Selecting problems, data, and methods

INFO 4940: Advanced NLP for Humanities Research

Selecting good problems

Good taste in problems

The ability to select good problems is perhaps the most important skill in research.

A “good” problem is:

- Important
- Consequential
- Tractable
- Timely?

What do those terms mean?

Important means **important to someone**. Who *specifically* is waiting for the answer to this question?

Consequential means that the results will open up new ground. Good problems lead to good future work.

Tractable means that you can solve the problem given practical constraints (data, time, compute, personnel, etc.)

Timely means that the problem matters right now …

Ask: Why has no one solved this problem and why can we do it?

Finding data

90% of research is data wrangling

Every project needs data.

If there’s a pre-cooked dataset for your exact problem, you probably don’t have a research problem at all.

So … you will either **assemble** a dataset from multiple sources or you will **build** a dataset from scratch. In practice, you’ll probably do some of each.

Data sources

Consult the existing literature for leads. Some defaults:

- **Books**: HathiTrust, Project Gutenberg, Internet Archive
- **Social media**: Web scraping, public scrapes
- **Academic publications**: JSTOR, arXiv, SSRN
- **Demographics**: IPUMS, MPC
- **Economics**: US Federal Reserve and member banks
- **Polling and elections**: Roper Center
- **Social sciences**: CCSS, ICPSR

About Kaggle

There’s nothing wrong with Kaggle, but see the previous statement about pre-cooked datasets and real problems

Selecting methods

Simple ?= better

There is no automatic bonus for using methods that are sophisticated or computationally complex

In fact, well-established methods have many benefits: we know that they work, we know how they work, we know how to interpret their outputs, we know where, when, and how they fail …

But **recent advances in NLP may strongly suggest the use of relatively recent methods**

Learn from examples

Consider the papers you’ve read. They are potential models.

- What methods did they use?
- Why did they use them?
- Could you realistically use the same methods?
- Do you have similar data? Similar compute access?
- Did the authors release code or libraries?
- Are human subjects involved?

About human subjects

Research on people can be ethically fraught. Researchers have developed standard protocols to ensure(?) that people who are the subjects of research are protected and that the benefits of the work outweigh any harms to those subjects.

In general, you need to think about such harms if you are interacting with people in any way. **Is *anything* happening in the lives of your subjects that would not have happened in the absence of your research?**

Post-facto observation is OK

If you do not collect personally identifying information and if the data you collect is obtained by observing public sources absent interaction with participants, you probably don’t need IRB approval.

Surveys *always* require IRB approval. They are also slow, small, and difficult to debias. **Survey methods are probably not appropriate for this class.**