Wetenschappelijk Onderzoek- en
Documentatiecentrum
Ministerie van Veiligheid en Justitie

datakwaliteit en de praktijk

Sunil Choenni

Content

- Data Quality and Dimensions
- Exploiting Domain Knowledge
- Obtaining and Implementing Domain Knowledge
- Conclusions

Data Quality and Dimensions

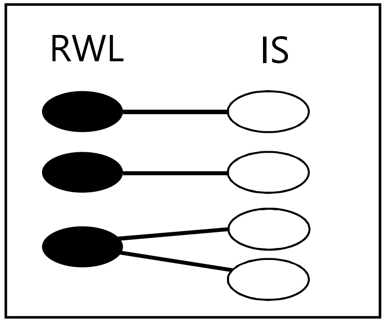
- Various definitions, ranging from defining some dimensions to more comprehensive definitions.
- Examples of the latter
  - Data should be a representation of (parts) of real-life
  - Fit for use
- To conclude: broad notion, subjective, and context/application dependent

Data Quality Dimensions

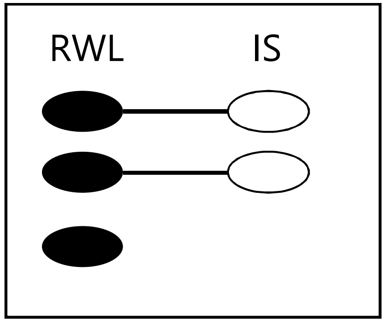
- Completeness
- Timeliness
- Accuracy
- Consistency
- Unambiguity
- Usability
- Relevance
- Presentation/Understandability
- ……..

Completeness

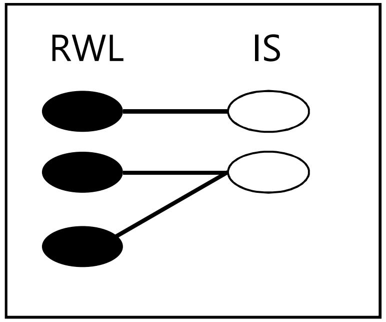
- Theoretical definition: represent every meaningful state of a real-world phenomenon



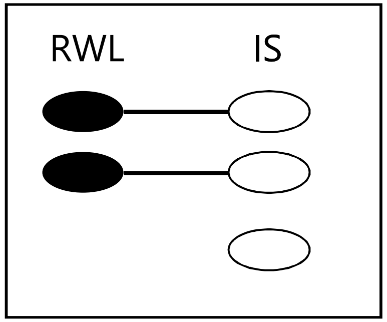
Correct



Incomplete



Ambiguous



Meaningless

Completeness

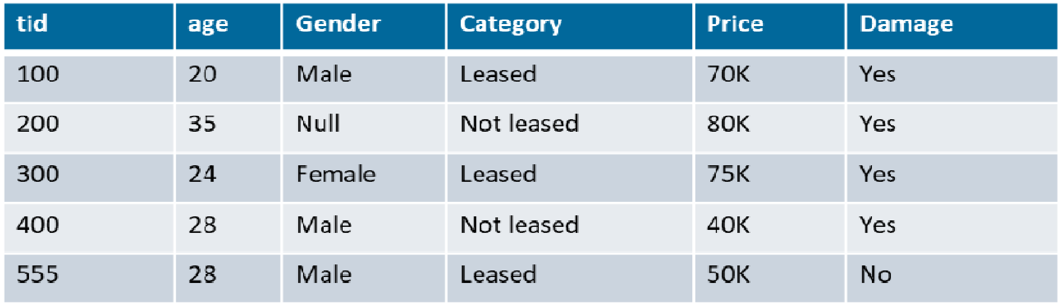
- Practical definition: percentage of values entered in data sources
- Null values! Value unknown, undefined, doesn’t exist, unknown whether it exists, NA, …
- Insight in null values is necessary to improve quality

Completeness

DBconstraint: #people = #male + #female;   ? #people

Optimizer has two choices:

- Count # tids (result: 5)
- Apply DBbconstraint (result: 4)
- Split along gender: 3 males and 1 female



tid	age	Gender	Category	Price	Damage
100	20	Male	Leased	70K	Yes
200	35	Null	Not leased	80K	Yes
300	24	Female	Leased	75K	Yes
400	28	Male	Not leased	40K	Yes
555	28	Male	Leased	50K	No

Timeliness

- No agreement wrt a definition. However it has to do with the velocity of processing updates.
- Two common indicators are
  - The delay between a change in a real world state and the resulting modification in the IS
  - Volatility (time period for which information is valid in the real world)

Timeliness

- Real world evolves over time
- Focus on obtaining resemblances between datasets and real world phenomena

Globe, World, Earth, Planet by Qimono on Pixabay| CC0 Public Domain

Semantic Level: Example

- Stored birthplace of an offender is USSR
- Today, USSR does not pertain a real-place
- DQ(country) in the past was fine but today poor → DQ degradation

The Soviet Union and it's satelite states and allies.png by Keeby101 on Wikimedia Commons | CC Attribution-Share Alike 3.0 Unported

Accuracy and Consistency

- The extent to which data are correct and reliable. Proximity of a value v (John) to another value v’ (Juhn)
- Violation of integrity rules
  - Marital status = married --→ age > 16
  - integrity constraints

Supplier

S#	Sname	City
S20	Fashion_Fox	Almere
S26	Cyber_Shop	Breukelen
S35	Orcam	Enschede

Part

P#	Pname	Price	Stock
P4	CHAIR	70	4000
P10	DRESS	120	100
P12	TABLE	50	1000
P15	LAMP	70	450

Deliver

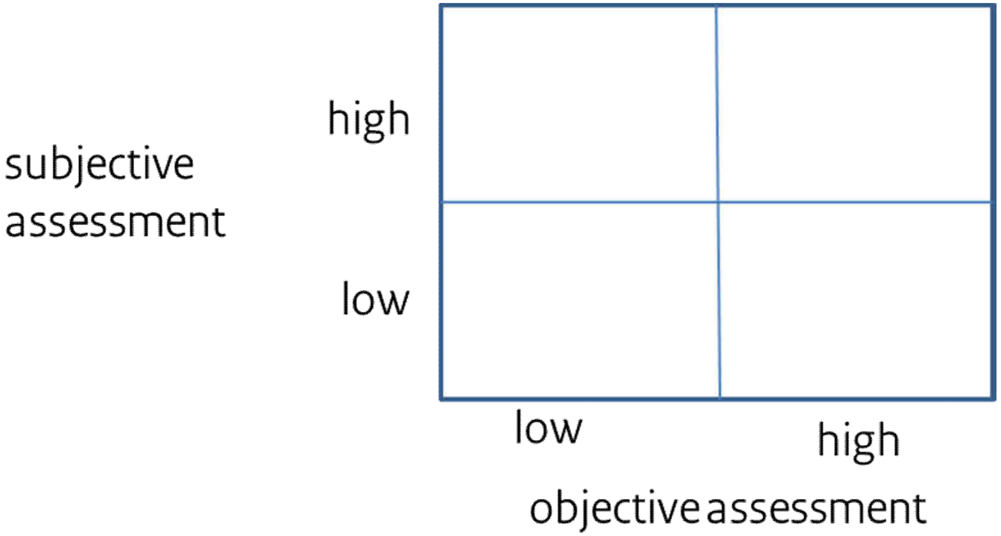
S#	P#
S20	P10
S26	P10
S26	P12
S35	P4
S35	P10
S35	P12

~~Insert (S30, P4)~~

~~Delete (S35, Orcam, Enschede)~~

Dimensions

- Determine relevant and viable dimensions in a domain
- Distinguish between “objective” and “subjective” measurable dimensions



Levels of Data Quality (DQ)

- Syntactic
- Semantic
- Pragmatic

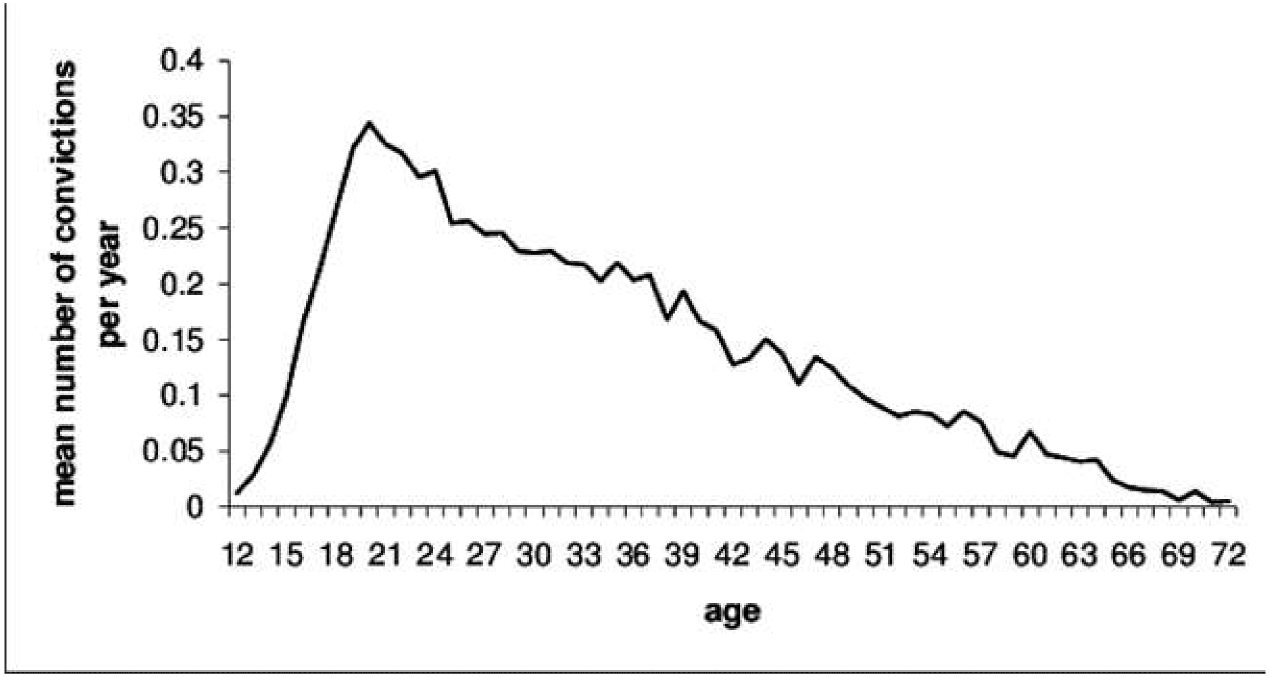
Businesswoman, Presentation, Suit by Geralt on Pixabay | CC0 Public Domain

Syntactic

- Degree to which stored data meets specified metadata
- Metadata: $12 \leq age \leq 99$
- If 90 out of 100 people meet the age constraint in our database then DQ(age) = 90%

Syntactic: Domain Knowledge

- Domain Knowledge: Age-Crime curve



Semantic Level

- Degree to which stored data corresponds to represented external phenomena

- Several dimensions to assess Data Quality (DQ) at the semantic level

- DQ is determined by the extent to which stored data adheres to these dimensions

Evolving Semantics: DQ degradation

- To prevent DQ degradation data evolvement and semantic changes should be handled adequately
- In practice, data evolvement may lead to unjustified trend reversals
  - reorganizations of municipals
  - rules and regulations are changing over time
  - …

Exploiting Dependencies

- Quantitative dependencies: Criminal Justice system
- Chain of police-prosecution-courts-execution

Illustration by Annemarie van der Linde

Exploiting Dependencies

- Qualitative dependencies
- Study dramatical changes
- Cannot be automated fully

Course, Graphic, Upward, Bar, Top by Peggy_Marco on Pixabay| CC0 Public Domain

Redundancy

- Different databases may store same kind of data → may cause inconsistencies

- Scrutinize overlapping data sets and search for inconsistencies

- Present inconsistencies to domain experts to select the most plausible value of an attribute

Semantic Level: Groups

- On the basis of domain knowledge we may define groups of rules to improve data quality. Some of these groups are
    - Rules to manage redundancy
    - Rules to deal with missing data
    - Rules to handle semantic changes in attributes
    - Rules to exploit dependencies
    - Rules to filter out results that should not be shown to the user.
    - Rules to determine whether large deviations exist between past and future data or between values from the same or different databases.

Domain knowledge

- How to obtain domain knowledge?
  - knowledge elicitation techniques
  - data mining technology and statistics
- How to implement domain knowledge?
  - knowledge representation techniques
  - form groups of rules
  - IF THEN ELSE formalism/state transition diagrams

Elicitation

- Task of knowledge engineers
  - Protocol analysis: experts are asked to solve a case in front of knowledge engineer
  - Interviews
  - literature

Example IF THEN ELSE

- Quantitative dependencies: Criminal Justice system

- Correlations between attributes

```
IF (date_comitted_crime = “unknown”) THEN
    IF (reported_crime_date ≠ “unknown”)
            THEN date_comitted_crime := reported_crime_date
            ELSE generate_alert().
```

Summary

- Data Quality is a broad notion which is subjective, time and context dependent

- Domain knowledge helps to enforce and to improve data quality

- Application of domain knowledge for data quality purposes requires structuring and organization of the knowledge in some formal system