CSE 564
Visualization & Visual Analytics

Visual Causality Analysis

Klaus Mueller and Jun Wang

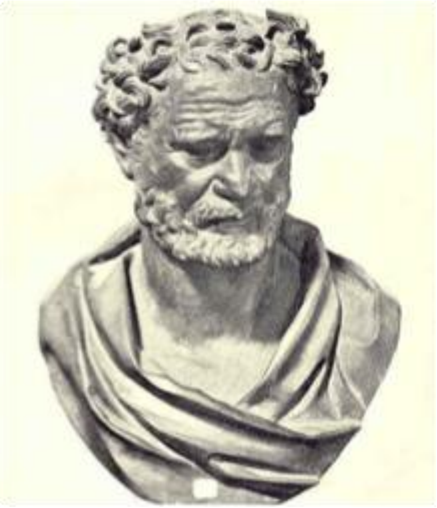
Computer Science Department
Stony Brook University

Causality – Fabric of Science

- Used as a utility to explain the observed world
  - ➢ Biology, Health Sciences, Psychology, Social Science, Economics, Environmental sciences, and many more…

Causality – Fabric of Science

- Being studied as a specialized topic in
  - Philosophy
  - Physics
  - Statistics
  - Computer Science



"I would rather discover one true cause than gain the kingdom of Persia"

- Democritus

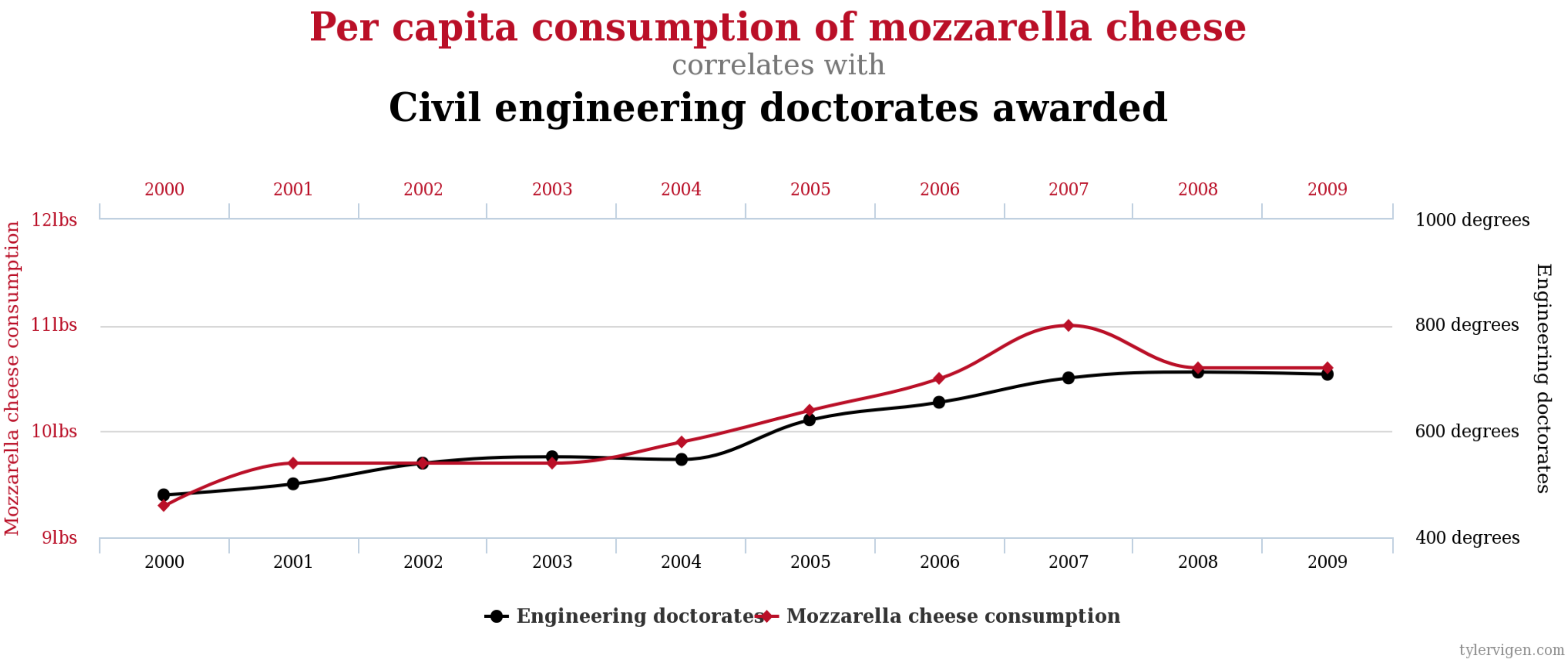
A bit of Philosophy

- Regularity – time and space constraint
  - A cause and its effect must both occur and be nearby in time and space, and a cause must precede its effect.
  - *"Day causes night"* ?
- Counterfactuals
  - *Had the cause not taken place, the effect would not have happened either*
  - $c \rightarrow e$ and $\neg c \rightarrow \neg e$
  - "If I didn't study hard, I would not get a good grade in the exam."

What is not causality – Correlation

- Correlation (Association) – a widely used evidence
  - Pearson’s correlation coefficient
  - Spearman's rank correlation coefficient
  - Joint probability and two-way Chi-squared statistics
  - Linear Regressions, etc.
- What correlation CAN tell you:
  - Two things are often observed happening together
  - Stained teeth and lung cancer are statistically associated
- What correlation CANNOT tell you:
  - *Counterfactual*
  - Can bleaching teeth reduces the chance of getting cancer?

What is not causality – Correlation

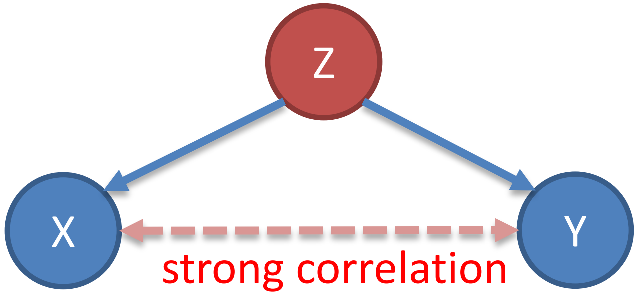


Correlation can be ambiguous or even spurious

What is not causality – Correlation

Even we know things are indeed related, correlation is still not causation. Why?

- Confounding – common cause of several events/variables

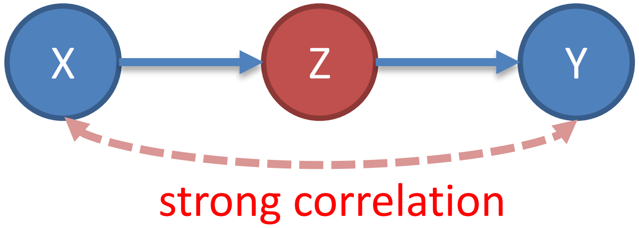


Example:

- Correlation of yellow teeth and lung cancer
  confounder: smoking
- Correlation of ice cream consumption and swimming population
  confounder: outdoor temperature
- Is smoking really causing lung cancer?

What is not causality – Correlation

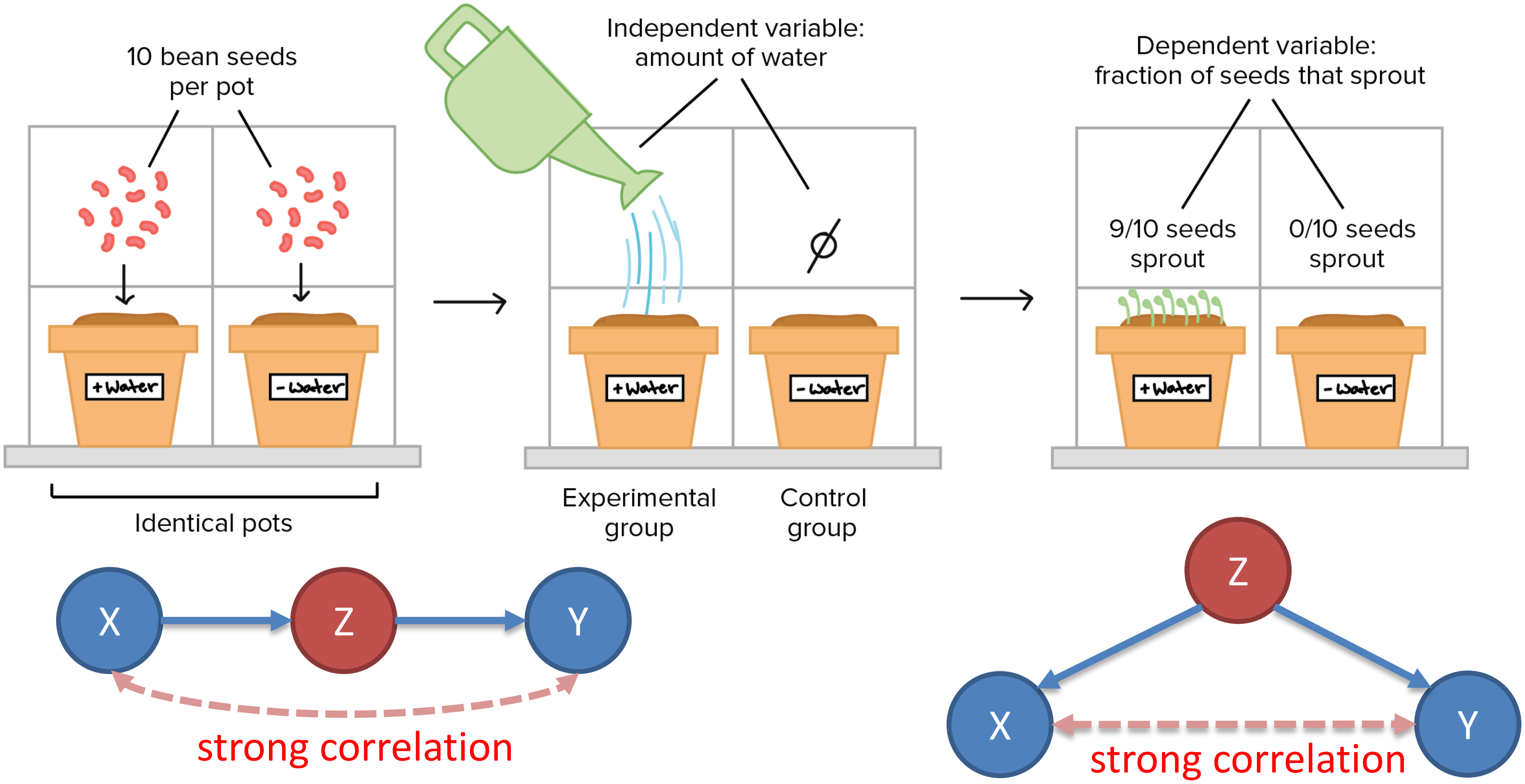
- Chaining



Example – Ice cream consumption (Y) is caused by outdoor temperature (Z), which is decided by position of the Earth relative to the Sun (X)

The Quest for Causality

Controlled experiments – get rid of other causes of Y



Controlled Studies are Difficult

Can we ask people to smoke so that we can find out if smoking is causing lung cancer?

Can we ask people to stop flossing to find out if it prevents gum disease and cavities?

Placebo studies are often difficult to morally justify

- should we deny a dying person the drug that could save him?

All we usually have is anecdotal and observational data

- lots of them
- not a controlled study, but even placebo studies can have side effects

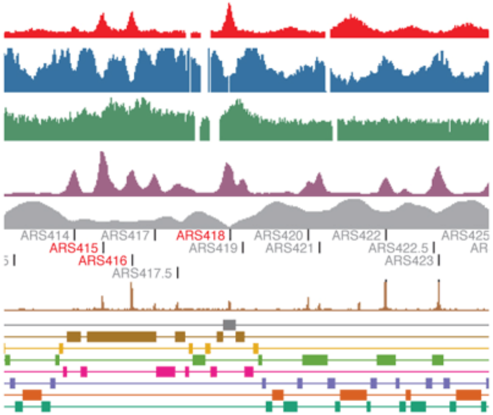
Thus, true causalities are difficult to determine

- combine statistical methods with expert knowledge & common sense

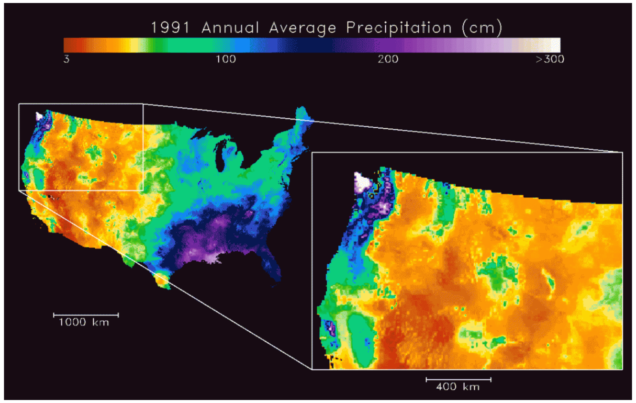
The Era of Big Data

BARC.L - Barclays PLC

Financial Data



Bioinformatics

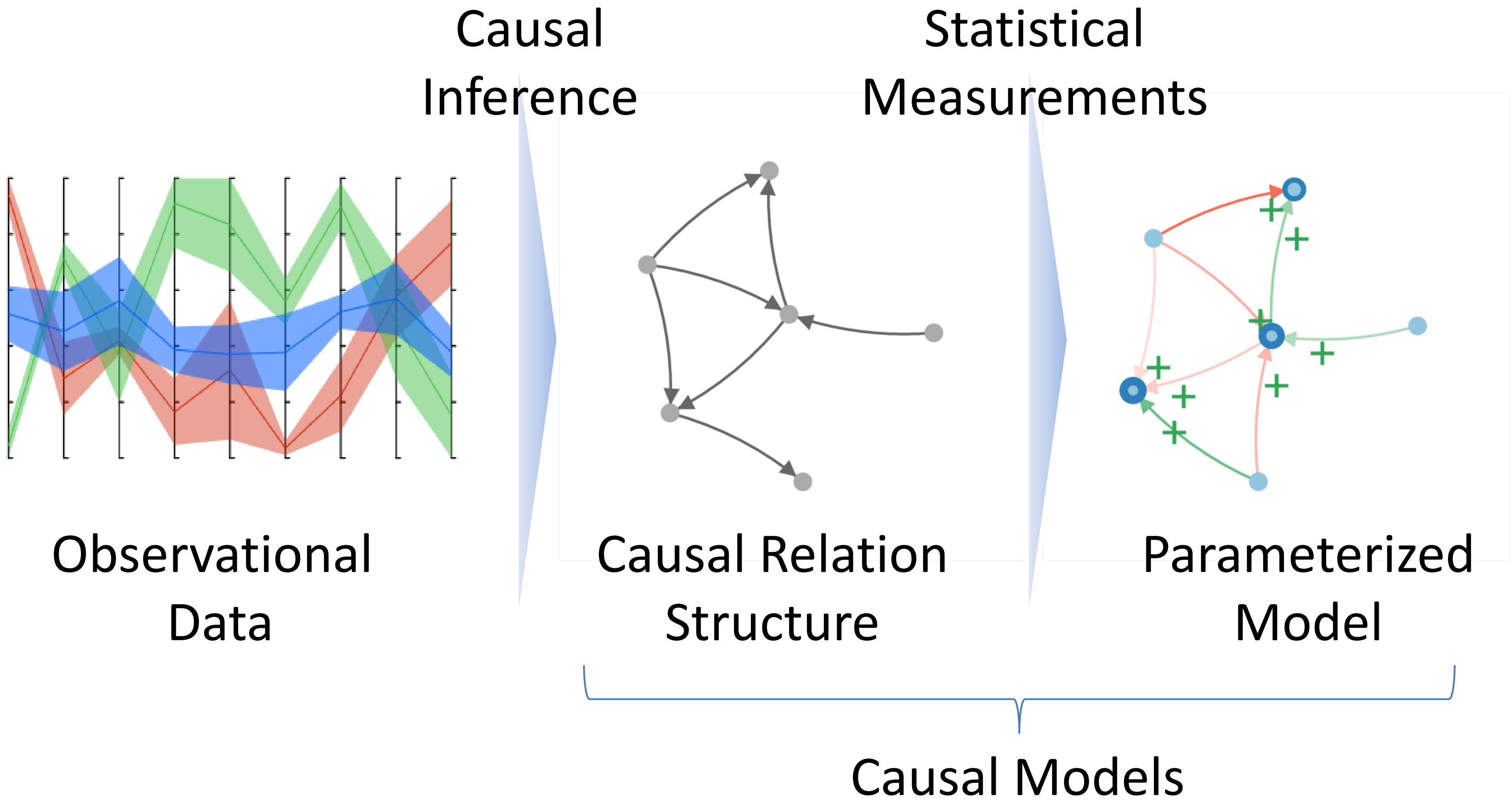


Scientific Data

Men
Women
.24
power society political human nation country liberty rights freedom state civil government america democracy thomas
.23
battles war victory won bands fought fighting defeated battle win fight meet enemy sword defeat
.31
sooo super excitement sooooooo ridiculously tomorrow uber excited yay tonight soooo
.30
soo love missing sooooo sooo guys soooo miss missed
.22
deserved bet lose win wi loses game winning loose wins winner streak losing loser chance
.21
coach game baseball football fantasy nfl players season basketball league team sports play playing player
.28
babe <3 ily boyfriend bestfriend besties bestie xoxo
.28
soo cute soooo awww sooo adorable puppy baby babies aww
.21
facts logic based philosophy opinions logical simply argument fact opinion beliefs moral political society
.21
income tax pay country taxes public benefits budget government obama economy state cuts
.26
celebrate brother happy years nephew special wishing birthday sister daughter wonderful niece son
.25
yayy (: :) gooood soo alltodayy goooddayy funn

Psychology Data

Pipeline of Causality Analysis



Causal Inference

Statistical Measurements

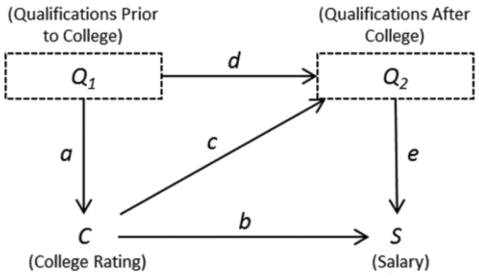
Observational Data

Causal Relation Structure

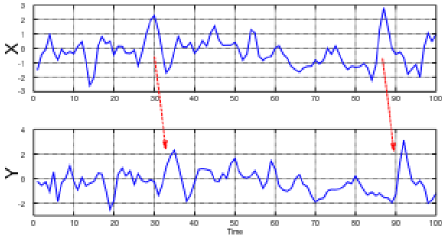
Parameterized Model

Causal Models

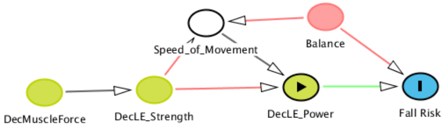
Causal Models


(Qualifications Prior to College)
(Qualifications After College)
Q_1
Q_2
d
a
c
e
b
C
(College Rating)
S
(Salary)

- Graph-based Model [Judea Pearl, 2000]
  - Probabilistic model: Bayesian Networks (BN)
  - Deterministic model: Causal Structure Model (CSM/SEM)
  - For time series: Dynamic BN



- Granger Causality [Clive W.J. Granger, 1969]
  - Specialized for time series data
  - Value of a variable at time *t* could depend on values of itself or other variables at any time no later than *t*

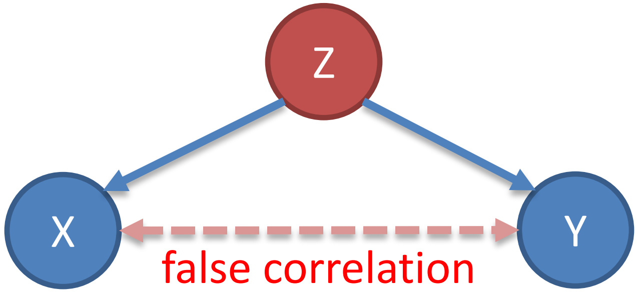


- Logic-based models [Sam. Kleinberg, 2010]
  - An event is defined by a set of propositions
  - A causal relation is a logic path between two events with certain time lags.

Causal Inference
From Data to BN Structure

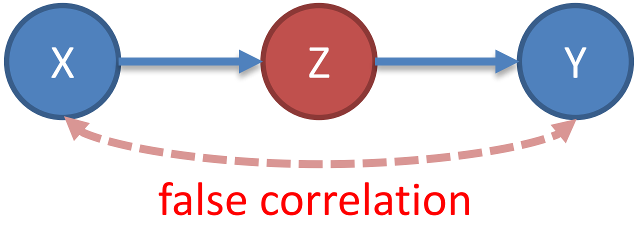
Why correlation is not causation?

- Confounding – common cause of several events/variables



Example – Correlation of ice cream consumption (X) and swimming population (Y) (*confounder*: outdoor temperature (Z))

- Chaining



Example – Outdoor temperature (Y) is decided by duration (Z) of sunshine (X)

Causal Inference
From Data to BN Structure

- What we see?
  - raw data
  - correlation/joint probability
- What we want?
  - test if it is causation or if it is confounding/chaining
- How?
  - Conditional Independence Test

Conditional Independence

Conditional Independence (CI)

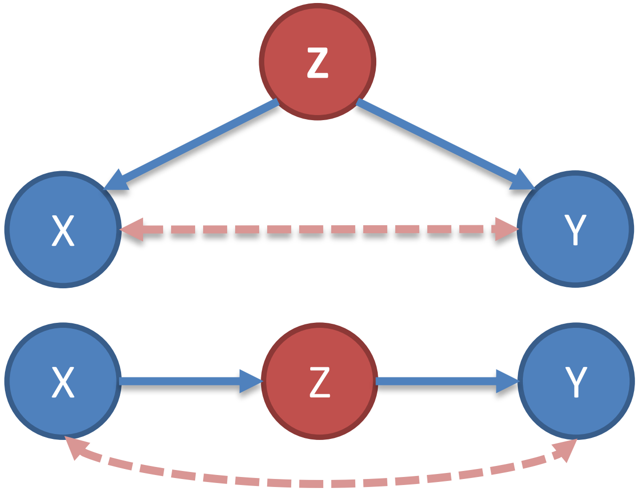
- A relation among three sets of variables:
  - two variables X and Y
  - a set of variables **Z**

So that $\boxed{P(X,Y|\boldsymbol{Z}) = P(X|\boldsymbol{Z})P(Y|\boldsymbol{Z})}$

or $\boxed{\rho(X,Y|\boldsymbol{Z}) = 0}$

ρ is the partial correlation*

Conditional Independent tests



- Written as $X \perp\!\!\!\perp Y \mid \boldsymbol{Z}$
- Read as *X* and *Y* are independent conditioning on ***Z***

**partial correlation of x and y conditioned on Z is the correlation of residuals of the regression of x on Z and the regression of y on Z.*

Conditional Independence

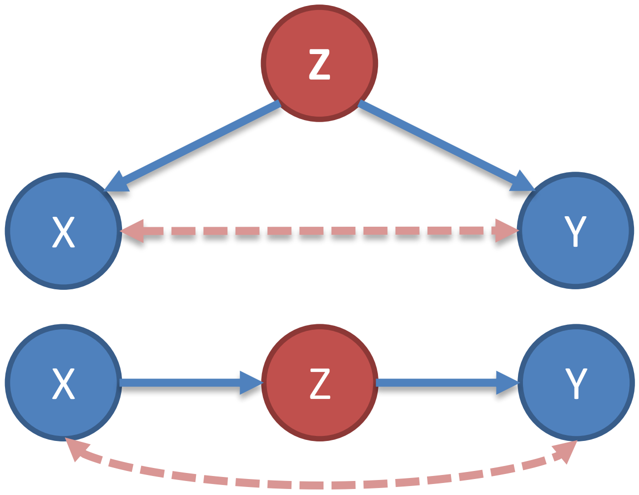
- Smoking (X) and lung cancer (Y)
  - $\boldsymbol{Z}$: gender, race, alcohol consumption, etc.
  - $P(X, Y|\boldsymbol{Z})$ and $P(X|\boldsymbol{Z})P(Y|\boldsymbol{Z})$
    under all possible combination of $\boldsymbol{Z}$
- Ice cream consumption (X) and swimming population (Y)
  - $\boldsymbol{Z}$: outdoor temperature
  - R_1- residuals of regression $X = \alpha Z + \beta$
    R_2- residuals of regression $Y = \alpha Z + \beta$
    Partial correlation ρ – correlation of $R_1$and R_2

Causal Inference – CI test

- *X* and *Y* are causally related only when $X \perp\!\!\!\perp Y \mid \boldsymbol{Z}$ is **NOT TRUE** no matter what $\boldsymbol{Z}$ is.
- To test if *X* and *Y* are causally dependent

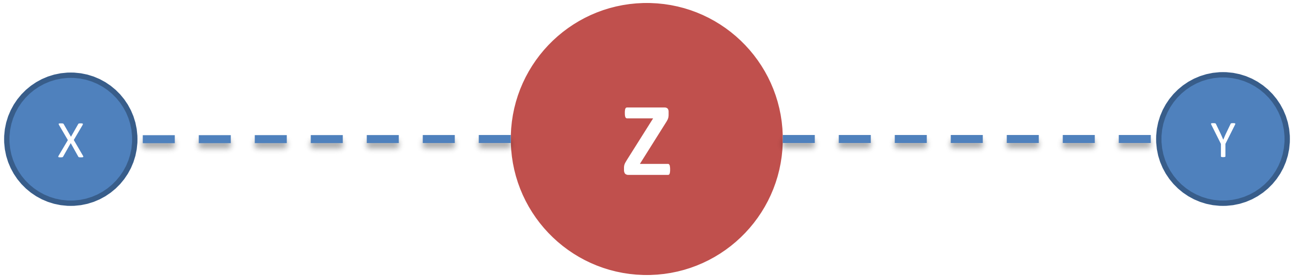
  EQUALS

  To search for the set $\boldsymbol{Z}$
- Brute force search *or* with some algorithms, both need a number of CI tests exponential to the number of variables.



This means the ice cream consumption and the swimming population may not be correlated when only looking at days with the same outdoor temperature.

Causal Inference

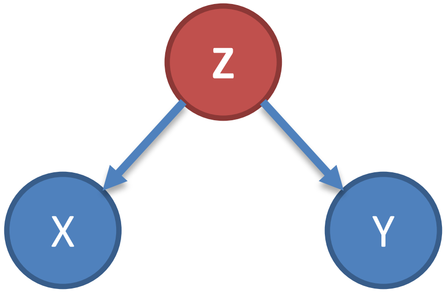


Are *X* and *Y* causally related? = Can we find the set ***Z***?

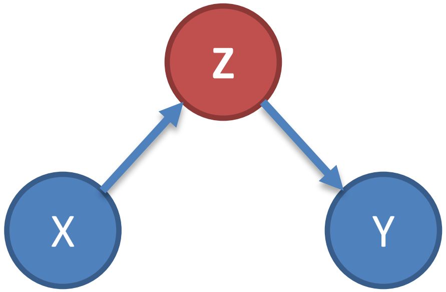
But how do we differentiate confounding and chaining?
How do we know the direction of the causal relation?
Based on the above information, We CANNOT!

Causal Inference – Colliders

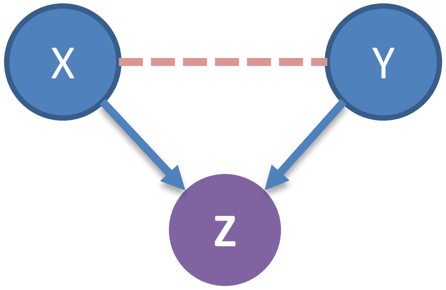
All possible relations between *X*, *Y* and ***Z***



Confounding



Chaining



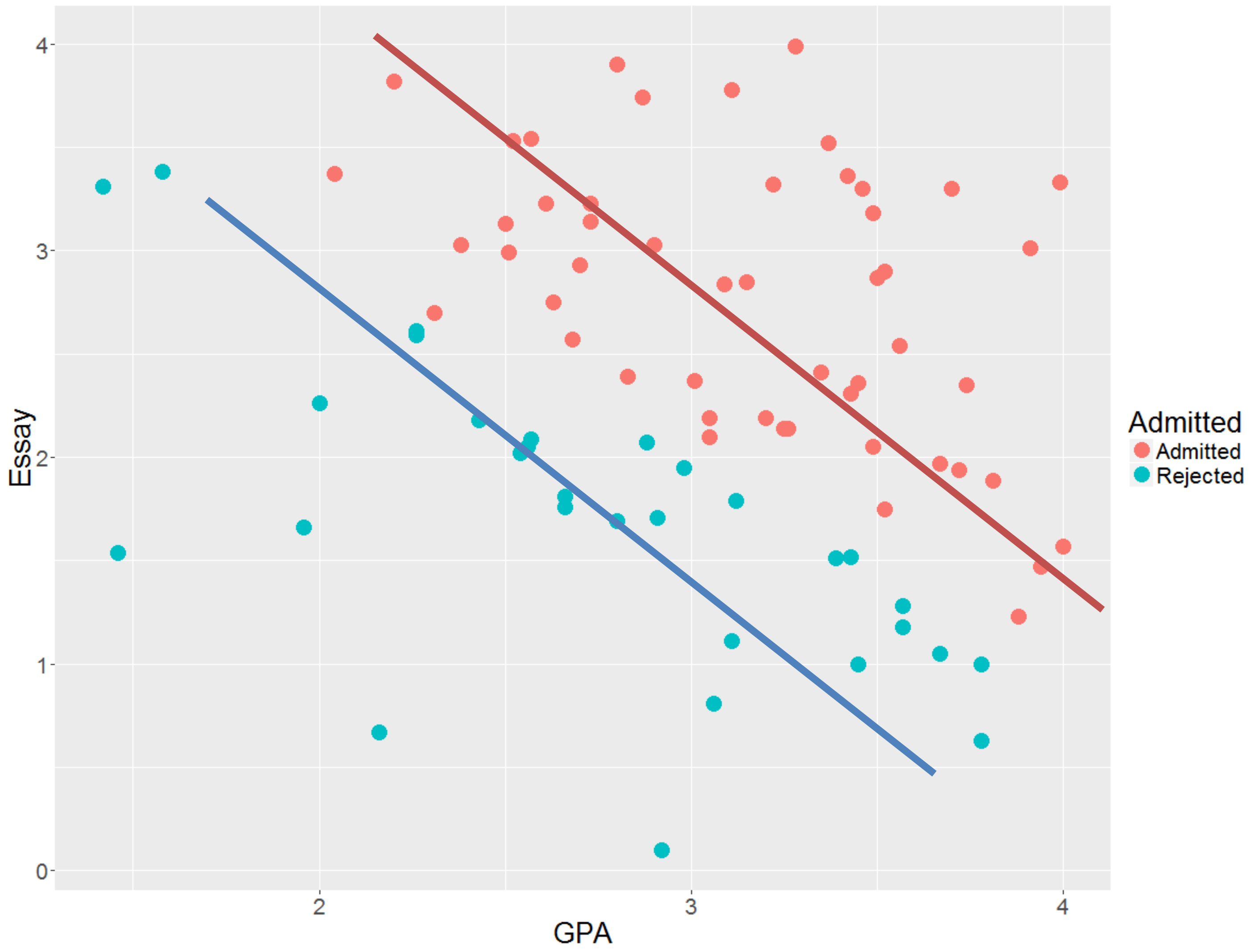
Collision

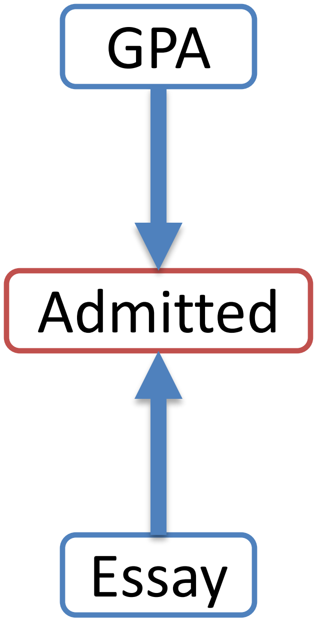
In the situation of collision, X and Y will not independent conditioning on ***Z*** (*colliders*)

which means…

Conditioning on ***Z*** will bring false connection

Example of Collision





GPA

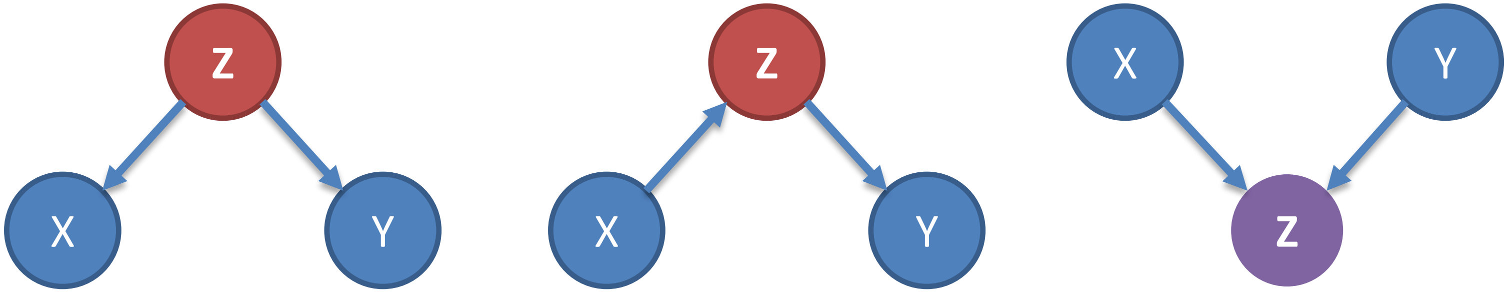
Admitted

Essay

Causal Inference

So…

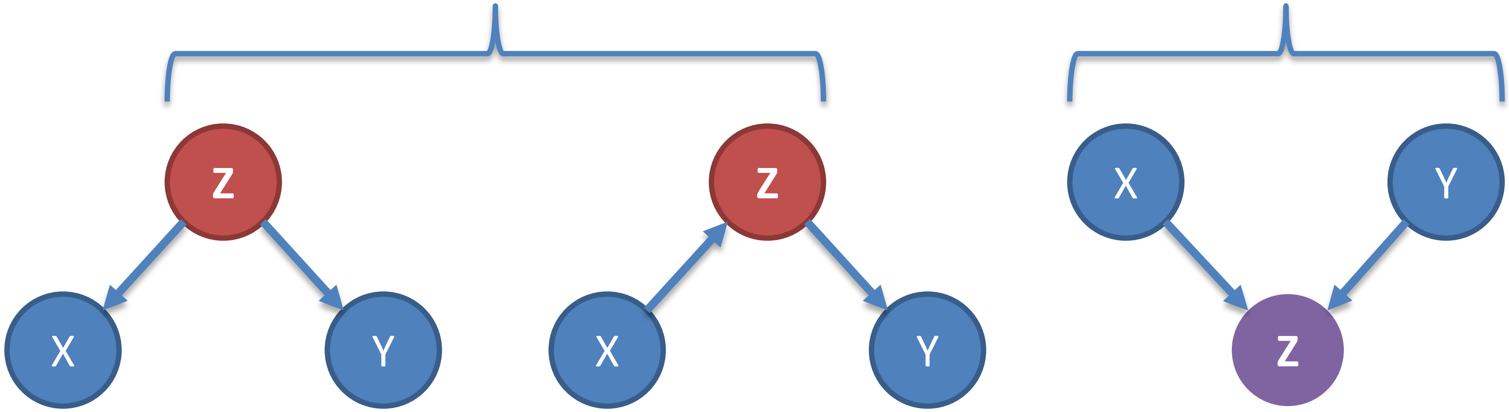
- To correctly recognize if X and Y are causally dependent, we have to search for the set ***Z***
- The set ***Z*** should contain confounders and variables chaining from *X* to *Y*
- The set ***Z*** should **NOT** include colliders of *X* and *Y*



Causal Inference

Z should be included in the d-separating set of *X* and *Y*

Z should not be included in the d-separating set of *X* and *Y*



Confounding

Chaining

Collision

Also…

- If we recognize colliders, we know edge directions

Causal Inference

Based on this, we can build a causal inference algorithm:

```
For each pair of potential relations $X-Y$ :
    looking for another set of variables $\boldsymbol{Z}$ so that $X \perp\!\!\!\perp Y \mid \boldsymbol{Z}$
    If $\boldsymbol{Z}$ exists:
        break the edge $X-Y$
        For each variable cannot be included in $\boldsymbol{Z}$:
            add $X \rightarrow Y \leftarrow Z$ to result
        End For
    Else:
        add $X-Y$ to result
    End If
End For
```

Causal Inference – Formal Expression

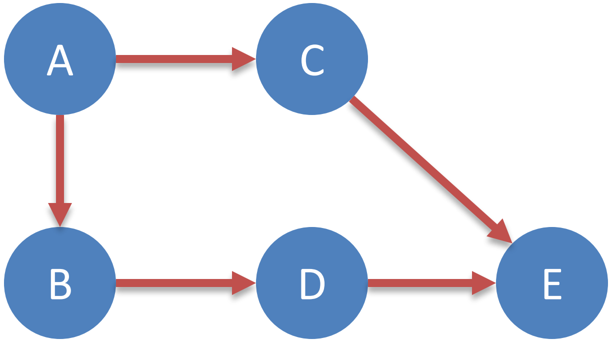
The ***Causal Markov Condition*** [Pear and Verma, 1991]

- *A variable is independent of all of its non-descendants conditioning on all of its direct causes (those that are connected to the node by one edge)*
- Corresponding to the ***d-separation*** in graph theory, in which
  - If set ***Z*** exists, we say ***Z*** is d-separating or blocking every path between *X* and *Y*, and ***Z*** is the *d-separating set* of *X* and *Y*
  - *X* and *Y* are causally dependent when they cannot be d-separated.
  - On the path $X \rightarrow Z_1 \leftarrow \mathrm{Z}_2 \rightarrow Z_3 \rightarrow Y$, ***Z*** can be either {}, $\{Z_2\}, \{Z_3\}, \{Z_2, Z_3\}$, as Z_1 is a collider on the path

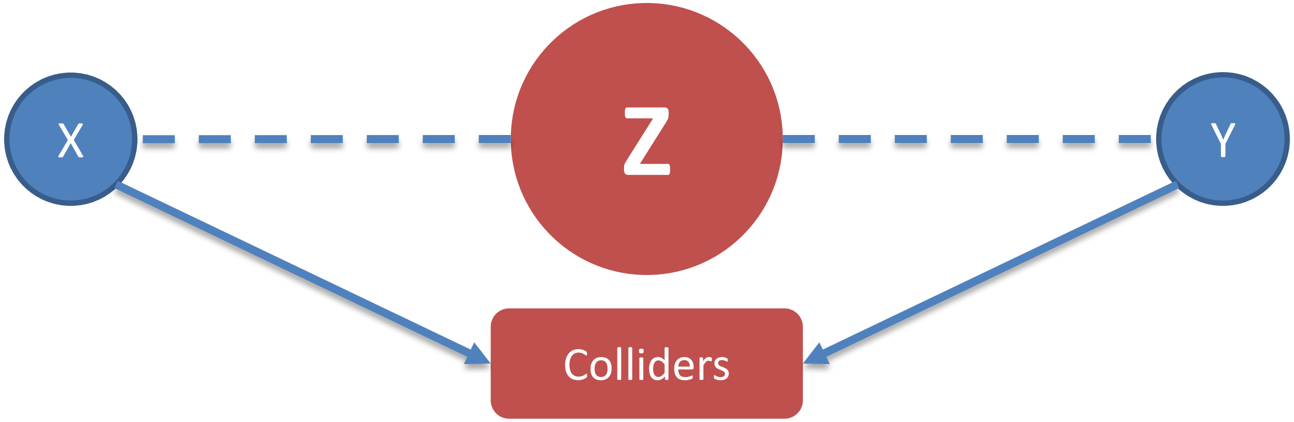
D-SEPARATION

Example:

- What is the d-separating set of C and D?
- Which variables can d-separate A and E?



Causal Inference



Causal Inference = Searching for d-separating sets and colliders

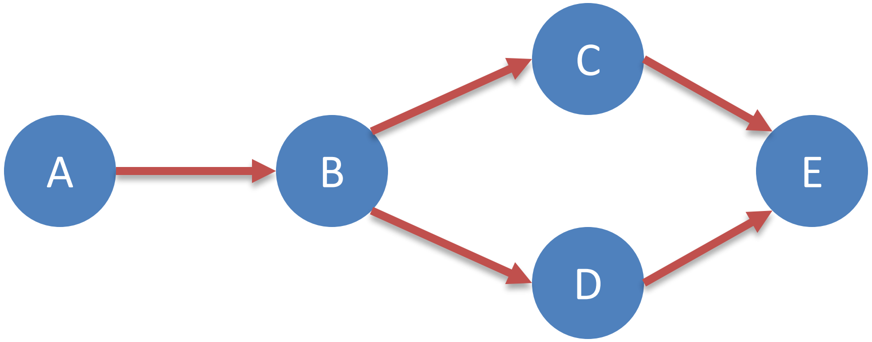
+ Build graph satisfying these constraints

The PC Algorithm (Phase 1)

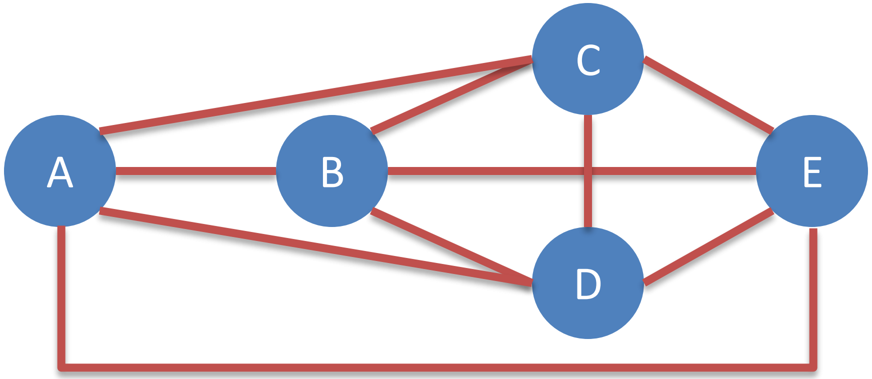
PC (Peter-Clark) [Spirtes, Glymour and Scheines, 1989]

```
Start with a complete undirected graph $G = (\boldsymbol{V}, \boldsymbol{E})$ where $\boldsymbol{V}$ is the node set and $\boldsymbol{E}$ is the edge set.
Set CI test order $n = 0$ (size of the separating set to test)
Repeat
    For each $Y \in V$:
        For each $Z \in Adjacencies(Y)$:
            For each subset $\boldsymbol{S} \subseteq Ajacencies(Y) \backslash \{Z\}$ with $|\boldsymbol{S}| = n$:
                (CI) test $Y \perp\!\!\!\perp Z \mid \boldsymbol{S}$
                If True:
                    Remove edge $Y - Z$ from $\boldsymbol{E}$
                    $\boldsymbol{Sepset}(Y - Z) = \boldsymbol{S}$
                    Break
    $n = n + 1$
Until all $|Adjacencies(Y) \backslash \{Z\}| < n$ or $n = n_{max}$
```

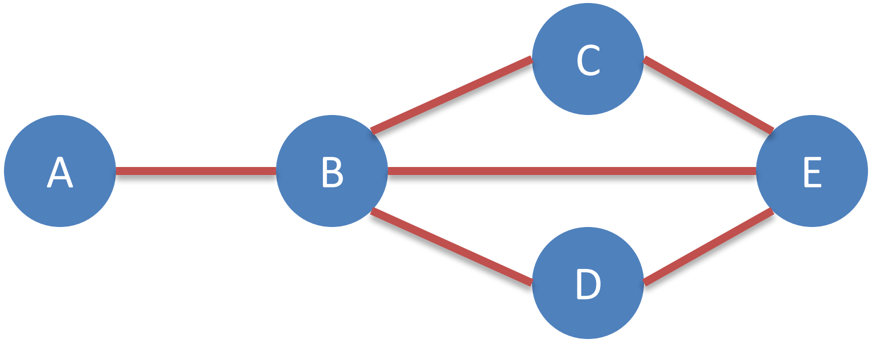
The PC Algorithm (Phase 1)



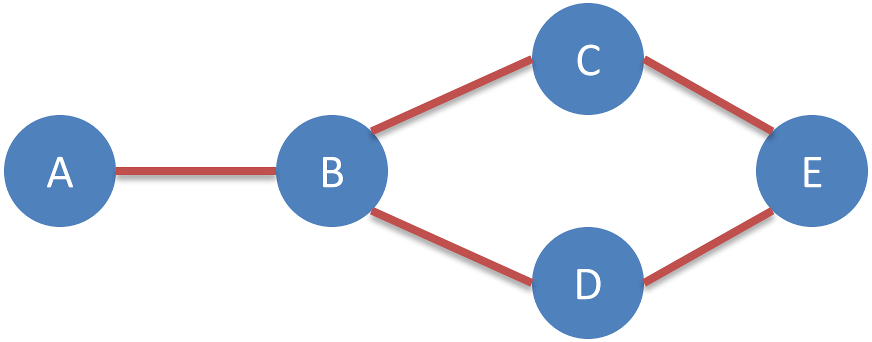
Ground Truth



Initial Graph & after n = 0



After n = 1



After n = 2

Constructs an undirected graph (***Skeleton***), finds all d-separating sets

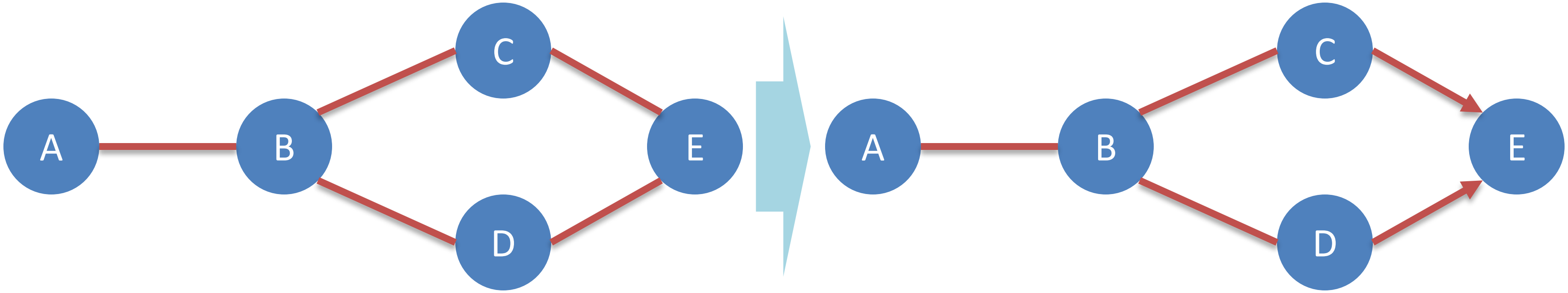
The PC Algorithm (Phase 2 & 3)

Orient edges

- For each triple of nodes $X - Y - Z$ and X is not adjacent to Z, orient as $X \rightarrow Y \leftarrow Z$ iff. $Y \notin \boldsymbol{Sepset}(X, Z)$

Propagation

- If $A \rightarrow B$, B and C are adjacent, A and C are not adjacent, and there is no arrowhead at B, then orient $B - C$ as $B \rightarrow C$.
- If there are a directed path from A to B and an undirected edge between A and B, orient $A - B$ as $A \rightarrow B$.



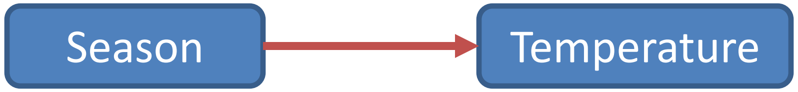
Not all edges can be oriented

Other Inference Algorithms

- SGS [Spirtes et al. 1989]
- TPDA [Cheng et al. 1997]
- Heuristic two-phase [Wang & Chan, 2010]
- TC [Pellet & Elisseeff, 2008]
- …

Practical Problems

What if there are mixed type of variables in dataset?



- No CI test method available for such situation.
    - ➢ G^2 test *or* test of $\rho(X, Y|Z) = 0$
- Possible solution
    - Discretize: Numeric -> Categorical (information loss)
    - **???**: Categorical -> Numeric
        - ➢ Pair-wise value mapping [Zhang et al., 2015]
        - ➢ Global-wise value mapping [Wang and Mueller, 2016]
        - ➢ Still an open problem

Practical Problems

Assumptions of these algorithms:

1. **Faithfulness** – exactly the CI relations found in the causal graph hold in real world; and no unobserved CI relations.
   - Violation: *unfaithful* population (data)
2. **Causal Sufficiency** – the set of measured variables includes all common causes of variable pairs in the set
   - If violated, spurious relation will exist in result.

Assumption of CI test

- For discrete data: enough data to fill the contingent table.
- For numeric data: variables are linearly related with Gaussian error.

Practical Problems

In Summary

- We assume perfect data measuring all variables of the observed system and their distributions.
- We usually don’t have such data

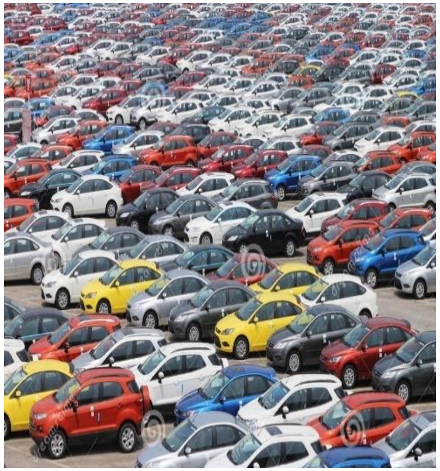
Solution

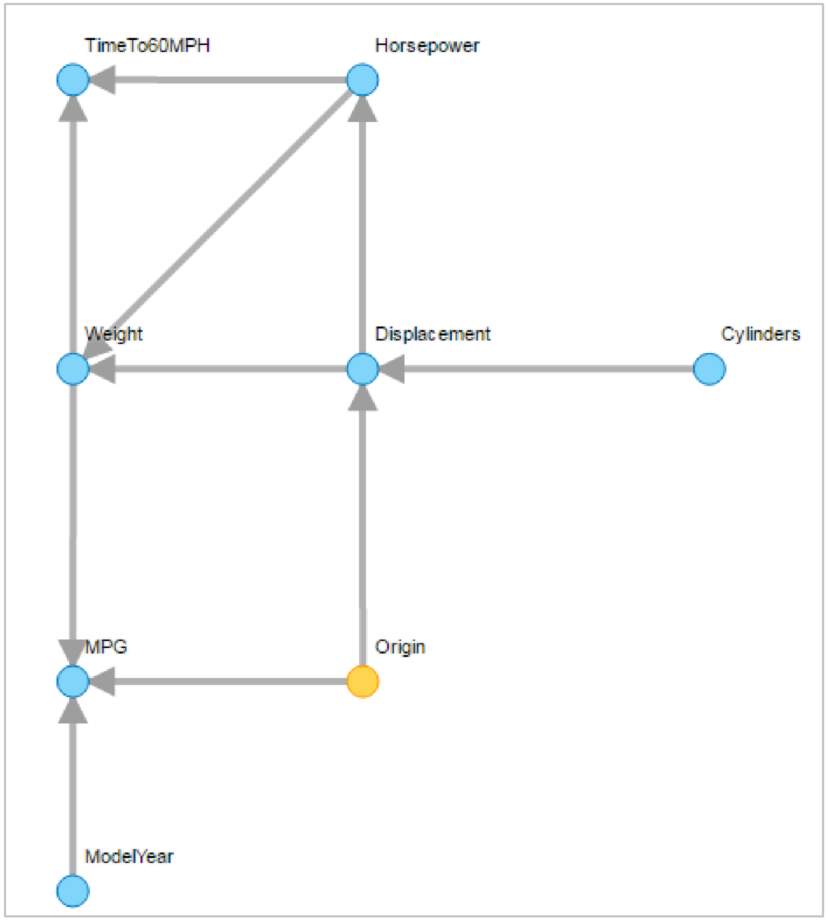
- Bring the domain researcher into the loop with a visual analytic system!

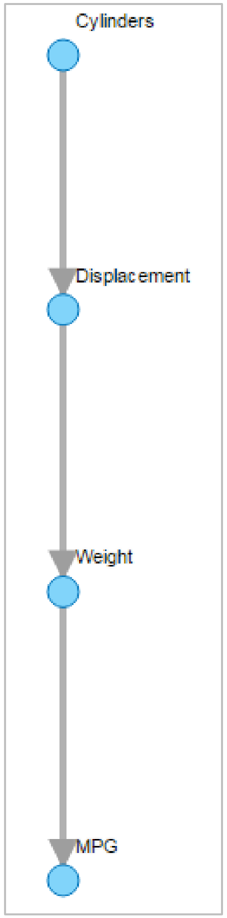
Visualization of the Graph Model

- Graph Visualization
    - nodes as variables
    - edges as causal relations (directed or undirected)
    - force directed layout (or other graph layouts)
- Interactions
    - Modify causal relations – create, delete, direct, reverse
    - Visualize selected part of the graph

Visualization of the Graph Model







*Ignore node color here

Measurements of Causality

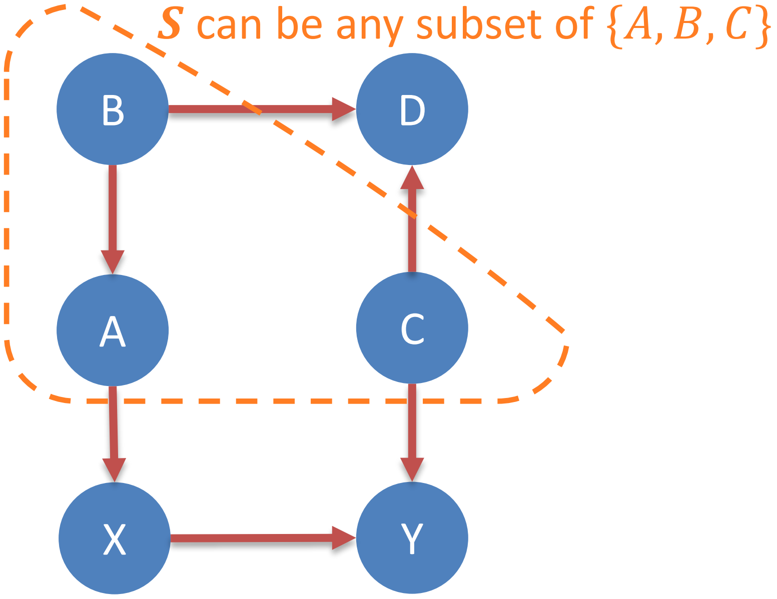
Now we have a causal graph structure, how do we measure these causal relationships?

- Probabilistic: Backdoor Criterion [Judea Pearl, 1993]
- Structural: Linear regression and Logistic regressions

Measurements of Causality

Back-door Criterion [Judea Pearl, 1993]

- When measuring $X \to Y$, we also need to consider the set $\boldsymbol{S}$ that can *block** other paths between X and Y



$$P(Y = y|do(X = x)) = \sum_{s} \frac{P(Y = y, X = x, S = s)}{P(X = x|S = s)}$$

The "do" calculus means assigning specific values to a variable, or *intervention* in causality terms.

*block is defined the same as d-separating

Measurements of Causality

Regressions – Equations from structures

- Linear regression measures the linear relationships between a dependent variable y and one or more explanatory variables $x_k, k = \{1, 2, \ldots, K\}$, taking the form

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \varepsilon_i$$

- Logistic regression analysis is actually a model of classification probabilities

$$\sigma(t) = \frac{1}{1 + e^{-t}}$$

$$\text{where } t = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \varepsilon_i$$

Measurements of Causality

A rich set of statistical measurements in regression

- Linear regression
  - ➢ Variable coefficients, t-statistics, p-values, standard errors
  - ➢ F-statistics, r-square, adj. r-square, p-value, BIC/AIC
- Logistic regression
  - ➢ Variable coefficients, t-statistics, p-values, standard errors
  - ➢ Likelihood Ratio, p-value, pseudo r-square, BIC/AIC
- Variable coefficients are used as metric of causal relations (positive and negative causes)

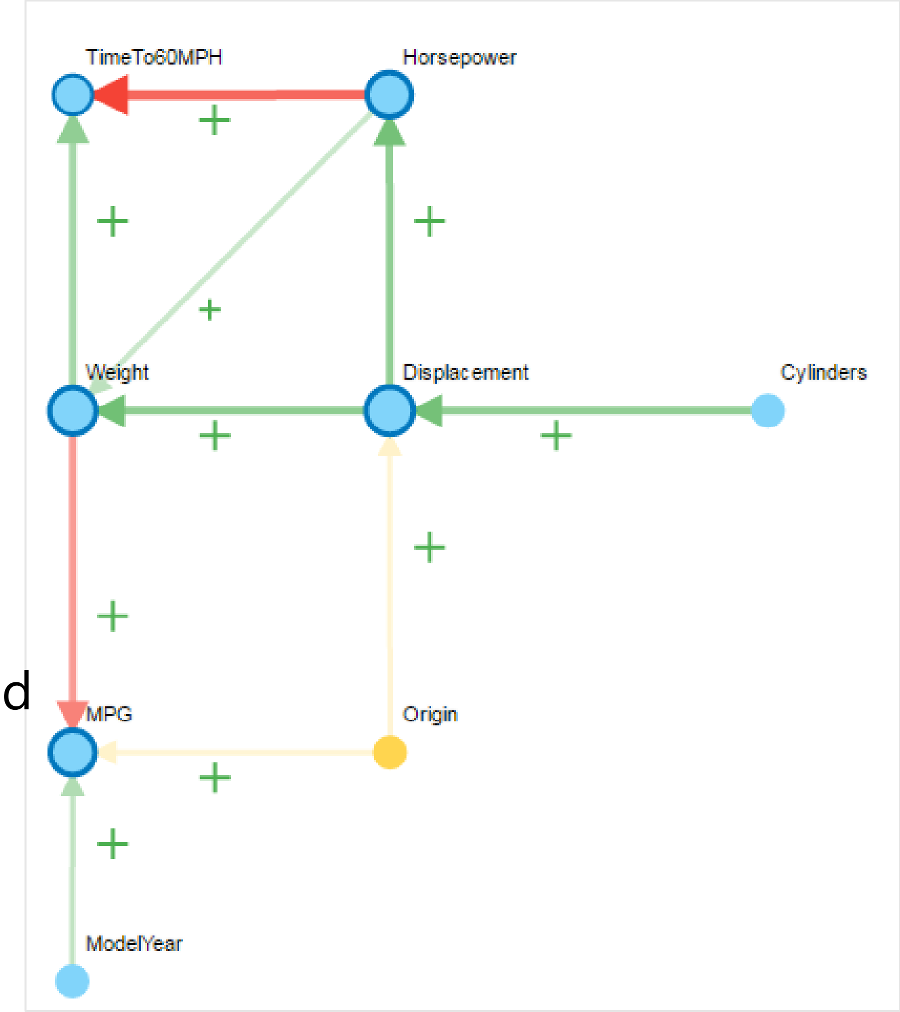
Graph Visualization

Nodes: variables

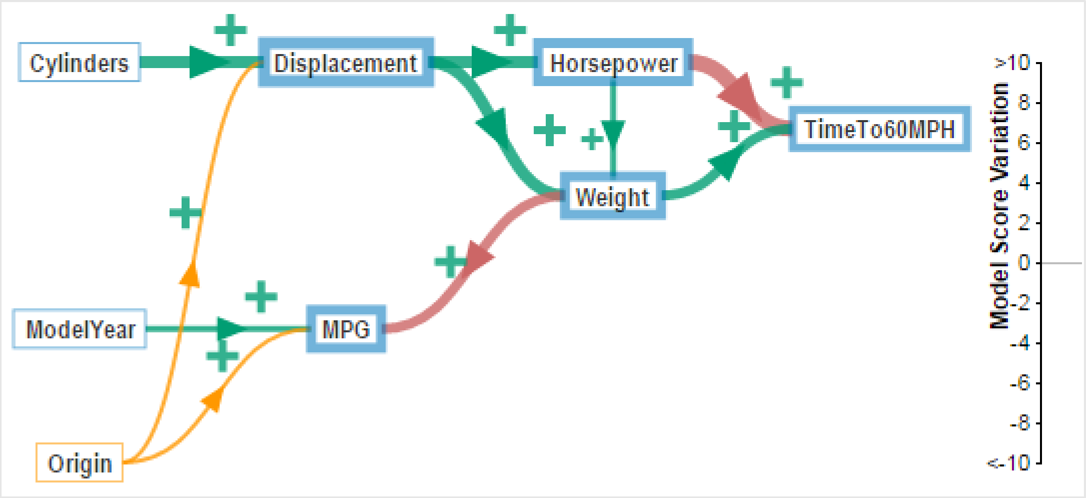
- Color: type of the variable
  (🔵 numerical 🟠categorical)
- Size: r-square
- Stroke: does or not have cause

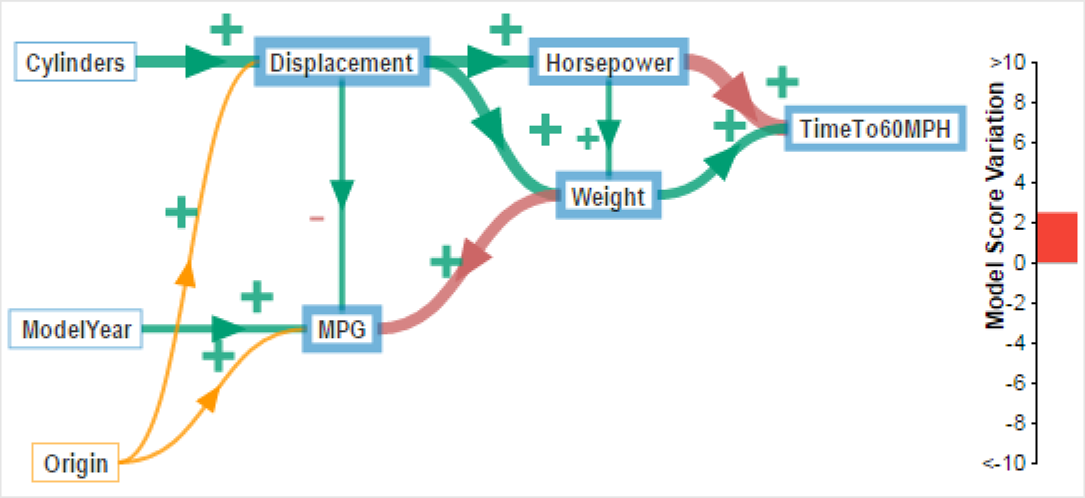
Edges: causal relations

- Marker: direction of relation
- Color: quality of relation
  ▶ positive▶ negative▶compound
- Opacity + width: causal strength
- Glyphs: BIC score change when deleted
  (+decrease— increase)



Graph Visualization





Graph Visualization

Possible Improvements

- Illustrative graph vs. statistical savvy graph?
- Visualization of separation sets?
- Better layout strategy?
- Other visualization approaches?
- Investigate causal relations with value bracketing?
- Management of inferred models?

Causal Inference

Theories of Causal Inference is still actively developing!

- CI test that can handle arbitrary data distributions?
- Involving time in inference?
- Methods to utilize data from different sources (experimental + observational) ?
- A faster inference algorithm (almost all are still exponential)?
- Application studies?