Digitalisation and organisational innovation

Lesson 9. **Big data**

Introduction

- In the **analog age**, collecting data about behavior (who does what, and when) was expensive and therefore relatively rare. Now, in the **digital age**, the behaviors of billions of people are recorded, stored, and analyzable.
- Because these types of data are a by-product of people's everyday actions, they are often called **digital traces**.
- In addition to these traces held by businesses, there are also large amounts of incredibly rich data held by governments. Together, these **business** and **government records** are often called **big data**.
- The ever-rising flood of big data means that we have moved from a world where behavioral data was scarce to one where it is **plentiful**.

Big data

- In the analog age, most of the data that were used for social research were created for the purpose of **doing research**.
- In the **digital age**, however, huge amounts of data are being created by companies and governments for purposes other than research, such as providing services, generating profit, and administering laws.
- While there are undoubtedly huge opportunities for repurposing, using data that were not created for the purposes of research also presents **new challenges**.
- Compare, for example, a social media service (such as Twitter), with a traditional public opinion survey (such as the General Social Survey).
- https://www.europeansocialsurvey.org/data-portal

- Twitter operates at a scale and speed that the GSS cannot match, but, unlike the General Social Survey, Twitter does not carefully sample users and does not work hard to maintain comparability over time.

- If you want hourly measures of global mood, Twitter is the best choice. On the other hand, if you want to understand **long-term changes** in the polarization of attitudes in the United States, then the GSS is best.

- When thinking about big data sources, many researchers immediately focus on **online data** created and collected by companies, such as search engine logs and social media posts. However, this narrow focus leaves out two other important sources of big data.

- First, increasingly, **corporate big data** sources come from digital devices in the physical world (supermarket checkout data, call records from mobile phones or billing data created by electric utilities).

- The second important source is data created by governments, which researchers call **government administrative records**, include things such as tax records, school records, and vital statistics records (e.g., registries of births and deaths).

Ten common characteristics of big data

- Rather than taking a platform-by-platform approach (e.g., here's what you need to know about Twitter, here's what you need to know about Google search data, etc.), we describe **10 general characteristics** of big data sources, which can be grouped into two categories:

1. Generally **helpful** for research: big, always-on, and nonreactive

2. Generally **problematic** for research: incomplete, inaccessible, nonrepresentative, drifting, algorithmically confounded, dirty, and sensitive.

1. Big

- The most widely discussed feature of big data sources is that they are BIG. Many papers, for example, start by discussing (and sometimes bragging) about how much data they analyzed.

- **Is that all that data really doing anything?**

- Too often researchers seem to treat the size of big data source as an end (“look how much data I can crunch”) rather than a means to some more important scientific objective.

- There are three specific scientific ends that large datasets tend to enable:

1. ***The study of rare events.***
2. ***The study of heterogeneity*** (i.e. on social mobility in the United States).
3. ***To detect small differences***.

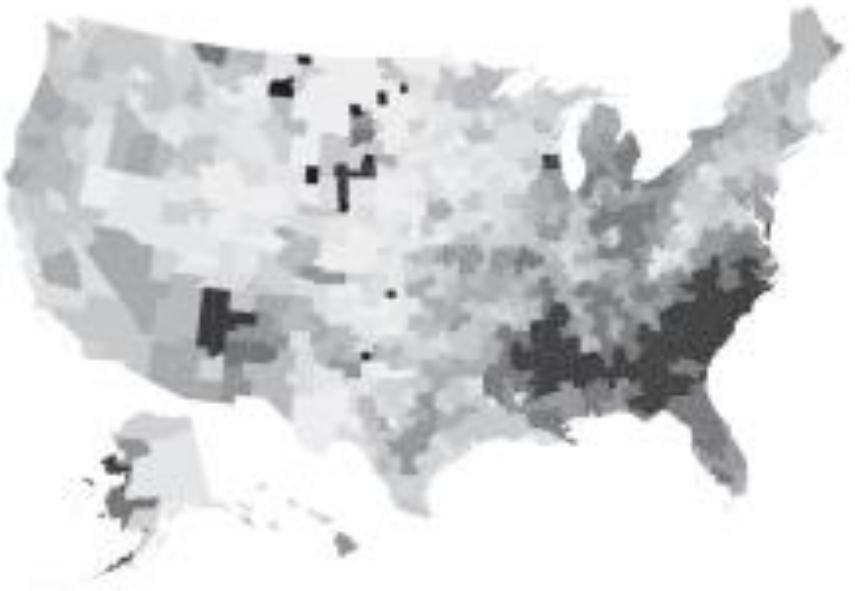
2. The study on social mobility in the United States (heterogeneity)

- Traditionally researchers have studied social mobility by comparing the life outcomes of parents and children.
- A consistent finding from this literature is that advantaged parents tend to have advantaged children, but the strength of this relationship varies over time and across countries.
- More recently, however, researchers were able to use the tax records from 40 million people to estimate the heterogeneity in intergenerational mobility across regions in the United States.
- They found that the probability that a child reaches the top quintile of the national income distribution, starting from a family in the bottom quintile is about 13% in San Jose (California) but only about 4% in Charlotte (North Carolina).

National

7.8%

Regional



>16.8%
12.9–16.8%
11.3–12.9%
9.9–11.3%
9.0–9.9%
8.1–9.0%
7.1–8.1%
6.1–7.1%
4.8–6.1%
<4.8%
Missing

- Researchers found that that high-mobility areas have less residential segregation, less income inequality, better primary schools, greater social capital, and greater family stability.

- These correlations alone do not show that these factors cause higher mobility, but they do suggest possible **mechanisms** that can be explored in further work.

- But the size of the data was really important in this project. If Chetty and colleagues had used the tax records of 40 thousand people rather than 40 million, they would not have been able to estimate regional heterogeneity.

3. To detect small differences

- Much of the focus on big data in industry is about these small differences: reliably detecting the difference between 1% and 1.1% click-through rates on an ad can translate into millions of dollars in extra revenue.

- In some scientific settings, however, such small differences might not be particular important, even if they are statistically significant.

- But, in some policy settings, they can become important when viewed in aggregate. For example, if there are two public health interventions and one is slightly more effective than the other, then picking the more effective intervention could end up saving thousands of additional lives.

Big data and conceptual error

- Although bigness is generally a good property when used correctly, it can sometimes lead to a conceptual error.

- For some reason, bigness seems to lead researchers to ignore how their data was generated.

- While bigness does reduce the need to worry about **random error** (errors in transcribing responses), it actually increases the need to worry about **systematic errors** (errori nella formulazione delle risposte), the kinds of errors that arise from biases in how data are created.

Conclusion

- Big datasets are not an end in themselves, but they can enable certain kinds of research, including the study of rare events, the estimation of heterogeneity, and the 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. Always-on

- Many big data systems are always-on; they are constantly collecting data. This always-on characteristic provides researchers with longitudinal data (i.e., data over time). Being always-on has two important implications for research.

- Always-on data systems enable researchers to study unexpected events and provide real-time information to policy makers.

- However, not always-on data systems are well suited for tracking changes over very long periods of time. That is because many big data systems are constantly changing.

3. Nonreactive

- One challenge of social research is that people can change their behavior when they know that they are being observed by researchers. Social scientists generally call this reactivity.

- For example, people can be more generous in laboratory studies than field studies because in the former they are very aware that they are being observed.

- Further, the behavior captured in big data sources is sometimes impacted by the goals of platform owners, an issue call algorithmic confounding.

- Finally, although nonreactivity is advantageous for research, tracking people’s behavior without their consent and awareness raises ethical concerns.

4. Incomplete

- Most big data sources are incomplete, in the sense that they don't have the information that you will want for your research. This is a common feature of data that were created for purposes other than research.
- Big data tends to be missing three types of information useful for social research: demographic information about participants, behavior on other platforms, and data to operationalize theoretical constructs (the hardest to solve).
- Theoretical constructs are abstract ideas, and operationalizing a theoretical construct means proposing some way to capture that construct with observable data. Social scientists call the match between theoretical constructs and data construct validity.
- ISTAT and social groups: https://www.istat.it/it/files/2018/02/GruppiSociali-nota.pdf

- To solve the other common types of incompleteness (incomplete demographic information and incomplete information on behavior on other platforms) there are two common solutions.

- The first solution is to do what data scientists call user-attribute inference and social scientists call imputation. In this approach, researchers use the information that they have on some people to infer attributes of other people.

- A second possible solution is to combine multiple data sources. This process is sometimes called record linkage.

5. Inaccessible

- Many sources of big data that would be useful are controlled and restricted by governments (e.g., tax data and educational data) or companies (e.g., queries to search engines and phone call meta-data).

- Therefore, even though these data sources exist, they are useless for the purposes of social research because they are inaccessible.

- Moreover, even if you are able to develop a partnership with a business or to gain access to restricted government data and you can “anonymize” the information, there are some others downsides.

- First, you will probably not be able to share your data with other researchers, which means that other researchers will not be able to verify and extend your results.

- Second, the questions that you can ask may be limited; companies are unlikely to allow research that could make them look bad.

- Finally, these partnerships can create at least the appearance of a conflict of interest, where people might think that your results were influenced by your partnerships.

6. Nonrepresentative

- Social scientists are accustomed to working with data that comes from a probabilistic random sample from a well-defined population. This kind of data is called representative data because the sample “represents” the larger population.

- To illustrate what can go wrong when researchers try to make an out-of-sample generalization from nonrepresentative data, we can use a study of the 2009 German parliamentary.

- By analyzing more than 100,000 tweets, they found that the proportion of tweets mentioning a political party matched the proportion of votes that party received in the parliamentary election. In other words, it appeared that Twitter data (which was essentially free), could replace traditional “public opinion surveys” (which are expensive because of their emphasis on representative data).

- But Germans on Twitter in 2009 were not a probabilistic random sample of German voters, and supporters of some parties might tweet about politics much more often than supporters of other parties.

- In fact the results were wrong: a follow-up paper pointed out that the original analysis had excluded the political party that had received the most mentions on Twitter: the Pirate Party, a small party that fights government regulation of the Internet.

- When the Pirate Party was included in the analysis, Twitter mentions becomes a terrible predictor of election results.

- Using nonrepresentative big data sources to do out-of-sample generalizations can go very wrong: lots of nonrepresentative data is still nonrepresentative.

- To conclude, many big data sources are not representative samples from some well-defined population. For questions that require generalizing results from the sample to the population from which it was drawn, this is a serious problem.

- 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.

7. Drifting

- Longitudinal data are very important for studying change. In order to reliably measure change, however, the measurement system itself must be stable: “if you want to measure change, don’t change the measure”.

- Unfortunately, many big data systems (especially business systems) are changing all the time, a process of drift.

- In particular, these systems change in three main ways: population drift (change in who is using them), behavioral drift (change in how people are using them), and system drift (change in the system itself).

8. Algorithmically confounded

- Although many big data sources are nonreactive, because people are not aware their data are being recorded, researchers should not consider behavior in these online systems to be “naturally occurring.”
- In reality, the digital systems that record behavior are highly engineered to induce specific behaviors such as clicking on ads or posting content.
- The ways that the goals of system designers can introduce patterns into data is called algorithmic confounding.
- Moreover, unlike some of the other problems with digital traces, algorithmic confounding is largely invisible. Dealing with algorithmic confounding is particularly difficult because many features of online systems are proprietary, poorly documented, and constantly changing.

- Rather than thinking of big data sources as observing people in a natural setting, a more apt metaphor is observing people in a casino.
- Casinos are highly engineered environments designed to induce certain behaviors, and a researcher would never expect behavior in a casino to provide an unfettered window into human behavior.
- Of course, you could learn something about human behavior by studying people in casinos, but if you ignored the fact that the data was being created in a casino, you might draw some bad conclusions.
- Algorithmic confounding means that we should be cautious about any claim regarding human behavior that comes from a single digital system, no matter how big.

9. Dirty

- Some researchers believe that big data sources, especially online sources, are pristine because they are collected automatically.

- In fact, people who have worked with big data sources know that they are frequently dirty. That is, they frequently include data that do not reflect real actions of interest to researchers.

- The ultimate source of this difficulty is that many of these big data sources were never intended to be used for research, and so they are not collected, stored, and documented in a way that facilitates data cleaning.

- Moreover, while dirty data that is created unintentionally can be detected by a reasonably careful researcher, there are also some online systems that attract intentional spammers.
- These spammers actively generate fake data, and (often motivated by profit)work very hard to keep their spamming concealed.
- For example, political activity on Twitter seems to include at least some reasonably sophisticated spam, whereby some political causes are intentionally made to look more popular than they actually are.
- Unfortunately, removing this intentional spam can be quite difficult.

10. Sensitive

- Health insurance companies have detailed information about the medical care received by their customers. This information could be used for important research about health, but if it became public, it could potentially lead to emotional harm (e.g., embarrassment) or economic harm (e.g., loss of employment).

- Many other big data sources also have information that is sensitive, which is part of the reason why they are often inaccessible.

- Unfortunately, it turns out to be quite tricky to decide what information is actually sensitive (Ohm, 2015), as was illustrated by the Netflix Prize.

- Moreover, collecting sensitive data without people’s consent raises ethical questions, even if no specific harm is caused.

- Much like watching someone taking a shower without their consent might be considered a violation of that person’s privacy, collecting sensitive information without consent creates potential privacy concerns.

Thanks for the attention

mbetti@unite.it

(Fai clic sulla freccia in modalità Presentazione)