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

### Maximizing Data Science Success with Information Quality (InfoQ) and JMP

Ron Kenett and Chris Gotwalt

https://community.jmp.com/t5/Discovery-Summit-Europe-2017/Plenary-Session-From-Quality-by-Design-to-Information-Quality-A/ta-p/37537



#### Plenary Session From Quality by Design (QbD) to Information Quality (InfoQ)



#### Ron S. Kenett

Research Professor Mathematics Department University of Turin



Here we show how InfoQ can help you achieve more with **JMP** 

Statistical Discovery - From SAS.

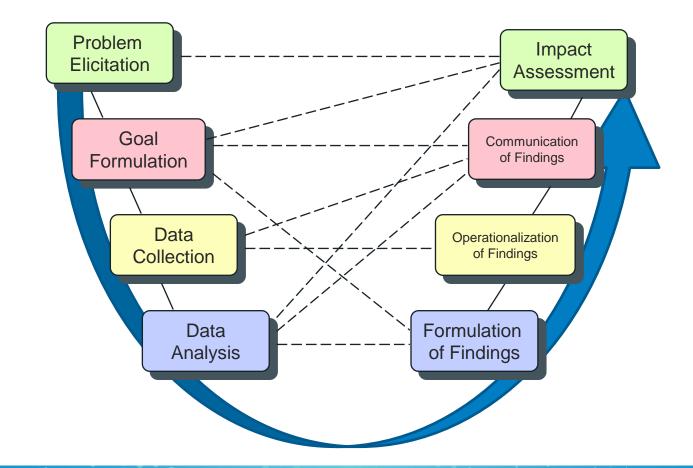
#### Abstract

# Information quality definition: The potential of a particular dataset to achieve a particular goal using a given empirical analysis method

Data analysis, from designed experiments to machine learning, is being deployed at an accelerating rate. At the same time, issues like the reproducibility and p-value controversy have made us increasingly aware that well intentioned statistical analyses can still lead to mistaken conclusions and bad decisions. For a data analysis project to fulfill its goals, one must assess the scope and strength of the conclusions possible given the data and tools available. This thinking process is best realized within a framework that isolates the components of the project: the goals, data collection procedure, data properties, the analysis provided, etc. The InfoQ Framework provides simple procedures for making this assessment and is easy to operationalize in JMP. In this presentation, we give an overview of InfoQ, and use case studies drawing from consumer research and pharmaceutical manufacturing to illustrate how JMP can be used to make an InfoQ assessment, highlighting situations of both high and low InfoQ. We also give tips showing how JMP can, in some cases, be used to increase information guality without acquiring more data.

### TOC

- Introduction to information quality (InfoQ)
- The case study
- An information quality assessment
- How JMP supports InfoQ



Kenett, R.S. (2015) A life cycle view of statistics Quality Engineering (with discussion), Vol. 27, No.1,pp. 111-129

# **Information Quality**

6

g

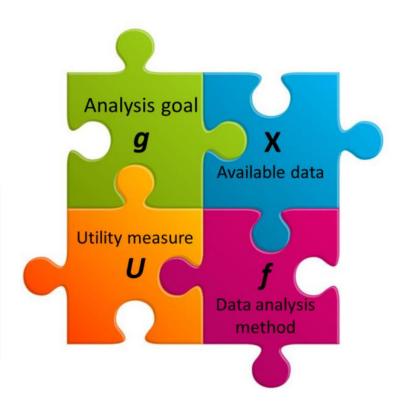
Х

U

The potential of a particular dataset to achieve a particular goal using a given empirical analysis method

A specific analysis goal The available dataset

- f An empirical analysis method
  - A utility measure



Introduction

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

Kenett, R.S. and Shmueli, G. (2013) On Information Quality. Journal of the Royal Statistical Society, Series A (with discussion), 176(4).

#### RON S. KENETT | THOMAS C. REDMAN

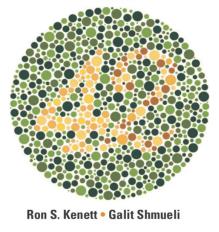
# THE REAL WORK OF

HOW TO TURN DATA INTO INFORMATION, BETTER DECISIONS, AND STRONGER ORGANIZATIONS



**Information Quality** 

The Potential of Data and Analytics to Generate Knowledge



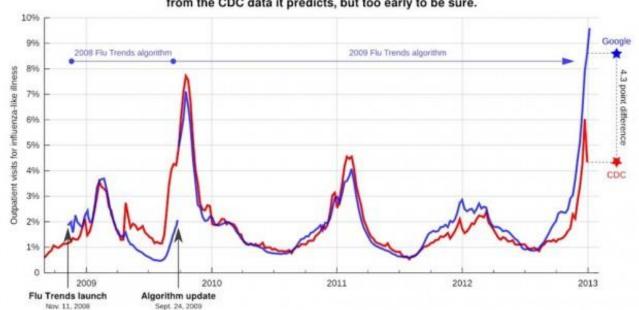
WILEY



- 1. Data resolution
- 2. Data structure
- 3. Data integration
- 4. Temporal relevance
- 5. Chronology of data and goal

- 6. Generalizability
- 7. Operationalization
- 8. Communication

#### **#1 Data Resolution**



Google Flu Trends U.S. may have diverged again from the CDC data it predicts, but too early to be sure.

Sources: http://www.google.org/fisteroly.up.cDC.U.reit.data.from.http:/lgis.cdc.giv/group/fisz/ew/Mispontalan/board.mml. Cook et al. (2011) Assessing Google Far Trends Performance in the United States during the 2009 influenza Vrius A (HUN1) Pandemic. PluoS ONE (90): +239530, doi:10.1371/bjournal.pome.0025510.

Introduction

Data as of Jan. 12, 2013. Keith Winstein (keithw@mit.eitu)

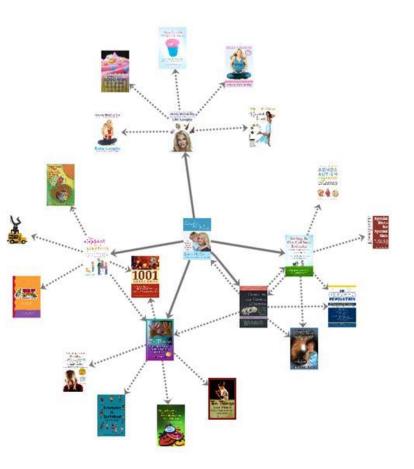
### **#2 Data Structure**

#### **Data Types**

- Time series, cross-sectional, panel
- Structured, semi-, non-structured
- Geographic, spatial, network
- Text, audio, video, semantic
- Discrete, continuous

#### **Data Characteristics**

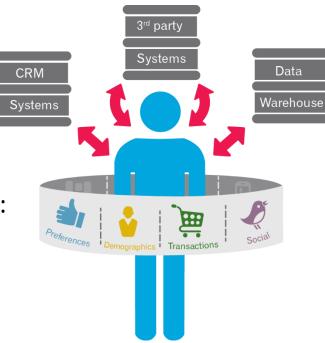
Corrupted and missing values due to study design or data collection mechanism



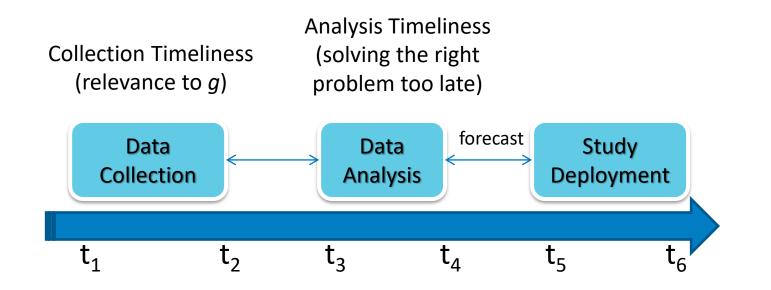
### **#3 Data Integration**



#### **Linkage, privacy-preserving methods**: Increase or decrease InfoQ?



### **#4 Temporal Relevance**



g: Prospective vs. retrospective; longitudinal vs. snapshot. Nature of X, complexity of f

# **#5 Chronology of Data & Goal**



AIR QUALITY INDEX

Air Quality Index (AQI) Values	Levels of Health Concern
0 to 50	Good
51-100	Moderate
101-150	Unhealthy for Sensitive Groups
151-200	Unhealthy
201-300	Very Unhealthy
301 to 500	Hazardous

#### Data: Daily AQI in a city

g<sub>1</sub>: Reverse-engineer AQI

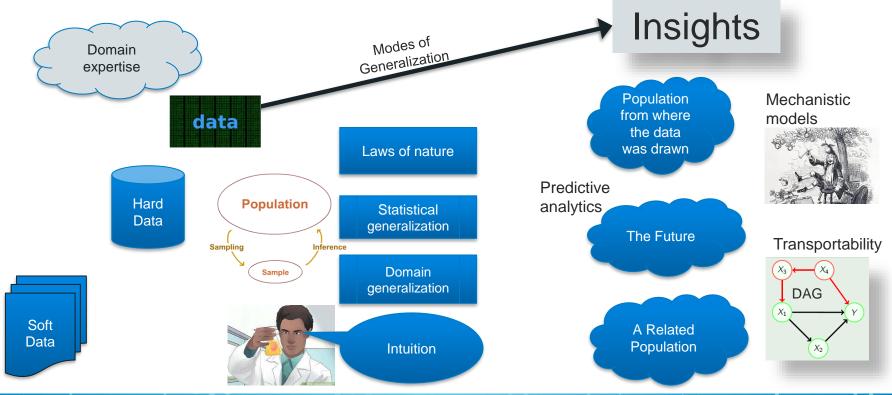
g<sub>2</sub>: Forecast AQI

Retrospective/prospective Ex-post availability Endogeneity

Introduction

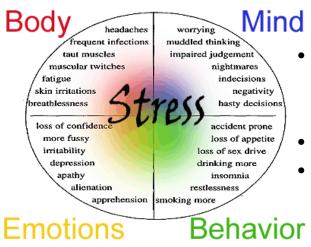
http://www.airnow.gov/?action=aqibasics.aqi

### **#6 Generalizability**



# **#7 (Construct) Operationalization**

X: construct X =  $\theta(\chi)$  operationalization (measur



 Causal explanation vs. prediction, description

- Theory vs. data
- Data: Questionnaire, physio measurement



# **#7 (Action) Operationalization**

"An operational definition consists of (1) a criterion to be applies to an object or a group of objects, (2) a test of compliance for the object or group and (3) a decision rule for interpreting the test results to whether the object or group is, or is not, in compliance"

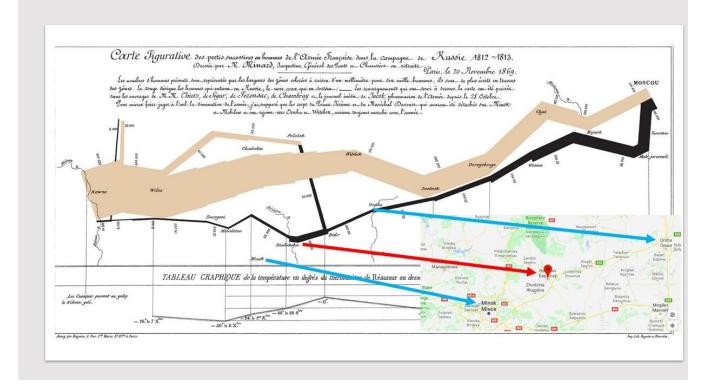


W.E. Deming (1982). Quality, Productivity and the Competitive Position, MIT Press

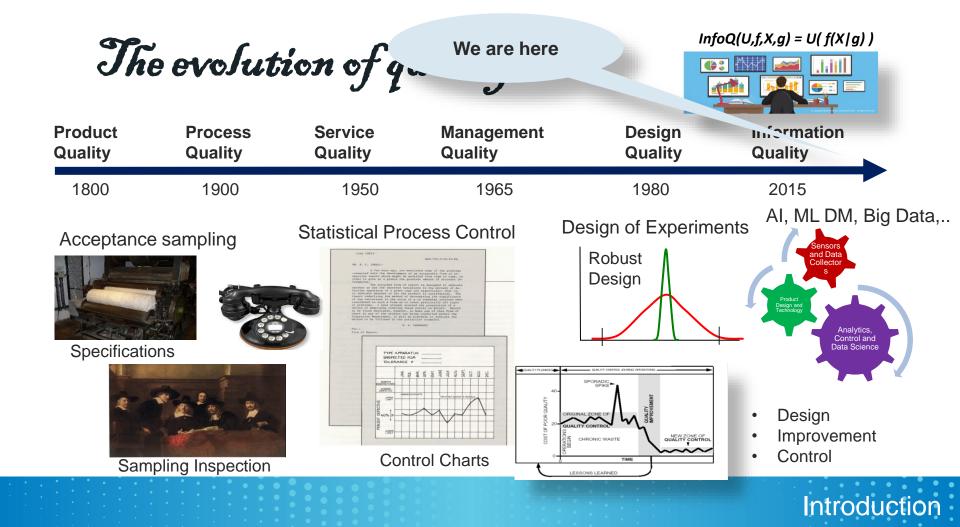
What do you want to accomplish?
By what method will you accomplish it?
How will you know when you have accomplished it?

### **#8 Communication**





https://www.nationalgeographic.org/thisday/jun24/napoleon-invades-russia/#:~:text=Minard-,On%20June%2024%2C%201812%2C%20the%20Grande%20Arm%C3%A9e%2C%20led%20by,more%20than%20500%2C000%20European%20troops.



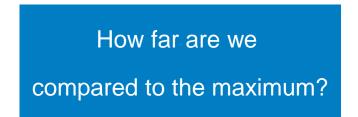
# **Assessing InfoQ**



#### Rating-based assessment (1-5 scale on each dimension)

#### InfoQ Score = $[d_1(Y_1) \ d_2(Y_2) \ \dots \ d_8(Y_8)]^{1/8}$

💱 InfoQ - JMP Pro	- 🗆 X
Help	Data Resolution
This is a rating-based approach to quantifying InfoQ that scores each of the eight dimensions. This coarse grained approach rates	Very Low 🖉 🕹 Very High
each dimension on a 5 point scale, with 5 indicating "Very High" schievement in that dimension.	Data Structure
	Very Low 🖉 Very High
he ratings are then normalized into a desirability function for each limension, which are then combined to produce an overall InfoQ.	Data Integration
core using the geometric mean of the individual desirabilities.	Very Low 🖉 Very High
ly dragging the slider handles, each dimension can be assigned a	Temporal Relevance
lausible range of ratings, or a specific rating.	Very Low 🖉 Very High
nfoQ	Chronology of Data and Goal
Lower Bound: Undefined Upper Bound: Undefined	Very Low 🛇 Very High
	Generalizability
	Very Low 🖉 Very High
	Operationalization
InfoQ score	Very Low 🖉 Very High
initia score	Communication
	Very Low 🛇 Very High



Information quality definition: The potential of a particular dataset to achieve a particular goal using a given empirical analysis method

Information quality assessment

#### The case study

The case study



