Tetrad

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

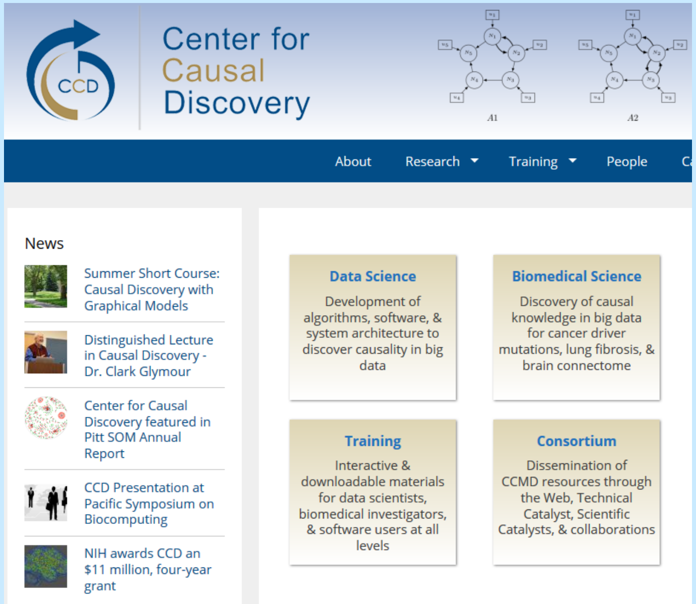
2) Download: http://www.phil.cmu.edu/projects/tetrad/current.html
   a) JNLP version: *Tetrad 5.3.0*
   b) Jar file: *Tetrad 5.3.0 (6/13/2016 Version 1)*

3) Data files:

   www.phil.cmu.edu/projects/tetrad_download/download/workshops/CCD/2016/Datasets/

Center for Causal Discovery:

Summer Short Course/Datathon - 2016



News

Summer Short Course: Causal Discovery with Graphical Models

Distinguished Lecture in Causal Discovery - Dr. Clark Glymour

Center for Causal Discovery featured in Pitt SOM Annual Report

CCD Presentation at Pacific Symposium on Biocomputing

NIH awards CCD an $11 million, four-year grant

Data Science
Development of algorithms, software, & system architecture to discover causality in big data

Biomedical Science
Discovery of causal knowledge in big data for cancer driver mutations, lung fibrosis, & brain connectome

Training
Interactive & downloadable materials for data scientists, biomedical investigators, & software users at all levels

Consortium
Dissemination of CCMD resources through the Web, Technical Catalyst, Scientific Catalysts, & collaborations

June 13-18, 2015

Carnegie Mellon University

Goals

1) Basic working knowledge of graphical causal models
2) Basic working knowledge of Tetrad V
3) Basic understanding of search algorithms
4) “Fully started” on using CCD algorithms/tools on real data, preferably your own.
5) Provide us with useful feedback on:
    1) The intro to graphical models/search with Tetrad segments
    2) The breakout sessions
    3) Follow up after the workshop: integrating CCD tools into your own research
6) Form community of researchers, users, and students interested in causal discovery in biomedical research

Monday: Basics of Graphical Causal Models, Tetrad

Morning: 9 AM – Noon, Baker Hall A51 : Giant Eagle Auditorium

1. Introduction
2. Representing/Modeling Causal Systems
    a) Causal Graphs/Interventions
    b) Parametric Models
    c) Instantiated Models

Afternoon: 1:30 PM – 4 PM, Baker Hall A51 : Giant Eagle Auditorium

1. Estimation, Inference, and Model fit
2. Case Study: Charitable Giving

Dinner: On your own

Tuesday: Basics of Search, Break-out Sessions

Morning: 9 AM – Noon, Baker Hall A51 : Giant Eagle Auditorium

1. D-separation & Model Equivalence
2. Searching for Causal Systems

Afternoon: 1:30 PM – 4 PM, Baker Hall A51 → breakout rooms

1. Break-out Session 1:
   - A. Brain/fMRI
   - B. Cancer
   - C. Lung Disease

Dinner: On your own

Wednesday: Latent Variables, etc., Break-out Sessions

Morning: 9 AM – Noon, Baker Hall A51 : Giant Eagle Auditorium

1. Latent Variable Model Search
2. Measurement

Afternoon: 1:30 PM – 3:30 PM, Baker Hall A51 → breakout rooms

1. Break-out Session 2

Evening: O'Hara Student Center (Pitt), 2nd Floor Ballroom

1. *5:30 – 6:15 Poster Session*
2. *6:15 – 8:00 Dinner (keynote speaker: Greg Cooper)*

Thursday: Research Area Overviews, Break-out Sessions

Morning: 9 AM – Noon, Baker Hall A51 : Giant Eagle Auditorium

1. fMRI – Brain
2. Cancer: Genomic Drivers
3. Lung Disease Pathways
4. Genetic Regulatory Network Search

Afternoon: 1:30 PM – 4 PM, Baker Hall A51 → breakout rooms

1. Break-out Sessions 3

Dinner: On your own

Friday: Wrap-up, DataThon

Morning: 9 AM – Noon, Baker Hall A51 : Giant Eagle Auditorium

1. Break-out Group Reports
2. General Debrief Q&A
3. Evaluations

Afternoon: 1:30 PM – 4 PM, Giant Eagle Auditorium: Datathon

1:00 Intro

1:30 Team Introductions

2:00 Data Prep

3:00 Supercomputing Resources

3:30 – 6:00 Data Analysis

Dinner: 6-8 PM: Pizza

Saturday: DataThon

Morning: 9 AM – Noon, Baker Hall A51 : Giant Eagle Auditorium

1. 9 AM: Breakfast and Q&A
2. 10AM – Noon: Data hacking

Noon – 1 PM: Lunch: on your own

Afternoon: 1:00 PM – 3 PM, Giant Eagle Auditorium

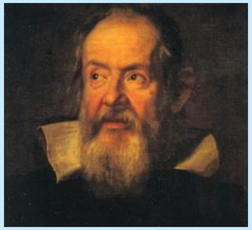
1:00 – 3:00: Data Hacking

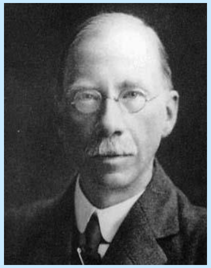
3:00: Participant Presentations

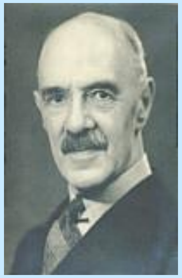
Questions?

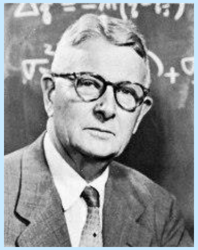
Causation and Statistics


Francis Bacon


Galileo Galilei


Udny Yule


Charles Spearman


Sewall Wright


Sir Ronald A. Fisher


Jerzy Neyman


Jamie Robins

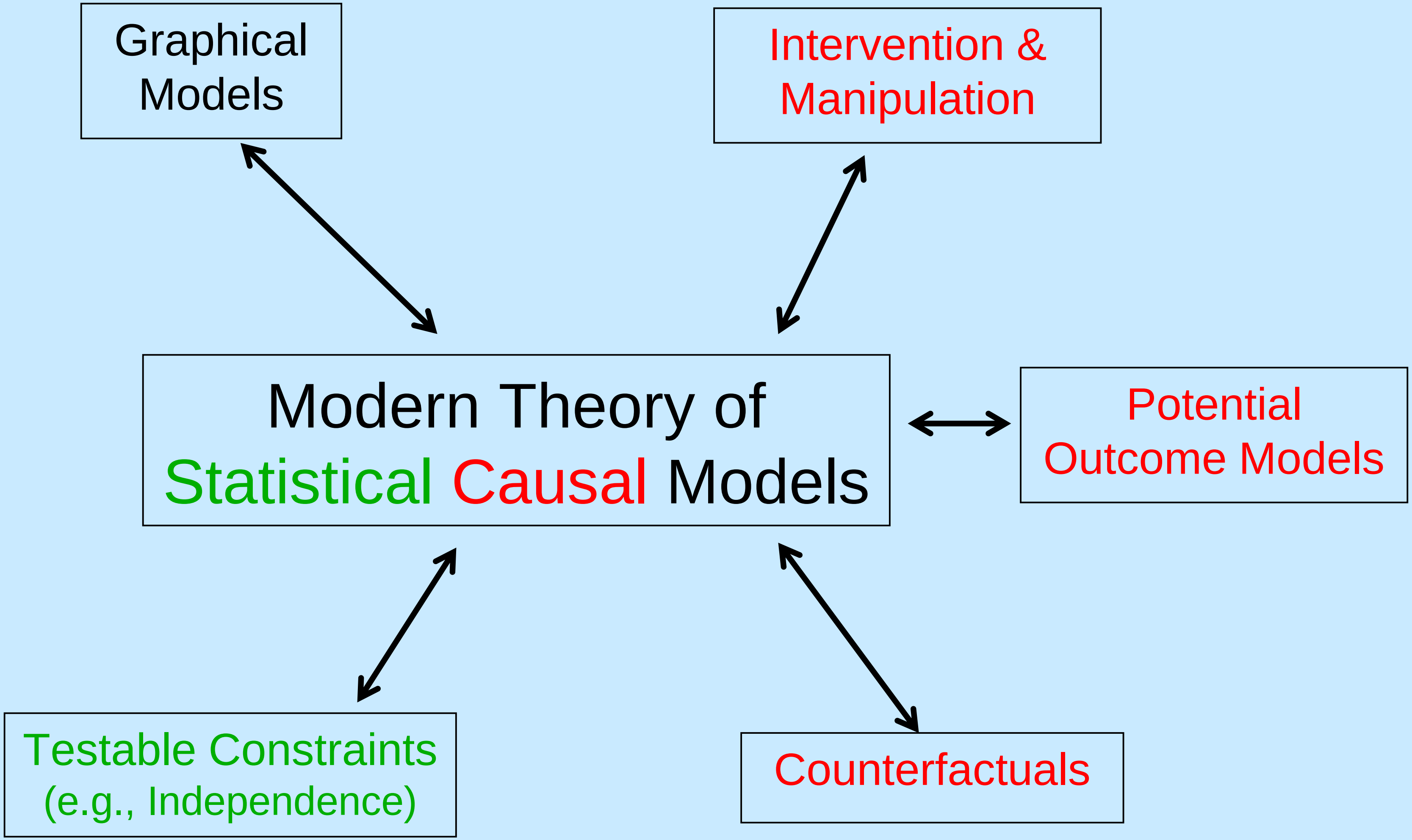

Don Rubin


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

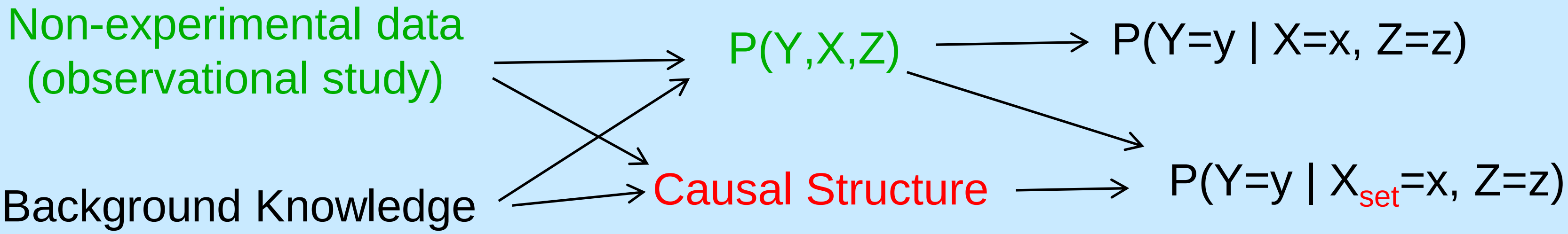
P(**Lung Cancer 1960 = y | Tar-stained fingers 1950 = no**)

≠

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

In general: P(**Y=y |** X=x**, Z=z**) ≠ P(**Y=y | X**$_{set}$=x, **Z=z**)

Causal Prediction vs. Statistical Prediction:



Non-experimental data (observational study)

Background Knowledge

P(Y,X,Z)

Causal Structure

P(Y=y | X=x, Z=z)

P(Y=y | X$_{set}$=x, Z=z)

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

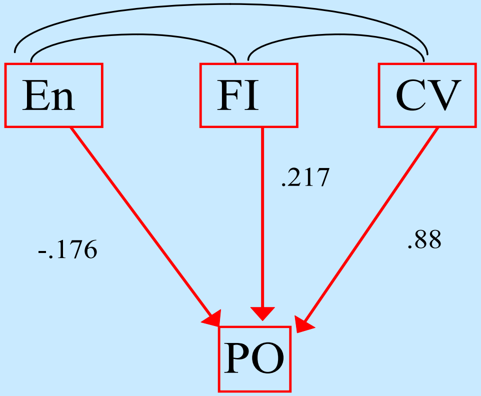
Regression Results

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

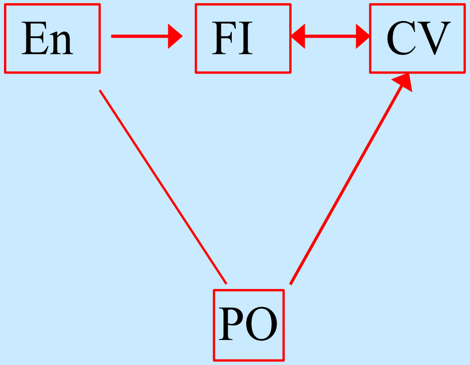
SE	(.058)	(.059)	(.060)
t	3.941	-2.99	14.6
P	$\boxed{.0002}$	.0044	.0000

Interpretation: foreign investment increases political repression

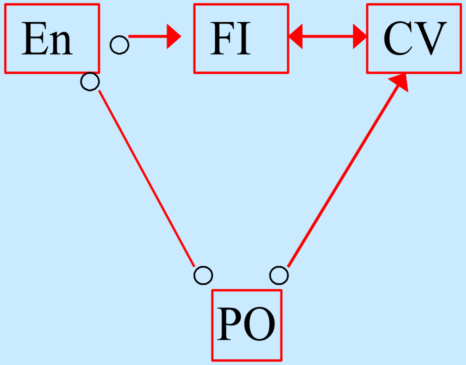
Case Study: Foreign Investment *Alternative Models*



Regression

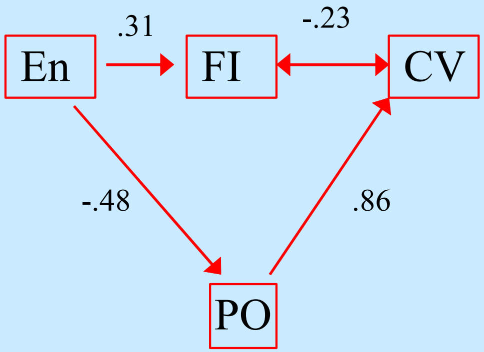


Tetrad - PC



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, $\chi2=0.12$,
p-value = .94

Outline

Representing/Modeling Causal Systems

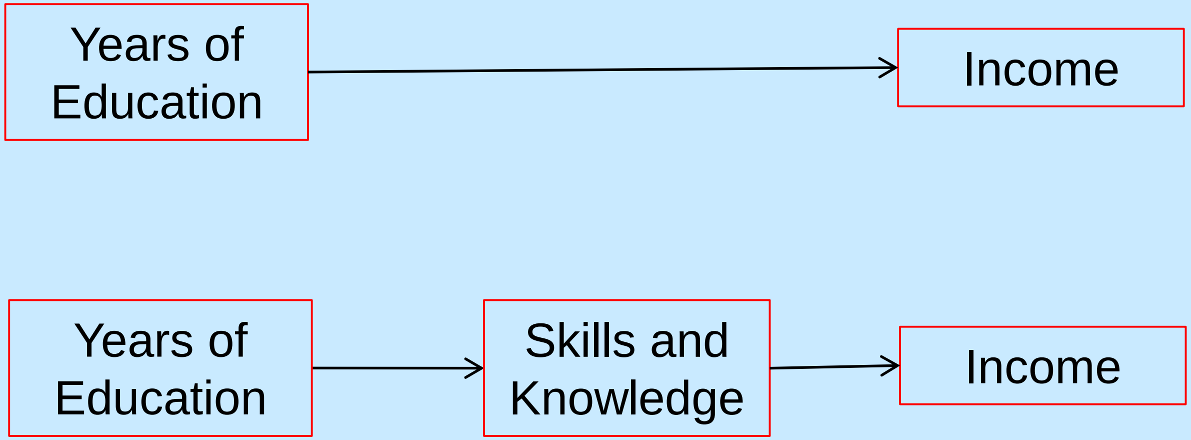
1) Causal Graphs
2) Parametric Models
    a) Bayes Nets
    b) Structural Equation Models
    c) Generalized SEMs

Causal Graphs

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

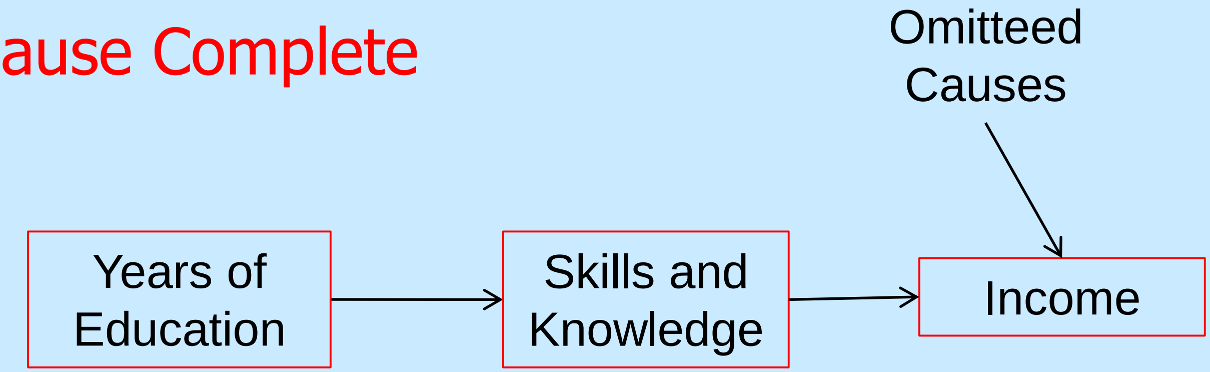
Each edge X → Y represents a direct causal claim:

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



Causal Graphs

Not Cause Complete



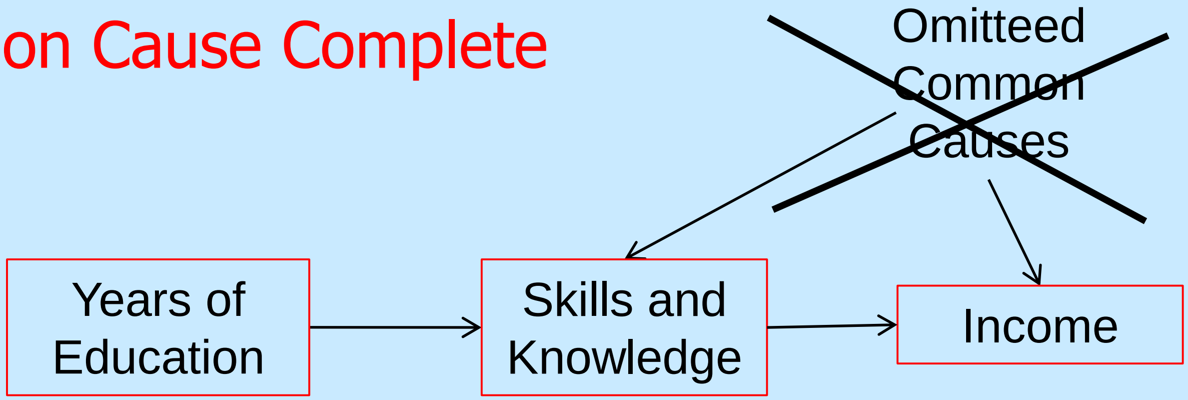
Omitteed
Causes

Years of
Education

Skills and
Knowledge

Income

Common Cause Complete



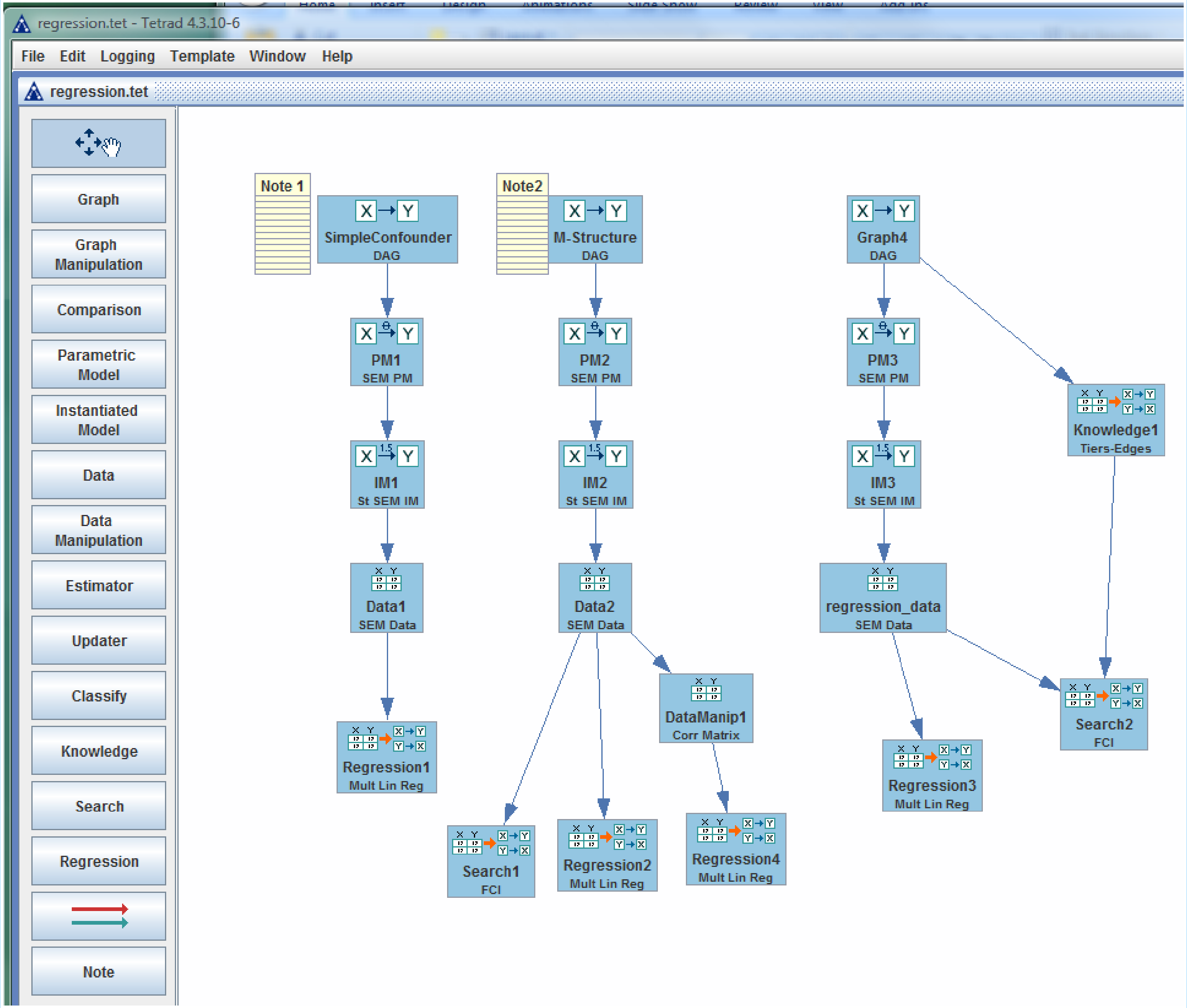
~~Omitteed
Common
Causes~~

Years of
Education

Skills and
Knowledge

Income

Tetrad: Complete Causal Modeling Tool



Tetrad

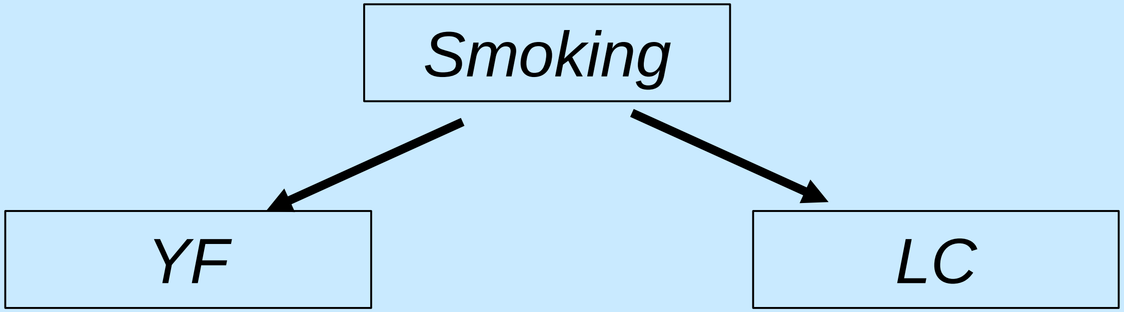
1) Main website: http://www.phil.cmu.edu/projects/tetrad/

2) Download: http://www.phil.cmu.edu/projects/tetrad/current.html
    a) JNLP version: *Tetrad 5.3.0*
    b) Jar file: *Tetrad 5.3.0* *(6/10/2016 Version 1)*

3) Data files:

   www.phil.cmu.edu/projects/tetrad_download/download/workshops/CCD/2016/Datasets/

Tetrad Demo & Hands-On



Smoking

YF

LC

Build and Save two acyclic causal graphs:

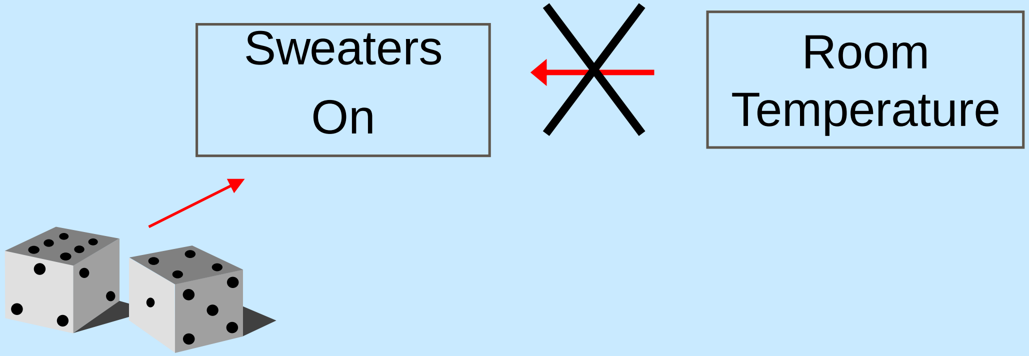
1) Build the Smoking graph picture above

2) Build your own graph with 4 variables

Modeling Ideal Interventions

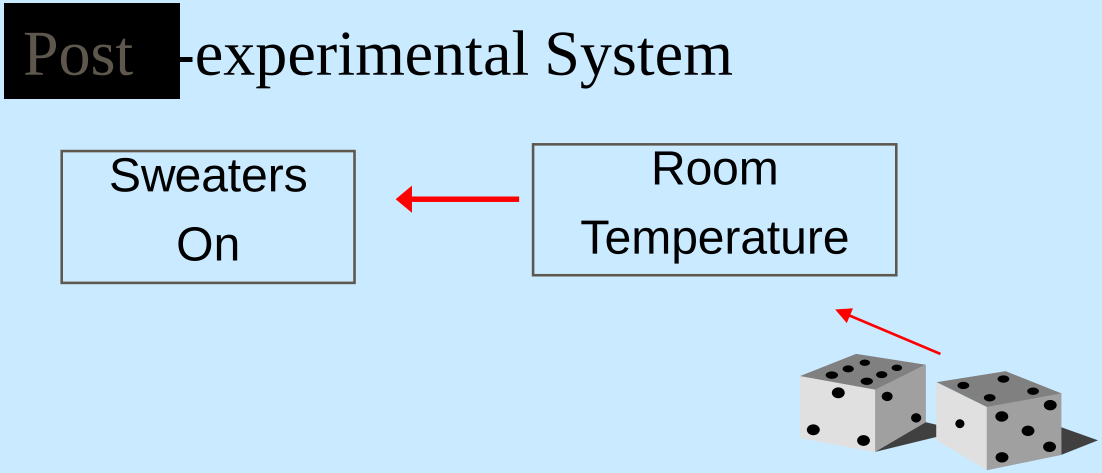
Interventions on the Effect

Post experimental System



Modeling Ideal Interventions

Interventions on the Cause



Post-experimental System

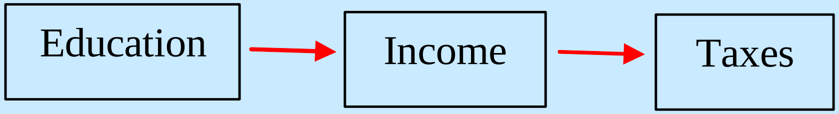
Sweaters On

Room Temperature

Interventions & Causal Graphs

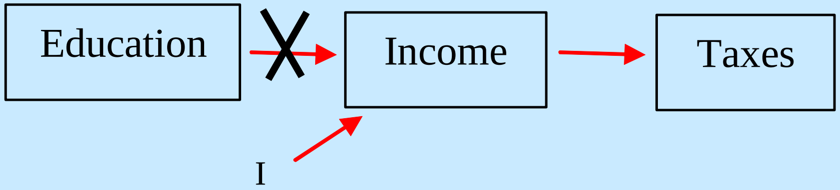
Model an ideal intervention by adding an “intervention” variable outside the original system as a direct cause of its target.

Pre-intervention graph

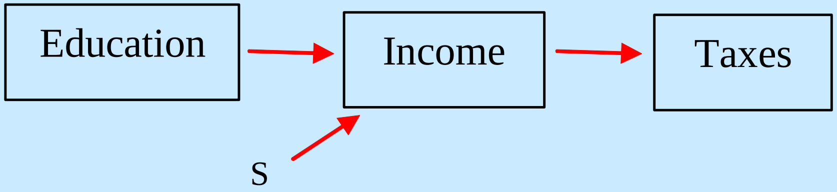


Intervene on *Income*

“Hard” Intervention



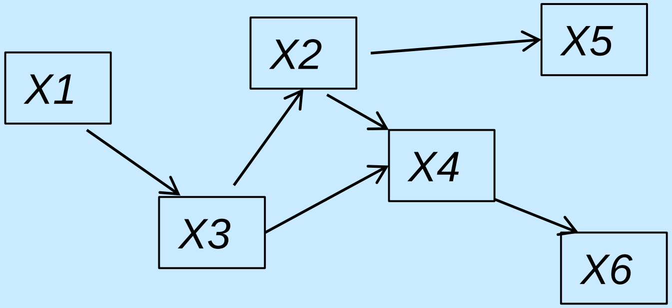
“Soft” Intervention



Interventions & Causal Graphs

Pre-intervention

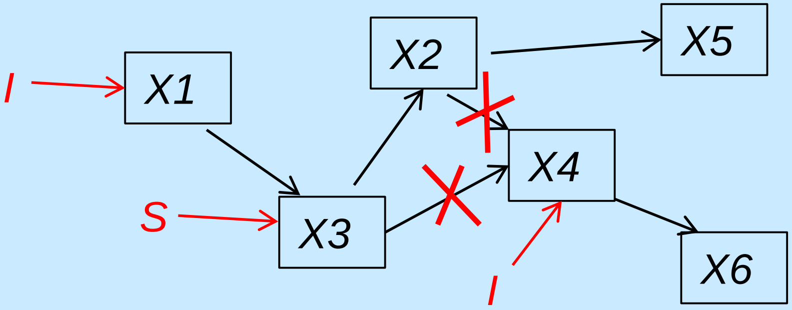
Graph



Intervention:

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

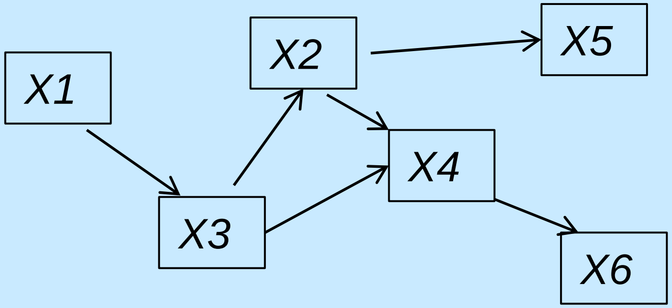
Post-Intervention
Graph?



Interventions & Causal Graphs

Pre-intervention

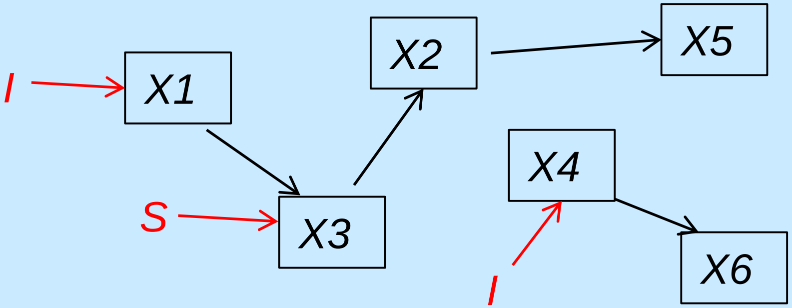
Graph



Intervention:

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

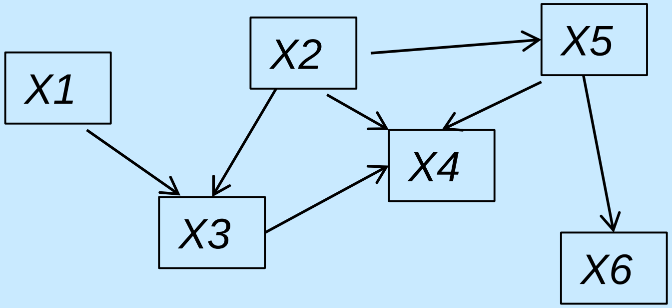
Post-Intervention
Graph?



Interventions & Causal Graphs

Pre-intervention

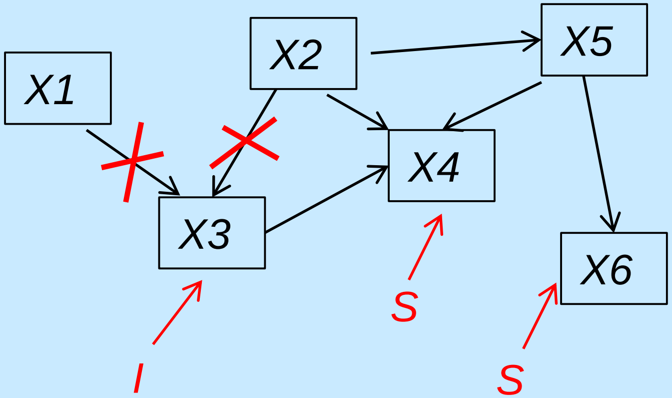
Graph



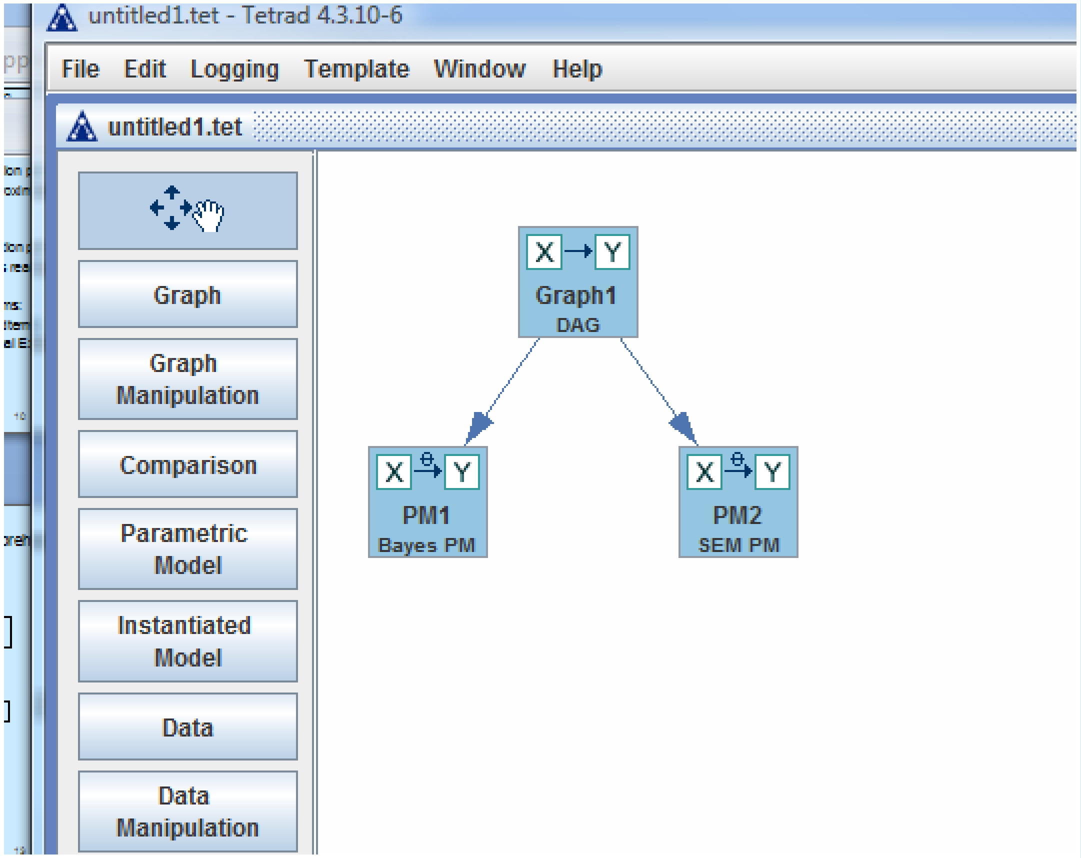
Intervention:

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

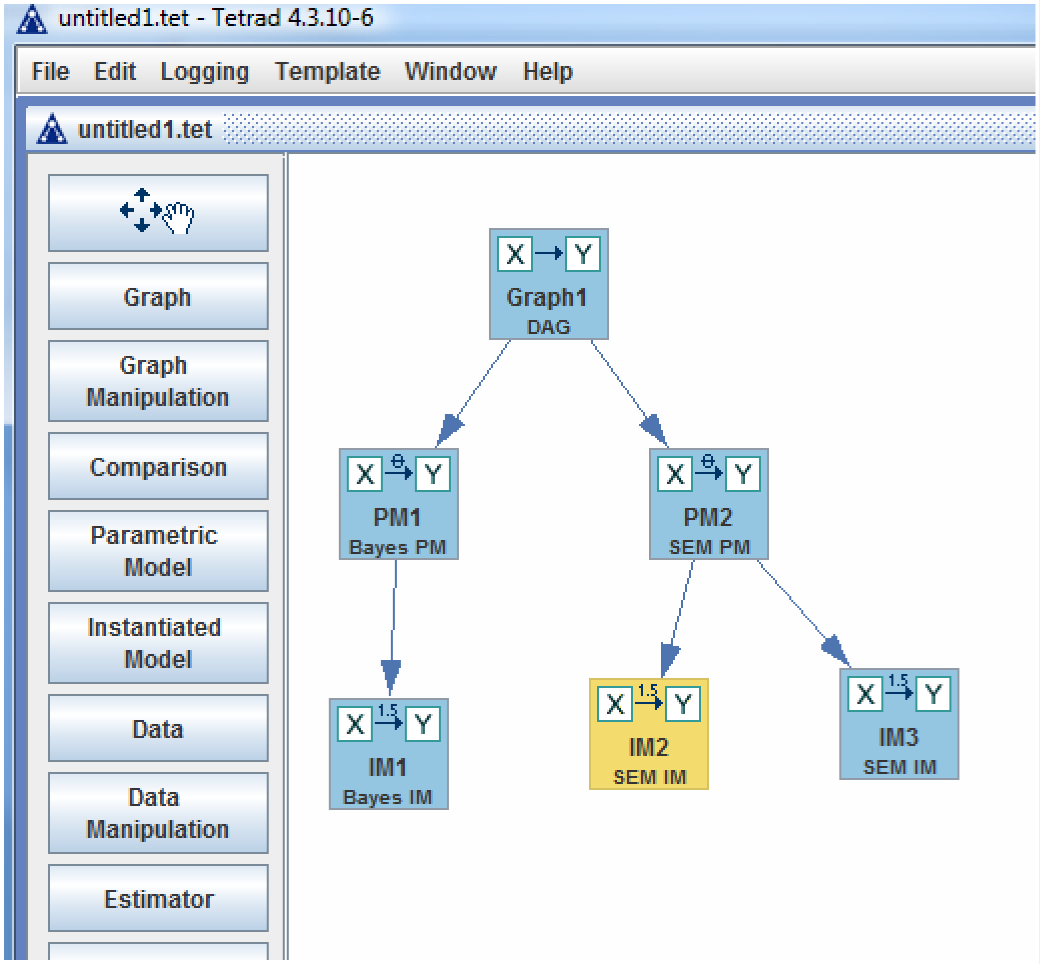
Post-Intervention
Graph?



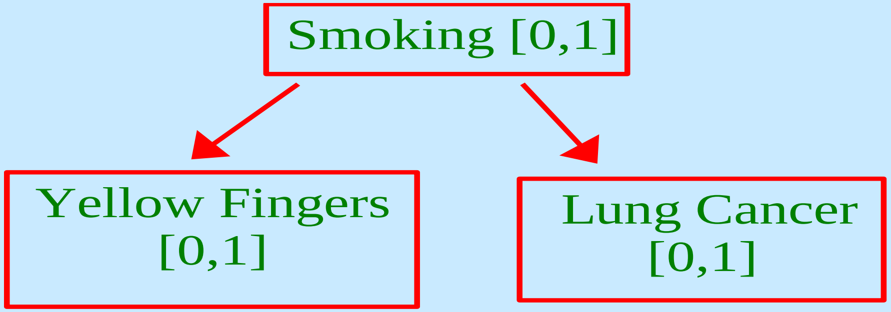
Parametric Models



Instantiated Models



Causal Bayes Networks



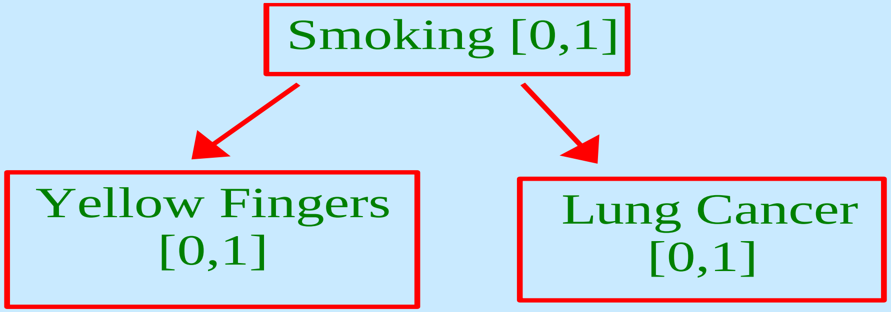
The Joint Distribution Factors

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 | S) P(LC | S)

Causal Bayes Networks



The Joint Distribution Factors

According to the Causal Graph,

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

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

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

$P(S = 0) = \theta_1$

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

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

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

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

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

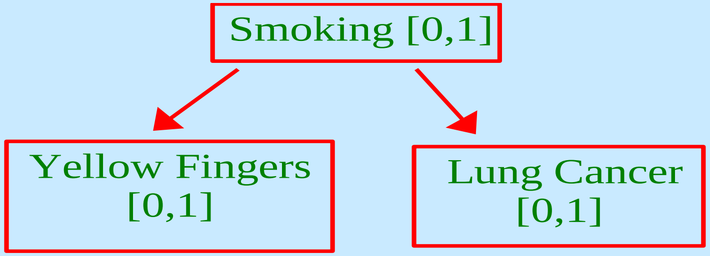
$P(LC = 0 \mid S = 0) = \theta_4$

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

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

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

Causal Bayes Networks



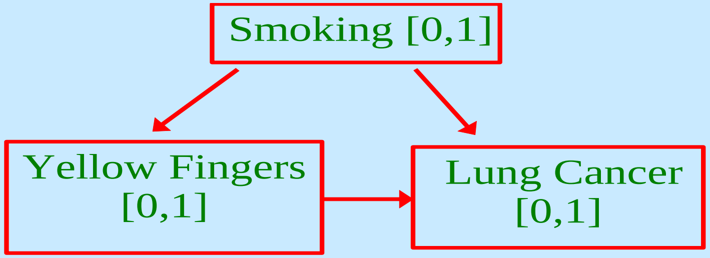
The Joint Distribution Factors

According to the Causal Graph,

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

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

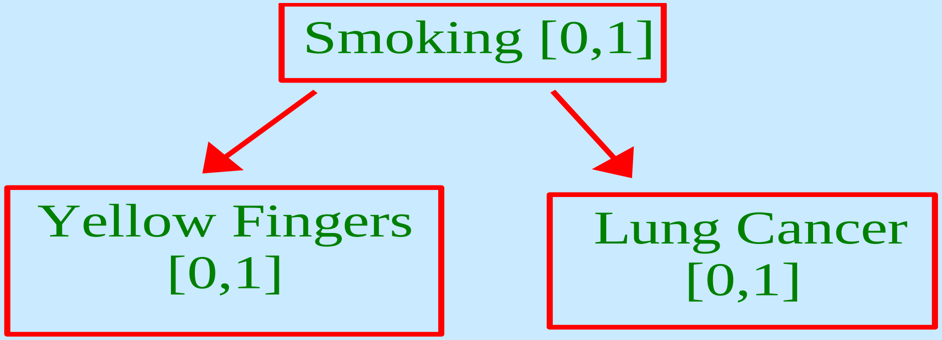
All variables binary [0,1]: $\boldsymbol{\theta} = \{\theta_{1,}\ \theta_{2,}\theta_{3,}\theta_{4,}\theta_{5,}\ \}$



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

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

Causal Bayes Networks


Smoking [0,1]

Yellow Fingers [0,1]

Lung Cancer [0,1]

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)

$P(S = 0) = .7$

$P(S = 1) = .3$

$P(YF = 0 \mid S = 0) = .99$

$P(YF = 1 \mid S = 0) = .01$

$P(YF = 0 \mid S = 1) = .20$

$P(YF = 1 \mid S = 1) = .80$

$P(LC = 0 \mid S = 0) = .95$

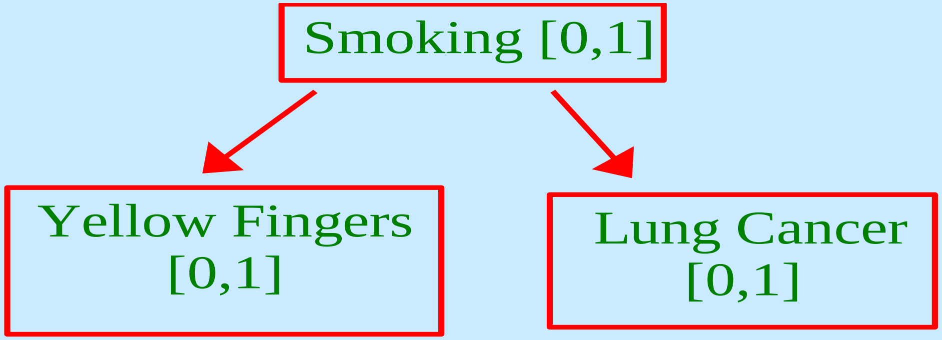
$P(LC = 1 \mid S = 0) = .05$

$P(LC = 0 \mid S = 1) = .80$

$P(LC = 1 \mid S = 1) = .20$

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)

$P(S = 0) = .7$

$P(S = 1) = .3$

$P(YF = 0 \mid S = 0) = .99$

$P(YF = 1 \mid S = 0) = .01$

$P(YF = 0 \mid S = 1) = .20$

$P(YF = 1 \mid S = 1) = .80$

$P(LC = 0 \mid S = 0) = .95$

$P(LC = 1 \mid S = 0) = .05$

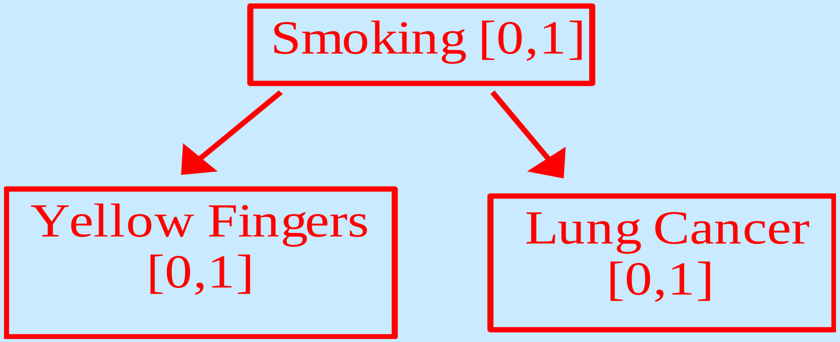
$P(LC = 0 \mid S = 1) = .80$

$P(LC = 1 \mid S = 1) = .20$

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

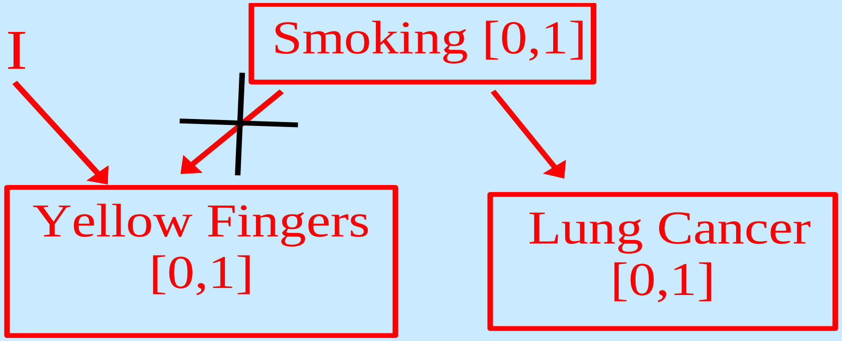
P(S=1,YF=1, LC=1) = .3 * .80 * .20 = .048

Calculating the effect of a hard interventions

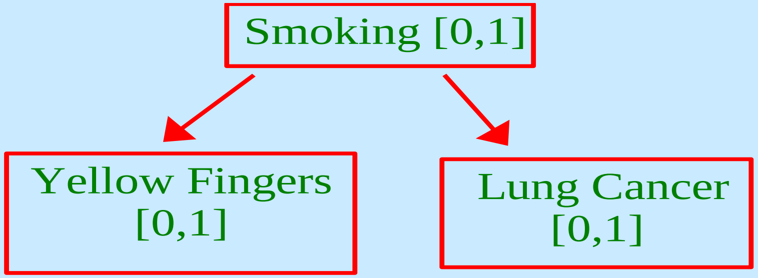


$$P(YF,S,L) = P(S)\ P(YF|S)\ P(L|S)$$

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



Calculating the effect of a hard intervention



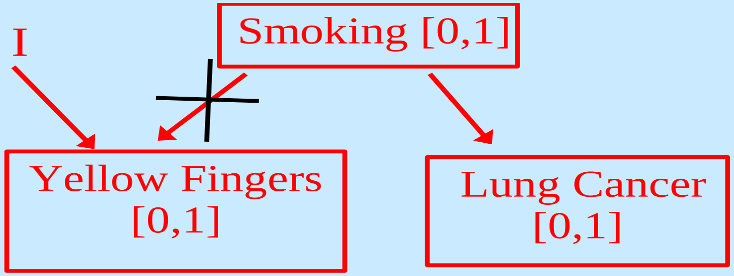
Smoking [0,1]

Yellow Fingers [0,1]

Lung Cancer [0,1]

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

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



I

Smoking [0,1]

Yellow Fingers [0,1]

Lung Cancer [0,1]

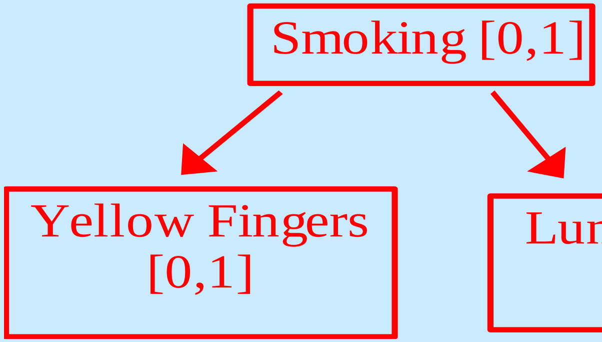
$P(YF = 1 \mid I) = .5$

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

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

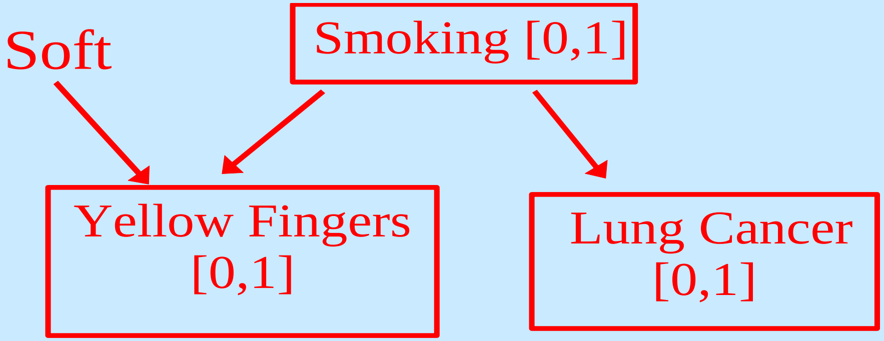
$P_m(S=1,YF_{set}=1, LC=1) = .3 * .5 * .2 = .03$

Calculating the effect of a soft intervention



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

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



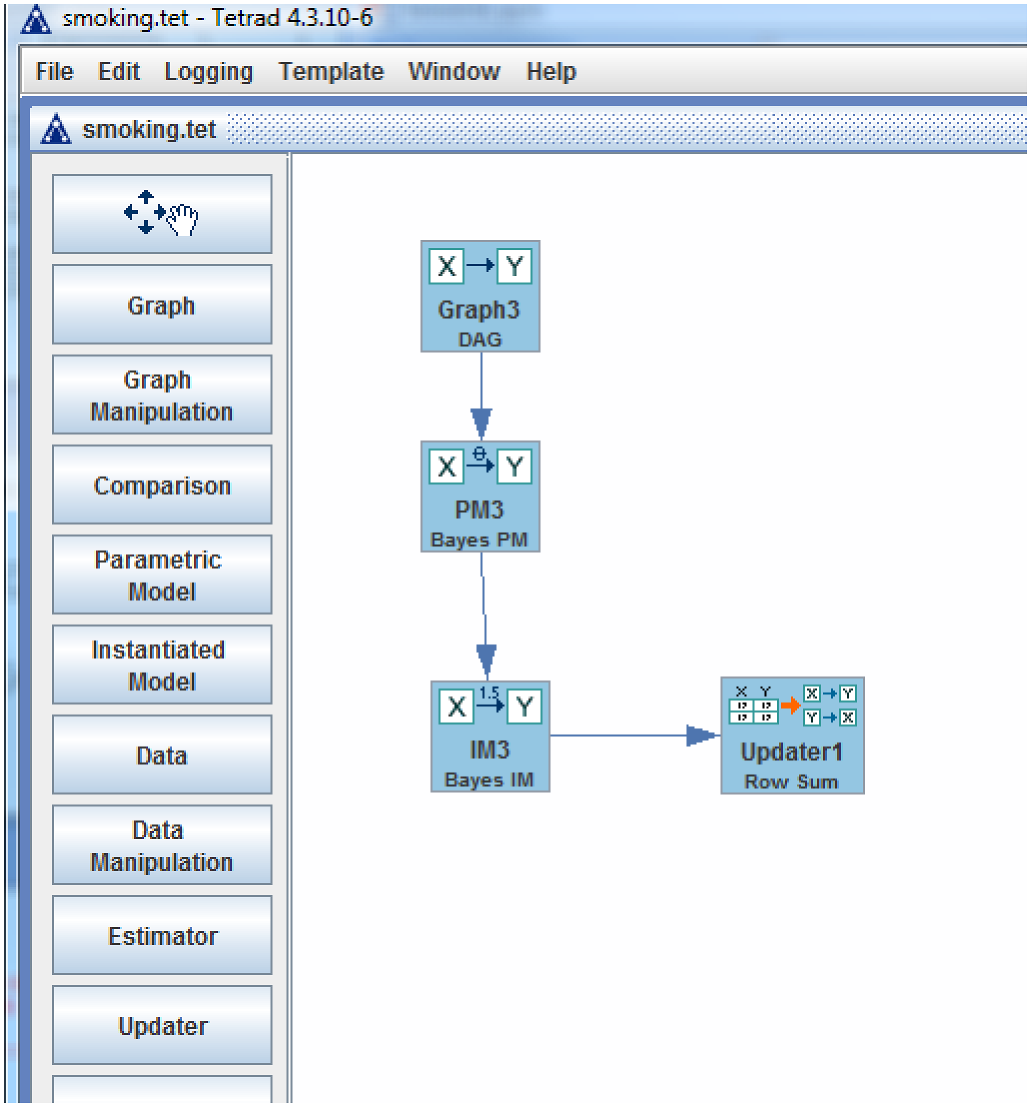
Tetrad Demo & Hands-On

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

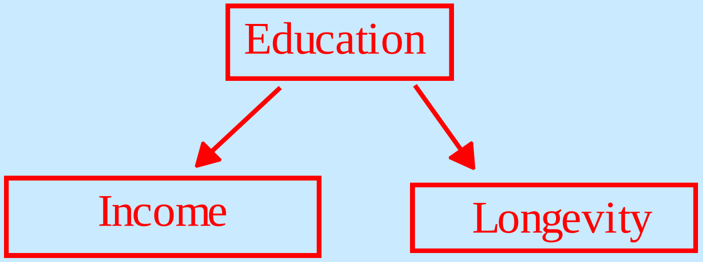

smoking.tet - Tetrad 4.3.10-6
File
Edit
Logging
Template
Window
Help
smoking.tet
Graph
Graph Manipulation
Comparison
Parametric Model
Instantiated Model
Data
Data Manipulation
Estimator
Updater
X
Y
Graph3
DAG
θ
PM3
Bayes PM
1.5
IM3
Bayes IM
Updater1
Row Sum

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

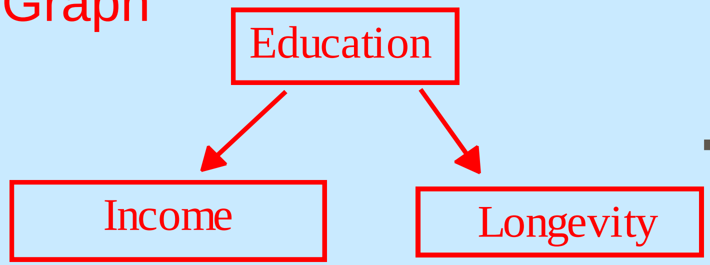
  For each variable X $\in$ **V**, an *assignment* equation:

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

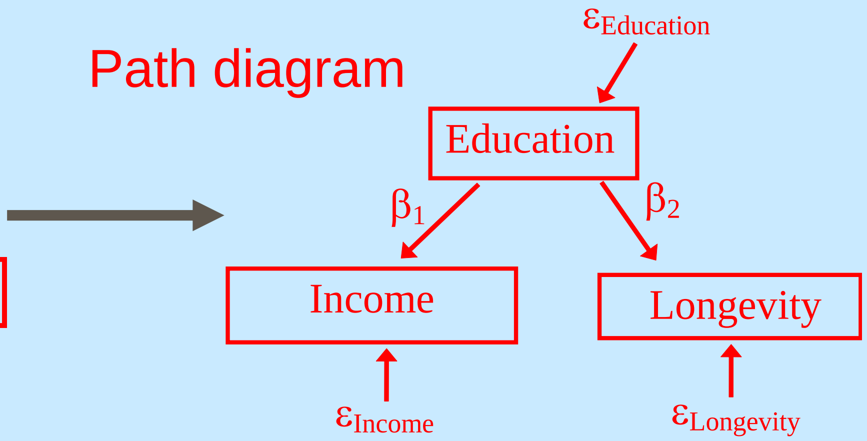
- Exogenous Distribution: Joint distribution over the exogenous vars : $P(\varepsilon)$

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(ε_{ed}, $\varepsilon_{\text{Income}}$,$\varepsilon_{\text{Income}}$)

- $\forall i \neq j \; \varepsilon_i \perp \varepsilon_j$ (pairwise independence)

- no variance is zero

E.g.

(ε_{ed}, $\varepsilon_{\text{Income}}$,$\varepsilon_{\text{Income}}$) ~N(0,Σ^2)

– Σ^2 diagonal,

- no variance is zero

Extra Slides

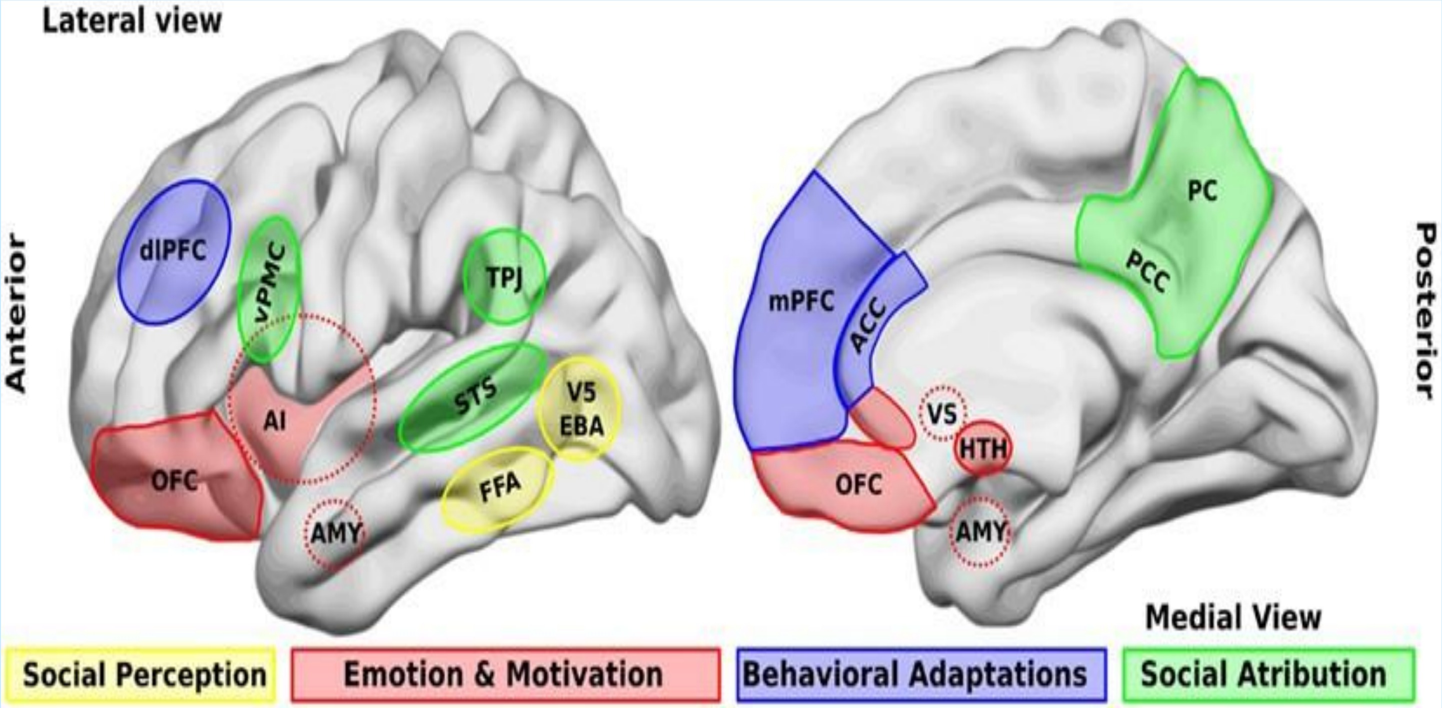
A Few Causal Discovery Highlights

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

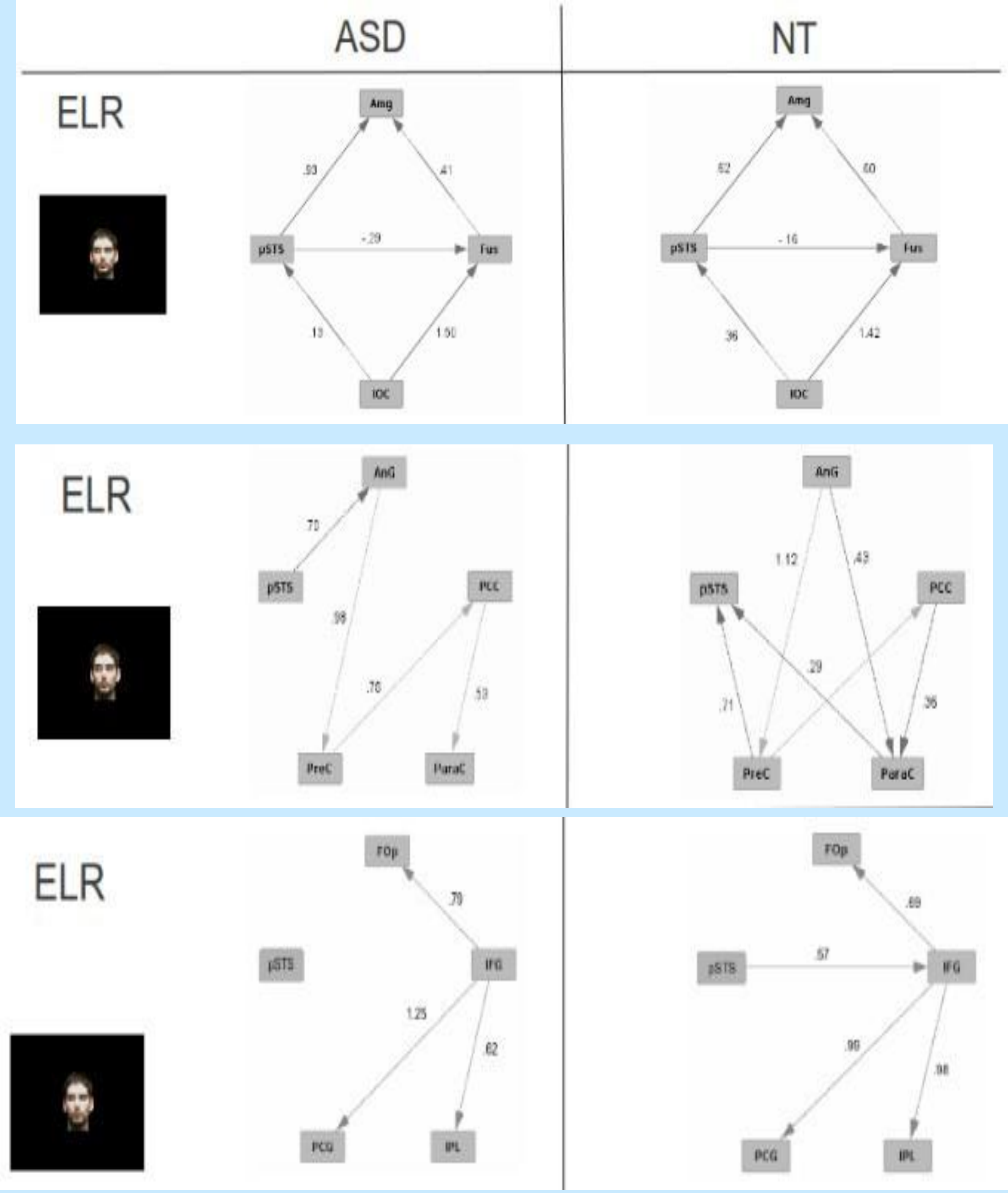
- Face processing network
- Theory of Mind network
- Action understanding network

Results

FACE

TOM

ACTION



What was Learned

face processing: ASD ≈ NT

Theory of Mind: ASD **≠** *NT*

action understanding: ASD ≠ NT
when faces involved

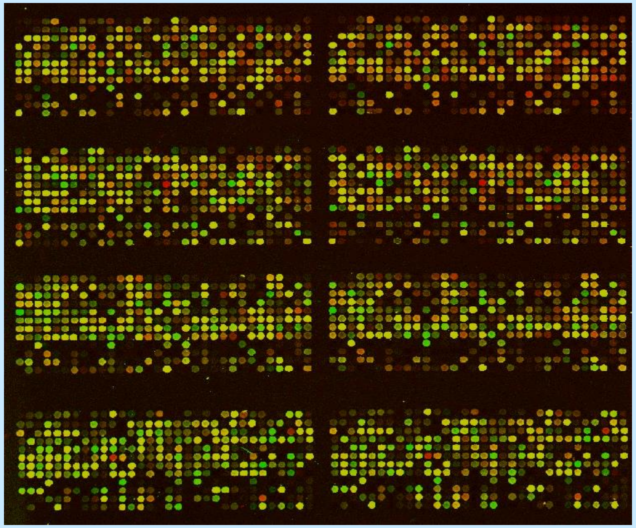
Genetic Regulatory Networks

Arbidopsis

Marloes Maathuis   ZTH (Zurich)

Genetic Regulatory Networks

Micro-array data
~25,000 variables



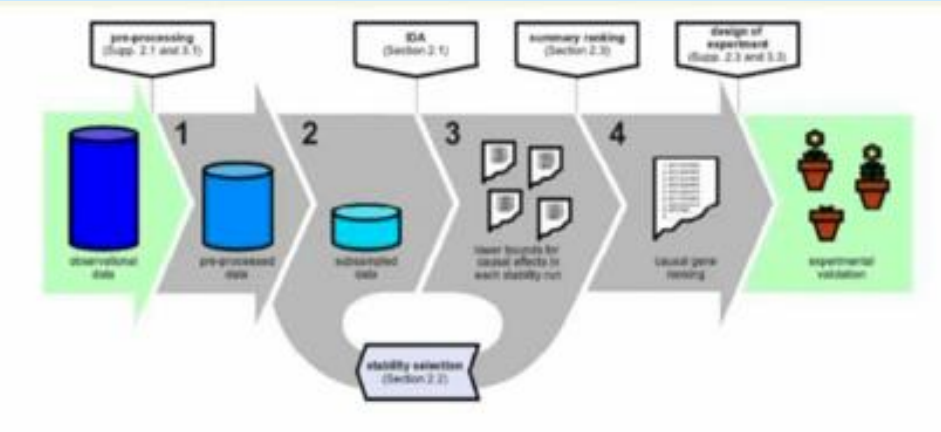
Causal Discovery

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 (the “Doer” effect)
  - 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. !