Data quality thoughts(?)

Sébastien Salva

LIMOS – Université d'Auvergne

Qui suis-je?

```
Public void setUp(){
Identity id=new Identity(‘’salva’’);}

Public void testid (){
assertEquals(id.surname, ’’sébastien’’);
assertEquals(id.name, ’’salva’’);
assertEquals(id.labo, ’’LIMOS’’);
assertEquals(id.city ’’Clermont-Ferrand’’);

assertArrayEquals(i.recherche, new String[] {‘’test basé modèle’’, ‘’sécurité’’, ’’inférence de modèles’’});
}
```

definition

- *Data quality* refers to the state of qualitative or quantitative pieces of information. There are many definitions of *data quality* but *data* is generally considered high *quality* if it is "fit for [its] intended uses in operations, decision making and planning".

- "*the degree of fulfilment of all those requirements defined for data, which is needed for a specific purpose*".

Some quality attributes

- Consistency
- Accuracy
- Completeness
- Auditability
- Orderliness
- Uniqueness
- Timeliness.

Audits?

- Verification
- Validation
- Self-assessment
- Internal audit
- External audit

ref

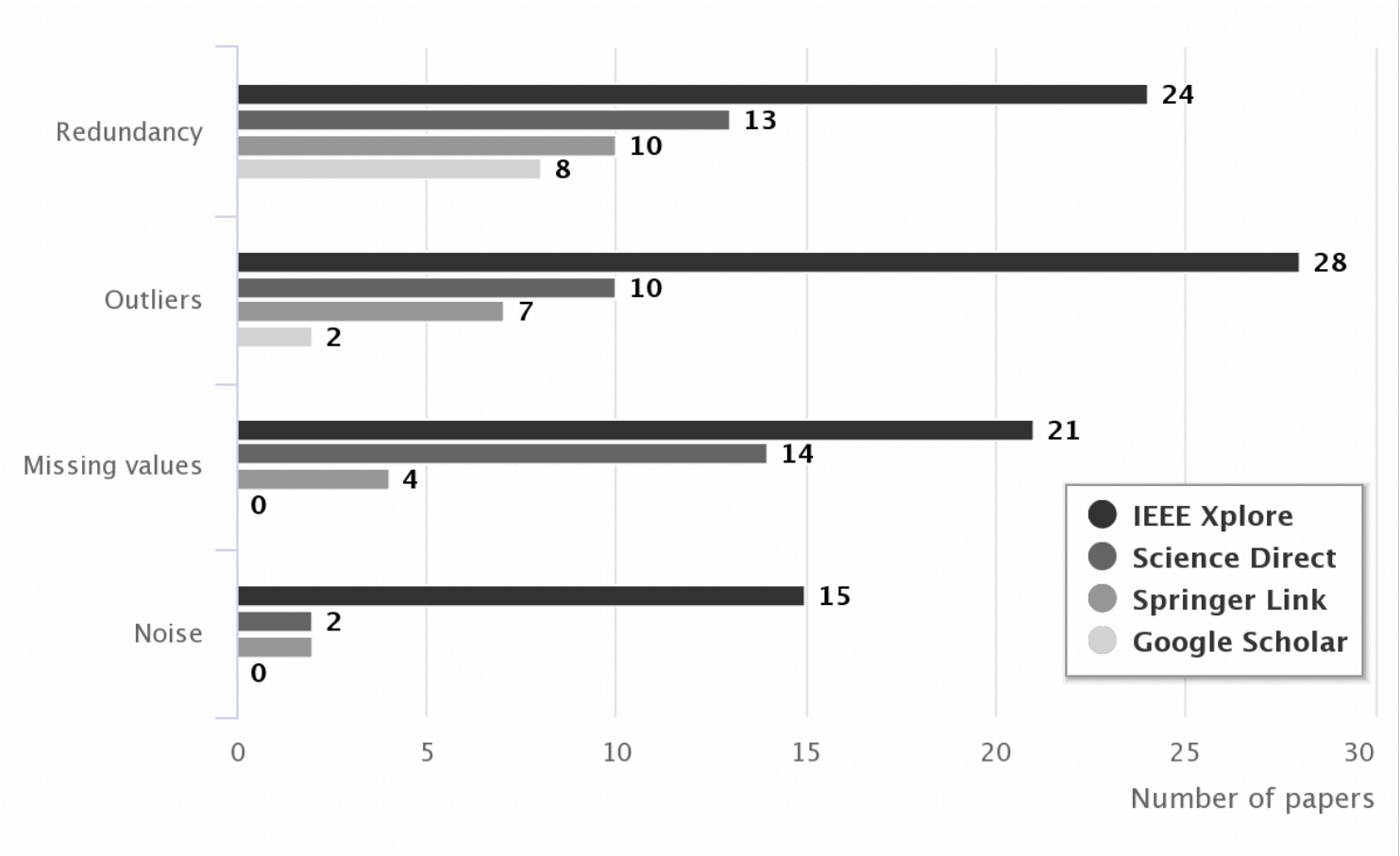
- Data quality assessment **2002**

- **Automating large-scale data quality verification 2018 (**leverages machine learning, e.g., for enhancing constraint suggestions, for estimating the 'predictability' of a column, and for detecting anomalies in historic data quality time series.

- **Does repeated measurement improve income data quality?**
- **Data Quality Measures and Data Cleansing for Research Information Systems**
- **A grounding-based ontology of data quality measures**
- **How to Address the Data Quality Issues in Regression Models: A Guided Process for Data Cleaning**
- A Review of Big Data Quality and an Assessment Method and features of Data Qualityfor Public Health Information Systems

ref

Number of papers found for each data quality issue



Issues ?

- Data source
- Collecting Data (missing data, inconsistencies, people, etc.)
- transcritions
- Analysis tools (issue in training, assuptions modifying data,

- => solutions ?
- Need of review processes, validation processes, audit,
- Data Cleansing or cleaning
- Different auto techniques,

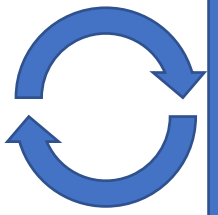
Some approaches

By hands:

- Data audit
- Data quality plan (risks, indicators, transcription, inconsistencies, etc.)
- Log reviews

Some approaches (literature)

Data monitoring



data cleaning

Data analysis algo.

Classifier

Research Information System

Etc.

Data audit verification
Repetitive quality measurement tools

Machine learning & prediction

Validation with onthologies

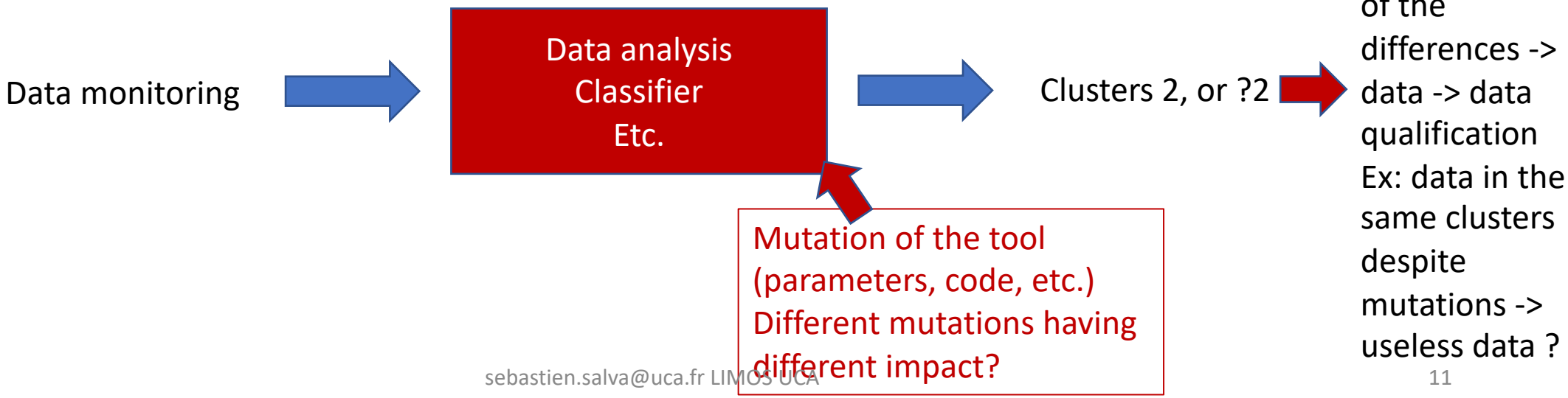
Testing with bad data (fuzzing, injections (security))

Mutation testing ?

Mutation testing ?

Used in soft. Eng. To evaluate quality of test cases -> could be used for data ?

Big picture:



And you ?