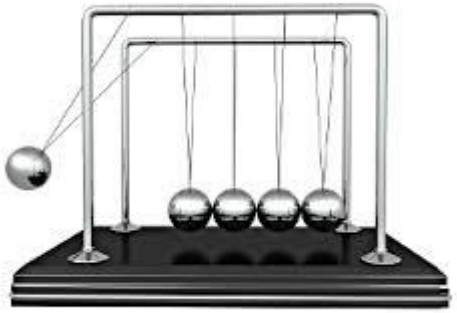
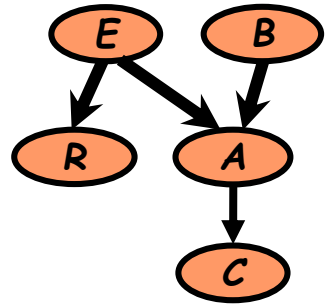
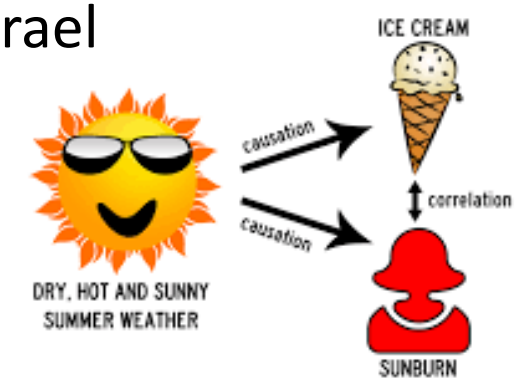




C A U S A L I T Y

Professor Ron S. Kenett

KPA Group, Israel and Neaman Institute, Technion, Israel



Quality Engineering, 27:111–121, 2015

ISSN: 0898-2112 print / 1532-4222 online
DOI: 10.1080/08982112.2015.968054

Taylor & Francis

Statistics: A Life Cycle View

Ron S. Kenett
KPA Group, Ra'anana, Israel

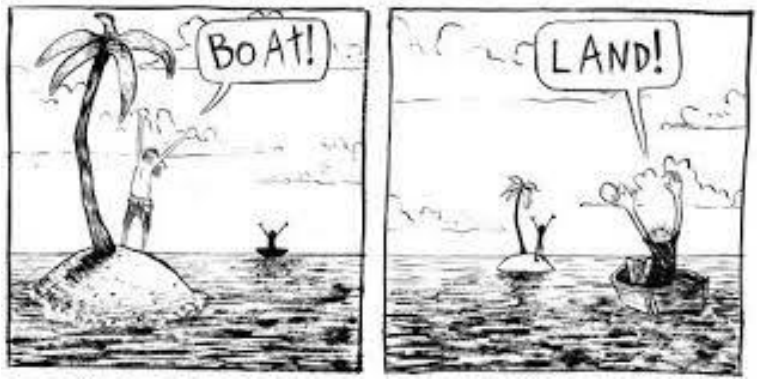
ABSTRACT Statistics has gained a r…d only on mathematical modeling, data collection,…le is about an expanded view of the role of statisti…ustry, and service organizations. Such an approac…he narrow view of statistics outlined above in an e… to reposition it as the leading profession in the analytics domain. Thi…w creates a need for focused research activities and the development o…n in collaboration of statisticians with experts …12). Specifically we discus here a "life cycle …lem elicitation, (2) goal formulation, (3) data …(5) …on of findings, (6) operationalizatio…findings, (7) communication, …impact assessment. These eight p…, with …tions …the inductive–deductive learning p… …overing these phases, beyond the data… …f statistical analysis and enhances the leve… …on quality. The envisaged overall approac… …nd practice of applied statistics needs to in…a trilogy combining …cycle view, (2) an analysis of impact, and (3)… …f the generated information and knowledge. We… …ng the problem, continue with a review of the i… …oncept presented in Kenett and Shmueli (2014), a… …n of the eight life cycle phases listed above. Adopting a life…e view of statistics has obvious implications to research, education, and sta…l practice. These are presented in the context of several examples. We conclude with a discussion of such implications.

Information quality

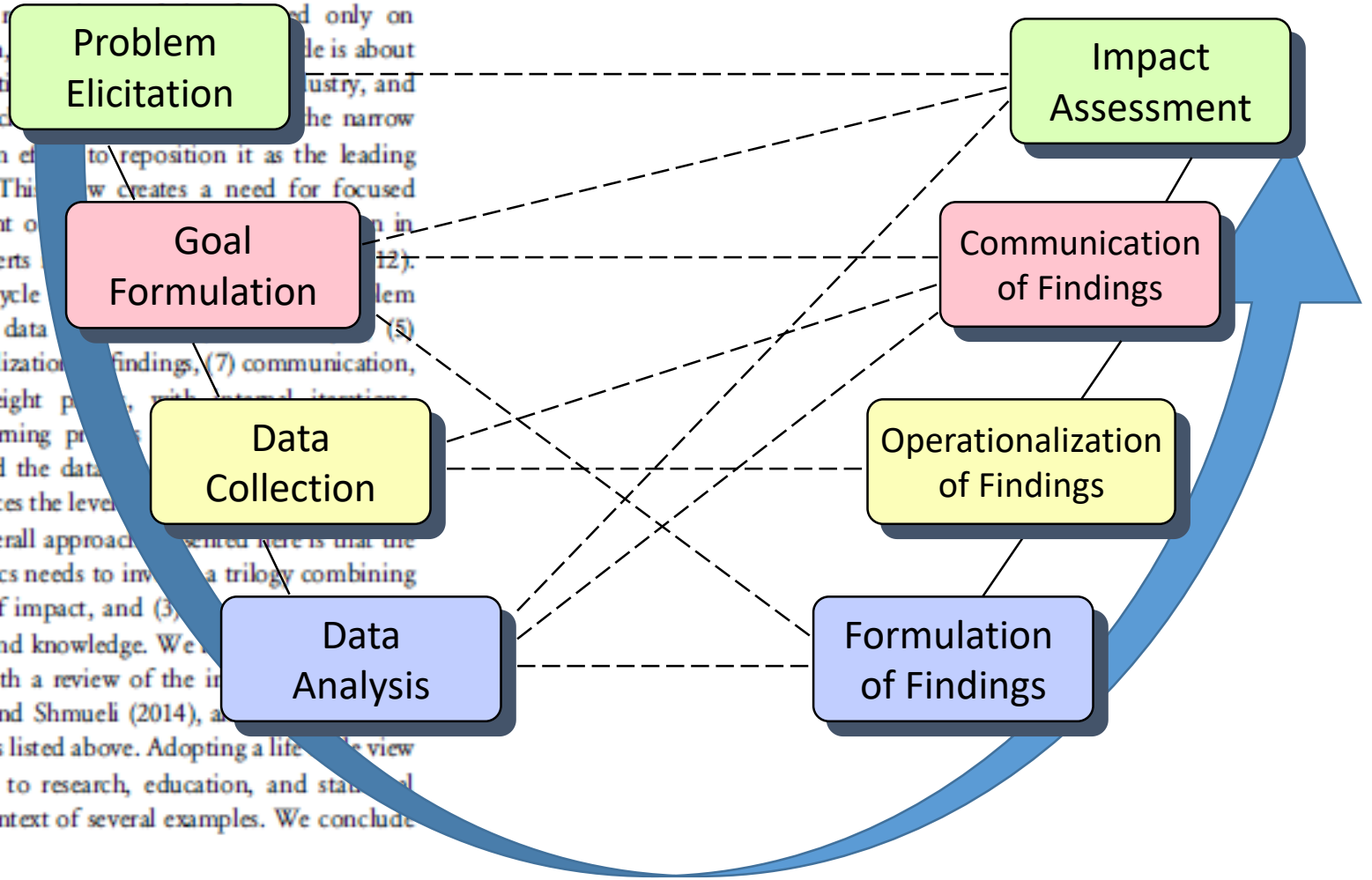
InfoQ(f,X,g,U) = U(f(X\|g))

InfoQ dimensions

- Data resolution
- Data structure
- Data integration
- Temporal relevance
- Chronology of data and goal
- Generalizability
- Operationalization
- Communication



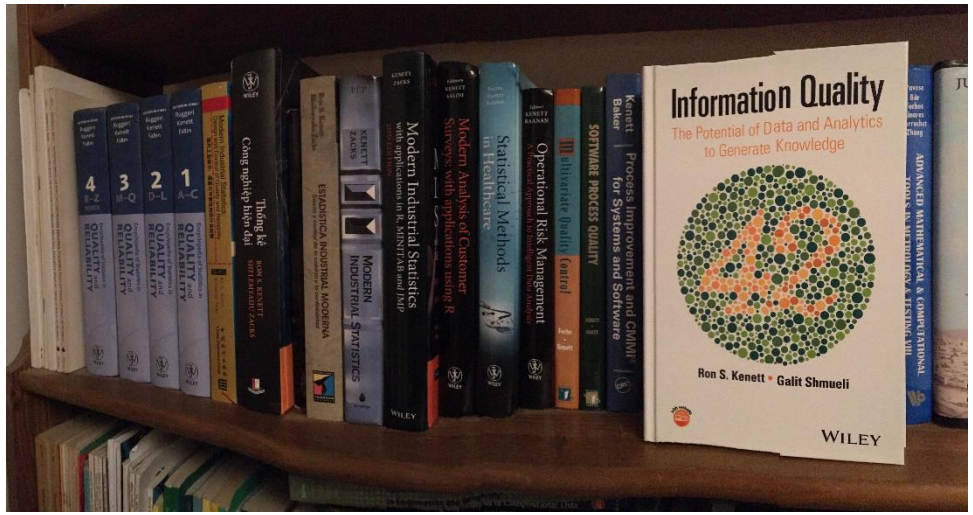
My point of view



"After all, it is all about information quality….."

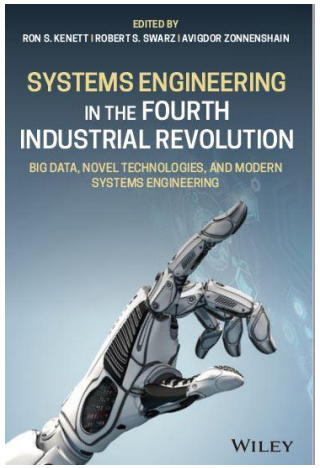
My motto

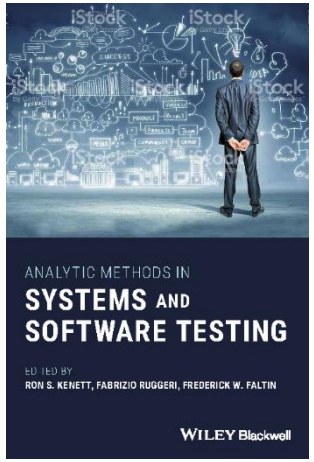
Applied statistics
is about meeting the challenge of
solving **real world problems**
with **mathematical tools**
and **statistical thinking**

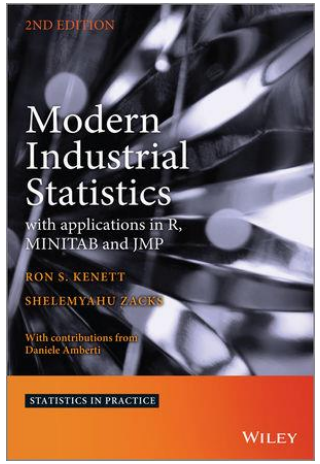


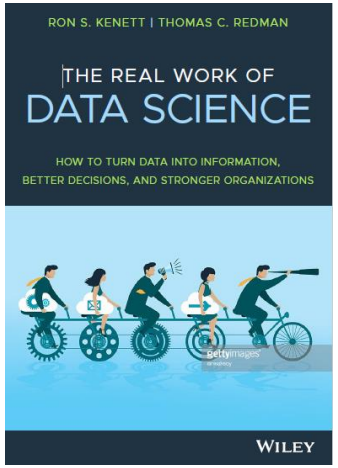
2018 ENBIS Box Medal

2013 RSS Greenfield Medal









Agenda

1. Background on causality in science and statistics
2. Fishbone cause and effect diagrams
3. Bayesian networks
4. Randomization in experimental designs
5. Propensity scores in observational studies
6. Counterfactuals and do calculus
7. Personalized medicine, condition based maintenance and Industry 4.0
8. Future research areas

Agenda

1. **Background on causality in science and statistics**
2. Fishbone cause and effect diagrams
3. Bayesian networks
4. Randomization in experimental designs
5. Propensity scores in observational studies
6. Counterfactuals and do calculus
7. Personalized medicine, condition based maintenance and Industry 4.0
8. Future research areas

"Statistics is important because it is conceived as contributing to a **causal understanding** ...
Statistics can indicate causality *even in the absence of a mechanistic understanding*.
But the traditional self-conception of statistics is that it can rarely say anything about causality.
This is a ***paradox***."

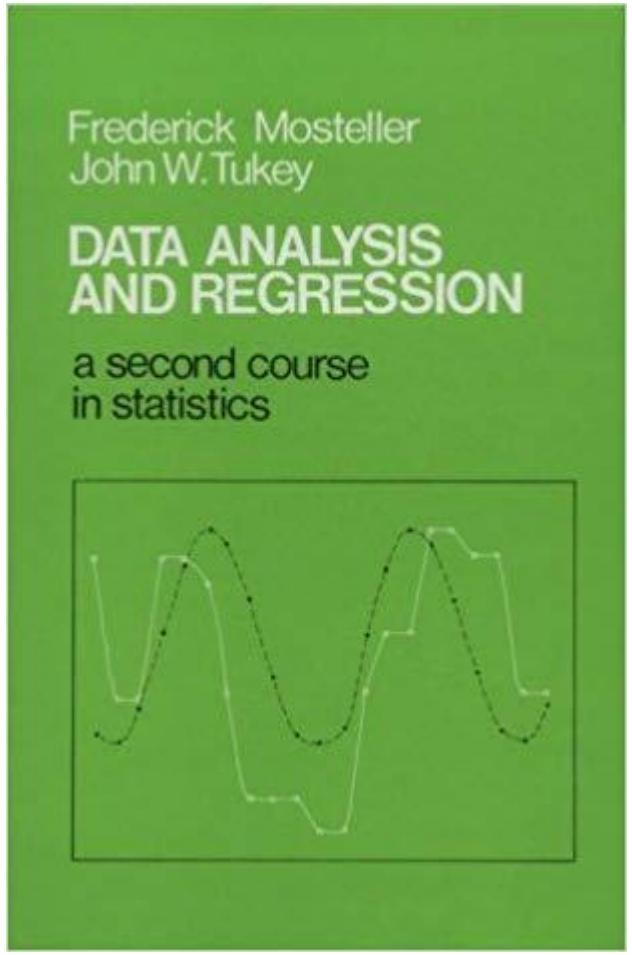
*Statistikk 50 År! Some remarks on causality**

Odd O. Aalen

*From a presentation celebrating 50 years to the establishment of a Masters Degree in Statistics in Norway, May 22, 2006

A journey back into the past

2006
1977
1953
1936
1931
1923
1921
1904
1886
1738



Data analysis and regression : a second course in statistics, Addison-Wesley, 1977

Frederick Mosteller
1916-2006

John Wilder Tukey
1915-2000

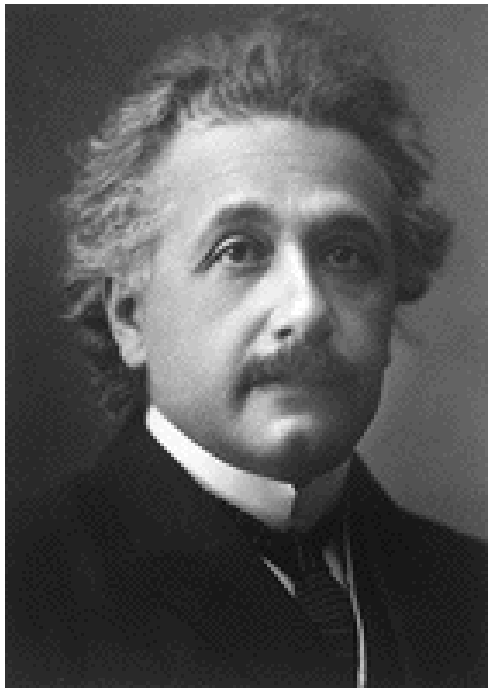
Causation:

1. Consistency
2. Responsiveness
3. A mechanism

"causation, though often our major concern, is usually not settled by statistical arguments"

Albert Einstein (1879-1955)

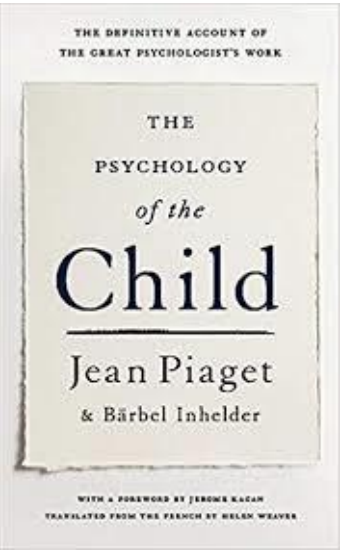
"Development of Western science is based on two great achievements: the invention of the formal logical system (in Euclidean geometry) by the Greek philosophers, and **the discovery of the possibility to find out causal relationships by systematic experiment** (during the Renaissance)."



A. Einstein, April 23, 1953

Jean Piaget (1896 – 1980)

Piaget's (1936) theory of cognitive development explains how a child constructs a mental model of the world. His contributions include a stage theory of child cognitive development, detailed observational studies of cognition in children, and a series of tests to reveal different cognitive abilities.



"The infant's hand hits a hanging toy. The infant sees it bob about, then repeats the gesture several times, later applying it to other objects as well, developing a striking schema for striking."

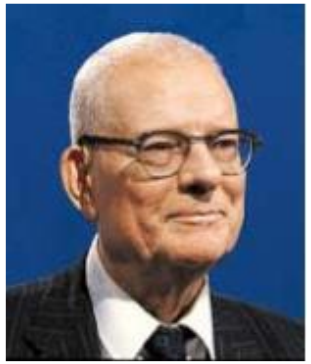
The notion of causality in the infant's model entails a primitive cause-effect relationship between actions and results. For example if Z = 'pull string hanging from bassinet hood' Y = 'toy shakes', the infant's model contains the causal relationship $Z \rightarrow Y$.

W. Edwards Deming (1900-1993)

"Tests of variables that affect a process are useful only if they **predict what will happen if this or that variable is increased or decreased**.

Statistical theory, as taught in the books, is valid and leads to operationally verifiable tests and criteria for an **enumerative study**. Not so with an **analytic problem**, as the conditions of the experiment will not be duplicated in the next trial.

Unfortunately, most problems in industry are analytic."*



*From preface to *The Economic Control of Quality of nufactured product* by W. Shewhart, 1931

Jerzy Neyman (1894-1981)

1990, Vol. 5, No. 4, 465–480

On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.

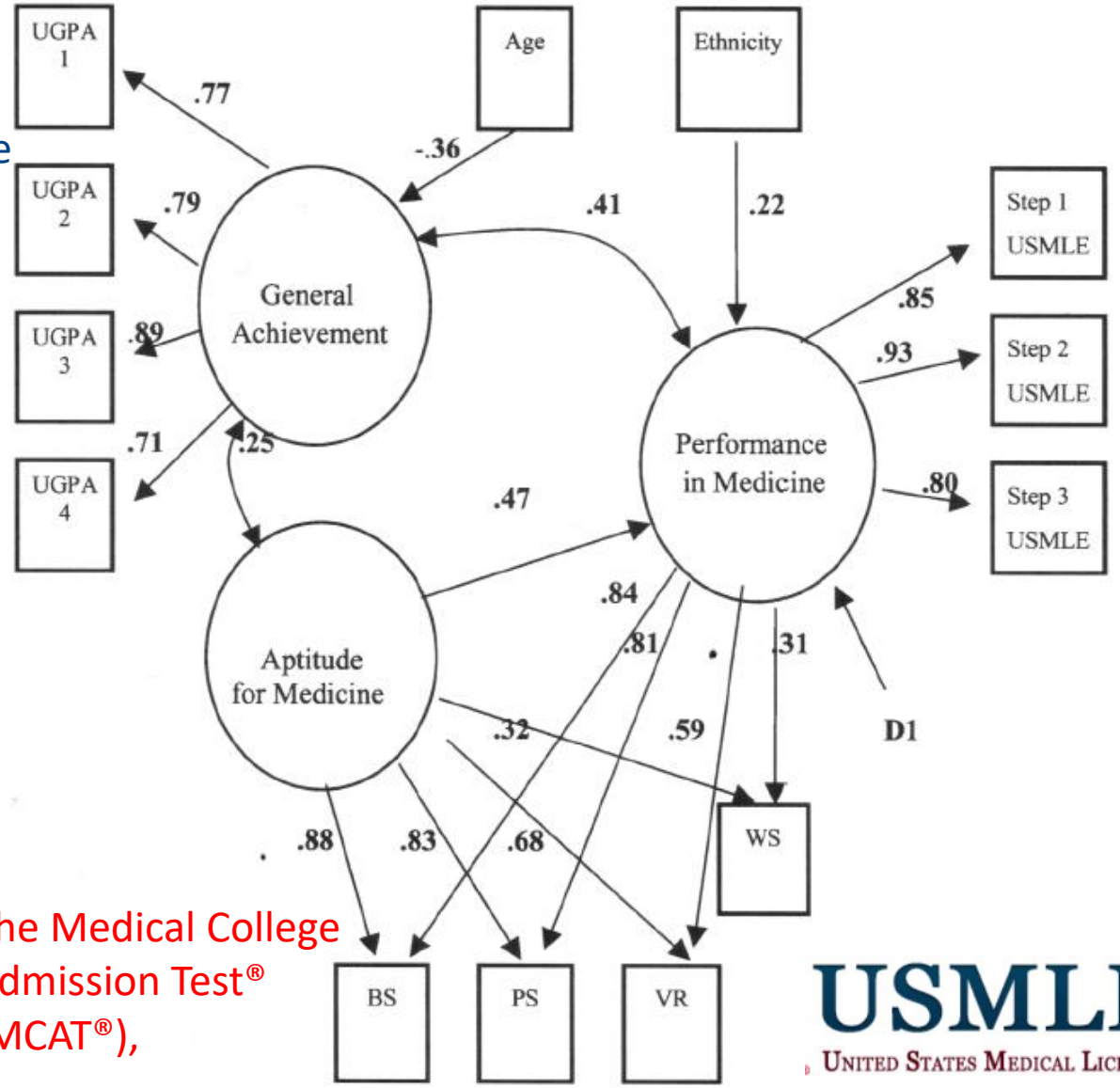
Jerzy Splawa-Neyman

Translated and edited by D. M. Dabrowska and T. P. Speed from the Polish original, which appeared in Roczniki Nauk Rolniczych Tom X (1923) 1–51 (Annals of Agricultural Sciences)

Abstract. In the portion of the paper translated here, Neyman introduces a model for the analysis of field experiments conducted for the purpose of comparing a number of crop varieties, which makes use of a double-indexed array of unknown potential yields, one index corresponding to varieties and the other to plots. The yield corresponding to only one variety will be observed on any given plot, but through an urn model embodying sampling without replacement from this doubly indexed array, Neyman obtains a formula for the variance of the difference between the averages of the observed yields of two varieties. This variance involves the variance over all plots of the potential yields and the correlation coefficient r between the potential yields of the two varieties on the same plot. Since it is impossible to estimate r directly, Neyman advises taking $r = 1$, observing that in practice this may lead to using too large an estimated standard deviation, when comparing two variety means.

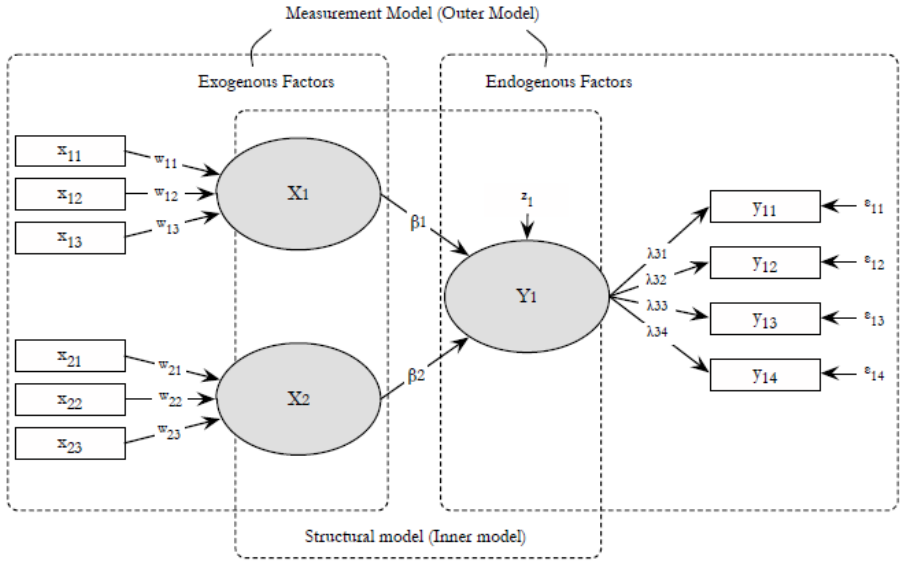
Potential outcomes

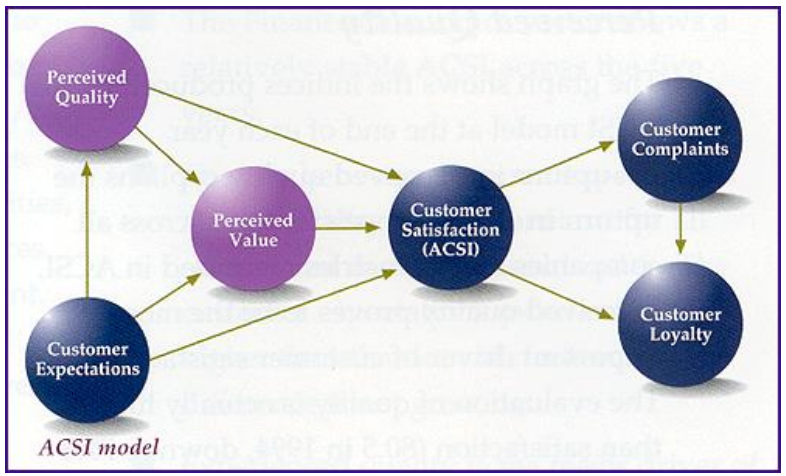
Undergraduate Grade Point Average (UGPA)



The Medical College Admission Test® (MCAT®),

USMLE®
UNITED STATES MEDICAL LICENSING EXAMINATION®





Structural Equation Models (SEM)

Figure 2. Latent variable path analysis model of UGPA, MCAT, and USMLE (Steps 1–3) latent variables employing ML estimation (n = 24,872). *Note*. Fit indexes: $\chi^2(55) = 11726.28$, $p < .001$ (CFI = .928, RMSEA = .025). UGPA1-4 = Undergraduate GPA Year 1–4; BS = Biological Sciences MCAT Subtest; PS = Physical Sciences MCAT Subtest; VR = Verbal Reasoning MCAT Subtest; WS = Writing Sample MCAT Subtest; Step 1–3 USMLE = United States Medical Licensing Exam Step 1–3.

Contingency tables

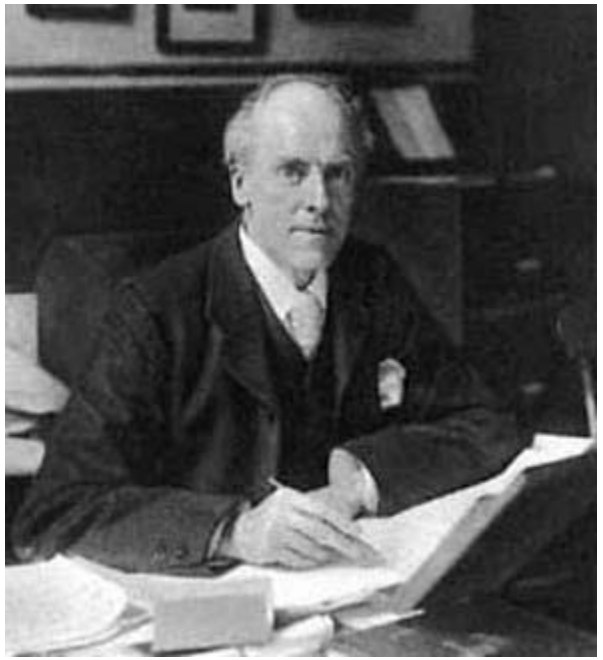
DEPARTMENT OF APPLIED MATHEMATICS,
UNIVERSITY COLLEGE, UNIVERSITY OF LONDON.

DRAPERS' COMPANY RESEARCH MEMOIRS.
BIOMETRIC SERIES, I.

MATHEMATICAL CONTRIBUTIONS TO THE THEORY OF EVOLUTION.

XIII. ON THE THEORY OF CONTINGENCY AND ITS RELATION TO ASSOCIATION AND NORMAL CORRELATION.

BY
KARL PEARSON, F.R.S.



$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}$$

Karl Pearson
1857-1936

The term **contingency table** was first used by Karl Pearson in "On the Theory of Contingency and Its Relation to Association and Normal Correlation", the Drapers' Company Research Memoirs Biometric Series I, published in 1904.

Contingency tables

(2.) *On the Conception of Contingency.*

In mathematical treatises on algebra a definition is usually given o
probability. If p be the probability of any event, and q the probabili
event, then the two events are said to be independent, if the prob
combined event be $p \times q$. Now let A be any attribute or character
classified into the groups $A_1, A_2, \ldots A_s$, and let the total number
examined be N, and let the numbers which fall into these groups be
respectively. Then the probability of an individual falling into one or
groups is given by $n_1/N, n_2/N, \ldots n_s/N$ respectively. Now supp
population to be classified by any other attribute into the groups B_1, I
the group frequencies of the N individuals to be $m_1, m_2, \ldots m_t$ resp
probability of an individual falling into these groups will be respectively $m_1/N, m_2/N,$
$m_3/N, \ldots m_t/N$. Accordingly the number of combinations of B_v with A_u to be
expected on the theory of independent probability if N pairs of attributes are
examined is

$$N \times \frac{n_u}{N} \times \frac{m_v}{N} = \frac{n_u \cdot m_v}{N} = \nu_{uv}, \text{ say.}$$

Brit. J. Phil. Sci. **34** (1983), 105–118 *Printed in Great Britain*

The Fisher/Pearson Chi-Squared Controversy: A Turning Point for Inductive Inference*

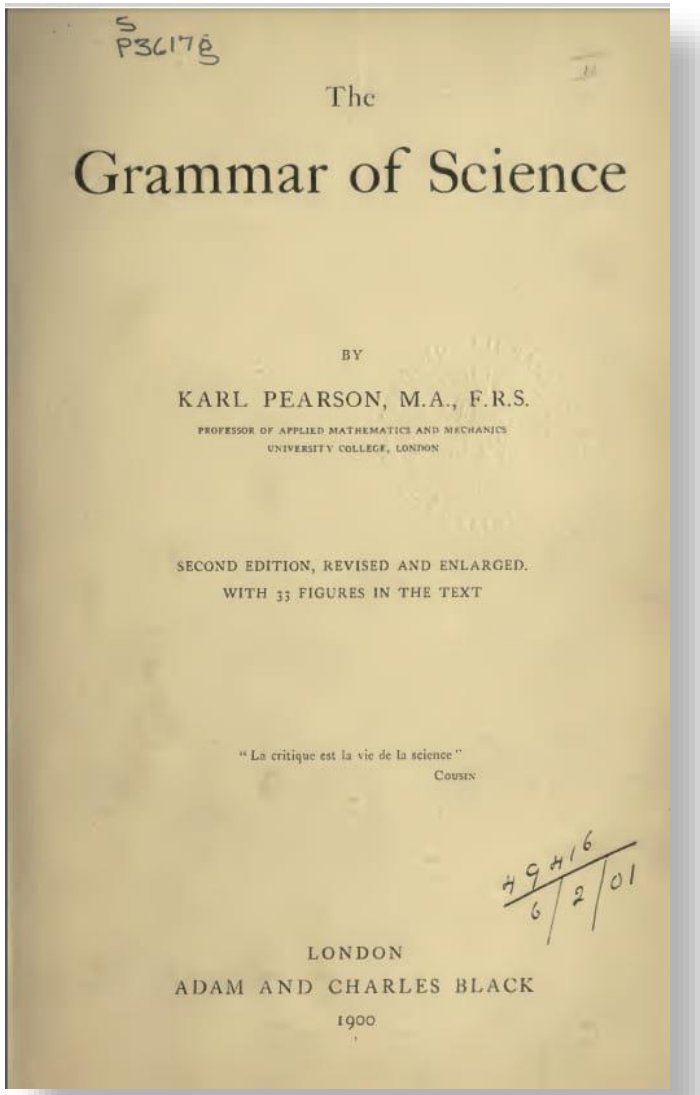
by DAVIS BAIRD

1 *The Chi-Squared Test*
2 *Yule and Greenwood's 1915 Paper*
3 *Fisher's Argument*
4 *Pearson's Reply*
5 *Assessment*
6 *Goodness of Fit and Closeness to Truth*
7 *Goodness of Fit and Information*
8 *Conclusion*

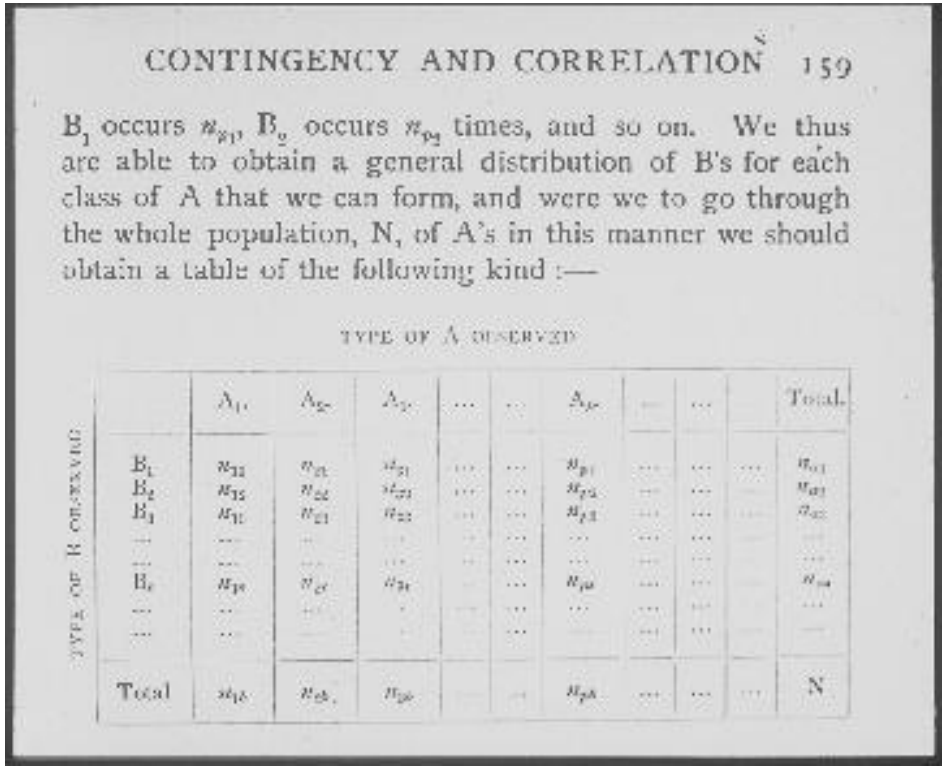
Contingency tables

Now it must be quite clear that if we make our measurement of contingency any function whatever of such quantities as $n_{uv} - \nu_{uv}$, its magnitude will be absolutely independent of the order of classification, *i.e.*, its value will be unchanged if we re-arrange the A's and the B's in any manner whatever. This is the fundamental gain of this new conception of contingency. But precisely as we can measure position or acceleration in a great variety of ways, so it is possible to measure contingency. We must try to select out of these ways those which: (a) bring contingency into line with the customary notions of correlation and association; and (b) permit of not too laborious calculations leading to the required measure.

Contingency tables



In the chapter **Contingency and correlation - the insufficiency of causation**, (The Grammar of Science, 1911), Pearson says: "Beyond such discarded fundamentals as 'matter' and 'force' lies still another fetish amidst the inscrutable arcana of modern science, namely, the category of cause and effect."



CONTINGENCY AND CORRELATION 159

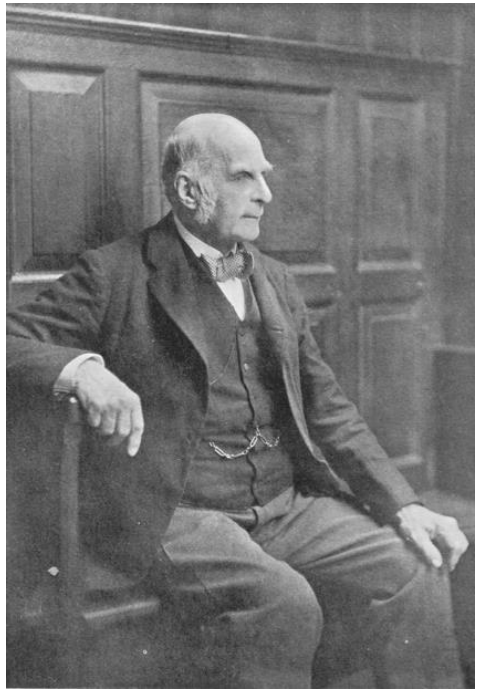
B_1 occurs n_{p1}, B_2 occurs n_{p2} times, and so on. We thus are able to obtain a general distribution of B's for each class of A that we can form, and were we to go through the whole population, N, of A's in this manner we should obtain a table of the following kind :—

TYPE OF A OBSERVED

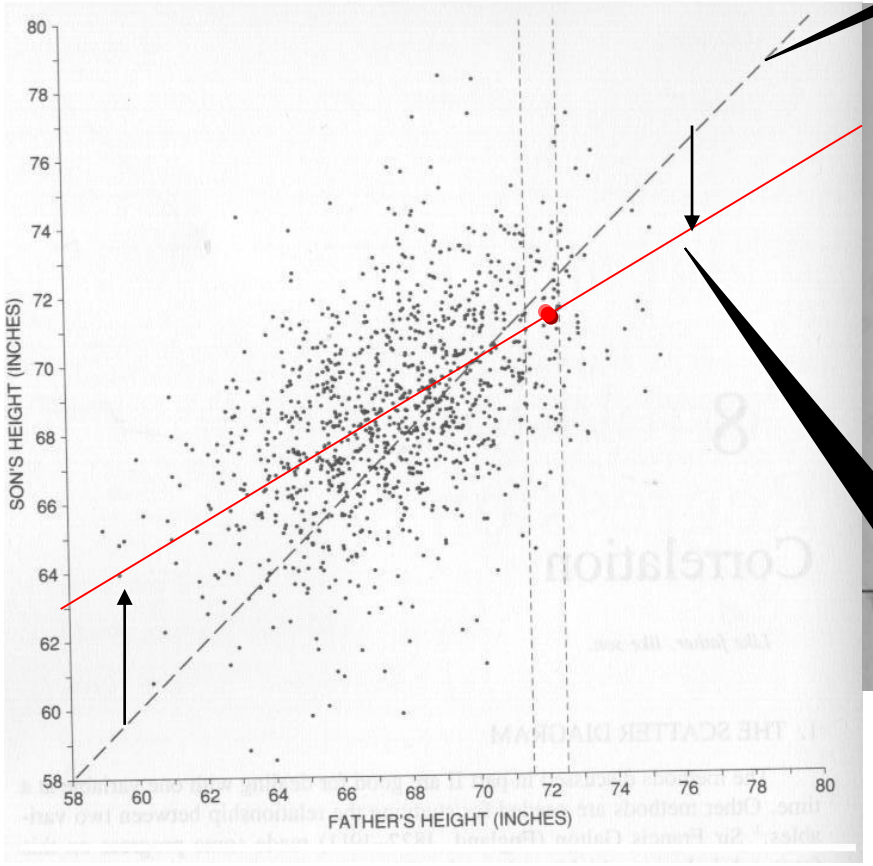
TYPE OF B OBSERVED	A_1	A_2	A_3	...	...	A_p	...	...		Total.
B_1	n_{11}	n_{21}	n_{31}	...	...	n_{p1}	...	...	...	n_{01}
B_2	n_{12}	n_{22}	n_{32}	...	...	n_{p2}	...	...	...	n_{02}
B_3	n_{13}	n_{23}	n_{33}	...	...	n_{p3}	...	...	...	n_{03}
...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...
B_q	n_{1q}	n_{2q}	n_{3q}	...	...	n_{pq}	...	...	...	n_{0q}
...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...
Total	n_{01}	n_{02}	n_{03}	...	...	n_{p0}	...	...	...	N

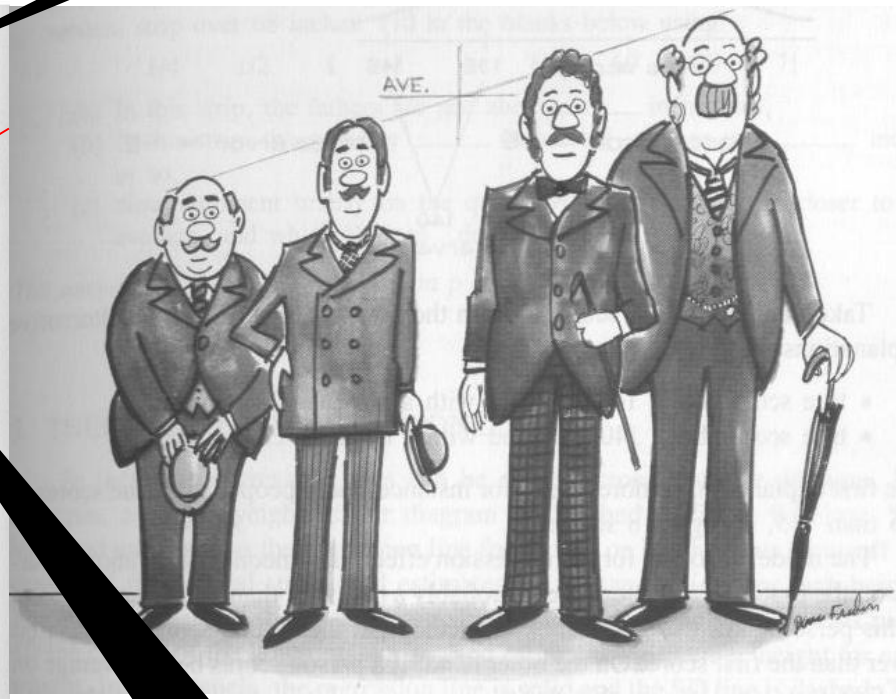
https://pure.mpg.de/…/item_2…/component/file_2368441/content

Regression towards the mean….



Sir Francis Galton (1822-1911)



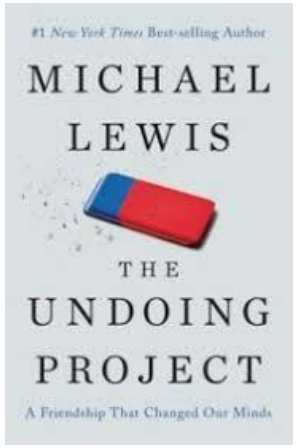


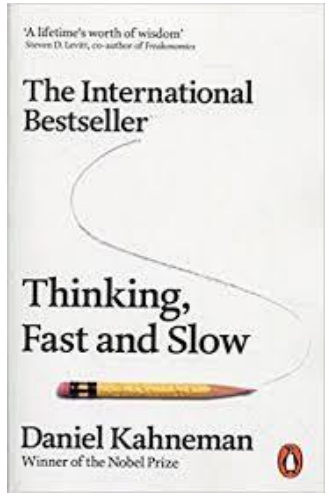
"It is easy to see that consequence of the co-relation must be the variation of the two organs being partly due to common causes"

Galton, F. (1886). "Regression towards mediocrity in hereditary stature". *The Journal of the Anthropological Institute of Great Britain and Ireland* 15: 246–263

Regression towards the mean….

1. Base rate neglect,
2. Overconfidence,
3. Anchoring,
4. Representativeness,
5. Availability,
6. **Regression towards the mean,**
7. Spurious correlation,
8. Framing.





Treatment to reduce high levels of a measurement

People with extreme values of the measurement, such as high blood pressure, may be treated to bring their values closer to the mean. If they are measured again we will observe that the mean of the extreme group is now closer to the mean of the whole population, that is, it is reduced. This should not be interpreted as showing the effect of the treatment.

Relating change to initial value

We may study the relation between the initial value of a measurement and the change in that quantity over time. In antihypertensive drug trials, for example, it may be postulated that the drug's effectiveness would be different (usually greater) for patients with more severe hypertension. This is a reasonable question, but, the regression towards the mean will be greater for the patients with the highest initial blood pressures, so that we would expect to observe the postulated effect even in untreated patients.

Regression towards the mean….

1. Base rate neglect,
2. Overconfidence,
3. Anchoring,
4. Representativeness,
5. Availability,
6. **Regression towards the mean,**
7. Spurious correlation,
8. Framing.

Comparison of two methods of measurement

When comparing two methods of measuring the same quantity researchers are sometimes tempted to regress one method on the other. The fallacious argument is that if the methods agree the slope should be 1. Because of the effect of regression towards the mean we expect the slope to be less than 1, even if the two methods agree closely.

https://www.ncbi.nlm.nih.gov/pubmed/16921578 .

Stephen Senn (2006), Change from baseline and analysis of covariance revisited, Stat Med.; 25(24):4334-44

Representativeness….

1. Base rate neglect,
2. Overconfidence,
3. Anchoring,
4. **Representativeness,**
5. Availability,
6. Regression towards the mean,
7. Spurious correlation,
8. Framing.

The hot hand fallacy

The "Hot Hand": Statistical Reality or Cognitive Illusion?

Amos Tversky and Thomas Gilovich

The Hot Hand in Basketball: On the Misperception of Random Sequences

THOMAS GILOVICH

Cornell University

AND

ROBERT VALLONE AND AMOS TVERSKY

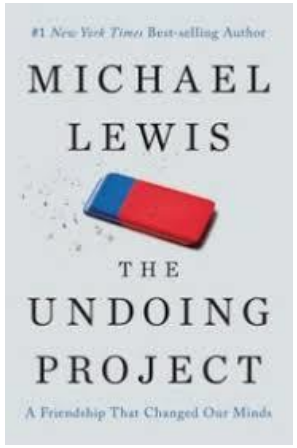
Stanford University

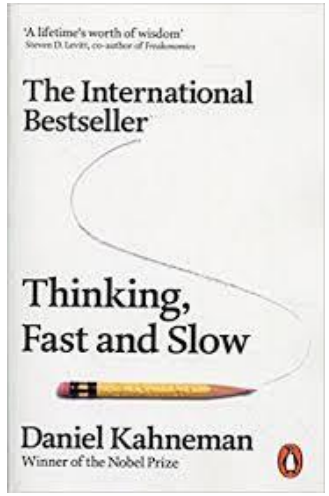
We investigate the origin and the validity of common beliefs regarding "the hot hand" and "streak shooting" in the game of basketball. Basketball players and fans alike tend to believe that a player's chance of hitting a shot are greater following a hit than following a miss on the previous shot. However, detailed analyses of the shooting records of the Philadelphia 76ers provided no evidence for a positive correlation between the outcomes of successive shots. The same conclusions emerged from free-throw records of the Boston Celtics, and from a controlled shooting experiment with the men and women of Cornell's varsity teams. The outcomes of previous shots influenced Cornell players' predictions but not their performance. The belief in the hot hand and the "detection" of streaks in random sequences is attributed to a general misconception of chance according to which even short random sequences are thought to be highly representative of their generating process.

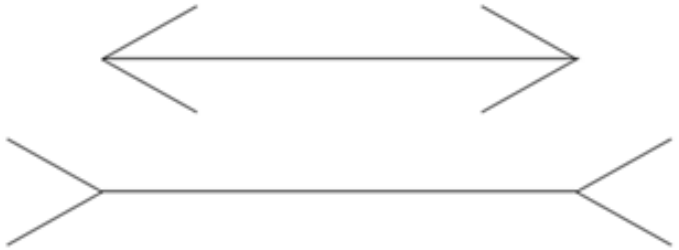
- 91% of the fans believe that a player has a better chance of making a shot after having just made his last two or three shots than he does after having just missed his last two or three shots
- 84% of the fans believe that it is important to pass the ball to someone who has just made several (two, three, or four) shots in a row

Framing….

1. Base rate neglect,
2. Overconfidence,
3. Anchoring,
4. Representativeness,
5. Availability,
6. Regression towards the mean,
7. Spurious correlation,
8. **Framing.**



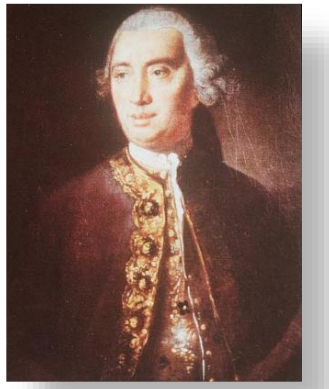




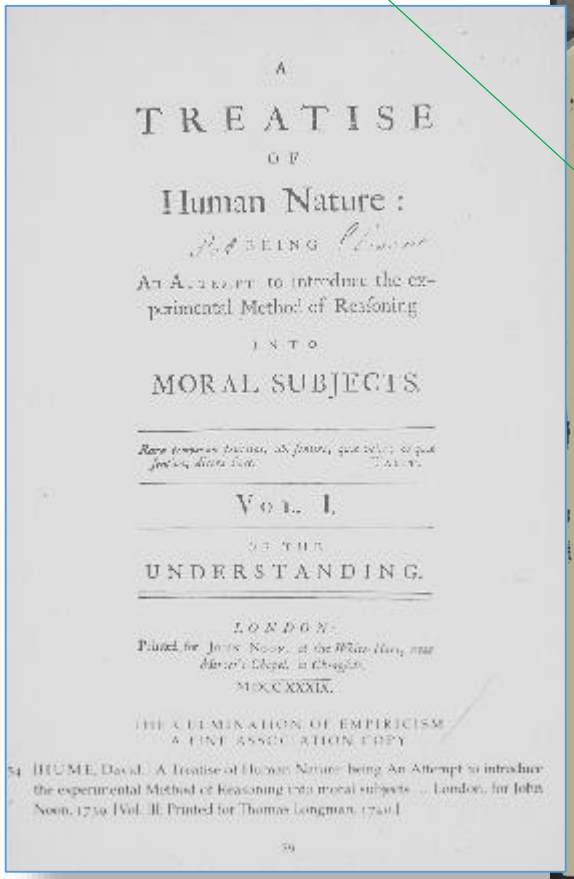
Muller-Lyer optical illusion

David Hume (1711-1776)

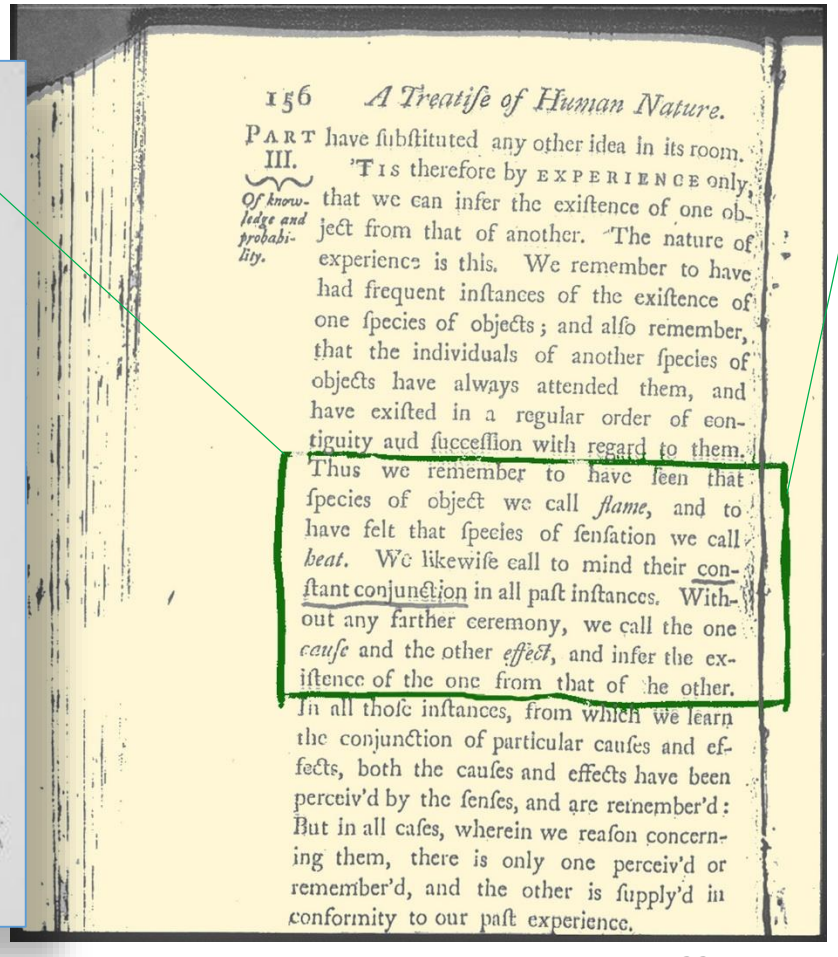
1. Analytical vs. empirical claims, *the former are product of thoughts, the latter matter of fact*
2. Causal claims are empirical
3. All empirical claims originate from experience.



"Thus we remember to have seen that species of object we call *flame*, and to have felt that species of sensation we call *heat*. We likewise call to mind their constant conjunction in all past instances. Without any farther ceremony, we call the one *cause* and the other *effect*, and infer the existence of the one from that of the other."



A
TREATISE
OF
Human Nature:
BEING
An Attempt to introduce the experimental Method of Reasoning
INTO
MORAL SUBJECTS.
VOL. I.
OF THE
UNDERSTANDING.
LONDON:
MDCCXXXIX.

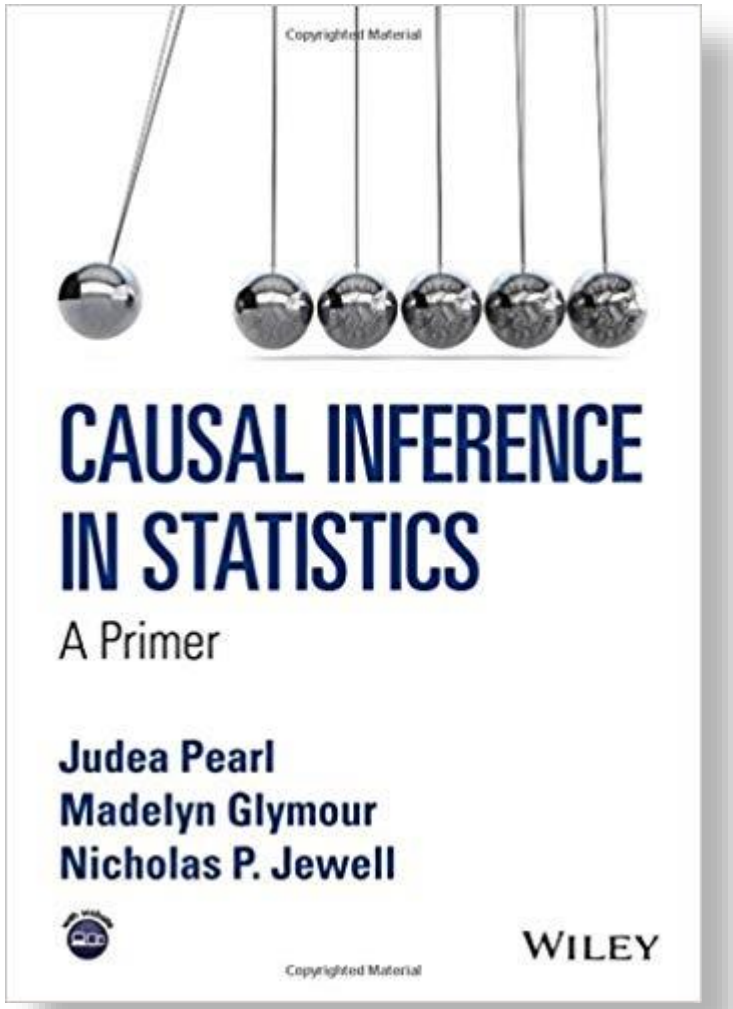


156 *A Treatise of Human Nature.*

PART III. *Of knowledge and probability.*

have ſubſtituted any other idea in its room. 'TIS therefore by EXPERIENCE only, that we can infer the exiſtence of one object from that of another. The nature of experience is this. We remember to have had frequent inſtances of the exiſtence of one ſpecies of objects; and alſo remember, that the individuals of another ſpecies of objects have always attended them, and have exiſted in a regular order of contiguity and ſucceſſion with regard to them. Thus we remember to have ſeen that ſpecies of object we call *flame*, and to have felt that ſpecies of ſenſation we call *heat*. We likewiſe call to mind their conſtant conjunction in all paſt inſtances. Without any farther ceremony, we call the one *cauſe* and the other *effect*, and infer the exiſtence of the one from that of the other. In all thoſe inſtances, from which we learn the conjunction of particular cauſes and effects, both the cauſes and effects have been perceiv'd by the ſenſes, and are remember'd: But in all caſes, wherein we reaſon concerning them, there is only one perceiv'd or remember'd, and the other is ſupply'd in conformity to our paſt experience.

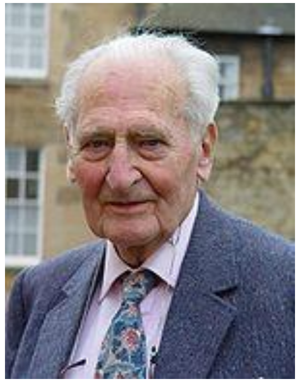
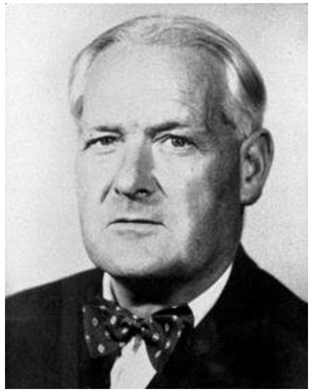
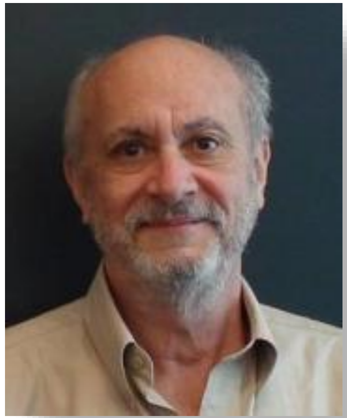
A journey back into the past



https://en.wikipedia.org/wiki/Newton%27s_cradle

2020
2006
1977
1953
1936
1931
1923
1921
1904
1886
1738

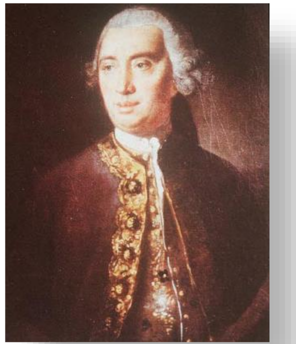
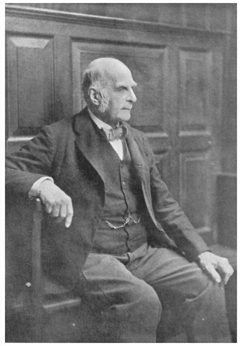


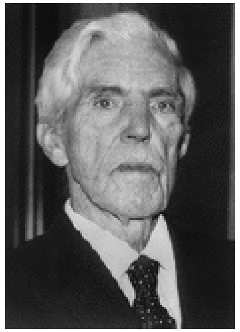




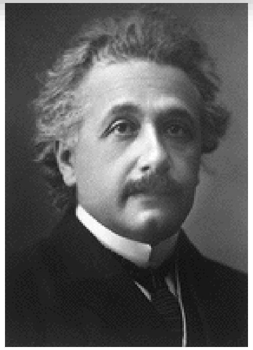
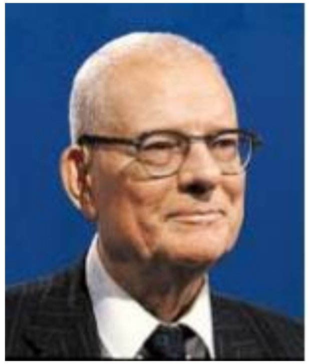








Agenda

1. Background on causality in science and statistics
2. **Fishbone cause and effect diagrams**
3. Bayesian networks
4. Randomization in experimental designs
5. Propensity scores in observational studies
6. Counterfactuals and do calculus
7. Personalized medicine, condition based maintenance and Industry 4.0
8. Future research areas

Wiley StatsRef:
Statistics Reference Online

Cause-and-Effect Diagrams

By *Ron S. Kenett*[1,2]

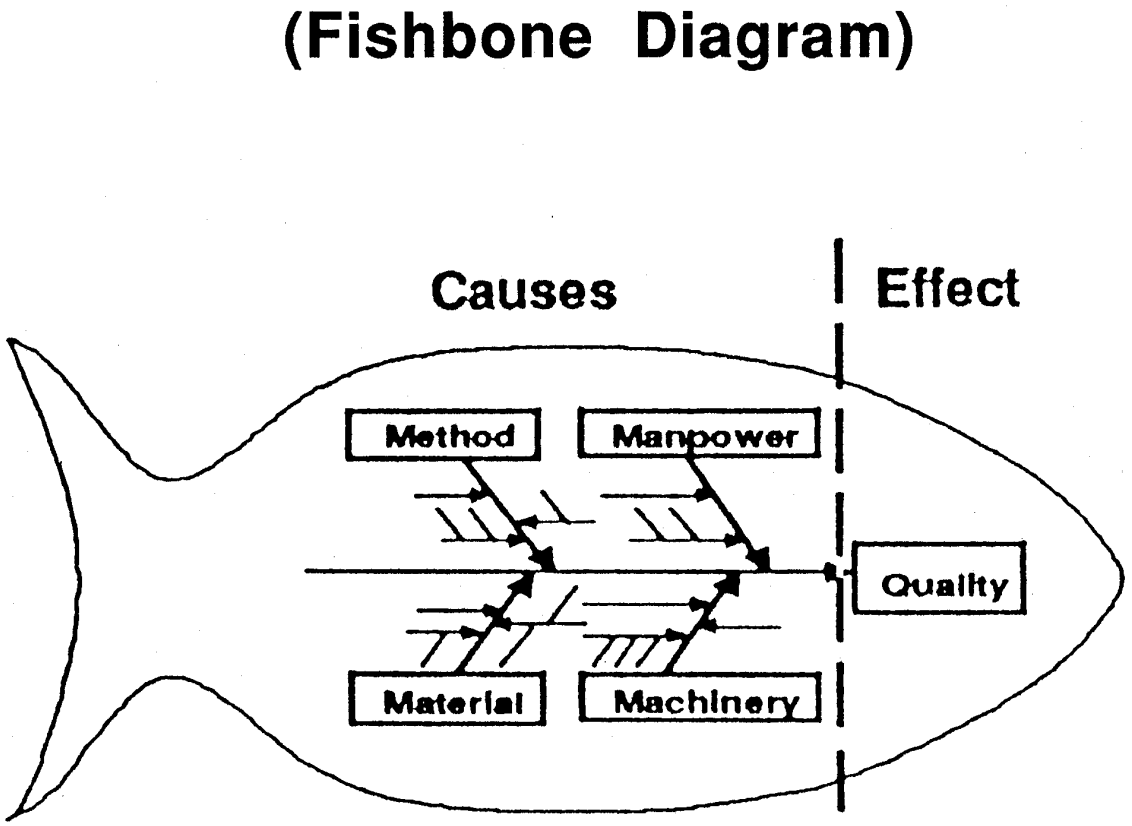
Keywords: *scatter plots, Ishikawa diagrams, structural equation models, Bayesian networks, integrated management models*

Abstract: Cause and effect is a basic knowledge driven by theoretical and empirical considerations. Several tools have been proposed to map cause and effect relationships, with some more heuristics some highly quantitative. In this section we cover the Ishikawa fishbone diagram, structural equation models, and Bayesian networks.

Cause-Effect Diagram

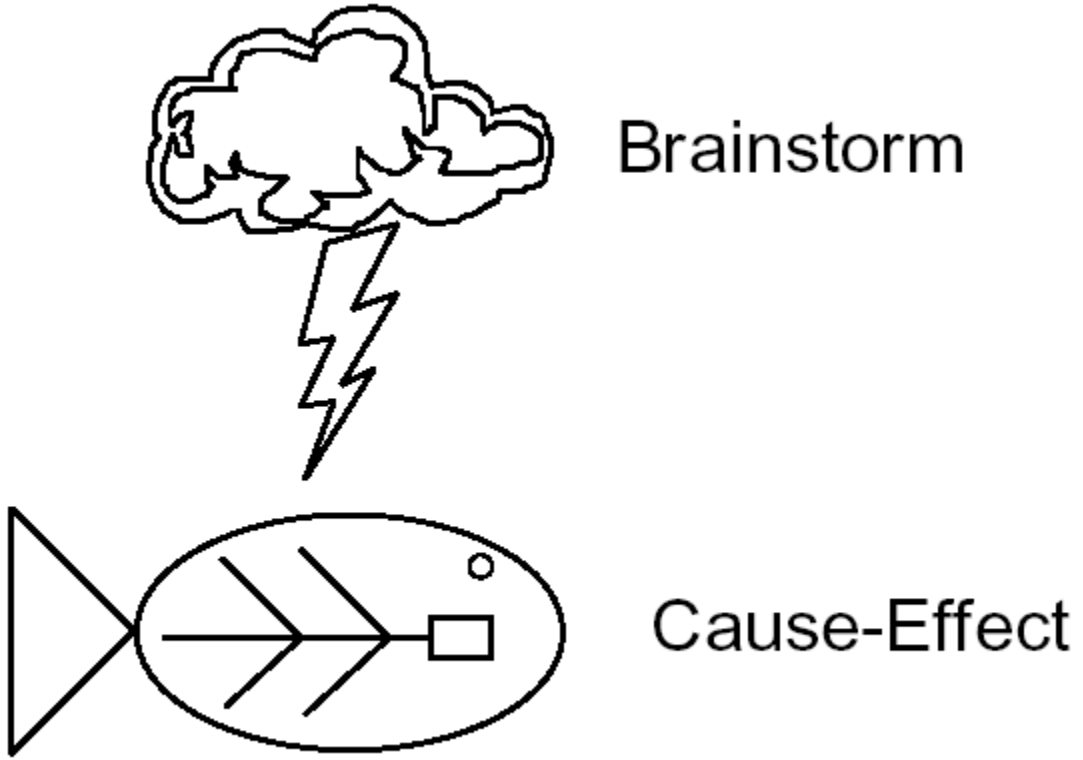
- Objectives: Visual presentation of relationships between **Effect** and possible **Causes**
- How?: **List** of possible Causes and their **Structure** (**Fishbone**)
- **Individual** and **Teamwork** tool for improvement program initiation
- Possibility to select critical Causes based on **Expert Knowledge**

Cause-Effect (Ishikawa) Diagram



Kaoru Ishikawa
1915 - 1989

Cause-Effect Diagram Methodology

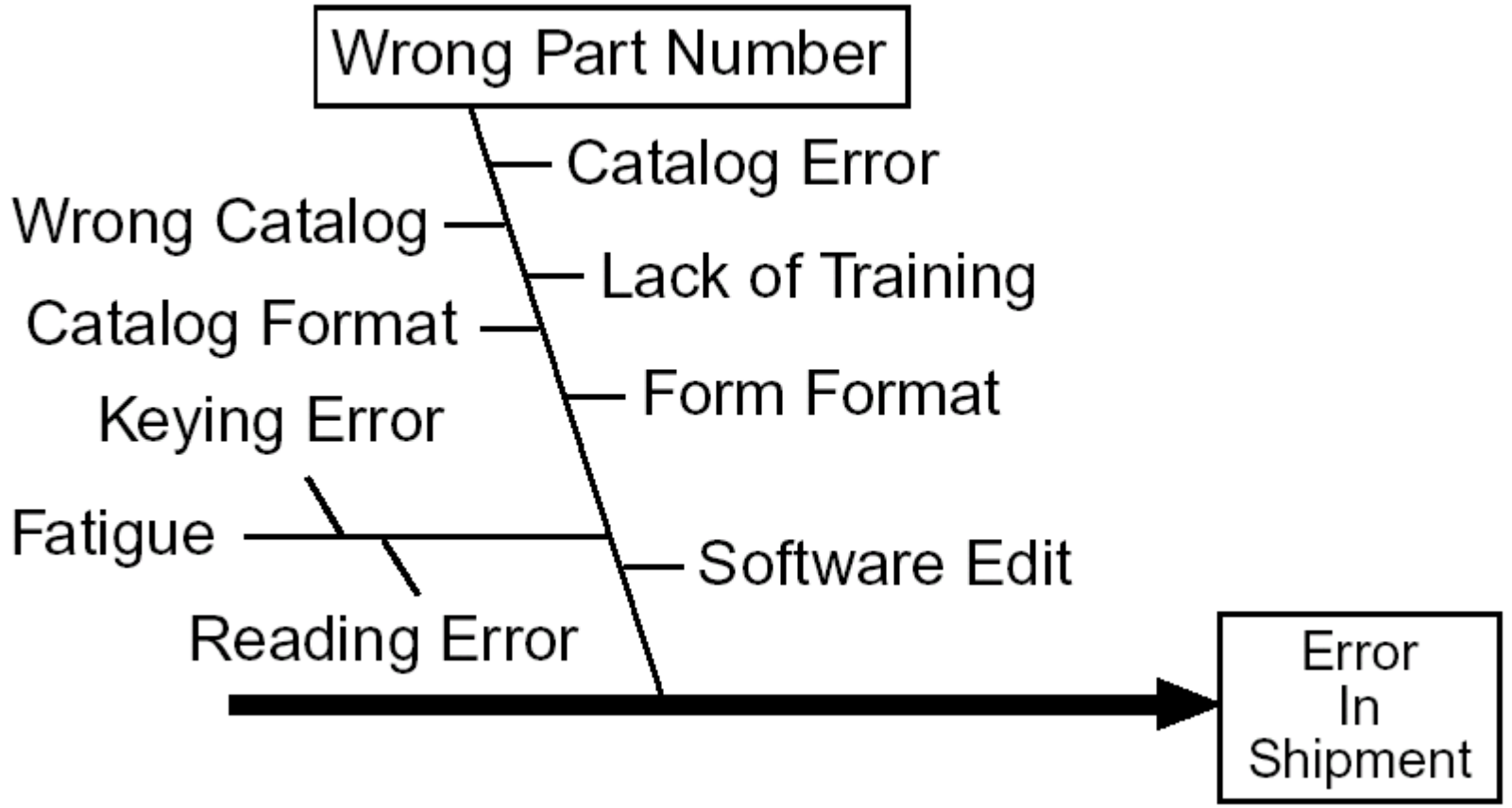


29

Round robin process

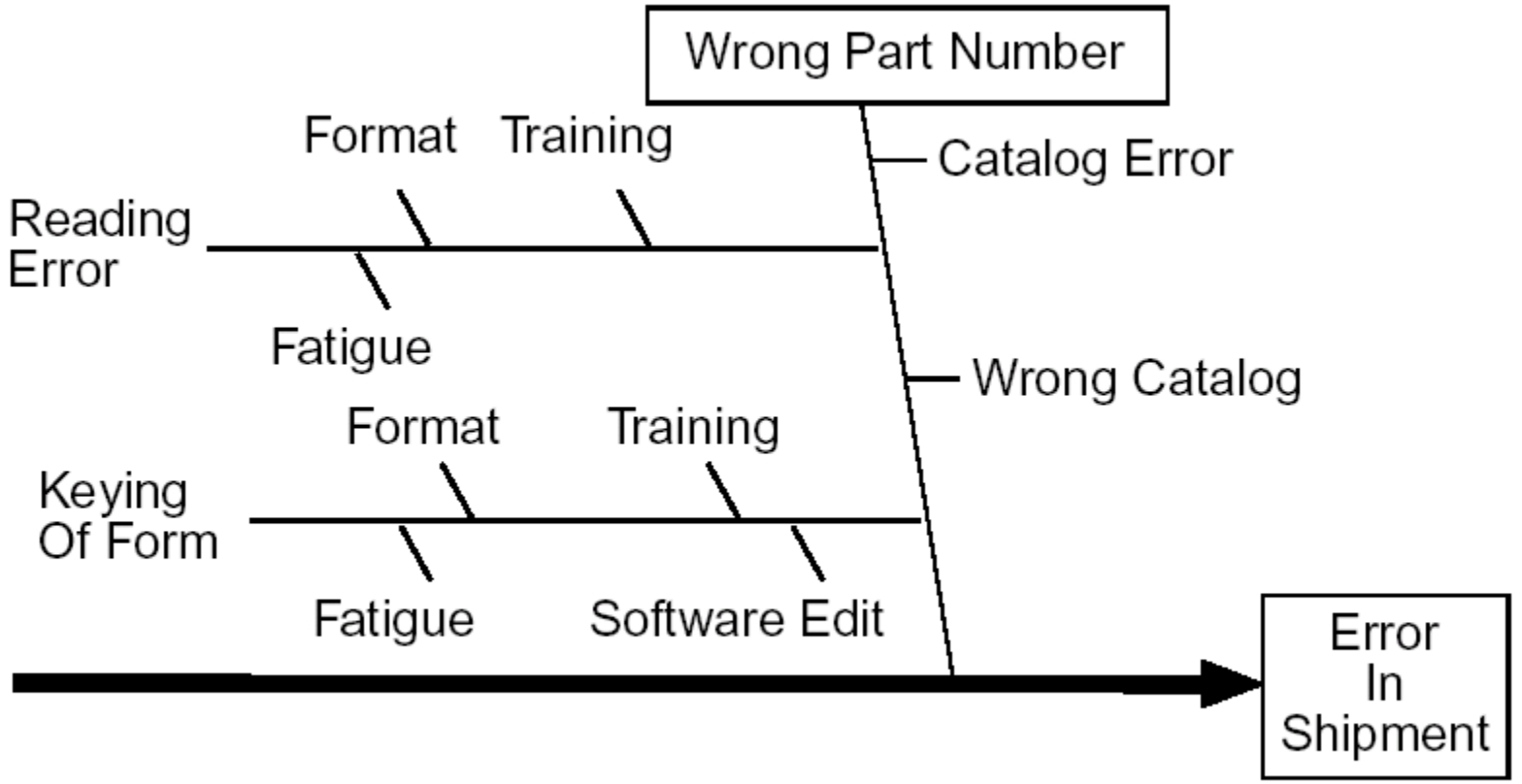
1. You can say “pass”
2. You can build on other’s ideas
3. No critique allowed (even self)
4. Indicate where to note the idea on the fishbone diagram

Why?


Wrong Part Number
Catalog Error
Wrong Catalog
Lack of Training
Catalog Format
Form Format
Keying Error
Fatigue
Software Edit
Reading Error
Error
In
Shipment

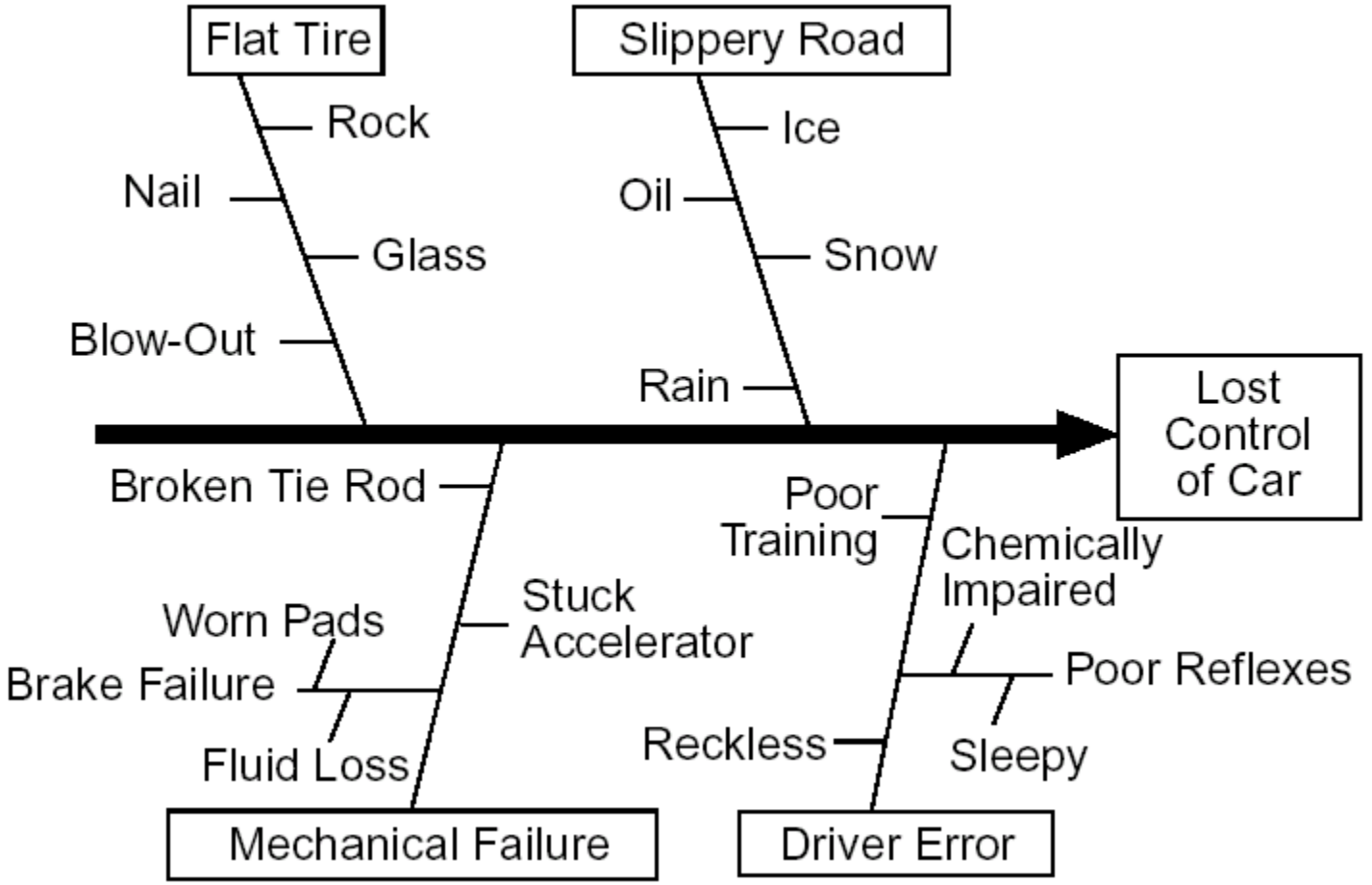
31

Why? Why? Why?



32

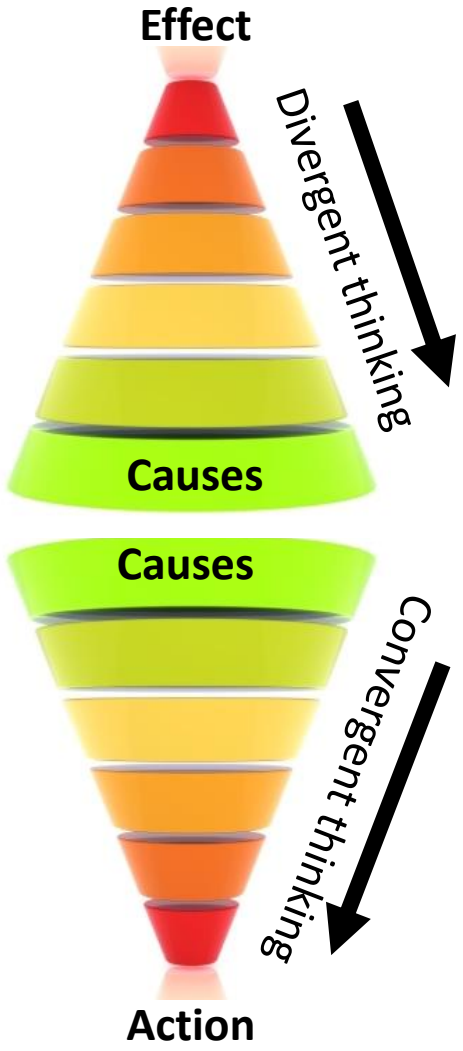
Lost control of a car



33

9 participants, 2 votes each to prioritize impact, cost and feasability

Lost control of a car – improvement priorities



Flat Tire

Nail

Rock

Blow-out

Glass

7

Slippery Road

Oil

Rain

Ice

Snow

2

Mechanical Failure

Brake failure

Broken tie rod

Stuck accelerator

5

Driver error

Reckless

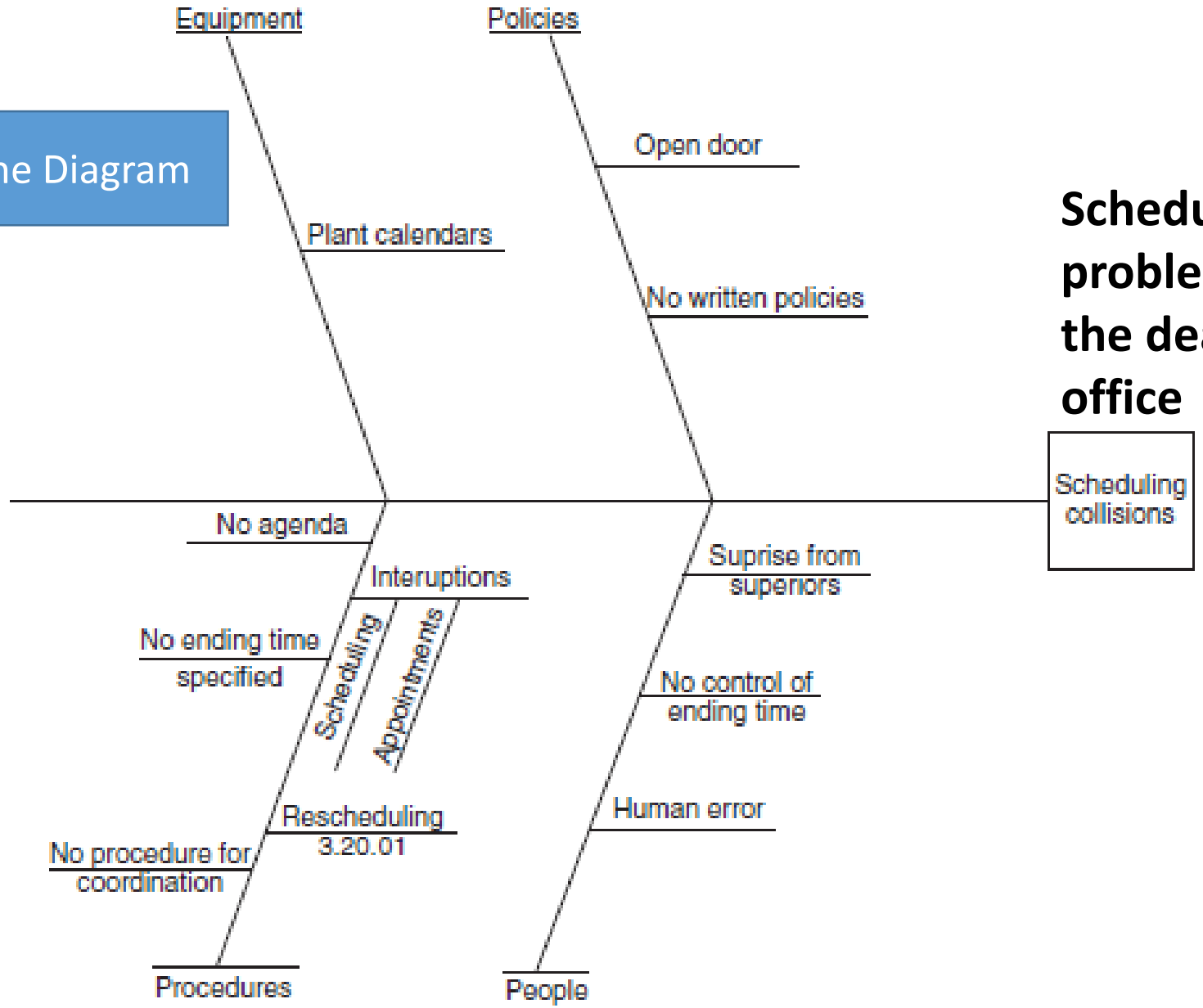
Poor training

Poor reflexes

4

To minimize the effect we will focus on the causes in green list

Fishbone Diagram



Scheduling problems at the dean's office

JURAN IMPRO CONFERENCE – ATLANTA, 1991

UALITY MANAGEMENT ALSO APPLIES TO A SCHOOL OF MANAGEMENT

Thomas F. Kelly
Dean and Professor

Ron Kenett
Professor

Elizabeth Newton
Secretary

Gary M. Roodman
Associate Dean and Professor

Angela Wowk
Secretary

School of Management
State University of New York at Binghamton
Binghamton, NY 13902-6000

ABSTRACT

lescribes a quality improvement project undertaken by the Deans' Office
l of Management at SUNY-Binghamton. The focus of the project was the
of the Dean's daily appointments. As the School has grown over the last
rs and demand for the Dean's time has increased, scheduling his
s has become a more and more complex task. The goal of this project was
s to put more predictability into the Dean's schedule and insure that his
ies actually unfold as planned. The project has brought improvements to
Office and has sparked TQM activities in other parts of the School.

Force Field Diagram

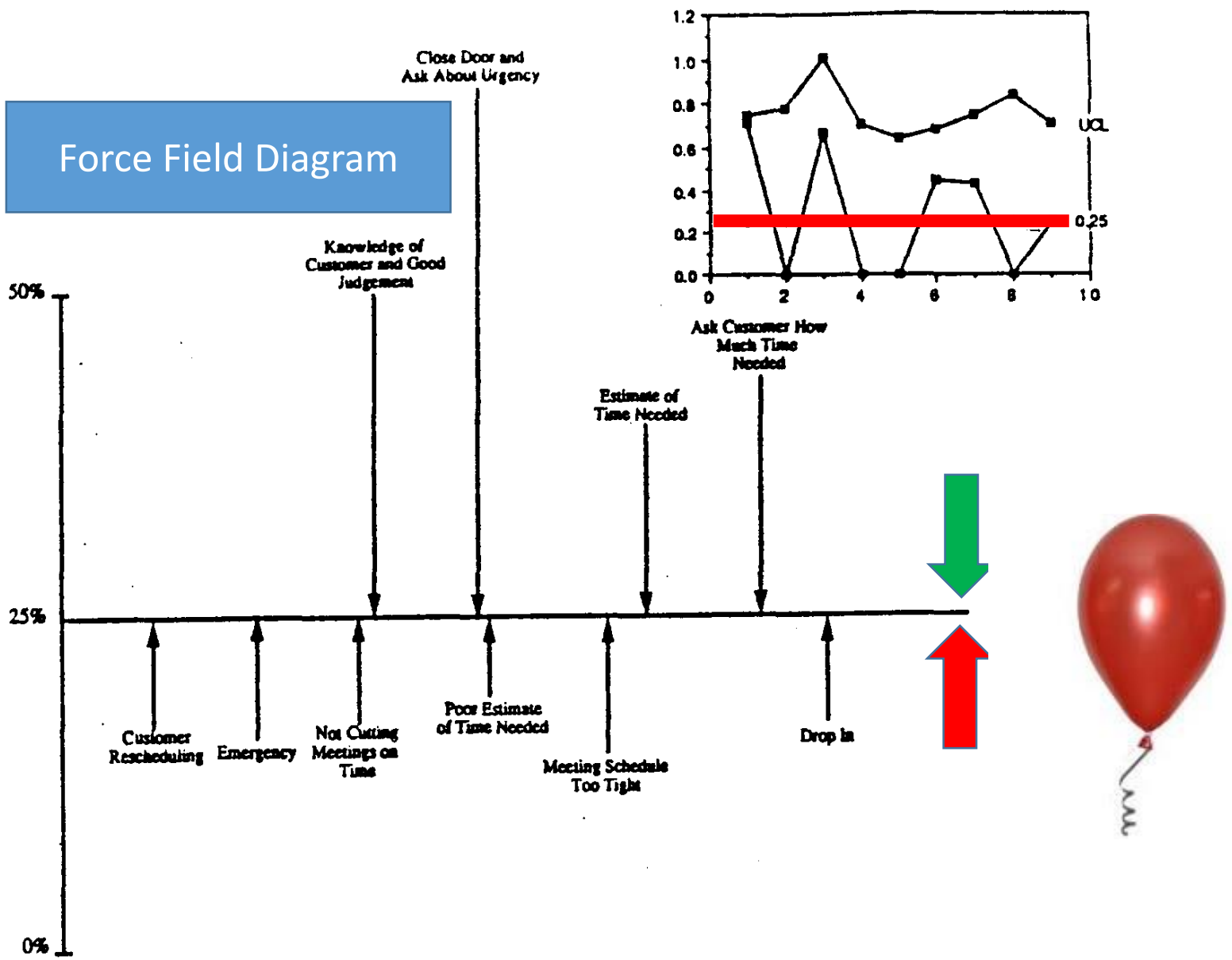


Figure 3

Flow Chart

Scheduling meetings at the dean's office

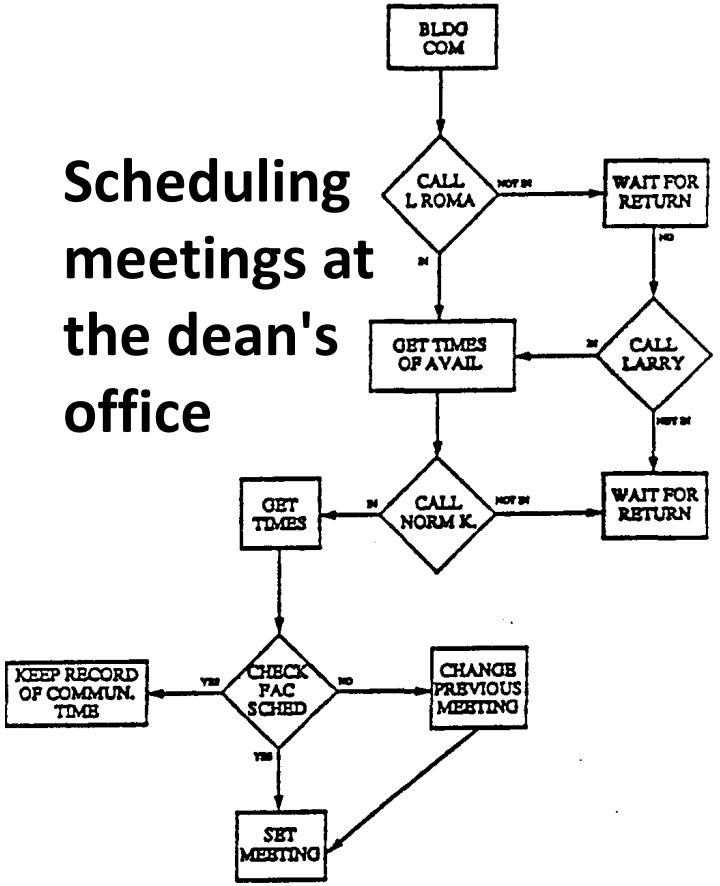
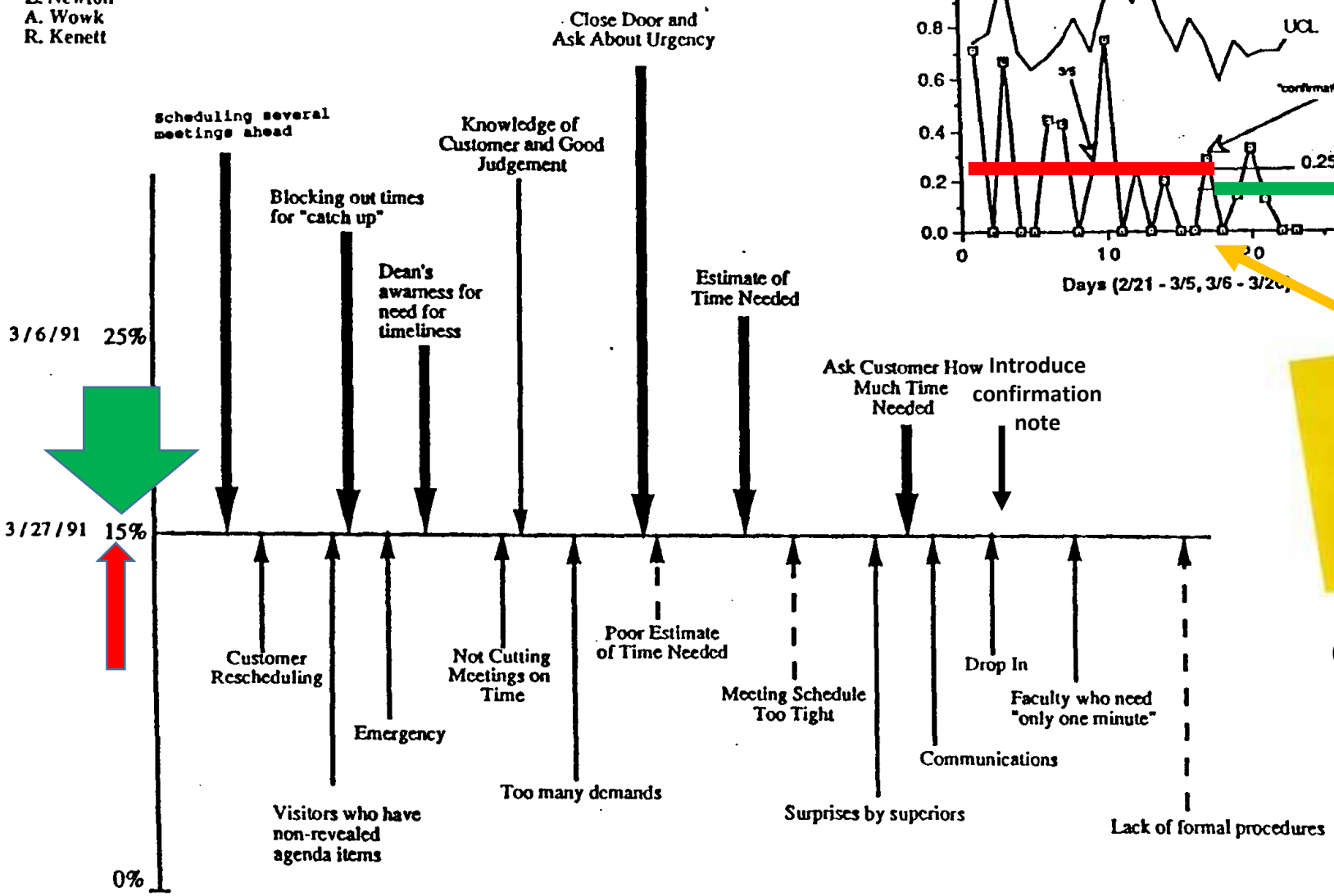


Figure 2

3/6/91, updated on 3/27/91
to indicate changes in forces
T. Kelly
G. Roodman
L. Newton
A. Wowk
R. Kenett



From 25% to 15% to conference calls

Cause and effect

Figure 4

Agenda

1. Background on causality in science and statistics
2. Fishbone cause and effect diagrams
3. **Bayesian networks**
4. Randomization in experimental designs
5. Propensity scores in observational studies
6. Counterfactuals and do calculus
7. Personalized medicine, condition based maintenance and Industry 4.0
8. Future research areas

Judea Pearl (1985)

R-43
April 1985

BAYESIAN NETWORKS: A MODEL OF SELF-ACTIVATED MEMORY FOR EVIDENTIAL REASONING*

Judea Pearl
Computer Science Department
University of California
Los Angeles, CA 90024
(judea@UCLA-locus)
(213) 825-3243

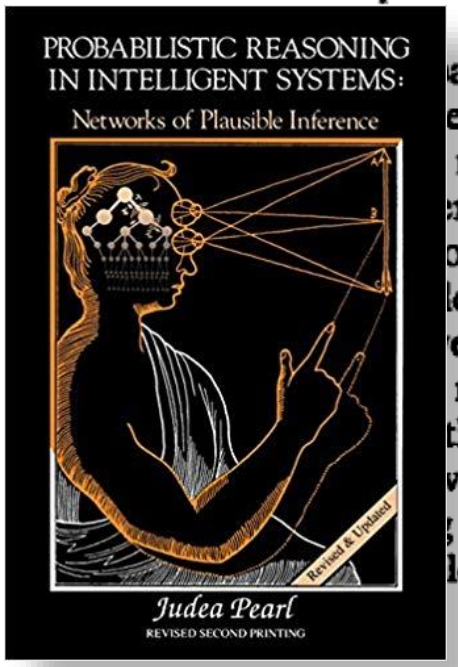
Topics: Memory Models
Belief Systems
Inference Mechanisms
Knowledge Representation

Submitted to the
Seventh Annual Conference of
the Cognitive Science Society
15-17 August 1985

Bayesian networks are directed acyclic graphs in which the nodes represent propositions (or variables), the arcs signify the existence of direct causal dependencies between the linked propositions, and the strengths of these dependencies are quantified by conditional probabilities. A network of this sort can be used to represent the deep causal knowledge of an agent or a domain expert and turns into a computational architecture if the links are used not merely for storing factual knowledge but also for directing and activating the data flow in the computations which manipulate this knowledge.

The first part of the paper defines the properties of Bayes networks which are necessary to guarantee completeness and consistency, and shows how dependencies and conditional-independence relationships can be tested using simple link-tracing operations.

art of the paper deals with the task of fusing and propagating the



Applicability of probabilistic methods to tasks requiring automated reasoning under uncertainty…. Application areas include diagnosis, forecasting, image understanding, multi-sensor fusion, decision support systems, plan recognition, planning and control, speech recognition – in short, almost any task requiring that conclusions be drawn from uncertain clues and incomplete information.

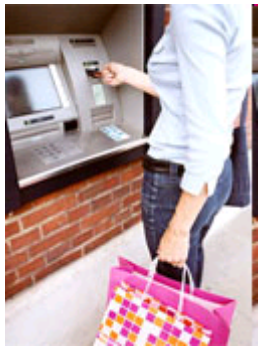
le develop causal models.

P. Spirtes, C. Glymour and R. Scheines, "Causality from Probability" Proceedings of the Conference on Advanced Computing for the Social Sciences, Williamsburg, Va. 1990.

https://www.sciencedirect.com/science/article/pii/B9780080514895500059

Monday

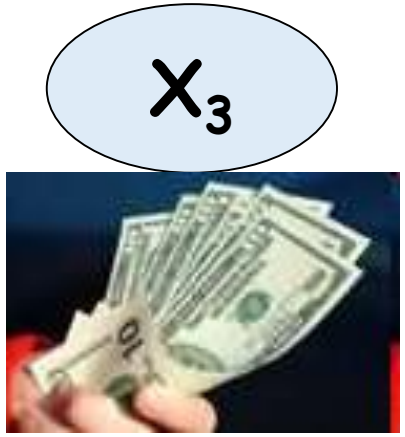
Tuesday

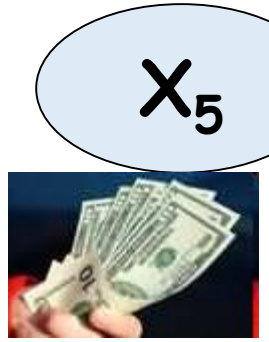


Wednesday

Thursday

Friday

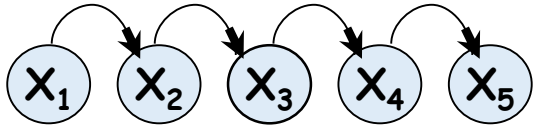




$$\mathbf{P(X_1\ X_2\ X_3\ X_4\ X_5) = ?}$$

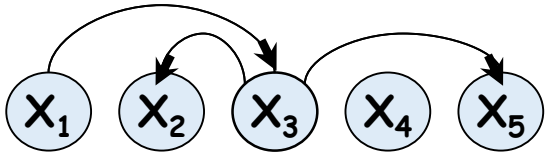
Independence

$$P(X_1 \cdots X_5) = P(X_1)P(X_2)P(X_3)P(X_4)P(X_5)$$



Markov Model

$$P(X_1 \cdots X_5) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_2)P(X_4 \mid X_3)P(X_5 \mid X_4)$$



Bayesian Network

$$P(X_1 \cdots X_5) = P(X_1)P(X_2 \mid X_3)P(X_3 \mid X_1)P(X_4)P(X_5 \mid X_3)$$

Five events

Earthquake

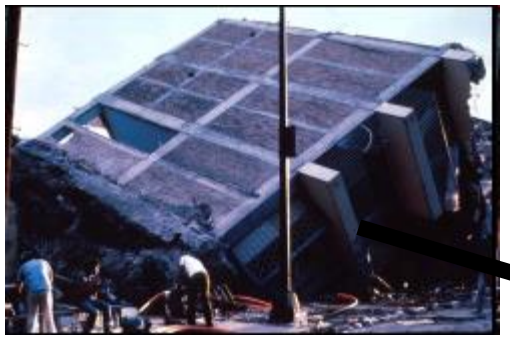
Burglary

Alarm

Radio

Call

Five events



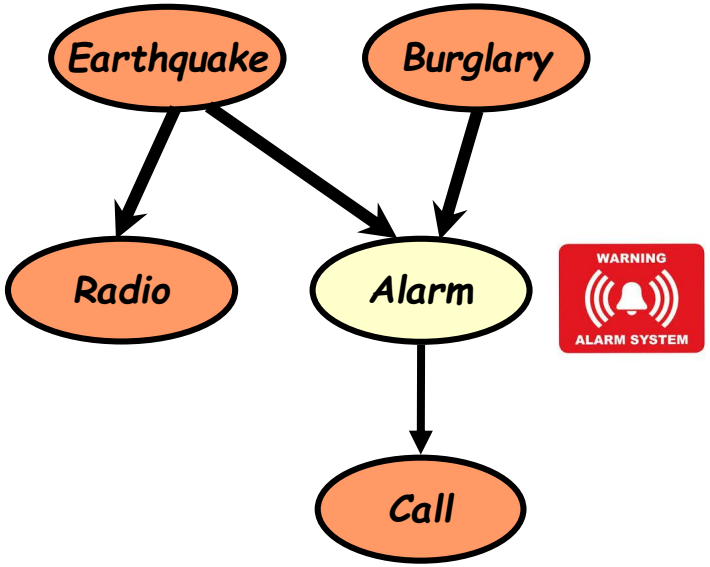
Earthquake

Burglary

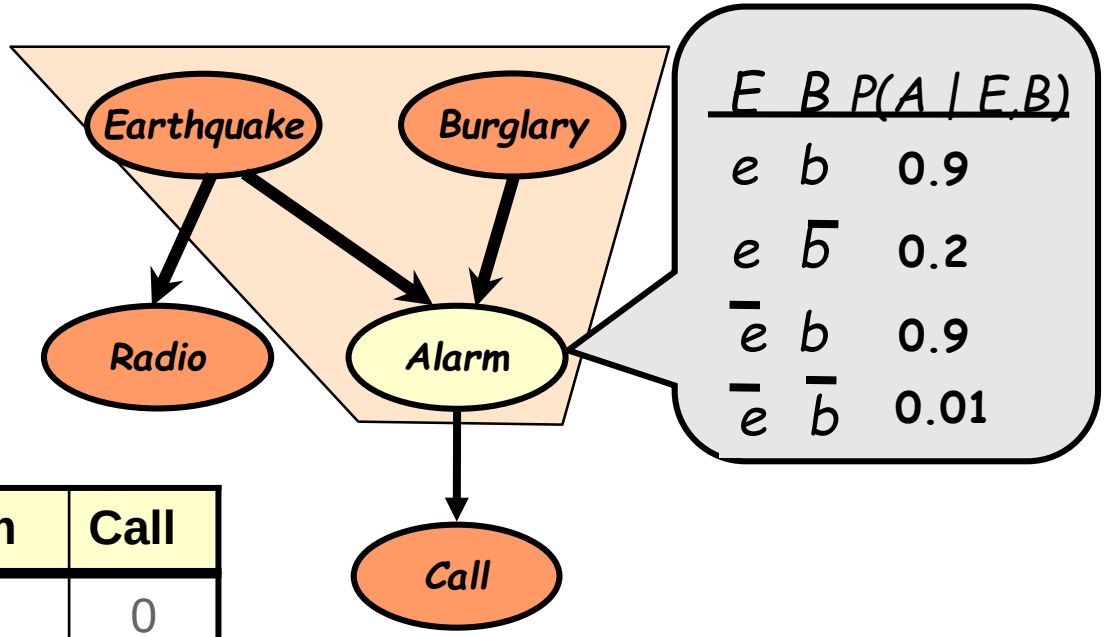
Radio

Call

Alarm

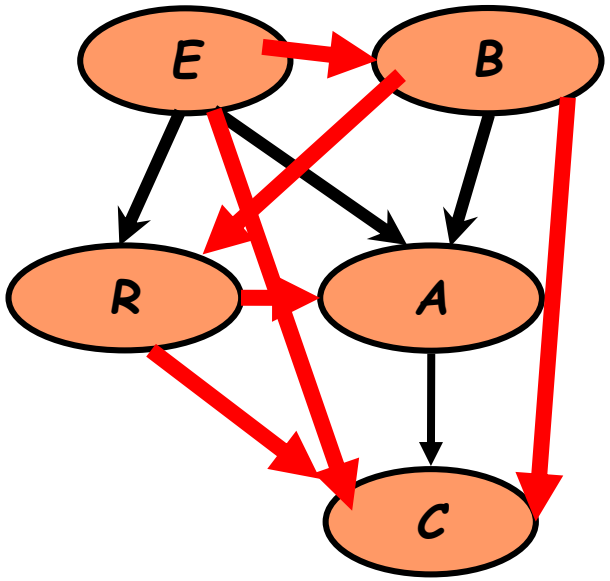


Five events, over time



time	Earthquake	Burglary	Radio	Alarm	Call
1	0	0	0	0	0
2	0	0	0	0	0
3	0	0	0	0	**1**
4	0	0	0	0	0
5	0	**1**	0	0	0
6	**1**	0	**1**	**1**	**1**
7	0	0	0	0	0

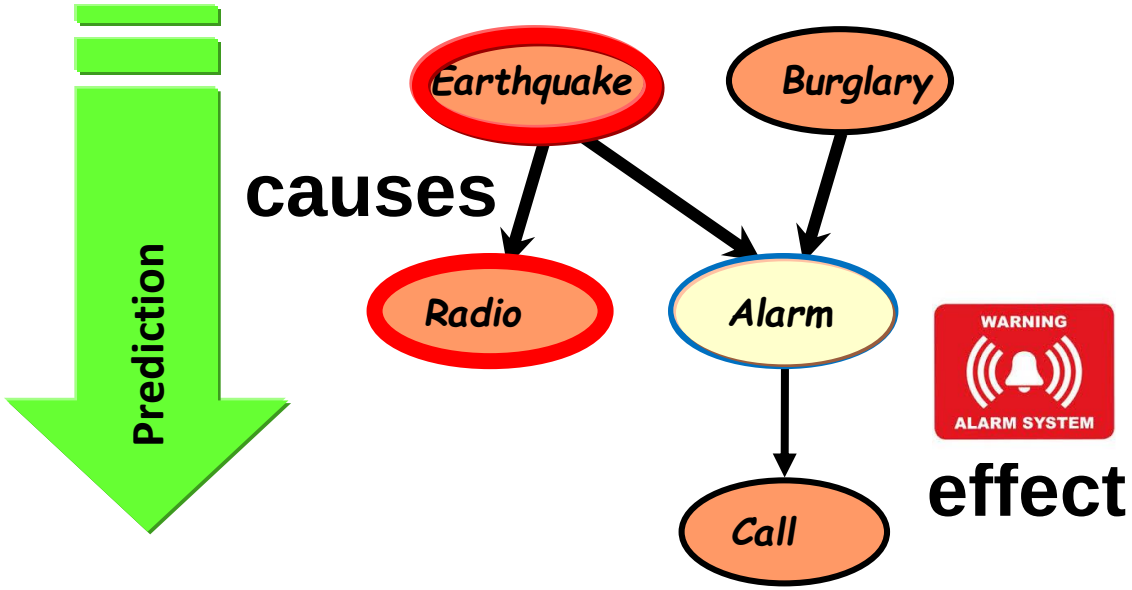
A Bayesian Network



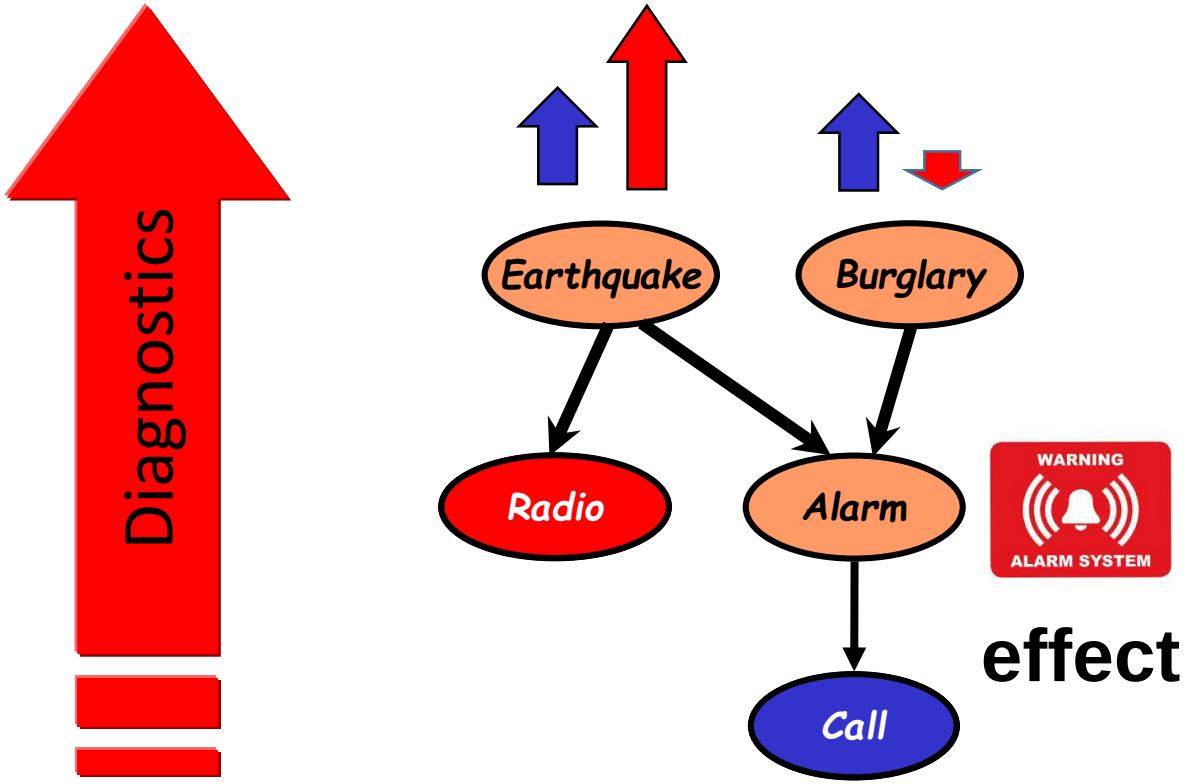
$$P(C,A,R,E,B) = P(B)*P(E|B)*P(R|E,B)*P(A|B,E,R)*P(C|A,R,B,E)$$

$$P(C,A,R,E,B) = P(B)*P(E)*P(R|E)*P(A|B,E)*P(C|A)$$

What is the effect of earthquake and radio on alarm?



$$P(Alarm \mid Earthquake, Radio) = P(Alarm \mid Earthquake)$$



What is causing the call?

The Law of Total Probability

Law of Total Probability

$$P(A) = \Sigma_B P(A, B)$$

$$= \Sigma_B P(A \mid B)\, P(B)$$ where B is any random variable

Why is this useful? given a joint distribution (e.g., P(A,B,C,D)) we can obtain any "marginal" probability e.g.,

$$P(B) = \Sigma_A \Sigma_C \Sigma_D P(A, B, C, D)$$

Less obvious: we can also compute <u>any conditional probability of interest</u> given a joint distribution,

e.g.,

$$P(c \mid b) = \Sigma_a \Sigma_d P(a, c, d \mid b)$$

$$= 1 / P(b) \; \Sigma_a \Sigma_d P(a, c, d, b)$$

where $1 / P(b)$ is just a normalization constant

Thus, the joint distribution contains the information we need to compute any probability of interest.

The Chain Rule

We can always write

$$P(a, b, c, \ldots z) = P(a \mid b, c, \ldots. z)\, P(b, c, \ldots z)$$

(by definition of joint probability)

Repeatedly applying this idea, we can write

$$P(a, b, c, \ldots z) = P(a \mid b, c, \ldots. z)\, P(b \mid c, .. z)\, P(c \mid .. z)..P(z)$$

This factorization holds for any ordering of the variables.

This is the chain rule for probabilities.

Conditional Independence

2 random variables A and B are conditionally independent given C iff

$$P(a, b \mid c) = P(a \mid c)\, P(b \mid c) \quad \text{for all values a, b, c}$$

More intuitive (equivalent) conditional formulation

A and B are conditionally independent given C iff

$P(a \mid b, c) = P(a \mid c)$ OR $P(b \mid a, c) = P(b \mid c)$ for all values a, b, c

Intuitive interpretation:

$P(a \mid b, c) = P(a \mid c)$ tells us that learning about b, given that we already know c, provides no change in our probability for a,

i.e., b contains no information about a beyond what c provides

Can generalize to more than 2 random variables

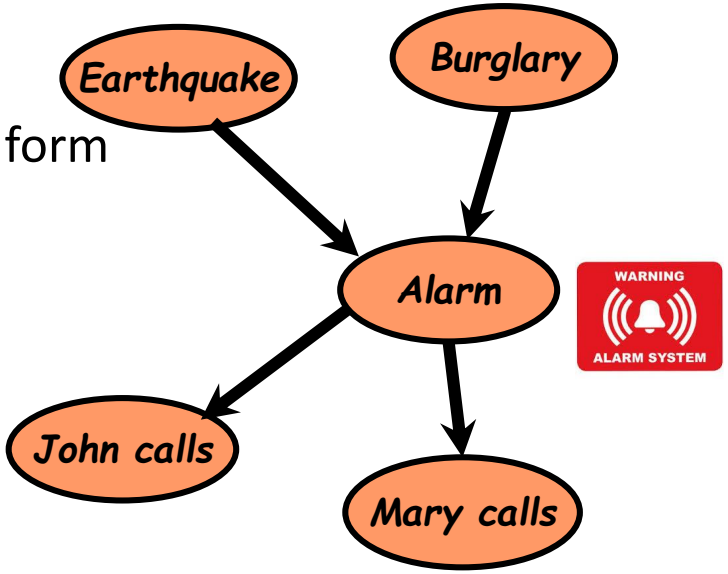
E.g., K different symptom variables $X_1, X_2, \ldots, X_K$, and C = disease

$P(X_1, X_2, \ldots, X_K \mid C) = \prod P(X_i \mid C)$

Also known as the naïve Bayes assumption

Bayesian Networks

- A Bayesian network specifies a joint distribution in a structured form
- Represent dependence/independence via a directed graph
  - Nodes = random variables
  - Edges = direct dependence
- Structure of the graph ⇔ Conditional independence relations



In general,

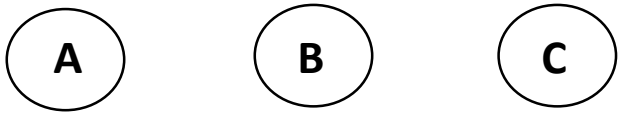
$$P(X_1, X_2, \ldots X_N) = \Pi\, P(X_i \mid \text{parents}(X_i))$$

The full joint distribution

The graph-structured approximation

- Requires that graph is acyclic (no directed cycles)
- 2 components to a Bayesian Network
  - The graph structure (conditional independence assumptions)
  - The numerical probabilities (for each variable given its parent)

A 3-way Bayesian Network

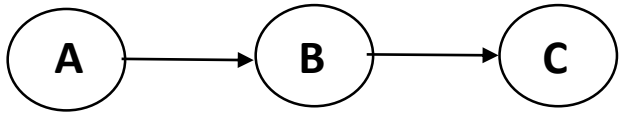


Marginal Independence:

P(A,B,C) = P(A) P(B) P(C)

A 3-way Bayesian Network

A chain

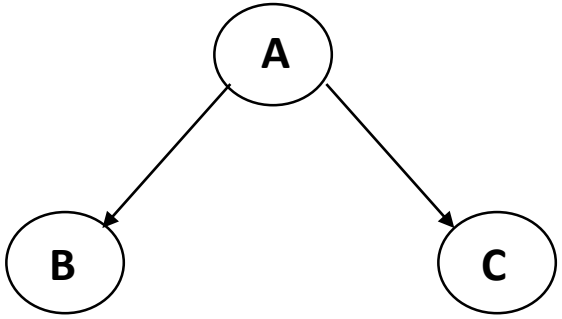


Markov dependence:

P(A,B,C) = P(C|B) P(B|A)P(A)

A 3-way Bayesian Network

A fork



Conditionally independent effects:

P(A,B,C) = P(B|A)P(C|A)P(A)

B and C are conditionally independent given A.

A 3-way Bayesian Network

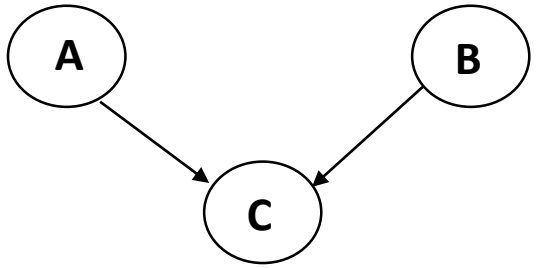
Battery

Fuel pump

A collider

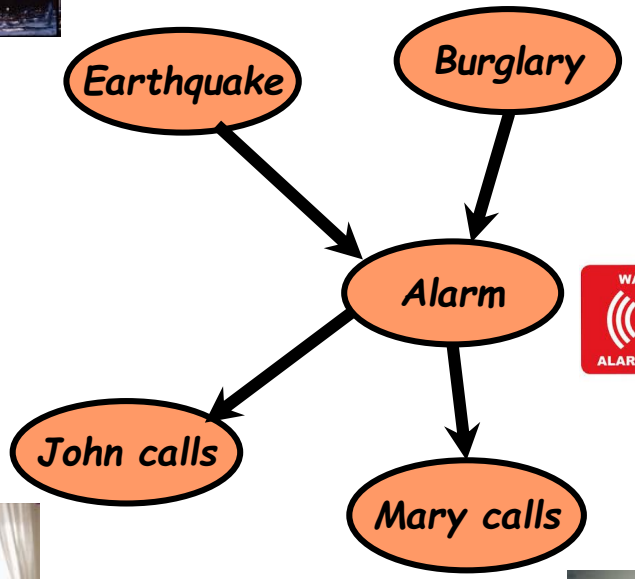
Independent Causes:

P(A,B,C) = P(C|A,B)P(A)P(B)



A car's engine can fail to start (C) due either to a dead battery (A) or due to a blocked fuel pump (B). Ordinarily, we assume that battery death and fuel pump blockage are independent events, because of the essential modularity of such automotive systems. Thus, in the absence of other information, knowing whether or not the battery is dead gives us no information about whether or not the fuel pump is blocked. However, if we happen to know that the car fails to start (i.e., we fix common effect (C), this information induces a dependency between the two causes battery death and fuel blockage. **Thus, knowing that the car fails to start, if an inspection shows the battery to be in good health, we can conclude that the fuel pump must be blocked.**

Burglary example revisited



Consider the following 5 binary variables:

B = a burglary occurs at your house

E = an earthquake occurs at your house

A = the alarm goes off

J = John calls to report the alarm

M = Mary calls to report the alarm

What is P(B | J, M) ?

- We can use the full joint distribution to answer this question

  This requires $2^5 = 32$ probabilities

- Alternatively, we can use prior domain knowledge to come up with a Bayesian Network with fewer probabilities

Constructing a Bayesian Network

Order the variables in terms of causality

e.g., {E, B} -> {A} -> {J, M}

$P(J, M, A, E, B) = P(J, M \mid A, E, B)\, P(A \mid E, B)\, P(E, B)$

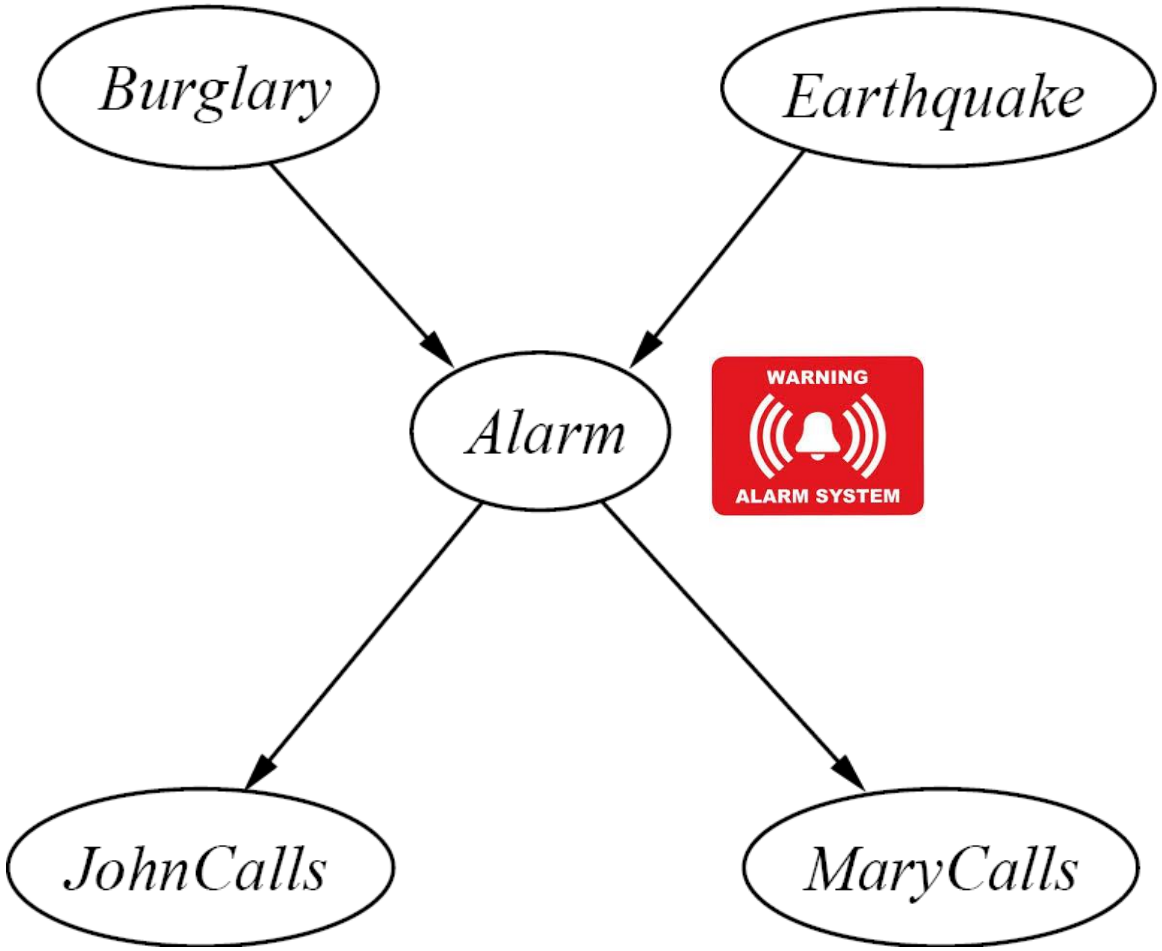
$\sim P(J, M \mid A)\, P(A \mid E, B)\, P(E)\, P(B)$

$\sim P(J \mid A)\, P(M \mid A)\, P(A \mid E, B)\, P(E)\, P(B)$

These causality assumptions are reflected in the graph structure of the Bayesian Network

Unconstrained joint distribution requires $O(2^n)$ probabilities. If we have a Bayesian network, with a maximum of k parents for any node, then we need O(n 2k) probabilities. Example: Full unconstrained joint distribution with n = 30 needs 10^9 probabilities for full joint distribution but binary Bayesian network with n = 30, k = 4, requires only 480 probabilities.

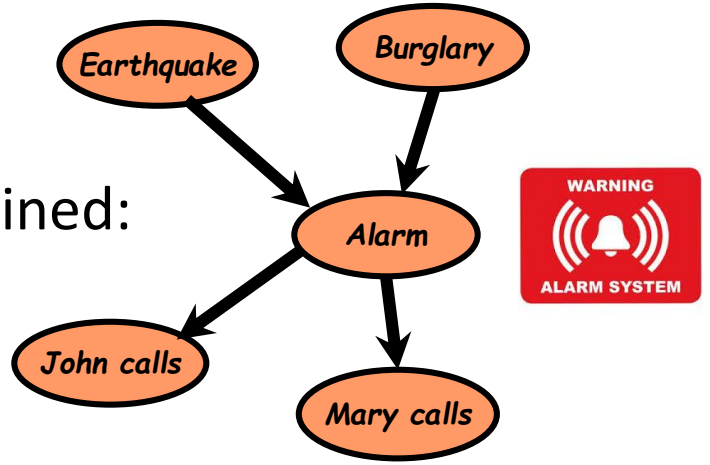
The Burglary Bayesian Network Structure



Constructing the Bayesian Network

P(J, M, A, E, B) =

P(J | A) P(M | A) P(A | E, B) P(E) P(B)



There are 3 conditional probability tables (CPDs) to be determined:
P(J | A), P(M | A), P(A | E, B)

Requiring 2 + 2 + 4 = 8 probabilities

And 2 marginal probabilities P(E), P(B) -> 2 more probabilities

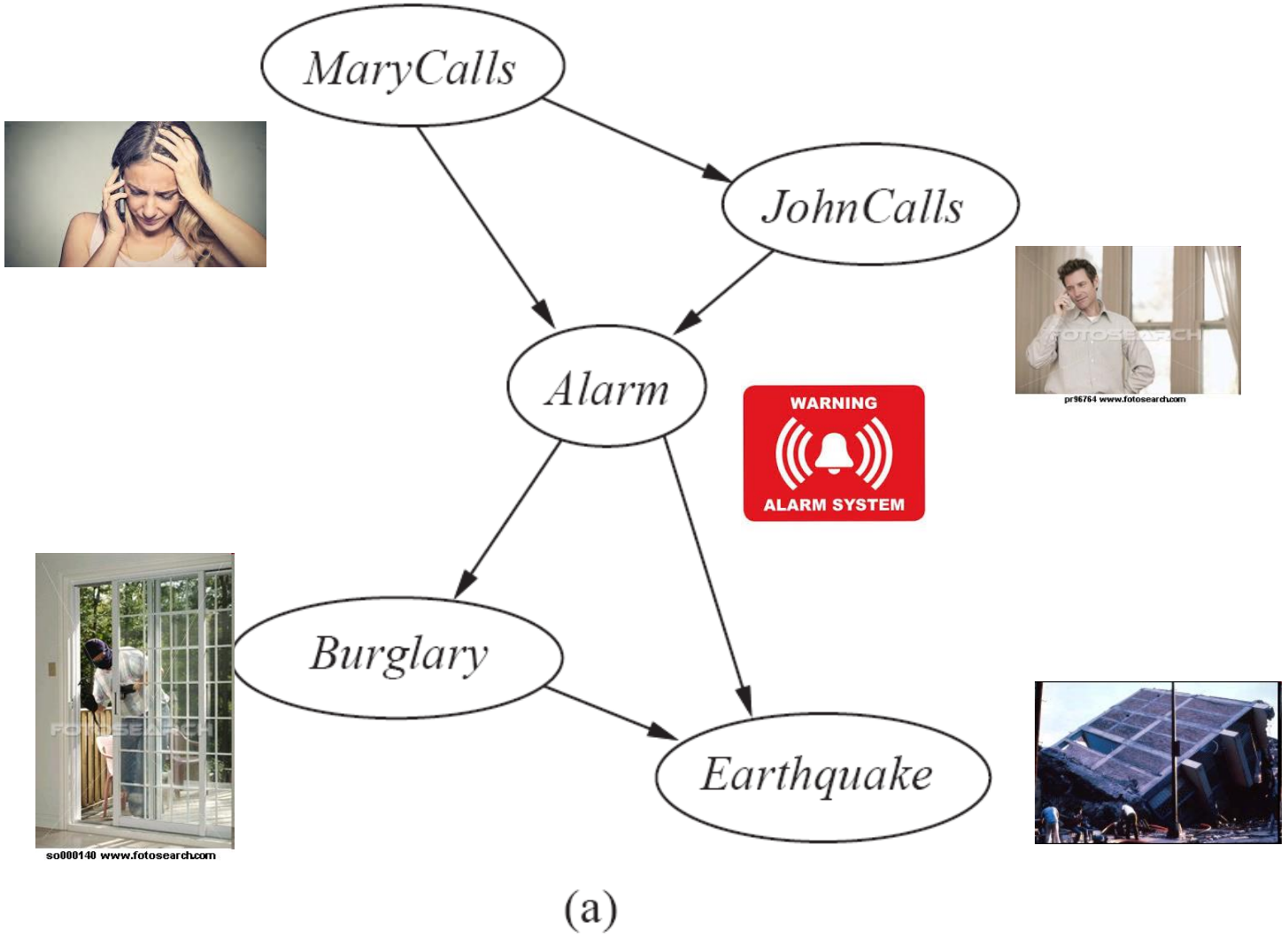
These probabilities come from

- Expert knowledge
- From data (relative frequency estimates)
- Or a combination of both

The Bayesian Network

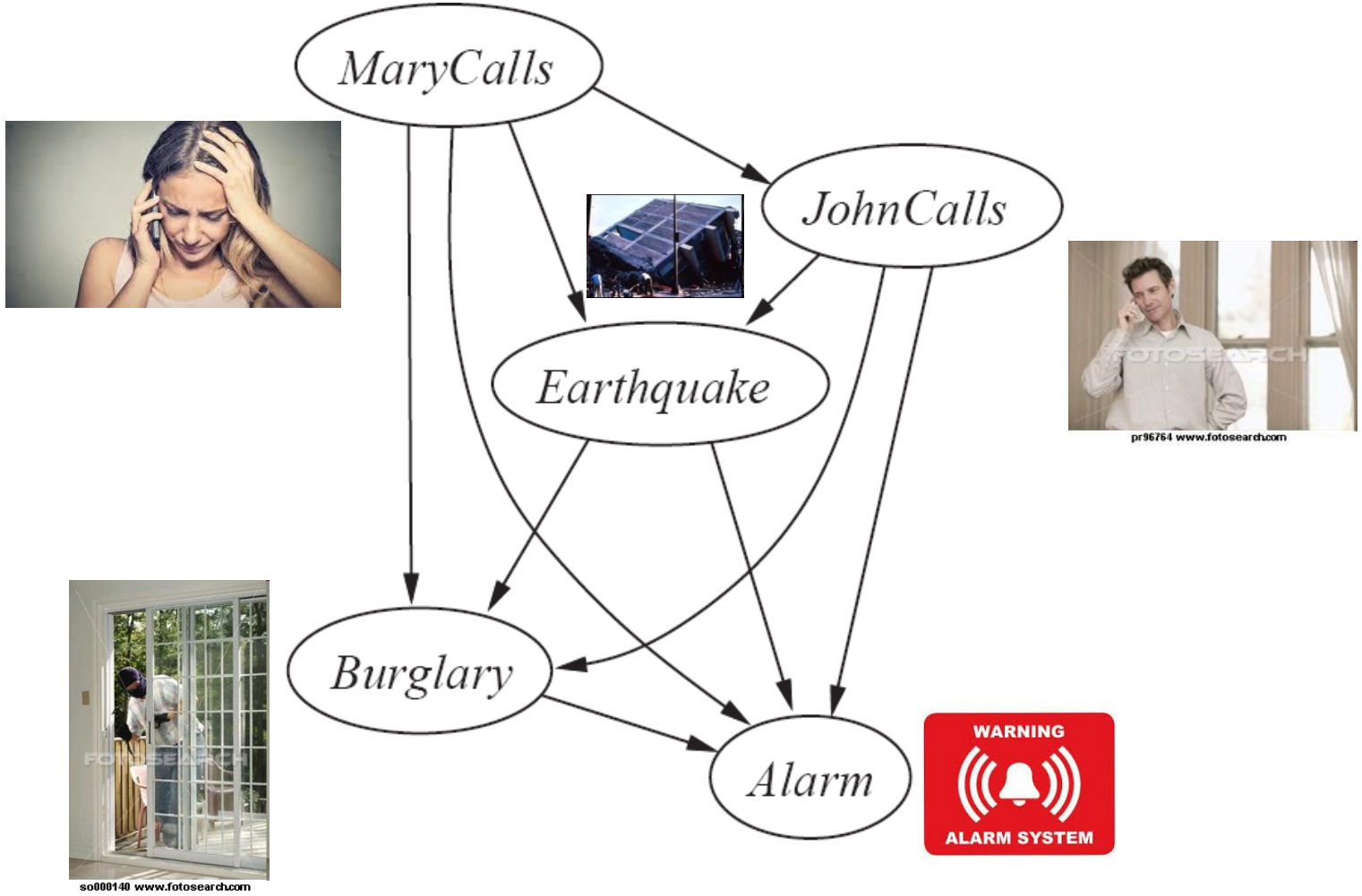
10 probabilities
Versus
2^5-1=32-1=31

The Bayesian Network for a different variable ordering



(a)

The Bayesian Network for a different variable ordering



(b)

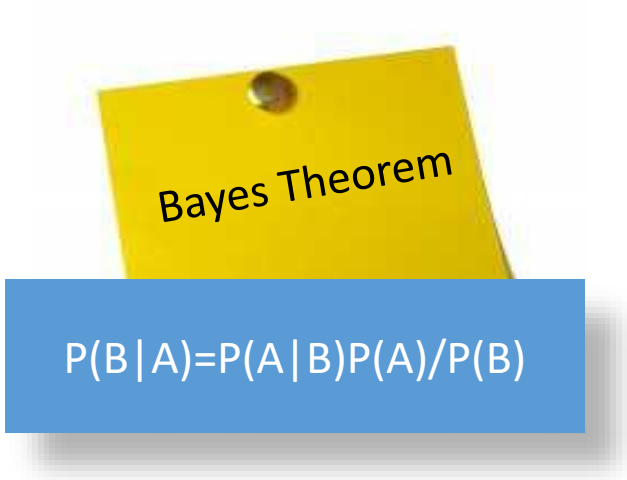
Inference (Reasoning) in Bayesian Networks

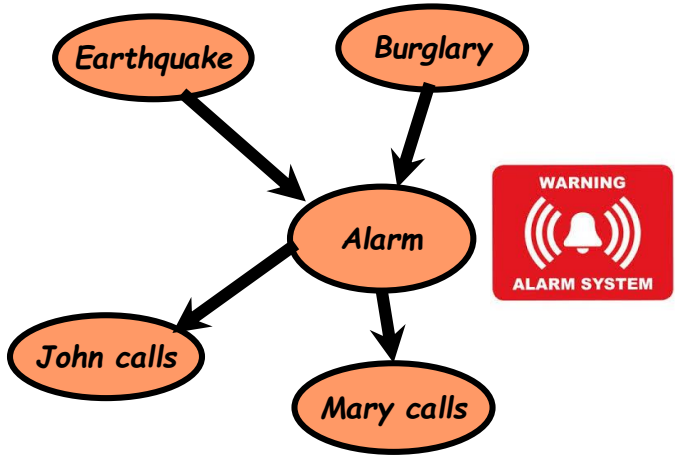
Consider answering a query in a Bayesian Network

- Q = set of query variables
- e = evidence (set of instantiated variable-value pairs)
- Inference = computation of conditional distribution P(Q|e)

Examples

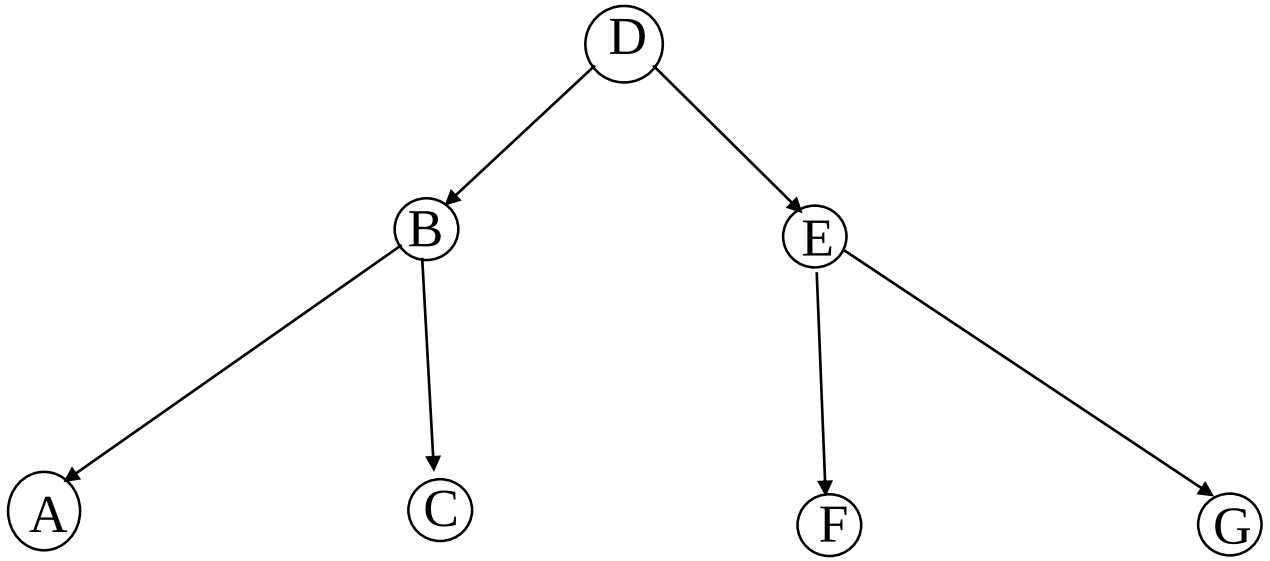
- P(Burglary | Alarm)
- P(Earthquake | JCalls, MCalls)
- P(JCalls, MCalls | Burglary, Earthquake)





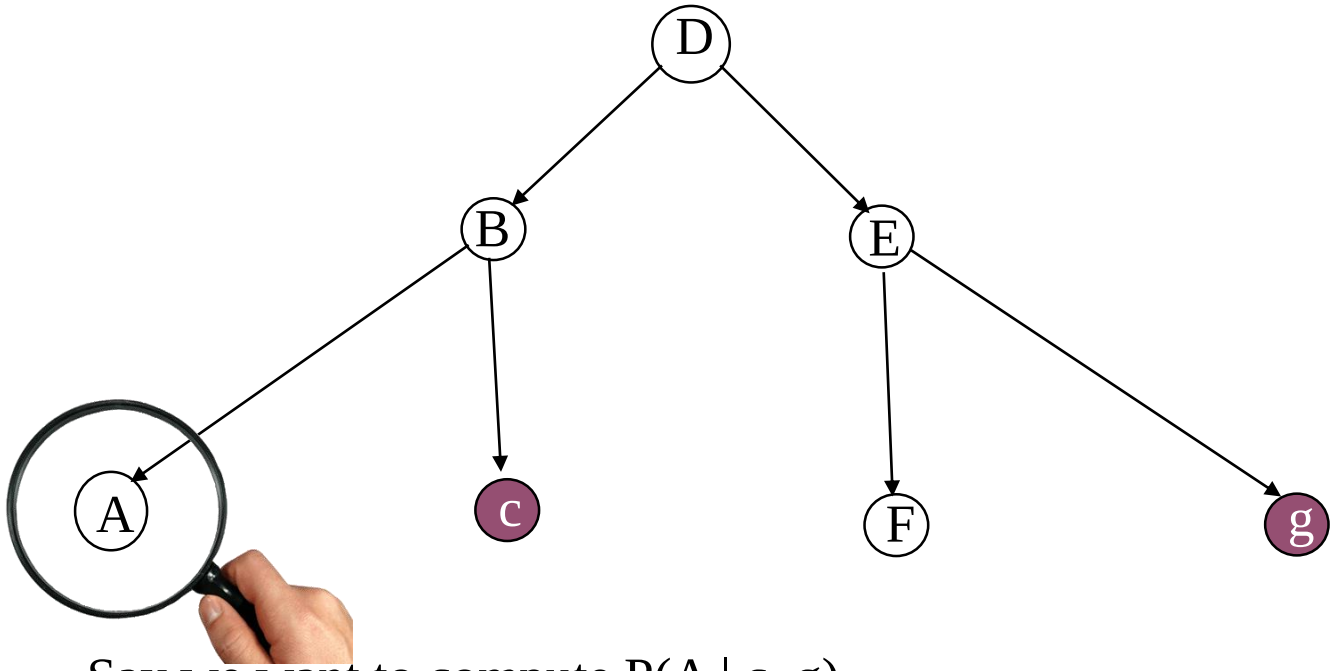
We can use the structure of the Bayesian Network to answer such queries efficiently

Example



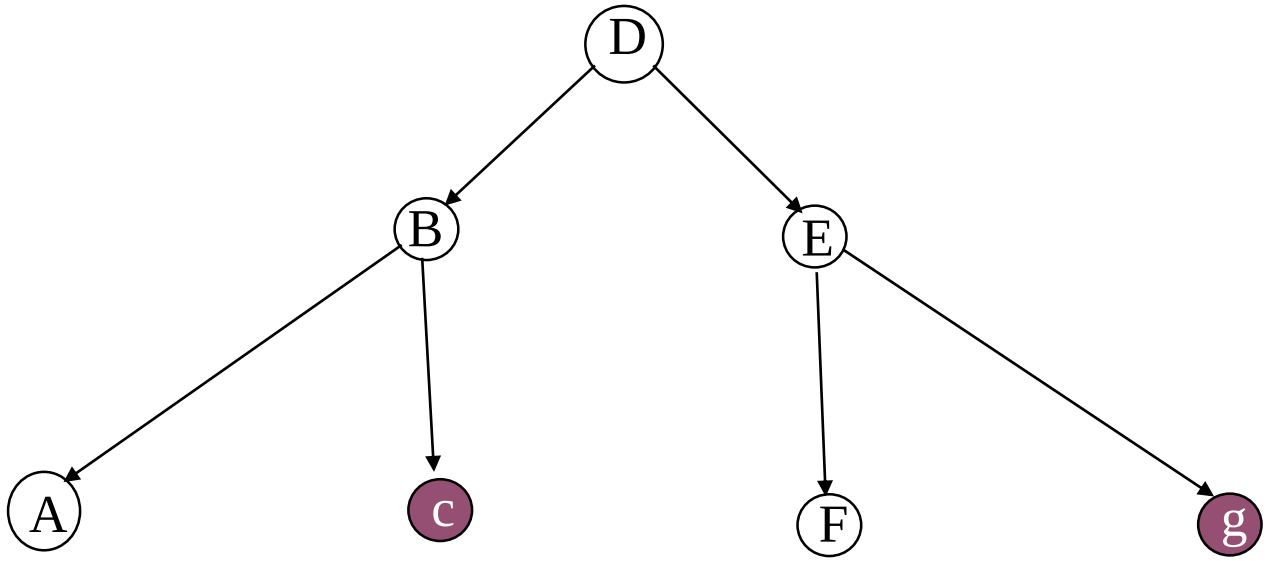
P(A, B, C, D, E, F, G) is modeled as P(A|B)P(C|B)P(F|E)P(G|E)P(B|D)P(E|D)P(D)

Example



Say we want to compute P(A | c, g)

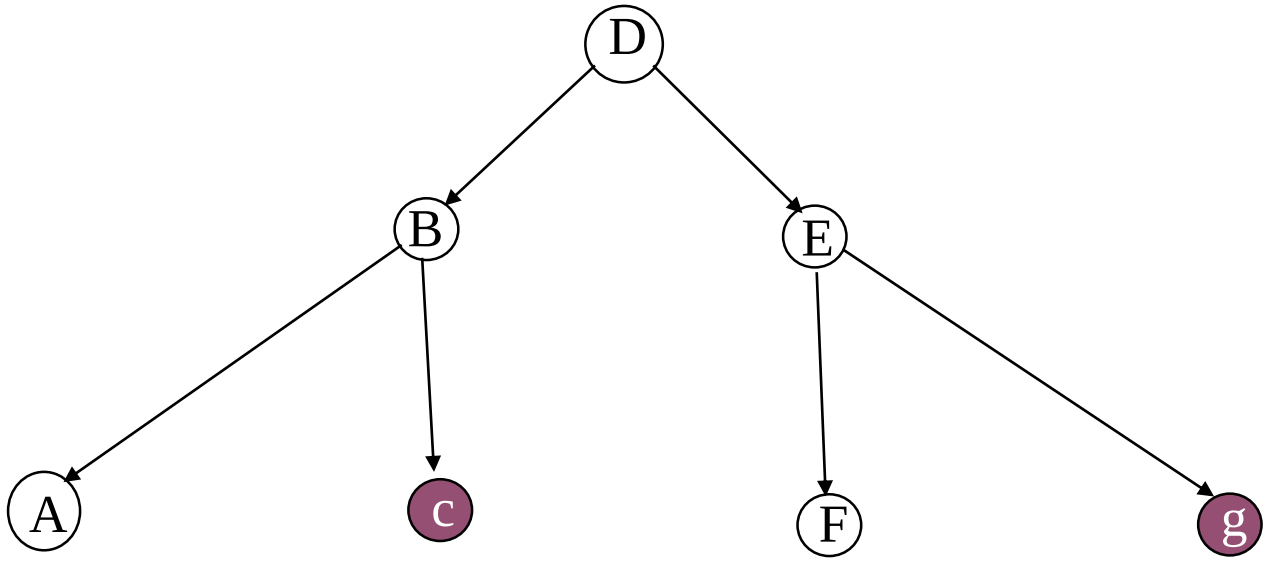
Example



Direct calculation: $P(A|c,g) = \sum_{BDEF} P(A,B,D,E,F \mid c,g)$

Complexity of the sum is $O(m^4)$

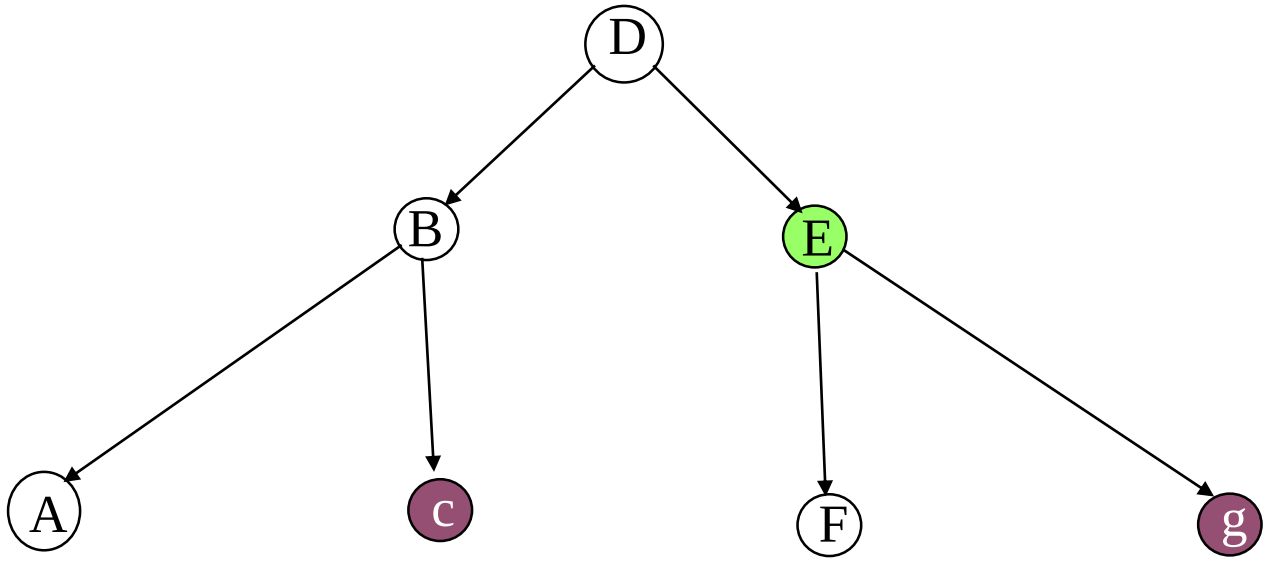
Example



Reordering:

$$\Sigma_D P(A|B) \Sigma_D P(B|D,c) \Sigma_E P(D|E) \Sigma_F P(E,F|g)$$

Example

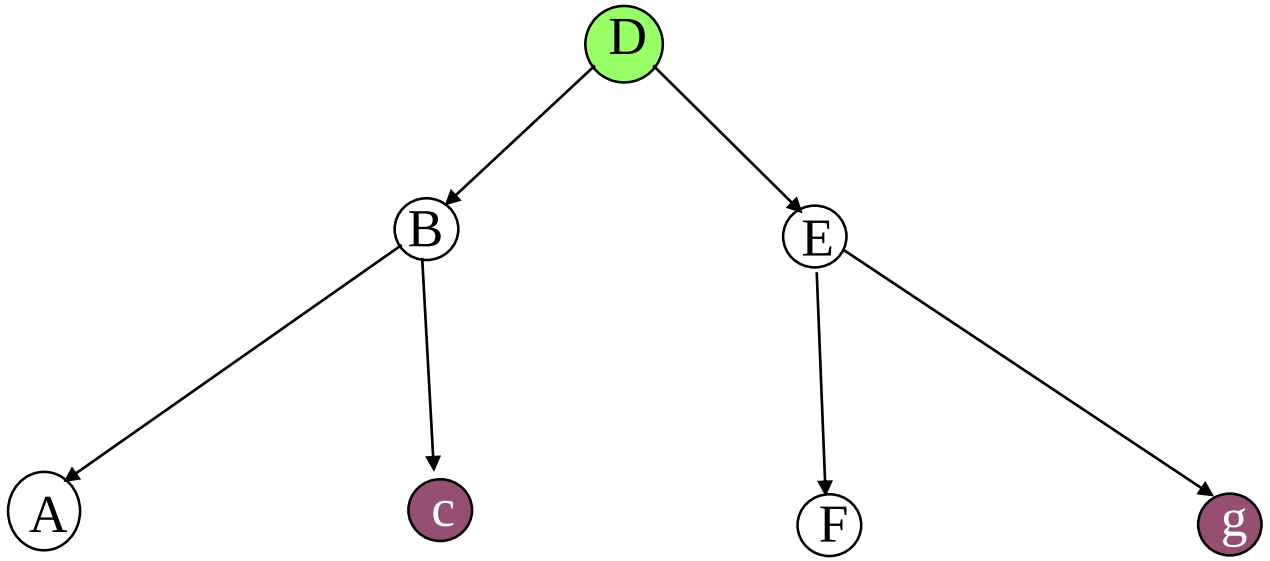


Reordering:

$$\Sigma_B\ P(A|B)\ \Sigma_D\ P(B|D,c)\ \Sigma_E\ P(D|E)\ \Sigma_F\ P(E,F\ |g)$$

$$P(E|g)$$

Example

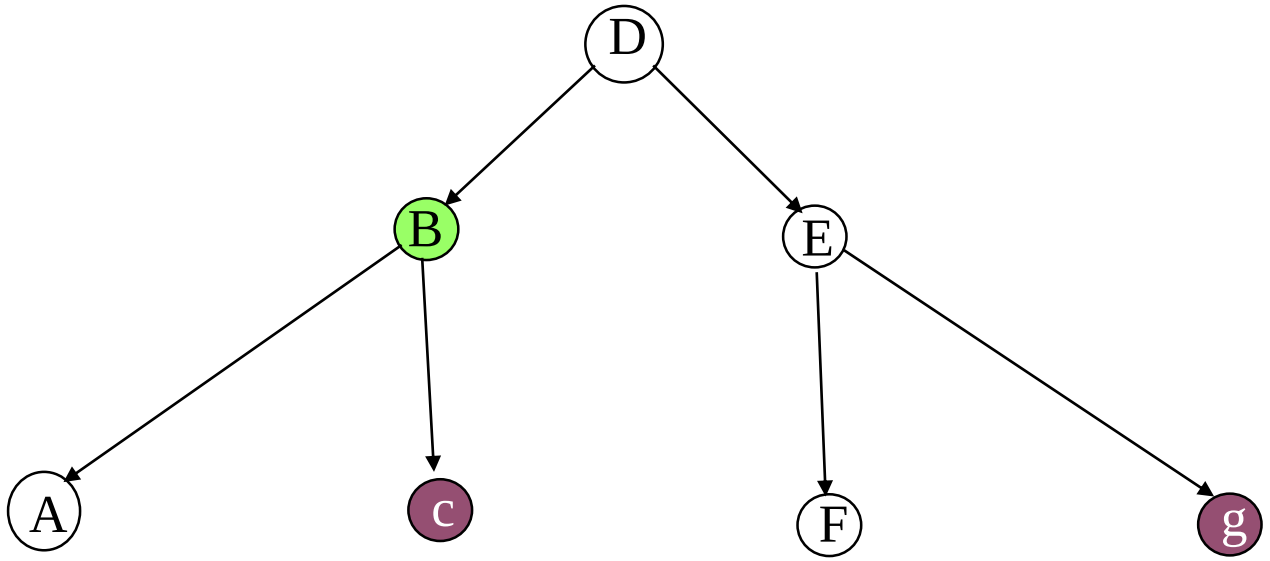


Reordering:

$$\Sigma_b\ p(a|b)\ \Sigma_d\ p(b|d,c)\ \Sigma_e\ p(d|e)\ p(e|g)$$

$$p(d|g)$$

Example

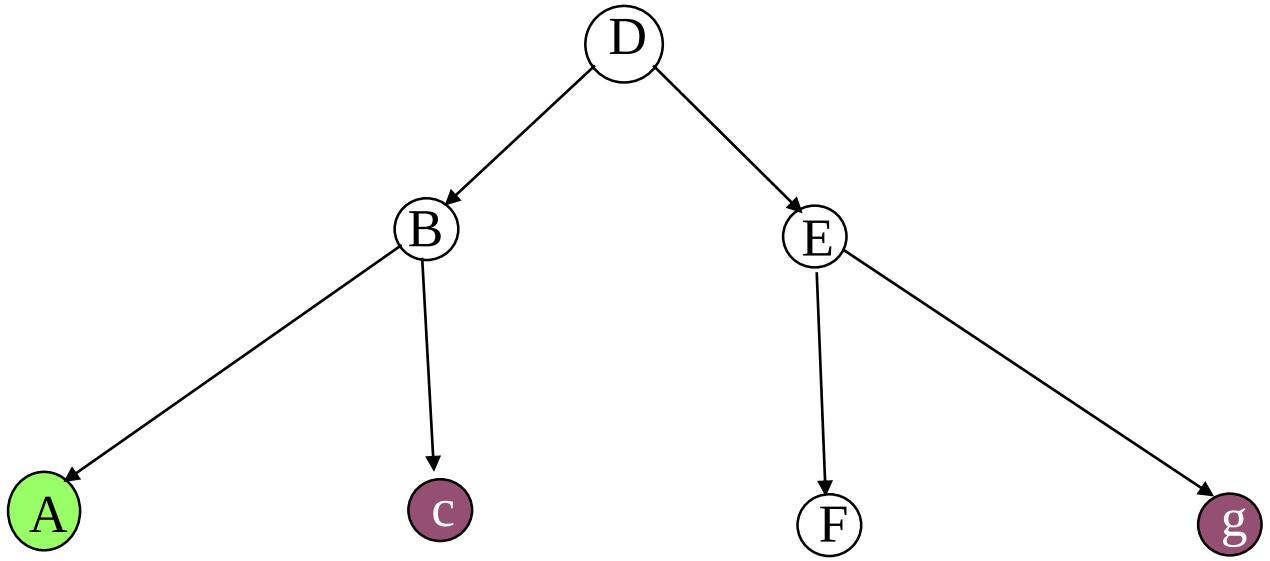


Reordering:

$$\sum_B P(A|B) \sum_D P(B|D,c) P(D|g)$$

$P(B|c,g)$

Example



Reordering:

$\sum_b P(A|B)\, P(B|c,g)$

$P(A|c,g)$

Complexity is $O(m)$, compared to $O(m^4)$

Real-valued Variables

Bayesian Networks can also handle Real-valued variables

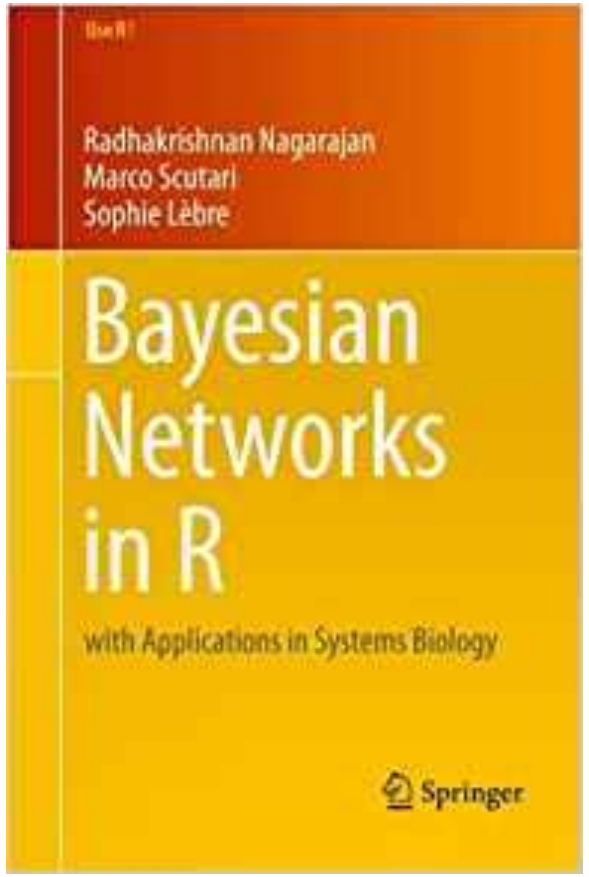
- If we can assume variables are Gaussian, then the inference and theory for Bayesian networks is well-developed,
  - E.g., conditionals of a joint Gaussian is still Gaussian, etc.
  - In inference we replace sums with integrals
- For other density functions it depends…
  - Can often include a univariate variable at the “edge” of a graph, e.g., a Poisson conditioned on day of week
- But for many variables there is little know beyond their univariate properties, e.g., what would be the joint distribution of a Poisson and a Gaussian? (its not defined)
- Common approaches in practice
  - Put real-valued variables at “leaf nodes” (so nothing is conditioned on them)
  - Assume real-valued variables are Gaussian or discrete
  - Discretize real-valued variables

Take home bullets

- Bayesian networks represent a joint distribution using a graph
- The graph encodes a set of conditional independence assumptions
- Answering queries (or inference or reasoning) in a Bayesian network amounts to efficient computation of appropriate conditional probabilities
- Probabilistic inference is intractable in the general case but can be carried out in linear time for Bayesian networks

Journal of Statistical Software

July 2010, Volume 35, Issue 3. http://www.jstatsoft.org/

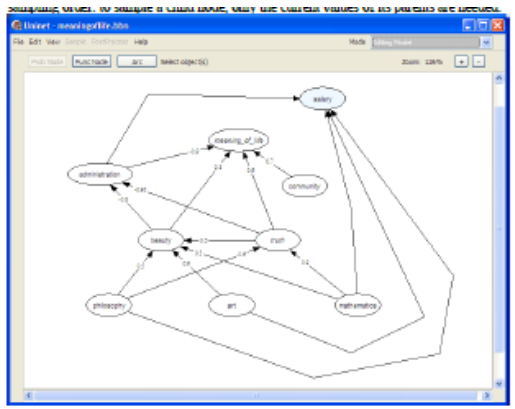


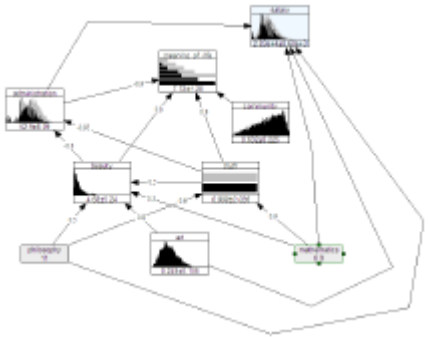
https://cran.r-project.org/web/packages/bnlearn

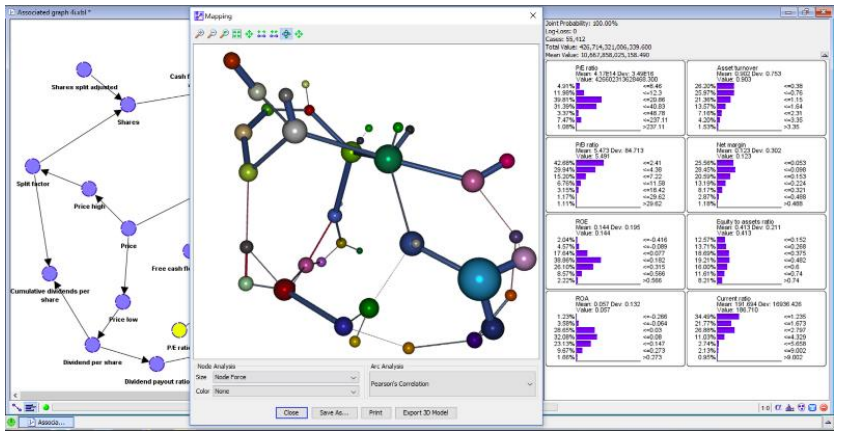
Learning Bayesian Networks with the bnlearn R Package

Marco Scutari
University of Padova

http://www.lighttwist.net/wp/uninet







BAYESIALAB

Microsoft Research

MSBNx
Bayesian Network Editor and Tool Kit

https://www.microsoft.com/en-us/download/confirmation.aspx?id=52299

MSBNx is a component-based Windows application for creating, assessing, and evaluating Bayesian Networks, created at Microsoft Research

https://msbnx.azurewebsites.net/msbnx/what_is_msbnx.htm

University of Pittsburgh

BAYESFUSION,LLC
Data Analytics, Modeling, Decision Support

GeNIe 2.1

Decision Systems Laboratory. Department of Information Science and Telecommunications and the Intelligent Systems Program at the University of Pittsburgh. www.bayesfusion.com

Causal probabilistic network modeling—An illustration of its role in the management of chronic diseases

by R. Hovorka
S. Andreassen
J. J. Benn
K. G. Olesen
E. R. Carson

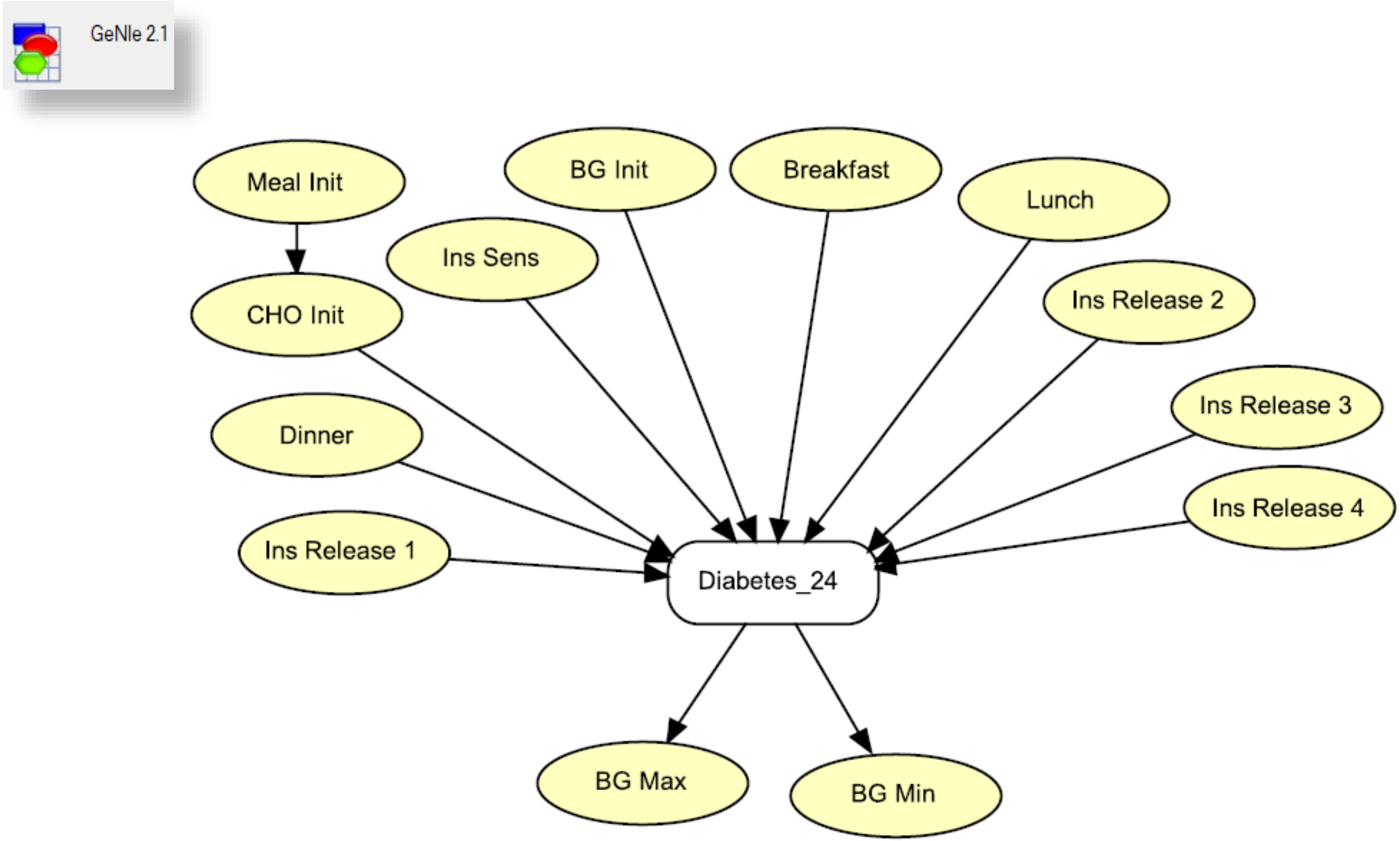
This paper describes the role of the novel technique of causal probabilistic network (CPN) modeling as an approach to tackling control system problems typified by that of the administration of treatment to the patient suffering from a chronic disease such as diabetes. Three roles of a CPN are discussed. First, since diabetes arises as a consequence of impaired control of carbohydrate metabolism, the ability of a CPN to represent the uncertainty of a physiologically-based model is described. Second, its ability to make robust estimates of the parameters of the metabolic model is presented, and finally, in conjunction with decision theory approaches, its ability to compare alternative therapies and advise on insulin therapy for patients with insulin-dependent diabetes mellitus is illustrated.

The management of chronic noncommunicable diseases such as diabetes (diabetes mellitus), raised blood pressure (hypertension), and elevated levels of cholesterol poses some difficult challenges for the clinician. In most cases, from an engineering or systems perspective, such diseases can be viewed as arising from a partial or complete failure of one or more of the multitude of feedback control loops of the human organism. The management of such diseases requires regu-

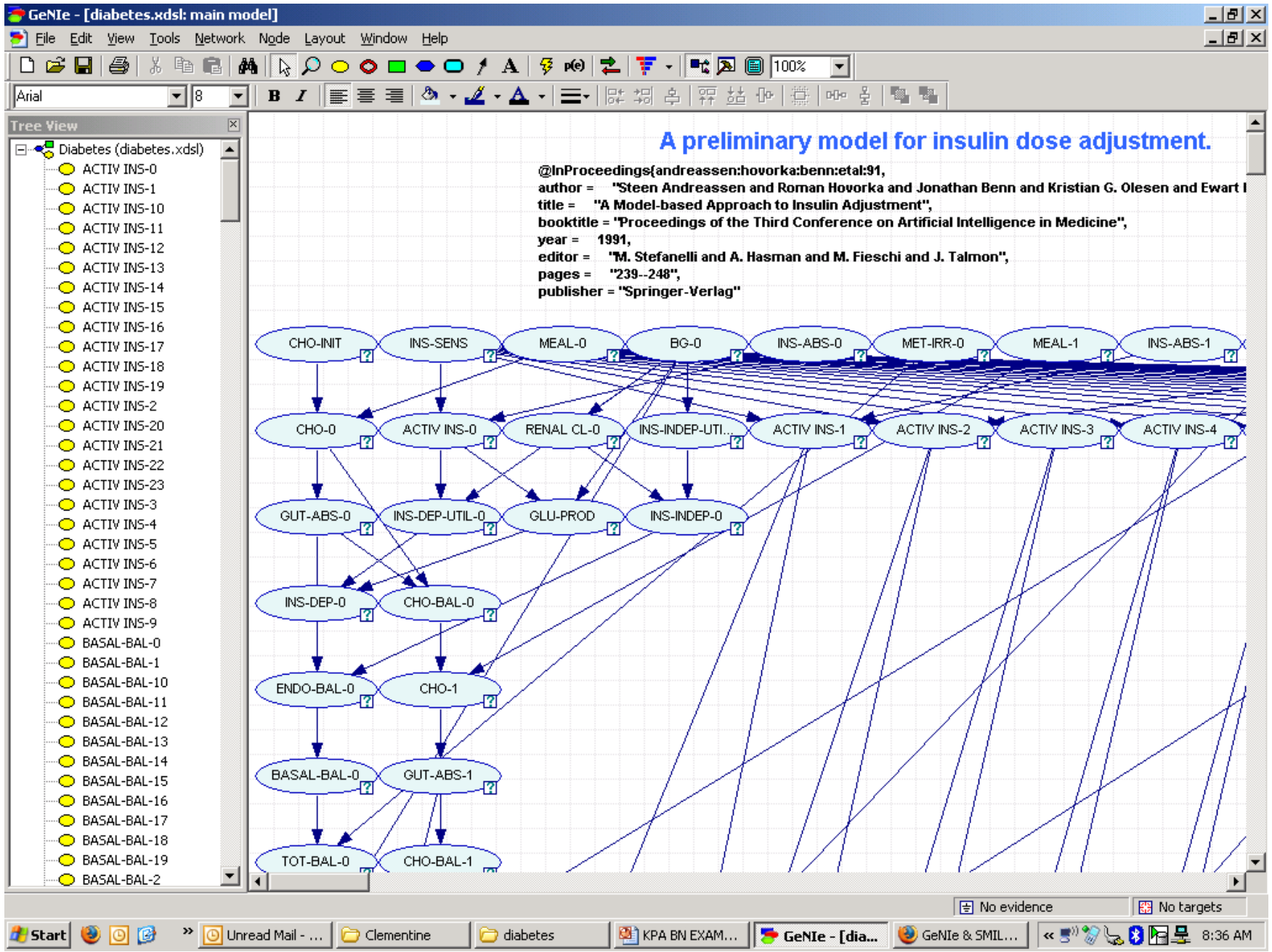
IBM SYSTEMS JOURNAL, VOL 31, NO 4, 1992

diabetes.xdsl

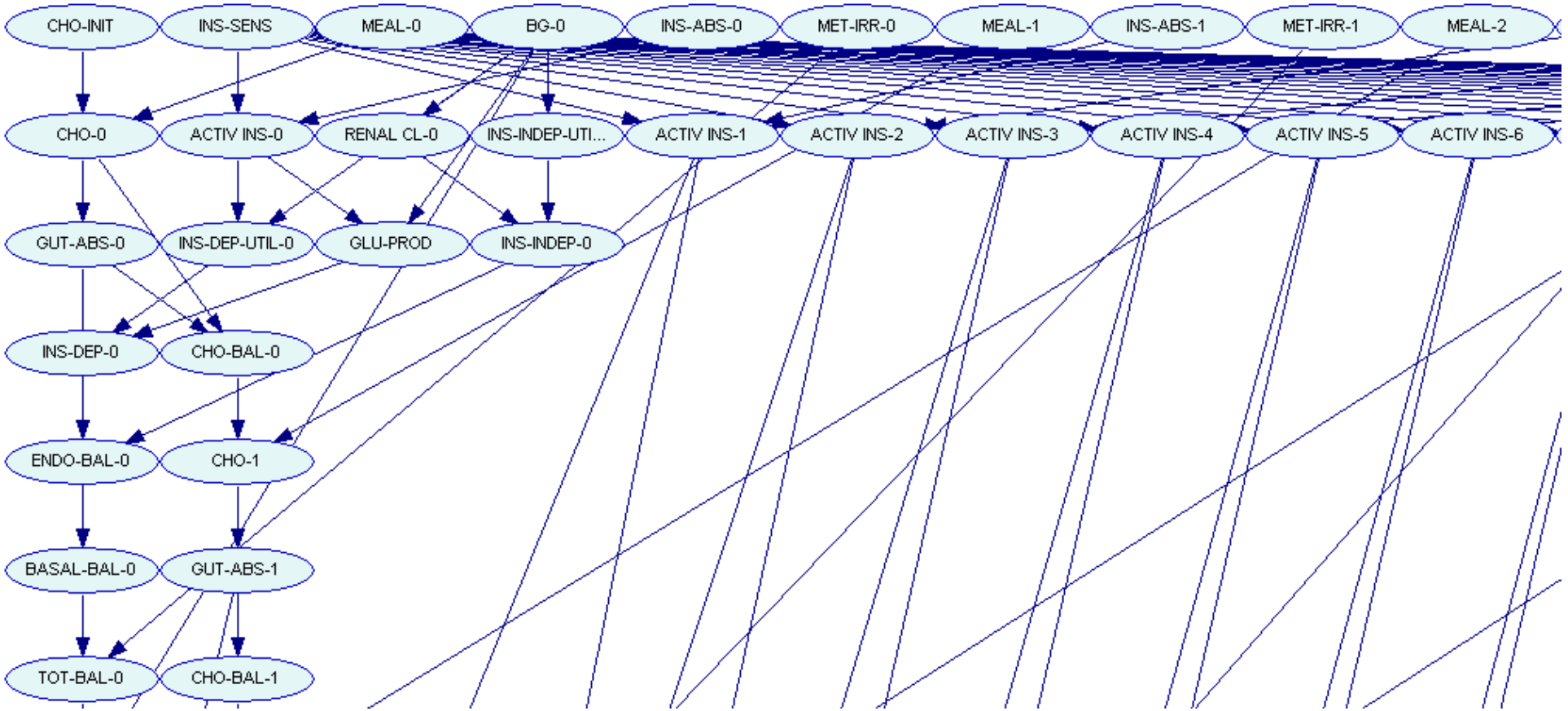
The basic building block of the system is a one hour model of the intake and utilization of food, blood glucose and insulin. The nodes BG and CHO acts as status variables denoting respectively the glucose in the blood stream and the glucose reservoir in the stomach. Intermediate nodes are primarily describing processes that utilizes the glucose

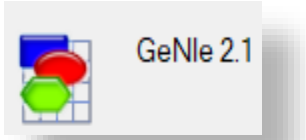

GeNIe 2.1
Meal Init
BG Init
Breakfast
Lunch
Ins Sens
CHO Init
Ins Release 2
Dinner
Ins Release 3
Ins Release 4
Ins Release 1
Diabetes_24
BG Max
BG Min

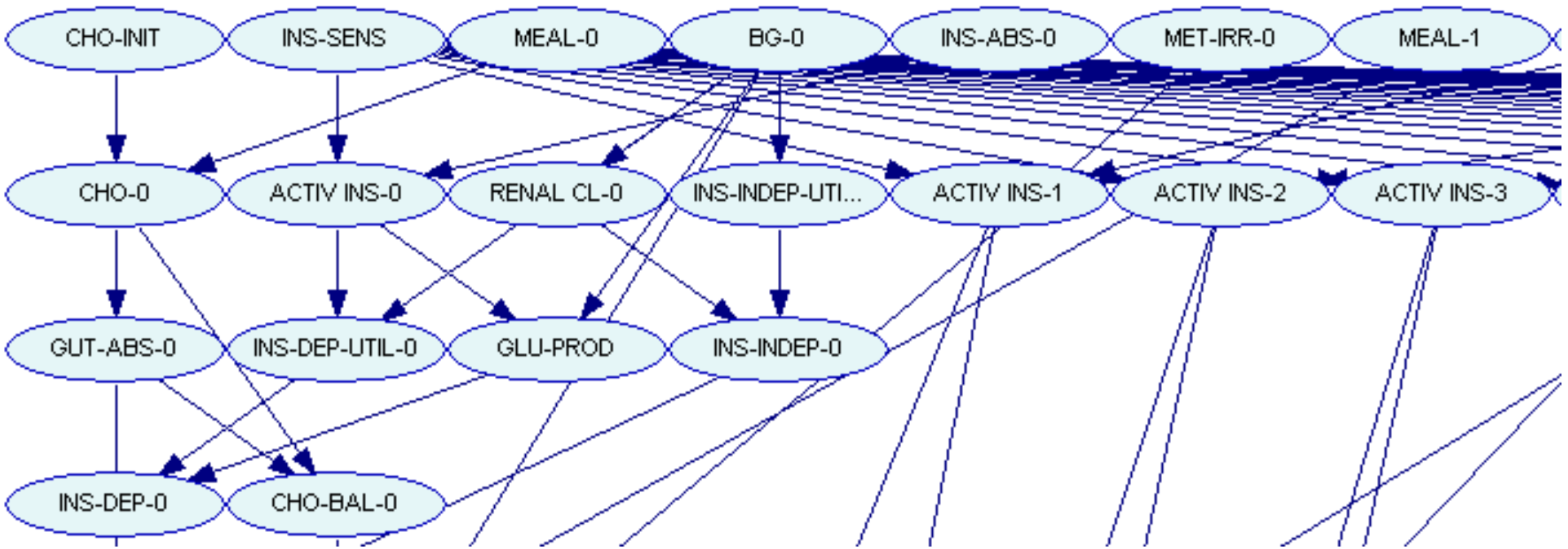
GeNIe 2.1


GeNIe - [diabetes.xdsl: main model]
File Edit View Tools Network Node Layout Window Help
Tree View
Diabetes (diabetes.xdsl)
ACTIV INS-0
ACTIV INS-1
ACTIV INS-10
ACTIV INS-11
ACTIV INS-12
ACTIV INS-13
ACTIV INS-14
ACTIV INS-15
ACTIV INS-16
ACTIV INS-17
ACTIV INS-18
ACTIV INS-19
ACTIV INS-2
ACTIV INS-20
ACTIV INS-21
ACTIV INS-22
ACTIV INS-23
ACTIV INS-3
ACTIV INS-4
ACTIV INS-5
ACTIV INS-6
ACTIV INS-7
ACTIV INS-8
ACTIV INS-9
BASAL-BAL-0
BASAL-BAL-1
BASAL-BAL-10
BASAL-BAL-11
BASAL-BAL-12
BASAL-BAL-13
BASAL-BAL-14
BASAL-BAL-15
BASAL-BAL-16
BASAL-BAL-17
BASAL-BAL-18
BASAL-BAL-19
BASAL-BAL-2
A preliminary model for insulin dose adjustment.
@InProceedings{andreassen:hovorka:benn:etal:91,
author = "Steen Andreassen and Roman Hovorka and Jonathan Benn and Kristian G. Olesen and Ewart
title = "A Model-based Approach to Insulin Adjustment",
booktitle = "Proceedings of the Third Conference on Artificial Intelligence in Medicine",
year = 1991,
editor = "M. Stefanelli and A. Hasman and M. Fieschi and J. Talmon",
pages = "239--248",
publisher = "Springer-Verlag"
CHO-INIT
INS-SENS
MEAL-0
BG-0
INS-ABS-0
MET-IRR-0
MEAL-1
INS-ABS-1
CHO-0
ACTIV INS-0
RENAL CL-0
INS-INDEP-UTI...
ACTIV INS-1
ACTIV INS-2
ACTIV INS-3
ACTIV INS-4
GUT-ABS-0
INS-DEP-UTIL-0
GLU-PROD
INS-INDEP-0
INS-DEP-0
CHO-BAL-0
ENDO-BAL-0
CHO-1
BASAL-BAL-0
GUT-ABS-1
TOT-BAL-0
CHO-BAL-1
No evidence
No targets
Start
Unread Mail - ...
Clementine
diabetes
KPA BN EXAM...
GeNIe - [dia...
GeNIe & SMIL...
8:36 AM

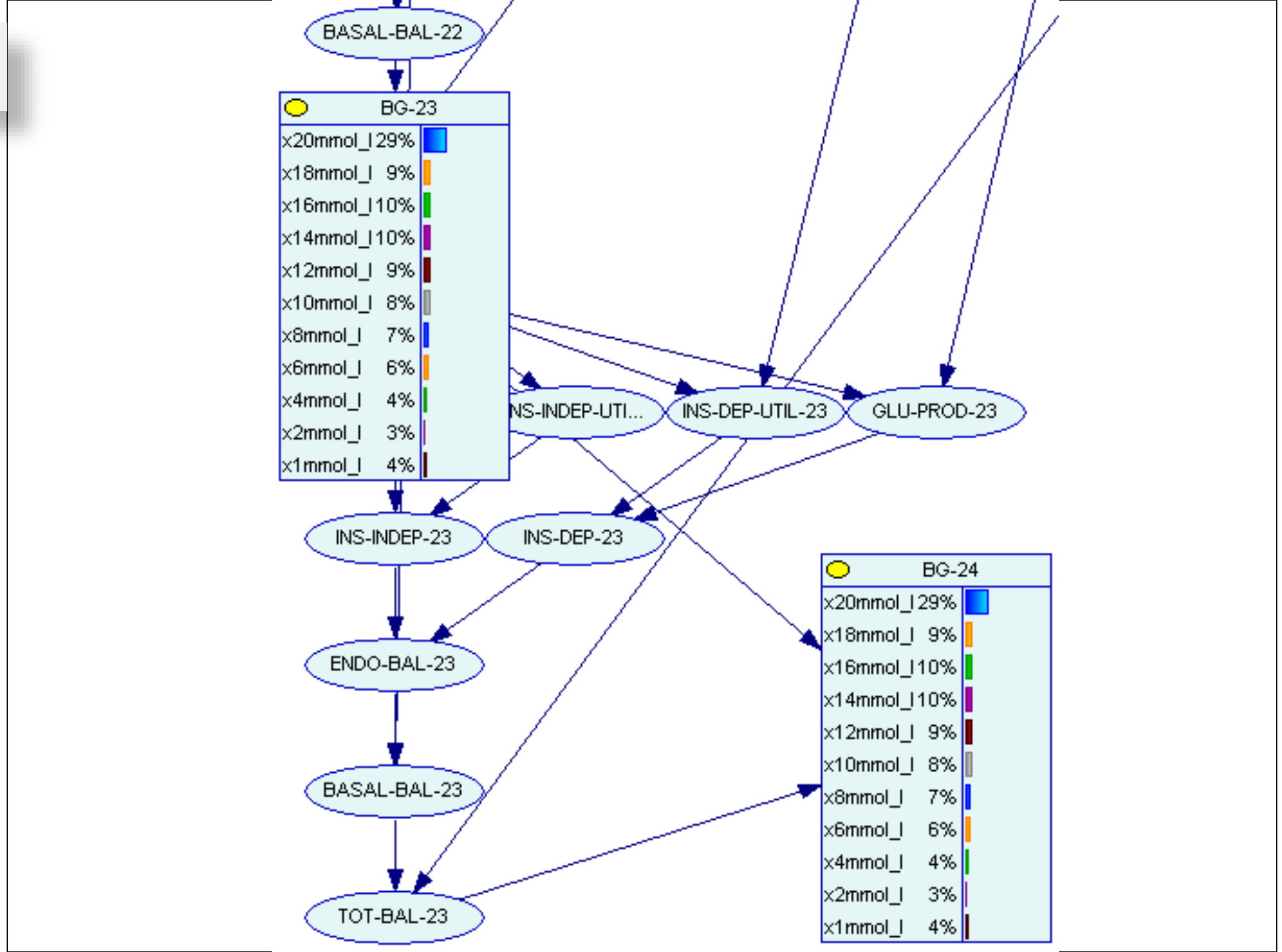
GeNIe 2.1

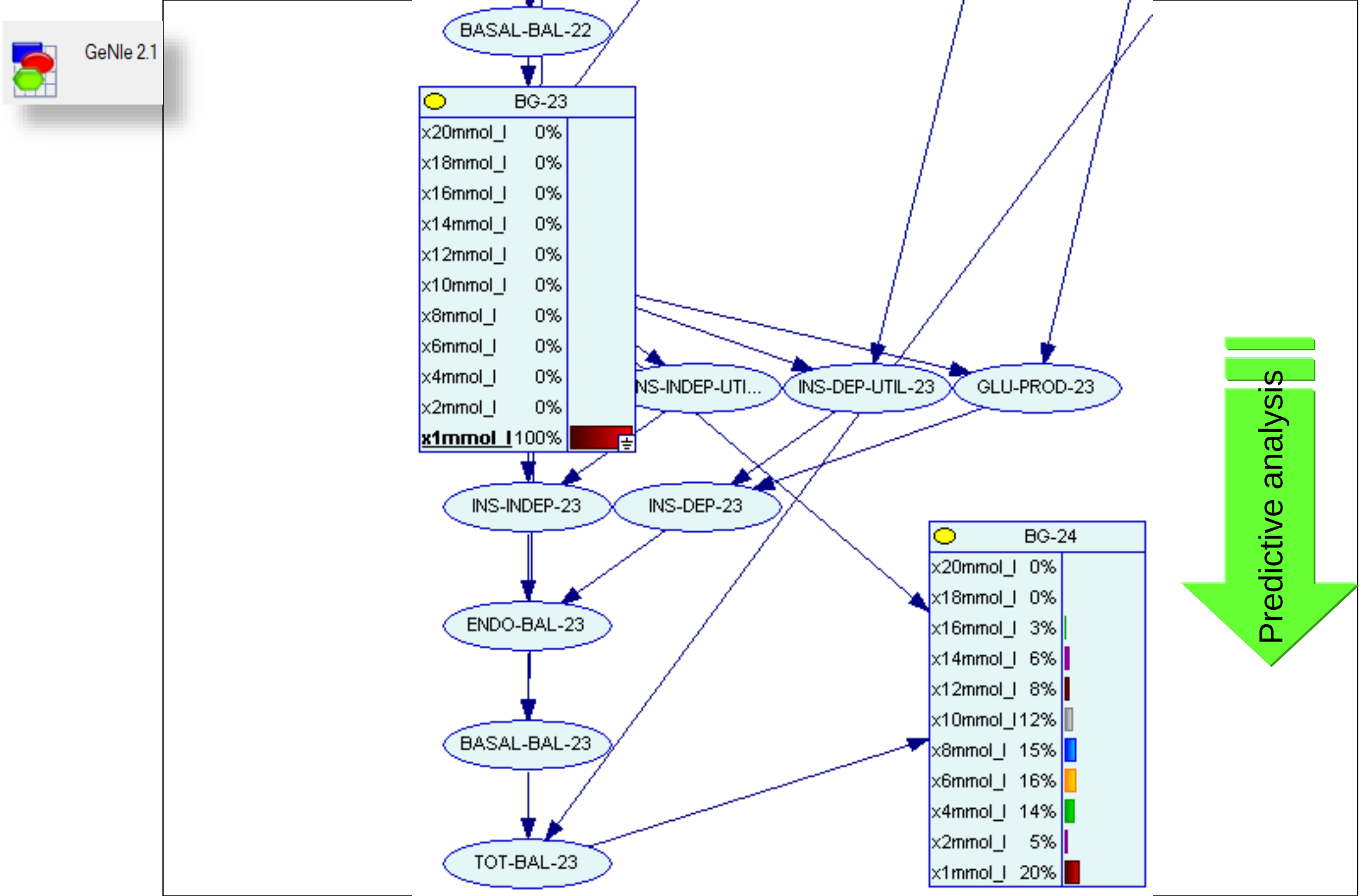

CHO-INIT
INS-SENS
MEAL-0
BG-0
INS-ABS-0
MET-IRR-0
MEAL-1
INS-ABS-1
MET-IRR-1
MEAL-2
CHO-0
ACTIV INS-0
RENAL CL-0
INS-INDEP-UTI...
ACTIV INS-1
ACTIV INS-2
ACTIV INS-3
ACTIV INS-4
ACTIV INS-5
ACTIV INS-6
GUT-ABS-0
INS-DEP-UTIL-0
GLU-PROD
INS-INDEP-0
INS-DEP-0
CHO-BAL-0
ENDO-BAL-0
CHO-1
BASAL-BAL-0
GUT-ABS-1
TOT-BAL-0
CHO-BAL-1

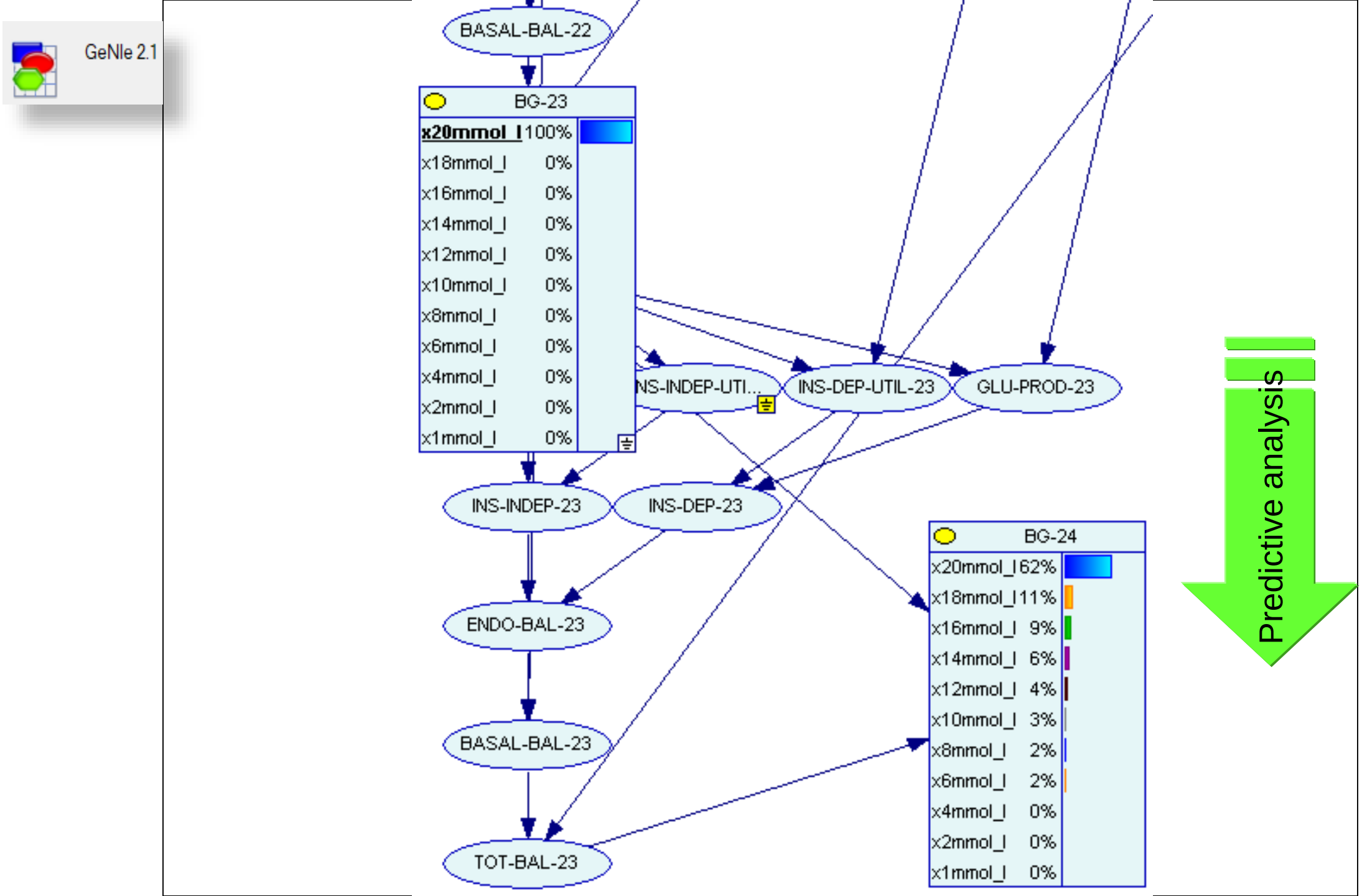

GeNIe 2.1

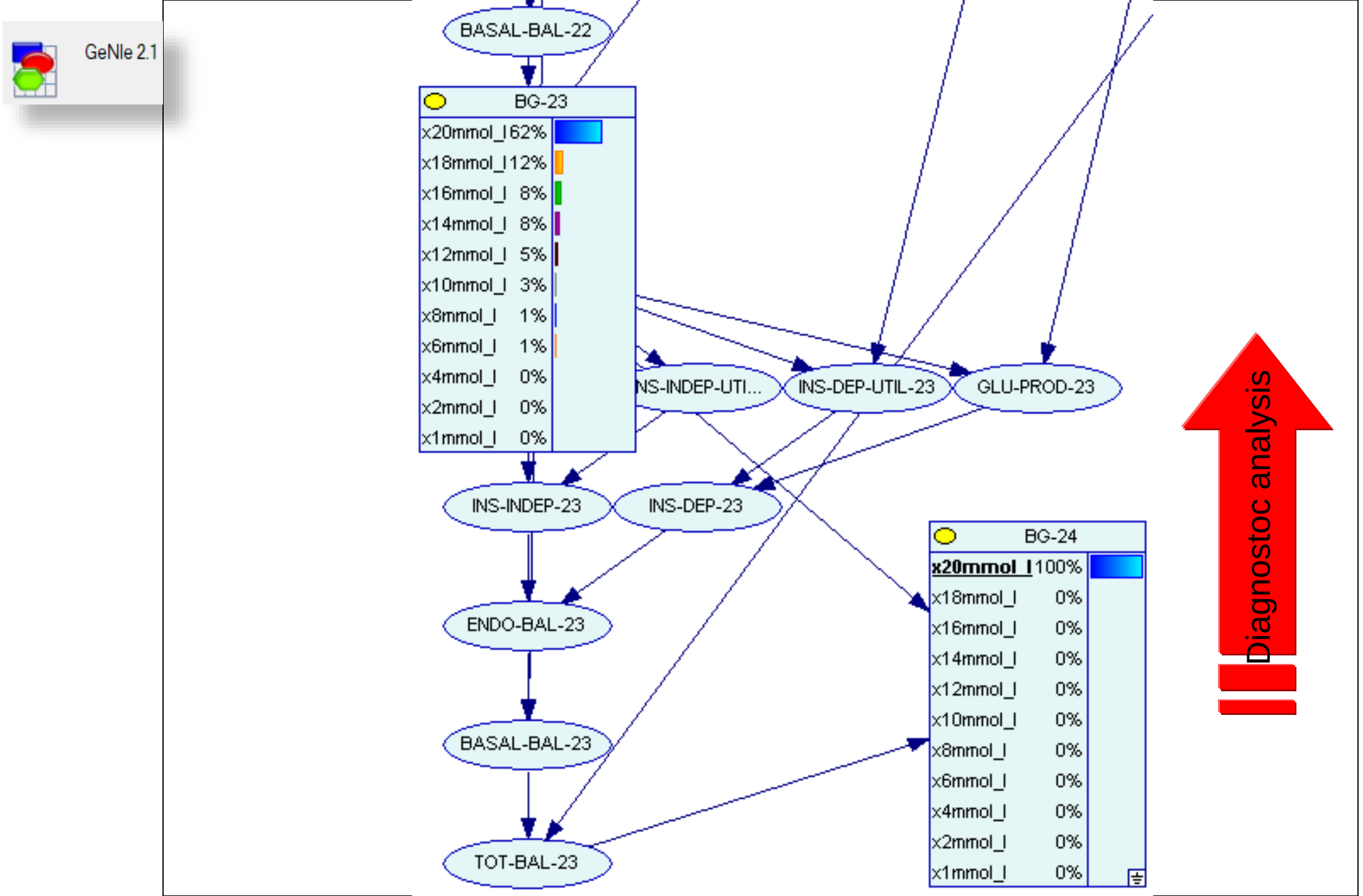

CHO-INIT
INS-SENS
MEAL-0
BG-0
INS-ABS-0
MET-IRR-0
MEAL-1
CHO-0
ACTIV INS-0
RENAL CL-0
INS-INDEP-UTI...
ACTIV INS-1
ACTIV INS-2
ACTIV INS-3
GUT-ABS-0
INS-DEP-UTIL-0
GLU-PROD
INS-INDEP-0
INS-DEP-0
CHO-BAL-0

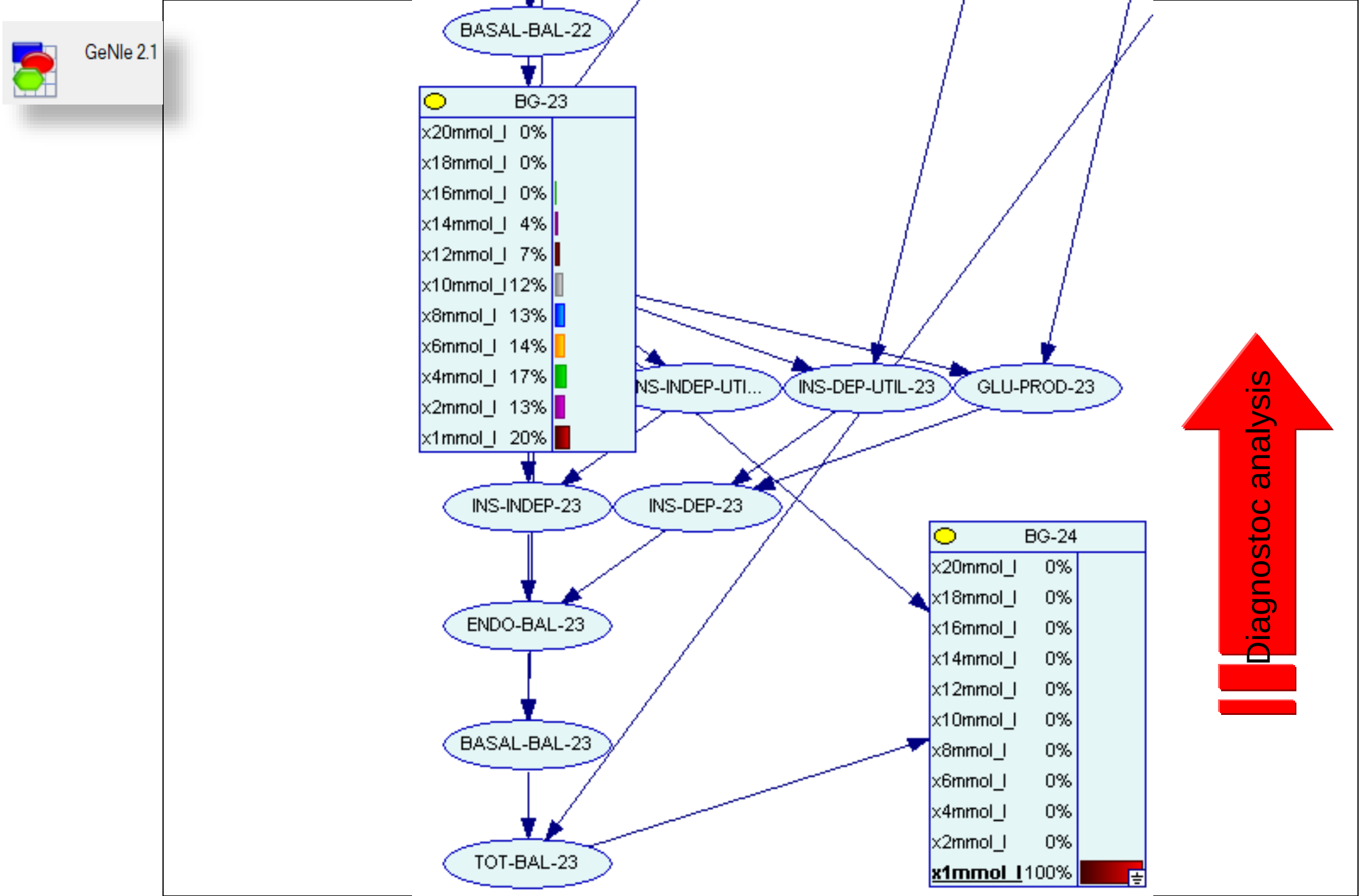
GeNIe 2.1

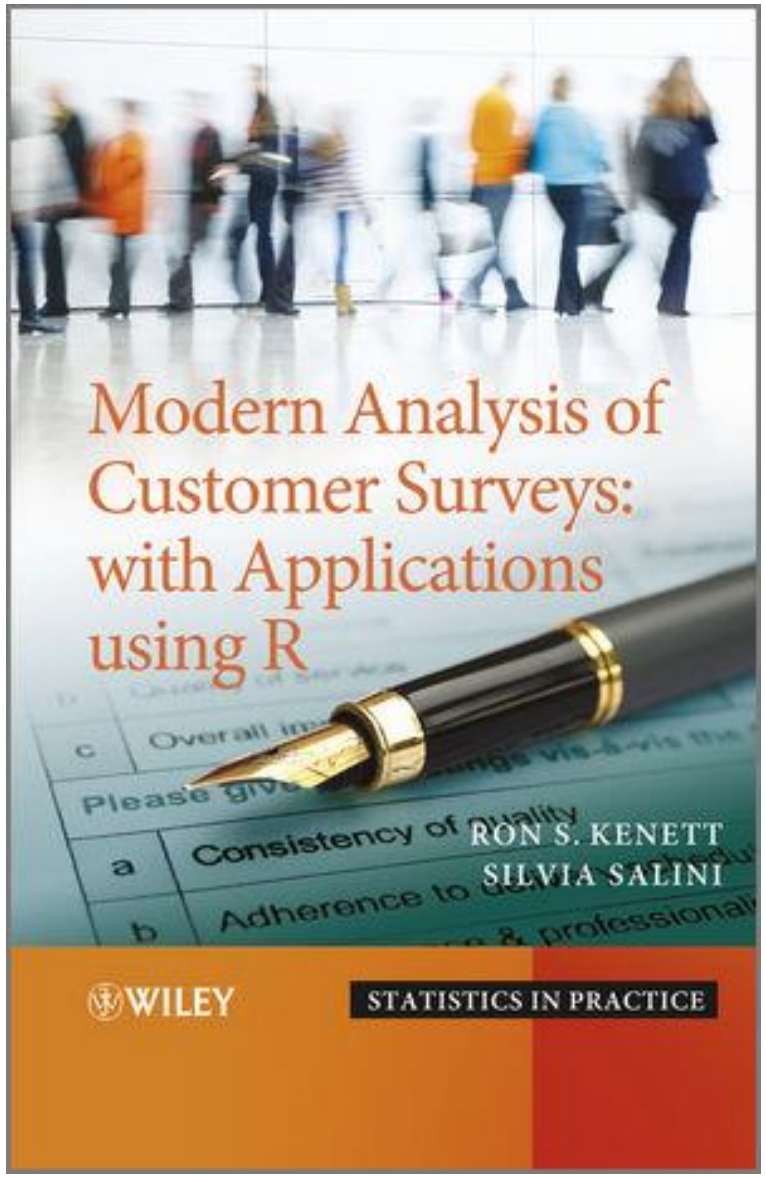

BASAL-BAL-22
BG-23
x20mmol_l 29%
x18mmol_l 9%
x16mmol_l 10%
x14mmol_l 10%
x12mmol_l 9%
x10mmol_l 8%
x8mmol_l 7%
x6mmol_l 6%
x4mmol_l 4%
x2mmol_l 3%
x1mmol_l 4%
NS-INDEP-UTI...
INS-DEP-UTIL-23
GLU-PROD-23
INS-INDEP-23
INS-DEP-23
ENDO-BAL-23
BASAL-BAL-23
TOT-BAL-23
BG-24
x20mmol_l 29%
x18mmol_l 9%
x16mmol_l 10%
x14mmol_l 10%
x12mmol_l 9%
x10mmol_l 8%
x8mmol_l 7%
x6mmol_l 6%
x4mmol_l 4%
x2mmol_l 3%
x1mmol_l 4%

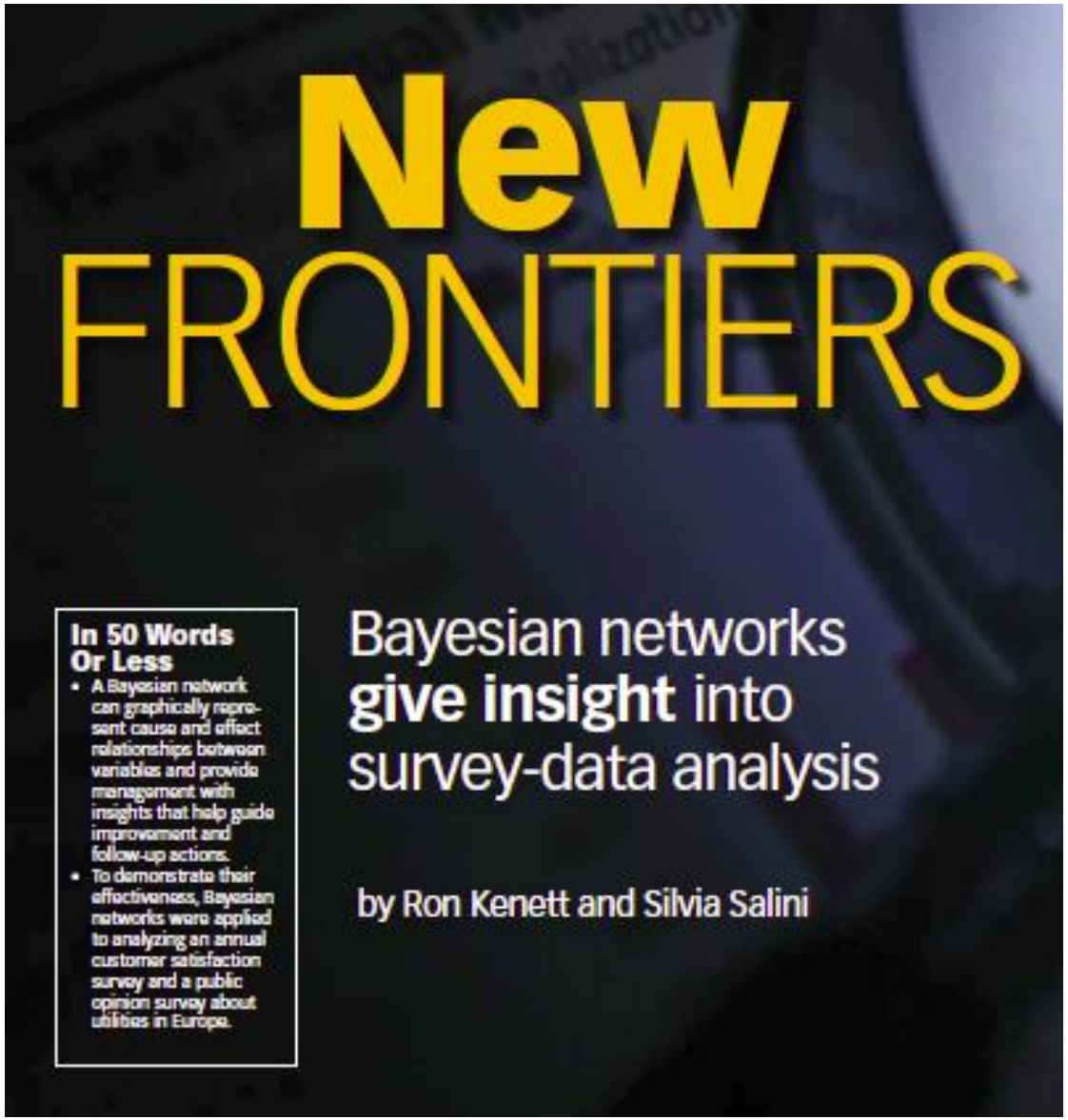

GeNIe 2.1
BASAL-BAL-22
BG-23
x20mmol_l 0%
x18mmol_l 0%
x16mmol_l 0%
x14mmol_l 0%
x12mmol_l 0%
x10mmol_l 0%
x8mmol_l 0%
x6mmol_l 0%
x4mmol_l 0%
x2mmol_l 0%
x1mmol_l 100%
NS-INDEP-UTI...
INS-DEP-UTIL-23
GLU-PROD-23
INS-INDEP-23
INS-DEP-23
ENDO-BAL-23
BASAL-BAL-23
TOT-BAL-23
BG-24
x20mmol_l 0%
x18mmol_l 0%
x16mmol_l 3%
x14mmol_l 6%
x12mmol_l 8%
x10mmol_l 12%
x8mmol_l 15%
x6mmol_l 16%
x4mmol_l 14%
x2mmol_l 5%
x1mmol_l 20%
Predictive analysis


GeNIe 2.1
BASAL-BAL-22
BG-23
x20mmol_l 100%
x18mmol_l 0%
x16mmol_l 0%
x14mmol_l 0%
x12mmol_l 0%
x10mmol_l 0%
x8mmol_l 0%
x6mmol_l 0%
x4mmol_l 0%
x2mmol_l 0%
x1mmol_l 0%
NS-INDEP-UTI...
INS-DEP-UTIL-23
GLU-PROD-23
INS-INDEP-23
INS-DEP-23
ENDO-BAL-23
BASAL-BAL-23
TOT-BAL-23
BG-24
x20mmol_l 62%
x18mmol_l 11%
x16mmol_l 9%
x14mmol_l 6%
x12mmol_l 4%
x10mmol_l 3%
x8mmol_l 2%
x6mmol_l 2%
x4mmol_l 0%
x2mmol_l 0%
x1mmol_l 0%
Predictive analysis

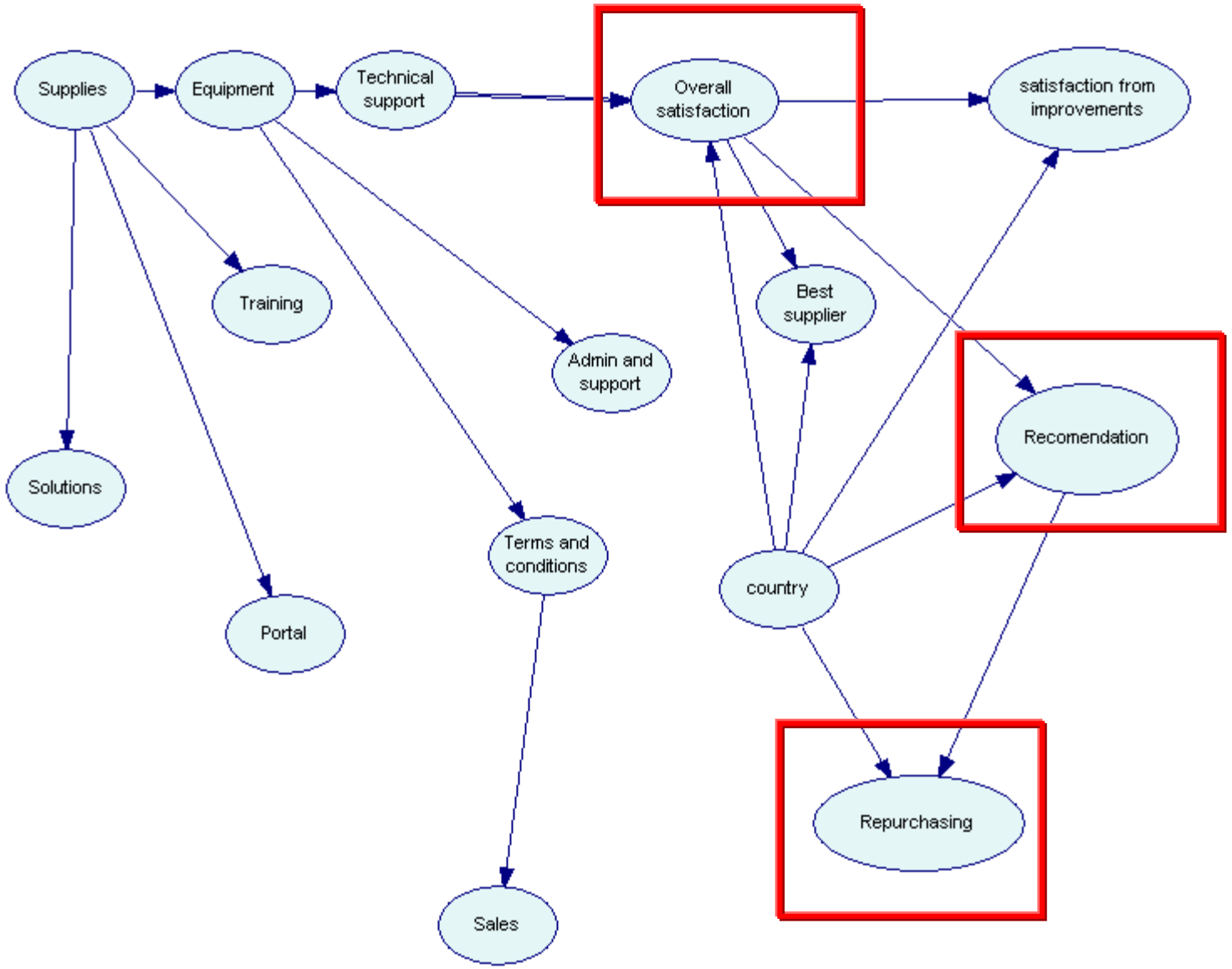

GeNIe 2.1
BASAL-BAL-22
BG-23
x20mmol_l 62%
x18mmol_l 12%
x16mmol_l 8%
x14mmol_l 8%
x12mmol_l 5%
x10mmol_l 3%
x8mmol_l 1%
x6mmol_l 1%
x4mmol_l 0%
x2mmol_l 0%
x1mmol_l 0%
NS-INDEP-UTI...
INS-DEP-UTIL-23
GLU-PROD-23
INS-INDEP-23
INS-DEP-23
ENDO-BAL-23
BASAL-BAL-23
TOT-BAL-23
BG-24
x20mmol_l 100%
x18mmol_l 0%
x16mmol_l 0%
x14mmol_l 0%
x12mmol_l 0%
x10mmol_l 0%
x8mmol_l 0%
x6mmol_l 0%
x4mmol_l 0%
x2mmol_l 0%
x1mmol_l 0%
Diagnostoc analysis

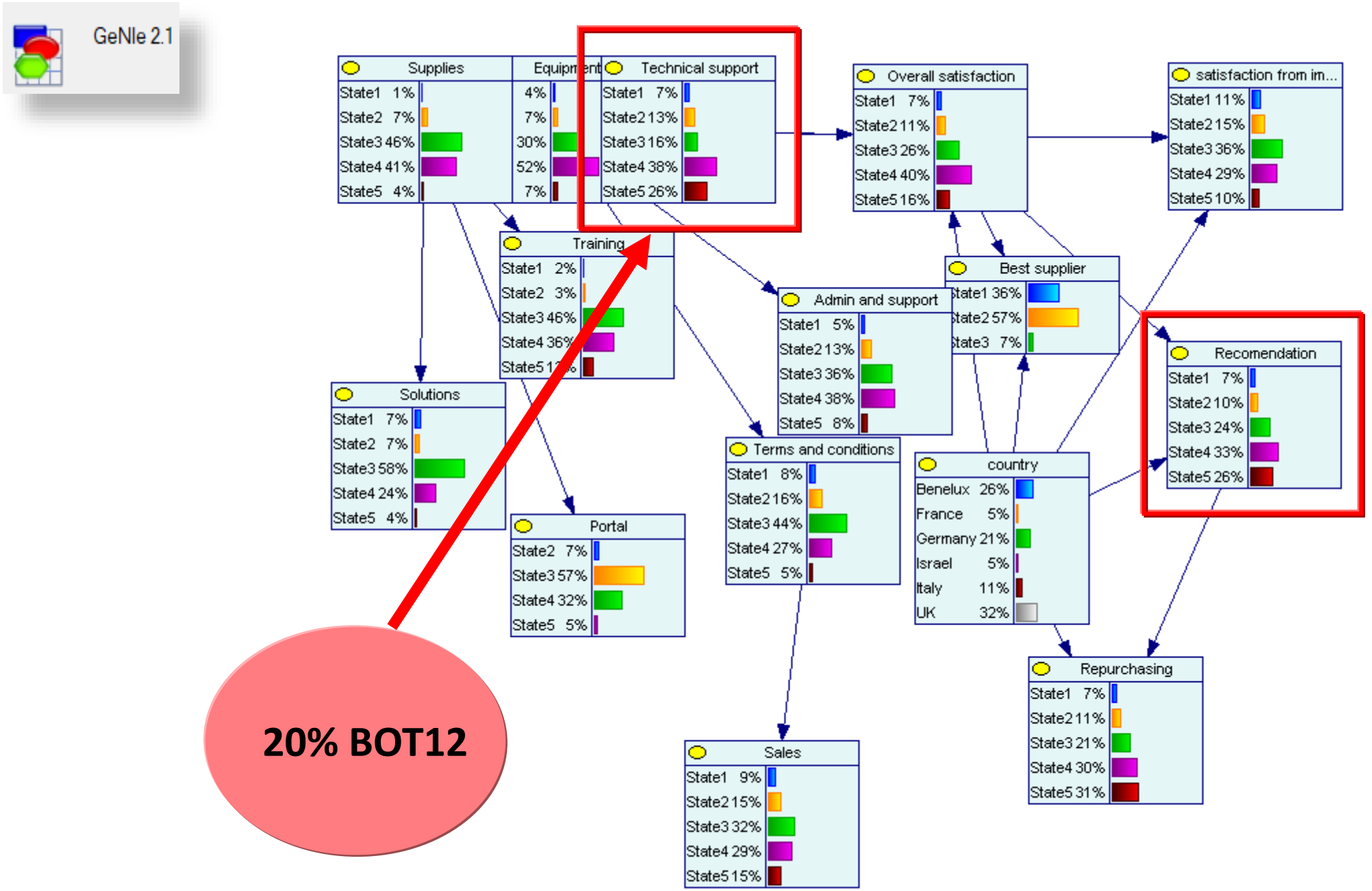

GeNIe 2.1
BASAL-BAL-22
BG-23
x20mmol_l 0%
x18mmol_l 0%
x16mmol_l 0%
x14mmol_l 4%
x12mmol_l 7%
x10mmol_l 12%
x8mmol_l 13%
x6mmol_l 14%
x4mmol_l 17%
x2mmol_l 13%
x1mmol_l 20%
NS-INDEP-UTI...
INS-DEP-UTIL-23
GLU-PROD-23
INS-INDEP-23
INS-DEP-23
ENDO-BAL-23
BASAL-BAL-23
TOT-BAL-23
BG-24
x20mmol_l 0%
x18mmol_l 0%
x16mmol_l 0%
x14mmol_l 0%
x12mmol_l 0%
x10mmol_l 0%
x8mmol_l 0%
x6mmol_l 0%
x4mmol_l 0%
x2mmol_l 0%
x1mmol_l 100%
Diagnostoc analysis

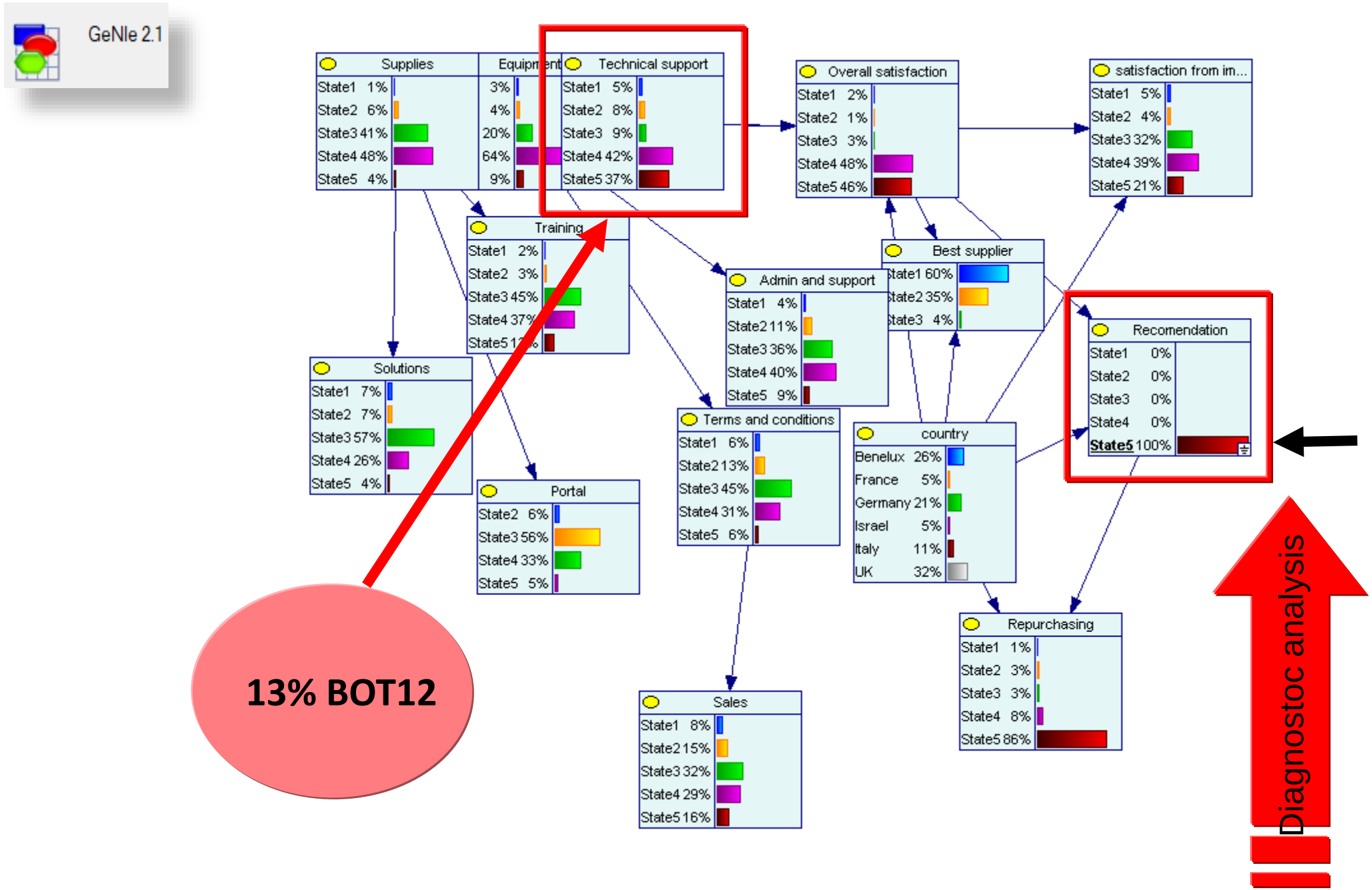

Modern Analysis of Customer Surveys: with Applications using R
RON S. KENETT
SILVIA SALINI
WILEY
STATISTICS IN PRACTICE

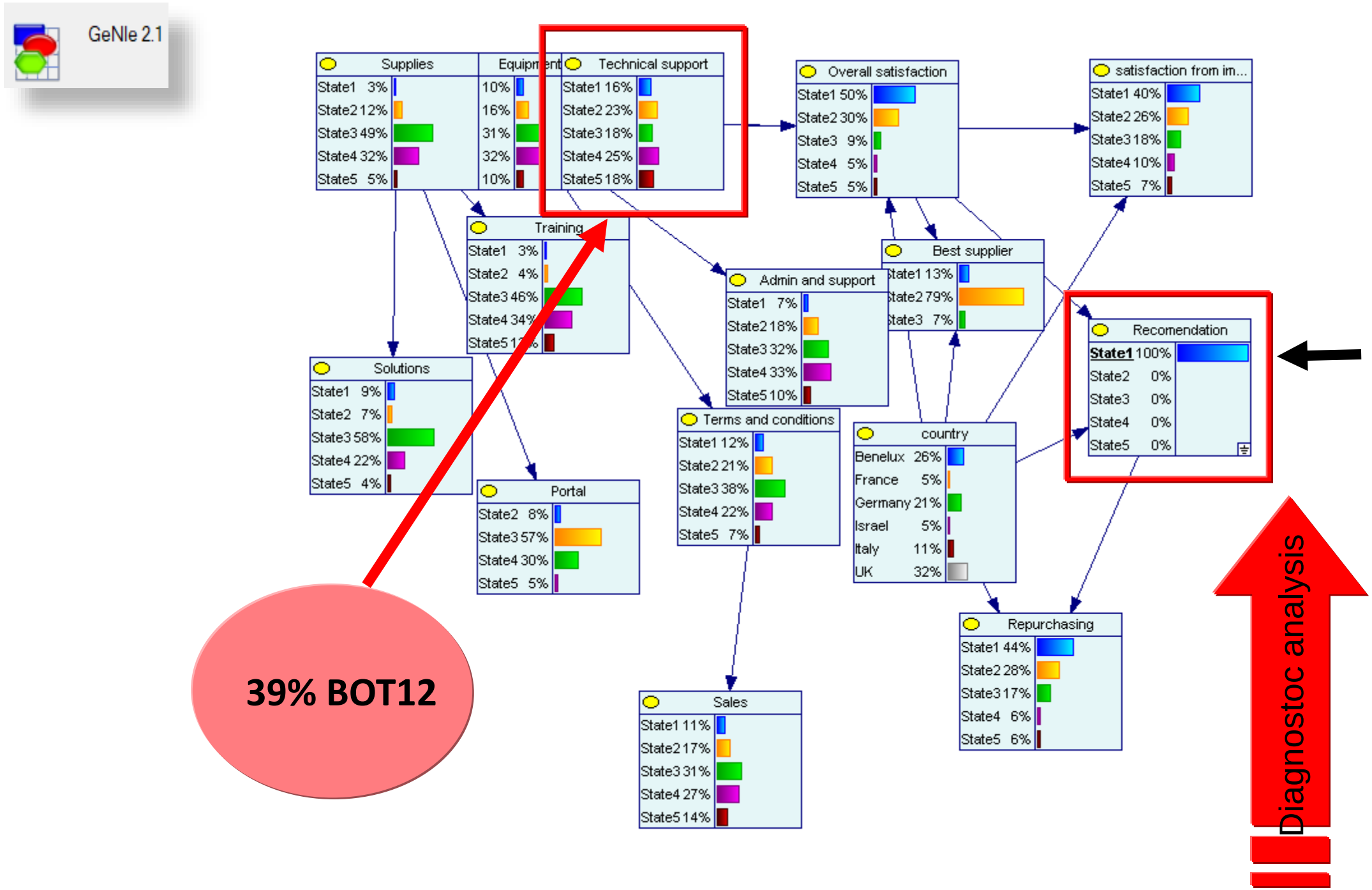

New FRONTIERS
In 50 Words Or Less
• A Bayesian network can graphically represent cause and effect relationships between variables and provide management with insights that help guide improvement and follow-up actions.
• To demonstrate their effectiveness, Bayesian networks were applied to analyzing an annual customer satisfaction survey and a public opinion survey about utilities in Europe.
Bayesian networks give insight into survey-data analysis
by Ron Kenett and Silvia Salini

GeNIe 2.1


Supplies
Equipment
Technical support
Overall satisfaction
satisfaction from improvements
Training
Admin and support
Best supplier
Recomendation
Solutions
Terms and conditions
country
Portal
Repurchasing
Sales


GeNIe 2.1
Supplies
State1 1%
State2 7%
State3 46%
State4 41%
State5 4%
Equipment
4%
7%
30%
52%
7%
Technical support
State1 7%
State2 13%
State3 16%
State4 38%
State5 26%
Overall satisfaction
State1 7%
State2 11%
State3 26%
State4 40%
State5 16%
satisfaction from im...
State1 11%
State2 15%
State3 36%
State4 29%
State5 10%
Training
State1 2%
State2 3%
State3 46%
State4 36%
Best supplier
State1 36%
State2 57%
State3 7%
Admin and support
State1 5%
State2 13%
State3 36%
State4 38%
State5 8%
Recomendation
State1 7%
State2 10%
State3 24%
State4 33%
State5 26%
Solutions
State1 7%
State2 7%
State3 58%
State4 24%
State5 4%
Terms and conditions
State1 8%
State2 16%
State3 44%
State4 27%
State5 5%
country
Benelux 26%
France 5%
Germany 21%
Israel 5%
Italy 11%
UK 32%
Portal
State2 7%
State3 57%
State4 32%
State5 5%
Repurchasing
State1 7%
State2 11%
State3 21%
State4 30%
State5 31%
Sales
State1 9%
State2 15%
State3 32%
State4 29%
State5 15%
20% BOT12


GeNIe 2.1
Supplies
State1 1%
State2 6%
State3 41%
State4 48%
State5 4%
Equipment
3%
4%
20%
64%
9%
Technical support
State1 5%
State2 8%
State3 9%
State4 42%
State5 37%
Overall satisfaction
State1 2%
State2 1%
State3 3%
State4 48%
State5 46%
satisfaction from im...
State1 5%
State2 4%
State3 32%
State4 39%
State5 21%
Training
State1 2%
State2 3%
State3 45%
State4 37%
Best supplier
State1 60%
State2 35%
State3 4%
Admin and support
State1 4%
State2 11%
State3 36%
State4 40%
State5 9%
Recomendation
State1 0%
State2 0%
State3 0%
State4 0%
State5 100%
Solutions
State1 7%
State2 7%
State3 57%
State4 26%
State5 4%
Terms and conditions
State1 6%
State2 13%
State3 45%
State4 31%
State5 6%
country
Benelux 26%
France 5%
Germany 21%
Israel 5%
Italy 11%
UK 32%
Portal
State2 6%
State3 56%
State4 33%
State5 5%
Repurchasing
State1 1%
State2 3%
State3 3%
State4 8%
State5 86%
Sales
State1 8%
State2 15%
State3 32%
State4 29%
State5 16%
13% BOT12
Diagnostoc analysis


GeNIe 2.1
Supplies
State1 3%
State2 12%
State3 49%
State4 32%
State5 5%
10%
16%
31%
32%
10%
Technical support
State1 16%
State2 23%
State3 18%
State4 25%
State5 18%
Overall satisfaction
State1 50%
State2 30%
State3 9%
State4 5%
State5 5%
satisfaction from im...
State1 40%
State2 26%
State3 18%
State4 10%
State5 7%
Training
State1 3%
State2 4%
State3 46%
State4 34%
Best supplier
State1 13%
State2 79%
State3 7%
Admin and support
State1 7%
State2 18%
State3 32%
State4 33%
State5 10%
Recomendation
State1 100%
State2 0%
State3 0%
State4 0%
State5 0%
Solutions
State1 9%
State2 7%
State3 58%
State4 22%
State5 4%
Terms and conditions
State1 12%
State2 21%
State3 38%
State4 22%
State5 7%
country
Benelux 26%
France 5%
Germany 21%
Israel 5%
Italy 11%
UK 32%
Portal
State2 8%
State3 57%
State4 30%
State5 5%
Repurchasing
State1 44%
State2 28%
State3 17%
State4 6%
State5 6%
Sales
State1 11%
State2 17%
State3 31%
State4 27%
State5 14%
39% BOT12
Diagnostoc analysis

But: Correlation is not causation…

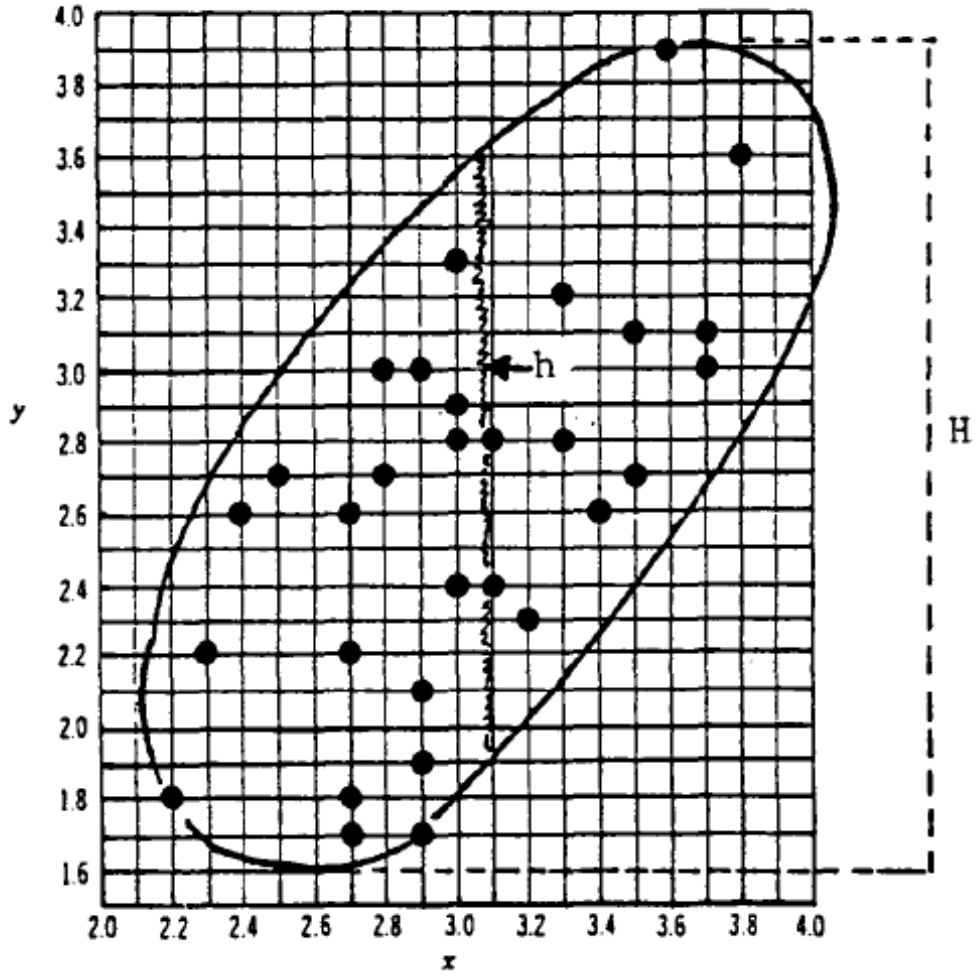
2. THE BALLOON IDEA

The basic idea consists of surrounding the sample plot with a kind of "birthday balloon" that is in fact an ellipse. But let us apply this method to an example taken from a well-known volume by Hoel (1971). The sample plot from page 189 of Hoel's book is reproduced in Figure 1.

First, we draw the balloon so as to surround all or most of the points and to fit the plot. Second, we measure the vertical height of the balloon at its center, h, and its vertical height at the extremes, H. Then we compute the formula

$$F = \sqrt{1 - \left(\frac{h}{H}\right)^2}.$$

If the points inside the balloon are "well distributed," then the result of the computation usually gives a fairly good idea of the value of Pearson's correlation coefficient.



Chatillon, G. (1984) The Balloon Rules for a Rough Estimate of the Correlation Coefficient, The American Statistician, 38(1), 58-60.

Correlation is not causation…

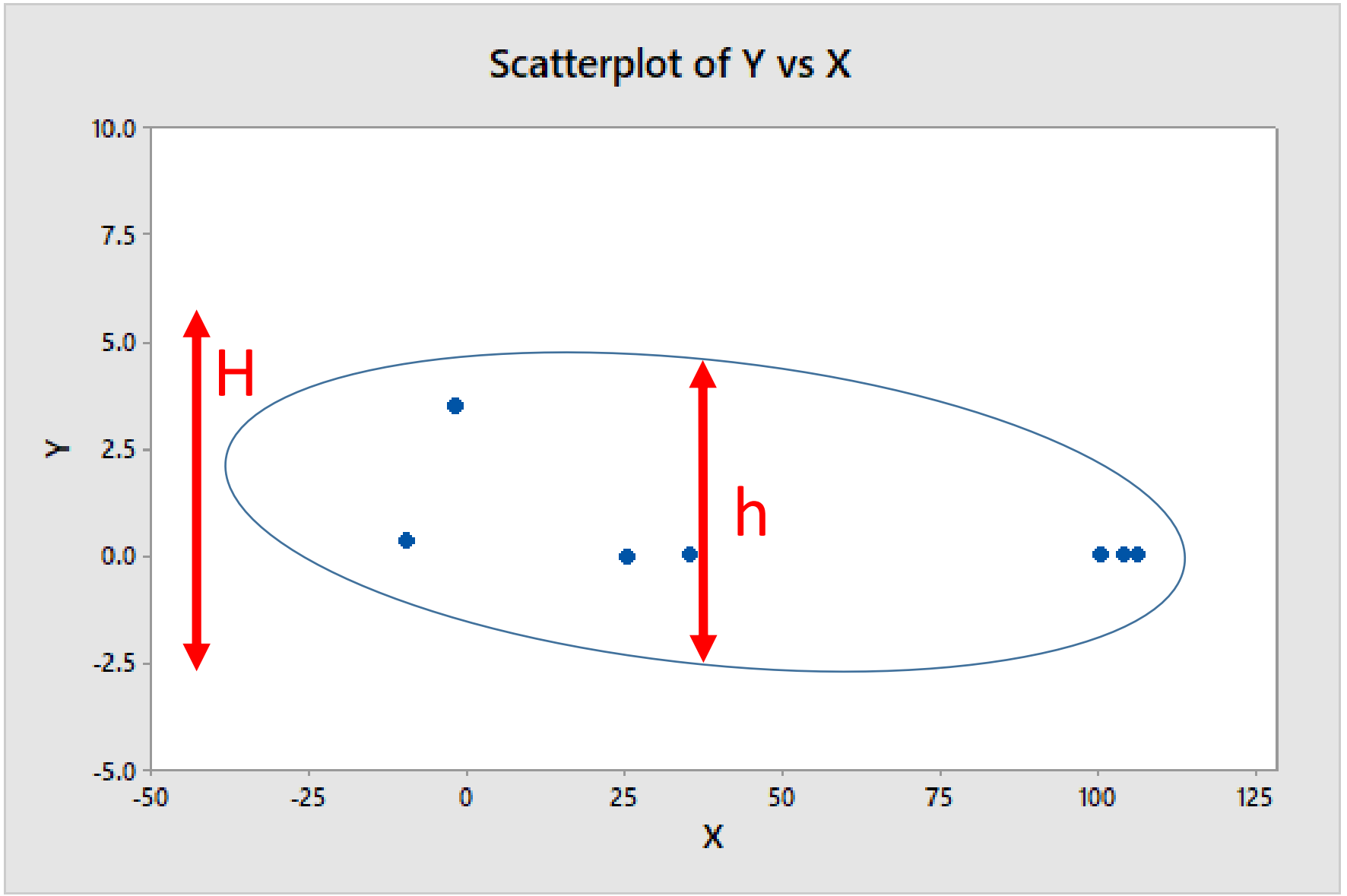
Correlations

Pearson correlation	-0.501
P-value	0.252

$$F = \sqrt{1 - \left(\frac{h}{H}\right)^2}.$$

$$Sqr\{[1 - \left(\frac{7.5}{8.75}\right)^2\}$$

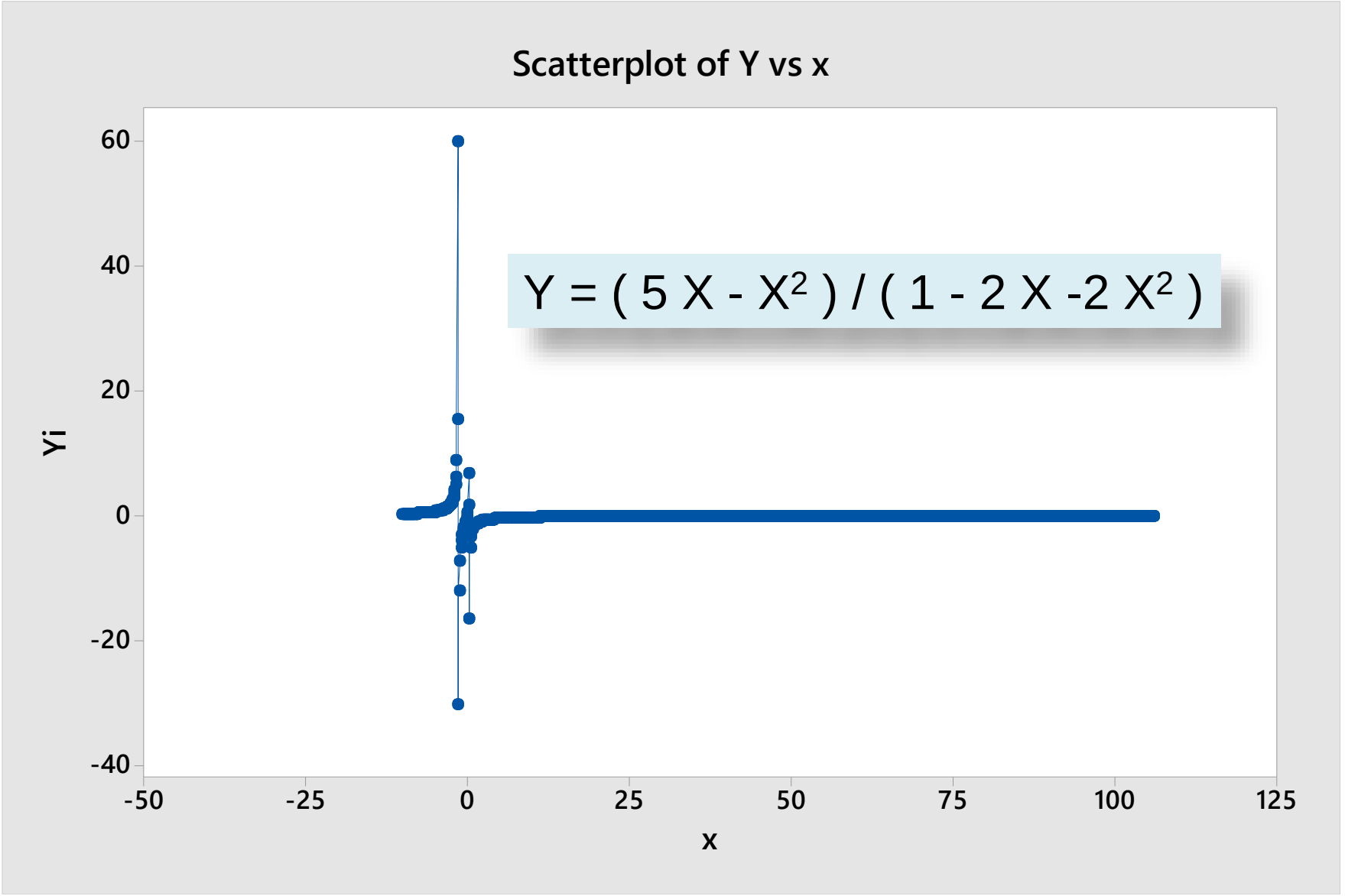
= 0.5



Correlation is not causation…

Correlations

Pearson correlation	-0.501
P-value	0.252

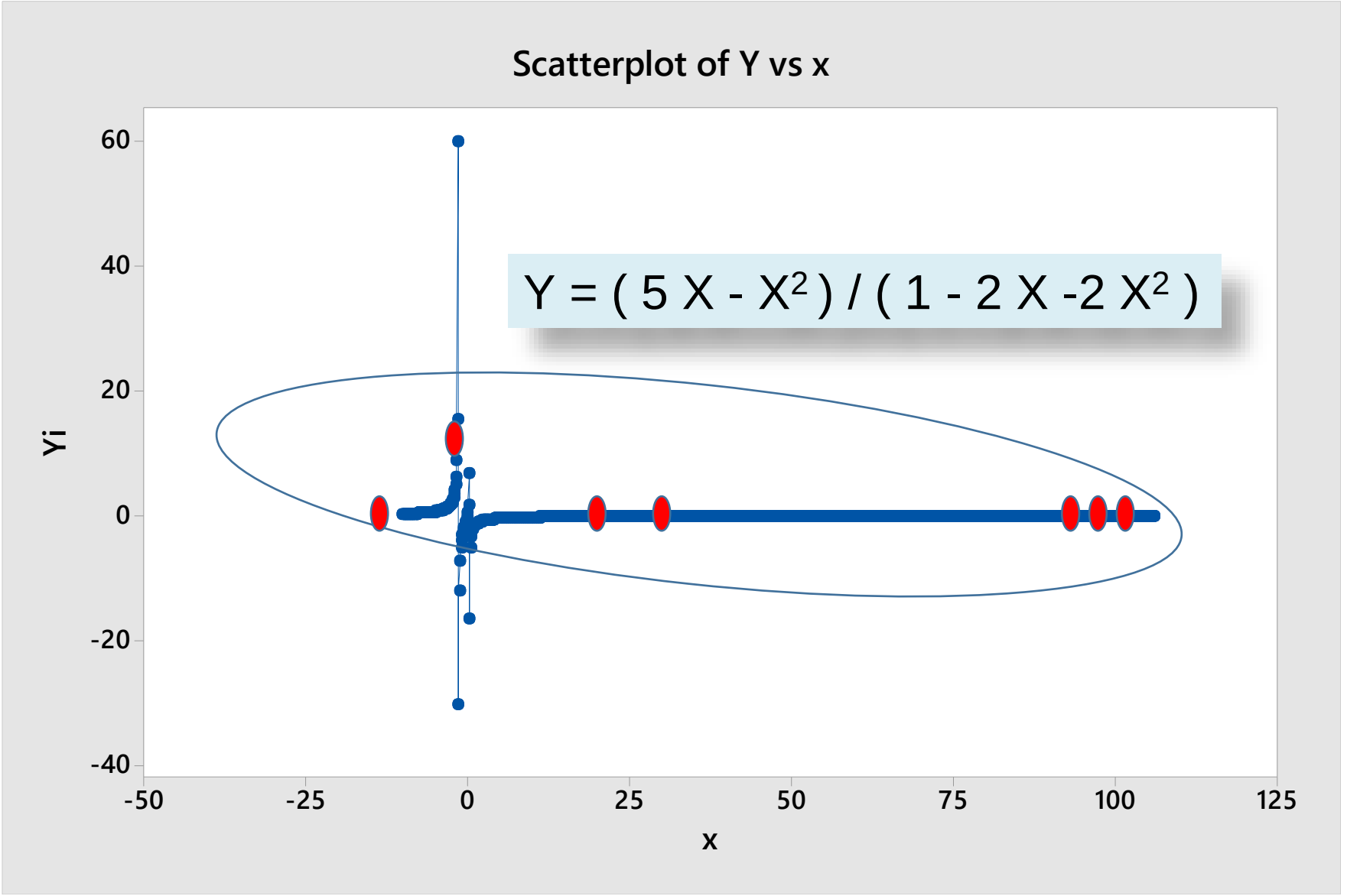


Correlation is not causation…

Correlations

Pearson correlation	-0.501
P-value	0.252

No correlation does not imply no causation

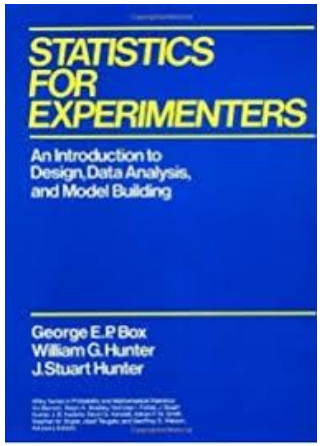


Correlation is not causation…

The population of Oldenburg in Germany and the number of observed storks in 1930-1936*

year	1930	1931	1932	1933	1934	1935	1936
Population in thousands	50	52	64	67	69	73	76
Number of storks	130	150	175	190	240	245	250

~~No~~ correlation does not imply ~~no~~ causation

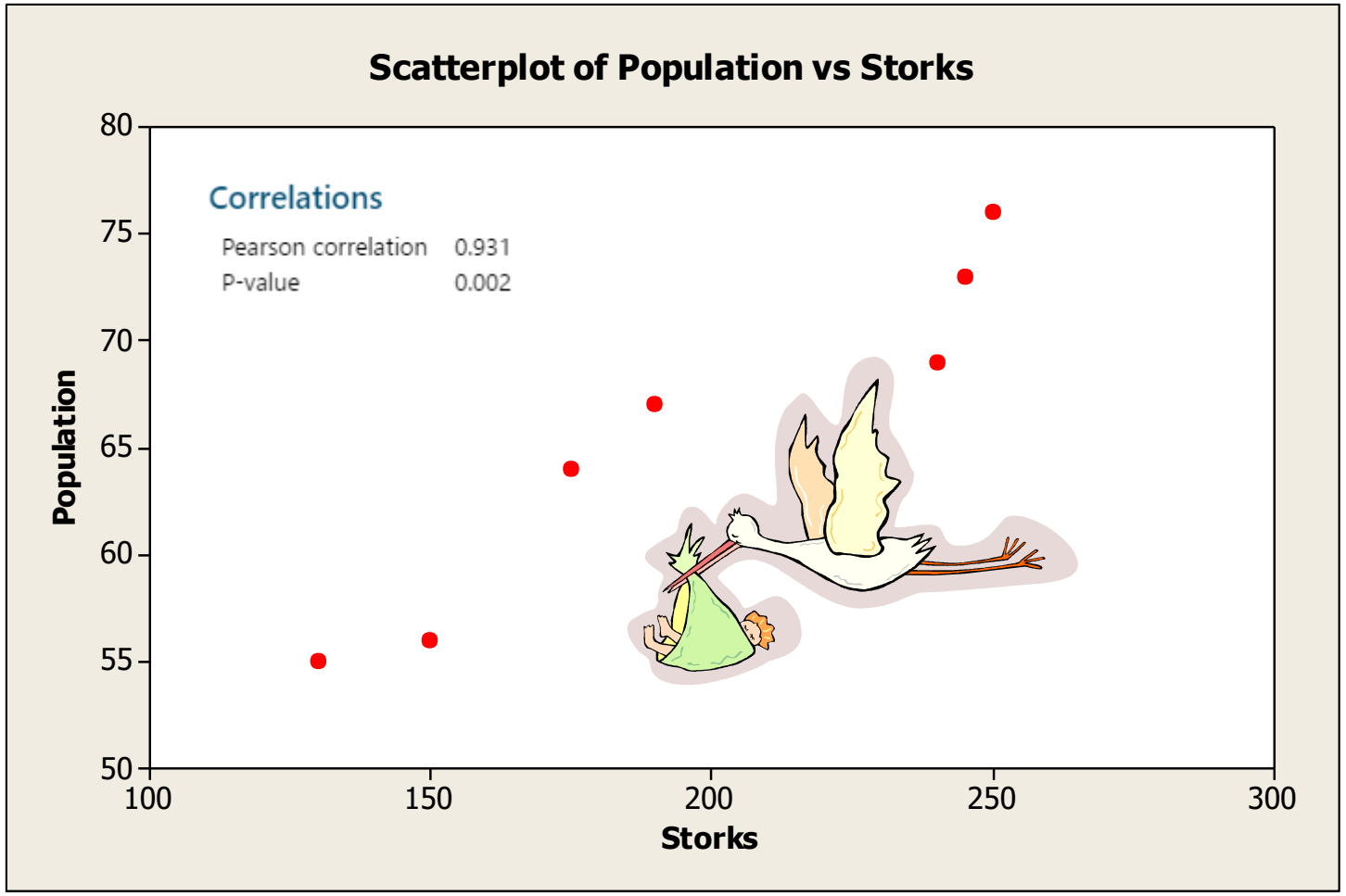


* Box, Hunter and Hunter, *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*, J. Wiley, 1978

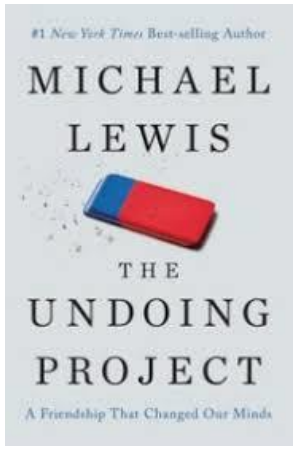
Spurious correlation

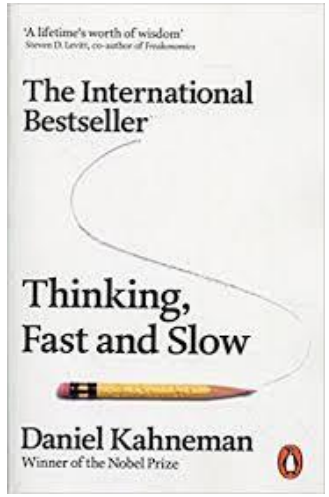
Time is a confounding variable

1. Base rate neglect,
2. Overconfidence,
3. Anchoring,
4. Representativeness,
5. Availability,
6. Regression towards the mean,
7. **Spurious correlation,**
8. Framing.

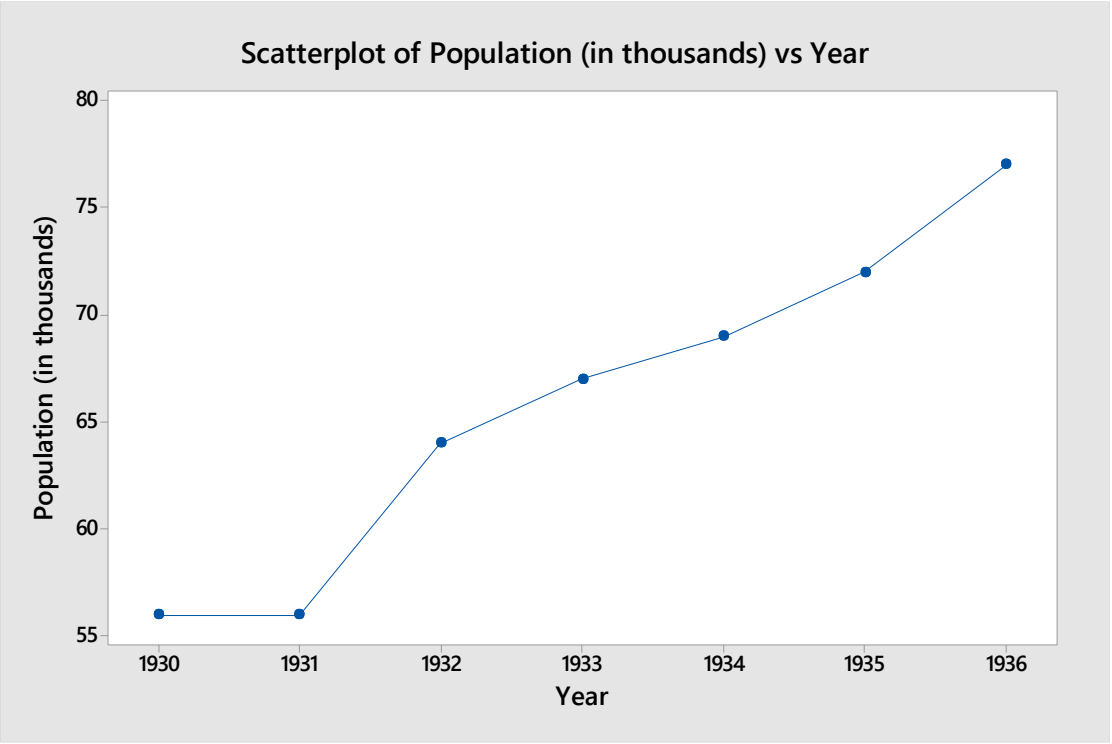


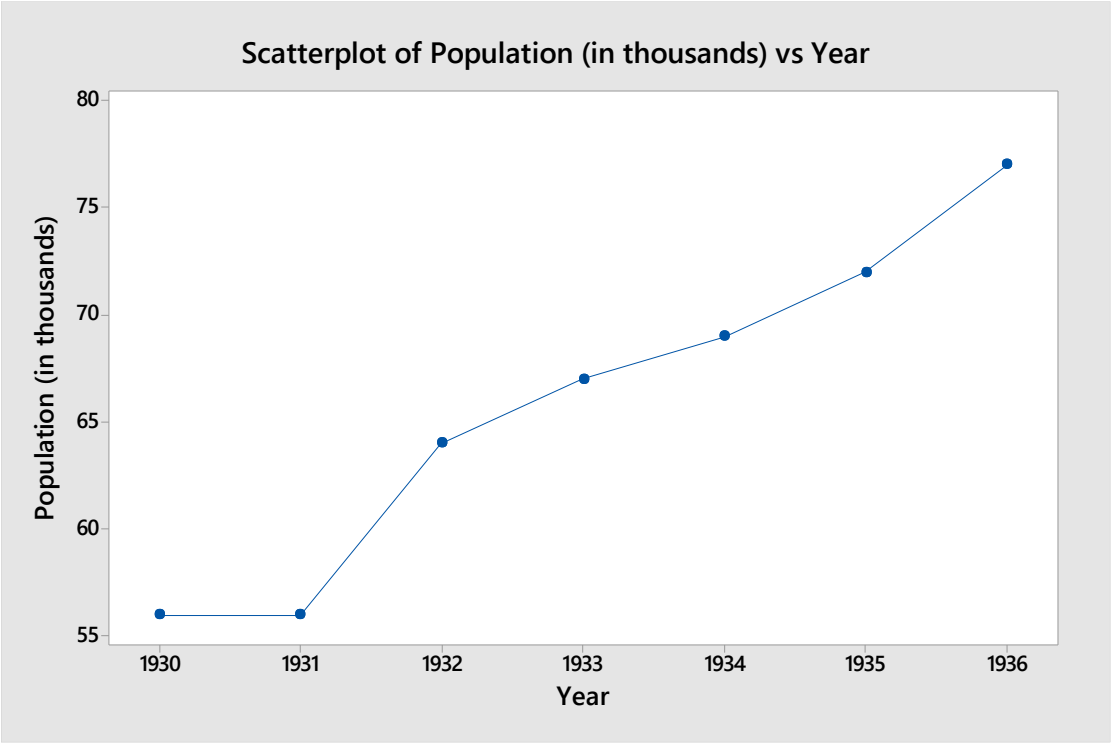
Correlation does not imply causation



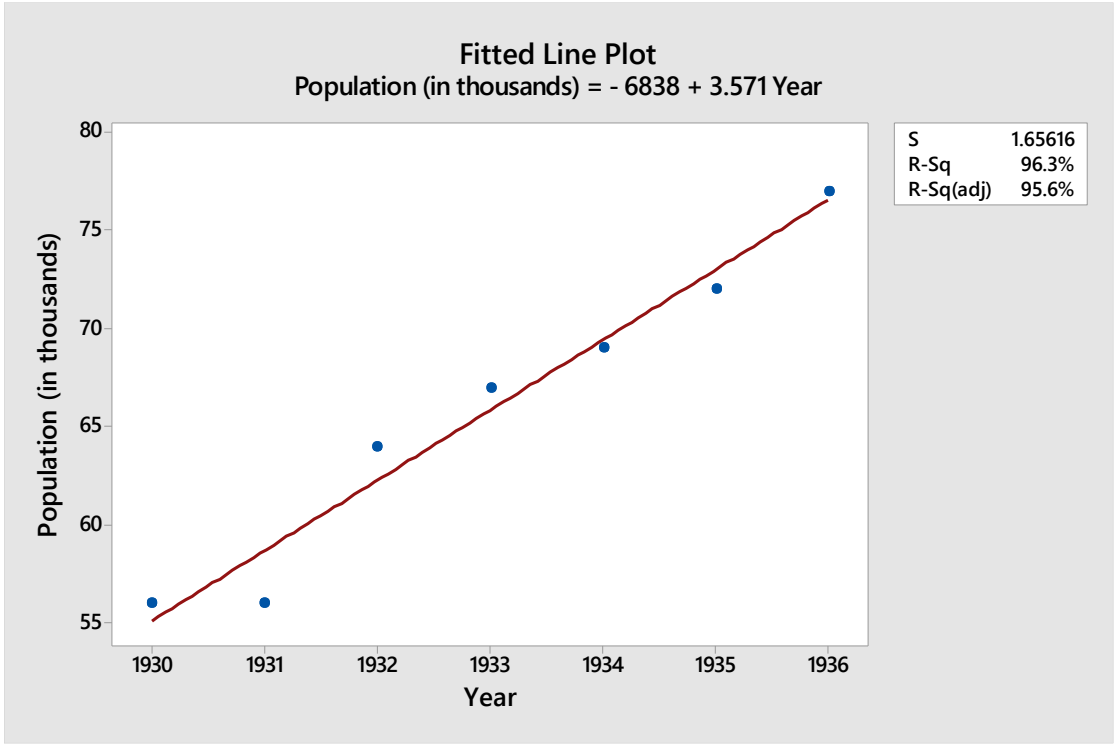


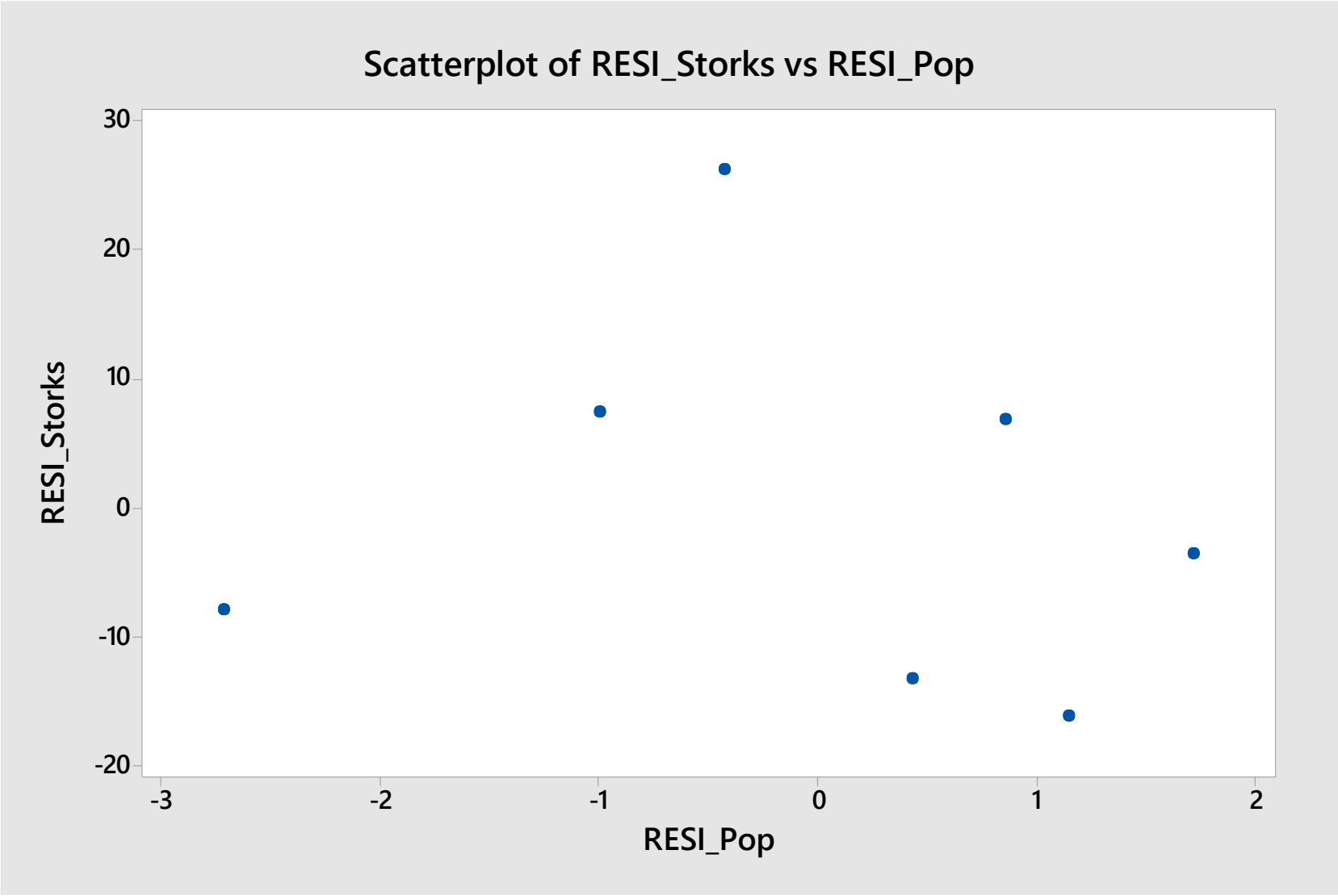
Spurious correlation

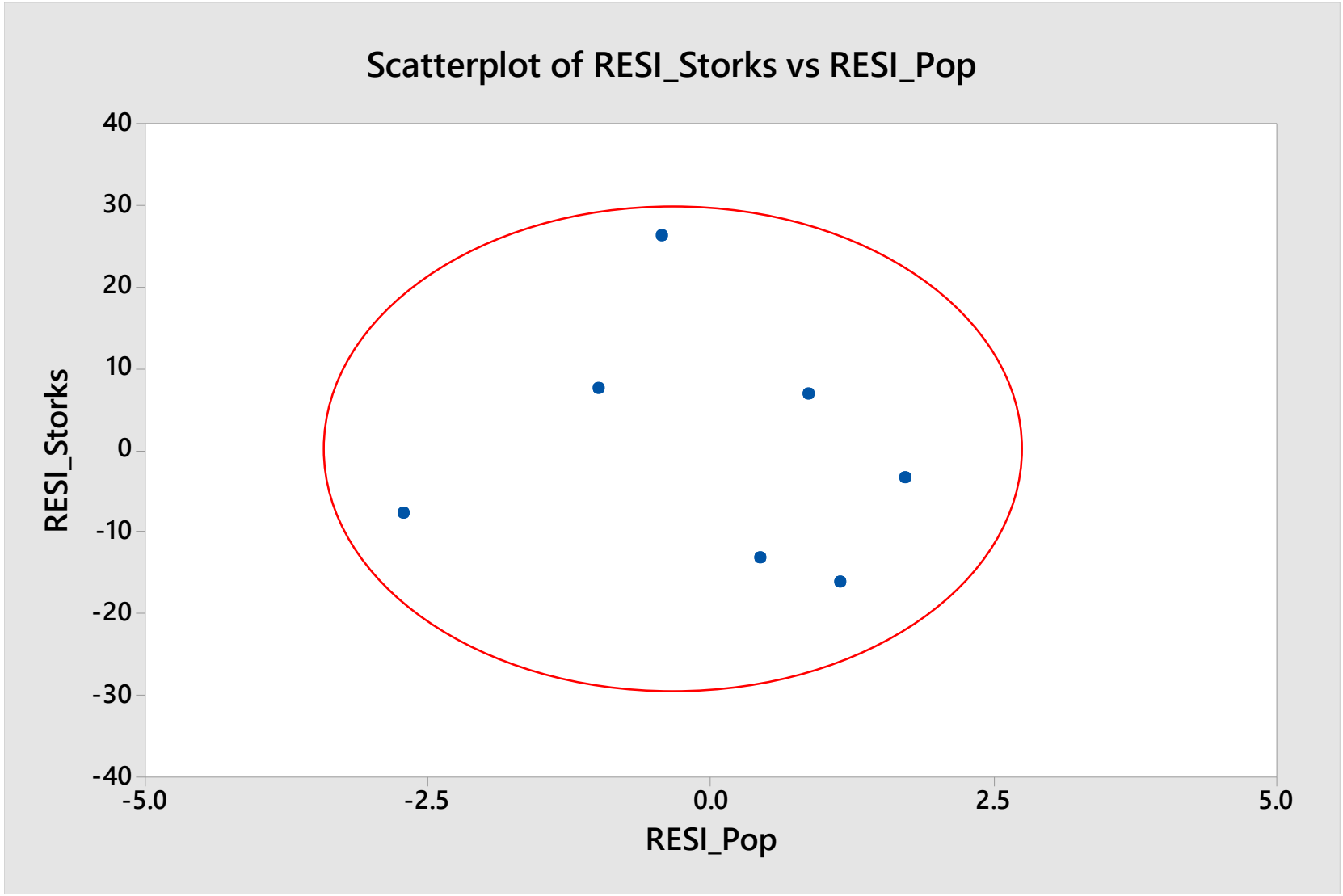




Spurious correlation




Scatterplot of RESI_Storks vs RESI_Pop
RESI_Storks
30
20
10
0
-10
-20
-3
-2
-1
0
1
2
RESI_Pop



Correlations

Pearson correlation	0.931
P-value	0.002

Correlations

Pearson correlation	-0.163
P-value	0.727

WebPower - Correlation Coefficie

https://webpower.psychstat.org/models/cor01/

Email link | RonLinks | LinkedIn | Facebook | News Ynet | בנק הפועלים – לקוח | Sci-Hub: removing ba | צבירת נקודות | Other bookmarks

Navigation

- WebPower
- Ask Power
- My Analyses
- New Analysis
- Tools
- Manual
- References
- What's new
- Workshop
- FAQ

Correlation Coefficient

Parameters (Help)	
Sample size	50
Correlation	0.3
# of vars partialed out	0
Significance level	0.05
Power	
H1	Two sided
Power curve	Show power curve
Note	Power for correlation

Calculate

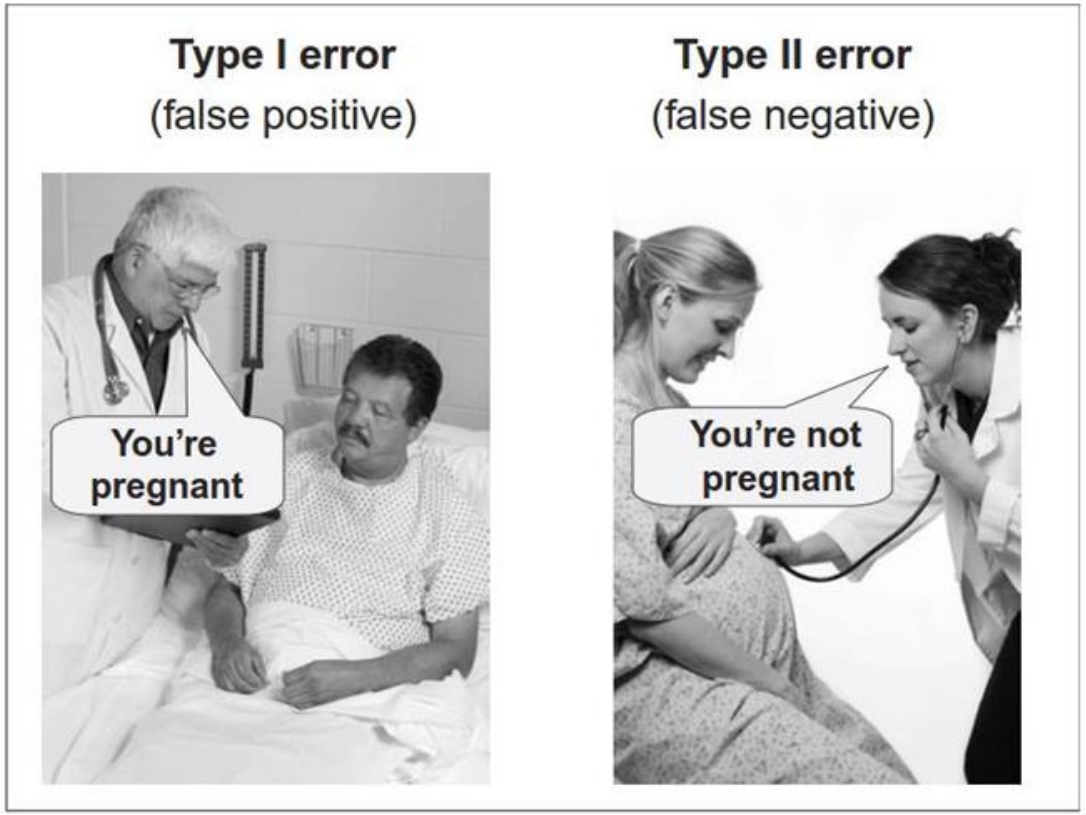
Output

```
Power for correlation

     n   r alpha  power
    50 0.3  0.05 0.5729

URL: http://psychstat.org/correlation
```

Power Curve

Download PDF figure

How many observations are needed to determine significant correlation?



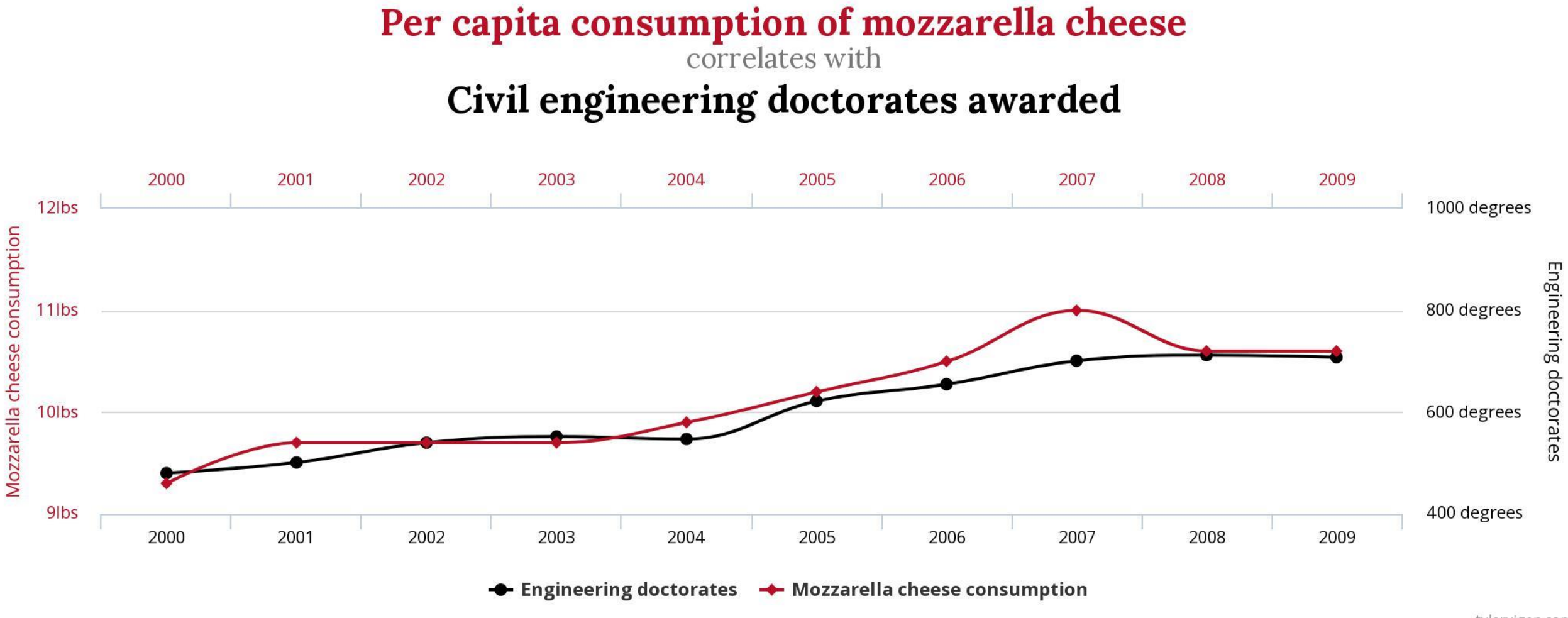
https://webpower.psychstat.org/models/cor01/
http://www.divms.uiowa.edu/~rlenth/Power/
https://cran.r-project.org/web/packages/WebPower/index.html

webpower_manual....pdf

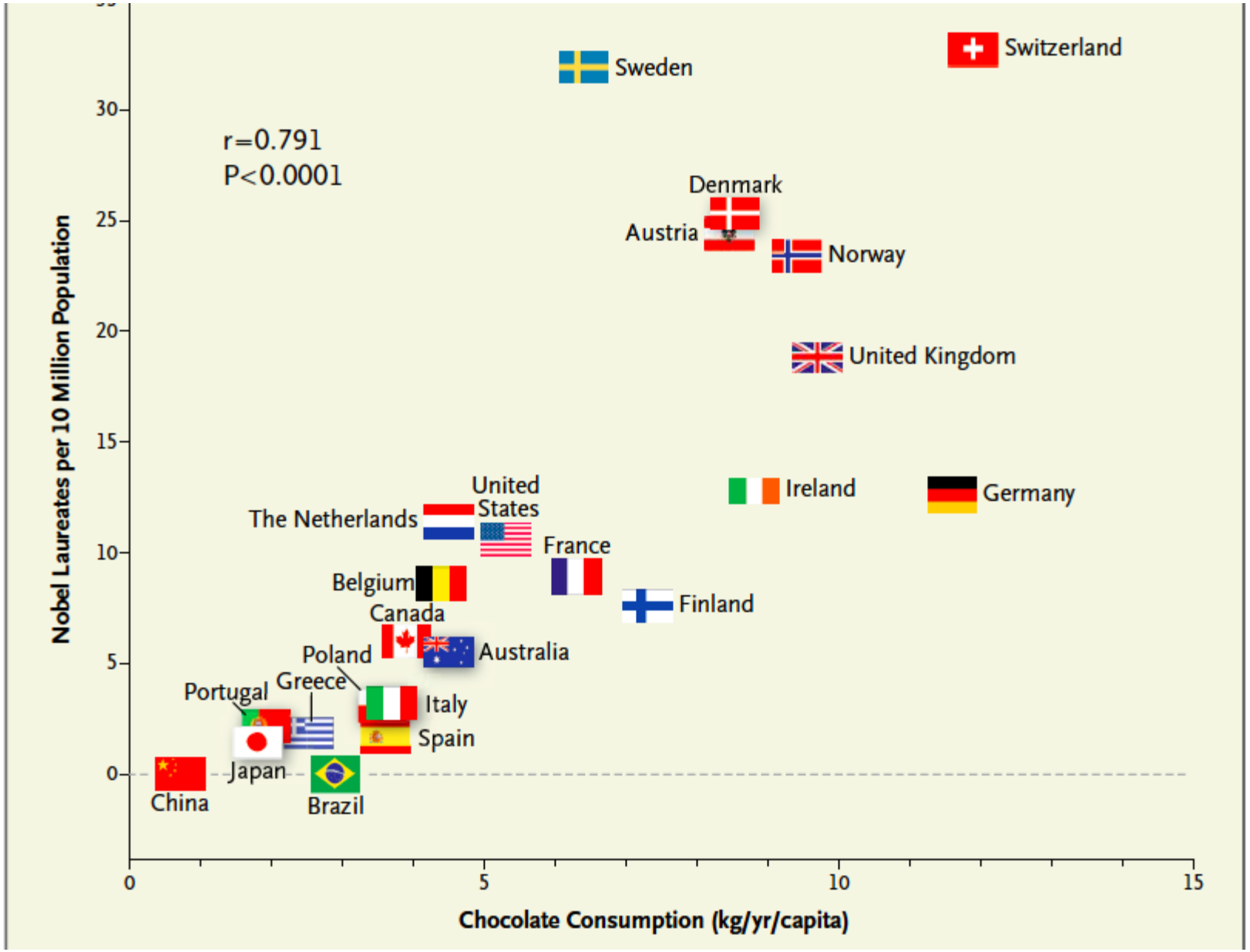
Show all

On spurious correlations

http://www.tylervigen.com/spurious-correlations



On spurious correlations



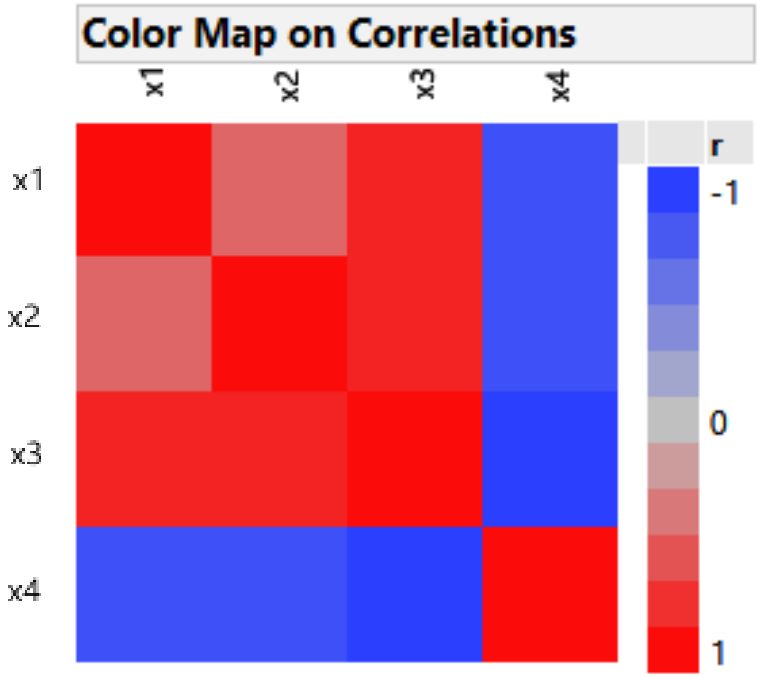
Causality effects?

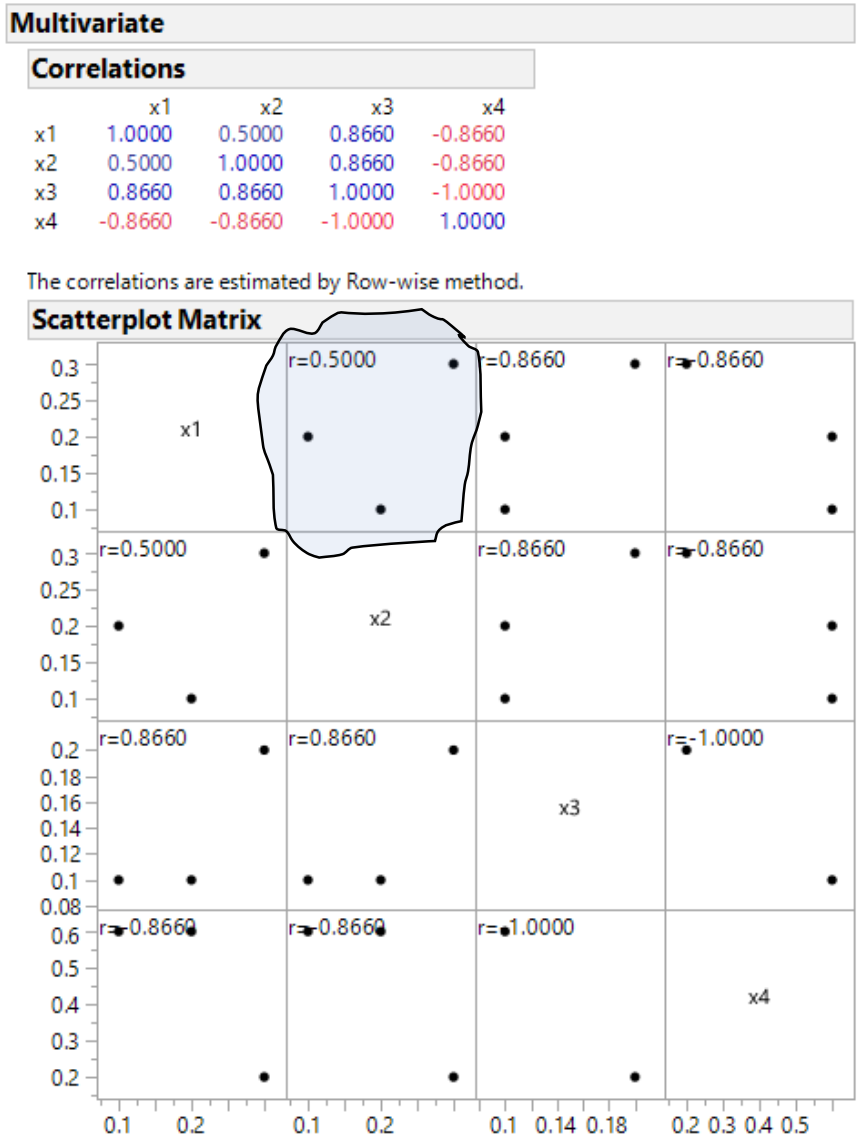
$$S^D=\{\mathbf{x}\in R_+^D:\ x_1+x_2+\cdots+x_D=k\}$$

Compositional data

On spurious correlations

x1	x2	x3	x4
0.1	0.2	0.1	0.6
0.2	0.1	0.1	0.6
0.3	0.3	0.2	0.2





subcompositional coherence: Using full composition or using subcomposition, one should make the same inference about relations within the common parts.

The correlation coefficient is not subcompositionaly coherent.

Pearson, K. (1897) Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London,* LX, 489-502.

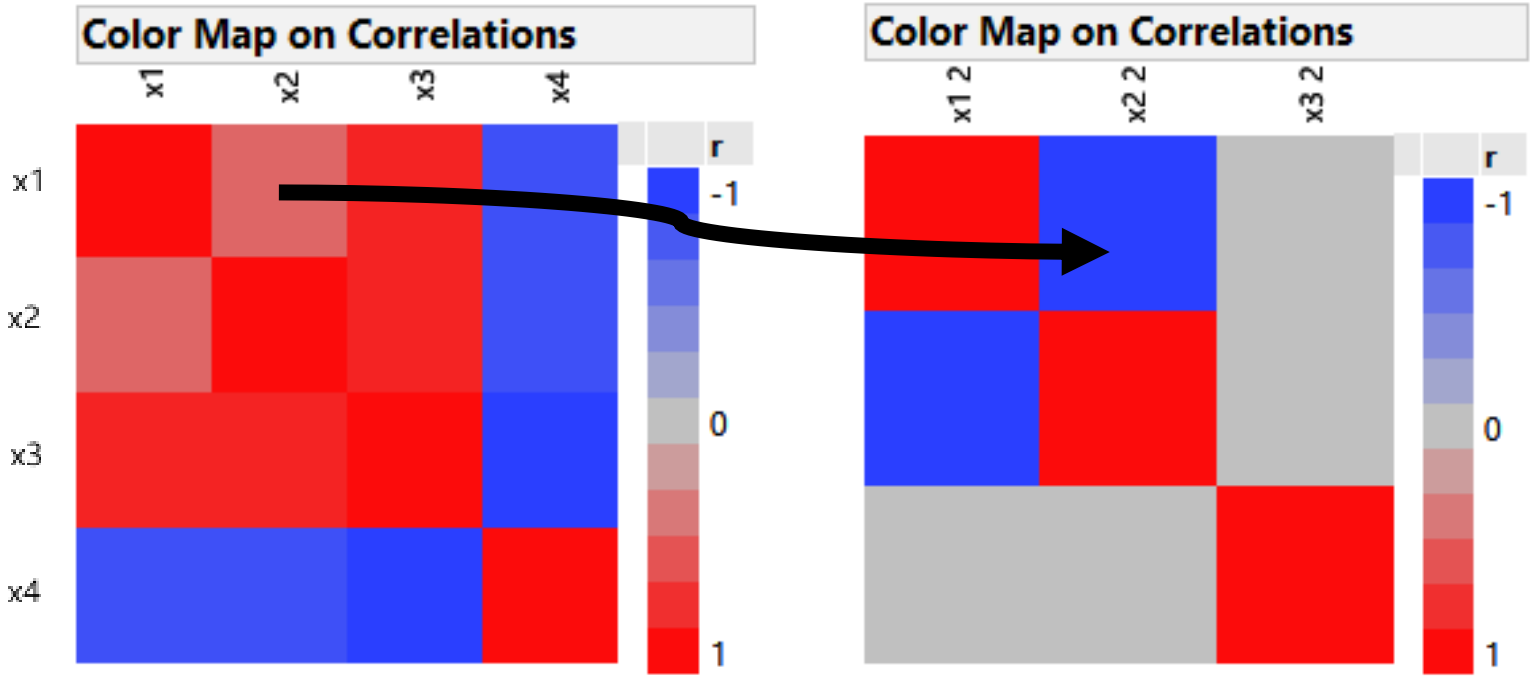
$$S^D=\{\mathbf{x}\in R_+^D:\ x_1+x_2+\cdots+x_D=k\}$$

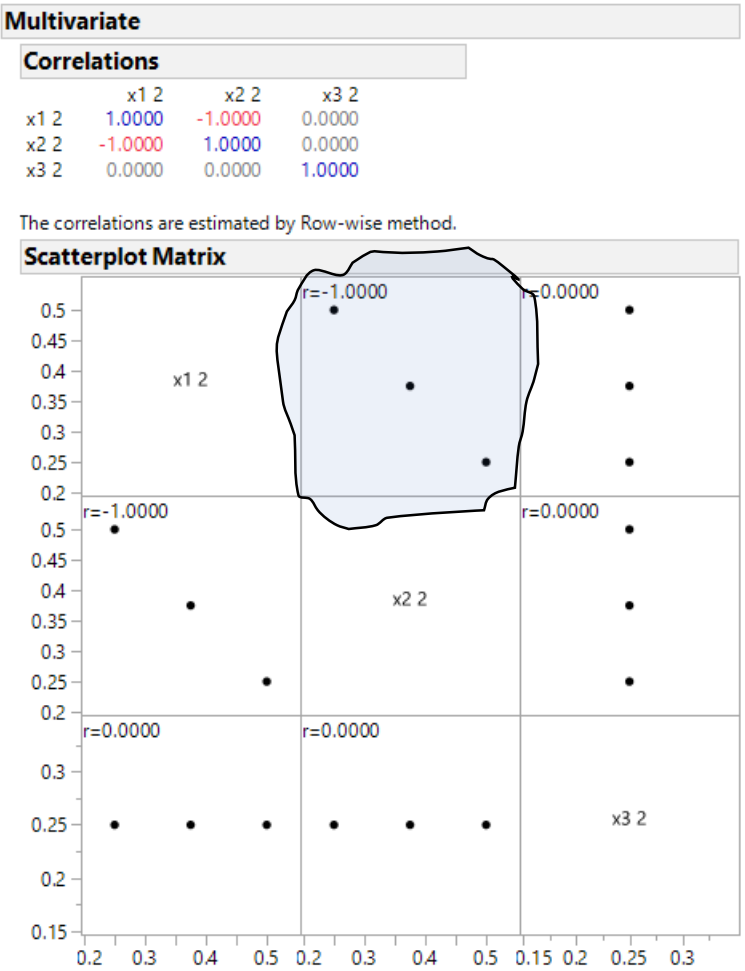
Compositional data

On spurious correlations

Full composition | Subcomposition

x1	x2	x3	x4	x1 2	x2 2	x3 2
0.1	0.2	0.1	0.6	0.25	0.5	0.25
0.2	0.1	0.1	0.6	0.5	0.25	0.25
0.3	0.3	0.2	0.2	0.375	0.375	0.25





Agenda

1. Background on causality in science and statistics
2. Fishbone cause and effect diagrams
3. Bayesian networks
4. **Randomization in experimental designs**
5. Propensity scores in observational studies
6. Counterfactuals and do calculus
7. Personalized medicine, condition based maintenance and Industry 4.0
8. Future research areas





“No aphorism is more frequently repeated in connection with field trials, than that we must ask Nature few questions, or, ideally, one question, at a time. The writer is convinced that this view is wholly mistaken. Nature, he suggests, will best respond to a logical and carefully thought out questionnaire. A factorial design allows the effect of several factors and interactions between them, to be determined with the same number of trials as are necessary to determine any one of the effects by itself with the same degree of accuracy.”

R.A. Fisher (1926). The arrangement of field experiments, *Journal of the Ministry of Agriculture of Great Britain* 33, 503–513.

The Design of Experiments

By

R. A. Fisher, Sc.D., F.R.S.

Formerly Fellow of Gonville and Caius College, Cambridge
Honorary Member, American Statistical Association
and American Academy of Arts and Sciences
Galton Professor, University of London

Oliver and Boyd
Edinburgh: Tweeddale Court
London: 33 Paternoster Row, E.C.
1937

LE seul moyen de prévenir ces écarts, consiste à supprimer, ou au moins à simplifier, autant qu'il est possible, le raisonnement qui est de nous, & qui peut seul nous égarer, à le mettre continuellement a l'épreuve de l'expérience; à ne conserver que les faits qui sont des vérités données par la nature, & qui ne peuvent nous tromper; à ne chercher la verité que dans l'enchaînement des expériences & des observations, sur-tout dans l'ordre dans lequel elles sont présentées, de la même manière que les mathématiciens parviennent à la solution d'un problême par le simple arrangement des données, & en réduisant le raisonnement à des opérations si simples, à des jugemens si courts, qu'ils ne perdent jamais de vue l'évidence qui leur sert de guide.

Methode de Nomenclature chimique,
A. L. LAVOISIER, 1787.

I have assumed, as the experimenter always does assume, that it *is* possible to draw valid inferences from the results of experimentation; that it is possible to argue from consequences to causes, from observations to hypotheses; as a statistician would say, from a sample to the population from which the sample was drawn, or, as a logician might put it, from the particular to the general.

An implicit definition of causal effects by Fisher is the following:

If we say, 'This boy has grown tall because he has been well fed,' we are not merely tracing out cause and effect in an individual instance; we are suggesting that he might quite probably have been worse fed, and that in this case he would have been shorter. ***We are, in fact, suggesting that existing differences of nutrition can account for differences of stature comparable to the standard deviation of stature****. Now this is just what is meant when we speak of nutrition as a cause of variability; we thereby mean that in a population absolutely uniform in regard to other causes, such as breeding and exercise, existing differences of nutrition would produce a certain variability—in fact, that a certain percentage of the variance must be ascribed to nutrition.*

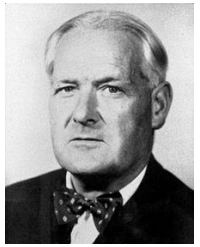
Fisher RA (1919) The causes of human variability. *The Eugenics Review*;10(4): 213-220.

In the 1920s RA Fisher presented **randomization** as an essential ingredient of his approach to the design and analysis of experiments, validating significance tests. In its absence, the experimenter had to rely on his judgement that the effects of biases could be discounted.

Twenty years later, Bradford Hill promulgated the **random assignment of treatments in clinical trials as the only means of avoiding systematic bias between the characteristics of patients assigned to different treatments**. The two approaches were complementary, Fisher appealing to statistical theory, Hill to practical needs. The two men remained on good terms throughout most of their careers.

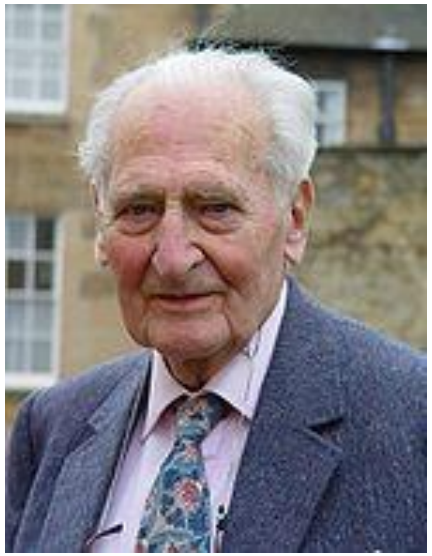
Peter Armitage (2003) Fisher, Bradford Hill, and randomization, *International Journal of Epidemiology* 32:925–928

Bradford Hill, A. (1953). Observation and experiment. New England Journal of Medicine 248:995-1001
Bradford Hill, A. (1965). The environment and disease: association or causation? Proceedings of the Royal Society of Medicine 58:295-300



Strength (effect size): A small association does not mean that there is not a causal effect, though the larger the association, the more likely that it is causal.
Consistency (reproducibility): Consistent findings observed by different persons in different places with different samples strengthens the likelihood of an effect.
Specificity: Causation is likely if there is a very specific population at a specific site and disease with no other likely explanation. The more specific an association between a factor and an effect is, the bigger the probability of a causal relationship.
Temporality: The effect has to occur after the cause (and if there is an expected delay between the cause and expected effect, then the effect must occur after that delay).
Biological gradient: Greater exposure should generally lead to greater incidence of the effect. However, in some cases, the mere presence of the factor can trigger the effect. In other cases, an inverse proportion is observed: greater exposure leads to lower incidence.[
Plausibility: A plausible mechanism between cause and effect is helpful (but Hill noted that knowledge of the mechanism is limited by current knowledge).
Coherence: Coherence between epidemiological and laboratory findings increases the likelihood of an effect. However, Hill noted that "... lack of such [laboratory] evidence cannot nullify the epidemiological effect on associations".
Experiment: "Occasionally it is possible to appeal to experimental evidence".
Analogy: The effect of similar factors may be considered.

Austin Bradford Hill
(1897-1991



Richard Doll
(1912 – 2005)

The Environment and Disease: Association or Causation?

by Sir Austin Bradford Hill CBE DSC FRCP(hon) FRS
(*Professor Emeritus of Medical Statistics, University of London*)

Amongst the objects of this newly-founded Section of Occupational Medicine are firstly 'to provide a means, not readily afforded elsewhere, whereby physicians and surgeons with a special knowledge of the relationship between sickness and injury and conditions of work may discuss their problems, not only with each other, but also with colleagues in other fields, by holding joint meetings with other Sections of the Society'; and, secondly, 'to make available information about the physical, chemical and psychological hazards of occupation, and in particular about those that are rare or not easily recognized'.

BRITISH MEDICAL JOURNAL

LONDON SATURDAY SEPTEMBER 30 1950

SMOKING AND CARCINOMA OF THE LUNG

PRELIMINARY REPORT

BY

RICHARD DOLL, M.D., M.R.C.P.
Member of the Statistical Research Unit of the Medical Research Council

AND

A. BRADFORD HILL, Ph.D., D.Sc.
Professor of Medical Statistics, London School of Hygiene and Tropical Medicine; Honorary Director of the Statistical Research Unit of the Medical Research Council

Br Med J. 1950 Sep 30; 2(4682): 739–748.
Smoking and Carcinoma of the Lung
Richard Doll and A. Bradford Hill

"U.S. Surgeon General Luther Terry holds a copy of the 387 page report of the Advisory Committee to the Surgeon General of the Public Health Service on the relationship of smoking to health Jan 11, 1964. He spoke at a Washington news conference at which the study was released. It termed smoking a health hazard calling for corrective action." (AP Photo/hwg)

US report ties smoking to cancer

Cornfield Inequality

Cornfield J (1956). **A statistical problem arising from retrospective studies**. *Proceedings 3rd Berkeley Symposium on Mathematical Statistics,* 4:135–48.

R_0 is the observed relative risk between an exposed and unexposed group, which could be explained by an unmeasured confounder, U. R_O is no greater than the ratio of the prevalence of U in the exposed to that in the unexposed population. $R_O \leq R_U$, where R_U is the ratio of risk in those with U compared to those without U.

Lung cancer in asbestos workers: relative risk of asbestos exposed workers dying from lung cancer is **6.8** times their expected number in general population.

60% of all males smoke, 80% of males in asbestos-related occupations. The prevalence ratio, 0.8/0.6 = **1.33**, is much less than R_o = 6.8, so Cornfield's inequality implies that smoking cannot explain the entire association between asbestos and lung cancer.

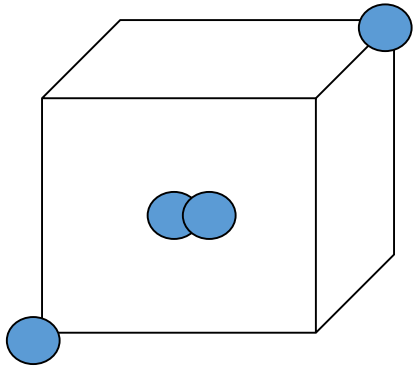
Cornfield Inequality

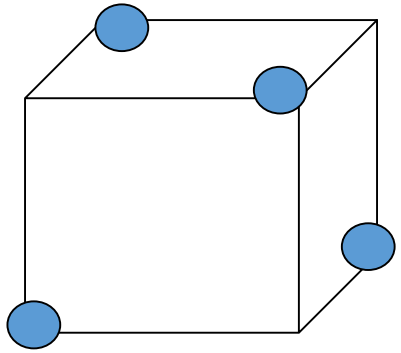
Cornfield J, Haenszel W, Hammond EC, Lilienfeld AM, Shimkin MB, Wynder EL (1954) **Smoking and lung cancer: recent evidence and a discussion of some questions**. J Natl Cancer Inst 1954;22:

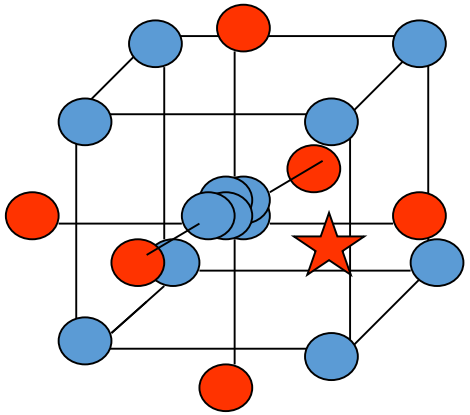
"The consistency of all the epidemiologic and experimental evidence also supports the conclusion of a causal relationship with cigarette smoking…results in animals are fully consistent with the epidemiologic findings in man.

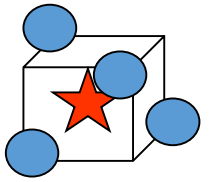
When a demonstrable parallelism exists between epidemiologic data and laboratory findings, greater significance accrues to both."

Design of Experiments Strategy









Scoping

Initial assessment

Screening

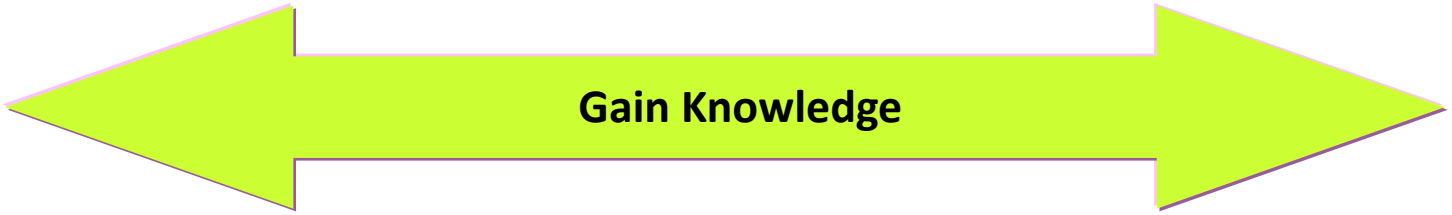
Fractional designs

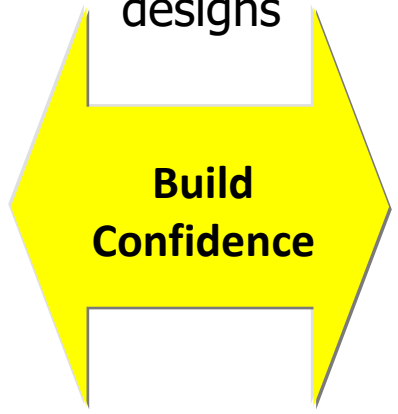
Optimizing

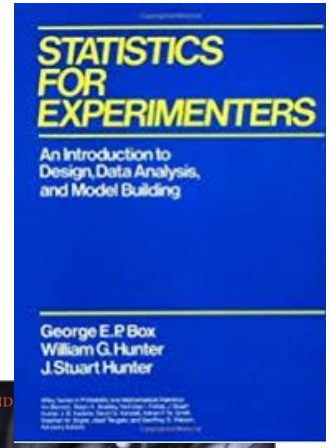
Response surfaces

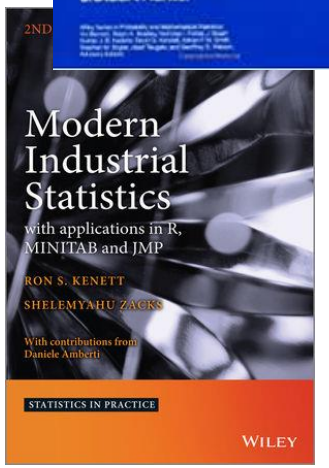
Robustness

Robust designs









Replicates and pseudo-replicates (Hurlbert):
https://web.ma.utexas.edu/users/mks/statmistakes/pseudorep.html

Experiment
Make it your motto day and night
Experiment
And it will lead you to the light
The apple on the top of the tree
Is never too high to achieve
So take an example from Eve
Experiment
Be curious
Though interfering friends may frown,
Get furious
At each attempt to hold you down
If this advice you'll only employ
The future can offer you infinite joy
And merriment
Experiment
And you'll see

Mabel Mercer sings Cole Porter

Rubin: What if, in a randomized experiment, the chosen randomized allocation exhibited substantial imbalance on a prognostically important baseline covariate?

Cochran: Why didn't you block on that variable?

Rubin: Well, there were many baseline covariates, and the correct blocking wasn't obvious; and I was lazy at that time.

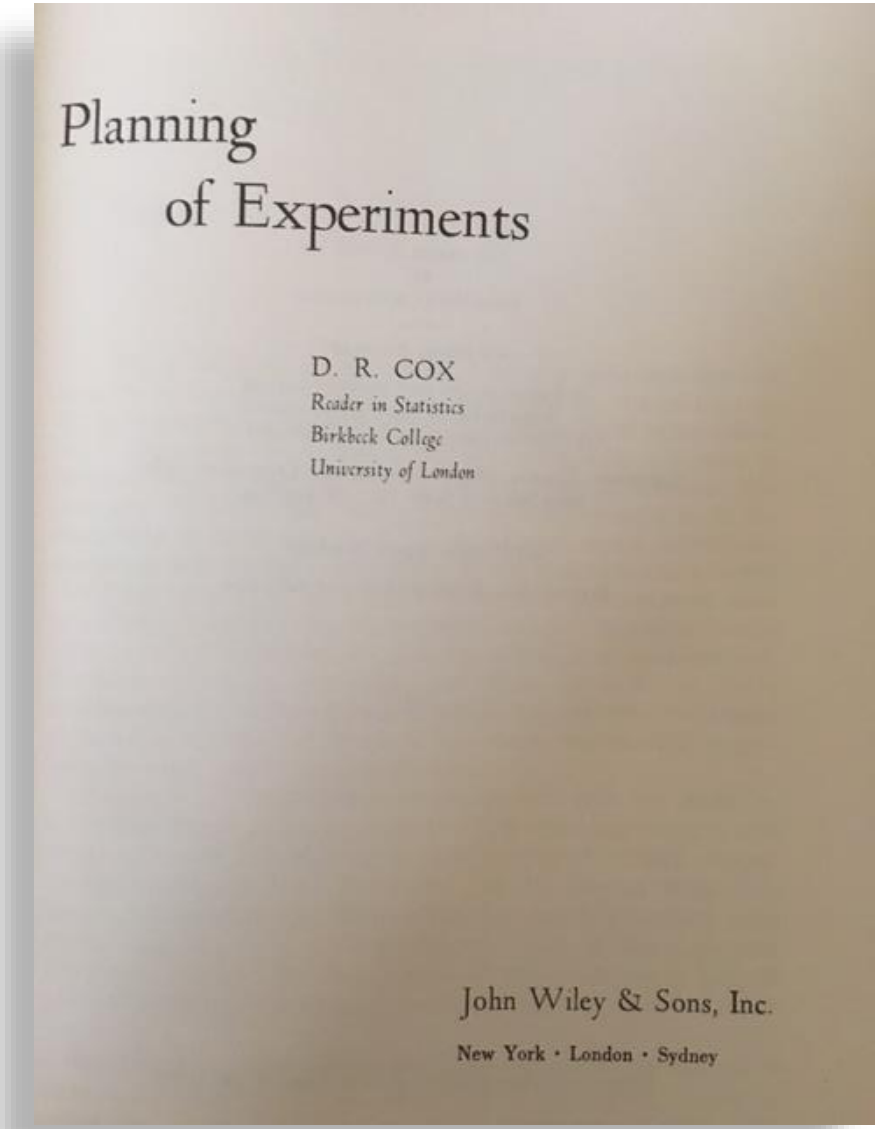
Cochran: This is a question that I once asked Fisher, and his reply was unequivocal:

Fisher (recreated via Cochran): Of course, if the experiment had not been started, I would rerandomize.

Don Rubin, Annual meeting of Israeli Statistical Association,31/5/ 2018

When asked: How you would handle a random order with a perceptible pattern? Fisher responded that he did not understand the question: "I would of course rerandomize"

D.R. Cox (personal communication, 26/2/2019)



On randomization and re-randomization

and $T_2\ T_1$, giving each equal probability. The full discussion of this process of randomization is deferred to Chapter 5.

A typical arrangement of treatments resulting from such a randomization is shown in Table 3.1 together with some fictitious observations. For each pair of units the difference between the observation on T_2 and the observation on T_1 is calculated. The treatment effect is estimated by $\bar{d}$, the mean of these differences, and the estimated standard error of $\bar{d}$, and a test of the statistical significance of $\bar{d}$ can be obtained by simple standard statistical calculations (Goulden, 1952, p. 51), the amount of the uncontrolled variation being estimated from the observed dispersion of the differences in the last column of Table 3.1.

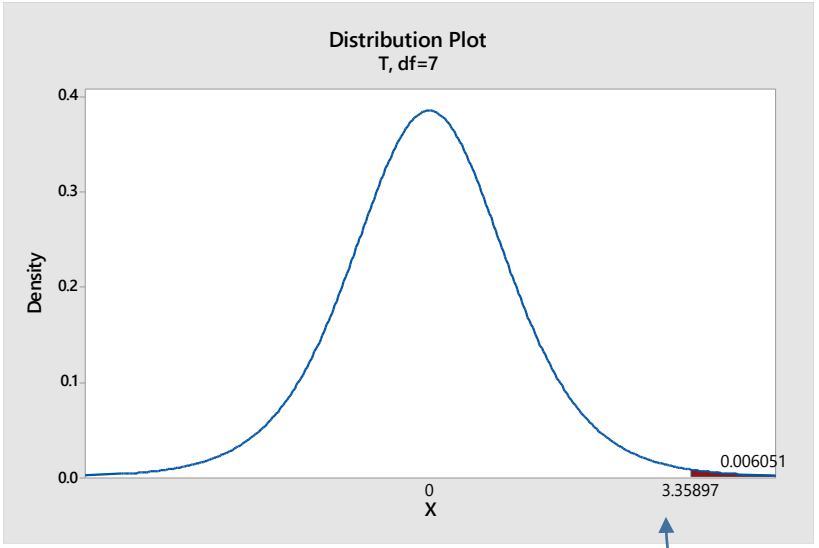
A treatment is applied as T1 or T2.
What is the treatment effect?
Is the effect at T2 greater than the effect at T1?

TABLE 3.1

PAIRED COMPARISON EXPERIMENT

Day	First Unit	Second Unit	Difference, d
1	T_1:2.8	T_2:3.2	0.4
2	T_2:3.1	T_1:3.1	0.0
3	T_2:3.4	T_1:2.9	0.5
4	T_1:3.0	T_2:3.5	0.5
5	T_2:2.7	T_1:2.4	0.3
6	T_2:2.9	T_1:3.0	−0.1
7	T_2:3.5	T_1:3.2	0.3
8	T_1:2.6	T_2:2.8	0.2

Mean, $\bar{d} = 0.262$
Estimated standard error = 0.078

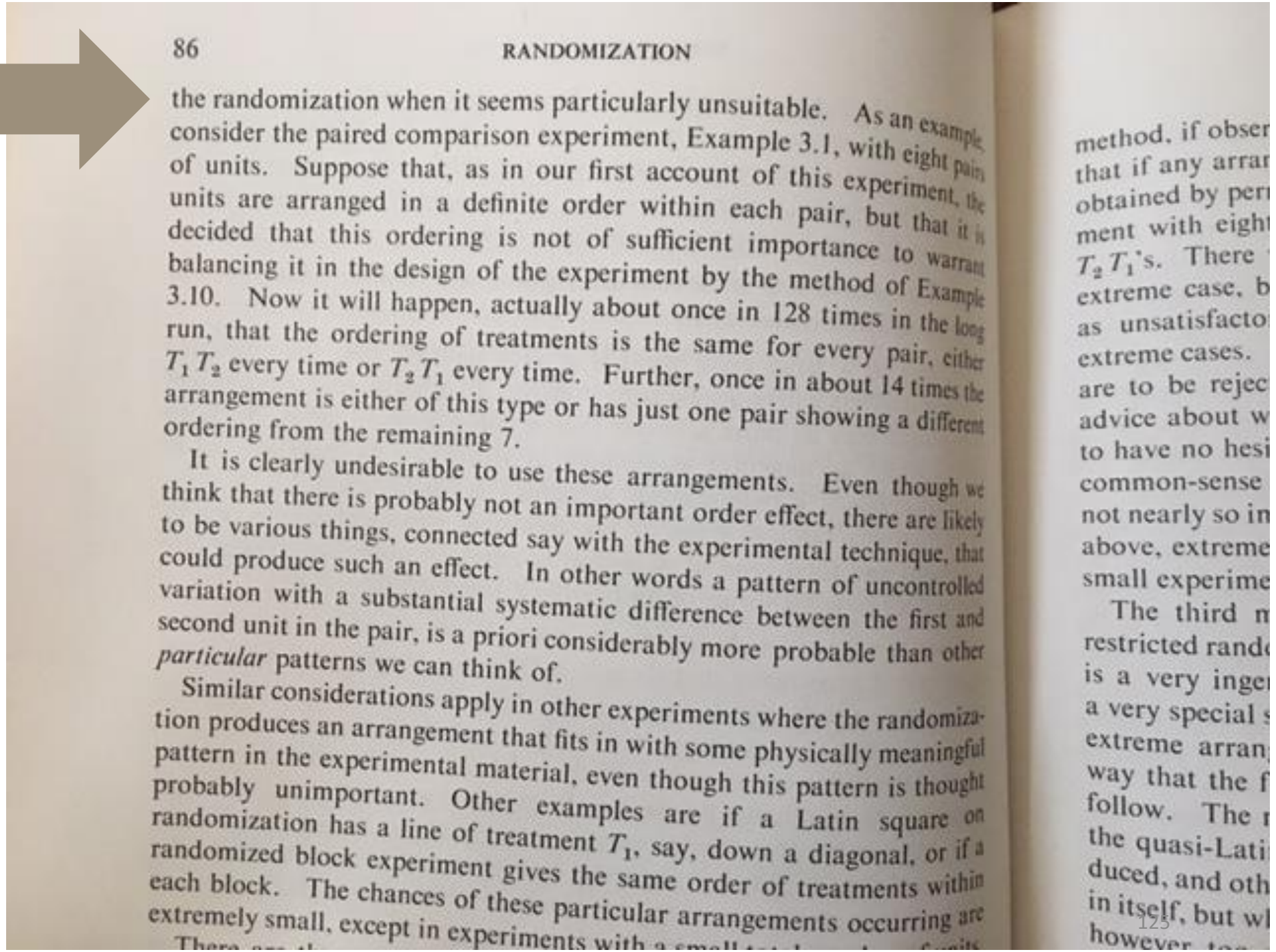


5.7 SOME FURTHER POINTS

There are some difficulties that arise in the application of randomization, articularly to small experiments, and these will now be discussed. The first point concerns the rejection of an arrangement produced by

Following this introduction,
Cox discusses three approaches marked below in red, green and blue.

Further discussion is marked in yellow.



the randomization when it seems particularly unsuitable. As an example, consider the paired comparison experiment, Example 3.1, with eight pairs of units. Suppose that, as in our first account of this experiment, the units are arranged in a definite order within each pair, but that it is decided that this ordering is not of sufficient importance to warrant balancing it in the design of the experiment by the method of Example 3.10. Now it will happen, actually about once in 128 times in the long run, that the ordering of treatments is the same for every pair, either $T_1 T_2$ every time or $T_2 T_1$ every time. Further, once in about 14 times the arrangement is either of this type or has just one pair showing a different ordering from the remaining 7.

It is clearly undesirable to use these arrangements. Even though we think that there is probably not an important order effect, there are likely to be various things, connected say with the experimental technique, that could produce such an effect. In other words a pattern of uncontrolled variation with a substantial systematic difference between the first and second unit in the pair, is a priori considerably more probable than other *particular* patterns we can think of.

Similar considerations apply in other experiments where the randomization produces an arrangement that fits in with some physically meaningful pattern in the experimental material, even though this pattern is thought probably unimportant. Other examples are if a Latin square on randomization has a line of treatment T_1, say, down a diagonal, or if a randomized block experiment gives the same order of treatments within each block. The chances of these particular arrangements occurring are extremely small, except in experiments with a small

method, if obser
that if any arran
obtained by per
ment with eight
$T_2 T_1$'s. There
extreme case, b
as unsatisfacto
extreme cases.
are to be rejec
advice about w
to have no hesi
common-sense
not nearly so in
above, extreme
small experime
The third m
restricted rand
is a very inge
a very special
extreme arran
way that the f
follow. The
the quasi-Lati
duced, and oth
in itself, but wh
however

extremely small, except in experiments with a small total number of units.

There are three ways of dealing with the difficulty, all depending on curtailing the randomization. The first method is to incorporate a condition about order into the formal design of the experiment, as was done in Example 3.10, where T_1 and T_2 each occurred four times in the first position. This is probably the best solution in the present case, but it is certainly not a general answer to the problem, since there are various reasons why it may be impracticable or undesirable to introduce further constraints into the design. For example we lose degrees of freedom for residual in eliminating a source of variation that is probably not important, we make the experiment more complicated and there may already be several different systems of grouping in the design, making the introduction of further conditions difficult or impossible.

The second method is to reject extreme arrangements whenever they occur, i.e., to rerandomize. For example in the paired comparison

curtailing the randomization

tion about order into the formal design of the experiment, as was done in Example 3.10, where T_1 and T_2 each occurred four times in the first position. This is probably the best solution in the present case, but it is certainly not a general answer to the problem, since there are various reasons why it may be impracticable or undesirable to introduce further constraints into the design. For example we lose degrees of freedom for residual in eliminating a source of variation that is probably not important, we make the experiment more complicated and there may already be several different systems of grouping in the design, making the introduction of further conditions difficult or impossible.

The second method is to reject extreme arrangements whenever they occur, *i.e.*, to rerandomize. For example in the paired comparison experiment, we may decide to reject all arrangements with seven or more pairs in the same order. A highly desirable condition in using this

method, if observer biases like those of Example 5.6 are to be avoided, is that if any arrangement is to be rejected, so must all other arrangements obtained by permuting the names of the treatments. Thus if the arrangement with eight $T_1 T_2$'s is rejected, so must the arrangement with eight $T_2 T_1$'s. There would be little likelihood of disagreement over such an extreme case, but since the decision as to what arrangements to regard as unsatisfactory is arbitrary, there could be disagreement with less extreme cases. The best plan is, if possible, to decide which arrangements are to be rejected before randomization. It is difficult to give general advice about which arrangements to reject, but the best rule is probably to have no hesitation in rejecting any arrangement that seems on general common-sense grounds to be unsatisfactory. Fortunately this matter is not nearly so important in practice as might be thought, since, as remarked above, extreme arrangements occur with appreciable chance only in very small experiments.

common-sense grounds to be unsatisfactory. Fortunately this matter is not nearly so important in practice as might be thought, since, as remarked above, extreme arrangements occur with appreciable chance only in very small experiments.

The third method is to use a special device, known technically as restricted randomization (Grundy and Healy, 1951; Youden, 1958). This is a very ingenious idea, in which a design is selected at random from a very special set of arrangements. The set is chosen to exclude both the extreme arrangements and the very balanced arrangements, in such a way that the full mathematical consequences of ordinary randomization follow. The method is probably of most value for a special design called the quasi-Latin square (Chapter 12), for which the method was first introduced, and otherwise in a series of small experiments, each of some interest in itself, but which also need to be considered collectively. The method is however too specialized to discuss here and its full implications have not yet been worked out; the nonstatistical reader requiring more information about it should consult a statistician.

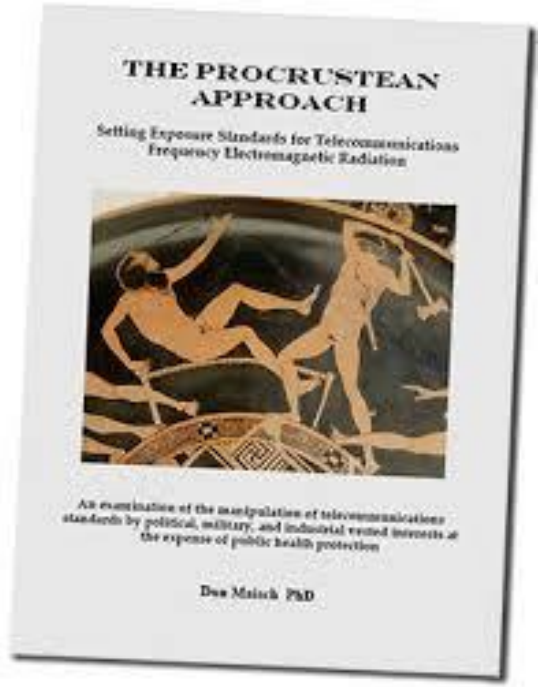
The reader may object that the second method, the rejection of extreme arrangements, will falsify the mathematical consequences of randomization described in § 5.6. This is true of the estimation of error, although

duced, and otherwise in a series of small experiments, each of some interest in itself, but which also need to be considered collectively. The method is however too specialized to discuss here and its full implications have not yet been worked out; the nonstatistical reader requiring more information about it should consult a statistician.

The reader may object that the second method, the rejection of extreme arrangements, will falsify the mathematical consequences of randomization described in § 5.6. This is true of the estimation of error, although not of the absence of bias in the treatment estimates themselves. The estimate of error will only be unbiased if there is in fact no systematic order effect. However in single small experiments the estimate of error is very inaccurate anyway. More importantly we have here a mathematical interpretation of randomization: that it leads to desirable properties in the long run, or on the average, and on the other hand a practical problem—namely the designing and drawing of useful conclusions from a particular single experiment that we are now in the process of considering. Usually the concept that our procedures will work out well in the long run is a very helpful one, both qualitatively and in giving a vivid physical picture of the meaning of probabilities calculated in connection with a

particular experiment. However to adopt arrangements that we suspect are bad, simply because things will be all right in the long run, is to force our behavior into the Procrustean bed of a mathematical theory. Our object is the design of individual experiments that will work well: good long-run properties are concepts that help us in doing this, but the exact fulfillment of long-run mathematical conditions is not the ultimate aim.

The second general matter is closely related to the first. Suppose that we design and carry out a randomized experiment, and that when we come to analyze and interpret the results we realize either that the arrangement we have used is probably an unfortunate one and should have been rejected, or, by inspection of the results, that there is some particular form of uncontrolled variation. For example, we might have the above paired comparison experiment with, say, six pairs receiving the order $T_1 T_2$ and two receiving the order $T_2 T_1$. Inspection of the results may suggest a substantial order effect comparable t



long-run properties are concepts that help us in doing this, but the exact

fulfillment of long-run mathematical conditions is not the ultimate aim.

The second general matter is closely related to the first. Suppose that we design and carry out a randomized experiment, and that when we come to analyze and interpret the results we realize either that the arrangement we have used is probably an unfortunate one and should have been rejected, or, by inspection of the results, that there is some particular form of uncontrolled variation. For example, we might have the above paired comparison experiment with, say, six pairs receiving the order $T_1 T_2$ and two receiving the order $T_2 T_1$. Inspection of the results may suggest a substantial order effect comparable to the treatment effect. Another example would be if an agricultural field trial arranged in randomized blocks shows a systematic trend from one end to the other of the experimental area. What do we do in such situations?

In some cases, possibly in the first, we may decide that the data s

be regarded with suspicion. Suppose, however, that we do wish to

what conclusions we can. The previous discussion shows that it

good enough to say that the long-run properties are valid whatev

form of the uncontrolled variation and on those grounds to analy

experimental results by the usual methods. On the other hand, to

duce modifications into the analysis based on inspection of the

and

obj

Fortunat

Another

reason wh

example a

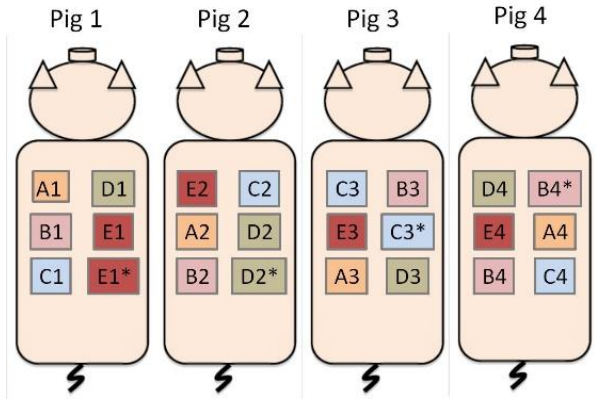
here is tha

planted a

For this

not occu

look f



Factors

Add Factor | Remove | Add N Factors 1

Split plot design

Name	Role	Changes	Values
Antibiotic	Categorical	Hard	A, B
Timing	Categorical	Easy	2, 4, 12
Concentration	Categorical	Easy	0, 2, 4, 6

REML Variance Component Estimates

Random Effect	Var Ratio	Var Component	Std Error	95% Lower	95% Upper	Pct of Total
Animal	0.3369366	0.0009742	0.0014479	-0.001864	0.003812	25.202
Residual		0.0028914	0.0011798	0.0014872	0.0078735	74.798
Total		0.0038656	0.0017459	0.0018763	0.0120769	100.000

-2 LogLikelihood = 5.2031109955

Note: Total is the sum of the positive variance components.

Total including negative estimates = 0.0038656

Fixed Effect Tests

Source	Nparm	DF	DFDen	F Ratio	Prob > F
Antibiotic	1	1	1.82	13.1960	0.0783
Timing	2	2	12.01	3.2182	0.0760
Concentration	3	3	12.06	9.2033	0.0019*
Antibiotic*Timing	2	2	12.01	0.3831	0.6898
Antibiotic*Concentration	3	3	12.03	0.7022	0.5687
Timing*Concentration	6	6	12.47	0.6574	0.6849

https://community.jmp.com/t5/JMP-Blog/The-QbD-Column-Split-plot-experiments/ba-p30716

duce modifications into the analysis
and on personal judgement about the design must lead to some loss of
objectivity. The following procedure is suggested.

(*a*) Work through the conventional analysis of the observations ignoring the suspected complication.

(*b*) Make a special statistical analysis of the observations taking account of the complication in whatever seems the most reasonable way. The reader who is not familiar with fairly advanced statistical methods will probably need statistical advice in this. The method will usually involve the analysis of what is known technically as a nonorthogonal least-squares situation.

(*c*) If the conclusions of the two analyses are for practical purposes equivalent there is no difficulty. If the conclusions do differ, care is needed. The assumptions underlying the second analysis should be carefully thought over, and if they seem reasonable, the second analysis should be regarded as correct.

(*d*) In reporting on the experiment, conclusions from both analyse

should be given, at any rate briefly. If the first analysis is rejected, reasons should be outlined. The general idea should be to make it clear to the reader what has been done and to give him the opportunity of forming his own conclusions as far as practicable.

Fortunately these difficulties tend to occur infrequently in practice.

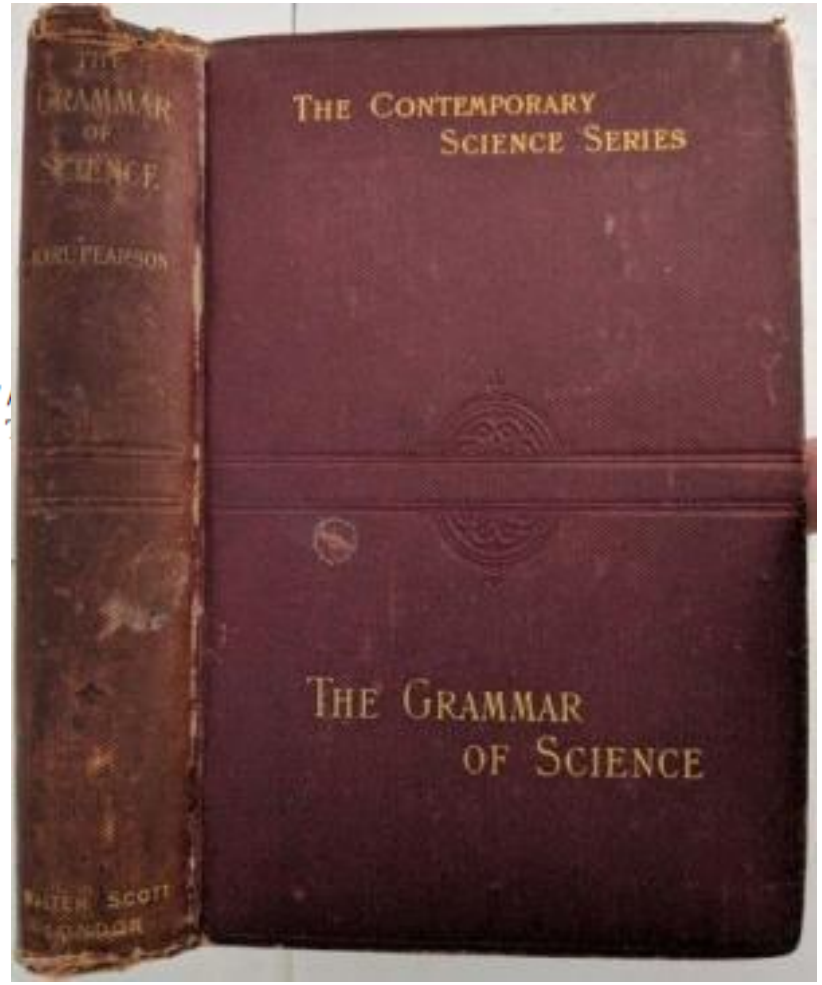
Another difficulty that occasionally arises is that there is some practical reason why certain treatment arrangements are not allowable. One example arises in raspberry variety trials (Taylor, 1950). The point here is that additional canes spring up near many of the canes originally planted and it is necessary to remove these new canes from each plot. For this to be possible varieties that resemble each other closely must not occur close together, thus restricting the randomization. Another example occurs in carpet wearing trials, in which dyed and undyed carpets are under comparison. An experimental carpet is formed by sewing together squares of carpet of different types and the whole carpet placed say in a busy corridor. It would often be desirable that the carpet should look presentable and this would preclude full randomization of the dyed and undyed sections. The procedure in such cases is either to do as much randomization as possible or to use a systematic arrangement taking whatever steps are practicable to avoid bias.

Eur J Epidemiol
DOI 10.1007/s10654-017-0288-1

CUTTER LECTURE

Statistical science: a grammar for research

David. R. Cox[1]



Received: 13 July 2017 /

Causality

There are broadly at least three views of causality in the literature; for a brief review, see Cox and Wermuth [6].

1

First, largely in the time series field, there is Wiener-Granger causality essentially about the ability of one time series to predict the future of another. Wiener was an outstanding MIT pure mathematician and Granger an econometrician.

2

The second and widely used definition involves the notion of an exposure being hypothetically changed, other things being equal. It can be regarded as underpinning the classical theory of randomized experiments and, generalized into broader settings, it has a large and rich literature.

3

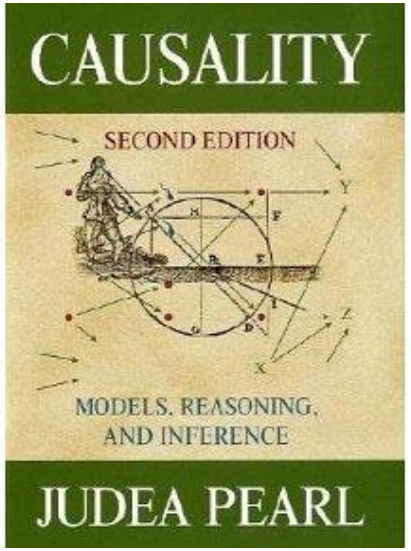
The third notion adds to the second some notion of evidence-based explanation in terms of an underlying process, biological or physical perhaps. Of course such explanations are not "ultimate". Their danger is that they can nearly always be manufactured after the event, but very much more than that is required, typically explicit independent evidence. Davey-Smith coined the term triangulation for this view of causality.

Agenda

1. Background on causality in science and statistics
2. Fishbone cause and effect diagrams
3. Bayesian networks
4. Randomization in experimental designs
5. **Propensity scores in observational studies**
6. **Counterfactuals and do calculus**
7. Personalized medicine, condition based maintenance and Industry 4.0
8. Future research areas

Cited by

	All
Citations	89932
h-index	98
i10-index	307



Judea Pearl

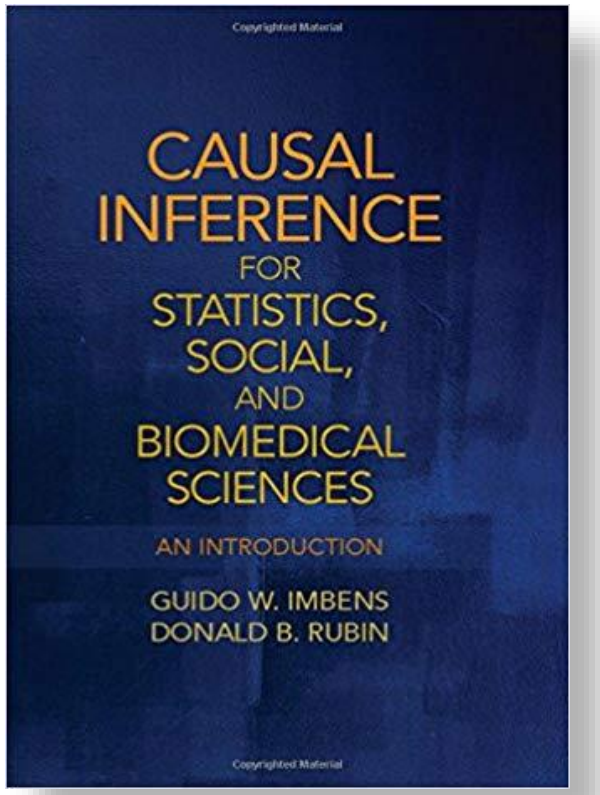
2011 Turing Award for fundamental contributions to artificial intelligence through the development of a calculus for probabilistic and causal reasoning

Maximum likelihood from incomplete data via the EM algorithm AP Dempster, NM Laird, DB Rubin Journal of the royal statistical society. Series B (methodological), 1-38	55477	1977
Statistical analysis with missing data RJA Little, DB Rubin John Wiley & Sons	24728	2014
Bayesian data analysis A Gelman, J Carlin, H Stern, DB Rubin CRC press	22694 *	2004
The central role of the propensity score in observational studies for causal effects PR Rosenbaum, DB Rubin Biometrika 70 (1), 41-55	21341	1983

Don Rubin

Cited by

	All
Citations	279566
h-index	128
i10-index	364



Causal Analysis

For causal questions, we need to infer aspects of the data generation process.

We need to be able to deduce:

- the likelihood of events under *static conditions*, (as in standard statistical analysis) and also
- the dynamics of events under *changing conditions.*

Causal Analysis

"dynamics of events under changing conditions" includes:

- Predicting the effects of interventions.
- Predicting the effects of spontaneous changes.
- Identifying causes of reported events.

Population & Outcome Variable

Define the population by U.
Each unit in U is denoted by u.

The outcome of interest is Y.
Also called the response variable.

For each $u \in U$,
there is an associated value $Y(u)$.

Causes/Treatment

Causes are those things that could be treatments or conditions in hypothetical experiments.

For simplicity, we assume that there are just two possible states:

- Unit *u* is exposed to treatment/condition and
- Unit *u* is exposed to comparison.

The Treatment/Condition Variable

Let *D* be a variable that indicates the state to which each unit in *U* is exposed.

$$D = \begin{cases} 1 & \text{If unit } u \text{ is exposed to treatment/condition} \\ 0 & \text{If unit } u \text{ is exposed to comparison} \end{cases}$$

Where does *D* come from?

- In a controlled study:
constructed by the experimenter.
- In an uncontrolled study:
determined by factors beyond the experimenter's control.

Linking Y and D

Potential outcomes

Y = response variable

D = treatment/condition variable

The response *Y* is potentially affected by whether *u* receives treatment or not.

Thus, we need two response variables:

$Y_1(u)$ is the outcome if unit *u* is exposed to treatment.

$Y_0(u)$ is the outcome if unit *u* is exposed to comparison.

The Effect of Treatment/Condition on Outcome

Treatment variable *D*

$$D = \begin{cases} 1 & \text{If unit } u \text{ is exposed to treatment} \\ 0 & \text{If unit } u \text{ is exposed to comparison} \end{cases}$$

Response variable *Y*

$Y_1(u)$	is the outcome if unit *u* is exposed to treatment
$Y_0(u)$	is the outcome if unit *u* is exposed to comparison

$$\delta_u = Y_1(u) - Y_0(u)$$

Le nez de Cléopâtre: s'il eut été plus court, toute la face de la terre aurait change. Pascal (1669)

Counterfactuals

For any unit *u*, treatment causes the effect

$$\delta_u = Y_1(u) - Y_0(u)$$

Fundamental problem of causal inference

For a given unit *u*, we observe either $Y_1(u)$ *or* $Y_0(u)$, it is impossible to observe the effect of treatment on *u* by itself!

We do not observe the **counterfactual**

If we give *u* treatment, then we cannot observe what would have happened to *u* in the absence of treatment.

The **propensity score (PS)** is the probability of treatment assignment conditional on observed baseline characteristics. The **propensity score** allows one to design and analyze an observational (nonrandomized) study so that it mimics some of the particular characteristics of a randomized controlled trial.

Warning: PS for an incomplete blocks design is identical to a completely randomized design

https://onlinelibrary.wiley.com/doi/10.1002/sim.3133

Bridging observational studies and randomized experiments by embedding the former in the latter

Statistical Methods in Medical Research
0(0) 1–21

DOI: 10.1177/0962280217740609
journals.sagepub.com/home/smm
SAGE

Marie-Abele C Bind and Donald B Rubin

Abstract

Consider a statistical analysis that draws causal inferences from a
being valid in the standard frequentist senses; i.e. the analysis pr
valid in the sense of rejecting true null hypotheses at the nom
which are presented as having at least their nominal coverage fo
statements, the analysis must embed the observational study in
observed data, or a subset of that hypothetical randomized data
involves: (1) a purely *conceptual stage* that precisely formulate th
experiment where the exposure is assigned to units; (2) a *de*
before any outcome data are observed, (3) a *statistical analysis st*
and non-exposed units of the hypothetical randomized experim
statistical evidence for the sizes of possible causal effects. Stage
the effort, whereas Stage 1 demands careful scientific argume
readers of the proffered statistical analysis. Otherwise, the resu
a presentation of scientifically meaningless arithmetic calculation
most scientifically interesting to the dedicated researcher a
perspective is rarely implemented with any rigor, for example,
approach using an example examining the effect of parental smok
in East Boston in the 1970s.

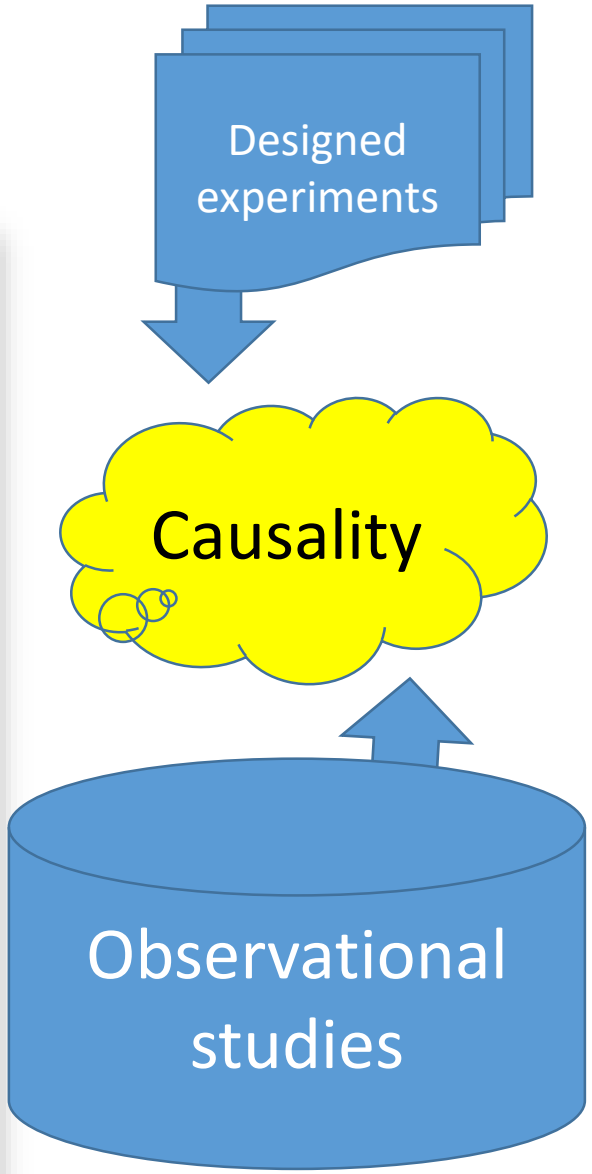
CAUSAL INFERENCE IN RETROSPECTIVE STUDIES

PAUL W. HOLLAND
Research Statistics Group
Educational Testing Service
DONALD B. RUBIN
Harvard University

Philosophical discussions of causality often emphasize the *meaning* of causation. Scientists are usually concerned with *understanding* causal mechanisms. Purely statistical discussions of causality are substantially more limited in scope, because the unique contribution of statistics is to *measuring* causal effects and not to the understanding of causal mechanisms or to the meaning of causation. This distinction is sometimes expressed as "statistics can establish correlation, but not causation." We feel our emphasis on *measurement* is more appropriate, because it focuses on what statistical theory *can* contribute to discussions of causality. Measuring causal effects accurately without any understanding whatsoever of the causal mechanisms

AUTHORS' NOTE: *A version of this article titled "Causal Inference in Prospective and Retrospective Studies" was delivered at the Jerome Cornfield Memorial Session of the American Statistical Association, August 1980, in Houston. The topic of the article was especially appropriate for that session since many important contributions to the study of health effects from prospective and retrospective studies were made by Jerome Cornfield.*

EVALUATION REVIEW, Vol. 12 No. 3, June 1988 203-231



Biometrika (1995), **82**, 4, *pp.* 669–710
Printed in Great Britain

R-218-B

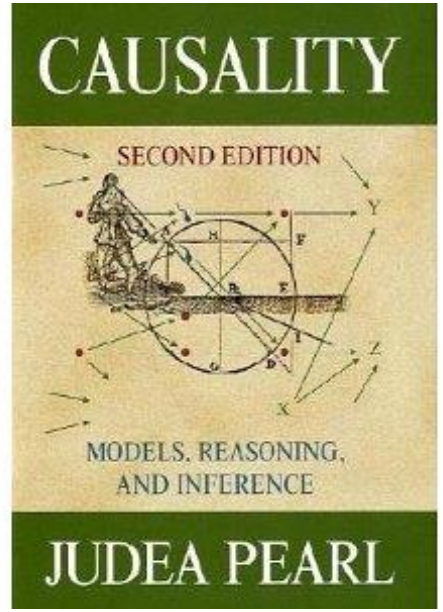
Causal diagrams for empirical research (With Discussions)

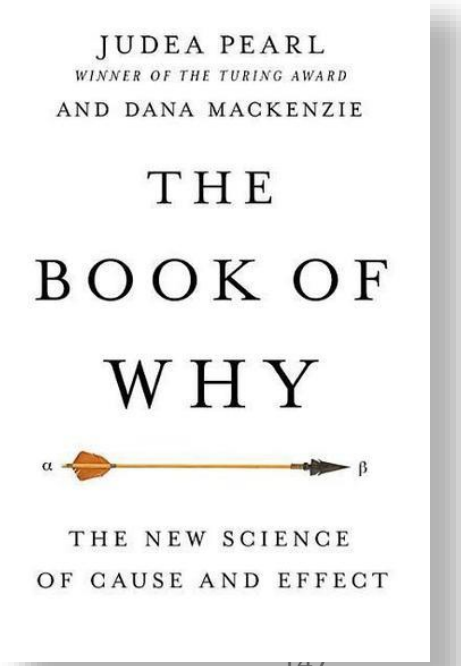
BY JUDEA PEARL

Cognitive Systems Laboratory, Computer Science Department, University of California, Los Angeles, California 90024, U.S.A.

SUMMARY

The primary aim of this paper is to show how graphical models can be used as a mathematical language for integrating statistical and subject-matter information. In particular, the paper develops a principled, nonparametric framework for causal inference, in which diagrams are queried to determine if the assumptions available are sufficient for identifying causal effects from nonexperimental data. If so the diagrams can be queried to produce mathematical expressions for causal effects in terms of observed distributions; otherwise, the diagrams can be queried to suggest additional observations or auxiliary experiments from which the desired inferences can be obtained.





Imagining

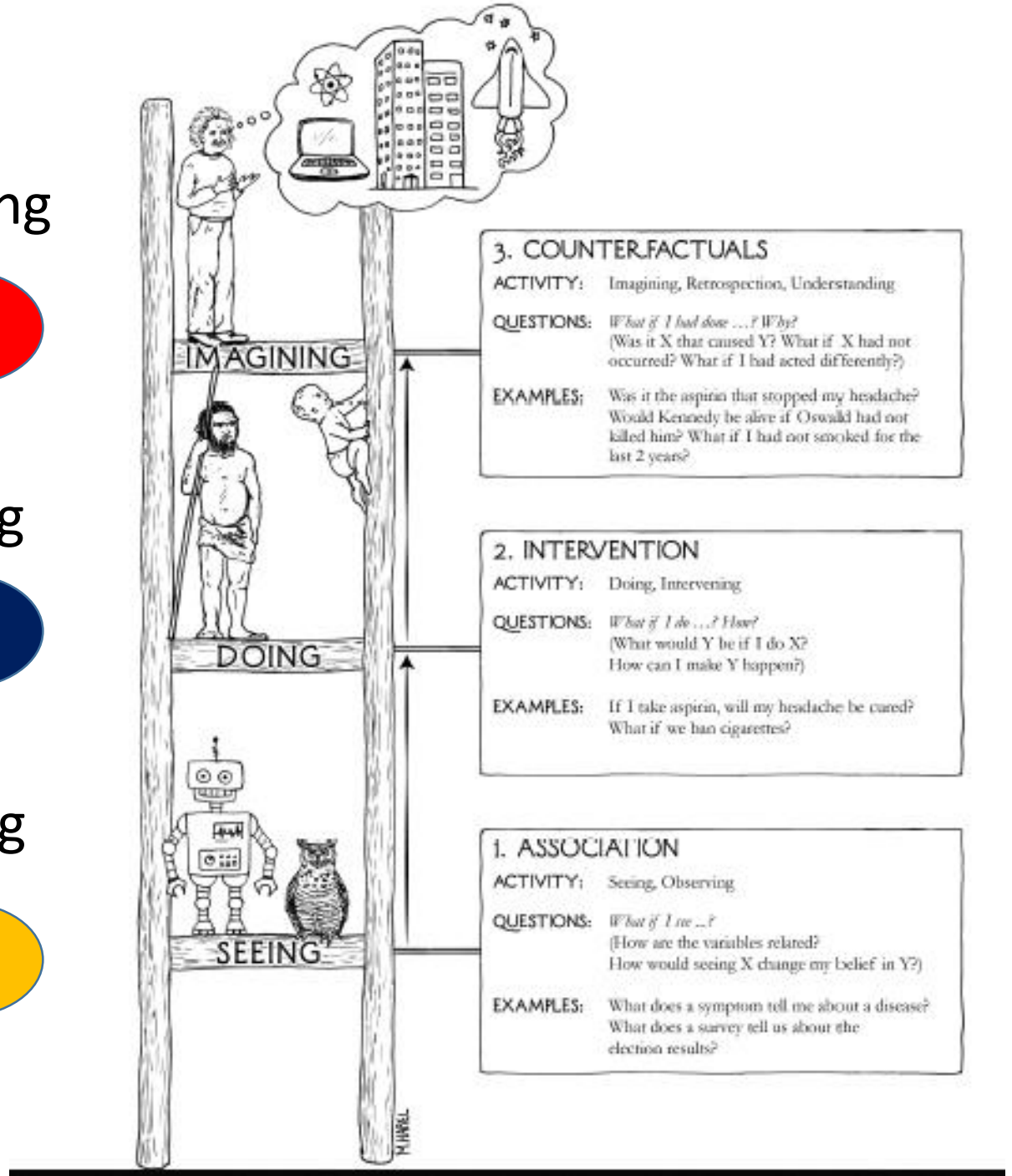
3

Doing

2

Seeing

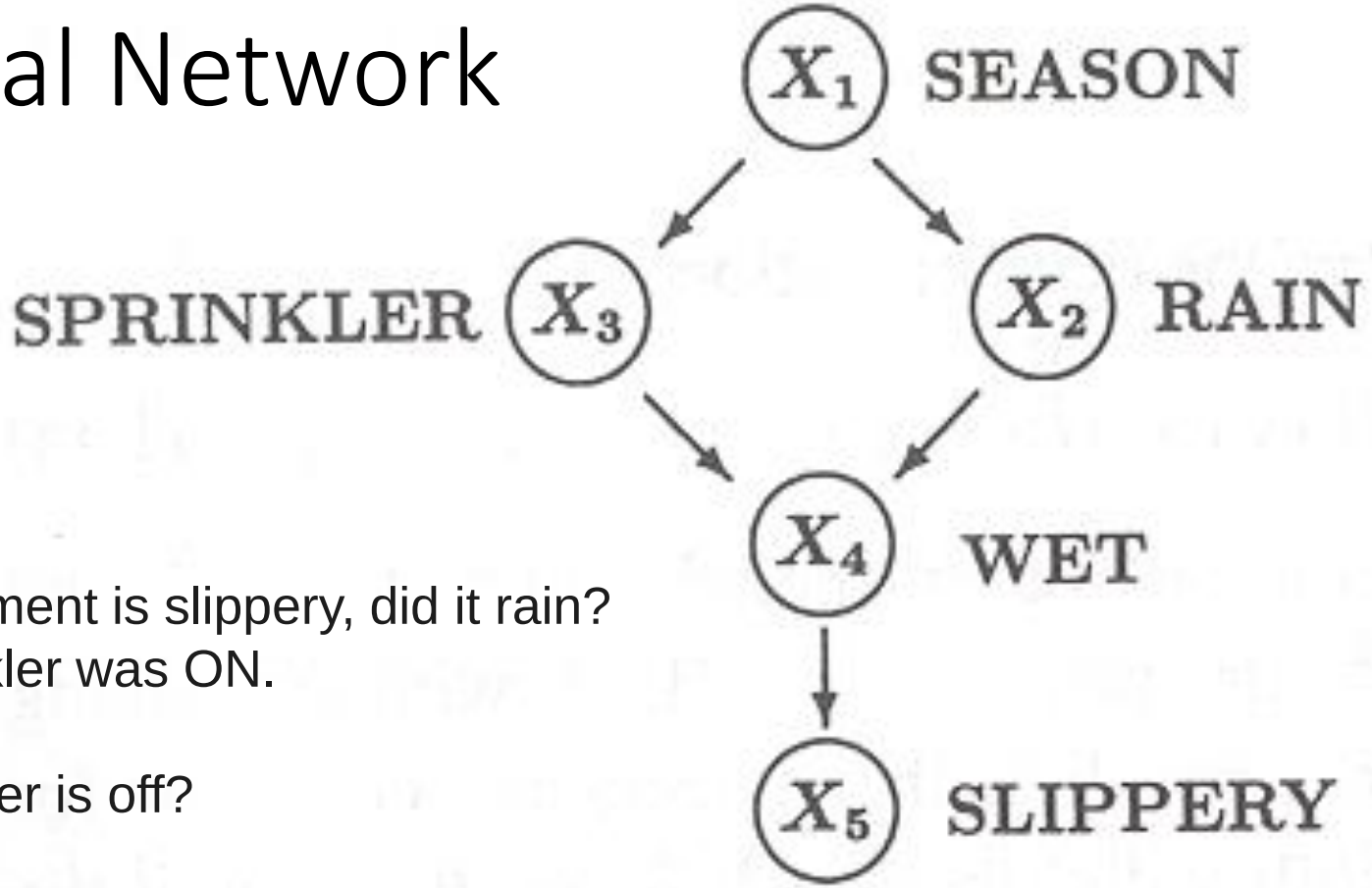
1



Graph Terminology

- Nodes – vertices on a graph (X_i)
- Edge – line or arrow connecting two nodes
- Adjacent – two variables connected by an edge
- Path – sequence of edges (p)
- Directed Path – arrows at the end of every edge
- Acyclic – No loops
- DAG – directed acyclic graph (G)
- Parents, children, descendants, etc.

A Structural Causal Network



Q1: If the season is dry, and the pavement is slippery, did it rain?
A1: Unlikely, it is more likely the sprinkler was ON.

Q2: But what if we see that the sprinkler is off?
A2: Then it is more likely that it rained

Q3: Do you mean that if we actually turn the sprinkler off, the rain will be more likely?
A3: No, the likelihood of rain would remain the same

From Bayesian Networks to Causal Graphs

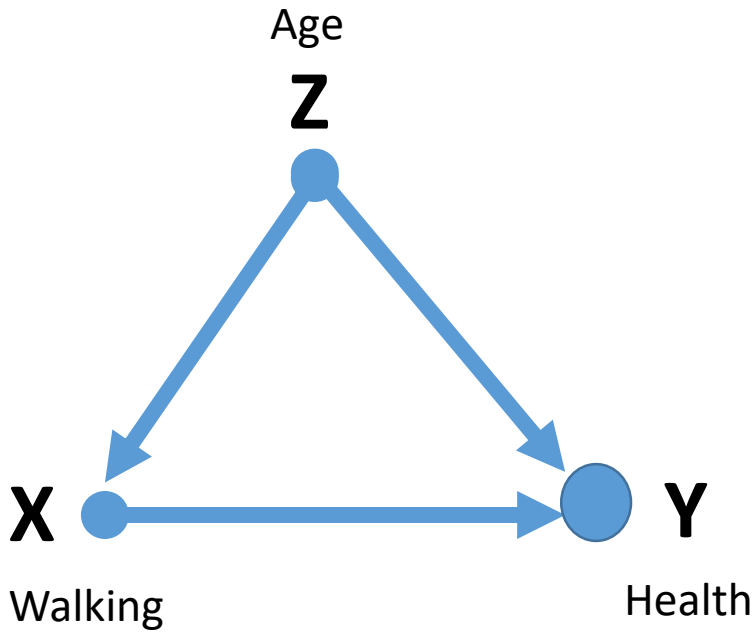
A DAG G is a **causal graph or structural causal network (SCN)** if,

for each node X_i, with parents PA_i,

we have $X_i = f_i(PA_i, e_i)$,

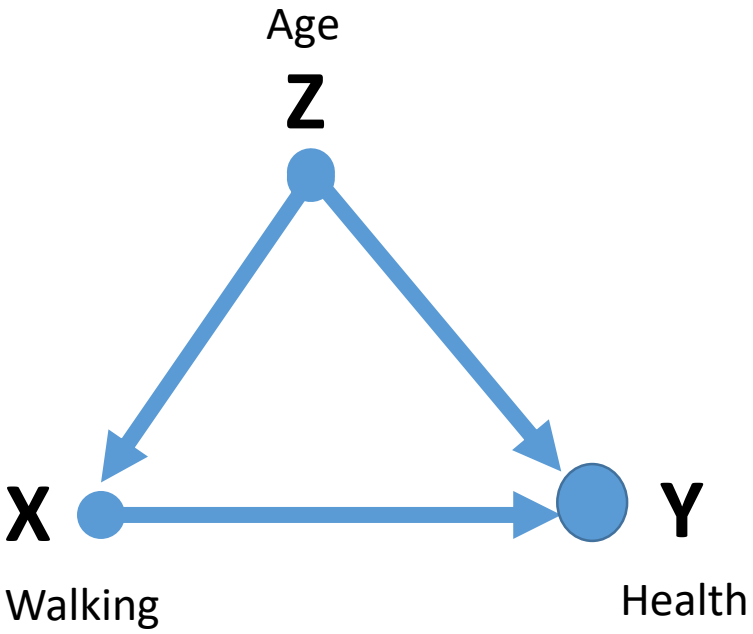
e_i independent random variables

and f_i a deterministic function.



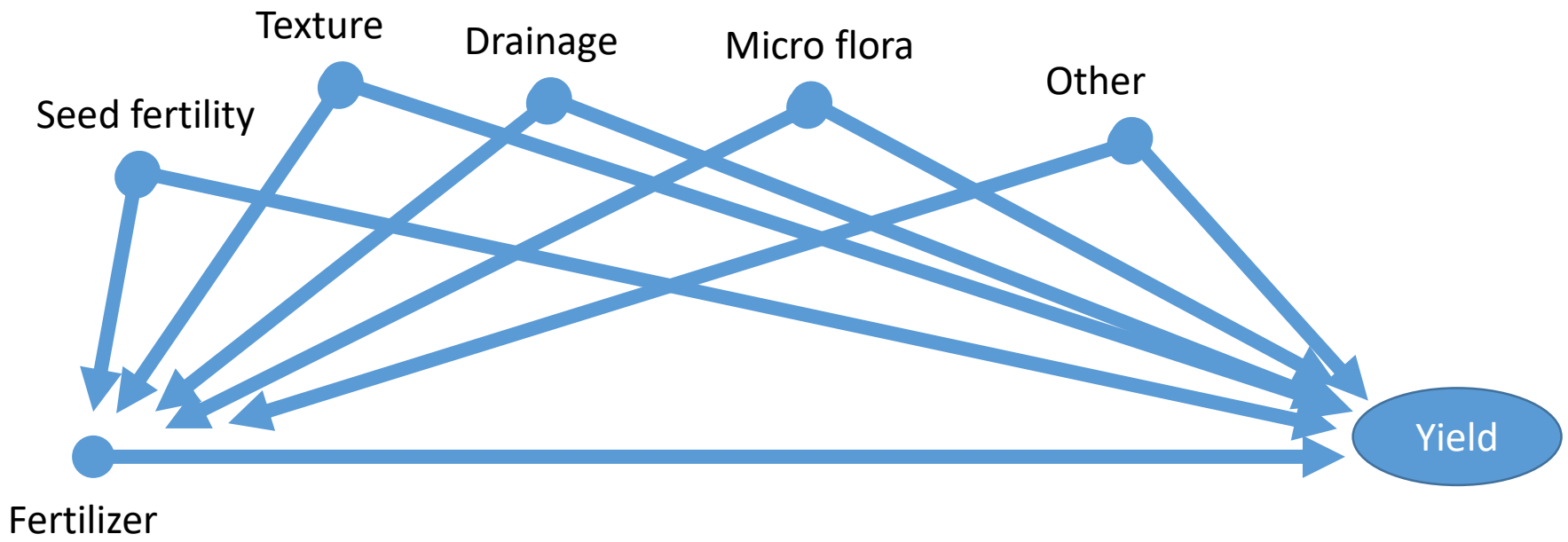
A Structural Causal Network

Z is a confounder of the causal relationship between X and Y

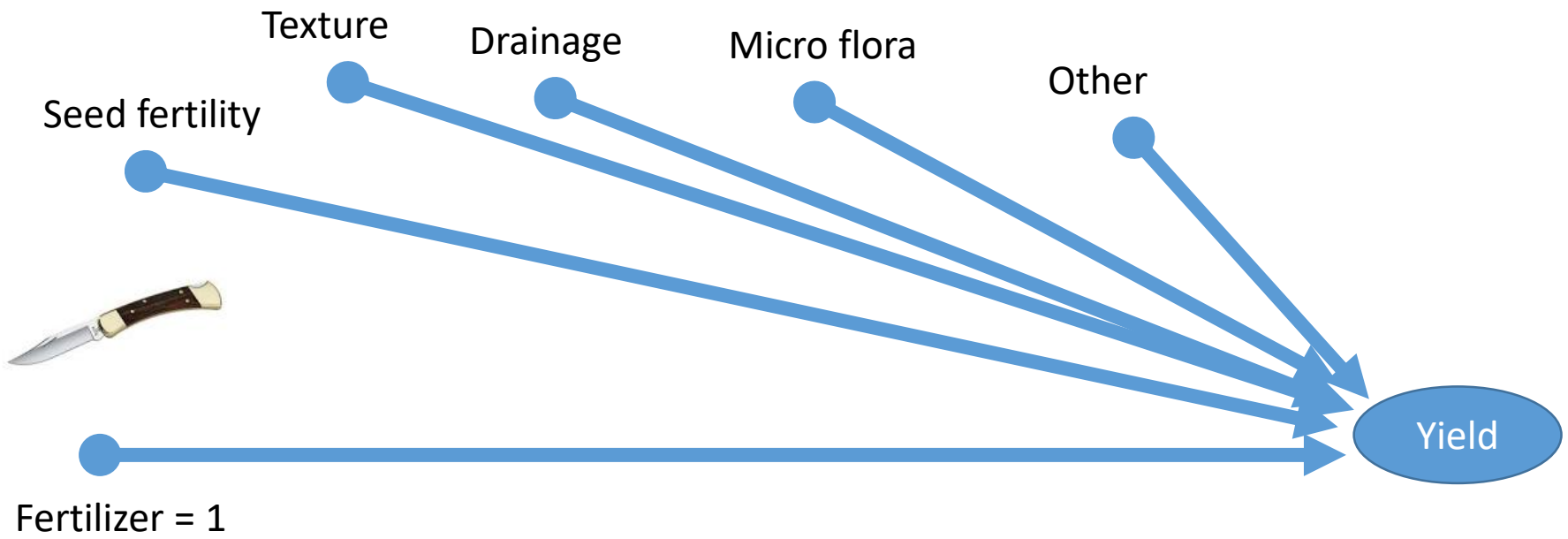


Is walking good for your health?

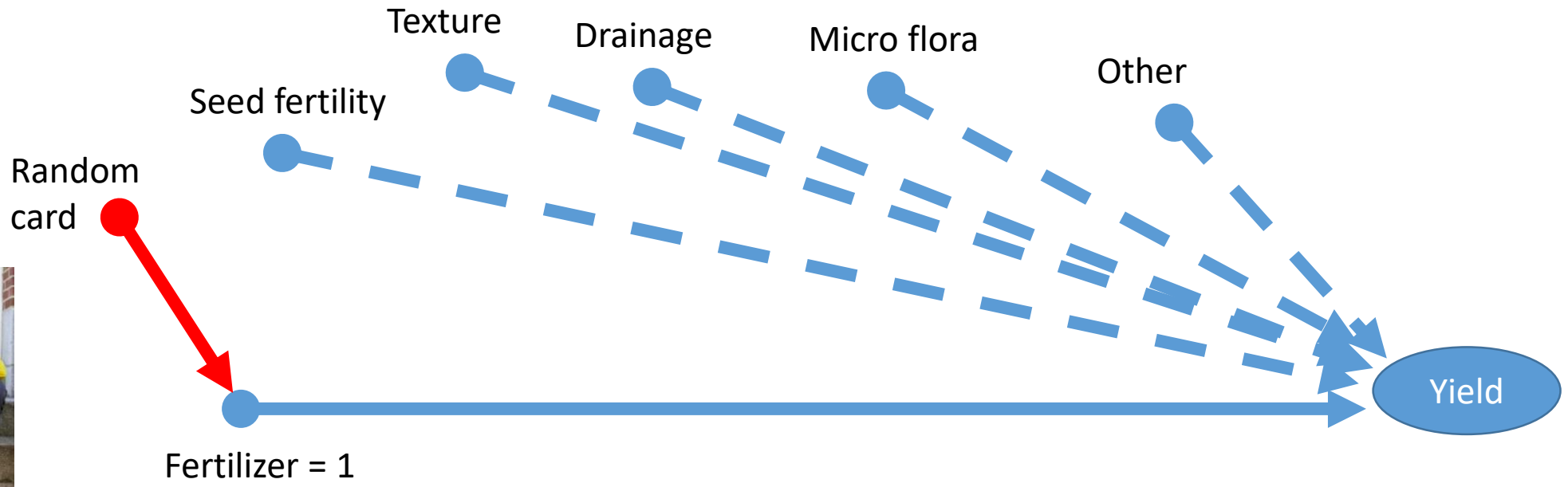
What we have

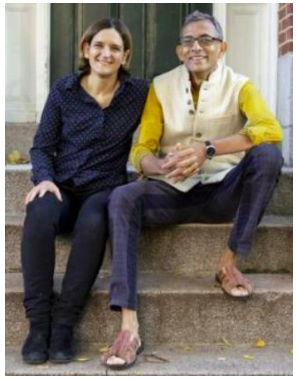

Texture
Drainage
Micro flora
Other
Seed fertility
Yield
Fertilizer

What we want


Texture
Drainage
Micro flora
Other
Seed fertility
Yield
Fertilizer = 1

What we get with randomizaton



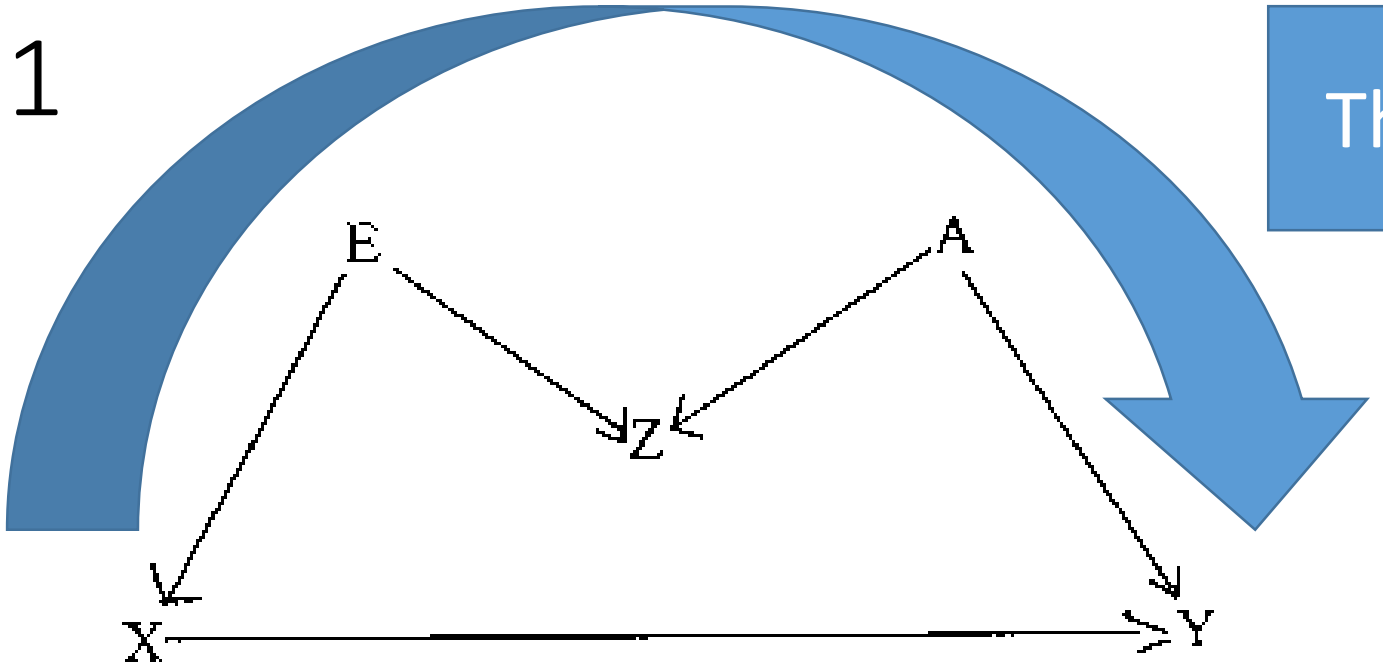


https://foreignpolicy.com/2019/10/22/economics-development-rcts-esther-duflo-abhijit-banerjee-michael-kremer-nobel/

Abhijit Banerjee and Esther Duflo: The Nobel couple fighting poverty

The team pioneered “randomized controlled trials”, or RCTs, in economics. https://www.bbc.com/news/world-asia-india-50048519

Example 1

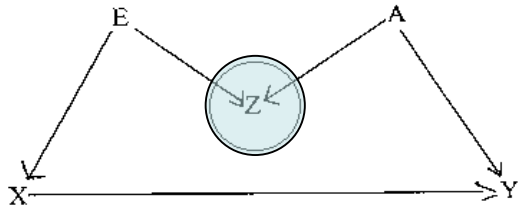


The M-bias

X = exposure of interest
Y = disease
E = education
Z = type of car owned by patient
A = age

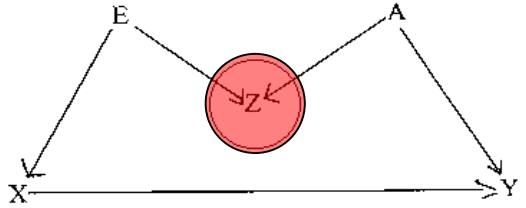
Back door path blocked by Z so that there is no need to control for anything

The M-bias



X = exposure of interest
Y = disease
E = education
Z = type of car owned by patient
A = age

- The back-door criterion suggests that the effect of X and Y is not confounded by A, Z or E.
- The only arrow into X is the one traversing (X, E, Z, A, Y) and this path contains two arrows pointing head-to-head at Z.



X = exposure of interest
Y = disease
E = education
Z = type of car owned by patient
A = age

The M-bias

Consequences of Adjusting for Z

- Statistically adjusting for Z, when estimating the effect of X on Y, will give a biased effect estimate.
- Thus, one should not necessarily “control” for every variable that is related to both the disease and the treatment/exposure of interest.

Example 2

The expanded M-bias

What variables one needs to adjust to get unconfounded effect of X_i (risk variable) on X_j (outcome).

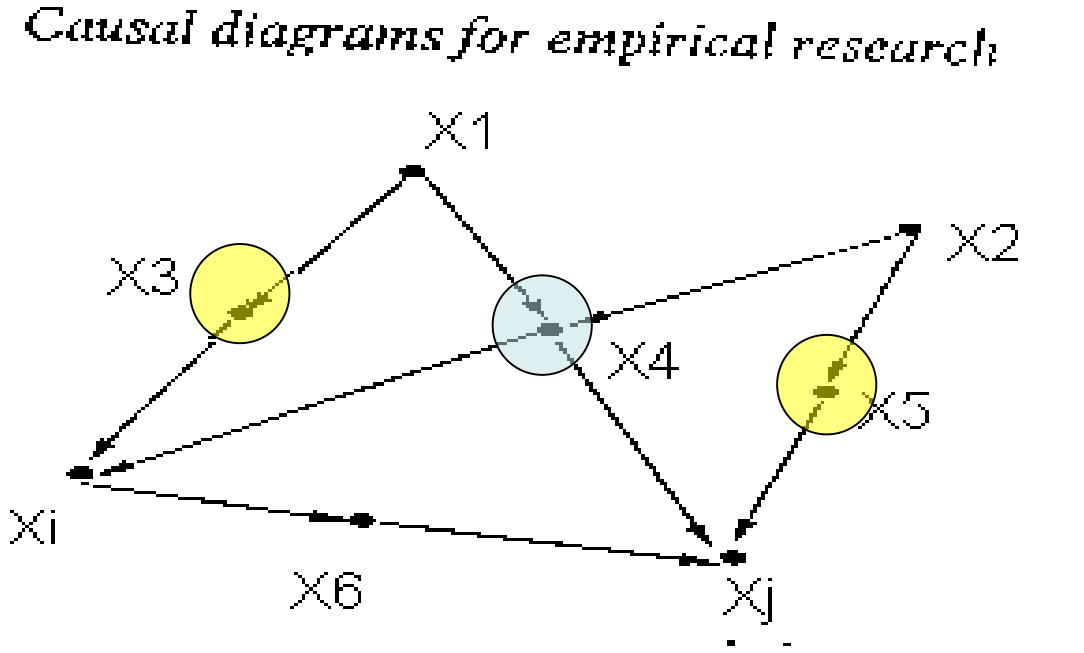
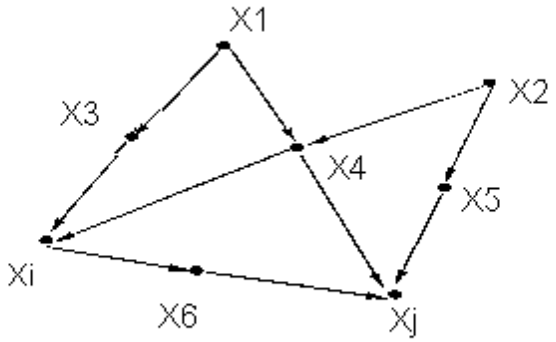
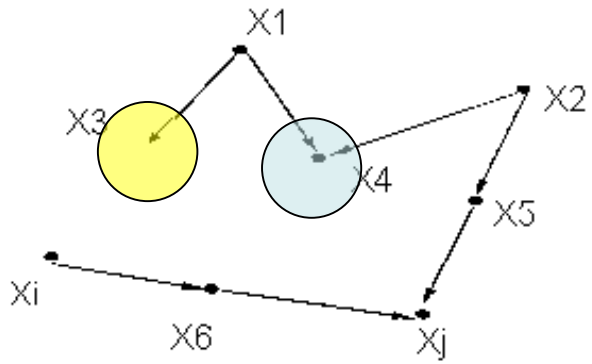


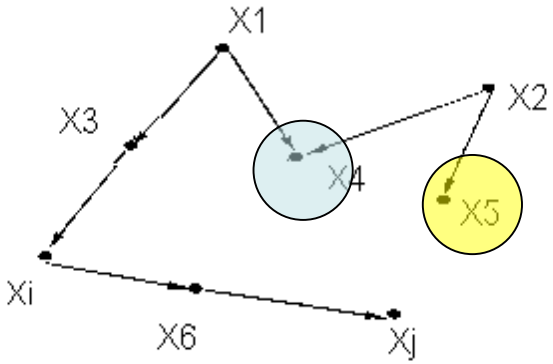
Fig. 2. A diagram representing the back-door criterion; adjusting for variables $\{X_3, X_4\}$ or $\{X_4, X_5\}$ yields a consistent estimate of $pr(x_j|\hat{x}_i)$.



The expanded M-bias

In this case, one could adjust for $\{X_3,X_4\}$ or $\{X_4,X_5\}$ but not just for $\{X_4\}$.





Rules to control information from A to C

1. In a **chain**, A -> B -> C, controlling for B prevents information about A, through B, from getting to C

2. In a **fork**, A <- B -> C, controlling for B prevents information about A, through B, from getting to C

3. In a **collider**, A -> B <- C, the opposite holds. A and C start independent so that information about A tells nothing about C, but, controlling for B, causes information, through B, to flow

4. Controlling for descendants is partially controlling for the variable itself. Controlling for a descendant of a collider partially opens the information flow.

d (directional) - separation vs. d - connected

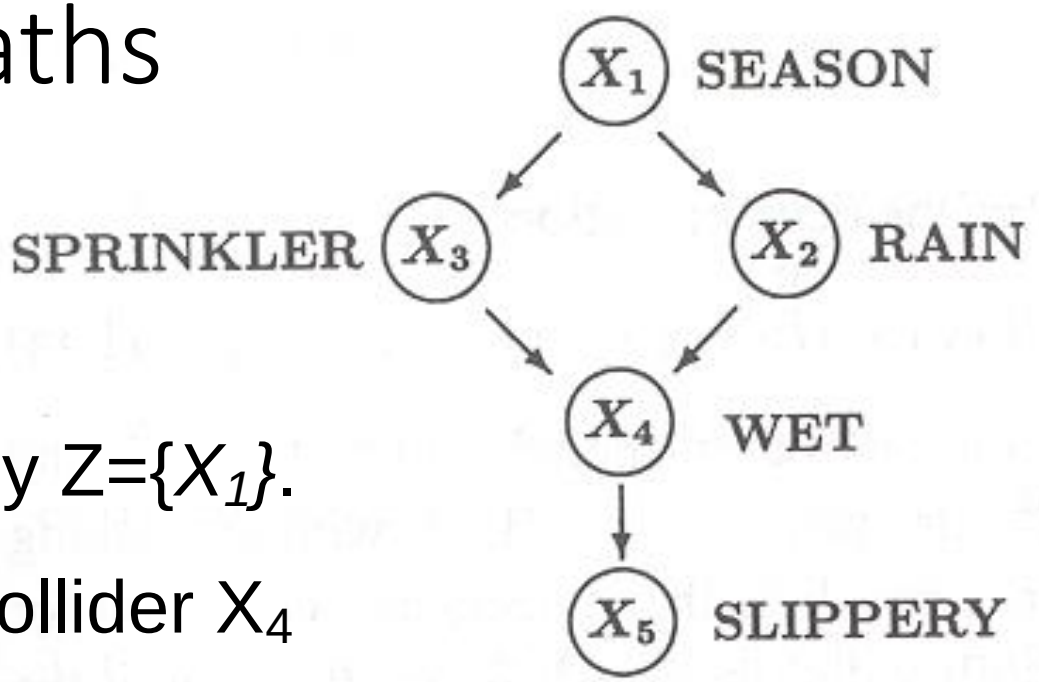
A path, p, is said to be **d-separated** by a set of nodes Z if and only if:

1. *p* contains a fork $i \leftarrow m \rightarrow j$ or a chain $i \rightarrow m \rightarrow j$ such that the middle node *m* is in *Z*, or
2. *p* contains an inverted fork (or collider) $i \rightarrow m \leftarrow j$ such that the middle node *m* is not in *Z* and such that no descendent of *m* is in *Z*.

A set Z is said to d-separate X from Y if and only if Z blocks every path from a node in X to a node in Y.

A pair of d-separated nodes are independent.

Example of d-separated paths



$X=\{X_2\}$ and Y=$\{X_3\}$ are d-separated by Z=$\{X_1\}$.

The path $X_2 – X_4 – X_3$ is blocked by collider X_4

However, X and Y are not d-separated by Z'=$\{X_1, X_5\}$ since X_5 is a descendant of the collider, X_4.

So, knowing X_5 causes X_2 and X_3 to be dependent

Interventions in Causal Graphs

When we **intervene** on a variable in a model, we fix its value and change the system. Values of other variables often change as a result. When we **condition** on a variable, we change nothing only narrow focus on a subset of cases.

The causal effect of a variable (node) X_i can be defined as how the outcome, Y, changes when this variable is set to some value, thereby breaking the influence of predecessors.

This basic insight translates into the G-estimation algorithm of Robins (1986).

After intervening in the graph, by setting $X_i = x_i'$, then the joint distribution of the data becomes:

$$P(x_1, x_2, \ldots, x_n \mid x_{i'}) = \begin{cases} P(x_1, x_2, \ldots, x_n) / P(x_i \mid pa_i) & \text{if } x_i = x_{i'} \\ 0 & \text{if } x_i \neq x_{i'} \end{cases}$$

ACE=$P(Y=1|do(X=1))-P(Y=1|do(X=0))$

Rule 1: Ignoring observations

$P(y \mid do\{x\}, z, w) = P(y \mid do\{x\}, w)$

if $(Y \perp\!\!\!\perp Z|X,W)_{G_{\overline{X}}}$

Rule 2: Action/observation exchange

$P(y \mid do\{x\}, do\{z\}, w) = P(y \mid do\{x\},z,w)$

if $(Y \perp\!\!\!\perp Z|X,W)_{G_{\overline{X}\underline{Z}}}$

Rule 3: Ignoring actions

$P(y \mid do\{x\}, do\{z\}, w) = P(y \mid do\{x\}, w)$

if $(Y \perp\!\!\!\perp Z|X,W)_{G_{\overline{X}\overline{Z(W)}}}$

Do calculus

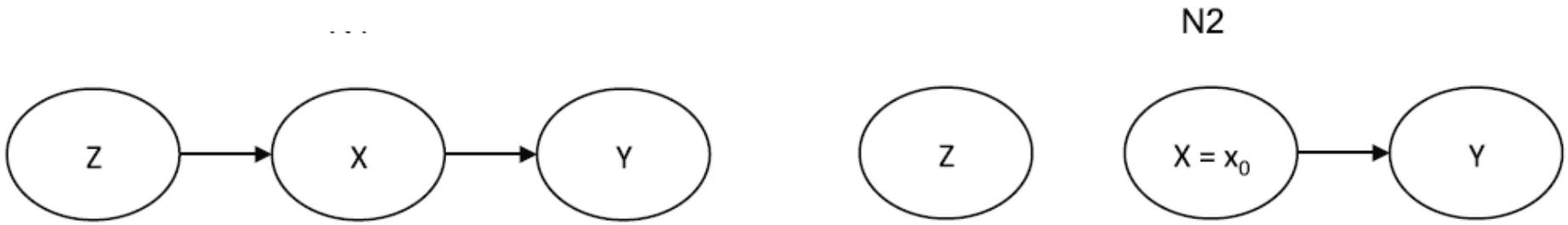


FIGURE 1. Network Pre and Post Intervention.

Adjustment formula $P(Y = y|do(X = x)) = SumP(Y = y|X = x, Z = z)P(Z = z)$

$$\text{ACE}=P(Y=1|do(X=1))-P(Y=1|do(X=0))$$

Recovery rates with and without drug [(Y=1)/n]

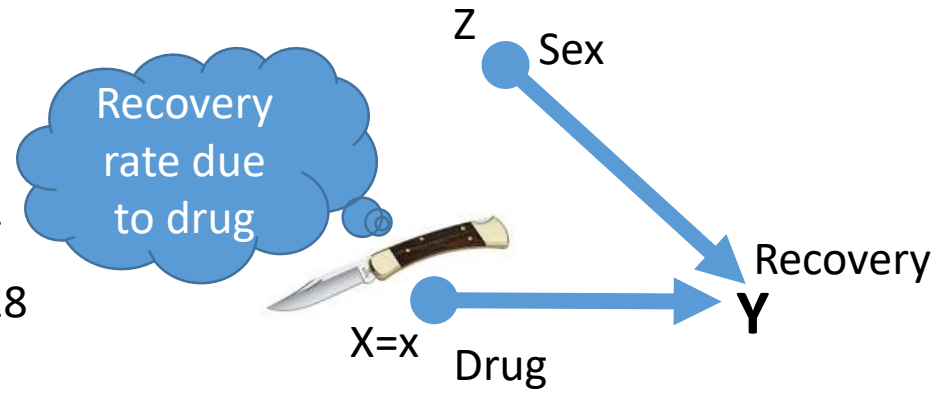
	Drug (X=1)	No Drug (X=0)
Men (Z=0)	81/87 (93%)	234/270 (87%)
Women (Z=1)	192/263 (73%)	55/80 (69%)
Total	273/350 (78%)	289/350 (83%)

Do calculus

$P(Y=1|do(X=1))$=P(Y=1|X=1, Z=1)P(Z=1)+P(Y=1|X=1, Z=0)P(Z=0)

$P(Y=1|do(X=1))$=(0.93(87+270))/700 + (0.73(263+80)/700) = 0.832

$P(Y=1|do(X=0))$=(0.87(87+270))/700 + (0.69(263+80)/700) = 0.7818



ACE=$P(Y=1|do(X=1))-P(Y=1|do(X=0))$= 0.832 - 0.7818 = 0.0502

Adjustment formula $P(Y = y|do(X = x)) = SumP(Y = y|X = x, Z = z)P(Z = z)$

ACE=$P(Y{=}1|do(X{=}1))$-$P(Y{=}1|do(X{=}0))$

The Causal Effect Rule

Do calculus

Given a graph G in which a set of variables PA are designed as the parents of X, the causal effect of X on Y is given by

$$P(Y = y|do(X = x) = SumP(Y = y|X = x, PA = z)P(PA = z)$$

Where z ranges over all the combinations of values that the variable PA can take.

$$P(Y = y|do(X = x) = SumP(X = x, Y = y, PA = z)/P(X = x|PA = z)$$

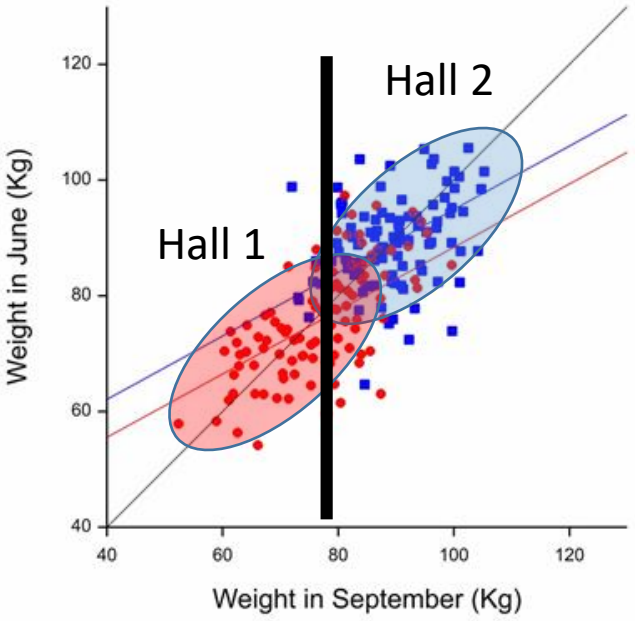
Propensity score = $P(X = x|PA = z)$

Adjustment formula $P(Y = y|do(X = x)) = SumP(Y = y|X = x, Z = z)P(Z = z)$

Lord's Paradox and Causal Graphs

"A large university is interested in investigating the effects on the students of the diet provided in the university dining halls. Various types of data are gathered. In particular, the sex and weight of each student at the time of his arrival in September and his weight the following June are recorded." (Lord, 1967). Lord posits two statisticians who use different but respected statistical methods to reach opposite conclusions about the effects of the diet provided in the university dining halls on students' weights.

One statistician does not adjust for initial weight or sex; using analysis of variance (ANOVA), and treating **gain scores** (June - September) as the outcome, he finds no significant difference between dining halls and states that there is no evidence of any effect of diet on student weights. The second statistician adjusts for initial weight; using analysis of covariance (ANCOVA), and treating **June weights as the outcome**, he finds a significant difference between the two dining halls.



	Baseline	Outcome
Diet A (Hall 1)	$\bar{X}_A$	$\bar{Y}_A$
Diet B (Hall 2)	$\bar{X}_B$	$\bar{Y}_B$

$Y_B - Y_A = X_B - X_A = D.$ Is there an effect?

$(Y_B - Y_A) - r(X_B - X_A) = D - rD = (1 - r)D,$

$(Y_B - X_B) - (Y_A - X_A) = (Y_B - Y_A) - (X_B - X_A) = D - D = 0.$ ← Who is right?

In neither halls students gain weight but in each stratum Hall 2 tend to gain more weight than Hall 1

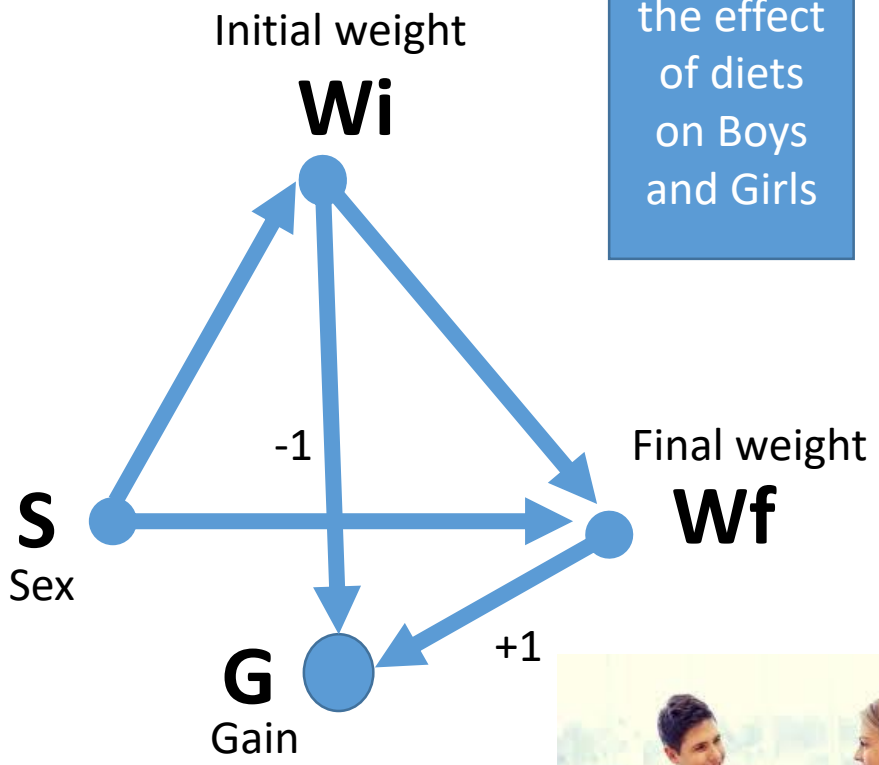
Lord's Paradox and Causal Graphs (Original)

Consult the story behind the data. Account for S. The variable of interest is G.

$$G = Wf - Wi$$

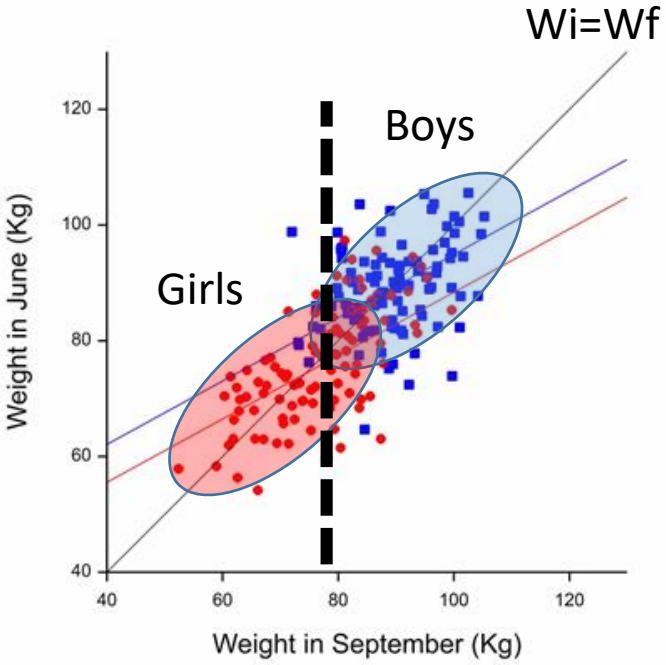
No backdoor between S and G need to be blocked so the aggregated data provides the answer (statistician one).

Wi is a mediating variable of S and G, and controlling for Wi provides the direct effect of S on G.



What is the effect of diets on Boys and Girls

Sex strongly affects the percentages of students in each stratum

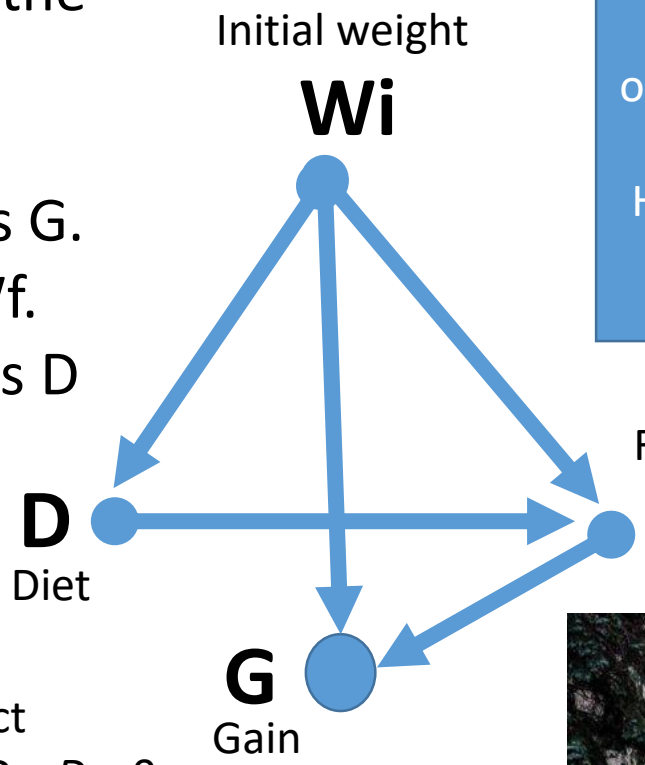


http://causality.cs.ucla.edu/blog/index.php/2019/08/13/lords-paradox-the-power-of-causal-thinking/

Lord's Paradox and Causal Graphs (Adapted)

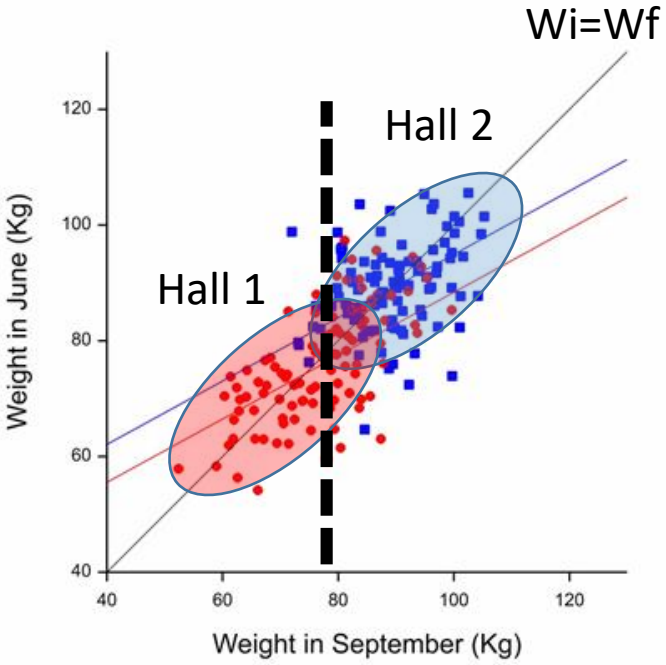
Consider another story behind the data. Account for Hall (Diet).

Again, the variable of interest is G. Wi is a confounder for D and Wf. Controlling for Wi de-confounds D and Wf, as well as D and G.



What is the effect (G) on students of diet in Hall 1 and Hall 2

The Hall very strongly affects the percentages of students in each stratum



$P(Gain|Diet=A) = P(Gain|Diet=B) \neq P(Gain|do(Diet=A)) = P(Gain|do(Diet=B))$

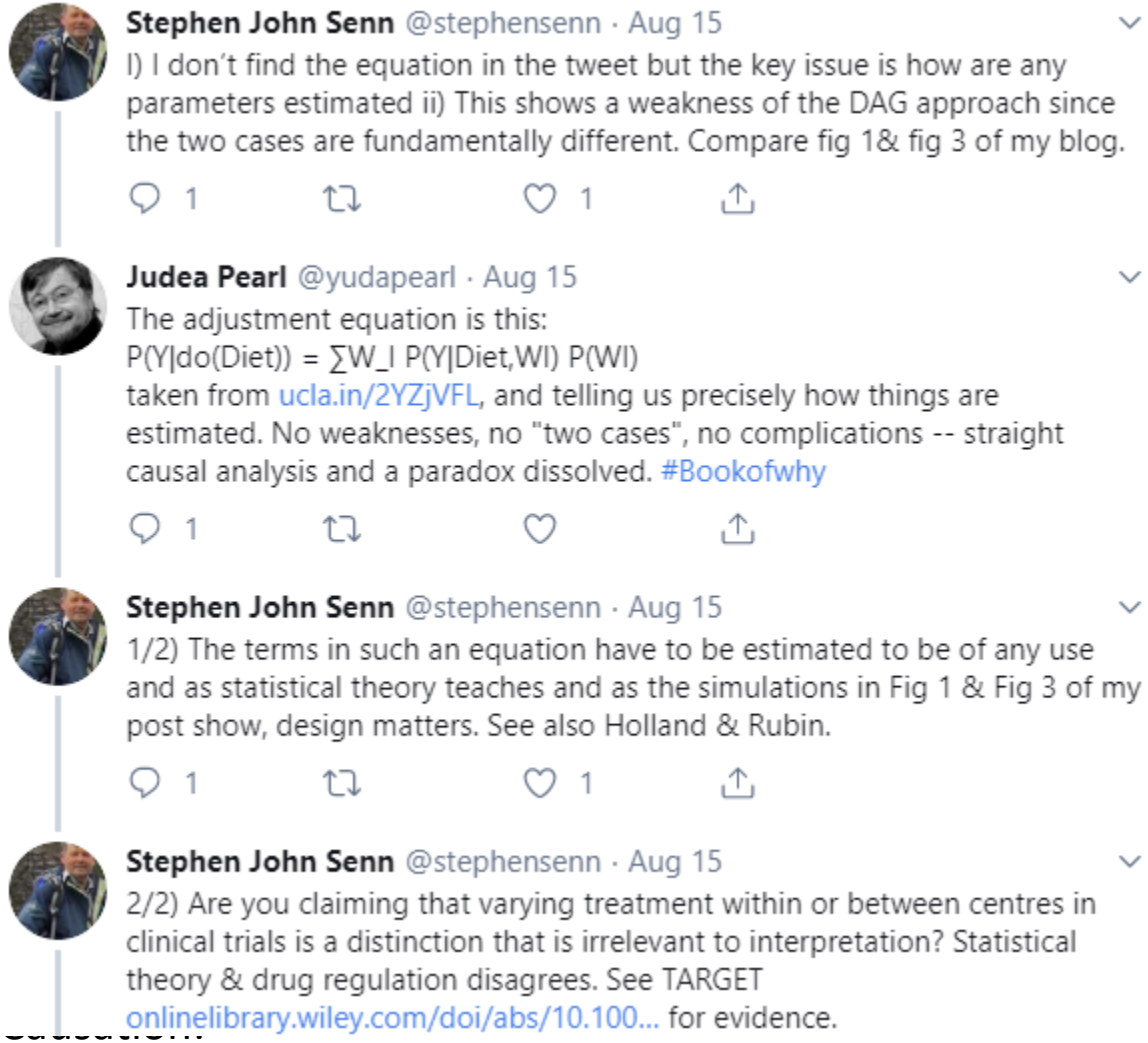
Association:
Switching from Diet A to Diet B has no effect
$(Y_B - X_B) - (Y_A - X_A) = (Y_B - Y_A) - (X_B - X_A) = D - D = 0.$

$P(G|do(Diet) = \sum\{Wi\} P(G|Diet, Wi) P(Wi)$

Causation:
Comparing gains in Diet A versus Diet B, for students with same intial weight, shows higher gain with Diet B

https://errorstatistics.com/2019/08/02/s-senn-red-herrings-and-the-art-of-cause-fishing-lords-paradox-revisited-guest-post/
https://errorstatistics.com/2018/11/11/stephen-senn-rothamsted-statistics-meets-lords-paradox-guest-post/amp/

Lord's Paradox and Causal Graphs



Stephen John Senn @stephensenn · Aug 15
I) I don't find the equation in the tweet but the key issue is how are any parameters estimated ii) This shows a weakness of the DAG approach since the two cases are fundamentally different. Compare fig 1& fig 3 of my blog.
1 1

Judea Pearl @yudapearl · Aug 15
The adjustment equation is this:
P(Y|do(Diet)) = ∑W_I P(Y|Diet,WI) P(WI)
taken from ucla.in/2YZjVFL, and telling us precisely how things are estimated. No weaknesses, no "two cases", no complications -- straight causal analysis and a paradox dissolved. #Bookofwhy
1

Stephen John Senn @stephensenn · Aug 15
1/2) The terms in such an equation have to be estimated to be of any use and as statistical theory teaches and as the simulations in Fig 1 & Fig 3 of my post show, design matters. See also Holland & Rubin.
1 1

Stephen John Senn @stephensenn · Aug 15
2/2) Are you claiming that varying treatment within or between centres in clinical trials is a distinction that is irrelevant to interpretation? Statistical theory & drug regulation disagrees. See TARGET onlinelibrary.wiley.com/doi/abs/10.100... for evidence.

```
BLOCKSTRUCTURE Hall/Student
TREATMENTSTRUCTURE Diet
COVARIATE Base
ANOVA Weight
```

Back door adjustment formula

Average causal effect of an interventions by first estimating its effect at each level of the de-confounder.

Then, compute a weighted average of those levels, where each level is weighted according to its prevalence in the population.

Comparing gains in Diet A versus Diet B, for students with same intial weight, shows higher gain with Diet B

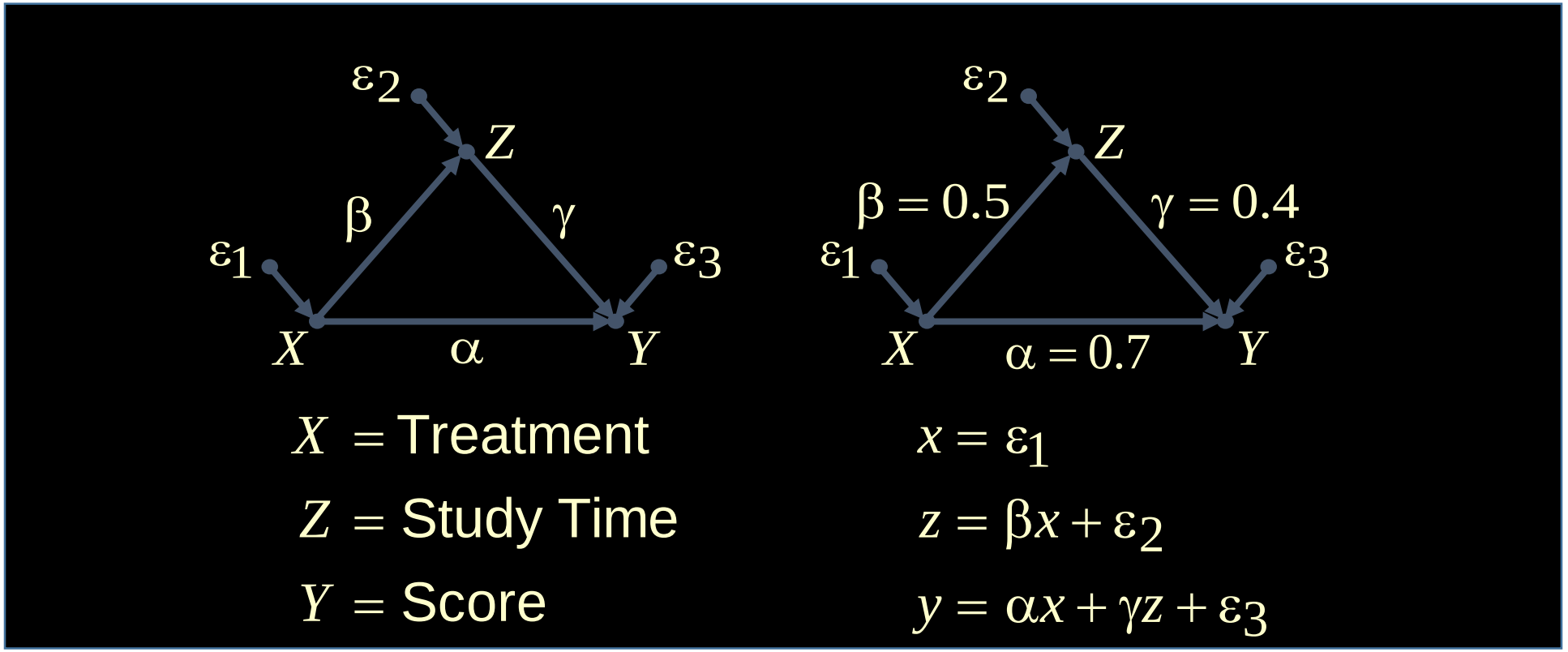
A Structural Causal Model

Definition: A structural causal model is a 4-tuple $\langle V, U, F, P(u) \rangle$, where

- $V = \{V_1,\ldots,V_n\}$ are endogeneus variables
- $U = \{U_1,\ldots,U_m\}$ are background variables
- $F = \{f_1,\ldots, f_n\}$ are functions determining V,
  $v_i = f_i(v, u)$ e.g., $y = \alpha + \beta x + u_Y$
- $P(u)$ is a distribution over U

$P(u)$ and F induce a distribution $P(v)$ over observable variables

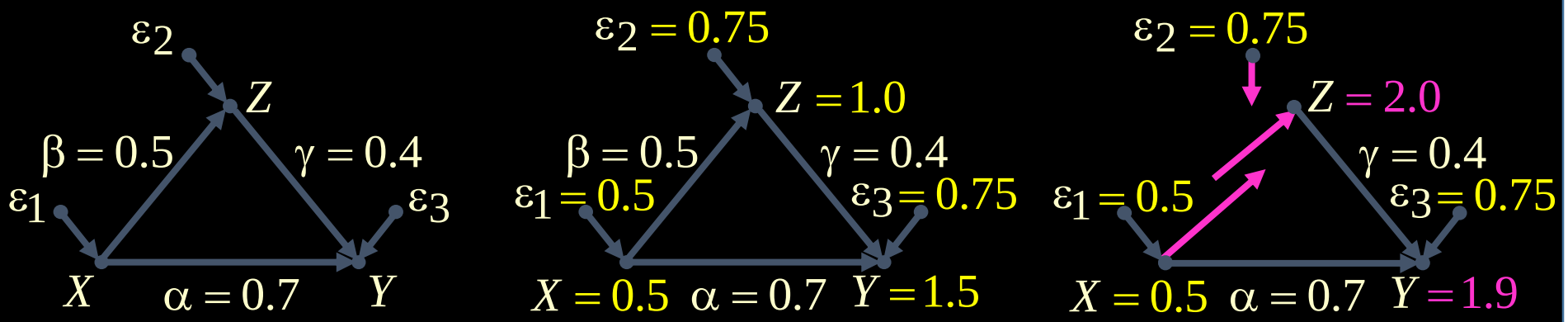
A Structural Causal Network



Data shows: α = 0.7, β = 0.5, γ = 0.4

A student named Joe, measured X=0.5, Z=1, Y=1.5

Q_1: What would Joe's score be, had he doubled his study time?



Q_1: What would Joe's score be had he doubled his study time?

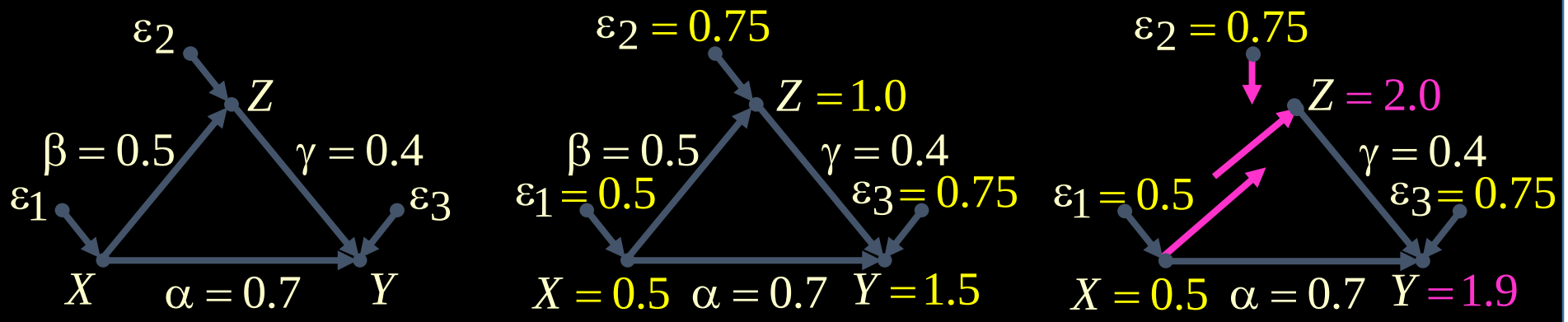
Answer: Joe's score would be 1.9

Or,

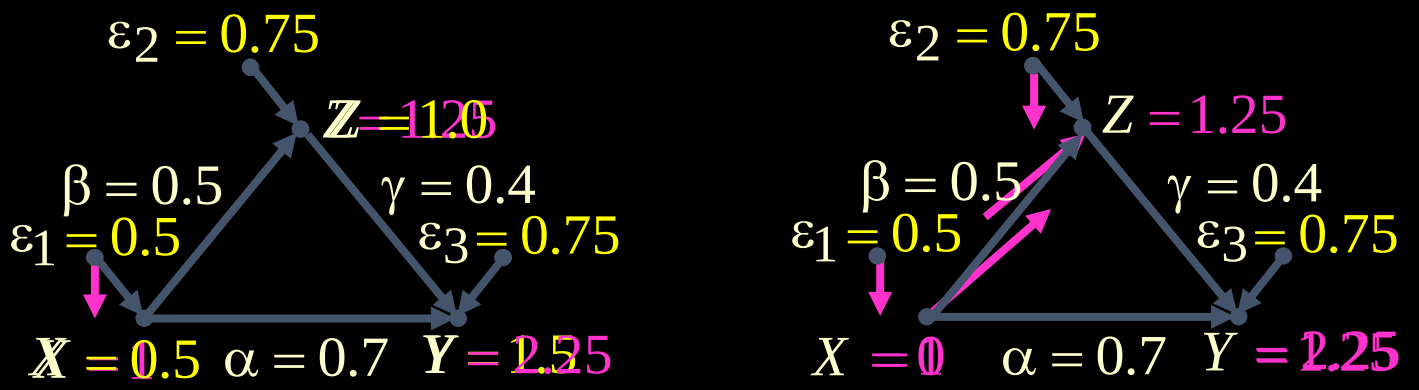
In counterfactual notation:

$$Y_{Z=2}(u) = 1.9$$

"do" calculus example



Q_2: What would Joe's score be, had the treatment been 0, and had he studied at whatever level he would have studied had the treatment been 1?



Internal and external validity

RANDOMIZED EXPERIMENTS AND OBSERVATIONAL STUDIES: CAUSAL INFERENCE IN STATISTICS

PAUL R. ROSENBAUM

ABSTRACT. This talk describes the theory of causal inference in randomized experiments and nonrandomized observational studies, using two simple theoretical/actual examples for illustration. Key ideas: causal effects, randomized experiments, adjustments for observed covariates, sensitivity analysis for unobserved covariates, reducing sensitivity to hidden bias using design strategies.

1. SEVEN KEY CONTRIBUTIONS TO CAUSAL INFERENCE

1.0.1. ***Ronald A. Fisher (1935).*** *The Design of Experiments.* Edinburgh: Oliver & Boyd. Although Fisher had discussed his randomized experiments since the early 1920's, his most famous discussion appears in Chapter 2 of this book, in which Fisher's exact test for a 2×2 table is derived from randomization alone in the experiment of the 'lady tasting tea.'

1.0.2. ***Jerzy Neyman (1923).*** On the application of probability theory to agricultural experiments. Essay on principles. Section 9. (In Polish) Roczniki *Nauk Roiniczych,* Tom X, pp1-51. Reprinted in English in *Statistical Science*, 1990, 5, 463-480, with discussion by T. Speed and D. Rubin. In this paper, Neyman writes the effects caused by treatments as comparisons of potential outcomes under alternative treatments.

Generalizability and transportability

TECHNICAL REPORT
R-452
July 2015

DE GRUYTER J. Causal Infer. 2015; 3(2): 259–266

Causal, Casual and Curious

Judea Pearl*

Generalizing Experimental Findings

DOI 10.1515/jci-2015-0025

Abstract: This note examines one of the most crucial questions in causal inference: "How generalizable are randomized clinical trials?" The question has received a formal treatment recently, using a non-parametric setting, and has led to a simple and general solution. I will describe this solution and several of its ramifications, and compare it to the way researchers have attempted to tackle the problem using the language of ignorability. We will see that ignorability-type assumptions need to be enriched with structural assumptions in order to capture the full spectrum of conditions that permit generalizations, and in order to judge their plausibility in specific applications.

Keywords: generalizability, transportability, selection bias, admissibility, ignorability

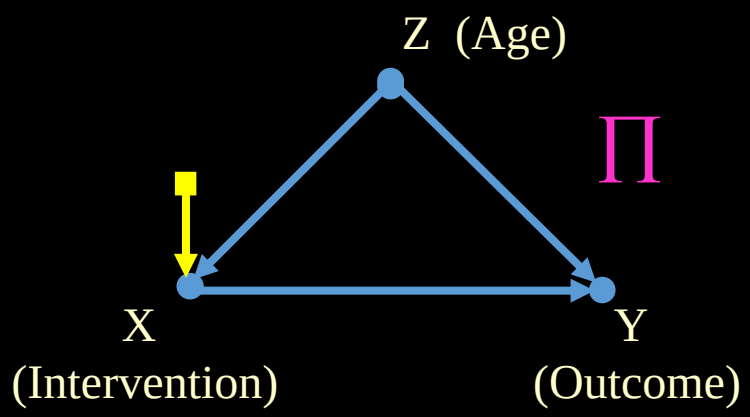
1 Transportability and selection bias

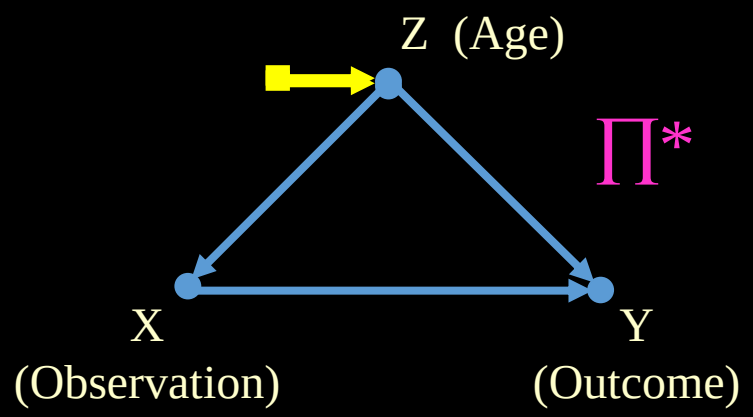
The long-standing problem of generalizing experimental findings from the trial sample to the population as a whole, also known as the problem of "sample selection-bias" [1, 2], has received renewed attention in the past decade, as more researchers come to recognize this bias as a major threat to the validity of experimental findings in both the health sciences [3] and social policy making [4]. Since participation in a randomized trial cannot be mandated, we cannot guarantee that the study population would be the same as the population of interest. For example, the study population may consist of volunteers, who respond to financial and medical incentives offered by pharmaceutical firms or experimental teams, so, the distribution of outcomes in the study may differ substantially from the distribution of outcomes under the policy of interest.

Query of interest: $Q = P^*(y \mid do(x))$

Target population

Arkansas Survey data available	New York Survey data Resembling target	Los Angeles Survey data Youngish population
Boston Age not recorded Mostly successful lawyers	San Francisco High post-treatment blood pressure	Texas Mostly Spanish subjects High attrition
Toronto Randomized trial College students	Utah RCT, paid volunteers, unemployed	Wyoming RCT, young athletes





Experimental study in LA

Measured: $P(x, y, z)$

$P(y \mid do(x), z)$

Observational study in NYC

Measured: $P^*(x, y, z)$

$P^*(z) \neq P(z)$

Needed: $P^*(y \mid do(x)) = ? = \sum_z P(y \mid do(x), z) P^*(z)$

Transport Formula (calibration): $F(P, P_{do}, P^*)$

https://cran.r-project.org/web/packages/causaleffect/causaleffect.pdf

Pearl, J. 2009. *Causality: Models, Reasoning, and Inference*. Cambridge University Press

JASA, 2000, 95(450), pp. 441-448

Causal Inference Without Counterfactuals

A. P. DAWID

A popular approach to the framing and answering of causal questions relies on the idea of counterfactuals: outcomes that would have been observed had the world developed differently; for example, if the patient had received a different treatment. By definition, one can never observe such quantities, nor assess empirically the validity of any modeling assumptions made about them, even though one's conclusions may be sensitive to these assumptions. Here I argue that for making inference about the likely effects of applied causes, counterfactual arguments are unnecessary and potentially misleading. An alternative approach, based on Bayesian decision analysis, is presented. Properties of counterfactuals *are* relevant to inference about the likely causes of observed effects, but close attention then must be given to the nature and context of the query, as well as to what conclusions can and cannot be supported empirically. In particular, even in the absence of statistical uncertainty, such inferences may be subject to an irreducible degree of ambiguity.

KEY WORDS: Average causal effect; Causes of effects; Causation; Determinism; Effects of causes; Metaphysical model; Potential response; Treatment-unit additivity.

Fact–Fiction. Are counterfactuals to be regarded as genuine features of the external world, or are they purely theoretical terms?

Real–Instrumental. Can any inferences based on counterfactuals be allowed, or should they be restricted to those that could in principle be formulated without mention of counterfactuals?

Clear–Vague. Do counterfactual terms in a model have a clear relationship with meaningful aspects of the problem addressed? Can counterfactual constructions and arguments help to clarify understanding?

Helpful–Dangerous. Can use of counterfactuals streamline thinking and assist analyses, or do they promote misleading lines of argument and false conclusions?

Dimensions for assessing counterfactuals

Causality

Statistical Perspectives and Applications

Edited by

Carlo Berzuini • Philip Dawid • Luisa Bernardinelli
Statistical Laboratory, Centre for Mathematical Sciences
University of Cambridge, Cambridge, UK

Statistical Causality from a Decision-Theoretic Perspective

A. Philip Dawid

Centre for Mathematical Sciences, University of Cambridge, Cambridge CB3 0WB, United Kingdom; email: apd@statslab.cam.ac.uk

4.9 Postscript

The existence of a variety of different formal explications of statistical causality is somewhat embarrassing – we can only pray for the arrival of a messianic figure who (just as Kolmogorov did for probability theory) will sweep away the confusion and produce a single theory that everyone can accept. Meanwhile, let us put a positive gloss on this babel of different languages: since different people seem to find different approaches naturally appealing, there may be something for everyone. In that understanding I suggest that DT deserves careful attention from those who currently choose to think about statistical causality in other terms.

Journal of Business & Economic Statistics

ISSN: 0735-0015 (Print) 1537-2707 (Online) Journal homepage: https://www.tandfonline.com/loi/ubes20

Causal Interpretations of Black-Box Models

Qingyuan Zhao & Trevor Hastie

To cite this article: Qingyuan Zhao & Trevor Hastie (2019): Causal Interpretations of Black-Box Models, Journal of Business & Economic Statistics, DOI: 10.1080/07350015.2019.1624293

To link to this article: https://doi.org/10.1080/07350015.2019.1624293

Partial Dependence Plot and Causality

J. H. Friedman. Greedy function approximation: a gradient boosting machine. Annals of Statistics, 29(5):1189{1232, 2001.

- Given the output $g(x)$ of a machine learning algorithm (commonly estimates $\mathrm{E}[Y|X=x]$), the PDP of g on a subset of variables $X_{\mathcal{S}}$ is defined as (let $\mathcal{C}$ be the complement set of $\mathcal{S}$)

$$g_{\mathcal{S}}(x_{\mathcal{S}}) = \mathrm{E}_{X_{\mathcal{C}}}[g(x_{\mathcal{S}}, X_{\mathcal{C}})] = \int g(x_{\mathcal{S}}, x_{\mathcal{C}}) \mathrm{d}P(x_{\mathcal{C}}).$$

- This is different from the conditional expectation

$$\mathrm{E}[g(X_{\mathcal{S}}, X_{\mathcal{C}})|X_{\mathcal{S}} = x_{\mathcal{S}}] = \int g(x_{\mathcal{S}}, x_{\mathcal{C}}) \mathrm{d}P(x_{\mathcal{C}}|X_{\mathcal{S}} = x_{\mathcal{S}}).$$

Back-door adjustment

If the causal relationship of (X, Y) can be represented by a graph and $X_{\mathcal{C}}$ satisfies a graphical *back-door criterion*, then

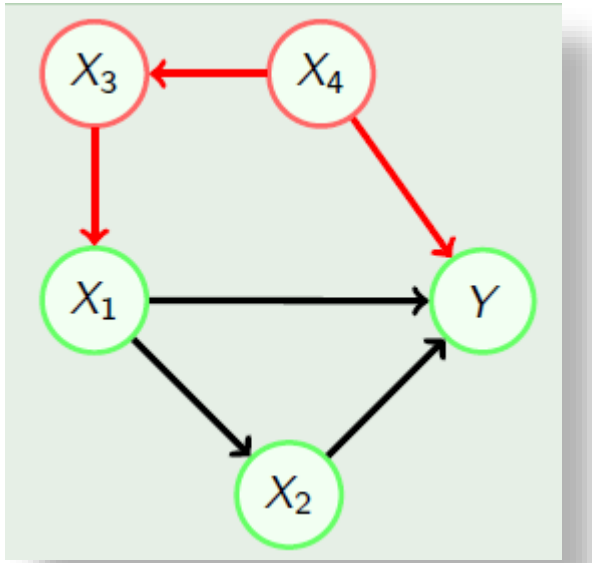
$$\begin{aligned}\mathrm{P}(y|do(X_{\mathcal{S}} = x_{\mathcal{S}})) &= \int \mathrm{P}(y|do(X_{\mathcal{S}} = x_{\mathcal{S}}), X_{\mathcal{C}} = x_{\mathcal{C}}) \, \mathrm{d}P(x_{\mathcal{C}}) \\ &= \int \mathrm{P}(y|X_{\mathcal{S}} = x_{\mathcal{S}}, X_{\mathcal{C}} = x_{\mathcal{C}}) \, \mathrm{d}P(x_{\mathcal{C}}).\end{aligned}$$

Here $\mathrm{P}(y|do(X_{\mathcal{S}} = x_{\mathcal{S}}))$ stands for the distribution of Y after we make an intervention on $X_{\mathcal{S}}$ that sets it equal to $x_{\mathcal{S}}$.

- $\mathrm{E}(y|do(X_{\mathcal{S}} = x_{\mathcal{S}}))$ is essentially $\mathrm{E}[Y(x_{\mathcal{S}})]$ in the Neyman-Rubin potential outcome framework.

PDP is the same as Pearl's back-door adjustment formula!

A set of variables X_C satisfies the back-door criterion with respect to X_S and Y if
1) No node in X_C is a descendant of X_S, and
2) X_C blocks (d-separates) every back-door path between X_S and Y (contains an arrow into X_S).



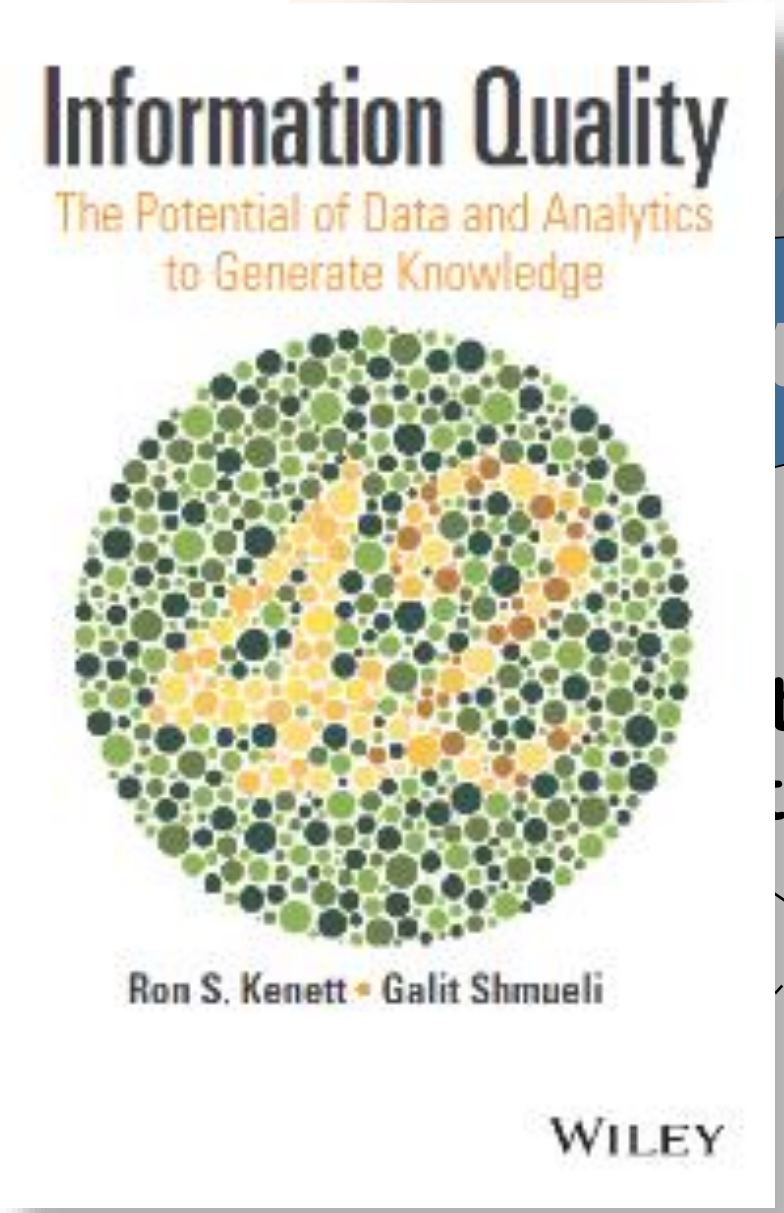
$X_S = X_1$,

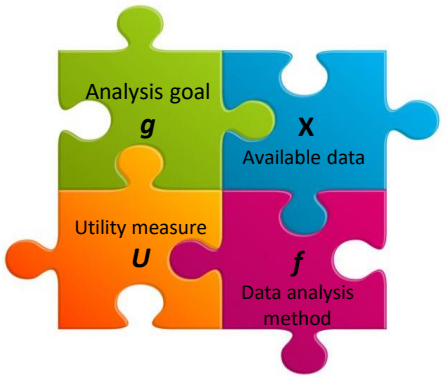
$X_C = \{X_3\}$, $\{X_4\}$ or $\{X_3; X_4\}$.

Agenda

1. Background on causality in science and statistics
2. Fishbone cause and effect diagrams
3. Bayesian networks
4. Randomization in experimental designs
5. Propensity scores in observational studies
6. Counterfactuals and do calculus
7. **Personalized medicine, condition based maintenance and Industry 4.0**
8. Future research areas

$$InfoQ(U,f,X,g) = U(f(X|g))$$





What

Analytic Space

InfoQ Dimensions

Goals

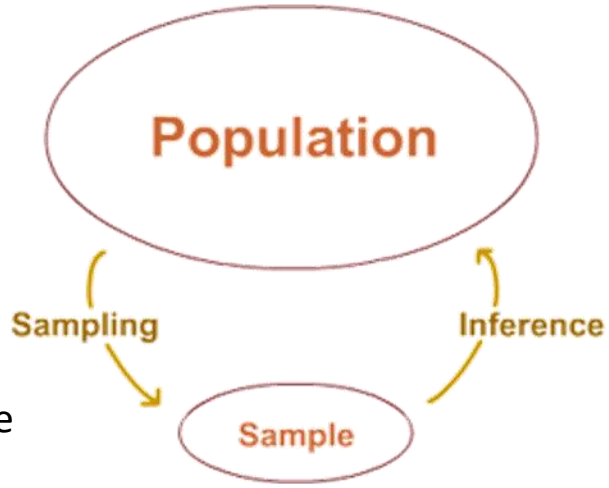
Analysis Quality

How

1. Data resolution
2. Data structure
3. Data integration
4. Temporal relevance
5. Chronology of data and goal
6. **Generalizability**
7. Operationalization
8. Communication

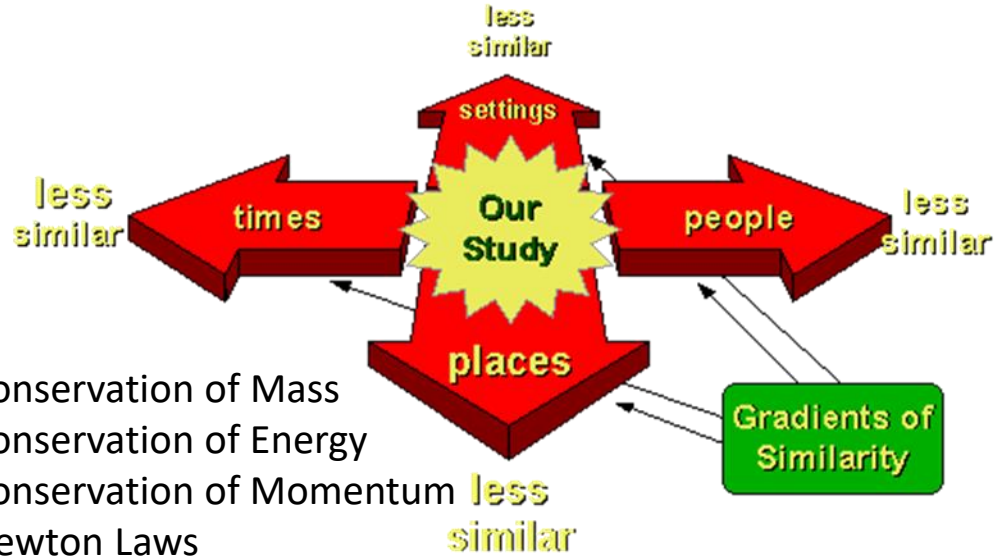
Generalizability

Statistical generalizability



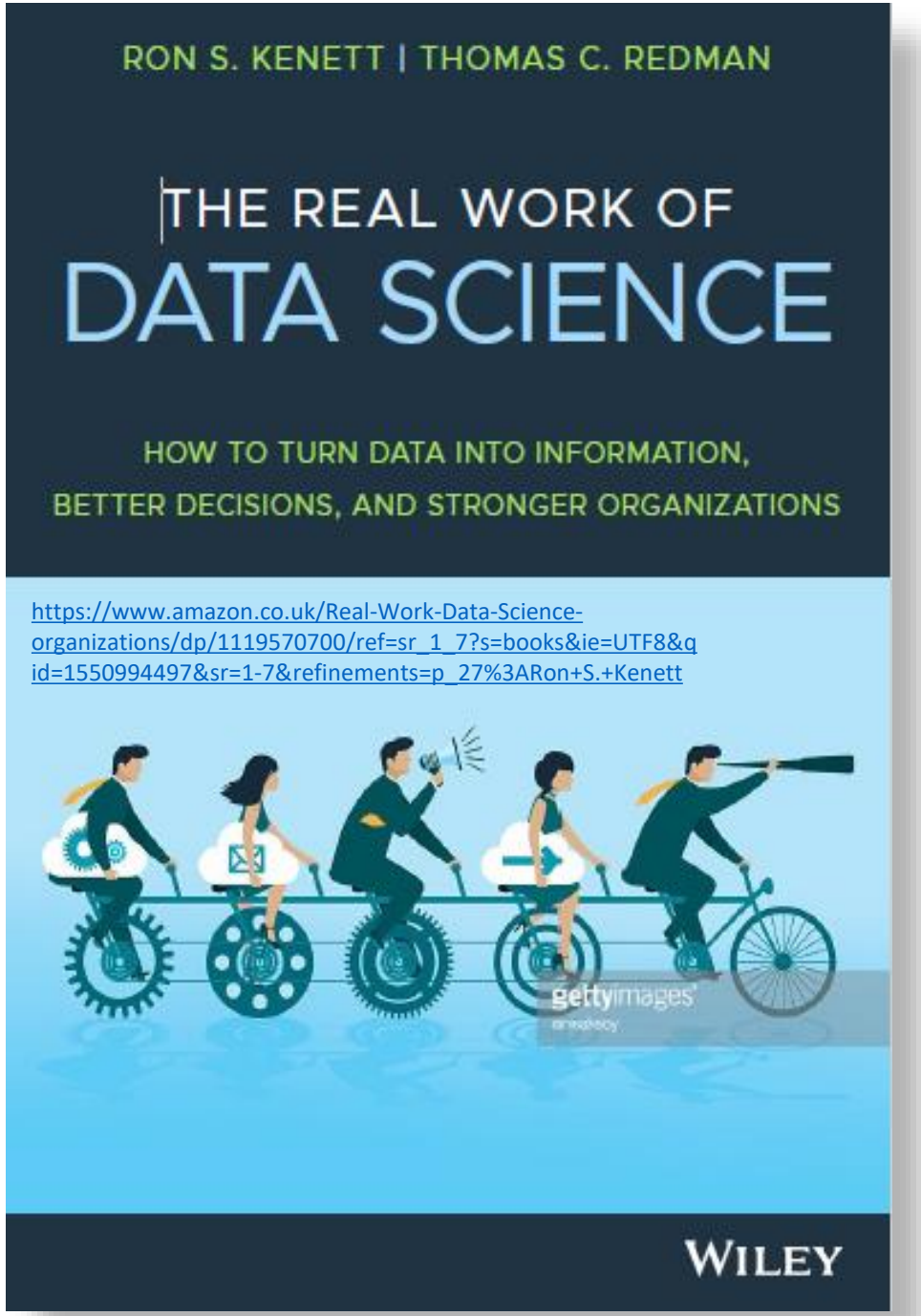
Statistical inference
Regression models
Predictive analytics
…..

Scientific generalizability



Conservation of Mass
Conservation of Energy
Conservation of Momentum
Newton Laws
Principle of least action
Laws of thermodynamics
Maxwell's equations
…..

Intuition



https://www.amazon.co.uk/Real-Work-Data-Science-organizations/dp/1119570700/ref=sr_1_7?s=books&ie=UTF8&qid=1550994497&sr=1-7&refinements=p_27%3ARon+S.+Kenett

1. “A higher calling.”
2. “The difference between a good data scientist and a great one.”
3. “Learn the business.”
4. “Understand the real problem.”
5. “Get out there.”
6. “Sorry, but you can’t trust the data. Deal with it.”
7. “Make it easy for people to understand your insights.”
8. “When the data leaves off and your intuition takes over.”
9. “Take accountability for results.”
10. “What does it mean to be ‘data-driven,’”
11. “Rooting out bias in decision-making.”
12. “Teach, teach, teach.”
13. “Evaluating data science outputs more formally”
14. “Educating senior management.”
15. “Putting data science, and data scientists, in the right spots.”
16. “Moving up the analytics maturity ladder.”
17. “The industrial revolutions and data science.”
18. Epilogue

Level 5: *Learning and discovery* - This is where attention is paid to information quality. Data from different sources is integrated. Chronology of Data and Goal and Generalization is a serious consideration in designing analytic platforms. **Leverage causality models.**

Level 4: *Quality by Design* - Experimental thinking is introduced. The data scientist suggests experiments, like A/B testing, to help determine which website is better. **Develop causality analysis**.

Level 3: *Process focus* - Probability distributions are part of the game. The idea that changes are statistically significant, or not, is introduced. Some attention is given to model fitting. **Introduce causality analysis.**

Level 2: *Descriptive statistics level* – Management asks to see histograms, bar charts and averages. **Models are not used, data is analyzed in rather basic ways.**

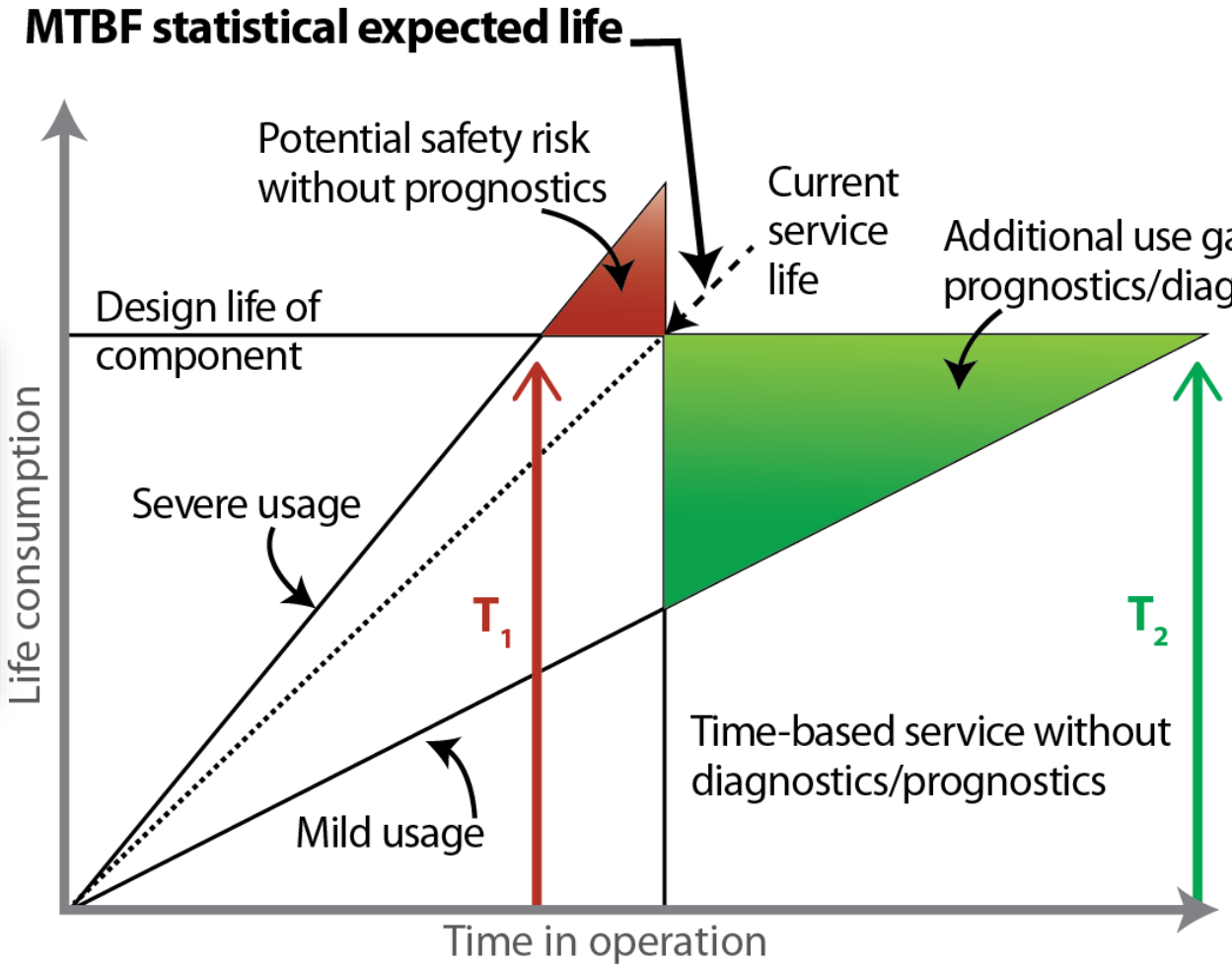
Level 1: *Random demand for reports driven by firefighting* - New reports address questions such as: How many components of type X did we replace last month or how many people in region Y applied for a loan?

Condition Based Maintenance (CBM)
Health and Usage Monitoring Systems (HUMS)
Prognostics and Health Management (PHM)

- Monitoring

- Diagnostics
- Prognostics
- Prescriptive

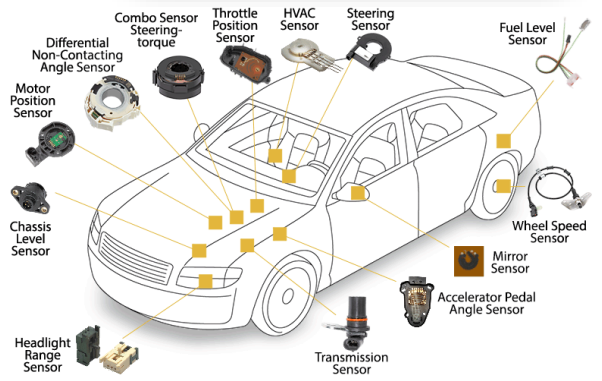
Causality



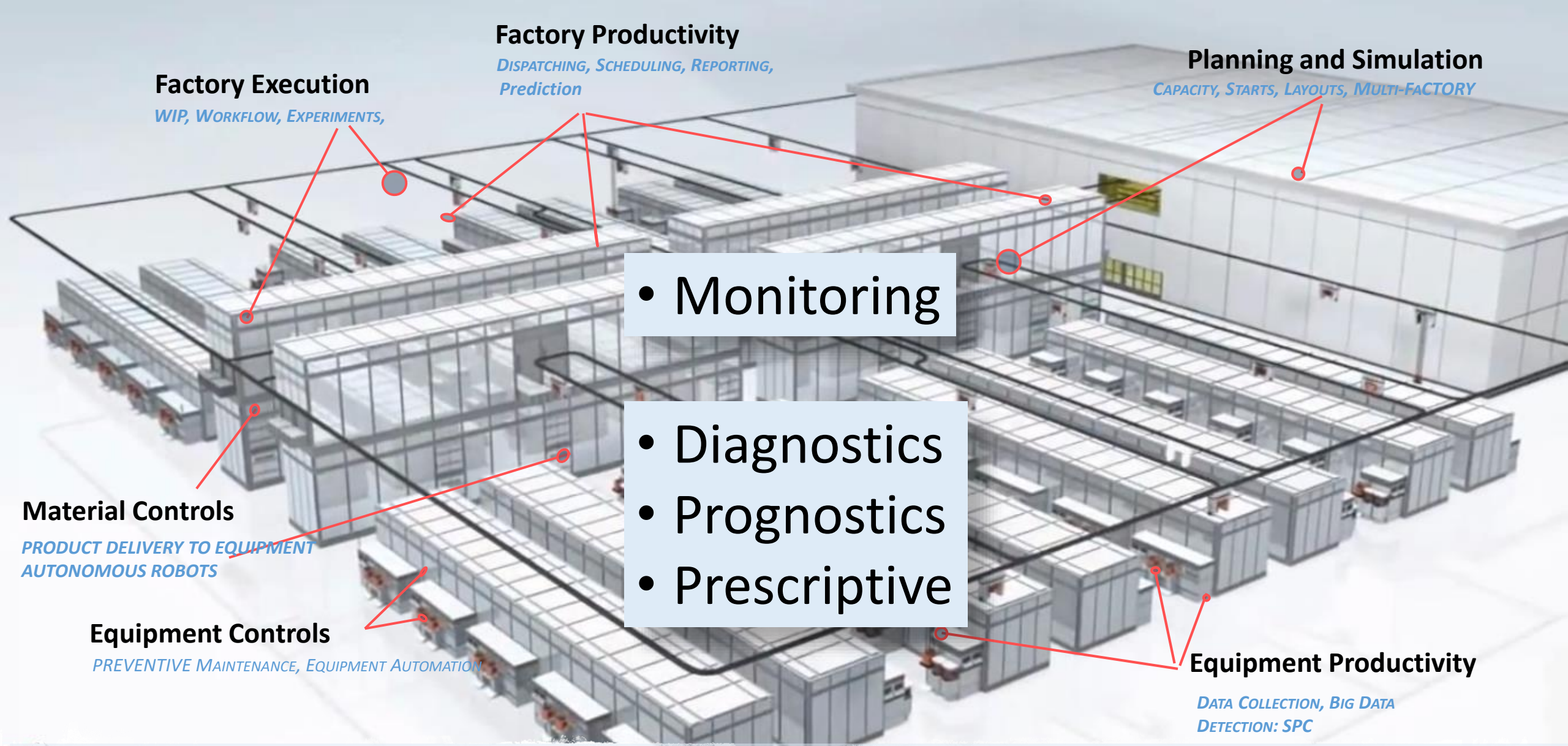
Source: Economic and Safety Benefits of Diagnostics & Prognostics (Romero et al. 1996)

Analytics

Sensor technology



00514



The Industry 4.0 Factory

ANALYTIC METHODS IN
SYSTEMS AND SOFTWARE TESTING

EDITED BY
RON S. KENETT, FABRIZIO RUGGERI, FREDERICK W. FALTIN

WILEY

EDITED BY
RON S. KENETT | ROBERT S. SWARZ | AVIGDOR ZONNENSHAIN

SYSTEMS ENGINEERING IN THE FOURTH INDUSTRIAL REVOLUTION

BIG DATA, NOVEL TECHNOLOGIES, AND MODERN SYSTEMS ENGINEERING

WILEY

2ND EDITION

Modern Industrial Statistics

with applications in R, MINITAB and JMP

RON S. KENETT

SHELEMYAHU ZACKS

With contributions from Daniele Amberti

STATISTICS IN PRACTICE

WILEY

- Monitoring

- Diagnostics
- Prognostics
- Prescriptive

Dr. Ran Balicer at Exponential Medicine

Picture this: instead of going to a physician with your ailments, your doctor calls you with some bad news: "Within six hours, you're going to have a heart attack. So why don't you come into the clinic and we can fix that." Crisis averted.

ATLASENSE BIOMED
Monitoring anytime, anywhere

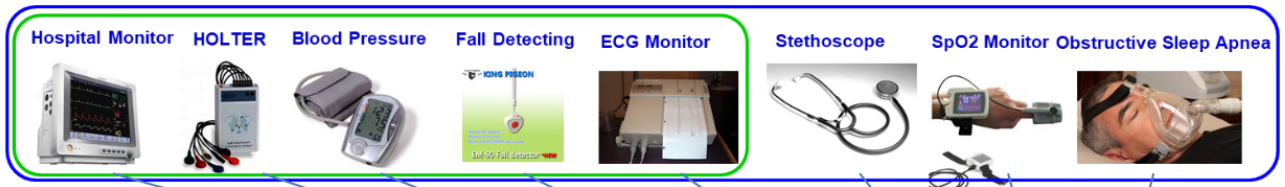
Home | Company | Solutions | Contact us

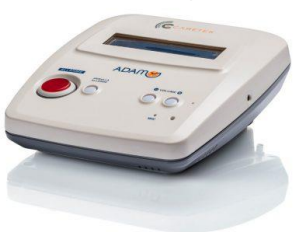
Monitoring and Diagnostics

Continuous Remote Patient Monitoring Solutions

Mobile Monitoring and Analysis Software for Early Detection of Life-threatening Conditions and Trends

Continuous Remote Triage and Monitoring of Multiple Diseases and Conditions From a Single Wearable Solution





ATLASense Biomed Cloud

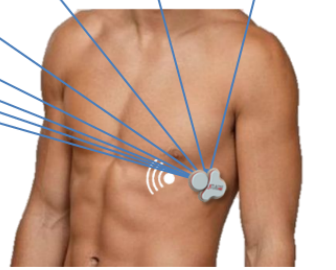
AI, Predictive analytics, Advanced algorithms, Big data

Monitored data

Predictive analytics, trends and alerts

Customization And thresholds

Mobile gateway



Patented Technology, Universal Interface for all diagnostic algorithms

Alerts and notifications

Healthcare providers

Dashboard Administration and **remote control**

Prognostic analysis

ORIGINAL ARTICLE

Change in Systolic Blood Pressure During Stroke, Functional Status, and Long-Term Mortality in an Elderly Population

Avraham Weiss,[1,*] Yichayaou Beloosesky,[1,*] Ron S. Kenett,[2] and Ehud Grossman[3]

BACKGROUND
Elevated systolic blood pressure (SBP) recorded by 24-hour blood pressure monitoring (24H BPM) on the first day of acute stroke is associated in elderly patients, with an unfavorable outcome. Herein, we assessed, by 24H BPM, the impact of the change in SBP levels during the first week of stroke on short-term functional status and long-term mortality in elderly patients.

METHODS
One hundred and fifty acute stroke patients (69 males), mean age at admission 83.6 ± 5.5 years, 82% with ischemic stroke, were investigated. 24H BP M was recorded within 24 hours of admission and 1 week later. After 7 days, patients were assessed for functional status according to the modified Rankin scale (mRS) and were subsequently followed for mortality up to 7.5 years (mean 3.16 ± 2.29).

RESULTS
After 7 days, SBP decreased from 147 ± 21 to 140 ± 20 mm Hg ($P < 0.001$). Functional status improved and mRS decreased from 4.2 to 3.7. During follow-up, 58 patients (17 males and 41 females) had died. Mortality rate was higher in females (69% vs. 45%; $P < 0.01$) and in patients with a history of congestive heart failure. The average admission SBP predicted short-term functional status and long-term mortality. However, the change in SBP corrected for admission levels, gender, age and other variables was not associated with short-term functional status and long-term mortality.

CONCLUSION
There is no evidence of association between change in SBP during the first week of stroke and short-term functional status and long-term mortality in this group of stroke patients.

Keywords: acute phase stroke; blood pressure; change in blood pressure; elderly; functional status; hypertension; morbidity; mortality.

doi:10.1093/ajh/hpv118

Elevated blood pressure (BP) is commonly observed during an acute stroke and usually returns to normal within a few days.[1–4] Elevated BP during an acute ischemic stroke might be advantageous by improving cerebral perfusion to the ischemic tissue or detrimental by exacerbating edema and hemorrhagic transformation of the ischemic tissue.[1] It is unclear whether high BP should be treated in acute ischemic stroke.[3]

Most intervention studies have failed to discover an association between the fall in BP and outcome.[4] Castillo *et al.* found that a fall in systolic BP (SBP) of more than 20 mm Hg was associated with a poor outcome.[2] Leira *et al.* observed that in elderly patients (>70 years) reductions in SBP was associated with a worse prognosis.[3] Glantin *et al.*, in a small study using 24H BP monitoring (BPM), found an inverse correlation between the decrease of 24H BP within the first week and functional status in elderly patients with acute stroke.[9]

It should be noted that most of the studies were limited by their small size.[3] Recently, a meta-regression analysis showed that modest BP reductions may reduce death and dependency, whereas large falls or increases in BP are associated with a worse outcome. However, these results were not adjusted for baseline BP values or for other factors such as previous hypertensive disease and previous antihypertensive treatment.

Several randomized controlled trials evaluated the therapeutic BP lowering during the acute phase of stroke outcome.[7–9,12] Most data regarding the association the change in BP during the acute phase of stroke come were derived from interventional studies that different agents, evaluated short-term outcome, used BP levels and included only very few elderly patients.

We recently evaluated the effect of admission SBP and long-term outcome in very elderly stroke patients. the present study, we extended the follow-up and assessed relationship between a spontaneous change in BP, as 24H BPM, and short- and long-term outcomes.

METHODS

Study population

All study patients diagnosed with acute stroke were consecutively admitted and evaluated on admission to

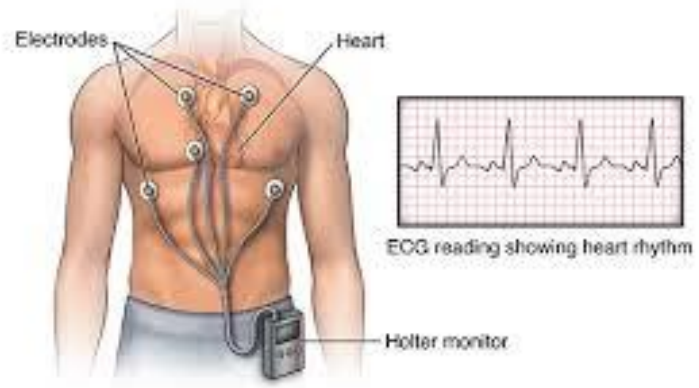
Correspondence: Ehud Grossman (grossm@tauex.tau.ac.il).

Initially submitted April 16, 2015; date of first revision June 20, 2015; accepted for publication June 24, 2015.

[1]Geriatric Ward, Rabin Medical Center, Petach Tikvah, Israel; [2]University of Turin, Italy; [3]Department of Internal Medicine D, Hypertension Unit, The Chaim Sheba Medical Center, Tel-Hashomer, Affiliated to Sackler School of Medicine, Tel-Aviv University, Israel.
*These authors contributed equally to this work as co-first authors.

Casual BP test versus Holter

Holter monitor with ECG reading



Stroke

American Heart Association | American Stroke Association

Systolic Blood Pressure During Acute Stroke Is Associated With Functional Status and Long-term Mortality in the Elderly

Avraham Weiss, Yichayaou Beloosesky, Ron S. Kenett and Ehud Grossman

Stroke. 2013;44:2434-2440; originally published online July 18, 2013;
doi: 10.1161/STROKEAHA.113.001894
Stroke is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231

Print ISSN: 0039-2499. Online ISSN: 1524-4628

Agenda

1. Background on causality in science and statistics
2. Fishbone cause and effect diagrams
3. Bayesian networks
4. Randomization in experimental designs
5. Propensity scores in observational studies
6. Counterfactuals and do calculus
7. Personalized medicine, condition based maintenance and Industry 4.0
8. **Future research areas**

Traditional Data Sources

Small volume – low statistical power
Limited variety – Biased estimates
Low velocity – estimates may not be valid in the future

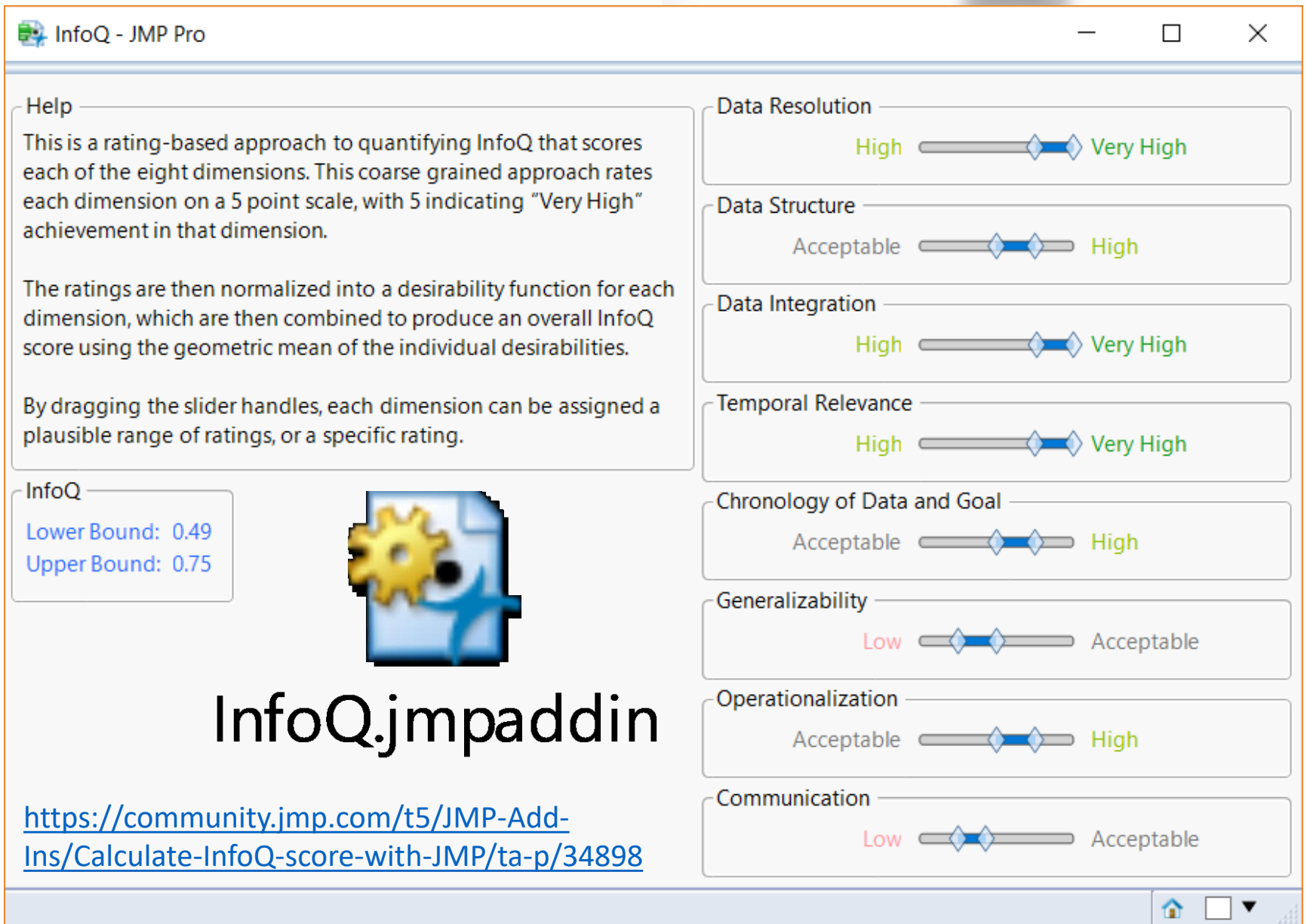
Untapped Sources

High volume – high statistical significance - small p value
High variety – small bias
High velocity – dynamic update of estimates

Information Quality (InfoQ)

Information Quality (InfoQ)

How

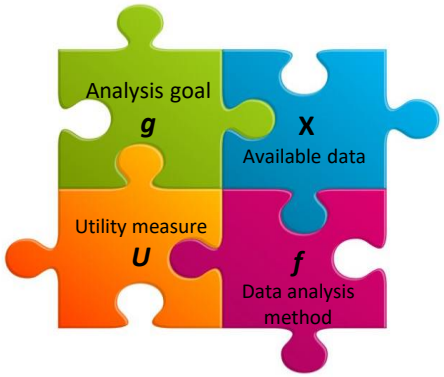


https://community.jmp.com/t5/JMP-Add-Ins/Calculate-InfoQ-score-with-JMP/ta-p/34898

What

InfoQ Components

$$InfoQ(f,X,g) = U(f(X|g))$$



InfoQ Score

#	Dimension	Note	Value	Index
1	Data resolution		5	1.0000
2	Data structure		4	0.7500
3	Data integration		5	1.0000
4	Temporal relevance		5	1.0000
5	Generalizability		3	0.5000
6	Chronology of data and goal		5	1.0000
7	Concept operationalization		2	0.2500
8	Communication		3	0.5000
	InfoQ Score = 0.68			

InfoQ=68%

Big data

Unbiased Estimates

Data integration

Timely Estimates

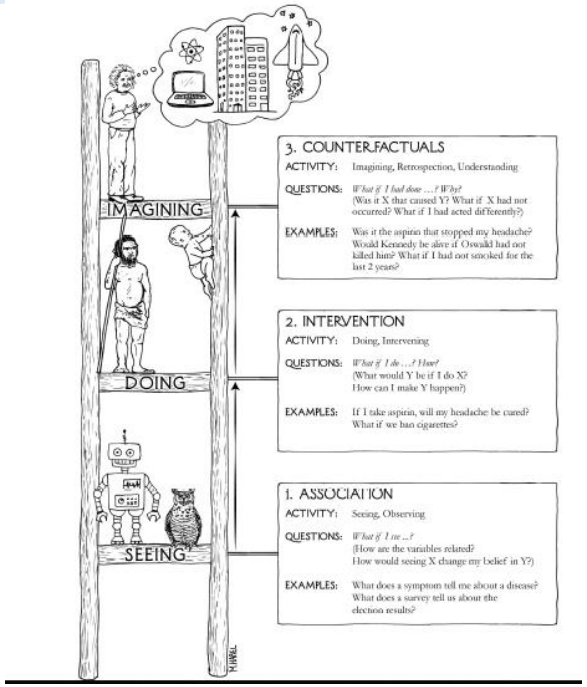
Chronology of data and goal

Variety

Velocity

Volume

Causality



Statistical Significance

Operationalization of findings

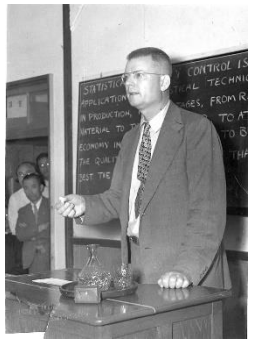
An historical perspective

1. **Product Quality**
2. **Process Quality**
3. **Service Quality**
4. **Management Quality**
5. **Design Quality**
6. **Information Quality**

1800 1900 1950 1960 1980 2010

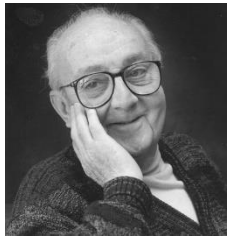
Eli Whitney
(1765 – 1825)

Walter Shewhart
(1891 – 1961)



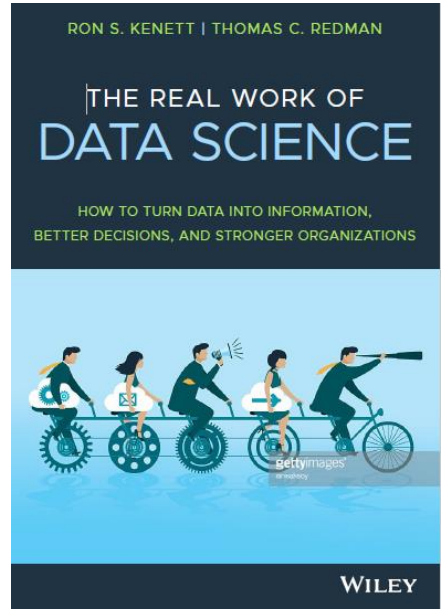
Edwards Deming
(1900 – 1993)

Joseph Juran
(1904 - 2008)



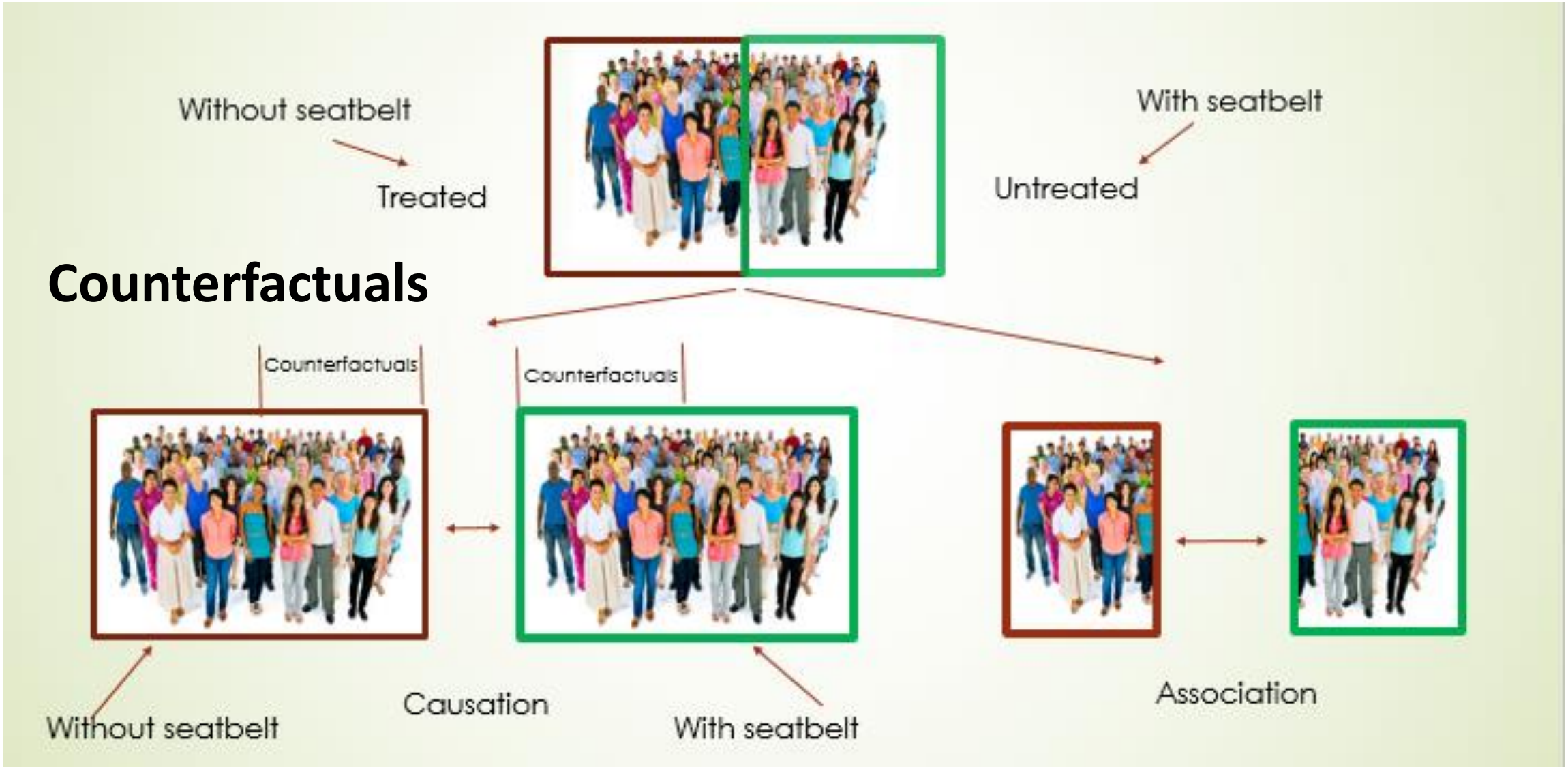
George Box
(1919 – 2013)

Genichi Taguchi
(1924 – 2012)



- Monitoring

- Diagnostics
- Prognostics
- Prescriptive


Without seatbelt
Treated
With seatbelt
Untreated
Counterfactuals
Counterfactuals
Counterfactuals
Without seatbelt
Causation
With seatbelt
Association

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3033588

Applications and Theoretical Results of Association Rules and Compositional Data Analysis: A Contingency Table Perspective

Marina Vives-Mestres*, Josep Antoni Martín-Fernández* Santiago Thió-Henestrosa* Ron S. Kenett**

Abstract: Association rule mining was originally developed for basket analysis. To generate an association rule, the collection of more frequent itemsets must be detected. The association rules are then ranked using measures of interestingness. Using the associaton rule expression as a contingency table a representation on the unit simplex is appropiate. Compositional data analysis provides nice properties such as subcompostional coherence and scalability. We explore here the implication of compositional data analysis to association rule mining in large databases and big data and propose compositional measures of interestingness. Visualization of compositional measures on a simplicial representation of the itemsets gives new insights in association rule mining. The case study used here to demonstrate our approach is derived from a medical data set of side effects from Nicardipine.

	B	**^{c}B**
A	x_1	x_2
^{c}A	x_3	x_4

	B	**^{c}B**
A	$x_1\sqrt{x_2x_3}$	$x_2\sqrt{x_1x_4}$
^{c}A	$x_3\sqrt{x_1x_4}$	$x_4\sqrt{x_2x_3}$

$\mathbf{T}_{\text{ind}}$ of AR $\{\mathbf{A} \Rightarrow \mathbf{B}\}$

	B	**^{c}B**
A	$1/\sqrt{x_2x_3}$	$1/\sqrt{x_1x_4}$
^{c}A	$1/\sqrt{x_1x_4}$	$1/\sqrt{x_2x_3}$

$\mathbf{T}_{\text{int}}$ of AR $\{\mathbf{A} \Rightarrow \mathbf{B}\}$

ilr-coordinates	ilr_1	ilr_2	ilr_3
T	$\frac{1}{2}\ln\left(\frac{x_1x_4}{x_2x_3}\right)$	$\frac{\sqrt{2}}{2}\ln\left(\frac{x_1}{x_4}\right)$	$\frac{\sqrt{2}}{2}\ln\left(\frac{x_2}{x_3}\right)$
$\mathbf{T}_{\text{ind}}$	0	$\frac{\sqrt{2}}{2}\ln\left(\frac{x_1}{x_4}\right)$	$\frac{\sqrt{2}}{2}\ln\left(\frac{x_2}{x_3}\right)$
$\mathbf{T}_{\text{int}}$	$\frac{1}{2}\ln\left(\frac{x_1x_4}{x_2x_3}\right)$	0	0

	B	^{c}B	
A	p_{11}	p_{10}	p_{1+}
^{c}A	p_{01}	p_{00}	p_{0+}
	p_{+1}	p_{+0}	1

- **SUPPORT** How frequent is the itemset {A, B}?
  support {A,B}= *S*{A,B}= p_{11}
- **CONFIDENCE** Among the antecedent A, how frequent is the consequent B?
  confidence {A → B}= C{A → B}= p_{11} / p_{1+}
- **LIFT** Deviation of the support from that expected under independence
  lift {A → B} = *S* {A,B} / (*S* {A}·*S* {B}) = p_{11} / (p_{1+} p_{+1})

lift {A → B}
- < 1 → when A holds, support of B decreases
- = 1 → No association between A and B
- > 1 → when A holds, support of B increases

Causal interpretation

	B	^{c}B
A	x_1	x_2
^{c}A	x_3	x_4

	B	^{c}B
A	$x_1\sqrt{x_2x_3}$	$x_2\sqrt{x_1x_4}$
^{c}A	$x_3\sqrt{x_1x_4}$	$x_4\sqrt{x_2x_3}$

$\mathbf{T}_{\text{ind}}$ of AR $\{\mathbf{A} \Rightarrow \mathbf{B}\}$

	B	^{c}B
A	$1/\sqrt{x_2x_3}$	$1/\sqrt{x_1x_4}$
^{c}A	$1/\sqrt{x_1x_4}$	$1/\sqrt{x_2x_3}$

$\mathbf{T}_{\text{int}}$ of AR $\{\mathbf{A} \Rightarrow \mathbf{B}\}$

ilr-coordinates	ilr_1	ilr_2	ilr_3
$\mathbf{T}$	$\frac{1}{2}\ln\left(\frac{x_1x_4}{x_2x_3}\right)$	$\frac{\sqrt{2}}{2}\ln\left(\frac{x_1}{x_4}\right)$	$\frac{\sqrt{2}}{2}\ln\left(\frac{x_2}{x_3}\right)$
$\mathbf{T}_{\text{ind}}$	0	$\frac{\sqrt{2}}{2}\ln\left(\frac{x_1}{x_4}\right)$	$\frac{\sqrt{2}}{2}\ln\left(\frac{x_2}{x_3}\right)$
$\mathbf{T}_{\text{int}}$	$\frac{1}{2}\ln\left(\frac{x_1x_4}{x_2x_3}\right)$	0	0

From association to causation

Discover **causal rules** from large databases of binary variables

A	B	C	D	E	F	Y	#repeats
1	1	1	1	1	1	1	14
1	0	1	1	1	1	1	8
1	1	0	1	0	1	1	15
0	1	1	1	1	1	1	8
0	1	0	0	0	0	0	5
0	0	0	0	1	0	1	6
1	0	0	0	0	1	0	4
1	0	1	1	1	0	0	3
0	1	0	1	1	0	0	3
0	1	0	0	1	0	0	5

A → Y
C → Y
BF → Y
DE → Y

Transactions on Machine Learning
and Data Mining
Vol.1, No 2 (2008) 83-96
© ISSN: 1865-6781 (Journal),
IBaI Publishing ISSN 1864-9734

ibai Publishing
www.ibai-publishing.org

Relative Linkage Disequilibrium Applications to Aircraft Accidents and Operational Risks

Ron S. Kenett[1] and Silvia Salini[2]

[1]KPA Ltd., Raanana, Israel and University of Torino, Torino, Italy
ron@kpa.co.it
[2]Department of Economics, Business and Statistics, University of Milan, Italy
silvia.salini@unimi.it

From association to causation

Discover **causal association rules** from large databases of binary variables

A	B	C	D	E	F	Y
1	1	1	1	1	1	1
1	0	1	1	1	1	1
1	1	0	1	0	1	1
0	1	1	1	1	1	1
0	1	0	0	0	0	0
0	0	0	0	1	0	1
1	0	0	0	0	1	0
1	0	1	1	1	0	0
0	1	0	1	1	0	0
0	1	0	0	1	0	0

A → Y

A	B	C	D	E	F	Y
1	1	1	1	1	1	1
1	0	1	0	1	1	1
1	1	0	1	0	1	0
1	0	1	0	1	0	0
0	1	1	1	1	1	0
0	0	1	0	1	1	0
0	1	0	1	0	1	1
0	0	1	0	1	0	1

Fair dataset

From association to causation

Discover **causal association rules** from large databases of binary variables

Fair dataset

A	B	C	D	E	F	Y
1	1	1	1	1	1	1
1	0	1	0	1	1	1
1	1	0	1	0	1	0
1	0	1	0	1	0	0
0	1	1	1	1	1	0
0	0	1	0	1	1	0
0	1	0	1	0	1	1
0	0	1	0	1	0	1

- A: ***Exposure variable***
- {B,C,D,E,F}: ***controlled variable set***.
- Rows with the same color for the controlled variable set are called ***matched record pairs***.

	A=0	
A=1	Y=1	Y=0
Y=1	n_{11}	n_{12}
Y=0	n_{21}	n_{22}

$$OddsRatio_{D_f}(A \rightarrow Y) = \frac{n_{12}}{n_{21}}$$

An association rule A → Y is a causal association rule if: $OddsRatio_{D_f}(A \rightarrow Y) >> 1$

From association to causation

Discover **causal association rules** from large databases of binary variables

A	B	C	D	E	F	G	Y
1	1	1	1	1	1	0	1
…	…						…
1	1	0	1	0	1	0	1

A	B	C	D	E	Y
1	1	1	1	1	1
…	…				…
0	1	1	1	1	0
…					…

1. Remove irrelevant variables (support, local support, association)

For each association rule (e. g. A → Y)

2. Find the exclusive variables of the exposure variable (support, association), i.e. G, F. The controlled variable set = {B, C, D, E}.

3. Find the fair dataset. Search for all ***matched record pairs***
4. Calculate the odds-ratio to identify if the testing rule is causal
5. Repeat 2-4 for each variable which is the combination of variables. Only consider combination of non-causal factors.

QUALITY AND RELIABILITY ENGINEERING INTERNATIONAL
Qual. Reliab. Engng. Int. 2007; **23**:653–663
Published online 12 June 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/qre.861

Special Issue

Joseph M. Juran, a Perspective on Past Contributions and Future Impact

A. Blanton Godfrey[1,¶] and Ron S. Kenett[2,*,†,‡,§]

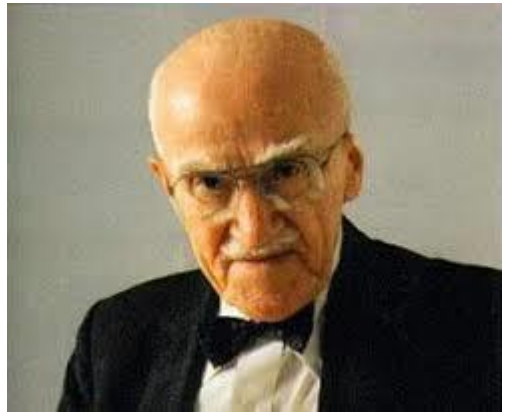
[1]*College of Textiles, North Carolina State University, U.S.A.*
[2]*KPA Ltd., Israel*

This paper combines presentations by the authors in a special session dedicated to the work of Joseph M. Juran at the sixth annual conference of the European Network for Business and Industrial Statistics in Wroclaw, Poland. The paper offers an historical perspective of the contributions of J. M. Juran to management science emphasizing aspects of cause and effect relationships and Integrated Models. Specifically, the paper presents the Juran concepts of Management Breakthrough, the Pareto Principle, the Juran Trilogy® and Six Sigma. The impact of these contributions, put in an historical perspective of key thinkers who investigated cause and effect relationships, is then discussed. The impact of these contributions to modern Integrated Models is then assessed. Copyright © 2007 John Wiley & Sons, Ltd.

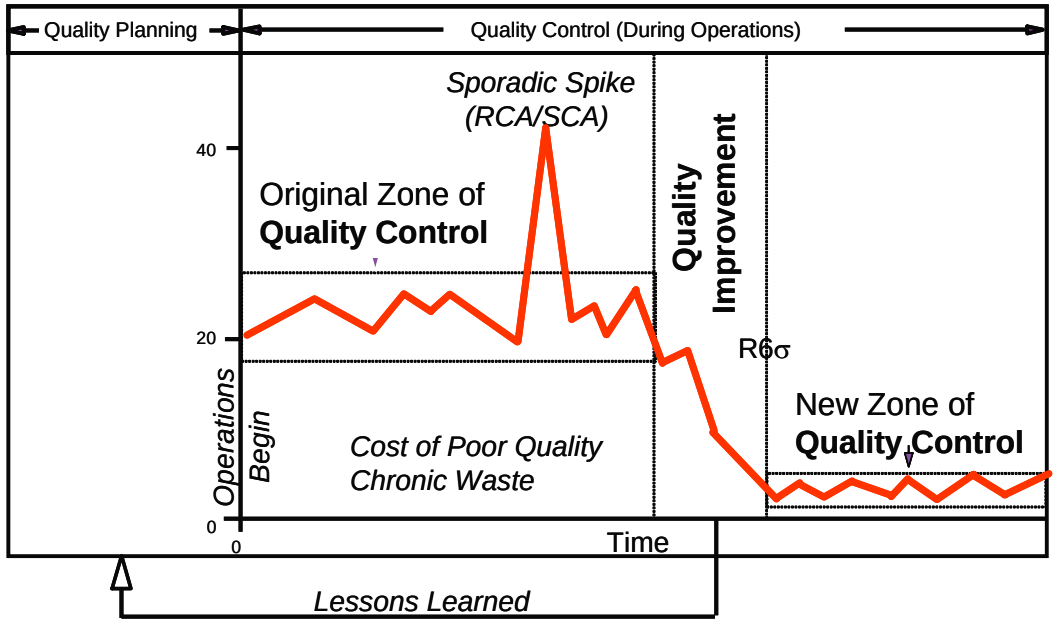
Received 29 January 2007; Revised 16 April 2007; Accepted 18 April 2007

KEY WORDS: J. M. Juran; the Juran Trilogy®; Management Breakthroughs; the Pareto Principle; Six Sigma; quality systems; Integrated Models; cause and effect relationships

Causality in management



The Juran Trilogy®



Quality Planning, Quality Control, and Quality Improvement

5. CAUSE AND EFFECT MODELS

At a presentation celebrating 50 years to the establishment of a Masters Degree in Statistics in Norway Odd O. Aalen has been quoted as stating that: '*Statistics is important because it is conceived as contributing to a* [illegible] *sence of a mechanistic understanding.* [illegible] *y anything about causality. This is a*

Causality in science

[illegible] with causality (e.g. Cox[9]). A famous [illegible] ermany and the number of observed [illegible] ing variable, time[10]. Sketch a scatter plot of population size versus number of storks in the table below and you will see what we mean, if a cause and effect relationship is implied by the data. This simple plot has been used in hundreds of statistics courses—and now in almost every Six Sigma course—to warn students of the dangers of assuming causality too quickly.

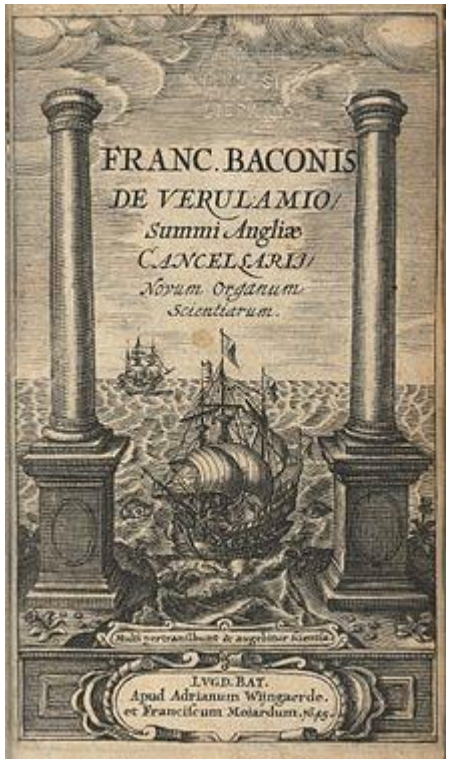
Year	1930	1931	1932	1933	1934	1935	1936
Population in thousands	50	52	64	67	69	73	76
Number of storks	130	150	175	190	240	245	250

Causality is a basic component of the scientific method and general learning. Establishment of causality relies on a combination of axiomatic thought and empirical evidence derived from observational studies of designed experiments. A review of key thinkers and writers in this area covers many centuries and continents.

Sir Francis Bacon (1561–1626) was the chief figure of the English Renaissance and an influential advocate of 'active science.' He writes in Novum Organum. (New Method, 1620) '*… the true method of experience … first lights the candle, then by means of the candle shows the way; commencing as it does with experience duly ordered and digested, not bungling or erratic, and from it educing axioms, and from established axioms again new experiments…*'

Science should start with what Bacon called Tables of Investigation. The Table of Presence lists instances in which the phenomenon being studied occurs. The Table of Absence in Proximity includes the important negative instances; these are the ones most like the positive instances. The Table of Comparison compares the degrees of the phenomenon. Interpretation begins with a brief survey which will suggest the correct explanation of the phenomenon. Although this 'anticipation' resembles a hypothesis, there is in Bacon's discussions no clear indication that he recognized the central scientific importance of devising and testing hypotheses. He goes on to consider 'prerogative instances', those most likely to facilitate interpretation, of which he classifies 27 different types. By following the method outlined, scientific investigation is supposed to produce, almost mechanically, a gradually increasing generality of understanding, a 'ladder of axioms' upon which the mind can climb up or down.

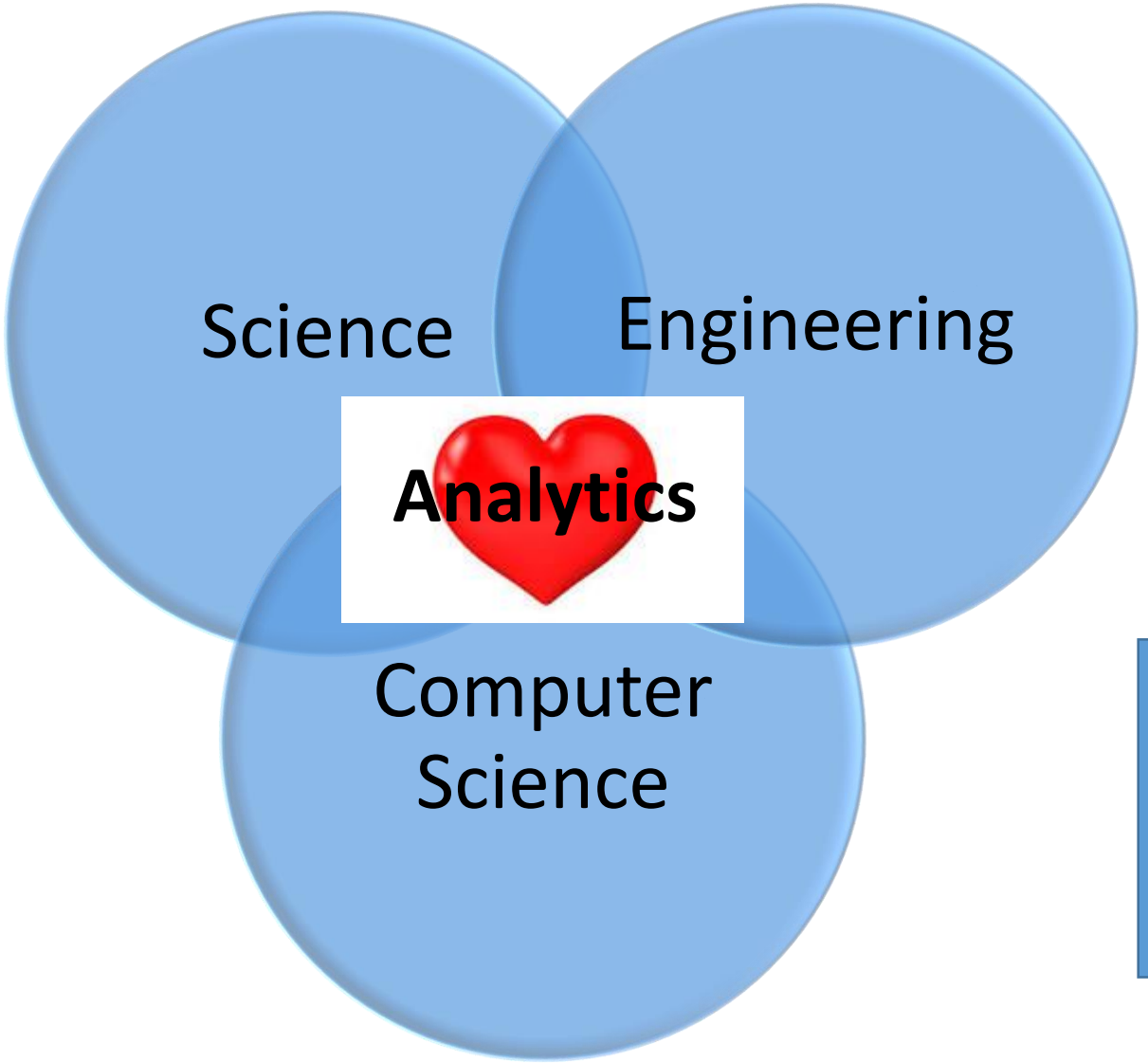
Sir Francis Bacon
1561 - 1626



…the true method of experience. . . first lights the candle, then by means of the candle shows the way; commencing as it does with experience duly ordered and digested, not bungling or erratic, and from it educing axioms, and from established axioms again new experiments. .

So, what did we cover

1. Background on causality in science and statistics
2. Fishbone cause and effect diagrams
3. Bayesian networks
4. Randomization in experimental designs
5. Propensity scores in observational studies
6. Counterfactuals and do calculus
7. Personalized medicine, condition based maintenance and Industry 4.0
8. Future research areas



Thank you for your attention