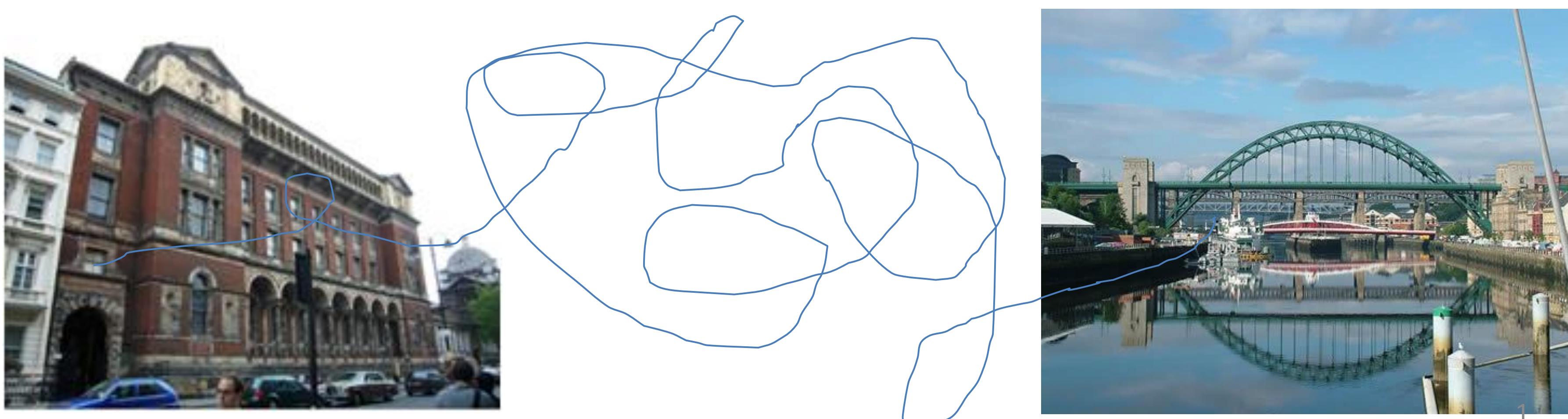
on doing fine work in statistics



Old Huxley Building

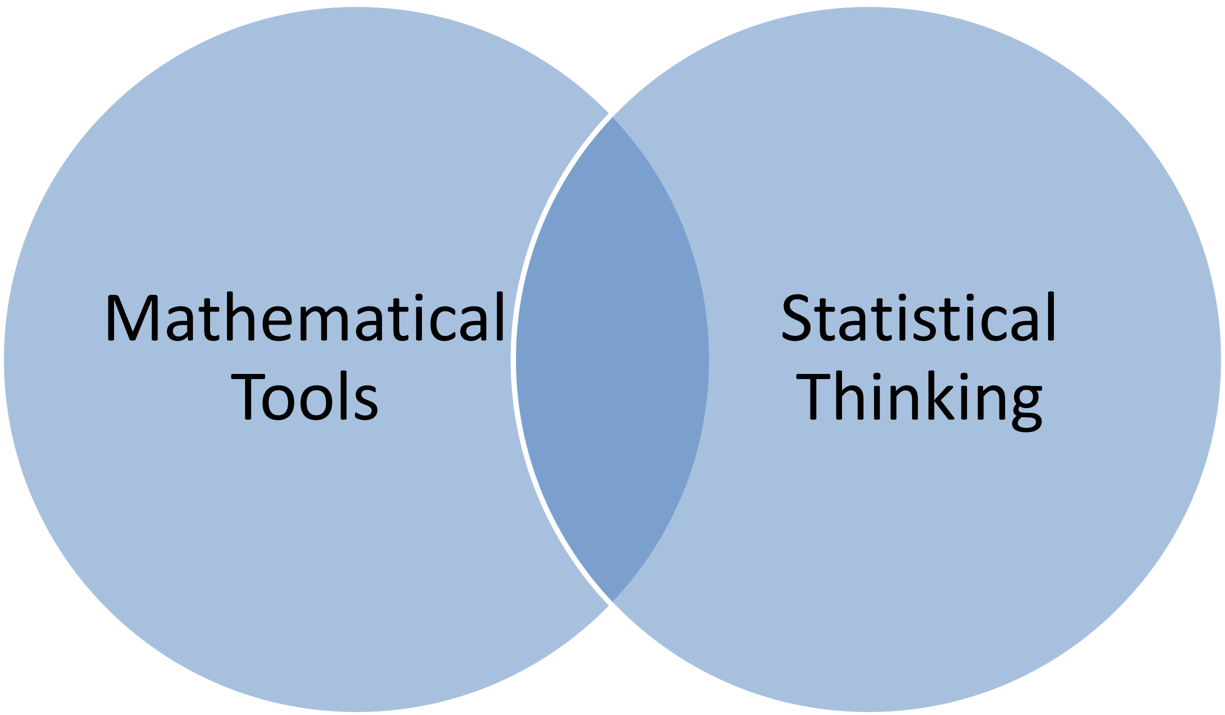
RSS 2013

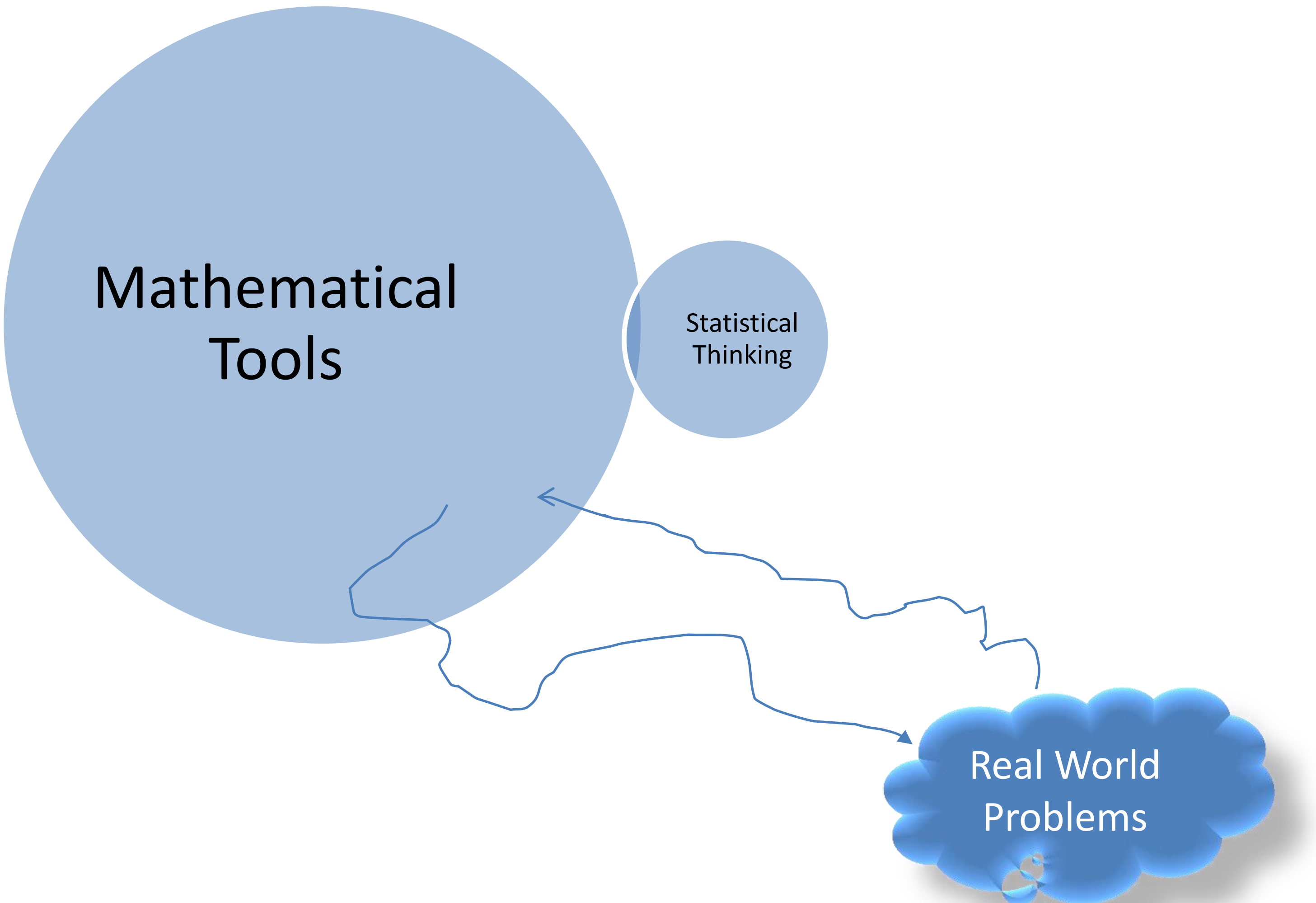
"Much fine work in statistics involves minimal mathematics; some bad work in statistics gets by because of its apparent mathematical content.“

David Cox (1981),
Theory and general principle in statistics, JRSS(A), 144, pp. 289-297.

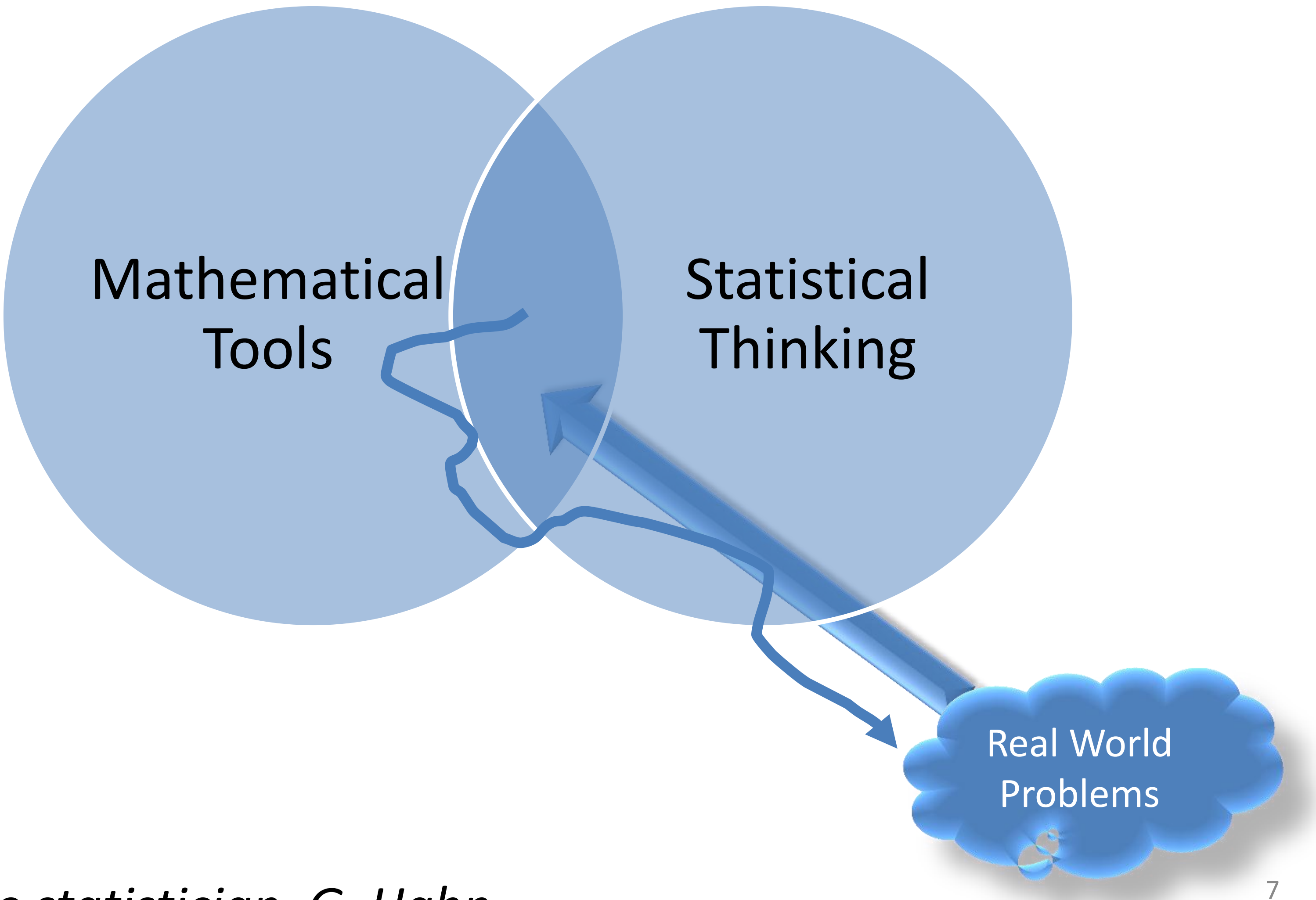
Wordle of
Ron Kenett CV

The challenge of **solving real world problems** with **mathematical tools** and **statistical thinking**

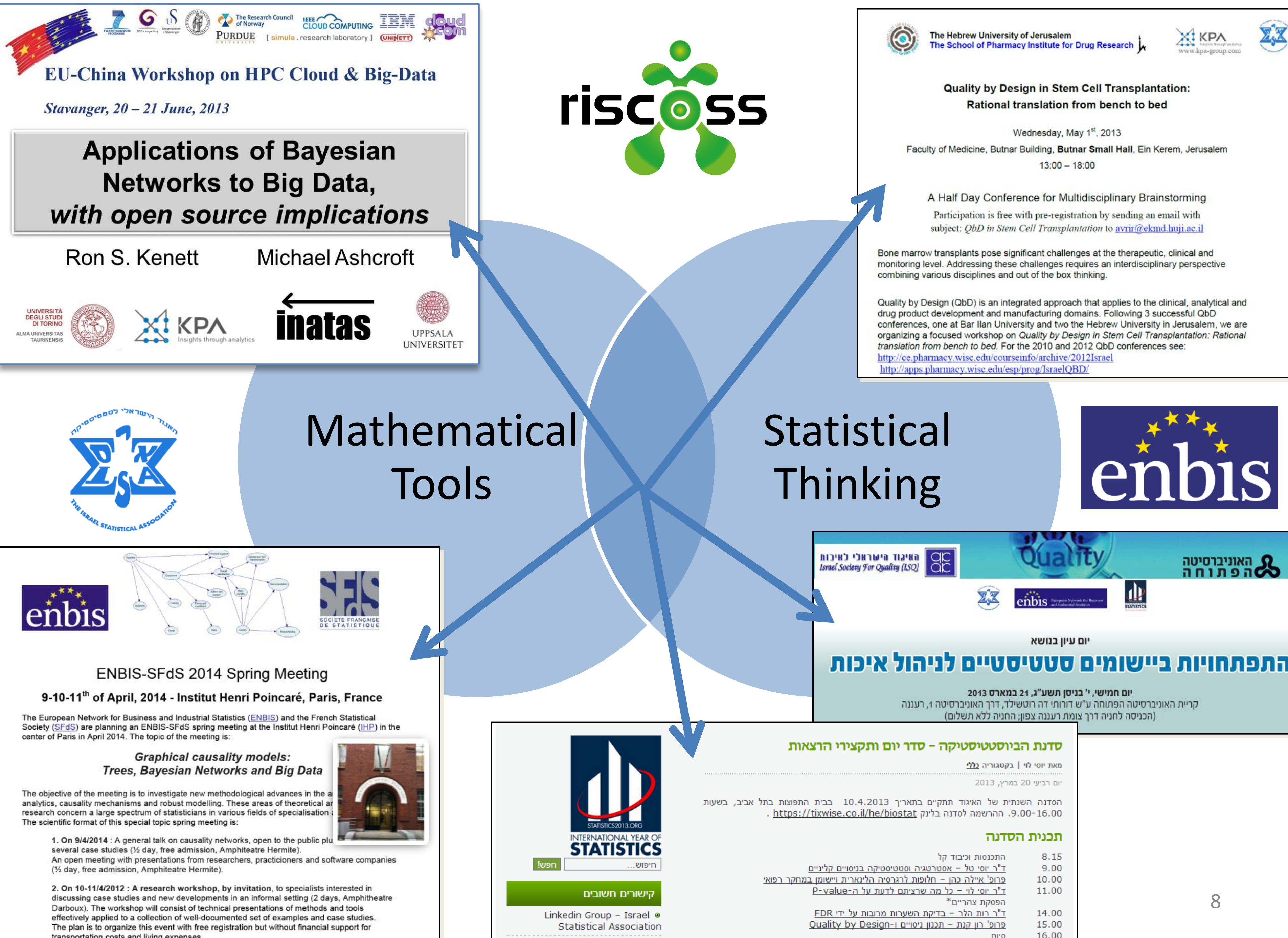

Mathematical Tools
Statistical Thinking



The mathematical statistician



The proactive statistician, G. Hahn



EU-China Workshop on HPC Cloud & Big-Data
Stavanger, 20 – 21 June, 2013
Applications of Bayesian Networks to Big Data, with open source implications
Ron S. Kenett
Michael Ashcroft
riscoss
The Hebrew University of Jerusalem
The School of Pharmacy Institute for Drug Research
Quality by Design in Stem Cell Transplantation: Rational translation from bench to bed
Wednesday, May 1st, 2013
Faculty of Medicine, Butnar Building, Butnar Small Hall, Ein Kerem, Jerusalem
13:00 – 18:00
A Half Day Conference for Multidisciplinary Brainstorming
Participation is free with pre-registration by sending an email with subject: QbD in Stem Cell Transplantation to avrir@ekmd.huji.ac.il
Bone marrow transplants pose significant challenges at the therapeutic, clinical and monitoring level. Addressing these challenges requires an interdisciplinary perspective combining various disciplines and out of the box thinking.
Quality by Design (QbD) is an integrated approach that applies to the clinical, analytical and drug product development and manufacturing domains. Following 3 successful QbD conferences, one at Bar Ilan University and two the Hebrew University in Jerusalem, we are organizing a focused workshop on Quality by Design in Stem Cell Transplantation: Rational translation from bench to bed. For the 2010 and 2012 QbD conferences see:
http://ce.pharmacy.wisc.edu/courseinfo/archive/2012Israel
http://apps.pharmacy.wisc.edu/esp/prog/IsraelQBD/
Mathematical Tools
Statistical Thinking
enbis
THE ISRAEL STATISTICAL ASSOCIATION
ENBIS-SFdS 2014 Spring Meeting
9-10-11th of April, 2014 - Institut Henri Poincaré, Paris, France
The European Network for Business and Industrial Statistics (ENBIS) and the French Statistical Society (SFdS) are planning an ENBIS-SFdS spring meeting at the Institut Henri Poincaré (IHP) in the center of Paris in April 2014. The topic of the meeting is:
Graphical causality models: Trees, Bayesian Networks and Big Data
1. On 9/4/2014 : A general talk on causality networks, open to the public plu... several case studies (½ day, free admission, Amphiteatre Hermite). An open meeting with presentations from researchers, practicioners and software companies (½ day, free admission, Amphiteatre Hermite).
2. On 10-11/4/2012 : A research workshop, by invitation, to specialists interested in discussing case studies and new developments in an informal setting (2 days, Amphitheatre Darboux). The workshop will consist of technical presentations of methods and tools effectively applied to a collection of well-documented set of examples and case studies. The plan is to organize this event with free registration but without financial support for transportation costs and living expenses.
Israel Society For Quality (ISQ)
יום עיון בנושא
התפתחויות ביישומים סטטיסטיים לניהול איכות
STATISTICS2013.ORG
INTERNATIONAL YEAR OF STATISTICS
Linkedin Group – Israel Statistical Association
סדנת הביוסטטיסטיקה – סדר יום ותקצירי הרצאות
תכנית הסדנה

J. R. Statist. Soc. A, (1980),
143, *Part* 4, *pp.* 383–430

Sampling and Bayes' Inference in Scientific Modelling and Robustness

By George E. P. Box

University of Wisconsin–Madison

[Read before the Royal Statistical Society at a meeting organized by the South W... Group on Thursday, May 15th, 1980, the President Sir Claus Moser in the C...

statistician and the investigator are the same person but it is still of value to separate his dual functions

Warning

We do not teach tools and methods for doing that

"This is not as the analyst of a single set of data, nor even as the designer and analyzer of a single experiment, but rather as a colleague working with an investigator throughout the whole course of iterative deductive-inductive investigation."

Problem elicitation 1 | RLD 2 | CED 3 | PSE 4 | InfoQ 5

Are we generating knowledge? (InfoQ)

Are we making an impact? (PSE)

Statistical Thinking

Data integration (ETL, data fusion)

Problem elicitation (cognitive science)

Visualization and data exploration (HMI)

Causality (CS, BN)

Statistical education (concept science)

Confidentiality

Integrated Models

Unstructured data (semantic data, networks)

Reproducible research (Sweave, CFR Part11)

Other Disciplines

1

Problem Elicitation

Greenfield, T. (1987) Consultant's cameos: A chapter of encounters. pp 11-25 in Hand, D.J. and B.S. Everitt eds, *The statistical consultant in action*, Cambridge University Press

1

Problem Elicitation

‘The teacher acts as a mentor in training the student in unstructured problem solving while the computer stores and retrieves information. This sets the mind free to do what it does best – be inductively creative’ (Box, 1997)

‘Most iterative investigations involve multi-dimensional learning: in particular, how to study the problem is a necessary precursor to how to solve the problem’ (Box, 2000)

The Theory of Applied Statistics

Kenett, R.S. (2012) A Note on the Theory of Applied Statistics . Available at SSRN: http://ssrn.com/abstract=2171179 or http://dx.doi.org/10.2139/ssrn.2171179

Visualization: Some (old) references

- **The Visual Display of Quantitative Information**, E. Tufte, Graphics Press, 1983.
- **Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods**, W. Cleveland and R. McGill, *Journal of the American Statistical Association*, 79 (387):531-554, 1984.
- **Graphs in Scientific Publications**, W. Cleveland, *The American Statistician*, 38 (4):261-269, 1984.
- **How to Display Data Badly**, H. Wainer, *The American Statistician* 38(2):137-147, 1984.
- **Data-Based Graphics: Visual Display in the Decades to Come**, J. Tukey, *Statistical Science*, 5(3): 327-339 , 1990.
- **Communicating Statistics,** T. Greenfield, *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 156 (2): 287-297, 1993.

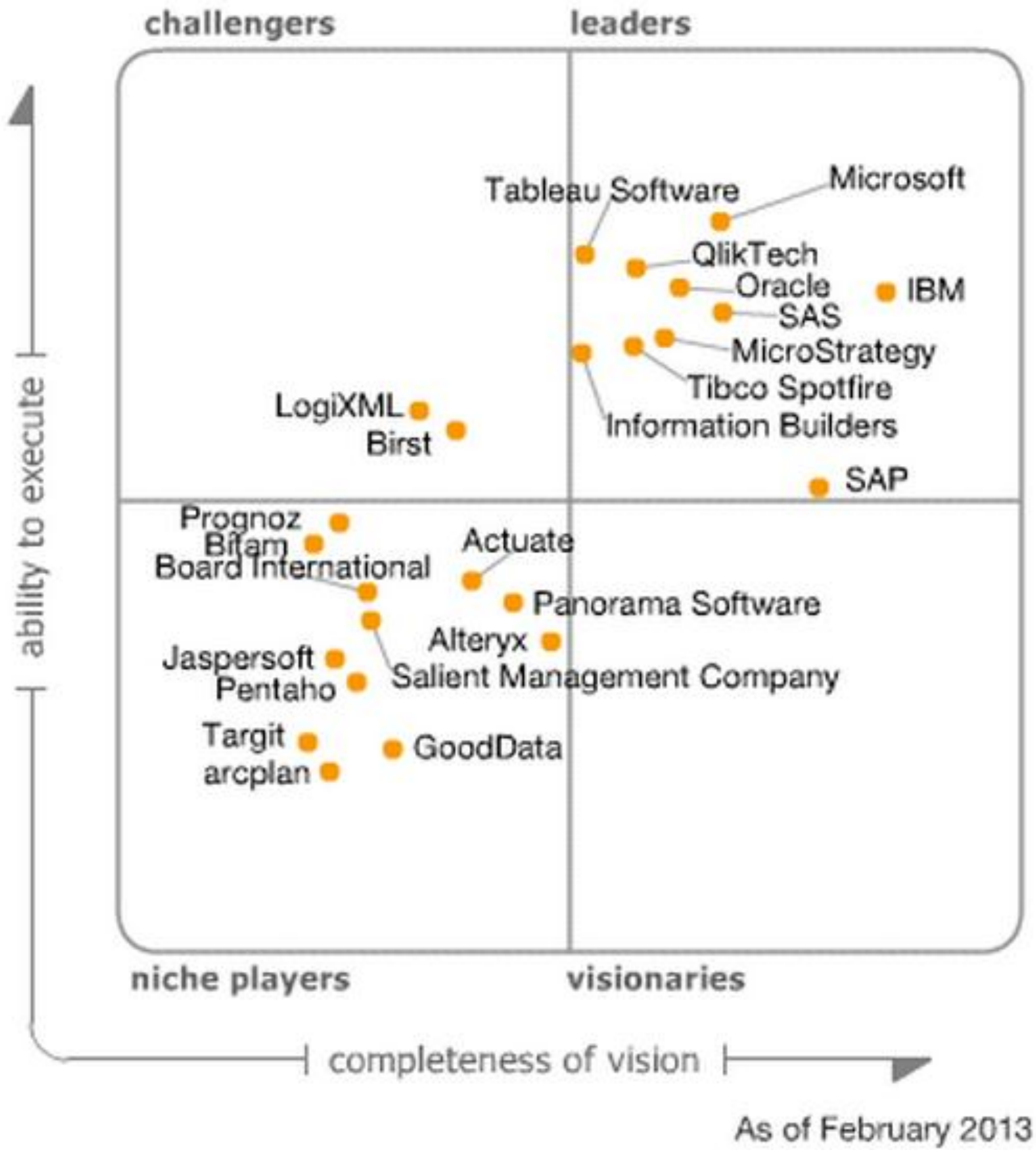
2

Business Intelligence and Analytics Platforms Market Segment

"The dominant theme of the market in 2012 was that **data discovery** became a mainstream BI and analytic architecture. The market also saw increased activity in **real time, content and predictive analytics**." Gartner, 2013

2

Figure 1. Magic Quadrant for Business Intelligence and Analytics Platforms



Gartner's Magic Quadrant for Business Intelligence and Analytics Platforms
5 February, 2013.
Analyst(s): K. Schlegel, R. Sallam, D/ Yuen, J. Tapadinhas

Web Search Interest: obama

2

Worl... 4 - present

Cate... s & Current Events (50-75%), Entertainment (0-10%), Society (0-10%) , more...

Totals

obama 10

Interest over time Forecast News headlines

Learn what these numbers mean

Embed this chart

Do they interact?

Web Search Interest: depression

Worldwide, 2004 - present

Categories: Health (50-75%), Arts & Humanities (10-25%), Society (0-10%), Lifestyles (0-10%) , more...

Totals

depression 63

Interest over time Forecast News headlines

Learn what these numbers mean

Embed this chart

2

The Simplex

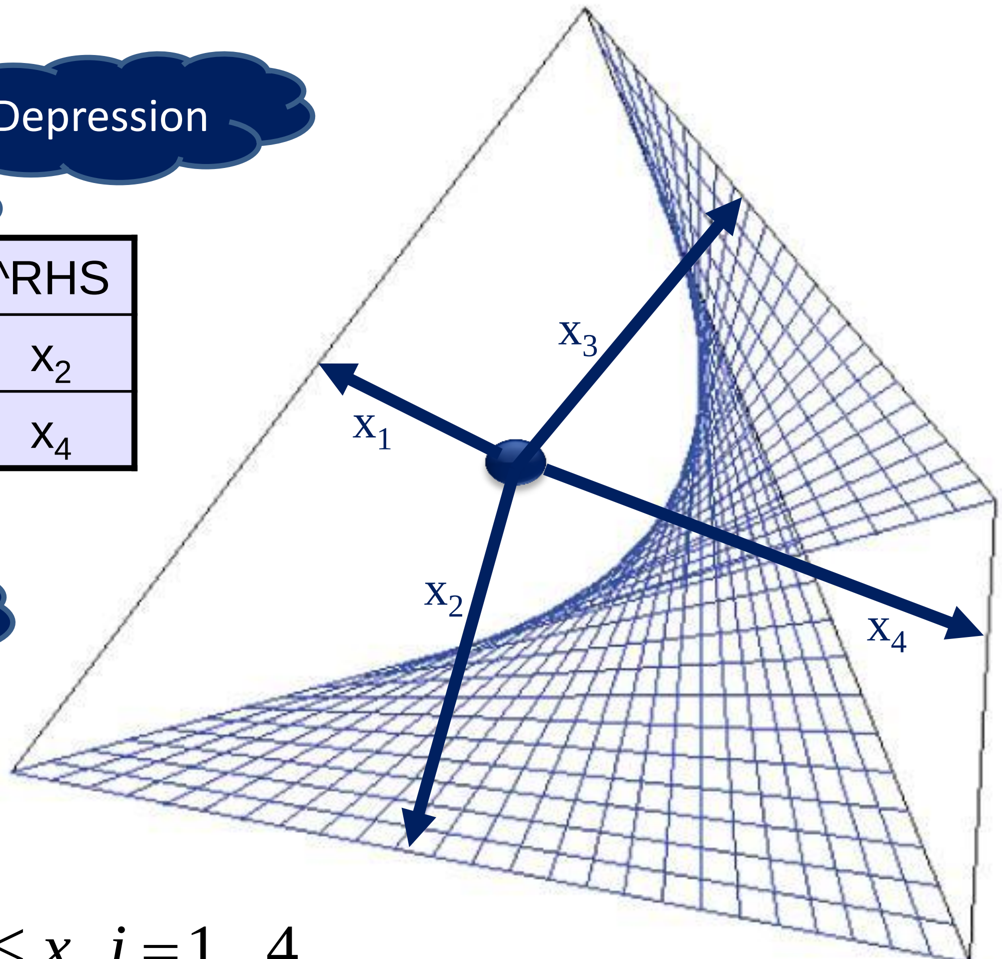
Bishop, Fienberg, & Holland, 1975

$D = x_1 x_4 - x_2\, x_3 = 0$

Depression

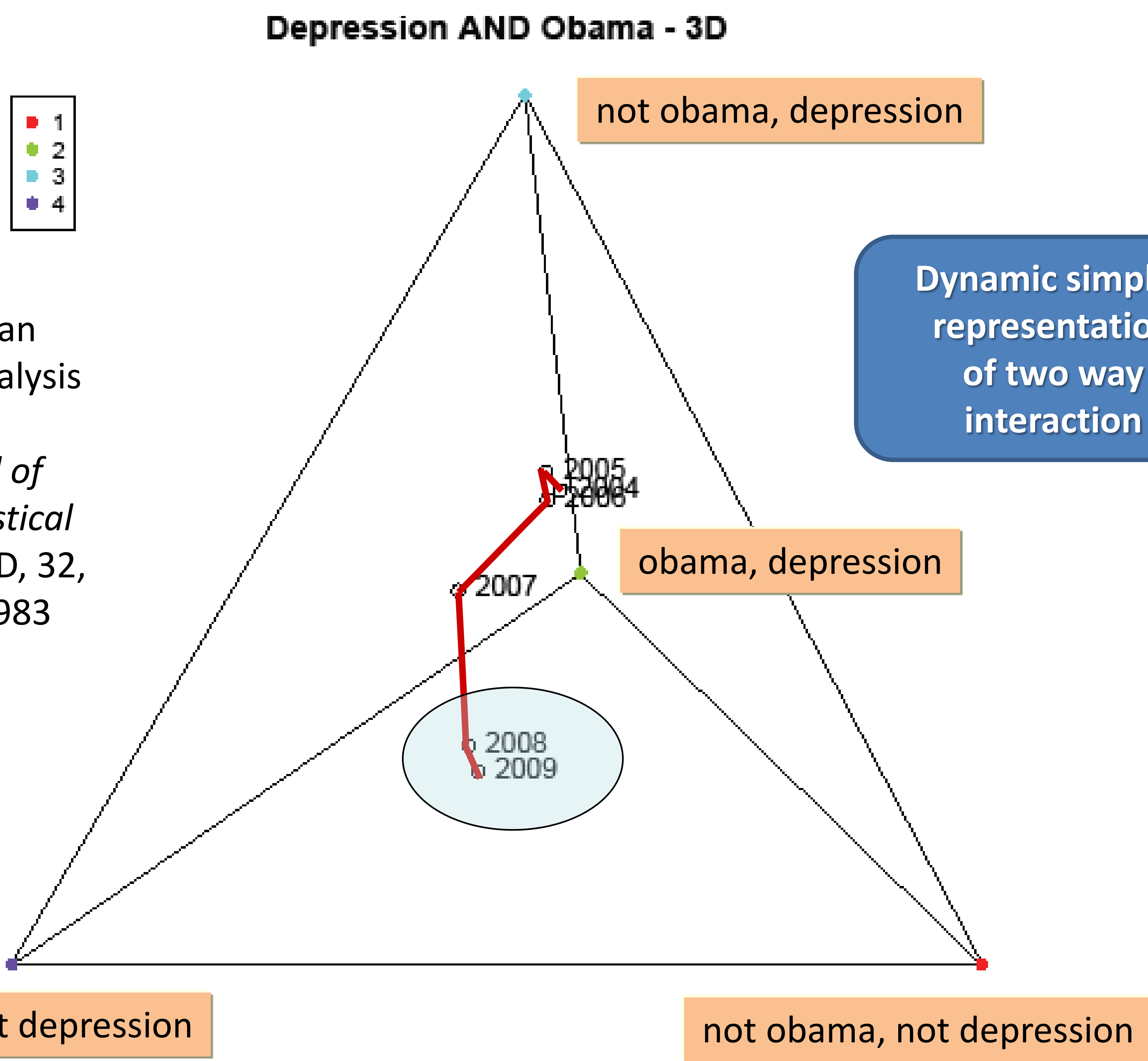
	RHS	^RHS
LHS	x_1	x_2
^LHS	x_3	x_4

Obama



$$\sum_{i=1}^{4} x_i = 1,\ 0 \le x_i, i = 1 \ldots 4.$$

2



Kenett, R. "On an Exploratory Analysis of Contingency Tables" *Journal of the Royal Statistical Society,* Series D, 32, pp. 395-403, 1983

Dynamic simplex representation of two way interaction

Linkage Disequilibrium

$$\boxed{D = x_1 x_4 - x_2 x_3}$$

$$x_1 = fg + D$$

$$x_2 = (1 - f)g - D$$

$$x_3 = f(1 - g) - D$$

$$x_4 = (1 - f)(1 - g) + D$$

	RHS	^RHS
LHS	x_1	x_2
^LHS	x_3	x_4

$$g = x_1 + x_2$$

$$f = x_1 + x_3$$

$$\sum_{i=1}^{4} x_i = 1, \; 0 \le x_i, i = 1 \ldots 4.$$

D can be extended to more dimensions…

Linkage Disequilibrium

$$\underline{X} = \underline{f} \otimes \underline{g} + D\underline{e} \otimes \underline{e}$$

where

$$\underline{X} = (x_1, x_2, x_3, x_4)$$

$$\sum_{i=1}^{4} x_i = 1,\ 0 \le x_i, i = 1\ldots4.$$

$$\underline{f} = (f, 1 - f)$$

$$\underline{g} = (g, 1 - g)$$

$$\underline{e} = (1, -1)$$

$$f = x_1 + x_3$$

$$g = x_1 + x_2$$

An algebraic observation…

2

Relative Linkage Disequilibrium

D is the distance from the point corresponding to the contingency table in the simplex, to the surface D=0 in the e⊗e direction.

$$RLD = \frac{D}{D_M}$$

D_M is the distance from the point corresponding to the contingency table on the surface D=0 in the e⊗e direction, to the surface of the simplex, in that direction.

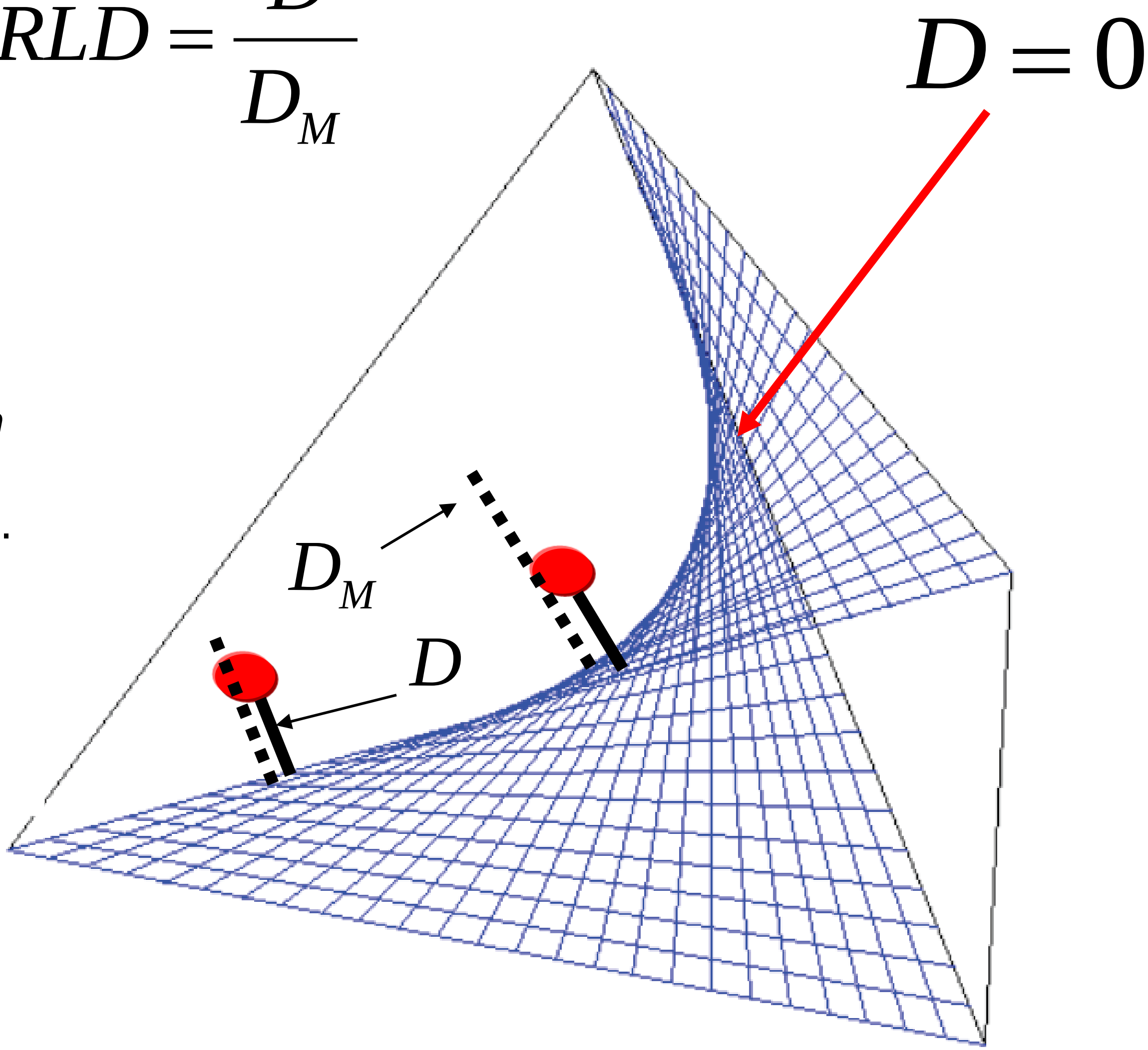
$$\begin{aligned}
&If \quad D > 0 \\
&then \\
&if \quad x_3 < x_2 \\
&\qquad then \qquad RLD = \frac{D}{D + x_3} \\
&\qquad else \qquad RLD = \frac{D}{D + x_2}
\end{aligned}$$

$$\begin{aligned}
&else \\
&if \quad x_1 < x_4 \\
&\qquad then \qquad RLD = \frac{D}{D - x_1} \\
&\qquad else \qquad RLD = \frac{D}{D - x_4}
\end{aligned}$$

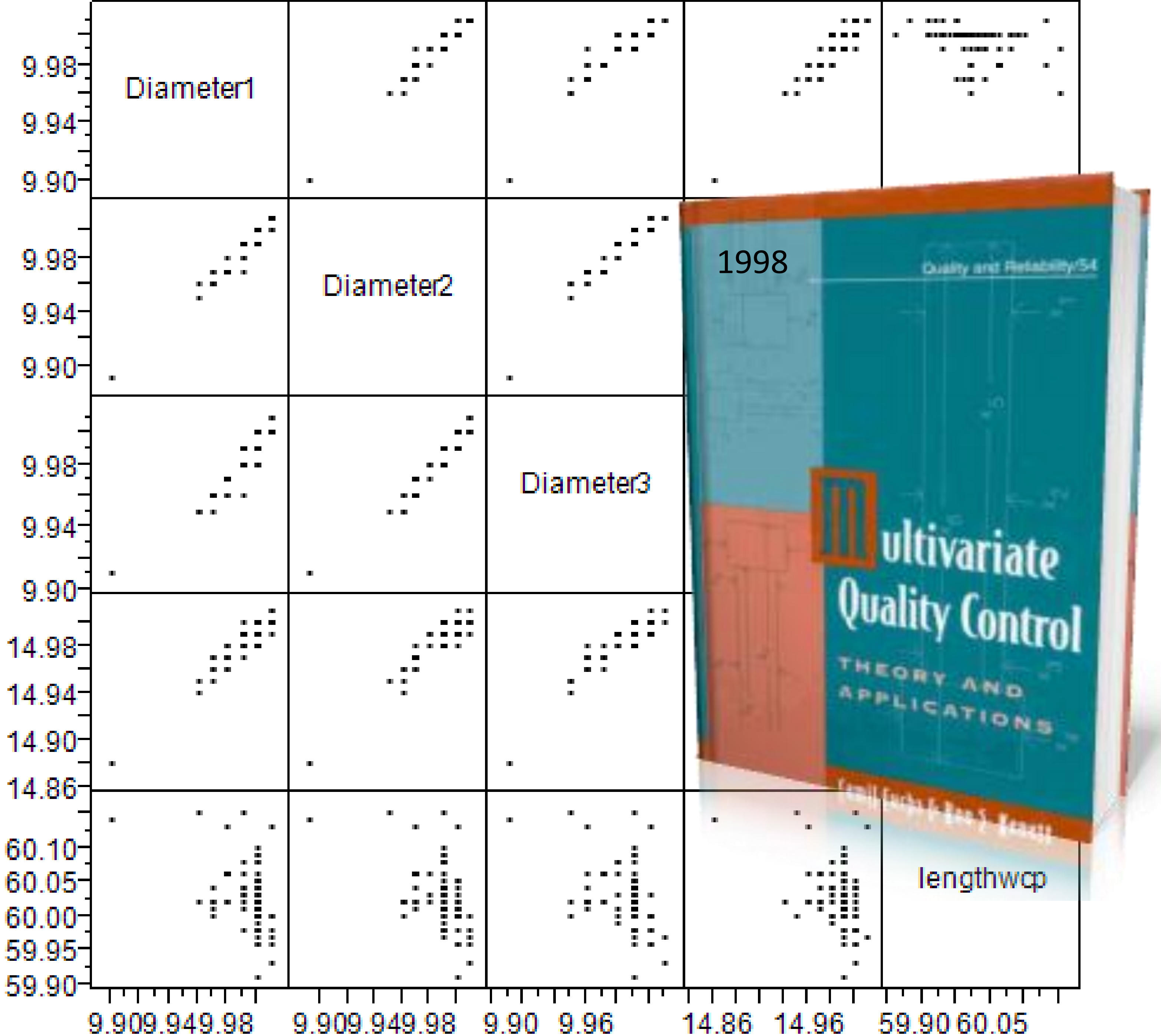
$$RLD = \frac{D}{D_M}$$

Kenett, R. and Salini, S., "Relative Linkage Disequilibrium Applications to Aircraft Accidents and Operational Risks" . *Trans. on Machine Learning and Data Mining*, 1,(2), pp. 83-96, 2008.

Implemented in arules R Package. Version 0.6-6, Mining Association Rules and Frequent Itemsets

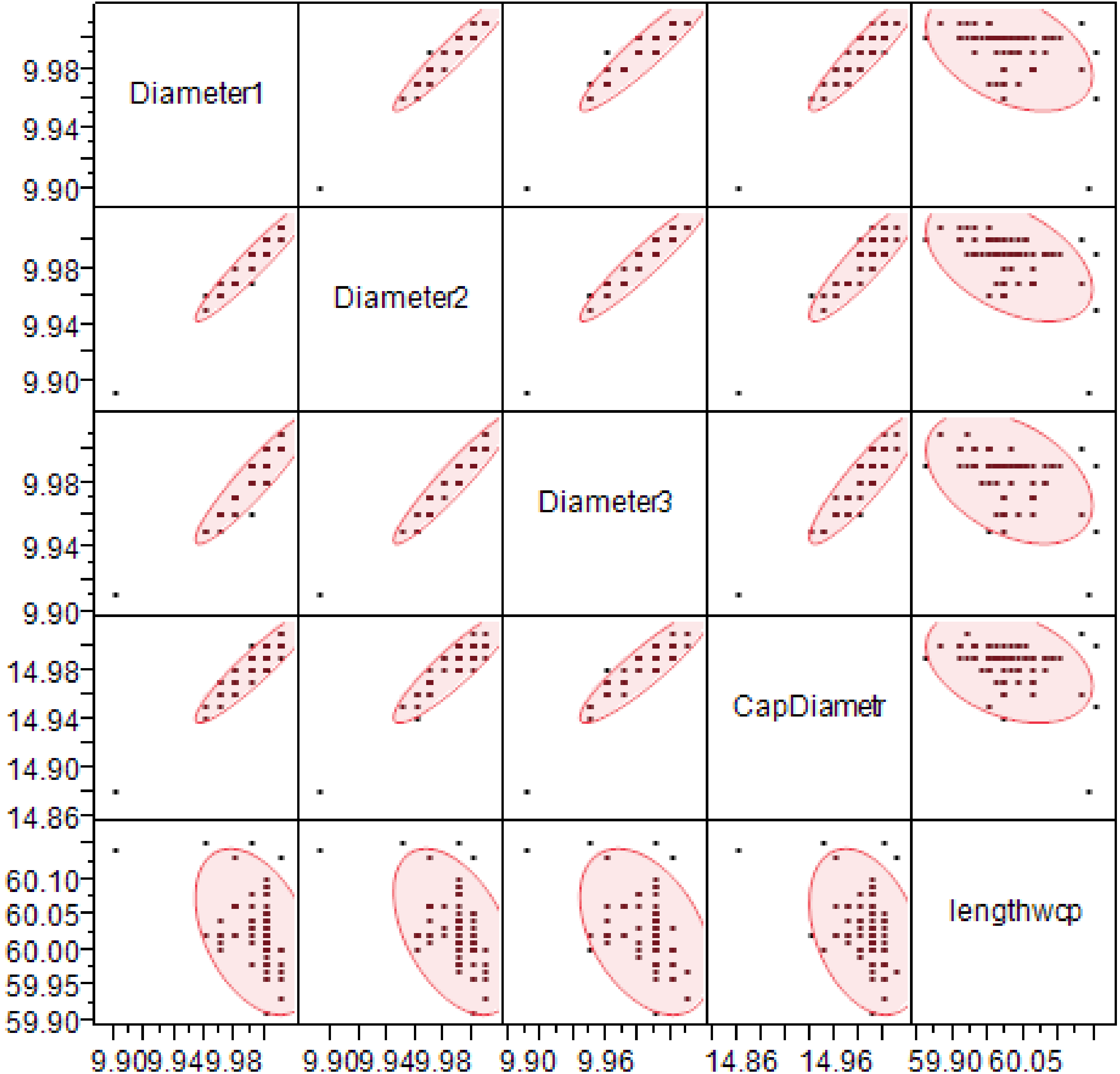


2

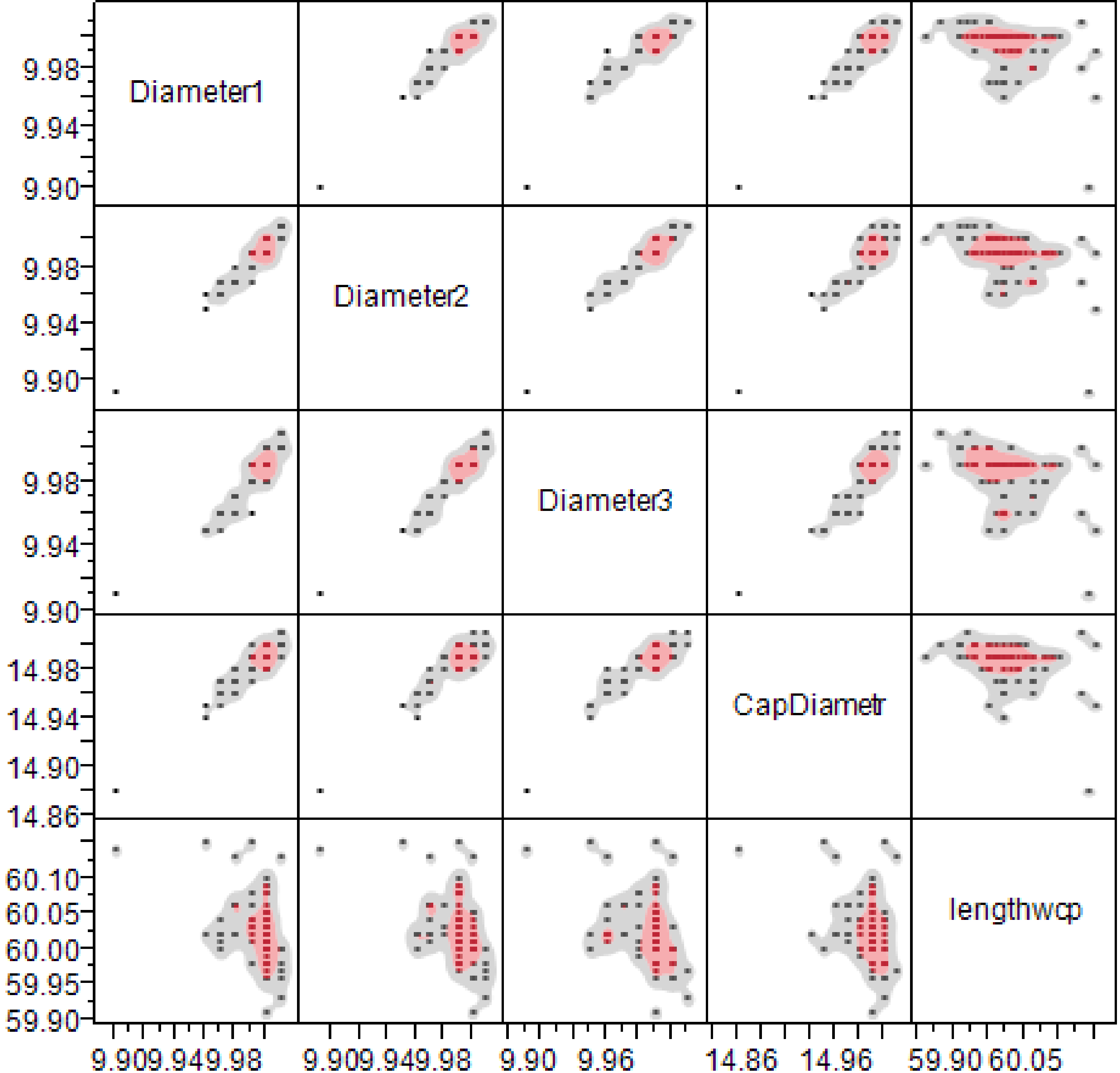


Diameter1
Diameter2
Diameter3
lengthwcp
9.98
9.94
9.90
14.98
14.94
14.90
14.86
60.10
60.05
60.00
59.95
59.90
9.909.949.98
9.909.949.98
9.90 9.96
14.86 14.96
59.90 60.05
1998
Quality and Reliability/54
Multivariate Quality Control
THEORY AND APPLICATIONS

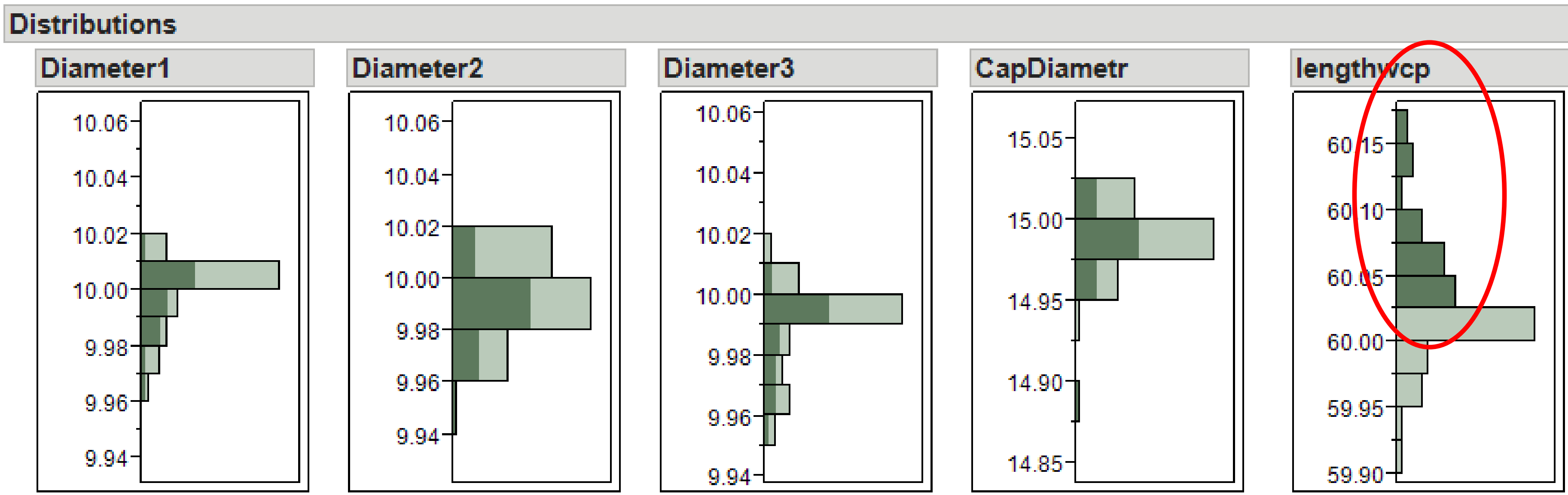
2


Diameter1
Diameter2
Diameter3
CapDiametr
lengthwcp
9.98
9.94
9.90
14.98
14.94
14.90
14.86
60.10
60.05
60.00
59.95
59.90
9.909.949.98
9.909.949.98
9.90 9.96
14.86 14.96
59.90 60.05

2

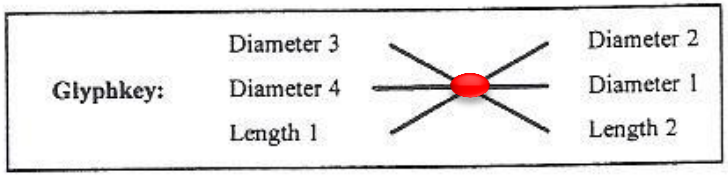

Diameter1
Diameter2
Diameter3
CapDiametr
lengthwcp
9.98
9.94
9.90
14.98
14.94
14.90
14.86
60.10
60.05
60.00
59.95
59.90
9.909.949.98
9.909.949.98
9.90 9.96
14.86 14.96
59.90 60.05

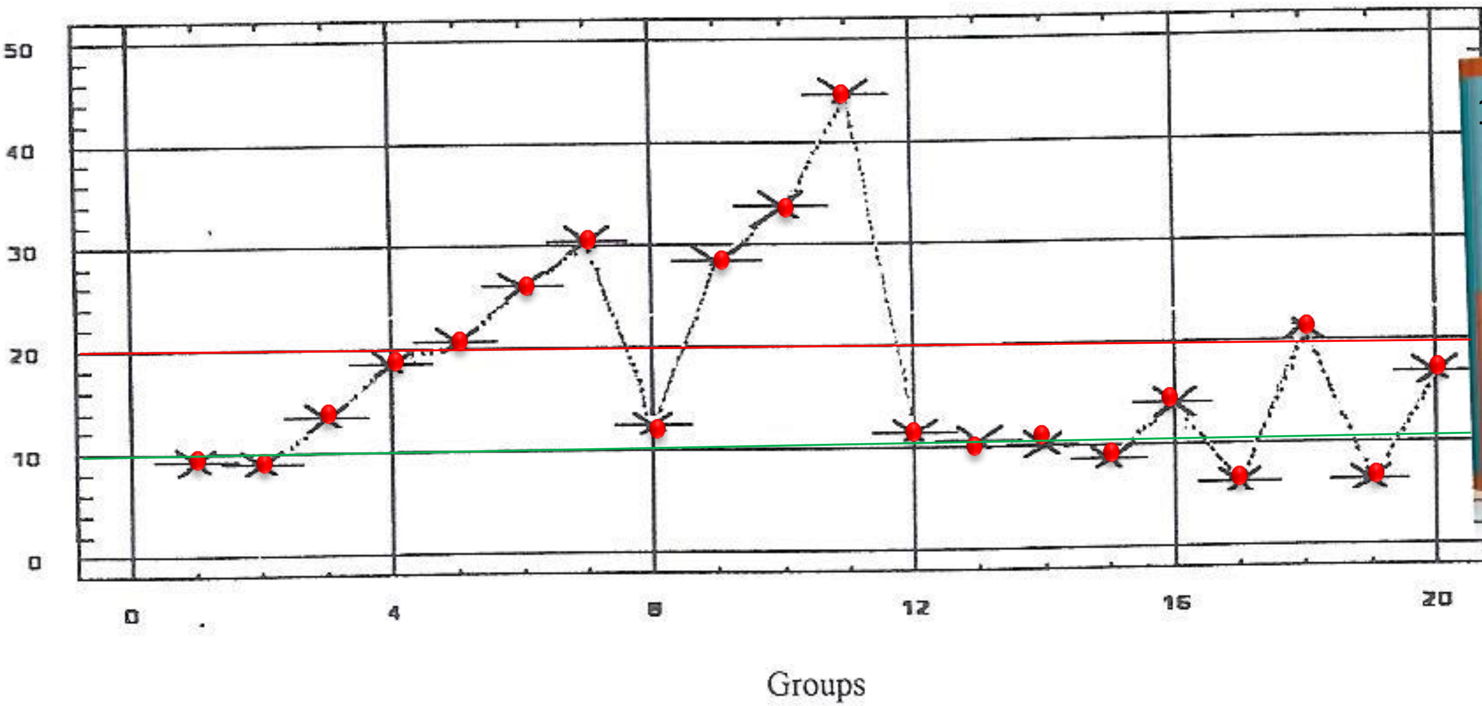
2



Distributions
Diameter1
Diameter2
Diameter3
CapDiametr
lengthwcp
10.06
10.04
10.02
10.00
9.98
9.96
9.94
15.05
15.00
14.95
14.90
14.85
60.15
60.10
60.05
60.00
59.95
59.90

2





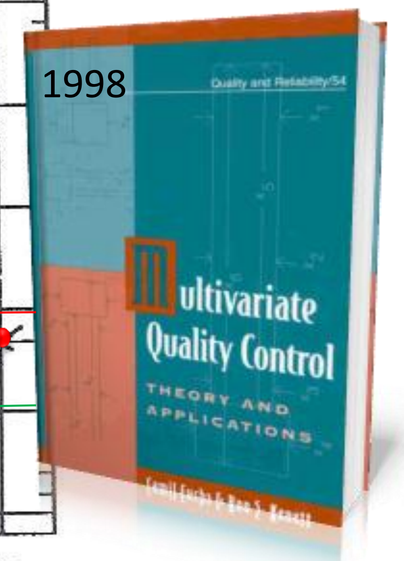


Figure 11.2: Starplot of the 6 variables from Case Study 1.

Starplots

The starplot was apparently first developed at the SCS corporation as an enhancement to the Multivariate Control Charts available through

3

(ARL_0 and ARL_1) or (PFA and CED)?

Special Issue Article

Quality and Reliability Engineering International

(wileyonlinelibrary.com) DOI: 10.1002/qre.1436

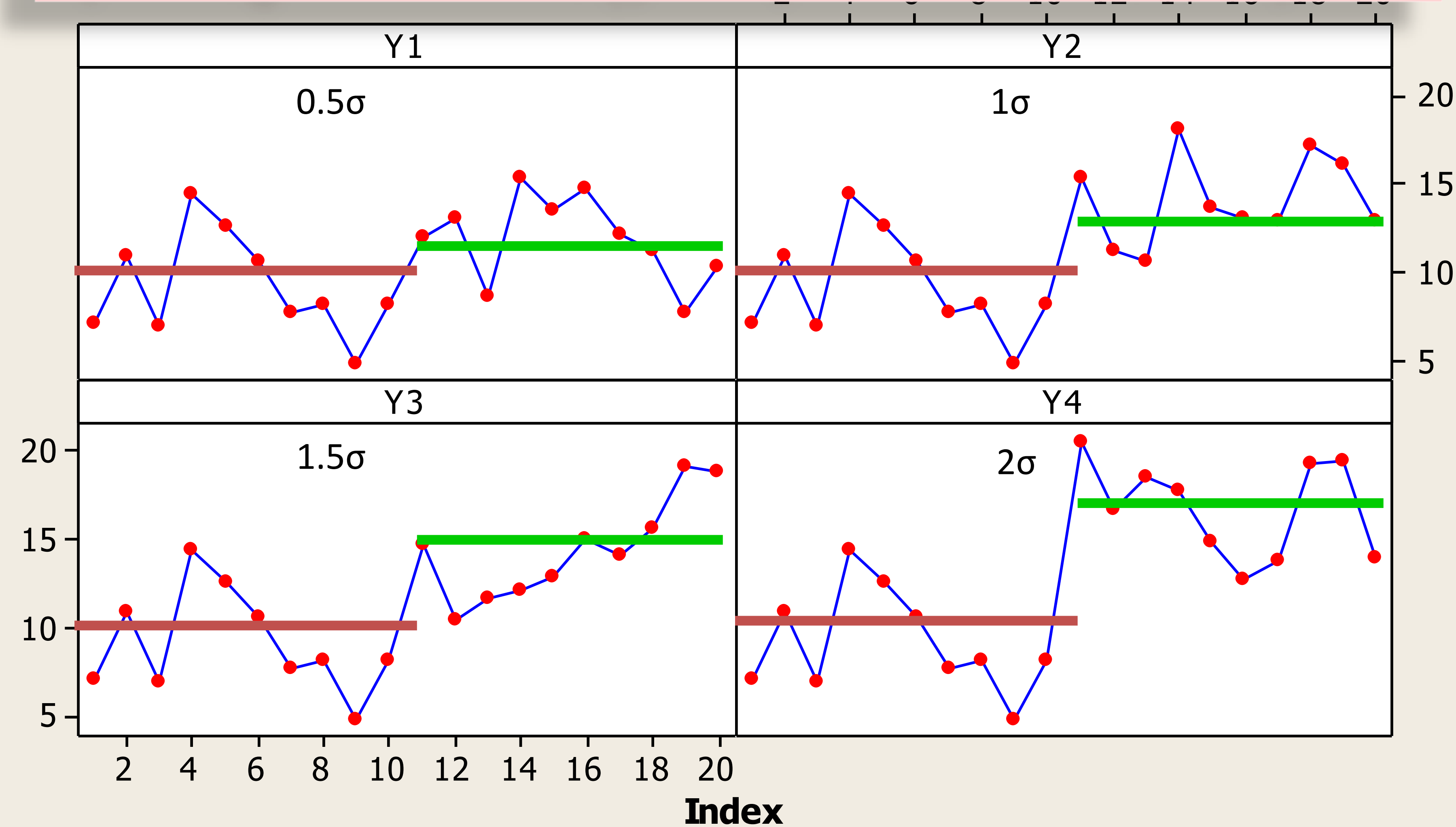
Published online in Wiley Online Library

On Assessing the Performance of Sequential Procedures for Detecting a Change

Ron S. Kenett[a,b*†] and Moshe Pollak[c]

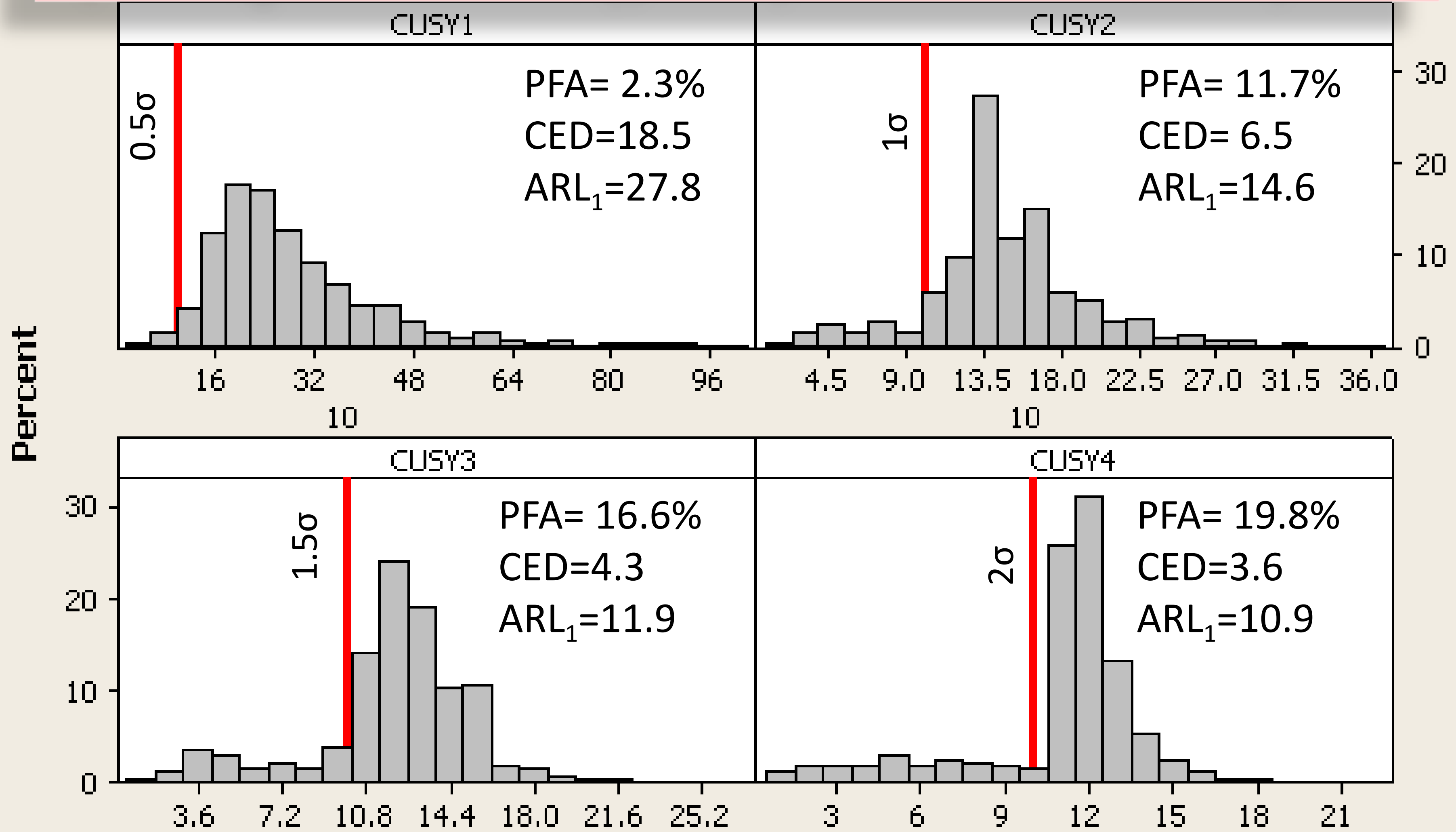
The literature on statistical process control has focused mostly on the average run length (ARL) to an alarm, as a performance criterion of sequential schemes. When the process is in control, this is the ARL to false alarm, generally denoted by ARL_0, and represents the in-control operating characteristic of the procedure. The ARL from the occurrence of a change to its detection represents an out-of-control operating characteristic and is typically embodied by ARL_1, the ARL to detection assuming that the change occurs at the very start of surveillance. However, these indices do not tell the whole story, and at times they are not defined well by a single number. We review the role of various operating characteristics in assessing performance of sequential procedures in comparison with ARL_0 and ARL_1. Copyright © 2012 John Wiley & Sons, Ltd.

$(ARL_0$ and $ARL_1)$ or (PFA and CED)?



Kenett, R.S. and Pollak, M. (2012) On Assessing the Performance of Sequential Procedures for Detecting a Change , *Quality and Reliability Engineering International*, 28, pp. 500–507

(ARL_0 and ARL_1) or (PFA and CED)?



Kenett, R.S. and Pollak, M. (2012) On Assessing the Performance of Sequential Procedures for Detecting a Change , *Quality and Reliability Engineering International*, 28, pp. 500–507

4

Are we making an impact?

Practical Statistical Efficiency (PSE)

PSE = E{R} x T {I} x P {I} x V {PS} x P {S} x V {P} x V {M} x V {D}

V{D} = value of the data actually collected

V{M} = value of the statistical method employed

V{P} = value of the problem to be solved

P{S} = probability that the problem actually gets solved

V{PS} = value of the problem being solved

P{I} = probability the solution is actually implemented

T{I} = time the solution stays implemented

E{R} = expected number of replications

Kenett, Coleman, Stewardson

E.J.G. Pitman

C. Eisenhart

B. Hoadley

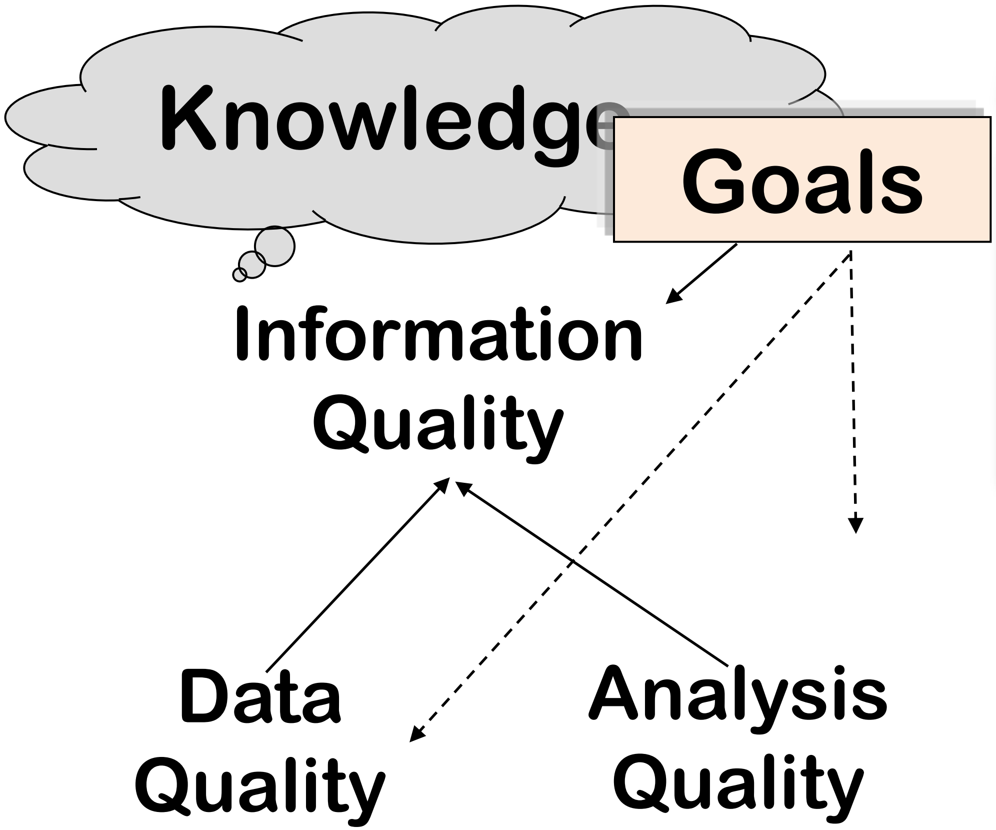
A.B. Godfrey

Kenett, R.S., Coleman, S.Y. and Stewardson, D. (2003). Statistical Efficiency: The Practical Perspective, *Quality and Reliability Engineering International*, 19: 265-272.

5

Are we generating knowledge?

Information Quality (InfoQ)



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

g	A specific analysis goal
X	The available dataset
f	An empirical analysis method
U	A utility measure

What

How

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

Primary Data
- Experimental
- Observational

Secondary Data
- Experimental
- Observational

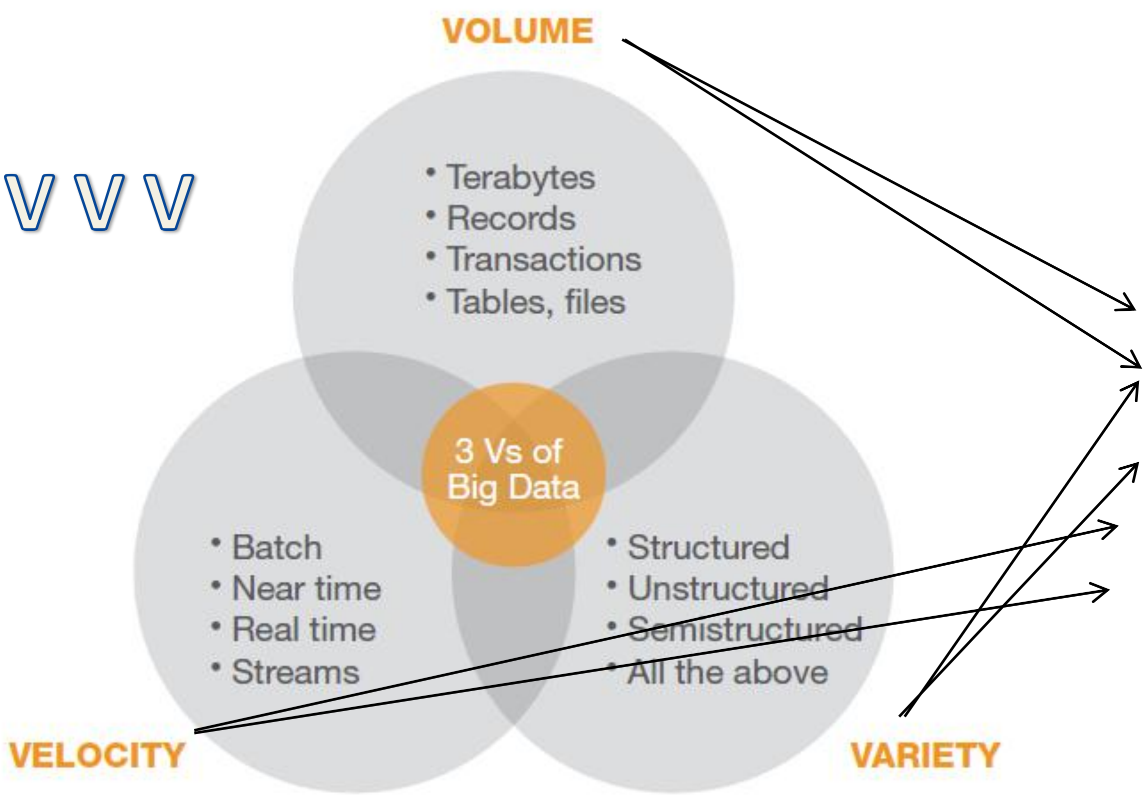
Kenett, R.S. and Shmueli , G. (2013) On Information Quality, http://ssrn.com/abstract=1464444
Journal of the Royal Statistical Society, Series A (with discussion), 176(4).

5

Big Data
VVV and InfoQ

Power	Prefix
10^9	Giga
10^{12}	Tera
10^{15}	Peta
10^{18}	Exa
10^{21}	Zetta
10^{24}	Yotta

V V V



InfoQ

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

Russom, P. (2011) *Big Data Analytics,* TDWI Best Practices Report, Q4.

5

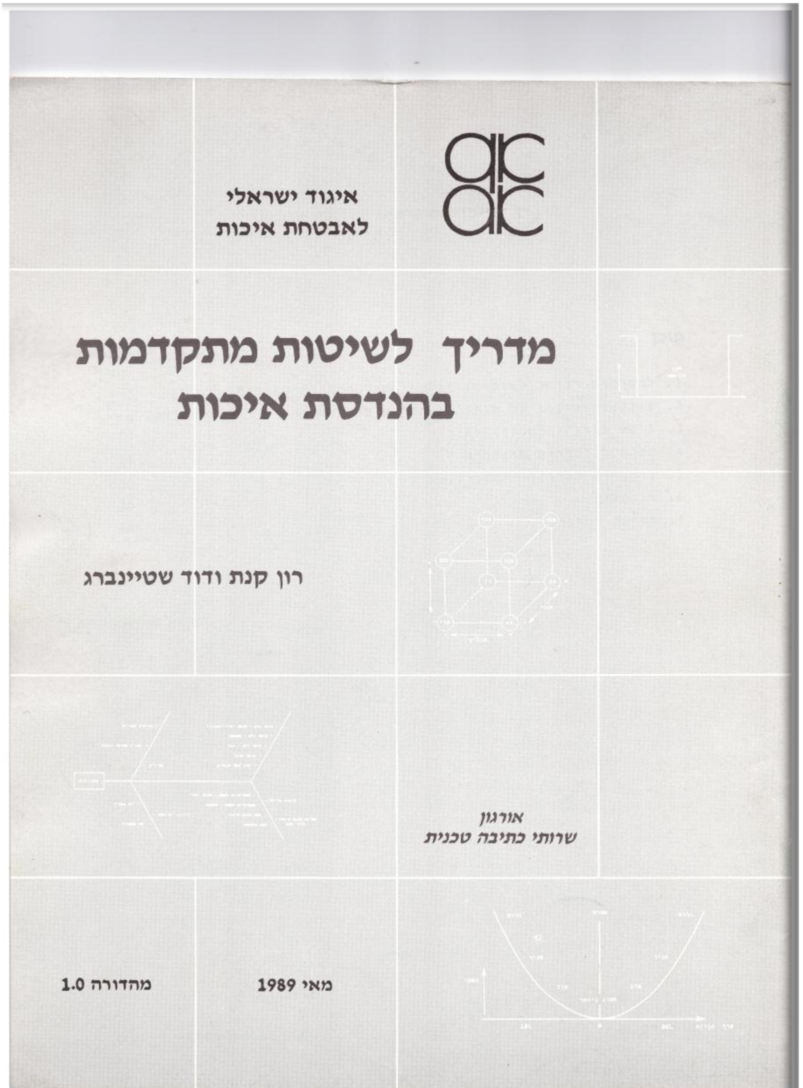
Dimensions of Information Quality (InfoQ)

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

Mallows
Hand
Huber
Cox
Juran
Box
Deming
Greenfield

Kenett, R.S. and Shmueli , G. (2013) On Information Quality, http://ssrn.com/abstract=1464444
Journal of the Royal Statistical Society, Series A (with discussion), 176(4).


Modern Industrial Statistics
Design and Control of Quality and Reliability
Ron S. Kenett
Shelemyahu Zacks
Wiley
Encyclopedia of Statistics in
QUALITY and RELIABILITY
EDITORS
Fabrizio Ruggeri
Ron S. Kenett
Frederick W. Faltin
Statistical Methods in Healthcare
Quality and Reliability/54
Multivariate Quality Control
Theory and Applications
Modern Industrial Statistics with R, MINITAB and JMP
ESTADÍSTICA INDUSTRIAL MODERNA
Diseño y control de la calidad y la confiabilidad
SOFTWARE PROCESS QUALITY
MANAGEMENT AND CONTROL
DISTRIBUTION OF REVIEW RATE
Too Slow
Too Fast
RON S. KENETT
EMANUEL R. BAKER
Process Improvement and CMMI for Systems and Software
Ron Kenett
Emanuel Baker
CRC Press
Taylor & Francis Group
AN AUERBACH BOOK
Editors
RON KENETT
YOSSI RAANAN
Operational Risk Management
A practical approach to intelligent data analysis
WILEY
STATISTICS IN PRACTICE
Modern Analysis of Customer Surveys: with Applications using R
RON S. KENETT
SILVIA SALINI
WILEY
STATISTICS IN PRACTICE


איגוד ישראלי
לאבטחת איכות
מדריך לשיטות מתקדמות
בהנדסת איכות
רון קנת ודוד שטיינברג
אורגון
שרותי כתיבה טכנית
מאי 1989
מהדורה 1.0


קמיל פוקס שאול בר־לב
אמת
וסקר
קולות החיילים יכריעו
חשיבותם של סקרים
פרס מוביל ב־3 אחוזים

RSS Greenfield Medalists

Year	Name	Year	Name
1991	D Price	1992	T P Davis & D M Grove
1995	A Racine-Poon	1998	R Caulcutt & M Gerson
2000	L Furlong	2005	Susan Lewis
2007	S J Morrison	2010	D Montgomery

CREATE IMPACT

GENERATE KNOWLEDGE