Causal Discovery by Computer

Clark Glymour

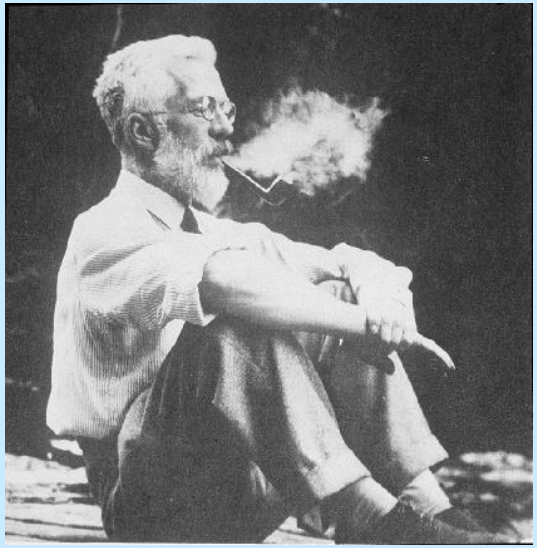
Carnegie Mellon University

Outline

1. A century of mistakes about causation and discovery:
    1. Fisher
    2. Yule
    3. Spearman/Thurstone
2. Search for causes is statistical estimation
3. Strategies and algorithms
4. Examples
5. Prospects: The Center for Causal Discovery

A Conversation with R.A. Fisher

Science is about causation. Causes can only be discovered by experiment.



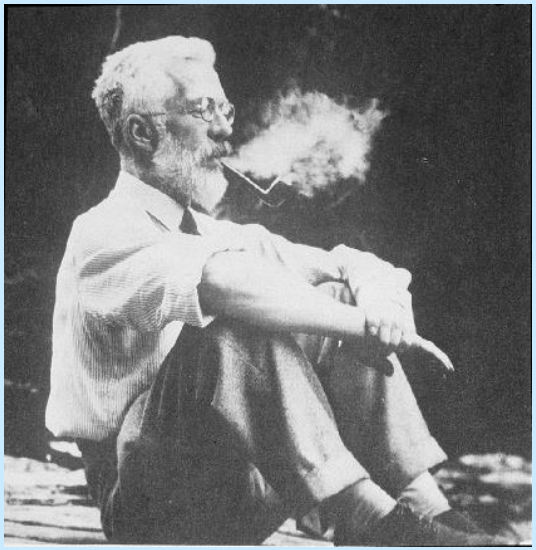
So what did you study? And what did you research?

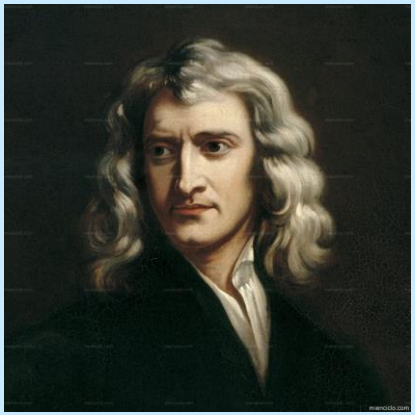
Physics and astronomy. Then statistics and biology, especially evolution.

Say again!?

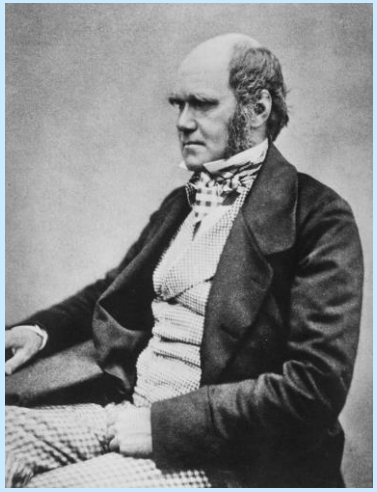
Continued…

Very well: Science is about causation. Causes can only be discovered by experiment.



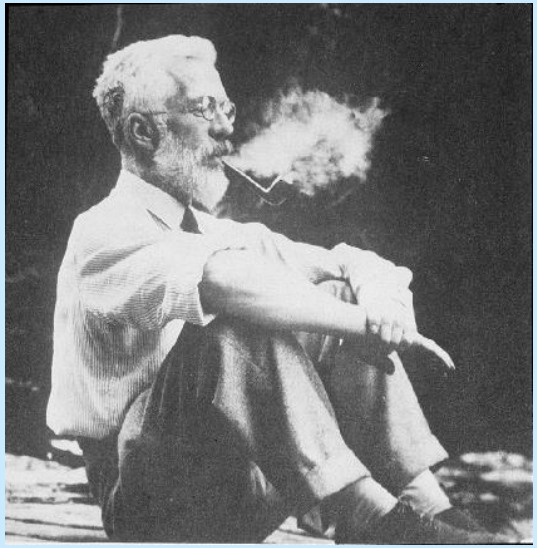


So much then for my discovery of the cause of the motions of the planets, and the cause of the tides.



So much as well for my discovery of the origin of species.

I have a proof



Lung Cancer is associated with smoking. The association can equally be explained in three ways:

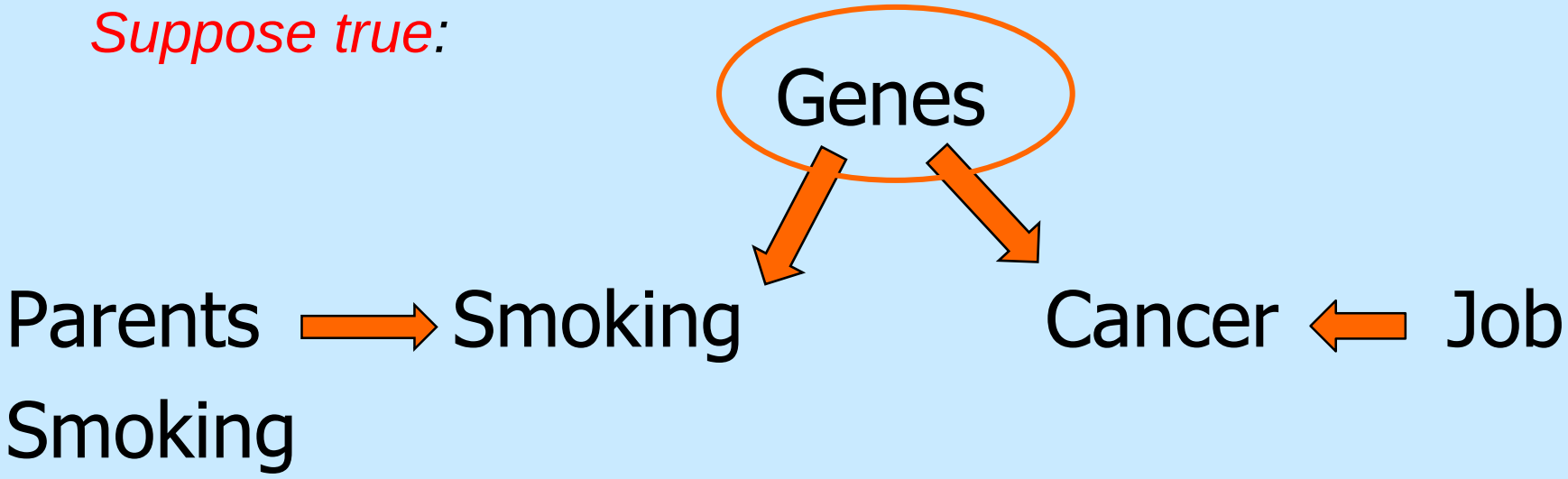
1. *Early undetected cancers cause smoking*
2. *Genetics causes both smoking and cancer*
3. *Smoking causes cancer*

The data cannot distinguish these explanations.

The argument is quite general: Associations cannot distinguish cause from effect from joint effects of the unobserved

Even Genius Makes Mistakes

Suppose true:



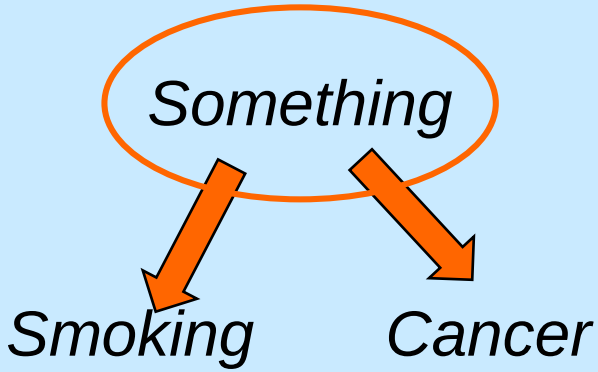
Suppose Data:

- *Parents Smoking and Smoking are not associated with Job*
- *Parents Smoking is not associated with Cancer*
- *Smoking and Cancer are associated*
- *Parents Smoking and Smoking are associated*
- *Job and Cancer are associated*

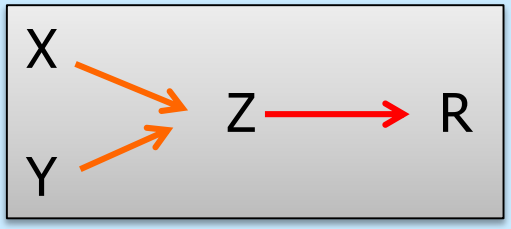
Suppose Data:

- *Parents Smoking and Smoking are not associated with Job*
- *Parents Smoking is not associated with Cancer*
- *Smoking and Cancer are associated*
- *Parents Smoking and Smoking are associated*
- *Job and Cancer are associated*

*Then **every causal explanation** of the data entails that*



Another example:

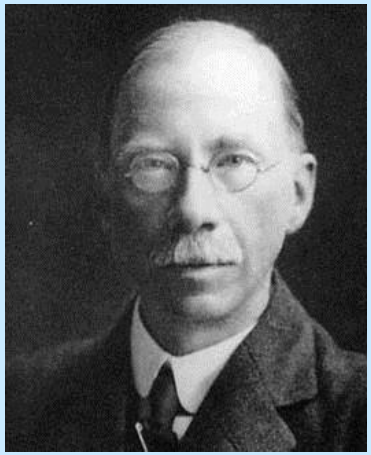


Suppose data were to show:

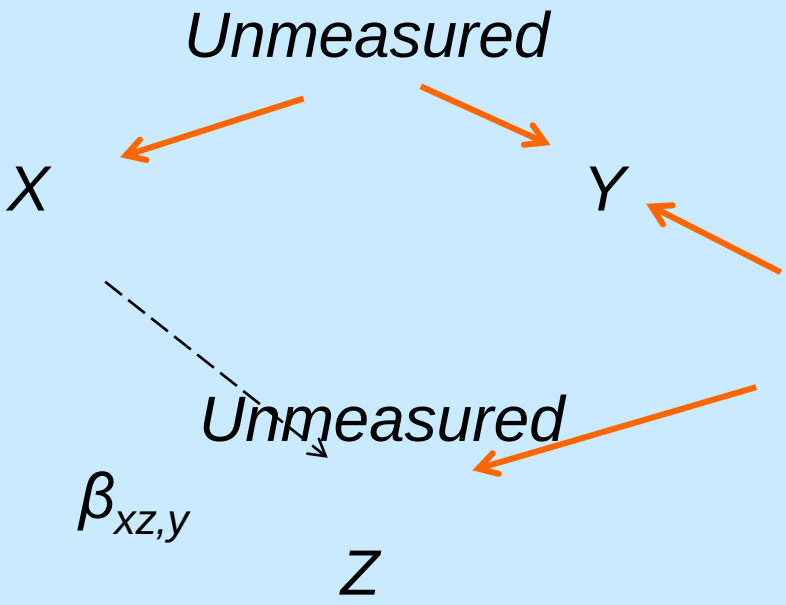
- X and Y are not associated
- Z is associated with X and Y
- X and Y *are* associated conditional on a value of Z
- X, Y, Z are associated with R
- X, Y are not associated with R conditional on Z

Then Z is a cause of R and there are no confounding common causes of Z and R.

George Udney Yule and Regression

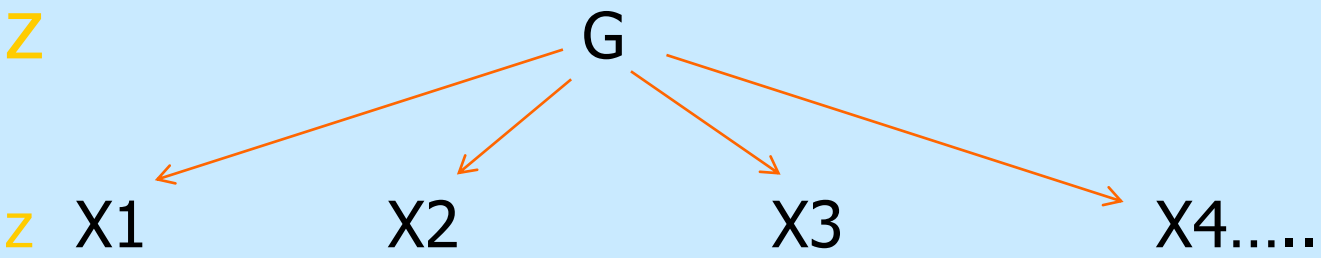


Yule championed multiple regression as a means to estimate causal effects. He was aware of some, but not all, of its foibles:



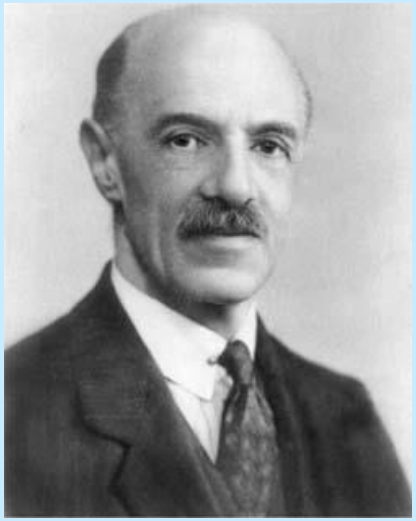
Although X and Z have no unmeasured common cause, and X has no influence on Z, regression of Z on X,Y results in a

Spearman, Thurstone and Factor Analysis



Thurstone's Equivocation:

"Vectors of Mind" = Linear reduction of correlations



Upshot for the 20th Century

- Causal inferences should not be made from non-experimental data
- But if you must, use regression
- Or factor analysis
- Or "potential outcomes"

Fear of Search

Terry Speed: There may be better search methods than regression, but only regression should be used.

~ 1990: Rethinking:

- Relationships between causal hypotheses and probability hypotheses
- Search

Search for causal relations is statistical estimation

Hypothesis Space:

	A	B	C
A	?	?	?
B	?	?	?
C	?	?	?

Data: A $\perp\!\!\!\perp$ C , A $\not\perp\!\!\!\perp$ B, B $\not\perp\!\!\!\perp$ C, A $\not\perp\!\!\!\perp$ C | B

Statistical inference:

A -> B <- C:

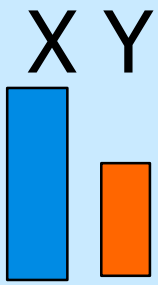
	A	B	C
A	0	1	0
B	0	0	1
C	0	1	0

What's the Difference?

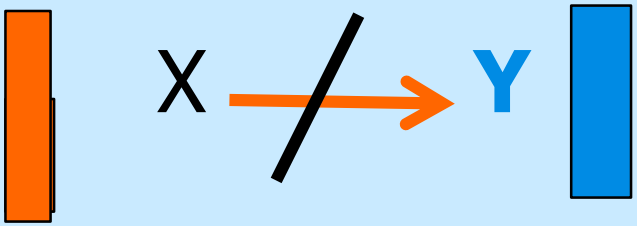
In conventional estimation we are estimating a *probability distribution*—an unobserved distribution covering present and potential future observations.

In causal estimation, we are estimating both a current probability distribution *and* the *probability distributions that would result from various interventions.*

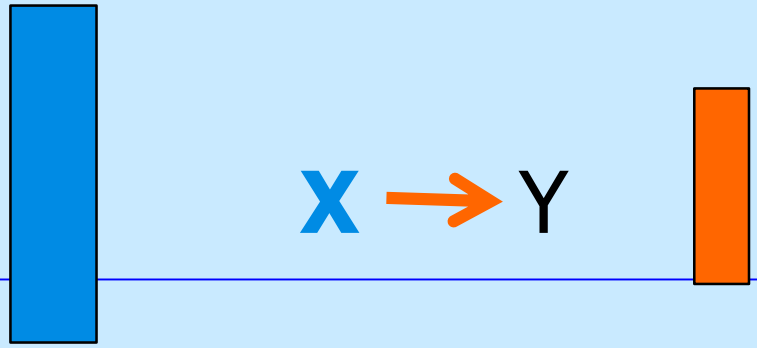
Interventions Change Probabilities



Observed



Force a value on Y; distribution of X doesn't change. Dependence of Y on X is broken



Force a value on X; distribution of Y changes

Goals are the same as in Experimental Design

*First Goal: To estimate whether there are **some** values of other variables **C** such that **some** intervention that changes the distribution of A will change the distribution of B when **C** variables are forced to have those values.*

Secondary Goals: to estimate the signs or strengths of effects.

	A	B	C
A	?	?	?
B	?	?	?
C	?	?	?

Conventional Statistical Estimation Has Assumptions, E.G.

- Distribution Family?
- I.I.D Sampling?
- Censored Data?
- Stationary Time Series?
- Variance Known?
- Prior Probabilities?
- Likelihood Function (i.e., the "model")

Standards for Statistical Estimation Are Standards for Search

- Asymptotic convergence to true information under explicit assumptions.
- Unbiased (Expected value is the true value)
- Finite sample error probabilities.
- Tests of assumptions.
- Robustness to "small" violations of assumptions.

Representation

- Represent causal relations on a collection **V** of variables by directed edges X -> Y

- A directed edge indicates that some intervention on X would change the distribution of Y if all other variables in **V** (that are not effects of X) were held constant at some values.

What Extra Assumptions Are Needed for Causal Inference?

Answer: Generalizations of principles of experimental design:

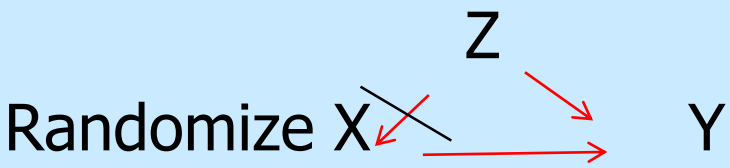
Markov Condition

Faithfulness Condition

Experimental Principles

Why do we randomize in experiments?

Because we expect (or hope) that on average randomizing X will remove any association between X and Y due to common causes of X and Y.



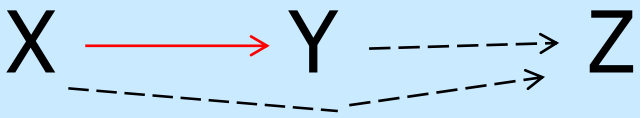
And *thus*, the association (or it's absence) of X and Y in the experiment will measure the effect (or it's absence) of X on Y.

Why in experiments do we "stratify" values of potential confounders, Z, (i.e. arrange subsamples in which Z is constant)?

One reason: Because we expect that the association of X and Y when Z is constant (i.e., conditioned on) will measure the effect of X on Y.

Experimental Principles: Blinding

When we think Y is an effect of X and we do not know whether Z is an effect of Y, and we want to know whether X causes Z



- We do **not** force Y to be constant when we randomize X.
- We do **not** condition on Y when we randomize X.
- But if we want to know if there is a direct influence of X on Z, we **do** condition on Y.

Generalizations

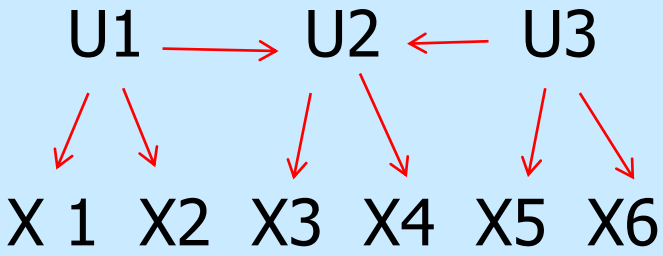
- Markov: X is independent of all other variables (except for effects of X) conditional on the direct causes of X.
- Faithfulness: All independencies and conditional independencies among variables in a system of variables follow from the Markov condition for the true causal graph of the system.

A Flow of Mathematical Results Since 1990

- *Sufficient* conditions for recovering causal information:
- Markov
- Acyclic
- Faithfulness
- I.I.D sampling
- No unrecorded common causes

And a host of alternative sufficient conditions.

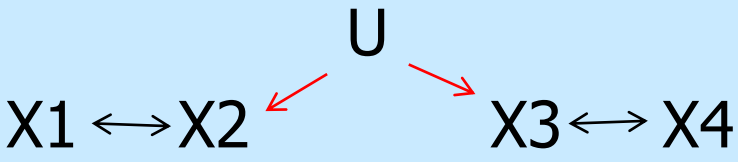
Examples for which There Are Correct Search Procedures

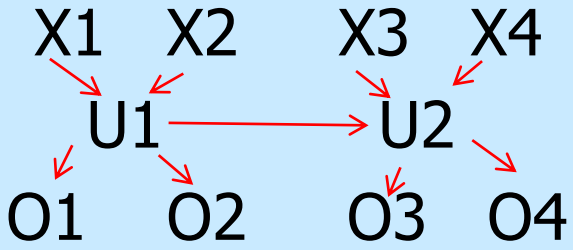


Linearity of X variables

Non-Gaussian

Linear

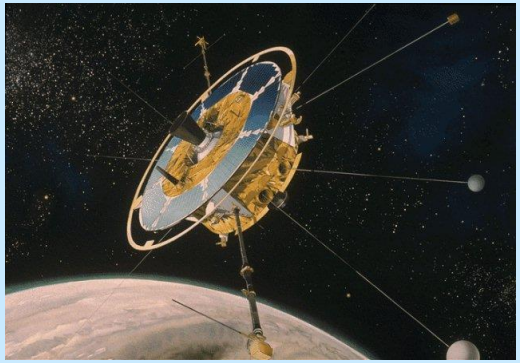




The Zen of Search: Don’t Look (where you don’t need to)

Applications

In flight recalibration



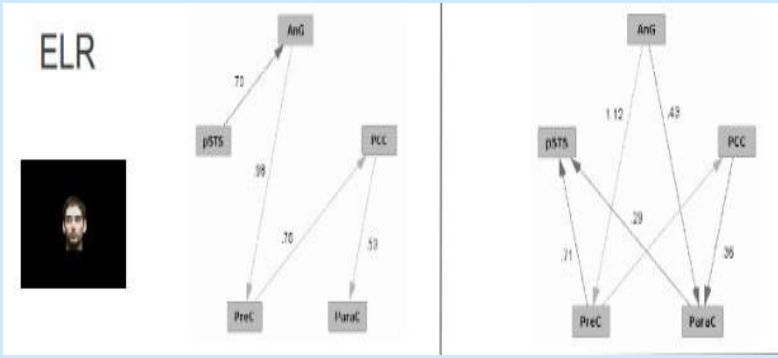
Acid Rain

College Dropouts

Lead & IQ

Gene Regulators

Autism Spectrum



Center for Causal Discovery

1. New Algorithms
2. Moving from small discovery problems to huge problems
3. Making software easily usable
4. Training
5. Science Applications:
   - Cellular pathways in breast cancer
   - .....in lung disease
   - Brain processing in normals and autistics

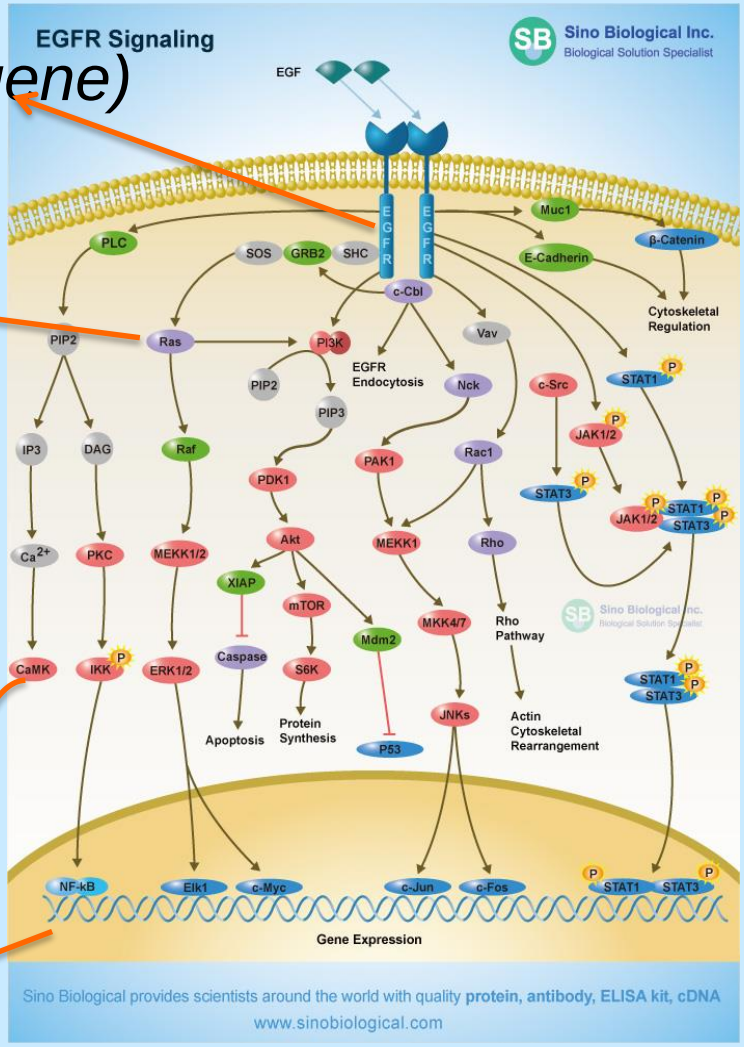
Cancer Signaling and Big Data

Protein (gene) mutations

Protein quantity

Protein modification

RNA transcription



- Genomic data
  - Somatic mutations
  - Somatic copy number alterations
  - DNA methylation
- Proteomic data
  - Quantity of certain signaling proteins
  - Chemical modification (phosphorylation) of proteins
- Transcriptomic data
  - mRNA
  - miRNA
  - lncRNA

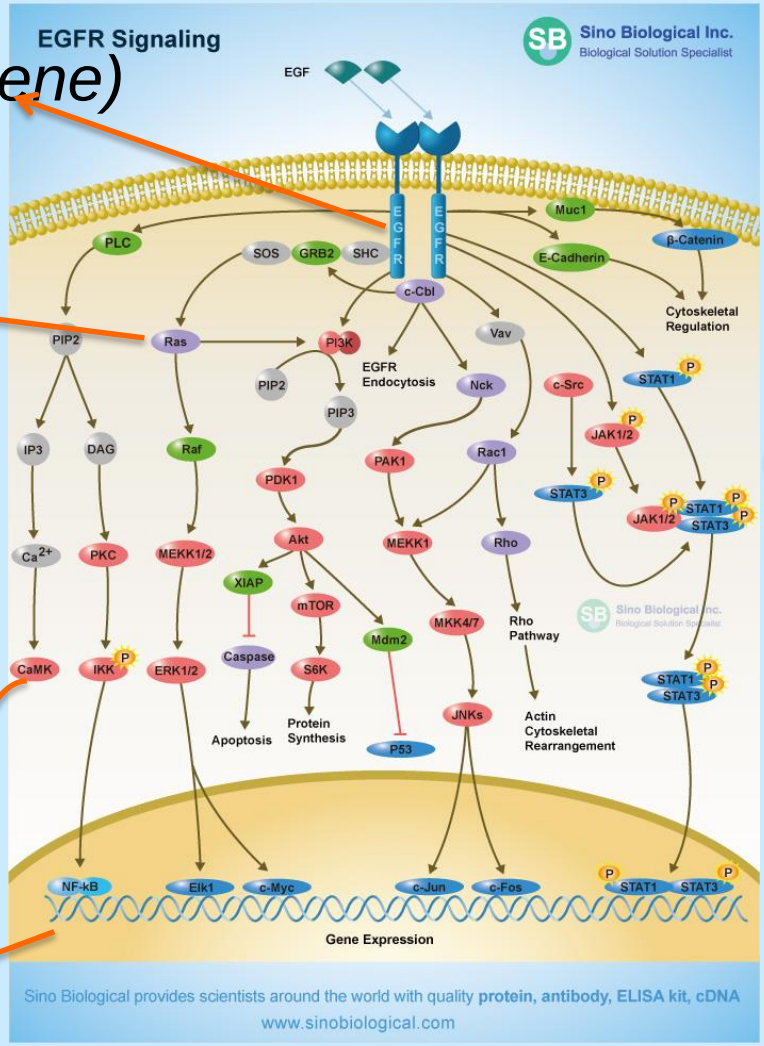
Signaling Pathway and Causal Network

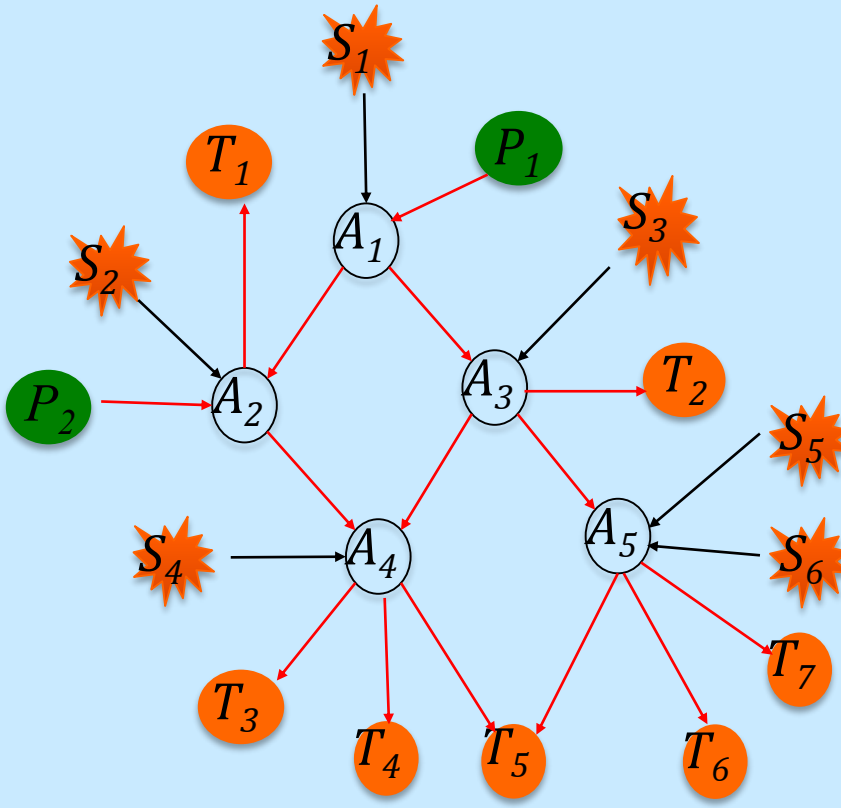
Protein (gene) mutations

Protein quantity

Protein modification

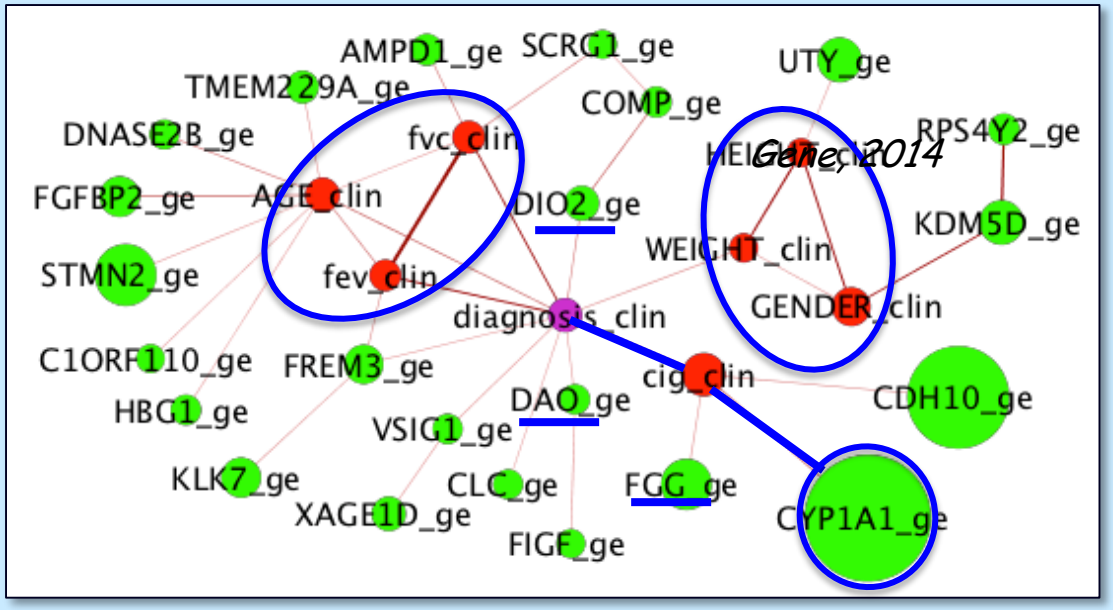
RNA transcription





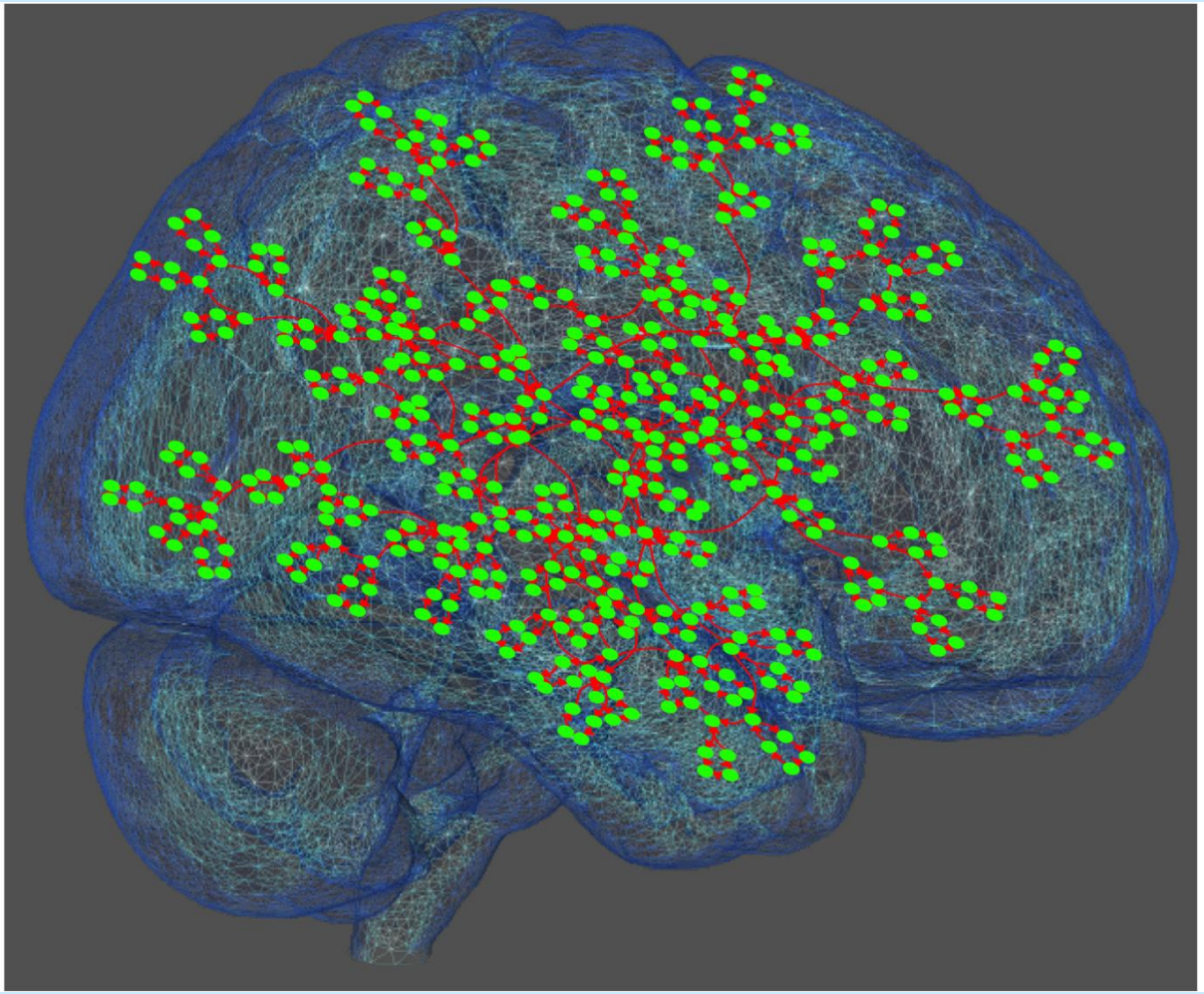
Identify associations between clinical features, omics and lung disease using graphical models (undirected)

AJ Sedgewick

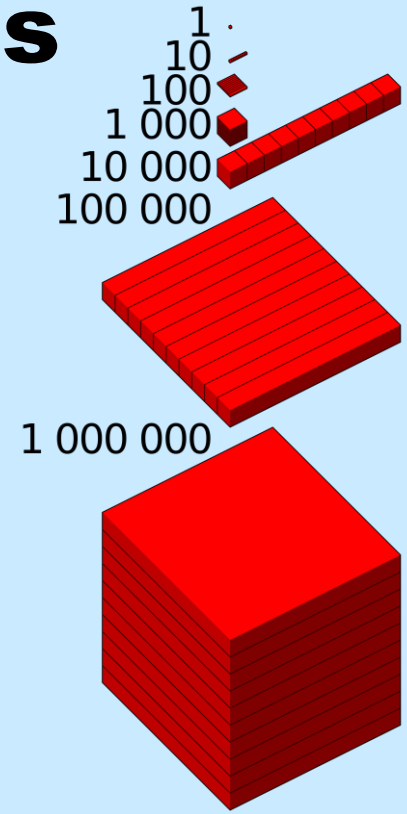


Gene, 2014

Finding the Neural Mechanisms in Autism



5 Months of Algorithmic Progress



Size:

.(Thanks PSC)

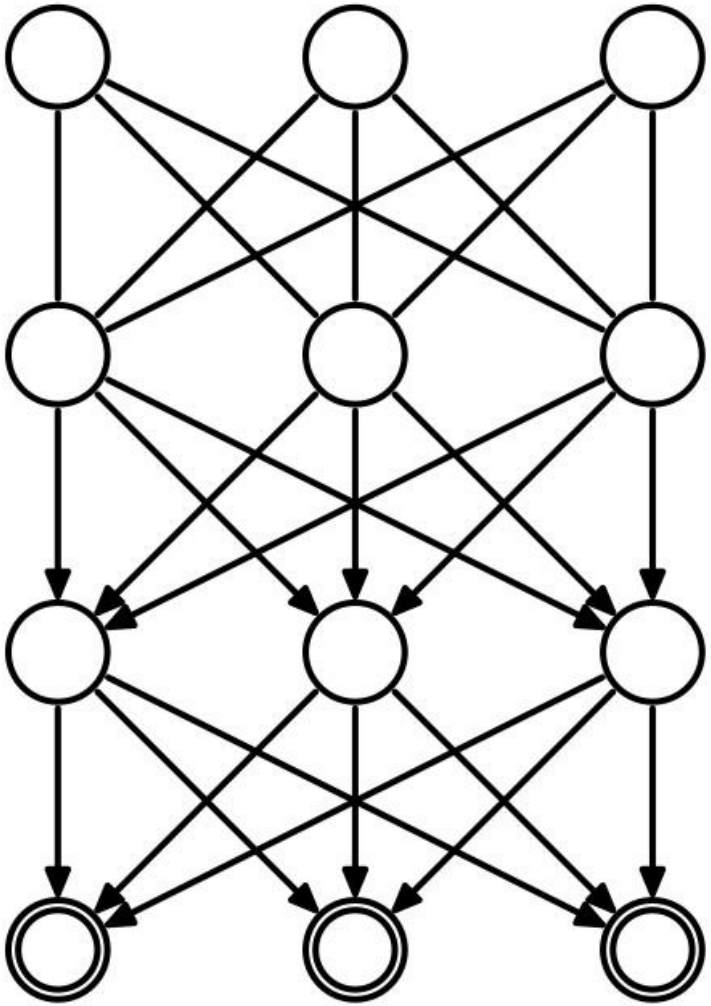
Variable Types:

Mixed categorical and continuous variables

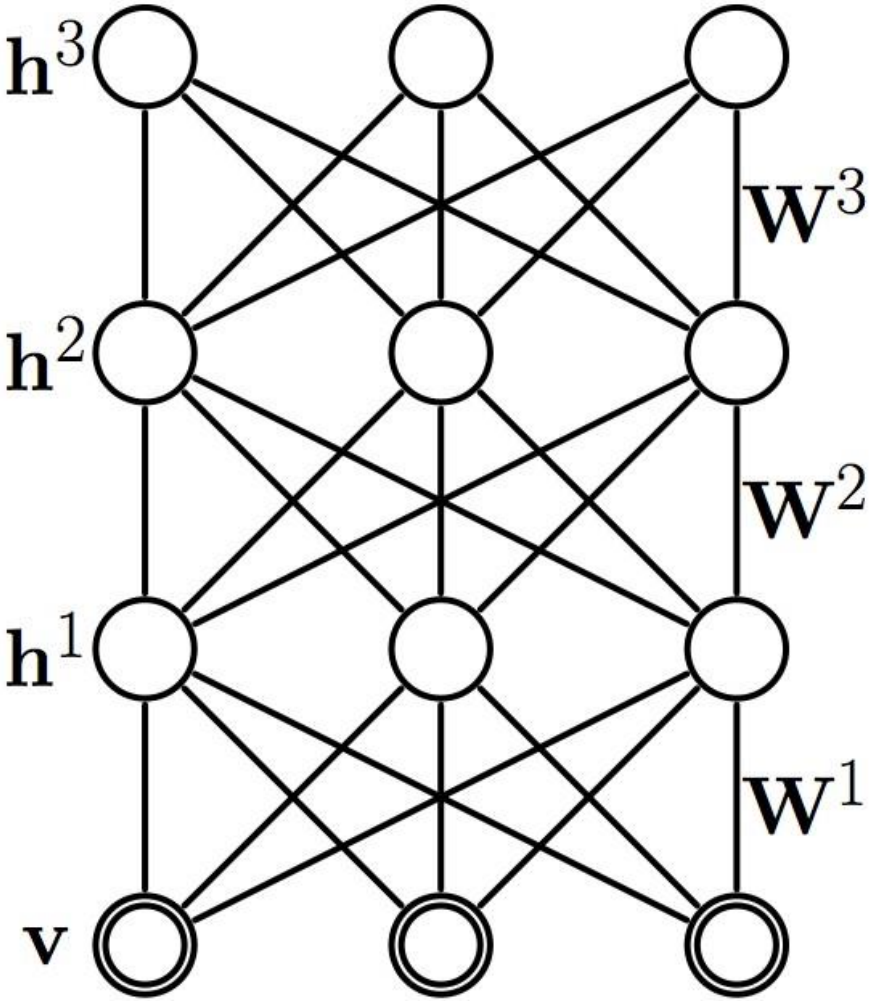
Model Types

Endogenous latent variables: X -> U -> Y

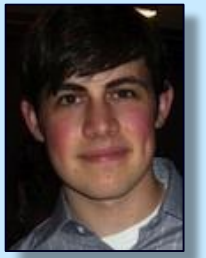
Deep Belief Network



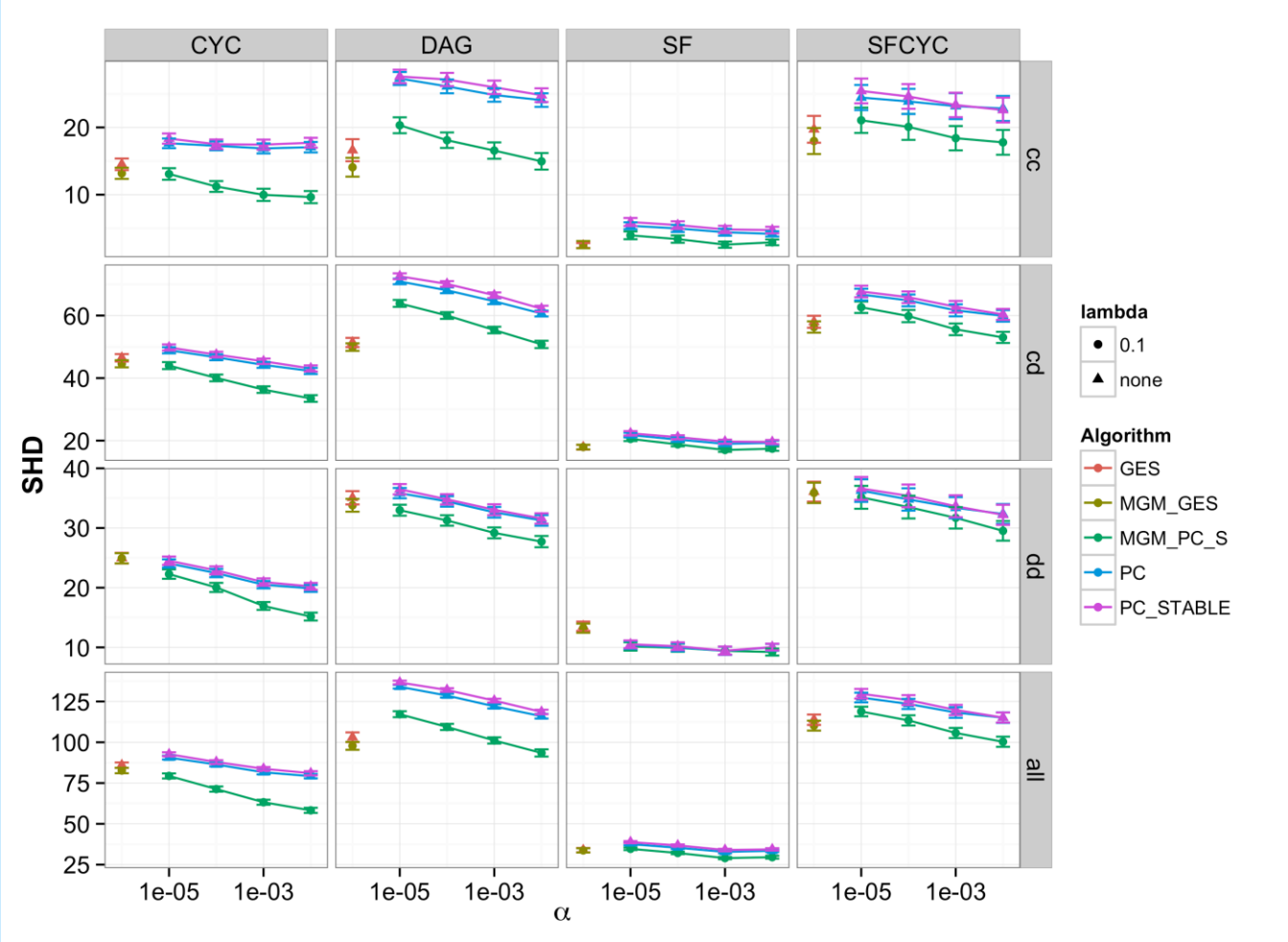
Deep Boltzmann Machine



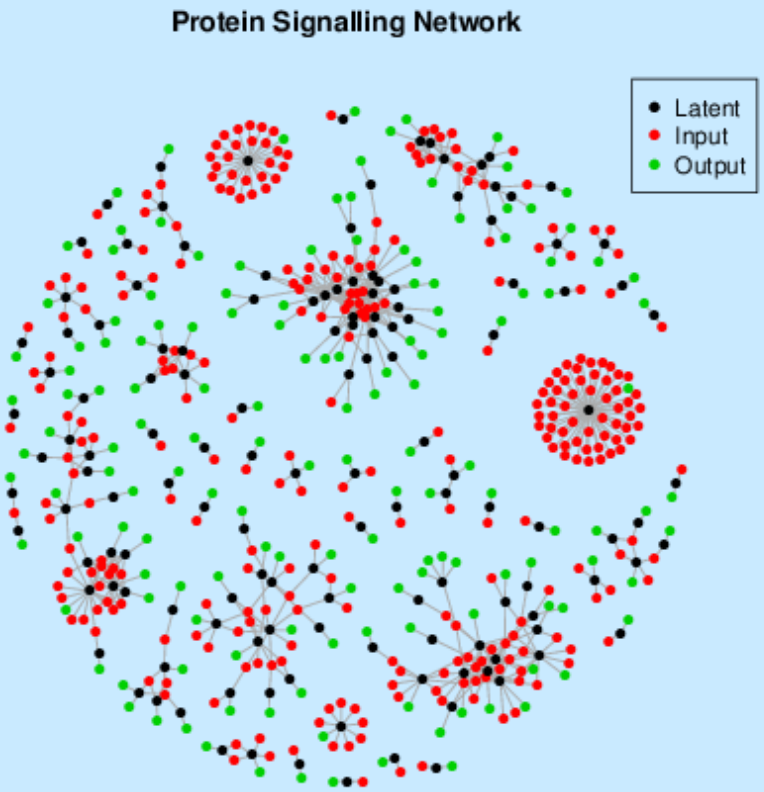
Comparison of MGM-Learn flavors to other structure learning methods



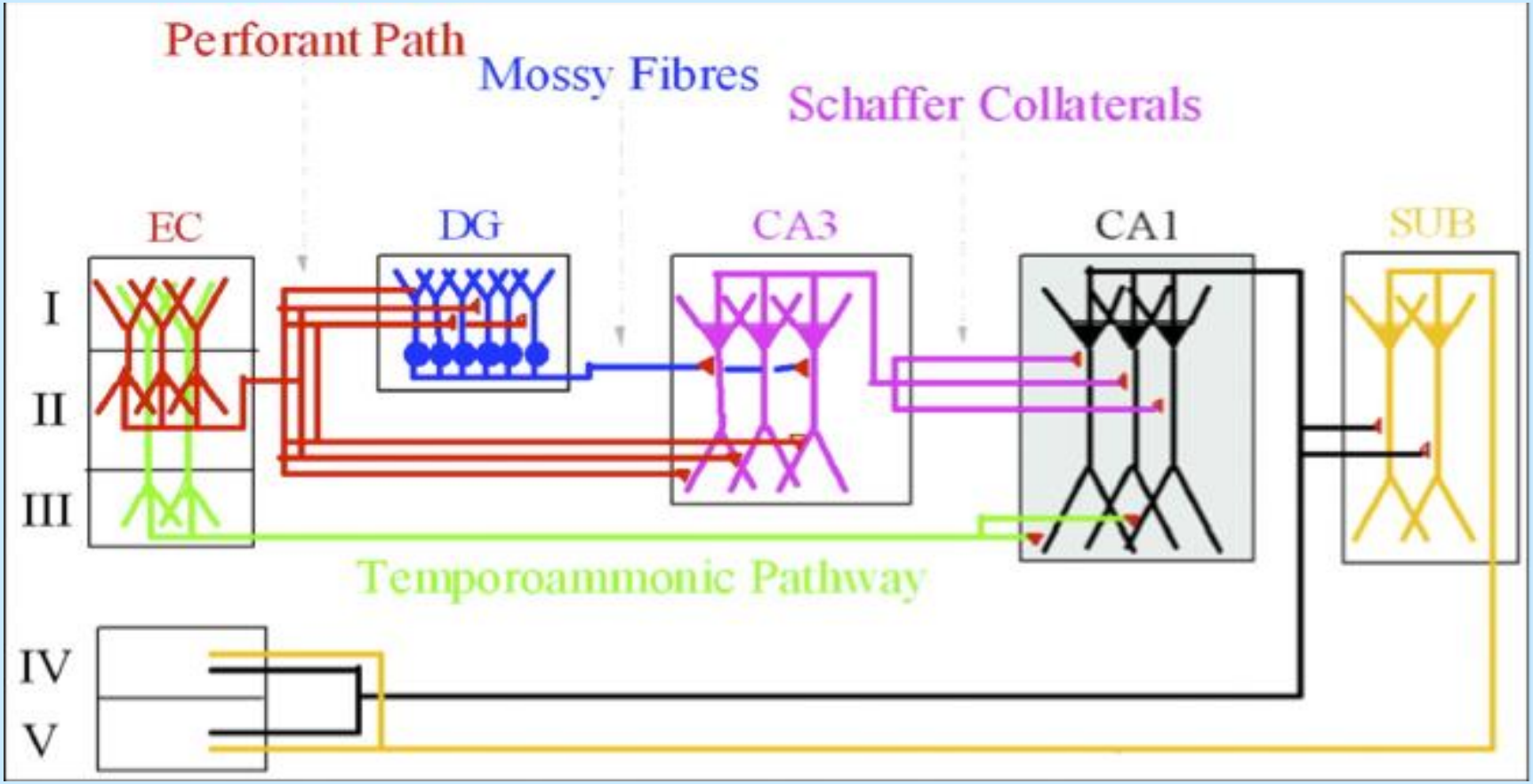
AJ Sedgewick



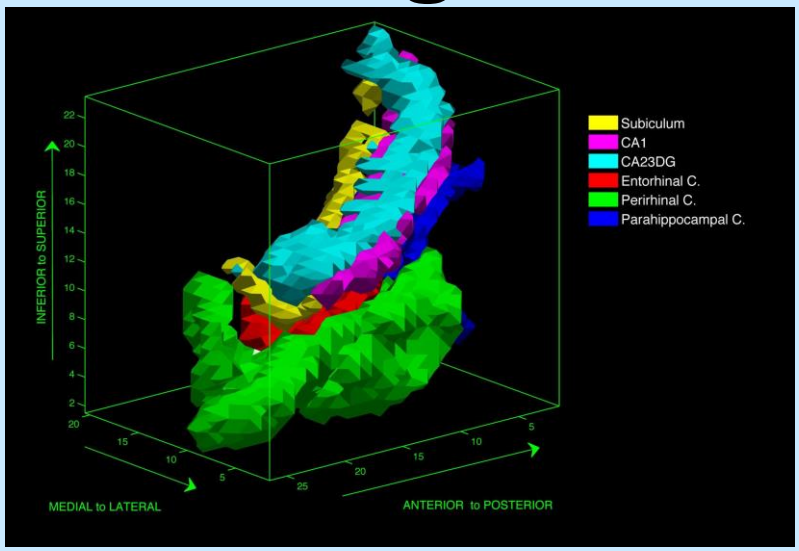
Inferring Unobserved Intermediate Causes—Between Somatic Gene Anomalies and mRNA Transcripts

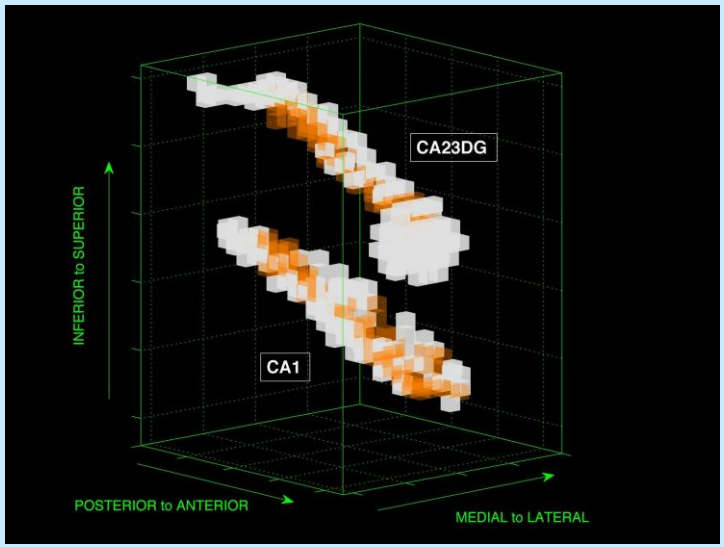


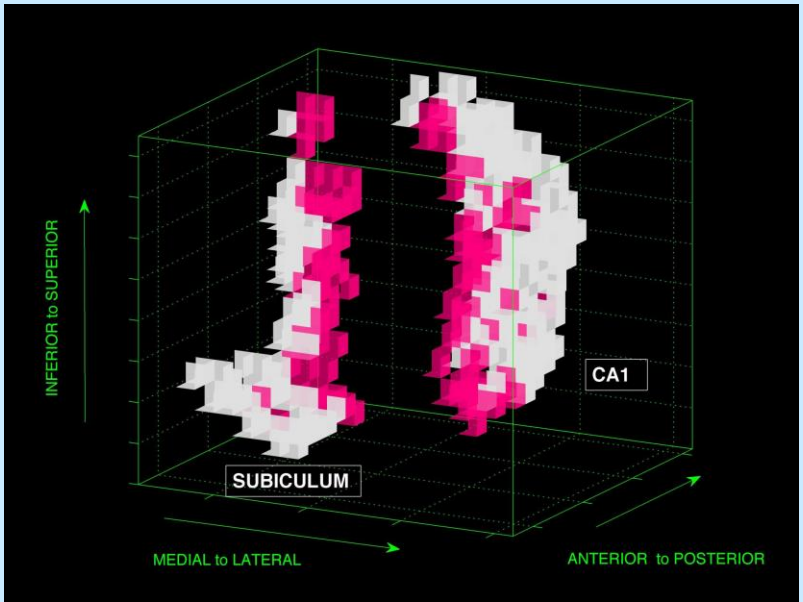
Hippocampal Pathways

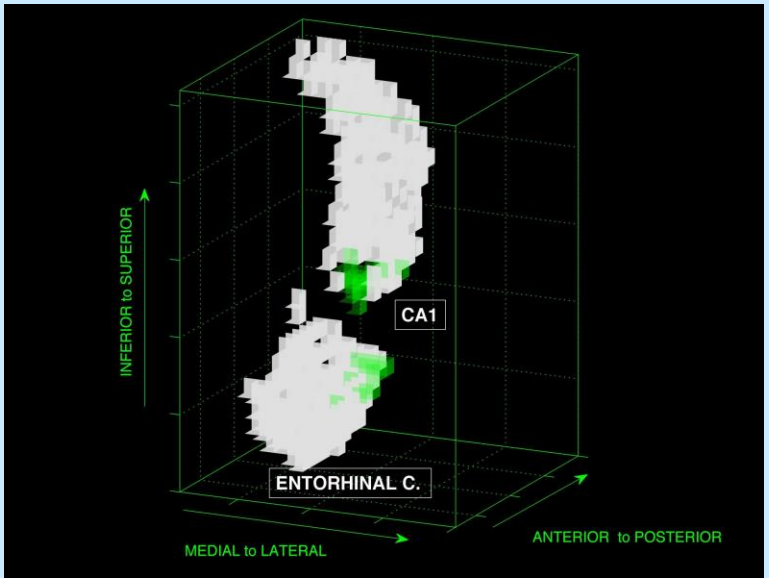


Identification of Communicating Sub-Regions of the Hippocampus

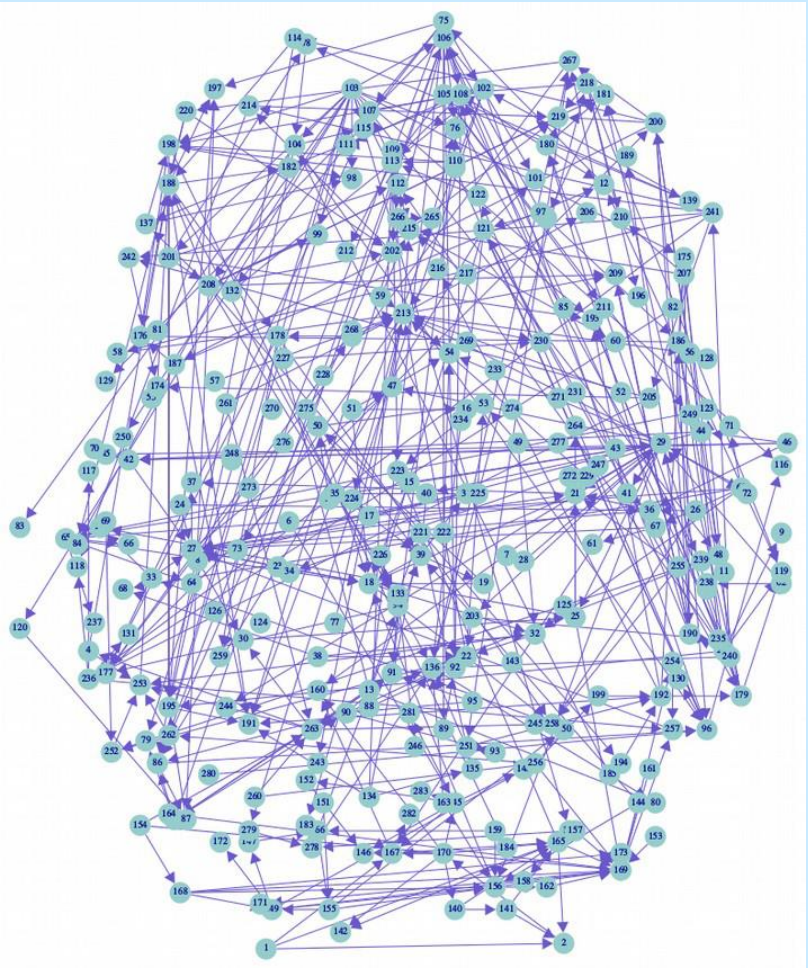








Distinguishing Autistics by Graphical Search with fMRI



Predicted	Actual	
	NT	AT
NT	.92	.01
AT	.08	.99

The Big Task for Big Data

- Showing that we can use big biomedical data to discover novel, important causal relations that can be experimentally confirmed.