The principle of independent conditionals

and the Arrow of Time

Dominik Janzing

Max Planck Institute for Intelligent Systems

Tübingen, Germany

Note:

this talk is about causal inference in classical statistics, nothing quantum

- there's so much to discover still
- quantum analog of our results could be exciting

Moreover, content of this talk is inspired by quantum information theory...

What I learned from quantum information theory

The fact that a field exists since decades
does not imply
that the most elementary questions are already solved

People started understanding entanglement in $\mathbb{C}^2 \otimes \mathbb{C}^2$
after almost one century of quantum theory...

“All questions about finite dimensional quantum systems are trivial”

A quantum theory postdoc in 1996

Some work on quantum causality

1. D.J. and T. Decker: How much is a quantum controller controlled by the controlled system? AAECC 2008.
2. D.J. and T. Beth: On the potential influence of quantum noise on measuring effectiveness in clinical trials. IJQI 2006.
3. D.J: Is there a physically universal cellular automaton or Hamiltonian? ArXiv 2010.

. . . but I won't talk about it

Goal of causal inference

predict the effect of interventions on the world
from **passive** observations only

$\Rightarrow$ requires assumptions
$\Rightarrow$ no purely mathematical justification possible

Example: tricky link between statistical relations and causal relations

Paradox result of a recent study

- coffee drinking increases life expectancy

  (causal statement)

- coffee drinking is negatively correlated with life expectancy

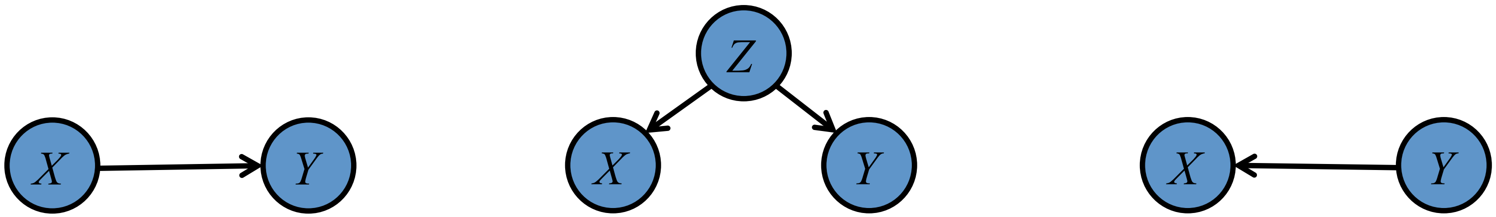
  (statistical statement)

explanation: coffee drinkers die earlier **despite** drinking coffee because they tend to have unhealthy habits in addition

Reichenbach's Principle of Common Cause

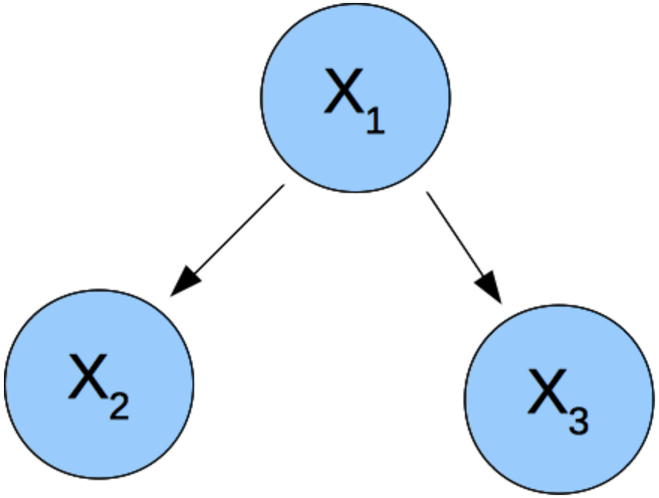
postulates that every statistical dependence has a causal explanation:

If two quantities X and Y are statistically dependent then at least one of the following cases is true:



Causal inference from statistical data: formal setting

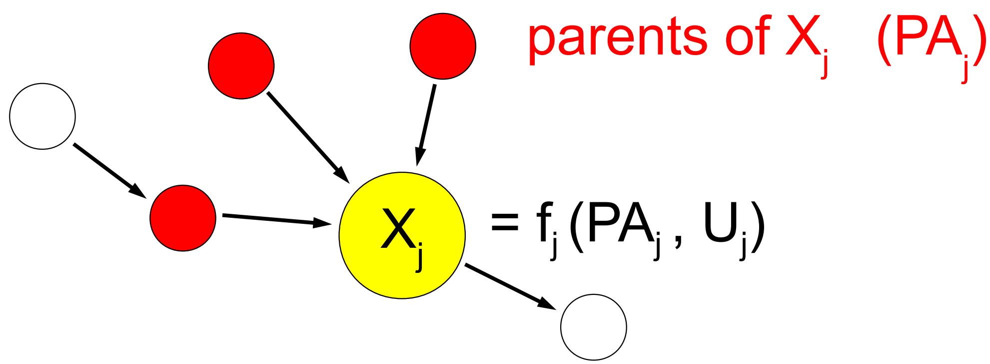
- given the random variables $X_1, \ldots, X_n$ and a data matrix of observations
- infer the causal directed acyclic graph (DAG)



Postulate 1: Functional causal model

(Pearl 2000)

- every variable X_j is a function of its parents (direct causes) and an unobserved noise term U_j
- the U_j are jointly statistically independent

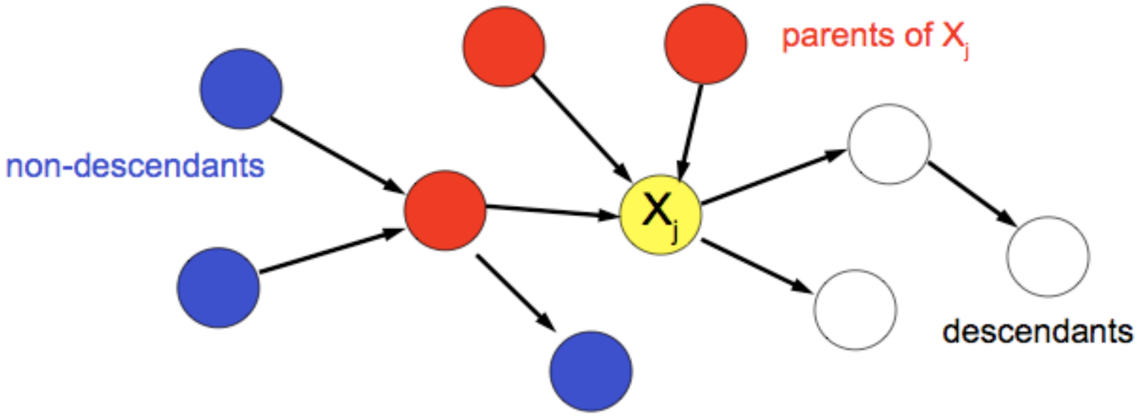


"local hidden variable model"

Markov Condition

Theorem: the functional model implies the following 3 equivalent conditions:

- **Local Markov condition:** X_j statistically independent of non descendants, given its parents



- **Global Markov condition:** d-separation implies conditional independence
- **Factorization:** $p(X_1, \ldots, X_n) = \prod_{j=1}^{n} p(X_j | PA_j)$

(equivalence subject to technical conditions, see Lauritzen 1996)

Interpretation of 3 Versions

- **Local Markov Condition:**
  every information exchange with non-descendants involves the parents

- **Global Markov Condition:**
  characterizes the set of all independences implied by the local version

- **Factorization:**
  each causal conditional $p(x_j|pa_j)$ represents a causal mechanism

(ideas for quantum Markov conditions: Poulin & Leifer 2008,
compare also causal/acausal quantum states by Leifer & Spekkens 2007)

Postulate 2: Causal Faithfulness

(Spirtes, Glymour, Scheines 1993)

p is called faithful relative to G if only those independences hold true that are implied by the Markov condition, i.e.,

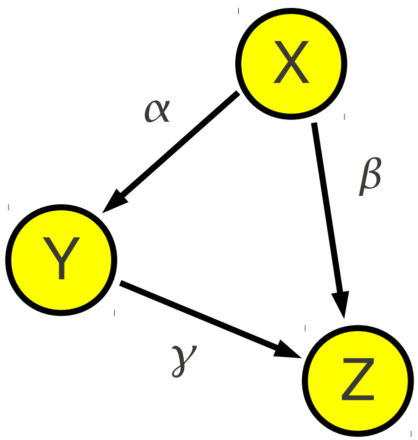
$$X \perp\!\!\!\perp Y \,|Z \quad \Rightarrow \quad Z \text{ d-separates } X \text{ and } Y$$

Recall: Markov condition reads

$$X \perp\!\!\!\perp Y \,|Z \quad \Leftarrow \quad Z \text{ d-separates } X \text{ and } Y$$

Unfaithful distributions, Example (1)

cancellation of direct and indirect influence in linear models



$$
\begin{aligned}
X &= U_X \\
Y &= \alpha X + U_Y \\
Z &= \beta X + \gamma Z + U_Z
\end{aligned}
$$

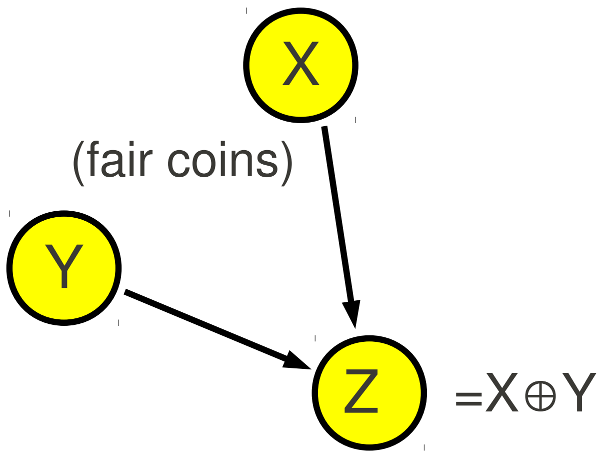
with independent noise terms U_X, U_Y, U_Z

$$\beta + \alpha\gamma = 0 \quad \Rightarrow \quad X \perp\!\!\!\perp Z$$

Unfaithful distributions, Example (2)

binary causes with XOR as effect

- for $p(X), p(Y)$ uniform: $X \perp\!\!\!\perp Z, Y \perp\!\!\!\perp Z$.
  i.e., unfaithful (since X, Z and Y, Z are connected in the graph).
- for $p(X), p(Y)$ non-uniform: $X \not\perp\!\!\!\perp Z, Y \not\perp\!\!\!\perp Z$.
  i.e., faithful



unfaithfulness considered unlikely because it only occures for non-generic parameter values

Conditional-independence based causal inference

(Spirtes, Glymour, Scheines and Pearl)

causal Markov condition + causal faithfulness:

- accept only those DAGs as causal hypotheses for which

$$Z \text{ d-separates } X \text{ and } Y \quad \Leftrightarrow \quad X \perp\!\!\!\perp Y \,|Z$$

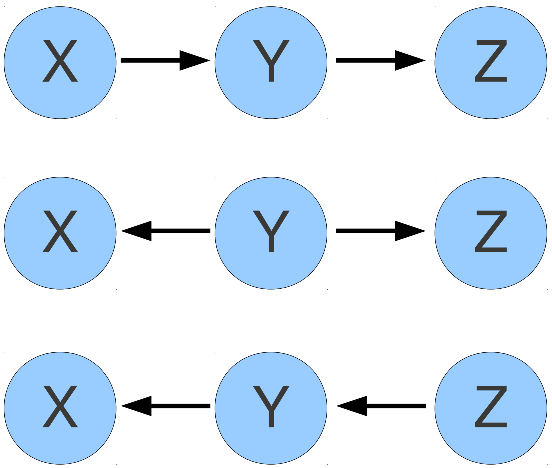
- identifies causal DAG up to Markov equivalence class (DAGs that imply the same conditional independences)

Markov Equivalence Class

Theorem (Verma and Pearl, 1990): two DAGs are Markov equivalent iff they have the same skeleton and the same v-structures.

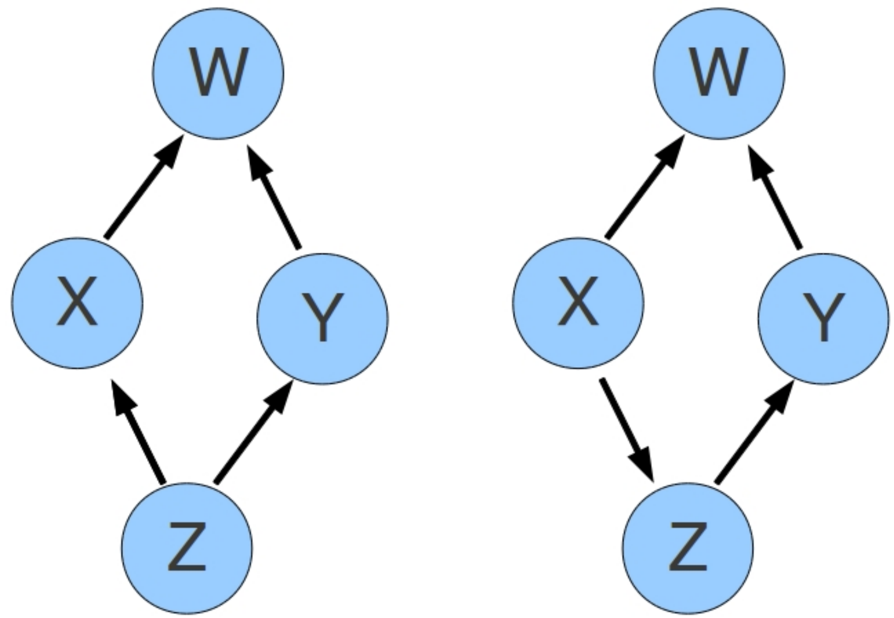
- **skeleton:** corresponding undirected graph
- **v-structure:** substructure $X \to Y \leftarrow Z$ with no edge between X and Z

Markov equivalent DAGs (1)



- same skeleton, no v-structure
- only independence: $X \perp\!\!\!\perp Z \mid Y$

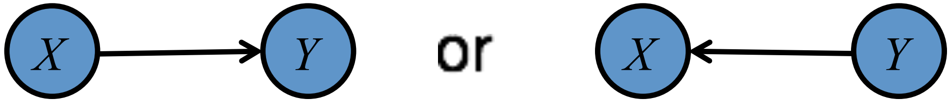
Markov equivalent DAGs (2)



same skeleton, v-structure at W

Limitations of Independence-based Approach

- Markov equivalence classes can be large
- Most elementary problem unsolvable:



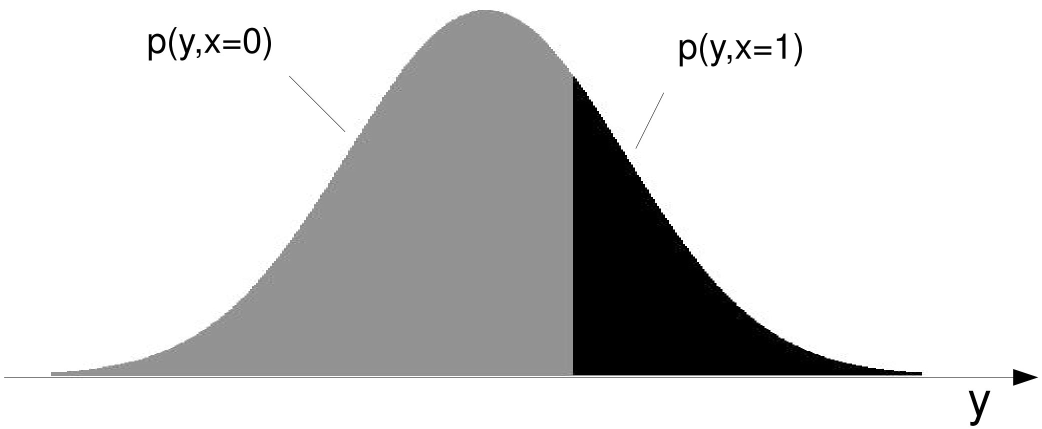
- probability distributions contain interesting information other than independences

$\Rightarrow$ new inference rules desirable

Example:

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

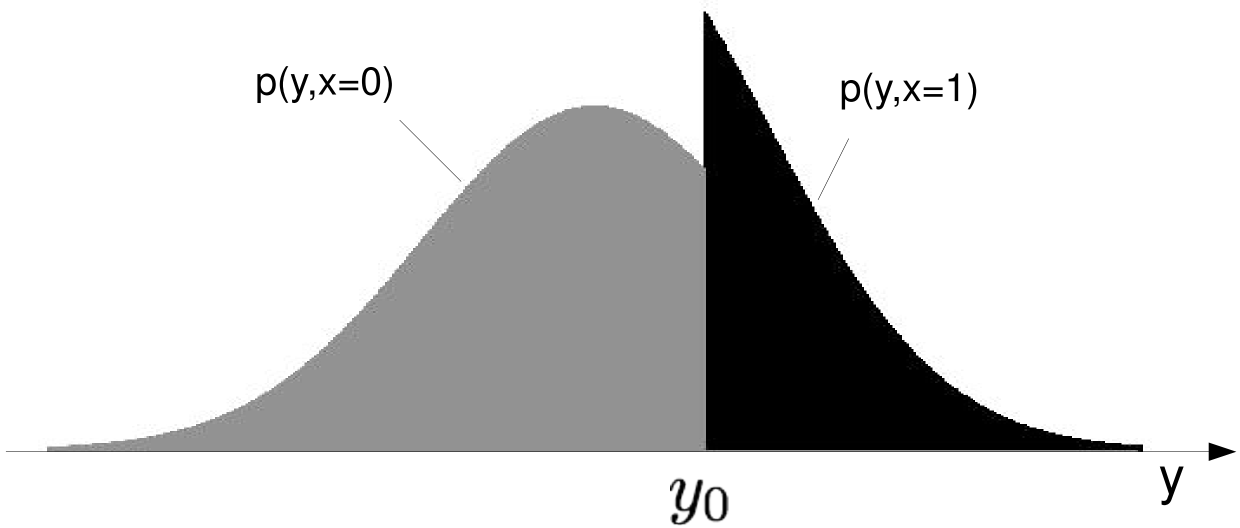


- $Y \to X$ is plausible: simple thresholding mechanism

- $X \to Y$ requires a strange mechanism: $P(Y|X = 0)$ and $P(Y|X = 1)$ are truncated Gaussians

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

this happens if we change $P(X)$ to $P'(X)$



- $P(X)$ is the unique distribution that generates Gaussian output
- $P(X)$ seems 'to know' $P(Y|X)$

Goal: invent an inference rule that rejects $X \to Y$ for this reason

Algorithmic independence of conditionals (IC)

(Lemeire & Dirkx 2006, Janzing & Schölkopf 2010, Lemeire & Janzing 2012)

New postulate for causal inference:

- if $X \to Y$ then $P(X)$ and $P(Y|X)$ are algorithmically independent
- the shortest description of $P(X,Y)$ is given by describing $P(X)$ and $P(Y|X)$ separately
- violated in the example above

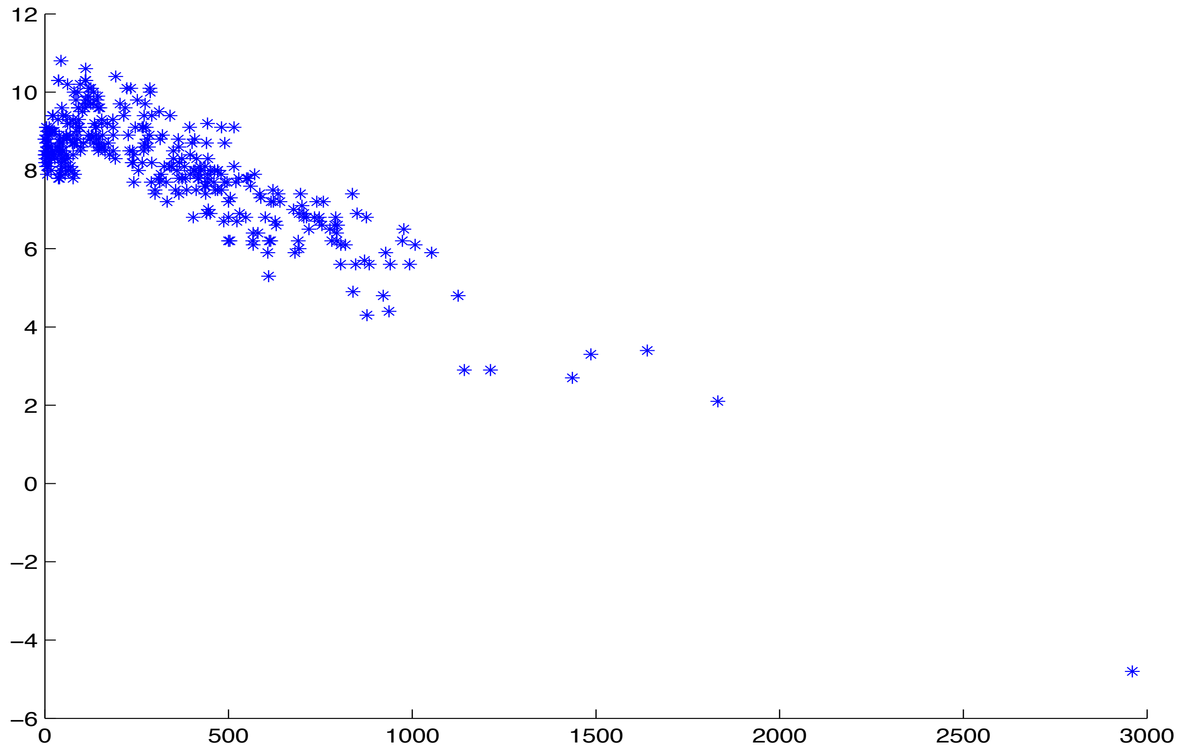
- actually phrased for n variables

Raises 3 questions:

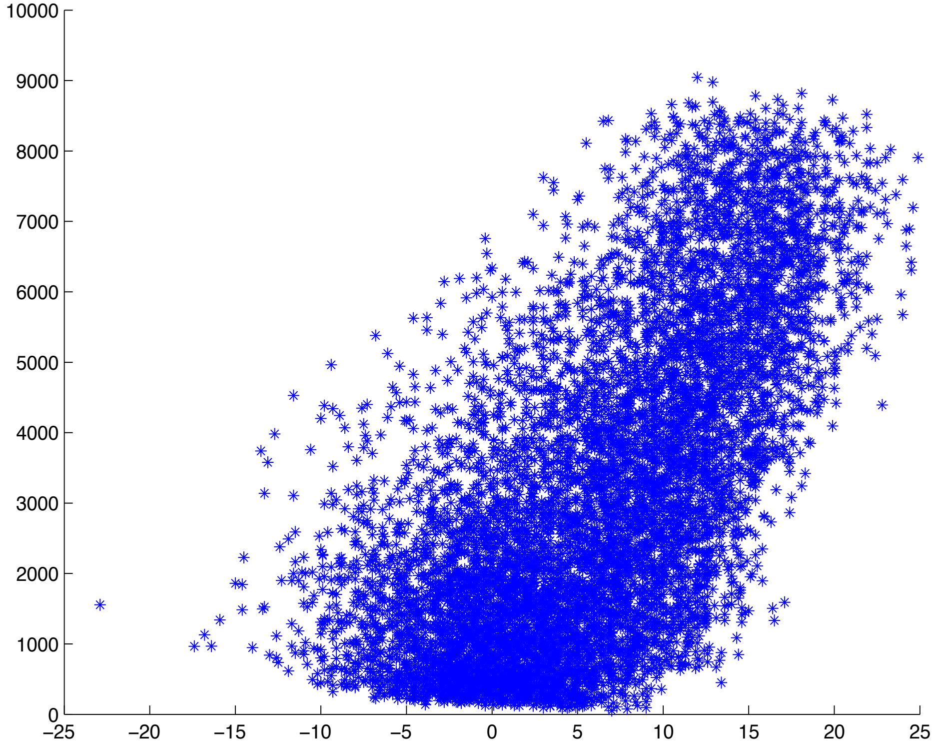
1. are these asymmetries observable for real data?
2. why is description length related to causality?
3. what's the relation to the arrow of time?

   (asymmetry between cause and effect should be related to asymmetry between past and future)

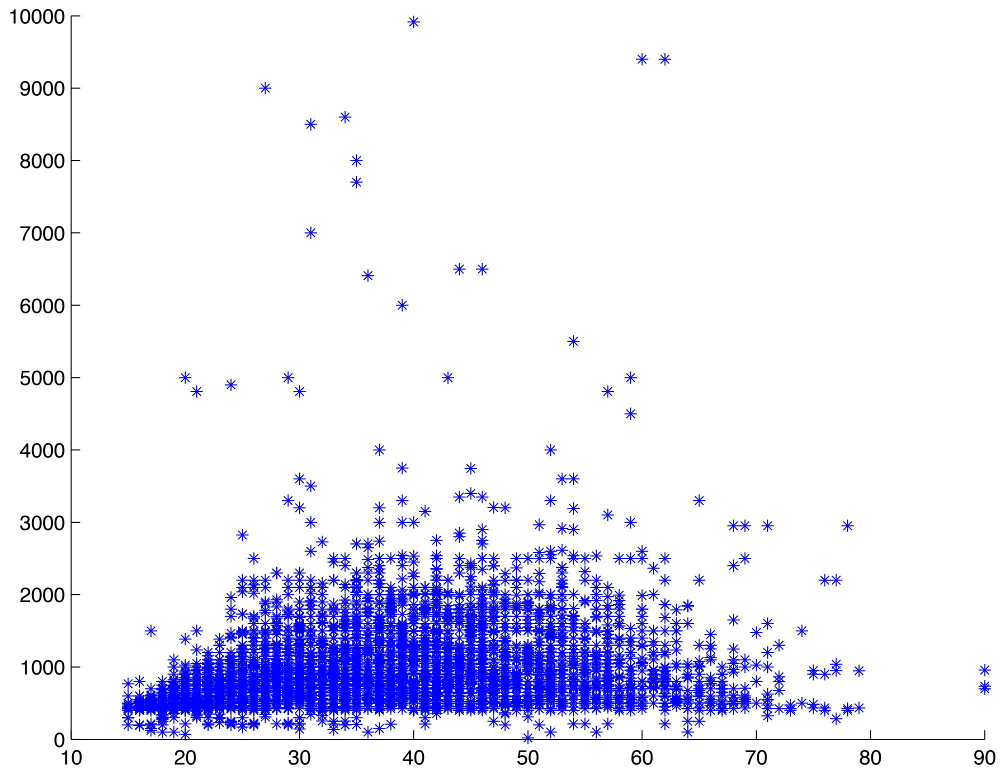
Infer cause and effect from scatter plot



Infer cause and effect from scatter plot



Infer cause and effect from scatter plot



Novel causal inference algorithms

implement rudimentary versions of the above principle

- Linear additive noise models: Kano, Shimizu, 2004
- Additive noise models: Hoyer, DJ, . . . NIPS 2008,
- Post-nonlinear models: Zhang, Hyvarinen, UAI 2009.
- Information-Geometric Causal Inference: Daniusis, DJ, . . . , UAI 2010, DJ et al, AI 2012.

achieve classification rates of about 70-80 % on real data

Why is causality related to description length?

Forget about statistics for the moment –

how do we draw causal conclusions in real life?

Causal inference for individual objects

Janzing & Schölkopf 2010

Similarities between single objects also indicate causal relations:

However, if similarities are too simple there need not be a common cause:

Consider a binary sequence

Experiment:
2 persons are instructed to write down a string with 1000 digits

Result:
Both write 1100100100001111110110101010001…
(all 1000 digits coincide)

The naive statistician concludes...

"There must be an agreement between the subjects"

- correlation coefficient 1 (between digits) is highly significant for sample size 1000 !
- reject statistical independence, assume causal relation

Some other mathematician recognizes...

11.0010010000111111011010101001...

$= \pi$

- subjects may have come up with this number independently because it follows from a simple law
- superficially strong similarities are not necessarily significant if the pattern is too simple

How do we measure complexity

of patterns/objects?

Kolmogorov complexity

(Kolmogorov, Chaitin, Solomonoff)

of a binary string x

- $K(x) :=$ length of the shortest program with output x (on a Turing machine)
- interpretation: number of bits required to describe the rule that generates x
- equality "=" is always understood up to string-independent additive constants

- $K(x)$ is uncomputable
- probability-free definition of information content

Conditional Kolmogorov complexity

- $K(y \mid x)$: length of the shortest program that generates y from x
- number of bits required for describing y if x is given

- $K(y|x^*)$: length of the shortest program that generates y from the shortest description of x
- note: x can be generated from its shortest description but not vice versa because there is no algorithmic way to find the shortest compression

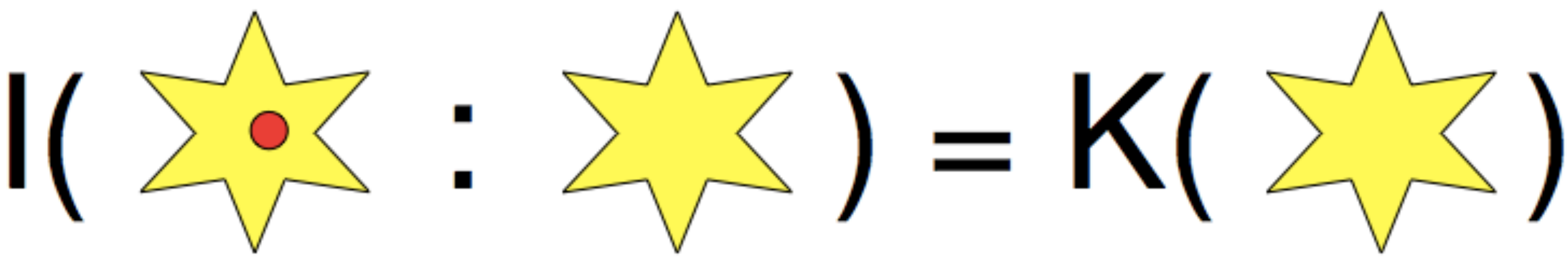
Algorithmic mutual information

(Chaitin, Gacs)

Information of x about y

- $$\begin{aligned} I(x:y) &:= K(x) + K(y) - K(x,y) \\ &= K(x) - K(x\,|\,y^*) = K(y) - K(y\,|\,x^*) \end{aligned}$$
- Interpretation: number of bits saved when compressing x, y jointly rather than independently
- Algorithmic independence $x \perp\!\!\!\perp y$: $\iff$ $I(x:y) = 0$

Algorithmic mutual information (example)



Conditional algorithmic mutual information

Information that x has on y (and vice versa) when z is given

- $I\left(x:y\,|\,z\right) := K\left(x\,|\,z\right) + K\left(y\,|\,z\right) - K\left(x,y\,|\,z\right)$
- Analogy to statistical mutual information:

$$I\left(X:Y\,|\,Z\right) = S\left(X\,|\,Z\right) + S\left(Y\,|\,Z\right) - S\left(X,Y\,|\,Z\right)$$

- Conditional algor. independence $x \perp\!\!\!\perp y\,|\,z :\Longleftrightarrow I\left(x:y\,|\,z\right) = 0$

Algorithmic analog of Reichenbach's principle

- Reichenbach argued that every **statistical** dependence indicates a causal relation

- We argued that every **algorithmic** dependence indicates a causal relation

Question:

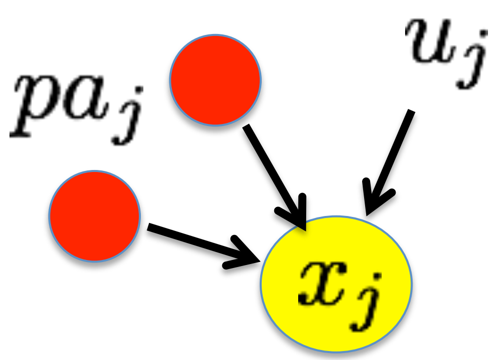
Do **conditional** algorithmic (in)dependences
tell us s.th.
about the causal DAG?

Algorithmic model of causality

(Janzing & Schölkopf IEEE TIT 2010)

Given n causality related strings $x_1, \ldots, x_n$

- each x_j is computed from its parents pa_j and an unobserved string u_j from a Turing machine T



$$= T(pa_j, u_j)$$

- all u_j are algorithmically independent
- u_j describe the mechanism that generate x_j from pa_j
- u_j are he analog of noise in the statistical functional model

Relation to Church-Turing Principle

- **Church-Turing:**

  every mechanism in nature can be simulated by a program on a universal Turing machine

- **Algorithmic causal model:**

  independent causal mechanisms are simulated by algorithmically independent programs

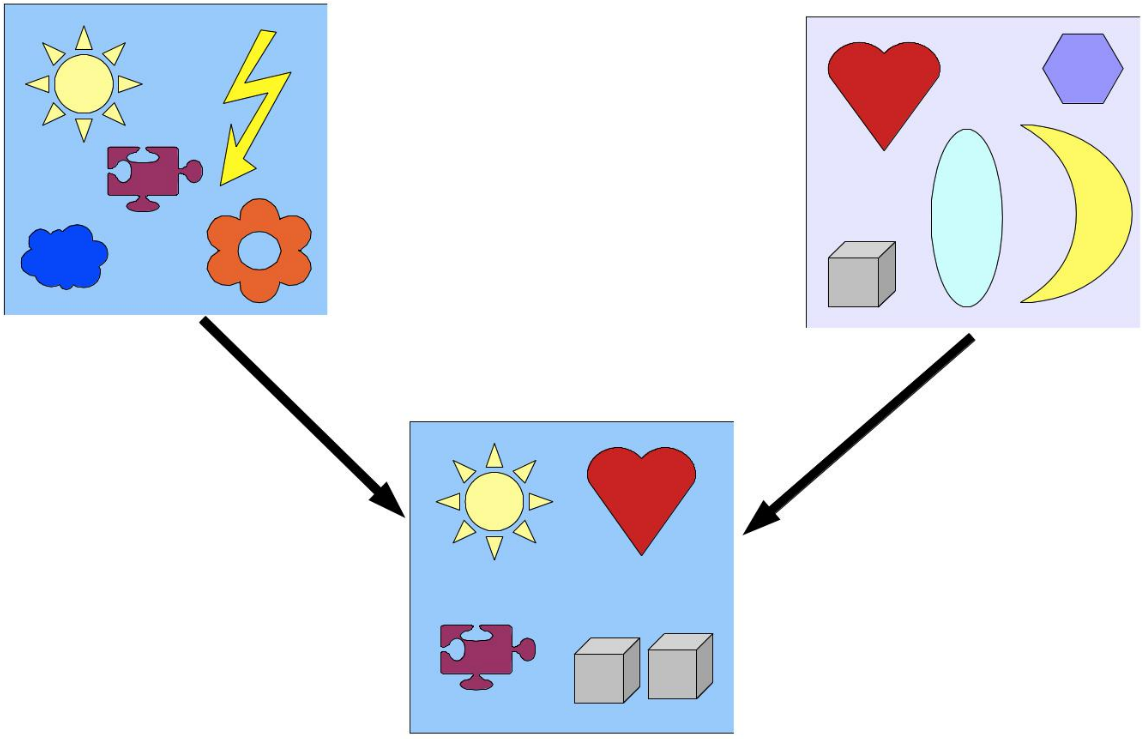
Theorem:

(Janzing & Schölkopf IEEE TIT 2010)

the algorithmic model implies the following 3 equivalent conditions

- **Local Markov:** $x_j \perp\!\!\!\perp nd_j \,|pa_j^*$
- **Global Markov:** d-separation implies algorithmic independence
- **Additivity:** $K(x_1, \ldots, x_n) = \sum_{j=1}^{n} K(x_j|pa_j^*)$

Example: 3 carpet designs



Statistical vs. algorithmic causal Markov condition

- **Nodes:** random variables vs. single objects (represented by binary words)
- **Dependence measure:** Shannon mutual information vs. algorithmic mutual information
- **Justification:** function model vs. algorithmic functional model

algorithmic Markov condition more general:

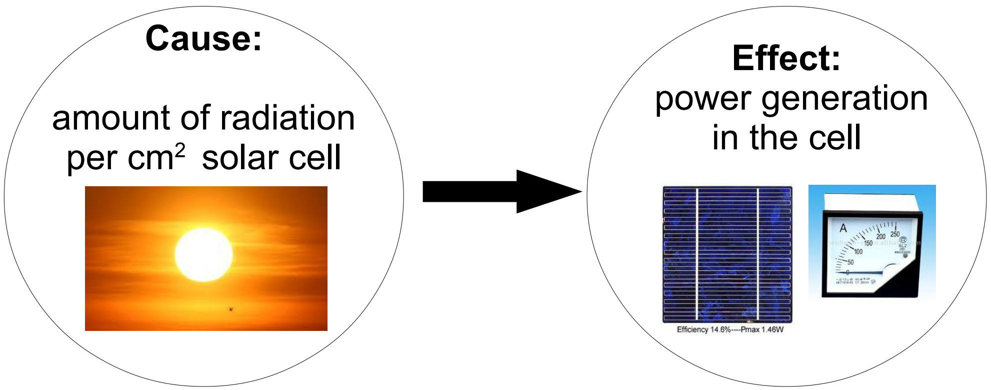
- if objects $x_1, \ldots, x_n$ denote k iid samples from joint distribution $P(X_1, \ldots, X_n)$ then algorithmic information per k converges to Shannon entropy
- limit, however, blurs non-statistical dependences

Revisiting algorithmic independence of conditionals

- if $X \rightarrow Y$ then $P(X)$ and $P(Y|X)$ contain no algorithmic information about each other
- follows from algorithmic Markov condition if we believe that $P(X)$ and $P(Y|X)$ are generated by causally unrelated mechanisms

(why) do we believe that nature generates $P(\text{cause})$ and $P(\text{effect}|\text{cause})$ independently?

Justifying independence of conditionals



Changes affecting $P(\text{cause})$

- move the solar cell to a more/less shady place
- mount it at a different angle to the sun

Changes affecting $P(\text{effect}|\text{cause})$

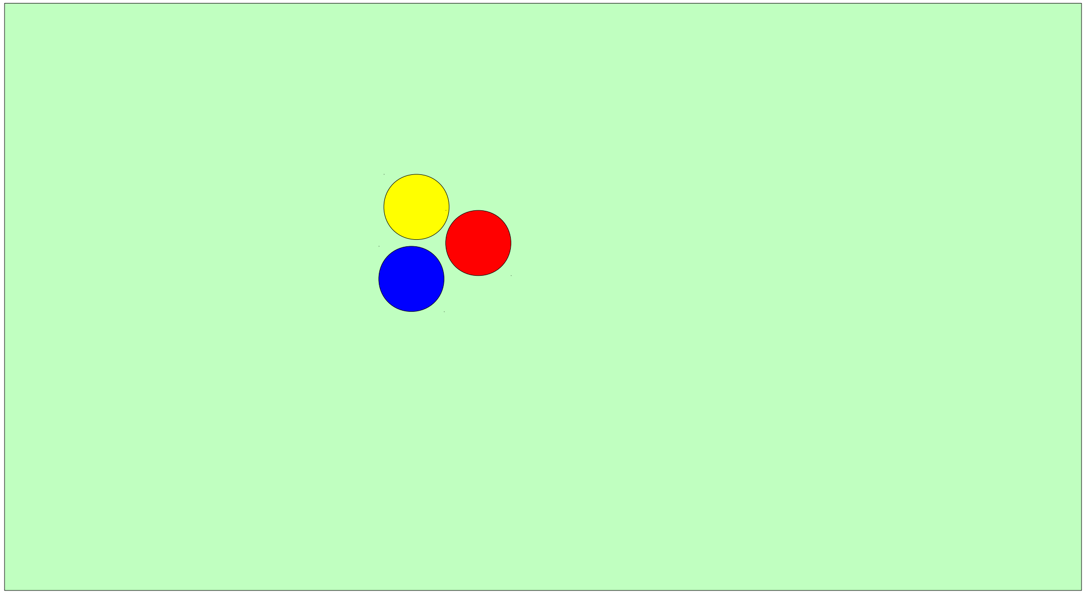
- use less/more efficient cells
- change temperature

Justifying independence of conditionals

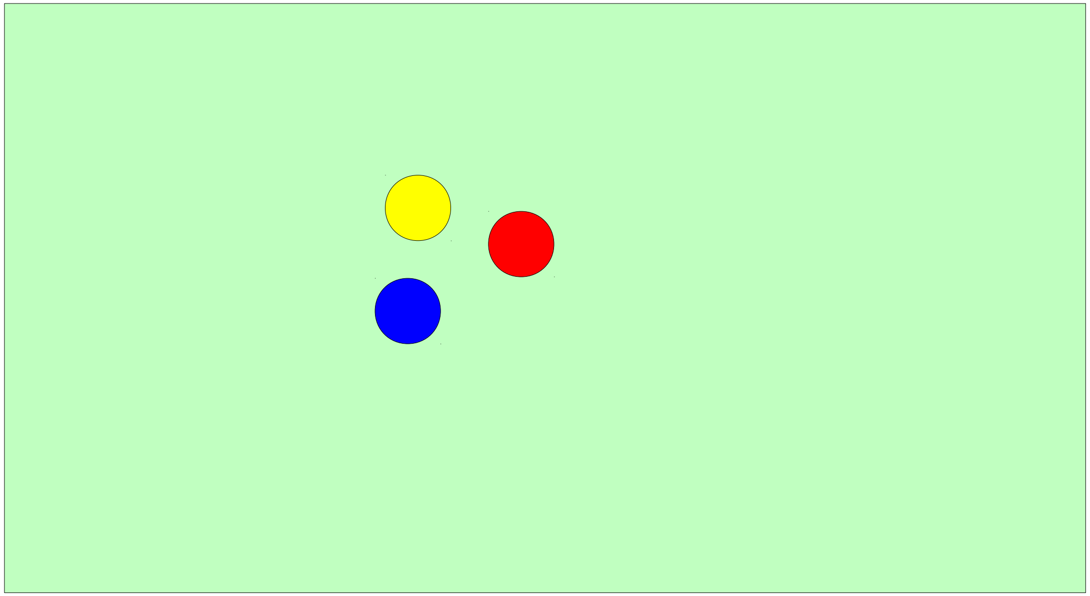
changes under operations / different background conditions:

- some operations change $P(\mathrm{cause})$ only
- some change $P(\mathrm{effect}|\mathrm{cause})$ only
- some change both
- hard to find operations that change $P(\mathrm{effect})$ without affecting $P(\mathrm{cause}|\mathrm{effect})$ or vice versa

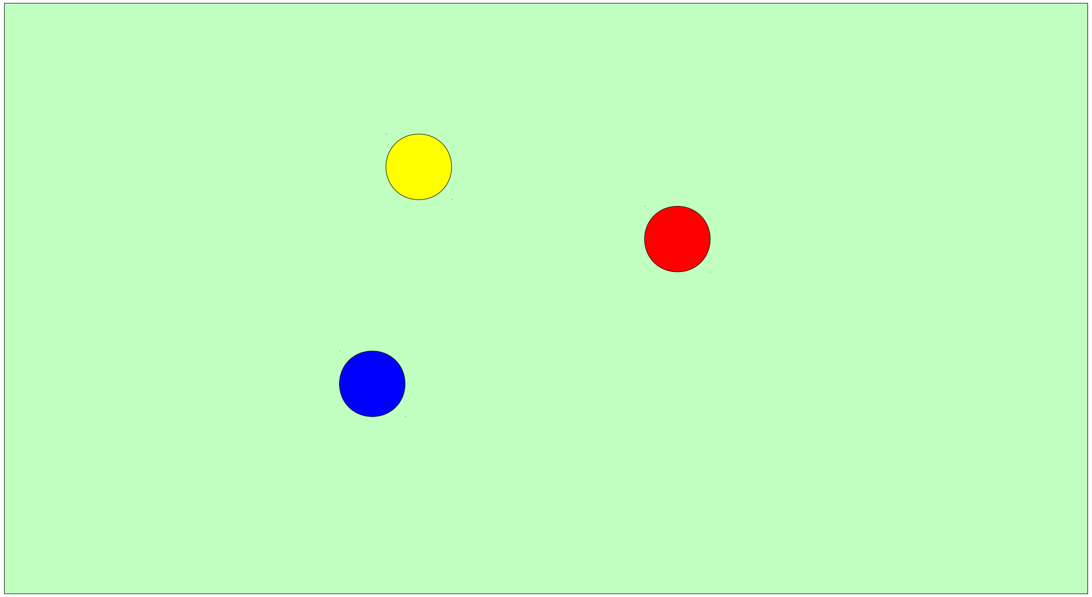
Arrow of time



Arrow of time



Arrow of time



Arrow of time

- **typical closed system dynamics:**

$$\text{simple state} \to \text{complex state}$$

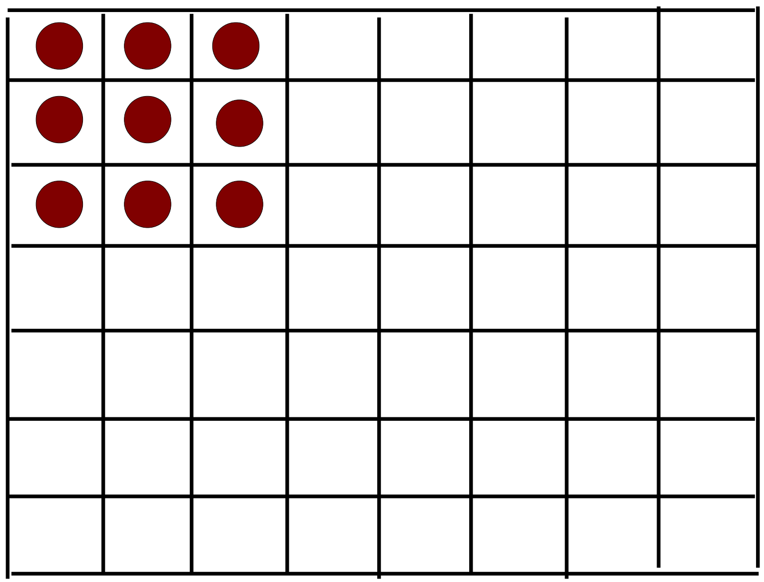
- **unlikely:**

$$\text{complex state} \to \text{simple state}$$

(thermodynamic entropy = Kolmogorov complexity?)

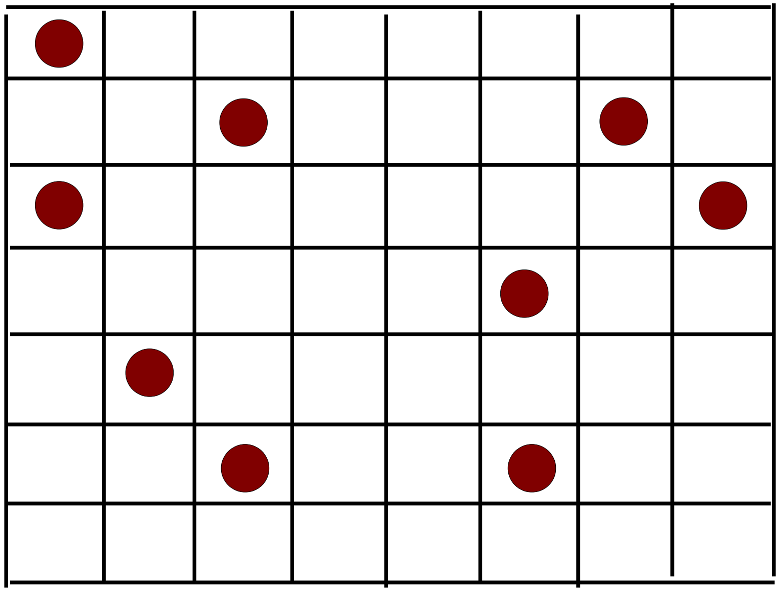
Zurek: Algorithmic randomness and physical entropy, PRA 1989

Discrete dynamical system



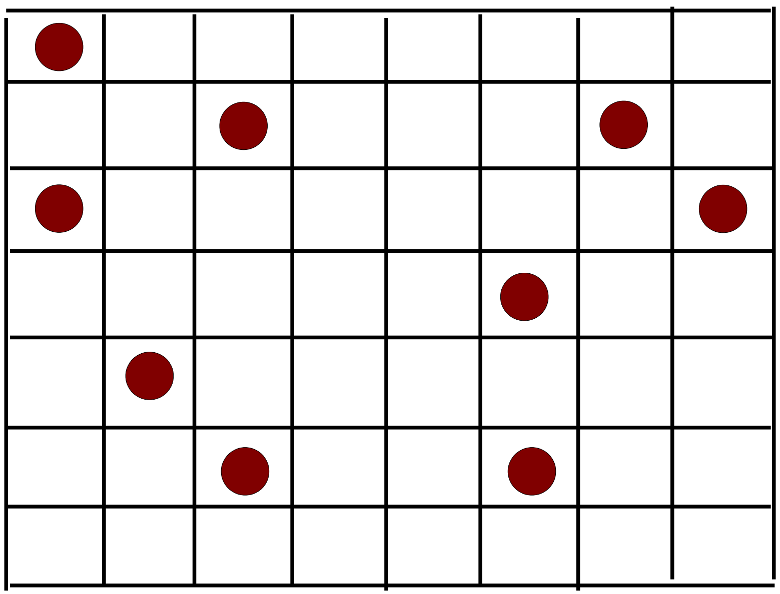
initial state s with low description length $K(s)$

Discrete dynamical system



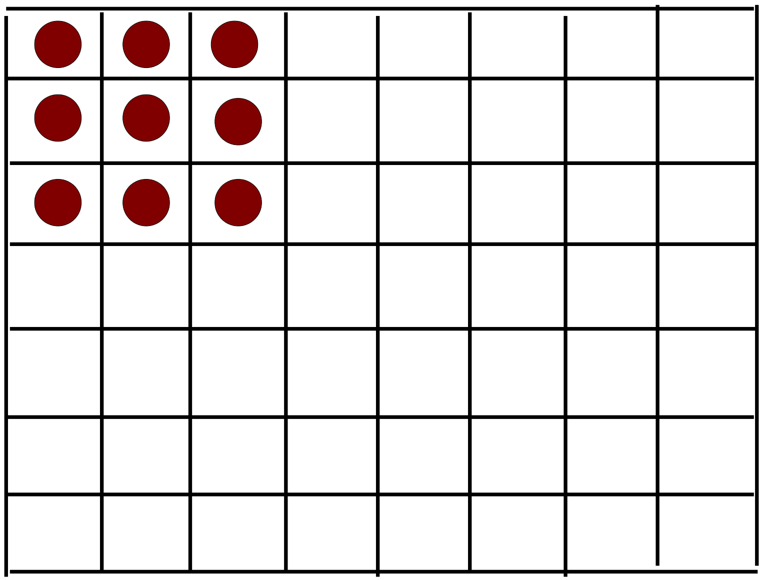
state $D(s)$ with large description length after applying bijective dynamics D

Time reversed scenario



initial state s with large description length $K(s)$

Time reversed scenario



final state $D(s)$ with low description length $K(D(s))$

Independence between input and dynamics induces Arrow of Time

initial state s, bijective dynamics D

- assume $K(D(s)) < K(s)$

- then $K(s|D) \stackrel{+}{=} K(D(s)|D) \stackrel{+}{\leq} K(D(s)) < K(s)$

- hence, s contains algorithmic information about D

Independence between input and dynamics more general than Arrow of Time

Postulate: $K(s|D) \stackrel{+}{=} K(s)$ (also for non-bijective D)

- implication $K(D(s)) \geq K(s)$ only holds for bijective D
- lower bounds for $K(D(s))$ in terms of non-bijectivity of D
- postulate makes also sense if D is probabilistic

- replace $s \equiv P(\mathrm{cause})$ and $D \equiv P(\mathrm{effect}|\mathrm{cause})$

Wrong approach to distinguishing cause and effect

"Variable with lower entropy is the cause" (motivated by thermodynamics)

- Cause may be continuous, effect binary
- entropy depends on scaling
- application of non-linear functions tends to decrease entropy

Take home messages

- **new inference principle:**

  algorithmic independence between a causal mechanism and its input

- Related to **Arrow of Time**

- justified by our general theory of inferring causal relations from algorithmic dependences

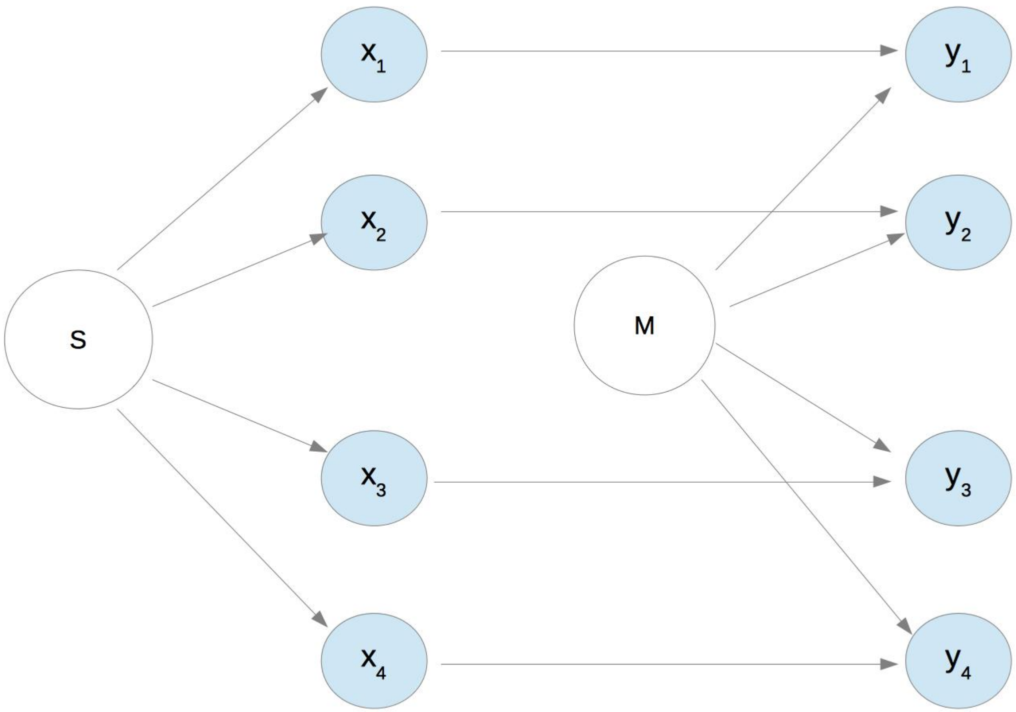
Thanks for your attention!

References

- D.J. and B. Schölkopf: Causal inference using the algorithmic Markov condition, IEEE TIT 2010
- Schölkopf, DJ, . . . : On causal and anticausal learning, ICML 2012.
- D.J. and B. Steudel: Justifying additive-noise-based causal discovery via algorithmic information theory, OSID 2010.
- J. Lemeire and D.J.: Replacing causal faithfulness with the algorithmic independence of conditionals, Minds & Machines 2012.
- D. Janzing: On the entropy production of time series with unidirectional linearity, J. Stat. Phys. 2010.

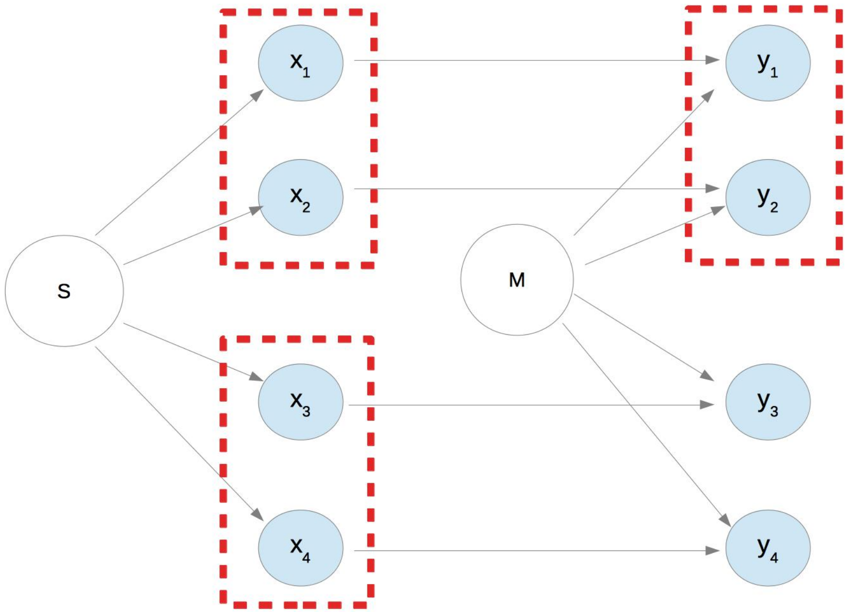
Probability free version:

Observations $(x_1, y_1), \ldots, (x_m, y_m)$ from $P(X, Y)$ define causal structure with $n = 2m + 2$ objects:



Probability free version:

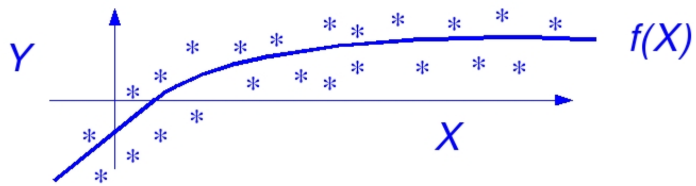
Algorithmic Markov condition implies e.g. $x_3, x_4 \perp\!\!\!\perp y_1, y_2 \,|x_1, x_2$



- additional x-values do not help for predicting y from x
- semisupervised learning does not help in causal direction
  Schölkopf, Janzing,. . . 2012

Causal inference with additive noise models

(Hoyer, Janzing, Mooij, Peters, Schölkopf 2008)



- Assume the effect is a function of the cause up to an additive noise term that is independent of the cause:

$$Y = f(X) + U_Y \quad \text{with } U_Y \perp\!\!\!\perp X$$

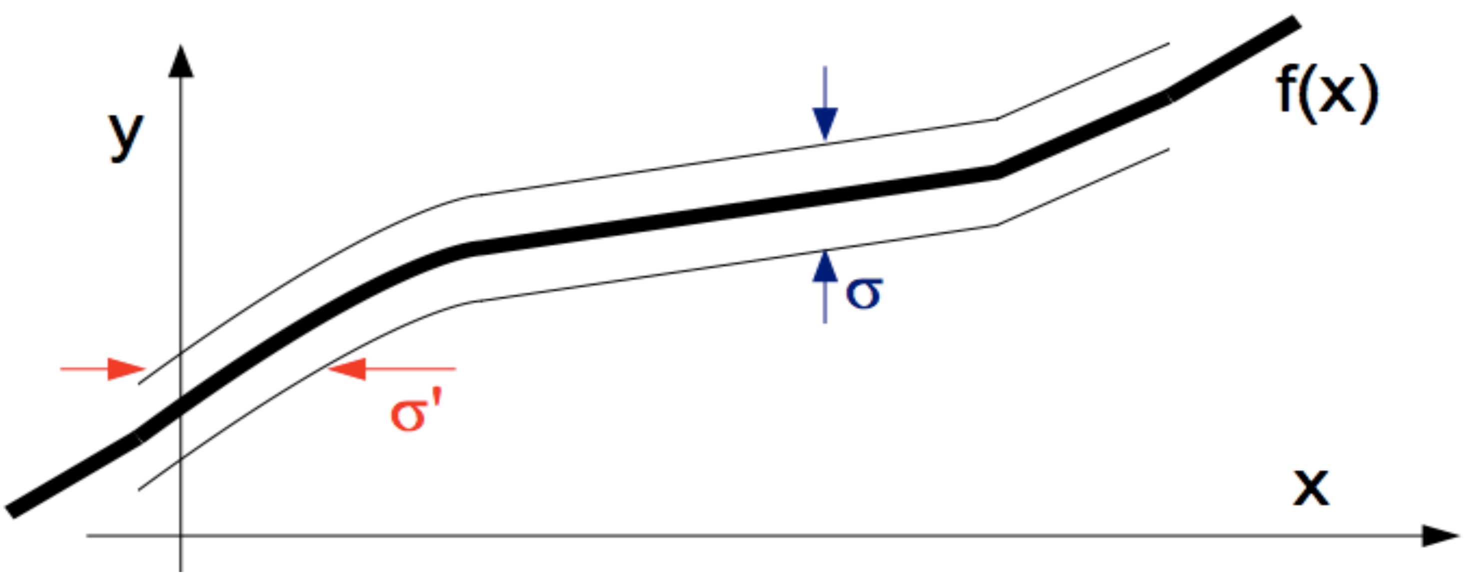
- there is, in the generic case, no model

$$X = g(Y) + U_X \quad \text{with } U_X \perp\!\!\!\perp Y \,,$$

even if f is invertible (proof non-trivial)

Intuition:

- assume noise of bounded range
- additive noise model implies range of Y around f is constant
- for nonlinear f, range of X around backward function non-constant



Inference rule

Infer $X \to Y$ if there is an additive noise model from X to Y
but not vice versa

Implementation:

- compute a function f as non-linear regression of Y on X function of
- compute the residual
$$U := Y - f(X)$$
- check whether U and X are statistically independent

Results:

- performed above chance level on our real-world cause-effect pairs $\sim 70\%$
- ratio of correct answers tends to 1 for conservative decisions

Justification of AN-based inference via IC condition
(Janzing & Steudel 2010)

Assume there is an additive noise model from X to Y

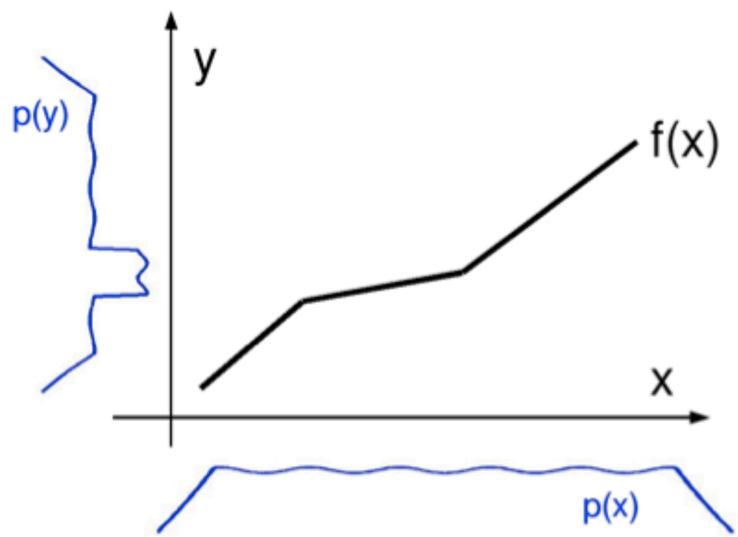
- $P(Y)$ and $P(X|Y)$ satisfy the equation

$$\frac{\partial^2}{\partial y^2} \log p(y) = -\frac{\partial^2}{\partial y^2} \log p(x|y) - c\frac{\partial^2}{\partial x \partial y} \log p(x|y)$$

- $P(Y)$ can "almost" be computed from $P(X|Y)$
- $Y \rightarrow X$ is unlikely because $P(Y)$ contains algorithmic information about $P(X|Y)$ unless $P(Y)$ is simple

Inferring deterministic causal relations

- 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}



Daniusis, DJ, . . . : UAI 2010