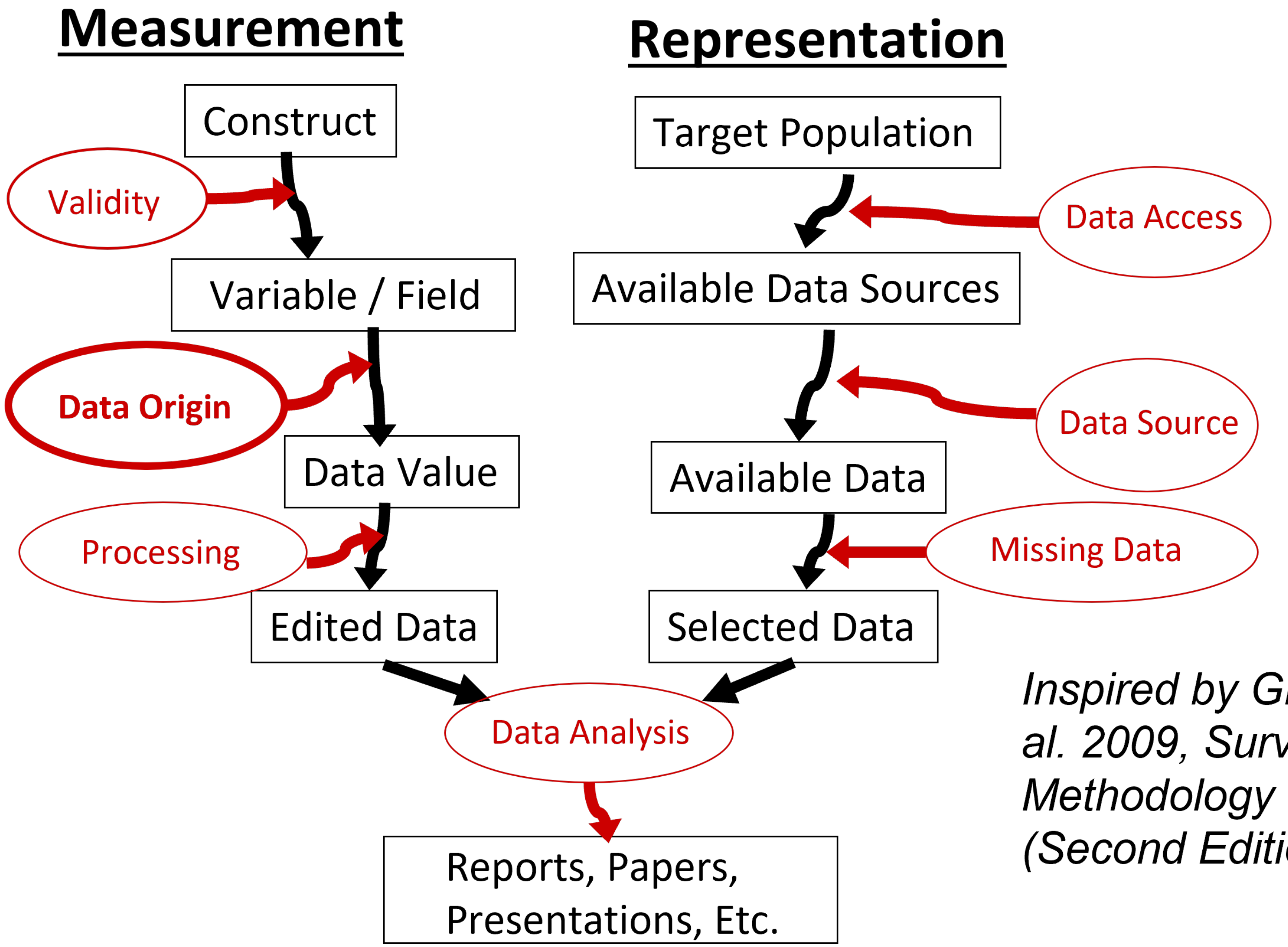
Data Origin Threats for Gathered Data
By Trent D. Buskirk

Data Origin Definition

- **Data Origin:** How were the individual values / data points for a given variable (or field) recorded, captured, gathered, computed or represented?
- Were there any errors in the process that ultimately produced the value for the variable?
- **Designed Data:** interviewer effects, social desirability, reporting error, etc., in survey responses
- **Gathered Data:** issues with the process of creating or retrieving the values for fields of interest

Dimensions of TDQ



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

Threats Concerning Data Origin for Gathered Data (1)

- Data Related Threats:
  - Incorrect data aggregation / variable computation
  - Reporting or recording errors in administrative data
  - Mixing of units of measurement: e.g., one scrapes Zillow and gets lot size data, but any unit less than **1** is in ***acres***, when most lots are measured in ***square feet***

Threats Concerning Data Origin for Gathered Data (2)

- **Technology/Processing Related Threats**
  - Instrumentation error
  - Variance in technology as it relates to recorded or tracked information, such as location
  - Annotator / Translator / Transcriber error when working with organic qualitative data

Additional Manifestations of Errors related to Data Origin for Gathered Data

- Errors in consistency resulting from lack of metadata regarding variable definitions or meanings of recorded information.

- Errors in consistency when combining information across multiple sources (e.g., county-level test results across multiple health systems using different diagnostic tests or screening criteria).

- Temporal validity issues related to data recorded at different time points for various fields.

- Data integrity issues – representation of information is not consistent over time or across sources.

Threats Concerning Data Origin for Gathered Data: Data Representation

- In the Gathered or Big Data space, “representation” takes a different meaning than in the Designed Data space.
- For gathered data, much effort is placed on database design and aspects of representation refer to how data are ***formatted and stored in databases***.
  - Are all data represented as text/ASCII/UTF?
  - Are numbers stored as bytes and character strings in ASCII?
  - Are dates stored without formatting? If not, what kind of formatting?

Data Representation

- Data representation errors related to encodings can occur
  - For example: if data are stored with an ASCII representation and are then sourced using UTF.
- Date storage can also pose problems if such data are represented and stored in one database as a number (computed as days since a reference data). When this variable is analyzed using another software platform, the field is treated as numeric without any metadata or formatting to suggest it actually is a time related variable.
  - However, from a data storage and compression perspective, storing a data as numeric with a tag for time requires less memory/space compared to storing the date in long format (June 12, 2020, or 06/12/2020).

Examples: Data Origin Threats for Gathered Data (2)

- Wouters et al. (2017) look at how cost estimates at a product level are difficult to measure for a variety of reasons, due to human errors in timesheets and machine errors.
- Understanding the data generating process was critical to working with the data!
  - For example, in vendor related reports of square footage of housing units, the values are presented as numeric, but are rounded to the nearest 100.
  - So the distribution of this type of data will indicate a "clumping" effect similar to what we might see in surveys, with a slider bar only being ticked on whole multiples of 5 or 10.
  - Knowing that this measurement is actually rounded *a priori*, one could incorporate this information into comparisons across data types by applying a consistent transformation to designed data.

Examples: Data Origin Threats for Gathered Data (3)

- Misra et al. (2016) report on how "noisy human labels" create problems for models aimed at classifying images, which is essentially a rater reliability/validity issue.
- There is also recent evidence of annotator bias in studies of crowdsourcing for natural language processing: as shown in this article
- This is very similar to the problem with interviewer effects in designed survey data using human interviewers.

What’s Next?

- In the next segment of the course, we are asking you to review a few articles in the popular press that illustrate threats related to Data Origin for COVID-19 related testing and tracking that have been gathered by various state and local health departments.

- While you are reading these posts from various popular press sources, we ask that you think about:
  - What was the error made?
  - Why would you consider it an error related to data origin?
  - What implications do you think an unresolved error like this can have or did have in these examples?

UNIVERSITY OF MICHIGAN