Data Quality in Environmental and Sustainability Sciences

Laure Berti-Equille
IRD
laure.berti@ird.fr

Institut de Recherche pour le Développement
FRANCE

Outline

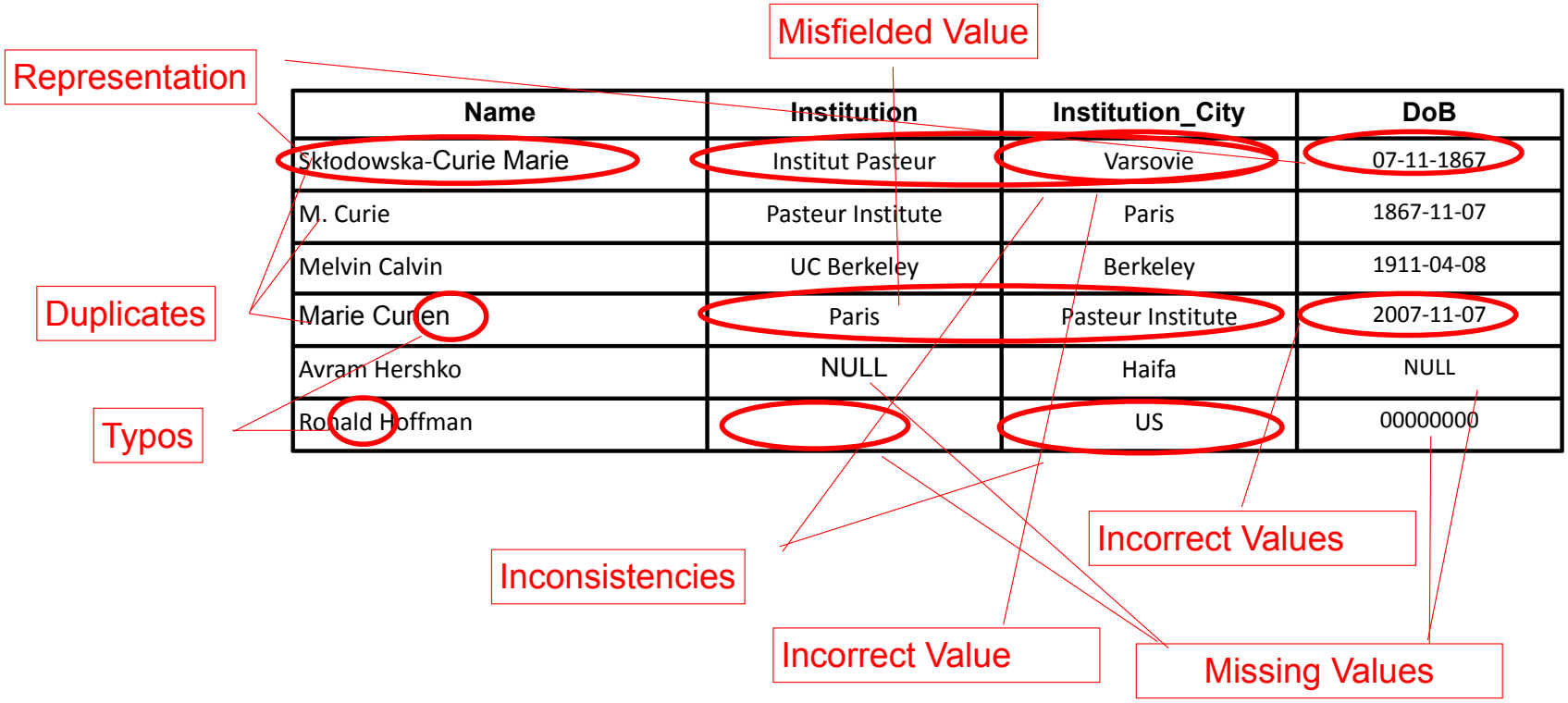
I. **Data Quality Problems**
   - Illustrative Examples
   - Importance of Profiling Data Quality

II. **Data Cleaning and Preparation**
   - Generic data pipeline
   - Automating data preparation

1. Examples of Data Quality Problems

Relational data quality problems

Nobel Laureates in Chemistry



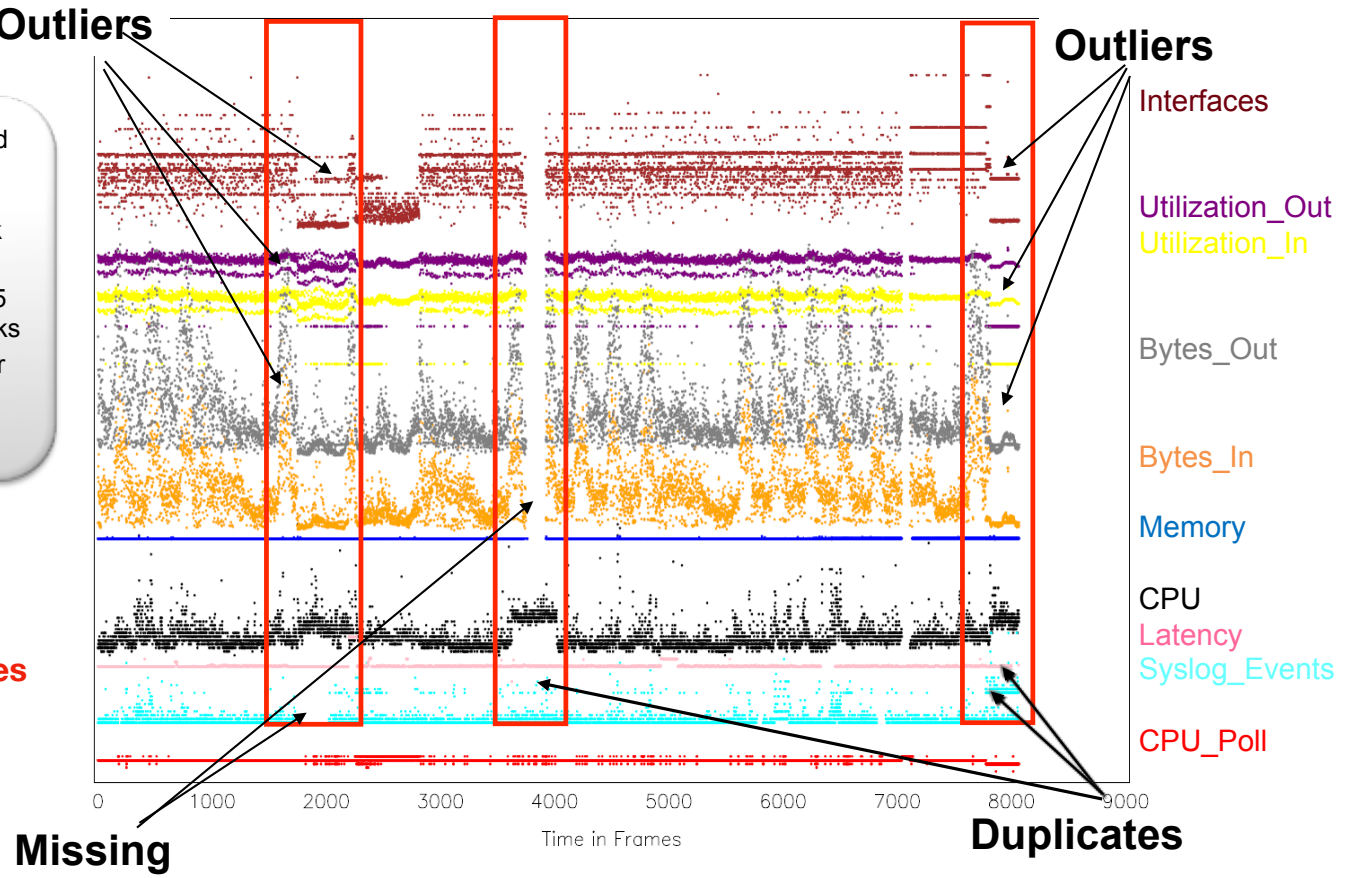
Name	Institution	Institution_City	DoB
Skłodowska-Curie Marie	Institut Pasteur	Varsovie	07-11-1867
M. Curie	Pasteur Institute	Paris	1867-11-07
Melvin Calvin	UC Berkeley	Berkeley	1911-04-08
Marie Curien	Paris	Pasteur Institute	2007-11-07
Avram Hershko	NULL	Haifa	NULL
Ronald Hoffman		US	00000000

2. Examples of Sensor Data Quality Problems

SNMP Data

- Periodic inbound and outbound traffic measurements from interfaces of network devices
- 10 attributes, every 5 minutes, over 4 weeks
- Axes transformed for plotting

Importance of Patterns of Anomalies



3. PFAS Data Quality Problems (1/2)

1. Inconsistent and Incomplete Identification

- **Incomplete DB coverage**
  Current databases often cover only a fraction of existing PFAS compounds.
- **Lack of standardized naming conventions**
  Misidentification, duplication, and difficulty in combining data from different sources.
- **Unknown/Unreported PFAS**
  Many PFAS in use or present in the environment may be proprietary or simply unknown.

2. Variable Data Formats and Metadata: Lack of Standardization

- **Inconsistent Formats**
  PFAS Data stored in different formats (spreadsheets, scientific publications, proprietary databases) with various conventions: e.g., PFAS concentrations in parts per million (ppm), while another uses parts per billion (ppb), hindering integration for analysis.
- **Missing Metadata**
  Crucial details about samples (location, collection date, analytical method, etc.) are often incomplete or missing.

3. PFAS Data Quality Problems (2/2)

3. Limited Accessibility, Interoperability, and Reusability

- **Fragmented Storage**
  PFAS data is scattered across government agencies, academic institutions, industry, etc. There's no central, comprehensive repository.

- **Manual Curation**
  Maintaining and updating data often requires manual curation, which can be time-consuming and prone to human error.

4. Measurement and Analytical Gaps

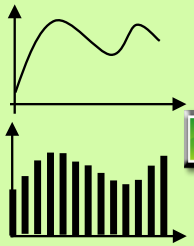
- **Detection Limits**
  Analytical techniques for PFAS are constantly evolving. Older data may have high detection limits, meaning low levels of contamination weren't captured at the time.

- **Nontarget Analysis**
  Targeted analysis looks for known PFAS. Nontarget analysis, which could identify novel PFAS, is less common and often less accessible.

Data Quality Profiling

DATA

TYPE

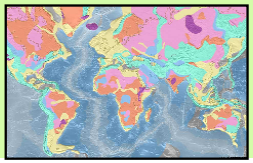


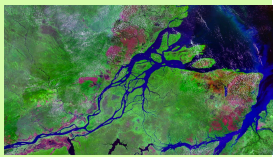
0101010101

ACACGTGT

John Doe

High
Medium
Low





RELATIONSHIP

- Structural (record)
- Sequential
- Graph-based
- Temporal
- Spatial
- Spatio-Temporal

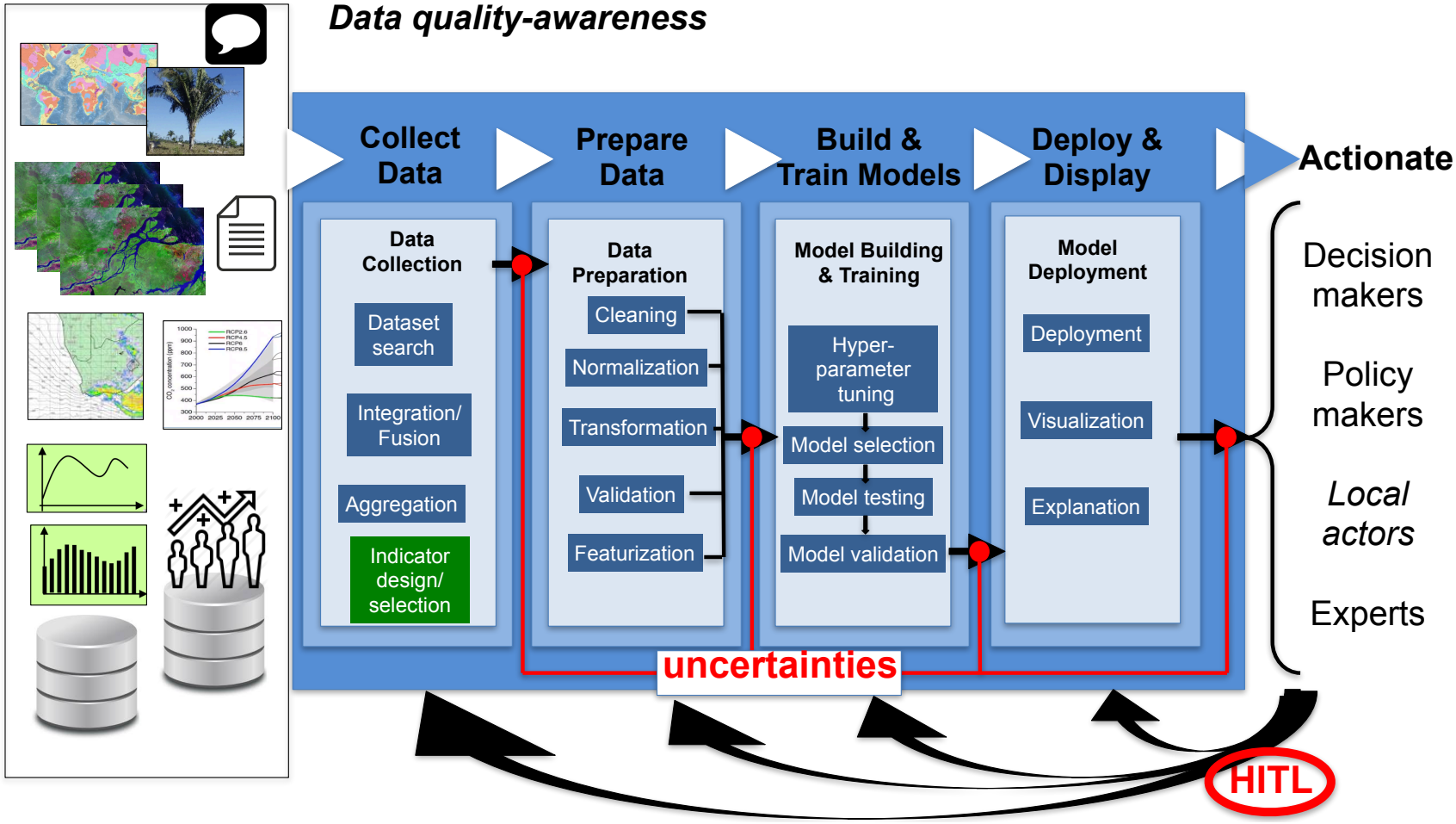
DATA QUALITY PROBLEMS

TYPE	CARDINALITY
Missing data	Single-Point
Anomalous data	Collection
Duplicate data	
Inconsistent data	
Obsolete data	
Incorrect data	

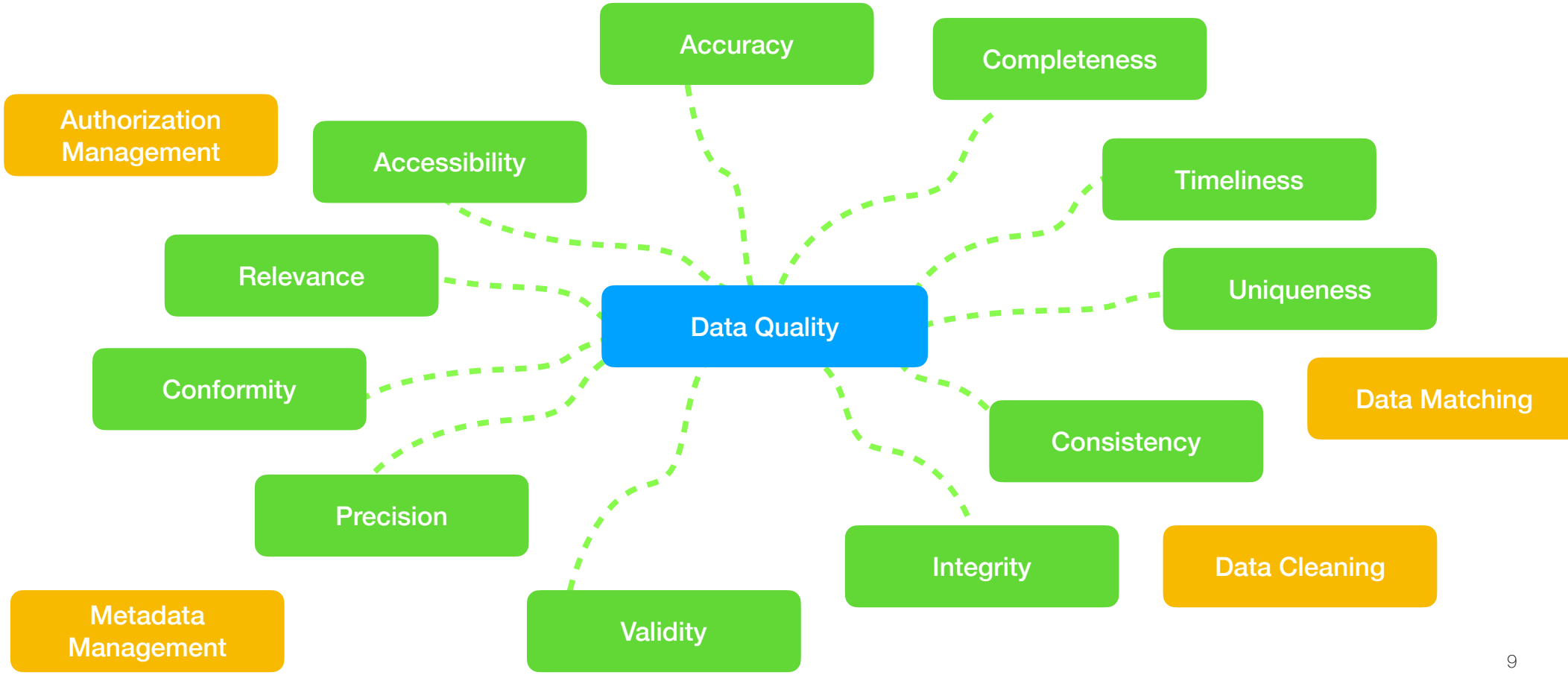
DETECTION MODE

- Model-based
- Data distribution-based
- Constraint-based
- Pattern-based

Data Science Pipeline



Data Quality Profiling: Reporting Indicators



Outline

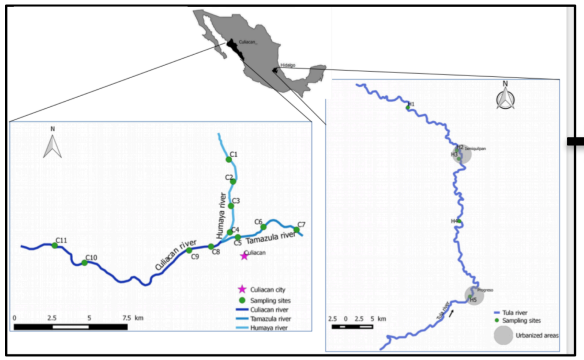
I. **Data Quality Problems**
   - Illustrative Examples
   - Profiling Data Quality

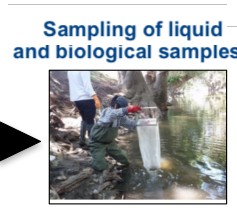
II. **Data Cleaning and Preparation**
   - Impact of Data Prep
   - Generic Pipeline

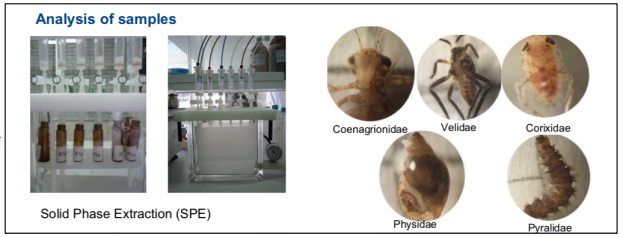
Data Preprocessing Impact (1/2)

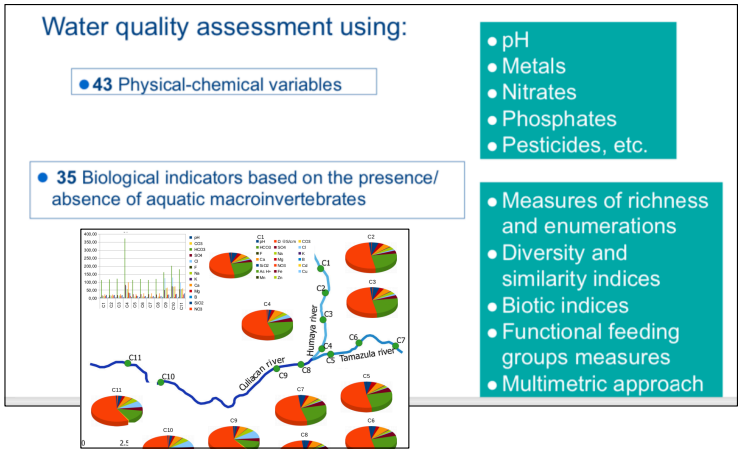
Are the Mexican rivers of Humaya, Tula, Tamazula, and Culiacan polluted ?

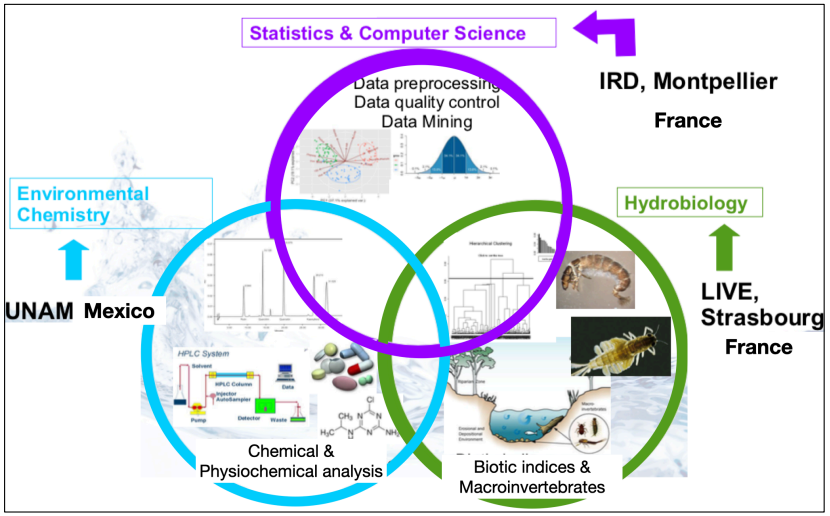
[Serrano Balderas et al. 2015, 2017]





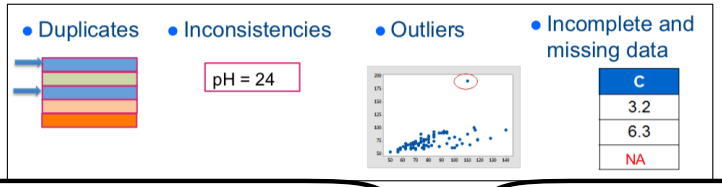




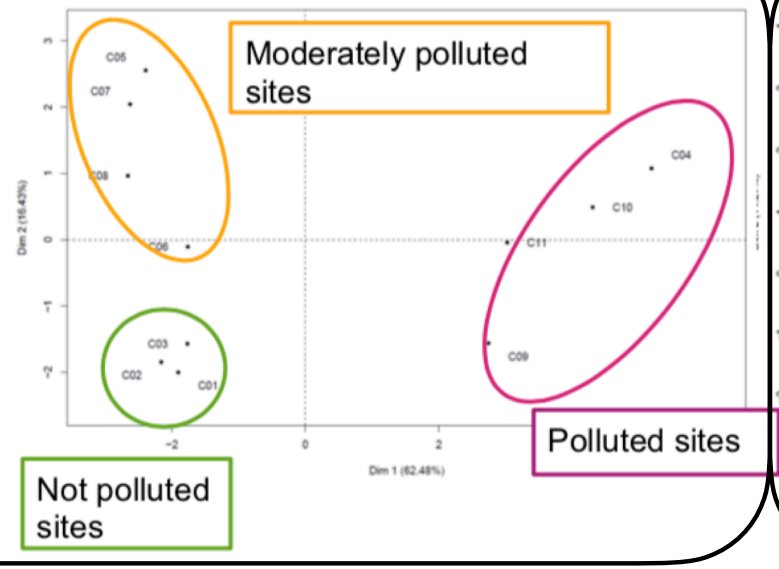


Data Preprocessing Impact (2/2)

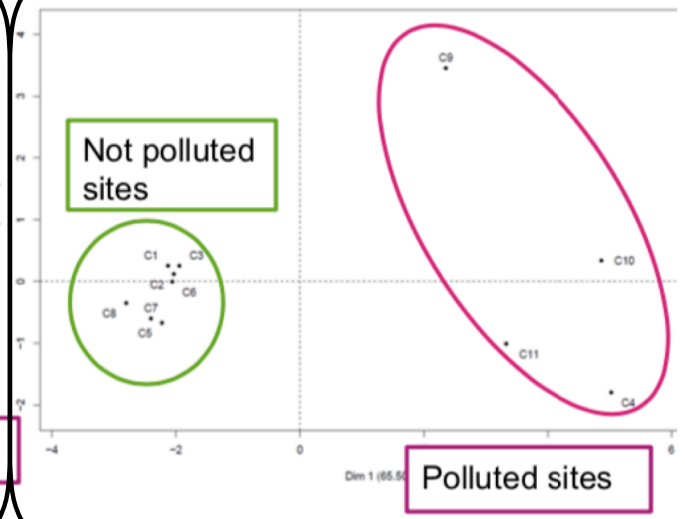
Various data cleaning/preparation strategies lead to different and misleading conclusions



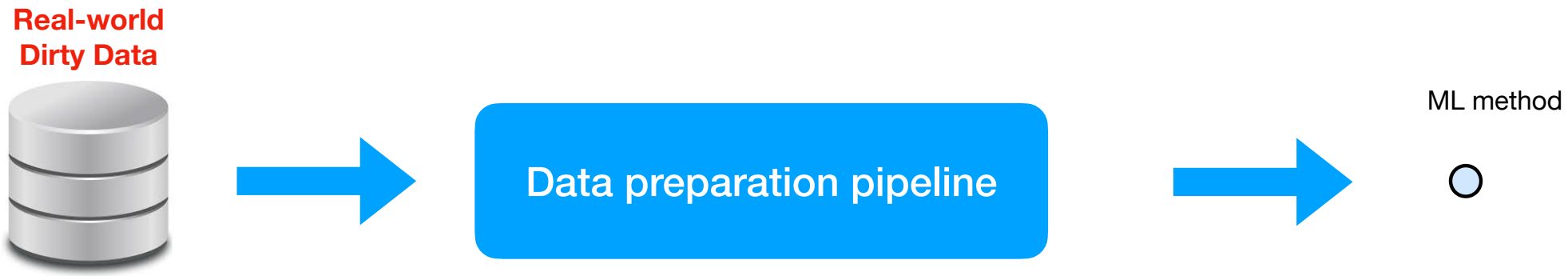
Preprocessing 1



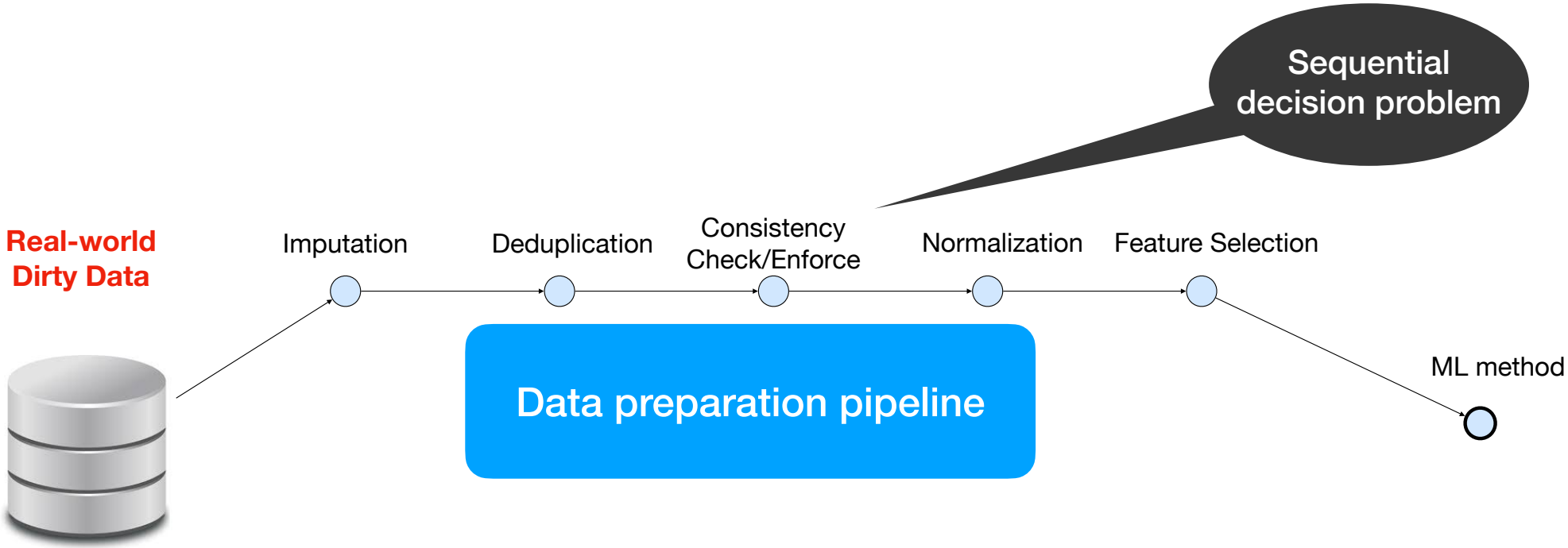
Preprocessing 2



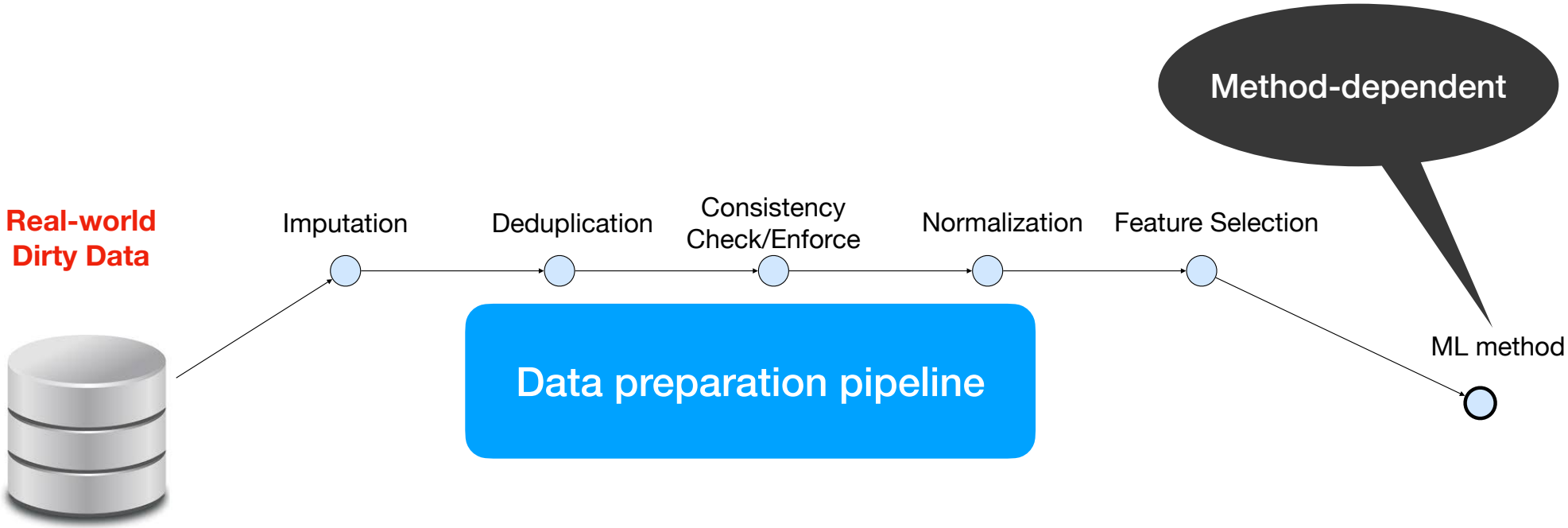
Data Preprocessing pipeline



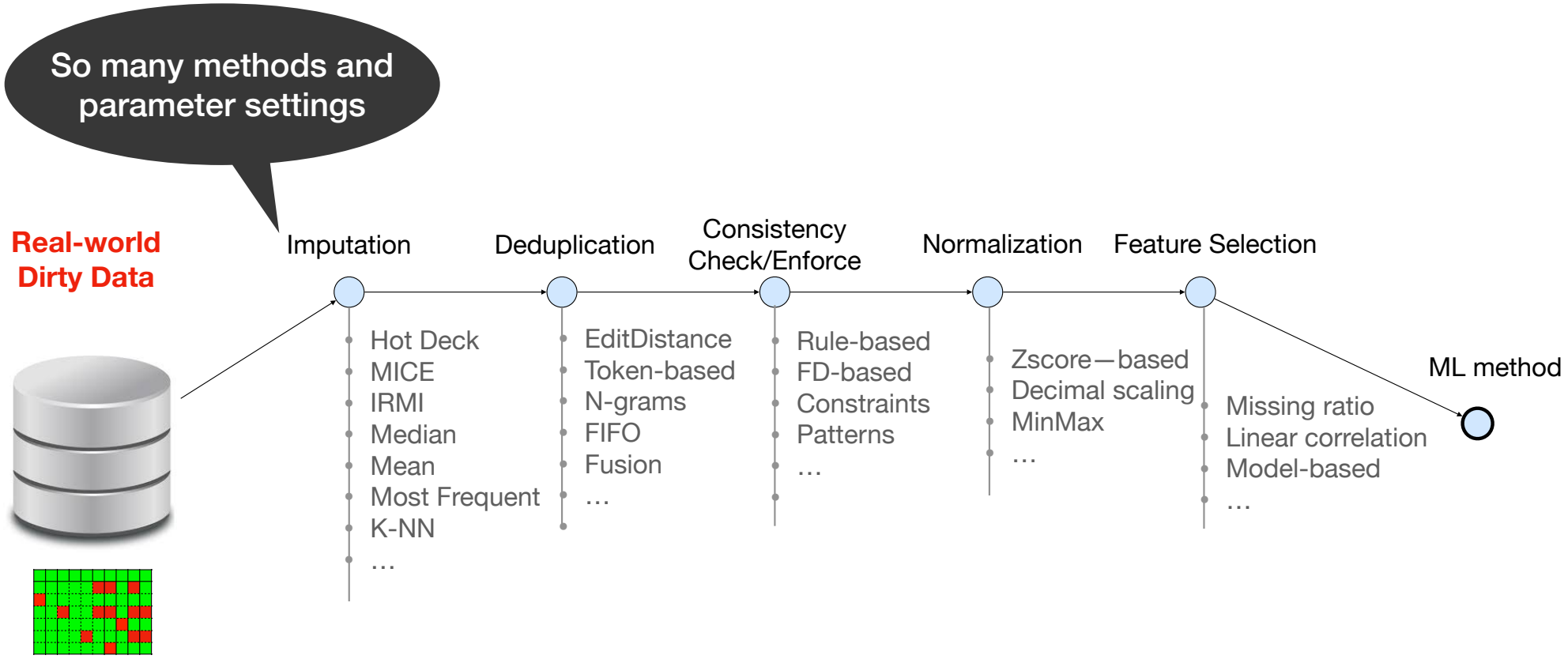
Data Preprocessing pipeline



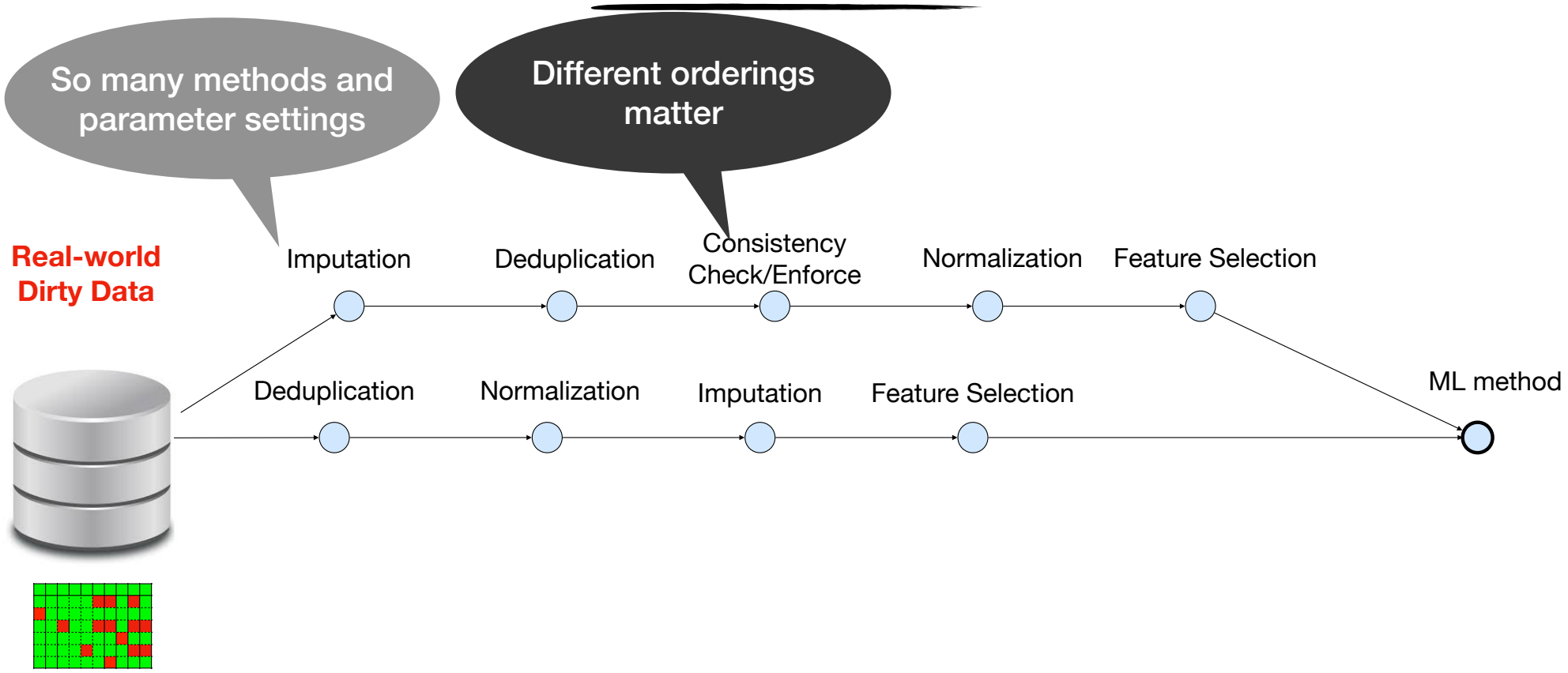
Data Preprocessing pipeline



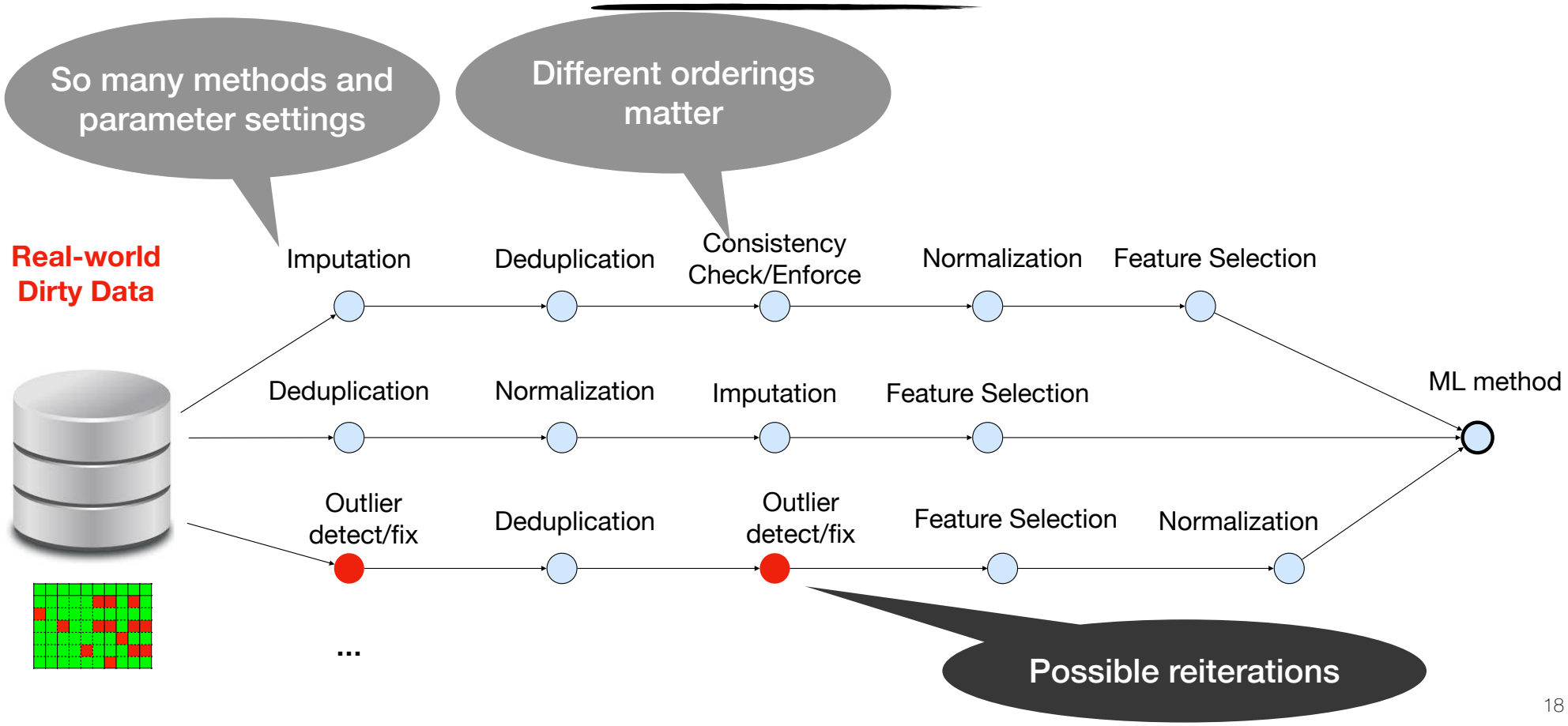
Data Preprocessing pipeline



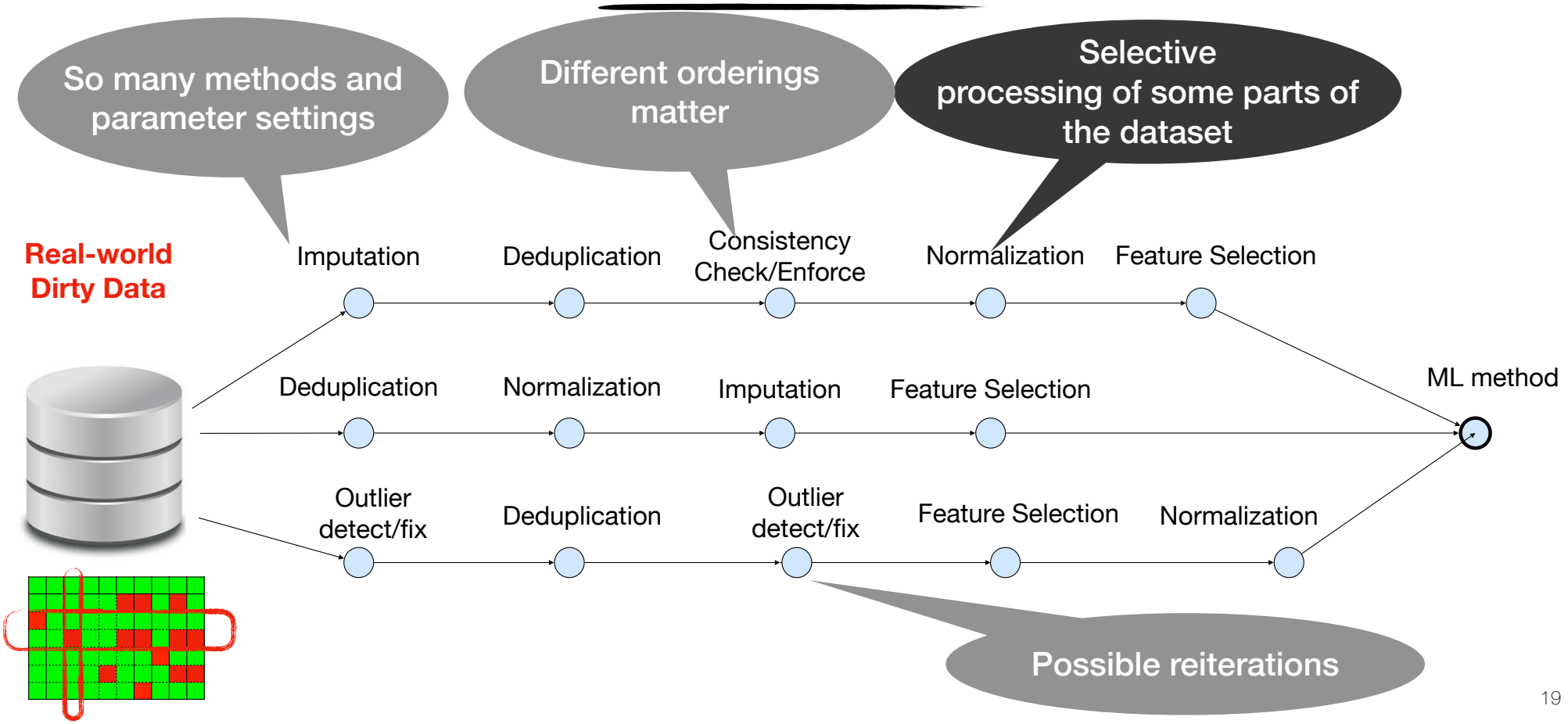
Data Preprocessing pipeline



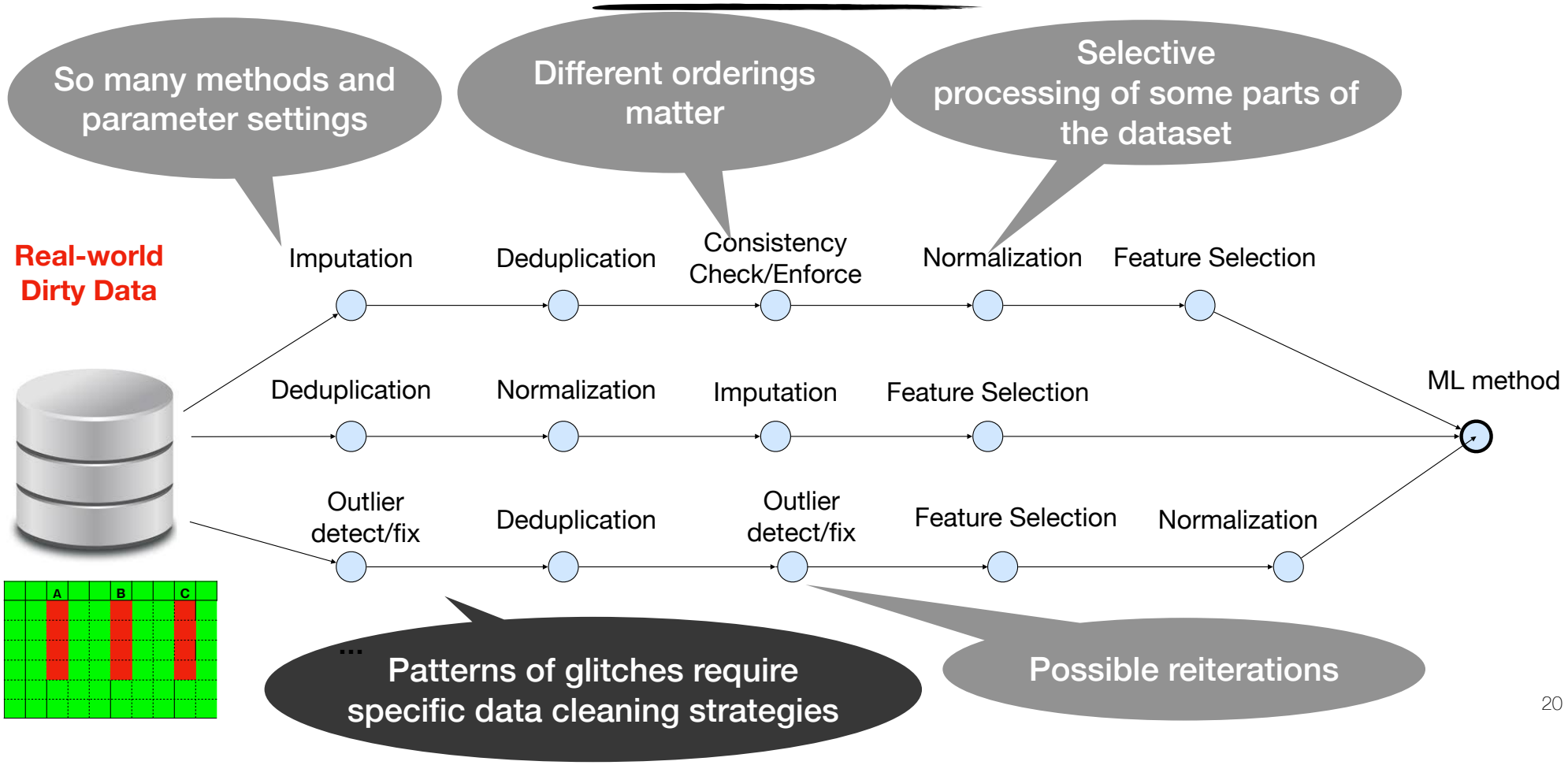
Data Preprocessing pipeline



Data Preprocessing pipeline



Data Preprocessing pipeline



Conclusions

- PFAS Data standardization and data quality profiling are mandatory starting points.
- Principled data cleaning and data preparation are essential and have a great impact on the analysis results.
- Many data preprocessing tasks require seamless integration of automated tasks and Human-in-the-Loop expertise.
- Many tools and R&D solutions exist for quantifiying & monitoring DQ

Thanks!

Laure Berti-Equille
contact: laure.berti@ird.fr
https://laureberti.github.io/website/