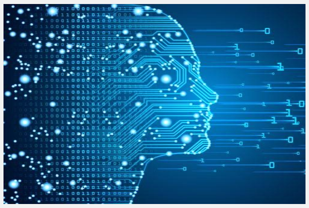
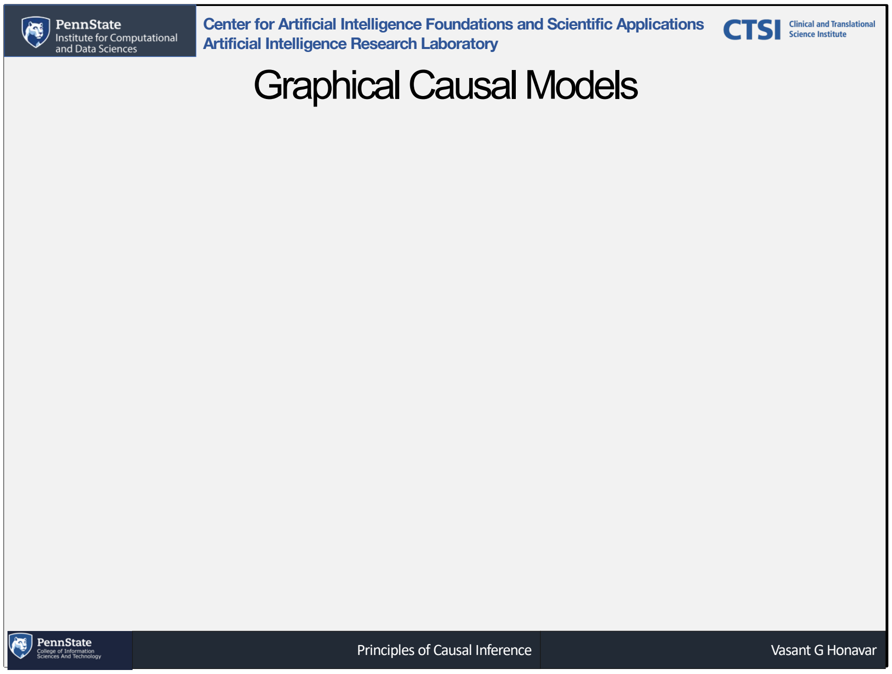


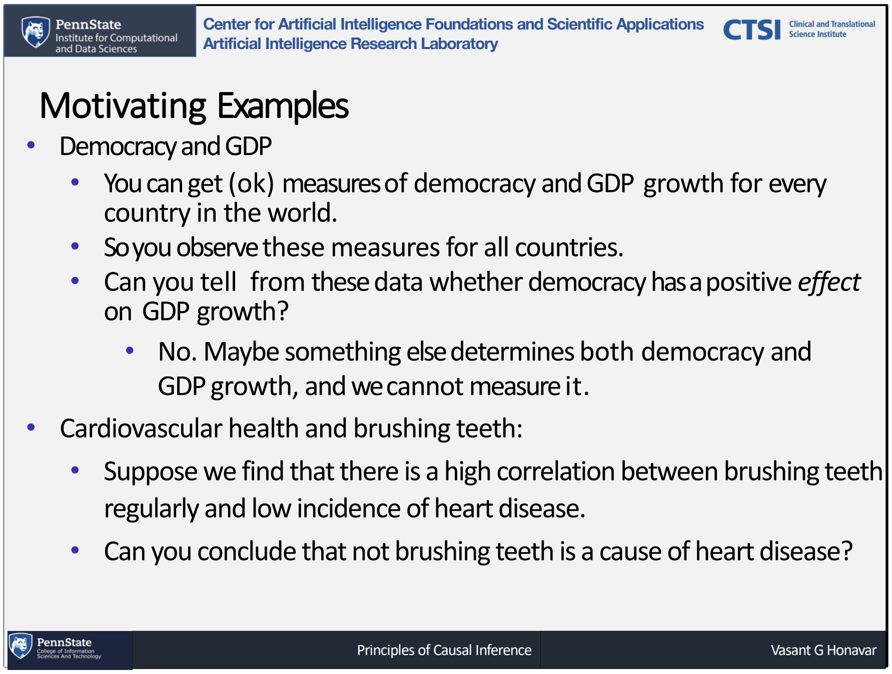
Principles of Causal Inference

Vasant G. Honavar

Dorothy Foehr Huck and J. Lloyd Huck Chair in Biomedical Data Sciences and Artificial Intelligence
Professor of Data Sciences, Informatics, Computer Science and Engineering, Bioinformatics & Genomics,
Public Health Sciences and Neuroscience
Center for Artificial Intelligence Foundations and Scientific Applications
Institute for Computational and Data Sciences
Huck Institutes of the Life Sciences
Clinical and Translational Sciences Institute
Northeast Big Data Hub
Pennsylvania State University
vhonavar@psu.edu
http://faculty.ist.psu.edu/vhonavar
http://ailab.ist.psu.edu

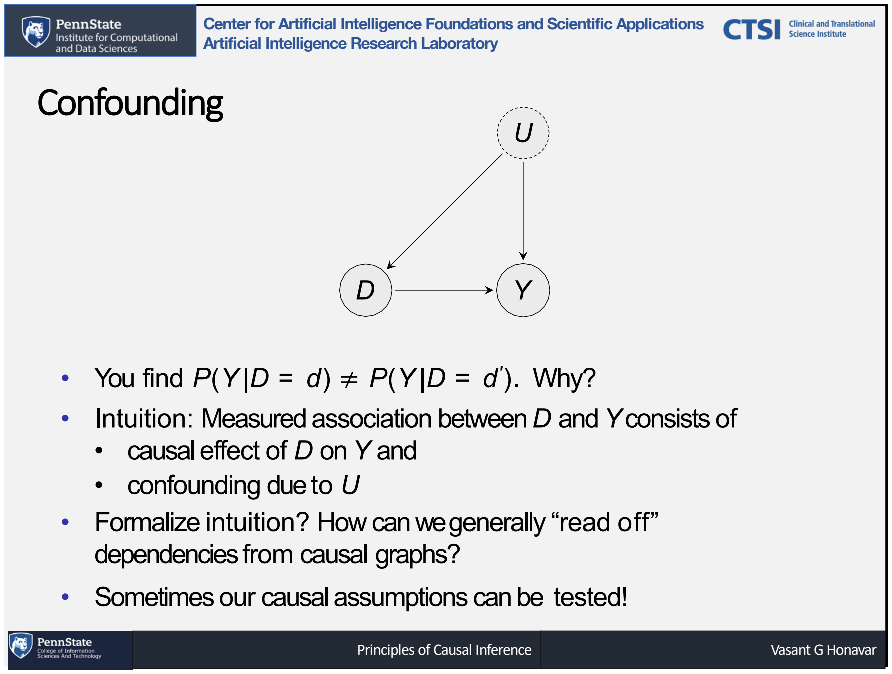


Graphical Causal Models



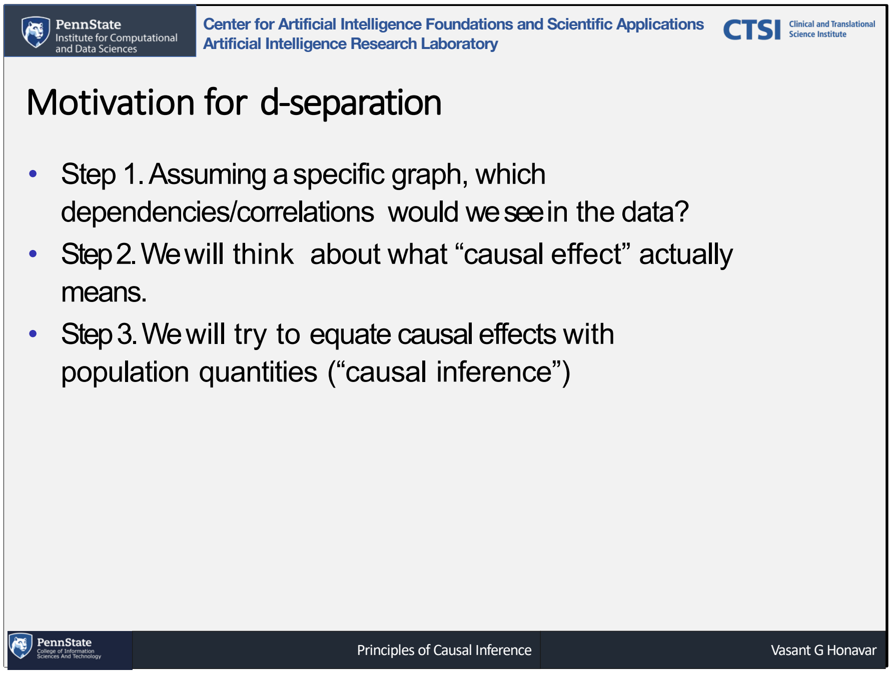
Motivating Examples

- Democracy and GDP
  - You can get (ok) measures of democracy and GDP growth for every country in the world.
  - So you observe these measures for all countries.
  - Can you tell from these data whether democracy has a positive *effect* on GDP growth?
    - No. Maybe something else determines both democracy and GDP growth, and we cannot measure it.
- Cardiovascular health and brushing teeth:
  - Suppose we find that there is a high correlation between brushing teeth regularly and low incidence of heart disease.
  - Can you conclude that not brushing teeth is a cause of heart disease?



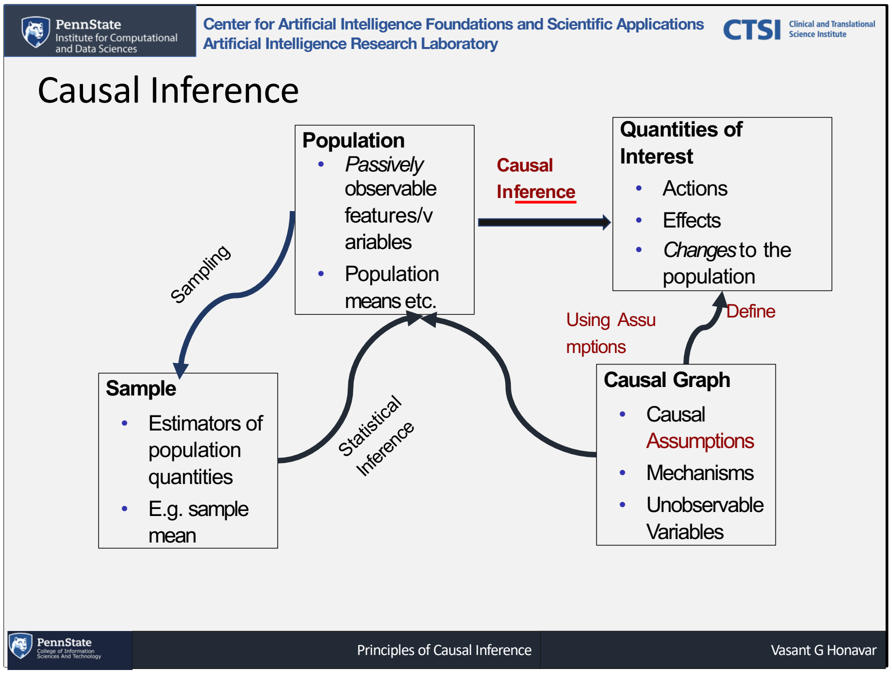
Confounding

- You find $P(Y|D = d) \neq P(Y|D = d')$. Why?
- Intuition: Measured association between *D* and *Y* consists of
  - causal effect of *D* on *Y* and
  - confounding due to *U*
- Formalize intuition? How can we generally “read off” dependencies from causal graphs?
- Sometimes our causal assumptions can be tested!



Motivation for d-separation

- Step 1. Assuming a specific graph, which dependencies/correlations would we see in the data?
- Step 2. We will think about what “causal effect” actually means.
- Step 3. We will try to equate causal effects with population quantities (“causal inference”)



PennState Institute for Computational and Data Sciences

Center for Artificial Intelligence Foundations and Scientific Applications
Artificial Intelligence Research Laboratory

CTSI Clinical and Translational Science Institute

Causal Inference

Population
- *Passively* observable features/variables
- Population means etc.

Causal Inference

Quantities of Interest
- Actions
- Effects
- *Changes* to the population

Sampling

Sample
- Estimators of population quantities
- E.g. sample mean

Statistical Inference

Using Assumptions

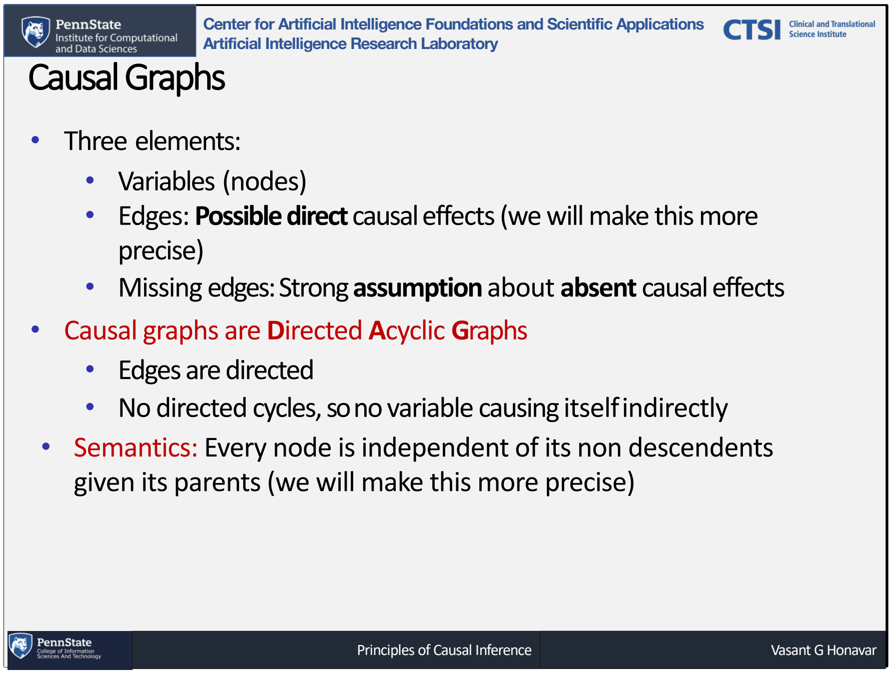
Define

Causal Graph
- Causal Assumptions
- Mechanisms
- Unobservable Variables

PennState College of Information Sciences and Technology

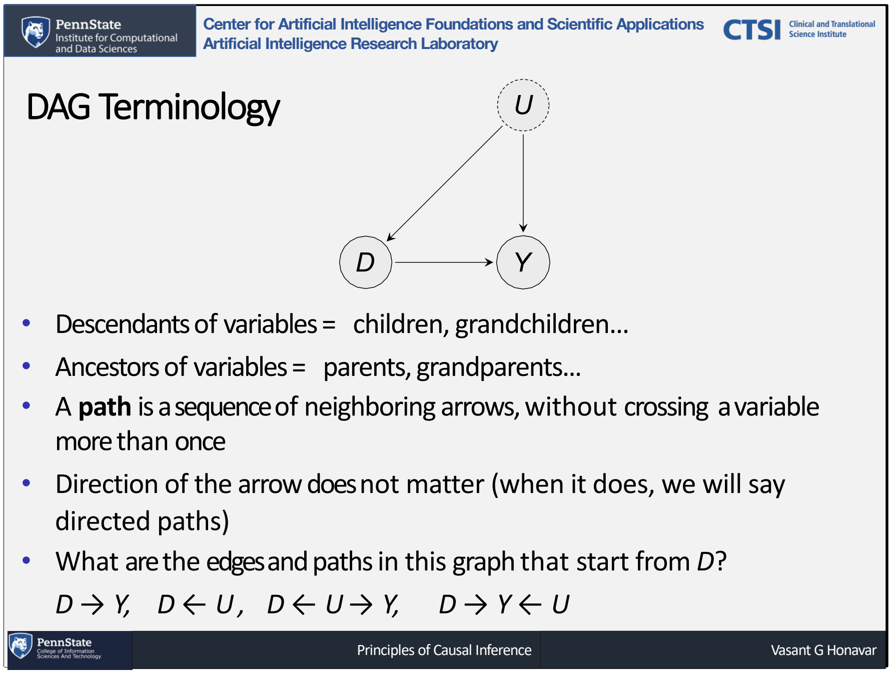
Principles of Causal Inference

Vasant G Honavar



Causal Graphs

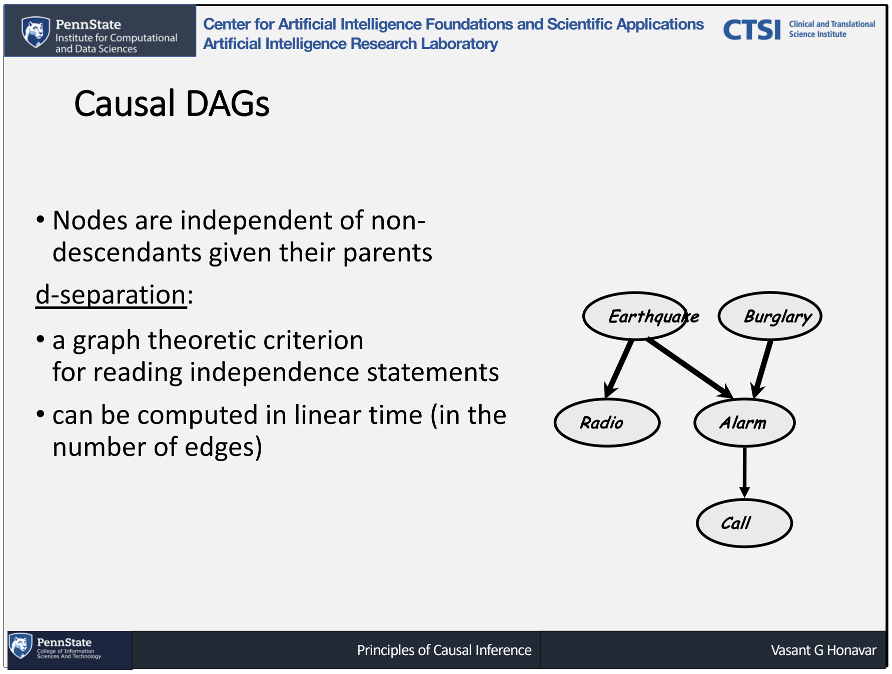
- Three elements:
  - Variables (nodes)
  - Edges: **Possible direct** causal effects (we will make this more precise)
  - Missing edges: Strong **assumption** about **absent** causal effects
- Causal graphs are **D**irected **A**cyclic **G**raphs
  - Edges are directed
  - No directed cycles, so no variable causing itself indirectly
- Semantics: Every node is independent of its non descendents given its parents (we will make this more precise)



DAG Terminology

- Descendants of variables = children, grandchildren…
- Ancestors of variables = parents, grandparents…
- A **path** is a sequence of neighboring arrows, without crossing a variable more than once
- Direction of the arrow does not matter (when it does, we will say directed paths)
- What are the edges and paths in this graph that start from *D*?

  $D \to Y$, $D \leftarrow U$, $D \leftarrow U \to Y$, $D \to Y \leftarrow U$

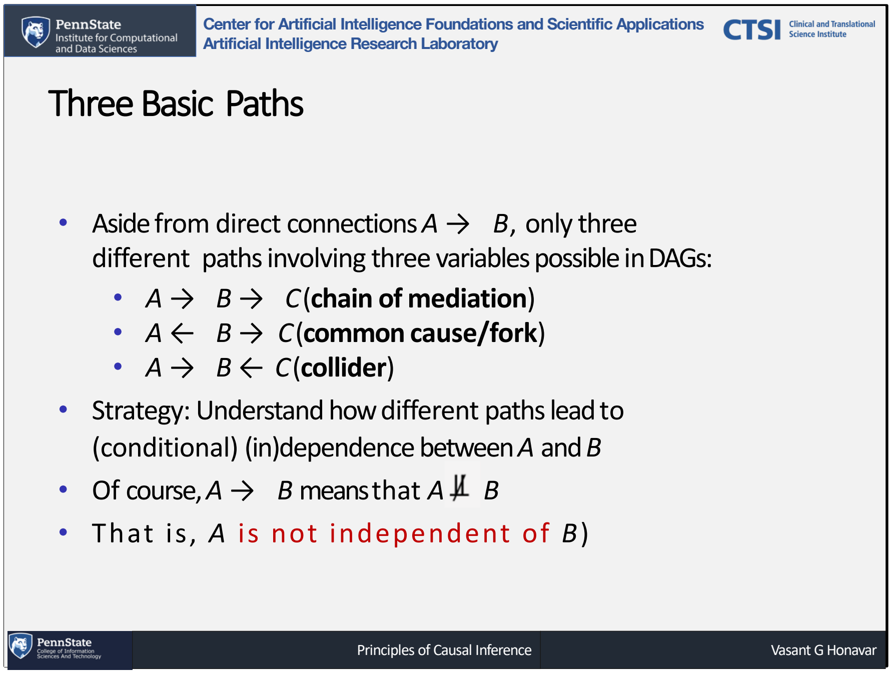


Causal DAGs

- Nodes are independent of non-descendants given their parents

<u>d-separation</u>:

- a graph theoretic criterion for reading independence statements
- can be computed in linear time (in the number of edges)



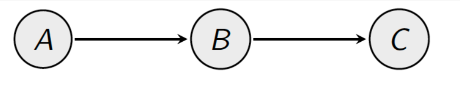
Three Basic Paths

- Aside from direct connections $A \rightarrow B$, only three different paths involving three variables possible in DAGs:
  - $A \rightarrow B \rightarrow C$ (**chain of mediation**)
  - $A \leftarrow B \rightarrow C$ (**common cause/fork**)
  - $A \rightarrow B \leftarrow C$ (**collider**)
- Strategy: Understand how different paths lead to (conditional) (in)dependence between A and B
- Of course, $A \rightarrow B$ means that $A \not\perp\!\!\!\perp B$
- That is, A is not independent of B)

$a \perp\!\!\!\perp b \mid c$

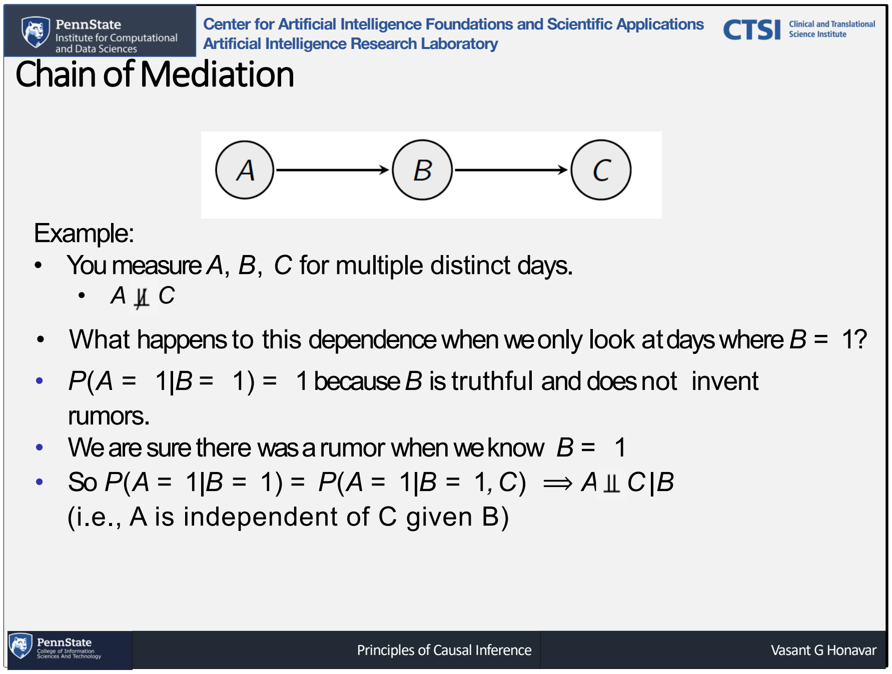
$a \not\perp\!\!\!\perp b \mid c$

Chain of Mediation



- Rumor A on a given day ($P(A = 1) = 0.5$),
- Person B knows the rumor sometimes ($B = 1$),
- Person B sometimes spreads rumor to person C ($C = 1$).
- B or C do not invent rumors beyond A
- You measure A, B, C for multiple distinct days.
  - Will C be informative about A?
    - Yes, when C knows a rumor, a rumor A definitely occurred
    - $P(A = 1|C = 1) = 1 > P(A = 1) = 0.5$ or $A \not\perp\!\!\!\perp C$

- What happens to this dependence when we only look at days when $B = 1$?

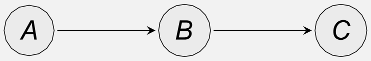


Chain of Mediation

Example:

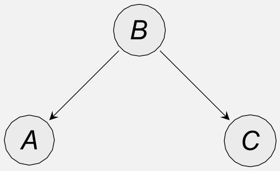
- You measure A, B, C for multiple distinct days.
  - $A \not\perp\!\!\!\perp C$
- What happens to this dependence when we only look at days where $B = 1$?
- $P(A = 1|B = 1) = 1$ because B is truthful and does not invent rumors.
- We are sure there was a rumor when we know $B = 1$
- So $P(A = 1|B = 1) = P(A = 1|B = 1, C) \implies A \perp\!\!\!\perp C|B$
  (i.e., A is independent of C given B)

Chain of Mediation



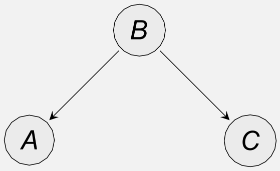
- In chain of mediation: $A \not\perp\!\!\!\perp C$, but $A \perp\!\!\!\perp C|B$
- We say this path is **open** unconditionally, but conditional on the middle node it is **blocked**
- As in “blocking the information flow”
- Note that $P(A|C = 1) > P(A)$, so that C predicts A, while the causal influence actually flows along $A \rightarrow B \rightarrow C$.

Common Cause

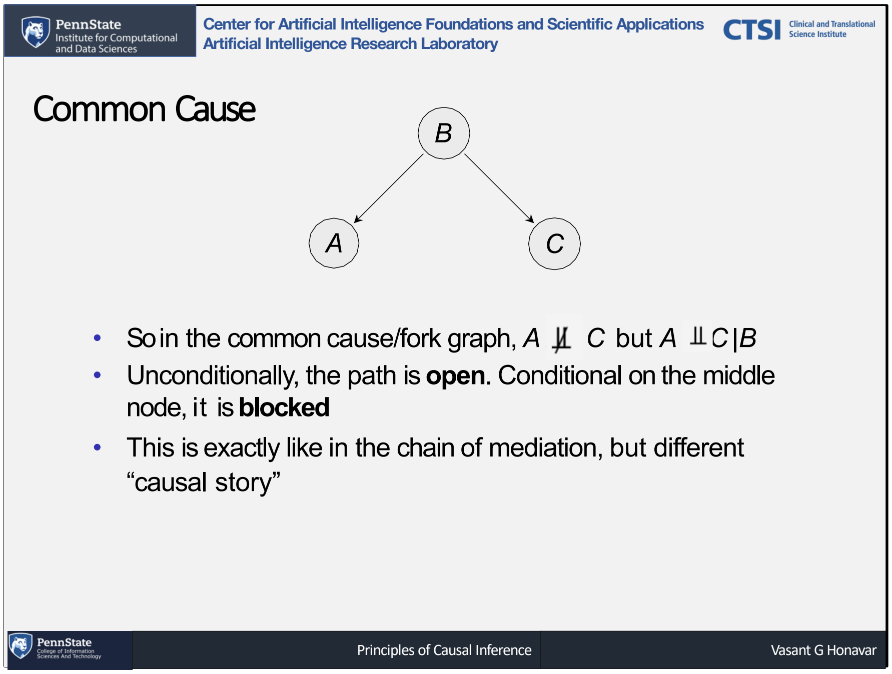


- Example: *B* denotes Season; A, daily ice cream sales in the city; and C, daily number of drownings in the city
- Question: Does knowing *C* help you predict *A*?
    - *B* is a common cause of both *A* and *C*
    - *A* and *C* are correlated
    - Knowing the number of drownings is high, implies probably *B* = summer, and that means ice cream sales *C* are relatively high
    - So $P(A) \neq P(A|C)$, or $A \not\perp C$

Common Cause



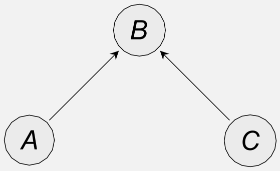
- Example: *B* denotes Season; A, daily ice cream sales in the city; and C, daily number of drownings in the city. $A \not\perp\!\!\!\perp C$
- What if we look at a subset of the data where season is held constant, e.g. $B =$ summer?
  - During summer, you see still variation in ice cream sales and drownings
  - But once you fix *B*, changes in B cannot influence A or C.
  - $A \perp\!\!\!\perp C | B$



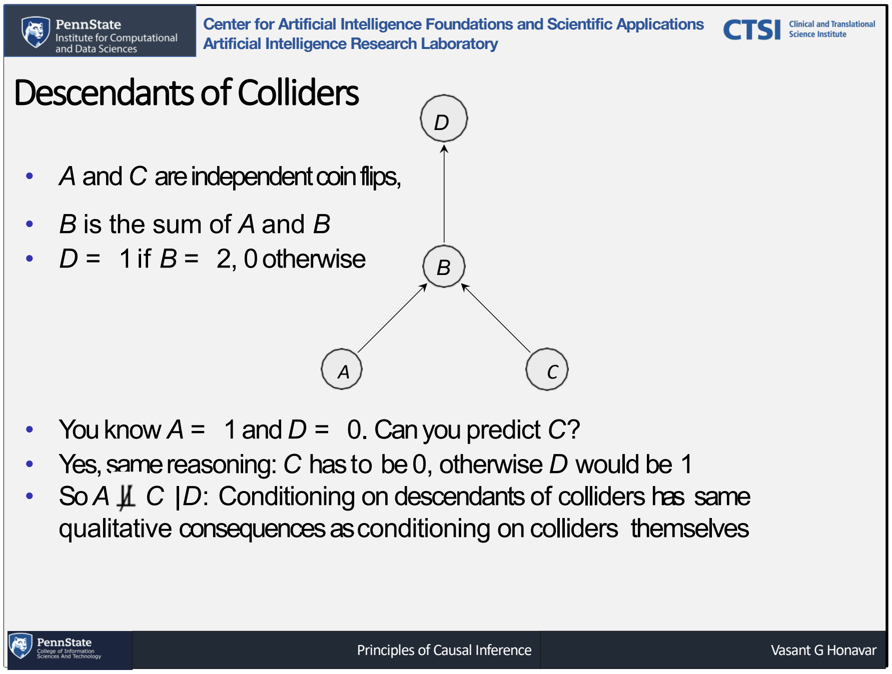
Common Cause

- So in the common cause/fork graph, $A \not\perp\!\!\!\perp C$ but $A \perp\!\!\!\perp C | B$
- Unconditionally, the path is **open**. Conditional on the middle node, it is **blocked**
- This is exactly like in the chain of mediation, but different “causal story”

Collider



- Example: *A* and *C* denote independent coin flips (results are 0 or 1).
  - Suppose *B* is the sum of A and B, so 0, 1, or 2
- Can you predict *A* when you know *C* ?
  - No! Because $A \perp\!\!\!\perp C$
- But suppose you know one coin *A* = 1 and *B* = 1. Can you then predict the other coin *C* ?
  - Yes, it HAS to be 0
- So here $A \perp\!\!\!\perp C$, but $A \not\perp\!\!\!\perp C | B$

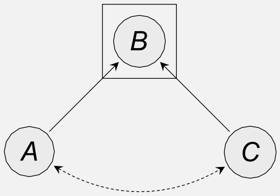


Descendants of Colliders

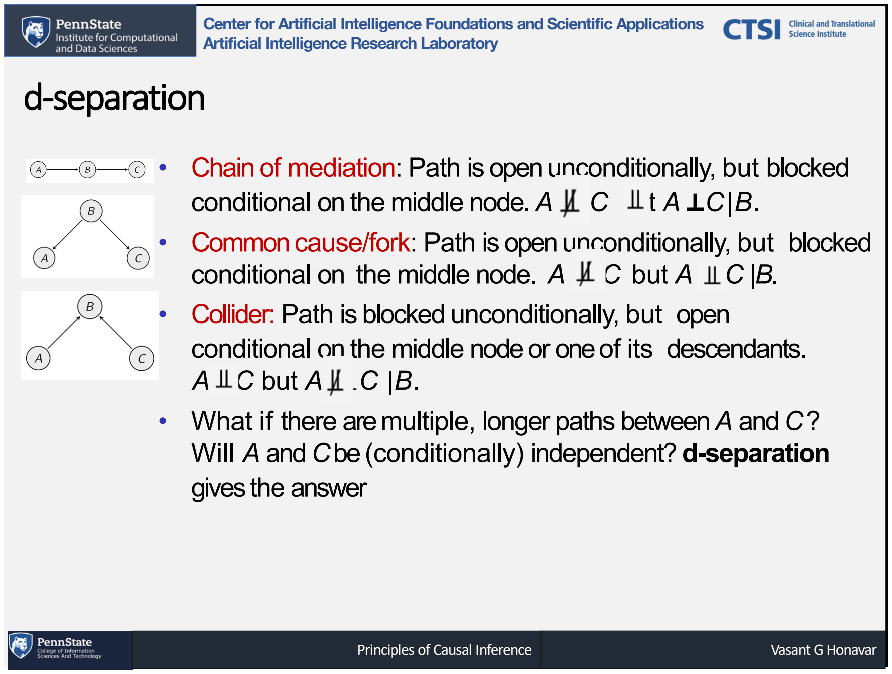
- *A* and *C* are independent coin flips,
- *B* is the sum of *A* and *B*
- $D = 1$ if $B = 2$, 0 otherwise

- You know $A = 1$ and $D = 0$. Can you predict *C*?
- Yes, same reasoning: *C* has to be 0, otherwise *D* would be 1
- So $A \not\perp\!\!\!\perp C \mid D$: Conditioning on descendants of colliders has same qualitative consequences as conditioning on colliders themselves

Conditioning on Colliders: Visualization

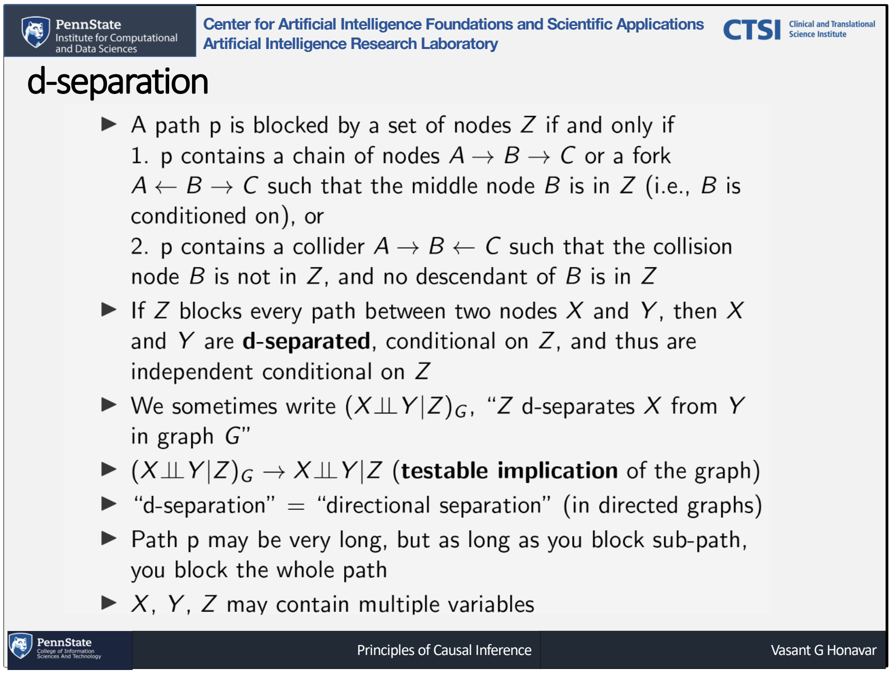


- Sometimes helpful: Boxed variable indicates we condition on it
- Bi-directed arc between the start and end node indicates associations created by conditioning on collider
- **Does not indicate cyclic causation**; rather equivalent to an additional common cause of the nodes connected by the bi-directed arc
- Is then treated as a normal path



d-separation

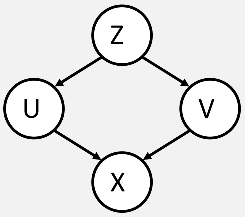
- Chain of mediation: Path is open unconditionally, but blocked conditional on the middle node. $A \not\perp\!\!\!\perp C$ $\perp\!\!\!\perp$t $A \perp C|B$.
- Common cause/fork: Path is open unconditionally, but blocked conditional on the middle node. $A \not\perp\!\!\!\perp C$ but $A \perp\!\!\!\perp C|B$.
- Collider: Path is blocked unconditionally, but open conditional on the middle node or one of its descendants. $A \perp\!\!\!\perp C$ but $A \not\perp\!\!\!\perp .C|B$.
- What if there are multiple, longer paths between A and C? Will A and C be (conditionally) independent? **d-separation** gives the answer



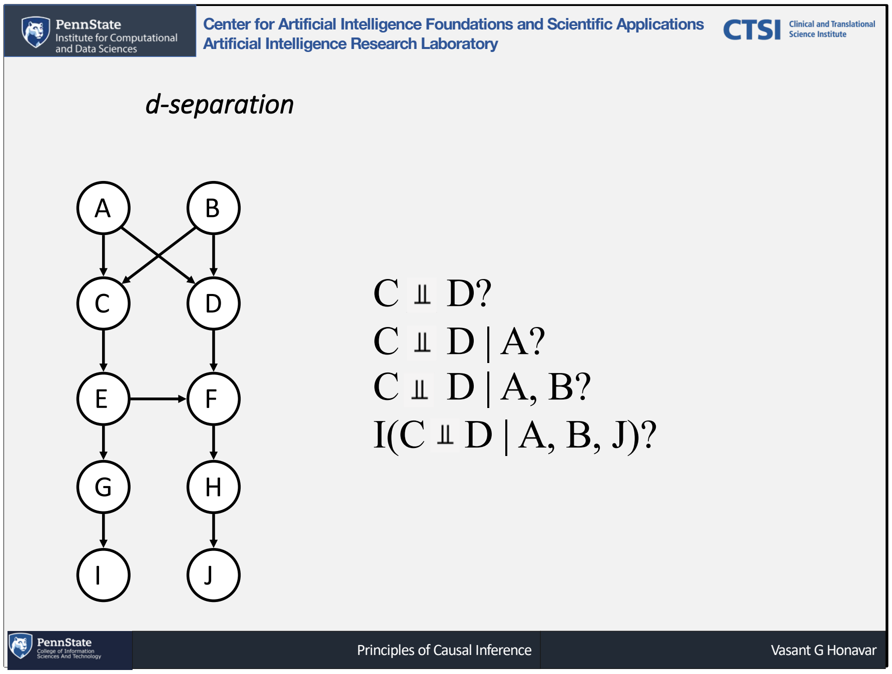
d-separation

- A path p is blocked by a set of nodes Z if and only if
  1. p contains a chain of nodes $A \rightarrow B \rightarrow C$ or a fork $A \leftarrow B \rightarrow C$ such that the middle node B is in Z (i.e., B is conditioned on), or
  2. p contains a collider $A \rightarrow B \leftarrow C$ such that the collision node B is not in Z, and no descendant of B is in Z
- If Z blocks every path between two nodes X and Y, then X and Y are **d-separated**, conditional on Z, and thus are independent conditional on Z
- We sometimes write $(X \perp\!\!\!\perp Y|Z)_G$, "$Z$ d-separates X from Y in graph G"
- $(X \perp\!\!\!\perp Y|Z)_G \rightarrow X \perp\!\!\!\perp Y|Z$ (**testable implication** of the graph)
- "d-separation" = "directional separation" (in directed graphs)
- Path p may be very long, but as long as you block sub-path, you block the whole path
- X, Y, Z may contain multiple variables

d-separation?



- $X \perp\!\!\!\perp Z$? No.
- $X \perp\!\!\!\perp Z \mid U$? No
- $X \perp\!\!\!\perp Z \mid U, V$? Yes



PennState Institute for Computational and Data Sciences

Center for Artificial Intelligence Foundations and Scientific Applications
Artificial Intelligence Research Laboratory

CTSI Clinical and Translational Science Institute

d-separation

$C \perp\!\!\!\perp D$?

$C \perp\!\!\!\perp D \mid A$?

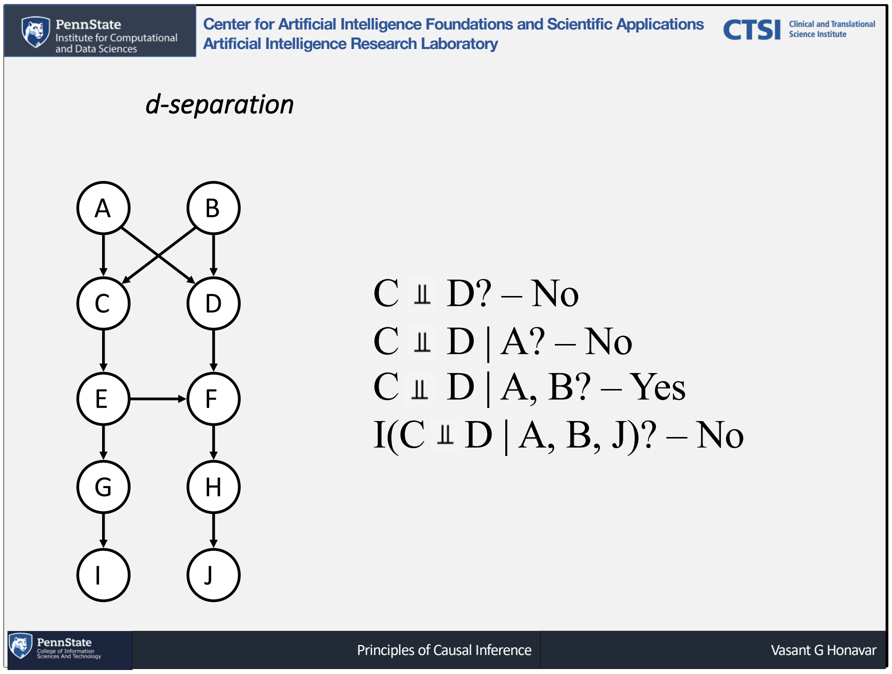
$C \perp\!\!\!\perp D \mid A, B$?

$I(C \perp\!\!\!\perp D \mid A, B, J)$?

PennState College of Information Sciences And Technology

Principles of Causal Inference

Vasant G Honavar



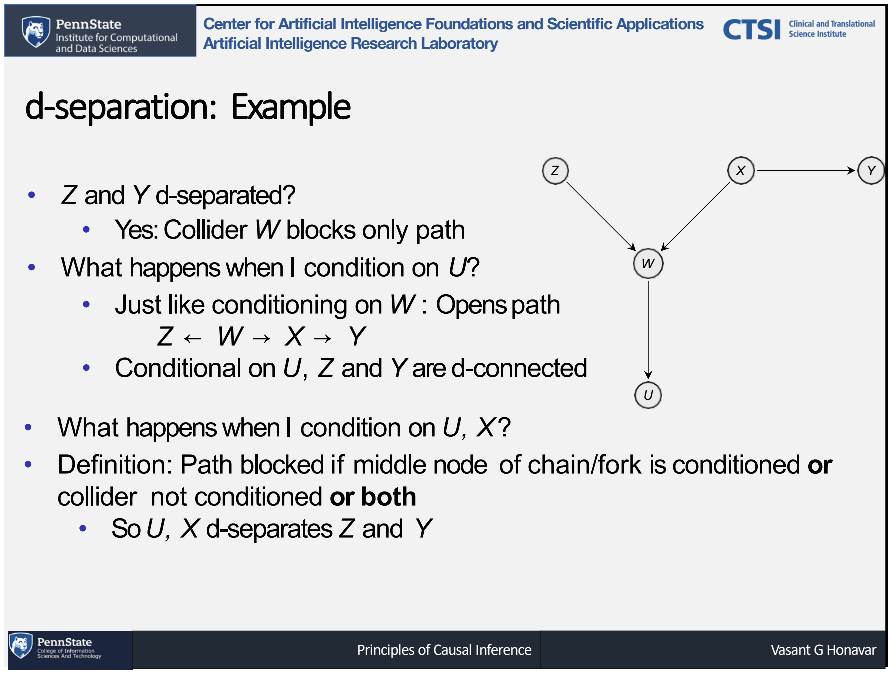
d-separation

$C \perp\!\!\!\perp D$? – No

$C \perp\!\!\!\perp D \mid A$? – No

$C \perp\!\!\!\perp D \mid A, B$? – Yes

$I(C \perp\!\!\!\perp D \mid A, B, J)$? – No

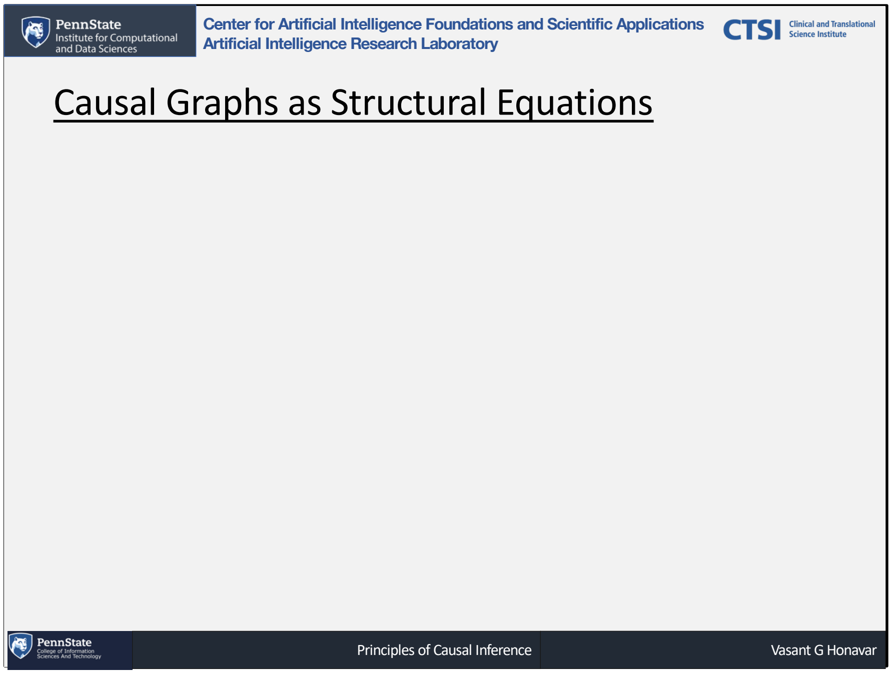


d-separation: Example

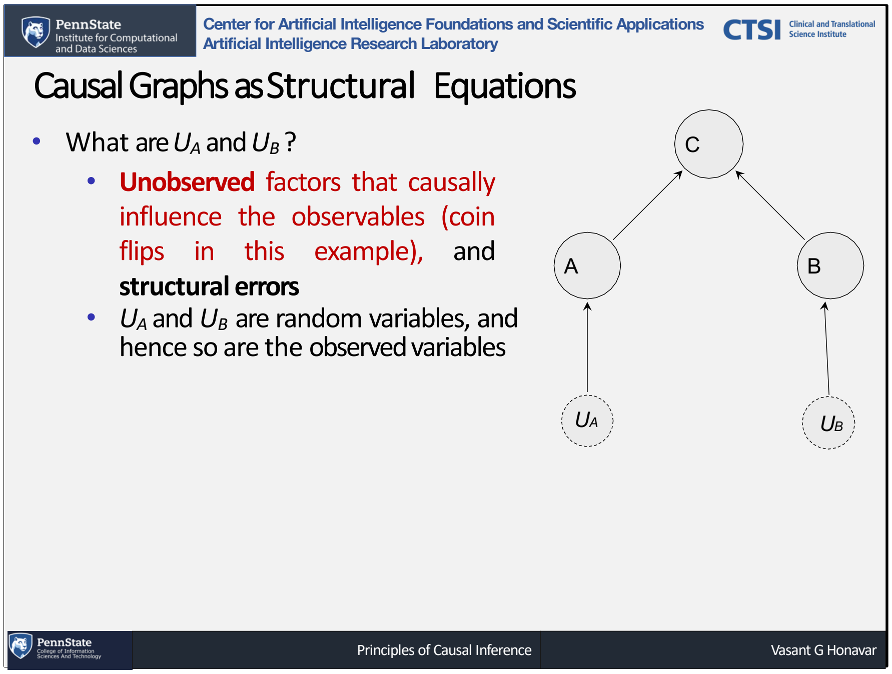
- *Z* and *Y* d-separated?
  - Yes: Collider *W* blocks only path
- What happens when I condition on *U*?
  - Just like conditioning on *W* : Opens path $Z \leftarrow W \rightarrow X \rightarrow Y$
  - Conditional on *U*, *Z* and *Y* are d-connected
- What happens when I condition on *U, X*?
- Definition: Path blocked if middle node of chain/fork is conditioned **or** collider not conditioned **or both**
  - So *U, X* d-separates *Z* and *Y*

d-separation

- Theorem [Verma & Pearl, 1998]: If a set of evidence variables *E* *d*-separates *X* and *Z* in an SCM then $X \perp\!\!\!\perp Z \mid E$
- *d*-separation can be computed in linear time using a depth-first search like algorithm.
- *d*-separation can be used to test whether finding out about the value of one variable might give us any additional hints about some other variable, given what we already know.

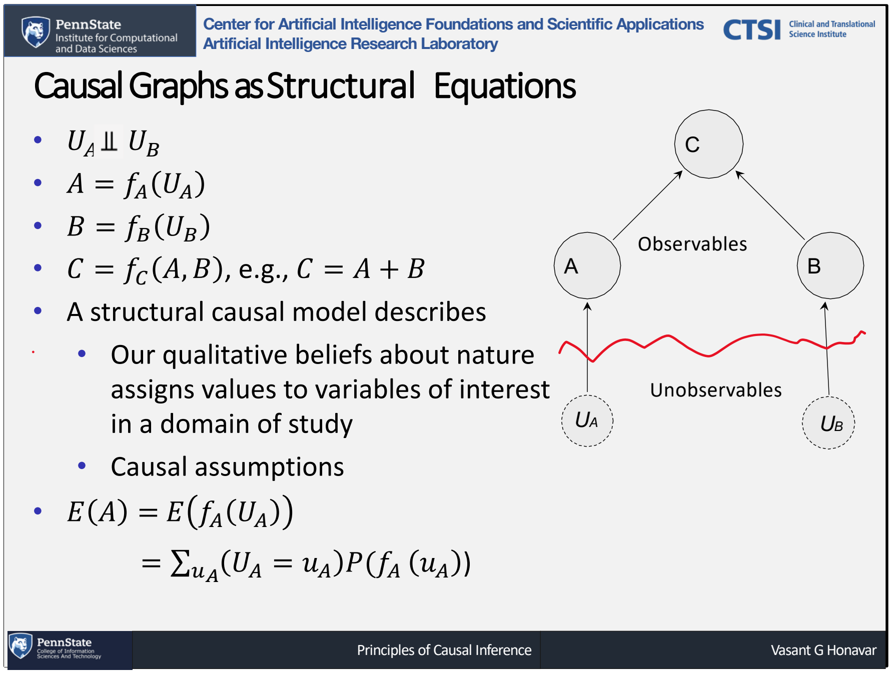


Causal Graphs as Structural Equations



Causal Graphs as Structural Equations

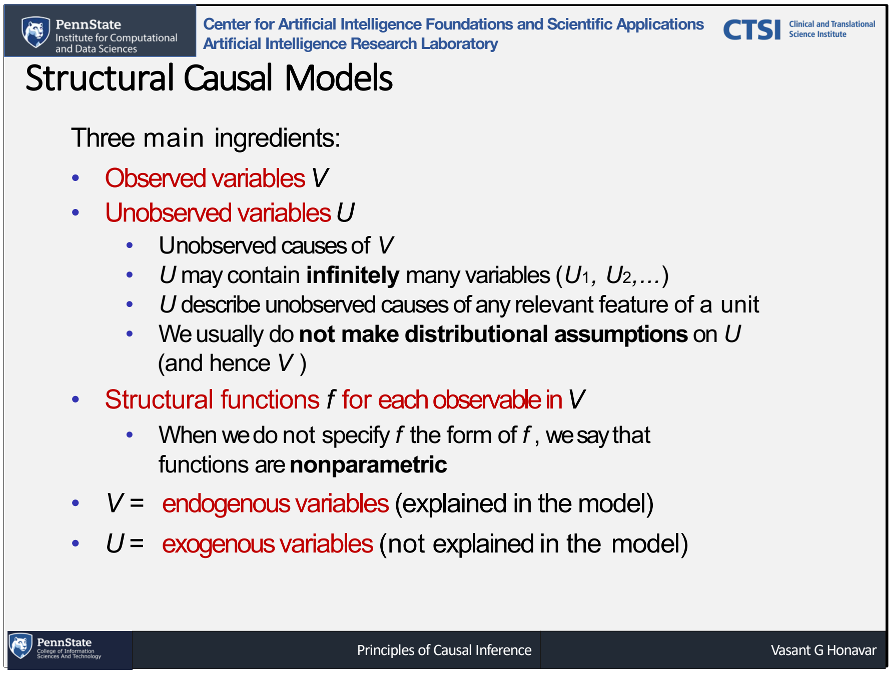
- What are U_A and U_B ?
  - **Unobserved** factors that causally influence the observables (coin flips in this example), and **structural errors**
  - U_A and U_B are random variables, and hence so are the observed variables



Causal Graphs as Structural Equations

- $U_A \perp\!\!\!\perp U_B$
- $A = f_A(U_A)$
- $B = f_B(U_B)$
- $C = f_C(A, B)$, e.g., $C = A + B$
- A structural causal model describes
  - Our qualitative beliefs about nature assigns values to variables of interest in a domain of study
  - Causal assumptions
- $E(A) = E\big(f_A(U_A)\big)$

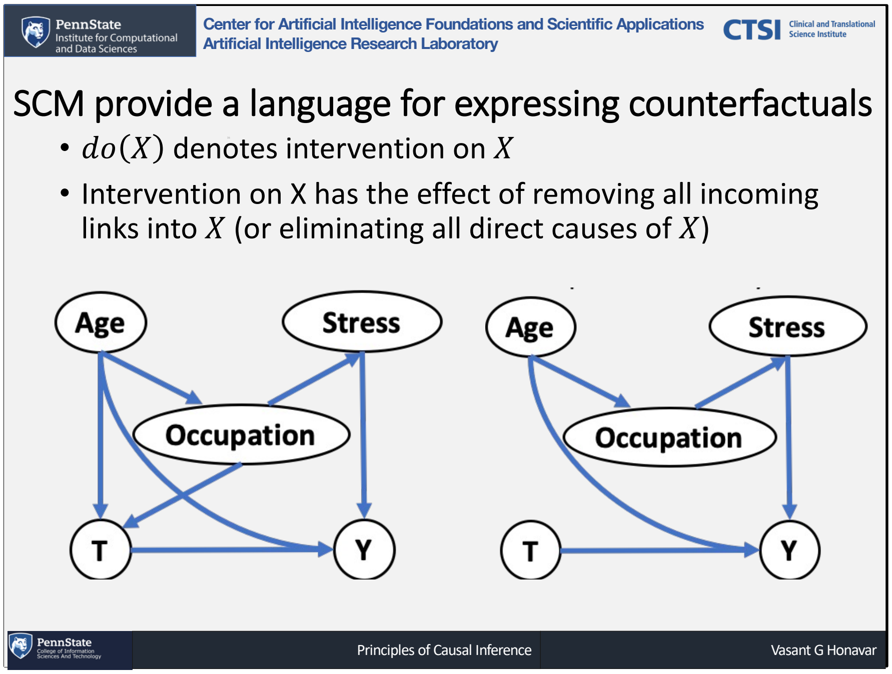
$$= \sum_{u_A}(U_A = u_A)P(f_A\,(u_A))$$



Structural Causal Models

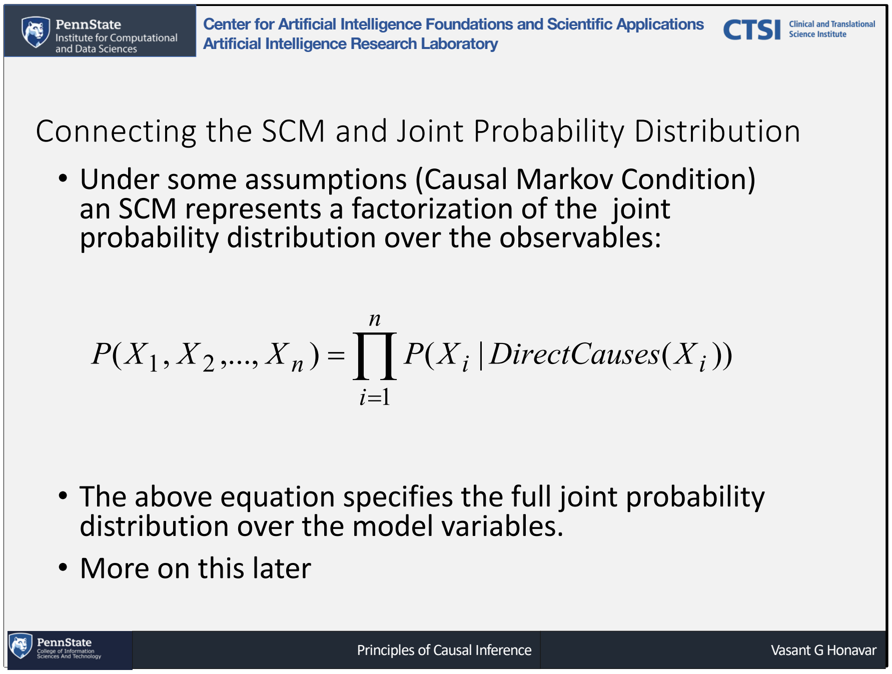
Three main ingredients:

- Observed variables *V*
- Unobserved variables *U*
  - Unobserved causes of *V*
  - *U* may contain **infinitely** many variables ($U_1, U_2, \ldots$)
  - *U* describe unobserved causes of any relevant feature of a unit
  - We usually do **not make distributional assumptions** on *U* (and hence *V*)
- Structural functions *f* for each observable in *V*
  - When we do not specify *f* the form of *f* , we say that functions are **nonparametric**
- *V* = endogenous variables (explained in the model)
- *U* = exogenous variables (not explained in the model)



SCM provide a language for expressing counterfactuals

- $do(X)$ denotes intervention on X
- Intervention on X has the effect of removing all incoming links into X (or eliminating all direct causes of X)

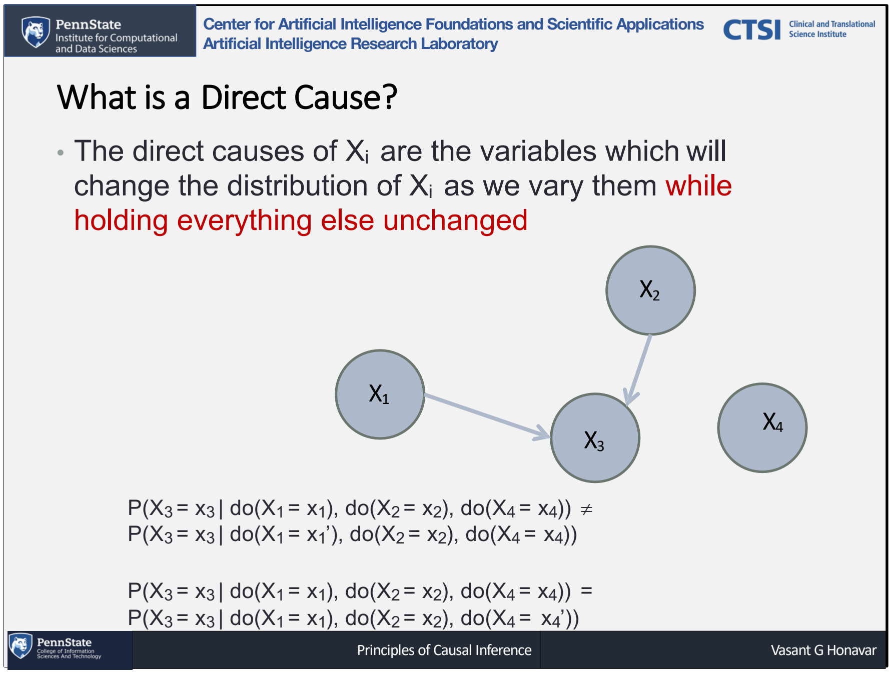


Connecting the SCM and Joint Probability Distribution

- Under some assumptions (Causal Markov Condition) an SCM represents a factorization of the joint probability distribution over the observables:

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i \mid DirectCauses(X_i))$$

- The above equation specifies the full joint probability distribution over the model variables.
- More on this later



What is a Direct Cause?

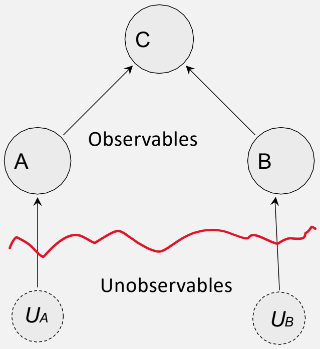
- The direct causes of X_i are the variables which will change the distribution of X_i as we vary them while holding everything else unchanged

$P(X_3 = x_3 \mid do(X_1 = x_1), do(X_2 = x_2), do(X_4 = x_4)) \neq$
$P(X_3 = x_3 \mid do(X_1 = x_1'), do(X_2 = x_2), do(X_4 = x_4))$

$P(X_3 = x_3 \mid do(X_1 = x_1), do(X_2 = x_2), do(X_4 = x_4)) =$
$P(X_3 = x_3 \mid do(X_1 = x_1), do(X_2 = x_2), do(X_4 = x_4'))$

Causal Graphs as Structural Equations

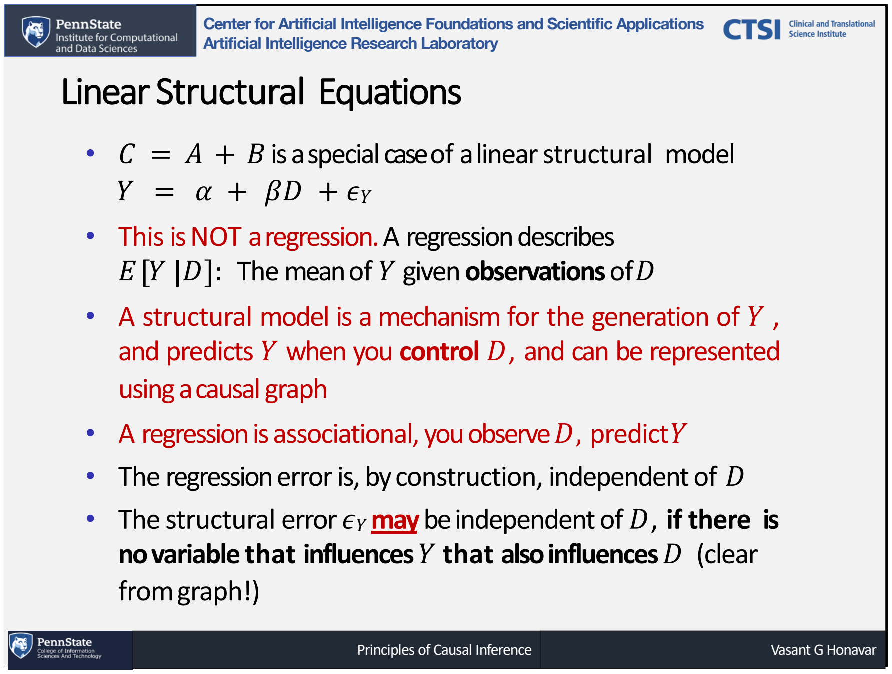
- $U_A \perp\!\!\!\perp U_B$
- $A = f_A(U_A)$
- $B = f_B(U_B)$
- $C = f_C(A, B)$, e.g., $C = A + B$
- A structural causal model describes
  - Our qualitative beliefs about nature assigns values to variables of interest in a domain of study
  - Causal assumptions



- Suppose $C = A + B$.
- Intuitively, what's the **causal effect** of A on C?
- $C^{a=1} - C^{a=0} = (1 + B) - (0 + B) = 1$

Linear Structural Models

- Structural causal model is NOT a regression.
- It is not an algebraic equation
- It is a causal model: LHS is caused by RHS!
  - Rearranging $C = A + B$ to obtain $A = C - B$ does not make sense!
- However,
  - You can use observed B and observed C to predict A (perfectly) $E[A\,|B, C] = C - B$
  - Even perfect regression fit does not tell you anything about causation!
- When we use an equation, we need to state whether it is structural causal model or a regression
- We will use
  - $Y = f_Y(\ldots)$ for structural models and
  - $E[Y\,|D] = f(D)$ for regressions



Linear Structural Equations

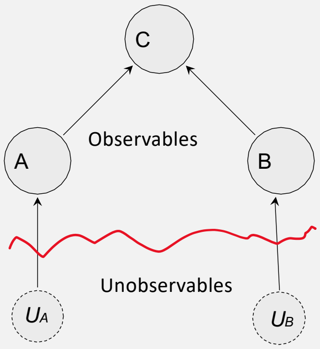
- $C = A + B$ is a special case of a linear structural model $Y = \alpha + \beta D + \epsilon_Y$
- This is NOT a regression. A regression describes $E[Y|D]$: The mean of Y given **observations** of D
- A structural model is a mechanism for the generation of Y, and predicts Y when you **control** D, and can be represented using a causal graph
- A regression is associational, you observe D, predict Y
- The regression error is, by construction, independent of D
- The structural error ϵ_Y **may** be independent of D, **if there is no variable that influences** Y **that also influences** D (clear from graph!)

Parametric versus Nonparametric Structural Equations

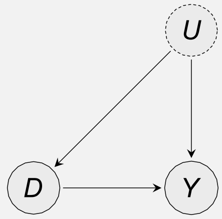
- In general, causal graphs are representations of
  - non-parametric structural equations (as general as possible)
- Causal graphs based on parametric e.g., *linear* structural equations are very restrictive, but easier to understand
- We will often use linear models to highlight basic ideas and then (try to) generalize them to nonparametric models

Causal Effects as Interventions

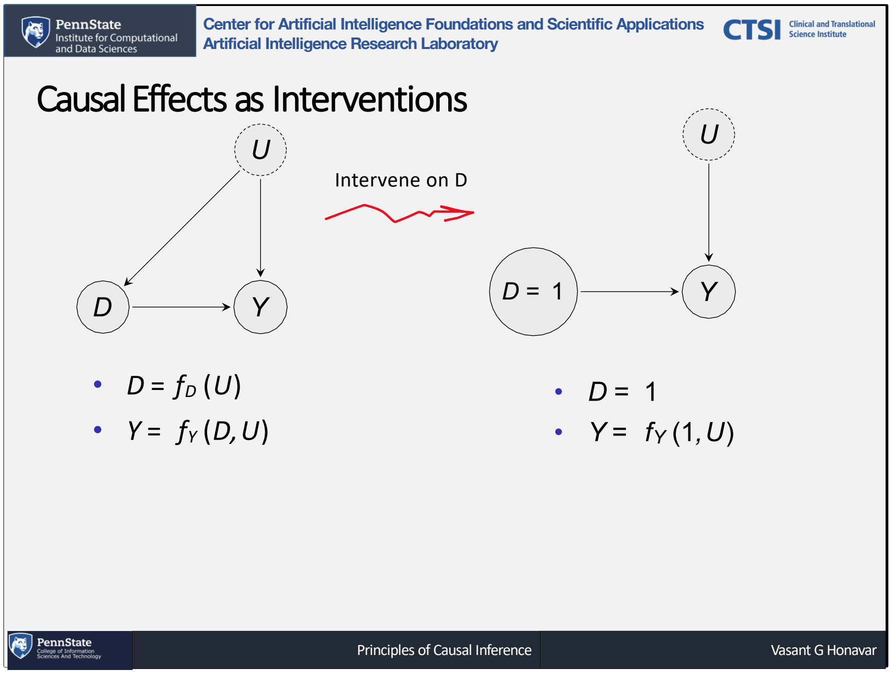
- Suppose $C = A + B$.
- How can we obtain the causal effect of A on C?
- By intervening on A independent of other variables and comparing $C^{a=1}$ with $C^{a=0}$
- $C^{a=1} - C^{a=0} = 1$



Causal Effects as Interventions



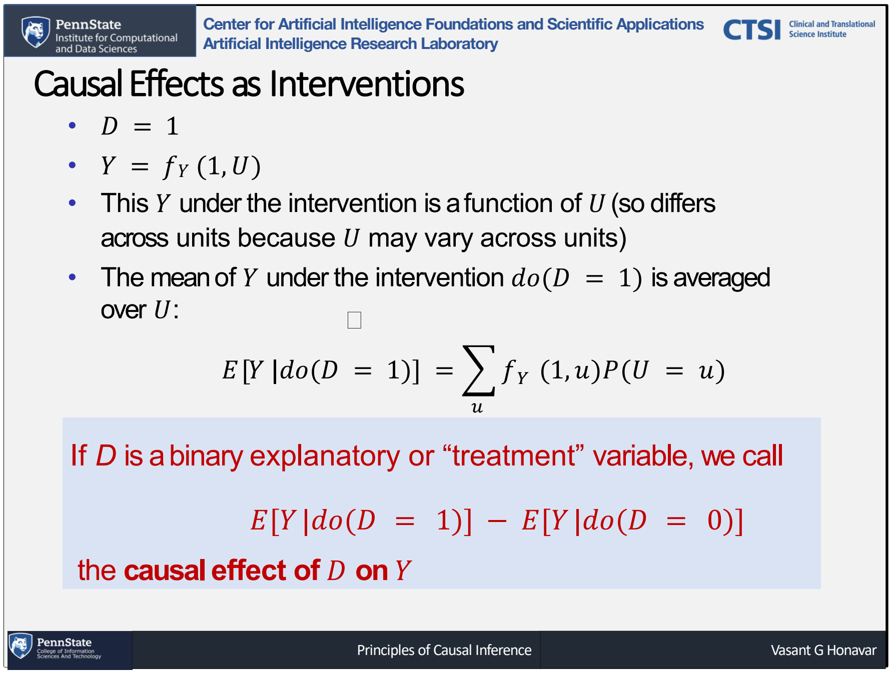
- What structural equation does the above model describe?
- $D = f_D(U)$
- $Y = f_Y(D, U)$
- What happens to the causal graph and the structural equations when we intervene on D i.e., "do" $D = 1$?



Causal Effects as Interventions

- $D = f_D(U)$
- $Y = f_Y(D, U)$

- $D = 1$
- $Y = f_Y(1, U)$



Causal Effects as Interventions

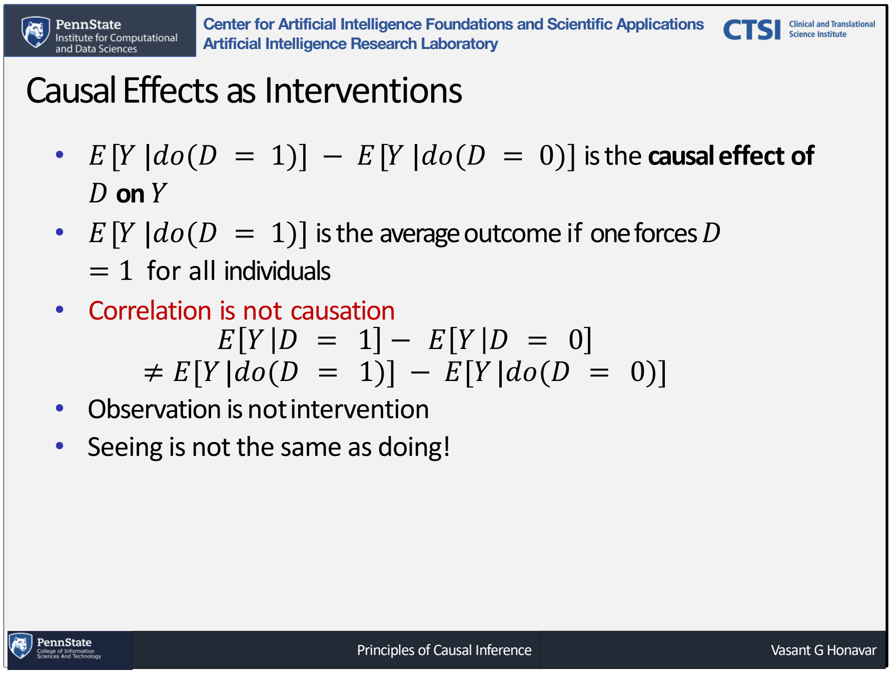
- $D = 1$
- $Y = f_Y(1, U)$
- This Y under the intervention is a function of U (so differs across units because U may vary across units)
- The mean of Y under the intervention $do(D = 1)$ is averaged over U:

$$E[Y\,|do(D = 1)] = \sum_{u} f_Y(1, u)P(U = u)$$

If *D* is a binary explanatory or "treatment" variable, we call

$$E[Y\,|do(D = 1)] - E[Y\,|do(D = 0)]$$

the **causal effect of** D **on** Y

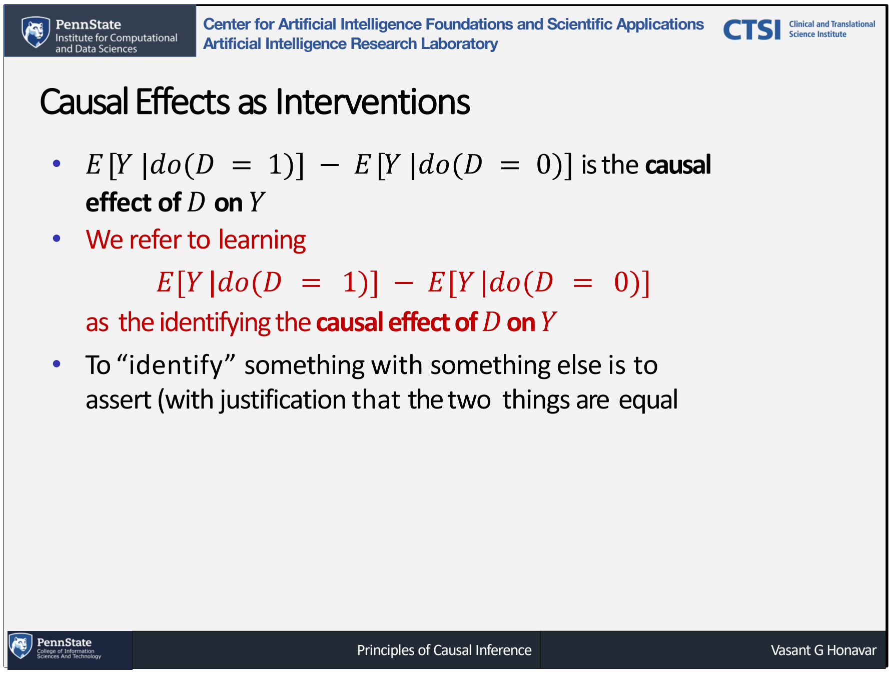


Causal Effects as Interventions

- $E[Y\,|do(D\ =\ 1)]\ -\ E[Y\,|do(D\ =\ 0)]$ is the **causal effect of** D **on** Y
- $E[Y\,|do(D\ =\ 1)]$ is the average outcome if one forces $D = 1$ for all individuals
- Correlation is not causation

$$E[Y\,|D\ =\ 1] - E[Y\,|D\ =\ 0]$$
$$\neq E[Y\,|do(D\ =\ 1)]\ -\ E[Y\,|do(D\ =\ 0)]$$

- Observation is not intervention
- Seeing is not the same as doing!

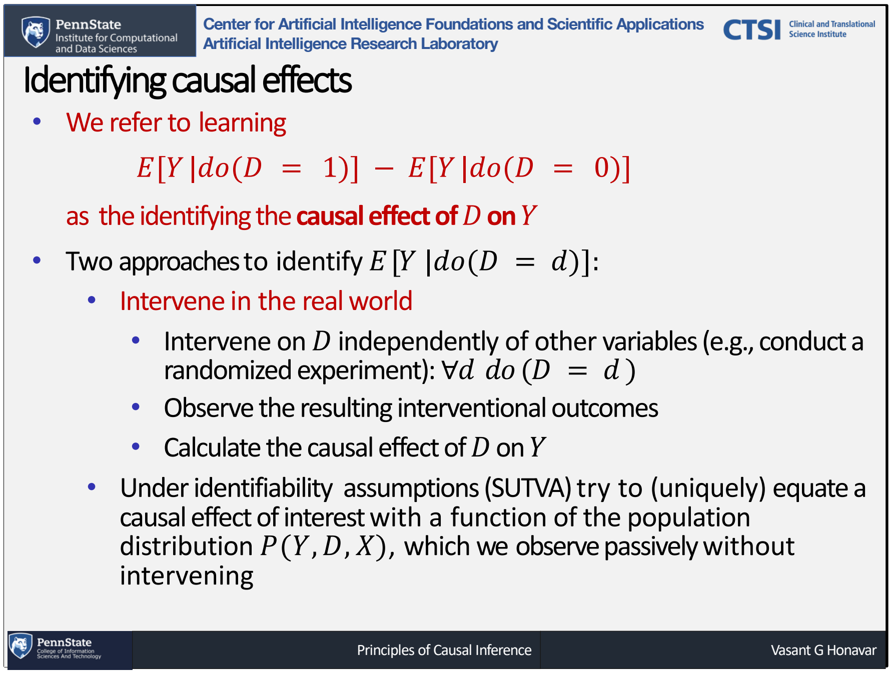


Causal Effects as Interventions

- $E[Y\,|do(D\;=\;1)]\;-\;E[Y\,|do(D\;=\;0)]$ is the **causal effect of D on Y**
- We refer to learning

  $$E[Y\,|do(D\;=\;1)]\;-\;E[Y\,|do(D\;=\;0)]$$

  as the identifying the **causal effect of D on Y**
- To "identify" something with something else is to assert (with justification that the two things are equal



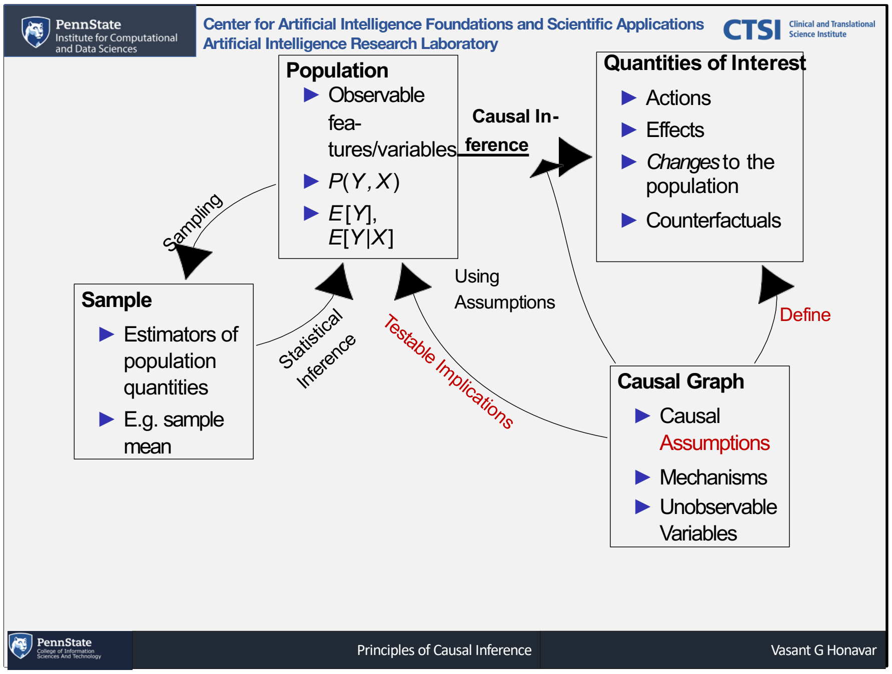
Identifying causal effects

- We refer to learning

  $$E[Y\,|do(D\ =\ 1)]\ -\ E[Y\,|do(D\ =\ 0)]$$

  as the identifying the **causal effect of D on Y**

- Two approaches to identify $E\,[Y\ |do(D\ =\ d)]$:
  - Intervene in the real world
    - Intervene on D independently of other variables (e.g., conduct a randomized experiment): $\forall d\ \ do\,(D\ =\ d\,)$
    - Observe the resulting interventional outcomes
    - Calculate the causal effect of D on Y
  - Under identifiability assumptions (SUTVA) try to (uniquely) equate a causal effect of interest with a function of the population distribution $P\,(Y\,,D\,,X)$, which we observe passively without intervening



PennState Institute for Computational and Data Sciences

Center for Artificial Intelligence Foundations and Scientific Applications
Artificial Intelligence Research Laboratory

CTSI Clinical and Translational Science Institute

Population

- Observable features/variables
- $P(Y, X)$
- $E[Y]$, $E[Y|X]$

Quantities of Interest

- Actions
- Effects
- *Changes* to the population
- Counterfactuals

Sample

- Estimators of population quantities
- E.g. sample mean

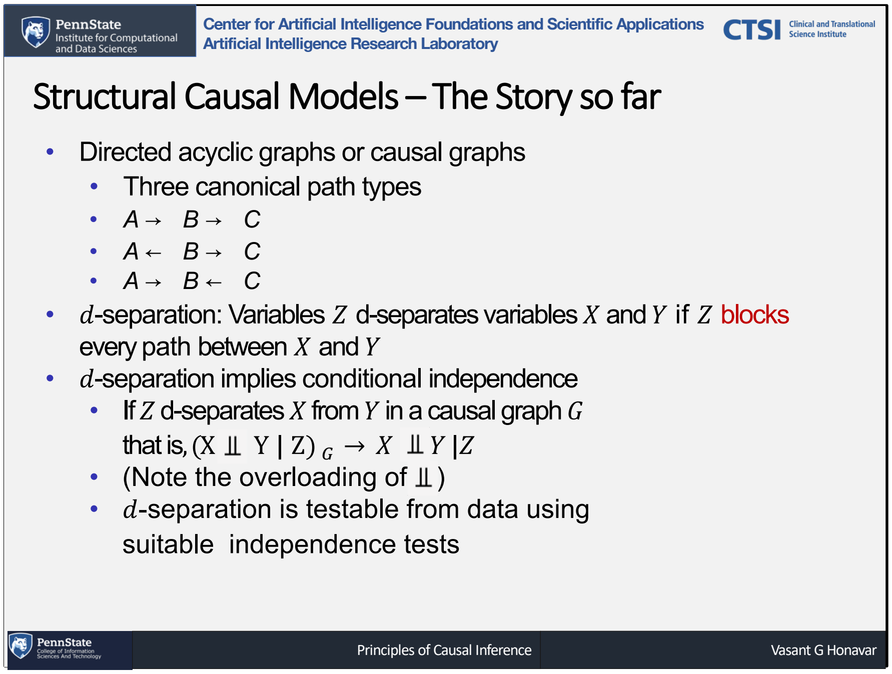
Causal Graph

- Causal Assumptions
- Mechanisms
- Unobservable Variables

PennState College of Information Sciences And Technology

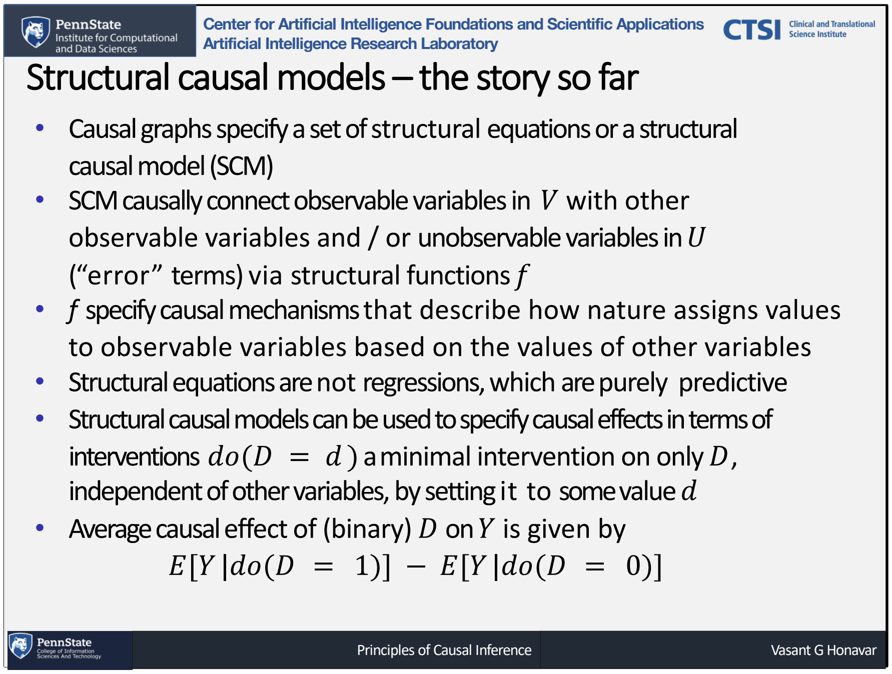
Principles of Causal Inference

Vasant G Honavar



Structural Causal Models – The Story so far

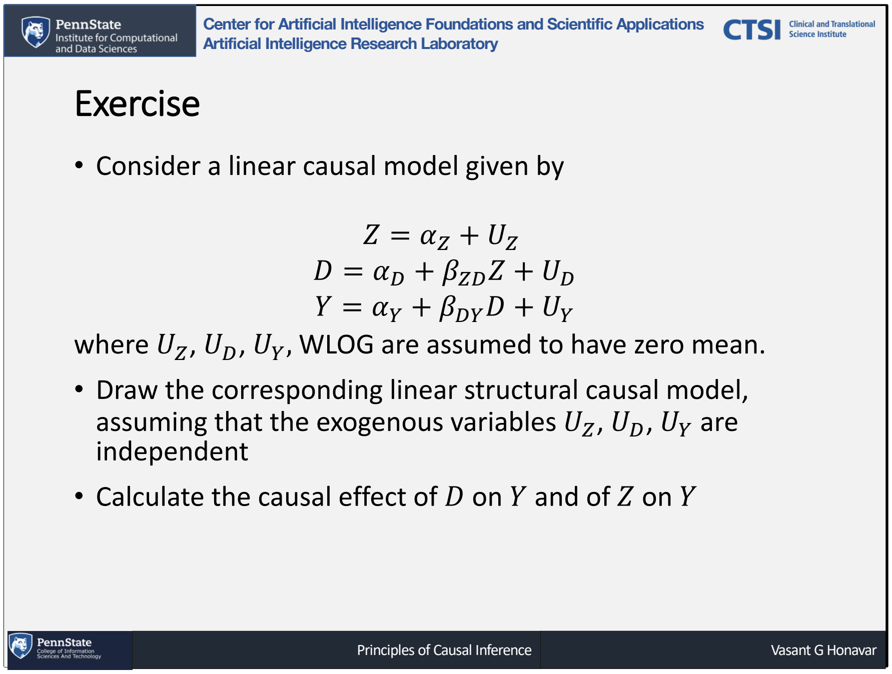
- Directed acyclic graphs or causal graphs
  - Three canonical path types
  - $A \rightarrow B \rightarrow C$
  - $A \leftarrow B \rightarrow C$
  - $A \rightarrow B \leftarrow C$
- d-separation: Variables Z d-separates variables X and Y if Z blocks every path between X and Y
- d-separation implies conditional independence
  - If Z d-separates X from Y in a causal graph G that is, $(X \perp\!\!\!\perp Y \mid Z)_G \rightarrow X \perp\!\!\!\perp Y \mid Z$
  - (Note the overloading of $\perp\!\!\!\perp$)
  - d-separation is testable from data using suitable independence tests



Structural causal models – the story so far

- Causal graphs specify a set of structural equations or a structural causal model (SCM)
- SCM causally connect observable variables in V with other observable variables and / or unobservable variables in U ("error" terms) via structural functions f
- f specify causal mechanisms that describe how nature assigns values to observable variables based on the values of other variables
- Structural equations are not regressions, which are purely predictive
- Structural causal models can be used to specify causal effects in terms of interventions $do(D = d)$ a minimal intervention on only D, independent of other variables, by setting it to some value d
- Average causal effect of (binary) D on Y is given by

$$E[Y\,|do(D = 1)] - E[Y\,|do(D = 0)]$$



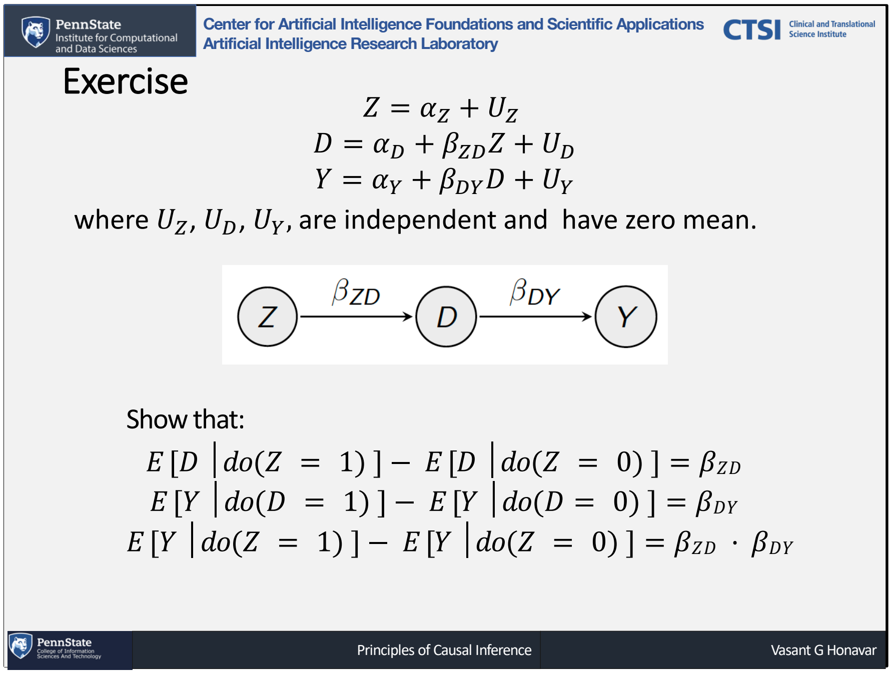
Exercise

- Consider a linear causal model given by

$$
\begin{gathered}
Z = \alpha_Z + U_Z \\
D = \alpha_D + \beta_{ZD} Z + U_D \\
Y = \alpha_Y + \beta_{DY} D + U_Y
\end{gathered}
$$

where U_Z, U_D, U_Y, WLOG are assumed to have zero mean.

- Draw the corresponding linear structural causal model, assuming that the exogenous variables U_Z, U_D, U_Y are independent
- Calculate the causal effect of D on Y and of Z on Y



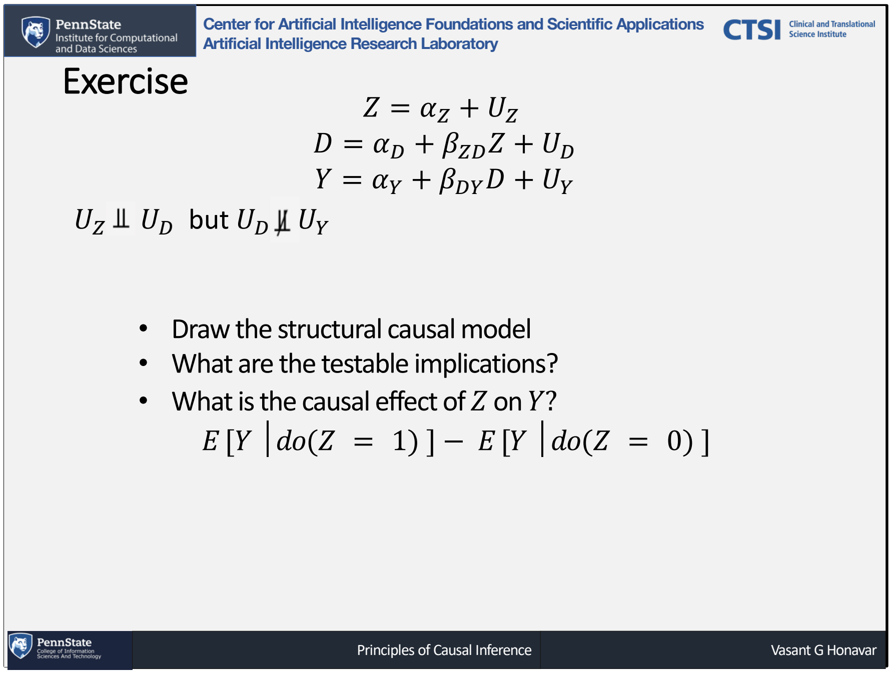
Exercise

$$Z = \alpha_Z + U_Z$$
$$D = \alpha_D + \beta_{ZD} Z + U_D$$
$$Y = \alpha_Y + \beta_{DY} D + U_Y$$

where U_Z, U_D, U_Y, are independent and have zero mean.

Show that:

$$E\,[D\;|do(Z\;=\;1)\,] - E\,[D\;|do(Z\;=\;0)\,] = \beta_{ZD}$$
$$E\,[Y\;|do(D\;=\;1)\,] - E\,[Y\;|do(D=\;0)\,] = \beta_{DY}$$
$$E\,[Y\;|do(Z\;=\;1)\,] - E\,[Y\;|do(Z\;=\;0)\,] = \beta_{ZD}\;\cdot\;\beta_{DY}$$



Exercise

$$Z = \alpha_Z + U_Z$$
$$D = \alpha_D + \beta_{ZD} Z + U_D$$
$$Y = \alpha_Y + \beta_{DY} D + U_Y$$

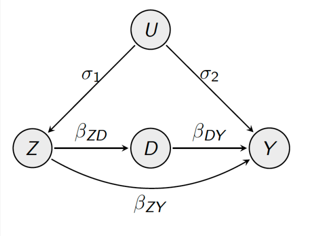
$U_Z \perp\!\!\!\perp U_D$ but $U_D \not\perp\!\!\!\perp U_Y$

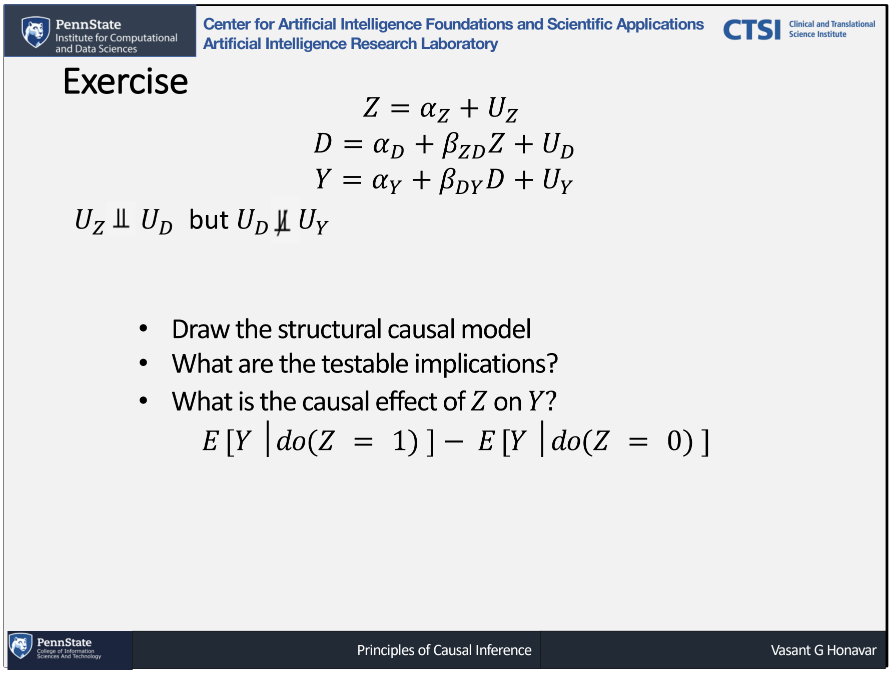
- Draw the structural causal model
- What are the testable implications?
- What is the causal effect of Z on Y?

$$E\,[Y \mid do(Z = 1)\,] - E\,[Y \mid do(Z = 0)\,]$$

Exercise

- Consider the structural causal model shown
- Write down the structural equations
- Calculate the causal effect of Z on Y





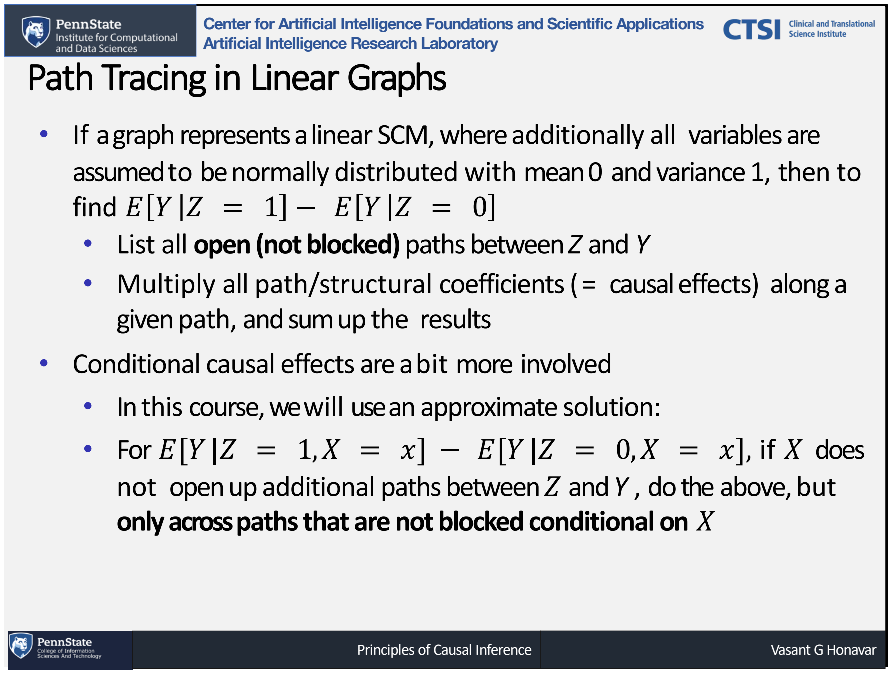
Exercise

$$Z = \alpha_Z + U_Z$$
$$D = \alpha_D + \beta_{ZD} Z + U_D$$
$$Y = \alpha_Y + \beta_{DY} D + U_Y$$

$U_Z \perp\!\!\!\perp U_D$ but $U_D \not\perp\!\!\!\perp U_Y$

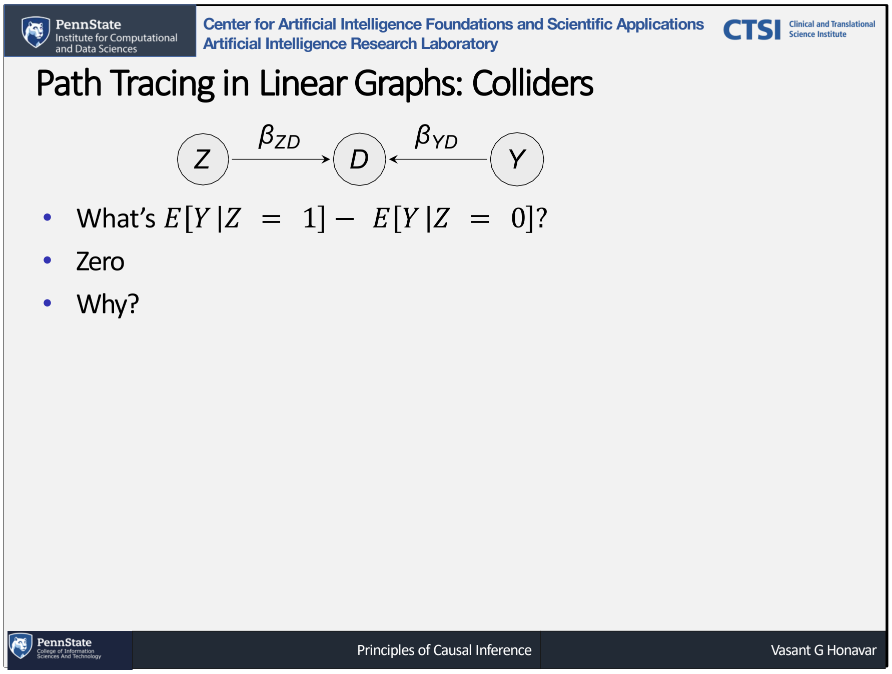
- Draw the structural causal model
- What are the testable implications?
- What is the causal effect of Z on Y?

$$E\left[Y \mid do(Z = 1)\right] - E\left[Y \mid do(Z = 0)\right]$$



Path Tracing in Linear Graphs

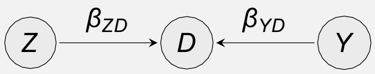
- If a graph represents a linear SCM, where additionally all variables are assumed to be normally distributed with mean 0 and variance 1, then to find $E[Y|Z = 1] - E[Y|Z = 0]$
  - List all **open (not blocked)** paths between Z and Y
  - Multiply all path/structural coefficients (= causal effects) along a given path, and sum up the results
- Conditional causal effects are a bit more involved
  - In this course, we will use an approximate solution:
  - For $E[Y|Z = 1, X = x] - E[Y|Z = 0, X = x]$, if X does not open up additional paths between Z and Y, do the above, but **only across paths that are not blocked conditional on X**



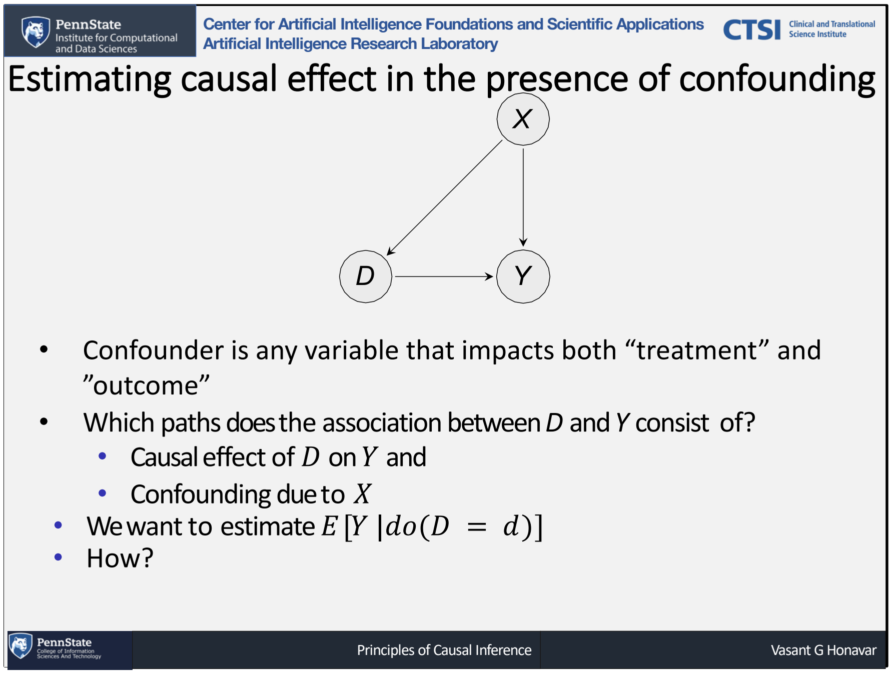
Path Tracing in Linear Graphs: Colliders

- What’s $E[Y \,|Z \;=\; 1] - \; E[Y \,|Z \;=\; 0]$?
- Zero
- Why?

Path Tracing in Linear Graphs: Colliders

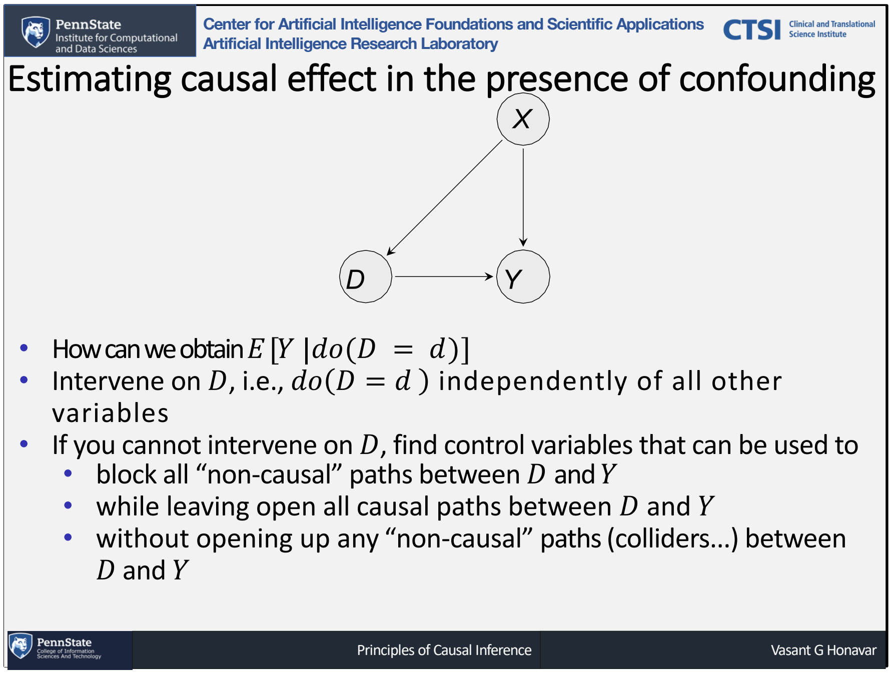


- $E[Y|Z = 1] - E[Y|Z = 0] = 0$
- Let's say β_{ZD} and β_{DY} are positive.
- Is $E[Y|Z = 1, D = 1] - E[Y|Z = 0, D = 1]$ positive?
  - Look at units with same $D = 1$, but different Z.
  - If you have $Z = 0$ but still $D = 1$, that must be because Y makes up for lack of Z
- So mean difference is **negative**
- $E[Y|Z = 1, D = 1] - E[Y|Z = 0, D = 1] \approx -\beta_{ZD} \cdot \beta_{DY}$
- We will return to this. See section 3.8, Pearl, Glymour, Jewell (2016) and Pearl (2013): "Linear Models: A Useful Microscope for Causal Analysis"



Estimating causal effect in the presence of confounding

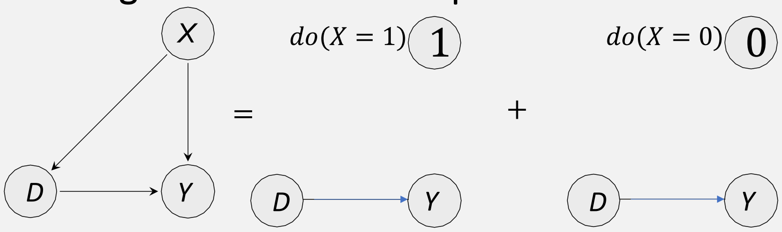
- Confounder is any variable that impacts both “treatment” and ”outcome”
- Which paths does the association between *D* and *Y* consist of?
  - Causal effect of D on Y and
  - Confounding due to X
- We want to estimate $E\,[Y\;|do(D\;=\;d)]$
- How?



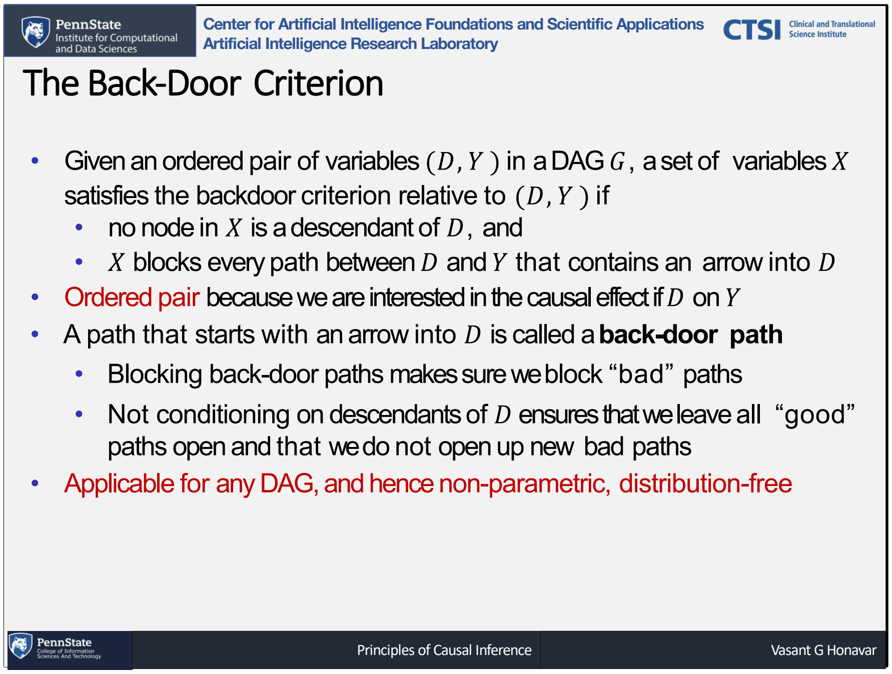
Estimating causal effect in the presence of confounding

- How can we obtain $E\,[Y\ |do(D\ =\ d)]$
- Intervene on D, i.e., $do(D = d\ )$ independently of all other variables
- If you cannot intervene on D, find control variables that can be used to
  - block all “non-causal” paths between D and Y
  - while leaving open all causal paths between D and Y
  - without opening up any “non-causal” paths (colliders…) between D and Y

Estimating causal effect in the presence of confounding

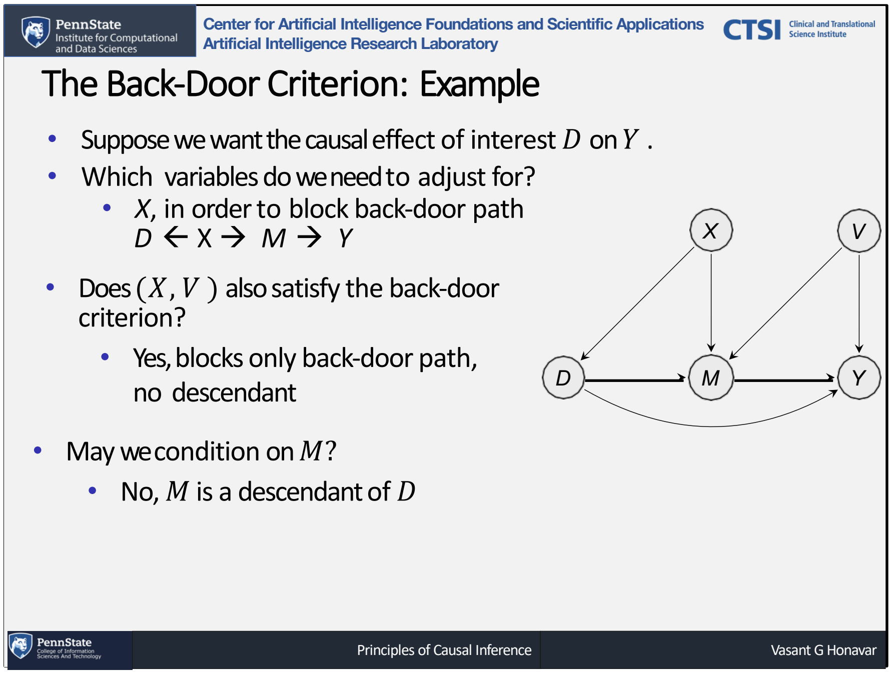


- Suppose we control for X
- We block the non-causal paths between D and Y without eliminating the causal paths or introducing any non-causal paths
- Now we can estimate the causal effect of D on Y separately from observational data with $X = 0$ and with $X = 1$ and take a weighted average of the two effects where the weights correspond to $P(X = 1)$ and $P(X = 0)$



The Back-Door Criterion

- Given an ordered pair of variables (D, Y) in a DAG G, a set of variables X satisfies the backdoor criterion relative to (D, Y) if
  - no node in X is a descendant of D, and
  - X blocks every path between D and Y that contains an arrow into D
- Ordered pair because we are interested in the causal effect if D on Y
- A path that starts with an arrow into D is called a **back-door path**
  - Blocking back-door paths makes sure we block "bad" paths
  - Not conditioning on descendants of D ensures that we leave all "good" paths open and that we do not open up new bad paths
- Applicable for any DAG, and hence non-parametric, distribution-free

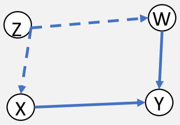


The Back-Door Criterion: Example

- Suppose we want the causal effect of interest D on Y .
- Which variables do we need to adjust for?
  - X, in order to block back-door path $D \leftarrow X \rightarrow M \rightarrow Y$
- Does (X, V) also satisfy the back-door criterion?
  - Yes, blocks only back-door path, no descendant
- May we condition on M?
  - No, M is a descendant of D

Backdoor criterion: Example

- Suppose we want to estimate the causal effect of a drug X on recovery Y

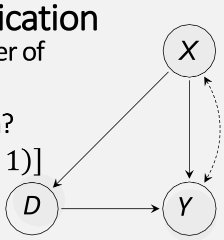


- X, Y, W are observed, Z is not.
- Is there any unobserved confounder?
- Yes, Z is an unobserved confounder
- How do we de-confound the causal effect X on Y?
- Look for an observed variable that satisfies the backdoor criterion
- W is such a variable – it blocks the backdoor path X ⍰ Z → W → Y, W is not a descendent of X
- Upon adjusting for W we have

$$P(Y = y|do(X = x)) = \sum_{w} P(Y = y|X = x, W = w)P(w)$$

- Hence, the causal effect of X on Y is identifiable from observational data

Using the Back-Door Criterion for Identification

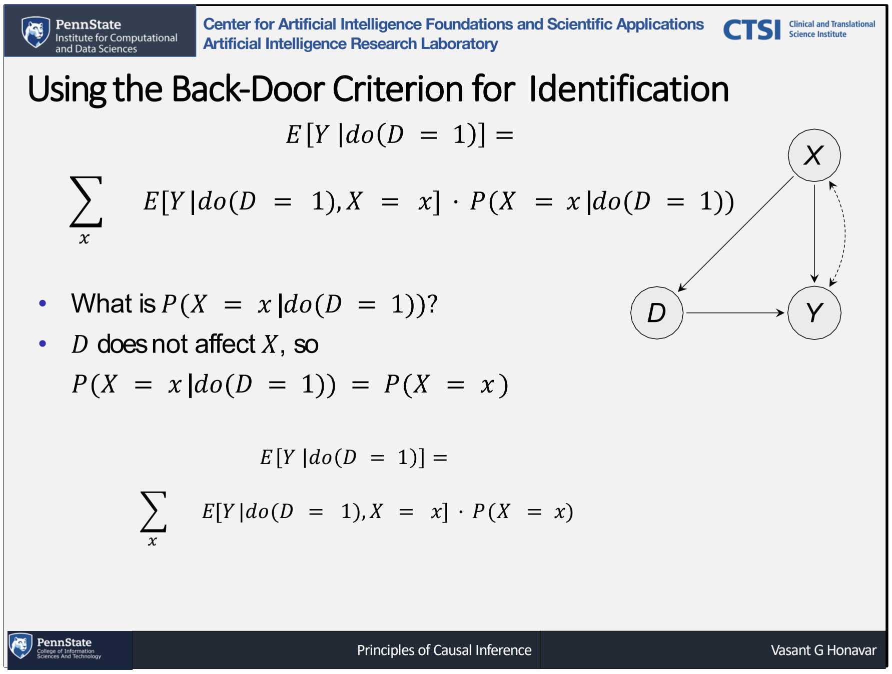


- Bi-directed arc is additional unobserved confounder of X and Y and hence D and Y
- Does X fulfill the BDC wrt D and Y in this graph?
- We want $E[Y \mid do(D = 1)] - E[Y \mid do(D = 1)]$
- We measure $P(Y, D, X)$.
- We somehow need to condition on X
- Question: How can we re-express $E[Y \mid do(D = 1)]$ as something that is conditional on X without making additional assumptions?

- Law of Iterated Expectations!

$E[Y \mid do(D = 1)] =$

$$\sum_x E[Y \mid do(D = 1), X = x]\, P(X = x \mid do(D = 1))$$



Using the Back-Door Criterion for Identification

$$E[Y\,|do(D\ =\ 1)]\ =$$

$$\sum_{x} E[Y\,|do(D\ =\ 1), X\ =\ x]\cdot P(X\ =\ x\,|do(D\ =\ 1))$$

- What is $P(X\ =\ x\,|do(D\ =\ 1))$?
- D does not affect X, so $P(X\ =\ x\,|do(D\ =\ 1))\ =\ P(X\ =\ x)$

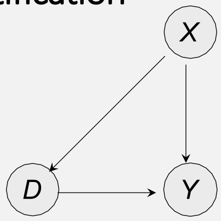
$$E[Y\,|do(D\ =\ 1)]\ =$$

$$\sum_{x} E[Y\,|do(D\ =\ 1), X\ =\ x]\cdot P(X\ =\ x)$$

Using the Back-Door Criterion for Identification

$$E[Y\,|do(D\ =\ 1)] =$$

$$\sum_x E[Y\,|do(D\ =\ 1), X\ =\ x]\cdot P(X\ =\ x)$$



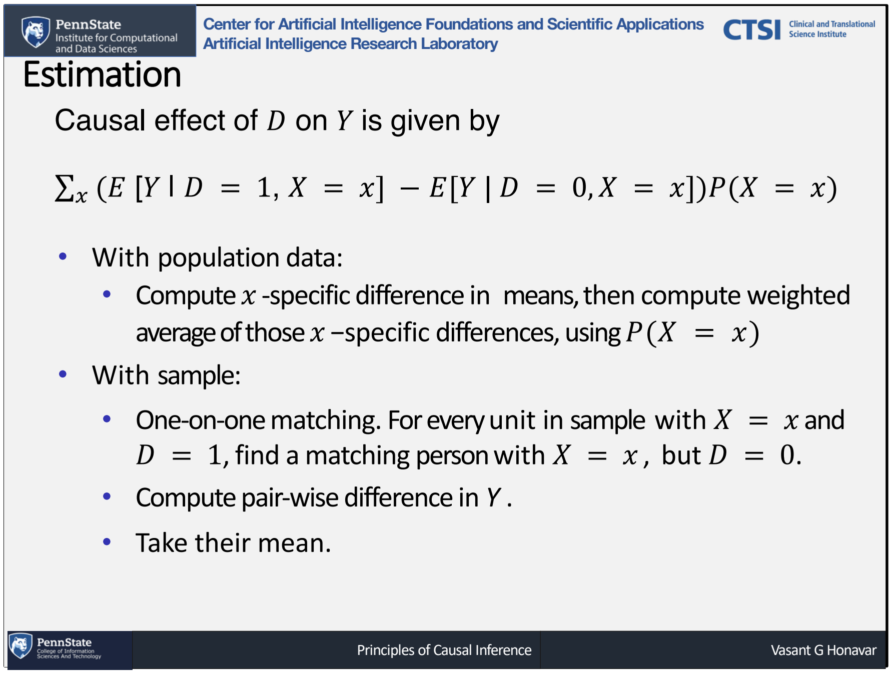
- How to get rid of the other $do(D\ =\ 1)$?
- Conditional on X , observing $D\ =\ 1$ is the same as **do**ing $D\ =\ 1$, at least with respect to Y

$$E[Y\,|do(D\ =\ 1)] =$$

$$\sum_x E[Y\,|D = 1, X\ =\ x]\cdot P(X\ =\ x)$$

Hence the causal effect of D on Y is given by

$$\sum_x (E\,[Y \mid D\ =\ 1, X\ =\ x] - E[Y \mid D\ =\ 0, X\ =\ x])P(X\ =\ x)$$

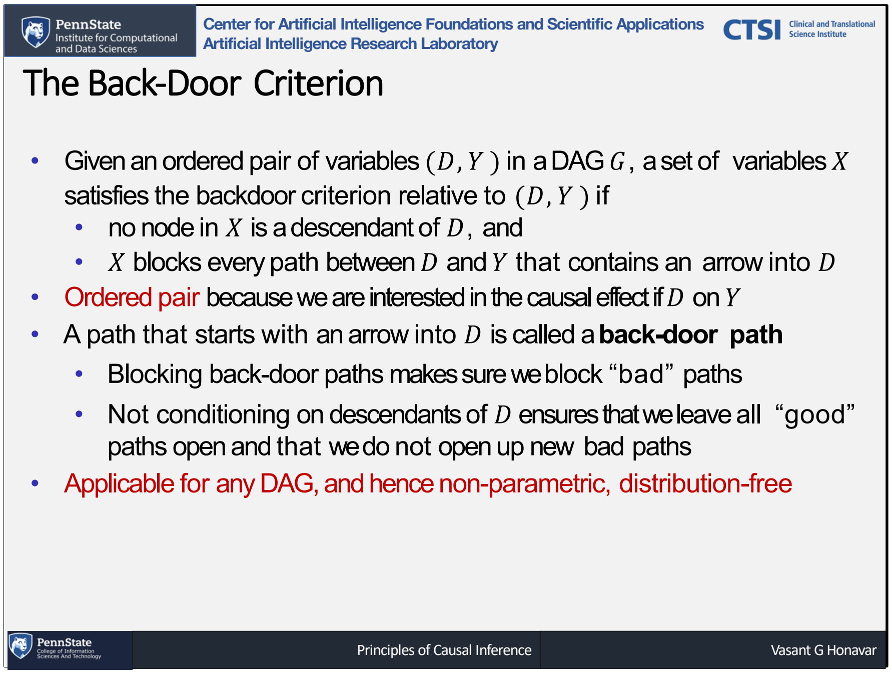


Estimation

Causal effect of D on Y is given by

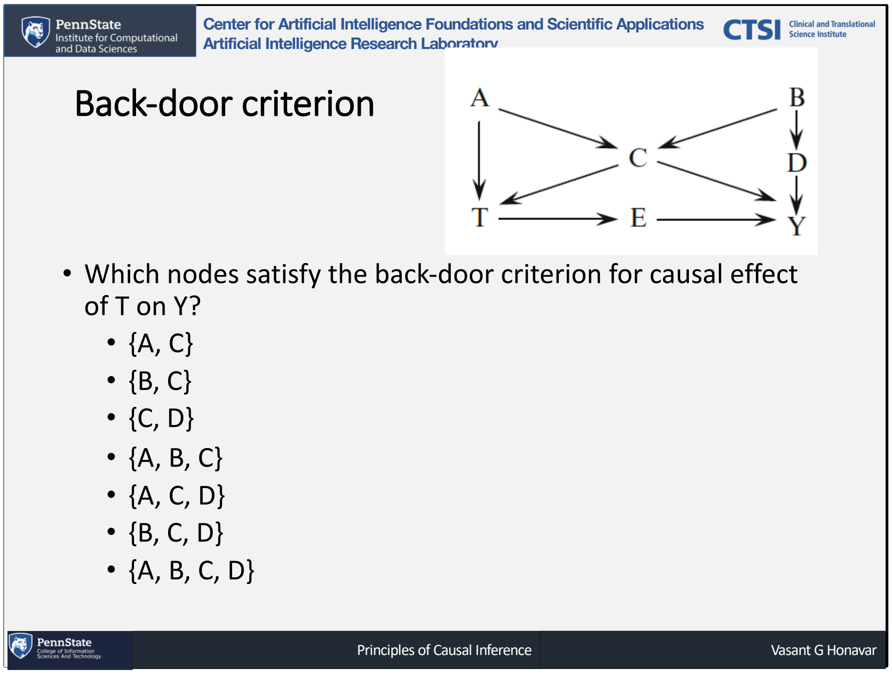
$$\sum_x (E\,[Y \mid D = 1, X = x] - E[Y \mid D = 0, X = x])P(X = x)$$

- With population data:
  - Compute x -specific difference in means, then compute weighted average of those x −specific differences, using $P(X = x)$
- With sample:
  - One-on-one matching. For every unit in sample with $X = x$ and $D = 1$, find a matching person with $X = x$, but $D = 0$.
  - Compute pair-wise difference in Y.
  - Take their mean.



The Back-Door Criterion

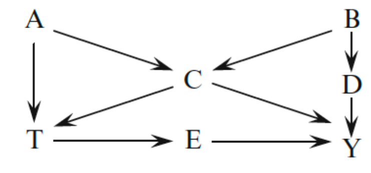
- Given an ordered pair of variables (D, Y) in a DAG G, a set of variables X satisfies the backdoor criterion relative to (D, Y) if
  - no node in X is a descendant of D, and
  - X blocks every path between D and Y that contains an arrow into D
- Ordered pair because we are interested in the causal effect if D on Y
- A path that starts with an arrow into D is called a **back-door path**
  - Blocking back-door paths makes sure we block “bad” paths
  - Not conditioning on descendants of D ensures that we leave all “good” paths open and that we do not open up new bad paths
- Applicable for any DAG, and hence non-parametric, distribution-free



Back-door criterion

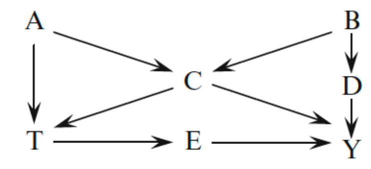
- Which nodes satisfy the back-door criterion for causal effect of T on Y?
  - {A, C}
  - {B, C}
  - {C, D}
  - {A, B, C}
  - {A, C, D}
  - {B, C, D}
  - {A, B, C, D}

Back-door criterion Dissected

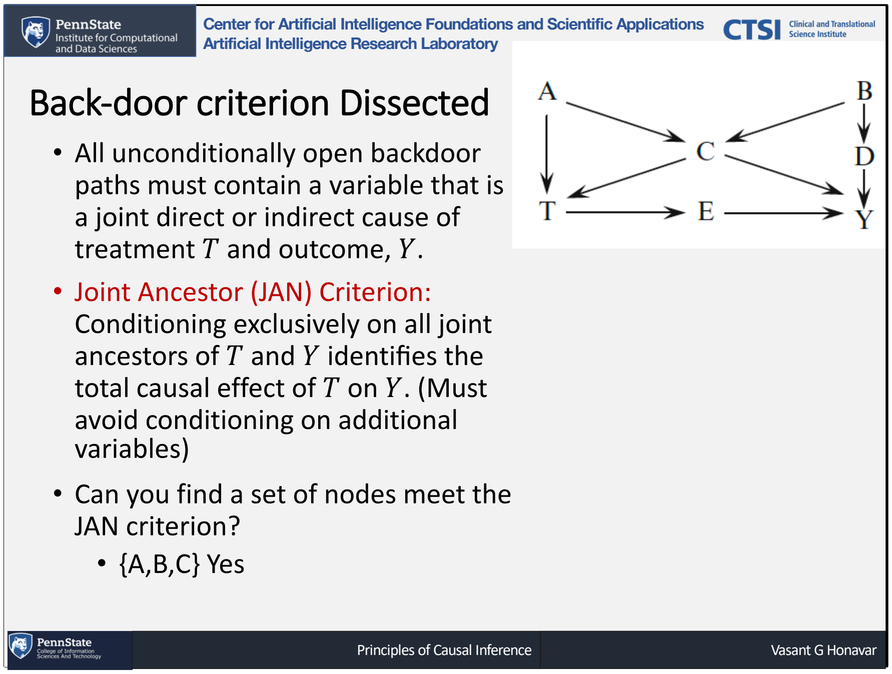


- All backdoor paths start with an arrow into the treatment variable
- This implies they can be blocked by conditioning on parents of the treatment, if they are observed
- Parents of treatment (POT) criterion: Any causal effect of T on Y can be identified by conditioning on all of the parents of T if they are observed.
- Sets of nodes meeting POT criterion:
  - {A, C}

Back-door criterion Dissected



- Parents of the outcome (POO) criterion: If no backdoor path shares a node with any causal path (other than T and Y), then conditioning on all parents of the outcome Y (if they are observed) that do not lie on a causal path from T to Y identifies the total causal effect of T on Y.
- Do these sets of nodes meet the POO criterion?
  - {C,D} Yes
  - {C, D, E} No. (Why?)

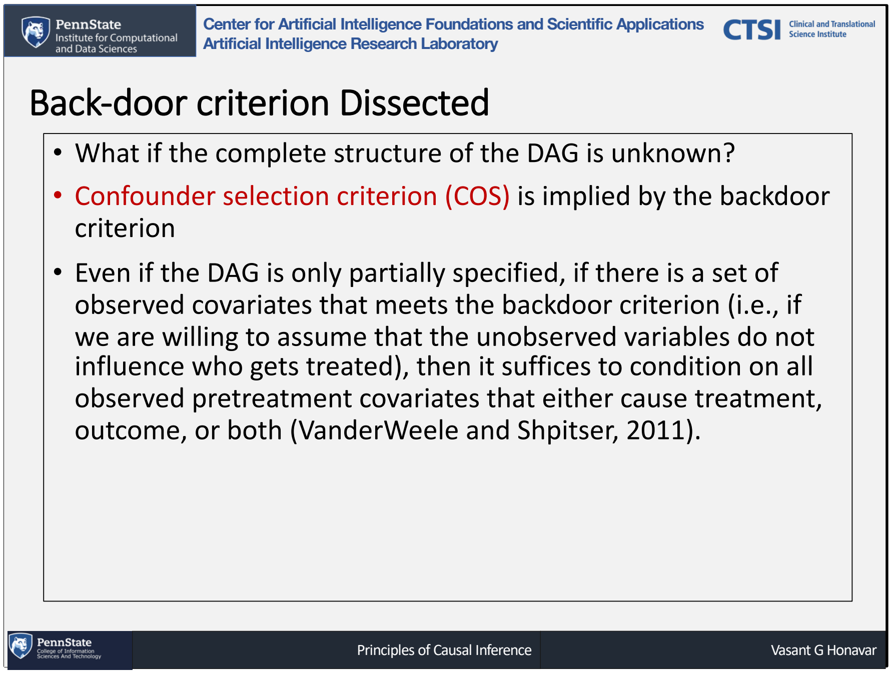


Back-door criterion Dissected

- All unconditionally open backdoor paths must contain a variable that is a joint direct or indirect cause of treatment T and outcome, Y.
- Joint Ancestor (JAN) Criterion: Conditioning exclusively on all joint ancestors of T and Y identifies the total causal effect of T on Y. (Must avoid conditioning on additional variables)
- Can you find a set of nodes meet the JAN criterion?
  - {A,B,C} Yes

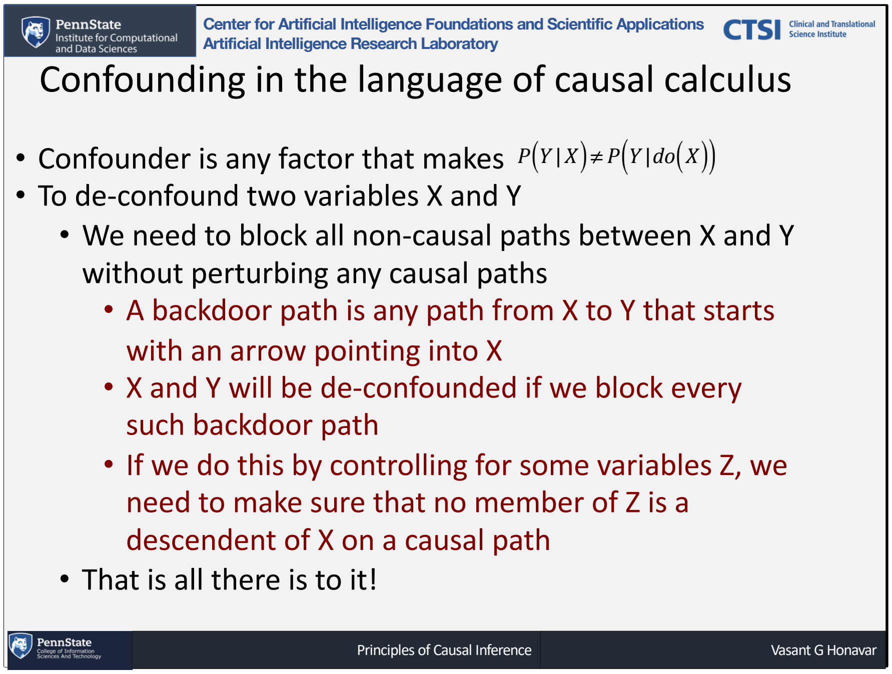
Back-door criterion Dissected

- Identifiability using the backdoor criteria or any of its special cases (POT, POO, JAN) assume that there exist at least one set of covariates that satisfy the criteria are indeed observed.
- Backdoor criterion implies ignorability (of Rubin's potential outcomes framework) and identifies a set of covariates that when controlled and adjusted for, ensure ignorability when the causal graph is correctly specified (meaning all of the relevant variables are included, and no edges that should be present are omitted).



Back-door criterion Dissected

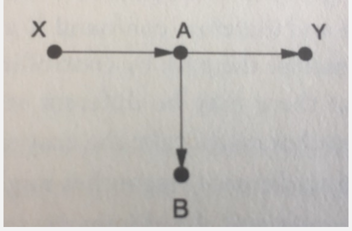
- What if the complete structure of the DAG is unknown?
- Confounder selection criterion (COS) is implied by the backdoor criterion
- Even if the DAG is only partially specified, if there is a set of observed covariates that meets the backdoor criterion (i.e., if we are willing to assume that the unobserved variables do not influence who gets treated), then it suffices to condition on all observed pretreatment covariates that either cause treatment, outcome, or both (VanderWeele and Shpitser, 2011).



Confounding in the language of causal calculus

- Confounder is any factor that makes $P(Y|X) \neq P(Y|do(X))$
- To de-confound two variables X and Y
  - We need to block all non-causal paths between X and Y without perturbing any causal paths
    - A backdoor path is any path from X to Y that starts with an arrow pointing into X
    - X and Y will be de-confounded if we block every such backdoor path
    - If we do this by controlling for some variables Z, we need to make sure that no member of Z is a descendent of X on a causal path
  - That is all there is to it!

Confounding through the lens of causal calculus



Source: The book of Why, Pearl and Mackenzie

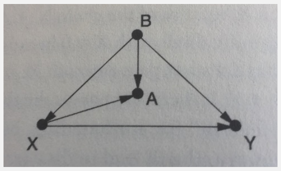
What do we need to control for in order to de-confound X and Y?

- Nothing!
  - There is no backdoor path into X
  - A, B are descendants of X (and hence should not be controlled for)

However,

- B passes a classical epidemiological definition of confounding
- But if we control for B, we introduce confounding rather than eliminating it!

Confounding through the lens of causal calculus

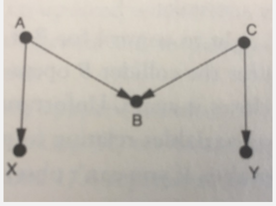


Source: The book of Why, Pearl and Mackenzie

What do we need to control for in order to de-confound X and Y?

- There is a backdoor path X ←B →Y
- We can block it only by blocking B
- If B is observable, we are all set
- If B is unobservable
  - We cannot control for it, so there is no way we can de-confound X and Y, so there is no way to estimate the causal effect of X on Y without running a RCT
  - Current statistical practice would advocate controlling for A, a proxy of B – but this only partially eliminates the confounding bias and introduces a collider biasl

Confounding through the lens of causal calculus

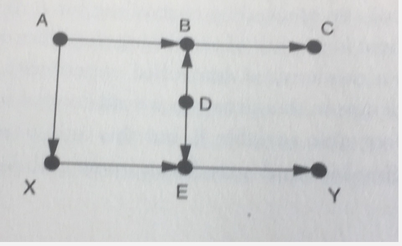


Source: The book of Why, Pearl and Mackenzie

What do we need to control for in order to de-confound X and Y?

- There is a backdoor path X ←A →B ←C → Y which is already blocked by B
- Some of the correlation based statistical definitions of confounding would identify B as a confounder!
- B becomes a confounder when we control for it!
- Example
  - B – Seatbelt use, X – Smoking, A – Attitude towards societal norms, C – Attitude towards safety and health related measures, Y – lung cancer
  - A 2006 study found B to be correlated with both X and Y

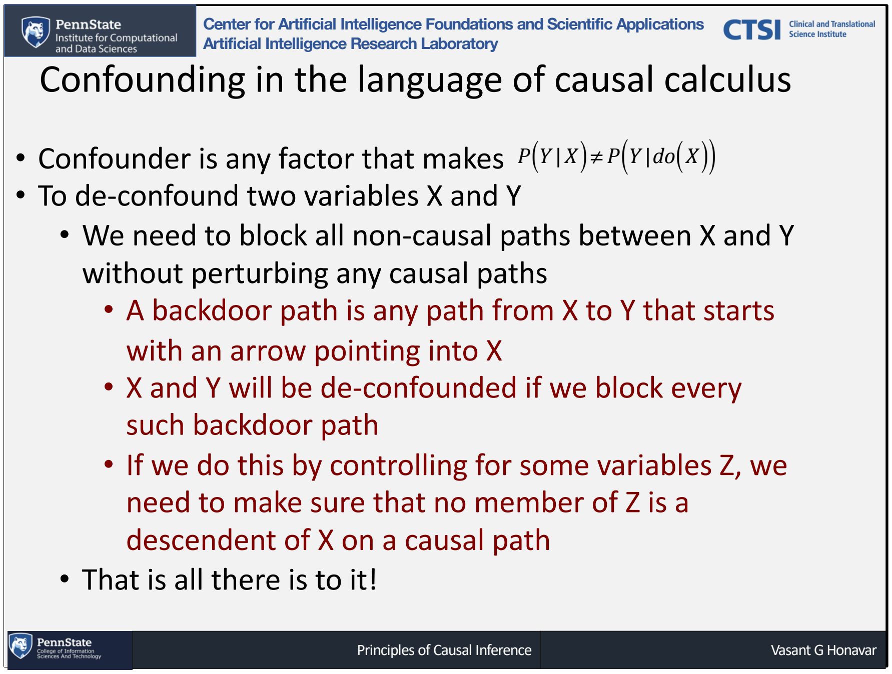
Confounding through the lens of causal calculus



Source: The book of Why, Pearl and Mackenzie

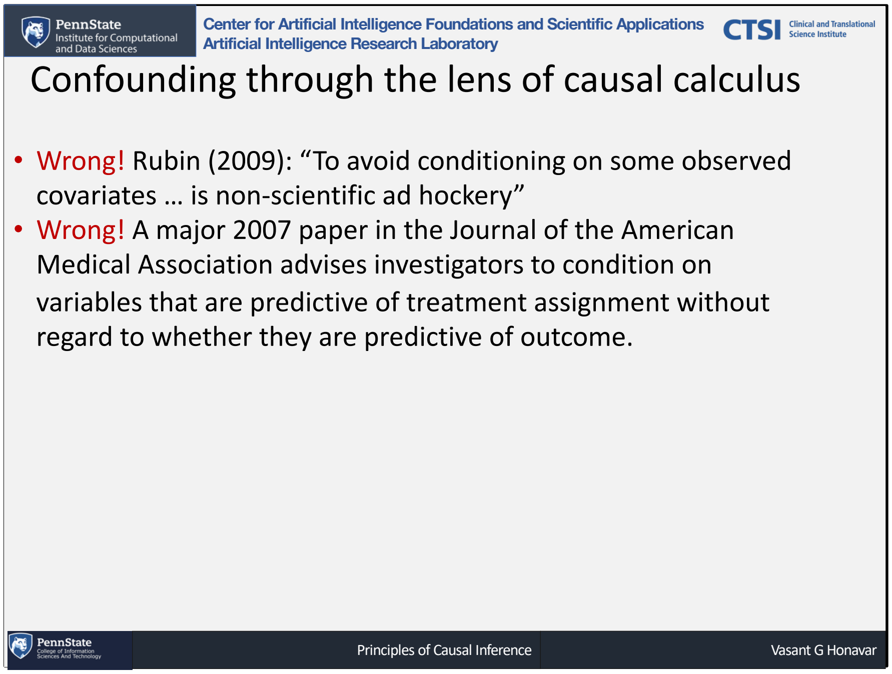
What do we need to control for in order to de-confound X and Y?

- A, B, C, D are pre-treatment variables, X is the treatment
- The only backdoor path X ←A → B ←D→ E→Y is already blocked by the collider B, so no need to control for anything!
- Standard statistical practice would be to control for B and C
- Reinforced by Rubin (2009): "To avoid conditioning on some observed covariates … is non-scientific ad hockery"
- Controlling for B and C introduces confounding (unless we control for A or D as well)



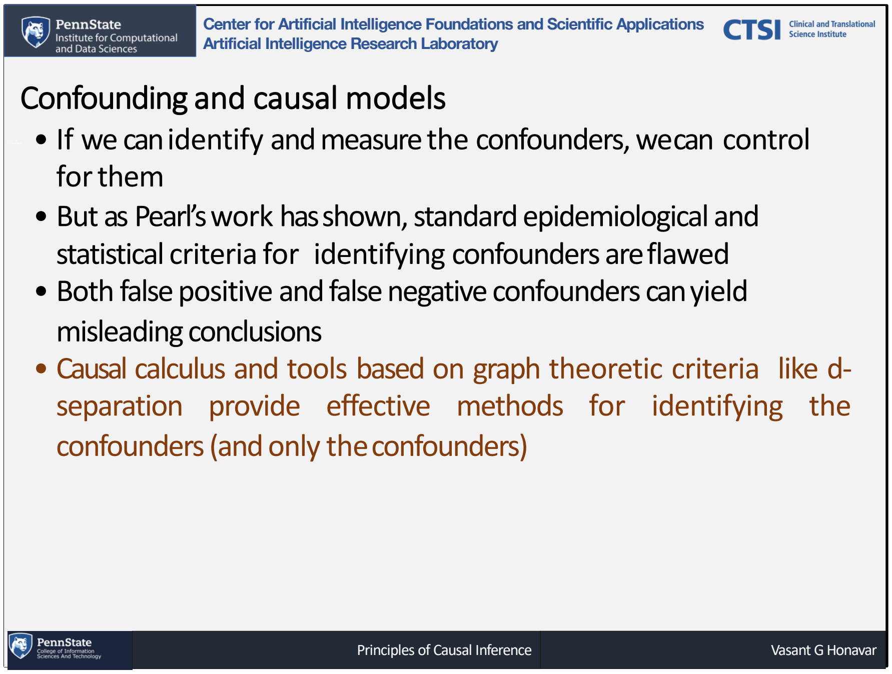
Confounding in the language of causal calculus

- Confounder is any factor that makes $P(Y|X) \neq P(Y|do(X))$
- To de-confound two variables X and Y
  - We need to block all non-causal paths between X and Y without perturbing any causal paths
    - A backdoor path is any path from X to Y that starts with an arrow pointing into X
    - X and Y will be de-confounded if we block every such backdoor path
    - If we do this by controlling for some variables Z, we need to make sure that no member of Z is a descendent of X on a causal path
  - That is all there is to it!



Confounding through the lens of causal calculus

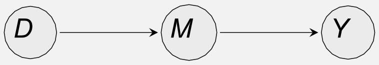
- Wrong! Rubin (2009): “To avoid conditioning on some observed covariates … is non-scientific ad hockery”
- Wrong! A major 2007 paper in the Journal of the American Medical Association advises investigators to condition on variables that are predictive of treatment assignment without regard to whether they are predictive of outcome.



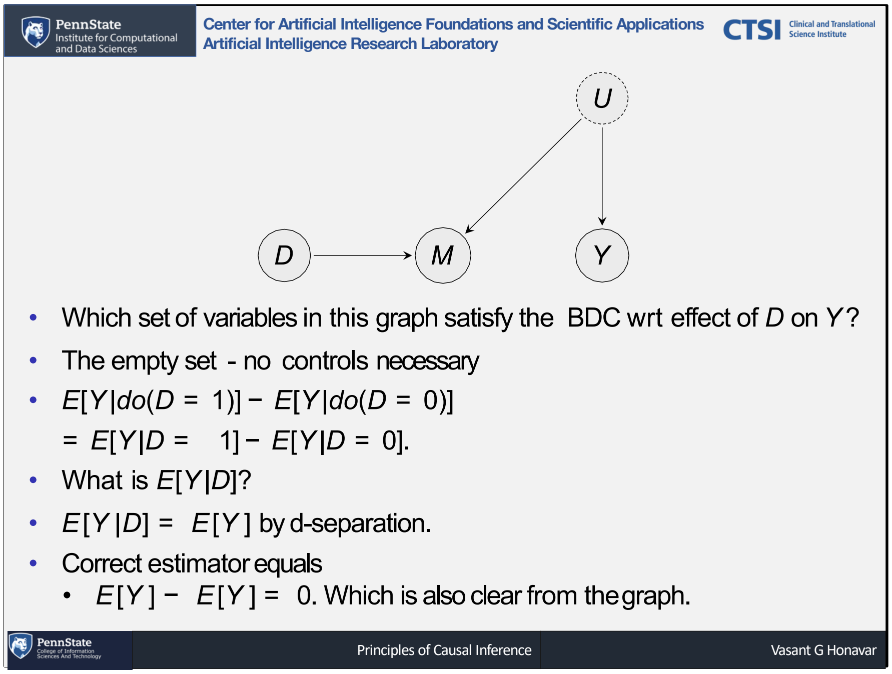
Confounding and causal models

- If we can identify and measure the confounders, wecan control for them
- But as Pearl's work has shown, standard epidemiological and statistical criteria for identifying confounders are flawed
- Both false positive and false negative confounders can yield misleading conclusions
- Causal calculus and tools based on graph theoretic criteria like d-separation provide effective methods for identifying the confounders (and only the confounders)

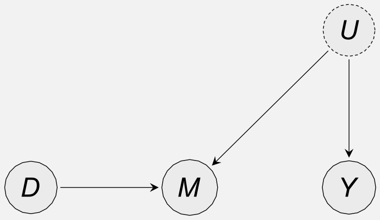
Review exercises – Back door criterion



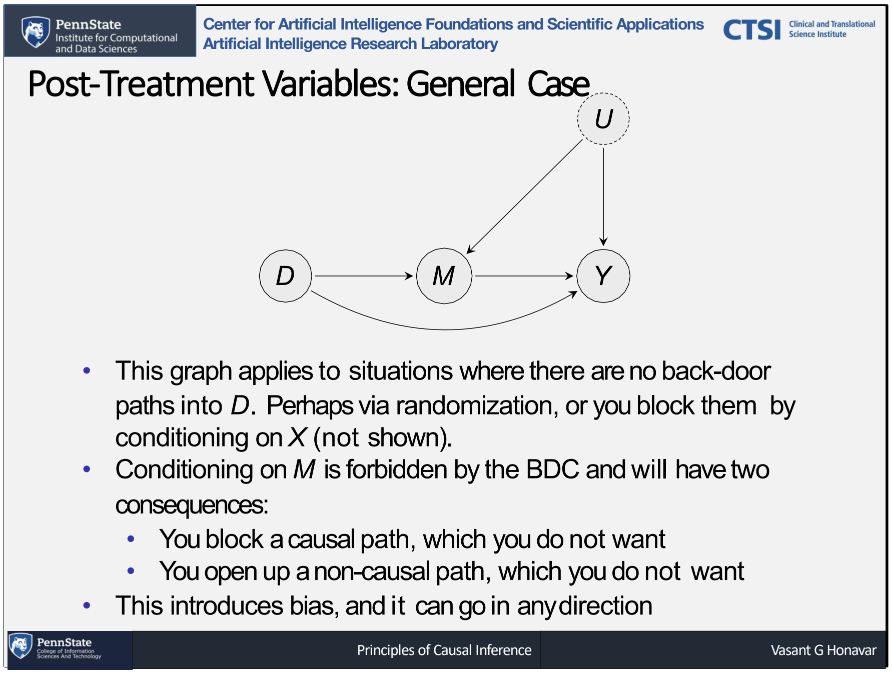
- Which set of variables in this graph satisfy the BDC with respect to the causal effect of D on Y?
- The empty set!
- $E[Y \mid do(D = 1)] - E[Y \mid do(D = 0) = E[Y \mid D = 1] - E[Y \mid D = 0]$ (correlation is causation)
  - No paths into D - just like we intervened on it
  - But you may have learned in statistics that…
- “M correlates with D and Y, so you need to control for it. Otherwise, you have omitted-variable bias”
- Bad idea: Conditional on M, D and Y are *d*-separated!
- Montgomery et al. 2018 AJPS estimate that 50% of political science studies suffer from this problem (of controlling for post-treatment variables)



- Which set of variables in this graph satisfy the BDC wrt effect of D on Y?
- The empty set - no controls necessary
- $E[Y|do(D = 1)] - E[Y|do(D = 0)]$
  $= E[Y|D = 1] - E[Y|D = 0]$.
- What is $E[Y|D]$?
- $E[Y|D] = E[Y]$ by d-separation.
- Correct estimator equals
  - $E[Y] - E[Y] = 0$. Which is also clear from the graph.

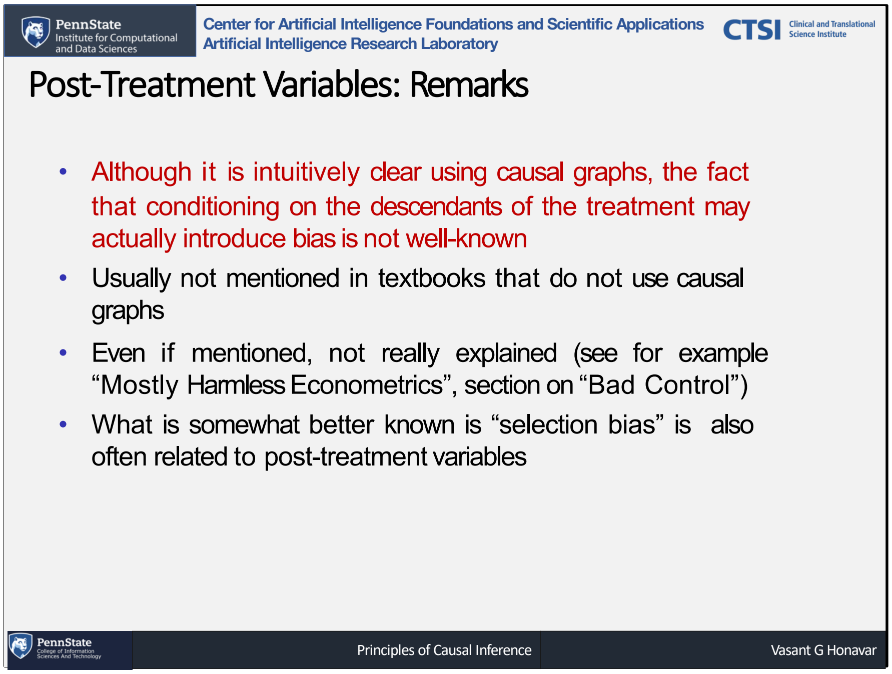


- "*M* correlates with *D* and *Y* . You may have learned in stats that you need to control for *M;* or else suffer from omitted variable bias"
- Bad idea: Conditional on *M*, *D* and *Y* are d-connected! Collider!
- $E[Y|D = 1, M = m] \neq E[Y|D = 1]$



Post-Treatment Variables: General Case

- This graph applies to situations where there are no back-door paths into *D*. Perhaps via randomization, or you block them by conditioning on *X* (not shown).
- Conditioning on *M* is forbidden by the BDC and will have two consequences:
  - You block a causal path, which you do not want
  - You open up a non-causal path, which you do not want
- This introduces bias, and it can go in any direction



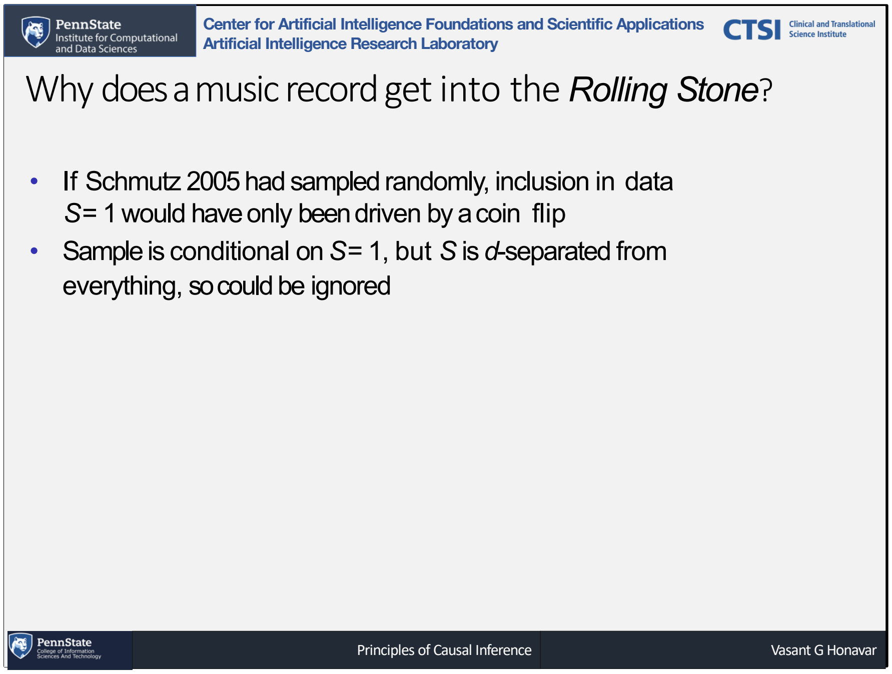
Post-Treatment Variables: Remarks

- Although it is intuitively clear using causal graphs, the fact that conditioning on the descendants of the treatment may actually introduce bias is not well-known
- Usually not mentioned in textbooks that do not use causal graphs
- Even if mentioned, not really explained (see for example “Mostly Harmless Econometrics”, section on “Bad Control”)
- What is somewhat better known is “selection bias” is also often related to post-treatment variables

Why does a music record get into the ***Rolling Stone***?

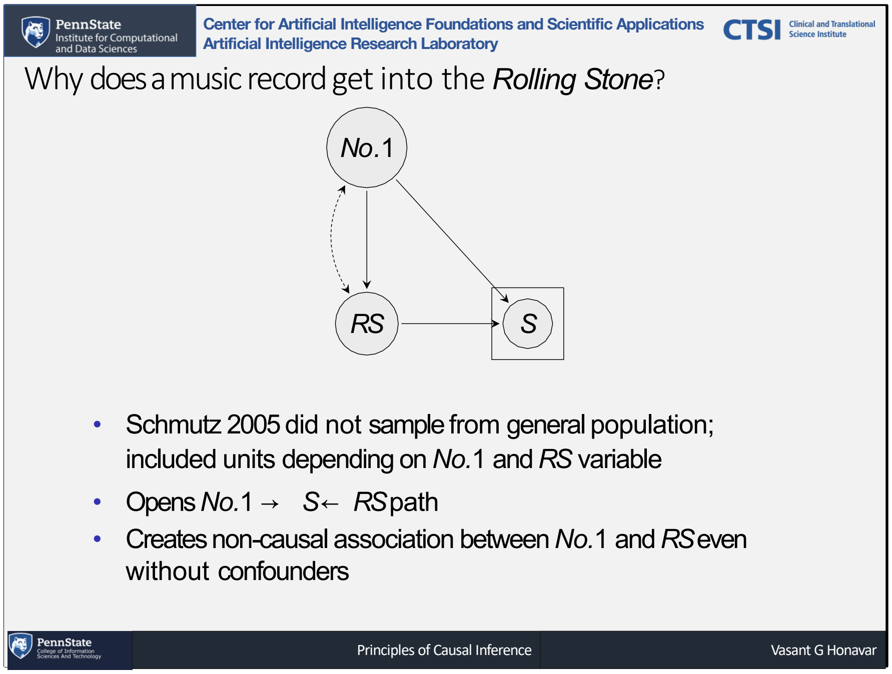
- Schmutz 2005: Why do music records get into the *Rolling Stone 500 Greatest Albums of All Time*?
- Compare these 500 records to 1,200 additional, also successful records (e.g. no. 1 records)
- Result: Strong *negative* association between being a no. 1 record and being included in *Rolling Stone* list
- Maybe *Rolling Stone* journalists are snobby and disregard commercial success?
- Or it's **selection bias**...

Source: Elwert and Winship 2014, "Endogenous Selection Bias"



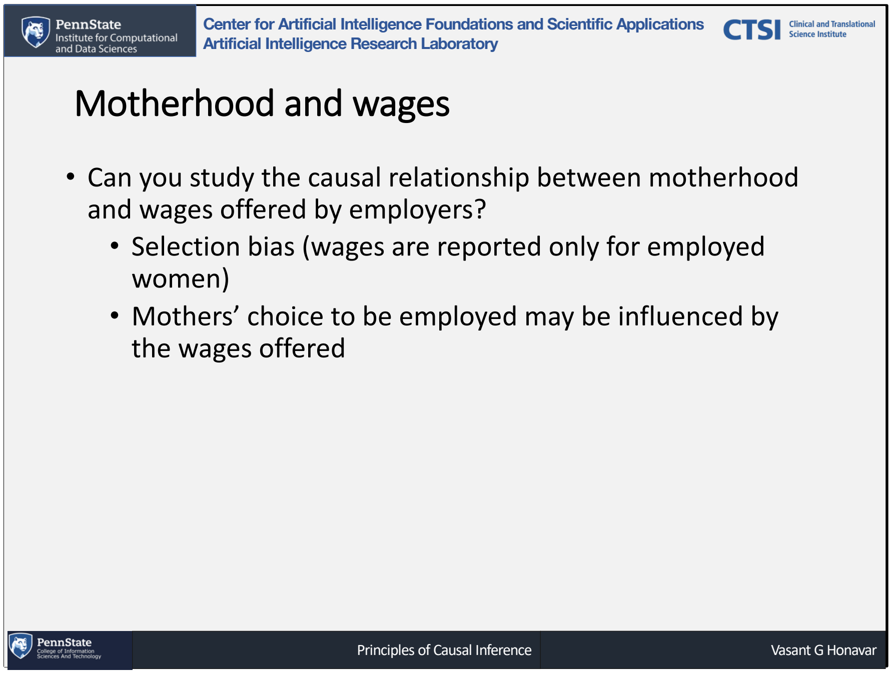
Why does a music record get into the *Rolling Stone*?

- If Schmutz 2005 had sampled randomly, inclusion in data $S = 1$ would have only been driven by a coin flip
- Sample is conditional on $S = 1$, but S is *d*-separated from everything, so could be ignored



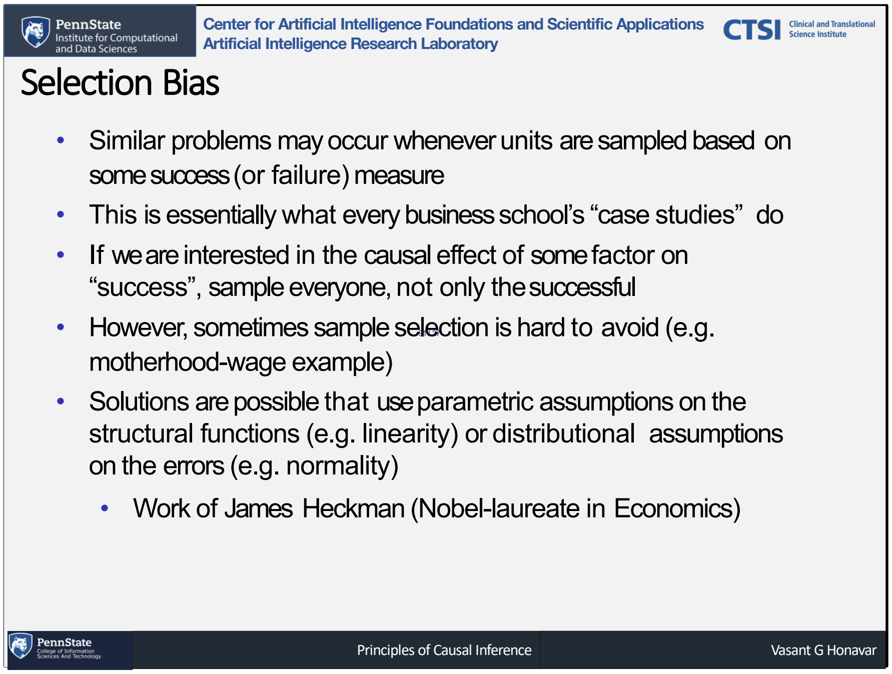
Why does a music record get into the *Rolling Stone*?

- Schmutz 2005 did not sample from general population; included units depending on *No.*1 and *RS* variable
- Opens *No.*1 → *S* ← *RS* path
- Creates non-causal association between *No.*1 and *RS* even without confounders



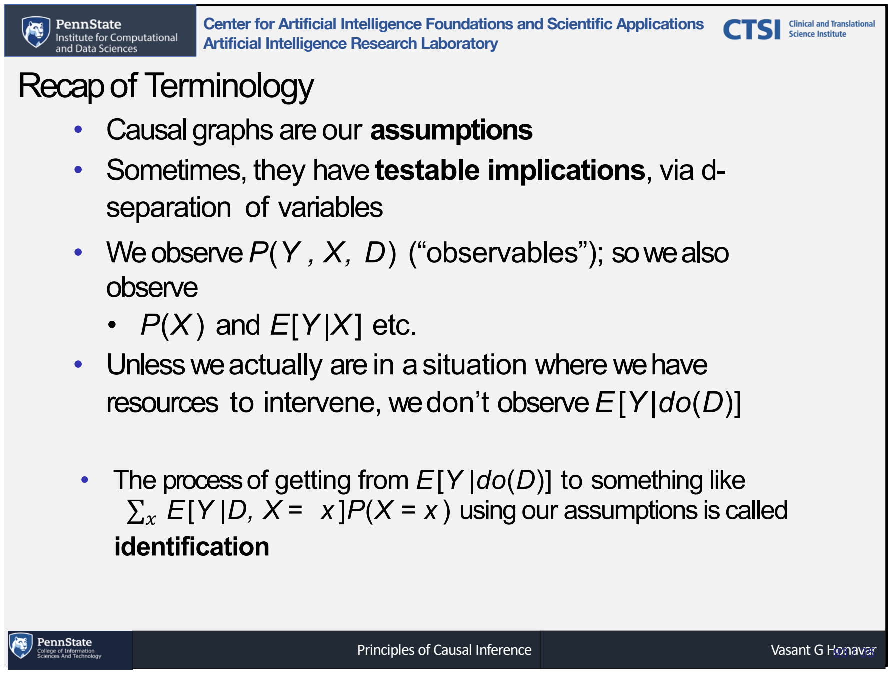
Motherhood and wages

- Can you study the causal relationship between motherhood and wages offered by employers?
  - Selection bias (wages are reported only for employed women)
  - Mothers' choice to be employed may be influenced by the wages offered



Selection Bias

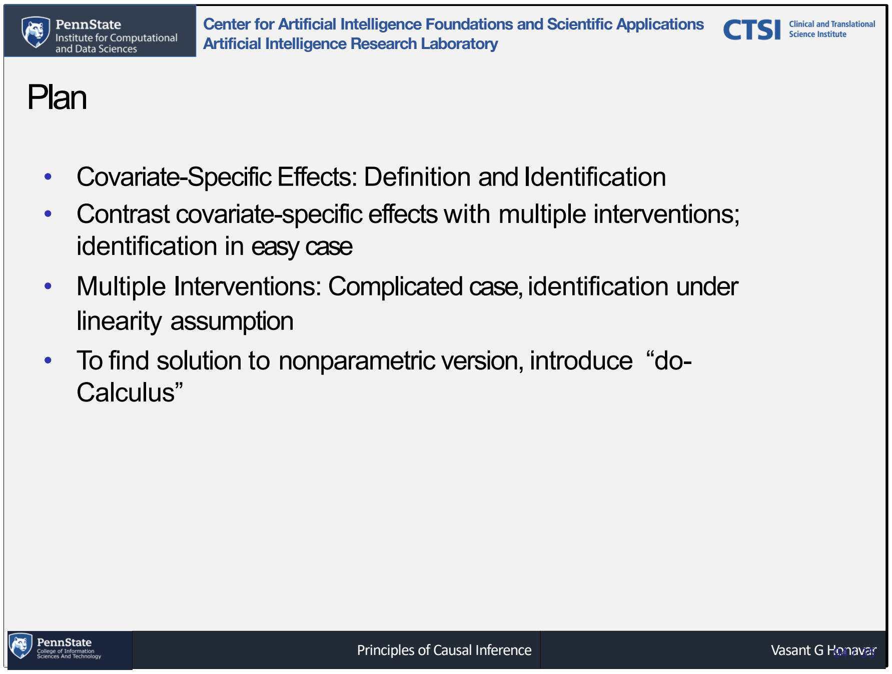
- Similar problems may occur whenever units are sampled based on some success (or failure) measure
- This is essentially what every business school's "case studies" do
- If we are interested in the causal effect of some factor on "success", sample everyone, not only the successful
- However, sometimes sample selection is hard to avoid (e.g. motherhood-wage example)
- Solutions are possible that use parametric assumptions on the structural functions (e.g. linearity) or distributional assumptions on the errors (e.g. normality)
  - Work of James Heckman (Nobel-laureate in Economics)



Recap of Terminology

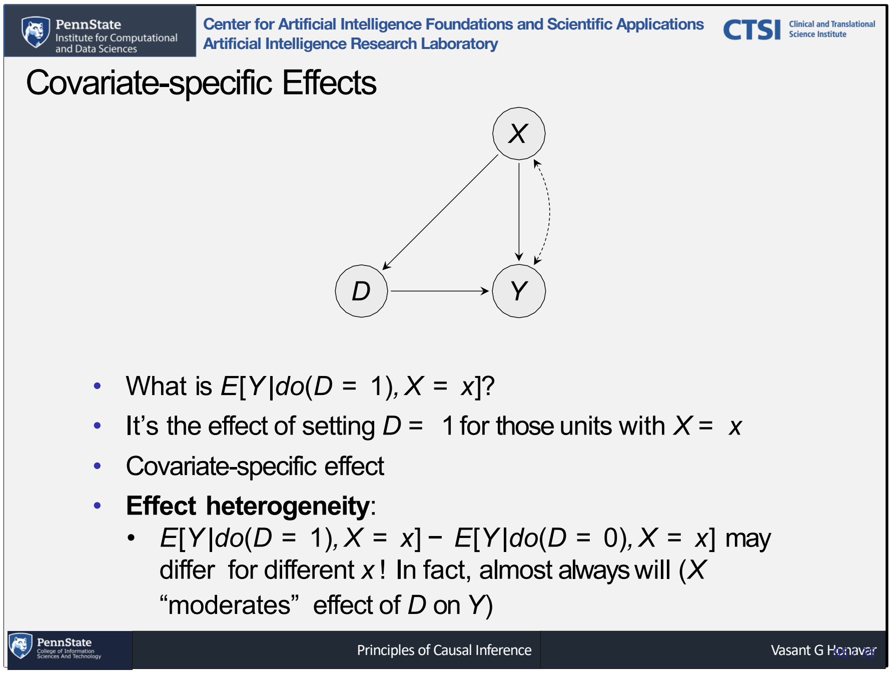
- Causal graphs are our **assumptions**
- Sometimes, they have **testable implications**, via d-separation of variables
- We observe $P(Y, X, D)$ ("observables"); so we also observe
  - $P(X)$ and $E[Y|X]$ etc.
- Unless we actually are in a situation where we have resources to intervene, we don't observe $E[Y|do(D)]$

- The process of getting from $E[Y|do(D)]$ to something like $\sum_x E[Y|D, X = x]P(X = x)$ using our assumptions is called **identification**



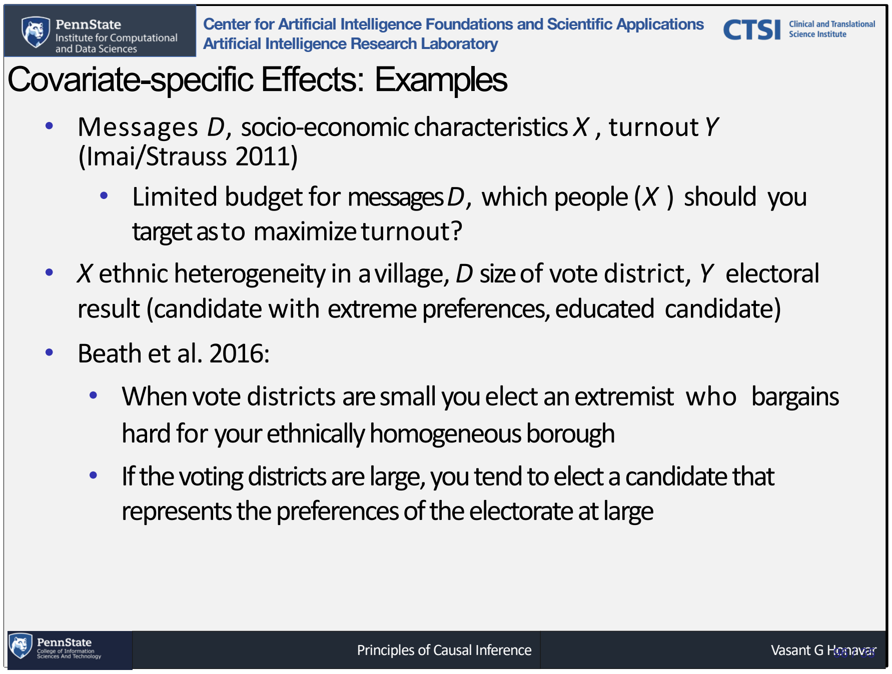
Plan

- Covariate-Specific Effects: Definition and Identification
- Contrast covariate-specific effects with multiple interventions; identification in easy case
- Multiple Interventions: Complicated case, identification under linearity assumption
- To find solution to nonparametric version, introduce "do-Calculus"



Covariate-specific Effects

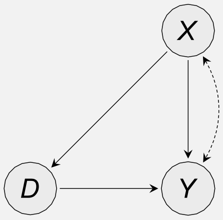
- What is $E[Y|do(D = 1), X = x]$?
- It's the effect of setting $D = 1$ for those units with $X = x$
- Covariate-specific effect
- **Effect heterogeneity**:
  - $E[Y|do(D = 1), X = x] - E[Y|do(D = 0), X = x]$ may differ for different x! In fact, almost always will (X "moderates" effect of D on Y)



Covariate-specific Effects: Examples

- Messages D, socio-economic characteristics X , turnout Y (Imai/Strauss 2011)
  - Limited budget for messages D, which people (X) should you target as to maximize turnout?
- X ethnic heterogeneity in a village, D size of vote district, Y electoral result (candidate with extreme preferences, educated candidate)
- Beath et al. 2016:
  - When vote districts are small you elect an extremist who bargains hard for your ethnically homogeneous borough
  - If the voting districts are large, you tend to elect a candidate that represents the preferences of the electorate at large

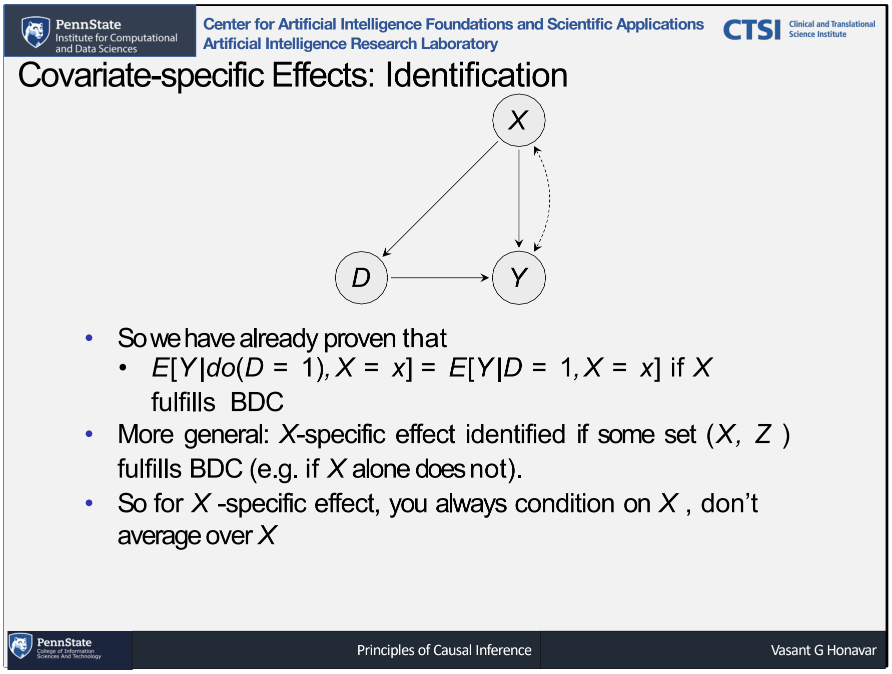
Covariate-specific Effects: Identification



- When we used the BDC, we first wrote $E[Y|do(D = 1)] =$

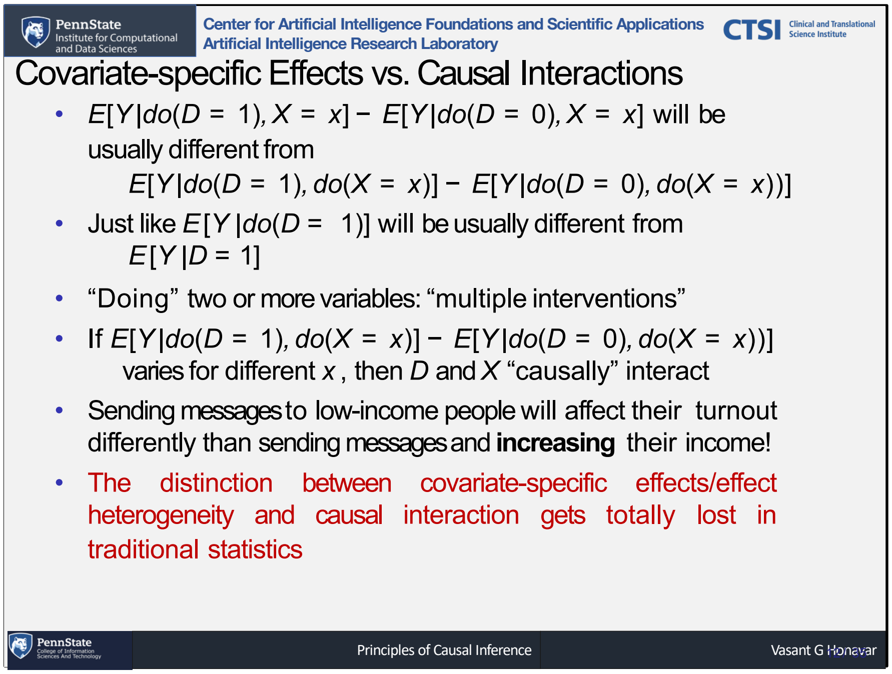
$$\sum_x E[Y|do(D = 1), X = x]P(X = x|do(D = 1))$$

- What were the next two steps?
  - $P(x = x|do(D = 1)) = P(X = x)$ because if X fulfills BDC, it contains no descendants of D
  - $E[Y|do(D = 1), X = x] = E[Y|D = 1, X = x]$:
  - Conditional on X, doing D is like observing D



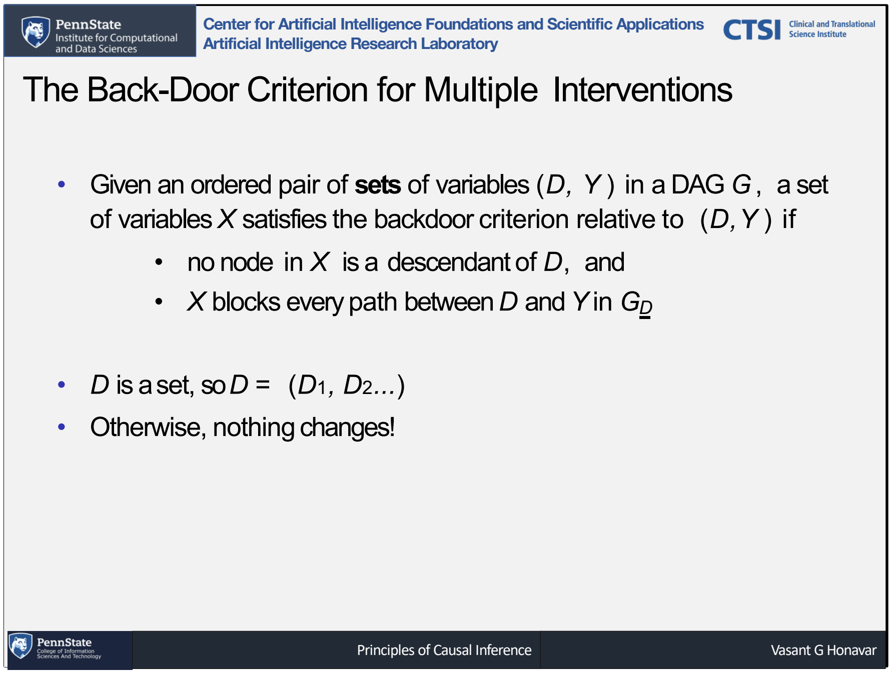
Covariate-specific Effects: Identification

- So we have already proven that
  - $E[Y|do(D = 1), X = x] = E[Y|D = 1, X = x]$ if X fulfills BDC
- More general: X-specific effect identified if some set (X, Z) fulfills BDC (e.g. if X alone does not).
- So for X -specific effect, you always condition on X , don't average over X



Covariate-specific Effects vs. Causal Interactions

- $E[Y|do(D = 1), X = x] - E[Y|do(D = 0), X = x]$ will be usually different from
  $E[Y|do(D = 1), do(X = x)] - E[Y|do(D = 0), do(X = x))]$
- Just like $E[Y|do(D = 1)]$ will be usually different from
  $E[Y|D = 1]$
- "Doing" two or more variables: "multiple interventions"
- If $E[Y|do(D = 1), do(X = x)] - E[Y|do(D = 0), do(X = x))]$
  varies for different x , then D and X "causally" interact
- Sending messages to low-income people will affect their turnout differently than sending messages and **increasing** their income!
- The distinction between covariate-specific effects/effect heterogeneity and causal interaction gets totally lost in traditional statistics

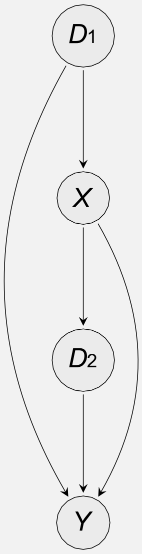


The Back-Door Criterion for Multiple Interventions

- Given an ordered pair of **sets** of variables (D, Y) in a DAG G, a set of variables X satisfies the backdoor criterion relative to (D, Y) if
  - no node in X is a descendant of D, and
  - X blocks every path between D and Y in $G_{\underline{D}}$
- D is a set, so $D = (D_1, D_2 \ldots)$
- Otherwise, nothing changes!

Multiple Interventions: Where the BDC fails

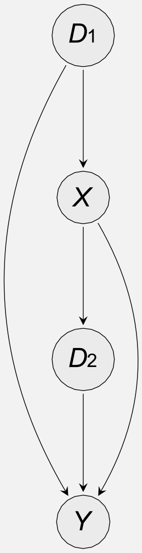
- Let’s first analyze this graph:
  - D_1 affects X , D_2 affects Y
  - X affects D_2 and Y
  - D_2 affects Y
- Example:
  - Medical treatment D_1 at time 1.
  - X health status after a while.
  - If healthy, stop treatment. If not, give treatment D_2 at time 2.
  - Then check health Y again

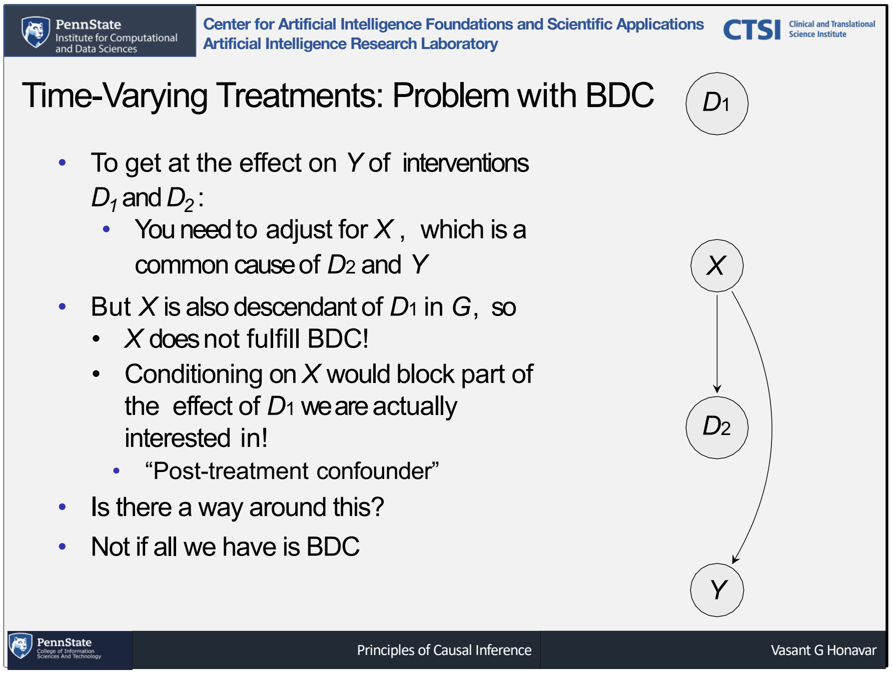


Source: Biostatistics, James Robins

Time-Varying Treatments: Problem with BDC

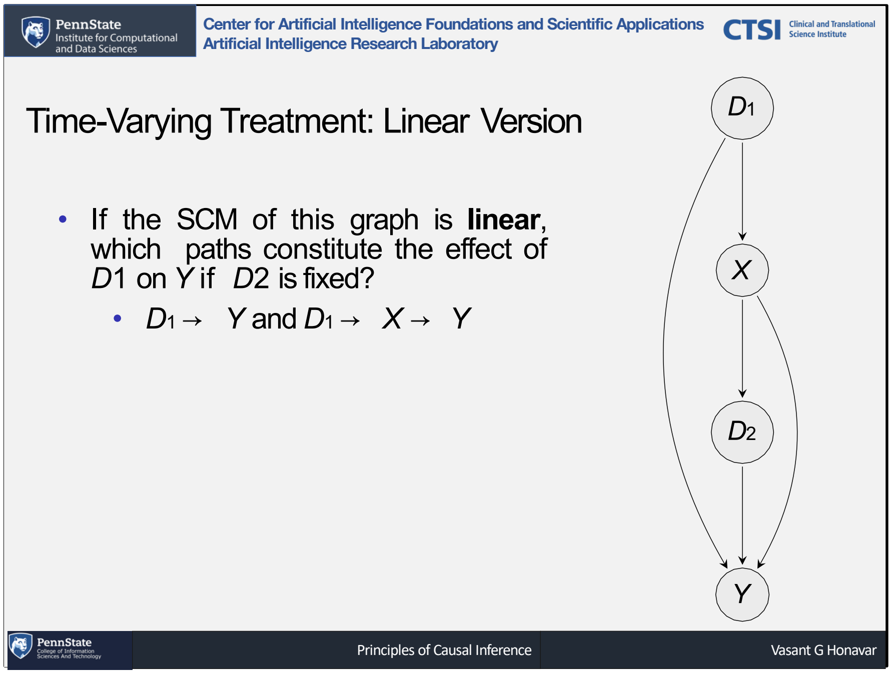
- Question: What's the effect of a given "strategy" (D_1, D_2) on Y?
- Put differently, can we identify the joint direct effect of (D_1, D_2)?
- Which variables do we need to adjust for?





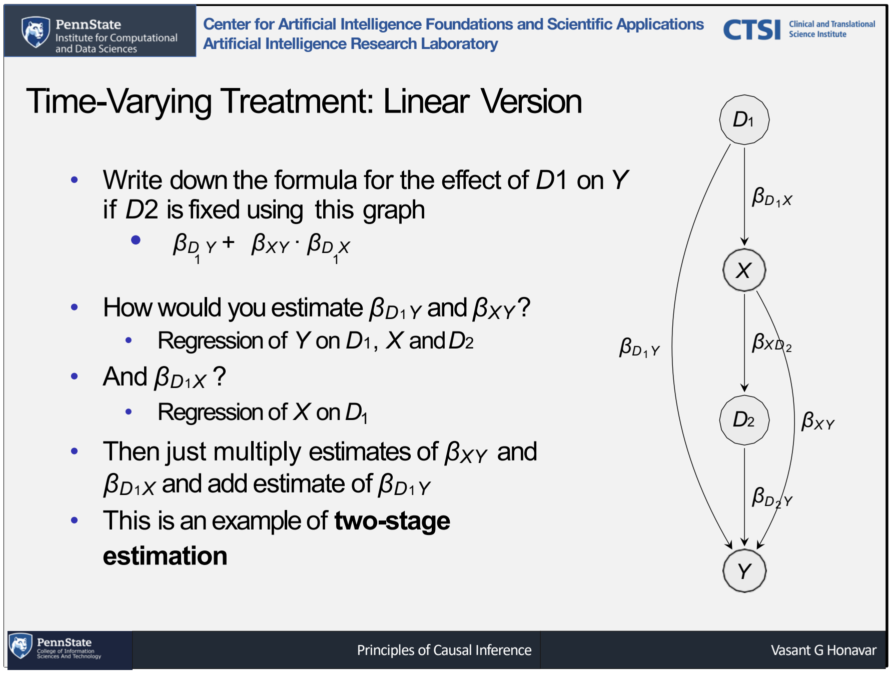
Time-Varying Treatments: Problem with BDC

- To get at the effect on *Y* of interventions D_1 and D_2:
  - You need to adjust for *X* , which is a common cause of D_2 and *Y*
- But *X* is also descendant of D_1 in *G*, so
  - *X* does not fulfill BDC!
  - Conditioning on *X* would block part of the effect of D_1 we are actually interested in!
    - “Post-treatment confounder”
- Is there a way around this?
- Not if all we have is BDC



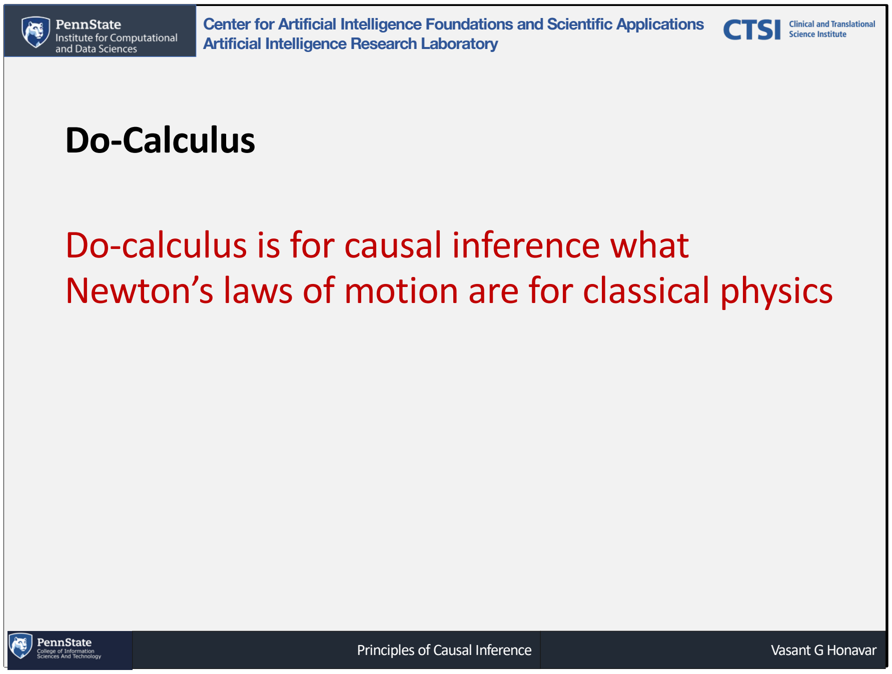
Time-Varying Treatment: Linear Version

- If the SCM of this graph is **linear**, which paths constitute the effect of *D*1 on *Y* if *D*2 is fixed?
  - $D_1 \rightarrow Y$ and $D_1 \rightarrow X \rightarrow Y$



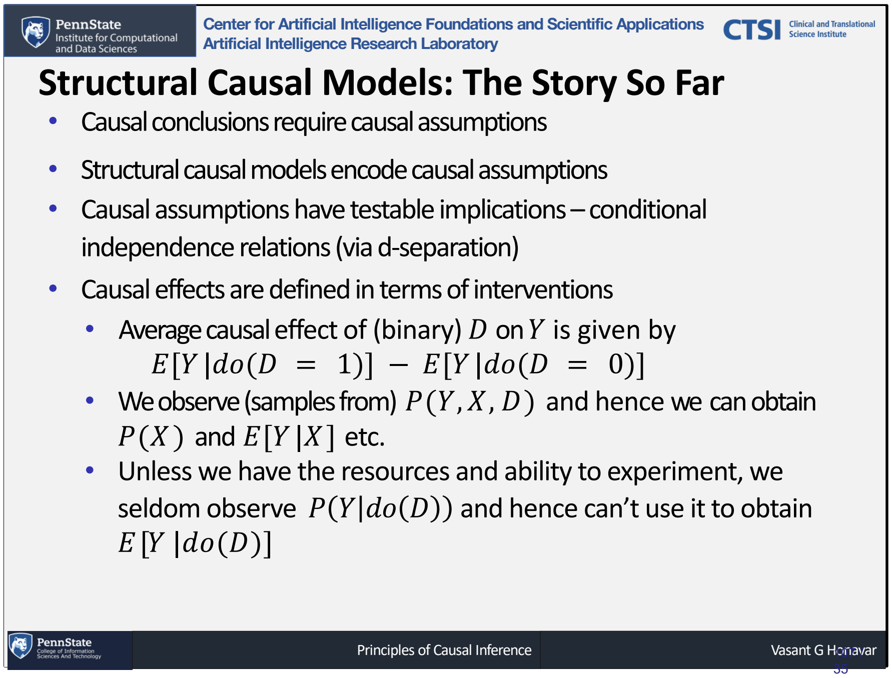
Time-Varying Treatment: Linear Version

- Write down the formula for the effect of $D1$ on Y if $D2$ is fixed using this graph
  - $\beta_{D_1Y} + \beta_{XY} \cdot \beta_{D_1X}$
- How would you estimate β_{D_1Y} and β_{XY}?
  - Regression of Y on D_1, X and D_2
- And β_{D_1X} ?
  - Regression of X on D_1
- Then just multiply estimates of β_{XY} and β_{D_1X} and add estimate of β_{D_1Y}
- This is an example of **two-stage estimation**



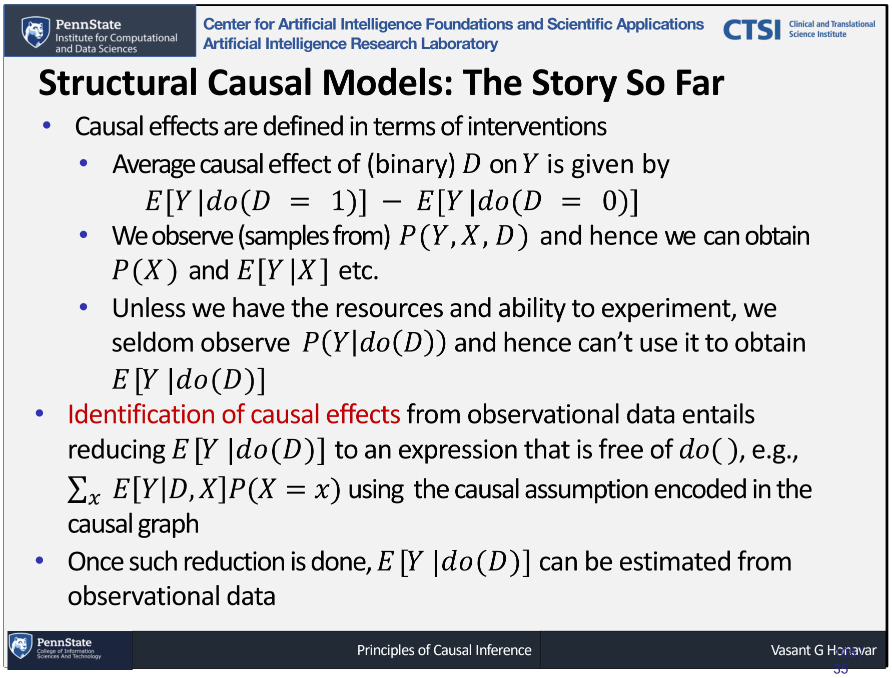
Do-Calculus

Do-calculus is for causal inference what Newton’s laws of motion are for classical physics



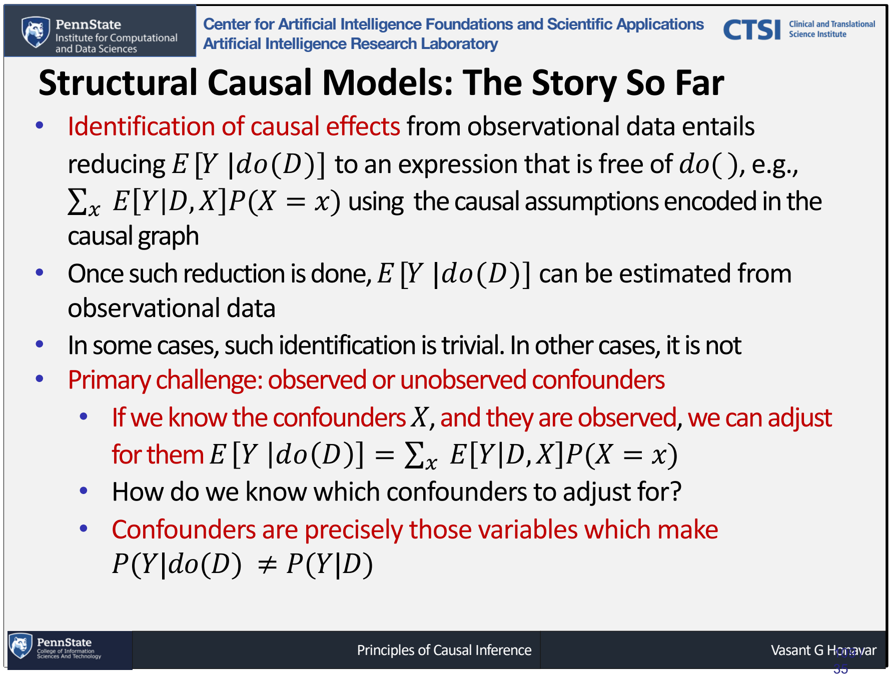
Structural Causal Models: The Story So Far

- Causal conclusions require causal assumptions
- Structural causal models encode causal assumptions
- Causal assumptions have testable implications – conditional independence relations (via d-separation)
- Causal effects are defined in terms of interventions
  - Average causal effect of (binary) D on Y is given by
    $E[Y|do(D = 1)] - E[Y|do(D = 0)]$
  - We observe (samples from) $P(Y, X, D)$ and hence we can obtain $P(X)$ and $E[Y|X]$ etc.
  - Unless we have the resources and ability to experiment, we seldom observe $P(Y|do(D))$ and hence can't use it to obtain $E[Y|do(D)]$



Structural Causal Models: The Story So Far

- Causal effects are defined in terms of interventions
  - Average causal effect of (binary) D on Y is given by
    $E[Y|do(D = 1)] - E[Y|do(D = 0)]$
  - We observe (samples from) $P(Y, X, D)$ and hence we can obtain $P(X)$ and $E[Y|X]$ etc.
  - Unless we have the resources and ability to experiment, we seldom observe $P(Y|do(D))$ and hence can't use it to obtain $E[Y|do(D)]$
- Identification of causal effects from observational data entails reducing $E[Y|do(D)]$ to an expression that is free of $do(\,)$, e.g., $\sum_x E[Y|D,X]P(X = x)$ using the causal assumption encoded in the causal graph
- Once such reduction is done, $E[Y|do(D)]$ can be estimated from observational data



PennState Institute for Computational and Data Sciences

Center for Artificial Intelligence Foundations and Scientific Applications
Artificial Intelligence Research Laboratory

CTSI Clinical and Translational Science Institute

Structural Causal Models: The Story So Far

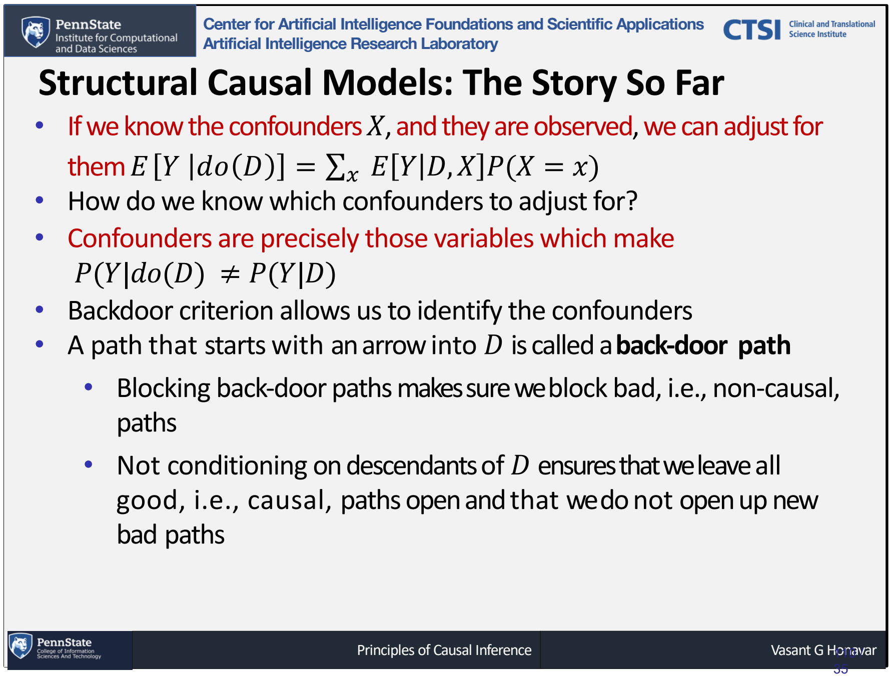
- Identification of causal effects from observational data entails reducing $E\,[Y\;|do(D)]$ to an expression that is free of $do(\,)$, e.g., $\sum_x E[Y|D,X]P(X=x)$ using the causal assumptions encoded in the causal graph
- Once such reduction is done, $E\,[Y\;|do(D)]$ can be estimated from observational data
- In some cases, such identification is trivial. In other cases, it is not
- Primary challenge: observed or unobserved confounders
  - If we know the confounders X, and they are observed, we can adjust for them $E\,[Y\;|do(D)] = \sum_x E[Y|D,X]P(X=x)$
  - How do we know which confounders to adjust for?
  - Confounders are precisely those variables which make $P(Y|do(D)\;\neq P(Y|D)$

PennState College of Information Sciences and Technology

Principles of Causal Inference

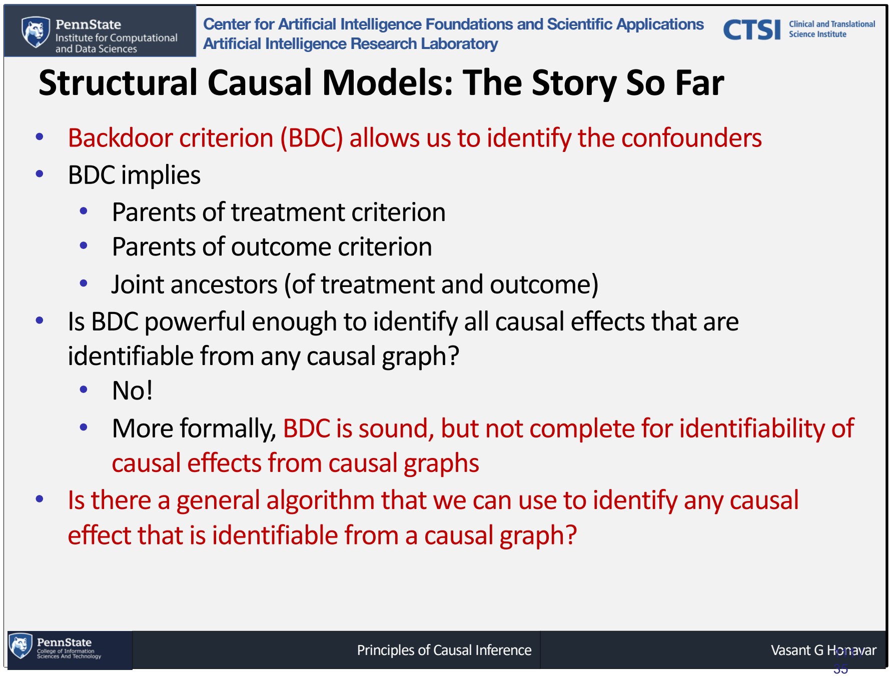
Vasant G Honavar

- Unless we actually are in a situation where we have resources to intervene, we don't observe $E[Y|do(D)]$



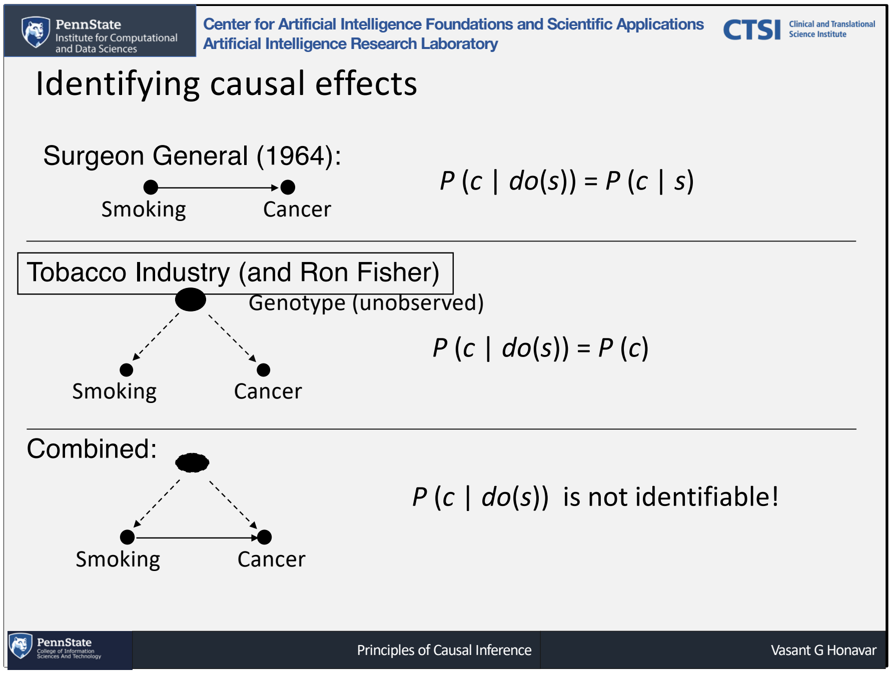
Structural Causal Models: The Story So Far

- If we know the confounders X, and they are observed, we can adjust for them $E\,[Y\ |do(D)] = \sum_x\ E[Y|D,X]P(X = x)$
- How do we know which confounders to adjust for?
- Confounders are precisely those variables which make $P(Y|do(D)\ \neq P(Y|D)$
- Backdoor criterion allows us to identify the confounders
- A path that starts with an arrow into D is called a **back-door path**
    - Blocking back-door paths makes sure we block bad, i.e., non-causal, paths
    - Not conditioning on descendants of D ensures that we leave all good, i.e., causal, paths open and that we do not open up new bad paths



Structural Causal Models: The Story So Far

- Backdoor criterion (BDC) allows us to identify the confounders
- BDC implies
  - Parents of treatment criterion
  - Parents of outcome criterion
  - Joint ancestors (of treatment and outcome)
- Is BDC powerful enough to identify all causal effects that are identifiable from any causal graph?
  - No!
  - More formally, BDC is sound, but not complete for identifiability of causal effects from causal graphs
- Is there a general algorithm that we can use to identify any causal effect that is identifiable from a causal graph?



Identifying causal effects

Surgeon General (1964):

$P\ (c \mid do(s)) = P\ (c \mid s)$

Tobacco Industry (and Ron Fisher)

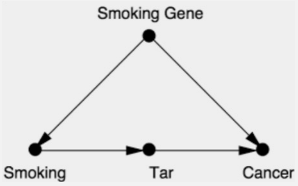
$P\ (c \mid do(s)) = P\ (c)$

Combined:

$P\ (c \mid do(s))$ is not identifiable!

The Front-Door Criterion (FDC)

Suppose we assume:



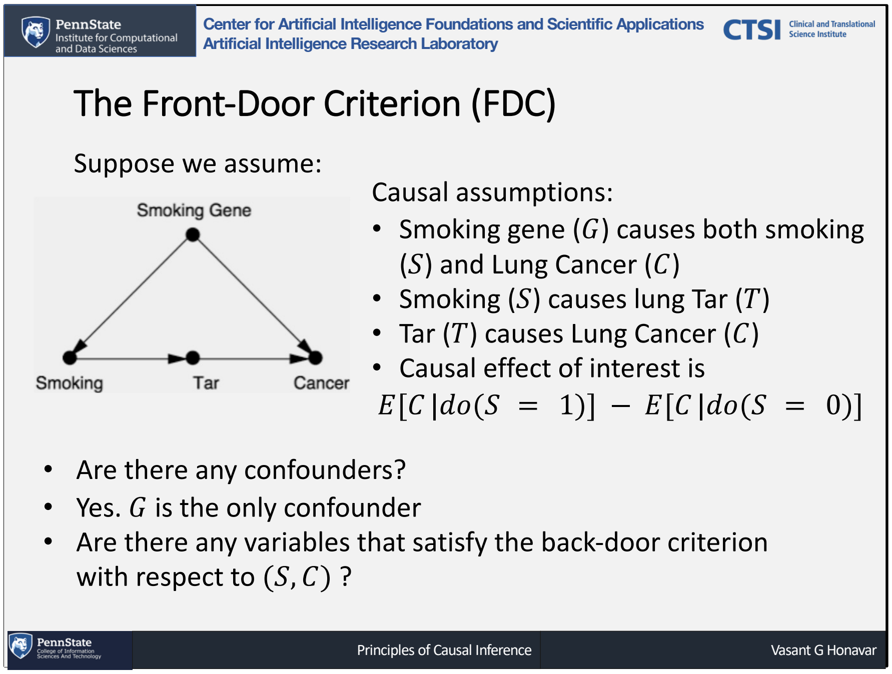
Causal assumptions:

- Smoking gene (G) causes both smoking (S) and Lung Cancer (C)
- Smoking (S) causes lung Tar (T)
- Tar (T) causes Lung Cancer (C)
- Causal effect of interest is

$$E[C\,|do(S\ =\ 1)]\ -\ E[C\,|do(S\ =\ 0)]$$

Suppose

- We have collected observational data on S, T, C for a set of individuals
- We cannot collect data for G because we do not know if a smoking gene exists



The Front-Door Criterion (FDC)

Suppose we assume:

Causal assumptions:

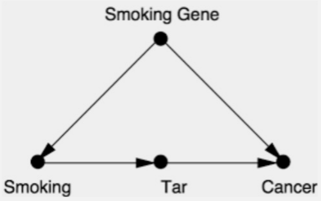
- Smoking gene (G) causes both smoking (S) and Lung Cancer (C)
- Smoking (S) causes lung Tar (T)
- Tar (T) causes Lung Cancer (C)
- Causal effect of interest is

$$E[C\,|do(S = 1)] - E[C\,|do(S = 0)]$$

- Are there any confounders?
- Yes. G is the only confounder
- Are there any variables that satisfy the back-door criterion with respect to (S, C) ?

The Front-Door Criterion (FDC)

Suppose we assume:



Causal assumptions:

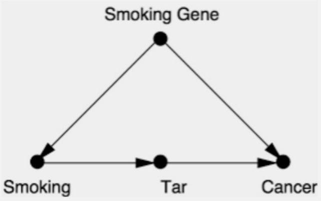
- Smoking gene (G) causes both smoking (S) and Lung Cancer (C)
- Smoking (S) causes lung Tar (T)
- Tar (T) causes Lung Cancer (C)
- Causal effect of interest is

$$E[C\,|do(S = 1)] - E[C\,|do(S = 0)]$$

- Are there any variables that satisfy the back-door criterion with respect to (S, C)?
- S, T and C are not candidates
- What about G?
- G would satisfy the backdoor criterion if it were observed!
- But it is not!

The Front-Door Criterion (FDC)

Suppose we assume:



Causal assumptions:

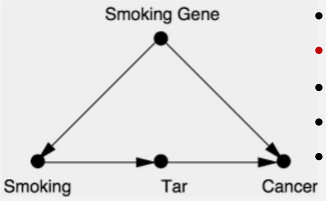
- Smoking gene (G) causes both smoking (S) and Lung Cancer (C)
- Smoking (S) causes lung Tar (T)
- Tar (T) causes Lung Cancer (C)
- Causal effect of interest is

$$E[C\,|do(S\ =\ 1)]\ -\ E[C\,|do(S\ =\ 0)]$$

- We cannot use the back door to adjust for G
- Is there another way?

The Front-Door Criterion (FDC)

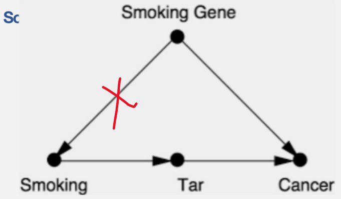
Suppose we assume: $E[C\,|do(S = 1)] - E[C\,|do(S = 0)]$



- We cannot use BDC to adjust for G
- Is there another way?
- We can get the causal effect of S on T
- We can get the causal effect of T on C
- And combine them to get the causal effect of S on C

The Front-Door Criterion (FDC)

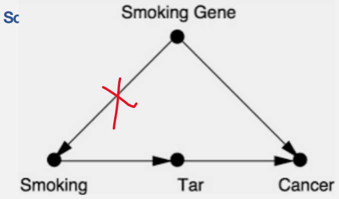
$E[C\,|do(S\ =\ 1)]\ -\ E[C\,|do(S\ =\ 0)]$



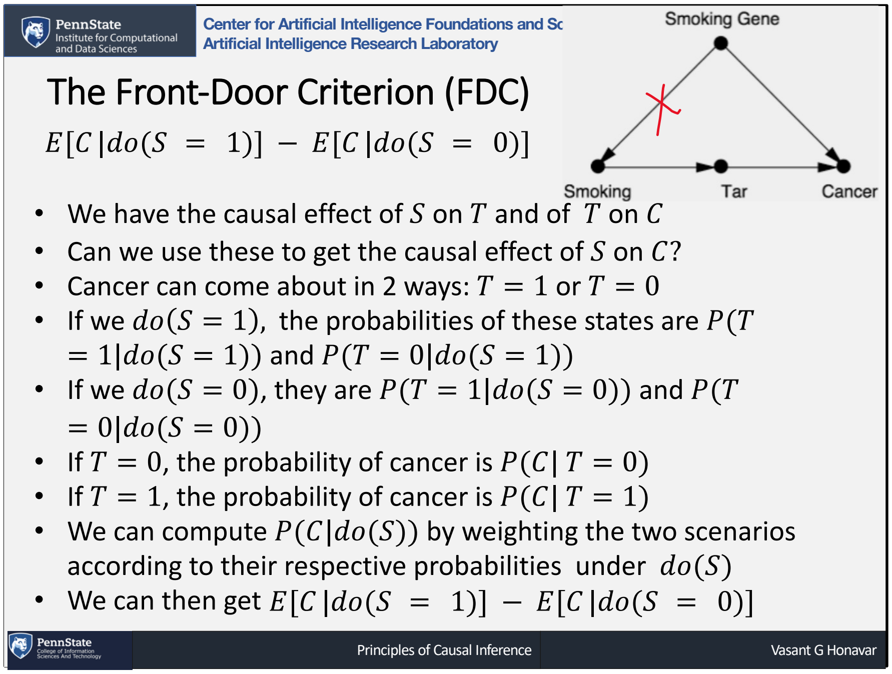
- We can get the causal effect of S on T
- Why?
- When we condition on S, There is no unblocked backdoor path from S to C because $S \leftarrow G \rightarrow C \leftarrow T$ is already blocked by the collider C
- We can observe $P(T|S=1)\ -\ P(T|S=0)$ to get the causal effect of S on T

The Front-Door Criterion (FDC)

$E[C\,|do(S\ =\ 1)] - E[C\,|do(S\ =\ 0)]$



- We can get the causal effect of T on C
- How?
- We can block the backdoor path into T which is T ← S ← G → C by adjusting for S
- We can get $P(C|do(T=1)) - P(S|do(T=0))$ using the backdoor adjustment formula



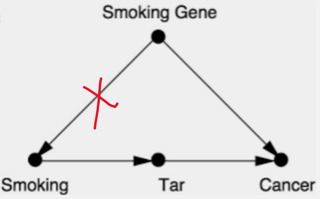
The Front-Door Criterion (FDC)

$E[C\,|do(S\ =\ 1)]\ -\ E[C\,|do(S\ =\ 0)]$

- We have the causal effect of S on T and of T on C
- Can we use these to get the causal effect of S on C?
- Cancer can come about in 2 ways: $T = 1$ or $T = 0$
- If we $do(S = 1)$, the probabilities of these states are $P(T = 1|do(S = 1))$ and $P(T = 0|do(S = 1))$
- If we $do(S = 0)$, they are $P(T = 1|do(S = 0))$ and $P(T = 0|do(S = 0))$
- If $T = 0$, the probability of cancer is $P(C|\,T = 0)$
- If $T = 1$, the probability of cancer is $P(C|\,T = 1)$
- We can compute $P(C|do(S))$ by weighting the two scenarios according to their respective probabilities under $do(S)$
- We can then get $E[C\,|do(S\ =\ 1)]\ -\ E[C\,|do(S\ =\ 0)]$

The Front-Door Criterion (FDC)

$E[C\,|do(S\ =\ 1)]\ -\ E[C\,|do(S\ =\ 0)]$



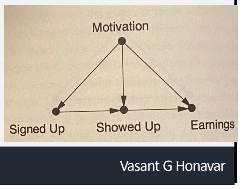
- What did we do?
- To obtain the causal effect of S on C, we adjusted for S and T which lie on the front-door path from S to C

$$P(C|do(S)) = \sum_t P(T = t, S) \sum_s P(C|S = s, T = t)P(S = s)$$

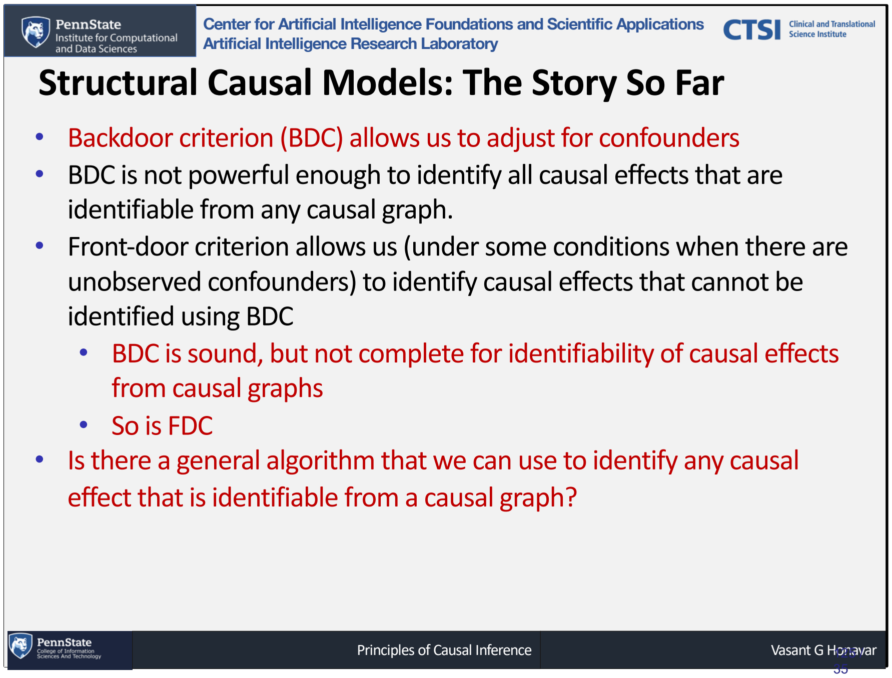
- There is do on the LHS but no do on the RHS!
- G, the unobserved confounder does not appear in the RHS
- If the causal graph shown is an accurate model of causal mechanism of cancer, the controversy about whether and to what extent smoking causes cancer could have been answered by an observational study that measured S, T, and C

Front-Door Criterion

- Glynn and Kashin (2018) studied the effect of job training services on earnings (without assuming a SCM)
- Data from RCT compared with an observational study
- They did not use a causal diagram but compared RCT with results of back-door and front-door adjustments
- Motivation is unobserved – back door criterion can't be applied, but Glynn and Kashin did with other potential confounders like age, gender, race, education
- Strictly speaking, front-door criterion can also not be applied exactly – because of the direct link from Motivation to Showed up
- Adjustment using FDC beats that using BDC in terms of agreement with RCT
- Study shows the power of FDC

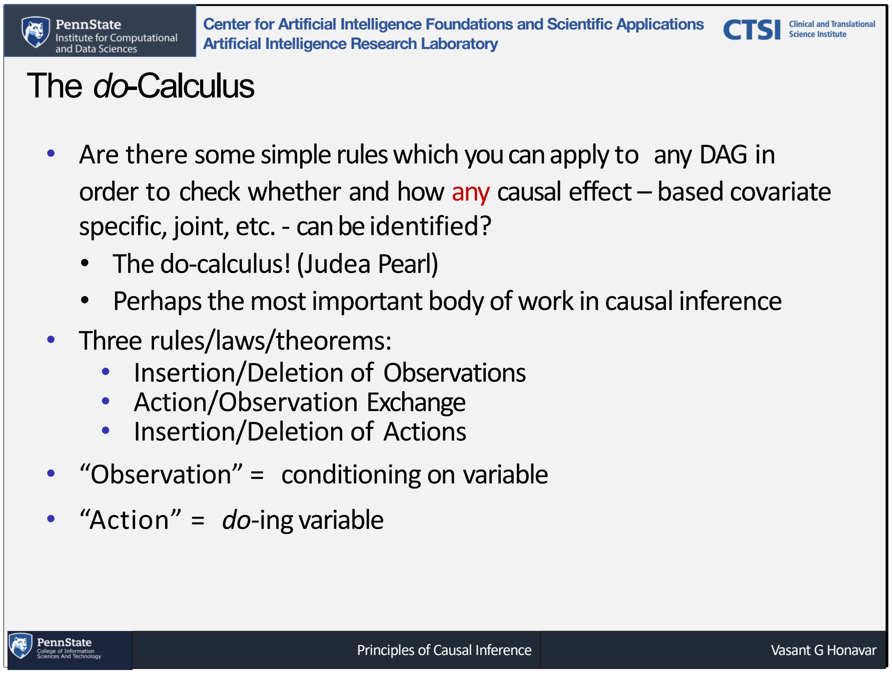


Glynn, Adam N., and Konstantin Kashin. "Front-door versus back-door adjustment with unmeasured confounding: Bias formulas for front-door and hybrid adjustments with application to a job training program." *Journal of the American Statistical Association* 113, no. 523 (2018): 1040-1049.



Structural Causal Models: The Story So Far

- Backdoor criterion (BDC) allows us to adjust for confounders
- BDC is not powerful enough to identify all causal effects that are identifiable from any causal graph.
- Front-door criterion allows us (under some conditions when there are unobserved confounders) to identify causal effects that cannot be identified using BDC
  - BDC is sound, but not complete for identifiability of causal effects from causal graphs
  - So is FDC
- Is there a general algorithm that we can use to identify any causal effect that is identifiable from a causal graph?

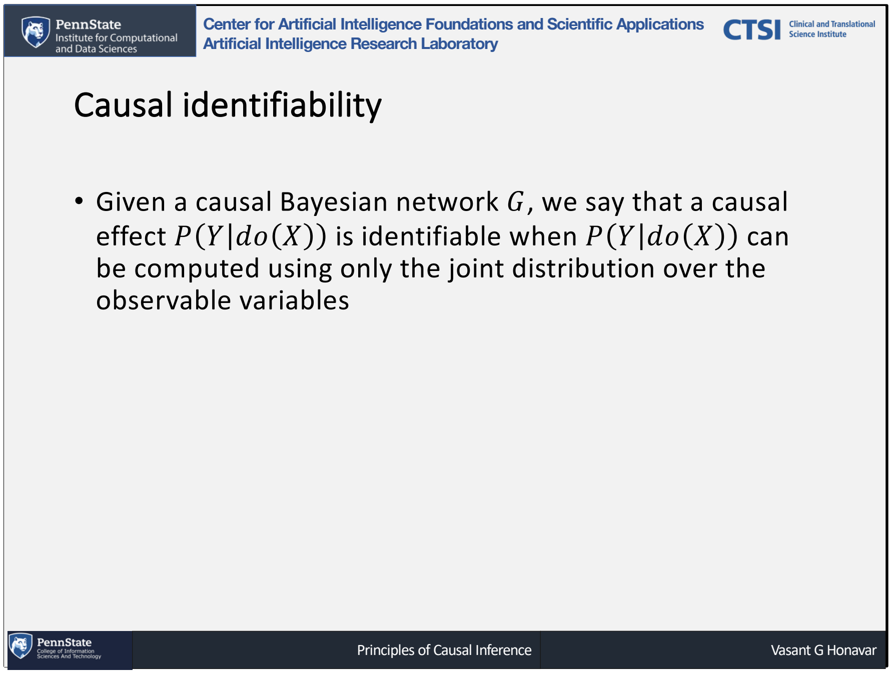


The *do*-Calculus

- Are there some simple rules which you can apply to any DAG in order to check whether and how any causal effect – based covariate specific, joint, etc. - can be identified?
  - The do-calculus! (Judea Pearl)
  - Perhaps the most important body of work in causal inference
- Three rules/laws/theorems:
  - Insertion/Deletion of Observations
  - Action/Observation Exchange
  - Insertion/Deletion of Actions
- “Observation” = conditioning on variable
- “Action” = *do*-ing variable

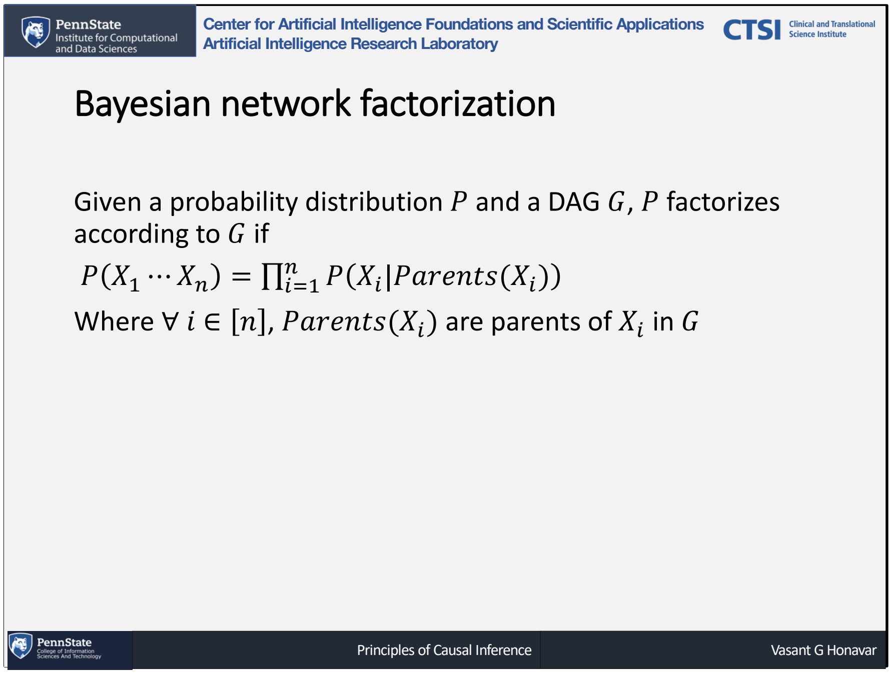
Identifying causal effects

- Postulate a causal graph (causal assumptions)
- See how the estimand (causal effect of interest) can be written as a function of the postulated causal graph
- Check whether this function can be calculated from observations (using do calculus)



Causal identifiability

- Given a causal Bayesian network G, we say that a causal effect $P(Y|do(X))$ is identifiable when $P(Y|do(X))$ can be computed using only the joint distribution over the observable variables

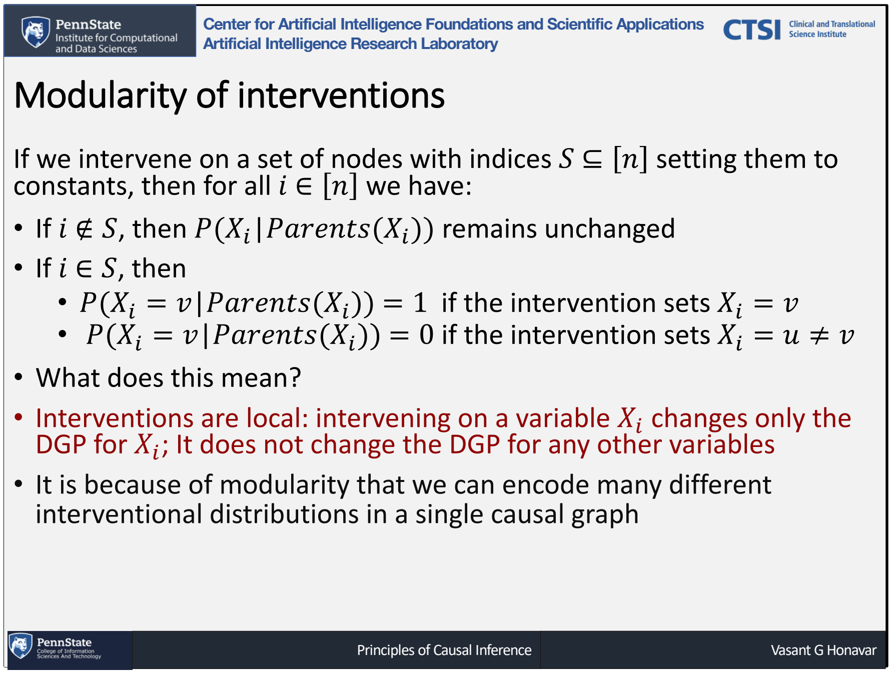


Bayesian network factorization

Given a probability distribution P and a DAG G, P factorizes according to G if

$$P(X_1 \cdots X_n) = \prod_{i=1}^{n} P(X_i|Parents(X_i))$$

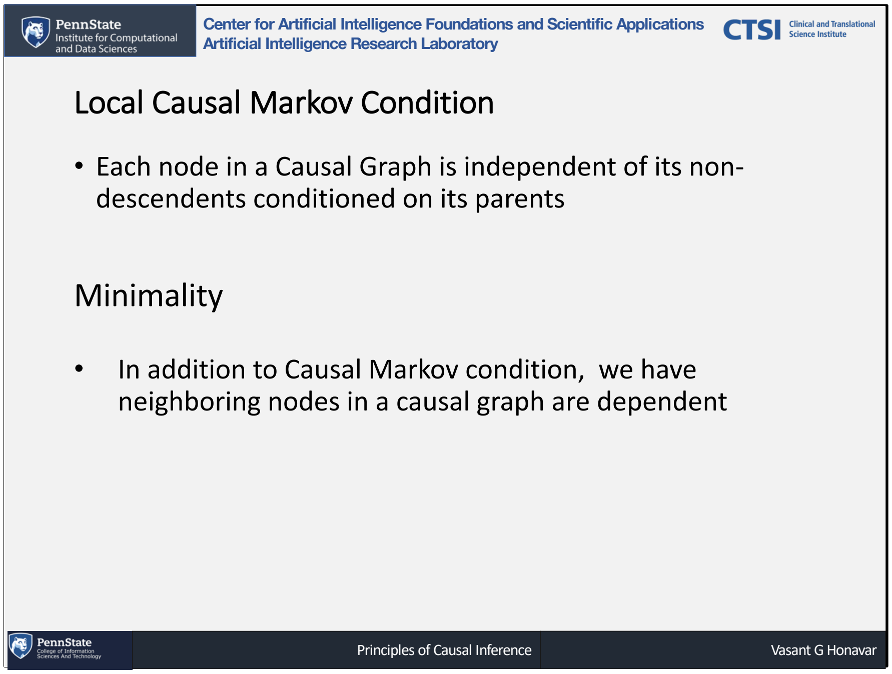
Where $\forall\ i \in [n]$, $Parents(X_i)$ are parents of X_i in G



Modularity of interventions

If we intervene on a set of nodes with indices $S \subseteq [n]$ setting them to constants, then for all $i \in [n]$ we have:

- If $i \notin S$, then $P(X_i | Parents(X_i))$ remains unchanged
- If $i \in S$, then
  - $P(X_i = v | Parents(X_i)) = 1$ if the intervention sets $X_i = v$
  - $P(X_i = v | Parents(X_i)) = 0$ if the intervention sets $X_i = u \neq v$
- What does this mean?
- Interventions are local: intervening on a variable X_i changes only the DGP for X_i; It does not change the DGP for any other variables
- It is because of modularity that we can encode many different interventional distributions in a single causal graph



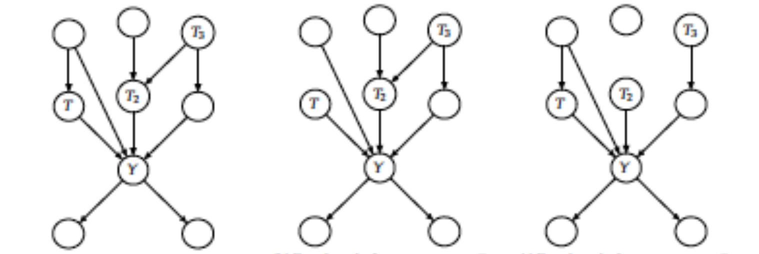
Local Causal Markov Condition

- Each node in a Causal Graph is independent of its non-descendents conditioned on its parents

Minimality

- In addition to Causal Markov condition, we have neighboring nodes in a causal graph are dependent

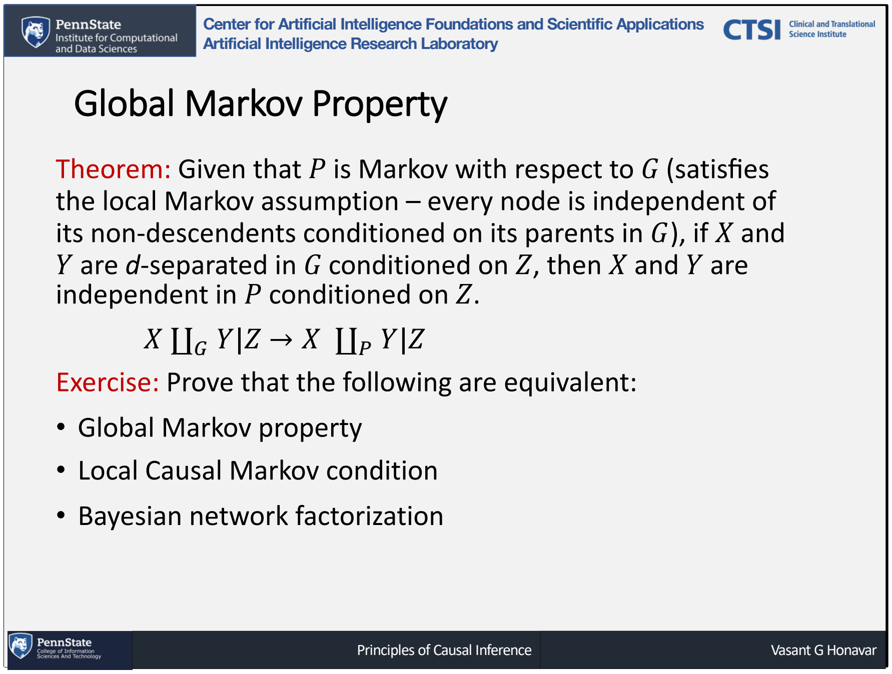
Because of modularity, we can encode many interventional distributions in a single causal graph



(a) Causal graph for observational distribution

(b) Causal graph after intervention on T (interventional distribution)

(c) Causal graph after intervention on T_2 (interventional distribution)



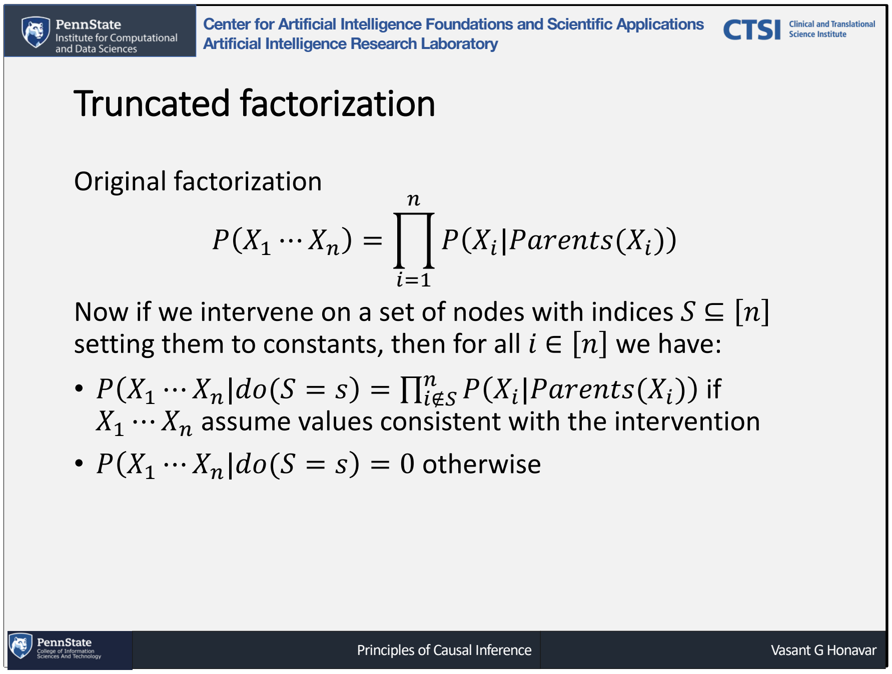
Global Markov Property

Theorem: Given that P is Markov with respect to G (satisfies the local Markov assumption – every node is independent of its non-descendents conditioned on its parents in G), if X and Y are *d*-separated in G conditioned on Z, then X and Y are independent in P conditioned on Z.

$$X \coprod_G Y|Z \rightarrow X \ \coprod_P Y|Z$$

Exercise: Prove that the following are equivalent:

- Global Markov property
- Local Causal Markov condition
- Bayesian network factorization



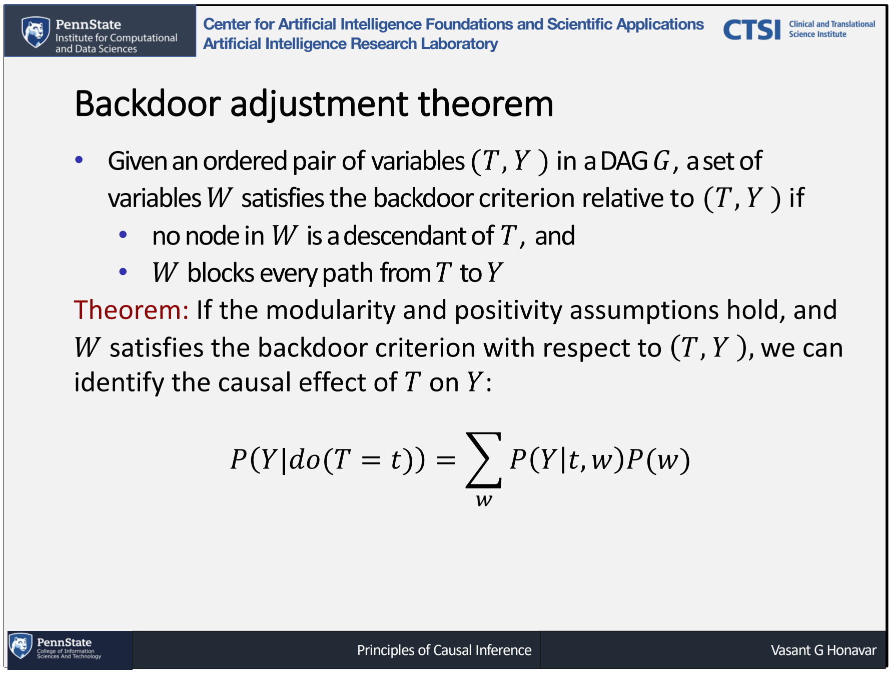
Truncated factorization

Original factorization

$$P(X_1 \cdots X_n) = \prod_{i=1}^{n} P(X_i | Parents(X_i))$$

Now if we intervene on a set of nodes with indices $S \subseteq [n]$ setting them to constants, then for all $i \in [n]$ we have:

- $P(X_1 \cdots X_n | do(S = s) = \prod_{i \notin S}^{n} P(X_i | Parents(X_i))$ if $X_1 \cdots X_n$ assume values consistent with the intervention
- $P(X_1 \cdots X_n | do(S = s) = 0$ otherwise

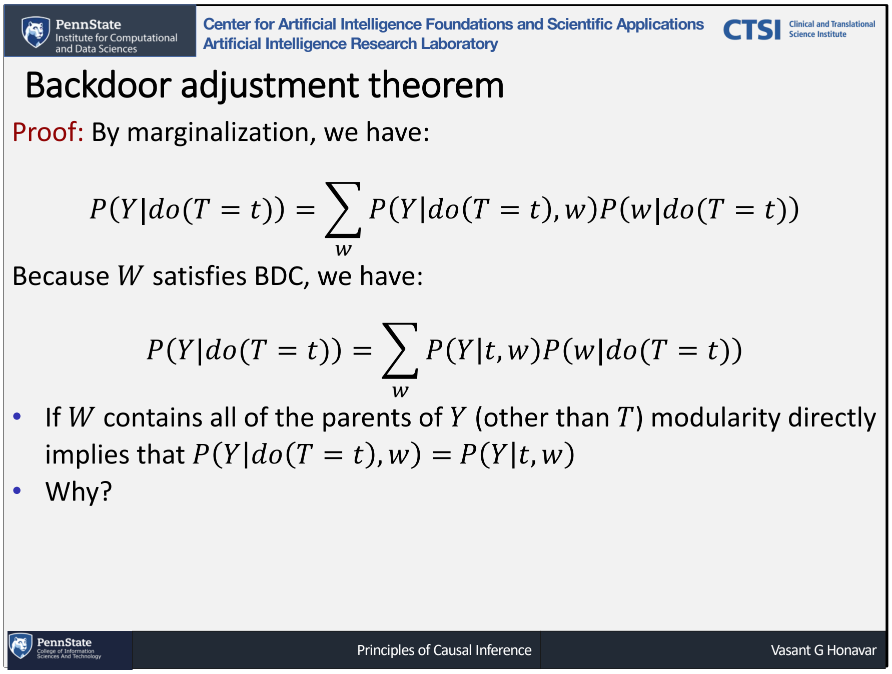


Backdoor adjustment theorem

- Given an ordered pair of variables (T, Y) in a DAG G, a set of variables W satisfies the backdoor criterion relative to (T, Y) if
  - no node in W is a descendant of T, and
  - W blocks every path from T to Y

Theorem: If the modularity and positivity assumptions hold, and W satisfies the backdoor criterion with respect to (T, Y), we can identify the causal effect of T on Y:

$$P(Y|do(T=t)) = \sum_{w} P(Y|t, w)P(w)$$



Backdoor adjustment theorem

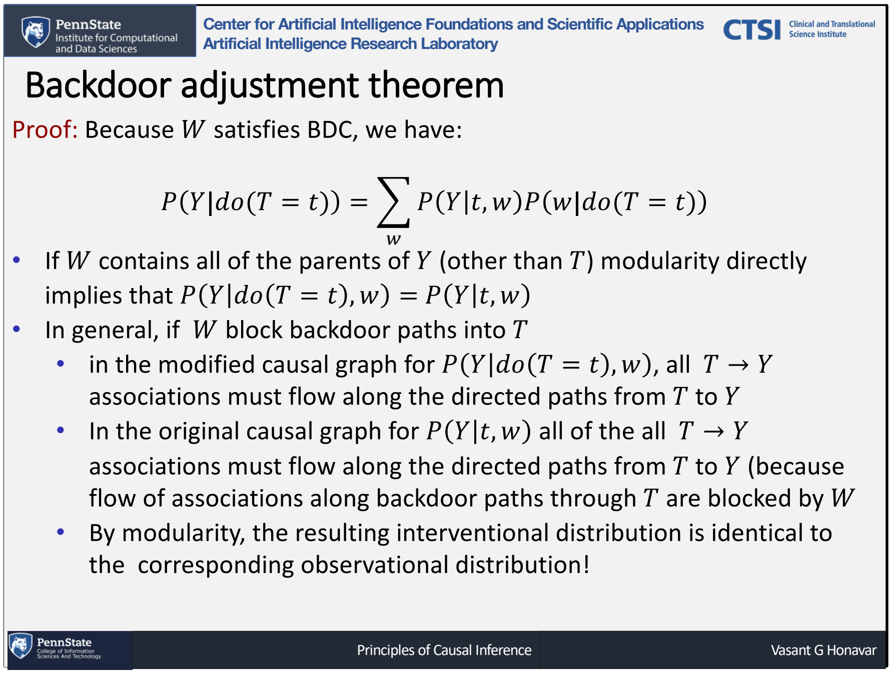
Proof: By marginalization, we have:

$$P(Y|do(T=t)) = \sum_{w} P(Y|do(T=t), w) P(w|do(T=t))$$

Because W satisfies BDC, we have:

$$P(Y|do(T=t)) = \sum_{w} P(Y|t, w) P(w|do(T=t))$$

- If W contains all of the parents of Y (other than T) modularity directly implies that $P(Y|do(T=t), w) = P(Y|t, w)$
- Why?

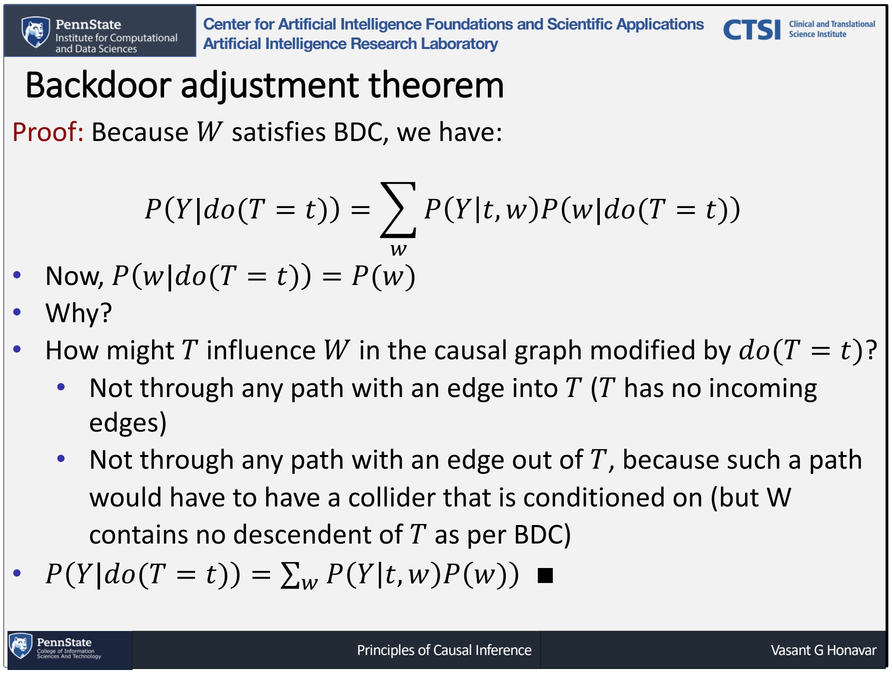


Backdoor adjustment theorem

Proof: Because W satisfies BDC, we have:

$$P(Y|do(T=t)) = \sum_{w} P(Y|t,w)P(w|do(T=t))$$

- If W contains all of the parents of Y (other than T) modularity directly implies that $P(Y|do(T=t),w) = P(Y|t,w)$
- In general, if W block backdoor paths into T
  - in the modified causal graph for $P(Y|do(T=t),w)$, all $T \rightarrow Y$ associations must flow along the directed paths from T to Y
  - In the original causal graph for $P(Y|t,w)$ all of the all $T \rightarrow Y$ associations must flow along the directed paths from T to Y (because flow of associations along backdoor paths through T are blocked by W
  - By modularity, the resulting interventional distribution is identical to the corresponding observational distribution!



Backdoor adjustment theorem

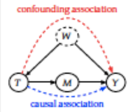
Proof: Because W satisfies BDC, we have:

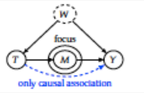
$$P(Y|do(T=t)) = \sum_{w} P(Y|t,w)P(w|do(T=t))$$

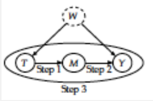
- Now, $P(w|do(T=t)) = P(w)$
- Why?
- How might T influence W in the causal graph modified by $do(T=t)$?
  - Not through any path with an edge into T (T has no incoming edges)
  - Not through any path with an edge out of T, because such a path would have to have a collider that is conditioned on (but W contains no descendent of T as per BDC)
- $P(Y|do(T=t)) = \sum_{w} P(Y|t,w)P(w))$ ■

Frontdoor Adjustment

- How can we identify the causal effect of T on Y in a causal graph even when we can't adjust for the confounder W because it is unobserved?
- If there is a mediator(s) like M along the causal path between T and Y, we can isolate the association that flows through M as the only causal association between T and Y (association flowing along directed paths from T to Y).

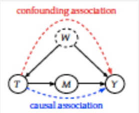




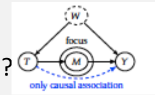


Frontdoor Adjustment

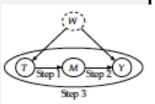
1. Identify the causal effect $P(M \mid do(T=t))$ of T on M. How?
   - Because Y is a collider f T that is not conditioned on, Y blocks backdoor paths into T
   - So using BDA, we have:
     $P(M \mid do(T=t)) = P(M|T=t)$



2. Identify the causal effect $P(Y \mid do(M=m))$ of M on Y. How?
   - Since T blocks the backdoor path into M, we can use BDA to adjust for T



$$P(Y \mid do(M=m)) = \sum_t P(Y|M=m, T=t)P(T=t)$$



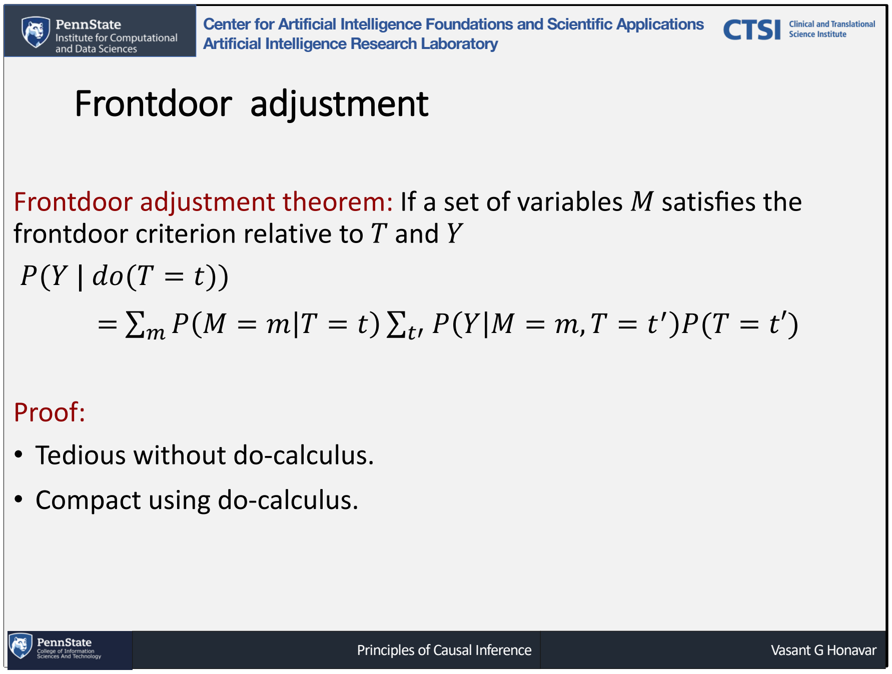
3. Combine the above steps to identify the causal effect of T on Y (through M):

$$P(Y \mid do(T=t)) = \sum_m P(M=m|do(T=t))P(Y|do(M=m)$$

$$= \sum_m P(M=m|T=t) \sum_{t'} P(Y|M=m, T=t')P(T=t')$$

Frontdoor criterion

- A set of variables M satisfies the frontdoor criterion relative to T and Y if:
    - M completely mediates the effect of T on Y (all causal paths from T to Y go through M).
    - There is no unblocked backdoor path from T to M.
    - All backdoor paths from M to Y are blocked by T



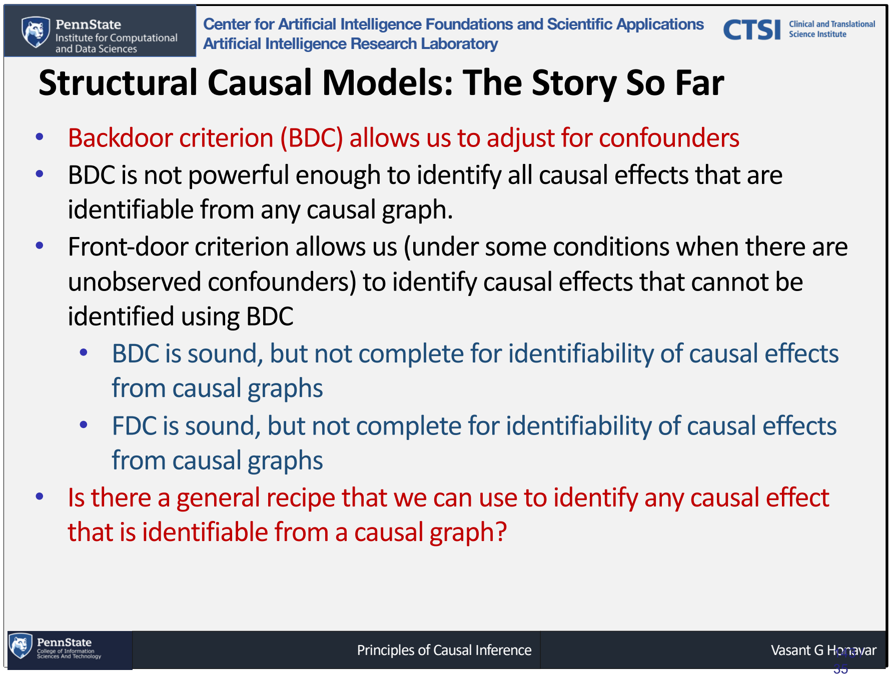
Frontdoor adjustment

Frontdoor adjustment theorem: If a set of variables M satisfies the frontdoor criterion relative to T and Y

$$P(Y \mid do(T = t))$$
$$= \sum_m P(M = m|T = t) \sum_{t'} P(Y|M = m, T = t')P(T = t')$$

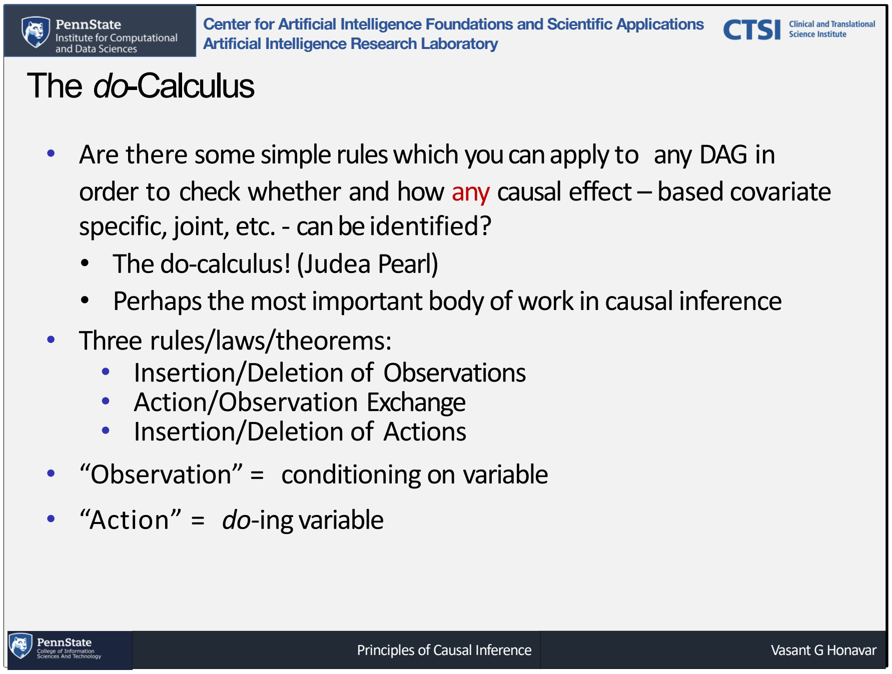
Proof:

- Tedious without do-calculus.
- Compact using do-calculus.



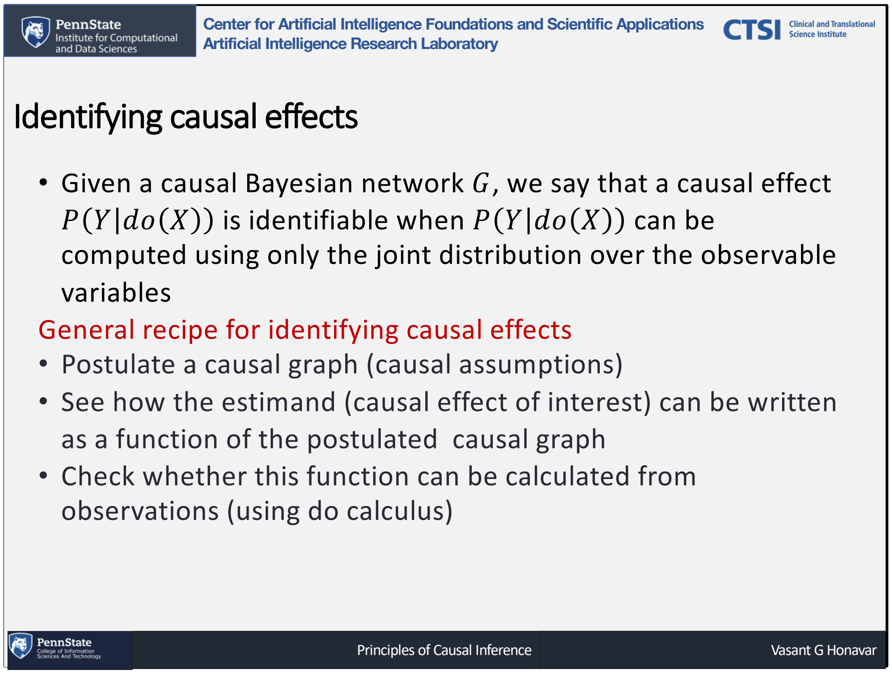
Structural Causal Models: The Story So Far

- Backdoor criterion (BDC) allows us to adjust for confounders
- BDC is not powerful enough to identify all causal effects that are identifiable from any causal graph.
- Front-door criterion allows us (under some conditions when there are unobserved confounders) to identify causal effects that cannot be identified using BDC
  - BDC is sound, but not complete for identifiability of causal effects from causal graphs
  - FDC is sound, but not complete for identifiability of causal effects from causal graphs
- Is there a general recipe that we can use to identify any causal effect that is identifiable from a causal graph?



The *do*-Calculus

- Are there some simple rules which you can apply to any DAG in order to check whether and how any causal effect – based covariate specific, joint, etc. - can be identified?
  - The do-calculus! (Judea Pearl)
  - Perhaps the most important body of work in causal inference
- Three rules/laws/theorems:
  - Insertion/Deletion of Observations
  - Action/Observation Exchange
  - Insertion/Deletion of Actions
- "Observation" = conditioning on variable
- "Action" = *do*-ing variable



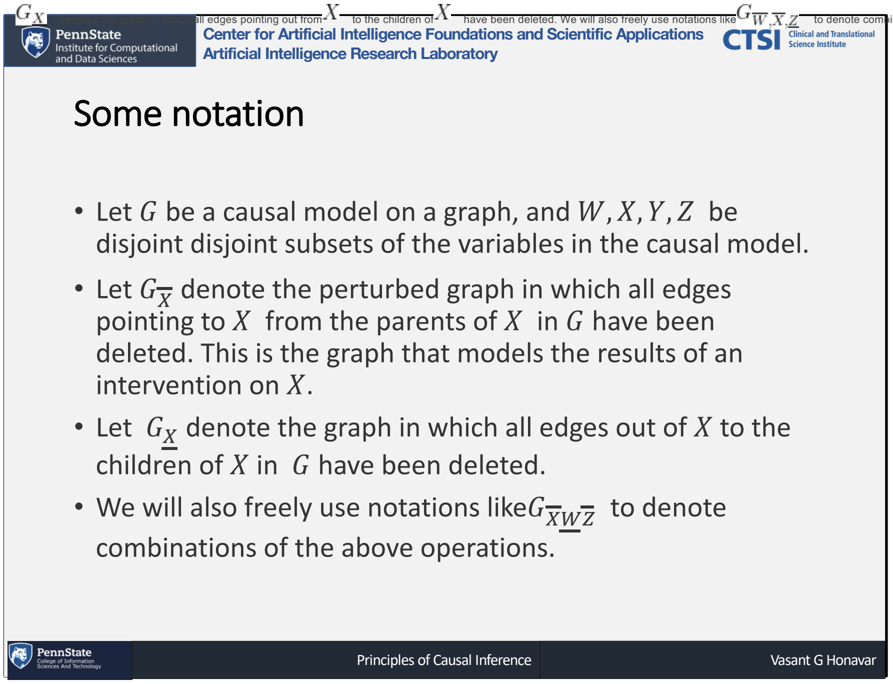
Identifying causal effects

- Given a causal Bayesian network G, we say that a causal effect $P(Y|do(X))$ is identifiable when $P(Y|do(X))$ can be computed using only the joint distribution over the observable variables

General recipe for identifying causal effects

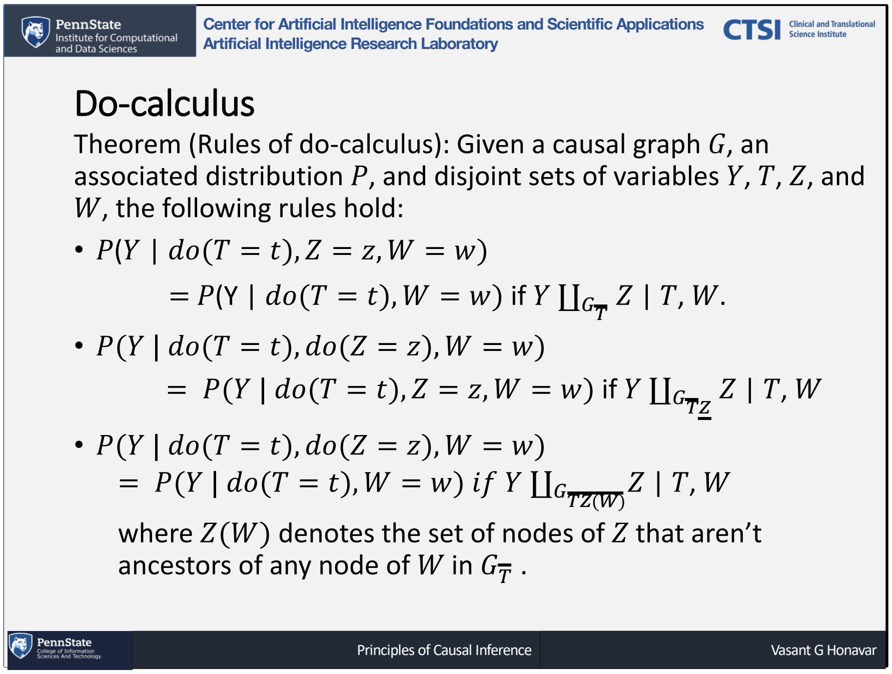
- Postulate a causal graph (causal assumptions)
- See how the estimand (causal effect of interest) can be written as a function of the postulated causal graph
- Check whether this function can be calculated from observations (using do calculus)

G_X denotes the graph in which all edges pointing out from X to the children of X have been deleted. We will also freely use notations like $G_{\overline{W},\overline{X},\underline{Z}}$ to denote combinations of these operations.



Some notation

- Let G be a causal model on a graph, and W, X, Y, Z be disjoint disjoint subsets of the variables in the causal model.
- Let $G_{\overline{X}}$ denote the perturbed graph in which all edges pointing to X from the parents of X in G have been deleted. This is the graph that models the results of an intervention on X.
- Let $G_{\underline{X}}$ denote the graph in which all edges out of X to the children of X in G have been deleted.
- We will also freely use notations like$G_{\overline{X}\underline{W}\overline{Z}}$ to denote combinations of the above operations.

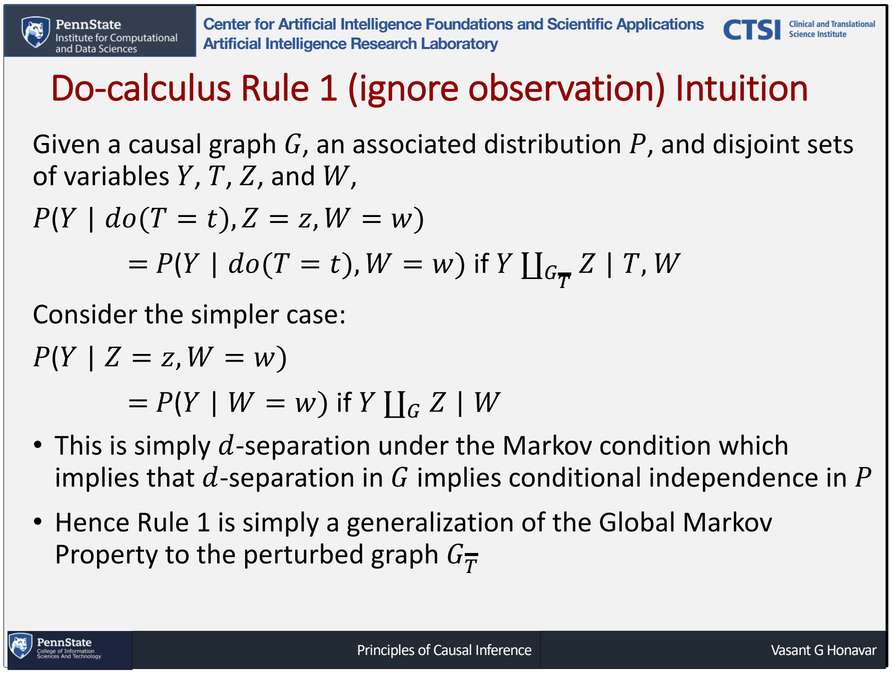


Do-calculus

Theorem (Rules of do-calculus): Given a causal graph G, an associated distribution P, and disjoint sets of variables Y, T, Z, and W, the following rules hold:

- $P(Y \mid do(T = t), Z = z, W = w)$
  $= P(\mathrm{Y} \mid do(T = t), W = w)$ if $Y \coprod_{G_{\overline{T}}} Z \mid T, W$.
- $P(Y \mid do(T = t), do(Z = z), W = w)$
  $= \; P(Y \mid do(T = t), Z = z, W = w)$ if $Y \coprod_{G_{\overline{T}\underline{Z}}} Z \mid T, W$
- $P(Y \mid do(T = t), do(Z = z), W = w)$
  $= \; P(Y \mid do(T = t), W = w) \; if \; Y \coprod_{G_{\overline{TZ(W)}}} Z \mid T, W$

  where $Z(W)$ denotes the set of nodes of Z that aren't ancestors of any node of W in $G_{\overline{T}}$.



Do-calculus Rule 1 (ignore observation) Intuition

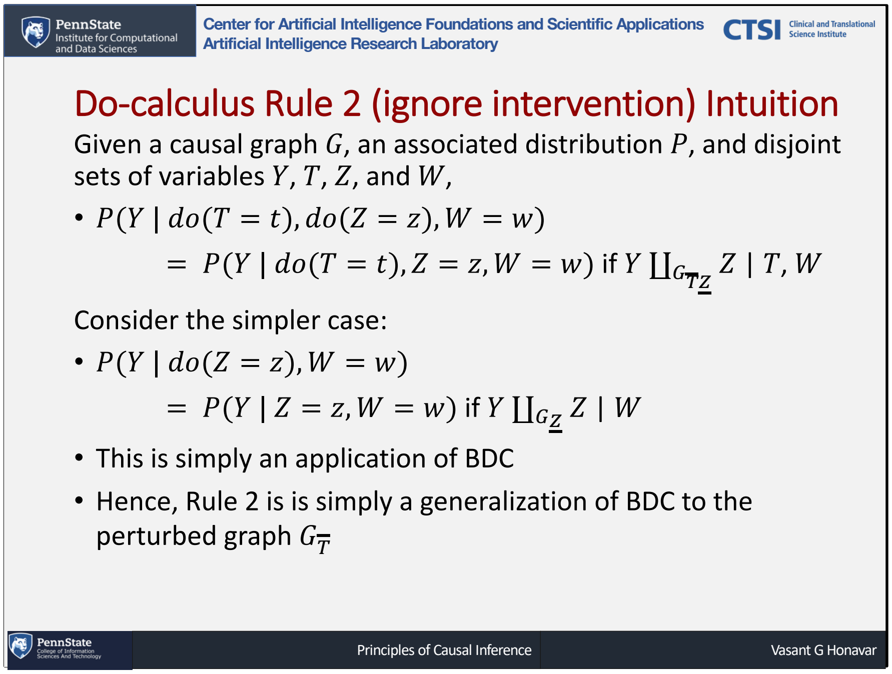
Given a causal graph G, an associated distribution P, and disjoint sets of variables Y, T, Z, and W,

$$P(Y \mid do(T=t), Z=z, W=w)$$
$$= P(Y \mid do(T=t), W=w) \text{ if } Y \coprod_{G_{\overline{T}}} Z \mid T, W$$

Consider the simpler case:

$$P(Y \mid Z=z, W=w)$$
$$= P(Y \mid W=w) \text{ if } Y \coprod_{G} Z \mid W$$

- This is simply d-separation under the Markov condition which implies that d-separation in G implies conditional independence in P
- Hence Rule 1 is simply a generalization of the Global Markov Property to the perturbed graph $G_{\overline{T}}$



Do-calculus Rule 2 (ignore intervention) Intuition

Given a causal graph G, an associated distribution P, and disjoint sets of variables Y, T, Z, and W,

- $P(Y \mid do(T = t), do(Z = z), W = w)$

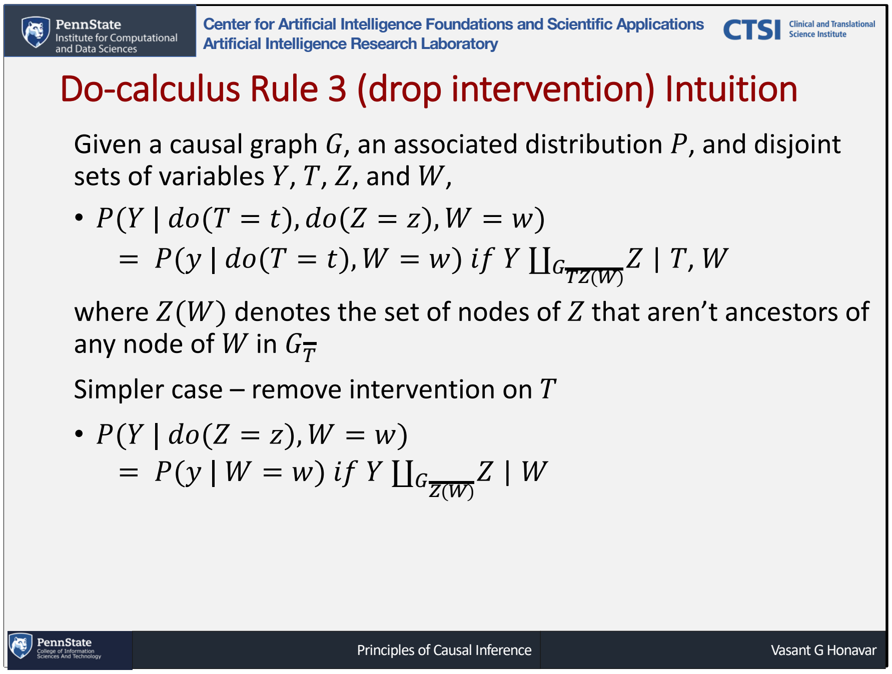
  $= \ P(Y \mid do(T = t), Z = z, W = w)$ if $Y \amalg_{G_{\overline{T}\underline{Z}}} Z \mid T, W$

Consider the simpler case:

- $P(Y \mid do(Z = z), W = w)$

  $= \ P(Y \mid Z = z, W = w)$ if $Y \amalg_{G_{\underline{Z}}} Z \mid W$

- This is simply an application of BDC
- Hence, Rule 2 is is simply a generalization of BDC to the perturbed graph $G_{\overline{T}}$



Do-calculus Rule 3 (drop intervention) Intuition

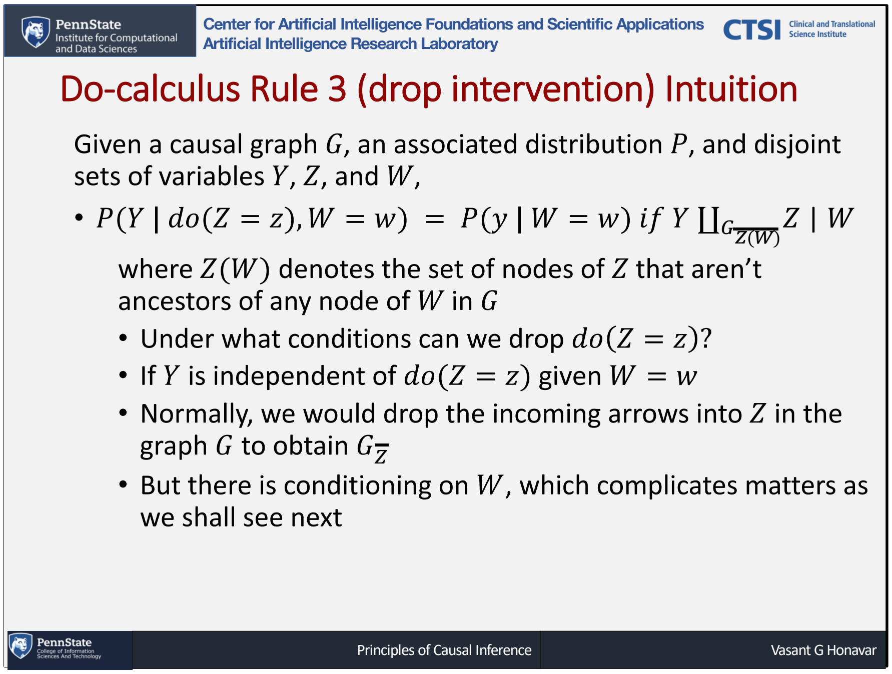
Given a causal graph G, an associated distribution P, and disjoint sets of variables Y, T, Z, and W,

- $P(Y \mid do(T=t), do(Z=z), W=w)$
  $= P(y \mid do(T=t), W=w) \; if \; Y \coprod_{G_{\overline{TZ(W)}}} Z \mid T, W$

where $Z(W)$ denotes the set of nodes of Z that aren't ancestors of any node of W in $G_{\overline{T}}$

Simpler case – remove intervention on T

- $P(Y \mid do(Z=z), W=w)$
  $= P(y \mid W=w) \; if \; Y \coprod_{G_{\overline{Z(W)}}} Z \mid W$



Do-calculus Rule 3 (drop intervention) Intuition

Given a causal graph G, an associated distribution P, and disjoint sets of variables Y, Z, and W,

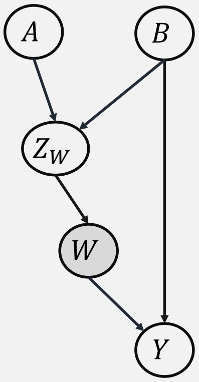
- $P(Y \mid do(Z = z), W = w) \; = \; P(y \mid W = w) \; if \; Y \coprod_{G_{\overline{Z(W)}}} Z \mid W$

  where $Z(W)$ denotes the set of nodes of Z that aren't ancestors of any node of W in G

  - Under what conditions can we drop $do(Z = z)$?
  - If Y is independent of $do(Z = z)$ given $W = w$
  - Normally, we would drop the incoming arrows into Z in the graph G to obtain $G_{\overline{Z}}$
  - But there is conditioning on W, which complicates matters as we shall see next

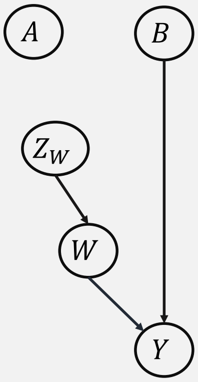
Do-calculus Rule 3 (drop intervention) Intuition

- Suppose Z_W is a node in Z that is an ancestor of a node in W
- When we condition on W, A, Z_W and become *d*-connected in G!
- Consequently, we cannot drop the conditioning on $do(Z = z)$ because $Z_W \in Z$!



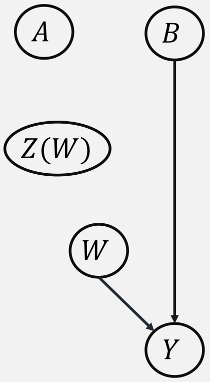
Do-calculus Rule 3 (drop intervention) Intuition

- If Z_W is a node in Z that is an ancestor of a node in W, when we condition on W, A, Z_W and B become *d*-connected in G!
- If we drop the edges into Z, resulting in $G_{\overline{Z}}$, Z will still affect the distribution of Y
- We cannot drop interventions on nodes in Z that are ancestors of any node in W.
- $Y \coprod_{G_{\overline{Z}}} Z \mid W$ does not guarantee that $P(Y \mid do(Z = z), W = w) = P(y \mid W = w)$
- When can we drop $do(Z = z)$ and not affect the distribution of Y?
- Only when we exclude nodes in Z that are ancestors of any node in W



Do-calculus Rule 3 (drop intervention) Intuition

- When can we drop $do(Z = z)$ and not affect the distribution of Y?
- When d-separation between Z and Y holds in the manipulated graph resulting from dropping arrows into nodes in Z that are NOT ancestors of any node in W.



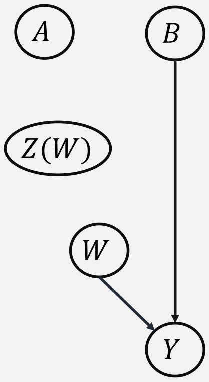
Do-calculus Rule 3 (drop intervention) Intuition

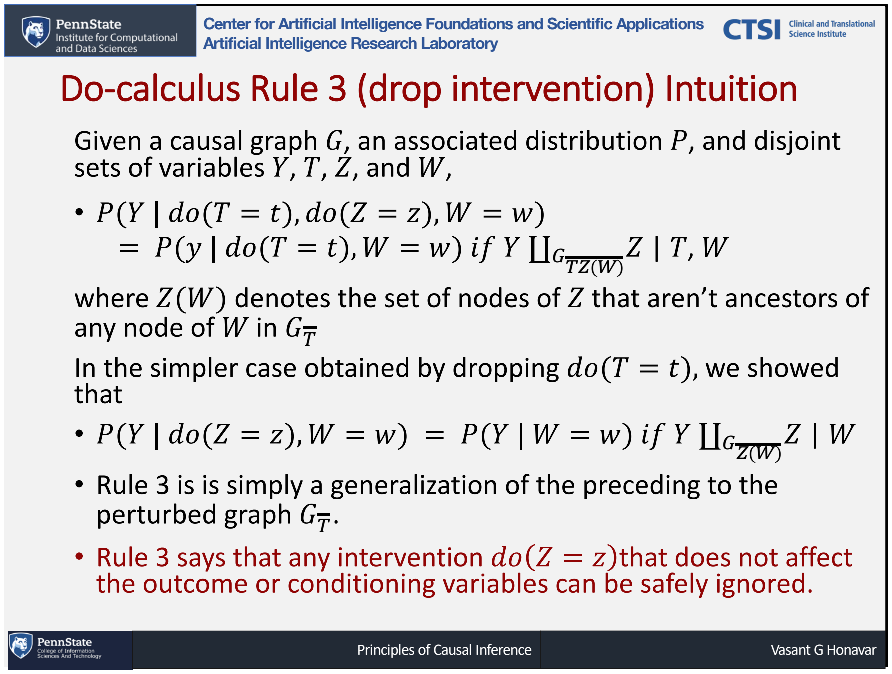
Given a causal graph G, an associated distribution P, and disjoint sets of variables Y, Z, and W,

$P(Y \mid do(Z = z), W = w)\ =\ P(y \mid W = w)\ if\ Y \amalg_{G_{\overline{Z(W)}}} Z \mid W$

where $Z(W)$ denotes the set of nodes of Z that aren't ancestors of any node of W in G

- We can drop $do(Z = z)$ and not affect the distribution of Y when Z are d-separated from Y in the manipulated graph $G_{\overline{Z(W)}}$.
- That is, we drop the interventions $do(Z = z)$ only when doing so does not causally impact Y when conditioning on (controlling for) W!





Do-calculus Rule 3 (drop intervention) Intuition

Given a causal graph G, an associated distribution P, and disjoint sets of variables Y, T, Z, and W,

- $P(Y \mid do(T = t), do(Z = z), W = w)$
  $= \ P(y \mid do(T = t), W = w) \ if \ Y \coprod_{G_{\overline{TZ(W)}}} Z \mid T, W$

where $Z(W)$ denotes the set of nodes of Z that aren't ancestors of any node of W in $G_{\overline{T}}$

In the simpler case obtained by dropping $do(T = t)$, we showed that

- $P(Y \mid do(Z = z), W = w) \ = \ P(Y \mid W = w) \ if \ Y \coprod_{G_{\overline{Z(W)}}} Z \mid W$
- Rule 3 is is simply a generalization of the preceding to the perturbed graph $G_{\overline{T}}$.
- Rule 3 says that any intervention $do(Z = z)$that does not affect the outcome or conditioning variables can be safely ignored.

Do-calculus Rule 3 (drop intervention) Intuition

Given a causal graph G, an associated distribution P, and disjoint sets of variables Y, T, Z, and W,

- $P(Y \mid do(T=t), do(Z=z), W=w)$
  $= P(y \mid do(T=t), W=w) \; if \; Y \perp\!\!\!\perp_{G_{\overline{TZ(W)}}} Z \mid T, W$

where $Z(W)$ denotes the set of nodes of Z that aren't ancestors of any node of W in $G_{\overline{T}}$

Rule 3 says that any intervention that does not affect the outcome or conditioning variables can be safely ignored.

Do-calculus

Theorem (Rules of do-calculus): Given a causal graph G, an associated distribution P, and disjoint sets of variables Y, T, Z, and W, the following rules hold:

- $P(Y \mid do(T = t), Z = z, W = w)$

  $= P(\mathrm{Y} \mid do(T = t, W = w)$ if $Y \coprod_{G_{\overline{T}}} Z \mid T, W.$

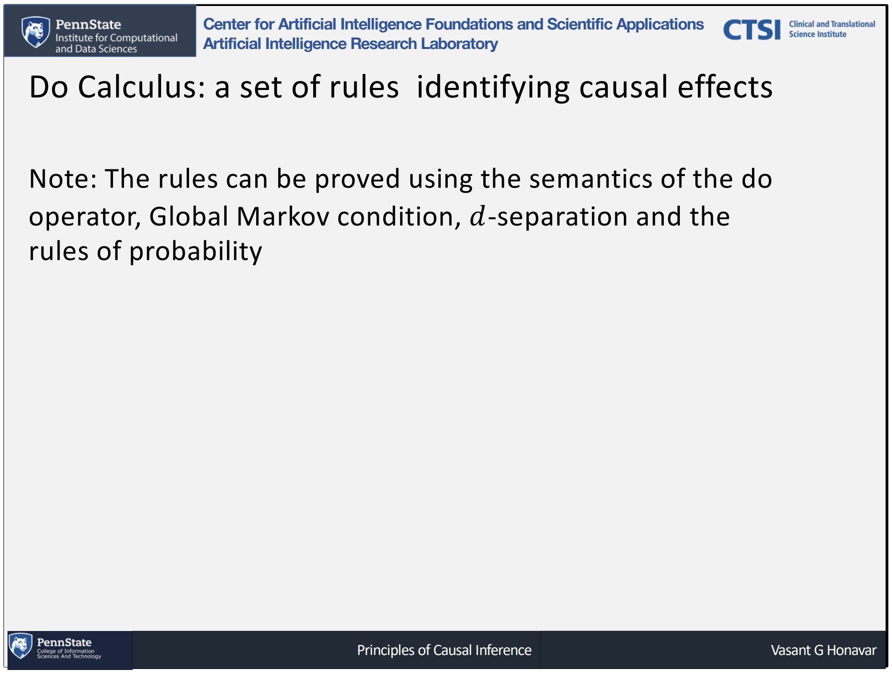
- $P(Y \mid do(T = t), do(Z = z), W = w)$

  $= \ P(Y \mid do(T = t), Z = z, W = w)$ if $Y \coprod_{G_{\overline{T}\underline{Z}}} Z \mid T, W$

- $P(Y \mid do(T = t), do(Z = z), W = w)$

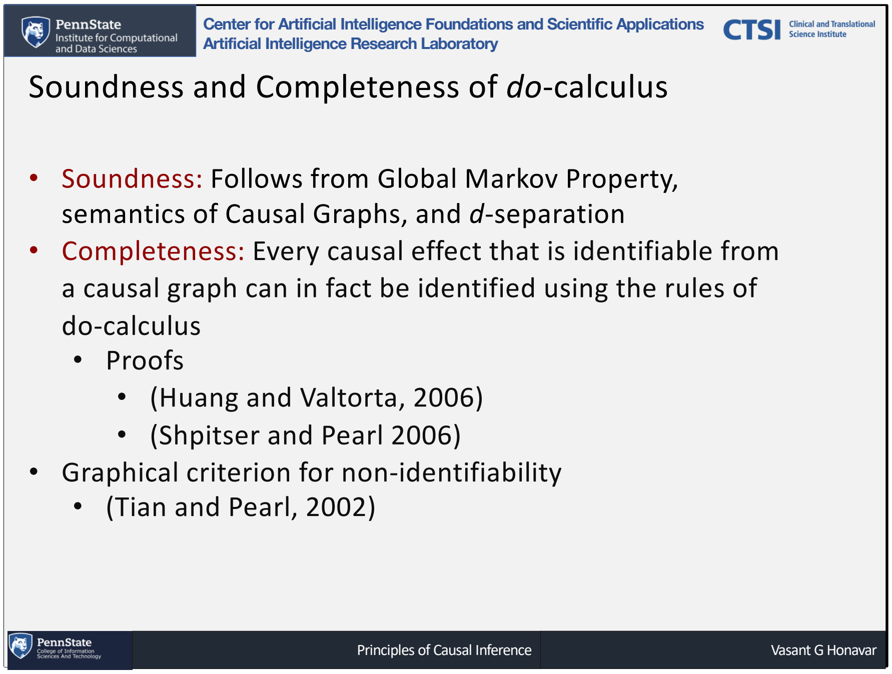
  $= \ P(y \mid do(T = t), W = w) \ if \ Y \coprod_{G_{\overline{TZ(W)}}} Z \mid T, W$

  where $Z(W)$ denotes the set of nodes of Z that aren't ancestors of any node of W in $G_{\overline{T}}$.



Do Calculus: a set of rules identifying causal effects

Note: The rules can be proved using the semantics of the do operator, Global Markov condition, d-separation and the rules of probability



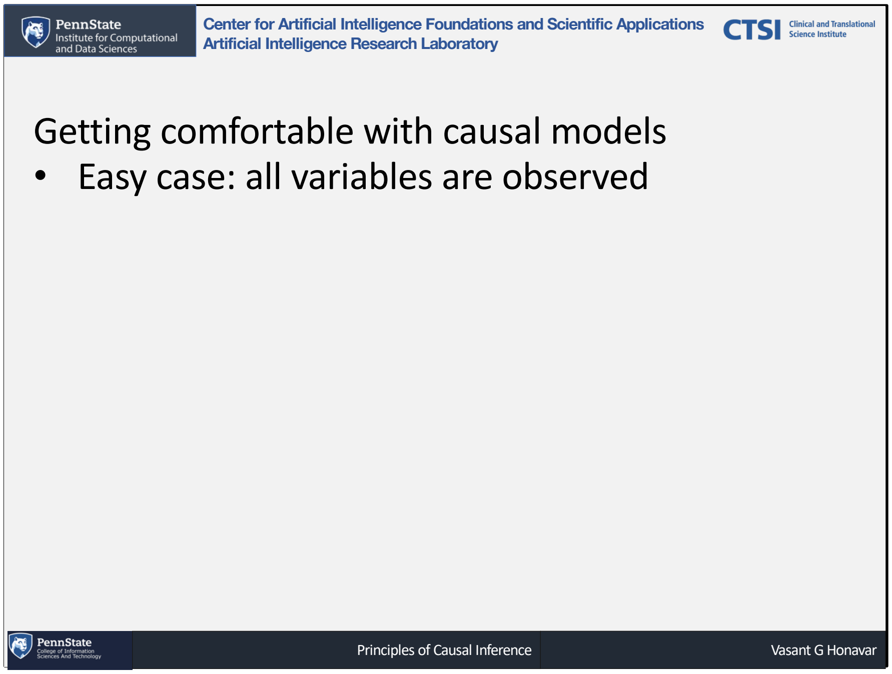
Soundness and Completeness of *do*-calculus

- Soundness: Follows from Global Markov Property, semantics of Causal Graphs, and *d*-separation
- Completeness: Every causal effect that is identifiable from a causal graph can in fact be identified using the rules of do-calculus
  - Proofs
    - (Huang and Valtorta, 2006)
    - (Shpitser and Pearl 2006)
- Graphical criterion for non-identifiability
  - (Tian and Pearl, 2002)

Exercises

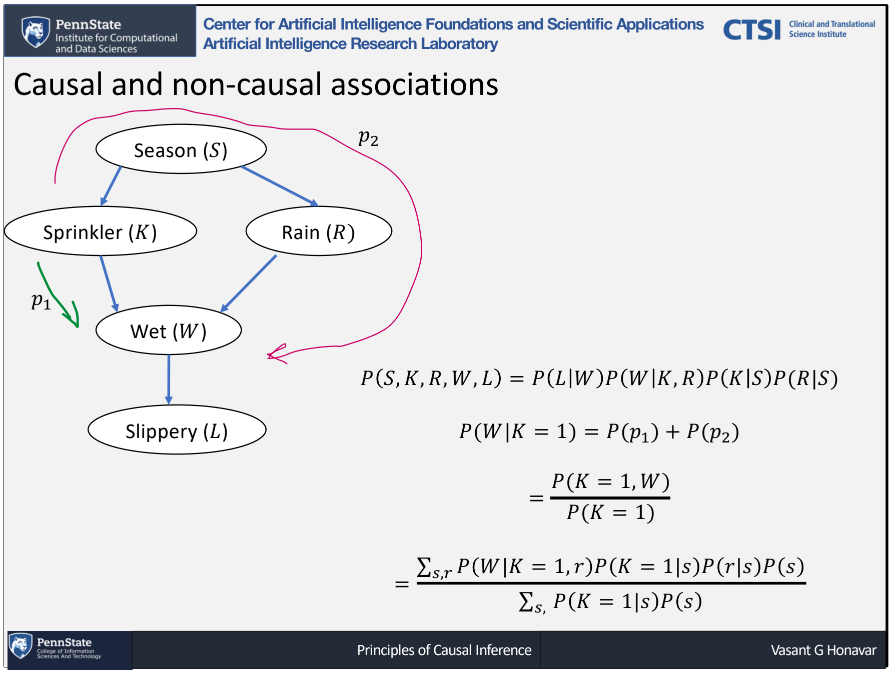
- Use do-calculus to prove
  - Backdoor adjustment formula
  - Frontdoor adjustment formula

Getting comfortable with causal models



Getting comfortable with causal models

- Easy case: all variables are observed



Causal and non-causal associations

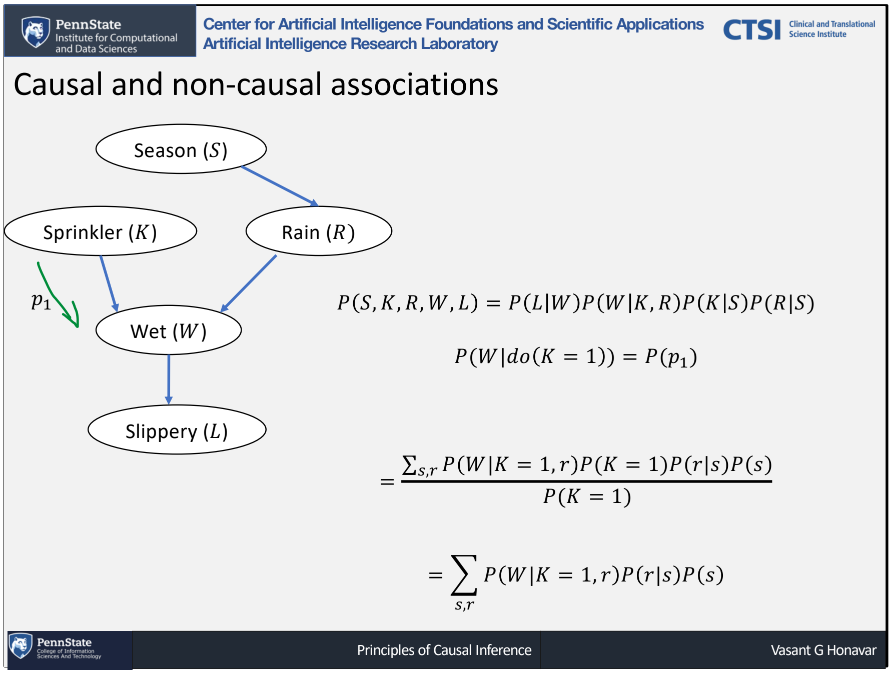
$$P(S,K,R,W,L) = P(L|W)P(W|K,R)P(K|S)P(R|S)$$

$$P(W|K=1) = P(p_1) + P(p_2)$$

$$= \frac{P(K=1,W)}{P(K=1)}$$

$$= \frac{\sum_{s,r} P(W|K=1,r)P(K=1|s)P(r|s)P(s)}{\sum_{s,} P(K=1|s)P(s)}$$

Type equation here.



Causal and non-causal associations

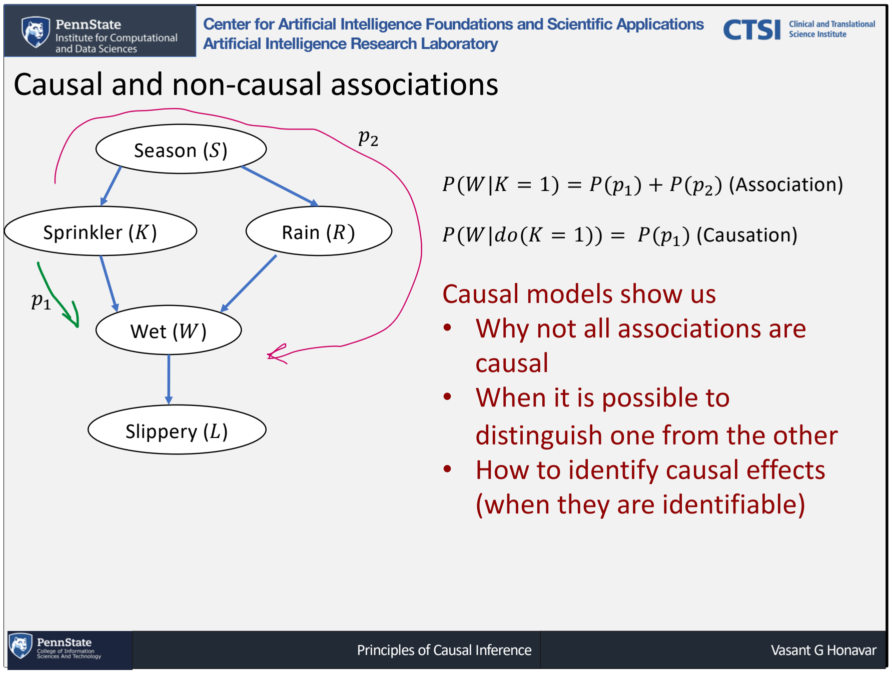
$$P(S,K,R,W,L) = P(L|W)P(W|K,R)P(K|S)P(R|S)$$

$$P(W|do(K=1)) = P(p_1)$$

$$= \frac{\sum_{s,r} P(W|K=1,r)P(K=1)P(r|s)P(s)}{P(K=1)}$$

$$= \sum_{s,r} P(W|K=1,r)P(r|s)P(s)$$

Type equation here.



Causal and non-causal associations

$P(W|K = 1) = P(p_1) + P(p_2)$ (Association)

$P(W|do(K = 1)) = P(p_1)$ (Causation)

Causal models show us

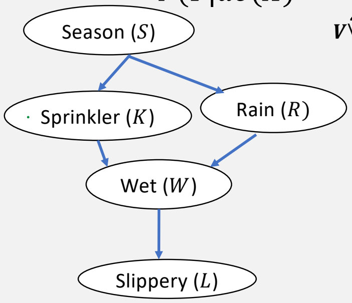
- Why not all associations are causal
- When it is possible to distinguish one from the other
- How to identify causal effects (when they are identifiable)

Type equation here.

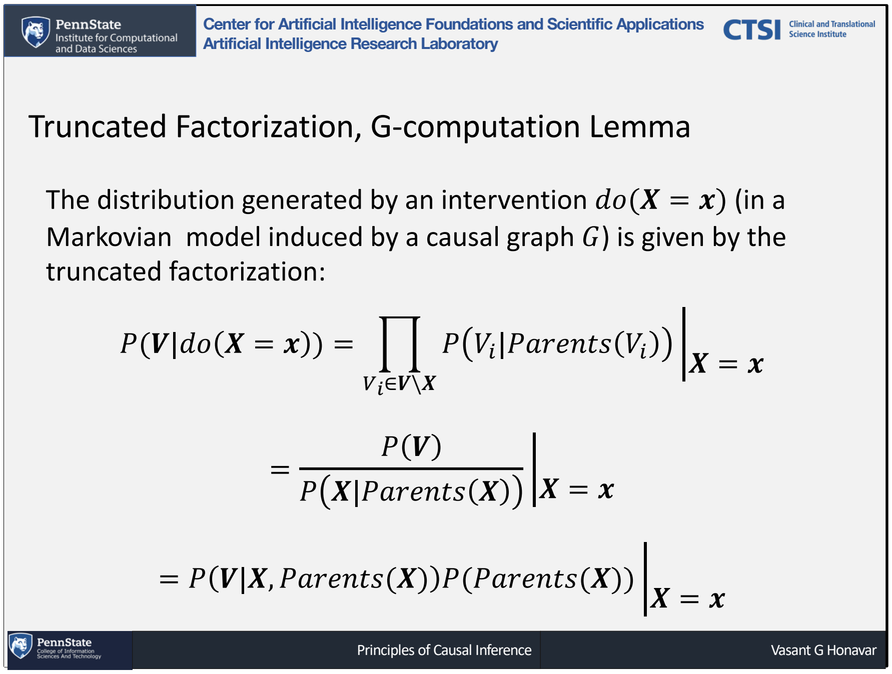
Identification from (Markovian) Causal Models

Theorem: Given the causal diagram *G* of any Markovian model that all variables $\boldsymbol{V}$ are measured, the causal effect $Q \ = \ P(\boldsymbol{Y} \mid do(\boldsymbol{X}))$ is identifiable for every subsets of variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ and is obtained from the truncated factorization, i.e.,

$$P(\boldsymbol{Y}|do(\boldsymbol{X})) = \sum_{\boldsymbol{V}\setminus\boldsymbol{X}\cup\boldsymbol{Y}} \prod_{V_i \in \boldsymbol{V}\setminus\boldsymbol{X}} P\big(V_i|Parents(V_i)\big)$$



$$P(W|do(K=1)) = \sum_{s,r} P(W|K=1,r)P(r|s)P(s)$$



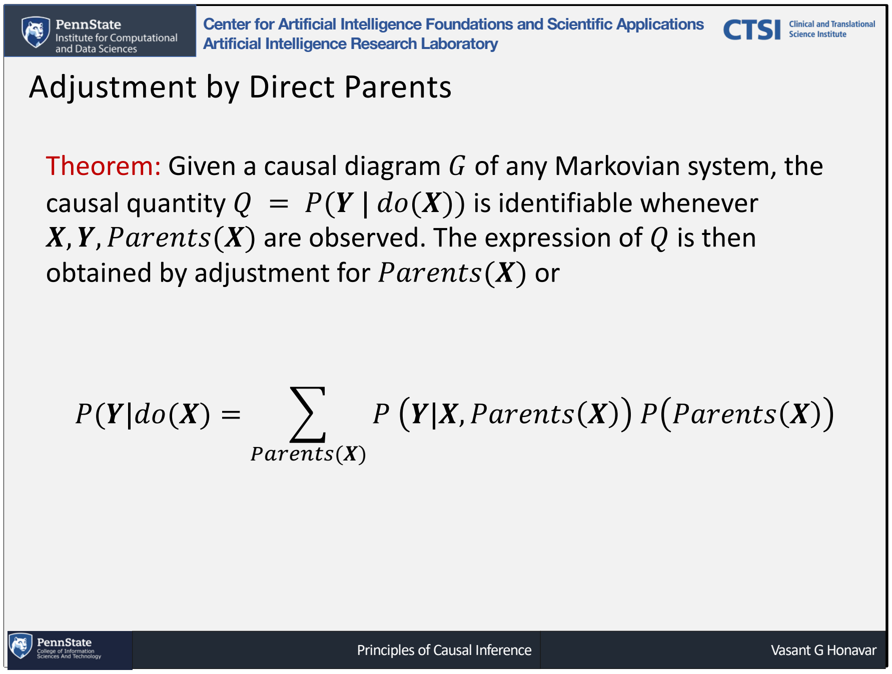
Truncated Factorization, G-computation Lemma

The distribution generated by an intervention $do(\boldsymbol{X} = \boldsymbol{x})$ (in a Markovian model induced by a causal graph G) is given by the truncated factorization:

$$P(\boldsymbol{V}|do(\boldsymbol{X} = \boldsymbol{x})) = \prod_{V_i \in \boldsymbol{V} \setminus \boldsymbol{X}} P\big(V_i|Parents(V_i)\big)\Big|_{\boldsymbol{X} = \boldsymbol{x}}$$

$$= \frac{P(\boldsymbol{V})}{P\big(\boldsymbol{X}|Parents(\boldsymbol{X})\big)}\Bigg|_{\boldsymbol{X} = \boldsymbol{x}}$$

$$= P(\boldsymbol{V}|\boldsymbol{X}, Parents(\boldsymbol{X}))P(Parents(\boldsymbol{X}))\Big|_{\boldsymbol{X} = \boldsymbol{x}}$$



Adjustment by Direct Parents

Theorem: Given a causal diagram G of any Markovian system, the causal quantity $Q \;=\; P(\boldsymbol{Y} \mid do(\boldsymbol{X}))$ is identifiable whenever $\boldsymbol{X}, \boldsymbol{Y}, Parents(\boldsymbol{X})$ are observed. The expression of Q is then obtained by adjustment for $Parents(\boldsymbol{X})$ or

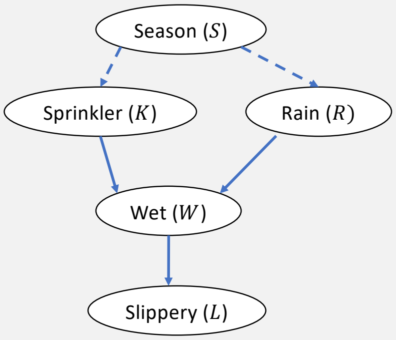
$$P(\boldsymbol{Y}|do(\boldsymbol{X}) = \sum_{Parents(\boldsymbol{X})} P\left(\boldsymbol{Y}|\boldsymbol{X}, Parents(\boldsymbol{X})\right) P\left(Parents(\boldsymbol{X})\right)$$

What if not all direct parents of treatment are observed?

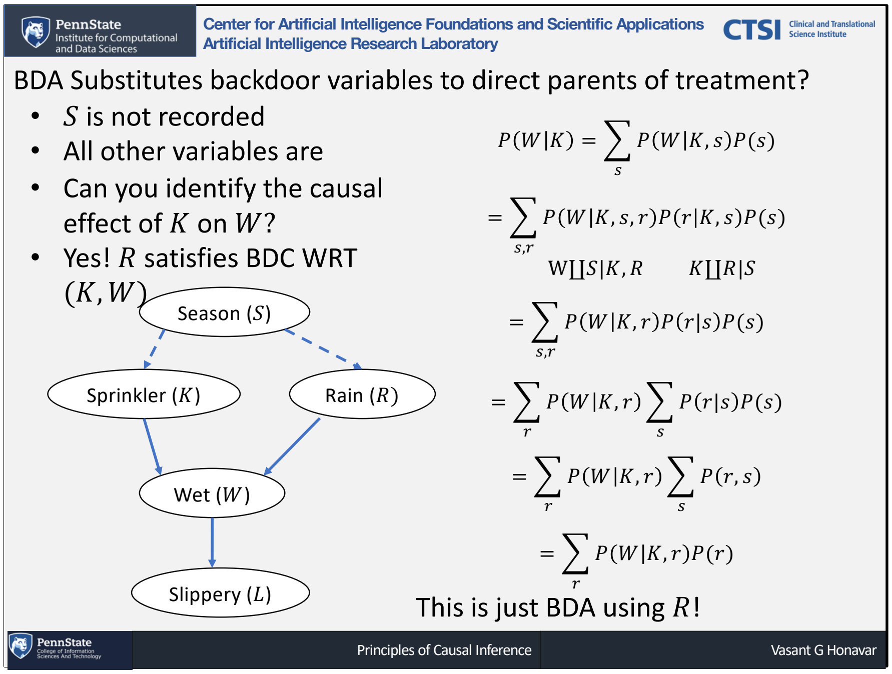
Can we still identify the causal effect of treatment?

Yes, e.g., using BDA, FDA, and more generally, do-calculus!

- BDA can use backdoor variables that substitute for direct parents of treatment



- S is not recorded
- All other variables are
- Can you identify the causal effect of K on W?



BDA Substitutes backdoor variables to direct parents of treatment?

- S is not recorded
- All other variables are
- Can you identify the causal effect of K on W?
- Yes! R satisfies BDC WRT (K, W)

$$P(W|K) = \sum_{s} P(W|K,s)P(s)$$

$$= \sum_{s,r} P(W|K,s,r)P(r|K,s)P(s)$$

$W \amalg S|K,R \qquad K \amalg R|S$

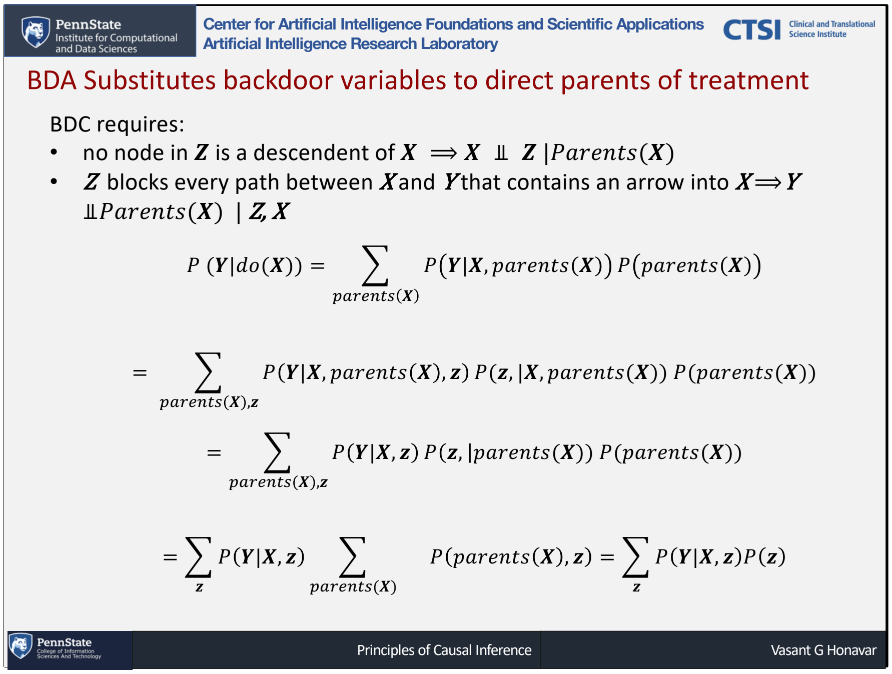
$$= \sum_{s,r} P(W|K,r)P(r|s)P(s)$$

$$= \sum_{r} P(W|K,r) \sum_{s} P(r|s)P(s)$$

$$= \sum_{r} P(W|K,r) \sum_{s} P(r,s)$$

$$= \sum_{r} P(W|K,r)P(r)$$

This is just BDA using R!



BDA Substitutes backdoor variables to direct parents of treatment

BDC requires:

- no node in $\boldsymbol{Z}$ is a descendent of $\boldsymbol{X}$ $\Longrightarrow \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Z} \,|Parents(\boldsymbol{X})$
- $\boldsymbol{Z}$ blocks every path between $\boldsymbol{X}$and $\boldsymbol{Y}$that contains an arrow into $\boldsymbol{X} \Longrightarrow \boldsymbol{Y}$ $\perp\!\!\!\perp Parents(\boldsymbol{X}) \ | \boldsymbol{Z}, \boldsymbol{X}$

$$P\ (\boldsymbol{Y}|do(\boldsymbol{X})) = \sum_{parents(\boldsymbol{X})} P\big(\boldsymbol{Y}|\boldsymbol{X}, parents(\boldsymbol{X})\big)\, P\big(parents(\boldsymbol{X})\big)$$

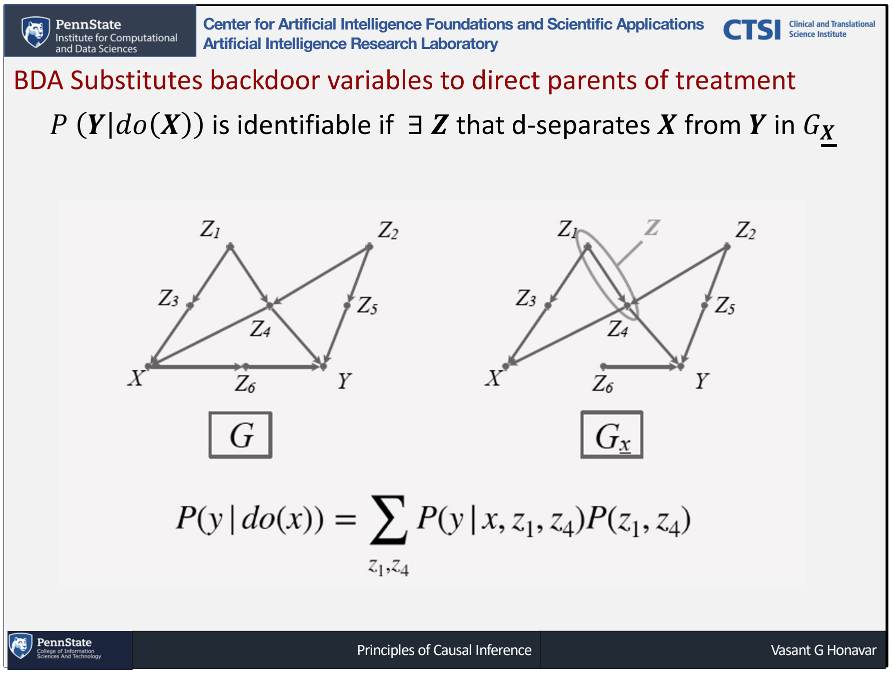
$$= \sum_{parents(\boldsymbol{X}),\boldsymbol{z}} P(\boldsymbol{Y}|\boldsymbol{X}, parents(\boldsymbol{X}), \boldsymbol{z})\, P(\boldsymbol{z}, |\boldsymbol{X}, parents(\boldsymbol{X}))\, P(parents(\boldsymbol{X}))$$

$$= \sum_{parents(\boldsymbol{X}),\boldsymbol{z}} P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{z})\, P(\boldsymbol{z}, |parents(\boldsymbol{X}))\, P(parents(\boldsymbol{X}))$$

$$= \sum_{\boldsymbol{z}} P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{z}) \sum_{parents(\boldsymbol{X})} P(parents(\boldsymbol{X}), \boldsymbol{z}) = \sum_{\boldsymbol{z}} P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{z}) P(\boldsymbol{z})$$

Then:

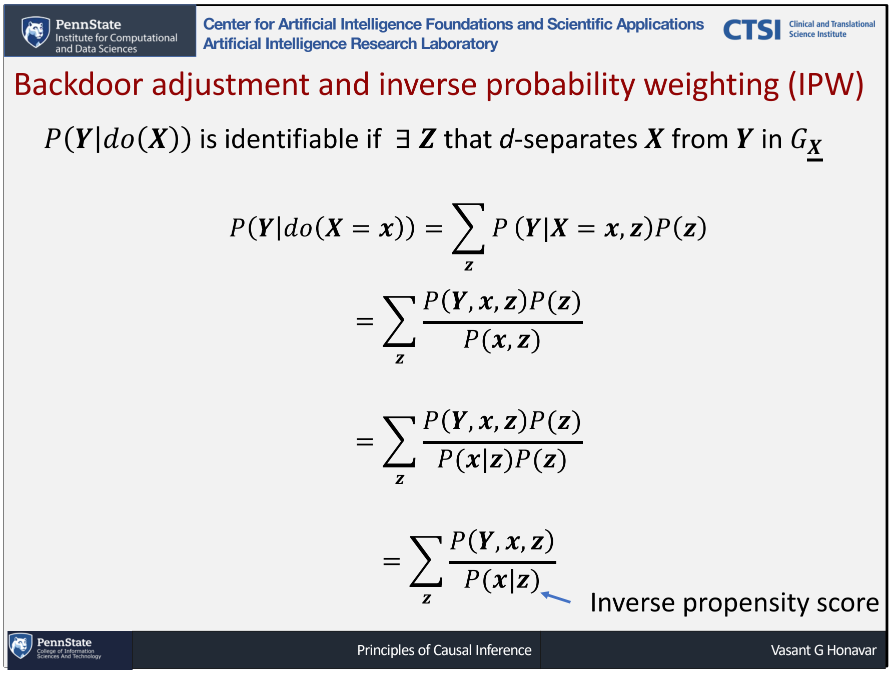
$$\begin{aligned} P(\mathbf{y} \,|\, do(\mathbf{x})) &= \sum_{pa_{\mathbf{x}}} P(\mathbf{y} \,|\, \mathbf{x}, pa_{\mathbf{x}}) P(pa_{\mathbf{x}}) \\ &= \sum_{\mathbf{z}, pa_{\mathbf{x}}} P(\mathbf{y} \,|\, \mathbf{x}, pa_{\mathbf{x}}, \mathbf{z}) P(\mathbf{z} \,|\, \mathbf{x}, pa_{\mathbf{x}}) P(pa_{\mathbf{x}}) \\ &= \sum_{\mathbf{z}, pa_{\mathbf{x}}} P(\mathbf{y} \,|\, \mathbf{x}, \mathbf{z}) P(\mathbf{z} \,|\, pa_{\mathbf{x}}) P(pa_{\mathbf{x}}) \\ &= \sum_{\mathbf{z}} P(\mathbf{y} \,|\, \mathbf{x}, \mathbf{z}) \sum_{pa_{\mathbf{x}}} P(\mathbf{z}, pa_{\mathbf{x}}) = \boxed{\sum_{\mathbf{z}} P(\mathbf{y} \,|\, \mathbf{x}, \mathbf{z}) P(\mathbf{z})} \end{aligned}$$



BDA Substitutes backdoor variables to direct parents of treatment

$P\,(\boldsymbol{Y}|do(\boldsymbol{X}))$ is identifiable if $\exists\,\boldsymbol{Z}$ that d-separates $\boldsymbol{X}$ from $\boldsymbol{Y}$ in $G_{\underline{\boldsymbol{X}}}$

$$P(y\,|\,do(x)) = \sum_{z_1,z_4} P(y\,|\,x,z_1,z_4)P(z_1,z_4)$$



Backdoor adjustment and inverse probability weighting (IPW)

$P(\boldsymbol{Y}|do(\boldsymbol{X}))$ is identifiable if $\exists\, \boldsymbol{Z}$ that *d*-separates $\boldsymbol{X}$ from $\boldsymbol{Y}$ in $G_{\underline{\boldsymbol{X}}}$

$$P(\boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x})) = \sum_{\boldsymbol{z}} P\,(\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{z})P(\boldsymbol{z})$$

$$= \sum_{\boldsymbol{z}} \frac{P(\boldsymbol{Y},\boldsymbol{x},\boldsymbol{z})P(\boldsymbol{z})}{P(\boldsymbol{x},\boldsymbol{z})}$$

$$= \sum_{\boldsymbol{z}} \frac{P(\boldsymbol{Y},\boldsymbol{x},\boldsymbol{z})P(\boldsymbol{z})}{P(\mathbf{x}|\mathbf{z})P(\boldsymbol{z})}$$

$$= \sum_{\boldsymbol{z}} \frac{P(\boldsymbol{Y},\boldsymbol{x},\boldsymbol{z})}{P(\mathbf{x}|\mathbf{z})}$$

Inverse propensity score

Backdoor adjustment and inverse probability weighting (IPW)

$P(\boldsymbol{Y}|do(\boldsymbol{X}))$ is identifiable if $\exists\, \boldsymbol{Z}$ that d-separates $\boldsymbol{X}$ from $\boldsymbol{Y}$ in $G_{\underline{\boldsymbol{X}}}$

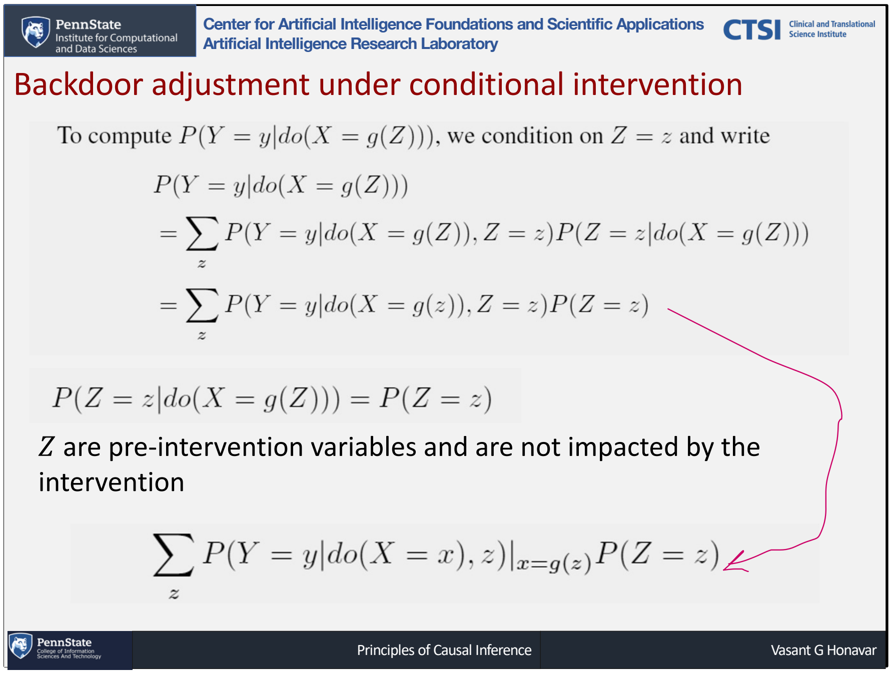
$$P(\boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x})) = \sum_{\boldsymbol{z}} P\,(\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{z})P(\boldsymbol{z})$$

$$= \sum_{\boldsymbol{z}} \frac{P(\boldsymbol{Y},\boldsymbol{x},\boldsymbol{z})}{P(\boldsymbol{x}|\boldsymbol{z})}$$

- IPW has the effect of estimating the interventional probability from a suitably resampled data to mimic an interventional distribution!

Backdoor adjustment under conditional intervention

- Suppose we contemplate an age-dependent policy whereby dosage X of drug is to be administered to patients, depends on their age Z. We write it as $do(X = g(Z))$.
- To find out the distribution of outcome Y that results from this policy, we seek to estimate $P(Y = y|do(X = g(Z)))$.
- We can often get it via Z-specific effect of $P(Y|do(X = x), Z = z)$



Backdoor adjustment under conditional intervention

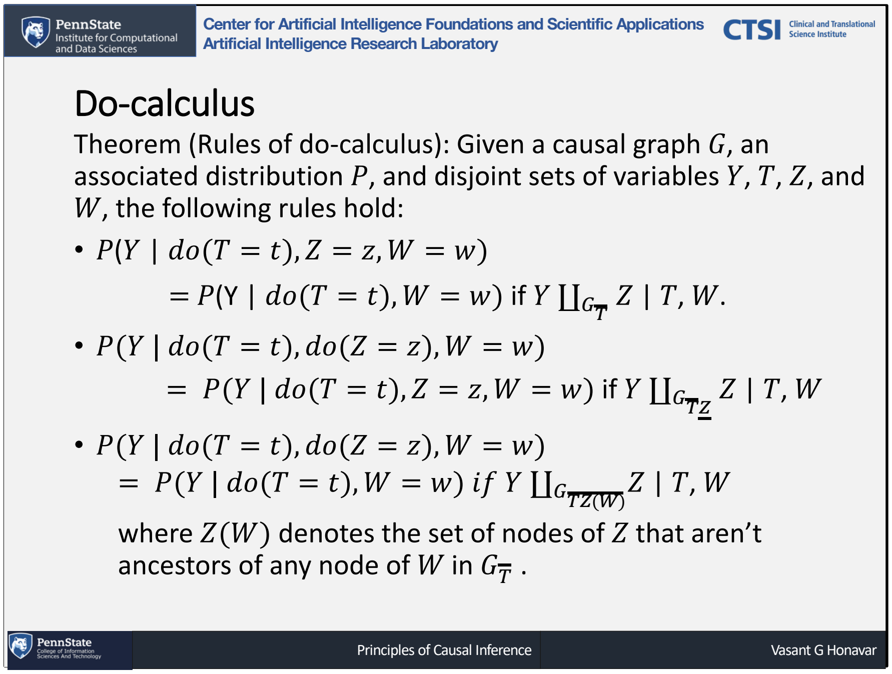
To compute $P(Y=y|do(X=g(Z)))$, we condition on $Z=z$ and write

$$\begin{aligned}
&P(Y=y|do(X=g(Z))) \\
&= \sum_z P(Y=y|do(X=g(Z)), Z=z)P(Z=z|do(X=g(Z))) \\
&= \sum_z P(Y=y|do(X=g(z)), Z=z)P(Z=z)
\end{aligned}$$

$$P(Z=z|do(X=g(Z))) = P(Z=z)$$

Z are pre-intervention variables and are not impacted by the intervention

$$\sum_z P(Y=y|do(X=x),z)|_{x=g(z)}P(Z=z)$$



Do-calculus

Theorem (Rules of do-calculus): Given a causal graph G, an associated distribution P, and disjoint sets of variables Y, T, Z, and W, the following rules hold:

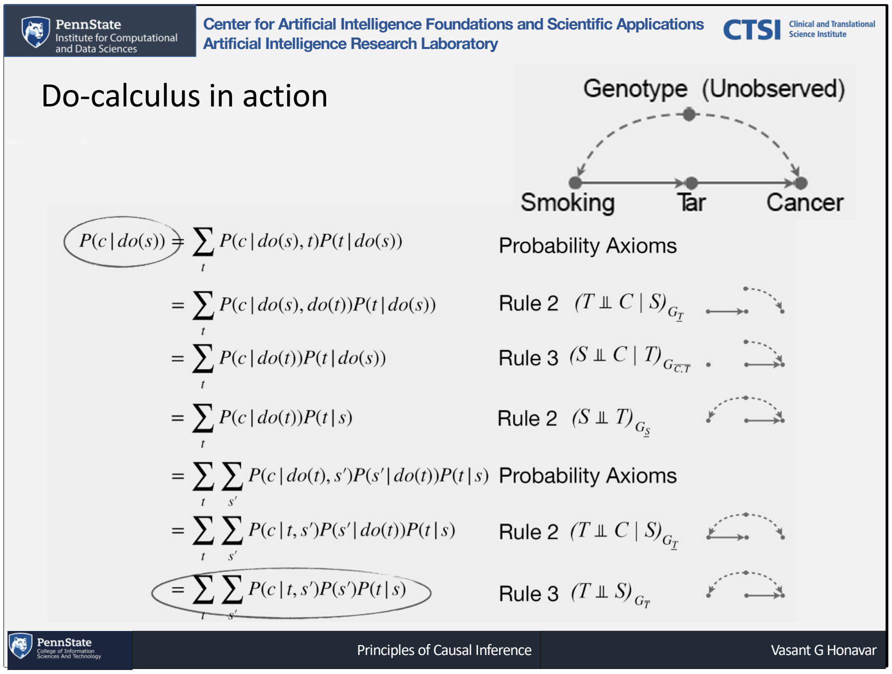
- $P(Y \mid do(T = t), Z = z, W = w)$

  $= P(\mathrm{Y} \mid do(T = t), W = w)$ if $Y \coprod_{G_{\overline{T}}} Z \mid T, W.$
- $P(Y \mid do(T = t), do(Z = z), W = w)$

  $= \ P(Y \mid do(T = t), Z = z, W = w)$ if $Y \coprod_{G_{\overline{T}\underline{Z}}} Z \mid T, W$
- $P(Y \mid do(T = t), do(Z = z), W = w)$

  $= \ P(Y \mid do(T = t), W = w) \ if \ Y \coprod_{G_{\overline{TZ(W)}}} Z \mid T, W$

  where $Z(W)$ denotes the set of nodes of Z that aren't ancestors of any node of W in $G_{\overline{T}}$.

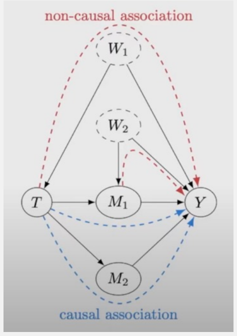


Do-calculus in action

Genotype (Unobserved)
Smoking
Tar
Cancer

$$
\begin{aligned}
\boxed{P(c \mid do(s))} &= \sum_t P(c \mid do(s), t) P(t \mid do(s)) && \text{Probability Axioms} \\
&= \sum_t P(c \mid do(s), do(t)) P(t \mid do(s)) && \text{Rule 2 } (T \perp\!\!\!\perp C \mid S)_{G_{\underline{T}}} \\
&= \sum_t P(c \mid do(t)) P(t \mid do(s)) && \text{Rule 3 } (S \perp\!\!\!\perp C \mid T)_{G_{\overline{C,T}}} \\
&= \sum_t P(c \mid do(t)) P(t \mid s) && \text{Rule 2 } (S \perp\!\!\!\perp T)_{G_{\underline{S}}} \\
&= \sum_t \sum_{s'} P(c \mid do(t), s') P(s' \mid do(t)) P(t \mid s) && \text{Probability Axioms} \\
&= \sum_t \sum_{s'} P(c \mid t, s') P(s' \mid do(t)) P(t \mid s) && \text{Rule 2 } (T \perp\!\!\!\perp C \mid S)_{G_{\underline{T}}} \\
&\boxed{= \sum_t \sum_{s'} P(c \mid t, s') P(s') P(t \mid s)} && \text{Rule 3 } (T \perp\!\!\!\perp S)_{G_{\overline{T}}}
\end{aligned}
$$

Identifiability of causal effect from graph



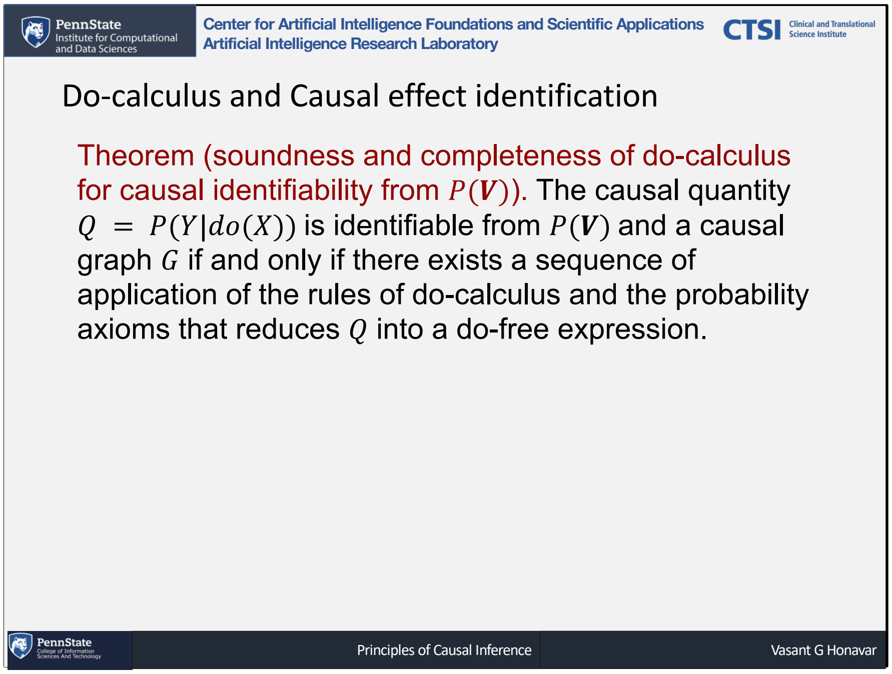
Back-Door Criterion: A set of variables Z is said to satisfy the back-door criterion relative to an ordered pair of variables (T, Y), if:

- Z intercepts all backdoor paths from T to Y (paths that contain an arrow into T)
- No node in Z is a descendent of T

Front-Door Criterion: A set of variables Z is said to satisfy the front-door criterion relative to an ordered pair of variables (T, Y), if:

- Z intercepts all directed paths from T to Y
- There is no unblocked backdoor path from T to Z
- All backdoor paths from Z to Y are blocked by T

- Is the back-door criterion satisfied with respect to (T, Y)?
  - No, because we cannot block the backdoor path into T because W_1 is unobserved.
- Is the front-door criterion satisfied with respect to (T, Y)?
  - No (because W_2 is unobserved, so there is non-causal association between M_1 and Y)



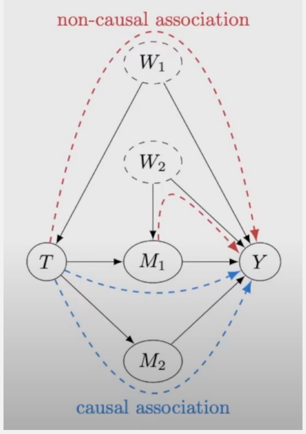
Do-calculus and Causal effect identification

Theorem (soundness and completeness of do-calculus for causal identifiability from $P(\boldsymbol{V})$). The causal quantity $Q = P(Y|do(X))$ is identifiable from $P(\boldsymbol{V})$ and a causal graph G if and only if there exists a sequence of application of the rules of do-calculus and the probability axioms that reduces Q into a do-free expression.

Identifiability of causal effect from graph

Unconfounded children criterion

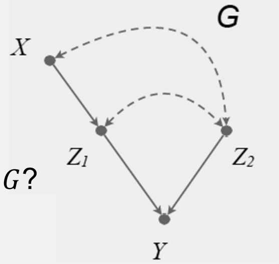
- The unconfounded children criterion is satisfied if it is possible to block all backdoor paths from the treatment variable T to all of its children that are ancestors of Y with a single conditioning set (Tian & Pearl, 2002).
- Unconfounded children criterion
  - Sufficient (but not necessary) condition for identifiability when T is a single treatment variable.
  - Generalizes the back-door and front-door criteria

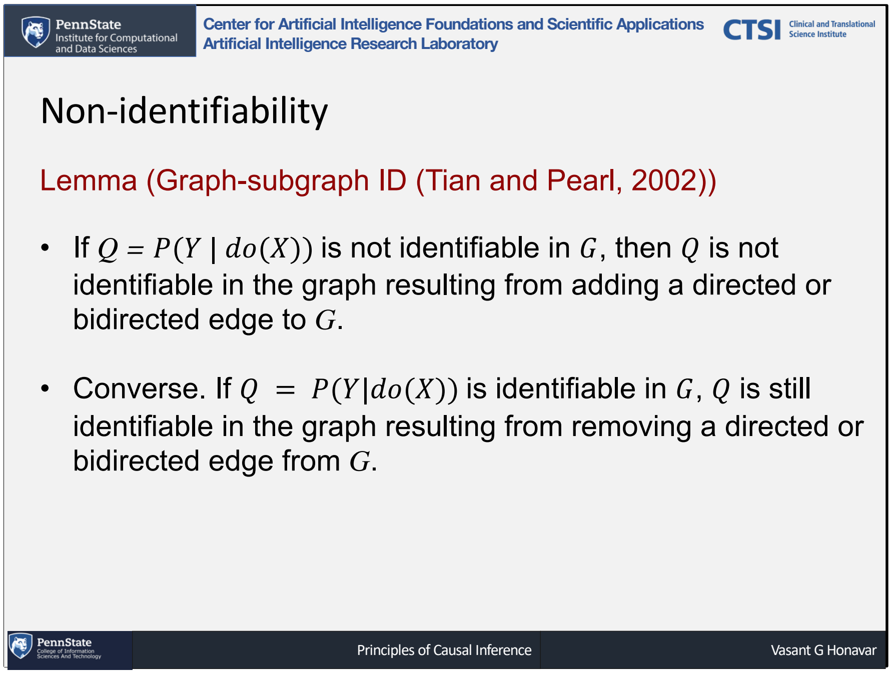


Non-identifiability

- Two models have the same distribution $P(\boldsymbol{V})$ over the observable variables $\boldsymbol{V}$
- Yet they differ in $P(Y|do(X)$
- In other words, $P(\boldsymbol{V})$ does not uniquely specify $P(Y|do(X)$

- Is $P(Y \mid do(X))$ identifiable from G?
- Is $P(Y \mid do(X), Z_1)$ identifiable from G?
- Is $P(Y \mid do(X), Z_2)$ identifiable from G?
- Is $P(Y \mid do(X), Z_1, Z_2)$ identifiable from G?

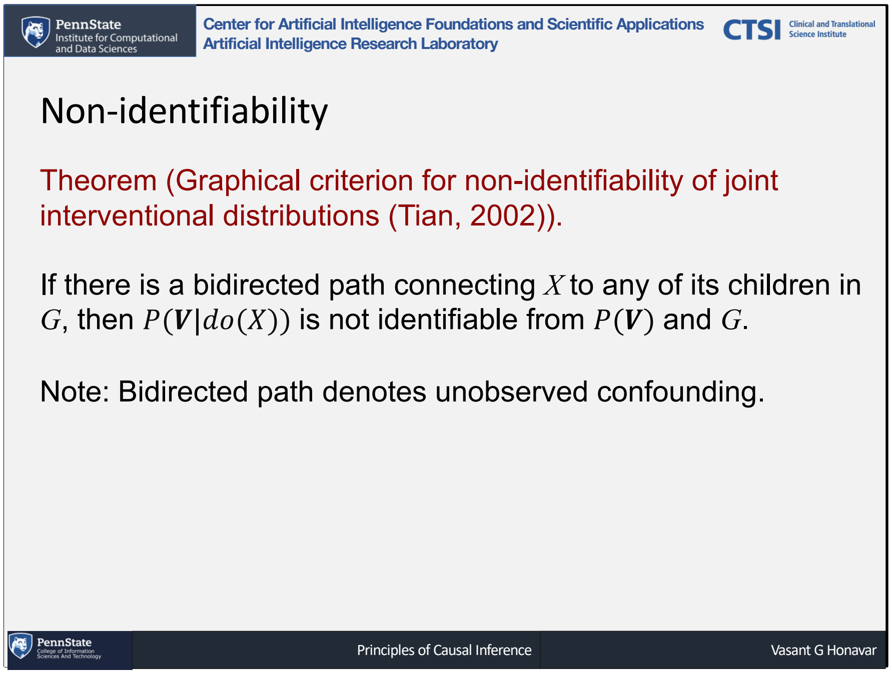




Non-identifiability

Lemma (Graph-subgraph ID (Tian and Pearl, 2002))

- If $Q = P(Y \mid do(X))$ is not identifiable in G, then Q is not identifiable in the graph resulting from adding a directed or bidirected edge to G.
- Converse. If $Q = P(Y|do(X))$ is identifiable in G, Q is still identifiable in the graph resulting from removing a directed or bidirected edge from G.



Non-identifiability

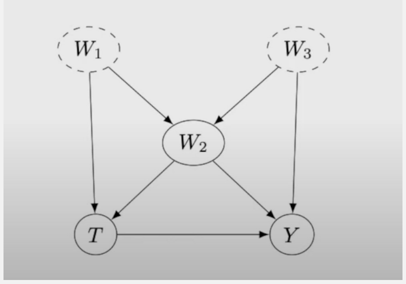
Theorem (Graphical criterion for non-identifiability of joint interventional distributions (Tian, 2002)).

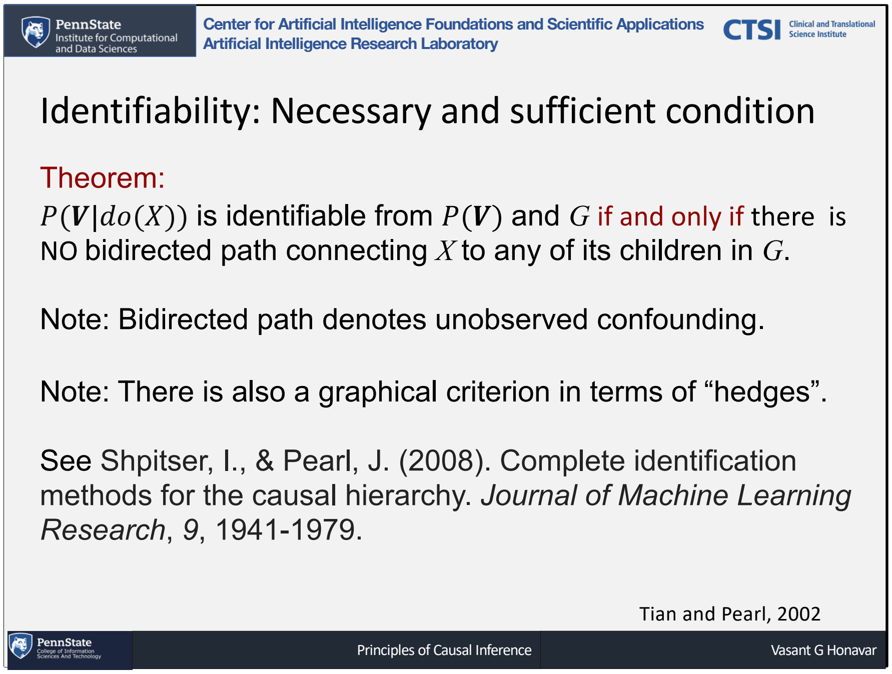
If there is a bidirected path connecting X to any of its children in G, then $P(\boldsymbol{V}|do(X))$ is not identifiable from $P(\boldsymbol{V})$ and G.

Note: Bidirected path denotes unobserved confounding.

Necessary condition for identifiability

For each backdoor path from T to any child M of T that is an ancestor of Y, it is possible to block that path (Pearl, 2009)





Identifiability: Necessary and sufficient condition

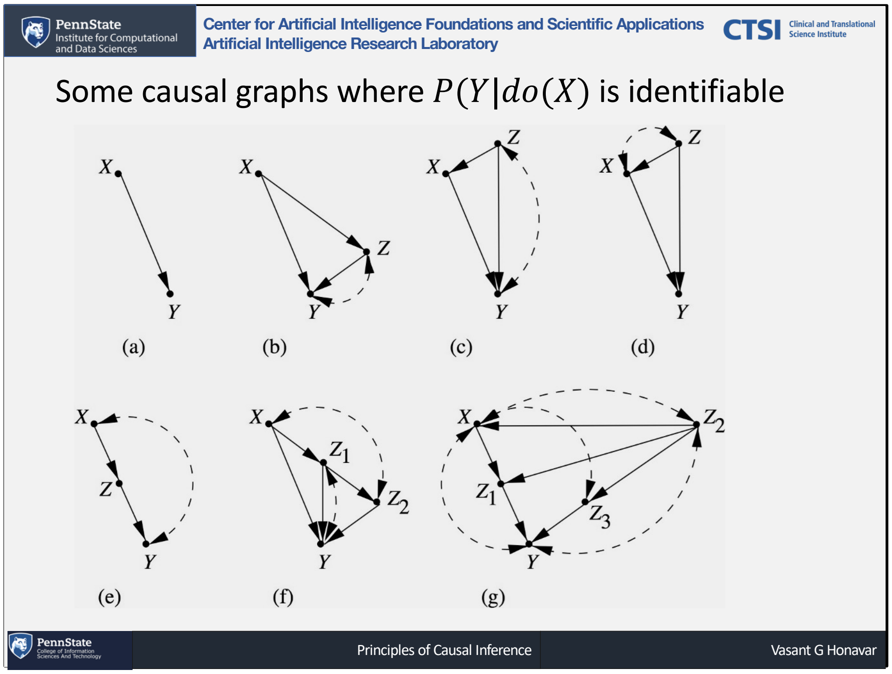
Theorem:
$P(\boldsymbol{V}|do(X))$ is identifiable from $P(\boldsymbol{V})$ and G if and only if there is NO bidirected path connecting X to any of its children in G.

Note: Bidirected path denotes unobserved confounding.

Note: There is also a graphical criterion in terms of “hedges”.

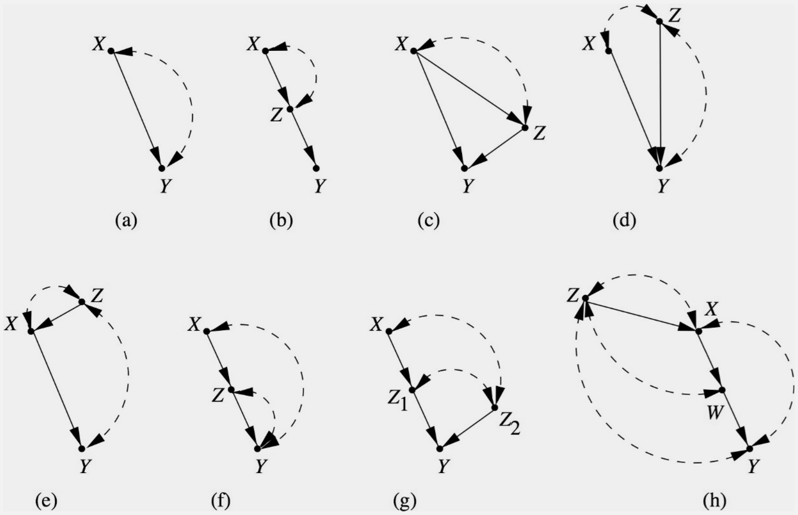
See Shpitser, I., & Pearl, J. (2008). Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, *9*, 1941-1979.

Tian and Pearl, 2002



Some causal graphs where $P(Y|do(X)$ is identifiable

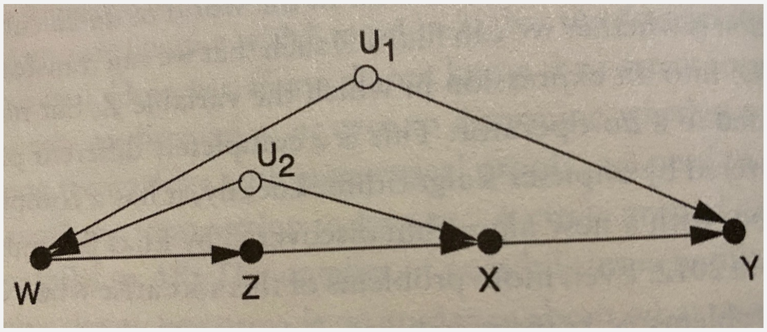
Some causal graphs where $P(Y|do(X)$ is not identifiable

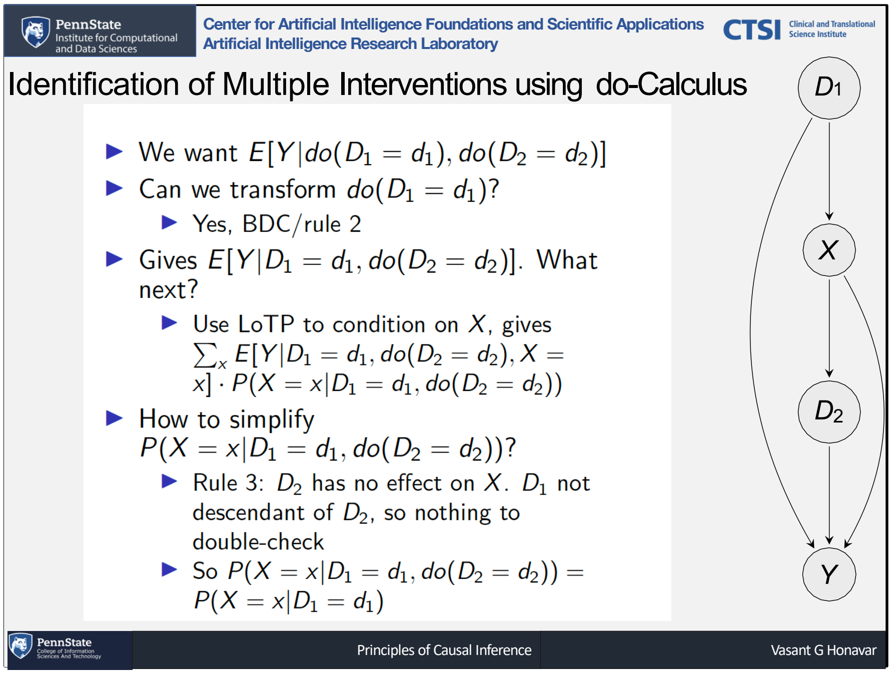


(a) (b) (c) (d)

(e) (f) (g) (h)

Exercise: Apply do calculus to identify the causal effect of X on Y (U_1 and U_2 are unobserved)

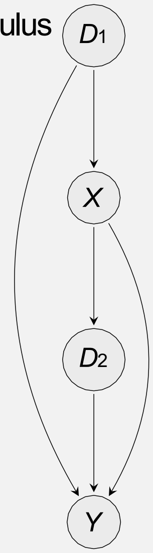




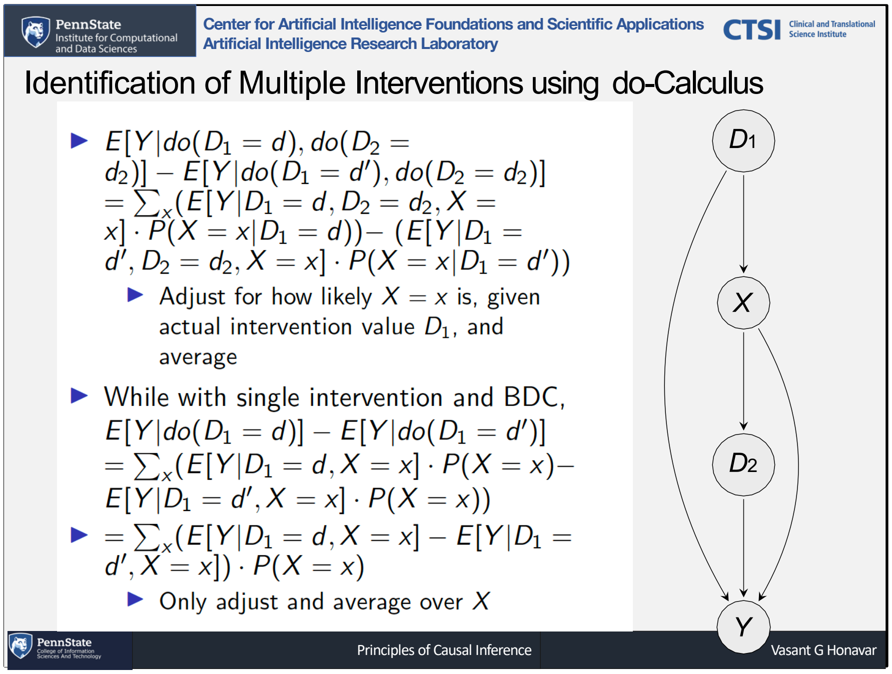
Identification of Multiple Interventions using do-Calculus

- We want $E[Y|do(D_1 = d_1), do(D_2 = d_2)]$
- Can we transform $do(D_1 = d_1)$?
  - Yes, BDC/rule 2
- Gives $E[Y|D_1 = d_1, do(D_2 = d_2)]$. What next?
  - Use LoTP to condition on X, gives $\sum_x E[Y|D_1 = d_1, do(D_2 = d_2), X = x] \cdot P(X = x|D_1 = d_1, do(D_2 = d_2))$
- How to simplify $P(X = x|D_1 = d_1, do(D_2 = d_2))$?
  - Rule 3: D_2 has no effect on X. D_1 not descendant of D_2, so nothing to double-check
  - So $P(X = x|D_1 = d_1, do(D_2 = d_2)) = P(X = x|D_1 = d_1)$

Identification of Multiple Interventions using do-Calculus

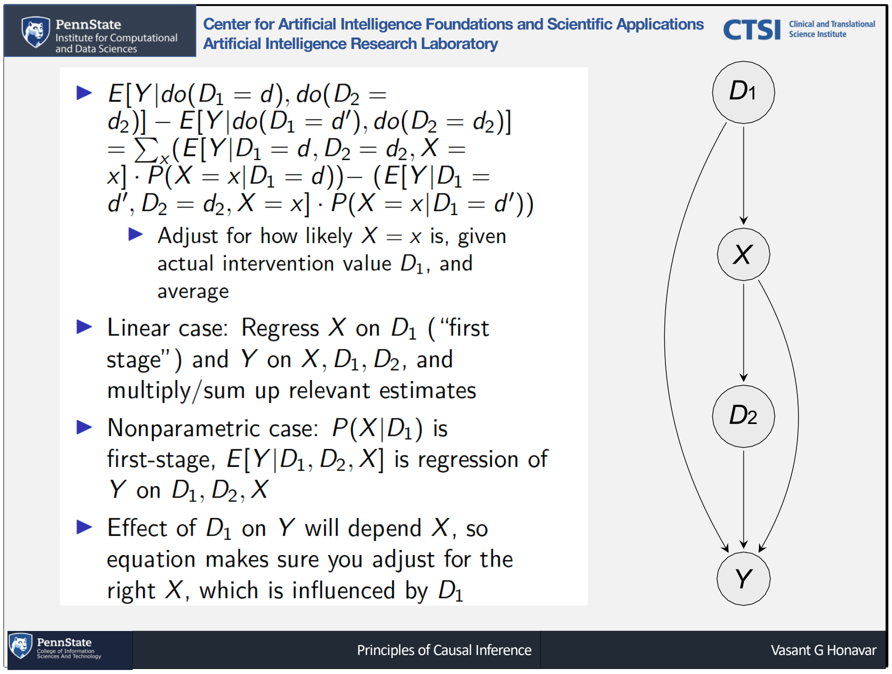


- So far we have $E[Y|do(D_1 = d_1), do(D_2 = d_2)] = \sum_x E[Y|D_1 = d_1, do(D_2 = d_2), X = x] \cdot P(X = x|D_1 = d_1)$
- What can we do about remaining $do(D_2 = d_2)$?
  - BDC/rule 2, gives $\sum_x E[Y|D_1 = d_1, D_2 = d_2, X = x] \cdot P(X = x|D_1 = d_1)$ No *do* left, identification!
- This means that $E[Y|do(D_1 = d), do(D_2 = d_2)] - E[Y|do(D_1 = d'), do(D_2 = d_2)] = \sum_x (E[Y|D_1 = d, D_2 = d_2, X = x] \cdot P(X = x|D_1 = d)) - (E[Y|D_1 = d', D_2 = d_2, X = x] \cdot P(X = x|D_1 = d'))$



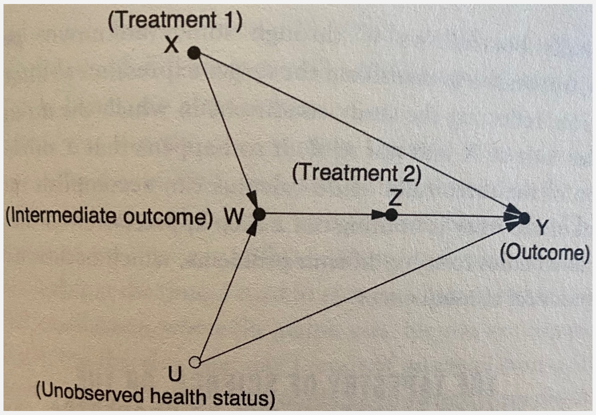
Identification of Multiple Interventions using do-Calculus

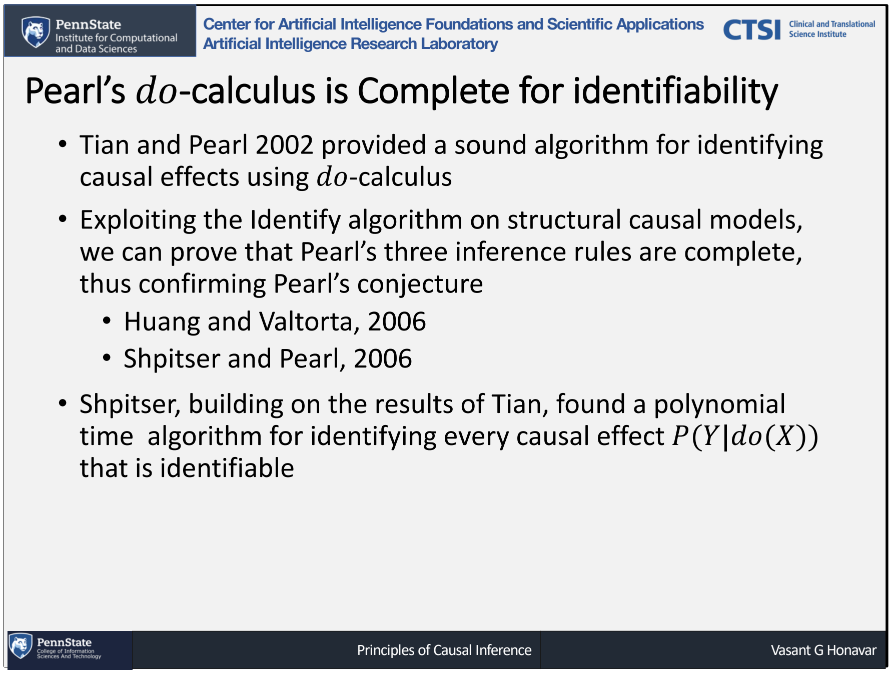
- $E[Y|do(D_1 = d), do(D_2 = d_2)] - E[Y|do(D_1 = d'), do(D_2 = d_2)] = \sum_x (E[Y|D_1 = d, D_2 = d_2, X = x] \cdot P(X = x|D_1 = d)) - (E[Y|D_1 = d', D_2 = d_2, X = x] \cdot P(X = x|D_1 = d'))$
  - Adjust for how likely $X = x$ is, given actual intervention value D_1, and average
- While with single intervention and BDC, $E[Y|do(D_1 = d)] - E[Y|do(D_1 = d')] = \sum_x (E[Y|D_1 = d, X = x] \cdot P(X = x) - E[Y|D_1 = d', X = x] \cdot P(X = x))$
- $= \sum_x (E[Y|D_1 = d, X = x] - E[Y|D_1 = d', X = x]) \cdot P(X = x)$
  - Only adjust and average over X



- $E[Y|do(D_1 = d), do(D_2 = d_2)] - E[Y|do(D_1 = d'), do(D_2 = d_2)] = \sum_x (E[Y|D_1 = d, D_2 = d_2, X = x] \cdot P(X = x|D_1 = d)) - (E[Y|D_1 = d', D_2 = d_2, X = x] \cdot P(X = x|D_1 = d'))$
  - Adjust for how likely $X = x$ is, given actual intervention value D_1, and average
- Linear case: Regress X on D_1 ("first stage") and Y on X, D_1, D_2, and multiply/sum up relevant estimates
- Nonparametric case: $P(X|D_1)$ is first-stage, $E[Y|D_1, D_2, X]$ is regression of Y on D_1, D_2, X
- Effect of D_1 on Y will depend X, so equation makes sure you adjust for the right X, which is influenced by D_1

Exercise: Apply do calculus to identify the causal effect of X on Y (U is unobserved)





Pearl’s do-calculus is Complete for identifiability

- Tian and Pearl 2002 provided a sound algorithm for identifying causal effects using do-calculus
- Exploiting the Identify algorithm on structural causal models, we can prove that Pearl’s three inference rules are complete, thus confirming Pearl’s conjecture
  - Huang and Valtorta, 2006
  - Shpitser and Pearl, 2006
- Shpitser, building on the results of Tian, found a polynomial time algorithm for identifying every causal effect $P(Y|do(X))$ that is identifiable