

An Ongoing Effort Towards an Integrative Computational Research Framework for Quantitative Social Science (Manuscript Incomplete)

Wanhong HUANG^{1*}

^{1*}Independent Researcher.

Corresponding author(s). E-mail(s):
huangwanhong.g.official@gmail.com;

Abstract

Quantitative social science research increasingly demands sophisticated computational capabilities. However, existing research practices remain constrained by a structural misalignment between the epistemic organization of the research life-cycle and available computational tools, resulting in fragmented tool and data ecosystems, limited workflow integration, steep learning curves, growing data complexity, and persistent reproducibility challenges.

This paper addresses the following research question: How can the epistemic structure of quantitative social science research be systematically encoded into an integrative, modular, and reproducible computational framework? To answer this question, we present a comprehensive software research framework built around a unified, modular architecture that supports the full research lifecycle, including literature review and meta-analysis, experimental and survey research, document analysis, and secondary data analysis.

The core idea of the proposed framework is to introduce a shared and integrative structural layer that mediates between heterogeneous data sources, analytical methods, and research outputs. By defining consistent intermediate representations of research data and analytical states, the framework enables different components of the workflow to interoperate without requiring direct coupling between specific tools. This approach supports extensible pipelines that can accommodate multiple data modalities, analytical environments, and target outputs such as figures, tables, and LaTeX manuscripts, while reducing redundancy, manual transformation, and workflow fragmentation.

We present the system’s preliminary design and implementation, demonstrate its use through a representative case study, and discuss current limitations and future development directions. The source code is publicly available at: <https://github.com/wwwanhonghuang/Quantitative-Social-Science-Research-framework>.

Keywords: Quantitative Social Science, Computational Social Science, Research Framework, Research Workflow Automation, Reproducible Research

1	Contents	
2	1 Introduction	4
3	2 A Comprehensive Overview of Quantitative Social Science Research	8
4	2.1 Overview of Quantitative Social Science Research Lifecycle	9
5	2.2 Methodologies of Quantitative Social Science Research Lifecycle	11
6	2.2.1 Question Discovery	11
7	2.2.2 Research Design Phase	11
8	2.2.3 Data Collection Phase	12
9	2.2.4 Data Processing Phase	12
10	2.2.5 Data Analysis Phase	13
11	2.2.6 Dissemination Phase	16
12	2.3 Research Schemes in Quantitative Social Science	17
13	2.3.1 Literature Review and Meta-Analysis	17
14	2.3.2 Experimental Research	18
15	2.3.3 Survey Research	20
16	2.3.4 Document and Content Analysis	23
17	2.3.5 Secondary Data Analysis	25
18	2.4 Data Analysis Methods	26
19	2.4.1 Statistics-Based Methods	26
20	2.4.2 Information Theory-Based Methods	28
21	2.4.3 Machine Learning and Deep Learning Methods	29
22	2.4.4 Graph Theory and Network Analysis	30
23	2.4.5 Interdisciplinary-Integrated Methods	30
24	2.5 Visualization and Reporting Tools	32
25	2.6 Summary: Toward an Integrative Computational Research Framework	33
26	3 Framework Architecture	35
27	3.1 Architectural Design Overview	35
28	4 Illustrative Case Studies	40
29	4.1 Illustrative Example: Automated Data Analysis and LaTeX Code Gen-	
30	eration	40
31	4.1.1 Project Initialization and Structure	40
32	4.1.2 Data Processing Configuration	41

33	4.1.3 Statistical Analysis Configuration	41
34	4.1.4 Automated Execution and Output Generation	43
35	5 Related Work	44
36	5.1 Statistical Software Packages	45
37	5.2 Qualitative and Mixed-Methods Tools	45
38	5.3 Computational Social Science Platforms	46
39	5.4 Workflow Management Systems	47
40	5.5 LLM-Augmented Research Tools	48
41	6 Discussion	49
42	6.1 Key Findings and Contributions	49
43	6.2 Limitations	49
44	6.3 Ethical Considerations	50
45	6.4 Future Directions	50
46	7 Conclusion	51

1 Introduction

Quantitative social science has emerged as one of the central methodological approaches in contemporary social research, providing systematic empirical tools for investigating social phenomena at scale [1, 2]. Rooted historically in empiricist and positivist traditions within the development of sociology and related social sciences, quantitative social analysis emphasizes observation, measurement, and statistical inference as primary means of generating knowledge about social structures and processes. At the same time, its epistemological foundations have long been subject to sustained philosophical critique. Classical challenges to empiricism and positivism have questioned whether general causal regularities or future outcomes can be legitimately inferred from historical observations alone, most notably through the problem of induction articulated by David Hume. These critiques highlight the inherent limits of moving from empirical regularities to claims of necessity or causal certainty. Nevertheless, despite such foundational concerns, quantitative approaches have remained indispensable in modern social science practice.

Against this background, research practices in quantitative social science have continued to evolve over several decades. In addition to classical approaches centered on structured surveys, controlled experiments, and manual content analysis, quantitative social science since the latter half of the twentieth century has made extensive use of statistical modeling, secondary data analysis, and systematic exploitation of administrative and official statistical data. These approaches are themselves methodologically sophisticated, involving complex inferential frameworks, model-based reasoning, and nontrivial data-processing pipelines, and they constitute an essential foundation of modern quantitative social research.

In the contemporary digital era, these established quantitative traditions have become increasingly intertwined with computational techniques. On the one hand, researchers now routinely draw on large-scale digital trace data generated by online platforms, such as social media activity records, online interaction logs, and digitized archival sources [1, 3, 4]. On the other hand, social relations, interaction patterns, and collective behaviors are frequently represented in network form, allowing researchers to model relational structures among individuals, organizations, or other social entities and to analyze their structural properties [5]. At the same time, multi-modal datasets that combine text, images, and behavioral signals have become increasingly common [3], while methods drawn from machine learning and artificial intelligence are employed to support large-scale data processing, pattern discovery, and automated analysis [6, 7].

These computationally oriented approaches have substantially expanded the range of social phenomena that can be investigated in a systematic manner. In certain domains, they involve network-based modeling frameworks and, in more specialized contexts, diffusion process models to characterize how information, behaviors, or innovations propagate through social systems [8]. However, the expansion of research scope has been accompanied by a marked increase in the technical and organizational demands placed on researchers. Beyond mastering diverse data sources and analytical methods, researchers must coordinate heterogeneous tools, manage complex workflows, and integrate activities across multiple stages of the research process. As a result,

contemporary quantitative social science increasingly confronts challenges that are as much infrastructural and organizational as they are methodological [9, 10].

A critical challenge facing quantitative social scientists is the fragmentation of the research tool ecosystem. A typical research project may require researchers to navigate Web of Science or Scopus for literature search, Zotero or Mendeley for reference management, Qualtrics or SurveyMonkey for survey administration, Python libraries such as pandas and scikit-learn for data analysis [11, 12], R statistical packages for hypothesis testing [13], and Gephi or Cytoscape for network visualization [14, 15].

These tools typically operate independently, each relying on distinct data formats, interfaces, and workflow conventions. As a result, researchers are frequently required to manually export, transform, and re-import data when moving between stages of the research process. This fragmentation introduces multiple points at which errors may arise and substantially increases the cognitive effort required to manage and coordinate research activities [16].

Beyond efficiency concerns, the lack of shared data representations and standardized interfaces across tools makes it difficult to maintain a coherent record of analytical decisions throughout the research workflow. Intermediate results, parameter choices, and processing steps are often scattered across disconnected systems, complicating the construction of transparent and reproducible analysis pipelines [17, 18]. While tools such as LaTeX or Microsoft Word remain essential for final manuscript preparation, the underlying challenge lies in systematically linking upstream data collection and analysis processes to these downstream research outputs. We refer to this challenge as **Data and Tool Fragmentation**, emphasizing the lack of shared data representations and standardized interfaces across heterogeneous research tools.

The second major challenge is the limited integration between different phases of the research lifecycle. Literature review, data collection, analysis, and reporting are often treated as disconnected stages with minimal information flow between them [19]. Insights from literature review that should inform variable operationalization are lost when transitioning to data collection. Exploratory findings from initial analyses that could reshape research questions rarely feed back into systematic literature searches. Visualization and statistical results must be manually copied into manuscripts, creating versioning problems when analyses are updated. This disconnect (1) reduces efficiency, and (2) impedes the iterative nature of scientific inquiry, where later stages should naturally inform revisions to earlier conceptualizations, and makes it more difficult to maintain reproducible research workflows across different phases of the research lifecycle [20]. We refer to this challenge as **Research Lifecycle Discontinuity**, highlighting the absence of formalized and integrated pipelines across different stages of the research process.

Third, the difficulty in reproducing quantitative social science research has become a pressing concern for the field’s credibility and Third, difficulties in reproducing quantitative social science research have emerged as a major structural concern for the credibility of the field and the accumulation of cumulative knowledge [21, 22]. Despite growing awareness of the reproducibility crisis and increasing journal requirements for data and code sharing [23, 24], such measures have often proven insufficient to address reproduction failures in practice. We refer to this challenge as **Structural**

Reproducibility Deficits, to stress that reproducibility failures arise from structural properties of research workflows rather than from isolated technical issues.

A key reason is that reproducibility is not attributable to any single technical component, but rather arises from the overall structure of the research workflow. As discussed above, fragmentation across tools and data, together with the lack of formal integration between different phases of the research lifecycle, makes it difficult to represent, record, and track analytical processes in a unified manner. In the absence of shared data representations, explicit workflow conventions, and stable interfaces, critical analytical decisions, intermediate states, and transformation steps are frequently embedded in implementation-specific details rather than expressed as explicit, inspectable research objects.

Under such conditions, even when identical data, algorithms, and computational settings are used, reproducing research outcomes often depends on implicit understanding and experiential reconstruction of the original analysis process, rather than on systematic reproduction grounded in clearly defined structures and contracts. Reproducibility thus becomes contingent on individual researchers' experience and careful execution, instead of being an intrinsic property supported by research infrastructure. This structural deficiency contributes to highly variable reproducibility practices across quantitative social science research [25].

Fourth, researchers face steep learning curves for increasingly complex emerging technologies and analytical methods. Modern quantitative social science demands proficiency not only in traditional statistical techniques but also in programming languages (Python, R), version control systems (Git), machine learning frameworks (scikit-learn, PyTorch, TensorFlow) [26, 27], natural language processing pipelines (spaCy, NLTK, Hugging Face transformers) [28, 29], and network analysis libraries (NetworkX, igraph) [30, 31]. Large language models have emerged as potentially powerful tools for literature synthesis, qualitative coding, and content analysis [32, 33], yet their integration into rigorous social science workflows remains methodologically unclear [34]. Each new technology requires substantial time investment to master, diverting attention from substantive research questions and creating barriers to entry for graduate students and early-career researchers [35]. We refer to this challenge as **Steep Learning Curves from Insufficient Abstraction**, underscoring how the lack of shared abstractions shifts integration and maintenance burdens onto individual researchers.

Fifth, researchers must contend with growing data complexity and scale. Social science datasets now routinely involve diverse modalities including structured survey responses, unstructured social media posts, network edge lists, time-stamped behavioral logs, images, videos, and geospatial coordinates [36, 37]. Data sources have proliferated beyond traditional surveys to encompass web APIs (Twitter, Reddit, Facebook), web scraping, sensor networks, administrative records, and digital archives [4]. Acquisition methods range from one-time downloads to real-time streaming, from manual data entry to automated pipelines. Managing this heterogeneity requires sophisticated data integration, harmonization, and quality control procedures that most researchers are ill-equipped to implement [38]. The sheer volume of data (social media datasets with millions of posts, network graphs with billions of edges,

longitudinal panels tracking thousands of individuals over decades) exceeds the capabilities of desktop statistical software and demands distributed computing approaches unfamiliar to most social scientists [39]. We refer to this challenge as **Data Heterogeneity and Scale**, reflecting the combined effects of increasing data diversity and volume on research workflows.

To address the challenges of **Data and Tool Fragmentation**, **Research Lifecycle Discontinuity**, **Structural Reproducibility Deficits**, **Steep Learning Curves from Insufficient Abstraction**, and **Data Heterogeneity and Scale** defined above, this paper presents the design and preliminary implementation of an integrative computational research framework for quantitative social science.

Our approach is guided by a set of structural design principles that directly respond to these challenges. First, to mitigate data and tool fragmentation (Problem **Data and Tool Fragmentation**), the framework introduces unified data representations and standardized processing interfaces, enabling heterogeneous tools and analytical components to interoperate within a coherent workflow structure.

Second, to address discontinuities across the research lifecycle (Problem **Research Lifecycle Discontinuity**), we emphasize the formalization of end-to-end research workflows, providing infrastructural support for linking literature review, data collection, analysis, and reporting within a single, traceable pipeline.

Third, to reduce the burden imposed by steep learning curves (Problem **Steep Learning Curves from Insufficient Abstraction**), the framework abstracts and integrates commonly used analytical models and methods spanning traditional statistical analysis as well as emerging interdisciplinary techniques, while retaining extensibility through modular components and well-defined interfaces. This design allows researchers to leverage shared abstractions without constraining the incorporation of specialized or advanced methods.

Fourth, to cope with increasing data heterogeneity and scale (Problem **Data Heterogeneity and Scale**), the framework enables researchers to use and define explicit data, processing, and storage contracts, supporting flexible yet standardized handling of large-scale, multi-modal datasets. In addition, the framework provides unified and systematic interfaces for literature search and data collection across major scholarly repositories, including PubMed, arXiv, OSF, CrossRef, SocArXiv, and Zenodo.

Taken together, these design choices aim to improve the reproducibility of quantitative social science research (Problem **Structural Reproducibility Deficits**) by shifting reproducibility from an implicit expectation placed on individual researchers to a structural property supported by shared representations, formalized workflows, and explicit interfaces.

This work makes five contributions.

First, we articulate a modular research framework centered on shared data representations and standardized processing interfaces, with the aim of enabling heterogeneous analytical tools to interoperate within a coherent research workflow. This approach allows researchers to continue using established external tools, such as LaTeX for manuscript preparation, while benefiting from more integrated data handling and transformation processes.

226 Second, we outline a framework-level approach to structuring research workflows
227 across different stages of the research lifecycle. By making the relationships between
228 literature collection, data acquisition, analysis, and the production of formal research
229 outputs explicit, this work seeks to support more systematic and traceable research
230 processes.

231 Third, we advance a perspective in which reproducibility is treated as a property
232 emerging from the structure of research workflows rather than as an implicit require-
233 ment placed on individual researchers. Emphasizing shared representations, explicit
234 workflow conventions, and stable interfaces, the proposed framework aims to provide
235 infrastructural support for reproducible analytical practices.

236 Fourth, we introduce a layered abstraction strategy that integrates commonly used
237 analytical models and methods within a unified interface, spanning traditional statis-
238 tical analysis as well as selected advanced interdisciplinary computational approaches.
239 In addition, the framework allows large language models to be incorporated as auxil-
240 iary modules within the research workflow, supporting tasks such as literature review,
241 preliminary coding, and conceptual structure organization. This design seeks to reduce
242 learning and integration costs while preserving methodological control.

243 Fifth, we describe mechanisms that allow researchers to use and define explicit con-
244 tracts for data representation, processing, and storage. These mechanisms are intended
245 to support flexible yet principled handling of increasingly heterogeneous and large-
246 scale datasets, and are complemented by unified interfaces for systematic literature
247 search and data collection across multiple scholarly repositories.

248 An open-source prototype implementation, together with documentation and
249 illustrative examples, is made available to facilitate critical evaluation, community
250 feedback, and collaborative extension of the proposed framework.

251 It is important to acknowledge the preliminary nature of this work. We view this
252 paper as presenting both completed work and a research agenda, openly documenting
253 our design rationale and implementation progress to invite community engagement
254 and collaborative refinement.

255 The remainder of this paper is organized as follows. Section 2 provides a
256 comprehensive overview of quantitative social science research methodologies and pro-
257 cesses, establishing the requirements that our computational framework addresses.
258 Section 3 presents the system architecture and core design principles guiding our
259 implementation.

260 Section 4 demonstrates the framework’s current capabilities through case studies.
261 Section 5 discusses related work and positions our contribution. Section 7 concludes
262 with limitations, ethical considerations, and future development directions.

263 2 A Comprehensive Overview of Quantitative Social 264 Science Research

265 This section provides a systematic overview of quantitative social science research to
266 establish the foundational requirements for our computational framework. We begin
267 by describing the general research lifecycle common to most quantitative studies,

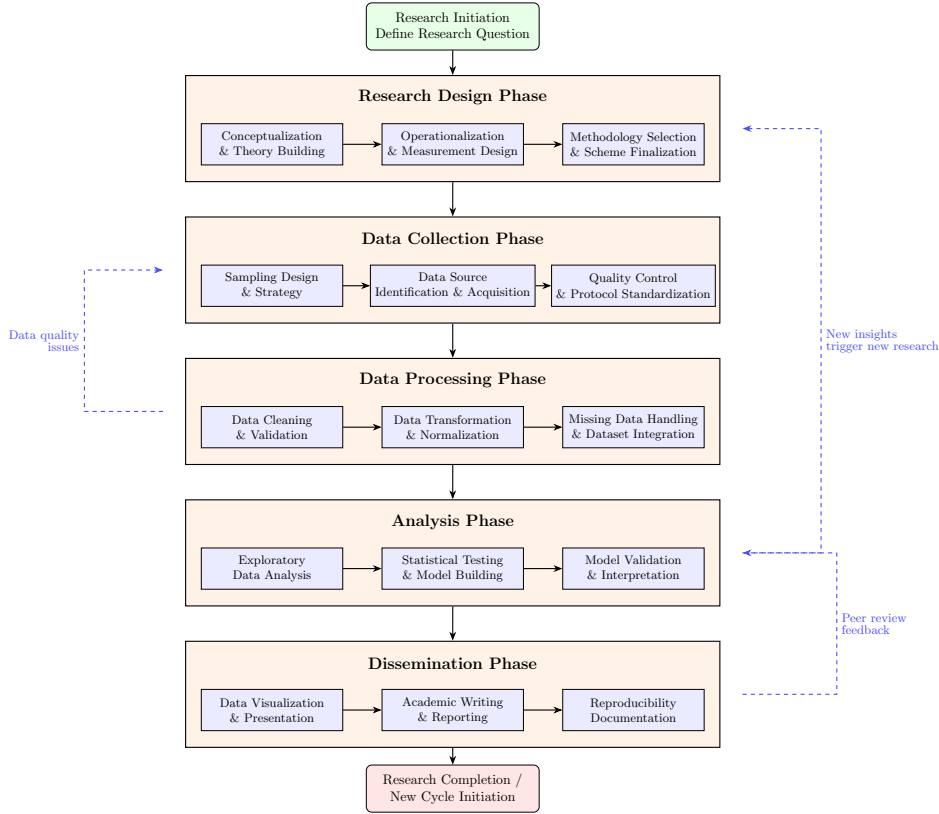


Fig. 1 Quantitative Social Science Research Lifecycle and Iterative Workflow

then detail the major research schemes employed in the field, followed by an examination of analytical methods and dissemination practices. Throughout, we identify computational support needs that inform our computational framework design.

2.1 Overview of Quantitative Social Science Research Lifecycle

Quantitative social science research is typically understood as a structured and iterative research process consisting of multiple interconnected stages that interact and provide feedback to one another in research practice [20, 40].

In this paper, we conceptualize this research process as a research lifecycle comprising five core stages, and explicitly incorporate problem discovery as the first stage of the research lifecycle. This delineation is made because the problem discovery stage plays a critical role in research practice by establishing the research context, motivation, and formalizing research objectives, which will constrain all subsequent research design, data processing, and analytical activities.

Understanding this complete research lifecycle is fundamentally important for designing computational tools that can support the entire research process. Figure 1

283 illustrates the relationships among the various stages in the research lifecycle, as well
284 as the key constituent elements of each stage.

285 In the research design stage, researchers need to identify and define theoretical
286 concepts relevant to the research question. This process is typically accomplished
287 through developing or adopting existing theoretical frameworks. Due to the highly
288 abstract nature of theoretical concepts, subsequent quantitative analysis requires an
289 operationalization process to transform these abstract concepts into measurable and
290 observable variables [41]. Building upon this foundation, researchers further design
291 research protocols for data collection, including measurement instruments, sampling
292 strategies, and data acquisition methods.

293 In the data collection stage, researchers implement the acquisition of empirical data
294 according to the established research design. This process typically requires select-
295 ing appropriate sampling methods and data collection strategies based on the target
296 characteristics of the research task and resource and contextual constraints under real-
297 world conditions. Researchers need to determine specific data sources and conduct
298 actual data collection activities such as questionnaire surveys, experiments, observa-
299 tions, or archival data extraction. This stage often also includes quality control and
300 process standardization to ensure consistency of data across different time points,
301 locations, or personnel [42].

302 The objective of the data processing stage is to prepare appropriate data rep-
303 resentations for subsequent analysis. Researchers typically need to clean, transform,
304 and integrate raw data to address issues such as noise, outliers, missing data, and
305 inconsistent data structures [43]. When research involves multiple data types, or even
306 multimodal data, data integration procedures are also required to merge data from
307 different sources into a consistent analytical dataset.

308 In the data analysis stage, researchers employ various methodological frameworks
309 to extract research findings from the processed data. Modern quantitative social sci-
310 ence, beyond traditional statistical methods, extensively adopts analytical frameworks
311 such as machine learning methods, information-theoretic approaches, graph-theoretic
312 models, and systems theory [44, 45]. This stage typically includes exploratory data
313 analysis, modeling and validation, and interpretation of analytical results within the
314 theoretical context.

315 The final stage of the research lifecycle is the dissemination stage. In this stage,
316 researchers communicate research findings through various forms such as papers,
317 reports, and conference presentations, combining text, visualizations, and other media
318 to suit different audiences [46]. This stage also typically includes systematic documen-
319 tation and organization of the research process and analytical results, such as making
320 data, code, and methodological descriptions publicly available, to support verification
321 and reuse of research results [47].

322 These five stages together constitute an iterative research process characterized
323 by feedback loops and stage-to-stage interaction. Findings from subsequent stages
324 often prompt researchers to return to earlier stages to revise the research question,
325 research design, or data processing strategies. For example, the analysis stage may
326 reveal deficiencies in data collection, preliminary results may trigger theoretical recon-
327 sideration, and feedback obtained during the dissemination process may also generate

new research questions. This iterative characteristic highlights the importance of maintaining coherence across stages and consistency of research products throughout the entire research lifecycle.

2.2 Methodologies of Quantitative Social Science Research Lifecycle

2.2.1 Question Discovery

Problem discovery constitutes a foundational stage in quantitative social science research, serving as the prerequisite for subsequent methodological and analytical decisions. From a methodological perspective, this stage primarily involves two fundamental questions: how research questions are generated, and how to judge whether a research question merits further investigation.

Relevant methodological literature indicates that research questions may originate from researchers' observations of social reality, personal or professional experiences, and systematic reading of existing research literature [48]. These sources are typically integrated through the researcher's theoretical imagination and methodological judgment. Existing research has also proposed a series of criteria for evaluating research questions, including importance, originality, feasibility, and appropriateness.

Among these, importance reflects the potential value of the research question at the levels of theoretical contribution and social significance; originality focuses on its degree of novelty relative to existing research; feasibility concerns whether the research question can be systematically addressed under existing theories, methods, and real-world conditions; appropriateness reflects the degree of alignment between the research question and the researcher's own background, professional capabilities, and available resources.

These considerations indicate that before formally entering quantitative analysis, researchers typically need to invest substantial intellectual and methodological work. Problem discovery is therefore an active, creative, and evaluative stage that largely shapes the scope, direction, and feasibility of the entire research lifecycle.

2.2.2 Research Design Phase

The core task of the research design stage is to transform the research question into an implementable research protocol. In this stage, researchers first need to clarify the research subjects and units of analysis, and formulate the research question formally so that it can be answered through empirical data. This process is typically based on a systematic literature review, which is used to identify relevant theoretical frameworks, existing research conclusions, and methodological approaches that can be drawn upon.

A key component of the research design stage is variable identification and operationalization. Researchers need to identify core concepts involved in the research question and transform abstract concepts into measurable variables or indicators through the operationalization process. Operationalization methods may include adopting existing scales, constructing new measurement indicators, or mapping theoretical concepts to observable objects such as behavioral frequencies, attitude ratings, structural characteristics, or event occurrence patterns. In this process, researchers

370 need to address issues such as construct validity, content validity, and measurement
371 reliability.

372 After variables are clarified, researchers typically formulate research hypotheses
373 based on theoretical derivation, explicitly specifying the expected relationships among
374 variables in a testable form. Research may pursue hypothesis testing or exploratory
375 objectives; in both cases, preliminary conceptions of variable relationships help guide
376 subsequent research design.

377 The research design stage also includes the selection and planning of research
378 schemes. Common schemes include questionnaire survey research, experimental or
379 quasi-experimental research, secondary data analysis, literature and text analysis,
380 comparative research, and mixed-methods research, among others. Different schemes
381 differ significantly in data sources, sample structure, inference approaches, and appli-
382 cable question types. With the development of computational methods, some research
383 also adopts more complex design forms, such as nested designs, multilevel designs, or
384 integrated research protocols combining multiple data sources.

385 Additionally, researchers need to conduct overall planning at this stage for sampling
386 strategies, measurement instruments, data collection procedures, and preliminary ana-
387 lytical methods. These decisions collectively constitute the blueprint for research
388 implementation and directly impact the feasibility and research quality of subsequent
389 stages.

390 2.2.3 Data Collection Phase

391 The data collection stage transforms the research design into concrete empirical data.
392 In this stage, researchers need to select appropriate sampling methods according to
393 the research protocol to ensure sample representativeness of the target population
394 as much as possible under resource constraints. Probability sampling methods (such
395 as simple random sampling, stratified sampling, cluster sampling, and multistage
396 sampling) provide the foundation for statistical inference, while nonprobability sam-
397 pling methods (such as convenience sampling, purposive sampling, quota sampling,
398 and snowball sampling) are commonly used for exploratory research or hard-to-reach
399 research populations.

400 In the actual data collection process, researchers need to attend to various poten-
401 tial sampling biases and systematic errors, such as coverage bias, nonresponse bias,
402 self-selection bias, and measurement error. Some research also employs methods such
403 as weighting adjustments, post-stratification, or sensitivity analysis to quantitatively
404 assess and correct sampling errors.

405 The choice of data sources depends on the research protocol and may include
406 questionnaire platforms, experimental facilities, archival data, administrative records,
407 network interfaces, or sensor systems, among others. The data collection process typi-
408 cally needs to be controlled through standardized procedures, accompanied by quality
409 control measures such as pretesting, real-time validation, and data auditing to reduce
410 human errors and systematic biases.

411 Additionally, the data collection stage requires detailed documentation of the data
412 acquisition process, including sampling decisions, implementation deviations, and field
413 conditions, to support subsequent data interpretation and reproducibility assessment.

2.2.4 Data Processing Phase

The objective of the data processing stage is to transform raw data into data representations suitable for systematic analysis while controlling information loss and structural biases introduced during the processing as much as possible. Researchers typically begin with data cleaning and validation, including correcting entry errors, identifying outliers, handling logical inconsistencies, and verifying whether variable values fall within reasonable ranges [49]. These operations, while procedural in form, often exert substantial influence on subsequent analytical results.

Missing data handling is an important component of the data processing stage [43]. Researchers need not only to identify the proportion and distribution of missing data but also to analyze its underlying mechanisms and select appropriate handling methods accordingly. Different handling strategies involve trade-offs between bias reduction and information retention, with implications for effective sample size and data integrity.

In the process of variable transformation and recoding, information loss control is likewise a key consideration. Discretization of continuous variables, aggregation of time series, simplified representations of text data, and threshold processing of network structures may all sacrifice fine-grained information while improving computational feasibility. Researchers typically need to balance interpretability, computational complexity, and information retention, and when necessary, preserve raw data or intermediate representations to support sensitivity analysis [50].

When research involves multi-source or multimodal data, the data integration process may introduce additional information loss risks, such as temporal asynchrony, inconsistent measurement scales, or semantic mismatches. To mitigate these risks, researchers employ various alignment strategies. For instance, when integrating survey data with administrative records, researchers may use temporal alignment by matching records by date ranges or event sequences, and semantic mapping by standardizing occupation codes or geographic classifications across different coding systems [51]. Metadata management throughout the integration process helps document transformation decisions and ensures traceability. Consistency assessments compare overlapping measurements from different sources or validate integrated variables against known benchmarks to help identify potential integration errors.

Furthermore, in research involving high-dimensional data, unstructured data, or privacy-sensitive data, the data processing stage may include complex operations such as dimensionality reduction, feature extraction, structural reconstruction, or anonymization. Common dimensionality reduction techniques include *principal component analysis* and *factor analysis* [52]. Feature extraction may involve generating text embeddings [53] or extracting image features [54]. Structural reconstruction includes network inference from behavioral data [55]. Anonymization approaches include *k-anonymity* [56] and *differential privacy* mechanisms [57]. While these operations improve analytical feasibility, they often involve irreversible information compression and therefore require clarification of their potential impacts on research conclusions through method selection, parameter control, and robustness testing.

2.2.5 Data Analysis Phase

The data analysis stage aims to extract empirical findings that can answer research questions from processed data through systematic analytical methods. This stage typically begins with exploratory data analysis, which uses descriptive statistics, data visualization, and pattern recognition to help researchers understand data structure, variable distributions, and potential associational relationships, providing a basis for subsequent modeling [58].

In inferential analysis, researchers commonly employ statistical modeling methods to characterize relationships among variables. Regression analysis is widely used to estimate the strength and direction of associations among variables, with specific forms including *ordinary least squares regression*, *logistic regression*, *Poisson regression*, *generalized linear models* [59], *generalized additive models* [60], and *multilevel models* or *hierarchical models* for nested data structures [61]. Beyond simple bivariate or multivariate associations, researchers often need to analyze more complex effect structures. *Mediation analysis* examines whether and how an independent variable affects a dependent variable through intermediate mechanisms, estimating both direct and indirect effects [62, 63]. *Moderation analysis*, also known as *interaction analysis*, investigates whether the relationship between variables varies across levels of a third variable, identifying conditional effects and boundary conditions [64]. Some research also employs *moderated mediation* or *mediated moderation* models to characterize even more complex theoretical relationships [65].

In causal analysis contexts, researchers may employ various identification strategies to isolate causal effects under observational data conditions. *Difference-in-differences* methods compare changes over time between treatment and control groups to account for time-invariant confounders [66]. *Instrumental variable methods* use exogenous variables that affect the treatment but not the outcome directly to address endogeneity bias [67]. *Propensity score methods*, including matching, weighting, and stratification, balance treatment and control groups on observed covariates [68]. *Regression discontinuity designs* exploit threshold-based treatment assignment rules [69]. *Synthetic control methods* construct counterfactual comparisons by combining control units [70]. Panel data methods, such as *fixed effects models* and *first-difference models*, control for time-invariant unobserved heterogeneity [71].

With the development of computational social science, machine learning methods have been widely introduced in the data analysis stage [44, 72]. Supervised learning methods are used for predictive modeling and classification tasks, including *decision trees*, *random forests* [73], *gradient boosting machines* [74], *support vector machines* [75], *neural networks*, and *deep learning models* [76]. Unsupervised learning methods help discover latent patterns and structures in data, including *k-means clustering*, *hierarchical clustering*, *Gaussian mixture models*, and dimensionality reduction techniques such as *principal component analysis*, *t-SNE* [77], and *UMAP* [78]. *Ensemble methods* combine multiple models to improve predictive performance [79]. These methods typically emphasize predictive accuracy, which calls for careful alignment with research objectives regarding interpretability and inference.

In contexts where research subjects possess relational or interactive structures, network analysis methods model and analyze nodes, edges, and overall structural characteristics [55, 80]. Centrality measures characterize node importance, including *degree centrality*, *betweenness centrality*, *closeness centrality*, and *eigenvector centrality* [81]. Community detection algorithms identify cohesive subgroups through approaches such as *modularity optimization* [82], *spectral clustering* [83], and *stochastic block models* [84]. *Network motifs* [85] and *graphlets* [86] characterize local structural patterns. *Exponential random graph models* (ERGMs) [87] and *stochastic actor-oriented models* (SAOMs) [88] enable statistical inference about network formation processes. Dynamic network models capture temporal evolution of relational structures, including discrete-time and continuous-time approaches [89].

For temporally related data, time series analysis methods study patterns of variable changes over time. *Autoregressive models*, *moving average models*, and *autoregressive integrated moving average (ARIMA) models* capture temporal dependencies [90]. *State-space models* and *structural time series models* decompose temporal patterns into trends, seasonal components, and irregular fluctuations [91]. *Vector autoregression (VAR) models* analyze multivariate time series systems [92]. *Event history analysis*, also known as *survival analysis* or *duration analysis*, examines time-to-event processes using methods such as *Kaplan-Meier estimation*, *Cox proportional hazards models*, and parametric survival models [93, 94].

Beyond the above methods, some research employs information-theoretic methods to analyze uncertainty, information transfer, and structural complexity. *Shannon entropy* measures uncertainty in probability distributions [95]. *Mutual information* quantifies statistical dependencies between variables without assuming functional forms [96]. *Transfer entropy* [97] and *Granger causality* [98] assess directional information flow in temporal or spatial systems. Complexity measures, such as *Kolmogorov complexity* and *algorithmic information theory*, characterize structural regularity and randomness [99]. These methods are particularly useful in analyzing high-dimensional, multivariate, or dynamical systems where traditional parametric assumptions may be overly restrictive.

In research involving explicit theoretical structures, researchers may also introduce domain-specific structural models that directly embed theoretical assumptions into model structures. *Agent-based models* simulate individual-level decision rules and emergent collective patterns [100, 101]. *Compartmental models*, such as *SIR models* in epidemiology, represent flow between discrete states [102]. *System dynamics models* use differential equations to represent feedback loops and stock-flow relationships [103, 104]. These structured modeling approaches enable researchers to formalize theoretical mechanisms and explore implications of different theoretical assumptions through simulation and analytical derivation.

Regardless of which analytical method is employed, model validation and robustness testing are important components of the data analysis stage. Researchers typically assess the stability and generalizability of results through *cross-validation*, *out-of-sample testing*, alternative model comparison, parameter sensitivity analysis, or subsample analysis [105]. For causal inference, researchers conduct *placebo tests*, *falsification tests*, and sensitivity analyses for unobserved confounding [106, 107]. The

545 interpretation of analytical results needs to be combined with theoretical background,
546 clarifying the applicable scope of model assumptions, and carefully distinguishing
547 among statistical correlation, predictive capability, and causal interpretation.

548 **2.2.6 Dissemination Phase**

549 The dissemination stage transforms research findings into forms that enable academic
550 and social impact. The primary objective is to communicate research contributions to
551 appropriate audiences while ensuring transparency and reproducibility of the research
552 process.

553 Academic journal publication establishes formal scholarly records and enables peer
554 review for quality assurance. Journal articles typically follow discipline-specific con-
555 ventions in presenting research questions, theoretical frameworks, methods, results,
556 and implications. Researchers select journals based on fit with their research topic,
557 methodological approach, and target readership. The peer review process provides
558 critical feedback and validation of research quality.

559 Conference presentations enable rapid knowledge exchange and direct engagement
560 with research communities. Academic conferences provide opportunities for early-stage
561 feedback, networking, and staying current with emerging research directions. Presen-
562 tation formats include oral presentations, poster sessions, and panel discussions, each
563 suited to different types of content and interaction goals.

564 Preprint releases accelerate scientific communication by making research publicly
565 available before formal peer review. Preprint repositories such as arXiv ¹, SSRN ², and
566 SocArXiv ³ establish priority claims while enabling broader access. This approach is
567 particularly valuable for rapidly evolving research areas or when timely dissemination
568 serves public interest.

569 Policy briefs and stakeholder reports translate research findings into accessi-
570 ble formats for decision-makers and practitioners. These communications emphasize
571 actionable implications while maintaining scientific integrity. Effective policy commu-
572 nication requires adapting technical content to non-specialist audiences, highlighting
573 practical relevance, and acknowledging limitations and uncertainties.

574 Public communication activities, including media engagement, public lectures, and
575 educational outreach, broaden research impact beyond academic communities. These
576 activities fulfill accountability to funding sources and society while contributing to
577 public understanding of social science research.

578 Contemporary dissemination increasingly integrates open science practices. Data
579 sharing through repositories such as ICPSR ⁴, Dataverse ⁵, and OSF ⁶ enables ver-
580 ification and secondary analysis. Code sharing through platforms such as GitHub ⁷

¹<https://arxiv.org>

²<https://www.ssrn.com>

³<https://osf.io/preprints/socarxiv>

⁴<https://www.icpsr.umich.edu>

⁵<https://dataverse.org>

⁶<https://osf.io>

⁷<https://github.com>

and GitLab ⁸ documents analytical procedures and supports computational reproducibility. Computational environment documentation through tools such as Docker ⁹ preserves the complete software stack. Detailed methodological documentation, including protocols, instruments, and decision logs, supports replication and extension of research.

These various dissemination forms serve complementary functions in the research ecosystem. Feedback obtained during dissemination often reveals new questions, methodological refinements, or directions for improvement, thereby driving new research cycles. The choice among dissemination strategies depends on research objectives, target audiences, institutional requirements, and available resources.

2.3 Research Schemes in Quantitative Social Science

Research schemes in this work are comprehensive approaches that structure the entire research process from problem formulation through data collection to analysis and interpretation [40]. Each scheme represents a distinct way of conducting social science inquiry and can employ multiple specific analytical methods. The choice of research scheme depends on the research question, theoretical framework, available resources, and epistemological commitments. Below we describe the major research schemes in quantitative social science, emphasizing their distinct workflows and computational support needs.

2.3.1 Literature Review and Meta-Analysis

Systematic literature reviews and meta-analyses synthesize existing research to identify patterns, gaps, and theoretical frameworks [108]. Systematic literature reviews employ explicit, reproducible methods for searching, screening, and synthesizing literature, distinguishing them from narrative review approaches. Meta-analysis extends systematic review by quantitatively combining effect sizes across studies to estimate overall effects with greater precision than individual studies can provide [109].

As a research scheme in quantitative social science, the literature review lifecycle begins with question discovery, followed by research design or planning, as indicated in Section 2.1. It should be noted that researchers may define new specific research questions beyond the initial ones during the review process. They also establish clear inclusion and exclusion criteria, often using frameworks such as PICO representing Population, Intervention, Comparison, and Outcome, and develop a priori protocols to minimize bias [110]. Search strategies must be comprehensive and reproducible, specifying keywords, Boolean operators, database selections, and search syntax. Advanced techniques include citation chaining and hand-searching key journals. Screening proceeds in stages, beginning with title and abstract review to identify potentially relevant studies, followed by full-text review applying detailed eligibility criteria. Multiple independent reviewers typically conduct screening to ensure reliability.

Data extraction involves systematically coding study characteristics such as design, sample, and setting, along with methods including measurement and analysis, and

⁸<https://gitlab.com>

⁹<https://www.docker.com>

findings covering effect sizes, confidence intervals, and p-values. Standardized extraction forms and codebooks ensure consistency across reviewers. Quality assessment evaluates methodological rigor and risk of bias using validated instruments like the Cochrane Risk of Bias tool or GRADE framework. Synthesis can be narrative, qualitatively describing patterns across studies, or quantitative, involving meta-analysis through effect size aggregation, meta-regression for moderator analysis, and publication bias assessment via funnel plots and statistical tests. Interpretation draws conclusions about the state of knowledge, identifies theoretical and empirical gaps, and provides recommendations for future research. Finally, reporting follows established guidelines such as *PRISMA* or *MOOSE* to ensure transparency and reproducibility.

Literature reviews employ diverse analytical methods beyond simple narrative synthesis. Descriptive statistics summarize study characteristics and sample demographics. Effect size calculation standardizes findings across studies using metrics like Cohen’s d, odds ratios, or correlation coefficients. Meta-regression examines how study characteristics moderate effects, while subgroup analysis compares effects across predefined categories. Publication bias assessment utilizes funnel plots, Egger’s test, trim-and-fill methods, and p-curve analysis. More recently, citation network analysis maps intellectual structures and identifies influential works, and topic modeling using techniques like *Latent Dirichlet Allocation* or *BERTopic* can identify thematic patterns across large literature corpora.

However, researchers conducting systematic reviews face substantial challenges. Information overload continues to worsen as scholarly production accelerates. Selection bias can occur if search strategies miss relevant studies or if eligibility criteria inadvertently exclude important work. Publication bias, which favors statistically significant findings, potentially distorts meta-analytic estimates. Cross-database redundancy requires careful deduplication to avoid counting the same study multiple times. Heterogeneity in study designs, populations, and measures complicates synthesis and limits the applicability of pooled estimates. The time and resource intensity of comprehensive reviews makes them challenging for individual researchers to complete without substantial support.

These challenges create clear computational support needs. Effective tools for literature reviews require automated search across multiple databases with unified query interfaces, intelligent duplicate detection using fuzzy matching and DOI resolution, AI-assisted screening to accelerate title and abstract review while maintaining accuracy, citation network analysis tools to identify seminal works and research communities, and statistical meta-analysis software implementing fixed-effect, random-effects, and mixed-effects models. Advanced features might include automated data extraction from PDFs, inter-rater reliability calculators, and seamless integration with reference management systems. Visualization tools such as PRISMA flow diagrams are also essential for documenting the screening process.

2.3.2 Experimental Research

Experimental research establishes causal relationships through controlled manipulation of independent variables while observing effects on dependent variables [111]. This paradigm emphasizes internal validity through experimental control, randomization,

and comparison groups, providing the strongest evidence for causality by isolating treatment effects from confounding variables.

Experimental design requires selecting an appropriate experiment structure. *Randomized controlled trials*, abbreviated as RCTs, randomly assign participants to treatment and control conditions, establishing the gold standard for causal inference. *Quasi-experiments* lack randomization but employ techniques like matching, regression discontinuity, or difference-in-differences to approximate experimental conditions when randomization is infeasible [112]. *Factorial designs* manipulate multiple factors simultaneously to examine main effects and interactions. *Repeated measures designs* reduce error variance by using participants as their own controls. Planning involves power analysis to determine required sample sizes for detecting effects of theoretical interest [113], developing detailed randomization procedures including simple, block, stratified, or adaptive randomization, and preparing standardized treatment protocols. Implementation requires careful treatment assignment, intervention delivery with fidelity monitoring, appropriate control group management including placebo controls, and blinding procedures to prevent expectancy effects. Measurement includes pre-treatment baseline assessments, manipulation checks to verify treatment mechanisms, and continuous outcome monitoring with quality control.

Beyond simple treatment effect estimation, experimental research employs sophisticated statistical methods for understanding causal mechanisms—techniques valuable across all research paradigms, not merely experiments. *Mediation analysis* examines indirect effects through which independent variables influence outcomes via intermediate variables [62]. For instance, a stress reduction intervention may improve academic performance through decreased anxiety as a mediating mechanism. *Moderation analysis* identifies conditions under which treatment effects vary, revealing for whom or under what circumstances interventions work best. Adjustment methods control for confounding variables that may bias treatment effect estimates, using techniques like ANCOVA, propensity score matching [68], or inverse probability weighting. *Interaction effect analysis* explores synergistic or antagonistic combinations of multiple factors. *Difference-in-differences estimators* leverage pre-post comparisons across treatment and control groups to isolate causal effects [114]. *Instrumental variable methods* address unmeasured confounding when natural experiments or randomization encourage but do not determine treatment assignment. *Regression discontinuity designs* exploit threshold-based treatment assignment to estimate local average treatment effects. These causal analysis frameworks, while particularly powerful in experimental contexts, extend to observational, survey, and qualitative research paradigms when appropriate assumptions hold.

The primary difficulty in experimental research lies in designing studies that adequately represent real-world complexity while maintaining experimental control. Researchers must navigate numerous interacting factors that can influence results. Trade-offs between internal validity referring to control and precision and external validity referring to generalizability require careful consideration [111]—highly controlled laboratory settings may not reflect messy real-world conditions. Ethical constraints limit permissible manipulations, particularly with vulnerable populations. Participant attrition threatens validity if dropout patterns differ between conditions

or correlate with outcomes. Contamination occurs when control group participants receive treatment elements through communication or diffusion. Despite randomization, confounding may persist in small samples or when randomization fails to balance unmeasured variables. Hawthorne effects cause behavioral changes from participants' awareness of being studied. Demand characteristics lead participants to infer study hypotheses and alter responses accordingly. Treatment fidelity varies when interventions cannot be delivered uniformly across sites or providers. Temporal factors like maturation, history effects, or regression to the mean complicate causal inference. Measurement reactivity occurs when repeated assessments themselves influence outcomes. Complex interactions between individual characteristics, contextual factors, and treatment components may require large samples and sophisticated designs to detect. Longitudinal experiments face compounding challenges as time-varying confounders and dynamic treatment regimens introduce additional complexity.

Computational tools support experimental research through randomization algorithms implementing various allocation schemes such as simple, block, stratified, or minimization methods, power analysis calculators for sample size determination considering effect sizes and statistical tests, causal inference estimation methods such as propensity scores, instrumental variables, or regression discontinuity approaches, mediation and moderation analysis frameworks, treatment effect visualization with uncertainty quantification, and compliance tracking systems monitoring adherence and protocol deviations.

2.3.3 Survey Research

Survey research gathers systematic primary data from populations through structured data collection instruments to describe characteristics, attitudes, behaviors, or relationships [115]. This scheme emphasizes representativeness through probability sampling and generalizability through statistical inference. Surveys efficiently collect standardized data from large samples, enabling population estimates and examination of associations between variables.

Survey designs can be classified along multiple dimensions. By temporal structure, *cross-sectional surveys* collect data at a single time point, providing a snapshot of population characteristics but precluding causal inference about change. *Longitudinal surveys* involve repeated measurements over time, enabling analysis of temporal patterns, trends, and causal processes. Longitudinal designs include *trend studies*, which sample different individuals from the same population at multiple time points to track population-level changes; *cohort studies*, which follow a specific cohort such as a birth cohort or graduating class through repeated measurements to examine life course trajectories; and *panel studies*, which repeatedly survey the same individuals, providing the strongest design for analyzing individual-level change and within-person dynamics [116]. Panel studies enable growth curve modeling, fixed effects estimation, and dynamic causal analysis, though they face challenges from panel attrition and conditioning effects from repeated measurement.

Survey data collection has evolved substantially beyond traditional questionnaires to encompass diverse instruments and modalities. Traditional approaches include *structured questionnaires* administered via paper, web, or mobile platforms, and

754 *structured interviews* conducted through telephone, face-to-face, or video formats.
755 Contemporary technological developments have expanded data collection possibilities.
756 *Ecological momentary assessment*, abbreviated as EMA, uses smartphones to collect
757 real-time data in naturalistic settings, reducing recall bias [117]. *Wearable devices*
758 *and sensors* including fitness trackers, smartwatches, and environmental sensors pas-
759 sively collect behavioral and physiological data when integrated into research designs
760 with informed consent. *Digital trace data* from social media, web browsing, or mobile
761 applications can be collected through research partnerships or participant-authorized
762 data donations [118]. Some survey research also incorporates linkage to *administrative*
763 *records* such as tax records, educational transcripts, or health insurance claims, or to
764 *electronic health records*, though this represents a hybrid approach between primary
765 data collection and secondary data analysis.

766 It is important to distinguish survey research from secondary data analysis based on
767 the researcher’s role in data generation. Survey research involves researchers actively
768 designing and implementing data collection protocols, even when leveraging digital
769 technologies or administrative systems. The researcher determines sampling frames,
770 measurement instruments, timing, and consent procedures. In contrast, secondary
771 data analysis examines datasets originally collected for other purposes such as exist-
772 ing EHR systems, government statistics, or archived surveys. The boundary can blur
773 when researchers design linkage studies connecting survey data to pre-existing admin-
774 istrative records, but the defining characteristic of survey research is the intentional
775 design of primary data collection to address specific research questions.

776 Survey research begins with conceptualization, where researchers define constructs
777 of interest and identify relevant variables. Instrument development involves crafting
778 clear, unbiased questions, selecting appropriate response scales including *Likert scales*,
779 *semantic differential scales*, and *visual analog scales*, structuring questionnaire flow
780 with attention to question order effects, and conducting cognitive interviewing to iden-
781 tify comprehension problems [119]. For technology-based data collection, instrument
782 development also includes designing sensor protocols, defining event triggers for EMA,
783 or specifying data streams for passive collection. Pilot testing on small samples assesses
784 reliability including test-retest reliability and internal consistency, validity including
785 content validity, criterion validity, and construct validity, and identifies problems with
786 wording, skip patterns, or administration.

787 Sampling design defines the target population, selects sampling methods based on
788 resources and research goals, and determines required sample sizes considering preci-
789 sion requirements and anticipated response rates [120]. Probability sampling methods
790 including *simple random sampling*, *stratified sampling*, *cluster sampling*, and *multi-*
791 *stage sampling* enable statistical inference to populations, whereas non-probability
792 methods such as *convenience sampling*, *purposive sampling*, *quota sampling*, and *snow-*
793 *ball sampling* may be appropriate when probability sampling is infeasible, though they
794 limit generalizability. For longitudinal designs, sampling must also address panel main-
795 tenance strategies, refreshment samples to address attrition, and rotation patterns for
796 repeated cross-sections.

797 Data collection implements surveys through various modes with distinct advan-
798 tages and limitations. *Online surveys* offer cost efficiency and rapid deployment but

799 face coverage bias toward internet-connected populations. *Mail surveys* reach diverse
800 populations but suffer from low response rates and slow turnaround. *Telephone sur-*
801 *veys* enable probability sampling through random digit dialing but face declining
802 response rates and mobile-only household challenges. *Face-to-face interviews* achieve
803 high response quality and can include complex tasks or biomarker collection, but
804 require substantial resources. *Mixed-mode designs* combine approaches to balance cov-
805 erage, cost, and data quality [121]. For technology-based collection, researchers must
806 address informed consent for passive data collection, data security and privacy pro-
807 tections, participant burden from frequent assessments, and technical support for
808 device issues. Response monitoring tracks completion rates with real-time dashboards
809 and implements follow-up procedures to maximize participation while minimizing
810 nonresponse bias.

811 Data preparation includes data entry or import, coding of open-ended responses
812 through manual coding or text analysis, and missing data handling. Validation involves
813 reliability testing using *Cronbach's alpha* for internal consistency [122], test-retest reli-
814 ability for temporal stability, and construct validity assessment through *confirmatory*
815 *factor analysis* [123]. Scale refinement drops poor-performing items based on item-total
816 correlations, factor loadings, and reliability improvement. Analysis includes descriptive
817 statistics, inferential tests, and multivariate modeling. Interpretation generalizes find-
818 ings to the population, acknowledges sampling and non-sampling errors, and considers
819 limitations. Reporting documents sampling procedures, response rates, measurement
820 properties, and weighting adjustments following standards such as AAPOR reporting
821 guidelines.

822 Survey analysis employs a range of analytical methods appropriate to the research
823 design. Descriptive statistics provide univariate and bivariate summaries, while cor-
824 relation analysis examines linear associations between variables. Regression models
825 estimate relationships while controlling for confounders, including *linear regression*
826 for continuous outcomes, *logistic regression* for binary outcomes, *ordinal regression*
827 for ordered categories, and *multilevel modeling* for nested data structures [124]. For
828 longitudinal data, *growth curve models* and *latent growth curve models* examine tra-
829 jectories of change [125], *fixed effects models* control for time-invariant unobserved
830 heterogeneity, *random effects models* efficiently estimate between-person and within-
831 person effects, and *dynamic panel models* incorporate lagged dependent variables.
832 *Factor analysis*, both exploratory and confirmatory, assesses measurement struc-
833 ture and construct validity. *Structural equation modeling* simultaneously estimates
834 measurement models and structural relationships [126]. *Cluster analysis* identifies
835 homogeneous subgroups, while *latent class analysis* discovers categorical latent vari-
836 ables [127]. Text mining analyzes open-ended responses using sentiment analysis, topic
837 modeling, or word frequencies. Weighting procedures adjust for unequal selection prob-
838 abilities, nonresponse, and coverage error to improve population representativeness
839 through techniques such as *raking*, *post-stratification*, and *calibration weighting*. Miss-
840 ing data imputation techniques including *multiple imputation* and *maximum likelihood*
841 *estimation* address item nonresponse [43].

842 Survey research nevertheless faces numerous challenges to data quality and infer-
843 ence. Response biases include social desirability bias in which respondents present

themselves favorably, acquiescence bias in which respondents agree with statements regardless of content, and extreme response bias in which respondents favor endpoint responses. Sampling error arises from studying samples rather than complete populations, while coverage error occurs when sampling frames incompletely represent target populations. Question wording effects can substantially influence responses through framing, question order, or leading language [128]. Nonresponse bias threatens validity when nonrespondents differ systematically from respondents, a challenge exacerbated by declining response rates across survey modes. Mode effects introduce systematic differences based on survey administration method such as web, telephone, or face-to-face administration, complicating mixed-mode designs [121]. For technology-based data collection, additional challenges include device heterogeneity stemming from different sensors, operating systems, and data quality standards; digital divide issues excluding populations without technology access; participant burden from frequent assessments leading to fatigue or dropout, and privacy concerns about passive data collection. Longitudinal surveys face specific challenges from panel attrition involving selective dropout over time, panel conditioning in which repeated measurement changes behavior, and time-varying confounding in causal analysis.

Computational support needs for survey research include survey design platforms enabling questionnaire construction with skip logic, randomization, and mobile optimization, sampling calculators for determining sample sizes and allocations across strata or clusters, scale validation tools implementing reliability and validity analyses including factor analysis and structural equation modeling, EMA and sensor data management systems handling high-frequency temporal data, paradata analysis examining response times and patterns to detect data quality issues, natural language processing for open-ended response analysis including sentiment analysis and topic modeling, response pattern analysis detecting satisficing, straight-lining, or other data quality issues, survey weighting tools implementing raking, post-stratification, and calibration methods, and longitudinal data analysis frameworks supporting growth curve models, fixed effects estimation, and dynamic modeling. For technology-based collection, additional computational needs include secure data transmission and storage infrastructure, device synchronization and data integration across platforms, real-time monitoring dashboards for participant engagement and data quality, and privacy-preserving data processing implementing differential privacy or federated learning when appropriate.

2.3.4 Document and Content Analysis

Document and content analysis systematically analyzes textual, visual, or multimedia content to identify patterns, themes, and meanings in communication artifacts [129]. This scheme treats naturally occurring content including documents, images, videos, and social media posts as data, enabling unobtrusive study of communication, culture, and discourse.

Content analysis begins by formulating specific research questions about content characteristics, patterns, or meanings. Corpus definition involves identifying and collecting relevant materials from archives, databases, websites, or social media platforms. Sampling may be necessary when the complete corpus is too large, using methods like

888 simple random sampling, stratified sampling by time period or source, or purposive
889 sampling for specific characteristics.

890 Coding scheme development defines categories, units of analysis such as words,
891 sentences, themes, or documents, and coding rules in a detailed codebook [130]. Coder
892 training familiarizes human coders with the coding scheme through practice materials
893 and group discussion to establish shared understanding. Content coding can employ
894 manual coding by trained humans, automated annotation using rule-based systems
895 or machine learning, or hybrid approaches combining automated preprocessing with
896 human validation. Reliability assessment calculates intercoder reliability using *Cohen's*
897 *kappa* for two coders or *Krippendorff's alpha* for multiple coders and any number of
898 values, identifying and resolving disagreements [131].

899 Analysis examines frequency patterns, conducts thematic analysis identifying
900 recurring patterns, performs discourse analysis investigating language use and social
901 meaning, and compares content across sources, time periods, or conditions. Interpretation connects observed patterns to research questions and theoretical frameworks, considering historical and social contexts. Reporting documents the coding scheme, demonstrates reliability, describes analysis procedures, and presents representative examples alongside quantitative summaries.

906 Content analysis employs both traditional and computational methods. Frequency
907 counts tabulate occurrences of categories or themes, while chi-square tests examine
908 associations between content characteristics and sources or time periods. Content
909 categorization assigns texts to predefined or emergent categories. Topic modeling
910 using *LDA* or *BERTopic* discovers latent thematic structures in large text collections [132, 133]. Sentiment analysis classifies emotional valence using lexicon-based or machine learning approaches [134]. Named entity recognition identifies mentions of people, organizations, locations, and other entities [135]. Semantic network analysis maps conceptual relationships, while discourse analysis examines language patterns and social construction. Frame analysis identifies interpretive schemas in texts. Automated text classification using machine learning or large language models can scale coding to massive corpora while maintaining accuracy through active learning and validation [6].

919 Content analysis nevertheless faces significant methodological challenges. Inter-
920 coder reliability can be difficult to achieve when categories require interpretation
921 or judgment. Subjective interpretation may introduce researcher bias into category
922 definition and application. Validity of coding schemes depends on theoretical ground-
923 ing and operational definitions. Large-scale corpus handling strains manual coding
924 capacity, motivating computational approaches. Context preservation becomes prob-
925 lematic when analyzing excerpts or aggregating across documents. Multimodal content
926 integration requires methods spanning text, image, audio, and video analysis.

927 Computational tools for content analysis therefore include natural language
928 processing pipelines performing tokenization, part-of-speech tagging, and syntactic
929 parsing as preprocessing steps, topic modeling implementations including *LDA*, *NMF*,
930 and *BERTopic*, sentiment analysis using lexicons or deep learning models, named
931 entity recognition using conditional random fields or neural networks, automated cod-
932 ing with large language models fine-tuned on labeled examples, reliability calculators

for various coefficients and data structures, and visualization tools for text patterns including word clouds, topic evolution over time, and semantic networks.

2.3.5 Secondary Data Analysis

Secondary data analysis examines existing datasets collected for other purposes to answer new research questions [136]. This scheme emphasizes efficiency by leveraging existing data infrastructure and enables research on phenomena requiring large samples, long time periods, or resources beyond individual researchers' capacity.

Secondary analysis begins by developing research questions answerable with existing data, recognizing both opportunities and constraints imposed by available measures and designs. Data source identification searches repositories including ICPSR¹⁰, UK Data Service¹¹, government statistical agencies, and institutional archives. Data acquisition involves obtaining access through registration or restricted-use agreements, downloading datasets and documentation, and verifying data integrity.

Data quality assessment evaluates completeness referring to the extent of missing data, accuracy through comparison with external sources when possible, consistency in terms of logical relationships between variables, and documentation quality including codebooks, methodology reports, and questionnaires. Variable mapping aligns existing variables with research constructs, often requiring compromise between theoretical ideals and available measures. Data preparation includes recoding variables to appropriate formats and scales, creating derived variables through transformations or combinations, subsetting to relevant subsamples, and merging datasets from different waves or sources while maintaining proper linkages.

Analysis applies appropriate statistical or computational methods considering the data structure such as cross-sectional, longitudinal, hierarchical, or network structures. Validation conducts sensitivity analyses examining robustness to modeling choices and robustness checks comparing results across subsamples or specifications. Interpretation considers the original data collection context, purpose, and methods, recognizing how these may influence applicability to new research questions. Reporting acknowledges data sources with proper citation, documents all transformations and analytical decisions, and discusses how original collection context may affect interpretation. It should be noted that cross-sectional and longitudinal designs refer to temporal data structures rather than research schemes per se. In this taxonomy, they are discussed under survey research because survey research is defined by the intentional design of primary data collection, regardless of whether measurements are taken at one or multiple time points

Secondary analysis employs the full range of statistical methods depending on data structure. For cross-sectional data, researchers use regression, structural equation modeling, and propensity score methods. For longitudinal data, growth curve modeling, fixed effects models, difference-in-differences, and event history analysis are appropriate [125]. For multilevel data, hierarchical linear models and mixed-effects models account for nesting [61]. For time series, ARIMA models, vector autoregression, and Granger causality tests are employed [90]. For spatial data, spatial regression and

¹⁰<https://www.icpsr.umich.edu>

¹¹<https://ukdataservice.ac.uk>

geographically weighted regression address spatial dependencies [137]. Data integration and harmonization techniques address differences in variable definitions, response categories, or population coverage across datasets. Missing data methods handle item and unit nonresponse. Complex survey analysis accounts for sampling weights, stratification, and clustering in survey datasets [138].

Secondary analysis faces unique challenges arising from using data collected for other purposes. Data documentation quality varies substantially across sources, with some datasets lacking comprehensive codebooks, questionnaires, or methodology reports. Missing metadata complicates understanding variable definitions, skip patterns, or quality flags. Variable measurement differences across datasets limit comparability and integration. Contextual understanding of original collection requires substantial background research into study design, historical context, and intended uses. Temporal and geographic limitations may restrict applicability to other settings or time periods. Data access restrictions including privacy and proprietary constraints can prevent replication or limit research possibilities.

Computational support for secondary analysis therefore requires tools for data integration combining datasets with different structures and schemas, metadata management systems organizing codebooks and documentation, automated documentation generation extracting variable information into searchable databases, data harmonization algorithms mapping variables across datasets using semantic matching, format conversion utilities translating between data formats such as SAS, Stata, SPSS, R, and Python, and exploratory data analysis tools rapidly examining variable distributions and relationships to assess fitness for research questions.

2.4 Data Analysis Methods

While research schemes provide overarching structures for inquiry, specific analytical methods transform data into findings. This section catalogs major categories of analytical methods employed across quantitative social science, recognizing that the boundaries between categories are fluid and many contemporary analyses integrate multiple approaches.

2.4.1 Statistics-Based Methods

Statistical methods form the foundation of quantitative social science, providing tools for description, inference, and modeling grounded in probability theory [139].

Descriptive statistics summarize data distributions and relationships through measures of central tendency including mean, median, and mode, and measures of dispersion including variance, standard deviation, and interquartile range. Distribution analysis examines skewness and kurtosis to assess normality assumptions and guide transformation decisions. Cross-tabulations and contingency tables describe relationships between categorical variables.

Inferential statistics enable generalization from samples to populations through hypothesis testing and confidence intervals [140]. *T-tests* compare means between two groups, while *paired t-tests* handle within-subject comparisons. *Analysis of variance*

(ANOVA) compares means across multiple groups through partitioning total variance into between-group and within-group components. *One-way ANOVA* examines a single factor, *factorial ANOVA* investigates multiple factors and their interactions, and *repeated measures ANOVA* handles within-subject designs. Post-hoc tests such as Tukey’s HSD or Bonferroni corrections address multiple comparisons. *Analysis of covariance* (ANCOVA) combines ANOVA with regression by controlling for continuous covariates. *Multivariate analysis of variance* (MANOVA) extends ANOVA to multiple dependent variables simultaneously, testing whether group differences exist across the multivariate outcome space.

Chi-square tests examine associations in contingency tables, with *Pearson’s chi-square* testing independence and *McNemar’s test* for paired categorical data. *Fisher’s exact test* provides inference for small samples. Nonparametric alternatives include *Mann-Whitney U test* for two independent samples, *Wilcoxon signed-rank test* for paired samples, and *Kruskal-Wallis test* for multiple groups, which make fewer distributional assumptions than parametric tests.

Regression analysis estimates relationships between dependent and independent variables. *Linear regression* models continuous outcomes, with *multiple regression* incorporating multiple predictors, *polynomial regression* capturing nonlinear relationships, and *hierarchical regression* examining sequential variable entry to assess incremental variance explained. *Generalized linear models* extend regression to non-normal outcomes through link functions, including *logistic regression* for binary outcomes, *Poisson regression* for count data, and *negative binomial regression* for overdispersed counts. *Ordinal regression* models ordered categorical outcomes. *Multi-level models* or *hierarchical linear models* account for nested data structures such as students within schools, accommodating both fixed and random effects [61].

Time series analysis addresses temporal dependencies. *Autoregressive models*, *moving average models*, and *autoregressive integrated moving average* (ARIMA) models capture patterns for forecasting. *Seasonal decomposition* separates time series into trend, seasonal, and irregular components [90]. *Vector autoregression* (VAR) models analyze systems of multiple time series. *Granger causality tests* assess predictive relationships in temporal data [98].

Survival analysis examines time-to-event outcomes. *Kaplan-Meier estimation* provides nonparametric survival curve estimates. *Cox proportional hazards regression* models hazard rates while adjusting for covariates without specifying baseline hazard distribution [93]. *Parametric survival models* including *exponential*, *Weibull*, and *log-normal models* specify distributional forms. *Competing risks models* handle multiple potential event types.

Multivariate methods address relationships among multiple variables simultaneously. *Factor analysis* examines latent structure underlying observed variables. *Exploratory factor analysis* discovers factor structures, while *confirmatory factor analysis* tests hypothesized measurement models, reducing dimensionality and assessing construct validity [123]. *Principal component analysis* creates orthogonal linear combinations explaining maximum variance, often used for dimensionality reduction. *Structural equation modeling* simultaneously estimates measurement models linking

latent constructs to observed indicators and structural models specifying relationships among latent variables [126].

Cluster analysis identifies homogeneous subgroups using *k-means clustering* partitioning observations into *k* clusters, *hierarchical clustering* creating nested groupings through agglomerative or divisive approaches, or *model-based clustering* using finite mixture models. *Discriminant analysis* classifies cases into groups based on predictor variables through *linear discriminant analysis* or *quadratic discriminant analysis*. *Canonical correlation analysis* examines relationships between two sets of variables. *Multidimensional scaling* creates low-dimensional representations preserving pairwise distances.

From an analytical purpose perspective, statistical methods serve several distinct research objectives. *Effect decomposition analysis* examines mechanisms through which variables influence outcomes. *Mediation analysis* tests indirect effects via intervening variables, estimating direct and indirect pathways using approaches such as the Sobel test, bootstrap methods, or structural equation modeling frameworks [62, 141]. *Moderation analysis* investigates conditional effects, testing whether relationships vary across levels of a third variable through interaction terms in regression or through subgroup analysis [64]. Combined models examine *moderated mediation*, where mediation pathways differ across moderator levels, or *mediated moderation*, where moderation effects operate through mediators [65].

Causal inference aims to establish causal relationships from observational or experimental data. Beyond randomized experiments providing the gold standard, quasi-experimental designs approximate experimental conditions. *Propensity score methods* balance treatment and control groups on observed covariates through matching, weighting, or stratification [68]. *Instrumental variable estimation* addresses endogeneity by leveraging exogenous variation affecting treatment but not outcomes directly [67]. *Difference-in-differences* compares changes over time between treatment and control groups [66]. *Regression discontinuity designs* exploit threshold-based treatment assignment [69]. *Interrupted time series analysis* assesses intervention effects through temporal pattern changes. Sensitivity analyses assess robustness to unmeasured confounding [106].

Prediction and classification focus on forecasting outcomes rather than parameter interpretation. Methods optimize predictive accuracy through cross-validation, balancing bias-variance tradeoffs. *Model comparison and selection* use information criteria such as *Akaike Information Criterion* (AIC) or *Bayesian Information Criterion* (BIC) that penalize complexity [142]. *Effect size estimation* quantifies substantive importance beyond statistical significance through measures like Cohen’s *d*, odds ratios, or R-squared [113]. *Power analysis* determines sample sizes needed to detect effects of theoretical interest with adequate statistical power.

2.4.2 Information Theory-Based Methods

Information theory provides a mathematical framework for quantifying uncertainty, information content, and dependencies between variables [95, 96]. While less commonly employed than classical statistics, information-theoretic methods offer unique insights into complex systems and causal relationships. Central to this approach are

1104 entropy measures, which quantify uncertainty or randomness in probability distribu-
1105 tions. *Shannon entropy* measures average information content, with higher entropy
1106 indicating greater uncertainty. *Conditional entropy* measures uncertainty in one vari-
1107 able given knowledge of another. *Mutual information* quantifies dependencies between
1108 variables without assuming specific functional forms, making it suitable for detecting
1109 nonlinear relationships. Unlike correlation, mutual information is zero if and only if
1110 variables are statistically independent.

1111 Information flow analysis enables causal discovery in time series using *transfer*
1112 *entropy*, which measures directed information transfer from one process to another
1113 beyond what can be predicted from each process’s history [97]. This permits distin-
1114 guishing causation from mere correlation in temporal systems. Complexity measures
1115 approximate algorithmic complexity, quantifying the minimum description length of a
1116 dataset. Applications include network analysis using mutual information to infer rela-
1117 tionships, causal inference in observational data, information diffusion studies, and
1118 model selection using information criteria such as AIC or BIC that balance fit against
1119 complexity [142].

1120 2.4.3 Machine Learning and Deep Learning Methods

1121 Machine learning methods discover patterns through algorithmic learning from exam-
1122 ples [105, 143]. These methods have become increasingly prevalent for prediction,
1123 classification, dimensionality reduction, and discovering complex nonlinear relation-
1124 ships [7]. Within the machine learning paradigm, supervised learning learns mappings
1125 from inputs to outputs using labeled training data. Classification methods predict
1126 categorical outcomes using *support vector machines* finding optimal separating hyper-
1127 planes [75], *random forests* combining multiple decision trees [73], *gradient boosting*
1128 sequentially fitting trees to residuals [74], and *neural networks* learning hierarchi-
1129 cal representations [76]. Regression methods predict continuous outcomes through
1130 *ridge regression* adding L2 regularization, *lasso regression* using L1 regularization for
1131 variable selection [144], *elastic net* combining penalties, and *deep neural networks*
1132 modeling highly nonlinear relationships. Ensemble methods combine multiple models
1133 through bagging, boosting, or stacking to improve prediction accuracy and robustness.

1134 In contrast, unsupervised learning discovers structure in unlabeled data with-
1135 out predefined outcomes. Clustering groups similar observations using *DBSCAN*
1136 for density-based clusters [145], *Gaussian mixture models* representing distributions
1137 as component combinations, and *spectral clustering* using graph-theoretic meth-
1138 ods [83]. Dimensionality reduction creates low-dimensional representations through
1139 *t-SNE* preserving local neighborhoods [77], *UMAP* preserving global structure [78],
1140 and *autoencoders* learning compressed representations. Anomaly detection identifies
1141 unusual observations using *isolation forests*, *one-class SVM*, or reconstruction error
1142 from autoencoders.

1143 Deep learning represents a particularly powerful class of methods employing multi-
1144 layer neural networks that learn hierarchical representations [76]. For natural language
1145 processing, *transformers*, *BERT* for contextual embeddings, and *GPT* for generation
1146 have revolutionized text analysis [33, 146]. Computer vision applications use *convolu-*
1147 *tional neural networks* for image analysis in social contexts [147]. Sequence modeling

1148 employs *LSTMs* or *GRUs* for capturing temporal dependencies in longitudinal social
1149 data [148]. Applications span predictive modeling of social outcomes, pattern recog-
1150 nition in complex behavioral data, automated content analysis at scale, and social
1151 media analysis investigating networks and influence [1].

1152 2.4.4 Graph Theory and Network Analysis

1153 Network analysis studies relational structures connecting entities such as individuals,
1154 organizations, or concepts [55, 80]. Network construction creates graph representa-
1155 tions including social networks mapping interpersonal relationships, citation networks
1156 representing scholarly influence, and co-occurrence networks linking entities appearing
1157 together. Construction methods include correlation-based approaches, partial correla-
1158 tion networks controlling for confounders, and graphical models estimating conditional
1159 independence structures [149].

1160 Once networks are constructed, centrality measures identify structurally impor-
1161 tant nodes. *Degree centrality* counts direct connections, identifying well-connected
1162 actors. *Betweenness centrality* measures how often nodes lie on shortest paths
1163 between others [150], revealing bridges and brokers who control information flow.
1164 *Closeness centrality* averages distance to all other nodes, identifying efficient commu-
1165 nicators. *Eigenvector centrality* weights connections by the importance of connected
1166 nodes, recursively identifying influential actors. Community detection methods iden-
1167 tify densely connected subgroups through *modularity optimization* [82], the *Louvain*
1168 *algorithm* for efficient large-scale detection [151], *label propagation* using neighbor
1169 influence, and *stochastic block models* providing probabilistic frameworks [84].

1170 Network dynamics examine how networks evolve over time through temporal net-
1171 work analysis tracking edge formation and dissolution. Diffusion processes model how
1172 information, behaviors, or diseases spread through networks using threshold models
1173 where adoption occurs after sufficient peer exposure, cascade models where influ-
1174 ence propagates sequentially, or epidemic models such as SIR or SIS borrowed from
1175 epidemiology [152]. Network comparison assesses similarity between networks using
1176 graph kernels, network alignment finding node correspondences across networks, or
1177 distributional comparisons of network statistics. Applications include social influence
1178 analysis examining peer effects and contagion, collaboration network studies investi-
1179 gating scientific productivity and teamwork patterns, information diffusion tracking
1180 how news and innovations spread, and organizational structure analysis examining
1181 communication and authority patterns.

1182 2.4.5 Interdisciplinary-Integrated Methods

1183 Computational social science increasingly integrates methods across traditional dis-
1184 ciplinary boundaries, drawing on diverse theoretical frameworks and analytical
1185 techniques.

1186 *Agent-based modeling* simulates social systems through autonomous agents fol-
1187 lowing behavioral rules, enabling exploration of how micro-level behaviors generate
1188 macro-level patterns [153]. Agents possess internal states, decision rules, and inter-
1189 action protocols, situated within environments that constrain behavior. Emergent
1190 phenomena arise from agent interactions without central coordination. Applications

1191 include modeling segregation dynamics [154], market dynamics, collective behavior,
1192 and organizational processes. Model validation compares simulated patterns against
1193 empirical data through pattern-oriented modeling or statistical calibration.

1194 *Field-theoretic models* from physics provide powerful frameworks for understand-
1195 ing social behavior emergence. Drawing on concepts from electromagnetic field theory
1196 and quantum field theory, these approaches conceptualize social influence as operat-
1197 ing through continuous social fields rather than discrete interactions [155]. *Social field*
1198 *theory* models individuals as embedded in multidimensional social spaces where posi-
1199 tions determine experienced forces. Field potentials represent latent tendencies toward
1200 behaviors or attitudes, with gradients indicating directional influences. *Dynamic field*
1201 *models* incorporate temporal evolution, treating social change as field dynamics gov-
1202 erned by differential equations analogous to physical field equations [156]. Opinion
1203 dynamics models such as *Ising models* borrowed from statistical physics represent dis-
1204 crete opinion states with interaction energies determining alignment tendencies [157].
1205 *Continuous opinion models* like the Deffuant model or Hegselmann-Krause model
1206 describe opinion convergence through bounded confidence mechanisms. Phase transi-
1207 tions in social systems, where small parameter changes produce qualitative behavioral
1208 shifts, parallel physical phase transitions and can be analyzed using critical phenomena
1209 theory [158]. These approaches enable rigorous mathematical treatment of collective
1210 behavior, social influence propagation, and emergence of social order from individual
1211 interactions.

1212 *Natural experiments* leverage digital trace data for causal inference when random-
1213 ized experiments are infeasible, exploiting naturally occurring variation in treatment
1214 exposure. Examples include policy changes affecting different regions at different
1215 times, platform design changes affecting random subsets of users, or exogenous shocks
1216 creating quasi-random treatment assignment. *Computational text analysis* combines
1217 natural language processing techniques with substantive social theory to analyze com-
1218 munication at unprecedented scales [6]. Methods integrate semantic analysis, discourse
1219 analysis, and social theory to understand meaning-making processes.

1220 Mixed-methods integration combines quantitative and qualitative approaches in
1221 complementary ways. *Sequential explanatory designs* use quantitative analysis to
1222 identify patterns followed by qualitative investigation to understand mechanisms.
1223 *Sequential exploratory designs* reverse this order, using qualitative exploration to gen-
1224 erate hypotheses for quantitative testing. *Triangulation* employs multiple data sources
1225 and methods to validate findings and provide richer understanding. *Digital ethnog-*
1226 *raphy* combines computational analysis of online behavior with interpretive study of
1227 meaning and culture, bridging macro-level pattern detection with micro-level meaning
1228 interpretation.

1229 Causal inference methods have developed sophisticated approaches for establish-
1230 ing causality in observational data beyond those already discussed. *Synthetic control*
1231 *methods* create counterfactual comparisons by constructing weighted combinations of
1232 control units matching treated units on pre-treatment characteristics [70]. *Mediation*
1233 *analysis with sensitivity to unmeasured confounding* extends traditional mediation to
1234 assess robustness [159]. *Structural causal models* using directed acyclic graphs formal-
1235 ize causal assumptions and guide identification strategies [160]. *Causal forests* extend

1236 random forests for heterogeneous treatment effect estimation, identifying subgroups
1237 with varying causal effects [161].

1238 2.5 Visualization and Reporting Tools

1239 Effective presentation of research findings requires appropriate visualization and
1240 reporting tools that transform complex analytical results into accessible formats for
1241 diverse audiences.

1242 Data visualization creates visual representations of quantitative patterns and
1243 relationships. *Statistical graphics* display distributions, correlations, and group com-
1244 parisons through histograms, scatterplots, box plots, and density plots. *Network*
1245 *visualizations* reveal relational structures using node-link diagrams, adjacency matri-
1246 ces, or force-directed layouts. *Temporal visualizations* show time series patterns,
1247 event sequences, and trend evolution over time. *Geospatial visualizations* map spa-
1248 tial distributions and geographic patterns using choropleth maps, point maps, or heat
1249 maps [162]. *Interactive dashboards* enable exploratory interaction with results, allowing
1250 users to filter, zoom, and reconfigure visualizations dynamically. Common tools include
1251 Plotly ¹² and D3.js ¹³ for web-based interactive graphics, Tableau ¹⁴ for business intel-
1252 ligence dashboards, and ggplot2 [163] or matplotlib [164] for publication-quality static
1253 graphics.

1254 Academic reporting follows disciplinary conventions to communicate findings to
1255 scholarly audiences. The *IMRaD format* providing Introduction, Methods, Results,
1256 and Discussion sections offers standard structure for empirical reports. Tables and
1257 figures follow style guidelines such as APA ¹⁵, ASA ¹⁶, or journal-specific requirements.
1258 Statistical reporting conventions include effect sizes quantifying practical significance,
1259 confidence intervals indicating precision of estimates, and significance tests with
1260 appropriate corrections for multiple comparisons [165].

1261 Reproducibility tools enable verification and extension of research findings. *Code*
1262 *sharing* through repositories such as GitHub ¹⁷, GitLab ¹⁸, or institutional archives
1263 documents analytical procedures. *Computational notebooks* including Jupyter ¹⁹ and
1264 R Markdown ²⁰ integrate code, results, and narrative explanation in executable docu-
1265 ments [166]. *Containerization* using Docker ²¹ preserves complete computational
1266 environments including software versions and dependencies. *Data repositories* such as
1267 those mentioned in Section ?? enable data sharing when ethically permissible and
1268 legally feasible, following FAIR principles ensuring data are Findable, Accessible, Inter-
1269 operable, and Reusable [167]. *Version control systems* track analytical evolution and
1270 document decision points, creating audit trails of methodological choices [168].

¹²<https://plotly.com>

¹³<https://d3js.org>

¹⁴<https://www.tableau.com>

¹⁵<https://apastyle.apa.org>

¹⁶<https://www.asanet.org/publications-0/style-guide>

¹⁷<https://github.com>

¹⁸<https://gitlab.com>

¹⁹<https://jupyter.org>

²⁰<https://rmarkdown.rstudio.com>

²¹<https://www.docker.com>

1271 These technical tools support the broader dissemination strategies discussed in
1272 Section ??, providing infrastructure for transparent, reproducible, and accessible
1273 research communication.

1274 2.6 Summary: Toward an Integrative Computational Research 1275 Framework

1276 This overview of quantitative social science research reveals the challenges defined in
1277 Section ?? and demonstrates how they manifest across research schemes and lifecycle
1278 stages. Quantitative social science encompasses diverse research schemes including
1279 literature review and meta-analysis, experimental research, survey research, document
1280 and content analysis, and secondary data analysis, among others. While these schemes
1281 share a common research lifecycle structure—problem discovery, research design, data
1282 collection, data processing, data analysis, and dissemination—each scheme instantiates
1283 this lifecycle with distinct methodological emphases, data requirements, and analytical
1284 approaches.

1285 **Data and Tool Fragmentation** permeates the quantitative social science land-
1286 scape. As documented in this section, statistical methods span classical hypothesis
1287 testing to complex multilevel models and causal inference frameworks; machine learn-
1288 ing encompasses supervised and unsupervised learning, deep learning architectures,
1289 and ensemble methods; network analysis employs graph-theoretic algorithms and
1290 diffusion models; information-theoretic approaches quantify uncertainty and informa-
1291 tion flow. Each methodological tradition operates within distinct software ecosystems
1292 with incompatible data structures and workflows. Researchers conducting literature
1293 reviews may use Zotero ²² and R’s metafor package [169], experimental researchers
1294 may employ SPSS ²³ and G*Power [170], survey researchers may use Qualtrics ²⁴ and
1295 Stata ²⁵, content analysts may use NVivo ²⁶ and Python’s NLTK [171], and secondary
1296 data analysts may use SAS ²⁷ and institutional data portals. This fragmentation
1297 forces researchers to invest substantial effort learning multiple disconnected tools and
1298 repeatedly translating data and results across platforms.

1299 **Research Lifecycle Discontinuity** emerges from the lack of integrated sup-
1300 port spanning all six lifecycle stages. Current tools typically address isolated stages
1301 or tasks: reference managers for literature organization, survey platforms for data
1302 collection, statistical packages for analysis, and LaTeX or Word for manuscript prepa-
1303 ration. Transitions between stages require manual data export and import, format
1304 conversions, and workflow reconfigurations. Information created in one stage—such
1305 as variable definitions during research design, data quality assessments during col-
1306 lection, or transformation decisions during processing—often fails to propagate to
1307 subsequent stages. This discontinuity creates opportunities for errors, duplicates effort,
1308 and obscures the connections between research questions, methodological decisions,
1309 and analytical results.

²²<https://www.zotero.org>

²³<https://www.ibm.com/spss>

²⁴<https://www.qualtrics.com>

²⁵<https://www.stata.com>

²⁶<https://www.qsrinternational.com/nvivo>

²⁷<https://www.sas.com>

1310 **Structural Reproducibility Deficits** stem from inadequate documentation and
1311 workflow management infrastructure. As illustrated across research schemes, con-
1312 temporary quantitative social science involves numerous methodological choices at each
1313 lifecycle stage: sampling strategies, operationalization decisions, missing data handling
1314 approaches, model specifications, and robustness checks. These decisions collectively
1315 determine research findings, yet they are often documented informally in researcher
1316 notes or incompletely described in methods sections. Current tools rarely provide sys-
1317 tematic mechanisms for tracking decisions, versioning analytical workflows, or linking
1318 results to the specific procedures that generated them. Consequently, reproducing pub-
1319 lished findings—or even replicating one’s own earlier analyses—remains challenging
1320 despite increasing emphasis on computational reproducibility.

1321 **Steep Learning Curves from Insufficient Abstraction** create barriers
1322 to adopting sophisticated methods. Statistical software exposes low-level procedu-
1323 ral details requiring users to manage data structures, implement algorithms, and
1324 coordinate multiple operations manually. Machine learning frameworks demand under-
1325 standing of tensor operations, optimization procedures, and model architectures.
1326 Network analysis tools require graph data structure manipulation and algorithm
1327 parameterization. While this flexibility benefits experts, it impedes researchers seeking
1328 to apply established methods to substantive questions. Current tools provide limited
1329 abstraction layers that would hide unnecessary complexity while preserving analyti-
1330 cal rigor, forcing researchers to choose between accessible but limited point-and-click
1331 interfaces and powerful but complex programming environments.

1332 **Data Heterogeneity and Scale** challenges intensify as research increasingly
1333 involves multi-source data integration, combining survey responses with administrative
1334 records, digital trace data with qualitative interviews, or textual corpora with net-
1335 work structures. Multimodal data encompassing text, images, audio, video, and sensor
1336 streams require specialized processing pipelines. Scale challenges emerge as datasets
1337 grow from thousands to millions of observations, and text corpora expand from hun-
1338 dreds to billions of tokens. Temporal complexity arises in longitudinal panel data,
1339 high-frequency event sequences, and dynamic networks. Spatial complexity appears
1340 in geographic data with hierarchical structures and spatial dependencies. These com-
1341 plexities demand sophisticated data representation, storage, processing, and analytical
1342 capabilities that exceed what specialized single-purpose tools typically provide.

1343 These five challenges, manifesting differently across research schemes and life-
1344 cycle stages, point toward requirements for an integrative computational research
1345 framework. Such a framework should provide several core capabilities:

1346 *Unified data architecture* must support flexible representation and storage of
1347 diverse data types including structured tabular data, unstructured text and multime-
1348 dia, relational network data, spatial and temporal data, and hierarchical or nested
1349 structures. The architecture should enable seamless integration across data sources
1350 and modalities while preserving metadata and provenance throughout the research
1351 lifecycle, addressing **Data Heterogeneity and Scale** challenges.

1352 *Standardized workflow abstractions* should formalize common patterns in the
1353 research lifecycle across different schemes, addressing **Research Lifecycle Discon-**
1354 **tinuity**. These abstractions define interfaces for data collection, processing, analysis,

1355 and dissemination stages while permitting scheme-specific instantiations. Workflow
1356 specifications should be machine-readable, version-controlled, and shareable, enabling
1357 researchers to document, reproduce, and adapt analytical procedures. Modular work-
1358 flow components should compose flexibly, allowing researchers to construct custom
1359 pipelines appropriate to their research questions while maintaining compatibility with
1360 shared infrastructure.

1361 *Integrated analytical methods* should provide unified interfaces to diverse method-
1362 ological traditions, addressing **Data and Tool Fragmentation**. Rather than
1363 requiring researchers to learn multiple disconnected tools, the framework should offer
1364 consistent APIs for statistical modeling, machine learning, network analysis, text
1365 analysis, and visualization. Method selection and parameterization should be guided
1366 by research objectives and data characteristics rather than by tool availability or
1367 researcher familiarity.

1368 *Computational reproducibility infrastructure* must address **Structural Repro-**
1369 **ducibility Deficits**. Version control should track not only code but also data
1370 transformations, methodological decisions, and parameter selections throughout the
1371 analytical process. Environment management should capture software dependencies,
1372 versions, and configurations ensuring that analyses produce consistent results across
1373 platforms and over time. Workflow documentation should automatically generate audit
1374 trails linking research questions to data operations to analytical results, making the
1375 complete research process transparent and verifiable.

1376 *Intelligent assistance through large language models and AI* combined with *pro-*
1377 *gressive abstraction layers* can address **Steep Learning Curves from Insufficient**
1378 **Abstraction**. The framework should offer multiple interaction modes: graphical
1379 interfaces for exploratory analysis and rapid prototyping, scripting interfaces for repro-
1380 ducible workflows and customization, and programmatic APIs for method developers
1381 and advanced users. LLMs can assist in literature synthesis through automated sum-
1382 marization, content analysis through intelligent coding assistance, and manuscript
1383 preparation through writing support. Built-in but extensible algorithm libraries reduce
1384 tool learning costs by providing common methods while allowing integration of spe-
1385 cialized techniques. Progressive disclosure of complexity allows novices to accomplish
1386 common tasks easily while providing experts with fine-grained control.

1387 In summary, an integrative computational research framework addresses the
1388 five core challenges identified in Section ?? by providing unified data architecture,
1389 standardized workflow abstractions, integrated analytical methods, reproducibility
1390 infrastructure, and intelligent assistance with progressive abstraction. Such a frame-
1391 work supports the complete research lifecycle across diverse quantitative social science
1392 schemes, reducing barriers to sophisticated analysis while enhancing transparency,
1393 reproducibility, and cumulative knowledge building. The following sections present our
1394 ongoing efforts toward realizing this vision through concrete architectural designs and
1395 prototype implementations.

1396 3 Framework Architecture

1397 3.1 Architectural Design Overview

1398 The proposed research framework adopts a modular, workflow-oriented architecture
1399 designed to support the full lifecycle of quantitative social science research. The frame-
1400 work conceptualizes the research process as a sequence of stages with distinct research
1401 objectives and operational emphases, including problem discovery, research design,
1402 data collection, data processing, analysis, and dissemination (see Section 2.1).

1403 The architecture is organized into three complementary layers that jointly support
1404 research workflows. At the highest level is the *Research Workflow Layer*, which func-
1405 tions as a conceptual layer representing the progression of research across different
1406 phases. This layer does not prescribe specific computational implementations; instead,
1407 it provides an explicit structural representation of research stages and their relations.

1408 The Research Workflow Layer interfaces with the *Functional API Layer*, which
1409 exposes a set of high-level functions corresponding to common research activities.
1410 These functions are implemented by the *Infrastructure Module Layer*, which contains
1411 modular computational components that provide the underlying operational support.
1412 The separation between conceptual structure, functional interfaces, and infrastructural
1413 modules is intended to promote loose coupling, extensibility, and adaptability across
1414 diverse research contexts.

1415 The framework provides eight primary functional modules that address distinct
1416 research needs, as shown in Figure 2. These include an *LLM Q&A* module that
1417 facilitates interactive question-answering using large language models for literature
1418 exploration and research ideation. The *Data Collection & Review* module offers sys-
1419 tematic literature search across multiple academic databases including *arXiv*, *PubMed*,
1420 *CrossRef*, *OSF*, *SocArXiv*, and *Zenodo*, with automated metadata extraction and bib-
1421 liography management capabilities. For primary data collection, the *Questionnaire*
1422 *Design* module supports the creation of complex survey instruments with LaTeX-based
1423 rendering for professional presentation.

1424 For data analysis and visualization, the framework includes a *Data Visualiza-*
1425 *tion* module that generates publication-ready figures including statistical plots, word
1426 clouds, and frequency distributions, all with automatic LaTeX integration. The *Data*
1427 *Processing* module implements configurable data transformation pipelines that han-
1428 dle cleaning, validation, normalization, and feature engineering while maintaining
1429 comprehensive provenance tracking. Statistical analysis is supported through the
1430 *Data Analysis* module, which provides a comprehensive suite of methods including
1431 descriptive statistics, regression models, causal inference techniques such as difference-
1432 in-differences, instrumental variables, and propensity score matching, and multilevel
1433 modeling, and survey analysis with design-based inference.

1434 The framework also includes a LaTeX *Code Generation* module that automati-
1435 cally produces LaTeX tables and figures from statistical results, ensuring consistency
1436 between computational outputs and manuscript presentation. Finally, the *Contract*
1437 *Definition* module allows researchers to specify formal schemas for data struc-
1438 tures, analysis workflows, visualization specifications, and output formats, facilitating
1439 automation and reproducibility across the research process.

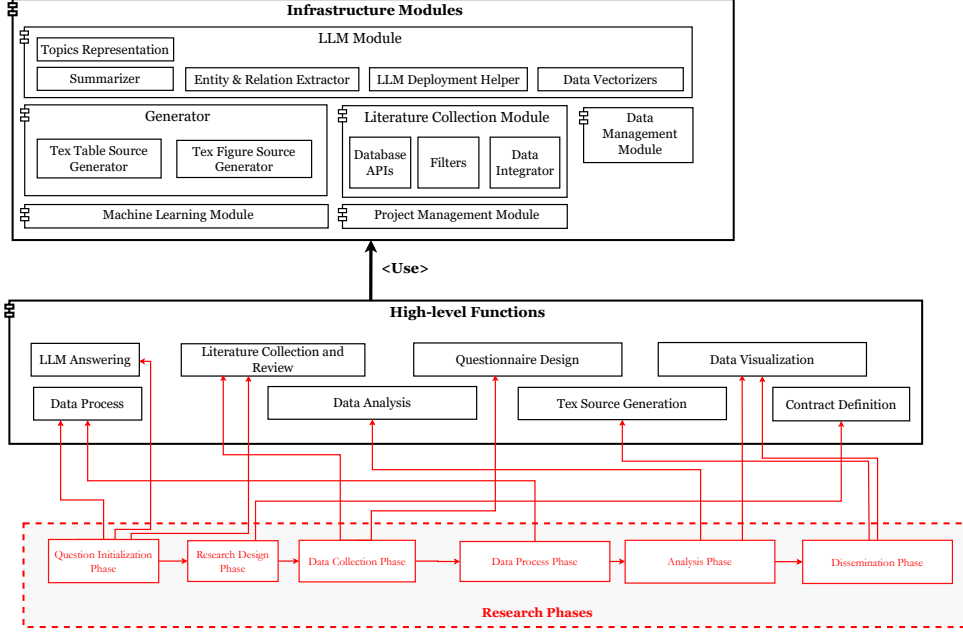


Fig. 2 High-level system architecture of the quantitative social science research computational framework. The red arrows illustrate the mapping between research phases and their primary high-level functions. These functions are implemented using multiple infrastructure modules designed for loose coupling and extensibility.

As illustrated in Figure 2, the research workflow progresses through several interconnected stages, each supported by specific functional modules. During the problem discovery stage, researchers primarily utilize the LLM Q&A module for interactive exploration of research questions, complemented by the Data Collection & Review module for literature discovery and the Data Processing module for preliminary exploration of existing datasets. This stage emphasizes exploratory capabilities that help identify research gaps and formulate initial hypotheses.

The research design stage centers around the Contract Definition module, which enables researchers to formally specify data schemas, analysis workflow specifications, visualization templates, output format requirements, and validation rules. These contracts serve as executable specifications that guide subsequent automation while providing machine-readable documentation of research intentions for reproducibility. The contract-based approach ensures methodological transparency and enables systematic validation throughout the research process.

During data collection, researchers employ the Data Collection & Review module to systematically gather literature from academic databases, while the Questionnaire Design module supports primary data collection through survey instruments. All collected data are stored according to predefined contracts, ensuring structural consistency and metadata completeness from the outset. The data processing stage

1459 then applies researcher-specified transformation pipelines through the Data Process-
1460 ing module, which executes sequences of data processors on raw inputs to produce
1461 standardized representations while maintaining contract-based validation to ensure
1462 processed data meet specified quality criteria.

1463 The analysis stage leverages the Data Analysis module, which provides access to
1464 diverse statistical and computational methods including traditional statistical infer-
1465 ence, causal modeling, machine learning approaches, and domain-specific analytical
1466 frameworks. This module integrates with the Data Visualization module to generate
1467 diagnostic plots and substantive visualizations that support both analytical explo-
1468 ration and result presentation. Finally, during the publication stage, the LaTeX Code
1469 Generation module automatically translates analysis results into publication-ready
1470 tables and figures, while the Data Visualization module produces graphics, reduc-
1471 ing transcription errors and accelerating the transition from analysis to manuscript
1472 preparation.

1473 The architecture embodies several central design principles that guide its imple-
1474 mentation and use. Loose coupling ensures that infrastructure modules maintain
1475 minimal interdependencies, communicating through well-defined interfaces rather than
1476 direct coupling. This design permits independent module development and testing,
1477 enables substitution of alternative implementations without system-wide changes, and
1478 supports graceful degradation when optional modules are unavailable. The frame-
1479 work’s extensibility is realized through multiple extension points: researchers can
1480 implement custom data processors by adhering to the processor interface contract,
1481 integrate new statistical techniques through standardized input/output schemas,
1482 incorporate alternative language models via the model abstraction layer, add cus-
1483 tom plot types through the visualization registry, and integrate additional academic
1484 databases through the unified API interface.

1485 The system adopts a contract-driven development approach through the Contract
1486 Definition module, which allows researchers to specify formal contracts that document
1487 data structure expectations and constraints, enable automated validation and error
1488 detection, facilitate code generation and automation, and serve as machine-readable
1489 research protocol documentation. Reproducibility is built into the system’s design,
1490 with every computational operation maintaining provenance information including
1491 version identifiers for code, data, and models; parameter configurations and ran-
1492 dom seeds; execution timestamps and computational environment specifications; and
1493 dependency graphs showing data lineage.

1494 The framework supports progressive automation, accommodating a spectrum from
1495 fully manual to highly automated workflows. Researchers can execute individual func-
1496 tions interactively for exploration, chain functions into semi-automated workflows
1497 with human validation checkpoints, or define fully automated pipelines via Makefiles
1498 for routine analyses. This flexibility makes the system suitable for both exploratory
1499 research requiring human judgment and confirmatory analyses that benefit from
1500 automation.

1501 This section describes the current organization of the infrastructure layer and the
1502 module structure of the proposed framework. The architecture is under active devel-
1503 opment, and what is presented here reflects a partial but stable implementation that

1504 has been used in practice. The purpose of this section is to clarify the engineering
1505 logic and organizational principles of the framework at its current stage, rather than
1506 to document a complete or finalized system.

1507 At the implementation level, the framework adopts an engineering-oriented,
1508 configuration-driven design. Research projects are organized as independent
1509 workspaces, which can be initialized through automated scripts that generate a stan-
1510 dard project skeleton. Core functionalities are not hard-coded into fixed pipelines;
1511 instead, they are largely driven by explicit configuration files that specify how data are
1512 acquired, processed, analyzed, stored, and transformed into research outputs. These
1513 outputs include structured datasets, statistical tables, figures, and LaTeX source files.
1514 This approach allows different research projects to share a common infrastructure
1515 while retaining flexibility in workflow design.

1516 As a representative example, the literature review components provide uni-
1517 fied mechanisms for accessing and managing scholarly data from multiple sources.
1518 The framework defines explicit conventions for literature retrieval, metadata stor-
1519 age, deduplication, and structured organization, enabling collected materials to be
1520 systematically reused for tasks such as literature review, secondary analysis, and
1521 documentation. At this stage, the framework focuses on enforcing consistent data
1522 representation and storage rather than automating interpretive judgments about
1523 content.

1524 For data collection, the framework supports both existing datasets and primary
1525 data acquisition. In the case of existing data, it provides explicit conventions for data
1526 ingestion and storage, ensuring that datasets from different sources are managed under
1527 a unified structure suitable for subsequent analysis. For primary data collection, the
1528 framework implements tools for structured questionnaire definition, response collec-
1529 tion, and result storage. The emphasis is on maintaining consistent variable definitions,
1530 data schemas, and machine-readable representations, rather than on presentation or
1531 layout concerns.

1532 Data analysis functionality is organized around shared conventions for data
1533 representation and processing. The framework provides modular components for
1534 data cleaning, transformation, and analysis, all of which adhere to standardized
1535 input–output interfaces. Analytical results are stored in structured formats and remain
1536 explicitly linked to their source data and processing steps, allowing results to be traced,
1537 reused, and incorporated into later stages of the research workflow.

1538 At the stage of result production, the framework supports the automated genera-
1539 tion of research artifacts, including statistical tables, figures, and LaTeX source files
1540 suitable for direct inclusion in manuscripts. By embedding result generation into the
1541 same configuration-driven workflow, the framework reduces manual transcription and
1542 helps maintain consistency between analysis and presentation.

1543 Overall, the infrastructure organization described here is intended to demonstrate
1544 the core organizational logic of the framework at its current stage: a configuration-
1545 driven, modular approach that integrates data acquisition, processing, analysis,
1546 storage, and output generation within a single structured workflow. The implemen-
1547 tation should be understood as evolving, with the emphasis placed on architectural
1548 clarity and practical feasibility rather than completeness.

1549 4 Illustrative Case Studies

1550 Case studies in this section are intended as an illustrative example rather than a
1551 fully realized empirical application. It demonstrates how the proposed framework is
1552 designed to support configuration-driven analysis workflows and automated artifact
1553 generation, using a simplified and partially implemented scenario.

1554 4.1 Illustrative Example: Automated Data Analysis and LaTeX 1555 Code Generation

1556 This section demonstrates the practical application of the propose framework through
1557 a case study on automated data analysis and LaTeX code generation. The case study
1558 illustrates how researchers can leverage the framework's configuration-driven approach
1559 to establish reproducible analysis workflows while minimizing manual coding efforts.

1560 4.1.1 Project Initialization and Structure

1561 The analysis workflow begins with project initialization using the framework's automa-
1562 tion scripts. Researchers execute a simple command that creates a standardized project
1563 structure, establishing the foundation for organized and reproducible analysis:

```
1564 python create_project.py --name "education_achievement_study" \  
1565                             --template "quantitative_social_science"
```

1566 This command generates the following project structure:

```
1567 education_achievement_study/  
1568 |-- data/  
1569 |   |-- config/           # Configuration files for data processing  
1570 |   |-- processed/        # Cleaned and transformed data  
1571 |   |-- raw/              # Original, unmodified data  
1572 |-- results/  
1573 |   |-- figures/          # Generated visualizations  
1574 |   |-- models/           # Statistical model objects  
1575 |   |-- statistics/       # Numerical results and summaries  
1576 |   |-- tables/           # Publication-ready LaTeX tables  
1577 |-- statistics/  
1578     |-- config/           # Analysis specification files
```

1579 The standardized structure ensures consistency across projects, facilitates collab-
1580 oration, and enables automated processing through well-defined input and output
1581 locations.

1582 4.1.2 Data Processing Configuration

1583 Within the `data/config/` directory, researchers define their data processing pipeline
1584 using JSON configuration files. These configurations specify the sequence of transfor-
1585 mations to be applied to raw data, along with detailed variable specifications:

```
1586 // data/config/processing_pipeline.json
1587 {
1588   "steps": [
1589     "handle_missing",
1590     "encode_categorical",
1591     "scale_numeric",
1592     "extract_text_features",
1593     "save_processed"
1594   ]
1595 }

1596 // data/config/variable_specification.json
1597 {
1598   "variables": {
1599     "id": {"type": "ratio", "missing": "drop"},
1600     "age": {"type": "interval", "missing": "mean", "scaling": "standard"},
1601     "gender": {"type": "nominal", "missing": "most_frequent"},
1602     "income": {"type": "ratio", "missing": "mean", "scaling": "minmax"},
1603     "education": {"type": "ordinal", "missing": "most_frequent",
1604                  "order": ["Bachelor", "Master", "PhD"]},
1605     "comments": {"type": "descriptive", "missing": "most_frequent"}
1606   }
1607 }
```

1608 These configuration files enable researchers to declaratively specify their data pro-
1609 cessing requirements without writing procedural code. The system interprets these
1610 specifications and executes the corresponding transformations, ensuring consistent
1611 application of preprocessing steps across different datasets and analysis runs.

1612 4.1.3 Statistical Analysis Configuration

1613 The heart of the automated analysis workflow resides in the `statistics/config/`
1614 directory, where researchers define their analytical plans. The framework supports a
1615 comprehensive range of statistical methods through structured configuration files. For
1616 multilevel modeling of educational achievement data, researchers might specify:

```
1617 // statistics/config/multilevel_models.json
1618 {
1619   "random_intercept_students": {
1620     "type": "random_intercept",
```

```

1621     "dependent": "test_score",
1622     "fixed_effects": ["student_ses", "gender", "prior_achievement"],
1623     "group": "school_id",
1624     "output_file": "random_intercept_students.json"
1625 },
1626
1627 "growth_curve_reading": {
1628     "type": "growth_curve",
1629     "dependent": "reading_score",
1630     "time": "wave",
1631     "fixed_effects": ["intervention", "baseline_score"],
1632     "group": "student_id",
1633     "time_as_random": true,
1634     "output_file": "growth_curve_reading.json"
1635 }
1636 }

```

1637 For hypothesis testing and survey analysis, additional configurations define reliability assessment, factor analysis, and validity testing:

```

1639 // statistics/config/survey_analysis.json
1640 {
1641     "reliability_job_satisfaction": {
1642         "type": "reliability",
1643         "scales": {
1644             "job_satisfaction": ["js1", "js2", "js3", "js4", "js5"],
1645             "organizational_commitment": ["oc1", "oc2", "oc3", "oc4", "oc5", "oc6"]
1646         },
1647         "calculate_if_deleted": true,
1648         "output_file": "reliability_analysis.json"
1649     }
1650 }

```

1651 The framework also includes validation configurations for assessing data quality and statistical assumptions:

```

1653 // statistics/config/validation_checks.json
1654 {
1655     "data_quality_check": {
1656         "type": "quality_report",
1657         "variables": ["income", "age", "education", "test_score"],
1658         "output_file": "quality_report.json"
1659     },
1660
1661     "normality_tests": {

```

```

1662     "type": "normality",
1663     "variables": ["income", "test_score", "age", "experience"],
1664     "tests": ["shapiro", "anderson", "jarque_bera"],
1665     "output_file": "normality_tests.json"
1666 }
1667 }

```

1668 These configuration files serve as executable documentation of the analysis plan,
1669 making research intentions explicit and reproducible. Researchers can modify analyses
1670 by editing configuration files rather than rewriting code, reducing the potential for
1671 errors and ensuring consistency across replications.

1672 4.1.4 Automated Execution and Output Generation

1673 Once configurations are defined, researchers execute the complete analysis pipeline
1674 with a single command:

```

1675 make analyze # Executes all configured analyses sequentially

```

1676 The framework processes each configuration file, applying the specified statistical
1677 methods to the preprocessed data. Results are automatically saved in structured for-
1678 mats within the `results/` directory. For example, a regression analysis produces both
1679 machine-readable JSON output and human-readable text summaries:

```

1680 // results/statistics/regression_income.json
1681 {
1682     "model_type": "OLS",
1683     "dependent": "income",
1684     "independents": ["age", "education", "gender"],
1685     "r_squared": 0.484,
1686     "adj_r_squared": -1.064,
1687     "n_observations": 5,
1688     "coefficients": {
1689         "const": {"coef": -0.2604, "std_err": 1.201, "p_value": 0.864},
1690         "age": {"coef": -0.0104, "std_err": 0.469, "p_value": 0.986},
1691         "education": {"coef": 0.4896, "std_err": 0.885, "p_value": 0.678},
1692         "gender_M": {"coef": 0.5104, "std_err": 0.885, "p_value": 0.667}
1693     }
1694 }

```

1695 Simultaneously, the framework generates LaTeX code for inclusion in research
1696 manuscripts. For descriptive statistics, it produces formatted tables:

```

1697 % results/tables/descriptive_statistics.tex
1698 \begin{table}[ht]
1699 \centering

```

```

1700 \caption{Descriptive Statistics by Gender}
1701 \label{tab:descriptive_gender}
1702 \begin{tabular}{lccccccc}
1703 \toprule
1704 & \multicolumn{4}{c}{Age} & \multicolumn{3}{c}{Income} \\
1705 \cmidrule(lr){2-5} \cmidrule(lr){6-9}
1706 Gender & N & Mean & SD & Range & N & Mean & SD & Range \\
1707 \midrule
1708 Female & 2 & 0.50 & 1.41 & [-0.5, 1.0] & 2 & 0.47 & 0.04 & [0.44, 0.50] \\
1709 Male & 3 & -0.33 & 1.04 & [-1.5, 0.5] & 3 & 0.42 & 0.52 & [0.00, 1.00] \\
1710 \bottomrule
1711 \end{tabular}
1712 \end{table}

```

1713 For regression results, it creates publication-ready coefficient tables:

```

1714 % results/tables/regression_results.tex
1715 \begin{table}[ht]
1716 \centering
1717 \caption{OLS Regression Results for Income Prediction}
1718 \label{tab:regression_income}
1719 \begin{tabular}{lccc}
1720 \toprule
1721 & Coefficient & Std. Error & p-value \\
1722 \midrule
1723 Constant & -0.260 & 1.201 & 0.864 \\
1724 Age & -0.010 & 0.469 & 0.986 \\
1725 Education & 0.490 & 0.885 & 0.678 \\
1726 Gender (Male) & 0.510 & 0.885 & 0.667 \\
1727 \midrule
1728 Observations & \multicolumn{3}{c}{5} \\
1729 R-squared & \multicolumn{3}{c}{0.484} \\
1730 Adjusted R-squared & \multicolumn{3}{c}{-1.064} \\
1731 \bottomrule
1732 \end{tabular}
1733 \end{table}

```

1734 5 Related Work

1735 The proposed framework builds upon and extends several categories of existing
1736 research infrastructure. This section reviews statistical software packages, quali-
1737 tative and mixed-methods tools, computational social science platforms, workflow
1738 management systems, and emerging LLM-augmented research tools, positioning our
1739 contributions within this landscape.

1740 5.1 Statistical Software Packages

1741 Statistical software packages form the foundation of quantitative social science
1742 research. R, a free and open-source environment for statistical computing and graph-
1743 ics, has become dominant in academic research due to its extensive package ecosystem
1744 (over 19,000 packages on CRAN), flexible programming capabilities, and strong com-
1745 munity support [172, 173]. R’s strengths lie in its comprehensive statistical methods,
1746 publication-quality visualization through packages like ggplot2 [163], and reproducible
1747 research workflows via R Markdown [174]. However, R’s steep learning curve, incon-
1748 sistent package quality, and memory limitations for large datasets present challenges
1749 for novice users and computationally intensive tasks.

1750 Commercial packages offer alternative approaches with different trade-offs. SPSS
1751 (Statistical Package for the Social Sciences) provides an intuitive graphical interface
1752 emphasizing point-and-click operations, making it accessible for researchers without
1753 programming backgrounds [175]. Its menu-driven approach and extensive documen-
1754 tation support straightforward analyses but limit flexibility for custom methods or
1755 complex workflows. Stata combines command-line and GUI interfaces, offering robust
1756 panel data and econometric capabilities particularly valued in economics and political
1757 science [176]. Its integrated data management, consistent syntax, and excellent docu-
1758 mentation facilitate reproducible research, though its proprietary license and limited
1759 extensibility compared to R constrain adoption in some fields. SAS (Statistical Analy-
1760 sis System) dominates in pharmaceutical research and large-scale enterprise analytics
1761 due to its reliability, performance with massive datasets, and regulatory compliance
1762 features [177]. However, its high licensing costs, dated programming paradigms, and
1763 limited community ecosystem restrict its use in academic social science.

1764 Our framework distinguishes itself from these packages in several key dimen-
1765 sions. While traditional statistical software focuses primarily on data analysis phases,
1766 our system integrates support across the entire research lifecycle from question dis-
1767 covery through dissemination. Rather than replacing these established tools, the
1768 framework provides orchestration capabilities that leverage their analytical strengths
1769 while addressing gaps in literature synthesis, research design support, multimodal
1770 data integration, and automated documentation generation. The LLM-augmented
1771 components offer intelligent assistance for tasks poorly supported by conventional
1772 statistical packages—literature review synthesis, concept operationalization guidance,
1773 and interpretation of complex analytical results in theoretical context.

1774 5.2 Qualitative and Mixed-Methods Tools

1775 Qualitative data analysis software addresses research paradigms beyond quantitative
1776 methods. NVivo supports systematic coding and analysis of textual, audio, video,
1777 and image data through features including hierarchical coding schemes, query tools
1778 for pattern discovery, and visualization of thematic relationships [178]. Its strengths
1779 in handling unstructured data and supporting grounded theory approaches make it
1780 valuable for interpretive research, though its proprietary format and limited quanti-
1781 tative capabilities constrain integration with statistical workflows. MAXQDA offers
1782 similar qualitative analysis features while emphasizing mixed-methods integration

1783 through tools connecting qualitative codes with quantitative variables, enabling joint
1784 displays and mixed-methods matrices [179]. Its flexibility across paradigms appeals
1785 to researchers combining interpretive and quantitative approaches, but effective use
1786 requires expertise in both methodological traditions.

1787 Atlas.ti provides powerful tools for coding, memo-writing, and network visualiza-
1788 tion of qualitative data, with particular strength in visualizing conceptual relationships
1789 through its network view feature [180]. Like its competitors, Atlas.ti excels at orga-
1790 nizing and analyzing unstructured data but offers limited capabilities for quantitative
1791 analysis beyond basic descriptive statistics. All three platforms share common limita-
1792 tions: steep learning curves, high licensing costs, proprietary data formats hindering
1793 interoperability, and weak integration with quantitative analytical pipelines.

1794 Our framework addresses integration gaps between qualitative and quantitative
1795 traditions. While we do not replicate the sophisticated coding and interpretive analy-
1796 sis features of dedicated qualitative software, we provide mechanisms for incorporating
1797 qualitative insights into mixed-methods workflows. LLM components can assist in
1798 initial coding of textual data, identify themes for deeper qualitative analysis, and facil-
1799 itate triangulation between qualitative findings and quantitative results. The system
1800 supports importing coded qualitative data from specialized tools for integration with
1801 quantitative analyses, enabling genuine mixed-methods approaches where insights
1802 from different paradigms inform each other systematically rather than remaining in
1803 isolated analytical silos.

1804 5.3 Computational Social Science Platforms

1805 Specialized platforms have emerged to support computational approaches to social sci-
1806 ence research. LIWC (Linguistic Inquiry and Word Count) provides dictionary-based
1807 text analysis quantifying psychological and linguistic dimensions in written or spoken
1808 language [181]. Its extensive validated dictionaries capturing emotional tone, cogni-
1809 tive processes, and social concerns enable efficient analysis of large text corpora, with
1810 applications ranging from personality assessment to mental health screening. However,
1811 LIWC’s dictionary approach lacks contextual sensitivity and cannot capture nuanced
1812 meanings or domain-specific language patterns, limiting validity for specialized corpora
1813 or languages beyond English.

1814 Platforms for social media analysis address distinct analytical needs. Crimson
1815 Hexagon offers enterprise-grade social media monitoring and sentiment analysis
1816 through machine learning classifiers trained on large datasets [182]. Its strengths in
1817 processing massive social media streams and providing real-time insights appeal to
1818 corporate and political applications, but high costs and proprietary algorithms limit
1819 transparency and academic adoption. Netlytic provides more accessible social media
1820 analytics emphasizing network analysis and visualization of online communities [183].
1821 Its free tier and academic focus support social network research, though limited scal-
1822 ability and analytical depth compared to commercial platforms restrict applications
1823 requiring sophisticated modeling.

1824 Our framework complements these specialized platforms through integration rather
1825 than replication. We incorporate capabilities for basic text analysis and social network

metrics while providing interfaces to external platforms like LIWC for specialized analyses. The key distinction lies in situating computational methods within broader research workflows—our system helps researchers identify when computational approaches are appropriate, select suitable methods and tools, interpret results in theoretical context, and integrate computational findings with traditional social science methods. LLM components assist in operationalizing theoretical constructs as computational measures, translating between social science concepts and computational implementations, and synthesizing findings across computational and conventional analytical approaches.

5.4 Workflow Management Systems

Reproducible research platforms have transformed how researchers document and share analytical processes. Jupyter notebooks combine code, visualizations, and narrative text in interactive documents supporting exploratory analysis and communication [184]. Their language-agnostic architecture (supporting Python, R, Julia, and others), interactive execution model, and rich output formats make them popular for data science and computational research. However, Jupyter’s limitations include challenges in version control due to JSON-based format, difficulties managing computational dependencies, weak support for long-running or distributed computations, and limited collaboration features beyond sequential editing.

R Markdown addresses some reproducibility challenges through literate programming documents combining R code with narrative text, rendering to multiple output formats including HTML, PDF, and Word [174]. Its tight integration with R, plain text format facilitating version control, and flexible output options support reproducible reporting. Extensions like bookdown and blogdown enable complex documents and websites. However, R Markdown’s restriction to the R ecosystem, limited interactivity compared to Jupyter, and lack of built-in dependency management constrain some applications.

Galaxy represents a different approach emphasizing graphical workflow construction for bioinformatics analyses [185]. Its visual workflow builder, extensive tool integration, and built-in reproducibility features support complex analytical pipelines without programming. Galaxy’s track record in genomics demonstrates the value of workflow systems for complex, multi-tool analyses, though its domain specificity and limited flexibility for novel methods restrict application beyond bioinformatics.

Our framework extends these workflow management approaches in several directions. While Jupyter and R Markdown excel at documenting analyses, they provide limited support for earlier research phases like literature review, question formulation, and research design. We integrate workflow documentation capabilities with tools supporting the complete research lifecycle. Our LLM-augmented components offer intelligent assistance that workflow platforms lack—suggesting appropriate analytical approaches based on research questions, identifying potential confounds in research designs, and generating interpretation frameworks for results. Unlike Galaxy’s domain-specific focus, our framework maintains flexibility across diverse social science research traditions while providing more structure than general-purpose notebook systems. We also address collaboration challenges through integrated version control, annotation

1870 systems, and mechanisms for sharing research artifacts with appropriate context and
1871 documentation.

1872 5.5 LLM-Augmented Research Tools

1873 Recent years have witnessed the emergence of large language model (LLM) applica-
1874 tions for research support. Elicit positions itself as an AI research assistant helping
1875 researchers find relevant papers, extract key information, and synthesize findings
1876 across documents [186]. Its strengths include rapid literature searches beyond key-
1877 word matching, automatic extraction of study characteristics (methods, sample sizes,
1878 findings), and synthesis capabilities generating summaries across multiple papers.
1879 However, Elicit focuses narrowly on literature review tasks without supporting sub-
1880 sequent research phases, occasionally produces inaccurate extractions or hallucinated
1881 citations, and provides limited customization for domain-specific needs or specialized
1882 methodologies.

1883 Semantic Scholar offers AI-powered search and recommendation emphasizing cita-
1884 tion context and influential papers through features like citation intent classification
1885 and research feed personalization [187]. Its comprehensive coverage of scientific liter-
1886 ature, free access model, and useful features like paper TLDRs (brief summaries) and
1887 citation contexts benefit researchers. However, like traditional databases, Semantic
1888 Scholar primarily supports literature discovery rather than synthesis or downstream
1889 research activities, and its recommendation algorithms can create filter bubbles
1890 reinforcing existing research directions.

1891 Connected Papers visualizes citation networks helping researchers discover related
1892 work through graphical representations of similarity relationships [188]. Its intuitive
1893 visualizations effectively surface relevant papers that keyword searches might miss,
1894 particularly useful for exploring unfamiliar research areas. However, the tool addresses
1895 only the literature discovery phase, provides limited analytical capabilities beyond
1896 visualization, and lacks integration with other research workflow components.

1897 Our framework makes several novel contributions beyond these specialized LLM
1898 tools. First, we provide integrated support across the entire research lifecycle rather
1899 than isolated tasks. LLM assistance for literature review connects directly to research
1900 design tools, suggesting operationalizations based on prior measurement approaches
1901 and identifying design patterns from reviewed studies. Second, we emphasize domain-
1902 specific customization for social science methodologies. While general-purpose AI
1903 assistants lack understanding of social science research norms, statistical methods, and
1904 theoretical traditions, our system incorporates domain knowledge enabling contextually
1905 appropriate suggestions. Third, we prioritize transparency and verification. Rather
1906 than presenting LLM outputs as authoritative, we provide mechanisms for users to
1907 verify suggestions against source materials, understand reasoning processes, and main-
1908 tain critical evaluation. Fourth, we address the complete research workflow including
1909 data collection, processing, analysis, and dissemination—phases unsupported by cur-
1910 rent LLM research tools. Finally, we provide integration architectures enabling LLM
1911 components to work synergistically with established statistical software and special-
1912 ized analytical tools, leveraging their complementary strengths rather than attempting
1913 to replace proven methods with AI alternatives.

1914 The combination of these related work categories reveals a fragmented landscape
1915 where researchers must integrate multiple disconnected tools to support complete
1916 research projects. Our framework addresses this fragmentation through comprehensive
1917 lifecycle support, intelligent assistance via LLM components, and integration architec-
1918 tures connecting specialized tools into coherent workflows. The following sections detail
1919 our design principles and implementation strategies for achieving these objectives.

1920 6 Discussion

1921 6.1 Key Findings and Contributions

1922 This work presents an integrative perspective on the design of computational infras-
1923 tructure for quantitative social science research and translates this perspective into a
1924 preliminary, working research framework. Rather than proposing a single algorithm
1925 or analytical method, the primary contribution lies in the systematic integration of
1926 heterogeneous research activities, which literature exploration, research design, data
1927 processing, statistical analysis, and result dissemination within a unified architectural
1928 framework.

1929 This work proceeds from the premise that fragmented workflows and weak align-
1930 ment between methodological intent and computational execution constitute an
1931 important source of friction in quantitative social science research, particularly with
1932 respect to efficiency and reproducibility. By adopting declarative configurations, mod-
1933 ular pipelines, and automated result generation, the proposed system demonstrates
1934 how methodological transparency and computational automation can coexist.

1935 The unique value proposition of this work lies in its methodological alignment.
1936 The system architecture is explicitly designed to mirror the epistemic structure of
1937 quantitative research rather than imposing a purely software-driven abstraction. In
1938 addition, the integration of LLM-assisted reasoning as an auxiliary component—rather
1939 than as a fully autonomous agent—provides a pragmatic and responsible approach to
1940 leveraging recent advances in artificial intelligence within social science research.

1941 6.2 Limitations

1942 Despite its contributions, this work has several important limitations. From a technical
1943 perspective, the current system requires researchers to explicitly specify dataset struc-
1944 tures, preprocessing logic, and analytical pipelines through configuration files. While
1945 this approach enhances reproducibility and auditability, it introduces a non-trivial
1946 learning curve, particularly for researchers with limited programming experience.

1947 Methodologically, the system focuses primarily on quantitative and computational
1948 research paradigms. Qualitative methods, mixed-method designs, and interpretive
1949 workflows are only indirectly supported and would require additional conceptual and
1950 technical extensions. Moreover, the current implementation prioritizes flexibility and
1951 extensibility over performance optimization, which may limit scalability for extremely
1952 large datasets or real-time analysis scenarios.

1953 In terms of scope, this work should be understood as a foundational framework
1954 rather than a comprehensive solution. Many design choices remain open, and the

1955 system does not aim to standardize research practices or replace existing domain-
1956 specific tools. Instead, it provides a flexible infrastructure that can coexist with, and
1957 build upon, existing methodological ecosystems.

1958 **6.3 Ethical Considerations**

1959 The increasing automation of research workflows raises important ethical considera-
1960 tions, particularly in the context of social science. Data privacy and security are critical
1961 concerns, as social science datasets often contain sensitive personal or institutional
1962 information. The proposed system emphasizes explicit data specification and controlled
1963 data processing pipelines, but responsibility for ethical data handling ultimately
1964 remains with researchers and institutions.

1965 Bias in automated analysis is another important issue. Configuration-driven
1966 pipelines and LLM-assisted reasoning may inadvertently encode methodological biases,
1967 assumptions, or errors if not carefully designed and reviewed. To mitigate this risk, the
1968 system is intentionally designed to expose intermediate representations and analytical
1969 decisions, enabling inspection and critical reflection rather than opaque automation.

1970 With respect to responsible AI usage, this work adopts a conservative stance.
1971 LLMs are positioned as supportive tools for reasoning, summarization, and explora-
1972 tion, rather than as authoritative decision-makers. This design choice reflects the
1973 view that scientific judgment, hypothesis formulation, and interpretation should
1974 remain human-centered, particularly in domains with significant social and ethical
1975 implications.

1976 **6.4 Future Directions**

1977 Several promising directions for future work emerge from this study. A primary objec-
1978 tive is to reduce the cognitive and technical overhead associated with configuration-
1979 based workflows. Potential approaches include the development of graphical user
1980 interfaces, interactive configuration assistants, and higher-level abstractions that
1981 preserve transparency while improving usability.

1982 From a technical standpoint, future enhancements may include support for
1983 heterogeneous execution environments, improved resource management for LLM com-
1984 ponents, and tighter integration with emerging computational technologies. Extending
1985 the system to better support mixed-method research and cross-disciplinary workflows
1986 also represents an important avenue for development.

1987 Finally, the long-term success of an integrative research framework depends on
1988 community engagement. Future work will explore open development models, shared
1989 configuration libraries, and collaborative governance structures that allow researchers
1990 to contribute modules, best practices, and methodological extensions. Through collec-
1991 tive development, the proposed framework has the potential to evolve into a shared
1992 infrastructure supporting diverse forms of quantitative social science research.

7 Conclusion

This paper presents an ongoing effort towards the design and implementation of an integrative research framework for quantitative social science. Motivated by the increasing complexity, scale, and computational intensity of contemporary social science research, we aim to provide a unified infrastructure that supports the full research lifecycle, from literature exploration and research design to data analysis and dissemination.

We first examine the methodological structure of quantitative social science research and identify a set of recurring structural challenges in contemporary research workflows. These include fragmentation across tools and data representations, weak integration between different phases of the research lifecycle, difficulties in maintaining reproducibility as a structural property of workflows, steep learning curves associated with increasingly complex analytical methods, and growing challenges in managing heterogeneous and large-scale data. Based on these perspectives, we propose a modular, pipeline-oriented system architecture that emphasizes extensibility, transparency, and reproducibility. The system design integrates automated data processing, statistical analysis, LLM-assisted reasoning, and publication-ready output generation within a coherent framework. Through representative case studies, we demonstrate the feasibility of configuration-driven analysis pipelines and automated LaTeX result generation, illustrating the potential of the proposed approach to streamline quantitative research workflows.

It is important to emphasize that the proposed system should be understood as a preliminary and evolving framework rather than a finalized solution. Several challenges remain. First, the current design requires researchers to explicitly describe datasets, preprocessing steps, and analytical pipelines through structured configurations. While such explicit specification supports transparency and auditability, it also introduces an initial learning cost. Reducing this barrier through improved abstractions, better defaults, or graphical user interfaces represents an important direction for future work. Second, the present LLM integration relies on local deployment and is therefore constrained by available computational resources. Extending support to heterogeneous execution environments and improving resource efficiency remain open technical questions.

Despite these limitations, this work establishes a foundational framework for exploring integrative and automated research infrastructure in quantitative social science. By aligning system architecture with methodological practice and emphasizing declarative specification, modular organization, and automated artifact generation, the proposed framework provides a flexible basis for further development. We hope that this work can serve both as a practical reference point for researchers experimenting with structured computational workflows and as a conceptual contribution to ongoing discussions at the intersection of computational infrastructure, methodological rigor, and social science research.

2034 References

- 2035 [1] Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Chris-
2036 takis, N., Contractor, N., Fowler, J., Gutmann, M., *et al.*: Computational social
2037 science. *Science* **323**(5915), 721–723 (2009)
- 2038 [2] Edelman, A., Wolff, T., Montagne, D., Bail, C.A.: Computational social science
2039 and sociology. *Annual Review of Sociology* **46**, 61–81 (2020)
- 2040 [3] Salganik, M.J.: *Bit by Bit: Social Research in the Digital Age*. Princeton
2041 University Press, Princeton, NJ (2018)
- 2042 [4] Ruths, D., Pfeffer, J.: Social media for large studies of behavior. *Science*
2043 **346**(6213), 1063–1064 (2014)
- 2044 [5] Borgatti, S.P., Mehra, A., Brass, D.J., Labianca, G.: Network analysis in the
2045 social sciences. *Science* **323**(5916), 892–895 (2009)
- 2046 [6] Grimmer, J., Stewart, B.M.: Text as data: The promise and pitfalls of automatic
2047 content analysis methods for political texts. *Political Analysis* **21**(3), 267–297
2048 (2013)
- 2049 [7] Molina, M., Garip, F.: Machine learning for sociology. *Annual Review of*
2050 *Sociology* **45**, 27–45 (2019)
- 2051 [8] Watts, D.J., Dodds, P.S.: Influentials, networks, and public opinion formation.
2052 *Journal of Consumer Research* **34**(4), 441–458 (2007)
- 2053 [9] King, G.: Ensuring the data-rich future of the social sciences. *Science* **331**(6018),
2054 719–721 (2011)
- 2055 [10] Howison, J., Herbsleb, J.D.: Scientific software production: incentives and collab-
2056 oration. In: *Proceedings of the ACM 2011 Conference on Computer Supported*
2057 *Cooperative Work*, New York, NY, pp. 513–522 (2011). ACM
- 2058 [11] McKinney, W.: Data structures for statistical computing in python. In: *Pro-*
2059 *ceedings of the 9th Python in Science Conference (SciPy 2010)*, pp. 51–56
2060 (2010)
- 2061 [12] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel,
2062 O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., *et al.*: Scikit-learn:
2063 Machine learning in python. *Journal of Machine Learning Research* **12**, 2825–
2064 2830 (2011)
- 2065 [13] R Core Team: *R: A Language and Environment for Statistical Computing*. R
2066 Foundation for Statistical Computing, Vienna, Austria (2013). R Foundation for
2067 Statistical Computing

- 2068 [14] Bastian, M., Heymann, S., Jacomy, M.: Gephi: an open source software for
2069 exploring and manipulating networks. In: Proceedings of the International AAAI
2070 Conference on Web and Social Media, vol. 3, pp. 361–362 (2009). AAAI Press
- 2071 [15] Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin,
2072 N., Schwikowski, B., Ideker, T.: Cytoscape: a software environment for integrated
2073 models of biomolecular interaction networks. *Genome Research* **13**(11), 2498–
2074 2504 (2003)
- 2075 [16] Kitchin, R.: Big data, new epistemologies and paradigm shifts. *Big Data &*
2076 *Society* **1**(1), 2053951714528481 (2014)
- 2077 [17] Peng, R.D.: Reproducible research in computational science. *Science* **334**(6060),
2078 1226–1227 (2011)
- 2079 [18] Stodden, V., Borwein, J., Bailey, D.H.: Setting the default to reproducible: repro-
2080 ducibility in computational and experimental mathematics. *SIAM News* **46**(5),
2081 4–6 (2013)
- 2082 [19] Boell, S.K., Cecez-Kecmanovic, D.: A hermeneutic approach for conducting lit-
2083 erature reviews and literature searches. *Communications of the Association for*
2084 *Information Systems* **34**, 257–286 (2014)
- 2085 [20] Grinnell Jr, R.M., Unrau, Y.A.: *The Practice of Social Research*. Oxford
2086 University Press, New York (2018)
- 2087 [21] Freese, J.: Replication in social science. *Contexts* **6**(3), 59–62 (2007)
- 2088 [22] Ioannidis, J.P.: Why most published research findings are false. *PLoS Medicine*
2089 **2**(8), 124 (2005)
- 2090 [23] Nosek, B.A., Alter, G., Banks, G.C., Borsboom, D., Bowman, S.D., Breckler,
2091 S.J., Buck, S., Chambers, C.D., Chin, G., Christensen, G., *et al.*: Promoting an
2092 open research culture. *Science* **348**(6242), 1422–1425 (2015)
- 2093 [24] Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K.M., Gerber, A.,
2094 Glennerster, R., Green, D.P., Humphreys, M., Imbens, G., *et al.*: Promoting
2095 transparency in social science research. *Science* **343**(6166), 30–31 (2014)
- 2096 [25] Christensen, G., Miguel, E.: Transparency, reproducibility, and the credibility
2097 of economics research. *Journal of Economic Literature* **56**(3), 920–980 (2018)
- 2098 [26] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen,
2099 T., Lin, Z., Gimelshein, N., Antiga, L., *et al.*: Pytorch: An imperative style,
2100 high-performance deep learning library. In: *Advances in Neural Information*
2101 *Processing Systems*, vol. 32 (2019)
- 2102 [27] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M.,

- 2103 Ghemawat, S., Irving, G., Isard, M., *et al.*: Tensorflow: A system for large-scale
2104 machine learning. In: 12th USENIX Symposium on Operating Systems Design
2105 and Implementation (OSDI 16), Savannah, GA, pp. 265–283 (2016). USENIX
2106 Association
- 2107 [28] Honnibal, M., Montani, I.: spaCy 2: Natural Language Understanding with
2108 Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing.
2109 Software documentation (2017). <https://spacy.io>
- 2110 [29] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P.,
2111 Rault, T., Louf, R., Funtowicz, M., *et al.*: Huggingface’s transformers: State-of-
2112 the-art natural language processing. arXiv preprint arXiv:1910.03771 (2019)
- 2113 [30] Hagberg, A., Swart, P., S Chult, D.: Exploring network structure, dynamics,
2114 and function using networkx. In: Proceedings of the 7th Python in Science
2115 Conference, vol. 2008. Pasadena, CA, pp. 11–15 (2008). SciPy
- 2116 [31] Csardi, G., Nepusz, T., *et al.*: The igraph software package for complex network
2117 research. InterJournal, complex systems **1695**(5), 1–9 (2006)
- 2118 [32] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P.,
2119 Neelakantan, A., Shyam, P., Sastry, G., Askell, A., *et al.*: Language models
2120 are few-shot learners. Advances in Neural Information Processing Systems **33**,
2121 1877–1901 (2020)
- 2122 [33] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep
2123 bidirectional transformers for language understanding. In: Proceedings of the
2124 2019 Conference of the North American Chapter of the Association for Com-
2125 putational Linguistics, Minneapolis, MN, pp. 4171–4186 (2019). Association for
2126 Computational Linguistics
- 2127 [34] Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., Yang, D.: Can large
2128 language models transform computational social science? Computational Lin-
2129 guistics **50**(1), 237–291 (2024)
- 2130 [35] Wing, J.M.: Computational thinking. Communications of the ACM **49**(3), 33–35
2131 (2006)
- 2132 [36] Lazer, D., Kennedy, R., King, G., Vespignani, A.: The parable of google flu:
2133 traps in big data analysis. Science **343**(6176), 1203–1205 (2014)
- 2134 [37] Boyd, D., Crawford, K.: Critical questions for big data: Provocations for a cul-
2135 tural, technological, and scholarly phenomenon. Information, Communication &
2136 Society **15**(5), 662–679 (2012)
- 2137 [38] Groves, R.M.: Three eras of survey research. Public Opinion Quarterly **75**(5),
2138 861–871 (2011)

- 2139 [39] Dean, J., Ghemawat, S.: Mapreduce: simplified data processing on large clusters.
2140 Communications of the ACM **51**(1), 107–113 (2008)
- 2141 [40] Creswell, J.W., Creswell, J.D.: Research Design: Qualitative, Quantitative, and
2142 Mixed Methods Approaches, 5th edn. SAGE Publications, Thousand Oaks, CA
2143 (2017)
- 2144 [41] Adcock, R., Collier, D.: Measurement validity: A shared standard for qualitative
2145 and quantitative research. American Political Science Review **95**(3), 529–546
2146 (2001)
- 2147 [42] Fowler, F.J.: Survey Research Methods, 5th edn. SAGE Publications, Thousand
2148 Oaks, CA (2013)
- 2149 [43] Little, R.J., Rubin, D.B.: Statistical Analysis with Missing Data, 3rd edn. John
2150 Wiley & Sons, ??? (2019)
- 2151 [44] Breiman, L.: Statistical modeling: The two cultures. Statistical Science **16**(3),
2152 199–231 (2001)
- 2153 [45] Watts, D.J., Strogatz, S.H.: Collective dynamics of ‘small-world’ networks.
2154 Nature **393**(6684), 440–442 (1998) <https://doi.org/10.1038/30918>
- 2155 [46] Brownson, R.C., Colditz, G.A., Proctor, E.K.: Dissemination and Implemen-
2156 tation Research in Health: Translating Science to Practice, 2nd edn. Oxford
2157 University Press, New York (2018)
- 2158 [47] Gentzkow, M., Shapiro, J.M.: Code and data for the social sciences: A prac-
2159 titioner’s guide. Journal of Economic Perspectives **28**(2), 75–90 (2014) <https://doi.org/10.1257/jep.28.2.75>
- 2160
- 2161 [48] Feng, X.: Social Research Methods, 5th edn. China Renmin University Press,
2162 Beijing (2018)
- 2163 [49] Buuren, S.: Flexible Imputation of Missing Data, 2nd edn. CRC Press, ???
2164 (2018)
- 2165 [50] Wickham, H., Grolemund, G.: R for Data Science: Import, Tidy, Transform,
2166 Visualize, and Model Data. O’Reilly Media, ??? (2016)
- 2167 [51] Harron, K., Goldstein, H., Dibben, C.: Methodological Developments in Data
2168 Linkage. John Wiley & Sons, ??? (2017)
- 2169 [52] Jolliffe, I.T., Cadima, J.: Principal Component Analysis, 2nd edn. Springer, ???
2170 (2016)
- 2171 [53] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word
2172 representations in vector space. In: International Conference on Learning

- 2173 Representations (ICLR) (2013)
- 2174 [54] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep
2175 convolutional neural networks. In: *Advances in Neural Information Processing*
2176 Systems, vol. 25 (2012)
- 2177 [55] Newman, M.: *Networks*. Oxford University Press, Oxford (2018)
- 2178 [56] Sweeney, L.: k-anonymity: A model for protecting privacy. *International Journal*
2179 *of Uncertainty, Fuzziness and Knowledge-Based Systems* **10**(05), 557–570 (2002)
- 2180 [57] Dwork, C., Roth, A.: The algorithmic foundations of differential privacy.
2181 *Foundations and Trends in Theoretical Computer Science* **9**(3–4), 211–407
2182 (2014)
- 2183 [58] Tukey, J.W.: *Exploratory Data Analysis*. Addison-Wesley, ??? (1977)
- 2184 [59] Nelder, J.A., Wedderburn, R.W.: Generalized linear models. *Journal of the Royal*
2185 *Statistical Society Series A: Statistics in Society* **135**(3), 370–384 (1972)
- 2186 [60] Hastie, T.J., Tibshirani, R.J.: Generalized additive models. *Statistical Science*
2187 **1**, 297–318 (1990)
- 2188 [61] Raudenbush, S.W., Bryk, A.S.: *Hierarchical Linear Models: Applications and*
2189 *Data Analysis Methods*, 2nd edn. Sage, ??? (2002)
- 2190 [62] Baron, R.M., Kenny, D.A.: The moderator–mediator variable distinction in
2191 social psychological research: Conceptual, strategic, and statistical considera-
2192 tions. *Journal of Personality and Social Psychology* **51**(6), 1173–1182 (1986)
2193 <https://doi.org/10.1037/0022-3514.51.6.1173>
- 2194 [63] Imai, K., Keele, L., Yamamoto, T.: A general approach to causal mediation
2195 analysis. *Psychological Methods* **15**(4), 309–334 (2010)
- 2196 [64] Aiken, L.S., West, S.G.: *Multiple Regression: Testing and Interpreting Interac-*
2197 *tions*. Sage, ??? (1991)
- 2198 [65] Preacher, K.J., Rucker, D.D., Hayes, A.F.: Addressing moderated media-
2199 tion hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral*
2200 *Research* **42**(1), 185–227 (2007)
- 2201 [66] Angrist, J.D., Pischke, J.-S.: *Mostly Harmless Econometrics: An Empiricist’s*
2202 *Companion*. Princeton University Press, ??? (2009)
- 2203 [67] Angrist, J.D., Imbens, G.W., Rubin, D.B.: Identification of causal effects using
2204 instrumental variables. *Journal of the American Statistical Association* **91**(434),
2205 444–455 (1996)

- 2206 [68] Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in
2207 observational studies for causal effects. *Biometrika* **70**(1), 41–55 (1983) <https://doi.org/10.1093/biomet/70.1.41>
2208
- 2209 [69] Imbens, G.W., Lemieux, T.: Regression discontinuity designs: A guide to
2210 practice. *Journal of Econometrics* **142**(2), 615–635 (2008)
- 2211 [70] Abadie, A., Diamond, A., Hainmueller, J.: Synthetic control methods for com-
2212 parative case studies: Estimating the effect of california’s tobacco control
2213 program. *Journal of the American Statistical Association* **105**(490), 493–505
2214 (2010) <https://doi.org/10.1198/jasa.2009.ap08746>
- 2215 [71] Wooldridge, J.M.: *Econometric Analysis of Cross Section and Panel Data*, 2nd
2216 edn. MIT Press, ??? (2010)
- 2217 [72] Mullainathan, S., Spiess, J.: Machine learning: an applied econometric approach.
2218 *Journal of Economic Perspectives* **31**(2), 87–106 (2017)
- 2219 [73] Breiman, L.: Random forests. *Machine Learning* **45**(1), 5–32 (2001) <https://doi.org/10.1023/A:1010933404324>
2220
- 2221 [74] Friedman, J.H.: Greedy function approximation: A gradient boosting machine.
2222 *Annals of Statistics* **29**(5), 1189–1232 (2001) <https://doi.org/10.1214/aos/1013203451>
2223
- 2224 [75] Cortes, C., Vapnik, V.: Support-vector networks. *Machine Learning* **20**(3), 273–
2225 297 (1995) <https://doi.org/10.1007/BF00994018>
- 2226 [76] LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* **521**(7553), 436–444
2227 (2015) <https://doi.org/10.1038/nature14539>
- 2228 [77] Maaten, L., Hinton, G.: Visualizing data using t-sne. *Journal of Machine*
2229 *Learning Research* **9**, 2579–2605 (2008)
- 2230 [78] McInnes, L., Healy, J., Melville, J.: UMAP: Uniform manifold approximation
2231 and projection. *arXiv preprint arXiv:1802.03426* (2018)
- 2232 [79] Dietterich, T.G.: Ensemble methods in machine learning. *International Work-*
2233 *shop on Multiple Classifier Systems*, 1–15 (2000)
- 2234 [80] Wasserman, S., Faust, K.: *Social Network Analysis: Methods and Applications*.
2235 Cambridge University Press, Cambridge (1994)
- 2236 [81] Freeman, L.C.: Centrality in social networks conceptual clarification. *Social*
2237 *Networks* **1**(3), 215–239 (1978)
- 2238 [82] Newman, M.E.J.: Modularity and community structure in networks. *Proceedings*
2239 *of the National Academy of Sciences* **103**(23), 8577–8582 (2006) <https://doi.org/10.1073/pnas.0606164103>

- 2240 [org/10.1073/pnas.0601602103](https://doi.org/10.1073/pnas.0601602103)
- 2241 [83] Luxburg, U.: A tutorial on spectral clustering. *Statistics and Computing* **17**(4),
2242 395–416 (2007) <https://doi.org/10.1007/s11222-007-9033-z>
- 2243 [84] Holland, P.W., Laskey, K.B., Leinhardt, S.: Stochastic blockmodels: First steps.
2244 *Social Networks* **5**(2), 109–137 (1983) [https://doi.org/10.1016/0378-8733\(83\)](https://doi.org/10.1016/0378-8733(83)90021-7)
2245 [90021-7](https://doi.org/10.1016/0378-8733(83)90021-7)
- 2246 [85] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U.:
2247 Network motifs: simple building blocks of complex networks. *Science* **298**(5594),
2248 824–827 (2002)
- 2249 [86] Przulj, N.: Biological network comparison using graphlet degree distribution.
2250 *Bioinformatics* **23**(2), 177–183 (2007)
- 2251 [87] Lusher, D., Koskinen, J., Robins, G.: Exponential Random Graph Models for
2252 Social Networks: Theory, Methods, and Applications. Cambridge University
2253 Press, ??? (2013)
- 2254 [88] Snijders, T.A., Bunt, G.G., Steglich, C.E.: Introduction to stochastic actor-based
2255 models for network dynamics. *Social Networks* **32**(1), 44–60 (2010)
- 2256 [89] Holme, P., Saramäki, J.: Temporal networks. *Physics Reports* **519**(3), 97–125
2257 (2012)
- 2258 [90] Box, G.E.P., Jenkins, G.M., Reinsel, G.C., Ljung, G.M.: Time Series Analysis:
2259 Forecasting and Control, 5th edn. Wiley, Hoboken, NJ (2015)
- 2260 [91] Durbin, J., Koopman, S.J.: Time Series Analysis by State Space Methods, 2nd
2261 edn. Oxford University Press, ??? (2012)
- 2262 [92] Lütkepohl, H.: New Introduction to Multiple Time Series Analysis. Springer,
2263 ??? (2005)
- 2264 [93] Cox, D.R.: Regression models and life-tables. *Journal of the Royal Statistical*
2265 *Society: Series B (Methodological)* **34**(2), 187–220 (1972) [https://doi.org/10.](https://doi.org/10.1111/j.2517-6161.1972.tb00899.x)
2266 [1111/j.2517-6161.1972.tb00899.x](https://doi.org/10.1111/j.2517-6161.1972.tb00899.x)
- 2267 [94] Allison, P.D.: Event History and Survival Analysis, 2nd edn. Sage, ??? (2014)
- 2268 [95] Shannon, C.E.: A mathematical theory of communication. *The Bell System*
2269 *Technical Journal* **27**(3), 379–423 (1948)
- 2270 [96] Cover, T.M., Thomas, J.A.: Elements of Information Theory, 2nd edn. Wiley,
2271 New York (1999)
- 2272 [97] Schreiber, T.: Measuring information transfer. *Physical Review Letters* **85**(2),

- 2273 461–464 (2000) <https://doi.org/10.1103/PhysRevLett.85.461>
- 2274 [98] Granger, C.W.: Investigating causal relations by econometric models and cross-
2275 spectral methods. *Econometrica* **37**(3), 424–438 (1969)
- 2276 [99] Li, M., Vitányi, P.: An Introduction to Kolmogorov Complexity and Its
2277 Applications, 3rd edn. Springer, ??? (2008)
- 2278 [100] Epstein, J.M., Axtell, R.: Growing Artificial Societies: Social Science from the
2279 Bottom Up. MIT Press, ??? (1996)
- 2280 [101] Railsback, S.F., Grimm, V.: Agent-based and Individual-based Modeling: a
2281 Practical Introduction. Princeton University Press, ??? (2011)
- 2282 [102] Kermack, W.O., McKendrick, A.G.: A contribution to the mathematical theory
2283 of epidemics. *Proceedings of the Royal Society of London Series A* **115**(772),
2284 700–721 (1927)
- 2285 [103] Forrester, J.W.: Industrial Dynamics. MIT Press, ??? (1961)
- 2286 [104] Sterman, J.D.: Business Dynamics: Systems Thinking and Modeling for a
2287 Complex World. McGraw-Hill, ??? (2000)
- 2288 [105] Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning:
2289 Data Mining, Inference, and Prediction, 2nd edn. Springer, New York (2009)
- 2290 [106] Rosenbaum, P.R.: Observational Studies, 2nd edn. Springer, ??? (2002)
- 2291 [107] Imbens, G.W., Rubin, D.B.: Causal Inference in Statistics, Social, and Biomed-
2292 ical Sciences. Cambridge University Press, Cambridge (2015)
- 2293 [108] Cooper, H.: Research Synthesis and Meta-Analysis: A Step-by-Step Approach,
2294 5th edn. SAGE Publications, Los Angeles (2015)
- 2295 [109] Borenstein, M., Hedges, L.V., Higgins, J.P.T., Rothstein, H.R.: Introduction to
2296 Meta-Analysis, 2nd edn. Wiley, Chichester (2021)
- 2297 [110] Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mul-
2298 row, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., *et al.*: The
2299 prisma 2020 statement: an updated guideline for reporting systematic reviews.
2300 *International Journal of Surgery* **88**, 105906 (2021)
- 2301 [111] Shadish, W.R., Cook, T.D., Campbell, D.T.: Experimental and Quasi-
2302 experimental Designs for Generalized Causal Inference. Houghton Mifflin,
2303 Boston, MA (2002)
- 2304 [112] Cook, T.D., Campbell, D.T.: Experimental and Quasi-Experimental Designs for
2305 Generalized Causal Inference. Houghton Mifflin, Boston (2002)

- 2306 [113] Cohen, J.: Statistical Power Analysis for the Behavioral Sciences, 2nd edn.
2307 Lawrence Erlbaum Associates, Hillsdale, NJ (1988)
- 2308 [114] Angrist, J.D., Pischke, J.-S.: Mostly Harmless Econometrics: An Empiricist's
2309 Companion. Princeton University Press, Princeton, NJ (2008)
- 2310 [115] Groves, R.M., Fowler Jr, F.J., Couper, M.P., Lepkowski, J.M., Singer, E.,
2311 Tourangeau, R.: Survey Methodology. John Wiley & Sons, Hoboken, NJ (2009)
- 2312 [116] Lynn, P.: Methodology of Longitudinal Surveys. John Wiley & Sons, ??? (2009)
- 2313 [117] Shiffman, S., Stone, A.A., Hufford, M.R.: Ecological momentary assessment.
2314 Annual Review of Clinical Psychology **4**, 1–32 (2008)
- 2315 [118] Salganik, M.J.: Bit by Bit: Social Research in the Digital Age. Princeton
2316 University Press, ??? (2017)
- 2317 [119] Tourangeau, R., Rips, L.J., Rasinski, K.: The Psychology of Survey Response.
2318 Cambridge University Press, Cambridge (2000)
- 2319 [120] Cochran, W.G.: Sampling Techniques, 3rd edn. John Wiley & Sons, New York
2320 (1977)
- 2321 [121] Dillman, D.A., Smyth, J.D., Christian, L.M.: Internet, Phone, Mail, and Mixed-
2322 mode Surveys: The Tailored Design Method, 4th edn. John Wiley & Sons, ???
2323 (2014)
- 2324 [122] Cronbach, L.J.: Coefficient alpha and the internal structure of tests. Psychome-
2325 trika **16**(3), 297–334 (1951) <https://doi.org/10.1007/BF02310555>
- 2326 [123] Brown, T.A.: Confirmatory Factor Analysis for Applied Research, 2nd edn.
2327 Guilford Press, New York (2015)
- 2328 [124] Gelman, A., Hill, J.: Data Analysis Using Regression and Multilevel/Hierarchical
2329 Models. Cambridge University Press, Cambridge (2006)
- 2330 [125] Singer, J.D., Willett, J.B.: Applied Longitudinal Data Analysis: Modeling
2331 Change and Event Occurrence. Oxford University Press, New York (2003)
- 2332 [126] Kline, R.B.: Principles and Practice of Structural Equation Modeling, 4th edn.
2333 Guilford Press, New York (2015)
- 2334 [127] Collins, L.M., Lanza, S.T.: Latent Class and Latent Transition Analysis: With
2335 Applications in the Social, Behavioral, and Health Sciences. John Wiley & Sons,
2336 ??? (2009)
- 2337 [128] Schuman, H., Presser, S.: Questions and Answers in Attitude Surveys: Experi-
2338 ments on Question Form, Wording, and Context. SAGE Publications, Thousand

- 2339 Oaks, CA (1996)
- 2340 [129] Krippendorff, K.: Content Analysis: An Introduction to Its Methodology. Sage
2341 publications, Thousand Oaks, CA (2018)
- 2342 [130] Schreier, M.: Qualitative Content Analysis in Practice. SAGE Publications,
2343 London (2012)
- 2344 [131] Krippendorff, K.: Computing krippendorff’s alpha-reliability. Departmental
2345 Papers (ASC), University of Pennsylvania (2011)
- 2346 [132] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of
2347 Machine Learning Research **3**, 993–1022 (2003)
- 2348 [133] Grootendorst, M.: Bertopic: Neural topic modeling with a class-based tf-idf
2349 procedure. arXiv preprint arXiv:2203.05794 (2022)
- 2350 [134] Liu, B.: Sentiment Analysis and Opinion Mining. Morgan & Claypool, San
2351 Rafael, CA (2012)
- 2352 [135] Nadeau, D., Sekine, S.: A survey of named entity recognition and classification.
2353 Linguisticae Investigationes **30**(1), 3–26 (2007) [https://doi.org/10.1075/li.30.1.](https://doi.org/10.1075/li.30.1.03nad)
2354 03nad
- 2355 [136] Smith, E.: Using Secondary Data in Educational and Social Research. McGraw-
2356 Hill Education, Maidenhead (2008)
- 2357 [137] Anselin, L.: Spatial Econometrics: Methods and Models. Kluwer Academic
2358 Publishers, Dordrecht (1988)
- 2359 [138] Heeringa, S.G., West, B.T., Berglund, P.A.: Applied Survey Data Analysis, 2nd
2360 edn. CRC Press, Boca Raton, FL (2017)
- 2361 [139] Agresti, A., Finlay, B.: Statistical Methods for the Social Sciences. Pearson,
2362 Boston, MA (2018)
- 2363 [140] Wasserman, L.: All of Statistics: A Concise Course in Statistical Inference.
2364 Springer, ??? (2013)
- 2365 [141] Preacher, K.J., Hayes, A.F.: Asymptotic and resampling strategies for assessing
2366 and comparing indirect effects in multiple mediator models. Behavior Research
2367 Methods **40**(3), 879–891 (2008)
- 2368 [142] Burnham, K.P., Anderson, D.R.: Model Selection and Multimodel Inference: A
2369 Practical Information-Theoretic Approach, 2nd edn. Springer, New York (2002)
- 2370 [143] Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge,
2371 MA (2016)

- 2372 [144] Tibshirani, R.: Regression shrinkage and selection via the lasso. *Journal of the*
2373 *Royal Statistical Society: Series B* **58**(1), 267–288 (1996)
- 2374 [145] Ester, M., Kriegel, H.-P., Sander, J., Xu, X.: A density-based algorithm for
2375 discovering clusters in large spatial databases with noise. In: *Proceedings of the*
2376 *Second International Conference on Knowledge Discovery and Data Mining*, pp.
2377 226–231 (1996). AAAI Press
- 2378 [146] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N.,
2379 Kaiser, Ł., Polosukhin, I.: Attention is all you need. *Advances in Neural*
2380 *Information Processing Systems* **30** (2017)
- 2381 [147] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied
2382 to document recognition. *Proceedings of the IEEE* **86**(11), 2278–2324 (1998)
2383 <https://doi.org/10.1109/5.726791>
- 2384 [148] Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Computation*
2385 **9**(8), 1735–1780 (1997) <https://doi.org/10.1162/neco.1997.9.8.1735>
- 2386 [149] Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation
2387 with the graphical lasso. *Biostatistics* **9**(3), 432–441 (2008) [https://doi.org/10.](https://doi.org/10.1093/biostatistics/kxm045)
2388 [1093/biostatistics/kxm045](https://doi.org/10.1093/biostatistics/kxm045)
- 2389 [150] Freeman, L.C.: A set of measures of centrality based on betweenness. *Sociometry*
2390 **40**(1), 35–41 (1977) <https://doi.org/10.2307/3033543>
- 2391 [151] Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of
2392 communities in large networks. *Journal of Statistical Mechanics: Theory and*
2393 *Experiment* **2008**(10), 10008 (2008) [https://doi.org/10.1088/1742-5468/2008/](https://doi.org/10.1088/1742-5468/2008/10/P10008)
2394 [10/P10008](https://doi.org/10.1088/1742-5468/2008/10/P10008)
- 2395 [152] Easley, D., Kleinberg, J.: *Networks, Crowds, and Markets: Reasoning About a*
2396 *Highly Connected World*. Cambridge University Press, Cambridge (2010)
- 2397 [153] Epstein, J.M.: *Generative Social Science: Studies in Agent-based Computational*
2398 *Modeling*. Princeton University Press, Princeton, NJ (2006)
- 2399 [154] Schelling, T.C.: Dynamic models of segregation. *Journal of Mathematical*
2400 *Sociology* **1**(2), 143–186 (1971)
- 2401 [155] Lewin, K.: *Field Theory in Social Science: Selected Theoretical Papers*. Harper
2402 & Brothers, ??? (1951)
- 2403 [156] Schöner, G., Spencer, J.P., Group, D.R., et al.: *Dynamic thinking: A primer on*
2404 *dynamic field theory* (2015)
- 2405 [157] Castellano, C., Fortunato, S., Loreto, V.: Statistical physics of social dynamics.
2406 *Reviews of Modern Physics* **81**(2), 591–646 (2009)

- 2407 [158] Scheffer, M., Bascompte, J., Brock, W.A., Brovkin, V., Carpenter, S.R., Dakos,
2408 V., Held, H., Van Nes, E.H., Rietkerk, M., Sugihara, G.: Early-warning signals
2409 for critical transitions. *Nature* **461**(7260), 53–59 (2009)
- 2410 [159] VanderWeele, T.J.: *Explanation in Causal Inference: Methods for Mediation and*
2411 *Interaction*. Oxford University Press, New York (2015)
- 2412 [160] Pearl, J.: *Causality*. Cambridge University Press, Cambridge (2009)
- 2413 [161] Wager, S., Athey, S.: Estimation and inference of heterogeneous treatment effects
2414 using random forests. *Journal of the American Statistical Association* **113**(523),
2415 1228–1242 (2018)
- 2416 [162] Tufte, E.R.: *The Visual Display of Quantitative Information*, 2nd edn. Graphics
2417 Press, Cheshire, CT (2001)
- 2418 [163] Wickham, H.: *Ggplot2: Elegant Graphics for Data Analysis*. Springer, New York
2419 (2016)
- 2420 [164] Hunter, J.D.: Matplotlib: A 2d graphics environment. *Computing in Science &*
2421 *Engineering* **9**(3), 90–95 (2007)
- 2422 [165] Wilkinson, L., Statistical Inference, T.F., *et al.*: Statistical methods in psy-
2423 chology journals: Guidelines and explanations. *American Psychologist* **54**(8),
2424 594–604 (1999)
- 2425 [166] Rule, A., Birmingham, A., Zuniga, C., Altintas, I., Huang, S.-C., Knight, R.,
2426 Moshiri, N., Nguyen, M.H., Rosenthal, S.B., Pérez, F., *et al.*: Ten simple rules
2427 for writing and sharing computational analyses in jupyter notebooks. *PLoS*
2428 *Computational Biology* **15**(7), 1007007 (2019)
- 2429 [167] Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M.,
2430 Baak, A., Blomberg, N., Boiten, J.-W., Silva Santos, L.B., Bourne, P.E., *et*
2431 *al.*: The fair guiding principles for scientific data management and stewardship.
2432 *Scientific Data* **3**(1), 1–9 (2016)
- 2433 [168] Ram, K.: Git can facilitate greater reproducibility and increased transparency
2434 in science. *Source Code for Biology and Medicine* **8**(1), 1–8 (2013)
- 2435 [169] Viechtbauer, W.: Conducting meta-analyses in r with the metafor package.
2436 *Journal of Statistical Software* **36**(3), 1–48 (2010)
- 2437 [170] Faul, F., Erdfelder, E., Lang, A.-G., Buchner, A.: G* power 3: A flexible statis-
2438 tical power analysis program for the social, behavioral, and biomedical sciences.
2439 *Behavior Research Methods* **39**(2), 175–191 (2007)
- 2440 [171] Bird, S., Klein, E., Loper, E.: *Natural Language Processing with Python:*
2441 *Analyzing Text with the Natural Language Toolkit*. O’Reilly Media, ??? (2009)

- 2442 [172] R Core Team: R: A Language and Environment for Statistical Computing. R
2443 Foundation for Statistical Computing, Vienna, Austria (2023). R Foundation for
2444 Statistical Computing. <https://www.R-project.org/>
- 2445 [173] Wickham, H.: Advanced R, 2nd edn. CRC Press, Boca Raton, FL (2019)
- 2446 [174] Xie, Y., Allaire, J.J., Golemund, G.: R Markdown: The Definitive Guide. CRC
2447 Press, Boca Raton, FL (2018)
- 2448 [175] Pallant, J.: SPSS Survival Manual: A Step by Step Guide to Data Analysis Using
2449 IBM SPSS, 7th edn. Routledge, London (2020)
- 2450 [176] Kohler, U., Kreuter, F.: Data Analysis Using Stata, 3rd edn. Stata Press, College
2451 Station, TX (2012)
- 2452 [177] Delwiche, L.D., Slaughter, S.J.: The Little SAS Book: A Primer, 6th edn. SAS
2453 Institute, Cary, NC (2019)
- 2454 [178] Jackson, K., Bazeley, P.: Qualitative Data Analysis with NVivo, 3rd edn. SAGE
2455 Publications, London (2018)
- 2456 [179] Kuckartz, U., Rädiker, S.: Qualitative Text Analysis: A Systematic Approach.
2457 SAGE Publications, London (2019)
- 2458 [180] Frieze, S.: Qualitative Data Analysis with ATLAS.ti, 3rd edn. SAGE Publica-
2459 tions, London (2019)
- 2460 [181] Pennebaker, J.W., Boyd, R.L., Jordan, K., Blackburn, K.: The development and
2461 psychometric properties of liwc2015. Technical report, University of Texas at
2462 Austin (2015). <https://www.liwc.app>
- 2463 [182] Golder, S.A., Macy, M.W.: Methods for analyzing social media. Annual Review
2464 of Sociology **40**, 153–178 (2014)
- 2465 [183] Gruzdt, A.: Netlytic: Software for automated text and social network analysis. In:
2466 Proceedings of the International Conference on Social Media & Society, London
2467 (2016). ACM
- 2468 [184] Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic,
2469 J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., *et al.*: Jupyter notebooks—a
2470 publishing format for reproducible computational workflows. In: Positioning and
2471 Power in Academic Publishing: Players, Agents and Agendas, pp. 87–90 (2016).
2472 IOS Press
- 2473 [185] Afgan, E., Baker, D., Batut, B., Van Den Beek, M., Bouvier, D., Cech, M.,
2474 Chilton, J., Clements, D., Coraor, N., Grünig, B.A., *et al.*: The galaxy platform
2475 for accessible, reproducible and collaborative biomedical analyses: 2018 update.
2476 Nucleic Acids Research **46**(W1), 537–544 (2018)

- 2477 [186] Ought Inc.: Elicit: The AI Research Assistant. Accessed: 2024-01-15 (2023).
2478 <https://elicit.org>
- 2479 [187] Lo, K., Wang, L.L., Neumann, M., Kinney, R., Weld, D.S.: S2orc: The semantic
2480 scholar open research corpus. In: Proceedings of the 58th Annual Meeting of the
2481 Association for Computational Linguistics, pp. 4969–4983 (2020). Association
2482 for Computational Linguistics
- 2483 [188] Connected Papers: Connected Papers: Visual Tool to Help Researchers Find and
2484 Explore Papers. Accessed: 2024-01-15 (2023). [https://www.connectedpapers.](https://www.connectedpapers.com)
2485 [com](https://www.connectedpapers.com)