Table of contents

AI and Data Literacy

Student Edition

By Kimberly Seefeld, EdD

Published under **DataJoyAI** imprint

Companion resources for this book are available at:

DataJoyAI

https://datajoyai.com

Edition: First Edition
ISBN: 978-1-969233-33-3

Chapter 1

Data, AI, and the Modern World

1.1 Learning Outcomes

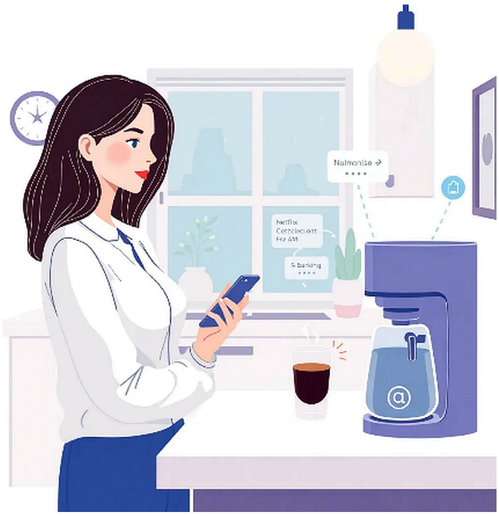
1-1: Define data, information, and artificial intelligence using precise, practical terminology that distinguishes between these related but distinct concepts.

1-2: Distinguish artificial intelligence from automation by identifying key characteristics, capabilities, and limitations of each approach.

1-3: Identify and analyze everyday AI applications across multiple domains, explaining their input-output relationships and practical impact.

1-4: Explain the importance of AI and data literacy for personal decision-making, professional development, and civic participation in the digital age.

1.2 Introduction



Sarah Martinez starts her day like millions of others around the world. Her smartphone alarm wakes her at precisely 6:30 AM, having learned from weeks of her sleep patterns that this timing optimizes her rest cycles. As she reaches for her phone, Netflix sends her a notification about a new documentary series that aligns perfectly with her recent viewing history of environmental content. Her banking app alerts her to an unusual spending pattern detected overnight that a fraudulent charge was automatically flagged and blocked before she even knew about it.

By the time Sarah has her first cup of coffee, she has already interacted with artificial intelligence systems at least five times, yet she might not realize it. This scenario, replicated across billions of lives daily, illustrates a fundamental truth about our modern world: we are living in an age where data and artificial intelligence have become as essential to daily life as electricity or running water.

The transformation has been remarkably swift. Just two years ago, most people had never heard of ChatGPT. Today, according to recent Pew Research data, 34% of American adults have used the AI chatbot—a figure that has doubled since 2023 [1]. Among adults under 30, that number jumps to 58%. According to Forbes analysis of global AI trends, 78% of organizations now use AI in some capacity, up from 55% just one year earlier [2].

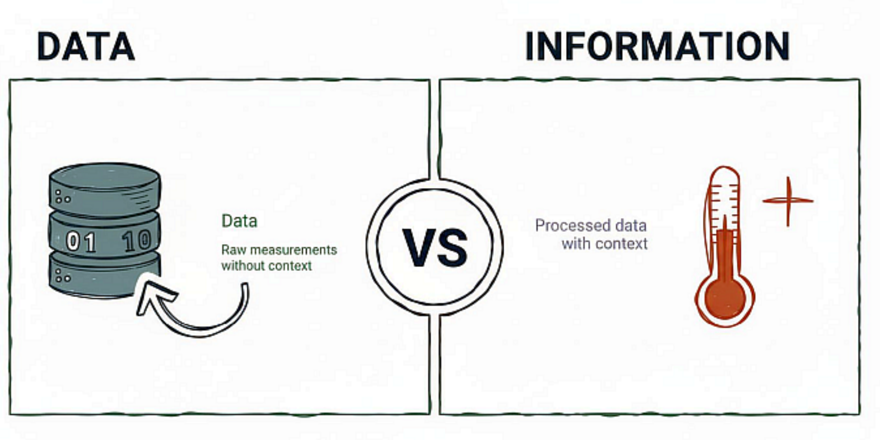
This rapid adoption creates both challenges and opportunities. As AI systems become more sophisticated and pervasive, the ability to understand, evaluate, and effectively use these tools becomes essential for personal and professional success.

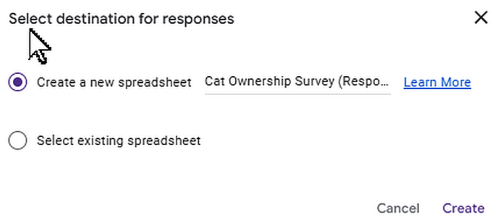
This chapter serves as your foundation for understanding this new reality. We will explore what data and artificial intelligence actually are, moving beyond the hype and headlines to develop practical, working knowledge of these technologies. More importantly, we will examine why developing literacy in these areas has become as crucial as traditional literacy was in previous centuries. The goal is not to make you a technical expert, but to provide you with the conceptual framework and practical understanding necessary to navigate, evaluate, and leverage these tools effectively in whatever field you choose to pursue.

1.3 Understanding Data in the Digital Age

1.3.1 The Foundation: What Data Actually Means

Before we can understand artificial intelligence, we must first establish a clear understanding of data itself. The term "data" has become so commonplace that we often use it without considering what it actually represents. In its most fundamental sense, data consists of raw facts, observations, or measurements that have been recorded but not yet processed or interpreted. Think of data as the building blocks of knowledge—individual pieces that, when properly assembled and analyzed, can reveal patterns, insights, and understanding.





2.8.3.3 Step 3: Analyzing Sample Data

Review how survey responses appear in spreadsheet format.

Timestamp	Q1	Q2	Q3	Q4	Q5
1/19/2026 17:23:24	Yes	12.8	Outdoor	2	3
1/19/2026 17:23:33	No	—	—	—	8
1/19/2026 17:23:46	Yes	2	Indoor	1	1
1/19/2026 17:24:02	Yes	6	Indoor	2	6

2.8.3.4 Data Quality Issues

Survey data often contain inconsistencies and ambiguities. This dataset illustrates several common challenges:

1. **Inconsistent or invalid entries:** For example, the age value "12,8" may represent multiple interpretations.

2. **Missing responses:** Non-cat owners left cat-related fields blank, requiring careful handling during analysis.

3. **Scale interpretation differences:** Respondents may interpret rating scales differently.

4. **Logical ambiguity:** Respondents with multiple cats may not know which cat's age to report.

5. **Data type issues:** Numeric fields may contain text, preventing accurate calculations.

3.3 Understanding Data Storage Formats

3.3.1 The Evolution of Data Storage: From Files to Formats

The history of data storage reflects the ongoing tension between human readability and machine efficiency. In the early days of computing, data was stored in formats that humans could easily read and modify—simple text files with values separated by commas or tabs. As datasets grew larger and analytical requirements became more sophisticated, the limitations of these human-readable formats became apparent, driving the development of more efficient but less accessible binary formats.

Let's first look at the human-readable formats, then the binary formats, and then the end discusses how modern systems attempt to bridge this gap. This progression reflects not just technological advancement, but also changing organizational needs around data collaboration, performance, and scale.

Human-readable formats like CSV and JSON prioritize accessibility and interoperability over performance. These formats can be opened in any text editor, easily inspected and modified by humans, and processed by virtually any programming language or tool. This accessibility makes them ideal for data exchange between different systems, debugging data quality issues, and situations where non-technical stakeholders need to examine or modify data directly.

However, human readability comes with significant costs. Text-based formats

3.8 Lab 3A Data Profiling

3.8.1 Introduction

Data profiling is like being a detective for your data. Before you can analyze data or make decisions with it, you need to understand what you're working with. Is the data complete? Are there errors? What patterns exist? This lab will teach you how to use R packages to automatically examine datasets and identify potential quality issues.

Think of data profiling like inspecting a used car before buying it - you want to know about any problems before you commit to using it for important decisions.

Data profiling plays a critical role in real-world applications. Marketing teams analyze customer survey data to guide campaign strategies, while financial analysts examine transaction records to identify errors before reporting. In healthcare, researchers validate patient information to ensure the accuracy of their studies, and retail companies review sales data to support precise inventory decisions. With poor data quality costing U.S. businesses more than $3 trillion each year, mastering data profiling has become an essential skill.

Data profiling is the process of examining data to understand its structure, quality, and content. It identifies what data exists, the types of variables, patterns and distributions, missing values, and any unusual or suspicious entries. Profiling helps uncover hidden issues that could affect decisions. For example, a retail company discovered that 30% of customer email addresses were invalid, preventing wasted marketing spend. By detecting errors early, data profiling saves time, improves confidence in decision-making, and ensures data quality and compliance.

3.8.2 Demonstration

Let's run through a simple data profile using R package `skimr` and dataset `food_coded.csv`. The code file is `lab3a_demo.R`.

- Data validity: Are there any impossible or suspicious values?
- Data consistency: Do the distributions look reasonable for each variable type?
- Recommendations: What data cleaning steps would you recommend before analysis?

This exercise will help you develop the critical eye needed to assess data quality in real-world scenarios and make informed decisions about data preparation and analysis approaches.

3.9 Lab 3B Introduction to APIs

3.9.1 Introduction

An **API** (Application Programming Interface) is like ordering food at a restaurant:

1. You (R) ask for something from the menu (website)
2. The waiter (API) takes your order to the kitchen (server)
3. The waiter brings you back your food (data)

APIs let you get fresh data from websites automatically instead of copying and pasting.

3.9.1.1 Why Do Data Analysts Use APIs?

In the real world, companies store their data on servers, which are specialized computers designed to hold and manage large amounts of information. Because this data is private and often sensitive, you cannot simply walk into a company and request access to its database. Instead, organizations use APIs (Application Programming Interfaces) as the official and secure way to share data with authorized users. Rather than downloading a static file one time, APIs allow you to request data whenever you need it, ensuring that you are always working with the most up-to-date information.

Businesses use APIs for many common data tasks. Sales APIs allow analysts to retrieve recent sales figures directly from company systems. Customer data can be accessed through APIs connected to CRM (Customer Relationship Manage-

Chapter 4

How AI Learns from Data

4.1 Learning Outcomes

4-1: Distinguish between supervised and unsupervised learning approaches and identify appropriate use cases for each

4-2: Explain the concepts of training data, features, and labels in machine learning contexts

4-3: Describe the training, validation, and testing process used to develop AI models

4-4: Recognize common challenges in AI learning, including overfitting and data quality issues

4.2 Introduction

Sarah Martinez settles into her couch after a long day, opens Netflix, and immediately sees a personalized homepage filled with recommendations: "Because you watched *The Crown*," "Trending in Documentaries," and "New Releases for You." She clicks on a documentary about climate change that Netflix suggested, and it turns out to be exactly what she was in the mood for. How did Netflix know?

This scenario, playing out millions of times daily across the globe, represents one of the most successful applications of machine learning in our everyday lives. Netflix's recommendation system processes viewing data from over 230 million subscribers worldwide, analyzing not just what people watch, but when they pause, rewind, or abandon shows entirely [1]. The system learns from this massive dataset to predict what each individual user might enjoy next.

But here's the fascinating part: Netflix's AI doesn't just memorize what Sarah has watched before. Instead, it has learned patterns from the collective behavior of millions of users, identifying subtle connections between viewing preferences, demographic factors, and content characteristics. This is machine learning in action—the process by which artificial intelligence systems learn to make predictions and decisions by finding patterns in data.

The story of how AI learns from data is fundamentally about pattern recognition at scale. Whether we're talking about Netflix recommendations, medical diagnosis systems, or autonomous vehicles, the underlying process remains remarkably similar: feed the system examples, let it identify patterns, and then use those patterns to make predictions about new, unseen situations.

So far in this course, we have examined what data is, how it flows through organizations, and where it gets stored. Now we turn to perhaps the most exciting question: How does artificial intelligence actually learn from all this data? The answer lies in understanding machine learning—the field that has transformed AI from science fiction into the practical tools we use every day.

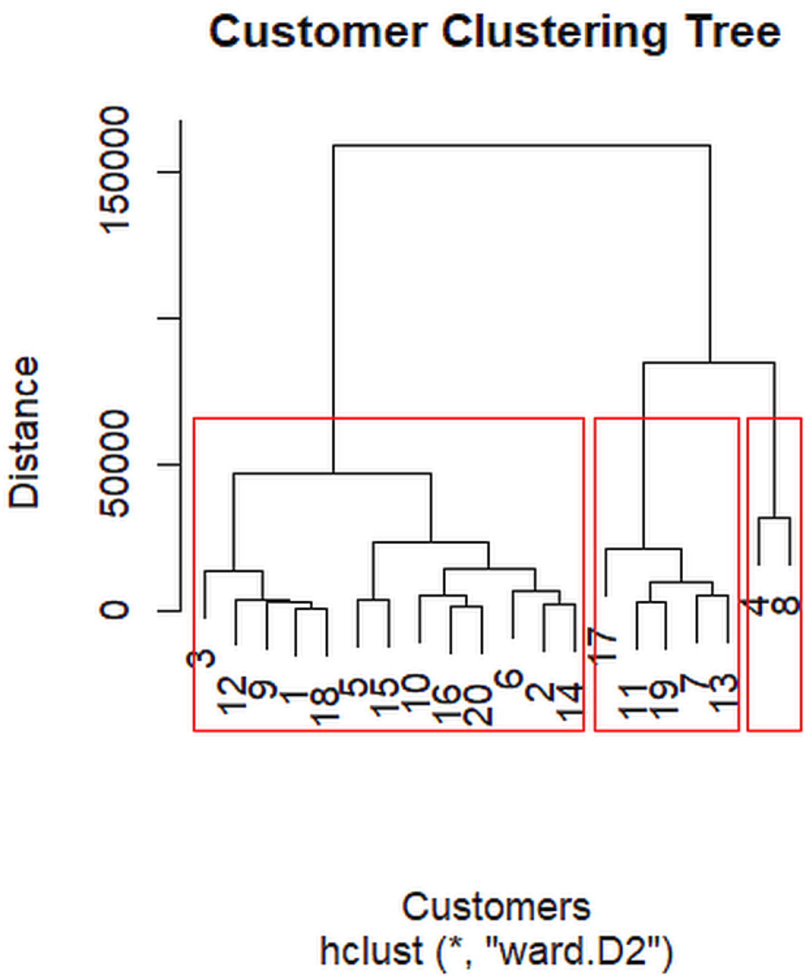
4.3 The Learning Spectrum: Supervised vs. Unsupervised

4.3.1 Conceptual Distinction

The trick with machine learning is to understand that not all learning happens the same way. Just as humans learn differently depending on whether they have a teacher providing answers or must discover patterns on their own, AI systems use different approaches based on the type of guidance they receive during training.

4.3.2 Supervised Learning: Learning with a Teacher

Supervised learning represents the most straightforward approach to machine learning, where we provide the AI system with both questions and answers during training. Think of it like studying for an exam with a comprehensive answer key—the system learns by examining thousands or millions of examples where the correct answer is already known.




```r
# Perform hierarchical clustering
hierarchical_result <- hclust(customer_distances, method = "ward.D2")

# Draw the dendrogram
plot(hierarchical_result, main = "Customer Clustering Tree",
     xlab = "Customers", ylab = "Distance")

# Highlight 3 main groups on the tree
rect.hclust(hierarchical_result, k = 3, border = "red")

# Get cluster assignments for 3 groups
```

the direction and rate of change.

5.4.2.1 The Power of Temporal Visualization

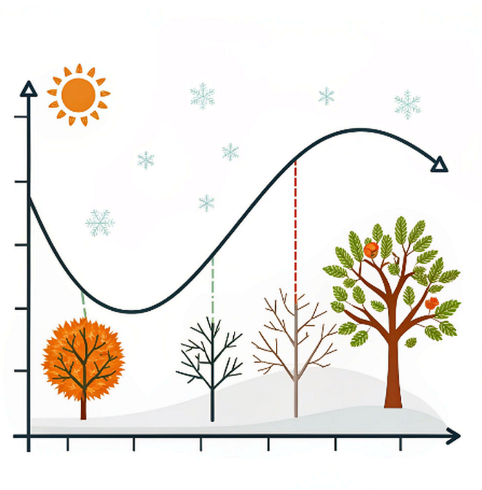
Google Analytics processes over 3 billion website interactions daily, and their primary interface relies heavily on line charts to help website owners understand traffic patterns, user behavior trends, and the impact of marketing campaigns over time [11].

Line charts prove particularly powerful when comparing multiple trends simultaneously. A company might overlay revenue, expenses, and profit margins on a single chart to understand the relationships between these metrics and identify periods where efficiency improved or declined.

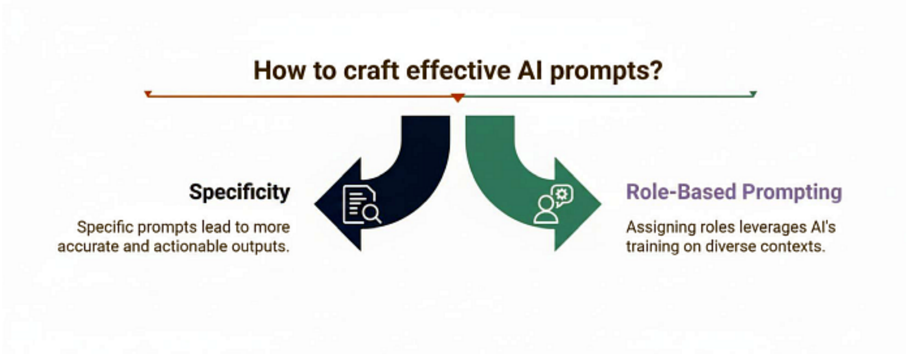
5.4.2.2 Multiple Line Series: Comparing Trends

When displaying multiple lines on a single chart, careful design becomes crucial. Each line should be clearly distinguishable through color, line style, or markers. However, research suggests limiting multiple line charts to 5-7 series maximum to maintain readability [12].

5.4.2.3 Seasonal Patterns and Cyclical Trends



Line charts excel at revealing seasonal patterns that might not be apparent in summary statistics. E-commerce companies use line charts to identify seasonal



6.3.2.1 Methodological Focus: Specificity Over Generality

The most common mistake in AI prompting involves being too vague or general. Consider the difference between these two prompts for a healthcare application:

Vague prompt: "Tell me about patient care."

Specific prompt: "You are a healthcare quality improvement specialist. Analyze the attached patient satisfaction survey data from our 200-bed hospital. Identify the top 3 factors contributing to low satisfaction scores in the emergency department and provide evidence-based recommendations for improvement, including estimated implementation costs and expected timeline for results."

Research from 2024 shows that specific prompts with clear context and constraints produce outputs that are 73% more accurate and 85% more actionable than generic prompts [2]. The specific prompt provides role context, defines the task clearly, specifies the data source, sets expectations for the analysis depth, and requests actionable outcomes.

6.3.2.2 Role-Based Prompting: The Power of Persona

One of the most effective prompt engineering techniques involves assigning specific roles or personas to the AI system. This approach leverages the AI's training on diverse professional contexts and helps focus its responses on relevant expertise and communication styles.

Examples of Effective Role-Based Prompts:

- **Financial Analysis:** "You are a senior financial analyst at a Fortune 500 company with 15 years of experience in risk assessment..."

2. What are the benefits and potential pitfalls of using role-based prompting in professional AI applications?
3. How can chain-of-thought and few-shot prompting techniques improve AI performance in complex tasks?
4. What strategies can you use to detect hallucinations in AI outputs, and why is each important?
5. How do different types of AI bias (demographic, cultural, temporal, confirmation) impact decision-making in professional contexts?
6. Why is it important to scale verification efforts according to the stakes of an AI application?
7. How can humans and AI systems complement each other in professional workflows, and what factors determine the best collaboration model?
8. In what ways can iterative prompt refinement and systematic evaluation improve AI adoption in organizations?
9. How might emerging techniques like multimodal or adaptive prompting change the way humans interact with AI in the next five years?
10. Considering the example of Maria Rodriguez, what lessons can you draw about the relationship between prompt engineering skill and the practical value derived from AI tools?

6.10 Lab 6A Understanding Prompt Engineering

6.10.1 Introduction

Prompt engineering is the practice of crafting clear, specific instructions to get better results from AI systems like ChatGPT, Claude, or other language models. Think of it as learning how to communicate effectively with AI - just like you might give different instructions to a child versus a professional colleague, you need to adjust how you "talk" to AI systems to get the results you want.

Unlike traditional computer programs that follow exact commands, AI systems interpret natural language and can understand context, but they need guidance to produce useful outputs. The quality of your prompt directly determines the

Chapter 7

Ethics, Fairness, and Responsible AI

7.1 Learning Outcomes

7-1: Recognize bias and fairness issues in AI systems and datasets, and understand their real-world implications

7-2: Understand privacy, consent, and surveillance concerns in AI applications

7-3: Identify AI misuse including deepfakes, misinformation, and manipulation tactics

7-4: Apply ethical frameworks for responsible AI development and deployment