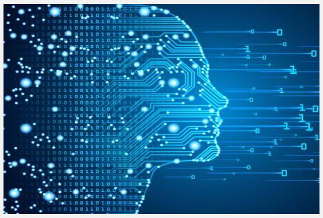
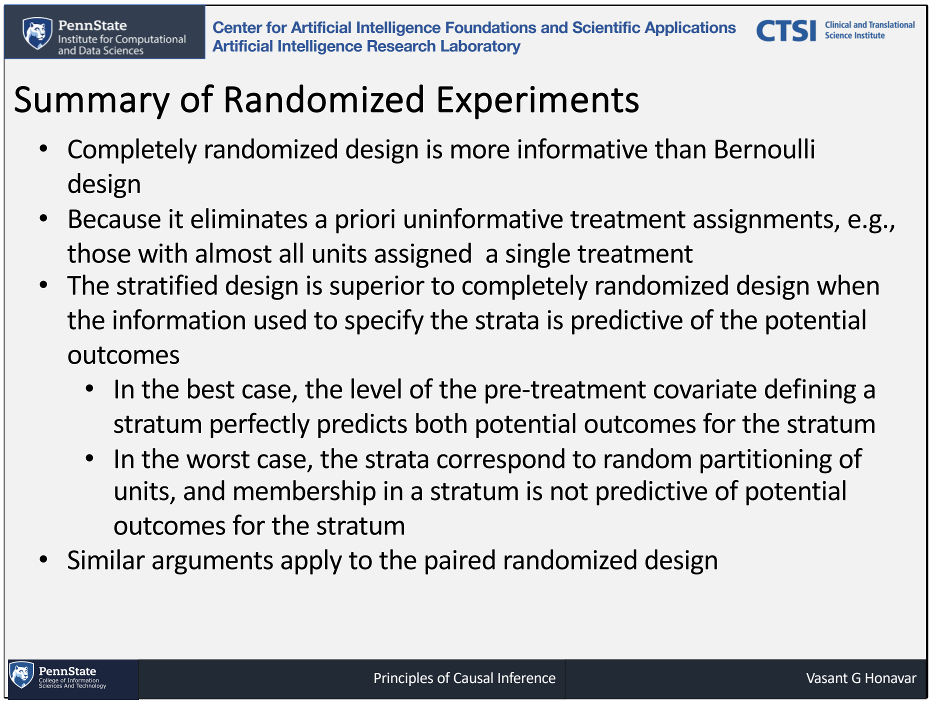


Principles of Causal Inference

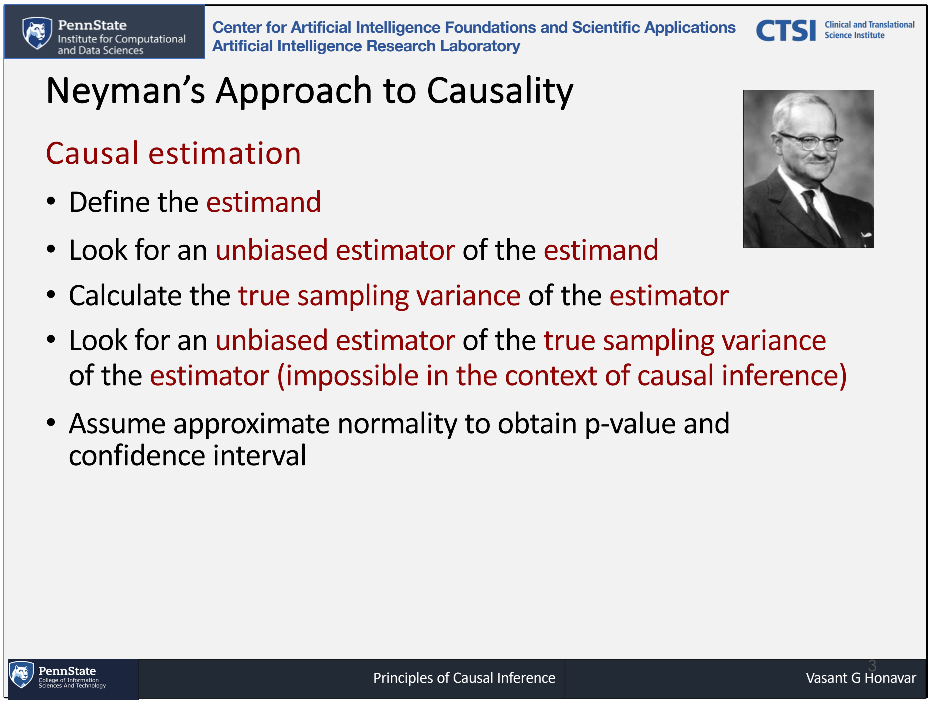
Vasant G. Honavar

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



Summary of Randomized Experiments

- Completely randomized design is more informative than Bernoulli design
- Because it eliminates a priori uninformative treatment assignments, e.g., those with almost all units assigned a single treatment
- The stratified design is superior to completely randomized design when the information used to specify the strata is predictive of the potential outcomes
  - In the best case, the level of the pre-treatment covariate defining a stratum perfectly predicts both potential outcomes for the stratum
  - In the worst case, the strata correspond to random partitioning of units, and membership in a stratum is not predictive of potential outcomes for the stratum
- Similar arguments apply to the paired randomized design



Neyman's Approach to Causality

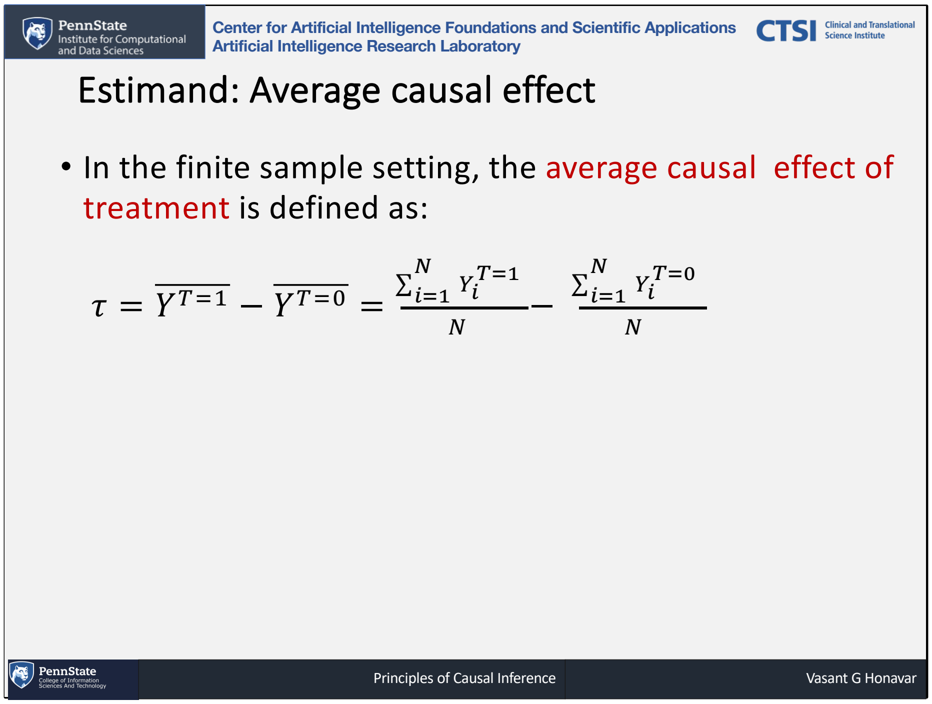
Causal estimation

- Define the estimand
- Look for an unbiased estimator of the estimand
- Calculate the true sampling variance of the estimator
- Look for an unbiased estimator of the true sampling variance of the estimator (impossible in the context of causal inference)
- Assume approximate normality to obtain p-value and confidence interval

Finite Sample versus Super Population

- Finite sample inference:
  - Only concerned with units in the sample
  - Only source of randomness is random assignment to treatment groups

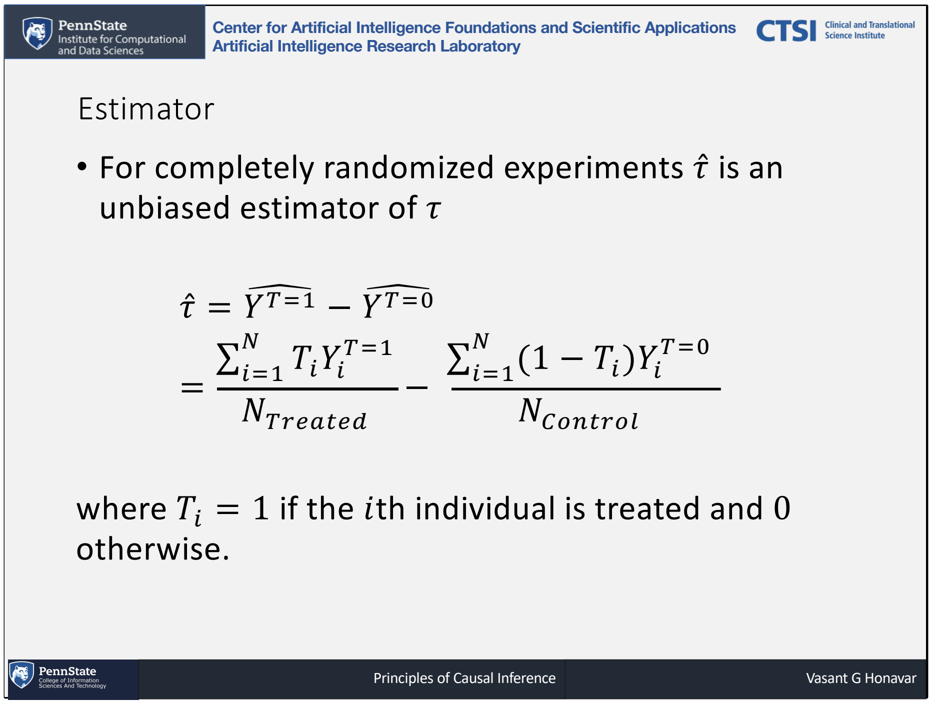
- Super population inference:
  - Extend inferences to greater population
  - Two sources of randomness: random sampling, random assignment
  - “repeated sampling”



Estimand: Average causal effect

- In the finite sample setting, the average causal effect of treatment is defined as:

$$\tau = \overline{Y^{T=1}} - \overline{Y^{T=0}} = \frac{\sum_{i=1}^{N} Y_i^{T=1}}{N} - \frac{\sum_{i=1}^{N} Y_i^{T=0}}{N}$$



Estimator

- For completely randomized experiments $\hat{\tau}$ is an unbiased estimator of τ

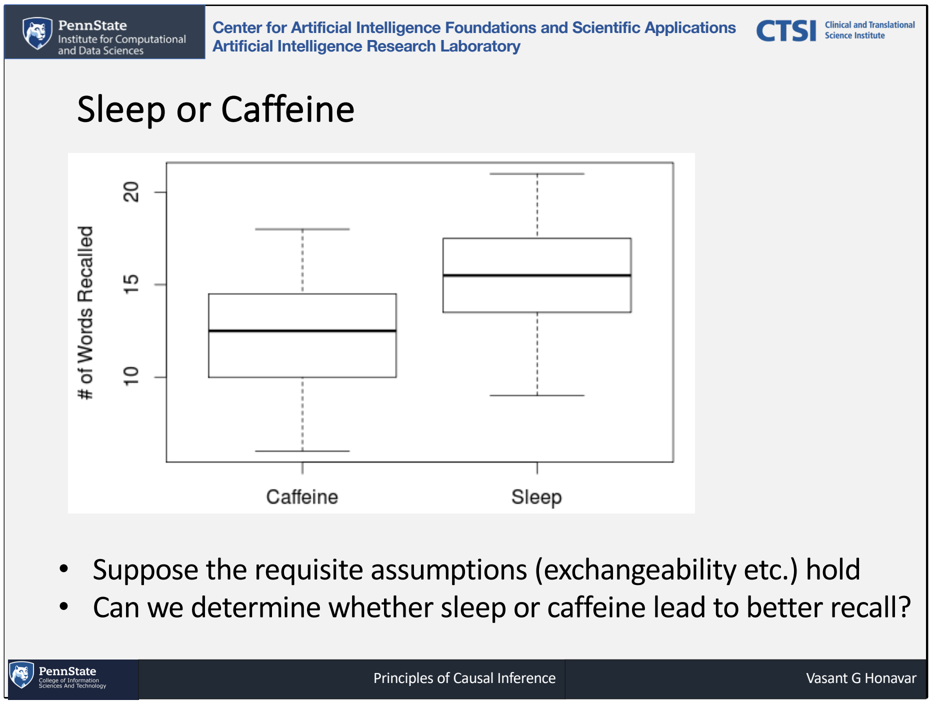
$$\hat{\tau} = \widehat{Y^{T=1}} - \widehat{Y^{T=0}}$$
$$= \frac{\sum_{i=1}^{N} T_i Y_i^{T=1}}{N_{Treated}} - \frac{\sum_{i=1}^{N} (1 - T_i) Y_i^{T=0}}{N_{Control}}$$

where $T_i = 1$ if the ith individual is treated and 0 otherwise.

Example: Sleep or Caffeine?

- Is sleep or caffeine better for memory?
- 24 adults were given a list of words to memorize, then randomly divided into two groups and sent over to take a break
- During the break one group took a nap for an hour and a half, while the other group stayed awake and then took a caffeine pill after an hour
- Y: number of words recalled

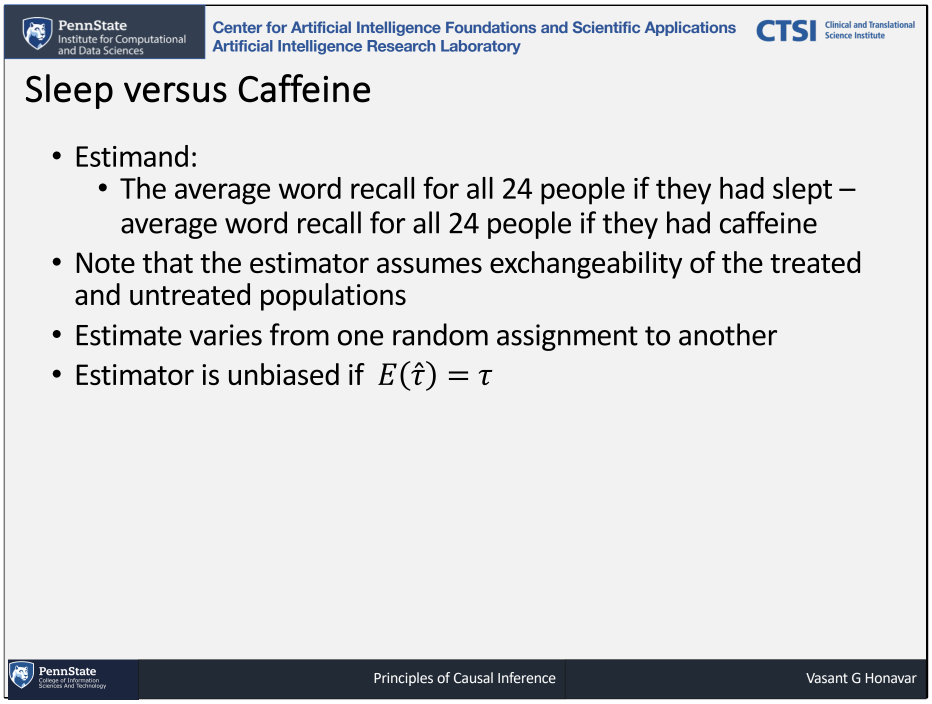
Mednick et al., "Comparing the benefits of caffeine, naps and placebo on verbal, motor and perceptual memory", Behavioral Brain Research, 2008; 193: 79-86.



Sleep or Caffeine

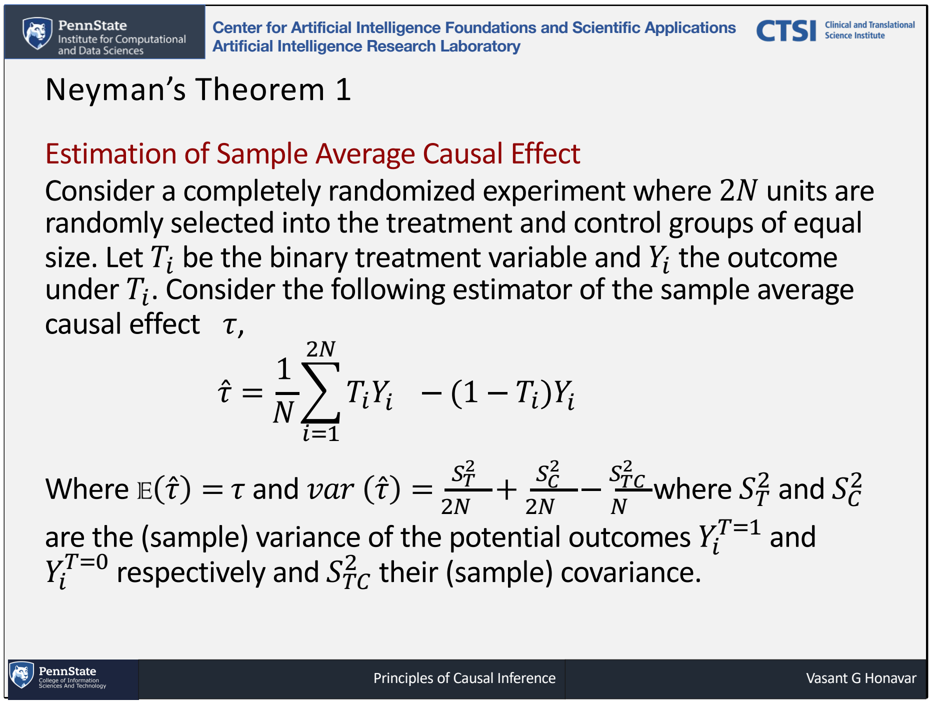
# of Words Recalled
10
15
20
Caffeine
Sleep

- Suppose the requisite assumptions (exchangeability etc.) hold
- Can we determine whether sleep or caffeine lead to better recall?



Sleep versus Caffeine

- Estimand:
  - The average word recall for all 24 people if they had slept – average word recall for all 24 people if they had caffeine
- Note that the estimator assumes exchangeability of the treated and untreated populations
- Estimate varies from one random assignment to another
- Estimator is unbiased if $E(\hat{\tau}) = \tau$



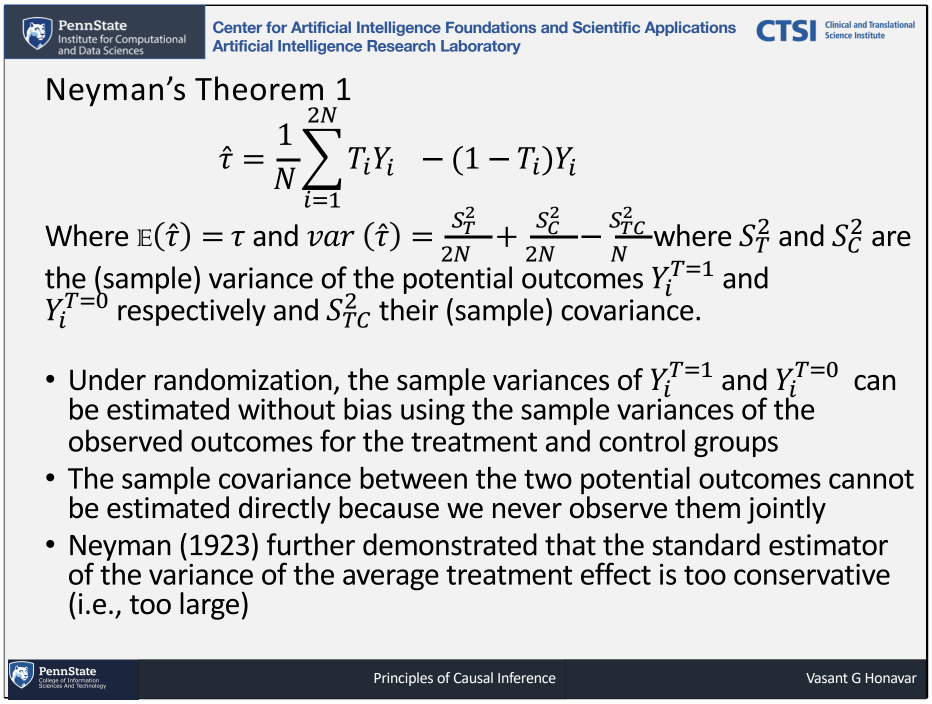
Neyman's Theorem 1

Estimation of Sample Average Causal Effect

Consider a completely randomized experiment where $2N$ units are randomly selected into the treatment and control groups of equal size. Let T_i be the binary treatment variable and Y_i the outcome under T_i. Consider the following estimator of the sample average causal effect τ,

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{2N} T_i Y_i \quad - (1 - T_i) Y_i$$

Where $\mathbb{E}(\hat{\tau}) = \tau$ and $var\ (\hat{\tau}) = \frac{S_T^2}{2N} + \frac{S_C^2}{2N} - \frac{S_{TC}^2}{N}$ where S_T^2 and S_C^2 are the (sample) variance of the potential outcomes $Y_i^{T=1}$ and $Y_i^{T=0}$ respectively and S_{TC}^2 their (sample) covariance.

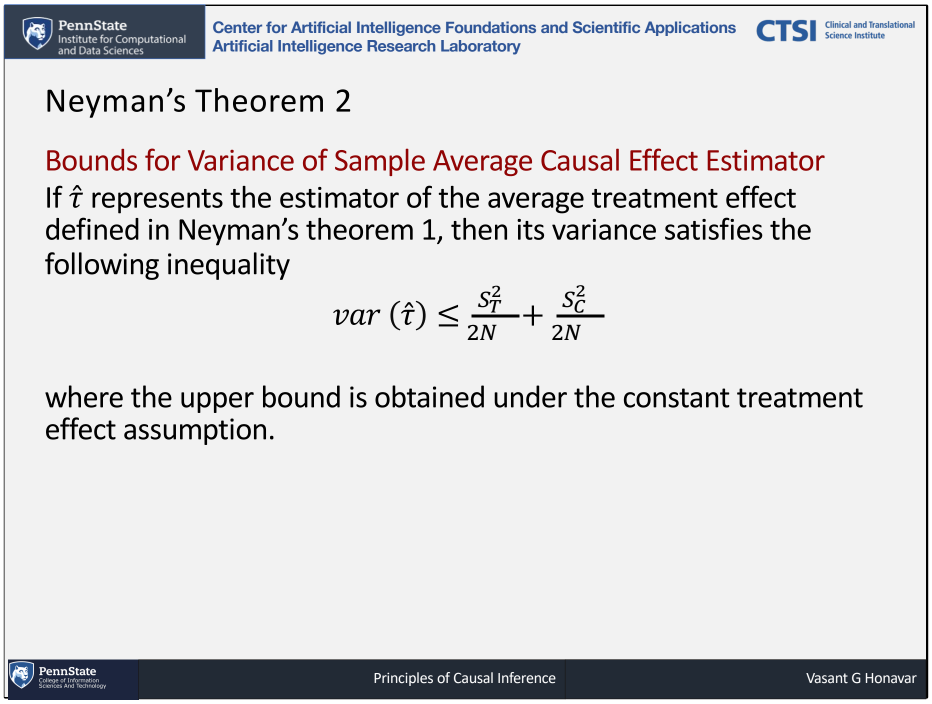


Neyman’s Theorem 1

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{2N} T_i Y_i \quad - (1 - T_i) Y_i$$

Where $\mathbb{E}(\hat{\tau}) = \tau$ and $var\ (\hat{\tau}) = \frac{S_T^2}{2N} + \frac{S_C^2}{2N} - \frac{S_{TC}^2}{N}$ where S_T^2 and S_C^2 are the (sample) variance of the potential outcomes $Y_i^{T=1}$ and $Y_i^{T=0}$ respectively and S_{TC}^2 their (sample) covariance.

- Under randomization, the sample variances of $Y_i^{T=1}$ and $Y_i^{T=0}$ can be estimated without bias using the sample variances of the observed outcomes for the treatment and control groups
- The sample covariance between the two potential outcomes cannot be estimated directly because we never observe them jointly
- Neyman (1923) further demonstrated that the standard estimator of the variance of the average treatment effect is too conservative (i.e., too large)



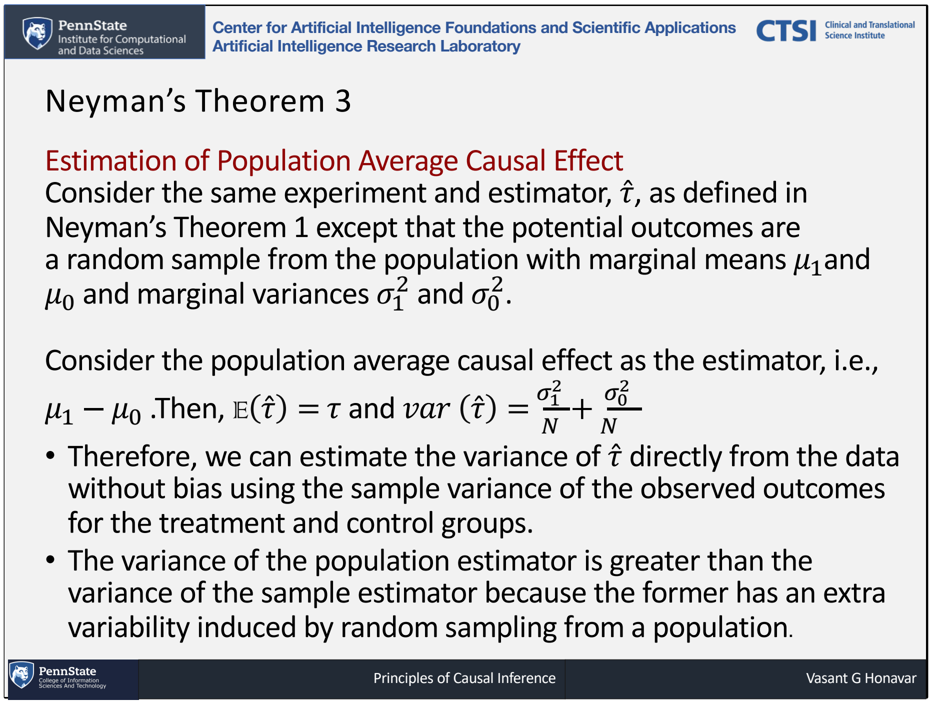
Neyman’s Theorem 2

Bounds for Variance of Sample Average Causal Effect Estimator

If $\hat{\tau}$ represents the estimator of the average treatment effect defined in Neyman’s theorem 1, then its variance satisfies the following inequality

$$var\ (\hat{\tau}) \leq \frac{S_T^2}{2N} + \frac{S_C^2}{2N}$$

where the upper bound is obtained under the constant treatment effect assumption.



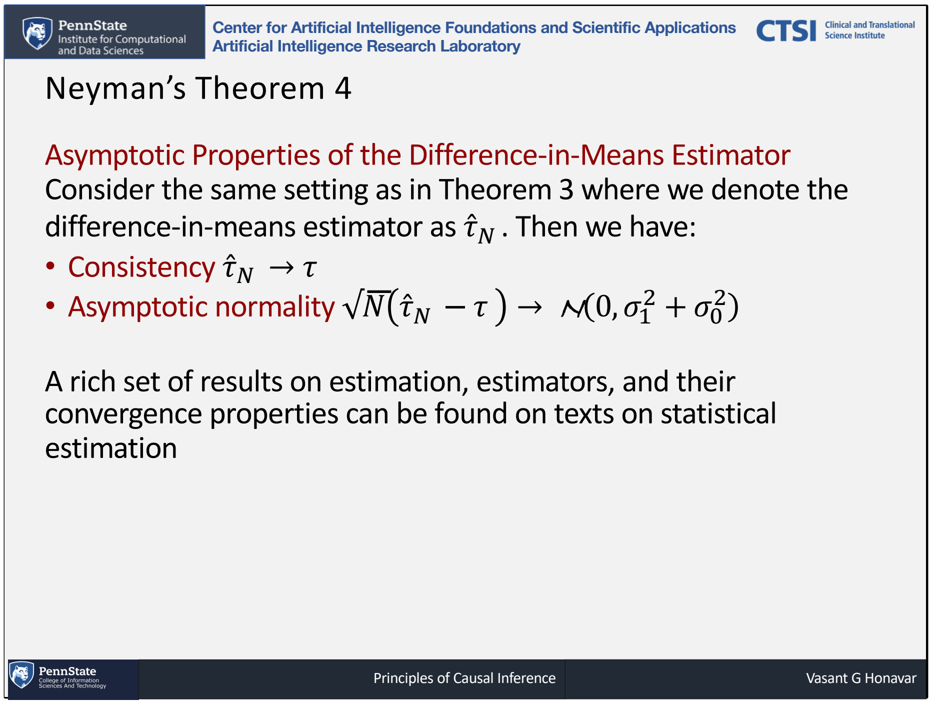
Neyman’s Theorem 3

Estimation of Population Average Causal Effect

Consider the same experiment and estimator, $\hat{\tau}$, as defined in Neyman’s Theorem 1 except that the potential outcomes are a random sample from the population with marginal means $\mu_1$and μ_0 and marginal variances σ_1^2 and σ_0^2.

Consider the population average causal effect as the estimator, i.e., $\mu_1 - \mu_0$ .Then, $\mathbb{E}(\hat{\tau}) = \tau$ and $var\ (\hat{\tau}) = \frac{\sigma_1^2}{N} + \frac{\sigma_0^2}{N}$

- Therefore, we can estimate the variance of $\hat{\tau}$ directly from the data without bias using the sample variance of the observed outcomes for the treatment and control groups.
- The variance of the population estimator is greater than the variance of the sample estimator because the former has an extra variability induced by random sampling from a population.



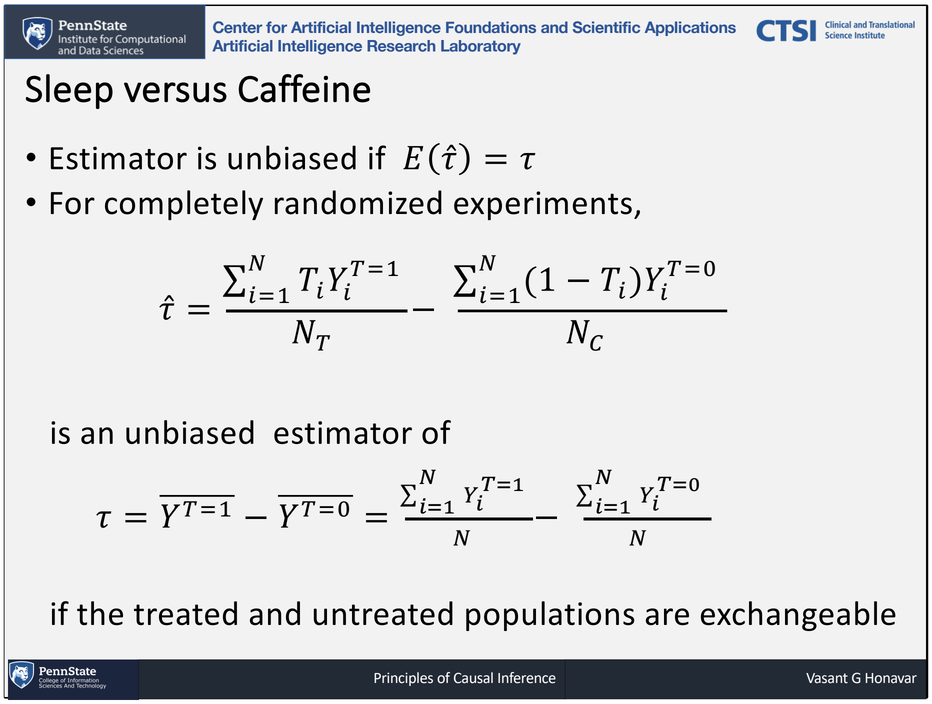
Neyman’s Theorem 4

Asymptotic Properties of the Difference-in-Means Estimator

Consider the same setting as in Theorem 3 where we denote the difference-in-means estimator as $\hat{\tau}_N$. Then we have:

- Consistency $\hat{\tau}_N \to \tau$
- Asymptotic normality $\sqrt{N}\left(\hat{\tau}_N - \tau\right) \to \mathcal{N}(0, \sigma_1^2 + \sigma_0^2)$

A rich set of results on estimation, estimators, and their convergence properties can be found on texts on statistical estimation



Sleep versus Caffeine

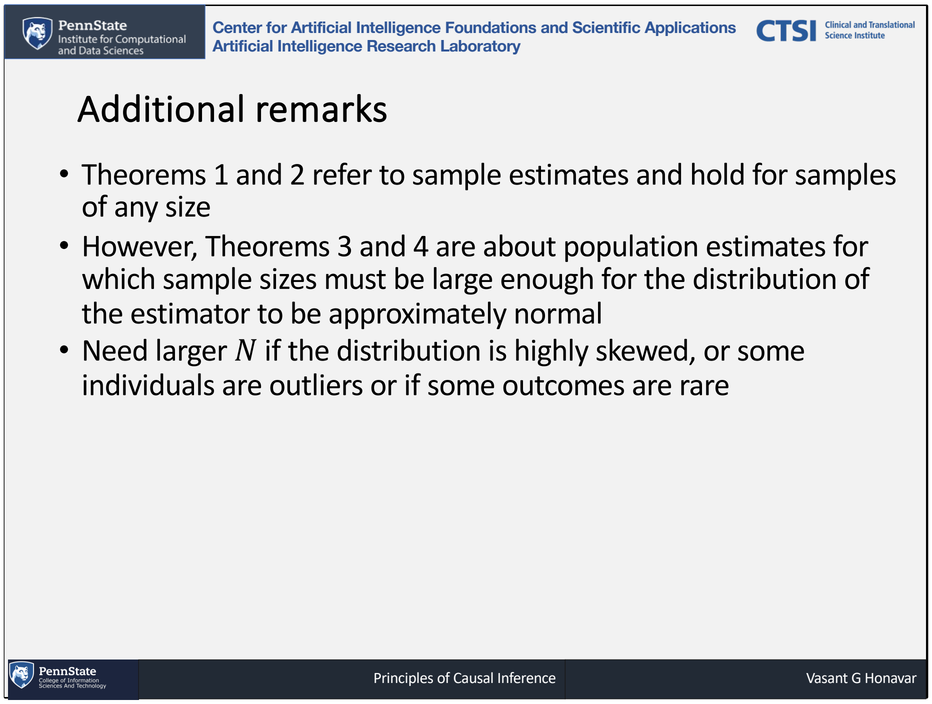
- Estimator is unbiased if $E(\hat{\tau}) = \tau$
- For completely randomized experiments,

$$\hat{\tau} = \frac{\sum_{i=1}^{N} T_i Y_i^{T=1}}{N_T} - \frac{\sum_{i=1}^{N} (1 - T_i) Y_i^{T=0}}{N_C}$$

is an unbiased estimator of

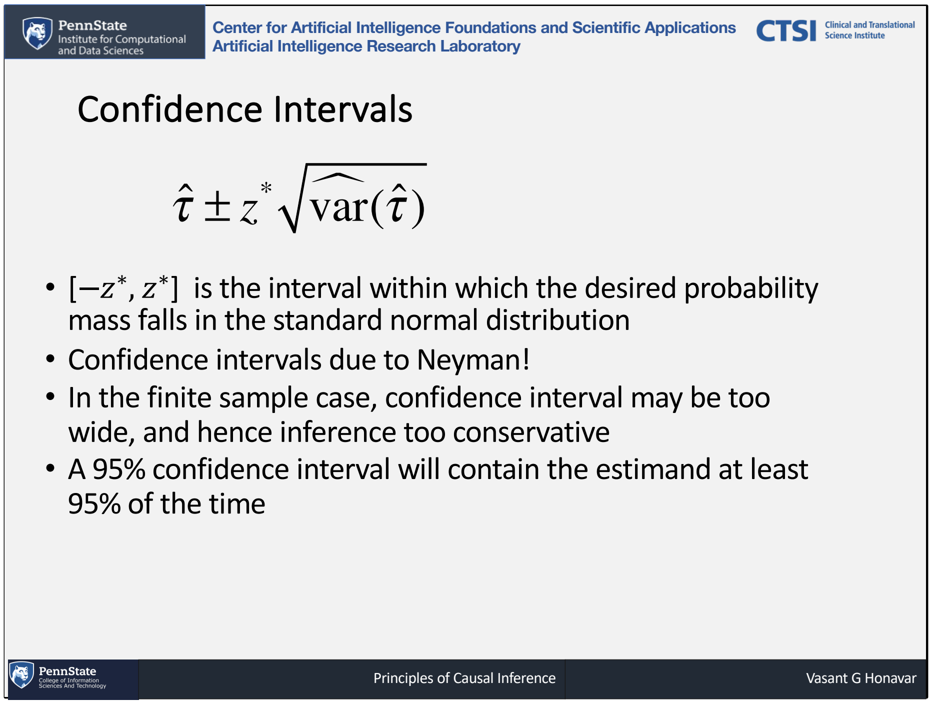
$$\tau = \overline{Y^{T=1}} - \overline{Y^{T=0}} = \frac{\sum_{i=1}^{N} Y_i^{T=1}}{N} - \frac{\sum_{i=1}^{N} Y_i^{T=0}}{N}$$

if the treated and untreated populations are exchangeable



Additional remarks

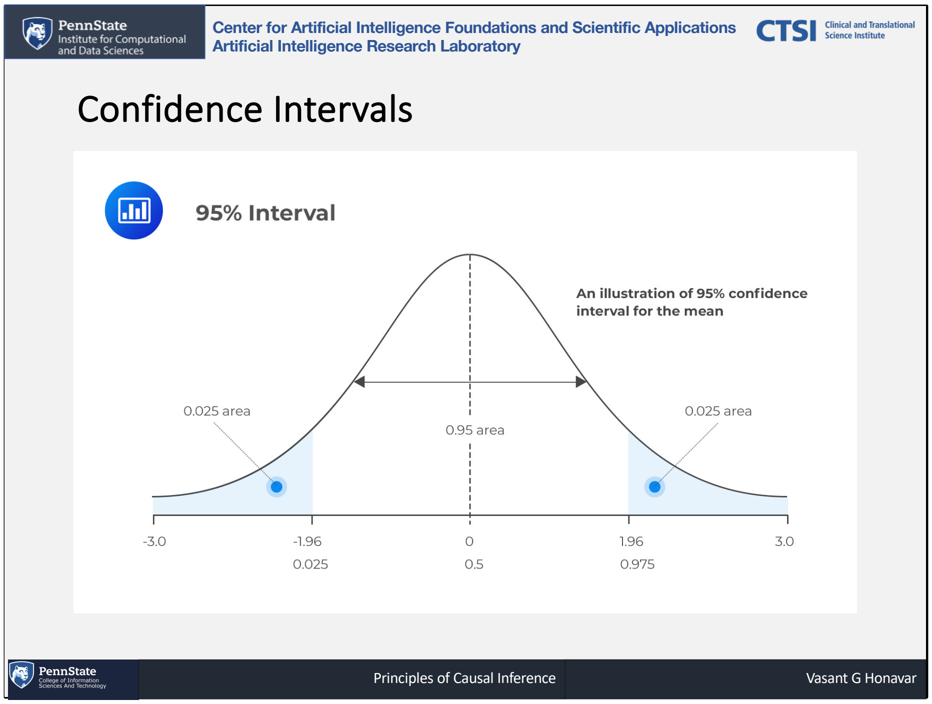
- Theorems 1 and 2 refer to sample estimates and hold for samples of any size
- However, Theorems 3 and 4 are about population estimates for which sample sizes must be large enough for the distribution of the estimator to be approximately normal
- Need larger N if the distribution is highly skewed, or some individuals are outliers or if some outcomes are rare



Confidence Intervals

$$\hat{\tau} \pm z^{*}\sqrt{\widehat{\text{var}}(\hat{\tau})}$$

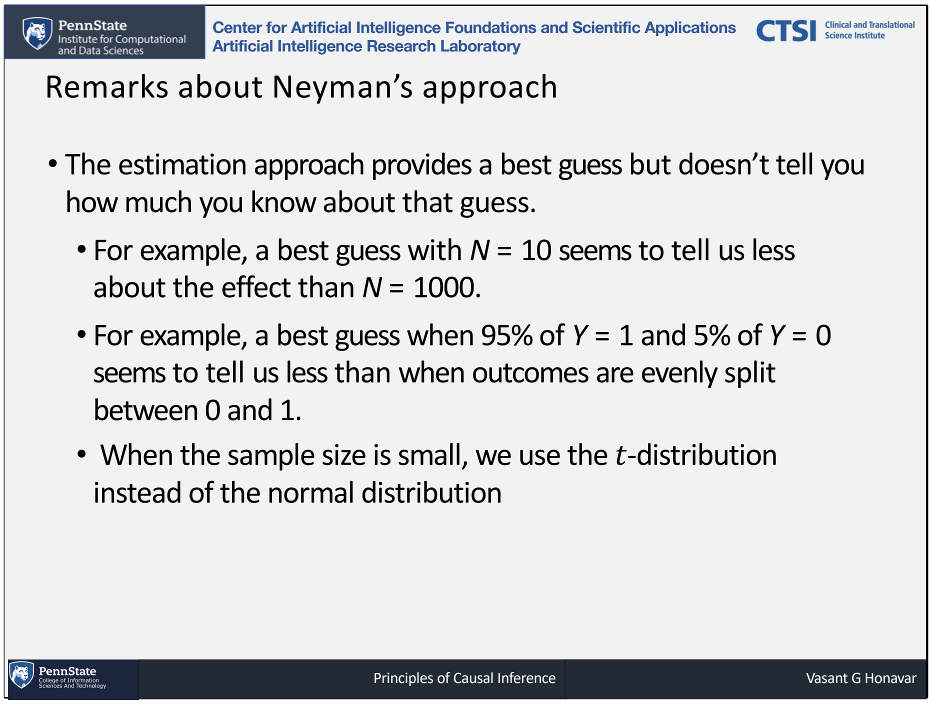
- $[-z^{*}, z^{*}]$ is the interval within which the desired probability mass falls in the standard normal distribution
- Confidence intervals due to Neyman!
- In the finite sample case, confidence interval may be too wide, and hence inference too conservative
- A 95% confidence interval will contain the estimand at least 95% of the time



Confidence Intervals

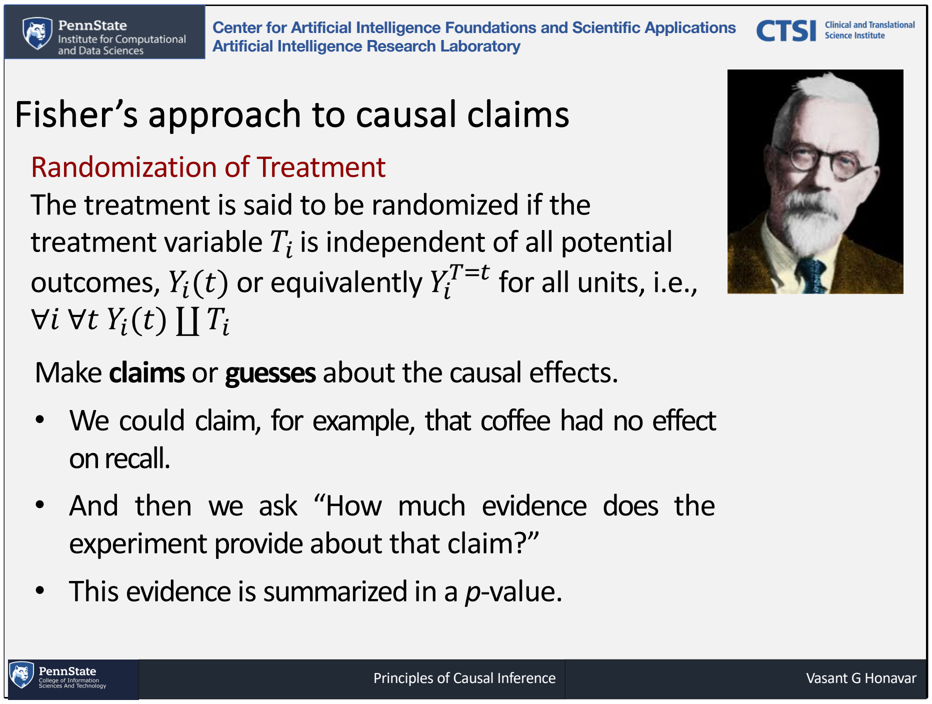
95% Interval

An illustration of 95% confidence interval for the mean



Remarks about Neyman’s approach

- The estimation approach provides a best guess but doesn’t tell you how much you know about that guess.
  - For example, a best guess with *N* = 10 seems to tell us less about the effect than *N* = 1000.
  - For example, a best guess when 95% of *Y* = 1 and 5% of *Y* = 0 seems to tell us less than when outcomes are evenly split between 0 and 1.
  - When the sample size is small, we use the t-distribution instead of the normal distribution



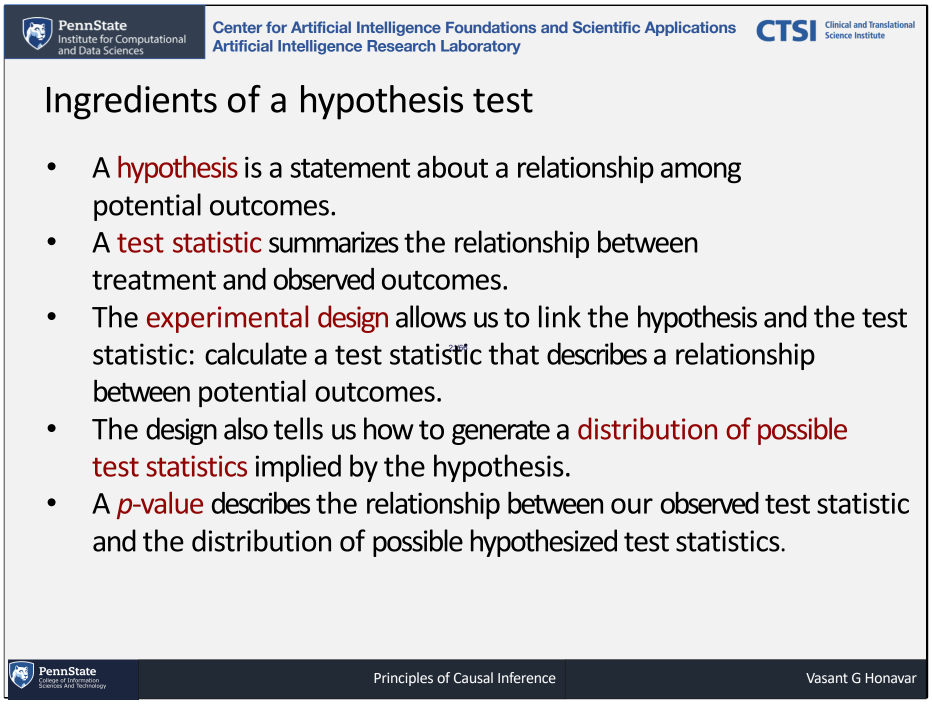
Fisher's approach to causal claims

Randomization of Treatment

The treatment is said to be randomized if the treatment variable T_i is independent of all potential outcomes, $Y_i(t)$ or equivalently $Y_i^{T=t}$ for all units, i.e., $\forall i\ \forall t\ Y_i(t) \coprod T_i$

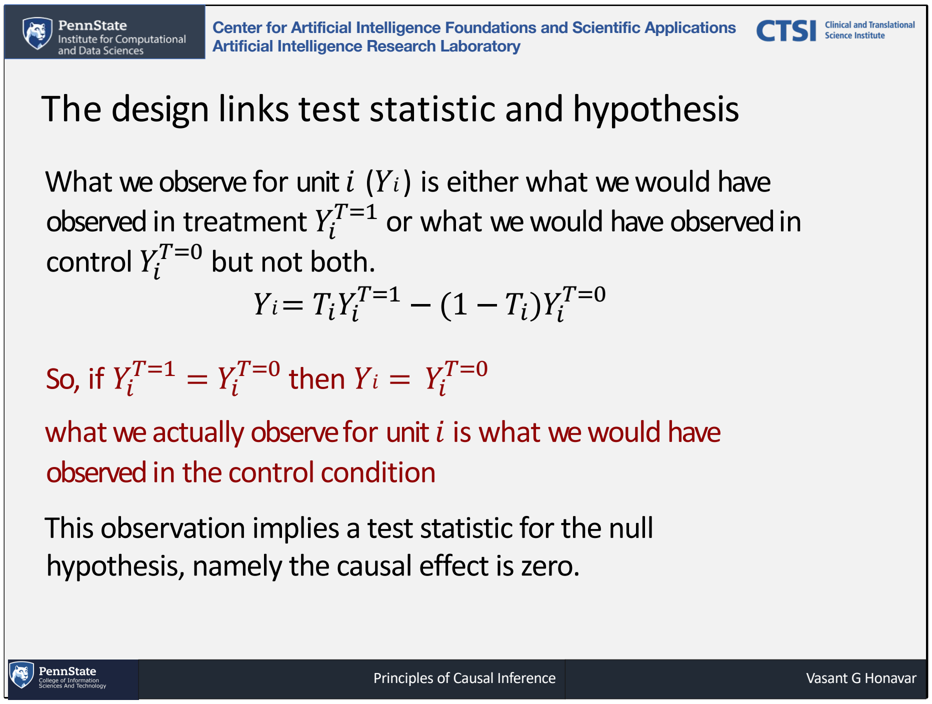
Make **claims** or **guesses** about the causal effects.

- We could claim, for example, that coffee had no effect on recall.
- And then we ask "How much evidence does the experiment provide about that claim?"
- This evidence is summarized in a *p*-value.



Ingredients of a hypothesis test

- A hypothesis is a statement about a relationship among potential outcomes.
- A test statistic summarizes the relationship between treatment and observed outcomes.
- The experimental design allows us to link the hypothesis and the test statistic: calculate a test statistic that describes a relationship between potential outcomes.
- The design also tells us how to generate a distribution of possible test statistics implied by the hypothesis.
- A *p*-value describes the relationship between our observed test statistic and the distribution of possible hypothesized test statistics.



The design links test statistic and hypothesis

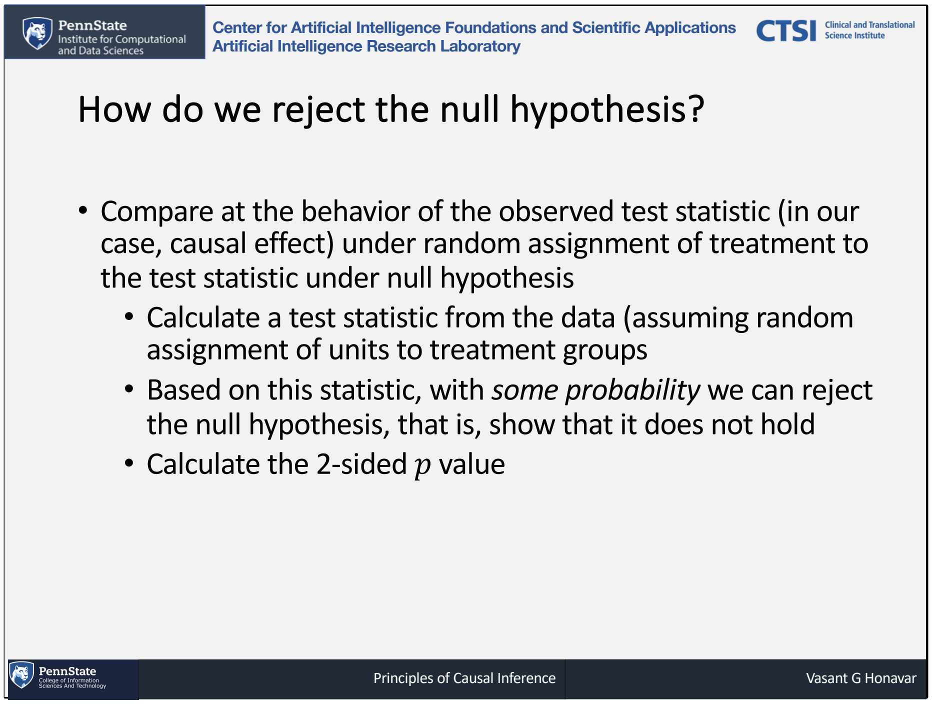
What we observe for unit i (Y_i) is either what we would have observed in treatment $Y_i^{T=1}$ or what we would have observed in control $Y_i^{T=0}$ but not both.

$$Y_i = T_i Y_i^{T=1} - (1 - T_i) Y_i^{T=0}$$

So, if $Y_i^{T=1} = Y_i^{T=0}$ then $Y_i = Y_i^{T=0}$

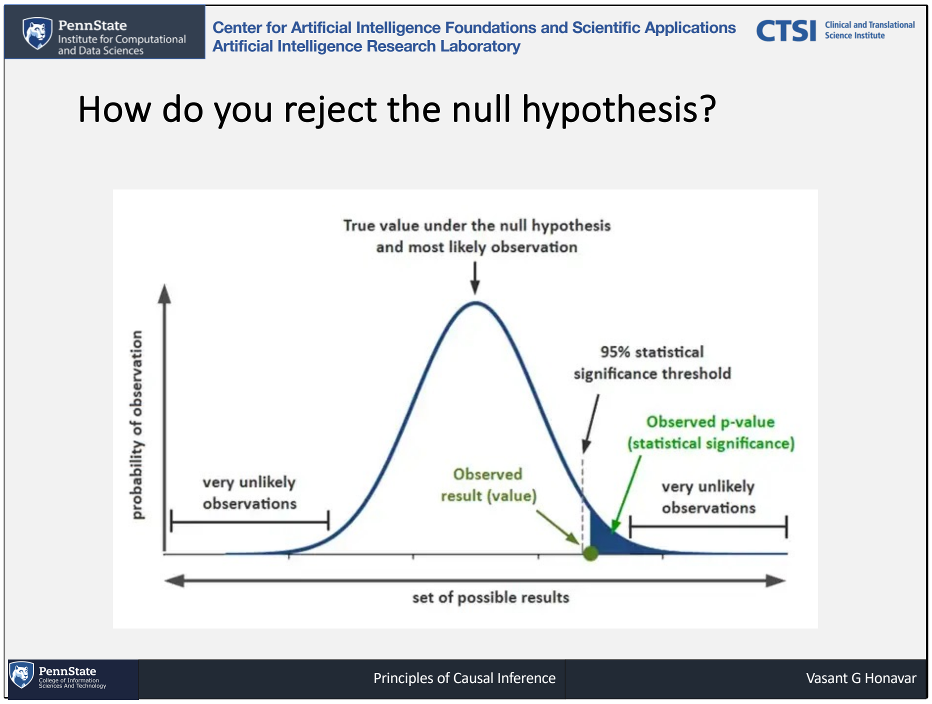
what we actually observe for unit i is what we would have observed in the control condition

This observation implies a test statistic for the null hypothesis, namely the causal effect is zero.



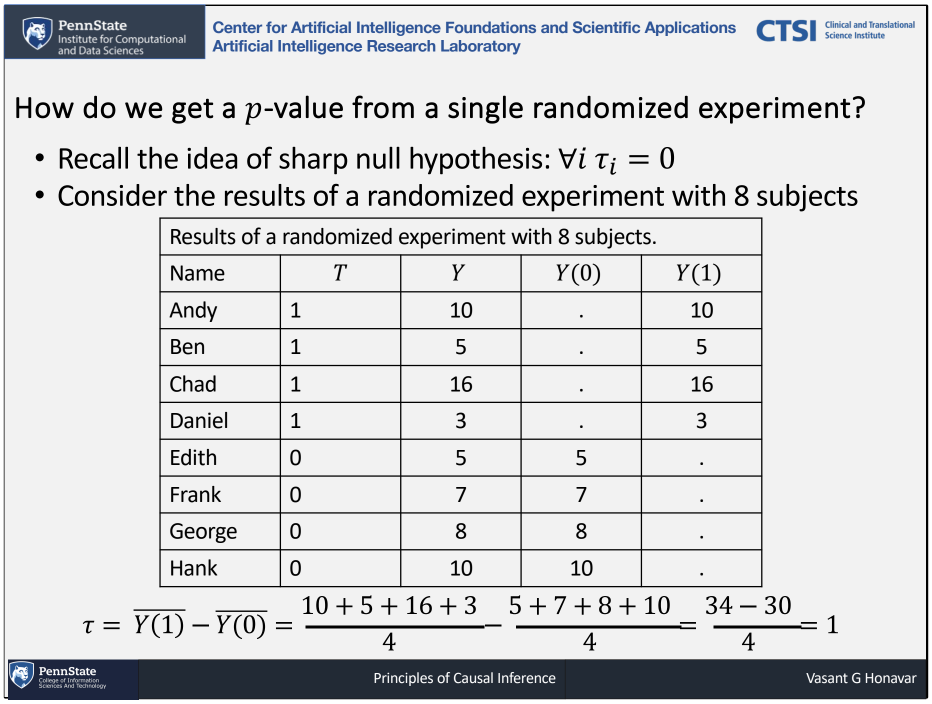
How do we reject the null hypothesis?

- Compare at the behavior of the observed test statistic (in our case, causal effect) under random assignment of treatment to the test statistic under null hypothesis
  - Calculate a test statistic from the data (assuming random assignment of units to treatment groups
  - Based on this statistic, with *some probability* we can reject the null hypothesis, that is, show that it does not hold
  - Calculate the 2-sided p value



How do you reject the null hypothesis?

True value under the null hypothesis and most likely observation
probability of observation
95% statistical significance threshold
Observed p-value (statistical significance)
very unlikely observations
Observed result (value)
very unlikely observations
set of possible results



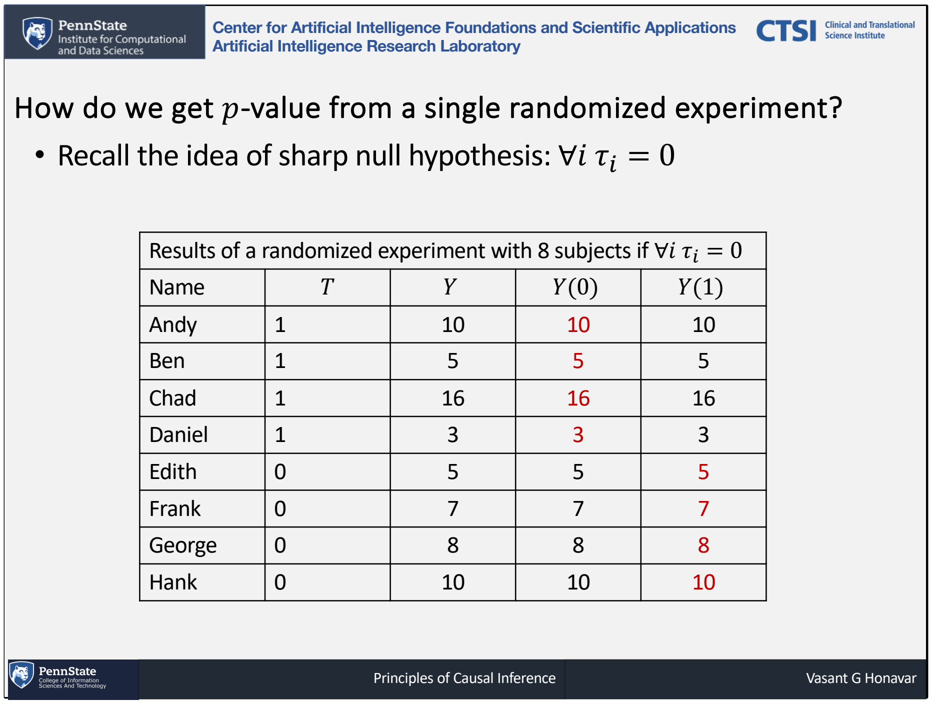
How do we get a p-value from a single randomized experiment?

- Recall the idea of sharp null hypothesis: $\forall i \; \tau_i = 0$
- Consider the results of a randomized experiment with 8 subjects

Results of a randomized experiment with 8 subjects.

Name	T	Y	$Y(0)$	$Y(1)$
Andy	1	10	.	10
Ben	1	5	.	5
Chad	1	16	.	16
Daniel	1	3	.	3
Edith	0	5	5	.
Frank	0	7	7	.
George	0	8	8	.
Hank	0	10	10	.

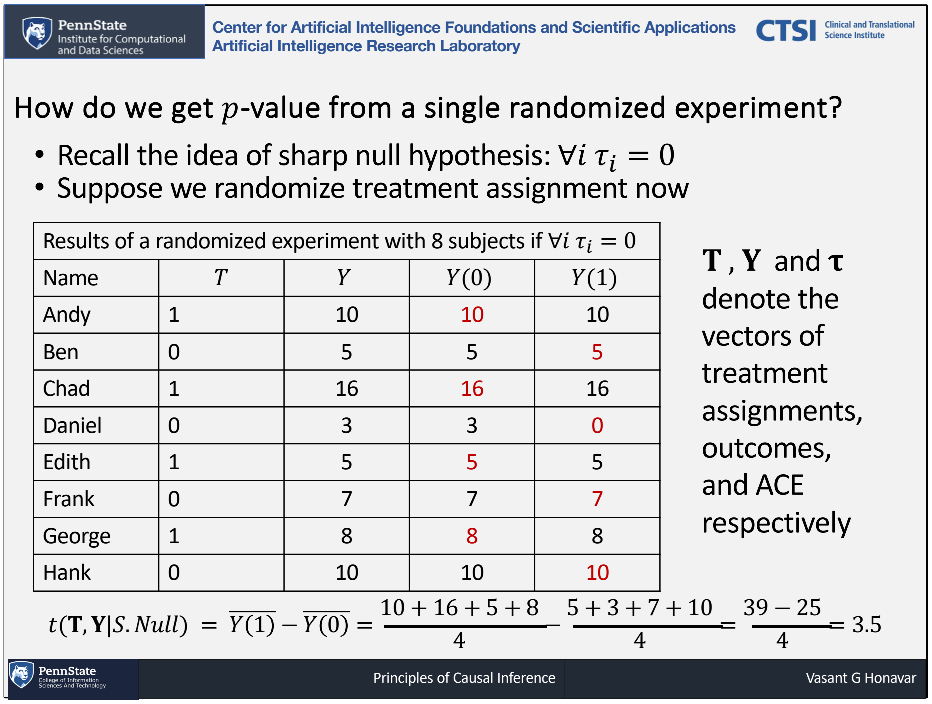
$$\tau = \overline{Y(1)} - \overline{Y(0)} = \frac{10+5+16+3}{4} - \frac{5+7+8+10}{4} = \frac{34-30}{4} = 1$$



How do we get p-value from a single randomized experiment?

- Recall the idea of sharp null hypothesis: $\forall i \ \tau_i = 0$

Results of a randomized experiment with 8 subjects if $\forall i \ \tau_i = 0$				
Name	T	Y	$Y(0)$	$Y(1)$
Andy	1	10	10	10
Ben	1	5	5	5
Chad	1	16	16	16
Daniel	1	3	3	3
Edith	0	5	5	5
Frank	0	7	7	7
George	0	8	8	8
Hank	0	10	10	10



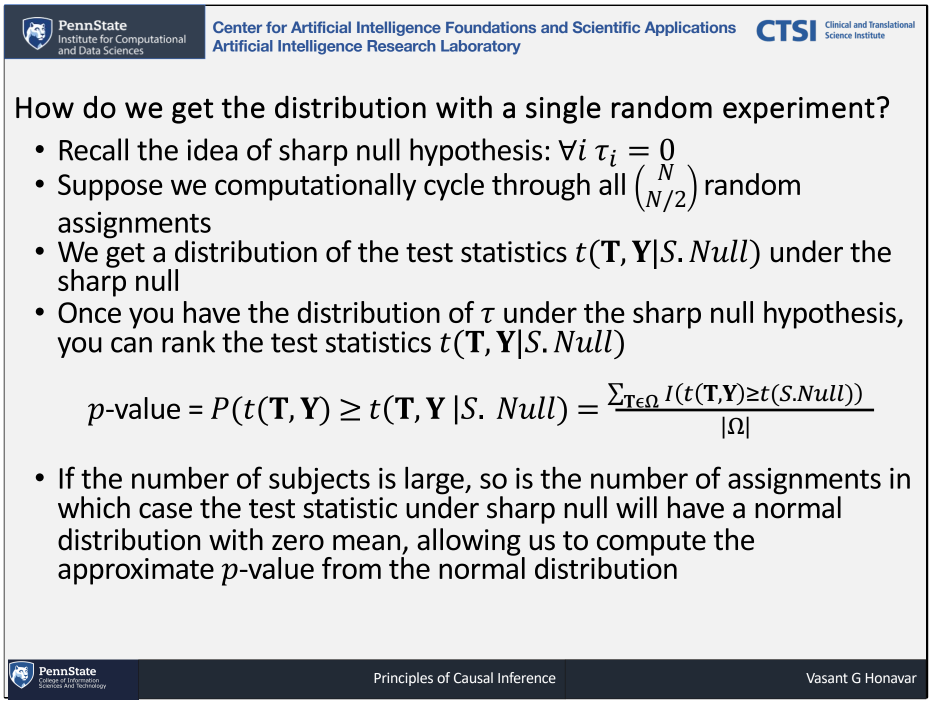
How do we get p-value from a single randomized experiment?

- Recall the idea of sharp null hypothesis: $\forall i\ \tau_i = 0$
- Suppose we randomize treatment assignment now

Results of a randomized experiment with 8 subjects if $\forall i\ \tau_i = 0$				
Name	T	Y	$Y(0)$	$Y(1)$
Andy	1	10	10	10
Ben	0	5	5	5
Chad	1	16	16	16
Daniel	0	3	3	0
Edith	1	5	5	5
Frank	0	7	7	7
George	1	8	8	8
Hank	0	10	10	10

$\mathbf{T}$, $\mathbf{Y}$ and $\boldsymbol{\tau}$ denote the vectors of treatment assignments, outcomes, and ACE respectively

$$t(\mathbf{T},\mathbf{Y}|S.Null) = \overline{Y(1)} - \overline{Y(0)} = \frac{10+16+5+8}{4} - \frac{5+3+7+10}{4} = \frac{39-25}{4} = 3.5$$

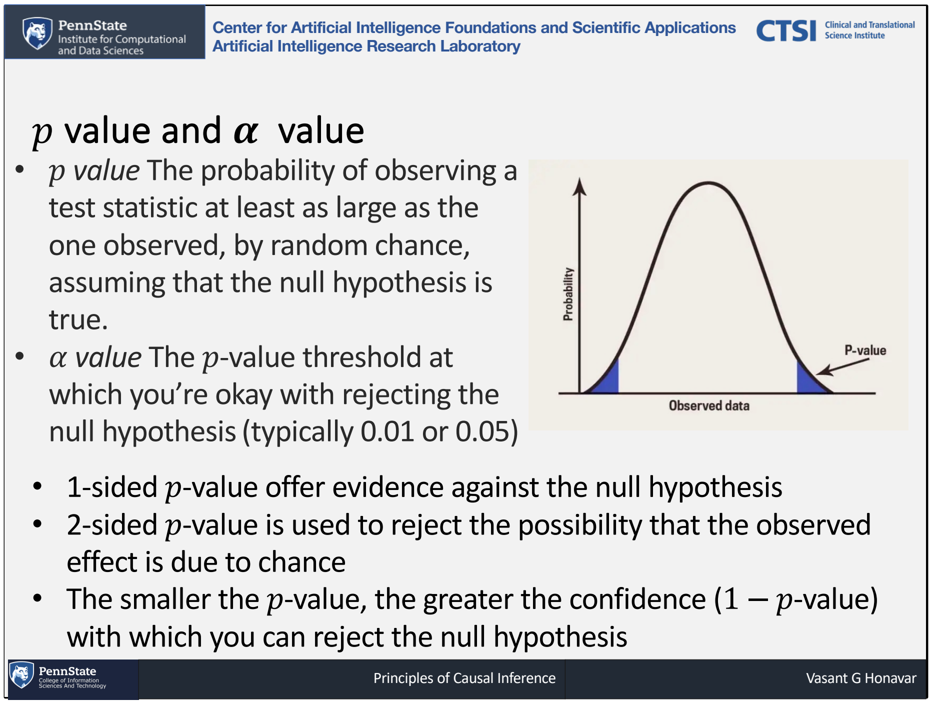


How do we get the distribution with a single random experiment?

- Recall the idea of sharp null hypothesis: $\forall i\ \tau_i = 0$
- Suppose we computationally cycle through all $\binom{N}{N/2}$ random assignments
- We get a distribution of the test statistics $t(\mathbf{T}, \mathbf{Y}|S.Null)$ under the sharp null
- Once you have the distribution of τ under the sharp null hypothesis, you can rank the test statistics $t(\mathbf{T}, \mathbf{Y}|S.Null)$

$$p\text{-value} = P(t(\mathbf{T}, \mathbf{Y}) \geq t(\mathbf{T}, \mathbf{Y}\ |S.\ Null) = \frac{\sum_{\mathbf{T}\epsilon\Omega} I(t(\mathbf{T},\mathbf{Y}) \geq t(S.Null))}{|\Omega|}$$

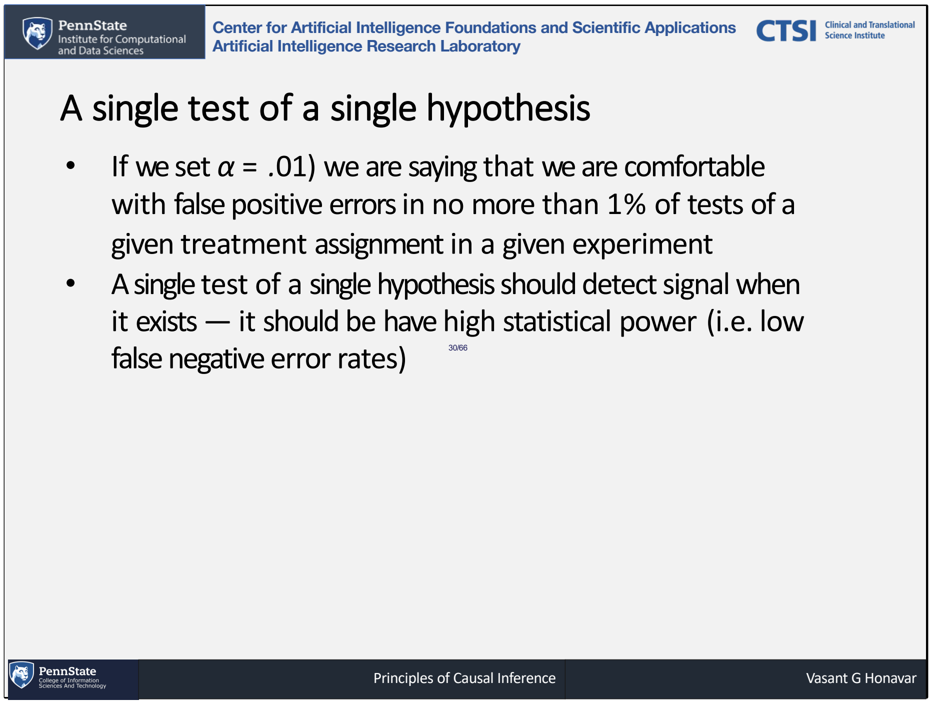
- If the number of subjects is large, so is the number of assignments in which case the test statistic under sharp null will have a normal distribution with zero mean, allowing us to compute the approximate p-value from the normal distribution



p value and $\boldsymbol{\alpha}$ value

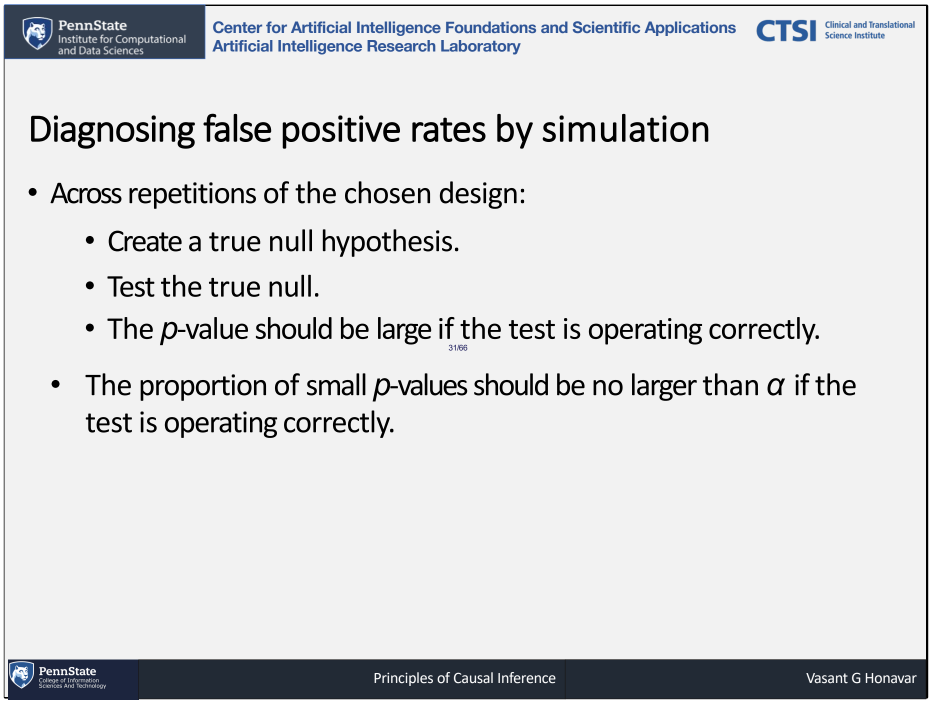
- p *value* The probability of observing a test statistic at least as large as the one observed, by random chance, assuming that the null hypothesis is true.
- α *value* The p-value threshold at which you're okay with rejecting the null hypothesis (typically 0.01 or 0.05)

- 1-sided p-value offer evidence against the null hypothesis
- 2-sided p-value is used to reject the possibility that the observed effect is due to chance
- The smaller the p-value, the greater the confidence $(1 - p$-value) with which you can reject the null hypothesis



A single test of a single hypothesis

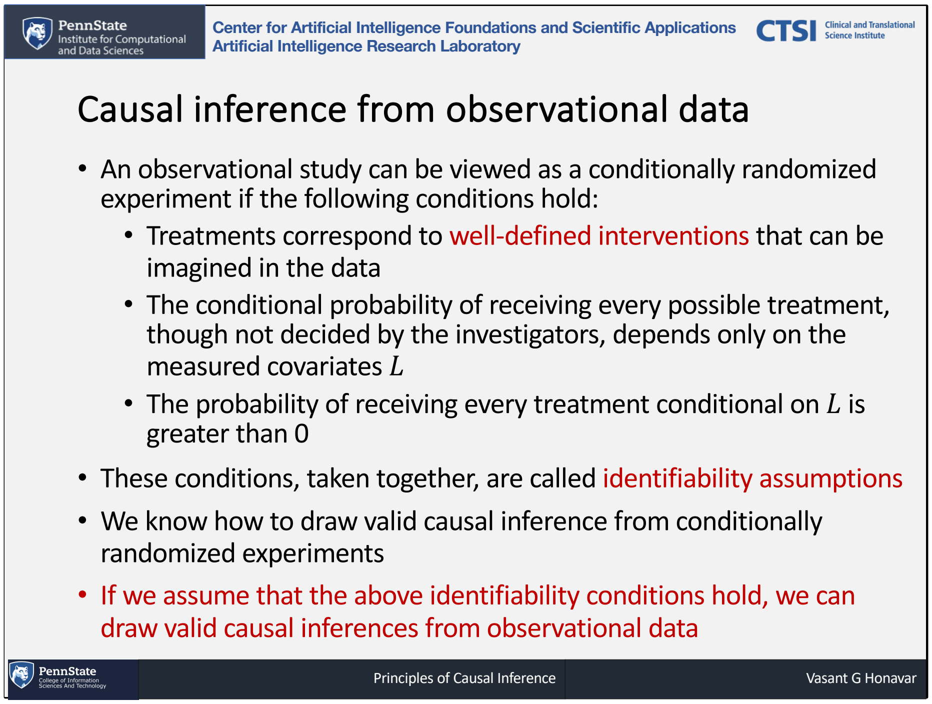
- If we set α = .01) we are saying that we are comfortable with false positive errors in no more than 1% of tests of a given treatment assignment in a given experiment
- A single test of a single hypothesis should detect signal when it exists — it should be have high statistical power (i.e. low false negative error rates) 30/66



Diagnosing false positive rates by simulation

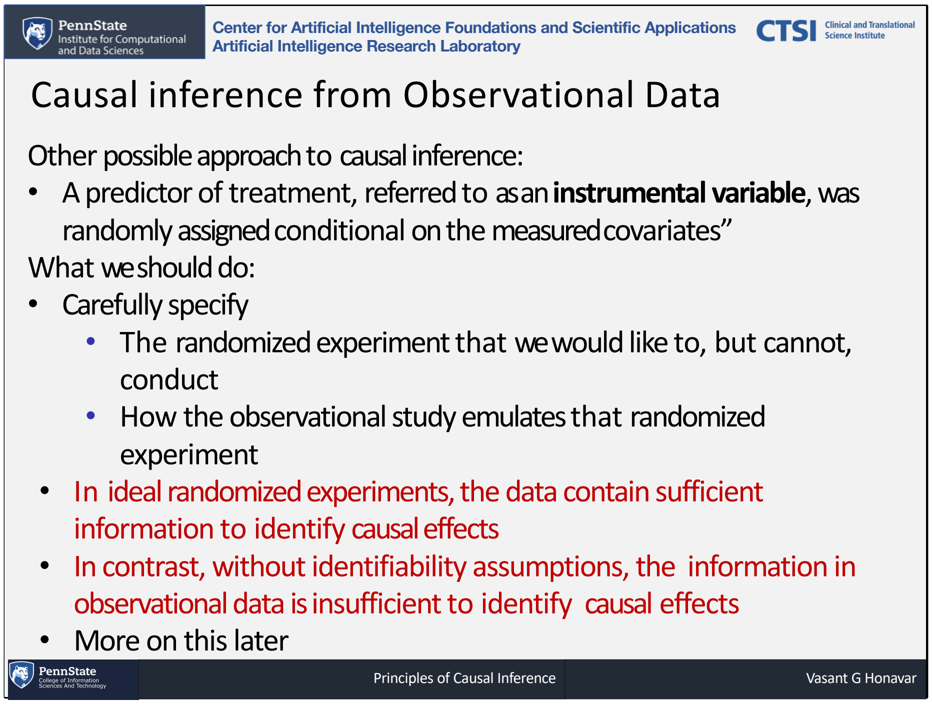
- Across repetitions of the chosen design:
  - Create a true null hypothesis.
  - Test the true null.
  - The *p*-value should be large if the test is operating correctly.

- The proportion of small *p*-values should be no larger than α if the test is operating correctly.



Causal inference from observational data

- An observational study can be viewed as a conditionally randomized experiment if the following conditions hold:
  - Treatments correspond to well-defined interventions that can be imagined in the data
  - The conditional probability of receiving every possible treatment, though not decided by the investigators, depends only on the measured covariates L
  - The probability of receiving every treatment conditional on L is greater than 0
- These conditions, taken together, are called identifiability assumptions
- We know how to draw valid causal inference from conditionally randomized experiments
- If we assume that the above identifiability conditions hold, we can draw valid causal inferences from observational data



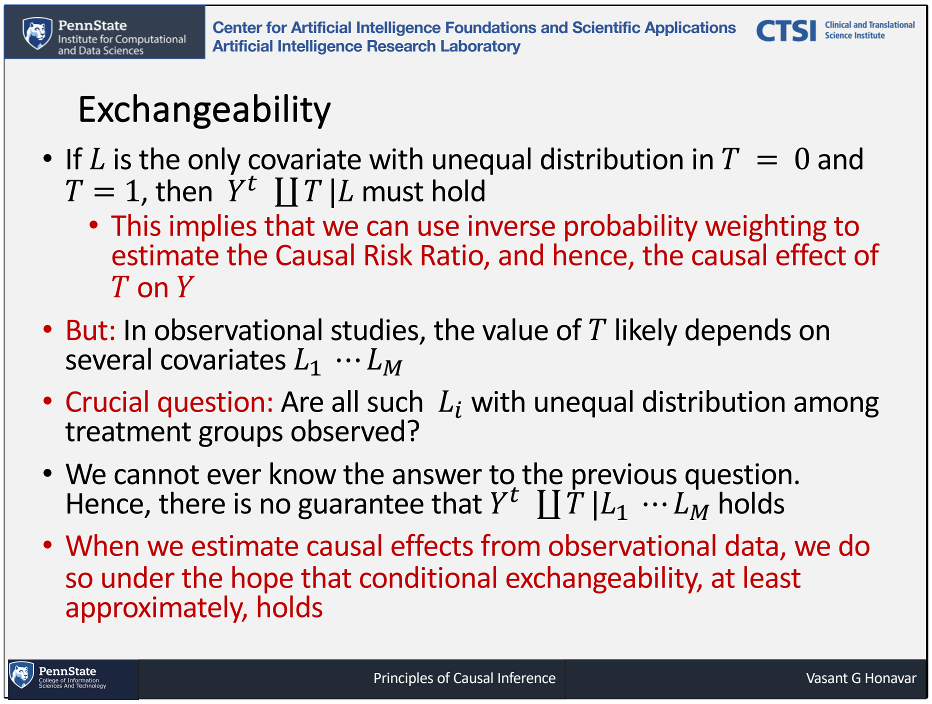
Causal inference from Observational Data

Other possible approach to causal inference:

- A predictor of treatment, referred to as an **instrumental variable**, was randomly assigned conditional on the measured covariates”

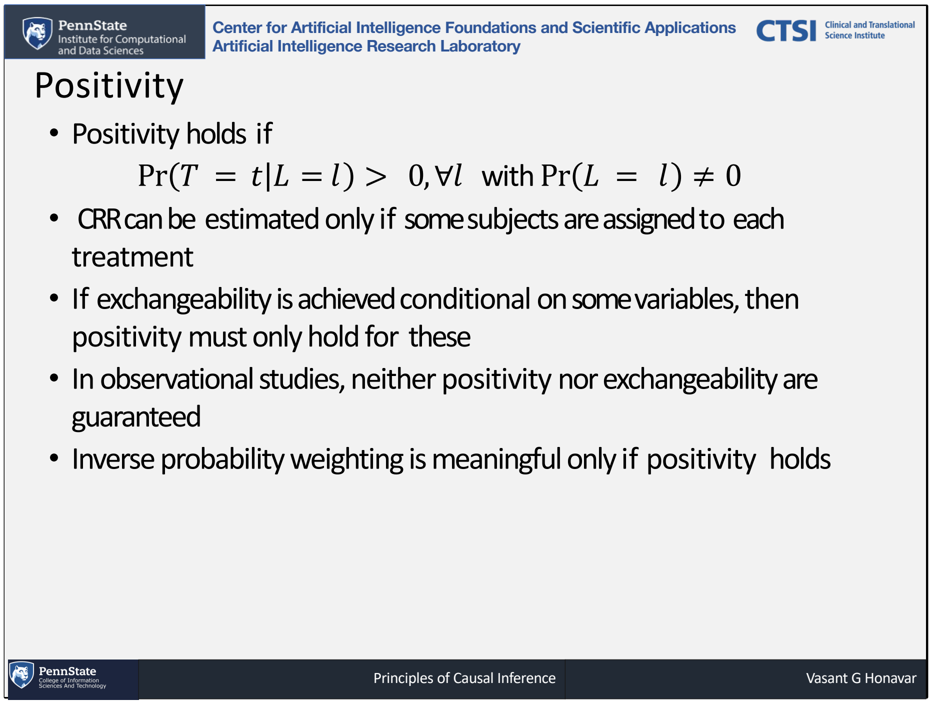
What we should do:

- Carefully specify
  - The randomized experiment that we would like to, but cannot, conduct
  - How the observational study emulates that randomized experiment
- In ideal randomized experiments, the data contain sufficient information to identify causal effects
- In contrast, without identifiability assumptions, the information in observational data is insufficient to identify causal effects
- More on this later



Exchangeability

- If L is the only covariate with unequal distribution in $T = 0$ and $T = 1$, then $Y^t \coprod T \mid L$ must hold
  - This implies that we can use inverse probability weighting to estimate the Causal Risk Ratio, and hence, the causal effect of T on Y
- But: In observational studies, the value of T likely depends on several covariates $L_1 \cdots L_M$
- Crucial question: Are all such L_i with unequal distribution among treatment groups observed?
- We cannot ever know the answer to the previous question. Hence, there is no guarantee that $Y^t \coprod T \mid L_1 \cdots L_M$ holds
- When we estimate causal effects from observational data, we do so under the hope that conditional exchangeability, at least approximately, holds

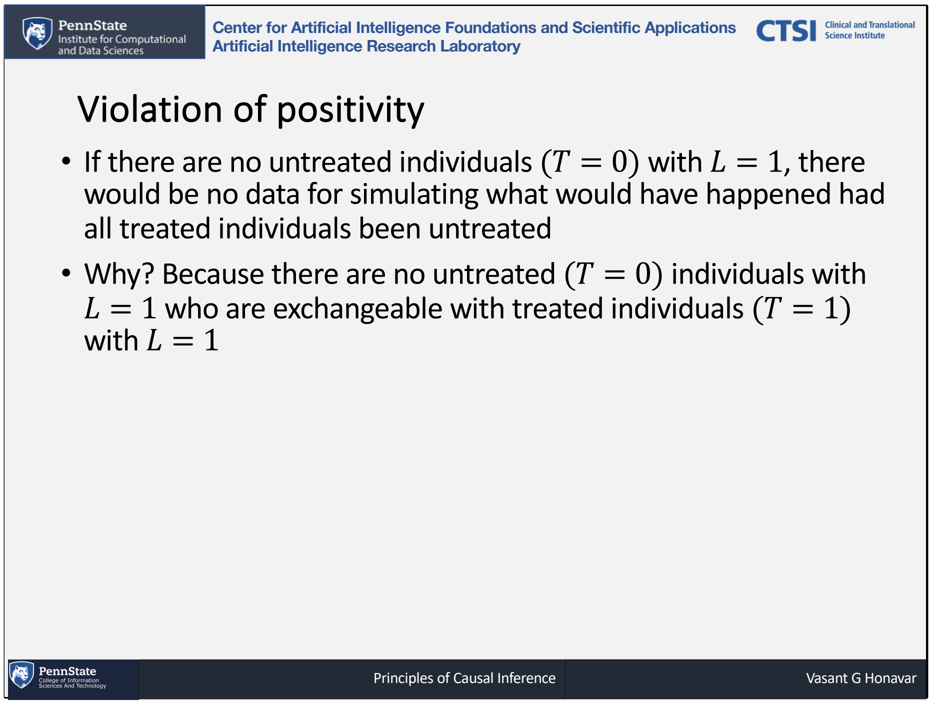


Positivity

- Positivity holds if

$$\Pr(T = t | L = l) > 0, \forall l \text{ with } \Pr(L = l) \neq 0$$

- CRR can be estimated only if some subjects are assigned to each treatment
- If exchangeability is achieved conditional on some variables, then positivity must only hold for these
- In observational studies, neither positivity nor exchangeability are guaranteed
- Inverse probability weighting is meaningful only if positivity holds

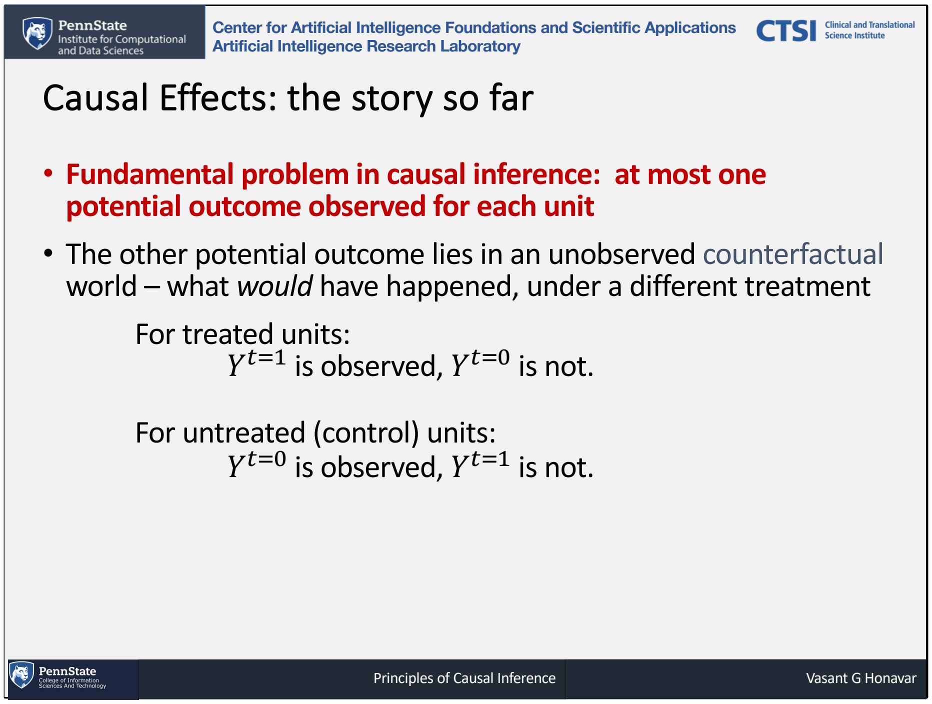


Violation of positivity

- If there are no untreated individuals $(T = 0)$ with $L = 1$, there would be no data for simulating what would have happened had all treated individuals been untreated
- Why? Because there are no untreated $(T = 0)$ individuals with $L = 1$ who are exchangeable with treated individuals $(T = 1)$ with $L = 1$

Consistency

- Consistency requires that $Y^t = Y$ for every individual with $T = t$
  - The observed outcome for every treated (untreated) individual equal her outcome had she been treated (not treated)

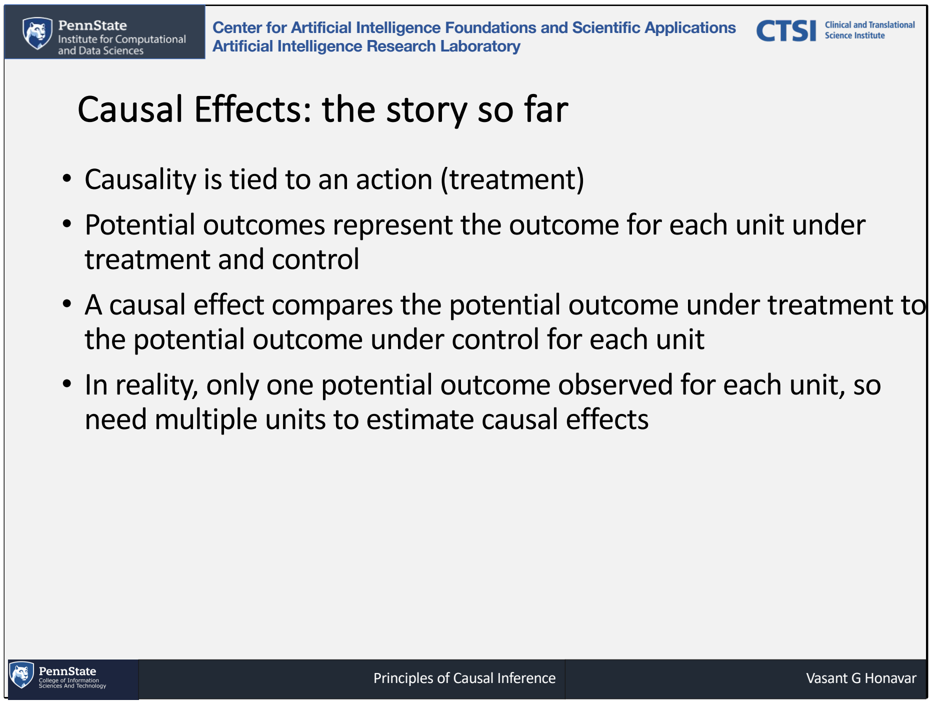


Causal Effects: the story so far

- **Fundamental problem in causal inference: at most one potential outcome observed for each unit**
- The other potential outcome lies in an unobserved counterfactual world – what *would* have happened, under a different treatment

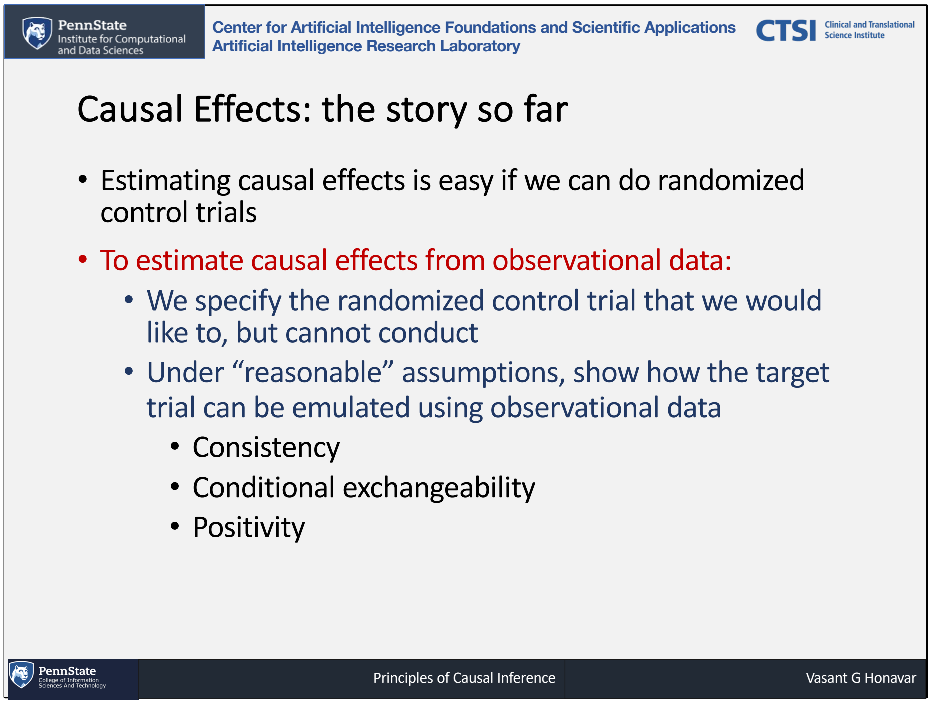
For treated units:
$Y^{t=1}$ is observed, $Y^{t=0}$ is not.

For untreated (control) units:
$Y^{t=0}$ is observed, $Y^{t=1}$ is not.



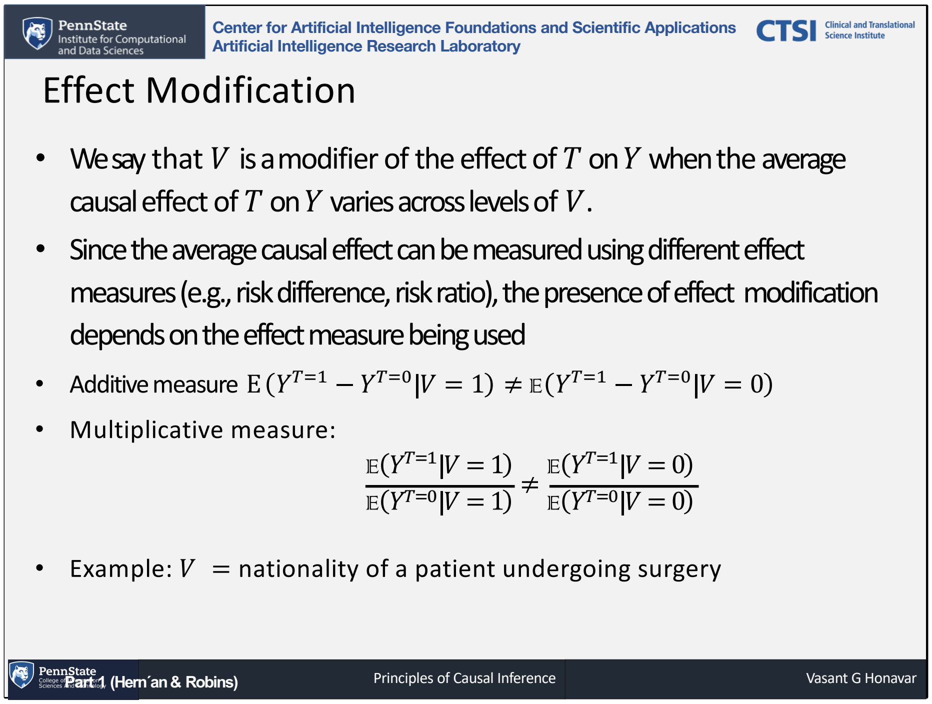
Causal Effects: the story so far

- Causality is tied to an action (treatment)
- Potential outcomes represent the outcome for each unit under treatment and control
- A causal effect compares the potential outcome under treatment to the potential outcome under control for each unit
- In reality, only one potential outcome observed for each unit, so need multiple units to estimate causal effects



Causal Effects: the story so far

- Estimating causal effects is easy if we can do randomized control trials
- To estimate causal effects from observational data:
  - We specify the randomized control trial that we would like to, but cannot conduct
  - Under "reasonable" assumptions, show how the target trial can be emulated using observational data
    - Consistency
    - Conditional exchangeability
    - Positivity

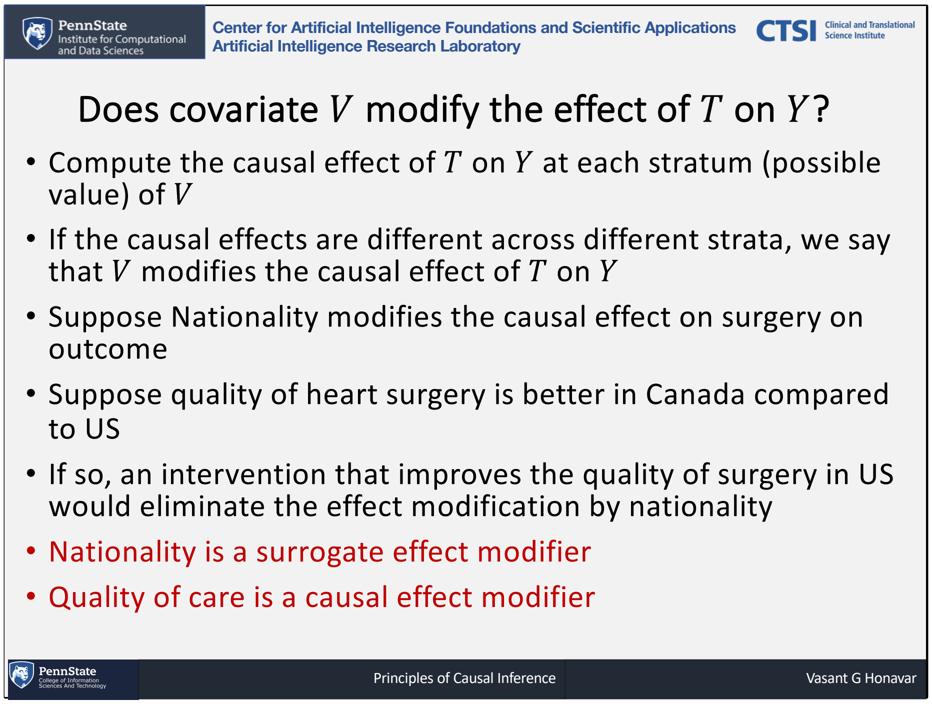


Effect Modification

- We say that V is a modifier of the effect of T on Y when the average causal effect of T on Y varies across levels of V.
- Since the average causal effect can be measured using different effect measures (e.g., risk difference, risk ratio), the presence of effect modification depends on the effect measure being used
- Additive measure $\mathrm{E}\,(Y^{T=1} - Y^{T=0}|V = 1) \neq \mathbb{E}(Y^{T=1} - Y^{T=0}|V = 0)$
- Multiplicative measure:

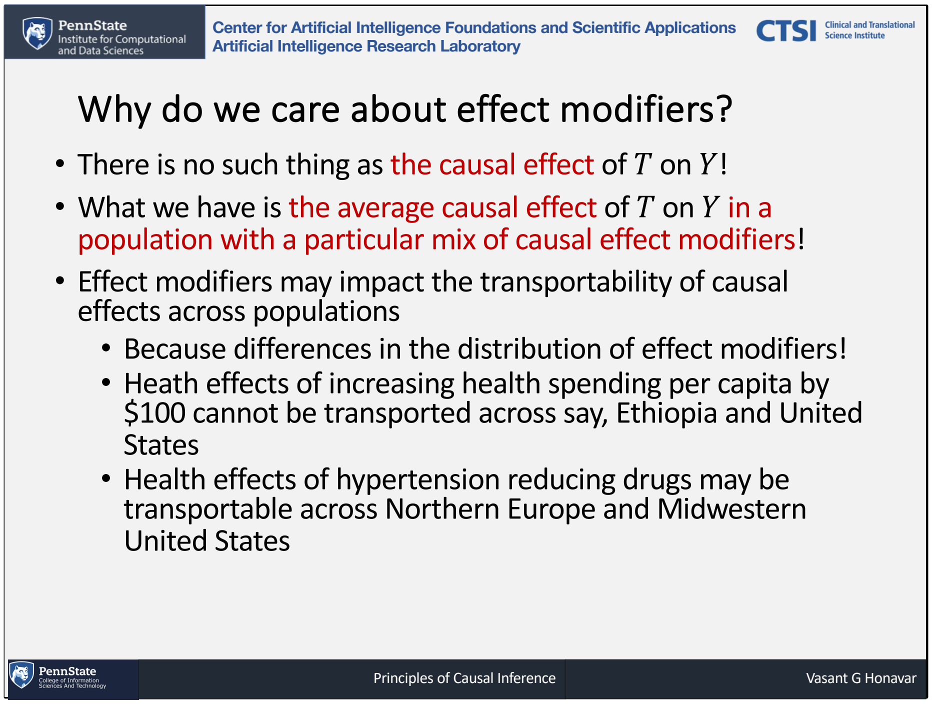
$$\frac{\mathbb{E}(Y^{T=1}|V=1)}{\mathbb{E}(Y^{T=0}|V=1)} \neq \frac{\mathbb{E}(Y^{T=1}|V=0)}{\mathbb{E}(Y^{T=0}|V=0)}$$

- Example: V = nationality of a patient undergoing surgery



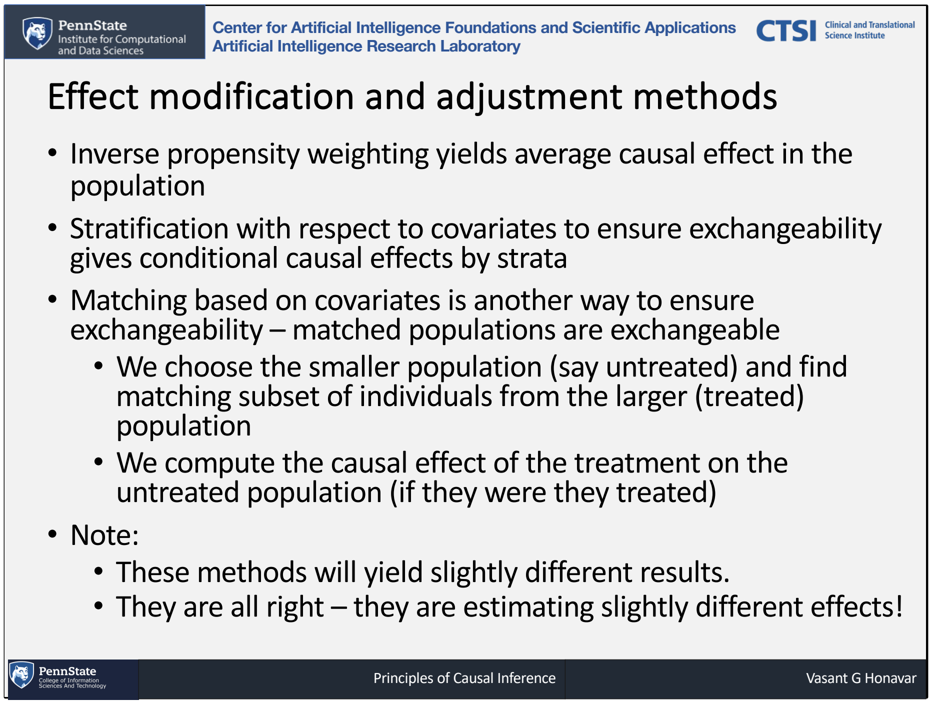
Does covariate V modify the effect of T on Y?

- Compute the causal effect of T on Y at each stratum (possible value) of V
- If the causal effects are different across different strata, we say that V modifies the causal effect of T on Y
- Suppose Nationality modifies the causal effect on surgery on outcome
- Suppose quality of heart surgery is better in Canada compared to US
- If so, an intervention that improves the quality of surgery in US would eliminate the effect modification by nationality
- Nationality is a surrogate effect modifier
- Quality of care is a causal effect modifier



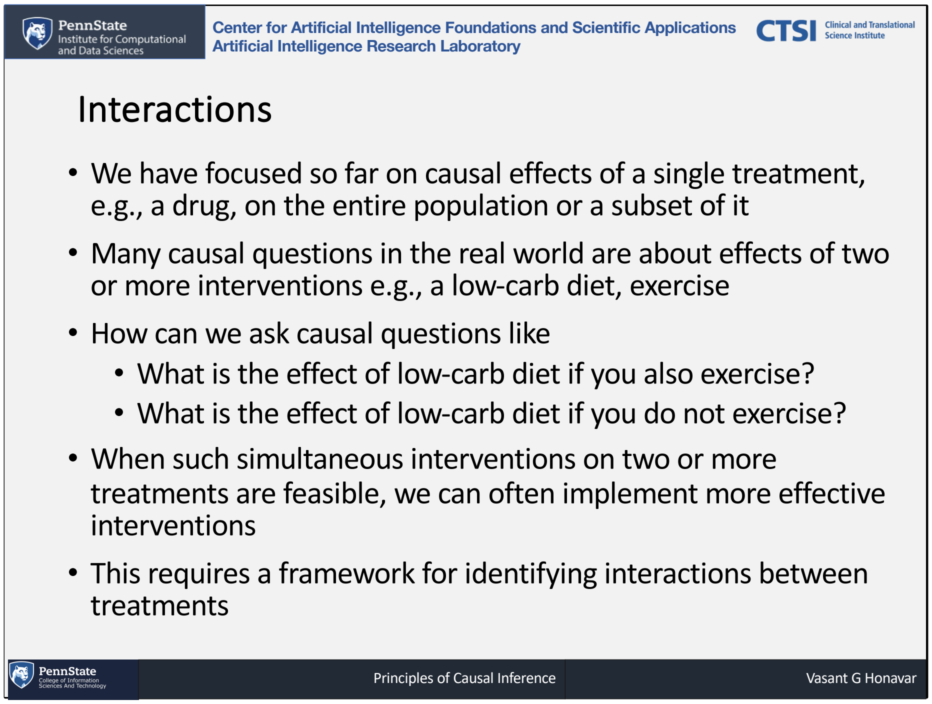
Why do we care about effect modifiers?

- There is no such thing as the causal effect of T on Y!
- What we have is the average causal effect of T on Y in a population with a particular mix of causal effect modifiers!
- Effect modifiers may impact the transportability of causal effects across populations
  - Because differences in the distribution of effect modifiers!
  - Heath effects of increasing health spending per capita by $100 cannot be transported across say, Ethiopia and United States
  - Health effects of hypertension reducing drugs may be transportable across Northern Europe and Midwestern United States



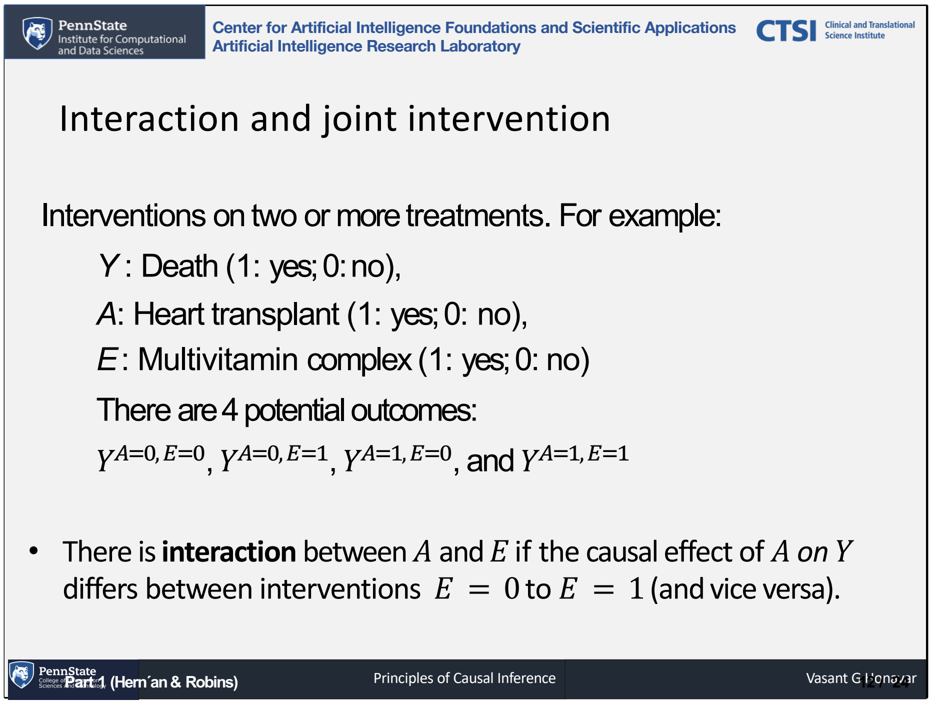
Effect modification and adjustment methods

- Inverse propensity weighting yields average causal effect in the population
- Stratification with respect to covariates to ensure exchangeability gives conditional causal effects by strata
- Matching based on covariates is another way to ensure exchangeability – matched populations are exchangeable
  - We choose the smaller population (say untreated) and find matching subset of individuals from the larger (treated) population
  - We compute the causal effect of the treatment on the untreated population (if they were they treated)
- Note:
  - These methods will yield slightly different results.
  - They are all right – they are estimating slightly different effects!



Interactions

- We have focused so far on causal effects of a single treatment, e.g., a drug, on the entire population or a subset of it
- Many causal questions in the real world are about effects of two or more interventions e.g., a low-carb diet, exercise
- How can we ask causal questions like
  - What is the effect of low-carb diet if you also exercise?
  - What is the effect of low-carb diet if you do not exercise?
- When such simultaneous interventions on two or more treatments are feasible, we can often implement more effective interventions
- This requires a framework for identifying interactions between treatments



Interaction and joint intervention

Interventions on two or more treatments. For example:

Y : Death (1: yes; 0: no),

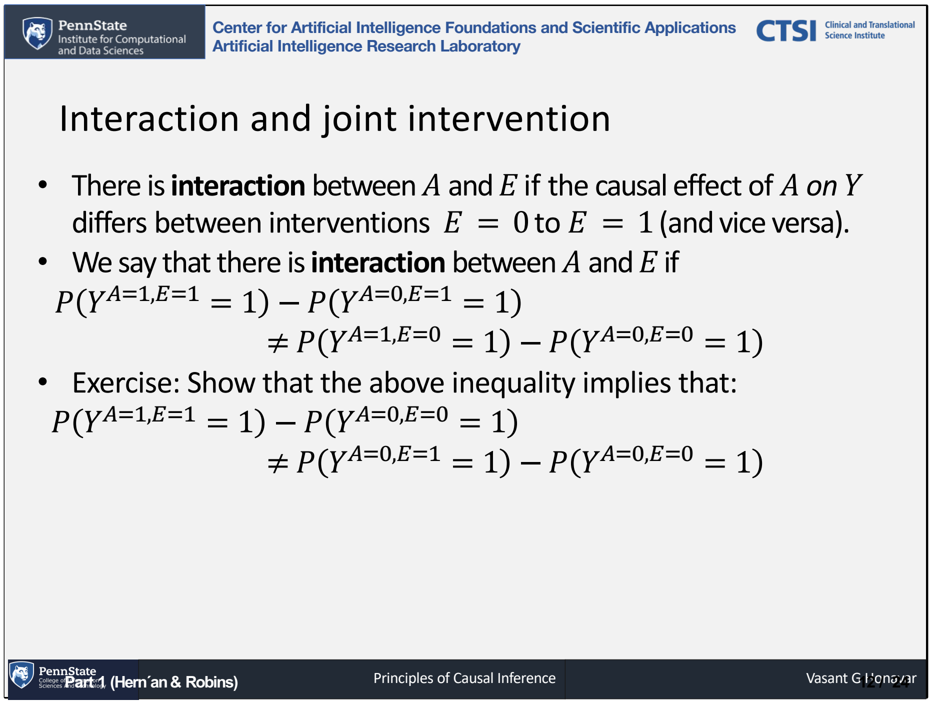
A: Heart transplant (1: yes; 0: no),

E: Multivitamin complex (1: yes; 0: no)

There are 4 potential outcomes:

$Y^{A=0,E=0}$, $Y^{A=0,E=1}$, $Y^{A=1,E=0}$, and $Y^{A=1,E=1}$

- There is **interaction** between A and E if the causal effect of A *on* Y differs between interventions $E = 0$ to $E = 1$ (and vice versa).



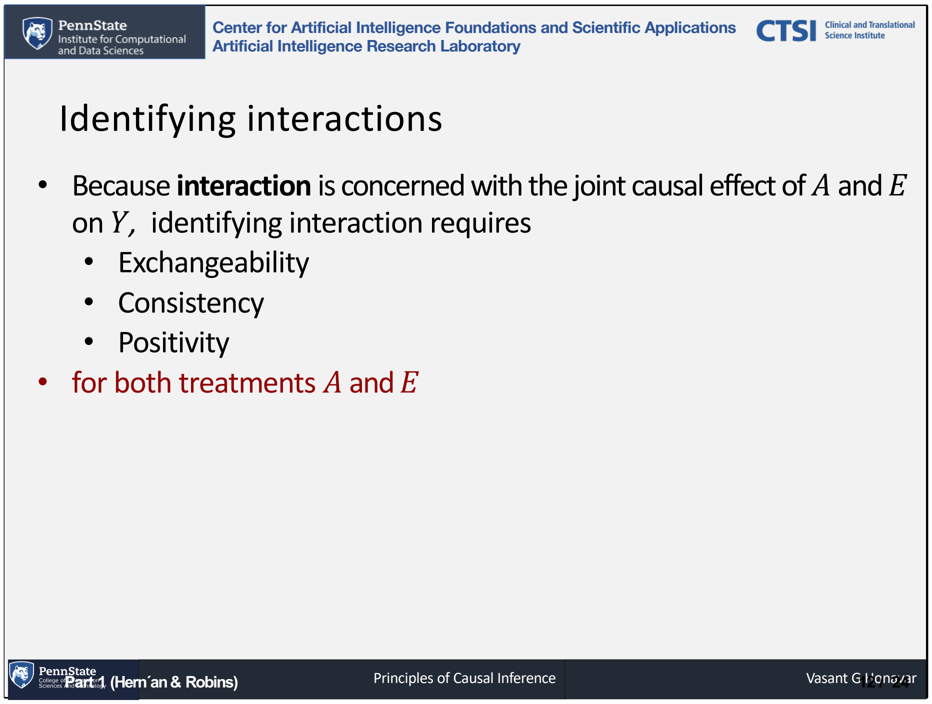
Interaction and joint intervention

- There is **interaction** between A and E if the causal effect of A *on* Y differs between interventions $E = 0$ to $E = 1$ (and vice versa).
- We say that there is **interaction** between A and E if

$$P(Y^{A=1,E=1} = 1) - P(Y^{A=0,E=1} = 1) \neq P(Y^{A=1,E=0} = 1) - P(Y^{A=0,E=0} = 1)$$

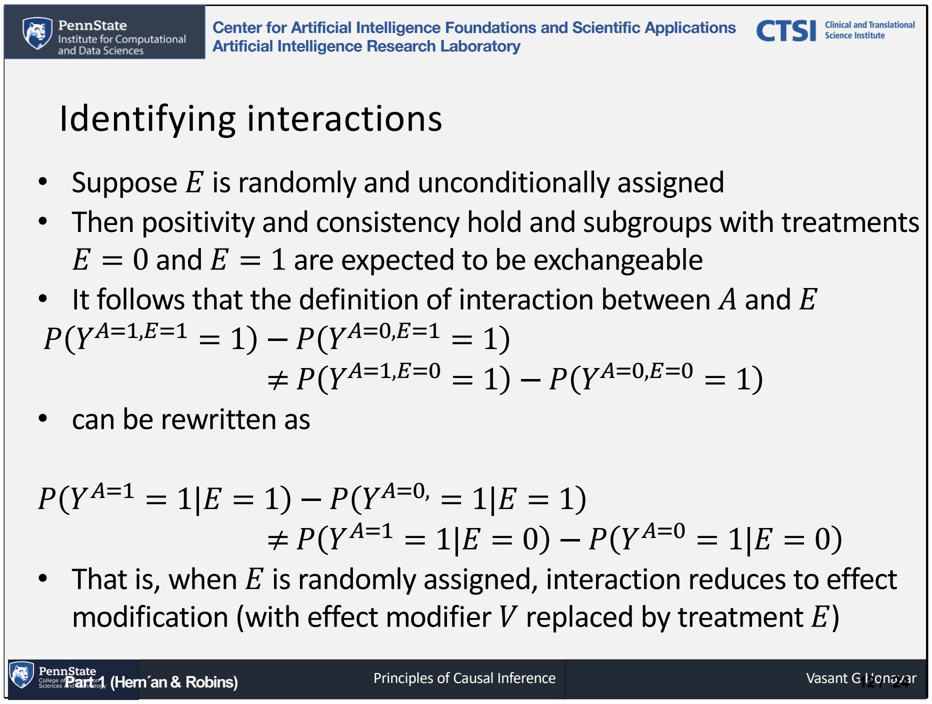
- Exercise: Show that the above inequality implies that:

$$P(Y^{A=1,E=1} = 1) - P(Y^{A=0,E=0} = 1) \neq P(Y^{A=0,E=1} = 1) - P(Y^{A=0,E=0} = 1)$$



Identifying interactions

- Because **interaction** is concerned with the joint causal effect of A and E on Y, identifying interaction requires
  - Exchangeability
  - Consistency
  - Positivity
- for both treatments A and E



Identifying interactions

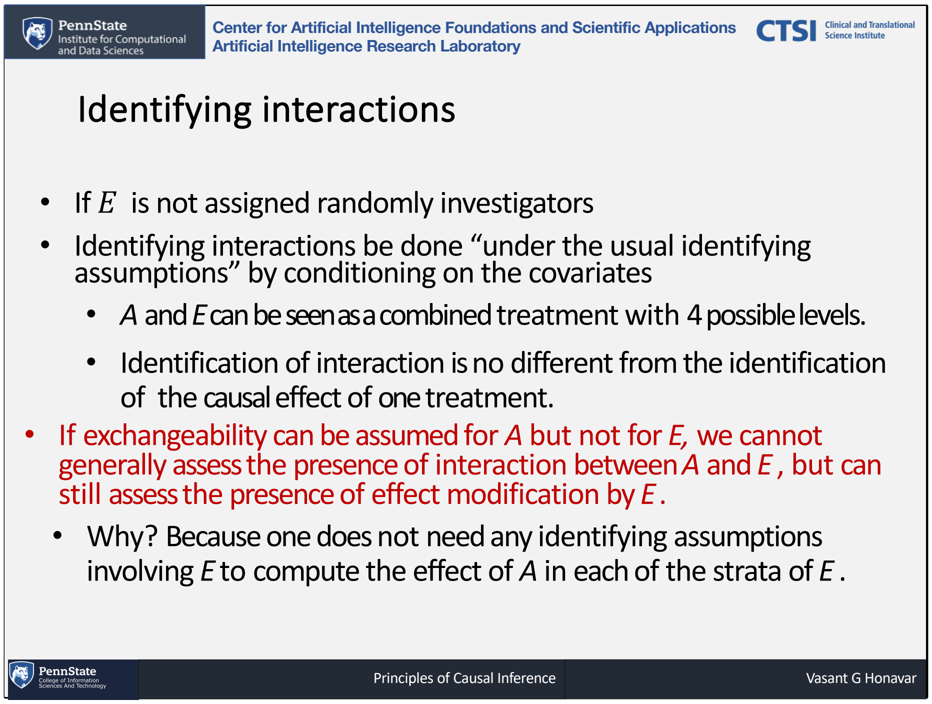
- Suppose E is randomly and unconditionally assigned
- Then positivity and consistency hold and subgroups with treatments $E = 0$ and $E = 1$ are expected to be exchangeable
- It follows that the definition of interaction between A and E

$$P(Y^{A=1,E=1} = 1) - P(Y^{A=0,E=1} = 1) \neq P(Y^{A=1,E=0} = 1) - P(Y^{A=0,E=0} = 1)$$

- can be rewritten as

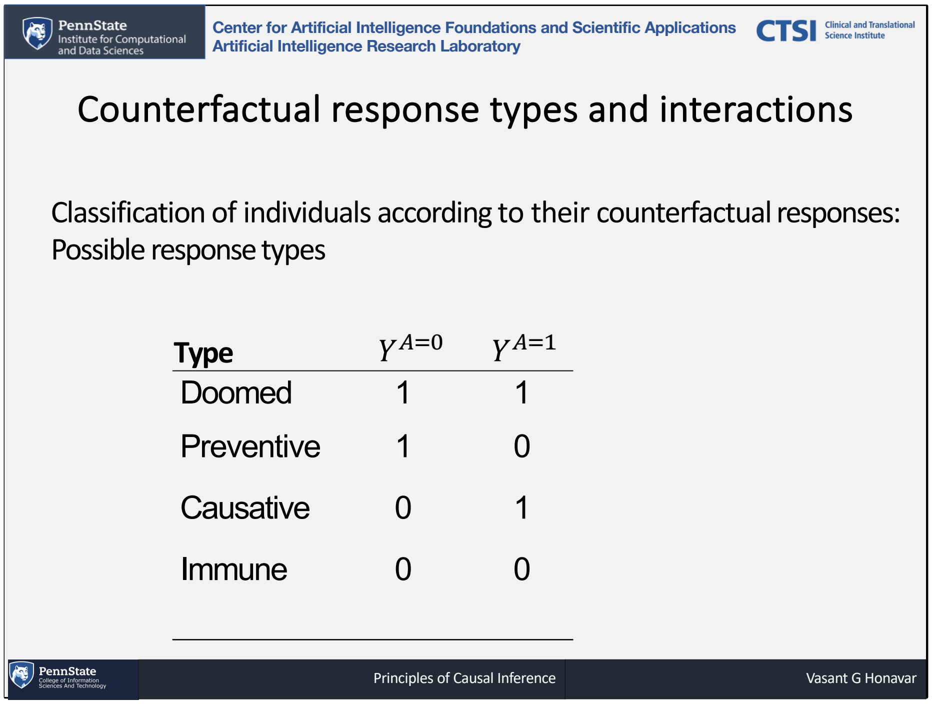
$$P(Y^{A=1} = 1|E = 1) - P(Y^{A=0,} = 1|E = 1) \neq P(Y^{A=1} = 1|E = 0) - P(Y^{A=0} = 1|E = 0)$$

- That is, when E is randomly assigned, interaction reduces to effect modification (with effect modifier V replaced by treatment E)



Identifying interactions

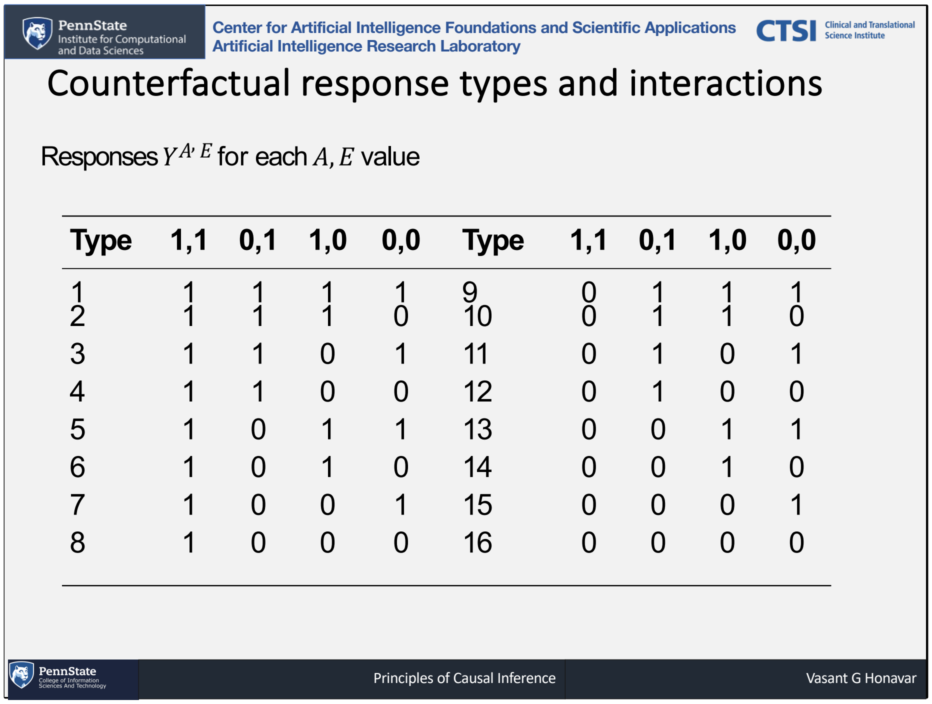
- If E is not assigned randomly investigators
- Identifying interactions be done “under the usual identifying assumptions” by conditioning on the covariates
  - *A* and *E* can be seen as a combined treatment with 4 possible levels.
  - Identification of interaction is no different from the identification of the causal effect of one treatment.
- If exchangeability can be assumed for *A* but not for *E,* we cannot generally assess the presence of interaction between *A* and *E* , but can still assess the presence of effect modification by *E* .
  - Why? Because one does not need any identifying assumptions involving *E* to compute the effect of *A* in each of the strata of *E* .



Counterfactual response types and interactions

Classification of individuals according to their counterfactual responses: Possible response types

Type	$Y^{A=0}$	$Y^{A=1}$
Doomed	1	1
Preventive	1	0
Causative	0	1
Immune	0	0



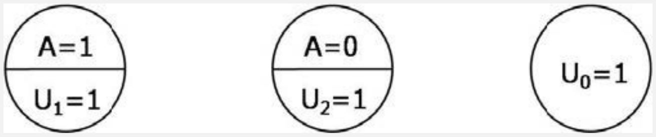
Counterfactual response types and interactions

Responses $Y^{A,E}$ for each A, E value

Type	1,1	0,1	1,0	0,0	Type	1,1	0,1	1,0	0,0
1	1	1	1	1	9	0	1	1	1
2	1	1	1	0	10	0	1	1	0
3	1	1	0	1	11	0	1	0	1
4	1	1	0	0	12	0	1	0	0
5	1	0	1	1	13	0	0	1	1
6	1	0	1	0	14	0	0	1	0
7	1	0	0	1	15	0	0	0	1
8	1	0	0	0	16	0	0	0	0

Sufficient Causes

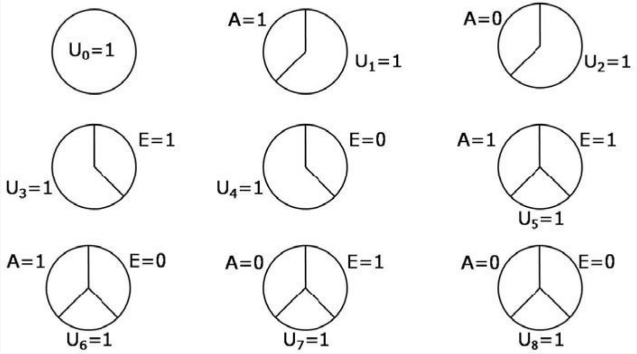
- Help represent the causal mechanisms involved in the interaction between two treatments.
- An oversimplified example:
  - $A = 1$ and set of background factors $U_1 = 1$ cause death,
  - $A = 0$ and set of background factors $U_2 = 1$ cause death,
  - "Doomed" individuals: $U_0 = 1$ cause death (regardless of treatment)



Heman & Robins: Figure5.1

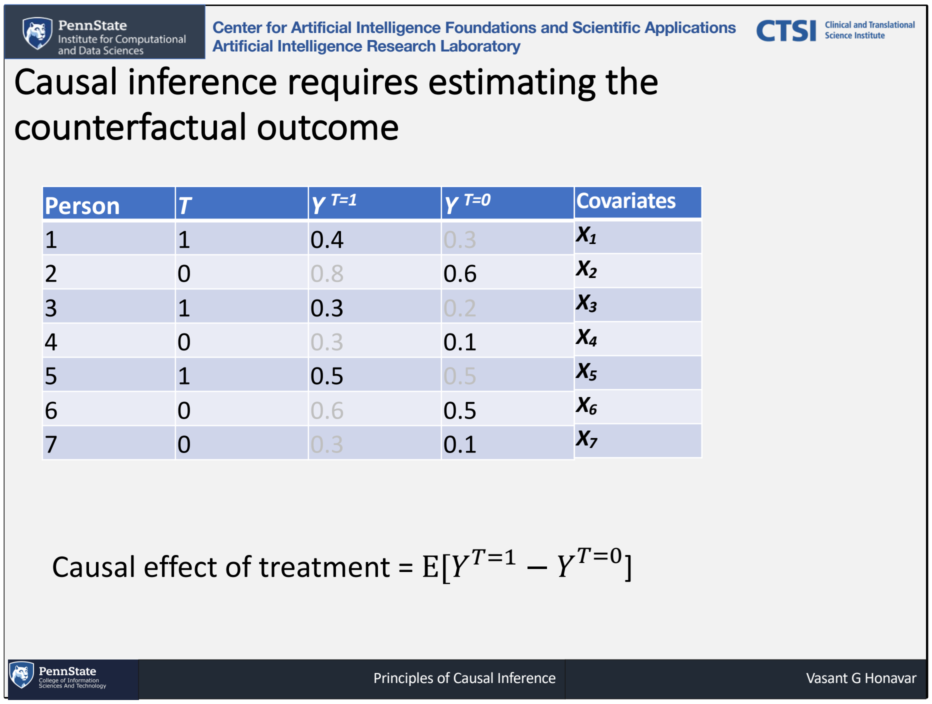
Sufficient Causes

In case of two treatments, there are nine possible sufficient causes (not all of them necessarily exist)



Hernan and Robins: Figure5.2

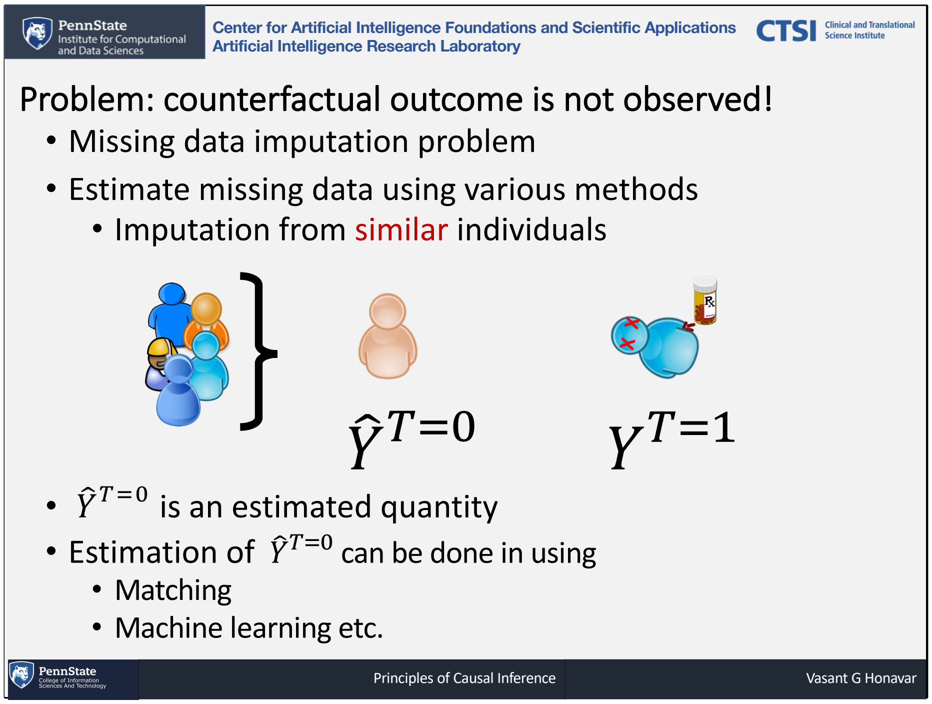
Back to causal effect estimation



Causal inference requires estimating the counterfactual outcome

Person	T	$Y^{T=1}$	$Y^{T=0}$	Covariates
1	1	0.4	0.3	X_1
2	0	0.8	0.6	X_2
3	1	0.3	0.2	X_3
4	0	0.3	0.1	X_4
5	1	0.5	0.5	X_5
6	0	0.6	0.5	X_6
7	0	0.3	0.1	X_7

Causal effect of treatment = $\mathrm{E}[Y^{T=1} - Y^{T=0}]$



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

Problem: counterfactual outcome is not observed!

- Missing data imputation problem
- Estimate missing data using various methods
  - Imputation from similar individuals

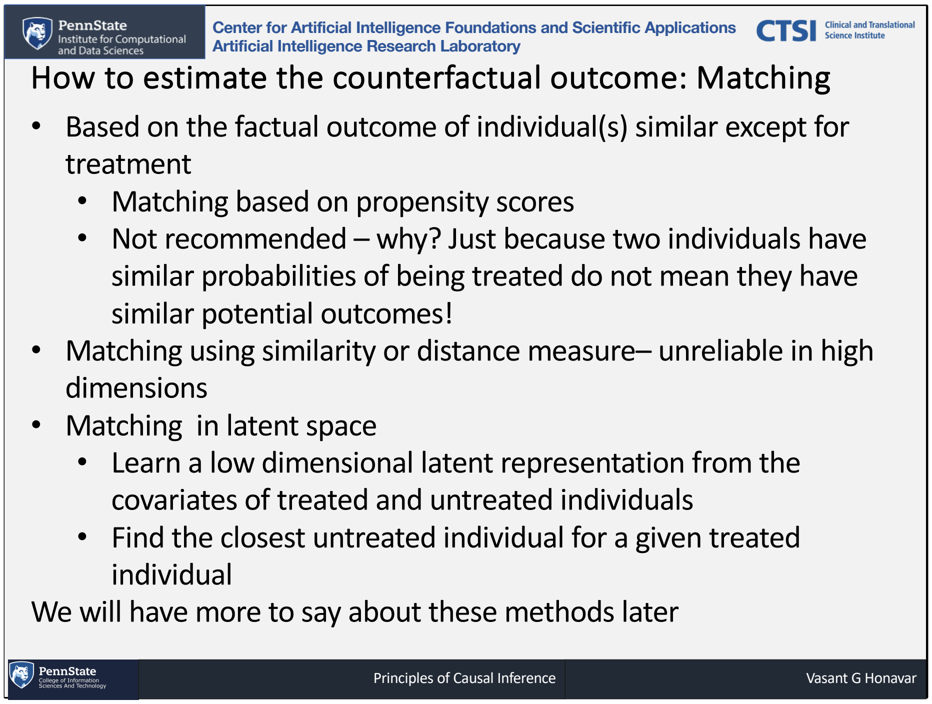
$$\hat{Y}^{T=0} \qquad Y^{T=1}$$

- $\hat{Y}^{T=0}$ is an estimated quantity
- Estimation of $\hat{Y}^{T=0}$ can be done in using
  - Matching
  - Machine learning etc.

PennState College of Information Sciences And Technology

Principles of Causal Inference

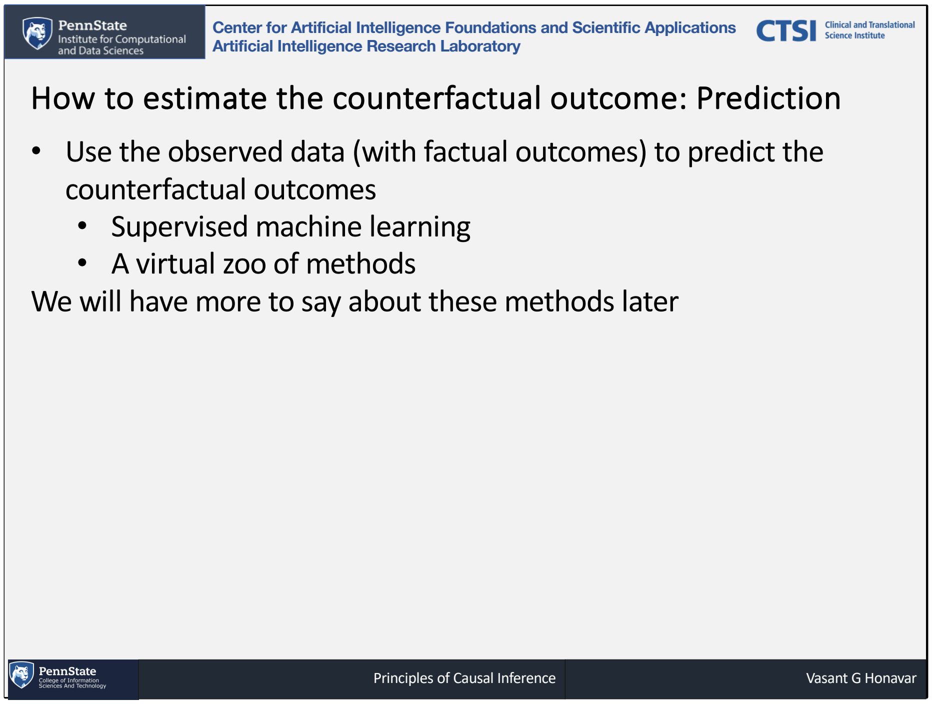
Vasant G Honavar



How to estimate the counterfactual outcome: Matching

- Based on the factual outcome of individual(s) similar except for treatment
  - Matching based on propensity scores
  - Not recommended – why? Just because two individuals have similar probabilities of being treated do not mean they have similar potential outcomes!
- Matching using similarity or distance measure– unreliable in high dimensions
- Matching in latent space
  - Learn a low dimensional latent representation from the covariates of treated and untreated individuals
  - Find the closest untreated individual for a given treated individual

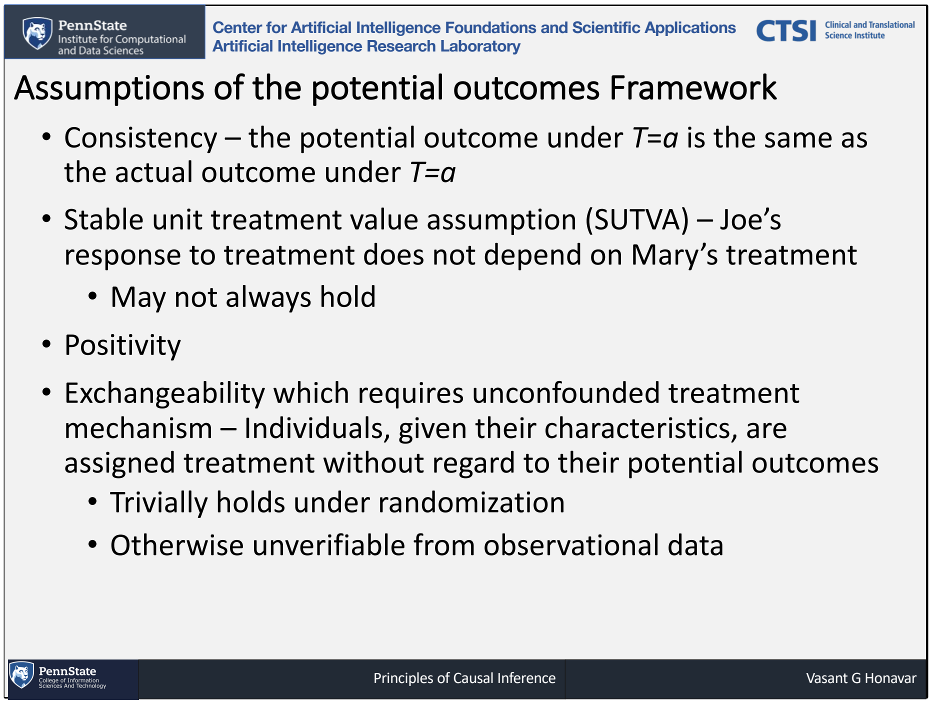
We will have more to say about these methods later



How to estimate the counterfactual outcome: Prediction

- Use the observed data (with factual outcomes) to predict the counterfactual outcomes
  - Supervised machine learning
  - A virtual zoo of methods

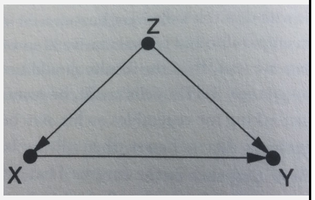
We will have more to say about these methods later



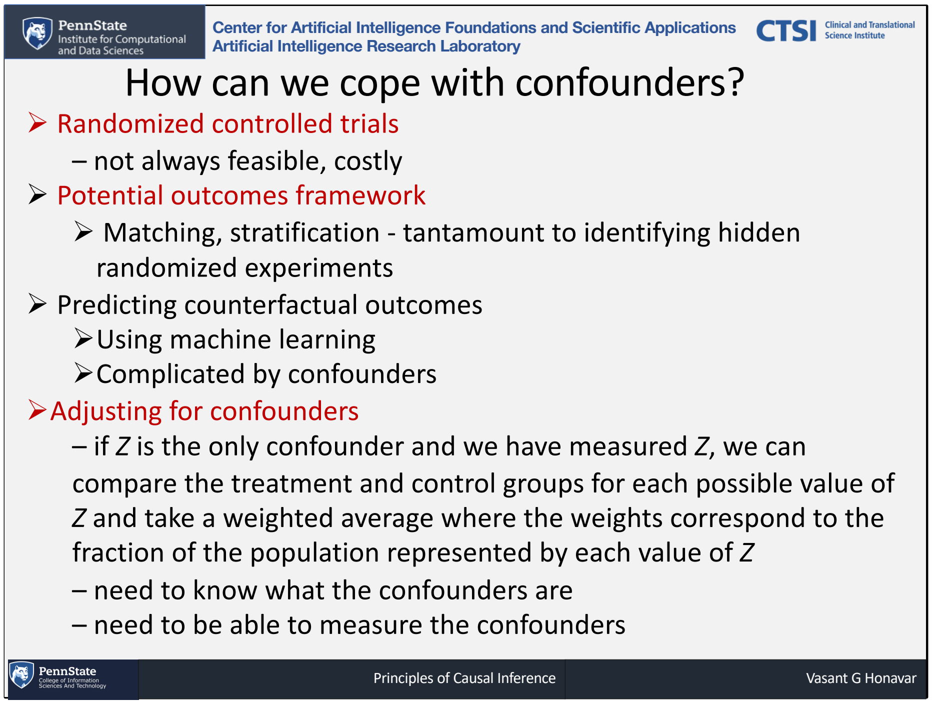
Assumptions of the potential outcomes Framework

- Consistency – the potential outcome under *T=a* is the same as the actual outcome under *T=a*
- Stable unit treatment value assumption (SUTVA) – Joe's response to treatment does not depend on Mary's treatment
  - May not always hold
- Positivity
- Exchangeability which requires unconfounded treatment mechanism – Individuals, given their characteristics, are assigned treatment without regard to their potential outcomes
  - Trivially holds under randomization
  - Otherwise unverifiable from observational data

Confounding revisited



- Confounding bias arises whenever a variable influences both who is selected for treatment and the outcome of the experiment
  - Sometimes the confounders are known
  - Sometimes the confounders are suspected
- The most basic version of confounding
  - The true causal effect X → Y is mixed with the spurious correlation induced by the fork X ←Z→ Y
  - Example:
    - We are testing a drug but give it to patients who are older, but not to those who are younger
    - Age becomes a confounder



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

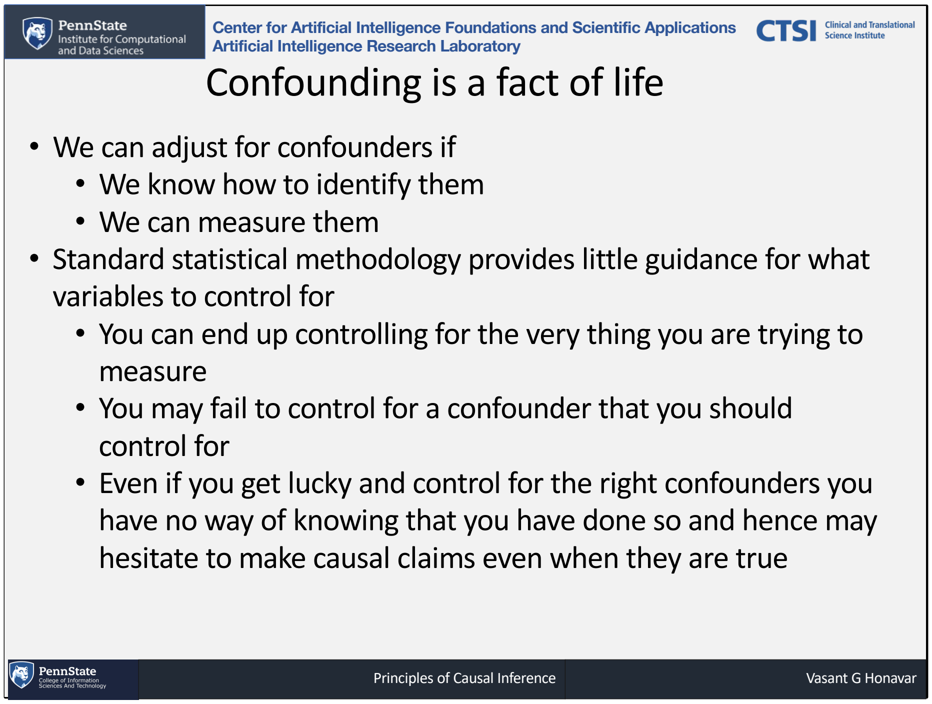
How can we cope with confounders?

- Randomized controlled trials
  - – not always feasible, costly
- Potential outcomes framework
  - Matching, stratification - tantamount to identifying hidden randomized experiments
- Predicting counterfactual outcomes
  - Using machine learning
  - Complicated by confounders
- Adjusting for confounders

  – if Z is the only confounder and we have measured Z, we can compare the treatment and control groups for each possible value of Z and take a weighted average where the weights correspond to the fraction of the population represented by each value of Z

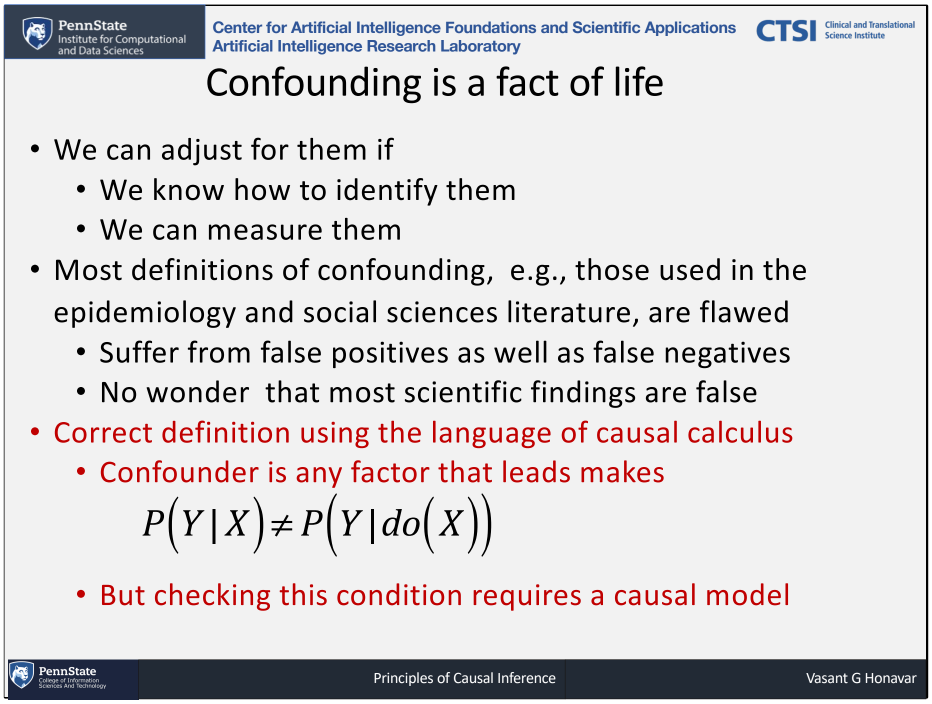
  – need to know what the confounders are

  – need to be able to measure the confounders



Confounding is a fact of life

- We can adjust for confounders if
  - We know how to identify them
  - We can measure them
- Standard statistical methodology provides little guidance for what variables to control for
  - You can end up controlling for the very thing you are trying to measure
  - You may fail to control for a confounder that you should control for
  - Even if you get lucky and control for the right confounders you have no way of knowing that you have done so and hence may hesitate to make causal claims even when they are true

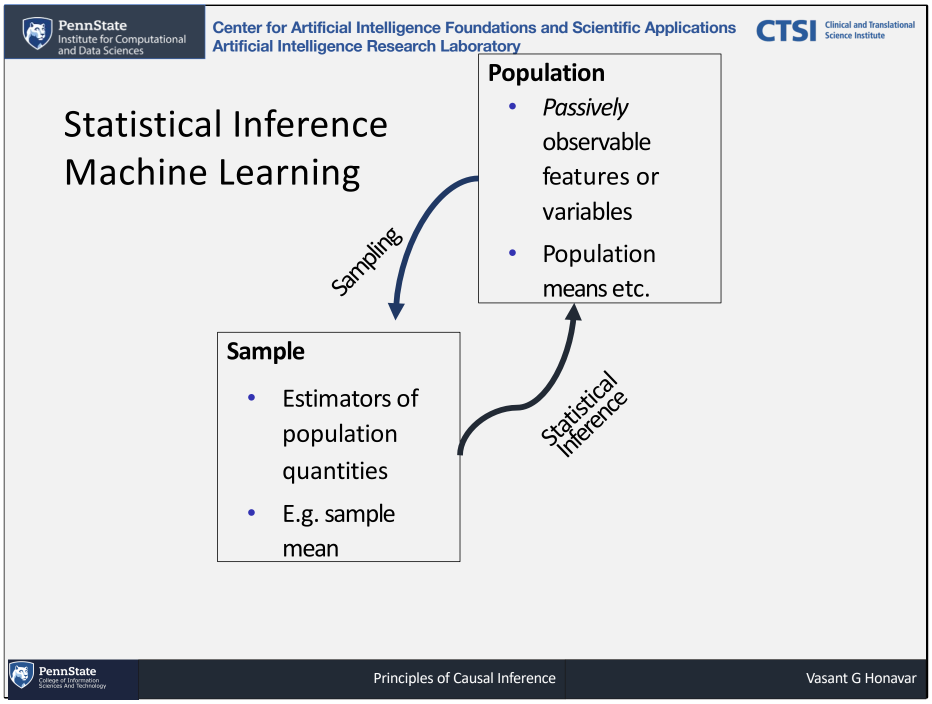


Confounding is a fact of life

- We can adjust for them if
  - We know how to identify them
  - We can measure them
- Most definitions of confounding, e.g., those used in the epidemiology and social sciences literature, are flawed
  - Suffer from false positives as well as false negatives
  - No wonder that most scientific findings are false
- Correct definition using the language of causal calculus
  - Confounder is any factor that leads makes

$$P(Y|X) \neq P(Y|do(X))$$

  - But checking this condition requires a causal model



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

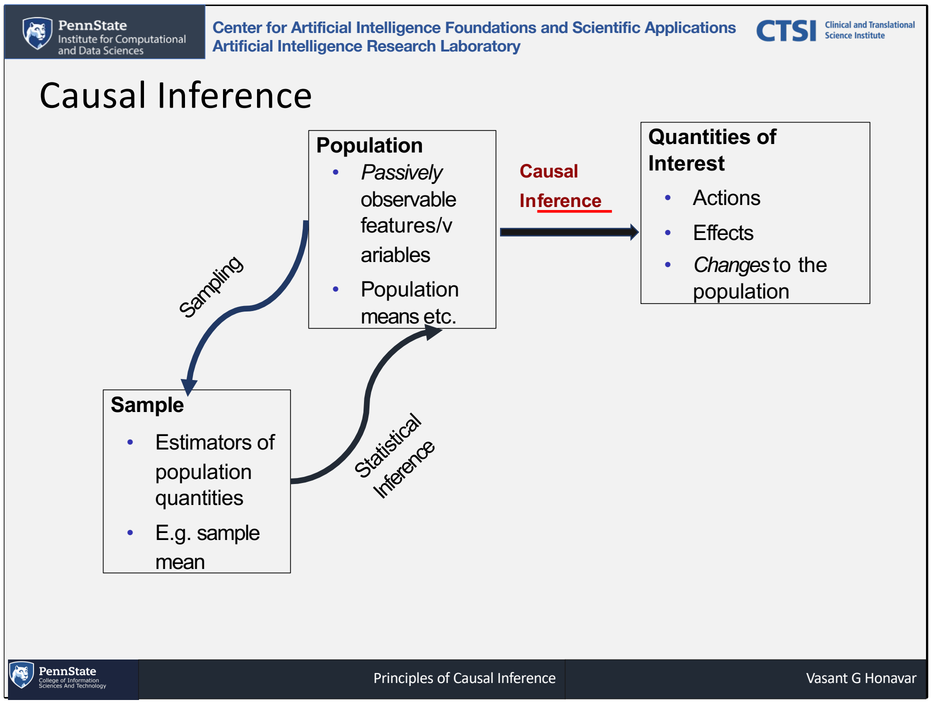
Statistical Inference
Machine Learning

Population

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

Sample

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



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

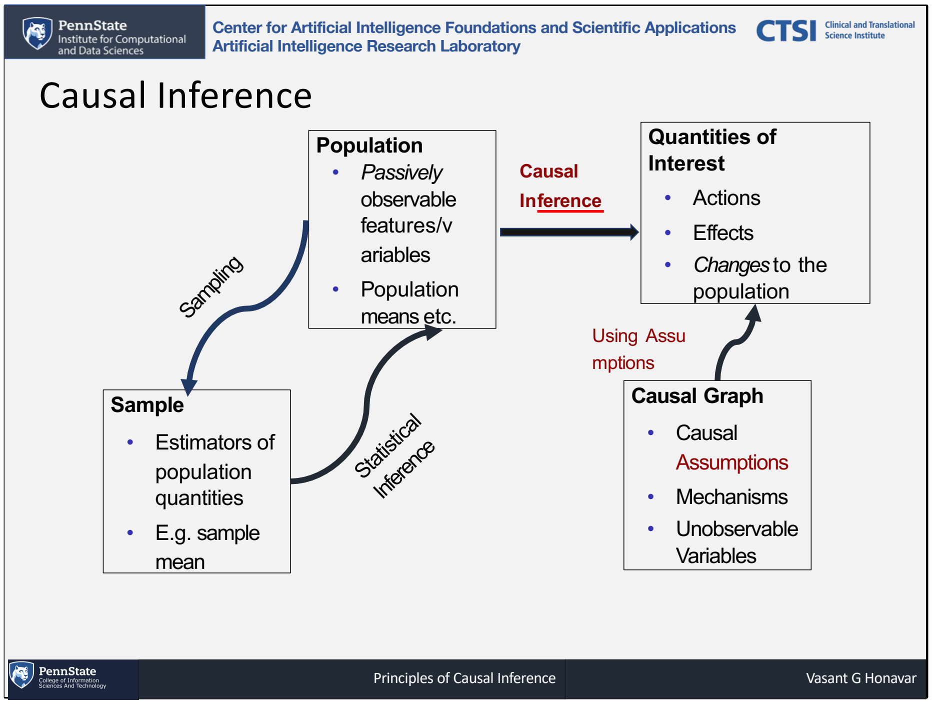
Sample
- Estimators of population quantities
- E.g. sample mean

Statistical Inference

PennState College of Information Sciences And Technology

Principles of Causal Inference

Vasant G Honavar



Causal Inference

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

Sampling

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

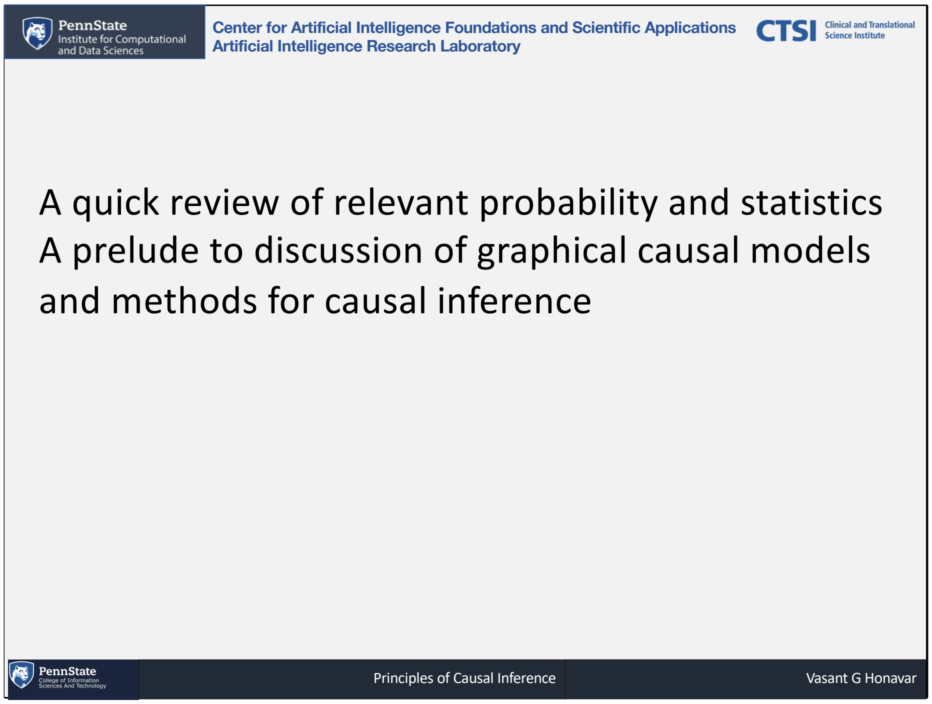
Statistical Inference

Causal Inference

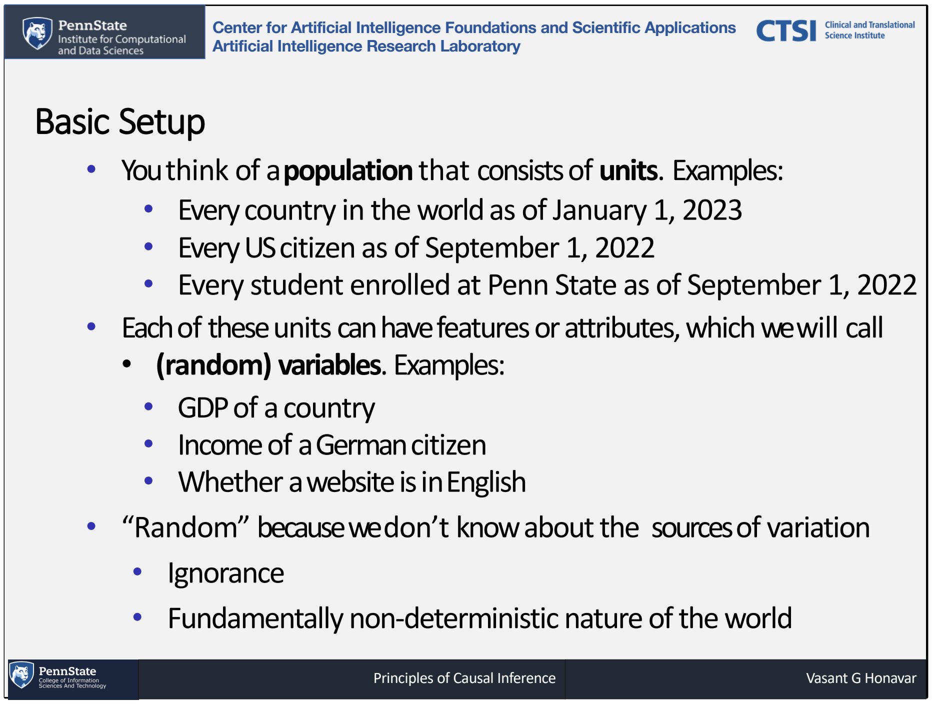
Quantities of Interest
- Actions
- Effects
- *Changes* to the population

Using Assumptions

Causal Graph
- Causal Assumptions
- Mechanisms
- Unobservable Variables

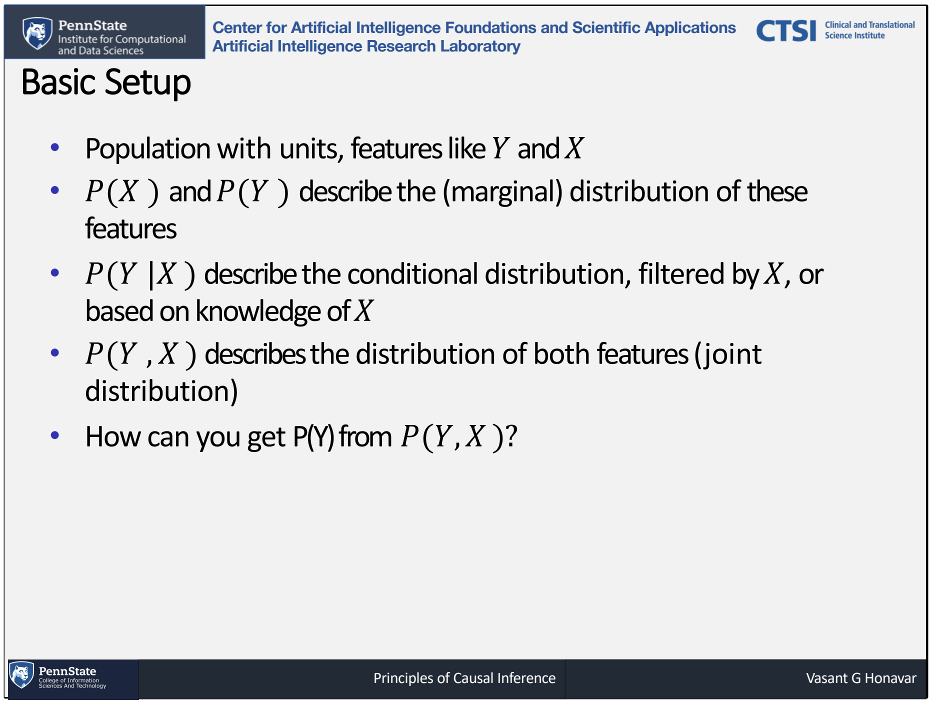


A quick review of relevant probability and statistics
A prelude to discussion of graphical causal models
and methods for causal inference



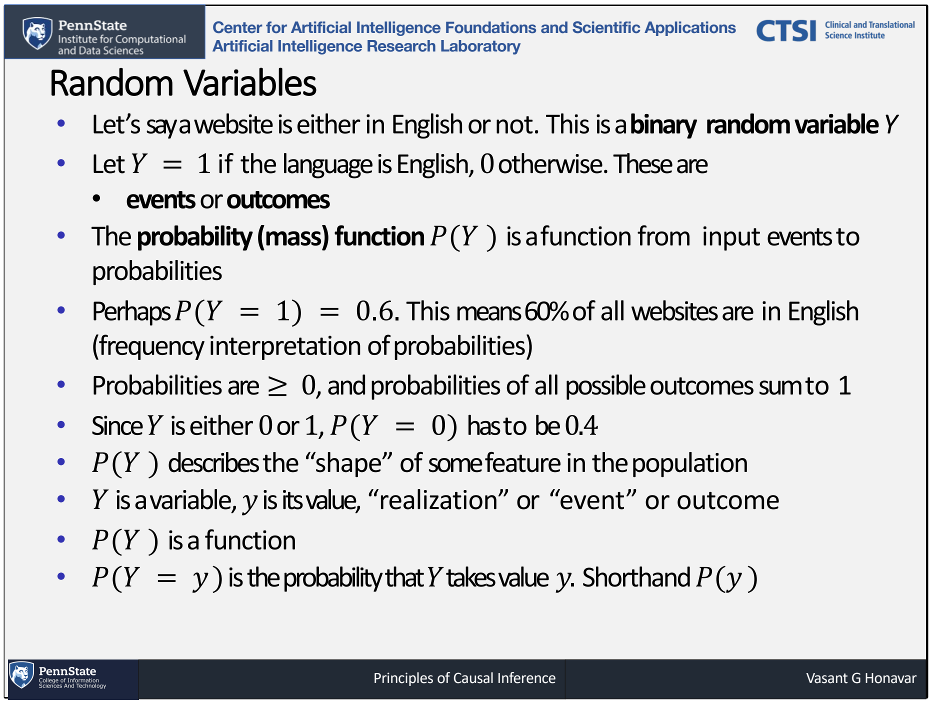
Basic Setup

- You think of a **population** that consists of **units**. Examples:
  - Every country in the world as of January 1, 2023
  - Every US citizen as of September 1, 2022
  - Every student enrolled at Penn State as of September 1, 2022
- Each of these units can have features or attributes, which we will call
  - **(random) variables**. Examples:
  - GDP of a country
  - Income of a German citizen
  - Whether a website is in English
- “Random” because we don’t know about the sources of variation
  - Ignorance
  - Fundamentally non-deterministic nature of the world



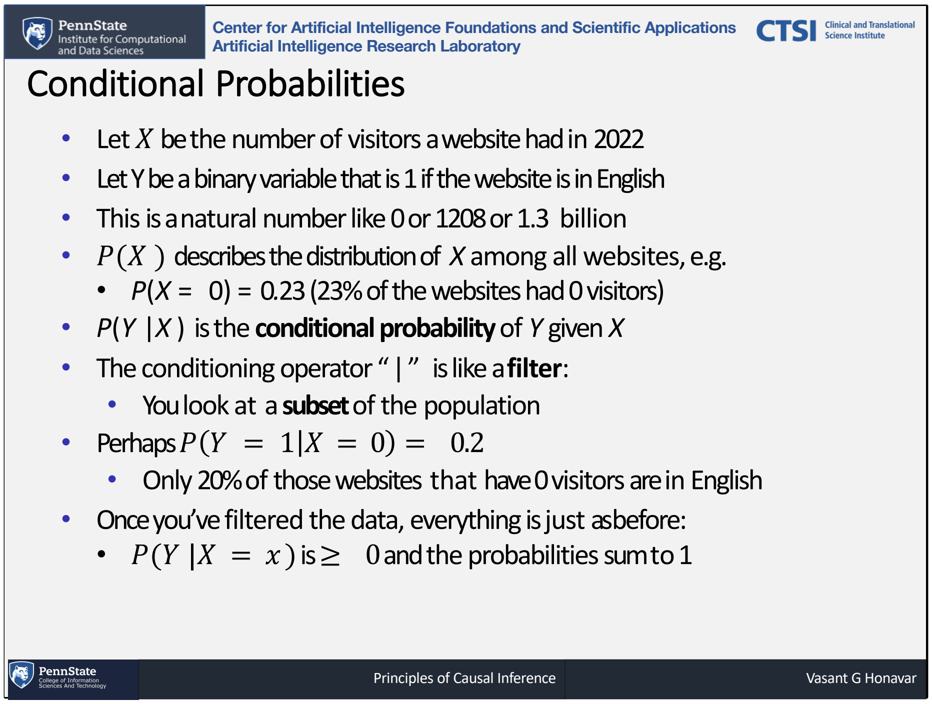
Basic Setup

- Population with units, features like Y and X
- $P(X)$ and $P(Y)$ describe the (marginal) distribution of these features
- $P(Y|X)$ describe the conditional distribution, filtered by X, or based on knowledge of X
- $P(Y,X)$ describes the distribution of both features (joint distribution)
- How can you get P(Y) from $P(Y,X)$?



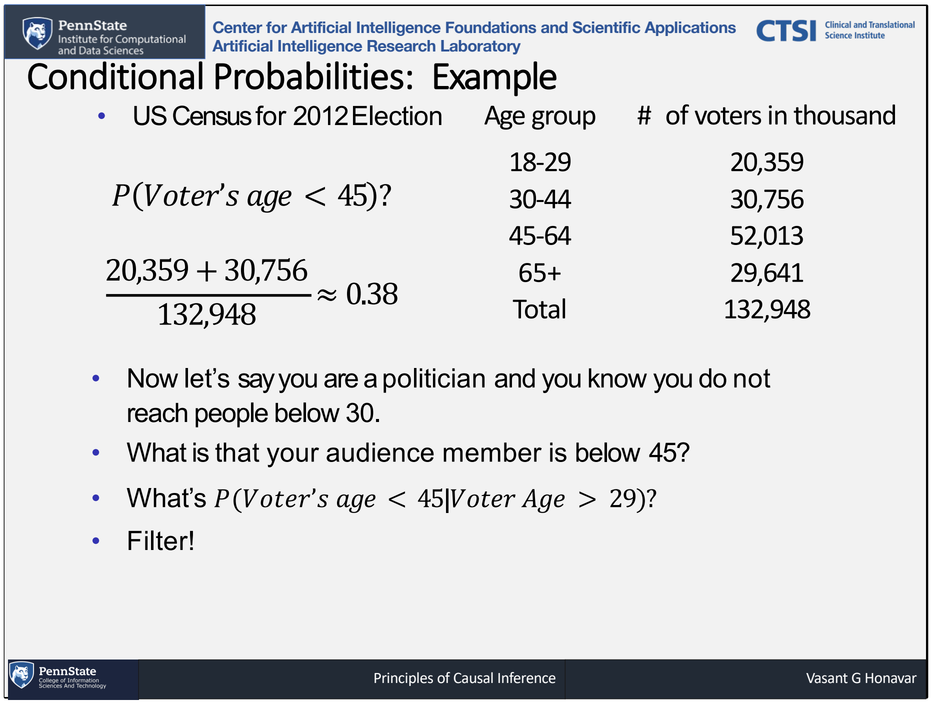
Random Variables

- Let’s say a website is either in English or not. This is a **binary random variable** Y
- Let $Y = 1$ if the language is English, 0 otherwise. These are
  - **events** or **outcomes**
- The **probability (mass) function** $P(Y)$ is a function from input events to probabilities
- Perhaps $P(Y = 1) = 0.6$. This means 60% of all websites are in English (frequency interpretation of probabilities)
- Probabilities are ≥ 0, and probabilities of all possible outcomes sum to 1
- Since Y is either 0 or 1, $P(Y = 0)$ has to be 0.4
- $P(Y)$ describes the “shape” of some feature in the population
- Y is a variable, y is its value, “realization” or “event” or outcome
- $P(Y)$ is a function
- $P(Y = y)$ is the probability that Y takes value y. Shorthand $P(y)$



Conditional Probabilities

- Let X be the number of visitors a website had in 2022
- Let Y be a binary variable that is 1 if the website is in English
- This is a natural number like 0 or 1208 or 1.3 billion
- $P(X)$ describes the distribution of X among all websites, e.g.
  - $P(X = 0) = 0.23$ (23% of the websites had 0 visitors)
- $P(Y \mid X)$ is the **conditional probability** of Y given X
- The conditioning operator " | " is like a **filter**:
  - You look at a **subset** of the population
- Perhaps $P(Y = 1|X = 0) = 0.2$
  - Only 20% of those websites that have 0 visitors are in English
- Once you've filtered the data, everything is just as before:
  - $P(Y \mid X = x)$ is ≥ 0 and the probabilities sum to 1



Conditional Probabilities: Example

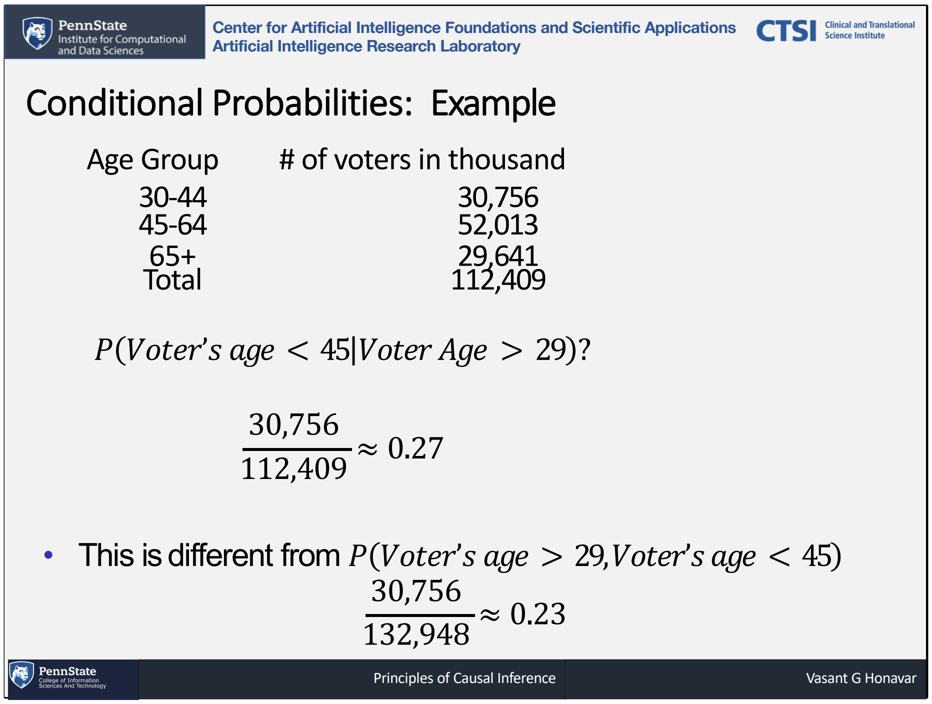
- US Census for 2012 Election

$P(Voter's\ age < 45)?$

$$\frac{20{,}359 + 30{,}756}{132{,}948} \approx 0.38$$

Age group	# of voters in thousand
18-29	20,359
30-44	30,756
45-64	52,013
65+	29,641
Total	132,948

- Now let’s say you are a politician and you know you do not reach people below 30.
- What is that your audience member is below 45?
- What’s $P(Voter's\ age < 45 | Voter\ Age > 29)$?
- Filter!



Conditional Probabilities: Example

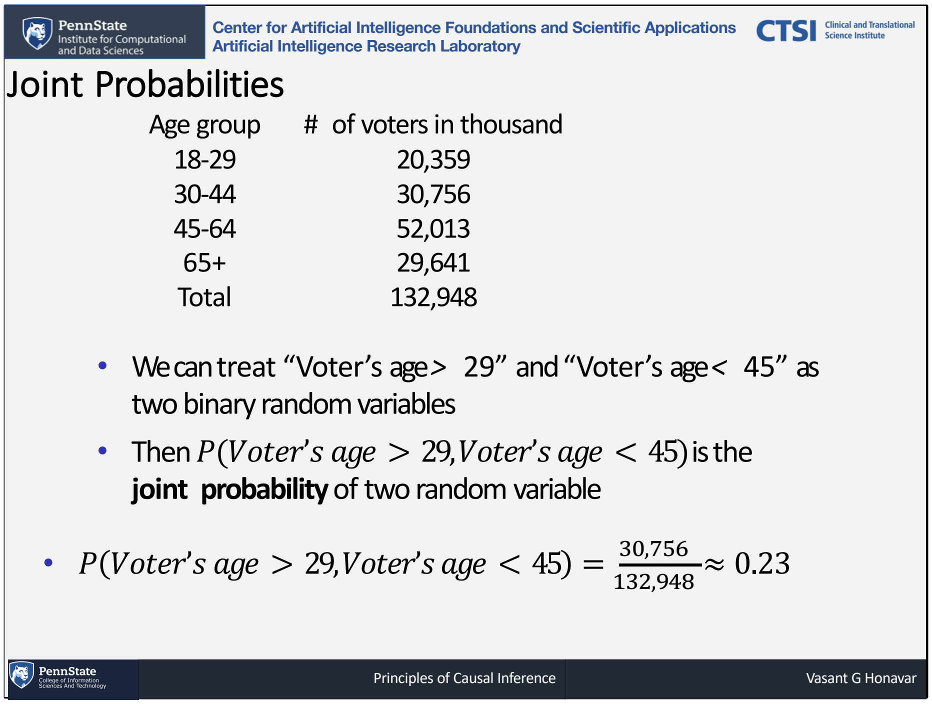
Age Group	# of voters in thousand
30-44	30,756
45-64	52,013
65+	29,641
Total	112,409

$$P(Voter\text{'}s\ age\ < 45 | Voter\ Age\ > 29)?$$

$$\frac{30{,}756}{112{,}409} \approx 0.27$$

- This is different from $P(Voter\text{'}s\ age\ > 29, Voter\text{'}s\ age\ < 45)$

$$\frac{30{,}756}{132{,}948} \approx 0.23$$

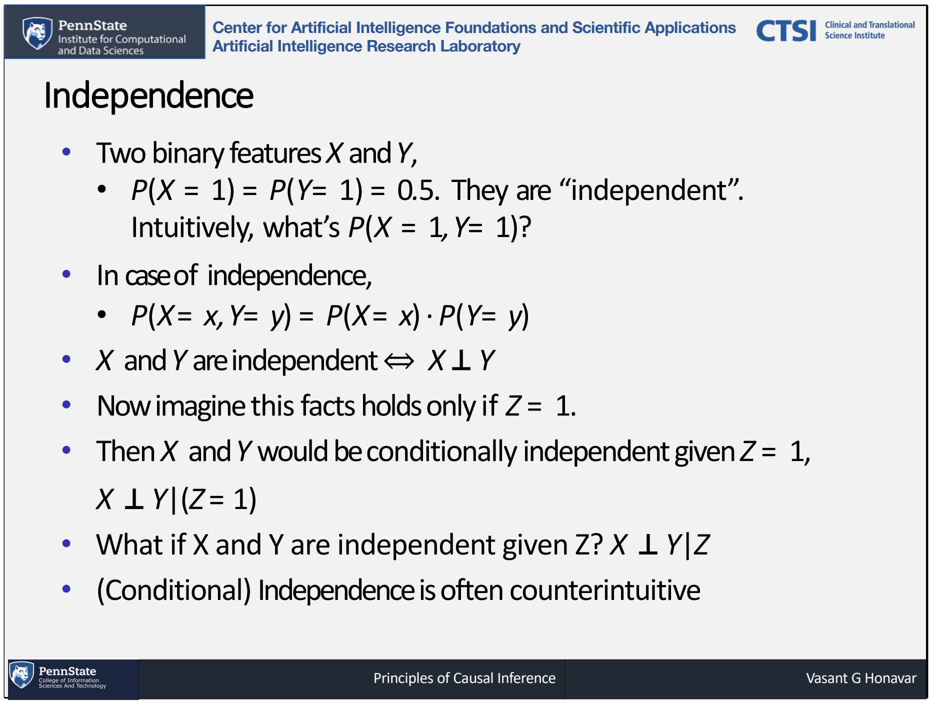


Joint Probabilities

Age group	# of voters in thousand
18-29	20,359
30-44	30,756
45-64	52,013
65+	29,641
Total	132,948

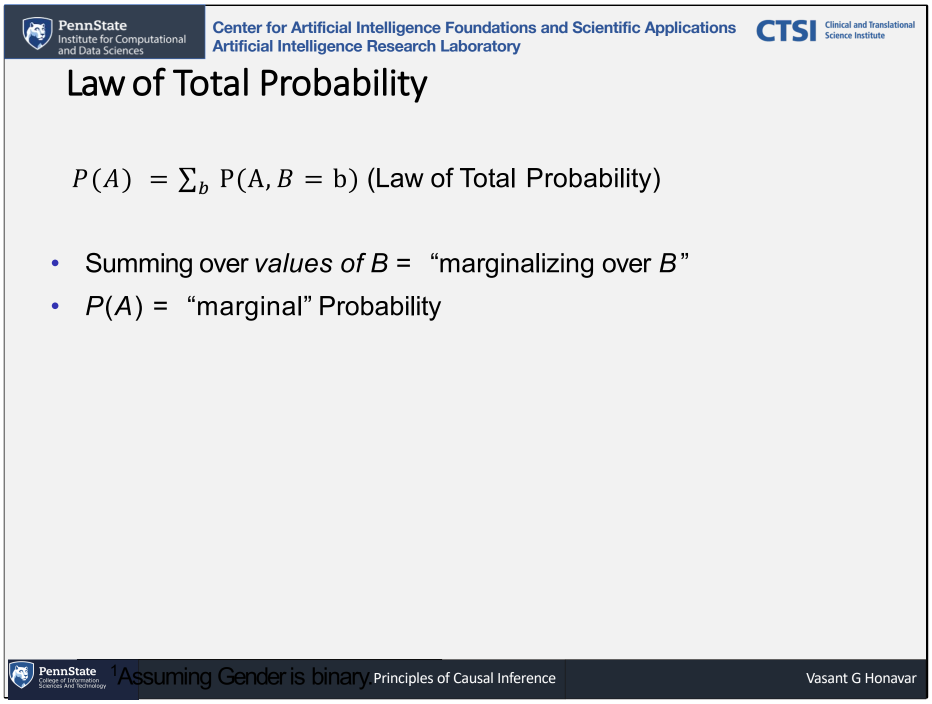
- We can treat "Voter's age > 29" and "Voter's age < 45" as two binary random variables
- Then $P(Voter's\ age > 29, Voter's\ age < 45)$ is the **joint probability** of two random variable

- $P(Voter's\ age > 29, Voter's\ age < 45) = \frac{30{,}756}{132{,}948} \approx 0.23$



Independence

- Two binary features X and Y,
  - $P(X = 1) = P(Y = 1) = 0.5$. They are “independent”. Intuitively, what’s $P(X = 1, Y = 1)$?
- In case of independence,
  - $P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$
- X and Y are independent $\Leftrightarrow$ $X \perp Y$
- Now imagine this facts holds only if $Z = 1$.
- Then X and Y would be conditionally independent given $Z = 1$, $X \perp Y|(Z = 1)$
- What if X and Y are independent given Z? $X \perp Y|Z$
- (Conditional) Independence is often counterintuitive

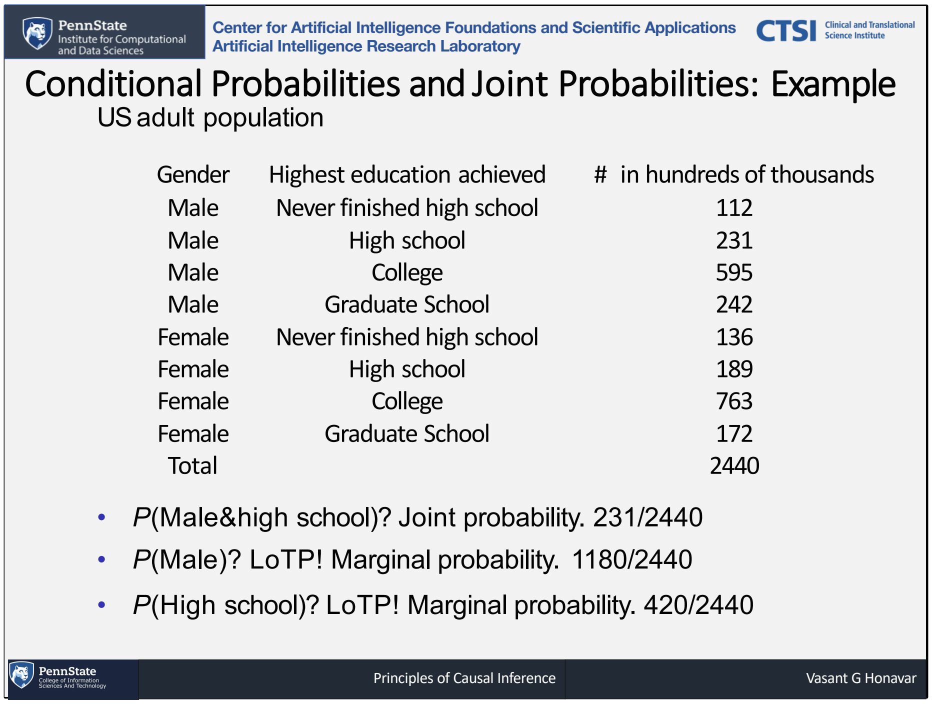


Law of Total Probability

$P(A) \; = \sum_b \mathrm{P}(\mathrm{A}, B = \mathrm{b})$ (Law of Total Probability)

- Summing over *values of B* = “marginalizing over *B*”
- *P*(*A*) = “marginal” Probability

[1] Assuming Gender is binary



PennState Institute for Computational and Data Sciences

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

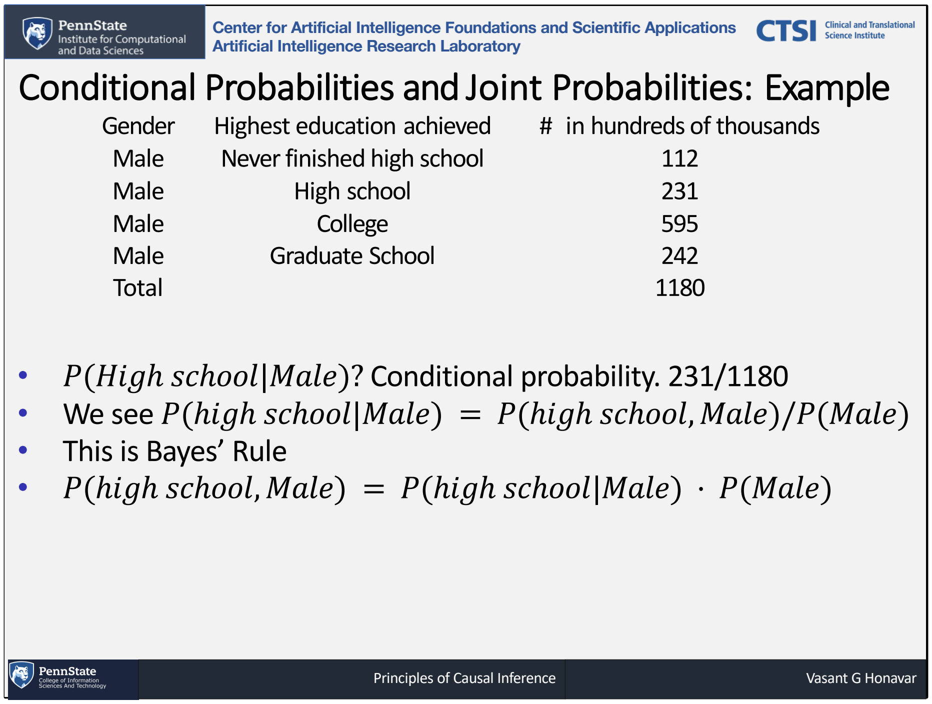
CTSI Clinical and Translational Science Institute

Conditional Probabilities and Joint Probabilities: Example

US adult population

Gender	Highest education achieved	# in hundreds of thousands
Male	Never finished high school	112
Male	High school	231
Male	College	595
Male	Graduate School	242
Female	Never finished high school	136
Female	High school	189
Female	College	763
Female	Graduate School	172
Total		2440

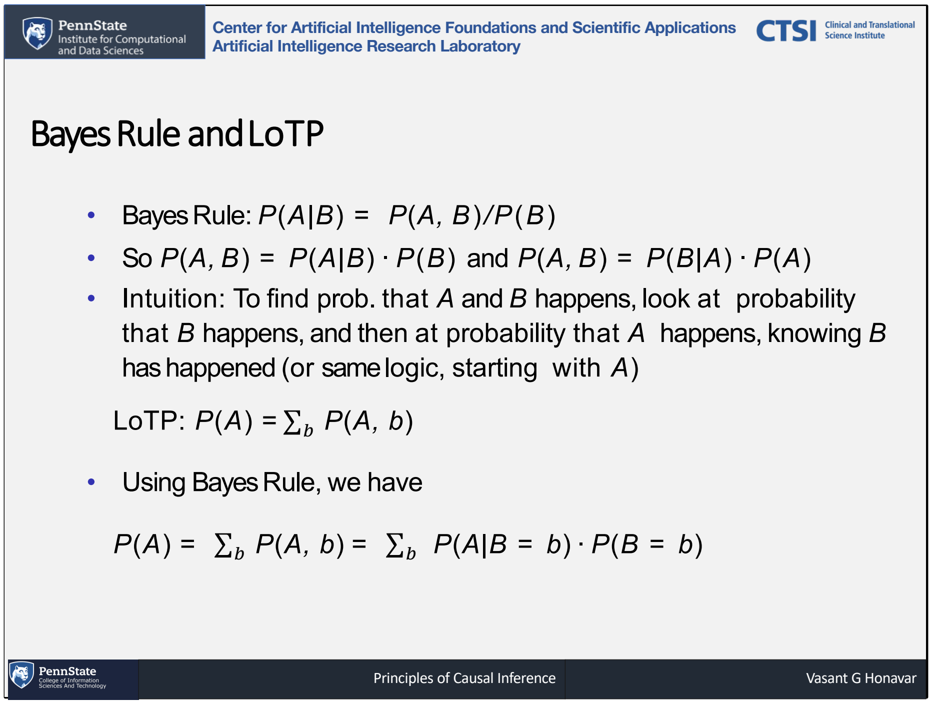
- *P*(Male&high school)? Joint probability. 231/2440
- *P*(Male)? LoTP! Marginal probability. 1180/2440
- *P*(High school)? LoTP! Marginal probability. 420/2440



Conditional Probabilities and Joint Probabilities: Example

Gender	Highest education achieved	# in hundreds of thousands
Male	Never finished high school	112
Male	High school	231
Male	College	595
Male	Graduate School	242
Total		1180

- $P(High\ school|Male)$? Conditional probability. 231/1180
- We see $P(high\ school|Male) = P(high\ school, Male)/P(Male)$
- This is Bayes' Rule
- $P(high\ school, Male) = P(high\ school|Male) \cdot P(Male)$



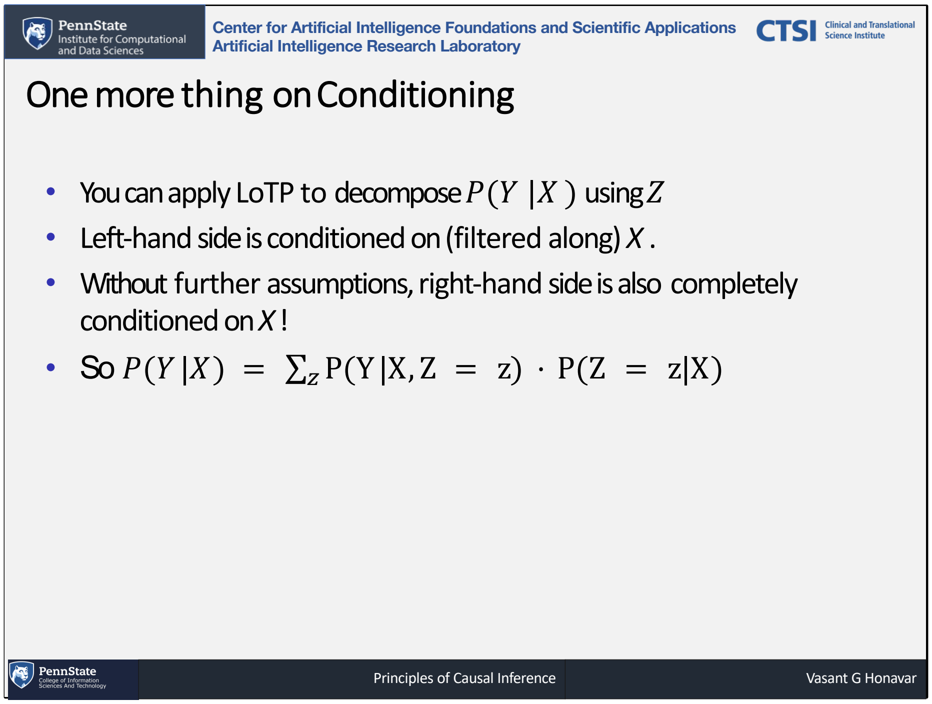
Bayes Rule and LoTP

- Bayes Rule: $P(A|B) = P(A, B)/P(B)$
- So $P(A, B) = P(A|B) \cdot P(B)$ and $P(A, B) = P(B|A) \cdot P(A)$
- Intuition: To find prob. that A and B happens, look at probability that B happens, and then at probability that A happens, knowing B has happened (or same logic, starting with A)

LoTP: $P(A) = \sum_b P(A, b)$

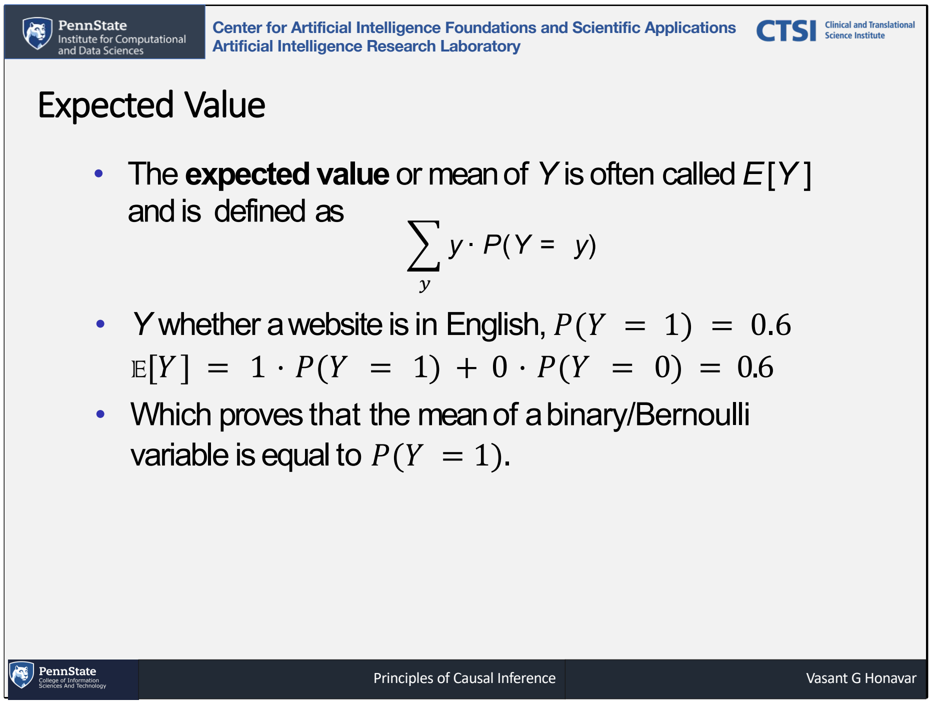
- Using Bayes Rule, we have

$$P(A) = \sum_b P(A, b) = \sum_b P(A|B = b) \cdot P(B = b)$$



One more thing on Conditioning

- You can apply LoTP to decompose $P(Y\ |X\ )$ using Z
- Left-hand side is conditioned on (filtered along) X .
- Without further assumptions, right-hand side is also completely conditioned on X !
- So $P(Y\,|X) = \sum_z \mathrm{P(Y\,|X, Z = z) \cdot P(Z = z|X)}$



Expected Value

- The **expected value** or mean of Y is often called $E[Y]$ and is defined as

$$\sum_y y \cdot P(Y = y)$$

- Y whether a website is in English, $P(Y = 1) = 0.6$
  $\mathbb{E}[Y] = 1 \cdot P(Y = 1) + 0 \cdot P(Y = 0) = 0.6$
- Which proves that the mean of a binary/Bernoulli variable is equal to $P(Y = 1)$.

Expected Value

- The expected value of a function $f(Y)$ of a random variable Y is

$$\sum_y f(y) \cdot P(Y = y)$$

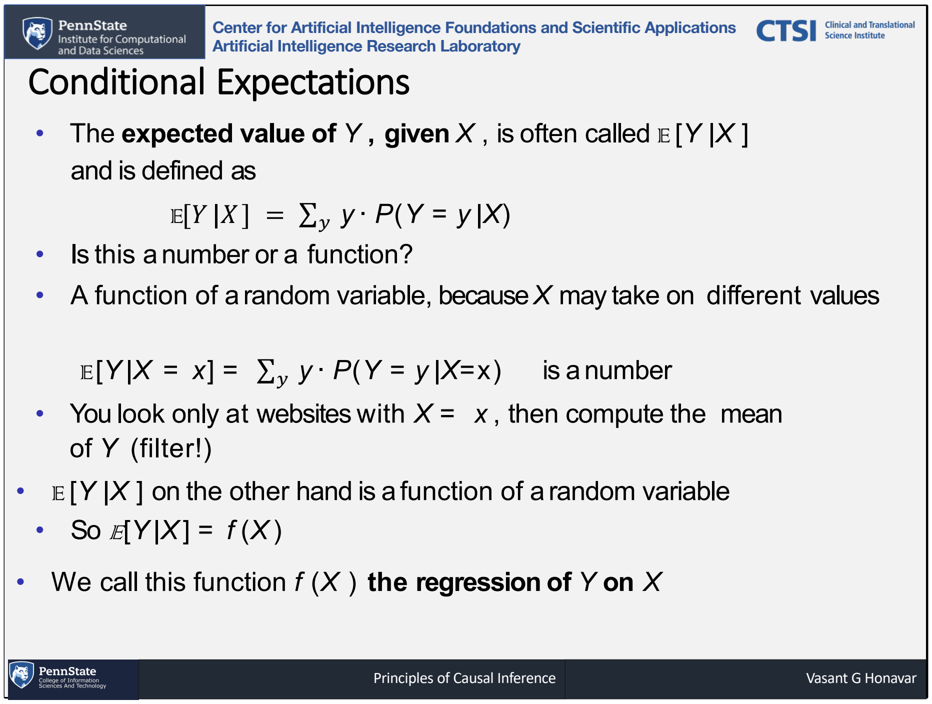
- "Law of the unconscious statistican"
- So no need to find $P(f(Y))$ to find mean of $f(Y)$

$$\mathbb{E}[Y^2 - 1] = \sum_y (y^2 - 1) \cdot P(Y = y)$$

$$= (1^2 - 1) \cdot P(Y = 1) + (0^2 - 1) \cdot P(Y = 0)$$

$$= (0)(0.6) + (-1)(0.4)$$

$$= -0.4$$



Conditional Expectations

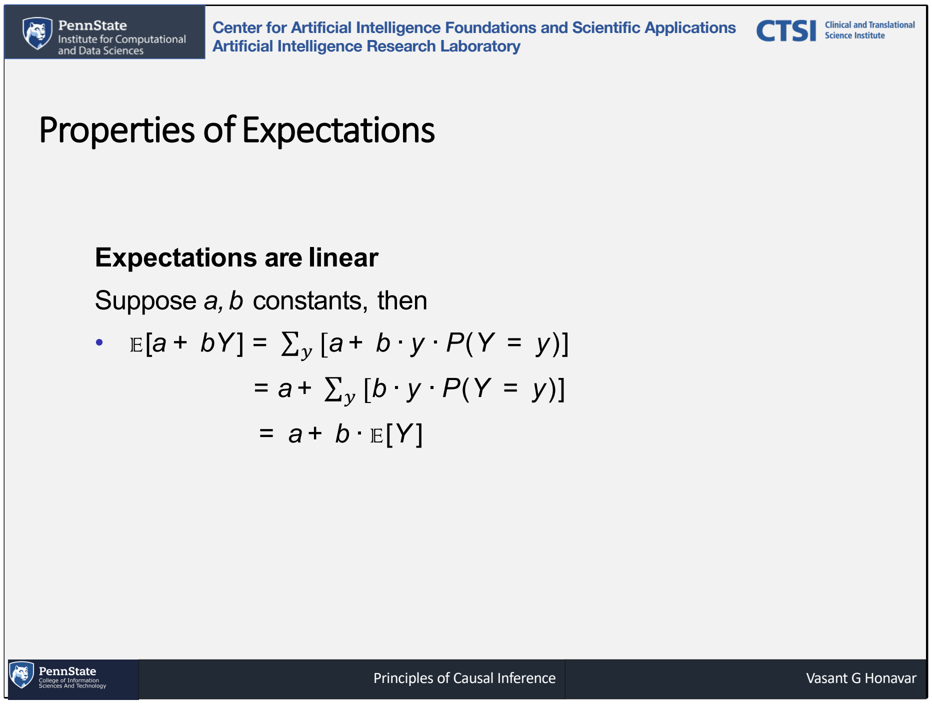
- The **expected value of** Y **, given** X , is often called $\mathbb{E}[Y|X]$ and is defined as

$$\mathbb{E}[Y|X] = \sum_y y \cdot P(Y = y|X)$$

- Is this a number or a function?
- A function of a random variable, because X may take on different values

$$\mathbb{E}[Y|X = x] = \sum_y y \cdot P(Y = y|X{=}x) \quad \text{is a number}$$

- You look only at websites with $X = x$, then compute the mean of Y (filter!)
- $\mathbb{E}[Y|X]$ on the other hand is a function of a random variable
  - So $\mathbb{E}[Y|X] = f(X)$
- We call this function $f(X)$ **the regression of** Y **on** X



Properties of Expectations

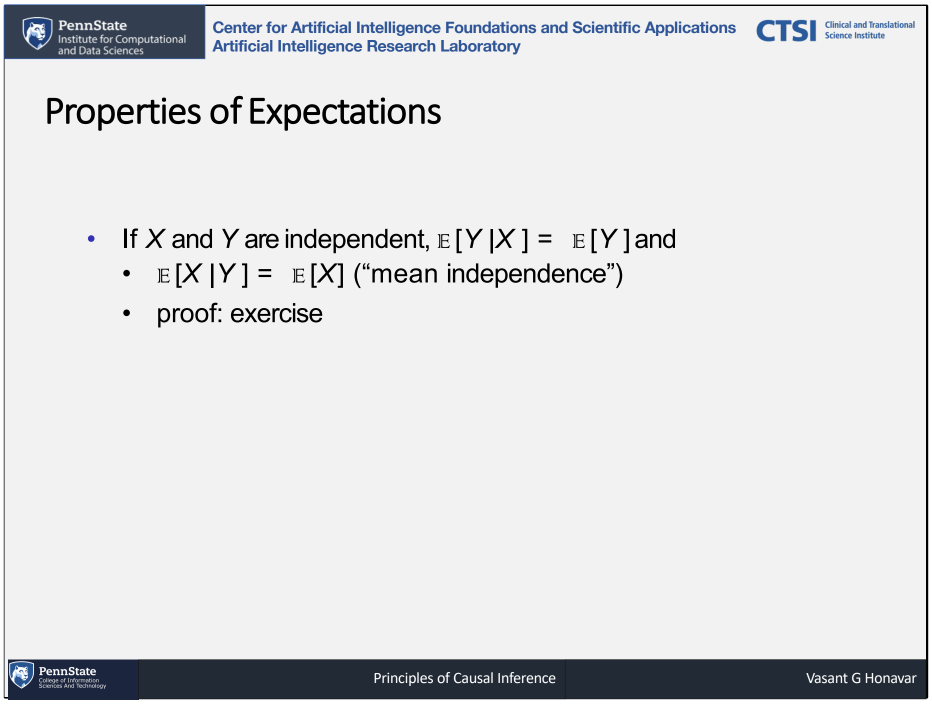
Expectations are linear

Suppose *a, b* constants, then

- $\mathbb{E}[a + bY] = \sum_y [a + b \cdot y \cdot P(Y = y)]$

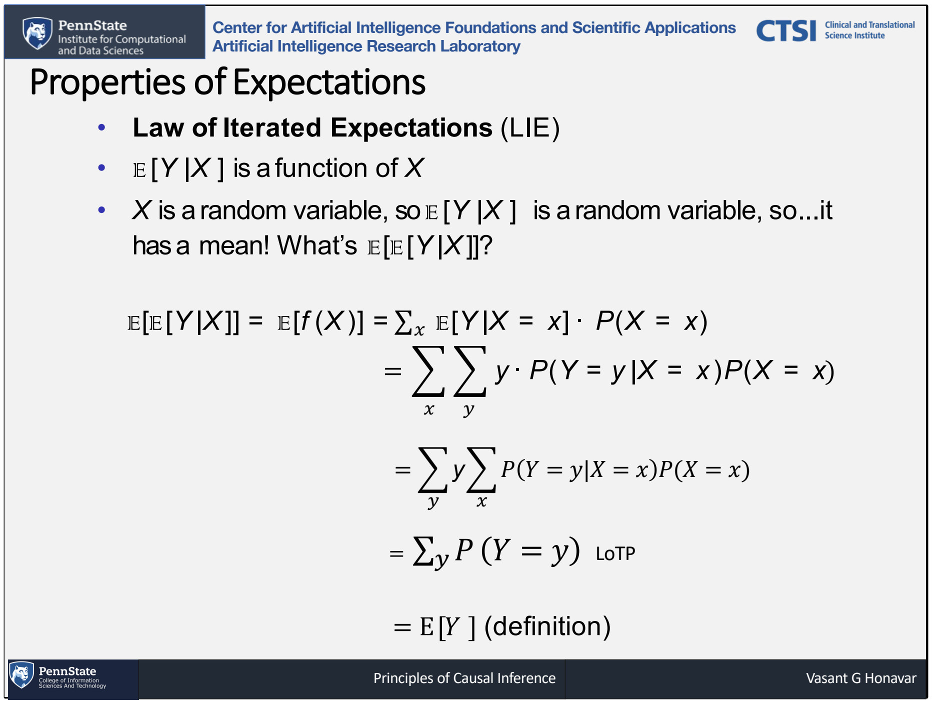
  $= a + \sum_y [b \cdot y \cdot P(Y = y)]$

  $= a + b \cdot \mathbb{E}[Y]$



Properties of Expectations

- If X and Y are independent, $\mathbb{E}[Y|X] = \mathbb{E}[Y]$ and
  - $\mathbb{E}[X|Y] = \mathbb{E}[X]$ (“mean independence”)
  - proof: exercise



Properties of Expectations

- **Law of Iterated Expectations** (LIE)
- $\mathbb{E}[Y|X]$ is a function of X
- X is a random variable, so $\mathbb{E}[Y|X]$ is a random variable, so...it has a mean! What's $\mathbb{E}[\mathbb{E}[Y|X]]$?

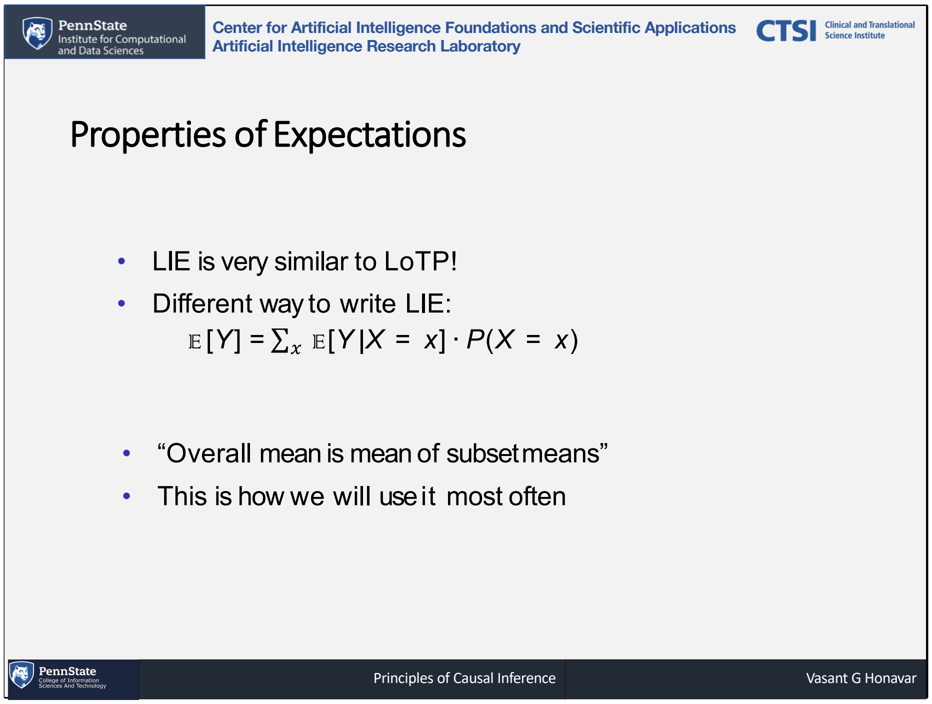
$$\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[f(X)] = \textstyle\sum_x \mathbb{E}[Y|X = x] \cdot P(X = x)$$

$$= \sum_x \sum_y y \cdot P(Y = y|X = x)P(X = x)$$

$$= \sum_y y \sum_x P(Y = y|X = x)P(X = x)$$

$$= \textstyle\sum_y P(Y = y) \quad \text{LoTP}$$

$$= \mathrm{E}[Y] \text{ (definition)}$$

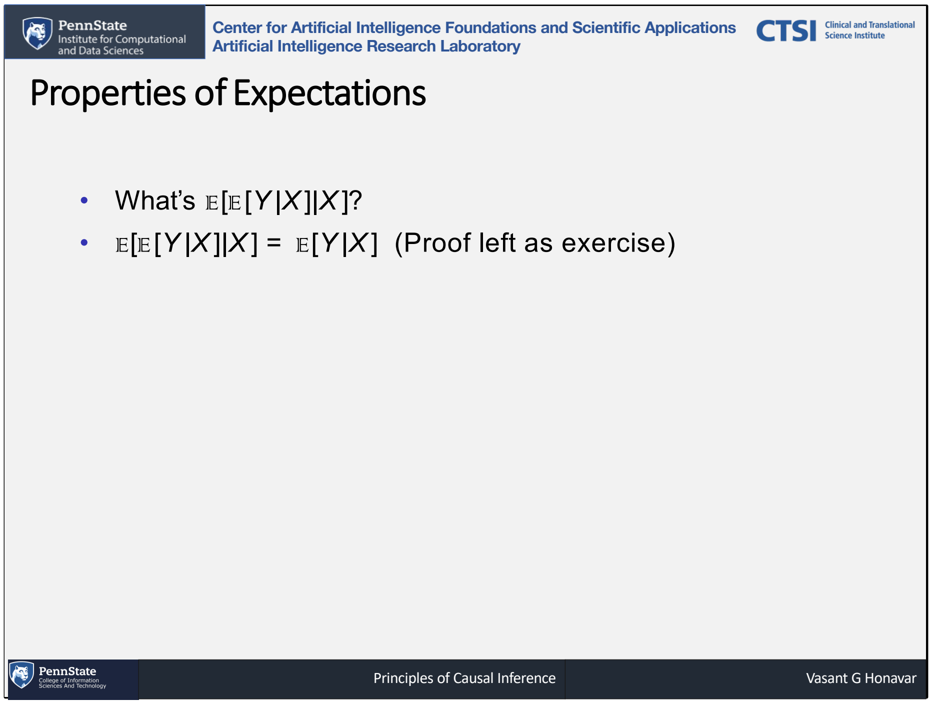


Properties of Expectations

- LIE is very similar to LoTP!
- Different way to write LIE:

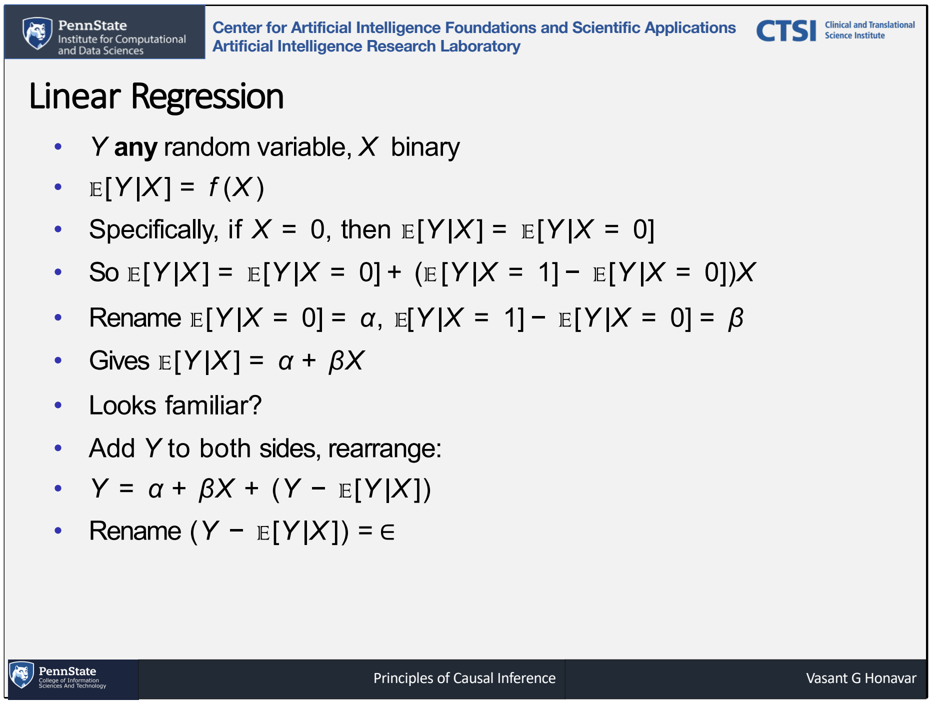
$$\mathbb{E}[Y] = \sum_x \mathbb{E}[Y|X = x] \cdot P(X = x)$$

- “Overall mean is mean of subset means”
- This is how we will use it most often



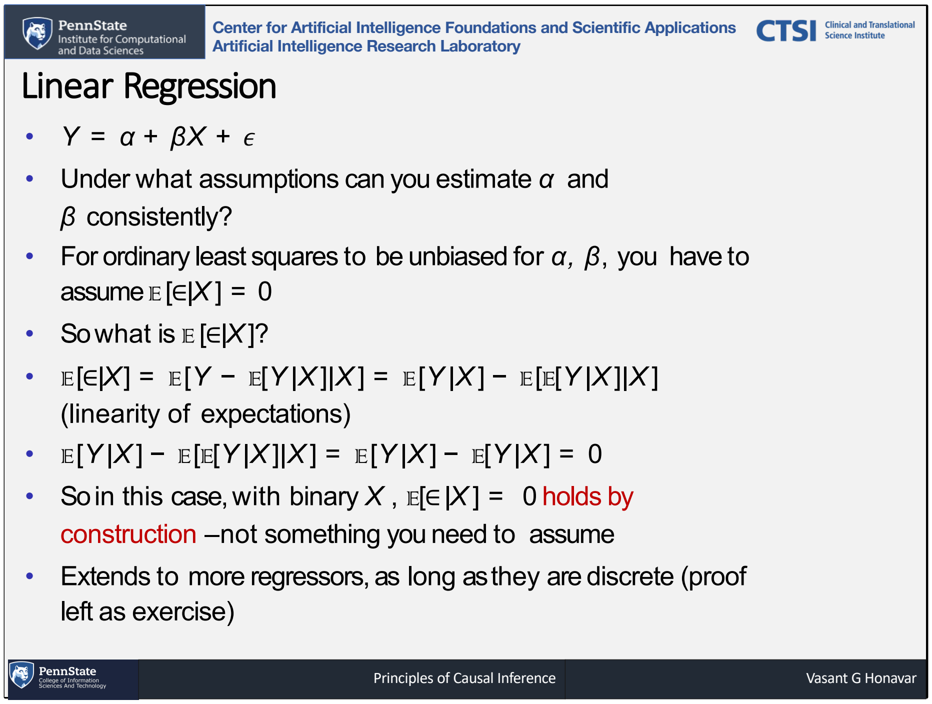
Properties of Expectations

- What's $\mathbb{E}[\mathbb{E}[Y|X]|X]$?
- $\mathbb{E}[\mathbb{E}[Y|X]|X] = \mathbb{E}[Y|X]$ (Proof left as exercise)



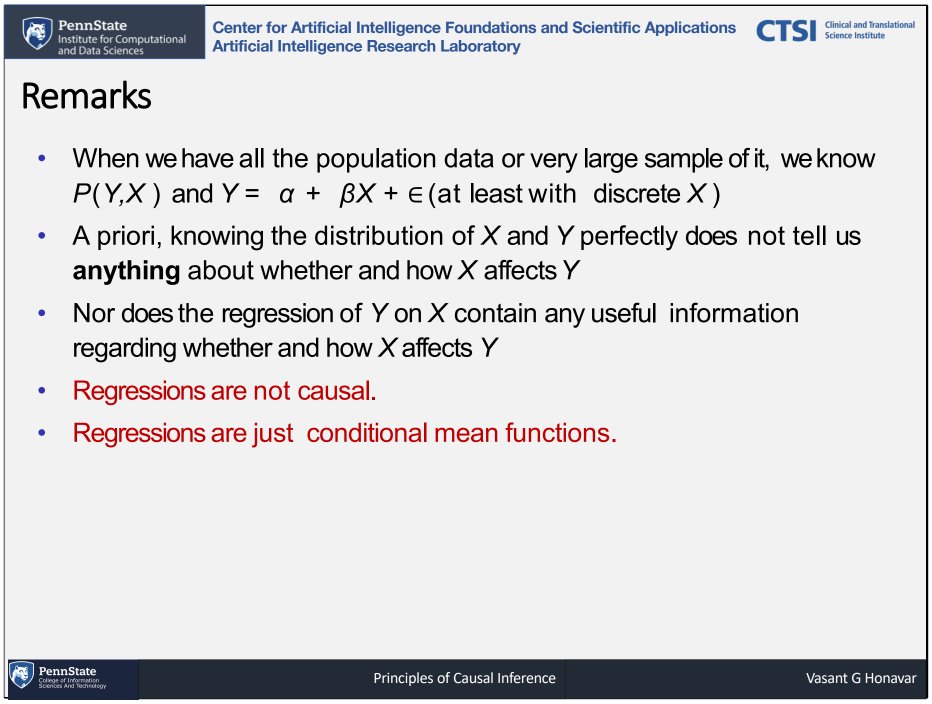
Linear Regression

- Y **any** random variable, X binary
- $\mathbb{E}[Y|X] = f(X)$
- Specifically, if $X = 0$, then $\mathbb{E}[Y|X] = \mathbb{E}[Y|X = 0]$
- So $\mathbb{E}[Y|X] = \mathbb{E}[Y|X = 0] + (\mathbb{E}[Y|X = 1] - \mathbb{E}[Y|X = 0])X$
- Rename $\mathbb{E}[Y|X = 0] = \alpha$, $\mathbb{E}[Y|X = 1] - \mathbb{E}[Y|X = 0] = \beta$
- Gives $\mathbb{E}[Y|X] = \alpha + \beta X$
- Looks familiar?
- Add Y to both sides, rearrange:
- $Y = \alpha + \beta X + (Y - \mathbb{E}[Y|X])$
- Rename $(Y - \mathbb{E}[Y|X]) = \in$



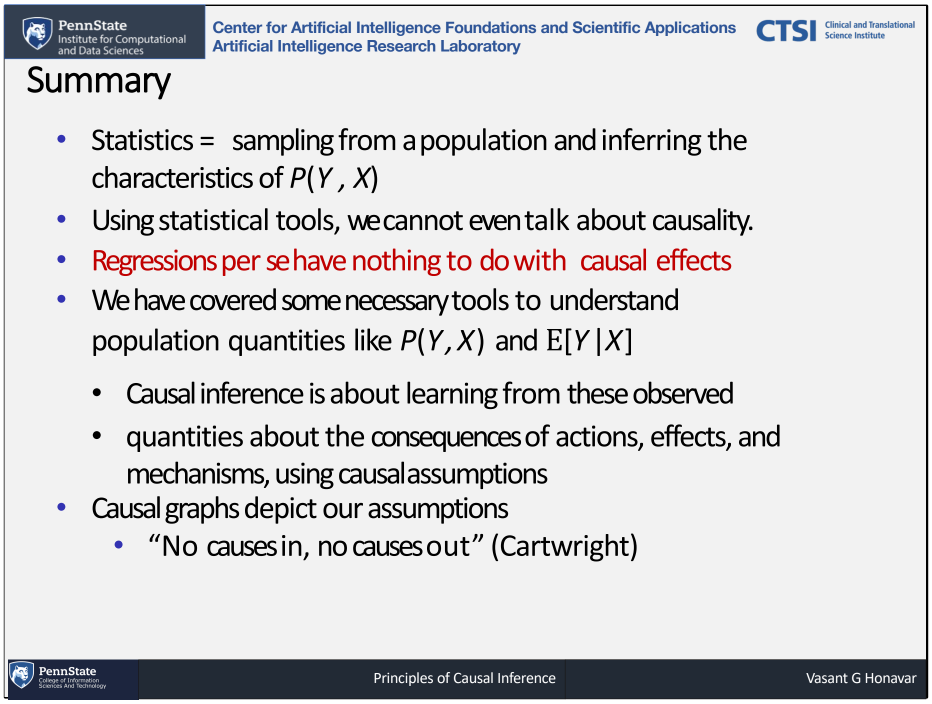
Linear Regression

- $Y = \alpha + \beta X + \epsilon$
- Under what assumptions can you estimate α and β consistently?
- For ordinary least squares to be unbiased for α, β, you have to assume $\mathbb{E}[\in|X] = 0$
- So what is $\mathbb{E}[\in|X]$?
- $\mathbb{E}[\in|X] = \mathbb{E}[Y - \mathbb{E}[Y|X]|X] = \mathbb{E}[Y|X] - \mathbb{E}[\mathbb{E}[Y|X]|X]$ (linearity of expectations)
- $\mathbb{E}[Y|X] - \mathbb{E}[\mathbb{E}[Y|X]|X] = \mathbb{E}[Y|X] - \mathbb{E}[Y|X] = 0$
- So in this case, with binary X, $\mathbb{E}[\in|X] = 0$ holds by construction –not something you need to assume
- Extends to more regressors, as long as they are discrete (proof left as exercise)



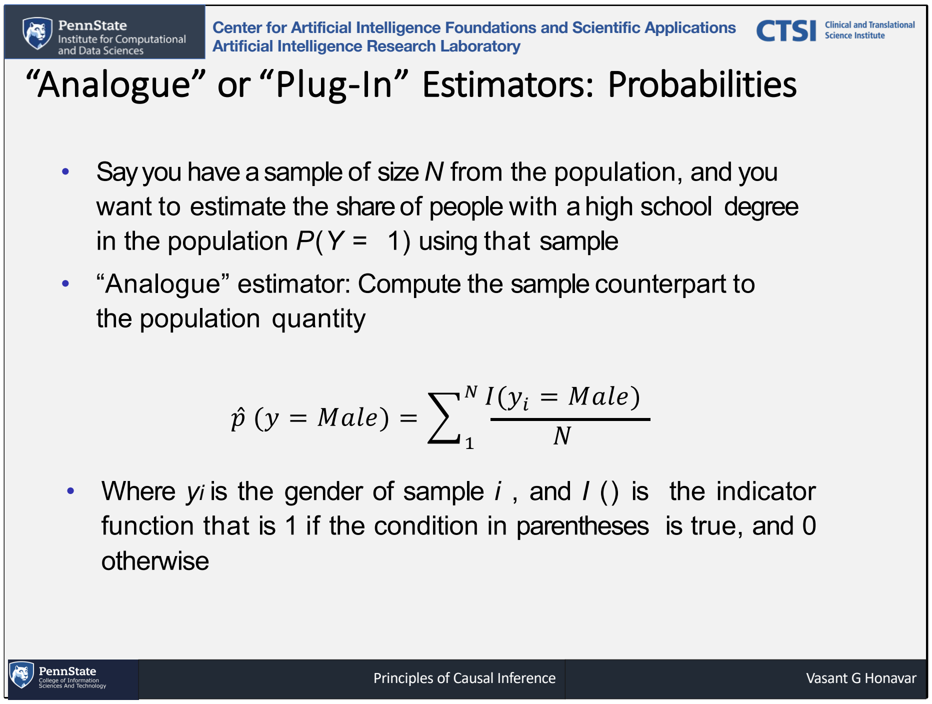
Remarks

- When we have all the population data or very large sample of it, we know $P(Y,X)$ and $Y = \alpha + \beta X + \in$ (at least with discrete X)
- A priori, knowing the distribution of X and Y perfectly does not tell us **anything** about whether and how X affects Y
- Nor does the regression of Y on X contain any useful information regarding whether and how X affects Y
- Regressions are not causal.
- Regressions are just conditional mean functions.



Summary

- Statistics = sampling from a population and inferring the characteristics of $P(Y\ ,\ X)$
- Using statistical tools, we cannot even talk about causality.
- Regressions per se have nothing to do with causal effects
- We have covered some necessary tools to understand population quantities like $P(Y, X)$ and $\mathrm{E}[Y\,|X]$
  - Causal inference is about learning from these observed
  - quantities about the consequences of actions, effects, and mechanisms, using causal assumptions
- Causal graphs depict our assumptions
  - "No causes in, no causes out" (Cartwright)

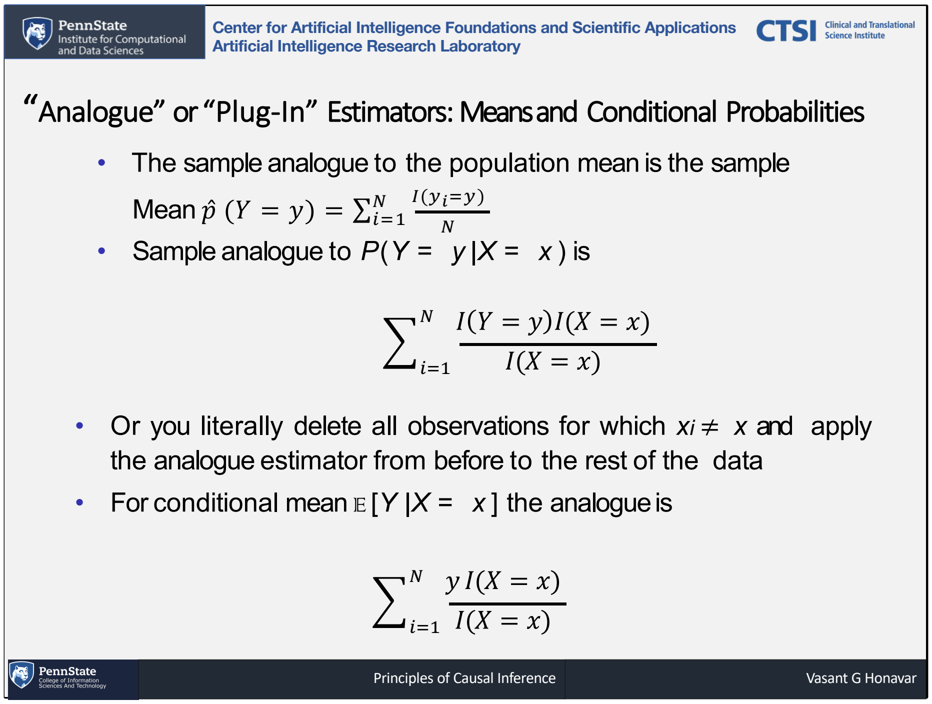


“Analogue” or “Plug-In” Estimators: Probabilities

- Say you have a sample of size *N* from the population, and you want to estimate the share of people with a high school degree in the population *P*(*Y* = 1) using that sample
- “Analogue” estimator: Compute the sample counterpart to the population quantity

$$\hat{p}\ (y = Male) = \sum_{1}^{N} \frac{I(y_i = Male)}{N}$$

- Where *yi* is the gender of sample *i* , and *I* () is the indicator function that is 1 if the condition in parentheses is true, and 0 otherwise



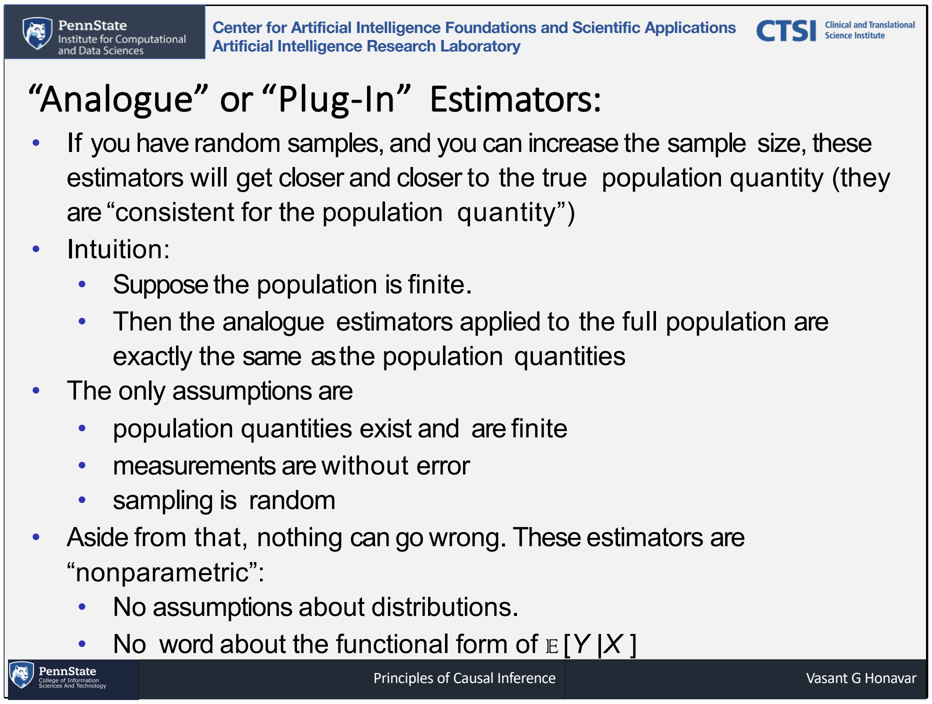
“Analogue” or “Plug-In” Estimators: Means and Conditional Probabilities

- The sample analogue to the population mean is the sample Mean $\hat{p}\ (Y = y) = \sum_{i=1}^{N} \frac{I(y_i=y)}{N}$
- Sample analogue to $P(Y = y | X = x)$ is

$$\sum\nolimits_{i=1}^{N} \frac{I(Y = y) I(X = x)}{I(X = x)}$$

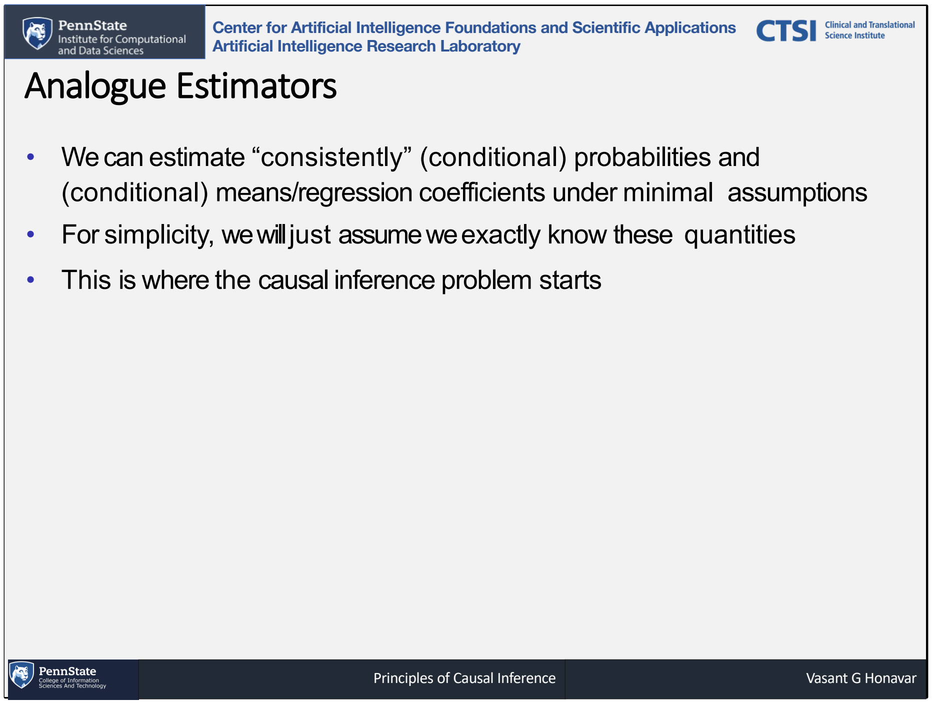
- Or you literally delete all observations for which $x_i \neq x$ and apply the analogue estimator from before to the rest of the data
- For conditional mean $\mathbb{E}[Y | X = x]$ the analogue is

$$\sum\nolimits_{i=1}^{N} \frac{y\, I(X = x)}{I(X = x)}$$



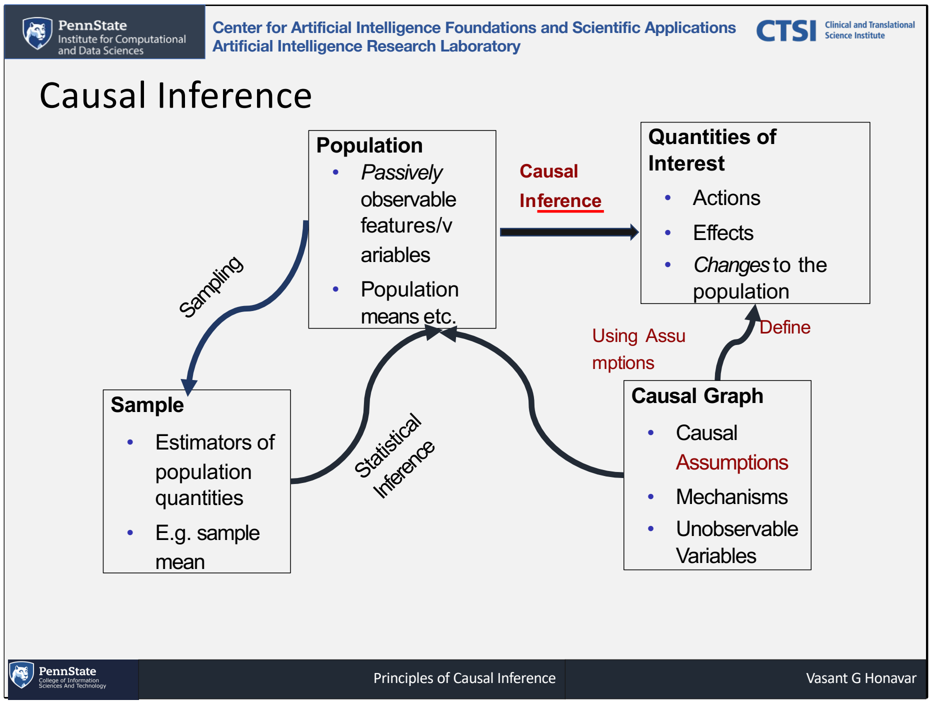
“Analogue” or “Plug-In” Estimators:

- If you have random samples, and you can increase the sample size, these estimators will get closer and closer to the true population quantity (they are “consistent for the population quantity”)
- Intuition:
  - Suppose the population is finite.
  - Then the analogue estimators applied to the full population are exactly the same as the population quantities
- The only assumptions are
  - population quantities exist and are finite
  - measurements are without error
  - sampling is random
- Aside from that, nothing can go wrong. These estimators are “nonparametric”:
  - No assumptions about distributions.
  - No word about the functional form of $\mathbb{E}[Y|X]$



Analogue Estimators

- We can estimate “consistently” (conditional) probabilities and (conditional) means/regression coefficients under minimal assumptions
- For simplicity, we will just assume we exactly know these quantities
- This is where the causal inference problem starts



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