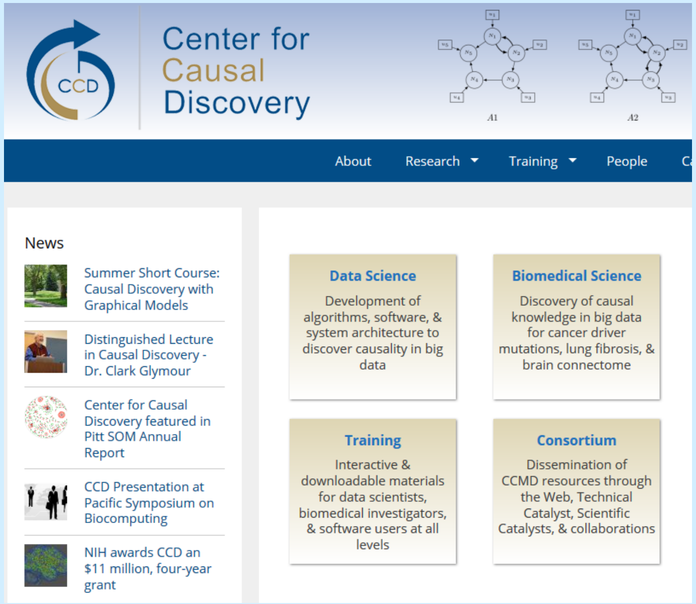
Center for Causal Discovery:

Summer Workshop - 2015



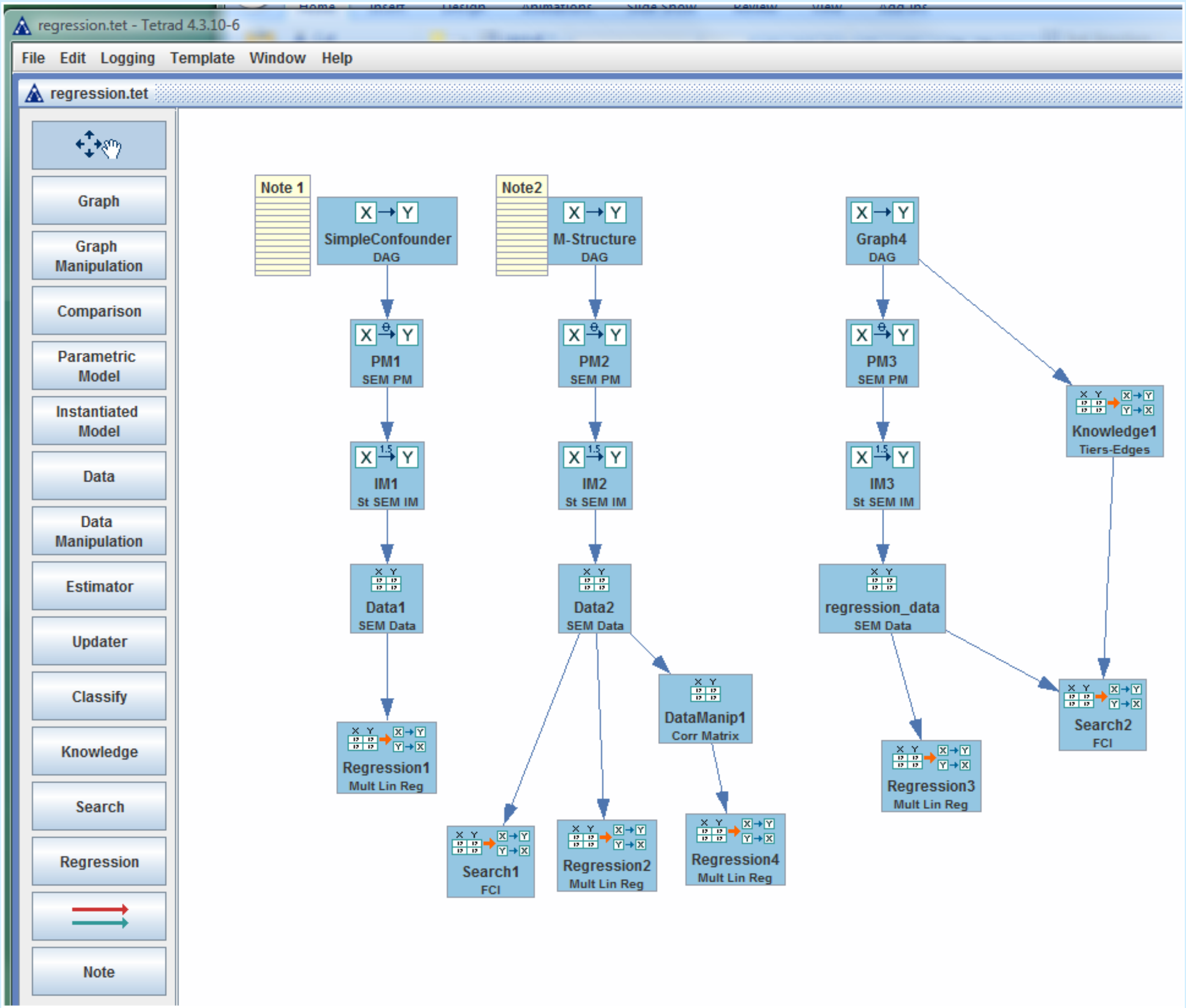
June 8-11, 2015

Carnegie Mellon University

Goals

1) Working knowledge of graphical causal models
2) Basic working knowledge of Tetrad V
3) Basic understanding of search algorithms
4) Basic understanding of several applications:
    a) fMRI
    b) Lung Disease
    c) Cancer
    d) Genetic Regulatory Networks
5) Form community of researchers, users, and students interested in causal discovery in biomedical research

Tetrad: Complete Causal Modeling Tool



Tetrad

1) Main website: http://www.phil.cmu.edu/projects/tetrad/

2) Download: http://www.phil.cmu.edu/projects/tetrad/current.html
    a) Previous version you downloaded: tetrad-5.1.0-6
    b) Newer version with several bug-fixes: tetrad-5.2.1-0

3) Data files:
    www.phil.cmu.edu/projects/tetrad_download/download/workshop/Data/

Outline

Day 1: Graphical Causal Models, Tetrad

1. Introduction
   a) Overview of Graphical Causal Models
   b) Tetrad
2. Representing/Modeling Causal Systems
   a) Parametric Models
   b) Instantiated Models
3. Estimation, Inference, Updating and Model fit
4. Tiny Case Studies: Charity, Lead and IQ

Outline

Day 2: Search

1. D-separation
2. Model Equivalence
3. Search Basics (PC, GES)
4. Latent Variable Model Search
    a) FCI
    b) MIMbuild
5. Examples

Outline

Day 3: Examples

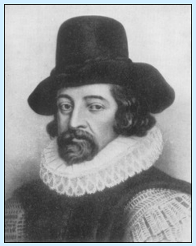
1. Overviews
   a) fMRI
   b) Cancer
   c) Lung Disease
   d) Genetic Regulatory Networks
2. Extra Issues
   a) Measurement Error
   b) Feedback and Time Series

Outline

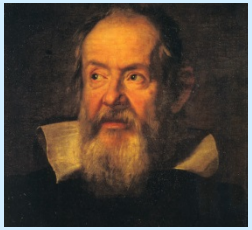
Day 4: Breakout Sessions

1. Morning
   a) fMRI
   b) Cancer
   c) Lung Disease
   d) Genetic Regulatory Networks
2. Afternoon
   a) Overview of Algorithm Development (Systems Group)
   b) Group Discussion on Data and Research Problems

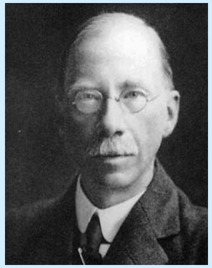
Causation and Statistics



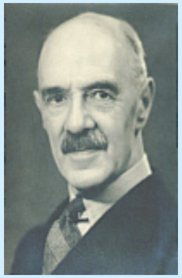
Francis Bacon



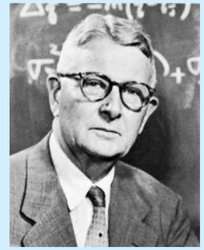
Galileo Galilei



Udny Yule



Charles Spearman



Sewall Wright

Sir Ronald A. Fisher

Jerzy Neyman

Jamie Robins

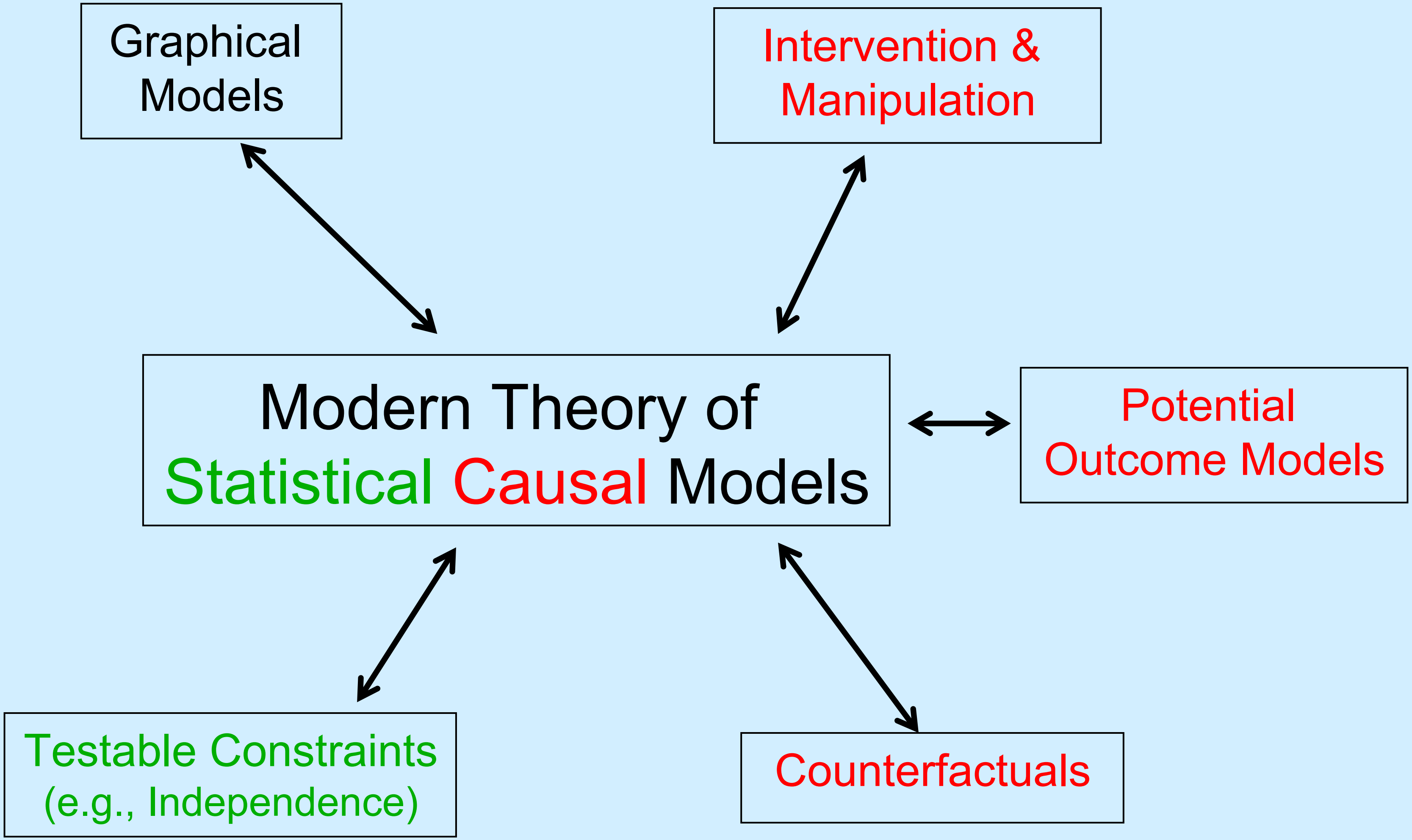
Don Rubin

Judea Pearl

Graphical Causal Models

Potential Outcomes

1500 1600 ….. 1900 1930 1960 1990


Graphical Models
Intervention & Manipulation
Modern Theory of Statistical Causal Models
Potential Outcome Models
Testable Constraints (e.g., Independence)
Counterfactuals

Causal Inference Requires More than Probability

Prediction from Observation ≠ Prediction from Intervention

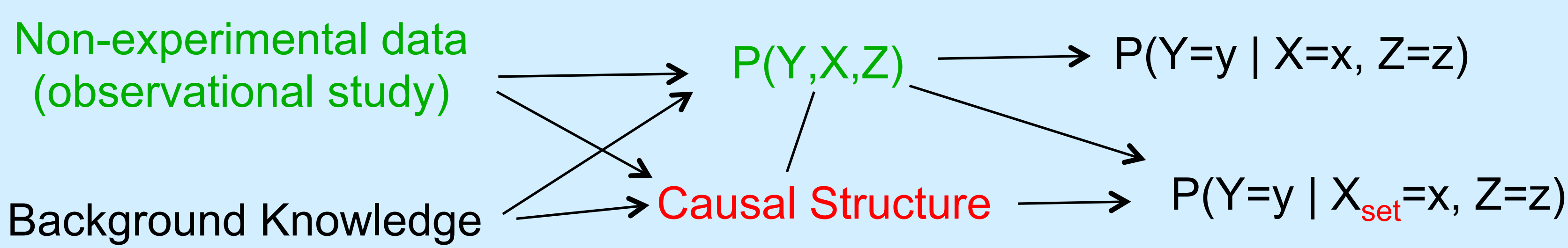
P(**Lung Cancer 1960 = y | Tar-stained fingers 1950 = no**)

≠

P(**Lung Cancer 1960 = y | Tar-stained fingers 1950**$_{set}$ **= no**)

In general: $P(\mathbf{Y{=}y} \mid X{=}x, \mathbf{Z{=}z}) \neq P(\mathbf{Y{=}y} \mid \mathbf{X}_{set}{=}x, \mathbf{Z{=}z})$

Causal Prediction vs. Statistical Prediction:



Estimation vs. Search

Estimation (Potential Outcomes)

- *Causal Question*: Effect of Zidovudine on Survival among HIV-positive men (Hernan, et al., 2000)
- *Problem*: confounders (CD4 lymphocyte count) vary over time, and they are dependent on previous treatment with Zidovudine
- *Estimation method discussed*: marginal structural models
- *Assumptions*:
  - Treatment measured reliably
  - Measured covariates sufficient to capture major sources of confounding
  - Model of treatment given the past is accurate
- *Output*: Effect estimate with confidence intervals

Fundamental Problem: estimation/inference is conditional on the model

Estimation vs. Search

Search (Causal Graphical Models)

- *Causal Question*: which genes regulate flowering in Arbidopsis
- *Problem*: over 25,000 potential genes.
- *Method*: graphical model search
- *Assumptions*:
  - RNA microarray measurement reasonable proxy for gene expression
  - Causal Markov assumption
  - Etc.
- *Output*: Suggestions for follow-up experiments

Fundamental Problem: model space grows super-exponentially with the number of variables

Causal Search

Causal Search:

1. Find/compute *all* the causal models that are indistinguishable given background knowledge and data
2. Represent features common to all such models

Multiple Regression is often the *wrong* tool for Causal Search:

Example: Foreign Investment & Democracy

Foreign Investment

Does Foreign Investment in 3rd World Countries inhibit Democracy?

Timberlake, M. and Williams, K. (1984). Dependence, political exclusion, and government repression: Some cross-national evidence. American Sociological Review 49, 141-146.

N = 72

PO	degree of political exclusivity
CV	lack of civil liberties
EN	energy consumption per capita (economic development)
FI	level of foreign investment

Foreign Investment

Correlations

	po	fi	en	cv
po	1.0			
fi	-.175	1.0		
en	-.480	0.330	1.0	
cv	0.868	-.391	-.430	1.0

Case Study: Foreign Investment

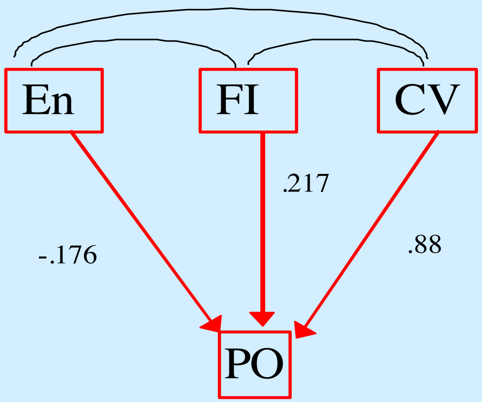
Regression Results

$$po = \boxed{.227*fi} - .176*en + .880*cv$$

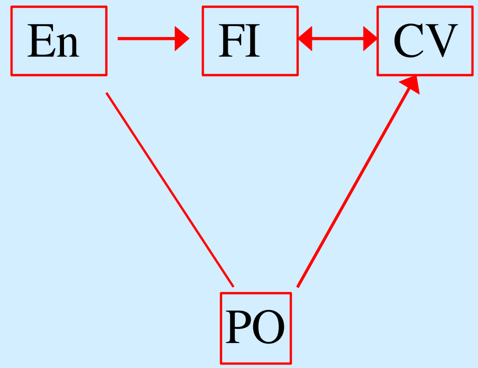
SE	(.058)	(.059)	(.060)
t	3.941	-2.99	14.6
P	$\boxed{.0002}$	.0044	.0000

Interpretation: foreign investment increases political repression

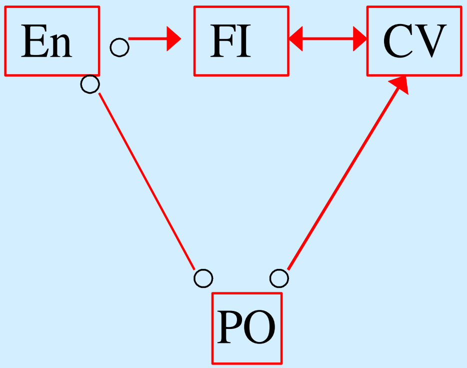
Case Study: Foreign Investment *Alternative Models*



Regression

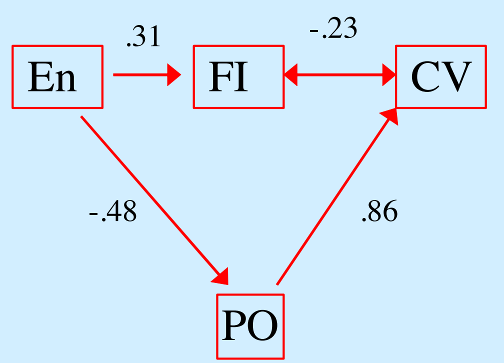


Tetrad - PC



Tetrad - FCI

There is no model with testable constraints (df > 0) that is not rejected by the data, in which FI has a positive effect on PO.

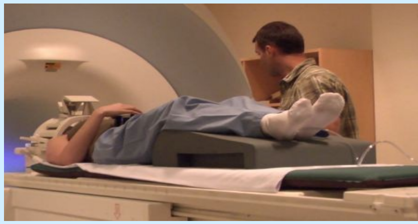


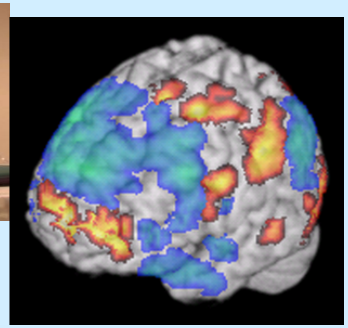
Fit: df=2, $\chi 2$=0.12,
p-value = .94

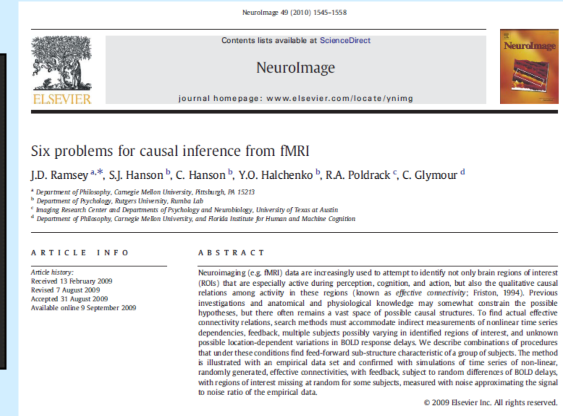
A Few Causal Discovery Highlights

fMRI

(~44,000 voxels)







NeuroImage 49 (2010) 1545–1558

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Six problems for causal inference from fMRI

J.D. Ramsey [a,*], S.J. Hanson [b], C. Hanson [b], Y.O. Halchenko [b], R.A. Poldrack [c], C. Glymour [d]

[a] Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA 15213
[b] Department of Psychology, Rutgers University, Rumba Lab
[c] Imaging Research Center and Departments of Psychology and Neurobiology, University of Texas at Austin
[d] Department of Philosophy, Carnegie Mellon University, and Florida Institute for Human and Machine Cognition

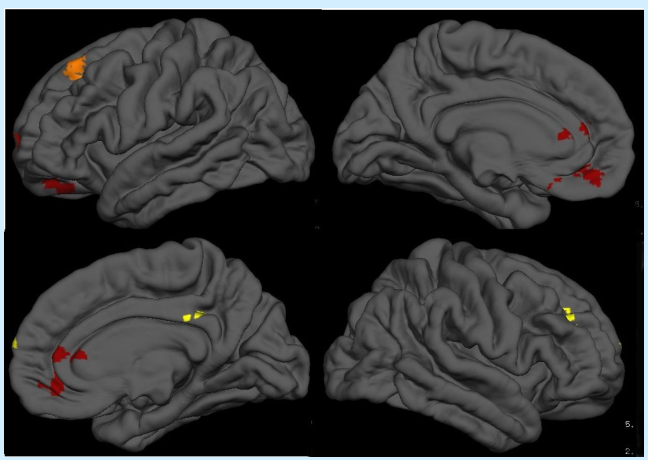
A R T I C L E I N F O

Article history:
Received 13 February 2009
Revised 7 August 2009
Accepted 31 August 2009
Available online 9 September 2009

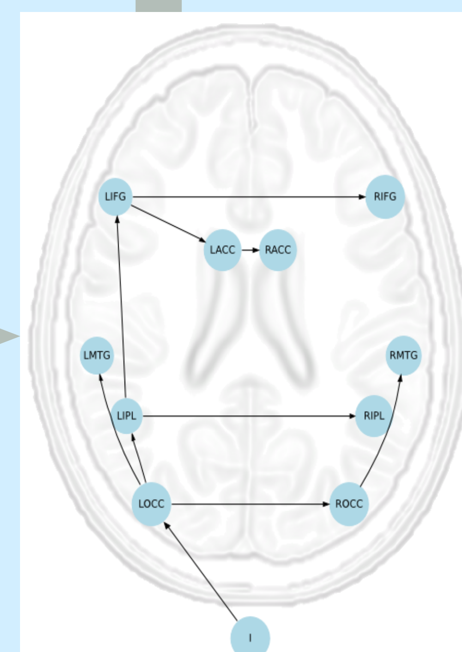
A B S T R A C T

Neuroimaging (e.g. fMRI) data are increasingly used to attempt to identify not only brain regions of interest (ROIs) that are especially active during perception, cognition, and action, but also the qualitative causal relations among activity in these regions (known as *effective connectivity*; Friston, 1994). Previous investigations and anatomical and physiological knowledge may somewhat constrain the possible hypotheses, but there often remains a vast space of possible causal structures. To find actual effective connectivity relations, search methods must accommodate indirect measurements of nonlinear time series dependencies, feedback, multiple subjects possibly varying in identified regions of interest, and unknown possible location-dependent variations in BOLD response delays. We describe combinations of procedures that under these conditions find feed-forward sub-structure characteristic of a group of subjects. The method is illustrated with an empirical data set and confirmed with simulations of time series of non-linear, randomly generated, effective connectivities, with feedback, subject to random differences of BOLD delays, with regions of interest missing at random for some subjects, measured with noise approximating the signal to noise ratio of the empirical data.

Clark Glymour, Joe Ramsey, Ruben Sanchez CMU



Causal Discovery



(ROI)

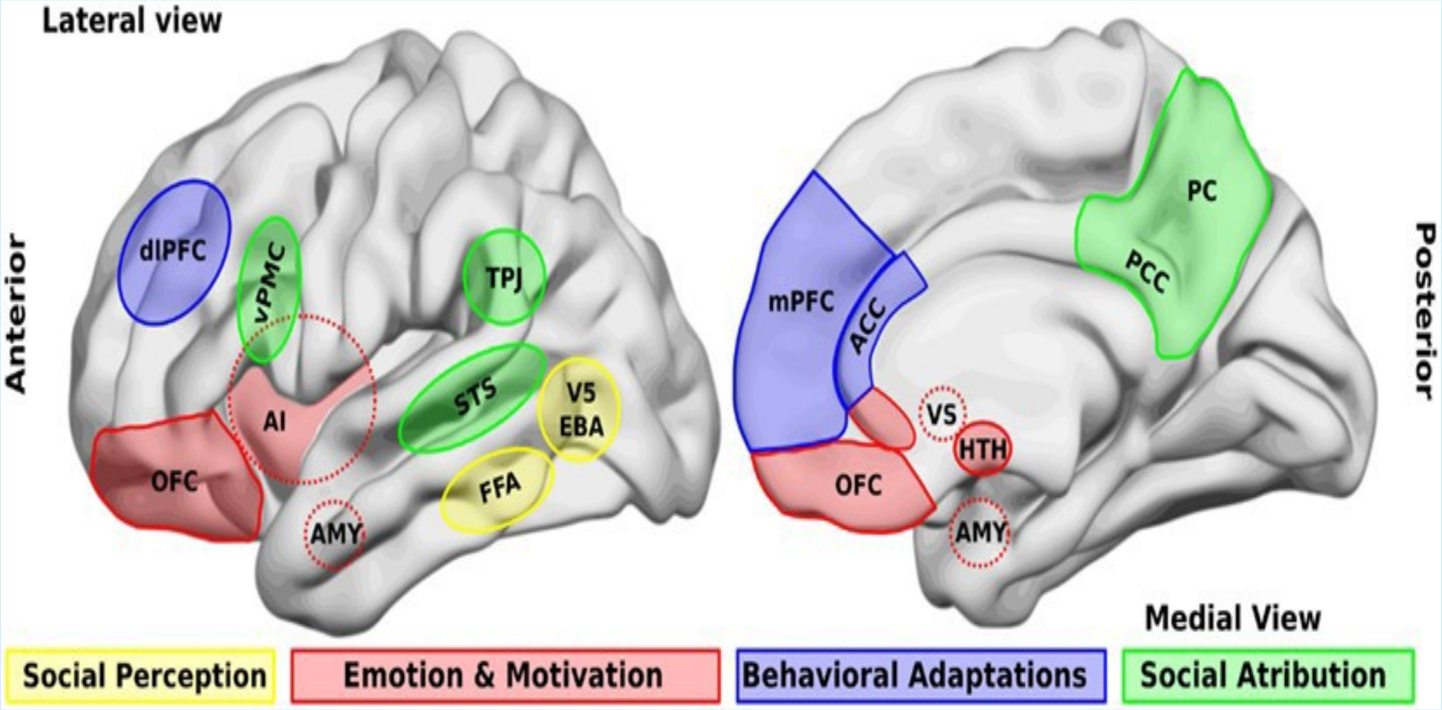
~10-20 Regions of Interest

Autism

Catherine Hanson, Rutgers

ASD vs. NT

Usual Approach:
Search for differential recruitment of brain regions



ASD vs. NT

Causal Modeling Approach:

Examine connectivity of ROIs

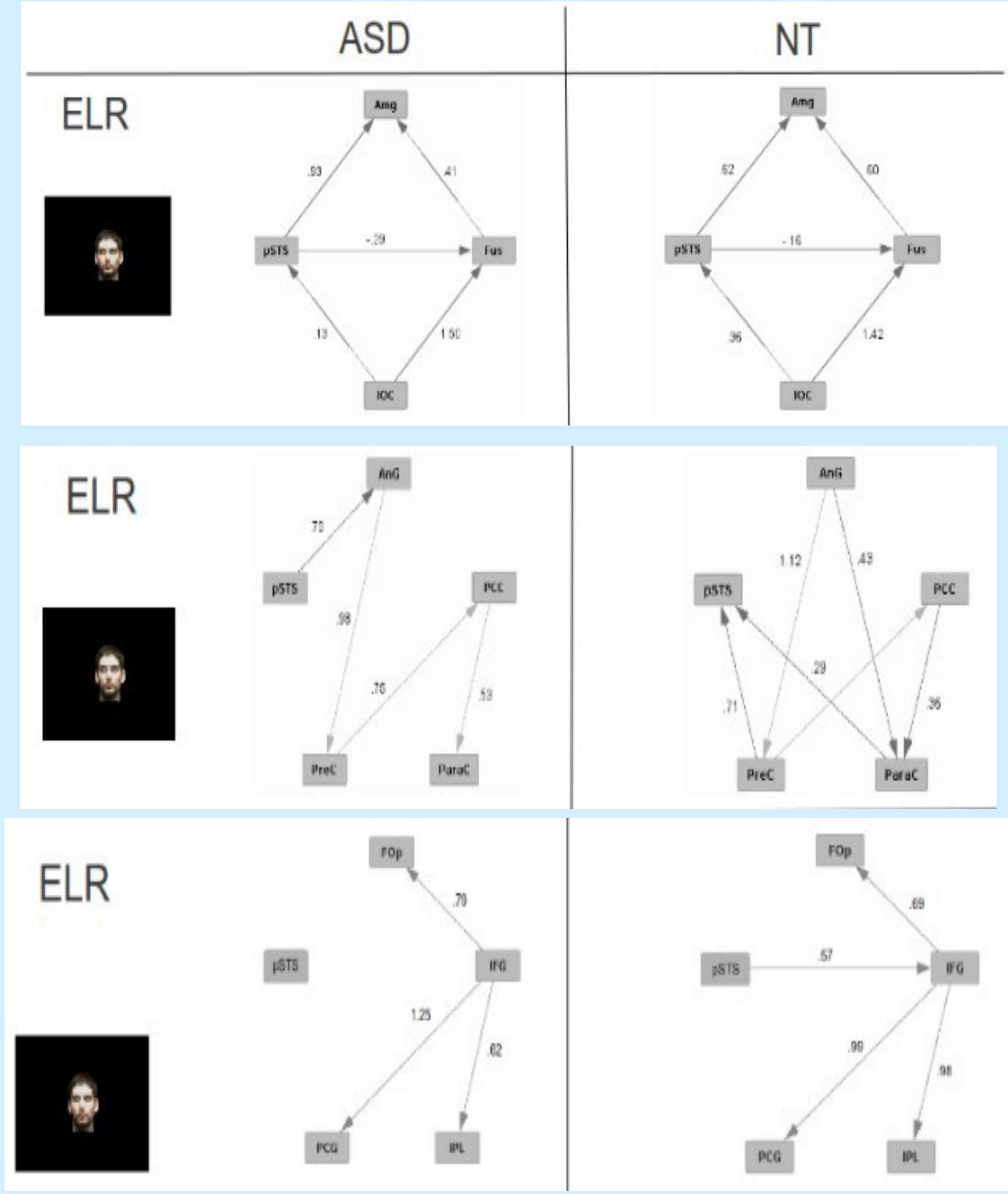
- Face processing network
- Theory of Mind network
- Action understanding network

Results

FACE

TOM

ACTION



What was Learned

face processing: $ASD \approx NT$

Theory of Mind: $ASD \neq NT$

action understanding: $ASD \neq NT$
when faces involved

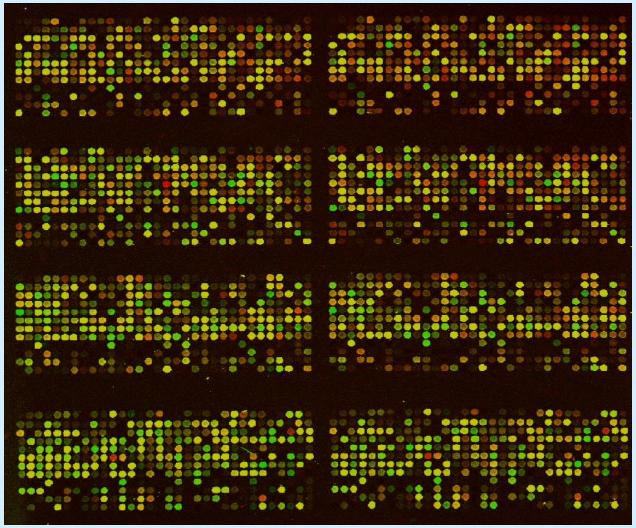
Genetic Regulatory Networks

Arbidopsis

Marloes Maathuis   ZTH (Zurich)

Genetic Regulatory Networks

Micro-array data
~25,000 variables



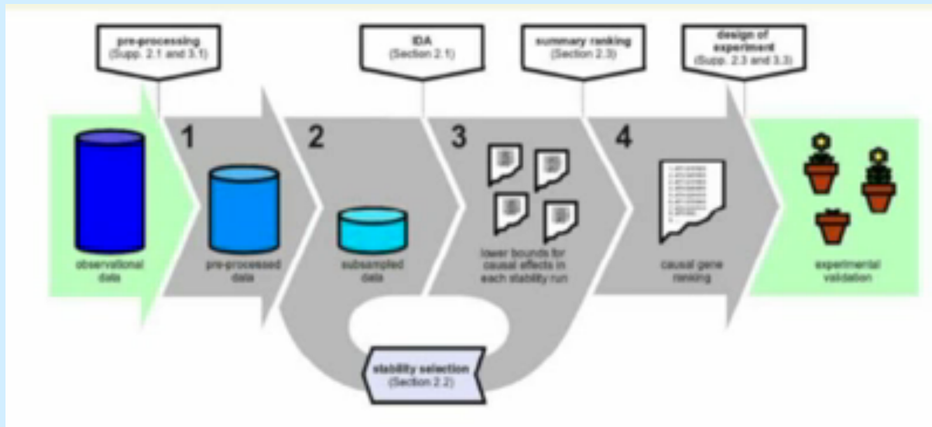
Causal Discovery

Candidate Regulators of Flowering time

Greenhouse experiments on flowering time

Genetic Regulatory Networks

Which genes affect flowering time in Arabidopsis thaliana?
(Stekhoven et al., *Bioinformatics*, 2012)



- ~25,000 genes
- Modification of PC (stability)
- Among 25 genes in final ranking:
  - 5 known regulators of flowering
  - 20 remaining genes:
    - For 13 of 20, seeds available
    - 9 of 13 yielded replicates
    - 4 of 9 affected flowering time
- Other techniques are little better than chance

Other Applications

- Educational Research:
  - Online Courses,
  - MOOCs,
  - Cog. Tutors
- Economics:
  - Causes of Meat Prices,
  - Effects of International Trade
- Lead and IQ
- Stress, Depression, Religiosity
- Climate Change Modeling
- The Effects of Welfare Reform
- Etc. !

Outline

Representing/Modeling Causal Systems

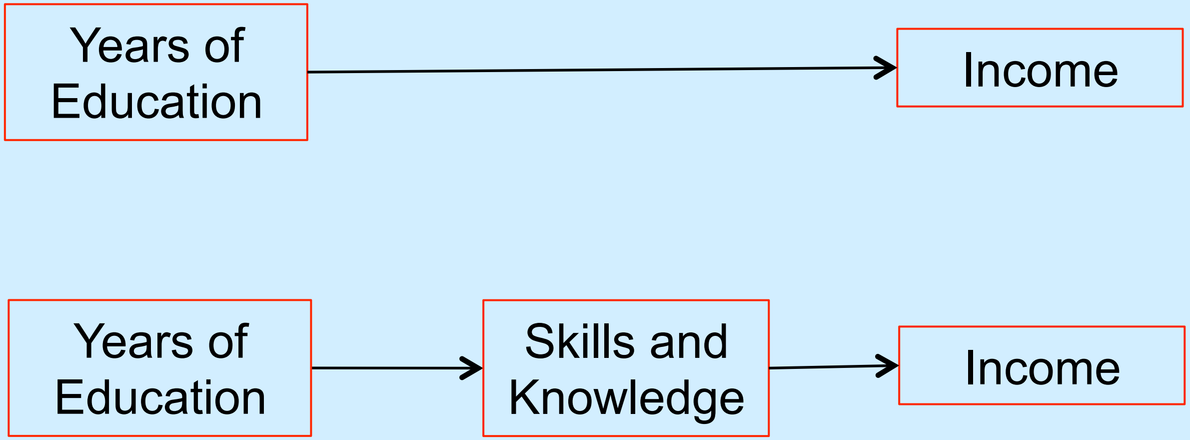
1) Causal Graphs
2) Parametric Models
    a) Bayes Nets
    b) Structural Equation Models
    c) Generalized SEMs

Causal Graphs

Causal Graph G = {**V**,**E**}

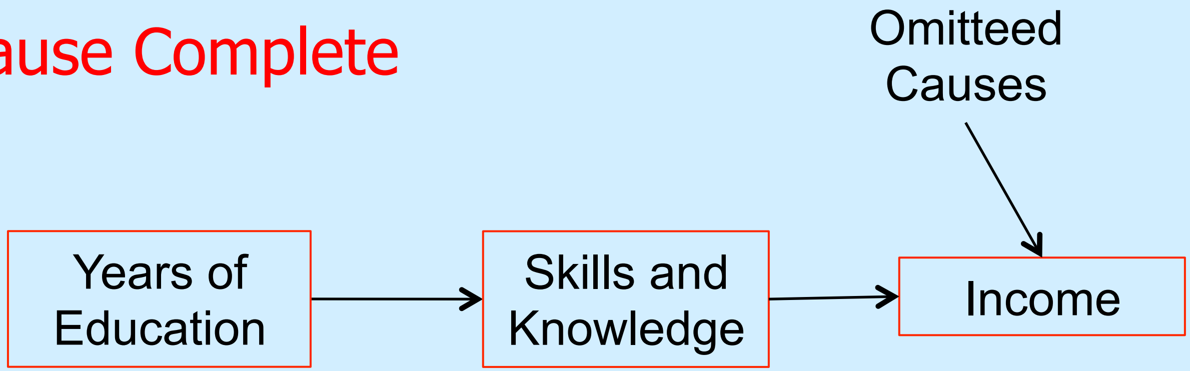
Each edge X → Y represents a direct causal claim:

X is a direct cause of Y relative to **V**

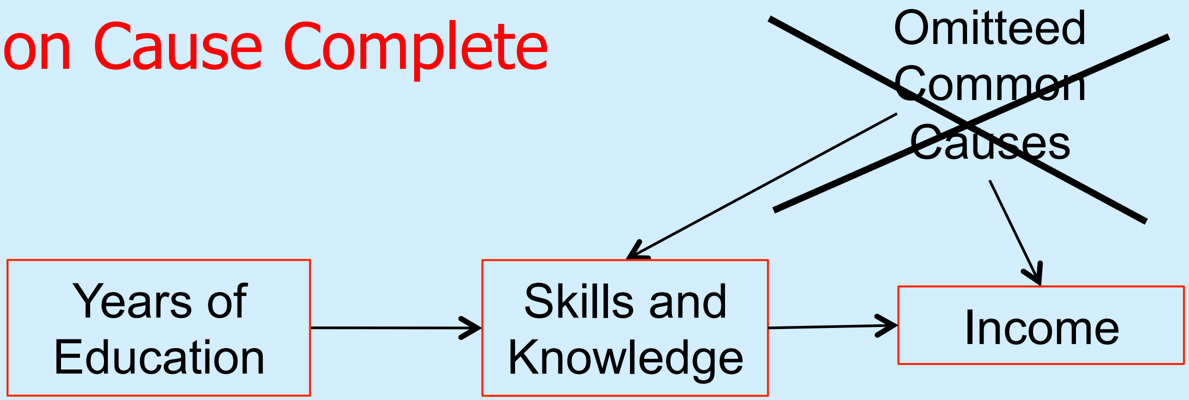


Causal Graphs

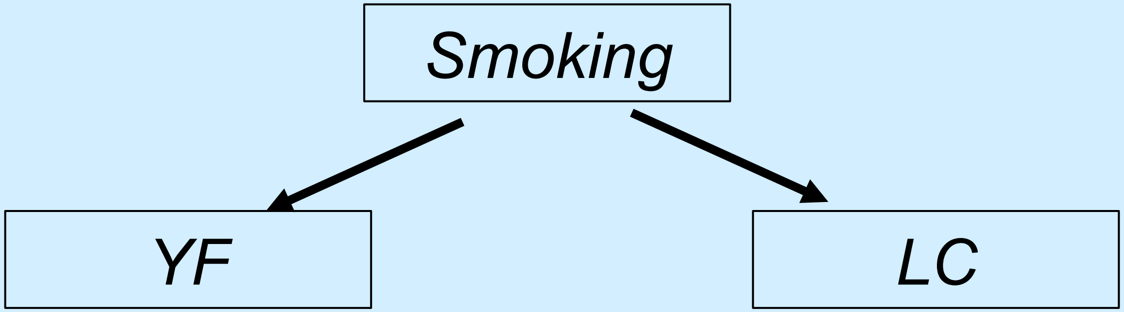
Not Cause Complete



Common Cause Complete



Tetrad Demo & Hands-On



Build and Save two acyclic causal graphs:

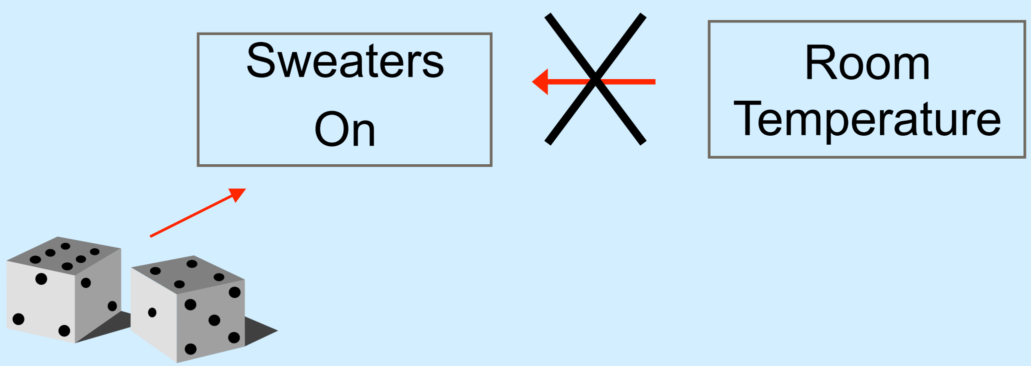
1) Build the Smoking graph picture above

2) Build your own graph with 4 variables

Modeling Ideal Interventions

Interventions on the Effect

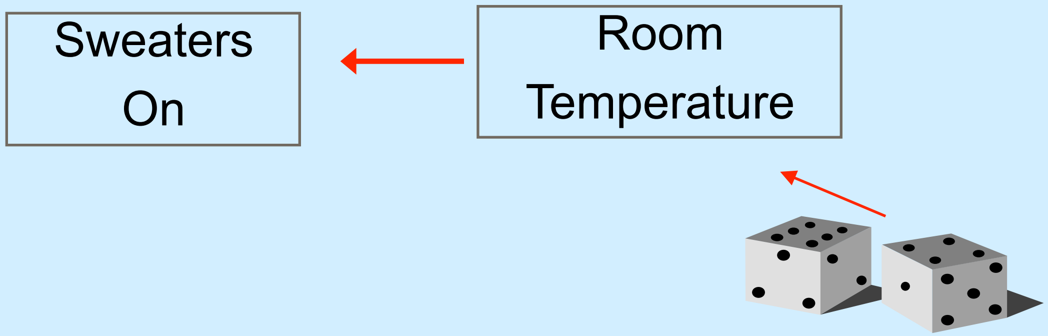
Post experimental System



Modeling Ideal Interventions

Interventions on the Cause

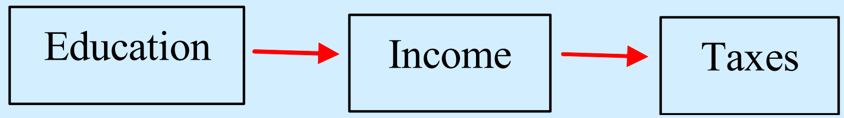
Post-experimental System



Interventions & Causal Graphs

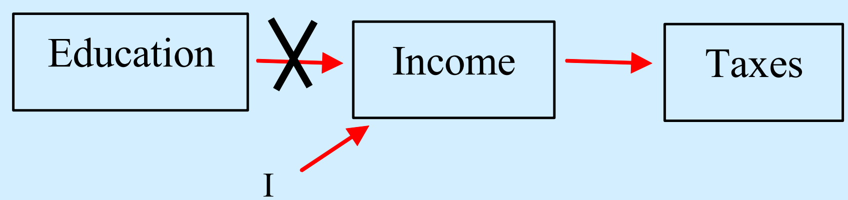
Model an ideal intervention by adding an "intervention" variable outside the original system as a direct cause of its target.

Pre-intervention graph

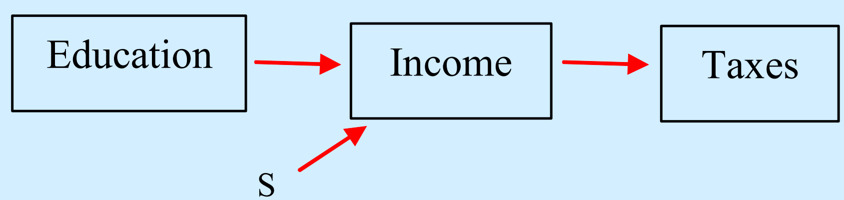


Intervene on *Income*

"Hard" Intervention



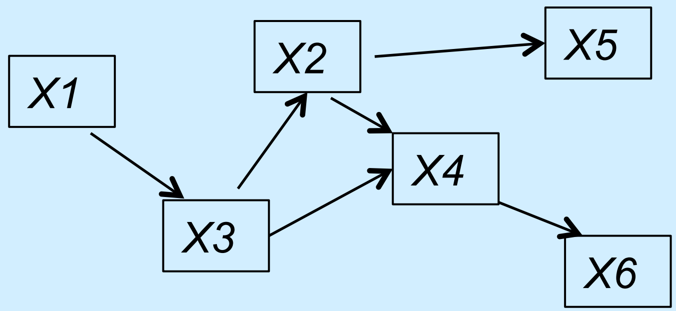
"Soft" Intervention



Interventions & Causal Graphs

Pre-intervention

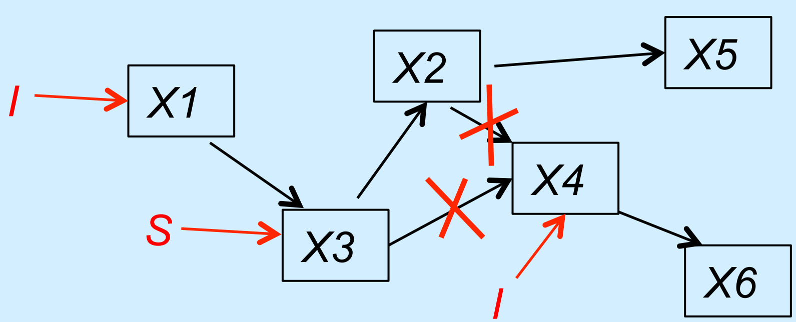
Graph



Intervention:

- hard intervention on both X1, X4
- Soft intervention on X3

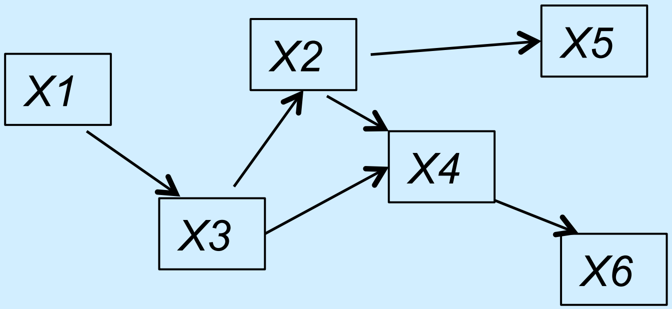
Post-Intervention
Graph?



Interventions & Causal Graphs

Pre-intervention

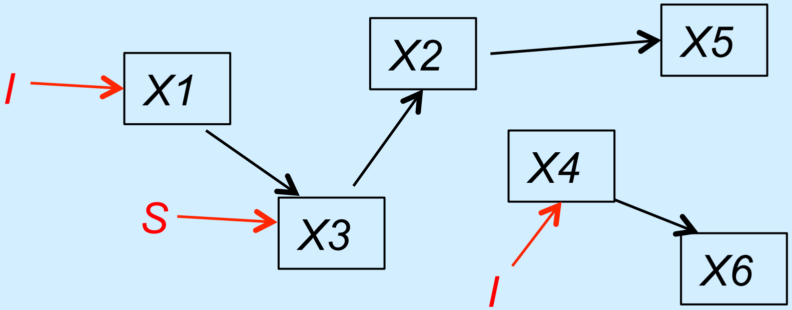
Graph



Intervention:

- hard intervention on both X1, X4
- Soft intervention on X3

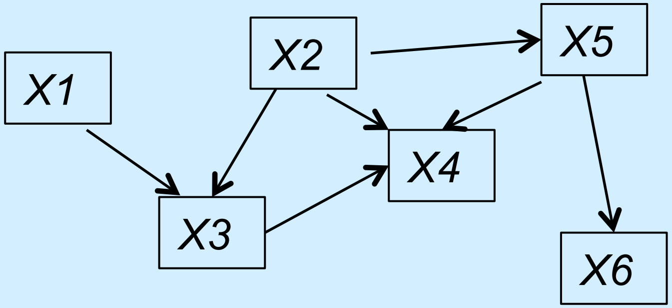
Post-Intervention
Graph?



Interventions & Causal Graphs

Pre-intervention

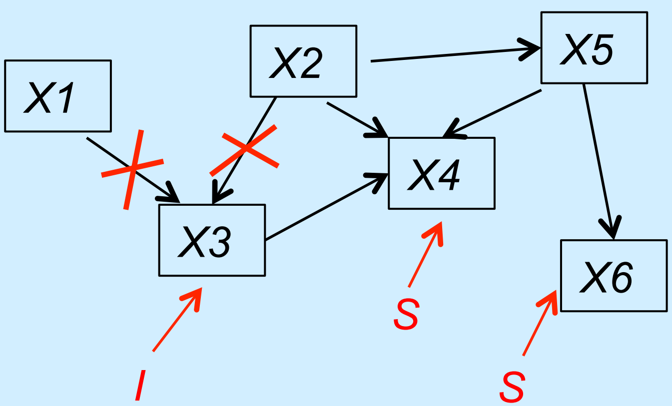
Graph



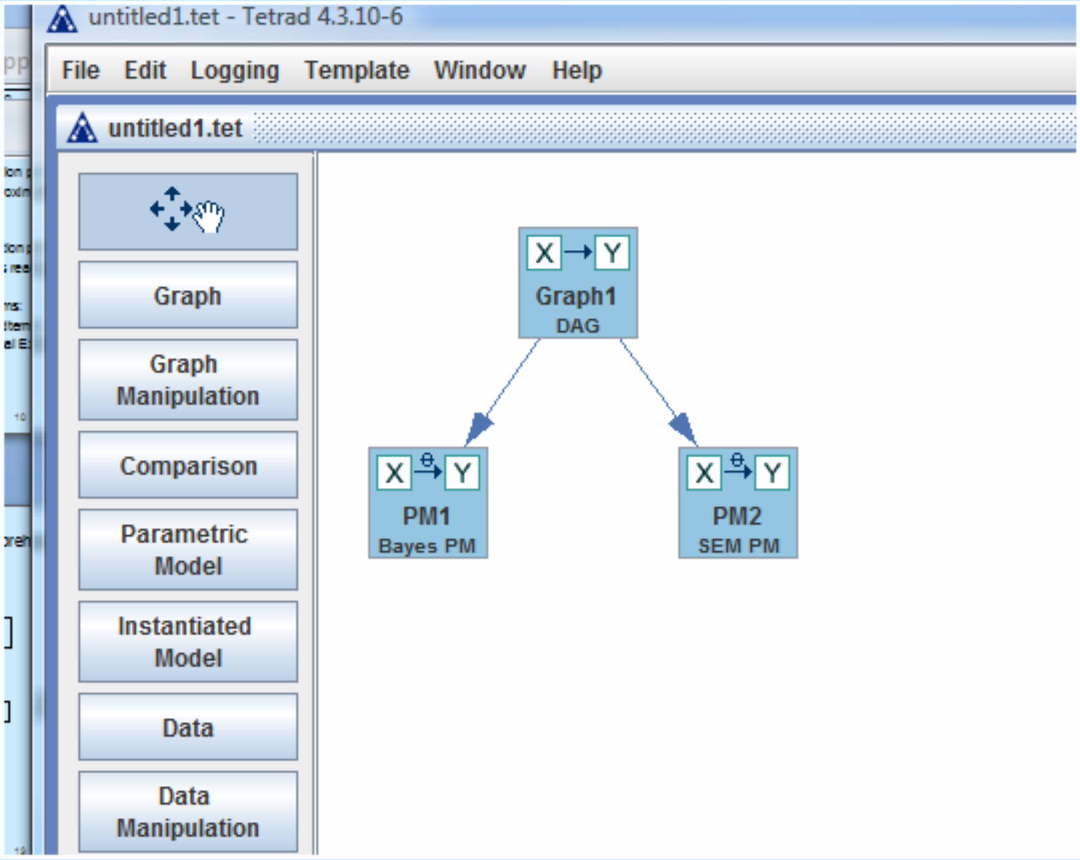
Intervention:

- hard intervention on X3
- Soft interventions on X6, X4

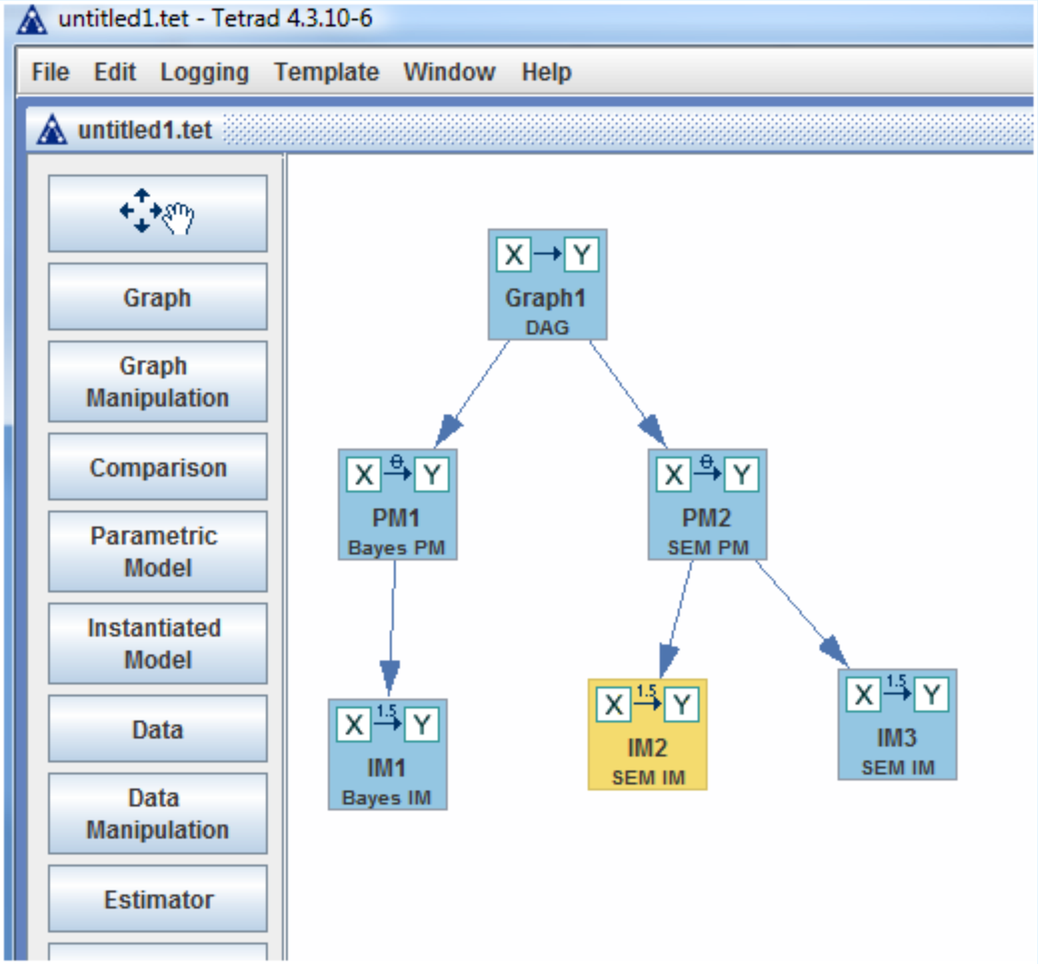
Post-Intervention
Graph?



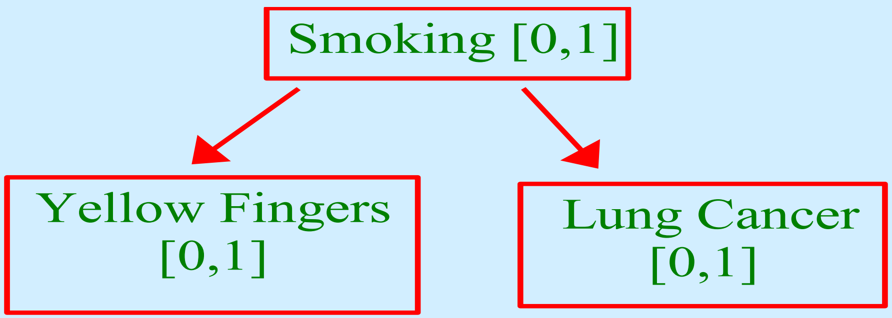
Parametric Models



Instantiated Models



Causal Bayes Networks



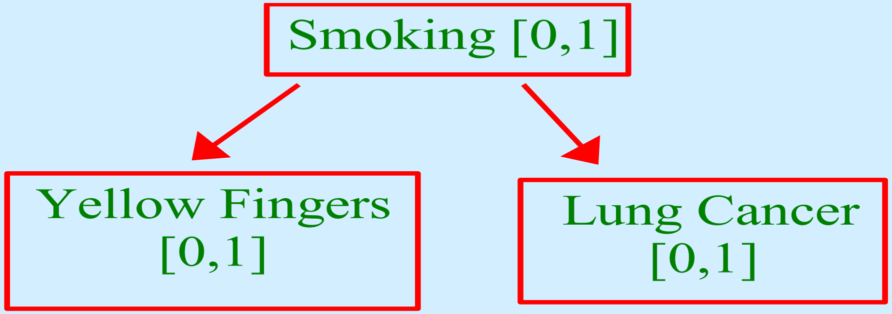
The Joint Distribution Factors

According to the Causal Graph,

$$P(V) = \prod_{x \in V} \mathbf{P}(X \mid Direct\_causes(X))$$

$$P(S,YF,L) = P(S)\ P(YF \mid S)\ P(LC \mid S)$$

Causal Bayes Networks



The Joint Distribution Factors

According to the Causal Graph,

$$P(V) = \prod_{x \in V} \mathbf{P}(X \mid Direct\_causes(X))$$

P(S) P(YF | S) P(LC | S) = f(θ)

All variables binary [0,1]: $\theta = \{\theta_1, \theta_2, \theta_3, \theta_4, \theta_5\}$

$P(S = 0) = \theta_1$

$P(S = 1) = 1 - \theta_1$

$P(YF = 0 \mid S = 0) = \theta_2$

$P(YF = 1 \mid S = 0) = 1 - \theta_2$

$P(YF = 0 \mid S = 1) = \theta_3$

$P(YF = 1 \mid S = 1) = 1 - \theta_3$

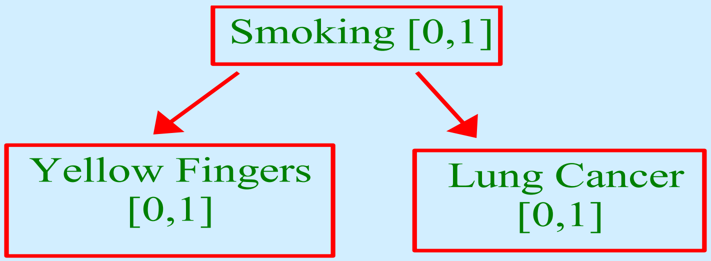
$P(LC = 0 \mid S = 0) = \theta_4$

$P(LC = 1 \mid S = 0) = 1 - \theta_4$

$P(LC = 0 \mid S = 1) = \theta_5$

$P(LC = 1 \mid S = 1) = 1 - \theta_5$

Causal Bayes Networks



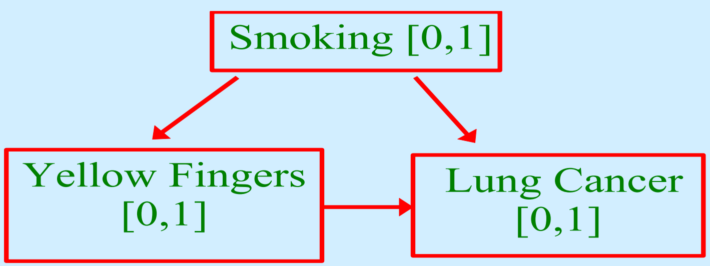
The Joint Distribution Factors

According to the Causal Graph,

$$P(V) = \prod_{x \in V} \mathbf{P}(X \mid Direct\_causes(X))$$

$$P(S,YF, LC) = P(S)\ P(YF \mid S)\ P(LC \mid S) = f(\theta)$$

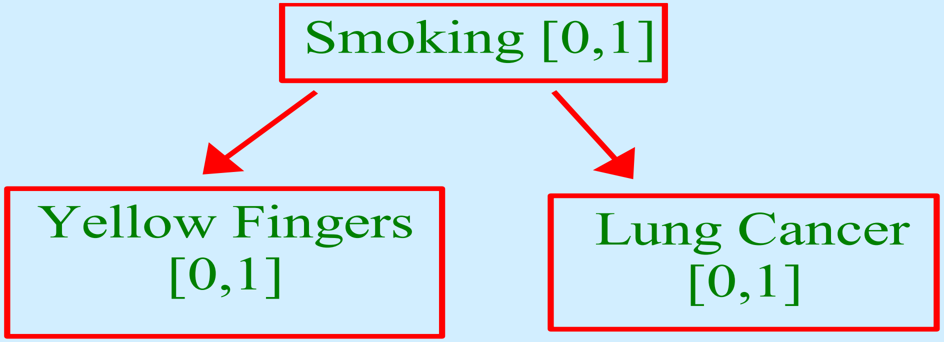
All variables binary [0,1]: $\theta = \{\theta_{1,}\ \theta_{2,}\theta_{3,}\theta_{4,}\theta_{5,}\ \}$



$$P(S,YF, LC) = P(S)\ P(YF \mid S)\ P(LC \mid \boxed{YF,} S) = f(\theta)$$

All variables binary [0,1]: $\theta = \{\theta_{1,}\ \theta_{2,}\theta_{3,}\theta_{4,}\theta_{5,}\ \theta_{6,}\theta_{7,}\ \}$

Causal Bayes Networks



The Joint Distribution Factors

According to the Causal Graph,

$$P(V) = \prod_{x \in V} \mathbf{P}(X \mid Direct\_causes(X))$$

P(S,YF, L) = P(S) P(YF | S) P(LC | 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

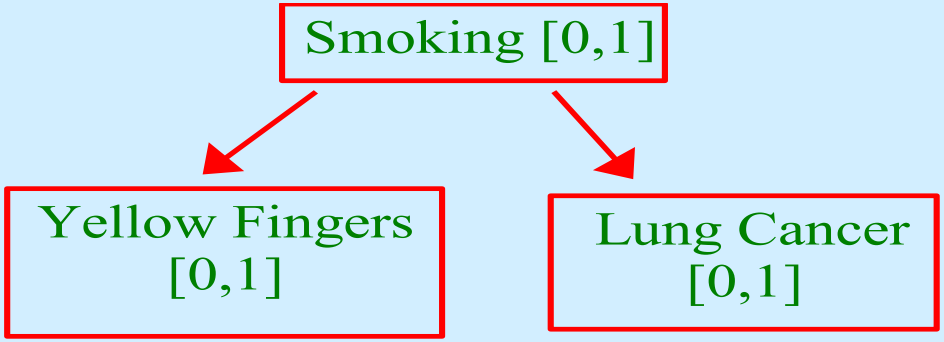
P(LC = 1 | S = 0) = .05

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

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

P(S=1,YF=1, LC=1) = ?

Causal Bayes Networks



The Joint Distribution Factors

According to the Causal Graph,

$$P(V) = \prod_{x \in V} \mathbf{P}(X \mid Direct\_causes(X))$$

P(S,YF, L) = P(S) P(YF | S) P(LC | S)

P(S = 0) = .7

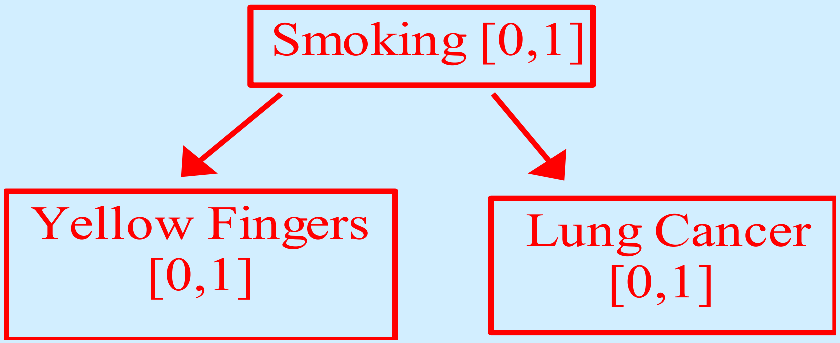
P(S = 1) = .3

P(YF = 0 | S = 0) = .99	*P(LC = 0 | S = 0) = .95*
P(YF = 1 | S = 0) = .01	*P(LC = 1 | S = 0) = .05*
P(YF = 0 | S = 1) = .20	*P(LC = 0 | S = 1) = .80*
P(YF = 1 | S = 1) = .80	*P(LC = 1 | S = 1) = .20*

P(S=1,YF=1, LC=1) = P(S=1) P(YF=1 | S=1) P(LC = 1 | S=1)

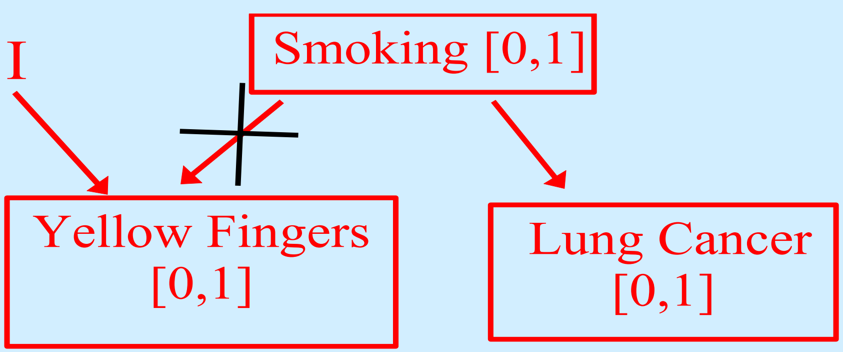
P(S=1,YF=1, LC=1) = .3 * .80 * .20 = .048

Calculating the effect of a hard interventions

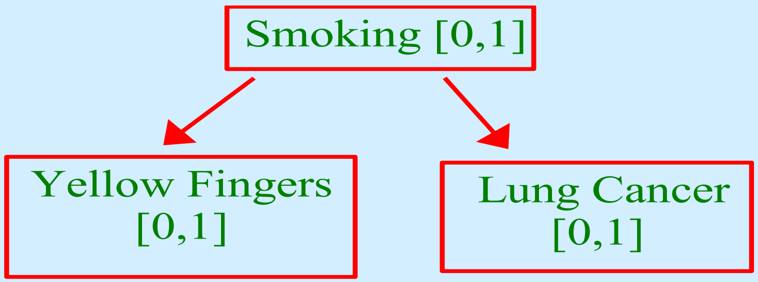


$$P(YF,S,L) = P(S)\,P(YF|S)\,P(L|S)$$

$$P_m(YF,S,L) = P(S)\;P(YF|\,I)\;P(L|S)$$

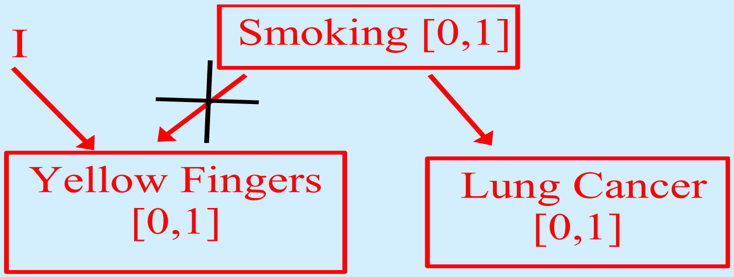


Calculating the effect of a hard intervention



P(S,YF, L) = P(S) P(YF | S) P(LC | S)

P(S=1,YF=1, LC=1) = .3 * .8 * .2 = .048



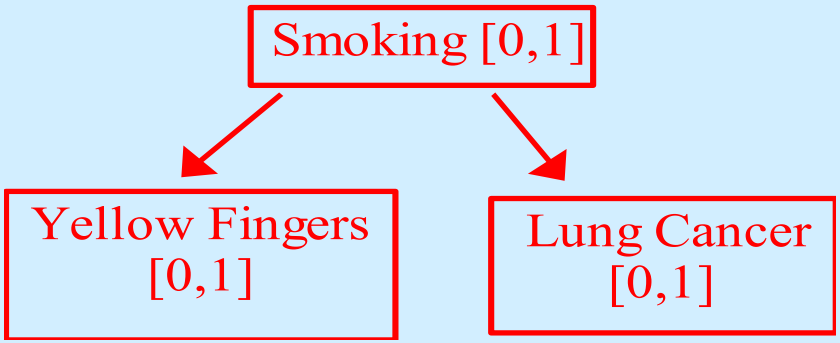
P(YF =1 | I) = .5

$P_m(S=1, YF_{set}=1, LC=1)$ = ?

$P_m(S=1, YF_{set}=1, LC=1)$ = P(S) P(YF | I) P(LC | S)

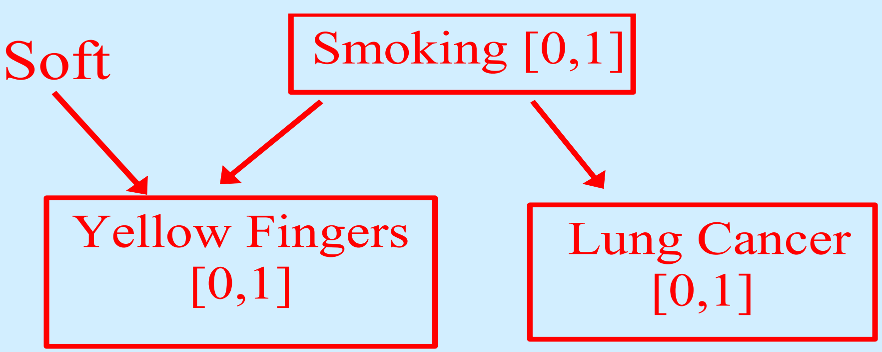
$P_m(S=1, YF_{set}=1, LC=1)$ = .3 * .5 * .2 = .03

Calculating the effect of a soft intervention



$$P(YF,S,L) = P(S)\,P(YF|S)\,P(L|S)$$

$$P_m\,(YF,S,L) = P(S)P(YF|\,S,\,\text{Soft})\,P(L|S)$$



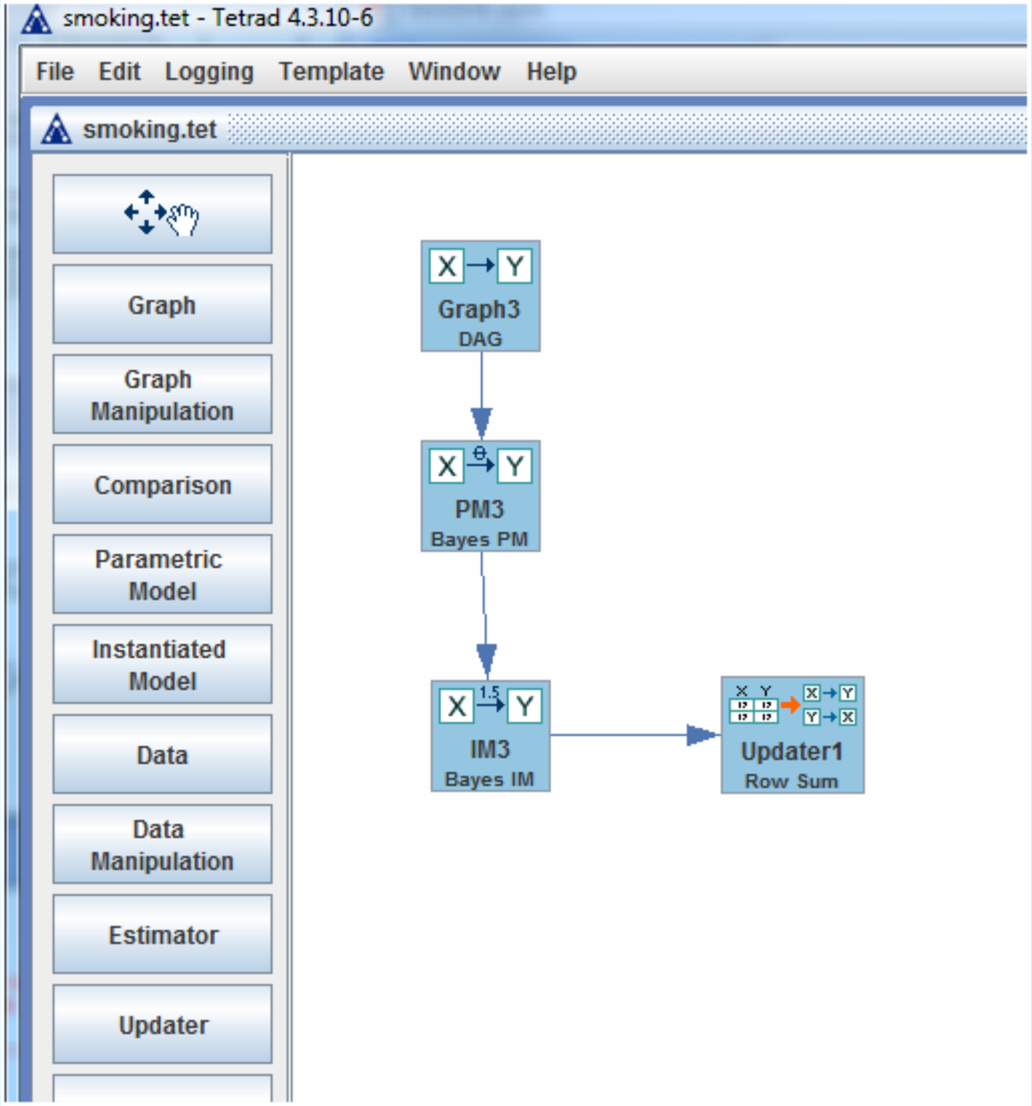
Tetrad Demo & Hands-On

1) Use the DAG you built for Smoking, YF, and LC
2) Define the Bayes PM (# and values of categories for each variable)

3) Attach a Bayes IM to the Bayes PM
4) Fill in the Conditional Probability Tables

   (make the values plausible).

Updating



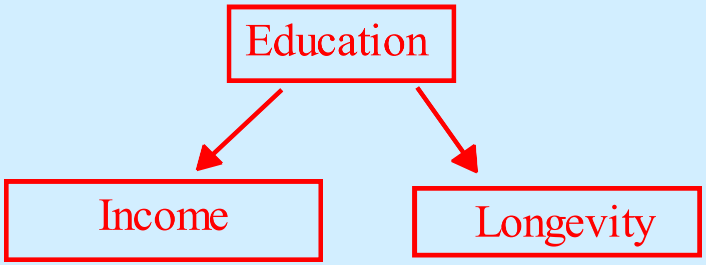
smoking.tet - Tetrad 4.3.10-6
File
Edit
Logging
Template
Window
Help
smoking.tet
Graph
Graph Manipulation
Comparison
Parametric Model
Instantiated Model
Data
Data Manipulation
Estimator
Updater
X
Y
Graph3
DAG
X
θ
Y
PM3
Bayes PM
X
1.5
Y
IM3
Bayes IM
Updater1
Row Sum

Tetrad Demo

1) Use the IM just built of Smoking, YF, LC
2) Update LC on evidence: YF = 1
3) Update LC on evidence: $YF_{set} = 1$

Structural Equation Models

Causal Graph



- Structural Equations

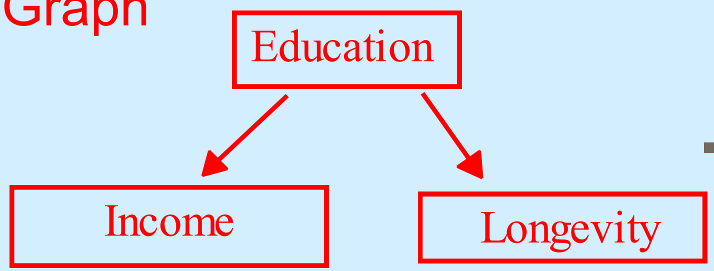
  For each variable $X \in \mathbf{V}$, an *assignment* equation:

  $$X := f_X(\text{immediate-causes}(X), \varepsilon_X)$$

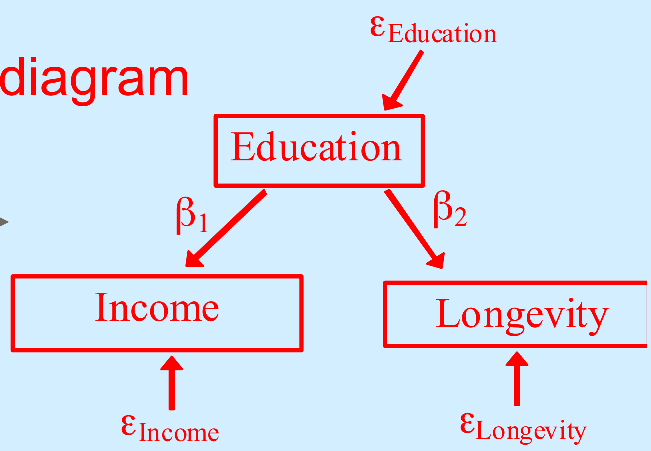
- Exogenous Distribution: Joint distribution over the exogenous vars : $P(\varepsilon)$

Linear Structural Equation Models

Causal Graph



Path diagram



Equations:

Education := $\varepsilon_{\text{Education}}$

Income := β_1 Education + $\varepsilon_{\text{income}}$

Longevity := β_2 Education + $\varepsilon_{\text{Longevity}}$

Structural Equation Model:

V = B**V** + **E**

Exogenous Distribution:

$P(\varepsilon_{ed}, \varepsilon_{Income}, \varepsilon_{Income})$

- $\forall i \neq j \; \varepsilon_i \perp \varepsilon_j$ (pairwise independence)
- no variance is zero

E.g.

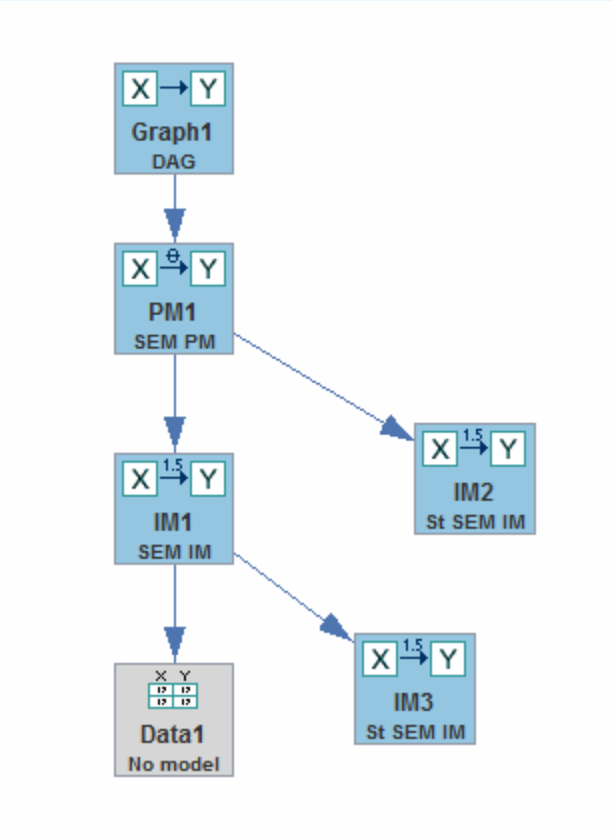
$(\varepsilon_{ed}, \varepsilon_{Income}, \varepsilon_{Income}) \sim N(0, \Sigma^2)$

- Σ^2 diagonal,
- no variance is zero

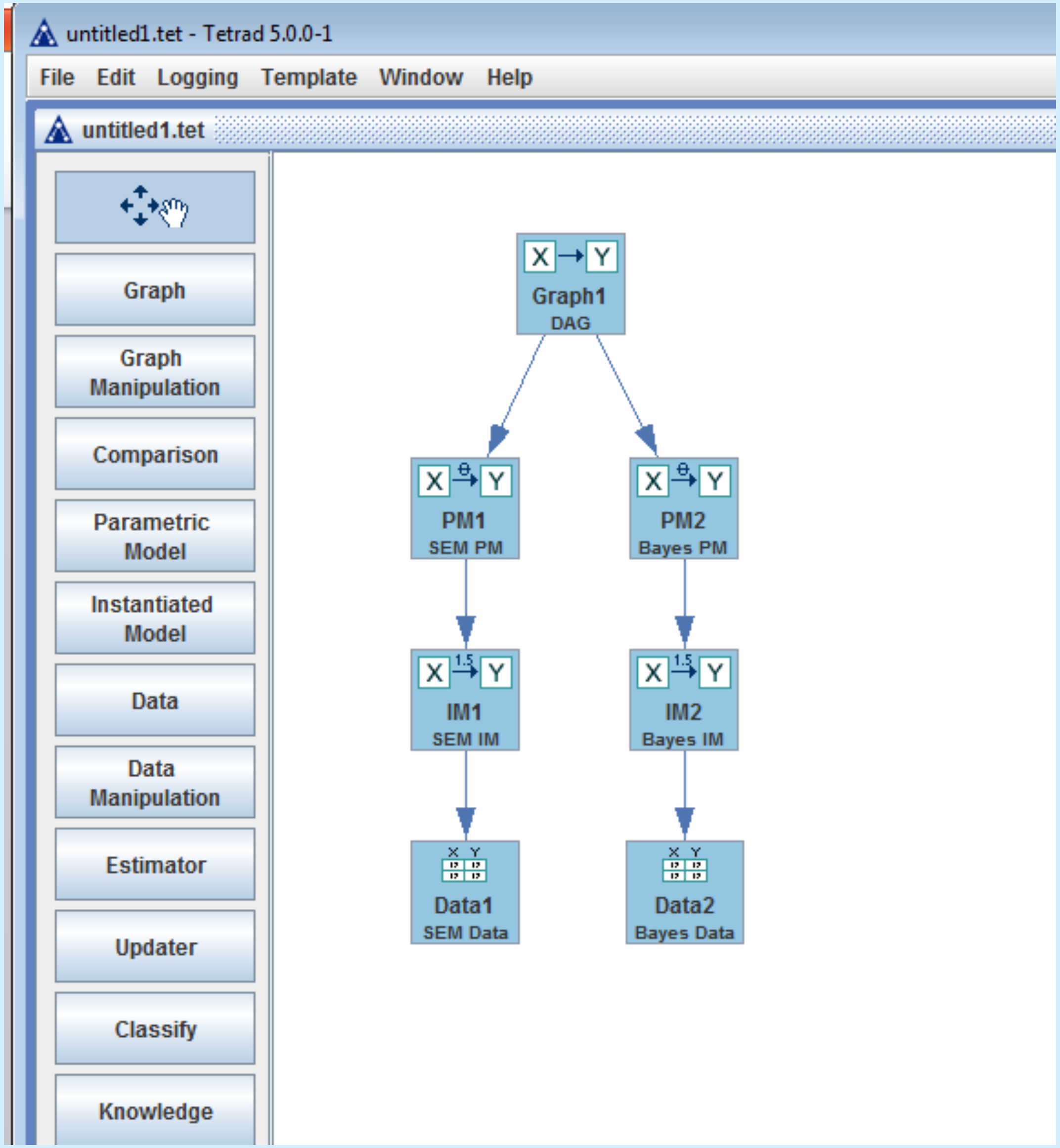
Tetrad Demo & Hands-On

1) Attach a SEM PM to your 3-4 variable graph
2) Attach a SEM IM to the SEM PM
3) Change the coefficient values.

4) Attach a Standardized SEM IM to the SEM PM, or the SEM IM



Simulated Data



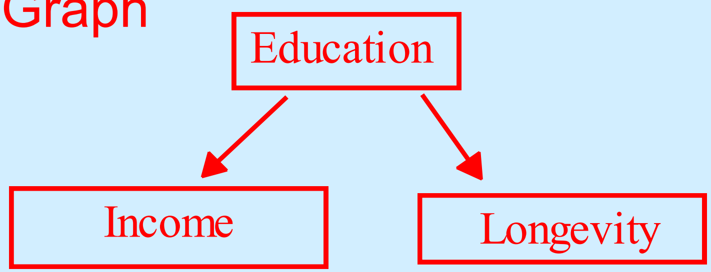
Tetrad Demo & Hands-On

1) Simulate Data from both your SEM IM and your Bayes IM

Generalized SEM

1) The Generalized SEM is a generalization of the linear SEM model.
2) Allows for arbitrary connection functions
3) Allows for arbitrary distributions
4) Simulation from cyclic models supported.

Causal Graph



SEM Equations:

$\text{Education} := \varepsilon_{\text{Education}}$

$\text{Income} := \beta_1 \text{ Education} + \varepsilon_{\text{income}}$

$\text{Longevity} := \beta_2 \text{ Education} + \varepsilon_{\text{Longevity}}$

$P(\varepsilon_{\text{ed}}, \varepsilon_{\text{Income}}, \varepsilon_{\text{Income}}) \sim N(0, \Sigma^2)$

Generalized SEM Equations:

$\text{Education} := \varepsilon_{\text{Education}}$

$\text{Income} := \beta_1 \text{ Education}^2 + \varepsilon_{\text{income}}$

$\text{Longevity} := \beta_2 \text{ In(Education)} + \varepsilon_{\text{Longevity}}$

$P(\varepsilon_{\text{ed}}, \varepsilon_{\text{Income}}, \varepsilon_{\text{Income}}) \sim U(0,1)$

Hands On

1) Create a DAG.
2) Parameterize it as a Generalized SEM.
3) In PM – select from Tools menu “show error terms”

   Click on error term, change its distribution to Uniform
4) Make at least one function non-linear
5) Make at least one function interactive
6) Save the session as “generalizedSEM”.