Computational & AI-Assisted Methods for Social Sciences

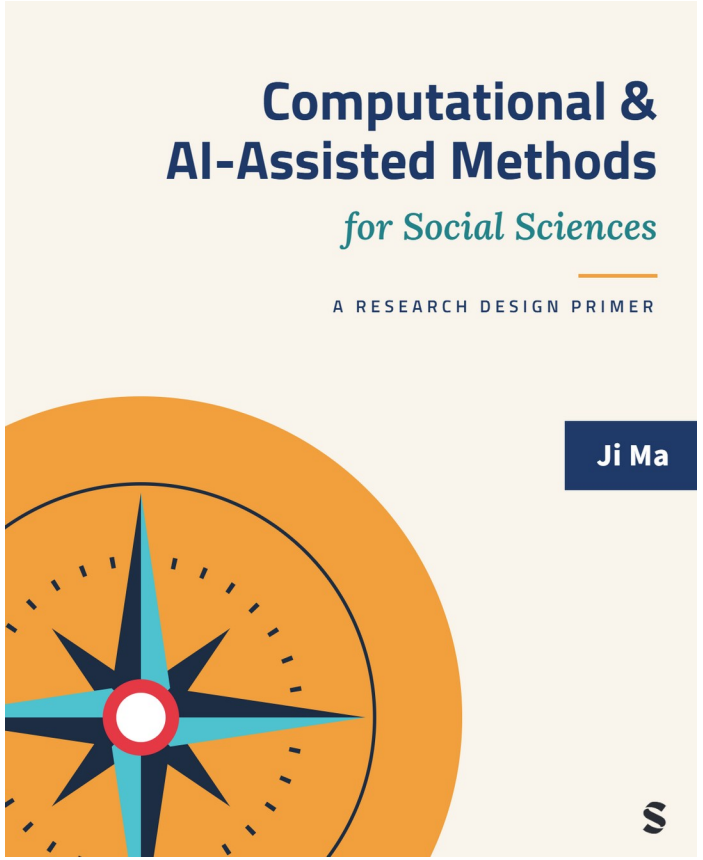
A Research Design Primer

Chapter 3

Data as Relational and Process

Ji Ma

cssprimer.org

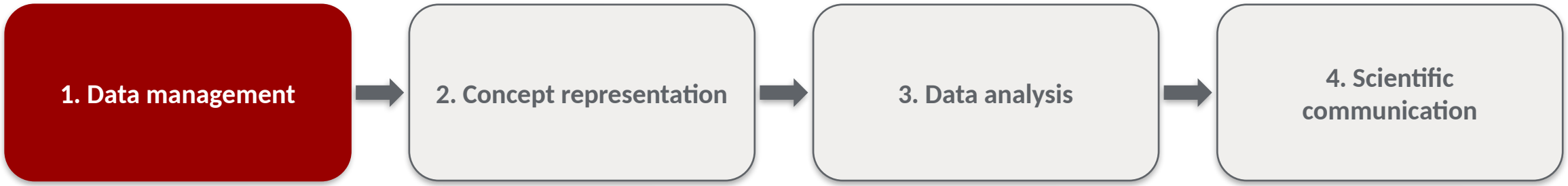


Learning objectives

By the end of this chapter, you can:

- Understand different theoretical perspectives on data: representational, relational, and process
- Recognize the importance of data source, structure, and standardization in social science research
- Use the 3S framework to evaluate data management practices in empirical studies

Stage 1 of 4: data management



This chapter and the next develop the first stage: organizing, storing, and validating research data.

Takeaway: Every later stage inherits the decisions made here.

§3.1

Nature of data: representation, relation, and process

What data is — a window, a web of connections, or a trail of decisions.

The representational view: data as a window

- Data treated as reflecting the social world “out there”
- Survey responses read as a straightforward portrait of public opinion
- Big data extends the logic: more traces, fuller picture

The relational view: data as interconnected points

- Context shapes the data and is shaped by them (Leonelli 2020)
- A post's meaning includes retweets, likes, timing, geography, ties
- Not a static mirror: an ecosystem of mutual influence

Takeaway: Analyzing points without their relations discards part of the phenomenon.

The process view: data carry their own history

§3.1

- Collection, cleaning, and transformation decisions are part of the data's meaning
- Platform prompts, missing-data handling, anonymization all shape the record
- Boyd: “enriched evidence” assembled from multiple lines of evidence

Takeaway: Ask not only what the data show but how they came to show it.

Three views converge on three questions: the 3S

- **Source**: where the data come from and how you acquired them
- **Structure**: how the data are organized
- **Standardization**: whether the workflow around them is repeatable

§3.2

Source: repurposing and acquisition

The expanding universe of usable data — and what using it costs.

What counts as data keeps expanding

- Digitized archives, platform logs, sensors, administrative records
- Born-digital versus produced: digitization, scraping, APIs, surveys, hand coding
- Each acquisition route shapes missingness, measurement error, and ethics

Repurposing: research from data made for something else

- Most digital traces were recorded for commerce or navigation, not research
- Azucar: social media footprints predict Big Five traits, correlations 0.29–0.40
- Gebru: 50M Street View images; sedans predict Democrat (88%), pickups Republican (82%)

Repurposing reaches analog sources too

- OCR converts books, newspapers, manuscripts, epitaphs into machine-readable text
- China Biographical Database: funerary biographies and poems become structured data
- The pipeline: OCR, manual transcription, computational parsing, human review

What repurposing costs: scale, validity, equity

- Scale: millions of entries exceed conventional tools; “data fishing” without theory
- Validity: curated online selves; benchmarks that inherit self-report bias
- Equity and privacy: unequal access skews samples; reidentification persists

Takeaway: Every repurposed source imports the biases of its original purpose.

Classify a source before you commit to it

§3.2

Axis	Values	Cost it predicts	Question to ask
Format	non-digital → unstructured → featurized	Conversion labor; whose measurement decisions are baked in	Who turned this into variables, and for what purpose?
Security	published → controlled → confidential	Permission: what you may analyze and share	Can I publish enough for someone else to check this?
Interface	compiled → API → scraped	Reproducibility: same request, same data later?	If I ran this query next year, would I get this corpus back?

↗ Data-management tools · cssprimer.org/tools/data

Takeaway: Three independent axes, three different currencies of cost.

The axes vary independently: four worked cases

§3.2

- OpenAlex: featurized, published, API plus bulk — cheap on all three
- District administrative records: featurized and compiled, but confidential
- Digitized newspaper archive: heavy labor, light permissions
- Scraped platform data: unstructured and volatile, terms unresolved

Takeaway: The classification tells you where effort will go before you invest weeks.

§3.3

Structure: tidy and relational design

Two principles that impose order — and when to stop imposing it.

Two principles impose order on diverse data

Tidy data (Wickham)

- Each variable in its own column
- Each observation in its own row
- Each unit of analysis in its own table

Relational database

- Tables linked by primary and foreign keys
- Less duplication, graceful updates
- Faster queries, better consistency

Takeaway: Tidy within each table; relational across tables.

The untidy table: one wide sheet mixes units

- Each new tweet adds columns; the table grows sideways
- Gender and age repeat on every tweet — redundancy invites inconsistency
- Violates all three tidy principles at once

Takeaway: A table that mixes tweets and users answers questions about neither cleanly.

The tidy fix: Tweets and Users, joined by a key

- Tweets table: tweet ID, timestamp, sentiment, text
- Users table: demographics stored once per user
- Join on User_ID to ask: do older users tweet more positively?
- Normalization minimizes redundancy and protects integrity

Takeaway: The join replaces repetition; the key preserves the connection.

RICF lessons: serve the research, not the engineering

- Design general purpose and flexible before questions are fixed
- Center the units of analysis most scholars will need
- Collaborate across disciplines; a database matters only if used
- Open data work deserves academic reward
- Data creators see first what data can — and cannot — reveal

Let AI draft the schema; audit it yourself

- Describe your sources and units in plain language; get candidate tables and keys in seconds
- Drafting is where the model's usefulness ends
- Audit: defensible units of analysis? keys that identify *your* records? design serving *your* questions?

Takeaway: Draft-then-audit is the pattern; Chapter 4 practices it on an ERD.

§3.4

Workflow standardization

Research as a repeatable script, not a heroic one-time effort.

Why ad hoc methods break down

- The data: large, messy, multi-source, continuously updated
- The methods: multi-stage pipelines — crawl, clean, merge, model, visualize
- One missed step or parameter derails the entire project

Takeaway: A workflow is a step-by-step script coordinating the whole procedure (Goble 2009).

Four payoffs of a standardized workflow

- **Effectiveness**: automation frees attention for interpretation and theory
- **Reproducibility**: every phase documented, retraceable, verifiable
- **Reusability**: scripts and intermediates transfer to new projects
- **Traceability**: the provenance of every figure can be pinpointed

Takeaway: Standardization is not overhead; it keeps a growing project correct.

§3.5

Cheap data, costly validation

The 3S covers origin, organization, and process. None of it guarantees correctness.

Acquiring data has never been easier. Ensuring its accuracy and credibility remains a costly, labor-intensive process.

A “universe” of observations is not safety

- All posts on a platform: which universe, representing which population?
- Selection mechanisms decide who appears in the data at all
- Uncertainty includes coverage error, selection bias, measurement error — not just sampling error

Takeaway: Complete data on the wrong universe still answers the wrong question.

Gold mining: sift for value, not for purity

- Discard sand when extraction costs exceed the gold's worth
- Kept nuggets are not pure; they pass a quality threshold
- Even “sand” records convert to “gold” through standardization, correction, imputation

Validate a random sample to estimate error rates

- First learn the data's history — its process nature
- A standard proportion formula sizes the check: N = 10,000 → about 370 records
- Categorize the failures, then correct, remove, or impute

Three recurring problems, three remedies (Ma 2020)

- Incorrect values: logical rules — totals must equal the sum of components
- Missing values: categorize plausible reasons; infer from related records where possible
- Inconsistent reliability: Chow test for structural breaks; split the analysis if found

Or: focus on a cleaner, representative subset

§3.5

- 344 policy journals varied widely in record and journal quality (Ma 2025)
- Top 30 journals: over 70% of policy citations, two-thirds of scholarly citations
- Inclusion criteria dropped records that made key variables incomputable

Takeaway: Target a promising patch of ground rather than panning every inch of sand.

Robustness tests: where bias can enter (Ma 2025)

Research stage	Possible causes of bias
Data quality	(1) Missing data by year; (2) Missing institutional affiliation data
Research design	(3) Heterogeneity of team composition; (4) Heterogeneity of policy citations
Statistical methods	(5) Model selection; (6) Model specification

Takeaway: Robustness tests do not rescue bad data; they show what a conclusion can bear.

LLMs scale the checks humans cannot

- Zero-shot: describe the anomaly in the prompt; no examples needed
- Few-shot: a handful of annotated cases sharpens precision
- Tasks: irregular formatting, error detection and correction, cross-referencing external records

Human-in-the-loop turns validation into estimation

§3.5

- 200 flagged entries → hand-check 20 → 18 confirmed: an estimated error rate
- Structured error biases estimates and falsely narrows confidence intervals (Rister et al. 2025)
- Protocol: define the check, write a promptbook, audit before scaling, probe structured error, decide

Why big data works: precision against random noise

- Observed value = signal + random noise + systematic error
- SE = s/√N: idiosyncratic errors cancel as N grows
- Caveats: dependent observations shrink effective N; precision is not external validity

Confidently wrong: what large N cannot fix

Random noise

- Varies unpredictably across observations
- Cancels out as N grows
- Sampling distribution narrows onto the truth

Systematic error

- Correlated with the phenomenon itself
- More biased observations repeat the same error
- Distribution narrows onto the wrong value — displaced by bias b

Takeaway: Huge N makes tiny biases look highly significant; read effect sizes, not p-values.

Use scale to diagnose and correct bias

§3.5

- Granular subgroup checks against external benchmarks like the census
- Reweight toward the target population: post-stratification
- Panel data: within-unit changes difference out stable bias
- Drifting bias survives differencing — stress-test the data-generating process

Takeaway: Big data's best use is not shrinking standard errors but interrogating how the data were made.

Practice: audit two studies with the 3S

- Pick two empirical studies that interest you
- Source: where did the data come from, and how were they acquired?
- Structure: how were the data organized — and was that structure defended?
- Standardization: could you retrace the workflow from what is reported?

You will need

- The CSS Empirical Studies Database, filtered to data-management practice, at cssprimer.org/ch03
- Two studies from your own field

Deliverable

A one-page annotated bibliography critiquing each study's data management in terms of source, structure, and standardization.

↗ **Studies filtered to data management** · cssprimer.org/ch03/

Discussion

- Pick a dataset you have used: what assumptions does treating it as a straightforward representation impose? What does a process view make visible?
- For that dataset, name its source, structure, and standardization. Where do you expect the largest threats to validity?
- Hand your project to a collaborator six months from now: what must be documented for them to reproduce your results?
- You suspect systematic error: how would you combine sampling checks, a cleaner subset, and robustness tests to diagnose it?

Key concepts from Chapter 3

Representational view: Data as a snapshot or mirror of the real world, emphasizing direct measurement

Relational view: Data as interconnected points, shaped by and shaping their context

Process view: Data as the dynamic product of collection, cleaning, and transformation decisions

3S framework: Evaluate any dataset by its source, structure, and workflow standardization

Data repurposing: Using data collected for one goal to investigate a new research question

Tidy data: Each variable a column, each observation a row, each unit its own table

Workflow standardization: A repeatable task sequence ensuring traceability, reproducibility, reuse

Robustness test: Checking whether findings hold under alternative assumptions, samples, and models

Materials on the companion site

Studies filtered to data management cssprimer.org/ch03/

The Practice assignment's slice of the studies browser, pre-filtered to this stage

Data-management tools cssprimer.org/tools/data

The maintained tool list for gathering, storing, and documenting data

CSS Empirical Studies Database cssprimer.org/studies/

The full database, when students want to look beyond this stage

Where to go next

- Read: Chapter 4 — build your literature corpus and apply the 3S hands-on
- Browse: the chapter hub — studies database, slides, and references at cssprimer.org/ch03
- Try: classify one data source from your field on the format, security, and interface axes