"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

23/3/2017

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

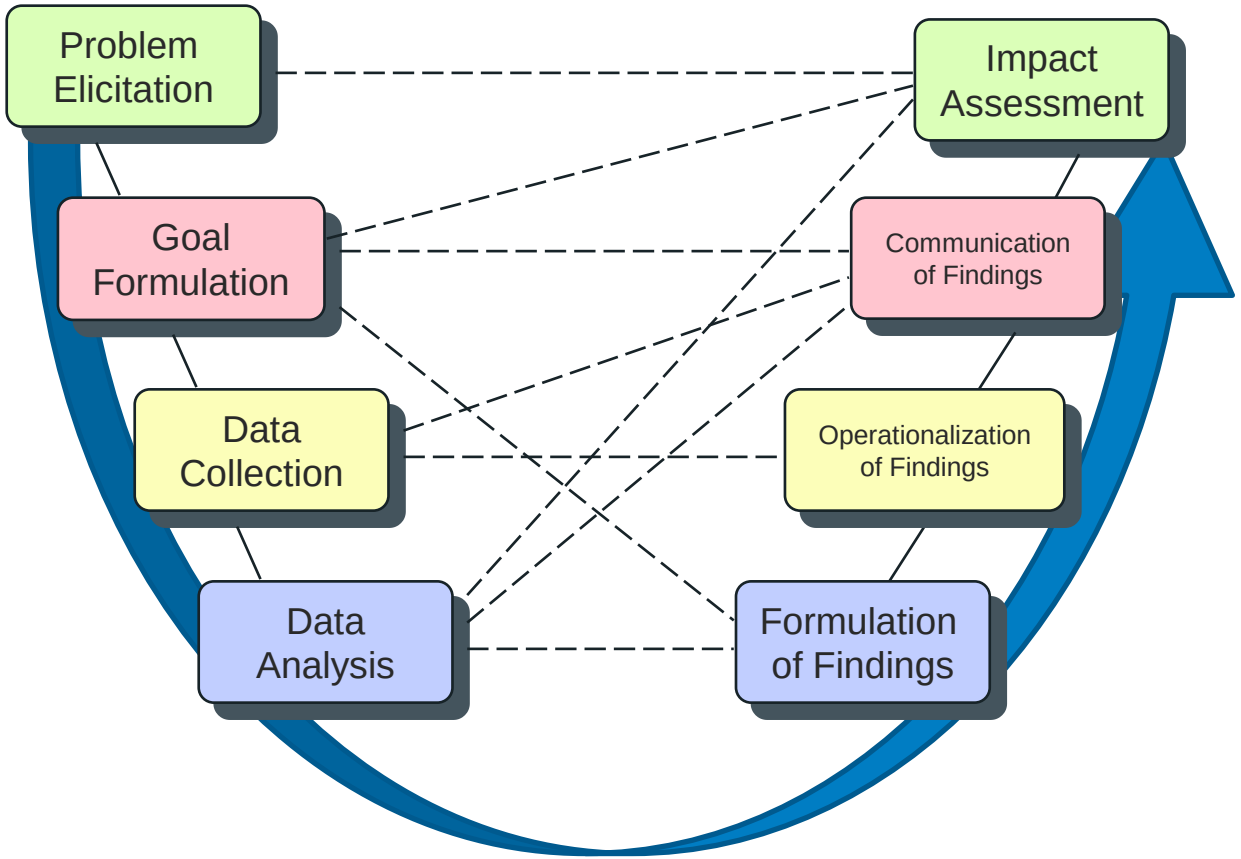
Abstract

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 quality without acquiring more data.

TOC

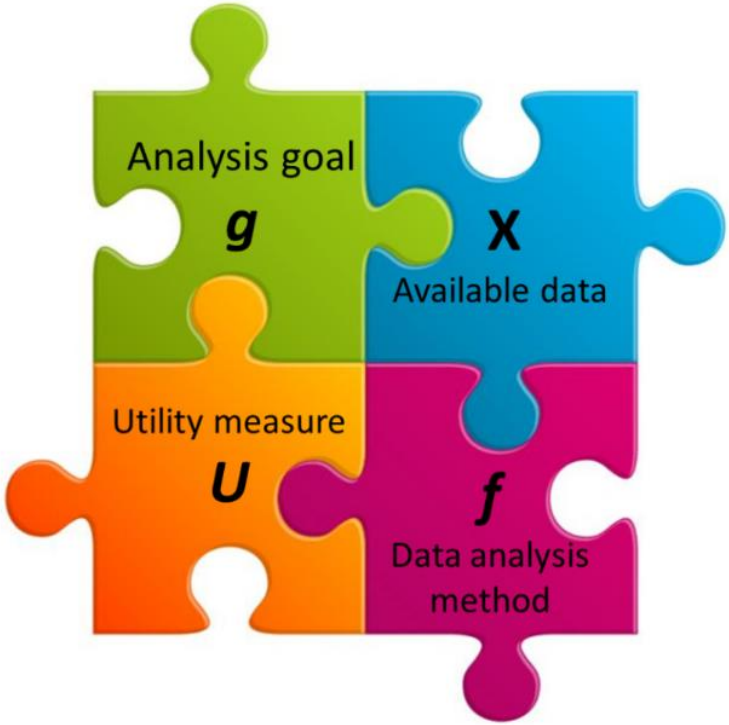
- Introduction to information quality (InfoQ)
- The case study
- An information quality assessment
- How JMP supports InfoQ


Problem Elicitation
Impact Assessment
Goal Formulation
Communication of Findings
Data Collection
Operationalization of Findings
Data Analysis
Formulation of Findings

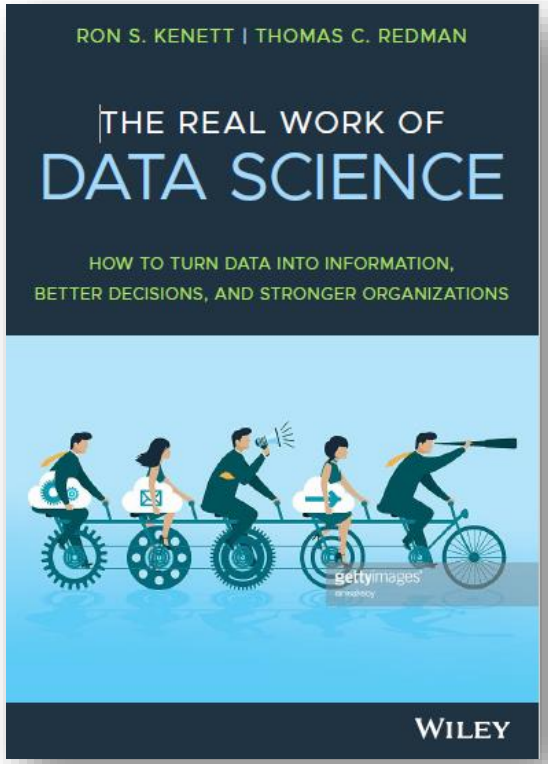
Information Quality

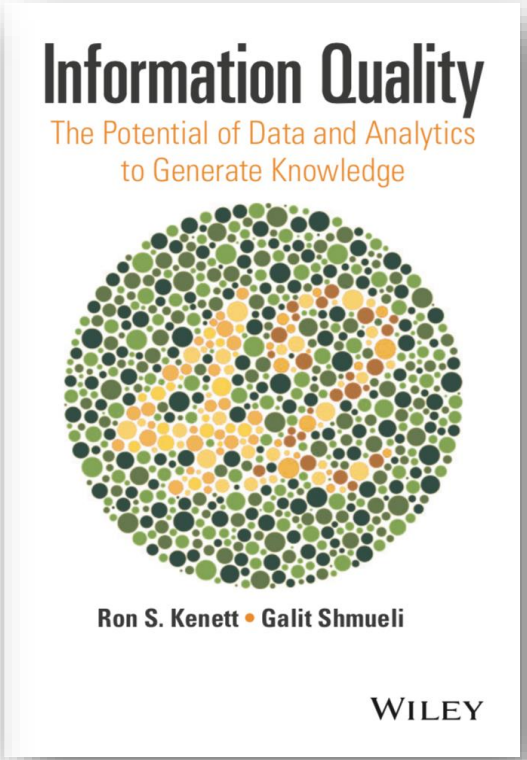
The potential of a particular dataset to achieve a particular goal using a given empirical analysis method

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



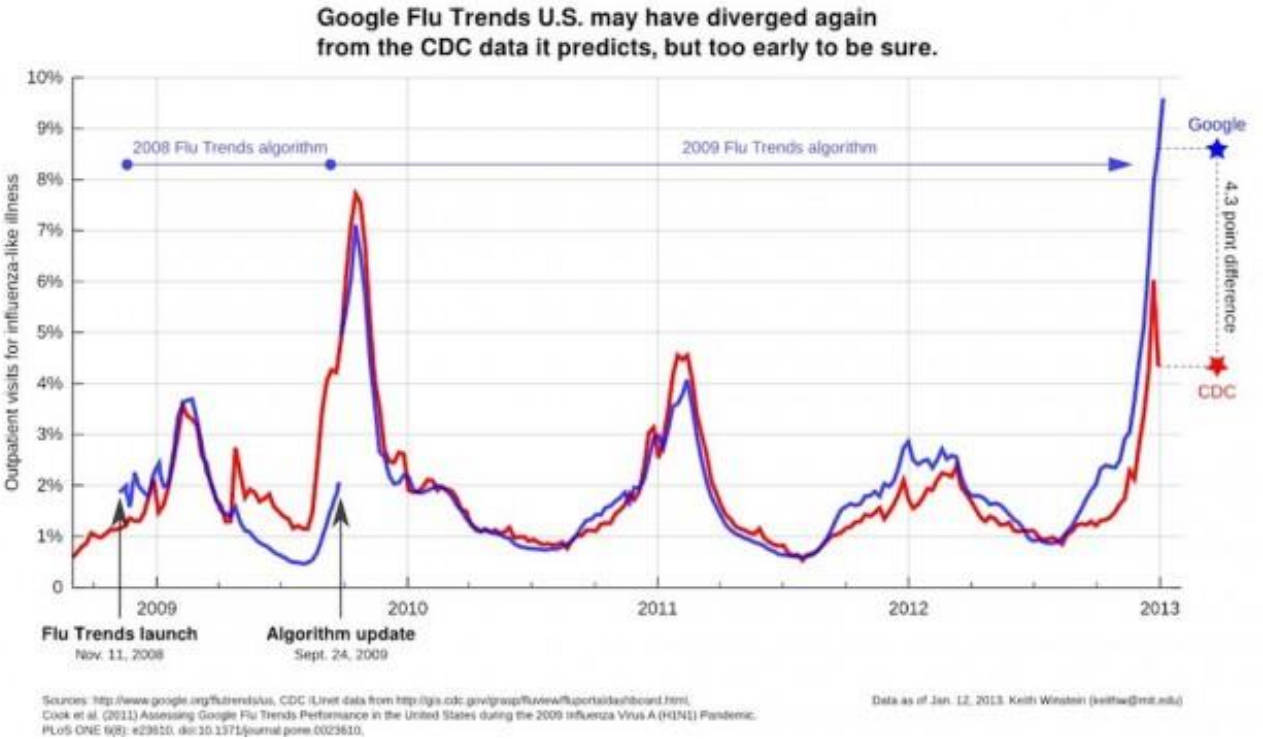
$$InfoQ(U,f,X,g) = U(f(X|g))$$





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



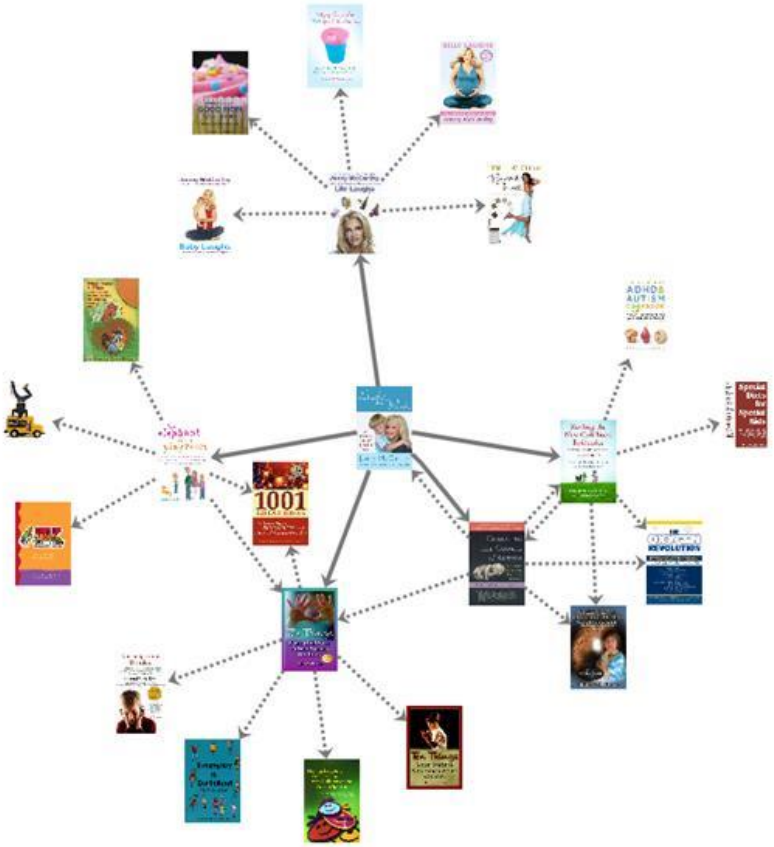
#2 Data Structure

Data Types

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

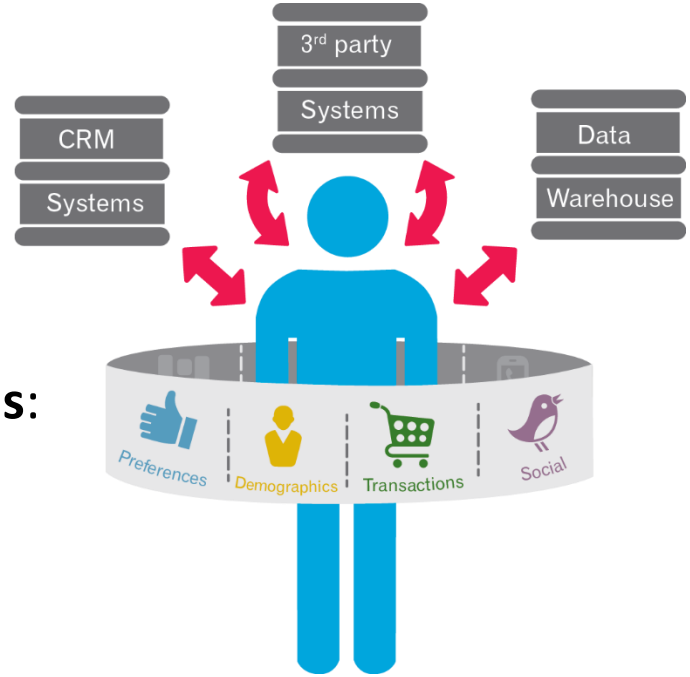
Data Characteristics

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

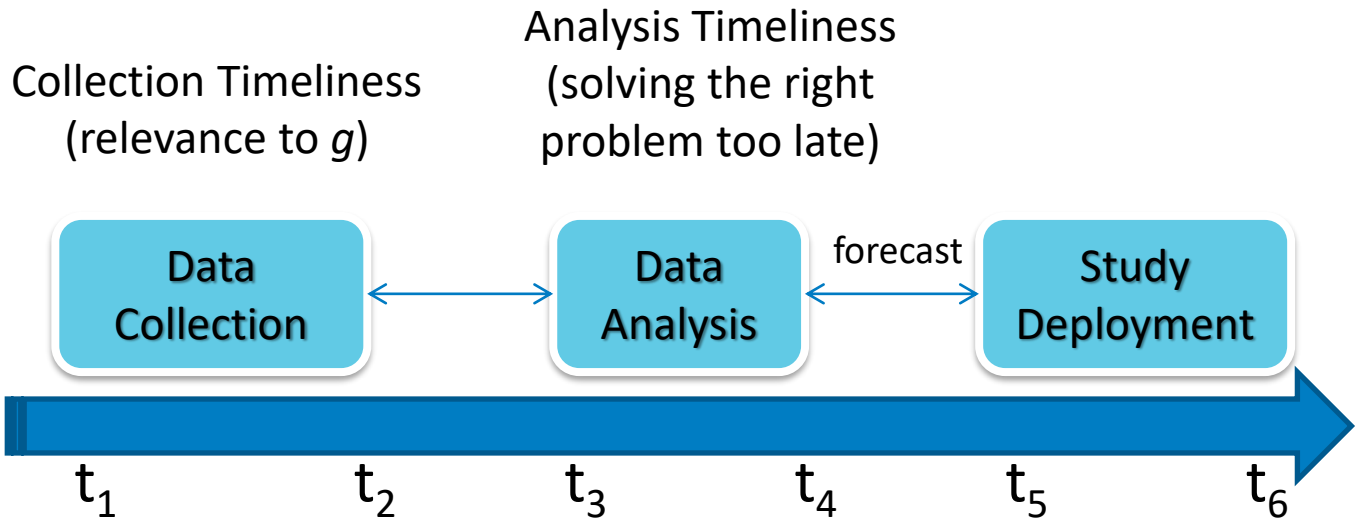


#3 Data Integration

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 (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

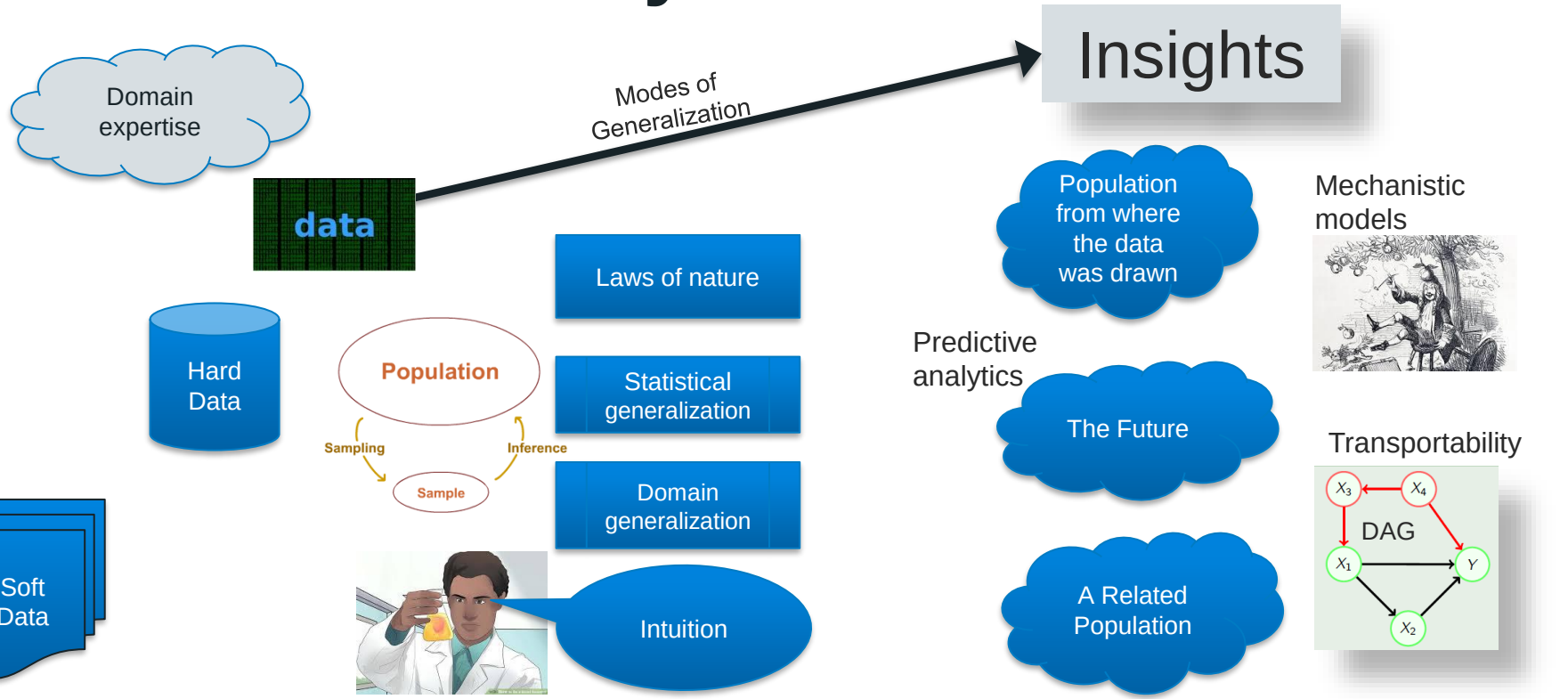
g_1: Reverse-engineer AQI

g_2: Forecast AQI

Retrospective/prospective
Ex-post availability
Endogeneity

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

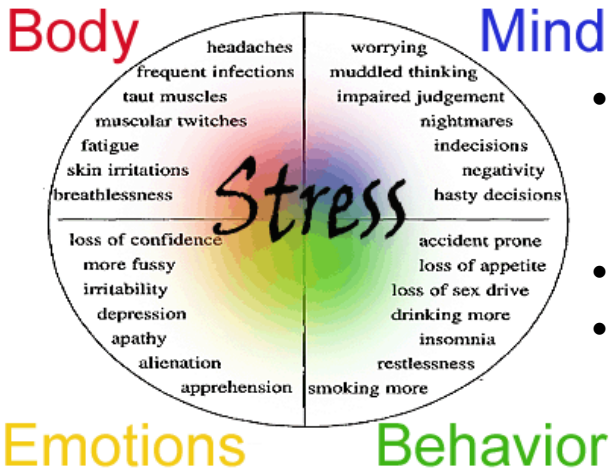
#6 Generalizability



#7 (Construct) Operationalization

X: construct

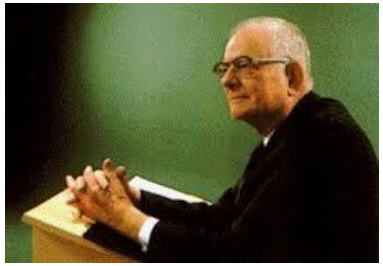
X = θ(χ) operationalization (measur



- 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”



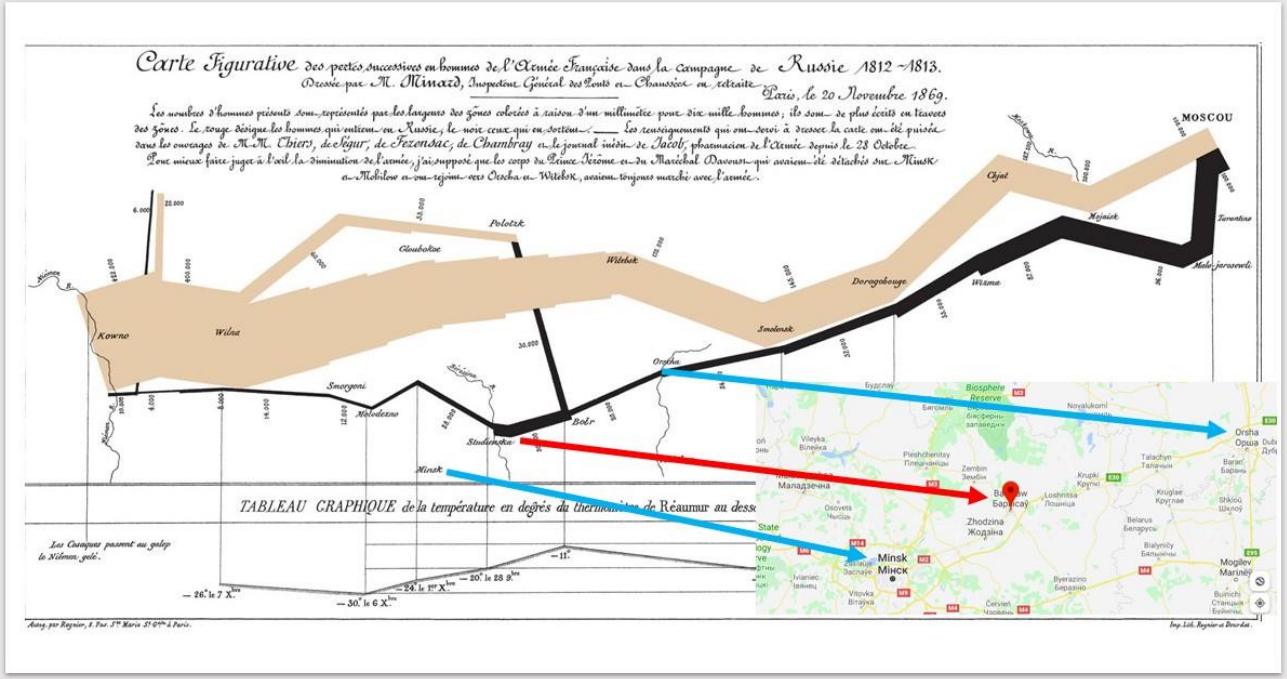
W.E. Deming (1982). *Quality, Productivity and the Competitive Position*, MIT Press

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

http://www.spcpress.com/pdf/DJW187.pdf

#8 Communication

Napoleons March.jmp



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.

The evolution of quality

We are here

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

Product Quality — 1800

Process Quality — 1900

Service Quality — 1950

Management Quality — 1965

Design Quality — 1980

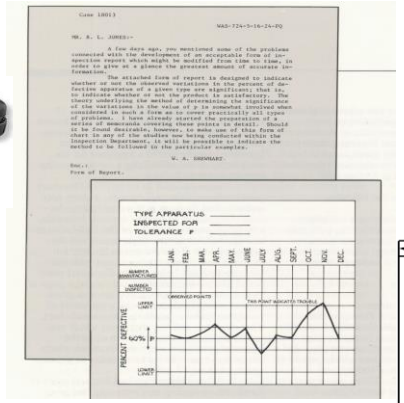
Information Quality — 2015

Acceptance sampling

Specifications

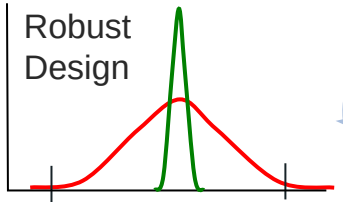
Sampling Inspection

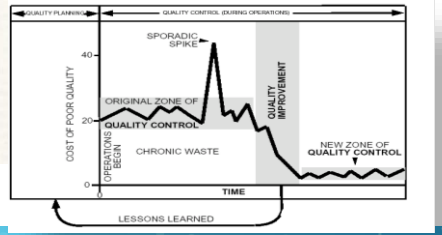
Statistical Process Control



Control Charts

Design of Experiments





AI, ML DM, Big Data,..

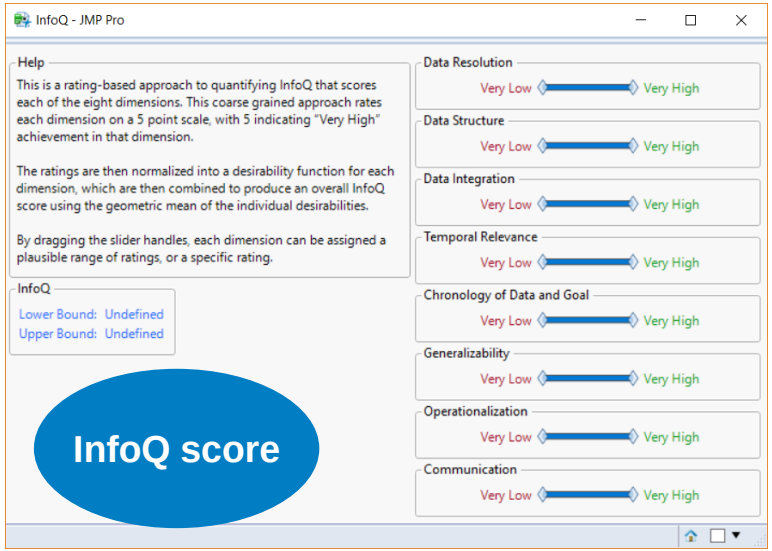
- Design
- Improvement
- Control

Assessing InfoQ

InfoQ.jmpaddin

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

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



How far are we

compared to the maximum?

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

The case study


jmp
Statistical Discovery™ From SAS.
Tailored analysis
Information quality