A Brief Discussion on Causal Inference

Nilavra Pathak,
Applied Researcher

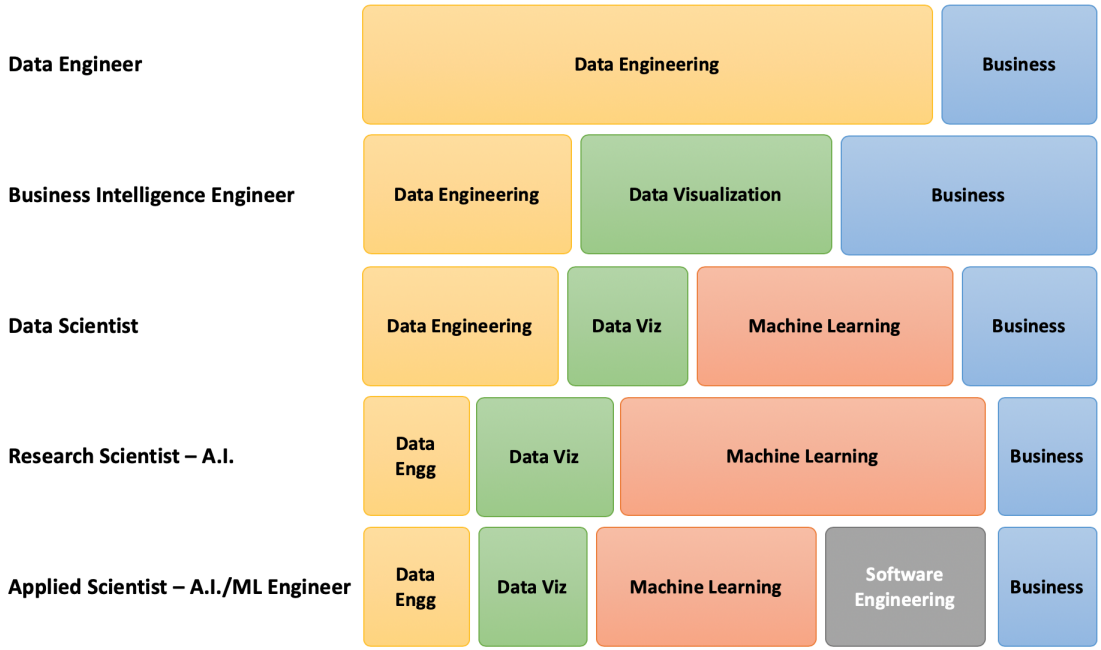
Expedia Group

May 7, 2020

About Me

Why do we need to care about causal Inference?

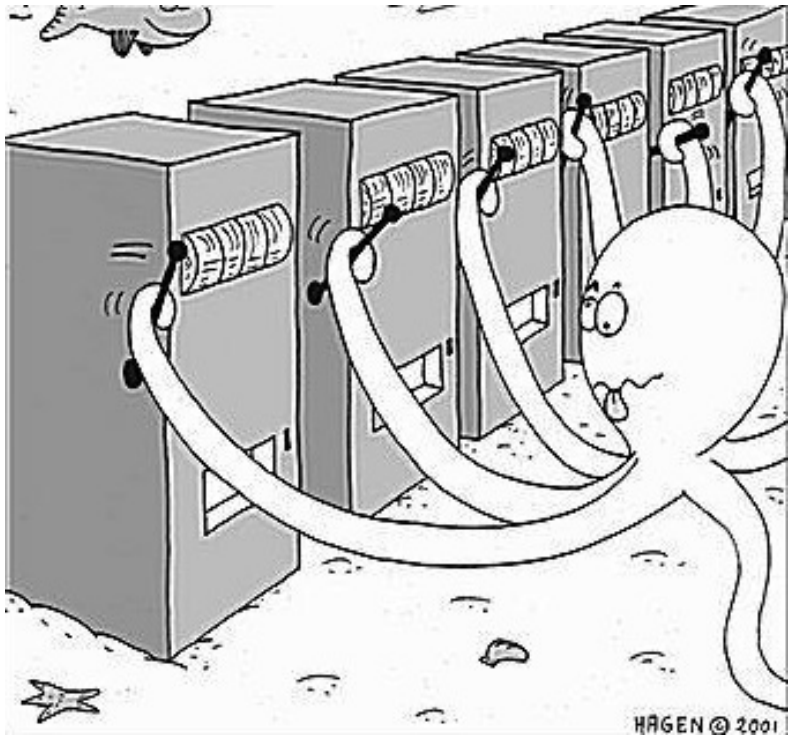
- Started Working in Expedia from July 2019
- Primary work on online bidding
- Crazy about Bandits !!
- Spend time on Bayesian Stats, PML, Causal Inference and Control Theory
- Spend even more time on Data Engineering and big data stuff
- Still don't care to learn about Deep Learning or NLP/Computer Vision



What are Bandit Algorithms?

A brief digression. I swear it's connected to causal inference.

- Bandits are stateless reinforcement learning
- General framework – N choices find the optimal choice that minimizes some regret
- Tied to the intervention and counterfactual notion in causal inference



What is Causal Inference?

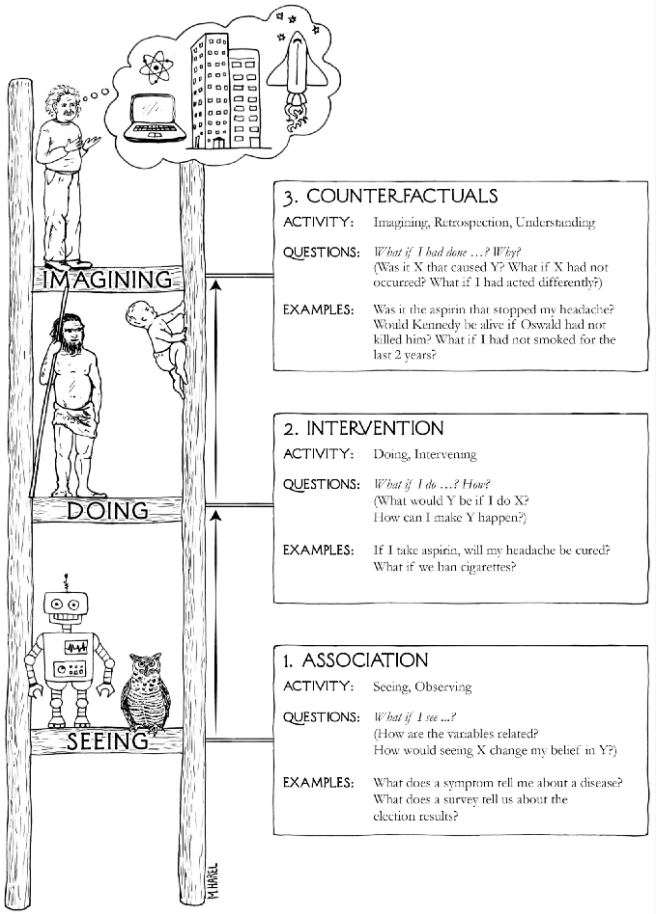
Why do we need to care about causal Inference?

- *Causal Inference is prediction under intervention*
- **Identifiability:** Estimating an identifiable causal effect from the data
- **Causal Discovery:** Discovering the causal model from the data
- **Counterfactuals:** How will the system perform if we did X instead of doing Y ?
- "*Say you're in a hospital and you have 250 million electronic health records of what medicines people received and what happened to those people. It seemed silly to say that it is impossible to learn, say that Advil helps headaches*" – David Blei.

Ladder of Causation

What kind of questions that each class is capable of answering?

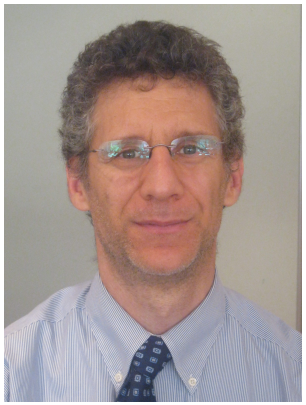
- **Association** : $P(y|x)$
  - Most modern ML reside in this layer
  - No true causality is established
- **Intervention** $P(y|do(x))$
  - True causation can only be established by intervention
  - Modern big data can utilize a lot
- **Introspection** $P(y|do(x'), x, y)$
  - Thinking machines

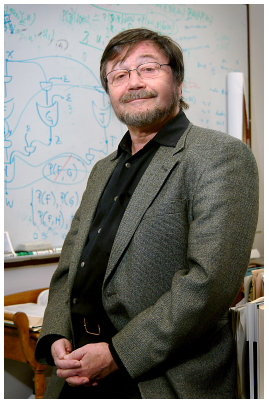


State of Modern Causal Inference

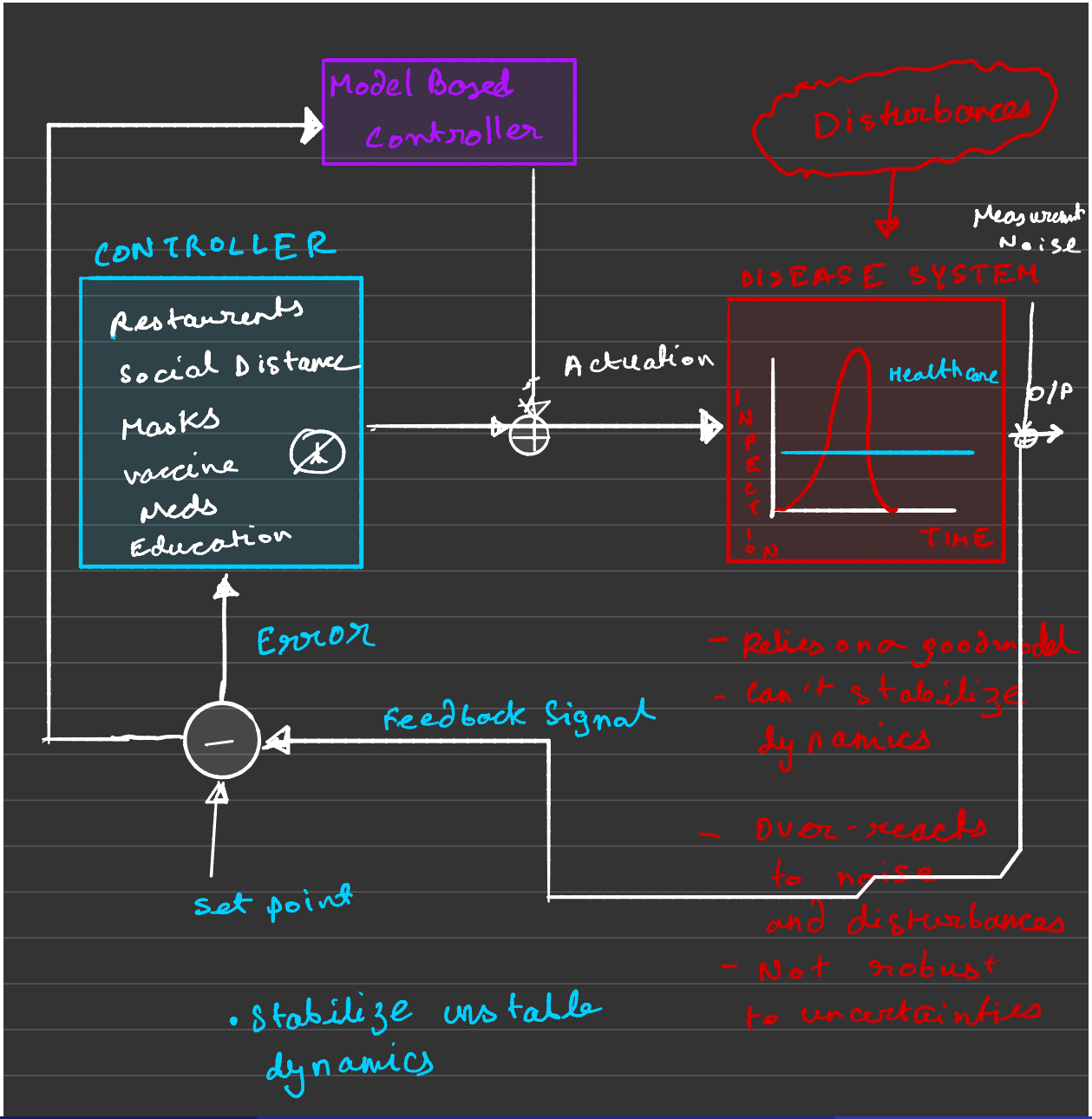
What are the schools of thoughts?

- Fischer: Randomized Controlled Trials
  - Gold Standards
  - Experimental Approach
  - Expensive/ Time Consuming
- Econometric: SEM, IV, Diff-Diff
  - Prevalent
  - Slightly Confusing
- Systems Identification/Control Theory: I have no clue (I can make some connections)
- Stats: Potential Outcomes Framework
- Pearl's AI/Math: Structural Causal Models



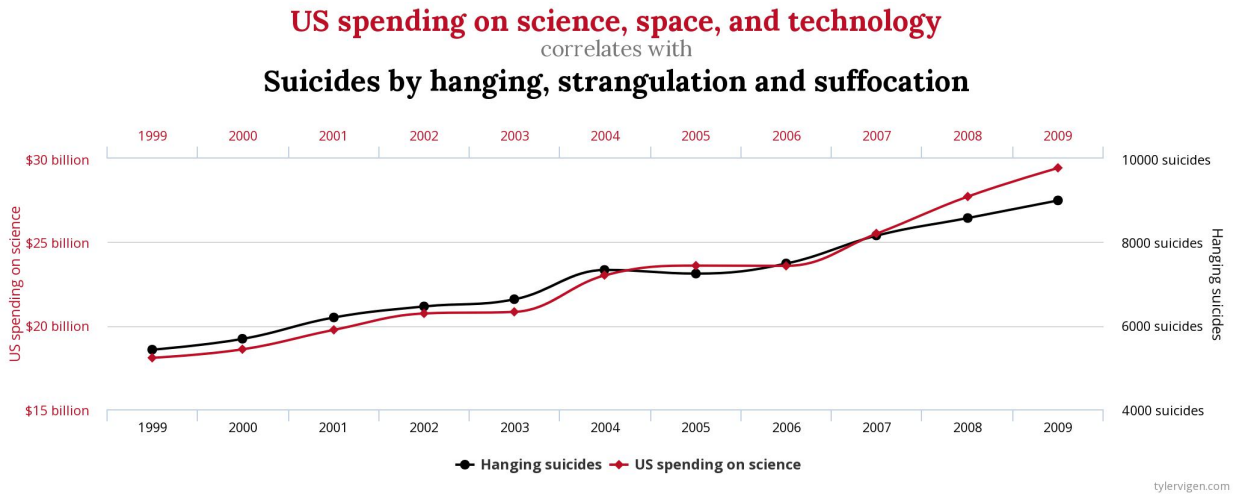


Control Example



Motivation

- Experimentation needs to account for all controlled factors
- Sometimes experimentation is impossible: cost, feasibility, time
- Big data explosion and history of running some experiments
- In most cases question scientific question deals with why and what will happen if we did something ?
- Spurious co-relations
  `https://www.tylervigen.com/spurious-correlations`



Seven Tools of Causal Inference

It's sometimes called Pillars and sometimes it becomes eight. Pearl keeps changing things.

- Encoding Causal Assumptions – Graphical Models and establishing testability
- Do-calculus and the control of confounding
- Causal Discovery
- The Algorithmization of Counterfactuals
- Mediation Analysis and the Assessment of Direct and Indirect Effects
- Adaptability, External Validity and Sample Selection Bias
- Recovering from Missing Data

Encoding Causal Assumptions – Graphical Models

Definition: A causal theory is a 4-tuple (Structural Causal Model):

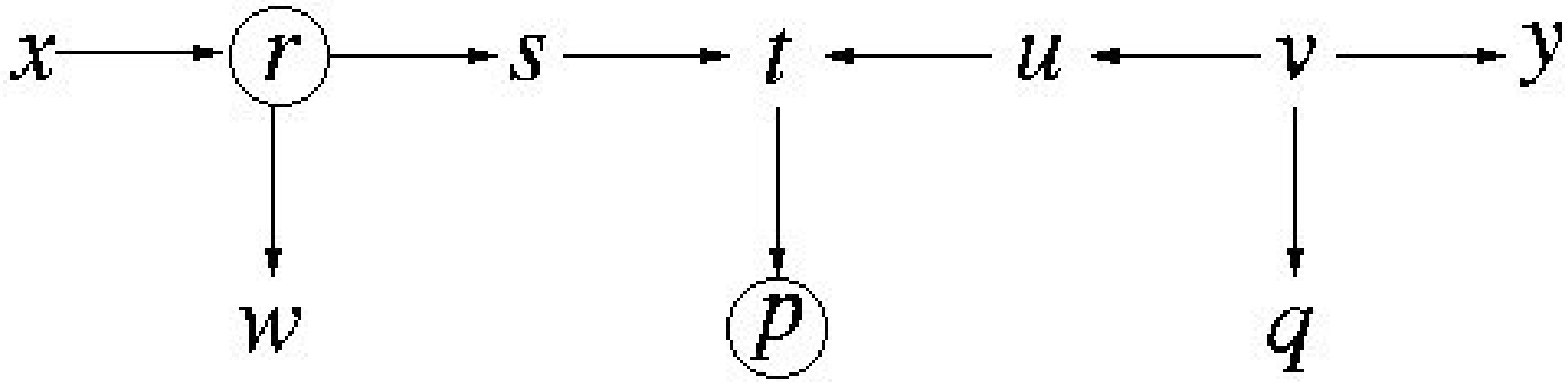
$$< V, U, P(U), f_i >$$

- V is the set of observables
- U is a set of unobserved variables that represent disturbances, assumptions, abnormalities, etc. (Noise Terms)
- P(U) is the distribution over U
- $X_i = f_i(V_1, V_2, ..., U_1, ...)$ is a set of functions that represents the causal nature
- **Magic: ★**

Encoding Causal Assumptions – Testability

- **d-separation:** *is a criterion for deciding, from a given a causal graph, whether a set X of variables is independent of another set Y, given a third set Z. The idea is to associate "dependence" with "connectedness" (i.e., the existence of a connecting path) and "independence" with "unconnected-ness" or "separation".*
- Causal influence flows one way through a DAG
- Statistical information can flow in either direction
- Application: Compute the regression co-efficients for

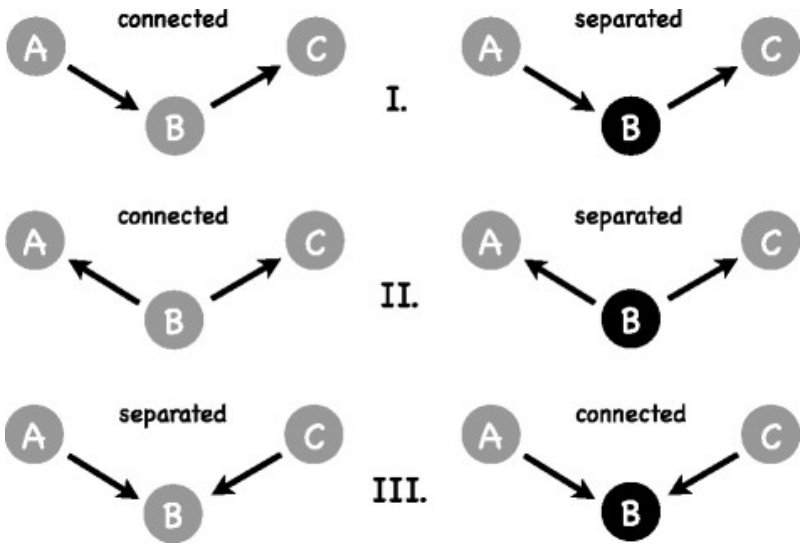
$$y = c_1 x + c_2 r + c_3 p$$

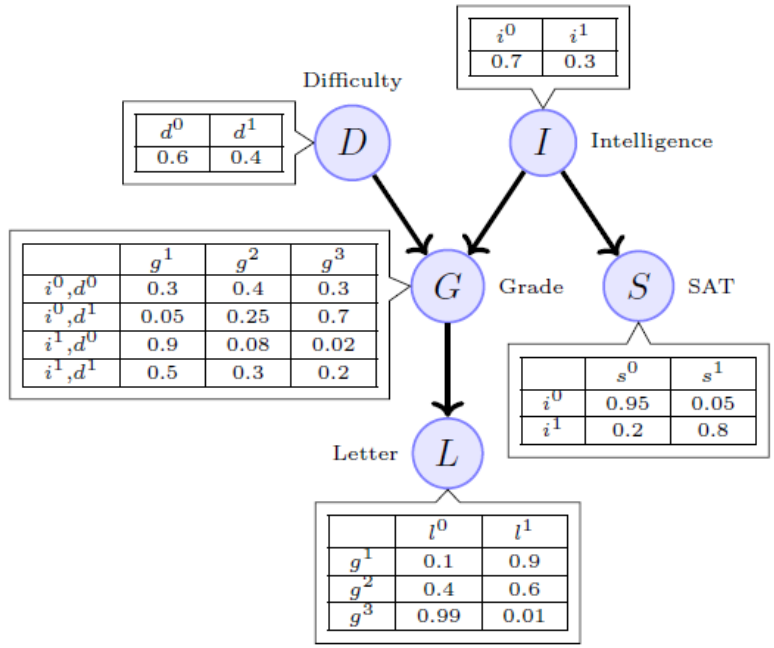


d-separation : Testability for Independence

Two exercise & Examples

- Given Letter of recommendation was given is there an association between SAT and Difficulty of the Exam ? D ⫫ S | L
- Problem of confounders : Grade can be correlated to SAT scores as they have the same parent Intelligence. Now say I gave a Grade “F” to a student. Can we determine the SAT score ?





- Intervention : $P(Y|do(X = x))$
- Easy way is experimentation, but we need to make sure we are controlling for X and not Z i.e. $P(Y|do(X))$ not $P(Y|do(X), Z)$ or $P(Y|do(X), do(Z))$
- Alternative mathematical rules
  - Identification: An aspect of a statistical model is identifiable when it cannot be changed without there also being some change in the distribution of the observable variables.

Identification

Example: $P(Y|X = x) \neq P(Y|do(X = x))$

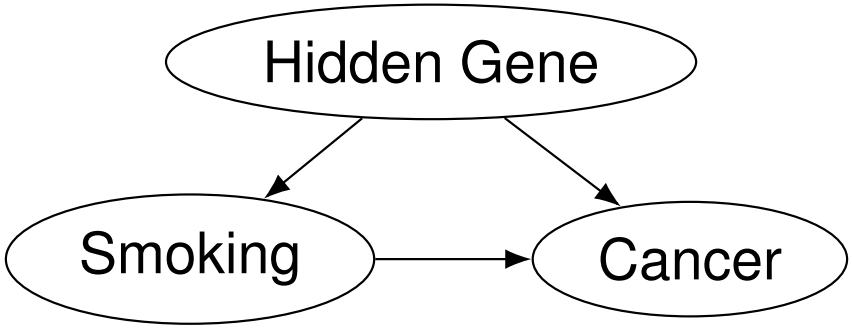


Figure: Casual Graph of Smoking Causes Cancer

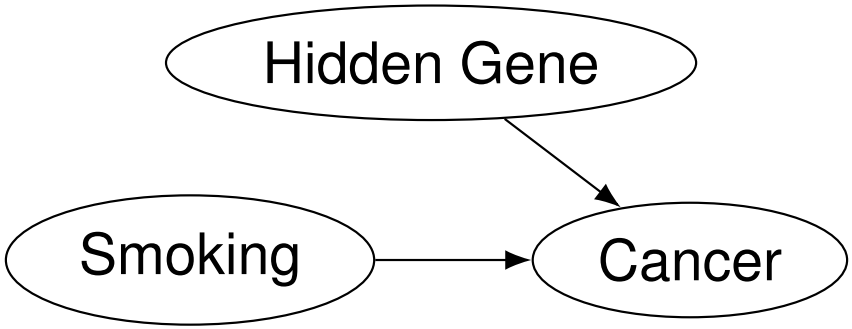


Figure: do(Smoking) Cause Cancer

Identification Strategies: Control of Confounding

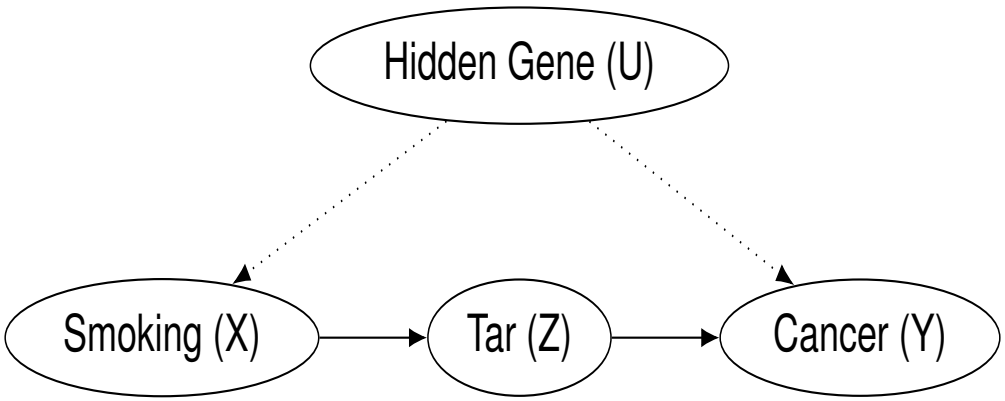
Calculate $P(Y|do(X = x))$

- Back-door criterion: ★
  $Pr(Y|do(X = x)) = \sum_s Pr(Y|X = x, S = s)Pr(S = s)$
  A set of conditioning variables or controls S satisfies the back-door criterion when
  - We know the causal graph
  - S blocks every back-door path between X and Y
  - S does not have any descendent of X
- Disjunctive cause criterion: ★ Control for all parents of the treatment variable, the effect variable (that are not descendants of the treatment), or both.

Front-door criterion ★

$$Pr(Y|do(X=x)) = \sum_s Pr(S=s|X=x) \sum_{x'} Pr(Y|X=x', S=s) Pr(X=x')$$

- S blocks all directed paths from X to Y
- There are no unblocked back-door paths from X to S
- X blocks all back-door paths from S to Y

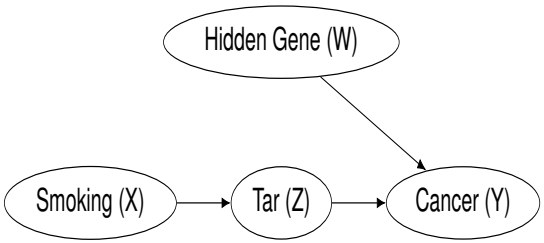


do-Calculus

Rules to calculate $P(Y|do(X = x))$

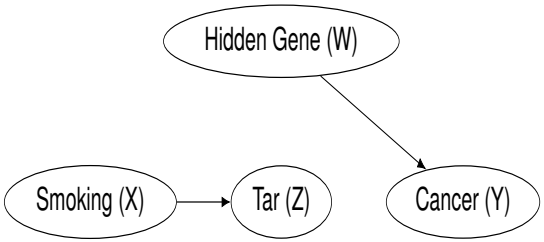
A graph G, W , X , Y , Z are disjoint subsets of the variables. $G_{\bar{X}}$ denotes the perturbed graph in which all edges pointing to X have been deleted, and $G_{\underline{X}}$ denotes the perturbed graph in which all edges pointing from X have been deleted. Z(W) denote the set of nodes in Z which are not ancestors of W.

- Insertion/deletion of observations ★: If $(Y \perp\!\!\!\perp Z \mid X, W)_{G_{\bar{X}}}$, then $P(y \mid do(x), Z, W) = p(Y|do(X), W)$

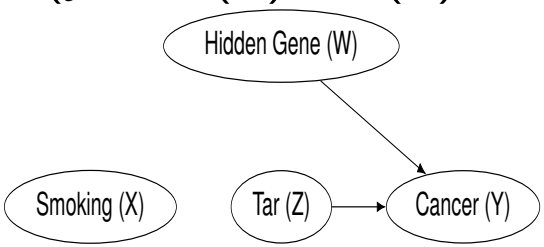


do-Calculus

- Action/observation exchange ★: If $(Y \perp\!\!\!\perp Z \mid X, W)_{G_{\bar{X},\underline{Z}}}$, then $P(y \mid do(x), do(z), W) = p(Y|do(X), z, W)$



- Insertion/deletion of actions ★: If $(Y \perp\!\!\!\perp Z \mid X, W)_{G_{\bar{X},\overline{Z(W)}}}$, then $P(y \mid do(x), do(z), W) = p(Y|do(X), W)$



Identification Task

- The objective is to solve for $P(Y \mid do(X))$.
- The three rules of do-Calculus can be applied to reduce $P(Y \mid do(X))$ to a form without the do-operator
- The rules of do-calculus do not themselves indicate the order in which they should be applied.
- *Existing algos: Tian J, Pearl J (2003). "On the Identification of Causal Effects." Technical report, Department of Computer Science, University of California, Los Angeles. R-290-L.*
- *Shpitser I, Pearl J (2006b). "Identification of Joint Interventional Distributions in Recursive Semi-Markovian Causal Models." In Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2, pp. 1219–1226. AAAI Press.*

Stochastic Interventions

A Calculus for Stochastic Interventions: Causal Effect Identification and Surrogate Experiments – Correa & Bareinboim AAAI 2020

- The paper came out a few days back
- Stochastic intervention generalizes over atomic interventions
- σ-Calculus

Causal Discovery

I am really interested in.

- Revealing Causal information by analyzing purely observational data, known as causal discovery
- Traditionally algorithms for identification of causal effects, or inferences about the effects of interventions, when the causal relations are completely or partially known, address a different class of problems (Pearl 2000)
- What asymmetries in the data gives you hints about the model ?
- Under certain circumstances under weak assumptions, causal queries can be estimated.

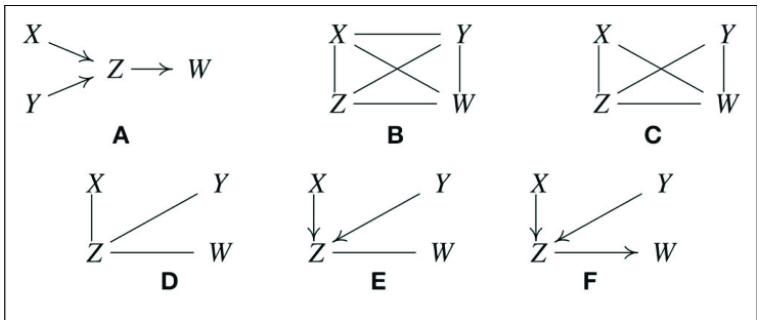
Causal Discovery from Data is a difficult problem

Equivalence classes

- They output (independence) equivalence classes, i.e., a set of causal structures satisfying the same conditional independences. Number of Equivalence class DAGs for n nodes
- Test for conditional independence i.e. $X \perp Y|Z \Leftrightarrow Pr(Y|X,Z) = Pr(Y|Z)$
- A Markov equivalence class is a set of DAGs that encode the same set of conditional independencies (Faithfulness assumptions). They have the same number of d-separations.

Constrained Based Method

- Check for all conditional independences i.e. check partial correlations
- Faithfulness/Markov assumption to find markov equivalence classes of all DAGs with same d-separation (Inductive Causation)
- Algorithms: IC, PC, FCI, RFCI, FCI



Restricted Structural Causal Models

A generalization over SCM

- A fundamental issue is given two variables, how to distinguish cause from effect.
- Independence based methods fail
- $\exists f, N_y$ and $\exists g, M_x$
  - $y = f(x, N_y)$ and $N_Y \perp\!\!\!\perp X$
  - $X = g(Y, M_x)$ and $M_Y \perp\!\!\!\perp X$
- $M_x \sim U(.)$ and g = inverse-cdf(X | Y)
- In general this will be applicable if f and g are complex

Restricted Structural Causal Models

A generalization over SCM

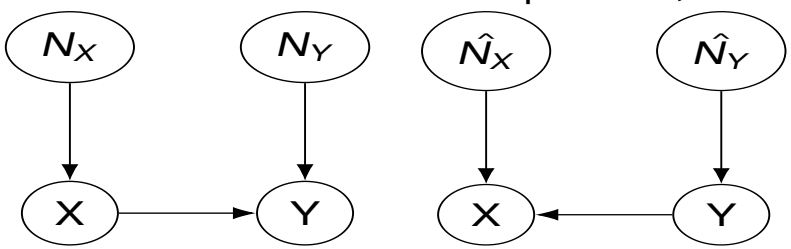
- What if we considered f and g to be simple ? i.e. Linear.

  S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. J. Kerminen. A linear non-gaussian acyclic model for causal discovery. Journal of Machine Learning Research, 7:2003-2030, 2006.

  - $Y = \alpha X + N_Y$ and $N_Y \perp\!\!\!\perp X$
  - If (N_Y, X) is non-Gaussian then X $\neq \beta$X + M_x i.e. Identifiability
  - In the inverse case the noise is not independent of the input
  - Regress Y $\sim$ f(X), test the independence between residuals and input
  - **Proof:**
    - $y = \beta X + N_Y$
    - $x = \theta Y + \hat{N_X}$
    - $\hat{N_X} = (1 - \theta\beta)X - \theta N_Y$
    - ★ Let X_1 , . . . , X_n be independent, non-degenerate random variables. If for two linear combinations

      $$l_1 = \sum a_i X_i, a_i \neq 0$$

      $l_2 = \sum b_i X_i, b_i \neq 0$
      are independent, then each Xi is normally distributed.



Counterfactuals

What if we had done that ?

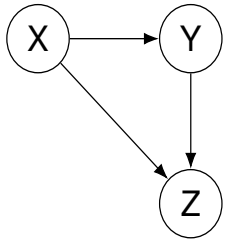
Counterfactual Fairness example – *Given that Alice did not get promoted in her job, and given that she is a woman, and given everything else we can observe about her circumstances and performance, what is the probability of her getting a promotion if she was a man instead?*

$$P(Promotion = True | Alice = Woman, Promotion = False, do(Alice = Man))$$

Computing Counterfactuals

$$X = b \times N_X$$
$$Y = a \times X + N_Y$$
$$Z = 2 \times Y + X + N_Z$$

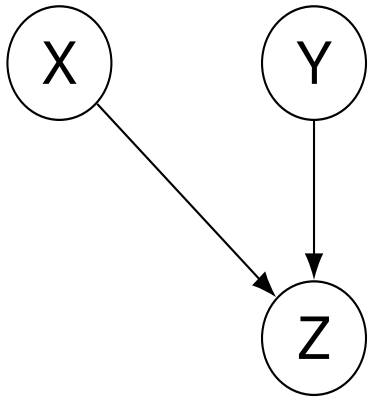
$$N_X, N_Y, N_Z \sim U(-1, 1, 2) \sim (\frac{1}{2}, \frac{1}{3}, \frac{1}{6})$$
$$(X, Y, Z) = (1, 2, 4)$$
$$(a, b) = 1$$
$$(N_X, N_Y, N_Z) = (1, 2, 1)$$



Steps: P($Y_{X=2} > 3$)

- Compute Y = 1 $\times 2 + N_y > 3$
- Y > 3 only if $N_y = 2$
- $P(Y_{X=2} > 3) = \frac{1}{6}$

Exercise: P($Z_{X=2} > 4$)



Counterfactuals: Markov Factor Replacement

Importance Sampling Approach

- Causal graph can be factorized using Markov factorization : $P(Z|Y,X)P(Y|X)P(X)$
- Intervention amounts to replacing the factor $P^*(Y|X)$
- ω be a shorthand for all variables appearing in the Markov factorization
- Compute counterfactual expectation with Importance sampling $\int_\omega zP^*(\omega) = \int_\omega z\frac{P^*(Y|X)}{P(Y|X)}P(\omega) = \frac{1}{n}\sum_{n=1}^{N} z_i\frac{P^*(Y_i|X_i)}{P(Y_i|X_i)}$
- More things to do: Compute Confidence Intervals and interpreting them, Different bound calculation – *Counterfactual Reasoning and Learning Systems– Bottou et. al.*

- **Task:** Given data and model quantify the causal pathway
- **Result:** Taking use of counterfactuals one can say when direct and indirect effects are estimable from data only.

Transfer Learning

I really don't know what's going on here

Checkout this guy – Elias Bareinboim

Checkout what your neighbor, Ilya Shiptser, and Karthika Mohan are upto.

References

What you should start with – Books

- *Advanced Data Analysis from an Elementary Point of View* Cosma Rohilla
- *Peters, Jonas, Dominik Janzing, and Bernhard Schölkopf. Elements of causal inference: foundations and learning algorithms. MIT press, 2017.*
- *Pearl, Judea. Causality. Cambridge university press, 2009.*
- *Pearl, Judea, Madelyn Glymour, and Nicholas P. Jewell. Causal inference in statistics: A primer. John Wiley Sons, 2016.*

References

Survey

- Judea Pearl. 2018. The Seven Tools of Causal Inference with Reflections on Machine Learning.
- Pearl, Judea. Causal inference in statistics: An overview. Statistics surveys 3 (2009): 96-146.

Causal Discovery

- LiNGAM
- *Spirtes, Peter, and Kun Zhang. "Causal discovery and inference: concepts and recent methodological advances." Applied informatics. Vol. 3. No. 1. Springer Berlin Heidelberg, 2016.*
- *Introduction to the Foundations of Causal Discovery* Frederick Eberhardt

Counterfactuals

- Bottou, Léon, et al. "Counterfactual reasoning and learning systems: The example of computational advertising." The Journal of Machine Learning Research 14.1 (2013): 3207-3260.