

# Foundations of Quantitative Social Science Research

## : Data Processing (Skeleton Draft)

Wanhong Huang

Spring 2026

### Contents

|          |                                                                          |          |
|----------|--------------------------------------------------------------------------|----------|
| <b>1</b> | <b>Learning Objectives</b>                                               | <b>5</b> |
| <b>2</b> | <b>Introduction</b>                                                      | <b>5</b> |
| <b>3</b> | <b>Learning Objectives</b>                                               | <b>7</b> |
| <b>I</b> | <b>Philosophical Foundations</b>                                         | <b>7</b> |
| <b>4</b> | <b>The Ontological Status of Data: Being vs. Representation</b>          | <b>8</b> |
| 4.1      | Data as Entity vs. Data as Relation . . . . .                            | 8        |
| 4.2      | The Map-Territory Distinction in Data . . . . .                          | 8        |
| 4.3      | Realism, Constructivism, and Instrumentalism . . . . .                   | 8        |
| 4.4      | Digital Objects and Their Modes of Existence . . . . .                   | 8        |
| <b>5</b> | <b>Epistemology of Transformation: Knowledge Construction and Loss</b>   | <b>8</b> |
| 5.1      | What Can We Know Through Data? . . . . .                                 | 8        |
| 5.2      | Transformation as Knowledge Construction . . . . .                       | 8        |
| 5.3      | The Inevitability of Loss: What Transformations Hide . . . . .           | 8        |
| 5.4      | Reversibility, Sufficiency, and Information Preservation . . . . .       | 8        |
| <b>6</b> | <b>Ethics and Axiology: Values, Bias, and the Morality of Processing</b> | <b>8</b> |
| 6.1      | Value-Laden Choices in Data Processing . . . . .                         | 8        |
| 6.2      | Algorithmic Bias and Structural Injustice . . . . .                      | 8        |
| 6.3      | Fairness Criteria and Their Tensions . . . . .                           | 8        |
| 6.4      | Responsibility and Accountability in Automated Decisions . . . . .       | 8        |
| <b>7</b> | <b>Logic and Semantics: Meaning, Types, and Symbolic Systems</b>         | <b>8</b> |
| 7.1      | Formal Logic and Data Constraints . . . . .                              | 8        |
| 7.2      | Type Systems: Syntax and Semantics . . . . .                             | 8        |
| 7.3      | The Meaning of Symbols and Grounding Problem . . . . .                   | 8        |
| 7.4      | Compositionality and Semantic Preservation . . . . .                     | 8        |

## II Theoretical Frameworks 8

### 8 Mathematical Foundations 8

|       |                                                                       |   |
|-------|-----------------------------------------------------------------------|---|
| 8.1   | Algebraic Structures: Sets, Groups, and Type Theory . . . . .         | 8 |
| 8.1.1 | Set Theory and Relations . . . . .                                    | 8 |
| 8.1.2 | Groups, Rings, and Fields . . . . .                                   | 8 |
| 8.1.3 | Type Theory and Lambda Calculus . . . . .                             | 8 |
| 8.1.4 | Algebraic Data Types . . . . .                                        | 8 |
| 8.2   | Category Theory: Morphisms, Functors, and Composition . . . . .       | 8 |
| 8.2.1 | Categories, Objects, and Morphisms . . . . .                          | 8 |
| 8.2.2 | Functors and Natural Transformations . . . . .                        | 8 |
| 8.2.3 | Universal Properties and Limits . . . . .                             | 8 |
| 8.2.4 | Monads and Kleisli Categories . . . . .                               | 8 |
| 8.3   | Topological Foundations: Spaces, Connectivity, and Homeomorphisms . . | 8 |
| 8.3.1 | Topological Spaces and Continuity . . . . .                           | 8 |
| 8.3.2 | Connectedness and Compactness . . . . .                               | 8 |
| 8.3.3 | Simplicial Complexes and Cell Decompositions . . . . .                | 8 |
| 8.3.4 | Homotopy and Homology . . . . .                                       | 8 |
| 8.4   | Geometric Foundations: Manifolds, Metrics, and Distance . . . . .     | 8 |
| 8.4.1 | Metric Spaces and Distance Functions . . . . .                        | 8 |
| 8.4.2 | Normed and Hilbert Spaces . . . . .                                   | 8 |
| 8.4.3 | Manifolds and Differential Structures . . . . .                       | 8 |
| 8.4.4 | Riemannian Geometry and Curvature . . . . .                           | 8 |

### 9 Statistical Foundations 8

|       |                                                                            |   |
|-------|----------------------------------------------------------------------------|---|
| 9.1   | Probability Spaces and Measure Theory . . . . .                            | 8 |
| 9.1.1 | Measure Spaces and $\sigma$ -Algebras . . . . .                            | 8 |
| 9.1.2 | Probability Measures and Lebesgue Integration . . . . .                    | 8 |
| 9.1.3 | Conditional Probability and Independence . . . . .                         | 8 |
| 9.1.4 | Filtrations and Adapted Processes . . . . .                                | 8 |
| 9.2   | Random Variables and Distributions . . . . .                               | 8 |
| 9.2.1 | Random Variables and Their Distributions . . . . .                         | 8 |
| 9.2.2 | Expectation, Variance, and Moments . . . . .                               | 8 |
| 9.2.3 | Common Probability Distributions . . . . .                                 | 8 |
| 9.2.4 | Joint, Marginal, and Conditional Distributions . . . . .                   | 8 |
| 9.2.5 | Copulas and Dependence Structures . . . . .                                | 8 |
| 9.3   | Statistical Inference: Estimation, Testing, and Uncertainty Quantification | 8 |
| 9.3.1 | Point Estimation and Properties of Estimators . . . . .                    | 8 |
| 9.3.2 | Confidence Intervals and Credible Regions . . . . .                        | 8 |
| 9.3.3 | Hypothesis Testing and p-Values . . . . .                                  | 8 |
| 9.3.4 | Likelihood Theory and Maximum Likelihood Estimation . . . . .              | 8 |
| 9.3.5 | Bayesian Inference and Prior Specification . . . . .                       | 8 |
| 9.4   | Sampling Theory and Survey Methodology . . . . .                           | 8 |
| 9.4.1 | Sampling Designs: Simple, Stratified, and Cluster Sampling . . .           | 8 |
| 9.4.2 | Sampling Distributions and Central Limit Theorem . . . . .                 | 8 |
| 9.4.3 | Survey Weights and Design Effects . . . . .                                | 8 |
| 9.4.4 | Non-response and Selection Bias . . . . .                                  | 8 |
| 9.5   | Regression, Correlation, and Dependence Structures . . . . .               | 8 |

|            |                                                                              |          |
|------------|------------------------------------------------------------------------------|----------|
| 9.5.1      | Correlation and Causation . . . . .                                          | 8        |
| 9.5.2      | Linear Regression and Least Squares . . . . .                                | 8        |
| 9.5.3      | Generalized Linear Models . . . . .                                          | 8        |
| 9.5.4      | Time Series and Spatial Dependencies . . . . .                               | 8        |
| 9.5.5      | Graphical Models and Conditional Independence . . . . .                      | 8        |
| <b>10</b>  | <b>Information-Theoretic Foundations</b>                                     | <b>8</b> |
| 10.1       | Entropy and Encoding: The Limits of Compression . . . . .                    | 8        |
| 10.1.1     | Shannon Entropy and Differential Entropy . . . . .                           | 8        |
| 10.1.2     | Source Coding Theorem . . . . .                                              | 8        |
| 10.1.3     | Kolmogorov Complexity . . . . .                                              | 8        |
| 10.1.4     | Minimum Description Length . . . . .                                         | 8        |
| 10.2       | Information Loss: Divergence, Bottlenecks, and Irreversibility . . . . .     | 8        |
| 10.2.1     | Kullback-Leibler Divergence . . . . .                                        | 8        |
| 10.2.2     | f-Divergences and Statistical Distance . . . . .                             | 8        |
| 10.2.3     | Data Processing Inequality . . . . .                                         | 8        |
| 10.2.4     | Information Bottleneck Theory . . . . .                                      | 8        |
| 10.3       | Relational Information: Mutual Information and Dependencies . . . . .        | 8        |
| 10.3.1     | Mutual Information and Conditional Mutual Information . . . . .              | 8        |
| 10.3.2     | Transfer Entropy and Directed Information . . . . .                          | 8        |
| 10.3.3     | Information Geometry . . . . .                                               | 8        |
| 10.3.4     | Maximal Correlation and Dependence Measures . . . . .                        | 8        |
| <b>11</b>  | <b>Dynamical Systems Foundations</b>                                         | <b>8</b> |
| 11.1       | State Spaces: Trajectories, Attractors, and Evolution . . . . .              | 8        |
| 11.1.1     | Phase Space and State Variables . . . . .                                    | 8        |
| 11.1.2     | Flows and Orbits . . . . .                                                   | 8        |
| 11.1.3     | Fixed Points and Limit Cycles . . . . .                                      | 8        |
| 11.1.4     | Attractors and Basins of Attraction . . . . .                                | 8        |
| 11.2       | Stochastic Processes: Noise, Uncertainty, and Latent Dynamics . . . . .      | 8        |
| 11.2.1     | Markov Chains and Transition Matrices . . . . .                              | 8        |
| 11.2.2     | Brownian Motion and Diffusion Processes . . . . .                            | 8        |
| 11.2.3     | Hidden Markov Models . . . . .                                               | 8        |
| 11.2.4     | Stochastic Differential Equations . . . . .                                  | 8        |
| 11.3       | Stability and Bifurcation: Sensitivity in Transformation Pipelines . . . . . | 8        |
| 11.3.1     | Lyapunov Stability . . . . .                                                 | 8        |
| 11.3.2     | Bifurcation Theory and Catastrophes . . . . .                                | 8        |
| 11.3.3     | Chaos and Sensitive Dependence . . . . .                                     | 8        |
| 11.3.4     | Robustness and Structural Stability . . . . .                                | 8        |
| <b>III</b> | <b>Practices with Integrated Theory</b>                                      | <b>8</b> |
| <b>12</b>  | <b>Data Integrity and Cleaning</b>                                           | <b>8</b> |
| 12.1       | Error Theory and Detection . . . . .                                         | 8        |
| 12.1.1     | Theoretical Lens: Multi-Perspective Error Analysis . . . . .                 | 8        |
| 12.1.2     | Practice: Detection and Validation Algorithms . . . . .                      | 9        |
| 12.2       | Missing Data . . . . .                                                       | 9        |
| 12.2.1     | Theoretical Lens: The Nature of Absence . . . . .                            | 9        |

|           |                                                               |           |
|-----------|---------------------------------------------------------------|-----------|
| 12.2.2    | Practice: Imputation Algorithms . . . . .                     | 9         |
| 12.3      | Entity Resolution and Deduplication . . . . .                 | 9         |
| 12.3.1    | Theoretical Lens: Identity Across Transformations . . . . .   | 9         |
| 12.3.2    | Practice: Record Linkage Algorithms . . . . .                 | 10        |
| 12.4      | Outlier Detection . . . . .                                   | 10        |
| 12.4.1    | Theoretical Lens: Deviations and Anomalies . . . . .          | 10        |
| 12.4.2    | Practice: Anomaly Detection Algorithms . . . . .              | 10        |
| <b>13</b> | <b>Structural Transformations</b>                             | <b>11</b> |
| 13.1      | Symmetry-Preserving Transformations . . . . .                 | 11        |
| 13.1.1    | Theoretical Lens: Invariance and Gauge Freedom . . . . .      | 11        |
| 13.1.2    | Practice: Normalization and Standardization . . . . .         | 11        |
| 13.2      | Reshaping and Pivoting . . . . .                              | 11        |
| 13.2.1    | Theoretical Lens: Structural Homeomorphisms . . . . .         | 11        |
| 13.2.2    | Practice: Data Reshaping Operations . . . . .                 | 11        |
| <b>14</b> | <b>Feature Engineering and Representation Learning</b>        | <b>12</b> |
| 14.1      | Feature Construction . . . . .                                | 12        |
| 14.1.1    | Theoretical Lens: Projections and Embeddings . . . . .        | 12        |
| 14.1.2    | Practice: Feature Generation Techniques . . . . .             | 12        |
| 14.2      | Discretization and Encoding . . . . .                         | 12        |
| 14.2.1    | Theoretical Lens: Quantization and Information Loss . . . . . | 12        |
| 14.2.2    | Practice: Binning and Encoding Schemes . . . . .              | 12        |
| 14.3      | Dimensionality Reduction . . . . .                            | 13        |
| 14.3.1    | Theoretical Lens: Manifold Learning and Compression . . . . . | 13        |
| 14.3.2    | Practice: Reduction Algorithms . . . . .                      | 13        |
| 14.4      | Topological Feature Extraction . . . . .                      | 13        |
| 14.4.1    | Theoretical Lens: Persistent Topology and Shape . . . . .     | 13        |
| 14.4.2    | Practice: Topological Data Analysis (TDA) . . . . .           | 14        |
| <b>A</b>  | <b>Mathematical Appendix</b>                                  | <b>15</b> |
| A.1       | Category Theory Primer . . . . .                              | 15        |
| A.2       | Topology for Data Scientists . . . . .                        | 15        |
| A.3       | Differential Geometry Essentials . . . . .                    | 15        |
| A.4       | Sheaf Theory Basics . . . . .                                 | 15        |
| <b>B</b>  | <b>Case Studies</b>                                           | <b>15</b> |
| B.1       | Social Network Data Processing . . . . .                      | 15        |
| B.2       | Text Corpus Processing . . . . .                              | 15        |
| B.3       | Geospatial Data Integration . . . . .                         | 15        |
| B.4       | Longitudinal Survey Processing . . . . .                      | 15        |

# 1 Learning Objectives

## 2 Introduction

Data processing encompasses activities such as transformation, cleaning, restructuring, and representation of information, and it occupies a central position in quantitative social science. Many existing treatments approach data processing primarily as a technical endeavor. Emphasis is often placed on procedural competence through the application of specific algorithms, while the deeper questions concerning the justification of these operations and the philosophical commitments they embody receive comparatively limited attention.

In contrast to the topics addressed in previous chapters, the methods adopted in data processing evolve with particular rapidity in the contemporary era. This chapter therefore emphasizes generative forms of knowledge that support flexible reasoning about data processing as methodological conditions continue to change. To cultivate such intellectual adaptability, the discussion proceeds from philosophical foundations through formal theory to concrete practice, guided by the conviction that operational skill gains durability when its underlying rationale is understood.

The architecture of the chapter follows a generative progression from philosophy to theory and from theory to practice. These levels operate on distinct temporal horizons. Practical techniques evolve rapidly as computational capacity expands and methodological innovation accelerates. Theoretical frameworks typically exhibit greater stability, offering durable lenses through which data operations may be interpreted, even as those frameworks themselves undergo periods of gradual refinement punctuated by conceptual reconfiguration. Philosophical inquiry operates on an even longer timescale. Questions concerning the ontological status of data, the epistemology of transformation, and the ethical dimensions of algorithmic choice persist across technological eras and provide generative conditions from which theories emerge.

Beginning from philosophy therefore serves a constructive purpose. Readers are equipped to engage with present methods while developing the intellectual orientation required to interpret and create methods that may arise under future paradigms.

The landscape of quantitative social science has already undergone substantial transformation within a single decade. Computational social science has expanded the analytical repertoire to include large-scale network analysis, agent-based modeling, and advanced natural language processing. Causal inference has been reshaped through graphical models, potential outcomes, and renewed attention to identification strategies. Machine learning has progressed from predictive applications toward representation learning and integration with causal reasoning.

Further shifts in representational and computational substrates remain plausible. One may anticipate environments shaped by quantum data representations, neuromorphic architectures capable of processing temporally sparse signals, memory systems designed to preserve topological structure, probabilistic computing hardware that operates directly on distributions, bio-digital data generated through advanced sensing technologies, and fault-tolerant quantum computation grounded in topological principles. Developments of this kind reorganize familiar methodological questions within new material and computational constraints rather than rendering prior knowledge obsolete.

To engage productively with such transformations, readers benefit from understanding several enduring dimensions of inquiry: what it means for something to exist as data, how

knowledge is constructed and potentially lost through transformation, and how formal frameworks such as category theory, topology, geometry, and information theory illuminate structural properties shared across diverse techniques. The intellectual demands of the present era extend beyond the adoption of new tools toward the capacity to evaluate, critique, and generate them.

These relationships may be interpreted through a hierarchy of evolutionary timescales. Algorithms and implementation practices often shift within a matter of years. Theoretical structures tend to persist across decades, even as they are occasionally reorganized through conceptual breakthroughs. Philosophical orientations frequently endure for centuries, shaping the space within which theoretical innovation becomes possible.

Grounding practice in theory and theory in philosophy cultivates intellectual robustness across these timescales. Such robustness enables continuity of understanding even as the surrounding technical landscape transforms.

The chapter is organized into three parts that reflect this generative architecture. The first part examines philosophical foundations by addressing the nature of data, the epistemic consequences of transformation, the value commitments embedded in processing decisions, and the semantic conditions under which symbols acquire meaning. These discussions establish the conceptual space within which data processing becomes intelligible.

The second part develops formal languages for reasoning about data, including algebraic structure, category-theoretic composition, topological connectivity, geometric representation, statistical inference, information-theoretic constraints, and dynamical behavior. Each framework functions as a lens that reveals distinct yet complementary aspects of the same phenomena. Missing data, for example, may be interpreted through the semantics of optional types, the topology of incomplete spaces, statistical mechanisms governing absence, and the informational structure of uncertainty. The coexistence of multiple perspectives reflects the complexity of the phenomena under investigation rather than redundancy of explanation.

The final part turns toward concrete activities such as cleaning, transformation, and feature construction. Each practice is examined in explicit dialogue with the philosophical and theoretical foundations developed earlier. Algorithms appear as instantiations of principles, methodological choices as expressions of commitments, and case studies as opportunities to observe how distinct theoretical lenses converge upon shared practical problems.

Readers are invited into a demanding intellectual project. Patience with abstraction supports slower forms of reasoning that sustain long-term understanding. Openness to multiple modes of explanation allows a single operation to be interpreted through several theoretical perspectives. Active integration encourages continuous reflection on the philosophical assumptions and structural principles expressed within technical procedures.

The aim of this chapter is generative. It seeks to cultivate deep understanding, intellectual independence, creative capacity, and fluency across paradigms. When future computational or representational shifts occur, readers should be prepared to recognize their structural implications and situate them within an expanded conceptual landscape.

Given the chapter's scope, several reading paths remain viable. A linear progression from philosophy through theory to practice develops the strongest cumulative understanding. Readers motivated by a concrete problem may begin with a practical section and trace its theoretical and philosophical dependencies. Others may focus on building new formal lenses before returning to applied contexts. Regardless of the chosen path, the value of the chapter resides in the integration of its parts, since insight emerges

most fully when philosophical reflection, theoretical structure, and practical reasoning illuminate one another.

### 3 Learning Objectives

*“Data is not a thing in itself, but a representation shaped by choices—philosophical, mathematical, and practical. To process data is to navigate the tension between preservation and transformation, between what is and what we choose to see.”*

## Introduction

This chapter develops a multi-layered understanding of data processing, moving from abstract philosophical questions to concrete algorithmic practice. We begin by examining the ontological and epistemological status of data itself, then build formal theoretical frameworks from mathematics, statistics, information theory, and dynamical systems. Finally, we apply these frameworks to practical challenges in data cleaning, transformation, and feature engineering.

The structure follows a deliberate progression: *Why* (philosophical foundations) → *What* (theoretical tools) → *How* (practices). Each practical section integrates multiple theoretical perspectives, demonstrating how abstract concepts illuminate concrete algorithmic choices.

## Part I

# Philosophical Foundations

## 4 The Ontological Status of Data: Being vs. Representation

### 4.1 Data as Entity vs. Data as Relation

### 4.2 The Map-Territory Distinction in Data

### 4.3 Realism, Constructivism, and Instrumentalism

### 4.4 Digital Objects and Their Modes of Existence

## 5 Epistemology of Transformation: Knowledge Construction and Loss

### 5.1 What Can We Know Through Data?

### 5.2 Transformation as Knowledge Construction

### 5.3 The Inevitability of Loss: What Transformations Hide

### 5.4 Reversibility, Sufficiency, and Information Preservation

## 6 Ethics and Axiology: Values, Bias, and the Morality of Processing

### 6.1 Value-Laden Choices in Data Processing

### 6.2 Algorithmic Bias and Structural Injustice

### 6.3 Fairness Criteria and Their Tensions

### 6.4 Responsibility and Accountability in Automated Decisions

## 7 Logic and Semantics: Meaning, Types, and Symbolic Systems

### 7.1 Formal Logic and Data Constraints

### 7.2 Type Systems: Syntax and Semantics

### 7.3 The Meaning of Symbols and Grounding Problem

### 7.4 Compositionality and Semantic Preservation

## Part II

# Theoretical Frameworks

- Information-theoretic noise and channel capacity (§10.1)
- Systemic bias and structural errors (§6)
- Type violations and semantic errors (§7)

### **12.1.2 Practice: Detection and Validation Algorithms**

- Constraint-based validation rules
- Statistical outlier detection for errors
- Cross-field consistency checks
- Error correction codes and checksums

## **12.2 Missing Data**

### **12.2.1 Theoretical Lens: The Nature of Absence**

- Ontology of absence: nothingness vs. unknown vs. inapplicable (§4)
- Type theory: nullable types, Maybe monad, optional values (§8.1)
- Missing data mechanisms: MCAR, MAR, MNAR (§9.3)
- Topology of incomplete spaces: open vs. closed sets (§8.3)
- Information-theoretic view: entropy of missingness patterns (§10.1)

### **12.2.2 Practice: Imputation Algorithms**

- Simple imputation: mean, median, mode
- k-Nearest Neighbors imputation
- Matrix completion and low-rank approximation
- Multiple imputation and uncertainty quantification
- Deep learning approaches: autoencoders, GANs
- Case study: Handling missing laboratory values in clinical data

## **12.3 Entity Resolution and Deduplication**

### **12.3.1 Theoretical Lens: Identity Across Transformations**

- Equivalence relations and quotient structures (§8.1)
- Identity through categorical isomorphisms (§8.2)
- Probabilistic matching and likelihood ratios (§9.1)
- Metric spaces and similarity measures (§8.4)
- Information-theoretic record comparison (§10.3)

### 12.3.2 Practice: Record Linkage Algorithms

- Deterministic matching: exact and rule-based methods
- Fellegi-Sunter probabilistic model
- String similarity metrics: edit distance, Jaro-Winkler, n-grams
- Blocking and indexing strategies
- Machine learning approaches: supervised and active learning
- Clustering-based deduplication
- Case study: Patient matching across healthcare systems

## 12.4 Outlier Detection

### 12.4.1 Theoretical Lens: Deviations and Anomalies

- Statistical deviation: tail probabilities and extreme values (§9.2)
- Geometric distance from manifold (§8.4)
- Topological isolation and connectivity (§8.3)
- Information-theoretic surprise (§10.2)
- Dynamical systems: transient vs. persistent anomalies (§11.1)

### 12.4.2 Practice: Anomaly Detection Algorithms

- Univariate methods: Z-scores, modified Z-scores, IQR
- Multivariate methods: Mahalanobis distance, elliptic envelope
- Density-based: Local Outlier Factor (LOF), DBSCAN
- Isolation-based: Isolation Forest, Extended Isolation Forest
- Distance-based: k-NN distance, distance to k-th neighbor
- Model-based: Autoencoders, one-class SVM
- Robust statistics: M-estimators, RANSAC
- Time series: ARIMA residuals, change point detection
- Case study: Fraud detection in financial transactions

## 13 Structural Transformations

### 13.1 Symmetry-Preserving Transformations

#### 13.1.1 Theoretical Lens: Invariance and Gauge Freedom

- Group actions and orbits (§8.1)
- Invariants and equivariance
- Location-scale families in statistics (§9.2)
- Gauge transformations and degrees of freedom
- Information preservation under symmetry (§10.2)

#### 13.1.2 Practice: Normalization and Standardization

- Z-score standardization
- Min-max normalization
- Robust scaling (median, IQR)
- Box-Cox and Yeo-Johnson transformations
- Unit vector normalization
- Whitening transformations
- Case study: Feature scaling in machine learning pipelines

### 13.2 Reshaping and Pivoting

#### 13.2.1 Theoretical Lens: Structural Homeomorphisms

- Topological equivalence: wide vs. long formats (§8.3)
- Categorical duality: products and coproducts (§8.2)
- Information preservation in reshaping (§10.2)
- Tidy data principles and normal forms

#### 13.2.2 Practice: Data Reshaping Operations

- Pivot: wide to long conversion
- Melt and stack operations
- Unpivot: long to wide conversion
- Unstack and spread operations
- Multi-index and hierarchical data
- Aggregation and grouping
- Case study: Restructuring panel data for analysis

## 14 Feature Engineering and Representation Learning

### 14.1 Feature Construction

#### 14.1.1 Theoretical Lens: Projections and Embeddings

- Basis functions and function spaces (§8.1)
- Kernel methods and reproducing kernel Hilbert spaces (§8.4)
- Sufficient statistics and information preservation (§9.3)
- Feature maps as functors (§8.2)

#### 14.1.2 Practice: Feature Generation Techniques

- Polynomial features and interactions
- Domain-specific transformations (e.g., Fourier, wavelets)
- Ratio and difference features
- Aggregation features: rolling windows, exponential smoothing
- Automated feature engineering: featuretools, tsfresh
- Case study: Creating predictive features for time series forecasting

### 14.2 Discretization and Encoding

#### 14.2.1 Theoretical Lens: Quantization and Information Loss

- Cell complexes and partition theory (§8.3)
- Quantization and rate-distortion theory (§10.1)
- Information loss in binning (§10.2)
- Type conversion and semantic shifts (§7)

#### 14.2.2 Practice: Binning and Encoding Schemes

- Equal-width and equal-frequency binning
- Entropy-based discretization (MDL, MDLP)
- Tree-based binning (CART, chi-merge)
- One-hot encoding and dummy variables
- Target encoding and weight of evidence
- Binary encoding and hash encoding
- Learned embeddings: entity embeddings, word2vec
- Case study: Encoding high-cardinality categorical variables

## 14.3 Dimensionality Reduction

### 14.3.1 Theoretical Lens: Manifold Learning and Compression

- Manifold hypothesis (§8.4)
- Information bottleneck principle (§10.2)
- Variance preservation and eigenspaces (§9.5)
- Metric preservation and distortion (§8.4)
- Rate-distortion tradeoffs (§10.1)

### 14.3.2 Practice: Reduction Algorithms

- Linear methods:
  - Principal Component Analysis (PCA)
  - Singular Value Decomposition (SVD)
  - Independent Component Analysis (ICA)
  - Factor Analysis
  - Linear Discriminant Analysis (LDA)
- Nonlinear methods:
  - t-Distributed Stochastic Neighbor Embedding (t-SNE)
  - Uniform Manifold Approximation and Projection (UMAP)
  - Isomap and Locally Linear Embedding (LLE)
  - Autoencoders and variational autoencoders
- Sparse methods: sparse PCA, dictionary learning
- Random projections and Johnson-Lindenstrauss lemma
- Case study: Visualizing high-dimensional genomic data

## 14.4 Topological Feature Extraction

### 14.4.1 Theoretical Lens: Persistent Topology and Shape

- Persistent homology and barcodes (§8.3)
- Morse theory and critical points
- Reeb graphs and contour trees
- Shape statistics and topological signatures

### 14.4.2 Practice: Topological Data Analysis (TDA)

- Computing persistent homology: Vietoris-Rips, Čech complexes
- Persistence diagrams and bottleneck distance
- Mapper algorithm for visualization
- Vectorizing topological features: persistence images, landscapes
- Software tools: GUDHI, Ripser, giotto-tda
- Integration with machine learning pipelines
- Case study: Analyzing sensor network data with TDA

## Conclusion

This chapter has traversed multiple levels of abstraction—from philosophical questions about what data *is*, through formal mathematical and statistical frameworks, to concrete algorithms for cleaning, transforming, and extracting features from data. The recurring theme is that **every practical choice encodes theoretical commitments**.

When we impute missing values, we make ontological assumptions about absence. When we detect outliers, we invoke geometric, statistical, and information-theoretic notions of deviation. When we reduce dimensionality, we navigate tradeoffs between compression and fidelity formalized by rate-distortion theory.

The integration of theory and practice is not merely pedagogical—it is essential for principled data science. Understanding the formal foundations enables us to:

- **Reason about validity:** Which transformations preserve the information we care about?
- **Navigate tradeoffs:** What are we losing when we discretize, aggregate, or project?
- **Design better algorithms:** Can category theory suggest new data structures? Can topology reveal hidden patterns?
- **Communicate decisions:** Explaining choices in formal terms clarifies assumptions and constraints.

As data processing increasingly shapes consequential decisions—in medicine, policy, justice, and everyday life—the responsibility to understand not just *how* algorithms work, but *why* they are justified, becomes inescapable. This chapter provides the conceptual tools for that understanding.

## Further Reading

### Philosophical Foundations:

- Floridi, L. (2010). *Information: A Very Short Introduction*. Oxford University Press.

- Leonelli, S. (2016). *Data-Centric Biology: A Philosophical Study*. University of Chicago Press.

### **Mathematical and Statistical Foundations:**

- Mac Lane, S. (1998). *Categories for the Working Mathematician*. Springer.
- Casella, G., & Berger, R. L. (2002). *Statistical Inference* (2nd ed.). Duxbury.
- Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory* (2nd ed.). Wiley.

### **Topological Data Analysis:**

- Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society*, 46(2), 255-308.
- Oudot, S. Y. (2015). *Persistence Theory: From Quiver Representations to Data Analysis*. American Mathematical Society.

### **Practical Methods:**

- Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, 59(10), 1-23.
- Little, R. J., & Rubin, D. B. (2019). *Statistical Analysis with Missing Data* (3rd ed.). Wiley.
- Zheng, A., & Casari, A. (2018). *Feature Engineering for Machine Learning*. O'Reilly Media.

## **A Mathematical Appendix**

### **A.1 Category Theory Primer**

### **A.2 Topology for Data Scientists**

### **A.3 Differential Geometry Essentials**

### **A.4 Sheaf Theory Basics**

## **B Case Studies**

### **B.1 Social Network Data Processing**

### **B.2 Text Corpus Processing**

### **B.3 Geospatial Data Integration**

### **B.4 Longitudinal Survey Processing**