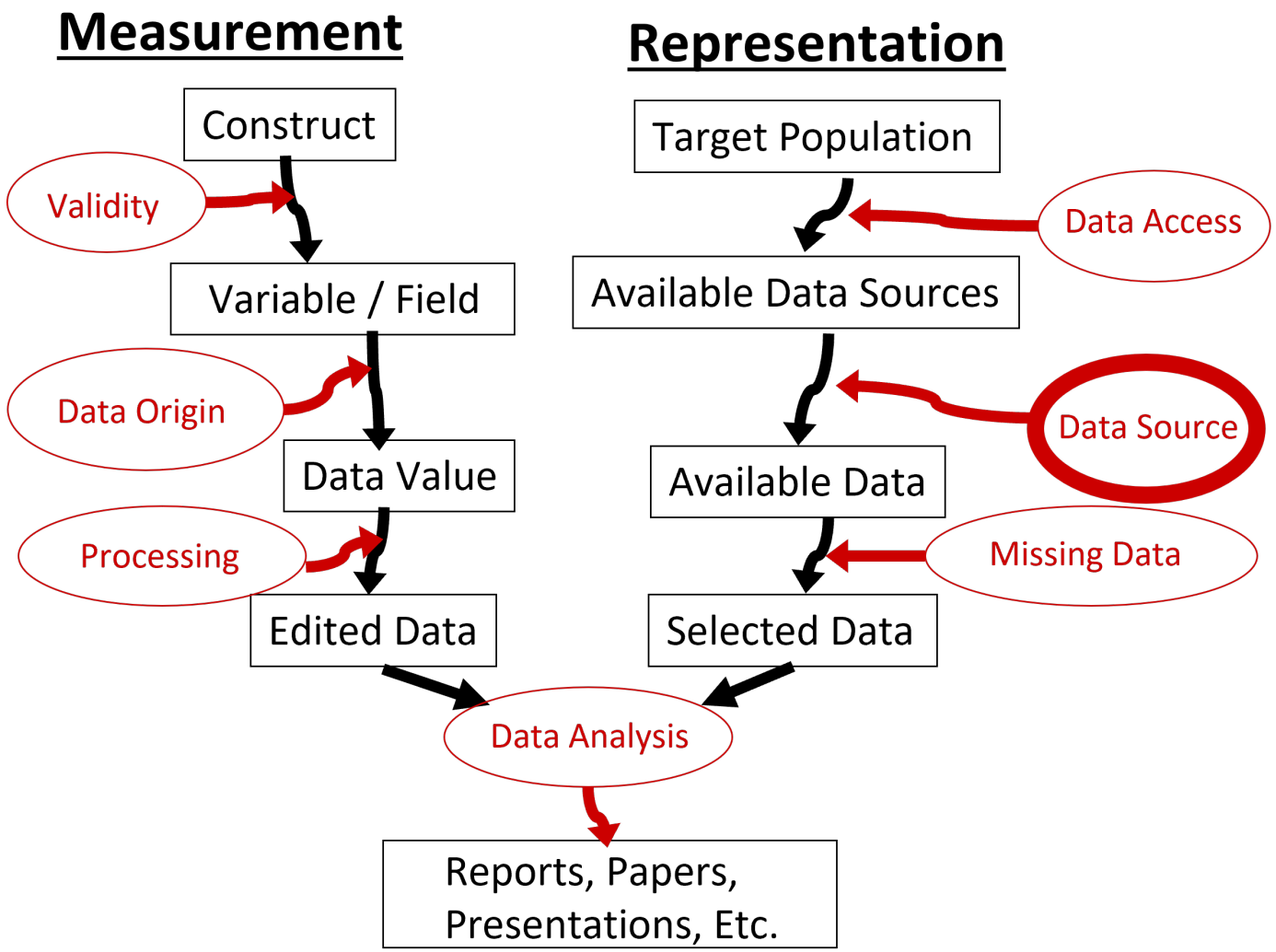
Data Source Definition
By James Wagner

Dimensions of TDQ: The Big Picture!



Data Source Definition (1)

- The **data source error** in designed data is the result of **sampling** from a population.
  - In order to reduce costs, designed data often sample from the population and then measure just the sample.
  - This approach introduces **sampling error.**
  - There is a whole subfield of statistics devoted to accurately estimating sampling error.
  - Sampling error is often reported as the **variance or standard error** of a statistic.

Data Source Definition (2)

- For gathered data, **data source errors** result from **sampling** data.
  - As with designed data, the source can be a set of materials (e.g., corpus of images) or…
  - a process (e.g., tweets).

Data Source Definition (3)

- Errors can result from sampling or selection of gathered data.
  - In some situations, with gathered data, there is no sampling.
    - Example: All of the electronic medical records for all current participants in Medicare may be available.
  - In other situations, sampling does occur.
    - Example: tweets are sampled by keyword and in time.
    - A corpus of labelled images are a sample of all images available on the internet.

Data Source Definition (4)

- Gathered data often contain designed elements.
  - For example, training data (e.g. a corpus of labelled images) may contain labels or may be generated by annotators.
  - Training data then train models that are used to label new data that do not have labels.
  - The selection of the training data can impact the results.

What’s Next?

- We will look at **data source threats** to designed data and gathered data.

UNIVERSITY OF MICHIGAN