What are Gathered Data?
By Trent D. Buskirk

What are Gathered Data?

- **Gathered data:**
  - Data obtained from an existing source most likely not designed for research purposes or distribution as such.
  - Data generated from some external procedure, tool, meter or sensor
  - Often referred to as “organic data” or “Big Data”
  - Social Media Data (e.g. Twitter); Sensor Data (e.g. Bluetooth wearable); Web-scraped or Open Data (e.g. Zillow.com)

Big Data as a Process

Big Data Generation	**Big Data Management**	**Big Data Analytics**
Data Origin: • IoT Sensors • Social Media • Administrative • Open Data • Other	**Data Acquisition:** • Data Collection • Recording	**Modeling:** • Feature Engineering • Feature Reduction • Model Construction • Cross Validation
Data Structure: • Structured • Unstructured • Semi-Structured	**Data Processing:** • Extraction • Cleaning • Transformation • Annotation • Curation	**Analysis:** • Visualization • Hypothesis Testing • Feature Importance • Evaluation
Data Attributes: • Volume • Velocity • Variety • Veracity • And others **Data Elements:** • Metadata • User Tags • Log files • Other	**Data Access & Storage:** • Integration • Record Linkage • Aggregation • Representation • Load • Privacy	**Interpretation:** • Inference • Insight Extraction

Gathered Data… or Not? (1)

- **Public Use Survey Data Files**
  - These records are publicly available and often housed and accessed online
    - For example, Current Population monthly supplement survey data can be accessed and downloaded from census.gov
    - These data were designed for research or estimation purposes and while possibly a secondary data source for your research these data were designed and specific variables were intentionally included and processed for release in the public data files.
    - These data would not fall under the Gathered Data category.

Gathered Data… or Not? (2)

- **Administrative Records**
  - While data designs can leverage information housed within administrative records, more often than not these types of data files, while structured, are broadly considered a type of "Big Data."
    - Housing Sales data for a given county are often available from a county assessor's office either in tabular form that can be scraped or as a tabular file that can be downloaded or through an Application Program Interface.
    - While these data are structured they were likely not created or generated for research purposes *a priori*.
    - There is considerable variety in not only the types of administrative records that are available but also in how they can be accessed.

Gathered Data… or Not? (3)

- Application Programming Interfaces (APIs) allow users to more easily access web data stored, curated or created from online venues.
- Generally speaking data accessed through an API call would be considered gathered data
  - Unless the data source itself was presenting designed data (such as a call to generate tables created from the current population survey’s online survey tabulation tool).
- APIs and Web-scraping from the same website may not produce equivalent data and we explore this in an upcoming lecture.

Gathered Data: Summary

- Gathered data
  - Generated for a purpose other than analysis
  - Sourced from sensors, meters, web pages, social media, transactions, digital footprints and the like.
  - Data are opportunistically used for research or estimation purposes and are likely not the intended purpose of their generation.
  - While gathered data are not designed for research they can present viable auxiliary sources of information for investigating issues of interest to researchers and practitioners

What’s Next?

- In the next lecture, you will have a chance to **learn how to scrape data from a simple table presented online.**

Readings

- Ward and Barker (2013) on Big Data Definitions

UNIVERSITY OF MICHIGAN

© Faculty Presenter
Except where otherwise noted, this work is licensed under CC BY-NC 4.0