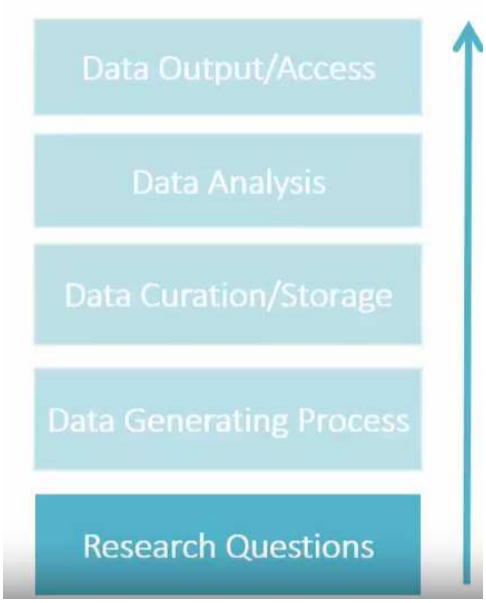
Framework for Data Collection and Analysis

Week 1 Research Question Design

Big Flow of survey design and analysis….



Three Types of Research Questions

1 Description
We're trying to show means, percentages, for certain subgroups. You're trying to put out some numbers about the world, about a piece of the world, describing certain variables or features as they appear in the world.

2 Causality
So this one here captures all things that have to do with wanting to know whether a certain treatment, for example, taking some pain medication helps against headache, and makes you smile.
With humans, it's often hard, if not impossible, to see the counterfactual.

3 Predictions

➔ Some kind of "**inferential**" goal in mind: So having a smaller set of data from a larger population in which case you want to do inference to the larger population.

For description, I would argue, that what you need is a positive and known selection probability
[In survey sampling, the term probability of selection refers to the chance (i.e. the probability from 0 to 1) that a member (element) of a population can be chosen for a given survey]
➔ Useful for dealing with missing data

Now, for the other two, this causation and prediction, the known selection probabilities are a little less important. Maybe not important at all, actually. What I do want to know though, is that I still have a positive selection probability of everybody.

Data Generating Process

Types of Data

1 Experiments (Laboratory)

- Systematic variation of treatment between one or more groups
- Ideally units are randomized to groups with the goal of equal distribution of characteristics across groups (no cofounders)
- Often smaller in scale
- Designed for a specific research purpose
- Often difficult to balance the strong internal validity and the often weaker external validity

2 Survey Interview Data

- Data collected with specific research purpose in mind
- e.g. Political polling, national crime, and victimization surveys, health surveys…
- Large variation in size and quality
- Timeliness depends in part on mode of data collection
- Often cross-section (on-off) sometimes panel studies (repeated data collection on the same units)
- Usually a sample not a Census

3 Found / Organic Data

- Appear as part of a process or part of a largely uncontrolled data collection effort
- e.g. Boston Street Bump app
- Has all the Big Data characteristics ➔ volume (terabytes of data), velocity (streaming data), data in many forms (structured, unstructured, text, multimedia), data in doubt (uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception etc.)

4 Administrative Data

- Data created as side effect of program administrations of any kind
- e.g. insurance plans, welfare benefits, social security systems, tax forms, reimbursement for services
- generated with a purpose but primary purpose is not research, thus quality often best for program relevant (mandatory) variables (fields)
- Delay in reporting and dissemination (i.e. slow)
- Includes (most) units in program/process
- The number of variables is rather small

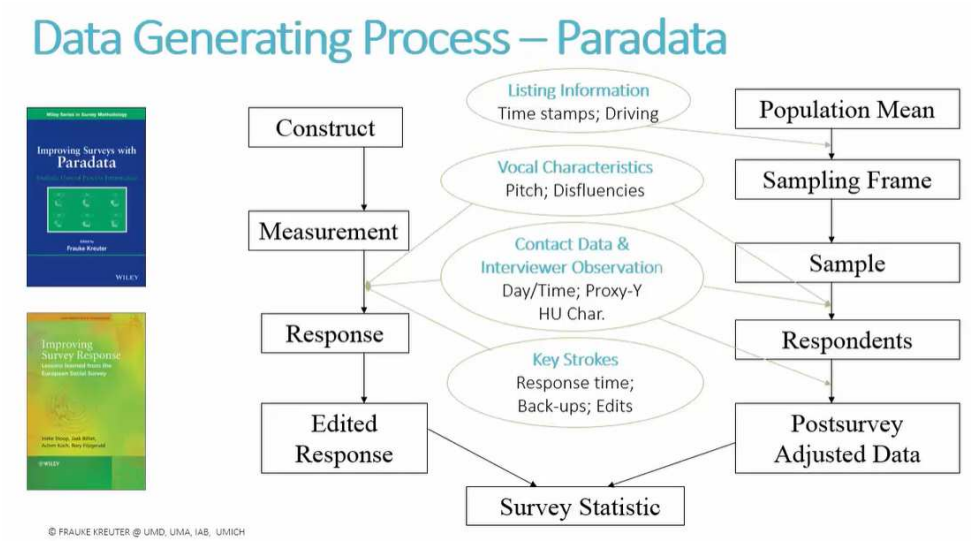
Examples of Found Data

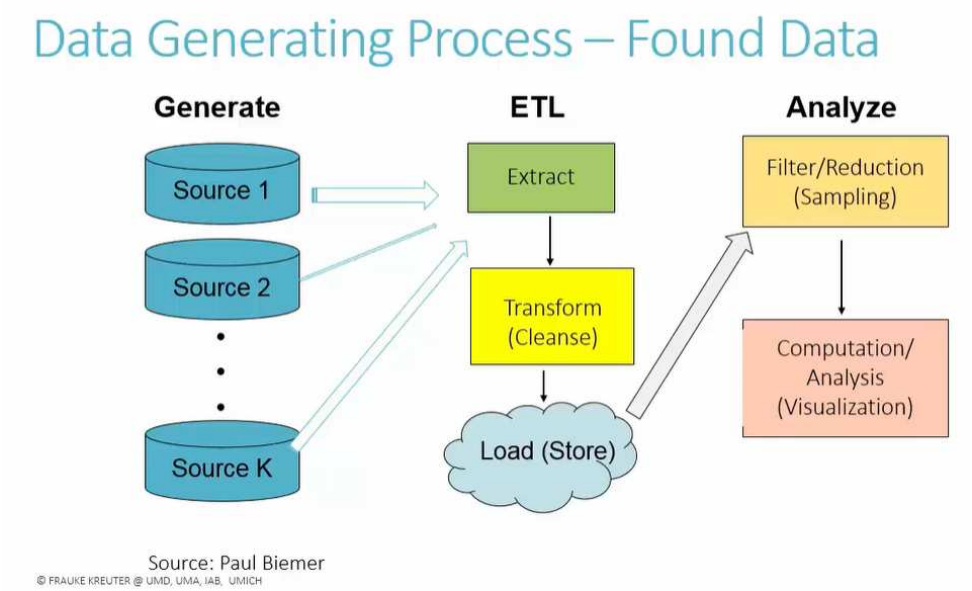
ref. UNECE Big Data, MIT billion price project

- Consumer price indices: experimenting with the computation of price indexes
- Mobile telephone data: statistics on tourism and daily commuting
- smart meters: statistics on power consumption using data collected from smart meter readings
- traffic loops: traffic statistics using data from traffic loops
- social media: using twitter data to analyze sentiment and to tourism flows
- job portals: computing statistics on job vacancies
- web scraping: tested methods for automatically collecting data from web sources

Visualizing the Data Generation Process

Who? What? Why? Who is missing? Who is counted repeatedly? What is not said / measured? … and why?





Data Curation

Post Processing
- Editing; De-identification
- Data Entry
- Coding
- Error Checking
- Data set construction
- Codebook construction
- Variable (feature) generation
    - ** Continuous (interval or ratio scale) – systolic blood pressure (mmHg)
    - ** Categorical (ordered or nominal scale)
    - ** Multiple Responses
    - ** Indicator Variables

- Building weights
- Imputing missing values

Data Analysis

How to combine / link different data sources?
How to properly capture variability?

Many samples…
Distribution of means….

Confidence intervals for clustered samples tend to be wider

Data Access

Access Issues
Privacy v.s. Public Good
Consent v.s. Confusion
Convenience v.s. Accusation
Privacy v.s. Data Quality
Identifiable v.s. Reachable
Europe v.s. USA

Access Resources

UK Data Archive
ICPSR
GESIS
Data.gov
eurostat
The Open Data Foundation (ODaF)

Key Ingredients for Valid Inference

Data Generating Process needs to be known
Framework as a tool to identify errors
Model or Break Confounders (through design)
Know your inferential Goal

American Association for Public Opinion Research AAPOR (2015)
https://www.aapor.org/AAPOR_Main/media/Task-Force-Reports/BigDataTaskForceReport_FINAL_2_12_15_b.pdf

Surveys and Big Data are complementary data sources not competing data sources.
Big Data offers the possibility to study tails of distributions.

The MIT Billion Prices Projects
Academic initiative using prices collected daily from hundreds of online retailers around the world to conduct economic research. One statistical product is the estimation of inflation in the US. Changes in inflation trends can be observed sooner in PriceStats than in the monthly Consumer Price Index (CPI).

City of Boston – Traffic / Road data
Issued a smart phone application available to anybody, which is designed to automatically detect pavement problems. Anyone who downloads the mobile app creates data about the smoothness of the ride. According to their website, these data provide the City with real-time information it uses to fix problems and plan long term investments.

Consumer Confidence Index
Produced every month by Statistics Netherlands using survey data. The index measures households' sentiments on

their financial situation and on the economic climate in general. Daas and Puts (2014) studied social media messages to see if they could be used to measure social media sentiment. They found that the correlation between social media sentiment (mainly Facebook data) and consumer confidence is very high. Social media messages (in this case Twitter data) form the basis of the University of Michigan Social Media Job Loss Index, with the goal of generating early predictions of Initial Claims for Unemployment Insurance. The predictions are based on a factor analysis of social media messages mentioning job loss and related outcomes (Antenucci et al. 2014).

Big Data Hubris
Occurs when the Big Data researcher believes that the volume of the data compensates for any of their deficiencies, thus obviating the need for traditional, scientific analytic approaches. As Lazer et al. (2014:2) note, Big Data hubris fails to recognize that "… quantity of data does not mean that one can ignore foundational issues of measurement and construct validity and reliability…

Total Survey Error (TSE) Framework
Framework that identifies all the major sources of error contributing to data validity and estimator accuracy (see, for example, Biemer 2010). The TSE framework also attempts to describe the nature of the error sources and what they may suggest about how the errors could affect inference. The framework parses the total error into bias and variance components which, in turn, may be further subdivided into subcomponents that map the specific types of errors to unique components of the total mean squared error.

Big Data Total Error (BDTE) framework
Include additional error sources that are unique to Big Data and can create substantial biases and uncertainties in Big Data products.

Couper, M.P. (2013). Is the sky falling? New technology, changing media, and the future of surveys. Survey Research Methods, 7, 145-165.

Two Types of Biases in Big Data

Selection Bias
Big data tends to focus more on the "haves" and less on the "have-nots". This may also be true of much market research, but social research has traditionally been more interested in the "have-nots"

Measurement Bias
Again, this is something that is well known to survey researchers, but has tended to be ignored in the heady rush to exploit the volume of organic data becoming available. social media is primarily about impression management (see Boyd & Ellison, 2008). To what extent do people's posts represent their "true" values, beliefs, behaviors, etc.? Similarly, if we counted the number of Facebook friends one has as an indicator of true social network size, we may be seriously wrong. The average Facebook user is estimated to have 229 "friends" (Hampton, Goulet, Rainie, & Purcell, 2011).