Towards a Learning Theory of Cause-Effect Inference

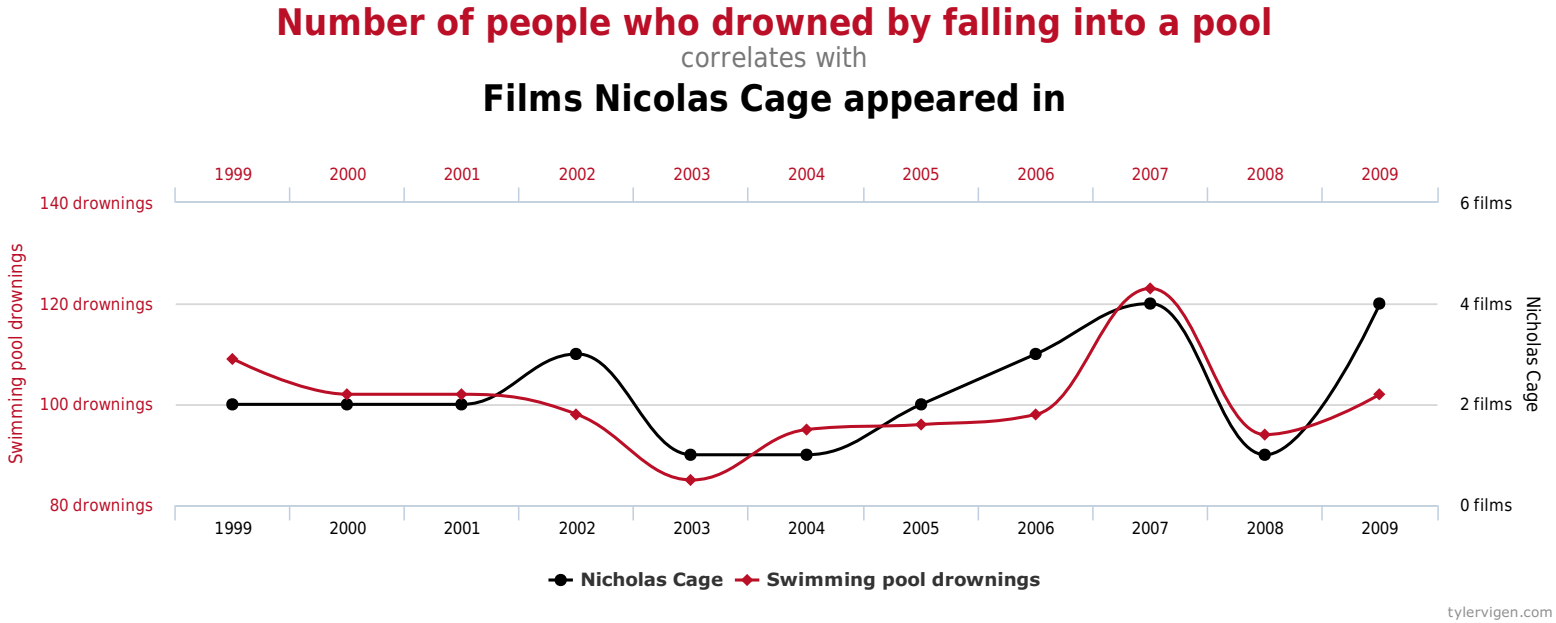
Lopez–Paz[1,2], Muandet[1], Schölkopf[1], Tolstikhin[1]

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany

[2]University of Cambridge, Cambridge, United Kingdom

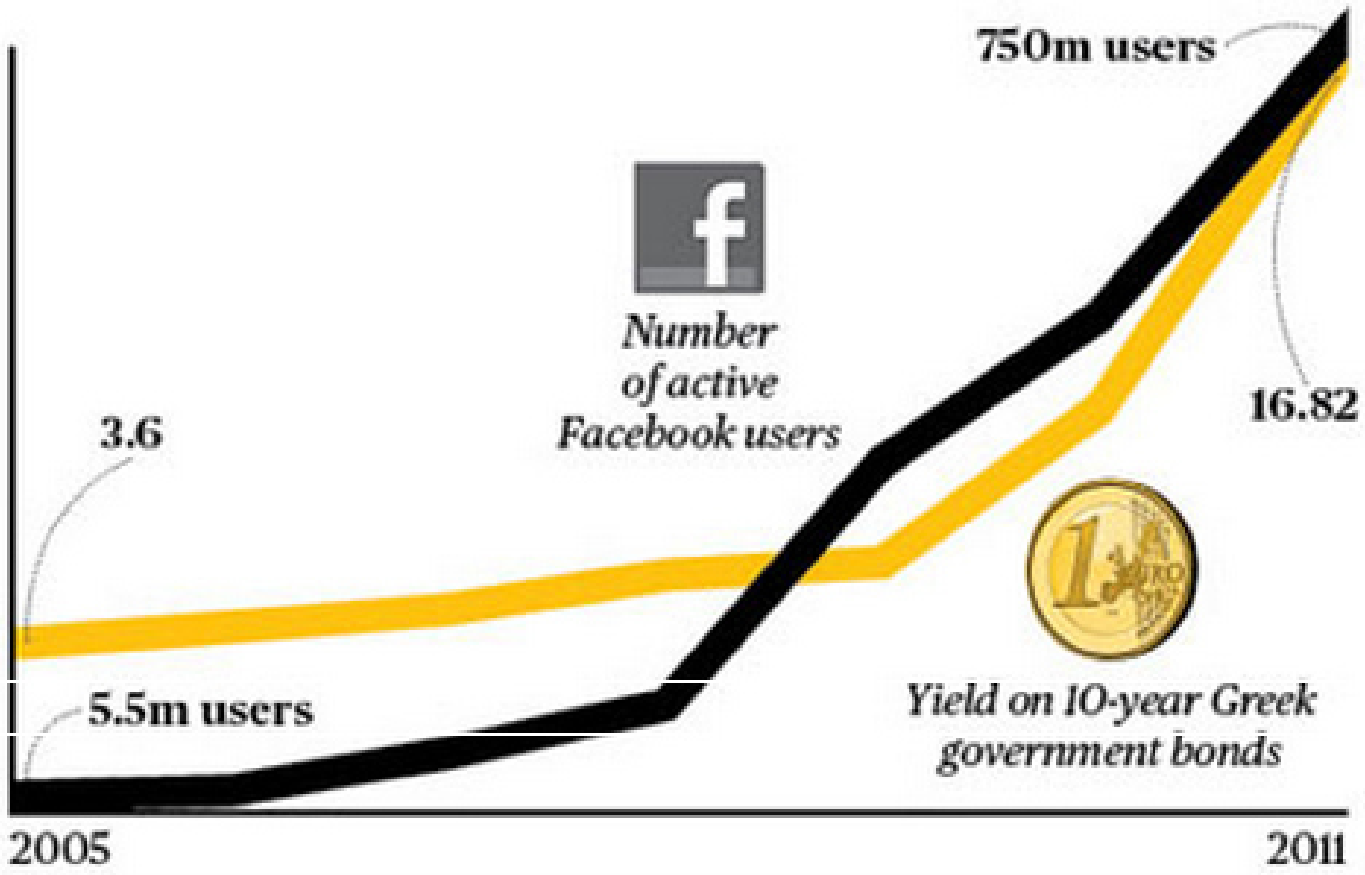
correlation $\not\Rightarrow$ causation

Observe Intervene



[0]http://www.tylervigen.com/spurious-correlations

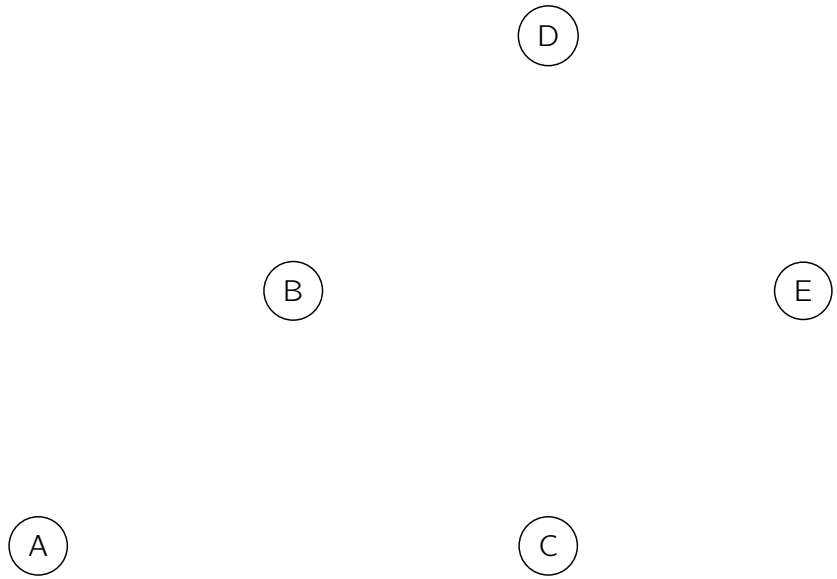
Is Facebook driving the Greek debt crisis?



[0] Bloomberg BusinessWeek.com

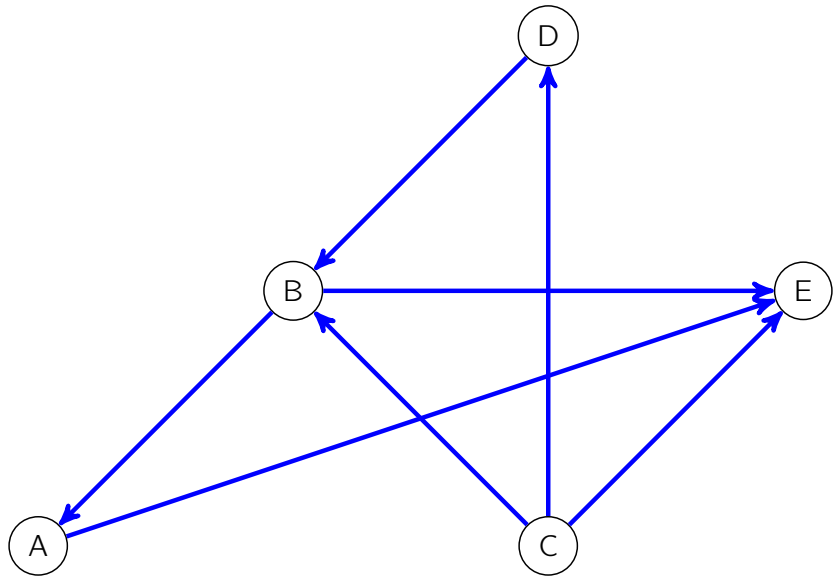
Causal Inference in a Nutshell

Given the i.i.d. sample $\{(A_i, B_i, C_i, D_i, E_i)\}_{i=1}^n$



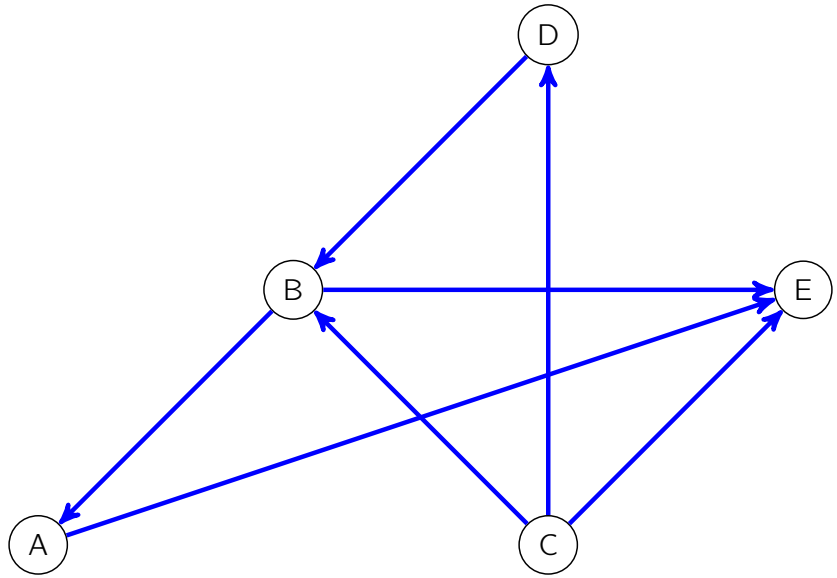
Causal Inference in a Nutshell

Given the i.i.d. sample $\{(A_i, B_i, C_i, D_i, E_i)\}_{i=1}^n$



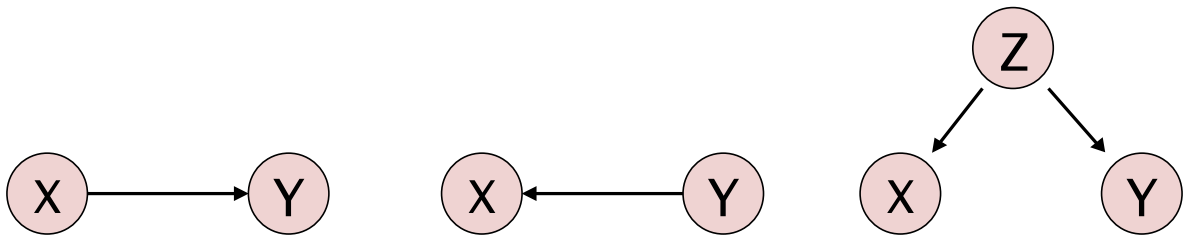
Causal Inference in a Nutshell

Given the i.i.d. sample $\{(A_i, B_i, C_i, D_i, E_i)\}_{i=1}^n$



Reichenbach's Principle

A dependency between X and Y implies



Previous Works, SGS/PC[1]

Idea: given a universe of variables $\mathcal{U} = \{X_1, \ldots, X_d\}$, X_i and X_j $(i \neq j)$ are causally related iff $\nexists\, S \subseteq (\mathcal{U} \setminus \{X_i, X_j\})$ st $X_i \not\perp\!\!\!\perp X_j | S$.

Assumptions:

- *Markov*: $d-$separation implies cond. ind.
- *faithfulness*: no cond. ind. other than those from $d-$seps.
- *sufficiency*: no latent variable

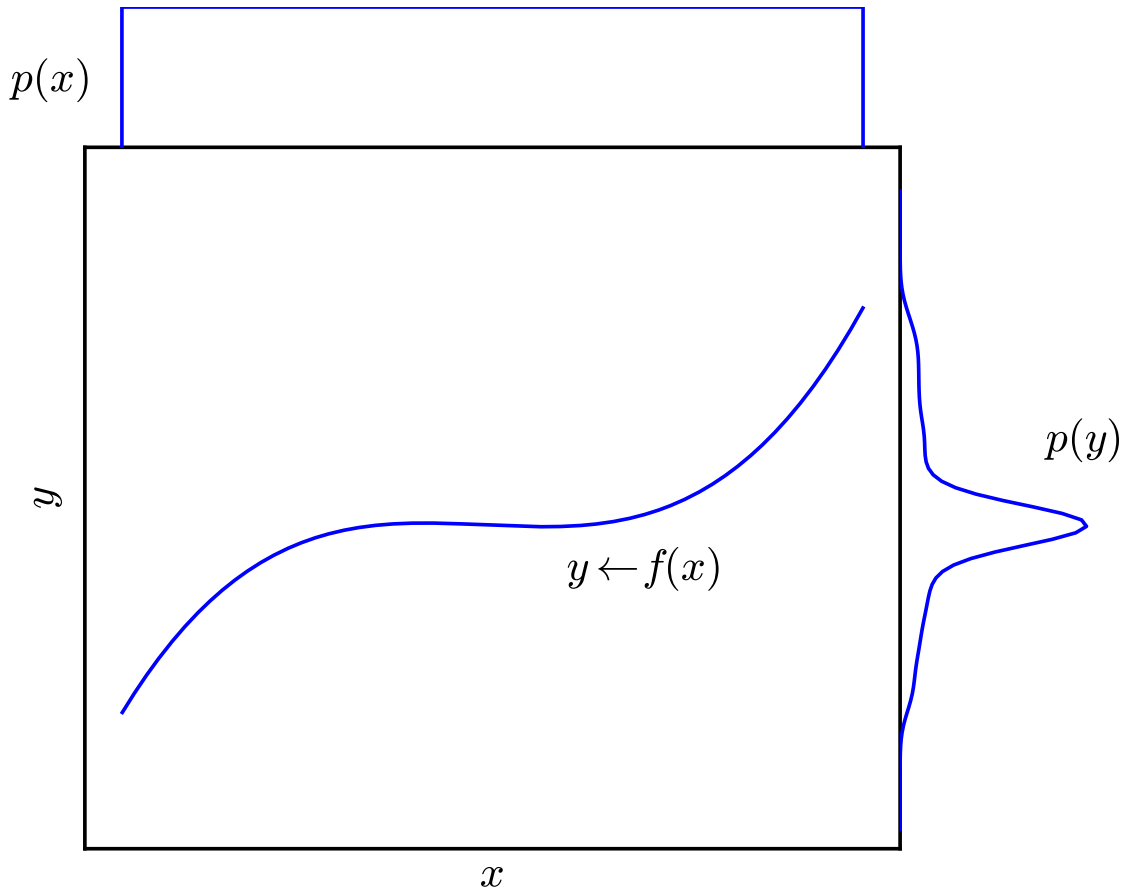
Limitations:

- equivalences $X \to Y \to Z$, $X \leftarrow Y \leftarrow Z$, $X \leftarrow Y \to Z$
- high-dimensional conditional independence tests
- not applicable to the two-variable case

[1] Spirtes et al., *Causation, prediction, and search*, 2000

Previous Works, IGCI[2]

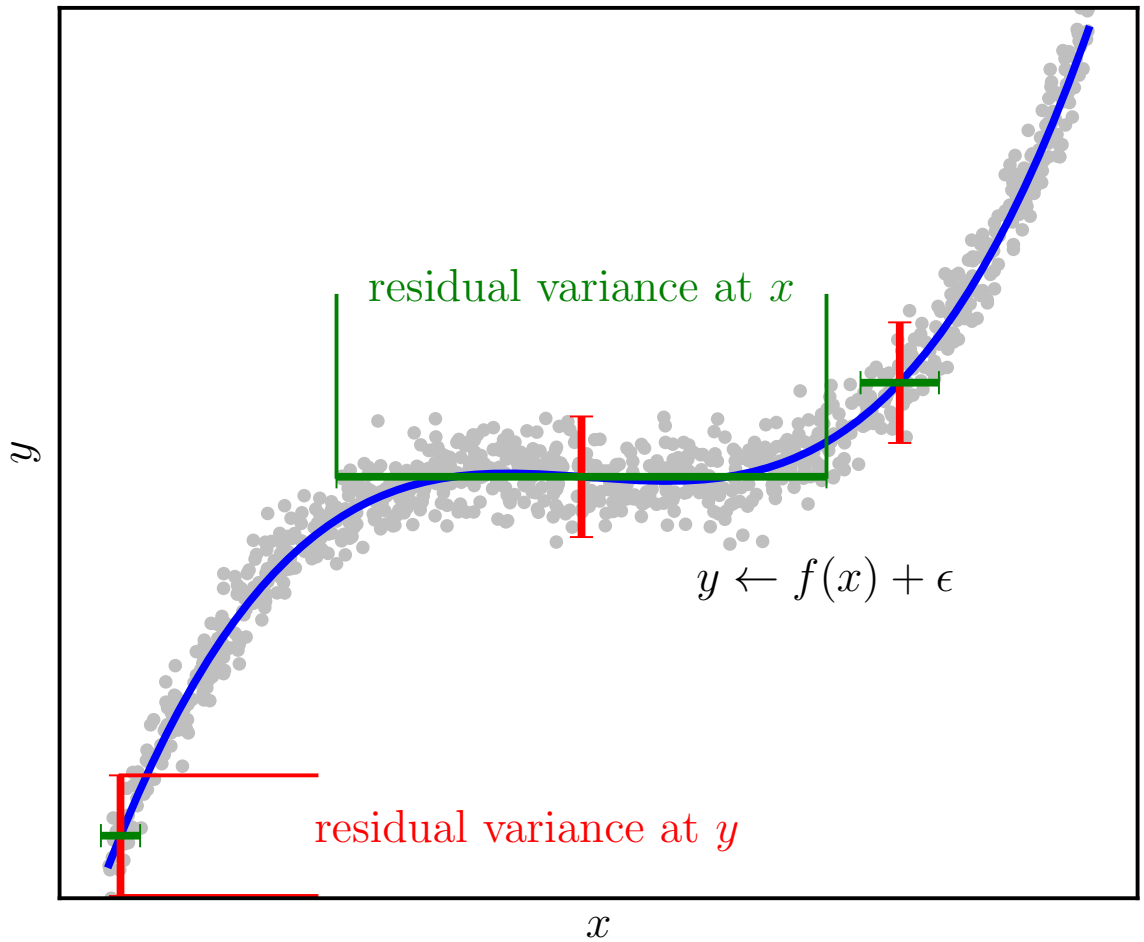
cause distribution versus **slope of the mechanism**



[2] Daniusis et al., *Inferring deterministic causal relationships*, 2010

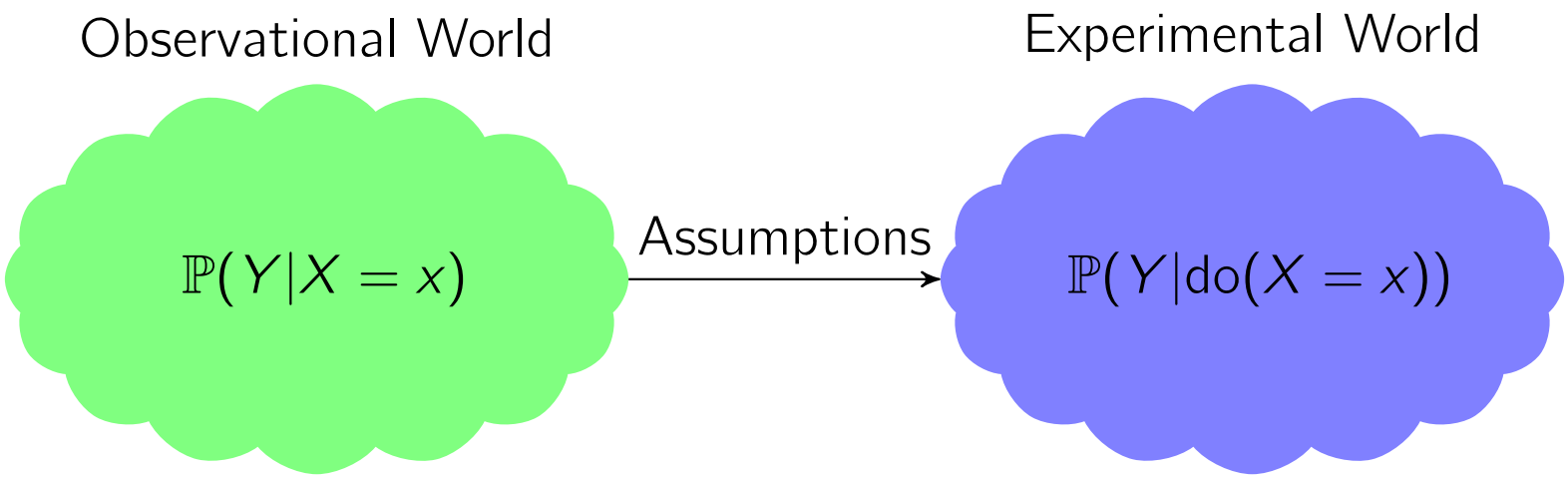
Previous Works, Additive Noise Model (ANM)[3]

cause distribution versus **residual distribution**



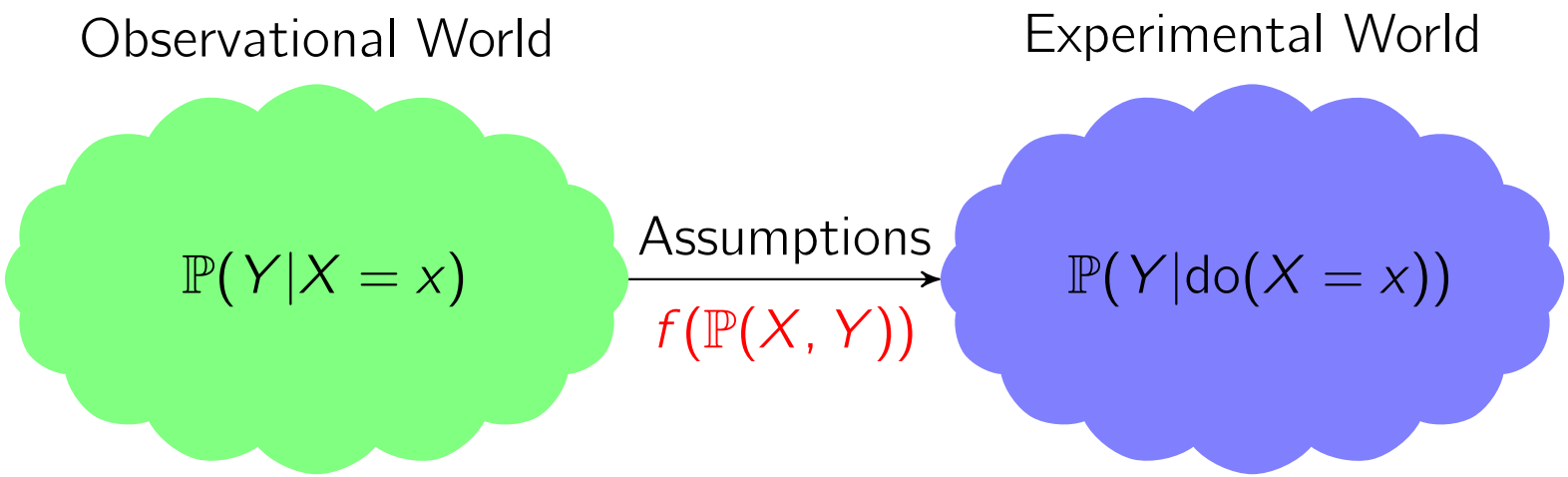
[3]Hoyer et al., Nonlinear causal discovery with additive noise models, 2009

Big Picture



Can we learn **the bridge, i.e., a causal footprint,** between two worlds given some information from the experimental world?

Big Picture



Can we learn **the bridge, i.e., a causal footprint,** between two worlds given some information from the experimental world?

Guyon Kaggle Competition Approach[4]

Cause-effect inference $\equiv$ Classification problem:

1. A collection of labeled causal samples $\{(S_i, l_i)\}_{i=1}^{n}$:
   - $S_i = \{(x_{ij}, y_{ij})\}_{j=1}^{n_i}$,
   - $l_i = +1$ when $X_i \to Y_i$, and $l_i = -1$ when $Y_i \to X_i$.
2. For each S_i, extract a feature vector $m(S_i) \in \mathbb{R}^m$:
3. Train a classifier $f : \mathbb{R}^m \to \{-1, +1\}$ on $\{(m(S_i), l_i)\}_{i=1}^{n}$.
4. Classify new test samples S_* as $f(m(S_*))$.

[4] https://www.kaggle.com/c/cause-effect-pairs/

Guyon Kaggle Competition Approach[4]

Cause-effect inference ≡ Classification problem:

1. A collection of labeled causal samples $\{(S_i, l_i)\}_{i=1}^n$:
   - $S_i = \{(x_{ij}, y_{ij})\}_{j=1}^{n_i}$,
   - $l_i = +1$ when $X_i \to Y_i$, and $l_i = -1$ when $Y_i \to X_i$.
2. For each S_i, extract a feature vector $m(S_i) \in \mathbb{R}^m$:
3. Train a classifier $f : \mathbb{R}^m \to \{-1, +1\}$ on $\{(m(S_i), l_i)\}_{i=1}^n$.
4. Classify new test samples S_* as $f(m(S_*))$.

Limitations:

1. The pre-defined sets of features may not be sufficient.
2. Handcrafting features is computationally expensive.
3. Handcrafting features makes theoretical analysis difficult.

[4] https://www.kaggle.com/c/cause-effect-pairs/

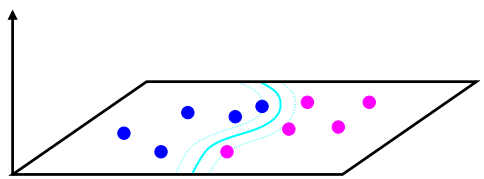
Our Approach: Learning from Distributions

- takes as inputs $\mathbb{P}(X, Y)$ from which $S = \{(x_j, y_j)\}_{j=1}^{n}$ are drawn.
- automates feature extraction by representing each joint distribution $\mathbb{P}(X, Y)$ by the **kernel mean embedding**.

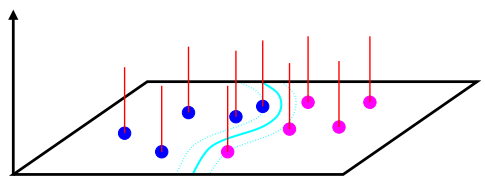
$$\mu_{\mathbb{P}(X,Y)} := \int k(z, \cdot)\, \mathrm{d}\mathbb{P}(z), \qquad z := (x, y).$$

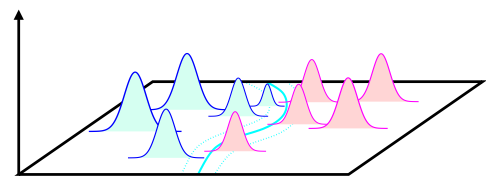
- It can be estimate efficiently from the sample:

$$\hat{\mu}_{\mathbb{P}(X,Y)} := \frac{1}{n} \sum_{i=1}^{n} k(z_i, \cdot), \qquad z_i := (x_i, y_i).$$



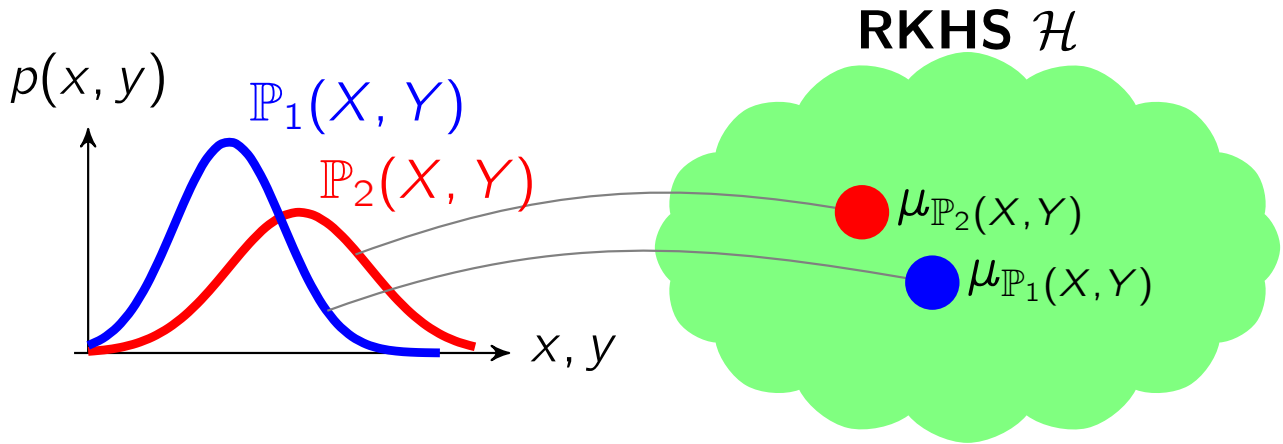
$x \longmapsto k(x, \cdot)$



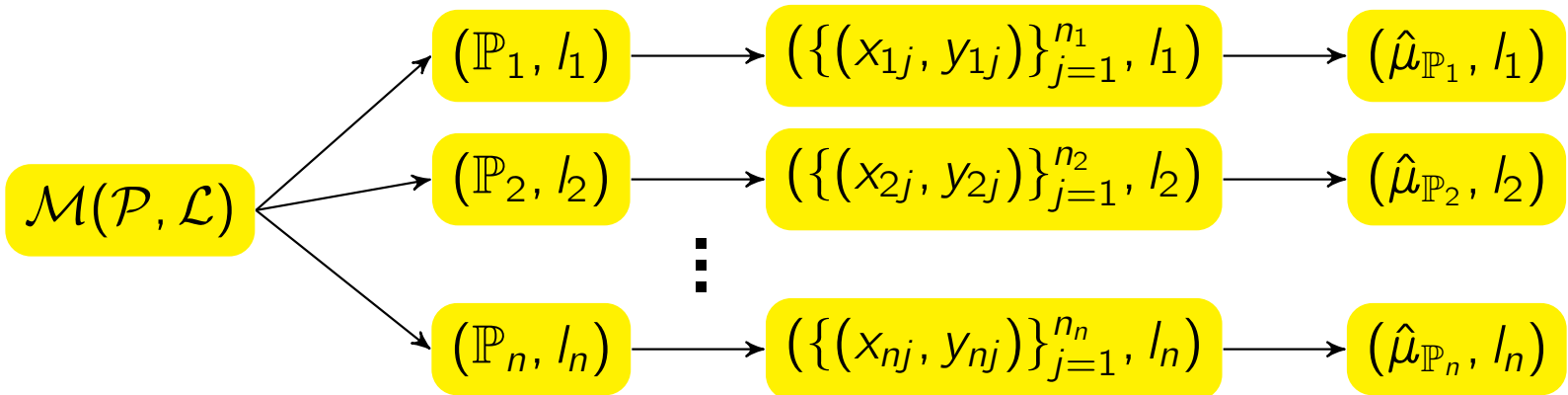


$\mathbb{P} \longmapsto \int k(x, \cdot)\, \mathrm{d}\mathbb{P}(x)$

Our Approach: Learning from Distributions



Cause-effect inference $\equiv$ Classifying distributions



Consistency Results: The Rate

With probability not less than $1-\delta$:

$$R_\varphi(\tilde{f}_n) - R_\varphi(f^*) \leq 4L_\varphi R_n(\mathcal{F}_k) + 2B\sqrt{\frac{\log(2/\delta)}{2n}}$$
$$+ \frac{4L_\varphi L_{\mathcal{F}}}{n} \sum_{i=1}^{n} \left(\sqrt{\frac{\mathbb{E}_{z\sim P_i}[k(z,z)]}{n_i}} + \sqrt{\frac{\log \frac{2n}{\delta}}{2n_i}} \right),$$

Remark

- The algorithm is consistent as $n, n_i \to \infty$.
- $\log n/n_i = o(1)$, i.e., n_i can grow as slow as $\log n$.
- For more theoretical results, come talk to us at our poster.

Experimental Protocol

- Gaussian kernel $k = k_\gamma$, with γ chosen via median-heuristic.
- Feature map that embeds marginals and joint as

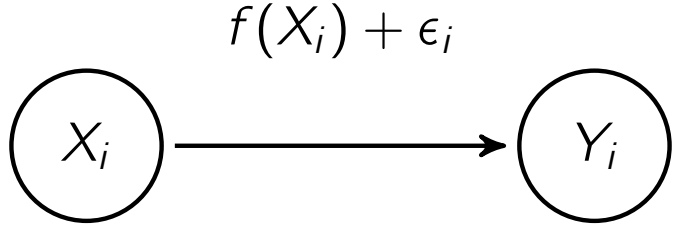
$$\nu(S) := (\hat{\mu}_{k,m}(P_{S_x}), \hat{\mu}_{k,m}(P_{S_y}), \hat{\mu}_{k,m}(P_{S_{xy}})) \in \mathbb{R}^{3m},$$

  where $\hat{\mu}_{k,m}$ approximates $\hat{\mu}_k$ with $m = 1000$ **random Fourier features**.[5]
- Random forest from `sklearn-0.16-git` as classifier.

For more experimental results, come talk to us at our poster.

[5] Rahimi, A. and Recht, B. Random features for large-scale kernel machines. *NIPS, 2007*

Experiments: Nonlinear+Gaussian (Synthetic)



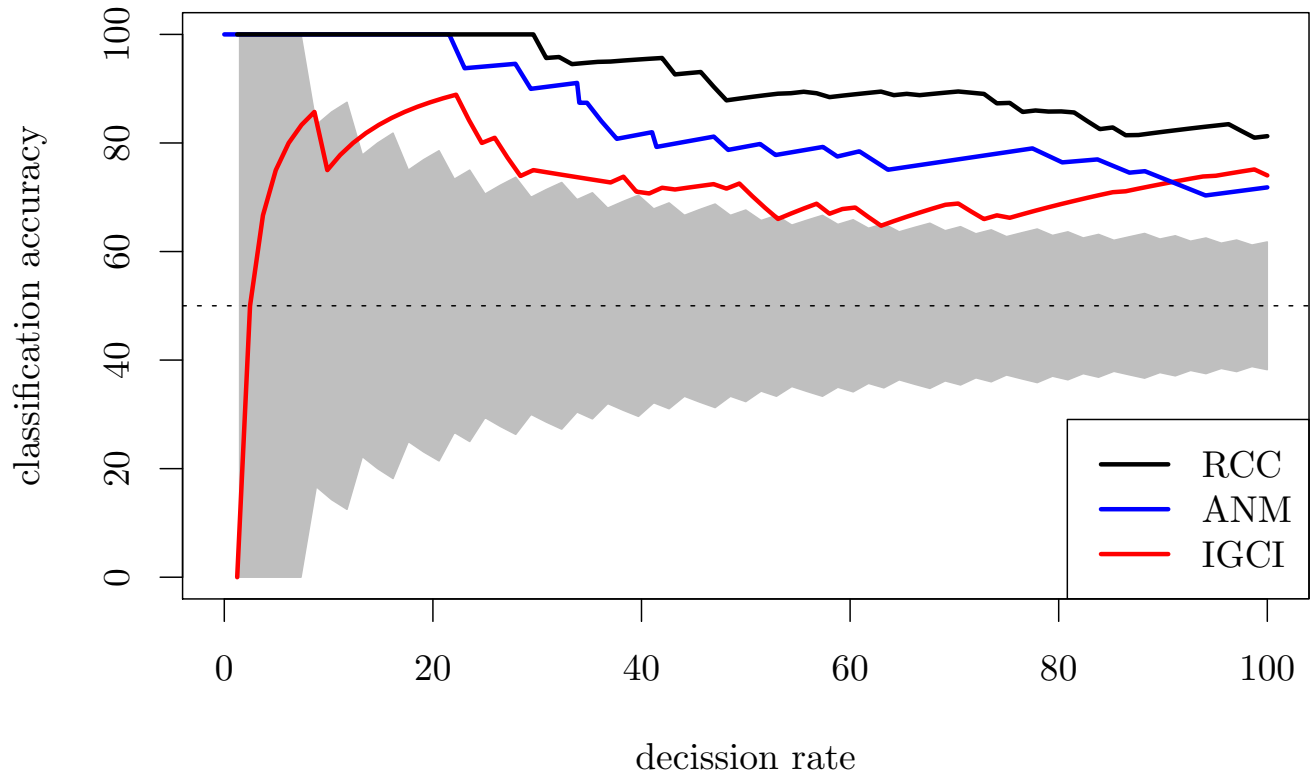
Setup:

- $X_i \sim \text{GMM}(c, \sigma_1, \sigma_2)$[6],
- $f(X)$ spline w/ d_f random knots
- $\epsilon_i \sim \mathcal{N}(0, \sigma_3)$, $\sigma \sim \mathcal{U}[0, 1]$
- $Y_i \leftarrow f(X_i) + \epsilon_i$

Accuracy: 97% ☺

[6] A Gaussian Mixture model with c components, mixing weights sampled from $\mathcal{U}[0, 1]$ and normalized to sum one, means as samples from $\mathcal{N}(0, \sigma_1)$, and standard deviations as positive samples from $\mathcal{N}(0, \sigma_2)$

Experiments: Tübingen Cause-Effect Pairs (Real)



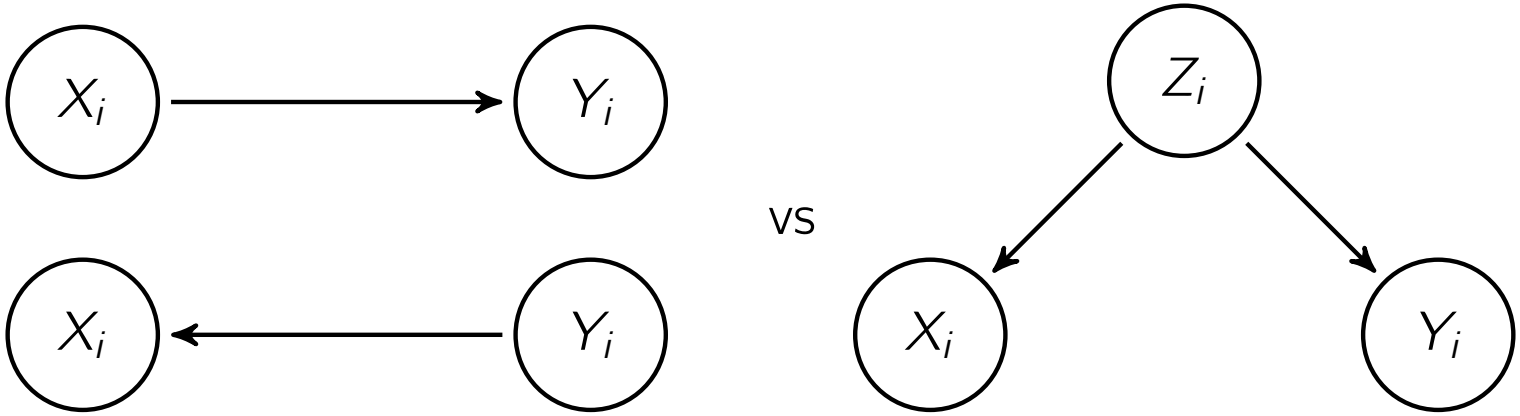
Training data: as in previous slide, aligned to test data

Test data: real-world Tübingen cause-effect pairs[7]

Accuracy: 81%, state-of-the-art ☺

[7] `http://webdav.tuebingen.mpg.de/cause-effect/`

Experiments: Detecting Confounders (Real)



Data: ChaLearn's challenge set[8], 20,000 causal samples

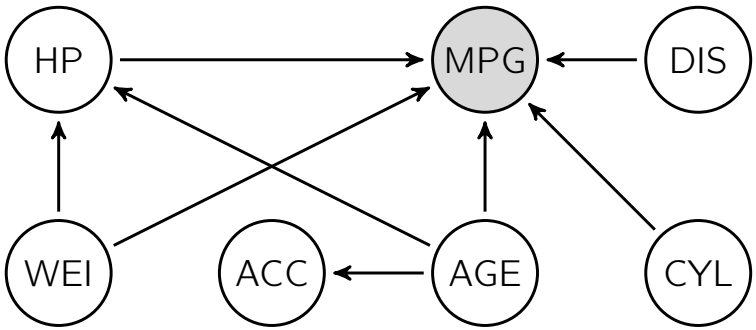
Accuracy: 80%

Related problem: Independence test, 88% accuracy ($\mathbb{P}(X, Y) = \mathbb{P}(X)\mathbb{P}(Y)$).

[8] https://www.codalab.org/competitions/1381

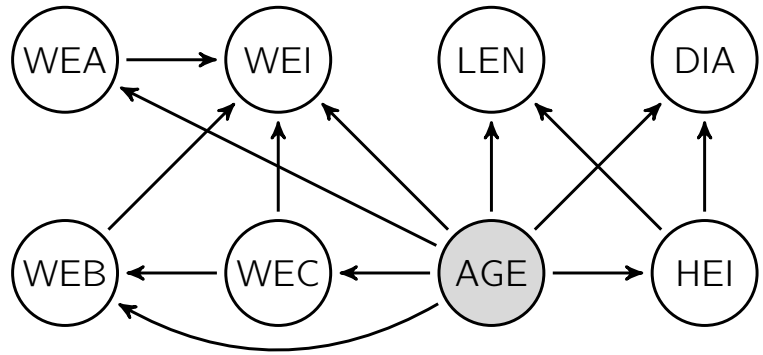
Experiments: Multivariate Variables (DAG)

AutoMPG (causal)



- HP: Horsepower
- WEI: Weight
- ACC: Acceleration
- AGE: Release year
- CYL: Cylinders
- DIS: Engine displacement
- **MPG: Miles per gallon**

Abalone (anti-causal)



- WEA: Meat weight
- WEB: Gut weight
- WEC: Shell weight
- WEI: Total weight
- LEN: Length
- DIA: Diameter
- HEI: Height
- **AGE: Age**

Take Home Message

- The causal footprint that links observational world to experimental world can be learnt.
- We formulate the causal inference between X and Y as a classification problem on $\mathbb{P}(X, Y)$.
- Kernel mean embedding incorporates information about $\mathbb{P}(X, Y)$ in an efficient and meaningful way.
- The idea can be extended to DAG over random variables.