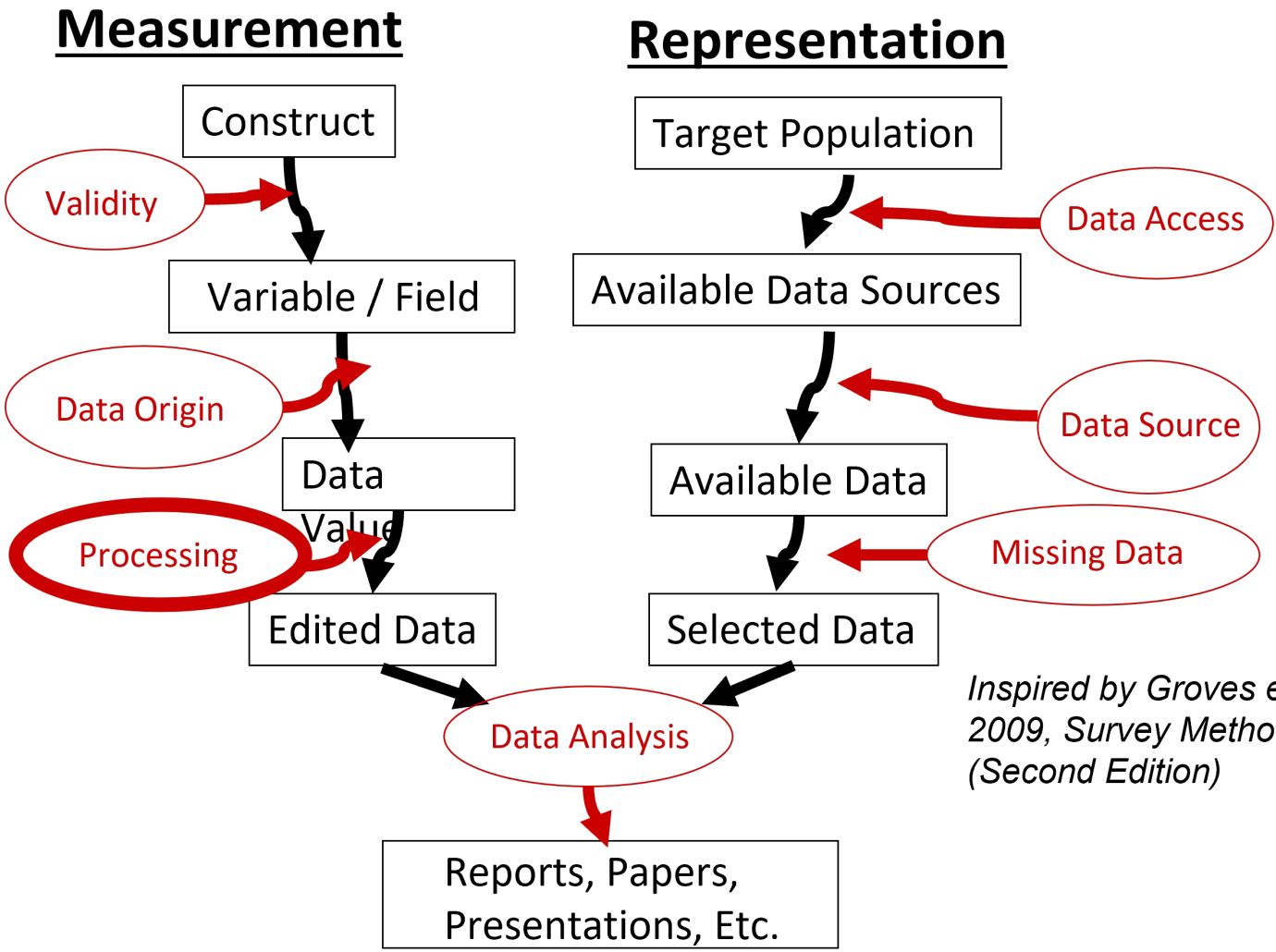
Data Processing Threats for Gathered Data

By Jinseok Kim

Dimensions of TDQ: The Big Picture!



Inspired by Groves et al. 2009, Survey Methodology (Second Edition)

Data Processing Threats (1): Parsing Errors

- Gathered data are often recorded in machine-readable formats (e.g., HTML, JSON, or XML).
- Better to be transformed into a human-readable format before analysis → 'data parsing'
  - e.g., Sequences in JSON → a table with rows and columns
- Data parsing can produce errors (e.g., missing data field).
  - Mainly due to user's lack of parsing experience or incorrect choice of parsers; or data provider's incautious or ill-defined practices.

Threats (2): Inaccurate Encoding

- Gathered data often contain characters and symbols from diverse cultures and languages (e.g., Chinese letters or ‘é’).
- Better to be converted into standard character formats before analysis → ‘Character Encoding’
  - e.g., American Standard Code for Information Interchange (ASCII)
- Text editors (e.g., word processing tools like MS Word) may not support a certain type of encoding → Some characters appear as strange symbols (e.g., I’ve → Iâ€™ve).

Threats (3): Inconsistent Identifiers

- In datasets, distinct entities like persons, places, and organizations need to be represented by unique identifiers.
  - E.g., A Social Security Number is associated with one person.
- Gathered data may have multiple data entries for the same unique entities → duplicates or split of entities.
  - E.g., A user may have two or more user ids on the same website.
- Information of multiple entities may be recorded in the same entry → contaminated or merged entities.

Threats (4): Other Common Threats

- Input Errors: inconsistency, typos, or invalid values
  - E.g., DD/MM/YY vs DD/MM/YYYY vs YYYY/MM/DD
  - E.g., numbers recorded in a person name (text string) field
- Outdated or contradicted Information
  - E.g., Birthday (Feb-18-1980) vs Age (45) as of 2020
- Missing values, Outliers, Irregular cardinality (e.g., cardinal feature mislabeled with categorical value), etc.

Threats (5): Too big to comb through

- The aforementioned common threats (e.g., missing values) are not specific to gathered data but can be more serious than in designed data.
- Why?
- ***Curse of Scale***: The large-scale-ness of gathered data can obscure our ability to properly detect data quality problems.
  - Too many variables and instances to scan (quadratic increase of search space) (**Hazen et al., 2014**).
  - Unknown (i.e., not pre-defined) errors are undetectable.

What’s Next?

- We will take a look at **data processing threats** for gathered data with a case study.

UNIVERSITY OF MICHIGAN