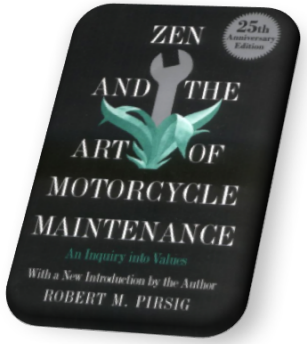
Theory and Practice of Data Cleaning

What is *Data Quality*? How can we improve it?

Data Quality (DQ) and Data Cleaning in Context

- **Data Cleaning/Wrangling**
  - A much needed, underappreciated **phase before data analysis can begin**
  - **Low-quality data** causes significant **costs** (whether we clean data or not)
  - Apply **early** in the **data life-cycle** (and apply often, as needed..)
- **Data Errors**
  - Many different types: quantitative (outliers) and **qualitative**:
    - **Syntax** / **format** errors (pattern violations)
    - **Semantic** / **schema** errors (integrity constraints)
- **Data Quality**
  - **… what is data quality?**
  - *"Even though quality cannot be defined, you know what it is."*



- Robert Pirsig

Data Quality Defined

- **Data Quality as Fitness for Use / Fitness for Purpose:**
  - Data are of **high-quality** if they are ***fit for use*** in their uses (by customers) in operations, decision-making, and planning.
  - They are fit for use when they are ***free of defects*** and possess the features needed to complete the operation, make the decision, or complete the plan.

Redman, Thomas C. ***Data Quality Management Past, Present, and Future: Towards a Management System for Data****. In Handbook of Data Quality, 15–40. Springer, 2013.*

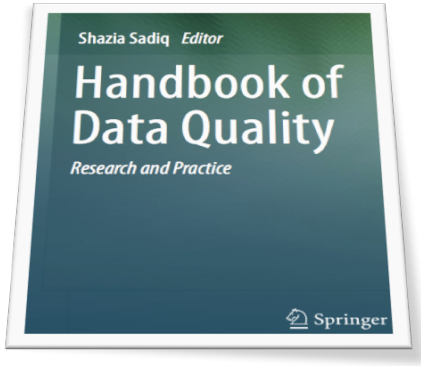
Data Quality: Fitness for Use

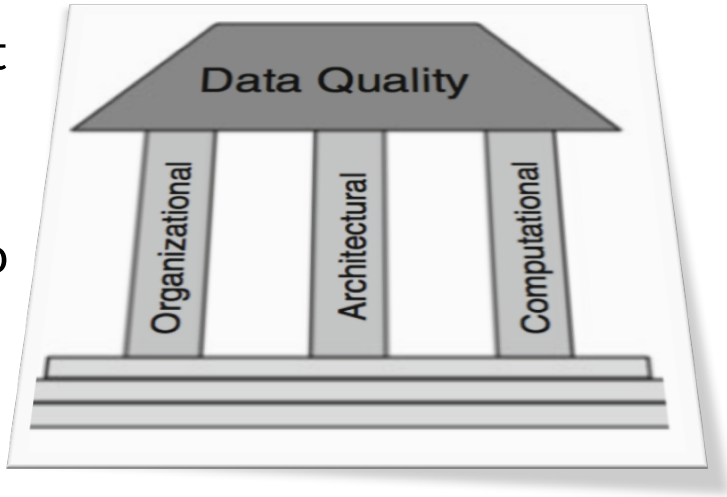
- **Where this comes from**:
  - What do you want to do with the data?
  - What are the ***questions*** you're trying to answer?
  - Do you even need this table / column / field?
    - e.g. analyzing census data per region, state, county, ..

- **Where this gets tricky:**
  - If you don't (yet) know what you want to ask of the data …
  - Interesting challenge for:
    - digital archivists (e.g., digital librarians) and research data librarians
    - data curators (e.g., at a natural history museum)

Pillars of Data Quality

- **Organizational**
  - **Data quality objectives** for the organization; **strategies** to establish roles, processes, policies, standards required to manage and ensure DQ objectives are met.
- **Architectural**
  - **Technology landscape** to deploy DQ management processes, standards, policies.
- **Computational**
  - **IT tools** and computational **techniques** required to meet DQ objectives
    - syntax and format *normalization*,
    - *integrity constraints*,
    - *provenance*,
    - *duplicate detection*, ...





Readings: [Sad13] Sadiq, S. ***Research and Practice in Data Quality Management****. Handbook of Data Quality, 2013.*

Common Phases, Steps in DQ Management

- **Context Reconstruction**
  - collect **context information** on organizational processes, services, data management procedures, quality issues, costs
  - (skip if context info is available from previous analyses)
- **Assessment/Measurement**
  - **measures** DQ along relevant **dimensions**
  - **assess** DQ by comparing with ***reference values***
  - enable ***diagnosis*** of DQ: find causes of poor DQ
- **Improvement**
  - concerns the selection of the ***steps***, ***strategies***, and ***techniques*** for reaching new data quality targets

*Readings: [BCF+09] Batini et al. **Methodologies for Data Quality Assessment and Improvement**. ACM Computing Surveys, 2009.*

Dimensions of Data Quality

Table 2 Dimensions of data quality

Conceptual view/associated metadata	**Data values**
Appropriate use	Accuracy
Areas covered	Completeness
Attribute granularity	Consistency
Clear definition	Timeliness
Comprehensiveness	
Essentialness	***Presentation quality***
Flexibility	Appropriateness
Homogeneity	Ease of interpretation
Identifiability	Formats
Naturalness	Format precision
Obtainability	Flexibility
Precision of domains	Handling of null values
Relevancy	Language
Robustness	Portability
Semantic consistency	Representation Consistency
Sources	Use of storage
Structural consistency	

- **DQ Dimensions**
  - **Accuracy**
  - **Completeness**
  - **Consistency**
  - **Timeliness**

Redman, Thomas C. ***Data Quality Management Past, Present, and Future: Towards a Management System for Data****. In Handbook of Data Quality, 15–40. Springer, 2013.*

Dimensions of Data Quality

- **Accuracy**
  - extent to which data are correct, reliable correspond to ground truth
  - often focus on syntax and patterns (e.g. regex matching for dates)
- **Completeness**
  - degree to which a dataset includes necessary information about relevant objects
- **Consistency**
  - *satisfaction* or *violation* of **schema** or **semantic rules**
  - in relational databases: **integrity constraints** (ICs), often in the form of ***denials***
- **Timeliness (also: Currency, Volatility)**
  - data change over time
  - answers questions such as:
    - what is the delay between change in the world and in the database?
    - how long data is valid in the real world?
    - is the data still appropriate?

*Readings: [BCF+09] Batini et al. **Methodologies for Data Quality Assessment and Improvement**. ACM Computing Surveys, 2009.*

Summary

- **Data Quality**: ***Fitness for Use***

- **Pillars** of Data Quality**:**
  - Organizational, Architectural, ***Computational***

- **Data Quality** Management **Phases**:
  - Context Reconstruction**,** ***Assessment*** ***(Measurement)*****,** ***Improvement***

- **Data Quality Dimensions**:
  - *Accuracy, Completeness, Consistency, Timeliness*