Title: The Algorithmic Markov Condition as a Foundation of Causal Inference

Date: Apr 10, 2012 03:30 PM

URL: http://pirsa.org/12040060

Abstract: I present our work on inferring causality in the classical world and encourage the audience to think about possible generalizations to the quantum world. Statistical dependences between observed quantities X and Y indicate a causal relation, but it is a priori not clear whether X caused Y or Y caused X or there is a common cause of both. It is widely believed that this can only be decided if either one is able to do interventions on the system, or if X and Y are part of a larger set of variables. In the latter case, conditional statistical independences contain some information on causal directions, formalized by the Causal Markov Condition on directed acyclic graphs. Contrary to this belief, we have shown that empirical joint distributions of just two variables often indicate the causal direction. The observed asymmetry between cause and effect is, on the one hand, related to the thermodynamic arrow of time. On the other hand, it can be derived from a new principle that we have postulated: the Algorithmic Causal Markov Condition, which relates Kolmogorov complexity to causality.
Literature: [1] Janzing, Schoelkopf: Causal inference using the algorithmic Markov condition, IEEE TIT 2010.
[2] Daniusis, Janzing,...: Inferring deterministic causal relations, UAI 2010.
[3] Janzing: On the entropy production of time-series with uni-directional linearity.Journ. Stat. Phys. 2010.

The algorithmic Markov condition as a foundation of causal inference

Dominik Janzing
Max Planck Institute for Intelligent Systems
Tübingen, Germany

Perimeter Institute Waterloo, 10 April 2012

One may argue whether the topic of this talk is **physics**,

if not, I think it should become part of it...

Correlation and Causality

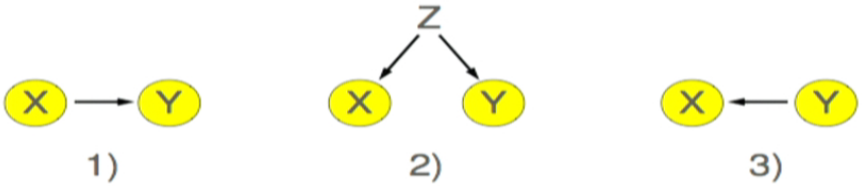
- Recently I read that children who visited a farm in their early childhood are less likely to get allergies

Many statistical observations raise causal questions.

Some people claim: "Correlation says nothing about causality" – is this true?

Reichenbach's principle of common cause (1956)

If two variables X and Y are statistically dependent then either

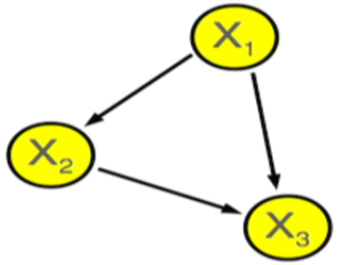


- in case 2) Reichenbach postulated $X \perp\!\!\!\perp Y \,|\, Z$.
- every statistical dependence is due to a causal relation, we also call 2) "causal".
- distinction between 3 cases is a key problem in scientific reasoning and the focus of this talk.

Causal inference problem, general form

Spirtes, Glymour, Scheines, Pearl

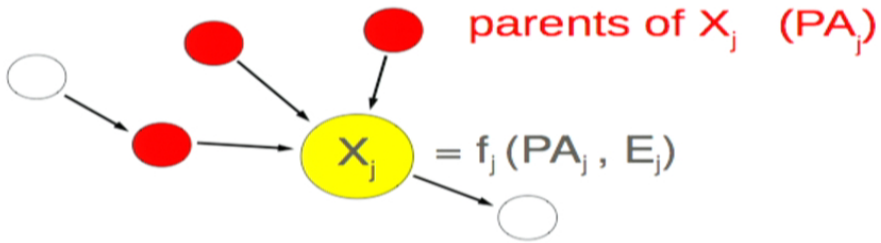
- Given variables $X_1, \ldots, X_n$
- infer causal structure among them from n-tuples iid drawn from $P(X_1, \ldots, X_n)$
- causal structure = directed acyclic graph



Functional model of causality

Pearl

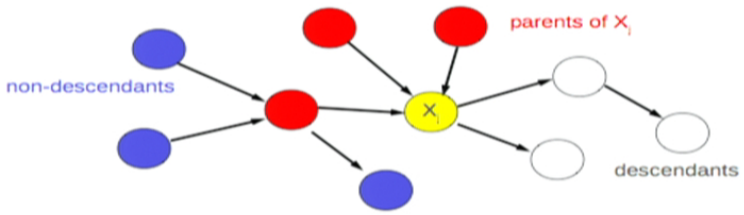
- every node X_j is a function of its parents and an unobserved noise term



- all noise terms E_j are statistically independent (causal sufficiency)
- which properties of $P(X_1, \ldots, X_n)$ follow?

Causal Markov condition (4 equivalent versions) Lauritzen et al, Pearl

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



(information exchange with non-decendants involves parents)

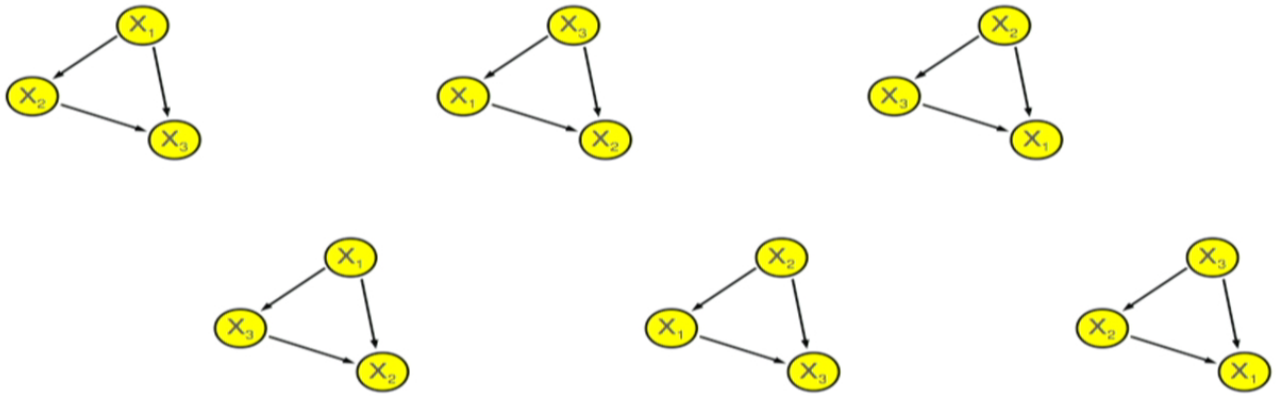
- **global Markov condition:** describes all ind. via d-separation
- **Factorization:** $P(X_1, \ldots, X_n) = \prod_j P(X_j|PA_j)$

(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



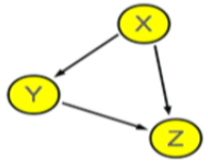
- reasonable rules for prefering simple DAGs required

Causal faithfulness

Spirtes, Glymour, Scheines

Prefer those DAGs for which the Markov condition implies every conditional independence that is observed

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



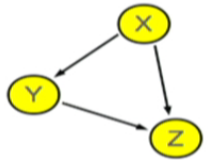
- not faithful, direct and indirect influence compensate
- **Application:** PC and FCI algorithm infer causal structure from conditional statistical independences

Causal faithfulness

Spirtes, Glymour, Scheines

Prefer those DAGs for which the Markov condition implies every conditional independence that is observed

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



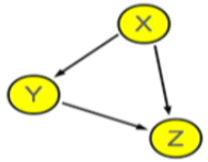
- not faithful, direct and indirect influence compensate
- **Application:** PC and FCI algorithm infer causal structure from conditional statistical independences

Causal faithfulness

Spirtes, Glymour, Scheines

Prefer those DAGs for which the Markov condition implies every conditional independence that is observed

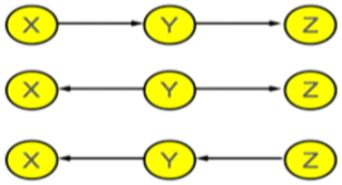
- **Idea:** generic choices of parameters yield faithful distributions
- **Example:** let $X \perp\!\!\!\perp Z$ for the DAG



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

Limitation of independence based approach:

- many DAGs impose the same set of independences

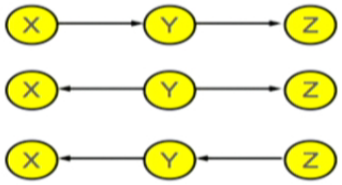


$X \perp\!\!\!\perp Z \,|\, Y$ for all three cases ("Markov equivalent DAGs")

- method fails if there are no conditional independences, unable to infer whether $X \to Y$ or $Y \to X$
- ignores important information:
  only uses yes/no decisions "conditional dependent or not"
  without accounting for the kind of dependences...

Limitation of independence based approach:

- many DAGs impose the same set of independences



$X \perp\!\!\!\perp Z \,|\, Y$ for all three cases ("Markov equivalent DAGs")

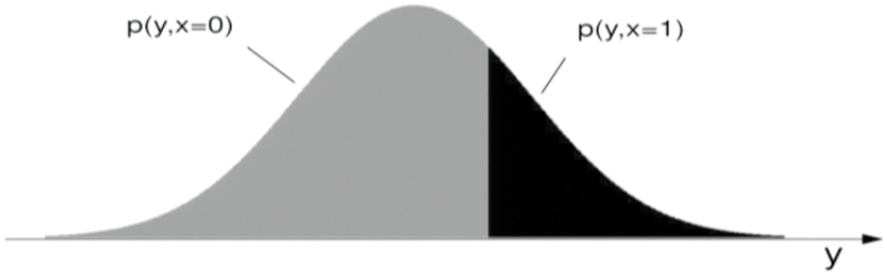
- method fails if there are no conditional independences, unable to infer whether $X \to Y$ or $Y \to X$

- ignores important information:
only uses yes/no decisions "conditional dependent or not"
without accounting for the kind of dependences...

$X \longrightarrow Y$ or $Y \longrightarrow X$? (toy 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.

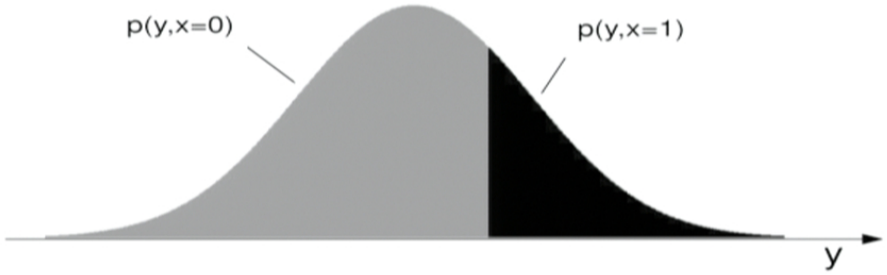


- $Y \longrightarrow X$ is plausible: simple thresholding mechanism
- $X \longrightarrow Y$ requires a strange mechanism: look at $P(Y|X = 0)$ and $P(Y|X = 1)$!

$X \to Y$ or $Y \to X$? (toy example)

Let X be binary and Y real-valued.

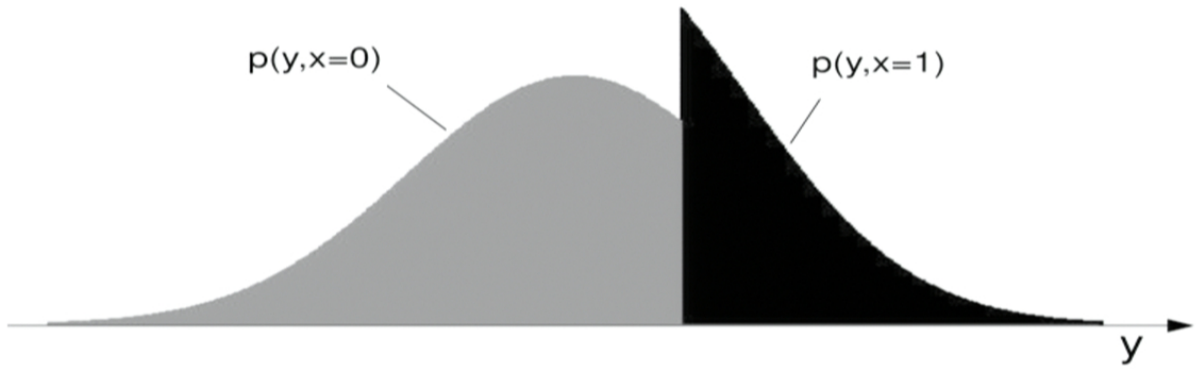
- Let Y be Gaussian and $X = 1$ for all y above some threshold and $X = 0$ otherwise.



- $Y \to X$ is plausible: simple thresholding mechanism
- $X \to Y$ requires a strange mechanism: 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 $P(x)$:



$P(X)$ and $P(Y|X)$ seem to be adjusted to each other!

General idea:

we only accept the hypothesis $X \longrightarrow Y$ if

$P(X)$ and $P(Y|X)$ look like "independently chosen" by nature

Challenge:

Develop practial inference rules and algorithms

that use this assumption for guessing whether $X \longrightarrow Y$ or $Y \longrightarrow X$

Example (1): Linear non-Gaussian models

If

$$Y = \alpha X + E \quad \text{with } E \perp\!\!\!\perp X,$$

we cannot have

$$X = \beta Y + \tilde{E} \quad \text{with } \tilde{E} \perp\!\!\!\perp Y,$$

unless $P(X, Y)$ is bivariate Gaussian

Inference rule : prefer $X \longrightarrow Y$ if this is the case

Kano & Shimizu 2003, graphics inspired by Hoyer

Example (1): Linear non-Gaussian models

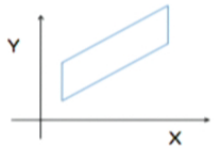
If

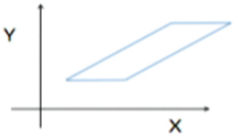
$$Y = \alpha X + E \quad \text{with } E \perp\!\!\!\perp X,$$

we cannot have

$$X = \beta Y + \tilde{E} \quad \text{with } \tilde{E} \perp\!\!\!\perp Y,$$

unless $P(X, Y)$ is bivariate Gaussian





Inference rule : prefer $X \longrightarrow Y$ if this is the case

Kano & Shimizu 2003, graphics inspired by Hoyer

Example (1): Linear non-Gaussian models

If

$$Y = \alpha X + E \quad \text{with } E \perp\!\!\!\perp X,$$

we cannot have

$$X = \beta Y + \tilde{E} \quad \text{with } \tilde{E} \perp\!\!\!\perp Y,$$

unless $P(X, Y)$ is bivariate Gaussian

Inference rule : prefer $X \longrightarrow Y$ if this is the case

Kano & Shimizu 2003, graphics inspired by Hoyer

Example (1): Linear non-Gaussian models

If

$$Y = \alpha X + E \quad \text{with } E \perp\!\!\!\perp X,$$

we cannot have

$$X = \beta Y + \tilde{E} \quad \text{with } \tilde{E} \perp\!\!\!\perp Y,$$

unless $P(X, Y)$ is bivariate Gaussian

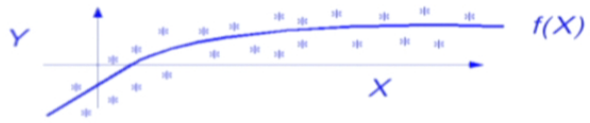
Inference rule : prefer $X \longrightarrow Y$ if this is the case

Kano & Shimizu 2003, graphics inspired by Hoyer

Example (2): Nonlinear additive noise models

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

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



- there will, in the generic case, be no model

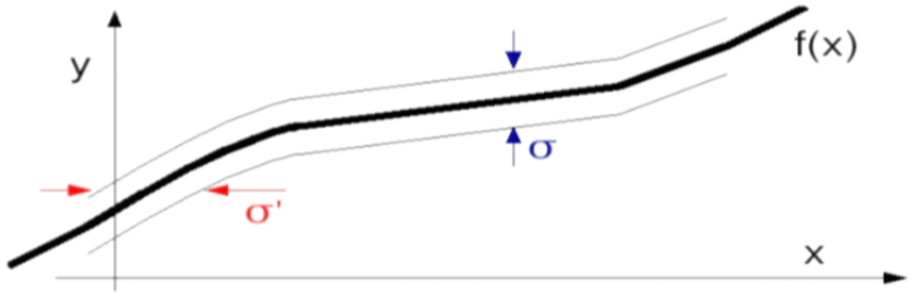
$$X = g(Y) + \tilde{E} \quad \text{with} \quad \tilde{E} \perp\!\!\!\perp Y\,,$$

even if f is invertible! (proof is non-trivial)

Hoyer, DJ,... NIPS 2008

Intuition

- additive noise model from X to Y imposes that the width of noise is constant in x.
- for non-linear f, the width of noise won't be constant in y at the same time.



Causal inference method:

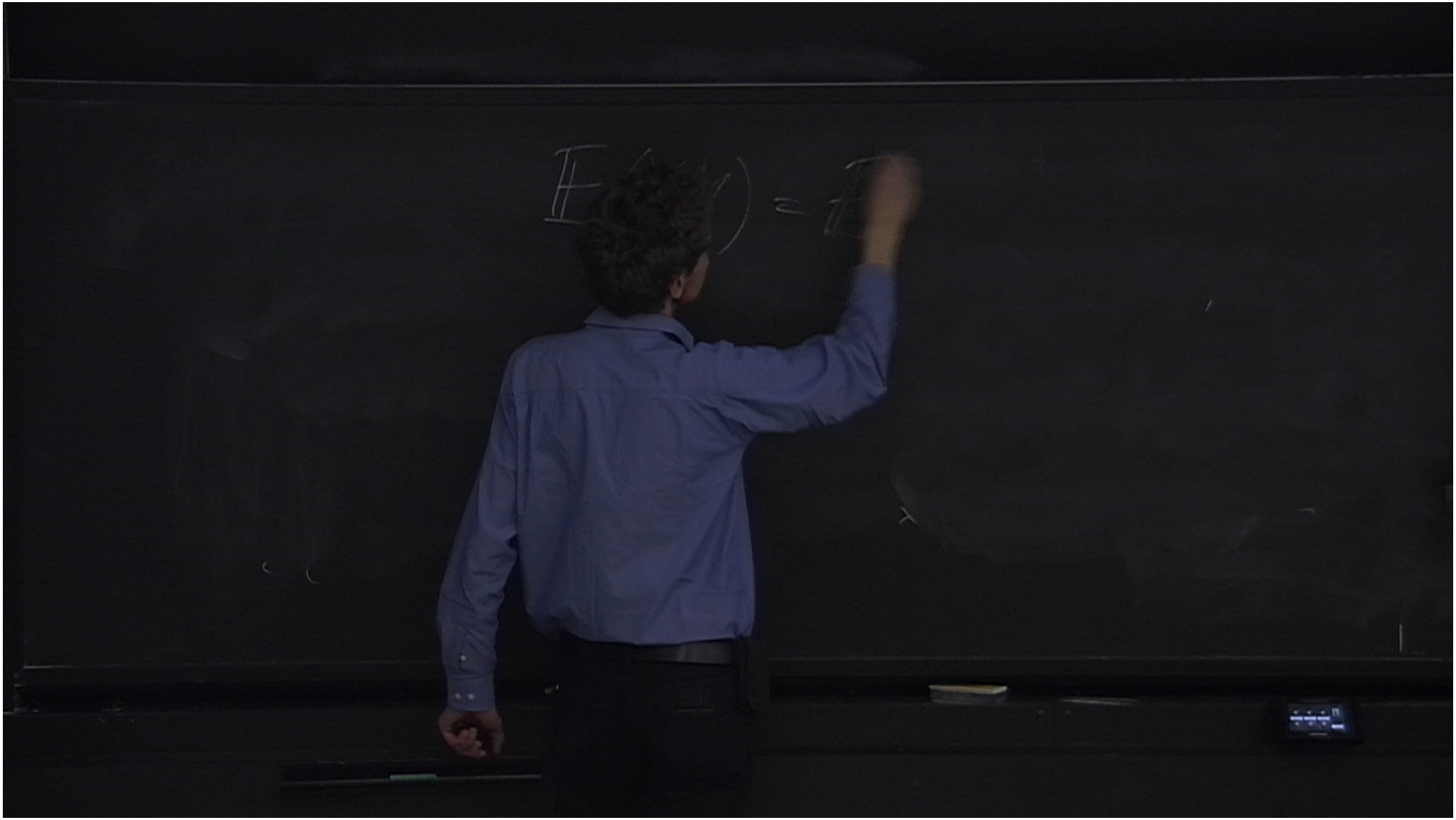
Prefer the causal direction that can better be fit with an additive noise model.

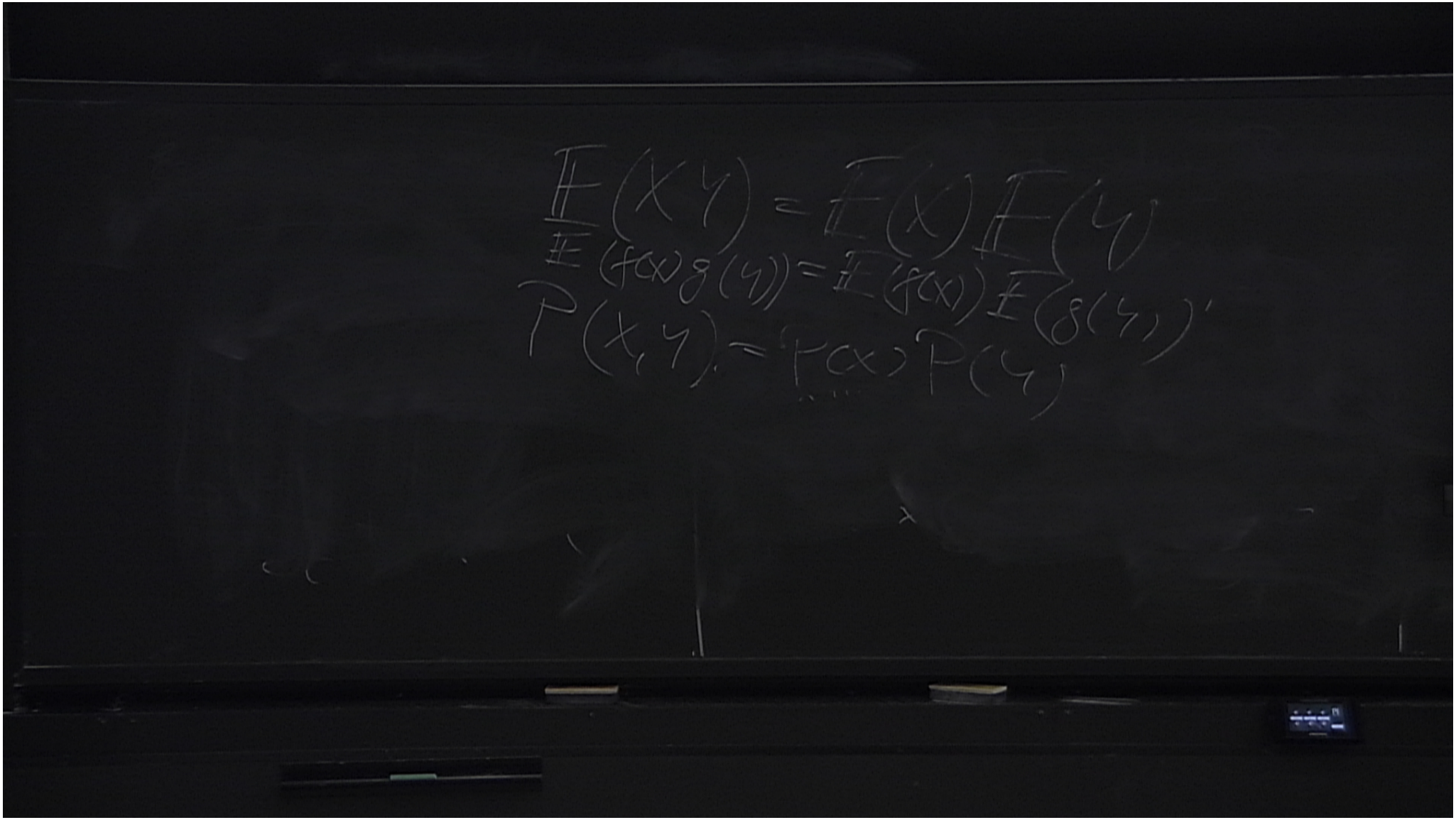
Implementation:

- Compute a function f as non-linear regression of Y on X
- Compute the residual

$$E := Y - f(X)$$

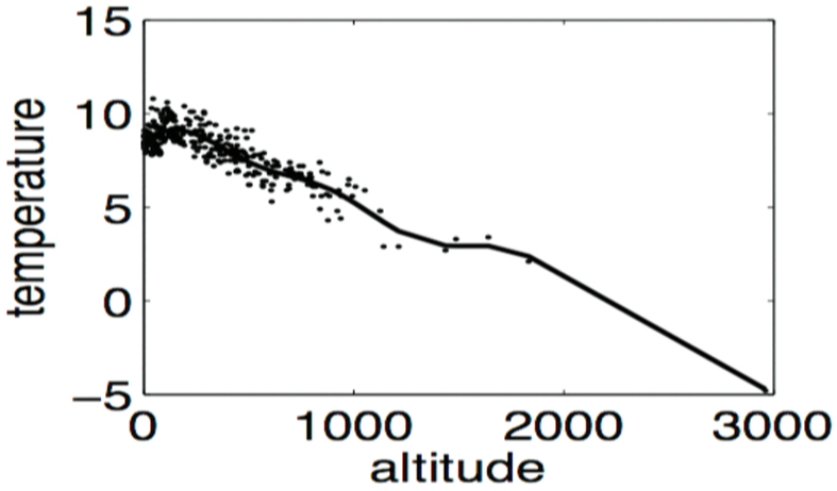
- check whether E and X are statistically independent (uncorrelated is not sufficient, method requires tests that are able to detect higher order dependences)




E(XY) = E(X)E(Y)
E(f(X)g(Y)) = E(f(X))E(g(Y))
P(X,Y) = P(X)P(Y)

Experiments with real data

Relation between altitude (cause) and temperature (effect) of places in Germany



Additive noise based inference...

- about 70% correct decisions for 70 cause-effect pairs with known ground truth
- fraction even better if we allow "no decision"
- we do not claim that noise is always additive in real life, but if it is for one direction this is unlikely to be the wrong one...

Again: independence principle

- If $P(X, Y)$ admits an AN model from X to Y, then $P(Y)$ and $P(X|Y)$ satisfy the DE

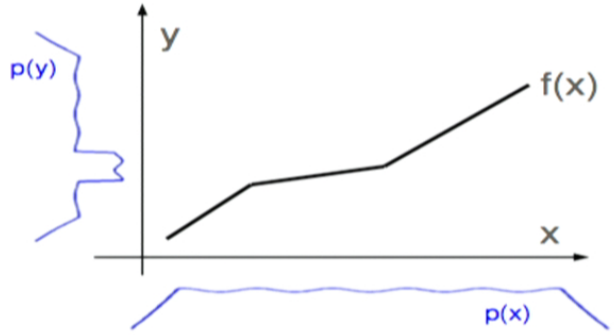
$$\frac{\partial^2}{\partial y^2} \log p(y) = -\frac{\partial^2}{\partial y^2} \log p(x|y) - \frac{1}{f'(x)} \frac{\partial^2}{\partial x \partial y} \log p(x|x) .$$

- unlikely if $P(Y)$ and $P(X|Y)$ are chosen independently

DJ and Steudel OSID 2010

Example (3): Inferring deterministic causality

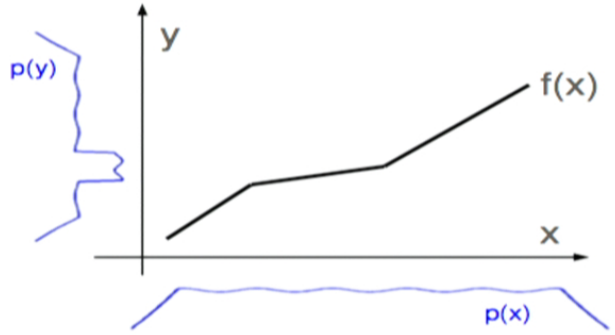
- Problem: infer whether $Y = f(X)$ or $X = f^{-1}(Y)$ is the right causal model
- Idea: if $X \longrightarrow Y$ then f and the density p_X are chosen independently "by nature"
- Hence, peaks of p_X do not correlate with the slope of f
- Then, peaks of p_Y correlate with the slope of f^{-1}



Daniusis, DJ,... UAI 2010

Example (3): Inferring deterministic causality

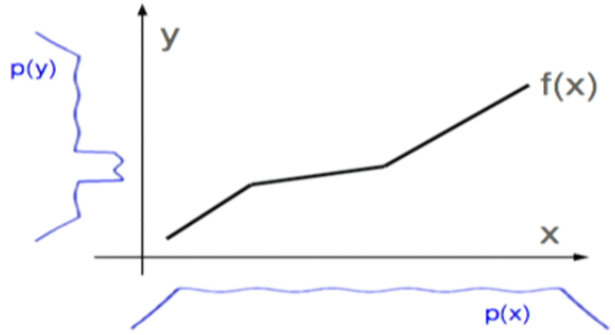
- Problem: infer whether $Y = f(X)$ or $X = f^{-1}(Y)$ is the right causal model
- Idea: if $X \longrightarrow Y$ then f and the density p_X are chosen independently "by nature"
- Hence, peaks of p_X do not correlate with the slope of f
- Then, peaks of p_Y correlate with the slope of f^{-1}



Daniusis, DJ,... UAI 2010

Example (3): Inferring deterministic causality

- Problem: infer whether $Y = f(X)$ or $X = f^{-1}(Y)$ is the right causal model
- Idea: if $X \longrightarrow Y$ 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

Formalization

Let f be a montonously increasing bijection of $[0, 1]$

- **Postulate:**

$$\int_0^1 \log f'(x) P(x) dx = \int_0^1 \log f'(x) dx \text{ (approximately)}$$

- **Idea:** averaging log of slope of f over P is the same as averaging over Lebesgue measure

- **Implication:**

$$\int_0^1 \log f^{-1'} P(y) dy \geq \int_0^1 \log f^{-1'} dy$$

Further implications: additivity of entropies

Let $\overrightarrow{P}(Y)$ be the distribution obtained by applying f to uniformly distributed X

$$S(P(Y)) = S(P(X)) + S(\overrightarrow{P}(Y)) .$$

entropy of output $=$
entropy of input $+$ entropy of image of uniform distribution under f

Further implications: additivity of entropies

Let $\overrightarrow{P}(Y)$ be the distribution obtained by applying f to uniformly distributed X

$$S(P(Y)) = S(P(X)) + S(\overrightarrow{P}(Y)) .$$

entropy of output $=$
entropy of input $+$ entropy of image of uniform distribution under f

Experiments

- **Our cause-effect pairs:**
  about 78% correct decisions

- **Rhine data:**
  - water levels at 22 cities measured in 15 minutes intervals from 1990 to 2008,
  - pick 231 random pairs and decide which one is "upstream"
  - 87% correct decisions

Hence...

empirical distributions $P(X, Y)$ are often asymmetric with respect to swapping cause and effect

– how does this relate to the arrow of time?

Arrow of time in stationary stochastic processes

Peters, DJ, Gretton, Schölkopf ICML 2009

- **Theorem:** If $(X_t)_{t\in\mathbb{Z}}$ has an autoregressive model

$$X_t = \sum_{j=1}^{p} \alpha_j X_{t-j} + E_t \qquad \text{with independent } E_t$$

  there is no such autoregressive model for (X_{-t}), unless E_t is Gaussian or $\alpha_j = 0$.

- **Experiment:** infer the direction of real-world time series (finance, EEG...)

- **Result:** more often linear in forward than in backward direction

smells like an arrow of time, right?

Arrow of time in stationary stochastic processes

Peters, DJ, Gretton, Schölkopf ICML 2009

- **Theorem:** If $(X_t)_{t\in\mathbb{Z}}$ has an autoregressive model

$$X_t = \sum_{j=1}^{p} \alpha_j X_{t-j} + E_t \qquad \text{with independent } E_t$$

  there is no such autoregressive model for (X_{-t}), unless E_t is Gaussian or $\alpha_j = 0$.

- **Experiment:** infer the direction of real-world time series (finance, EEG...)

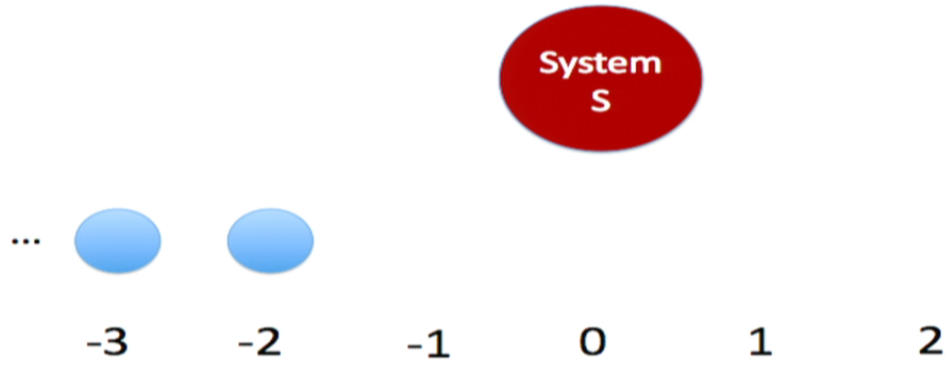
- **Result:** more often linear in forward than in backward direction

smells like an arrow of time, right?

Physical toy model for $X_t = \alpha X_{t-1} + E_t$

DJ, Journ. Stat. Phys. 2010

- X_t: physical observable of a fixed system S at time t.
- noise term provided by propagating particle beam (shift on $\mathbb{Z}$)



Model and its implications

Assumptions:

- interaction is rotation on phase space of S and particle at position 0
- incoming particles statistically independent

Implications:

- outgoing particles are dependent (except for Gaussian states)
- coarse-grained entropy increased
- $P(X_t|X_{t-1})$ is linear, but not $P(X_{t-1}|X_t)$

Model and its implications

Assumptions:

- interaction is rotation on phase space of S and particle at position 0
- incoming particles statistically independent

Implications:

- outgoing particles are dependent (except for Gaussian states)
- coarse-grained entropy increased
- $P(X_t|X_{t-1})$ is linear, but not $P(X_{t-1}|X_t)$

Model and its implications

Assumptions:

- interaction is rotation on phase space of S and particle at position 0
- incoming particles statistically independent

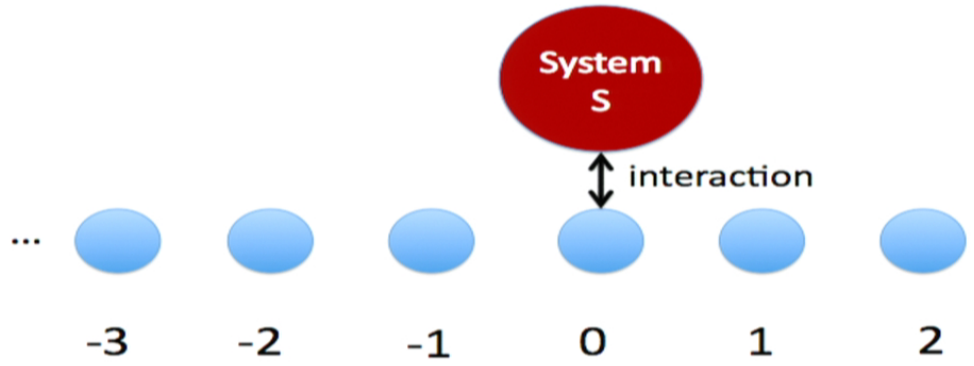
Implications:

- outgoing particles are dependent (except for Gaussian states)
- coarse-grained entropy increased
- $P(X_t|X_{t-1})$ is linear, but not $P(X_{t-1}|X_t)$

Physical toy model for $X_t = \alpha X_{t-1} + E_t$

DJ, Journ. Stat. Phys. 2010

- X_t: physical observable of a fixed system S at time t.
- noise term provided by propagating particle beam (shift on $\mathbb{Z}$)



Forget about statistics for the moment

how do we draw causal conclusions in real life?

what kind of similarities require an explanation?

here we would *not* assume that anyone has copied the design...

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
- infer the existence of a causal relation

a clever one recognizes...

$$11.001001000011111101101010100 1\ldots = \pi$$

- 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

Kolmogorov complexity

Kolmogorov 1965, Chaitin 1966, Solmonoff 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

Kolmogorov complexity

Kolmogorov 1965, Chaitin 1966, Solmonoff 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

$I($  $:$  $) = K($  $)$

Conditional algorithmic mutual information

$$I(x : y|z) = K(x|z) + K(y|z) - K(x, y|z)$$

- information of x on y (and vice versa) when z is already given
- formal analogy to statistical mutual information:

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

- define conditional independence:

$$x \perp\!\!\!\perp y|z \quad :\Leftrightarrow \quad I(x : y|z) \stackrel{+}{=} 0$$

Implication: algorithmic Markov condition

DJ, Schölkopf 2010

3 equivalent formulations

- **local version:** Given its direct causes pa_j, every x_j is conditionally algorithmically independent of its non-effects:

$$x_j \perp\!\!\!\perp nd_j \mid pa_j^*$$

- **global version:** d-separation implies independence
- **recursion formula:**

$$K(x_1, \ldots, x_n) \stackrel{+}{=} \sum_{j=1}^{n} K(x_j | pa_j^*) \, .$$

Conclusion and questions:

- given that nature chooses $P(cause)$ and $P(effect|cause)$ independently, then they are algorithmically independent because there are no algorithmic dependences without causal connections

- is the algorithmic independence between initial state and dynamics of a system also at the heart of the thermodynamic arrow of time?

- challenge: deeper understanding of the relation

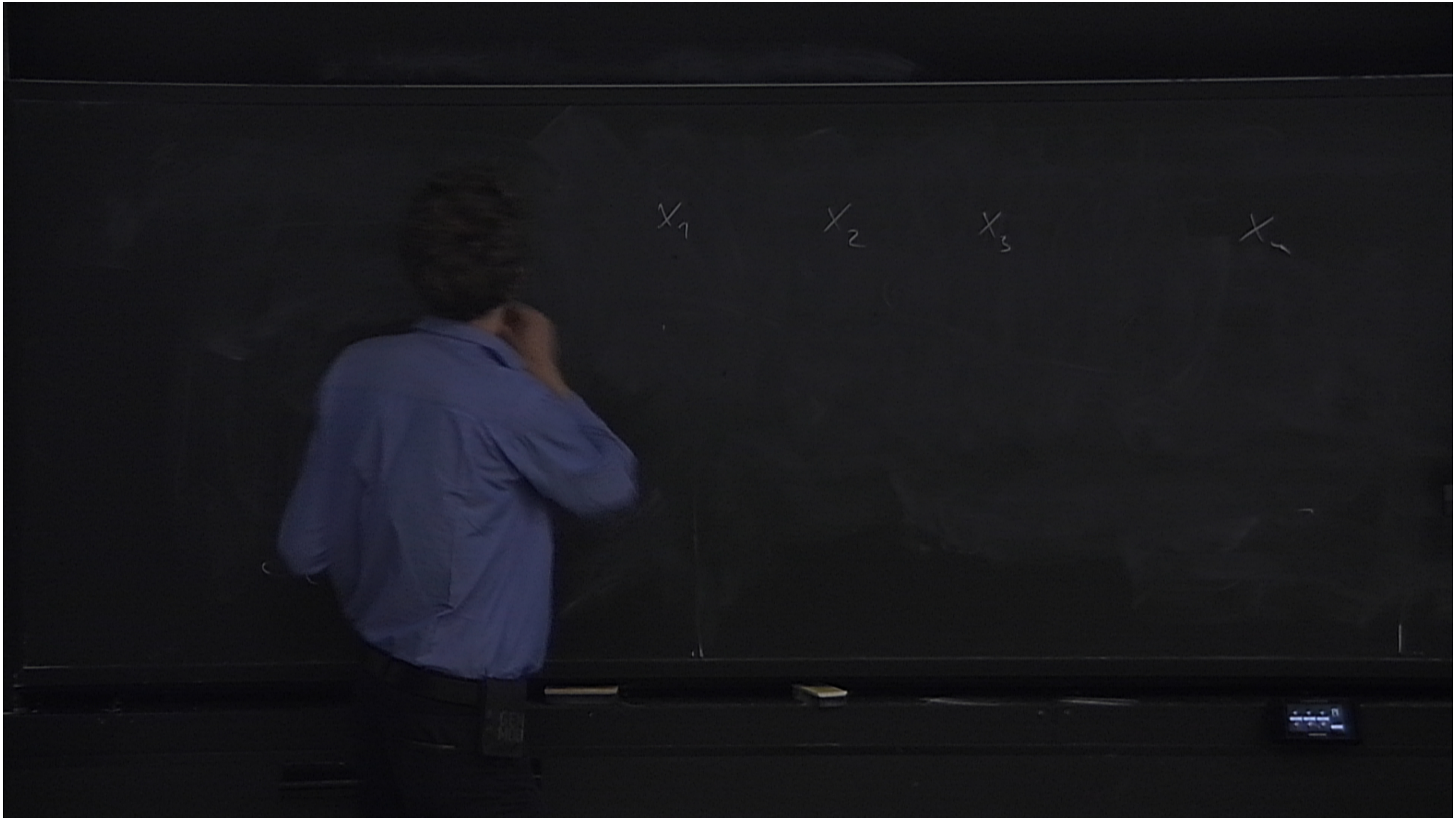
  Causality $\leftrightarrow$ Thermodynamics $\leftrightarrow$ Algorithmic Information

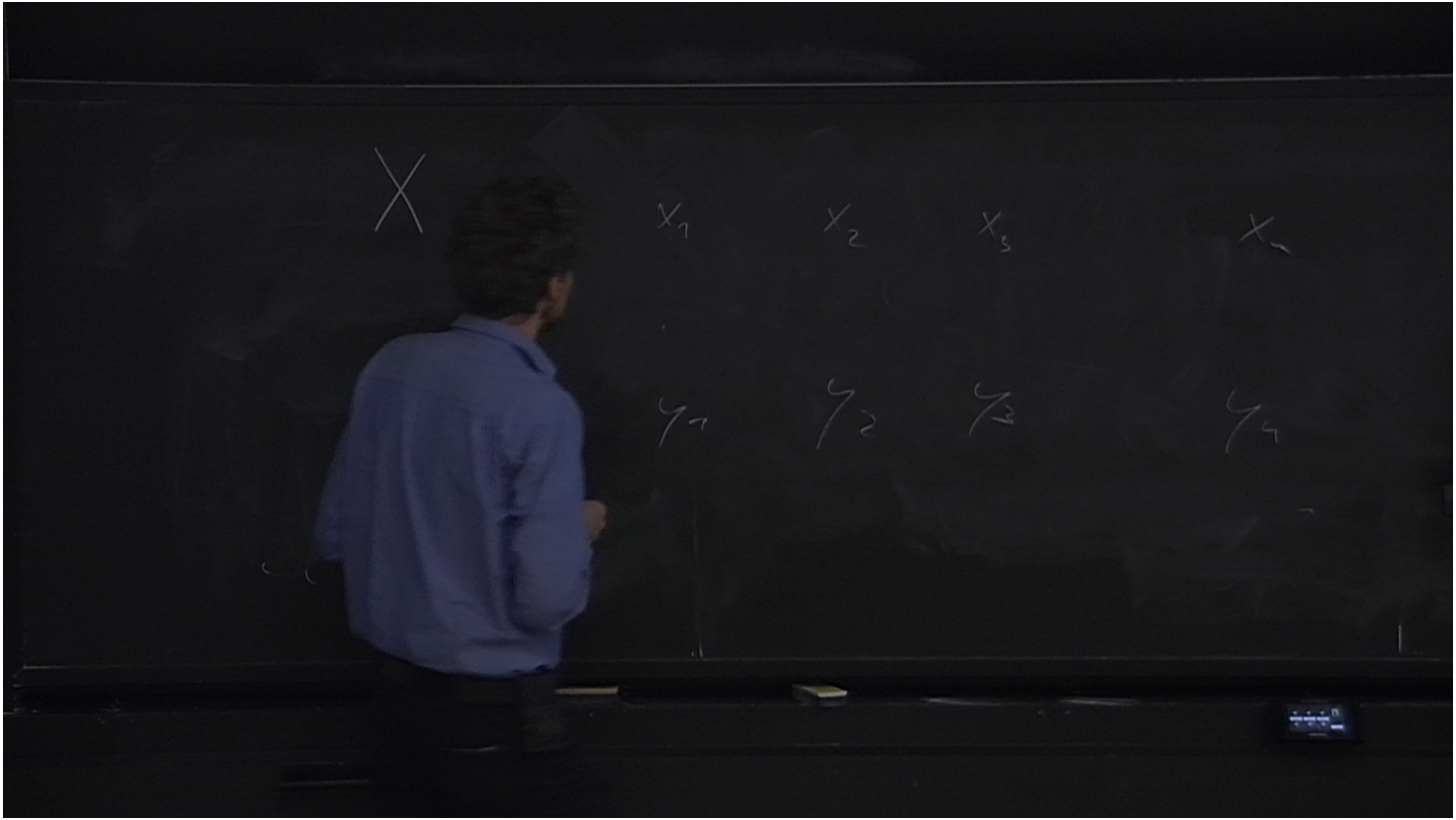
Conclusion and questions:

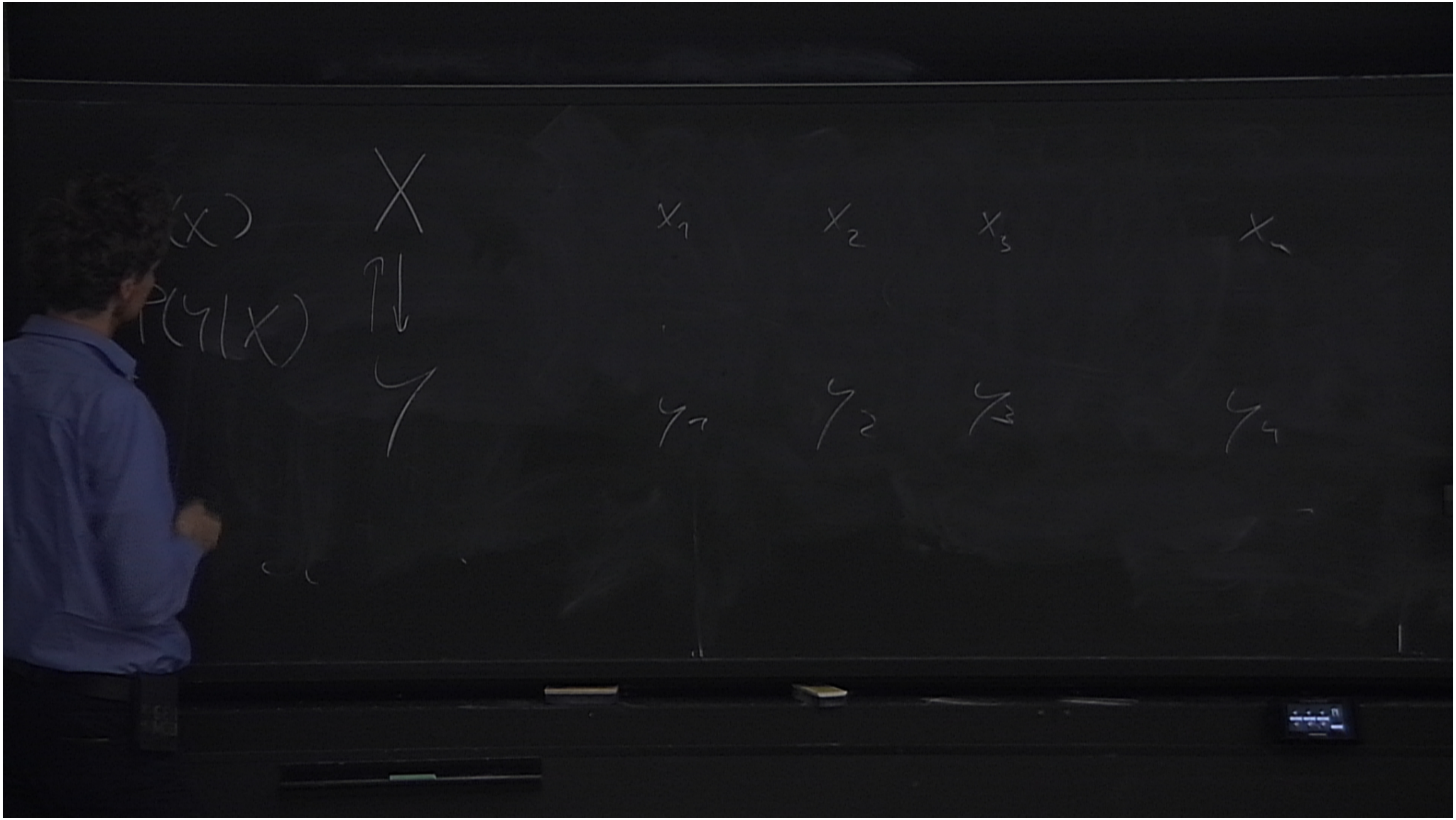
- given that nature chooses $P(cause)$ and $P(effect|cause)$ independently, then they are algorithmically independent because there are no algorithmic dependences without causal connections
- is the algorithmic independence between initial state and dynamics of a system also at the heart of the thermodynamic arrow of time?
- challenge: deeper understanding of the relation

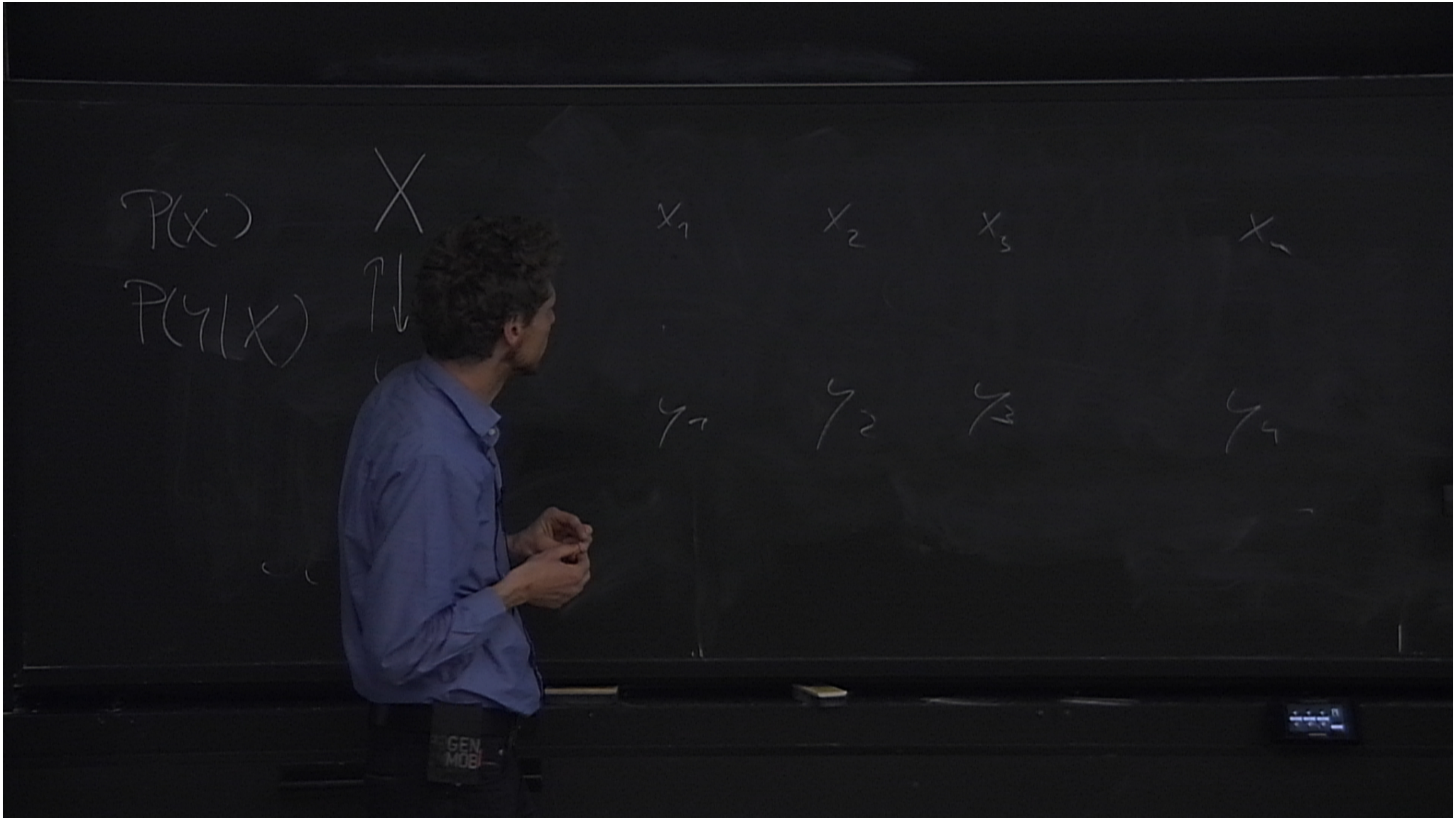
  Causality $\leftrightarrow$ Thermodynamics $\leftrightarrow$ Algorithmic Information

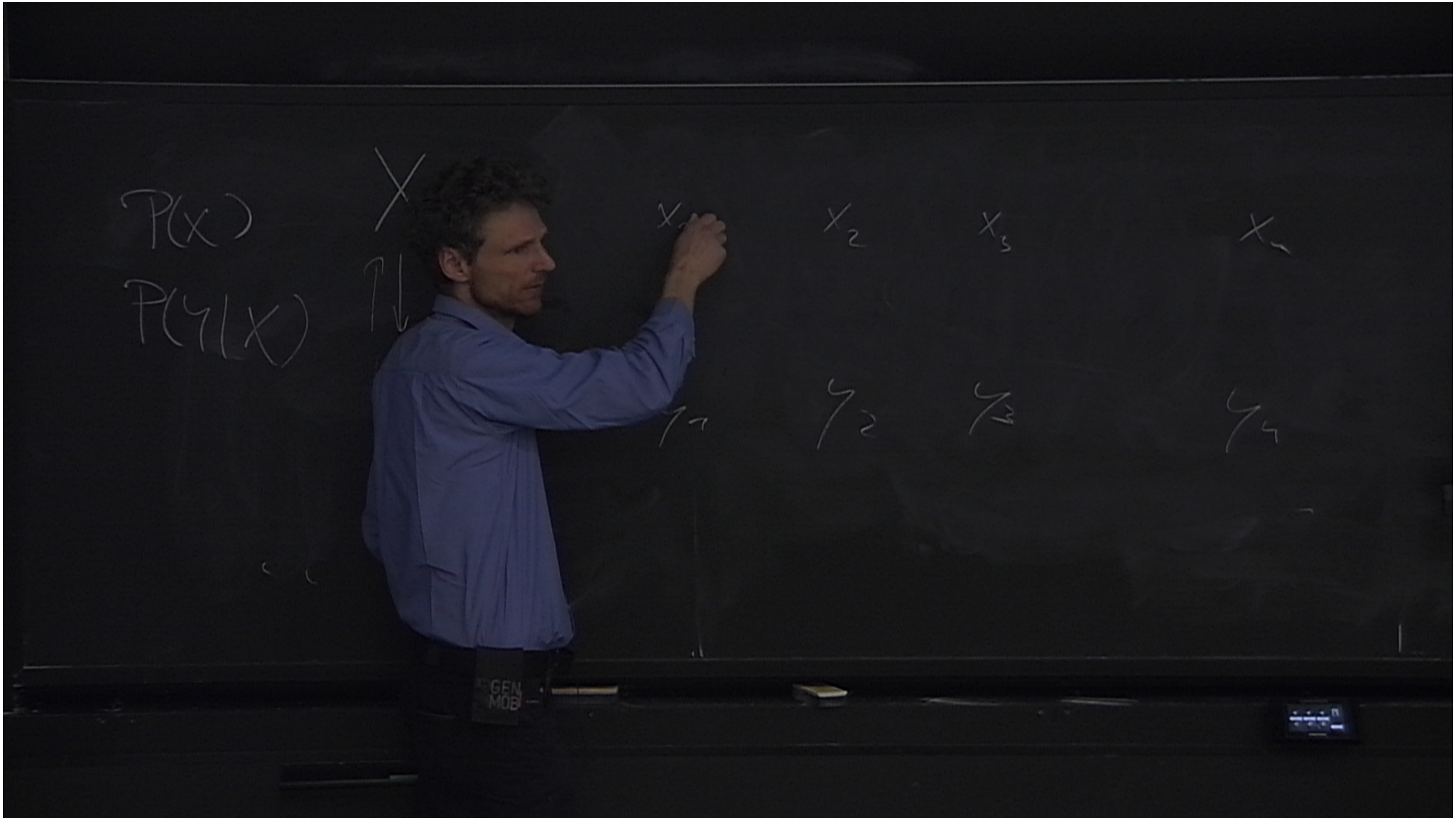
Thanks for your attention

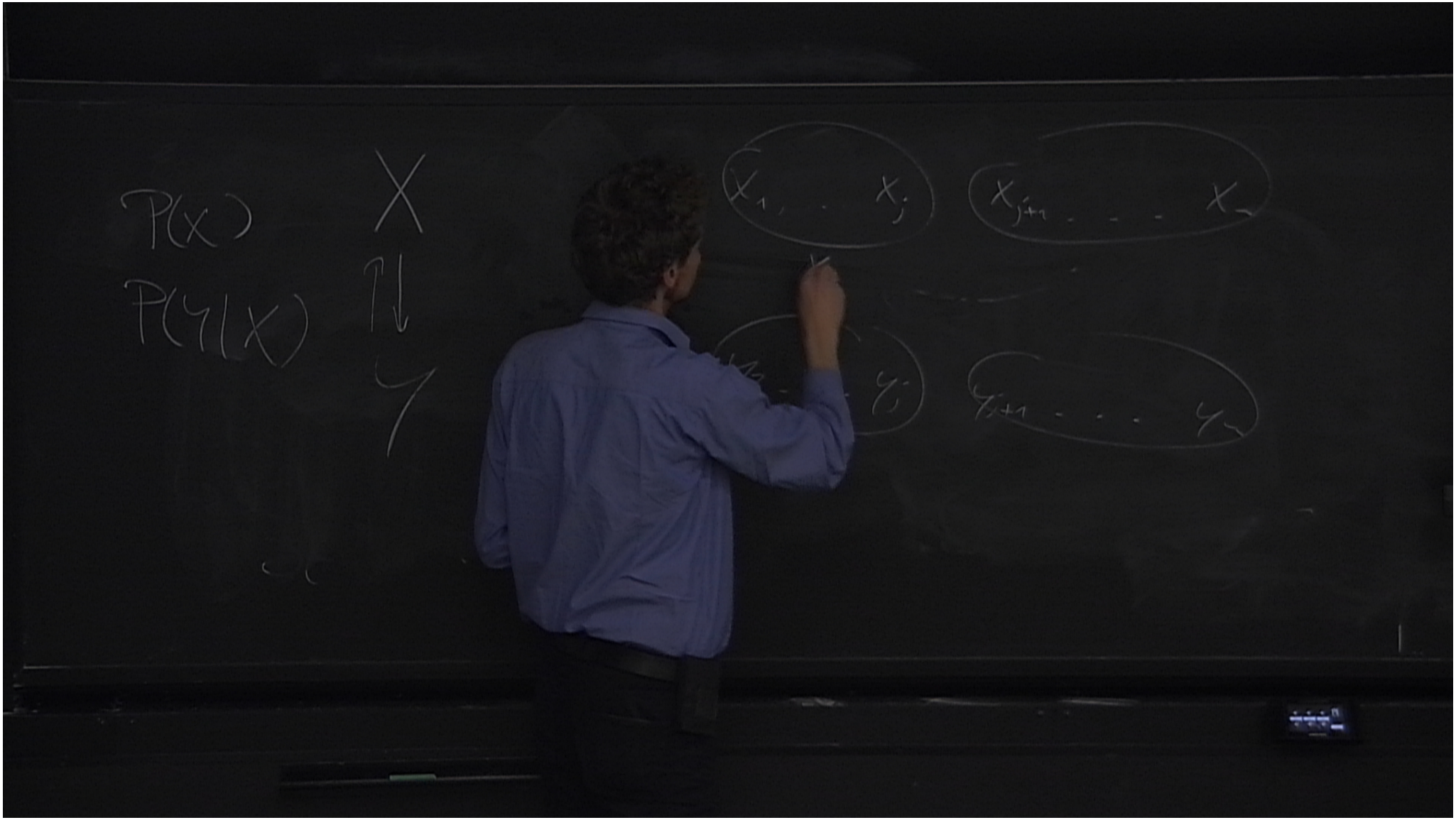

X_1
X_2
X_3
X_4

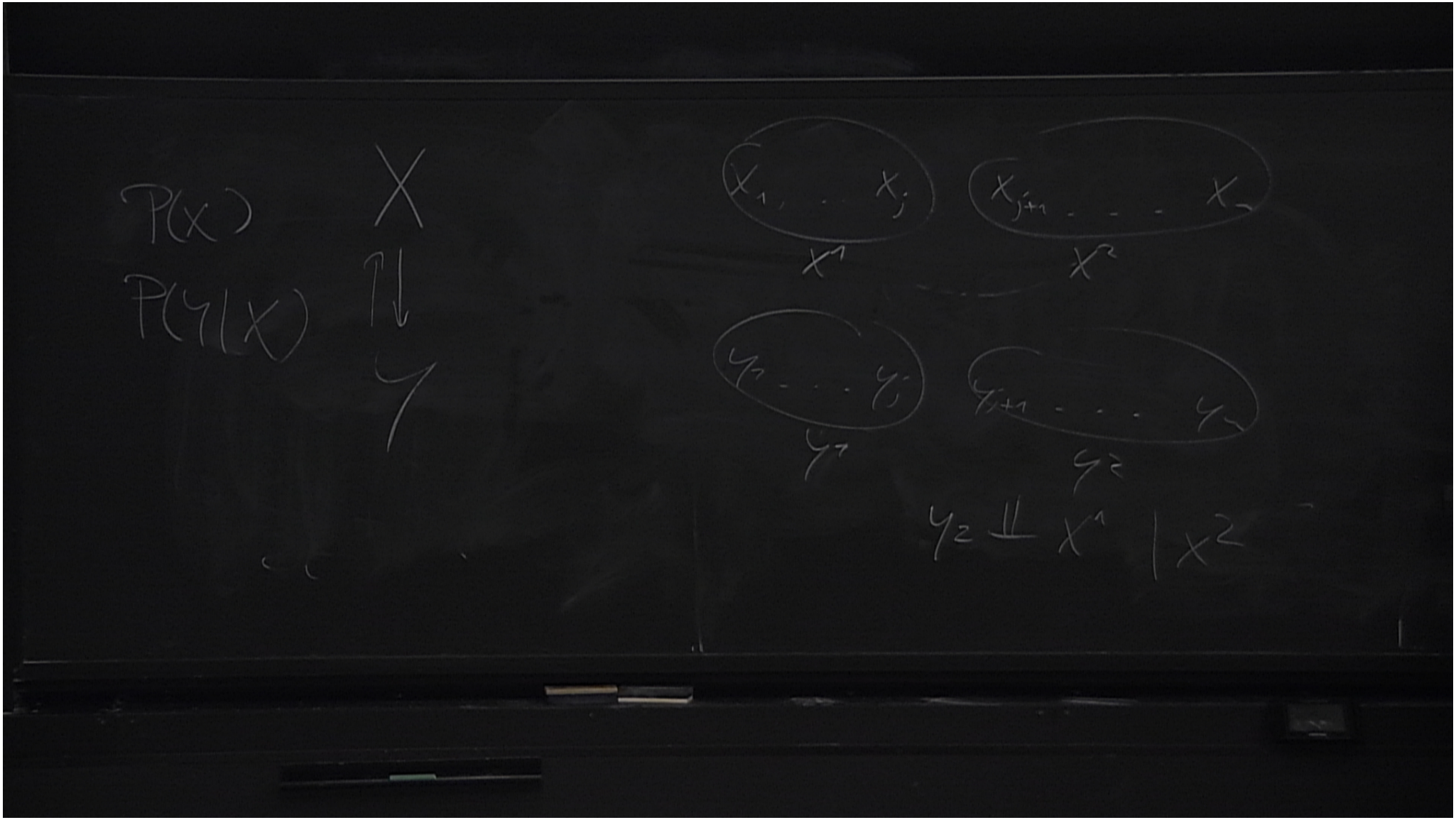

X
X_1
X_2
X_3
Y_1
Y_2
Y_3
Y_4


X
x_1 x_2 x_3 x_4
y_1 y_2 y_3 y_4


P(x)
X
P(y|x)
x_1
x_2
x_3
x_4
y_1
y_2
y_3
y_4


P(x)
P(y|x)
X
x_1
x_2
x_3
x_4
y_1
y_2
y_3
y_4


P(x)
P(y|x)
X
Y


P(X)
P(Y|X)
X
Y
$x_1 \dots x_j$
$x_{j+1} \dots x_n$
x^1
x^2
$y_1 \dots y_j$
$y_{j+1} \dots y_n$
y^1
y^2
$y_2 \perp x^1 \mid x^2$