Big Social Data: Approaches and Issues

Jared Edgerton

Framing the Course

This course examines how social data are:

- Created
- Collected
- Curated
- Linked
- Validated
- Shared

At scale, and under real research constraints.

What Counts as Social Data

Social data include information about (or arising from):

- Human interaction
- Institutions and organizations
- Digital platforms
- Text, networks, spatial traces, and sensors

Often at scales that strain conventional workflows.

Data as a Research Object

Core premise:

- Data are **constructed**, not found
- Measurement choices are theoretical choices
- Collection decisions shape downstream inference

Data pipelines are part of the substantive argument.

How the Course Is Organized

Each class session follows a consistent structure built around real research practice.

The goal is to connect:

- Tools
- Judgment
- Substantive inference

Rather than treating methods in isolation.

Component 1: Research Workflow Talk

Each session begins with:

- A research presentation by an outside speaker
- Focus on data generation and workflow, not just results
- Attention to choices, tradeoffs, and constraints

The emphasis is on *how* research is done.

Component 2: Coding Lab

The core of each session is a live coding lab:

- Instructor-prepared demo code
- Student-led walkthrough
- Live execution and debugging
- Emphasis on rerunnable pipelines

Code is treated as a shared research artifact.

Component 3: Application Paper

Each session closes with:

- Discussion of a social-science paper using the week's tools
- Focus on data construction, validity, and limitations
- Translation between method and substantive claim

The goal is critical evaluation, not replication.

Research Pipelines We Emphasize

End-to-end workflows:

- Acquisition
- Storage
- Cleaning
- Linkage
- Documentation
- Validation

With explicit provenance at each step.

Tooling Is Methodology

Tools shape what is easy, hard, or invisible:

- Programming languages
- Version control
- Databases
- Containers
- Automation

Methodological commitments are embedded in tooling.

Validation and Failure Modes

Modern data work requires:

- Sanity checks
- Audits and spot checks
- Inter-coder reliability
- Monitoring drift and bias

Validation is treated as a first-class task.

Human-in-the-Loop Workflows

Machine assistance does not remove judgment:

- LLMs for extraction and labeling
- Humans for validation and adjudication
- Documentation of uncertainty and failure modes

The focus is augmentation, not replacement.

What You Should Be Able to Do

By the end of the course, you should be able to:

- Design a defensible data pipeline
- Evaluate data quality and bias
- Reproduce and audit workflows
- Communicate limitations clearly

Discussion

- Where do you think bias enters most often?
- Which steps of a pipeline feel least transparent?
- What makes a data product trustworthy?