# Data Quality

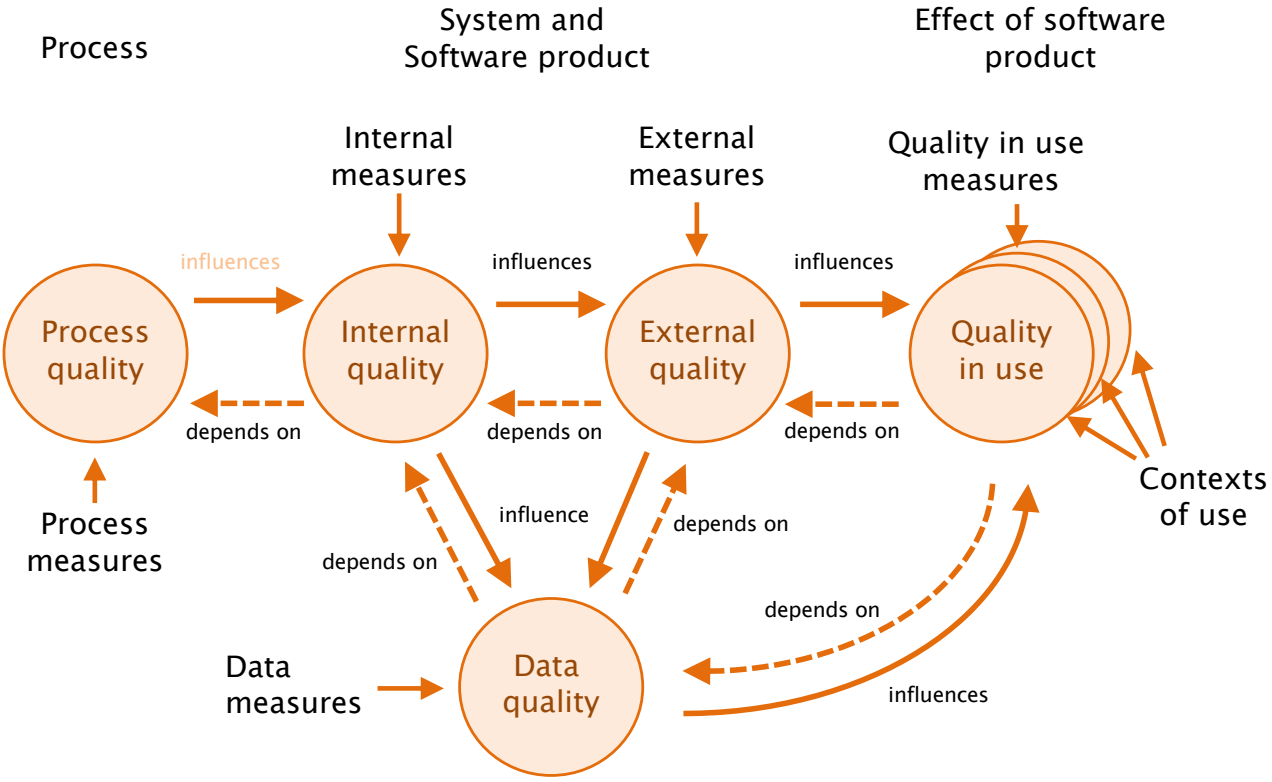
## Data Management and Visualization

SOftEng
http://softeng.polito.it



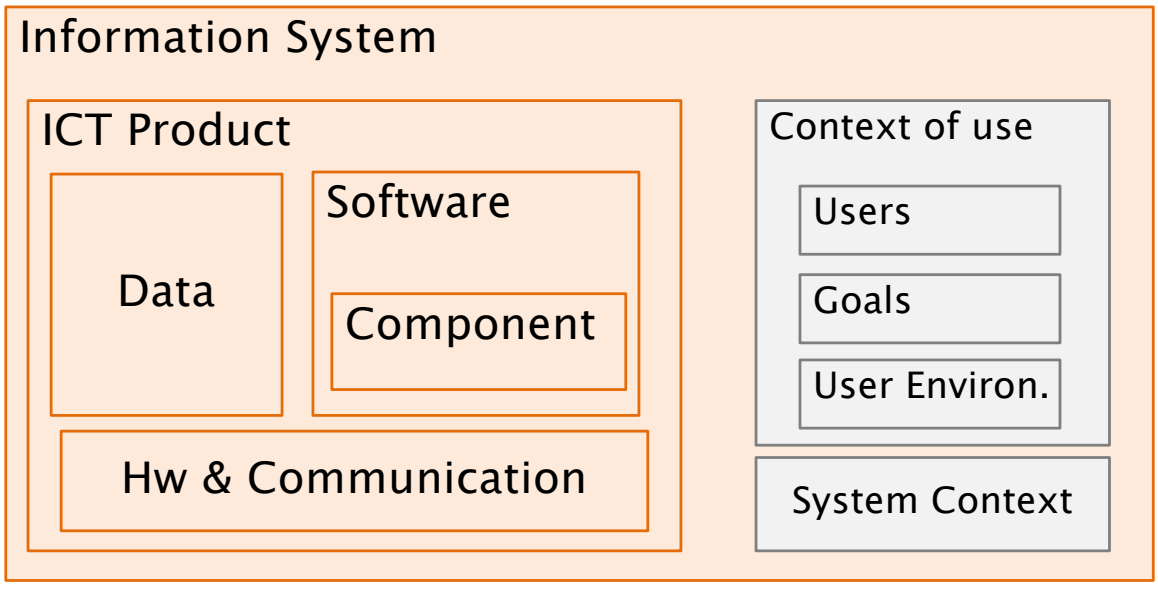
# Licensing Note

# Software Qualities

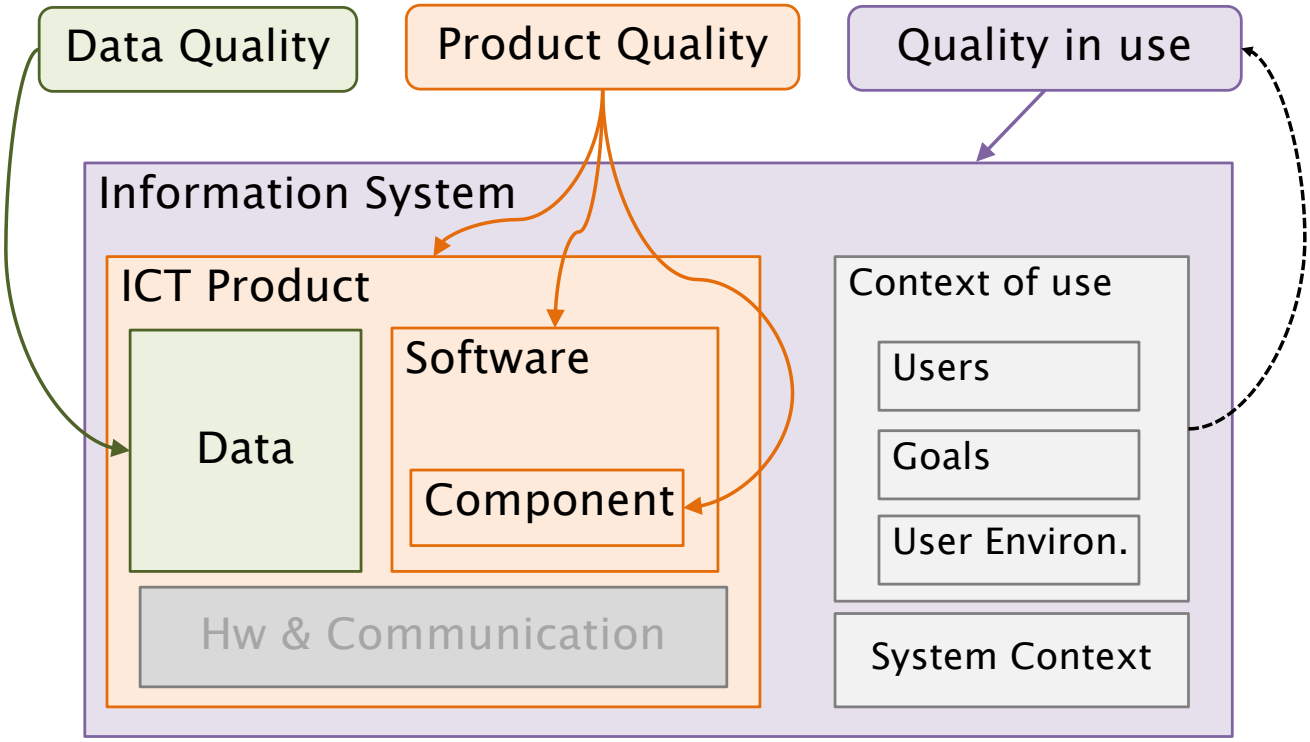



Adapted from ISO/IEC 25020

# Target entities



# Target entities vs. Q. Models

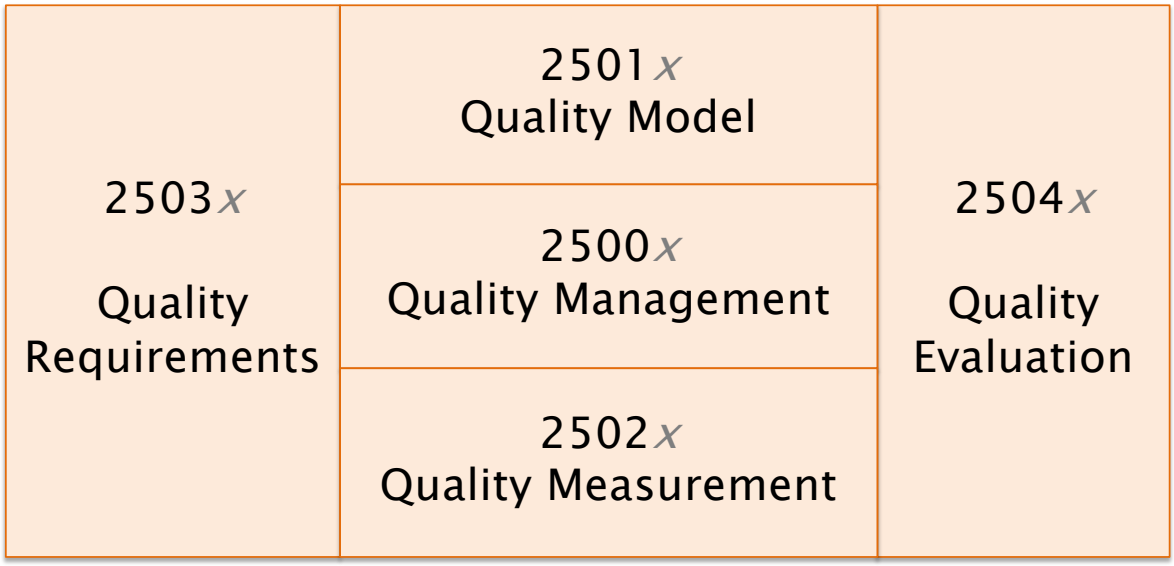





# Software Product Quality

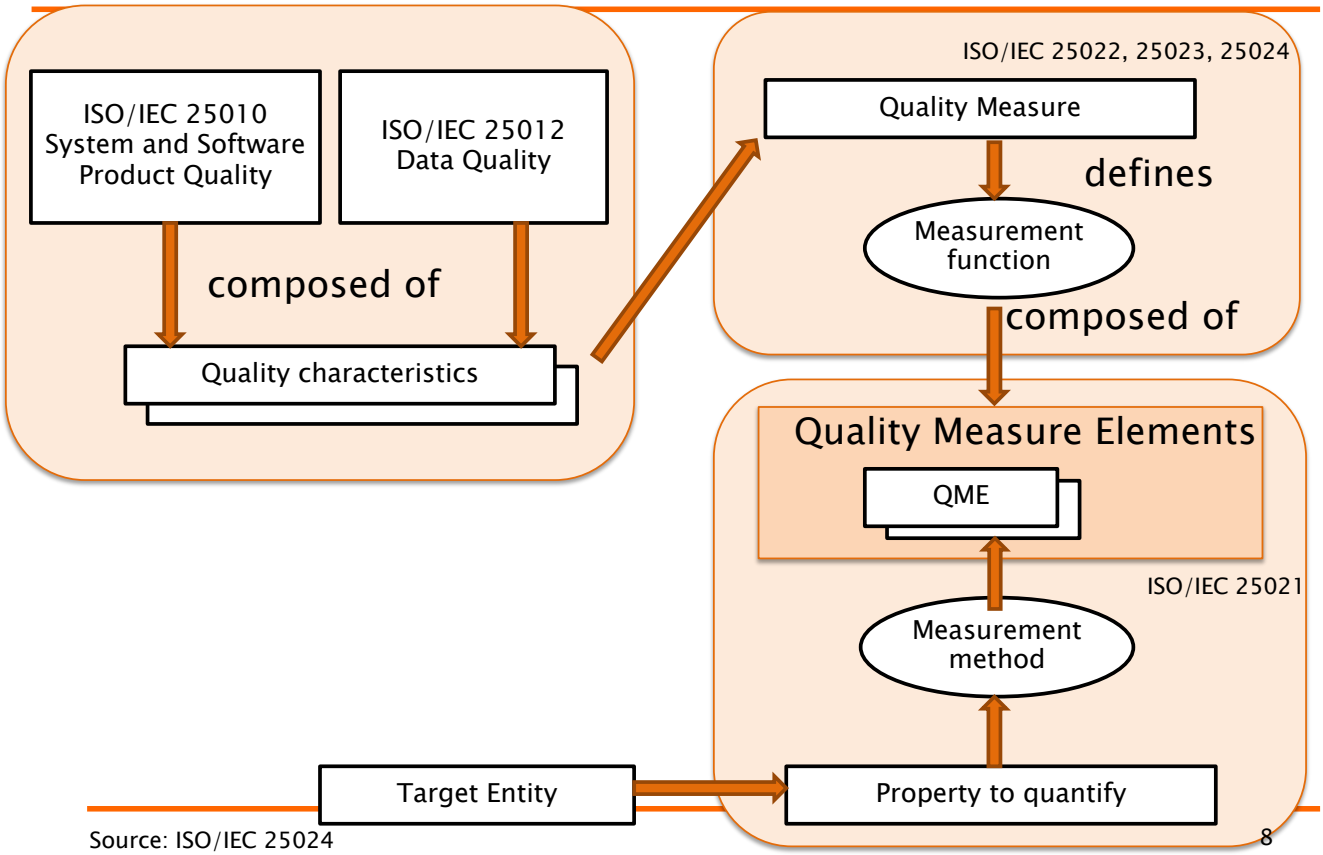
- ISO/IEC 9126: Issued 1991, revised 2001
  - Being retired
- ISO/IEC 250xx – **SQuaRE**
  - Software product Quality Requirements and Evaluation
  - Family of standards
    - in development

# ISO SQuaRE – Standard Family



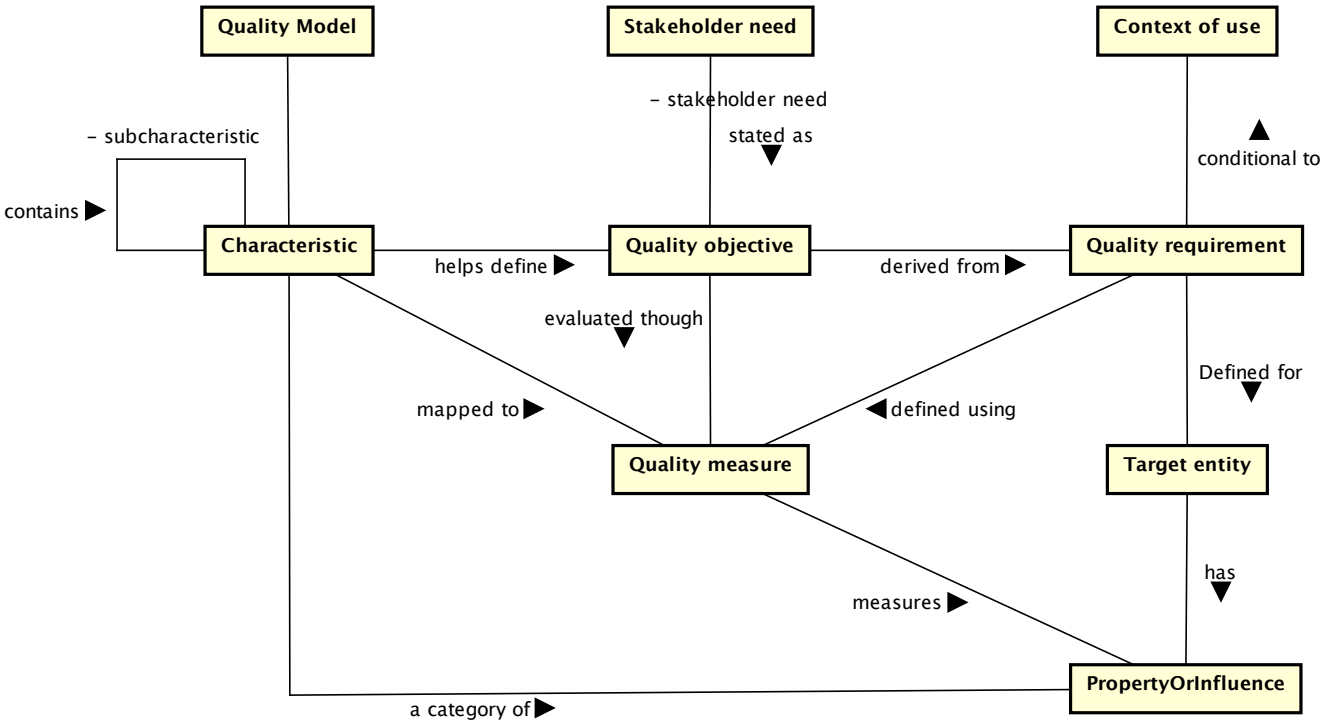



# Relationships among standards




Source: ISO/IEC 25024

# Quality conceptual model




Adapted from ISO/IEC 25010-1



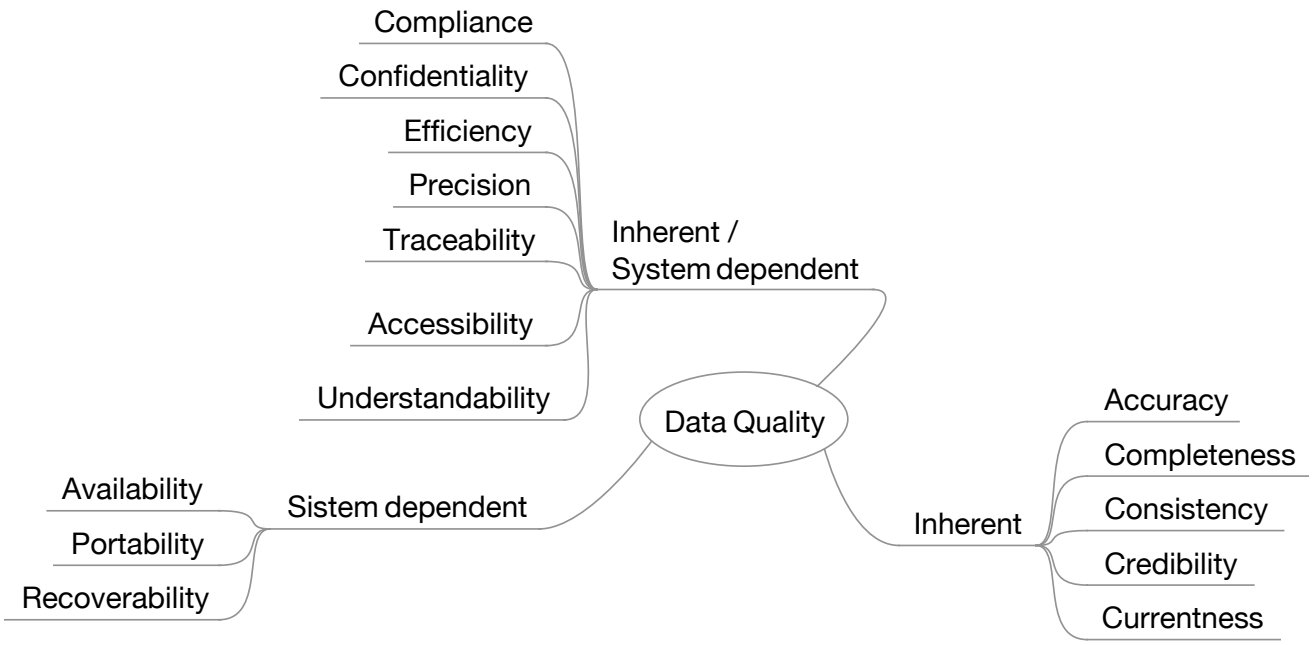
# Model structure

- Characteristic
  - Main aspects, e.g., usability
- Sub-Characteristic
  - Specific aspects, e.g. accessibility
- Measure
  - Measurement function to evaluate a specific (sub)-characteristic
- Measure element
  - Fundamental

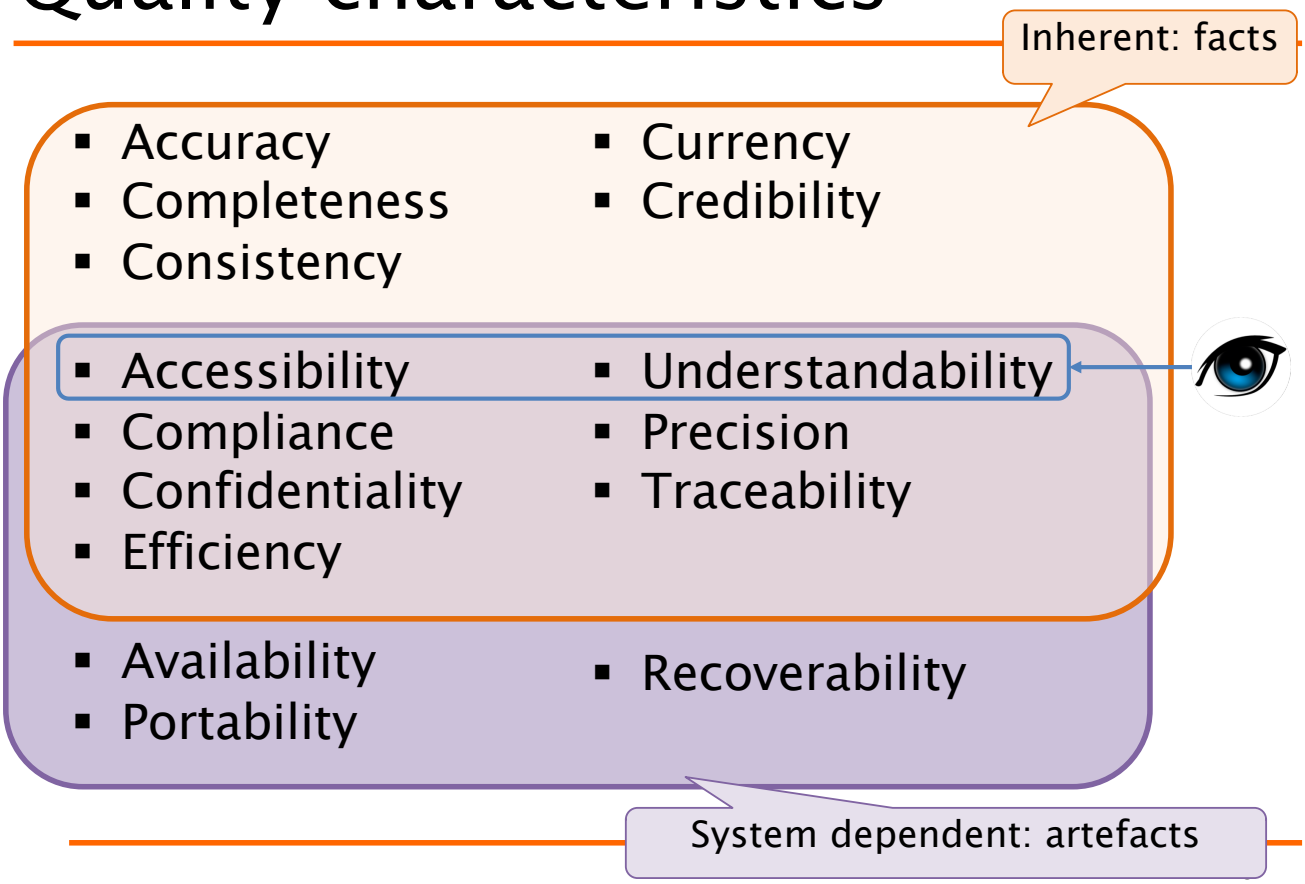
# DATA QUALITY



# Data Quality Model



# Quality characteristics




# Quality characteristics

- Accuracy
- Completeness
- Consistency
- Currency
- Credibility

- Accessibility
- Compliance
- Confidentiality
- Efficiency
- Understandability
- Precision
- Traceability

- Availability
- Portability
- Recoverability

# Accuracy

- Correspondence between data and reality
  - Syntactic
    - It belongs to a set of validated information
  - Semantic
    - The meaning (the content) corresponds to the reality



# Open or Closed Wordl?

- Closed World (CWA):
  - The knowledge represented in the data (and its schema) is complete
  - E.g., if a code appears in the list of valid codes it is correct, otherwise it is wrong
- Open World (OWA):
  - The knowledge represented in the data is (knowingly) incomplete
  - E.g., if a code appears in the list of valid codes it is correct, otherwise it is not possible to tell for sure

# CWA – Accuracy : Genomics

- Human genes are known and coded, each has a predefined symbol
- Any code not included in those predefined represents a syntactic accuracy error
- E.g. code '**SEPT2**'(Septin-2) when imported into  is automatically turned into 'February 2'



# OWA – Accuracy

How to decide what is accurate?

- Rules that define what is syntactically correct
  - E.g. regular expressions
- Constraints to define what values are semantically acceptable
  - E.g. validity interval

# Where do rules come from?

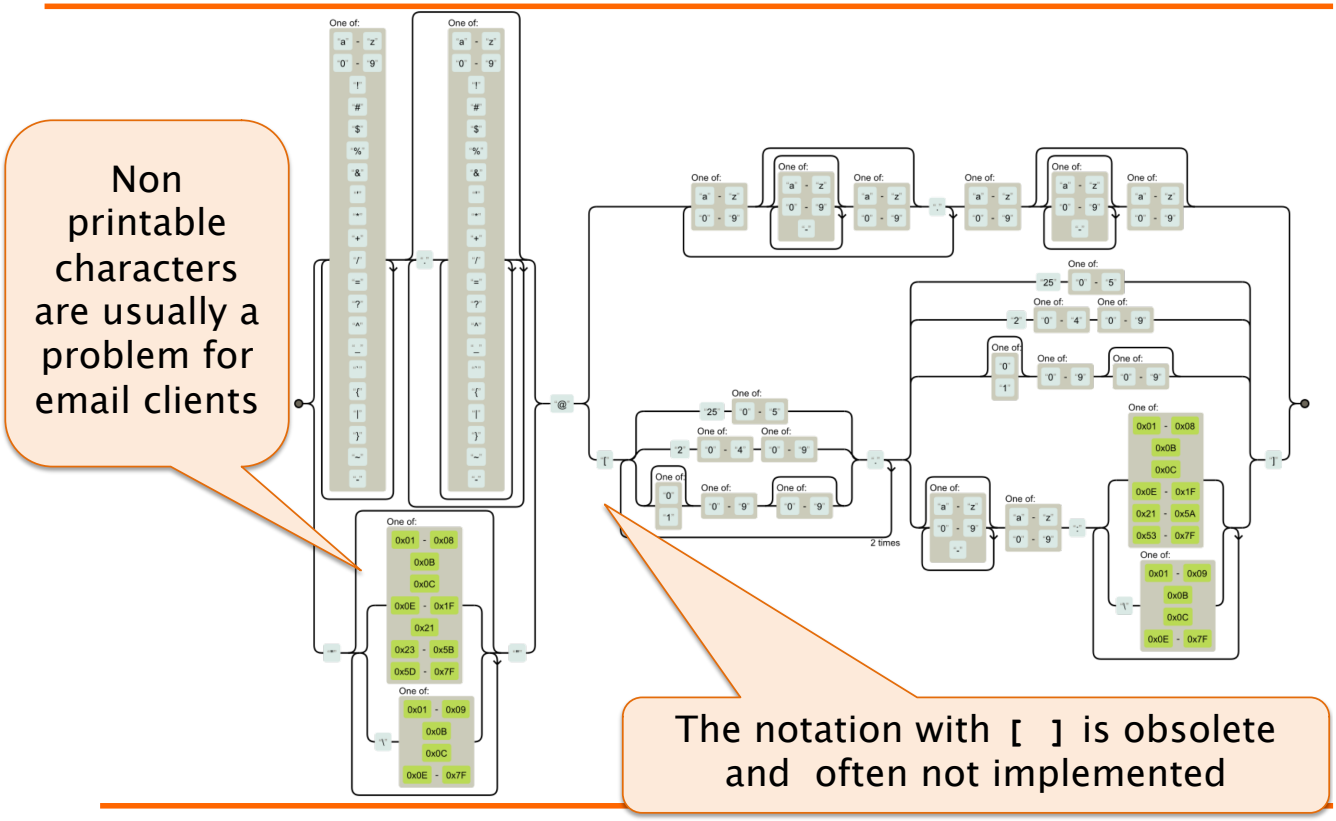
- Standard
- Domain knowledge
- Similar data
- Past data



# OWA: Email per RFC-5322

```
\A(?:[a-z0-9!#$%&'*+/=?^_`{|}~-]+(?:\.[a-z0-
9!#$%&'*+/=?^_`{|}~-]+)*
 |  "(?:[\x01-\x08\x0b\x0c\x0e-\x1f\x21\x23-\x5b\x5d-
\x7f]
      |  \\[\x01-\x09\x0b\x0c\x0e-\x7f])*")
@ (?:(?:[a-z0-9](?:[a-z0-9-]*[a-z0-9])?\.)+[a-z0-
9](?:[a-z0-9-]*[a-z0-9])?
  |  \[(?:(?:25[0-5]|2[0-4][0-9]|[01]?[0-9][0-
9]?)\.){3}
 (?:25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?|[a-z0-9-]*[a-
z0-9]:
           (?:[\x01-\x08\x0b\x0c\x0e-\x1f\x21-\x5a\x53-
\x7f]
           |  \\[\x01-\x09\x0b\x0c\x0e-\x7f])+)
     \])\z
```

# OWA: Email per RFC-5322





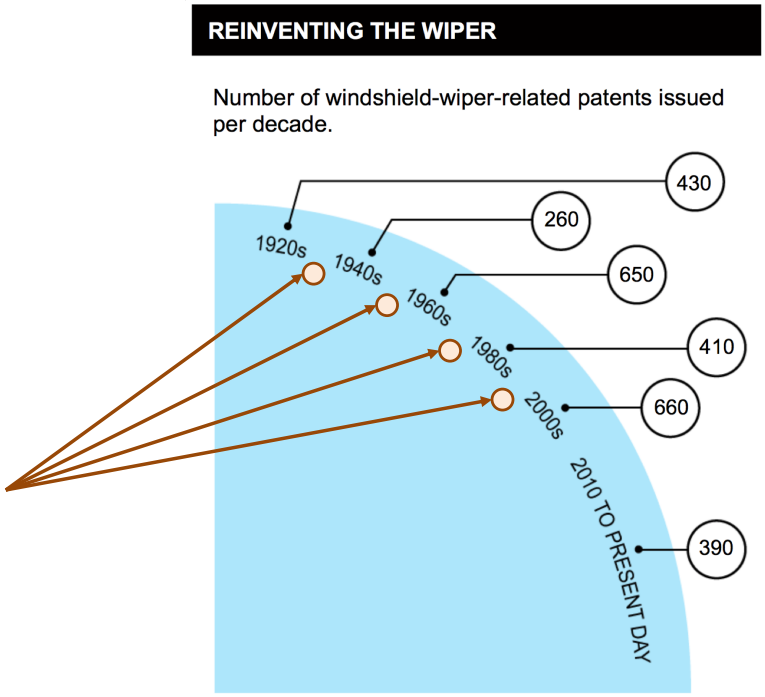

# Completeness

- Computer: presence of all necessary values
  - Both to entity occurrences and to attributes of a single occurrence
  - Note: not all missing values constitute a completeness issue
- User: how much the available data is capable of satisfying the needs

# Completeness

What about 1930s, 1950s, 1970s, 1990s ?

A possible hypothesis, another one considered later




Source: http://www.nytimes.com/2014/09/14/magazine/who-made-that-windshield-wiper.html?_r=0



# Consistency

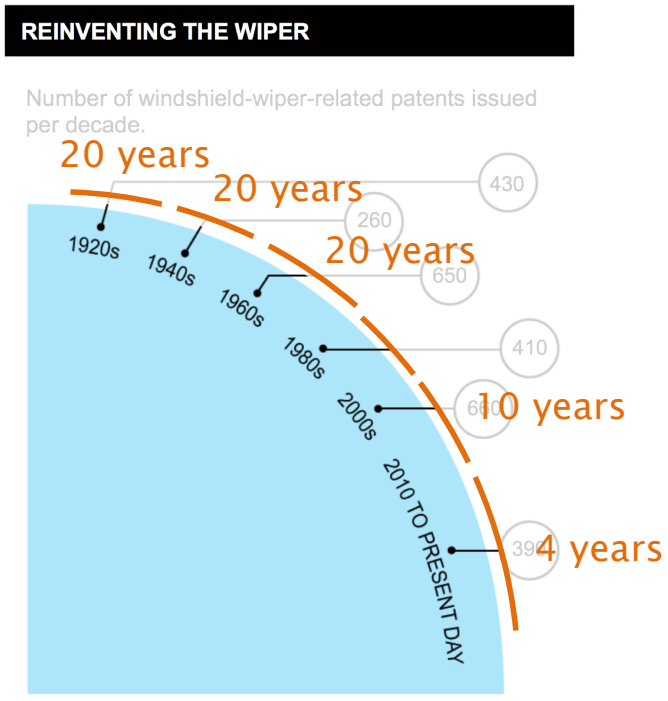
- Absence of contradictions in the data
  - Referential integrity
    - Often guaranteed in RDBMS
  - Duplication
    - Increase the risk of inconsistency on update
  - Semantic
    - E.g. birth date must be before death date

# Consistency in graph data

- Values in a series of data encoded with visual attributes must be comparable
  - Consistent aggregation level
  - Consistent measurement method
  - Consistent target entities



# Aggregation level




Count on of events on periods of different length are not comparable

A possible hypothesis, another one considered earlier

Source: http://www.nytimes.com/2014/09/14/magazine/who-made-that-windshield-wiper.html?_r=0

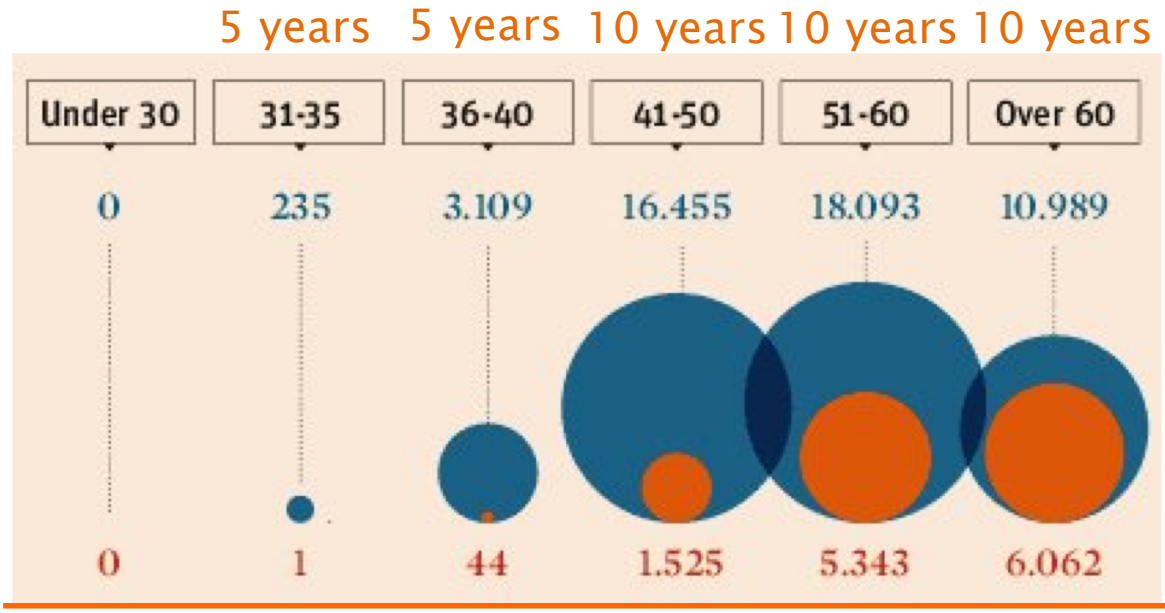
# Aggregation level

| Period | Duration [years] | Patents | Pat. per year |
|---|---|---|---|
| 1920s | 20 | 430 | 21.5 |
| 1940s | 20 | 260 | 13.0 |
| 1960s | 20 | 650 | 32.5 |
| 1980s | 20 | 410 | 20.5 |
| 2000s | 10 | 660 | 66.0 |
| 2010 to present | 4 | 390 | 97.5 |

When comparing values corresponding to entities or categories with different *size*, normalized values (i.e. densities) are comparable, absolute values are not!



# Aggregation level




Source: Corriere della Sera, 09 Settembre 2017

# Aggregation level

| Range | Size | Count | Density |
|---|---|---|---|
| 31-35 | 5 | 235 | 47.0 |
| 36-4 | 5 | 3109 | 621.8 |
| 41-50 | 10 | 16455 | 1645.5 |
| 51-60 | 10 | 18093 | 1809.3 |
| Over 60 | 10 | 10989 | 1098.9 |
| | Ratios: | 5.3 | 2.6 |

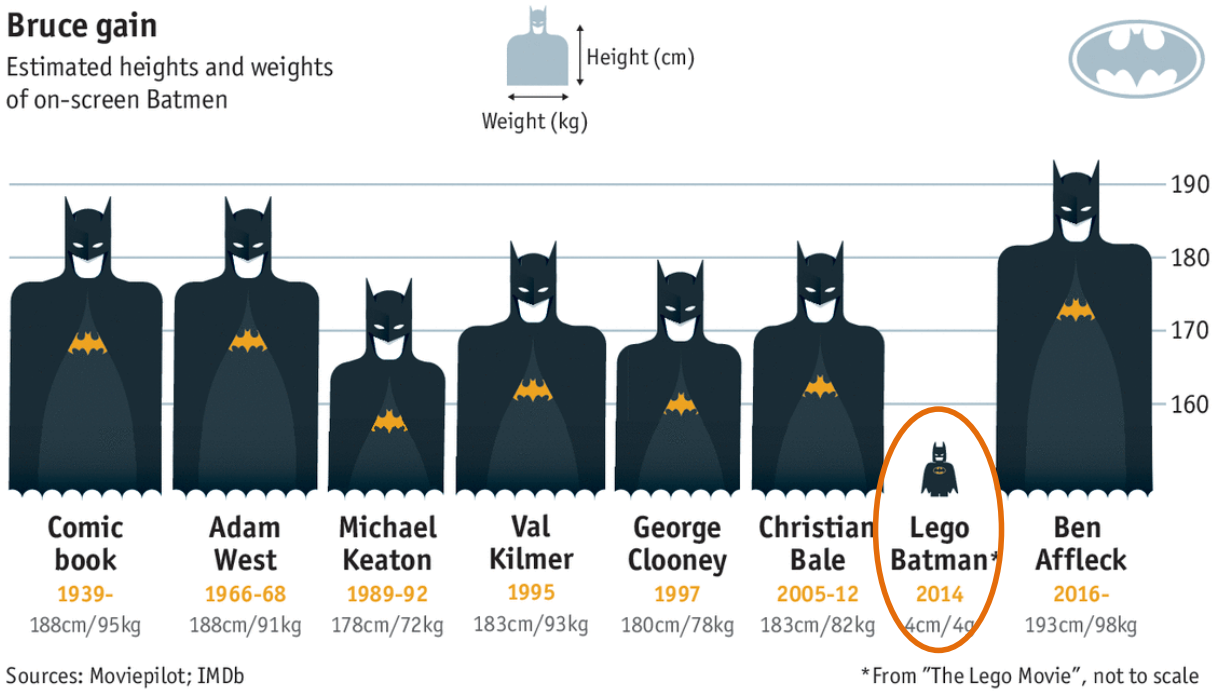
Lie factor = 2



# Consistent method

- A series of values that are not measured using the same method might not be directly comparable
  - estimate vs. actual, projection vs. final
  - periodic samples collected at different possibly non equivalent times
    - e.g. different period of year, week, day

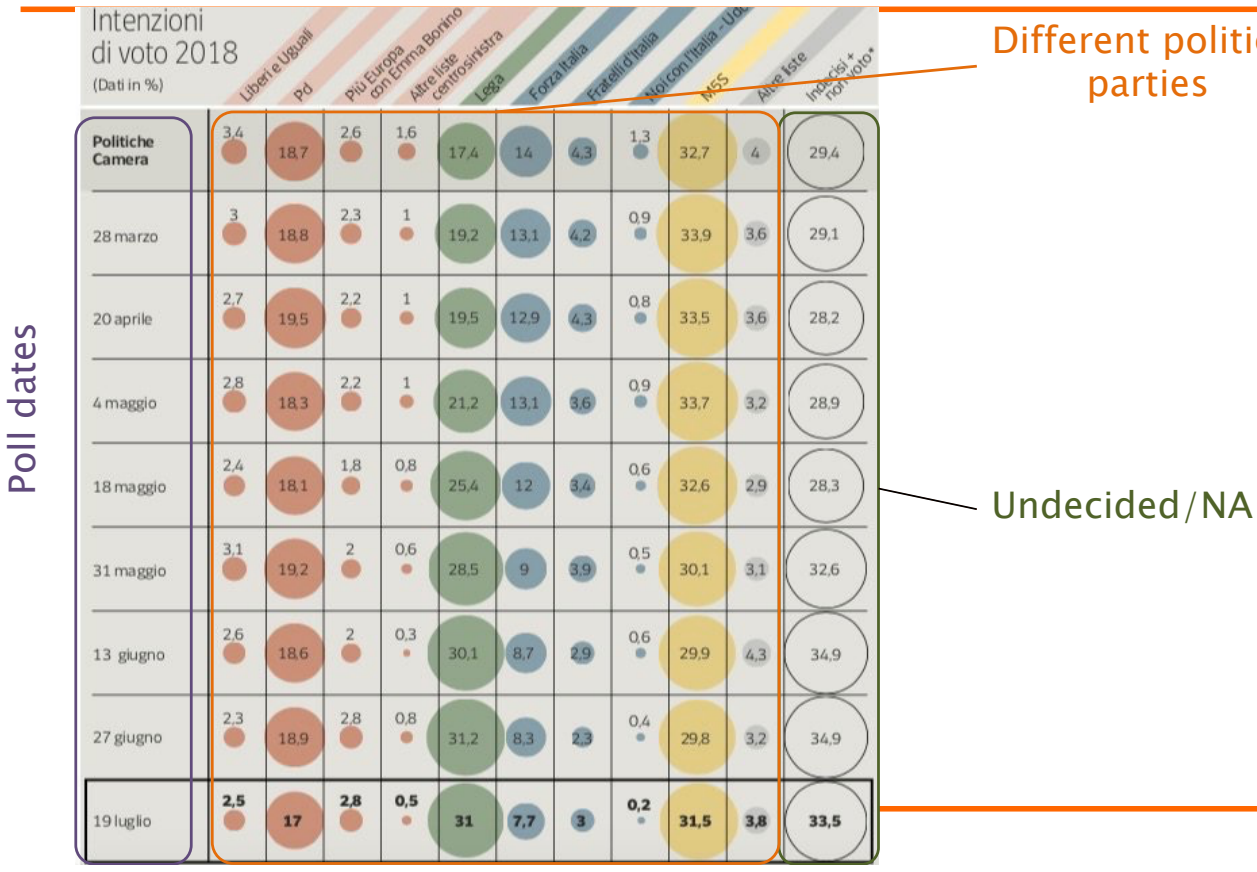
# Consistent target entities




Economist.com



# Consistent target



# Consistent target

- Proportions computed on different reference wholes

$$Undecided = \frac{n_{undec} + n_{NA}}{N_{sample}}$$

$$P_i = \frac{n_p i}{N_{sample} - n_{undec} - n_{NA}}$$



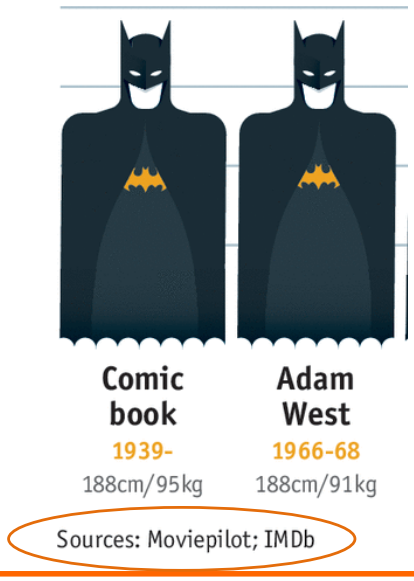
# Currency

- Currency is the extent to which data is up-to-date
  - With reference to the reality and
  - With reference to the task at hand
- Lack of information to establish currency is an Understandability issue

# Credibility

- The extent to which data are regarded as true and credible by users

- What is the source of the data showed in the graph?



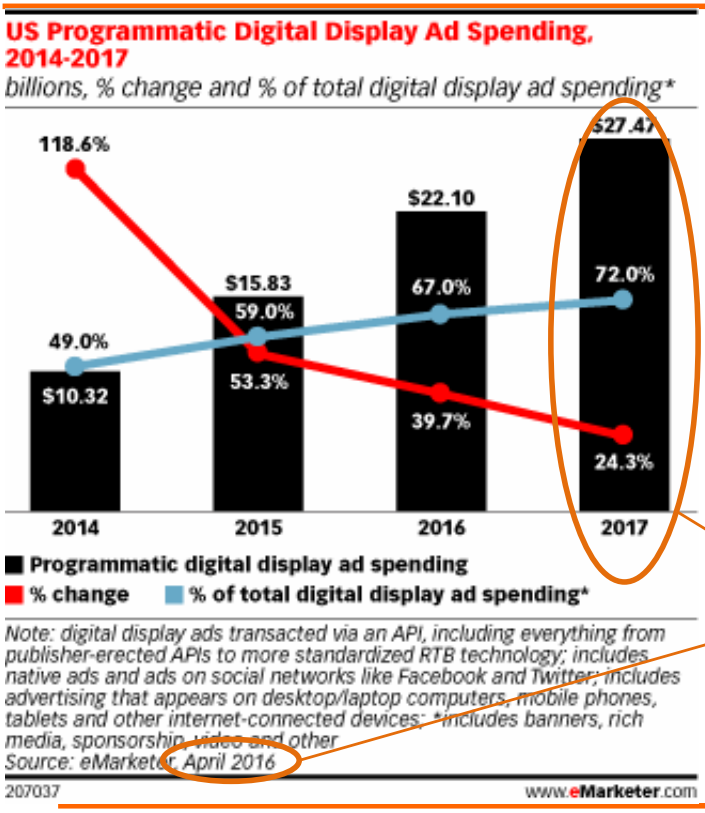



# Understandability

- The extent to which data can be read and interpreted by users

- How is data measured? Is there a track of how values are collected, measured or estimated?
  - If multiple multiple methods are used that might represent an inconsistency issue.

# Understandability

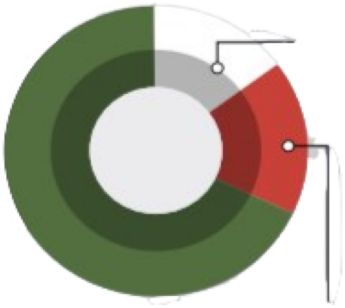



Data from 2016 including values for 2017. Undeclared mix of projections and final data.

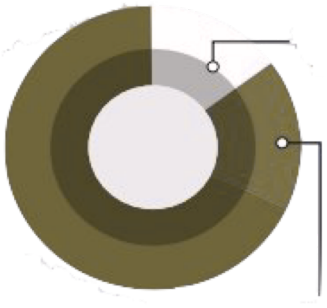


# Accessibility

- The capability of data to be accessed, particularly by people who need supporting technology or special configuration because of some disability



Original



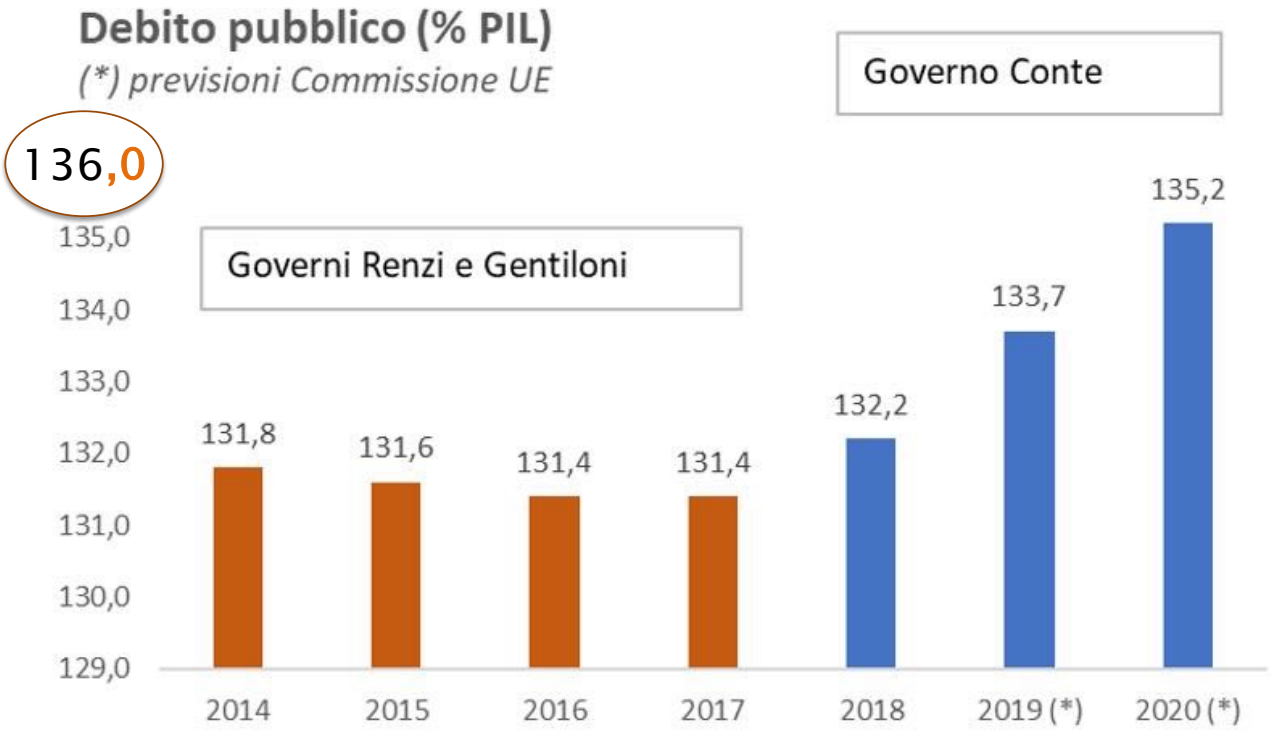
Color-blind simulation

# Precision

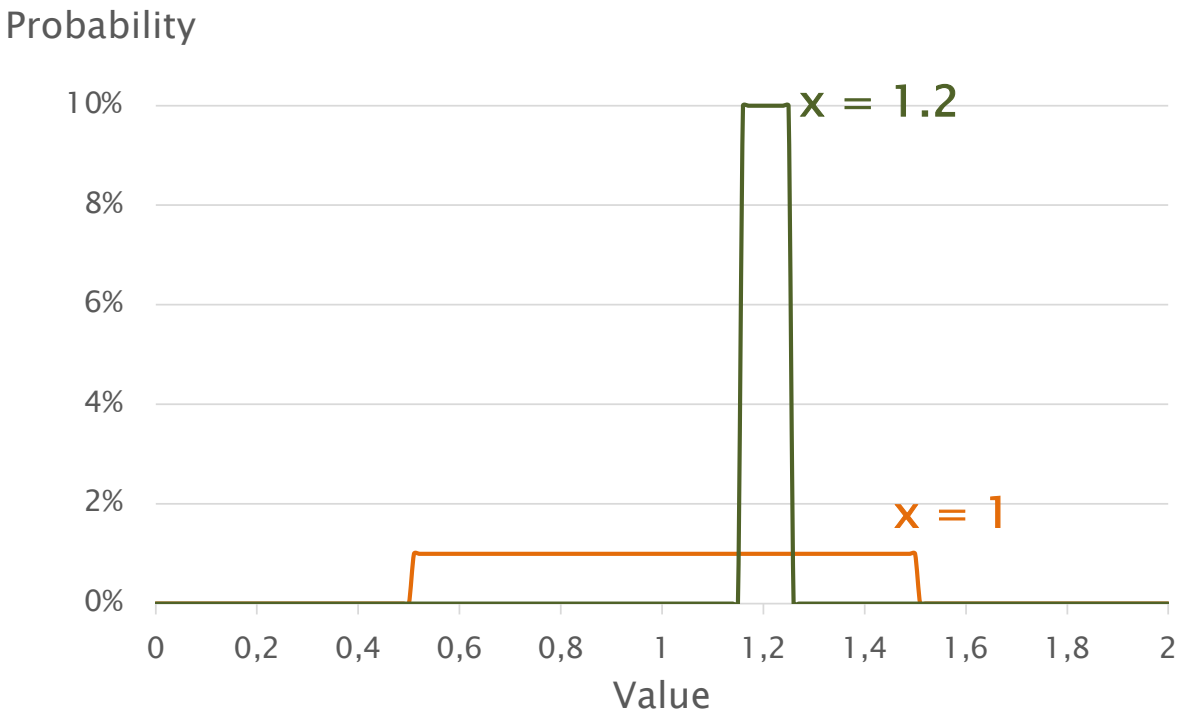
- The capability to provide the degree of information needed in a stated context of use
  - Enough information to allow discriminate
  - Not too much to overload reader
    - Related to "Utility"



# Precision



# Precision






# References

- ISO/IEC 25010 – System and software quality models
- ISO/IEC 23012 – Data Quality model
- ISO/IEC 25024 – Measurement of data quality