DATA COLLECTION

WEB SCRAPING

WORLD WIDE WEB

The way we **share**, **collect**, and **publish** data has changed over the past few years due to the ubiquity of the *World Wide Web* (WWW).

Private businesses, **government**, and **individual users** are posting and sharing all kinds of data and information.

At every moment, new channels generate vast amounts of data on human behaviour.

WORLD WIDE WEB

There was a time in the recent past where both scarcity and inaccessibility of data was a problem for researchers and decision-makers. That is **emphatically** not the case anymore.

Data abundance carries its own set of problems:

- tangled masses of data;
- traditional data collection methods and classical (small) data analysis techniques may not be sufficient anymore.

DATA SOURCES (TRADE-OFFS)

Automated vs. Traditional

Accuracy vs. Completeness

Coverage vs. Validity

Speed vs. Cost

etc.

WEB DATA SCRAPING EXAMPLE – NEW PHONE

Let's say you want to know what people think of a new phone. Standard approach: market research (e.g. telephone survey, reward system, etc.)

Pitfalls:

- unrepresentative sample: the selected sample might not represent the intended population
- systematic non-response: people who don't like phone surveys might be less (or more) likely to dislike the new phone
- coverage error: people without a landline can't be reached, say
- measurement error: are the survey questions providing suitable info for the problem at hand?

WEB DATA QUALITY – NEW PHONE

These solutions can be **costly**, **time-consuming**, **ineffective**.

Proxies – indicators that are strongly related to the product's popularity, without measuring it directly.

If **popularity** is defined as large groups of people preferring one product over a competitor, then sales statistics on a commercial website may provide a proxy for popularity.

Rankings on Amazon could provide a more **comprehensive** view of the phone market vs. traditional survey.

POTENTIAL ISSUES – NEW PHONE

Representativeness of the **listed products**

- Are all phones listed?
- If not, is it because that website doesn't sell them?
- Is there some other reason?

Representativeness of the **customers**

- Are there specific groups buying/not-buying online products?
- Are there specific groups buying from specific sites?
- Are there specific groups leaving/not-leaving reviews?

Truthfulness of customers and **reliability** of reviews.

IS WEB SCRAPING LEGAL?

Ethical Guidelines:

- Work as transparently as possible
- Document data sources at all time
- Give credit to those who originally collected and published the data
- If you did not collect the information, you probably need permission to reproduce it
- Don’t do anything illegal.

Crawling another company’s information to process and resell it is a common complaint.

IS WEB SCRAPING LEGAL?

What is a spider?

- Programs that graze or crawl the web for information rapidly
- Jumps from one page to another, grabbing the entire page content

Scraping is taking specific information from specific websites (which is the goal): how are these **different**?

"Scraping inherently involves **copying**, and therefore one of the most obvious claims against scrapers is copyright infringement."

FRIENDLY COOPERATION WITH API

Application program interface (API) are sets of routines, protocols, and tools for building software applications.

Many APIs restrict the user to a certain amount of API calls per day (or some other limits).

These limits should be obeyed.

DATA COLLECTION

WEB SCRAPING