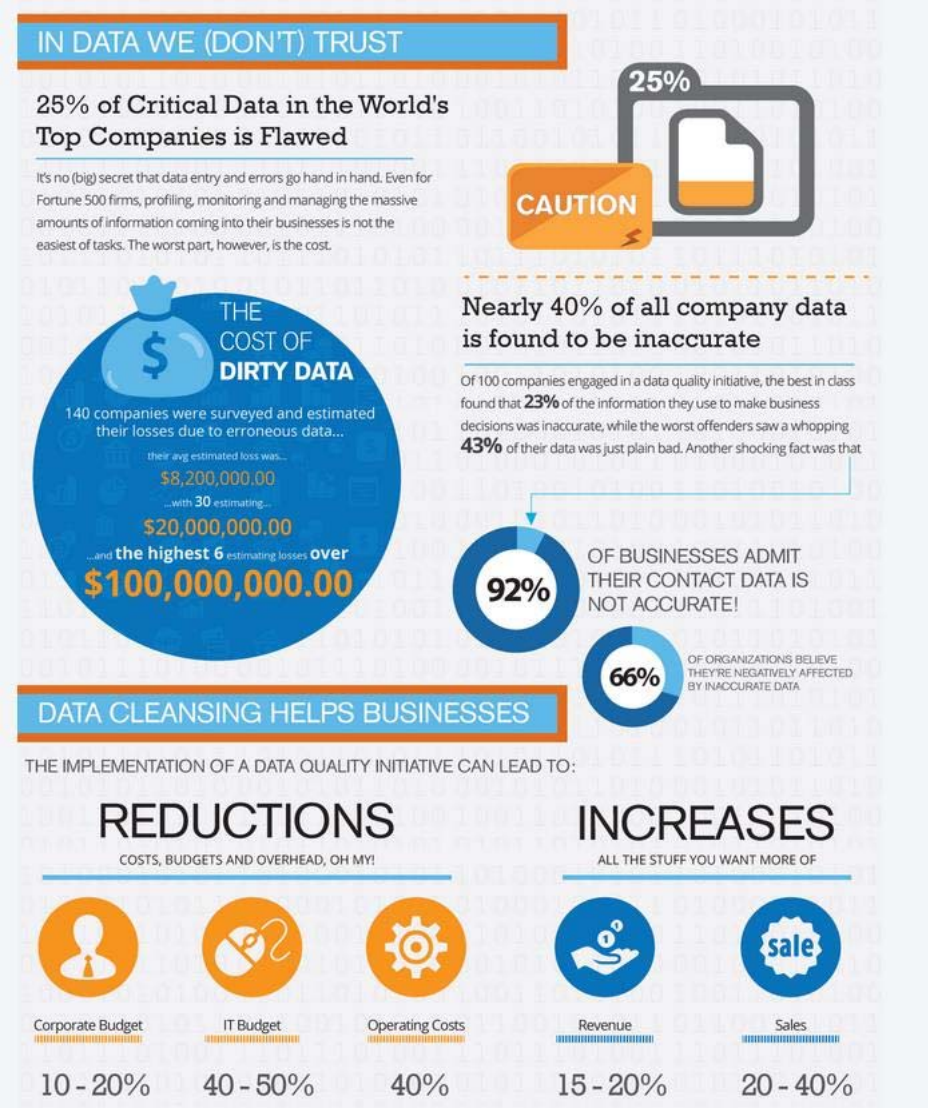
Data Quality

Knowledge objectives

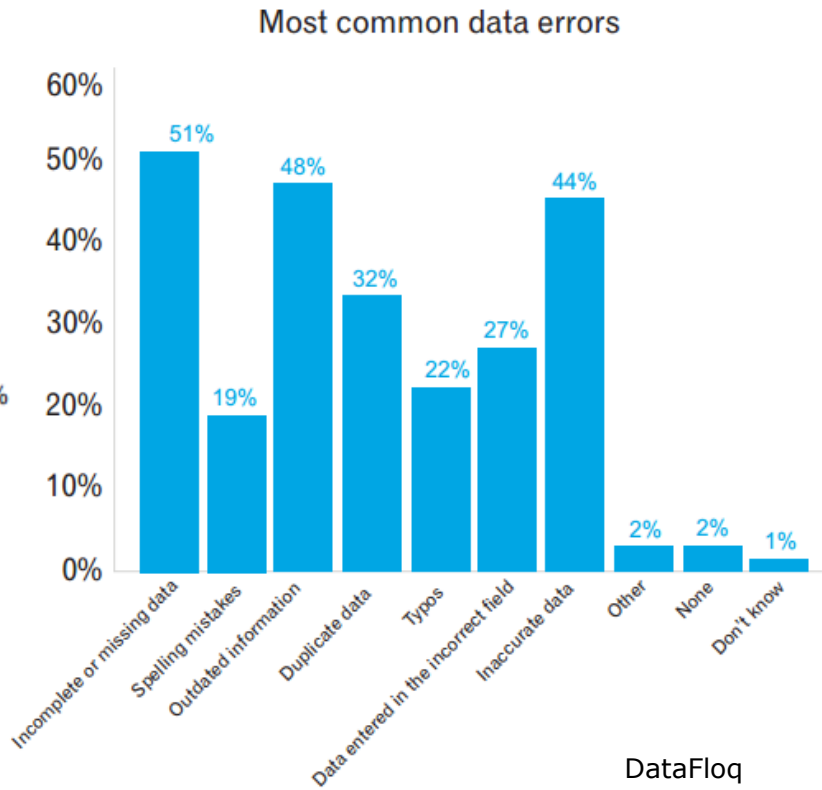
1. Explain what data quality is
2. Exemplify the causes of data quality problems
3. Classify the data conflicts depending on:
   a) They affect only the schema or also the instances
   b) They can happen in a single data source or need many
4. Calculate the value of the most prominent data quality measures (i.e., Completeness, Accuracy, Consistency, and Timeliness)

Motivation (I)



Halo Business Intelligence

Motivation (II)



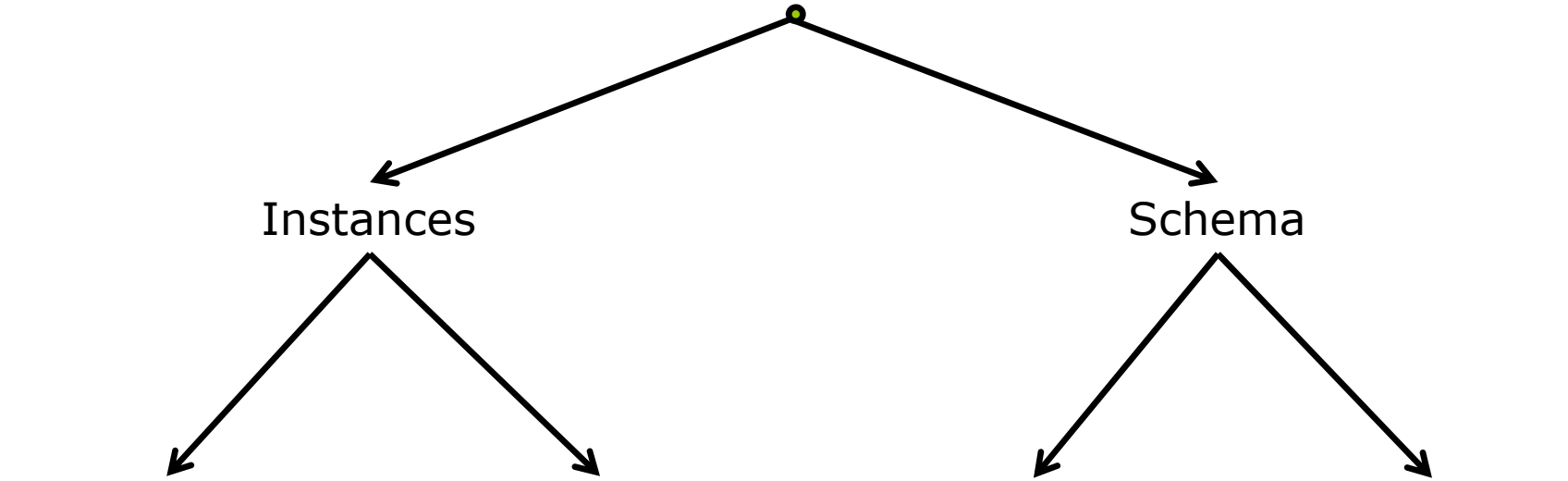
DataFloq

Fitness for use

"A user can only assess the level of quality of a set of data for a particular task to be executed in a **specific context**, according to a set of criteria, thus determining whether or not these data can be used **for that purpose**."

William Edwards Deming

Data conflicts



Single source

(Data entry errors)

- Misspelling
- Redundancy
- Contradictions
- …

Multi-source

(Overlapping, contradicting and inconsistent data)

- Duplicates
- Inconsistent aggregation
- Inconsistent timing
- …

Single source

(Lack of integrity constraints, poor schema design)

- Uniqueness
- Referential integrity
- …

Multi-source

(Heterogeneous data models and schema designs)

- Naming conflicts
- Structural conflicts
- …

Typical measures

- Completeness
- Accuracy
- Consistency
- Timeliness
- Relevance
- Response time
- Latency

Completeness

"The degree to which a given collection of data describes the corresponding set of real-world objects."

- Missing entities
- Missing values

$$Q_{Cm}(A_i) = |R(NotNull(A_i))|/|R|$$

$$Q_{Cm}(R) = |R(\bigwedge_{A_i \in R} NotNull(A_i))|/|R|$$

Accuracy

"The extent to which data are correct, reliable and certified error free."

- Free of typing errors
- Appropriate precision

$$e_A = |v_A - v_{RealWorld}|$$

$$Q_A(A_i) = |R(e_{A_i} \leq \varepsilon)|/|R|$$

$$Q_A(R) = |R(\bigwedge_{A_i \in R} e_{A_i} \leq \varepsilon)|/|R|$$

Consistency

"The degree of violation of semantic rules defined over a set of data items."

- Integrity constraints
  - Entity
  - Domain
  - Referential
  - User-defined
- Coincidence of copies
  - Temporal
  - Permanent

$$Q_{Cn}(R,B) = |R(\wedge_{rule \in B} rule(A_1,..,A_n))|/|R|$$

Timeliness (Freshness)

"How old the stored value of an attribute is with regard to the current value in the real world."

$$age(v) = now - TransactionTime$$

$f_u(v)$ = updates per time unit

$$Q_T(v) = (1 + f_u(v) \cdot age(v))^{-1}$$

$$Q_T(A_i) = Avg_{v \in R[Ai]} Q_T(v)$$

$$Q_T(R) = Avg_{Ai \in R} Q_T(A_i)$$

Summary

- Quality measures:
  - Accuracy
  - Completeness
  - Consistency
  - Timeliness

Bibliography

- Carlo Batini, et al. *Methodologies for data quality assessment and improvement*. ACM Compututing Surveys, 41(3), 2009
- Arkady Maydanchik. *Data Quality Assessment*. Technics Publications, 2007
- H. Garcia-Molina et al. *Database Systems*. Prentice Hall, 2009
- J. Bleiholder and F. Naumann. Data Fusion. ACM Computing Surveys, 41(1), 2008