Bit By Bit: Social Research in the Digital Age

Chapter 02: Observing Behavior

2.1 Introduction

Business + Government records = Big Data

Big Data subset of **Observational data**

Observational data is any data that results from observing a social system without any intervention.

- everything that does not involve talking with people (e.g., surveys) or changing people's environments (e.g., experiments)
- includes newspaper articles text and satellite photos.

2.2 Big Data

Created and collected by companies and governments **for purposes other than research**.

To use for research requires **repurposing**

Definition: “3 Vs”

- Volume (lot of data)
- Variety (variety of formats)
- Velocity (being created constantly)

other “Vs” such as Veracity, Value, Vague and Vacuous.

By author-

- 5 “Ws”: Who, What, Where, When, and Why.
- Most influential is “Why”

2.2 Big Data Sources

Have positives and negatives

Focus on for which kinds of research questions big data sources have attractive properties and for which kinds of questions they might not be ideal

Two other important sources of big data (other than search engine logs and social media posts),

- First, any digital devices in the physical world
- Second, government administrative records(tax records, school records, healthcare records)

Two general advices for **repurposing**,

- Imagine the ideal data. Compare with what you have. Determine what you can and can not use it for. Get more data if necessary
- Understand the data generating process

2.2 Big Data Sources

Social Scientists and Data Scientists tend to approach repurposing very differently.

Social scientists -> quick to point out the problems with repurposed data while ignoring its strengths.

Data scientists -> are quick to point out the benefits of repurposed data while ignoring its weaknesses.

The **best** approach is a **hybrid**.

Researchers need to understand the characteristics of big data sources—both good and bad—and then figure out how to learn from them

e.g. Twitter data VS General Social Survey (GSS) data

2.3 Ten Common Characteristics of Big Data

Big data sources tend to have a number of characteristics in common

HELPFUL

- Big
- Always-on
- Nonreactive

PROBLEMATIC

- Incomplete
- Inaccessible
- No representative
- Drifting
- Algorithmically confounded
- Dirty
- Sensitive

2.3.1 Big

Big datasets are not an end in themselves.

Enable certain kinds of research

- Study of rare events
- Estimation of heterogeneity
- Detection of small differences

Big datasets also seem to lead some researchers to ignore how their data was created, which can lead them to get a precise estimate of an unimportant quantity

2.3.2 Always-on

Always-on data systems,

- enable researchers to study unexpected events
- provide real-time information to policy makers

NOT well suited for tracking changes over very long periods of time because many big data systems are constantly changing—a process called *drift*

2.3.3 Nonreactive

Participants are generally not aware that their data are being captured or they have become so accustomed to this data collection that it no longer changes their behavior.

Does **NOT** ensure that these data are direct reflection of people's behavior or attitudes. Meaning, they are **NOT** always free of *social desirability bias* (the tendency for people to want to present themselves in the best possible way)

2.3.4 Incomplete

Don’t have the information needed for research.

Big data tends to be missing 3 types of information useful for social research-

- Demographic information about participants
- Behavior on other platforms
- Data to operationalize theoretical constructs (hardest to solve; often accidentally overlooked)

Theoretical constructs are abstract ideas (e.g. norms, intelligence, social capital, democracy) that social scientists study

Operationalizing a theoretical construct means proposing some way to capture that construct with observable data

e.g. “People who are more intelligent earn more money” by “People on Twitter who used longer words are more likely to mention luxury brands” > invalid constructs

Solutions-

1. to actually collect the data you need (surveys)
2. 2. data scientists call user-attribute inference and social scientists call imputation- the information that they have on some people to infer attributes of other people
3. 3. combine multiple data sources

2.3.5 Inaccessible

Legal, business, and ethical barriers that prevent data access

Researchers can partner with companies to obtain data access, but this can create a variety of problems for researchers and companies-

1. Can not share data with other researchers to verify and extend result
2. What you can aske may be limited

Four ingredients of the successful partnerships:

1. researcher interest
2. researcher capability
3. company interest
4. company capability

2.3.6 Nonrepresentative

For some scientific questions, Nonrepresentative data can be quite effective, for others not well suited

To distinguish, whether the questions are about-

- within-sample comparisons
- out-of-sample generalizations

Many big data sources are not representative samples from some well-defined population

But for questions about within-sample comparisons, nonrepresentative data can be powerful, so long as researchers are clear about the characteristics of their sample and support claims about transportability with theoretical or empirical evidence.

2.3.7 Drift

Drift make it hard to use big data sources to study long-term trends

One of the great advantages of many big data sources is that they collect data over time. Social scientists call this kind of over-time data ***longitudinal data*** -are very important for studying change.

Big data systems—especially business systems—are changing all the time. Especially in 3 main ways-

1. Population drift (change in who is using them)
2. Behavioral drift (change in how people are using them)
3. System drift (change in the system itself)

2.3.8 Algorithmically confounded

- Although Nonreactive but not necessarily “naturally occurring”
- Highly engineered to induce specific behaviors (e.g. clicking on ads or posting content)
- The ways that the goals of system designers can introduce patterns into data is called algorithmic confounding.
- One form of system drift.
- We should be cautious about any claim regarding human behavior that comes from a single digital system, no matter how big

2.3.9 Dirty

- Frequently dirty. That is, they frequently include data that do not reflect real actions of interest to researchers
- No single statistical technique or approach to can sufficiently clean the dirty data.
- Best way to avoid being fooled by dirty data, is to understand as much as possible about how the data were created

2.3.10 Sensitive

- Can be very sensitive hence, inaccessible
- De-identification—can fail in surprising ways
- Raises ethical questions, even if no specific harm is caused

2.4 Research Strategies

3 main strategies for learning from big data sources-

- Counting things
- Forecasting things
- Approximating experiments

2.4.1 Counting Things

- Lots of social research is really just counting things
- What things are worth counting?
  For example, a student might say that many people have studied migrants and many people have studied twins, but nobody has studied migrant twins.
- Better strategy is to look for research questions that are important or interesting (or ideally both)
- Important research is that it has some measurable impact or feeds into an important decision by policy makers
- Relatively simple counting of big data sources can, in some situations, lead to interesting and important research. The researchers must bring interesting questions to the big data source; the data by itself was not enough

2.4 Forecasting and nowcasting

- The second main strategy researchers can use with observational data is forecasting
- Making guesses about the future is notoriously difficult, and perhaps for that reason, forecasting is not currently a large part of social research.
- Focus on a special kind of forecasting called nowcasting - to “predict the present”
- Epidemiology, is one setting where the need for timely and accurate measurement is essential
- When evaluating any forecast or nowcast, it is important to compare against a baseline
- Consider drift and algorithmic confounding.

2.4 Approximating experiments

Some important scientific and policy questions are causal. For example, what is the effect of a job training program on wages?

The concern about possible preexisting differences arises no matter how many workers are in your data.

In many situations, the strongest way to estimate the causal effect of some treatment, such as job training, is to run a randomized controlled experiment where a researcher randomly delivers the treatment to some people and not others.

Two strategies that can be used with non-experimental data. The first strategy depends on looking for something happening in the world that randomly (or nearly randomly) assigns the treatment to some people and not others. The second strategy depends on statistically adjusting non-experimental data in an attempt to account for preexisting differences between those who did and did not receive the treatment.

These approaches can go badly wrong, but when deployed carefully, these approaches can be a useful complement to the experimental approach

Questions ?

Thank you