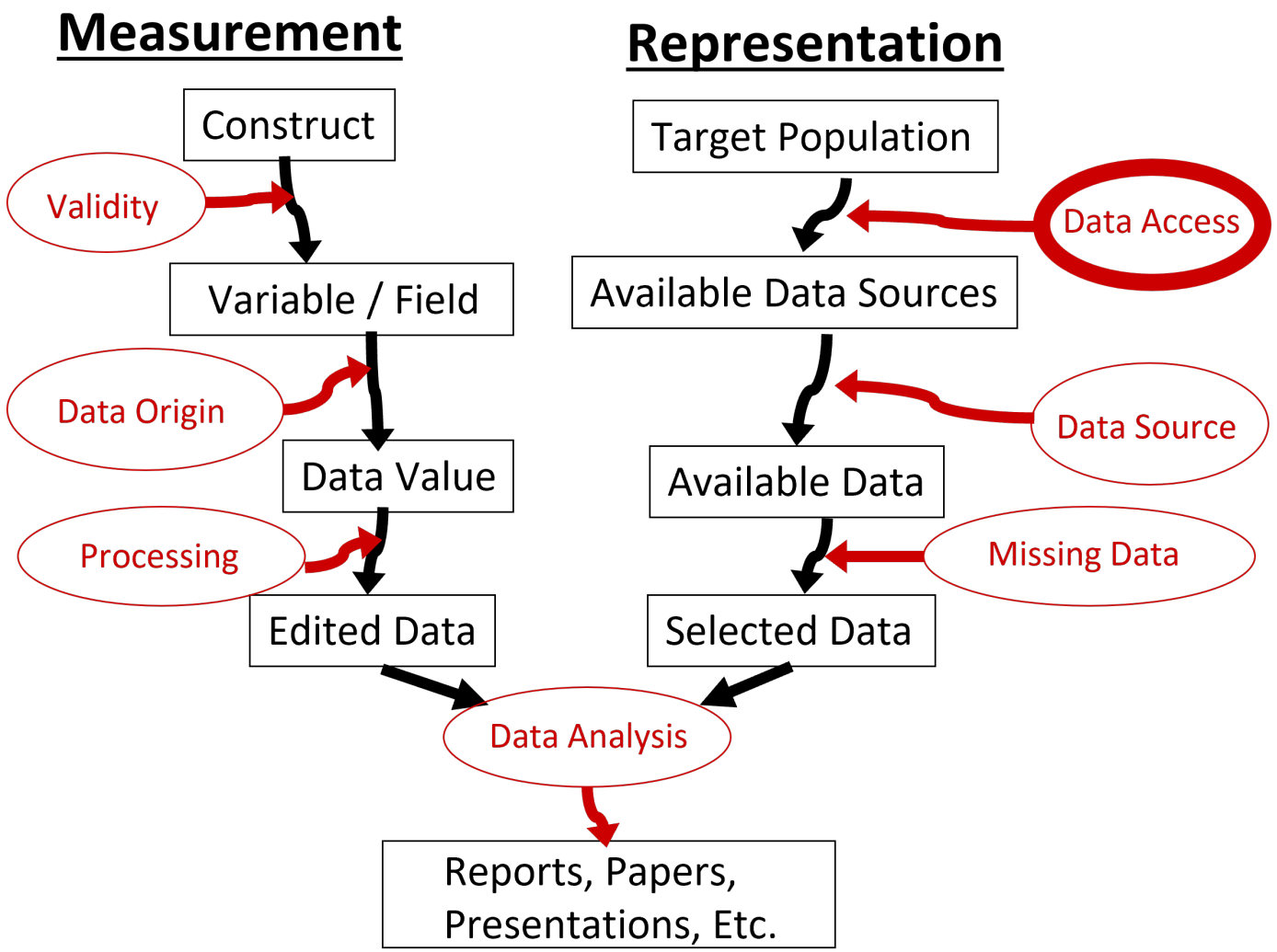
Data Access Threats for Gathered Data | Part 2

By Trent D. Buskirk

Data Access Definition

- **Data Access** refers to the methods or approaches or tools used to retrieve information from various sources.
- For gathered data, data access can be granted by using:
  - Web Scraping
  - Application Programming Interfaces (APIs)
  - Vendors and upstream data sources; apps, sensors and the like, among others…

Dimensions of TDQ



Threats to Quality Related to Data Access: APIs

- Platforms and their corresponding APIs set terms of use and declarations about the underlying quality of data that are often “as is.”
  - “Zillow provides the Zillow API, Zillow data, and Zillow brand & links "as is," "with all faults" and "as available," and the entire risk as to satisfactory quality, performance, accuracy, and effort is with you.”
- The main issue is it is often difficult to improve the underlying quality of the entire corpus of data directly.

Examples of Threats to Quality Related to Data Access (1)

- **González-Bailón et al. (2012)** concluded that there is a strong indication that Twitter returns tweets from those users who are more centrally located within the Twitter network of the Search API, compared to the Streaming API.
- Twitter data collected for the same set of hashtags can be different depending on the access choices of Twitter Search, Streaming API, Firehose (**Kim et al., 2020; Morstatter et al., 2013**)
  - Top 100 hashtags (majority overlapped) vs. Random 100 hashtags (most overlapped).

Examples of Threats to Quality Related to Data Access (2)

- **Morstatter et al. (2013)** comment that the amount of detail and the availability of documentation regarding how the API filters the database source can vary by platform.
  - For example:
    - Zillow.com has three different APIs for accessing real-estate information but each API has select fields that are available.
    - Twitter has various APIs that are free and paid and the fields/operators available for each of these differ.
    - The documentation can vary in availability for APIs and in the level of technical jargon/plain English that is used to explain their use.

Threats to Quality Related to Data Access (1)

- A growing body of research and thought pieces in the social sciences are raising concerns about a new digital divide that is brewing around gathered data and its access
  - **Boyd and Crawford (2012)** discuss differential access to social media platforms. Those with full access are in the best positions to understand relationships, limitations and sizes of populations. this situation leads to a new digital divide separating Big Data Rich from Big Data Poor.

Threats to Quality Related to Data Access (2)

- Researchers who are granted access to these platforms (either by affiliation or by paid memberships) may be able to retrieve more complete information that is not available publicly.

- Those with access are referred to as “embedded researchers” (**Ruths and Pfeffer, 2014**). These researchers may not be able to explore research not approved by the data owner or may be restricted in their ability to share their actual data sources making reproducibility challenging.

Threats to Data Access for Gathered Data: Platform Limits for the Twitter API (1)

- One very clear example of retrieval limits for the Twitter Platform is in the number of queries one can make to the Twitter Search API.
- Twitter calls these rate limits and they govern the amount of information that can be obtained per account or user within a given time period.
  - For example, using the Twitter Search API to retrieve tweets, there is a limit of a total of 450 query requests within a 15 minute time interval.
  - Limits also vary by type of information being retrieved with different rates being specified for queries seeking user information versus tweets, for example.

Threats to Data Access for Gathered Data: Platform Limits for the Twitter API (2)

- Twitter also limits the number of tweets archived per user to 3200.
  - This limit may impact the completeness of a data set or the expected sample yield.
  - This limit may impact the choice of access method one needs for a given study to maximize yield
  - This limit may also impact the quality of a given study in terms of its reproducibility

Threats to Data Access for Gathered Data: Platform Limits for the Twitter API

Cumulative Number of Tweets for user “GoBlue”

- 2000 by 2015
- 3200 by 2016
- 4800 by 2017
- 5200 by 2018
- 6400 by 2019

The most recent 3200 Tweets are Available for “GoBlue” are:

- 4800 by 2017
- 5200 by 2018
- 6400 by 2019

Study Seeks to Sample “GoBlue” and query this user profile for tweets posted in 2016. Because “GoBlue” has posted more than 3200 tweets after 2016 the query will not return any relevant tweets for this user for your study.

What's Next?

- Next up is a demonstration of accessing Twitter data using a search API. We have prepared an example Jupyter notebook to show you how you might access Twitter data using the Search API in R.

- To use the API yourself you will need to request and obtain a "token" and "key" from Twitter. We will discuss this more in depth in the demo coming up!

UNIVERSITY OF MICHIGAN