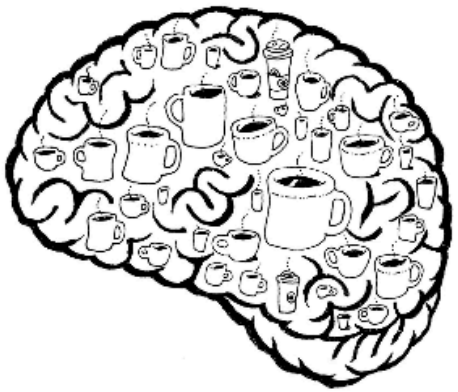
Causality

Evgeny Burnaev
Rodrigo Rivera

Skoltech

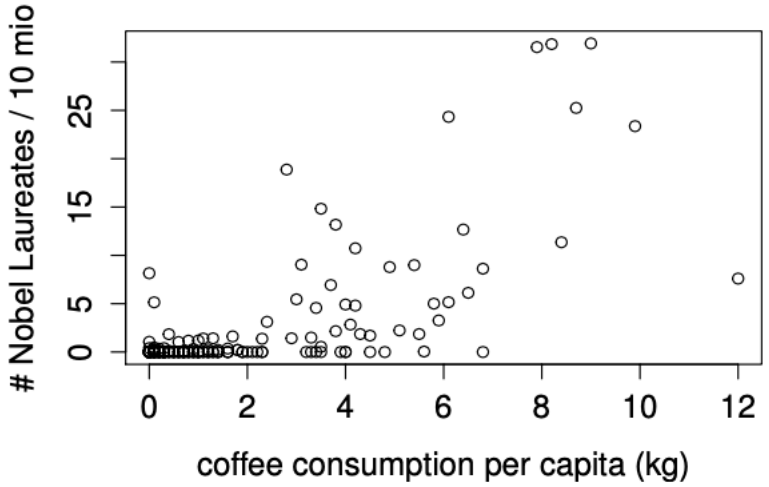
Example: chocolate

No (not enough) data for chocolate



... but we have data for coffee!

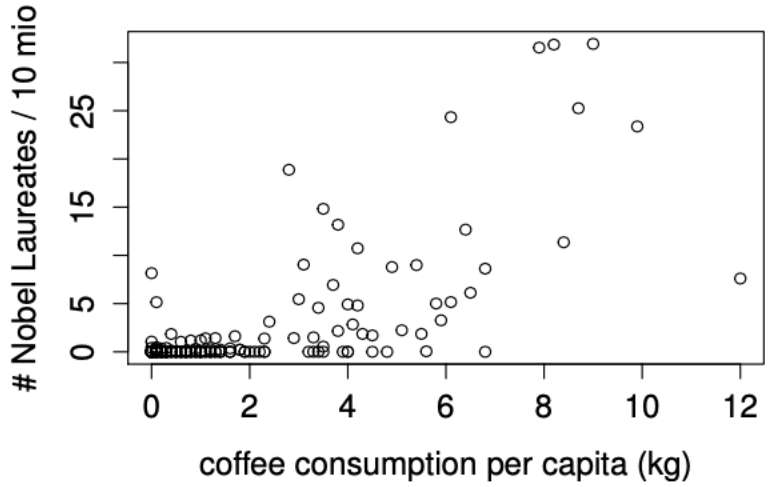
Example: chocolate



Correlation: 0.698

p-value: $< 2.2 \cdot 10^{-16}$

Example: chocolate

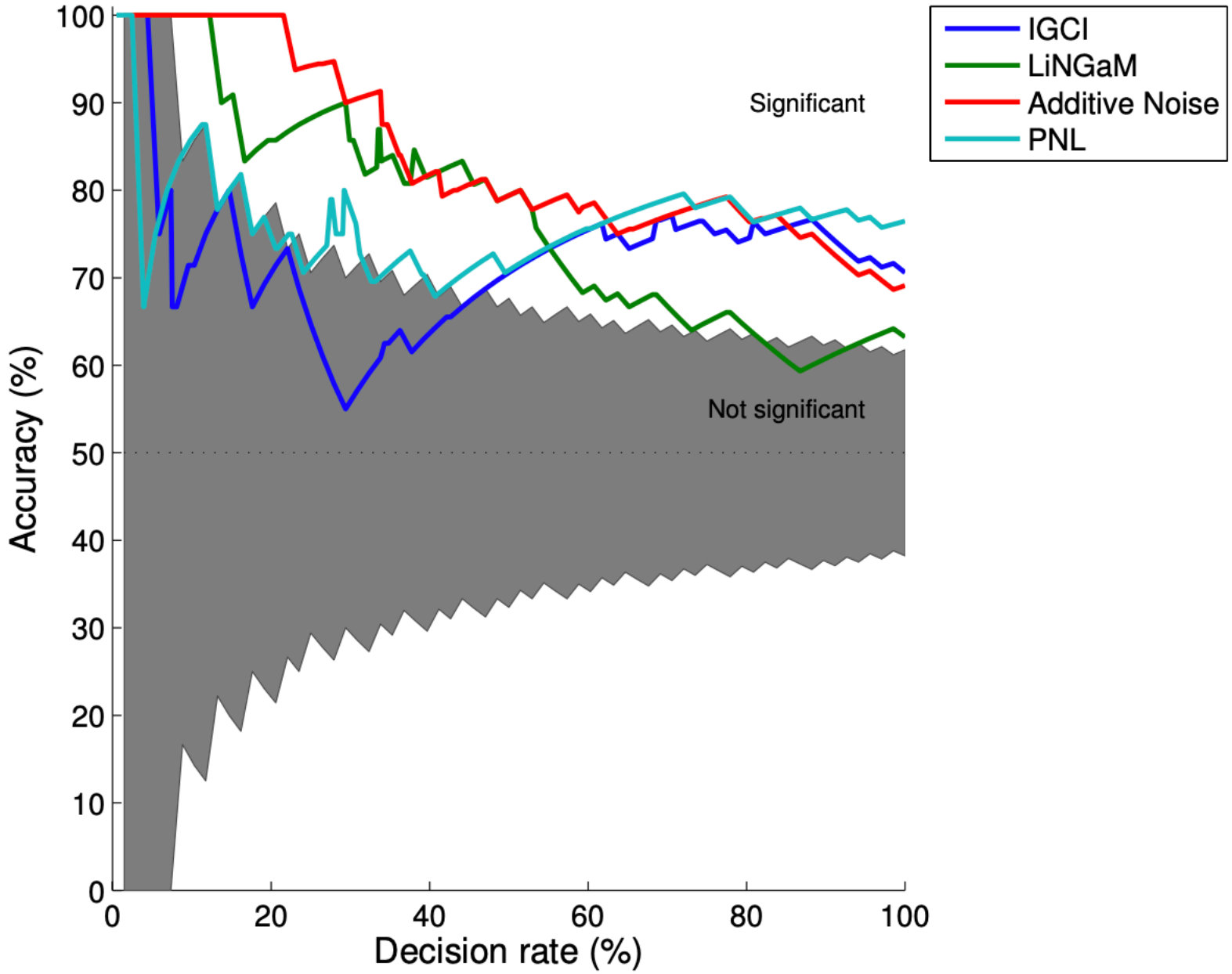


Correlation: 0.698
p-value: $< 2.2 \cdot 10^{-16}$

Coffee $\rightarrow$ Nobel Prize: Dependent residuals (p-value of $5.1 \cdot 10^{-78}$).
Nobel Prize $\rightarrow$ Coffee: Dependent residuals (p-value of $3.1 \cdot 10^{-12}$).

$\Rightarrow$ Model class too small? Causally insufficient?

Real Data: cause-effect pairs



Idea 2: restricted structural causal models

Slightly surprising:

identifiability for two variables $\rightsquigarrow$ identifiability for d variables

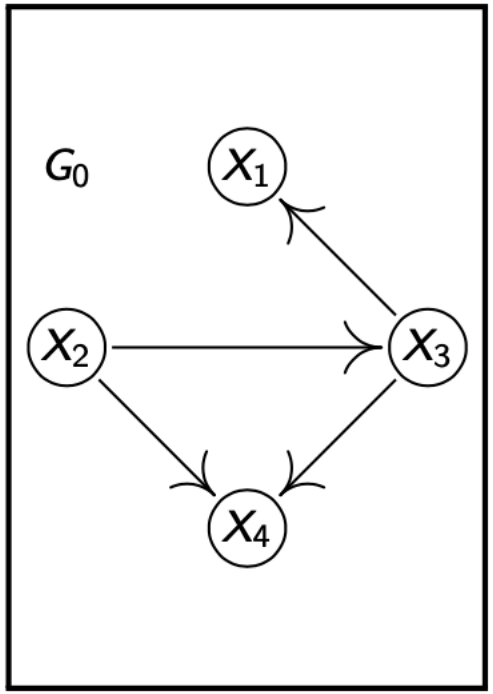
Peters et al.: *Identifiability of Causal Graphs using Functional Models*, UAI 2011

Idea 2: restricted structural causal models

Assume $P(X_1, \ldots, X_4)$ has been entailed by

$$X_1 = f_1(X_3, N_1)$$
$$X_2 = N_2$$
$$X_3 = f_3(X_2, N_3)$$
$$X_4 = f_4(X_2, X_3, N_4)$$

- N_i jointly independent
- G_0 has no cycles



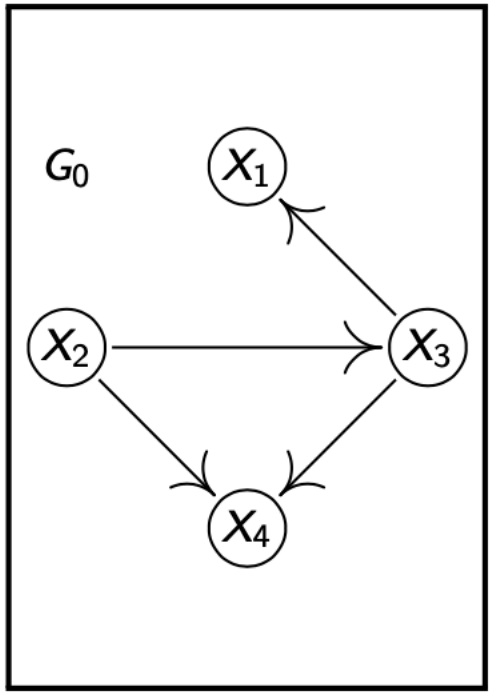
Structural causal model.
Can the DAG be recovered from $P(X_1, \ldots, X_4)$?

Idea 2: restricted structural causal models

Assume $P(X_1, \ldots, X_4)$ has been entailed by

$$X_1 = f_1(X_3, N_1)$$
$$X_2 = N_2$$
$$X_3 = f_3(X_2, N_3)$$
$$X_4 = f_4(X_2, X_3, N_4)$$

- N_i jointly independent
- G_0 has no cycles



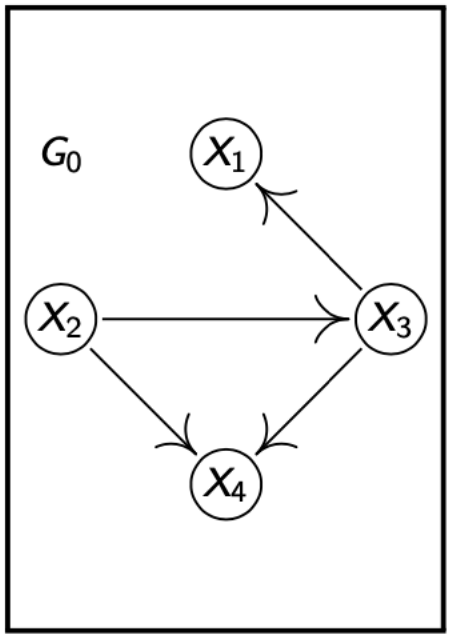
Structural causal model.
Can the DAG be recovered from $P(X_1, \ldots, X_4)$? **No.**

Idea 2: restricted structural causal models

Assume $P(X_1, \ldots, X_4)$ has been entailed by

$$X_1 = f_1(X_3) + N_1$$
$$X_2 = N_2$$
$$X_3 = f_3(X_2) + N_3$$
$$X_4 = f_4(X_2, X_3) + N_4$$

- $N_i \sim \mathcal{N}(0, \sigma_i^2)$ jointly independent
- G_0 has no cycles



Additive noise model with Gaussian noise.
Can the DAG be recovered from $P(X_1, \ldots, X_4)$? **Yes iff f_i nonlinear.**

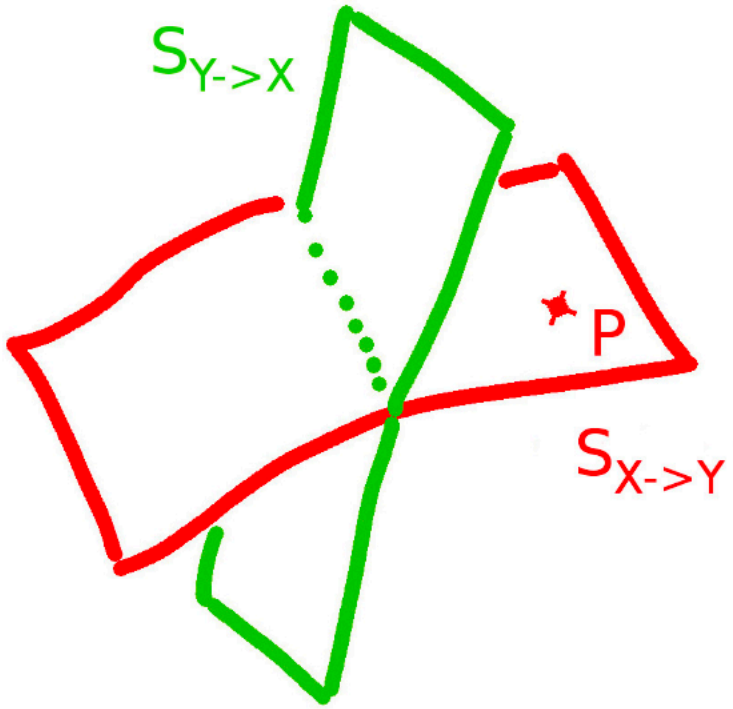
JP, J. Mooij, D. Janzing and B. Schölkopf: *Causal Discovery with Continuous Additive Noise Models, JMLR 2014*
P. Bühlmann, JP, J. Ernest: *CAM: Causal add. models, high-dim. order search and penalized regr.*, Annals of Statistics 2014

Let $P(X_1, \ldots, X_d)$ be entailed by an ...

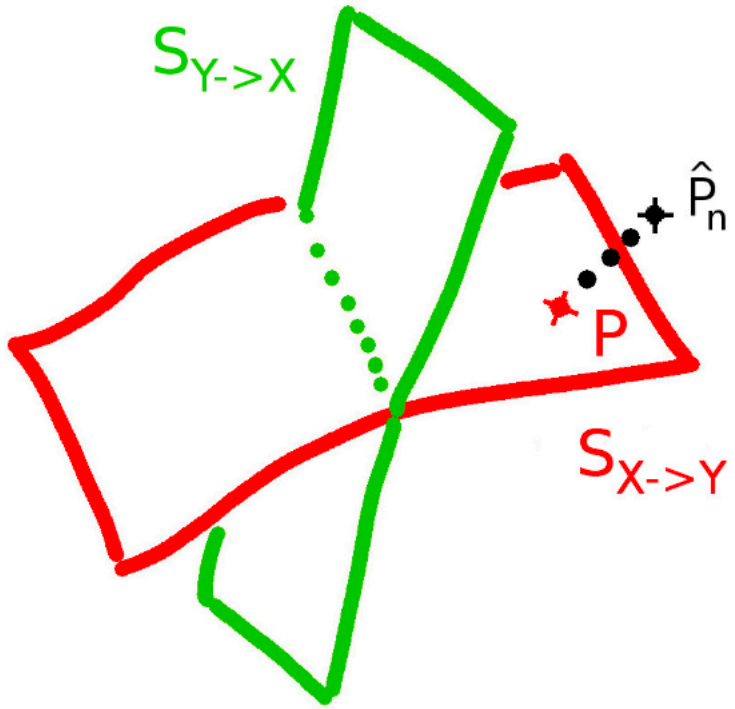
		conditions	identif.
structural causal model:	$X_i = f_i(X_{\mathbf{PA}_i}, N_i)$	-	✗
additive noise model:	$X_i = f_i(X_{\mathbf{PA}_i}) + N_i$	nonlin. fct.	✓
causal additive model:	$X_i = \sum_{k \in \mathbf{PA}_i} f_{ik}(X_k) + N_i$	nonlin. fct.	✓
linear Gaussian model:	$X_i = \sum_{k \in \mathbf{PA}_i} \beta_{ik} X_k + N_i$	linear fct.	✗

(results hold for Gaussian noise)

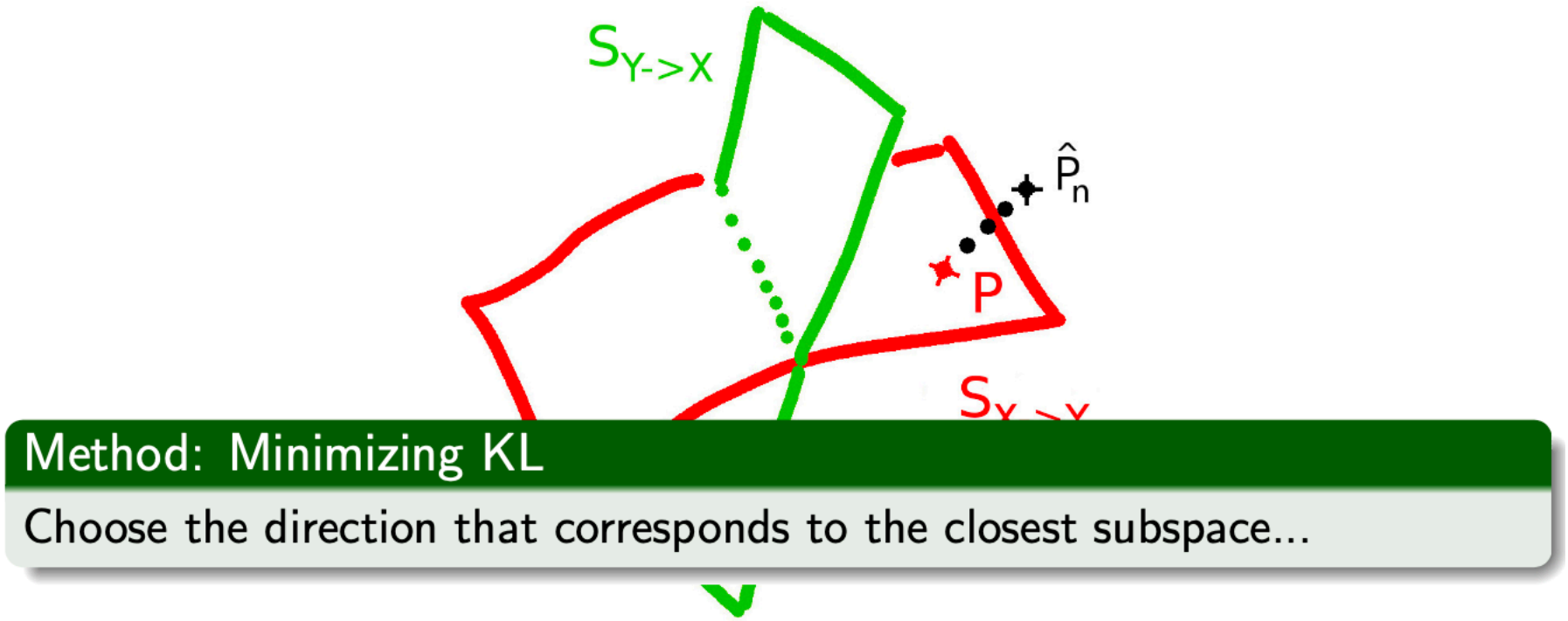
Idea 2: restricted structural causal models



Idea 2: restricted structural causal models



Idea 2: restricted structural causal models



Method: Minimizing KL

Choose the direction that corresponds to the closest subspace...

Idea 2: restricted structural causal models

Consider model classes

$$\mathcal{S}_G := \{Q : Q \text{ entailed by a causal additive model (CAM) with DAG } G\}$$

Define

$$\hat{G}_n := \underset{\text{DAG } G}{\operatorname{argmin}} \inf_{Q \in \mathcal{S}_G} \mathrm{KL}(\hat{P}_n \,||\, Q)$$

Idea 2: restricted structural causal models

Consider model classes

$$\mathcal{S}_G := \{Q : Q \text{ entailed by a causal additive model (CAM) with DAG } G\}$$

Define

$$\hat{G}_n := \underset{\text{DAG } G}{\operatorname{argmin}} \inf_{Q \in \mathcal{S}_G} \mathrm{KL}(\hat{P}_n \,||\, Q)$$

$$\overset{\text{max. likelihood}}{=} \underset{\text{DAG } G}{\operatorname{argmin}} \sum_{i=1}^{d} \log \hat{\mathrm{var}}(\mathrm{residuals}_{\mathbf{PA}_i^G \to X_i})$$

Idea 2: restricted structural causal models

Consider model classes

$$\mathcal{S}_G := \{Q : Q \text{ entailed by a causal additive model (CAM) with DAG } G\}$$

Define

$$\hat{G}_n := \underset{\text{DAG } G}{\operatorname{argmin}} \inf_{Q \in \mathcal{S}_G} \mathrm{KL}(\hat{P}_n \,||\, Q)$$

$$\overset{\text{max. likelihood}}{=} \underset{\text{DAG } G}{\operatorname{argmin}} \sum_{i=1}^{d} \log \hat{\mathrm{var}}(\mathrm{residuals}_{\mathbf{PA}_i^G \to X_i})$$

Wait, there is no penalization on the number of edges!

Idea 2: restricted structural causal models

Consider model classes

$$\mathcal{S}_G := \{Q : Q \text{ entailed by a causal additive model (CAM) with DAG } G\}$$

Define

$$\hat{G}_n := \underset{\text{DAG } G}{\operatorname{argmin}} \inf_{Q \in \mathcal{S}_G} \mathrm{KL}(\hat{P}_n \,||\, Q)$$

$$\overset{\text{max. likelihood}}{=} \underset{\text{DAG } G}{\operatorname{argmin}} \sum_{i=1}^{d} \log \hat{\mathrm{var}}(\mathrm{residuals}_{\mathbf{PA}_i^G \to X_i})$$

Wait, there is no penalization on the number of edges!
Wait again, there are too many DAGs!

Idea 2: restricted structural causal models

p	number of DAGs with p nodes
1	1
2	3
3	25
4	543
5	29281
6	3781503
7	1138779265
8	783702329343
9	1213442454842881
10	4175098976430598143
11	3160345939641891760742 5
12	521939651343829405020504063
13	1867660074443203518666481692 6721
14	1439428141044398334941790719839535103
15	2377252655534103549921802182862 76719253505
16	8375667077373332028769930304799641223522313830 3
17	6270792119692388989944645260249492190696355148267520 1
18	9942119532215951589522891459235452451655502687858830501478 3
19	3327719012271075917361775733112611258835830762584219025835467735 05
20	2344880451051088988152559855229099188899081192234291298795803236068491263
21	3469876828358875002875932843018108822231394454043860171902755911344658607767552 1
22	1075822921725761493652956179327624326573727662809185218104090000500559527511693495107583
23	6974332983728149264714154970024580487650427499051598589410910640154981198551095150137712207462 5

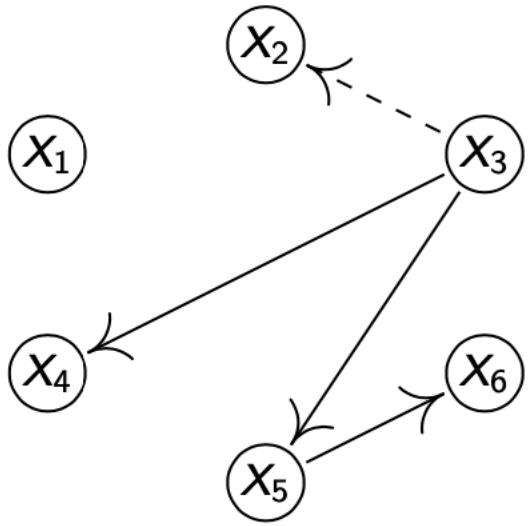
https://oeis.org/A003024/b003024.txt

Idea 2: restricted structural causal models

E.g. greedy search!

–	0.2	0.1	0.1	0.1	0.3
0.4	–	0.1	0.1	0.1	0.1
0.1	0.6	–	–	–	0.4
0.1	0.1	–	–	0.1	0.1
0.1	0.1	–	0.1	–	–
0.3	0.1	–	0.1	–	–

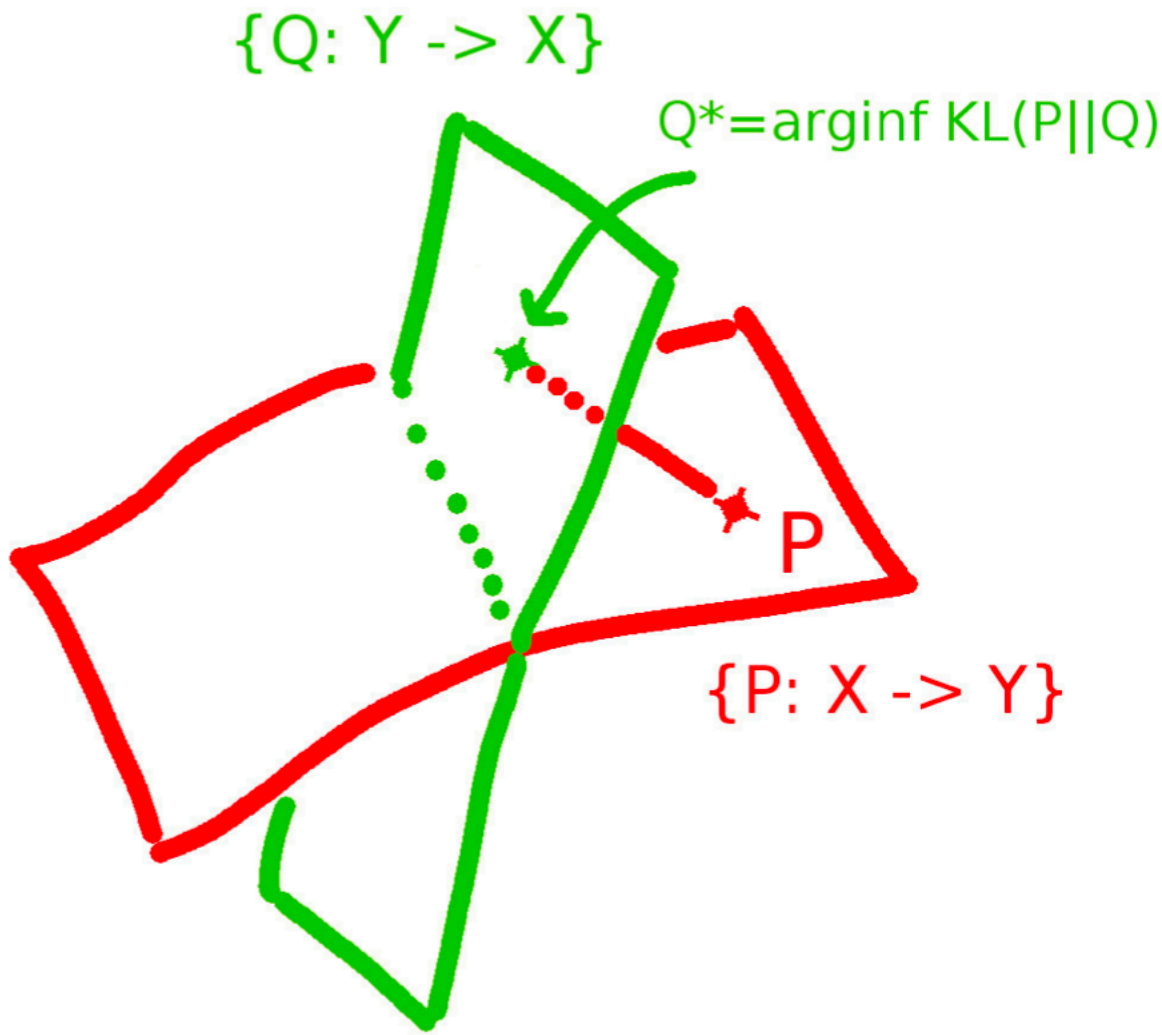
include best edge $\longrightarrow$ recompute column



Greedy Addition (e.g. Chickering 2002). Include the edge that leads to the largest increase of the log-likelihood.

Bühlmann, JP, Ernest: *CAM: Causal add. models, high-dim. order search and penalized regr.*, Annals of Statistics 2014

Can we characterize identifiability?



Can we characterize identifiability?

Proposition

Assume $P(X, Y)$ *is generated by*

$$Y = \beta X^2 + N_Y$$

with independent $X \sim \mathcal{N}(0, \sigma_X^2)$ *and* $N_Y \sim \mathcal{N}(0, \sigma_{N_Y}^2)$.

Can we characterize identifiability?

Proposition

Assume $P(X, Y)$ *is generated by*

$$Y = \beta X^2 + N_Y$$

with independent $X \sim \mathcal{N}(0, \sigma_X^2)$ *and* $N_Y \sim \mathcal{N}(0, \sigma_{N_Y}^2)$.

Then

$$\inf_{Q \in \{Q : Y \to X\}} \mathrm{KL}(P \,||\, Q) > 0 \qquad \text{if } \beta \neq 0\,.$$

Can we characterize identifiability?

Proposition

Assume $P(X, Y)$ *is generated by*

$$Y = \beta X^2 + N_Y$$

with independent $X \sim \mathcal{N}(0, \sigma_X^2)$ *and* $N_Y \sim \mathcal{N}(0, \sigma_{N_Y}^2)$.

Then

$$\inf_{Q \in \{Q: Y \to X\}} \mathrm{KL}(P \,||\, Q) = \frac{1}{2} \log \left(1 + 2\beta^2 \frac{\sigma_X^4}{\sigma_{N_Y}^2} \right)$$

Leonardo da Vinci: Mould of the Horses Head

Given an original drawing (left) and a copy. How good is the copy?

Leonardo da Vinci: Mould of the Horses Head

Given an original drawing (left) and a copy. How good is the copy?

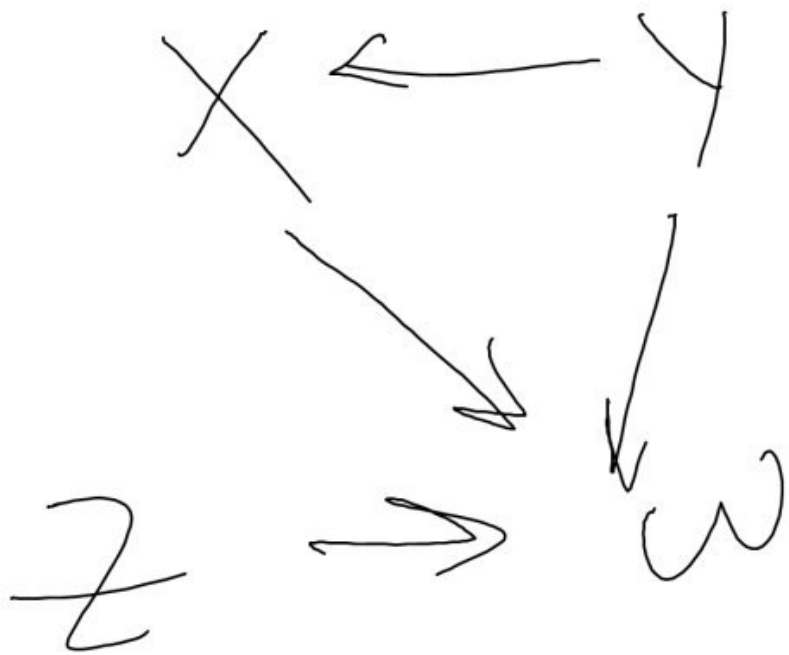
Given a true causal graph G and an estimate $\hat{G}$. How good is the estimate $\hat{G}$?

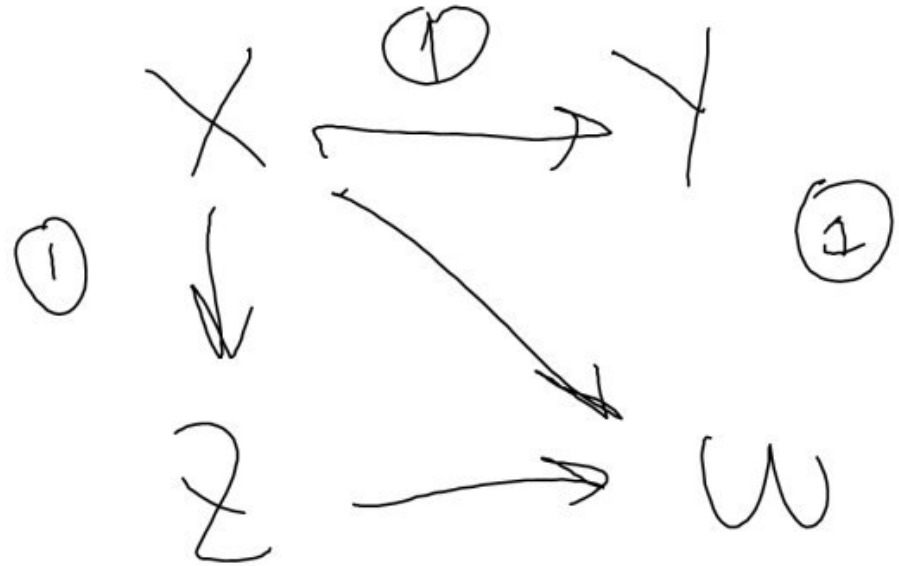
Leonardo da Vinci: Mould of the Horses Head

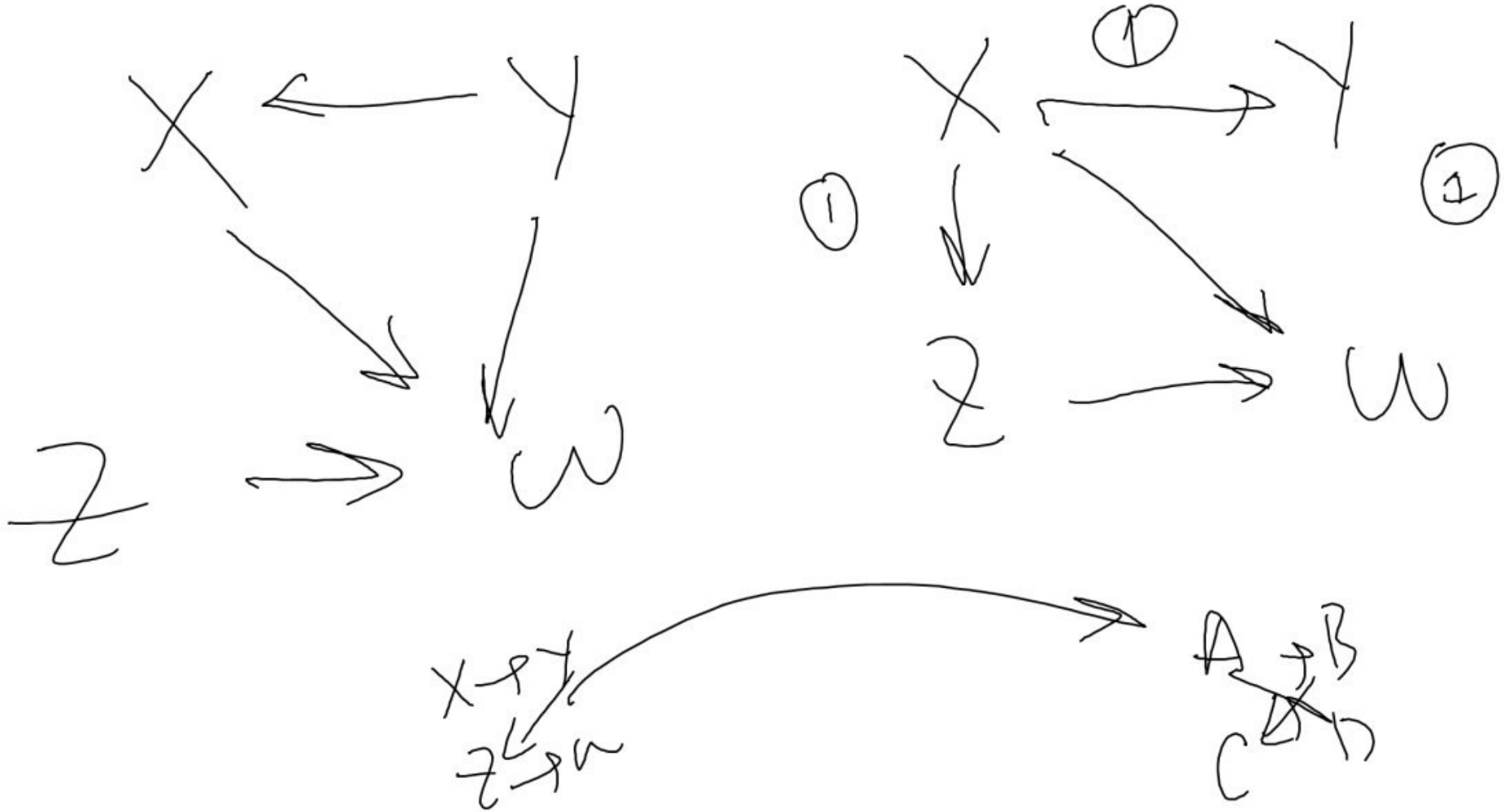
Given an original drawing (left) and a copy. How good is the copy?

Given a true causal graph G and an estimate $\hat{G}$. How good is the estimate $\hat{G}$?

What do we want do with it?


X
Y
Z
W


X
Y
Z
W
1
1
1


X
Y
Z
W
X
Y
Z
W
1
1
1
X × Y
Z → W
A → B
C

Definition: Structural Intervention Distance

For each pair (X, Y) check whether $\mathbf{PA}_X^{\hat{G}}$ is a valid adjustment set for (X, Y) in G for all distributions Markov w.r.t. G.

Definition: Structural Intervention Distance

For each pair (X, Y) check whether $\mathbf{PA}_X^{\hat{G}}$ is a valid adjustment set for (X, Y) in G for all distributions Markov w.r.t. G. - does not depend on P

$\mathrm{SID}(G, \hat{G})$ equals the number of pairs, for which this is not the case.

Definition: Structural Intervention Distance

For each pair (X, Y) check whether $\mathbf{PA}_X^{\hat{G}}$ is a valid adjustment set for (X, Y) in G for all distributions Markov w.r.t. G. - does not depend on P

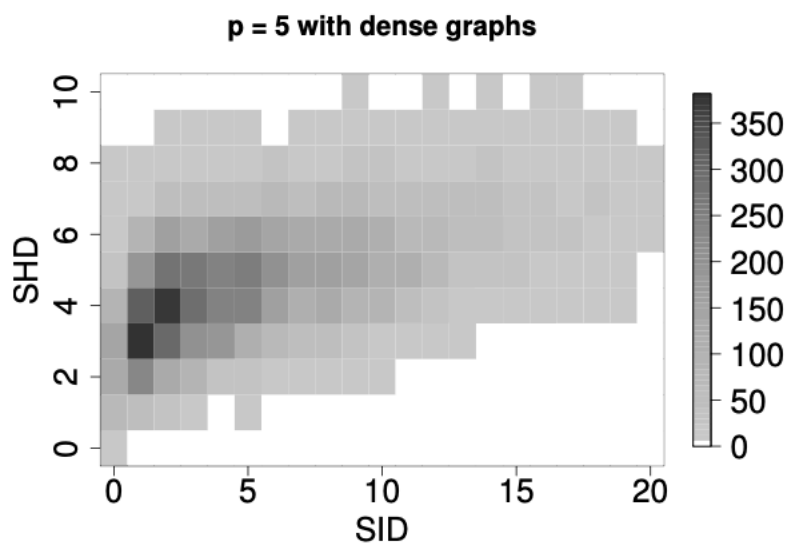
$\mathrm{SID}(G, \hat{G})$ equals the number of pairs, for which this is not the case.

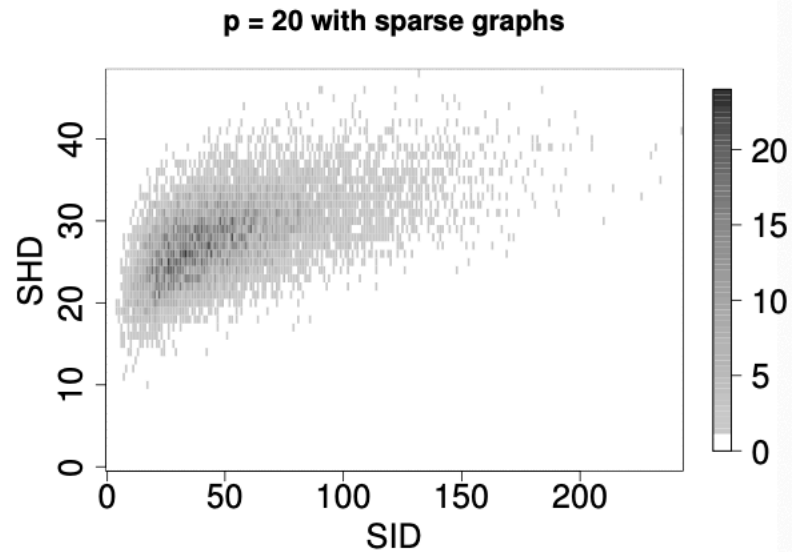
Graphical representation of the SID available!!

based on Shpitser et al: "On the validity of covariate adjustment for estimating causal effects", UAI 2010.

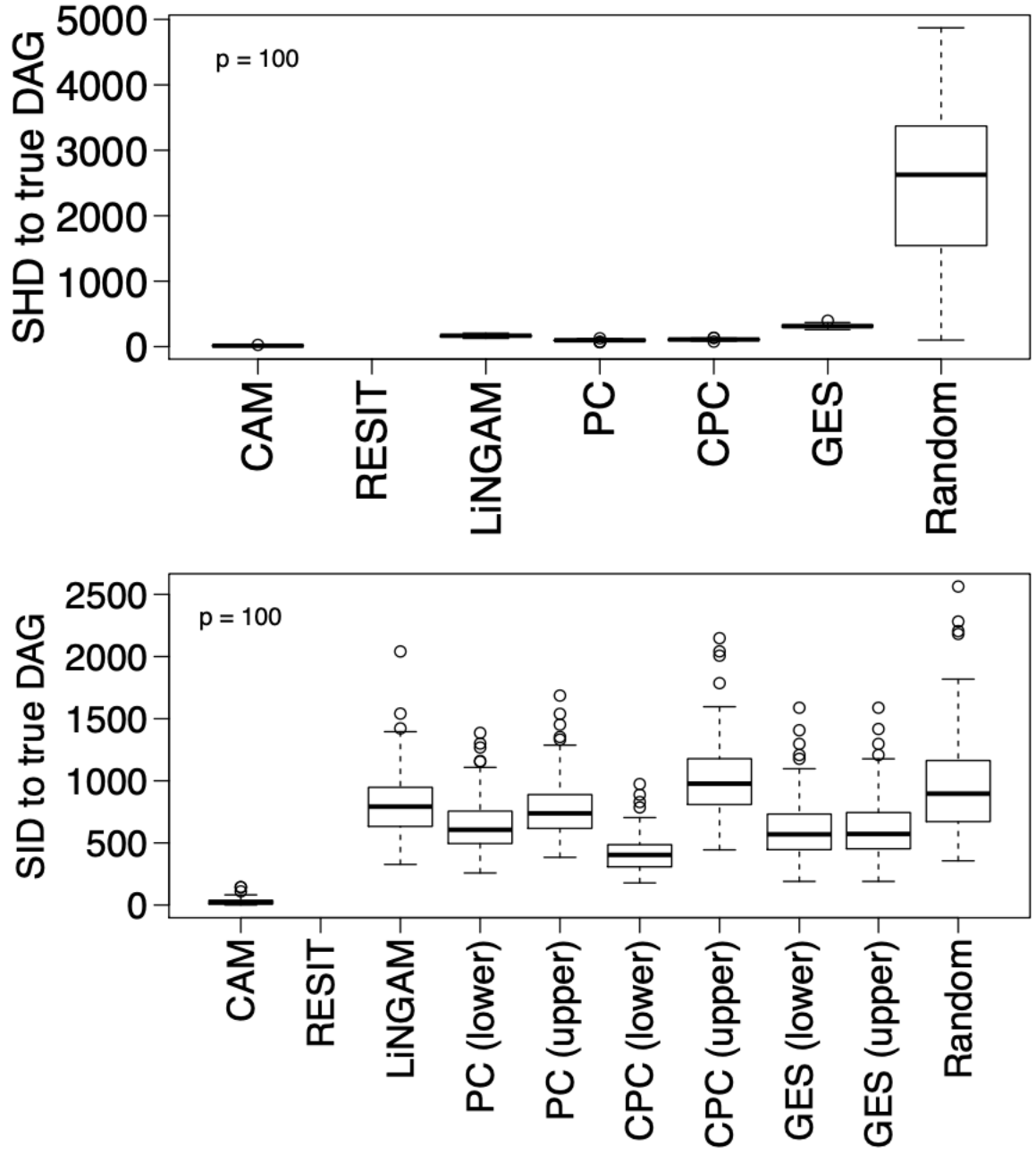
SHD and SID are quite different!

$10,000$ random DAGs

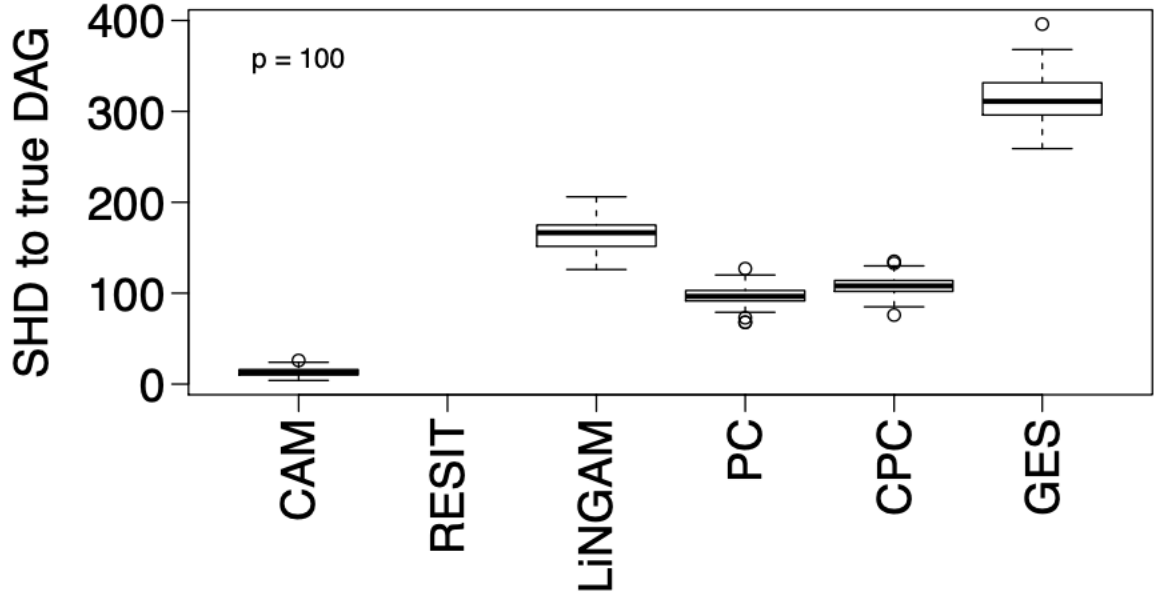


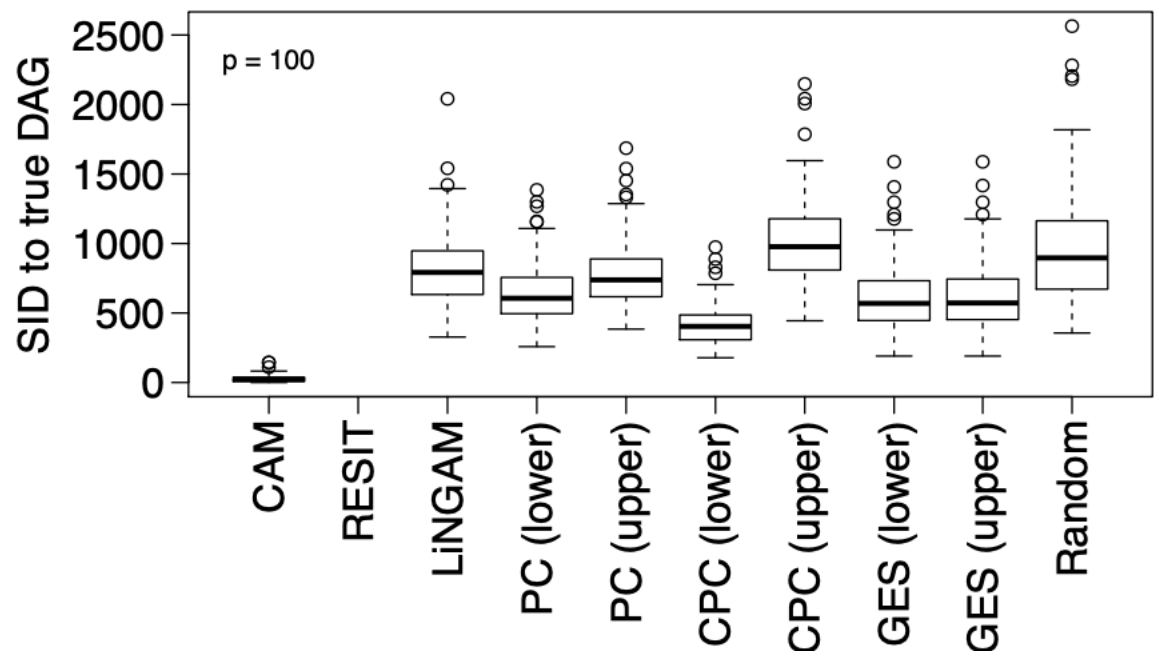


SHD and SID may lead to different conclusions (nonlinear, Gaussian).



SHD and SID may lead to different conclusions (nonlinear, Gaussian).

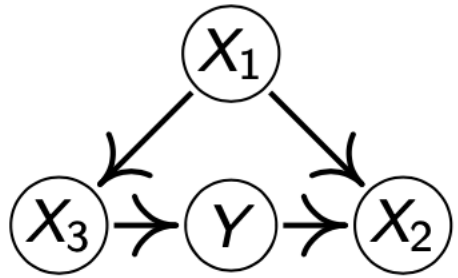


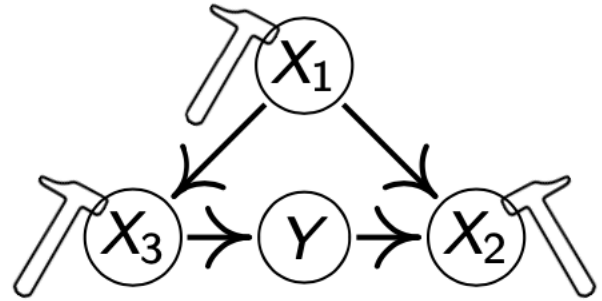


Idea 3: invariant causal prediction

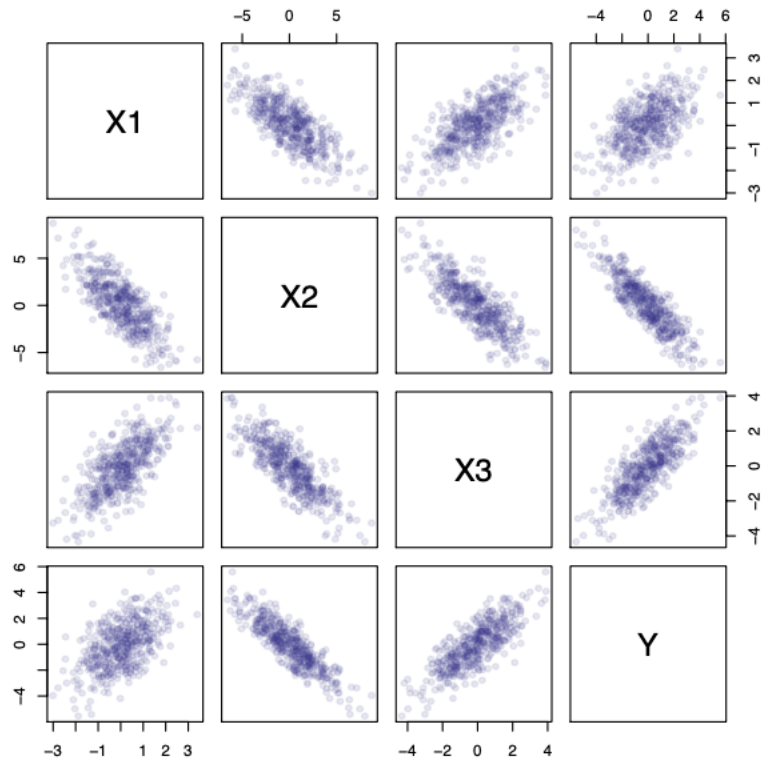
Concentrate on one target variable.

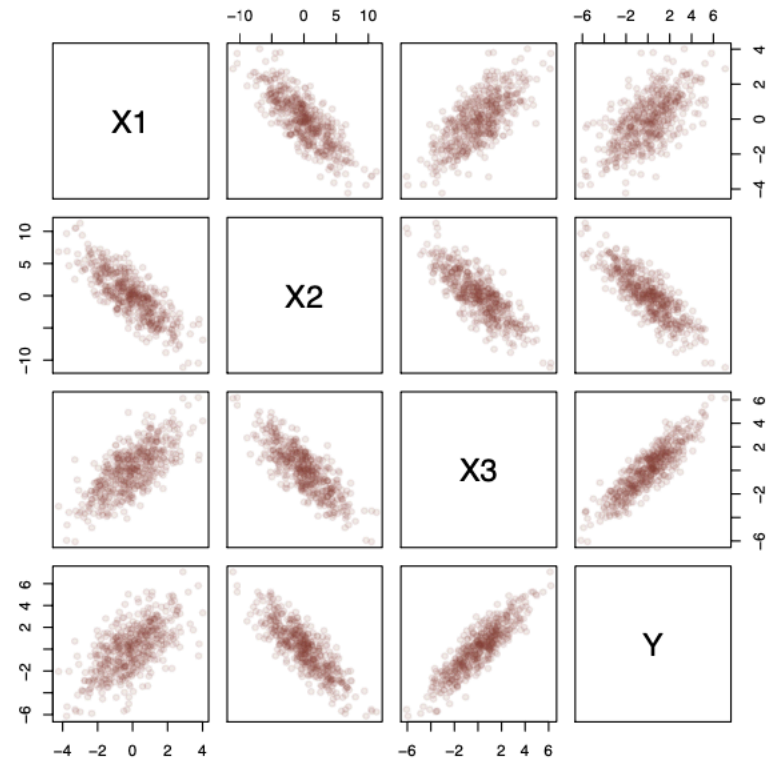
unknown:





known:





linear model

```
> linmod <- lm(Y~X)
> summary(linmod)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.000322   0.025858   0.012     0.99
X1          -0.444534   0.034306  -12.958   <2e-16 ***
X2          -0.402398   0.016471  -24.430   <2e-16 ***
X3           0.603502   0.025642   23.536   <2e-16 ***
```

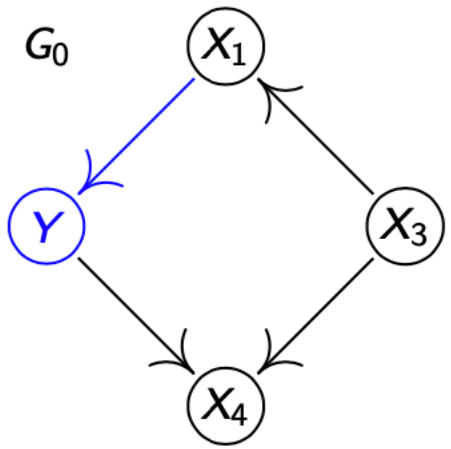
ICP (R-package InvariantCausalPrediction)

```
> ExpInd

      [1]1111111111111111111111111111111...22222222222222...

> icp <- ICP(X,Y,ExpInd)

             LOWER BOUND  UPPER BOUND  MAXIMIN EFFECT  P-VALUE
Variable_1         -0.11         0.10            0.00   1.0000
Variable_2         -0.33         0.00            0.00   1.0000
Variable_3          0.47         1.05            0.47   0.0012 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Key idea: MUTE ... or:
$P(Y \mid \mathbf{PA}_Y)$ remains invariant if the struct. equ. for Y does not change.

$$X_1 := f_1(X_3, N_1)$$
$$Y := f_2(X_1, N_2)$$
$$X_3 := f_3(N_3)$$
$$X_4 := f_4(Y, X_3, N_4)$$

- N_i jointly independent
- G_0 has no cycles

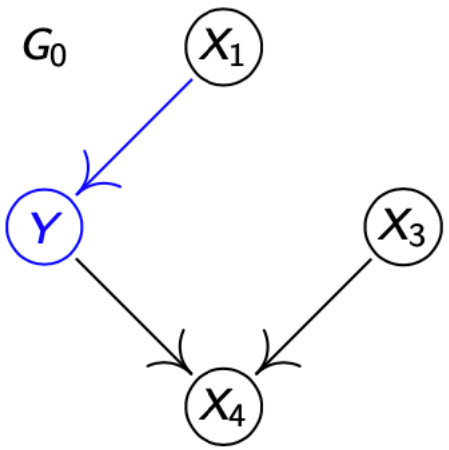


IMPORTANT: modularity, autonomy
Haavelmo 1944, Aldrich 1989, Pearl 2009, Schölkopf et al. 2012, Bareinboim et al. 2014, Hauser et al. 2014, ...

Key idea:
$P(Y \mid \mathbf{PA}_Y)$ remains invariant if the struct. equ. for Y does not change.

$$X_1 := \tilde{f}_1(\tilde{N}_1)$$
$$Y := f_2(X_1, N_2)$$
$$X_3 := f_3(N_3)$$
$$X_4 := f_4(Y, X_3, N_4)$$

- N_i jointly independent
- G_0 has no cycles

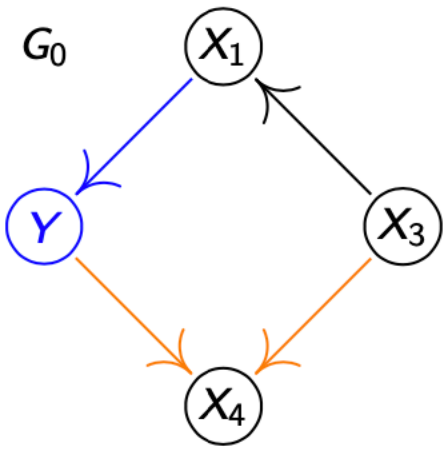


IMPORTANT: modularity, autonomy
Haavelmo 1944, Aldrich 1989, Pearl 2009, Schölkopf et al. 2012, Bareinboim et al. 2014, Hauser et al. 2014, ...

Key idea:
$P(Y \mid \mathbf{PA}_Y)$ remains invariant if the struct. equ. for Y does not change.

$$X_1 := f_1(X_3, N_1)$$
$$Y := f_2(X_1, N_2)$$
$$X_3 := f_3(N_3)$$
$$X_4 := \tilde{f}_4(Y, X_3, \tilde{N}_4)$$

- N_i jointly independent
- G_0 has no cycles

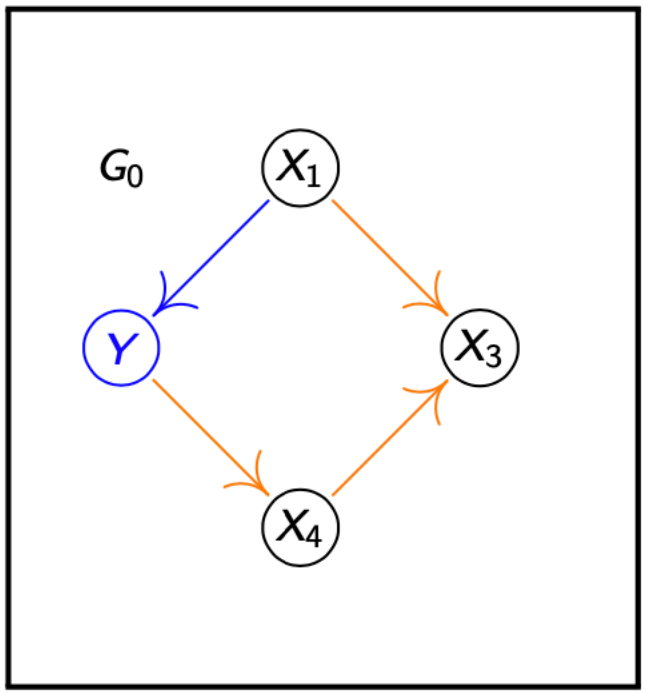


IMPORTANT: modularity, autonomy
Haavelmo 1944, Aldrich 1989, Pearl 2009, Schölkopf et al. 2012, Bareinboim et al. 2014, Hauser et al. 2014, ...

Key idea:
$P(Y \mid \mathbf{PA}_Y)$ remains invariant if the struct. equ. for Y does not change.

$$X_1 := \tilde{f}_1(\tilde{N}_1)$$
$$Y := f_2(X_1, N_2)$$
$$X_3 := \tilde{f}_3(X_1, X_4, \tilde{N}_3)$$
$$X_4 := \tilde{f}_4(Y, \tilde{N}_4)$$

- N_i jointly independent
- G_0 has no cycles



IMPORTANT: modularity, autonomy
Haavelmo 1944, Aldrich 1989, Pearl 2009, Schölkopf et al. 2012, Bareinboim et al. 2014, Hauser et al. 2014, ...

Given: Data from different environments $e \in \mathcal{E}$, e.g. interventions.

Given: Data from different environments $e \in \mathcal{E}$, e.g. interventions.

Proposition

Let $S^* = \mathbf{PA}_Y$. Then, $H_{0,S^*}(\mathcal{E})$ is true, i.e.,

$$\text{for all } e \in \mathcal{E}: \quad X^e \text{ has an arbitrary distribution and}$$
$$Y^e \,|\, X^e_{S^*} = x \quad \text{invariant}.$$

Given: Data from different environments $e \in \mathcal{E}$, e.g. interventions.

Proposition

Let $S^* = \mathbf{PA}_Y$. Then, $H_{0,S^*}(\mathcal{E})$ is true, i.e., there exists γ^* with support S^* that satisfies

$$\text{for all } e \in \mathcal{E}: \quad X^e \text{ has an arbitrary distribution and}$$

$$\sout{Y^e \mid X^e_{S^*} = x \quad \text{invariant}.}$$

$$Y^e = X^e \gamma^* + \varepsilon^e, \quad \varepsilon^e \sim F_\varepsilon \text{ and } \varepsilon^e \perp\!\!\!\perp X^e_{S^*}.$$

Given: Data from different environments $e \in \mathcal{E}$, e.g. interventions.

Proposition

Let $S^* = \mathbf{PA}_Y$. Then, $H_{0,S^*}(\mathcal{E})$ is true, i.e., there exists γ^* with support S^* that satisfies

$$\text{for all } e \in \mathcal{E}: \quad X^e \text{ has an arbitrary distribution and}$$

$$\sout{Y^e \mid X^e_{S^*} = x \quad \text{invariant}.}$$

$$Y^e = X^e \gamma^* + \varepsilon^e, \quad \varepsilon^e \sim F_\varepsilon \text{ and } \varepsilon^e \perp\!\!\!\perp X^e_{S^*}.$$

Goal: Find S^*.

Idea: Check $H_{0,S}(\mathcal{E})$ for several candidates S.

$$S(\mathcal{E}) := \bigcap_{S : H_{0,S}(\mathcal{E}) \text{ is true}} S$$

set	$\{3,5\}$	$\{3,7\}$	$S^* = \{1,3,6\}$	$\{2\}$	$\{3,8\}$	$\cdots$
inv. pred.	✓	✗	✓	✗	✓	$\cdots$

Theorem (PBM 2016)

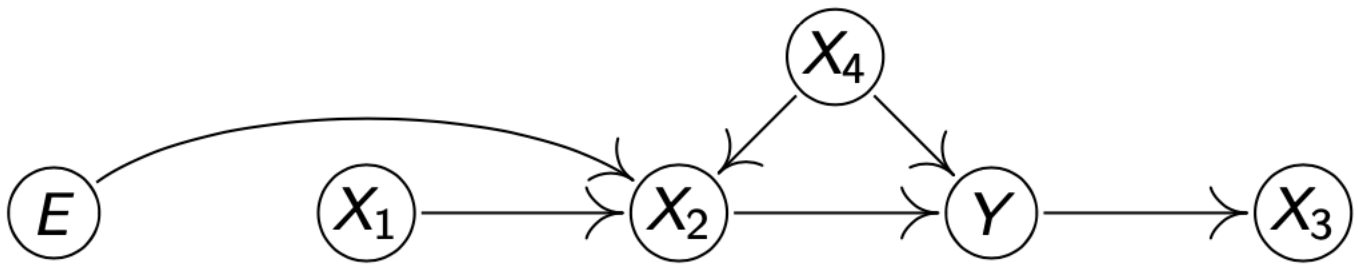
$$P(\hat{S}(\mathcal{E}) \subseteq S^*) \geq 1 - \alpha\,.$$

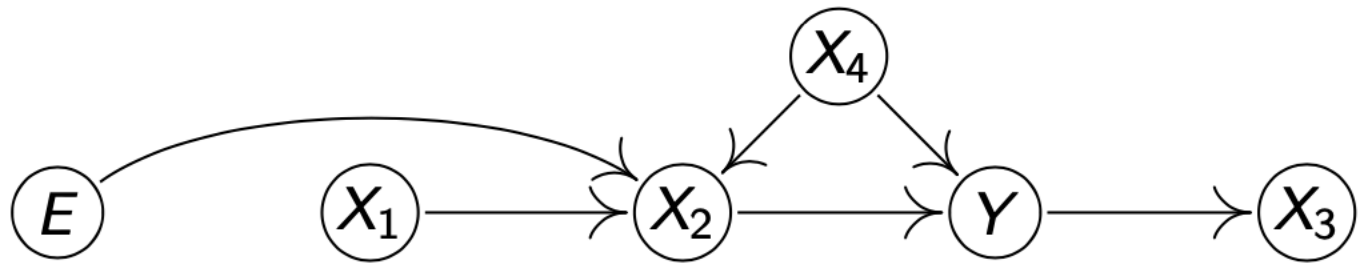
Theorem (PBM 2016)

$$P(\hat{S}(\mathcal{E}) \subseteq S^*) \geq 1 - \alpha\,.$$

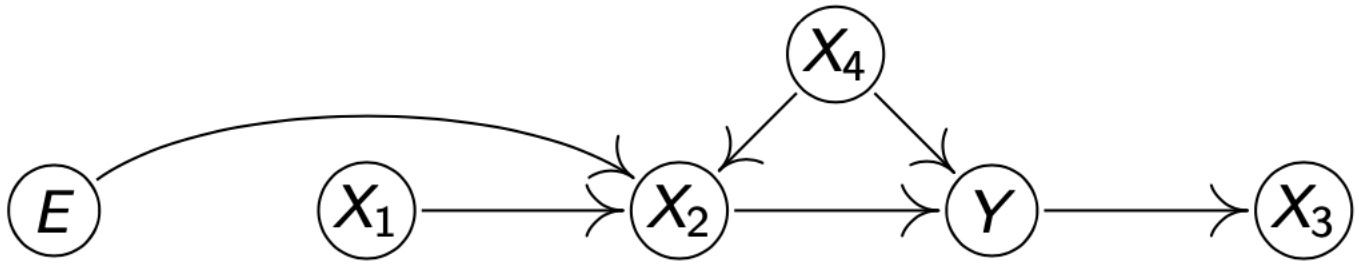
Identifiability improves if we have more and stronger interventions, at better places, more heterogeneity in the data.

JP, P. Bühlmann, N. Meinshausen: *Causal inference using invariant prediction: conf. interv.*, JRSS-B 2016 (w/ discussion).

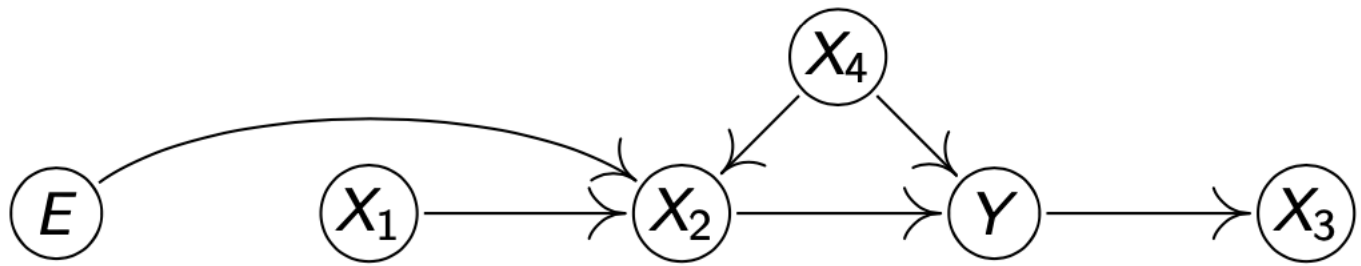

X_4
E
X_1
X_2
Y
X_3



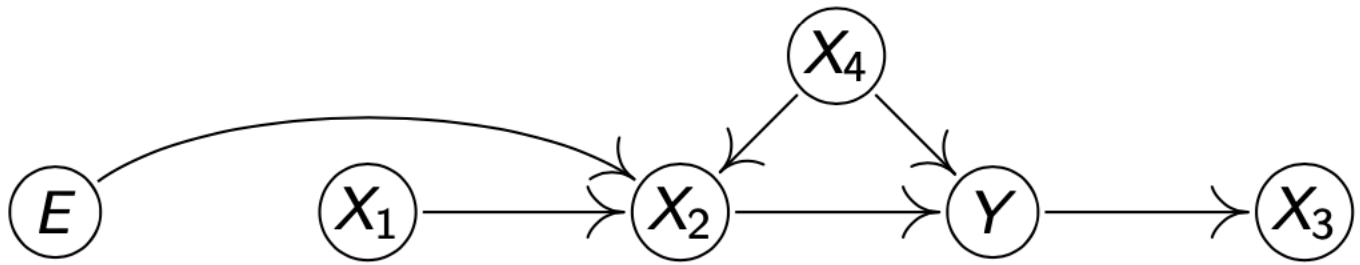

```
> Y <- X[,2] + X[,4] + noise
> ICP(X,Y,ExpInd)
```




```
> Y <- X[,2] + X[,4] + noise
> ICP(X,Y,ExpInd)

accepted set of variables: 2,4
accepted set of variables: 1,2,4
accepted set of variables: 2,3,4
accepted set of variables: 1,2,3,4

     LOWER BOUND  UPPER BOUND  MAXIMIN EFFECT  P-VALUE
X1         -0.03         0.01            0.00     0.48
X2          0.98         1.01            0.98  < 1e-09 ***
X3         -0.07         0.00            0.00     0.48
X4          0.95         1.01            0.95  2.6e-05 ***
```



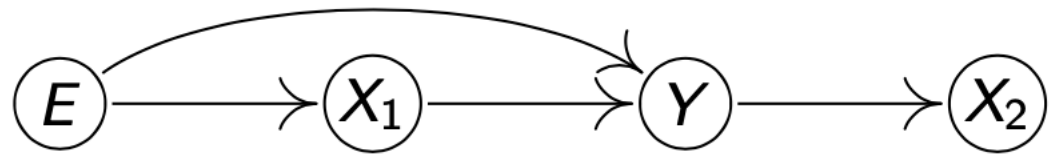

```
> Y <- X[,2]^2 + X[,4] + noise
> ICP(X,Y,ExpInd)
```



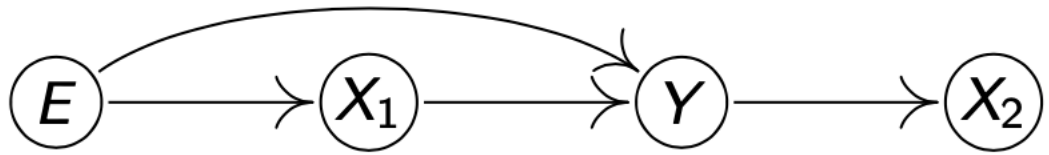

```
> Y <- X[,2]^2 + X[,4] + noise
> ICP(X,Y,ExpInd)

  empty set
  (all models rejected)
```

Model violation: nonlinear models

$\rightsquigarrow$ usually leads to loss of power, not coverage



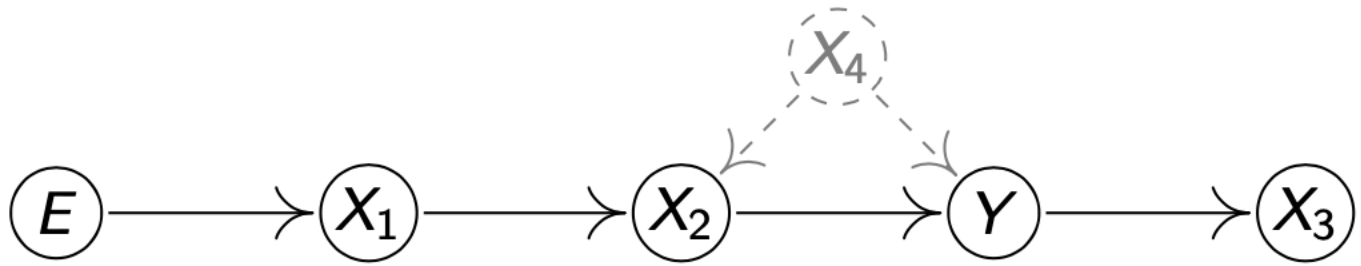

```
> Y <- X[,1] + E + noise
> ICP(X,Y,ExpInd)
```



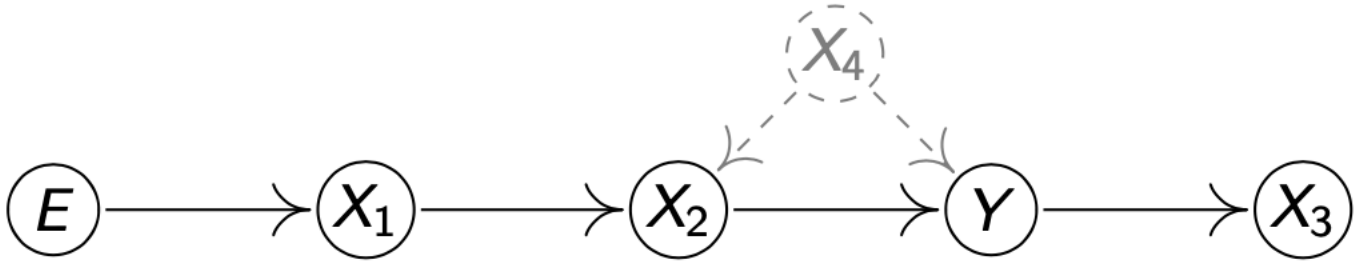

```
> Y <- X[,1] + E + noise
> ICP(X,Y,ExpInd)

  empty set
  (all models rejected)
```

Model violation: intervention on Y

$\rightsquigarrow$ usually leads to loss of power, not coverage



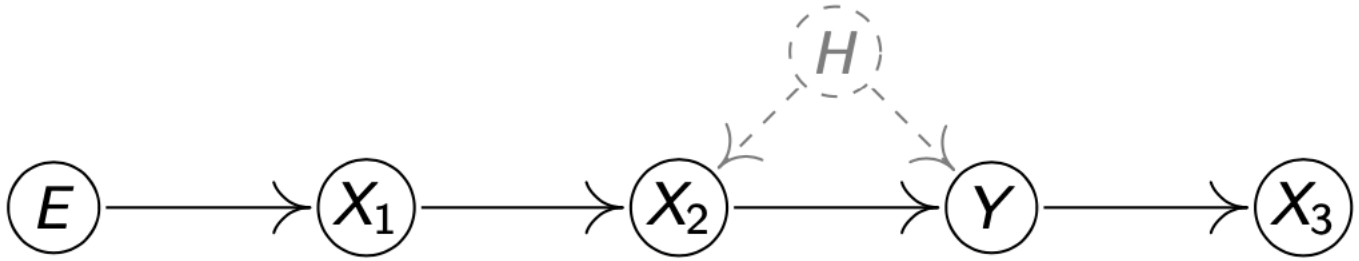

```
> Y <- X[,2] + X[,4] + noise
> ICP(X[,1:3],Y,ExpInd)
```




```
> Y <- X[,2] + X[,4] + noise
> ICP(X[,1:3],Y,ExpInd)

accepted set of variables: 1
accepted set of variables: 1,2
accepted set of variables: 1,3
accepted set of variables: 1,2,3

     LOWER BOUND  UPPER BOUND  MAXIMIN EFFECT  P-VALUE
X1         -0.87         1.05            0.00   <1e-09 ***
X2          0.00         1.86            0.00     1.00
X3         -1.61         0.00            0.00     0.73
```

Model violation: hidden variables

$\rightsquigarrow$ coverage still holds if we consider ancestors instead of parents



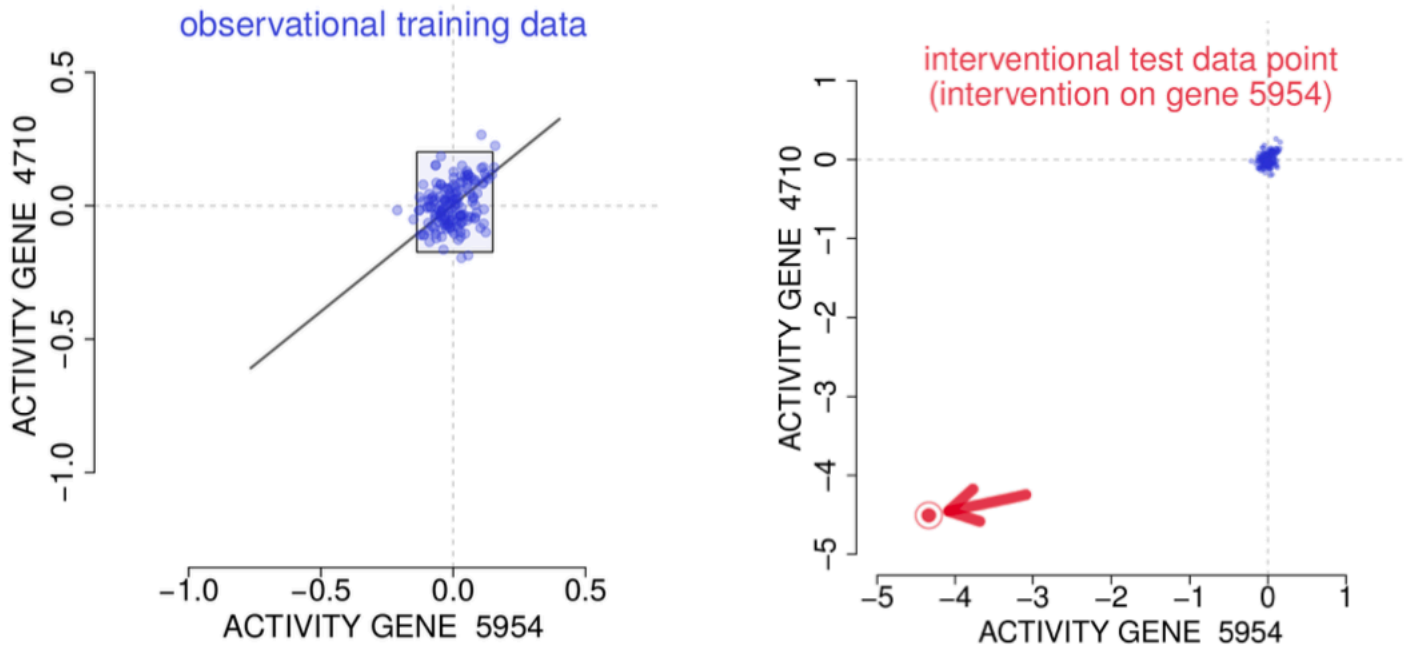
Theorem (PBM 2016)

Assume that the joint distribution over $(Y, X_1, \ldots, X_p, H_1, \ldots, H_q, E)$ *is faithful w.r.t. the augmented graph. Then*

$$S(\mathcal{E}) := \bigcap_{S\,:\,H_{0,S}(\mathcal{E})\ \text{is true}} S \subseteq \mathbf{AN}(Y) \cap \{X_1, \ldots, X_p\}.$$

Real data: genetic perturbation experiments for yeast (Kemmeren et al., 2014)

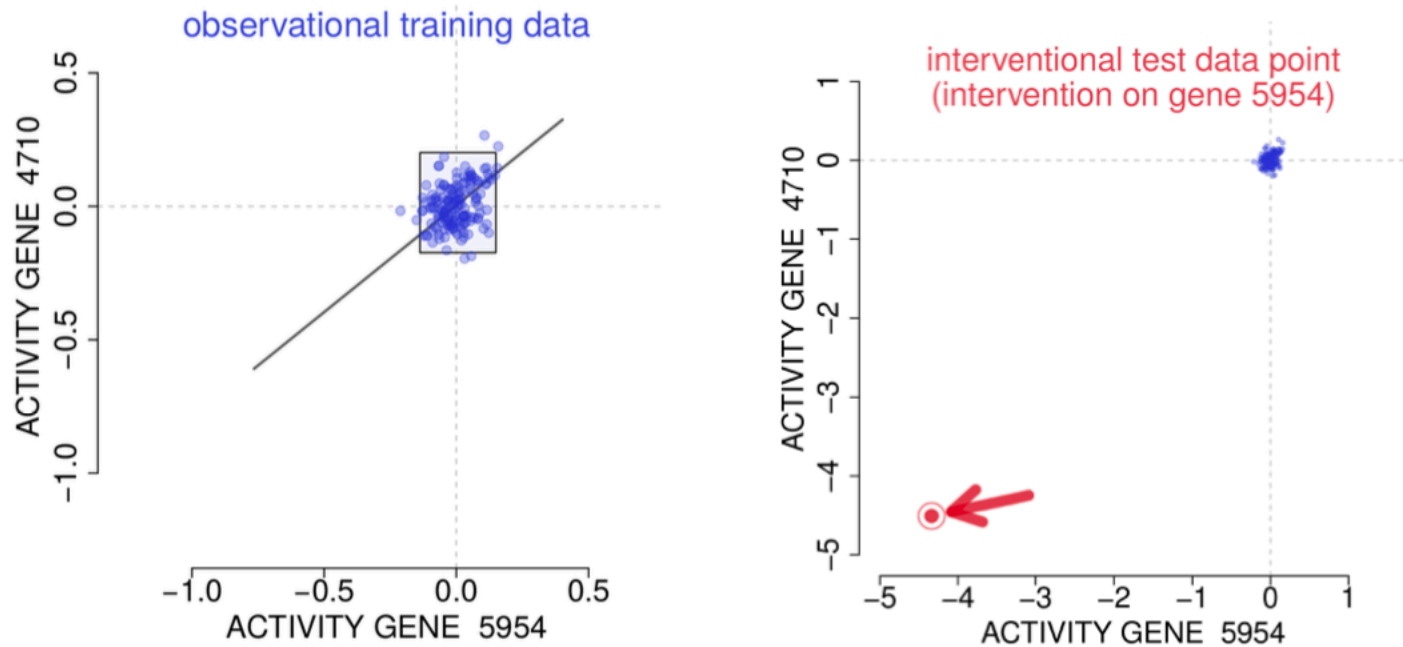
- $p = 6170$ genes
- $n_{obs} = 160$ wild-types
- $n_{int} = 1479$ gene deletions (targets known)



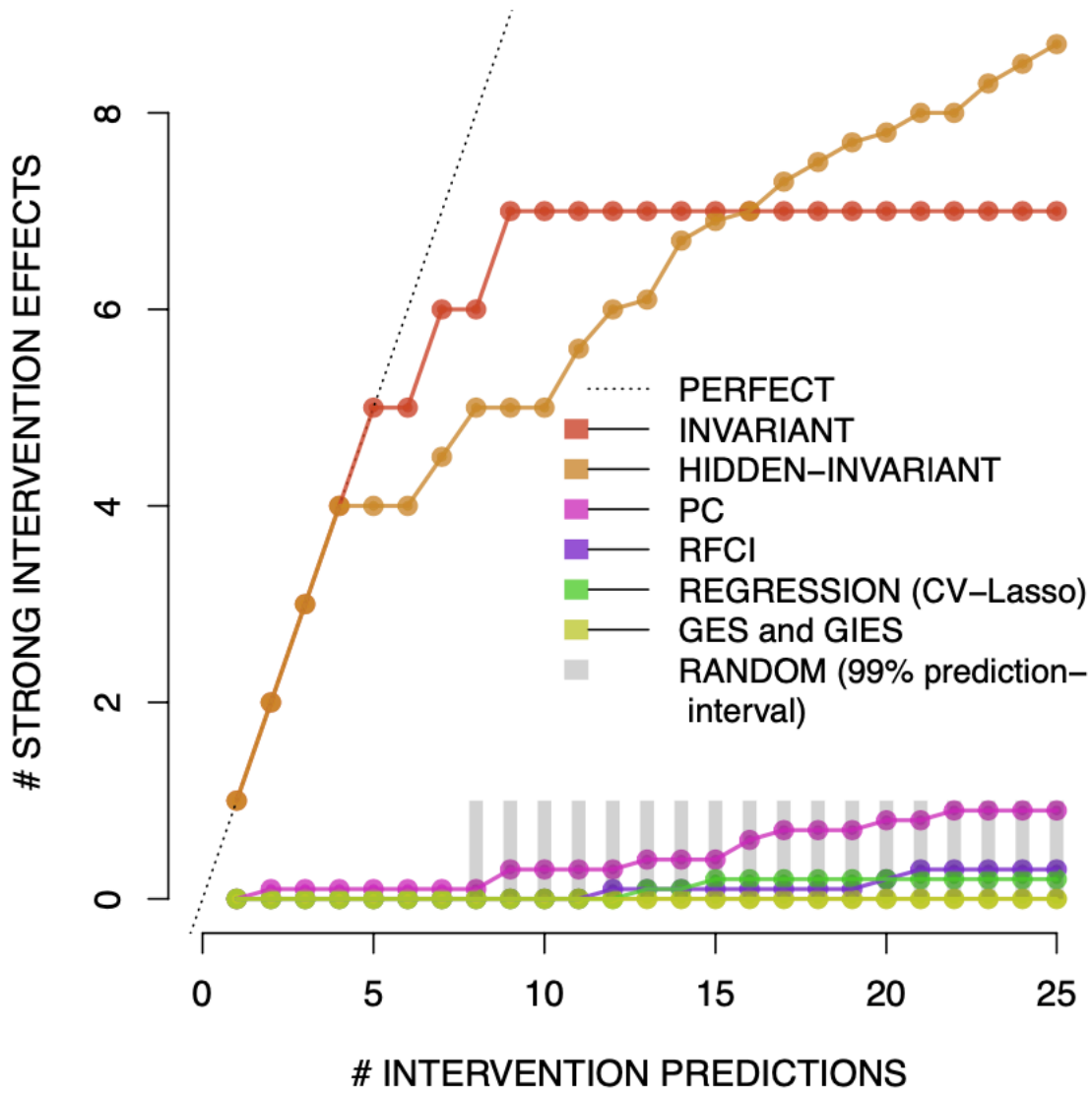
- true hits: $\approx 0.1\%$ of pairs

Real data: genetic perturbation experiments for yeast (Kemmeren et al., 2014)

- $p = 6170$ genes
- $n_{obs} = 160$ wild-types
- $n_{int} = 1479$ gene deletions (targets known)

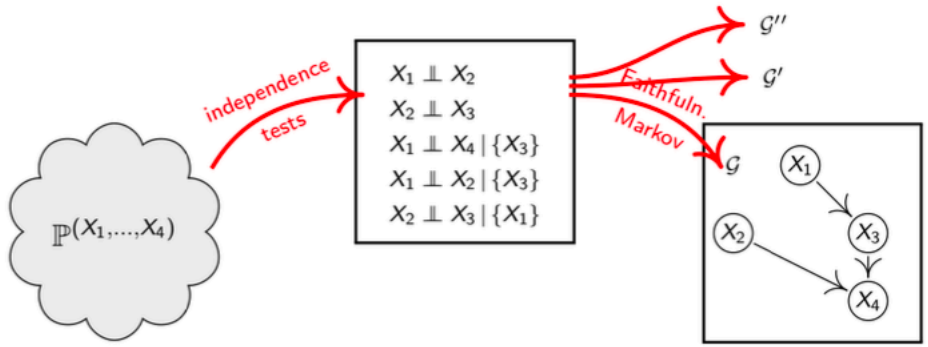


- true hits: $\approx 0.1\%$ of pairs
- our method: $\mathcal{E} = \{obs, int\}$


PERFECT
INVARIANT
HIDDEN−INVARIANT
PC
RFCI
REGRESSION (CV−Lasso)
GES and GIES
RANDOM (99% prediction−
interval)
STRONG INTERVENTION EFFECTS
0
2
4
6
8
0
5
10
15
20
25
INTERVENTION PREDICTIONS

Summary Part II:

- Idea 1: independence-based methods (single environment)



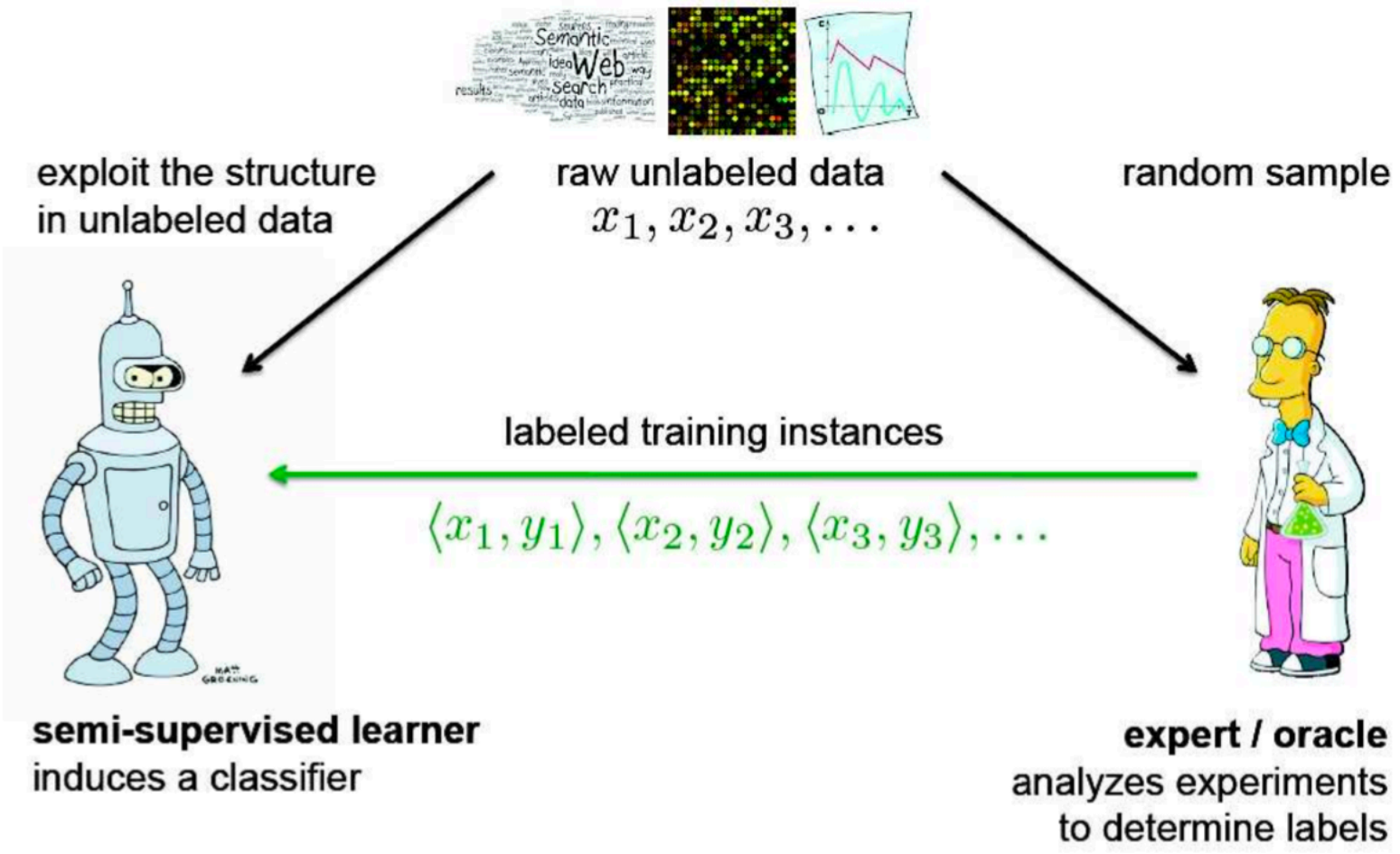
- Idea 2: additive noise (single environment)

$$X_1 = f_1(X_3) + N_1$$
$$X_2 = N_2$$
$$X_3 = f_3(X_2) + N_3$$
$$X_4 = f_4(X_2, X_3) + N_4$$

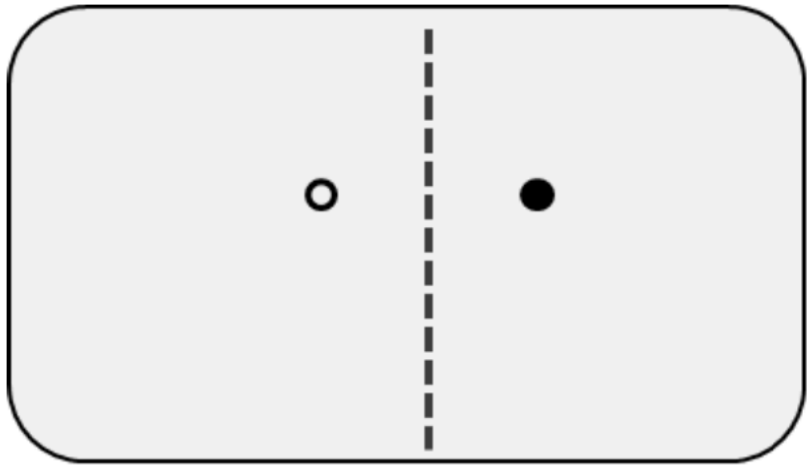
- Idea 3: invariant prediction (the more heterogeneity the better!)

Part III: Applications to Machine Learning

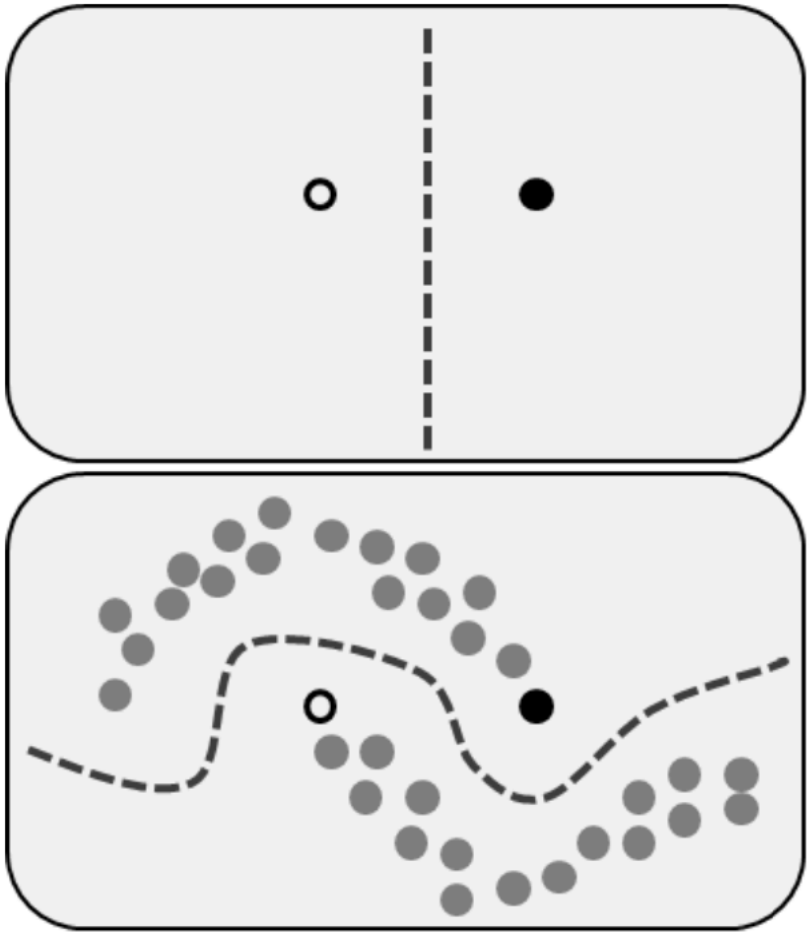
Semi-supervised learning



Semi-supervised learning



Semi-supervised learning



Idea 1: Semi-supervised learning

Consider a Markov factorization w.r.t. causal DAG:

$$p(x_1, \ldots, x_d) = \prod_{i=1}^{d} p(x_i \mid x_{pa(i)})$$

Idea 1: Semi-supervised learning

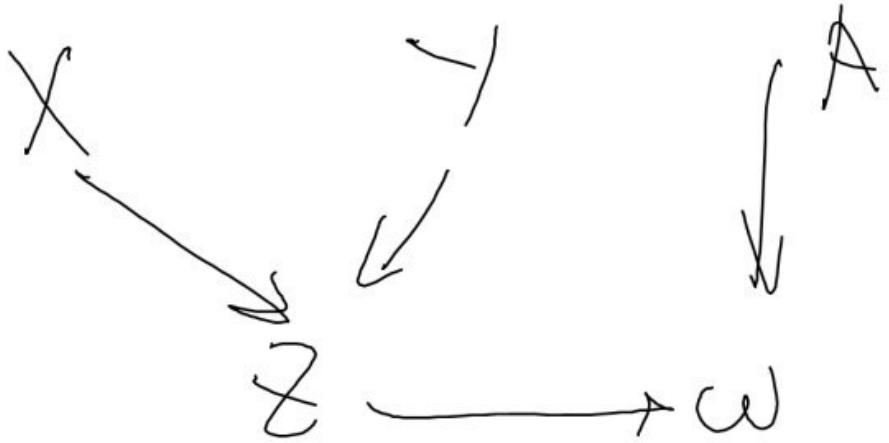
Consider a Markov factorization w.r.t. causal DAG:

$$p(x_1, \ldots, x_d) = \prod_{i=1}^{d} p(x_i \mid x_{pa(i)})$$

Modularity suggests:

$p(x_1 \mid x_{pa(1)}), \ldots, p(x_d \mid x_{pa(d)})$ are "independent"

Idea 1: Semi-supervised learning



$$p(A,X,Y,Z,W) =$$
$$= p(X)\,p(Y)\,\underline{p(Z|X,Y)}\cdot$$
$$\cdot p(A)\,\underline{p(W|A,Z)}$$

$$X = N_X$$
$$Z = f(X,Y) + N_Z$$
$$W = g(A,Z) + N_W$$

Idea 1: Semi-supervised learning

Consider a Markov factorization w.r.t. causal DAG:

$$p(x_1, \ldots, x_d) = \prod_{i=1}^{d} p(x_i \mid x_{pa(i)})$$

Modularity suggests:

$$p(x_1 \mid x_{pa(1)}), \ldots, p(x_d \mid x_{pa(d)}) \text{ are "independent"}$$

Special case:

$$p(\mathit{cause}), p(\mathit{effect} \mid \mathit{cause}) \text{ are "independent"}$$

Idea 1: Semi-supervised learning

Consider a Markov factorization w.r.t. causal DAG:

$$p(x_1, \ldots, x_d) = \prod_{i=1}^{d} p(x_i \mid x_{pa(i)})$$

Modularity suggests:

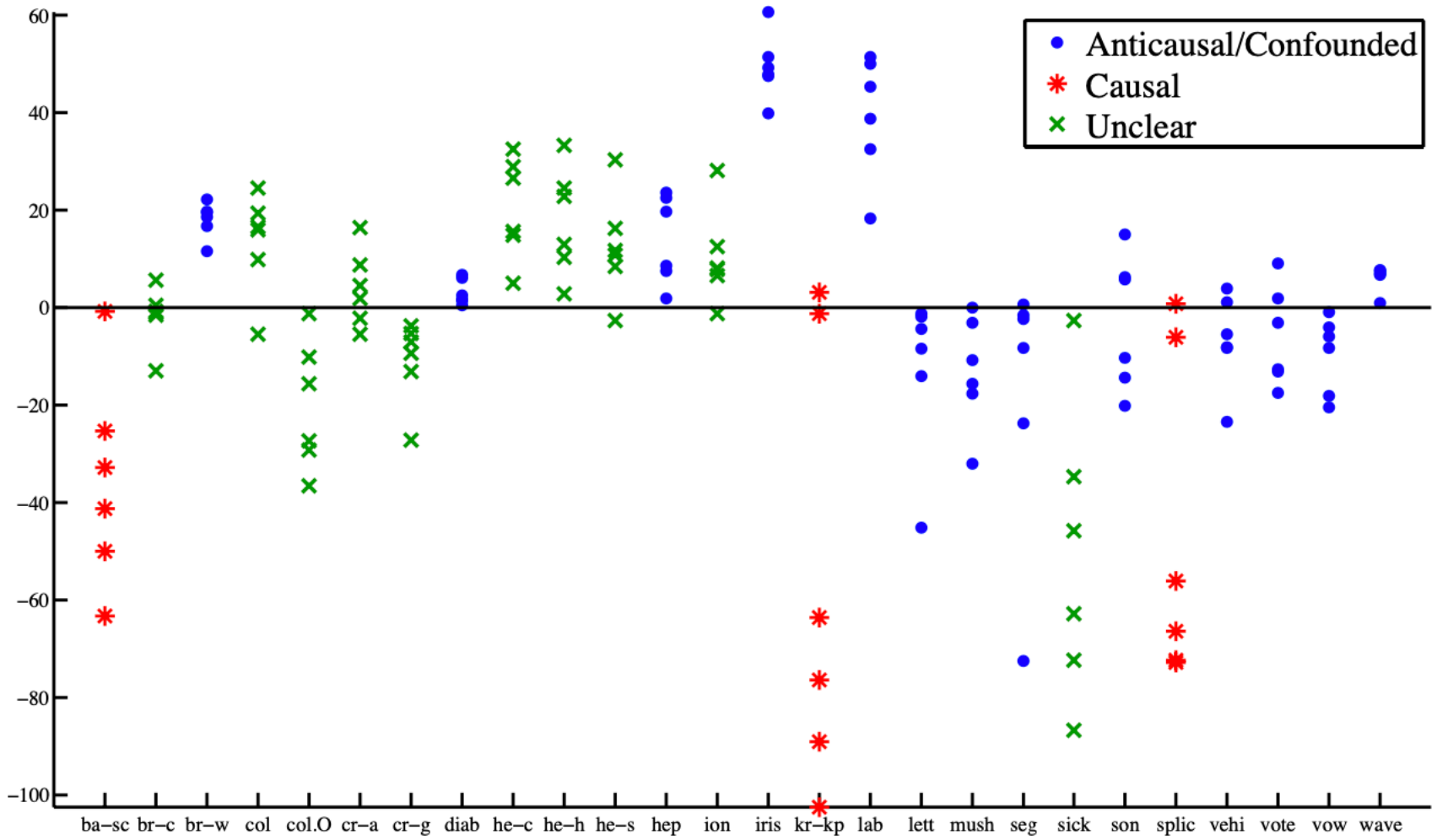
$$p(x_1 \mid x_{pa(1)}), \ldots, p(x_d \mid x_{pa(d)}) \text{ are "independent"}$$

Special case:

$$p(cause), p(effect \mid cause) \text{ are "independent"}$$

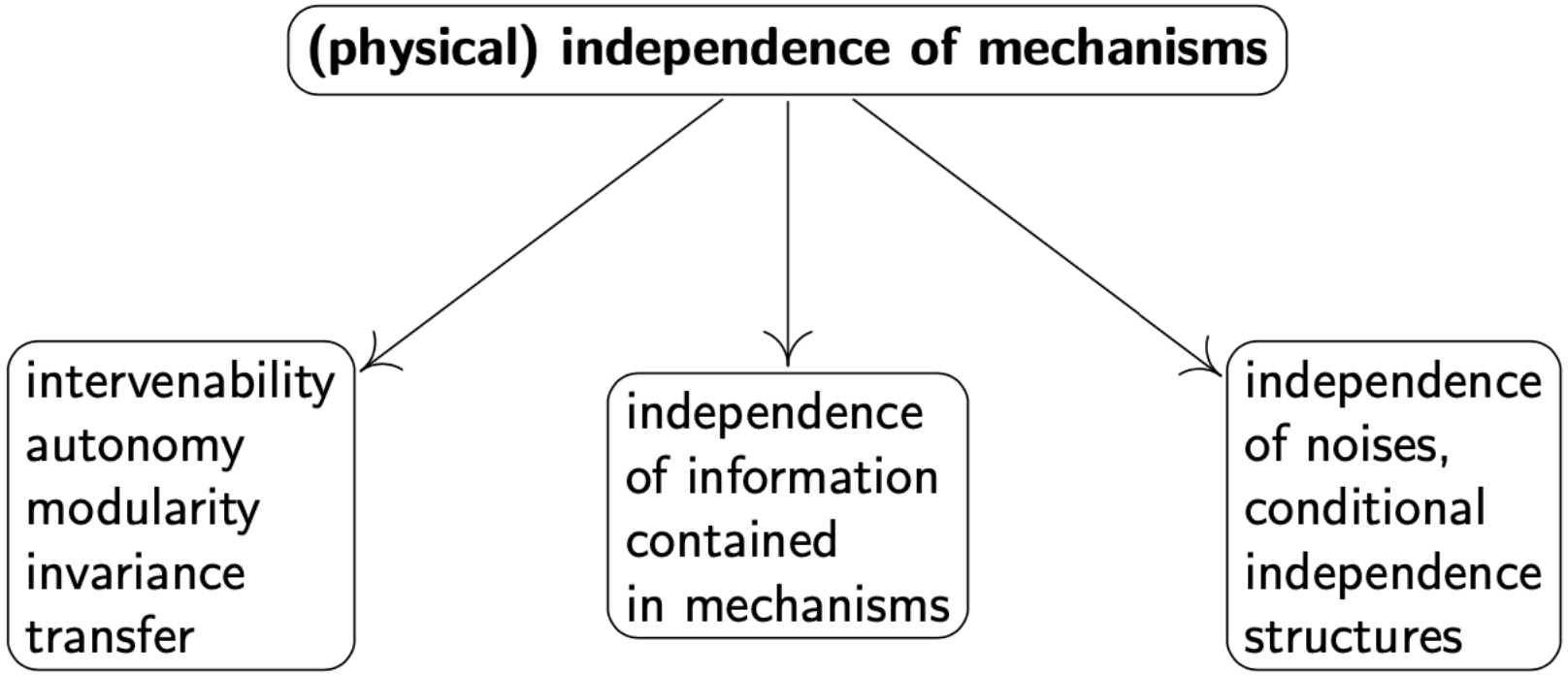
But then: Semi-supervised Learning does not work from cause to effect.

Idea 1: Semi-supervised learning

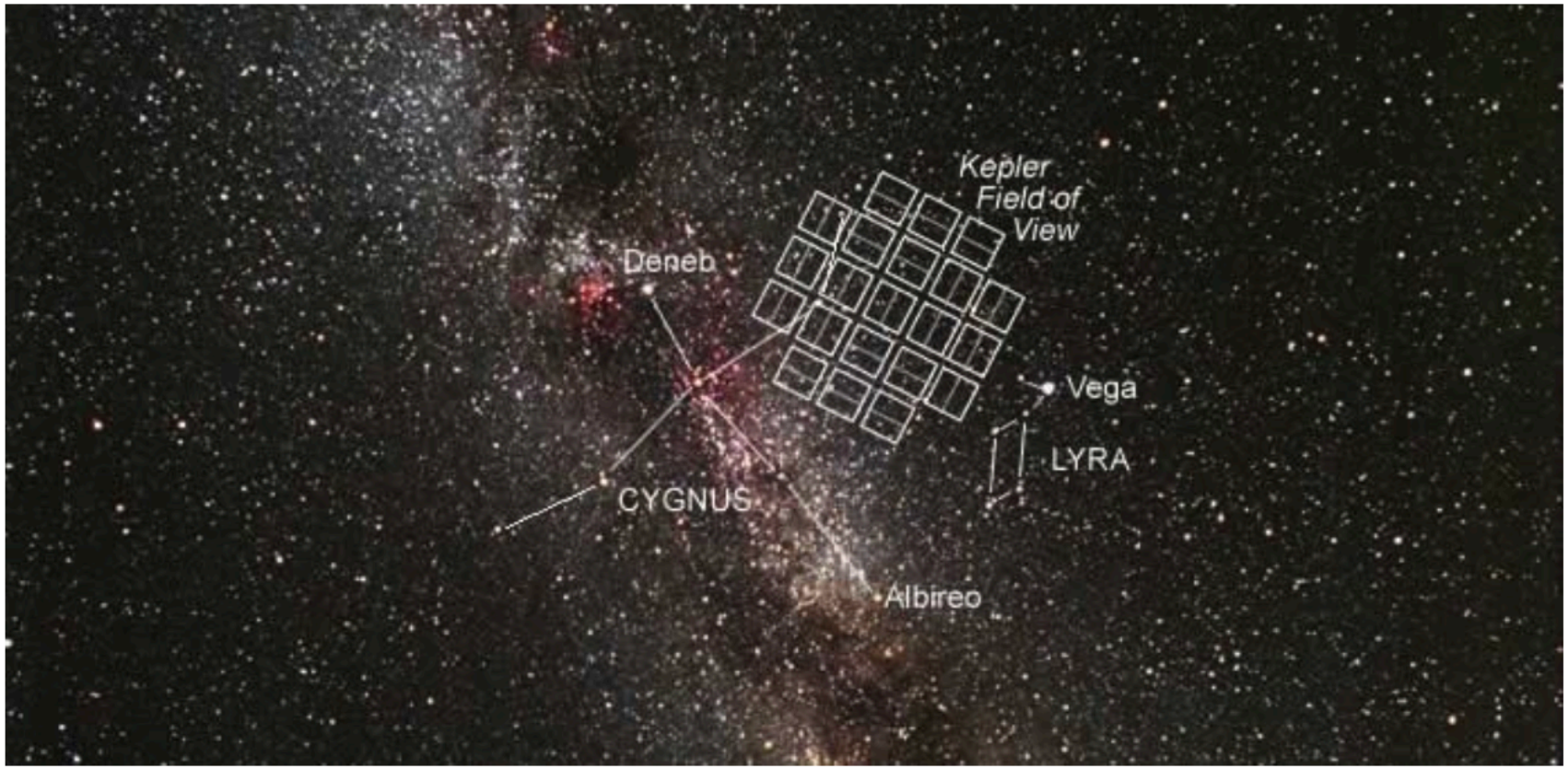


Schölkopf et al.: *On causal and anticausal learning*, ICML 2012

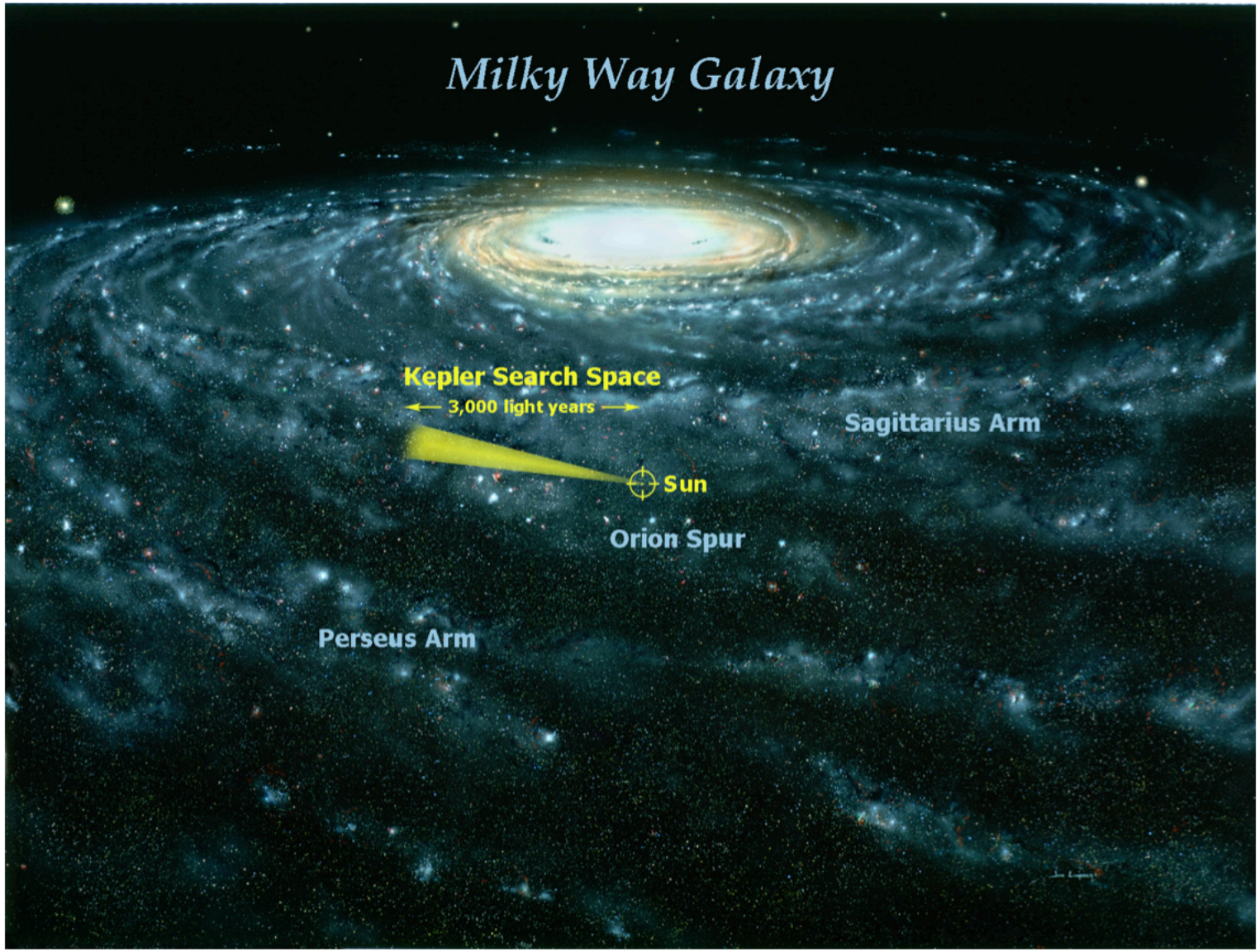
Idea 1: Semi-supervised learning



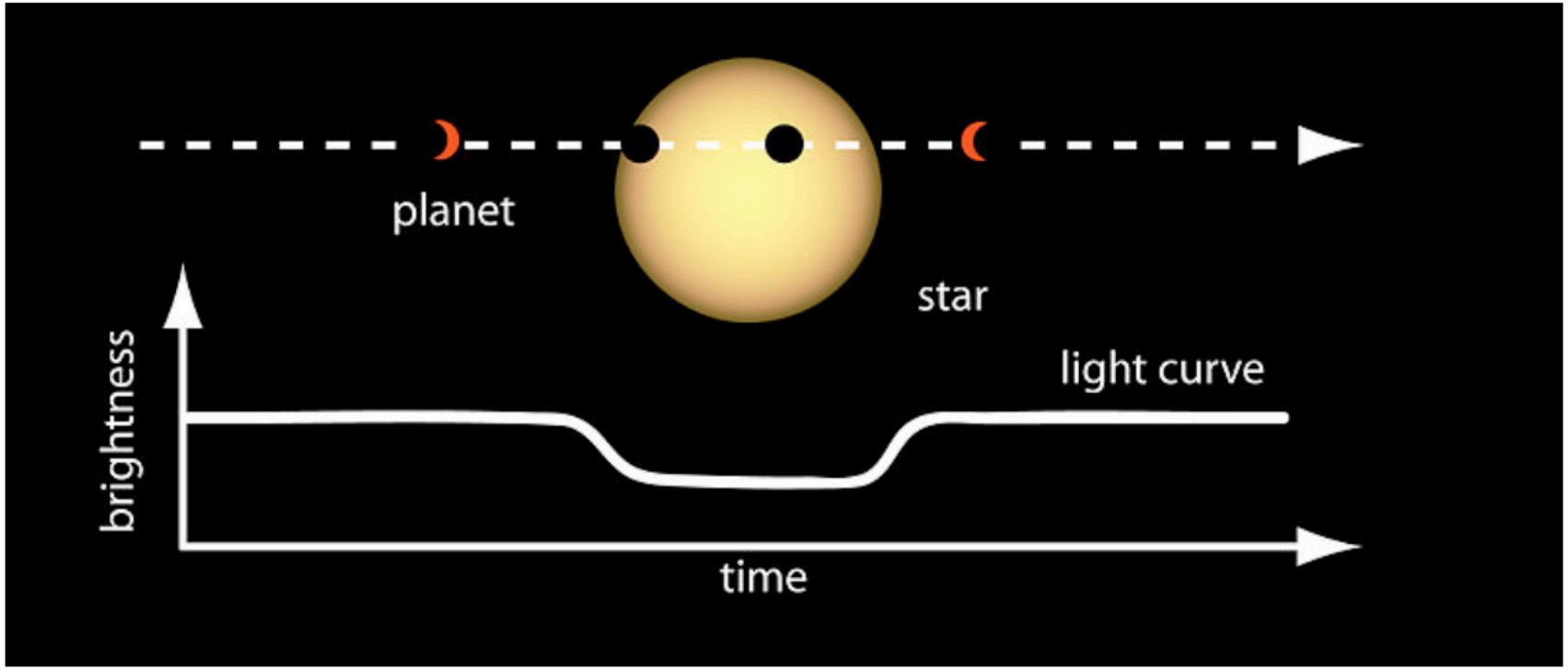
Idea 2: half-sibling regression



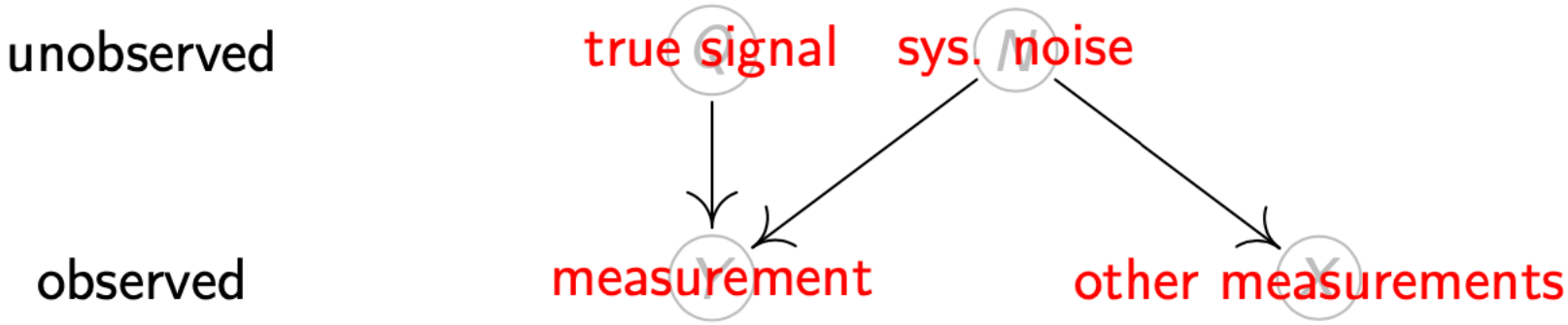
Idea 2: half-sibling regression



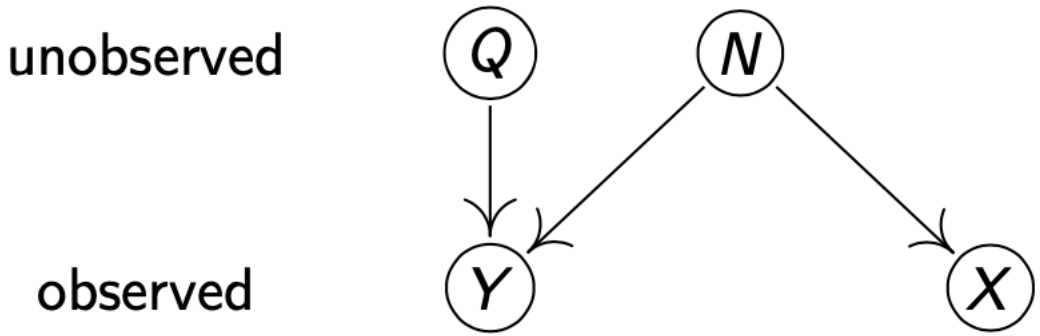
Idea 2: half-sibling regression



Idea 2: half-sibling regression

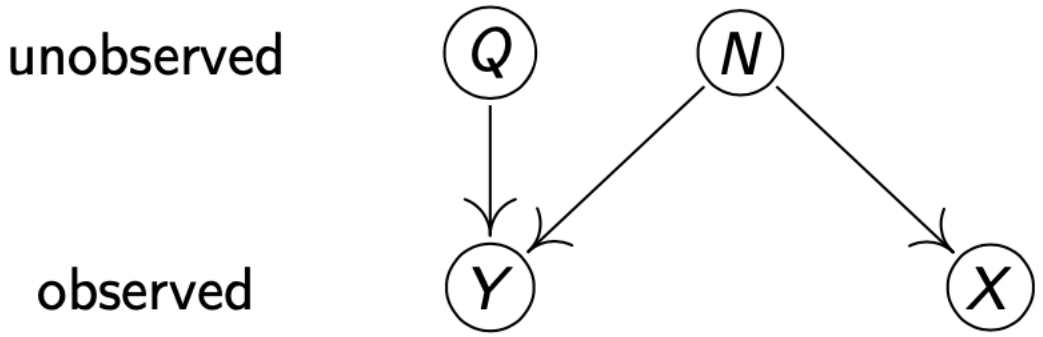


Idea 2: half-sibling regression



Assume $Y = f(N) + Q$.

Idea 2: half-sibling regression



Assume $Y = f(N) + Q$.

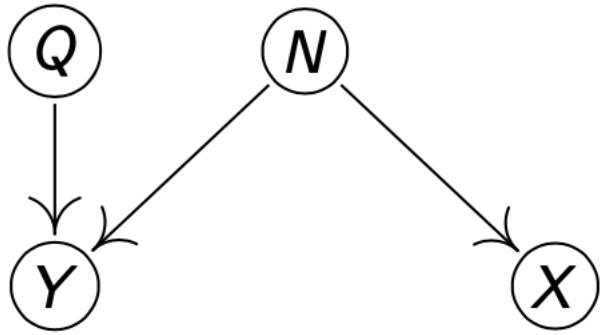
Proposed idea:

Remove everything from Y explained by X.

Idea 2: half-sibling regression

unobserved

observed



Assume $Y = f(N) + Q$.

Proposed idea:

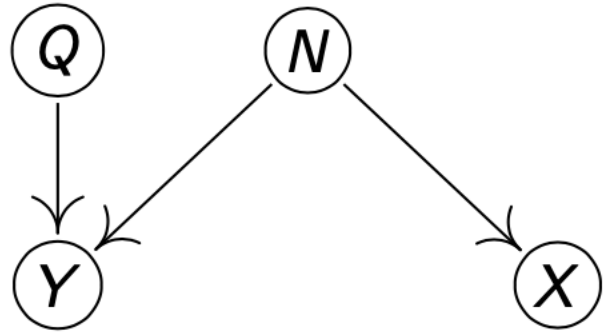
Remove everything from Y explained by X.

Or: $\hat{Q} := Y - \mathbf{E}[Y \mid X]$.

Idea 2: half-sibling regression

unobserved

observed



Assume $Y = f(N) + Q$.

Proposed idea:

Remove everything from Y explained by X.

Or: $\hat{Q} := Y - \mathbf{E}[Y \mid X]$.

Proposition

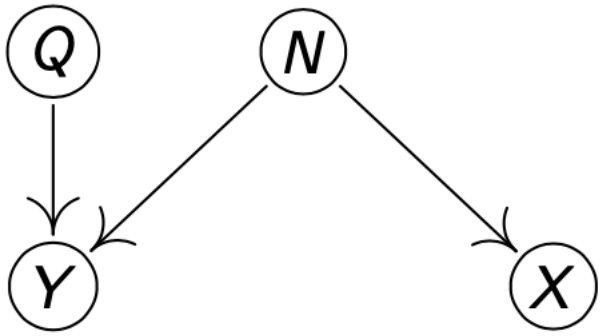
Convergence against "correct" signal Q (up to reparameterization) if

- *perfect reconstruction:* $\exists \psi$ *such that* $f(N) = \psi(X)$

Idea 2: half-sibling regression

unobserved

observed



Assume $Y = f(N) + Q$.

Proposed idea:

Remove everything from Y explained by X.

Or: $\hat{Q} := Y - \mathbf{E}[Y \mid X]$.

Proposition

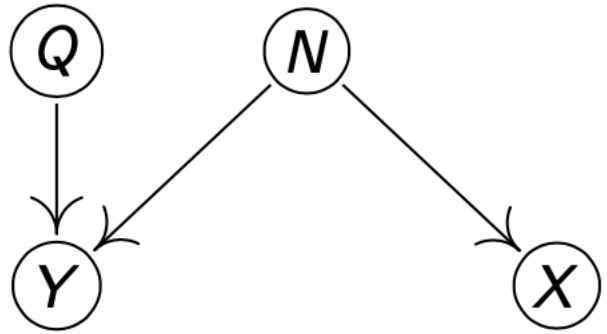
Convergence against "correct" signal Q (up to reparameterization) if

- *perfect reconstruction:* $\exists \psi$ *such that* $f(N) = \psi(X)$
- *low noise:* $X = g(N) + s \cdot R$ *and* $s \to 0$

Idea 2: half-sibling regression

unobserved

observed



Assume $Y = f(N) + Q$.

Proposed idea:

Remove everything from Y explained by X.

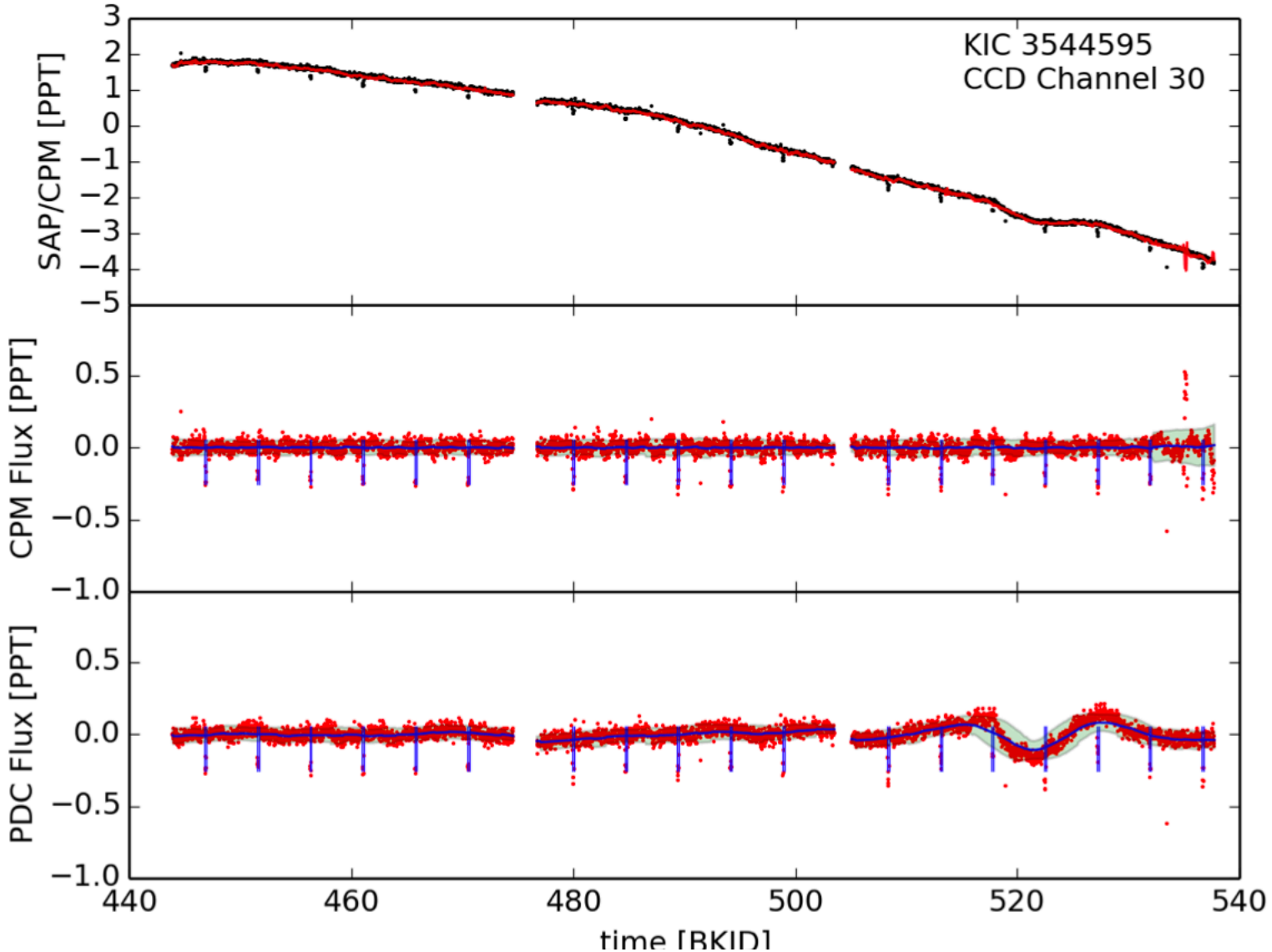
Or: $\hat{Q} := Y - \mathbf{E}[Y \mid X]$.

Proposition

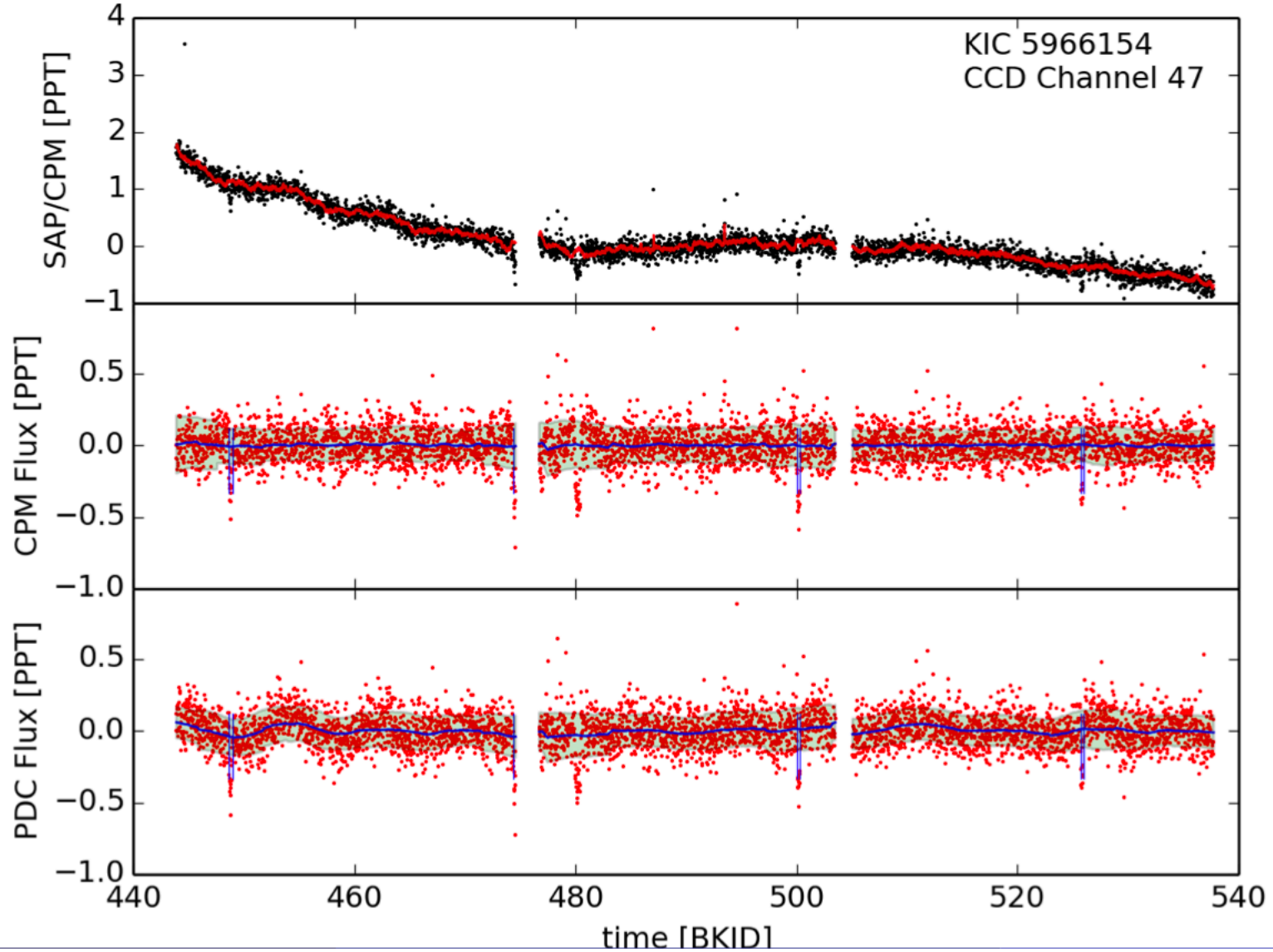
Convergence against "correct" signal Q (up to reparameterization) if

- *perfect reconstruction:* $\exists \psi$ *such that* $f(N) = \psi(X)$
- *low noise:* $X = g(N) + s \cdot R$ *and* $s \to 0$
- *limit of infinitely many X's:* $X_i = g_i(N) + R_i$, $i = 1, \ldots$

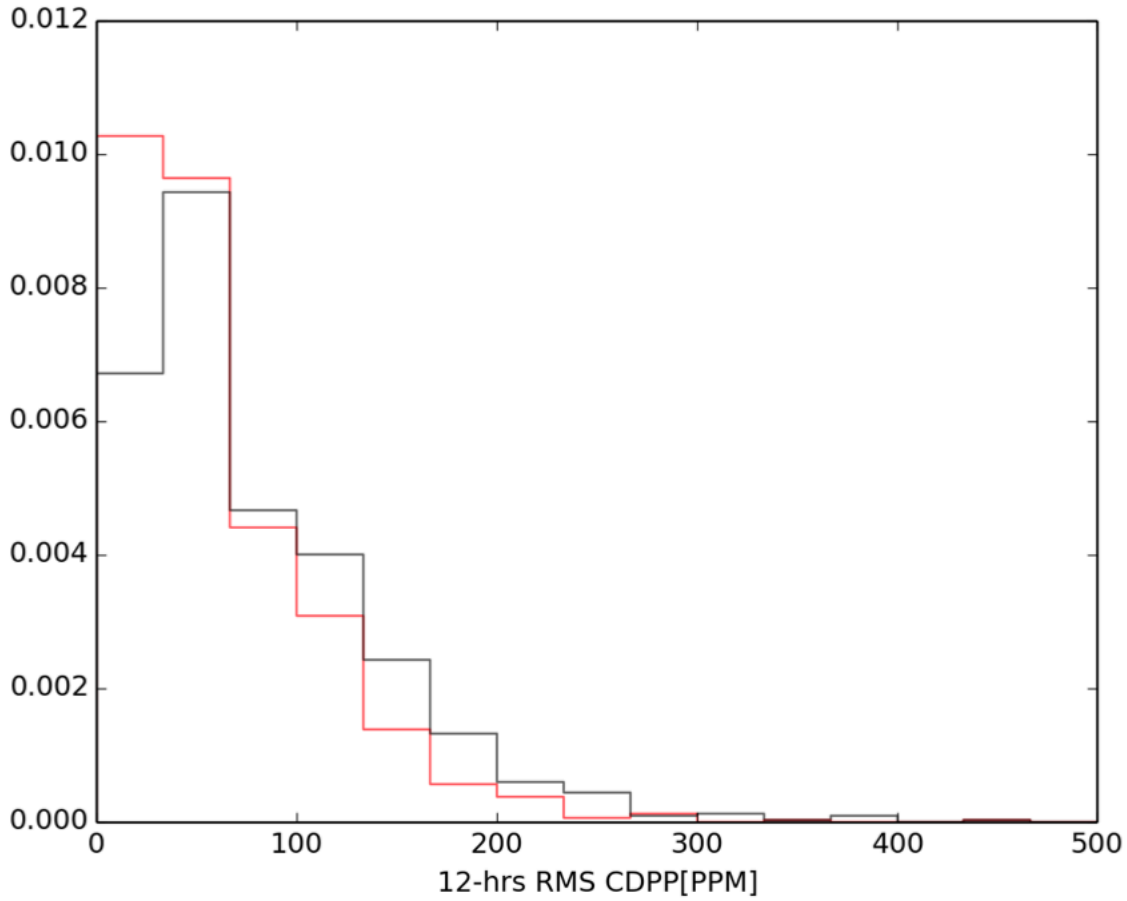
Idea 2: half-sibling regression



Idea 2: half-sibling regression



Idea 2: half-sibling regression



Schölkopf et al.: *Removing systematic errors for exoplanet search via latent causes*, ICML 2015

Schölkopf et al.: *Modeling Confounding by Half-Sibling Regression*, PNAS 2016