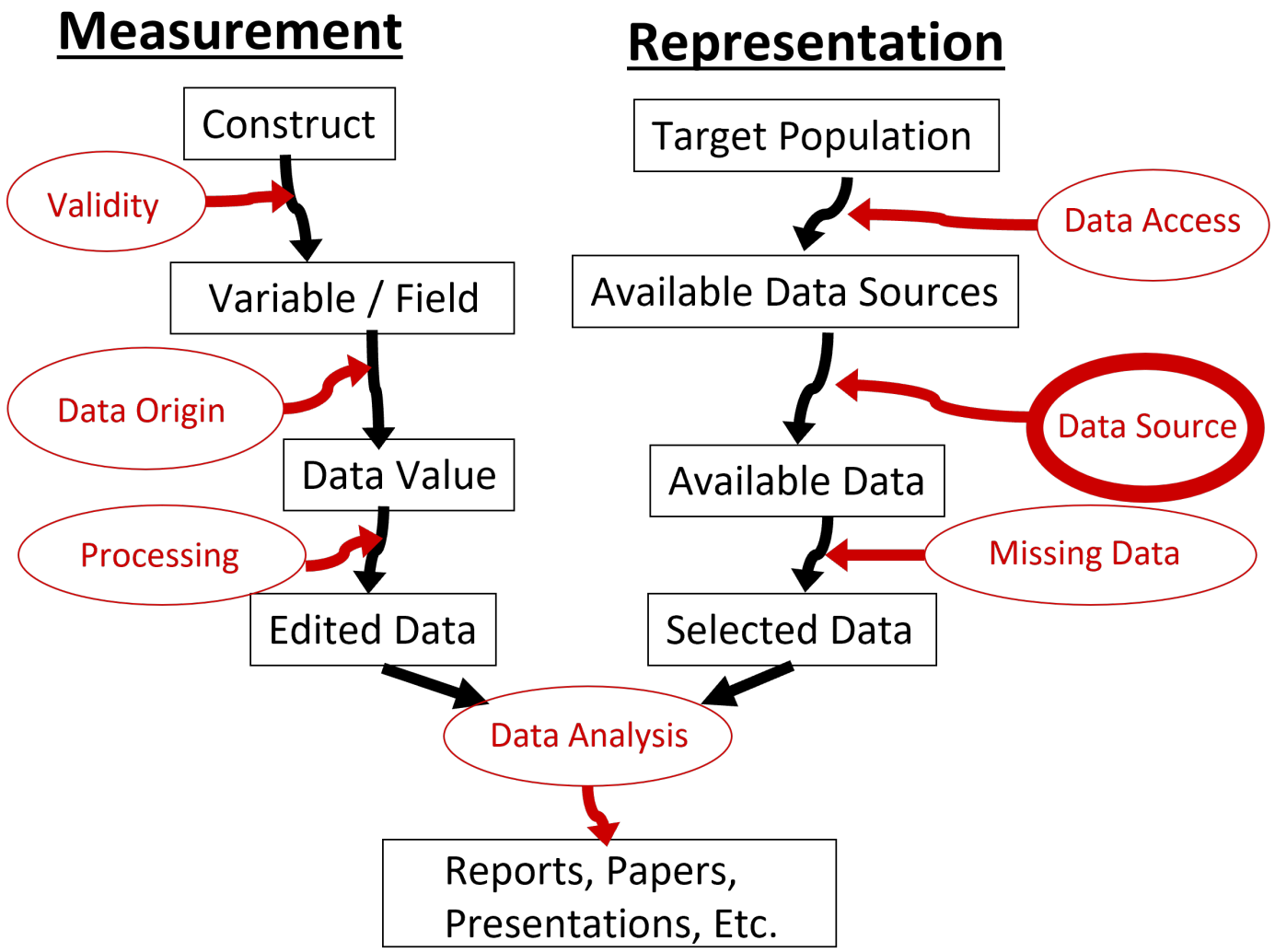
Data Source Threats for Gathered Data

By Trent D. Buskirk

Dimensions of TDQ



Data Source Definition

- The source of the data in designed data is a **selected sample** from a sampling frame.
- For gathered data, the source refers to both the collection of information **actually extracted** (or **selected**) and from which platform/repository the data was accessed.
    - For example, a sample of Tweets from the Full Corpus of Twitter.

Threats to Quality Related to Data Sources for Gathered Data: Algorithm Dynamics (1)

- Algorithm dynamics are the changes made by engineers to improve the commercial service and by consumers in using that service
  - If companies make changes to their platform(s) to improve services for customers, they may change the data-generating process.
    - Google modifies their search algorithm to improve new ISO methods.
    - Recommended searches may then change and potentially alter the relative magnitude of other searchers.

Threats to Quality Related to Data Sources for Gathered Data: Algorithm Dynamics (2)

- The changes services make over time to their own algorithms may create issues with reproducibility and repeatability (**Lazer et al. 2014**).
  - Google search results from “Covid-19” performed over the time period May 1, 2020 – May 31, 2020 on June 10, 2020 may not be the same if the same exact search is performed again on June 10, 2021, even if the same time period is requested.
  - Twitter data changes over time as public accounts go private, in which case all associated tweets are no longer represented in the corpus accessed using Twitter Search APIs.

Threats to Quality Related to Data Sources for Gathered Data: Proxy Population Mismatch

- Various sources curate information from distinct population groups (**Ruths and Pfeffer, 2014**)
- Proxy effects – Populations identified and studied based on self-reported identification may not be similar to the broader population.
  - Republicans who self identify on Twitter or Facebook may not be like other republicans in the population who do not identify themselves as such either on twitter or otherwise.
- Often assembled populations cannot be shared or archived and then reevaluated by other researchers as per terms of use.
  - UPDATE! Twitter just announced improved data release options for academic researchers via its Search API version 2.0.
    More information can be found at this webpage

Threats to Quality Related to Data Sources for Gathered Data: Population Biases

- Populations may not all be the same across Gathered Data Sources (Ruths and Pfeffer, 2014).
  - Instagram seems to be appealing to adults aged 18-29, African American, Latinos, women and Urban Residents.
  - Pinterest users seem to skew female between the ages of 25-34 with an average annual income of $100000.
  - Among Americans 65 and older, 46% use Facebook.
  - Approximately 66% of U.S. adults in rural regions use Facebook followed by YouTube with 64%, then Pinterest with about 26%.

Facebook Info Source

Threats to Quality Related to Data Sources for Gathered Data : Platform Limits

- Some data sources/gathering tools place limits on the amount of information that is available per user or that can be retrieved from APIs. And this can vary by platform and API tier (e.g. free versus paid).
- When retrieving tweets from Twitter for example, there are monthly caps on the total number of tweets that can be gathered and stored, even for those with access to the full firehose.
- Twitter also tethers the number of API calls that can be issued from a single user within specified units of time (e.g. no more than 180 API requests per 15 minute interval for the free search API).
- Twitter currently limits the number of tweets archived per twitter user to 3200.
- Web scraping platforms also limit the number of scraped pages per month as part of a tiered system.

Threats to Quality Related to Data Sources for Gathered Data: Technology

- The automation of collection can vary by source as can the orientation of mechanisms used for data collection (Vial, 2019).
  - Different machine learning algorithms might collect sound information to detect programming of a television show or song, but the sensitivity of these methods can be a function of whether an app was used, or a stand-alone meter placed in the house, for example (Nielsen Arbitron People Meters).
  - Orienting a sensor vertically or horizontally on a person or bicycle for tracking activities may impact the nature of the data collected.
  - GPS location can vary in accuracy depending on type of signal: Wi-Fi, Bluetooth or cellular networks, for example.

What’s Next?

- We will dive a bit deeper into how content characteristics may vary by user.
- We will also explore how platform limits imposed by some popular Social Media Data sites may impact quality for studies using these sources of gathered data.

UNIVERSITY OF MICHIGAN