

# **Causal Inference**

using the algorithmic Markov condition

Janzing. D., Schoelkopf. B.  
IEEE transaction on Information Theory 2010

Talk by: ashkan mokarian

# 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

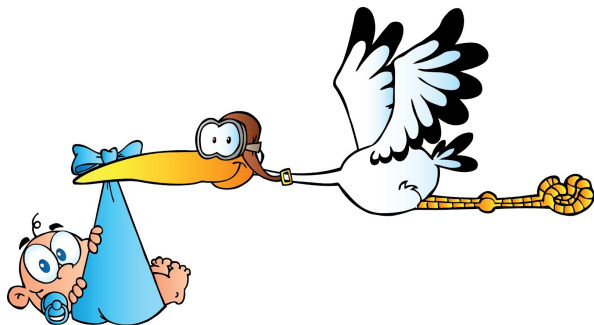
## Summary

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.

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

Correlation does not tell us anything about Causality. Better talk about dependence.



# Dependence vs. Causation

amazon.com Hello. Sign in to get personalized recommendations

Your Amazon.com Today's Deals

Shop All Departments Electronics Browse Brands Top Sellers

Search Electronics

Prime



**Mobile Edge Express**  
Other products by [Mobile Edge Express](#)

★★★★★ (18 customer reviews)


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

Availability: In Stock. [See details](#)

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 PC](#) from Hewlett-Packard today!



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

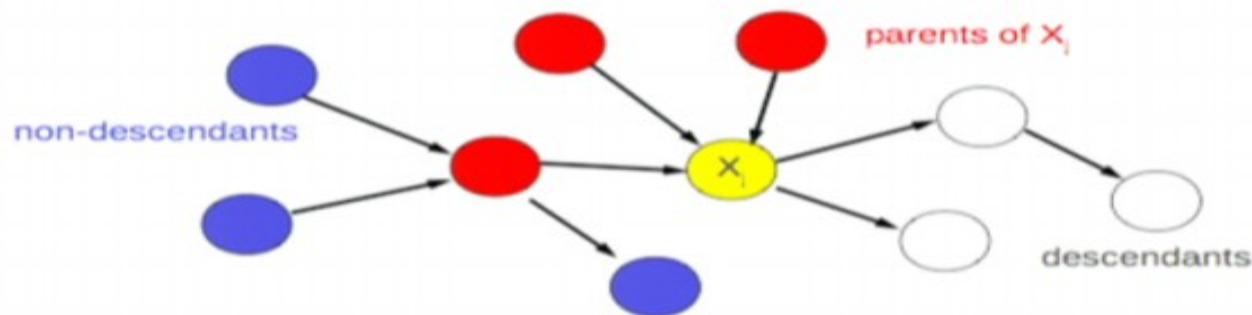
[Buy both now!](#)

# Outline

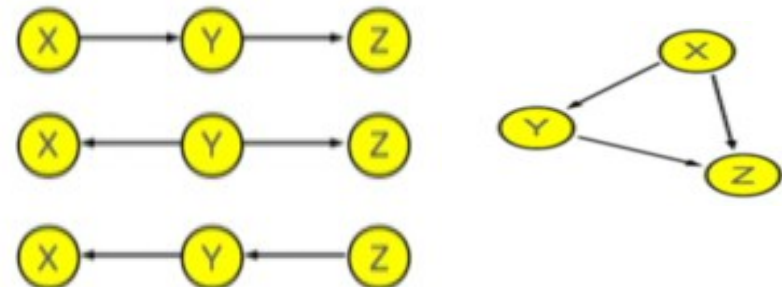
1. What is the causal inference problem
2. Probabilistic failure  
( how?, why it doesn't work)
3. Algorithmic success or How I Learnt to Love AIT

# Causal inference problem

- ✓ Given variables  $X_1, \dots, X_n$
- Given data: n-tuples drawn from  $P(X_1, \dots, X_n)$
- ✓ Infer causal structure = DAG



- Too many choice
- Reasonable rules for preferring a DAG is required.

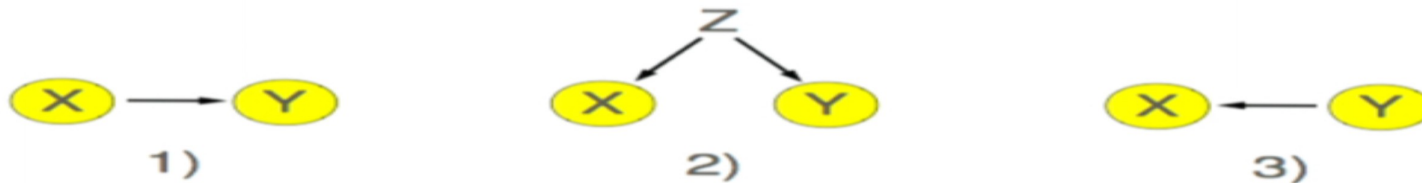


# Statistical Implications of Causality



Reichenbach's *Common Cause Principle* links **causality** and **probability**:

If two variables  $X$  and  $Y$  are statistically **dependent** then either



In case 2)  $Z$  screens  $X$  and  $Y$  from each other (given  $Z$ , the observables  $X$  and  $Y$  become **independent**)

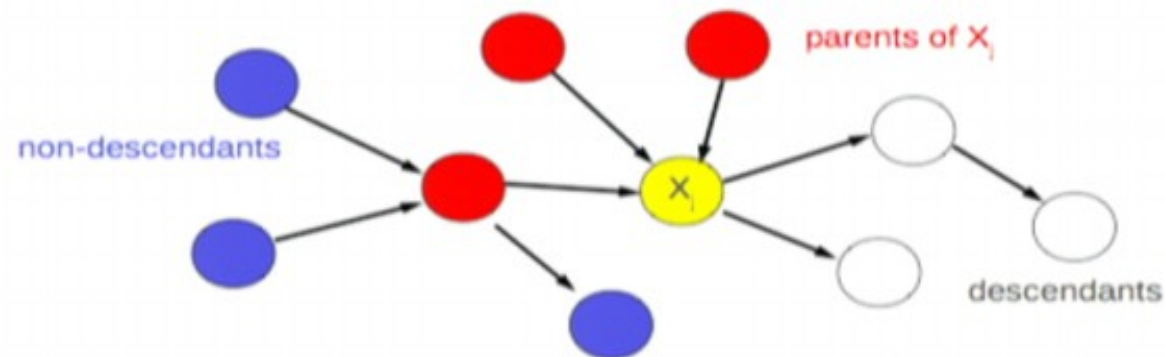
# Causal Markov condition (4 equivalent versions)

- ▶ **Local Markov condition:** every node is conditionally independent of its non-descendants, given its parents (Information exchange with non-descendants involves parents)

- ▶ **Factorization:**

$$P(X_1, \dots, X_n) = \prod_{j=1}^n P(X_j | PA_j)$$

- ▶ *Two other versions*

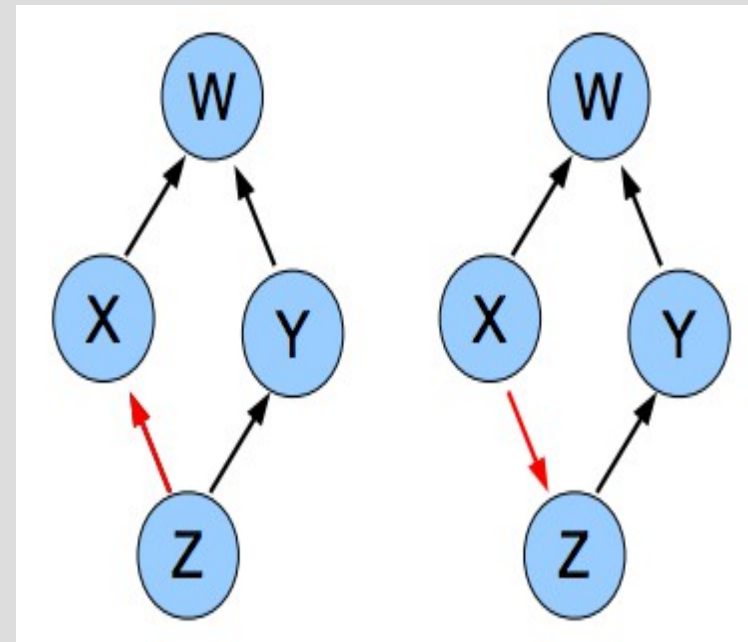


# Independence based Causal Inference

## Limitations:

- identifies DAGs up to **Markov equivalent** class

- Skeleton:** corresponding undirected graph
- V-structure:** sets of colliders, e.g.  
 $X \rightarrow W \leftarrow Z$





# Independence based Causal Inference

## Limitations:

- identifies DAGs up to Markov equivalent class
- Using only dependent/independent. Not considering strength of dependence

Strength of statistical dependencies is often measured in terms of **Mutual Information**. For  $X, Y, Z$ , MI is defined as:

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

# Independence based Causal Inference

## Limitations:

- identifies DAGs up to Markov equivalent class
- Dependence on **i.i.d.** sampling. Fails for **single** objects.
- ...

# Algorithmic mutual information

Common information between  $x$  and  $y$

$$\begin{aligned} \blacktriangleright I(x : y) &:= K(x) + K(y) - K(x, y) \\ &= K(x) - K(x \mid y^*) = K(y) - K(y \mid x^*) \end{aligned}$$

Example:  $I(\text{★} : \text{★}) = K(\text{★})$

$$\blacktriangleright I(x : y \mid z) := K(x \mid z) + K(y \mid z) - K(x, y \mid z)$$

$\blacktriangleright$  Conditional independence defined as:

$$x \perp y \mid z \Leftrightarrow I(x : y \mid z) \simeq 0$$

# 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  $I(X ; Y)$

# Analogy to statistics (in the single object case)

- Replace strings  $x, y$  (=objects) with random variables  $X, Y$
- ~~Replace Kolmogorov complexity with Shannon entropy~~

$$H(X) \leq \frac{1}{n} \mathbb{E}(K(\mathbf{x}|n)) \leq H(X) + \frac{|\mathcal{A}| \log n}{n} + \frac{c}{n}$$

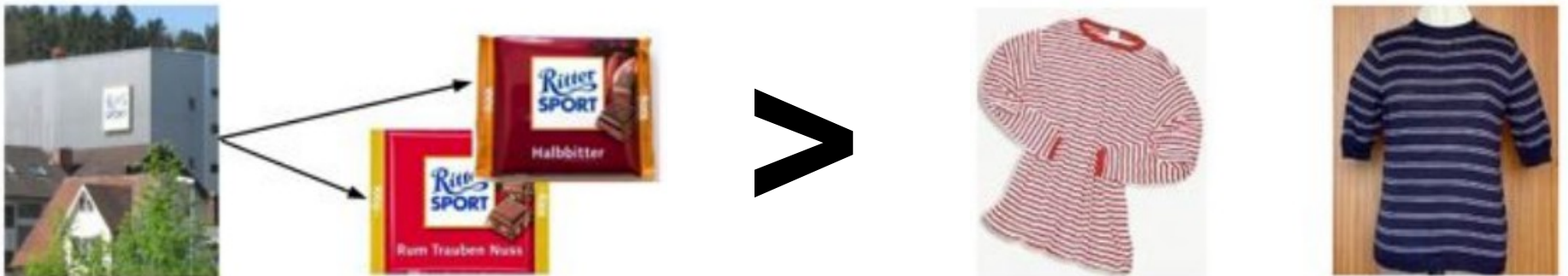
- ~~Replace algorithmic mutual information  $I(x : y)$  with statistical mutual information  $I(X ; Y)$~~

$$I(X; Y) - K(P) \stackrel{+}{\leq} \mathbb{E}(I(x : y)) \stackrel{+}{\leq} I(X; Y) + 2K(P)$$

# AMI for single objects

$$I(X;Y) - K(P) \stackrel{+}{\leq} \mathbb{E}(I(x:y)) \stackrel{+}{\leq} I(X;Y) + 2K(P)$$

- i) if  $K(P) \ll I(X;Y)$
- ii) o.w.  $I(x:y)$  shows the desired properties

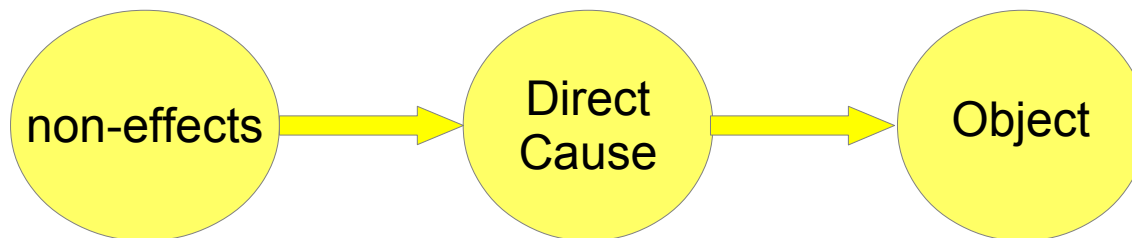


# causal Markov condition

- ▶ Recall the **(Local) Causal Markov condition**:
- ▶ Reformulation:  
given all direct causes of an observable, its non-effects provide no additional **statistical** information on it

# causal Markov condition

- Generalization:  
given all direct causes of an observable, its non-effects provide no additional ~~statistical~~ information on it
- **Algorithmic Causal Markov Condition:**  
given all direct causes of an object, its non-effects provide no additional **algorithmic** information on it





(4 equivalent versions)

# Algorithmic Causal Markov Condition

For  $n$  strings  $x_1, \dots, x_n$  the following conditions are equivalent

- ▶ Local Markov condition

$$I(x_j : nd_j | pa_j) = 0$$

- ▶ Recursive form:

$$K(x_1, \dots, x_n) = \sum_{j=1}^n K(x_j | pa_j)$$

- ▶ *Two other versions*

(proof in the paper)

# Independence of conditionals - Example

If  $X \rightarrow Y$  then,

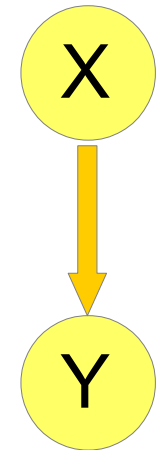
$$I(P(X):P(Y|X))=0$$

And equivalently,

$$K(P(X, Y))=K(P(X))+K(P(Y|X))$$

Which implies

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



# Use Heuristic - $X \rightarrow Y$ or $Y \rightarrow X$ ?

$$P(x, y) = \frac{1}{2} \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(y - \mu - x\lambda)^2}{2\sigma^2}}$$

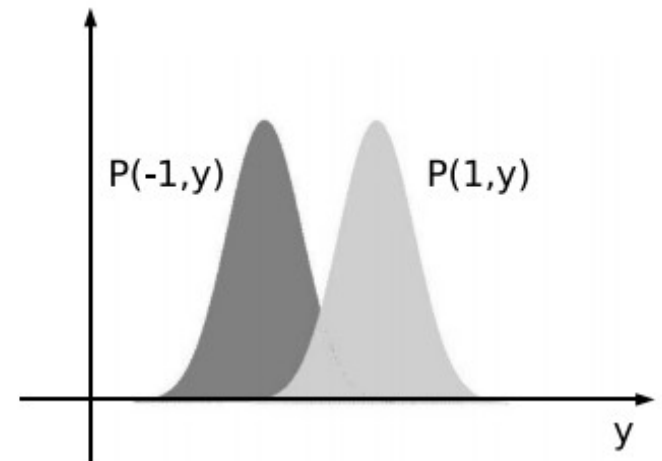
$$K(P(X, Y)) = K(\mu, \lambda, \sigma) = K(\mu) + K(\lambda) + K(\sigma)$$

---

Assumption: equally weighted modes

$$K(P(X)) = 0$$

$$K(P(Y|X)) = K(\mu) + K(\lambda) + K(\sigma)$$



$$P(y) = \frac{1}{2} \frac{1}{\sigma \sqrt{2\pi}} \left( e^{-\frac{(y - \mu + \lambda)^2}{2\sigma^2}} + e^{-\frac{(y - \mu - \lambda)^2}{2\sigma^2}} \right) \rightarrow K(P(Y)) = K(\mu) + K(\lambda) + K(\sigma)$$

$$P(X = 1|y) = \frac{1}{2} \left( 1 + \tanh \frac{\lambda(y - \mu)}{\sigma^2} \right) \rightarrow K(P(X|Y)) = K(\mu) + K(\lambda/\sigma^2)$$

# The power to reject(+)

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  $X \rightarrow Y$  and  $Y \rightarrow X$ . (e.g. The  $X \leftarrow Z \rightarrow Y$  is the true structure)

Whereas previous rules just choose the simplest one

# Decidable modifications

- ▶  $K(\cdot)$  uncomputable. **AMI** is **uncomputable**.
- ▶ estimate AMI by apprx.  $K(\cdot)$  with “conditional compression scheme”

$$\tilde{I}(x:y) = \text{Compress}(x|\varepsilon) - \text{Compress}(x|y)$$

GenCompress: optimal prefix coding with *Replace*, *Insert*, *Delete* operations.

Example:

$\text{Compress}(\text{gaccgtca}) = |10\ 00\ 01\ 01\ 10\ 11\ 01\ 00| = 16\text{bits}$

$\text{Compress}(\text{gacc}\mathbf{g}\text{tca}|\text{gac}\mathbf{c}\mathbf{t}\text{tca}) = \{(0, 7), (\mathbf{R}, 4, \mathbf{g})\}$

$\dots = |0\ 000\ 111\ 1\ 00\ 100\ 10| = 15\text{bits}$

# Summary

- ▶ Statistical causal inference results in DAGs only up to Markov equivalent classes
- ▶ by replacing AMI in CMC we can get the APMC and it's already enough for overcoming the limitations of statistical causal inference
  - ▶ Inference for single objects (no need for iid sampling)
  - ▶ Beyond Markov equivalent DAGs ( $X \rightarrow Y$ )
  - ▶ Novel inference rules

# Parts not covered from paper

- ✖ Two other novel inference rules derived from APMC
- ✖ More Ideas on how to replace Kolmogorov complexity with decidable criteria

THANKS FOR YOUR ATTENTION