

# Simply Inferring Causality

Jilles Vreeken



20 August 2015



# Questions of the day

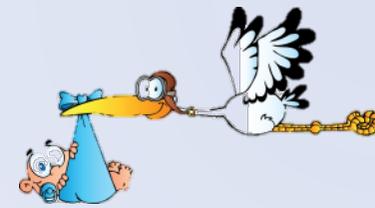
What is **causation**,  
how can we **measure** it,  
and how can **discover** it?



# Causality

*'the relationship between something  
that happens or exists  
and the thing that causes it'*

# Correlation vs. Causation



## Storks Deliver Babies ( $p = 0.008$ )

### KEYWORDS:

Teaching;  
Correlation;  
Significance;  
 $p$ -values.

*Robert Matthews*

Aston University, Birmingham, England.  
e-mail: rajm@compuserve.com

### Summary

This article shows that a highly statistically significant correlation exists between stork populations and human birth rates across Europe. While storks may not deliver babies, unthinking interpretation of correlation and  $p$ -values can certainly deliver unreliable conclusions.

**Correlation** does not tell us anything about **causality**

Instead, we should talk about **dependence**.

Country	Area (km <sup>2</sup> )	Storks (pairs)	Humans (10 <sup>6</sup> )	Birth rate (10 <sup>3</sup> /yr)
Albania	28,750	100	3.2	83
Austria	83,860	300	7.6	87
Belgium	30,520	1	9.9	118
Bulgaria	111,000	5000	9.0	117
Denmark	43,100	9	5.1	59
France	544,000	140	56	774
Germany	357,000	3300	78	901
Greece	132,000	2500	10	106
Holland	41,900	4	15	188
Hungary	93,000	5000	11	124
Italy	301,280	5	57	551
Poland	312,680	30,000	38	610
Portugal	92,390	1500	10	120
Romania	237,500	5000	23	367
Spain	504,750	8000	39	439
Switzerland	41,290	150	6.7	82
Turkey	779,450	25,000	56	1576

**Table 1.** Geographic, human and stork data for 17 European countries

# Dependence vs. Causation

amazon.com Hello. Sign in to get personalized recommendations.

Your Amazon.com Today's Deals

Shop All Departments Electronics Search

Electronics Browse Brands Top Sellers

Prime



**Mobile Edge Explorer**  
Other products by [Mobi](#)

★★★★★  (18 customer reviews)

List Price: \$49.99  
Price: **\$48.32**  
You Save: \$1.67 (3%)

Availability: In Stock.

Want it delivered Tuesday at checkout. [See details](#)

21 used & new available

[See larger image and other views](#)



[Share your own customer images](#)

---

**Better Together** (for amazon)

Buy this item with [HP Pavilion DV2610US 14.1" Entertainment Laptop](#) from Hewlett-Packard today!



+



Total List Price: \$1,123.99  
Buy Together Today: **\$898.31**

[Buy both now!](#)

# Causal Inference



# What is causal inference?

'reasoning to the conclusion that something is, or is **likely** to be, the cause of something else'

## **Godzillian** different definitions of 'cause'

- equally many inference frameworks
- not all with solid foundations;  
many highly specific, most require strong assumptions

# Naïve approach

If

$$P(\textit{cause})P(\textit{effect} \mid \textit{cause}) > P(\textit{effect})P(\textit{cause} \mid \textit{effect})$$

then  $\textit{cause} \rightarrow \textit{effect}$

# Naïve approach

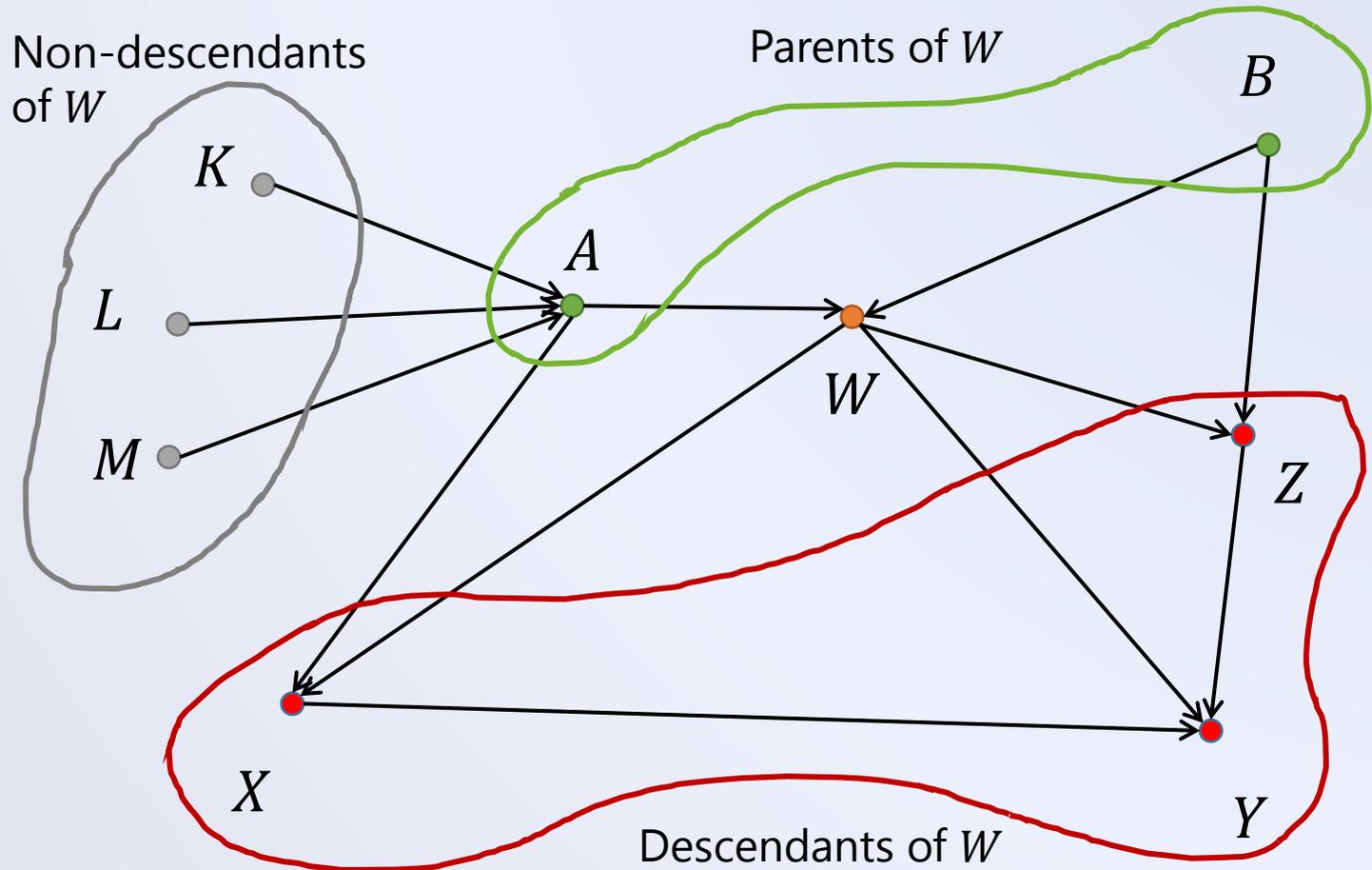
If

$$P(\textit{cause})P(\textit{effect} \mid \textit{cause}) > P(\textit{effect})P(\textit{cause} \mid \textit{effect})$$

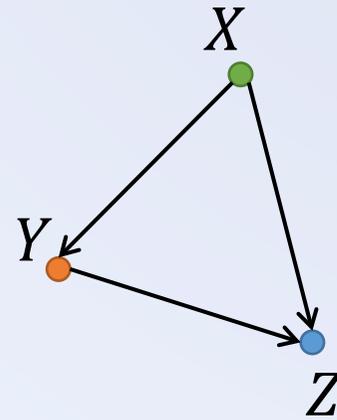
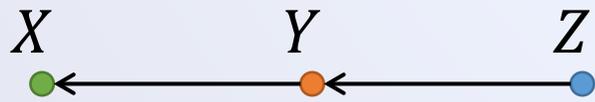
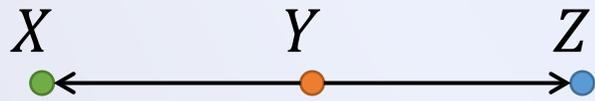
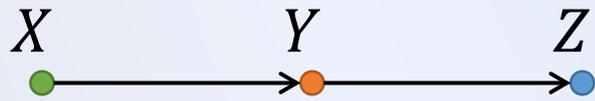
then  $\textit{cause} \rightarrow \textit{effect}$

(rough bastardization of Markov condition)

# Causal Graphs



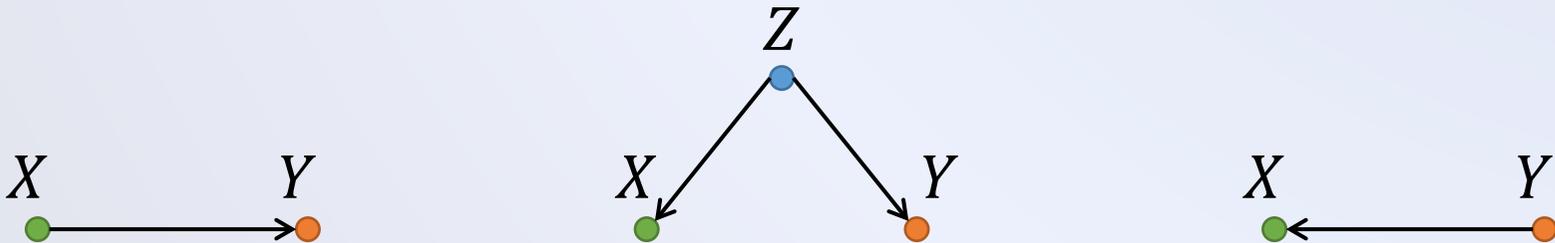
# Choices...



# Statistical Causality

Reichenbach's  
**common cause principle**  
links causality and probability

if  $X$  and  $Y$  are statistically dependent then either



When  $Z$  **screens**  $X$  and  $Y$  from each other,  
given  $Z$ ,  $X$  and  $Y$  become **independent**.

# Causal Markov Condition

Any distribution generated by a Markovian model  $M$

## **Endogenous variable:**

A factor in a causal model or causal system whose value is determined by the states of other variables in the system; contrasted with an **exogenous variable**

where  
and

of  $x_i$  in the causal diagram associated with  $M$

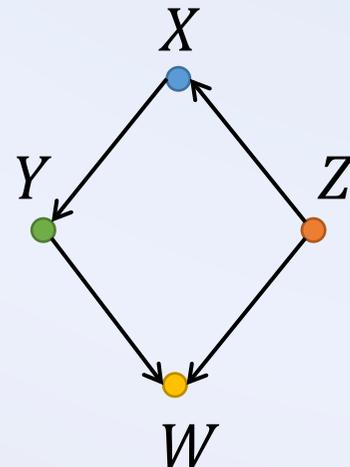
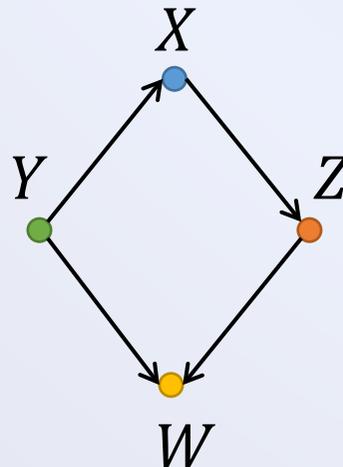
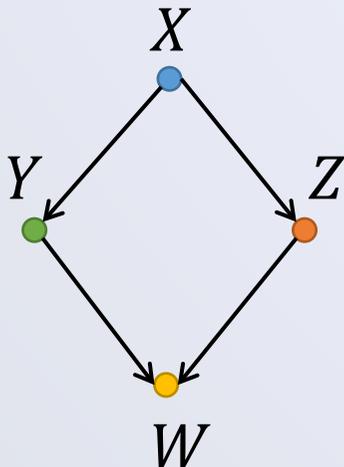
in  $M$ ,  
"

# In other words...

For all distinct variables  $X$  and  $Y$  in the variable set  $V$ , if  $X$  does not cause  $Y$ , then  $\Pr(X | Y, pa_X) = \Pr(X | pa_X)$

That is, we can **weed out** edges from a causal graph – we can identify DAGs *up to* Markov equivalence class.

We are **unable** to choose among these



# Three is a crowd

Traditional causal inference methods  
rely on **conditional independence tests**  
and hence require *at least* **three** observed variables

That is, they **cannot** distinguish between  
 $X \rightarrow Y$  and  $Y \rightarrow X$   
as  $p(x)p(y | x) = p(y)p(x | y)$   
are just factorisations of  $p(x, y)$

But, but, that's exactly what we want to know!

# Wiggle Wiggle

Let's take another look at the definition of causality.

*'the relationship between something that happens or exists and the thing that causes it'*

So, essentially, if  $X$  cause  $Y$ ,  
we can wiggle  $Y$  by wiggling  $X$ ,  
while when we cannot wiggle  $X$  by wiggling  $Y$ .

But... when we only have experimental data  
we cannot do any wiggling ourselves...

# Additive Noise Models

Whenever the joint distribution  $p(X, Y)$  **admits** a model in one direction, e.g.

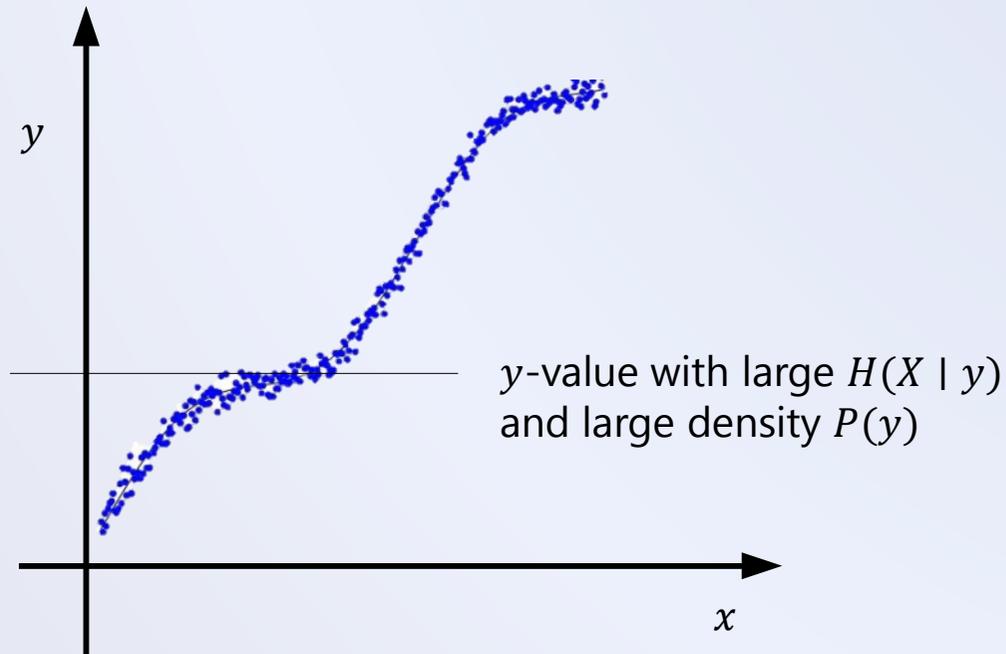
$$Y = f(X) + N \text{ with } N \perp\!\!\!\perp X,$$

but does **not admit** the reversed model,

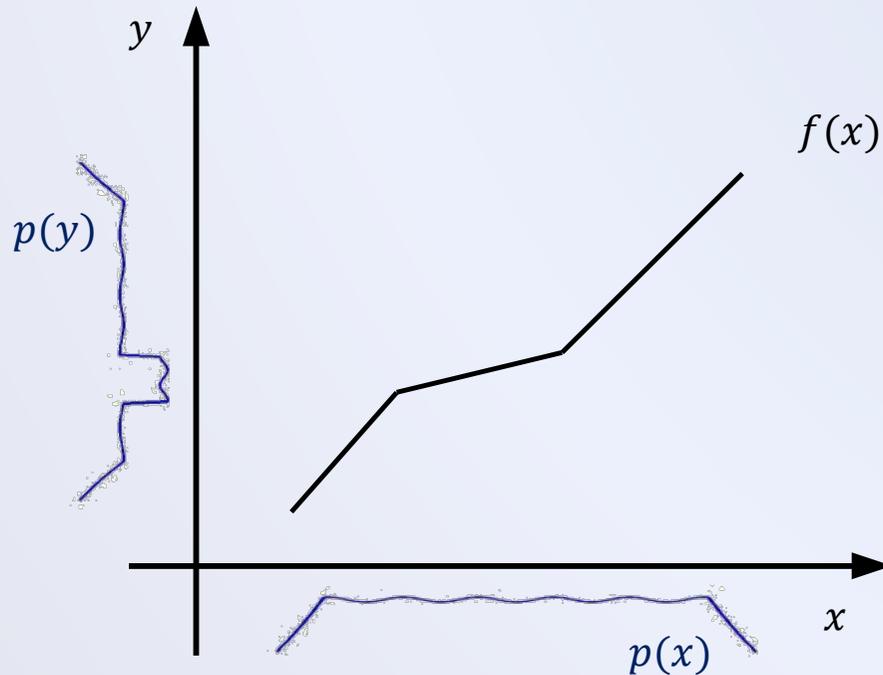
$$X = g(Y) + \tilde{N} \text{ with } \tilde{N} \perp\!\!\!\perp Y$$

We can infer  $X \rightarrow Y$

# May the Noise be With you



# May the Noise be With you



"If the **structure** of density of  $p(x)$  is not correlated with the slope of  $f$ , then the flat regions of  $f$  induce peaks in  $p(y)$ ."

The causal hypothesis  $Y \rightarrow X$  is thus implausible because the causal mechanism  $f^{-1}$  appears to be adjusted to the "input" distribution  $p(y)$ ."

# Plausible Markov Kernels

If  $p(\textit{cause})p(\textit{effect} \mid \textit{cause})$  is **simpler** than  $p(\textit{effect})p(\textit{cause} \mid \textit{effect})$  then  $\textit{cause} \rightarrow \textit{effect}$

but, how to measure 'simpler'?

what about having not  $p$  but  $\hat{p}$ ?

is model complexity alone enough?

is data complexity alone enough?

and, what if there *is* no distribution?

# My approach

Given two objects **X** and **Y** of your favorite types

- e.g. bags of *observations*, or two *objects* of arbitrary type

Say whether

- **X** and **Y** are independent,
- **X** causes **Y** (or vice versa), OR
- **X** and **Y** are correlated

on basis of **descriptive** complexity

Without parameters, without assuming distributions.

(assuming, for the time being, no hidden confounders)

# Kolmogorov Complexity

$$K(s)$$

The Kolmogorov complexity of a binary string  $s$  is the length of the shortest program  $k(s)$  for a universal Turing Machine  $U$  that generates  $s$  and **halts**.

# Elementary, my dear Watson

$k(s)$  is the *simplest* way to generate  $s$   
*algorithmically*

if  $s$  represents data that  
contains causal dependencies  
there will be evidence in  $k(s)$

how can we get this evidence out?  
by using *conditional two-part* complexity

# Conditional Complexity

$$K(s | t)$$

The **conditional** Kolmogorov complexity of a string  $s$  is the length of the shortest program  $k(s)$  for a universal Turing Machine  $U$  **that given string  $t$  as input** generates  $s$  and halts.

# Kolmo-causal

Serialising  $\mathbf{X}$  and  $\mathbf{Y}$  into a string  $s$  we have  $K(\mathbf{X}, \mathbf{Y})$   
for the length of the **shortest** program  $k(\mathbf{X}, \mathbf{Y})$   
that generates  $\mathbf{X}$  and  $\mathbf{Y}$

Intuitively,  
this will factor out *differently* depending  
on how  $\mathbf{X}$  and  $\mathbf{Y}$  are related,  
right?

# Wrench in the works

**Information, however, is symmetric**

$$\begin{aligned} K(\mathbf{X}, \mathbf{Y}) \\ &\triangleq K(\mathbf{X}) + K(\mathbf{Y} | \mathbf{X}) \\ &\triangleq K(\mathbf{Y}) + K(\mathbf{X} | \mathbf{Y}) \end{aligned}$$

(equality up to a constant, Zvonkin & Levin, 1970)

# Direction of information

Instead of factorizing  $K(\mathbf{X}, \mathbf{Y})$ ,  
we have to look at the **effect** of conditioning.

That is, if **knowing  $\mathbf{X}$**  makes the  
algorithmic description of  $\mathbf{Y}$  **easier**,  
 $\mathbf{X}$  is **likely** an (algorithmic) **cause** of  $\mathbf{Y}$

We have to identify the  
strongest **direction of information**  
between  $\mathbf{X}$  and  $\mathbf{Y}$

# Conditional Complexity

So, how about we just regard  
 $K(\mathbf{X} | \mathbf{Y})$  and  $K(\mathbf{Y} | \mathbf{X})$  ?

Close, but no cigar.

If  $K(\mathbf{X})$  is much larger than  $K(\mathbf{Y})$ ,  
directly comparing  $K(\mathbf{X}|\mathbf{Y})$  and  $K(\mathbf{Y}|\mathbf{X})$   
will be biased to the simplest 'cause'.

# Normalised

What we should therefore do,  
is **normalise** as then we can  
determine the strongest  
**direction of information**  
between **X** and **Y** by  
comparing

$$\Delta_{\mathbf{X} \rightarrow \mathbf{Y}} = \frac{K(\mathbf{Y} | \mathbf{X})}{K(\mathbf{Y})} \quad \text{and} \quad \Delta_{\mathbf{Y} \rightarrow \mathbf{X}} = \frac{K(\mathbf{X} | \mathbf{Y})}{K(\mathbf{X})}$$

# Characterising independence

If  $\mathbf{X}$  and  $\mathbf{Y}$  are *algorithmically independent*,

$$K(\mathbf{X}, \mathbf{Y}) \triangleq K(\mathbf{X}, \mathbf{Y} \mid \mathbf{X} \perp\!\!\!\perp \mathbf{Y}) \triangleq \\ K(\mathbf{X}) + K(\mathbf{Y})$$

we will see that

$$\Delta_{\mathbf{X} \rightarrow \mathbf{Y}} = \frac{K(\mathbf{Y} \mid \mathbf{X})}{K(\mathbf{Y})} \quad \text{and} \quad \Delta_{\mathbf{Y} \rightarrow \mathbf{X}} = \frac{K(\mathbf{X} \mid \mathbf{Y})}{K(\mathbf{X})}$$

both approach 1

as neither data will give much information about the other

# Characterising correlation

Last, when  $\mathbf{X}$  and  $\mathbf{Y}$  are only *algorithmically correlated*,

$$K(\mathbf{X}, \mathbf{Y}) \triangleq K(\mathbf{X}, \mathbf{Y} \mid \text{cor}(\mathbf{X}, \mathbf{Y}))$$

we will see

$$\Delta_{\mathbf{X} \rightarrow \mathbf{Y}} = \frac{K(\mathbf{Y} \mid \mathbf{X})}{K(\mathbf{Y})} \approx \Delta_{\mathbf{Y} \rightarrow \mathbf{X}} = \frac{K(\mathbf{X} \mid \mathbf{Y})}{K(\mathbf{X})}$$

as both carry approximately equal amounts  
of information about each other

# Characterising causation

However, if **X** *algorithmically causes* **Y**,  
we will see

$$\Delta_{\mathbf{X} \rightarrow \mathbf{Y}} = \frac{K(\mathbf{Y} | \mathbf{X})}{K(\mathbf{Y})} \quad \text{approach 0, and}$$

$$\Delta_{\mathbf{Y} \rightarrow \mathbf{X}} = \frac{K(\mathbf{X} | \mathbf{Y})}{K(\mathbf{X})} \quad \text{approach 1}$$

as for a large part **X** explains **Y**

# Inference by Complexity

$$\Delta_{\mathbf{X} \rightarrow \mathbf{Y}} < \Delta_{\mathbf{Y} \rightarrow \mathbf{X}}$$

**X** gives more information about **Y** than vice versa:  
there is an *algorithmic causal connection*!

distances to 0, 1, and each other tell us the **strength**

For objects *and* collections.  
Descriptive. No parameters. No priors.  
Catches any **algorithmic** dependency.

...but, can we actually implement this?

# Implementing our Rule

We can approximate  $K(s)$  by lossless **compression**

here we need compressors that incorporate conditional description of **models** and **data**

$$C(\mathbf{Y}' | \mathbf{X}), \quad C(\mathbf{Y} | \mathbf{X}, \mathbf{Y}')$$

Do we have such compressors?

not really – never any explicit use for

but... we can define them!

# ERGO

For sets of observations  $\mathbf{X}$  and  $\mathbf{Y}$

- high-dimensional continuous data
- fast, non-parametric, noise tolerant, **very good results**

Estimates directed information by entropy

- average number of bits to describe  $(x,y)$  assuming  $\mathbf{X} \rightarrow \mathbf{Y}$
- using normalised cumulative resp. Shannon entropy
- (non-)linear functional (non-)deterministic causation

# Shannon entropy for continuous

As discussed last week, to compute

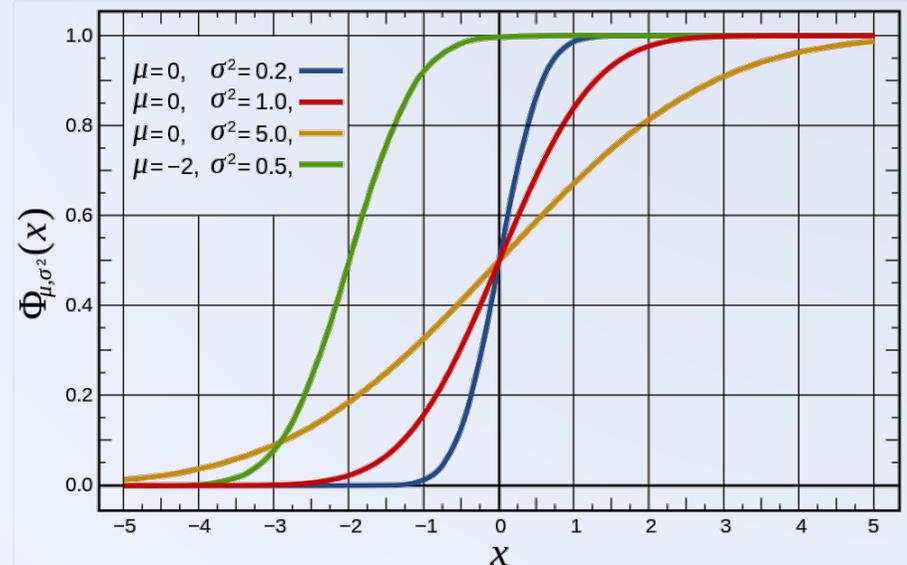
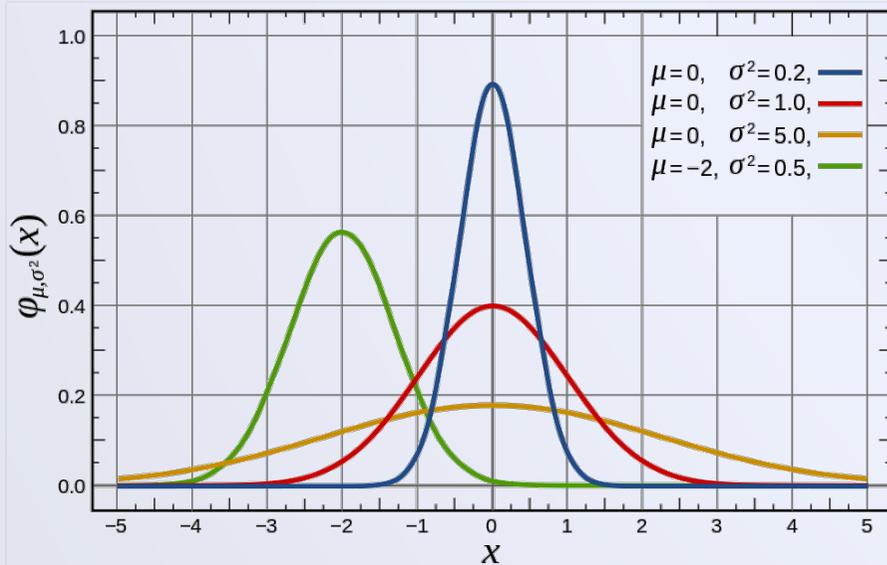
$$h(X) = - \int_{\mathbf{X}} f(x) \log f(x) dx$$

We need to estimate the probability density function, choose a step-size, and then hope for the best.

If we don't know the distribution, we can use kernel density estimation – which requires choosing a kernel and a bandwidth.

KDE is well-behaved for univariate, but estimating multivariate densities is very difficult, especially for high dimensionalities.

# Cumulative Distributions



$$F(x) = P(X \leq x)$$

**cdf** can be computed directly from data  
**no** assumptions necessary

# Cumulative Entropy

Entropy has been defined for cumulative distribution functions!

$$h_{CE}(X) = - \int_{dom(X)} P(X \leq x) \log P(X \leq x) dx$$

As  $0 \leq P(X \leq x) \leq 1$   
we obtain  $h_{CE}(X) \geq 0$   
(!)

# Cumulative Entropy

How do we compute  $h_{CE}(X)$  in practice?

Easy.

Let  $X_1 \leq \dots \leq X_n$  be i.i.d. random samples of continuous random variable  $X$

$$h_{CE}(X) = - \sum_{i=1}^{n-1} (X_{i+1} - X_i) \frac{i}{n} \log \frac{i}{n}$$

# Multivariate Cumulative Entropy

Cumulative entropy, however, is **only defined** for univariate variables.

We **estimate** the multivariate cumulative entropy of  $\mathbf{X}$  as

$$\hat{h}(\mathbf{X}) = h(\mathbf{X}_1) + h(\mathbf{X}_2|\mathbf{X}_1) + h(\mathbf{X}_3|\mathbf{X}_1, \mathbf{X}_2) + \dots h(\mathbf{X}_m|\mathbf{X}_1, \dots, \mathbf{X}_{m-1})$$

where we choose  $X_{i+1}$  such that  $h(\mathbf{X}_{i+1} | \mathbf{X}_i, \dots)$  is minimal when we optimally discretise  $\mathbf{X}_i$

Note, other factorisations are possible, and may give better approximations...

# Entropy-based Direction of Information

$$\Delta_{\mathbf{X} \rightarrow \mathbf{Y}} = \frac{h(\mathbf{Y} | \mathbf{X})}{h^u(\mathbf{Y})} + \frac{H(\mathbf{X}', \mathbf{Y}')}{H^u(\mathbf{X}') + H^u(\mathbf{Y}')}$$

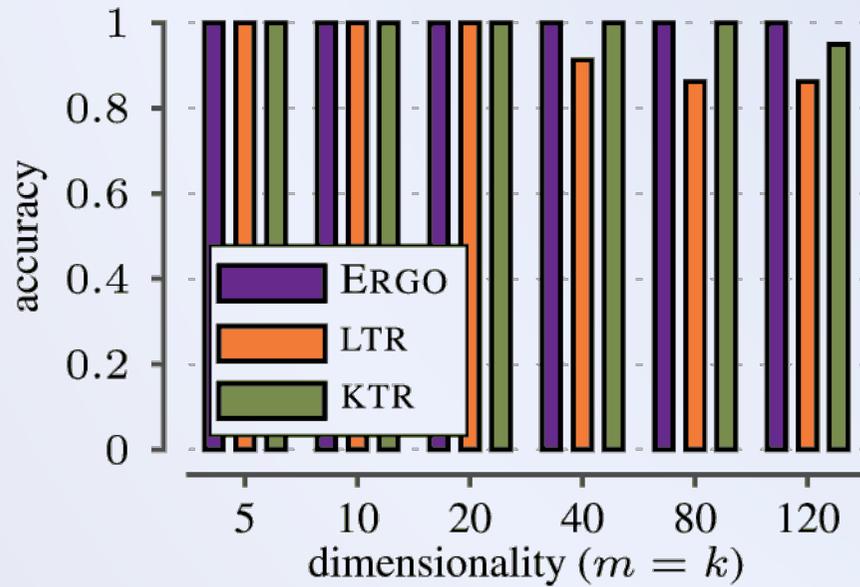
and accordingly for  $\Delta_{\mathbf{Y} \rightarrow \mathbf{X}}$

# Entropy-based Direction of Information

$$\Delta_{\mathbf{X} \rightarrow \mathbf{Y}} = \frac{\overbrace{h(\mathbf{Y} | \mathbf{X})}^{\text{Cost of the data (cumulative entropy)}}}{h^u(\mathbf{Y})} + \frac{\overbrace{H(\mathbf{X}', \mathbf{Y}')}^{\text{Cost of the model (Shannon entropy)}}}{H^u(\mathbf{X}') + H^u(\mathbf{Y}' )}$$

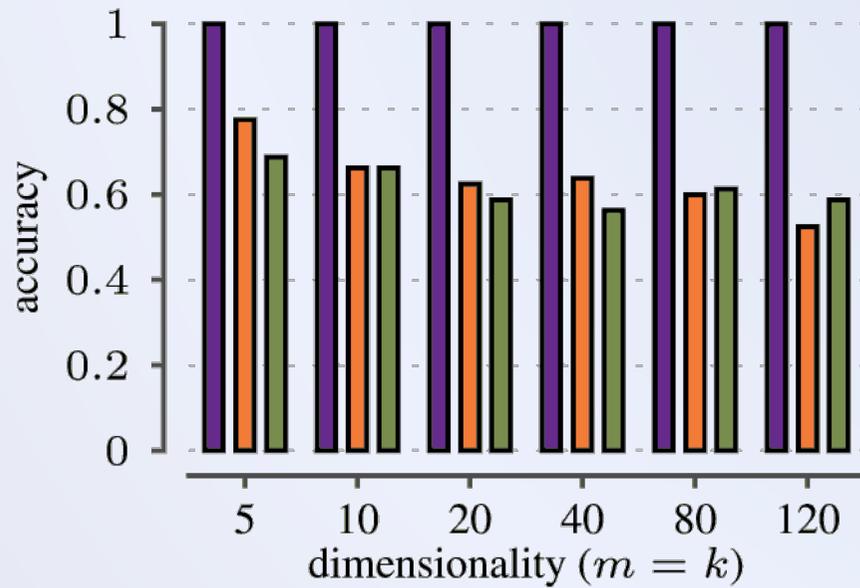
and accordingly for  $\Delta_{\mathbf{Y} \rightarrow \mathbf{X}}$

# Results



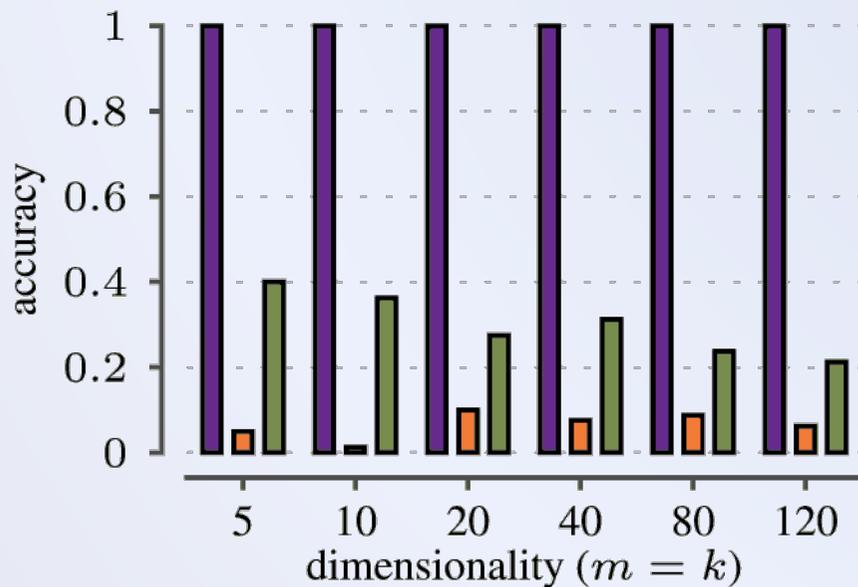
$$\tanh(2x) + \tanh(3x + 1) + \tanh(4x + 2)$$

# Results



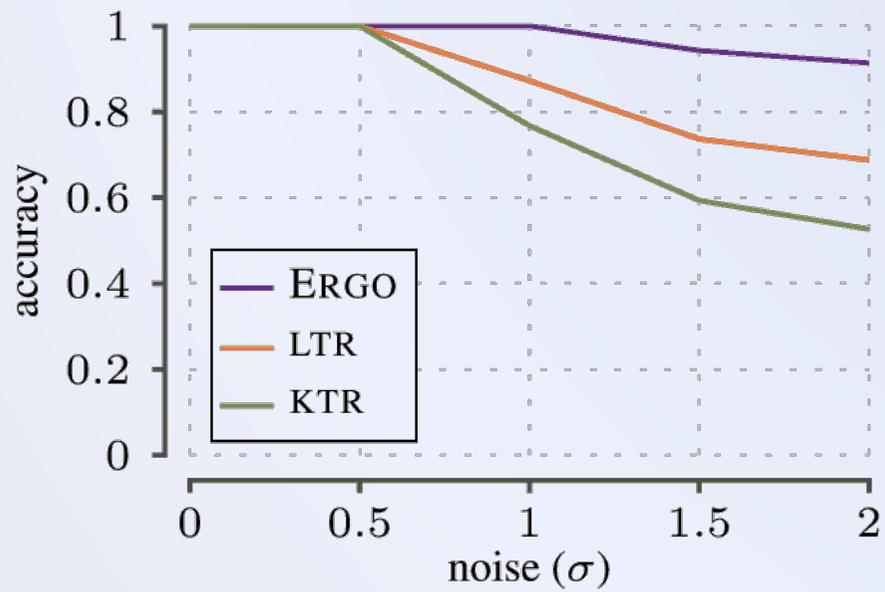
$$\sin(2x) + \sin(3x + 1)$$

# Results



$$\sin(2x) + \sin(3x + 1) + \frac{1}{3}(\tanh(2x) + \tanh(3x + 1) + \tanh(4x + 2))$$

# Results



# More results

## Benchmark data

### *Age → Marital Status*

$$\Delta_{age \rightarrow marital\ status} = 0.90$$

$$\Delta_{marital\ status \rightarrow age} = 1.36$$

### *Education level → Income*

$$\Delta_{education\ level \rightarrow income} = 1.13$$

$$\Delta_{income \rightarrow education\ level} = 1.62$$

### *Gender → Income*

$$\Delta_{gender \rightarrow income} = 0.55$$

$$\Delta_{income \rightarrow gender} = 0.96$$

## Real data

*# of Roman Catholic family members*

→

*# of married couples in the family*

*# of family members with high edu.*

→

*# of family members with high status*

*average income of whole family*

→

*# of home owners*

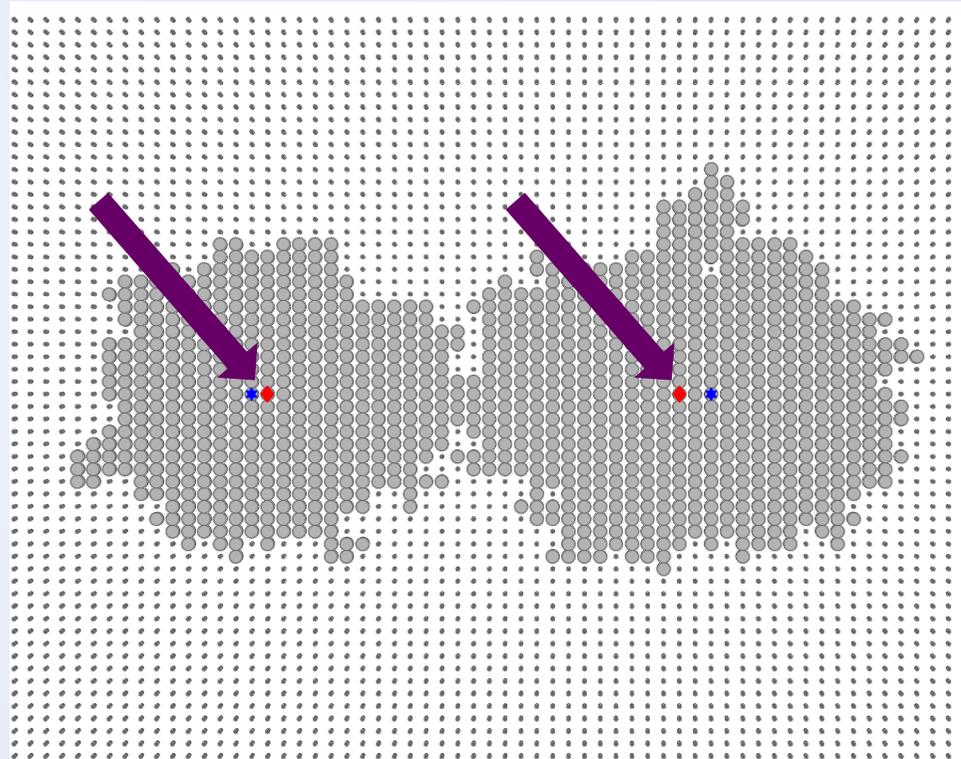
# Example: Who are the Culprits?

Suppose a graph in which an epidemic spreads

- who caused it?

## Main ideas

- uninfected neighbors **exonerate** you from being a culprit
- the more **easy** to reach the footprint, the **better**



# Conclusions

## Causal inference

- important, difficult, in rapid development

## Causal inference by **algorithmic complexity**

- solid foundations, clear interpretation, non-parametric
- for *any* pair of **objects** of *any* sort, for **type** and **token** causation
- instantiation for multivariate real-valued data works very well

## Ongoing

- how deep does the rabbit hole go?

Time's up!



Okay...  
so, what was this  
all about?

# Take home message

**data analysis** ↔ **communication**

transfer the data  
to the analyst  
in as **few** as possible bits

'induction by compression'



# Conclusion

## exploratory data analysis

'summarising, by algorithmic means,  
the main **structure** of a dataset  
in an easily understandable form'

information theory offers versatile tools for  
measuring **structure** as well as **simplicity**

*Thank you!*

exploratory data analysis

'summarising, by algorithmic means,  
the main **structure** of a dataset  
in an easily understandable form'

information theory offers versatile tools for  
measuring **structure** as well as **simplicity**