors

And the Elimination of Causally Irrelevant Variables Is Critical

Using Needleman's Covariates

With similar prior, the marginal posterior:



Very Sensitive to Prior Over Regressors

TETRAD eliminated

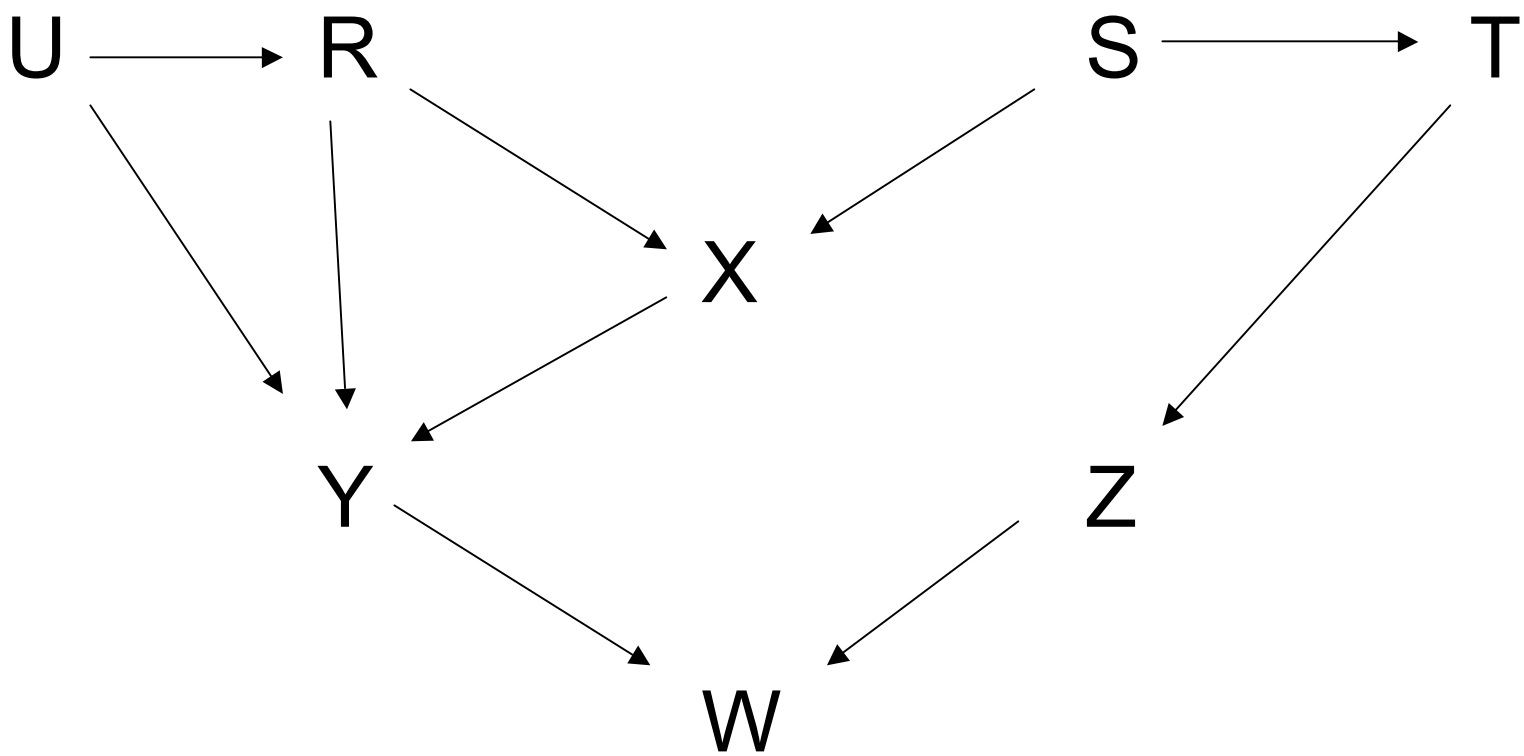
Bayes Nets In Classification

- Bayes nets provide a kind of solution to problems of relevant evidence.
- The Markov Blanket of a variable
  - MB of X in a graph is the parents of X; the children of X and the parents of children of X
  - If the joint distribution is Faithful to a DAG, all variables in the DAG are independent of X conditional on the values of the variables in the Markov Blanket for X.

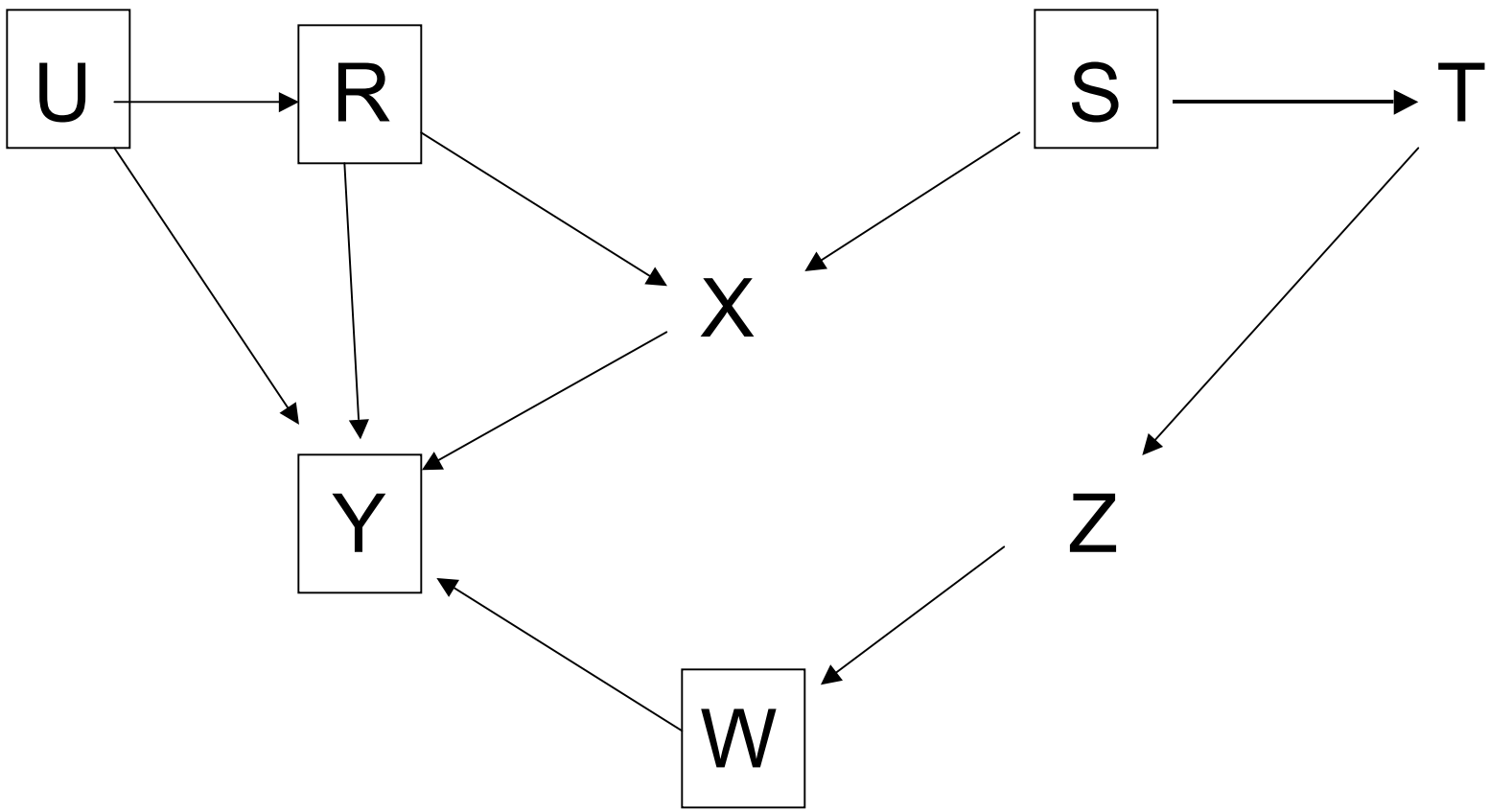
Markov Blanket is the “Ultimate Classifier”

- Everything else in the Bayes Net is independent of X conditional on the variables in the Markov Blanket of X

Causality and Relevant Evidence: The Markov Blanket of Variable X



Markov Blanket of X

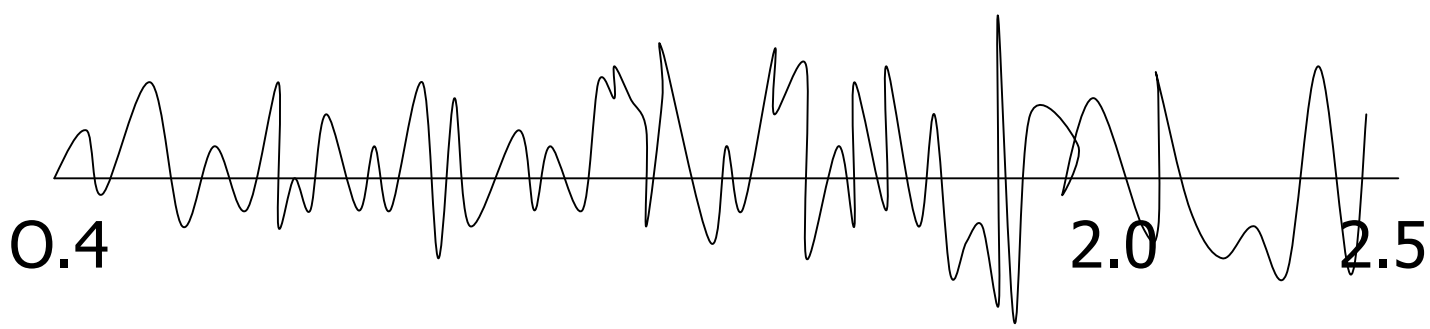


Example: Detecting Mineral Composition from Spectra

- Given: Reference Library of visible near/infrared spectra over 900 frequencies for 135 pure minerals
- Find: An algorithm that will identify new rock and soil samples that contain a specific mineral class, e.g., carbonates.

Basic Difficulties

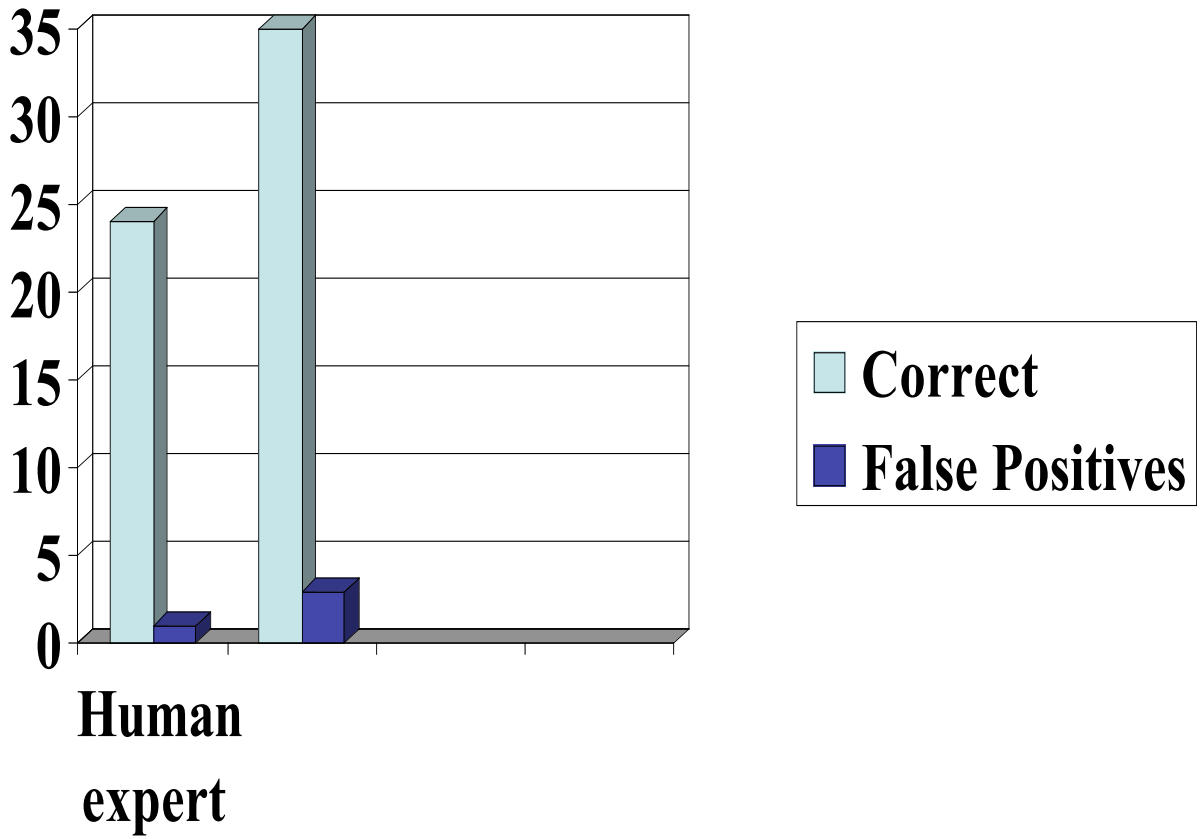
- Reference library is very small sample of pure minerals that occur in a variety of combinations in natural rocks and soils.
- The distinctive signal of each class of minerals is confined to a small segment of the spectrum—rest is “noise”—e.g., carbonate signal is between 2.0 and 2.5 nanometers.



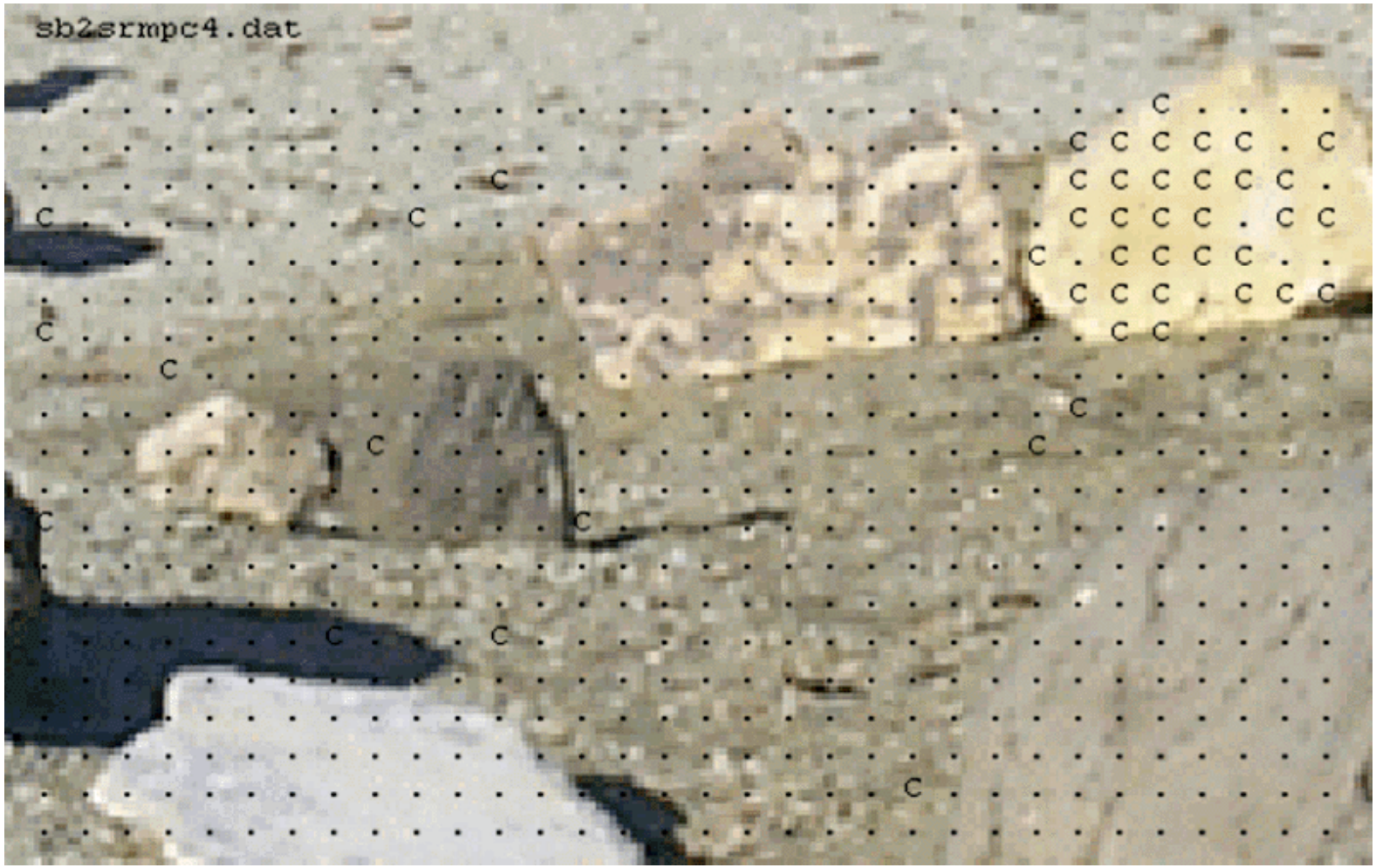
Methods for Carbonates

- Neural net classifiers fail because adequate training data is not available
- Regression classifiers fail, for reasons already noted, and because there are more Library minerals than signal frequencies between 2.0 and 2.5.
- Bayes net classifier succeeds better than human experts.

Finding 90 Carbonaceous Samples among 190 Samples: Bayes Net Algorithm Vs. Human Expert



NASA Ames Test: Regression Methods Find Carbonates Everywhere; Bayes Nets Find the Rock



White Rock in upper right hand corner is carbonate. No other carbonates are present.

Application: Dropouts

- Data from 1993-94 U.S. News and World Report surveys.
- Classical analysis (with TETRAD program) says

Everything else — *Average SAT* ⟶ *Dropout rate*

- 1994 CMU alters financial aid policies to increase average SAT scores of freshman class

Applications: Dropouts

- Dropout rate from 1994 on decreases monotonically with increasing average SAT of freshman class

Causal Structure of Sea Surface Temperature (SST) and Sea Level Pressure Teleconnections

3. Causal Analysis of Teleconnections (Chu, with some help from Silva, and using an idea of Spirtes')

- Sea Surface Temperature (SST) and SLP anomalies clustered by spatial region =>
- Time series of indices of SST and SLP by region
- For each time series for each region, another time series variable representing a one month lag of the original series is created—then time series variables for two month and three month lags.
- Correlations of all time series are computed.
- TETRAD IV search for causal model of the time series => almost unique causal model
- Residuals after regression of each variable on its direct causes analyzed by TETRAD IV search for "simultaneous" causation => Same causal structure between clusters as for time series
- Using General Additive Model (parameter free), the Markov Blanket of each time series cluster is estimated => Same causal structure.

Ocean Indices

- QBO (Quasi Biennial Oscillation): Regular variation of zonal stratospheric winds above the equator
- SOI (Southern Oscillation): Sea Level Pressure (SLP) anomalies between Darwin and Tahiti
- WP (Western Pacific): Low frequency temporal function of the ‘zonal dipole’ SLP spatial pattern over the North Pacific.
- PDO (Pacific Decadal Oscillation): Leading principal component of monthly Sea Surface Temperature (SST) anomalies in the North Pacific Ocean, poleward of 20° N
- AO (Arctic Oscillation): First principal component of SLP poleward of 20° N
- NAO (North Atlantic Oscillation) Normalized SLP differences between Ponta Delgada, Azores and Stykkisholmur, Iceland

- (From “Dicovery of Climate Indices Using Clustering”, Steinbach et al 2003) Thanks to Mike Steinbach for providing us with the original data.

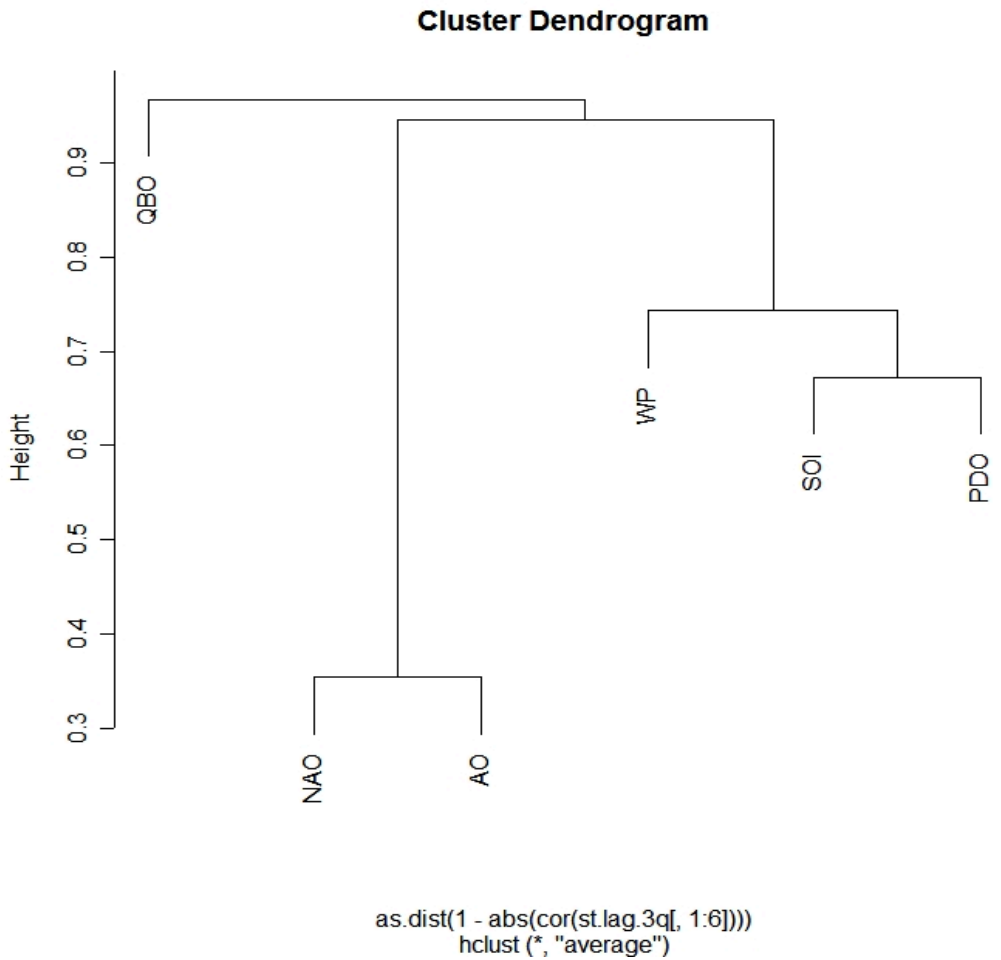
Original Data

Time Points	QBO	SOI	WP	PDO	AO	NAO
1st Month	q_1	s_1	w_1	p_1	a_1	n_1
2nd Month	q_2	s_2	w_2	p_2	a_2	n_2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
504th Month	q_{504}	s_{504}	w_{504}	p_{504}	a_{504}	n_{504}

Clustering of the six ocean indices

- Distance between *x* and *y* is defined as:

  1 – |corr(x, y)|
- Using the average pairwise distance between points in two subclusters as the distance between the two subclusters



Data transformed for causal inference

- $QBO_0 = \{q_1, q_2, \ldots, q_{501}\}$
- $QBO_1 = \{q_2, q_3, \ldots, q_{502}\}$
- $QBO_2 = \{q_3, q_4, \ldots, q_{503}\}$
- $QBO_3 = \{q_4, q_5, \ldots, q_{504}\}$
- $SOI_0 = \{s_1, s_2, \ldots, s_{501}\}$
- … … …
- $NAO_3 = \{n_4, n_5, \ldots, n_{504}\}$

General Additive Models (GAMs)

- While PC assumes that if X, Y, directly cause Z, they must be related by

$$Z = aX + bY + \varepsilon$$

where a, b are real constants and ε is a Normally distributed random variable,

GAMs assume that if X, Y directly cause Z, they must be related by

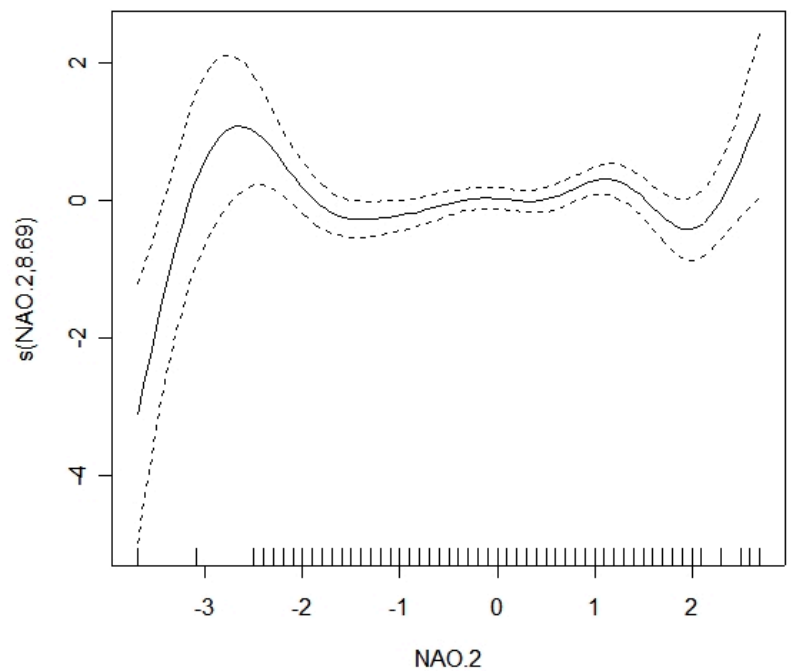
$$Z = f(X) + g(Y) + \varepsilon$$

where f, g are *any* continuous functions.

GAMs are much more general, but GAMs are not a search procedure, and require prior guesses as to which variables are causes and which are effects—provided by the PC algorithm.

Data is not perfect

- Collinearity:
  - *Corr(QBO_2, QBO_3) = 0.95*
  - Consequence: Conditional on QBO_2, QBO_3 and SOI_2 seems independent
  - Result: The edge between QBO_3 and SOI_2 is incorrectly removed
- Nonlinearities, Example:
  - Nonlinear relation between NAO_2 and WP_3
  - *Corr(WP_3, NAO_2) = 0.065*
  - Result: The edge between WP_3 and NAO_2 is incorrectly removed.
  - Could also result in incorrectly added edges in other situations



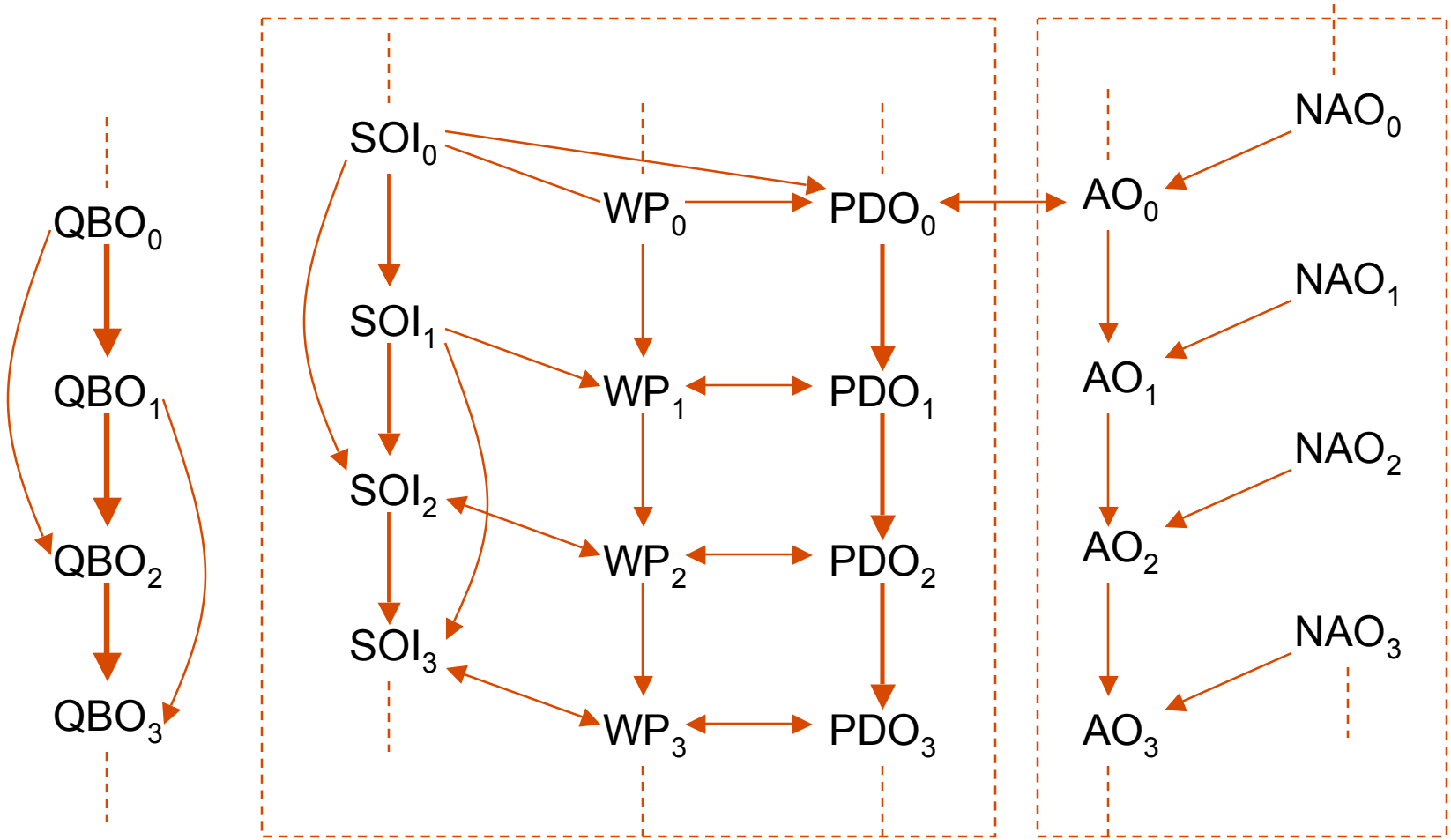
Our Approach

- Combine the parametric method (linear model) with semiparametric method (additive model):
  - First, assuming a linear model, generate a causal pattern using PC or FCI algorithms
  - Then, add edges incorrectly removed by the PC or FCI algorithms:
    - Regress with general additive modelQBO_3, SOI_3, …, NAO_3, respectively, against all the other variables. If a predictor variable is significant, then this predictor must be connected with the response variable conditional on all the other variables.
  - In this study, we did not find with general additive regression any edges incorrectly added by the PC or FCI algorithms.

Full Graph:

PC Algorithm Output

Time Direction



Find contemporary relations

- Since the time series are at monthly intervals, there may be causal connections that occur at more rapid rates and are missed by the time series analysis.
- Ideally, these connections would relate the variables QBO, SOI, WP, PDO, AO, NAO in the same way as the time series analysis.
- Granger 1999, Moneta and Spirtes,2003, Hoover, 2004
  - Assuming linear model, regress QBO_3, SOI_3, …, NAO_3, respectively, against all the variables in the previous time points
  - Feed the residuals into PC or FCI algorithms
  - The resulting pattern tells the contemporary causal relations
- Repeat the above three steps, but replace the linear regression by general additive model regression
  - Turns out we get exactly the same pattern as using linear regression.

Contemporary Graph:
PC Algorithm Output

Contemporary Graph:
FCI Algorithm Output

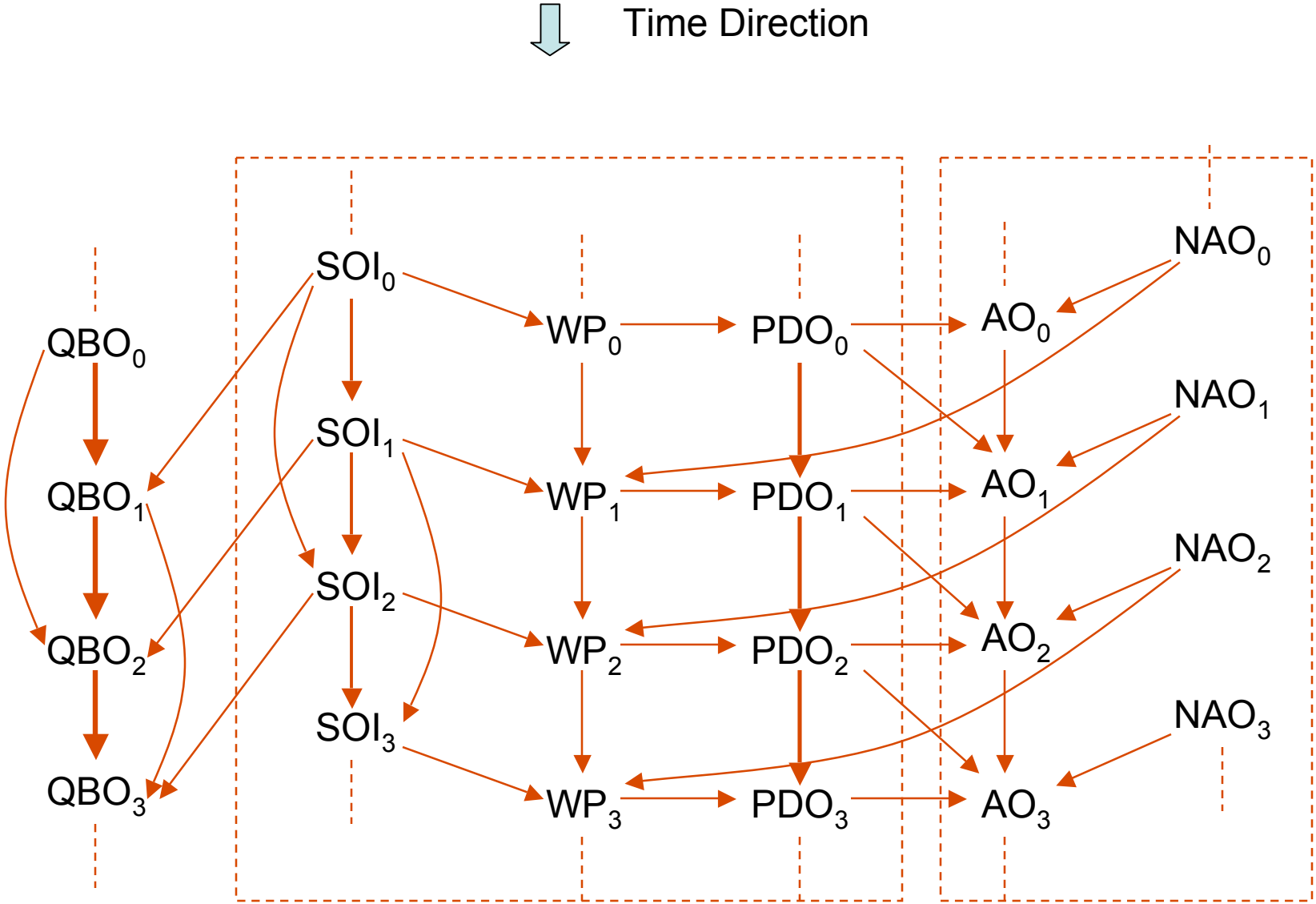
Adding missing edges

- Using general additive model, regress QBO_3, SOI_3, …, NAO_3, respectively, against all the other variables.
  - If a predictor variable is significant, then this predictor must be associated with the response variable conditional on all the other variables
  - E.g., with all the other variables present, PDO_2 is significant in predicting AO_3, hence PDO_2 and AO_3 must be associated given all the other variables
  - Caution: It is possible that, based on the pattern, two variables are associated conditional on other variables, but neither is a significant term, with the presence of all the remaining variables, in predicting the other variable
- No edges *removed* from the PC output by the GAM.

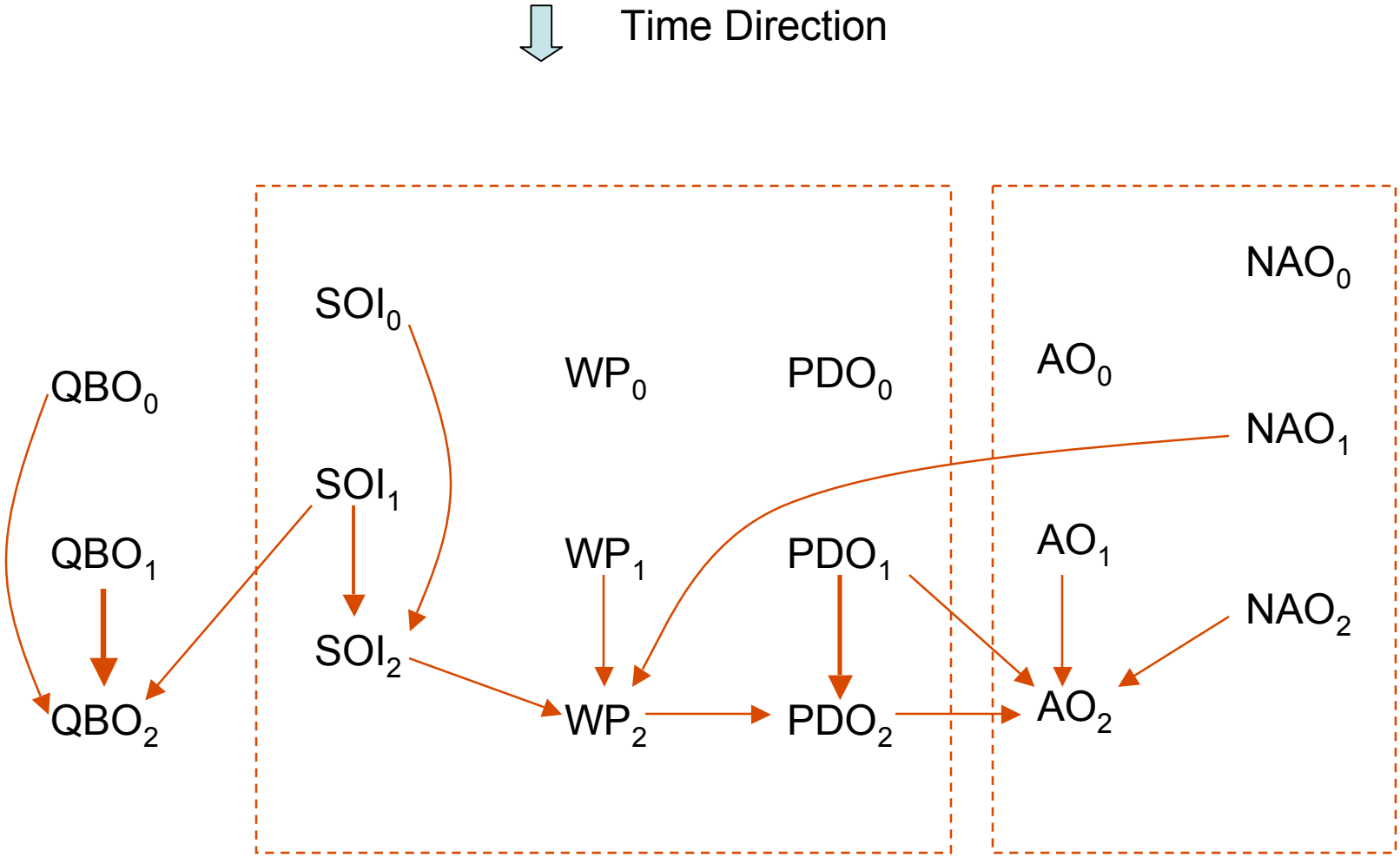
Correct Orientations

- As noted, PC may make errors because of non-linear relations.
- The General Additive Model does not assume linearity, and can be used to check directions of edges of PC output as well as edges
- In this case, the conditional independence relations found with the general additive model remove the double headed arrows in the PC output but do not change any other orientations.

Final Full Graph



Final Reduced Graph



How Do Children Learn Causal Relations?

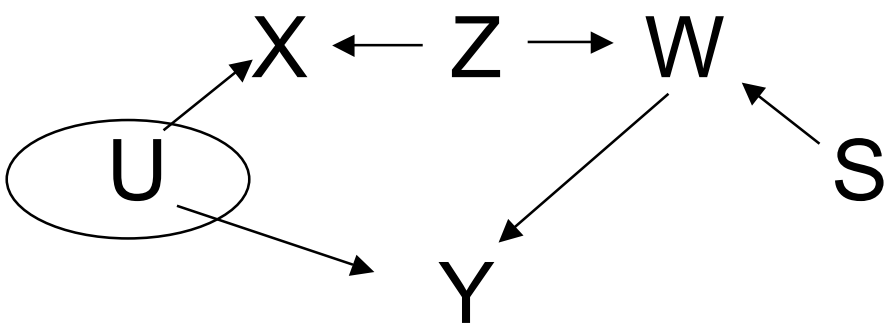
- Big open question, but
- Growing evidence from several laboratories that they use simplified versions of Bayes net algorithms that exploit the Markov condition.

Lessons: Common Big Mistakes

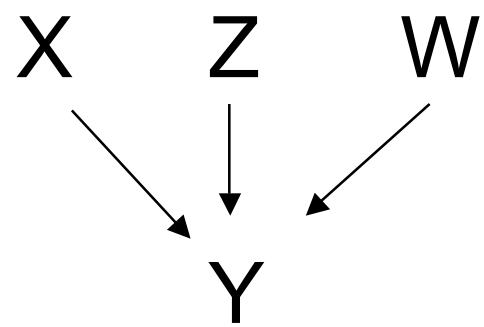
- Regression
- Scales
- Aggregation
- Conditioning on Aggregates

Regression Is an Unreliable Search Method for Causation

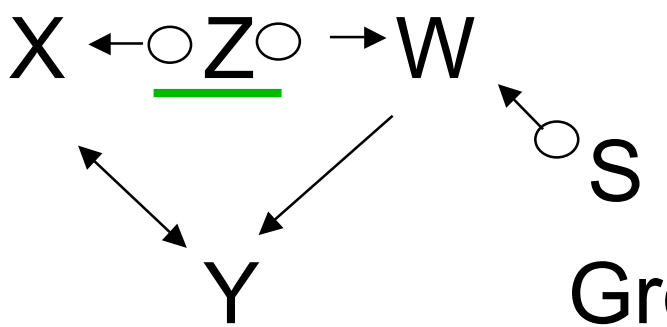
True Structure:



Regression Analysis



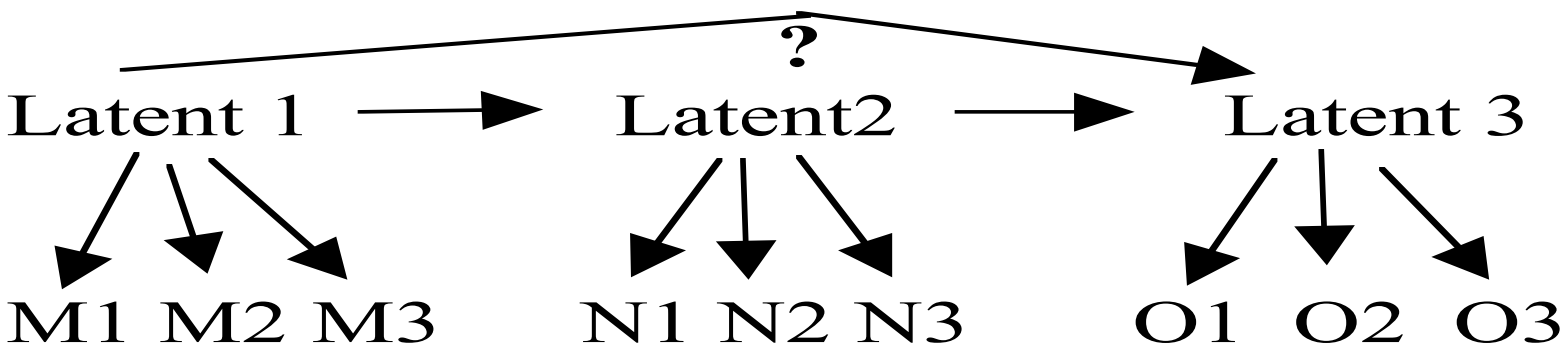
TETRAD Algorithm Result



means latent common cause

Green underline means the edges cannot collide at Z

Scales and Fatherhood: What My Daughter Had to Do for Her Ph.D at Harvard



$M = \Sigma Mi$ $N = \Sigma Ni$ $O = \Sigma Oi$

Regress O against M, N

If the regression coefficient for M is significant, conclude the direct connection from Latent 1 to Latent 2 exists.

(L1 = cognitive ability before stroke; L2 = cognitive ability immediately after stroke; L3 = cognitive ability 6 months after stroke)

What Could (and Should) Be Done to Determine Latent Structure in Linear, Normal Systems?

From a bunch of measured variables: X1….Xn, some of which

may influence each other

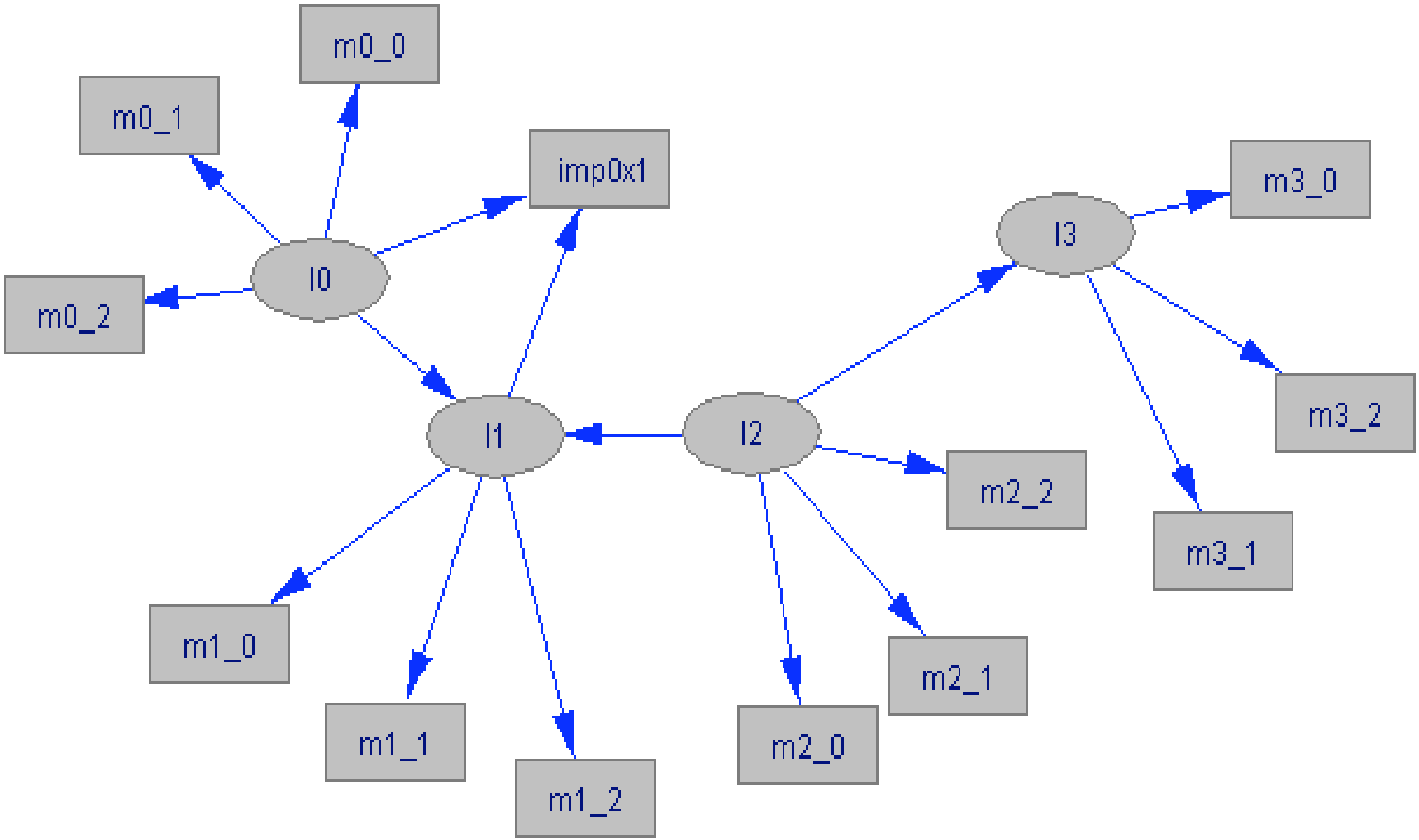
may have common unmeasured causes

and *no* other prior knowledge…

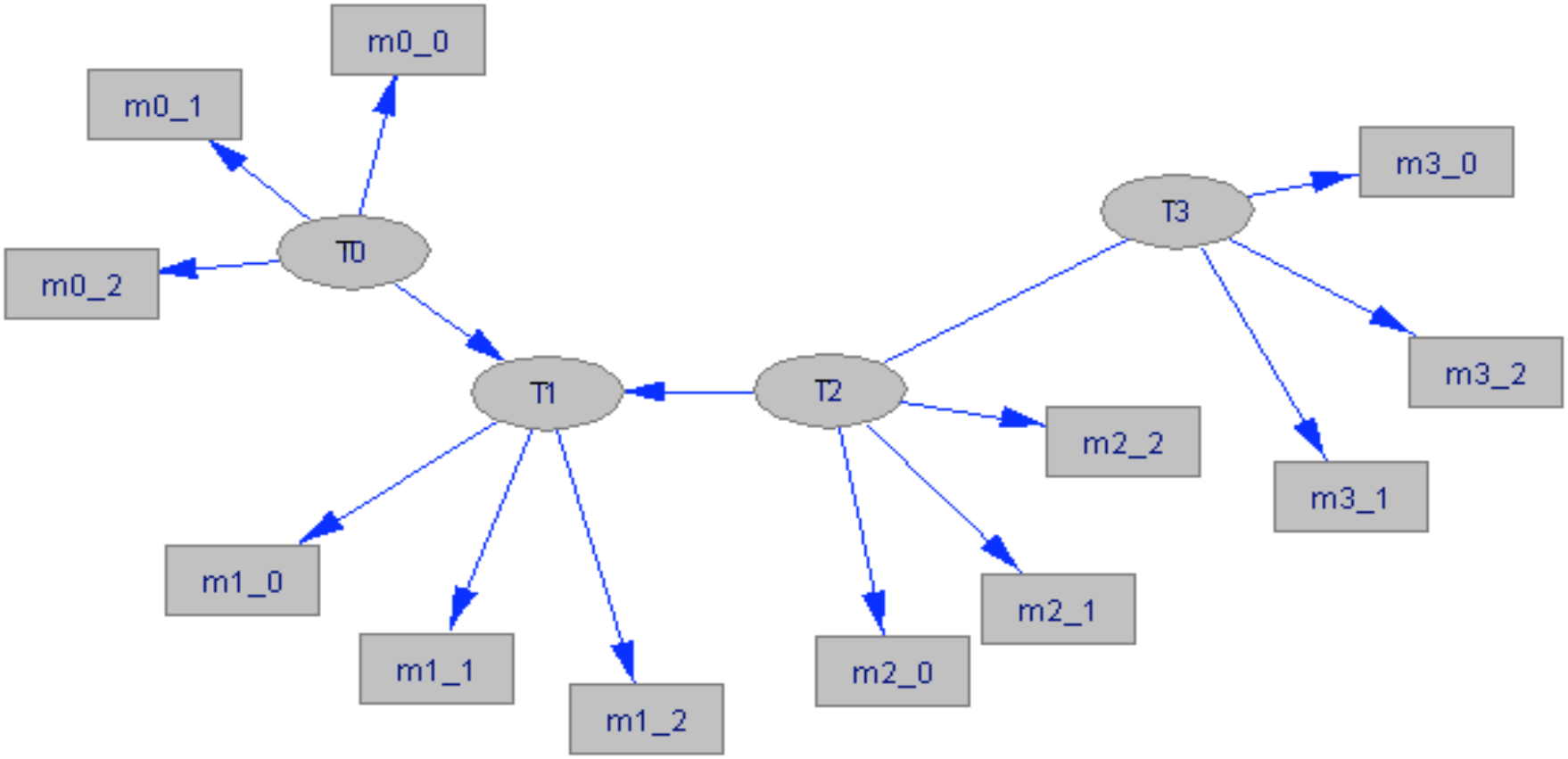
Find the causal structure (up to the Markov equivalence class) among the unmeasured common causes, and for each such latent variable, L, find a set of measured variables that have L as their only common causal source.

Yes, it can be done.

Example: Data (n = 2,000), **for *m* variables only**, from the following:



…when given to a program in TETRAD IV yields the result:



N.B. One edge between latents cannot be oriented

Aggregation Destroys Conditional Independence in Non-Linear Systems

- Suppose for each of several units, I, the causal structure is

X-> Y -> Z <- W

And the dependencies are not linear.

So X is independent of Z conditional on Y.

But ΣiX is not independent of ΣiZ conditional on ΣiY.

Kiss of death for attempts to discover gene regulation networks from correlations of gene expressions.

Conditioning on Aggregates Destroys Independence

- Suppose Xi is independent of Xj for "enions" ui, uj, i,j = 1…..n
- Let $X = (1/n)\Sigma_k X_k$

Then except for special distributions, Xi, Xj are not independent conditional on X.

Science is Changing

In area after area, science is drifting to semi-automated model search, often over huge, really, really, really huge, datasets, and to experimental designs that simultaneously seek multiple causal relations.

Science and Computation

The strategies of scientific inquiry have been historically limited by two things: data and computational power. Examples:

- **Legendre, comets and statistical inference**
- **Fisher, hypothesis testing and the design of experiments**
- **Thurstone, factor analysis and psychometrics**
- **Behaviorism in psychology**
- **Applications of quantum theory of matter in chemistry**
- **Prediction of forest fires**

The Force of Technology

- Cheaper and cheaper computation after 1960 prompts development of algorithms for searching data for
  - Rules for recognizing objects with a particular property from their other properties
  - Rules for classification
  - Causal relations among features

The Force of Technology

- Despite wide resistance, the accumulation of data sets too vast for humans to survey forces recourse to automated search.

Example: The MODIS Satellites

Satellites

- Provide continuous data from which to try to predict fires, as well as detection of fires and estimation of their size.
  - Measures of leaf cover
  - Measures of temperature
  - Measures of soil moisture

  Etc.

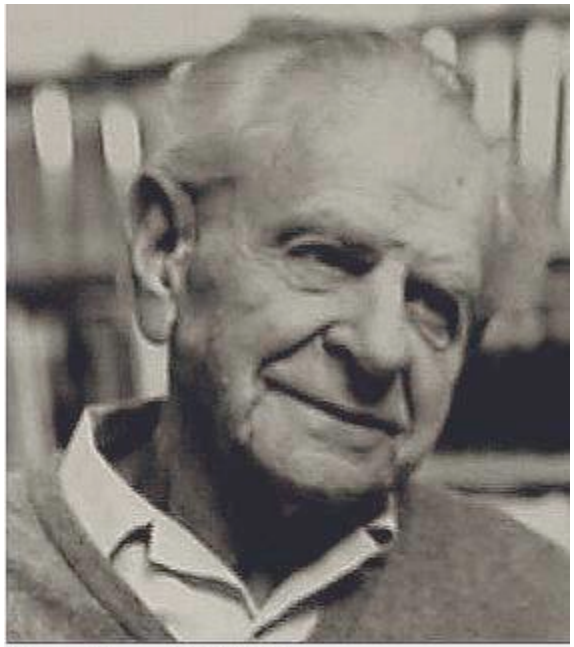
The Mystification of Inquiry

Philosophers: Scientific Discovery is Necessarily Uncomputable

- Karl Popper
- Rudolf Carnap
- Russell Hanson
- Tom Kuhn

Scientists: Scientific Discovery is Necessarily a Cottage Industry

- Albert Einstein
- Ronald Fisher



Sir Karl Popper (1902-1994)

Tradition versus Search

- Traditional scientific method: Someone conjectures a hypothesis; someone makes observations or conducts experiments to test it.
- Search: Huge data sets are collected; automated methods search for regularities in the data, generating and testing hypotheses as they go along.

Searching for nearly true models is like searching for….

Recommendations for Finding a Needle in a Haystack

- Popper; Draw a straw at random, test it for steel and eye; draw another; keep going.
- Mayo: Yeah, and whip that baby severely!
- Pomos: Whoa, meaning change Dude: *needle* = *made of straw*. Success!
- Data Mining: Run a magnet through the haystack.

You really should spend your summer learning about causal Bayes nets

Sources;

P. Spirtes, C. Glymour and R. Scheines, Causation, Prediction and Search, Springer Lecture Notes in Statistics, 1993; 2nd edition, MIT, 2000

J. Pearl, Causality, Oxford, 2000.

C. Glymour and G. Cooper, Causation, Computation and Discovery, MIT, 1999

B. Shipley, Cause and Correlation in Biology, Oxford, 1999.

Causal and Statistical Reasoning On-line course:

http:// www. phil.cmu.edu/projects/csr

Causality Lab: http:// www. phil.cmu.edu/projects/causalitylab

TETRAD Software: http:// www. phil.cmu.edu/projects/tetrad