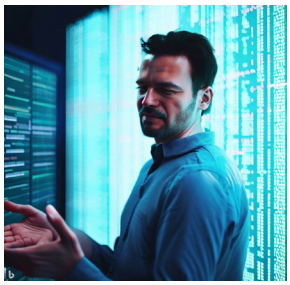

Chapter 1

Introduction to Data Science

"I have been impressed with the urgency of doing. Knowing is not enough; we must apply. Being willing is not enough; we must do"
—Leonardo da Vinci

LEARNING OBJECTIVES

By the end of chapter one, readers should be able to:

- Discuss why data science is an essential topic in all domains
- Discuss the multidisciplinary nature of data science
- Enumerate the typical skill set of a data scientist
- Compare the three main open-source programs, Orange, JASP, and BlueSky Statistics
- Discuss the differences between data science, data management, statistics, and business intelligence
- Discuss the various practical applications for data science in various industries

Chapter 1 Summary

This chapter delves into the exploration of data science, defining its key components, tracing its historical evolution, and highlighting its vital importance in our data-centric world. It also covers the rationale behind this textbook's creation, aimed at providing an enriching learning experience for both newcomers and established professionals in the field. We describe the differences between data science, business intelligence, and statistics as well as the important integration between data science and data management. In addition, we shed light on our decision to use open-source software as our primary educational platform due to its transparency, affordability, adaptability, and its ability to more rapidly train and engage more no-code data scientists on data science projects. We strive to offer professionals more career options at a time when many organizations of all sizes are struggling to find the talent they need to integrate predictive analytics with their business strategies. This chapter lays the foundation for future chapters where open-source tools will be applied to a variety of data types including tabular data, images, and text.

Key Principles 

- Data science is an interdisciplinary field that requires collaboration.
- There is no universally accepted definition of data science, nor is there a data science parent organization.
- No-code data science promotes the democratization of the field with no shortcuts or compromises in analysis capabilities.
- Data science requires a step-wise process that is robust, reliable and repeatable.

1.1 WHAT IS DATA SCIENCE?

Data science is pervasive in all industries. The digital revolution has generated an avalanche of data that requires expert analysis.[1] While the increasing emphasis is on machine learning and artificial intelligence, we should not lose sight of the importance of data preparation, exploration, and visualization. Approximately two-thirds of the time a data scientist spends on a project is related to

getting the data ready for modeling.[2] Therefore, there is a need for both basic and advanced data science skills.

KEY POINT: According to Wikipedia, data science can be defined as "an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from noisy, structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains." [3] Data science is a convenient umbrella terminology encompassing computational and statistical expertise. It is worth noting that there is no widely accepted definition of data science, nor is there a parent-professional organization.

This book will explore the multi-faceted realm of data science, with a focus on four pivotal analytical domains: Machine Learning for Tabular Data, Predictive Image Analytics, Predictive Text Mining, and Artificial Intelligence (AI). Table 1.1 provides a snapshot of a few select industries, showcasing how various data science application examples are deployed for those domains. [4-7]

Data science is a relatively new field with significant popularity and hype. For example, in 2012, the Harvard Business Review termed data science "the sexiest job of the 21st century." [8] This aphorism is overly glamorous, as data scientists spend the majority of time finding, cleaning, preparing, and exploring data, which can be arduous.

KEY POINT: No-code data science means a programming language is not required to perform a function, such as creating a box plot. Instead, the function uses a graphical user interface (GUI) to create the image. No-code data science leads to more people being involved in data science, or the democratization of the field. Software programs such as Orange and JASP are examples of no-code data science.

1.1.1 Background

In 1962, John Tukey, a renowned statistician, made a pronouncement that reverberated in the statistical world. He stated, "*I have come to feel that my central interest is in data analysis, which I take to include among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate*…." He clearly saw the analytical world larger than just statistics and could sense that computing would change data analysis forever. [9]

William Cleveland, a computer scientist, and statistician, was the first to publish the term "data science" in 2001. He also envisioned the field as more expansive than just statistics and dealt with much more than just theory. [10] Early data scientists worked for some of the most innovative Internet companies, such as Google, Facebook, LinkedIn, and Twitter, to assist them in gaining insights into the avalanche of new data. The term "data scientist" is attributed to DJ Patil (LinkedIn) and J Hammerbacher (Facebook) in 2008.[10] Data scientists generated new data use cases for the consumer and new business models and products. At this early stage, professionals in the field labeled themselves as data scientists, even without formal degrees. The advent of the Internet was perhaps the most significant source of the data explosion. For example, in 2023, there are, on average, approximately 8.5 billion Google searches per day. [11]

Almost every Internet activity today can be measured (*datafied or quantified*) and mined. In addition to the meteoric increase in data volume, there is also tremendous variety in the data, such as location data (geographic information system), survey data, image data, email data, tweet data, and sensor data. The data science field has been facilitated by faster computer processor speed (the addition of GPUs and TPUs), open-source software designed to process large volumes of data, and more expansive storage. In the biomedical domain, the electronic health record and genomic databases comprise the two largest data sources waiting to be mined. Data science has also benefited from the "open data" era, where industry and government have tried to make data available for the public, developers, and researchers.[12]

1.1.2 Artificial Intelligence and Machine Learning

Artificial intelligence (AI) has existed since the 1950s, but it has not been a significant aspect of data science until the last two decades. The term "*artificial intelligence"* is attributed to John McCarthy, an early computer scientist who introduced the term in 1956 at a Dartmouth conference.[13] Further historical details about AI can be found with these references.[14-15] Figure 1.1 outlines the history of modern AI and its contributing factors over three decades.[16]

The following are important definitions:

Artificial intelligence (AI)*: "refers to systems that display intelligent behaviour by analysing their environment and taking actions - with some degree of autonomy - to achieve specific goals."* [17]

Table 1.1 Various industry applications of data science

Industry	Machine Learning for Tabular Data	Predictive Image Analytics	Predictive Text Mining	Artificial Intelligence (AI)
Healthcare	**1**. Predicting patient readmission. **2.** Diagnosing diseases from electronic health records. **3.** Forecasting medicine demands by season	**1.** Tumor detection in X-rays. **2**. Analyzing skin lesions for disease identification. **3**. Brain scans for specific neurological disorders	**1.** Analyzing patient feedback and sentiment. **2.** Identifying trends and clusters from clinical notes. **3.** Predicting disease outbreaks from news articles	**1.** AI-driven diagnostic assistance. **2.** Robotic surgeries. **3.** Personalized treatment planning and suggestions
Manufacturing	**1.** Predicting equipment failures. **2.** Quality control analytics. **3.** Optimizing production process efficiency	**1.** Detecting manufacturing defects via cameras. **2.** Analyzing product images for quality control. **3.** Robot vision for part placement	**1.** Analyzing worker feedback. **2.** Mining equipment manuals for best practices. **3.** Predictive maintenance based on textual logs	**1.** Robots for assembly lines. **2.** AI-driven process optimization. **3.** Autonomous material handling systems
Service	**1.** Predicting customer service ticket volume. **2.** Service efficiency optimization based on historical data. **3.** Customer retention modeling	**1**. Facial emotion detection during service interactions. **2.** Analyzing service space layouts via images. **3.** Image analysis of service equipment for maintenance needs	**1.** Mining customer feedback for service improvements. **2.** Analyzing service-related complaints for root cause. **3.** Textual trend analysis from customer interactions	**1.** AI-driven chatbots for 24/7 customer support. **2.** Virtual assistants for service scheduling and reminders. **3.** Predictive algorithms for optimizing service resource allocation
Agriculture	**1.** Soil quality predictions based on historical data. **2.** Crop yield forecasting. **3.** Predictive models for pest outbreaks	**1.** Drone imagery for crop health analysis. **2.** Disease identification from plant images. **3.** Water stress identification in fields	**1.** Analyzing agricultural research papers for trends. **2.** Mining farmer feedback for crop improvement. **3.** Text analysis for market demand predictions	**1.** AI for precision irrigation systems. **2**. Robots for harvesting. **3.** AI-guided drones for crop monitoring

Machine Learning (ML): *"is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of computer programs that can change when exposed to new data."* [18]

Deep Learning (DL): *"discovers intricate structure in large datasets by using the backpropagation algorithm to indicate how a machine should change its internal parameters that are used to compute the representation in each layer from the representation in the previous layer."* [19] *Additionally, it is a type of machine learning that uses algorithms (neural networks) with additional hidden layers to handle very large and complex datasets.*

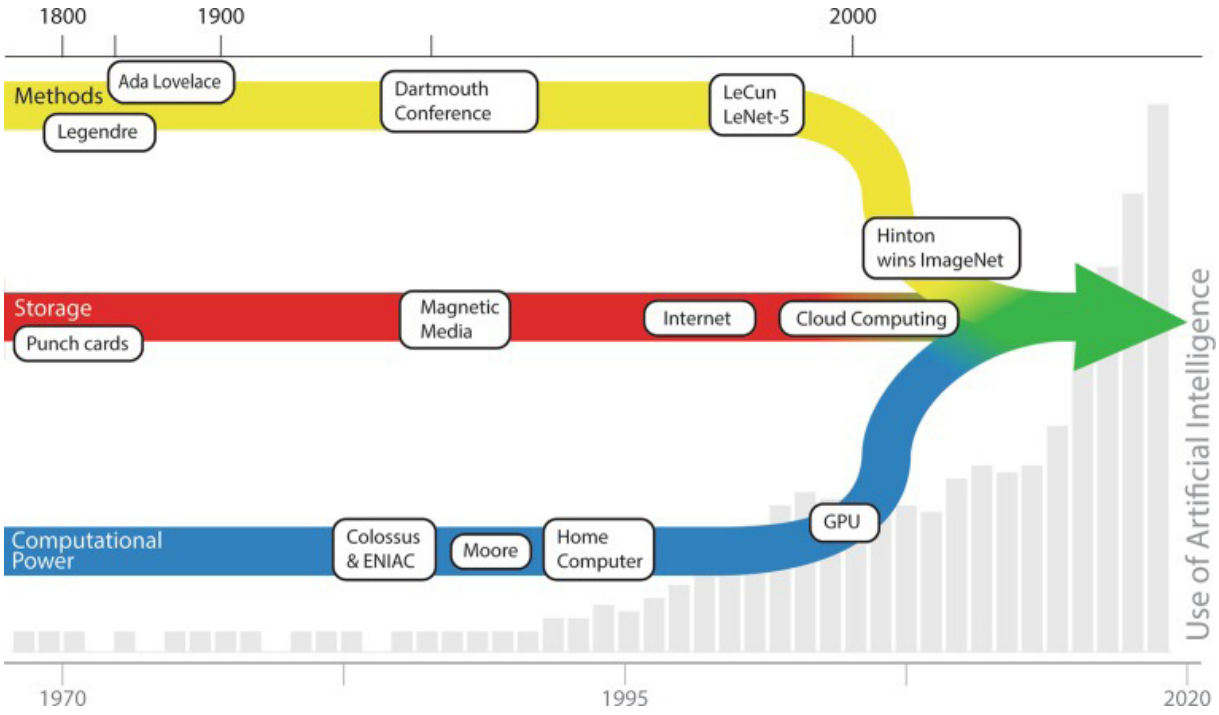

Figure 1.1 AI Temporal Progression

Artificial intelligence began as "expert systems," meaning rules created by experts were computerized for clinical decision support in the medical field. An example of this was the early MYCIN project that made recommendations regarding infectious diseases and choice of antibiotics.[20] The project ultimately failed for multiple reasons, but it should be noted that this initiative preceded personal computers, laptops, apps, and the Internet.

We have devoted chapters on image analytics (computer vision) and text mining as significant components of AI. This field can be characterized as narrow and general AI. Narrow AI functions as a narrow task, for example, language translation with Google Translate or evaluating a retinal image. General (strong) AI implies the technology thinks independently, like a human. We have not reached general AI at this point.

Machine learning is a subset of AI and pertains primarily to supervised and unsupervised learning using older statistical algorithms, such as linear regression, or newer ones, such as random forest. Machine learning focuses mainly on predictive analytics. This is discussed in much more detail in Chapter 4 on modeling.

Figure 1.2 shows the relationship between AI, machine, and deep learning. Many authorities would show this Venn diagram as shared by both the data science and computer science fields.

In this textbook, we will discuss machine learning and predictive analytics. Machine learning (ML) is commonly organized into three types:

- **Supervised learning**. The ML model is trained on labeled data (the outcome is known) and aims to predict an outcome based on predictive variables. This is a classification model if the outcome is categorical (e.g., cancer, benign, email spam, credit card fraud). If the outcome is numerical (e.g., length of stay in a hospital - in days, percent of customer complaints, percent of defective products manufactured), this is a regression model.

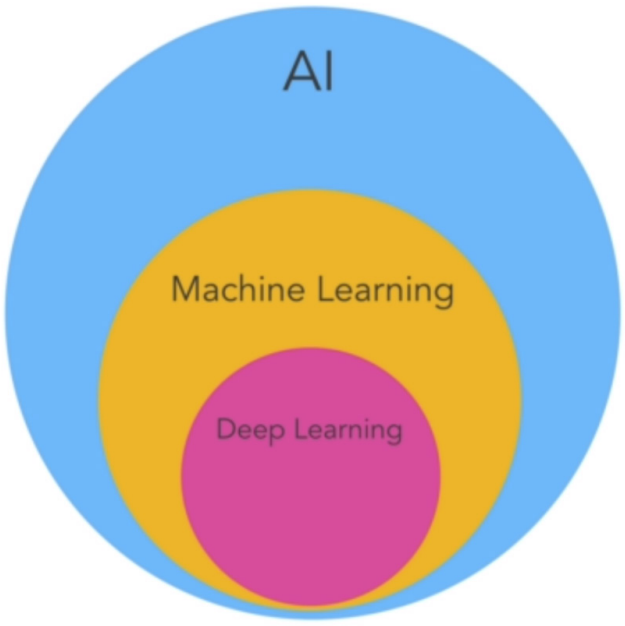

Figure 1.2 Venn diagram of AI, ML, and Deep learning

- **Unsupervised learning**. The ML model is trained on unlabeled data to identify hidden patterns, groups, or clusters in the dataset. Examples include medical image analysis, customer segmentation / clustering, recommendation systems, and identifying different groups of students with similar learning styles.
- **Reinforcement learning**. This method comes from the gaming industry, where an agent interacts with the environment with associated rewards and penalties to reach a goal. It learns through trial and error. Think about the game Pac-Man. Other examples are personalized medicine, personalized marketing, training robots to interact with their environment, and financial algorithmic trading.

1.1.3 Artificial Intelligence and Data Science

Up until recently there was only one way to practice data science and that was using a programming language, such as R or Python. This textbook is an example of how no-code data science can cover every major data wrangling, data prep, and machine learning task. No code and low-code options also exist to create AI applications.[21] AutoML low-code options are also available to perform advanced automated machine learning and visualizations tasks with minimal lines of code. Auto-Sklearn is an open-source AutoML Python library that automates the process of algorithm selection, hyperparameter tuning, and model selection.[22] H2O AutoML is an open-source AutoML platform that offers a no-code web interface, making it easier to create machine learning models. It automates algorithm selection, feature generation, hyperparameter tuning, iterative modeling, and model assessment. It provides a user-friendly interface for training and evaluating machine learning models.[23] Several other AutoML packages exist that require coding. Furthermore, AI has been integrated with Python packages and Integrated Development Environments (IDEs), as additional examples of low-code data science. [24-25]

ChaptGPT is a large language model (LLM) released by OpenAI in November 2022.[26] It was trained on massive general content, such as Wikipedia, books, Internet text resources and social media It is an example of "generative ai" which produces text, imagery, audio and synthetic data More details on LLMs and Chatbots is included in Chapter 11. Data analysis can be conducted with LLMs, and early hype suggested that it could replace data scientists. While it is likely that AI generated data science will improve in the future, there is no evidence it will replace data scientists.

Many LLMs allow the user to upload a dataset and then the user enters a prompt in natural language to request a data science task. It can also create a variety of predictive models including performance measures. The output often includes the Python code to show the workflow, so this is another example of low-code data science. Most LLMs will not produce data visualizations, so a user must cut and paste the Python code into an IDE, such as a Jupyter Notebook to see the images.[27]

A list of possible LLM use-cases is as follows:

- Function as a coding assistant to save time and keystrokes
- Review and summarize literature
- Evaluate genomic data
- Describe and summarize a dataset, discussing the columns and rows
- Use a prompt to determine if there are data challenges such as class imbalance, multicollinearity, outliers, and missing data
- Use a prompt to solve data challenges and report the rationale
- Use a prompt to discuss strengths and weaknesses of the data

A list of limitations and precautions is as follows:

- Plausible but incorrect results may occur, also known as "hallucinations."
- Results that may disagree with those run on another platform. The LLMs will not automatically perform pre-processing. That requires a human prompt.
- Automation bias where someone new to data science will not critically analyze the results
- Data bias due to the presence of systematic errors or prejudices in a dataset that can lead to inaccurate or unfair outcomes
- Lack of reasoning, intuition and common sense
- Lack of creativity
- Lack of oversight. A human must be in the loop.

1.1.4 Foundations and Frontiers: Integrating DMBoK Practices with Data Science

The Data Management Body of Knowledge (DMBoK) offers a comprehensive framework for effective data management in organizations, encompassing 11 core areas, which are Data Governance, Data Architecture, Data Modeling and Design, Data Storage and Operations, Data Security and Privacy, Data Integration and Interoperability, Document and Content Management, Reference and Master Data Management, Metadata Management, Data Quality Management, and Big Data and Analytics. Data science uses statistical and computational methods to derive insights. Integrating DMBoK with the data science process ensures data quality, context understanding, and compliance with privacy standards.

This ensures accurate and reliable analyses rooted in solid data management practices.[28]

The Data Management Body of Knowledge (DMBoK) and data science, while related, serve different primary functions, and their overlap is mainly in the domain of ensuring that data is ready and optimized for analysis.

Purpose:

- DMBoK: Focuses on providing guidelines for the management of data throughout its lifecycle, ensuring data quality, privacy, security, and availability. It's about laying the foundation and creating an environment where data can be efficiently used.[29]
- Data Science: Concentrates on analyzing data to derive insights. It utilizes various statistical, machine learning, and computational techniques to study data and make predictions or decisions.[30]

Overlap Areas:

- Data Quality: Before any data science activity, it's imperative that the data is clean, accurate, and relevant. DMBoK's guidelines can be instrumental in ensuring this.
- Data Governance and Privacy: DMBoK's focus on governance and privacy is crucial for data scientists, especially when dealing with personal or sensitive information.
- Data Architecture & Modeling: The structure and design of databases impact how easily data scientists can pull and manipulate data.
- Metadata Management: Understanding metadata can be key for data scientists to comprehend the context of the data they are working with. DMBoK's guidelines can be instrumental in ensuring this.
- Data Integration: Before analysis, data often needs to be integrated from various sources, an area DMBoK covers.

Distinct Areas:

- Techniques like machine learning, deep learning, predictive modeling, etc., which are core to data science, don't fall under DMBoK's purview.
- Conversely, areas like Document and Content Management, or Data Storage and Operations, which are in DMBoK, might not be directly relevant to a data scientist's everyday tasks but are essential for data infrastructure.

In summary, DMBoK sets the stage for data science activities. It ensures that data is well-curated, structured, integrated, and governed. While they have overlapping concerns about data quality and infrastructure, DMBoK is broader in terms of data management, whereas data science delves deeper into analysis and insight generation.

1.1.5 Statistical Foundations and Predictive Frontiers: A Data Voyage

Let's use the analogy of a car's large windshield and small rear-view mirror to compare statistics vs. Business intelligence (BI) vs. data science / predictive analytics:

Statistics:

- Rear View Mirror: Statistics can be seen as the "rear view mirror" of this analogy. It looks at data from the past and uses mathematical techniques to summarize, analyze, and interpret that data. Much like how a rear-view mirror provides information on where you've been, statistics provide a comprehensive look at what has happened.[31]

Business Intelligence (BI):

- Rear View Mirror, but a bit Wider: BI expands on the basic insights given by statistics. While still mostly looking at the past, it gives a broader view by combining data from different sources, often in real-time, to provide actionable insights. Imagine the rear-view mirror showing not just the road behind, but also a bit of the landscape and conditions. It helps companies understand their performance, sales, customer behavior, and more.[31]

Data Science / Predictive Analytics:

- Windshield Looking Forward: Data science and predictive analytics are like the large windshield of a car which provides a glimpse of what has occurred, with greater insights into what is likely to occur in the future. They try to foresee future events by leveraging historical data, algorithms, machine learning models, and other tools. Like a driver looking through the windshield and anticipating the curves, obstacles, and conditions ahead, data science and predictive analytics help businesses and researchers anticipate future events, opportunities and challenges.[32]

In summary, while statistics and BI largely focus on understanding and interpreting past data, data science, especially predictive analytics, looks ahead and tries to predict the future. All three are crucial; just as both the rear-view mirror and windshield are essential for a driver to navigate the road safely.

1.1.6 The 8-Step Data Science DISCOVER Process

KEY POINT: Many different acronyms and step charts have been created for the Data Science process. We have created our own process, as shown in Figure 1.3, which we call the DISCOVER process. We use DISCOVER as a word and an acronym for the eight steps in the no-code data science process, as defined in Figure

1.3. Any great methodology, such as Data Science, needs a stepwise process that is robust, detailed, reliable, and repeatable. It is also worth noting that this is an interactive process and not necessarily a strict linear process. Process repeatability can only be achieved with clearly stated specifics for each of the data science process steps, which is the purpose of this book.

Our 8-step DISCOVER process will be supplemented with more details in later chapters to include the required activities to unlock the power of your data.

Data Science DISCOVER Process Step 1: Define Research questions, Problem, and Goals. Data Science project success does not happen by accident. Their success requires adherence to rigorous and disciplined best practices at every step in the process. In the appropriate chapters of this book, these best practice details will be provided. A summary of all best practices will be listed in the Appendix. Listed below are examples of **"Define Research questions, Problem, and Goals** for the project": Step 1, with more details listed below:

1. What business, customer, clinical, or public health problem are you trying to solve?
2. What research questions should this project answer?
3. What are your expected project goals and deliverables?
4. How will this project create value, benefit patients, customers, or organizations, and/or improve the decision-making processes for healthcare and other professionals?
5. What is the scope and purpose of the study?
6. What is the economic impact of this project?
7. Which metric(s) will be used to define success for this project?
8. How will you ensure that ethical, fair, and bias-free analysis will not benefit one population group over another?
9. What are the known project limitations?
10. Which domain experts, clinical experts, and other resources are required to ensure the correct interpretation of the project results? [33-35]

It should be noted that the data science process can also be shown as a circular or iterative process. A circular representation emphasizes the iterative nature of data science, where the steps are not strictly sequential but can be revisited and refined based on insights gained from the analysis. It highlights the feedback loop between different stages, allowing for continuous improvement and refinement of the analysis. In the next chapter, we show a version of our 8-step DISCOVER process that displays many of the possible iterations and feedback loops present between data science process steps.

1.1.7 The 3 Levels of AI for Healthcare

AI is currently being used in e-commerce, healthcare, finance, self-driving cars, recommender systems and in so many different areas. In this section we explore different levels of AI applications for healthcare. Figure 1.4 displays some alternate definitions of the "AI" acronym for healthcare with some examples for each of the three levels of AI. AI is typically defined as general intelligence and narrow intelligence systems.

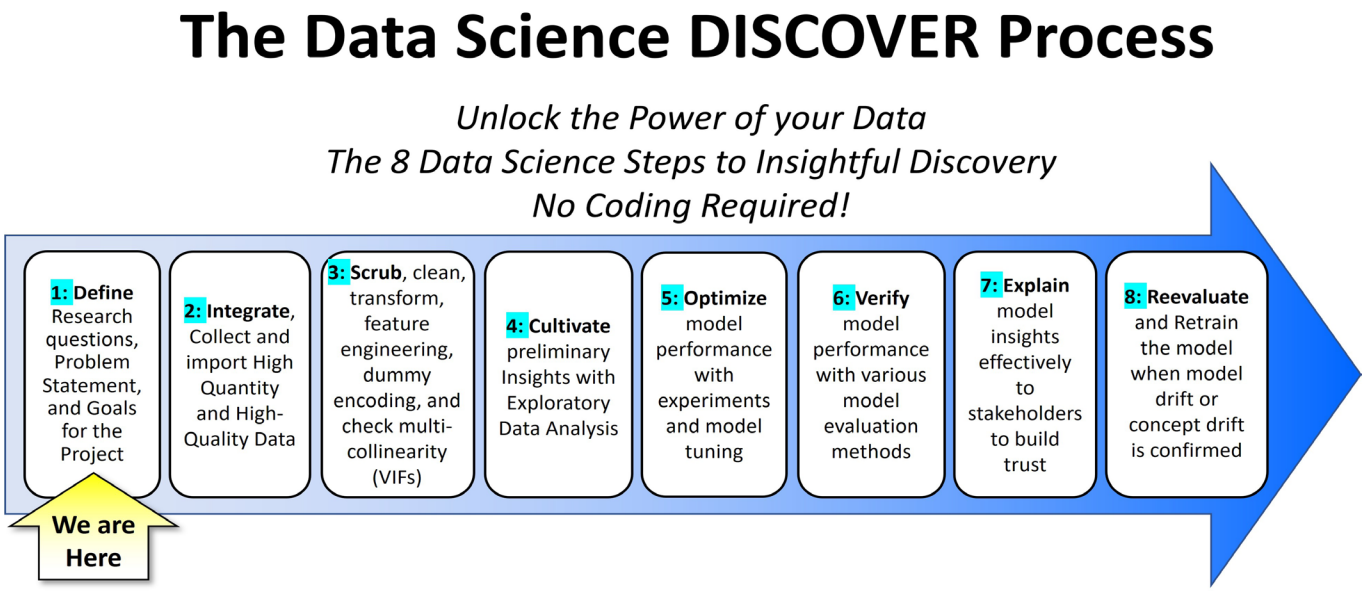

Figure 1.3 The 8-Step Data Science DISCOVER process

As shown in Figure 1.4, we define three levels of AI that the Data Science process can offer to benefit healthcare professionals and their patients, as described below.

1.1.7.1. Actionable Insights (AI Level 1)

Some refer to this level as "turning data into actionable insights with machine learning." [36] These insights are based on robust predictions that the data science process can provide if it has access to a larger data set from which the algorithms can be trained. Such insights can include soliciting and analyzing health risk factors to predict illnesses and diseases. How is this different from what happens in the current healthcare process? Today, every new doctor or clinic we visit requires that we fill out a health survey that asks certain health-related lifestyle questions, current symptoms, current medical issues, medication history, personal health history, and family health history questions. The answers to these questions are usually reviewed by an individual doctor who factors our responses to those questions into their plan for testing, diagnosis, treatment, and care.

The opportunity for AI applications at this level of the patient survey stage is as great as it is underutilized. If the survey questions were aligned with a national database that included the responses to such questions and the verified patient diagnosis, a learning algorithm could be created to predict outcomes within a specified degree of accuracy. We refer to such AI as Actionable Insights about health risks extracted from large national databases. Such insights are actionable since they can create a general risk assessment of a patient before collecting data from a personal diagnosis, lab tests, and vital sign checks.

A great example of this level of AI is the automated ML-driven clinical mortality risk scoring framework called AutoScore.[37] AutoScore was trained on almost 45,000 patient admission encounters in the ICU between 2001 and 2012. A software package in R was also developed to demonstrate and share this method. The method showed high-performance levels compared to more complex methods intended to predict which patients were at the most risk of experiencing adverse events or worsening health conditions. This method and study reported various study limitations. Still, it offers hope that a real-time mortality risk scoring system could be created and linked to an electronic health record system. This method's mortality risk score may provide an early warning system for high-risk patients that may avoid adverse events. This Level 1 of AI is also applicable to any other sector and industry outside of healthcare.

1.1.7.2. Augmented Intelligence (AI Level 2)

This level of artificial intelligence is deemed to be of a high enough level of accuracy that it can be used as a serious and silent second opinion for healthcare professionals. This term has other commonly used synonyms such as Intelligence Amplification, machine augmented intelligence and other terms.[38] The human still makes the final decision for the diagnosis or plan of care. Still, the AI model should be taken seriously if trained on a large database, and the predictive model offers high model accuracy levels. If the large or national database includes

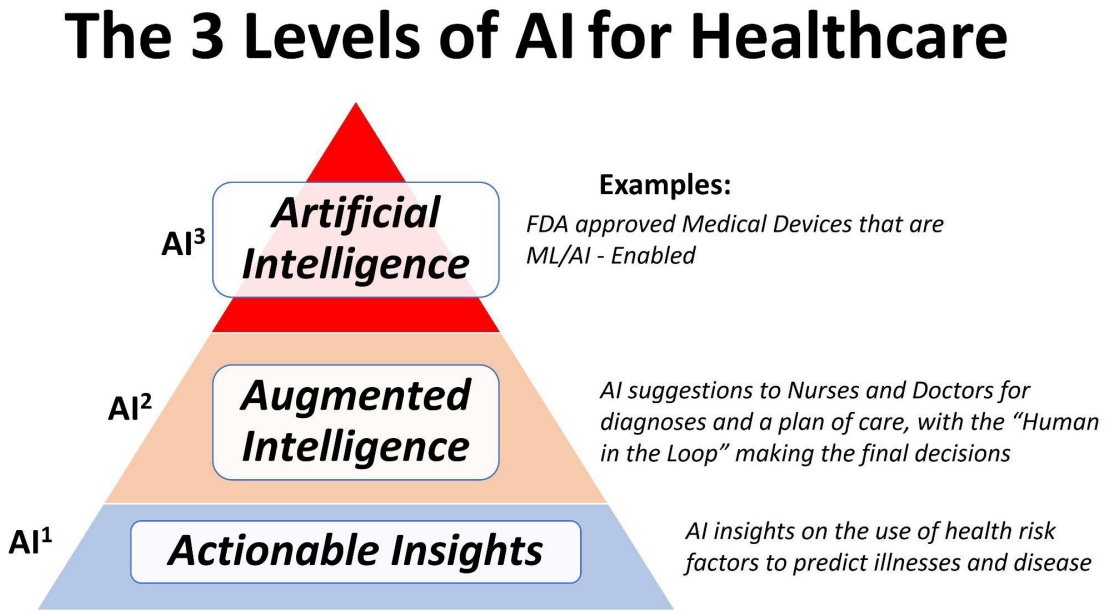

Figure 1.4 The three levels of AI for Healthcare

the effectiveness of specific care plans for illnesses and diseases, a learning algorithm could recommend such steps for patients that match the same criteria.

An example of augmented intelligence would be the application of a real-time AI algorithm trained on an extensive national database of home healthcare patients to determine the best plans of care to improve the Activity of Daily Life (ADL) scores for patients with a similar case mix. The AI algorithm would offer the healthcare professional recommendations, but the human would make the final decision. As mentioned, if the AI model reports very high prediction accuracy levels, the AI recommendations should be taken more seriously.

Another potential augmented intelligence example would be the development of a trained algorithm that factors in all treatment information, medication, ICD10 codes, lab data, vitals data, and real-time 24/7 telemetry data on many acute care heart disease patients under the care of a cardiologist. Suppose such an algorithm were improved to produce high accuracy levels for health condition predictions, responses to various plans of care, and medications. In that case, it should also be considered a reliable real-time advisor for the cardiologist who will make all final decisions. This Level 2 of AI is also applicable to any other sector and industry outside of healthcare.

1.1.7.3 Artificial Intelligence (AI Level 3)

This highest level of AI is reserved for those healthcare algorithms that perform at least as competitively as a trained healthcare professional. The FDA has approved over 520 medical devices that are ML/AI-enabled. 521 of these devices are listed on the FDA website with their note that the list is incomplete.[39] Figure 1.5 shows a Pareto chart listing the main uses for these medical devices that they have approved. The top four areas of 16 total sites make up 91.7% of all areas of ML/AI-enabled devices as listed here: Radiology (75.2%, Cardiovascular (10.9%), Hematology (2.9%), and Neurology (2.7%).

The FDA states that "One of the greatest potential benefits of ML resides in its ability to create new and important insights from the vast amount of data generated during the delivery of health care every day." [39] None of the successful applications of ML and AI in healthcare would have been possible without adherence to a rigorous data science process as described in Figure 1.4 and throughout this book. This Level 3 of AI is also applicable to any other sector and industry outside of healthcare.

1.2 WHAT DO DATA SCIENTISTS DO?

Statistics, mathematics, and programming are the cornerstones of data science, but there are many other essential requirements as listed below. Keep in mind that this skills list and priorities may vary based on the sector and company.

The fundamental skill sets and expertise required for data scientists are:

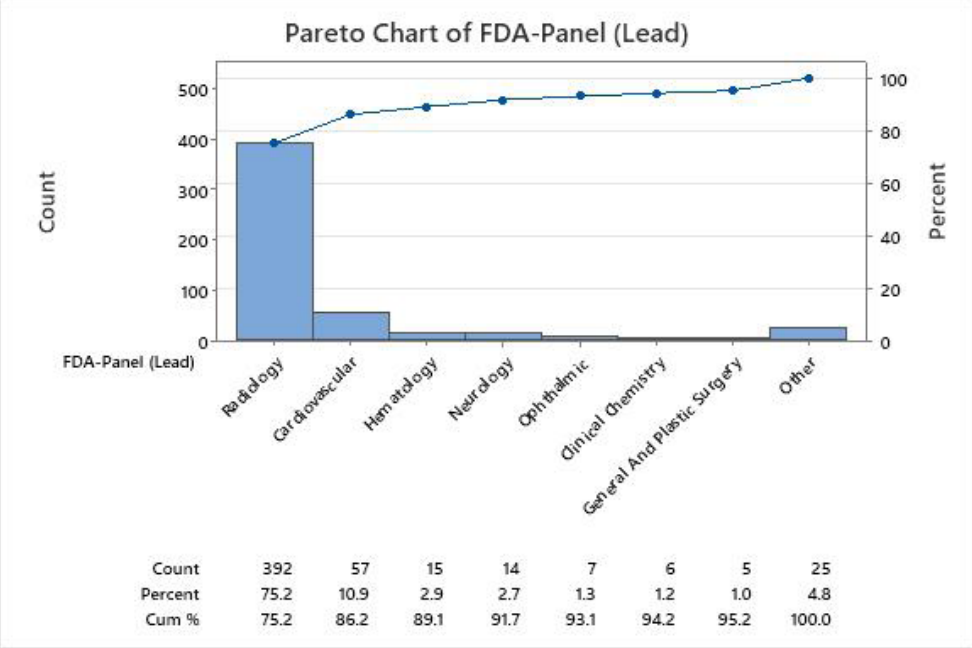

Figure 1.5 The types of ML/AI medical devices that are FDA approved

- Mathematics and statistics
- Domain expertise, e.g., business, manufacturing, service, healthcare, etc.
- Programming in multiple languages: R, Python, SQL, etc.
- Database management and data warehousing
- Predictive modeling and descriptive statistics
- Machine learning and artificial intelligence
- Big data
- Communication and presentation (soft skills or eSkills) [40]

Figure 1.6 shows the average time spent by data scientists on different activities based on a 2022 survey. Based on this survey, only 27% of the time is spent on modeling, whereas 67% of the time is spent on data preparation, cleaning, visualization, and reporting.[41] It bears noting that only 5% of survey respondents were in the healthcare sector, so these results may differ for healthcare.

Very few data scientists have all these skills at the beginning of their careers. Most require additional experience in a specific sector to become comfortable with the technology and the domain. Data scientists are involved in multiple processes, from finding and curating the data to building, deploying, and maintaining the models, and presenting the progress. *Data engineers* are considered different from data scientists. A data engineer has some of the skills of a data scientist, but the emphasis is on advanced programming and systems development. Data engineers create data pipelines and software solutions for data, whereas data scientists are more likely to devote more time to the statistics,analytics required, and modeling. [42] In any sector, there are data professionals in addition to data scientists and engineers, as displayed in Table 1.2. [43]

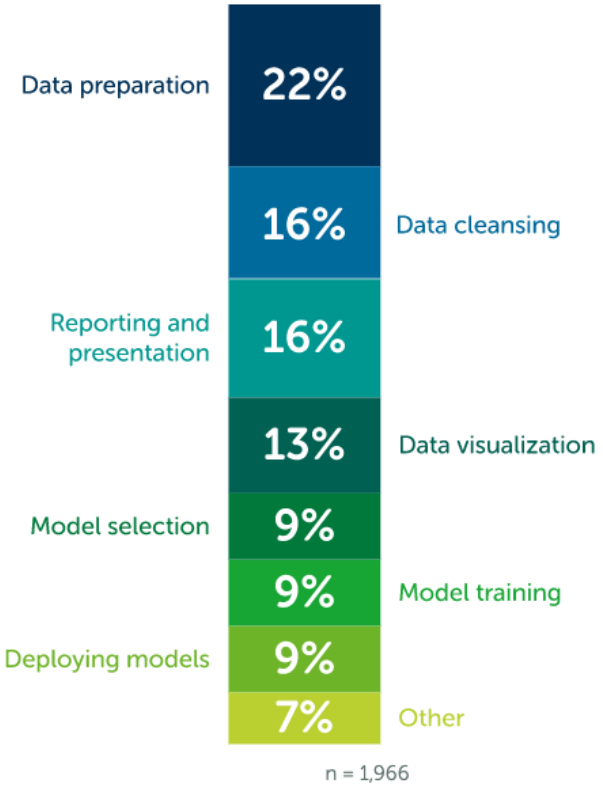

Figure 1.6 Time spent by data scientists [41]

1.3 WHY LEARN DATA SCIENCE?

Because we live and work in a data and tech-driven society, data literacy is becoming required for many domains. With the increasing use of machine learning, natural language processing, and artificial intelligence, the average professional must understand where it fits into their domain and how it will affect the future requirements of their profession.[44] Another benefit of understanding data science is applying the knowledge to better understand analytical research papers and scientific publications. It is common now to see predictive analytics based on machine learning algorithms instead of statistical methods in publications, so it is important to understand these methodologies. In medicine, the convergence of genomics and traditional medicine creates an increased need to understand newer approaches, such as unsupervised learning, and various deep learning strategies. In the manufacturing sector, data science enables manufacturers to optimize quality, predict machine breakdown events, manage inventories, and manage complex supply chains.[45] In the service sectors, data science enables improved customer insights that can improve the customer experience, improve demand forecasting, analyze customer sentiment analysis, and detect fraud. Learning data science has many career-related benefits due to the high demand for data science skills in many different industries.[46]

1.4 DATA SCIENCE EDUCATIONAL CHALLENGES

Data science is challenging since you must simultaneously learn various domains such as mathematics, statistics, a programming language, and computer science as is usually true with a master's level data science program. Several areas of data science are challenging for all data science students, such as understanding neural networks, image, and text analysis. Programming expertise is strongly recommended for anyone who will spend the majority of their time doing data science. However, many people desire to be proficient but not necessarily an expert in the field of data science. For such individuals, tools such as Orange and JASP make the data science

Table 1.2 Data Professionals

Title	Job Description
Data Scientist	A jack of all trades. They offer insights into the best solutions for a specific project while uncovering larger patterns and trends in the data. They also research and develop new algorithms and approaches.
Data Analyst	Responsible for different tasks such as visualizing, transforming, and manipulating the data. Sometimes they're also responsible for web analytics tracking and A/B testing analysis.
Data Engineer	Responsible for designing, building, and maintaining data pipelines. They need to test ecosystems for businesses and prepare them for data scientists to run their algorithms.
Data Architect	Data architects share common responsibilities with data engineers. They both need to ensure the data is well-formatted and accessible for data scientists and analysts and improve the data pipelines' performance.
Data Storyteller	A data storyteller needs to take data, simplify it to focus on a specific aspect of the data, analyze its behavior and then use their insights to create a compelling story that helps people (fellow teammates, customers, etc.) better understand a given phenomenon.
Machine Learning Scientist	Most often, when you see the term "scientist" in a job role, it indicates this job role requires researching to develop new algorithms and insights. In this case, a machine learning scientist researches new approaches to data manipulation to design new algorithms.
Machine Learning Engineer	In addition to designing and building machine learning systems, machine learning engineers must run tests (such as A/B tests) while monitoring the different systems' performance and functionality.
Business Intelligence Developer	They design strategies that allow businesses to find the information they need to make decisions quickly and efficiently.

learning journey easier to sign up for. Learning basic programming is not difficult for some individuals but it is a major roadblock for many. Programming complex neural networks such as convolutional neural networks or transformer-based language models can be challenging for experienced coders.[47] Furthermore, finding data of sufficient volume and quality can be difficult and frequently requires individuals trained in data science.

1.5 WHY DID WE CREATE THIS TEXTBOOK?

This textbook was originally drafted as training materials to teach data science workshops for clinicians in the US, as a function of the Medical Intelligence Society. This society promotes data science, machine learning, and artificial intelligence in the healthcare domain.[48] Since then, the book has been supplemented with training materials and data analysis examples that are now applicable to any industry. Our continued workshops walk participants through the book in an interactive manner with the aim to increase interest and competency in applied data science by using open-source software and performing analytical exercises on datasets from multiple industries. We also intend to offer certification levels for professionals interested in documenting their knowledge and application skills in no-code data science techniques. In this book, multiple links to external resources have been added to provide additional useful information. Similarly, valuable sources of datasets and additional statistical and machine-learning concepts have been added to this book. We hope to create an army of no-code data scientists who can maximize the power of open-source software to learn basic, intermediate, and advanced levels of applied data science with a limited need for higher math or statistical knowledge. We put a high priority on teaching professionals how to correctly create and interpret all of their data visualizations and analysis work while being able to identify and recommend the next logical analysis steps. A chart or analysis result should never be added to a presentation or research report without the addition of clear interpretation comments. You cannot build trust in your analysis and predictive models with stakeholders and customers unless you learn to be a master storyteller