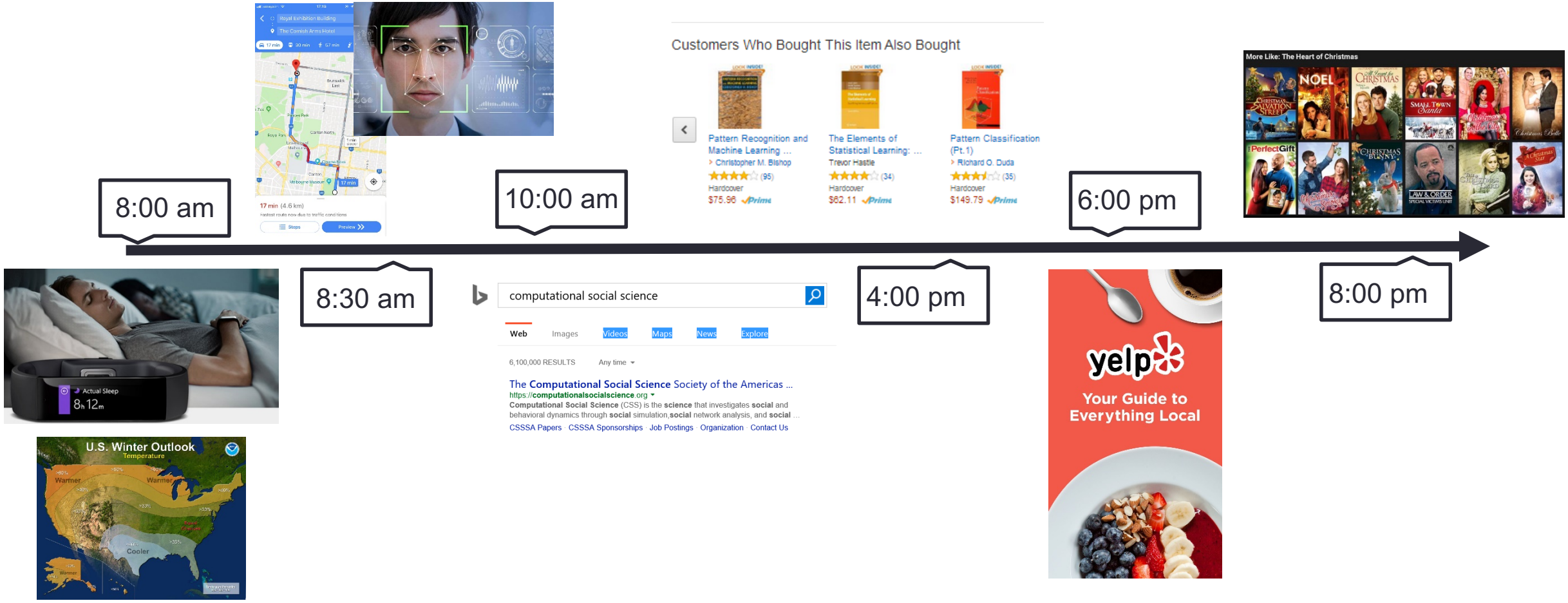
Causal Inference and Stable Learning

Peng Cui
Tsinghua University

Zheyan Shen
Tsinghua University

ML techniques are impacting our life

- A day in our life with ML techniques



Now we are stepping into risk-sensitive areas

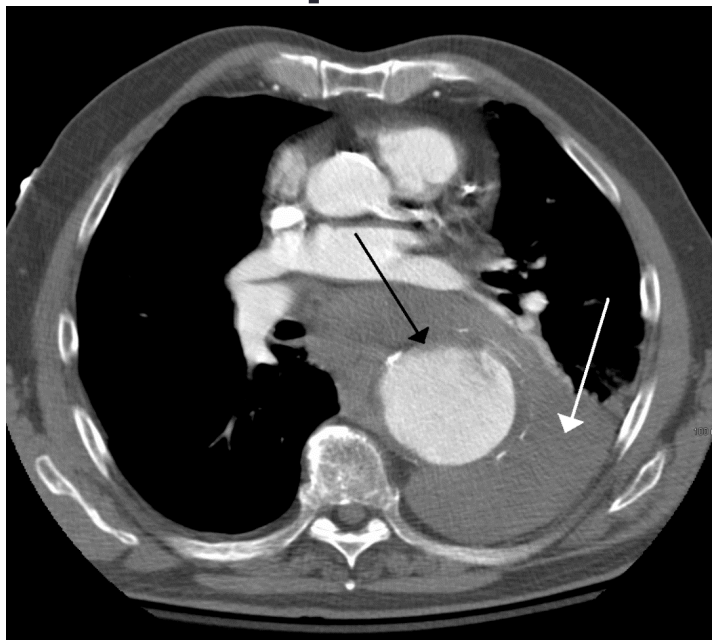
Human

Shifting from ***Performance Driven*** to ***Risk Sensitive***

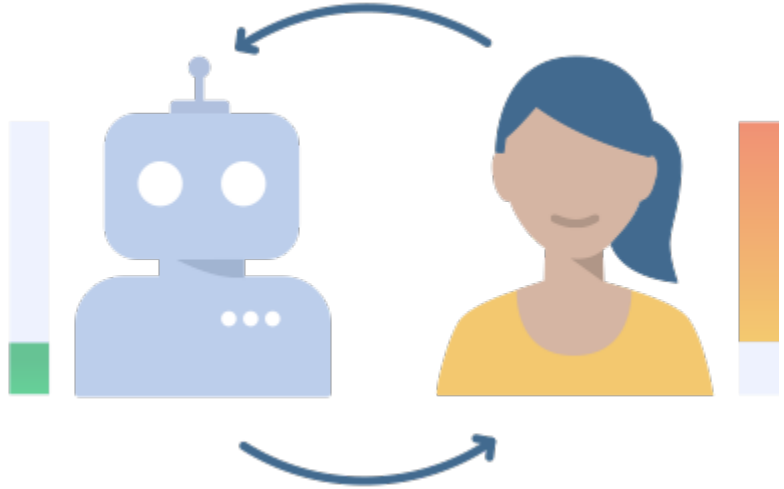
Problems of today's ML - ***Explainability***

Most machine learning models are black-box models

Unexplainable



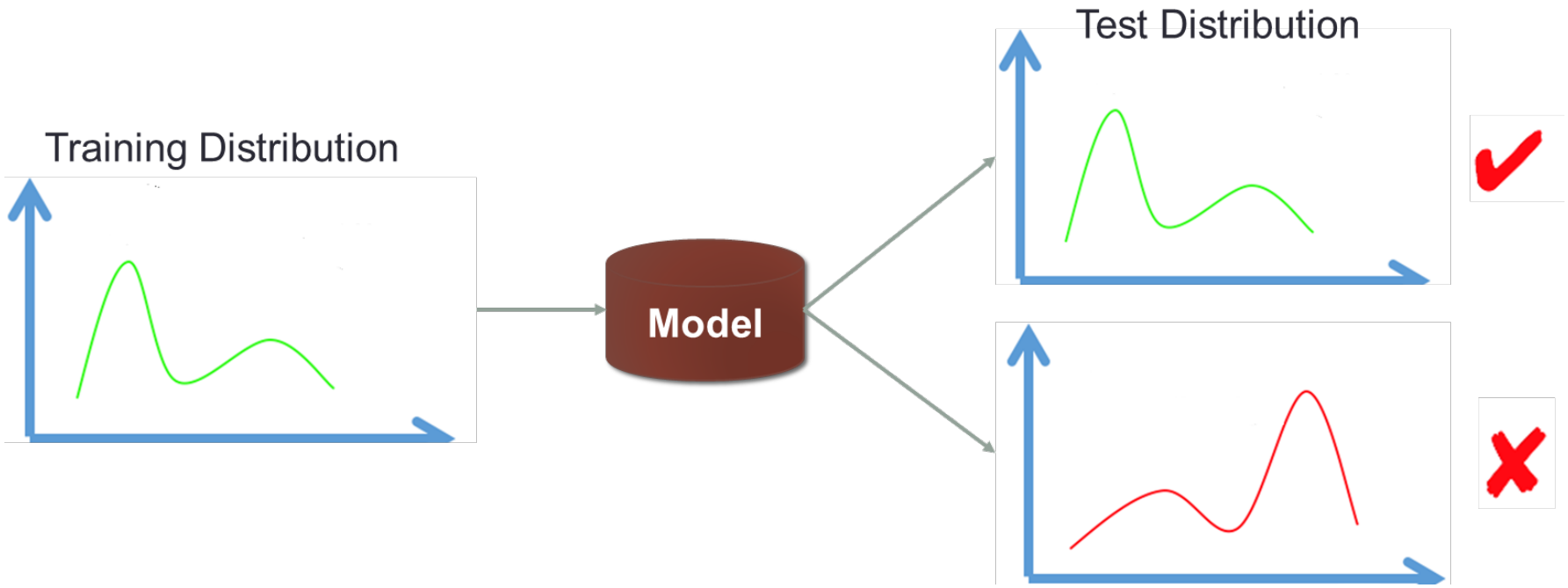
Human in the loop



Health Military Finance Industry

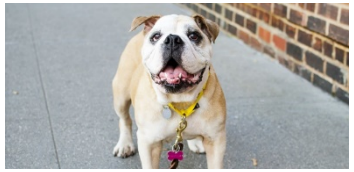
Problems of today's ML - ***Stability***

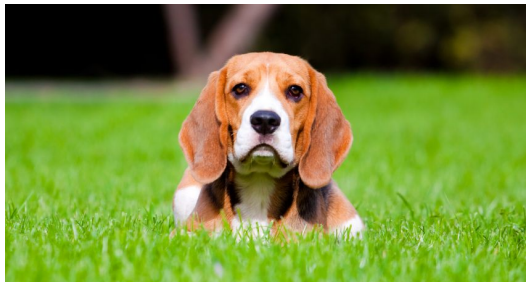
Most ML methods are developed under I.I.D hypothesis



Problems of today's ML - ***Stability***





Yes

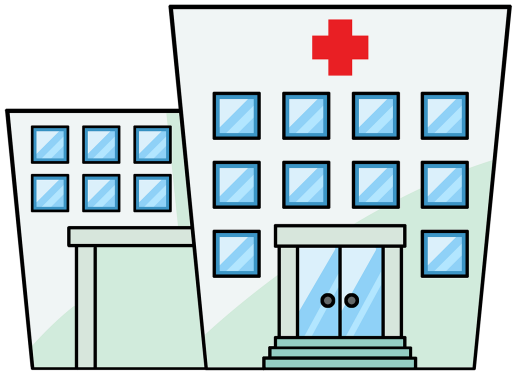
Maybe

No

Problems of today's ML - ***Stability***

- Cancer survival rate prediction

Training Data

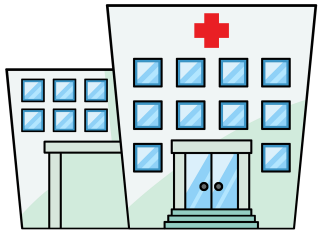


City Hospital

Higher income, higher survival rate.

Predictive Model

Testing Data



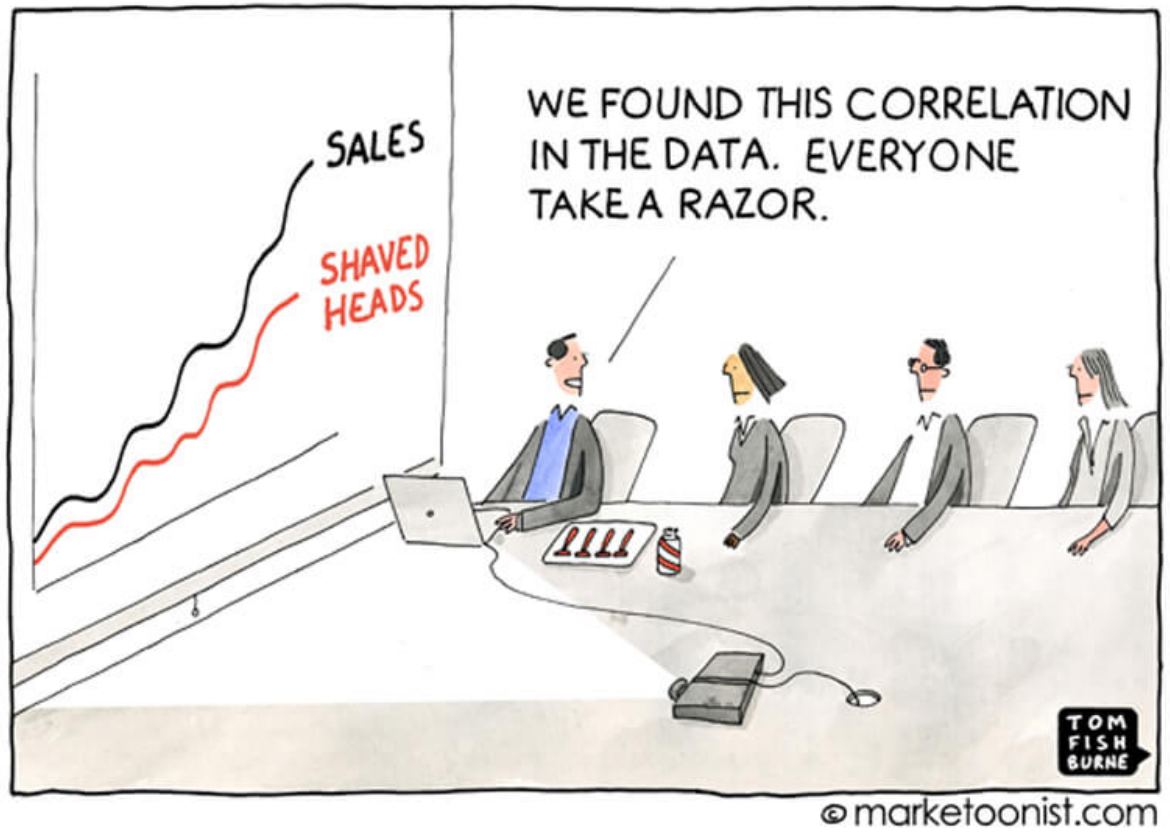
City Hospital

University Hospital

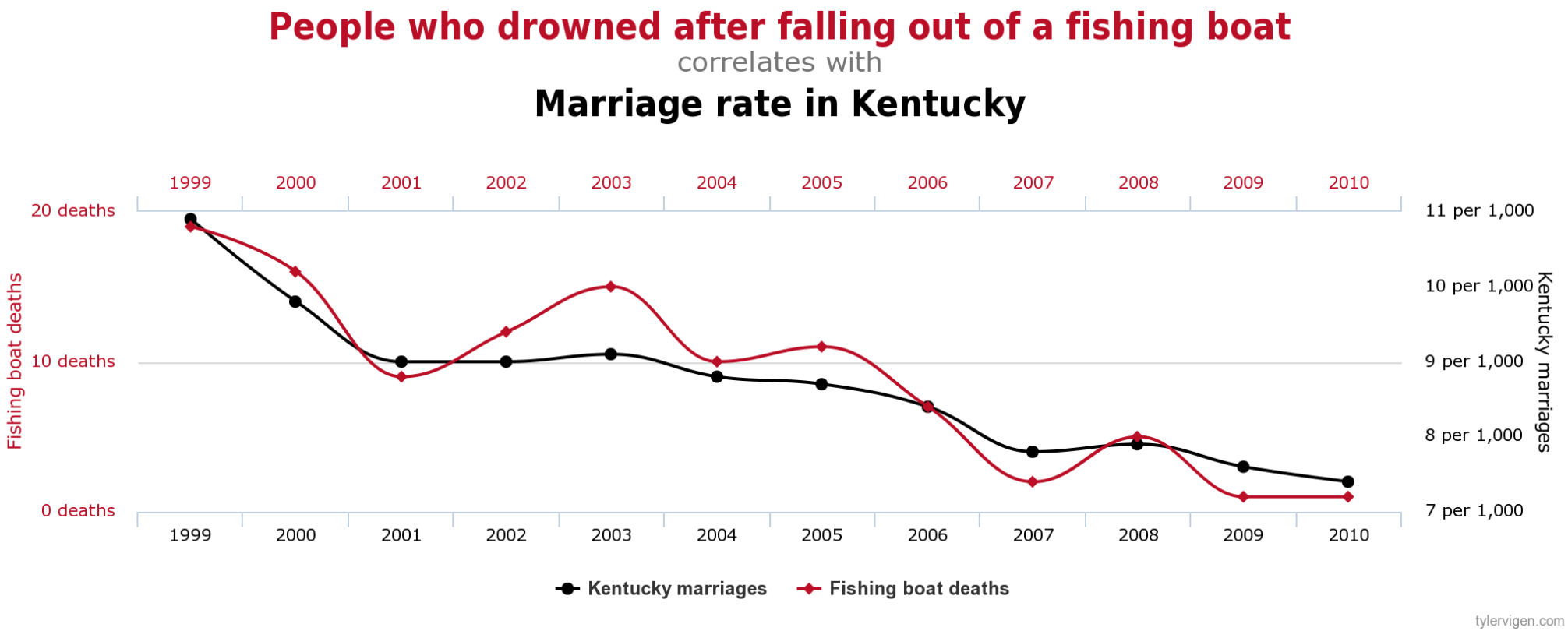
Survival rate is not so correlated with income.

A plausible reason: ***Correlation***

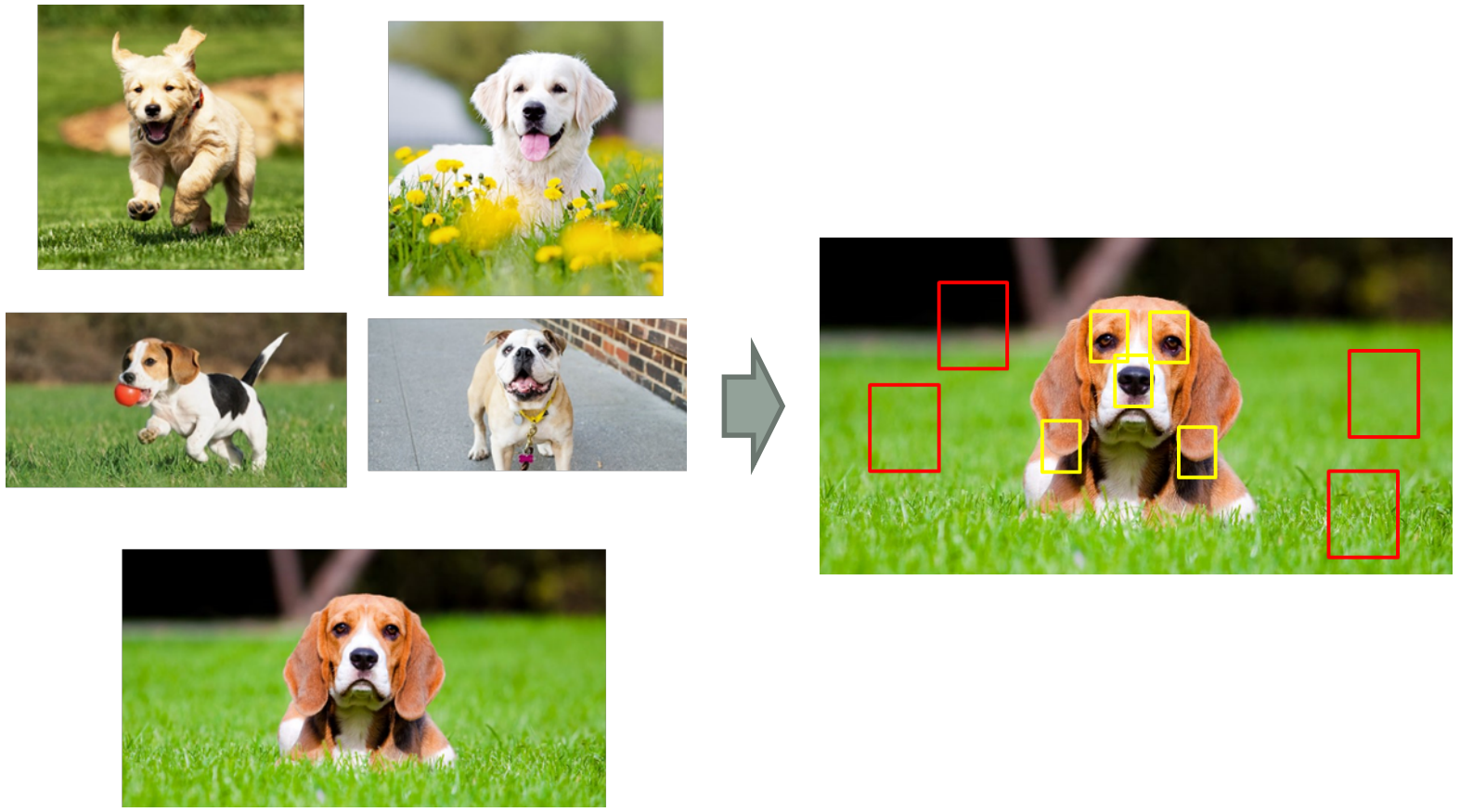
Correlation is the very basics of machine learning.



Correlation is not explainable

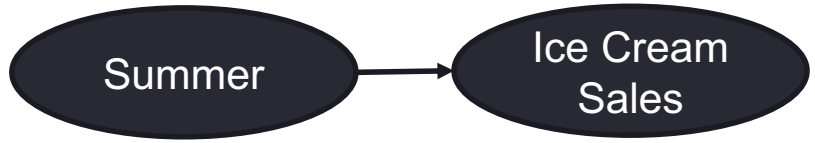


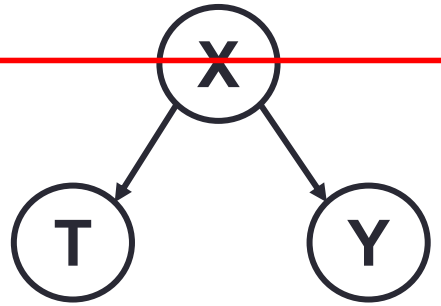
Correlation is '*unstable*'

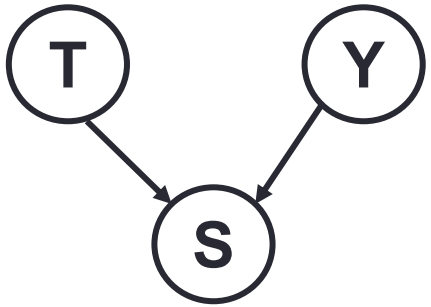


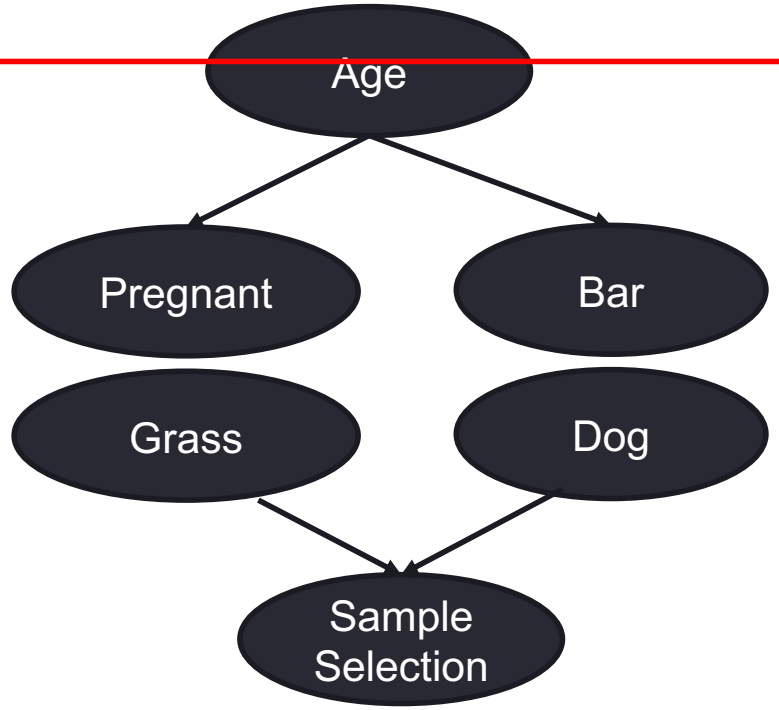
It's not the fault of ***correlation***, but the way we use it

- Three sources of correlation:
  - Causation
    - Causal mechanism
    - Stable and explainable
  - Confounding
    - Ignoring X
    - Spurious Correlation
  - Sample Selection Bias
    - Conditional on S
    - Spurious Correlation







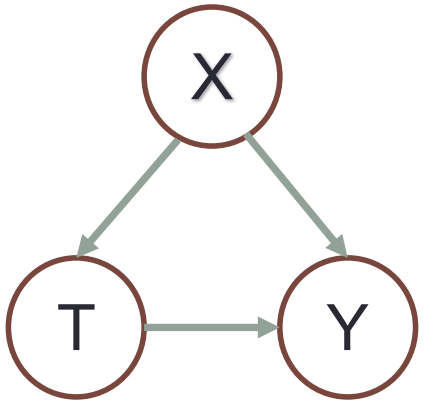


A Practical Definition of Causality

Definition: T causes Y if and only if

changing T leads to a change in Y,

while keeping everything else constant.



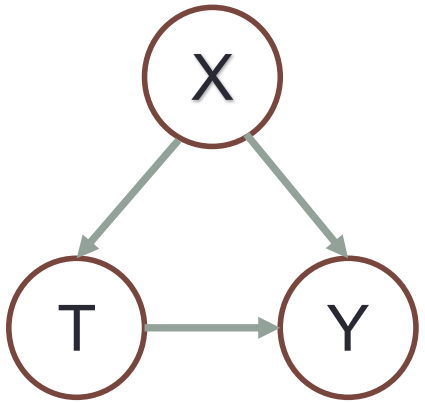
Causal effect is defined as the magnitude by which Y is changed by a unit change in T.

Called the “interventionist” interpretation of causality.

**Interventionist* definition [http://plato.stanford.edu/entries/causation-mani/]

The ***benefits*** of bringing causality into learning

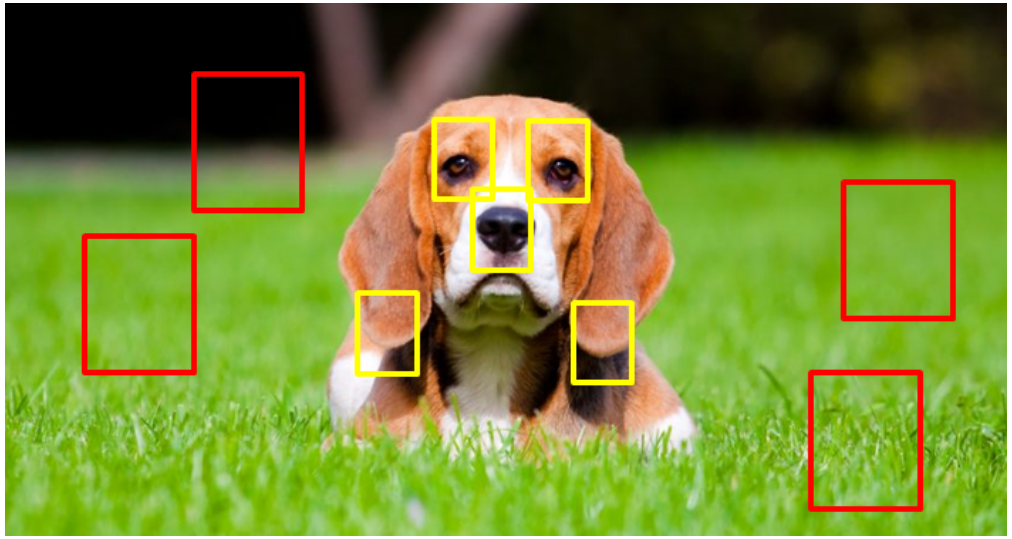
Causal Framework



T：grass
X：dog nose
Y：label

Grass—Label: Strong correlation
Weak causation

Dog nose—Label: Strong correlation
Strong causation



More ***Explainable*** and More ***Stable***

The ***gap*** between causality and learning

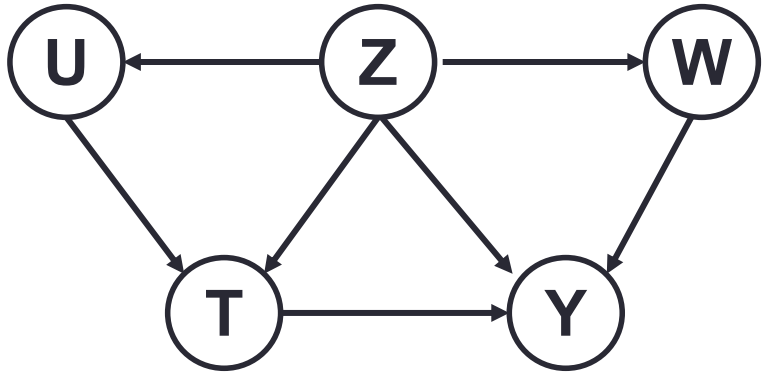
- How to evaluate the outcome?
- Wild environments
  - High-dimensional
  - Highly noisy
  - Little prior knowledge (model specification, confounding structures)
- Targeting problems
  - Understanding v.s. Prediction
  - Depth v.s. Scale and Performance

How to bridge the gap between *causality* and *(stable) learning*?

Paradigms - Structural Causal Model

A graphical model to describe the causal mechanisms of a system

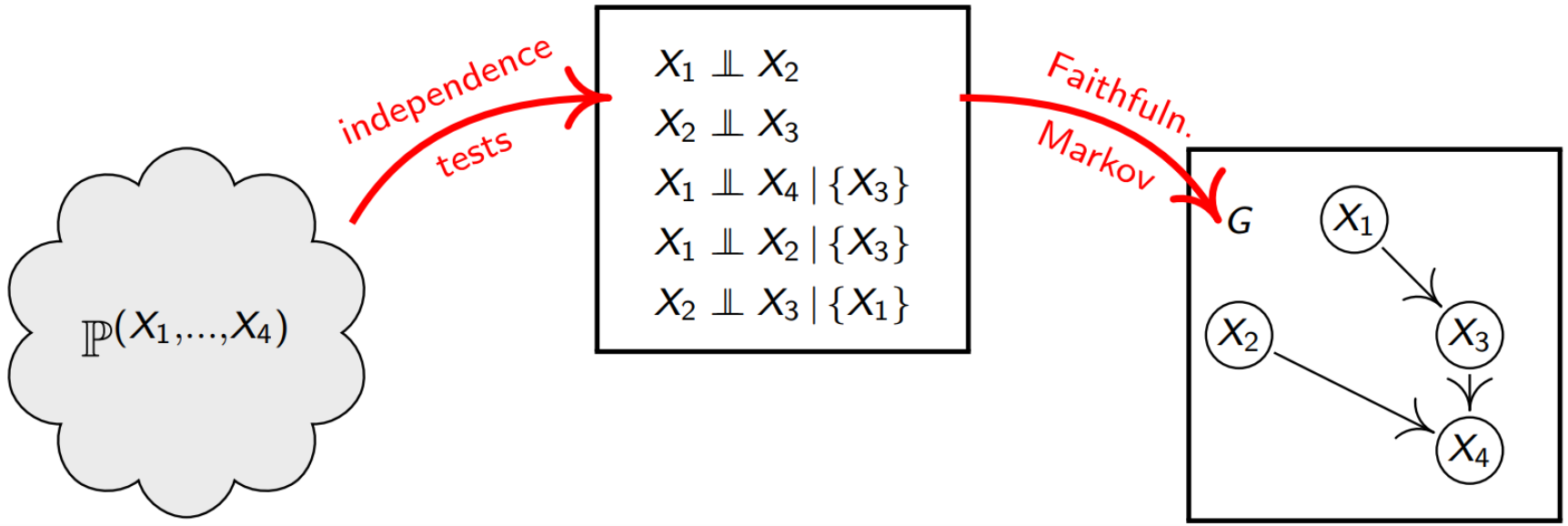
- Causal Identification with back door criterion
- Causal Estimation with do calculus



How to discover the causal structure?

Paradigms – Structural Causal Model

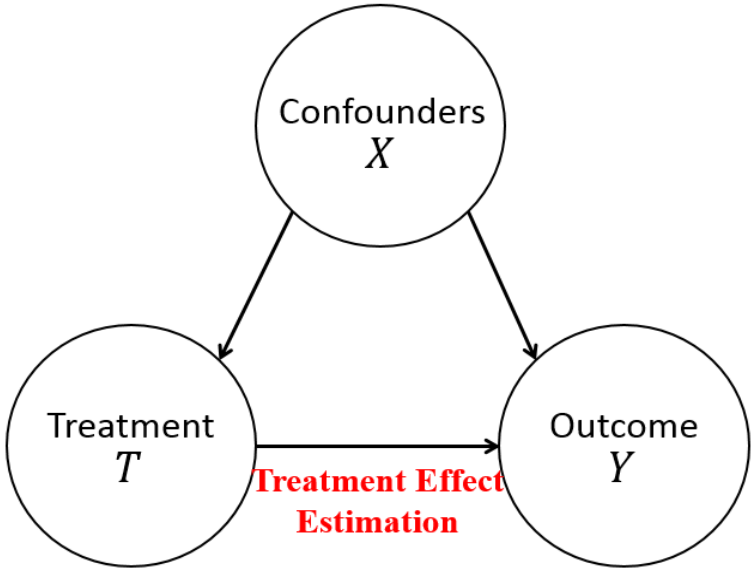
- Causal Discovery
  - Constraint-based: conditional independence
  - Functional causal model based



A ***generative*** model with strong expressive power.
But it induces high complexity.

Paradigms - Potential Outcome Framework

- A simpler setting
  - Suppose the confounders of T are known a priori



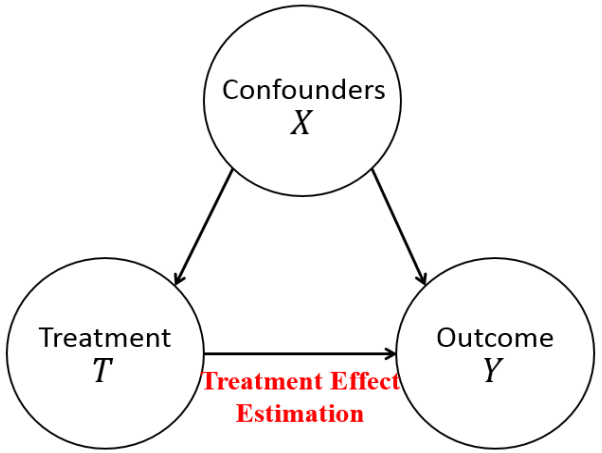
- The computational complexity is affordable
  - Under stronger assumptions
  - E.g. all confounders need to be observed

More like a ***discriminative*** way to estimate treatment's partial effect on outcome.

Causal Effect Estimation

- Treatment Variable: $T = 1$ or $T = 0$
- Treated Group $(T = 1)$ and Control Group $(T = 0)$
- Potential Outcome: $Y(T = 1)$ and $Y(T = 0)$
- Average Causal Effect of Treatment (ATE):

$$ATE = E[Y(T = 1) - Y(T = 0)]$$



Counterfactual Problem

Person	T	$Y_{T=1}$	$Y_{T=0}$
P1	1	0.4	?
P2	0	?	0.6
P3	1	0.3	?
P4	0	?	0.1
P5	1	0.5	?
P6	0	?	0.5
P7	0	?	0.1

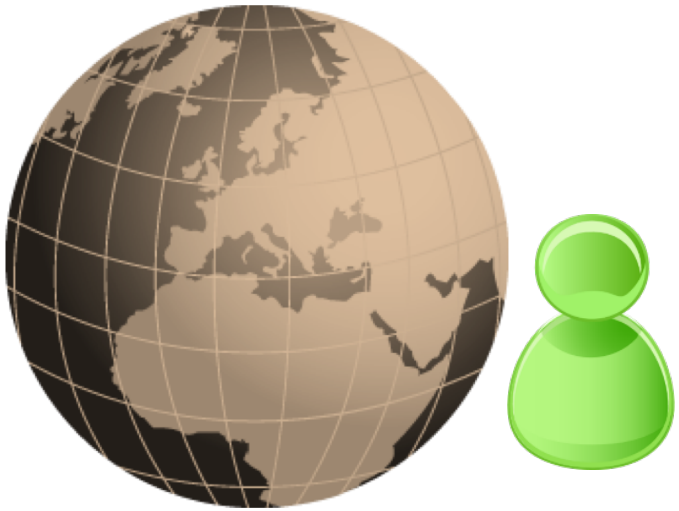
- Two key points for causal effect estimation
  - Changing T
  - Keeping everything else constant

- For each person, observe only one: either $Y_{t=1}$ or $Y_{t=0}$
- For different group (T=1 and T=0), something else are not constant

Ideal Solution: Counterfactual World

- Reason about a world that does not exist
- Everything in the counterfactual world is the same as the real world, except the treatment

$Y(T = 1)$



$Y(T = 0)$

Randomized Experiments are the "Gold Standard"

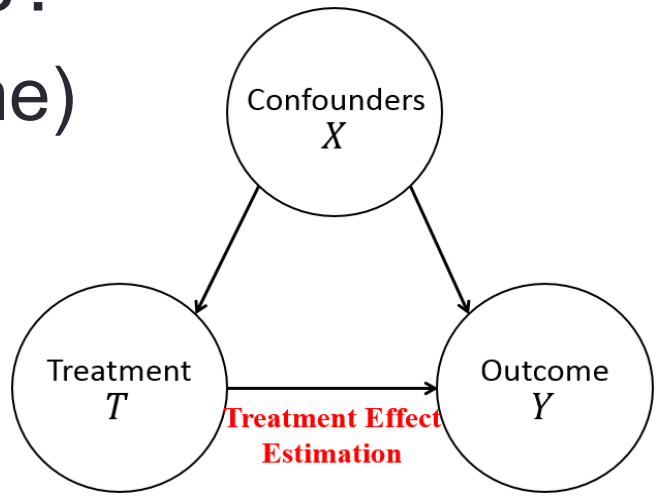
- Drawbacks
  - Cost
  - Unethical
  - Unrealistic

Causal Inference with Observational Data

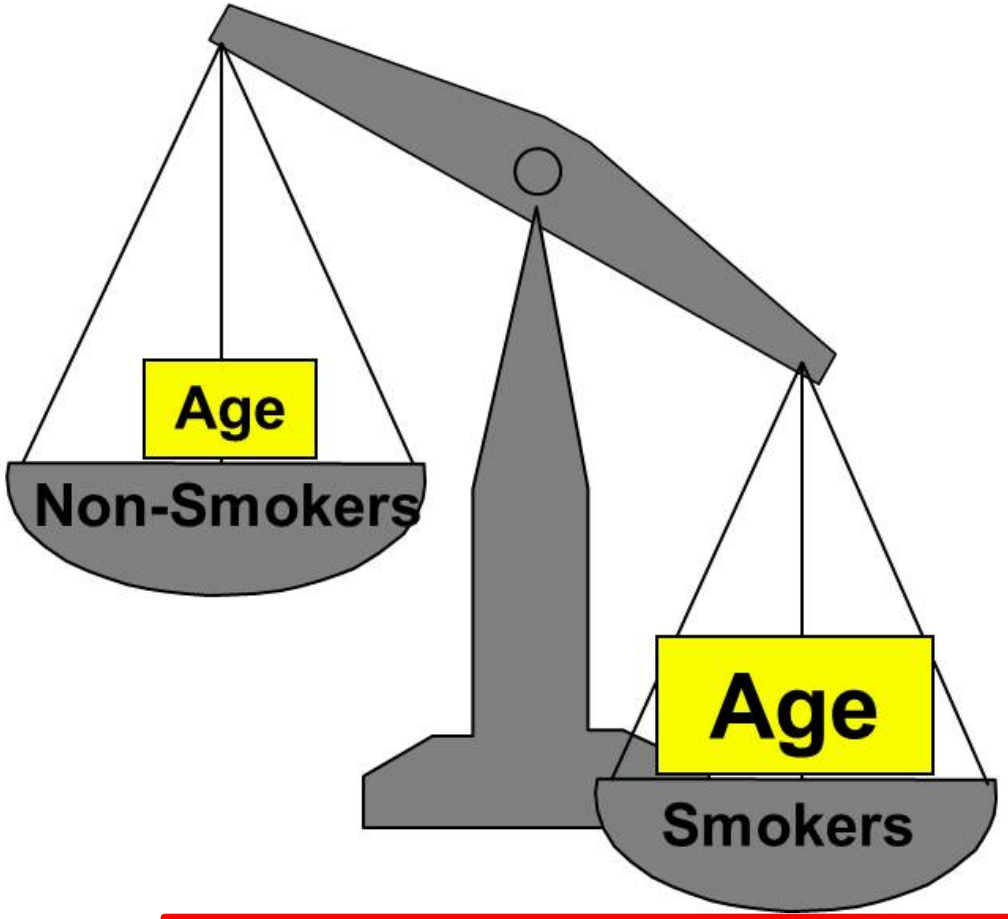
- Counterfactual Problem:

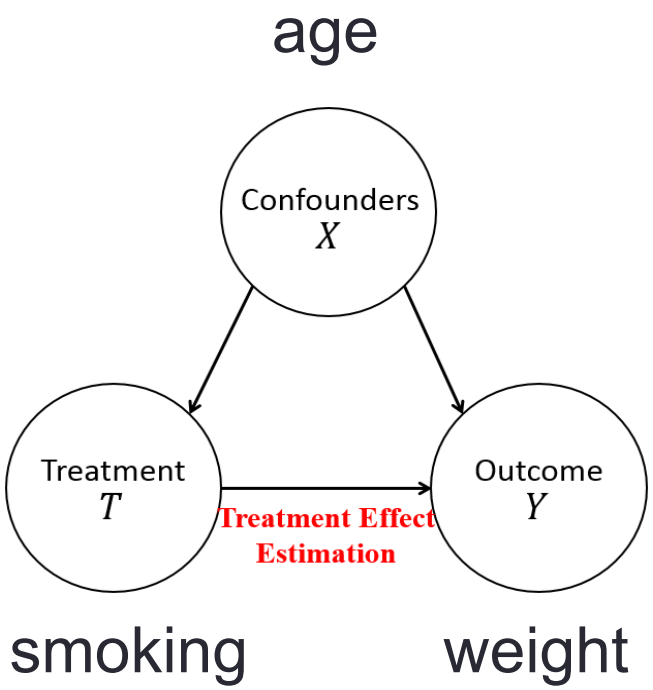
$$Y(T=1) \quad \text{or} \quad Y(T=0)$$

- Can we estimate ATE by directly comparing the average outcome between treated and control groups?
  - Yes with randomized experiments (X are the same)
  - No with observational data (X might be different)



Confounding Effect



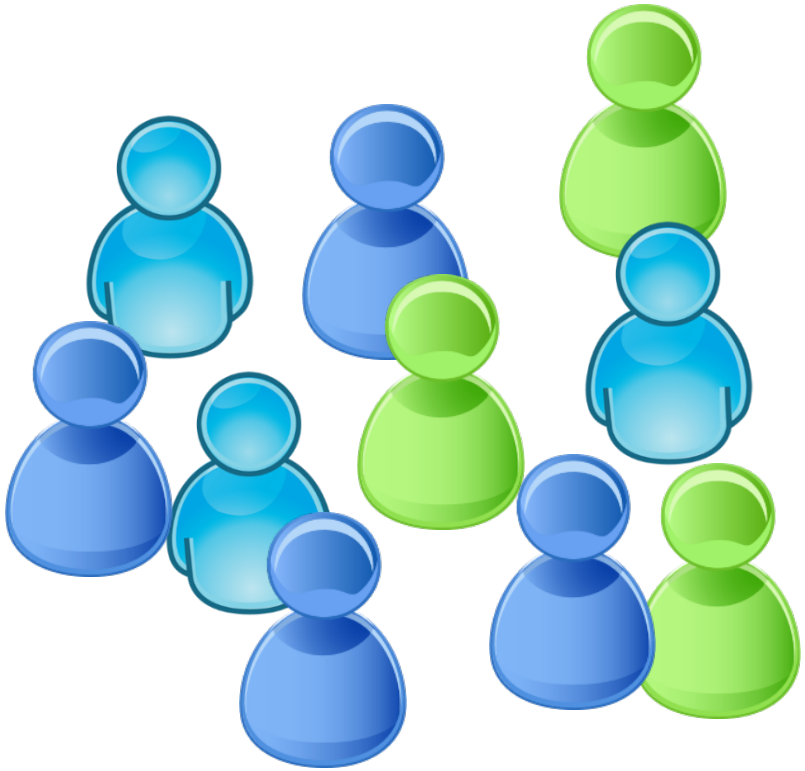


Balancing Confounders' Distribution

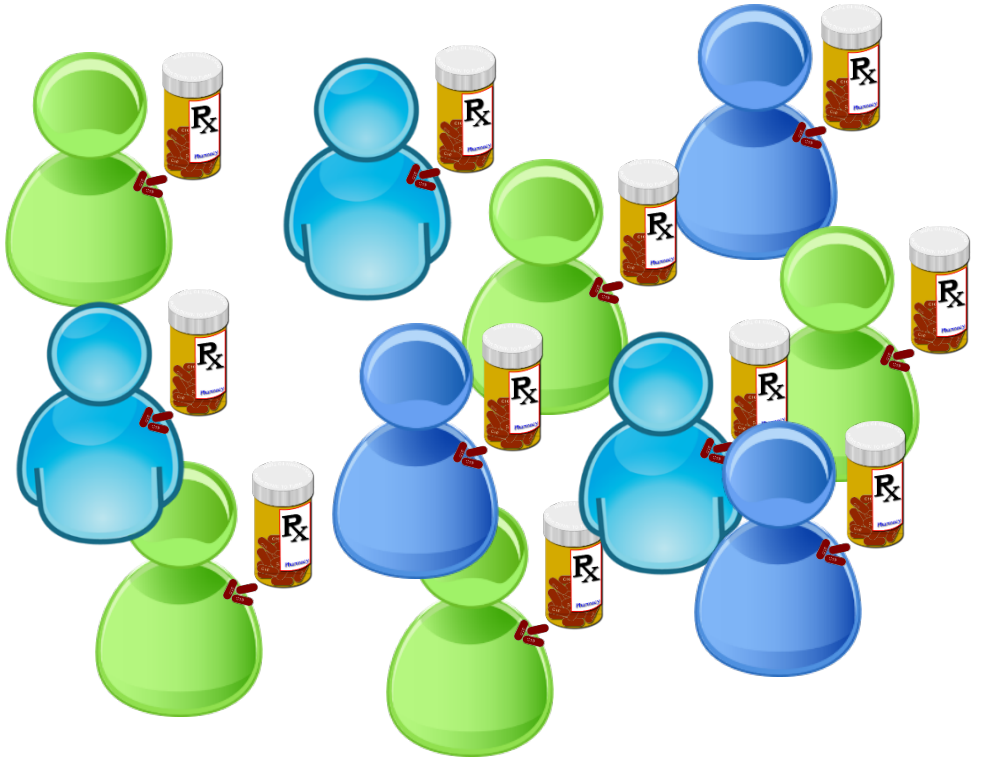
Methods for Causal Inference

- **Matching**
- **Propensity Score**
- **Directly Confounder Balancing**

Matching

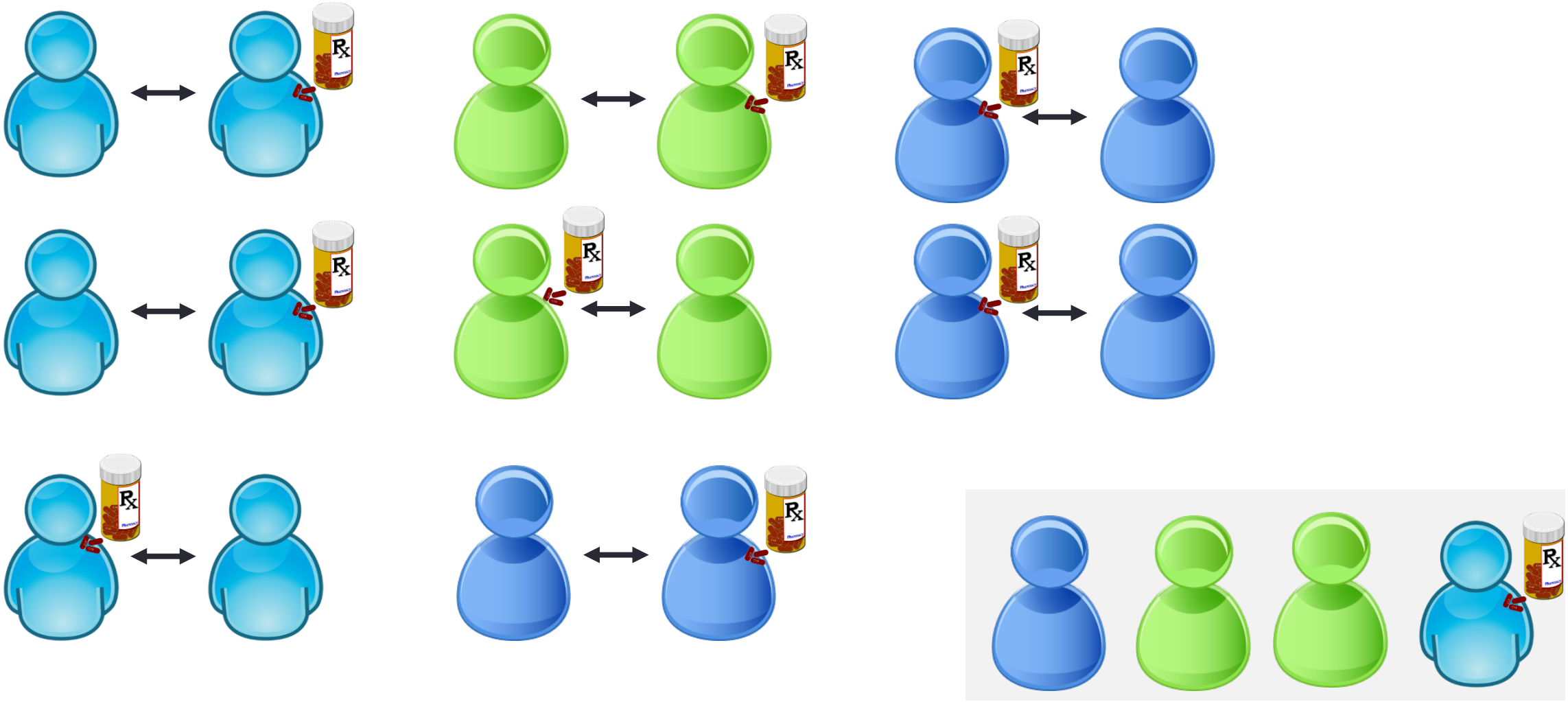


$T = 0$



$T = 1$

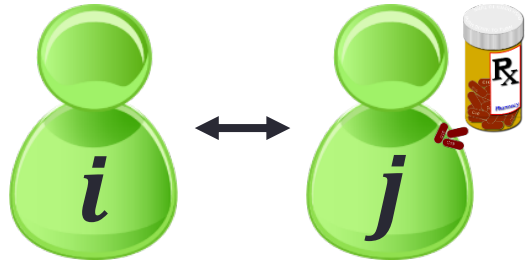
Matching



Matching

- Identify pairs of treated (T=1) and control (T=0) units whose confounders X are similar or even identical to each other

$$Distance(X_i, X_j) \leq \epsilon$$



- Paired units guarantee that the everything else (Confounders) approximate constant
- Small ϵ: less bias, but higher variance
- Fit for low-dimensional settings
- But in high-dimensional settings, there will be few exact matches

Methods for Causal Inference

- **Matching**
- **Propensity Score**
- **Directly Confounder Balancing**

Propensity Score Based Methods

- Propensity score $e(X)$ is the probability of a unit to get treated

$$e(X) = P(T = 1|X)$$

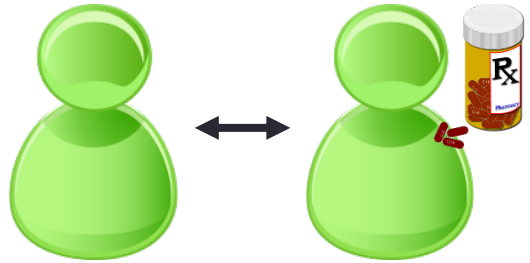
- Then, Donald Rubin shows that the propensity score is sufficient to control or summarize the information of confounders

$$T \perp\!\!\!\perp X \mid e(X) \quad \Rightarrow \quad T \perp\!\!\!\perp (Y(1), Y(0)) \mid e(X)$$

- Propensity scores cannot be observed, need to be estimated

Propensity Score Matching

- Estimating propensity score: $\hat{e}(X) = P(T = 1|X)$
  - **Supervised learning**: predicting a known label T based on observed covariates X.
  - Conventionally, use logistic regression



- Matching pairs by distance between propensity score:

$$Distance(X_i, X_j) \leq \epsilon$$

$$Distance(X_i, X_j) = |\hat{e}(X_i) - \hat{e}(X_j)|$$

- High dimensional challenge: from matching to PS estimation
- But this is a 'hard' solution.

P. C. Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate behavioral research, 46(3):399–424, 2011.

Inverse of Propensity Weighting (IPW)

- Why weighting with inverse of propensity score?
  - Propensity score induces the distribution bias on confounders X

$$e(X) = P(T = 1|X)$$

Unit	$e(X)$	$1 - e(X)$	#units	#units (T=1)	#units (T=0)
A	0.7	0.3	10	7	3
B	0.6	0.4	50	30	20
C	0.2	0.8	40	8	32

Distribution Bias

Unit	#units (T=1)	#units (T=0)
A	10	10
B	50	50
C	40	40

Confounders are the same!

Reweighting by inverse of propensity score: $w_i = \frac{T_i}{e_i} + \frac{1 - T_i}{1 - e_i}$

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1):41–55, 1983.

Inverse of Propensity Weighting (IPW)

- Estimating ATE by IPW [1]:

$$w_i = \frac{T_i}{e_i} + \frac{1 - T_i}{1 - e_i}$$

$$ATE_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\frac{T_i Y_i}{\hat{e}(X_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1 - T_i)Y_i}{1 - \hat{e}(X_i)}$$

  - Interpretation: IPW creates a pseudo-population where the confounders are the same between treated and control groups.
- But requires correct model specification for propensity score
- High variance when e is close to 0 or 1

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1):41–55, 1983.

Non-parametric solution

- Model specification problem is inevitable
- Can we directly learn sample weights that can balance confounders' distribution between treated and control groups?

Methods for Causal Inference

- **Matching**
- **Propensity Score**
- **Directly Confounder Balancing**

Directly Confounder Balancing

- **Motivation**: The collection of all the moments of variables uniquely determine their distributions.
- **Methods**: Learning sample weights by directly balancing confounders' moments as follows (ATT problem)

$$\min_{W} \| \overline{\mathbf{X}}_t - \mathbf{X}_c^T W \|_2^2$$

The first moments of X on the Treated Group

The first moments of X on the Control Group

With moments, the sample weights can be learned without any model specification.

J. Hainmueller. Entropy balancing for causal effects: A mul- tivariate reweighting method to produce balanced samples in observational studies. Political Analysis, 20(1):25–46, 2012.

Entropy Balancing

$$\begin{aligned} \min_{W} \quad & W \log(W) \\ s.t. \quad & \|\overline{\mathbf{X}}_t - \mathbf{X}_c^T W\|_2^2 = 0 \\ & \sum_{i=1}^{n} W_i = 1, W \succeq 0 \end{aligned}$$

- Directly confounder balancing by sample weights W
- Minimize the entropy of sample weights W

Either know confounders a priori or regard all variables as confounders .
All confounders are balanced equally.

Athey S, et al. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. Journal of the Royal Statistical Society: Series B, 2018, 80(4): 597-623.

Differentiated Confounder Balancing

- **Idea**: Different confounders make different confounding bias
- Simultaneously learn ***confounder weights β*** and ***sample weighs W***.

$$\min \quad \left(\beta^T \cdot (\overline{\mathbf{X}}_t - \mathbf{X}_c^T W)\right)^2$$

- ***Confounder weights*** determine which variable is confounder and its contribution on confounding bias.
- ***Sample weights*** are designed for confounder balancing.

Kun Kuang, Peng Cui, et al. 2017. Estimating Treatment Effect in the Wild via Differentiated Confounder Balancing, KDD 2017, 265–274.

Assumptions of Causal Inference

- **A1: Stable Unit Treatment Value (SUTV):** The effect of treatment on a unit is independent of the treatment assignment of other units

$$P\left(Y_i \middle| T_i, T_j, X_i\right) = P(Y_i | T_i, X_i)$$

- **A2: Unconfounderness:** The distribution of treatment is independent of potential outcome when given the observed variables

$$T \perp \left(Y(0), Y(1)\right) | \, X$$

No unmeasured confounders

- **A3: Overlap:** Each unit has nonzero probability to receive either treatment status when given the observed variables

$$0 < P(T = 1 | X = x) < 1$$

Sectional Summary

- Progress has been made to draw causality from big data.
- From single to group
- From binary to continuous
- Weak assumptions

Ready for Learning?