Data Quality

Dataforståelse

NIFR 2023

Agenda

Dataforståelse

- Data Quality Concepts
- Edit Data with MySQL
  - Update
- Identifying bad data

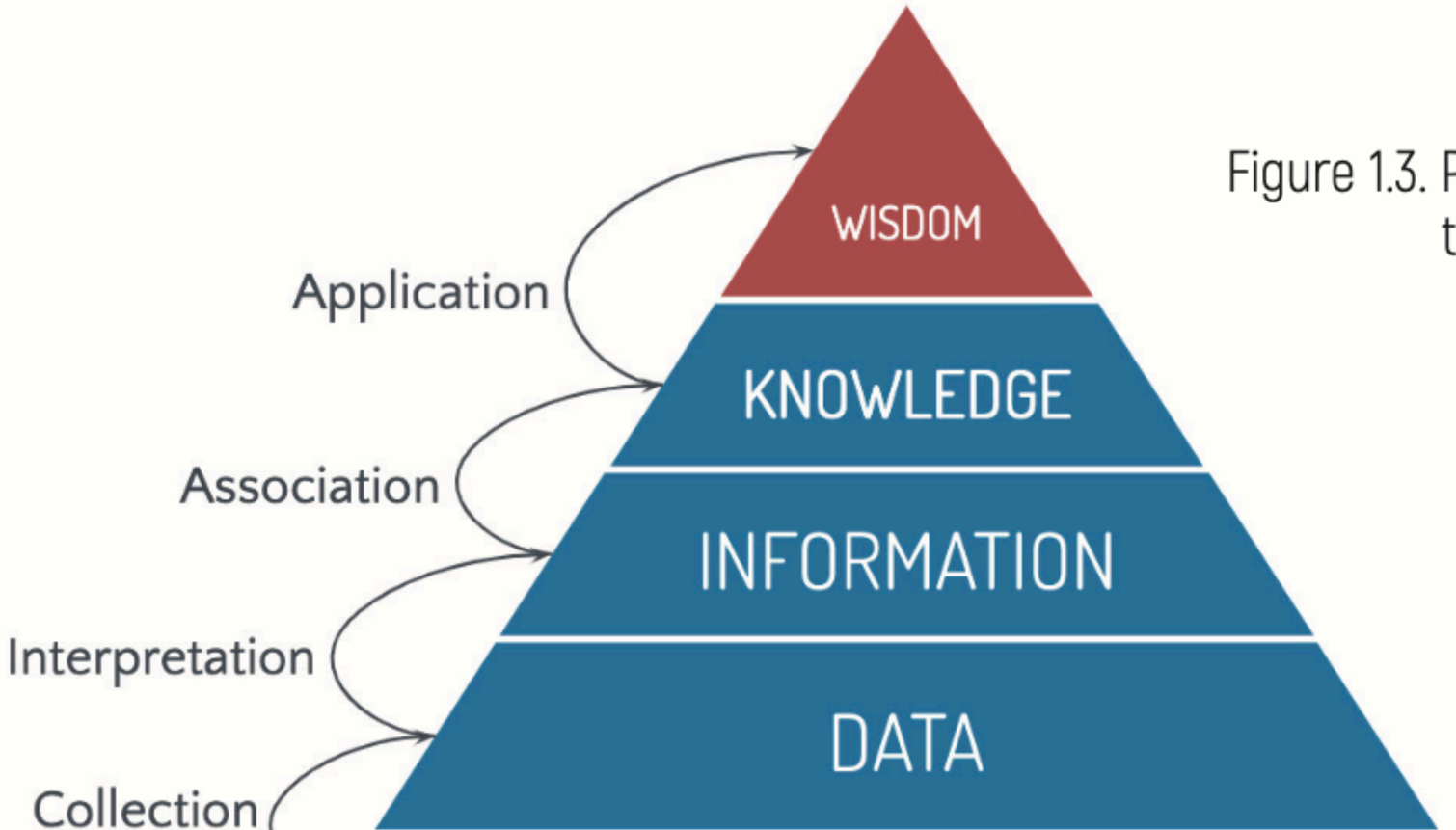
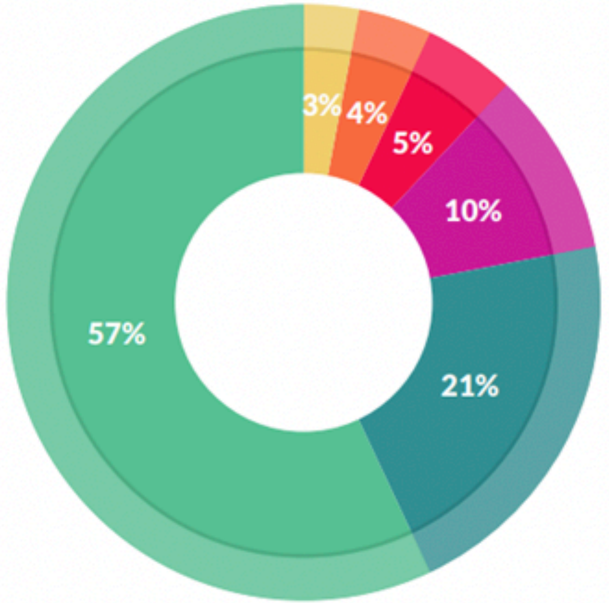


Figure 1.3. Progressing up the DIKW Pyramid.

According to a survey conducted by Figure Eight in 2016, almost 60% of Data Scientists' time is spent on cleaning and organizing data. You can find the survey results at here.



Multi Dimensional Nature of Data Quality

Parameters

- Accuracy
- Completeness
- Time-Related Dimensions
  - Currency, Timeliness & Volatility
- Consistency

Accuracy

Data Quality Dimension

For single data values, accuracy measures the distance between a value *v* and a value *v'* which is considered correct. Two kinds of accuracy can be identified, namely a *syntactic accuracy* and a *semantic accuracy*.

Accuracy

Data Quality Dimension

the distance between a value *v* and a value *v'* which is considered correct

V: Britney Spars

V':Britney Spears

Accuracy

Data Quality Dimension

the distance between a value *v* and a value *v'* which is considered correct

V: Britney Spars

V': Britney Spears

V: BriasdtneySpars

Semantic Accuracy

Data Quality Dimension

Song	Artist	Album
We Will Rock You	Robin	Neonlys

Syntactic Accuracy

Data Quality Dimension

Song	Artist	Album
We Will Rock You	Qrueen	News of the world

It is often possible to detect semantic inaccuracy in a record, and to provide an accurate value, by comparing the record with equivalent data in different sources.

Completeness can be generically defined as »the extent to which data are of sufficient breadth, depth and scope for the task at hand«

Completeness

Data Quality Dimension

- Column Completeness
  - Function of missing values in a column of a table
- Population Completeness
  - Missing Rows

Completeness

Data Quality Dimension

- Value completeness: null values for some attribute
- Tuple completeness: A whole tuple (row) completeness
- Attribute completeness: Number of null values of a specific attribute
- Relation completeness: Null values in the whole relation

ID	Name	Surname	BirthDate	Email
1	John	Smith	03/17/1974	smith@abc.it
2	Edward	Monroe	02/03/1967	NULL
3	Anthony	White	01/01/1936	NULL
4	Marianne	Collins	11/20/1955	NULL

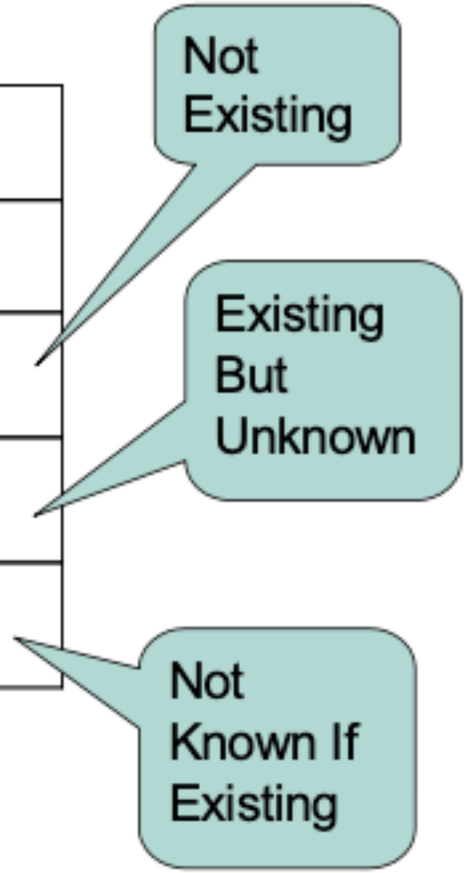


Fig. 2: Examples of different NULL value meanings

Time Related Dimensions

Data Quality Dimension

- Stable Data
- Time Variable Data

ID	Title	Director	Year	#Remakes	LastRemakeYear
1	Casablanca	Weir	1942	3	1940
2	Dead Poets Society	Curtiz	1989	0	NULL
3	Rman Holiday	Wylder	1953	0	NULL
4	Sabrina	NULL	1964	0	1985

Fig. 1: A relation *Movies* with data quality problems

Time Related Dimensions

Data Quality Dimension

- Currency
  - How often is data updated?
- Volatility
  - Frequency according to which data vary in time
- Timeliness
  - How current is data relative to usage?

Time Related Dimensions

Data Quality Dimension

- Stable Data
- Time Variable Data

ID	Title	Director	Year	#Remakes	LastRemakeYear
1	Casablanca	Weir	1942	3	1940
2	Dead Poets Society	Curtiz	1989	0	NULL
3	Rman Holiday	Wylder	1953	0	NULL
4	Sabrina	NULL	1964	0	1985

Fig. 1: A relation *Movies* with data quality problems

Consistency

Data Quality Dimension

The consistency dimension captures the violation of semantic rules defined over (a set of) data items.

Consistency

Data Quality Dimension

The movie production year must be compatible with the director's lifetime.

Ultima sequenza, L' (2003)
aka The Lost Ending
(International: English title)

Fellini: Je suis un grand
menteur (2002)
aka Federico Fellini: I'm a
Big Liar (USA: literal
English title)
aka Federico Fellini: Sono un gran
bugiardo (Italy)

Multiple violations of dimensions

Consistency & Completeness

Song	Artist	Album
We Will Rock you	Queen	NULL

UFO Exercise

Hvilke kvalitetskriterier overskrider datasettet?

Hvordan kan dette observeres?

Dato	Observationssted (postnr. og by)
fredag.09.1976	5610 og 5492 Voldbro og Bred
1999	5200 Odense V
10.01 og 15.01.2000 (usikkert)	over nordlige Tyskland (tror jeg)
ultimo 12.2018	7800 Skive
11.01.2019	5853 Ørbæk
25.01.2019	9293 Kongerslev

http://www.sufoi.dk/obs/obs-2019/obs19-k1.php

```sql
SELECT *
FROM pokemon
ORDER BY speed desc
LIMIT 10;
```

101	Electrode	140	80	80	70	50	60
150	Mewtwo	130	90	154	90	110	106
142	Aerodactyl	130	75	60	65	105	80
135	Jolteon	130	95	110	60	65	65
65	Alakazam	120	95	135	45	50	55
51	Dugtrio	120	70	50	50	80	35
121	Starmie	115	85	100	85	75	60
53	Persian	115	65	65	60	70	65
128	Tauros	110	70	40	95	100	75
26	Raichu	110	80	90	55	90	60

Identifying bad data

Data exploration

Correcting the values

Data cleaning

SQL Update

```
UPDATE employees
SET
    email = 'mary'
WHERE
    employeeNumber = 1056;
```

Table name

```
UPDATE employees
SET
    email = 'mary'
WHERE
    employeeNumber = 1056;
```

Column name

```
UPDATE employees
SET
    email = 'mary'
WHERE
    employeeNumber = 1056;
```

Constraint

```
UPDATE employees
SET
    email = 'mary'
WHERE
    employeeNumber = 1056;
```

IS NULL & DISTINCT

Data exploration

Par-øvelse

- Data exploration
  - Hvordan ser datasettet ud? Hvad beskriver det?
  - Identificer eventuelle kolonner med NULL værdier
  - Beskriv: Hvilke kolonner har problemer med fejlværdier?

Data exploration & Data Cleaning

Par-øvelse

- Ret 3 **NULL** fejl i datasettet vha. UPDATE
- Beskriv:
  - Hvordan fandt i fejlen
  - Hvordan rettede i fejlen
  - Hvordan kan denne fejl perspektiveres til viden/begreber ved. Data quality

Data exploration & Data Cleaning

Par-øvelse

- Ret 3 **umulige** værdier i datasettet vha. UPDATE
- Beskriv:
  - Hvordan fandt i fejlen
  - Hvordan rettede i fejlen
  - Hvordan kan denne fejl perspektiveres til viden/begreber ved. Data quality