Statistical Learning: Causal-oriented and Robust

Peter Bühlmann

Seminar for Statistics, ETH Zürich

A big "Thank You" to the Organizers

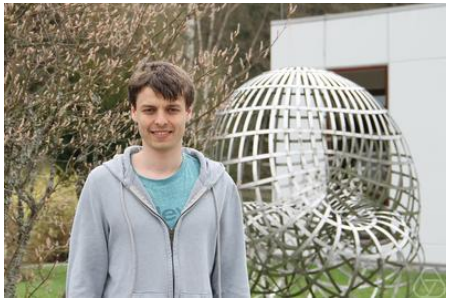
– for making this possible!

Supported in part by the European Research Council under the Grant Agreement

No. 786461 (CausalStats - ERC-2017-ADG)

Acknowledgments

Dominik Rothenhäusler
Stanford University



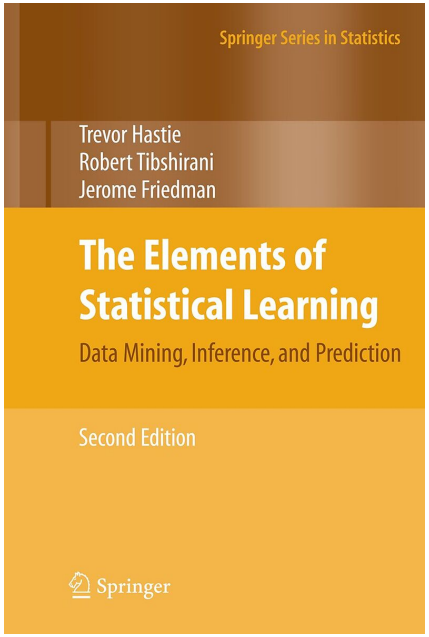
Niklas Pfister
Univ. Copenhagen

Yuansi Chen
Duke University

Jonas Peters
Univ. Copenhagen

Nicolai Meinshausen
ETH Zürich

Statistical Learning



Hastie, Tibshirani & Friedman (2000): have built a bridge between statistics and machine learning

the success of statistics and machine learning during the last 20 years is exceptional!

yet (and nothing in the book)...
... there are big open issues on

- stability, robustness
- generalizability and transferability to new populations
- insight and understanding in terms of “causality”
  (a “dangerous” word!)

$\rightsquigarrow$ want to highlight some aspects in making
statistical (& machine) learning
more “causal”-oriented and more robust

yet (and nothing in the book)...
... there are big open issues on

- stability, robustness
- generalizability and transferability to new populations
- insight and understanding in terms of “causality”
  (a “dangerous” word!)

$\rightsquigarrow$ want to highlight some aspects in making
statistical (& machine) learning
more “causal”-oriented and more robust

Statistical Machine Learning

multiple of terabytes of data...
and we want to extract something useful

- we want not just “black boxes”
- we want to understand and get new insight

interpretable Artificial Intelligence – in modern language!

Statistical Machine Learning

multiple of terabytes of data...
and we want to extract something useful

- we want not just “black boxes”
- we want to understand and get new insight

interpretable Machine Learning – in modern language!

Statistical Machine Learning

multiple of terabytes of data...
and we want to extract something useful

- we want not just “black boxes”
- we want to understand and get new insight

interpretable Statistical Inference – in modern context!

and as we will argue

better understanding exhibits increased robustness

no surprise in vague “scientifc folklore terms”
$\rightsquigarrow$ but needs to be mathematized!

and as we will argue

better understanding exhibits increased robustness

no surprise in vague “scientifc folklore terms”
$\rightsquigarrow$ but needs to be mathematized!

Setting the scene

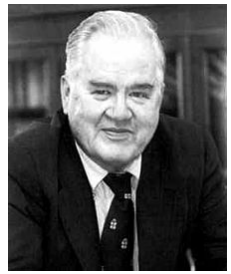
data as numbers alone are usually not informative

but: data as numbers in the context of (mathematical) models can be highly informative

the term “model” can be rather general (and e.g. infinite-dimensional)

Single data generating distribution & statist. inference

Mosteller and Tukey (1968): "One hallmark of the statistically conscious investigator is his firm belief that however the survey, experiment or observational program actually turned out, it could have turned out somewhat differently."

Mosteller Tukey

data (values) $z_1, z_2, \ldots, z_n$ are outcomes of random variables

$Z_1, Z_2, \ldots, Z_n$ $\underbrace{\text{i.i.d.}}_{\text{independent, identically distributed}}$ or stationary from distribution P_0

data-generating P_0 is from a possibly infinite-dimensional model (e.g. nonlinear regression or classification, from PDEs, etc.)

goal:

inference from data about (functional of) $\underbrace{P_0}_{\text{data-generating distr.}}$

Graunt & Petty (1662): first life table

Arbuthnot (1710), Bayes (1761), Laplace (1774), Gauss (1795, 1801, 1809), Quetelet (1796-1874),..., Karl Pearson (1857-1936), Fisher (1890-1962), Egon Pearson (1895-1980), Neyman (1894-1981), ...

$\rightsquigarrow$ most of this is about generalization to new data from the same distribution P_0 as the observed "training" data

e.g:

new patient w. "same characteristics" as in representative study

Generalization to new data generating distributions

classical framework does not allow to generalize
beyond the data-generating distribution P_0

setting:
observed data from distribution P_0

want to say something about new $P' \neq P_0$

transfer learning (Bozinovskio & Fulgosi, 1976; Pratt, 1993;...)
domain adaptation (Bridle & Cox, 1990; Crammer, Kearns & Wortmann, 2008;...)
transportability (causal) (Pearl & Bareinboim, 2011;...)

Generalization to new data generating distributions

classical framework does not allow to generalize
beyond the data-generating distribution P_0

setting:
observed data from distribution P_0

want to say something about new $P' \neq P_0$

transfer learning (Bozinovskio & Fulgosi, 1976; Pratt, 1993;...)
domain adaptation (Bridle & Cox, 1990; Crammer, Kearns & Wortmann, 2008;...)
transportability (causal) (Pearl & Bareinboim, 2011;...)

Generalization to new data generating distributions

a more general and often realistic setting:

heterogeneous data from various distributions P_e ($e \in \mathcal{E}$)
$\mathcal{E} = \{$ observed environments, scenarios, sub-populations,...,
..., "sources"$\}$

want to say something about new $P_{e'}$ ($e' \notin \mathcal{E}$)

Very many examples

multi-center study (e.g. of COVID-19 vaccine) with diverse subpopulations: want to generalize to

new different (outside study) subpopulations

Dahabreh, Petito, Robertson, Hernan, Steingrimsson (2000):

> *Instead, for decision-making, the users typically have a new target population in mind.*

trained on designed scenarios from $\mathcal{E}$

trained on designed scenarios from $\mathcal{E}$

new scenario $e' \notin \mathcal{E}$

if one ignores the generalization problem to new $P' \neq P_0$:

in many realistic applications, machine learning algorithms will have poor performance!

- adversarial attacks in autonomous systems
- personalized medicine for "new but somewhat different" individuals (when predicting disease outcome, drug efficacy, ...)
- ...

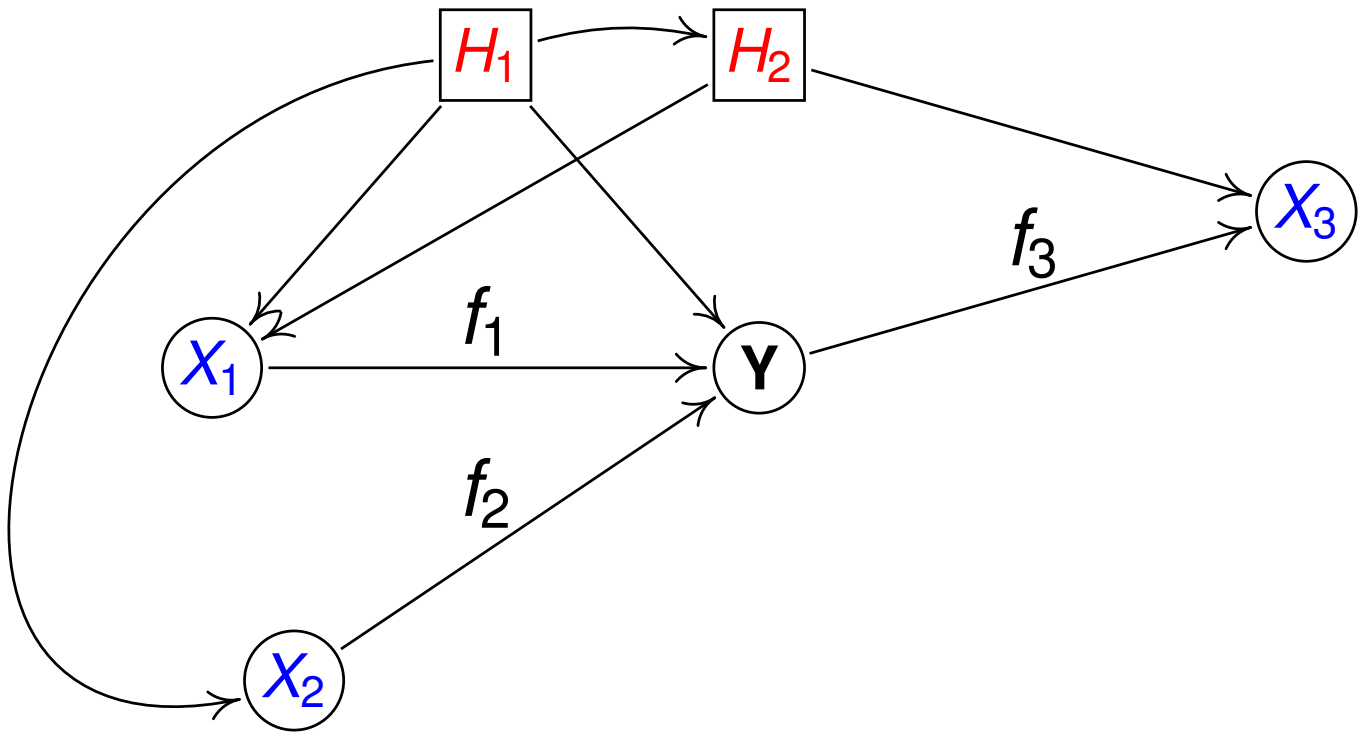
while standard machine learning often performs very well on new data from the same distribution P_0 used for training

- image classification for same type of images
- speech recognition for same/similar person
- ...

Model for entire system and perturbations

- a response/target variable (phenotype) Y of interest
- explanatory variables/features/covariates X
- many non-observed hidden variables H

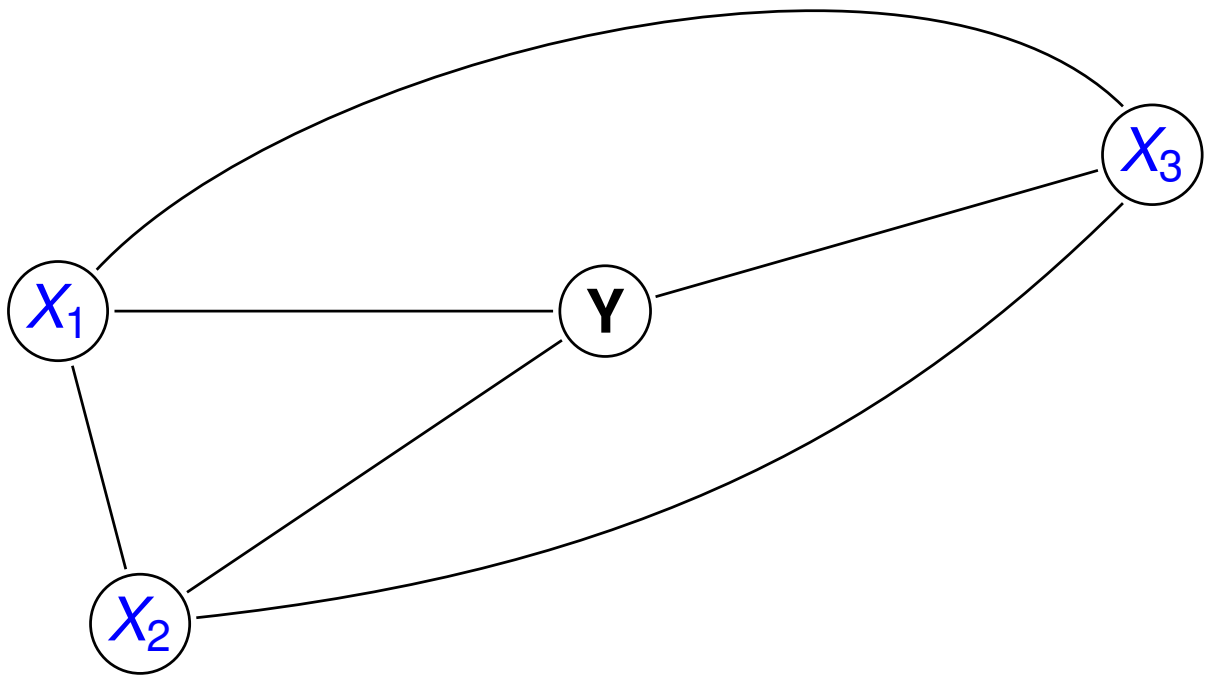
the unrealistic oracle world: entire system is known



Model for entire system and perturbations

- a response/target variable (phenotype) Y of interest
- explanatory variables/features/covariates X
- many non-observed hidden variables H

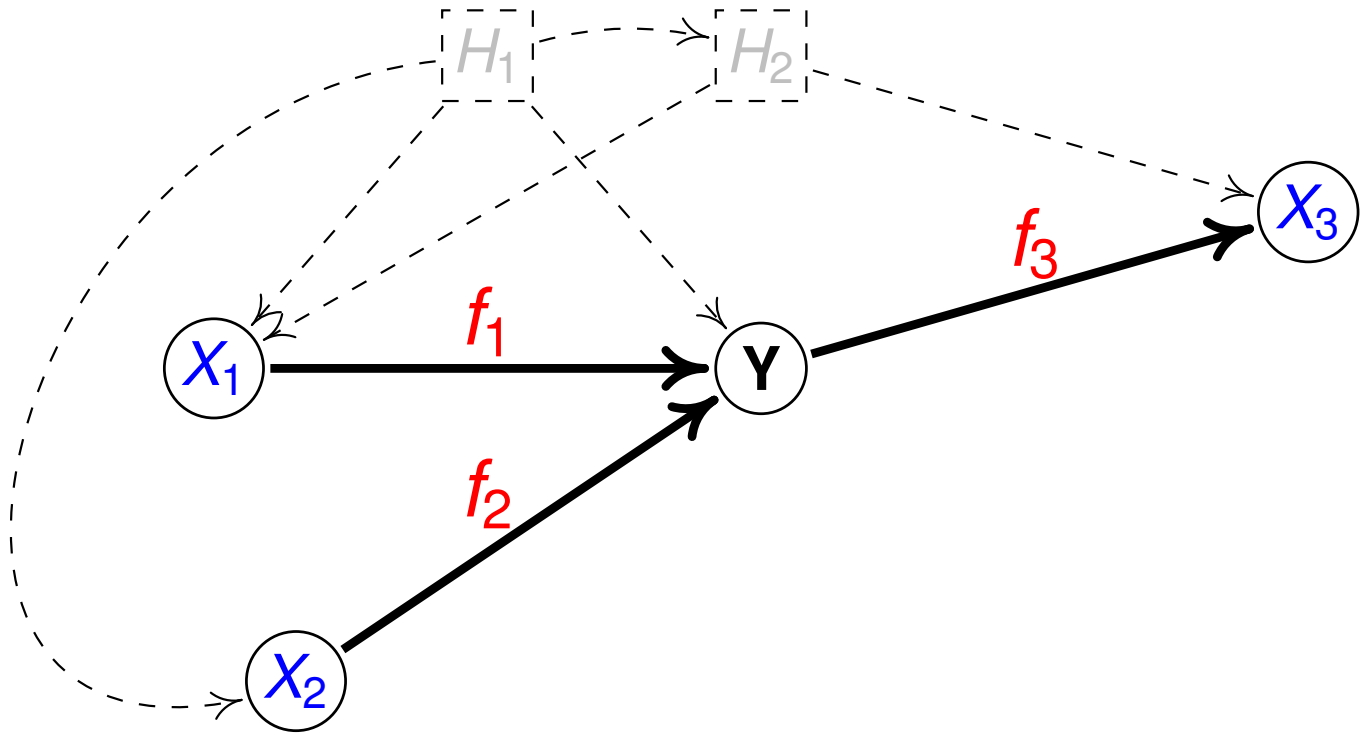
observed: regression association among observed variables
too many relations, no directionality



Model for entire system and perturbations

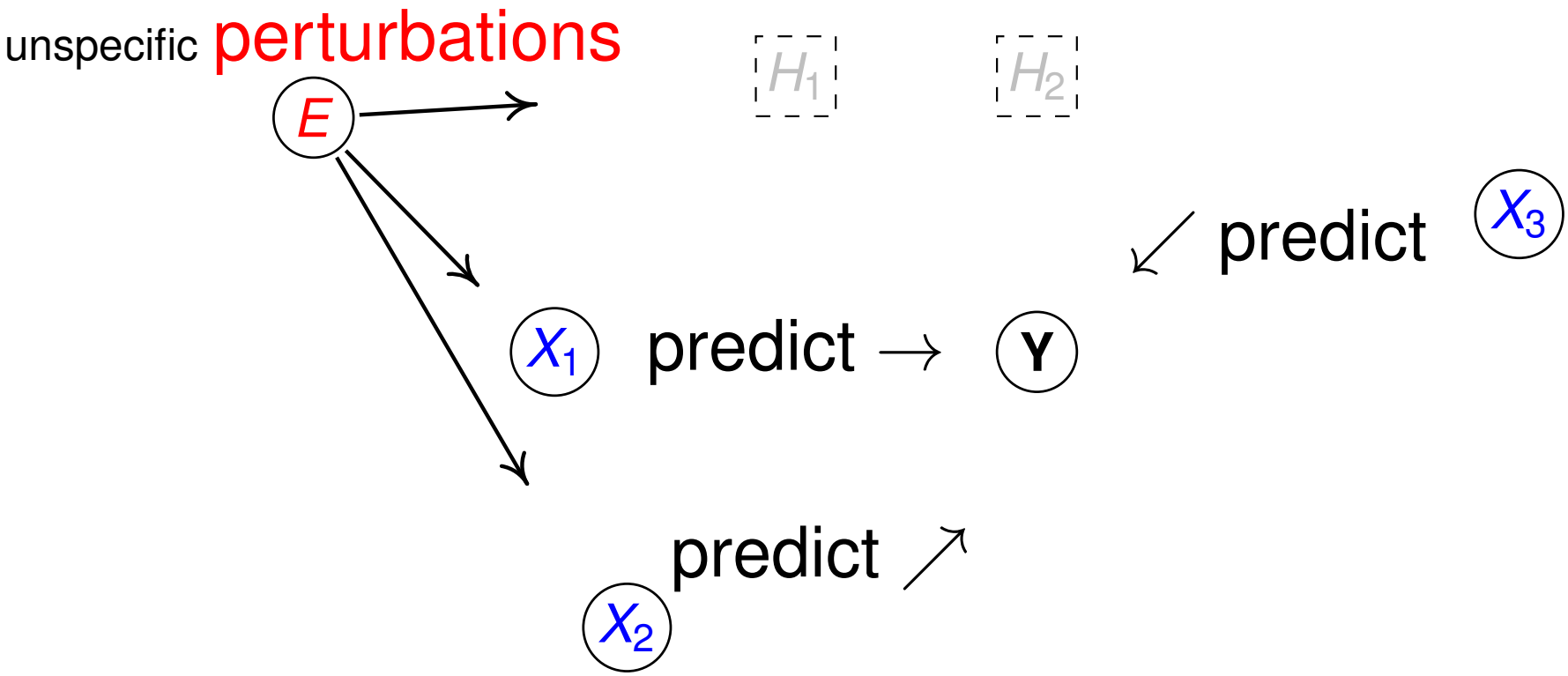
aim (1): infer true relations (& directionality) between X_j's and Y
$\rightsquigarrow$ very ambitious! (considered as a main task of "causality")

cannot be solved by regression or "standard machine learning"!



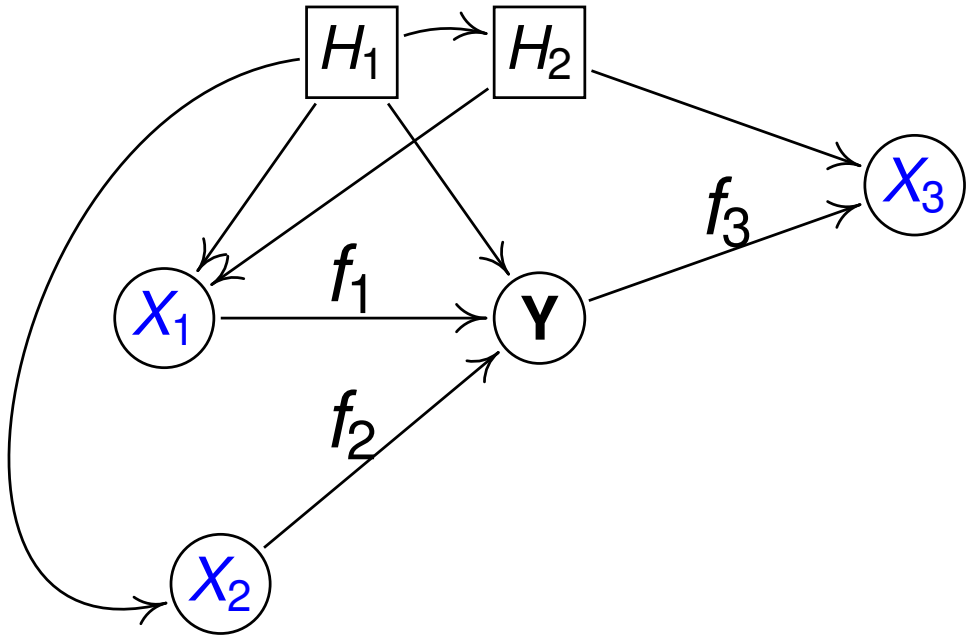
Model for entire system and perturbations

aim (2): robust prediction of Y from X_j's under perturbations/interventions/subpopulations to the system



Structural Equation Models (SEMs)

given a structure (DAG = Directed Acyclic Graph)



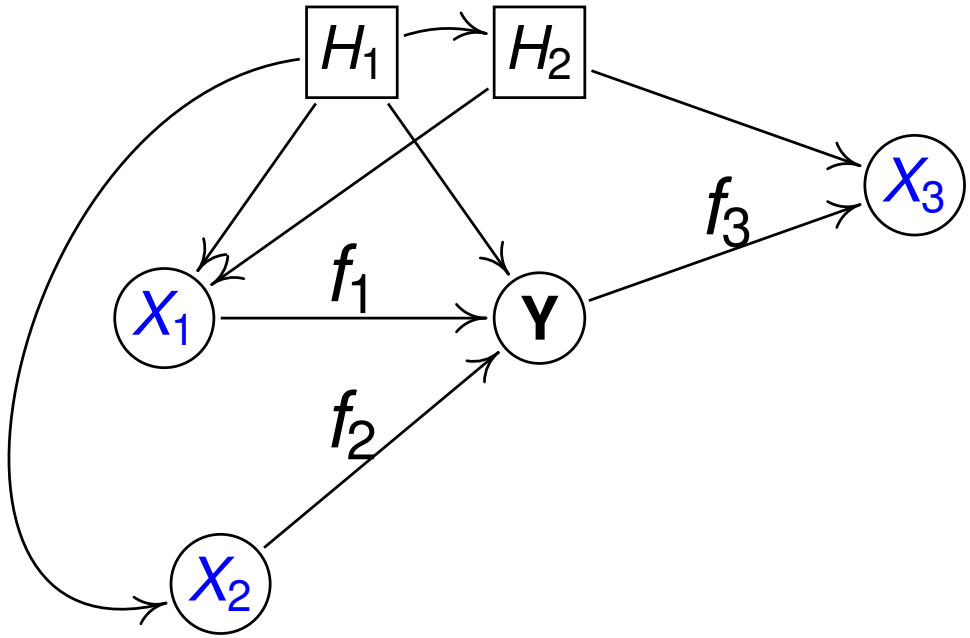
specify function models for every random variable:
for $W = (Y, X_1, \ldots, X_p, H_1, \ldots, H_q)$ we have

$$W_j \leftarrow f_j(W_{\mathrm{pa}(j)}, \varepsilon_j) \quad , j = 1, \ldots, p+q+1,$$

$\varepsilon_1, \ldots \varepsilon_{p+q+1}$ jointly independent

Structural Causal Models (SCMs) $\rightsquigarrow$ Judea Pearl

given a structure (DAG = Directed Acyclic Graph)



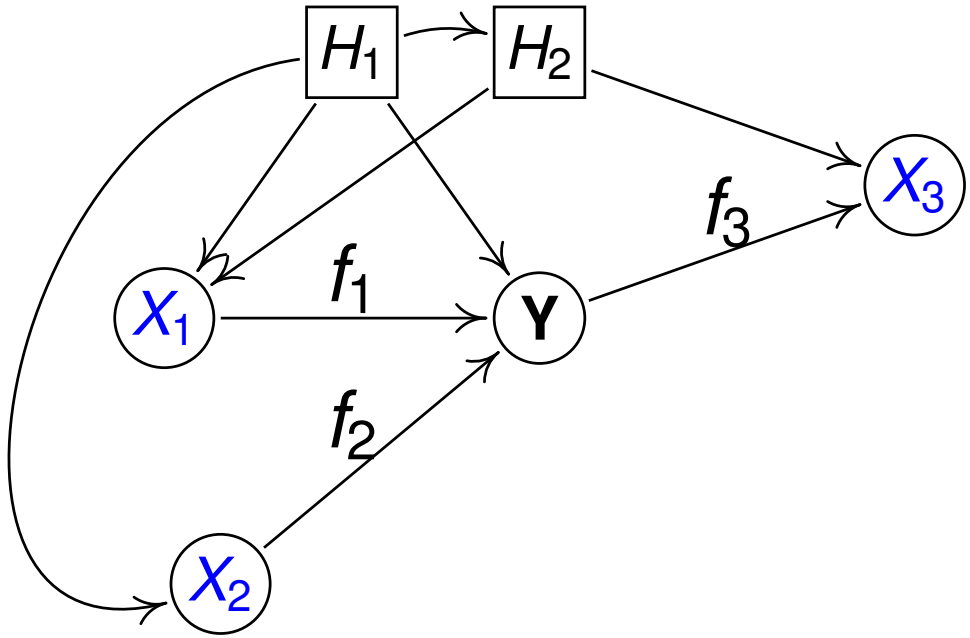
specify function models for every random variable:
for $W = (Y, X_1, \ldots, X_p, H_1, \ldots, H_q)$ we have

$$W_j \leftarrow f_j(W_{\mathrm{pa}(j)}, \varepsilon_j) \quad , j = 1, \ldots, p+q+1,$$

$\varepsilon_1, \ldots \varepsilon_{p+q+1}$ jointly independent

Structural Equation Models (SEMs)

given a structure (DAG = Directed Acyclic Graph)



specify function models for every random variable:
for $W = (Y, X_1, \ldots, X_p, H_1, \ldots, H_q)$ we have

$$W_j \leftarrow f_j(W_{\mathrm{pa}(j)}, \varepsilon_j) \quad , j = 1, \ldots, p+q+1,$$

$\varepsilon_1, \ldots \varepsilon_{p+q+1}$ jointly independent

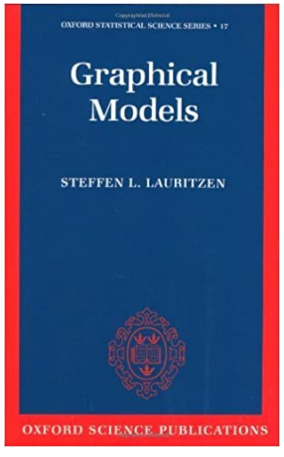
function models for every random variable:
for $W = (Y, X_1, \ldots, X_p, H_1, \ldots, H_q)$ we have

$$W_j \leftarrow f_j(W_{\mathrm{pa}(j)}, \varepsilon_j) \quad , j = 1, \ldots, p+q+1,$$

$\varepsilon_1, \ldots \varepsilon_{p+q+1}$ jointly independent

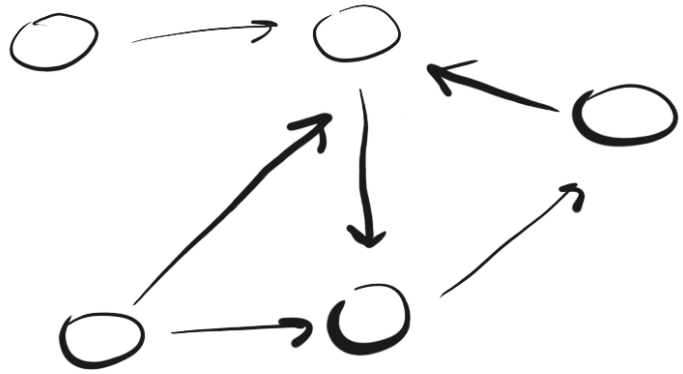
the joint distribution P of W satisfies the (local and global) Markov property w.r.t. graph (structure) of the SEM

$\rightsquigarrow$ see e.g. Lauritzen (1996)



important remark: learning the structure/DAG from observational data

P_0



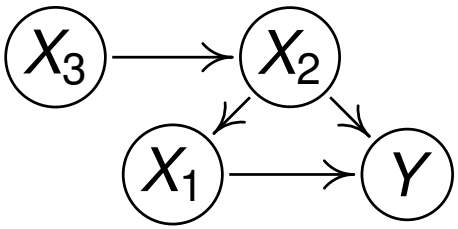
it's impossible in general without additional assumptions!

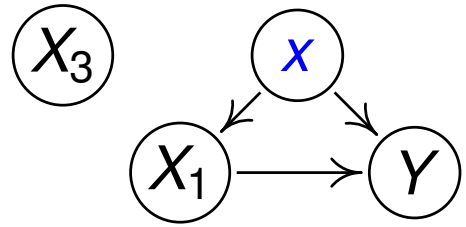
can (only) estimate a Markov equivalence class of graphs

Verma & Pearl (1990); Sprites, Glymour & Scheines (1993); Lauritzen (1996); Chickering (2002); Kalisch & PB (2007); ...

that is:
given data-generating distribution P_0, several DAG structures and corresponding function models generate P_0 in the SEM
$\rightsquigarrow$ learning directions ("causal relations") is troublesome!

Models for perturbations based on structural equations: Pearl's do-intervention as an example





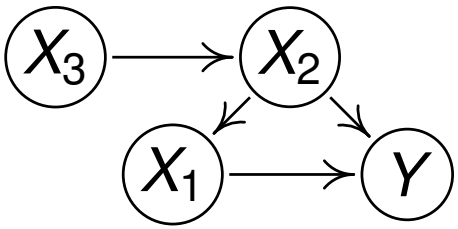
$$\mathrm{do}(X_2 = x)$$

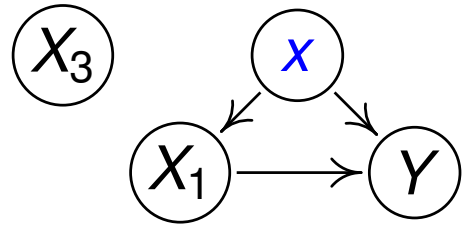
$$\rightsquigarrow$$

$$X_3 \leftarrow \varepsilon_3$$
$$X_2 \leftarrow f_2(X_3, \varepsilon_2)$$
$$X_1 \leftarrow f_1(X_2, \varepsilon_1)$$
$$Y \leftarrow f_Y(X_1, X_2, \varepsilon_Y)$$

$$X_3 \leftarrow \varepsilon_3$$
$$X_2 \leftarrow x$$
$$X_{1;x} \leftarrow f_1(X_2 = x, \varepsilon_1)$$
$$Y_x \leftarrow f_Y(X_{1,x}, X_2 = x, \varepsilon_Y)$$

Models for perturbations based on structural equations: Pearl's do-intervention as an example





$$\mathrm{do}(X_2 = x)$$
$$\rightsquigarrow$$

$$X_3 \leftarrow \varepsilon_3$$
$$X_2 \leftarrow f_2(X_3, \varepsilon_2)$$
$$X_1 \leftarrow f_1(X_2, \varepsilon_1)$$
$$Y \leftarrow f_Y(X_1, X_2, \varepsilon_Y)$$

$$X_3 \leftarrow \varepsilon_3$$
$$X_2 \leftarrow x$$
$$X_{1;x} \leftarrow f_1(X_2 = x, \varepsilon_1)$$
$$Y_x \leftarrow f_Y(X_{1,x}, X_2 = x, \varepsilon_Y)$$

dynamic propagation of interventions
same functions in structural equations

many other perturbation models based on structural equations
e.g. additive shift

all of them exhibiting

- dynamic propagation of interventions
- “parts of” the structural equations remain the same:
  autonomy assumption (which is a “huge” assumption...)

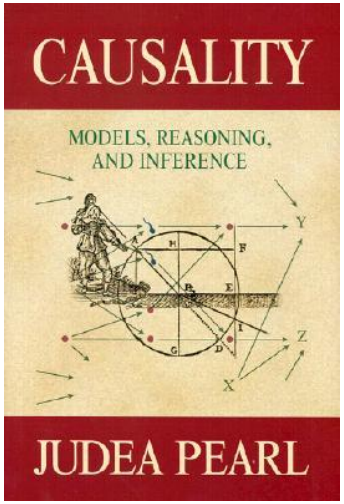
“Causal thinking” for generalization

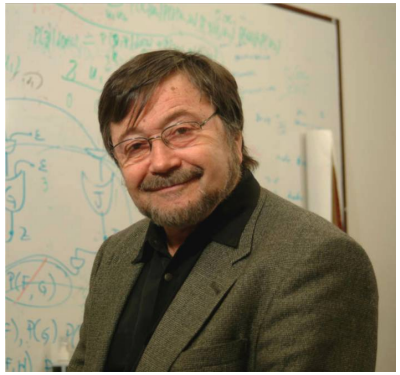
with perturbation models at hand

causality: is giving a prediction (a quantitative answer) to a “what if I do/perturb” question

but that perturbation (aka “new situation”) is not observed

see for example

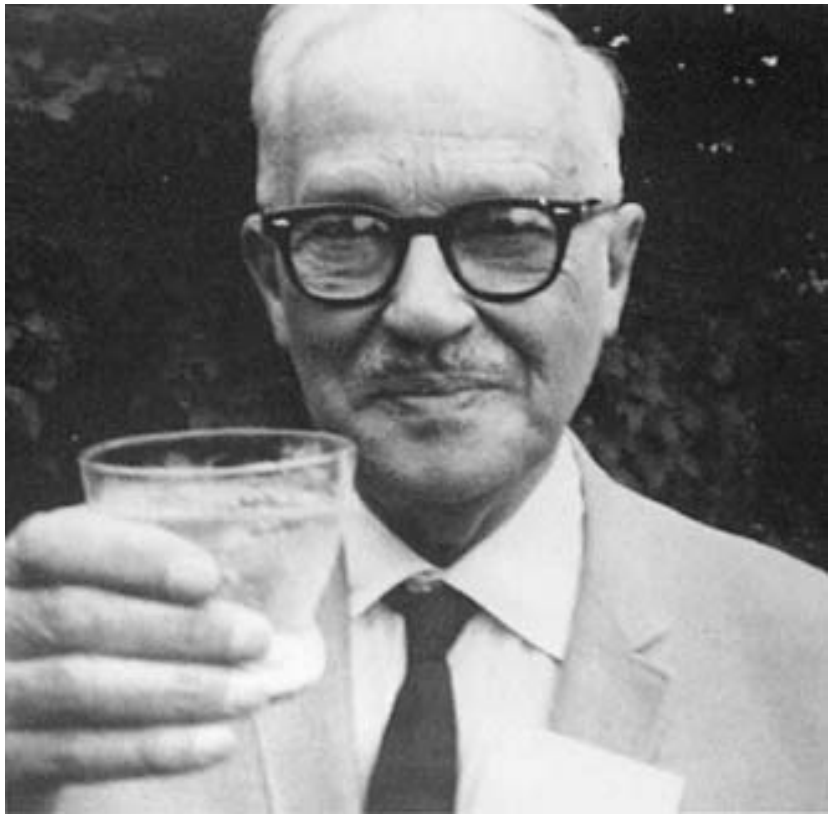




Judea Pearl

Turing Award, 2011

Neyman's potential outcome model: in his 1923 master thesis!



Jerzy Neyman

potential outcome: what would have happened if we would have assigned a certain treatment

Neyman's potential outcome model: a milestone!

many modern applications are faced with such prediction tasks:

- genomics: what would be the effect of knocking down (the activity of) a gene on the growth rate of a plant?

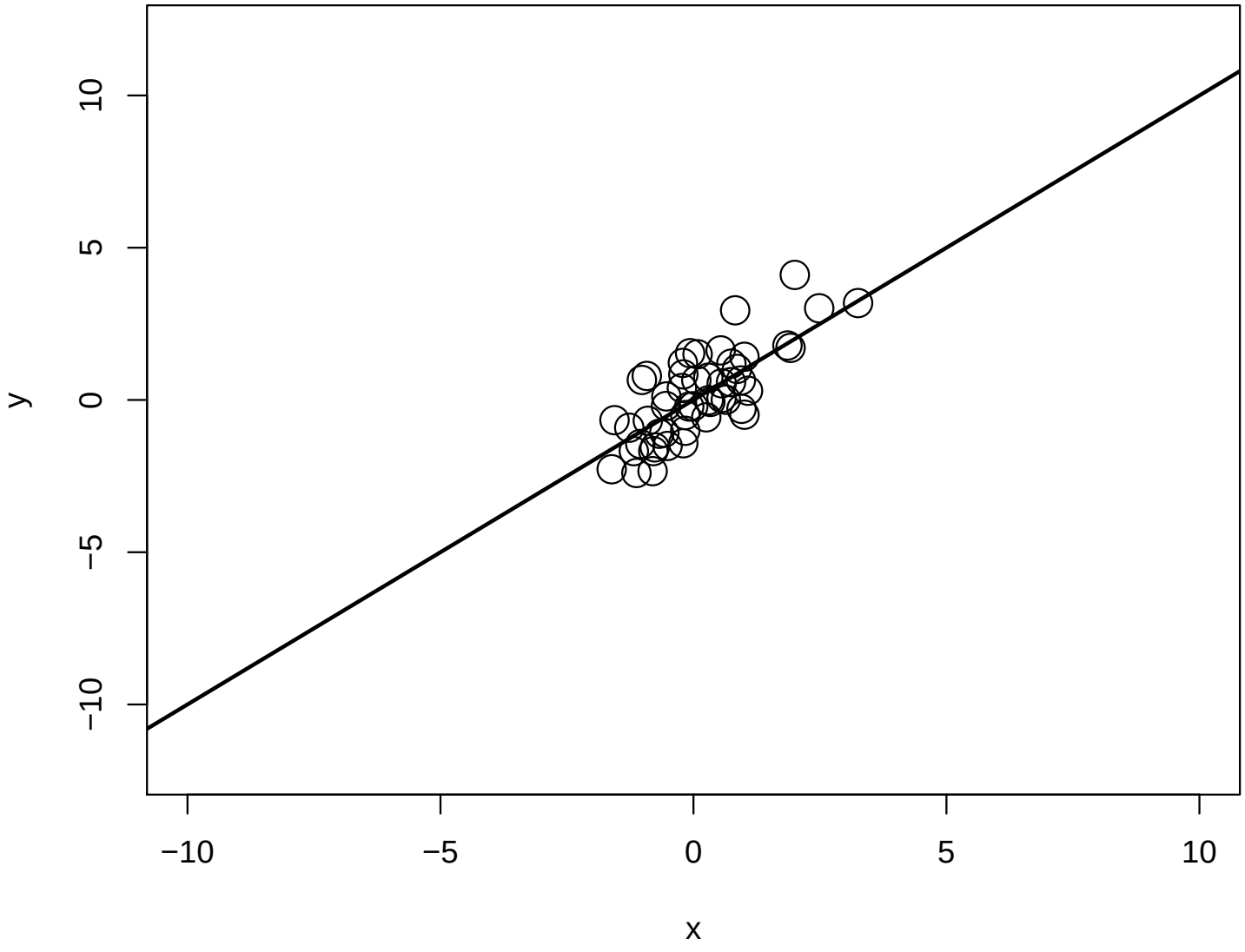
  we want to predict this without any data on such a gene knock-out (e.g. no data for this particular perturbation)

- advertising in E-commerce
- policy making
- algorithmic fairness
- ...

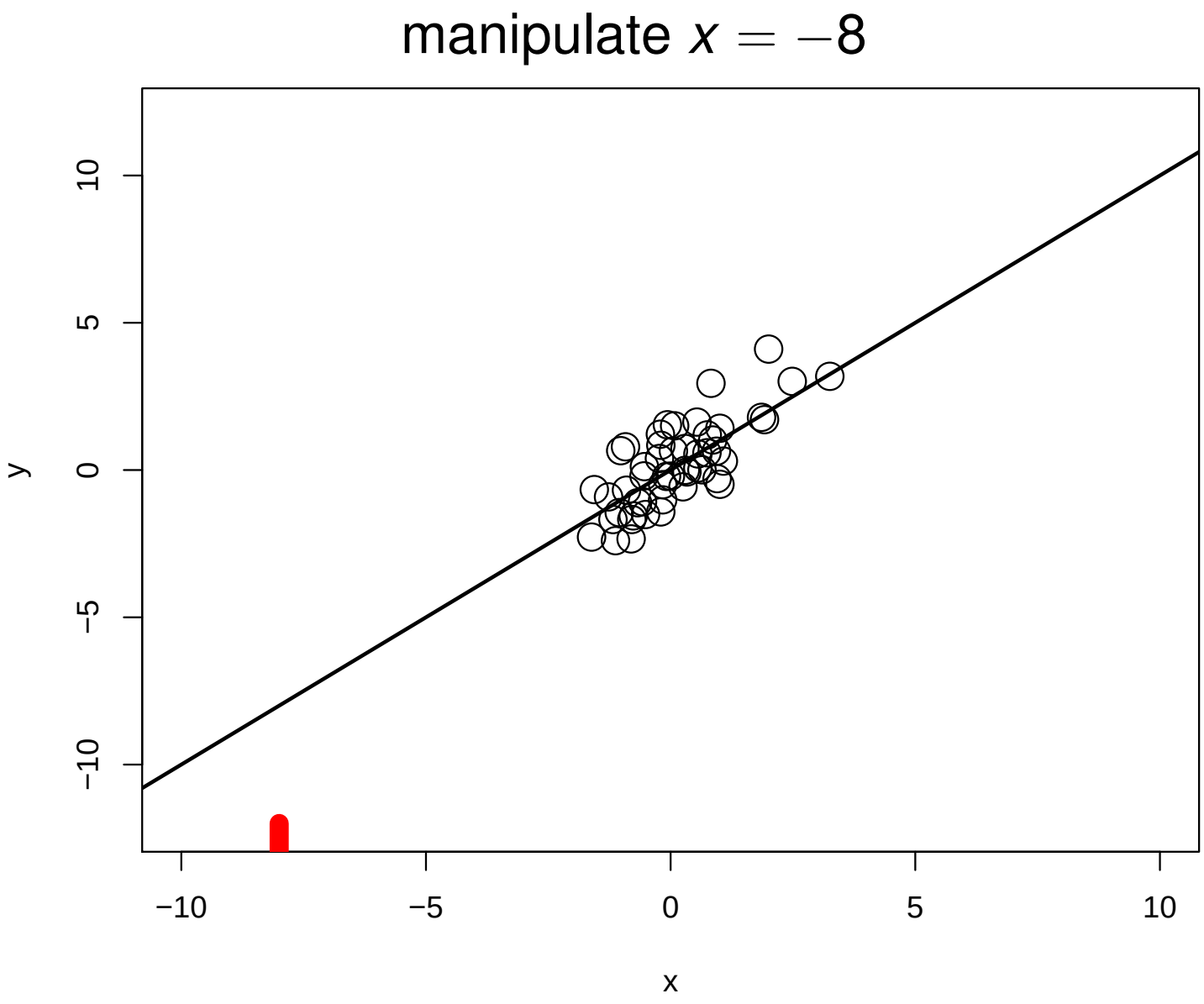
and therefore:
$\rightsquigarrow$ causality can be used to generalize to new scenarios

what would happen for a new patient who is different
(a “perturbed version”) from the ones in the study?

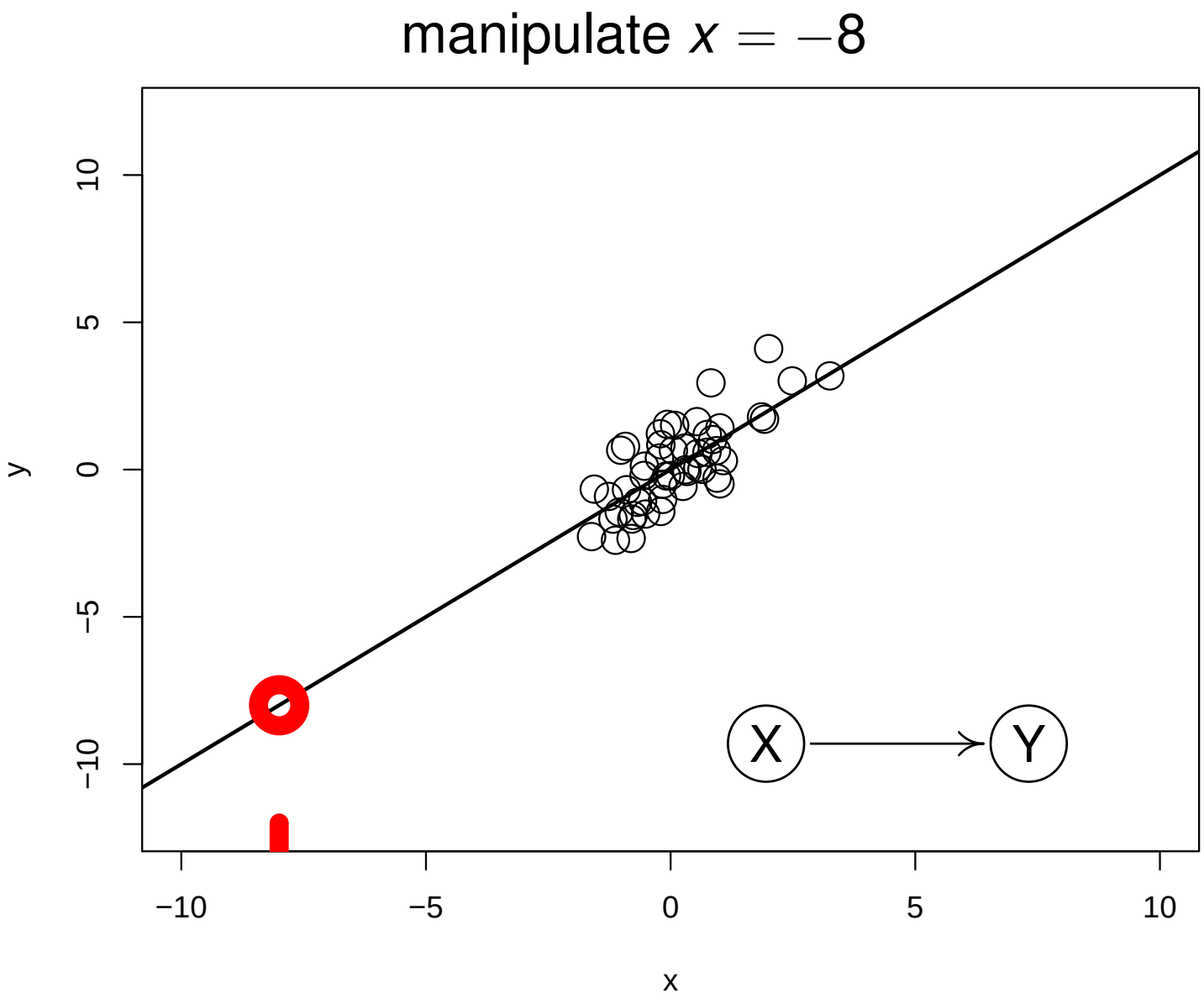
Predicting an intervention effect (synthetic data)



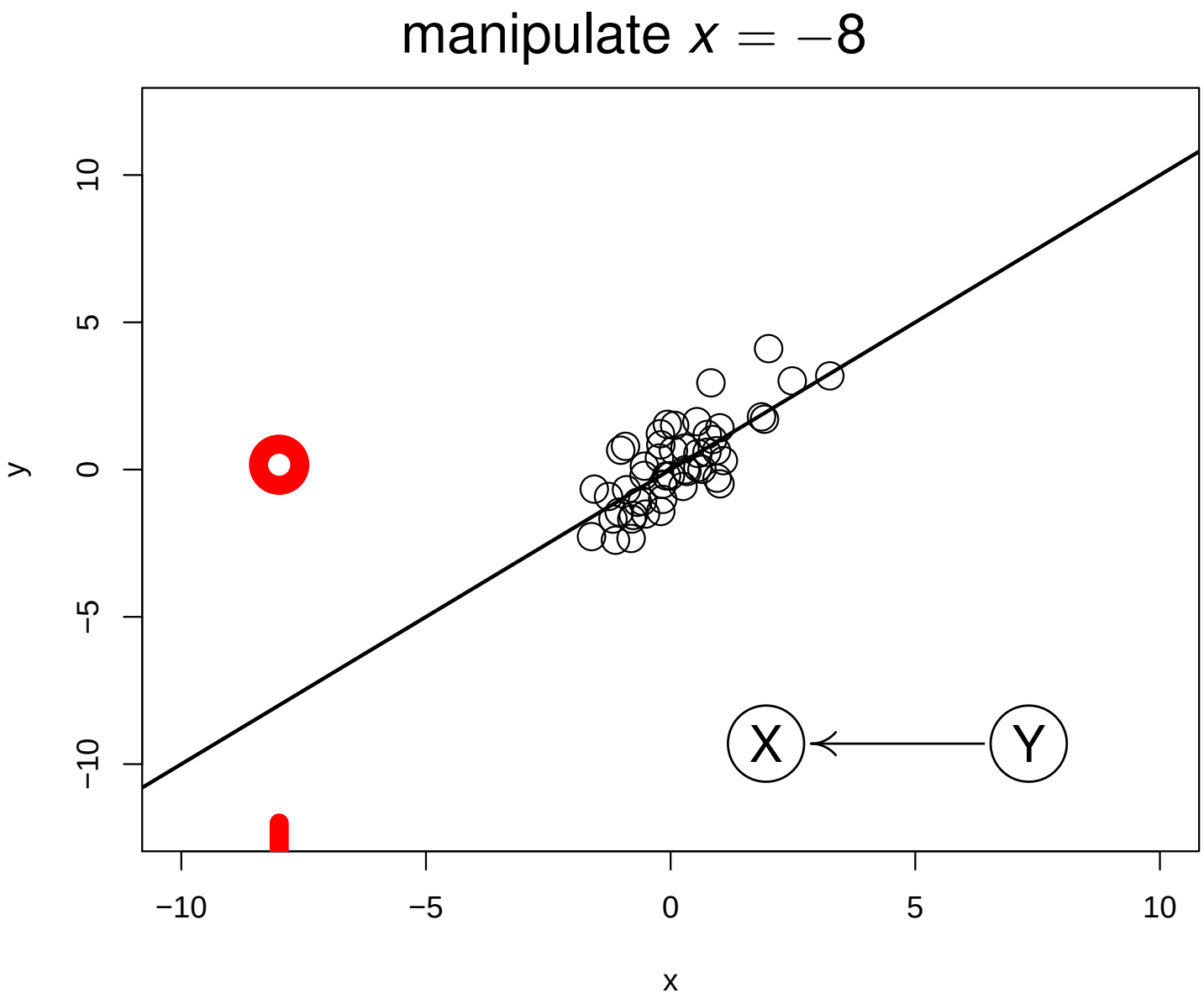
Predicting an intervention effect (synthetic data)



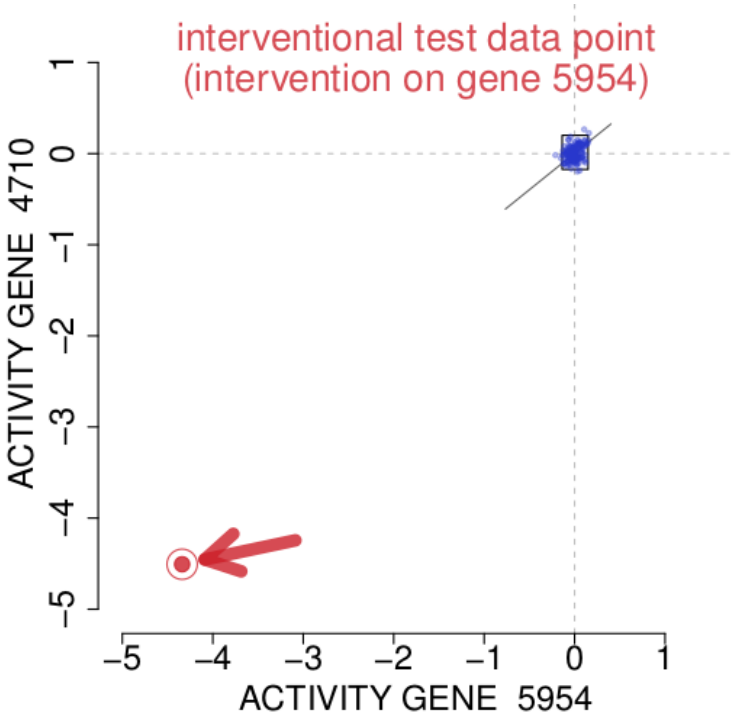
Predicting an intervention effect (synthetic data)

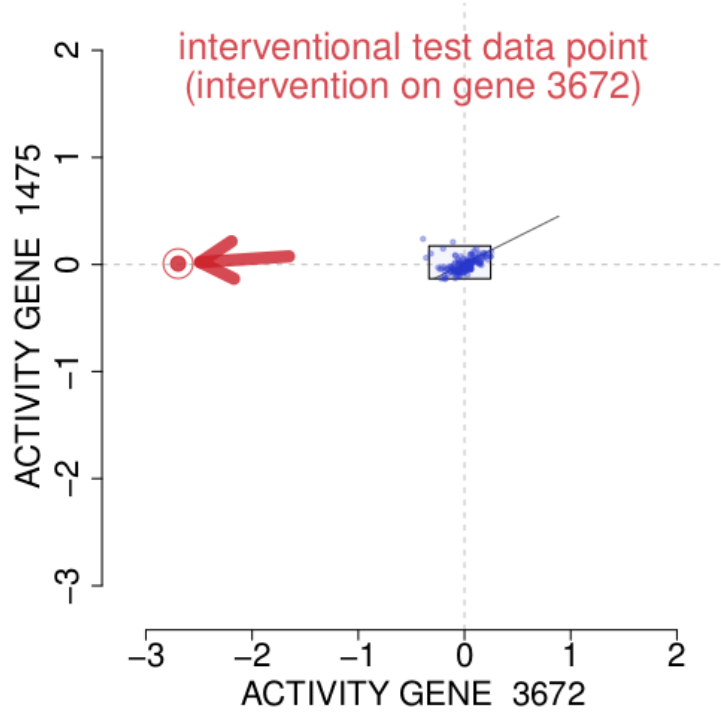


Predicting an intervention effect (synthetic data)



Predicting an intervention effect: real gene expression data

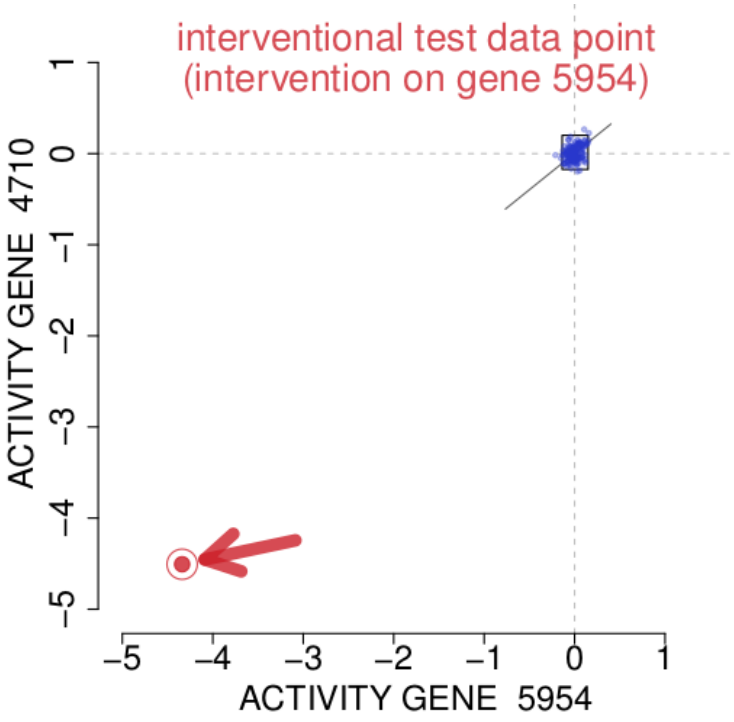


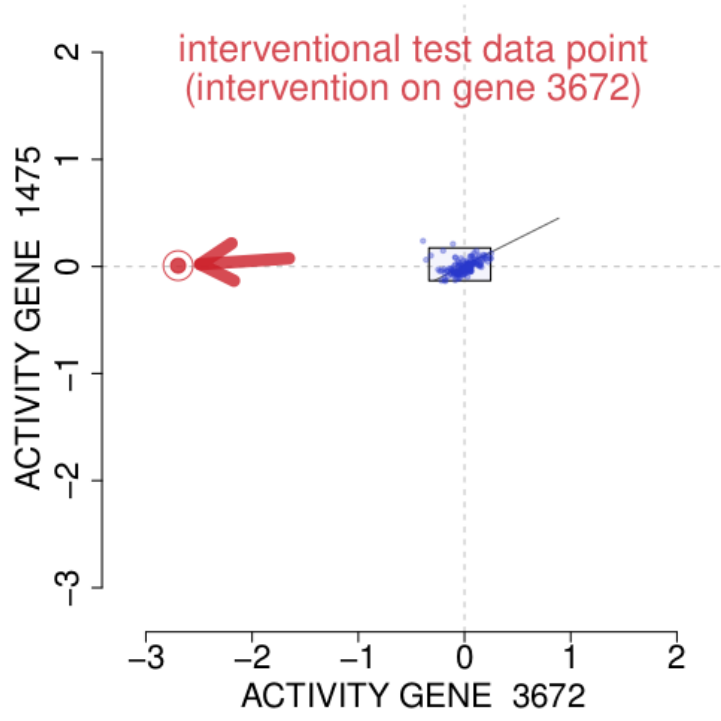


Challenge:

how to predict?

Predicting an intervention effect: real gene expression data

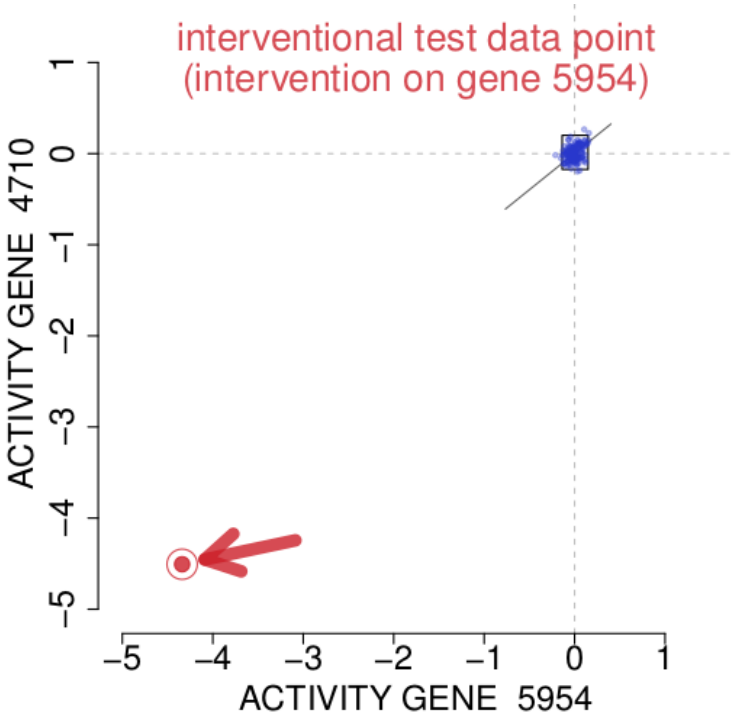


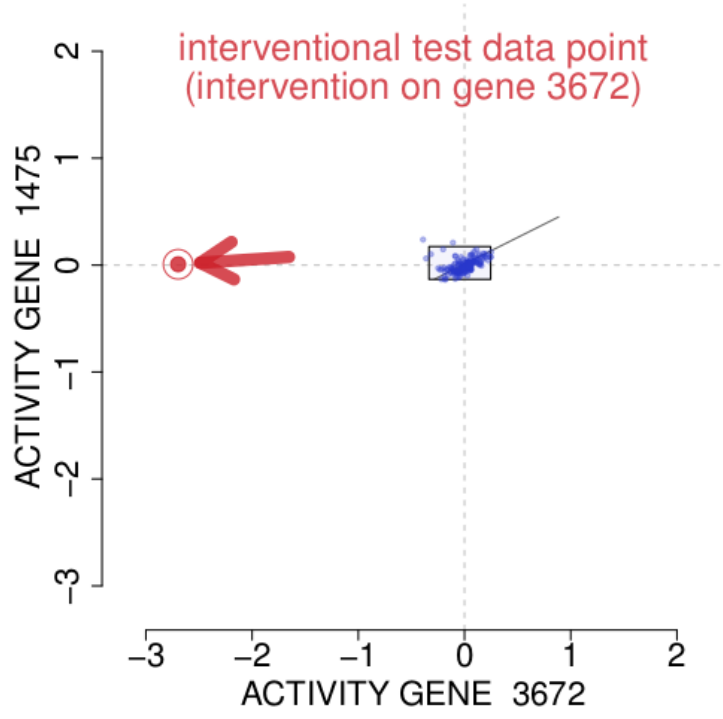


Challenge:

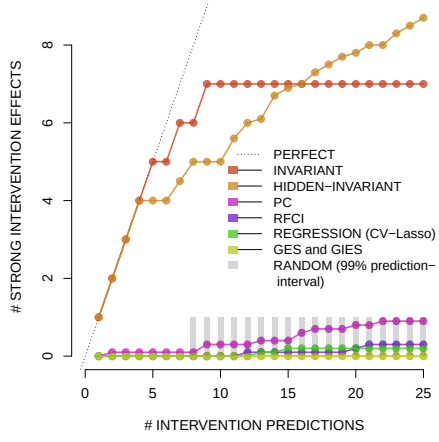
how to predict?

Predicting an intervention effect: real gene expression data





results (see later):



The word “causality”

the term “causality” is perhaps overly ambitious...

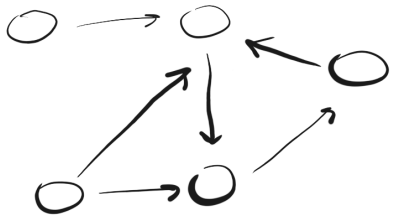
it’s really about predicting intervention effects:

what would happen if you would perturb/assign/do ...?

and you have never seen the

perturbation/assignment/do-action

A short intermediate “summary”

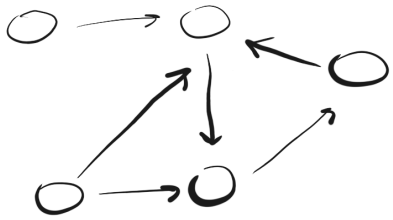


causal SEM model

+ intervention model

$\Longrightarrow$ prediction of interventions

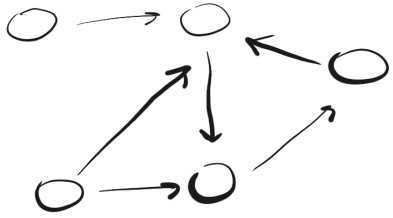
A short “summary”



causal SEM model

+ intervention model $\Longrightarrow$ prediction of interventions

what we want:



inferring
causal system &
enhancing robustness

$\Longleftarrow$ observing unspecific interventions from $\mathcal{E}$

Heterogeneity, Robustness and a bit of causality

assume heterogeneous data from different known observed environments or experimental conditions or perturbations or sub-populations $e \in \mathcal{E}$:

$$(X^e, Y^e) \sim P_e, \quad e \in \mathcal{E}$$

with response variables $\underbrace{Y^e}_{\text{target}}$ and $\underbrace{\text{predictor variables } X^e}_{\text{features}}$

examples:

- data from 10 different countries
- medical data from 13 different centers/hospitals
-  large-scale data applications

consider “many possible” but mostly non-observed environments/perturbations $\mathcal{F} \supset \underbrace{\mathcal{E}}_{\text{observed}}$

examples for $\mathcal{F}$:

- 10 countries and many other than the 10 countries
- 13 centers/hospitals and many new ones

problem:

predict Y given X such that the prediction works well (is “robust”/“replicable”) for *“many possible”* new environments $e \in \mathcal{F}$ based on data from much fewer environments from $\mathcal{E}$

a pragmatic prediction problem:

predict Y given X such that the prediction works well (is "robust"/"replicable") for *"many possible"* environments $e \in \mathcal{F}$ based on data from much fewer environments from $\mathcal{E}$

for example with linear models: find

$$\text{argmin}_{\beta \in \mathbb{R}^d} \max_{e \in \mathcal{F}} \mathbb{E}|Y^e - X^e \beta|^2$$

it is "robustness"

a pragmatic prediction problem:

predict Y given X such that the prediction works well (is "robust"/"replicable") for *"many possible"* environments $e \in \mathcal{F}$ based on data from much fewer environments from $\mathcal{E}$

for example with linear models: find

$$\text{argmin}_{\beta \in \mathbb{R}^d} \max_{e \in \mathcal{F}} \mathbb{E}|Y^e - X^e \beta|^2$$

it is "robustness"

a pragmatic prediction problem:

predict Y given X such that the prediction works well (is "robust"/"replicable") for *"many possible"* environments $e \in \mathcal{F}$ based on data from much fewer environments from $\mathcal{E}$

for example with linear models: find

$$\text{argmin}_{\beta \in \mathbb{R}^d} \max_{e \in \mathcal{F}} \mathbb{E}|Y^e - X^e \beta|^2$$

it is "robustness" and causality

Causality and worst case risk over perturbations

for linear models: in a nutshell

$$\text{argmin}_{\beta \in \mathbb{R}^d} \max_{e \in \mathcal{F}} \mathbb{E}|Y^e - X^e \beta|^2 = \text{ causal parameter}$$

where $\mathcal{F} = \ldots$

Haavelmo (1943); Peters, PB & Meinshausen (2016); Rojas-Carulla, Schölkopf, Turner & Peters (2018); Arjovsky, Bottou, Gulrajani, Lopez-Paz (2019), PB (2020), Rothenhäusler, PB, Peters & Meinshausen (2021), ...

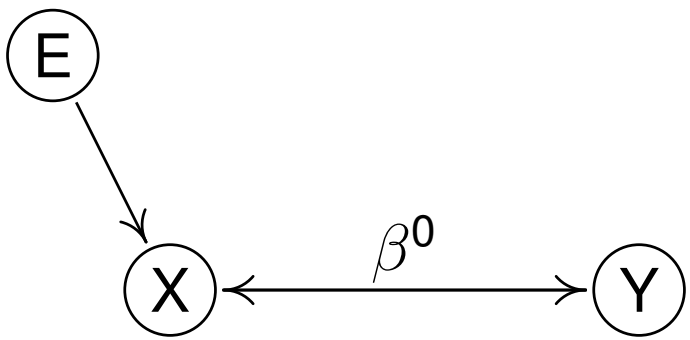
risk optimization $\longleftrightarrow$ causality

Causality and worst case risk over perturbations

for linear models: in a nutshell

for $\mathcal{F} = \{$all perturbations not acting on Y directly (but dynamically propagated)$\}$,

$\text{argmin}_{\beta \in \mathbb{R}^d} \max_{e \in \mathcal{F}} \mathbb{E}|Y^e - X^e\beta|^2 =$ causal parameter $= \beta^0$



β^0 the causal ("system") parameter"; in linear SEMs:

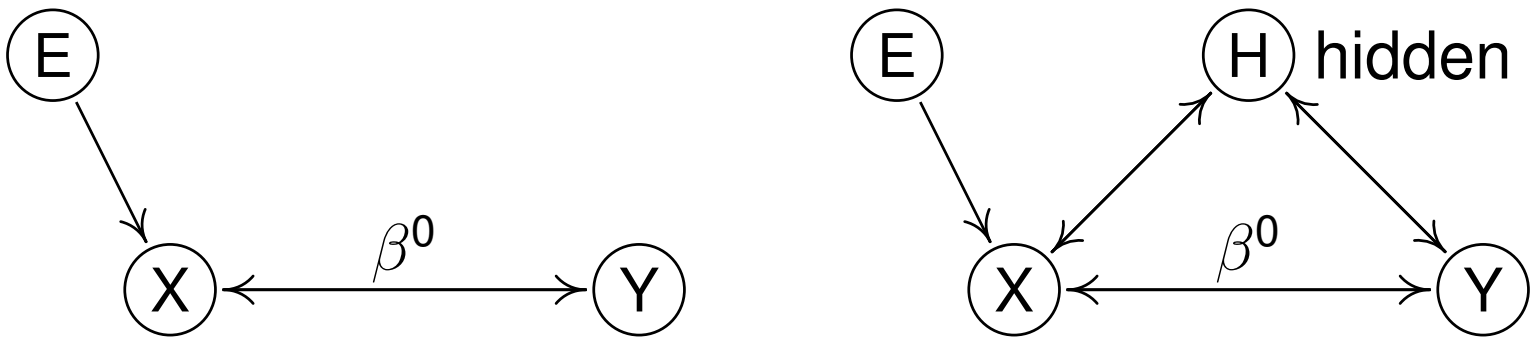
$$Y \leftarrow \sum_{j \in \text{pa}(Y)} \beta_j^0 X_j + \varepsilon_Y, \text{ causal parameter} := \{\beta_j^0;\ j \in \text{pa}(Y)\}$$

Causality and worst case risk over perturbations

for linear models: in a nutshell

for $\mathcal{F} = \{$all perturbations not acting on Y directly (but dynamically propagated)$\}$,

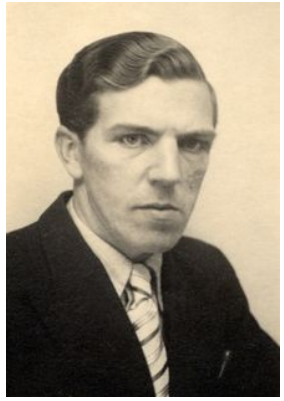
$\text{argmin}_{\beta \in \mathbb{R}^d} \max_{e \in \mathcal{F}} \mathbb{E}|Y^e - X^e\beta|^2 =$ causal parameter $= \beta^0$



β^0 the causal ("system") parameter"; in linear SEMs:

$$Y \leftarrow \sum_{j \in \text{pa}_X(Y)} \beta_j^0 X_j + \sum_{r \in \text{pa}_H(Y)} \gamma_r H_r + \varepsilon_Y, \text{ caus. par.} := \{\beta_j^0;\ j \in \text{pa}_X(Y)\}$$

causality and distributional robustness are intrinsically related (Haavelmo, 1943)



Trygve Haavelmo, Nobel Prize in Economics 1989

in fact: Haavelmo (1943) derived an invariance principle:

$$\mathcal{L}(Y^e - X^e \beta^0_{\text{causal}}) \text{ invariant across all } e \in \mathcal{F}$$

that is: causal $\implies$ invariance

Haavelmo (1943): causal $\Longrightarrow$ invariance

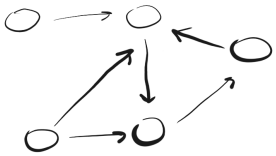
and we advocate to "invert" (Peters, PB & Meinshasuen, 2016):

causal $\Longleftarrow$ invariance

search (by statistical testing) for β^* such that

$\mathcal{L}(Y^e - X^e\beta^*)$ invariant across observed $e \in \mathcal{E}$

$\rightsquigarrow \ \beta^* = \beta^0_{\text{causal}}$ if observed $\mathcal{E}$ is sufficiently rich



inferring
causal system &
enhancing robustness

invariance $\Longleftarrow$

observing
unspecific
interventions
from $\mathcal{E}$

this “inverted” process: Invariant Causal Prediction (ICP)

(Peters, PB & Meinshausen, 2016)

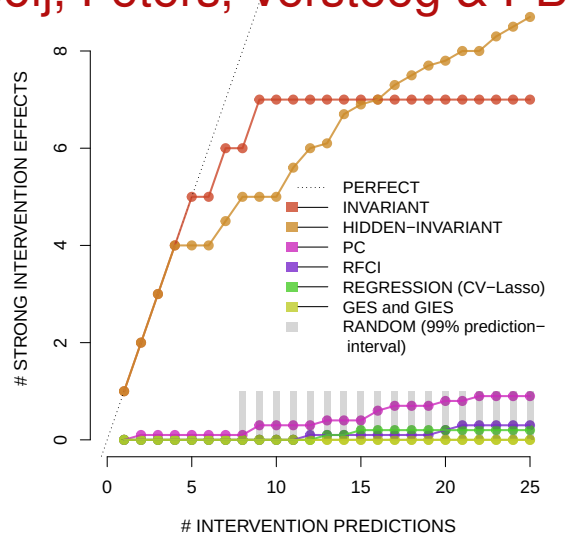
ICP has statistical error control (under some assumptions):

$$\mathbb{P}[\text{at least one false positive claim}] \leq \alpha$$

$\rightsquigarrow$ has led to interesting findings in gene network of *Saccharomyces Cerevisiae* (yeast) which were

biologically validated by gene knock-out experiments

Meinshausen, Hauser, Mooij, Peters, Versteeg & PB (2016)



Genome-wide mRNA expressions in yeast: $p = 6170$ genes

- $n_{obs} = 160$ "observational" samples of wild-types
- $n_{int} = 1479$ "interventional" samples
  corresponding to a single gene deletion strain

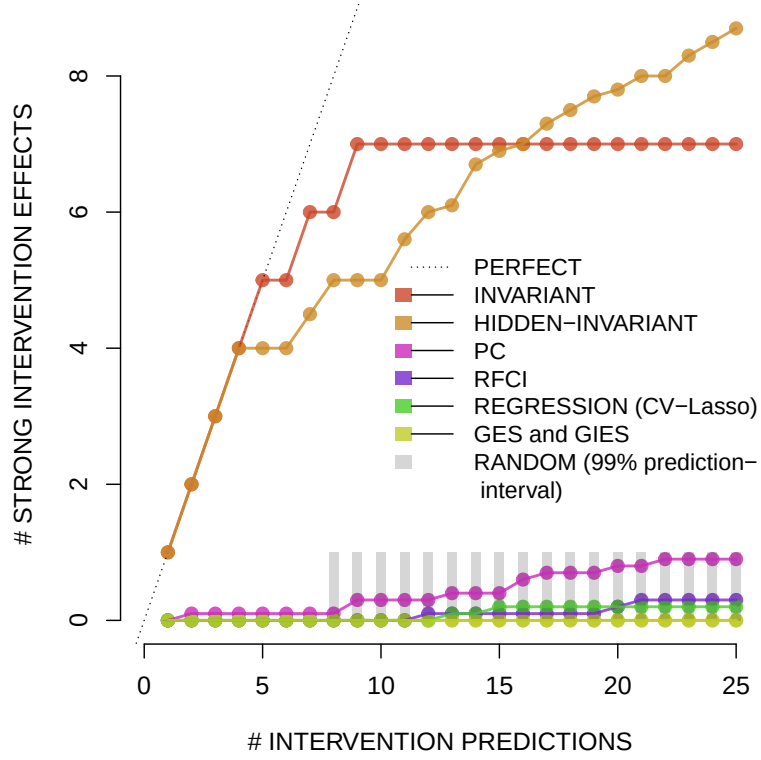
fit invariant linear model and

rank genes according to p-values for invariance

(one gene expression is response; all others the covariates)

that is: observed gene knock-outs ("environments") are used to predict new gene knock-out effects

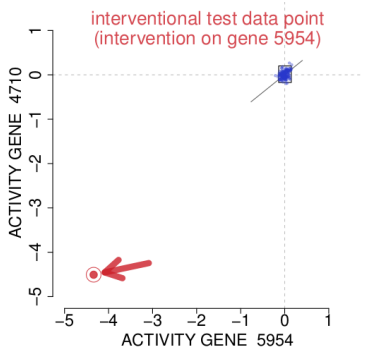
validation: 1/3 of the interventional data are set aside
check whether a predicted intervention effect actually exhibited a strong true effect

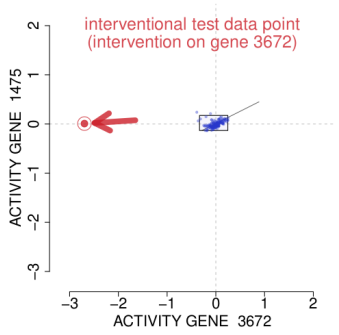


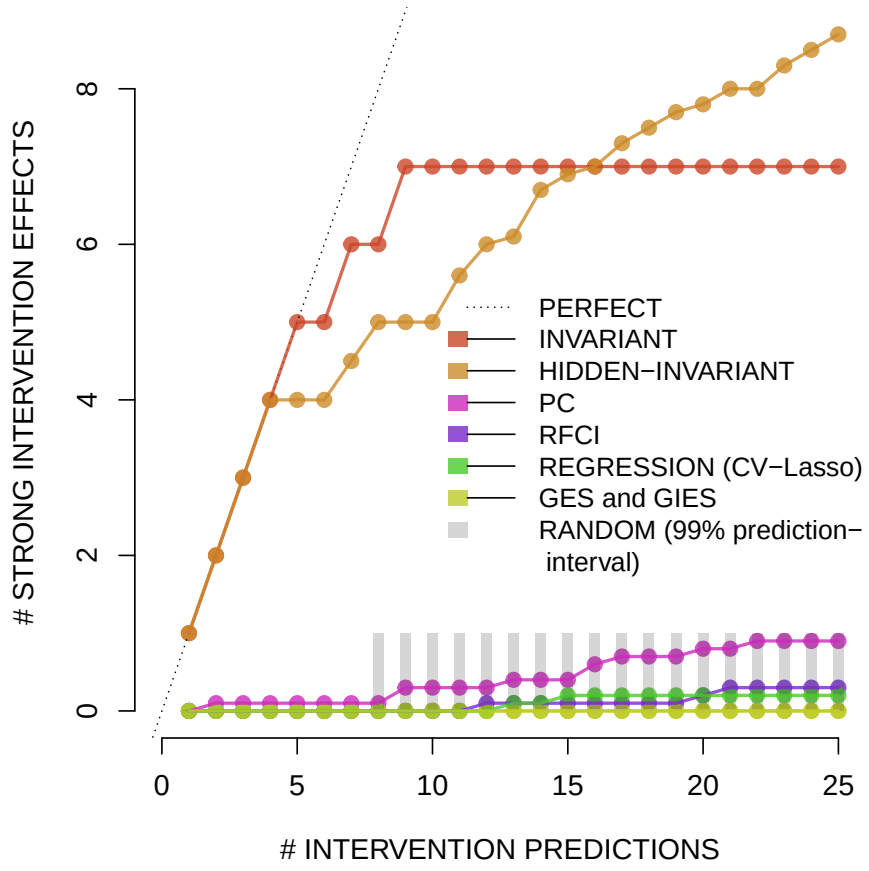
I : invariant causal prediction method

H: invariant causal prediction with some hidden variables

for predicting interventions (the red dot)







I : invariant causal prediction method

H: invariant causal prediction with some hidden variables

we can prioritize future experiments!

Causal regularization

encouraging various degree of invariance, from small to large

$$\mathcal{L}(Y^e - X^e\beta^*) \equiv \text{ const. w.r.t. } e$$

the environments $e \in \mathcal{E}$ from before:
now outcomes of a q-dimensional random variable A ("anchor")

plausible construction for estimator based on n data points $Y_{n\times 1}$, $X_{n\times d}$, $A_{n\times q}$:

$$\text{argmin}_{\beta\in\mathbb{R}^d}(\|Y - X\beta\|_2^2/n + \xi\| \underbrace{A^T(Y - X\beta)}_{\text{"correlations"}} \|_2^2)$$

$\xi > 0$ a regularization parameter

encouraging "correlation invariance"

i.e., residuals being uncorrelated of A (aka envs.)

$$\text{argmin}_{\beta \in \mathbb{R}^d} \left(\| Y - X\beta \|_2^2 / n + \xi \| A^T (Y - X\beta) \|_2^2 \right)$$

causal regularization:

$$\hat{\beta} = \text{argmin}_\beta \left(\| (I - \Pi_A)(Y - X\beta) \|_2^2 / n + \gamma \| \Pi_A (Y - X\beta) \|_2^2 / n \right)$$

$\Pi_A = A(A^T A)^{-1} A^T$ (projection onto column space of A)

- for $\gamma = 1$: least squares
- for $0 \leq \gamma < \infty$: general causal regularization

$$\text{argmin}_{\beta \in \mathbb{R}^d} \left(\|Y - X\beta\|_2^2/n + \xi \|A^T(Y - X\beta)\|_2^2 \right)$$

causal regularization:

$$\hat{\beta} = \text{argmin}_\beta \left(\|(I - \Pi_A)(Y - X\beta)\|_2^2/n + \gamma \|\Pi_A(Y - X\beta)\|_2^2/n + \lambda \|\beta\|_1 \right)$$

$\Pi_A = A(A^T A)^{-1} A^T$ (projection onto column space of A)

- for $\gamma = 1$: least squares + ℓ_1-penalty
- for $0 \leq \gamma < \infty$: general causal regularization + ℓ_1-penalty

convex optimization problem

the framework also encompasses nonlinear function estimation (e.g. Random Forests, deep neural networks, etc.) PB (2020)

$$\text{argmin}_{f \in \mathbf{F}} \left(\|(I - \Pi_A)(Y - f(X))\|_2^2/n + \gamma \|\Pi_A(Y - f(X))\|_2^2/n \right)$$

the framework also encompasses nonlinear function estimation (e.g. Random Forests, deep neural networks, etc.) PB (2020)

$$\begin{aligned}\text{argmin}_{f\in\mathbf{F}}\big(\quad & \|(I-\Pi_A)(Y-f(X))\|_2^2/n + \xi\ell_{\text{physical}}(Y,X,f) \\ + \quad & \gamma\|\Pi_A(Y-f(X))\|_2^2/n)\big)\end{aligned}$$

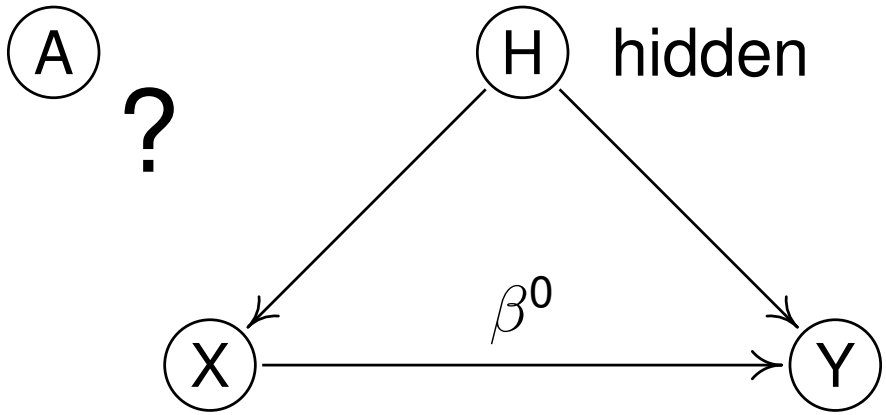
e.g. Quarteroni's physical heart model

(8th ECM first plenary talk)

can "add invariance" to a broad class of mathematical models!
and perhaps one "should"...

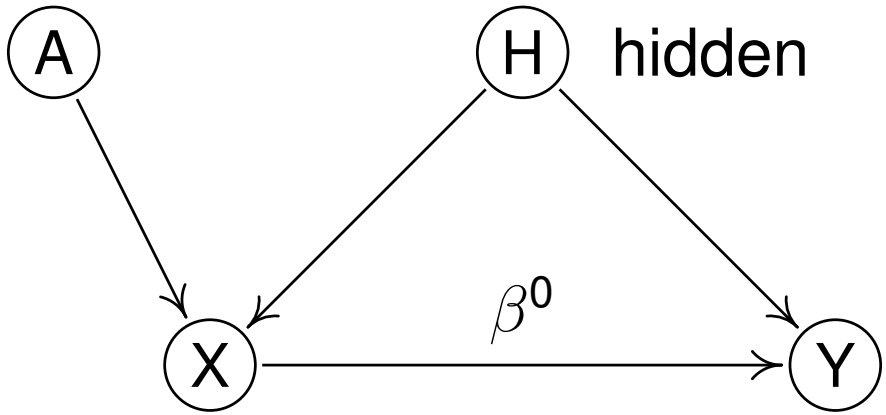
Causal regularization: nice mathematical guarantees

here for simplicity for linear models



Causal regularization: nice mathematical guarantees

here for simplicity for linear models



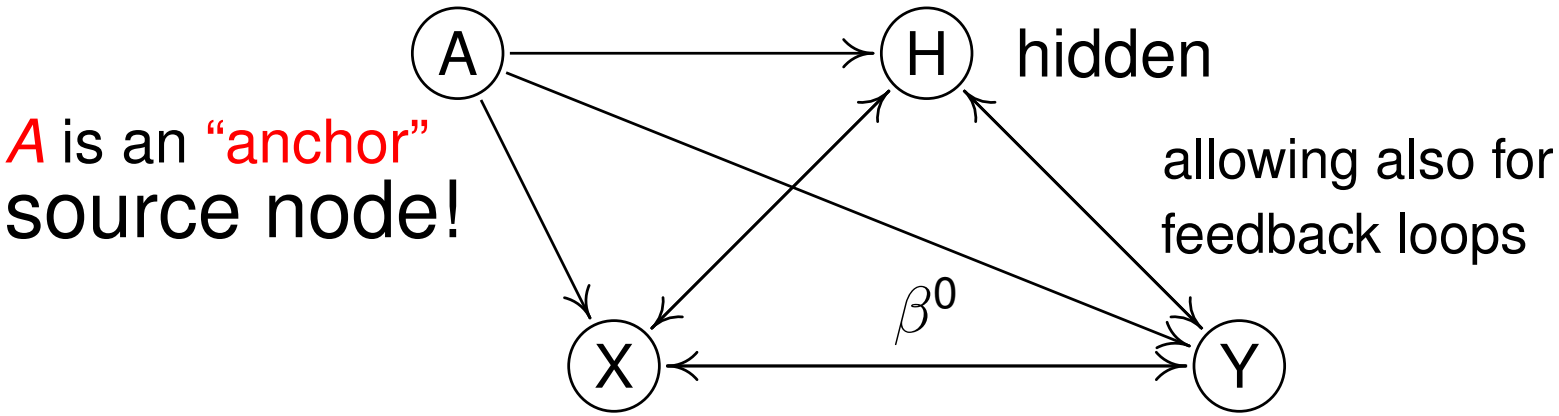
$$Y \leftarrow X\beta^0 + H\delta + \varepsilon_Y,$$
$$X \leftarrow A\alpha + H\gamma + \varepsilon_X$$

Instrumental variables regression model

(cf. Angrist, Imbens, Lemieux, Newey, Rosenbaum, Rubin,...)

Causal regularization and Anchor regression

(Rothenhäusler, PB, Peters & Meinshausen, 2021)



$\rightsquigarrow$ called Anchor regression

$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} \leftarrow B \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + \varepsilon + MA$$

allow that A acts on Y and H

$\rightsquigarrow$ there is a fundamental identifiability problem

cannot identify β^0

this is the price for more realistic assumptions than IV model

... but Causal Regularization offers something

but causal regularization solves for

$$\text{argmin}_{\beta \in \mathbb{R}^d} \max_{e \in \mathcal{F}} \mathbb{E}|Y^e - X^e \beta|^2$$

for a certain class of shift perturbations $\mathcal{F}$

Model for $\mathcal{F}$: shift perturbations

model for observed heterogeneous data ("corresponding to $\mathcal{E}$")

$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} = B \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + \varepsilon + MA$$

model for shift perturbations $\mathcal{F}$ (in test data)
shift vectors v

$$\begin{pmatrix} X^v \\ Y^v \\ H^v \end{pmatrix} = B \begin{pmatrix} X^v \\ Y^v \\ H^v \end{pmatrix} + \varepsilon + v$$

$v \in C_\gamma \subset \text{ span}(M)$, γ measuring the size of v

i.e. $v \in C_\gamma = \{v;\ v = Mu \text{ for some } u \text{ with } \mathbb{E}[uu^T] \preceq \gamma \mathbb{E}[AA^T]\}$

$$v \in C_\gamma = \{v;\ v = Mu \text{ for some } u \text{ with } \mathbb{E}[uu^T] \subset \text{span}(M)$$

in folklore:

perturbations (v) in new test data have the same direction (span(M)) as the observed heterogeneity/perturbations in the training data

but they could be much stronger in the test data (γ large)

- the most natural extrapolation!
- the more perturbations and heterogeneity you see, the better we can extrapolate (larger span(M))!

  this is controversial in classical statistics – but it has fundamental consequences for robustness

A fundamental duality theorem

(Rothenhäusler, Meinshausen, PB & Peters, 2018)

P_A the population projection onto A: $P_A \bullet = \mathbb{E}[\bullet|A]$

For any β

$$\max_{v \in C_\gamma} \mathbb{E}[|Y^v - X^v\beta|^2] = \mathbb{E}\left[\left|(\mathrm{Id} - P_A)(Y - X\beta)\right|^2\right] + \gamma\mathbb{E}\left[\left|P_A(Y - X\beta)\right|^2\right]$$

$$\approx \underbrace{\|(I - \Pi_A)(Y - X\beta)\|_2^2/n + \gamma\|\Pi_A(Y - X\beta)\|_2^2/n}_{\text{objective function on data}}$$

worst case shift interventions $\longleftrightarrow$ regularization!

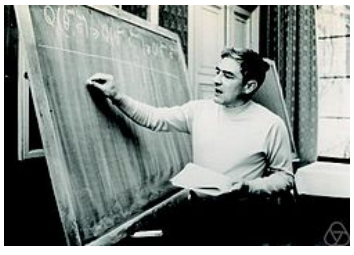
in the population case

$\rightsquigarrow$ just regularize! (instead of l.h.s. which is a difficult object)

robustness $\longleftrightarrow$ causal regularization

Robust Statistics
Robust Optimization
Adversarial Learning

Causality

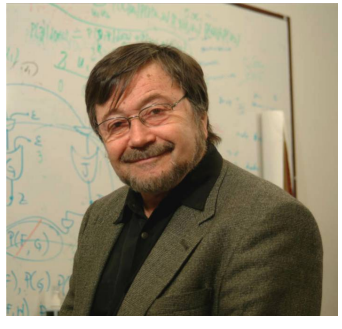


Peter Huber Arkadi Nemirovski

Aharon Ben-Tal

Ian Goodfellow


Phil Dawid Peter Spirtes



Judea Pearl

distributional robustness $\longleftrightarrow$ causal regularization

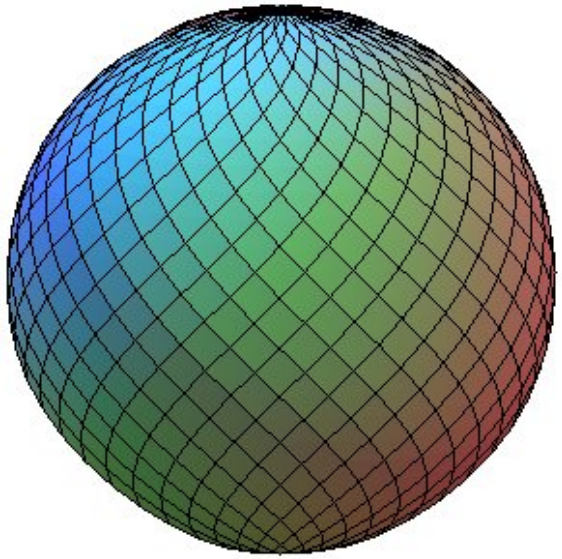
$$\sup_{d(P,P_0)\leq\rho} \mathbb{E}_P[(Y - X\beta)^2] \quad = \quad \max_{v\in C_\gamma} \mathbb{E}[|Y^v - X^v\beta|^2]$$

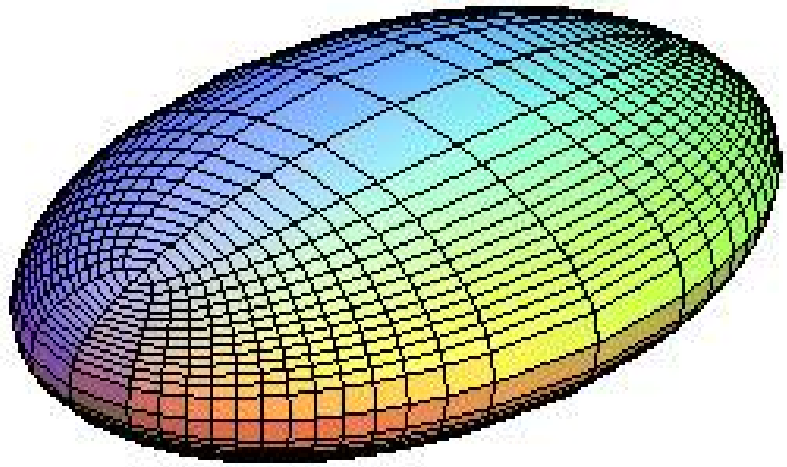
- P_0 is the observational distribution (steady state)
- for Gaussian case:
  $d(.,.) = d_M(.,.)$ is Wasserstein between Gaussians with interventional data-dependent covariances, potentially degenerate

  $\rightsquigarrow$ learned metric from the observed perturbations, especially useful in high dimensions

distributional robustness $\longleftrightarrow$ causal regularization

$$\sup_{d(P,P_0)\leq\rho} \mathbb{E}_P[(Y - X\beta)^2] \quad = \quad \max_{v\in C_\gamma} \mathbb{E}[|Y^v - X^v\beta|^2]$$





(mathcurve.com)

postulated sphere $\longleftrightarrow$ ellipsoid metric learned from data

robustness $\longleftrightarrow$ causality

the languages are rather different:

- metric for robustness
  Wasserstein, f-divergence
- optimal transport and
  adversarial robustness
- minimax optimality
- regularization
- ...

- causal graphs
- Markov properties on
  graphs
- perturbation models
- identifiability of systems
- transferability of systems
- ...

mathematics allows to classify equivalences and differences

$\rightsquigarrow$ can be exploited for better methods and algorithms

taking “the good” from both worlds!

and indeed, one can improve prediction
with causal-type regularization

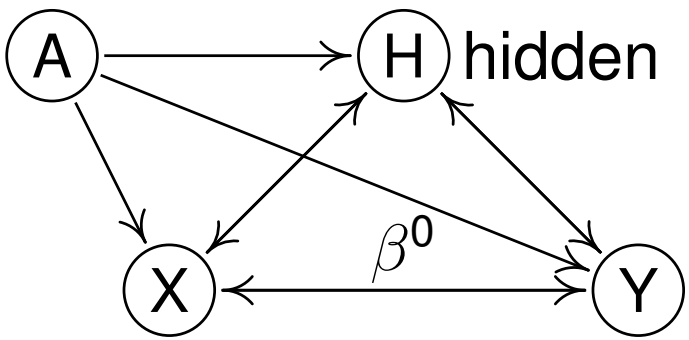
- causal-robust machine learning
  Leon Bottou et al. since 2013 (Microsoft and now Facebook)
- CNN-based classification with conditional invariance
  (Heinze-Deml and Meinshausen, 2017)
- invariant risk minimization
  (Arjovsky, Bottou, Gulrajani & Lopez-Paz, 2019)
- causal domain adaptation
  (Magliacane et al., 2017; Chen & PB, 2020; ...)

Towards interpretation:
Science aims for causal understanding

... but this may be a bit ambitious...

causal inference necessarily requires (often untestable) additional assumptions

e.g. in anchor regression model: we cannot find/identify the causal ("systems") parameter β^0



Invariance and "diluted causality"

by the fundamental duality for causal regularization:

$$\beta^\gamma = \text{argmin}_{\beta\in\mathbb{R}^d} \left(\mathbb{E}[|(I - P_A)(Y - X\beta)|^2] + \gamma\mathbb{E}[|P_A(Y - X\beta)|^2]\right)$$

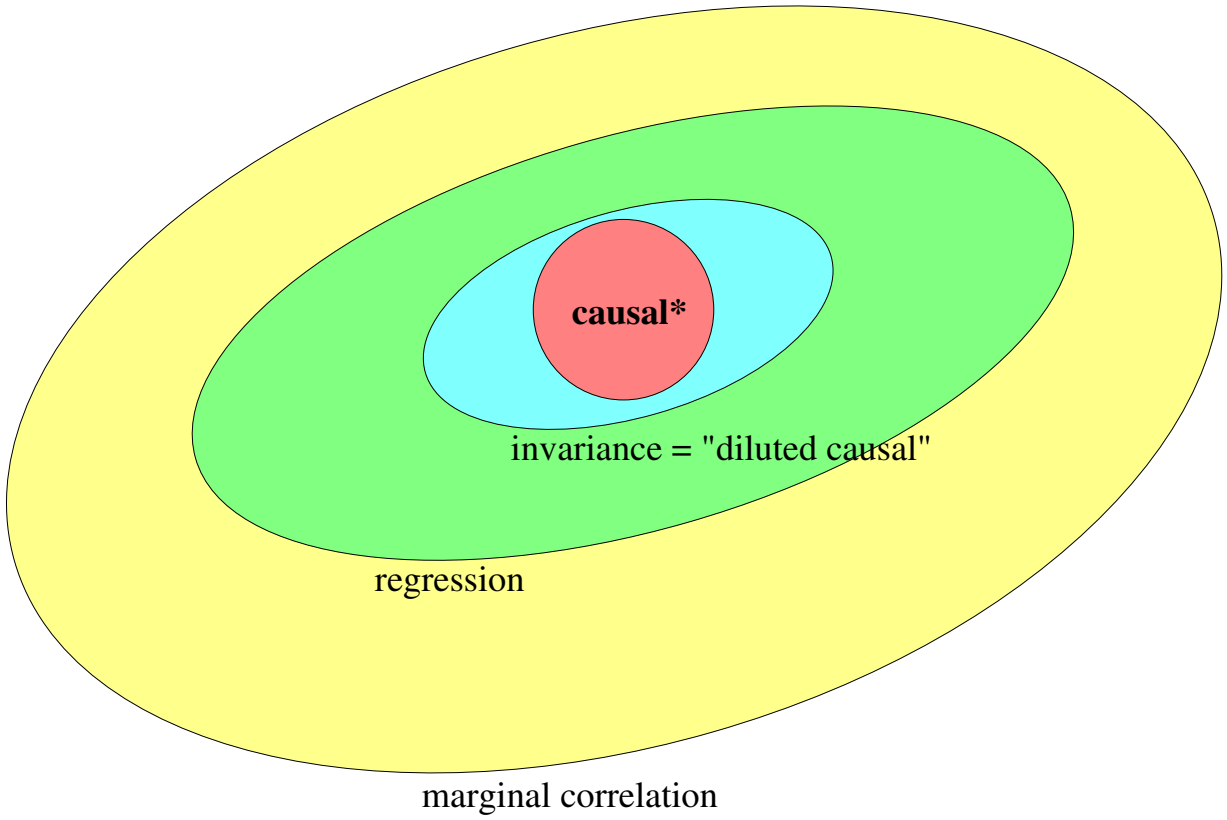
$\gamma \to \infty$ leads to shift invariance of residuals

$Y - X\beta^{\to\infty}$ has the same distribution over shift perturbations

$\beta^{\to\infty}$ is generally not the causal parameter
but because of shift invariance: call it "diluted causal"

note: causal = invariance w.r.t. very many perturbations

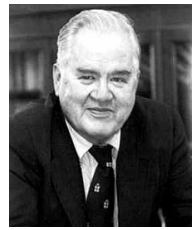
notions of associations



invariance = “diluted causal”

under faithfulness conditions, the figure is valid (causal* are the causal variables as in e.g. large parts of Dawid, Pearl, Robins, Rubin, ...)

Stabilizing

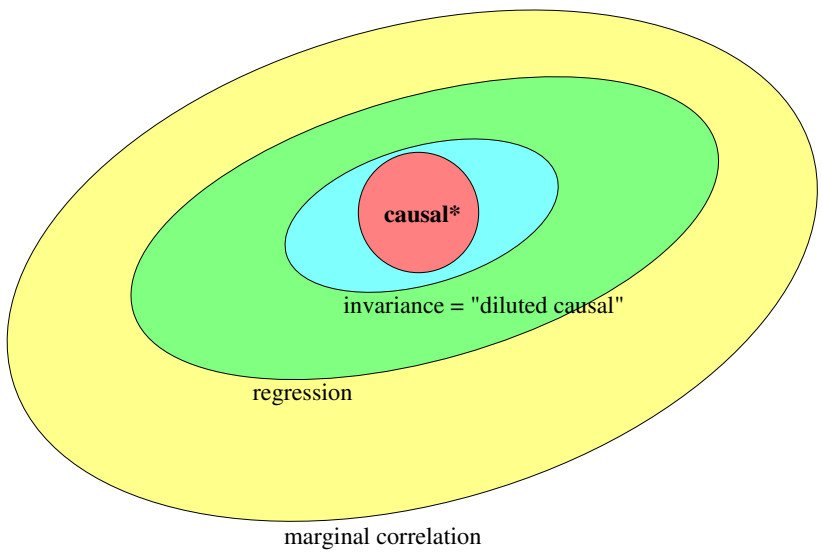


John W. Tukey (1915 – 2000), also co-inventor of FFT

Tukey (1954)

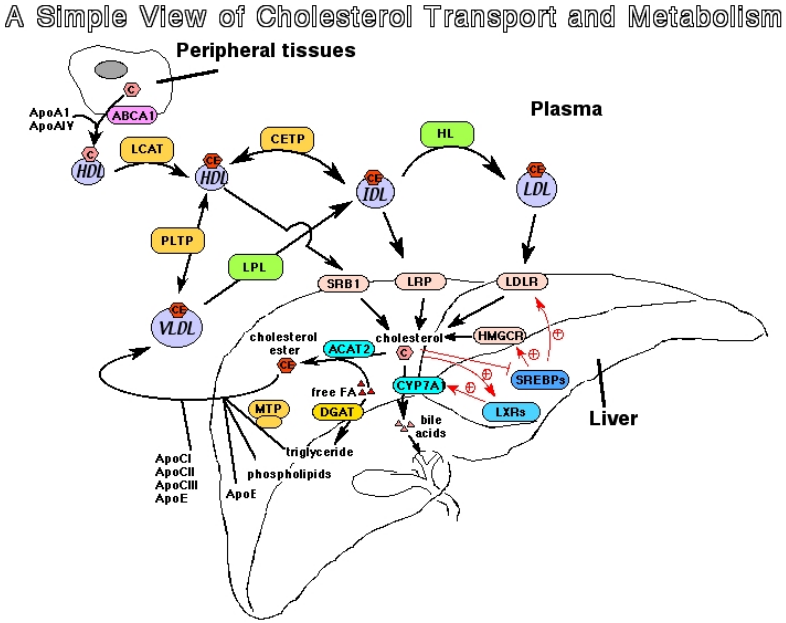
"One of the major arguments for regression instead of correlation is potential stability. We are very sure that the correlation cannot remain the same over a wide range of situations, but it is possible that the regression coefficient might. ...

We are seeking stability of our coefficients so that we can hope to give them theoretical significance."



“Diluted causality”: important proteins for cholesterol

Ruedi Aebersold, ETH Zürich



3934 other proteins
which of those are
“diluted causal”
for cholesterol

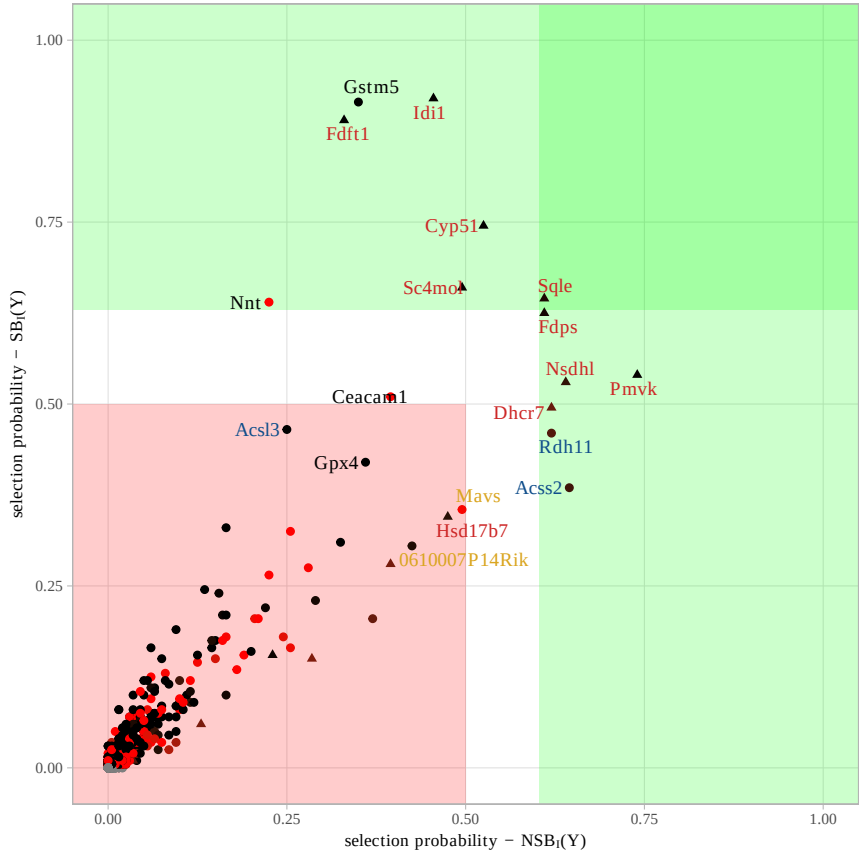
experiments with mice: 2 environments with fat/low fat diet

high-dimensional regression, total sample size $n = 270$

Y = cholesterol pathway activity, X = 3934 protein expressions

x-axis: importance w.r.t regression but non-invariant

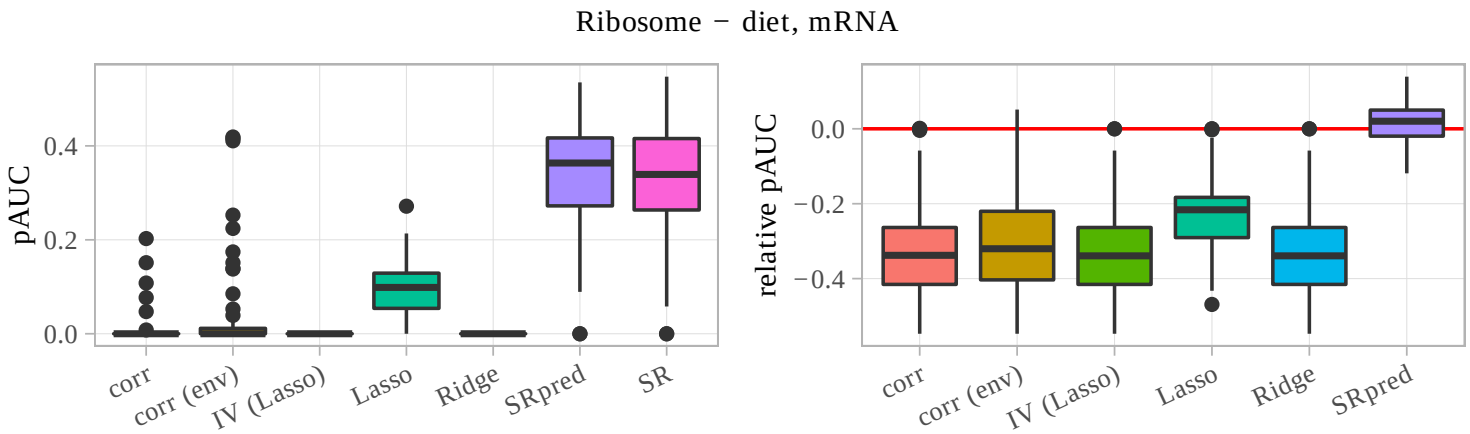
y-axis: importance w.r.t. invariance



... and we have some forms of validation

e.g. with respect to finding known pathways

for Ribosome pathway



invariance performs better than regression association!

for inferring known pathway structure

... and the story would go on

causal regularization leads to

- better distributional replicability of findings in new datasets in genomics Rothenhäusler, Peters, PB & Meinshausen (2018)
- large-scale kinetic systems based on metabolomics (Pfister, Bauer and Peters, 2019)

- finding more promising proteins and genes: based on high-throughput proteomics
- prediction of gene knock-downs: based on transcriptomics

Conclusions

- interpretation/“causality” and robustness are related to each other!

  at least in the outlined framework of causality and robustness
- stabilizing and finding suitable invariances in large datasets are powerful and can make a relevant difference in practice

  in the context of many mathematical models
- we hear nowadays quite often about “causal AI”

  overly ambitious wording and “true(ish) causality” is far away

  but in fact, it is mostly about “robust and reliable prediction”

robustness – causality