Information Quality (InfoQ)
Background

Figure 3: Index of Changing Work Tasks in the U.S. Economy 1960-2009

- Working with New Information
- Solving Unstructured Problems

The Skill Content of Recent Technical Change: An Empirical Investigation (updated data)
Data
From numbers to insights

- data
- statistical analysis
- findings
- information
- insight
Insights

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4619681/
Insights

$r = 0.791$
$P < 0.0001$

Nobel Laureates per 10 Million Population vs. Chocolate Consumption (kg/yr/capita)
More insights

The scatter plot shows the relationship between the number of Nobel Laureates per 10 million population and the number of IKEA stores per 10 million population across various countries. The correlation coefficient $r = 0.82$ is statistically significant at $P < 0.0001$. This suggests a strong positive correlation between the two variables.
How it is shown on the media:

Missing Data
How do you impute the missing data?
Applied statistics is about meeting the challenge of solving real world problems with mathematical tools and statistical thinking.
Why Research

The creation of new knowledge.

This includes

- new theory
- new methods
- assessing the practical value of methods
- creative application of existing methods
Research Motivation

• We find an interesting problem.
  – It requires novel use of old ideas.
  – It requires completely new ideas.

• We are stimulated by research that we read.
  – We have a better solution to the problem.
  – We explore the properties of the method.

• We have a customer.
Research Topics in Statistics/Quality

**SPC**
- Profile monitoring
- Image monitoring
- Multivariate SPC

**Reliability**
- Multi-layer, multi-scale surveillance
- Tests with multiple accelerating factors
- Degradation Models
Research Topics in Statistics/Quality

Design of Experiments
- Experiments for mixtures/Compositional Data
- Computer experiments
- Accelerated testing

Statistical Strategy
- Integrated models
- Life cycle views
- Information quality
Research Topics in Statistics/Quality

**Systems Thinking: Foundation, Uses and Challenges**
Nova Science Publishers, Moti Frank, Sigal Kordova and Haim Shaked (editors)

**Integrated Models in Healthcare Systems**

*Ron S. Kenett*
KPA Group, Raanana, Israel, Hebrew University School of Medicine, Jerusalem, Israel, Center for Risk Engineering, NYU-Poly, New York, USA and University of Turin, Italy

*Yifat Lavi*
University of Turin, Italy

---

**Integrated Models**

- Voice of the PATIENT
- Voice of the PROCESS
- Voice of WORKFORCE
Research Topics in Statistics/Quality

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

Unstructured problems

Academic courses

A life cycle view

Research Topics in Statistics/Quality

Information Quality
Information Quality: The Potential of Data and Analytics to Generate Knowledge

Part I: The Information Quality Framework
1. Introduction to information quality
2. Quality of data, quality of analysis
3. The dimensions of InfoQ
4. InfoQ at the study-design stage
5. InfoQ at the post-data collection stage

Part II: Applications of InfoQ
6. Education
7. Customer surveys
8. Healthcare
9. Risk management
10. Official statistics

Part III: Implementing InfoQ
11. InfoQ and reproducible research
12. InfoQ in review processes of scientific publications
13. Integrating InfoQ into applied statistics and data mining
14. Information quality support with R
15. Information quality support with MINITAB
16. Information quality support with JMP
Information Quality (InfoQ)

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

\[ \text{InfoQ}(f, X, g) = U \left( f(X \mid g) \right) \]

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

\[ \text{InfoQ}(f, X, g) = U(f(X|g)) \]

**InfoQ Dimensions**

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

**Domain Space**

- Data Quality
- Analysis Quality

**Analytic Space**

- Information Quality
- Goals

**Insights**

- How
- What

**What**

- Analysis goal \( g \)
- Available data \( X \)
- Utility measure \( U \)
- Data analysis method

**How**

- Analysis goal \( g \)
- Available data \( X \)
- Utility measure \( U \)
- Data analysis method

**Notes**

- How and What:
  - How: Analysis goal \( g \)
  - What: Available data \( X \), Utility measure \( U \), Data analysis method
InfoQ Assessment

Rating-based assessment
1-5 scale or desirability function on each dimension:

\[ \text{InfoQ Score} = \left[ d_1(Y_1) \ d_2(Y_2) \ ... \ d_8(Y_8) \right]^{1/8} \]

Experience from research methods courses:

- Preparing a PhD research proposal (FELU, Univ. of Ljubljana, Sant’Anna School of Advanced Studies, Pisa)
- Post-hoc evaluation of completed studies (CMU, goo.gl/erNPF)
\[ \text{InfoQ}(f,X,g) = U(f(X|g)) \]

**InfoQ Dimensions**

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

**How**

<table>
<thead>
<tr>
<th>#</th>
<th>Dimension</th>
<th>Note</th>
<th>Value</th>
<th>Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Data resolution</td>
<td>5</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Data structure</td>
<td>4</td>
<td>0.7500</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Data integration</td>
<td>5</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Temporal relevance</td>
<td>5</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Generalizability</td>
<td>3</td>
<td>0.5000</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Chronology of data and goal</td>
<td>5</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Concept operationalization</td>
<td>2</td>
<td>0.2500</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Communication</td>
<td>3</td>
<td>0.5000</td>
<td></td>
</tr>
</tbody>
</table>

**InfoQ Score = 68%**
InfoQ at the Study Design Stage

• Primary vs. Secondary Data
• Experimental vs. Observational Data
• Designed Experiments
• Computer Experiments
• Surveys: Pilot the questionnaire, plan initiatives to increase response rates

InfoQ at the Post-Data-Collection Stage

- Data Cleaning
- Preprocessing
- Reweighting
- Bias Adjustment
- Meta-Analysis
- Retrospective Experimental Design Analysis
- Censoring and Truncation
- **Surveys: Determine representativeness of returns and decide if responses should be weighted**

http://www.amstat.org/sections/srms/pamphlet.pdf
Chapter 13

Integrating InfoQ into Data Science/Analytics Programs, Research Methods Courses and More

Synopsis

The InfoQ components and dimensions presented in the previous chapters were applied to a wide range of domains such as education, healthcare, surveys and official statistics. In this chapter, the focus is on education programs in areas such as data science, business analytics or statistical methods. In this context, the focus is on practice oriented, practice based learning with InfoQ as an underlying, unifying theme. The chapter presents examples of integrating InfoQ in existing courses. It includes specific suggestions for designing or redesigning new courses and for designing a one day InfoQ workshop for analysts in industrial and service organizations.
The Information Quality Framework for Evaluating Data Science Programs

Shirley Coleman¹ and Ron S. Kenett²

¹ISRU, Newcastle University, UK, shirley.coleman@newcastle.ac.uk
²KPA, Israel and University of Turin, Italy, ron@kpa-group.com

Designing a new Analytics program requires not only identifying needed courses, but also tying the courses together into a cohesive curriculum with an overriding theme. Such a theme helps determine the proper sequencing of courses and create a coherent linkage between different courses often taught by faculty staff from different domains. It is common to see a program with some courses taught by computer science faculty, other courses taught by faculty and staff from the statistics department, and others from operations research, economics, information systems, marketing or other disciplines. Applying an overriding theme not only helps students organize their learning and course planning, but it also helps the teaching faculty in designing their materials and choosing terminology. The InfoQ framework introduced by Kenett and Shmueli provides a theme that focuses the attention of faculty and students on the important question of the value of data and its analysis with flexibility that accommodates a wide range of data analysis topics. In this chapter we review a number of programs focused on analytics and data science content from an InfoQ perspective. Our goal is to show, with examples, how the InfoQ dimensions are addressed in existing programs and help identify best practices for designing and improving such programs. We base our assessment on information derived from the program’s web site.
Challenges Ahead

• Bridge the gap between theory and applications

• Strengthen the position of statistical thinking in business, industrial, educational and academic applications

• Expand research on information quality (InfoQ)
Thank you for your attention