Causal Inference:
prediction, explanation, and intervention

Lecture 7: Causality in time series (part II)

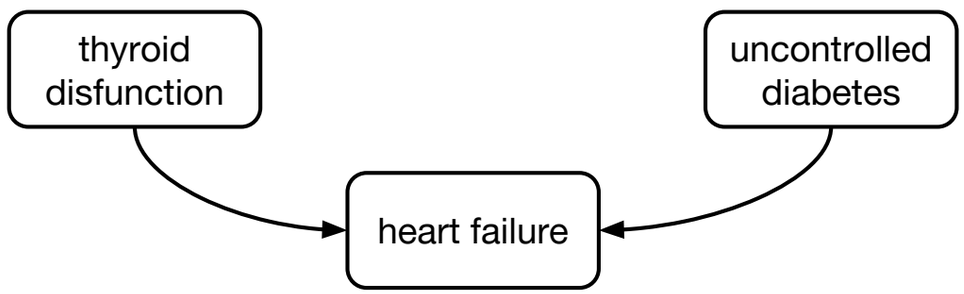
Samantha Kleinberg
samantha.kleinberg@stevens.edu

Upcoming events

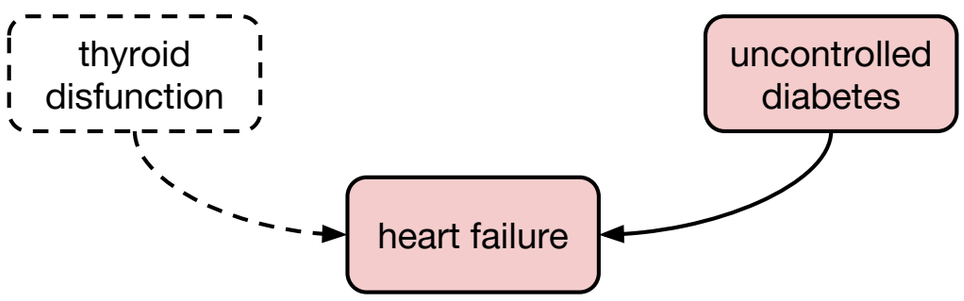
- Next week: midterm
- Two weeks: project proposals due

Explanation and Inference

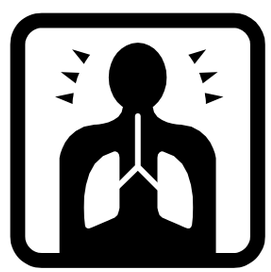
- Causal inference operates on the type-level



- Causal explanation explains particular events



Timing



Intuition

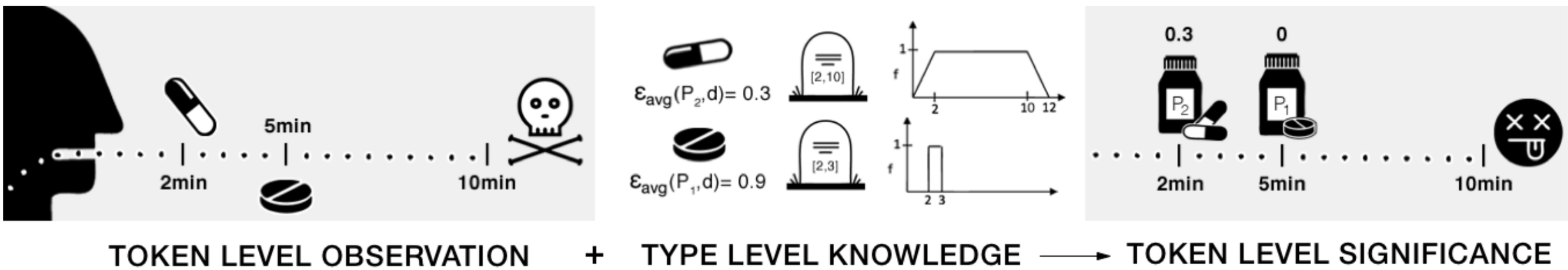
- Type level relationships are candidates
- Strictly applying type-level relationships is problematic
- Without background knowledge that timing is a strong constraint, do not want this to strictly limit explanations

Goals

- Quantify significance of an explanation
  - Allows for incomplete knowledge
  - Incorporates temporal uncertainty
- Algorithm for finding the significance of explanations

General approach

- Infer causal relationships from data
- Observe token sequence
  - E.g. Series of medical tests or continuous recordings
- Iterate over relationships to calculate significance for each explanation
- Result is ranking of possible explanations


5min
2min
10min
$\varepsilon_{avg}(P_2,d)= 0.3$
$\varepsilon_{avg}(P_1,d)= 0.9$
[2,10]
[2,3]
1
f
2
10 12
2 3
0.3
0
P_2
P_1
2min
5min
10min
TOKEN LEVEL OBSERVATION
+
TYPE LEVEL KNOWLEDGE
TOKEN LEVEL SIGNIFICANCE

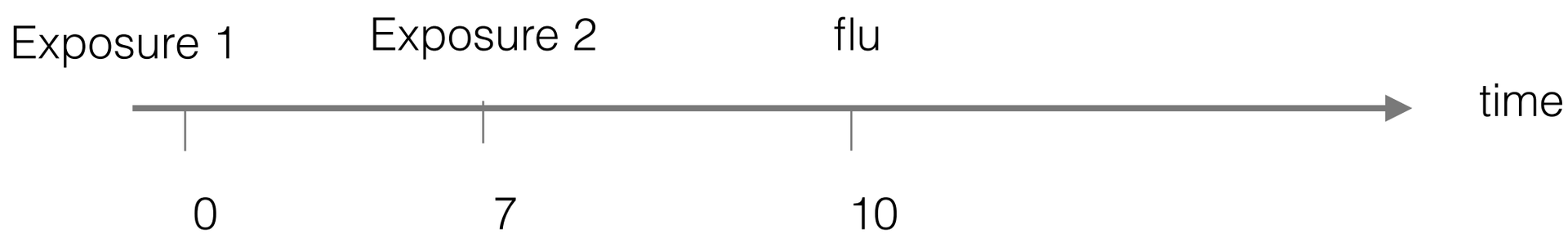
- Relax link between type and token: weight how strongly token event matches type-level knowledge
- Assess significance for each explanation

$$S(c_{t'}, e_t) = \varepsilon_{avg}(c_{r-s}, e) \times P(c_{t'}|\mathcal{V}) \times f(c_{t'}, e_t, r, s)$$

- Propose hypotheses for unexplained events

M. Zheng and S. Kleinberg, FLAIRS 2017

Significance of hypotheses



- When cause and effect occur consistent with type-level information
  - Significance = type-level significance
- When timing differs
  - Weight by difference
- When unknown if c occurs
  - Weight by probability of occurrence

Recall connecting principle!

Assessing significance

With $c \overset{\geq r, \leq s}{\underset{\geq p}{\sim}} e$

$$S(c_{t'}, e_t) = \varepsilon_{avg}(c,e) \times P(c_{t'} \mid V) \times f(t', t, r, s)$$

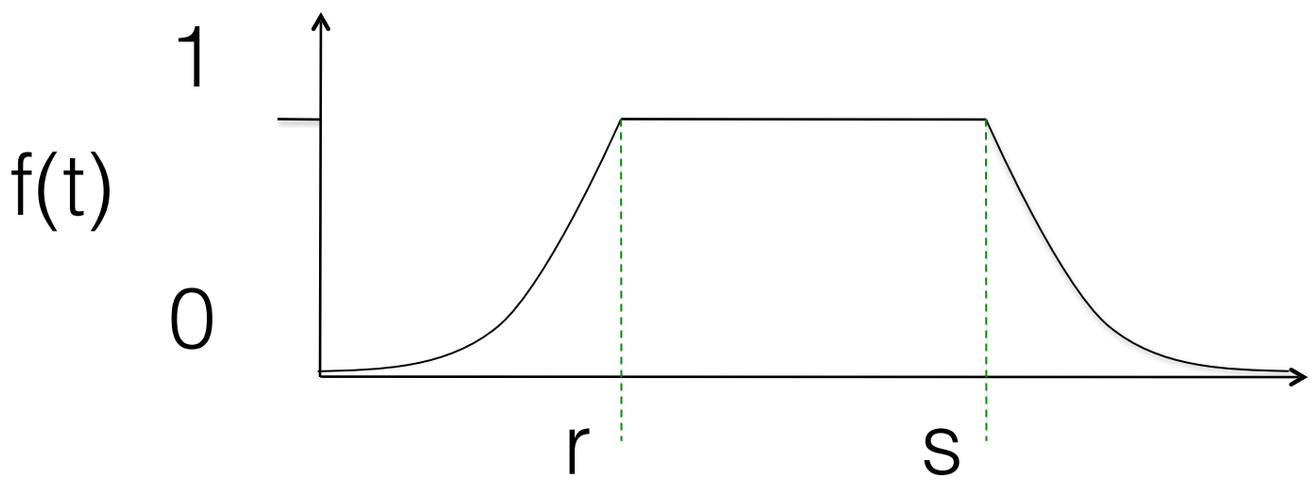
=1 if occurs at time t'

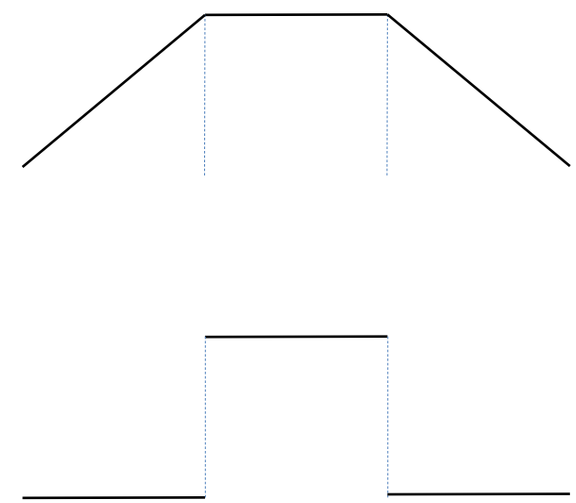
=1 if c in window [r,s]

V is observation sequence

Accounting for uncertainty

$$S(c_{t'}, e_t) = \varepsilon_{avg}(c,e) \times P(c_{t'} \mid V) \times f(t', t, r, s)$$





- Slope may be dependent on underlying mechanisms
  - i.e. when is cause still capable of producing effect
- Can learn function from data (see FLAIRS 2017)

Calculating P

$$S(c_{t'}, e_t) = \varepsilon_{avg}(c, e) \times P(c_{t'} \mid V) \times f(t', t, r, s)$$

- Where do probabilities come from?
  - Initial dataset
  - Model
- Conditional on token-level observations

Algorithm

1. $X \leftarrow$ type-level genuine and just so causes of e

   $\mathcal{V} \leftarrow [\mathcal{V}_0, \mathcal{V}_1 \ldots \mathcal{V}_t]$

   $EX \leftarrow \emptyset$

2. For each $x \in X$, where $x = y \rightsquigarrow^{\geqslant r, \leqslant s} e$, for $i \in [t-s, t-r]$ if $\mathcal{V}_i \models y$ the significance of y_i for e_t is $\varepsilon_{avg}(y_{r-s}, e)$ and $EX = EX \bigcup \{x\}$

3. For each $x \in X \setminus EX$, where $x = y \rightsquigarrow^{\geqslant r, \leqslant s} e$, and

   $$i = \underset{j \in [0,t)}{\operatorname{argmax}}(P(y_j|\mathcal{V}) \times f(j, t, r, s)),$$

   the significance of y_i for e_t is

   $$S(y_i, e_t) = \varepsilon_{avg}(y_{r-s}, e) \times P(y_i|\mathcal{V}) \times f(i, t, r, s).$$

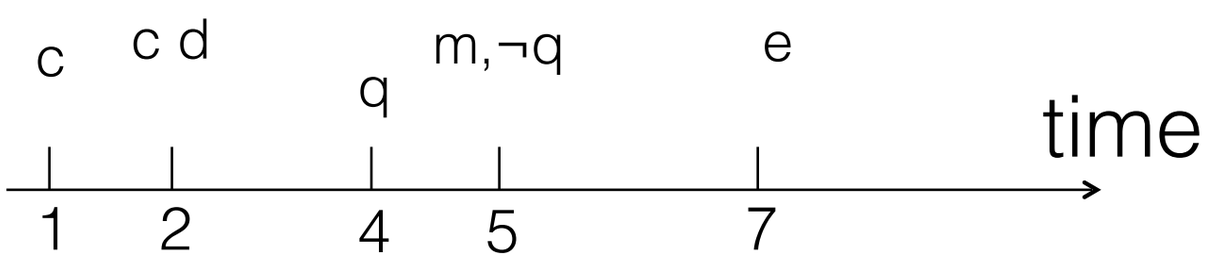
Example of procedure

- Type-level relationships (X)

$$c \wedge d \overset{\geq 3, \leq 5}{\sim} e \qquad \varepsilon_{avg}(c \wedge d, e)$$

$$m \overset{\geq 1, \leq 1}{\sim} e \qquad \varepsilon_{avg}(m, e)$$

- Token observation (V)



Example of procedure

- Type-level relationships (X)

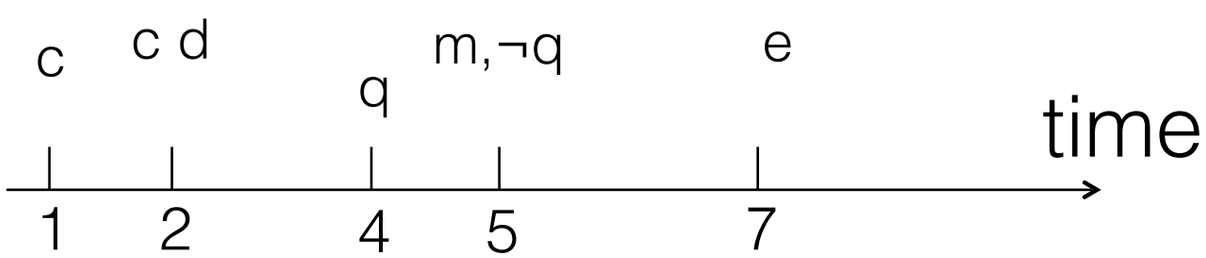
$$c \wedge d \overset{\geq 3, \leq 5}{\sim} e \qquad \varepsilon_{avg}(c \wedge d, e)$$

$$m \overset{\geq 1, \leq 1}{\sim} e \qquad \varepsilon_{avg}(m, e)$$

$$\varepsilon_{avg}(m, e) \times f(5,7,1,1)$$

- Token observation (V)

$$\varepsilon_{avg}(m, e) \times P(m_6 \mid V)$$

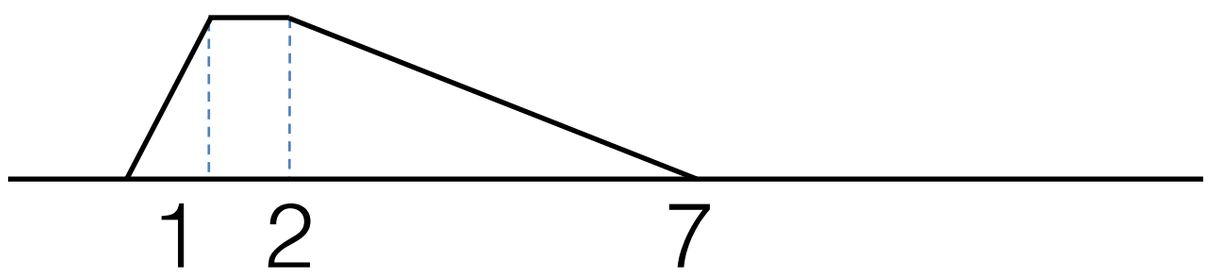


Example: multiple instances

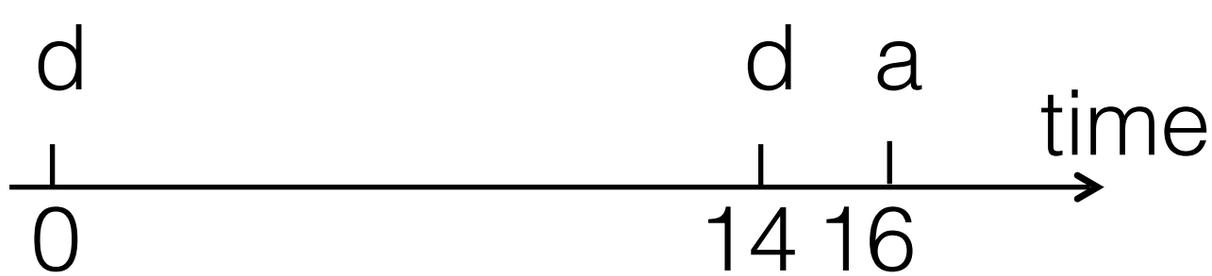
- Type-level relationships (X), drug (d), adverse event (a)

$$d \overset{\geq 1, \leq 2}{\rightsquigarrow} a$$

- f

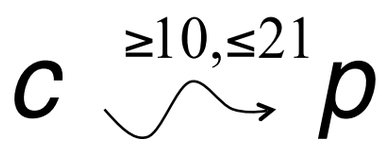


- Token observation (V)

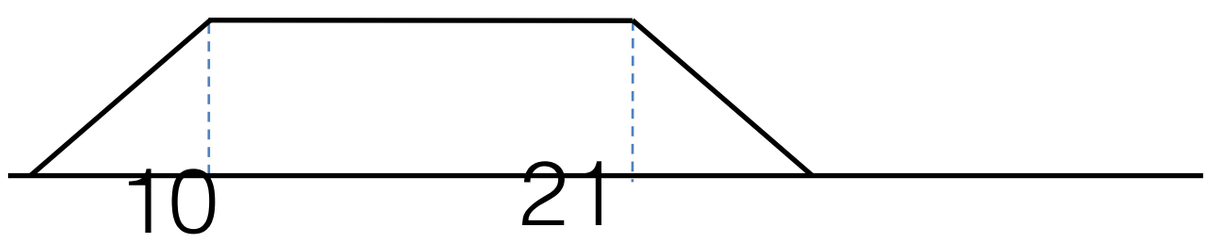


Example: Overdetermination

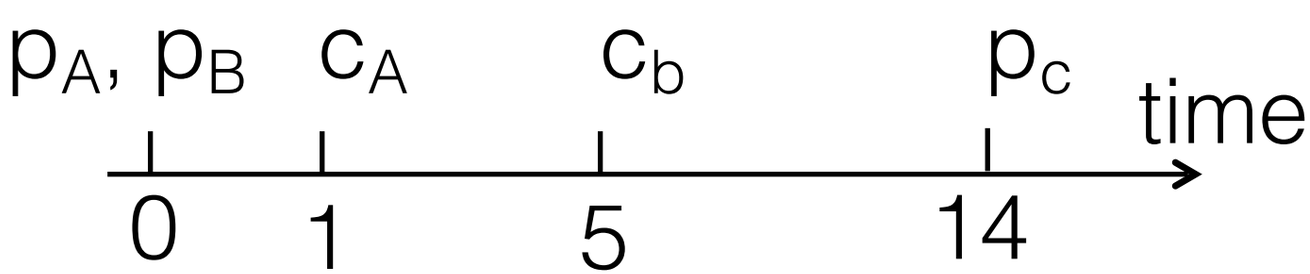
- Type-level relationships (X), contact(c), chickenpox(a)



- f



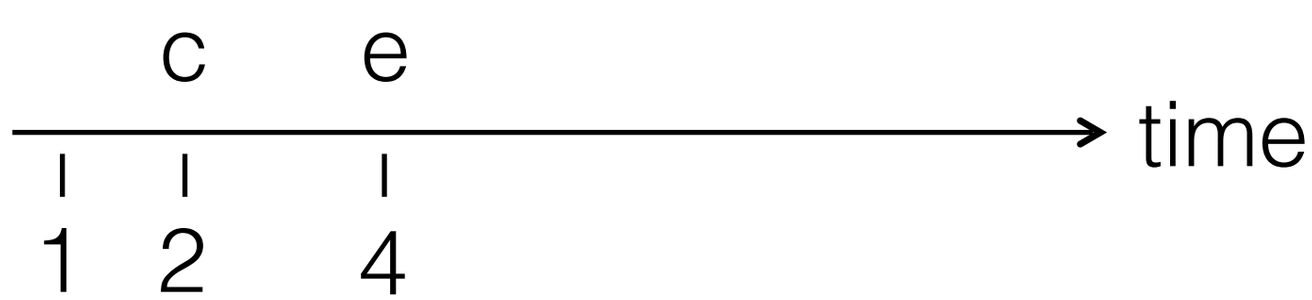
- Token observation (V)



Example: Transitivity

$$c \overset{\geq 1, \leq 2}{\rightsquigarrow} d$$

$$d \overset{\geq 1, \leq 1}{\rightsquigarrow} e$$

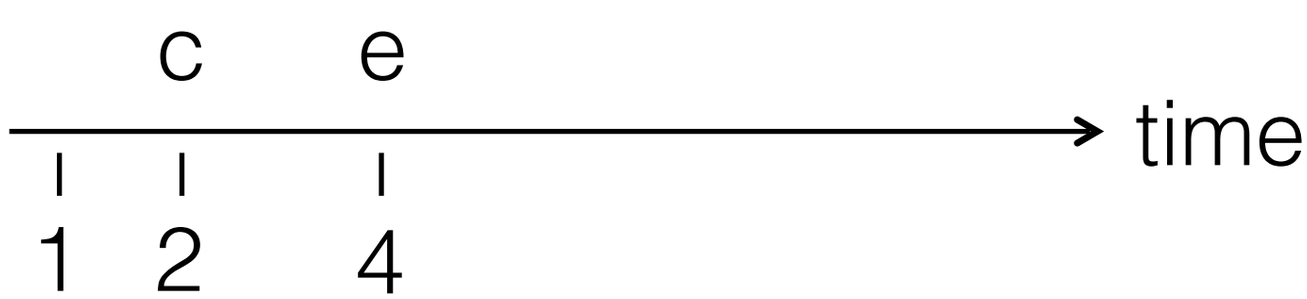


Example: Transitivity

$$c \overset{\geq 1, \leq 2}{\rightsquigarrow} d$$

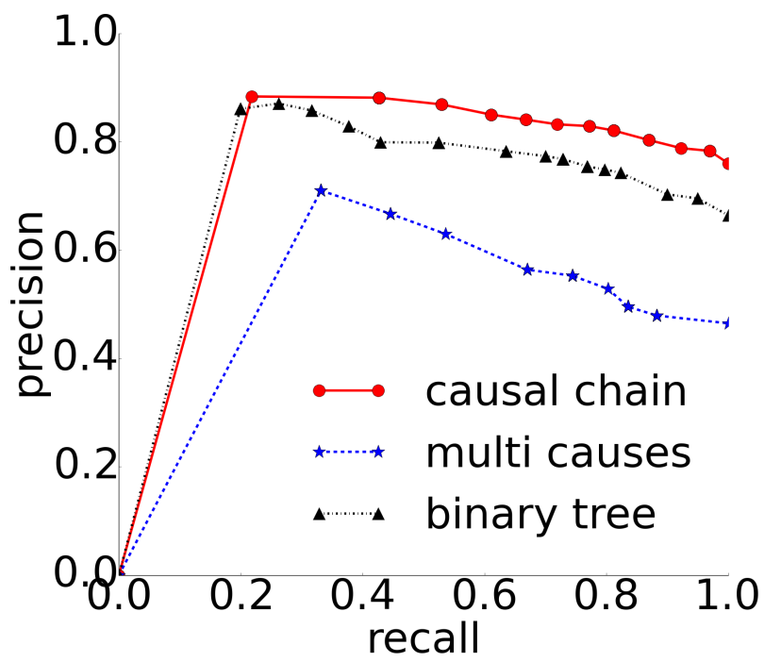
$$d \overset{\geq 1, \leq 1}{\rightsquigarrow} e$$

$$\varepsilon_{avg}(d,e) \times P(d_3 \mid c_2, e_4) \times 1$$

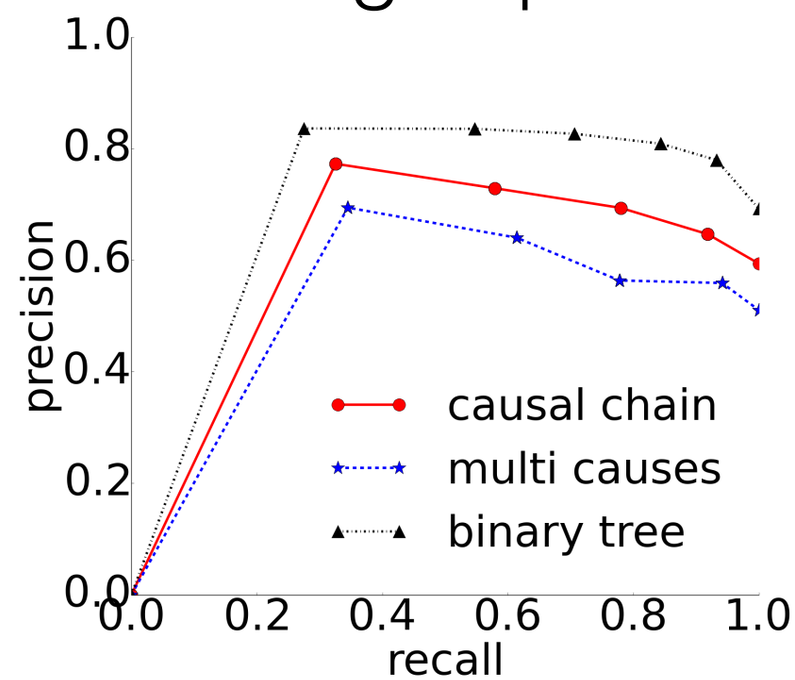


Key results

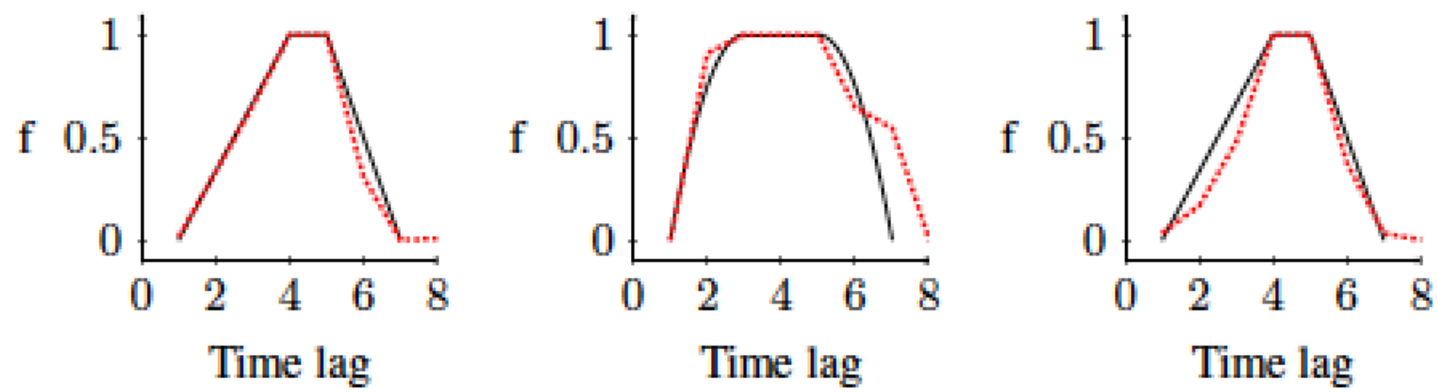
Using type-level knowledge



Discovering explanations



- On UCI bike sharing found 95% of rental increases caused by good weather, 67% of decreases by poor weather



(a) RMSD=0.070 (b) RMSD=0.111 (c) RMSD=0.042

Wiener

Causality amounts to increase in predictability

Wiener, N. (1956). The theory of prediction. Modern Mathematics for Engineers. New York: McGraw-Hill, 165-190.

Granger causality

Basic idea:

Information contained in cause that helps predict effect (which happens after) – that is not contained in other variables

Granger causality - definition

$X_1(t)$ value of X_1 at time t

$X_1^*(t)$ values of X_1 <u>up to</u> time t

$W^*(t)$ <u>all knowledge</u> up to time t

X_2 Granger-causes X_1 if

$$P(X_1(t+1) \mid W^*(t)) \neq P(X_1(t+1) \mid W^*(t) - X_2^*(t))$$

Feedback

X_2 Granger-causes X_1 if

$$P(X_1(t+1) \mid W^*(t)) \neq P(X_1(t+1) \mid W^*(t) - X_2^*(t))$$

Can have X_1 causes X_2 and X_2 causes X_1

Theory versus practice

- Ideal: use all variables over infinitely long timescale
- Reality: limited data, limited computational power necessitate simplifications

bivarate vs. multivariate

- Bivariate: $W = X_1, X_2$
- Multivariate: W = set of variables

Testing for Granger-causality

X_2 Granger-causes X_1 if

$$P(X_1(t+1) \mid W^*(t)) \neq P(X_1(t+1) \mid W^*(t) - X_2^*(t))$$

Basic approach: regression of X_1 on W. If X_2 terms have nonzero coefficients, then they provide info on X_1.

Vector autoregression (VAR)

$$X(t) = A(1)X(t-1) + A(2)X(t-2)\ldots A(m)X(t-m) + \varepsilon(t)$$

m = model order (how many past values are included)

bivariate VAR

$$X_1(t) = \sum_{j=1}^{m} A_{11}(j) X_1(t-j) + \sum_{j=1}^{m} A_{12}(j) X_2(t-j) + \varepsilon_1(t)$$

$$X_1(t) = \sum_{j=1}^{m} A_{11}(j) X_1(t-j) + \sum_{j=1}^{m} A_{12}(j) X_2(t-j) + \varepsilon_1(t)$$

Testing if X_2 causes X_1

-null hypothesis is that $A_{12}=0$ (no causality)

-test if terms significantly differ from zero

-test if epsilon is reduced with inclusion of X_2

Common cause of two effects

X causes both Y and Z

$$Y(t) = \sum_{j=1}^{m} A_{11}(j)Y(t-j) + \sum_{j=1}^{m} A_{12}(j)X(t-j) + \varepsilon_1(t)$$

$$Y(t) = \sum_{j=1}^{m} A_{11}(j)Y(t-j) + \sum_{j=1}^{m} A_{12}(j)Z(t-j) + \varepsilon_1(t)$$

Causal Chain

$$Z(t) = \sum_{j=1}^{m} A_{11}(j) Z(t-j) + \sum_{j=1}^{m} A_{12}(j) X(t-j) + \varepsilon_1(t)$$

X causes Y, Y causes Z

Directionality

Probability raising is symmetric. Is Granger causality?

Multivariate VAR

$$X_V(t) = \sum_{j=1}^{m} A(j) \times X_V(t-j) + \varepsilon_V(t)$$

X_V=vector of size V (values of the V variables)

A=VxV matrix

Complexity

For multivariate:

X is vector of size V (V=number of variables)

A is VxV matrix

m lags means sum over m VxV matrices

For bivariate:

2x2 instead of VxV

Chickens and eggs: Part 1 - chickens

Part 1: Did the Chicken Come First?

The following equation was estimated by OLS:

$$Eggs_t = \mu = \sum_{i=1}^{L} \alpha_i \, Eggs_{t-i} = \sum_{i=1}^{L} \beta_i \, Chickens_{t-i} + \epsilon_t;$$

$H_o : \beta_1 = \ldots \beta_L = 0$ (chickens do not Granger cause eggs).

L = no. of lags	F-statistic	P-value	R^2 of the regression
1	.04	.85	.96
2	1.71	.19	.97
3	1.10	.36	.97
4	.79	.54	.97

Thurman WN, Fisher ME (1988) Chickens, Eggs, and Causality, or Which Came First. American Journal of Agricultural Economics 70(2):237-238.

Chickens and eggs: Part 2 - eggs

Part 2: Did the Egg Come First?

The following equation was estimated by OLS:

$$Chickens_t = \mu + \sum_{i=1}^{L} \alpha_i \, Chickens_{t-i} + \sum_{i=1}^{L} \beta_i \, Eggs_{t-i};$$

$H_0 : \beta_1 = \ldots = \beta_L = 0$ (eggs do not Granger cause chichens).

L = no. of lags	F-statistic	P-value	R^2 of the regression
1	1.23	.27	.73
2	10.36	.0002	.81
3	5.85	.0019	.81
4	4.71	.0032	.82

Granger causality vs. causality

Why is it insufficient?

Common causes

Granger causality vs. causality

Why is it unnecessary?

Can have causality without Granger causality

-Timescale shorter than measured/tested

-Unfaithful distribution (canceling, control)

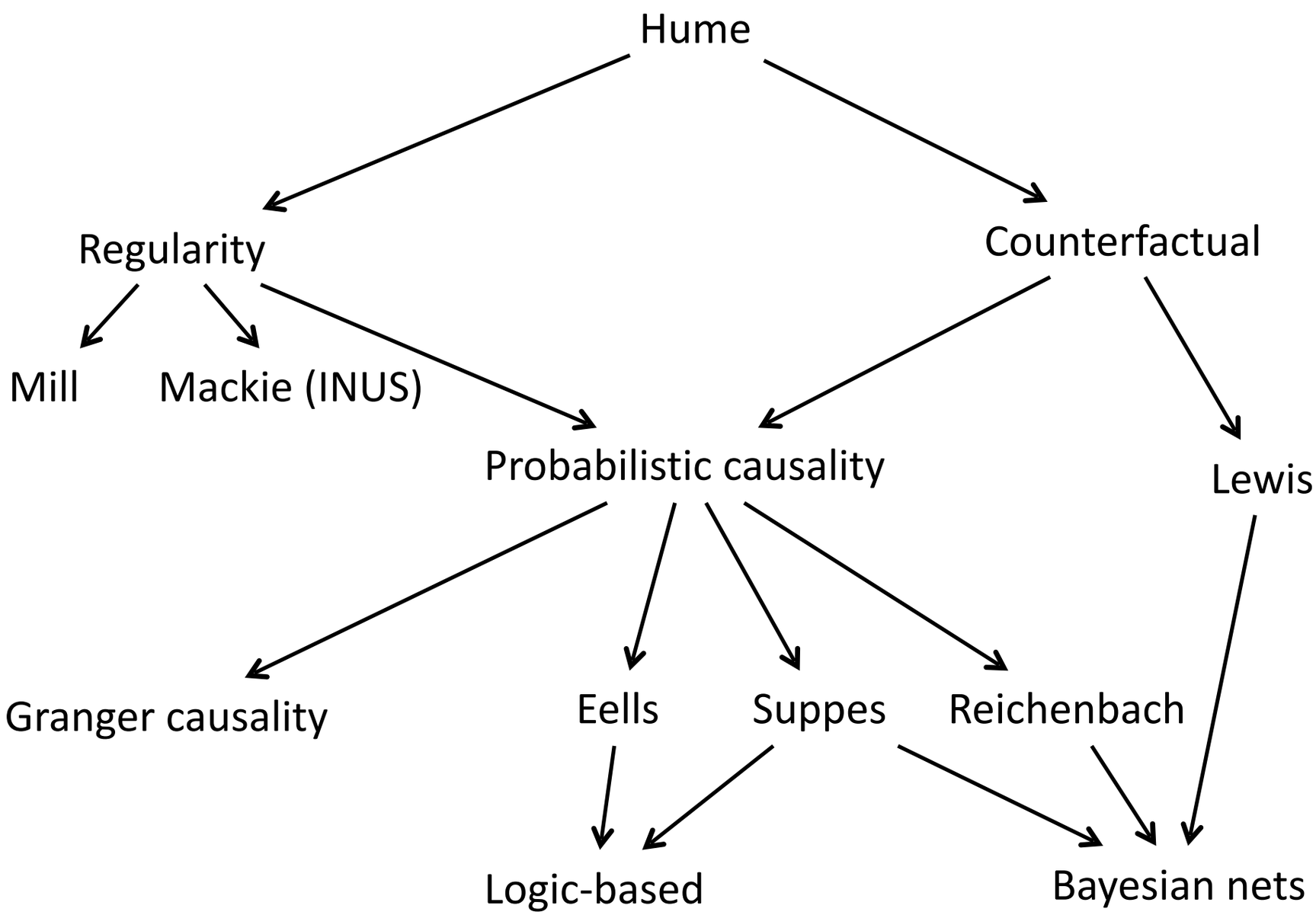
Software

- MSBVAR granger.test: bivariate (R package)
- MVGC Multivariate Granger Causality Toolbox (Matlab toolbox): bivariate and multivariate
  - http://www.sussex.ac.uk/sackler/mvgc/

	BN	**DBN**	**Granger**	**Temporal logic**
Results	Graph	Graph	Relationships	Relationships
Time	No	Set of lags	Single lag	Window
Data	C/D/M	C/D/M	C	D/M
Cycles	No	Yes	Yes	Yes
Latent vars	Yes	Yes	No	No
Prediction	Yes	Yes	No*	No
Token cause	Counterfactual-based	No	No	Probabilistic

*For bivariate

Journal club!



Problems

Nonstationarity

Preemption

Overdetermination

Determinism

Causal chains

Recap

- Concepts of causality
  - Regularities, Counterfactuals, Probabilistic
- Two levels
  - Type and token
- Three methods
  - Graphical models (BN, DBN), logic-based, Granger

Discussion example #1

We have a set of data on bike sharing: for each station, when each bike is rented; weather; miles of nearby bike lanes

1) How can we find what causes bike rentals?
2) How can we predict how many bikes will be rented at a particular station on a particular day?
3) What are limitations of this data? Can other datasets help?

https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset

Discussion example #2

You are a researcher at FooBar Corp, which produces widgets.

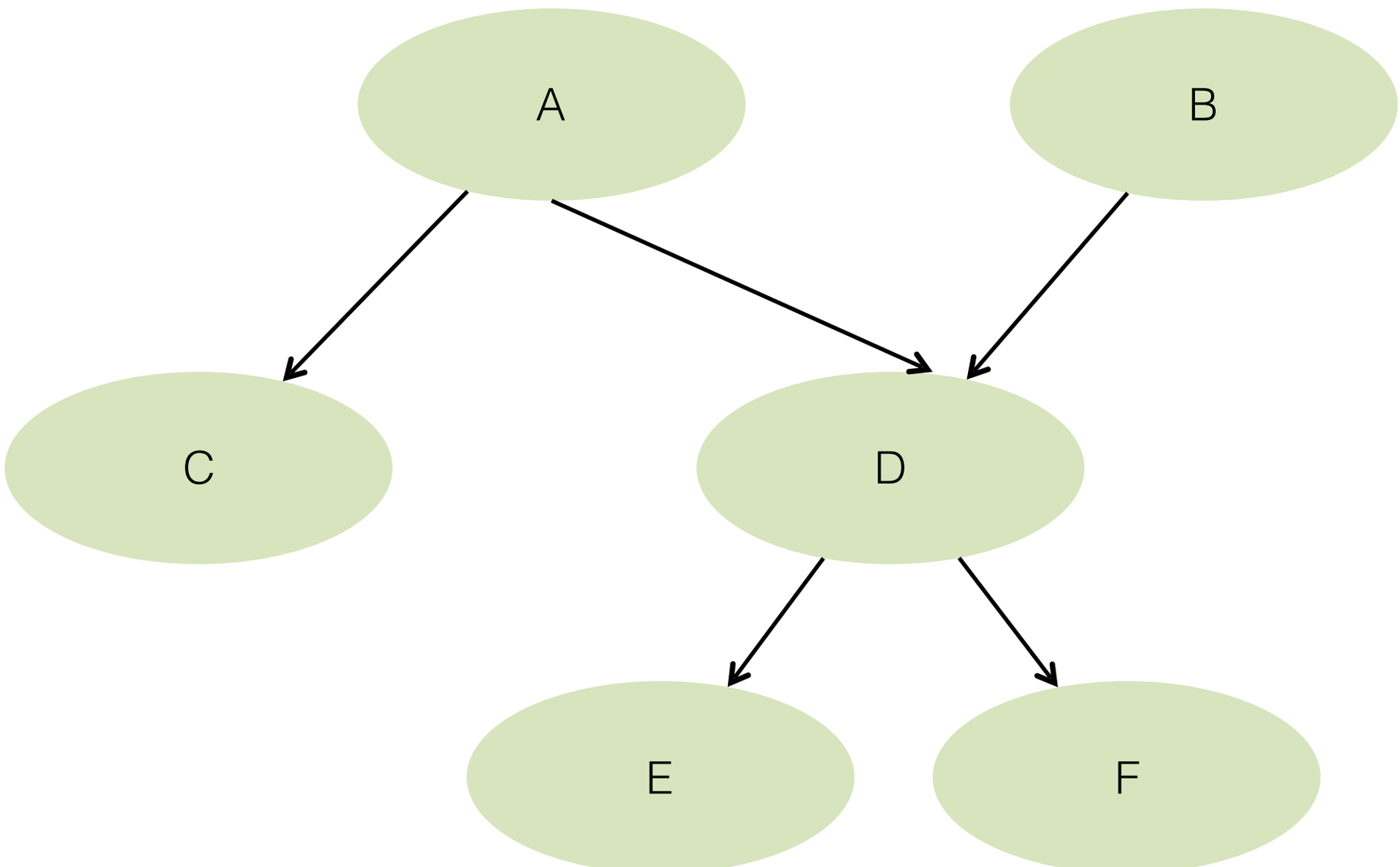
1) They have collected many years of detailed data on their employees and manufacturing process and want to use this data to improve efficiency (defined as the number of widgets that can be produced each hour).

What approach would you use to analyze their data, why would you choose that, and what information would FooBar Corp. need to best use these results?

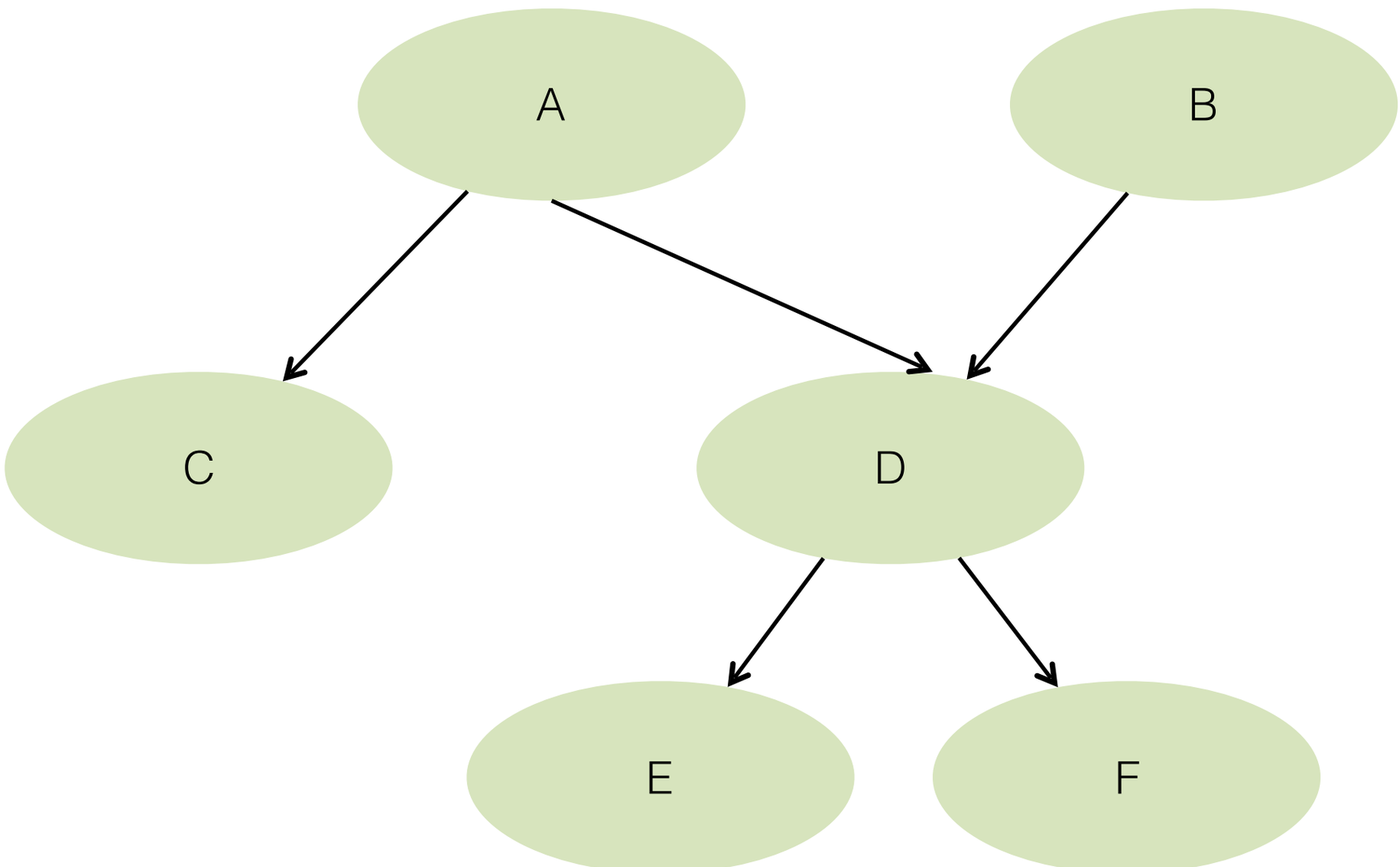
2) If FooBar Corp were willing to collect new data, what information would you seek?

Review questions

$P(E|do(D))?= P(E|D)$



$P(E|D,do(A))?=P(E|D)$



Bob likes multiple choice exams.

$P(H|M)=0.7$

What's the probability an exam is multiple choice if Bob's happy?

$P(H)=0.2$

$P(M)=0.1$

Answer: 0.35

Does an apple a day keep the doctor away?

Data: days apples eaten, # of doctor visits

Why can’t we go directly from type-level knowledge to token-level cases?

Jane’s study group all developed colds. Jane has been taking a vitamin D supplement and didn’t get a cold. Is the supplement responsible?

Types of problems

- Preemption
- Overdetermination
- Nonstationarity
- Indeterminism
- Hidden variables
- Unfaithful distribution
- Transitivity

Problem cases

- Regularity?
- Counterfactual?
- Probabilistic?

What is the key observation that helps us get causes from probabilities?

Learning graphical models (structure)

Two main approaches?

What set of relationships could <u>not</u> be represented with a BN?

What Bayes net(s) is consistent with these independencies?

$A \perp C$

$C \perp D \mid B$

$A \perp D \mid B$

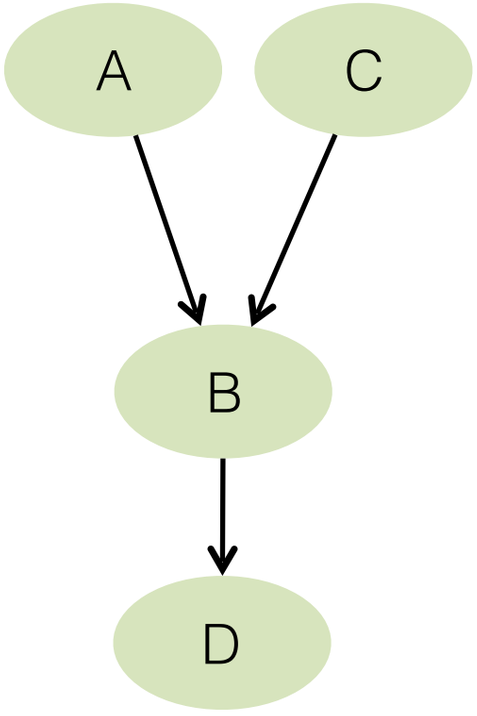
$A \not\perp C \mid D$

$A \not\perp B$

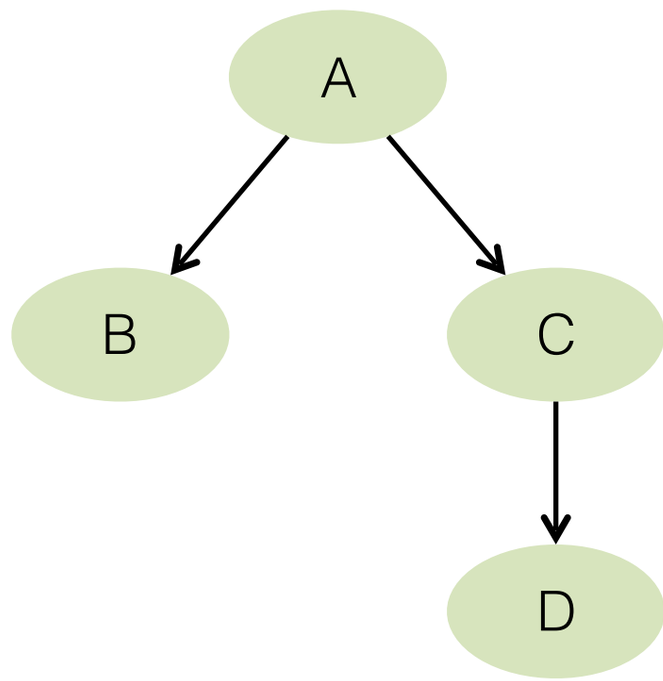
$C \not\perp B$

$D \not\perp B$

Answer:



What conditional independencies does this graph imply?

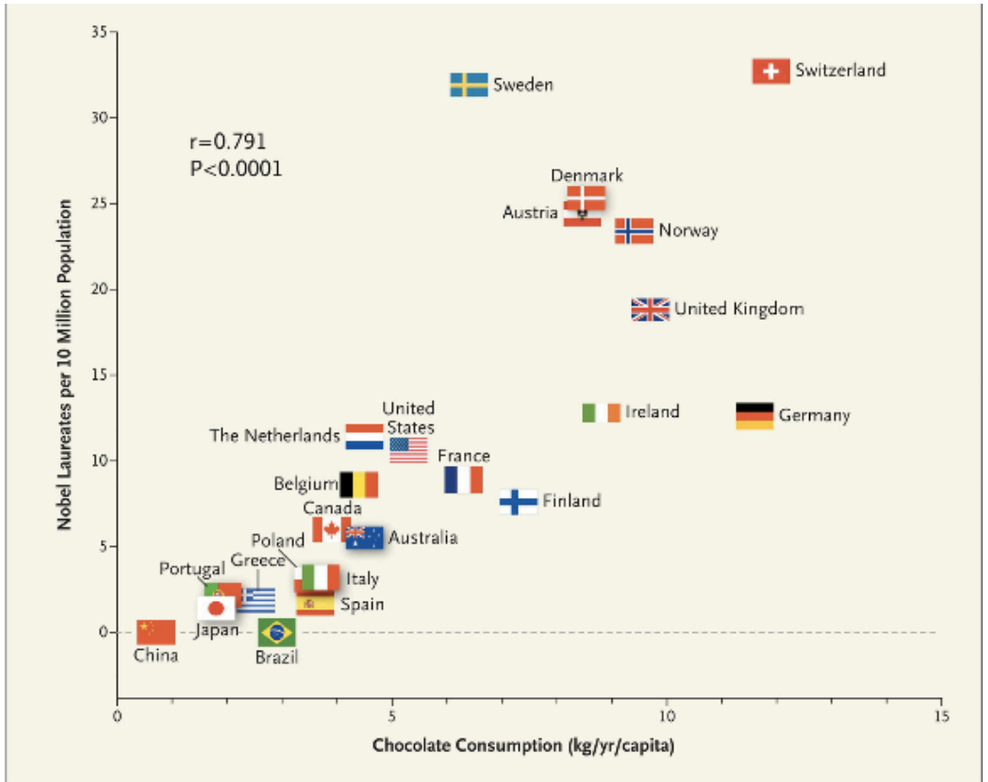


Countries that eat more chocolate win more Nobel prizes. The correlation is statistically significant and chocolate consumption seems to rise at the same rate as Nobel prize winners.

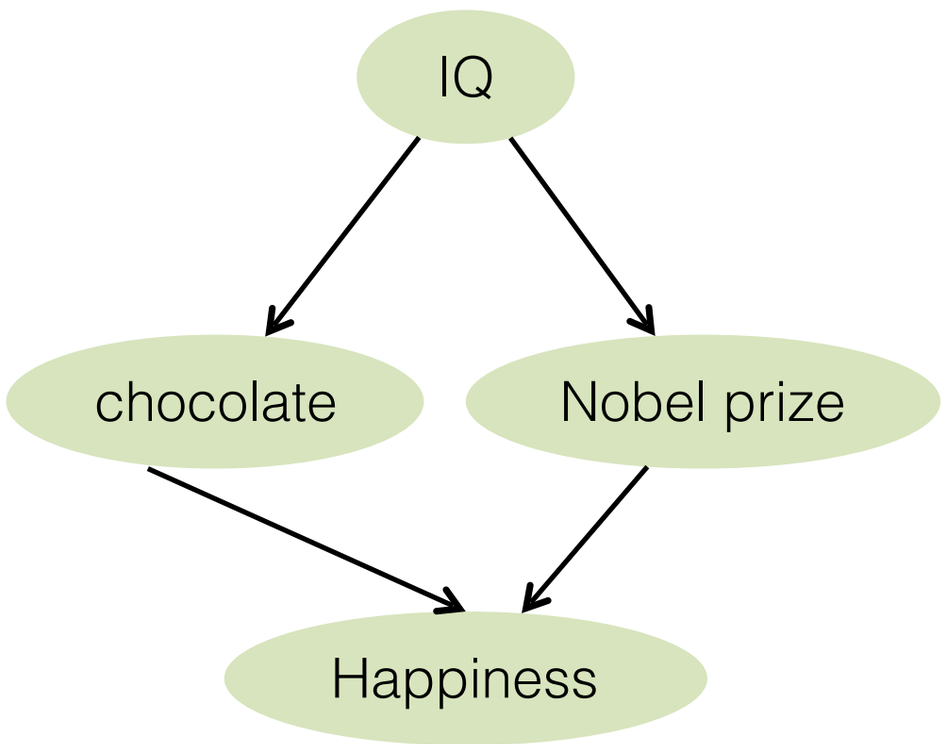
What relationships can explain this?

Should we

a) move somewhere with high chocolate consumption?
b) stay here but eat more chocolate?



Messerli FH (2012) Chocolate Consumption, Cognitive Function, and Nobel Laureates. N Engl J Med.



Now say…
H=C v N
C=I
N=I

If someone is happy, would they still be happy if they didn't have chocolate?

For next week: midterm

Two weeks: project proposal