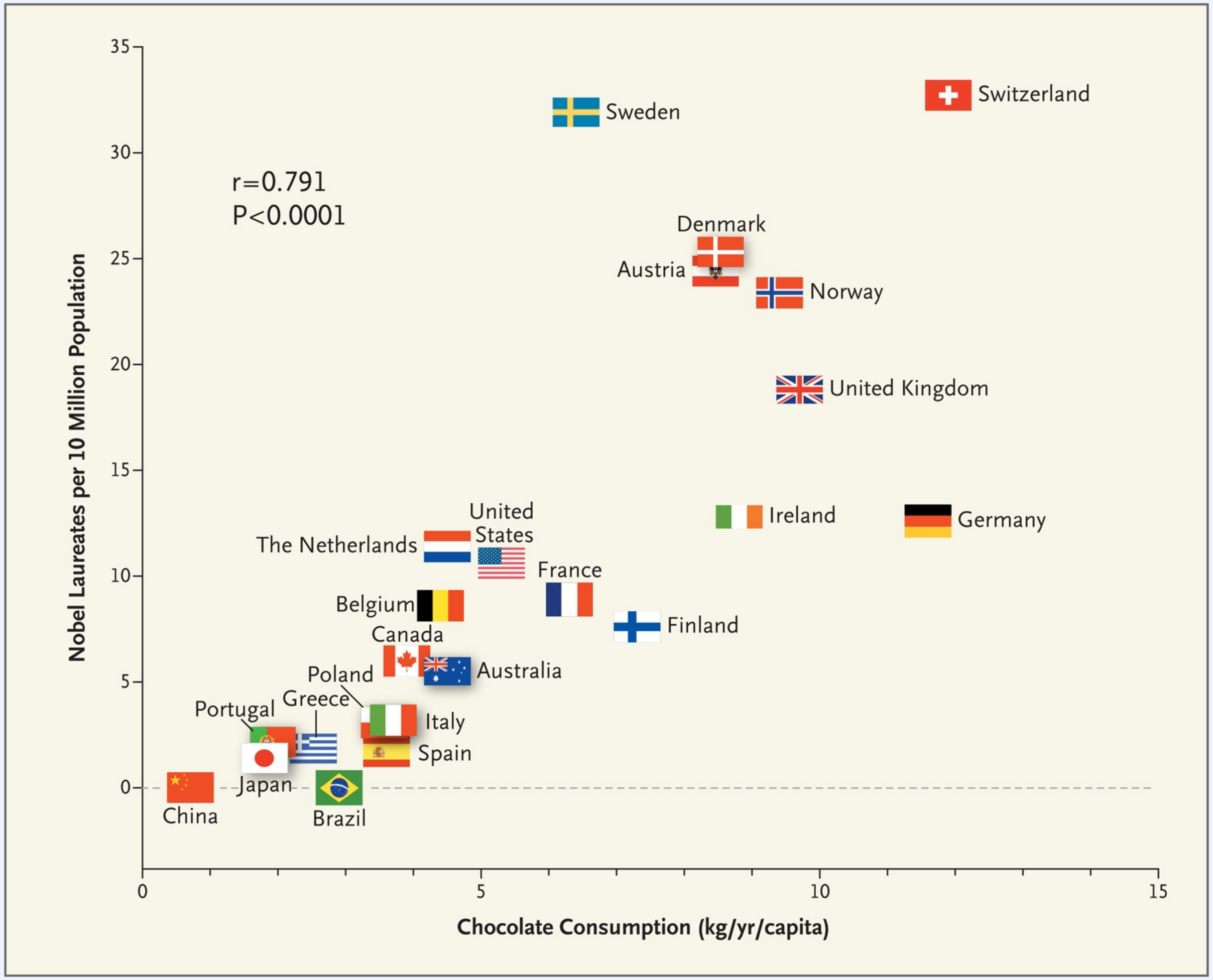
Telling Causal from Confounded

David Kaltenpoth
Jilles Vreeken

10 September 2019

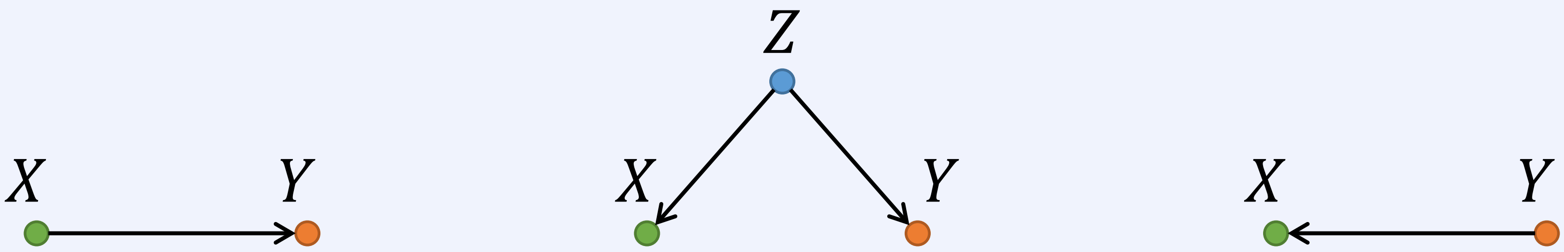
UNIVERSITÄT DES SAARLANDES
max planck institut informatik
CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY

Does Chocolate Consumption cause Nobel Prizes?



Reichenbach

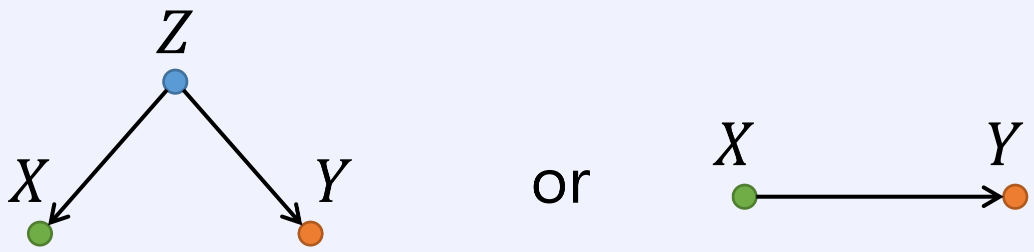
If X and Y are statistically dependent then either



How can we distinguish these cases?

Conditional Independence Tests

If we have measured everything relevant
then testing $X \perp\!\!\!\perp Y \mid Z$ for all possible Z
lets us decide whether



Problem: It's impossible to measure everything relevant

Why not just find a confounder?

We would like to be able to infer a $\hat{Z}$ such that

$$X \perp\!\!\!\perp Y \mid \hat{Z}$$

if and only if X and Y are actually confounded

Problem: Finding such a $\hat{Z}$ is **too easy**

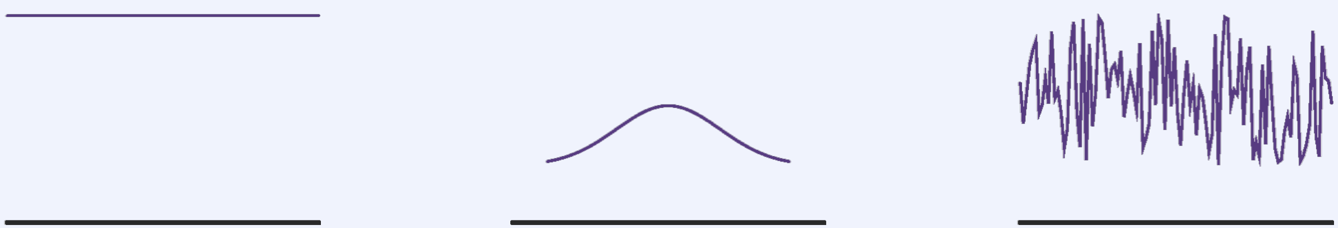
$\hat{Z} = X$ always works

Kolmogorov Complexity

$K(P)$ is the length of the shortest program computing P

$$K(P) = \min_{p} \left\{ |p| : p \in \{0,1\}^*, |\mathcal{U}(p, x, q) - P(x)| < \frac{1}{q} \right\}$$

This shortest program p^* is the best compression of P



From the Markov…

An admissible causal network for $X_1, \ldots, X_m$ is G satisfying

$$P(X_1, \ldots, X_m) = \prod_{i=1}^{m} P(X_i \mid PA_i)$$

Problem: How do we find a simple factorization?

…to the Algorithmic Markov Condition

The simplest causal network for $X_1, \dots, X_m$ is G^* satisfying

$$K(P(X_1, \dots, X_m)) = \sum_{i=1}^{m} K(P(X_i \mid PA_i^*))$$

Postulate: G^* corresponds to the true generating process

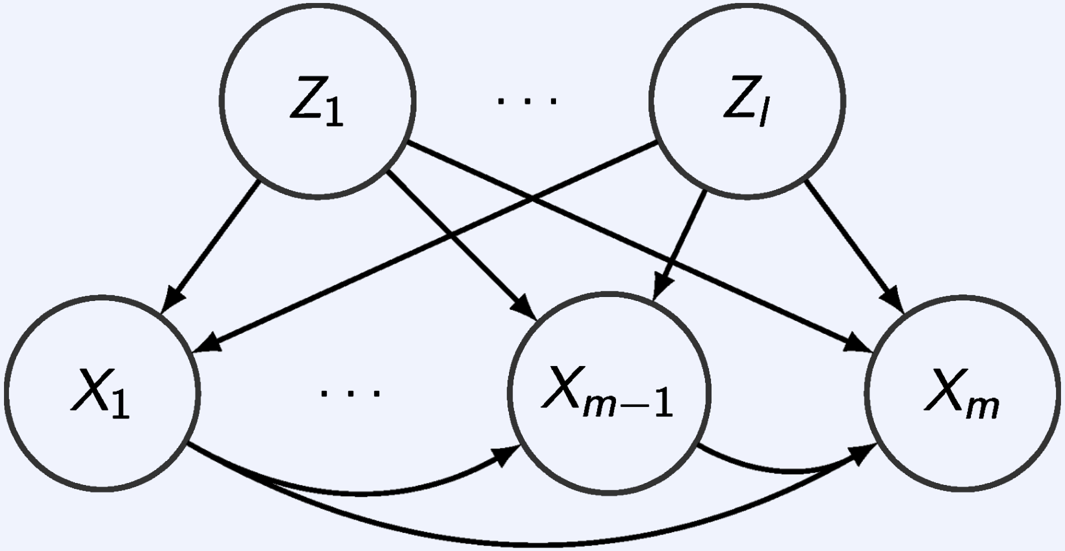
(Janzing & Schölkopf, 2010)

AMC with Confounding

We can also include latent variables

$$K\big(P(\boldsymbol{X},\boldsymbol{Z})\big) = \sum_{i=1}^{m} K\big(P(X_i \mid PA_i')\big) + \sum_{j=1}^{l} K\Big(P\big(Z_j\big)\Big)$$

We don't know $P(\cdot)$



$$P(\boldsymbol{X},\boldsymbol{Z}) = P(\boldsymbol{Z})\prod_{i=1}^{m} P(X_i \mid \boldsymbol{Z})$$

In particular, we will use probabilistic PCA

Kolmogorov is not computable

For data X, the **Minimum Description Length** principle identifies the best model $M \in \mathcal{M}$ by minimizing

$$L(X, M) = L(M) + L(X \mid M)$$

which provides a **computable** and **statistically sound approximation** to K

(Grünwald 2007)

Decisions, decisions

If

$$L(\boldsymbol{X}, Y, \mid \mathcal{M}_{co}) < L(\boldsymbol{X}, Y \mid \mathcal{M}_{ca})$$

then we consider $\boldsymbol{X}, Y$ to be confounded

Decisions, decisions

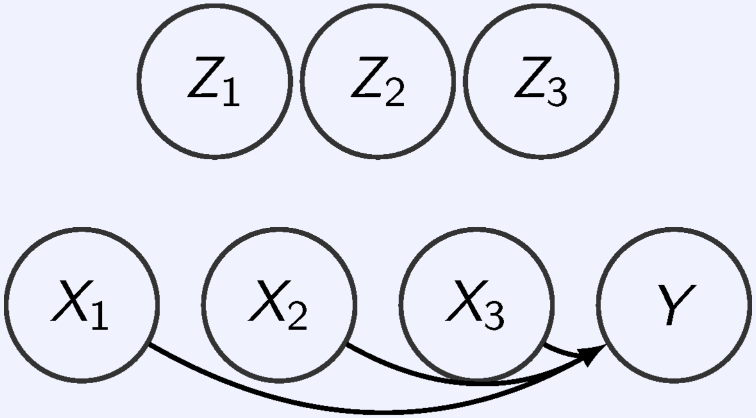
If

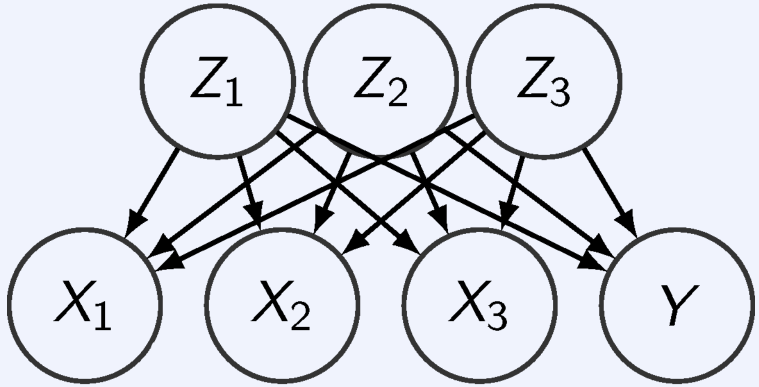
$$L(\boldsymbol{X}, Y, \mid \mathcal{M}_{co}) > L(\boldsymbol{X}, Y \mid \mathcal{M}_{ca})$$

then we consider $\boldsymbol{X}, Y$ to be causal

The difference can be interpreted as confidence

Confounding in Synthetic Data



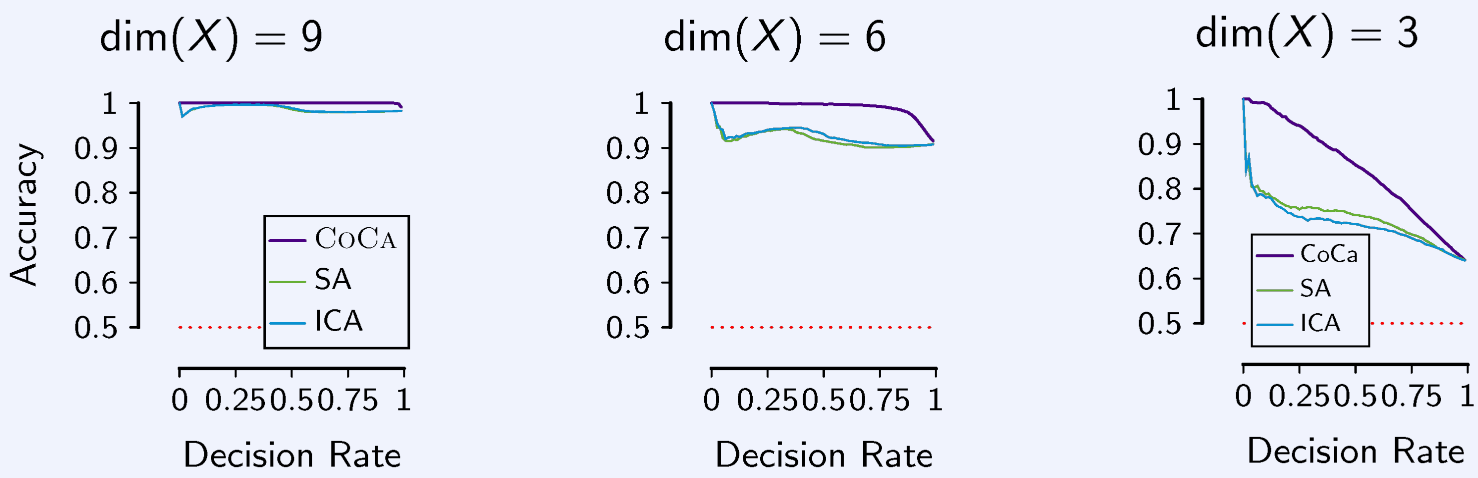


Synthetic Data: Results

There are only two other works directly related to ours

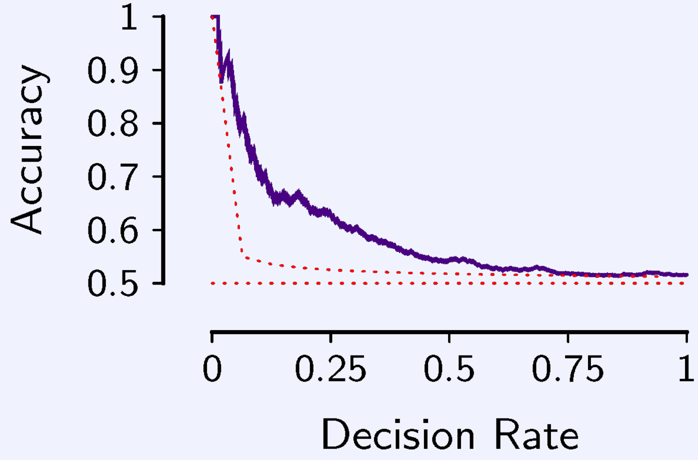
SA: Confounding strength in linear models using spectral analysis

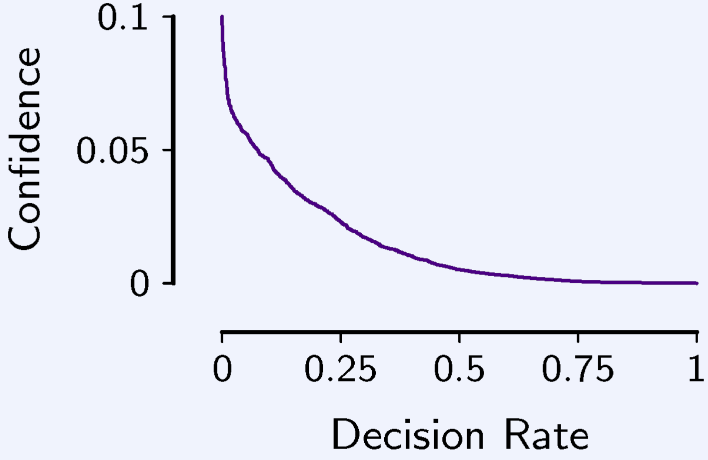
ICA: Confounding strength using independent component analysis



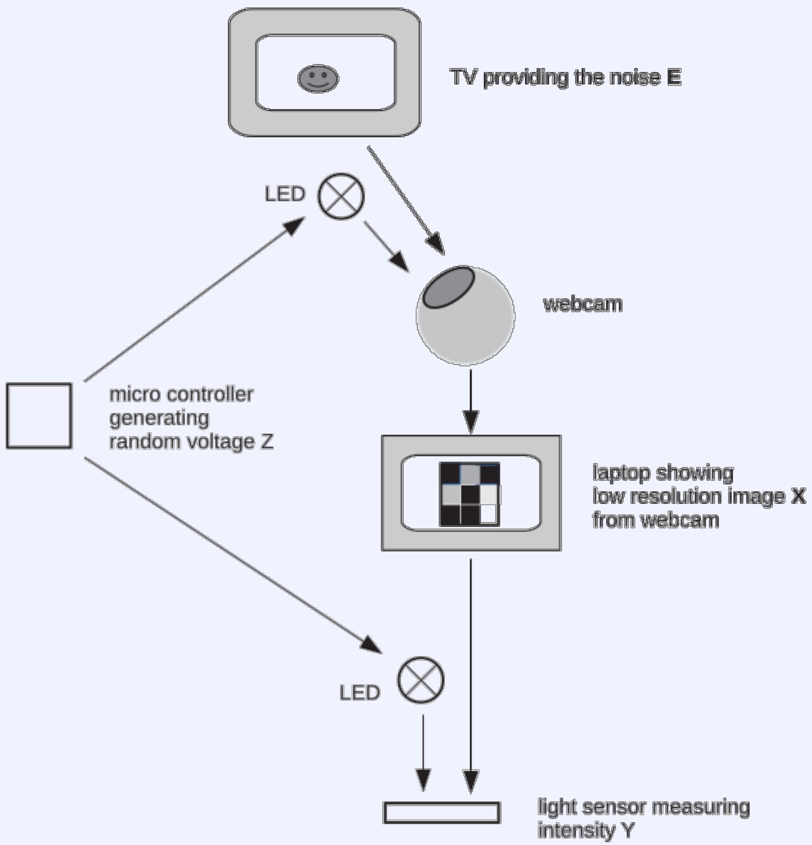
Confounding in Genetic Networks

More realistically, we consider gene regulation data



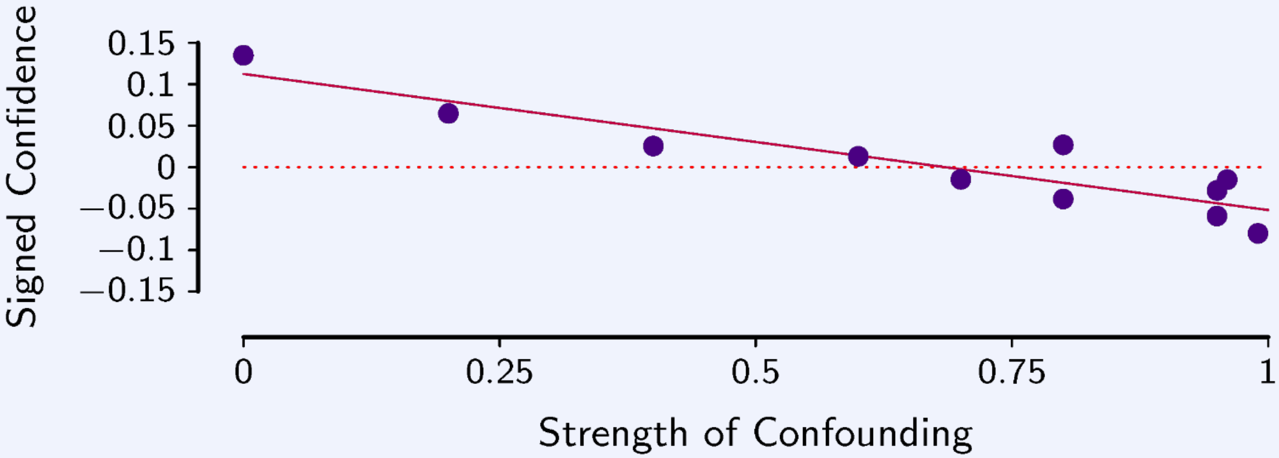


Optical Data

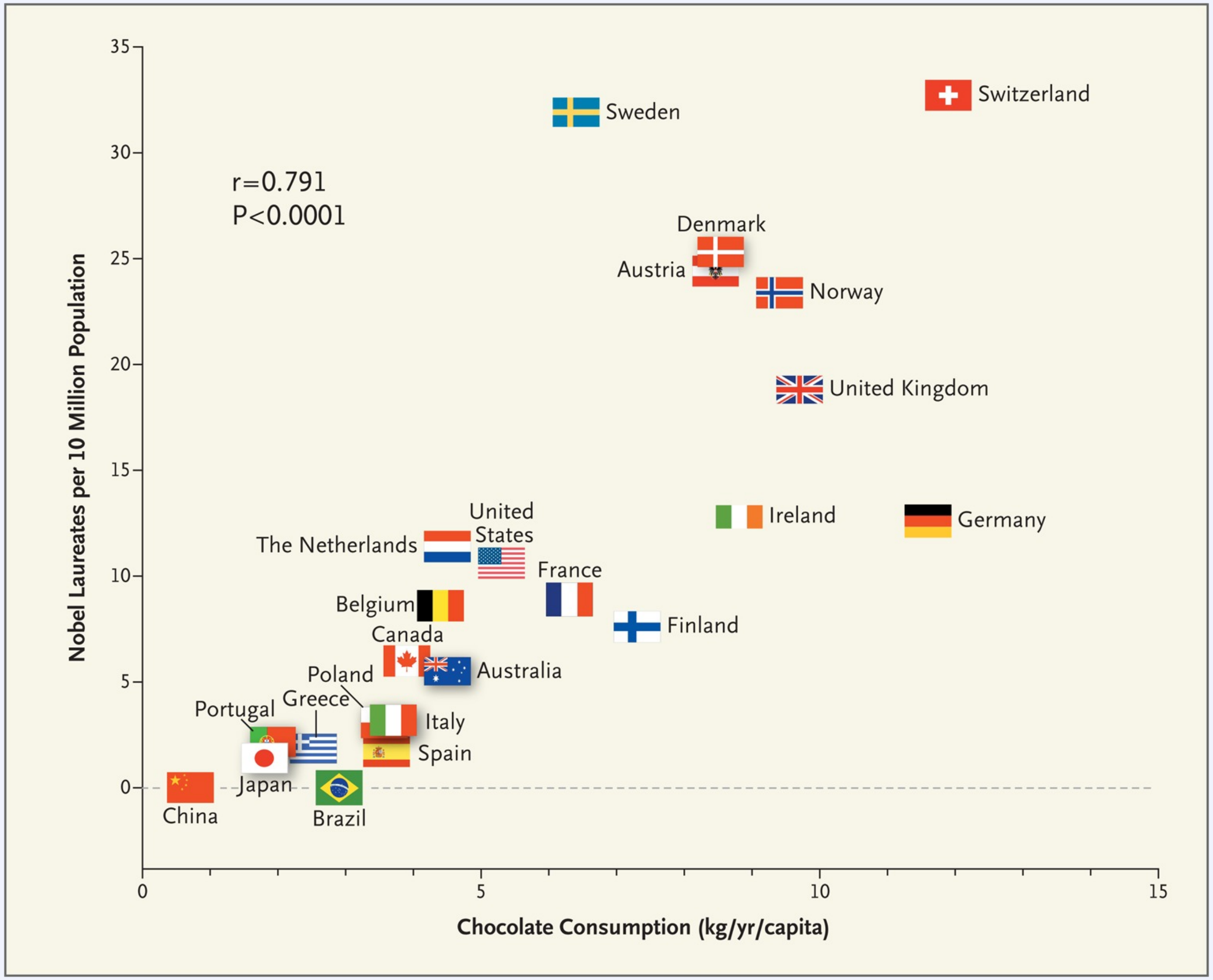


(Janzing & Schölkopf, 2017)

Optical Data



Wait! What about...



Conclusions

We looked into distinguishing causal from confounded

In particular, we

- generalized the AMC to include **latent variables**
- used a linear factor model and MDL to **instantiate** it
- showed that we obtain good results on synthetic and real data

In the future, we will

- work on a **significance test** for our score
- look into using **more complex factor models**
- apply our method to **real-world data**

Thank you!

We looked into distinguishing causal from confounded

In particular, we

- generalized the AMC to include **latent variables**
- used a linear factor model and MDL to **instantiate** it
- showed that we obtain good results on synthetic and real data

In the future, we will

- work on a **significance test** for our score
- look into using **more complex factor models**
- apply our method to **real-world data**