Independence of conditionals as causal inference principle

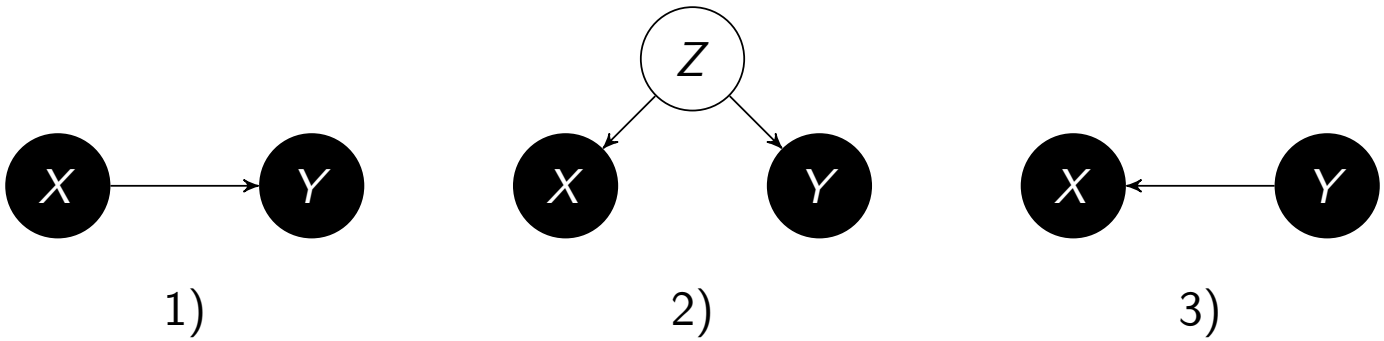
Dominik Janzing and Eleni Sgouritsa

Max Planck Institute for Intelligent Systems
Tübingen, Germany

06. June 2014

Reichenbach's principle of common cause (1956)

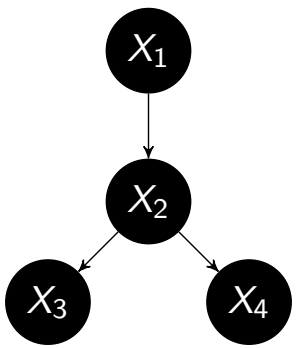
If two variables X and Y are statistically dependent then either



- in case 2) Reichenbach postulated $X \perp\!\!\!\perp Y \,|Z$.
- every statistical dependence is due to a causal relation, we also call 2) "causal".
- distinction between 3 cases is a key problem in scientific reasoning and the focus of this talk.

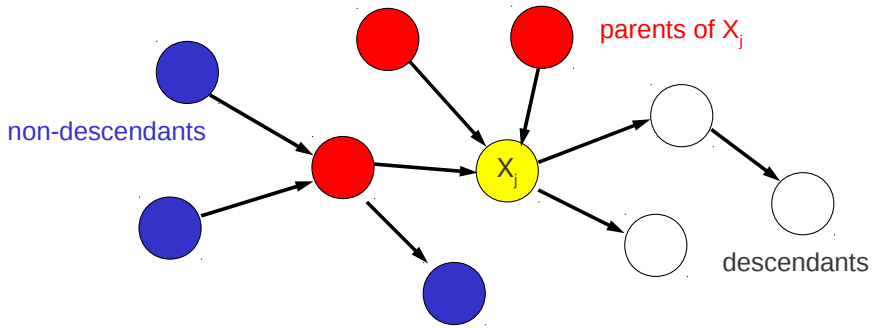
Causal inference problem, general form Spirtes, Glymour, Scheines, Pearl

- Given variables $X_1, \ldots, X_n$
- infer causal structure among them from n-tuples iid drawn from $P(X_1, \ldots, X_n)$
- causal structure = directed acyclic graph (DAG)



Causal Markov condition (3 equivalent versions) Lauritzen et al

- **local Markov condition:** every node is conditionally independent of its non-descendants, given its parents



(information exchange with non-descendants involves parents)

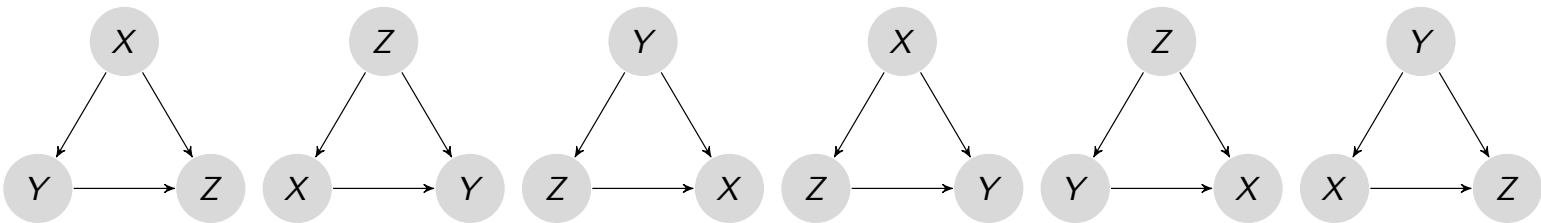
- **global Markov condition:** describes all ind. via d-separation
- **Factorization:** $P(X_1, \ldots, X_n) = \prod_j P(X_j|PA_j)$

(every $P(X_j|PA_j)$ describes a causal mechanism)

Causal inference from observational data

Can we infer G from $P(X_1, \ldots, X_n)$?

- MC only describes which sets of DAGs are consistent with P
- $n!$ many DAGs are consistent with any distribution



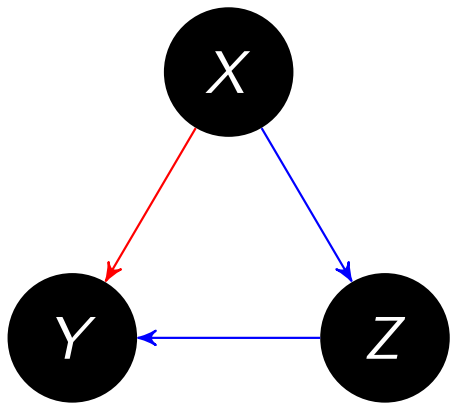
- reasonable rules for preferring simple DAGs required

Causal faithfulness

Spirtes, Glymour, Scheines, 1993

Prefer those DAGs for which all observed conditional independences are implied by the Markov condition

- **Idea:** generic choices of parameters yield faithful distributions
- **Example:** let $X \perp\!\!\!\perp Y$ for the DAG



- not faithful, direct and indirect influence compensate
- **Application:** PC and FCI infer causal structure from conditional statistical independences

Justifying faithfulness

Unfaithful distributions occur with probability zero if

- nature chooses each $P(X_j|PA_j)$ independently
- each $P(X_j|PA_j)$ is chosen from a probability density in parameter space (e.g. uniform distribution)

  here the parameter space of each conditional is a subset of $\mathbb{R}^k$ with $k := \{x_j\}^{\{pa_j\}}$

C. Meek: Strong completeness and faithfulness in Bayesian networks. (UAI 1995)

However, we will argue...

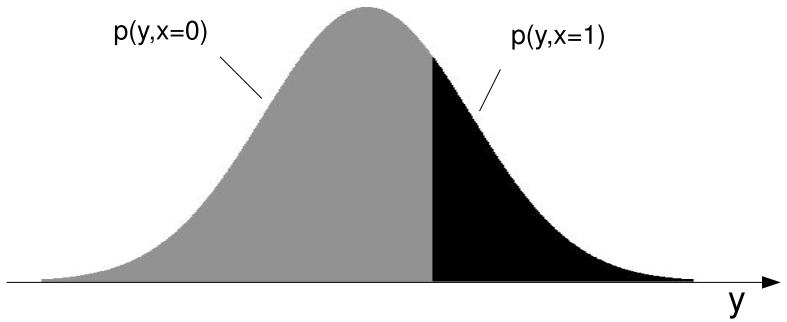
- There are cases of obvious parameter tuning that do not generate additional independences
  ($\Rightarrow$ faithfulness is too weak)

- Not every violation of faithfulness is due to parameter tuning since we do not believe in *densities* on the parameter space
  ($\Rightarrow$ faithfulness is too strong)

Why faithfulness is too weak

Let X be binary and Y real-valued.

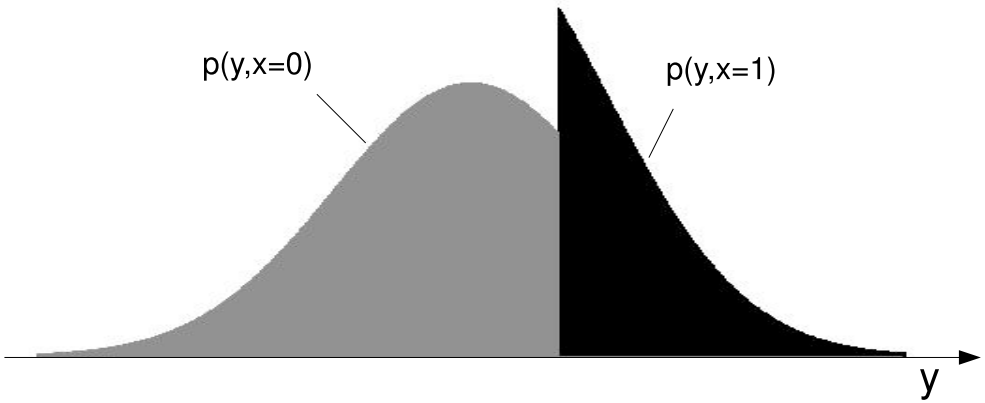
- Let Y be Gaussian and $X = 1$ for all y above some threshold and $X = 0$ otherwise.



- $Y \to X$ is plausible: simple thresholding mechanism
- $X \to Y$ requires a strange mechanism:
  look at $P(Y|X = 0)$ and $P(Y|X = 1)$!

not only $P(Y|X)$ itself is strange...

but also what happens if we change $P(X)$:



Hence, reject $X \to Y$ because it requires tuning of $P(X)$ relative to $P(Y|X)$. Faithfulness would accept both causal directions.

Why faithfulness is too strong

Consider deterministic relations

$$Y = f(X)$$

- unfaithful because $Y \perp\!\!\!\perp Z \,|X$
- but there is no adjustment between $P(X)$, $P(Y|X)$, $P(Z|Y)$
- only $P(Y|X)$ is 'non-generic'

We don't want to reject non-generic conditionals, we only want to reject non-generic **relations** between conditionals

Algorithmic independence of conditionals

The **shortest** description of $P(X_1, \ldots, X_n)$ is given by **separate** descriptions of $P(X_j|PA_j)$.

(Here, description length = Kolmogorov complexity)

Janzing, Schölkopf: Causal inference using the algorithmic Markov condition, IEEE TIT (2010).

Lemeire, Janzing: Replacing causal faithfulness with the algorithmic independence of conditionals, Minds & Machines (2012).

Short introduction into Kolmogorov complexity

Kolmogorov 1965, Chaitin 1966, Solomonoff 1964

of a binary string x

- $K(x) =$ length of the shortest program with output x
- interpretation: number of bits required to describe the rule that generates x
- neglect string-independent terms; use $\stackrel{+}{=}$ instead of $=$
- $K(x)$ is uncomputable
- probability-free definition of information content

Conditional Kolmogorov complexity

- $K(y|x)$: length of the shortest program that generates y from the input x.
- number of bits required for describing y if x is given
- $K(y|x^*)$ length of the shortest program that generates y from x^*, i.e., the shortest compression x.
- subtle difference: x can be generated from x^* but not vice versa because there is no algorithmic way to find the shortest compression

Algorithmic mutual information

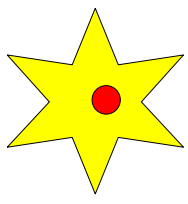
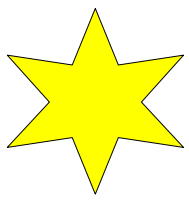
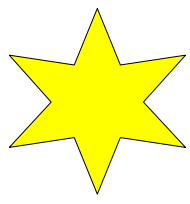
Chaitin, Gacs

Information of x about y (and vice versa)

- $I(x:y) := K(x) + K(y) - K(x,y)$
  $\stackrel{+}{=} K(x) - K(x|y^*) \stackrel{+}{=} K(y) - K(y|x^*)$

- Interpretation: number of bits saved when compressing x, y jointly rather than compressing them independently

Algorithmic mutual information: example

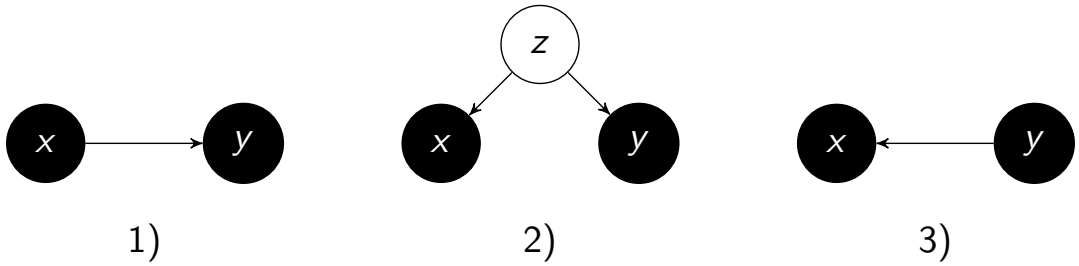
I( : ) = K()

Analogy to statistics:

- replace strings x, y (=objects) with random variables X, Y
- replace Kolmogorov complexity with Shannon entropy
- replace algorithmic mutual information $I(x : y)$ with statistical mutual information $I(X; Y)$

Causal Principle

If two strings x and y are algorithmically dependent then either



- every algorithmic dependence is due to a causal relation
- algorithmic analog to Reichenbach's principle of common cause
- distinction between 3 cases: use conditional independences on more than 2 objects

DJ, Schölkopf IEEE TIT 2010

Apply the causal principle to conditionals

$K(P(X_j|PA_j))$ denotes the length of the shortest program computing $P(x_j|pa_j)$ from (x_j, pa_j).

- If nature chooses each mechanism $P(X_j|PA_j)$ independently they are algorithmically independent, e.g.,

$$I(P(X_j|PA_j) : P(X_1|PA_1), P(X_2|PA_2), \dots) \stackrel{+}{=} 0 \quad \forall j\,.$$

- equivalent to

$$K(P(X_1, \dots, X_n)) \stackrel{+}{=} \sum_{j=1}^{n} K(P(X_j|PA_j))$$

(shortest description of the joint is given by separate descriptions of the causal conditionals)

Bivariate case

If $X \to Y$ then

$$I(P(X) : P(Y|X)) \stackrel{+}{=} 0$$

and, equivalently,

$$K(P(X,Y)) \stackrel{+}{=} K(P(X)) + K(P(Y|X))\,.$$

Relation to Occam's Razor

Note:

$$K(P(X,Y)) \stackrel{+}{=} K(P(X)) + K(P(Y|X))\,.$$

implies

$$K(P(X)) + K(P(Y|X)) \leq K(P(Y)) + K(P(X|Y))\,.$$

but not vice versa.

Principle of independent conditionals is stronger

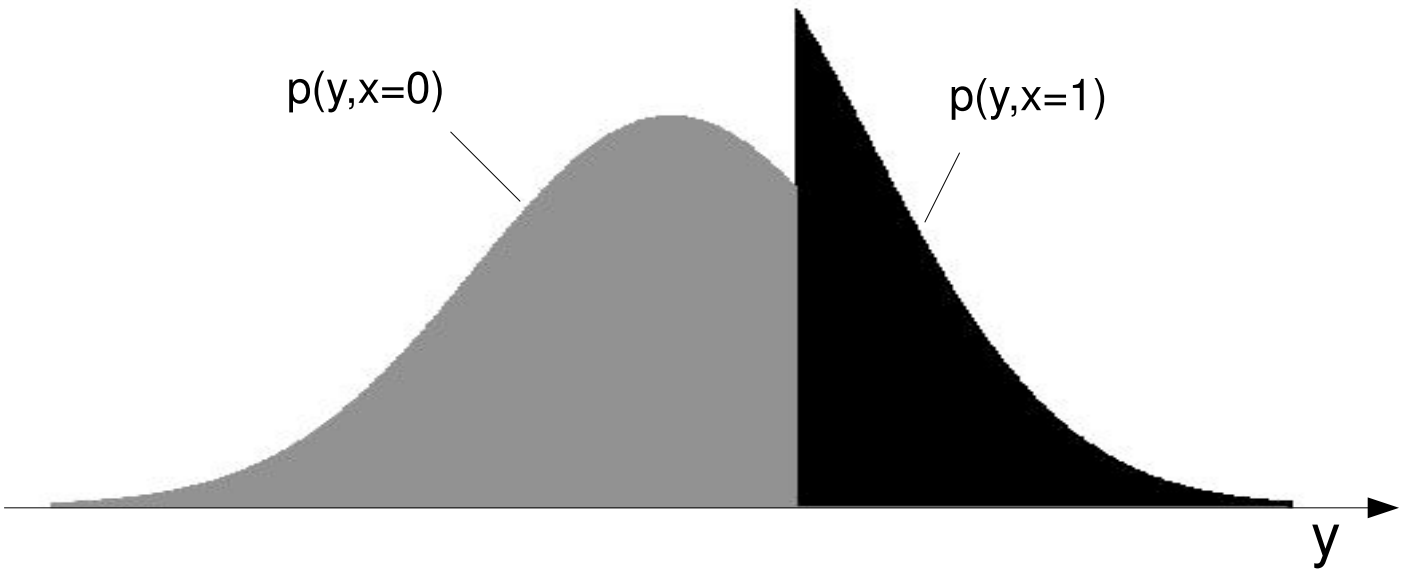
Assume

$$K(P(X,Y)) \leq \begin{cases} K(P(X)) + K(P(Y|X)) \\ K(P(Y)) + K(P(X|Y)) \end{cases}$$

Then we reject both DAGs $X \to Y$ and $Y \to X$.
(e.g. if $X \leftarrow Z \to Y$ is the true structure)

- Occam's Razor chooses the simplest model from the given class although the class may be inappropriate
- Our principle is (in principle) able to detect this

Revisiting the motivating example



Knowing $P(Y|X)$, there is a short description of $P(X)$, namely 'the unique distribution for which $\sum_x P(Y|x)P(x)$ is Gaussian'.

Bayesian view on Independent Conditionals vs. Faithfulness

Replace uniform prior with **Solomonoff's prior**:

- preferring simple structures is crucial for inference (e.g. if you are supposed to infer how 101010... continues)
- string x occurs with probability proportional to $2^{-K(x)}$
- the conditional $P(X_j|PA_j)$ occurs with probability $2^{-K(P(X_j|PA_j))}$
- simple conditionals get high probability.

Lemeire & Janzing: Replacing causal faithfulness with algorithmic independence of conditionals, Minds & Machines, 2012.

Then algorithmic dependences become unlikely:

For general objects:

- let x, y be strings describing two objects.
- if generated independentl, the pair (x, y) occurs with probability $2^{-K(x)}2^{-K(y)}$
- if generated jointly, it occurs with probability $2^{-K(x,y)}$
- hence $K(x, y) \ll K(x) + K(y)$ indicates generation in a joint process

For conditionals: let x and y be descriptions of $P(X_1|PA_1)$ and $P(X_2|PA_2)$, respectively

Example for simple conditionals

- **unbiased input:**

$$P(x) = 1/2$$

- **identity:**

$$P(y|x) = \begin{cases} 1 & \text{for } y = x \\ 0 & \text{otherwise} \end{cases}$$

- **AND-gate:**

$$P(z|x,y) = \begin{cases} 1 & \text{for } z = x \wedge y \\ 0 & \text{otherwise} \end{cases}$$

- **XOR-gate:**

$$P(z|x,y) = \begin{cases} 1 & \text{for } z = x \oplus y \\ 0 & \text{otherwise} \end{cases}$$

- **linear Gaussian model with simple parameter values:**

$$P(y|x) = \frac{1}{\sqrt{2\pi}} e^{-(y-x)^2}$$

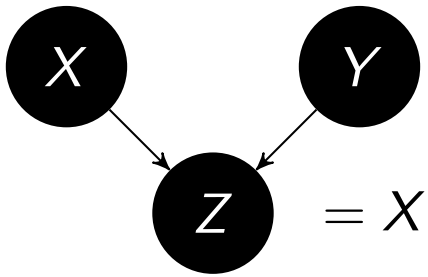
Simple conditionals do occur in nature

- **unbiased input:** physical two-level system with high temperature (each level occurs with probability $1/2$)
- **AND:** rainbow requires rain and sun
- **OR:** problems with one of our organs results in sickness
- **Determinism:** frequency of a pendulum determined by mass and length
- **Symmetries:** Let v be the velocity vector of a particle, then $p(y)$ is a rotation invariant Gaussian in thermal equilibrium

Likely violations of faithfulness

by combining *simple* ('non-generic') conditionals:

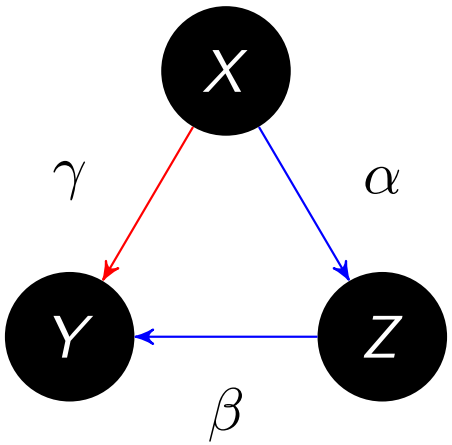
$P(x) = 1/2$



$= X \oplus Y$

yields $Z \perp\!\!\!\perp Y$

Similarities with faithfulness



$$Z = \alpha X + U_Z$$

$$Y = \gamma X + \beta Z + U_Y$$

- faithfulness forbids $\alpha\beta = -\gamma$
- also forbidden by algorithmic independence of conditionals *unless* α is simple, i.e., $\alpha = 1$ and $\beta = -\gamma$

Although Kolmogorov complexity is uncomputable...

we apply the principle of algorithmically independent conditionals:

- find notions of dependence of conditionals that capture essential aspects
- use it as a foundation/justification of new inference rules

Justifying additive noise based causal inference

Assume $Y = f(X) + E$ with $E \perp\!\!\!\perp X$ as in Jonas Peters' talk

- Then $P(Y)$ and $P(X|Y)$ are related:

$$\frac{\partial^2}{\partial y^2} \log p(y) = -\frac{\partial^2}{\partial y^2} \log p(x|y) - \frac{1}{f'(x)} \frac{\partial^2}{\partial x \partial y} \log p(x|y)\,.$$

  $\Rightarrow \frac{\partial^2}{\partial y^2} \log p(y)$ can be computed from $p(x|y)$ knowing $f'(x_0)$ for one specific x_0

- Given $P(X|Y)$, $P(Y)$ has a short description.
- We reject $Y \to X$ provided that $P(Y)$ is complex

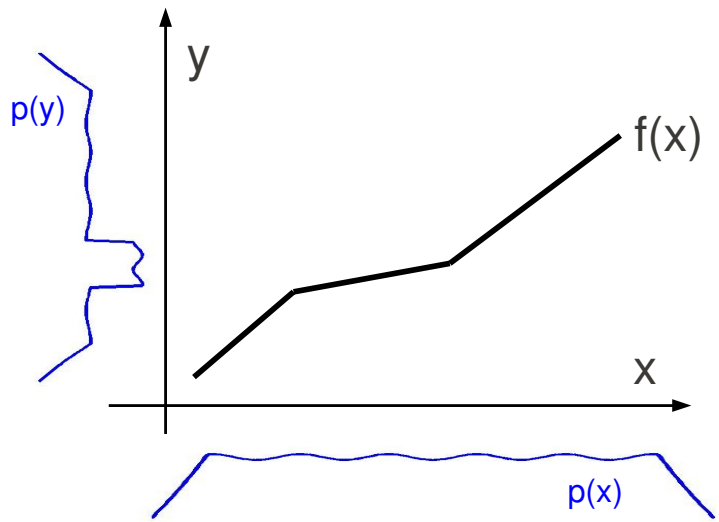
Hoyer, Janzing, Mooij, Peters, Schölkopf, NIPS (2008) (for the inference method)

Janzing, Steudel, OSID (2010) (for the justification)

Inferring deterministic causality

Daniusis, Janzing,... *UAI 2010*, Janzing et al. *AI 2012*

- Problem: infer whether $Y = f(X)$ or $X = f^{-1}(Y)$ is the right causal model
- Idea: if $X \to Y$ then f and the density p_X are chosen independently "by nature"
- Hence, peaks of p_X do not correlate with the slope of f
- Then, peaks of p_Y correlate with the slope of f^{-1}



- corresponding method called 'information geometric causal inference'

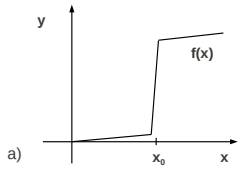
How IGCI is related to description length

Let $Y = f(X)$ with $f : [0,1] \to [0,1]$ monotonous and bijective.

- **IGCI considers f and $P(X)$ dependent** if

$$\int p(x) \log |f'(x)| dx \gg 0$$

- **Idea:** at most $1/c$ of x-values satisfy $|f'(x)| > c$



  at most 2^{-nc} of the n-tuples satisfy $\frac{1}{n}\sum_{j=1}^{n} \log |f'(x_j)| \geq c$.
- **description length** of $x_1, \ldots, x_n$ reduced by nc bits compared to a generic n-tuple
- f **contains information about** $P(X)$ because it belongs to a small set of distributions.

D. Janzing, B. Steudel, N. Shajarisales, B. Schölkopf: Justifying information-geometric causal inference, to appear in Chervonencis Festschrift

Common root of exististing causal inference methods

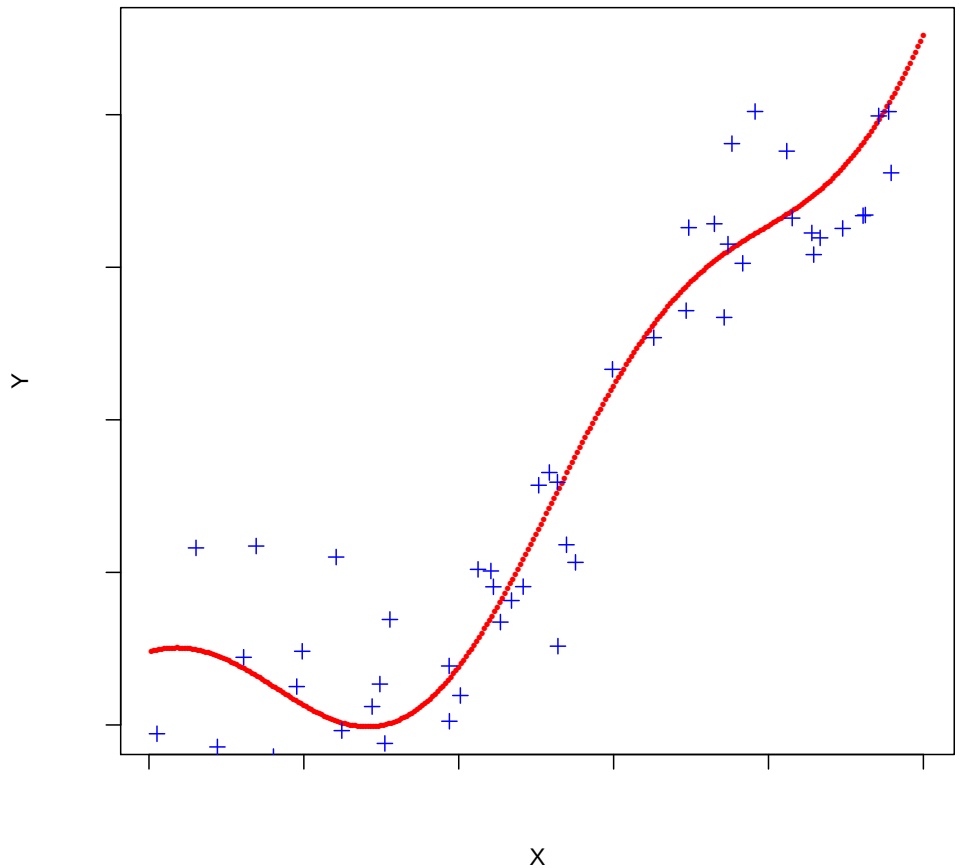
detect dependences between conditionals $P(X_j|PA_j)$

- **Independence-based approach:** reject DAGs that violate faithfulness
- **Additive noise models:** reject DAGs if other DAGs admit an additive noise model
- **Information-geometric causal inference:** reject $X \to Y$ with $Y = f(X)$ if $p(x)$ is large where f has large slope
- **future methods?**

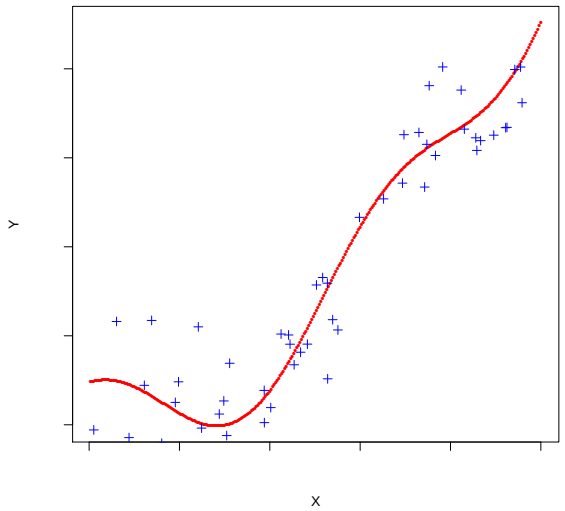
Goal: getting closer to 'algorithmic independence'?

How could standard machine learning benefit from causality?

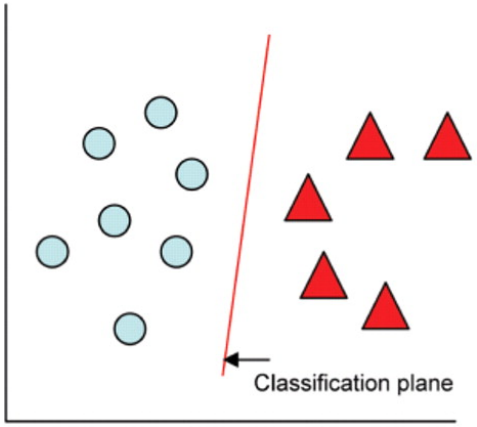
Task: predict y from x after observing some samples drawn from $P(X, Y)$.



- Regression: continuous label Y



- Classification: discrete label Y

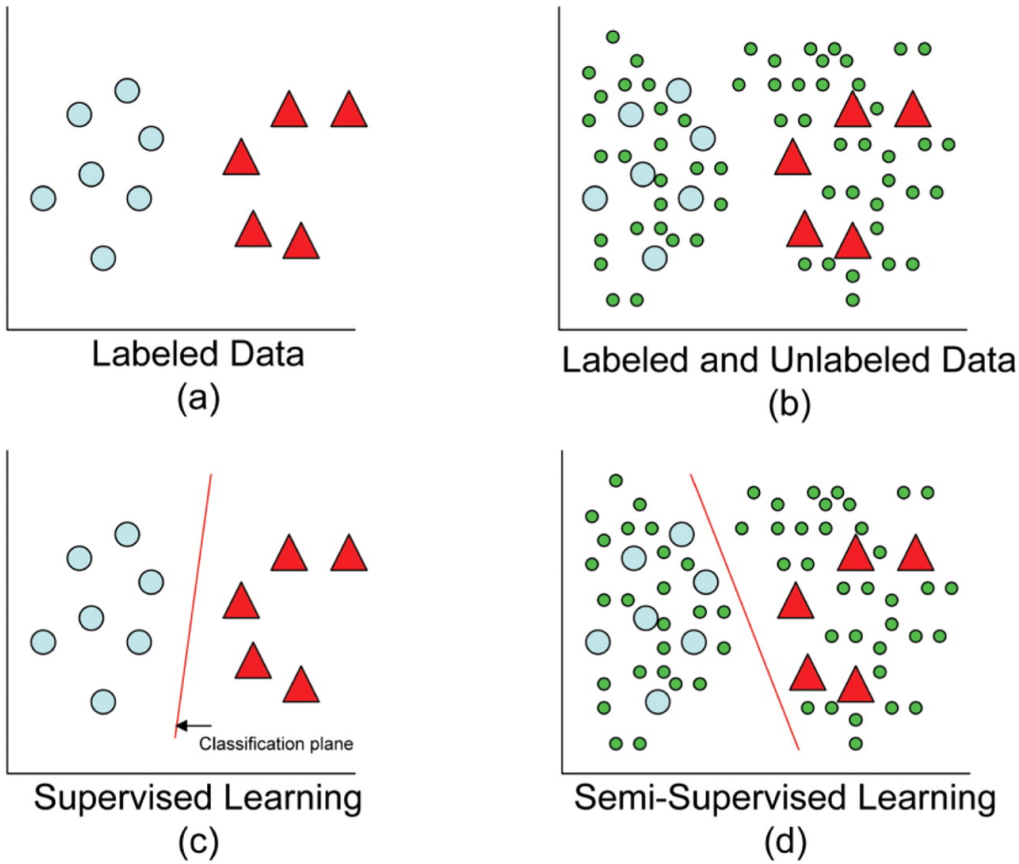


Semi-supervised learning (SSL)

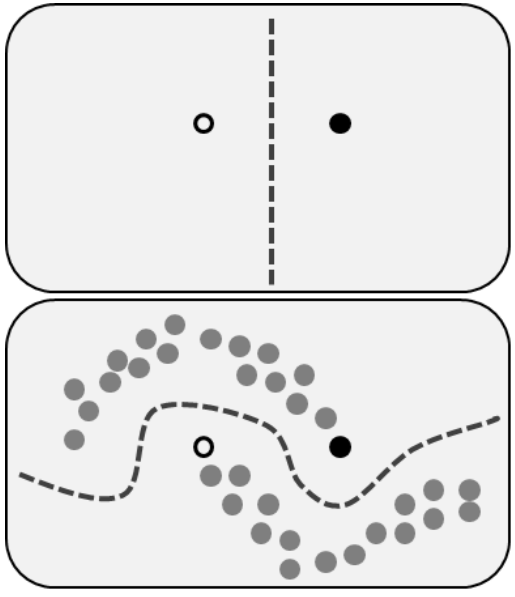
Goal: predict y from x after observing labeled data $(x_1, y_1), \ldots, (x_n, y_n)$ and **unlabeled** data $x_{n+1}, \ldots, x_{n+k}$

Semi-supervised learning (SSL)

Goal: predict y from x after observing labeled data $(x_1, y_1), \ldots, (x_n, y_n)$ and **unlabeled** data $x_{n+1}, \ldots, x_{n+k}$



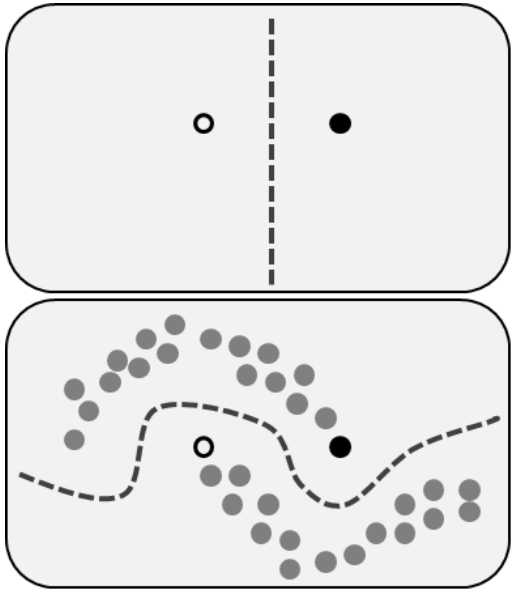
When can SSL work?



source: http://en.wikipedia.org/wiki/Semi-supervised_learning

- Can we have a more accurate prediction of Y by taking into account the unlabeled points?

When can SSL work?

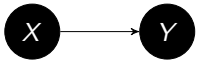


source: http://en.wikipedia.org/wiki/Semi-supervised_learning

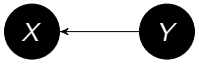
- Can we have a more accurate prediction of Y by taking into account the unlabeled points?
- The distribution of the unlabeled data $P(X)$ has to carry information relevant to the estimation of $P(Y|X)$.

SSL in Causal and Anti-Causal settings

The task is to predict Y from X



causal setting: predict effect from cause
e.g., predict splice sites from DNA sequence

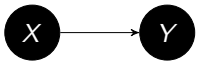


anticausal setting: predict cause from effect
e.g, breast tumor classification,
image segmentation

Schölkopf, Janzing, Peters, Sgouritsa, Zhang, Mooij: On causal and anticausal learning, ICML 2012

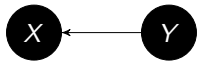
SSL in Causal and Anti-Causal settings

The task is to predict Y from X



causal setting: predict effect from cause
e.g., predict splice sites from DNA sequence

SSL pointless because $P(X)$ contains no information about $P(Y|X)$



anticausal setting: predict cause from effect
e.g, breast tumor classification, image segmentation

SSL can help because $P(X)$ and $P(Y|X)$ contain information about each other

Schölkopf, Janzing, Peters, Sgouritsa, Zhang, Mooij: On causal and anticausal learning, ICML 2012

Known SSL assumptions link $P(X)$ to $P(Y|X)$

- *SSL smoothness assumption*: $E(Y|X)$ should be smooth in regions where $P(X)$ is large.
- *Cluster assumption*: points lying in the same cluster are likely to have the same Y.
- *Low density separation*: The decision boundary should lie in a region where $P(X)$ is small.

The above assumptions can indeed be viewed as **linking properties of $P(X)$ to properties of $P(Y|X)$**.

Results: Meta-analysis

We didn't perform new experiments, instead checked our hypothesis analyzing results of other papers.

1. Dataset categorization as:
   - *Anticausal/Confounded*: (a) at least one X_i is an effect of Y, or (b) at least one X_i and Y are confounded.
   - *Causal*: Y is the effect of the X_i's.
   - *Unclear*: incomplete documentation or lack of domain knowledge.
2. Check performance of semi-supervised vs. supervised learning in each category.

Semi-supervised classification: 8 benchmark datasets

Table 1. Categorization of eight benchmark datasets of Section 5 (Semi-supervised classification) as Anticausal/Confounded, Causal or Unclear

Category	Dataset	Reason of categorization
Anticausal/ Confounded	g241c	The class causes the 241 features.
	g241d	The class (binary) and the features are confounded by a variable with 4 states.
	Digit1	The positive or negative angle and the features are confounded by the variable of continuous angle.
	USPS	The class and the features are confounded by the 10-state variable of all digits.
	COIL	The six-state class and the features are confounded by the 24-state variable of all objects.
Causal	SecStr	The amino acid is the cause of the secondary structure.
Unclear	BCI, Text	Unclear which is the cause and which the effect.

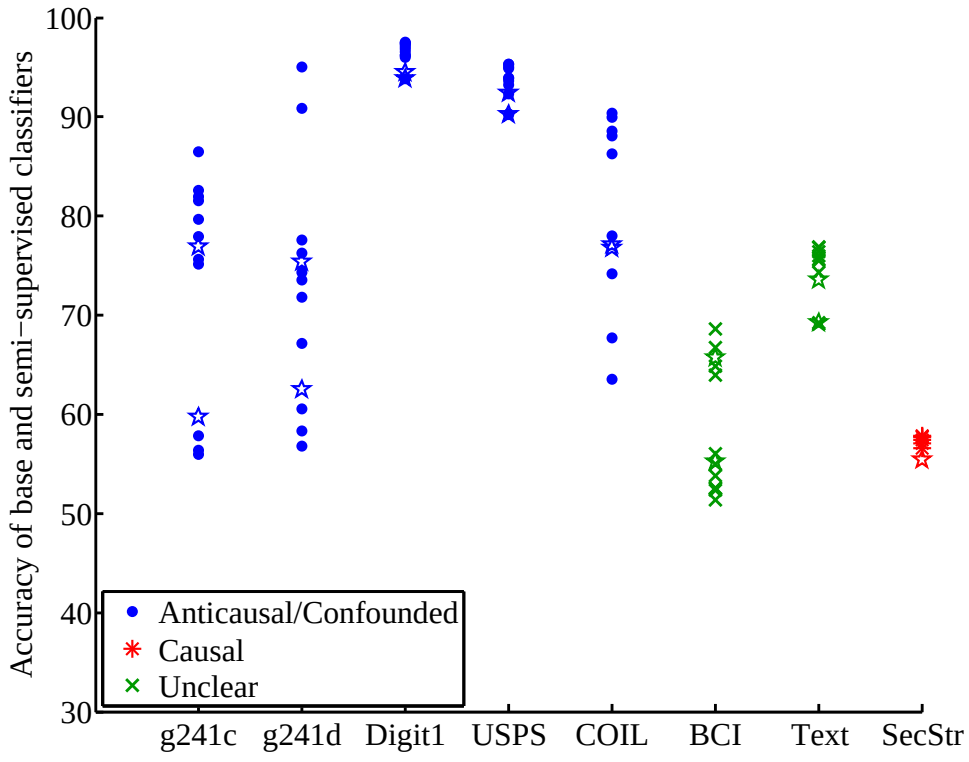
[O. Chapelle *et al.*, 2006] Semi-Supervised Learning.

Semi-supervised classification: 8 benchmark datasets

Table 1. Categorization of eight benchmark datasets of Section 5 (Semi-supervised classification) as Anticausal/Confounded, Causal or Unclear

Category	Dataset	Reason of categorization
Anticausal/ Confounded	g241c	The class causes the 241 features.
	g241d	The class (binary) and the features are confounded by a variable with 4 states.
	Digit1	The positive or negative angle and the features are confounded by the variable of continuous angle.
	USPS	The class and the features are confounded by the 10-state variable of all digits.
	COIL	The six-state class and the features are confounded by the 24-state variable of all objects.
Causal	SecStr	The amino acid is the cause of the secondary structure.
Unclear	BCI, Text	Unclear which is the cause and which the effect.

[O. Chapelle *et al.*, 2006] Semi-Supervised Learning.



Comparison of 11 SSL methods with the base classifiers 1-NN and SVM (star).

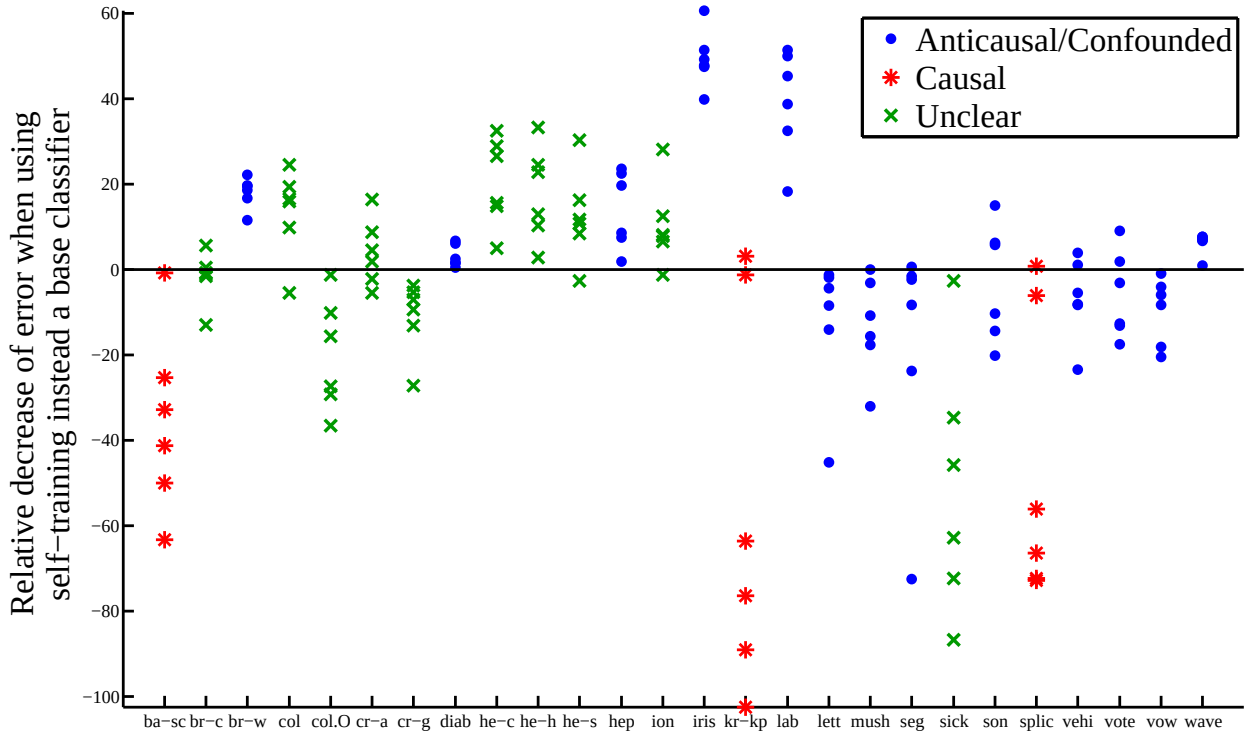
Semi-supervised classification: 26 UCI datasets

Table 2. Categorization of 26 UCI datasets of Section 5 (Semi-supervised classification) as Anticausal/Confounded, Causal or Unclear

Category	Dataset	Reason of categorization
Anticausal/ Confounded	breast-w	The class of the tumor (benign or malignant) causes some of the features of the tumor (e.g., thickness, size, shape etc.).
	diabetes	Whether or not a person has diabetes affects some of the features (e.g., glucose concentration, blood pressure), but also is an effect of some others (e.g. age, number of times pregnant).
	hepatitis	The class (die or survive) and many of the features (e.g., fatigue, anorexia, liver big) are confounded by the presence or absence of hepatitis. Some of the features, however, may also cause death.
	iris	The size of the plant is an effect of the category it belongs to.
	labor	Cyclic causal relationships: good or bad labor relations can cause or be caused by many features (e.g., wage increase, number of working hours per week, number of paid vacation days, employer's help during employee 's long term disability). Moreover, the features and the class may be confounded by elements of the character of the employer and the employee (e.g., ability to cooperate).
	letter	The class (letter) is a cause of the produced image of the letter.
	mushroom	The attributes of the mushroom (shape, size) and the class (edible or poisonous) are confounded by the taxonomy of the mushroom (23 species).
	segment	The class of the image is the cause of the features of the image.
	sonar	The class (Mine or Rock) causes the sonar signals.
	vehicle	The class of the vehicle causes the features of its silhouette.
	vote	This dataset may contain causal, anticausal, confounded and cyclic causal relations. E.g., having handicapped infants or being part of religious groups in school can cause one's vote, being democrat or republican can causally influence whether one supports Nicaraguan contras, immigration may have a cyclic causal relation with the class. Crime and the class may be confounded, e.g., by the environment in which one grew up.
	vowel	The class (vowel) causes the features.
	waveform-5000	The class of the wave causes its attributes.
Causal	balance-scale	The features (weight and distance) cause the class.
	kr-vs-kp	The board-description causally influences whether white will win.
	splice	The DNA sequence causes the splice sites.
Unclear	breast-cancer, colic, colic.ORIG, credit-a, credit-g, heart-c, heart-h, heart-statlog, ionosphere, sick	In some of these datasets, it is unclear whether the class label has been generated or defined based on the features (e.g., Ionoshpere, Credit Approval, Sick).

[Y. Guo *et al.*, 2006] An extensive empirical study on semi-supervised learning.

Semi-supervised classification: 26 UCI datasets



Comparison of self-training to its corresponding 6 base classifiers.

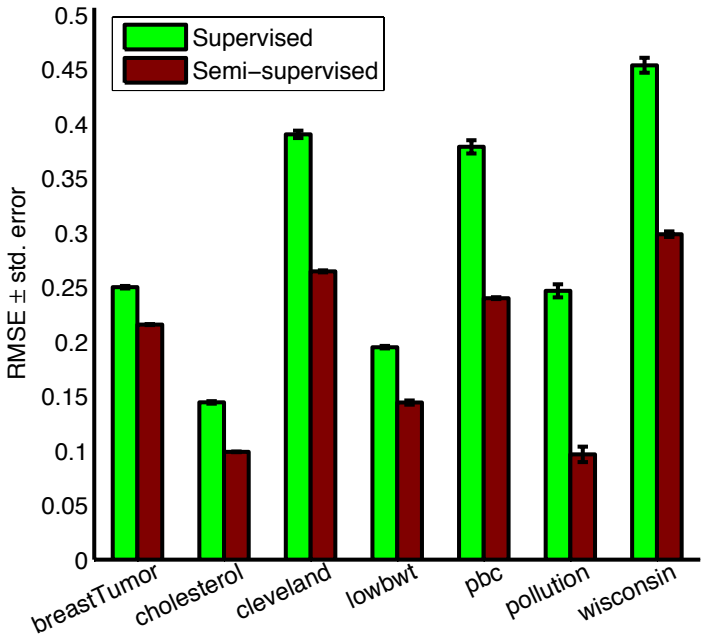
Semi-supervised regression: 31 UCI datasets

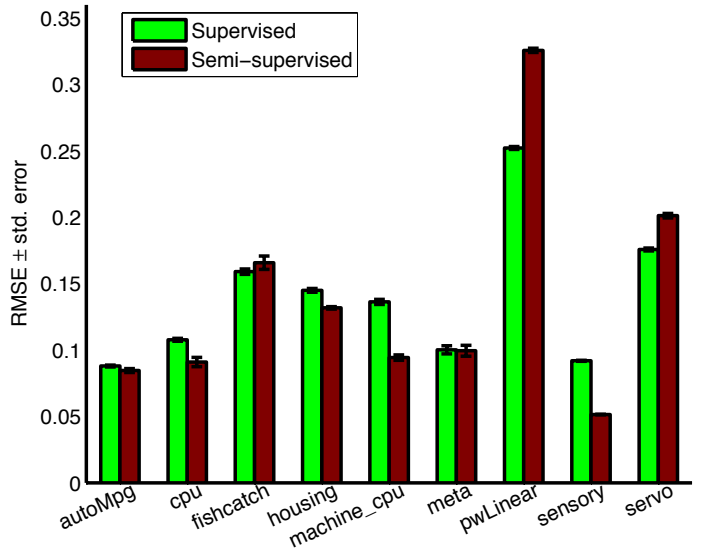
Table 3. Categorization of 31 datasets of Section 5 (Semi-supervised regression) as Anticausal/Confounded, Causal or Unclear

Category	Dataset	Target variable	Reason of categorization
Anticausal/ Confounded	breastTumor	tumor size	causing predictors such as inv-nodes and deg-malig
	cholesterol	cholesterol	causing predictors such as resting blood pressure and fasting blood sugar
	cleveland	presence of heart disease in the patient	causing predictors such as chest pain type, resting blood pressure, and fasting blood sugar
	lowbwt	birth weight	causing the predictor indicating low birth weight
	pbc	histologic stage of disease	causing predictors such as Serum bilirubin, Prothrombin time, and Albumin
	pollution	age-adjusted mortality rate per 100,000	causing the predictor number of 1960 SMSA population aged 65 or older
	wisconsin	time to recur of breast cancer	causing predictors such as perimeter, smoothness, and concavity
Causal	autoMpg	city-cycle fuel consumption in miles per gallon	caused by predictors such as horsepower and weight
	cpu	cpu relative performance	caused by predictors such as machine cycle time, maximum main memory, and cache memory
	fishcatch	fish weight	caused by predictors such as fish length and fish width
	housing	housing values in suburbs of Boston	caused by predictors such as pupil-teacher ratio and nitric oxides concentration
	machine_cpu	cpu relative performance	see remark on "cpu"
	meta	normalized prediction error	caused by predictors such as number of examples, number of attributes, and entropy of classes
	pwLinear	value of piecewise linear function	caused by all 10 involved predictors
	sensory	wine quality	caused by predictors such as trellis
	servo	rise time of a servomechanism	caused by predictors such as gain settings and choices of mechanical linkages
Unclear	auto93 (target: midrange price of cars); bodyfat (target: percentage of body fat); autoHorse (target: price of cars); autoPrice (target: price of cars); baskball (target: points scored per minute); cloud (target: period rainfalls in the east target); echoMonths (target: number of months patient survived); fruitfly (target: longevity of mail fruitflies); pharynx (target: patient survival); pyrim (quantitative structure activity relationships); sleep (target: total sleep in hours per day); stock (target: price of one particular stock); strike (target: strike volume); triazines (target: activity); veteran (survival in days)		

[U. Brefeld *et al.*, 2006] Efficient co-regularized least squares regression.

Semi-supervised regression: 31 UCI datasets





Comparison of coRLSR (brown) with regular RLSR (green).

Conclusion of the meta-study

- Accuracy is not significantly improved in causal datasets but the performance is increased in most of the anticausal/confounded datasets.

Conclusions

- independence of causal conditionals seems to be the leading principle for causal inference
- faithfulness is based on the same idea
- we propose algorithmic independence, but in practice we rely on computable notions of independence (including faithfulness?)
- recent causal inference algorithms already use computable notions of independence other than faithfulness
- failure of semi-supervised learning also defines independence

Thank you for your attention!