Computational & AI-Assisted Methods for Social Sciences

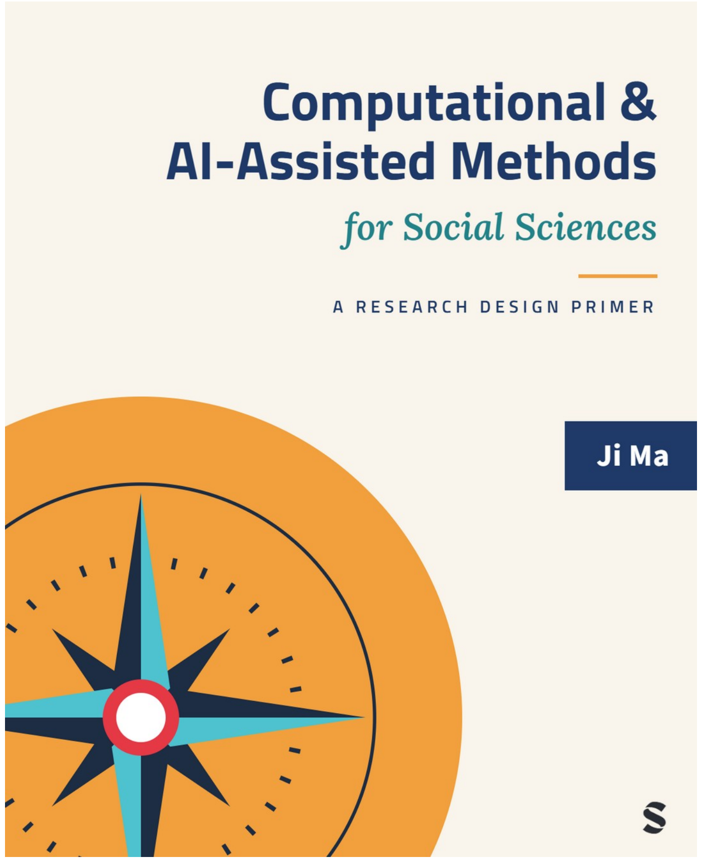
A Research Design Primer

Chapter 1

Introduction

Ji Ma

cssprimer.org



Learning objectives

By the end of this chapter, you can:

- Recognize how data-driven and theory-driven approaches complement each other
- Apply the four stages of empirical research to evaluate any method
- Locate your preferred roles within the CSS landscape
- Assess how AI technologies may reshape social science research

CSS is a methodology, not a buzzword

- Named in 2009; still sometimes dismissed as hype
- The skepticism has a real source: tools built in computer science, questions posed in social science
- This book's answer: view every method through a research design lens

Takeaway: The mismatch between tools and questions is what a research design lens repairs.

§1.1

“End of theory,” or new opportunities for theory?

What the big-data provocation got right, and what it got wrong.

The Fourth Paradigm and the end-of-theory provocation

- Anderson (2008): the data deluge makes hypothesize-test-conclude obsolete
- Hey et al.: a “Fourth Paradigm” of data-intensive science
- Patterns first, theory maybe later, was the promise

Takeaway: The provocation is worth taking seriously before rejecting it.

Prediction successes that fed the excitement

- Tweet frequency outperformed market-based box-office forecasts
- Twitter mood tracked stock market trends
- Google Flu Trends matched CDC surveillance with far less lag

Takeaway: Digital traces revealed patterns beyond the reach of traditional instruments.

Why pure prediction breaks down

- Overfitting; performance decays on unseen data and new contexts
- Patterns without explanatory value (Kitchin 2014)
- Bonikowski: richer data demand *more* theoretical reasoning, not less

Takeaway: A method is only as good as the theoretical insights it lets us generate.

Integration, not replacement, is the productive path

- Computational grounded theory: machine learning plus close qualitative reading (Nelson 2020)
- Voter-behavior predictions aligned with political theory (Theocharis & Jungherr)
- Narrative-driven frameworks grounding prediction in behavioral theory

Takeaway: Blending prediction with theory both improves forecasts and deepens explanation.

§1.2

Bridging “computational” and “social”

A definition, and the bridge the book is built on.

What makes a method computational?

- The book's test: inference that could not run without writing and executing code
- Regression in SPSS is computer-assisted, not computational
- Specifying the procedure in code buys iteration, non-linearity, high dimensionality

Takeaway: Using a computer is not the test; specifying the procedure in code is.

“The best way to create this powerful hybrid is not to focus on abstract social theory or fancy machine learning. The best place to start is research design.”

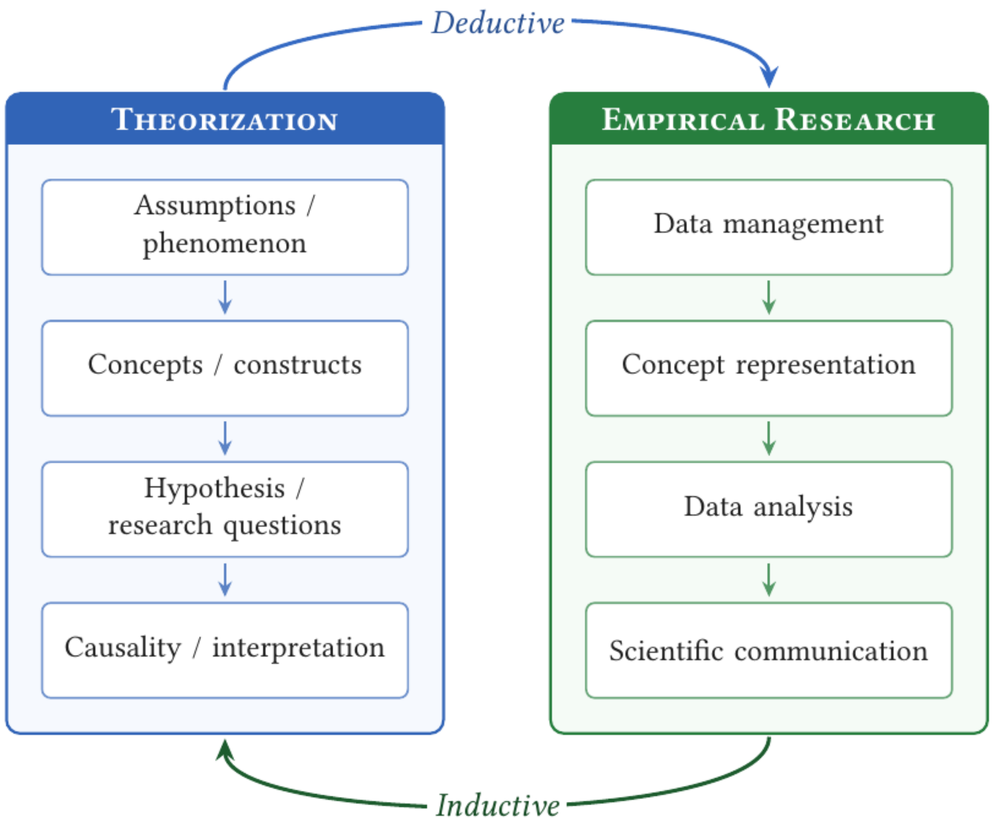
Matthew Salganik, Bit by Bit (2017)

§1.3

Research design: the backbone of methods

Theorization and empirical research, and the four stages inside the empirical half.

Every empirical study has two interlocking components

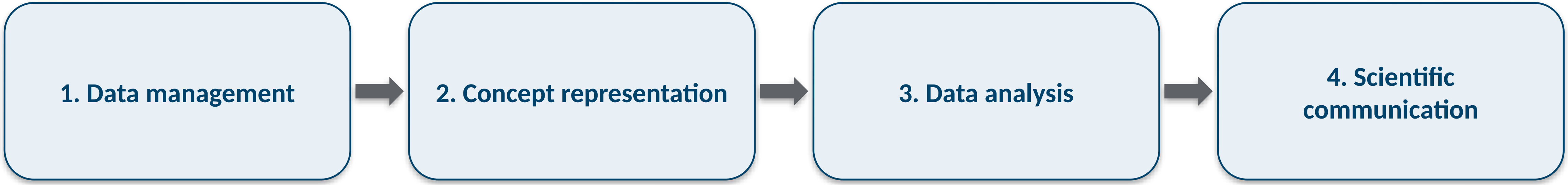


Theorization develops concepts and their relations; empirical research confronts them with evidence.

Adapted from Shoemaker et al. (2003) and Ragin (2011) by Ma (2023).

Takeaway: Theory guides empirical inquiry; empirical findings reshape theory.

The four stages of empirical research



The backbone of the book: each part develops one stage, in theory and in practice.

Takeaway: Any computational method ultimately serves one or more of these stages.

Data management: source, structure, standardization

- Source: repurposed digital footprints, administrative records, multimedia
- Structure: relational tables to JSON to raw text and images; tidy data principles
- Standardization: documented, reproducible workflows with traceable lineage

Takeaway: Repurposed data must be contextualized before they can answer research questions.

Concept representation: measures are theory-informed proxies

- Featurization turns raw records into socially meaningful variables
- Auxiliary instruments in between: networks, document-topic matrices
- Deductive coding, inductive discovery, or an integrative loop

Takeaway: A measure is not the concept itself; different goals justify different proxies.

Data analysis: capacities and epistemic purposes

- New capacities: fast iteration, non-linear relationships, high-dimensional data
- Same foundations: probability, hypothesis testing, regression baselines
- Four purposes: descriptive, explanatory, predictive, integrative

Takeaway: The four purposes are not a hierarchy; each is valuable matched to the right question.

Scientific communication: making work inspectable

- Sharing data and code; documenting workflows
- Visualization for high-dimensional results, including interactive forms
- Not mere dissemination: a social process of scrutiny and critique

Takeaway: Stochastic algorithms and API dependencies make reproducibility a deliberate effort.

§1.4

Comparing empirical approaches

Experiments, surveys, and computational methods through the same four-stage lens.

Three approaches, four stages

Approach	Data management	Concept representation	Data analysis	Communication
Experiment	Structured tabular files	Treatments and instruments	Design-based inference; regression	Tables and figures
Survey	Structured tabular files	Questionnaire items and scales	Descriptive and regression models	Tables and figures
Computational	Large-scale digital traces; text, networks, images	Automated measurement and coding	High-dimensional models; prediction	Reproducible code; interactive visuals

↗ CSS Empirical Studies Database · cssprimer.org/studies/

Takeaway: The same design questions arise everywhere; the answers differ by approach.

The comparison is a deliberate oversimplification

- Pairwise-comparison experiments collect high-dimensional behavioral data
- Multidimensional scaling and PCA long predate CSS
- Every approach keeps evolving; the boundaries are porous

Takeaway: High-dimensional thinking is older than computational social science.

Integration in practice: advocacy organizations on social media

- Automated text analysis of cognitive and emotional language at scale
- Surveys captured organization and audience context
- Each method supplied what the other lacked

Takeaway: Combining approaches at different stages yields more robust designs.

§1.5

"Know thyself": find your sweet spot in CSS

Three roles, usually overlapping, always collaborating.

Three roles in the CSS landscape

Role	Focus	Example work
Algorithmic methodologist	Creates and tests algorithms for social research	Structural topic model; community-detection modularity
Applied methodologist	Validates existing methods in a substantive domain	Benchmarking embeddings on political corpora; fine-tuned classifiers
Computational social scientist	Answers social science questions with computational tools	Word embeddings mapping the cultural meaning of class

Takeaway: Same field, three centers of gravity: build methods, validate methods, use methods.

The roles overlap, and that is the point

- Most projects need collaboration across all three roles
- Individuals move between roles over a career
- This book's vantage: a computational social scientist often working as applied methodologist

Takeaway: Choose a home base, not a cage.

§1.6

Doing social science research in the age of AI

What AI can assist, what it cannot replace, and how to use it responsibly.

What LLMs could already assist by 2023

Brainstorming, writing, and data reformatting rated most useful; the frontier has moved since.

TABLE 2
SUMMARY OF LLM CAPABILITIES AND RATING OF USEFULNESS

Category	Task	Usefulness
Ideation and Feedback	Brainstorming	●
	Feedback	◐
	Providing counterarguments	◐
Writing	Synthesizing text	●
	Editing text	●
	Evaluating text	●
	Generating catchy titles & headlines	●
	Generating tweets to promote a paper	●
Background Research	Summarizing Text	●
	Literature Research	○
	Formatting References	●
	Explaining Concepts	◐
Coding	Writing code	◐
	Explaining code	◐
	Translating code	●
	Debugging code	◐
Data Analysis	Creating figures	◐
	Extracting data from text	●
	Reformatting data	●
	Classifying and scoring text	◐
	Extracting sentiment	◐
	Simulating human subjects	◐
Math	Setting up models	◐
	Deriving equations	○
	Explaining models	◐

Notes: The third column reports my subjective rating of LLM capabilities as of September 2023:
○: Experimental; results are inconsistent and require significant human oversight.
◐: Useful; requires oversight but will likely save you time.
●: Highly useful; incorporating this into your workflow will save you time.

Source: Korinek (2023), Table 2, Journal of Economic Literature.

Takeaway: Capability keeps moving into tasks previously rated “not useful.”

Five task families where AI assists research

- Idea generation and hypothesis formation
- Research design and experimentation
- Data collection and preprocessing
- Simulation and modeling, including LLM-based agents
- Writing and dissemination

Takeaway: From ideation to dissemination, assistance is available at every step.

The expanding map of LLMs in scientific research

A field survey of LLMs across the scientific workflow, as of late 2024.

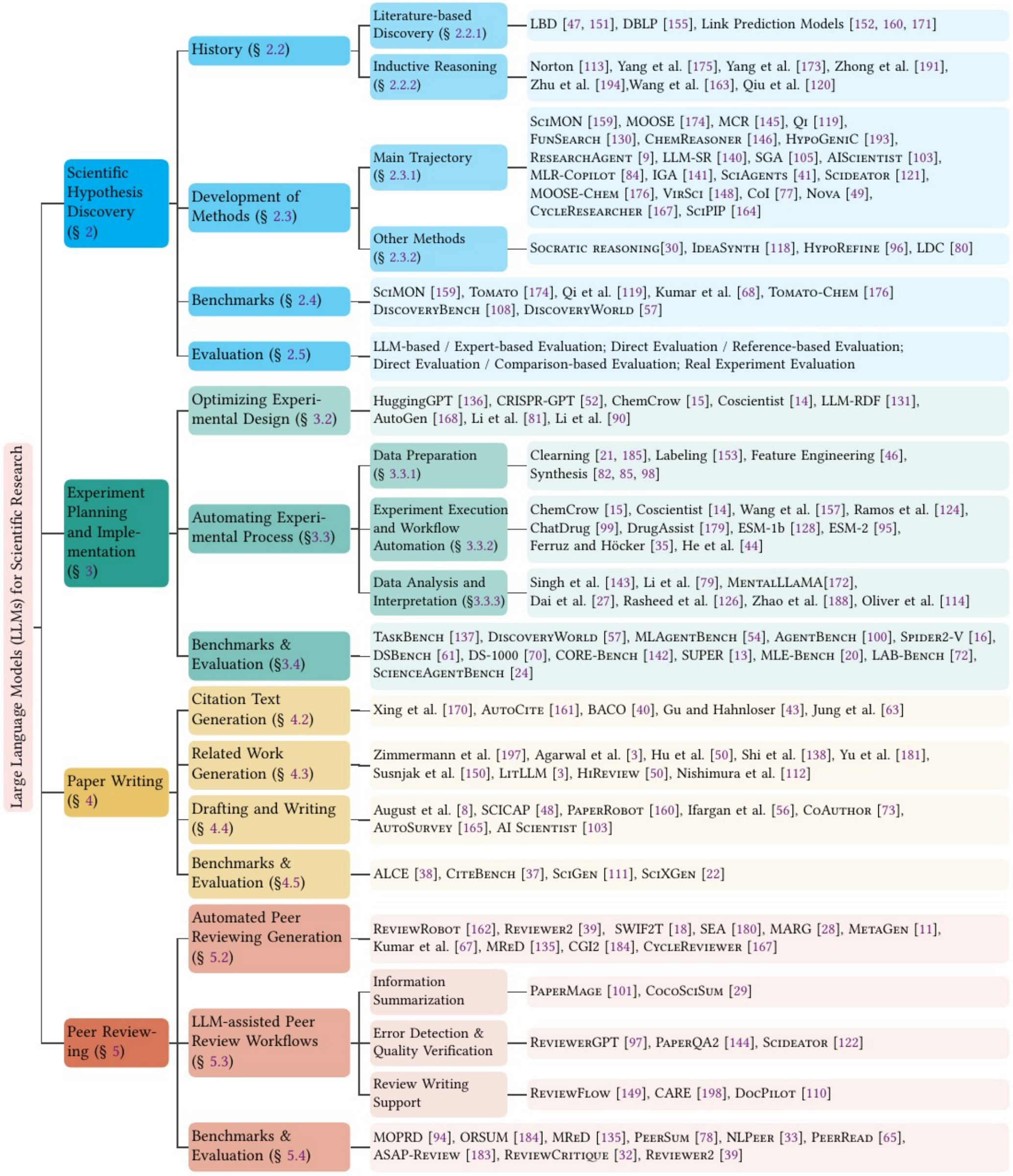


Fig. 2. The main content flow and categorization of this survey.

Source: Luo et al. (2025), Fig. 2, arXiv preprint. Included for instructional use.

Takeaway: Expect this landscape to change in months, not years.

Three functional roles of generative AI in research

- Measurement tool: turning text, images, audio into variables
- Prediction and simulation: synthetic respondents and agents
- Research environment: participants, coders, and institutions use AI too

Takeaway: The third role changes what counts as a human baseline.

Try out: three basic moves

- Summarize and interrogate a paper; demand supporting quotations
- Draft a literature search; run the strings yourself
- Critique a design; treat objections as hypotheses, not verdicts

↗ Getting started · cssprimer.org/setup/

Takeaway: Every move pairs with a verification habit; the habit is the skill.

What current AI still lacks for social inquiry

- Causal models that support explanation, not just pattern recognition
- Grounding in intuitive theories of the physical and social world
- Compositional learning that generalizes to genuinely new tasks
- Model collapse: training on synthetic output degrades diversity

Takeaway: Assigning a human-like identity to an LLM does not produce human-like behavior.

Ethical stakes are already visible

- "Junk science": convincing outputs without empirical grounding
- Contaminated data: 37% of an MTurk sample said yes to "are you an LLM?"
- Synthetic respondents pass attention checks at near-perfect rates
- Opaque models undermine reproducibility; biased training data propagate

Takeaway: Treat unsupervised online responses as potentially mixed-origin data.

Practical takeaways for working with AI

- Curate a verified knowledge base; constrain the model to it
- Keep intellectual ownership of questions, frameworks, designs
- Maintain a promptbook: the prompt is a research instrument
- Verify outputs: hand-audit samples, test every code snippet
- Patience: reflection must keep pace with generation

Takeaway: These five habits become a full project framework in Chapter 2.

What remains distinctly human

- Three pillars: domain knowledge, analytical reasoning, innovation
- AI streamlines the first; the other two stay ours to sharpen
- “Taste”: knowing when a result is elegant but wrong

Takeaway: Delegate execution and drafting; protect framing, mechanism, and inference.

Discussion

- How does theorization guide the collection and interpretation of data in a study you admire?
- Why treat the four stages as interconnected rather than isolated steps?
- Which of the three CSS roles fits you today, and what would move you toward another?
- What ethical considerations apply when AI tools enter your own research tasks?

Key concepts from Chapter 1

Theorization: Developing concepts and their interrelations; grounds research in established knowledge

Empirical research: Confronting concepts with evidence to support or refute the framework

Four stages: Data management, concept representation, data analysis, scientific communication

Computational method: Inference that could not be carried out without writing and executing code

Algorithmic methodologist: Creates, adapts, and tests algorithms for social science research

Applied methodologist: Validates existing computational methods within a substantive domain

Computational social scientist: Uses computational tools to answer questions of real-world significance

Human-like machines: Would need causal models, grounded learning, and compositional generalization

Materials on the companion site

Getting started cssprimer.org/setup/
Environment setup and the pages that close the first-week programming gap

Materials by chapter cssprimer.org/chapters/
One hub per chapter: the page to open at the start of each unit

CSS Empirical Studies Database cssprimer.org/studies/
122 annotated studies, filterable by design stage

Library cssprimer.org/library/
Primers and tool guides organized by function rather than by chapter

Where to go next

- Read: Chapter 2, the AI-assisted research framework (rules, memory, knowledge)
- Set up: the Getting Started guide and project skeleton at cssprimer.org/setup
- Try: the three basic moves on a paper from your own field