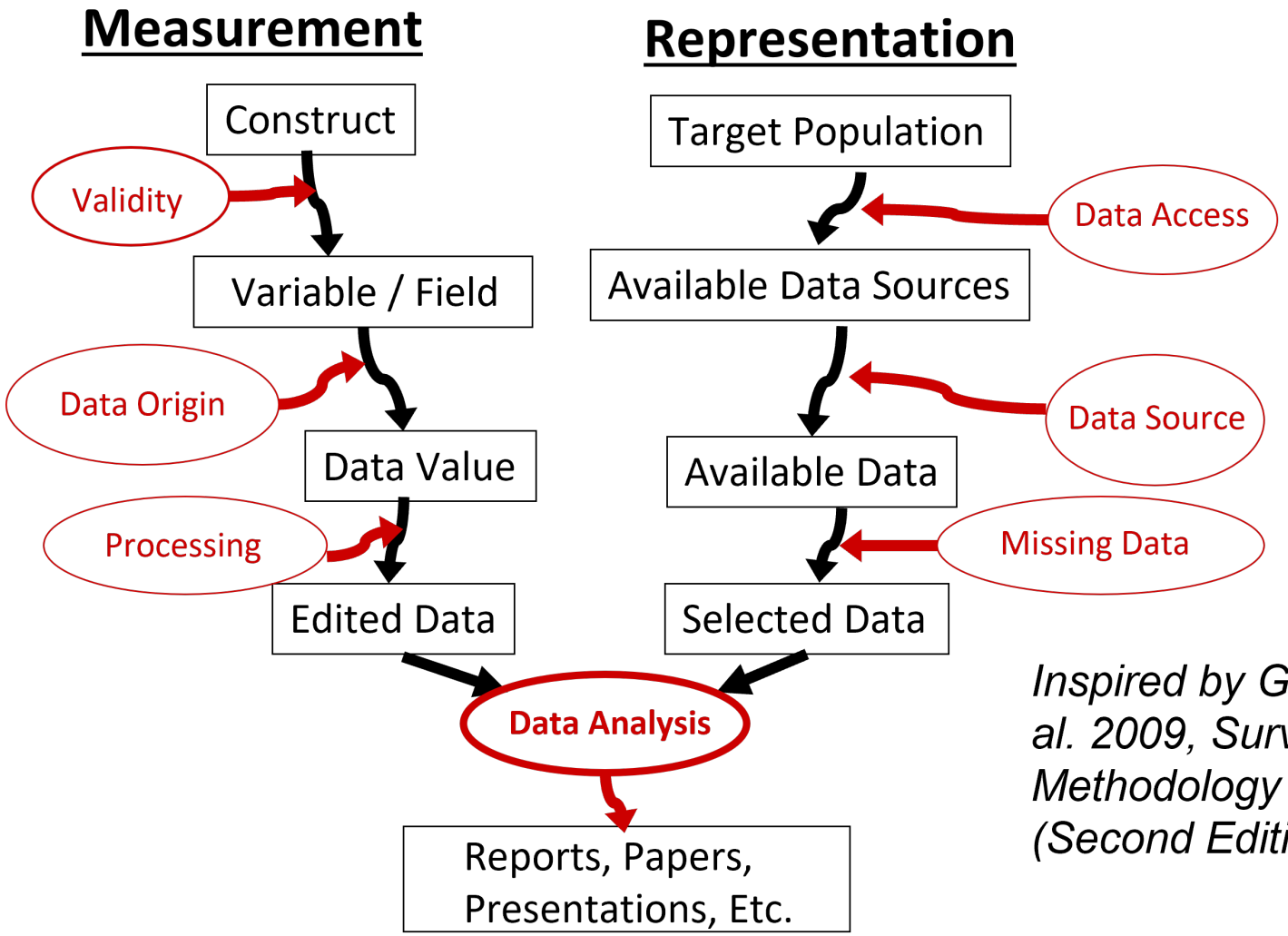
Threats Concerning Data Analysis for Gathered Data

By Trent D. Buskirk

Dimensions of TDQ



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

Data Analysis Focus

- We focus on the **quality** of the data analysis (specifically for gathered data)!
- Have appropriate models been specified for the types of variables being analyzed (valid for both types of data)?
- Are there biases inherent to the algorithms being used?

Threats Related to Data Analysis for Gathered Data (1)

- Incorrect model specification
- Incorrect use of adjustment procedures
- Algorithmic bias
- Failure to include relevant variables in an algorithm
- Possibility of overfitting (i.e., model too sensitive to training data)
- Inability to use the same version of a proprietary algorithm from software packages that are not open source
- Failure to completely specify the correct number of random seeds or to specify them correctly in the analysis may risk lack of reproducibility of the exact results

Threats Related to Data Analysis for Gathered Data (2)

- Incorrect use of a variable – some vendor variables come as ranges that can be mistaken for continuous variables rather than nominal.
- Incorrect interpretation of results of a model – especially from machine learning methods – so variable importance is not the same as statistical significance.
- There are no beta coefficients in many ML models!

Threats Related to Data Analysis for Gathered Data (3)

- Variable selection techniques may be improperly applied leading to model misspecification.
  - This is especially important when using large gathered data that may have many correlated variables.
- Typical variable importance measures can be biased for identification of important predictors whenever they are highly correlated.

Threats Related to Data Analysis for Gathered Data (4)

- Moreover, some methods for analyzing large data sets perform badly when the data are of mixed type (e.g. categorical versus continuous variables) which is important for survey-related applications
- Near Zero Variance variables should also be eliminated prior to modelling
- Without validation methods, many models may overfit the data and be overly complex to be reasonably applicable to external data sources about which predictions are desired.

Threats Related to Data Analysis for Gathered Data (5)

- Other issues with **feature engineering** - applying transformations on some variables may assume they are of one type when they are actually not (e.g. gathered income data from a vendor may represent income class rather than actual income level).

- The scale of some gathered data may not be gathered as part of the process and this may impede interpretation or clarity around the type of transformation or other feature engineering that would be reasonable.

Threats Related to Data Analysis for Gathered Data (6)

- **Madigan et al. (2014)** explore the impact of choices made in the data analysis process on observational studies based on medical records.
- They conclude that the current *ad hoc* investigator-based decisions lead to inflated p-values and incorrect conclusions
  - They suggest data-driven procedures as an alternative.

Threats Related to Data Analysis for Gathered Data (7)

- **Diesner (2015)** discusses the impact of choices of assumption and methods on data analysis in big data analytics.

- Exploratory analysis where one might visualize relationships among predictors using scatterplots, for example, may prove inconclusive when working with larger gathered data at scale (we get the “scatterblob”).

  - Hexbin plots may be an alternative for this.

- Statistical tests of correlation and association too reduce the dimensionality of a data set may also fail at scale because p-values will likely be mostly small.

What's Next?

- Next up, we will present a case-study that explores ALGORITHMIC BIAS in facial recognition algorithms that use image data to make classifications of gender and race.

UNIVERSITY OF MICHIGAN