Python Data Collection and Management for Public Policy Research

Day 7: Intro to Data, Data Ethics, Packages and Virtual Environments

Blake Miller[†]

9 July, 2024

[†] Assistant Professor, Department of Methodology, London School of Economics and Political Science (E-mail: b.a.miller@lse.ac.uk)

Agenda for Today

- What is Data?
    - Data Formats
    - Data Types
    - Data Storage
- Ethics of Computational Social Science

What is Data?

What is data?

- A representation of the world
- Data necessarily involve subjective decisions on how to:
    - How/what we decide to measure
    - How/what to sample.
    - How/what to share.
    - How to store.
- Storing data will necessarily involve some information loss.

Data Structures

- Tabular data: data in rectangular form, with rows and columns.
- Time series data: observations indexed in time order, often used in event analysis, and trend analysis.
- Graph data: Data representing relationships between entities (social networks, and citation networks)
- Hierarchical data: Data organized in a tree-like structure.

Tabular Data

Common file formats for storing tabular data:

- Comma- or tab-separated values (.csv, .tsv)
  - Each line is an observation
  - Variables are separated by a comma or tab
  - Free, wide support
- Proprietary data formats (.dta, .xlsx, etc.)
  - Difficult to read without closed-source software (Stata, Microsoft Excel).
  - Want to avoid to facilitate replication of our research!

Tabular Data

Relational databases (e.g., SQL)

- Data organized into tables (e.g., author, article, newspaper, etc. for a database of newspaper articles)
- Tables are related through "keys" (e.g., articles written by the same article will have a numeric key indicating the author record in the "author" table)
- Allows for fast retrieval of data from large datasets

Building on Data Storage Types

- Learn SQL (watch here)
- Use SQLite in Python (watch here)

Ethics of Computational Social Science (CSS)

Why are we learning about ethics in a Python course?

- Computational methods are powerful tools, misusing these tools can harm people.
- Many harms from misuse of data are unintentional, awareness is key!
- Goals and agendas of funding agencies, corporations do not align with research stakeholders.

Ethical Challenges Faced by CSS

Challenges faced by CSS [Leslie, 2023]:

- Treatment of research subjects
- Impacts of CSS research on affected individuals and communities
- Quality of CSS research and to its epistemological status
- Research integrity
- Research equity

Treatment of Research Subjects

We should aim to treat research subjects with:

1. Respect
   - Expectation of Privacy: Subjects may not expect their data will be used for research.
   - Personal Autonomy: Subjects may not want to be a part of our research.
2. Justice
   - Risks and benefits of research participation should not be unevenly distributed across groups (age, race, disability, gender, sexual orientation, etc.).
3. Care
   - Risks to subjects should be minimized.
   - Benefits to subjects should be maximized.

Treatment of Research Subjects: Anonymization

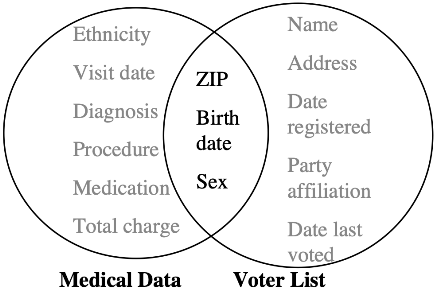


Figure 1 Linking to re-identify data

Source: [Sweeney, 2002]

- To protect subjects, it is good practice to remove personally identifiable information (PII)
- But is this enough? There is still a risk of re-identification through data linkage.

Treatment of Research Subjects: Harm Minimization Strategies

- Obscure/remove personally identifiable information (PII).
- Aggregate data to less specific units.
- Apply differential privacy: introduce noise to protect data while retaining usefulness.
- Obtain informed consent:
    1. Explain study's purpose and procedures.
    2. Encourage questions; ensure clear responses.
    3. Secure explicit consent, freely given.
    4. Confirm participants' right to withdraw anytime.

Impacts on Individuals and Communities

Consider individuals and communities as stakeholders in research:

- Think of who is likely to benefit from research (stakeholders). What do they care about?
- Interests and values of subjects/stakeholders are often not considered by researchers.
- Do funding agencies/corporations and subjects/stakeholders have mismatched agendas?
- Just allocation of risks and benefits of research.

Data Quality

- Challenges with algorithmic influences on data collection.
  - Companies like Meta, Google and ByteDance use algorithms to target delivery of content for engagement.
  - How does this affect the authenticity of social phenomena captured?
- The illusion of data veracity due to large volume of data.
  - Misconception that large data sets are inherently representative or accurate.
  - Overreliance on big data can obscure the need for robust methodological rigor and validation.

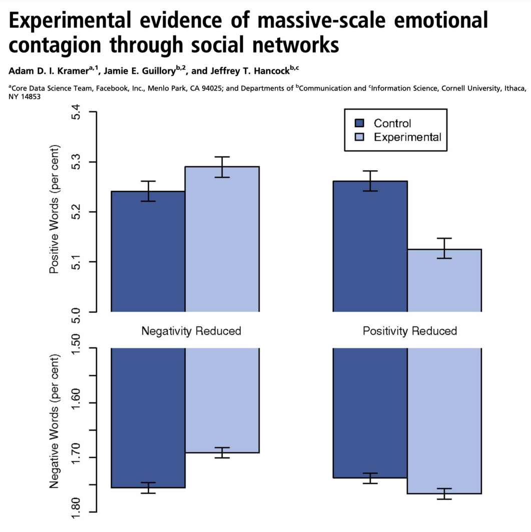
Research Integrity

- Asymmetrical resources and influence:
  - Disparities in access to data and computational resources can skew research outcomes.
  - Potential for dominant stakeholders to dictate research agendas and priorities.
- Dependence on corporations for resources and data.
  - Conflicts of interest may arise when corporate interests drive research directions.
  - Ethical dilemmas in maintaining research independence and integrity.

Research Equity

- Reinforcement of digital divides and data inequities.
  - Research driven by data predominantly collected from more privileged or accessible groups.
  - Potential to overlook marginalized populations, thus perpetuating inequality.
- Aggregation biases mask subgroup differences.
  - Generalized findings can obscure significant variations and perpetuate stereotypes.
  - Risk of policy and interventions failing to address or even exacerbating subgroup vulnerabilities.
- Global inequalities affect data sharing and collaboration.
  - Power imbalances between high- and low-resource settings can lead to exploitative data practices.
  - Inequitable distribution of research benefits and burdens across global divides.

Example 1: Emotional Contagion



Experimental evidence of massive-scale emotional contagion through social networks

Adam D. I. Kramer[a,1], Jamie E. Guillory[b,2], and Jeffrey T. Hancock[b,c]

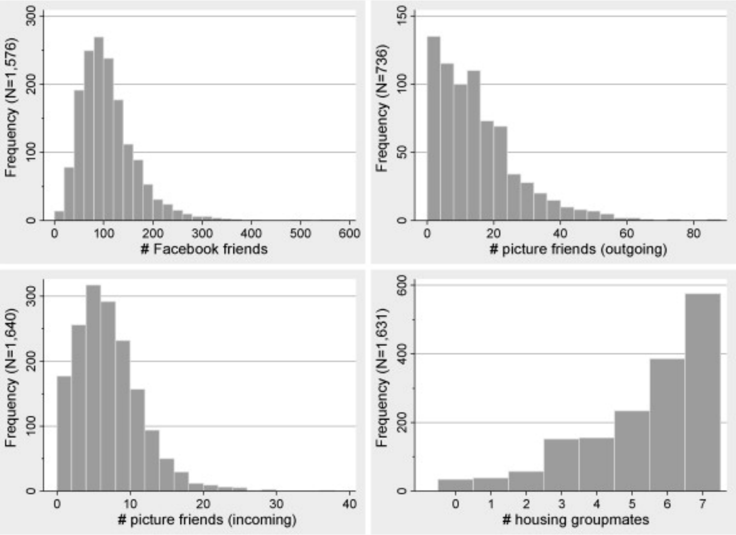
[a]Core Data Science Team, Facebook, Inc., Menlo Park, CA 94025; and Departments of [b]Communication and [c]Information Science, Cornell University, Ithaca, NY 14853

Source: [Kramer et al., 2014]

Summary and Ethical Concerns: Emotional Contagion

- **Study Overview:**
  - Researchers manipulated the news feeds of Facebook users to test if emotional states can be transferred to others via emotional contagion.
  - Participants were not informed they were part of an experiment.
- **Ethical Issues:**
  - Lack of informed consent.
  - Psychological manipulation without users' knowledge.
  - Potential emotional harm to participants.

Example 2: Tastes, Ties, and Time

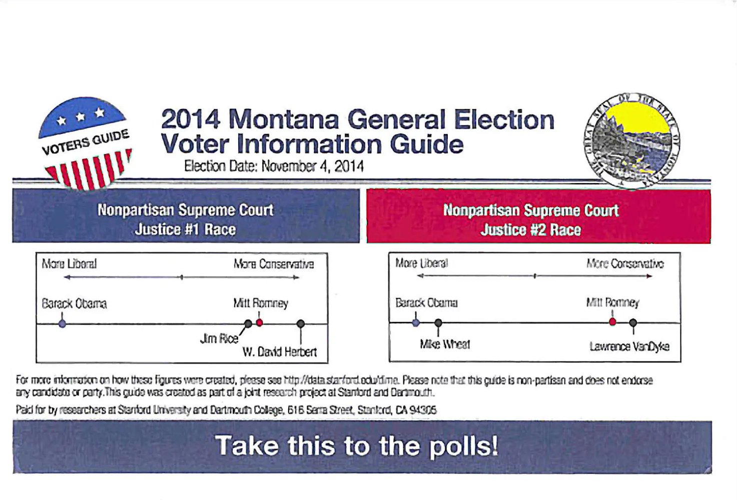


Source: [Lewis et al., 2008]

Summary and Ethical Concerns: Tastes, Ties, and Time

- **Study Overview:**
  - Used Facebook data to analyze the relationship between online behavior and offline social networks.
- **Ethical Issues:**
  - Privacy and data security.
  - Consent process and the extent to which participants were aware of the data usage.
  - Potential risks to participants' privacy given the sensitive nature of social network data.

Example 3: Montana Mail Study



VOTERS GUIDE

2014 Montana General Election Voter Information Guide

Election Date: November 4, 2014

Nonpartisan Supreme Court Justice #1 Race

More Liberal — More Conservative

Barack Obama
Mitt Romney
Jim Rice
W. David Herbert

Nonpartisan Supreme Court Justice #2 Race

More Liberal — More Conservative

Barack Obama
Mitt Romney
Mike Wheat
Lawrence VanDyke

For more information on how these figures were created, please see http://data.stanford.edu/dime. Please note that this guide is non-partisan and does not endorse any candidate or party.This guide was created as part of a joint research project at Stanford and Dartmouth.

Paid for by researchers at Stanford University and Dartmouth College, 616 Serra Street, Stanford, CA 94305

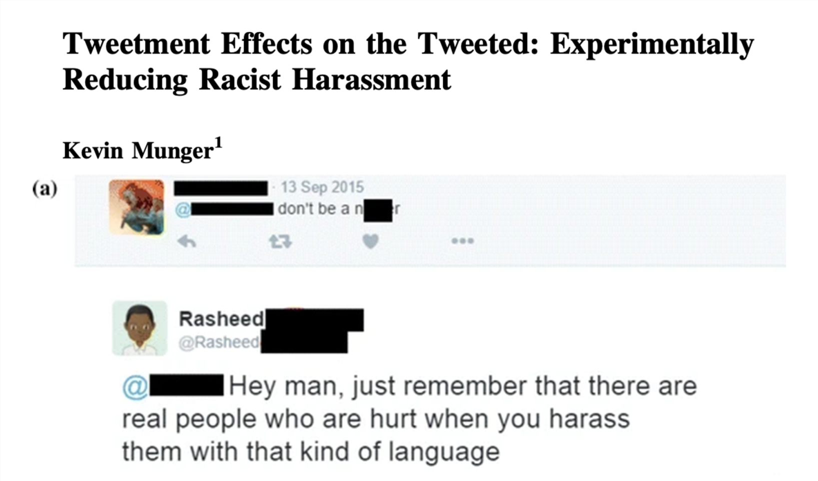
Take this to the polls!

Source: New York Times

Summary and Ethical Concerns: Montana Mail Study

- **Study Overview:**
  - Researchers sent political mailers to Montana voters, resembling official state election guides, to study political behavior.
- **Ethical Issues:**
  - Deception and misrepresentation.
  - Interference in a real election process without proper oversight.
  - Potential to influence voter behavior and outcomes.

Example 4: Tweetment Effects on the Tweeted



Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment

Kevin Munger[1]

(a)

[redacted] · 13 Sep 2015

@[redacted] don't be a n[redacted]r

Rasheed[redacted]

@Rasheed[redacted]

@[redacted] Hey man, just remember that there are real people who are hurt when you harass them with that kind of language

Source: [Munger, 2017]

Summary and Ethical Concerns: Tweetment Effects on the Tweeted

- **Study Overview:**
    - Examined the impact of automated counter-speech on racist Twitter users by sending messages from bot accounts.
- **Ethical Issues (?):**
    - Deception and manipulation.
    - Psychological impacts on participants.
    - Consent of targeted users.
    - But does the positive normative effect matter?

References i

Kramer, A. D., Guillory, J. E., and Hancock, J. T. (2014).
Experimental evidence of massive-scale emotional contagion through social networks.
Proceedings of the National academy of Sciences of the United States of America, 111(24):8788.

Leslie, D. (2023).
The ethics of computational social science.
In *Handbook of Computational Social Science for Policy*, pages 57–104. Springer International Publishing Cham.

References ii

- Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., and Christakis, N. (2008).
  Tastes, ties, and time: A new social network dataset using facebook. com.
  Social networks, 30(4):330–342.

- Munger, K. (2017).
  Tweetment effects on the tweeted: Experimentally reducing racist harassment.
  Political Behavior, 39:629–649.

References iii

Sweeney, L. (2002).
k-anonymity: A model for protecting privacy.
International journal of uncertainty, fuzziness and knowledge-based systems, 10(05):557–570.