

DSL Research in Causal Discovery

Constantin F. Aliferis M.D., Ph.D.

10-4-2002

Goals

- What is causality?
- Why causality is important?
- What is the main framework for causal discovery in biomedicine?
- What are heuristics for large-scale causal discovery?
- What are formal methods for causal discovery from observational data?
- Global vs local methods

Challenge: Define what it means
to say: “A causes B”

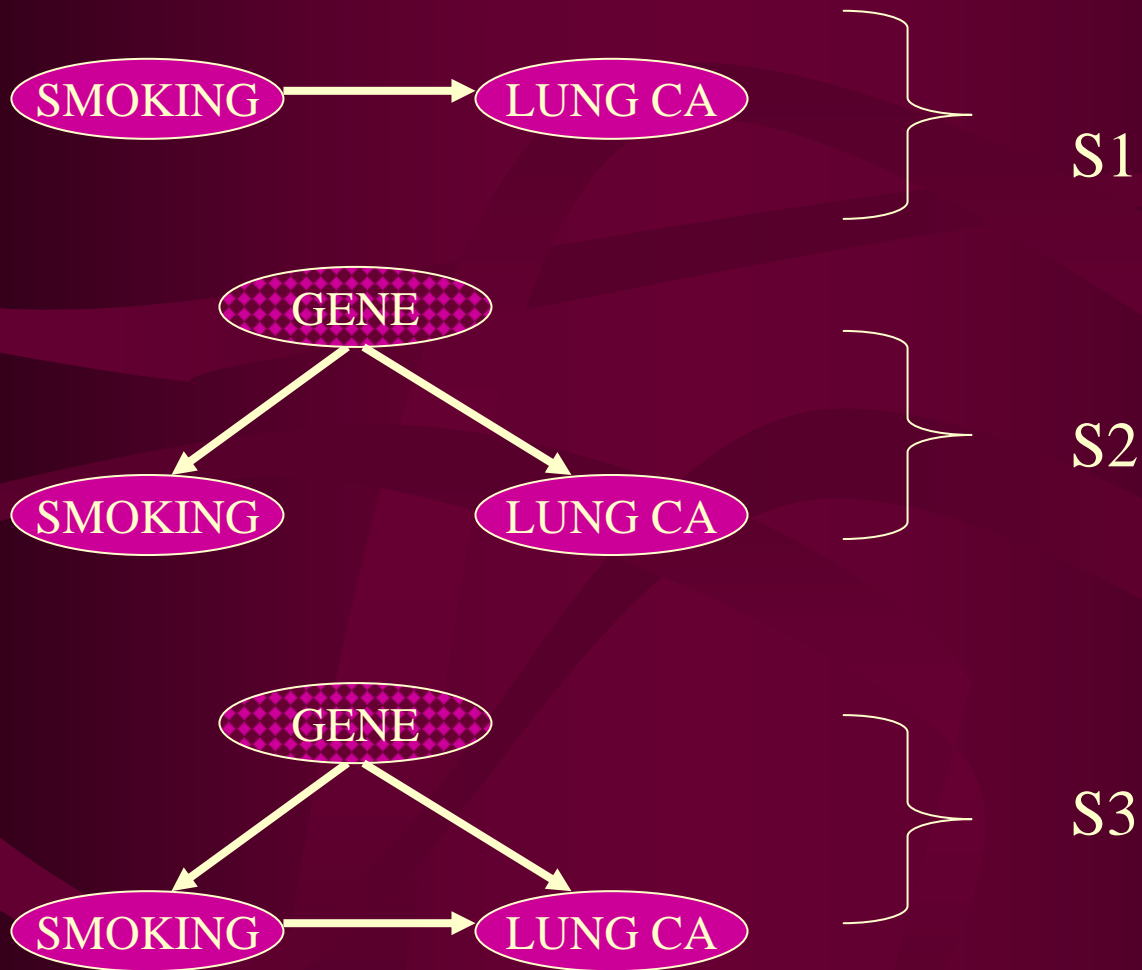
Why causal knowledge is important

- E.g., “does smoking cause Lung Cancer?”
- What are the implications of answering this question?
- If A does not cause lung cancer but is statistically associated with it, can it be useful for something? – give examples

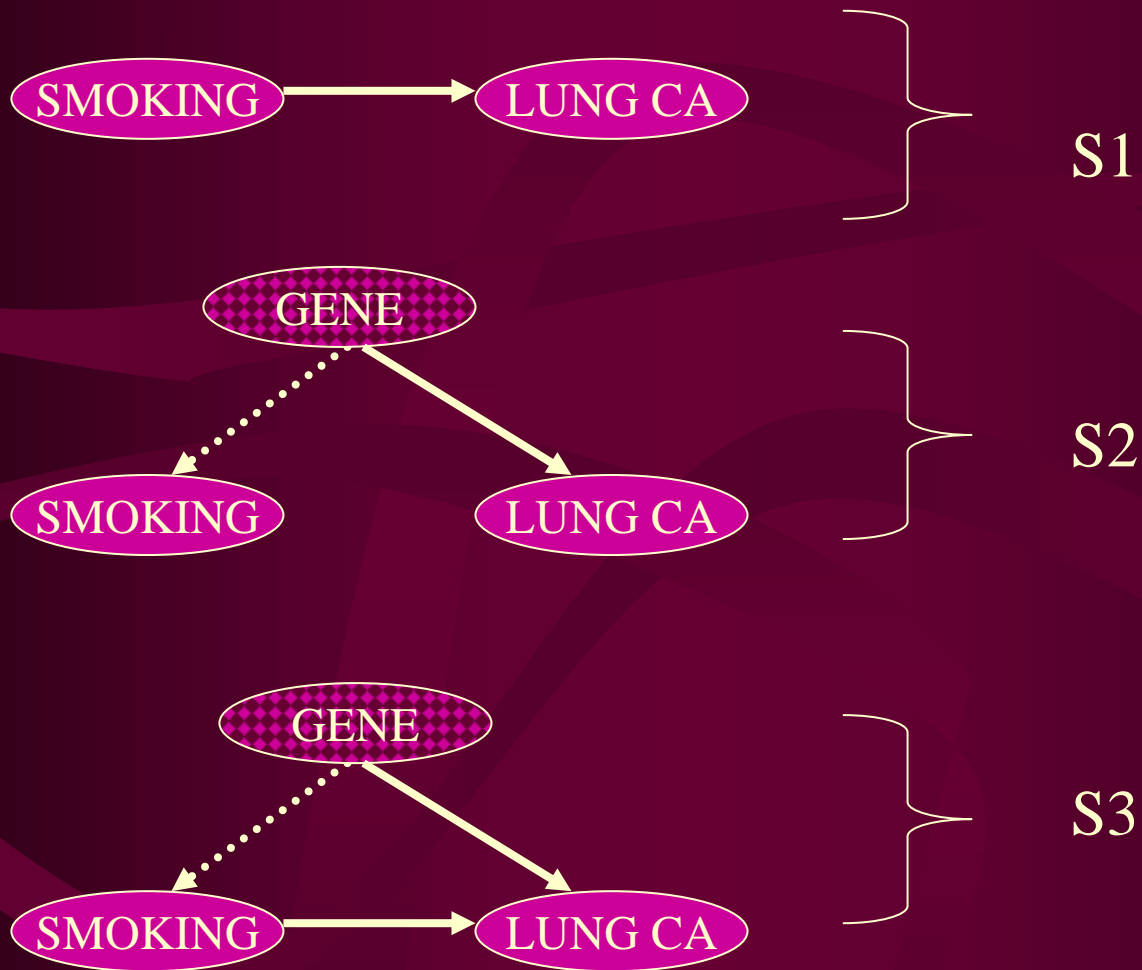
Causation and Association

- What is the relationship between the two?
- If A causes B are A, and B always associated?
- If A is associated with B are they always causes or effects of each other? (directly?, indirectly?, conditionally, unconditionally?)

Statistical Distinguishability



RANDOMIZED CONTROLLED TRIALS



RCTs *Are not* always feasible!

- Why?

Importance of Causal Discovery Today

- Questions:
 - what SNP combination causes what disease
 - how genes and proteins are organized in complex causal regulatory networks
 - how behaviour causes disease
 - how genotype causes differences in response to treatment
 - how the environment modifies or even supersedes the normal causal function of genes

How Can we do large-scale causal discovery without RCTs?

- Heuristics to the rescue...
- What is a heuristic?

Causal Heuristic #1

- Surgeon's General's "Epidemiological Criteria for Causality" [Surgeon General of the United States 1964]: *A* is causing *B* with high likelihood if:
 - (i) *A* precedes *B*;
 - (ii) *A* is strongly associated with *B*;
 - (iii) *A* is consistently associated with *B* in a variety of research studies, populations, and settings;
 - (iv) *A* is the only available explanation for *B* ("coherence");
 - (v) *A* is specifically associated with *B* (but with few other factors).

Causal Heuristic #2

‘If A is a robust and strong predictor of T then A is likely a cause of T ’

- Example: Feature selection
- Example: Predictive Rules

Causal Heuristic #3

- ‘The closer A and T are in a causal sense, the stronger their correlation’ (localizes causality as well)

Causal Heuristic #4

‘If they cluster together they have similar or related function’.

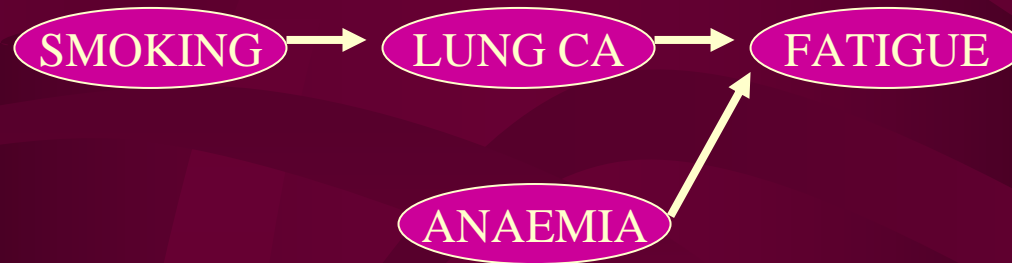
Specific examples: Causal claims expressed in a causally-neutral language

From a study of rule learning in epidemiology:

“...these classifiers...can guide them (researchers) in the formulation of plausible epidemiologic hypotheses”.

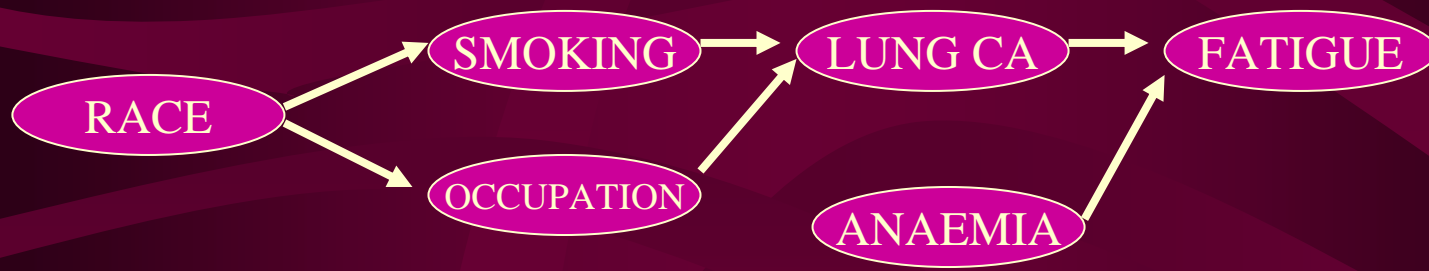
A question to ask is: what is exactly the nature of an “epidemiological hypothesis”? Armitage et al state that “*Epidemiology is concerned with the distribution of disease, or of a physiological condition, and of the factors that influence the distribution*”. Can we presume that these factors are causal factors? Is there any other way to influence things?

Strong predictive rules (or other classifiers):
good for causal discovery?



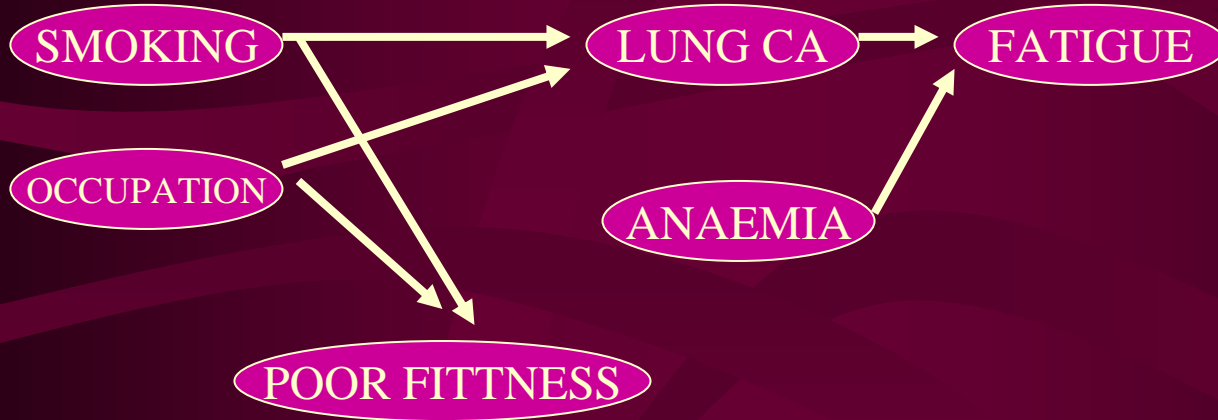
MOST PREDICTIVE SET: SMOKING, FATIGUE, ANAEMIA

Strong predictive rules (or other classifiers):
good for causal discovery?



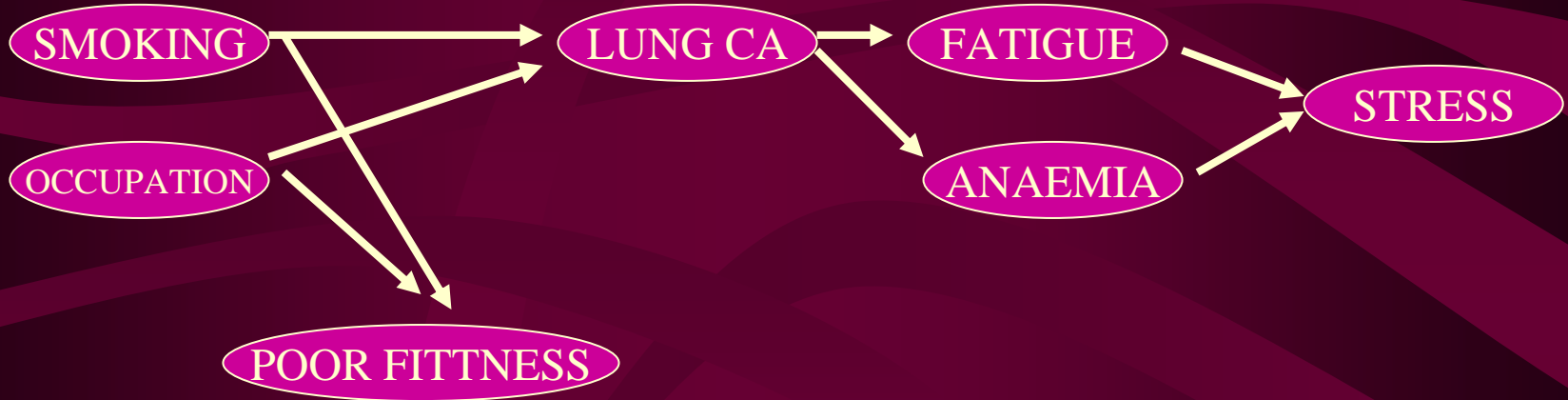
MOST PREDICTIVE SET: RACE, FATIGUE, ANAEMIA

Strong predictive rules (or other classifiers):
good for causal discovery?



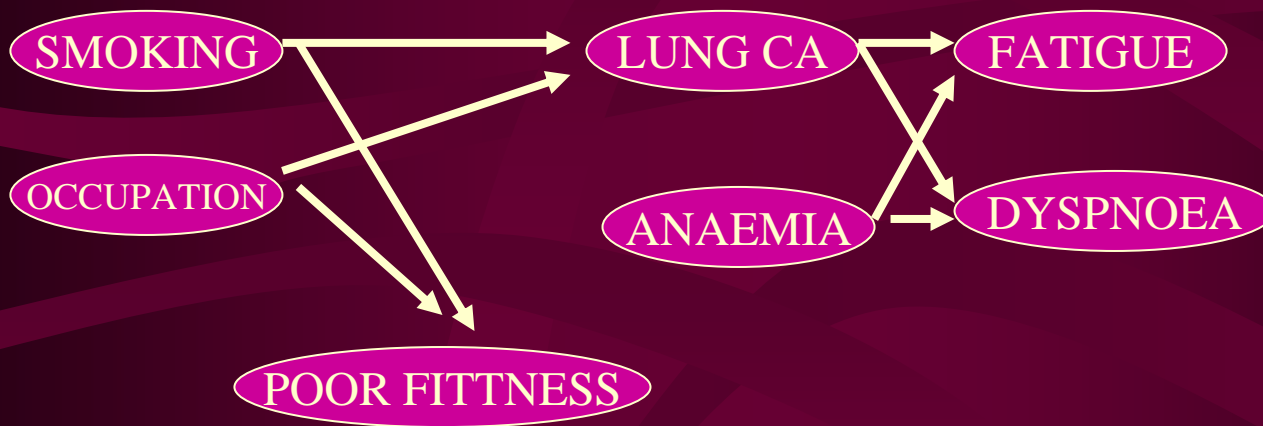
MOST PREDICTIVE SET: POOR FITNESS, FATIGUE, ANAEMIA

Strong predictive rules (or other classifiers):
good for causal discovery?



MOST PREDICTIVE SET: POOR FITNESS, STRESS

Strong predictive rules (or other classifiers):
good for causal discovery?



MOST PREDICTIVE SET: POOR FITNESS, ANAEMIA

Specific Examples: fuzzy causal claims

from the field of bioinformatics: the authors of a paper apply a heuristic search variable selection method for classification. An interesting aspect of their study is the *causal* interpretations of the strongest and most consistent gene predictors of leukemia histological subtypes:

“...*The CD2...plays an important role in mediating the interactions between human T lymphocytes and accessory cells...Blk may play an important role in B cell proliferation...Immunoglobulin-associated beta (B29), ..., belongs to family of surface adhesion molecules.*”

Specific Examples: fuzzy causal claims

Eisen et al 1998 state:

“Genes of similar function cluster together. ...strong tendency for these genes to share common roles in cellular processes”.

What is the basis of clustering? What is the problem with transitivity and cluster size?

Specific Examples: strength of association and closeness

Zhou et al, 2002:

“It is clear that in a biological pathway a gene is likely to show strong correlations with its neighbour genes, but not with genes that lie far apart in the pathway”.

what is a pathway?

what is a neighbour?

is this true?

OK. Maybe we should abandon the quest for causation then...

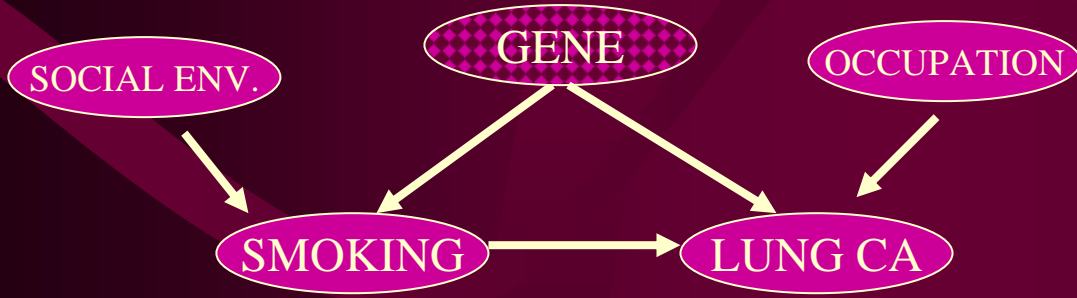
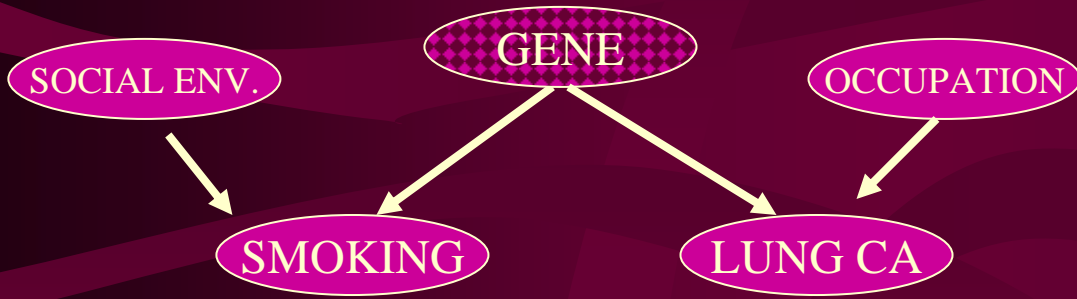
Spirtes et al 2000:

“The...looseness of causal claims has provided a reason for many people to dismiss the very idea of causation as prescientific. ...The skeptic about causality pushes the brake pedal to make his car slow, flips a switch to make a lamp glow, puts his money in the bank to collect interest.”

Conclusions so far

- Causal heuristics are unreliable
- Causation is difficult to define
- RCTs are good but not perfect and not always doable
- Causal heuristics are not reliable
- Major causal knowledge does not have RCT backing!

Is there a way around Fisher's problem?



$I(SE;LC|SM)?$

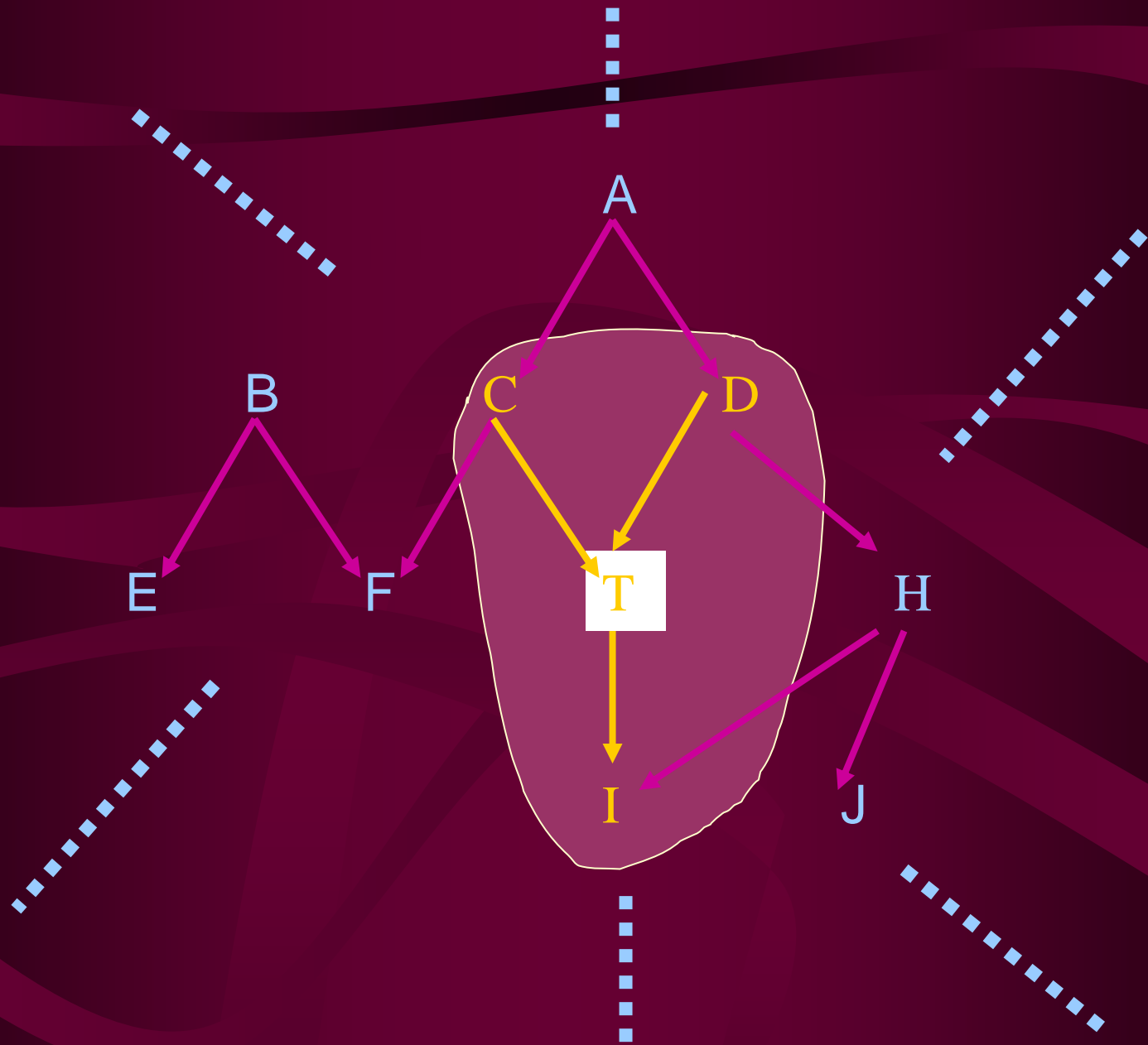
$I(SM;O)?$

Algorithms For Sound Causal Discovery From Observational data

- They do exist!
- They are based on a graphical-probabilistic language called “Causal Probabilistic Networks (a.k.a. “Causal Bayesian Networks”)
- They infer the full network of relationships
- They are reliable but intractable for $>$ a few hundred variables

Dealing With Scale

- But in biomedicine we have thousands or even hundred of thousands of variables!
- Solution: learn *local* portions of the causal network: direct causes and effects, or indirect up to degree k



Applications

- Learn gene and protein pathways
- Learn immediate causes and effects of diseases, histopathology types, clinical outcomes...
- Learn optimal predictor sets (more of this next time we meet...)
- Read tech report for more details...