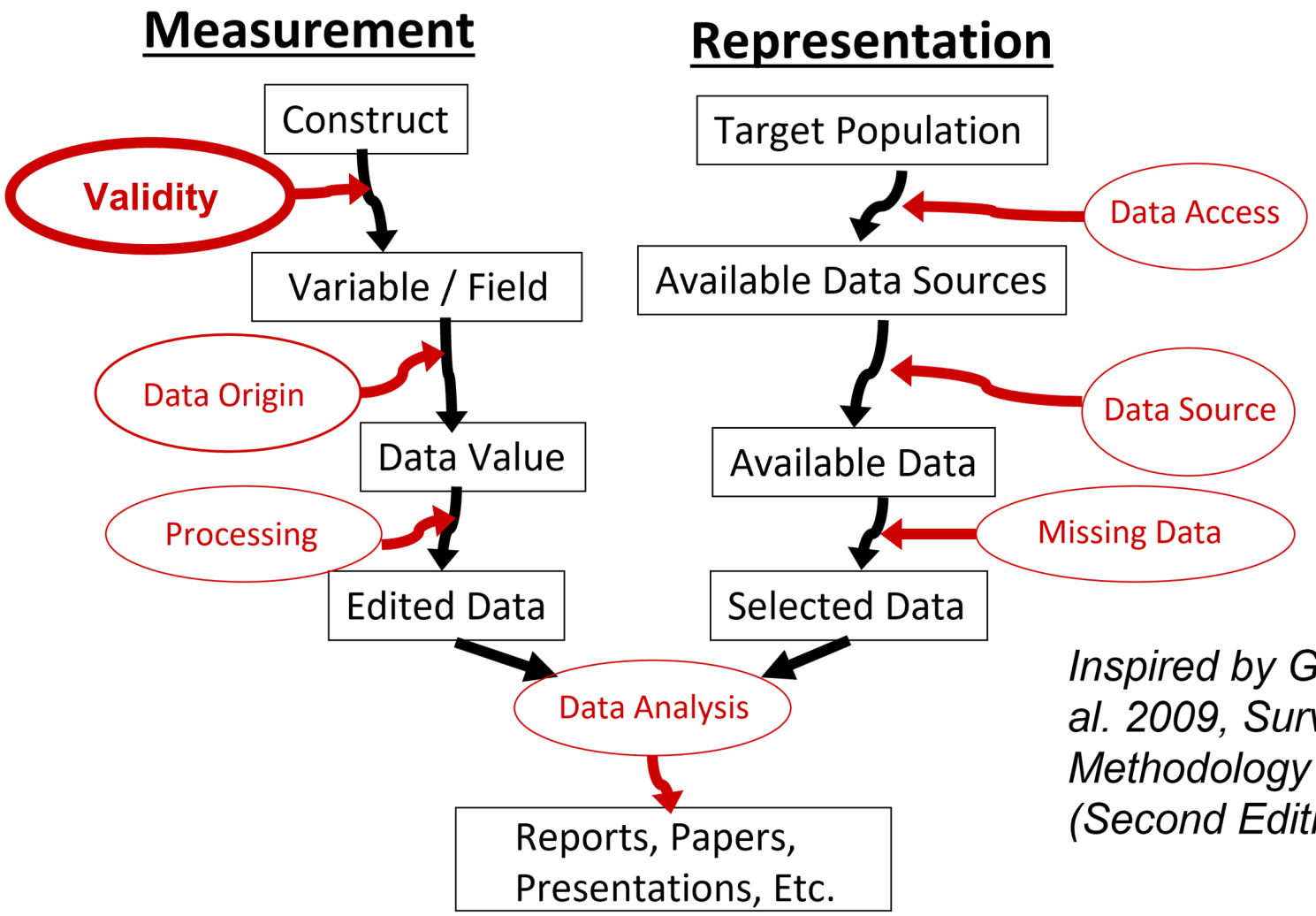
Threats to Validity for Gathered Data

By Trent D. Buskirk

Dimensions of TDQ: The Big Picture!



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

Threats to Validity for Gathered Data: The Costs in Context

- Erroneous Data costs US businesses over 600 billion dollars annually.
  - **These costs represent somewhere between 8-12% of annual revenue**
  - 40-60% of a service organization's expenses may be consumed as a result of poor data quality (Redman, 1998)
- Organizations typically find data error rates of between 1% and 5% but can be above 30% for some others (**Saha and Srivastava, 2014**).
  - Total errors in fields/All possible Fields

Threats to Validity for Gathered Data (1)

- Platform or Data Source dynamics and structure may limit the accurate reflection of human behavior (**Ruths and Pfeffer, 2014**).
  - Platform designers improve user experience based on key concepts:
    - Homophily (“birds of a feather”),
    - Transitivity (“a friend of a friend is a friend”)
    - Propinquity (“those close by form a tie”)

Threats to Validity for Gathered Data (2)

- Optimal user experience may not result in accurate measures based on data gathered from these sources.
  - For example, following users on Twitter may not be an adequate measure of true network size if the "following" is in one direction.
  - Correlation may be higher among “friend samples” recruited from Facebook because of recommendations made based on concepts described above.

Threats to Validity for Gathered Data (3)

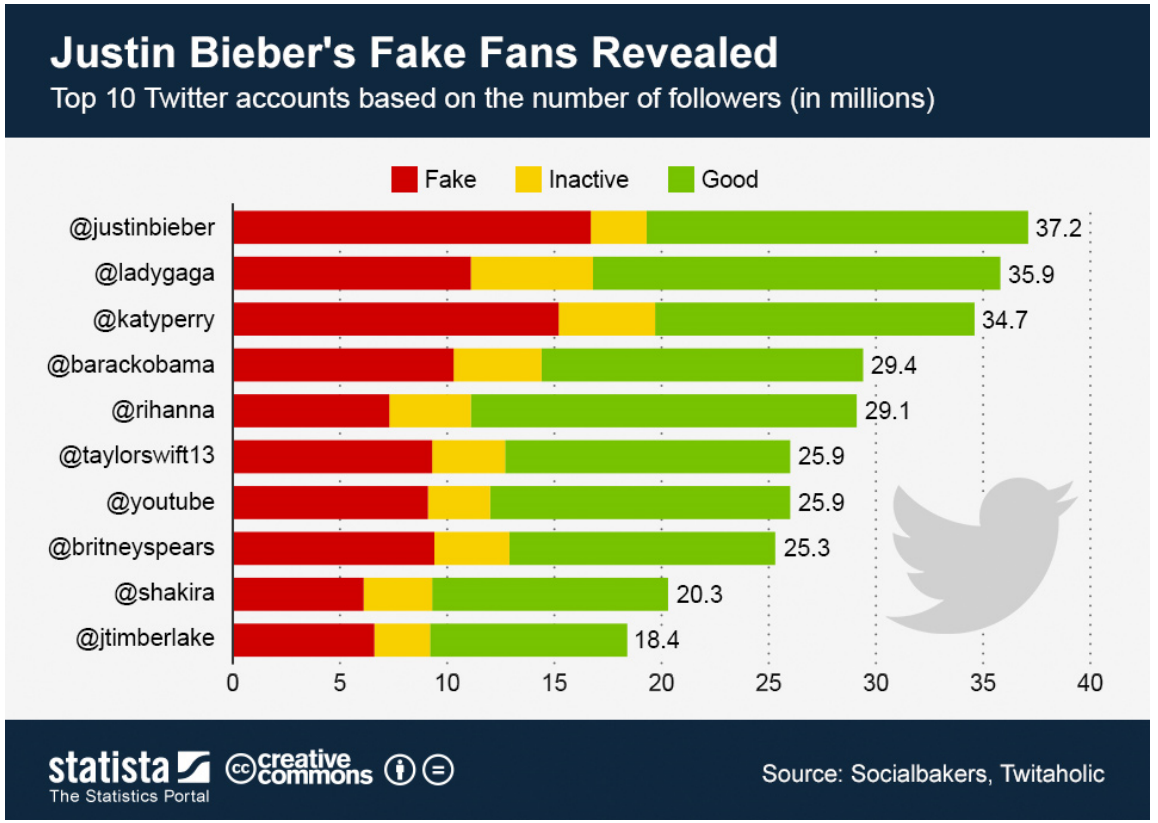
- Platform Technical Specifications and Processes may create Distortions in Measurement of Human Behavior (**Ruths and Pfeffer, 2014**).
  - Only the most recent 3200 tweets are shown in public accounts when a specific username is queried.
  - Google stores and reports final searches submitted *after auto-completion is done* as opposed to the actual text that was typed.
  - Twitter dismantles retweet chains back to the original user who posted the tweet.

Threats to Validity for Gathered Data (4)

- Gathered Data, even from Human-Oriented Platforms, can contain NON-HUMAN results (Ruths and Pfeffer, 2014).
  - Fake user profiles and Bots exist on virtually every platform.
    - Varol et al. (2017) estimated that between 9% and 15% of active Twitter accounts are bots
  - Platforms contain a mix of personal and business/organizational level users.
  - In 2018 Twitter released a new review policy for bot accounts that sought to limit the number on bad actors on their platform.

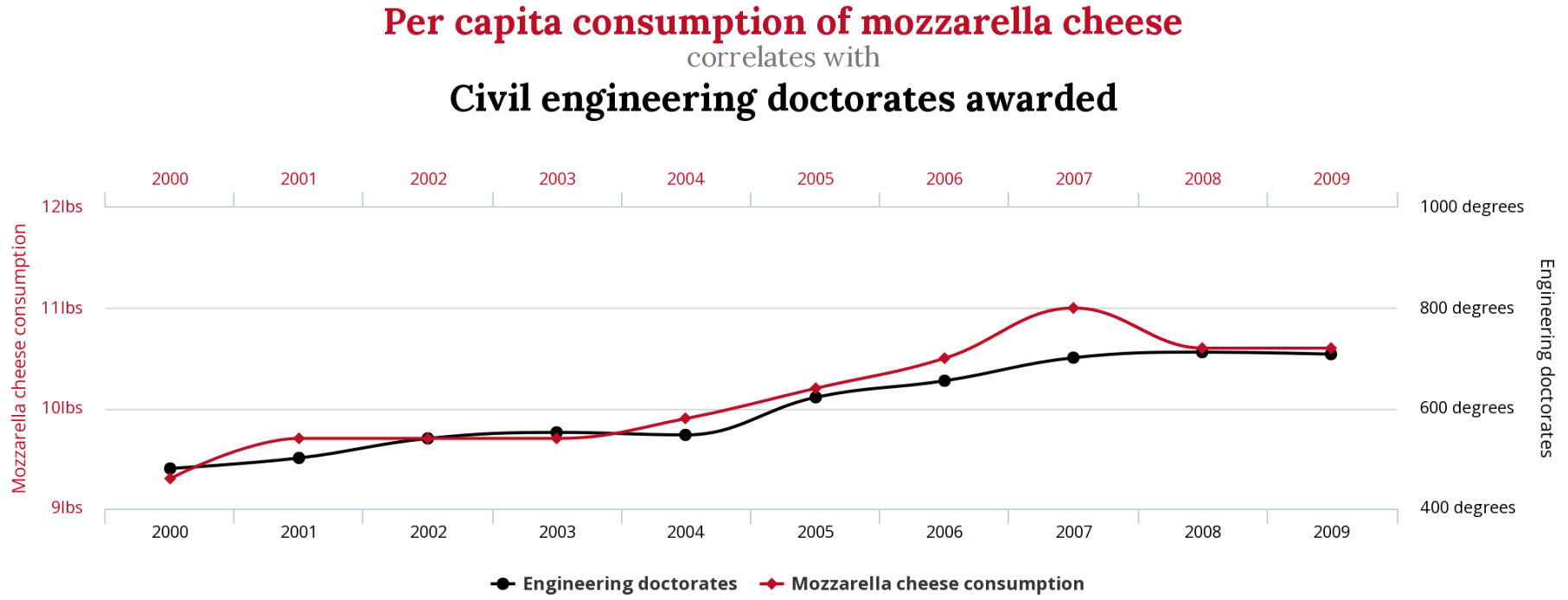
Examples of Threats to Validity for Gathered Data

- Is the number of Twitter followers a **valid** measure of interest / popularity / support / engagement?



Spurious Correlation is Real Threat to Validity of Gathered Data

- Correlation can be Spurious and it is not Causation!
  - Per capita consumption of Mozzarella Cheese (US) correlates positively with the Number of Civil Engineering Doctorates Awarded (US) (r=0.96)



What's Next?

- We will explore a rather historic example of threats to gathered data in the Google Flu Trends case study coming up next.
- In that case study we ask you to read two articles that show how the google flu trends tool worked for a while for predicting flu like illnesses until it stopped working.
- We then ask you to read a another article that discusses ways in which the google flu trends tool could be improved by combining its information with survey data.
- Leveraging multiple data sources is one of the ways we will explore later on for overcoming threats to gathered data.

UNIVERSITY OF MICHIGAN