



Data Quality

Marco Torchiano

Version 1.0.0 - May 2020

License

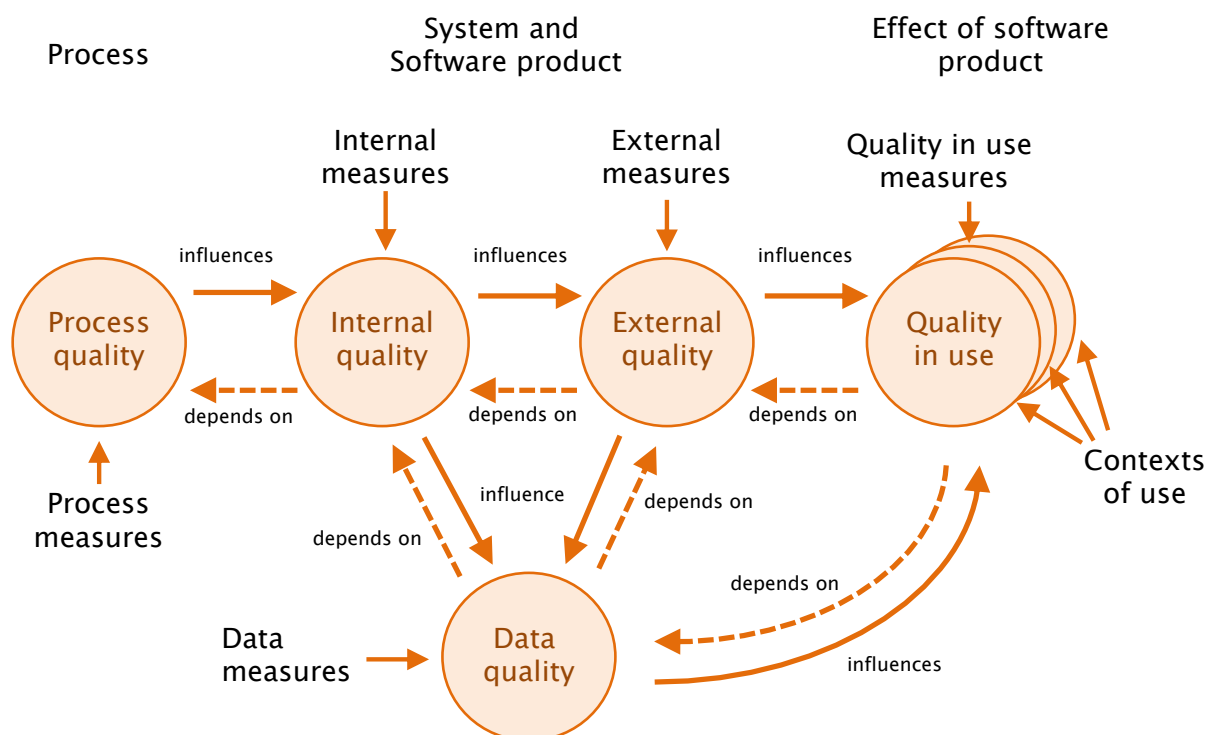
This work is licensed under a [Creative Commons Attribution-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-sa/4.0/).

- You are free to:
 - **Share** - copy and redistribute the material in any medium or format
 - **Adapt** - remix, transform, and build upon the materialfor any purpose, even commercially.

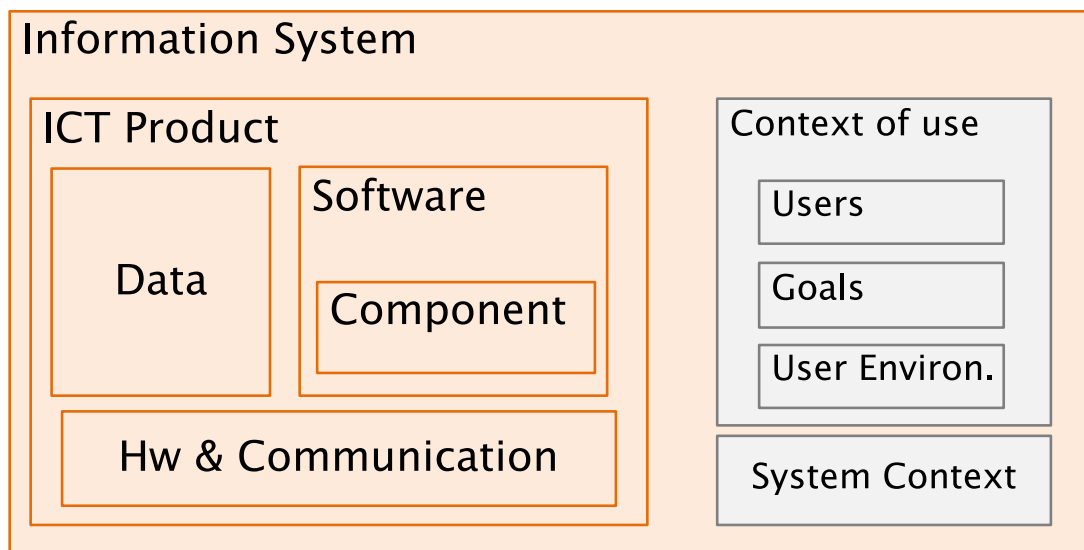
The licensor cannot revoke these freedoms as long as you follow the license terms.
- Under the following terms:
 - **Attribution** - You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
 - **ShareAlike** - If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.

Software Quality

Software Qualities

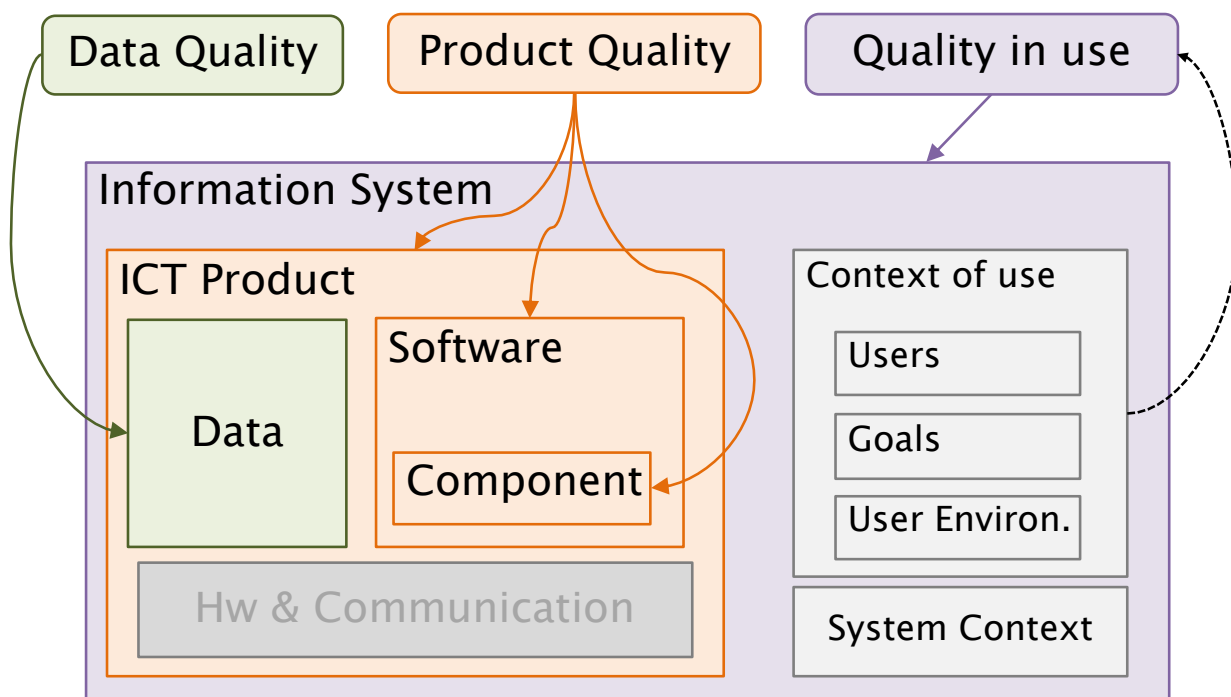


Target entities



5

Target entities vs. Qual Models



6

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

7

ISO SQuaRE – Standard Family

2503x Quality Requirements	2501x Quality Model	2504x Quality Evaluation
	2500x Quality Management	
	2502x Quality Measurement	

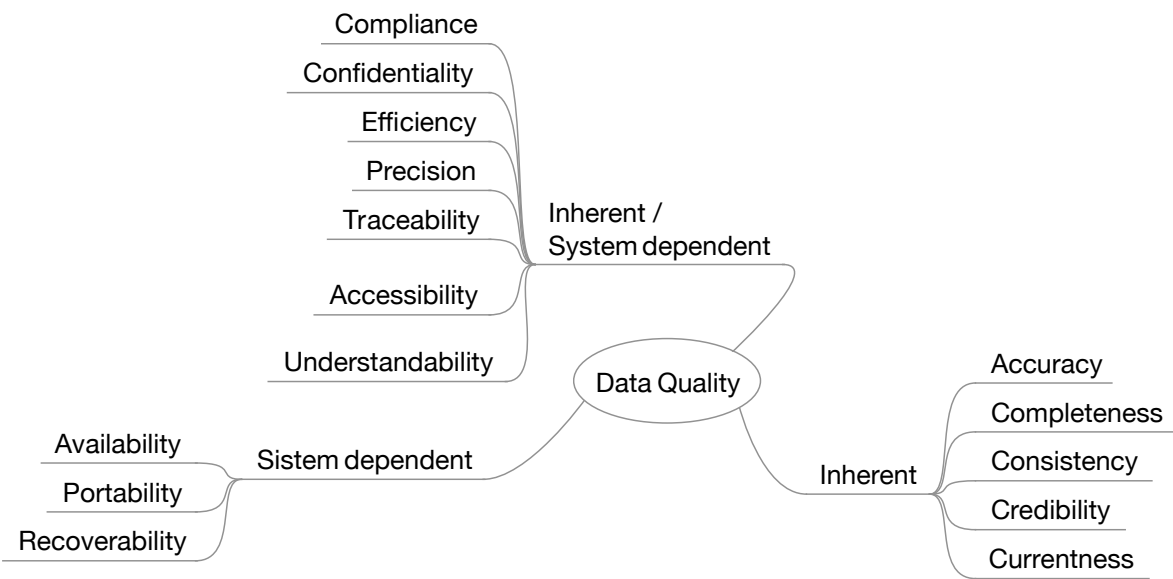
8

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



11

Quality characteristics

Inherent	Inh. / Sys. Dep.	Sys. Dep.
Accuracy	Accessibility	Availability
Completeness	Compliance	Portability
Consistency	Confidentiality	Recoverability
Currency	Efficiency	
Credibibility	Precision	
	Understandability	
	Traceability	

12

Accuracy

Correspondence between data and reality

- Syntactic
 - Value belongs to a set of validated information
- Semantic
 - The meaning (the content) corresponds to the reality

13

Accuracy: Open vs. Closed World

- Closed World Assumption (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 accurate, otherwise it is wrong
- Open World Assumption (OWA):
 - The knowledge represented in the data is (knowingly) incomplete
 - E.g., if a code appears in the list of valid codes it is accurate, otherwise it is not possible to immediately decide

14

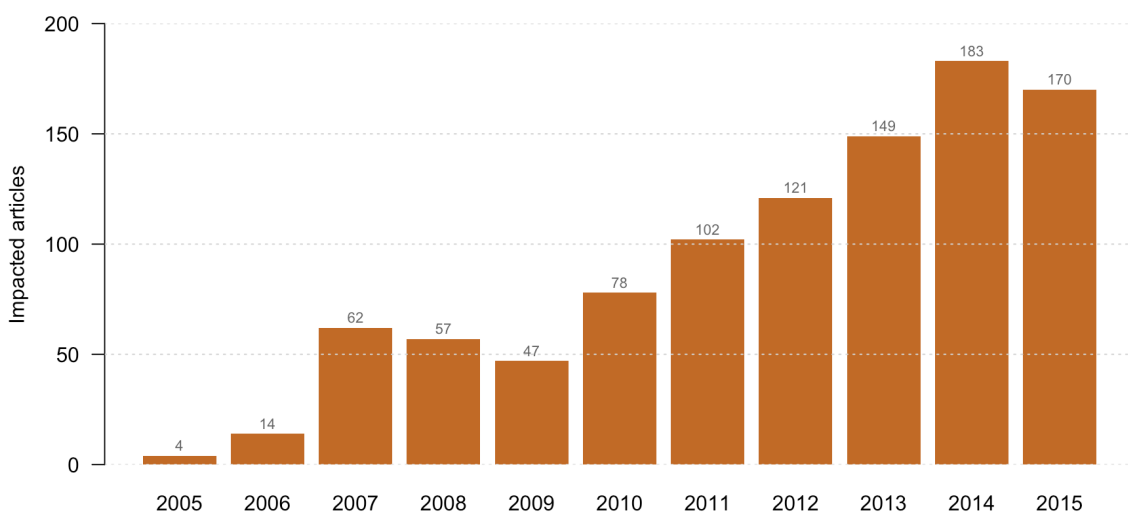
CWA – Accuracy Example : Genomics

- 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 a spreadsheet is automatically turned into 'February 2', a date.

15

CWA - Accuracy example : Genomics

Up to 20% of articles in genomics journals have errors



Source: [Ziemann et al. Genome Biology, \(2016\):17\(1\)](#)

16

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

17

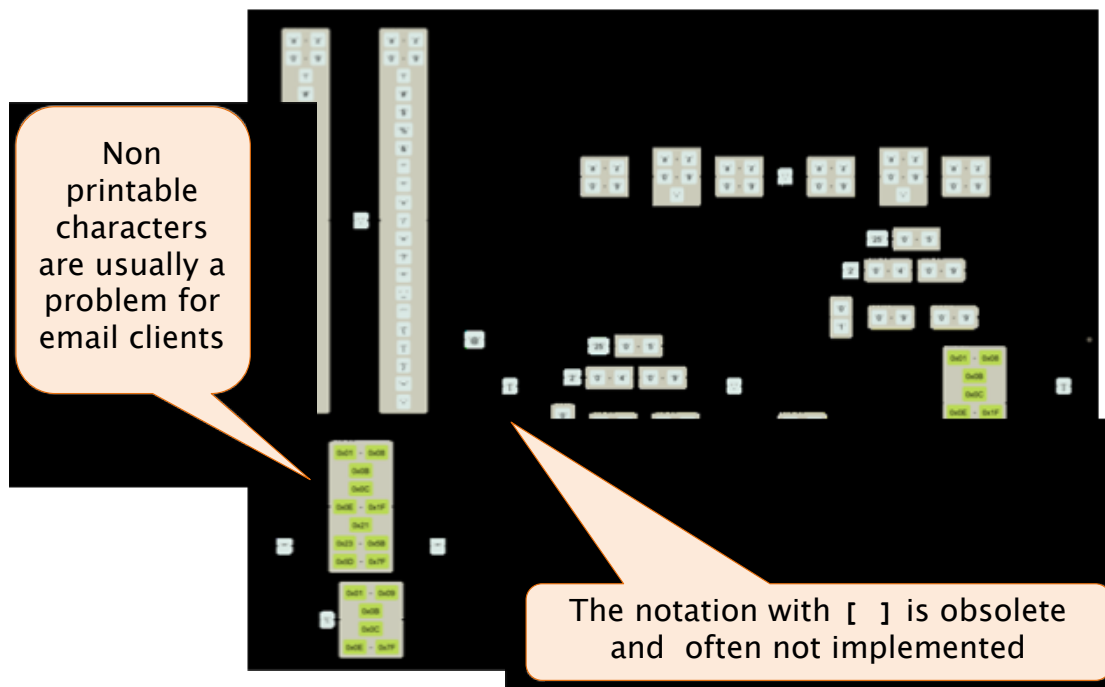
OWA - Accuracy

Where do rules come from?

- Standards
- Domain knowledge
- Similar data
- Past data

18

OWA: Email per RFC-5322



19

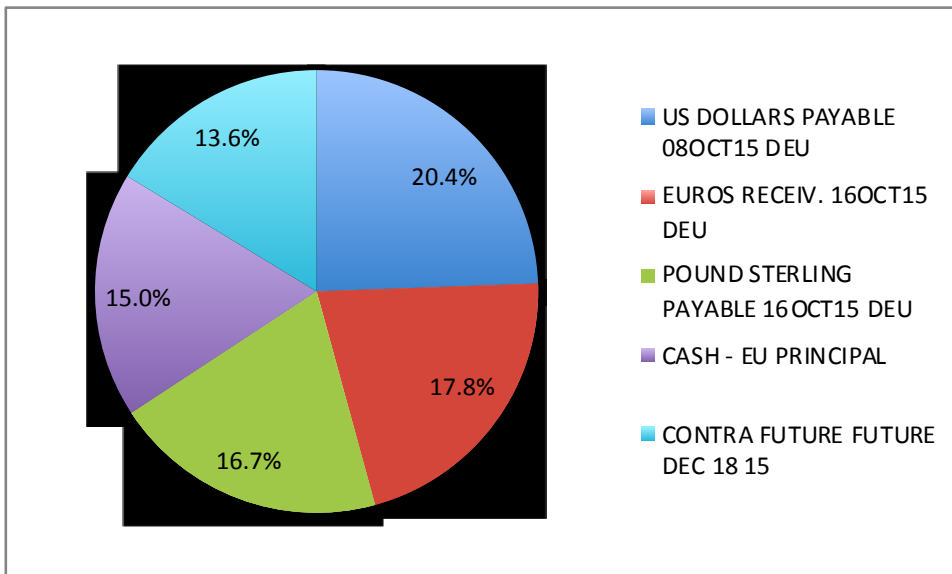
Completeness

Two different points of view:

- 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

20

Completeness



Sum of percentages: 83.5%
We miss the remaining 16.5%

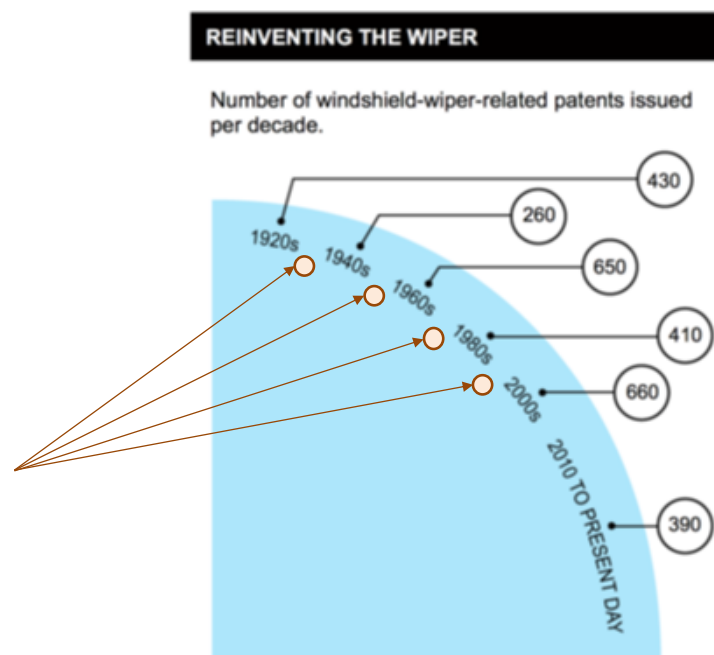
Also consistency:
expected 100%

21

Completeness

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

A possible hypothesis,
another one considered later



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

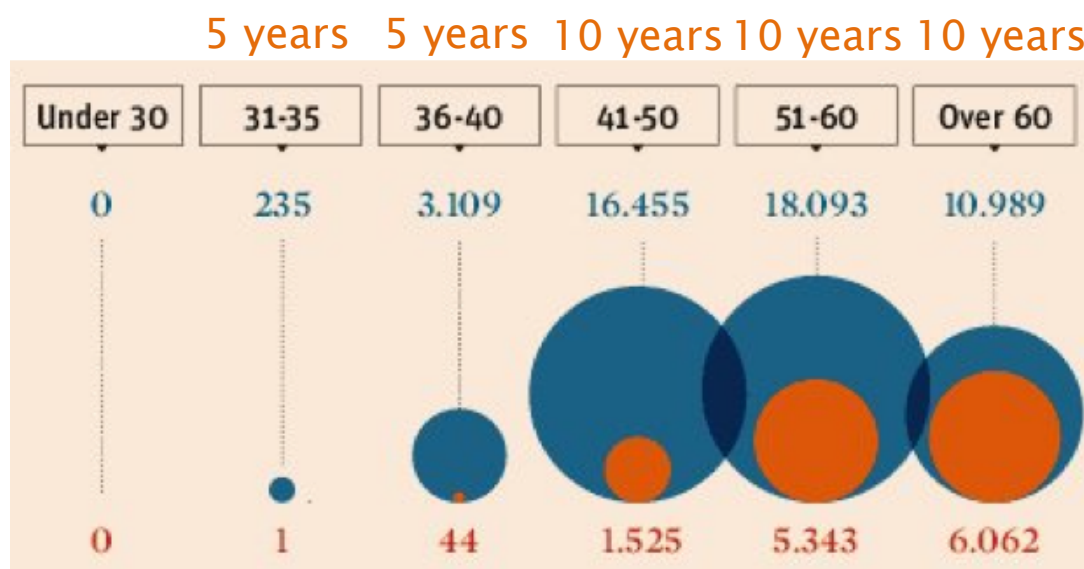
23

Consistency in graph data

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

24

Aggregation level



Source: Corriere della Sera, 09 Settembre 2017

25

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

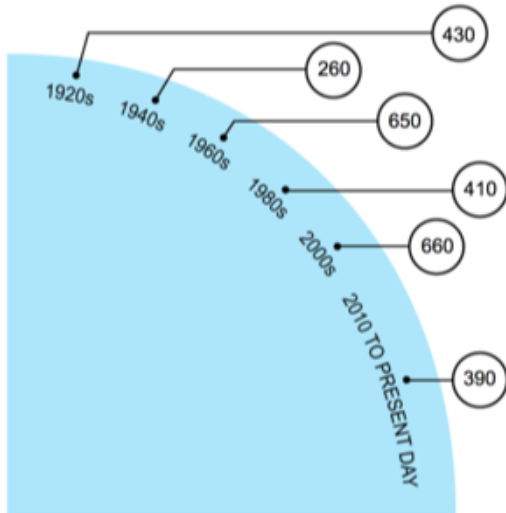
When entities or categories have different size, normalized values (i.e. densities) are comparable.

26

Consistent timeframe

REINVENTING THE WIPER

Number of windshield-wiper-related patents issued per decade.



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

27

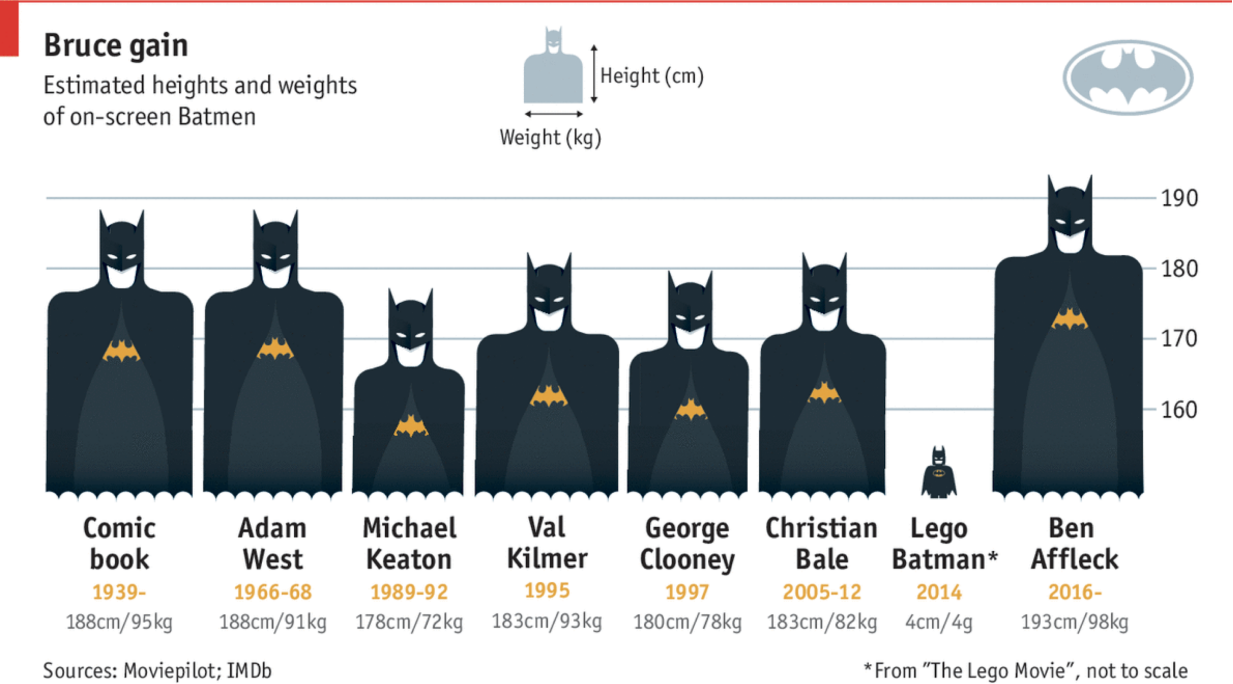
Consistent timeframe

Period	Duration	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

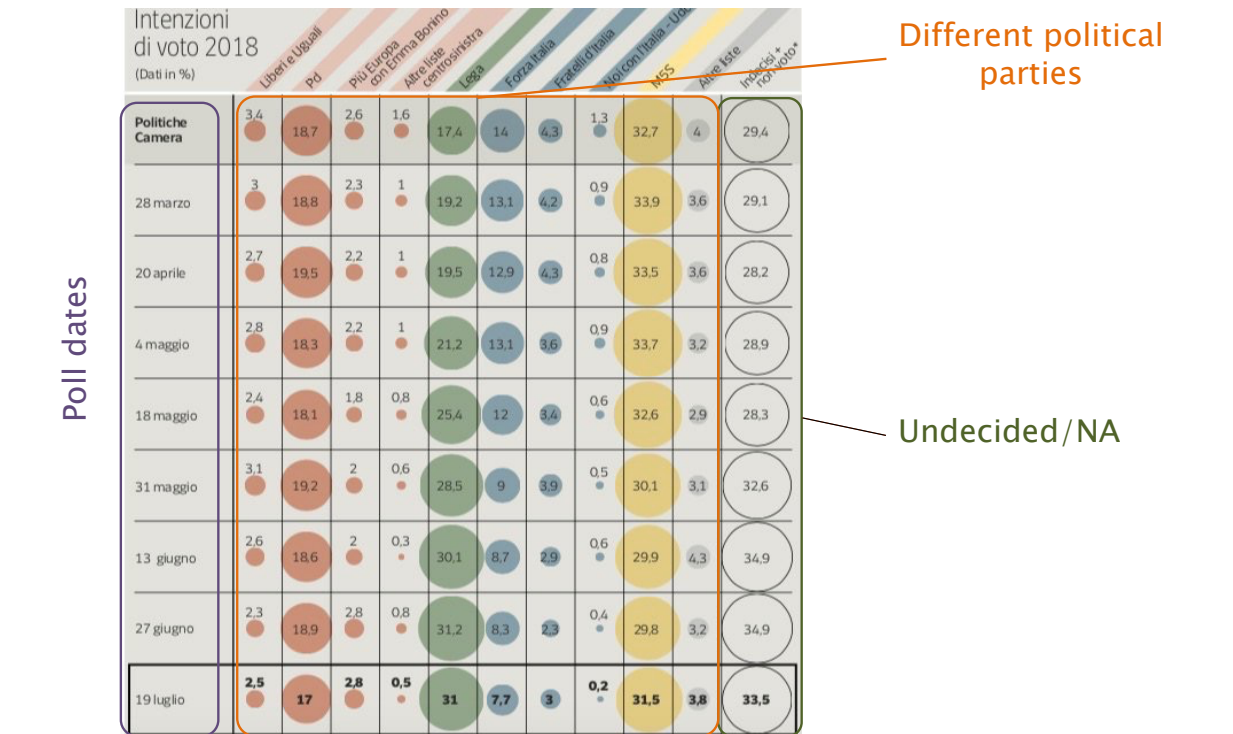
When comparing values corresponding to entities or categories with different size, normalized values (i.e. densities) are comparable, absolute values are not!

28

Consistent target entities



Consistent target



Consistent target

Proportions computed on different reference wholes

- Proportion of undecided refers to whole sample

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

- Party's proportions refer to non-undecided

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

31

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 nonequivalent times
 - e.g. different period of year, week, day

32

Understandability

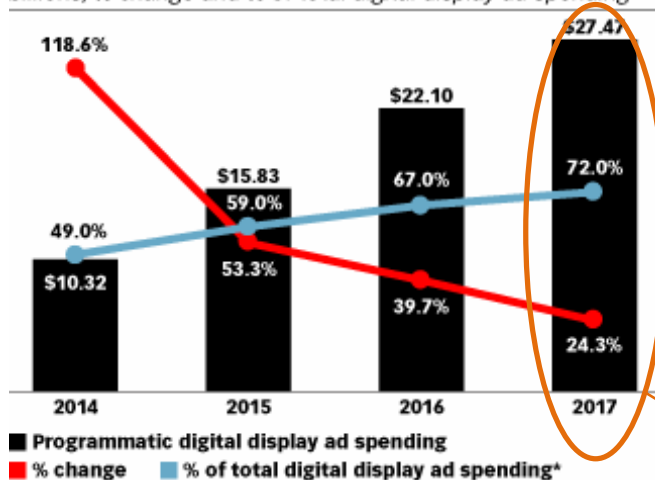
The extent to which data can be read and interpreted by users

- How is data measured?
 - Is there a track of how values are collected, measured or estimated?
 - When different methods are used that might represent an *Inconsistency* issue.

33

Understandability

US Programmatic Digital Display Ad Spending, 2014-2017
billions, % change and % of total digital display ad spending*



Note: digital display ads transacted via an API, including everything from publisher-erected APIs to more standardized RTB technology; includes native ads and ads on social networks like Facebook and Twitter; includes advertising that appears on desktop/laptop computers, mobile phones, tablets and other internet-connected devices; *includes banners, rich media, sponsorship, video and other
Source: eMarketer, April 2016

207037

www.eMarketer.com

Data from 2016 including values for 2017.

Undeclared mix of projections and final data.

34

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

35

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?

36

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”

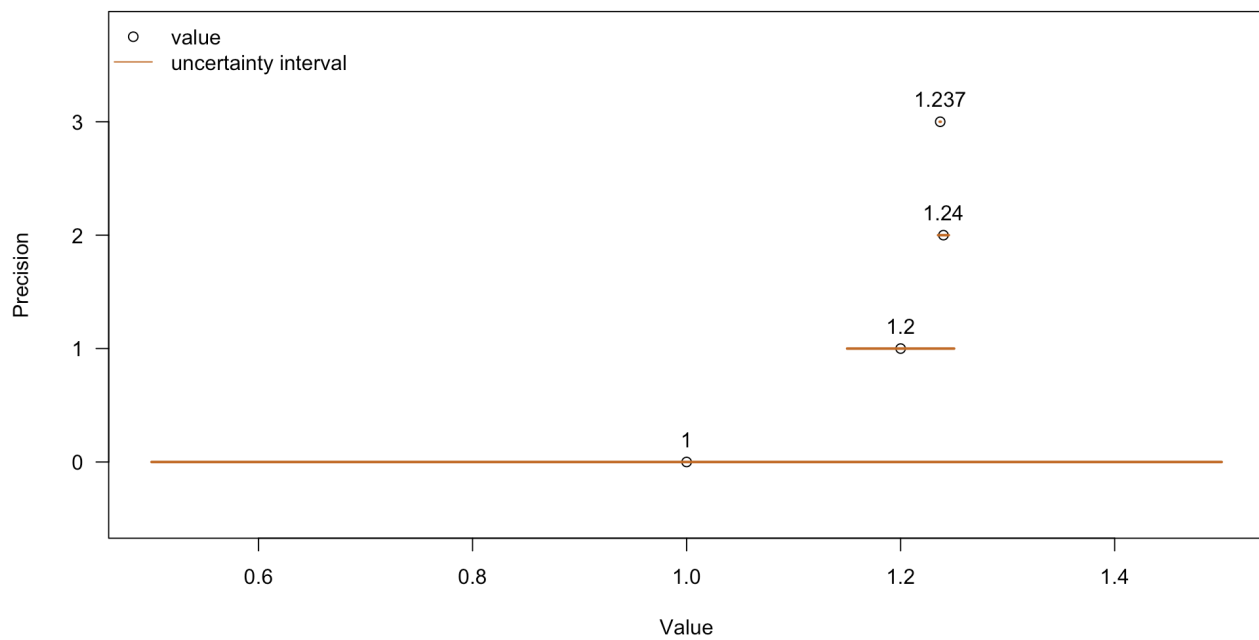
37

Precision



38

Precision and uncertainty



39

References

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

40