



Big Data - Big Deal or Bigger Deal Breaker: Sifting through the hype to uncover quality and conundrums with modern data sources

WAPOR Webinar
February 24, 2023

Trent D. Buskirk, Bowling Green State University
Novak Family Professor of Data Science

Today...I would like to...

- Define Big data – what is it, where is it?
- Identify Possible validity issues related to Big Data
- Show some examples of how Big Data validity can be maximized
- Show other example of how Big Data can be combined with designed data (e.g. surveys or experiments) to improve estimation/designs
- Entertain Questions

Shout Outs...

- Collaborators have enriched my thinking on these issues and have been working with me on projects that have honed some of the content included here today:
 - Brady West
 - James Wagner
 - Jinseok Kim
 - Antje Kirchner
 - Frauke Kreuter
 - Adam Eck

The BIG data revolution...

“Whilst there may be a ‘big data revolution’ underway, it is not the size or quantity of these data that is revolutionary.

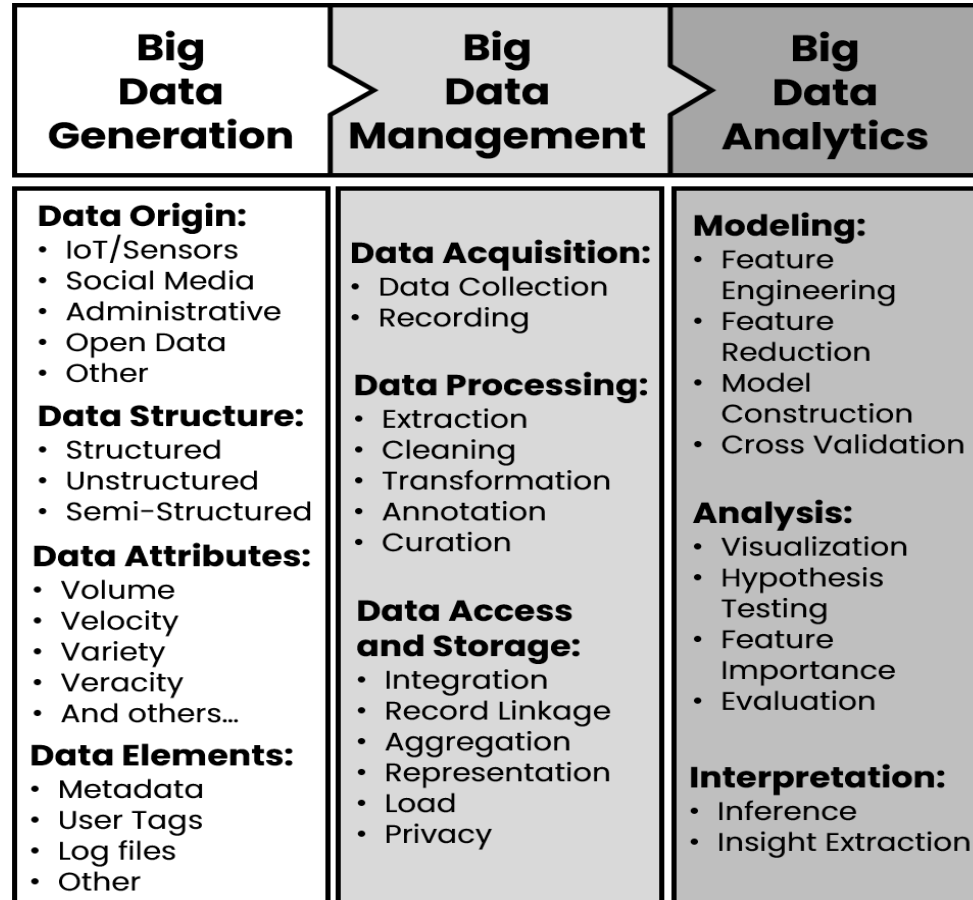
The revolution centers on the increased availability of new types of data which have not previously been available for social research.”

R. Connelly and Colleagues, 2016

<http://bit.ly/2k0V7GM>

Big Data

- Buskirk (2020) synthesized the work of Jagadish and colleagues (2014) and Curry (2016) to illustrate Big Data as a process moving from Data Generation to Management to Analysis
- Big Data, while likely designed and sourced at some point, for survey and social science researchers is more commonly thought of as “gathered” or “organic” rather than designed.



Potential for Big Data for Survey and Social Science Research

- Buskirk (2020) discusses the overall potential of Big Data in comparison to traditional survey data...
 - It may be all too easy for researchers to discount much of Big Data because it was not designed expressly for research purposes or because it lacks the rigor of experimental or designed observational data more traditionally used in social science research.
 - However, much of the appeal of big data as an ancillary data source for research is that there may be inherent value that can be mined from data that have yet to be used for research purposes.
 - For example, sensors can track daily activities of respondents and this passive approach might offer a viable alternative to traditional diaries for reducing measurement error.

Big Data – Long or Wide?

- A natural perspective within the Big Data paradigm is to think about Big Data in terms of the sheer number of CASES available.
- But in social science and modern survey research applications another perspective might not be how long, but rather how wide a data set can become.

Var 1

Var 2

Var 3

...

Var 1000

Var 1001

...

Var 10500

Case 1

Case 2

Case 3

...

Case
10,000

Case
10,001

...

Case
100,000

Case
100,001

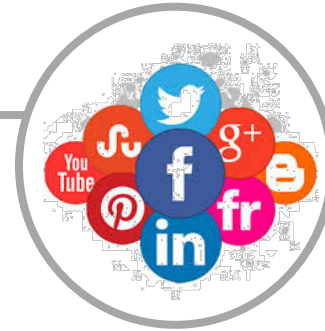
Types of Big Data...



Administrative and Open Data

**Administrative Records and Databases;
Population Registers**

**Open Access Data on Web (Tax Records;
Public Listings, etc.) and some image data**



Social Media Data

Facebook, Twitter, LinkedIn

Reddit

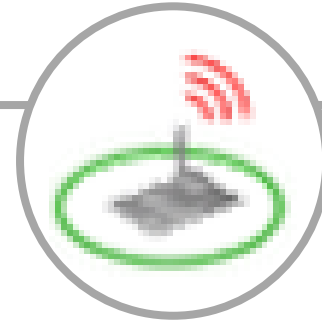
Instagram, YouTube

Types of Big Data...



Digital Trace Data

Online Transaction Data
Chat-bot generated content?
Internet Search Data
User Log File Data
Web Survey Paradata



Sensor Data

Bluetooth Wearable Sensor
Data (i.e. accelerometer)
Remote Sensors and Satellite Images
Household Sensors (i.e. Nest)



- Connelly and colleagues (2016) cautioned that Big Data should not just be synonymous with information collected through the Internet, noting that commercial transactions, medical records, and various other administrative data also fall within the big data umbrella;
- English et al. (2015) found an increased sensitivity for recruiting households with children under 5 from about 19% to just over 40% by combining Child flags from three Vendor Consumer Databases to create a more sensitive flag that results in a “Yes” if any one Database indicated Child Under 5.
- ten Bosch et al. (2018) describe how web scraping of open web data has been increasingly used in the production of official statistics including price statistics.
- Barcaroli, 2016 used web scraping to retrieve data about agritourism and used it to update and complete the national Farm Register in Italy.

Opportunities for Open Data and Survey Data Together



- Datta, Ugarte and Resnick (2020) explored linking data from the 2012 National Survey of Early Care (NSECE) and Education with commercially available proprietary real estate and property tax information from Zillow.com using the street address.
- The research team:
 - used each data source to evaluate the quality in each separate file along with the linked file.
 - leveraged property data available for both respondents and nonrespondents to assess possible patterns in nonresponse that would not have been possible with just the survey data.
 - offer a generalizable approach for investigating the quality of other data sources and illustrates how such investigation can inform and improve analyses using linked data.

Challenges with Administrative and Open Data...



- Administrative records may not be available for variables of interest or incomplete for others and these variables may be time sensitive and updated with different frequencies compared to field period needs
- Attributional error can occur with open and administrative data as not all variables are well described or defined.
- Open web data are volatile and can change in availability, format and content at times unforeseen by survey and social science researchers.
- APIs seem to provide more stable methods for accessing open data but these APIs are not ubiquitous across web sources and may not be as extensive (See West, Wagoner, Buskirk and Kim, 2020 for example).

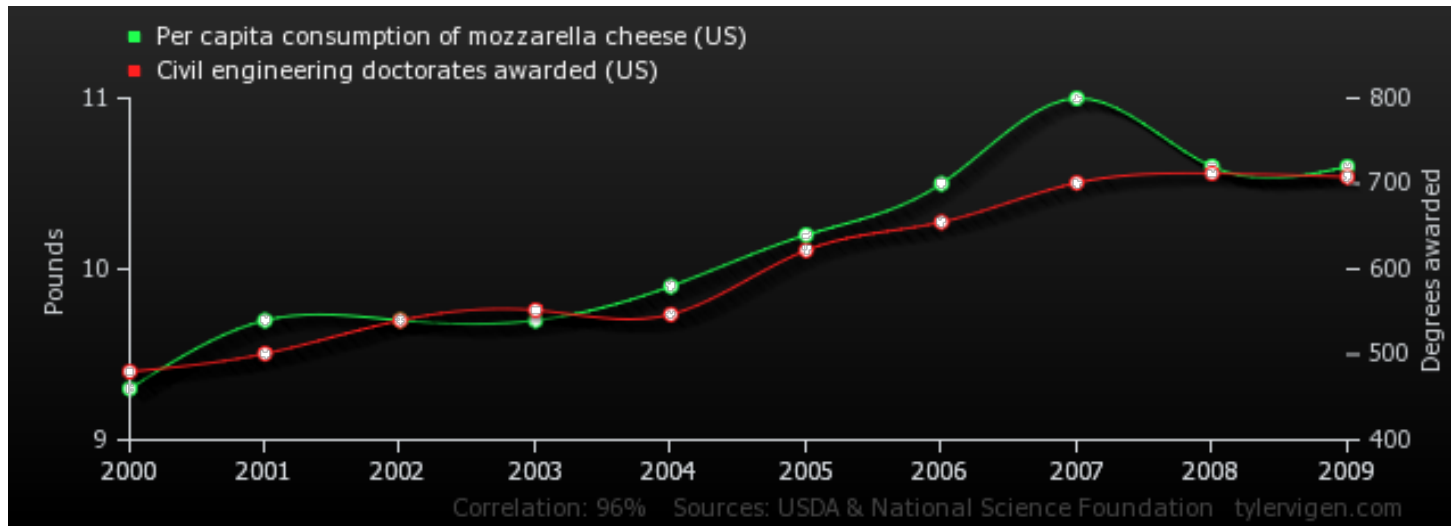
Spurious Correlation is Real Threat to Validity of Gathered Data

- Correlation can be **Spurious** and it is not **Causation!**



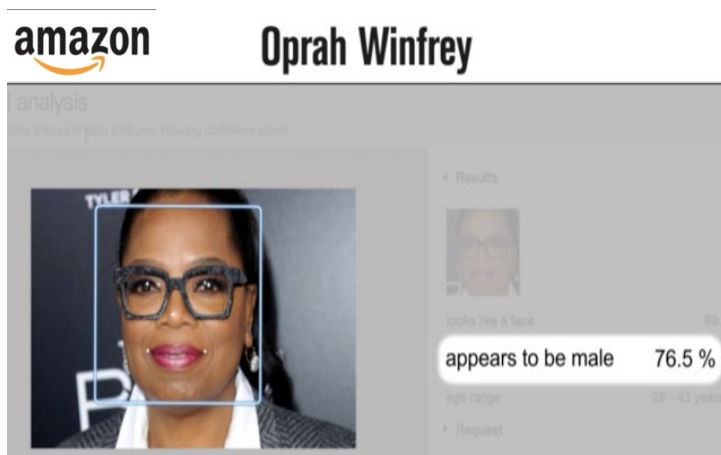
- Per capita consumption of Mozzarella Cheese (US) correlates positively with the Number of Civil Engineering Doctorates Awarded (US) ($r=0.96$)

Source: <http://www.tylervigen.com>



Example 1: Open Image Data...The Peril of Open Image Data?

- In a recent Time article, “Artificial Intelligence Has a Problem With Gender and Racial Bias. Here’s How to Solve It,” Joy Buolamwini (2019) report how facial recognition software used by Amazon, Microsoft and other large tech firms have little misclassification of gender for white users but this error soars to above 35% for darker skinned women, for example.
- <http://time.com/5520558/artificial-intelligence-racial-gender-bias/>



Example 1: Open Image Data: Unlocking the Power of Big Data

- Data representativeness criterion” (DRC), proposed by Schat et al. (2020) has been developed to assess how adequate a training data set is to prevent possible algorithmic biases related to omissions in training data.
- The work of Rolf et al. (2021) is rooted in the ideas of sub-group representation framed within a statistical sampling approach and presents some solutions for improving the representation of the training data.
- The work of Buskirk and Kern (2022, 2023) proposes to apply sample-based representation metrics from survey research to evaluate potential for biases in training data sets and relate these to fairness metrics.



- Burke-Garcia et al. (2018) explored how social media data create opportunities for not only sampling and recruiting specific populations but also for understanding a growing proportion of the population who are active on social media sites by mining the data that is present on such sites.
- Schneider and Harknett (2019) illustrate how targeted advertising on Facebook can be used to build an employer-employee matched data set where hundreds of employees at each of several specific companies are recruited and surveyed.
 - Provided a sampling frame (nested) that was not readily available and a fast and economical way to recruit respondents.
- Buskirk, Kreuter et al. (2023) explored how tweet content can be used to identify possible eligible respondents and tested different recruitment methods using Twitter's Direct Message capabilities.
 - Experiment was preregistered and more details can be found here: <https://osf.io/5fx9n>

Challenges with Social Media Data...



- While Social Media is popular, it is not ubiquitous. Pew (2018) estimates that about 70% of US adults use some social media. And not all users “use” at the same rates. In fact, most [Twitter] users rarely tweet, but the most prolific 10% create 80% of tweets from adult U.S. users (Pew, 2018).
- The entire population represented by a given data source may not be quantifiable or known at the time a sample is requested (e.g. # of twitter users constantly changing; unknown what the universe size of all tweets is without full access to twitter stream) (see: Lomborg and Bechmann, 2014 and Langer, 2014 (<https://abcnews.go.com/blogs/politics/2014/04/growing-doubts-about-big-data>)).
- Type of access to social media may impact the volume and content of samples. Kim, Nordgren and Emery (2020) note compiling different numbers of tweets across a 6-month field period. Streaming API collected 9% more tweets and the Search API 24% fewer tweets compared to the PowerTrack API which is the so-called “Full Fire Hose”.
- Recruitment methods using Social Media platforms are often black boxes with limited information available to the researcher on who sees and ad/invitation and who doesn't. And this selection process can change without notice, even within a field period (Schneider and Harknett, 2019).

Opportunities for Digital Trace Data...



- Stier et al. (2019) provide a synthesis of several ways digital trace data are being integrated with survey data to **improve substantive analyses**, **cross-validate and improve measurement** and **design survey experiments to generate useful digital trace data to directly measure treatment effects**.
- Moller et al. (2019) recruited a panel of participants to install a browser plug in on their devices that collected information about users' web activity for a series of media domains. This study allowed researchers to distinguish between several modes of online news use and found higher amounts of news use driven by internet searches.
- Wang and colleagues (2019) developed a comprehensive method for inferring demographic characteristics and used them in a weighting adjustment for deriving more accurate population based estimates based on digital trace data.
 - Use a Multi-lingual, multimodal, multi-attribute deep learning system for inferring demographics (age and gender and account type (pers. or business) based on profile picture, username, profile name and bio.

Challenges with Digital Trace Data...



- Typically digital trace data do not have accompanying information about the individual's behaviors or attributes making it more difficult for estimates of individual-level determinants of human behavior or public opinion to be based solely upon them (Stier et al., 2019).
- Studies employing digital trace data may have difficulty with replicability citing the rapid pace of changes, functionality and policies and privacy practices employed by SNS sites (Wells and Thorson (2015), Buskirk et al. (2010) and Dove (2015)).
- To date, research reports varying degrees of success for predicting demographics based on digital trace data (Hinds and Joinson (2018)) with Gender being most accurately predicted demographic (i.e. up to 80% accuracy) followed by Age (i.e. up to 70% accuracy). Political Orientation however was among the most difficult attributes to infer (up to 27% accuracy).

Threats to Validity for Big Data...Social Media and DT

- 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”)
 - This propensity may impact measurement of a users “real” friend network or the strength of the ties within it.

Threats to Validity for Big Data...Social Media and DT

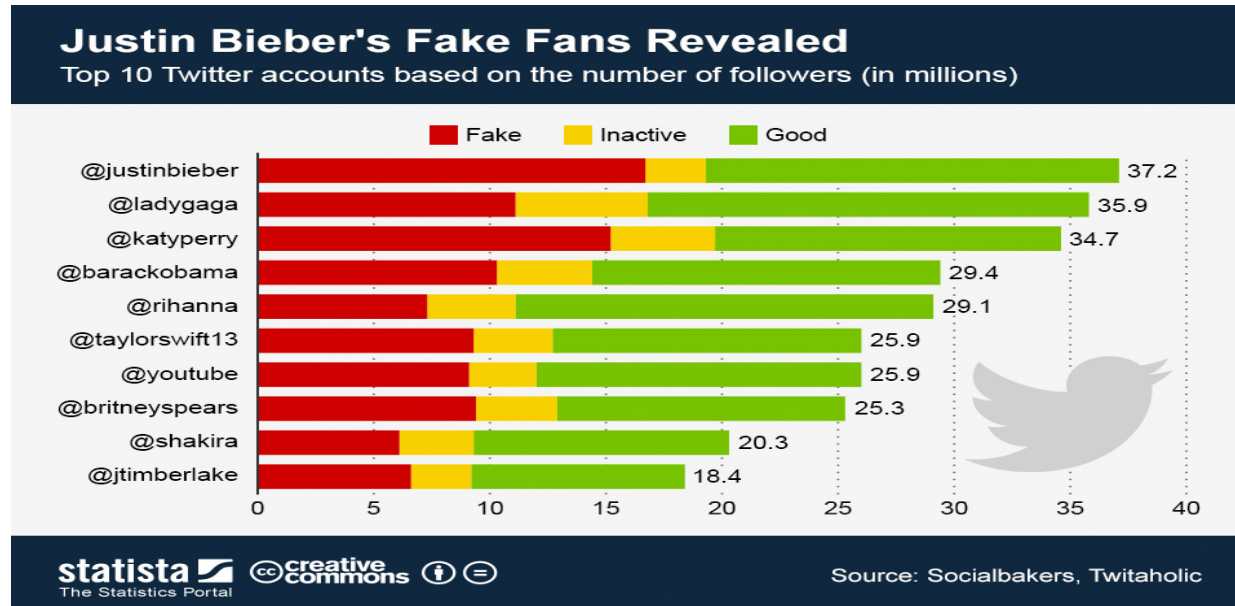
- Perceptions of the Platform and its culture may influence participation (Ruths and Pfeffer, 2014).
 - If users perceive Twitter as a place for political discourse and exchange, the posts made by users may be a more accurate reflection of their true opinions about politics.
 - These perceptions and platform cultural norms can change over time making threats for temporal validity when a study seeks to examine changes in opinion over time, for example.

Threats to Validity for Big Data...Social Media and DT

- 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
 - Posts may be created by humans –or– bots

Threats to Validity for Big Data...

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



<https://www.statista.com/chart/1031/top-10-twitter-accounts/>; <https://sparktoro.com/tools/fake-followers-audit>

Example 2: Google Flu Trends – The Promise of Big Data

- Ginsberg et al., 2009
- <http://dx.doi.org/10.1038/nature07634>

nature

Vol 457|19 February 2009|doi:10.1038/nature07634

LETTERS

Detecting influenza epidemics using search engine query data

Jeremy Ginsberg¹, Matthew H. Mohebbi¹, Rajan S. Patel¹, Lynnette Brammer², Mark S. Smolinski¹ & Larry Brilliant¹

Seasonal influenza epidemics are a major public health concern, causing tens of millions of respiratory illnesses and 250,000 to 500,000 deaths worldwide each year¹. In addition to seasonal influenza, a new strain of influenza virus against which no previous immunity exists and that demonstrates human-to-human transmission could result in a pandemic with millions of fatalities². Early detection of disease activity, when followed by a rapid response, can reduce the impact of both seasonal and pandemic influenza^{3,4}. One way to improve early detection is to monitor health-seeking behaviour in the form of queries to online search engines, which are submitted by millions of users around the world each day. Here we present a method of analysing large numbers of Google search queries to track influenza-like illness in a population. Because the relative frequency of certain queries is highly correlated with the percentage of physician visits in which a patient presents with influenza-like symptoms, we can accurately estimate the current level of weekly influenza activity in each region of the United States, with a reporting lag of about one day. This approach may make it possible to use search queries to detect influenza epidemics in areas with a large population of web search users.

By aggregating historical logs of online web search queries submitted between 2003 and 2008, we computed a time series of weekly counts for 50 million of the most common search queries in the United States. Separate aggregate weekly counts were kept for every query in each state. No information about the identity of any user was retained. Each time series was normalized by dividing the count for each query in a particular week by the total number of online search queries submitted in that location during the week, resulting in a query fraction (Supplementary Fig. 1).

We sought to develop a simple model that estimates the probability that a random physician visit in a particular region is related to an ILI; this is equivalent to the percentage of ILI-related physician visits. A single explanatory variable was used: the probability that a random search query submitted from the same region is ILI-related, as determined by an automated method described below. We fit a linear model using the log-odds of an ILI physician visit and the log-odds of an ILI-related search query: $\text{logit}(I(t)) = \alpha \text{zlogit}(Q(t)) + \varepsilon$, where $I(t)$ is the percentage of ILI physician visits, $Q(t)$ is the ILI-related query fraction at time t , α is the multiplicative coefficient, and ε is the error term. $\text{logit}(p)$ is simply $\ln(p/(1-p))$.

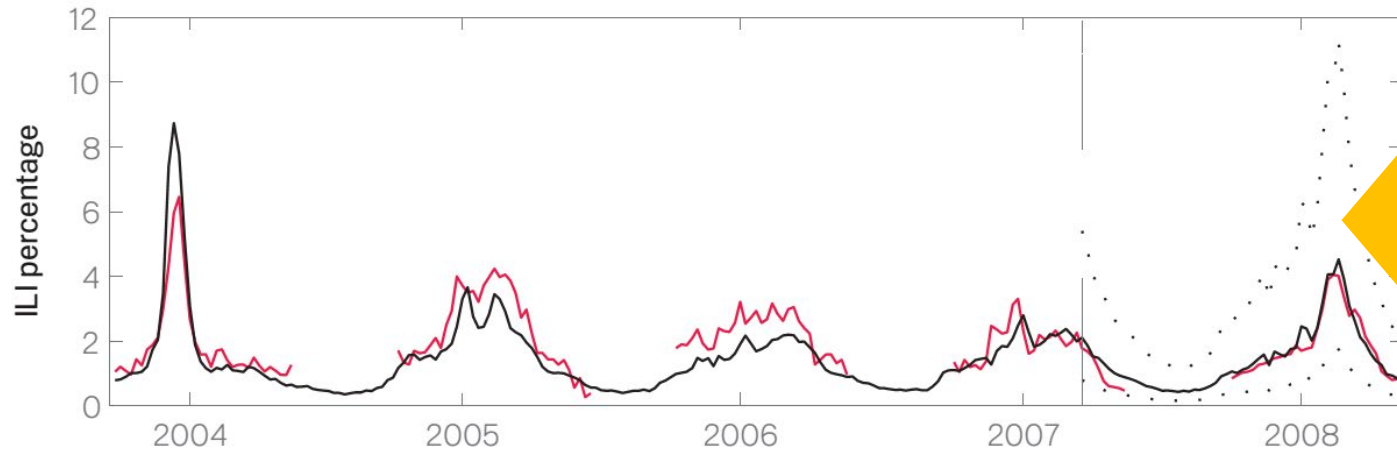
Publicly available historical data from the CDC's US Influenza

Example 2: Google Flu Trends – The Promise of Big Data

- One way to improve early detection is to monitor health-seeking behaviour in the form of queries to online search engines, which are submitted by millions of users around the world each day.
- The authors present a method of analyzing large numbers of Google search queries to track influenza-like illness in a population.
- Because the relative frequency of certain queries is highly correlated with the percentage of physician visits in which a patient presents with influenza-like symptoms, we can accurately estimate the current level of weekly influenza activity in each region of the United States, with a reporting lag of about one day.
- This approach may make it possible to use search queries to detect influenza epidemics in areas with a large population of web search users.

Ginsberg et al., 2009: <https://www.nature.com/articles/nature07634.pdf>

Example 2: Google Flu Trends – The Peril of Big Data...



The Confidence Bands began to widen greatly in 2008 suggesting less stable Google Trend Estimates.

Figure 2: A comparison of model estimates for the Mid-Atlantic Region (black) against CDC-reported ILI percentages (red), including points over which the model was fit and validated. A correlation of 0.85 was obtained over 128 points from this region to which the model was fit, while a correlation of 0.96 was obtained over 42 validation points. 95% prediction intervals are indicated.

Source: <http://dx.doi.org/10.1038/nature07634>

Example 2: Google Flu Trends – The Peril of Big Data...

- What happened? **Lazar and colleagues (2014)** discuss reasons why the Google Flu Trends method stopped working over time.
 - The reasons people searched for flu-related terms changed over time.
 - Therefore, the correlation between searching for flu-related terms and actual flu in the population changed (Cook et al. 2011)
 - So called “Big Data Hubris” where more data is better is not always the case...
 - Exclusion of potentially helpful keywords and lack of transparency around the final set of keywords used in the model (45 of them).
 - Algorithmic dynamics and changes to how search engines work over time (and user behavior).

Example 2: Google Flu Trends – Unleashing the Power of Big Data

- More recently, **Martin, Xu and Yasui (2015)** published an open source article examining how the Google Flu Trends data could be improved for the U.S. using a transformation based on CDC historical data.

Example 2: Google Flu Trends – Unleashing the Power of Big Data

- Martin and colleagues (2015) leveraged the flu trends preliminary estimates and final statistics from CDC along with weekly Google Flu Trend data (GFT) and created a transformation that assumed relative changes in CDC values from a prior week would be proportional to those from GFT (see <https://doi.org/10.1371/journal.pone.0109209>).

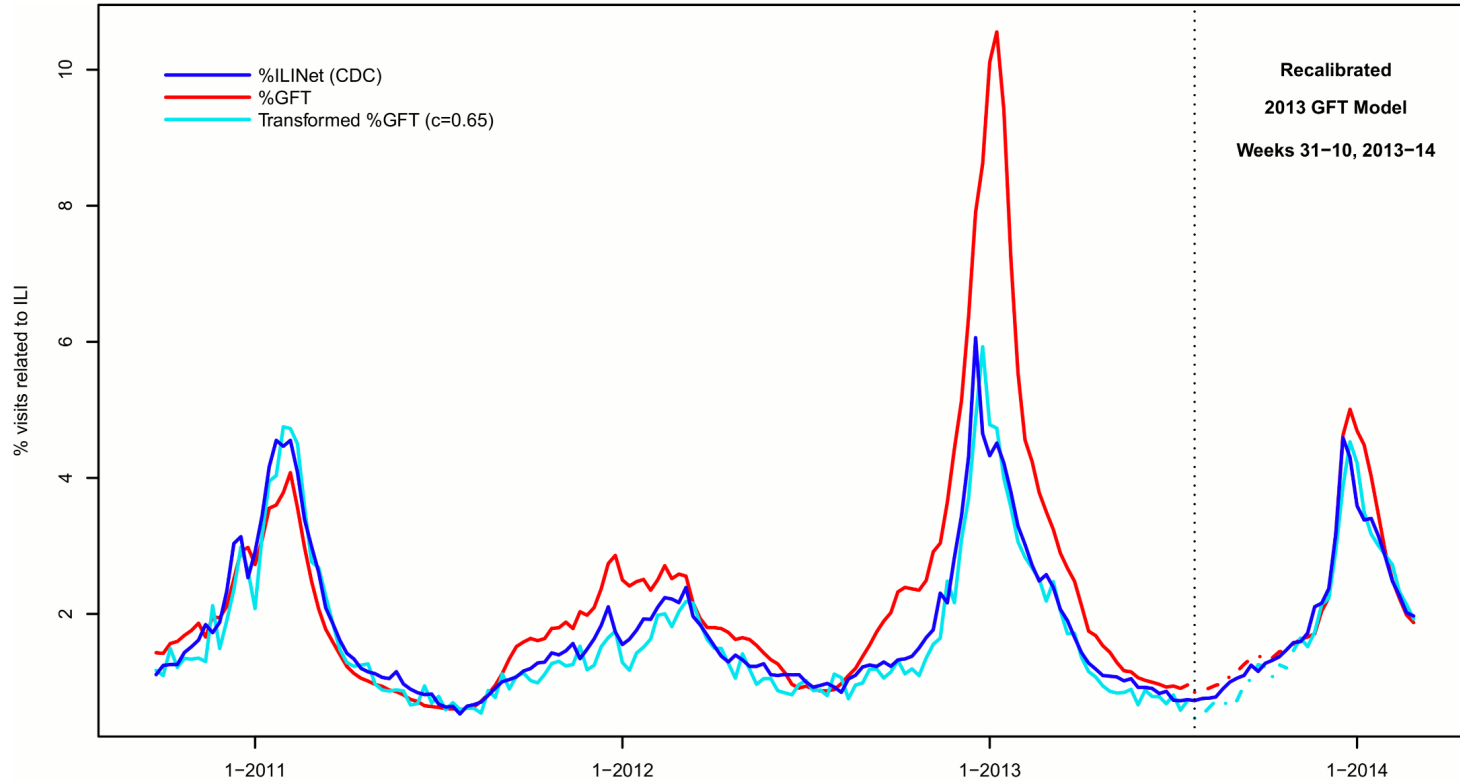
$$\frac{(f\%ILINet_i - p\%ILINet_{i-1})}{p\%ILINet_{i-1}} = c \frac{(\%GFT_i - \%GFT_{i-1})}{\%GFT_{i-1}}$$

Relative changes in CDC
values from a prior week

Relative changes in GFT
values from a prior week

They ultimately estimated the value of the proportionality constant, c to be 0.65

Example 2: Google Flu Trends – Unleashing the Power of Big Data



Martin LJ, Xu B, Yasui Y (2014) Improving Google Flu Trends Estimates for the United States through Transformation. PLOS ONE 9(12): e109209. <https://doi.org/10.1371/journal.pone.0109209>;

Example 2: Google Flu Trends – Unleashing the Power of Big Data

- In 2010–13, the transformed %GFT estimates were within ten percentage points of the true final values from CDC for 17 of the 29 weeks of “flu season”;
- The original %GFT estimates were within the same margin of error for only two of those 29 weeks.
- The sum of squared errors for the:
 - Original GFT was 177.4
 - CDC Original figure was 17.0
 - Transformed GFT was 12.1

Source: Martin LJ, Xu B, Yasui Y (2014) Improving Google Flu Trends Estimates for the United States through Transformation. PLoS ONE 9(12): e109209. <https://doi.org/10.1371/journal.pone.0109209>

Opportunities for Sensor Data...



- English et al. (2020) use data from 140 sensors throughout Chicago to investigate how air quality affects chronic health conditions
- Illic et al. (2020) collect picture data from the general population using smartphone sensors
 - Willingness to take pictures differs by type of request (heating 36% vs. 67% survey question, favorite place about 45% vs. 98% survey question)
 - 88% of pictures in line with the request (e.g., for heating elements 66% contained half or more of the element)
- Collect real-time data in everyday situations, usually passively collected without reactivity biases. Elevelt et al. (2019) use GPS smartphone location data to derive location the respondent is at. They show that home location can be derived reliably but respondent information is needed to derive other locations (work or school)




- Data access is limited to consenting individuals, for example, when collecting smart phone sensor/image data (e.g., Haas et al. 2020) with image data not necessarily conforming to the standards requested by the researchers (e.g., Illic et al. 2020).
- Sensor data from smart phones may be missing for various reasons: individuals run out of battery, have no signal, or other technical problems and sensors may be of different types for different phones/models (e.g., Bähr et al. 2020).
- Deriving measures from raw sensor data, such as functional geographic locations or travel paths, can be challenging and may contain measurement error (e.g., Abdulazim et al. 2013; Elevelt et al. 2019)

Some Resources for Thinking about Big Data Quality



- Total Data Error (Amaya, Biemer and Kinyon 2020)
 - Considers Error from Survey and Non-Survey Sources modelled after the Total Survey Error Framework
- Total Data Quality (West, Wagner, Buskirk, Kim, 2020).
 - Posits a Total Quality Assessment where we propose a unifying framework From Data Origin to Access to Sources to Processing and Analysis to examine total data quality for both designed and gathered data broadly defined using assessments of validity and accuracy relevant for a given data source.
 - For Gathered Data (e.g. Big Data) there is an expanded list of threats to that relate to coverage/representation as well as temporal validity and reproducibility including platform specific and technology related dimensions.
- Upcoming Special Issue of mda = Methods Data and Analysis looking at using big data sources for solving social science problems
 - This special issue will be peer reviewed but will contain a peer edited appendix that details in a reflective way choices and methods researchers had to use to overcome unexpected challenges with big data sourcing, processing or analysis.

Coursera Specialization on Total Data Quality: West, Wagner, Kim and Buskirk (2022)

coursera ▼ **Explore** ▼ 🔍 Online Degrees ▼ Find your New Career For Enterprise For Universities 🔔  Trent D Buskirk ▼

Filter by

Credit Eligible

Subject

Business

Data Science


Health

Information Technology

Skills

Accounting

15 results for "total data quality"




University of Michigan

Total Data Quality

Skills you'll gain: Entrepreneurship, Market Research, Research and Design

Beginner · Specialization · 1-3 Months




University of Michigan

The Total Data Quality Framework

Skills you'll gain: Entrepreneurship, Market Research, Research and Design

Beginner · Course · 1-4 Weeks



University of Michigan

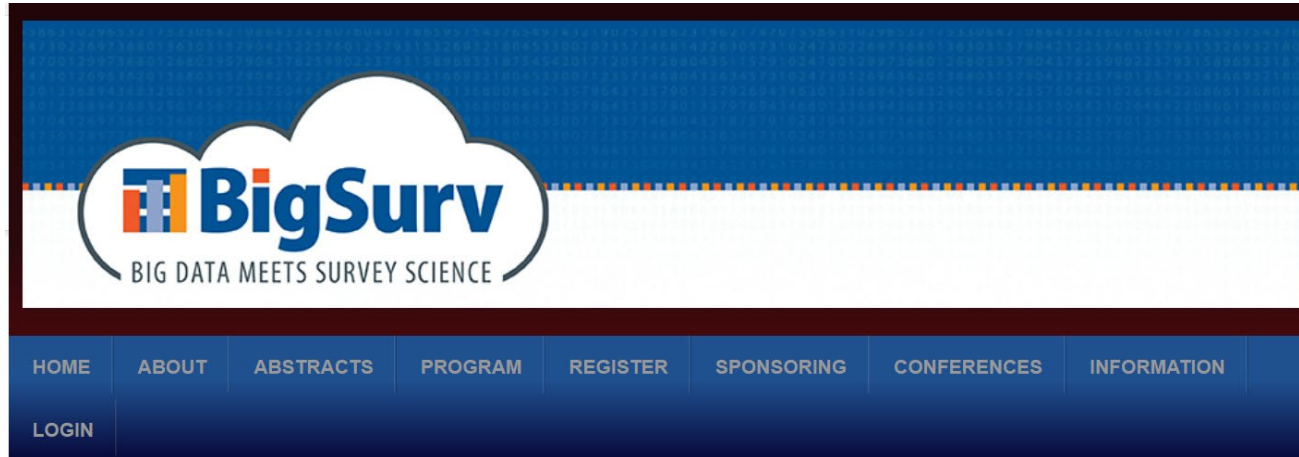
Measuring Total Data Quality

Beginner · Course · 1-4 Weeks

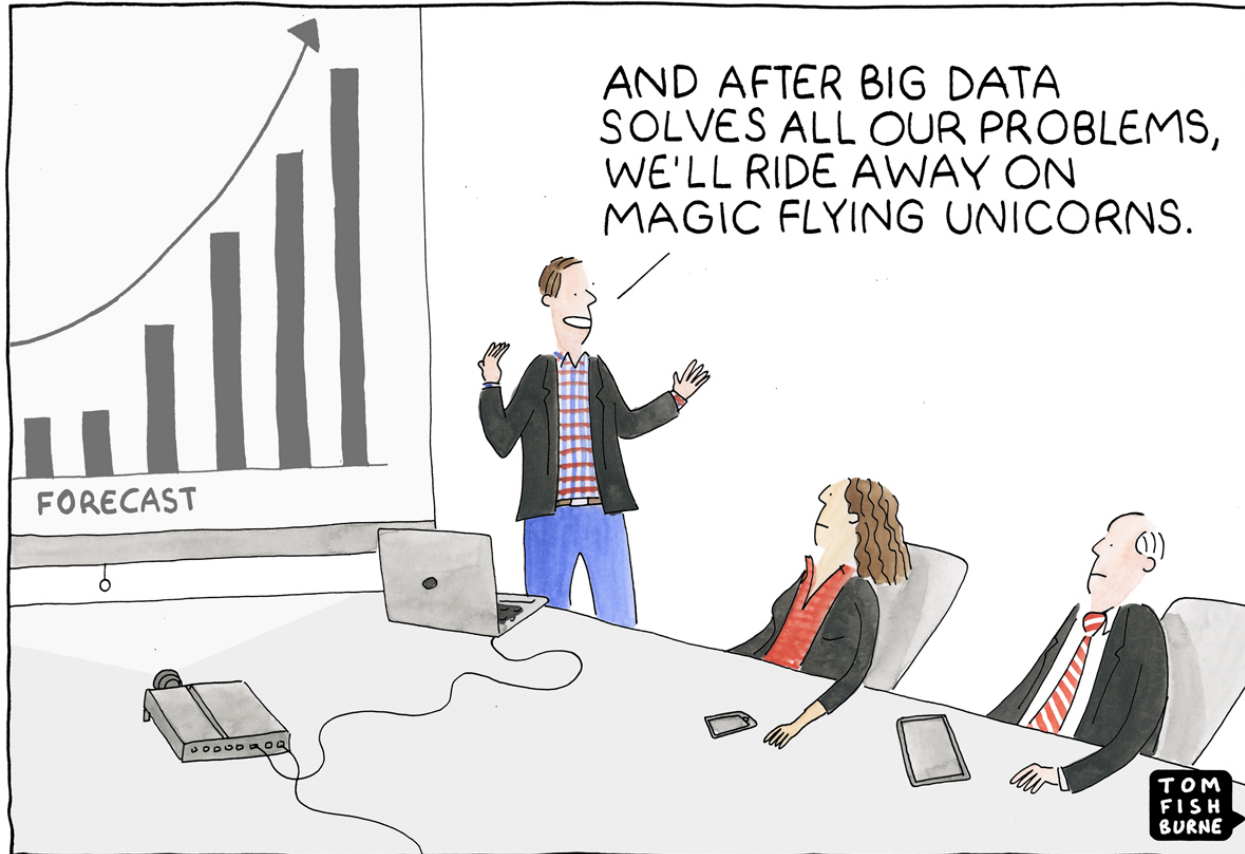
<https://bit.ly/total-data-quality>

From Big Data to Bigger Insights...

- Big Data sources offer much opportunity for survey and social science researchers in the Fourth Era to explore new domains and refine others.
- But their use is not without concerns – related to access, reproducibility, representation and quality to name a few.
- But out current statistical methods we have used for a long time are also not always well suited for Big Data sources. So machine learning methods will continue to emerge and be used in practice to leverage these alternate data sources.
 - Buskirk and Kirchner (2020) have discussed numerous examples of how machine learning is incorporated into the survey research process using big and small data alike.
- So as you proceed with your own Big Data sets consider the opportunities they present, the validity threats inherent in them and ways to leverage Big Data and other sources together to unlock the full potential.



The next conference is scheduled to take place in-person in Quito, Ecuador on October 26-29, 2023. The Conference will be hosted by the United Nations Association of Ecuador (UNA-Ecuador) in cooperation with the Data Science Institute of Universidad San Francisco de Quito. Visit <https://www.bigsurv.org/> for more details!



Thank you!



Trent D. Buskirk, PhD
buskirk@bgsu.edu

@trentbuskirk

References

- R. Groves (2011). *Public Opinion Quarterly*, Vol. 75, No. 5, pp. 861–871
- English, N. Ventura, I., Frasier, A. Buskirk, T.D. and Malarek, D. (2015). Can We Hit The Mark? Using Commercial and Publicly-Available Data to Target Specific Populations. *Paper presented at the Joint Statistical Meetings, Seattle, August 8-13, 2015.*
- ten Bosch, O., Windmeijer, D., van Delden, A., & van den Heuvel, G. (2018, October). Web scraping meets survey design: Combining forces. In *Big Data Meets Survey Science Conference, Barcelona, Spain.*
- Barcaroli G, et al. ISTAT Farm Register: Data Collection by Using Web Scraping for Agritourism Farms, ISTAT, ICASVII, Rome 2016.
- **Buskirk, T.D.** and Kirchner, A. (2020) “Why Machines Matter for Survey and Social Science Researchers: Exploring Applications of Machine Learning Methods to Design, Data Collection and Analysis,” in: *Big Data Meets Survey Science: A Collection of Innovative Methods, First Edition.* Edited by Craig A. Hill, Paul P. Biemer, **Trent D. Buskirk**, Lilli Japiec, Antje Kirchner, Stas Kolenikov, and Lars E. Lyberg. © 2020 John Wiley & Sons, Inc. Published 2020 by John Wiley & Sons, Inc.
- Amaya, A., Biemer, P. P., & Kinyon, D. (2020). Total Error in a Big Data World: Adapting the TSE Framework to Big Data. *Journal of Survey Statistics and Methodology*, 8(1), 89-119.
<https://doi.org/10.1093/jssam/smz056>
- Connelly, R., Playford, C. J., Gayle, V., & Dibben, C. (2016). The role of administrative data in the big data revolution in social science research. *Social Science Research*, 59, 1-12. doi: 10.1016/j.ssresearch.2016.04.015

References

- Datta, A. R., Ugarte, G., & Resnick, D. (2020). Linking Survey Data with Commercial or Administrative Data for Data Quality Assessment. *Big Data Meets Survey Science: A Collection of Innovative Methods*, 99-129.
- Stern, M. J., Bilgen, I., McClain, C., & Hunscher, B. (2017). Effective sampling from social media sites and search engines for web surveys: Demographic and data quality differences in surveys of Google and Facebook users. *Social Science Computer Review*, 35(6), 713-732.
- Illic, G., Struminskaya, B., Lugtig, P. (2020) Collecting picture data from the general population using smartphone sensors. Paper presented at the American Association for Public Opinion Research virtual conference.
- Bähr, S., Haas, G.-C., Keusch, F., Kreuter, F., Trappmann, M. (2020). Missing data and other measurement quality issues in mobile geolocation sensor data. *Social Science Computer Review*. DOI: 10.1177/0894439320944118.
- Elevelt, A., Bernasco, W., Lugtig, P., Ruiter, S., Toepoel, V. (2019). Where you at? Using GPS locations in an electronic time use diary study to derive functional locations. *Social Science Computer Review*. DOI: 10.1177/0894439319877872.
- Dove, J. (2015, April 29). Facebook shuts down friends data API. Retrieved from <http://thenextweb.com/dd/2015/04/29/facebook-shuts-down-friends-data-api-to-generate-more-trust-among-users/>.

References

- English, N., Zhao, C., Brown, K.L., Catlett, C., Cagney, K. (2020). Making sense of sensor data: How local environmental conditions add value to social science research. *Social Science Computer Review*. DOI: 10.1177/0894439320920601
- Abdulazim, A., Abdelgawad, H., Habib, K.M.N., Abdulhai, B. (2013). Using smartphones and sensor technologies to automate collection of travel data. *Transportation Research Record: Journal of the Transportation on Research Board*. 2383(1): 44-52). DOI: <https://doi.org/10.3141/2383-06>
- Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2020). Integrating survey data and digital trace data: key issues in developing an emerging field.
- Buskirk, T. D., Callegaro, M., & Rao, K. (2010). “N the Network?” Using Internet Resources for Predicting Cell Phone Number Status. *Social science computer review*, 28(3), 271-286.
- Lomborg, S., & Bechmann, A. (2014). Using APIs for data collection on social media. *The Information Society*, 30(4), 256-265.
- Wang, Z., Hale, S., Adelani, D. I., Grabowicz, P., Hartman, T., Flöck, F., & Jurgens, D. (2019). Demographic Inference and Representative Population Estimates from Multilingual Social Media Data. *Paper presented at the The World Wide Web Conference, San Francisco, CA, USA*. <https://doi.org/10.1145/3308558.3313684>.
- Wells, C., & Thorson, K. (2017). Combining Big Data and Survey Techniques to Model Effects of Political Content Flows in Facebook. *Social Science Computer Review*, 35(1), 33–52. <https://doi.org/10.1177/0894439315609528>

References

- Möller, J., van de Velde, R. N., Merten, L., & Puschmann, C. (2020). Explaining online news engagement based on browsing behavior: Creatures of habit?. *Social Science Computer Review*, 38(5), 616-632.
- Schneider, D., & Harknett, K. (2019). What's to like? Facebook as a tool for survey data collection. *Sociological Methods & Research*, 0049124119882477.
- Burke-Garcia, A., Edwards, B., & Yan, T. (2020). The Future Is Now: How Surveys Can Harness Social Media to Address Twenty-first Century Challenges. *Big Data Meets Survey Science: A Collection of Innovative Methods*, 63-97.
- Hinds, J., & Joinson, A. N. (2018). What demographic attributes do our digital footprints reveal? A systematic review. *PloS ONE*, 13(11).
- Curry, E. (2016). The big data value chain: definitions, concepts, and theoretical approaches. In *New horizons for a data-driven economy* (pp. 29-37). Springer, Cham.
- Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., & Shahabi, C. (2014). Big data and its technical challenges. *Communications of the ACM*, 57(7), 86-94.
- Rolf, E., Worledge, T., Recht, B., & Jordan, M. I. (2021). Representation Matters: Assessing the Importance of Subgroup Allocations in Training Data. arXiv preprint arXiv:2103.03399.
- Schat, E., van de Schoot, R., Kouw, W. M., Veen, D., & Mendrik, A. M. (2020). The data representativeness criterion: Predicting the performance of supervised classification based on data set similarity. *Plos one*, 15(8), e0237009.