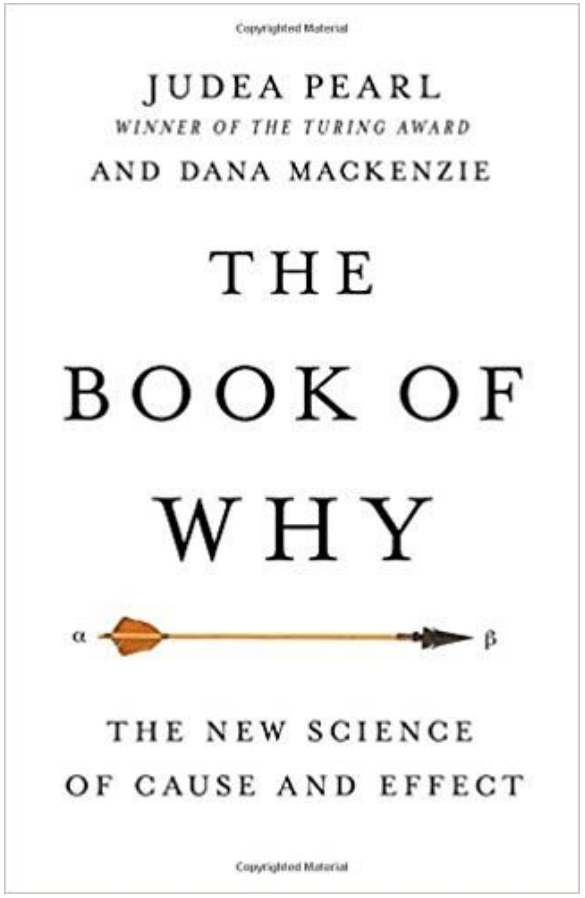
Historic anecdotes about (non-) causal thinking in statistics and artificial intelligence



published in May 2018

current amazon bestseller #1
in the category “statistics”
(followed by Elements of Statistical Learning)

Beate.Sick@zhaw.ch

Topics of today

- Humans and scientists want to understand the “WHY”
- Correlation: birth of statistics – end of causal thinking?
- (Causal) reasoning with Bayesian Networks
- Pearl’s ladder of causation
- Can our statistical and ML/DL models “only do curve fitting” ?
- Historic anecdotes in statistics and ML seen through a causal lens

Humans conscious rises the question of WHY?

God asks for WHAT

“Have you eaten from the tree which I forbade you?”

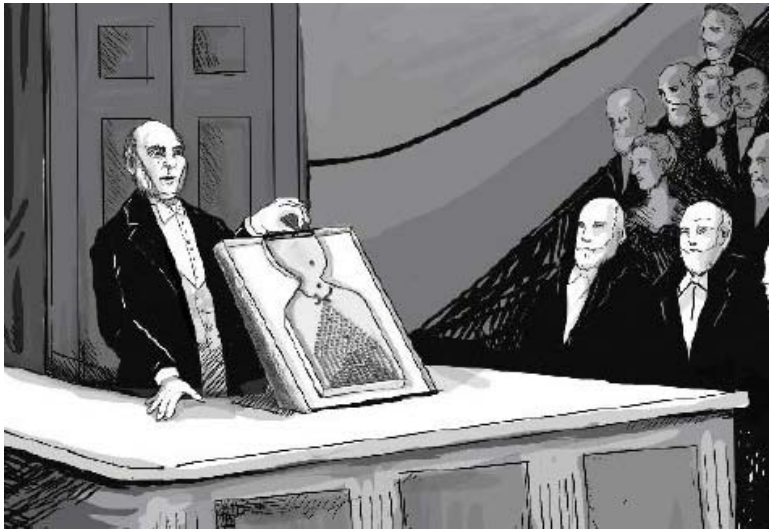
Adam answers with WHY

“The woman you gave me for a companion, she gave me fruit from the tree and I ate.”

“I would rather discover one cause than be the King of Persia.”

The ancient Greek philosopher Democritus (460–370 BC)

Galton on the search for causality



Galton in 1877 at the Friday Evening Discourse at the Royal Institution of Great Britain in London.

Francis Galton (first cousin of Charles Darwin) was interested to **explain** how traits like “intelligence” or “height” is passed from generation to generation.

Galton presented the “quincunx” (Galton nailboard) as causal model for the inheritance.

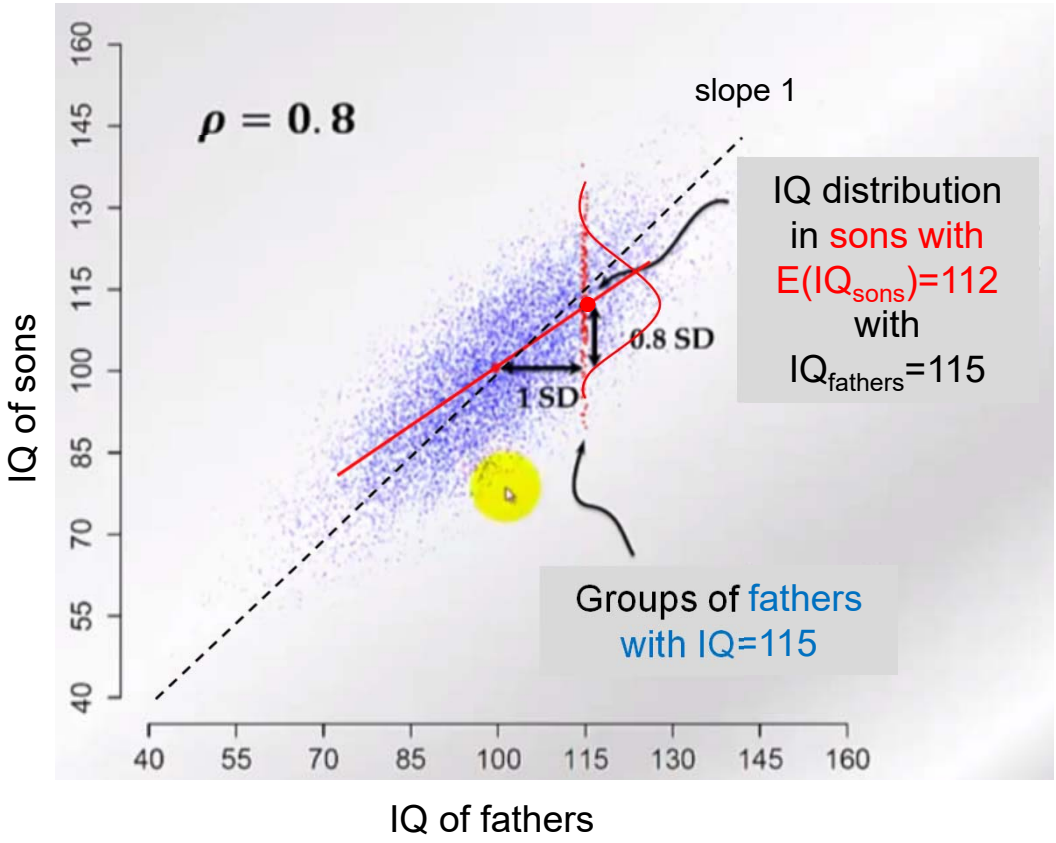
Balls “inherit” their position in the quincunx in the same way that humans inherit their stature or intelligence.

The stability of the observed spread of traits in a population over many generations contradicted the model and puzzled Galton for years.

Image credits: “The Book of Why”

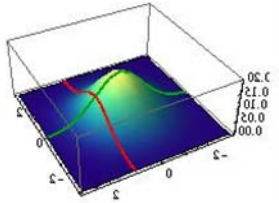
Galton's discovery of the regression line

Remark: Correlation of IQs of parents and children is only 0.42 https://en.wikipedia.org/wiki/Heritability_of_IQ



$$X1 \sim N\left(\mu_1 = 100,\ \sigma_1^2 = 15^2\right)$$
$$X2 \sim N\left(\mu_1 = 100,\ \sigma_1^2 = 15^2\right)$$

$$\begin{pmatrix} X1 \\ X2 \end{pmatrix} \sim N\left(\begin{pmatrix} 100 \\ 100 \end{pmatrix}, \begin{pmatrix} 15^2 & \mathrm{cov}(X1,X2) \\ \mathrm{cov}(X1,X2) & 15^2 \end{pmatrix}\right)$$

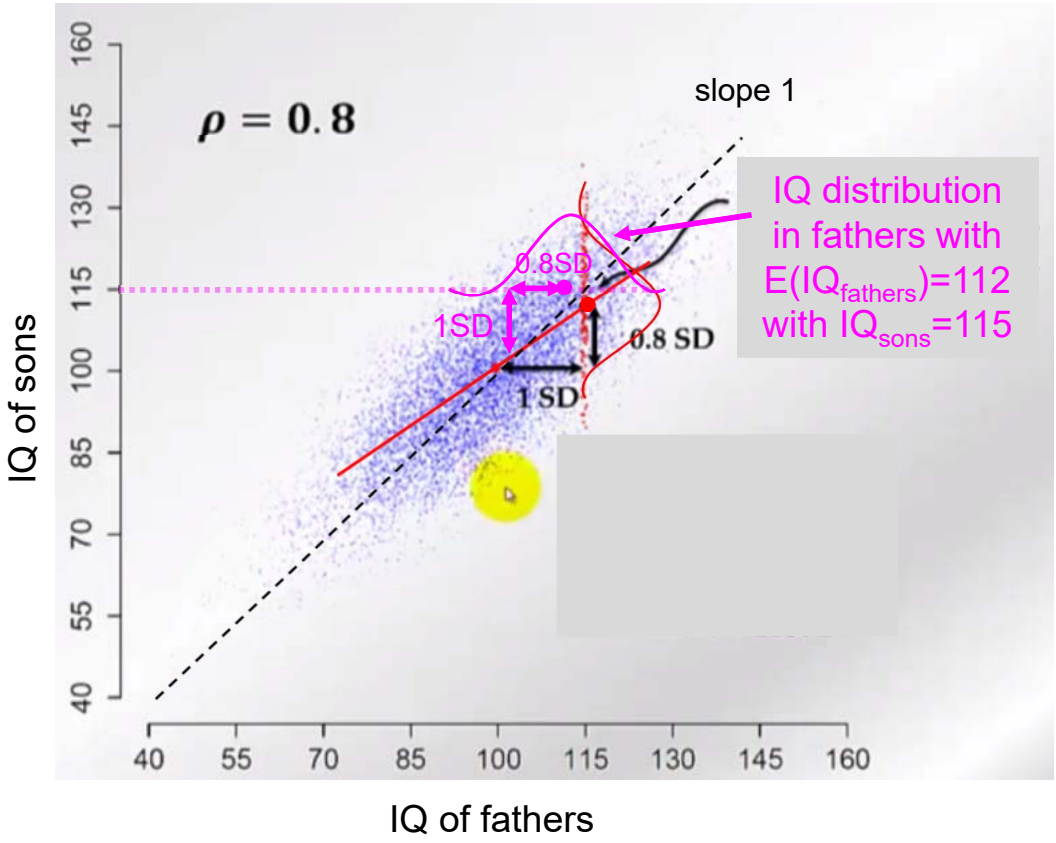


For each group of father with fixed IQ, the mean IQ of their sons is closer to the overall mean IQ (100) -> Galton aimed for a causal explanation.

All these predicted $E(IQ_{son})$ fall on a "regression line" with slope<1.

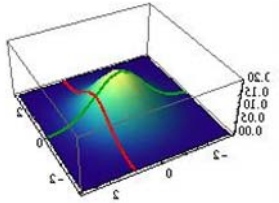
Image credits (changed): https://www.youtube.com/watch?v=aLv5cerjV0c

Galton's discovery of the regression to the mean phenomena



$$X1 \sim N\left(\mu_1 = 100,\ \sigma_1^2 = 15^2\right)$$
$$X2 \sim N\left(\mu_1 = 100,\ \sigma_1^2 = 15^2\right)$$

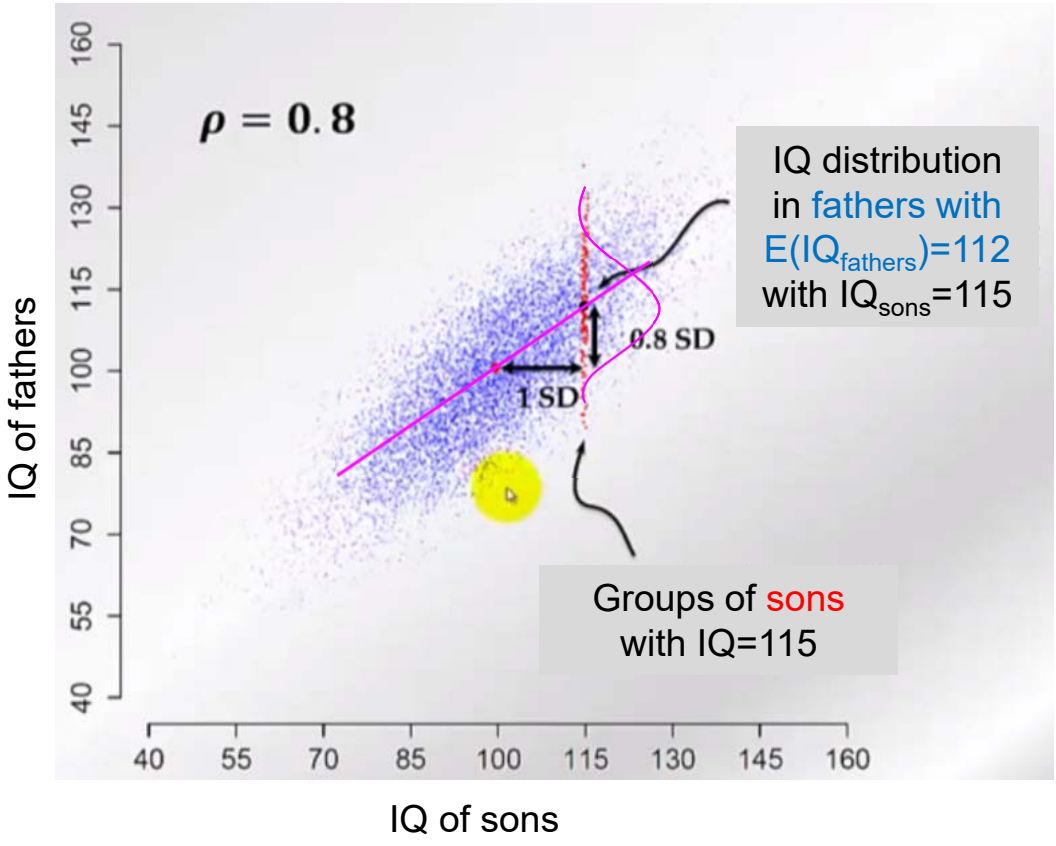
$$\begin{pmatrix} X1 \\ X2 \end{pmatrix} \sim N\left(\begin{pmatrix} 100 \\ 100 \end{pmatrix}, \begin{pmatrix} 15^2 & \mathrm{cov}(X1,X2) \\ \mathrm{cov}(X1,X2) & 15^2 \end{pmatrix}\right)$$



Also the mean of all fathers who have a son with IQ=115 is only 112.

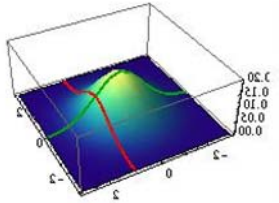
Image credits (changed): https://www.youtube.com/watch?v=aLv5cerjV0c

Galton's discovery of the regression to the mean phenomena



$$X1 \sim N\left(\mu_1 = 100,\ \sigma_1^2 = 15^2\right)$$
$$X2 \sim N\left(\mu_1 = 100,\ \sigma_1^2 = 15^2\right)$$

$$\begin{pmatrix} X1 \\ X2 \end{pmatrix} \sim N\left(\begin{pmatrix} 100 \\ 100 \end{pmatrix}, \begin{pmatrix} 15^2 & \text{cov}(X1,X2) \\ \text{cov}(X1,X2) & 15^2 \end{pmatrix}\right)$$

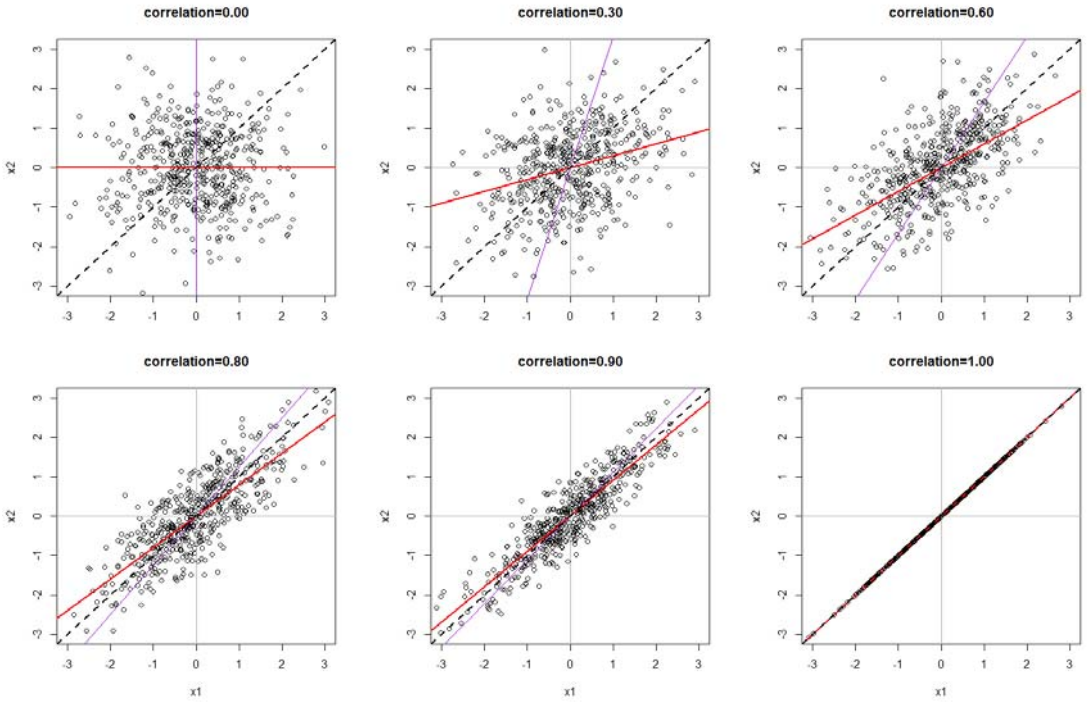


After switching the role of sons's IQ and father's IQ, we again see that E(IQ$_{fathers}$) fall on the regression line with the same slope <1.

There is no causality in this plot -> causal thinking seemed unreasonable.

Image credits (changed): https://www.youtube.com/watch?v=aLv5cerjV0c

Pearson's mathematical definition of correlation unmasks "regression to the mean" as statistical phenomena



The correlation c of a bivariate Normal distributed pair of random variables are given by the slope of the regression line after standardization!

c quantifies strength of linear relationship and is only 1 in case of deterministic relationship.

After standardization of the RV:

$$X1 \sim N\left(\mu_1 = 0,\ \sigma_1^2 = 1^2\right)$$
$$X2 \sim N\left(\mu_2 = 0,\ \sigma_2^2 = 1^2\right)$$

$$\begin{pmatrix} X1 \\ X2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{X1}^2 = 1 & c \\ c & \sigma_{X2}^2 = 1 \end{pmatrix}\right)$$

Regression line equation:

$$\hat{X}_2 = E\left(X_2 \mid X_1\right) = \beta_0 + \beta_1 \cdot X_1$$

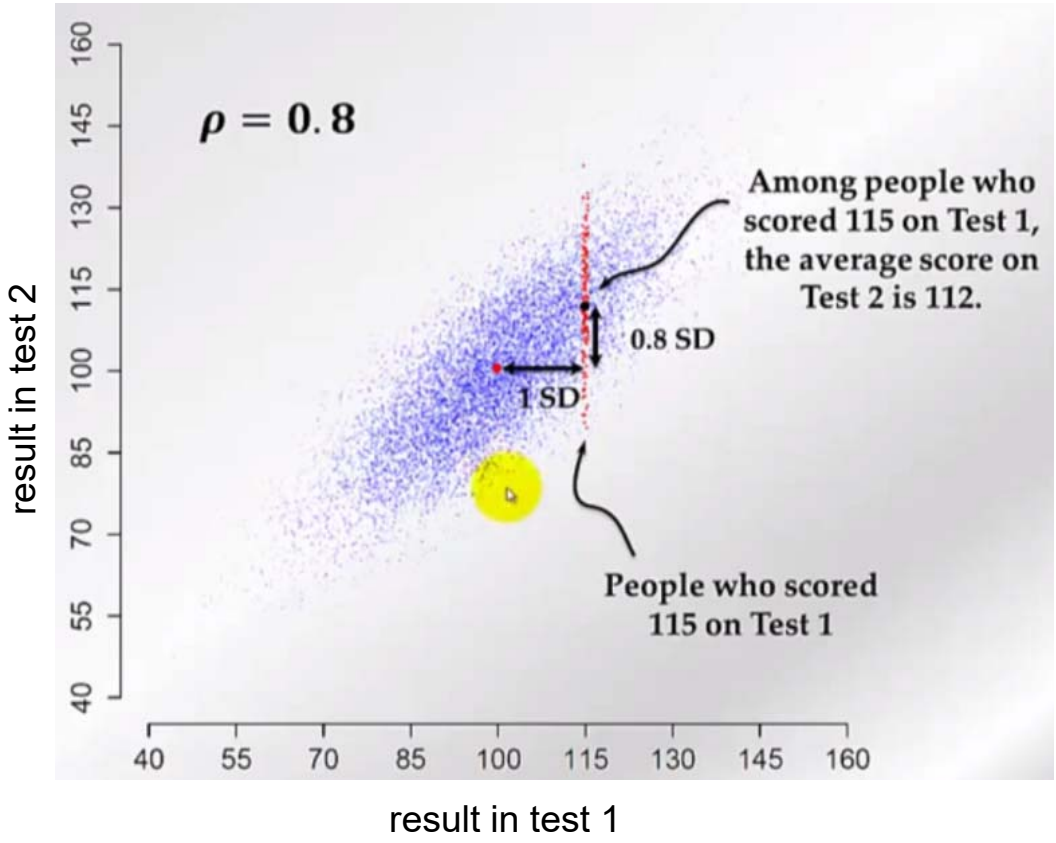
$$\beta_1 = c \cdot \frac{\sigma_2}{\sigma_1} = c$$

β_1 quantifies regression to the mean

$$\beta_0 = \mu_2 - \beta_1 \cdot \mu_1 = 0$$

$$c = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)\cdot(x_{i2} - \bar{x}_2)}{\mathrm{sd}(x_1)\cdot \mathrm{sd}(x_2)}$$

Regression to the mean occurs in all test-retest situations

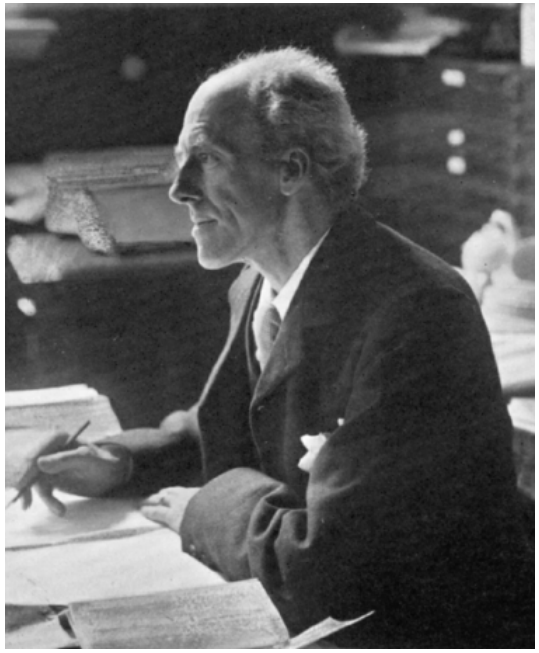


Retesting a extreme group (w/o intervention in between) in a second test leads in average to a results that are closer to the overall-mean -> to assess experimentally the effect of an intervention also a control group is needed!

With the correlation statistics was born and abandoned causality as "unscientific"

"the ultimate scientific statement of description of the relation between two things can always be thrown back upon… a contingency table [or correlation]."

Karl Pearson (1895-1936), The Grammar of Science



Pearl's rephrasing of Pearson's statment:
"**data is all there is to science**".

However, Pearson himself wrote several papers about "spurious correlation" vs "organic correlation" (meaning organic=causal?) and started the culture of "think: 'caused by', but say: 'associated with' "…

Quotes of data scientists

“Considerations of causality should be treated as they have always been in statistics: preferably not at all."

Terry Speed, president of the Biometric Society 1994

In God we trust. All others must bring data.

W. Edwards Deming (1900-1993), statistician and father of the total quality management

The world is one big data problem.

Andrew McAfee, Co-Rector MIT Initiative on the Digital Economy

Data without science is just data.

Elvis Murina, data scientist at ZHAW

See also http://bigdata-madesimple.com/30-tweetable-quotes-data-science/

Pearl's statements

Observing [and statistics and AI] entails detection of regularities

We developed [AI] tools that enabled machines to reason with uncertainty [Bayesian networks].. then I left the field of AI

Mathematics has not developed the asymmetric language required to capture our understanding that if *X* causes *Y* .

As much as I look into what's being done with deep learning, I see they're all stuck there on the level of associations. Curve fitting.

The book of Why
https://www.quantamagazine.org/to-build-truly-intelligent-machines-teach-them-cause-and-effect-20180515/

Probabilistic versus causal reasoning

Traditional statistics, machine learning, Bayesian networks

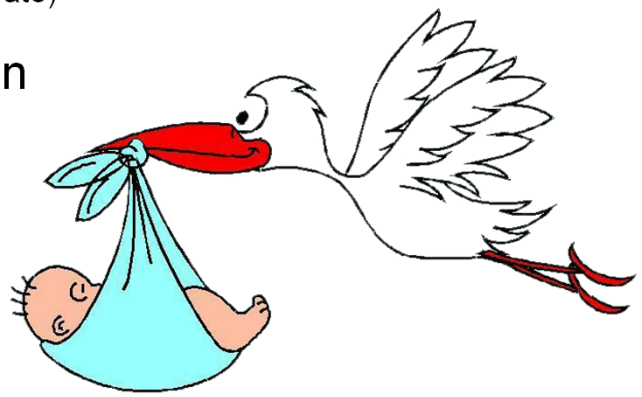
- About associations (stork population and human birth number per year are correlated)
- The dream is a models for the joined distribution of the data
- Conditional distribution are modeled by regression or classification

  (if we observe a certain number of storks, what is our best estimate of human birth rate?)

Causal models

- About causation (storks do not causally affect human birth rate)
- The dream is a models for the data generation
- Predict results of interventions

  (if we change the number of storks, what will happen with the human birth rate?)



Pearl's ladder of causality

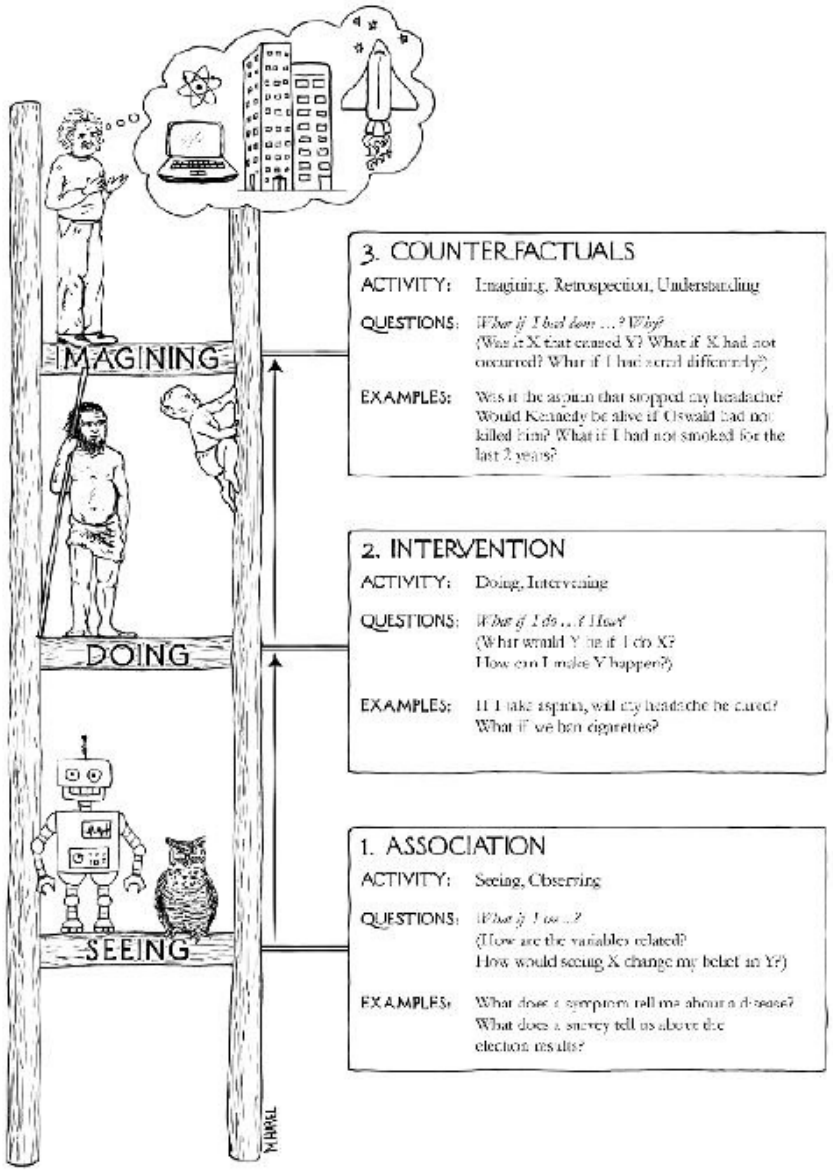
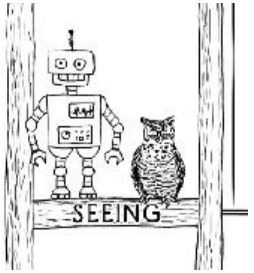


FIGURE 1.2. The Ladder of Causation, with representative organisms at each level. Most animals, as well as present-day learning machines, are on the first

Image credits: "The Book of Why"

On the first rung of the ladder
Pure regression can only model associations



P. Bühlman (ETH): “Pure regression is intrinsically the wrong tool”

(to understand causal relationships between predictors and outcome and to plan interventions based on observational data)”

Regression – the “statistical workhorse”: the wrong approach

we could use linear model (fitted from n observational data)

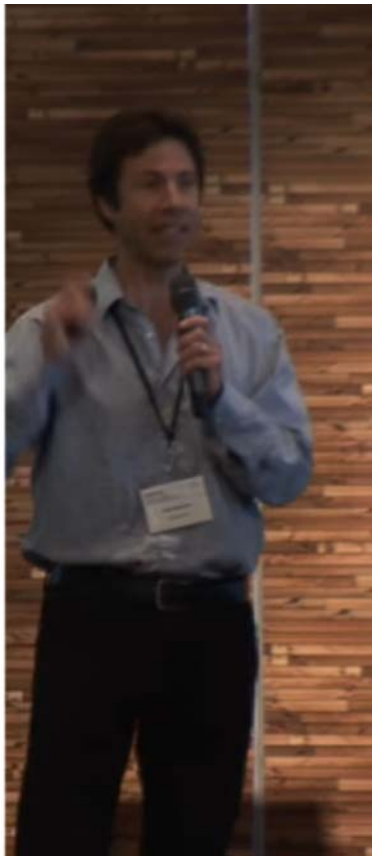
$$Y = \sum_{j=1}^{p} \beta_j X_j + \varepsilon,$$

$$\mathrm{Var}(X_j) \equiv 1 \text{ for all } j$$

$|\beta_j|$ measures the effect of variable X_j in terms of “association”

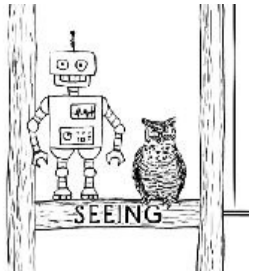
i.e. change of Y as a function of X_j when keeping all other variables X_k fixed

$\rightsquigarrow$ not very realistic for intervention problem
if we change e.g. one gene, some others will also change
and these others are not (cannot be) kept fixed



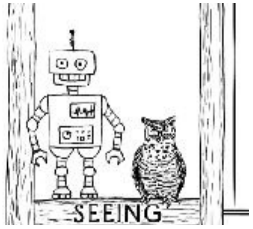
https://www.youtube.com/watch?v=JBtxRUdmvx4

How we work with rung-1 regression or ML models



xkcd.com

On the first rung of the ladder
DL is currently as good as a ensemble of pigeons ;-)



https://www.youtube.com/watch?v=NsV6S8EsC0E

PLOS ONE

RESEARCH ARTICLE

Pigeons (*Columba livia*) as Trainable Observers of Pathology and Radiology Breast Cancer Images

Richard M. Levenson[1]*, Elizabeth A. Krupinski[3], Victor M. Navarro[2], Edward A. Wasserman[2]*

1 Department of Pathology and Laboratory Medicine, University of California Davis Medical Center, Sacramento, California, United States of America, **2** Department of Psychological and Brain Sciences, The University of Iowa, Iowa City, Iowa, United States of America, **3** Department of Radiology & Imaging Sciences, College of Medicine, Emory University, Atlanta, Georgia, United States of America

* levenson@ucdavis.edu (RML); ed-wasserman@uiowa.edu (EAW)

Abstract

Pathologists and radiologists spend years acquiring and refining their medically essential visual skills, so it is of considerable interest to understand how this process actually unfolds and what image features and properties are critical for accurate diagnostic performance. Key insights into human behavioral tasks can often be obtained by using appropriate animal models. We report here that pigeons (*Columba livia*)—which share many visual system properties with humans—can serve as promising surrogate observers of medical images, a capability not previously documented. The birds proved to have a remarkable ability to distinguish benign from malignant human breast histopathology after training with differential food reinforcement; even more importantly, the pigeons were able to generalize what they had learned when confronted with novel image sets. The birds' histological accuracy, like that of humans, was modestly affected by the presence or absence of color as well as by degrees of image compression, but these impacts could be ameliorated with further training. Turning to radiology, the birds proved to be similarly capable of detecting cancer-relevant microcalcifications on mammogram images. However, when given a different (and for humans quite difficult) task—namely, classification of suspicious mammographic densities (masses)—the pigeons proved to be capable only of image memorization and were unable

OPEN ACCESS

Citation: Levenson RM, Krupinski EA, Navarro VM, Wasserman EA (2015) Pigeons (*Columba livia*) as Trainable Observers of Pathology and Radiology Breast Cancer Images. PLoS ONE 10(11): e0141357. doi:10.1371/journal.pone.0141357

Editor: Jonathan A Coles, Glasgow University, UNITED KINGDOM

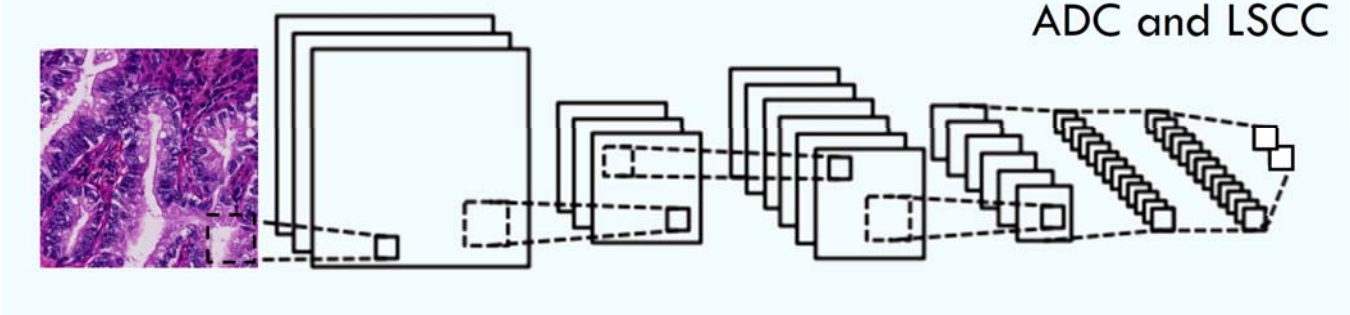
Received: August 25, 2015

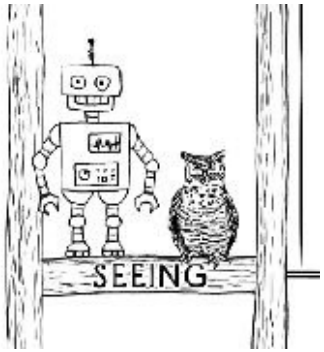
Accepted: October 7, 2015

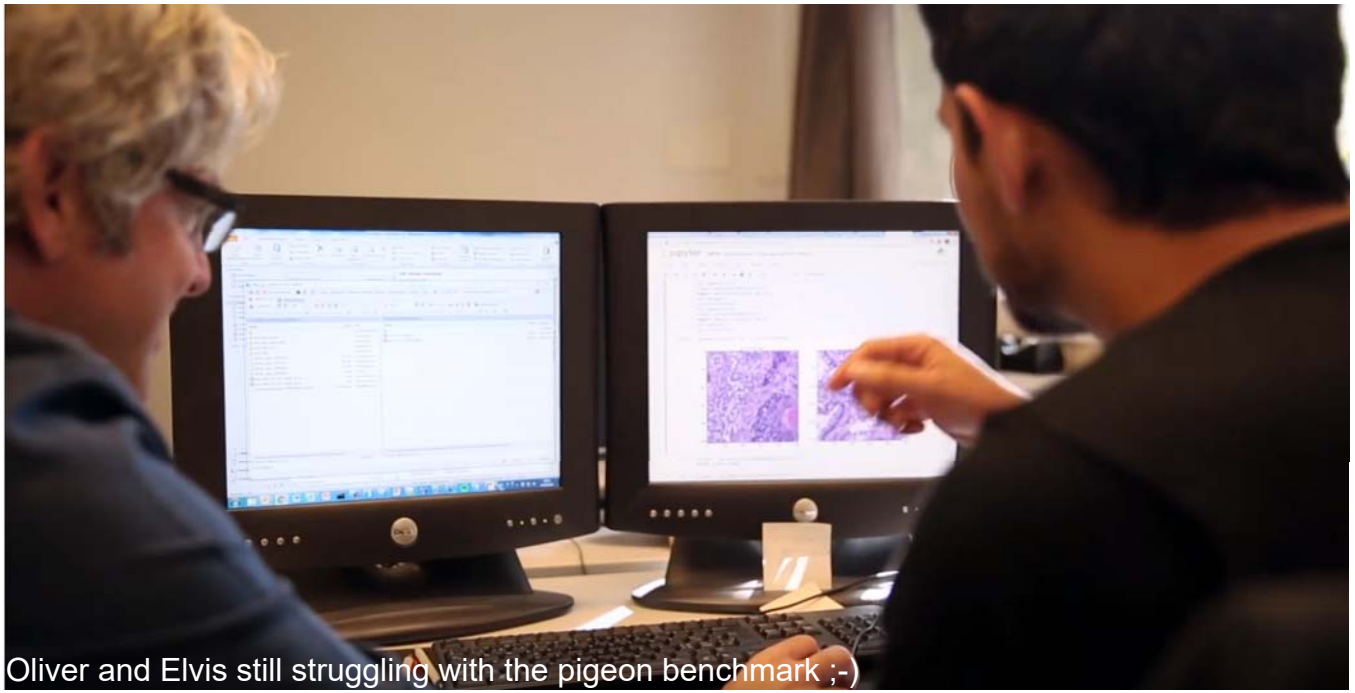
Published: November 18, 2015

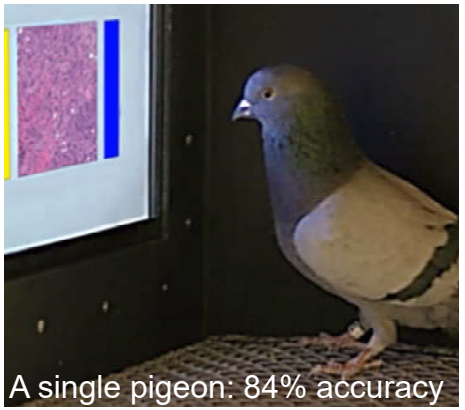
On the first rung of the ladder
DL is currently as good as an ensemble of pigeons

Elvis's DL model achieves ~90% accuracy on image level






Oliver and Elvis still struggling with the pigeon benchmark ;-)


A single pigeon: 84% accuracy

Can and should we try to learn about causal relationships?

If yes – what and how can we learn?

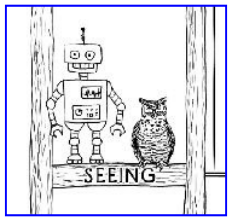
Ascending the second rung by going from "seeing" to "doing"

Research question:

What is the distribution of the blood pressure if people do not drink coffee?

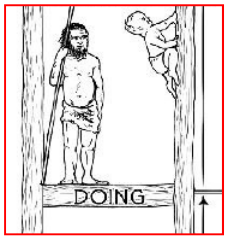
Conditioning:
Filter - restrict on non-coffee drinker

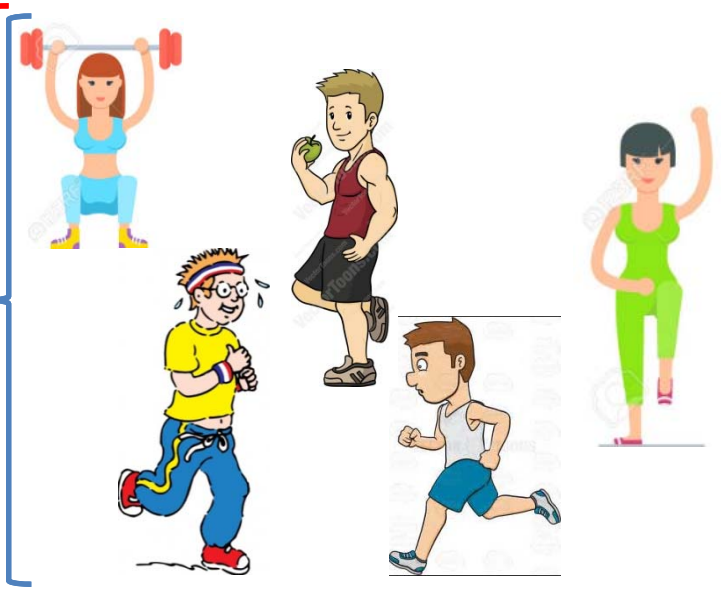
$$P(\mathrm{BP} \mid \mathrm{coffee} = 0)$$



"Do"-Operator:
Full population, after intervention that prohibits coffee consume

$$P(\mathrm{BP} \mid \mathrm{do}(\mathrm{coffee} = 0))$$

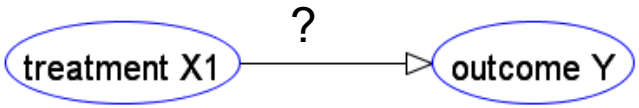


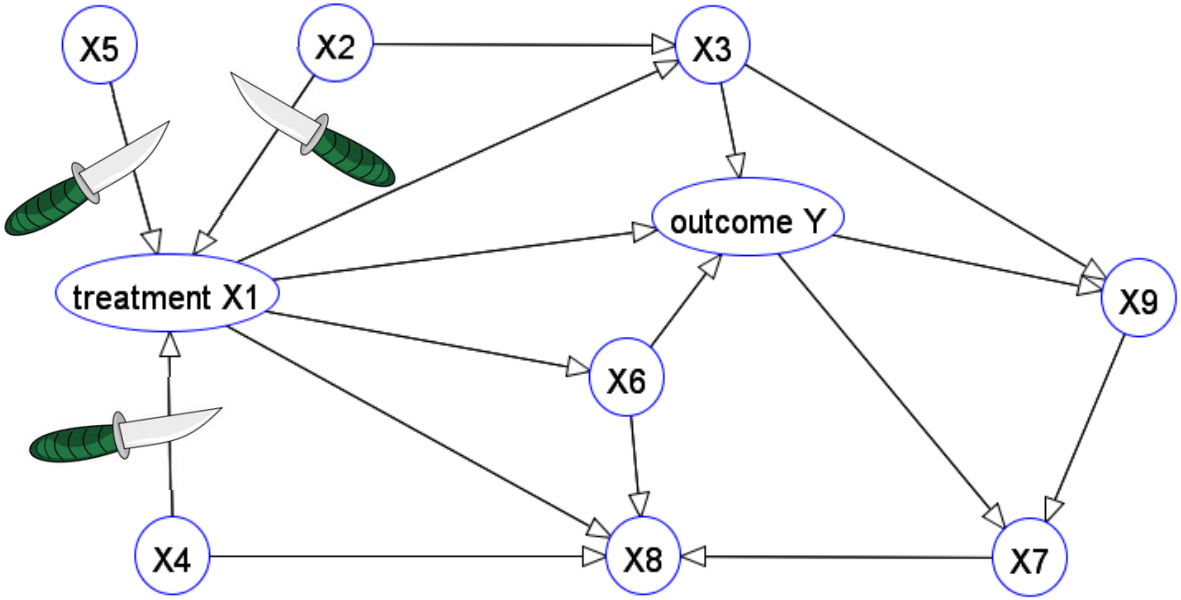


non-coffee drinker by choice

coffee drinker by choice

On the second "doing" rung of the ladder
Assessing the effect of intervention by randomized trials (RT)

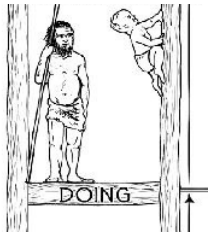




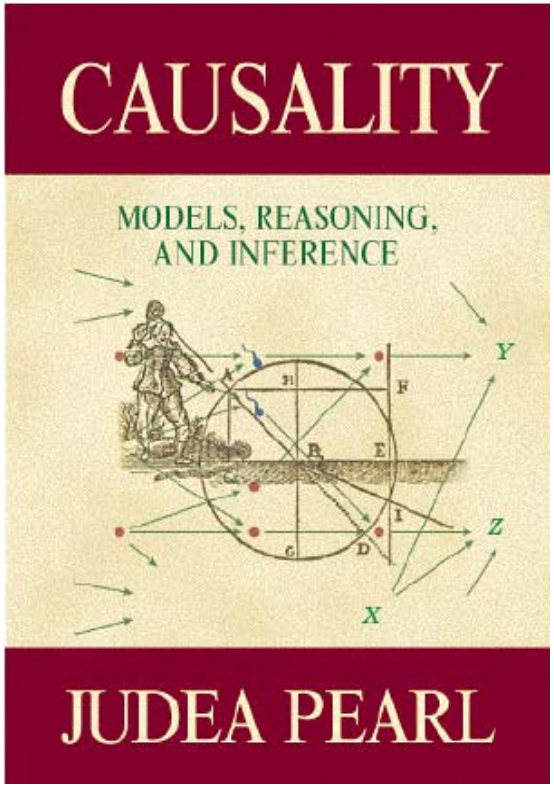
RCT through the lens of a causal graphical model

Since the treatment is assigned randomly to both treatment groups are exchangeable. Hence observed differences of the outcome in both groups is due to the treatment.

-> Model after collecting data from a RT: $outcome \sim treatment$



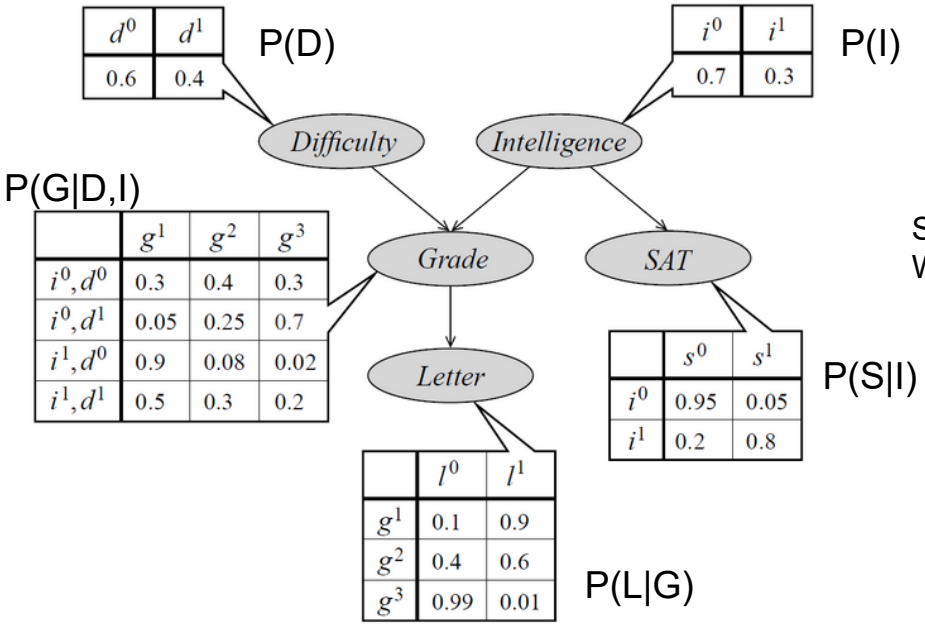
Judea Pearl broke with the taboo of causal reasoning based on observational data



ACM Turing Award 2011: “For fundamental contributions to artificial intelligence through the development of a calculus for probabilistic and causal reasoning."

Recap: BN interpretation

A probabilistic Bayesian network is a DAG about association where each node is a variable that is independent of all non-descendants given its parents



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

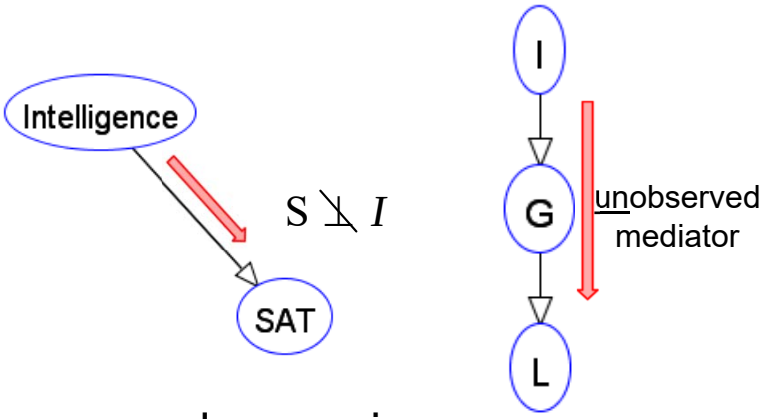
	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

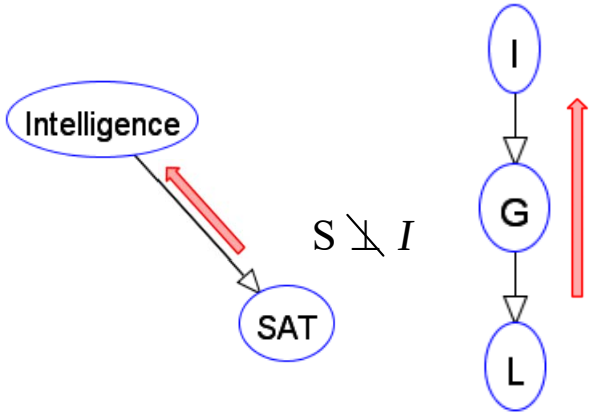
SAT (*Scholastic Assessment Test*)
Widely used for college admission

The example is taken from the great course of Daphne Koller on probabilistic graphical models.

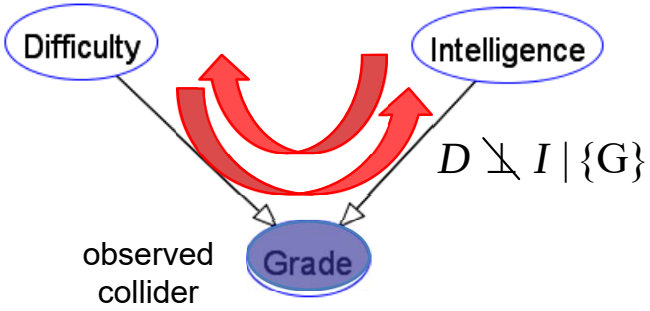
Recap: Open paths allowing belief to flow



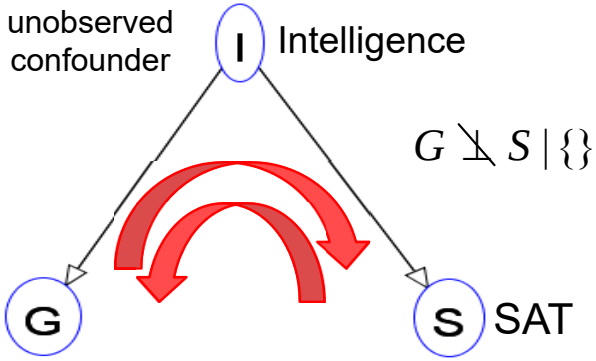
causal reasoning



evidence based reasoning



inter-causal reasoning

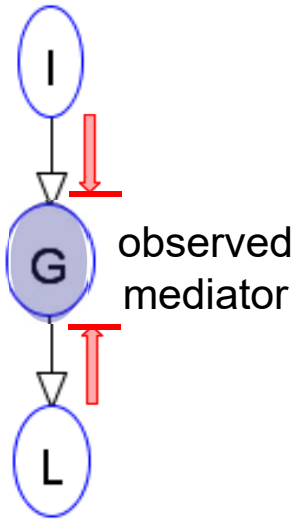


inter-effect reasoning

Recap: Closed paths not allowing belief to flow

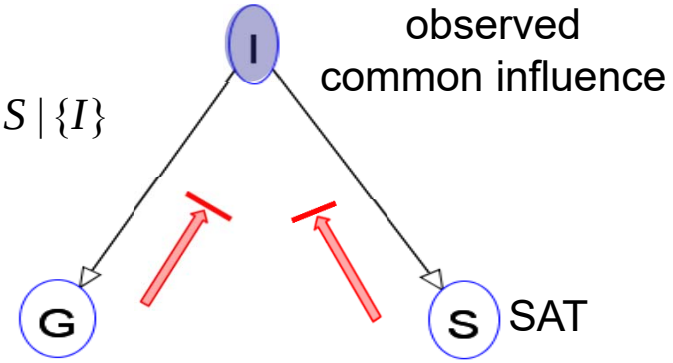
no flow of (causal) belief

$I \perp L \mid \{G\}$

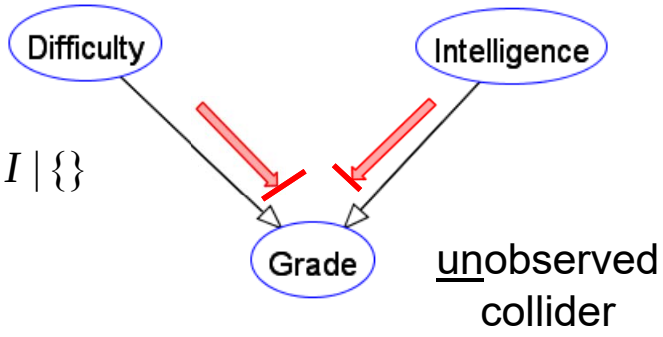


no flow of non-causal belief

$G \perp S \mid \{I\}$



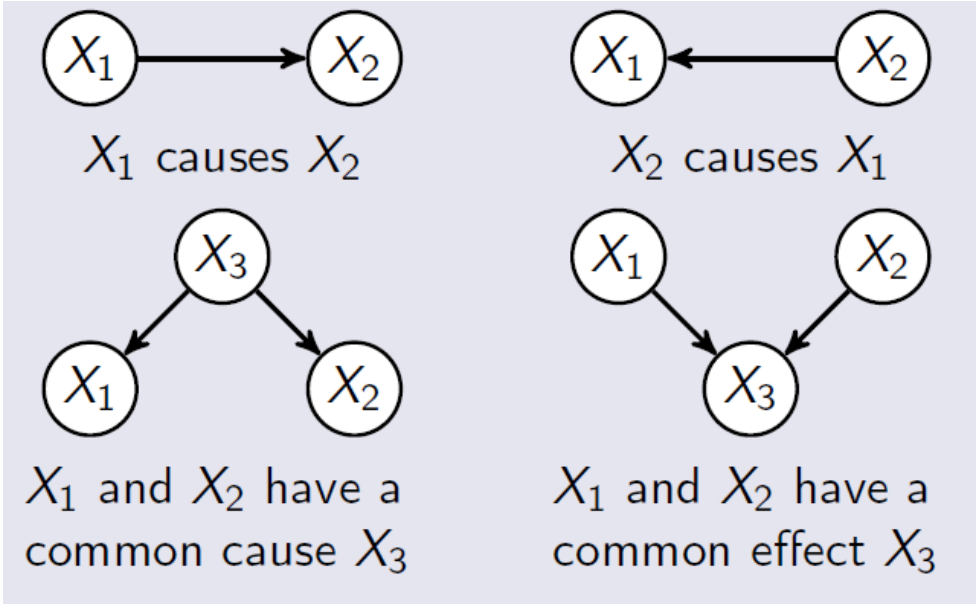
$D \perp I \mid \{\}$

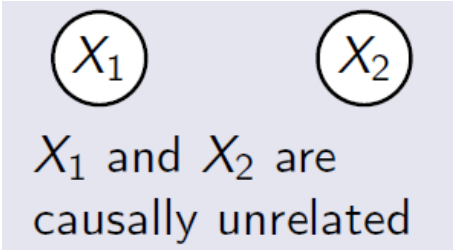


To avoid flow of non-causal belief - we must observe confounders!
- we must not observe colliders!

From Bayesian networks to causal Bayesian networks

A causal BN is a DAG about causal relationships where again nodes are variables, but a directed edge represents a potential causal effect.





Causal effects can only be transported along the direction of arrows!

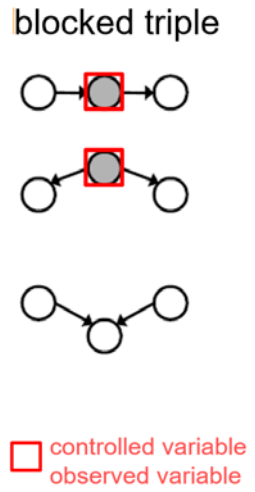
Pearl's backdoor criterion for causal Bayesian Networks

Backdoor paths between X and Y are not directed from X to Y and transported association is spurious.

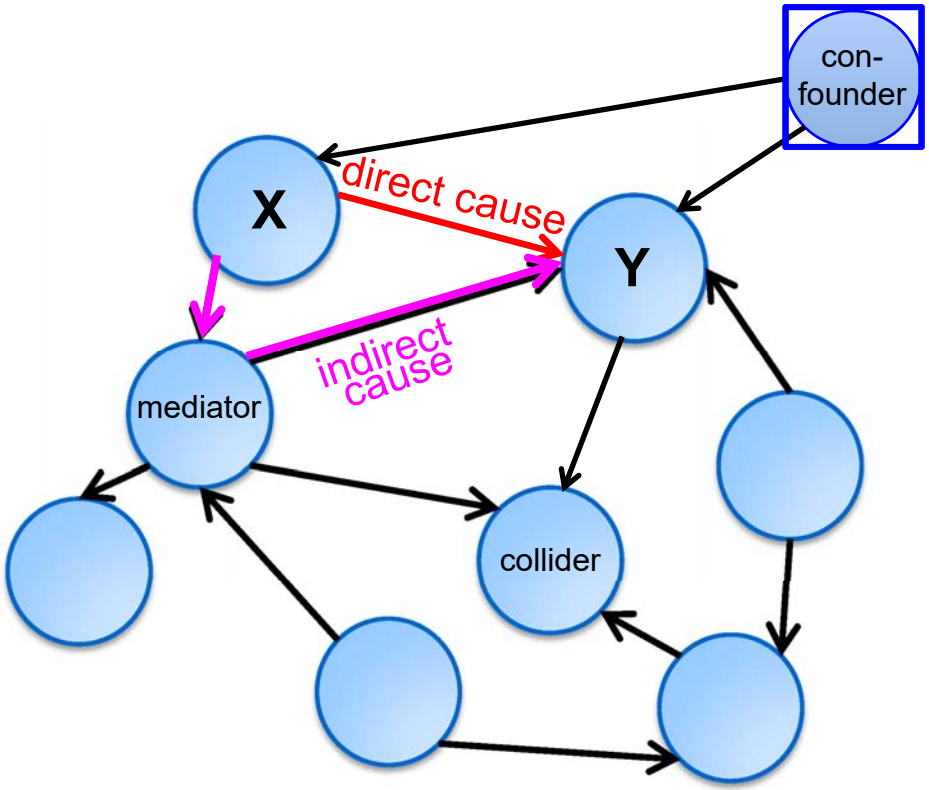
- We want to block all backdoor paths
- Determine a set S of "de-confounders" that closes all backdoor paths and control for these variables.

  Observe them and use them as co-variates in your model – the coefficient in front of X gives then the causal effect of X on Y!

A path is blocked if **1 single** triple-segment is blocked!



Causal effects are only transported along arrows from X to Y



To close all backdoor paths we must adjust for this confounder.

Here we have two paths along which a causal effect can be transported.
(If we add the direct and the indirect causal effect we get the total causal effect.)

All black paths do either transport non-causal belief or block the flow of belief.
(here only the upper right backdoor path is open as long as we do not adjust for the common cause of x and y, all other backdoor paths are blocked by unobserved colliders)

The classic epidemiological definition of confounding

A treatment X and outcome Y is confounded by a variable *Z* if

(1) Z associated with *X*

(2) Z associated with *Y* even if *X is fixed*.

only using statistical terms and not sufficient!

Simpsons addition in 1951

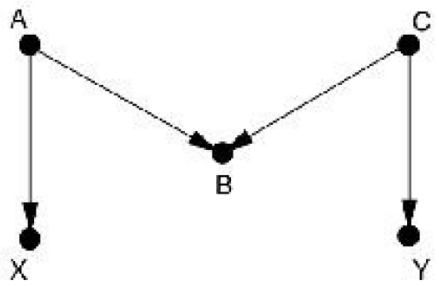
To avoid adjusting for a mediator this has been supplemented in recent years by

(3) *Z* should not be on the *causal* path between *X* and *Y*.

Added causal terms still not sufficient!

The classical confounding definition allows M bias

Backdoor path from X to Y



B fulfils all 3 confounder criteria:

- B is associated with X
- B is associated with Y (even if X is fixed)
- B does not lie on a causal path X to Y

However, controlling for B opens the backdoor path and introduces spurious association!

X: smoking
Y: lung disease
B: seat-belt usage
A: following social norms
C: health risk taking

A study conducted in 2006 investigating the effect of smoking (X) on lung diseases (Y) listed seat-belt usage (B) as one of the first variables to be controlled.

Pearl's valid definition of the concept "confounding"

Confounding, is anything that leads to a

discrepancy between the conditional probability

and the interventional probability between X and Y:

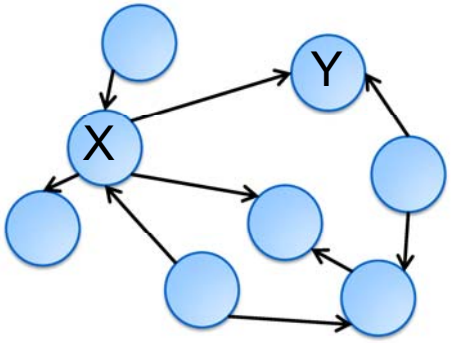
$$P(Y \mid X) \neq P(Y \mid do(X))$$

Can we do causal/intervential inference from observational data?

The very short answer: No!

Principle be Cartwright (1989): **No causes in – no causes out!**

observational data



Backdoor criterion or frontdoor criterion or 3 Rules of do-Calculus

$$P'(y \mid do(X = x_0))$$

$$=$$

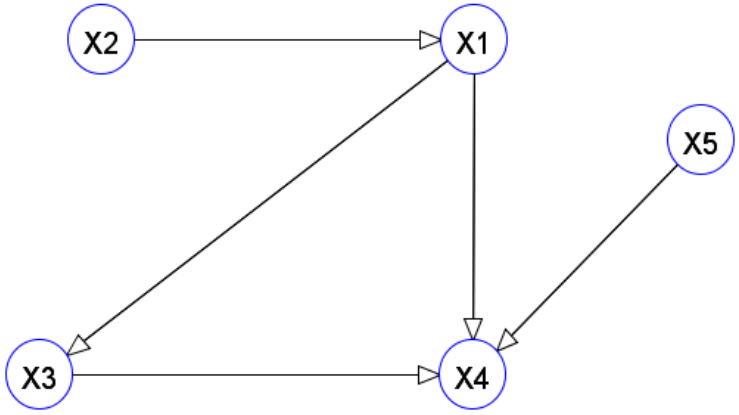
Expression without do (!!) which only uses information from observed JPD P

Ascending the third "imaging" rung of the ladder
Causal BN to predict intervention effect

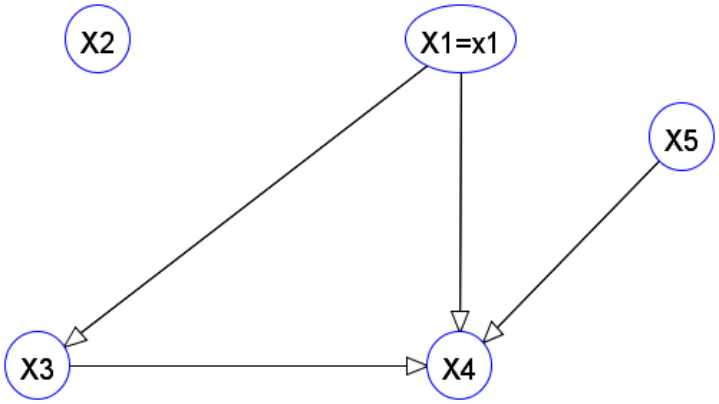
Intervention at variable X1:
do(X1=x1) implying that all arrows into X1 are deleted

Assumption: the remaining graphical model does not change under the intervention.

before intervention



after intervention on X1



$$P(X_1,X_2,X_3,X_4,X_5) \overset{\text{chain rule for BN}}{=}$$
$$P(X_2)\cdot P(X_1|X_2)\cdot P(X_3|X_1)\cdot P(X_4|X_1,X_3,X_5)\cdot P(X_5)$$

$$P(X_1,X_2,X_3,X_4,X_5 \mid do(X_1=x_1)) =$$
$$P(X_2)\cdot 1 \cdot P(X_3|X_1{=}x_1)\cdot P(X_4|X_1{=}x_1,X_3,X_5)\cdot P(X_5)$$
$$\text{II}$$
$$P(X_1=x_1)$$

On the third "imaging" rung of the ladder: imaging "do" operator opens the door to rung 3

What if??

How would the world look like if Dino's would have survived?

Would he live longer if he would always eat an apple instead of a cake?

Would we have earned more if we had doubled the price?

The unobserved outcome is called counterfactual.

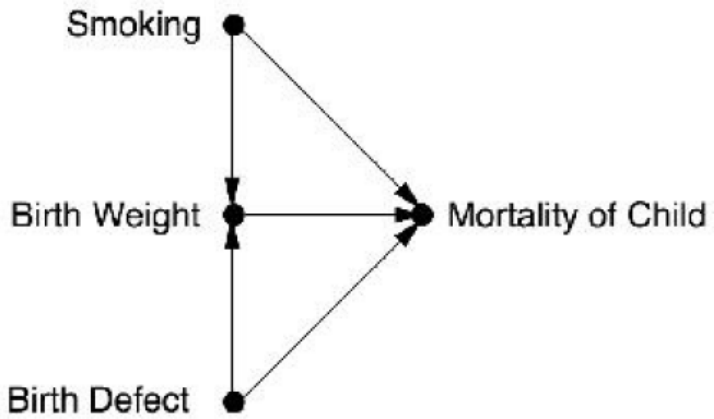
Historic anecdotes of of (non-) causal thinking

Are smoking mothers for underweighted newborns beneficial?

Since 1960 data on newborns showed consistently that low-birth-weight babies of smoking mothers had a better survival rate than those of nonsmokers.

This paradox was discussed for 40 years!

An article by Tyler VanderWeele in the 2014 issue of the *International Journal of Epidemiology* nails the explanation perfectly and contains a causal diagram:

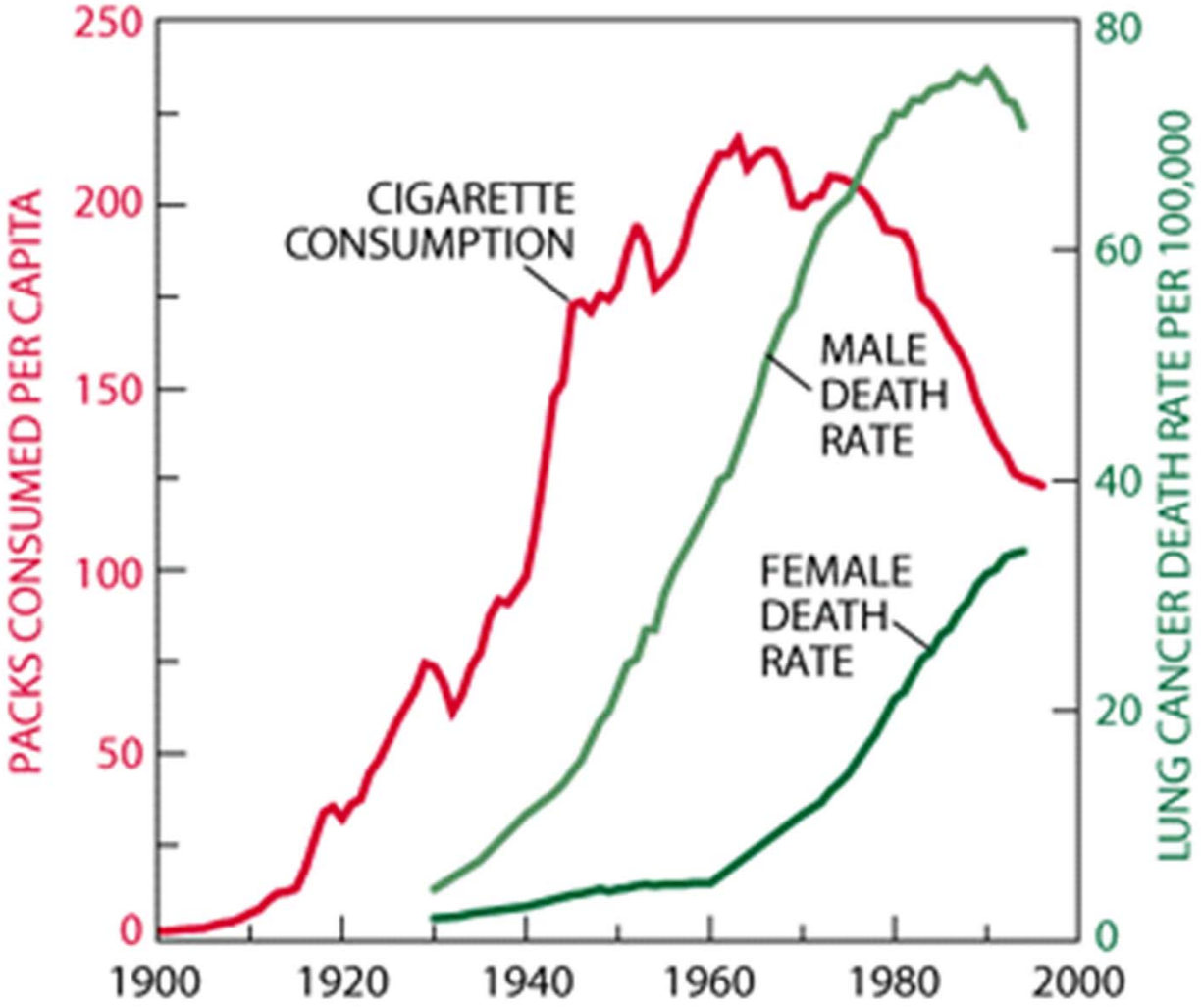


Association is due to a collider bias caused by conditioning on low birth weight.

Image credits: “The Book of Why”

BB Seminar ended here, discussion started

The smoking debate



1948, Doll and Bradford Hill investigated smoking as potential cause for lung cancer.

Marketing of the tobacco industry

Image credits: "The Book of Why"

George Weissman,
vice president of Philip Morris, 1954:

"If we had any thought or knowledge that in any way we were selling a product harmful to consumers, we would stop business tomorrow."

Observed association between lung cancer and smoking

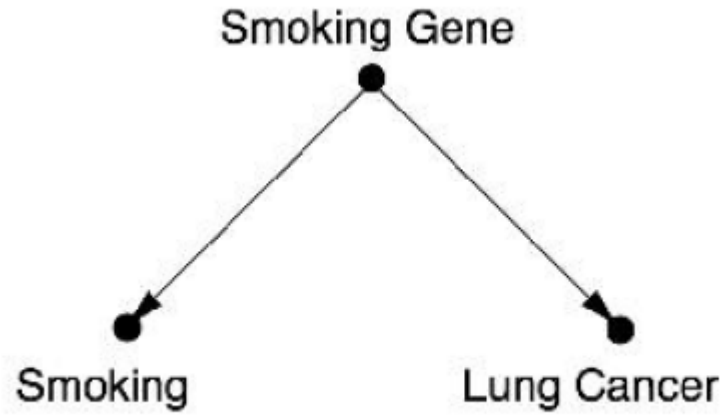
- 99.7% of lung cancer patients were smokers
  (retrospective study result)

- smokers have 30-times higher probability to die by lung-cancer within the next 5 years than non-smokers
  (Hill's 60,000 British physicians prospective study result)

- heavy smokers have 90-times higher probability to die by lung-cancer within the next 5 years than non-smokers
  (prospective study result)

Fisher's skeptics of the smoking-cancer connection

Ronald Fisher
(1890-1962)



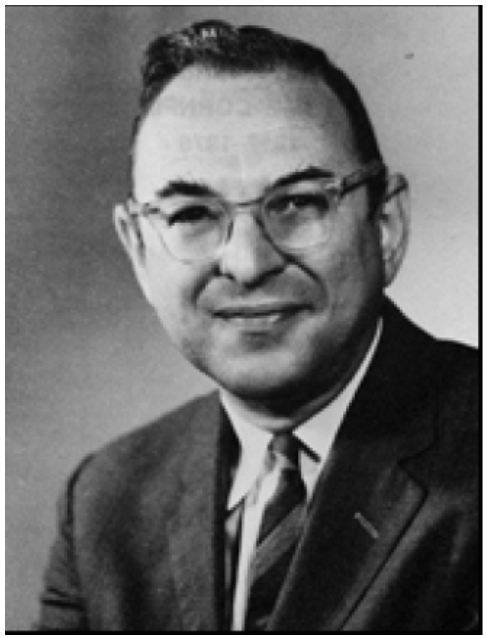
Fisher insisted, that the observed association could be due to an confounder such as smoking gene causing the longing for smoking and a higher risk for LC.

Cornfield's inequality

$$RR_{obs} \leq RR_{unobs}$$

$$RR_{obs} = \frac{q_1 \cdot RR_{unobs} + 1 - q_1}{q_0 \cdot RR_{unobs} + 1 - q_0} \leq \frac{q_1}{q_0} = \frac{P(U \mid E)}{P(U \mid \bar{E})}$$

Jerome Cornfield
(1912–1979)



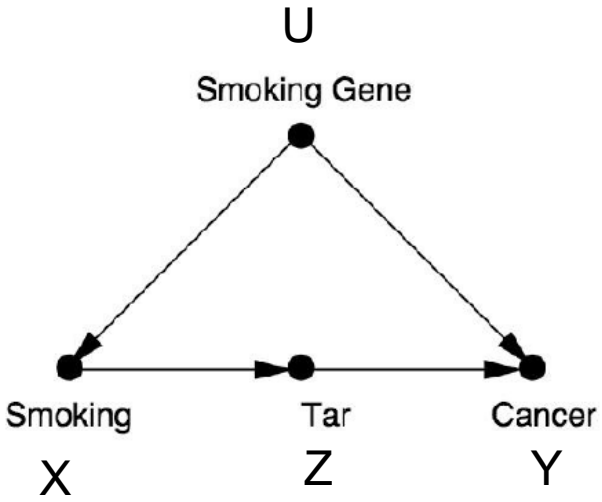
The unknown confounder U needs to be ≥K-times more common in smokers to explain a K-times higher risk for LC of smokers compared to non-smokers (RR=K).

If RR=10 and 10% of non-smokers have the "smoking gene," then 100% of the smokers would have to have it.

If 12% of non-smokers have the smoking gene, then it becomes impossible for the cancer gene to account fully for the association between smoking and cancer.

See also http://www.statlit.org/Cornfield.htm

Front-door criterion can handle unobserved confounder



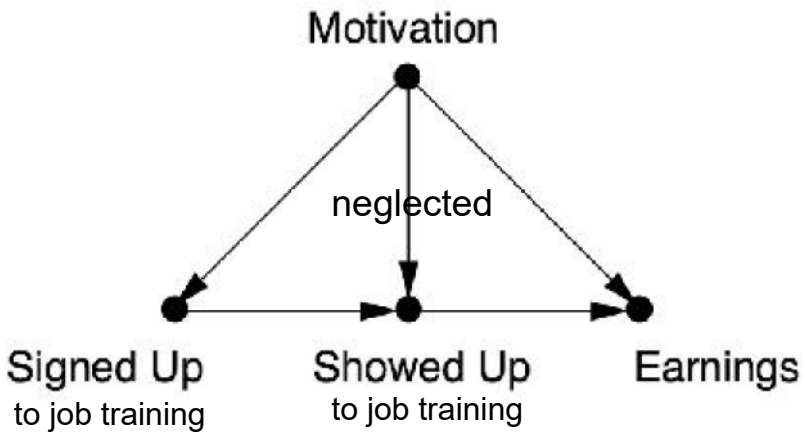
For a proof of the **front door approach** see figure 7.4 in "The Book of Why"
Anytime the causal effect of *X* on *Y* is confounded by one set of variables (*U*) and mediated by another (*Z*) and the mediating variables are shielded from the effects of *U*, then you can estimate *X*'s effect on Y from observational data.

In this way we can determine the causal effect of Smoking on LC.

The corresponding formula only requires observable probabilities:

$$P(Y \mid do(X)) = \sum_Z P(Z = z, X) \sum_X P(Y \mid X = x, Z = z)\, P(X = x)$$

Application: Effect estimation of a job training program



Observational data from Job Training Partnership Act (JTPA) Study1987-89.

After estimating the intervention effect from observational study data by using the front-door formula, a randomized trial was performed showing an effect that almost perfectly matched the predicted effect!

Glynn, A., and Kashin, K. (2018). Front-door versus back-door adjustment with unmeasured confounding: Bias formulas for front-door and hybrid adjustments. *Journal of the American Statistical Association.*

Image credits: “The Book of Why”

Pearl's statements about the future of AI

Interview question: What are the prospects for having machines that share our intuition about cause and effect?

Pearl's answer:

We have to equip machines with a [causal] model of the environment. If a machine does not have a model of reality, you cannot expect the machine to behave intelligently in that reality.

The first step, one that will take place in maybe 10 years, is that conceptual models of reality will be programmed by humans.

The next step will be that machines will postulate such models on their own and will verify and refine them based on empirical evidence.

https://www.quantamagazine.org/to-build-truly-intelligent-machines-teach-them-cause-and-effect-20180515/
https://www.acm.org/turing-award-50/video/neural-nets

Thanks for your attention!