Searching for Causal Relationships

Richard Scheines
Carnegie Mellon University

Outline

1) Philosophical Foundations

2) Methods → Causality

3) Search

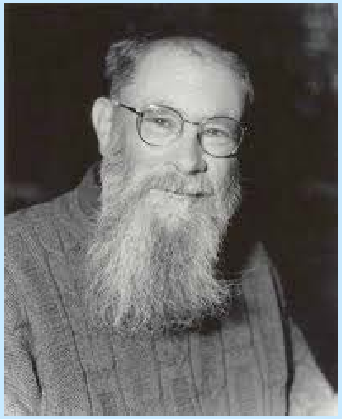
4) Regression vs. Causal Search

5) Example – Genetic Regulatory Network Discovery

Philosophical Foundations

1) Counterfactual Theories

David Hume



David Lewis

A caused B:

If A *had not* occurred, then

B *would not* have occurred

Similarity Metric over Possible Worlds

Vague: If John had not had tar stained fingers, then he would not have gotten lung cancer

Philosophical Foundations

2) Probabilistic Theories

Pat Suppes

A is a cause of B iff

- A is temporally prior to B, and
- P(B | A) > P(B), and
- No event C prior to A that screens off

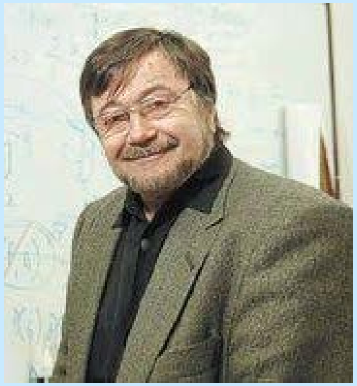
  A and B (i.e., A _||_ B | C)

Some definitions from the editor:
iff = if and only if
P(B | A) = probability of B given A
A _||_ B | C = no event C that makes B and A irrelevant to each other

Philosophical Foundations

3) Intervention Theories

Jim Woodward



Judea Pearl

X is a cause of Y iff

- There are x1 ≠ x2 s.t.

  P(Y | do(X=x1)) ≠ P(Y | do(X=x2))

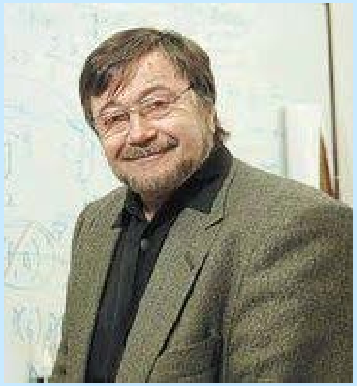
P(Lung Cancer | (Smoking=0)) ≠ P(Y | (Smoking=heavy))

P(Lung Cancer | (Tar Stains=0)) ≠ P(Y | (Tar Stains=heavy))

≠ = not equal to
s.t. = such that
P(Y | do(X=x1)) = probability of Y given that the value of X is x1

Philosophical Foundations

Jim Woodward



Judea Pearl

3) Intervention Theories

X is a cause of Y iff

- There are x1 ≠ x2 s.t.

  P(Y | do(X=x1)) ≠ P(Y | do(X=x2))

P(Lung Cancer | (Smoking=0)) ≠ P(Y | (Smoking=heavy))

P(Lung Cancer | do(Smoking=0)) ≠ P(Y | do(Smoking=heavy))

Smoking → Lung Cancer

P(Lung Cancer | (Tar Stains=0)) ≠ P(Y | (Tar Stains=heavy))

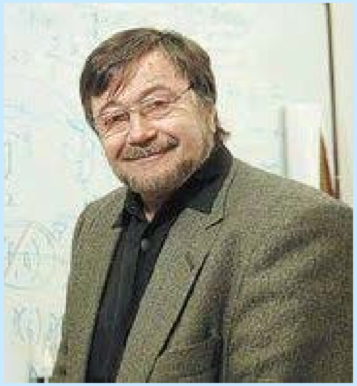
P(Lung Cancer | do(Tar Stains=0)) = P(Y | do(Tar Stains=heavy))

Tar Stains $\nrightarrow$ Lung Cancer

Philosophical Foundations

3) Intervention Theories

Jim Woodward



Judea Pearl

X is a cause of Y iff

- There are $x1 \neq x2$ s.t.

  $P(Y \mid do(X=x1)) \neq P(Y \mid do(X=x2))$

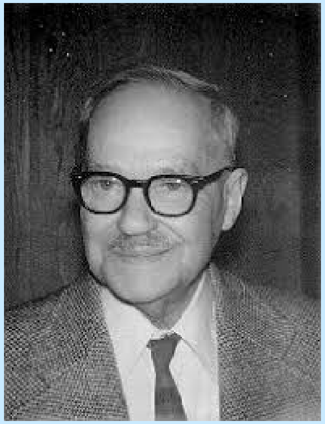
Do is tricky for variables like gender, age, race etc.

P(Wealthy | Race = white)) ≠ P(Wealthy | do(Race = white), Race = black)

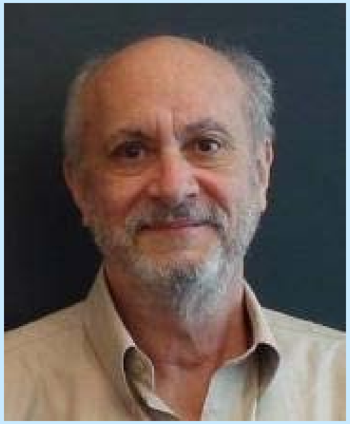
Philosophical Foundations

4) Potential Outcomes

- P(Y | had we done(X=x1), did(X=x2))
- P(Y | had we done(X=x1), X=x2)
- Etc.



Jerzey Neyman



Don Rubin

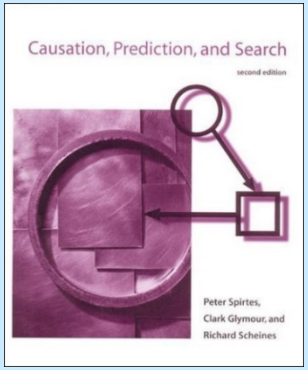
Mathematical Representation must include intervention

Predicting Interventions

- $P(Y \mid (X=x1))$
- $P(Y \mid do(X=x1))$
- $P(Y \mid do(X=x1), X=x, \mathbf{Z} = z)$

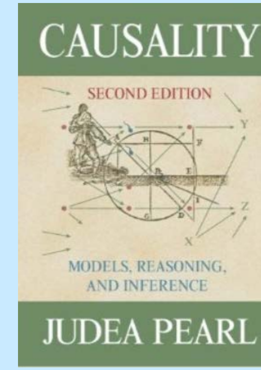
Counterfactuals

- $P(Y \mid \text{had we done}(X=x1), did(X=x2), X=x, \mathbf{Z} = z)$



Graphical Models

Intervention & Manipulation



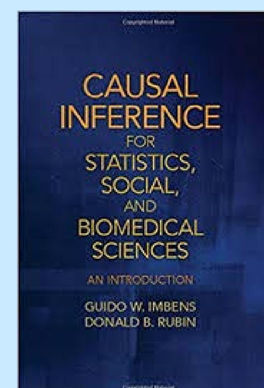
Search

Modern Theory of Statistical Causal Models

Testable Constraints (e.g., Independence)

Potential Outcome Models

Counterfactuals



Paths to Causality

- Human Experiments
- Animal Experiments
- Human Observational Studies
- Mechanistic Modeling
- Search

Paths to Causality

Human Experiments

- Randomize X=x, do(X=x)
- If do(X=x) $\not\!\perp\!\!\!\perp$ Y then X $\rightarrow$ Y
- With randomization, correlation *is* causation

Animal Experiments

- Experimentally determine X $\rightarrow$ Y (in animal model)
- Estimate relevance of animal model to humans

Paths to Causality

Observational Studies

- Establish Association ($X \not\perp\!\!\!\perp Y$)
- Eliminate alternative sources of association:
  - Reverse causation: $X \leftarrow Y$,
  - Confounding: $X \leftarrow C \rightarrow Y$
  - Sample Selection Bias: $X \rightarrow$ In sample $\leftarrow Y$

Strategies

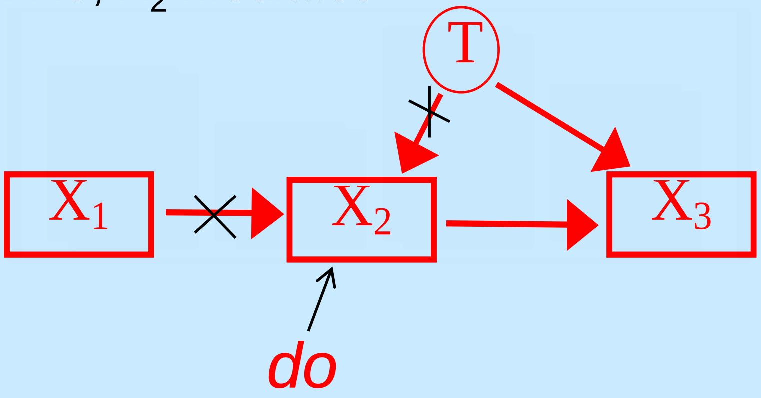
- Instrumental Variables
- Measure, model, and statistically control for confounding

Statistical Control ≠ Experimental Control

Question: Does X_1 directly cause X_3?

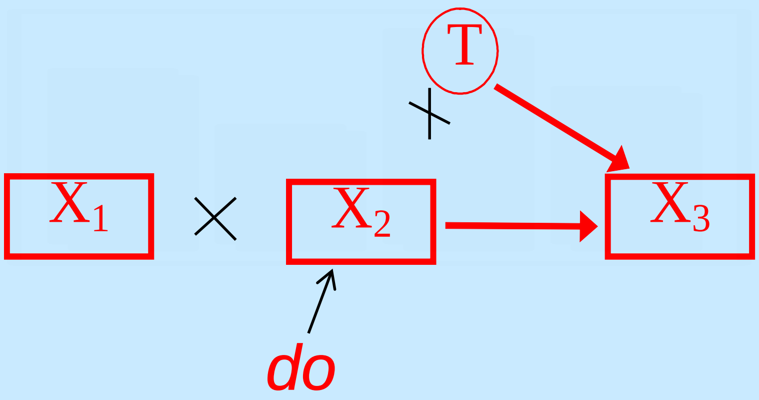
How to find out?

Truth: No, X_2 mediates



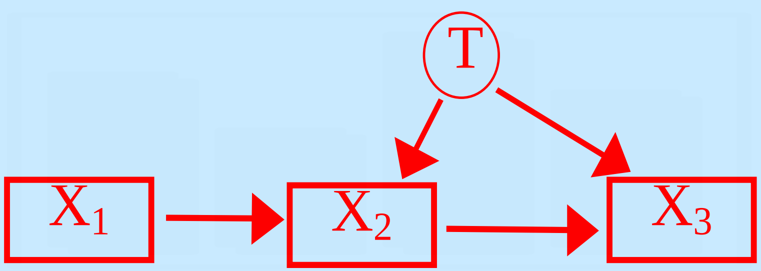
Experimentally control for X_2

Statistical Control ≠ Experimental Control



Experimentally control for X_2

$X_3 \perp\!\!\!\perp X_1 \mid do(X_2)$



Statistically control for X_2

$X_3 \not\perp\!\!\!\perp X_1 \mid X_2$

Paths to Causality

Mechanistic Models

- PBPK Models in Toxicology
- Mode of Action Models
- Mechanism of Action Models

Paths to Causality

Search

Causal Structure → Computable Relation → Testable Constraints

Testable Constraints

Background Knowledge

→ Discovery Algorithm → Causal Structures

Causal Features common to all members of the output

Model Evaluation

Causal Structure

Testable Statistical Predictions

Causal Graphs

e.g., Conditional Independence

$X \rightarrow Y \rightarrow Z$

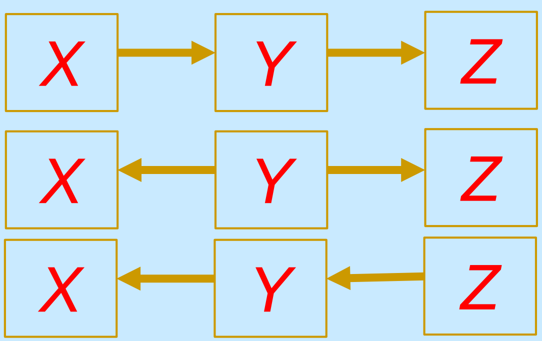
X _||_ Z | Y

Model Search

Causal Structure ← *Discovery Algorithm* ← Testable Statistical Predictions

Equivalence Class of Causal Graphs

Conditional Independence



X _||_ Z | Y

Regression vs. Causal Discovery

X → Y ??

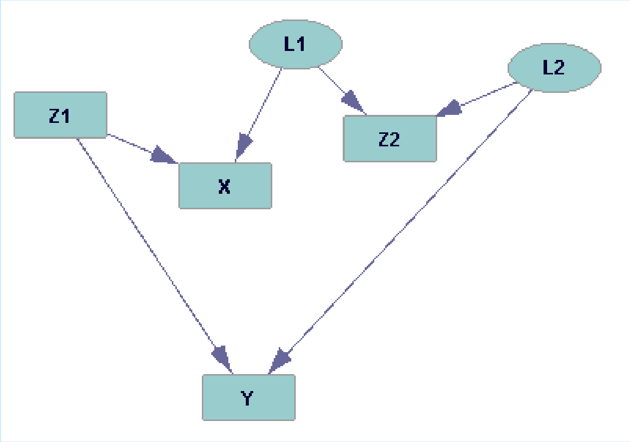
Observational Study Strategy:

Measure and control for all confounders Z:

- Prior to X
- Associated with X
- Associated with Y

Regression vs. Causal Discovery

X → Y ??



True Model

Data (Population)

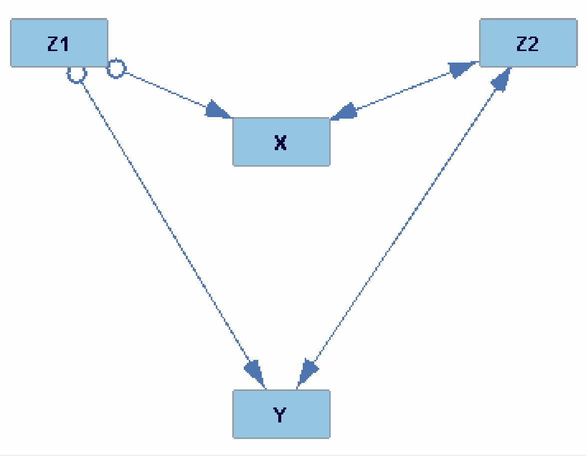
Multiple Regression

Y : Dependent Var.

X, Z1, Z2 : Independent Vars.

All coefficients ≠ 0

Causal Discovery Algorithm (FCI)



Search Example*
Genetic Regulation

Which genes regulate flowering time in *Arabidopsis thaliana*?

* Stekhoven DJ, et al. Causal stability ranking. *Bioinformatics* 28 (2012) 2819-2823.

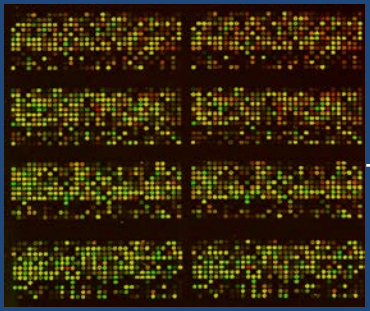
Observational Data

- n = 47 *Arabidopsis thaliana* gene expression profiles of 4-day old seedlings for which subsequent flowering time was also measured

- Affymetrix ATH1 arrays with expression measurements on 21,440 *A. thaliana* genes

Causal Network Analysis

Observational data:
Microarray measurements (**X**)
and flowering time (Y)

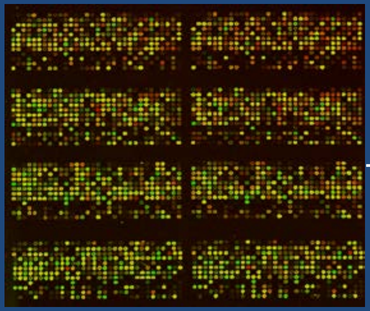


Causal Analysis

Causal networks on the variables in X and Y

Candidate Gene Selection

Observational data:
Microarray measurements (**X**)
and flowering time (Y)



Causal Analysis

Causal networks on the variables in X and Y

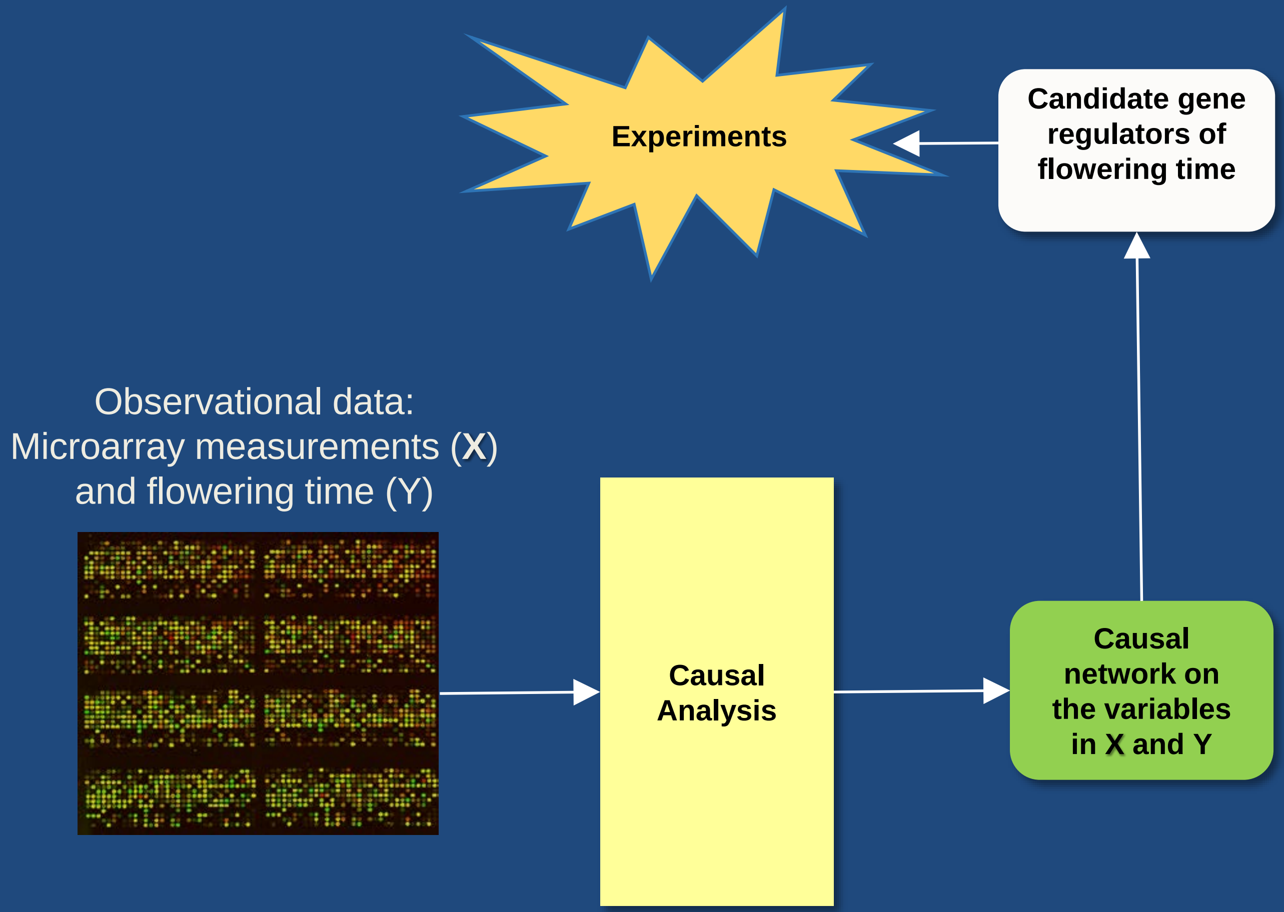
Select those genes that reliably cause the most substantial changes in flowering time (Y) according to the causal networks

Candidate gene regulators of flowering time

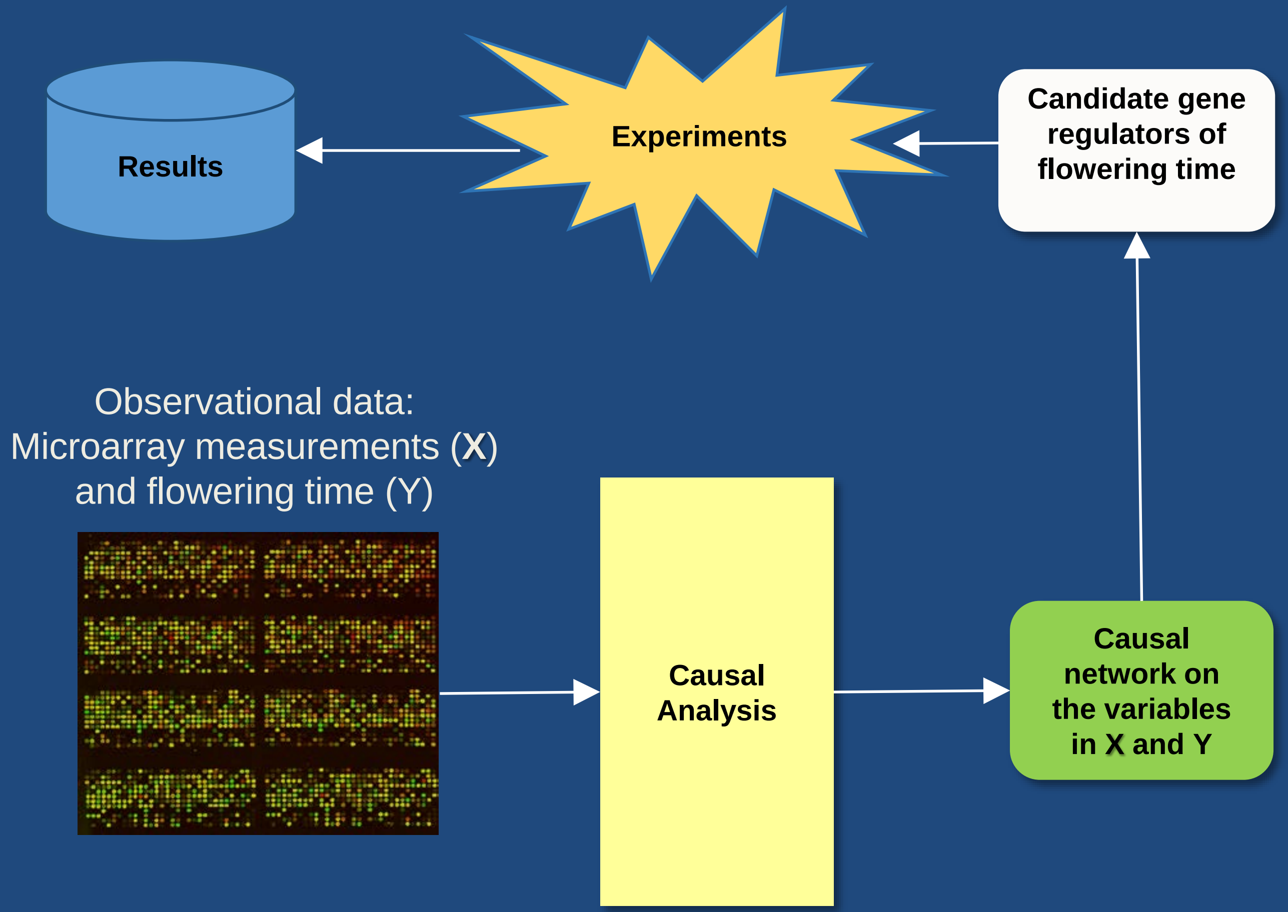
Candidate Regulators of Flowering Time

- Output: 25 genes → flowering time
- 5 of 25: known regulators of flowering
- Among remaining 20 not known to be regulators:
  13: mutant seeds available

Experimental Investigation



Experimental Results



Results

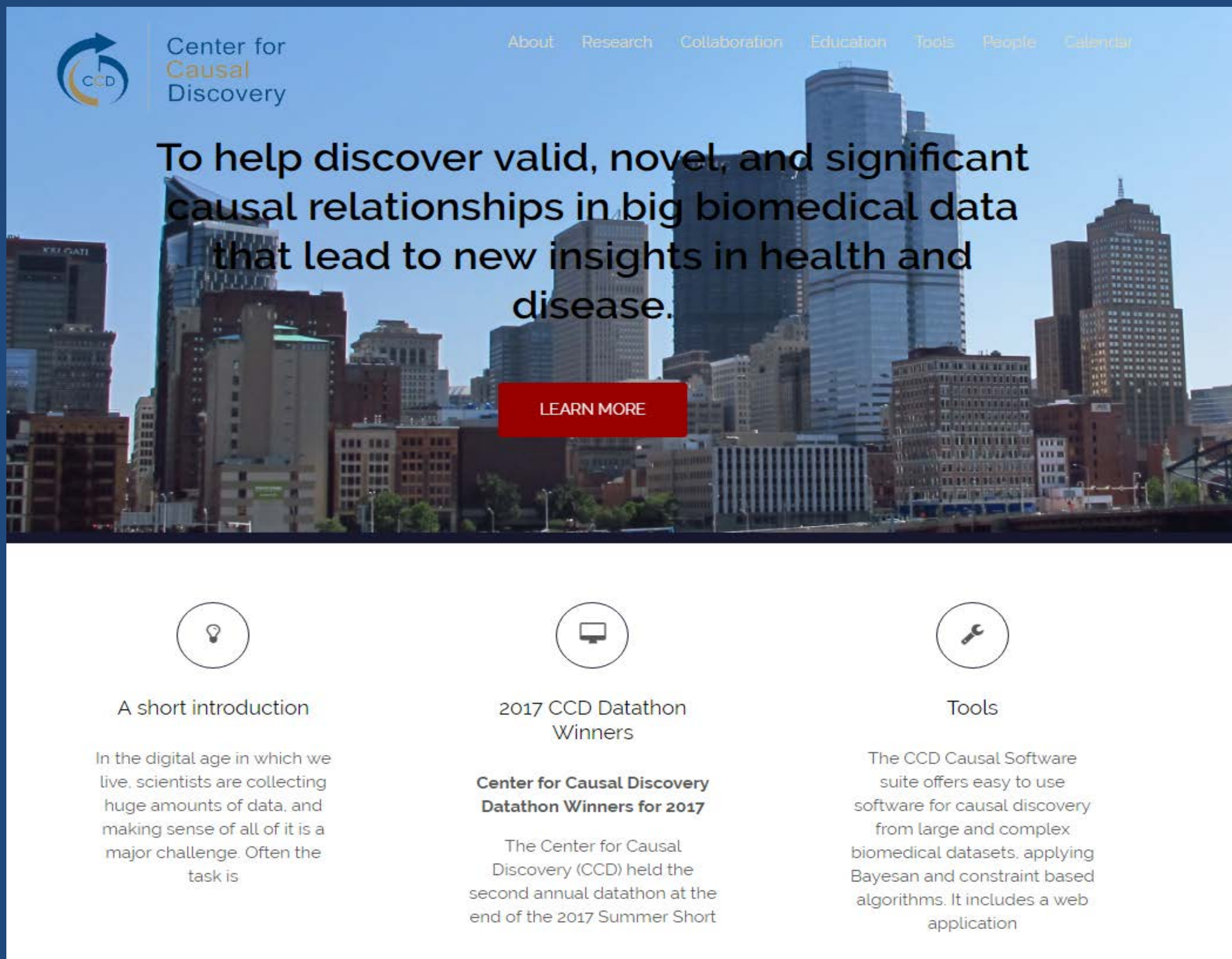
- 9 seed types, each with a single gene insertion, that yielded 4 or more plants

- 4 of 9: shorter mean flowering time ($p < 0.05$) than the control, wild-type plants

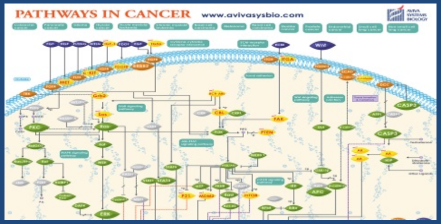
Greenhouse experiments on flowering time

NIH BD2K: Center for Causal Discovery

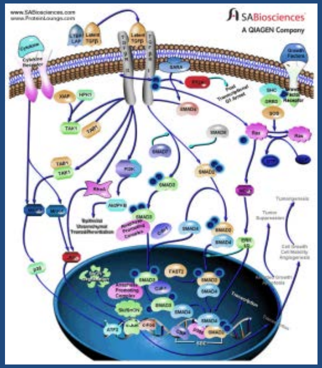
www.ccd.pitt.edu



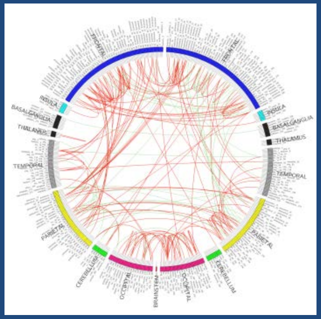
Driving Biomedical Projects (DBPs)



- Discover cell signaling networks in cancer



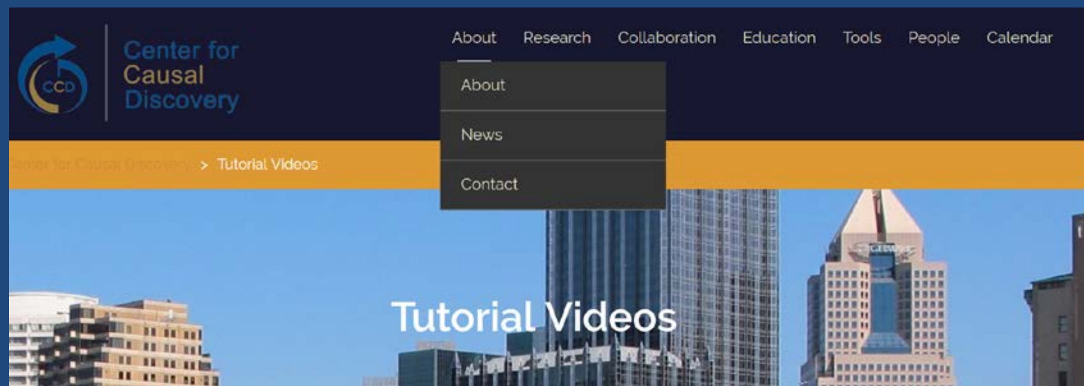
- Discover the mechanisms of disease onset and progression in chronic obstructive pulmonary disease and idiopathic pulmonary fibrosis

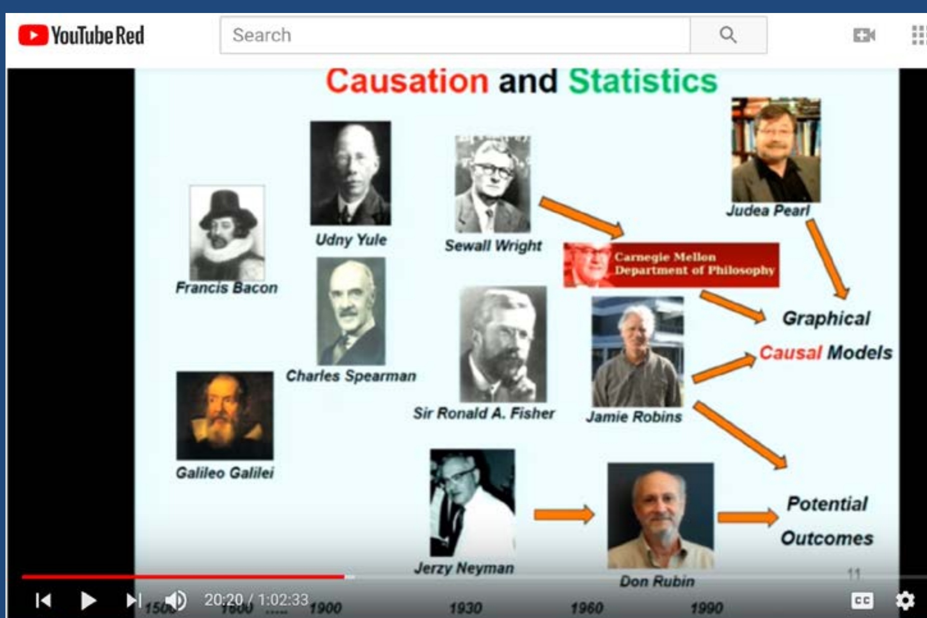


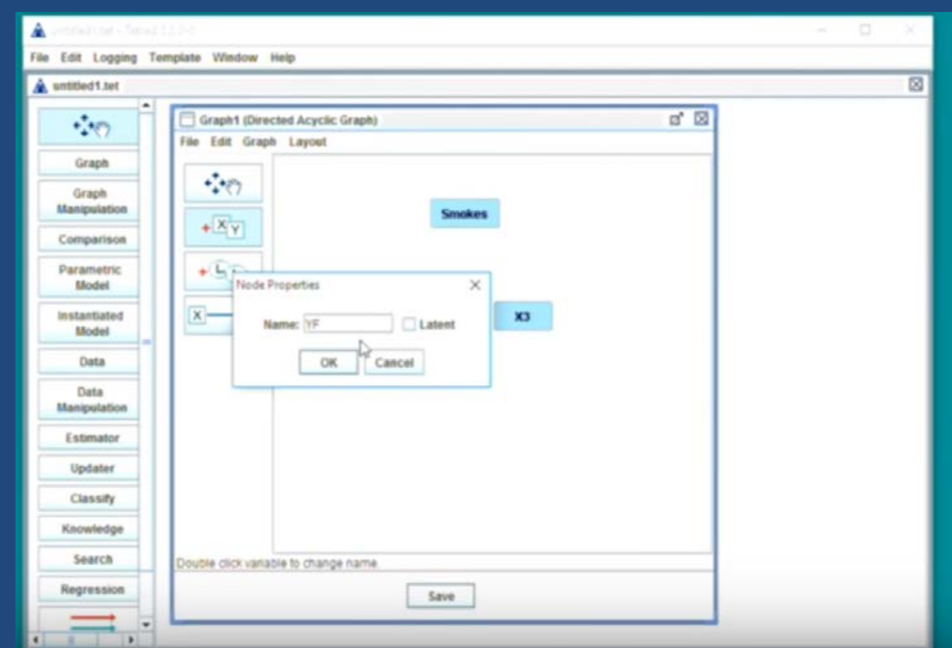
- Discover the functional (causal) connectivity of regions of the human brain from fMRI data

NIH BD2K: Center for Causal Discovery

www.ccd.pitt.edu







Summary

- Search is important when:
  - Low to moderate background knowledge theory
  - Many variables: astronomically many models
- Algorithms are now well developed, freely available
- Still hard to use intelligently without training

References

Lewis, David (1973). "Causation". *The Journal of Philosophy*. **70** (17).

Eberhardt, Frederick and Richard Scheines, 2007, "Interventions and Causal Inference", *Philosophy of Science*, 74(5): 981–995. doi:10.1086/525638

Pearl, Judea, 2009, *Causality*, New York: Cambridge University Press

Rubin, Donald B., 1974, "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies", *Journal of Educational Psychology*, 66(5): 688–701.

Sekhon, Jasjeet (2007). "The Neyman–Rubin Model of Causal Inference and Estimation via Matching Methods"(PDF). The Oxford Handbook of Political Methodology.

Spirtes, P., Glymour, C., Scheines, R. (2000). *Causation, Prediction, and Search*, 2nd Edition. MIT Press.

Stekhoven DJ, et al. Causal stability ranking. *Bioinformatics* 28 (2012) 2819-2823

Suppes, P. (1970) *A Probabilistic Theory of Causality*, Amsterdam: North-Holland Publishing

Woodward, James/Jim, 1997, "Explanation, Invariance, and Intervention", *Philosophy of Science*, 64(supplement): S26–S41

Woodward, James. 2003, *Making Things Happen: A Theory of Causal Explanation*, Oxford: Oxford University Press

Thank You!