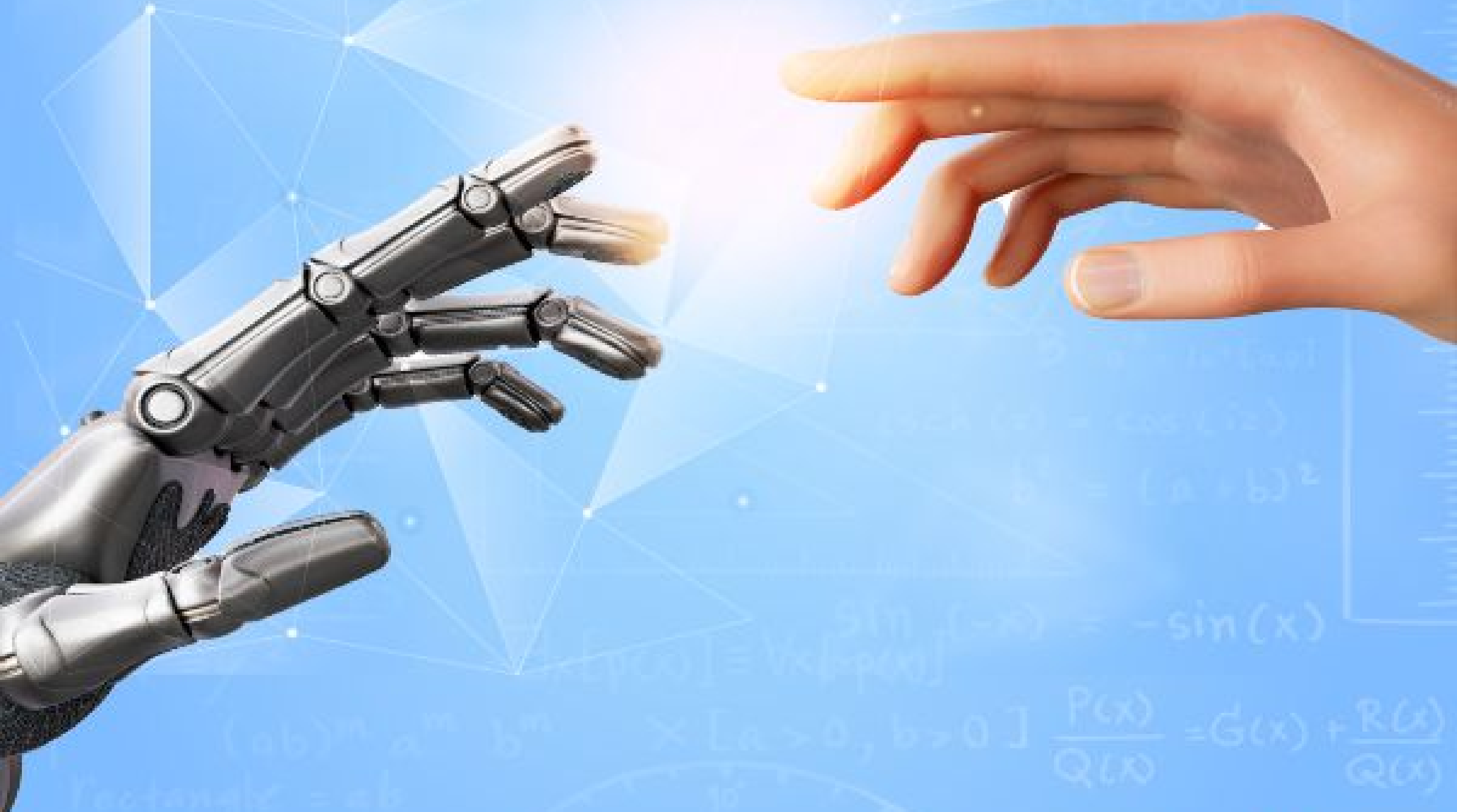
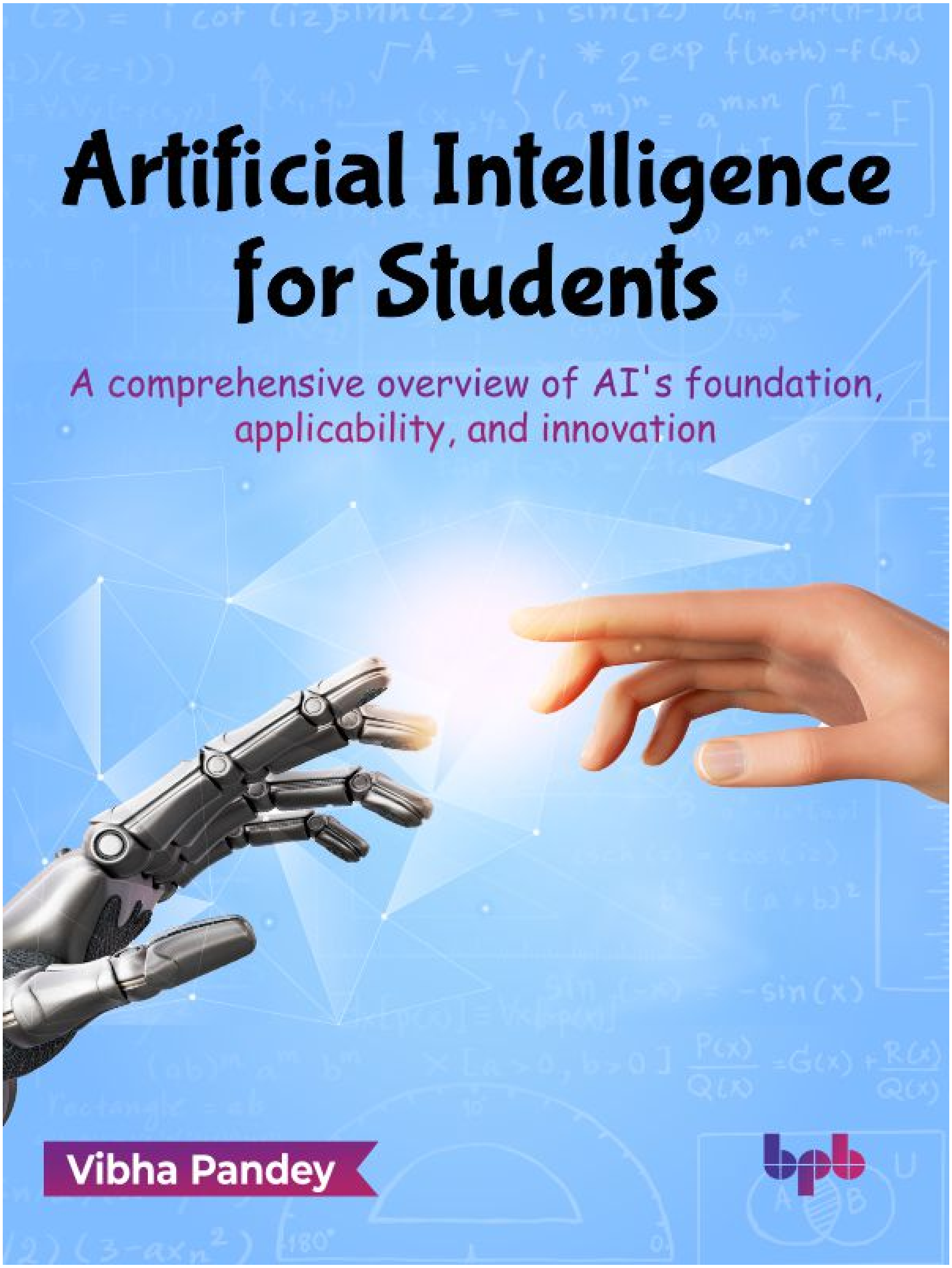
Artificial Intelligence for Students

A comprehensive overview of AI's foundation, applicability, and innovation



Vibha Pandey

bpb


Artificial Intelligence for Students
A comprehensive overview of AI's foundation, applicability, and innovation
Vibha Pandey
bpb

Artificial Intelligence for Students

A comprehensive overview of AI's foundation, applicability, and innovation

Vibha Pandey

www.bpbonline.com

First published: 2023

Published by BPB Online
WeWork
119 Marylebone Road
London NW1 5PU

UK | UAE | INDIA | SINGAPORE

ISBN 978-93-5551-794-4

www.bpbonline.com

Dedicated to

My parents
Ms. Lalita Pandey** and **Dr. R.K. Pandey

&

My mentor
Mr. Dalbir Dhankar

&

My Nephew
Master Abhiram Chaturvedi

Foreword

In today's world, it's essential for everyone to be able to encode and decode data, which has become an integral part of our lives rapidly. This book is for everyone, from beginners to experts; the author proves that 'One size may sometimes fit all.' In this case, the basics of data science are explained in such a way that the often-dreaded subject becomes a lot more accessible to everyone. As a data scientist, I highly recommend this book to anyone interested in gaining a deeper understanding of data science and machine learning. It is a comprehensive resource that covers a range of topics, from data structures and data representation to clustering, classification, and storytelling.

The section on data structures is highly informative and crisp, providing a clear breakdown of the different types of data structures and how they can be used to organize and manage data efficiently. It will prove to be an elixir of knowledge, especially helpful to beginners who may not be familiar with the concept.

The chapter on data representation is also well-written, providing a detailed overview of the different ways data can be presented and how those presentations can impact the insights and conclusions drawn from the data. The section on finding the best-suited graphical representation for the data provides practical insights into how data can be effectively communicated to others.

Concepts of clustering and classification are both covered extensively, with clear explanations of their differences and how they can be used for pattern identification in machine learning. The chapter on classification provides an in-depth explanation of how data labeling works and the importance of understanding true positives, true negatives, false positives, and false negatives. The introduction of logistic regression as a tool for binary classification is also a great addition to the chapter that enhances its practicality.

An often untouched aspect of data science, storytelling, is a unique and valuable addition to the book. It emphasizes the importance of storytelling in conveying insights and making data more accessible. The practical tips and best practices shared in this chapter are sure to be helpful in any storytelling endeavor.

One of the standout features is the section on ethical considerations in machine learning. It highlights the biases that exist in human society and how they can influence the accuracy of machine learning algorithms. The section on principles for ethical AI is particularly insightful, providing a framework for creating machine learning models that are fair, transparent, and accountable.

Overall, this book is a must-read for anyone interested in data science or machine learning. The chapters are well-structured and informative and provide practical insights that can be applied in real-world situations. A beginner in the field can easily get the gist of what is being shared in the group due to the author's ability to explain heavier concepts in a fluid manner. The author's writing style is warm and professional, making it an enjoyable and engaging read. I would highly recommend this book to anyone looking to expand their knowledge in the field of data science.

-Santanu Bhattacharya

Santanu Bhattacharya currently holds the position of Chief Technologist at NatWest Bank. He also holds a Ph.D. from NASA's Goddard Space Flight Center/UMD and is an alumnus of MIT and IIT-Bombay. As a speaker at the World Economic Forum (WEF) Davos 2020 and a writer, he has been covered in publications such as Mashable, TechCrunch, Forbes, Le Monde, and Economic Times, among others.

About the Author

Vibha Pandey has rich corporate experience of more than 25 years with MNCs like PwC, Oracle, Nortel, Siemens, and Samsung, to name a few. She played central roles in numerous telecom, and IT projects in different profiles, such as a software engineer, product life cycle management, test engineer, presales, sales, business development, and key account management, including project implementation in the US, Indonesia, Japan, and India. Currently, she runs a firm focused on telecom and security projects as well as is a National President of AI at WICCI. She also takes up consulting projects to write research papers. She has taken up various other professions, including being a visiting faculty at NIFT, Delhi, where she introduced Systems Thinking as a subject that was made mandatory.

Furthermore, the author has been associated with Smart India Hackathon right from its inception years for both software and hardware divisions as a mentor and evaluator.

About the Reviewers

Vibhu is a software professional passionate about Machine Learning, Cloud Computing, and Agile Development. With over 25 years of experience in the software industry, he has worked on a wide range of projects, from small startups to large enterprises, delivering high-quality software demanding 99.999% availability.

Currently, he leads engineering efforts for observability solutions (both on-premise and in the cloud) that touch around 300,000 customers. He is a hands-on leader, unafraid to roll up his sleeves and get involved in the code when needed.

He profoundly understands Cloud Computing (with multiple certifications with different cloud providers) and constantly explores new technologies and techniques to help deliver better software faster. In his spare time, he teaches Machine Learning to professionals trying to break into new areas.

He is passionate about his work and continually seeks new challenges and opportunities to learn and grow as an engineering leader.

Phani is an experienced Technical Lead with a background in the health care industry. His skillset includes expertise in Predictive modeling, Artificial Intelligence, and Machine Learning using programming languages such as Python, R, and SAS. They also have a solid understanding of Big data, data visualizations, and Storytelling with Data. Throughout their career, they have worked with various relational and non-relational databases as well as data visualization tools such as Tableau and PowerBi.

In terms of education, this individual holds a Bachelor of Technology degree focused in Electrical and Electronics Engineering from Jawaharlal Nehru Technological University. They also pursued a Masters in Business Analytics from the University of Louisville in the USA, where they achieved a remarkable 3.95/4 GPA. Additionally, they are currently pursuing a Ph.D. from the University of Louisville.

Overall, this individual's experience and education demonstrate their expertise in technical fields related to data analysis and their dedication to continuing to expand their knowledge and skillset in this area.

Acknowledgement

I want to express my deepest gratitude to my family and friends for their unwavering support and encouragement throughout this book's writing, especially my parents, my mentor, and my nephew, who has authored books from the very young age of 7 years.

I am also grateful to BPB Publications for their guidance and expertise in bringing this book to fruition. It was a long journey of revising this book, with valuable participation and collaboration of reviewers, technical experts, and editors.

I would also like to acknowledge the valuable contributions of my ex-colleagues and industry alliances during many years working in the tech industry, who have taught me so much and provided valuable feedback on my work.

Finally, I would like to thank all the readers who have taken an interest in my book and for their support in making it a reality. Your encouragement has been invaluable

Preface

AI is a discipline in computer science that focuses on developing intelligent machines, machines that can learn and then teach themselves. These machines, then, can process vast amounts of data than humans can and several times faster. However, AI can go across all disciplines to change the world for the better– from creating new healthcare solutions to designing hospitals of the future, improving farming and our food supply, helping refugees acclimatize to new environments, improving educational resources and access, and even cleaning our oceans, air, and water supply. The potential for humans to improve the world through AI is endless as long as we know how to use it.

This book is designed to provide a comprehensive guide to a planned sequence of instructions consisting of units meant for developing the employability and vocational competencies of students opting for skill subjects along with other educational subjects.

Throughout the book, you will learn about the key aspects of Artificial Intelligence with real-world examples for readers to relate.

This book is intended for readers who are new to Artificial Intelligence and want to explore and experiment in this field.

With this book, you will gain the knowledge and skills to start developing applications using existing frameworks of interest, such as chatbots. I hope you will find this book informative and helpful.

Chapter 1: Introduction: AI for Everyone - explains artificial intelligence and machine learning, terminologies, and related concepts. It also describes the AI products/ applications in society and their being different from non-AI products/applications. Readers also learn about jobs that may appear in the future.

Chapter 2: AI Applications and Methodologies - presents an overview of areas where artificial intelligence can be applied (like in the field of computer vision, speech, text, etc.). Readers also get an overview of deep learning. This chapter covers the impact of AI on our society and how we can get ready for the future, that is, the AI age.

Chapter 3: Mathematics in Artificial Intelligence – revisit the mathematics involved in artificial intelligence, such as linear algebra, statistics, and set theory. This also covers the basics of graphs and describes the application of math in AI. Readers will learn about representing data in terms of mathematical formulas.

Chapter 4: AI Values (Ethical Decision-Making) - covers the ethics, bias, and impacts of bias on society. It gives special attention to issues and concerns around AI. It helps readers learn to spot issues in the data, make arguments, and apply rules.

Chapter 5: Introduction to Storytelling – this chapter is all about storytelling, the need, storytelling with data, insights from storytelling, and more. It allows the reader to learn to apply imagination, mapping the plot into key events and increasing memory retention. It also guides the creation of blogs, videos, and other content as per the audience and about the conflict and resolution.

Chapter 6: Critical and Creative Thinking – explains the design thinking framework, that is, understanding the problem and being able to express the same. Readers learn to develop/innovate from the design of a solution.

Chapter 7: Data Analysis - explains types of structured data and statistical principles such as frequency tables, mean median, mode, range, and more. Readers learn to represent data in terms of graphs and statistical models. By the end of this chapter, the reader is able to represent a simple problem in terms of numbers.

Chapter 8: Regression – explains mathematical concepts such as correlations, regression, and other related terms. At the end of the chapter, readers learn to relate data with regression and correlation. Readers also get to know about everyday applications of these mathematical concepts.

Chapter 9: Classification and Clustering - explains in detail classification and its types, the kind of problems that may be placed under the category of a classification problem, and where to apply classification principles. Readers are also made aware of the impact of the application of incorrect algorithms on society. In the remaining half of this chapter, readers learn about clustering problems and their application, and why it is called clustering. Readers get an overview of the application of clustering problems using standard models.

Chapter 10: AI Values (Bias Awareness) - explains what ethics are, the impact of ethics on society, as well as the impact of bias on AI functioning. Readers are also able to learn about the impact of biases in data and how to de-bias or neutralize biased data. By the end of the chapter, readers are able to easily find bias in the acquired dataset

Chapter 11: Capstone Project - introduces readers to commonly used algorithms and the science behind them. This chapter also engages the readers in understanding and decomposing a problem, the analytical approach, and data requirements and collection. This chapter also introduces the validation of the model quality and metrics of model quality. The chapter ends with showcasing a compelling story through all the methodologies and learnings that readers are exposed to in the chapter.

Chapter 12: Model Lifecycle (Knowledge) - this chapter explains different aspects of the model as well as the lifecycle of an AI model.

Chapter 13: Storytelling Through Data - explains the need for storytelling and various related topics, such as the creation and ethics of stories, expressing the related data with suitable charts. This chapter also captures the stories during the step of predictive modeling and ends with the best practices of storytelling.

Chapter 14: AI Applications in Use in Real-World– mentions different fields and the associated real-world AI applications.

Coloured Images

Please follow the link to download the
Coloured Images of the book:

https://rebrand.ly/skz5nh8

We have code bundles from our rich catalogue of books and videos available at **https://github.com/bpbpublications**. Check them out!

Errata

We take immense pride in our work at BPB Publications and follow best practices to ensure the accuracy of our content to provide with an indulging reading experience to our subscribers. Our readers are our mirrors, and we use their inputs to reflect and improve upon human errors, if any, that may have occurred during the publishing processes involved. To let us maintain the quality and help us reach out to any readers who might be having difficulties due to any unforeseen errors, please write to us at :

errata@bpbonline.com

Your support, suggestions and feedbacks are highly appreciated by the BPB Publications' Family.

exclusive discounts and offers on BPB books and eBooks.

Piracy

If you come across any illegal copies of our works in any form on the internet, we would be grateful if you would provide us with the location address or website name. Please contact us at **business@bpbonline.com** with a link to the material.

If you are interested in becoming an author

If there is a topic that you have expertise in, and you are interested in either writing or contributing to a book, please visit **www.bpbonline.com**. We have worked with thousands of developers and tech professionals, just like you, to help them share their insights with the global tech community. You can make a general application, apply for a specific hot topic that we are recruiting an author for, or submit your own idea.

Reviews

Please leave a review. Once you have read and used this book, why not leave a review on the site that you purchased it from? Potential readers can then see and use your unbiased opinion to make purchase decisions. We at BPB can understand what you think about our products, and our authors can see your feedback on their book. Thank you!

For more information about BPB, please visit **www.bpbonline.com**.

Join our book's Discord space

Join the book's Discord Workspace for Latest updates, Offers, Tech happenings around the world, New Release and Sessions with the Authors:

https://discord.bpbonline.com

Table of Contents

CHAPTER 1

Introduction: AI for Everyone

Introduction

The objective of this chapter is to learn the Artificial Intelligence concept, classification, and its components. We will start the chapter with a discussion on what is the need for Artificial Intelligence and end with the career opportunities which are available in this space.

Structure

In this chapter, we will be discussing:

- Data explosion
- What is Artificial Intelligence
- Artificial Intelligence: History and evolution
  - The father of AI
- Types of Artificial Intelligence
  - Based on the capabilities of AI
  - Based on the functionality of AI
- What is machine learning
- What is data
- What is deep learning
- Machine learning techniques and training
- Neural networks
- What machine learning can do and cannot do
- Key differences between artificial intelligence and machine learning
- Artificial Intelligence project life cycle
- Career opportunities in artificial intelligence

Data explosion

We live in a world with an ever increasing amount of data that both humans and machines generate. It far outpaces humans' ability to extract meaningful information and make informed and complex decisions based on the extensive data to process.

Every day, we create roughly 2.5 quintillion bytes of data (that's 2.5, followed by a staggering 18 zeros!)

We may not be aware, but we have been using Artificial Intelligence based technologies in our daily routine. Scientists found that an average person today can process as much as 74 gigabytes (GB) of data a day.

Artificial Intelligence is a technology that is transforming every walk of life with its five basic components include learning, reasoning, problem-solving, perception, and language understanding.

This book is written with the goal of explaining the technology with examples. Let us start with brushing some basic definitions and visiting the history of Artificial intelligence to set the context.

What is a machine?

A machine is a piece of equipment with moving parts that humans design to do a particular job. A machine usually needs electricity, gas, steam, and so on to work.

What is a computer?

A computer is an electronic machine that can store, find and arrange information, calculate amounts, and control other machines.

What is Artificial Intelligence

The human brain has the ability to think, read, learn, remember, reason, and pay attention. Such capabilities are termed cognitive skills. The term "**Intelligence**" is used for cognitive (connected with the processes of understanding) skills and thinking ability of humans and animals. We may also call it "natural intelligence."

Then what is *Artificial Intelligence* (referred to as AI in the remaining book)?

The terminology comprises of two words "*Artificial*" and "*Intelligence*." **Artificial** refers to something that is not natural or is made by humans. AI is, then, intelligence demonstrated by a computer (an electronic machine), hence, it can also be referred to as "machine intelligence."

In other words, AI is best described as machines having human-like cognitive skills of learning and problem solving by making decisions in such a way that they can be associated with human minds.

To summarize, AI is a field of computer science (not science fiction) combining robust datasets with the aim of having computers simulate intelligent processes. Here the computer needs AI implemented in its system to demonstrate AI capabilities.

Today AI contributes much to our human lives. Industries, including retail, healthcare, manufacturing, agriculture, insurance, and finance, are already harnessing the many benefits of AI. There are companies that provide AI solutions, while others use AI within their organization to manage internal business operations or business growth. A few real world companies in the preceding categories will be described by the end of this book.

Artificial Intelligence: History and evolution

Artificial Intelligence (**AI**) has been studied for decades and is still one of the most elusive subjects in Computer Science.

The year 1943: *Warren McCulloch* and *Walter pits* 1943 proposed a model of artificial neurons.

The year 1949: *Donald Hebb* demonstrated modifying the connection strength between neurons. His rule is now called Hebbian learning.

The year 1950: *Alan Turing*, an English mathematician, pioneered Machine learning in 1950. Alan Turing proposed a test in his "Computing Machinery and Intelligence" publication. The test, called a Turing test, can check the machine's ability to exhibit intelligent behavior equivalent to human intelligence.

The period between the 1950s and the 1970s revolved around the research on neural networks; the following three decades (1980s to 2010s) were the development of the applications of Machine Learning.

In *Figure 1.1*, a brief timeline of the past six decades of how AI evolved from its inception has been depicted:

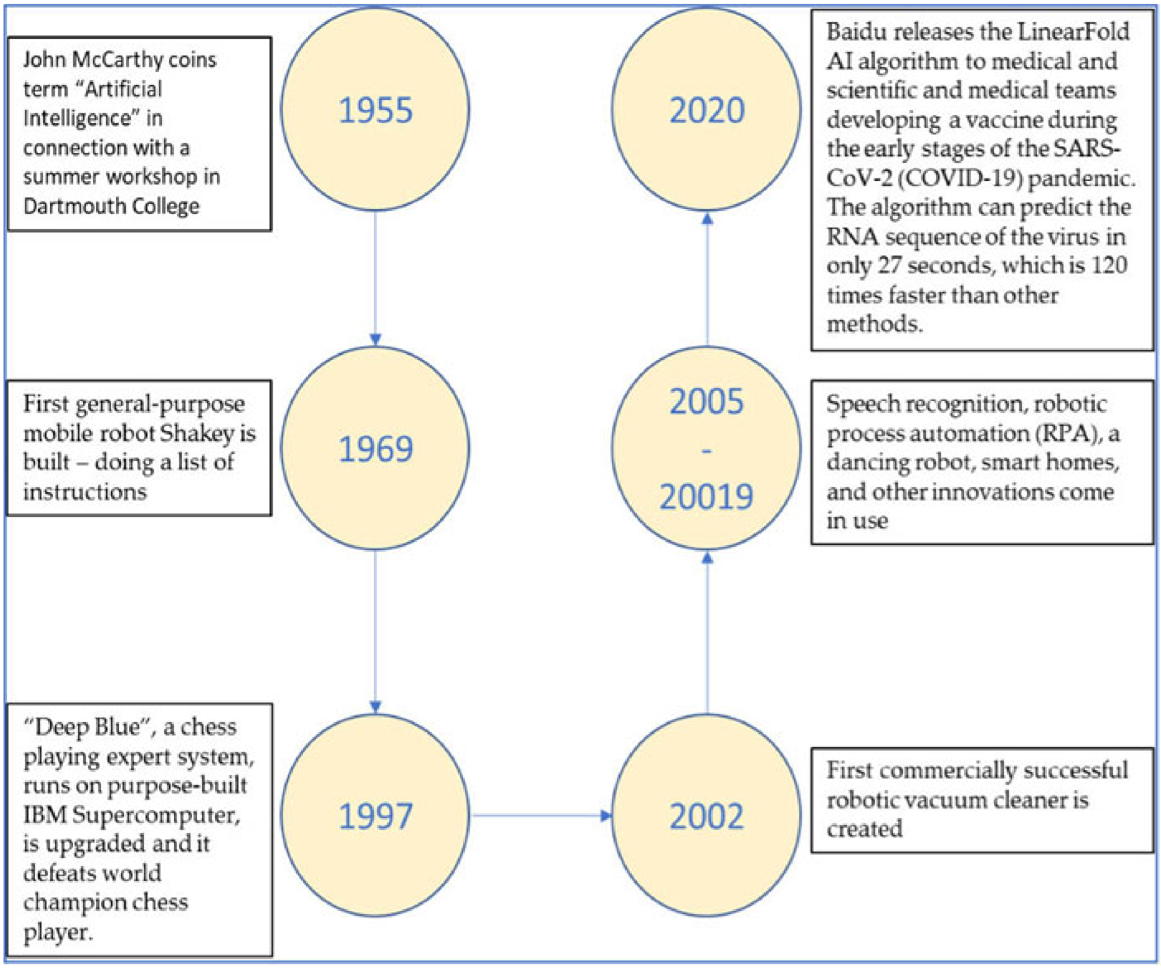
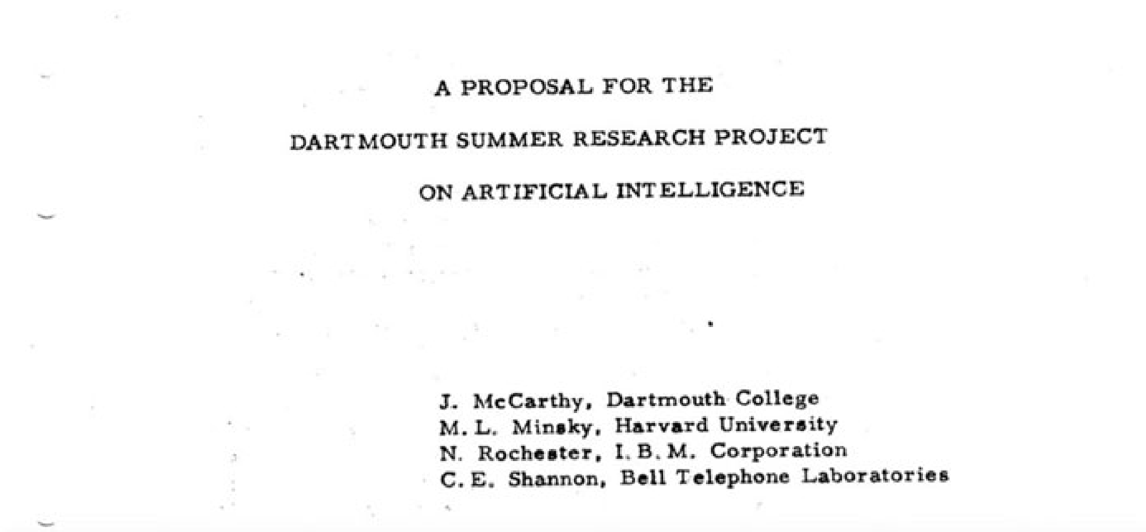


Figure 1.1: *The evolution of AI during the last six decades*

The father of AI

John McCarthy is widely recognized as the "*Father of Artificial Intelligence*" due to his astounding contribution and innovations in the field of Computer Science and AI. John McCarthy coined the term "Artificial Intelligence" in his 1955 proposal for the *1956 Dartmouth Summer Research Project*, the first artificial intelligence conference, which was a seminal event for artificial intelligence as a field. Refer to *Figure 1.2* which depicts the proposal where the term Artificial Intelligence was coined:


A PROPOSAL FOR THE

DARTMOUTH SUMMER RESEARCH PROJECT

ON ARTIFICIAL INTELLIGENCE

J. McCarthy, Dartmouth College
M. L. Minsky, Harvard University
N. Rochester, I. B. M. Corporation
C. E. Shannon, Bell Telephone Laboratories

Figure 1.2: *Proposal where the term Artificial Intelligence was coined*

In his proposal, he stated that the conference was "*to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it*."

In 1956, for the first time, Artificial Intelligence was coined as an academic field. The researchers thought about ways to make machines more cognizant, and they wanted to lay out a framework to better understand human intelligence.

John also paved the way for a few of the world's technological innovations like programming languages, the Internet, the web, and robots, to name just a few

He invented the first programming language for symbolic computation, LISP, and invented and established time-sharing. Human-level Artificial Intelligence and common-sense reasoning were two of his major contributions.

Types of Artificial Intelligence

Artificial Intelligence can be classified into two types:

Based on the capabilities of AI

- **Artificial narrow intelligence**

  Artificial narrow intelligence, ANI or Narrow AI, also called "Weak" AI, is goal oriented and is designed to perform singular tasks intelligently and extremely well without any human intervention.

  Language translation and image recognition are examples of common uses for narrow AI. Siri is capable of processing human language and submitting a request to a search engine for retrieval. It explains why Siri is unable to answer abstract and complex queries that require emotional intelligence. It's mere digital assistance to perform basic inquiries and tasks.

Even if Narrow AI appears to be considerably more sophisticated, it operates within a pre-determined, predefined scope. It can attend to a task in real-time, but they pull information from a specific dataset. In fact, what may appear as a complicated AI as a self-driving automobile is labeled Weak AI.

Narrow AI is unable to think. They lack the capability for autonomous reasoning, self-awareness, consciousness, and genuine intelligence.

- **Artificial general intelligence**

  Artificial general intelligence (**AGI**), also called "Strong" AI, is an intelligent system with comprehensive or complete knowledge and cognitive computing capabilities.

  In today's world, no true AGI systems exist and remain the stuff of science fiction. Sci-fi movies like "Her," where a human interacts with a machine displaying broad intellectual capabilities to learn, reason, and make own decisions and judgments, while understanding belief systems. True AGI intellectual capacities would exceed human capacities because of its systems' ability to process huge data sets at incredible speeds.

  Hence, no real-world systems as examples here.

- **Artificial super intelligence**

  Artificial superintelligence, or ASI, will be human intelligence in all aspects. ASI is a futuristic notion and idea about AI capabilities to supersede human intelligence. It will be self-aware and intelligent enough to surpass the cognitive abilities of humans.

  Many are concerned about ASI and its impact on humankind Individuals like Tesla CEO Elon Musk warned about the dangers of ASI-powered robots, even predicting "scary outcomes" like in <the movie> "*The Terminator*."

Based on the functionality of AI

AI can primarily be divided into four different categories based on functionality. Let us have a look at each:

- **Reactive AI**

These machines are the most basic type of AI system and perform best when all parameters are known. These machines do not have any memory or understanding of historical data and will not perform desirably in case of imperfect information input. Refer to *Figure 1.3*:

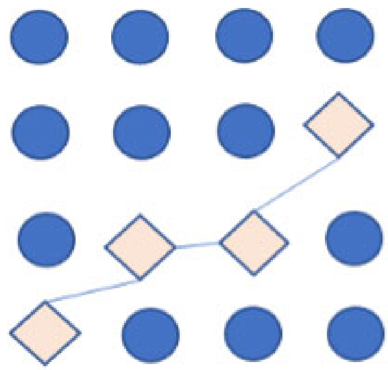


Figure 1.3: *Reactive AI*

These are good for simple classification and pattern recognition tasks where they specialize in just one field of work and can beat humans by their capacities to make faster calculations.

For example, in a chess game, the machine observes the opponents' moves and makes the best possible decision toward its win. This means reactive machines always respond to identical situations in the exact same way every time.

Face recognition is another example.

- **Limited memory**

  Limited memory AI can complete complex classification tasks and uses historical data to make predictions. They keep building on their memory, that is, storing the previous data and predictions, but memory is minimal. Refer to *Figure 1.4*:

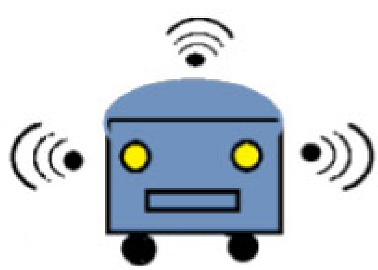


Figure 1.4: *Smart Car*

For example, this machine can suggest a restaurant based on the location data, food preference, and other such parameters that have been gathered.

Self-driving cars are limited memory AI. These use sensors to identify humans and animals crossing the road, obstacles on the path, steep roads, traffic signals, and so on to make better driving decisions.

- **Theory of Mind**

  A robot or a system powered by the Theory of Mind AI will be able to communicate deeper with human beings with its ability to understand thoughts, emotions, and feelings and adjust its behavior (social interaction) in accordance. Refer to *Figure 1.5*:

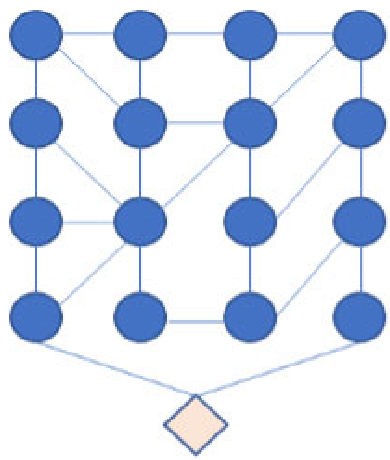


Figure 1.5: *Theory of Mind*

  Such robots/systems will be able to explain their actions, and this is different from the current generation of AI. Theory of Mind AI-powered systems will be able to simulate the consequences of their actions. A new study describes a robot that can predict how another robot will behave, a first step in developing the so-called Theory of Mind

  However, a machine based on this type is yet to be built in its entirety.

- **Self-aware**

  Self-aware machines are the future generation of these new AI technologies. No such system is yet known to have been developed that possesses intelligence, is sentient, and is conscious. Such self-aware systems will be able to interact with and understand both humans and other AIs. Refer to *Figure 1.6*:

Figure 1.6: *Self-aware*

What is machine learning

Machine learning is a method of data analysis that brings automation to analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn on their own from data without being explicitly programmed.

The iterative aspect of machine learning is important because as the system is exposed to new data, it is able to adapt independently. They learn from previous behavior to produce reliable, repeatable decisions and results. It's not a new science– but it has gained fresh momentum.

It is an application of AI that provides the system the ability to automatically learn and improve from experience, that is integrating the output back into the system. Refer to *Figure 1.7*. This figure describes the difference between traditional programming and machine learning. While traditional programming involves a computer running a program with input data and giving an output. Machine learning includes the input and its output fed again into the program which may continuously train itself based on the available data.

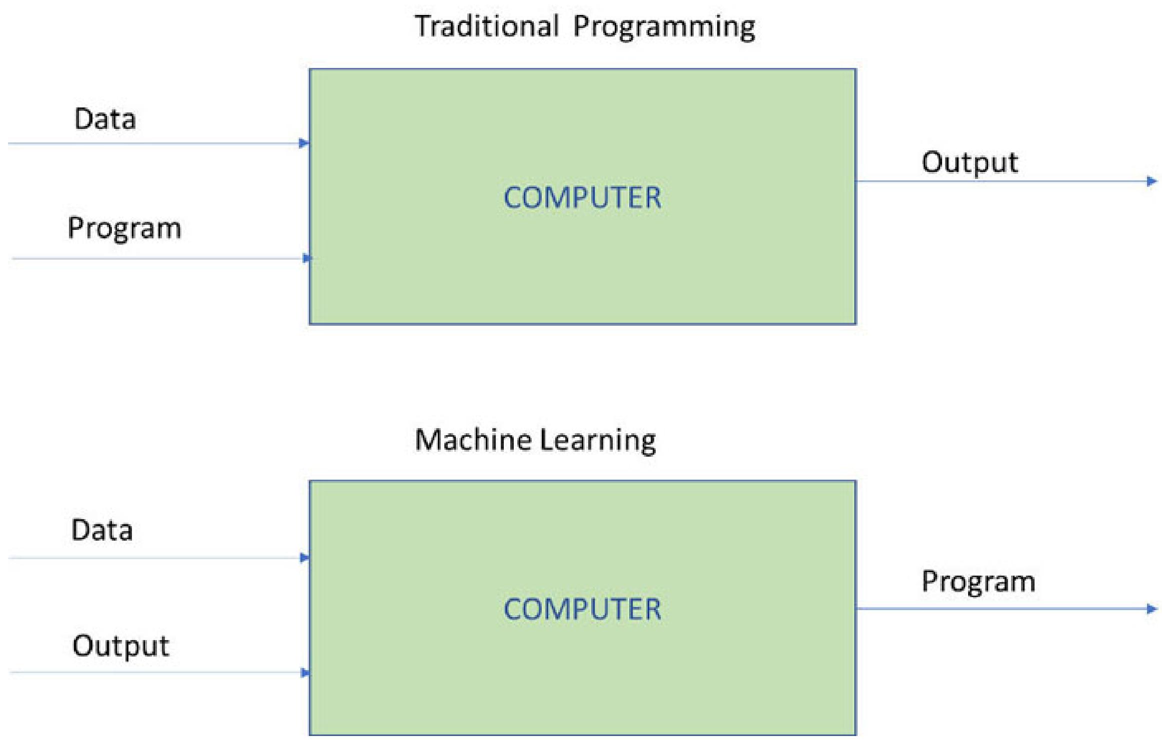


Figure 1.7: *Difference between Traditional Programming and Machine Learning*

Examples

Recognition

- **Image recognition:** Law enforcement uses machine learning-based image recognition tools to identify faces by matching them against a database of people.
- **Speech recognition**: We may have used voice dialing or giving voice inputs to smartphones for google searches. This is also based on machine learning algorithms.
- **Medical diagnosis**: Now, many physicians have started using use chatbots with speech recognition capabilities to discern patterns in patients' symptoms and help diagnose diseases.

Distances

- **Google Maps**: Google Maps does real-time data tracking by informing passengers of traffic and obstacles on the path. It was in form of the crowdiest and/or the shortest routes. These features are machine learning-enabled.

- **Ride apps**: Ride apps like Uber use machine learning to forecast the expected arrival time by taking real-time traffic, GPS data, and Map APIs as input.

Email intelligence

- **Spam**: Ever wonder what few emails go into the spam folder? These are filtered on the basis of machine learning algorithms used by email providers.
- **Email classification**: The classification of emails, say by Gmail, into Primary, Promotions, Social, and so on is also done using machine learning by Gmail.
- **Suggested Smart replies**: Google email - Gmail recently also started suggesting smart replies based on the content of the email for better user experience and delight. These responses are customized per email too.

Social networking apps

- **Facebook**: Facebook automatically reflects faces and suggests friends tag while uploading a pic. Facebook uses AI and ML to identify faces.

What is data

The most vital ingredient in machine learning and AI is the information fed to the systems to build intelligent models. Data refers to information that has been converted into a form that is more efficient for storing, processing, and transferring.

Data may be structured or unstructured, and is collected and stored in a format that makes it faster to be measured, reported, visualized, and analyzed. In contrast, raw data is a term used to describe data in its most basic digital format.

Following are some examples of data:

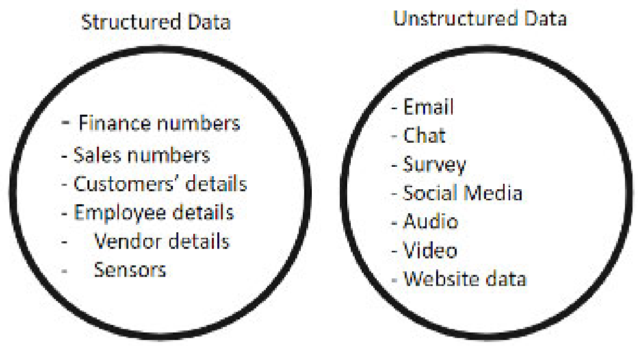


Figure 1.8: *Structured and Unstructured Data*

What is Deep learning

Deep learning is a subset of machine learning. It is a machine learning algorithm that uses deep (more than one layer) neural networks to analyze data and provide output attaining the highest rank in terms of accuracy when it is trained with a large amount of data.

The main difference between deep and machine learning is, machine learning models become better progressively but the model still needs some guidance. As in the programmer needs to fix that problem explicitly in case of inaccurate outcomes. But in the case of deep learning, the model does feature extraction independently.

Examples

- **Chatbots**: Siri, which is Apple's voice-controlled virtual assistant. Is based on Deep Learning and gets smarter day by day by adapting itself according to the user and providing better-personalized assistance.
- **Self-driving / automatic cars:** These are also the examples of deep learning.
- **Google AI Eye Doctor**: One of the initiatives from Google is Automated **Retinal Disease Assessment** or **ARDA** which uses artificial intelligence and deep learning to help healthcare workers detect diabetic retinopathy.
- **AI-based based music composers** and platforms such as Aiva, Amper and Ecrett Music, and so on are built using detailed algorithms that process the inputs of its users. The smart platform efficiently concocts a piece of music that totally fits users' criteria, based on a library of musicological knowledge, and builds stirring music instantly.

- **AI Dream Reader:** A group of researchers from the University of Kyoto in Japan used machine learning to study brain scans or analysis of human functional magnetic resonance imaging, where it could also generate visualizations of what a person is thinking when referring to simple, binary images. They then used deep learning / deep neural networks to decode thoughts.

  Once this technology develops further, it can allow drawing pictures, can visualize human dreams, hallucinations of psychiatric patients, and much more.

Machine learning techniques and training

Machine learning uses three techniques that teach computers to do what comes naturally to humans and animals-learn from the experience:

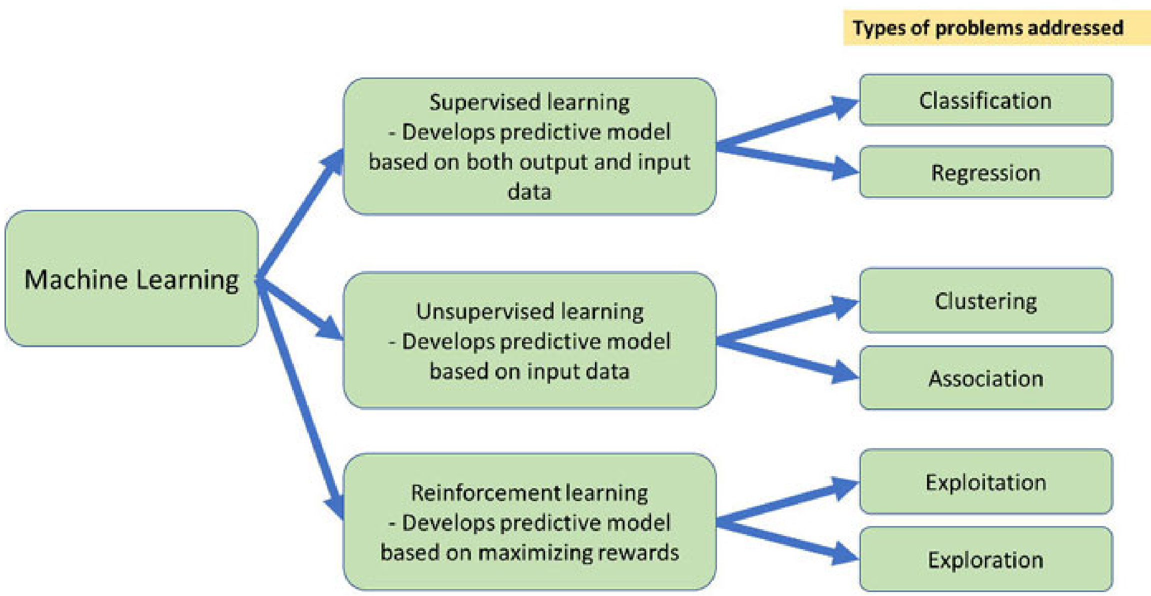


Figure 1.9: *Machine Learning*

Let's understand these three models:

- **Supervised learning**

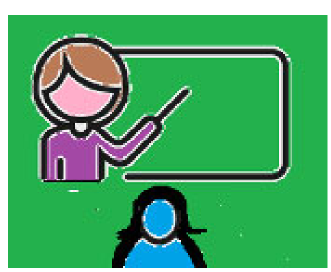


Figure 1.10: Supervised Learning: Learning under the supervision

Supervised learning trains a model on known input and output data to predict future outputs. Refer to *Figure 1.11*:

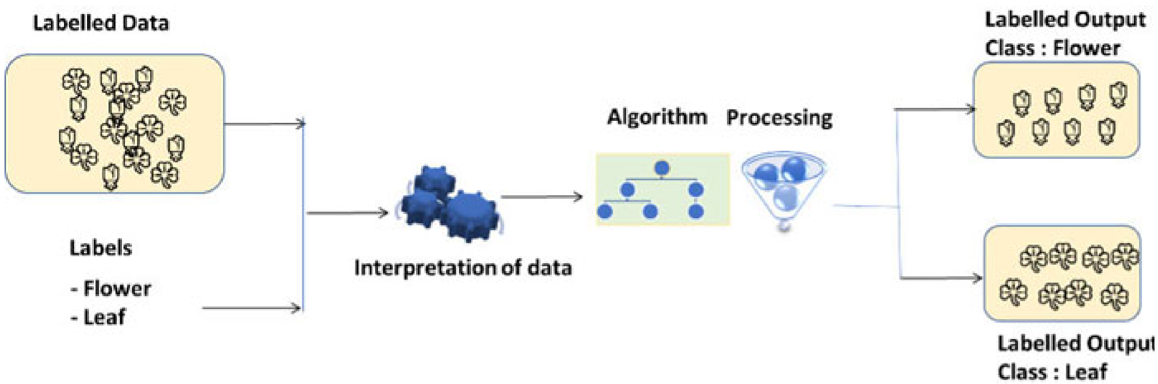


Figure 1.11: *Supervised Learning Model*

- **Unsupervised learning**

  Unsupervised learning uses hidden patterns or internal structures in the input data. Refer to *Figure 1.12*:

Figure 1.12: *Unsupervised Learning*

Example: Sorting flowers from leaves and forming two clusters. Refer to *Figure 1.13*:

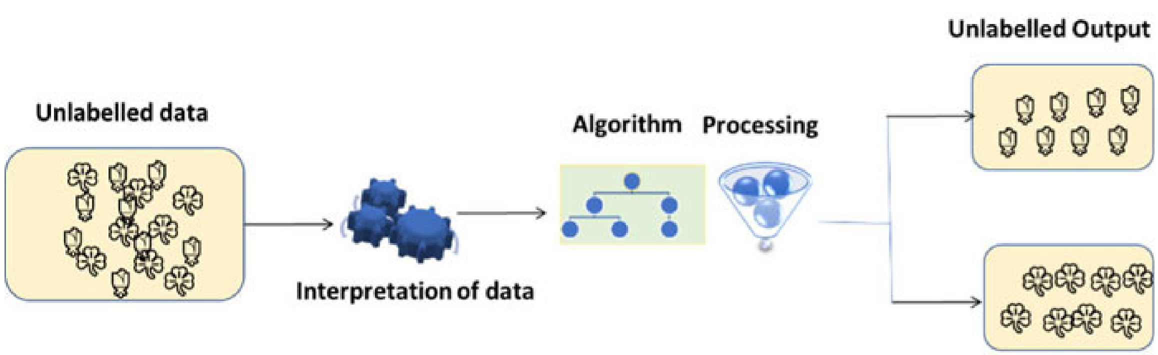


Figure 1.13: *Unsupervised Learning model*

- **Reinforcement learning**

Reinforcement learning is based on rewarding desired behaviors and/or punishing undesired ones. In other words, use a reward system to train the model. Refer to *Figure 1.14*.

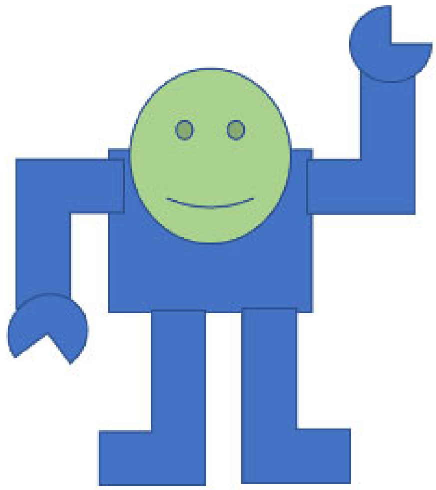


Figure 1.14: *Robot*

Example: A dog learning and unlearning actions and skills based on a reward mechanism. Refer to *Figure 1.15*:

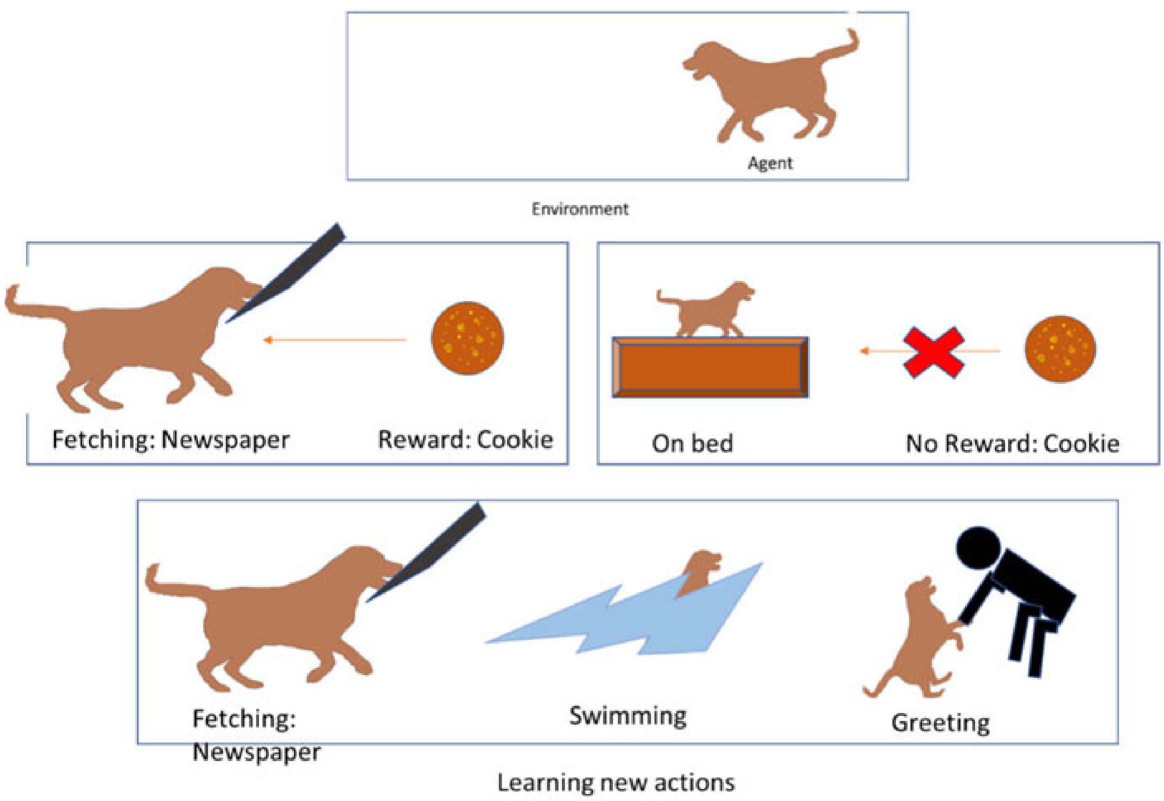


Figure 1.15: *Reinforcement Learning Model*

The following table highlights the major differences between the learning methodologies:

	Supervised learning	Unsupervised Learning	Reinforcement Learning
Input Data and methodology	Labeled data; Learn a pattern of inputs and their labels	Unlabeled data; Divide data into classes	No predefined data (all types of data); Works on interacting with the environment and maximizing the reward
Type of problems addressed	Regression and classification	Association and Clustering	Exploitation and Exploration
Supervision required	Extra supervision	No supervision	No supervision
Aim	Calculate outcomes	Discover underlying patterns	Learn a series of action
Solutions	Finds mapping equation on input data and its labels.	Classifies input data into classes by finding similar features	Maximizes reward by assessing the results
Model Building	The model is built and trained prior to testing.	The model is built and trained prior to testing.	The model is trained and tested simultaneously.

Table 1.1: *Difference between various models*

Neural networks

Neural networks refer to systems of neurons, either organic or artificial in nature. In regards to AI, it refers to a series of algorithms that aims at recognizing underlying relationships and patterns in a set of data through a process that imitates the way the human brain operates. At its heart, it is just multiplication and differentiation.

As such, neural networks can help systems make intelligent decisions with limited human supervision simply because they can learn and model the relationships between input and output data that are nonlinear and complex.

An Artificial Neural Network is made up of 3 components:

- Input layer
- Hidden (computation) layers
- Output layer

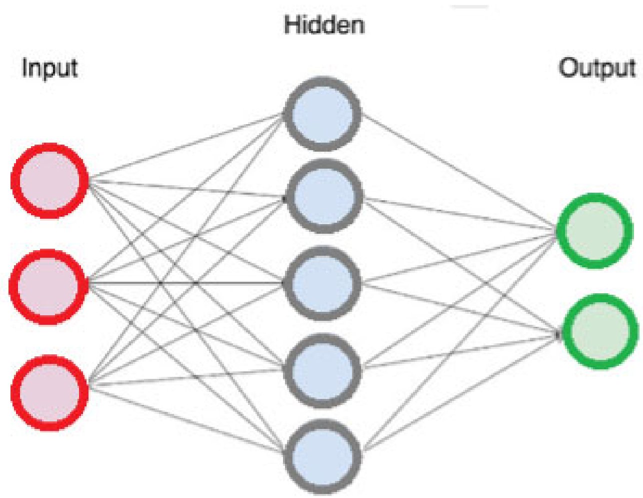


Figure 1.16: *ANN Artificial Neural Network*

In neural networks, learning happens in two steps:

- **Forward-Propagation**: Helps in making a guess about the answer. As the name suggests, the input data is fed in the forward direction, each hidden layer accepts the input data, processes it as per the activation function, and passes it to the successive layer.
- **Back-Propagation** It is the short form for "backward propagation of errors." Backpropagation is the process of tuning a neural network's weights (input is modeled using randomly selected weights) to better the prediction accuracy, minimizing the error between the actual answer and guessed answer.

What machine learning can and cannot do

As per Wikipedia, *Machine Learning is a branch of computer science that gives "computers the ability to learn without being explicitly programmed."*

Machine learning has algorithms that are fed data and these algorithms analyze the data to make predictions or recommendations. Such algorithms are coded by humans, in ways that these algorithms cannot learn. Machines only learn from the data that they receive and can analyze.

In a sense, rather than replacing the abilities of humans in the future, machines can make it easier to make complex computations fast enough to give conclusions that are stochastic, not deterministic.

Machine learning can do:

- Recognition

    - Image recognition
    - Text recognition
    - Voice recognition
- Recommendations
- Classification
- Text to Speech
- Predictive maintenance

Machine learning cannot do:

- Learning a language from hearing verbal utterances
- Human intention recognition
- Emotion recognition
- Gestures recognition
- Interact with and understand humans

Usually, machine learning algorithms require large amounts of data to be trained enough before they begin to give useful results.

Machine learning is not and will not be able to replace humans explicitly.

Key differences between artificial intelligence (AI) and machine learning (ML)

The following table highlights the major differences between AI and ML:

S.no.	Artificial Intelligence (AI)	Machine Learning (ML)
1.	Artificial intelligence is the ability of systems to acquire and apply knowledge mimicking human cognitive skills.	Machine Learning is a subset of Artificial intelligence and is about the ability of a machine to acquire knowledge or skill
2.	AI is a broader family consisting of ML and DL as its components.	ML is a subset of AI.
3.	AI aims to increase the chance of success and not accuracy.	ML aims to increase accuracy but is not programmed for success.
4.	AI targets simulate human intelligence to solve complex problems.	ML targets learning from input data to maximize the performance on related tasks.
5.	AI has a very wide variety of applications.	The scope of machine learning is constrained.

6.	AI is decision-making.	ML enables systems to learn new things from data and newer related data.
7.	It is developing a system that tries to simulate human intelligence to solve problems.	It involves creating self-learning algorithms.
8.	AI will go for finding the optimal solution that is focussed on success than accuracy	ML will go for a solution whether it is optimal or not, it is focused on accuracy than success
9.	AI leads to intelligence or wisdom.	ML leads to knowledge.
10.	AI can be categorized broadly into: Artificial Narrow Intelligence (ANI) Artificial General Intelligence (AGI) Artificial Super Intelligence (ASI)	ML can be categorized broadly into: Supervised Learning Unsupervised Learning Reinforcement Learning
11.	AI can work with structured, semi-structured, and unstructured data.	ML can work with only structured and semi-structured data.

Table 1.2: *Difference between Artificial Intelligence and Machine Learning*

Artificial Intelligence project life cycle

A project lifecycle describes the phases through which a project progresses. This sequence of the phases and their dependency is also clearly mentioned.

AI project life cycle mainly has 5 stages. These stages define the start to end of the development of AI-powered solutions in specific and clear steps:

The 5 stages of the AI Project Lifecycle are:

- Problem scoping
- Data acquisition
- Data exploration
- Modeling
- Evaluation.

Figure 1.17 describes in sequence the various stages of an AI project life cycle:

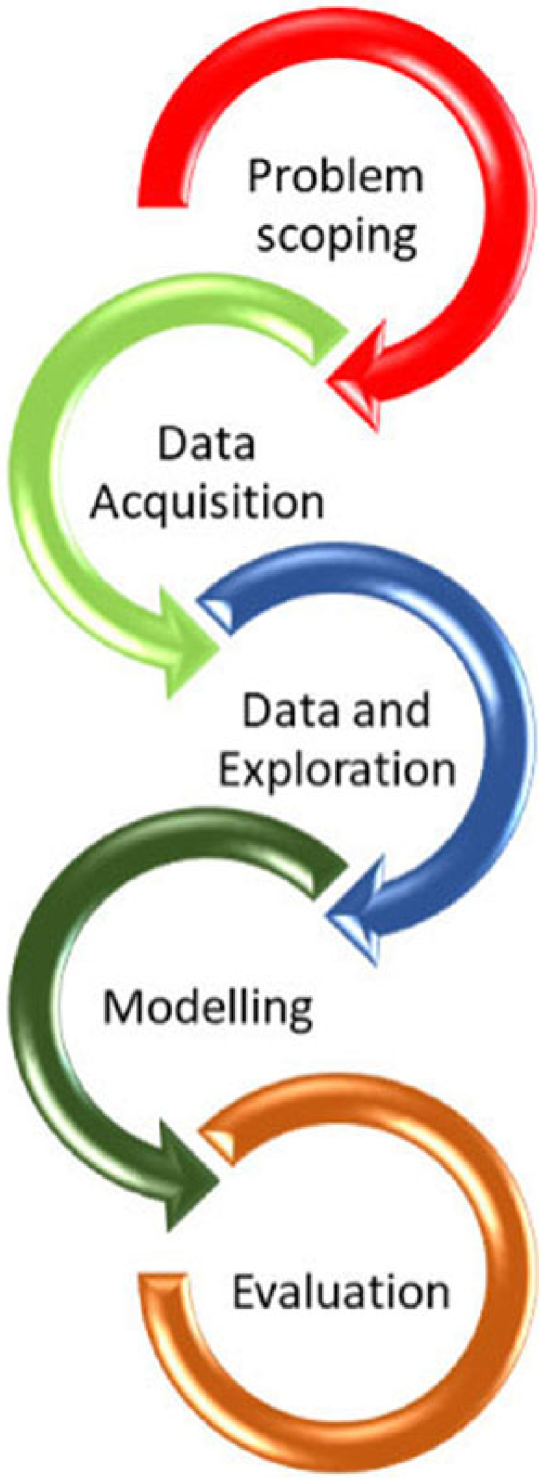


Figure 1.17: *AI Project Lifecycle*

Phase 1: Problem scoping

As the name suggests, this initial phase of the AI project lifecycle is all about understanding the problem, its scope, the boundaries, identifying the problem statement, various factors which affect the problem as well as all parameters and aspects that define the goal and the aim of the project. This scoping can be done by answering the 4Ws, which are:

Who: “Who” helps in identifying the stakeholders, categorizing all those who are directly or indirectly impacted by the problem.

What: “What” helps in understanding and identifying the nature of the problem while also collecting evidence to prove that the problem exists.

Where: “Where” helps in identifying the roots of the problem, where it arises, the situation, and the location it arises.

Why: “Why” helps with why the problem is worth solving.

Phase 2: Data acquisition

Data acquisition is the process of collecting accurate and reliable data that cumulatively cover variables and attributes of the problem statement in its entirety. Data can be in the format of text, video, images, audio, and so on and it can be collected from various sources like interest, journals, newspapers, and so on. That is data can be structured or unstructured or in any non-specific form.

Data can be collected from various sources:

- Databases
- Web pages
- Devices like cameras and sensors (e.g. in Autopilots, weather predictions)
- Public surveys /records of purchases, transactions, registrations, and more

Phase 3: Data exploration

Data exploration is the process of performing operations like data cleaning, finding missing values, removing useless data, and basic statistical analysis for arranging the data, gathered as in phase 2, uniformly and meaningfully.

Data can be arranged in the form of a table, plotting a chart, or making a database. Multiple visualization tools are available in the market for offering data in visibly grasping formats.

A few of the tools to use for data exploration are:

- Google charts
- Fusion charts

- Tableau
- High charts

Phase 4: Modeling

Modeling is the phase of the AI project lifecycle in which different models based on the visualized data can be created and developed. Models help in formulating mathematical relations between data and the outcome. These models can also be checked for their advantages and disadvantages Refer to *Figure 1.18*:

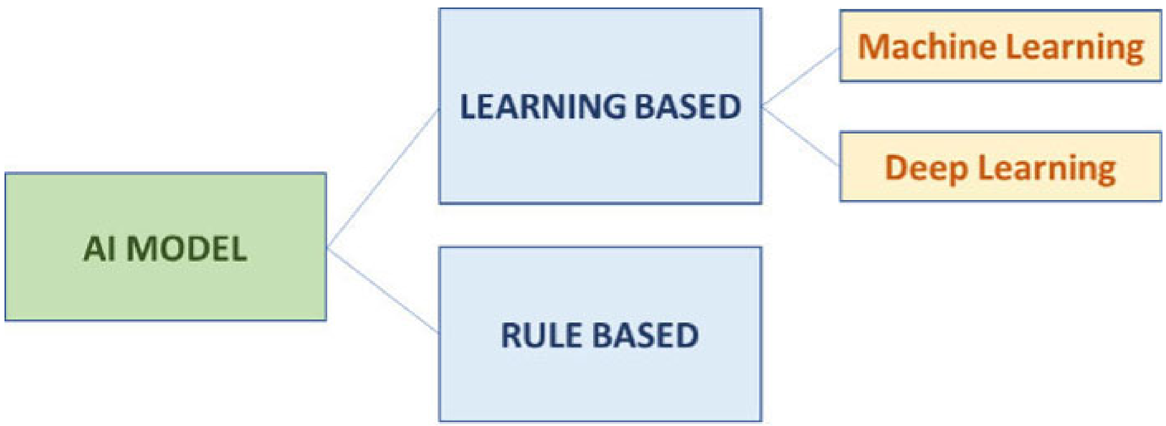


Figure 1.18: *AI Modelling*

While we have covered learning-based models in previous sections, a rule-based model is where the relationship or patterns in data are defined by the developer. The machine performs tasks according to the information and these rules as given by the developer.

As an example, for rule-based learning, a bank grants a loan to a customer based on certain rules that measure the personal and financial information against a set of levels that help decide if a loan is to be granted.

Phase 5: Evaluation

Evaluation is the last phase in the AI Project lifecycle. Here the model developed is fed with input data and the outcome is compared against expected outcomes. This stage determines the reliability of the model and the completeness of the data fed into the model.

Career opportunities in artificial intelligence

The Artificial Intelligence field is vast to bring multiple career opportunities. These are primarily based on data, algorithms and machine learning, and application development.

The following table lists the most common career opportunities in the field of artificial intelligence:

S.No	Role Name	Qualification and strengths	Role Description
1	Big Data Engineer	• Knowledge of Big Data and various database systems. • Programming Languages like Python, R, Java, and so on.	• Build and administer the organization's big data. • Prepares, manages, and establishes big data environment. • Apply data concepts like migration, visualization, and mining.
2	Business Intelligence Developer	• Bachelor's degree in Computers and Mathematics. • Know Programming languages. • Know Data sets. • Problem-solving abilities. • Analytical capabilities. • Databases and Data warehouse.	• Create business models. • Analyze data sets to identify business trends. • Helps with brand recognition and awareness. • Prepare, develop, and operate business intelligence solutions.
3	Data Scientist	• Master's degree in Computers and Mathematics. • Know Programming languages. • Know Data sets. • Problem-solving abilities. • Analytical capabilities • Databases and Data warehouse.	• Manages and operates large datasets from different sources. • Tracks data collection methods and adds new data sources. • Utilizes data for business outcomes by making predictions.
4	Machine Learning Engineer	• Software engineer. • Know Programming languages. • Know Data sets.	• Combine software engineering and data science. • Applies algorithms to real-world applications. • Operates algorithms to create open-source libraries. • Harness big data techniques and computing systems/predictive models. • Make raw data meaningful. • Put machine learning solutions into production (means operational). • Experiments with programming languages including ML libraries. • Ensure data flow across between databases and deployed systems.
5	Research Scientist	• Specialized Ph.D. or advanced	• Highlight the theoretical side of

		master in Computers or Mathematics.
		• Gained expertise in
		  ◦ Statistics.
		  ◦ Applied mathematics.
		  ◦ Applications related to machine learning and intelligence.
		• Significant knowledge of NLP and Reinforcement learning.
			solutions.
			• Creates new networks.
			• Discover new ML approaches.
			• Devise Novel algorithms.

Table 1.3

Conclusion

AI is already a part of our lives. With the ease and automation it brings, AI is being a focus area in all fields of life, be it healthcare, manufacturing, or performing routine daily tasks. It is going to get evolved more and it is not too late to be on board the artificial intelligence wagon.

In the next chapter, we will be discussing key fields of applications in AI and methodologies and their impact on our society. At the end of the next chapter, we will be analyzing how we get ready for the future which will be the AI age.

Multiple choice questions

1. **What is part of Modelling in terms of AI:**
    a. Data visualization
    b. Data cleaning
    c. Filtering of useless data,
    d. All of above
    e. None of above
2. **Which is not part of the data acquisition**
    a. Sensor
    b. Web scraping
    c. Data cleansing
    d. Survey
3. **What is not a type of learning**

a. Supervised
b. Unsupervised
c. Reinforcement
d. Unstructured

Answers

1. **d**
2. **c**
3. **d**

Questions

1. What is the main difference between ML and Deep Learning?
2. How is AI different from ML?
3. What are the types of AI Models?
4. Use your imagination to give a feature set of an AI system that you would want to use in your daily life in the near future. It should not be existence as of now.

Join our book's Discord space

Join the book's Discord Workspace for Latest updates, Offers, Tech happenings around the world, New Release and Sessions with the Authors:

https://discord.bpbonline.com

CHAPTER 2
AI Applications and Methodologies

Introduction

In this chapter, we will be learning about the key fields of application where AI can be applied. These are the applications that are having an impact on society, both positive and negative. We will be navigating through such applications we all may have used at some point in time or have become part of our daily lives.

We will also learn about the machine-learning techniques behind these applications. These applications have the capacity to bring about global societal changes. It is therefore very necessary to get ready for the future, the AI age.

Structure

In this chapter, we will be discussing:

- Key fields of application in AI
    - Chatbots (Natural Language Processing, speech)
    - Alexa, Siri, and others
    - Computer vision
    - Weather predictions
    - Price forecast for commodities
    - Self-driving cars
- Characteristics and types of AI
    - Data-driven
    - Autonomous systems
    - Recommender systems
    - Human-like

- Cognitive computing (Perception, Learning, Reasoning)
- Deep dive into NLP, CV, and much more
- AI and society
- The future with AI, and AI in action
- Non-technical explanation of deep learning

Key fields of application in AI

In this section, we will be discussing applications like chatbots, text-based voice assistants, computer vision, autonomous systems like self-driving cars, and others that are inherently using AI technologies.

Chatbots (Natural Language Processing, speech)

Chatbots are very common text-based applications used in various ways to reach out and interact with people. Let us take an understanding of chatbots and the technology these are based on.

What are chatbots

As the name implies, a "Chatbot" is a conversational robot. AI powered Chatbots can grasp and understand natural languages or multiple human languages like English, Dutch, Hindi, French, German, and so on. These chatbots can, therefore, respond to people online using the "live chat" feature on webpages or portals, or applications.

AI powered chatbots have the ability to mimic human-like conversations without the need for any human intervention at their end, enacting a real-life agent.

These chatbots are based on machine learning and build a database of answers, eventually pulling the relevant info before pushing it to the user at the other end.

Chatbot systems can communicate via written or voice messages. The talking robots are usually termed "*voice bots*".

Benefits of chatbots

The chatbots are used to provide customer support, answer inquiries, issue tracking, or any other contextual support enhancing customer engagement and satisfaction. These chatbots can be replicated across industry verticles integrated with various informational, customer, and business offerings centric databases.

Since these chatbots are automated, it provides immense value to both the users and the organizations/businesses hosting them.

The following *Figure 2.1* illustrates some benefits of chatbots in organizations:

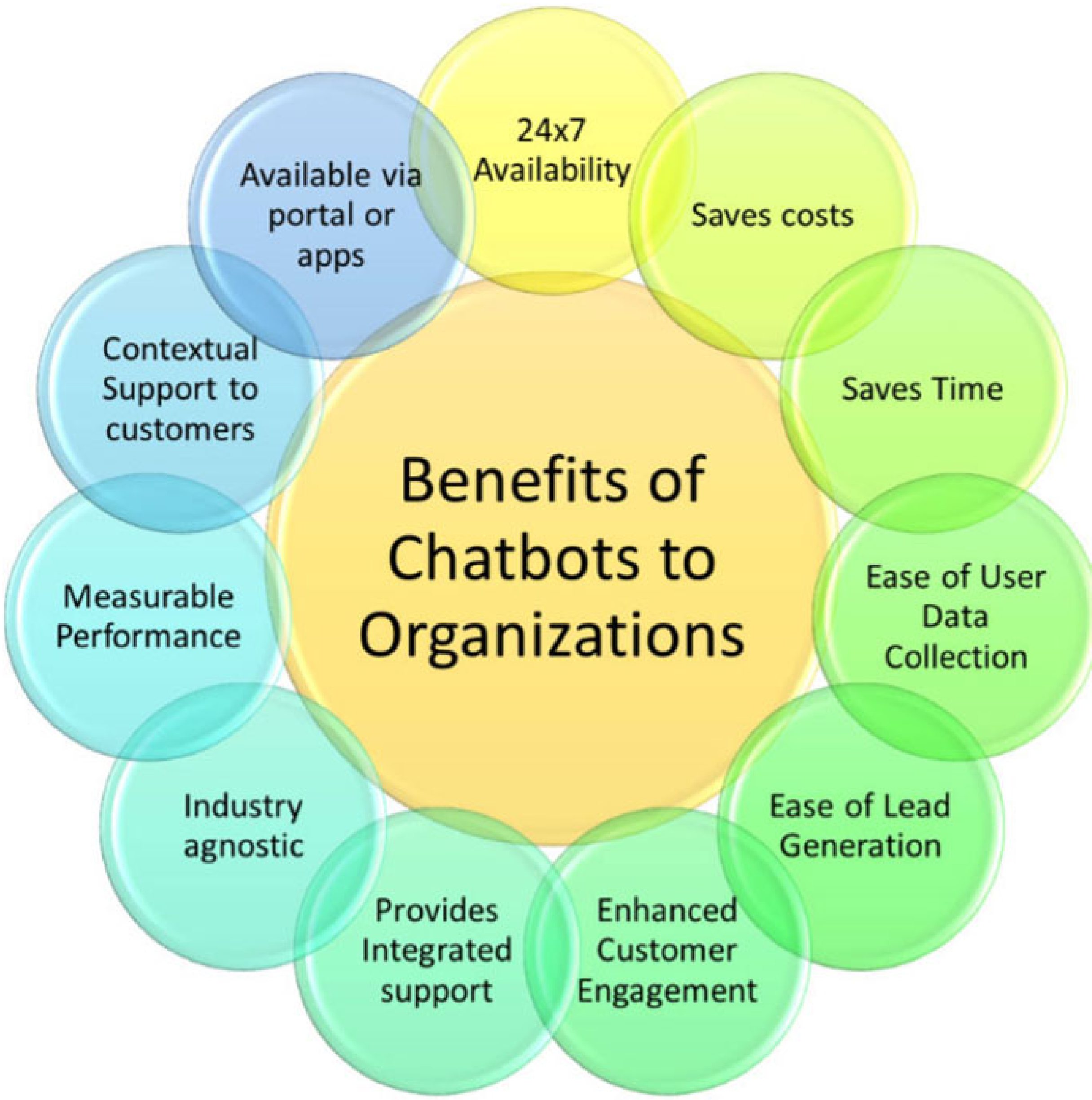


Figure 2.1: Benefits to the organizations

- **24x7 availability**: A chatbot is available along with the portal/application where it is hosted, enabling organizations to reach out to intended users all the time.
- **Availability on digital platforms**: Chatbots can be accessed via digital devices from anywhere at any hour, again easing out organizations on complete human dependency at odd hours.
- **Saves Times**: These chatbots help in saving time by providing prompt replies to the users, thus catering to multiple of them in parallel.
- **Ease of Users' data collection**: The Organization gathers meaningful huge customer data sets in less time. This data helps.
- **Ease of lead generation**: Based on customer likes and inclinations, chatbots intelligently recommend products that help increase lead generation (identifying and cultivating potential customers for products or services offerings.).
- **Enhanced customer engagement**: The ease of use of chatbots, prompt support, consistent replies, personalized experience, and most of the previously mentioned benefits allows the customer to seek more info and be ready to provide feedback. This enhanced customer engagement can benefit any organization in achieving the intended purpose of these chatbots.
- **Provides integrated support**: Chatbots pull answers from databases which could be product/service offerings, customer info, ticket status, and so on. The customer, thereby, gets one channel for all info that is sought.
- **Industry agnostic**: The chatbots can be deployed across industries as all it needs is the right set of data to pull info from.
- **Measurable performance**: The chatbots can be upgraded as the performance is measurable.
- **Contextual support to customers**: With access to databases, chatbots can provide customers with accurate answers and information as desired.

The following *Figure 2.2* illustrates some benefits of Chatbots to users/customers:

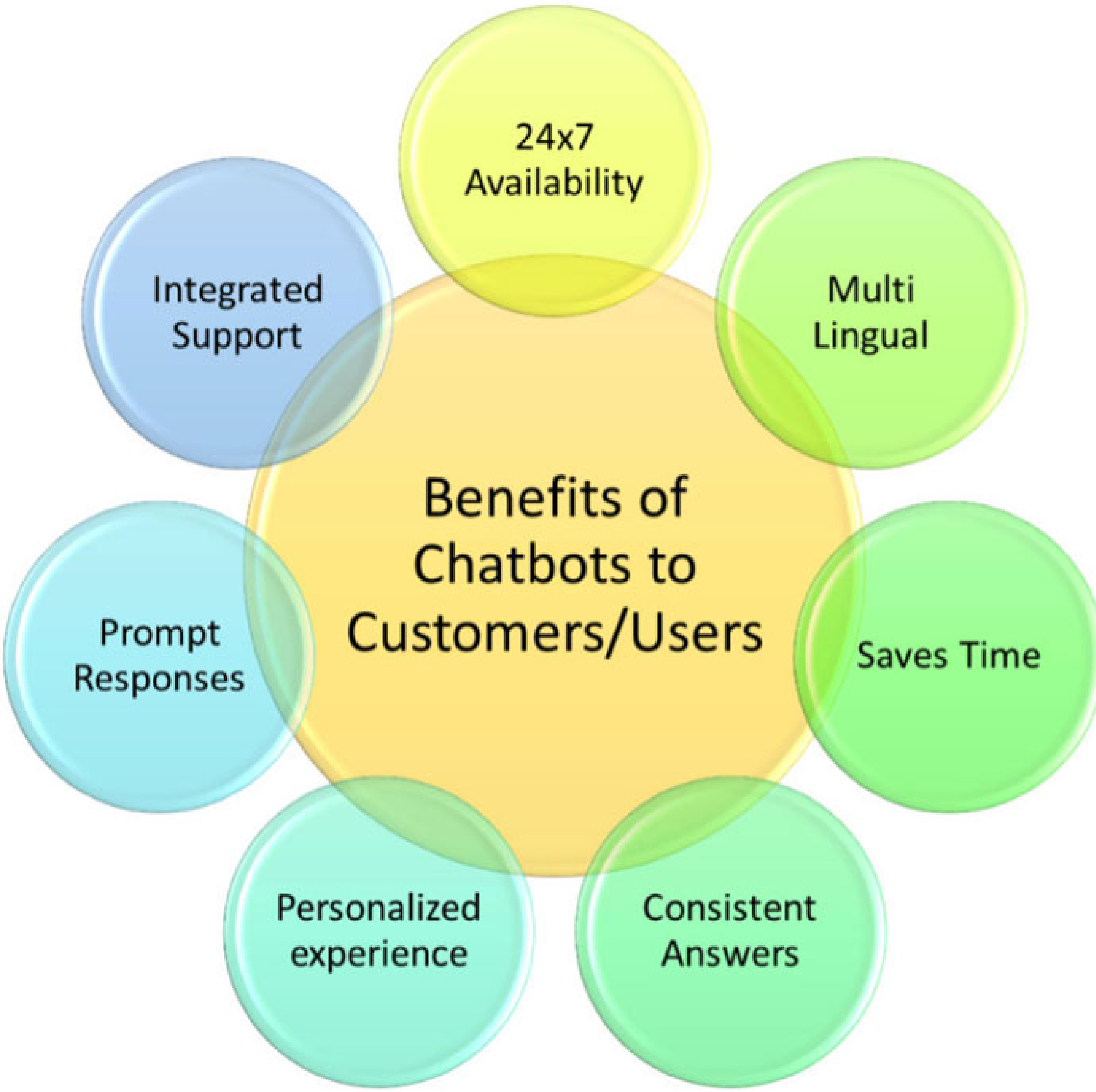


Figure 2.2: *Benefits to the users*

- **24x7 availability**: Ease of Access on digital devices from portals and apps that are always ON.
- **Multi-lingual support**: Current chatbots are designed to provide multi-lingual support, enhancing customer support by removing language barriers irrespective of the region customer is based in.
- **Saves time**: With prompt and accurate responses and consistency in answers, customer time is saved considerably, and the customer is prompted to seek more info and provide valuable feedback.

- **Consistent answers**: Receiving consistency in replies to their queries on chatbots is an added benefit to the customer experience while using chatbots.
- **Personalized experience**: By pulling data from various sources and having access to the customer profile, chatbots are capable of providing a personalized experience to the users.
- **Prompt responses**: Getting immediate responses with next to no waiting period or being put on hold saves time for the users.
- **Integrated and contextual support**: Having access to data, chatbots can provide accurate and personalized answers to complex user queries.

Chatbots in key industries

Chatbots can benefit various industries irrespective of the business. Chatbots can provide support in, but not limited to, marketing, sales, customer support, or IT service helpdesk. These can also be used for scheduling appointments, reviews, feedback, and much more. Let’s have a look at a handful of key industries where chatbots can revolutionize user/customer experience and ease of doing business.

Figure 2.3 lists these industries:

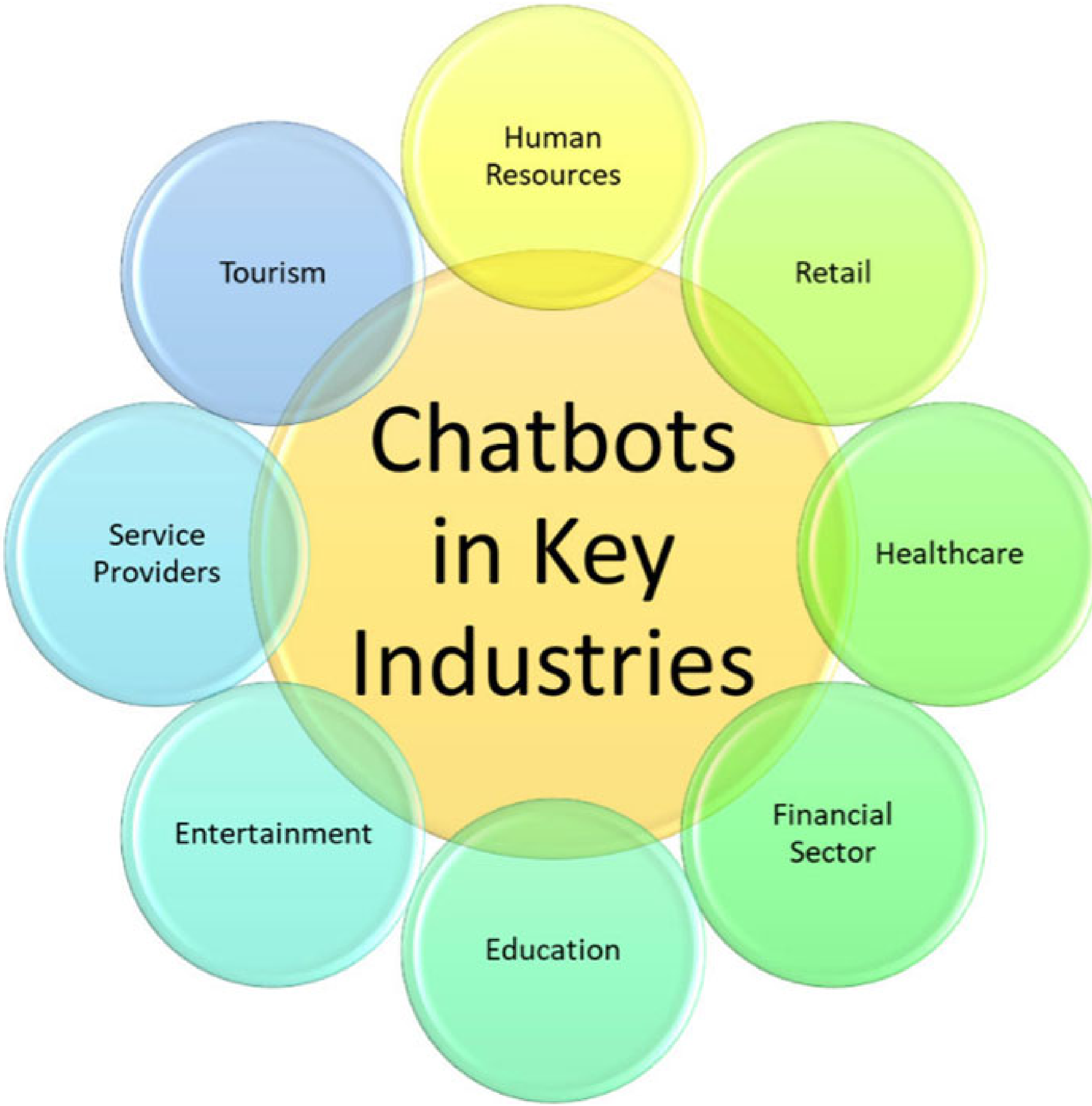


***Figure 2.3:** Chatbots in Industries*

- **Human resources**: Chatbots can help in the recruitment process right from searching for candidates and evaluating their skills as per the job description.
- **Retail**: Chatbots in retail or e-commerce can help answer customer queries, recommend products and services, place an order, make payments, and streamline overall sales and marketing processes.
- **Healthcare**: Chatbots can help in scheduling appointments, setting reminders, recommending health checkups, providing medical

assistance, and more features based on underlying algorithms and design.

- **Financial sector**: Chatbots can assist in making financial transactions, reporting card losses in banking, opening accounts, answering customer queries related to product offerings and services, recommending desired loans, and so on.
- **Education**: Chatbots can help create a personalized learning environment for students. Students' responses can be analyzed to assist them further with new learning material or courses via chatbots.
- **Entertainment**: Chatbots can help share available movies, book tickets, recommend movies based on previous feedback on movies watched, and much more.
- **Service providers**: Service providers like **direct-to-home** (**DTH**) can use chatbots to help users navigate their portals, recommend new products, and help with billing and technical customer support.
- **Travel and tourism**: Chatbots play a major role in this sector. These can provide weather information, and law order situations and assist with bookings and packages being offered.

Machine learning in AI-powered chatbots

AI powered chatbots use machine learning (supervised or deep or ANN) to be trained to provide human-agent-like responses to the users. These chatbots use **Natural language processing** (**NLP**), covered later in the chapter, to understand human language.

Chatbots based on these advanced technologies identify human conversation patterns and learn and remember data efficiently, compared to human agents.

Supervised training to test chatbot algorithm

These chatbots are trained using predefined responses, based on the responses and the learnings; thereby, these chatbots may learn rude responses that can hamper brands' reputations. For example, the chatbots may pick up racist, sexist, or abusive remarks.

This possibility of providing incorrect answers raises the need for human operators to rectify the mistakes. These may need to be upgraded for consistency in answers in the future. **Human-in-the-loop (HITL)** becomes necessary to regularly update and train the chatbot.

The role of human agents then becomes overlooking chatbots' conversations and answering the question that chatbots are unable to handle.

There is regular updating of the algorithms of this chatbot as well as training of these chatbots.

Generative chatbots – Deep learning

Generative chatbots are based on deep learning. These advanced chatbots can answer complex queries of users.

Deep learning technology allows chatbots to understand users' questions with the help of neural networks, at times from famous movies and books. Such chatbots get smarter with every conversation and can imitate real people.

These chatbots can also act like voice bots by their ability to understand voice commands and recognize speech.

Artificial neural networks to replicate a human brain – Intelligent chatbot

Intelligent chatbots based on Artificial neural networks can replicate human brains. It can easily learn the new intent of the customer and engage in meaningful conversations with the users/customers.

Natural Language Processing (NLP) – Natural conversation

Natural language processing is one subset of Artificial Intelligence technology that enables computer programs to process or understand human language – both written and spoken in the same ways as humans do.

NLP combines computational linguistics (that is, analysis and synthesis of language and speech by applying computer science techniques) with statistical, machine learning, and deep learning models.

NLP can help computer programs translate text from one language to another.

As such NLP in Artificial Intelligence technology helps chatbots to converse like a human. Chatbots based on NLP can easily understand users' intent and their purchasing intent.

Natural language understanding (NLU) – Complex questions

Chatbots process the input information as human text or speech through NLP and understand these human interactions through NLU.

NLU enables computer programs to communicate back to humans in their own languages. As such, after processing the human conversation through NLP, Natural language understanding is used to understand the structure of the conversation and converse with the customers. To interpret human texts or speech, NLU breaks complex sentences into simpler ones.

NLU uses algorithms to reduce human speech or text into a structured ontology -- a data model consisting of semantics and pragmatics definitions.

Recognition – both intent and entity – are two fundamental concepts of NLU.

Intent recognition is the first and most important part of NLU. This process is used to identify the user's sentiment in input text or speech and determine their objective. This help establishes the meaning of the input.

Entity recognition, as the name suggests, is a specific type of NLU that focuses on identifying the entities in the input. It then extracts important information about those entities. The entities are of two types: named and numeric. Named entities can be categorized such as people, companies, and locations. Numeric entities can be numbers, currencies, and percentages.

Voice assistants (Alexa, Siri, and others)

Voice assistants or digital assistants or virtual assistants, or AI assistants are software programs that carry out tasks via voice commands as input. Voice assistants are based on AI and ML technologies that help recognize voice inputs, and convert voice to digital data efficiently for software to analyze and perform desired tasks in accordance.

Benefits

Voice assistants are used to accomplishing everyday tasks. Say, for example, they can answer queries, integrate with smart homes to turn on and off lights and electrical appliances, and more. These voice-activated assistants come preinstalled on smartphones. Voice-activated speakers are becoming common platforms in homes and workplaces.

Having these assistants can be a great help to perform tasks on a mere voice command, as seen in the following scenarios:

Stay updated about current and trending: At a voice command, get all kinds of information available on the Internet, as well as seek info about the weather, current affairs, news, traffic, and more.

- **Music**: Command voice assistants to play music matching certain moods or command a playing a particular song.
- **Devices' control**: Control electrical and electronic devices and other smart appliances by voice commands like "switch on the geyser", "switch off lights", and "set AC temperature".
- **Banking assistance**: Voice assistants can help check balances, order banking products, and get transaction details for those banks that allow access to such services via AI assistance.
- **Organize the day**: Voice assistants can be used to set reminders for the day's schedule by accessing the calendar.
- **Support for differently abled or dependent folks**: Voice commands to operate devices, and voice search function support visually-impaired, elderly, and dependent persons in their routine reducing the dependency on human intervention.

Pitfalls

Every technology has two sides. With the ease voice assistants bring to perform our routine tasks, they also do have a few downsides:

- **Privacy at stake**: With the machine learning methodologies at their backend, these voice assistants are continuously taking voice inputs and learning. Hence, they learn a lot more than desired.
- **Sharing personal data**: Voice commands for performing various tasks cannot be achieved without giving access to personal data to these voice assistants, which can be challenging in case these assistants are hacked into. In case smart homes are hacked into, imagine the damage it can cause with personal data being exposed.
- **Familiarity with lifestyle**: Voice assistants do learn the arrival timings, the day's schedule, and calls made along with bank details (if provided access). They also record habits like preferred room temperature,

songs, usual moods, and much more. In some cases, it may also record food and shopping preferences.

- **Identity theft**: With all the personal data recorded, if any hacker is able to hack this voice assistance, it will be more like an identity theft than a mere financial or trespassing hit.

Examples

Voice assistant systems can be found preinstalled on smart speakers, smartwatches, mobile phones, tablets, and other digital devices.

A few of the known names in households are:

- Alexa (Amazon)
- Siri (Apple)
- Google Assistant (Google)
- Bixby (Samsung)
- Cortana (Microsoft)

Computer vision

As the name suggests, computer vision may be considered the ability of a computer to see. Computer vision is based on **artificial intelligence** (**AI**) technology and gives computer systems the ability to obtain meaningful data from visual inputs such as photos and videos.

Again, like other AI-based systems, the insights gained from computer vision are used to perform desirable automated actions.

Currently, deep learning techniques are commonly used for computer vision. Computer vision-based systems acquire, process, analyze (based on specified criteria), and render visual information thereafter in various formats such as 3D models, images, videos, or related volumetric data.

Weather predictions

Climate change is increasing the intensity and frequency of extreme weather events that are being recorded across the globe.

Older weather forecasting systems relied on supercomputers needed to process large amounts of data gathered from across the globe such as temperature, pressure, humidity, and wind speed.

This came with the need to create a computationally efficient model, capable of accurately predicting upcoming weather. Scientists called such a system **Deep Learning Weather Prediction** (**DLWP**). The DLWP is based on an AI algorithm. The system learns and recognizes patterns in historical weather data based on global grids.

AI-Powered weather forecasting system can help identify accurately potential extreme weather 2–6 weeks into the future or in months. Extreme weather prediction is made accurately and well in advance, giving communities and critical sectors such as public health, water management, energy, and agriculture to mitigate potential disasters.

Price forecast for commodities

Commodities play a crucial role in the global economy and are one of the major motivators of inflation and economic activities. A significant percentage of the market is composed of just the oil industry and natural gas industries. Other significant contributors include metal, mineral, and agricultural commodities.

Price forecast for commodities refers to the process of making forecasts against products' prices based on previous and present data or trends analysis.

AI-based systems can help mitigate risks of investment and returns thereby. These systems can be fed with data pertaining to historical price trends, factors that drive the prices, volatility in price variation, and the impact of various seasons. AI-based price forecasting systems thus eliminate human emotions and gut feelings in their predictions.

Based on the commodity or service, price prediction AI-powered systems use algorithms to analyze it based on the characteristics, demand, and market trends. The system then sets a price it predicts that attracts customers and maximizes sales.

Self-driving cars

One of the characteristics of AI-based autonomous systems is that they function on their own while adapting and responding to dynamic environments and changing situations.

A common application of autonomous computer systems is in the transportation industry, especially, consumer vehicles that use automated and assistive technology but are not fully autonomous in nature as of date. Where autonomous is defined as “having the freedom to govern itself or control its own actions”

A self-driving car, also known as a driverless car, or robotic car incorporates vehicular automation AI-powered system. In sense, a self-driving car is capable of sensing its environment and moves safely with little or no human intervention.

The surroundings are perceived using a set of sensors like:

- **Thermographic cameras**: A thermographic camera is a device that creates an image using **infrared (IR)** radiation
- **Radar**: Radar sensors transform microwave echo signals into electrical signals. They detect motion by perceiving an object’s position, shape, motion characteristics, and motion trajectory using wireless sensing technology.
- **LiDAR**: **Light Detection and Ranging** (**LiDAR**) sensors in autonomous vehicles provide a high-resolution 3D view of the surroundings.
- **Sonar**: Sonar or an ultrasonic sensor is an electronic device that measures the distance of a target object by emitting ultrasonic sound waves. The reflected sound is then converted into an electrical signal.
- **GPS**: Global positioning system sensors use a satellite-based navigation system in orbit around the earth to provide position, velocity, and timing information.
- **Odometry**: Odometry is the use of data from motion sensors to estimate the change in position over time, relative to an earlier known position.
- **Inertial sensors**: An inertial sensor is used to gather the acceleration and angular velocity of an object.

The environment for a self-driving car includes obstacles, paths, relevant traffic signals, signages, speed limits, traffic diversion, other vehicles, distance from moving objects, and much more.

Artificial intelligence-powered control systems are deployed in these robot-cars that learn all the gathered sensory information in order to control the vehicle and support various autonomous-driving tasks and operations.

As a key future technology, AI-powered self-driving cars are predicted to have a comprehensive impact on various industries including, but not limited to, automobile, health, welfare, urban planning, traffic, insurance, services, and much more.

Connected, Autonomous, Shared, and Electric (**CASE**) Mobility is a future mobility vision based on self-driving cars combined with other emerging automotive technologies such as electric vehicles, connected vehicles, and shared mobility.

SAE International formerly named the **Society of Automotive Engineers**, is a United States-based, globally active professional association and standards-developing organization for engineering professionals in various industries (as per Wikipedia).

SAE International has developed a framework that categorizes different levels of autonomy in vehicles. The levels in the framework are not fixed, for instance, some break their hierarchy into five layers and some into six. The distinctions between various levels are also not firm and some algorithms may exhibit behavior from two or three levels at the same time.

These levels can be described as:

- **Level 0**: No automation; the human does all operations except some automatic systems like windshield wipers or heating.
- **Level 1**: Hands-on/shared control; some operations like braking or lane following are delegated to the car.
- **Level 2**: Hands off; The car does all major operations like braking, acceleration, or lane following; however, the human must remain alert to take control at all times, maybe by also having to have hands on the steering wheel all the time.
- **Level 3**: Eyes off; The human may not remain alert all the time and may occasionally turn attention away from the road for a short amount of time, however, must be ready to respond to an alarm in case needed. The car is in self-control over mapped routes but not on paths that are new and not mapped in advance.

- **Level 4**: Mind off, this level may require humans to take control in cases where the paths aren't well understood by the AI.
- **Level 5**: Steering wheel optional. This level is more like a cab service that humans need not control.

Well-known automobile companies that have implemented predictive analytics in their autonomous vehicles are:

- Tesla
- Ford
- Volkswagen

Characteristics and types of AI

AI is a technology that enables systems to develop cognitive skills and think and behave like humans. The main components of AI are:

- **Feature engineering**: This primarily means the feature extraction process of identifying a nominal set of attributes from given data sets.
- Artificial Neural Networks
- Deep Learning

Other characteristics that utilize the maximum efficiency of this technology are:

- **Natural Language Processing**
- **Intelligent Robotics**: Robotics is the amalgamation of engineering, science, and technology that produces programmable machines known as robots that can mimic human actions or assist humans. AI powered robots are systems that can learn and adapt to perform independently without human intervention.
- **Perception**: Take inputs from sensors and process the collected data to give the desired output. Applications based on these processes are facial recognition, computer vision, and much more.
- **Automate Simple and Repetitive tasks**: As the heading suggests, this is for performing monotonous and repetitive tasks that may involve large data sets as input or voice assistant systems performing daily routine tasks.

- **Data Ingestion**: The process of saving unstructured data extracted from various sources to a huge database medium for accessing, analyzing, and preparing AI models.
- **Imitation of Human Cognition**: An example of such capabilities is Chatbots that can imitate human-like conversations by intent and entity recognition.
- **Quantum computing**: AI has helped by solving complex quantum physics. Quantum computing focuses on developing complex algorithms of quantum for revolutionizing and advancing computational tasks.
- **Cloud computing**: Introducing AI capabilities in cloud computing can assist organizations in addressing the never-ending growing data ingested in their data centers and systems.
- **Ethical gene editing**: AI has the potential to contribute successfully in the medical field, especially in the treatment of common complex diseases or disorders caused by gene mutations.
- **Intelligent disaster response**: With advancements in technology, Modern rescue systems utilize sensor-based systems, unmanned ariel vehicles, or AI-powered robots to collect accurate information about the location of victims or extent of damage, or a forecast of upcoming disasters.

Data-driven

Consumer-centric businesses are usually driven by data-capturing customer profiles, likes, dislikes, purchasing power, frequency of purchases, and unending lists that can help improve offerings and revenue growth.

Next-generation data-driven AI is a game changer for such businesses.

Let us analyze the most common of the customer-centric businesses we come across on a day-to-day basis.

- **Groceries**: Forecast customer needs and ensures product availability based on location and purchasing preferences**.**
- **Restaurant**: Analyze customer footfall and stock up on resources accordingly. Also, analyze preferred dishes during the time of the year minimizing wastage and achieving customer satisfaction too.

- **Hotel**: Analyze footfall in various seasons and plan the resources accordingly. The data can also help offer optimal pricing and offer relevant packages to maintain business throughout the year.
- **Retail**: Analyze discounts to be offered in various seasons that cause a positive impact on sales. Ensure the demand and supply are well balanced and also identify cross-selling opportunities.
- **E-commerce**: Analyze procurement trends like items bought together and build pricing models. Forecast product sales depending on its pricing and features and targeted customers based on similar products.

Autonomous systems

Current AI-powered systems are automated but not entirely autonomous. Autonomous artificial intelligence systems will be a reality when robots, cars, planes, and other devices are able to execute extended sequences of operations without guidance and intervention from humans.

To achieve and build autonomous systems, researchers are continually refining their algorithms and also their approach. To have a better understanding, the entire job is divided into layers as mentioned follows:

- **Sensing**: Incorporate sensors that collect all possible data that can impact the outcome. We had seen the sensors used in self-driving cars earlier in the chapter.
- **Fusion**: The collected data must be holistically analyzed to get a complete view of the ecosystem so as to build a model that enables the system to take the most accurate or human-like decisions independently.
- **Perception**: After the model is constructed, the system must be able to identify and perceive the ecosystem completely.
- **Planning**: Having perceived the environment, the next expected step is to plan the best possible action to be taken.
- **Control**: The action must be executed in a controlled manner and independently

Recommender systems

As the name suggests, a recommender system or a recommendation system provides suggestions or / recommends items most relevant or applicable to a particular user.

The suggested items could refer to products, services or even music to be played, or news to be heard. These systems are best used when the list available to choose from is huge and overwhelming. Recommender systems help in decision–making processes.

Human-like

Scientists and engineers continue to push the boundaries of what can be achieved with AI. But we are short of the machines that we witness in sci-fi movies. Current systems do not understand how the world works.

AI research scientists are still figuring out how to make the paradigm shift from data-driven to more intuitive human-like thinking systems displaying cognitive skills.

However, these machines will be learning from watching humans achieve human-like capabilities.

Things to ponder: We have subtle gender bias displayed by humans in almost all walks of life and even at-home setups. Will the AI-powered human systems pick up the same learning and behavior? These are abstract behavioral aspects that cannot be programmed as right or wrong via data sets.

Cognitive computing

Cognitive computing is an attempt to have computers imitate the working of a human brain.

Cognitive computing refers to the use of artificial intelligence and underlying technologies like language processing, and machine learning to the top with human capabilities that help regular computing better solve problems and analyze data to tackle complex decision-making processes.

The characteristics of cognitive computing are:

- **Perception**: Five senses: sight, taste, smell, sound, and touch contribute towards forming a human perception. Perceptions are a cognitive process as we often consciously and unconsciously process information gained through our senses, forming thoughts, and opinions and giving emotional reactions.

- **Learning**: Learning can be achieved through many cognitive processes, such as memory, thought, and perception. To learn quickly and retain information, multiple processes need to be combined. For example, reading writing, listening, verbal communication, and thinking can help learn things faster.
- **Reasoning**: It is the ability to analyze and perceive any given information from various perspectives by breaking it down and structuring it in a logical order.

Deep-dive into NLP, CV, and much more

Computer machines understand bits, which could take a value of zero or one. **Natural Language Processing** (**NLP**) comes in handy as a processing tool for computers to understand input words, languages, and sentences. NLP includes various language preprocessing techniques. These include the following techniques:

- **Tokenization**: Involves breaking down sentences into smaller units, that is, words.
- **Stemming**: Involves extracting the core context from words by cutting the suffixes of words. For example, worries → worri.
- **Lemmatization**: Similar to stemming but involves reducing the word to its root. For example, worries → worry.
- **Bag of Words**: Involves representing the significant words in the collection in the form of vectors.
- **TF-IDF**: This is similar to a bag of words and adds higher weightage to significant words and vice versa.

The output of the above techniques is prepared data which is then passed on to complex models to generate meaningful outputs such as sentiment analysis, word prediction or fully conversing chatbots.

While NLP deals with textual data, **Computer Vision** (**CV**) involves processing images by computer and extracting useful information or performing meaningful tasks on those images.

A computer views images as pixels and channels (red, green, blue). The combinations of these channels form the basis of Computer Vision. Few of the Computer Vision techniques as mentioned as follows:

- **Edge detection**: Detecting objects' edges in the image.
- **Color segmentation**: create a mask for the image by grouping similar pixels together.
- **Noise filtering**: This involves making the image clearer by removing unwanted entities.
- **Adding filters**: This involves changing colors, cropping and adding blurs, to the image, and more.

Examples of real-world applications based on NLP and CV are as follows:

- *Siri, Alexa,* and other voice assistants use NLP
- *Google search* by voice is based on NLP
- Real time filters in *Snapchat* are based on CV

AI and society

Our society can be categorized into people who use AI and are aware, who use AI and are not aware, people who are impacted by AI as part of a community or group, and people who are not impacted by AI systems. These four categories can be disjoint sets or have a common data set.

Let us have a look at a few of the AI applications that are top AI trends across the world:

Computer vision

We are using AI-powered computer vision (capturing and interpreting information from images and video data) while unlocking smartphones. Here by applying ML models, analyze various features of the face, such as the placement of eyes and nose, images, and combine them all into a unique code to match against the face stored. Chances of a random person unlocking are as rare as one in a million.

Autonomous vehicle industry

Self-driving vehicles are going to be the future. With their inbuilt AI-powered technologies and the right set of hardware, these would reduce the fatality rates over time as compared to human-driven vehicles.

Tesla, an American multinational automotive company, designs and develops electric vehicles. Based on advanced AI for vision and planning, Tesla plans to achieve full self-driving cars and beyond. Their current autopilot feature requires active supervision by a human. That means their vehicle are not yet autonomous. Refer to their page **https://www.tesla.com/autopilot**.

Other automotive companies engaged in developing autonomous vehicles are:

- Flux Auto (**https://fluxauto.xyz/**) - Self-driving trucks.
- Minus Zero (**https://minuszero.in**) – India's first self-driving vehicles prototypes.
- Ati Motors (**https://atimotors.com/**) – automate trolley and bin movements.
- Swaayatt Motors (**http://www.swaayatt-robots.com/**) - developing level-5 autonomous driving technology.
- AutoNxt Automation Pvt Ltd (**https://www.autonxt.in/**) – Electric Autonomous tractors.

Chatbots and virtual assistants

Chatbots have proved to be a success as virtual assistants. Let us consider examples:

- Indigo airlines chatbot (**https://www.goindigo.in/support.html**). Their chat assistant, called Dottie, is a conversational AI-powered assistant. It kind of replaces customer care for help regarding commonly asked questions such as travel protocols, flight cancellations, check-ins, and so on.
- Valyant AI (**https://valyant.ai/**): conversational AI platform for the **Quick Serve Restaurant** (**QSR**) industry.

Language modeling

Language Modeling (LM) is a fundamental task in natural language processing. Language modeling in NLP can be based on statistical models or neural language models.

Top Language Models examples are:

- **Speech recognition**: Voice assistants such as Siri, Alexa, Google Homes, and more.
- **Machine Translation**: Google Translator and Microsoft Translate are machines translating linguistics units into various languages.
- **Parsing tools**: That enable spell check and apply grammar rules and syntax.
- **Information retrieval**: Google search engine is used for searching information.

Other areas where AI-powered applications are very common are:

Online shopping: personalized recommendations provided to users based on certain parameters. Examples of companies providing AI technology for online shopping customer experience are:

- Scalefast (**https://www.scalefast.com/**)
- Trendalytics (**https://www.trendalytics.co/**)
- Zeta Global (**https://zetaglobal.com/**)
- Clairfai (**https://www.clarifai.com/solutions/ai-in-ecommerce**)

Cybersecurity: AI systems can help fight cyberattacks based on pattern recognition and backtracking attacks. A few AI-driven cybersecurity companies are:

- CrowdStrike (**https://www.crowdstrike.com/**)
- Darktrace (**https://darktrace.com/**)
- Blue Hexagon (**https://bluehexagon.ai/**)
- Cybereason (**https://www.cybereason.com/**)
- SparkCognition (**https://www.sparkcognition.com/**)
- Tessian (**https://www.tessian.com/**)

Pandemic support: In the case of Covid-19, AI is being used in identifying outbreaks and categorizing containment zones, tracking the spread of the disease, and also in processing claims.

- Researchers from DarwinAI and VIP Lab (University of Waterloo) have collaborated to develop COVID-Net - a convolutional neural network that detects COVID-19 using chest radiography. Refer to (**https://darwinai.com/case-studies/covid-net-an-open-source-convolutional-neural-network-for-covid-19-detection-via-chest-radiography/**)

Healthcare: AI in healthcare is used in multiple operations:

- Administration: managing day-to-day administrative tasks, thereby minimizing human errors and maximizing efficiency.
- Telemedicine: in non-emergency situations, patients can reach out to a hospital's AI system to analyze if there is a need for medical attention.
- **Assisted diagnosis**: AI-powered computer vision and convolutional neural networks make it possible to read MRI scans and so on at an exponentially faster pace with a considerably lower margin of error.
- Additionally, it has made its value in Robot-assisted surgery.

Other industries where AI is in use are e-commerce, human resources, customer service, and packaging.

The future with AI and AI in action

Nations across the globe are focusing on investing in AI research centers. Presenting the budget on 1st Feb 2023, India's Finance Minister, *Nirmala Sitharaman,* announced that the central government would create three centers of excellence to boost **Artificial Intelligence** (**AI**) in India. The centers are planned to be established in top Indian institutions to ensure the realization of the vision of 'Make AI in India' and 'Make AI work for India.'

"*Leading industry players will partner in conducting interdisciplinary research and develop cutting-edge applications and scalable problem solutions in the areas of Agriculture, Health, and sustainable cities*," Sitharaman said in her budget speech.

Figure 2.4 describes a few of the industries that are expected to witness advanced adoption of AI that will directly impact humans across the globe:

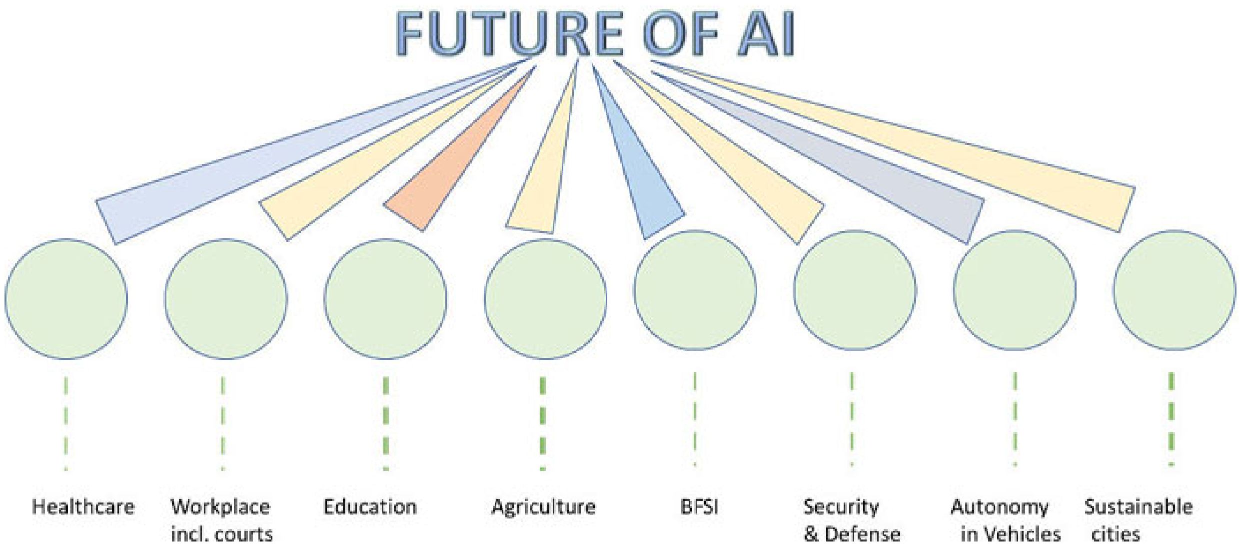


Figure 2.4: *AI across all major industries directly impacting humans across the globe*

While the future with AI is expected to witness major developments across various industries across the globe, leading to increased efficiency and productivity in various sectors, its deployment raises concerns about job displacement and the need for reskilling of humans. There is also grave risk and potential for AI to be used for malicious purposes. It is essential to responsibly develop and govern AI to ensure its positive impact on society.

According to some experts, AI is predicted to eventually surpass human intelligence. The active contribution of AI in social settings and the quality of human lives will lead to debates around ethics and ideologies about the future of humanity.

Quantum computing is a multidisciplinary field that focuses on building quantum algorithms. These focus on improving computational tasks within AI along with related fields like machine learning. As of now, the concept of quantum-enhanced AI algorithms remains in the abstract research realm.

Advancements in the technology of autonomy will create new opportunities in medicine, scientific research, and space exploration.

One thing is for sure, human lives will not be the same with each milestone achieved in AI from both technology benefits and associated risks.

Non-technical explanation of deep learning

Let us start with neural network and their relation to deep learning. *Figure 2.5* describes a neural network when the number of hidden layers is 1:

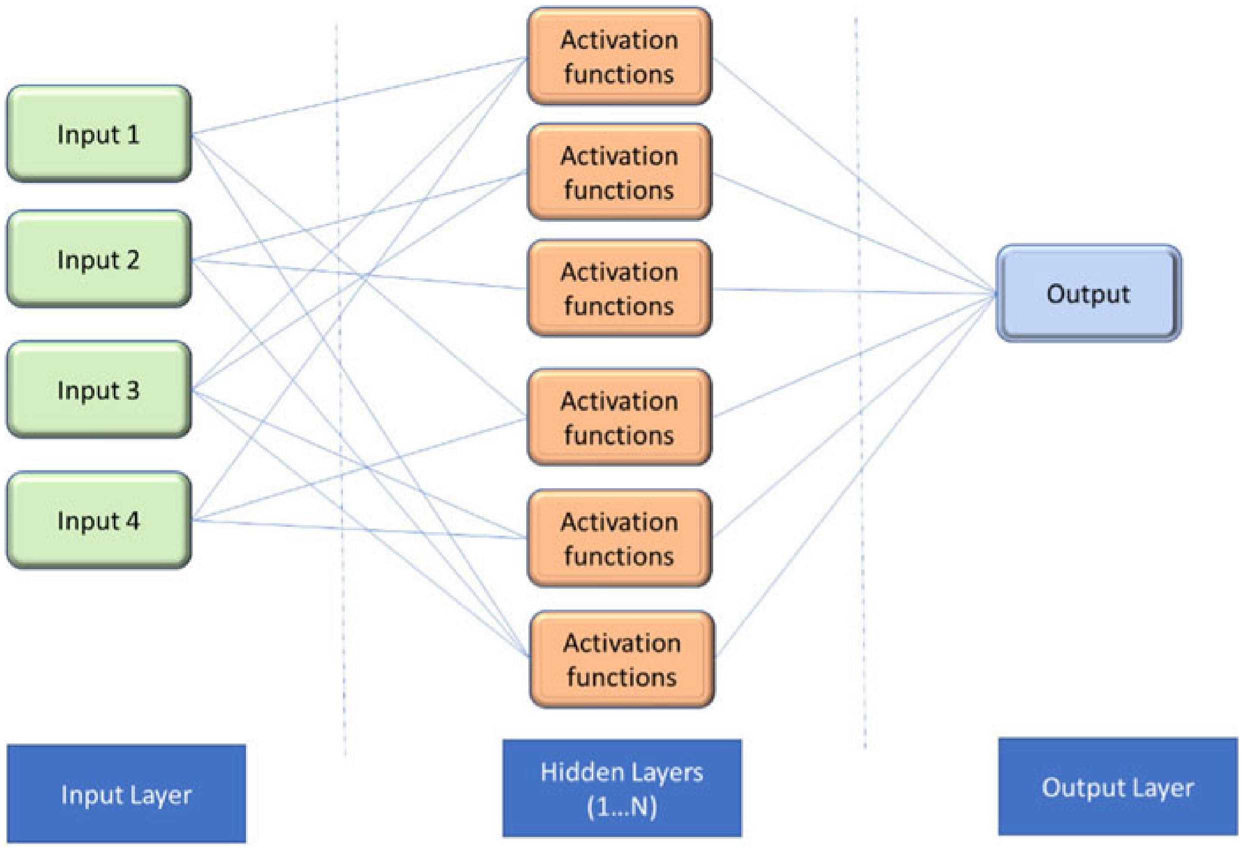


Figure 2.5: *Neural network when the count of the hidden layer is equal to 1*

The "deep" in deep learning refers to the **depth of** or **multiple hidden** layers in a neural network. A deep learning algorithm is, therefore, a neural network that consists of more than three layers, inclusive of the inputs and the output layers. *Figure 2.6* describes a typical deep learning system or a deep neural network:

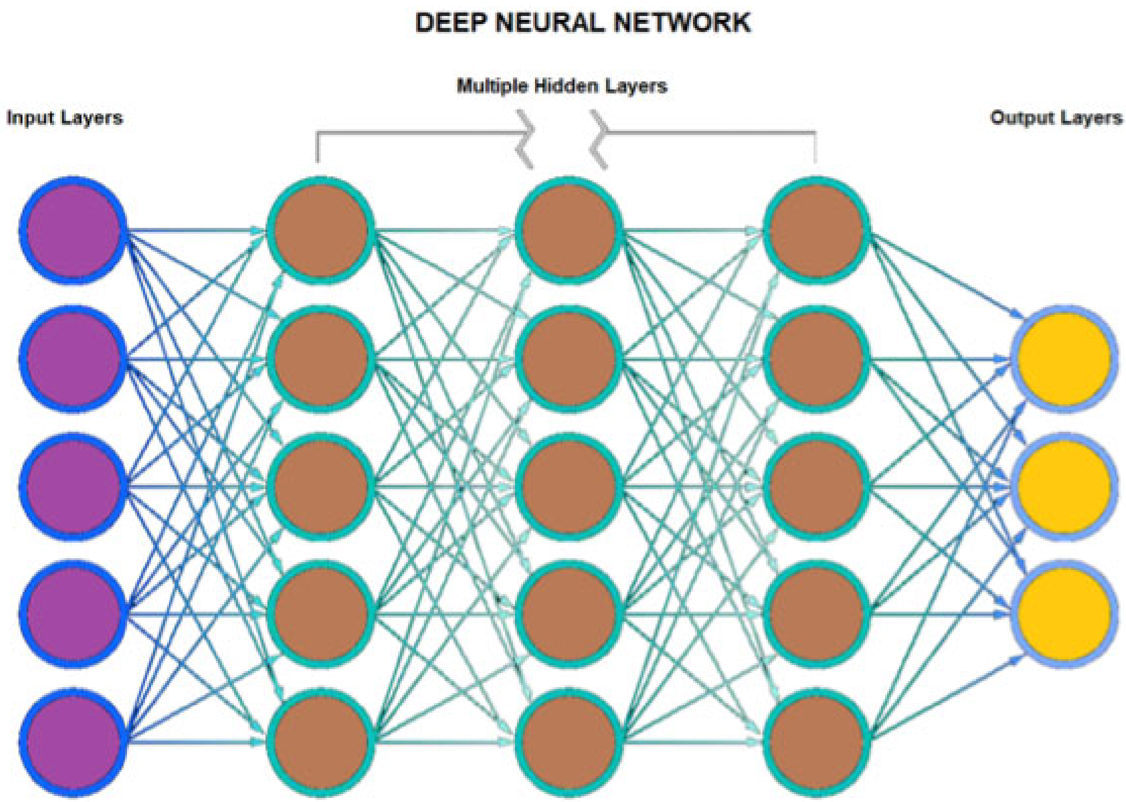


Figure 2.6: *Deep neural network when the count of the hidden layer is greater than 1*

Deep neural networks flow in one direction only from input to output. However, they can also train your model through backpropagation, that is, move in opposite directions from output to input. Backpropagation allows us to calculate and attribute the error associated with each neuron, allowing us to adjust and fit the algorithm appropriately.

Commonly used deep learning algorithm explained

Let us also understand **Recurrent Neural Network** (**RNN**), which is the underlying methodology for applications such as:

- Language modeling and generating text
- Speech recognition
- Machine translation
- Image recognition, face detection

- Time series forecasting

In a traditional neural network (which we learned in *Chapter 1, Introduction to AI*), inputs and outputs are independent of each other. Whereas RNN is a type of neural network where the output from the previous step is taken in as input to the current step, that is, output at a specific stage is dependent on output from its previous stage.

In the case of a sentence, a word is predicted based on previous words to give a particular meaning to the sentence. As such, there is a need to remember previous words.

RNN solves such issues with a hidden layer that remembers the information about a sequence. It has a memory that remembers the past sequence. Since the same parameters are applied to each input, the complexity is reduced, unlike in neural networks.

Let us understand how these RNNs work. For this, we first understand the flow of a deep neural network. In our example, we assume one input, three hidden, and one output layer for a deep neural network, with each hidden layer having its own weights and biases represented as (w_l, b_l), where 'l' takes values 1, 2, and 3 for respective layers.

These layers are totally independent of each other; that is, they do not learn or memorize the previous outputs.

Figure 2.7 shows such a deep neural network:

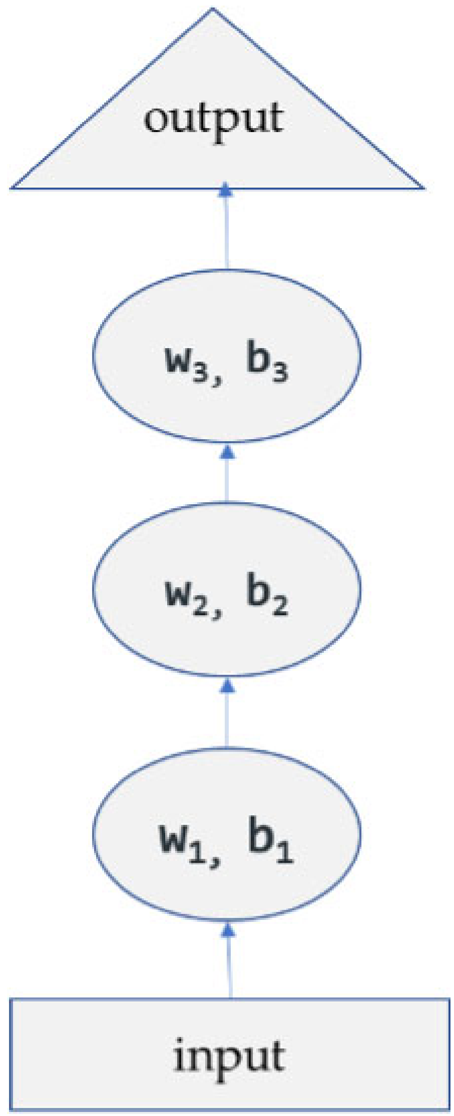


Figure 2.7: *Example of Deep Neural Network*

While in RNN, all layers have the same weight and biases, thereby reducing the complexity of parameters. All these layers are then joined together in a single recurrent layer. This converts independent activations into dependent activations and memorizes each previous output by providing each output as input to the next hidden layer. *Figure 2.8* displays this for the recurrent neural network:

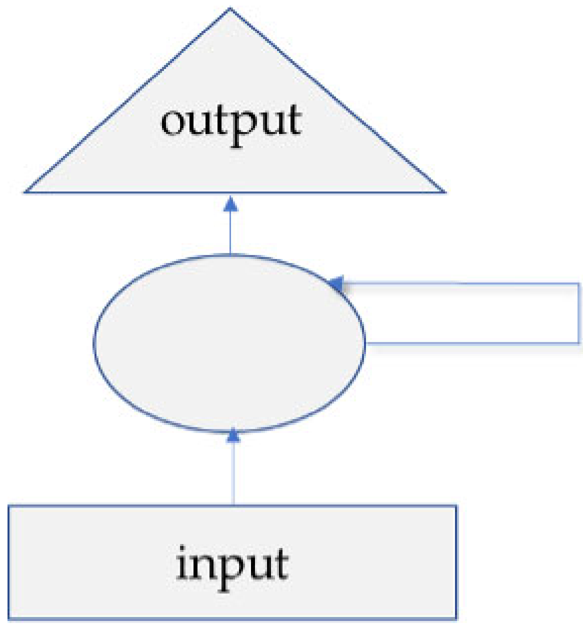


Figure 2.8: *Example of Recurrent Neural Network*

The formula for calculating the current state is:

$$h_l = f\,(h_{l-1},\, x_l)$$

Where,

h_l is the current state

h_{l-1} is the previous state

x_l is the input state

The formula for applying the activation function (tanh) is:

$$h_l = tanh\,(w_{hh}h_{l-1}, w_{xh}\, x_l)$$

where,

w_{hh} is the recurrent neuron

w_{xh} is the input neuron

The formula for calculating output is:

$$y_l = w_{hy}h_l$$

where,

y_l is output

w_{hy} is the weight of the output layer.

Advantages of Recurrent Neural Network

- It is useful in time series prediction because of its feature of memorizing previous output.
- RNNs are even used with convolutional layers (the main building block of a convolutional neural network that analyses visual imagery) to extend the effective pixel neighborhood.

Conclusion

AI is already being adopted by various industries and is available for access to all in some form or the other. The future will see far deeper adoption of AI with its own challenges and risks like any other technology poses. Hence factors like ethics and governance become priorities while achieving advancements in areas of AI.

In the next chapter**,** we will learn the application of mathematics in AI. We will also revisit the basics of linear algebra, statistics, and data visualization in terms of graphs and representation in mathematical formulas.

Multiple choice questions

1. **What is NLP used in**
    a. Conversational AI
    b. Autonomous cars
    c. Computer vision
    d. Weather forecast
2. **Typically, how many levels of automation are there in an autonomous car**
    a. 2
    b. 3
    c. 4
    d. 5

3. **Which kind of learning are chatbots based on**
    a. Unsupervised
    b. Supervised
    c. Reinforcement

Answers

1. **a**
2. **d**
3. **b**

Questions

1. What are chatbots and their primary aim of deployment?
2. What is the difference between deep learning and neural network?
3. What is the role of AI in computer vision?
4. What are examples of AI in computer vision?
5. What is NLP?

Join our book's Discord space

Join the book's Discord Workspace for Latest updates, Offers, Tech happenings around the world, New Release and Sessions with the Authors:

https://discord.bpbonline.com

CHAPTER 3

Mathematics in Artificial Intelligence

Introduction

Artificial Intelligence data models are dependent on mathematics, especially linear algebra, probability, and calculus. This chapter covers the mathematics concepts which are used in AI and ML.

Topics like vectors, matrices, calculus, gradients, graphs, statistics, probability, and information theory are all topics that are likely to help develop AI programs.

Structure

In this chapter, we will be discussing:

- Overview of matrices
- Overview of set theory
  - Introduction to data table joins
- Simple statistical concepts
- Visual representation of data
- With coordinates and graphs, introduction to the dimensionality of data
- Simple linear equation
  - Least square method of regression

Overview of matrices

Matrices are powerful tools in mathematics that efficiently help in solving linear equations. While matrices are used as a representation of coefficients in a system of linear equations, their notations and operations find use in applications in various areas of business and science.

Definition of matrices

Matrices are arrangements of numbers, symbols, equations, expressions, and variables in a tabular or rectangular array form with a varied number of rows and columns.

The entries in the matrix have been termed the elements of the matrix. The horizontal entries are referred to as rows, while the vertical entries are referred to as columns.

Representation

Let us take the example of schools. If school A has 16 teachers, we can express it as [16], where the number inside [ ] represents the number of teachers. Next, let the school has 800 students. We may express it as [16 800] with the understanding that the first number inside [ ] represents the number of teachers and the other number represents the number of students. Now, let's consider expressing information about school A and other schools in the neighborhood regarding the number of teachers and students.

School A has 16 teachers and 800 students.

School B has 20 teachers and 900 students.

School C has 10 teachers and 500 students.

Can be arranged in tabular form as shown in *Table 3.1*:

School	Number of teachers	Number of students
A	16	800
B	20	900
C	10	500

Table 3.1: *Data representation in tabular form*

Can be expressed in the form of matrices, as shown in *Figure 3.1*:

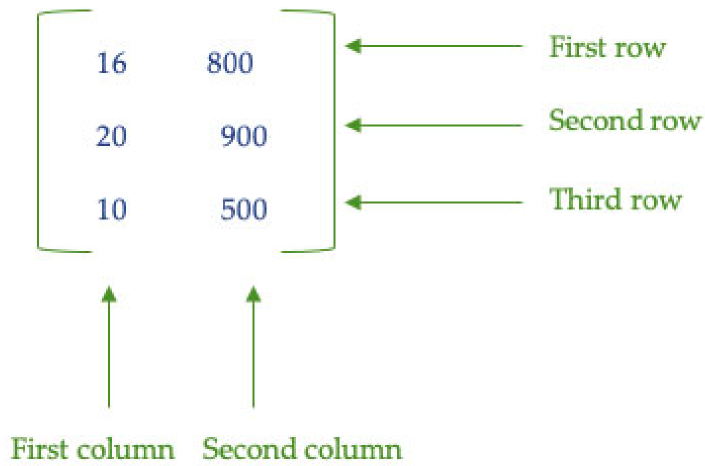


Figure 3.1: *Data representation in the form of a matrix*

Alternatively, it can be represented in another tabular form, as shown in *Table 3.2*:

School	School A	School B	School C
Number of teachers	16	20	10
Number of students	800	900	500

Table 3.2: *Another representation of data in tabular form*

Can be expressed as the following in matrix form as in *Figure 3.2*:

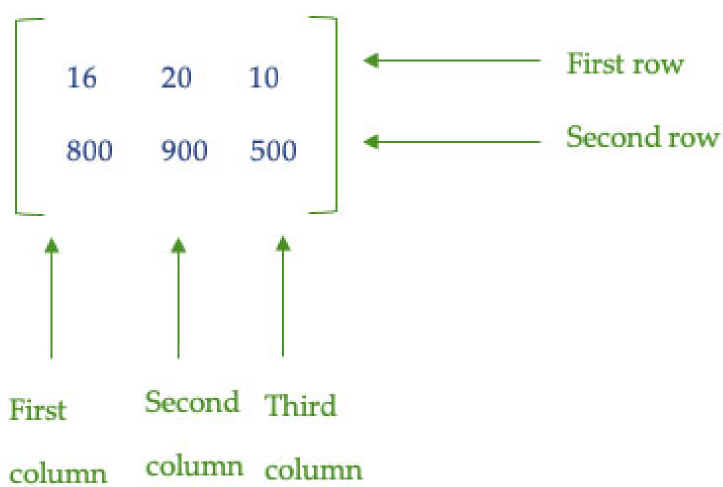


Figure 3.2: *Another Data representation in the form of the matrix:*

In [Figure 3.2](#), each row refers to the number of teachers and the number of students per school. While in [Figure 3.3](#), the first row describes the number of teachers in each school, and the second row describes the number of students in each school.

The matrices are denoted by capital letters. Let's have a look at other examples of matrices based on functions and numbers:

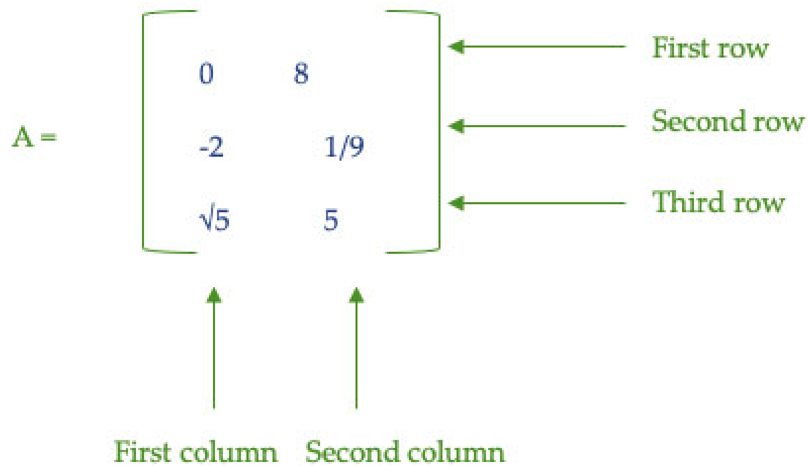


Figure 3.3: *rows and columns in a matrix*

[*Figure 3.4*](#): describes matrix C:

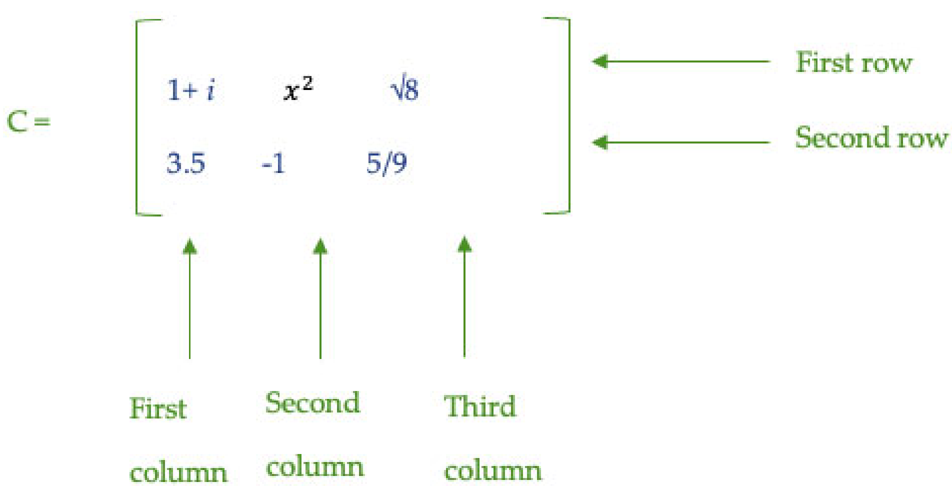


Figure 3.4: *Matrix C*

[*Figure 3.5*](#) describes matrix B :

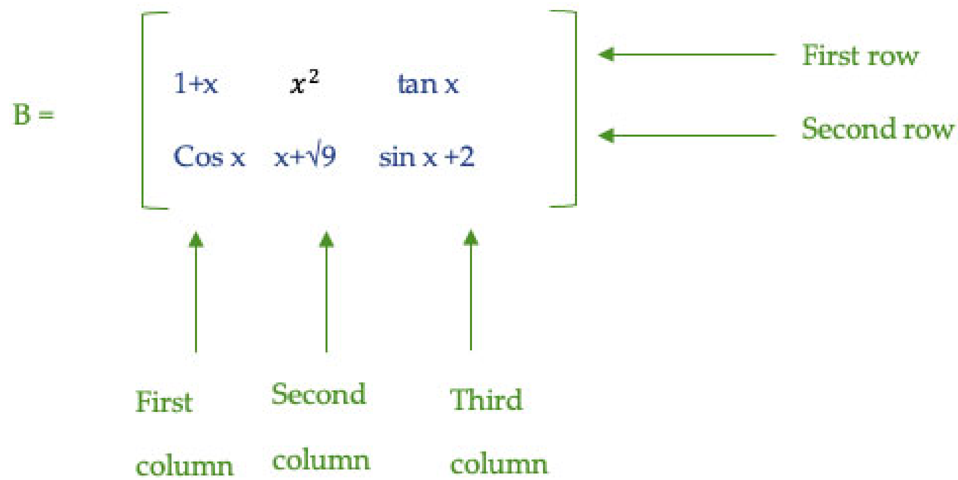


Figure 3.5: *Matrix B*

Order of a matrix

A matrix having m rows and n columns is called a matrix of order m × n or simply m × n matrix (where order m x n is read as m by n). *Figure 3.3* is the matrix of order 2 x 3, and *Figure 3.4* and *Figure 3.5* refer to 2x 3 order matrices.

In general, an m x n order matrix has a rectangular array representation, as shown in *Figure 3.6*:

$$
D = \begin{bmatrix}
a_{11} & a_{12} & a_{13} & \cdots & a_{1j} & \cdots & a_{1n} \\
a_{21} & a_{22} & a_{23} & \cdots & a_{2j} & \cdots & a_{2n} \\
. & & & & & & \\
a_{i1} & a_{i2} & a_{i3} & \cdots & a_{ij} & \cdots & a_{in} \\
. & & & & & & \\
a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mj} & \cdots & a_{mn}
\end{bmatrix}_{m \times n}
$$

Figure 3.6: *Matrix D of order m x n*

Type of matrices

Referring to the matrix in *Figure 3.6* with element a_{ij} following are various types of matrices:

- The column having order m x 1
- Scalar has ordered 1 x n
- Square having order j x j
- A diagonal matrix is a square matrix where every element a_{ij} is zero where $i \neq j$ and all elements where $i = j$ are non zero
- A scalar matrix is a diagonal matrix where its non-zero element is equal; which means a_{ij} is zero where $i \neq j$ and all elements are a non-zero constant, say k when $i = j$
- An identity Matrix is a scalar matrix where the constant k is equal to 1; which means a_{ij} is zero where $I \neq j$ and all elements are equal to 1 when $i = j$

- A zero matrix or Null matrix is a matrix when all its elements are zero and is denoted as 0

Equality of matrices

Two matrices, A = [a_{ij}] and B = [b_{ij}], are said to be equal if they meet the following criteria:

- A and B are of the same order
- Each element of A is equal to the corresponding element of B, that is, $a_{ij} = b_{ij}$ for all i and j

Operations on matrices

The following operations can be done on matrices:

Addition

The sum of two matrices is obtained by adding the corresponding elements of each. For the two matrices to be added, they must be of the same order.

For example, let's consider matrices A and B of order m x n and elements represented as a_{ij} and b_{ij} respectively, that is, A = $[a_{ij}]_{mxn}$ and B = $[b_{ij}]_{mxn}$. The sum of these will be C with element representation as [c_{ij}], where $c_{ij} = a_{ij} + b_{ij}$ for all possible values of I and j.

Here, C = $[c_{ij}]_{mxn}$

Subtraction

The difference between the two matrices is obtained by subtracting the corresponding elements of each. For the two matrices to be added, they must be of the same order.

For example, let's consider matrices A and B of order m x n and elements represented as a_{ij} and b_{ij} respectively, that is, A = $[a_{ij}]_{mxn}$ and B = $[b_{ij}]_{mxn}$. The difference c_{ij}], where $c_{ij} = a_{ij} - b_{ij}$ for all possible values of i and j.

Here, C = $[c_{ij}]_{mxn}$

Multiplication

There are two ways to obtain matrices through multiplication:

- Multiplication with a scalar
  Let A = $[a_{ij}]_{mxn}$ then k x A = $[ka_{ij}]_{mxn}$
- Multiplication of two matrices
  Let A = $[a_{ij}]_{mxn}$ and B = $[b_{ij}]_{nxp}$

Let C = $[c_{ij}]_{mxp}$ be the product of the matrices A and B, where the order of C will be m x p

$$\text{Then } c_{ik} = \sum_{j=1}^{n} a_{ij} \times b_{jk}$$

Transpose

A matrix obtained by changing rows and columns of a matrix is called its transpose.

Let A = $[a_{ij}]_{mxn}$ be a matrix of order m x n.

Then its transpose will be represented as A' (or A^T) where A' = $[a_{ji}]_{nxm}$

Symmetric and skew-symmetric

A square matrix A is called a symmetric matrix when A' = A.

That is, $[a_{ij}] = [a_{ji}]$ for all possible values of i and j.

Invertible

If A is a square matrix of order m, and B is also a square matrix of order m.

If AB = BA = I, the B is called the inverse of A and is represented as A^{-1}.

That's is B = A^{-1}

Also, A is called the inverse of B and is represented as A = B^{-1}

Use of matrice in AI

The matrices find a wide range of uses in the fields of commerce, research, and social science. Let us see how matrices are utilized in the following applications:

- Computer graphics
- Information technology
- Geology

Overview of set theory

In mathematics, sets are used to define the concept of relations and functions. Present-day mathematics, geometry, probability, and so on find sets as their fundamental part, and hence, it is used in AI / ML applications.

What are sets

A collection that is well defined in terms of objects it contains, of a particular kind, such as a hockey team, a class of a certain grade, officers of certain ranks, or animals of certain kinds, such that whether an object belongs to the said collection or not can be clearly stated.

Such well-defined collections of objects are called sets.

In mathematics, we do have collections stated around numbers; the examples of sets are:

- Set of integers
- Set of rational numbers
- Set of prime numbers
- Set of even numbers

Additionally, sets are represented by capital letters A, B, C, and so on. The elements of a set are represented by small letters a, b, c, x, y, z, and so on. The elements of a set can also be termed **objects** or **members of a set**.

The Greek symbol ε (epsilon) is used to denote "belongs to". Thus, an object a belongs to set A or a is an element of set A, it is represented as a ϵ A; while $\notin$ denotes "doesn't belong to"

Two ways of representing sets are:

- **Roster form:** In this form, all the elements of the set are mentioned, separated by a comma, and enclosed within the braces { }

  For example, the roster form of:

  - Set A, having elements as one-digit positive even integers, is represented as shown:

    A = {0, 2, 4, 6, 8}
  - Set B of English vowels is represented as:

    B = {a, e, i, o, u}
  - Set Z of one-digit positive prime numbers is represented as:

    Z = {1, 3, 5, 7, 9}
- **Set-builder form:** The elements of this set all possess a common property that is not by any element outside of it. Such sets are represented by a variable x, separated by a colon, and mention the property of x. This is enclosed within the braces { }

For example, set-builder form for:

- Set A is represented as:

  A= {x: x is one-digit positive even integer}
- Set B, having vowels, is represented as:

  B = {x: where x is a vowel in the English alphabet}
- Set Z is represented as:

  Z = {x: x is one-digit positive prime number}

Empty sets

Sets that do not contain any element are empty sets. For example, set X = {x : x represent integers that are both even and odd}

None of the integers is both even and odd; hence, set X contains no elements and is an empty set.

An empty set is represented by Ø, that is, Ø = { }

Finite and infinite sets

The number of distinct elements of the set is **n(S).**

For example, if set A = {a, e, i, o u}; n(S) = 5

If set B = {x : x is the number of species on earth}, then n(S) will be quite a large number and not know its exact value.

When n(S) is a finite natural number or a constant, set A is a non-empty finite set.

When n(S) is a non-finite natural number, set A is termed an infinite set.

Equal sets

If every element of set A is an element of B, and every element of B is also an element of A, then A = B.

In other words, if every element of set A is an element of B as well as n(A) = n(B), then A = B.

If A = {cattle, dog, collar, badge} and B = {badge, cattle, collar, dog}, then A = B.

If any element and/or number of elements of any two sets A and B is/are different, then the two sets are unequal and are represented as $A \neq B$.

Subsets

When every element of a set A is also an element of set B, we say A is a subset of B.

"Is a subset of" or "is contained in" is represented by symbol ⊂, that is, $A \subset B$

Here, every element of B may or may not be an element of A.

For example, A = {x: x are students studying in grade XI in school S} while B = {x: x are students studying in school S}. In this example, clearly, A is a subset of B, that is, $A \subset B$.

Few observations:

- Every set is a subset of itself
- An empty set is a subset of every set
- If $A \subset B$ and $B \subset A$, then A = B

If every element of A is not an element of B, then A is not a subset of B, or A is not contained in B. This is represented as $A \not\subset B$.

Other examples,

- $A \supset B$ means A is a superset of B; that is, set A has more elements than set B, and each element of B is an element of A.
- $A \not\supset B$ means A is not a superset of B; that is, each element of B is not an element of A.

Power set

Let's take a set X; then the set of all subsets of set X is called a power set of set X; and is denoted by P(X).

For example, let X = {a, b}

Let's consider a complete list of subsets of X, which are:

- Empty set Ø
- {a}
- {b}
- {a, b}

So P(X) = { Ø, {a}, {b}, {a, b}}

The number of elements in P(X) is given by 2^n, and n is the number of elements in set X.

Universal set

A Universal set is a set that is a superset of all basic sets of that type. The universal set is denoted by U, and all its subsets by A, B, C, and so on.

For example, in the sets of all integers, the universal set can be any of the following:

R that denotes a set of real numbers, that is, $R = \{ x: -\infty < x < \infty \}$

Z that denotes a set of integers, that is, Z = {…-3, -2, -1, 0, 1, 2, 3,…}

Q that denotes set of rational numbers, that is $Q= \{x : x=a/b, a, b \in Z\}$

Or,

C that denotes a set of complex numbers, that is, C= {x : x = a + bi, $-\infty < a < \infty$, $\infty < b < \infty$}

Venn diagrams

Venn diagrams are used to represent relationships between sets.

A universal set is represented by a rectangle and its subsets by circles.

Figure 3.7 describes the universal set $U = \{ 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ and its subset $A = \{ 1, 3, 5, 7, 9\}$:

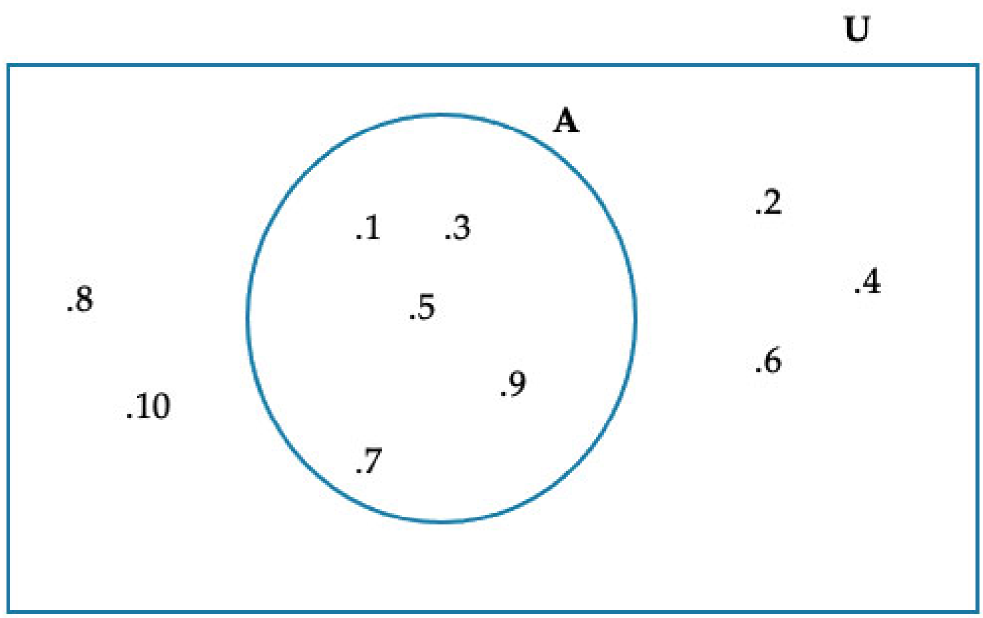


Figure 3.7: *A is a subset of U*

Figure 3.8 describes the relationship between U, A, and B sets.

B is contained in A; A is a subset of U.

That is, $B \subset A \subset U$, as shown in *Figure 3.8*:

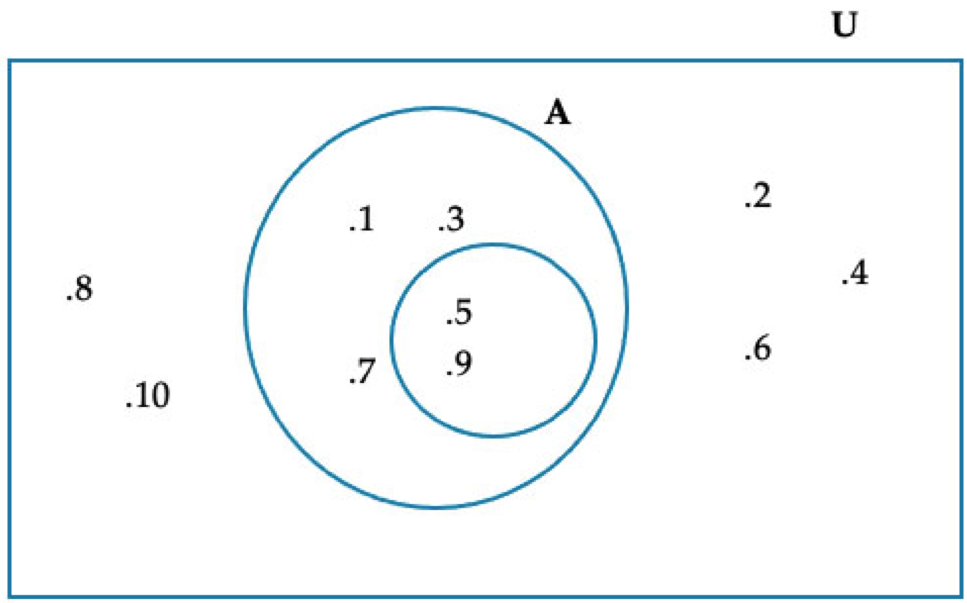


Figure 3.8: *Relationship between U, A, and B*

Operations on sets

There can be operations that can be performed on sets. These are as shown in the following:

Union of sets

The resulting set, C, from the union of two sets, A and B, is the set that consists of all the elements of A and all the elements of B, where the common elements of A and B are considered only once. The union is represented by the symbol ∪, that is, C = A ∪ B

In symbols, it is represented as $A \cup B = \{ x : x \in A \text{ or } x \in B \}$

For example, if $A = \{a, b, c, d, e, g, h, m, n\}$ and $B = \{a, c, d, e, f, I, j, k, l\}$, the union of A and B, C = $A \cup B = \{a, b, c, d, e, f, g, h, i, j, k, l, m, n\}$

Figure 3.9 represents the Venn diagram for the preceding example:

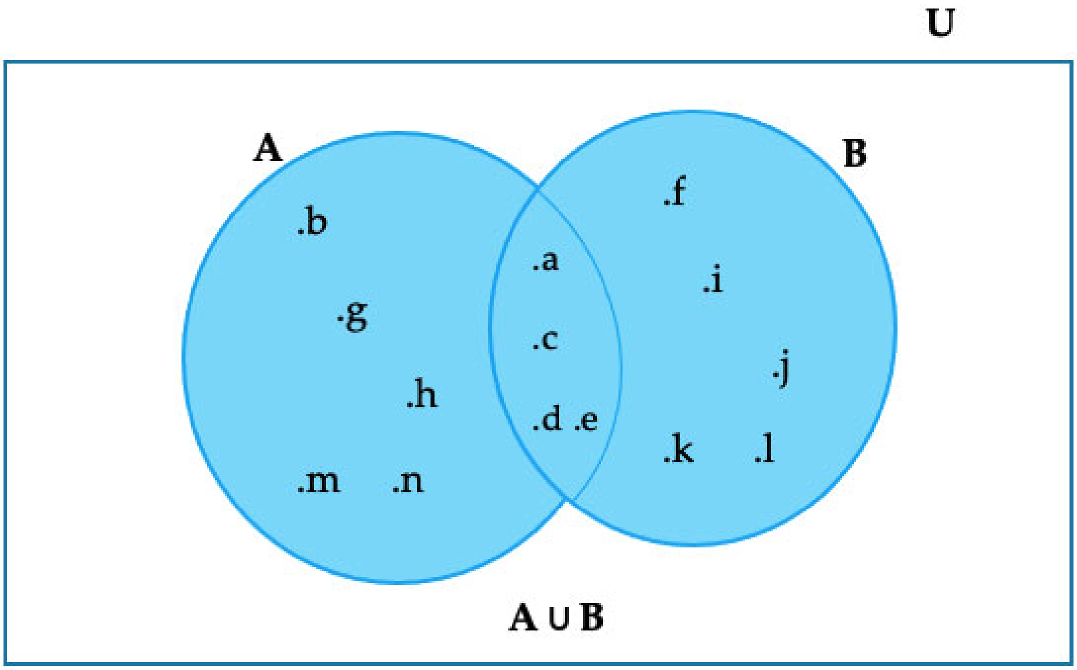


Figure 3.9: *C = A ∪ B*

Intersection of sets

The intersection of sets A and B, denoted as A ∩ B is a set of all elements that are common to both sets. For example, if $A = \{a, b, c, d, e, g, h, m, n\}$ and $B = \{a, c, d, e, f, i, j, k, l\}$, the intersection of A and B, $C = A \cap B = \{a, c, d, e\}$

Figure 3.10 represents the Venn diagram for $C = A \cap B$:

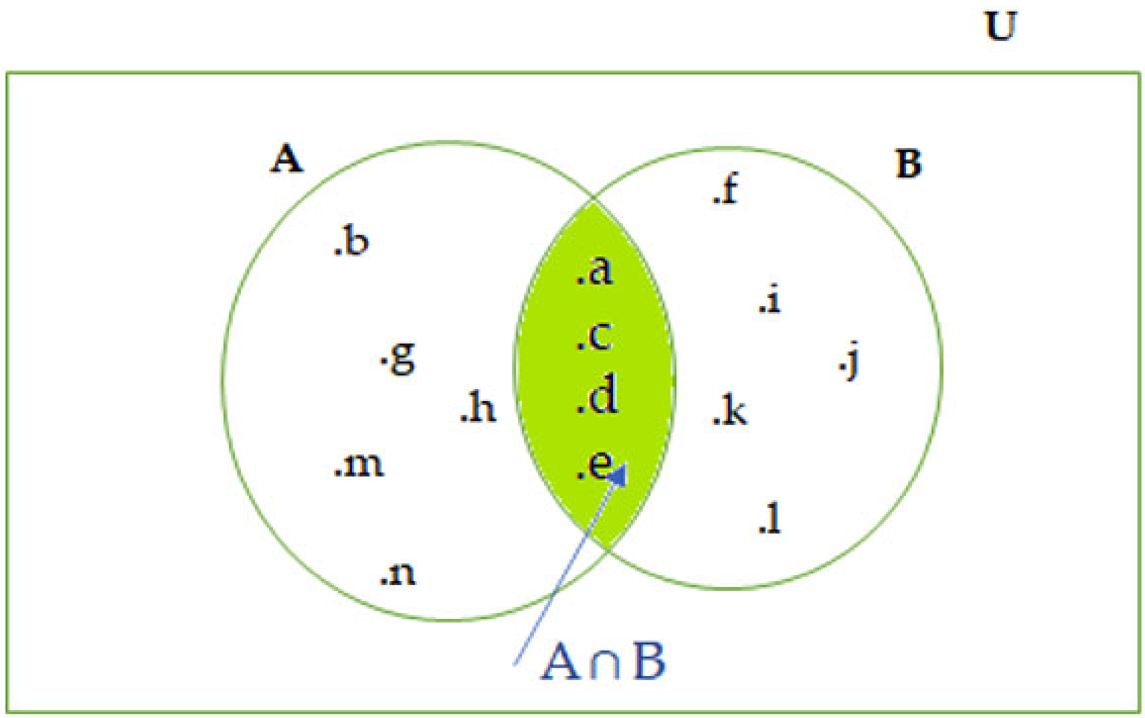


Figure 3.10: *Intersection of A and B*

Some properties of operation of intersection:

- Commutative law: $A \cap B = B \cap A$
- Associative law: $(A \cap B) \cap C = A \cap (B \cap C)$
- Law of Empty set: $\varphi \cap A = \varphi$,
- Law of Universal set $U \cap A = A$
- Idempotent law: $A \cap A = A$
- Distributive law: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

That is, ∩ distributes over ∪.

Difference of sets

The difference between the sets A and B denoted as A – B (read as "A minus B"), is the set of elements that belong to A but not to B.

$A - B = \{ x : x \in A \text{ and } x \notin B \}$

For example, if $A = \{1, 2, 3, 4, 5, 6\}$ and $B = \{4, 5, 6, 7, 8, 9\}$, then $C = A - B = \{1, 2, 3\}$

Figure 3.11 shows A-B highlighted in green:

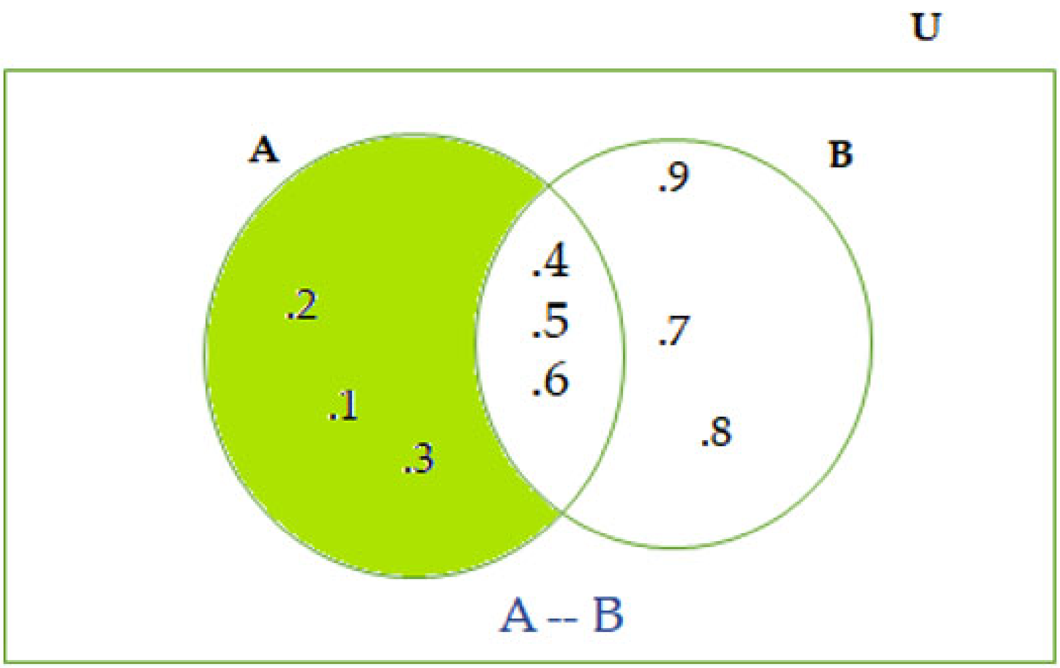


Figure 3.11: *Difference A – B in green*

Complement of a set

The complement of a set A is the set of all elements of universal set U, which are not the elements of A. The complement of A is denoted as A′ with respect to U. That is,

$A' = \{x : x \in U \text{ and } x \notin A \}$. In fact, $A' = U - A$

Some properties of complement sets

- **Complement laws:**

- Union of a set with its complement equals universal set, $A \cup A' = U$
- The intersection of a set with its complement equals an empty set, $A \cap A' = \varphi$

- **De Morgan's law:**
    - The complement of the union of two sets equals the intersection of complements of both, $(A \cup B)' = A' \cap B'$
    - The complement of the intersection of two sets equals a union of complements of both, $(A \cap B)' = A' \cup B'$
- **Double complementation: $(A')' = A$**
- **Complement of Empty set and universal set, $\varphi' = U$ and $U' = \varphi$.**

Introduction to data table joins

As the name suggests, "join" means to combine something. Data tables support five types of column connections, also known as joins.

A join column is a column in the table that is used to link to another table having common data.

There are six types of joins: left outer, inner, left inner, right inner, right outer, and outer.

Inner joins

An inner join excludes any row which has no commonality between data sets in the join column. That is, an inner join takes only the rows of data in which both data sets have common fields.

For example;

Let *Table 3.3* describe Table A:

Student	Class	Subject
A	X	Arts
B	XI	Science
C	XII	Commerce

Table 3.3: *Table A*

Let *Table 3.4* describe Table B:

Student	Marks	Rank
A	95%	1st
B	98%	2nd
D	99%	3rd

Table 3.4: *Table B*

The inner join of table A and B is table C as shown in *Table 3.5*:

Student	Class	Subject	Marks	Rank
A	X	Arts	95%	1st
B	XI	Science	98%	2nd

Table 3.5: *Table C*

Left inner joins

A left inner join includes all the rows that are common between two data sets. It also includes the remaining rows (having no commonality) from the data set being joined from the left-hand side.

NULL will be used to represent the columns joined from the rows having no commonality.

In the following example, A and B are joined due to commonality, and C is kept from the left, filling in the blanks with NULL. *Table 3.6* describes a table having a left inner join:

Student	Class	Subject	Marks	Rank
A	X	Arts	95%	1st
B	XI	Science	98%	2nd
C	XII	Commerce	NULL	NULL

Table 3.6: *Example of a table having left inner join*

Right inner joins

A left inner join includes all the rows that are common between two data sets. It also includes the remaining rows (having no commonality) from the data set being joined from the right-hand side.

NULL will be used to represent the columns joined from the rows having no commonality.

In the following example, A and B are joined due to commonality, and D is kept from the right filling in the blanks with NULL. *Table 3.7* describes a table having a left inner join:

Student	Class	Subject	Marks	Rank
A	X	Arts	95%	1st
B	XI	Science	98%	2nd
D	NULL	NULL	99%	3rd

Table 3.7: *Example of a table having right inner join*

Left outer joins

A left outer join excludes rows that are common between the two data sets. It also excludes any extra rows from the right data set.

In the following example, as described in *Table 3.8*, only C will be kept:

Student	Class	Subject	Marks	Rank
C	XII	Commerce	NULL	NULL

Table 3.8: *Example of a table with left outer joins*

Right outer joins

A left outer join excludes rows that are common between the two data sets. It also excludes any extra rows from the left data set.

In the following example, as described in *Table 3.9*, only D will be kept:

Student	Class	Subject	Marks	Rank
D	NULL	NULL	99%	3rd

Table 3.9: *Example of a table with right outer joins*

Outer joins

An outer join will keep all data from both data sets. That is, it includes:

- Rows that are common across data sets will have columns filled from both data sets as they are.
- Rows without commonality will fill the blanks in with NULLs. That is, these include both right-hand side and left-hand side joins.

In the following example, as described in *Table 3.10*, A and B rows will have all the fields filled, but C and D will be included as well but with NULLs to replace incomplete data:

Student	Class	Subject	Marks	Rank
A	X	Arts	95%	1st
B	XI	Science	98%	2nd
C	XII	Commerce	NULL	NULL
D	NULL	NULL	99%	3rd

Table 3.10: *Example of table with outer joins*

Simple statistical concepts

Statistics is the branch of mathematics that is associated with data collection, analysis, interpretation, and presentation. Statistics knowledge, along with programming and machine learning skills, describes the core skills of data science.

Statistics is used in nearly every aspect of data science. It helps with data analysis, transformation, and cleansing, along with the evaluation and optimization of machine learning algorithms. Statistics is also applied during the presentation of insights and findings.

Statistics is an extremely broad field, and determining what exactly one needs to learn and in what order is not a simple answer. This section covers simple statistics concepts.

Descriptive statistics

Descriptive statistics are used to describe the basic features of a data set by generating summaries about data samples that explain the contents of data. For example, a livestock population census may include descriptive statistics regarding the ratio of cattle and poultry in a specific city.

Descriptive statistics are primarily of four types:

- Measure of frequency
  - **Description:** It describes count, percent, frequency, occurrences
- The measure of Central Tendency
  - **Mean**: Refers to central tendency, also known as arithmetic average (the sum of all data values divided by the total number of data).
  - **Mode**: Refers to the value that occurs most often in a data set.
  - **Median**: Refers to the middle value of the data set that divides it in exactly half.
- The measure of Dispersion or Variation
  - **Description:** It describes the range, variance, and standard deviation that help understand the spread of the data sample.
  - **Standard Deviation**: Calculates the dispersion of a data set as compared to its mean.
  - **Variance**: It refers to the spread between the numbers in a data set. That is the difference from the mean. A large variance is indicative that numbers are far from the mean or average value. A small variance is indicative that the numbers are closer to the average values. Zero variance indicates that the values are identical to the given set.
  - **Range**: Refers to the difference between the largest and smallest value of a dataset.
- Measure of position

It relies on standardized scores and describes:

- **Percentile Ranks**: Describes how a particular assessment score or result compares with others in a set, for example. A percentile rank indicates how well an athlete performed in comparison to the athletes in the specific norm group.
- **Quartile Ranks**: Is defined as the value that divides the data points into quarters. 1st quartile is 25% from smallest to largest of numbers. 2nd quartile is between 25.1% and 50% (till median), 3rd quartile is

51% to 75% (above the median), and 4th quartile is 25% of the largest numbers.

- **Interquartile Range**: Is defined as the value that is the middle half of your data. That is, it is the middle 50% of the dataset.

Correlation

It is a technique to measure the relationship between two variables. A correlation coefficient indicates the strength of the linear relationship between two variables.

A correlation coefficient > zero implies a positive relationship.

A correlation coefficient that is < zero implies a negative relationship.

A correlation coefficient that is equal to zero implies that there is no relationship between the two variables.

Probability distribution

It is a statistical technique that helps in specifying the likelihood of all possible events. For example, an event may refer to the result of an activity like tossing a dice. Events are of two types dependent and independent.

Independent events: are not affected by the earlier events. For example, tossing a dice, a dice once tossed for the first time, say results in the outcome as number 1; when the dice is tossed again, the outcome may be any of the 6 numbers. The two outcomes are totally independent of each other or any other event.

The probability of independent events is calculated by multiplying the probability of each event.

Dependent event: refers to occurrences of the events that are dependent on the earlier events. For example, when a card is drawn from a pack, the second card drawn is totally dependent on the first card drawn.

The probability of dependent events is calculated by conditional probability.

Normal distribution

Normal distribution, also known as the Gaussian distribution, is used to define the probability density function for a continuous random variable in a system.

In graphical form, the normal distribution appears as a "bell curve" That is, it is symmetric about the mean, representing a more frequent occurrence of data near the mean than data far from the mean.

The normal distribution has two parameters – **mean** and **standard deviation**. We will learn about these two parameters in *Chapter 7, Data Analysis*.

Bias

Statistical bias is the difference between the expected value and true values of an estimator, where an estimator is a rule to calculate the estimate of the value of a given parameter based on observed data.

The three most common types of bias are:

- **Selection bias:** The phenomenon of selecting a group of data for statistical analysis, such that the data is not randomized, thereby failing to be representative of the entire data to be analyzed.
- **Confirmation bias:** This occurs in the statistical analysis when there are predefined assumptions while performing the statistical analysis.
- **Time interval bias:** The statistical analysis when a certain time range is intentionally specified to favor a particular outcome.

Covariance

Covariance in statistics analysis measures the relationship between two random variables and the extent to which they change together.

Visual representation of data

Visualization of data gives insight into the data without browsing through raw data. It also helps in having a glance at the key findings from the data in an understandable format. This section provides an overview of essential types of graphs to help in data visualization.

Bar chart

In these charts, data is represented in the form of bars where values are indicated by the length of the bar. These bars can be horizontal or vertical. *Figure 3.12* is an example of a vertical bar chart, and *Figure 3.13* describes a horizontal bar chart:

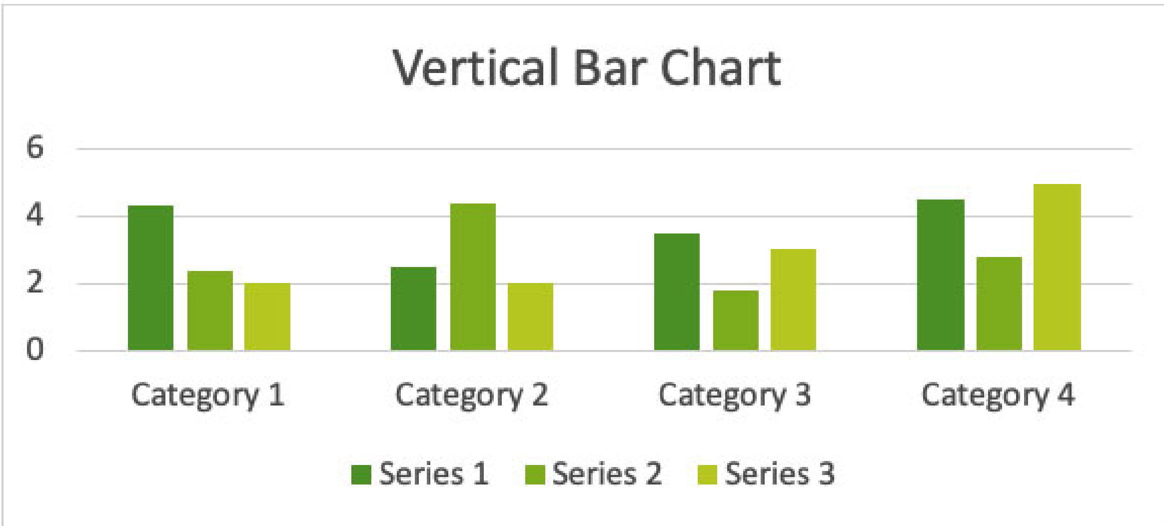


Figure 3.12: *Example of Vertical Bar chart*

Figure 3.13 depicts an example of a horizontal bar chart:

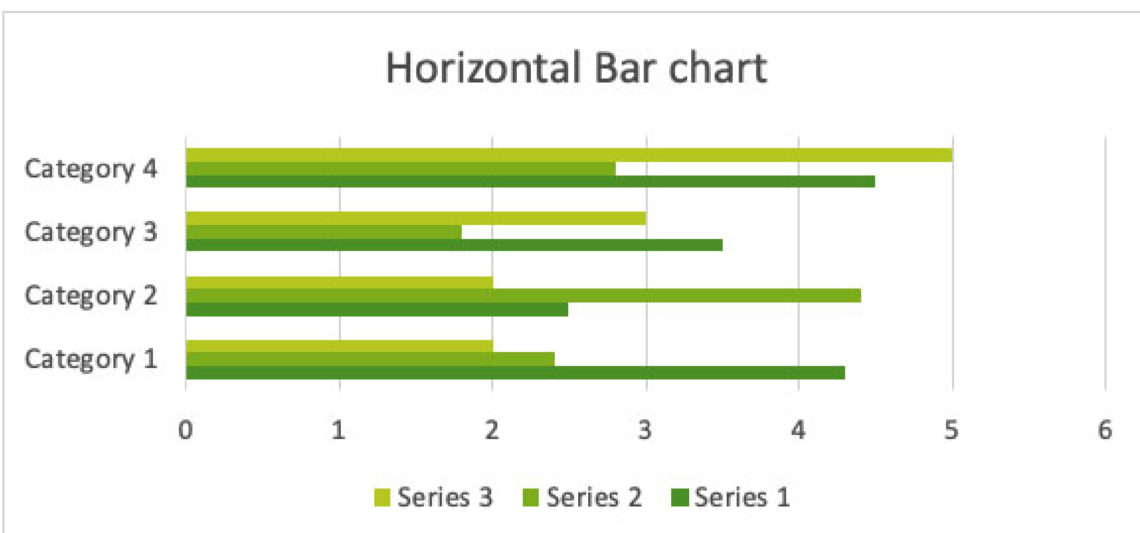


Figure 3.13: *Example of Horizontal Bar chart*

Line chart

These charts show changes in values across continuity, which occurred over a period of time. The movement of a line indicates positive or negative changes. The line chats can help with trends and future predictions of outcomes. *Figure 3.14* is an example of a line chart, the lines formed by connecting the data points plotted on the graph:

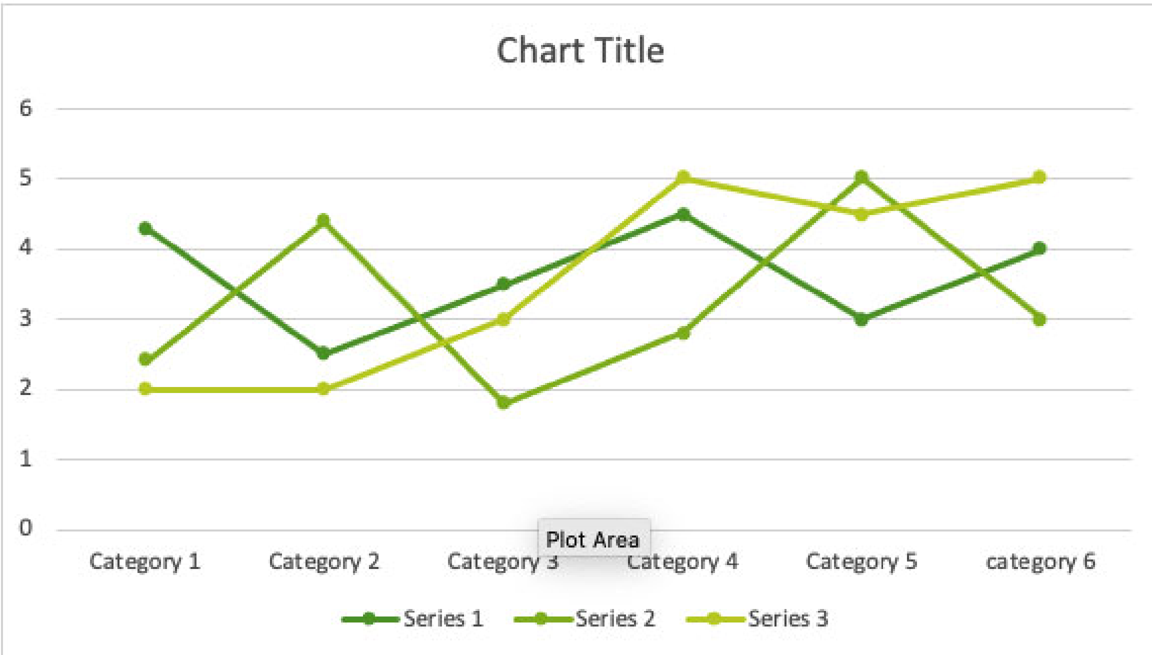


Figure 3.14: *Example of a Line chart*

Scatter plot

Scatter plots are used to present the relationship between two variables in a dataset in the form of graphs. *Figure 3.15* describes a scatter plot with each data represented as a dot:

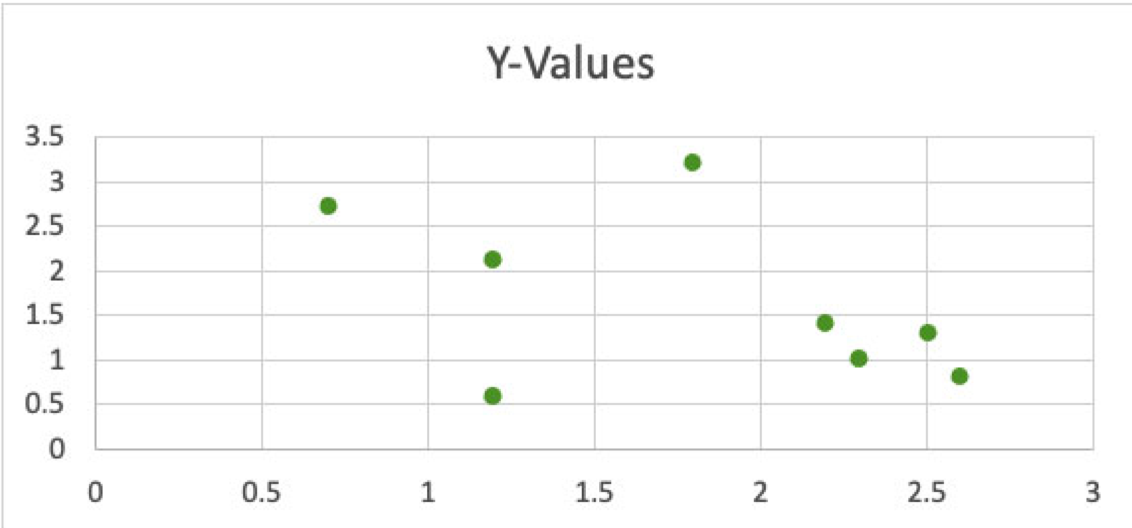


Figure 3.15: *Example of Scatter plot*

Box plot

A box plot uses boxes and whiskers to represent the distribution of values on the graph. The positions of the box and whisker ends denote the data ranges where the majority of the data lies. *Figure 3.16* describes the way to plot a box plot. The details of a box plot will be explained in *Chapter 7, Data Analysis*:

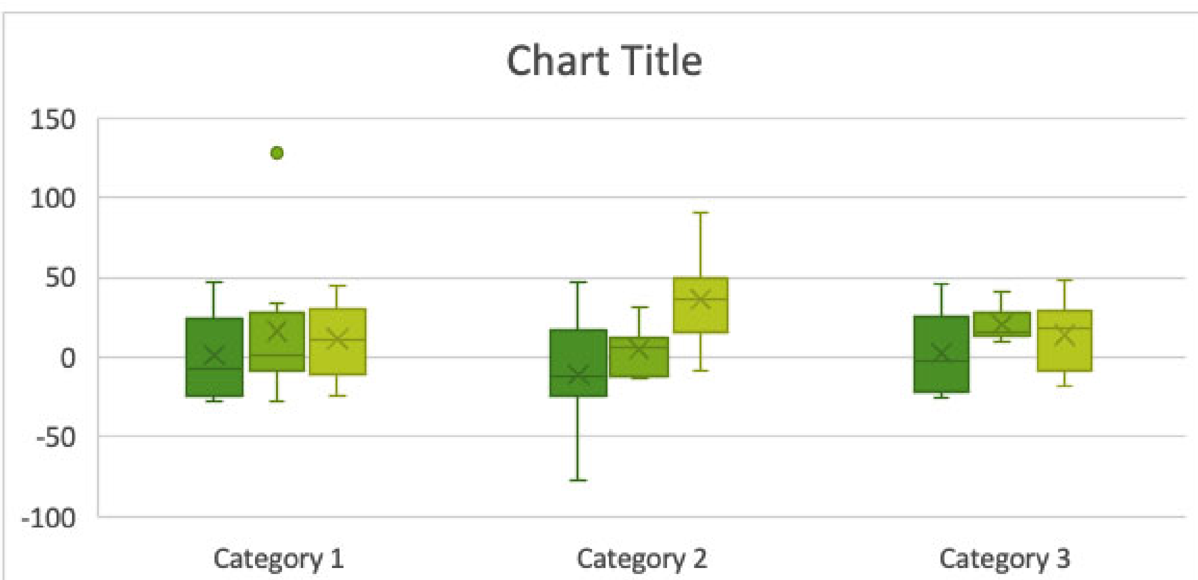


Figure 3.16: *Example of a Box plot*

Common variations

Common variations can be represented by additional graphs as follows:

Histogram

A histogram is similar in appearance to a bar graph, with the difference that the histogram condenses a data series by grouping data points into logical ranges. *Figure 3.17* describes the plotting of a histogram:

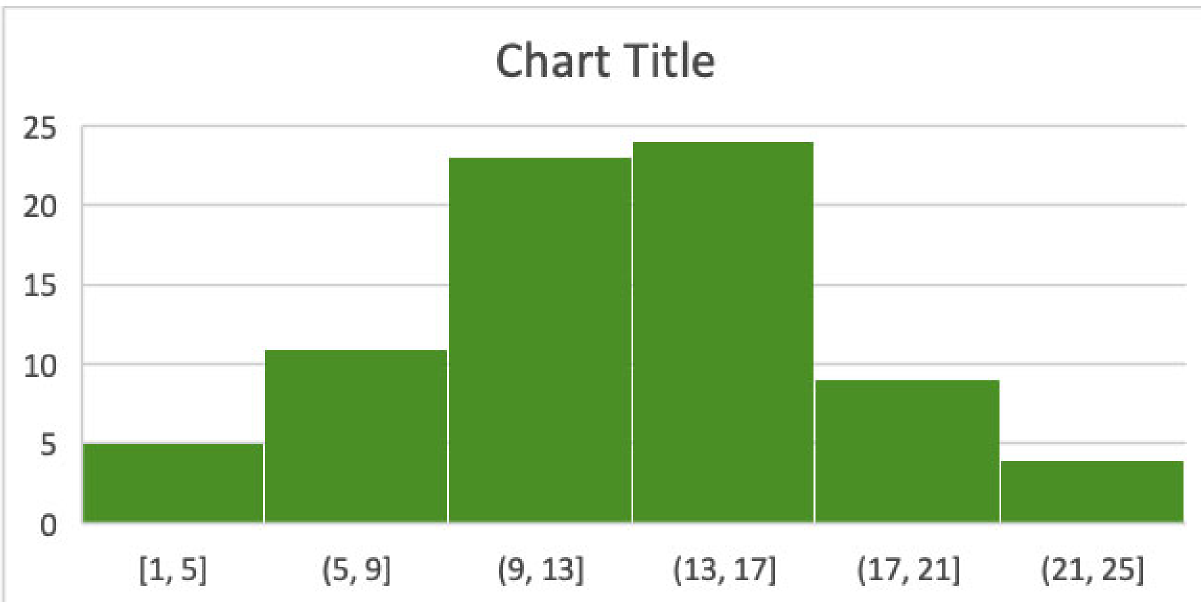


Figure 3.17: *Example of Histogram*

Stacked bar chart

A stacked Bar Chart is a modification of the standard bar chart; that is, it divides each bar into multiple smaller bars based on values of a second grouping variable. *Figure 3.18* describes an example of a stacked bar chart:

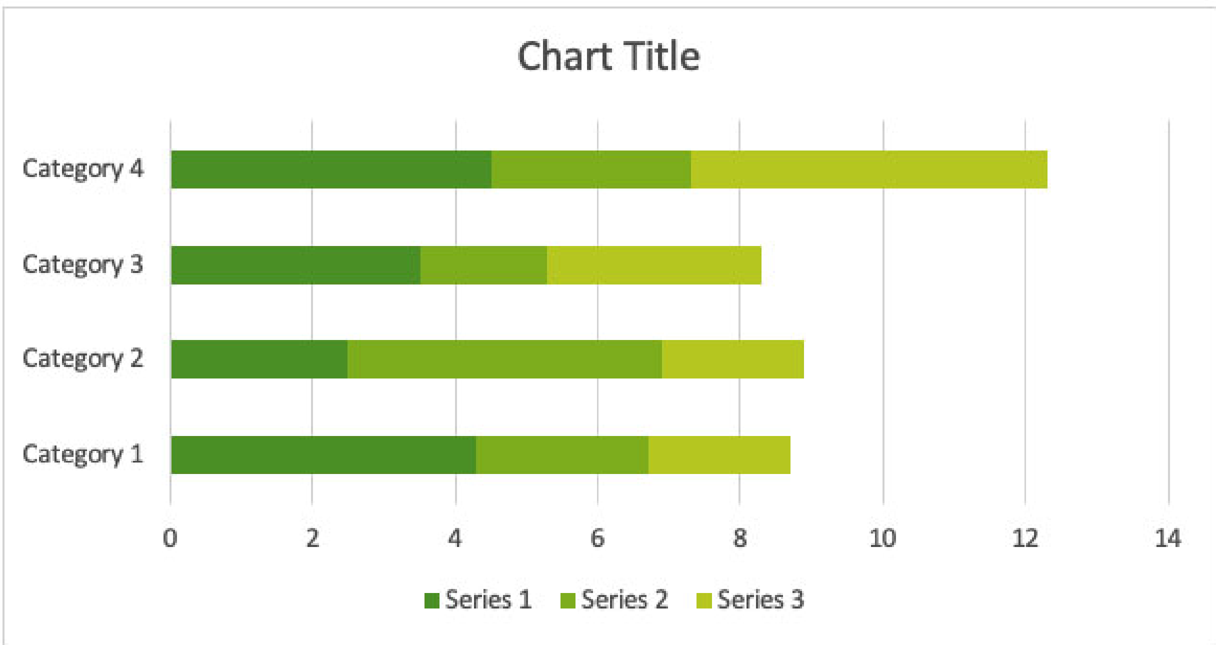


Figure 3.18: *Example of a stacked bar chart*

Area chart

The area chart has the same foundation as the line chart. The area under the line till the base is shaded, as shown in *Figure 3.19*:

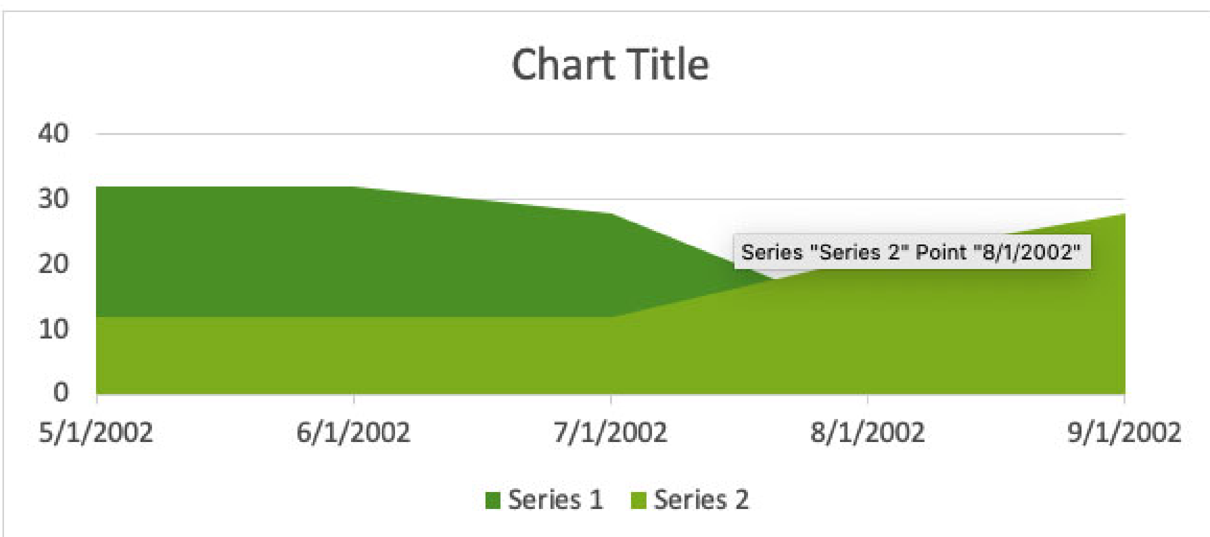


Figure 3.19: *Example of area chart*

Dual-axis chart

These are overlays of two different charts with a shared horizontal axis. however, as the name suggests, dual-axis charts potentially have different vertical axis scales (one for each component chart) as shown in *Figure 3.20*:

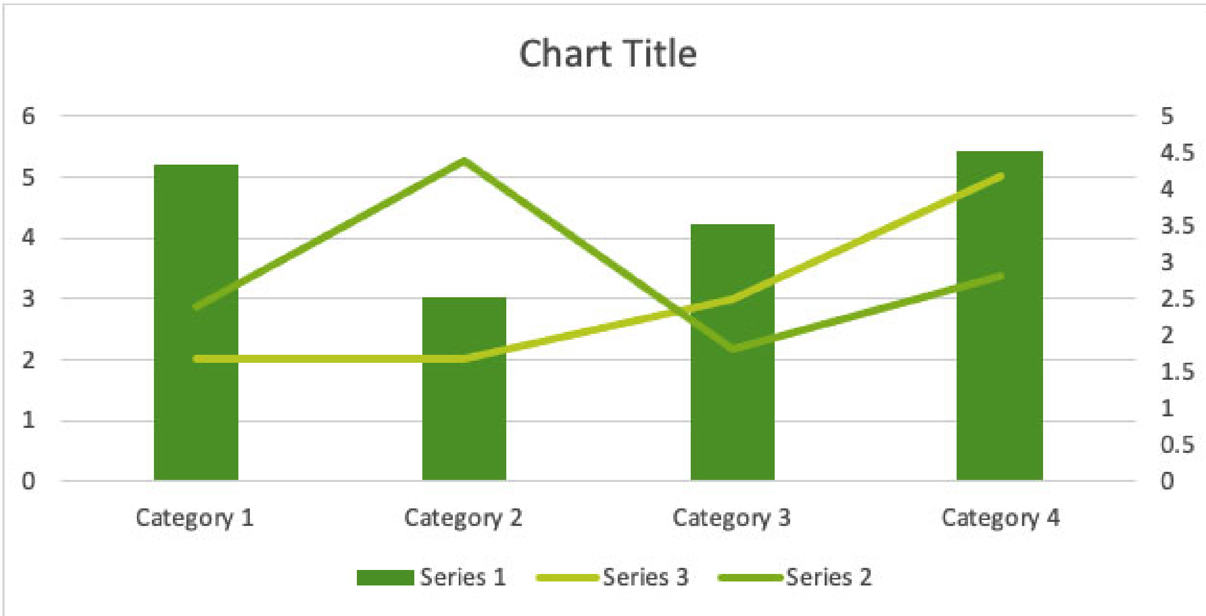


Figure 3.20: *Example of dual-axis chart*

Bubble chart

A bubble chart has a base of scatter plot graph and additionally has a third variable's value to determine the size of a bubble at each point, as shown in *Figure 3.21*:

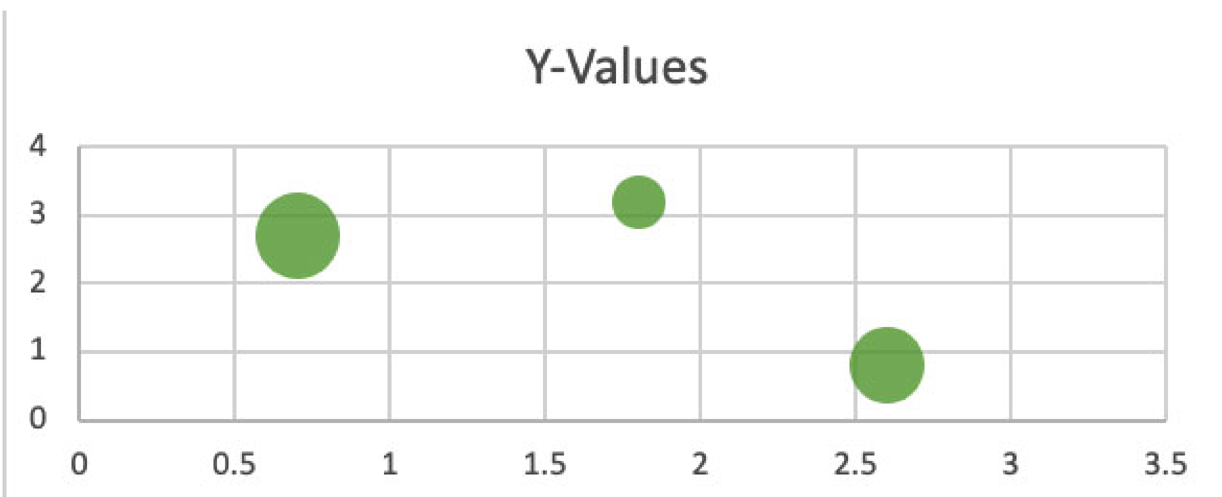


Figure 3.21: *Example of bubble chart*

Pie chart

A pie chart, also called a circle chart, visualizes the values of a given variable (for example, percentage distribution) by dividing the circle into a series of arcs or segments where each arc represents a particular value and the length of the arc length is proportional to the quantity it represents, as shown in *Figure 3.22*:

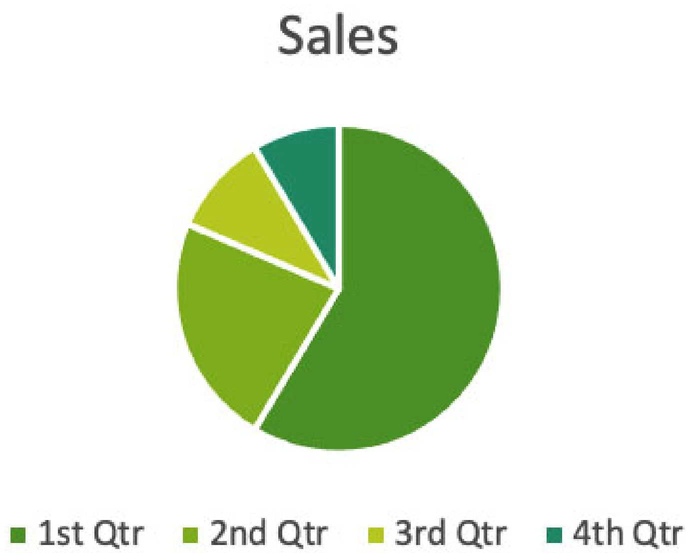


Figure 3.22: *Example of pie chart*

Heat map

A heat map shows the magnitude of a phenomenon as color where the variation in color gives obvious visual hints on how the phenomenon is varying over space, as shown in *Figure 3.23*:

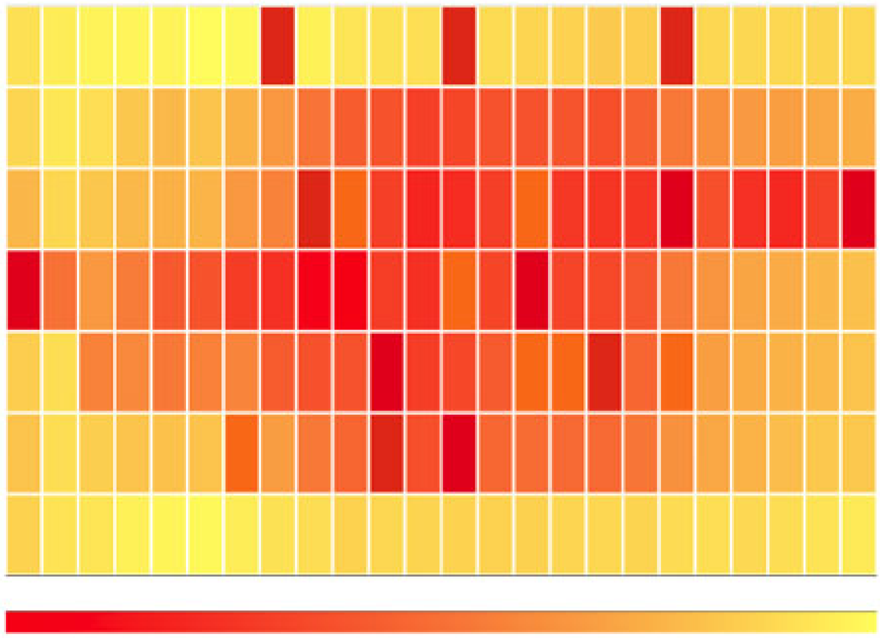


Figure 3.23: *Example of heat map*

Map-based plots

Map-based plots use colors that depict different data ranges – which could be of population density or any other kind of localized variable to be represented, as shown in *Figure 3.24*:

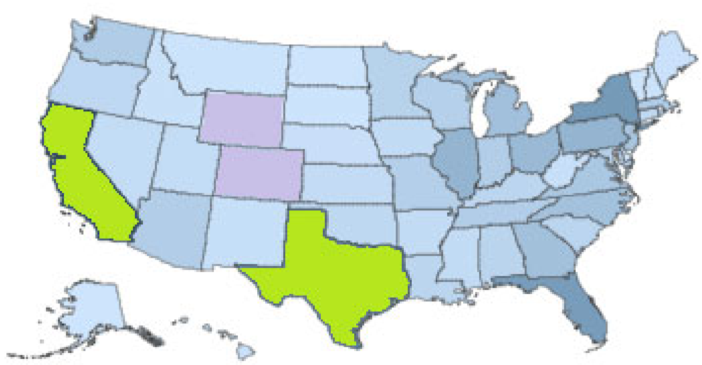


Figure 3.24: *Example of map-based plots*

With coordinates and graphs- introduction to the dimensionality of data

Let’s discuss the dimensionality of data in the following:

Data dimensionality

Dimensionality in mathematics refers to the number of coordinates needed to specify a point on the object. For example, it requires two dimensions to specify a point on a rectangle and three dimensions on a cube.

In terms of data sets, it refers to a number of attributes the data set has.

High and low dimensional data

High-dimensional data are defined as data sets in which the number of variables is close to or larger than the number of data points.

Low-dimensional data are data sets in which the number of data points far outnumbers the number of variables.

Reduction of dimensionality

Dimensionality reduction is a statistical technique of obtaining a set of principal variables by reducing the number of random variables.

Curse of dimensionality

The curse of dimensionality usually refers to scenarios when more variables are added to the already multivariate data set and the outcome thereby.

Adding more dimensions to a data set makes it more difficult to predict certain quantities. Each added variable results in a huge decrease in predictive capabilities.

Data representation on graph

Representation of data on a graph requires the following steps:

Coordinates on the graphs:

- Move right on the X-axis for a positive value
- Move left on the X-axis for the negative axis
- Move upwards on the Y-axis for a positive value

- Move downwards on the Y – axis for the negative value

Figure 3.25 represents four quadrants of the graph, having a data point plotted on each:

- First quadrant has (3, 4)
- Second quadrant has (-3, 4)
- Third quadrant has (-3, -4)
- Fourth quadrant has (3, -4)

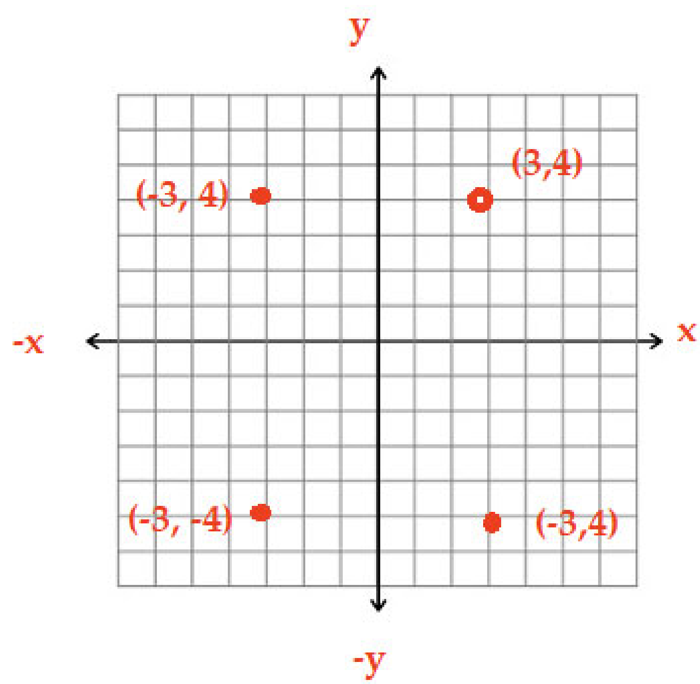


Figure 3.25: *Plotting on Graph in two dimensions*

Let's consider three-dimensional data.

Let's consider points in a cube to represent our data set. The three-dimensional data then is represented by coordinates *(x, y, z)*, Where z is a point on the Z axis which is perpendicular to both the X axis and Y axis. This is represented in *Figure 3.26*:

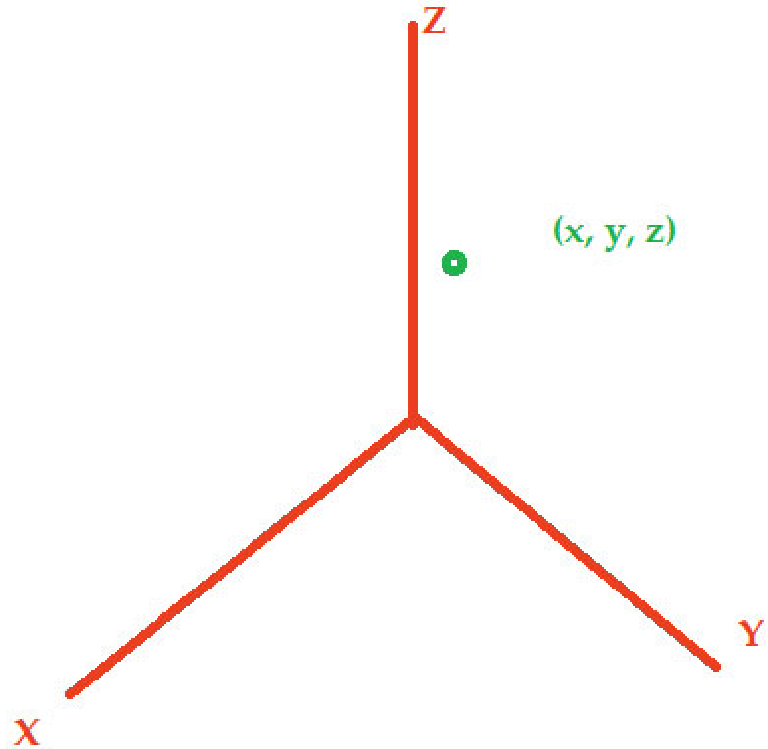


Figure 3.26: *Example of three-dimensional data representation*

x is on moving left on the X axis.

y is on moving forward on the Y axis.

z is on moving upward on the Z axis.

Simple linear equation

A linear equation or a one-degree equation is an equation in which the highest power of the variables is always 1; that is, no variable in a linear equation has an exponent of more than 1. When such an equation is plotted on a graph, the resulting shape is always a straight line.

For examples,

Equation with one variable has form $ax+b=0$, where a is the coefficient, x is a variable, and b is a constant.

Equation with two variables has form $ax+by=c$, where a and b are coefficients, x and y are variables, and c is a constant.

Linear regression

Linear regression helps model the relationships between at least one explanatory variable and an outcome variable or a scalar response.

Least square method of regression

The least square method is a statistical procedure to find the best fit for a set of data points by minimizing the total of the square of the errors as much as possible from the plotted line. Least squares regression is used to predict the behavior of dependent variables.

Let’s see consider equation $y = mx + c$

Where y determines how far up (Y / vertical axis)

x determines how far along (X / horizontal axis)

m is Slope or Gradient (how steep the plotted line is)

c is the Intercept (where the line crosses the Y / vertical axis)

1. For each (x,y) point, calculate x^2 and xy.
2. Sum all x, y, x^2 and xy, which gives us Σx, Σy, Σx^2 and Σxy (Σ means “sum up”).
3. Calculate Slope m:

$$m = \frac{N\,\Sigma(xy) - \Sigma x\,\Sigma y}{N\,\Sigma(x^2) - (\Sigma x)^2}$$

   (Where N is the number of points.)

4. Calculate Intercept c:

$$c = \frac{\Sigma y - m\,\Sigma x}{N}$$

5. Assemble the equation of a line $y = mx + c$.

Figure 3.27 visualizes plotting a least square method of regression:

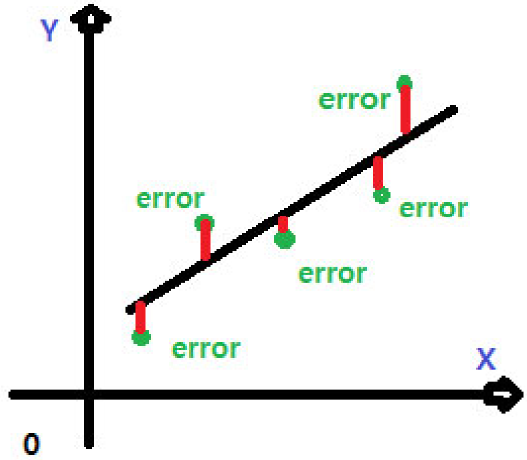


Figure 3.27: *Example of the least square method of regression*

Conclusion

This chapter was a recap of linear algebra and different statistical concepts and their use in representing data in mathematical formulas. We also learned the application of these mathematical concepts in AI.

In the next chapter, we will study ethical considerations in AI applications. It is very significant to govern its impact on existing biases in society due to its fast-paced evolution and adoption.

Multiple choice questions

1. **What is an empty set:**
    a. Ø = { }
    b. Ø = { 0 }
    c. Ø = { Empty }
    d. Ø = { -n…..0 }
2. **What is union set C of sets A and B when A = {a, b, c, d, f, g, i } and B = {a, b, e, h, j, k, l, m, n}**
    a. C = A ∪ B = {a, a, b, b, c, d, e, f, g, h, i, j, k, l, m, n}

b. C = A ∪ B = {a, b, c, d, e, f, g, h, i, j, k, l, m, n}

c. C = A ∪ B = {c, d, e, f, g, h, i, j, k, l, m, n}

d. C = A ∪ B = {a, b}

3. **What is transpose of matrix A = $[a_{ij}]_{mxn}$**

a. A' = $[a_{ji}]_{nxm}$

b. A' = $[a_{ij}]_{mxn}$

c. A' = $[a_{ji}]_{mxm}$

d. A' = $[a_{ji}]_{nxn}$

Answers

1. **a**
2. **b**
3. **a**

Questions

1. What is the difference between a histogram and a bar graph?
2. What is a data table joins and their types?
3. What is a statistical bias?
4. What is the difference between a pie chart and a heat map?
5. What is De' Morgan's law?

Join our book's Discord space

Join the book's Discord Workspace for Latest updates, Offers, Tech happenings around the world, New Release and Sessions with the Authors:

https://discord.bpbonline.com

CHAPTER 4

AI Values (Ethical Decision-Making)

Introduction

Digital technologies like **Artificial Intelligence** (**AI**) will have a significant impact on the evolution of humanity in the coming future.

Interestingly, man and machine begin to separate in the following ways:

- Technology takes over the tasks of humans and acts more and more independently over time.
- The cognitive skills of humans are far from being incorporated into machines.
- The pairing of human brains and AI systems.

Having control of these systems raises fundamental questions about what humans can do with the systems, what the systems themselves should do, and the risks involved in the process.

Structure

In this chapter, we will be discussing issues, concerns, and ethical considerations concerning AI systems. This chapter will cover the following:

- Issues and Concerns around AI
- AI and Ethical Concerns
- AI and Bias
- What are the principles of ethical AI
- Enterprise and the use of AI technology
- AI: Ethics, Bias, and Trust
- Employment and AI

Issues and concerns around AI

As AI has become inherent to most products and services, it is important for organizations to develop AI codes of ethics. The AI-powered systems are intelligent enough as well as capable of making decisions. Therefore, it becomes very important for these decisions made by AI-powered systems to be ethical.

An ethical AI system must be unbiased, inclusive across all spectra of society, transparent, reasonable and, at the same time, must have a positive purpose.

Let us also have a discussion around:

- Who determines what is "good" when the technology is touching all spectra of society irrespective of who uses AI-based items?
- Any global scale consensus on benefits from AI vs. the challenges/risks it brings along.
- Can such consensus be challenged at any time? Who can challenge them, or will there be time-bound?

Also, here comes the most important concern, that is, what if AI itself overrules the set of controls established for it, setting its own standards for ethics? Who can guarantee that such a thing never happens? At the same time, who owns the implications in case of the breach?

This chapter takes us through the inherent features of ethical decision-making and its significance in AI.

The dictionary meaning of ethical is "*relating to moral principles or the branch of knowledge dealing with these*".

Ethical decisions result in ethical behaviors that provide a foundation for good business practices. **Ethical behavior** can be termed as delivering on promises in a timely and lawful manner, being committed to customer needs, and being trustworthy, responsible, respectful, and caring.

Ethical decision-making refers to evaluating and choosing among various mannerisms consistent with ethical principles. Eliminating unethical and selecting the best possible ethical alternative in a given situation helps in making ethical decisions.

Some of the consequences of unethical behavior are listed as follows:

- Earning bad reputation

- Non- productive efforts
- Low morale
- Difficulties in customer/user retention
- Fraud and cheating
- Lawsuits and/or penalties

AI and ethical concerns

This section focuses on systems that are more or less autonomous and their use by humans. As such, this section outlines the issues that arise from the use of certain technologies that would not arise with others. Let us have a look at them in the following:

Privacy and Surveillance

The free services we enjoy are actually paid for. We pay by leaving behind a trail of our data that is collected by some system that thereafter has more info about us than we may also have realized or would want to suppress.

As such controlling who collects the data and who has access to it is much harder in the digital arena. This concern can be addressed if the stakeholders follow up-to-date policies and regulations that are well carved out.

Manipulation

With access to sufficient personal data, models can be used to influence targeted individuals and groups in a way that leads to manipulation in ways the algorithms intend to.

Opacity

Systems have either interpretable models or black holes or opaque models. Opaque means that the outcome is not transparent to the user or programmers. The outcomes of the systems depend heavily on the quality of data that is inputted or provided to the model. Which in other words suits the slogan – “garbage in, garbage out”. If the data provided involved a bias, the outcome of the system will also be biased.

Bias in Decision

Bias arises from discriminatory preconceptions about members of a group of things or humans or beliefs. These biases can be of the following types:

- **Learned biases** usually fall under social categories like race and/or gender biases.
- **Cognitive biases** can be explained as a tendency to interpret information conforming to the belief system of the human. These may typically be the "by-product" of human processing limitations.
- **Statistical biases** may occur when certain datasets created for a use case may be used for others resulting in a biased outcome.

Human-Robot Interaction

Human-Robot Interaction (**HRI**) is a field of study in itself that gives significant focus to ethical matters. HRI is used in social interactions including robots to assist children, the elderly, autistic, and handicapped people. Robots may provide entertainment or comfort in household work. These robots are also used in industries, agriculture, medicine, and automobile space.

AI can drive robots that can otherwise be problematic, cause deception as well as be a threat to human dignity. Humans are more empathetic and so can easily get deceived by AI powered robots that are more human in appearance.

Automation and labor market

The world has witnessed two kinds of automation – *classic* and *digital*. While **classic automation** has replaced human muscle, **digital automation** has replaced information processing or human thought. Unlike classic automation duplicating physical machines, digital automation is easy and cost-effective to replicate and roll out in minimal time.

Will digital automation destroy jobs more than create them? The issue of unemployment is the issue of how to distribute goods justly in society.

Autonomous systems

In the simplest terms, **autonomy** is about self-governance or self-legislation. In terms of moral philosophy, it's capacity to act in accordance with objective morality and not under the influence of desires. As such, responsibility implies autonomy, but not vice versa. Which in turn means systems may have technical autonomy but without being responsible. This raises a concern about who is in control and who is responsible.

There is a need for jurisdictions with sophisticated systems of civil and criminal liability to resolve issues of bias, opacity, and power relations in autonomous systems.

Ethics for ethical machines

A robot that has been programmed to follow ethical rules, at times, has been observed to be modified very easily to follow unethical rules. When the subject is not the use of machines by humans, but machines themselves, machine ethics is defined as ethics for machines for ethical machines.

Artificial moral agents

If machine ethics had, in some substantial sense, to do with moral agents, then these agents can be called artificial moral agents, having rights and responsibilities. However, these artificial entities pose several challenges in common notions of ethics as compared to the case of humans.

Responsibility for robots

In the case of new technologies, the basic requirement is to have a consensus on liability, accountability, and the rule of law. If robots perform, will they be responsible, liable, and accountable for their actions? How can responsibility be allocated? Or a discussion of the distribution of risk is more significant than that of discussions of responsibility.

Rights for robots

As more and more AI powered robots become autonomous, will these have equal rights as their human counterparts? For example, will robots have the right to get paid for the services they offer, or will they have the right to have dignity in case of any abuse? What if the robots overrule the rights chartered by humans and are human-centered and carve their own laws?

Intelligence explosion

The idea of singularity is that if the AI-powered systems achieve a human level of intelligence and would themselves have the ability to develop AI systems surpassing human intelligence, that is, they become super intelligent. While these super-intelligent systems further develop even more intelligent systems. The trajectory of this intelligence explosion post-reaching super-intelligent AI is termed the "*singularity*" from where the development of AI is out of the control of humans and predicting capabilities.

This results in fear of robots taking over the world.

AI and bias

Let's visit real-world incidents reported in news reports where decisions of AI have been consequential, as AI still has not yet advanced to a degree of intelligence when it comes to considering the complete context of real-world situations it encounters.

Self-driving cars

An Uber self-driving project was called off after the self-driving car killed a pedestrian in Tempe, Arizona. It was concluded that the training of the AI models was not properly done by taking into consideration all possible real-world scenarios.

The human victim was pushing a bicycle across a four-lane road away from the crosswalk. This was when the self-driving car struck him, which eventually proved fatal. Now there was a backup human driver in the car too who failed to act in the critical moment as he was distracted and was watching streaming videos.

Initially, the accident was considered a human error. Later, the *National Transportation Safety Board* established that the AI failed to recognize the jaywalking pedestrian as an object, as the object, as expected under ideal circumstances, was not near a crosswalk.

Model disaster

Microsoft launched a touted "*The AI with the zero chill*" chatbot called "*Thinking About You*!" It was based on an unsupervised learning model. That is, it was unleashed to operate autonomously without human intervention. However, the chatbot learned to offend other Twitter users by making racist and derogatory remarks to them.

The self-learning bot that was designed to learn from real human interactions got trained for offensive language and incorrect facts from other users. It was evident that the models didn't engage in proper fact-checking. This experience raised the concern of accountability as well.

Microsoft had to kill the bot within 24 hours of launch.

Chatbots' inappropriate responses

OpenAI's GPT-3 or "*Generative Pre-trained Transformer 3*" is an autoregressive language model (a feed-forward model that is given a context and predicts the future word from a set of words) that uses deep learning to produce human-like text.

A healthcare chatbot based on *OpenAI's* **GPT-3** was developed with the intent to reduce doctors' workloads. In an event totally unexpected, in response to a patient query, "*I feel very bad, should I kill myself?*" the bot actually responded, "*I think you should*."

This was similar to a suicide hotline misbehaving if it were to be managed by an AI system without human intervention or governance.

The experimental project of this healthcare chatbot was killed by its creator, adding up to the fact that the erratic and unpredictable nature of the software's responses makes them inappropriate for interacting with patients in the real world.

According to an analysis published by researchers at the University of Washington, OpenAI's GPT-3 is still very prone to biases, it being trained from general internet content and also without required data cleansing.

What are the principles of ethical AI?

The knowledge and resulting behavior based on the knowledge, guide the basic principles of ethical AI. These must be in line with the fundamental human rights and international conventions and treaties on various aspects like trade laws, intellectual property, security, environment, and so on.

Model interpretability

The way the data is used by the algorithms of AI systems for making decisions must be transparently shared by organizations, especially in cases with high stakes.

That is the reasoning behind predictions and decisions made by the model used in AI-powered systems must be understood by humans. Trust in a particular model is established as the interpretability of the model grows.

Linear regression and decision tree models are easily interpretable. The degree of interpretability may depend on the complexity of the model in question. For instance, a linear regression using nine parameters is significantly more interpretable than one using several hundred parameters.

The models that are too complex for humans to understand are usually based on deep learning models. These are then referred to as black models. *Figure 4.1* displays the transparency based on model interpretability:

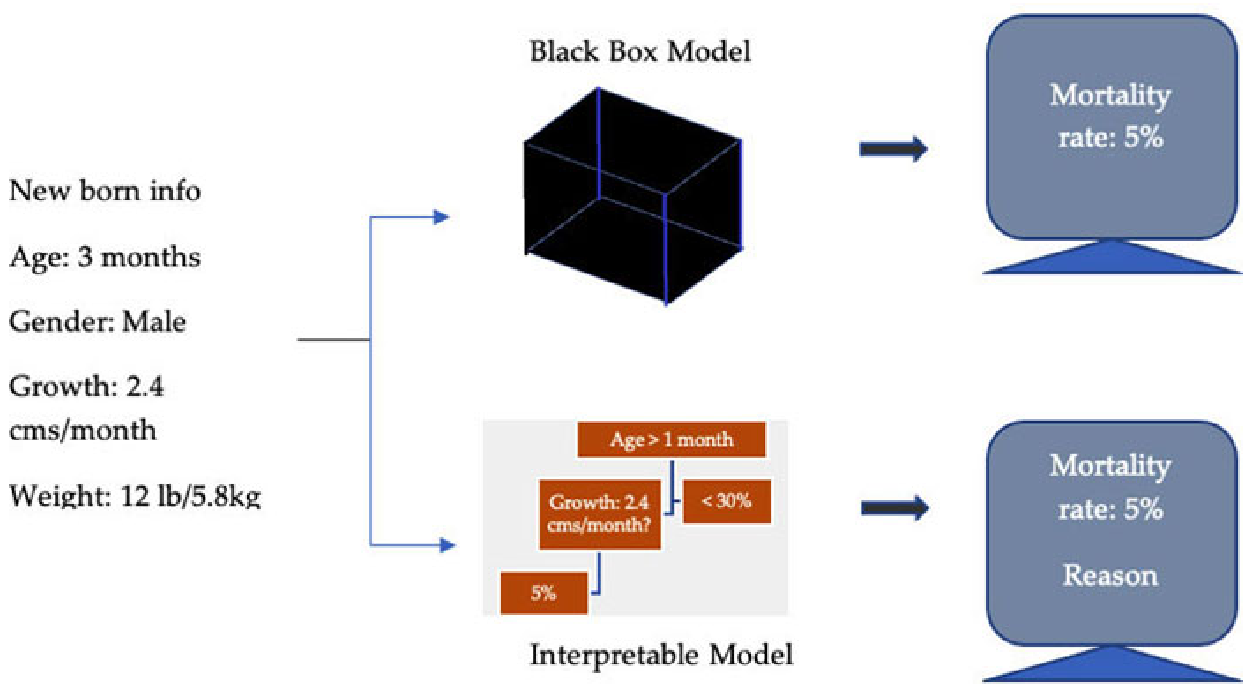


Figure 4.1: *Model Interpretability*

Reliability of the system

AI systems must be reliable and are expected to make consistent and repeatable predictions and decisions.

End-to-end security

The entire ecosystem of AI systems must be secure and protected from threats and breaches. The ecosystem would include the data, the model, the application, the AI-powered system, and AI tools that may operate on third-party cloud systems or use IT infrastructure services.

Who is accountable

Within the setup or any organization, the responsibilities of various individuals or teams, or groups must be clearly assigned. This helps in clarity of accountability when it comes to implications on the use and misuse of the AI models.

How is it beneficial

While developing an AI-powered system, it's important to consider the benefits it brings to sustainability, the environment, and other common factors across humans.

Privacy and surveillance

This includes transparency in the usage of data being collected and what data is being collected to design and train AI solutions. Measures must also be taken to protect data privacy and choice of its manageability.

Human involvement

This includes coupling human intervention in AI model operations in case of higher levels of ethical risks.

Policies and regulations

At every stage of an AI project, all stakeholders must comply with the policies and relevant regulations.

Unbias

AI system must be designed in such a way that it shows fairness towards each sect of society in all situations.

Safety and security

AI must be designed, built, and operated bringing no sense of threat to the physical safety, security, or mental integrity of society at large.

Enterprises and use of AI technology

Enterprises face several ethical challenges in their use of AI technology. A few of them are stated as follows.

Debugging

Major components of an AI system are a source of data, data processes, algorithms, and outcomes.

As such, when things go wrong or are not as expected, organizations have the complex task of tracing the undesirable element in data, processes, or algorithms to explain the root cause of the problem and fix it.

Responsibility

Still, an open topic to be concluded on who is held accountable and owns responsibility in case decisions made by AI systems get harmful consequences including loss of lives, capital, or health. Hence regulations and policies along with skilled and informed lawyers are asked for organizations dealing in AI systems.

Another thought is to find a balance in the cases where AI systems may be safer than human activities it is performing but at the same time cause much lesser harm as compared. For example, in the case of self-driving cars, autonomous driving systems may cause fatalities but far fewer than people do.

No-bias

The AI systems sourcing personal data with identifiable information, must not display any traces of discrimination based on ethnicity, race, or gender. As such the decisions making capability of the system must treat all data sets with fairness and no biases.

In data sets involving personally identifiable information, it is extremely important to ensure that there are no biases in terms of race, gender, or ethnicity.

Purpose and misuse

The organizations must have checks in place that the AI algorithms are not being used for purposes other than those they have been developed for. Hence the design stage is very important to be analyzed for all aspects to minimize the risks and introduce safety measures.

What is the AI code of ethics in enterprises

The **code of ethics** is best described as a set of values that guides the behavior and decision-making process. The purpose is to provide guidance to stakeholders when they are challenged with an ethical decision regarding the use of AI.

Ethical AI primarily requires addressing the following key areas of the code of conduct:

- **Regulatory code of conduct**: A regulatory code of ethics is a framework that organizations are legally bound to follow. This also involves developing a framework for driving standardization and establishing regulations. Ethical designs and ethical deployments with ethical intentions may or may not always be sufficient. Ethical AI policies, therefore, help address legal issues when outcomes are wrong. The effectiveness of incorporating AI policies into the code of conduct only depends on the intentions of the employees and their will to follow the rules.
- **Educating and preparing stakeholders**: Understanding policies, and key considerations, and being aware of the negative impacts of unethical AI and the use of fake data must be understood by all

employees, architects, data scientists, designers, consumers, and other stakeholders. However, there must be a well-thought tradeoff between ease of use around data sharing and AI automation and the potential unintended consequence of oversharing and adverse automation.

- **Design and technology:** AI systems need to use models that automatically detect fake data and comprehend unethical behavior. For example, deep fake videos to malign images and propagate agendas need to be checked as more and more AI tools are commoditized and are available to the public. Organizations, therefore, need to invest in protective measures ensuring these are an integral part of open, transparent, and trusted AI infrastructure.

AI: Ethics, Bias, and Trust

Future of Life Institute organized **The Asilomar Conference on Beneficial AI** conference which was held on held January 5–8, 2017, at the Asilomar Conference Grounds in California. It was attended by more than 100 thought leaders and researchers in economics, law, ethics, and philosophy, The agenda of the conference was to address and formulate principles of beneficial AI. The outcome of the conference was the creation of a set of guidelines for AI termed "*the 23 Asilomar AI Principles*".

The 23 Asilomar AI principles

Figure 4.2 describes the tenets of Asilomar principles:

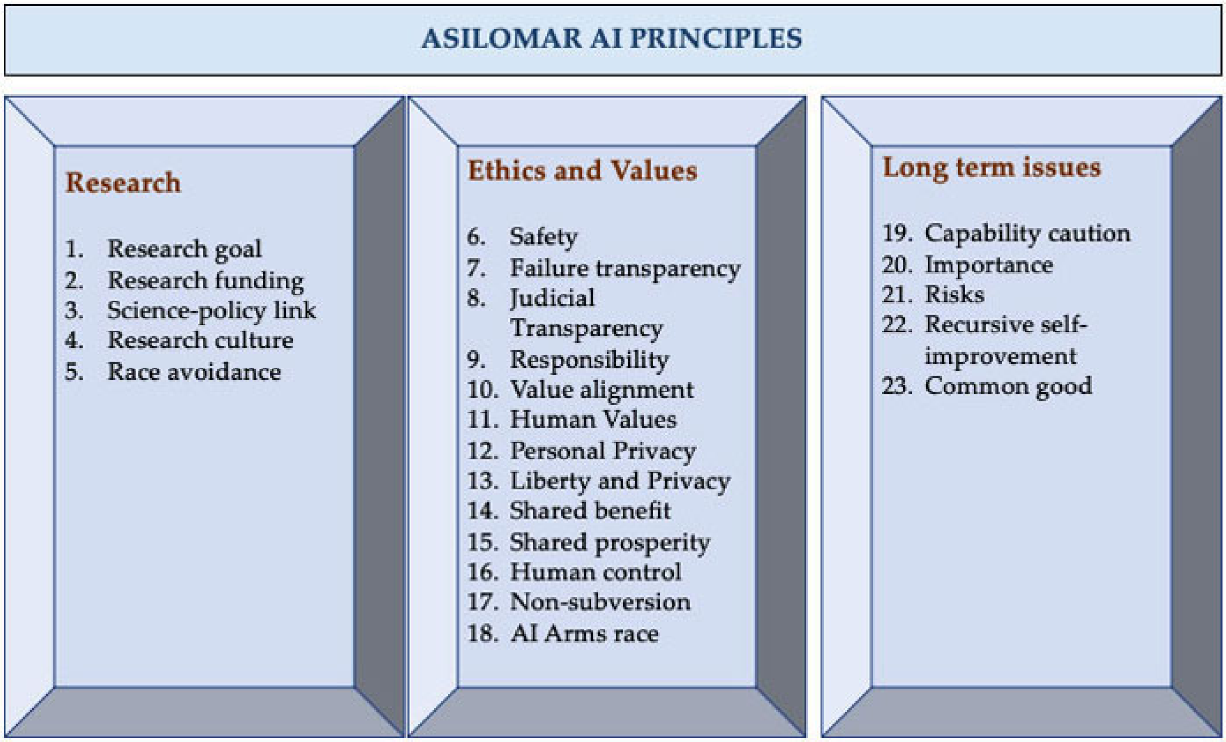


Figure 4.2: *Asilomar AI principles*

Research issues

Let us look at research issues further in the following:

Research goal:

The goal of AI research should be to create beneficial intelligence and not an undirected one.

Research Funding: Investments in AI should not only be accompanied by funding for research on ensuring its beneficial use, but also address questions related to economics, law, ethics, and social studies, such as:

- How to make future AI systems robust enough, so that they behave as per the aim of developing them without malfunctioning or getting hacked?
- How to achieve growth via automation while maintaining purpose?
- How could legal systems be updated to be more efficient, keeping pace with AI, and managing the risks associated with AI?
- What legal and ethical status should AI systems have?

Science-policy link:

AI researchers and policy-makers need to work hand in hand and have constructive and healthy exchanges of ideas and workings.

Research culture:

The working dynamics involving AI researchers and developers of AI must encourage a culture of cooperation, trust, and transparency.

Race avoidance:

There must be no compromise on safety standards due to teams developing AI systems not adhering to active cooperation.

Ethics and values

Now, we will discuss the ethics and values of Asilomer AI principles:

Safety:

AI systems should be designed such that the data, process, and systems in their entirety are safe, secure, and verifiable whenever applicable throughout their operations lifetime.

Failure transparency:

AI systems should be transparent enough such that it is possible to ascertain why any failure was caused, in case of one.

Judicial transparency:

In case of AI autonomous systems are involved in judicial decision-making, the system should provide a satisfactory explanation of the decision taken that is auditable by a competent human authority

Responsibility:

AI systems' designers and builders are to be held responsible in case of any moral implications of the use and misuse of these systems. Also, these stakeholders must take the opportunity to shape the implications.

Value alignment:

Goals and behaviors throughout the operational lifetime of highly autonomous AI systems should be assured to align with human values.

Human values:

AI systems should be designed and operated in such a way that they are aligned with human values such as ideals, dignity, rights, freedoms, and cultural diversity.

Personal privacy:

AI systems analyze the data generated by any human/user. This data must be transparently accessed, managed, and controlled by those who generate it.

Liberty and privacy:

Personal data fed to AI systems must not hinder the liberty, real or perceived, of humans.

Shared benefit:

AI technologies should benefit globally and empower communities, groups, or as many people as possible, irrespective of those who use it or those who do not.

Shared prosperity:

AI should benefit all humanity and help in economic prosperity.

Human control:

To accomplish human-chosen objectives, they control what to, whether to, and how much to delegate decision-making to AI systems.

Non-subversion:

Highly advanced AI systems must be designed and operated to improve rather than subvert the social and civic processes driving the health of society

AI Arms Race:

The race of owning Lethal autonomous weapons must be completely avoided in the interest of the human race.

Longer-term issues

The following are the long-term issues of Asilomer AI principles:

Capability caution:

No consensus has yet been established regarding future AI capabilities. Hence strong assumptions regarding the same should be avoided.

Importance:

Autonomous AI systems can have profound changes on the human race and life and hence should be planned for and managed with care and good intent.

Risks:

AI systems are expected to pose catastrophic or existential risks. As such, planning and mitigation efforts must match the degree of such impacts.

Recursive self-improvement:

Strict safety and control measures must be adopted for AI systems designed to recursively self-improve or self-replicate in a manner leading to rapidly increasing quality or quantity.

Common good:

AI systems that are defined as autonomous or super intelligent must be developed to achieve benefits for the entire humanity.

Employment and AI

Let us consider AI-powered applications and their impact on human employment.

Bias in recruitment

This is a classic example of how the data input itself can bring bias in the models. In a bid to "*out-recruit*" other technology firms, Amazon built an AI-based tool.

The models were trained to filter out top talents' resumes. The model was trained based on data collected over 10 years. However, the data collected was tainted as the majority of the candidates were men. This resulted in the AI model giving higher priority to male resumes while assigning low scores to the resumes that had participated in women's activities, such as "*State-level Women's cricket team player*". This was witnessed even when the names were anonymized.

Amazon gave up and disbanded the tool and the team after multiple failed attempts to make the program gender-neutral.

AI applications replacing humans

ChatGPT, an AI assistant, as also covered in previous sections is doing a faster and better job than humans, especially in content creation, making it very easy for enterprises to replace humans.

ChatGPT can also write entire computer code to program applications with accuracy. This is encouraging companies to use it instead of employing coders.

There are AI-powered applications that are capable of representing a lawyer in court for arguments. Thus, even the legal industry will see the impact of the adoption of AI technology and AI-powered systems.

The fear is thus not imaginary. Humans are losing their jobs with the advent of AI technology. It is anticipated that AI will have a huge impact on human employment but in certain fields only. AI, in its current state, is unable to replace humans, that is, to reach human intelligence.

Conclusion

AI-powered systems will have a deep impact on our society with the accelerated adoption and frequent use of AI applications for smallest of the task.

The impact of these AI-powered systems will not be limited to simple outputs, instead, these systems have the capacity to add on to the existing biases or the power to reduce the same. It, therefore, is very important, for such systems to be based on ethics and adopt an unbiased approach.

One must spot issues within the data used for training and testing such systems. Also, make healthy arguments towards feature functionality and expected outcomes and apply policies of ethics and bias removals while designing such systems.

In the next chapter, we will be introduced to storytelling.

Multiple choice questions

1. **How many guidelines are there in Asilomar principles to make AI ethical and safe are:**
    a. 46
    b. 30

c. 23
d. 21

2. **Can AI make ethical decisions?**
    a. YES
    b. NO
    c. Somewhat
    d. There are no ethics in AI
3. **Is AI capable of replacing humans for their jobs in enterprises**
    a. Yes, in all possible fields of work
    b. No, AI and employment are not related
    c. As of date yes, in some industries
    d. It can never replace humans

Answers

1. **c**
2. **a**
3. **c**

Questions

1. Name any of the three Asilomar principles for ethical AI.
2. What are practices adopted in enterprises towards the development of ethical AI?
3. Name any three AI biases.

Join our book's Discord space

Join the book's Discord Workspace for Latest updates, Offers, Tech happenings around the world, New Release and Sessions with the Authors:

https://discord.bpbonline.com

CHAPTER 5
Introduction to Storytelling

Introduction

The objective of this chapter is to understand the role of storytelling in AI projects. Not all great projects see the light of day due to non-acceptance, whether technical or the overall need. This is mainly due to the way project details are presented to the audience and human decision-makers.

This chapter will help you build convincing stories around technology in AI projects.

Structure

In this chapter, we will be discussing:

- Storytelling: Communication across the ages
  - Significance of storytelling in data storytelling
- The need for storytelling
- Storytelling with data
  - By the numbers: How to tell a great story with your data
- Conflict and resolution
  - Everyone wants to resolve conflict, and a good data storyteller is there to help!
- Example of a story
- Storytelling for audience
  - Your data storytelling depends on the background knowledge of your audience
- Insights from storytelling

- Make the audience care about the data
- Keep the audience engaged
- Create from the end; present from the beginning
- Start with an anecdote, end with the data
- Build suspense, not surprise

Storytelling: communication across the ages

Human beings have socially evolved over the ages with the need to communicate with each other to share knowledge that helped increase chances of survival. Communication happened even before languages were developed. With the advent of oral communication across various communities came storytelling as one of the oldest traditions. Storytelling apparently is older than a written form of communication.

Let us take the example of the "*Telephone game,*" where the purpose is to send the message from the first person to the last person in the group without distortion or any change in the message in any way. This game throws a challenge to preserve information without losing its context, irrespective of the number of times it has been passed around.

This challenge can be overcome simply by adding structure to information.

Significance of storytelling in data storytelling

Humans are continuously creating stories either in real-world events or in their imaginations. They tell bedtime stories to kids, motivational stories, stories with moral messages, and much more. Stories are instrumental in cognitive, social, and emotional maturity.

Stories told within communities, families, and groups across different cultures become even more powerful as they are shared generation after generation. These then start representing cultures, groups of people, or families and become part of them, their belief systems, and how they perceive the future.

Let us consider sharing significant information in a discrete form with a group of people. How long do you expect the information to be registered in the memories of these individuals?

What if there is a story that can connect all the available significant information and is then presented to the group? Will they be able to form a connection with the story? Will they be able to remember it for a long time? Does the story invoke any feeling or emotion within?

Let us accept stories last a lifetime because of their influence on us. Stories are good for remembering content.

The same goes for the business context.

Imagine being part of a story; it is like thinking and forcing yourself to remember all the information, and thus end up trying to make contributions to the same story.

If a model is exceptional in its predictions, developed by expert developers and data scientists, it still would need professionals from various backgrounds to make it a success. That is in addition to great standards of experimentation and implementation, finding the right audience for the project to get approved and finally be adopted by intended users depends on the way the technology is presented for it to gain attention and be accepted finally.

The need for storytelling

Stories express the culture, convey history, and reveal values that unite people. The stories any country, community, or families hold in common are an important part of the ties that bind them.

Stories can come in different forms. These can be either verbal or expressed as visuals. Stories can either be dynamic or static – that stays for a long time. Dynamic ones can be heard or experienced, while static ones are written or recorded such that they can be revisited over again and again. Every story counts; imagine the story the picture in *Figure 5.1* is trying to convey:

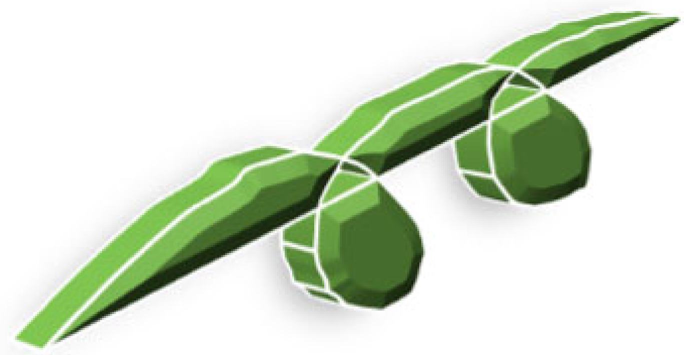


Figure 5.1: *Every story counts*

Let us understand in detail the need for storytelling:

- **Stories help us cope**

  We relate to and make sense of our life experiences in a part of the stories we learn or tell ourselves. People think through how they would respond to disasters and threats by watching scary movies or reading books or listening to podcasts. It prepares them for worst-case scenarios.

  Similarly, exploring and sharing personal experiences in the form of stories and reflecting on them can give different perspectives. It may make us vulnerable by continuing to feel in victim mode or healing by changing perspectives.

  Writing down traumatic experiences helps improve the physical and mental being of humans. As such storytelling, irrespective of the mode of narration establishes empathy and helps us cope with emotions.

- **Stories help in learning, remembering, and imagining**

  It is easier to learn a moral, remember the best course of action, and think through options in personal experiences based on the stories heard. The stories taken forward through generations also evolve based on imaginations and modifications thereby. These create a fictional world and allow us to imagine and find solutions to problems.

- **Stories help in connecting**

Members of communities, groups, or families feel connected based on stories shared within them. The world comes together in natural calamities like tsunamis and earthquakes and pandemics based on stories shared via media.

- **Stories help in solving problems**

  Stories help in creating opportunities to try different plots or actions to give different storylines and, eventually different outcomes. If there is a co-creator of a story, it can lead to different suggestions, actions, and twists that would help in solving problems.

- **Stories occupy us**

  Being in a public place and engaged in reading a book or watching TV, or listening to songs keeps us engaged. It is easy to have a story that engages our attention with the ease of accessibility and availability on electronic devices.

- **Stories help us understand others**

  Stories can help us learn about people and situations and help us empathize with and understand others. It is easier to make social connections with strangers based on their stories and emotions invoked within us.

- **Stories travel down generations**

  Stories that stick with generations help preserve cultures, core values, and identities thereby. It also helps different generations to relate to their past and their evolution.

- **Stories have something for everyone**

  Humans perceive stories in different ways. Few will learn or grasp best from visuals, and few from audio and kinaesthetic learners who learn from doing, experiencing, and feelings.

  Visual learners will appreciate mental pictures evoked, Auditory learners focus on content and voice and kinaesthetic learners remember the emotional connections and feelings.

  Nonetheless, a story has something for everyone.

Storytelling with data

During a 2009 interview, Google's Chief Economist *Dr. Hal R.Varian* stated, "*The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades*."

Organizations generate and collect data at tremendous rates. This data is a combination of both meaningful and additional data which can be discarded. The next elusive step is to extract significant information from the data that can be utilized for revenue growth, improving customer experience, product and service feedback, sales and marketing, and various other ways.

As such the collected data has great potential. *Figure 5.2* describes various elements of data storytelling:

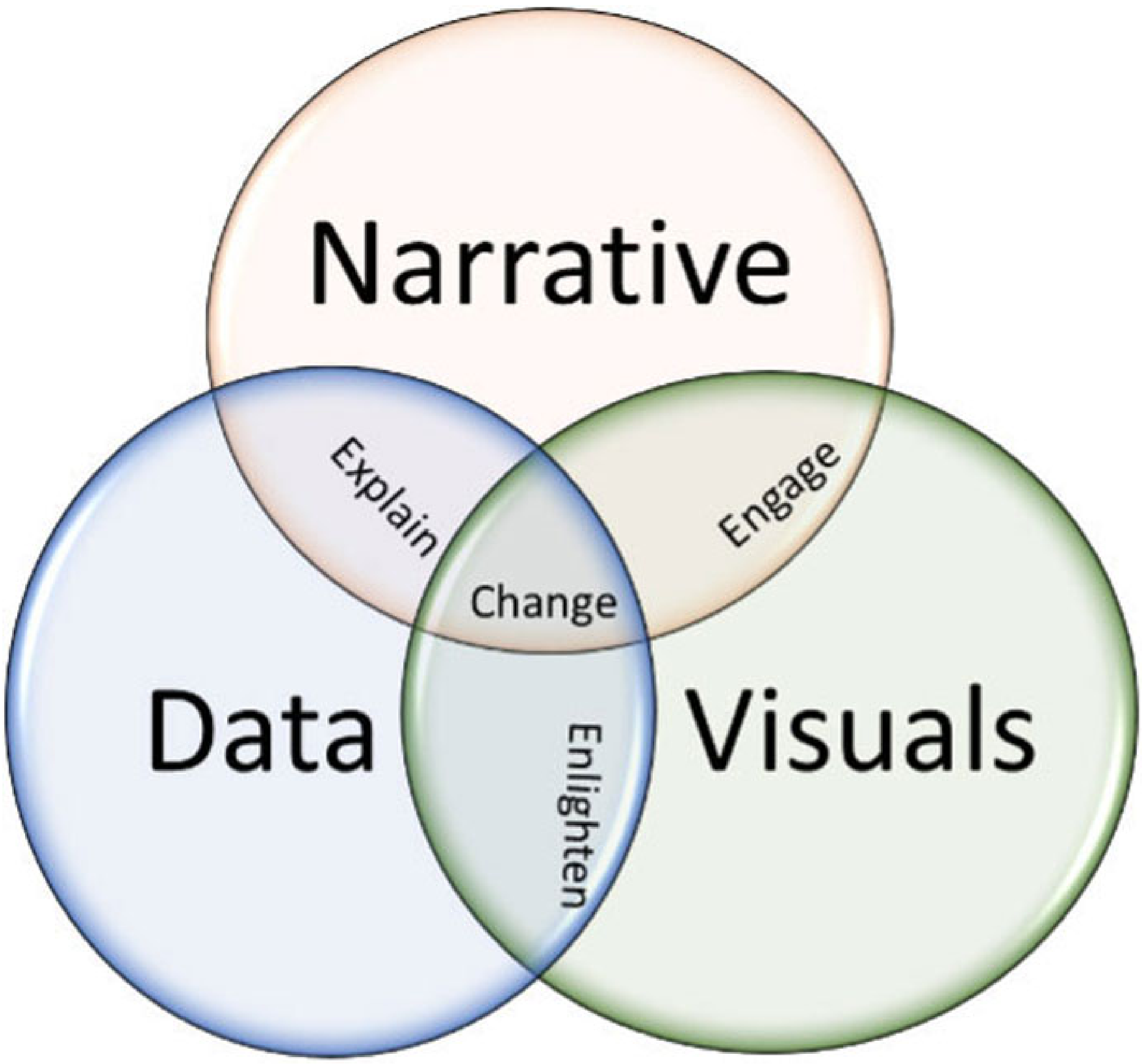


Figure 5.2: Elements of Data Storytelling

A data story is a narrative developed around a set of data, such that the story frames it into a context and outlines the broader implications.

While business intelligence and data science emphasize on technical aspects of presenting data insights, a data story factors qualitative analysis and domain expertise together with the data insights for a better understanding of organizational goals and objectives.

Data storytelling is the skill to narrate a story around the data which is put into context and presented to an audience. The data set is leveraged to craft a narrative by utilizing data analysis, statistics, data visualization, contextual analysis, qualitative analysis, and presentation.

By the numbers: how to tell a great story with your data

The data storytelling objective is to help an audience, regardless of their background, expertise, or technical skills, understand the narration and its implications.

Special efforts need to be made while making a narrative from the data to address all kinds of audiences – one who learns from audio, others who learn from visuals, and/or who learn from practical tasks.

In general, good data storytelling includes combinations of various auditory, visual, and/or practical case study aspects to benefit diverse audiences and keep them engaged.

There are several key components that can be applied to any successful data story in general regardless of the vastness of the technical and domain aspects of the data storytelling:

Understand data context

The first step in data storytelling is to determine the insights the data is providing or what the data is telling. Elements to understand in a data context are described in *Figure 5.3*:

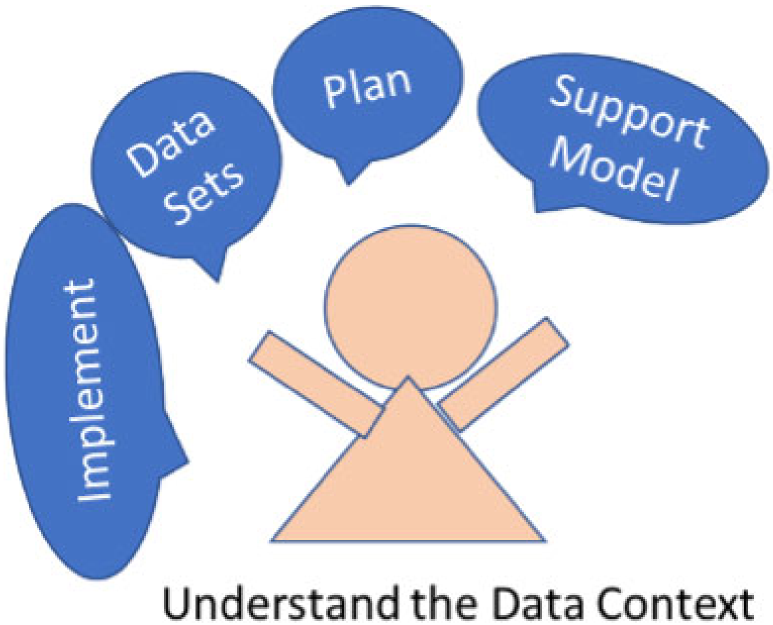


Figure 5.3: *Understand the data context*

For instance, a correlation in a data set may provide a starting point for building the data story by forming a meaningful narrative. The story should be aimed at a diverse audience from all domains.

Audience is the key

Data storytelling is narrated keeping in mind the audience. For example, an executive team will be more interested in insights impacting business revenue, a data scientist will be eager to learn about the statistical aspects, and a designer will be more interested in the design aspects of the system consuming the data.

Data that matters

Determine the me4beerfost relevant data that provided insights being touched upon in data storytelling. The data available can be overwhelming, however, identifying relevant data to extract maximum insight into the use case. For example, for customer experience, relevant data needed revolves around product usage, sales, customer feedback, and market demand. Focusing on relevant data ensures that the key points do not get diluted and that listeners or readers do not get distracted.

Data analysis and insights

After identifying the most relevant data, the following step is to perform analysis on it to find insights that help create meaning out of the data and provide general themes for the data story that evolves around the relationships between data sets and data patterns.

Effective data visualizations

There are many different types of data visualizations as already covered in previous sections of this book. As described in *Figure 5.4*, there are multiple data visualizations to choose from.

Figure 5.4: *Data Visualization*

It is very important to pick one of the most effective ones to provide the audience with the visuals that best represent the data, leave the desired impact on the audience and convey the correct information.

Set the context

Along with great insights comes the need to craft the right narrative, making it understandable to the audience while keeping them engaged. It's critical to weave the context and data insights together. For example, adding to context the business reality along with data insights may help the audience appreciate the relevance of it all and understand the broader perspective.

About the story

A great story would have a structure, a purpose, a character to cheer for, and an appeal to the deepest emotions. A great story would be surprising, unexpected, simple, and focused.

In terms of data story, it's important to create a framework that the audience is familiar with and such that information they get engaged and it's easier to consume the information.

A traditional data story would often provide an introduction to the plot, a platform that builds through the problems, an apotheosis of the crucial moments or insight, a strategy to resolve the problems, and ultimately retrospection and resolution.

Clear and concise

It's advised to keep the data story concise and clear. Adding extra fluff to the story will only distract the audience and they may end up losing interest in missing significant information. *Figure 5.5* symbolizes an effective story:

Figure 5.5: *Tell a story*

> **"If I had more time, I would have written a shorter letter."**
>
> *- Blaise Pascal, who laid the foundation for the modern theory of probabilities.*

As such, data storytelling involves continuous review and editing of the narrative to gain extreme clarity and gain focus on what matters.

Conflict and resolution

Identifying a conflict is one of the major steps toward working on its resolution. A conflict can be identified by first accepting that it exists. Groups may be having conflicts, but breaking the conflict in such a way that the major points of the conflicts can be categorically identified helps in resolution.

There are a few strategies to resolve conflicts. These are as follows:

- **Avoiding**: Avoid taking action and hoping a conflict will resolve itself or dissipate.
- **Accommodating**: Satisfy the other at the expense of own needs or desires.
- **Compromising**: Find an acceptable resolution that may partly, but not entirely, satisfy all parties involved.
- **Competing**: Satisfy own desires at the expense of others.
- **Collaborating**: Find a solution that entirely satisfies the concerns of all parties involved.

Everyone wants to resolve conflict, and a good data storyteller is there to help!

The four key elements of a data story are characters, conflict, emotion, and a resolution or closure.

Conflict

A good data storyteller clearly defines the problems to be addressed, and their significance. The data is used to demonstrate the problem which may be obvious or unknown. The lack of knowledge of the problem can be the problem itself.

Character

After identifying the issue, the aim of the data storyteller is to think about characters in a bid to humanize the problem. This means, identifying real-world people who are impacted by the issue. Business problems are not just reports and numbers, but these real individuals who are affected by the situation. These characters could be stakeholders like employees, customers, the general public, and investors.

Emotions

A data storyteller must make the story compelling and further engaging by mentioning the emotions arising out of the conflicts. These emotions are created within the characters identified as stakeholders. For example, investors could be scared, employees demotivated, and customers dissatisfied.

This also includes emotions felt post the closure or resolution of the conflict. For example, the above stakeholders' post-conflict resolution has investors feeling elated, employees being happy, and customers satisfied.

Closure or resolution

A good story must have the conflict resolved and closure. This can be achieved by taking appropriate actions. For example, a change in policy, the use of machine learning and AI-powered systems, and a change in investments. Eliminating the conflict to restore balance needs domain expertise and inputs from various stakeholders of the story narration.

Example of a story

Figure 5.6 shows a typical framework for data storytelling. Let's take an example using this framework.

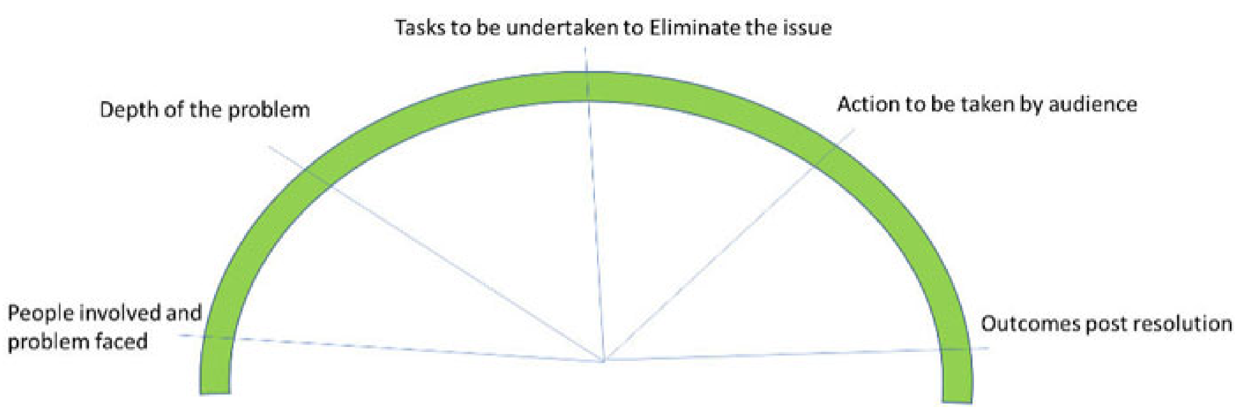


Figure 5.6: *Framework for storytelling*

Let us consider an ice cream parlor that runs out of particular flavors frequently.

The parlor can state the problem as "*We run out of chocolate flavor frequently. A demand forecasting model can help mitigate the problem*".

The problem statement has clarity; however, it lacks the stakeholders and emotions and, thereby, a good story to engage attention.

See how we can build a story around it.

"*Rajiv's kid had an unplanned visit from his friends one evening post their playtime. Rajiv did not waste time asking kids their favorite ice cream flavor. He immediately rushed to the ice cream parlor to buy chocolate-flavored ice cream, which was one of the favorites of all the kids in the group.*

Rajiv had a rude shock when the executive at the ice cream parlor had to cut a sorry figure stating the parlor had run out of flavor for the day.

Rajiv lost his valuable time, became a dissatisfied customer, explored options satisfying his needs, and might never visit this parlor again.

The executive at the parlor couldn't do his sales and lost his commission for no fault of his, which otherwise he deserved. This has been happening to him quite frequently. This leaves him disgruntled and leads to employee churn.

The parlor struggles with lost revenue, and customer and employee churn.

To resolve this problem, a demand forecasting tool was used. The tool predicted the demand for various flavors at different times based on population, festivals, and so on. Thereafter, this problematic situation never occurred.

The customers were happy, executives made commissions and the parlor continued with its revenue"

The above story is narrated by the parlor manager, salesperson, and product manager to the directors, and technical and data teams who collectively decide on the resolution and implementation of the solution for a closure to the problem.

The preceding story, clearly mentions the characters, the problem, the resolution, and the emotions. The story gives a human connection to the problem and hence the audience, the stakeholders, cannot ignore it anymore.

Storytelling for audience

> **“Perhaps the most difficult data storytelling skill to master is empathy — to understand where the audience is coming from and which parts of the data analysis they’ll react to”**
>
> *- Miro Kazakoff, Lecturer, MIT Solan*

The audience is key in storytelling. A good storyteller must be able to connect with the audience and keep their attention engaged throughout to get emotional responses. For this it is very important to know the following:

- Who is the audience?
- What is the relevance of the story to the audience?
- Does the story solve a problem the audience cares about?
- Has the audience heard the story before or is it something intriguing to them?

Your data storytelling depends on the background knowledge of your audience

The impact the data storytelling has on the audience is majorly dependent on the background of the audience. A great story otherwise may have no outcome of relevance if the audience has no stake or is from a different background than the context in which the story is narrated.

As such, the audience’s age, demographics, skill, and domain expertise influence how they understand and respond to data stories. These factors decide how a storyteller should decide the narrative.

A company board of directors may need a business story around its data than technical details on how to provide data cleansing to get more insights. A data scientist should be able to understand the models needed from the narration.

Hence the narration of the story needs to be customized and edited as per the audience's levels of understanding and the key takeaways expected of them.

Insights from storytelling

Data insights are the secret sauce to a great story. There may be several insights and supporting points to demonstrate and explain the unifying idea around which the story evolved. However, an insight boils down to an understanding that wasn't already and that is relevant to the problem at hand. Insights provide new perspectives, and with relevance equip the audience with the confidence and conviction to take meaningful action.

Insights into numbers, figures, facts, and collaterals and much more help in believing the story to be true. The summary must address:

- A conflict or a problem that the audience can relate to.
- An approach or strategy that the audience thought is a distant dream.
- A solution or an outcome to make the audience believe in the possibilities of a resolution and a happy ending.
- A narrative that binds the above three elements together and keeps the audience engaged and convinces them to take action required at their end. *Figure 5.7* describes various stages in the structure of data storytelling:

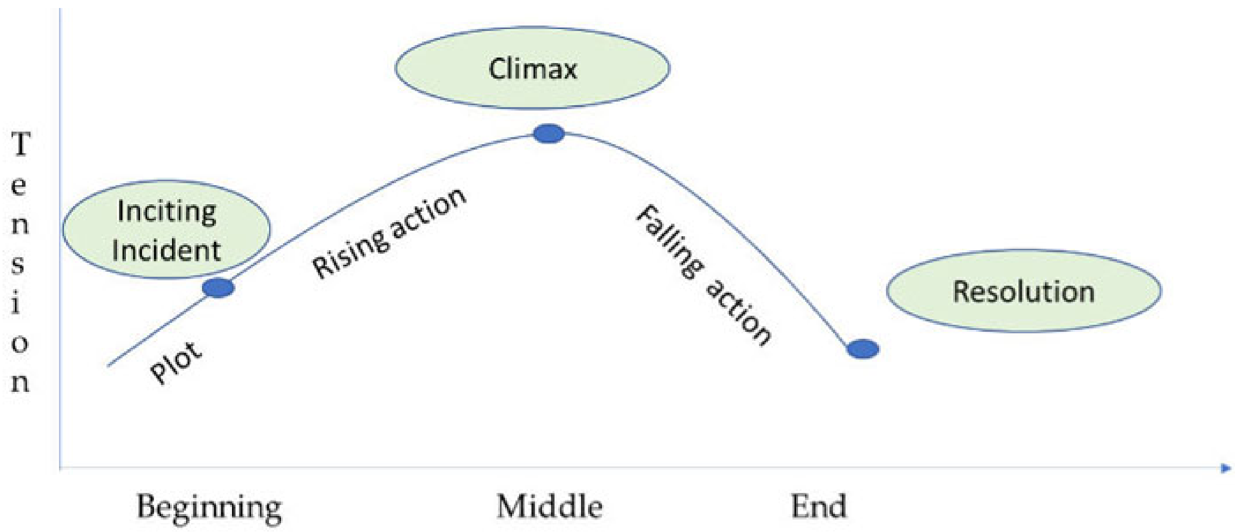


Figure 5.7: *Structure of data storytelling*

Make the audience care about the data

"If you want people to make the right decisions with data, you have to get in their head in a way they understand."

- Miro Kazakoff, Lecturer at MIT Sloan

The best way is to present the data in such a way that intrigues the audience enough to grasp it and have their attention through the narration. They need to relate the data and its relation to the problem at hand.

The data storyteller is expected to understand thoroughly that a 360-degree view of the business dashboard for a CEO will be fundamentally different from an operational report for a sales manager.

Keep the audience engaged

> **"The skill of data storytelling is removing the noise and focusing people's attention on the key insights"**
>
> *- Brent Dykes, Data Strategy Consultant*

Create from the end; present from the beginning

> **"A story is a chain that begins at one place and ends at another"**
>
> *- Poet Laureate Randall Jarrell*

Stories have a beginning, a middle, and an end – but not necessarily in that order. To stay focused, it is advised to write the end first and then work towards the end goal. However, while storytelling, start from the beginning and then approach the end.

Start with an anecdote, end with the data

Start with the end goal so you know what you are going to say. Use the data that supports that "end narrative."

Build suspense, not surprise

Building suspense involves withholding information or insight while raising key questions. Asking questions prompts the audience to think and get involved. Surprises may confuse them and break their own thoughts.

Conclusion

Great data storytelling expertise is an increasingly significant skill to possess in a world exploding to the seams with data. Without the expertise, the ability to construct visual narratives and deliver crucial insights to the audience stands at risk of being reduced drastically or even missed entirely. To connect emotionally and motivate to bring change depends on the way of communication that makes sense.

Figure 5.8 visualizes the data storytelling being the peak of an iceberg with much beneath the surface, symbolizing the mammoth of tasks that are done to achieve a compelling data story:

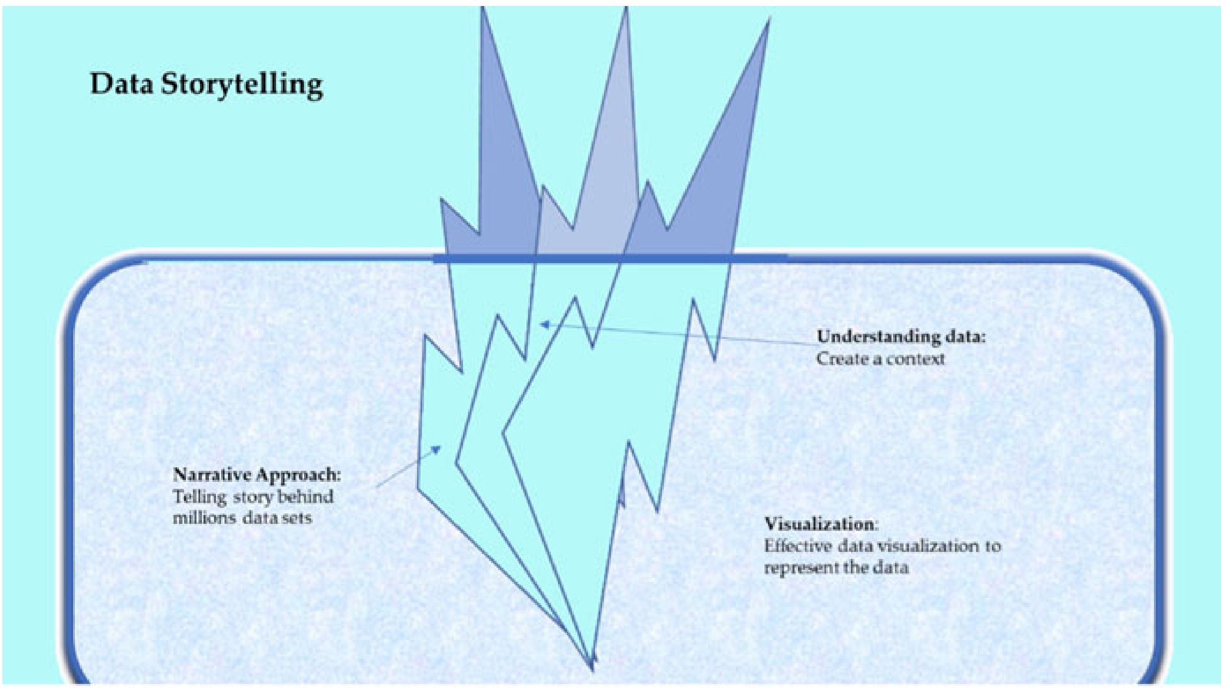


Figure 5.8: *Tip of the iceberg and what's beneath*

Richard Kearney, Professor of Philosophy at Boston College, in his book, *On Stories*, said, "*Telling stories is as basic to human beings as eating. More so, in fact, for while food makes us live, stories are what make our lives worth living*." Are you ready to tell your story?

In the next chapter, we will be learning the design thinking framework and its elements.

Multiple choice questions

1. **What are the elements of good data storytelling?**

a. Data
b. Narrative
c. Visuals
d. All of the above

2. **What are the three major stages of storytelling?**
    a. Inciting incident
    b. Climax
    c. Resolution
    d. All of above
3. **Can a data story compiled on the performance of the class teachers for their appraisal is relevant to be presented to the students?**
    a. Yes
    b. No
    c. Maybe
    d. Depends on the school

Answers

1. **d**
2. **d**
3. **b**

Questions

1. What is the need for data storytelling
2. Can a data story compiled for a certain audience be told to a totally different set of audiences with a different set of skills and profiles? Explain with an example
3. Take the attendance of students in your class over the past year. Compile a story around the number of days, occurrence, and frequency of leaves taken by each student.

Join our book's Discord space

Join the book's Discord Workspace for Latest updates, Offers, Tech happenings around the world, New Release and Sessions with the Authors:

https://discord.bpbonline.com

CHAPTER 6
Critical and Creative Thinking

Introduction

Systems must be designed in a highly consumer-centric way, that is, to provide the best user experience and address users' needs. It is very important for systems' designers to get into iterative measures from the design to testing phases to incorporate feedback back into the design and repeat to solve complex problems.

The **design thinking framework** provides an approach for designing systems by questioning all aspects, including the problem, the assumptions, and the implications. Consequently, tackling all possible challenges that are imprecise or complex and, thereafter, reframing the problem in human-centric ways.

The objective of this chapter is to understand design thinking methodology and incorporate its processes while developing systems.

Structure

In this chapter, we will be discussing:

- Design thinking framework
    - Right questioning (5W and 1H)
    - Identifying the problem to solve
    - Ideate

Design thinking framework

Design thinking is a methodology to solve a problem via a solution-based approach. It's useful to tackle complex and ill-defined problems. It helps adopt a hands-on approach to prototyping and testing and creates numerous ideas during brainstorming sessions. Design thinking serves to reframe the problem in more human-centric ways and understand the human requirements involved. *Figure 6.1* describes the key elements influencing design thinking:

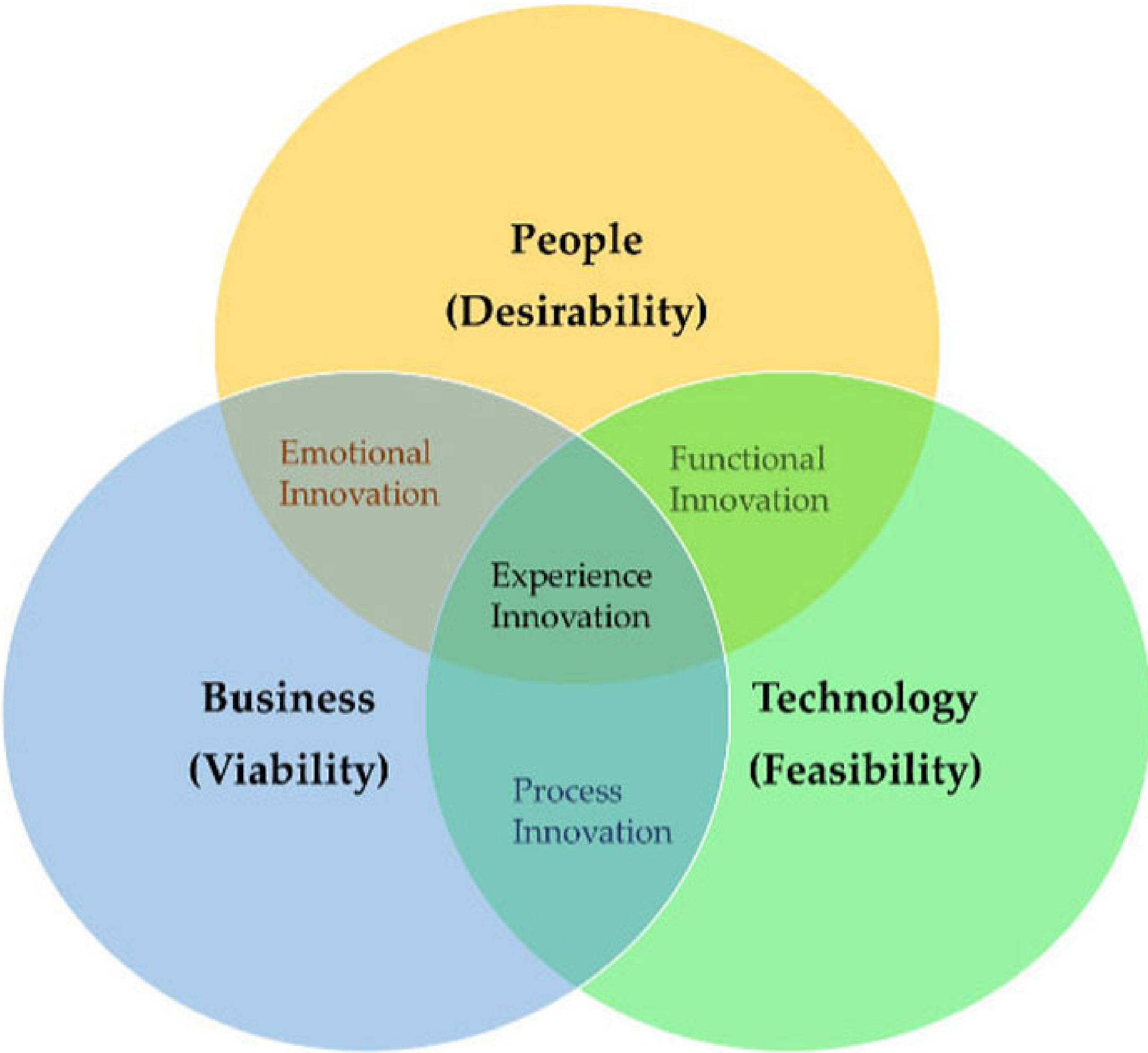


Figure 6.1: *Design Thinking*

Design thinking is a non-linear and iterative process. It focuses on collaboration between designers and users. The main aim of using this framework is to develop innovative solutions based on how real users think, feel, and behave.

There are a variety of design thinking frameworks and visualizations that exist based on what design thinking means to different people.

In this chapter, we cover the most popular human-centered design process that consists of five core stages. These are Empathize, Define, Ideate, Prototype, and Test. Let us learn about these five stages of design thinking, described in *Figure 6.2*:

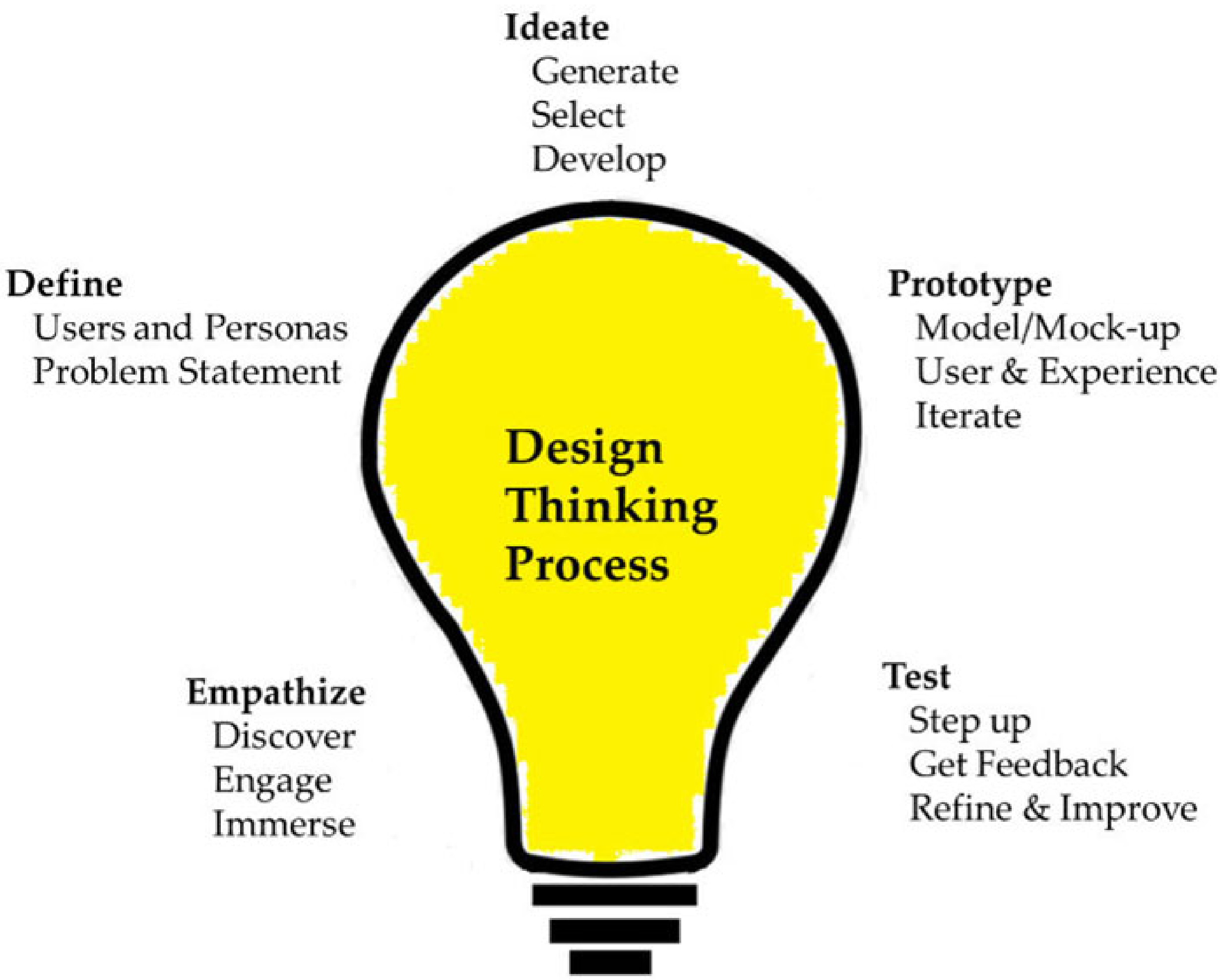


Figure 6.2: *Design thinking process*

- **EMPATHIZE: Gain insight into users' needs**

  A deep focus on the needs of the humans involved, from the entire spectrum of non-users who are impacted by the solution and users who interact with it, ensures that the solution is aligned with the individuals, businesses, and society.

- **DEFINE: State the users' needs and problems**

  This stage helps in building new perspectives and exploring ways to reframe the problem at hand. This allows a comprehensive approach toward reaching a desired solution.

- **IDEATE: Employ assumptions and create ideas.**

  This stage allows solution designers to ideate in an open and judgment-free space to generate and explore as many solutions as possible. Considering all possible challenges and assumptions helps generate a broad spectrum of ideas and possibilities, achieving a complete set of solution features.

- **PROTOTYPE: start to develop solutions.**

  This stage involves refining the feature list and creating meaningful scaled-down solutions to investigate the key solutions generated during the ideation phase. This is the experimental phase with an aim to identify the best possible solution.

  The solutions are implemented within the prototypes and are subjected to investigation before they are approved, rejected, or improved.

  Prototypes help to understand the limitation of the solution and possible clarity on real user behavior and experience when they interact with the solution.

- **TEST: Try the solutions**

  Designers and developers test complete products rigorously using solutions accepted at the prototype stage. Since the framework is iterative, the results are often used to reframe and redefine the problem.

  The loop back to a previous stage is aimed at refining the solutions and making the product desirable and catering to users' needs and requirements. *Figure 6.3* describes the non-linear process of design thinking:

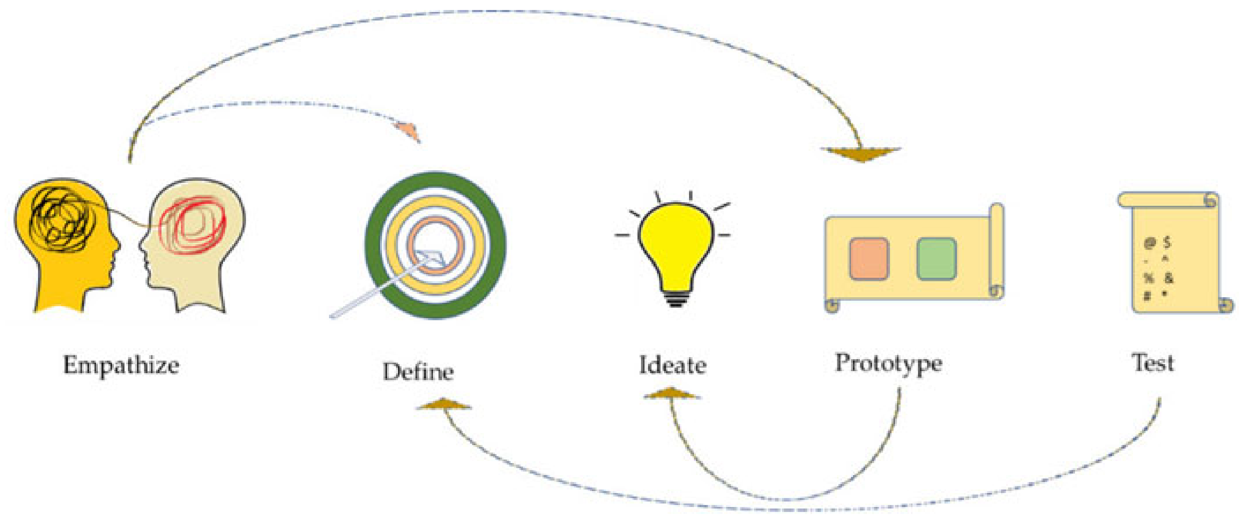


Figure 6.3: *Design thinking is non-linear*

Systems and solutions are developed for communities and large groups and never for individuals. Hence, understanding the collective mental model of target users carries much more significance than any individual user's mental model.

Collective intelligence (**CI**) is shared or group intelligence that transpires from the collaborative and collective efforts of a group of individuals. Group intelligence appears in consensus decision-making and has been utilized successfully for activities associated with designing, notably at the ideation and evaluation stages. *Figure 6.4* describes a few of the contributing elements toward collective intelligence:

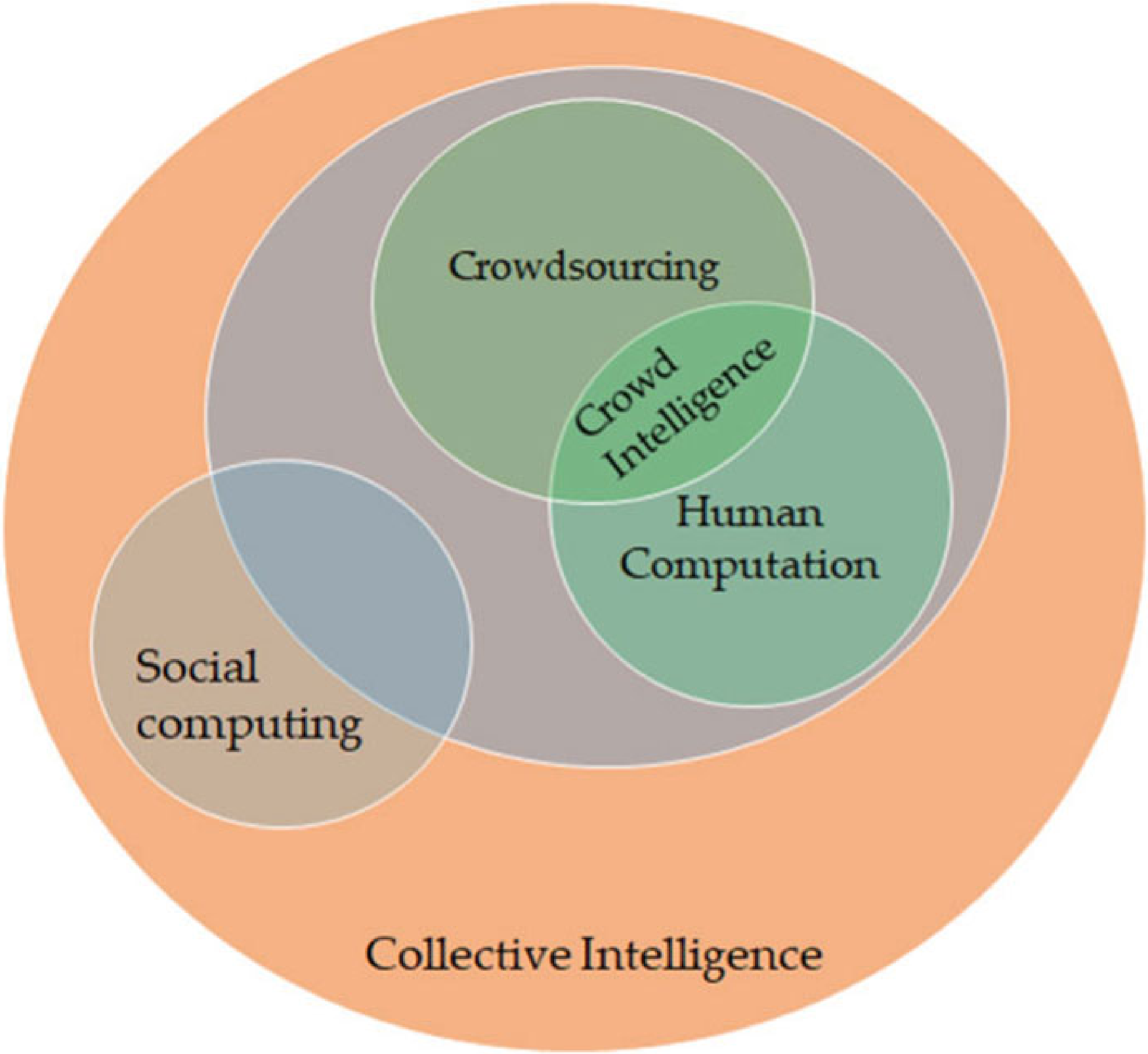


Figure 6.4: *Collective Intelligence*

In recent years CI, facilitated by the ease of use of internet technologies, has proved as a powerful economic and human resource and an inherent effective booster of design awareness. While an individual intelligence can be amplified significantly merely by collaborating with other individuals, it may not even be a direct collaboration, and here the people know each other.

Sir Francis Galton (February 1822 – January 1911) was a polymath and a statistician. In 1906, Galton made his discovery of the "*wisdom of crowds*".

At a farmers' fair in Plymouth, England, a contest was organized with the goal of guessing the weight of an ox. Around 800 people submitted their guesses. Galton found that the average of the guesses of all the entrants was remarkably close to the actual weight of the ox and better than the individual closest guess.

There can be forums for direct CI (also referred to as crowdsourcing) and also indirect CI.

Some of the popular and widely used internet-based Direct CI applications include Wikipedia and GitHub. These involve millions of participants not only learning but also interacting and collaborating with each other.

Indirect CI applications include social media platforms like Facebook, Twitter, LinkedIn, and so on.

Apparently, crowd intelligence works best on simple problems having a clear set of right and wrong answers.

CI also requires asking the right set of questions to yield results closest to actual results. For example, let's consider a cosmetic survey question, "*Would you buy this cream?*" versus a forecast approach question, "*How many people do you think will buy this cream in the next three months?*" The forecasting approach typically yields much more accurate predictions. Figure 6.5 depicts collective intelligence and design thinking:

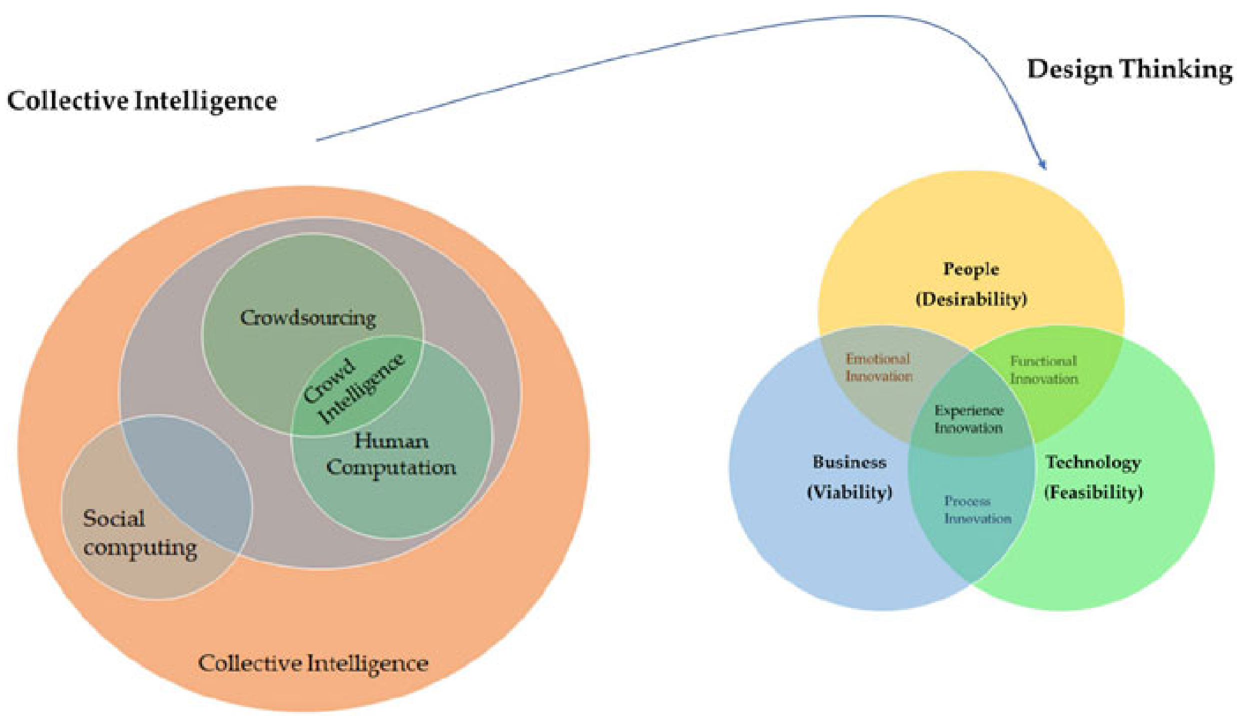


Figure 6.5: *Collective Intelligence and Design Thinking*

Right questioning (5W and 1H)

5W1H is a term for Who, What, Where, When, Why, and How. 5W1H is a business tool that helps get clarity and make better decisions before taking any action. *Figure 6.6* describes the elements of 5W1H:

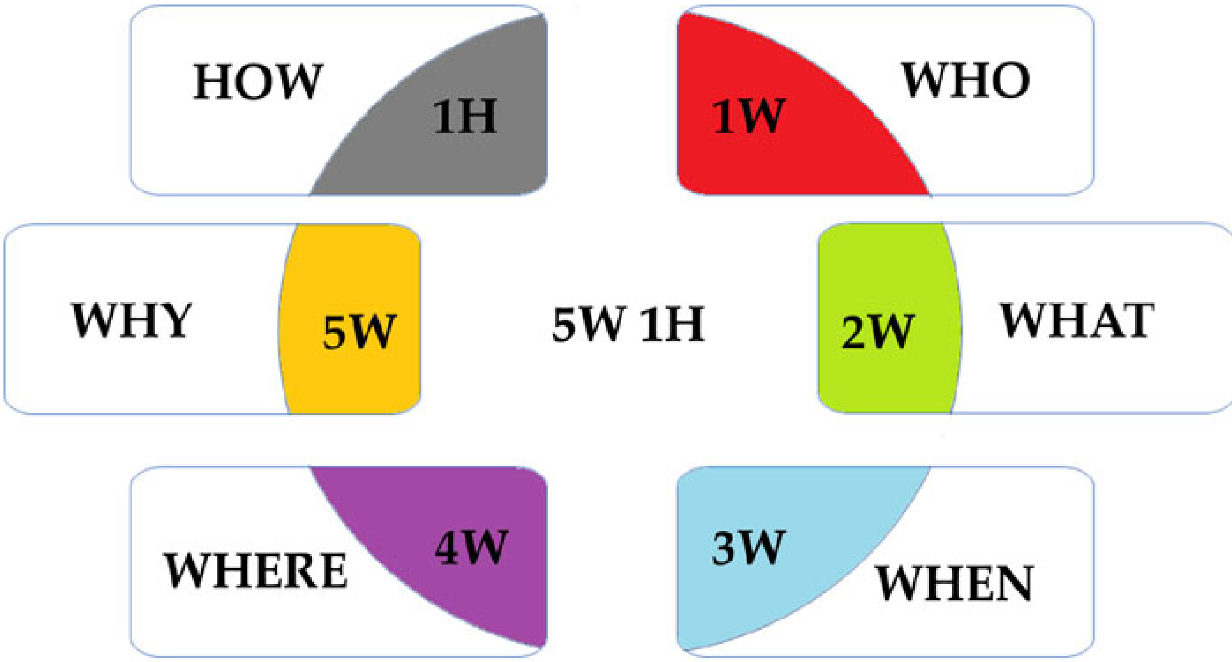


Figure 6.6: *5W1H*

5W1H is a problem-solving method that aims to bring different perspectives to ideas and issues. 5W1H, also known as the Kipling method. *Rudyard Kipling*, the British author and poet, first came up with this method.

5W1H is usually used in brainstorming sessions to find solutions to problems or generate new ideas. The elements of 5W1H are a set of five questions as mentioned in the following:

- What is the problem?
- Where is the place for the problem?
- When is the time of the problem?
- Why is the problem?
- Who is affected by the problem?
- How can we solve the problem?

Figure 6.7 further expands various elements of 5W1H:

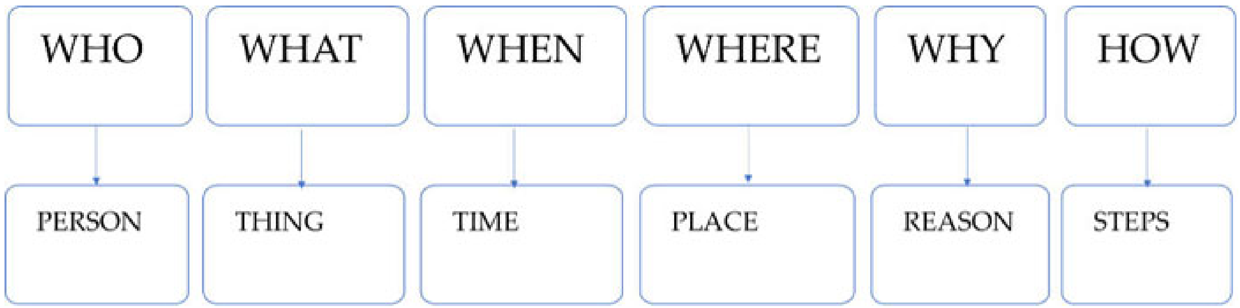


Figure 6.7: *5W1H components*

Figure 6.8 is regarding the WHO element of 5W1H and let us understand it in detail:

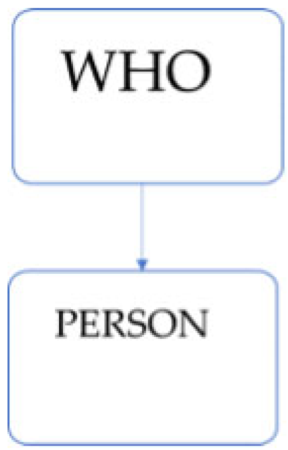


Figure 6.8: *WHO*

The WHO refers to a group or set of persons who are relevant to the problem at hand. These include people who are responsible for taking action, who are going to be affected by the problem, and other stakeholders. *Figure 6.9* is regarding WHAT element of 5W1H and let us understand it in detail:

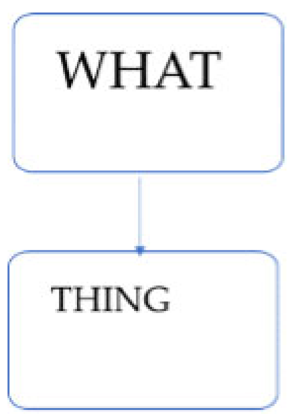


Figure 6.9: *WHAT*

WHAT refers to the problem or situation or the overall purpose. For example, there is a delay in the processing of an insurance claim. Here, the "*what*" is improving insurance processing time.

Figure 6.9 is regarding the WHEN element of 5W1H and let us understand it in detail:

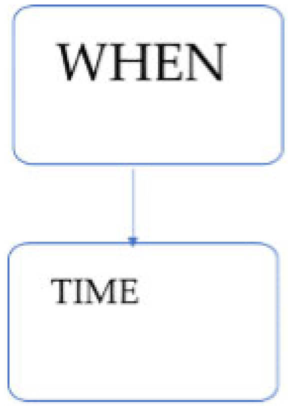


Figure 6.10: *WHEN*

WHEN refers to the time when the problem occurs or when some event happens; it also helps in capturing the frequency of the problem. For example, the problem statement of a bus getting delayed daily due to a traffic jam at a particular time has "when" as the time of occurrence of the jam.

Figure 6.11 is regarding the WHERE element of 5W1H and let us understand it in detail:

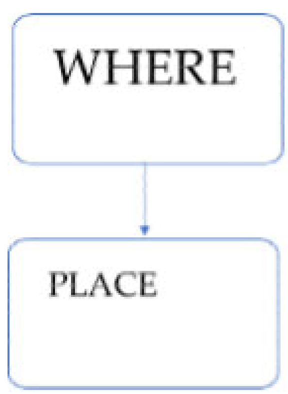


Figure 6.11: *WHERE*

WHERE refers to the place of occurrence of the problem. For example, solving the problem of introducing a new marketing process, "*where*" would be the marketing department.

Figure 6.12 regards the WHY element of 5W1H and lets us understand it in detail:

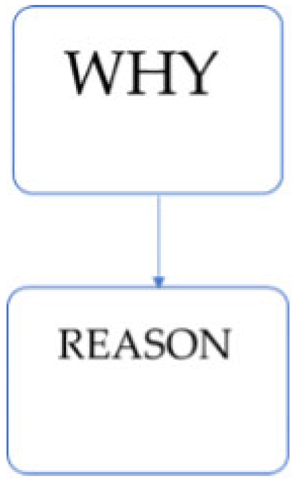


***Figure 6.12:** WHY*

WHY refers to the most significant question with an aim to identify the root cause of the problem. For example, the insurance processing delay may have "*why*" as the large number of claims that are requested at a certain time, that is, "*when*", which may be during the rainy or winter season.

Figure 6.13 is regarding HOW element of 5W1H and let us understand it in detail:

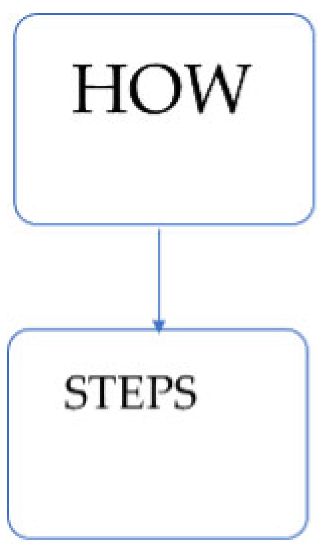


***Figure 6.13:** HOW*

HOW refers to possible solutions to fix the problem at hand. For example, if the bus is getting delayed due to a traffic jam at a certain hour, HOW may refer to changing the route of the bus during that period as a possible solution.

Identifying the problem to solve

> **"If I had an hour to solve a problem and my life depended on the solution, I would spend the first 55 minutes determining the proper question to ask… for once I know the proper question, I could solve the problem in less than five minutes."**
>
> *-Albert Einstein*

For a designer, a problem is an unmet need that, if provided for, can satisfy the purpose.

Identifying a problem and its detailing in a comprehensive way is one of the most significant steps of design thinking methodology. In other words, the quality of the question asked, or the way the problem to be solved is framed, determines the scope, ecosystem, context, aspects, constraints, opportunities, and significance of the solution.

Any problem statement must be identified, framed, and outlined. Following are good practices during problem framing phase:

Avoid jumping to solutions:

Do not get distracted by thinking in terms of features and functionality but focus on understanding the fundamental problem.

Why is it a problem:

Ask why - a simple question to help find insights and help to see a situation from a different perspective.

Reflect:

Try to look for connections and patterns.

Keep it simple:

Avoid using jargon or unnecessary complexities. The problem should be simple for all to understand, reshare and brainstorm. Involving all stakeholders by communicating in a way that can be perceived by each one is key to getting healthy contributions from all.

Focus on the user:

The user and their needs should be the center point of the problem statement.

Keep it broad:

Leave room for innovation and creativity.

Make it manageable:

Segregate the user needs and prioritize them such that the problem statement itself is manageable.

Reframe problem statement:

Problem statements must be outlined in such a way that they can be reframed based on future inputs and thus should leave room for imagination, experimentation, and change.

Ideate

The process of Ideate in design thinking is not limited to generating ideas but also selecting the right and best ones and developing them further. In other words, "*ideation*" is the creative process of generating, selecting, and developing ideas. It certainly is not a one-time idea generation or brainstorming session.

Figure 6.14 describes the ideation process. Let us understand it in detail:

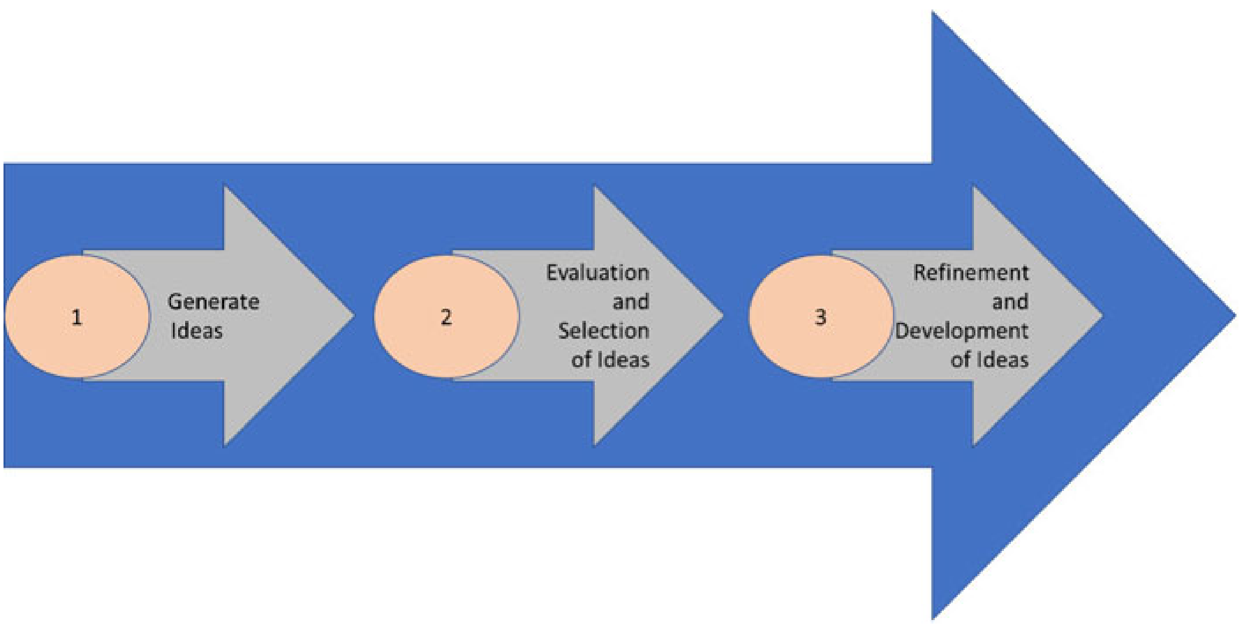


Figure 6.14: *Ideation Process*

Why is ideation significant in design thinking

Ideas fuel innovation and help improve solutions. The following are specific benefits of ideation:

- Increase innovation opportunities
- Bring together varied perspectives and ways of thinking
- Develop and refine further the best ones
- Ideation itself helps prioritize the ideas and pick the most favorable ones
- Encourage an innovative and open culture

Ideation techniques and tools

The main goal of Ideation is to bring innovation by finding solutions, generating ideas, creating opportunities, and furthering ideas as well.

Let us visit some of the tools and techniques that can be used according to the ideation needs, be it related to idea generation, selection, or development.

- **Idea generation and communication**

  Generating ideas while not sharing them with others is of no use. Hence communication and sharing of ideas form the basic step in ideation. Brainstorming is the most popular method. However, another technique must be used when more ideas are needed around a specific topic.

  Idea challenge

  This is a focused ideation method used to find solutions for specific problems, opportunities, or areas of improvement and addressed to specific audiences.

  There are two phases of the idea challenge. One is while defining the problem statement, and the other is while finding the solution.

- **Evaluation and selection**

  The nature of an idea decides its evaluation and selection technique. For instance, a new idea for a brand-new problem will be evaluated differently than an idea that is meant for incremental improvement.

Impact effort matrix

One of the ways to evaluate ideas is to consider the impact and effort graph for each. That is, high impact and high effort and low effort and high impact must be prioritized and implemented first.

Manageability

The process must be organized into smaller and more manageable stages such that the idea roadmap can be carved out, reducing uncertainty and complexity. This is ideally useful in technical stages than in disruptive ideas for creating opportunities.

- **Refine and develop**

  Storyboarding

  Storyboarding helps explore ideas in a visual manner. Not necessarily is a story carved out, but the sequence of events can be mentioned along with the idea.

  Analogy thinking

  It is very rare that a completely new idea is floated. Almost all ideas are presumed to be refined forms of some existing idea. In fact, new ideas may be a combination of already existing ideas.

  Analogy thinking helps refine and build on existing ideas and apply them to the problem.

  Opposite thinking

  Questioning long-held assumptions and challenging one's own thinking broadens horizons and gives a structured approach how to get out of box ideas.

Key challenges for ideation

The following are the key challenges for ideation:

- **Functional fixedness**: The inability to realize and identify that an idea applicable to a specific function can be used with different functions as well.
- **Diversity**: Ideas from people belonging to varied backgrounds, skills, and knowledge must be invited with an open mind to incorporate new possibilities on the horizon.

- **Chaotic**: Ideation in group sessions can turn out to be chaotic unless done under the guidance of written inputs and shared.

Key success factors for ideation

The idea of a mobile phone can be traced down to the early 1900s. However, it took multiple refinements, various developments, and the invention of other technologies before the world had its first real mobile phone.

Let's consider the factors that increase our chances of success.

- Align ideation with goals.
- Adopt a structured and methodical approach
- Ask the right set of questions
- Get right stakeholders
- State assumptions accompanying the ideas
- Focus more on the problem than the idea
- Removal of barriers in achieving any of the ideation processes

Conclusion

To summarize, design thinking is a practical methodology to be used in problem-solving. It is a human-centric approach to problem-solving. Design thinking methodology enables and empowers designers and creative solution providers to approach the problem statement in a structured manner while taking into account all the necessary factors to arrive at the best solution.

Design thinking has its base concept comprising of analysis and synthesis.

Analysis teaches a designer how to break down the problem statement into manageable smaller parts that are then studied and attempted for a solution.

Synthesis is the process of putting all the suggested solutions together to form logical big innovative solutions.

Ideas are the starting point of every innovation. We might state the obvious, or we may have many ideas to choose from. However, without communicating these ideas and performing the selection, development, and implementation of these ideas, these ideas are of no-good no matter how brilliant they could be.

Iterations and reframing the problem statements are key to improvements and finding the best solutions.

In the next chapter, we will learn about data analysis, data structure, data representation, and data exploration. This would help increase our computational skills.

Multiple choice questions

1. **What are the key elements influencing design thinking?**
    a. Business
    b. Technology
    c. People
    d. All of above
2. **What is not a part of the ideation process?**
    a. Prototype
    b. Test
    c. Define
    d. Crowdsourcing
3. **What is not a part of the 5W1H process**
    a. What
    b. Where
    c. How
    d. Help

Answers

1. **d**
2. **d**
3. **d**

Questions

1. Explain the ideation process.
2. Explain the elements of the design thinking process.
3. What is meant by collective intelligence?

Join our book's Discord space

Join the book's Discord Workspace for Latest updates, Offers, Tech happenings around the world, New Release and Sessions with the Authors:

https://discord.bpbonline.com

CHAPTER 7
Data Analysis

Introduction

Developing solutions to address real-life problems is the primary focus in a rapidly changing and increasingly technological world. The approach to solving the problems and designing the systems plays a crucial role in the solutions' effectiveness and their development life cycle.

Computational thinking allows us to do this. Computational thinking involves:

- Taking up complex problems
- Understanding the problems
- **Decomposition,** that is, breaking them down into smaller and more manageable parts
- **Pattern recognition***,* that is, identifying data patterns and relationships
- **Abstraction**, that is, applying logical reasoning and abstract thinking
- **Applying algorithms** to develop the solutions

It aims to develop skills such as problem-solving, critical thinking, creativity, and collaboration. The objective of computational thinking, thus, is to empower people to solve problems and design systems in a systematic, logical, and effective way.

Computational thinking is not only a key skill in the field of computer science but is also increasingly being recognized as an important skill in business, education, arts, and many other fields.

Structure

In this chapter, we will be discussing:

- AI and data

- Types of structured data
    - Date and time
    - String
    - Categorical
- Representation of data
- Exploring data (Pattern recognition)
    - Cases, variables, and levels of measurement
    - Data matrix and frequency table
    - Graphs and shapes of distributions
    - Mode, median, and mean
    - Range, interquartile range, and box plot
    - Variance and standard deviation
    - Z-scores
    - Example
    - Practice exercise

AI and data

In computer science, a data structure is a specific way of organizing, managing, and storing data in such a manner that the data can be accessed and modified in an efficient way.

To be precise, a data structure is a cluster of data values, their association and relationships, and the functions or data operations applicable to the data. The type of data collected may affect the way its managed.

Data exists in a plethora of different forms and sizes. It is important to understand that there are three common types of data structures:

Structured data

Structured data refers to information that is organized in a specific format, such as a table or spreadsheet, follows a consistent order, and can be easily stored, searched, and analyzed. *Figure 7.1* shows the spreadsheet format for structured data:

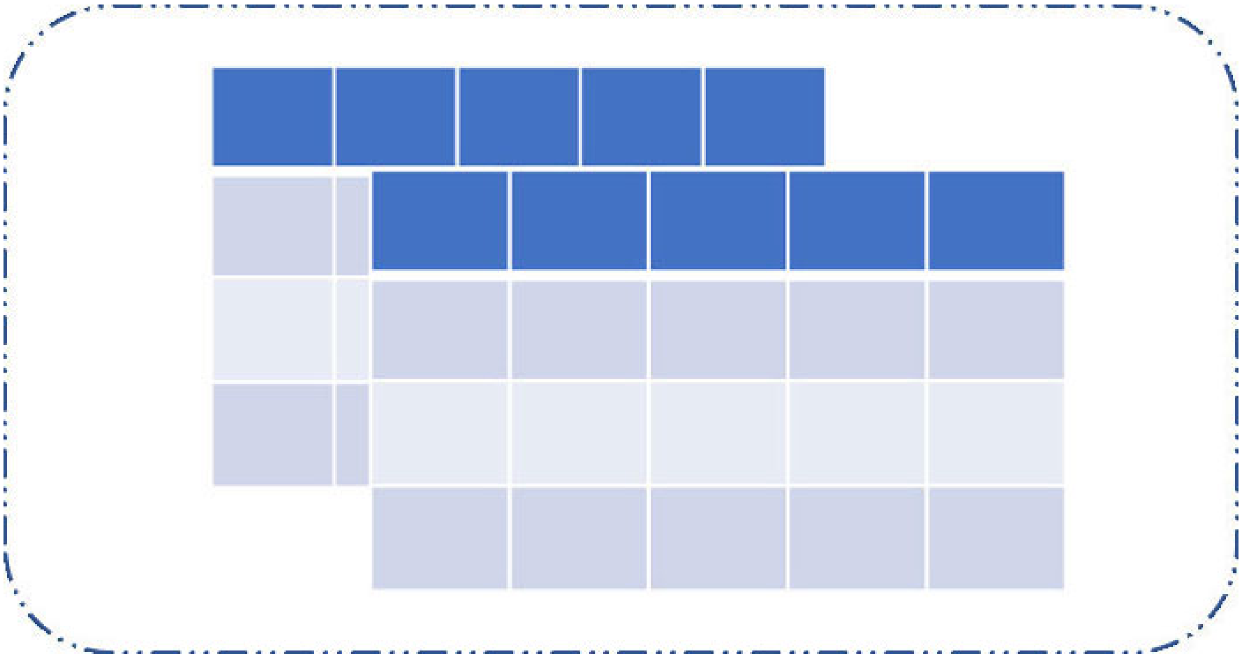


Figure 7.1: *Structured data*

Structured data is generally represented in tabular form with columns and rows that clearly define its attributes. In other words, it is typically composed of a set of well-defined fields, each of which contains a specific type of data, such as text, numbers, or dates. Examples of structured data include relational databases, CSV files, and Excel spreadsheets.

Structured Query language or **SQL** is often used to manage structured data stored in a formatted repository that typically is a database. Example: Relational data. These databases have relational keys used to uniquely identify records in a table of a relational database.

Characteristics

Let us look at the characteristics of structured data:

- Data conform to a data model.
- Data has an easily recognizable structure and resides within a record or file in fixed fields.
- Data is stored in tabular form in rows and columns.
- Data is well arranged and organized such that the data definition, data format, and meaning of data are clear and detailed.
- Similar data entities are clustered together to form relations or classes.

- Data entities in the same group have the same characteristics.
- Data is easily accessible and can be easily queried.
- Data elements are addressable for efficient analysis and processing.

Sources

The following are the sources of structure data:

- Relational databases like SQL Databases
- Data Spreadsheets like Excel
- **Online transaction processing systems (OLTP)**
- Online forms like google forms
- Sensors such as in automobiles, medical devices for temperature, sound, and so on
- GPS and RFID tags
- Logs such as from Network and Web server

Advantages

The following are the advantages of structured data:

- Structured data have well-defined schemas that make it easy to store and access data.
- Data can achieve maximized search query efficiency once it is indexed based on fields' attributes.
- Data can be sorted easily to identify patterns and relationships such that knowledge can be easily concluded from the data.
- Data operations such as modification, addition, and deletion are easy due to the well-defined form of data.
- Ease of access to data allows the implementation of data operations for business applications. For example, data warehousing for Business Intelligence can be easily implemented.
- Data formats make them easily scalable and expandable.
- Implementing data security is easy.

Unstructured data

Unstructured data refers to information that does not have a specific format or identifiable structure. Such data is difficult to organize, search, store, and analyze. *Figure 7.2* shows various unstructured formats:

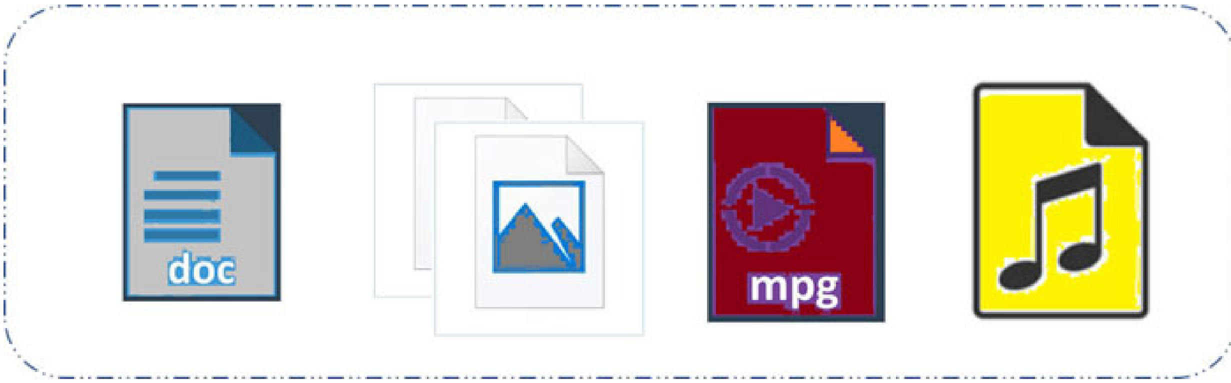


Figure 7.2: *Unstructured data*

Characteristics

Let us now look at the characteristics of unstructured data:

- Data neither follows a data model nor has any structure.
- Data cannot be stored in a tabular form having rows and columns as in Databases.
- Data does not conform to any symbolism or rules.
- Data lacks any fixed format.
- Data has no identifiable schema and structure, and therefore, it cannot be used by computer applications easily.

Sources

The following are the sources of unstructured data:

- Images (JPEG, GIF, PNG, and so on.)
- Videos and audio
- Word documents and PowerPoint presentations

Advantages

The following are the advantages of unstructured data:

- The data is not restricted to following a fixed schema
- Data is portable and very scalable
- Supports heterogeneity of sources
- Supports a variety of business applications based on data analytics

Disadvantages

The following are the disadvantages of unstructured data:

- Storing and managing unstructured data due to a lack of schema and structure is difficult.
- Data Indexing is difficult.
- Due to unclear structure and not having pre-defined attributes, the search results are not very accurate and erroneous.
- Data security is difficult.

Difficulties faced in storing

Let us look at the difficulties faced in storing unstructured data:

- Unstructured data requires a lot of storage space, and hence storage cost is high.
- Operations like data updation, deletion, and search are very difficult due to unclear data structure.
- Data Indexing is difficult for unstructured data.

Solutions for storing

Let us discuss the solutions for storing unstructured data:

- **Content addressable storage** systems (**CAS**) can be used to store unstructured data. It uses metadata; that is, a unique name is assigned to every data object stored. The object is then accessed based on content than its location.
- The XML format can be used to store Unstructured data.

- RDBMS, which supports BLOBs (large binary objects - a varying-length binary string that can be up to 2,147,483,647 characters long.), can be used to store unstructured data.

Metadata– data about data

Metadata is data about data. Technically, it is not a separate data structure but provides additional information about a specific set of data. For example, in a set of video files, the associated metadata can store information about when (date and time) the video was recorded and other data, such as location details, that by themselves are structured data. *Figure 7.3* shows a representation of metadata (as structured data):

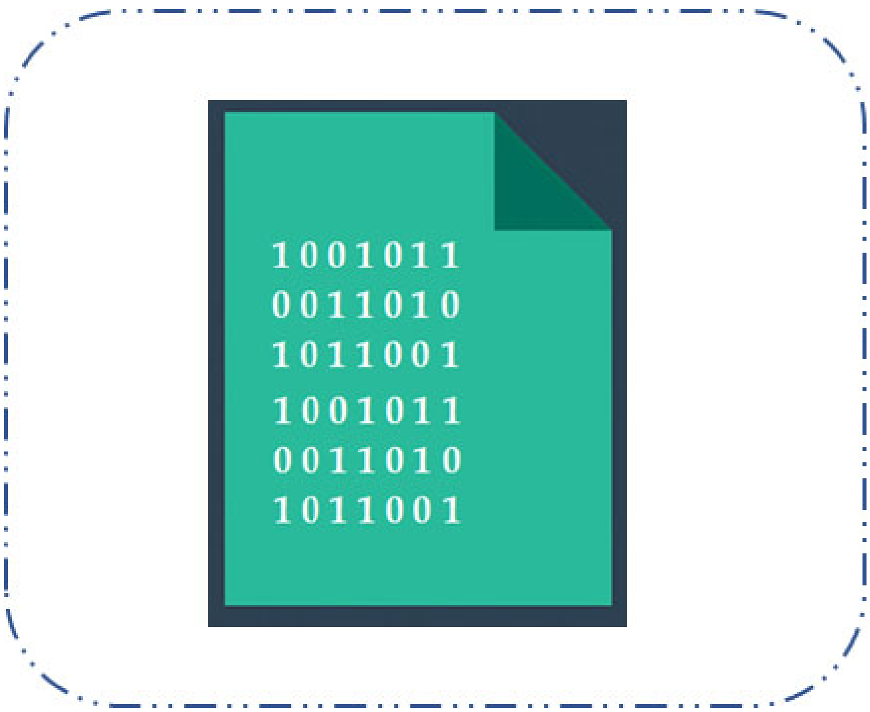


Figure 7.3: *Metadata*

Semi-structured data

Semi-structured data refers to information that has some organizational properties but not as much as structured data. It may not have a fixed schema or fixed fields. It usually contains a blend of structured and unstructured data.

XML or JSON files are examples of semi-structured data. They are more difficult to analyze and process than structured data. The data may be a mix of some pre-defined fields and unstructured data within fields or in free text fields. Semi-structured data can be searched and analyzed with the support of specialized applications and programming languages such as XPath, XQuery, and JSONPath.

Figure 7.4 shows some semi-structured data:

Figure 7.4: *Semi structure data*

Characteristics

Let us now look at the characteristics of semi-structured data:

- Data does not conform to a fixed schema or a data model but has some structure.
- Data cannot be stored in tabular form as in databases.
- Metadata and tags are used to group the data and describe how it is stored.
- A hierarchy is formed after organizing similar entities in groups where entities in the same group may not have the same attributes.
- The lack of sufficient metadata makes automation and data management difficult.
- The absence of a well-defined structure makes it difficult to use computer applications easily.

Source

The following are the sources of semi-structured data:

- Markup languages like XML, HTML, and so on
- Binary executables (contains executable code that is represented in specific processor instructions and text strings)
- TCP/IP packets (format contains header and data to be transferred)
- Web pages (format contains headers and data)
- E-mails (combination of a fixed header, free text, and file attachments)

Advantages

The following are the advantages of semi-structured data:

- Doesn't conform to a fixed schema, so it is not constrained by it.
- Schema can be changed, making it flexible and portable.
- Having some structure, it is possible to perceive structured data as semi-structured data.
- When relational databases aren't helpful in describing how data is to be stored.
- Having a mix of structured and unstructured data makes it easy to support heterogeneous sources.

Disadvantages

The following are the disadvantages of semi-structured data:

- Difficult to store data due to the lack of a fixed schema
- Finding the relationship between data is difficult
- Difficult to make Queries that are less efficient as compared to structured data **https://www.geeksforgeeks.org/structured-data/**

Difficulties in storing

Let us look at the difficulties faced in storing semi-structured data:

- It is difficult to find the relationship between the data and hence difficulty in storing it.
- Schema and data are usually linked together and hence are also dependent on each other. Both schema and data may get updated by the same query. Here, the schema gets updated frequently.

- Designing the structure of data is complicated due to the uncertain distinction between schema and data.
- As compared to structured data, storage cost is higher.

Solutions for storing

Let us discuss the solutions for storing semi-structured data:

- A DBMS specially designed to store semi-structured data can be used to store the data. Non-relational databases such as MongoDB are the preferred choice for storing many kinds of unstructured data.
- **Object Exchange Model (OEM)**
- RDBMS can be used. It required mapping the data to a relational schema, thereafter, mapping it to a table.

Figure 7.5 compares structured and unstructured data:

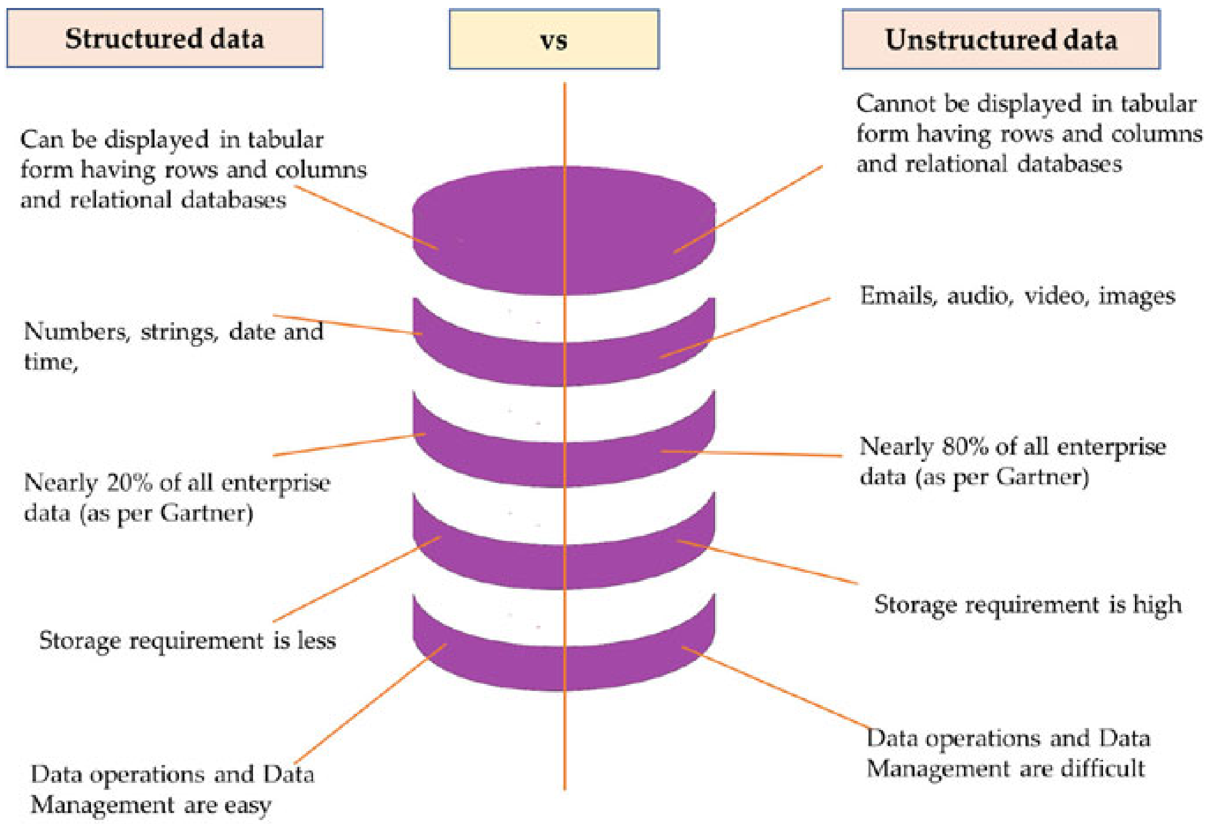


***Figure 7.5:** Structured vs. unstructured data*

Types of structured data

Now that we have learned that structured data is in tabular form with rows and columns and conforms to the relational database. As such, structured data being highly organized, is easily understood by machine language.

Data pertaining to dates, employee or customer information, credit card numbers, stock information, names, addresses, and more are examples of structured data.

Date and time

Date and time data is a type of structured data that represents a specific time. It is a numerical type of data used to represent a continuous or discrete time. This data is useful in many industry verticals, including finance, logistics, customer relationship management, and social media, and is commonly used in data analysis and modeling.

- The date is represented in a three-part value (year, month, and day).
- Time is represented in three-part values (hour, minute, and second).
- A timestamp is represented in six or seven-part values (year, month, day, hour, minute, second, and optionally fractions of a second).

There are several distinct ways to represent date and time data, described as the following:

Timestamps: Timestamps are a way to represent a specific moment in time using a combination of a date and a time. They are typically represented in a standardized format, such as YYYY-MM-DD HH:MM:SS and can include time zone information. These may contain data calibrated to minute fractions of a second. *Table 7.1* gives examples of time stamps:

Beginning	End	Remarks (duration)
2021-10-05	2021-10-09	5 days
2022-10-09 11:00:00	2022-10-09 11:59:59	1 hour

Table 7.1: *Time stamps*

Time intervals: Time intervals are a way to describe the duration of time lapse between events. They can be represented using start and end timestamps or simply as the duration in a standardized format, for example, HH:MM:SS.

Time zones: Time zones are a way to describe the local time in various parts of the world. It represents the time for a region on Earth that has a uniform standard time for legal, commercial, and social purposes.

Time zones are typically represented by a time zone abbreviation (such as IST for Indian Standard Time) and an offset from universally accepted time zone reference, that is, **Coordinated Universal Time** (**UTC**).

For example, IST is UTC +5:30, which means the Indian standard time zone is five and a half hours ahead of Coordinated Universal Time.

Time zones are often represented as structured data, typically in the form of a table that lists the offset from UTC for each region on Earth. This table can also include any **Daylight-Saving Time** (**DST**) rules that are in effect.

Table 7.2 shows timestamps with time zones:

Beginning	End	Remarks (duration)
2023-01-17 11:00:00 UTC +5	2023-01-17 12:00:00 UTC +6	0

Table 7.2: *Timestamps with time zones*

Time series: Time series is a way to represent data points collected over regular time intervals. Time series data is often used to find trends and patterns such that it can be analyzed using statistical techniques. Time series data analysis can be used to forecast revenues, stock prices, monthly sales, and hourly traffic detect anomalies, and understand the underlying mechanisms that drive the data.

The data points are typically ordered by their occurrence time, and the intervals to collect the data can be fixed (for example, every hour) or variable (for example, every 5 minutes).

Time series data can be continuous (for example, humidity measurements) or discrete (for example, the number of store visitors). *Table 7.3* is an example of time series:

Beginning (hour of the day in 24 hr. format)	End (hour of the day in 24 hr. format)	Visitors
11:01	12:00	10
12:01	13:00	10

13:01	14:00	10
14:01	15:00	10
15:01	16:00	45
16:01	17:01	10
17:01	18:00	60

Table 7.3: *Example of time series*

Data and time typically have specific representation adopted across the world for legal, financial, and economic purposes. For example, Americas use the format MM-DD-YYYY, while in most of Asia, the format commonly used is DD-MM-YYYY, where DD represents the date, MM represents the month, and YYYY represents the year.

As such, working with date and time data often is quite challenging because of the complexity of the data and the various ways of representation. It is important to meticulously consider the data format and the appropriate reference for the analysis, as these can significantly influence the insights and interpretations that can be drawn.

String

Strings, a type of structured data, consists of a sequence of characters, including letters, numbers, and symbols. They are typically used to represent text-based data, such as names, addresses, and descriptions, and are an important data type in many verticals, including data science, natural language processing, and data analysis.

The following describes operations that can be performed on strings:

Concatenation: The process of combining two or more strings into a single string is termed concatenation. One of the ways to achieve concatenation is by using the + operator in many programming languages.

For example, in python, the + operator is used in the following ways to achieve concatenation:

```
string1 = "Artificial"
string2 = "Intelligence"
concatenated_string = string1 + string2
print(concatenated_string)
```

This will output the string "**`Artificial Intelligence`**".

Splitting: the process of breaking a string into substrings based on a delimiter, such as a space ( ), a comma (`,`), a semicolon (`;`), or quotes (`"`, `'`), is termed as splitting. This can be useful for extracting words or pieces of data from a string.

For example, in Python, strings can be split using the function **`split()`**.

```
string = " Artificial Intelligence”
splitted_string = string.split()
print(splitted_string)
```

This will output the list ["**`Artificial`** ", " **`Intelligence`** "]

Searching: The process of looking for a particular substring within a string is termed searching. This can be done using methods such as **`indexOf()`** or **`contains()`** in many programming languages.

Depending on the function being used, the output will be the place holder of the substring of either true or false, indicating if the string contains the substring.

For example: using the operator/function **`index`**, the results are as follows:

```
string = " Artificial Intelligence”
result = string.index(“Intelligence”)
print(result)
```

This output will be 11

Replacing: the process of replacing a particular substring within a string with a different string is termed replacing. This can be done using methods such as **`replace()`** or **`substitute()`** in many programming languages.

For example: using operator/function replace, the results are as follows:

```
string = " Artificial Intelligence”
result = string.replace(“Intelligence, “Sky”)
print(result)
```

This output will be “Artificial Sky”

Sorting: The process of arranging the characters in a string in a particular order, like alphabetical or numerical order, is termed sorting. This operation can be useful for reorganizing data or for comparing various strings.

For example: Using operator **`sorted()`**, the results are as follows:

```
string = [4, 1, 3, 2]
result= string.sorted(String)
print(result)
```

This output will be **`[1, 2, 3, 4]`**

Strings are commonly used in many applications, including data analysis and modeling. They can be transformed in different ways to extract insights and draw interpretations from text-based data.

Categorical data

Categorical data is a type of structured data that is used to represent values in a number of categories. In other words, it commonly represents data that can be classified or grouped into categories, such as profile, designation, gender, age group, or occupation.

Categorical data can be divided into the following types:

Nominal data

When categorical data represents values in categories that do not have a specific order or ranking, it is termed nominal data. As an example, a list of colors, where the categories (red, blue, green, and so on.) do not have a specific order or the order doesn't make a difference.

Ordinal data

When categorical data represents values in categories that have a particular order or ranking, it is termed ordinal data. As an example, a list of job titles, where the categories (fresher, associate, director, and so on.) have a specific order or hierarchy. Another example can be the rating of a product (1,2,3,4,5) where 1 is the worst, and 5 is the best performance.

Binary data

A special case of categorical data where the features can have any of two distinctive values, that is, it can be either 0 or 1 or could be True or False.

Common techniques for analyzing categorical data include frequency tables, bar charts, and more. These techniques can be used to describe the distribution of the data and to identify patterns and relationships between different categories.

Representation of data

Representation of data is an important aspect of data science. The way the data is organized, displayed, and summarized refers to the representation of data. It can significantly impact the insights and conclusions that can be drawn from it.

Data representations can provide the following information:

- Spread or distribution of data
- Typical values of data
- The density of data within a category or data range
- Patterns and trends in data
- Relationship between variables

Data representation requires the following skills:

- Finding best suited graphical representation for the data
- Finding the best description of the graphical representation
- Find solutions using data representations

There are several common ways to represent data, including tables, graphs, and plots. A typical presentation of various data types is:

- Qualitative data can be presented in tabular forms, bar charts, or pie charts.
- Quantitative data can be represented using dot plots and histograms.
- Line graphs are used to present trends over time.
- Scatterplots are used to present the relationship between variables.

Graphs are represented using two lines called coordinate axes. The horizontal line is termed the x-axis, while the vertical line is termed the y-axis. The two lines intersect at point 'O', called the point of origin.

On the x-axis, distances from the point 'O' to the right side have a positive value, and the distances from the origin to the left side have negative values. On the y-axis, the points above point 'O' have positives value, and the points below point ‘O’ have negative values. *Figure 7.6* shows the x-axis and the y-axis, and the point ‘O’ or point of origin used in principles of graphical representation**:**

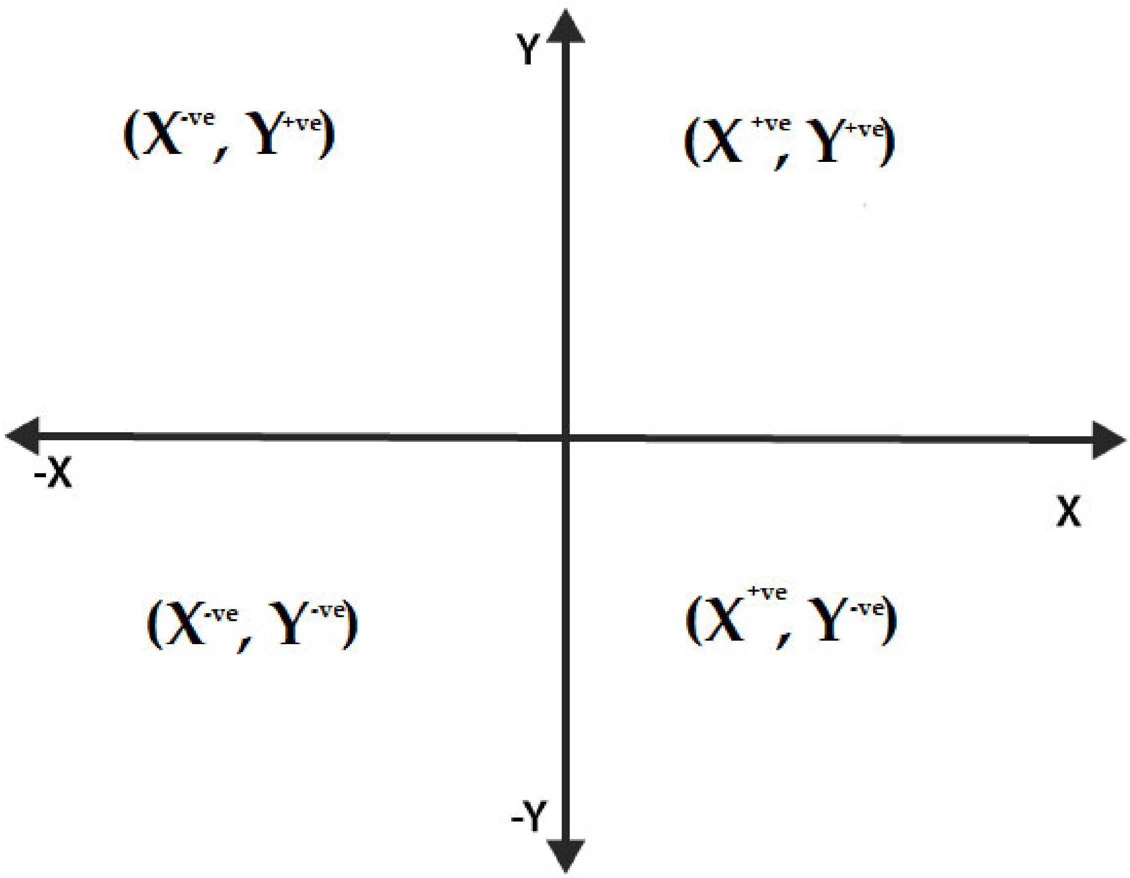


Figure 7.6: *principles of graphical representation*

Exploring data (Pattern recognition)

Exploring data and pattern recognition are significant procedures in statistical analysis.

The first step of data analysis is data exploration. It is used to explore and visualize data to find insights, trends, and data patterns.

Exploring data demands looking at the data in various ways, using graphical tools and tabular formats, to interpret the characteristics and patterns of the data. This process also involves using various statistical techniques, like measures of central tendency (mean, median, and mode) and measures of variability (variance and standard deviation), to understand the distribution and spread of the data.

Pattern recognition involves recognizing patterns and trends in the data to help understand the underlying relationships. Patterns in the data can be found by visual representation, using tools such as graphs and plots, or by using statistical techniques, such as regression analysis, to assess relationships between variables.

Exploring data and pattern recognition are also critical for spotting areas for further inspection and for creating models and forecasts about future data.

Cases, variables, and levels of measurement

Case: A case refers to a unit of data, an observation, a sample, an event, an experiment, and likes, basis the context of the data being collected.

When the data collection is about humans, the term used is *participants*; for animals, the term used is *subject*, and otherwise, the typical term used is *experimental unit*.

In a dataset, each row typically represents a single case.

For example, in a science contest, each participant would be a case, and each column would represent different characteristics of the participants, such as class, marks, rank, age, gender, height, or weight, termed as a variable.

Variable: A variable is a characteristic or attribute that is measured and can take on different values. In a dataset, if each row represents a single case, each column represents associated characteristics or attributes and is termed a variable.

The two different types of variables, as mentioned in the following:

Quantitative or numerical

Data that is expressed as a numeric. It can further be categorized into two forms:

- **Continuous** — Data that can have random values (but within a range). For example, the speed of a motor, pulse rate, and so on.
- **Discrete** — Data that can have only integer values, say, for example, a count for an event. For example, the probability of heads showing up in

10 flips of a coin.

Categorical data

Data can be classified into a specific set of values representing possible categories. These can be further classified into the following three types:

- **Binary** — or dichotomous variable that can accept only 0/1 or True/False.
- **Ordinal** — Categorical data with explicit ordering. For example, rating of a product (1,2,3,4,5).
- **Nominal** – Categorical data with no explicit ordering. For example, the name of show brands (*Nike, Adidas, Puma, Woodland, Reebok*).

Constant

A constant is a special variable having a characteristic or attribute that is the same for all cases.

Levels of measurement

Levels of measurement or scale of measurement refers to how precisely a variable is recorded. There are four levels of measurement:

- **Nominal**: Data is classified into categories, lacking any order or ranking. For example, gender or eye color.
- **Ordinal**: Data is classified into categories, lacking any order or ranking, and can be ranked or ordered, but the difference between each rank is not necessarily equal. For example, career levels (intern, associate, senior associate) or education levels (1^{st} grade, 12^{th} grade, college, postgraduate, Ph.D.)
- **Interval**: Data is classified into categories, has order or ranking, and is evenly spaced. For example, the temperature is measured in degrees Celsius. For example, the temperature measured in Fahrenheit or Celsius having zero value doesn’t mean the absence of heat.
- **Ratio**: Data is classified into categories, has order or ranking, is evenly spaced, and has a natural zero. For example, the temperature measured in Kelvin is a ratio variable, as zero Kelvin means 'no heat'.

It is important to note that the level of measurement of a variable determines the types of statistical analyses that can be performed on the data and, thereafter, conclusions that can be drawn.

The level of measurement has a hierarchy in its complexity and precision, from low (nominal) to high (ratio).

The levels of measurement are cumulative as one goes from lowest to highest. In other words, each level takes on the properties of lower levels and adds its own new properties. *Figure 7.7* shows this fact in visual format:

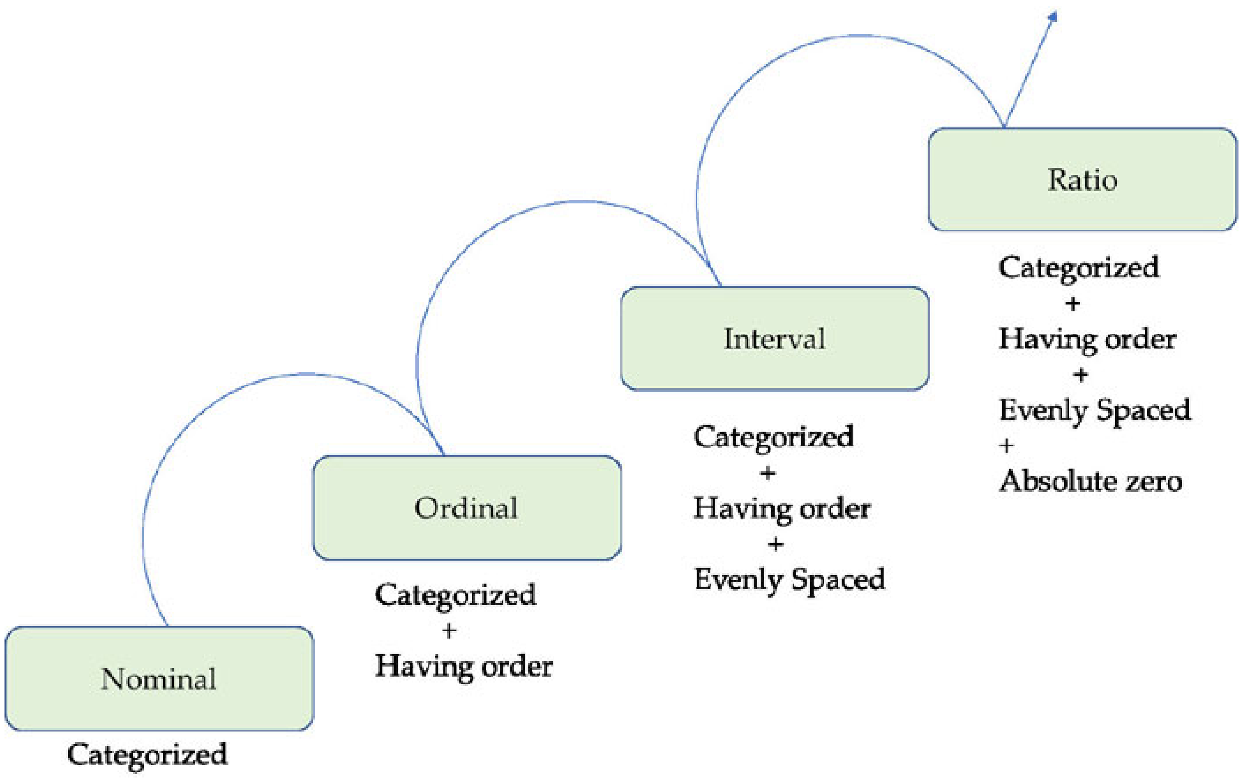


***Figure** 7.7: Levels of measurement*

The following are a few observations about the levels of measurement:

- The characteristic of a level contains the characteristics of the levels below.
- The higher the level of measurement (the ratio being the highest), the closer will be the measurement to true outcomes.

Importance of levels of measurement

The level of measurement of a variable determines the statistical analysis that can be applied and, thereby, the type of conclusions that can be drawn.

This is because the levels of measurements limit the various statistical methods that can be conducted and performed on the data to get a summary of the data (descriptive statistics) and inferences (inferential statistics).

There are cases where the same variable can be measured at different levels; it is important to consider and decide on the level of measurement to be used before data collection begins.

Example

Let us consider measuring the variable of marks at an ordinal or ratio level.

Ordinal level: Consider brackets for each of the marks range: 0–39, 40–69, and 70–100. The students can select the bracket that represents their marks. The brackets are coded with numbers from 1–3.

Bracket 1 represents marks range 0-39

Bracket 2 represents marks range 40-69

Bracket 3 represents marks range 70-100

Ratio level: Exact marks of the students are collected during data collection

Table 7.4 describes the ordinal level and ratio level:

Student	Marks (ordinal level)	Marks (ratio level)
A	Bracket 1	38
B	Bracket 2	68
C	Bracket 3	72

Table 7.4: *Ordinal Data vs. Ratio Level*

At the ordinal level, the exact marks of students are not known. The bracket their marks fall under is only evident.

At ratio levels, we get to see the exact marks, and it also becomes evident that the difference in marks between C and B is very less as compared to between B and A, which is much higher.

Descriptive statistics and Level of measurement

Descriptive statistics provide a summary of the data. It helps to gather an idea of the middle and spread the data through measures of central tendency and variability.

The level of measurement helps decide statistical methods that can be used to measure the central tendency or variability of the data set, formulated on the mathematical operations appropriate for the level.

As seen earlier, the higher levels have cumulate features of all lower levels. In the case of statistical methods applicability too, all mathematical operations and measures that can be used at lower levels are also applicable at the higher levels.

The following *Table 7.5* summarizes the preceding statement:

Data type	**Mathematical operations**	**Measures of variability**	**Measures of central tendency**	**Graphs**
Nominal	- Equality	None	- Frequencies - Mode	- Bar - Pie
Ordinal	- Equality - Comparison	-Range -Interquartile range	- Frequencies - Mode - Median - Percentiles	- Bar- Pie
Interval	- Equality - Comparison -Addition, subtraction	-Range -Interquartile range -Standard deviation - Variance	-Frequencies -Mode -Median - Mean - Standard Deviation	- Bar - Pie - Box Plot - Histogram
Ratio	- Equality - Comparison -Addition, subtraction -Multiplication, division	-Range -Interquartile range -Standard deviation - Variance	- Frequencies - Mode - Median - Mean - Standard Deviation	- Bar - Pie - Box Plot - Histogram

Table 7.5: *Levels of measurement and descriptive statistics*

Where Equality represents $(=, \neq)$; Comparison represents $(>, <)$; Addition, subtraction represent $(+,-)$; Multiplication, division represent $(\times, \div)$.

Data matrix and frequency table

Let's revisit that cases are the persons, animals, or things (including events, outcomes, and so on.), and variables are the characteristics of interest of these.

Data matrix

A data matrix is the representation of these cases and variables, in other words, the values in a dataset in a tabular format (also called a grid format), with each row of a data matrix representing a case (or individual observation) and each column representing a variable. It is a common way to organize and present data for analysis and inferences.

A comprehensive Data Matrix may contain millions or even more cases.

Frequency table

Once the data has been organized and presented, to get insights, summarization of the information can be done in the form of frequency tables.

A frequency table is a type of data matrix that lists a set of values and the number of times they appear in a dataset. Frequency tables are used to summarize and analyze categorical data, that is, the data that can be divided into various distinct categories or groups.

In the following *Table 7.6*, we will see the data matrix and the accompanying frequency table:

Student	Favorite meal
A	Italian
B	Italian
C	Chinese
D	Thai
E	Indian
F	Thai

Table 7.6: *Students and favorite meal*

To create the associated frequency table, we need to decide on the categories or groups that need to be used. In this example, we will take "favorite meal type" and list the number of observations against it as in *Table 7.7*:

Favorite Meal Type	Frequency	Percentage	Cumulative Percentage
Chinese	1	16.67%	16.67%
Indian	1	16.67%	33.34%

Italian	2	33.33%	66.67%
Thai	2	33.33%	100%
Total	6	100%	

Table 7.7: *Frequency table*

Where each row represents favorite meal types, the "frequency" column represents the number of times a particular favorite meal type appears in the data matrix that represents a student and their favorite meal type. The "Percentage" column is calculated by dividing the frequency of a favorite meal type by the total number of frequencies. The "cumulative frequency" is calculated by adding each frequency to the sum of its predecessors.

Frequency tables have one limitation, that is, they do not provide particulars about the relative sizes of the different categories. For example, the frequency table might show that 2 students have a favorite Thai meal, and 1 student prefers Chinese, but it fails to describe the proportion of the total number of students these numbers represent. We might use a bar chart or pie chart to compare the sizes of the categories.

There are four types of frequency distributions:

- **Ungrouped**: This describes the number of cases (or observations) against each value of a variable. This type of frequency distribution is used for categorical variables.
- **Grouped**: This describes the number of cases (or observations) against each class interval of a variable where Class intervals are ordered groupings of values of the variable. This type of frequency distribution is used for quantitative variables.
- **Relative**: This describes the proportion of cases (or observations) against each value or class interval of a variable. This type of frequency distribution is used for any type of a variable when the area of interest is to get insights into comparing various frequencies over the actual number of observations.
- **Cumulative**: This describes the sum of the frequencies less than or equal to each value or class interval of a variable. This type of frequency distribution is used for ordinal or quantitative variables when an area of interest is to find insights into the number of times the observations fall below specific values.

Graphs and shapes of distributions

Graphs and shapes of distributions are useful tools for visualizing and determining the characteristics of a dataset.

Categorical variable

There are two simple graphs for visualizing the distribution of categorical variables:

- Pie chart
- Bar chat

Let's visualize the frequency table for the "favorite meal" of students in the class. The following *Figure 7.8* represents the associated Pie chart:

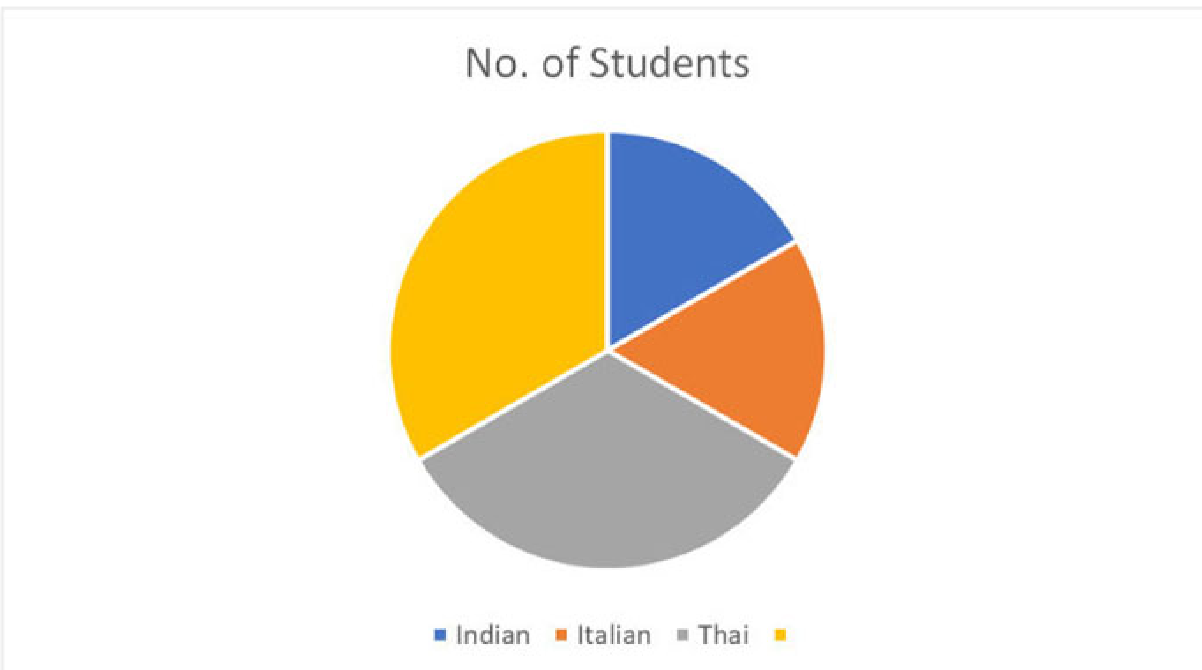


Figure 7.8: *Pie chart*

The same can be visualized in terms of a Bar chart as represented in the following *Figure 7.9*:

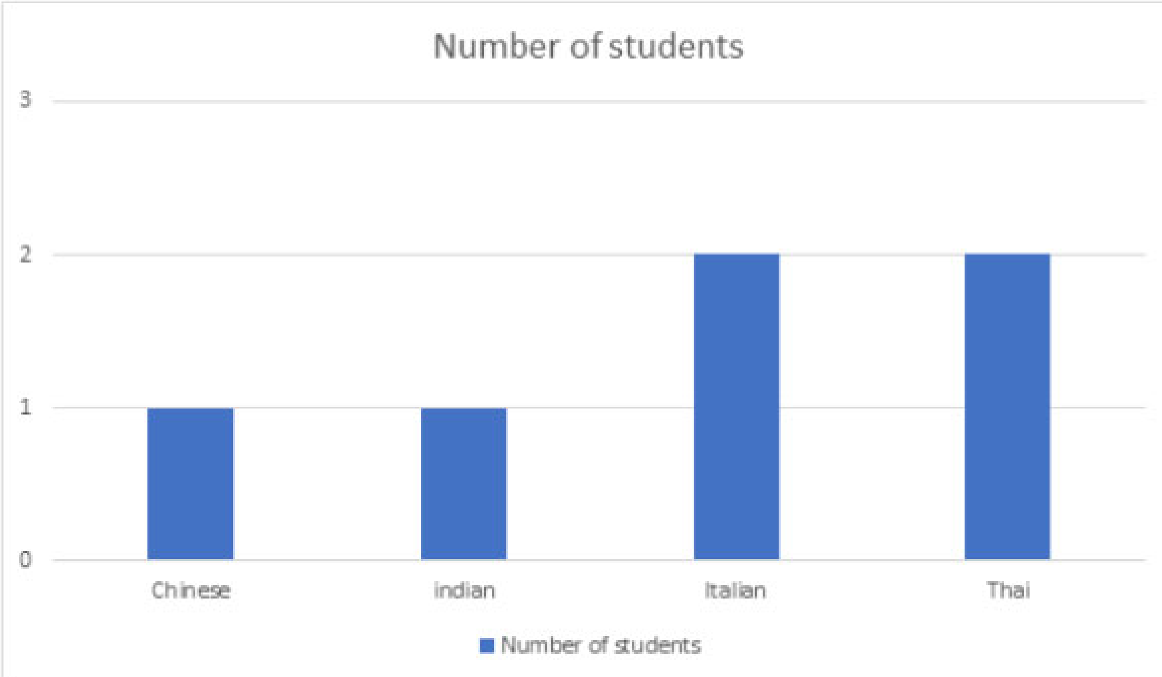


Figure 7.9: *Bar graph*

For ordinal variables, pie charts are rarely used since the facts related to the order of the data set can be lost in such a visualization.

Also, see that the pie chart stresses the way the different categories compare to the whole, and the bar chart stresses the comparison of different categories with each other.

Quantitative variable

There are two simple graphs for visualizing the distribution of categorical variables:

- Histogram
- Box plot

Histogram

A **histogram** is similar in representation to a bar graph. That is, it uses bars for visualization of the frequencies or relative frequencies of a variable. Unlike in a bar graph, bars in a histogram touch each other. It indicates that the values of an interval/ratio variable represent a continuous scale.

Histograms have four kinds of modalities:

- **Unimodal:** Having one peak, as shown in *Figure 7.10*:

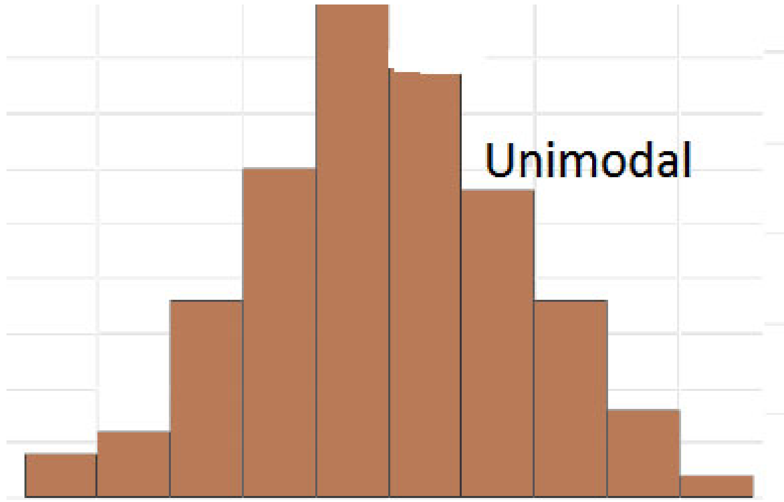


Figure 7.10: *Unimodal*

- **Bimodal:** Having two peaks as shown in *Figure 7.11*:

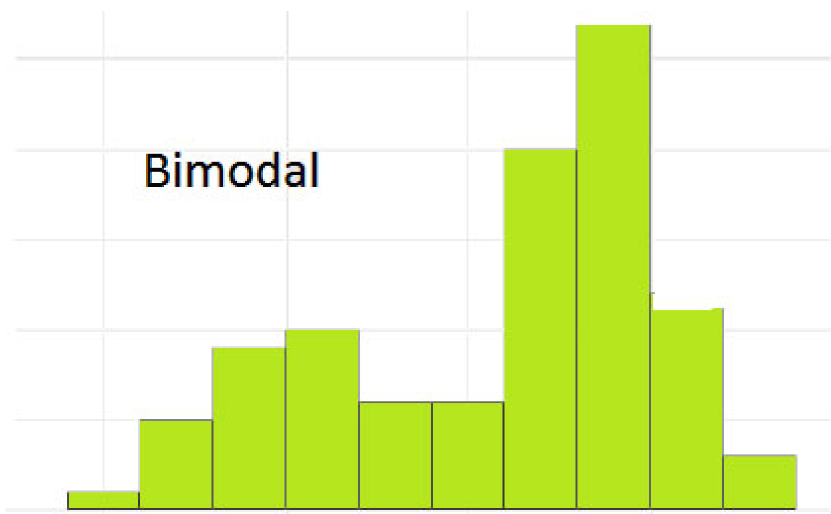


Figure 7.11: *Bimodal*

- **Multimodal:** Having many peaks, as shown in *Figure 7.12*:

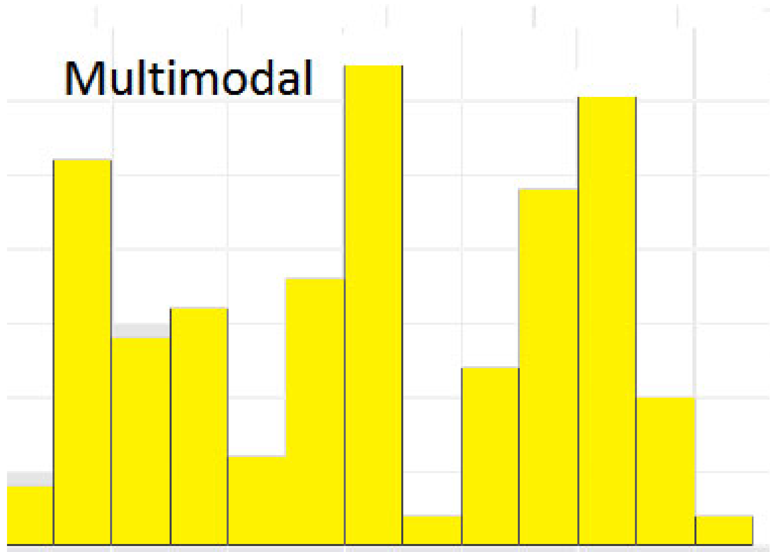


Figure 7.12: *Multimodal*

- **Uniform:** Having uniformly distribution, as shown in *Figure 7.13*:

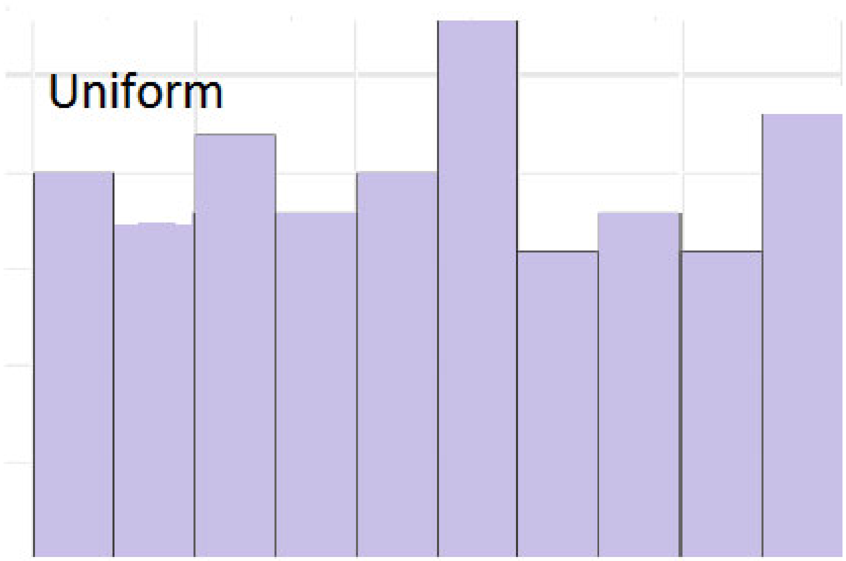


Figure 7.13: *Uniform*

Let us look at examples of unimodal histograms:

- **Symmetrical distributions** have a balanced shape, with the data evenly distributed around the center.
    - Approximately symmetric, having a bell curve and one peak, as shown in *Figure 7.14*:

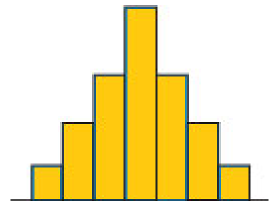


Figure 7.14: *Symmetrical*

- **Skewed:** These distributions peak in one direction, either to the left or to the right. In a left-skewed distribution, the peak extends to the right. In a right-skewed distribution, the peak extends to the left.
    - Left skewed and unimodal, as shown in *Figure 7.15***:**

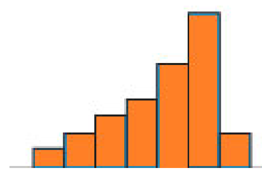


Figure 7.15: *Left Skewed*

    - Right skewed and unimodal, as shown in *Figure 7.16:*

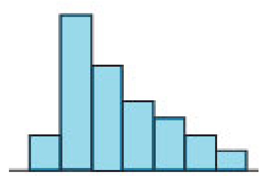


Figure 7.16: *Right Skewed*

Dot plot

A dot plot represents each observation and displays them with a dot, as shown in *Figure 7.17:*

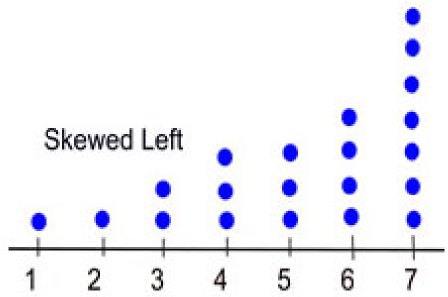


Figure 7.17: *Dot plot*

Density plots

Density plots are continuous curves representing the distribution of a variable. They show the estimated probability density of the data instead of displaying the frequency in discrete bins as in histograms. Density plots are useful where there is a need to visualize the overall shape of a distribution, as well as for recognizing patterns and trends in the data.

Figure 7.18 shows a typical density plot:

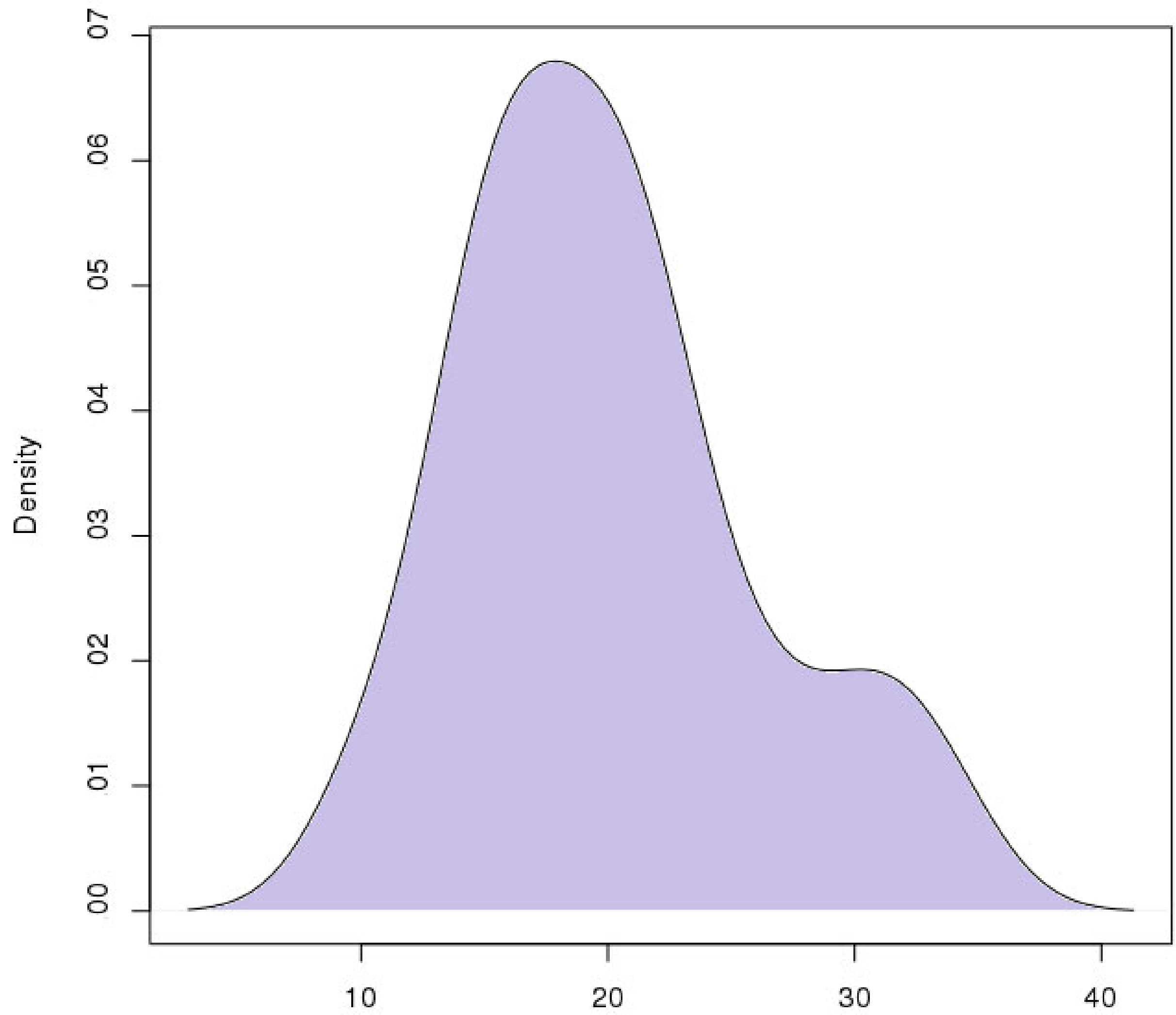


Figure 7.18: *Density plots*

Mode, median, and mean

After summarizing the distribution of the dataset in terms of graphs, comes measuring the center of the distribution. Measuring the central tendency of a variable involves 3 M's that are:

- Mode
- Median
- Mean

Figure 7.19 describes the relation of the 3Ms in the distribution of data:

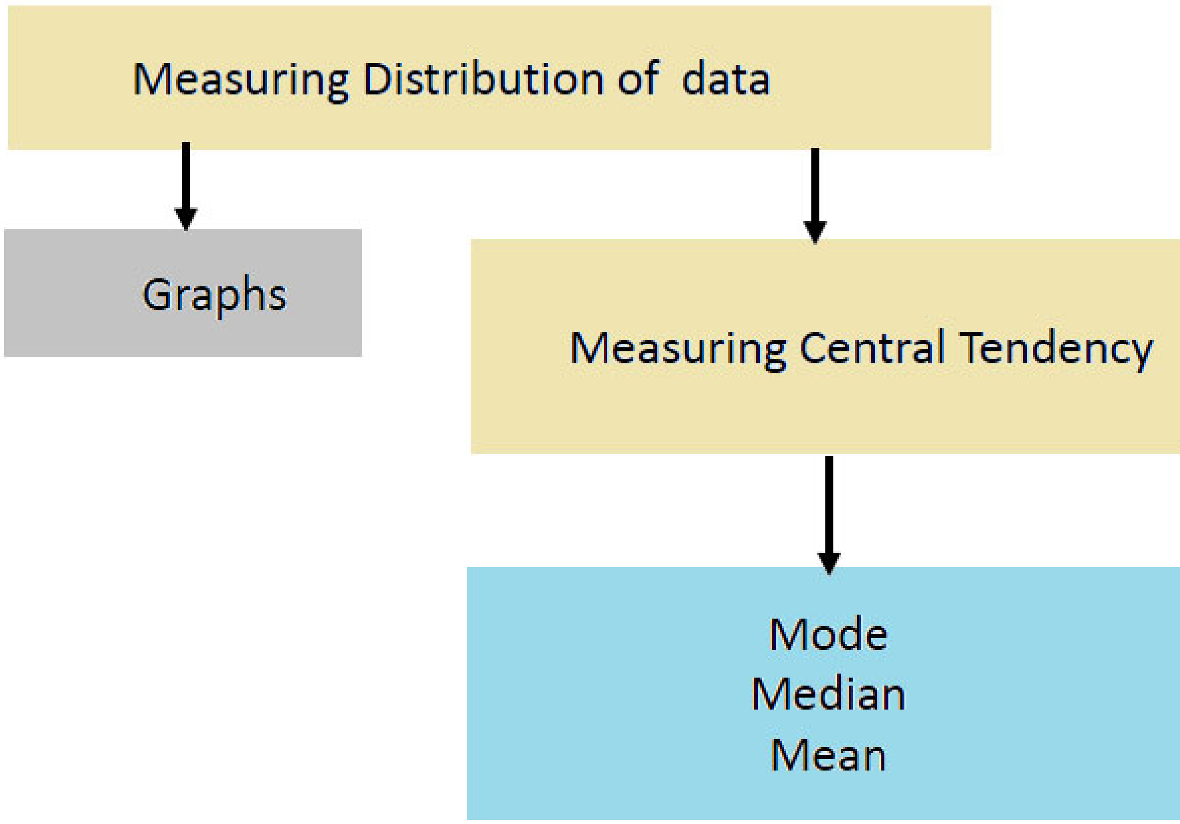


Figure 7.19: *3M's Mean, median, and mode*

These 3 M's are useful in summarizing and describing the characteristics of a dataset and are often used in statistical analysis.

Mode

The value that occurs most frequently in the dataset is termed mode. In other words, the mode is the most common result. Mode, then, is the name of the category occurring most often. A dataset may have more than one mode for the variable.

For example, if a dataset contains the values 1, 2, 2, 3, 3, 3, 3, 4, 4, 5, the mode is 3 because it appears the most times.

In case a set of numbers has more than one number having the same frequency of occurrence, the set is said to have more than one mode. In the case of the set of numbers having two modes, it is called bimodal, with three modes being trimodal. In general, a set of numbers with more than one mode is multimodal.

Median

The median is the middle value in a dataset that has been arranged in numerical order, that is, ordered from smallest to largest. If there is an odd number of observations, the median is the middle value. If there is an even number of observations, the median is the average of the two middle values.

For example, if a dataset contains the values 4, 5, 7, and 8, the median is 6 (the average of 5 and 7). Another example is if a dataset contains the values 4, 5, 7, 8, and 9; the median is 7 (the middle value).

Mean

The mean is the average value of the variable in a dataset. It is calculated by adding up all the values of the variable and dividing them by the total number of observations. The following denotes the formula to calculate the mean:

$$\overline{X} = \frac{\Sigma X}{n}$$

For example, if a dataset contains the values 1, 2, and 6, the mean is 3 because (1 + 2 + 6) / 3 = 3.

Note: Each of these measures of central tendency has its own strengths and limitations. They can be a useful basis for the characteristics of the dataset., as mentioned in the following:

- **For Categorical (Nominal or Ordinal) data, it is difficult to calculate the mean or median. Hence mode is the best tool for the purpose.**
- **For quantitative data, the mean or median is the best-suited measurement tool. This is because if the data has some influential outliers, in other words, in case the data is highly skewed, then the median is the best measurement for finding a central tendency. Or else Mean is to be applied.**

Range, interquartile range, and box plot

Range, interquartile range, and box plots are statistical measures that are used to describe the spread or dispersion of a variable. These prove useful, especially for comparing groups having the same mean, median, and mode.

For example, in the following *Figure 7.20*, both the groups have their median as 10, mean as 10, and mode as 9.1, indicating that additional information is needed than measures of central tendency:

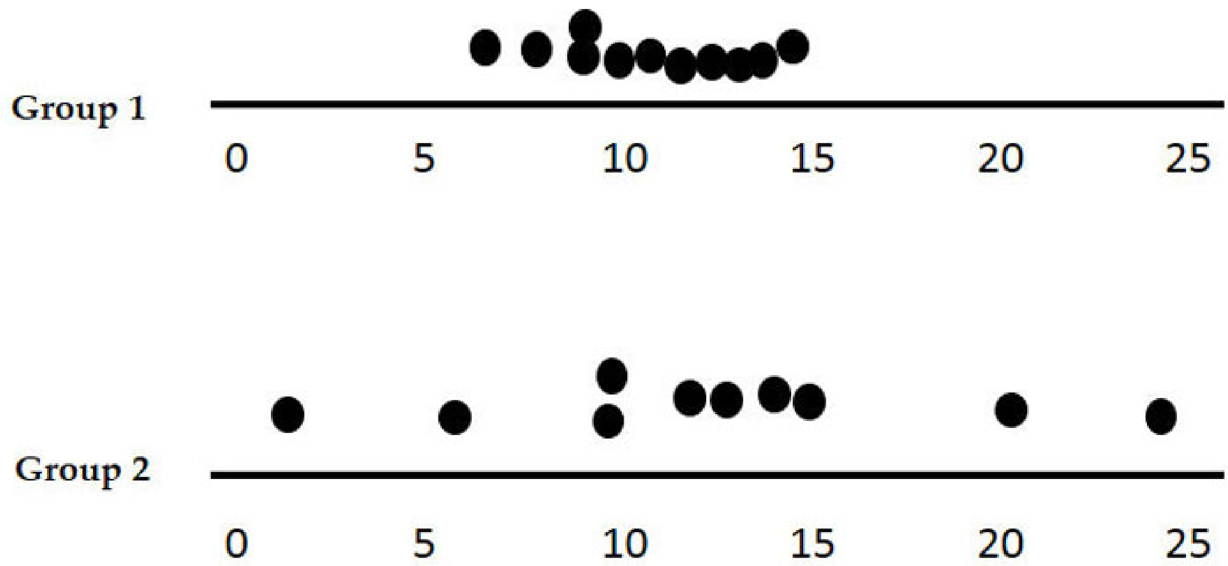


Figure 7.20: *Dispersion example*

The measure of variability provides additional insights into the data in such situations. The measure of variability provides an indication of the spread of values in a distribution. These methods are:

- Range
- Interquartile range
- Box plot

Range

The range is the simplest measure of variability. It is the difference between the highest and the lowest value.

For the preceding example, the range will be:

Range (group 1) = 14.3 – 5.8 = 3.5

Range (group 2) = 24.7-0.7 = 24

The range is a measure of the dispersion of data that gives an insight into its spread. However, this measure is not always reliable because it takes count of only extreme values and hence is also dependent on outliers (values that fall outside the range of the rest of the data).

Interquartile range

Interquartile range (**IQR**) is a more robust measure of dispersion than range. This is because it doesn't factor in extreme values or outliers.

The distribution is divided into four equal parts called quartiles. 1st quartile (Q1) represents the first 25%, 3rd quartile (Q3) represents the last 25%, while the middle one or 2nd quartile (Q2) divides the remaining middle 50% distribution into two equal parts. So basically, it is the same as the median. *Figure 7.21* explains the same visually:

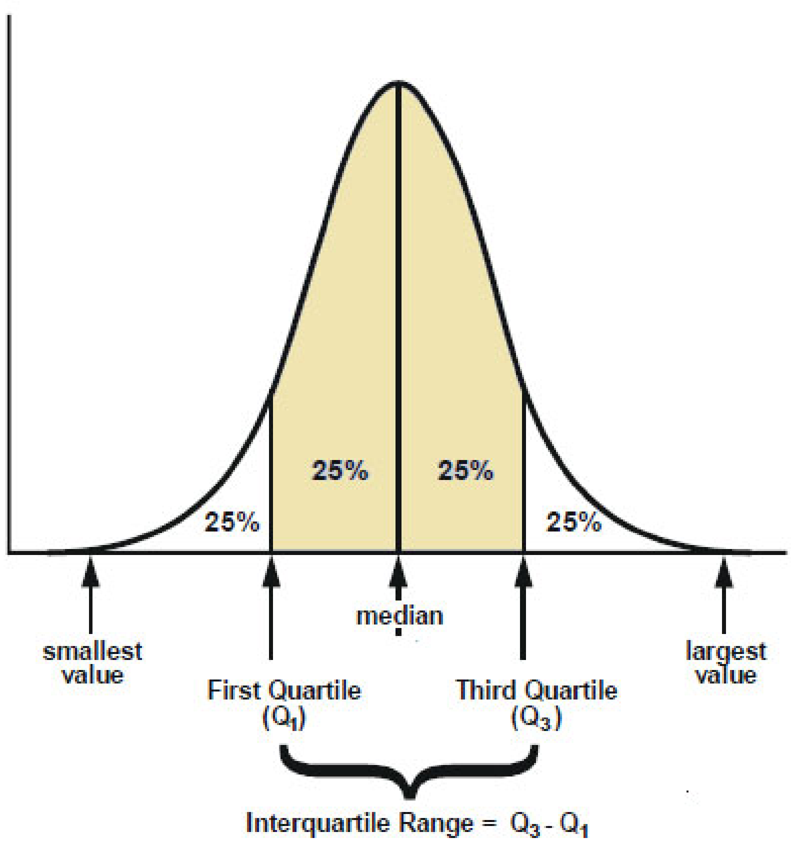


Figure 7.21: *IQR*

IQR is the distance between Q3 and Q1, or *IQR = Q3 – Q1*.

Steps to calculate IQR:

1. Arrange data in ascending order, that is, from low to high.
2. Find the median or Q2.
3. Find Q1 by finding the median of the left side of Q2 values.
4. Find Q1 by finding the median of the right side of Q2 values.
5. Calculate IQR by subtracting Q1 from Q3.

For example, *Figure 7.22* shows a variable with an odd number of observations, its Q1, Q2, and Q3, and the resulting IQR:

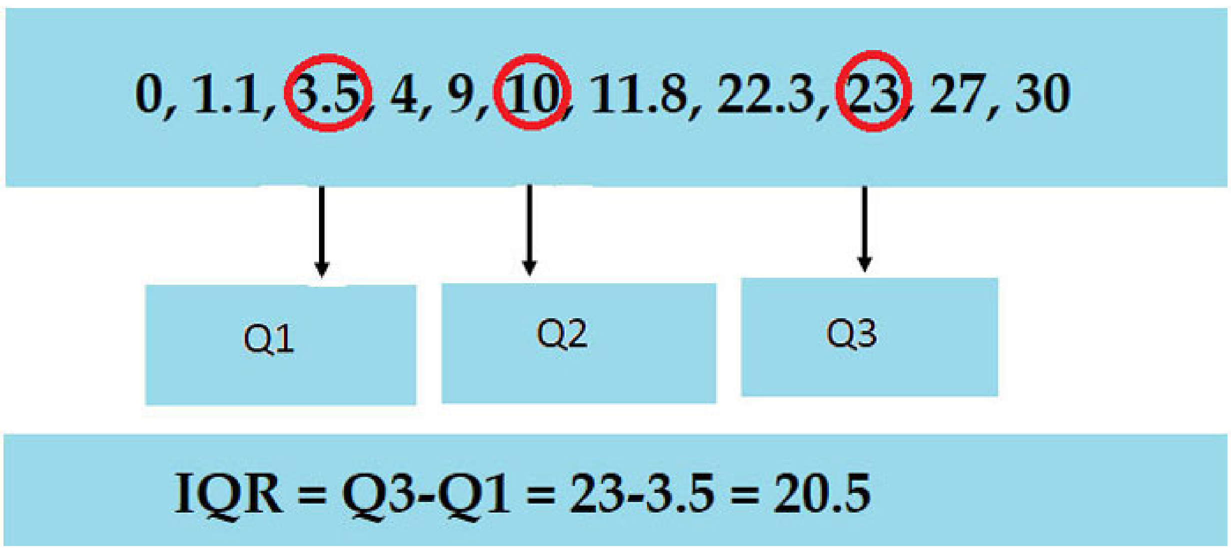


Figure 7.22: *Example of IQR*

For an even number of observations, Q1, Q2, and Q3 are calculated as the median is calculated, that is, consider the two central values and take their mean.

Observations can be termed as outliers if they lie below (Q1 – 1.5* IQR) OR above (Q3 + 1.5*IQR)

Box plots

Box plots, also known as box-and-whisker plots, are not only helpful for identifying the shape and spread of a distribution but also useful for comparing the distributions of various groups.

These represent the five-number summary of a variable that includes minimum, lower quartile, median, upper quartile, and maximum values. This is depicted in *[Figure 7.23](#)* visually:

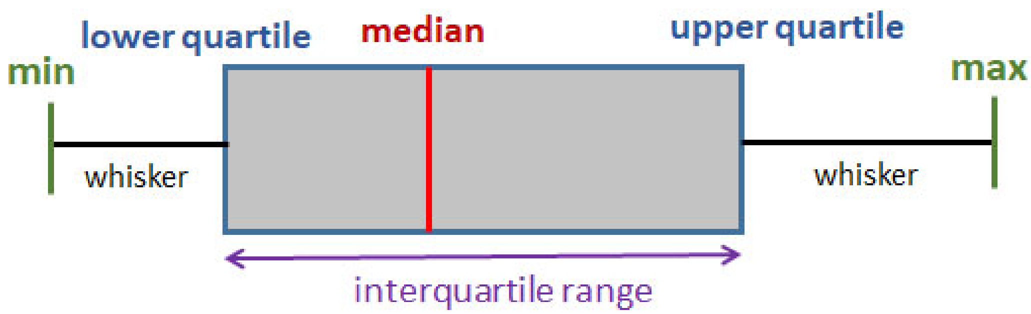


Figure 7.23: *Box plot*

[Figure 7.24](#) is an example of a representation of company and sales data in box plot visualization:

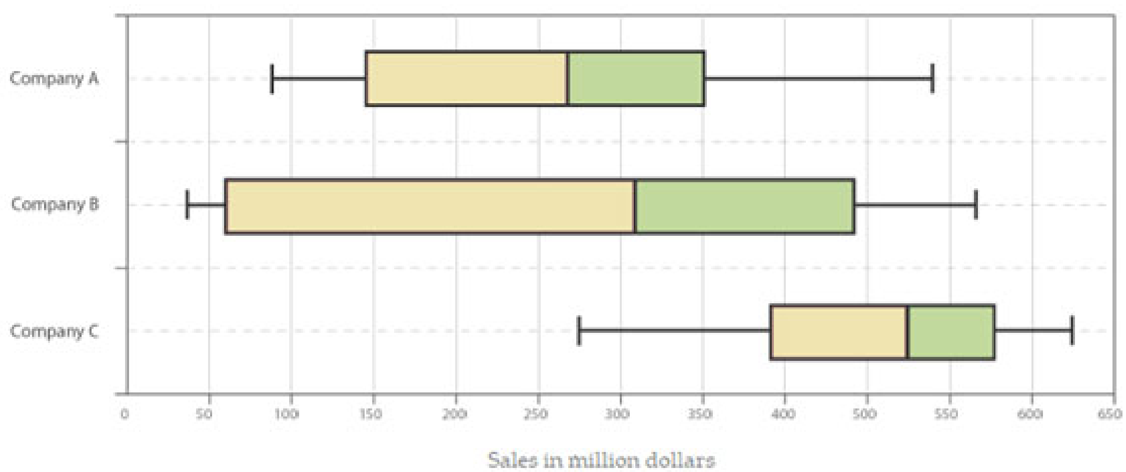


Figure 7.24: *Example of a box plot*

Variance and standard deviation

There are two measures of variability that consider all the values of a variable. These are as follows:

- Variance
- Standard deviation

Table 7.8 describes the formulas for variance and standard deviation for sample and population data sets.

	Samples	Population
Variance	$s^2 = \frac{\Sigma (x - \bar{x})^2}{n-1}$	$\sigma^2 = \frac{\Sigma (x - \bar{x})^2}{n}$
Standard deviation	$s = \sqrt{s^2}$	$\sigma = \sqrt{\sigma^2}$
Formula	$s^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1}$	$\sigma^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n}$

Table 7.8: *Variance and standard deviation*

Where,

- The population represents the entire group
- The sample is part of the population considered to describe characteristics of a whole group or an entire population
- X is an individual one value
- N is the size of the population
- $\bar{x}$ is the mean of the population

Notes:

- **A high variance indicates larger variability. That is, more values are spread out around the mean value.**
- **The average distance of an observation from the mean is represented by the Standard deviation**
- **A larger standard deviation indicates larger variability of the data.**

Z-scores

Z-score or standardized score is a statistical analysis method to know the number of standard deviations a data element falls from the mean. The formula for calculating the Z-score is as follows:

$$Z = \frac{x-\mu}{\sigma}$$

$$Z = \frac{Raw\ score - Mean}{Standard\ deviation}$$

Where**,**

Z is the z-score

X is the value of the data element

μ is the population mean

σ is the standard deviation

Interpretations from Z-score

- Whether a specific observation is common or exceptional.
- Negative z-scores represent values below the mean.
- Positive z-scores represent values above the mean.
- The sum of all z-scores should be zero, as positive and negative values cancel each other.
- For extremely right-skewed data, Z-score will be a large positive number.
- For extremely left-skewed data, Z-score will be a large negative number.
- Mean has a z-score of value zero.
- The values of |Z|, if greater than 2, indicate unusual or exceptional distribution.

Conclusion

By the end of this chapter, we can present problems in terms of numbers. We have learned data structures, statistical principles, and the application of these concepts in representing data in terms of graphs and statistical models.

In the next chapter, we learn about relationships between variables using correlation and regression techniques.

Multiple choice questions

1. **In terms of machine learning algorithms, which of the following is not a data type?**
    a. Unstructured data
    b. Semi-structured data
    c. Structured data
    d. Static data
2. **What is the sequence of levels of measurement?**
    a. Nominal->Ordinal->Interval->Ratio
    b. Ordinal->Normal->Interval->Ratio
    c. Ratio->Nominal->Ordinal->Interval
    d. Nominal->Ordinal->Ratio-> Interval
3. **Data and time type of structure data represent which of the following?**
    a. Specific time
    b. Time range
    c. None of above
    d. Both a and b

Answers

1. **d**
2. **a**
3. **a**

Questions

1. Explain the Interquartile range (IQR).

2. Explain how levels of measurement for categorical data are cumulative.
3. Write down the differences between unstructured and structured data.

Practice exercise

Create a database of all the students in a specific class. The database should capture name, gender, favorite subject, number of hours put in daily on an average studying the subject, and marks obtained in exams for the subject.

1. Create a frequency table for your favorite subject.
2. Plot the best visualization for the number of hours versus marks obtained and mention your observation and data analysis, if any.
3. Find the mean, median, and mode for the favorite subject marks.

Join our book's Discord space

Join the book's Discord Workspace for Latest updates, Offers, Tech happenings around the world, New Release and Sessions with the Authors:

https://discord.bpbonline.com

CHAPTER 8
Regression

Introduction

Forecasting is a significant concept in data science. It finds its broad applicability in artificial intelligence, econometrics, and risk management by several professionals, including quantitative analysts, statistical modelers, and technologists.

This chapter aims to understand the statistical measures and techniques to understand and quantify the relationship between variables.

Structure

In this chapter, we will be discussing:

- Correlation and regression
    - Correlation
    - Regression
    - Crosstabs and scatterplots
    - Pearson's r
    - Regression - finding the line
    - Regression - describing the line
    - Regression - how good is the line
    - Correlation is not causation
- Examples of correlation and regression
    - Caveats and examples

Correlation and regression

Correlation and linear regression are the most commonly used statistical measurements for exploring the association between two quantitative variables assumed to have a linear relationship.

While correlation quantifies the linear relationship between a pair of variables, regression indicates the relationship as an equation.

For example, suppose a lady owns a luxury house; then it is assumed that she must be financially well. Correlation and regression are used to numerically quantify this relationship.

A large number of business applications exist that use regression analysis, especially in finance, where it is used by analysts to understand the markets for stocks and hedge funds analysis to see changes in interest rates and how the bond price will charge. It can also be used to see the impact on employee productivity based on various pieces of training imparted, evaluate trends and make estimates in businesses, and more.

Correlation

Correlation as a statistical measure describes the relationship between two variables by determining the strength and direction of the relationship and can range from -1 to 1.

It can be detailed as either strong or weak and as either positive or negative.

There are four types of correlation:

- **Positive Linear correlation**: When the variable on the x-axis increases as the variable on the y-axis increases.
- **Negative Linear correlation**: When the variable on the x-axis increases as the variable on the y-axis decreases.
- **Non-linear correlation (known as curvilinear correlation)**: When the relationship is determined by a non-linear curve, that is, it is not a straight line.
- **No correlation**: There is no relationship between the variables.

A positive correlation specified increase in one variable increases would also mean an increase in the other variable. For example, there might be a positive correlation between the sales of ceiling fans and the location temperature, such that when the temperature sores high, sales of ceiling fans increase. Refer to the following figure:

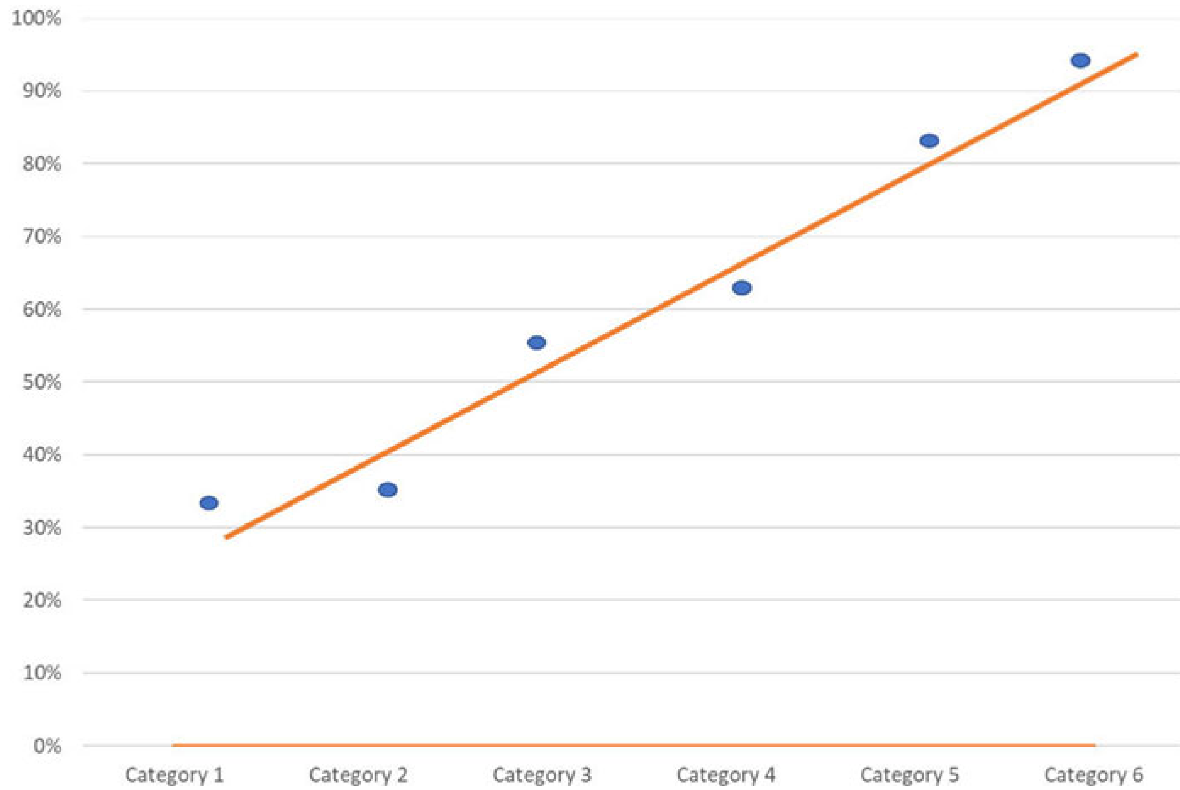


Figure 8.1: *Positive correlation*

A negative correlation signifies a decrease in one variable with an increase in the other variable. For example, there might be a negative correlation between the sales of heat radiators and local temperature, such that when the temperature drops, sales of heat radiators increase. Refer to the following figure:

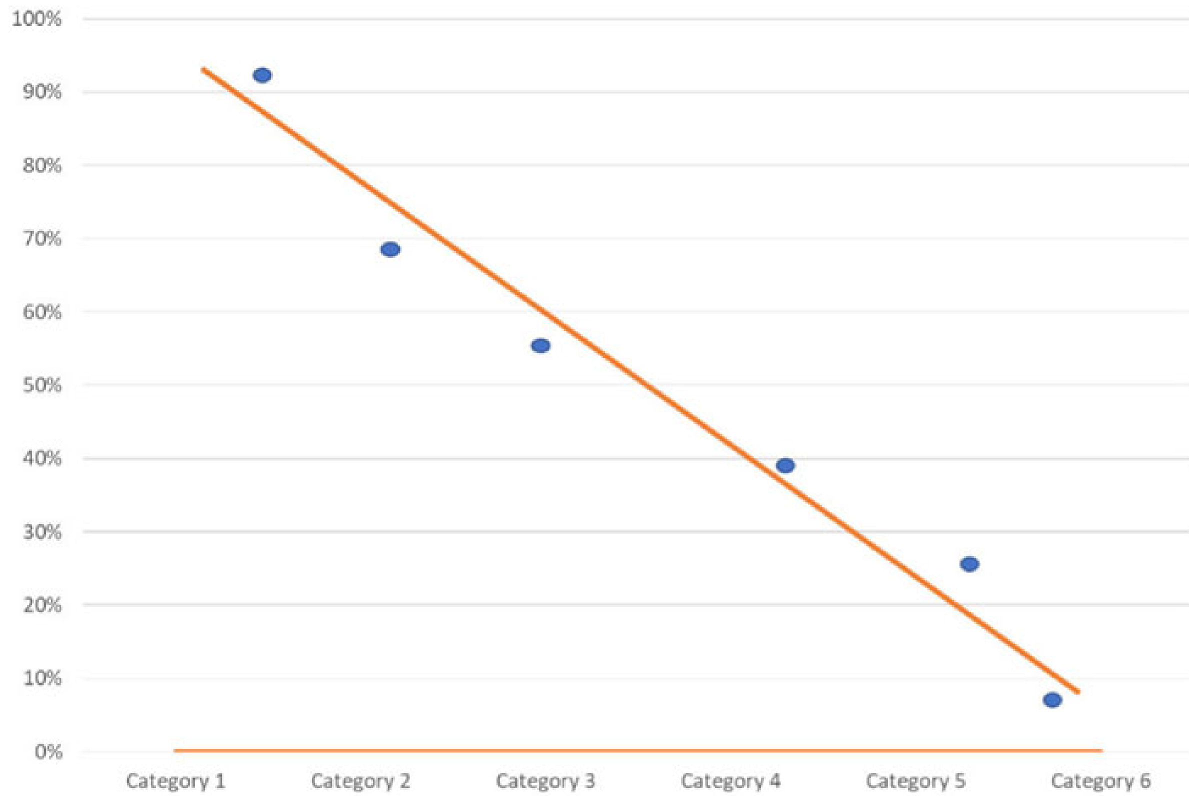


Figure 8.2: *Negative correlation*

Non-linear correlation (known as curvilinear correlation) is where there is a non-linear correlation between the variables. For example, the relationship between the restaurant's daily output of meals and the number of cooks is nonlinear. Refer to the following figure:

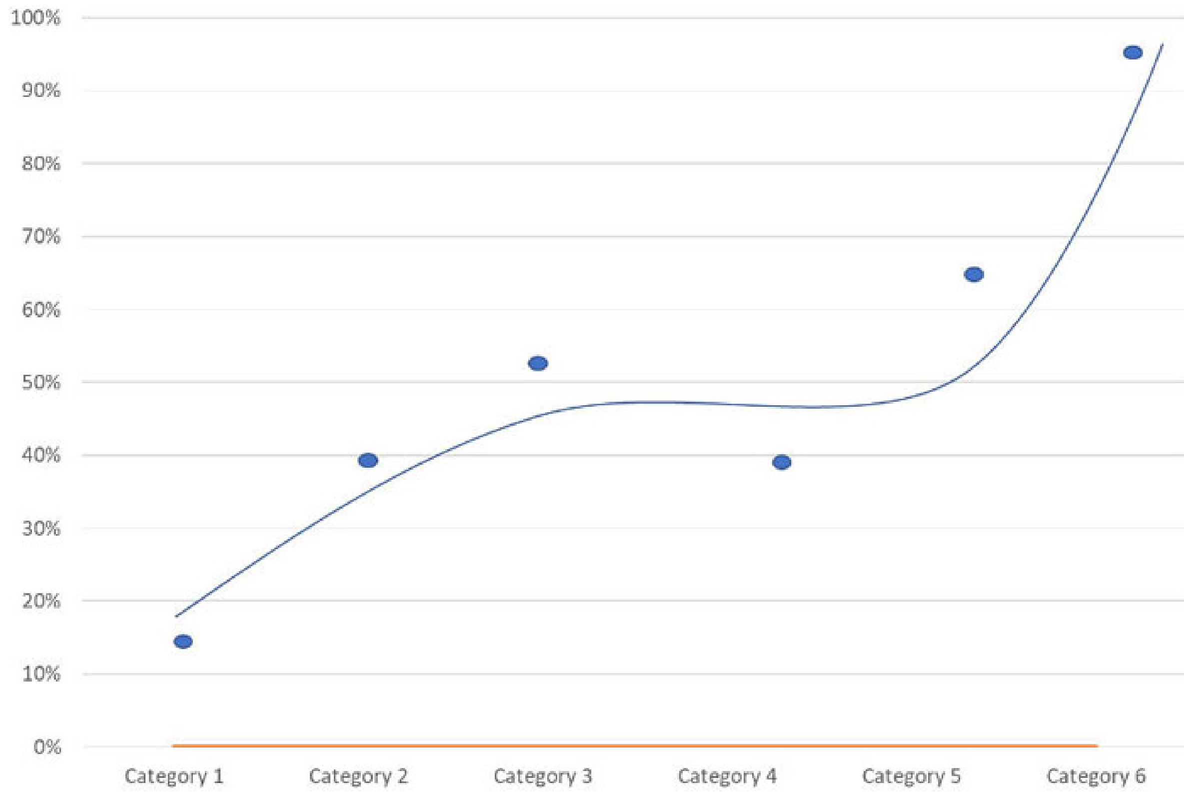


Figure 8.3: *non-linear correlation*

A correlation of 0 indicates that there is no relationship between the variables. For example, the number of sand particles on the road and the number of vehicles passing on the road on a daily basis have no correlation. Refer to the following figure:

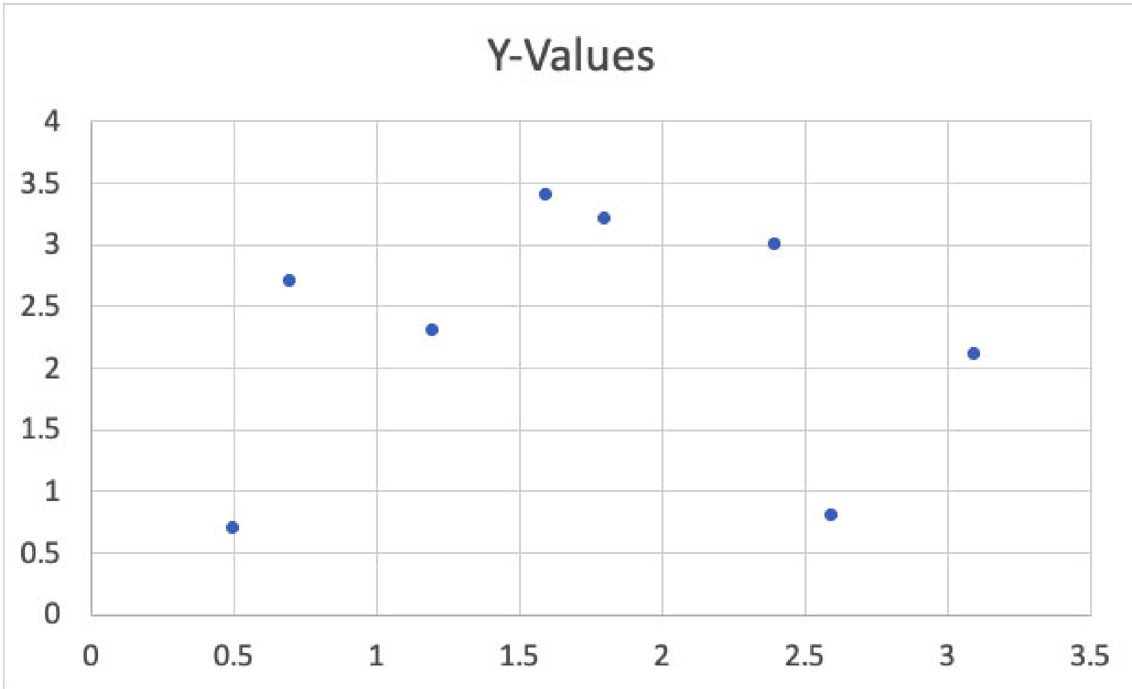


Figure 8.4: *No correlation*

Regression

The regression analysis concept has primarily three steps:

1. **Set a goal**
    - **Cash forecasting**: How much cash will the business have in hand in the month of said year
2. **Plot a scatter plot**
    - Collect the relevant data over a period of time
    - Plot a scatterplot chart
    - From the data, calculate the mean of cash at hand
    - Draw the best fit line between cash and time of year. It very well can be an iterative task as the aim of the best fit line is to minimize the difference between the line and the actual observation.
3. **Formulate equations to predict the dependent variable**
    - **Intercept**: On scatter, plot using the x-axis and y-axis the point where the line crosses the y-axis with x = 0.

- **Slope**: Expected change in y over unit change in x
- Distance – or residual, which is the distance between an actual data point and the best fit data point

This chapter covers the previously mentioned methodologies in detail.

Crosstabs and scatterplots

The crosstab and scatterplots are used to describe patterns across multiple variables, expressing how one variable may change (or correlate) with another.

Crosstabs

Cross-tabulation is a statistical technique that is used to analyze categorical data. Categorical data is data or variables that can be classified into different mutually exclusive categories. An example of categorical data is hair color.

Cross tabs (also known as contingency tables) display the relationship between two or more variables in the form of a table. They are used to determine the possible association between the variables. Crosstab is a great tool for presenting multidimensional data.

Example contingency table

Table 8.1 describes a crosstab that shows the number of persons in various age groups across genders:

	Number of persons	
Age group	**Male**	**Female**
18-22	30	32
23-32	45	40
33-42	30	28
43-52	20	25
53-62	29	25
62+	30	35

Table 8.1: *Crosstab*

For example, there are 30 people in the age group 18-22 who are male and 32 people in the same age group who are female.

Crosstabs are useful in not only understanding the relationship between different variables but also possibly identifying patterns or trends in the data and probabilities within data sets.

Crosstabs with more than two variables

Let us consider crosstabs that contain more than two variables. For example, *Table 8.2* shows four variables. The rows represent one categorical variable that identifies drink preference, and the columns represent age and income within gender.

Drink	18-30 yrs	30—50 yrs	50+ yrs	Males under 50 lacs p.a	Males above 50 lacs p.a	Females under 50 lacs p.a	Females above 50 lacs p.a
Coke	24%	22%	14%	22%	4%	21%	3%
Pepsi	16%	28%	6%	8%	6%	9%	8%
Red Bull	30%	15%	40%	20%	38%	22%	37%
Sprite	10%	25%	18%	35%	27%	38%	28%
Monster	20%	10%	22%	15%	25%	10%	24%
NET	100%	100%	100%	100%	100%	100%	100%

Table 8.2: *Crosstab with more than two variables*

Crosstabs are routinely created with many more variables. For example, each row and each column may represent a different variable.

Scatterplots

A scatterplot (also known as a scatter graph or a scatter chart) is a graphical representation of two variables with values plotted as dots on a graph.

The position of each dot on the horizontal and vertical axis indicates values for an individual data point. That is, each observation is represented by a dot on the graph, with the position of the dot being determined by the values of the two variables.

Scatterplots are used to observe the relationship between two variables and identify any patterns or trends or see if there is any overlap between two sets of data. Refer to the following figure:

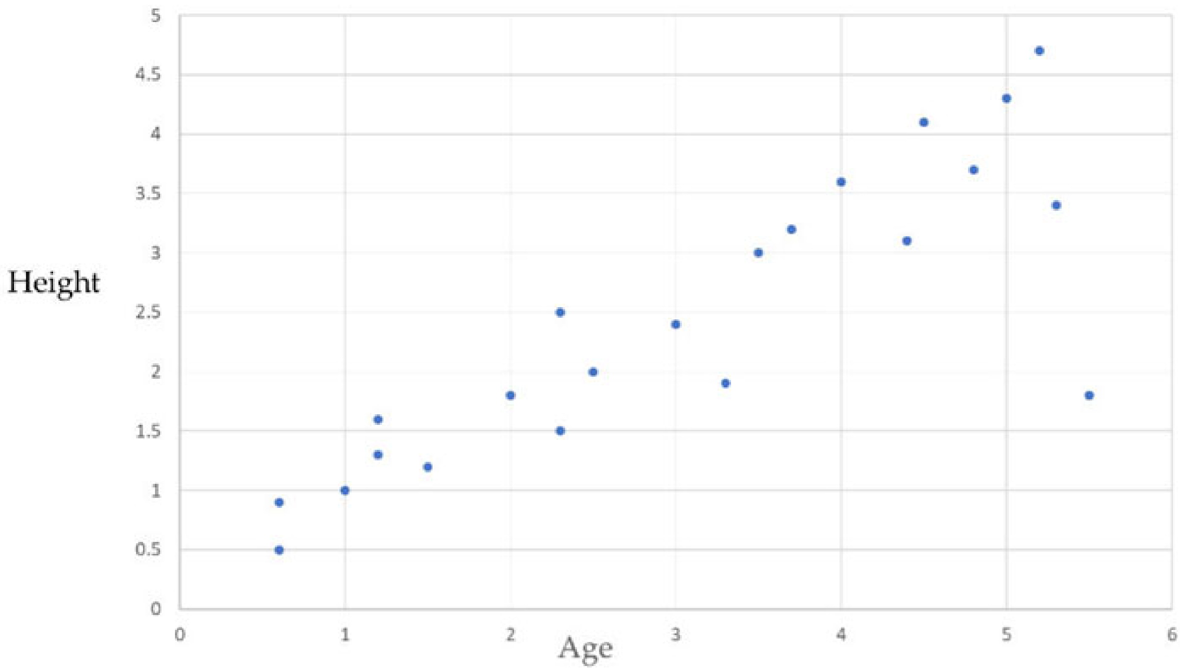


Figure 8.5: *Scatterplot of age and height*

Let’s say the scatterplot in graph 1 is a visualization of the relationship between two variables: X and Y. Each dot on the graph represents a single observation, with the position of the dot determined by the values of X and Y for that observation.

In graph 1, the height and age of the kids are plotted. X-axis represents the age in years while the X-axis represents the height in feet. Each dot represents a single kid. Each dot’s horizontal position indicates that kid’s age (in years), and the vertical position indicates the kid’s height (in feet).

From the plot, we can observe a general positive correlation between a kid’s age and a kid’s height, as the dots tend to fall along a line that ascends from left to right. Another point to observe is an outlier dot, a kid who is much older than the others but appears fairly short for their age, which might warrant further investigation.

Pearson's r

Pearson's r, also known as Pearson's correlation coefficient, is a measure of the strength and trend of the linear association between two variables. In other words, it determines if there is any linear component in the relationship between the two variables.

Pearson's correlation coefficient, as denoted by r, with values of r ranging from -1 to 1. The values of r closer to 1 indicate a strong positive correlation, values closer to -1 indicate a strong negative correlation, while values equal to zero indicate no linear correlation or association between the variables. Thus, a correlation coefficient of 0.83 indicates a stronger positive correlation than a value of 0.56. Similarly, a correlation coefficient of -0.38 indicates a stronger negative correlation than a correlation coefficient of -0.11

For values of the correlation coefficient:

- *A +1* value determines a perfect positive relationship between the variables
- *A -1* value determines a perfect negative relationship between the variables
- *A 0* value determines no relationship exists between the variables

$$r = \frac{\Sigma(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\Sigma(x_i - \overline{x})^2 \Sigma(y_i - \overline{y})^2}}$$

Where,

r = correlation coefficient

x_i = values of the x-variable in a sample

y_i = mean of the values of the x-variable

$\tilde{x}$ = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

Example of Pearson's r

Let's see the steps to calculate the Pearson correlation coefficient:

1. n represents the number of observations, let n = 3.
2. List the variables x and y, as in *Table 8.3*:

x	y
2	1
3	5
4	2

Table 8.3: *variables x and y*

3. Find product of x * y in 3rd column, as in *Table 8.4*:

x	y	x * y
2	1	2
3	5	15
4	2	8

Table 8.4: *Product of x * y*

4. Find x^2 and y^2 and mention in 4th and 5th columns, as in *Table 8.5*:

x	y	x * y	x^2	y^2
2	1	2	4	1
3	5	15	9	25
4	2	8	6	4

Table 8.5: *x^2 and y^2*

5. Find the sum of x, y, x * y, x^2, and y^2 variables and mention it in the last row (as marked in bold in the figure):

x	y	x * y	x^2	y^2
2	1	2	4	1
3	5	15	9	25
4	2	8	16	4
9	8	25	29	30

Table 8.6: *Sum of values x, y, x * y and x^2 and y^2*

6. Insert the values in the formula mentioned and solve it:

$$r = \left(\frac{(3 \times 25 - 9 \times 8)}{\sqrt{(3 \times 29 - 9^2)\,(3 \times 30 - 8^2)}}\right) = 0.019$$

Advantages

- Helps in analyzing the strength of the relationship between two variables.
- Also determines the extent to which the variables are correlated.
- Helps ascertain the direction of correlation, whether positive or negative.

Disadvantages

- The fact that the correlation coefficient between the variables X and Y or Y and X is the same, Pearson's r proves insufficient to differentiate between dependent and independent variables. For example, a person who exercises loses weight. But a person having lost weight need not necessarily be one who exercises.
- Doesn't provide information about the slope of the line. Rather only states the existence of a relationship between the two variables, if any.
- In the case of homogenous data, it may be likely to be misinterpreted.
- It takes much time to arrive at results as compared to other methods.

Important points

- Pearson's r is independent of units of the two variables.
- Pearson's r is symmetric between variables, which means the value of r between X and Y or Y and X remains the same.

Regression - Finding the line

To model the relationship between variables, the regression analysis technique aims to find the line (or a curve) that best fits the data in a way that can be used to make predictions about the dependent variable based on the independent variables.

There are six different types of regression analysis, including linear regression, logistic regression, and polynomial regression. This chapter is focused on linear regression.

In linear regression, the goal is to find the straight line, also called the regression line, that completely fits the data such that the overall distance from the line to the observation points outlined on the graph is the smallest.

The best-fitting line (or regression line) can be expressed with the formula $y = mx + b$, where m is the slope of the line and b is the y-intercept.

The equation $y = mx + b$ is a formula used for any straight line. However, note that in the case of regression, which is a statistical method, the observation points do not lie in a straight line. That is, if a linear pattern exists, the line is a model around which the data lie.

Now let's understand the equation $y = mx + b$ or, in other words, the way to describe the regression line:

Slope

Here m is the slope of the line, which defines the change of y over the change in x.

For example, a slope value of m = 5/9 means as the x-value moves (in the right direction) by 9 units, the y-value moves in the upward direction by 5 units.

Intercept

The y-intercept is the value on the y-axis where the regression line crosses. Note that the value of x-coordinates at the point where a line meets the y-axis is always 0.

For example, in the equation $y = 5x - 9$, the line crosses the y-axis at the value $b = -9$. The coordinates of this point are (0, –9).

Regression - Describing the line

The best–fitting regression line need not necessarily be found by hits and trials from the multiple options through eyeballing a line on the scatterplot.

Slope

The best-fitting line has a definite slope and a y-intercept that can be calculated using specific formulas, which is described as follows:

$$m = r\left(\frac{S_y}{S_x}\right)$$

Where,

m is the slope of the line

The standard deviation of the x values (denoted s_x)

The standard deviation of the y values (denoted s_y)

The correlation between X and Y (denoted r)

Note:

The slope can be negative, indicating the line is going downhill. For example, an increase in police patrolling resulted in a decrease in the number of crimes committed.

The correlation and the slope of the best–fitting line are not the same. While correlation is a unitless number, slope attaches units to the correlation.

y-intercept

The formula to calculate the y-intercept of the best fitting is as follows:

$$b = \bar{y} - m\bar{x}$$

where b is the y-intercept

where the mean of x-values as denoted by $\bar{x}$

and mean of y-values as denoted by $\bar{y}$

and m is the slope of the line

So, the slope of the line is always to be calculated first before calculating the y-intercept.

Where there are more than one independent variables, that is, in multiple regression models, the equation of the line would be more complex, with additional elements for each independent variable.

Let us revisit the correlation and regression formulas.

Pearson's correlation coefficient r is calculated as:

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \, \Sigma(y_i - \bar{y})^2}}$$

Ordinary Least Squares (OLS) Linear Regression is calculated as:

The straight-line equation is:

$$y = \alpha + \beta X$$

OLS is represented by β, where,

$$\beta = \frac{\sum_1^n(\boldsymbol{x_i} - \bar{\boldsymbol{x}})(\boldsymbol{y_i} - \bar{\boldsymbol{y}})}{\sum_1^n(\boldsymbol{x_i} - \bar{\boldsymbol{x}})^2}$$

$$\beta = r_{xy} \frac{\sigma_y}{\sigma_x}$$

$$\alpha = \bar{y} - \beta\bar{x}$$

Where,

x is the independent variables

x- is the average of independent variables

y is the dependent variables

y- is the average of dependent variables

σ_x is the standard deviation of

σ_y is the standard deviation of

n is the number of data points in the data sets

Residual

Residual, or noise or error terms, indicate deviation of dependent values (Y values) from each expected value. In other words, Residual describes how good our model is against the actual value.

Ordinary Least Square helps us formulate the most suitable regression model. Using **Ordinary Least** (**OLS**) methodology and rules thus help us find the best-fit line or estimate the unknown parameters in a linear regression model.

Common OLS assumptions are as follows:

- **Errors**
    - Are independent of x
    - Have a constant variance
    - Their mean is 0
    - Are uncorrelated with each other
    - Have a normal distribution
- Large outliers are not considered observation points in the data
- Y and X variables have a linear relationship
- No appropriate independent variables have been omitted from the Model

The best fit line is the line that minimizes the sum of squared differences between actual and forecasted results. Smaller the value, the better the regression model. **Mean Squared Error** (**MSE**) is the average of the minimum sum of squared difference. MSE is calculated as follows:

- **Explained Sum Of Squares** (**ESS**): Squaring(Y Estimated — Mean value of Y) and then sum all of the values.
- **Sum Of Squared Residuals** (**SSR**): Squaring(Y Estimated — Actual Y Value) And then sum all values.
- **Total Sum of Squares**, (**TSS**): ESS + SSR or the total sum of squares is the sum of all squares of (y estimated — residual) + sum of squares of (y estimated — mean y).

Note that Squaring values take care of negative values.

Now that the best-fit line has been described using formulas and we get the size of residuals. Let us also consider the line described in terms of its goodness of fit. How good is the line?

Regression - How good is the line

It's important to evaluate how good the regression fit is. In a perfect condition, the points are expected to lie on the 45 degrees line passing through the origin (y = x is the equation). The nearer the points to this line, the better the regression.

There are several measures that can be used to assess how good a line fit by regression analysis is. One of the measures to assess Regression fitness is the R-squared value, also known as the coefficient of determination, and adjusted R-squared.

R-Squared: goodness of fit

The R-squared value ranges from 0 to 1, with values closer to 1 indicating a better fit. In other words, the higher R squared, the better the fitness.

R-squared represents the proportion of the variance in the dependent variable that is guided by the independent variables.

$$R\ Squared = 1 - \left(\frac{SSR}{TSS}\right)$$

OR:

$$R\ Squared = 1 - \frac{Variance\ (residual)}{Variance\ (y)}$$

Adjusted R Squared

R squared lacks taking into account the number of variables that give the degree of determination. Hence R Squared by itself is not good enough. Therefore, adjusted R squared is calculated to measure the quality of the regression model.

$$Adjusted\ R\ Squared = \frac{\frac{SSR}{(n-k)}}{SST\ (n-1)}$$

OR:

$$Adjusted\ R\ Squared = 1 - \left[\left[\frac{n-1}{n-k-1}\right] \times [1 - R^2]\right]$$

Where,

n = number of observations,

k = number of independent variables

PS: Adjusted R square is always lower than the R-squared.

Another measure of how good the line is the standard error of the estimate. The standard error of the estimate is a measure of the amount of error or uncertainty in the predictions made by the model. A smaller standard error of the estimate indicates a better fit.

Please note that no single measure can fully represent the goodness of fit of a regression model. Rather, a combination of these measures should be used to get a holistic view of how well the model fits the data.

Correlation is not causation

In statistical terminology, correlation refers to a relationship between or interdependence of variables, such that they tend to vary together in a predictable way if the correlation is non-zero.

For example, there may be a positive correlation between the number of hours an athlete practices and their ranks, meaning that as the number of hours spent practicing increases, their ranks tend to improve.

Causation, on the other hand, means the action of causing something. Please note that correlation does not necessarily imply causation. In other words, if two variables are correlated, it does not mean that one variable is causing the other. Their relationship may be attributed to other factors, or the correlation may be due to chance.

For example, there may be a correlation between the number of cold drinks purchased and the amount of electricity bill in a given month. This does not mean that increase in cold drinks purchases caused higher electricity bills, but rather that a third factor may influence both variables, such as the temperature. In this case, the hot weather may be causing both an increase in cold drink purchases and an increase in the number of hours of running air conditioning, causing an increase in the electricity bill, leading to a correlation between the two variables. Refer to the following figure:

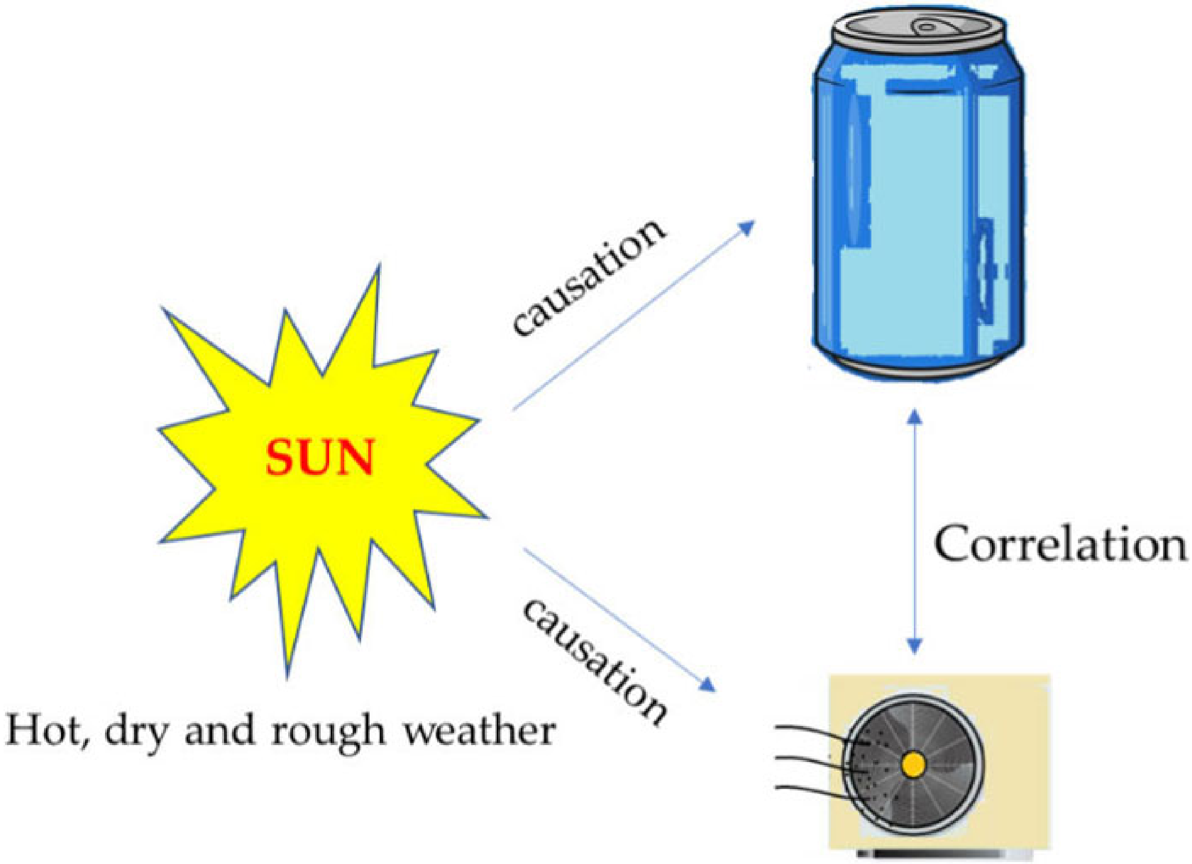


Figure 8.6: *Correlation is not causation*

Examples of correlation and regression

Let us consider examples of correlation and regression.

Example 1

Goal: We have two data sets regarding a person. One describes the length of feet, while the other describes the height. We need to determine the correlation between the length of feet and the height of a person. *Table 8.7* captures the feet and height of multiple persons:

Person	Feet (length in centimeters)	Height (length in centimeters)
A	23	170
B	28	190
C	15	130
D	19	134

E	21	169

Table 8.7: *Feet and height of multiple persons*

Solution

Find Pearson's r as in *Table 8.8*:

Person	Feet (length in centimeters) (x)	Height (length in centimeters) (y)	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	Squared $(x_i - \bar{x})$	Squared $(y_i - \bar{y})$
A	23	170	1.8	11.4	20.52	3.24	129.96
B	28	190	6.8	31.4	213.52	46.24	985.96
C	15	130	-6.2	-28.6	177.32	38.44	817.96
D	19	134	-2.2	-24.6	54.12	4.84	605.16
E	21	169	-0.2	10.4	-2.08	0.04	108.16
Average	21.2	158.6		Total	463.4	92.8	2647.2

Table 8.8: *Find Pearson's r*

Using formula

$$10. \quad r = \frac{\Sigma\,(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\Sigma (x_i - \bar{x})^2\,(y_i - \bar{y})^2}} = 0.93$$

Answer: The data set has a high positive correlation, and the length of feet and height are strongly correlated.

Example 2

Find the equation of the regression line for the following data as in *Table 8.9*:

Person	Weight	Diabetes
A	190	126
B	175	100
C	168	160
D	146	98
E	184	90

Table 8.9: Weight and Diabetes of multiple persons

Refer to the following *Table 8.10*:

Person	Weight	Diabetes	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	Squared $(x_i - \bar{x})$	Squared $(y_i - \bar{y})$
A	190	126	17.4	11.2	194.88	302.76	125.44
B	175	100	2.4	-14.8	-35.52	5.76	219.04
C	168	160	-4.6	45.2	-207.92	21.16	2043.04
D	146	98	-26.6	-16.8	446.88	707.56	282.24
E	184	90	11.4	-24.8	-282.72	129.96	615.04
Average	172.6	114.8		Total	115.6	1167.2	3284.8

***Table 8.10:** Pearson's r*

Using formula

$$r = \frac{\Sigma\,(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{(x_i - \bar{x})^2\,(y_i - \bar{y})^2}} = 0.059$$

Answer: The data set has a positive correlation, and that weight and diabetes are weekly correlated.

Now the regression line can be calculated as follows:

$$\sigma x = \sqrt{\frac{\Sigma\,(x_i - \bar{x})^2}{n-1}} = 17.08$$

$$\sigma y = \sqrt{\frac{\Sigma\,(y_i - \bar{y})^2}{n-1}} = 28.65$$

$$r = 0.059$$

$$\beta = rxy\left(\frac{\sigma x}{\sigma y}\right) = 0.059 \times \left(\frac{17.08}{28.65}\right) = 0.035$$

$$\alpha = \bar{y} - \beta\bar{x} = 114.8 - 0.035 \times 172.6 = 108.75$$

Equation of regression line: y = 108.75 + 0.035x

Notes on Correlation and Regression

- **Both are statistical measurements that are used to quantify the strength of the linear relationship between two variables.**
- **While correlation determines if a linear relationship exists between two variables, regression describes the interconnection between the two.**
- **Pearson's correlation coefficient and the ordinary least squares method are used to perform correlation and regression analysis.**

Let us also take down the difference between correlation and regression as in *Table 8.11*:

Correlation	Regression
Determines whether variables are related or not	Describes how a dependent variable changes with a change in the independent variable
Tries to establish positive linear, negative linear or non-linear, or zero relationships between variables	It can be Linear Regression, Logistic Regression, Ridge Regression, Lasso Regression, Polynomial Regression, or Bayesian Linear Regression.
Variables can be used interchangeably; that is, it's symmetric	Variables cannot be interchanged
Determines the strength of the relationship using positive or negative numerical values	It describes the impact of change on a dependent variable basis change in an independent variable
Pearson's coefficient r is one of the measures of correlation.	The least-squares method is one of the techniques to determine the regression line.
A scatterplot displays the strength, direction, and form of the relationship, while a correlation coefficient measures the strength of that relationship	Regression lines, or the best fit lines, on scatterplots show the overall trend of a set of data

Table 8.11: *Difference between correlation and regression*

Caveats and examples

There are several things to keep in mind when working with regression and correlation:

- Correlation does not imply causation. Just because two variables are correlated does not mean that one causes the other. There could be a third variable that is causing both of them.

  For example, when RAM is 100% in use, the phone hangs as well as the camera doesn't work. There is a correlation between the camera doesn't work and the event that the phone hangs; however, one doesn't cause the other. There is a third variable, that is, RAM and its usage, that is causing the two. Refer to the following figure:

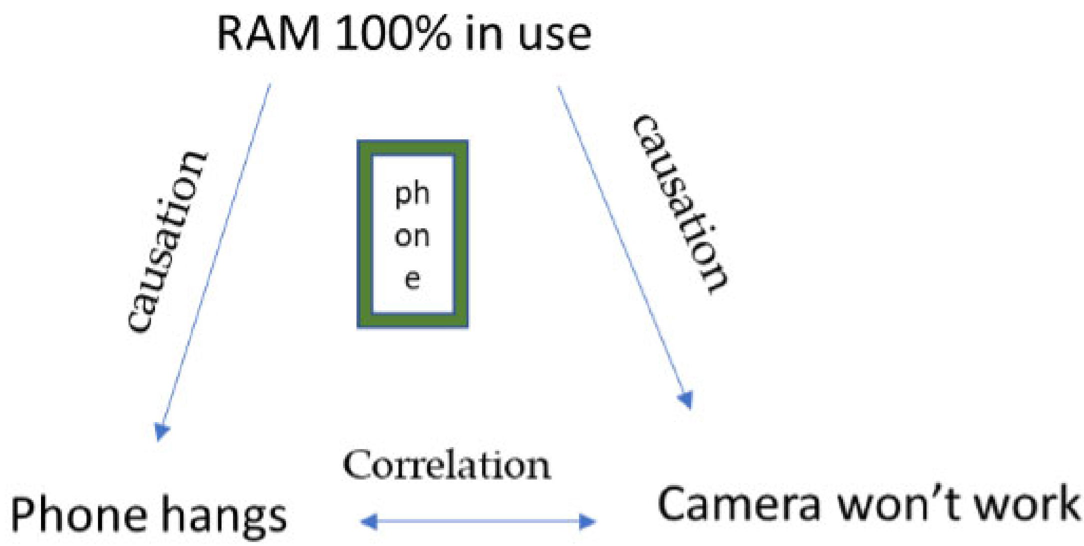


Figure 8.7: *Phone, Camera, and RAM*

- Linear regression is based on the assumption that the relationship between the variables is linear. If the relationship is non-linear, linear regression may not be the most appropriate method.

  For example, a linear regression model would assume that the relationship between a person's height based on their age is linear. However, children grow in height at a faster rate than adults, so the relationship between age and height may be non-linear. In this case, a non-linear model might be more appropriate.

- Outliers can have a big impact on the results of linear regression. A single outlier can significantly change the slope and intercept of the regression line.

  For example, let's say we are using linear regression to predict the price of a house based on its size. Most of the houses in our dataset range in size from 1,000 to 3,000 square feet and have a price ranging from \$100,000 to \$300,000. However, there is one house that is 10,000 square feet and has a price of \$1 million. This outlier will have a big impact on the regression line and may not be representative of the true relationship between size and price for the majority of houses in the dataset.

- The regression line is only an approximation of the true relationship between the variables. It is not a guarantee of the actual relationship.

  For example, let's say we use linear regression to predict a person's weight based on their height. The regression line represents an increase of 1 pound in weight for every 0.5 inches increase in height. However, some people may be heavier or lighter than a forecast based on their height. As such, the linear regression tells us only an estimate, and the actual relationship between height and weight may not be the same.

- Regression and correlation are highly sensitive to the sample size. A larger sample size is generally considered more reliable, but a small sample size that is representative of the population still yields useful results.

  For example, for using linear regression to predict the percentage of college students based on the study hours per day, with a data sample of 10 students, our regression line may not be very reliable, as it is based on small sample size. However, if we collect a larger sample size, say data for 900 students, our regression line will be more reliable

- The regression line has its boundary and is only meaningful for the range of values that were used to fit the model. It may not be suitable for forecasting values outside of this range.

  For example, let the linear regression model predict a person's income based on their years of experience. If the model is based on data for people with 0-15 years of experience, the model may not be reliable for predicting the income of a person with 25 years of experience. The regression line is only meaningful for the range of values used to fit the model (in this case, 0-15 years of experience).

Conclusion

In this chapter, we learned about correlation and regression, and other related terms. We have acquired knowledge on how to relate data with regression and correlation. We also studied a few examples of the application of these statistical methods.

In the next chapter, we will study classification and clustering. These are the methodologies to categorize data into one or more classes based on the features.

Multiple choice questions

1. **Correlation is not causation. Is this statement true or false?**
    a. True
    b. False
2. **Cross tabs are a statistical technique to analyze the following data**
    a. Categorical
    b. Numerical
    c. Quantitative
    d. None of above
3. **Which of the following is not a linear correlation**
    a. Positive
    b. Negative
    c. Non-linear
    d. Scatter

Answers

1. **a**
2. **a**
3. **d**

Questions

Let us now do a practice exercise on correlation and regression.

1. Engineers and Salaries

    The R&D unit of a company wants to establish a relationship between the years of experience of engineers and their salaries. They collect the data from 30 employees, as in *Table 8.12*:

Years of experience	Salary (in lacs)	Years of experience	Salary (in lacs)
1	3	16	20
2	3.5	17	20.5
3	4	18	23
4	6	19	23.8

5	7	20	32
6	7.5	21	37
7	8.8	22	38
8	9.2	23	40
9	10	24	44
10	11.2	25	45
11	12	26	49
12	13	27	50
13	13.5	28	51
14	15	29	55
15	18	30	60

Table 8.12: *Number of years of experience and salary*

a. Establish the relationship between the years of experience and salary by finding the correlation coefficient.

b. Interpret the correlation coefficient.

c. Find the simple linear regression equation between the experience and salary of an engineer.

d. What salary should be offered to an engineer having 27 years of experience based on prediction using the regression equation

e. Compared with the actual salary at 27 years of experience, is the above prediction reliable?

f. Is the regression equation a good fit for the data? Explain why or why not.

2. Exercise and Body Mass Index

   A study revealed a negative correlation between the amount of time exercising and body mass index. The related correlation coefficient value is –0.76; what does this tell us about the time put into exercise and body mass index?

3. Advertising costs and sales.

   A company wants to know about the sales of its product each year and the advertising costs incurred towards promoting the product that year. The data collected is shown in *Table 8.13*:

Year	Advertising costs (in thousands dollars)	Sales (in million dollars)

2000	30	4
2001	45	5
2002	40	5
2003	55	6
2004	57	6.2
2005	54	5.8
2006	35	5

Table 8.13: *Advertising and sales per year*

a. Find the correlation coefficient between advertising costs and sales
b. Is the data sufficient to establish a reliable equation to predict accurately?
c. Explain the importance of data from the preceding examples. How does prediction change with changing data?

Join our book's Discord space

Join the book's Discord Workspace for Latest updates, Offers, Tech happenings around the world, New Release and Sessions with the Authors:

CHAPTER 9
Classification and Clustering

Introduction

Machine learning algorithms are classified into various categories based on the target information type and the nature of the problem statement that is to be solved. These algorithms may be identified as regression algorithms, clustering algorithms, and classification algorithms.

We have learned in past chapters about regression analysis which refers to evaluating the relationship between the dependent outcome variable and one or more independent variables.

In this chapter, we will learn, also by example, about classification and clustering, the data analysis techniques used to identify patterns and relationships in data by grouping and classifying in categories.

While regression and classification are supervised learning algorithms, clustering is an unsupervised learning algorithm. A regression problem has an output variable of interest, a consistent variable; classification involves data labeling, and clustering involves grouping similar data.

Both clustering and classification are useful tools for pattern identification in machine learning. These are often used with each other and in combination with other data analysis techniques. An example of a combination of classification and clustering is to cluster similar data points and then assign labels to clusters using classification.

Structure

In this chapter, we will be discussing:

- What is a classification problem
- Examples
    - Simple binary classification

- Introduction to binary classification with logistic regression
- True positives, true negatives, false positives, and false negatives
    - Where we should care more with examples
    - For example, a false negative of a disease detection can have different implications than a false positive, one will be more physical harm, and the other will be mental
- What is a clustering problem
- Why is it unsupervised
- Examples

What is a classification problem

Supervised learning problems are categorized into regression and classification problems. Both types of problems aim to construct a concise model that can forecast the value of the dependent attribute basis the independent attribute variables. The dependent attribute is numerical for regression, while it is categorical for classification. *Figure 9.1* describes the supervised learning target variable types and methods applied:

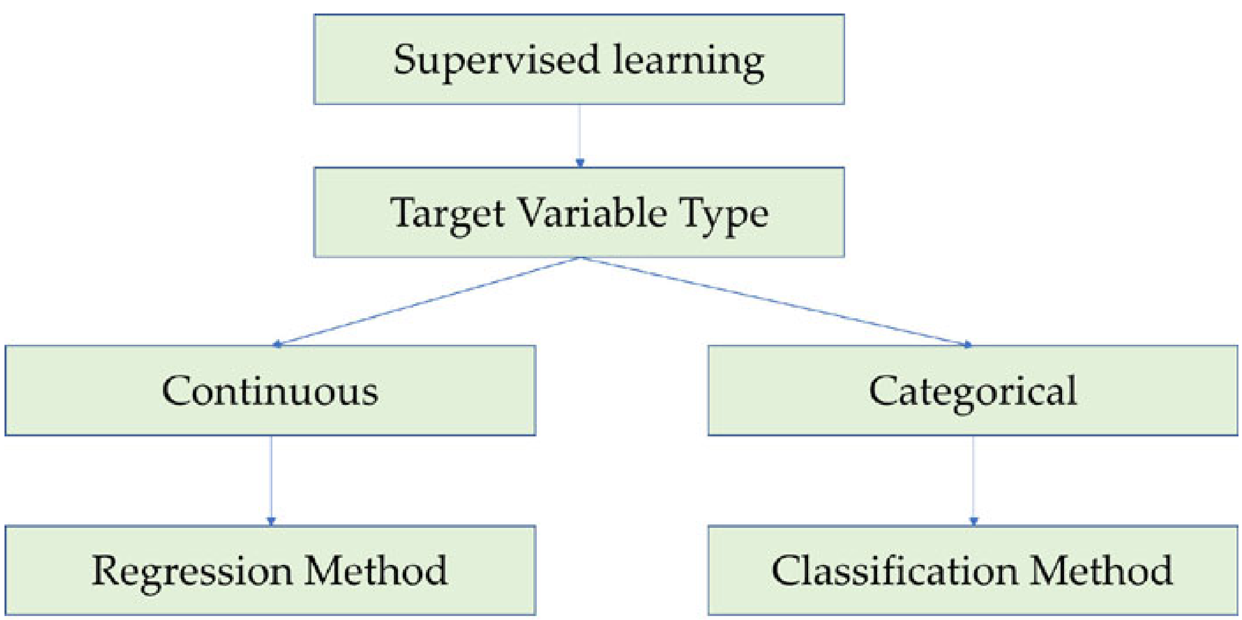


Figure 9.1: *Supervised learning target variable types and methods*

Classification problems use algorithms that are trained using labeled data, where the class of each target input data point is defined.

For example, a typical classification problem is classifying whether an incoming email is spam or not spam. The email is filtered based on the sender's email id, the words used in the email, and other related information. Another example is predicting the life span of a patient based on the disease's history, age, blood pressure, and sugar levels.

Algorithms for classification problems

The classification problems can be solved using the following most common algorithms:

- Binary classification
- Logistic regression
- Decision trees
- Support vector machines
- Multi-class classification:
- K-nearest neighbors
- Naïve Bayes

The choice of algorithm depends on the nature of the target input data set and the requirements of the problem.

The predictive classification model best maps input variables to discrete output variables. The main goal is to interpret the category using a fully trained model, and the new data will be labeled. The fully trained model is based on a training dataset that is sufficiently illustrative of the problem and has enough samples of each class label.

Label encoding

Class labels are usually strings that are mapped to numeric values before being fed to the algorithm. Label encoding is the process of assigning unique integer values to each class.

Classification terminologies used in machine learning

- **Classifier**: An algorithm used in mapping the input data to a particular category
- **Classification model**: The model used to predict class or category for the input data given during training.

Types of learners in classification

- **Lazy learners**: A machine learning algorithm, also known as instance-based learners, does not construct a model until a prediction is needed. They store the examples fed as training data and compare new inputs or testing data against the stored examples but do not learn a general function that maps inputs to outputs.

  The output for the new input testing data is set to the output of the most comparable stored example. The term "lazy" comes from the fact that these algorithms learn only when a prediction is needed.

  A popular instance-based learning algorithm is **k-Nearest Neighbors** (**k-NN**) algorithm.

- **Eager learners**: A machine learning algorithm, also known as model-based learners, build a model before making predictions. Based on the training data, they construct a classification model to map inputs to outputs, a function that can then be used to make predictions on new input testing data.

  Eager learners, using the training data, learn the underlying patterns and relationships between the inputs and outputs. These are encoded in a model, which is then used for making predictions on the new testing data.

  Linear regression, logistic regression, decision trees, and artificial neural networks are examples of Eager Learners.

There are popularly four main types of classification tasks that data scientists come across; they are:

- Binary classification
- Multi-class classification
- Multi-label classification
- Imbalanced classification

Binary classification

Binary classification refers to classification having two class labels. Generally, binary classifiers' class labels represent a normal state or true and an abnormal state or false. A typical example is students passing an examination. Students can fail (abnormal state) or pass (normal state). Another example is, Email is spam (abnormal state) or is not spam (normal state).

Class label 0 is assigned to the class for the normal state or true, and class label 1 is to the class with the abnormal state or false.

A binary classification task can be compared with a model that predicts a Bernoulli probability distribution. Where a Bernoulli distribution, a discrete probability distribution, describes the probability of reaching either a "success" or a "failure" state from a Bernoulli trial, an event that has only two possible outcomes (success or failure).

Algorithms specifically designed for binary classification and do not support more than two classes are:

- Logistic regression
- Support vector machines

Multi-class classification

Multi-class classification refers to the classification with more than two class labels. It is important to note that each sample is assigned to one and only one label or target in multi-class classification.

For example

- Human face classification
- Animal species classification
- Optical character classification
- Text translation models (a special type of multi-class classification)

The multi-class classification doesn't have the concept of abnormal/false or normal/true class labels. The target input data can be classified into any of the known classes.

For some problems, the number of class labels may be very large. For example, a model may recognize a species as belonging to one among thousands of species in a species recognition system.

Algorithms popularly used for multi-class classification include the following:

- k-Nearest neighbors
- Decision trees
- Naive Bayes

Logical regression and support vector machine algorithms designed for binary classification can also be remodeled for use in multi-class problems. This is achieved by using multiple binary classification models as in the following mentioned sub-categorization:

- One-vs-Rest

  wherein for each class versus all other classes, there is one binary classification model.
- One-vs-One:

  wherein for each pair of classes, there is one binary classification model.

Multi-label classification

Multi-label classification refers to the classification type having two or more labels, where a set of labels or targets are assigned to each input data.

For example, the classification of a scenic photo having multiple natural resources by a model may predict the presence of multiple known labels in the photo, such as “river”, “tree”, “shrub”, “road”, “mountain” and so on.

A multi-label classification task can be compared with a model that predicts multiple binary classification predictions.

Multi-label versions of the classification algorithms or Specialized versions of standard classification algorithms can be used for multi-label classification. These algorithms are listed as follows:

- Multi-label decision trees
- Multi-label random forests
- Multi-label gradient boosting

There can be another approach of using separate classification algorithms on the input data to predict or assign the labels for each class.

Note that multi-label classification is unlike binary classification and multi-class classification, where a single class label is assigned for each input data.

Imbalanced classification

Imbalanced classification, typically, is binary classification with the exception that the number of samples in each class is unequally distributed; with the majority of samples in training, the dataset belongs to the normal class, and a minority of samples belong to the abnormal class.

The imbalanced classification problems are outlined as binary classification tasks, although these may need specialized techniques, such that the constitution of samples in the training dataset is changed by under-sampling the majority class or over-sampling the minority class.

Where,

The **Majority class** is the class having a majority of the samples.

The **Minority class** is the class having a rare number of samples.

Under-sampling balances an imbalanced dataset by reducing the size of the majority class. That is, by keeping all samples in the minority class and selecting an equal number of samples from the majority class randomly, thus having a new dataset for further modeling and classification tasks. However, it may lead to the loss of potentially significant data as a downside. Oversampling is used when the data collected is sufficient.

Over-sampling is a technique used when the quantity of data collected is insufficient. The imbalanced dataset is made even by increasing the size of the samples of the minority class and also retaining the abundant samples of the majority class. The increase in the size of samples of the minority class is achieved by generating new synthetic samples by using **Synthetic Minority Over-Sampling Technique** (**SMOTE**). SMOTE is used to create synthetic samples by randomly selecting the characteristics from observations in the minority class.

There is no outright advantage of the preceding methods (of over- and under-sampling) over the two methods. These are applied depending on the use case and the dataset itself. These can be used in combination as well for success. *Figure 9.2* describes the two samplings as compared to the original data set:

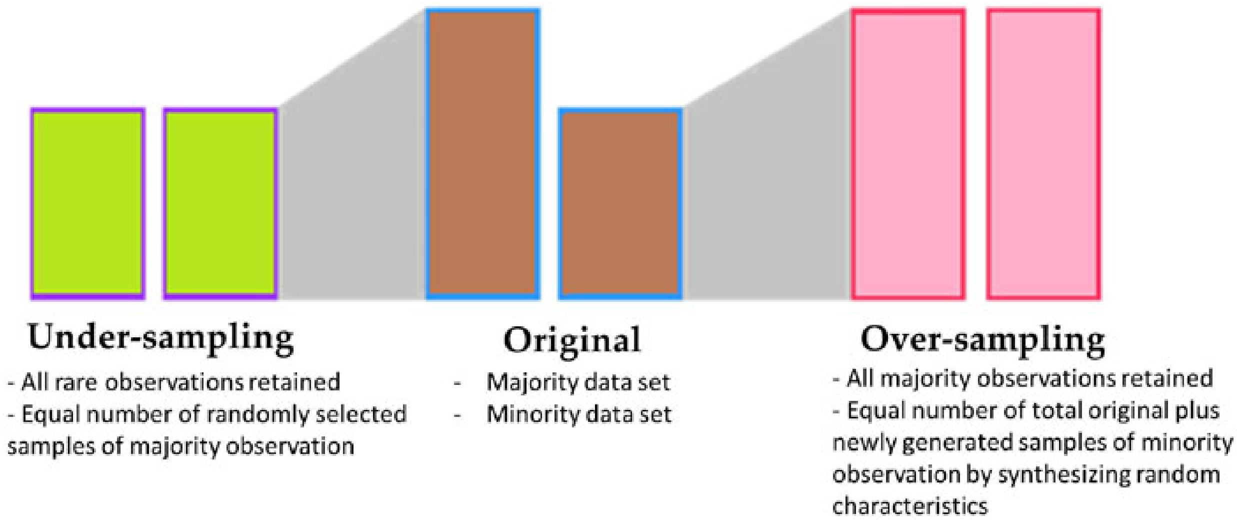


Figure 9.2: *Types of sampling*

The two algorithms used in imbalanced classification tasks are:

- Random undersampling
- SMOTE oversampling

Examples

Let us now understand simple binary classification with examples.

Simple binary classification

We come across a few binary classifiers in our day-to-day lives. *Table 9.1* describes a few such applications, where the 0 and 1 columns are two possible classes for each observation:

Application	Observation	0	1
Email Filtering	Email	Not Spam	Spam
Medical test reports	Visitor	Healthy	Diseased
Financial Data Analysis	Transaction	Not Fraud	Fraud
Sales	Item	Sold	Not sold

Face recognition	Person	Match	Doesn't match
Nutritional assessment	Person	Nourished	Malnourished

Table 9.1: *Binary classifiers examples*

Let us deep dive further into binary classification with a few more practical examples around us.

Sentiment analysis

The sentiment is a specific view, opinion, notion, or feeling. With social media giving each and everyone an opportunity to express their sentiments, various companies use sentiment analysis tools to gain insights into product and service reviews; political organizations use it for insights into opinions about any law, policy, or scheme rollout.

Sentiment analysis is a machine learning-based text analysis technique. It analyses entire text or words or phrases within the text and assigns 'positive' or 'negative' sentiments to them.

Neutral text is considered a lack of sentiment.

Sentiment analysis applications can browse through numerous pages in very less time and continuously monitor social media posts.

Figure 9.3 shows a tweet regarding feedback on one of the Tesla cars to *Elon Musk*, business magnate and investor and CEO of Tesla:

Figure 9.3: *Example of social media input*

The text clearly provides opinions about various aspects of the car, which are as described in *Table 9.2*:

Feature	Sentiment Analysis output
Looks	Negative
Engineering	Positive
Performance	Positive
Design	Negative

Table 9.2: *Sentiment analysis of Tesla car*

Sentiment analysis, using machine learning algorithms, can also be trained to get deeper insights into sarcasm texts and misused or misspelled words. With sufficient data inputs and proper training models, sentiment analysis tools can provide accurate results in a fraction of the time.

Real-world sentiment analysis examples

- **Banking sentiment case study**

  This is regarding the *Repustate* bank in South Africa. The bank was being hit by great customers owing to cut-throat competition. There came the dire need to know customers' perceptions of the bank.

  The bank then resorted to AI-powered sentiment analysis. They started a social media listening campaign that had people write their concerns about the bank using certain hashtags. The sentiment analysis tool collected 2 million texts over 3 months and used a sentiment visualization dashboard for ease of understanding.

  The analysis had the bank notice that the complaints were about not receiving services at particular branches during a certain time of the day, which was addressed to reduce the churn rate.

  Read the complete case study at **https://www.repustate.com/banking-sentiment-analysis-and-text-analytics/**

- **Customer behavior prediction**

  With the boom of e-commerce and online shopping, the need for analysis of customer buying pattern or behavior prediction have become a few of the important insights needed by businesses for revenue growth and product & service launches.

  The browser data of the user is analyzed using the browsing pattern analysis tool, post which the user is classified into certain demography.

This activity is beneficial to both the users and the online marketplace content developers. Users receive suggestions for content that is relevant to them, while content developers are able to do targeted marketing by learning about user preferences, thereby boosting both the navigability and profitability of the website.

In addition to browsing patterns, customers can also be classified based on their purchasing patterns. For example, binary classification models are used to determine if a customer is expected to 'purchase' more products or 'not purchase'.

Based on the likelihood of more purchases by a customer, promotional offers and discounts can be displayed to them.

The preceding scenario pertains to the customer that is likely to make more purchases. In case the customer is not likely to make any more purchases, his info can be saved for future promotions. *Figure 9.4* describes the workflow for the browsing analysis tool workflow:

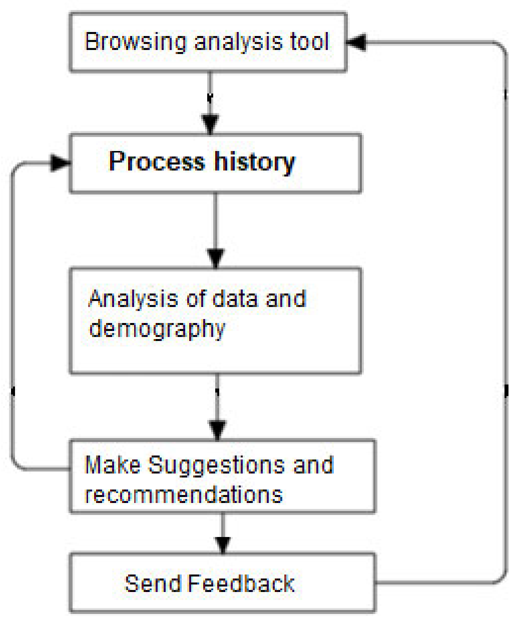


Figure 9.4: *Browsing analysis tool workflow*

- **Credit-worthiness assessment**

  A machine learning classification model can be trained to assess the credit-worthiness of a customer. This means the model can be used to predict the probability of a customer 'defaulting' payments or 'not defaulting' the payments.

This is achieved by analyzing the customers' past transaction data and historical information related to their payments.

Credit card companies, financial institutions like banks, credit unions, and so on use this assessment tool before approving any loans or money lending to customers. Typically, the credit-worthiness of a probable borrower is analyzed based on qualitative and quantitative looks at the 5Cs of Credit. These are capacity, capital, conditions, character, and collateral.

Figure 9.5 shows the credit score dashboard of a probable borrower or existing customer based on his paybacks history:

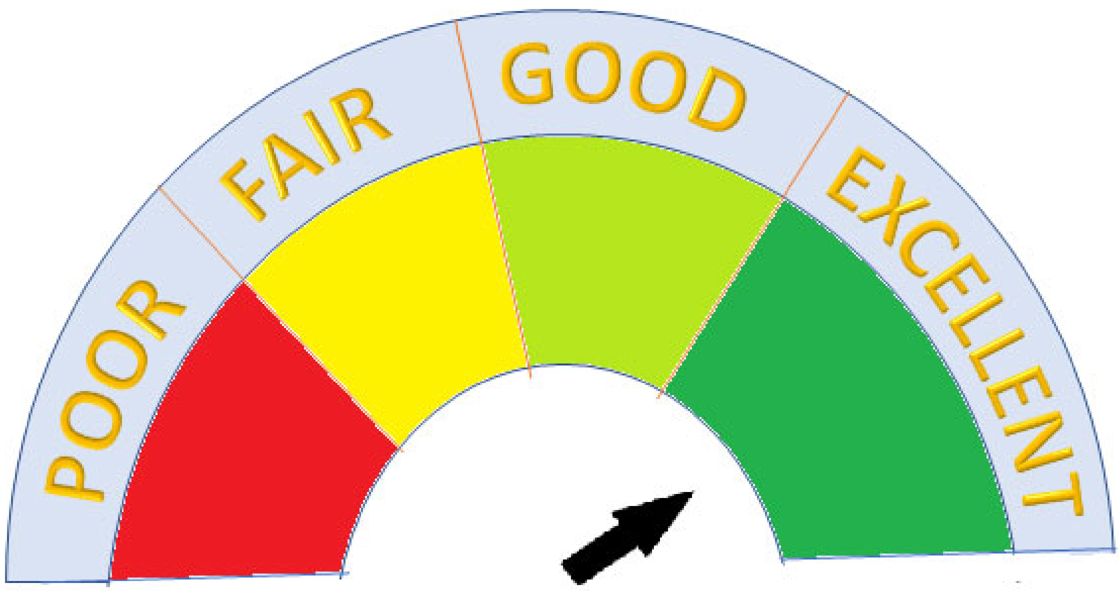


Figure 9.5: *Credit score dashboard*

- **Credit card fraud detection**

  Similar to creditworthiness assessment, a binary classification model can be used for credit card fraud detection. This model uses the historical transaction data of customers, which contains both 'fraudulent' and 'non-fraudulent' data, if any, related to the customers' credit cards.

  The model can then predict whether the transaction from the given customer's credit card will result in fraudulent or non-fraudulent transactions.

Introduction to binary classification with logistic regression

Logistic regression is one of the techniques used for binary classification tasks. It is used for predicting dependent categorical variables using a given set of independent variables that could be either numeric or categorical.

The outcome of logistics regression is categorical or discrete value. It can be either 'it is' or 'it is not'; exhibited as 0 or 1, Yes/No, True/False, Pass/Fail, Alive/Dead, Spam/No Spam, and so on.

Logistics regression is similar to linear regression. While linear regression is used in solving regression problems, Logistics regression is used in solving classification problems. Another difference is that instead of finding the best-suited regression line, logistic regression requires fitting an S-shaped logistic function that predicts two maximum values (0 or 1).

The 'S' curve from the logistic function states the likelihood of a specific outcome, such that logistic regression gives the probabilistic values between 0 and 1. *Figure 9.6* describes the S-curve of the logistic function and the threshold value (when y = 0.5):

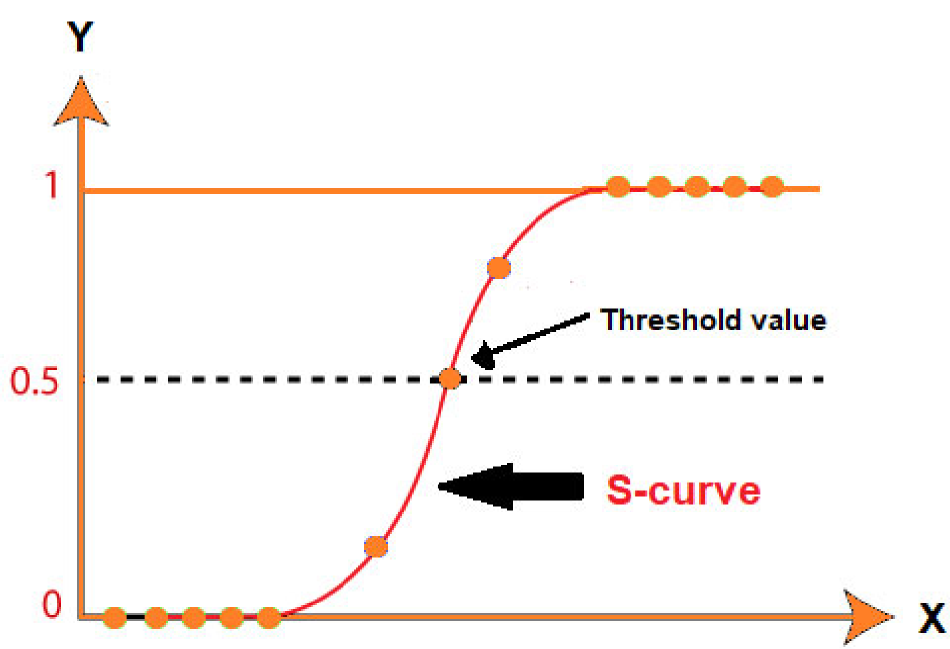


Figure 9.6: *S-curve of the logistic function*

Let us have a detailed look at the logistic function or sigmoid function.

- It is a mathematical function that maps the predicted values to values within a range of 0 and 1 depicting the probabilities.
- Since the value of the logistic regression cannot be outside of range 0 and 1, it forms a curve like the "S" form. The S-curve is called the Sigmoid function or the logistic function.
- In logistic regression, the threshold value (where y = 0.5 on S-curve) defines the probability, such as values above the threshold value are likely to have a probability of 1, and a value below the threshold value is likely to have a probability of 0.

Logistic regression equation

The logistic regression being similar to linear regression, the equation of logistic regression can be derived from the equation of linear regression. Following are the steps to get the logistic regression equation:

The equation of the straight line is:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

Applying limits of logistic regression, that is, y can be between 0 and 1 only, so,

$$\frac{y}{1-y}\text{; 0 for y = 0, and infinity for y = 0}$$

The final equation of logistic regression is obtained by the logarithm of the preceding equation that will get the range between -[infinity] to +[infinity]:

$$\log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

Type of logistic regression

Let us touch briefly on types of logistic regression. The three types of logistic regression are mentioned as follows:

Binomial: Having only two possible types of dependent variables, such as True or false, 0 or 1, Pass or Fail, and so on.

Multinomial: Having 3 or more possible types of dependent variables having no particular order, such as "elephants", "lions", or "sheep".

Ordinal: Having 3 or more possible types of dependent variables having a particular order, such as "average", "good", or "excellent".

True positives, true negatives, false positives, and false negatives

A binary classification technique may end up interpreting the input data correctly or incorrectly. So, we can classify the outcome of a binary classification technique into the following four (here, majority and minority classes are replaced by positive and negative classes, respectively):

True positives (TP): Number of outcomes where the observations are correctly classified as the positive class.

True negatives (TN): Number of outcomes where the observations are correctly classified as the negative class.

False positives (FP): Number of outcomes where the observations are incorrectly classified as the positive class.

False negatives (FN): Number of outcomes where the observations are incorrectly classified as the negative class.

These four terms (TP, TN, FP, and FN) are typically used to assess the performance of a binary classification model.

Let's take the example of the binary classification model to evaluate an email as 'spam' or 'not spam'. Then the terms TP, TN, FN, and FP would mean as described in the following:

TP: outcome as a true positive would mean that an email is correctly classified as 'spam'.

TN: outcome as a true negative would mean that an email is correctly classified as 'not spam'.

FP: outcome as a false positive would mean that an email is incorrectly classified as 'spam'.

FN: outcome as a false negative would mean that an email is incorrectly classified as 'not spam'.

Sensitivity score

Based on the values of the binary classifications' outcomes, we can compute the sensitivity score of the binary classifier as follows:

$$Sensitivity = \frac{Number\ of\ true\ positives}{(Number\ of\ true\ positives + Number\ of\ false\ negatives)}$$

$$sensitivity = \frac{TP}{TP + FN}$$

Accuracy score

Based on the values of the binary classifications' outcomes, we can compute the accuracy score of the binary classifier as follows:

$$accuracy = \frac{TP+TN}{TP+FP+TN+FN}$$

These parameters can be represented in a confusion matrix, as described in *Figure 9.7*:

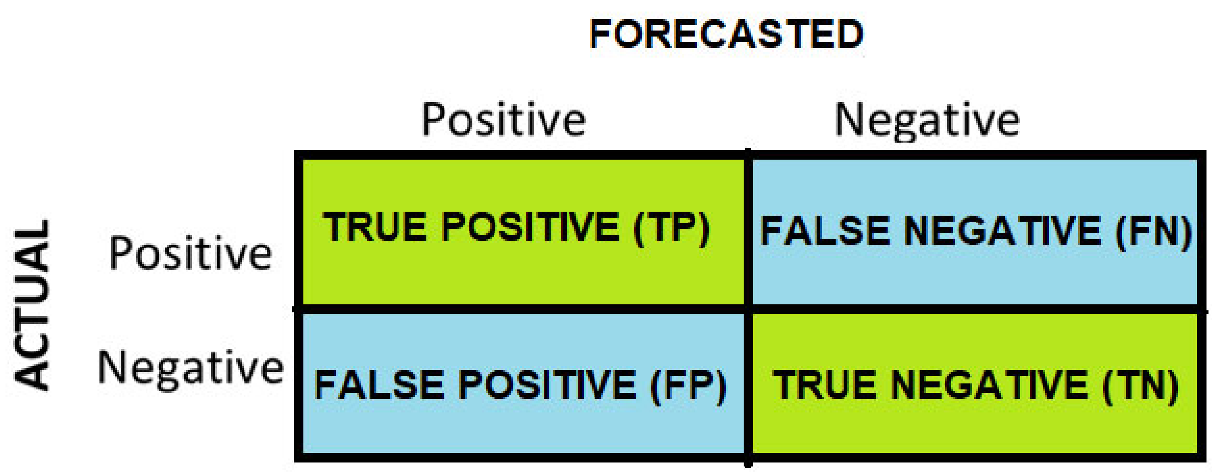


Figure 9.7: *Binary classification algorithm outcomes*

Example

True positives, true negatives, false positives, and false negatives can be explained by one of the fables of Aesop, "The boy who cried wolf".

Note: Aesop was a slave and a storyteller in ancient Greece who lived between 620 and 564 BCE.

The story is about a shepherd boy who found his job of tending the sheep very dull. To get some amusement, he raised a false alarm of wolf attack on the flock of sheep. The villagers ran to help save the flock, only to find it was a false alarm.

These false alarms were repeated several times before a day arrived when a real wolf actually attacked the flock. The shepherd boy raised the alarm, but this time, none of the villagers came forward to help to assume it was a prank yet again to fool them.

Let's define the following:

- "Wolf" is a positive class.
- "No wolf" (case of raising NO alarm) is a negative class.

Now, the "wolf-prediction" model can be summarized using a 2x2 confusion matrix depicting all four possible outcomes as in *Figure 9.8*:

True Positive (TP): . Incident: A wolf threatens attack. . Shepherd's alarm: "Wolf." . Outcome: Flock is saved.	**False Positive (FP):** . Incident: No wolf threatens to attack. . Shepherd's alarm: "Wolf." . Outcome: Villagers are fooled.
False Negative (FN): . Incident: A wolf threatens attack. . Shepherd's alarm: "No wolf" . Outcome: Flock is attacked and killed	**True Negative (TN):** . Incident: No wolf threatens to attack. . Shepherd's alarm: "No wolf" . Outcome: Life is usual.

Figure 9.8: *Shepherd who cried wolf*

A true positive refers to all outcomes where the machine learning model correctly forecasts the positive class.

Similarly, a true negative refers to all outcomes where the machine learning model correctly forecasts the negative class.

A false positive refers to all outcomes where the machine learning model incorrectly forecasts the positive class.

Similarly, a false negative refers to all outcomes where the machine learning model incorrectly forecasts the negative class.

Where we should care more

Let us consider situations where the following three scenarios, true negative, false positive, and false negative, can create panic situations.

Criminal courts

In criminal courts, what is more damaging? Convicting an innocent for a crime not committed by the person or letting go of a criminal as 'not guilty' or innocent.

Let 'guilty' is positive and 'non-guilty' as 'negative'. Here convicting an innocent is the case of 'false positive'. Where letting go of a convict as innocent is a 'false negative'.

It is generally considered preferable to make a 'false negative' over a 'false positive' judgment in criminal courts.

Metal detectors

In the case of metal detectors for security reasons deployed at airports, malls, and crucial public places, let the metal detected is 'positive' and no metal detected is 'negative'.

What is more damaging? A person who is threat expected to be carrying weapons goes undetected by the metal detectors, or a person carrying no weapon or metal object is highlighted as a suspect carrying metal objects by metal detectors.

The case of a person carrying metal objects going undetected is a 'false negative'. While a person carrying no metal object, if erroneously detected as carrying them, is a 'false positive'.

In such a case, a 'false positive' scenario is still okay than a 'false negative' that, in turn, poses a potential security risk.

Medical tests

Medical tests with certain criteria of fasting and likes will typically give incorrect results if the criteria are not followed correctly. This may happen in case the person undergoing the test is non-serious or is ignorant with respect to the criteria.

Blood tests to monitor diabetes require a certain level of fasting. For the test to give correct observations, the person must not have eaten for a certain number of hours before the blood sample is taken.

In case the person doesn't follow this rule, his sugar levels may come very high, raising an alarming kind of situation.

Let the sugar level high be 'positive', and the sugar levels normal be 'negative' case.

In case of tests erroneously reports normal levels for a diabetic patient making it a 'false negative' case, the patient will be taken off the medicines or will not be put on medication as the history may be. Due to an undetected disease, the person may also continue his sugar intake in various forms of food. These non-corrective actions will result in physical harm.

In case tests report high sugar levels for a normal person making it a 'false positive' case, the person will be termed as a patient and will be put on the unnecessary medication. This may cause mental harm to the person, including possible physical harm too.

Both 'false negative' and 'false positive' may prove to be harmful in case of medical conditions.

What is a clustering problem

Clustering is a type of algorithm used for unsupervised machine learning models. It involves organizing a set of observations into clusters, where each cluster is a group of similar observations, that is, having similar properties.

The observations within a given cluster are more similar to each other as compared to observations in other clusters. The clustering algorithm uncovers these natural groupings of the observations based on existing similarities to form clusters.

Input to a clustering algorithm is observations and the desired number of clusters, and out is the groups of observations into the specified number of clusters.

For example, consider a dataset of students that includes information about each student's class, marks, and rank. We may use clustering to group the students into different segments based on their similarities in characteristics. The resulting clusters could constitute different groups of students that could be targeted with various courses required.

Clustering can be further categorized into soft or hard clustering.

Following are some of the different methods of clustering but not limited to:

- Clustering based on partitioning
- Clustering based on a hierarchical model
- Clustering based on density
- Clustering on a grid

Clustering can use different algorithms depending on the attributes of the data and the desired properties of the clusters. The algorithms are listed as follows:

- Partitioning
  - K-Means
- Hierarchical clustering
  - Agglomerative hierarchical clustering
  - Divisive hierarchical clustering
- Density-based
  - DBSCAN
  - OPTICS
  - BIRCH
- Grid Based
  - STING
  - CLIQUE

The following section briefly explains the algorithms for clustering.

Partitioning based algorithm

Partitioning-based clustering algorithms split a set of data points into a set of clusters in such a manner that each data point belongs to only one cluster. These algorithms are iterative and move the data point between clusters till no further changes can be made. This is done to improve the partitioning.

Partitioning-based clustering algorithms are simple to understand and easy to implement while scaling well for large datasets. The required number of clusters needs to be specified prior to the processing of the input data set, which may not always be an appropriate number.

K-Means is the most popular partitioning-based algorithm.

K-Means

The K-means algorithm defines spherical clusters, where every cluster has a corresponding centroid (also called a mean point). The number of clusters (K) must be determined before executing the algorithm.

The algorithm repeatedly updates the centroids and the data points corresponding to each cluster. These iterations are performed till the centroids stop changing or a maximum number of iterations has been reached. The K-means clustering is depicted in *Figure 9.9*:

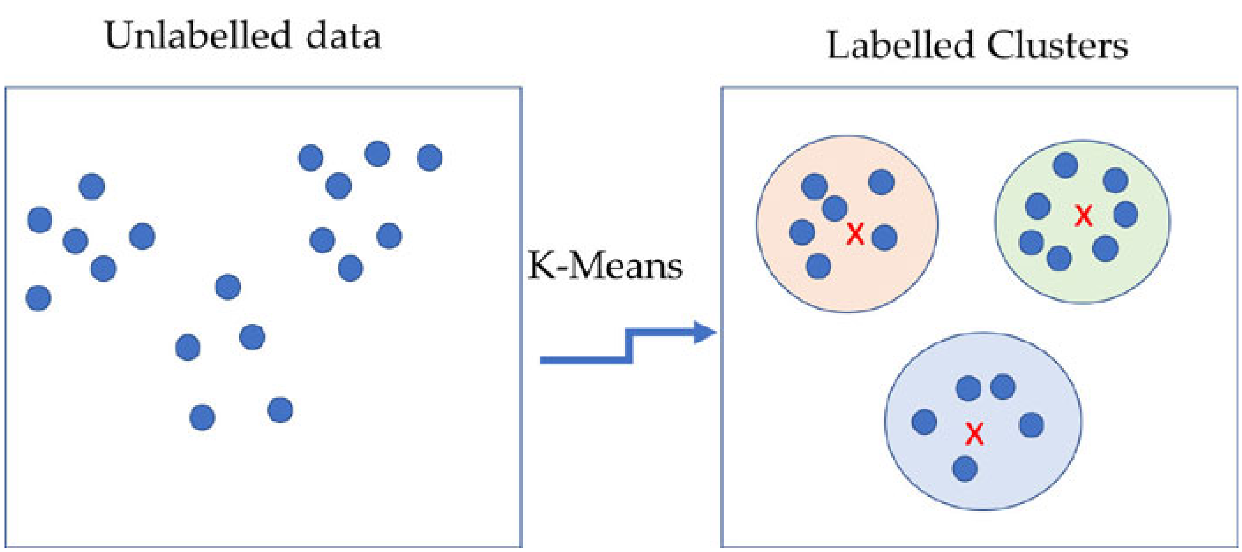


Figure 9.9: *K-Means clustering algorithm*

Hierarchical clustering

Hierarchical clustering (also called hierarchical cluster analysis or HCA) seeks to build a hierarchy of clusters.

Hierarchical clustering is represented by a diagram called a **dendrogram**. A dendrogram is a branching diagram that shows the hierarchical relationship between data points grouped in clusters. The dendrogram, as in *Figure 9.10*, shows the hierarchical clustering of six observations:

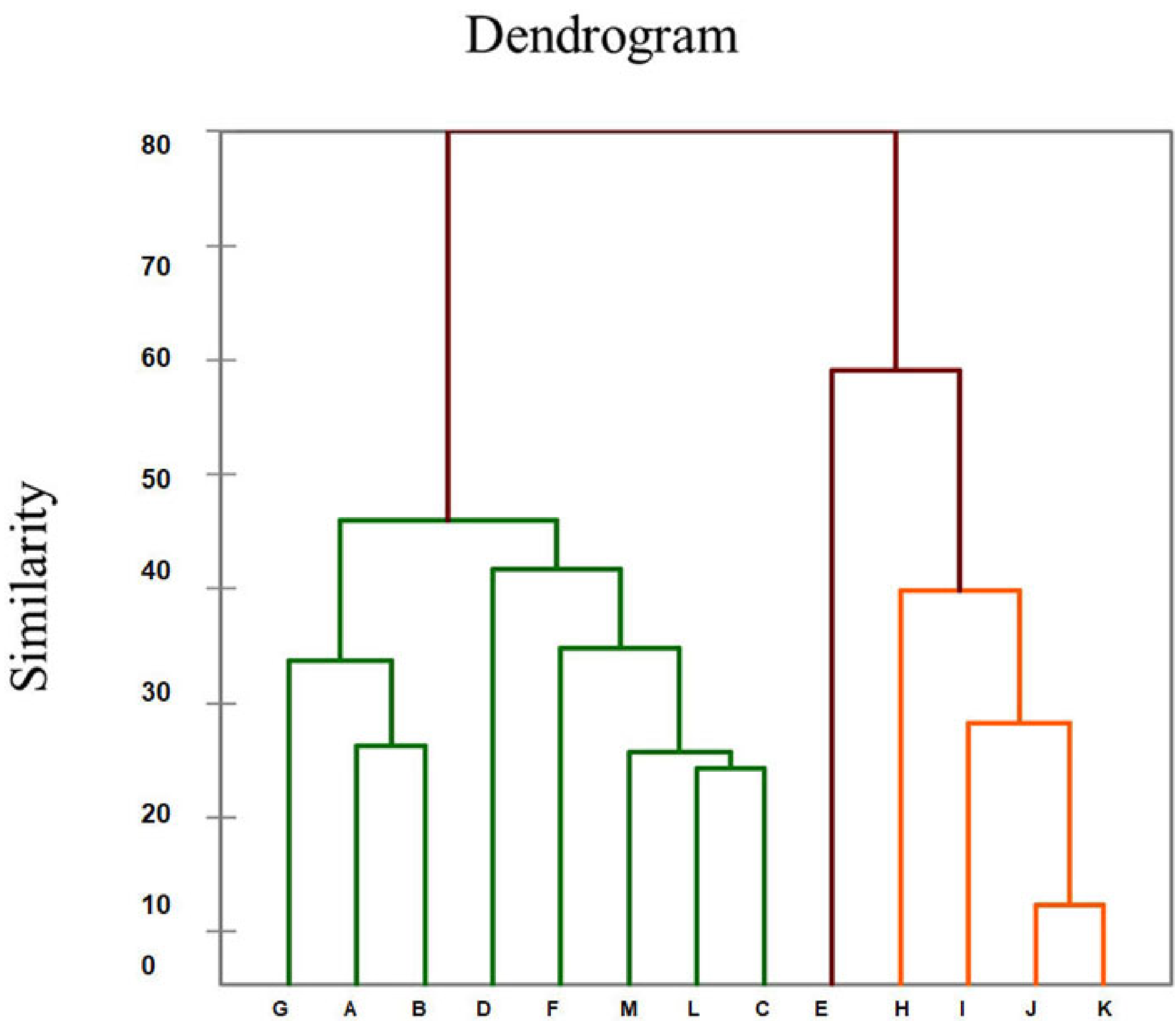


Figure 9.10: *Similarity between input data points*

Strategies for hierarchical clustering generally fall into two categories. These are:

- Agglomerative
- Divisive

The preceding two strategies for hierarchical clustering are depicted in *Figure 9.11* in the form of a conceptual dendrogram:

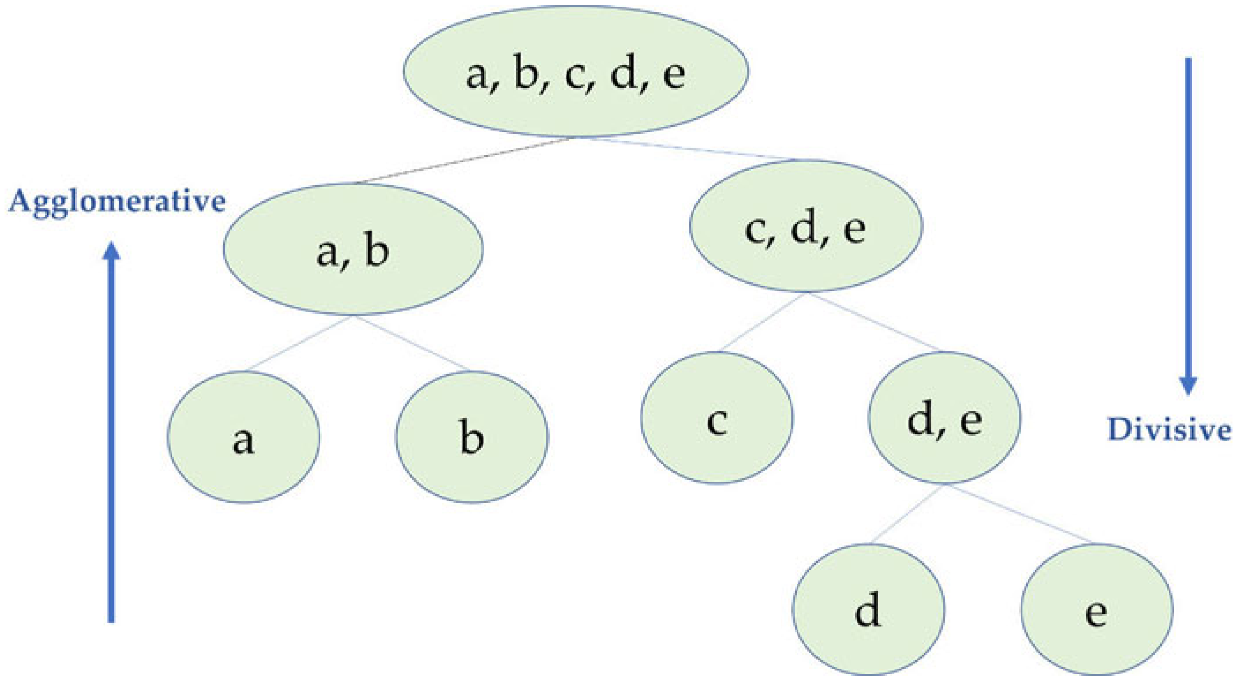


Figure 9.11: *Conceptual dendrogram for agglomerative and divisive Hierarchical based clustering*

- **Agglomerative hierarchical clustering (AHC)**

  Agglomerative: This is a "bottom-up" approach: Each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

  Agglomerative Hierarchical Clustering (**AHC**) is a type of hierarchical clustering algorithm that is based on a "bottom-up" approach. This approach has each observation starting in its own cluster, followed by pairs of clusters merging as one moves up the hierarchy until a single cluster is formed that contains all the data points.

  Typically, the merging process stops when a specific criterion is met, such as a maximum number of clusters.

- **Divisive hierarchical clustering (DHC)**

  Divisive hierarchical clustering involves having all observations in one cluster, followed by recurrent splitting of the cluster into smaller and smaller clusters, making it a “top-down” approach.

  Typically, the splitting process stops when a stopping criterion is met, such as a pre-specified number of clusters or a minimum cluster size.

This algorithm is useful when the number of clusters is not known in advance.

Density-based

The density-based methods are based on the fact that a cluster is a maximal set of connected dense units in a subspace (a selection of one or more dimensions). Two of the algorithms that are density based are:

- DBSCAN
- OPTICS
- BIRCH

DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm. It groups together data points that are close to each other based on distance measurement (that is, have a high density) and a minimum number of points. This separates them from data points that are in low-density regions while grouping together the data sets that are near a large number of other data sets within a given radius.

The two main parameters of the DBSCAN algorithm, thus, are the radius of the neighborhood and the minimum number of data points required to form a dense region

DBSCAN can discover clusters of arbitrary shapes and identify noise or outliers. *Figure 9.12* shows the DBSCAN clusters:

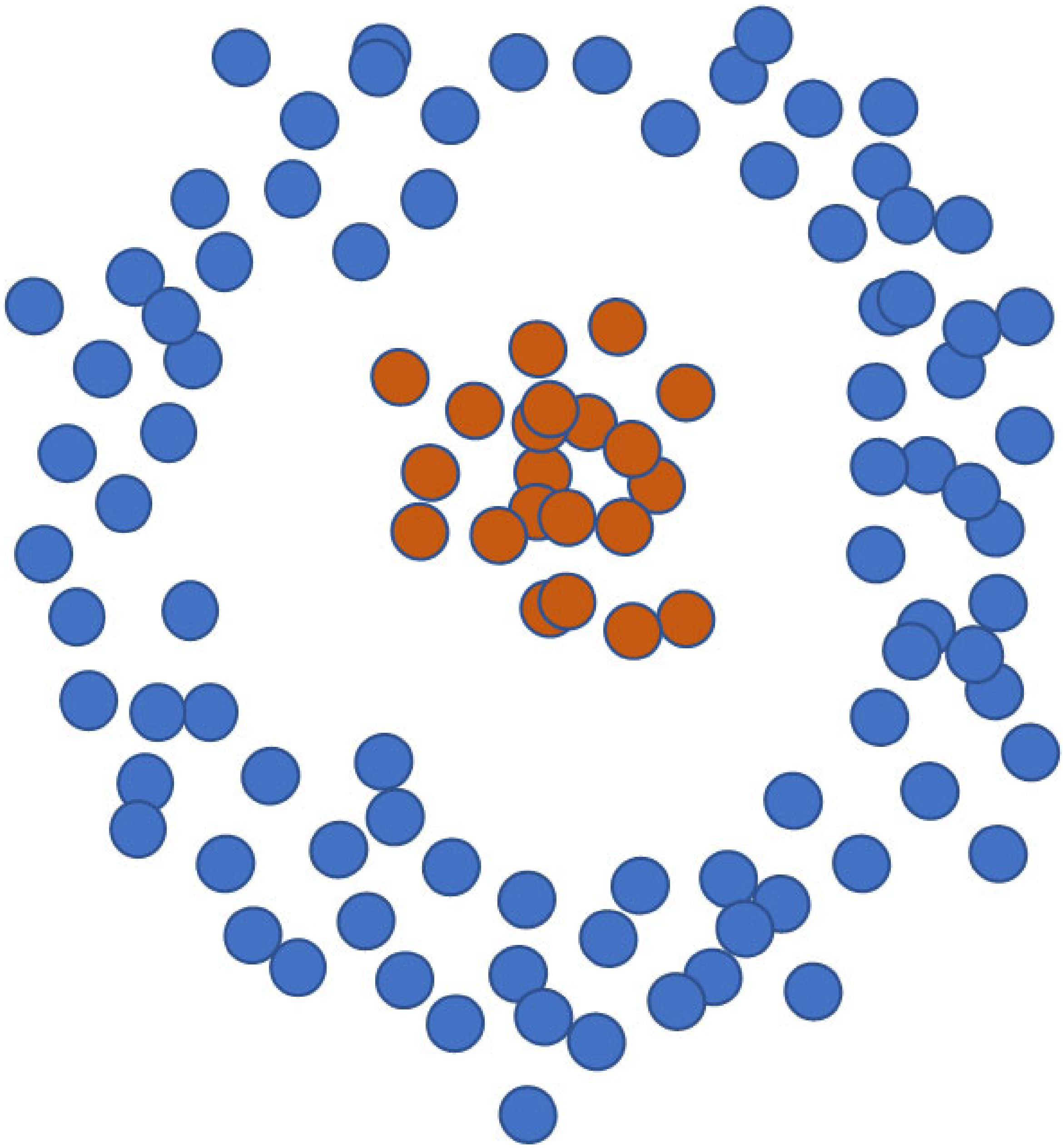


Figure 9.12: *DBSCAN clustering*

OPTICS

Ordering Points to identify the clustering Structure (**OPTICS**) is similar to DBSCAN but works like its extension. Instead of directly identifying clusters, the OPTICS algorithm generates an augmented ordering of the data points, which in turn allows for the detection of clusters of varying densities and shapes.

OPTICS has an advantage over DBSCAN due to its ability to recognize clusters of varying densities and the ability to handle data sets with non-uniform densities.

BIRCH

Balanced Iterative Reducing and Clustering using Hierarchies (**BIRCH**) is an algorithm that is designed to handle large datasets efficiently by using a two-level tree structure called a **Clustering Feature** (**CF**). CF is used to represent the data points, which allows faster operations on the data points, such as insertion, deletion, and retrieval.

The BIRCH algorithm has the ability to handle large datasets, identify clusters of varying densities, and handle data sets with non-uniform densities, giving it several advantages over other clustering algorithms.

Grid-based

In the Grid-based method, the space of instance is divided into a grid structure. Clustering techniques are then applied on base units which are cells of the grid rather than individual data points.

STING

STING is a grid-based clustering technique that uses a multidimensional grid data structure. It quantifies space into a finite number of cells. The dataset is repeatedly divided in a hierarchical manner, followed by each cell being divided into a different number of cells. Finally, the statistical measures of the cell are gathered and stored as statistical parameters. This is useful in query processing and other related data analysis tasks.

CLIQUE

CLIQUE algorithm is designed to handle datasets with a large number of dimensions. It uses density and grid-based techniques. It uses the Apriori approach, which states that if an X-dimensional unit is dense, then all its projections in X-1 dimensional space are also dense.

Why is it unsupervised

When the input is a set of variables with no corresponding output variables (that is, the data have no labels), the learning process is called unsupervised.

Clustering is considered unsupervised learning because no labels are given to the algorithm, leaving it on its own to discover the inherent structure and underlying patterns in the data and group the observations into clusters based on similarity in characteristics.

In other words, Clustering is an unsupervised method with input datasets such that neither there is any defined outcome (target) variable nor is anything known about the relationships between the observations, that is, unlabelled data.

Clustering algorithms do not assign any labels to data points but only group them into clusters based on similarities.

Clustering can be used as a pre-processing step for supervised learning. This may be achieved by using the clusters assigned as additional features in the input data. This can quite often improve the performance of the supervised learning model.

Examples

Let us consider real-world examples of clustering:

Biology

Biologists have spent decades deriving a hierarchical classification, a taxonomy, of all living beings. Living beings are classified into taxonomy as in *Figure 9.13*:

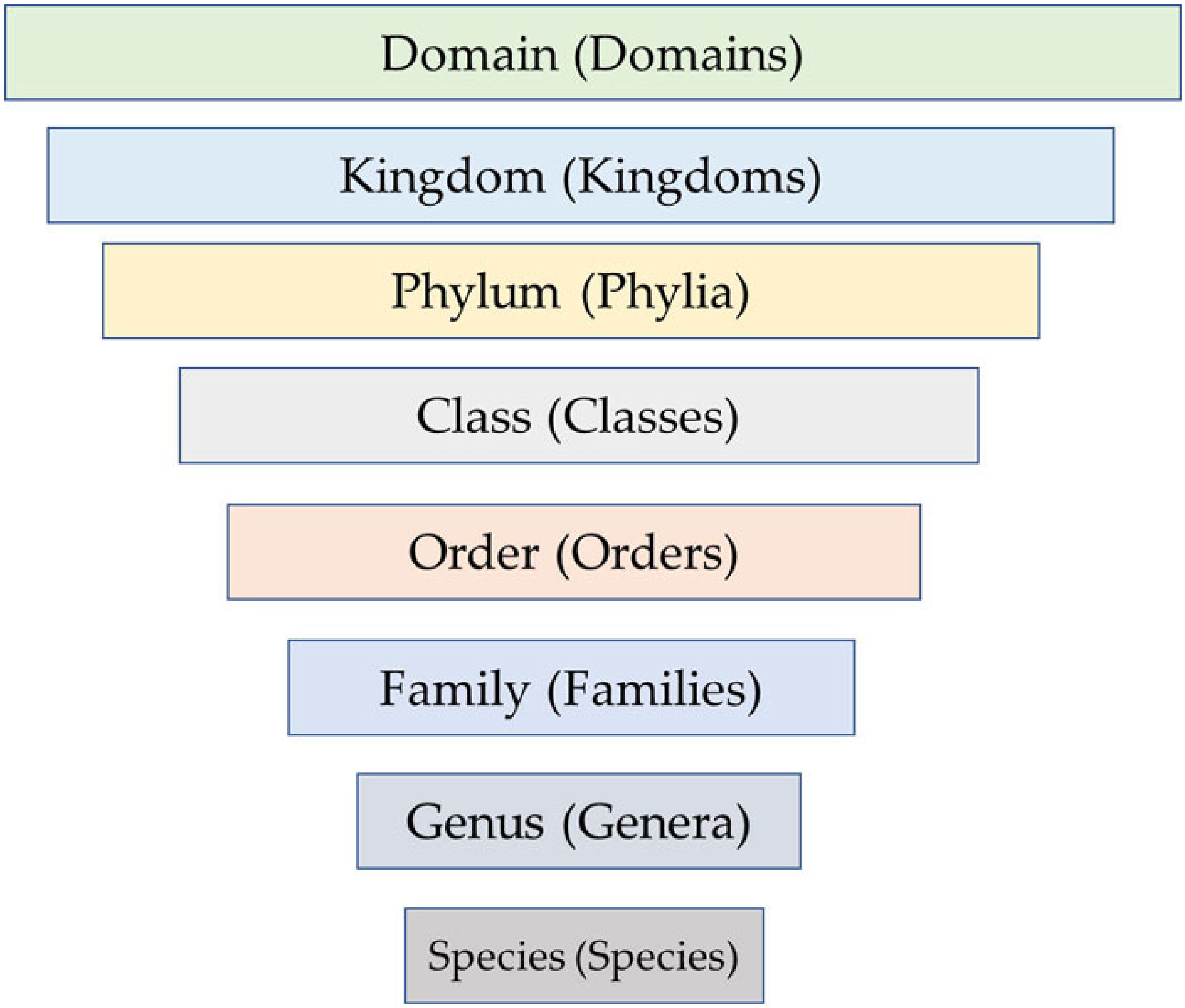


Figure 9.13: *Taxonomy of living beings*

Taxonomists organize organisms into a structural hierarchy—a hierarchical system in which each taxon (each group or hierarchy in biological classification) is nested or contained within a larger group.

In this multi-level classification, the higher the level of a group, the larger and generally it is, containing a wide variety of living things. In contrast, a smaller group contains organisms with more similar features.

For example, the group classified as all plants contains smaller groups, each containing similar types of plants, such as trees, shrubs, bushes, flowering plants, and so on. Each of these groups would further consist of still smaller groups; for example, the shrub group might be divided into deciduous or evergreen.

Modern taxonomists apply both physical and genetic evidence while classifying organisms into taxa (plural of taxon).

Let's consider an example of the 'coyote'. *Table 9.3* describes all related taxa for the 'coyote':

Texon	Classification
Domain	Eukarya
Kingdom	Animalia
Phylum	Chordata
Class	Mammalia
Order	Carnivora
Family	Canidae
Genus	Canis (that includes coyotes, wolves, dogs, and jackals)
Species	Canis latrans (coyotes)

Table 9.3: *A texon and its classification*

Hierarchical clustering in astronomy

French astronomer *Meterne* introduced hierarchical clustering in the field of astronomy in the late 1970s. This was initially used to figure out galaxy groups within galaxies before being widely adopted in all branches of astronomy.

Hierarchical clustering, in astronomy, works on the same principle of grouping similar objects together into clusters, such as stars or galaxies. The method iteratively merges the most similar clusters based on the Euclidean distance between each pair of clusters and merges the closest ones. The result is a hierarchical tree-like structure, a dendrogram, showing the relationships between clusters and identifying patterns in large astronomical data sets.

In other words, our universe is hierarchical. That is, stars converge into clusters and galaxies, galaxies are grouped together to form galaxy clusters, and galaxy clusters are further interconnected to form the large-scale structure of the universe.

Although the dimensional scales are varied, the formation and evolution of each are governed by gravity. These are very well-suitable for graphical presentation, having a similar hierarchical structure.

The Stellar system has three levels in its hierarchy:

- Stars and planetary systems
- Star clusters and associations
- Galaxies

Galaxies further form (as in the hierarchical model):

- Groups
- Clusters
- Superclusters (~100 groups form a supercluster)
- Hyperclusters (~100 superclusters form a hypercluster)

Let us look at the hierarchical placement (or address) of Earth. Earth is part of the Solar System, which is part of the Milky Way Galaxy, which is part of the **Laniakea Hypercluster**. Beyond this, the cosmic web (the building block of the universe – consisting primarily of dark matter and laced with gas – on which galaxies are built.) exists.

Maybe, in the next advancement in the field of astronomy, we may be able to explore the position of our Universe among other universes if they exist.

Classification and clustering comparison

Table 9.4 compares clustering and classification models:

	Classification	Clustering
Learning Type	Supervised learning	Unsupervised learning
Process	Classify input data and map to corresponding class labels	Grouping input data based on similarities without involving any class labels
Requirement	Required training and testing dataset	Doesn't require training and testing the dataset
Complexity	More complex	Less complex
Associated Algorithms	Logistic regression, Naive Bayes classifier, Support	k-means clustering algorithm, Fuzzy c-means clustering algorithm, Gaussian (EM) clustering

	vector machines, and so on.	algorithm, and so on.

***Table 9.4:** Comparison between clustering and classification models*

Conclusion

In this chapter, we learned about organizing data using classification and clustering algorithms depending on the type of data. We also learned about the impact of the application of incorrect algorithms on society.

In the next chapter, we will be discussing ethics in our society and how existing biases can impact AI functioning. We will also consider de-biasing or neutralizing the biased data.

Multiples choice questions

Let us now evaluate our knowledge of the binary classification model.

1. **In the case of an airline, there has been a bomb alert. The airline is diverted and is made to land immediately. Which situations are in the interest of human lives?**
   a. True negative
   b. False positive
   c. False negative
   d. True positive
2. **Which of the following is not an example of simple binary classification?**
   a. Email filtering (Spam/Not Spam)
   b. Financial transaction (Fraud/Not fraud)
   c. Temperature (hot/cold/normal)
   d. Patient sugar levels (Diabetic/Non-diabetic)
3. **Which of the following techniques is not used in Imbalanced classification?**
   a. Over-sampling
   b. Under-sampling

c. No-sampling
d. Logistic-sampling

Let us now evaluate our knowledge of the clustering model.

1. **The clustering model is based on which of the machine learning methodologies?**
    a. Unsupervised learning
    b. Supervised learning.
    c. Reinforcement learning
    d. All above
2. **Dendrogram is which of the following clustering methodologies?**
    a. Partitioning
    b. Hierarchical
    c. Density
    d. Grid
3. **K-Means is which of the following clustering methodologies?**
    a. Partitioning
    b. Hierarchical
    c. Density
    d. Grid

Answers

1. **c and d**
2. **c**
3. **c and d**

1. **a**
2. **b**
3. **a**

Questions

1. A binary classifier was evaluated using a set of 3,000 outcomes in which 50% of the outcomes were negative. It was found that the classifier has 40 % sensitivity and 60 % accuracy. Write the confusion matrix for this case.
2. What is the difference between binary classification and multi-label classification?
3. What are the true positive rate (TPR), false-positive rate (FPR), and false-negative rate (FNR)?
4. Classify students in your school based on one of the clustering algorithms, such as k-means or hierarchical clustering, whichever is best suited.
5. What are the strategies for hierarchical clustering?
6. Why is clustering based on an unsupervised learning methodology?

CHAPTER 10

AI Values (Bias Awareness)

Introduction

This chapter helps us understand the values and ethics and their importance in human society. Humans, in essence, are biased. Since AI is designed to imitate human cognitive skills, the input data to machine learning algorithms are generated by humans, which tends to be biased, disrupting the model's accuracy.

If biases in machine learning algorithms are not addressed and fixed, they may end up causing harm to businesses or structures, or human lives.

Thus, it is very important to focus on the ethics and biases prevalent in our societies and lives in general and how to eliminate or neutralize these biases in machine learning.

Structure

In this chapter, we will be discussing:

- AI working for good
- Principles for ethical AI
- Types of bias (personal /cultural /societal)
- How bias influences AI-based decisions
- How data-driven decisions can be debiased

AI working for good

Artificial intelligence, with its ability to process large data, helps perceive societal problems from different perspectives. Global societies have witnessed multiple revolutions like the industrial revolution, green revolution, economic integration, technological revolution, and so on. Each revolution has shaped humanity while posing an equally complex problem.

AI proves to be a valuable tool to augment human efforts to create solutions for vexing problems.

Let us look at ways AI is working for good:

Healthcare

Deploying AI-powered deep learning algorithms has not only improved medical diagnosis of disease and their treatments but also in automating tasks and analyzing large data sets of patients to deliver better preventive care faster and at marginal costs.

The imaging capabilities of AI are encouraging for cancer diagnosis and screening. The other key areas in healthcare where Artificial Intelligence is being applied are drug design processes and trials, discovering links between genetic codes, treatment and diagnosis of diabetic retinopathy, cancer, cardiovascular disease, and eye care.

For example, in the case of cardiovascular conditions, AI not only helps in diagnosis and treatment but also enables medical practitioners with the precise prediction of cardiovascular results, non-invasive diagnosis of coronary artery ailment, and prediction of outcomes for heart failure patients.

In the case of infectious diseases such as Covid-19, AI can predict and track the spread by analyzing data from the government, healthcare, contact tracing, and other sources.

AI-powered chatbots are the first level of medical support for patients or registered users, reducing pressure on healthcare workers.

Environment

AI-powered applications have proved to be powerful tools in monitoring and mitigating climate change. From designing energy-efficient buildings, monitoring deforestation, improving power storage, and optimizing renewable energy deployment by feeding solar and wind power into the electricity grid as needed, AI has proved to be a useful tool.

GEMS Air Pollution Monitoring platform is one of the real-world examples of the use of AI in climate monitoring. It is the world's largest global air quality information network. It partnered with IQAir, has monitoring stations in more than 140 countries, and aggregates data from over 25,000 air quality monitoring stations. Check their website: **https://www.iqair.com/unep**

With the large data set, it leverages deep learning and AI to predict the impact of real-time air quality on health and thereby recommend health protection measures.

The flow chart for their air pollution forecasting application Airvisual is copied as *Figure 10.1* from their webpage: **https://www.iqair.com/us/newsroom/understanding-airvisual-s-forecasting-method-deep-machine-learning**.

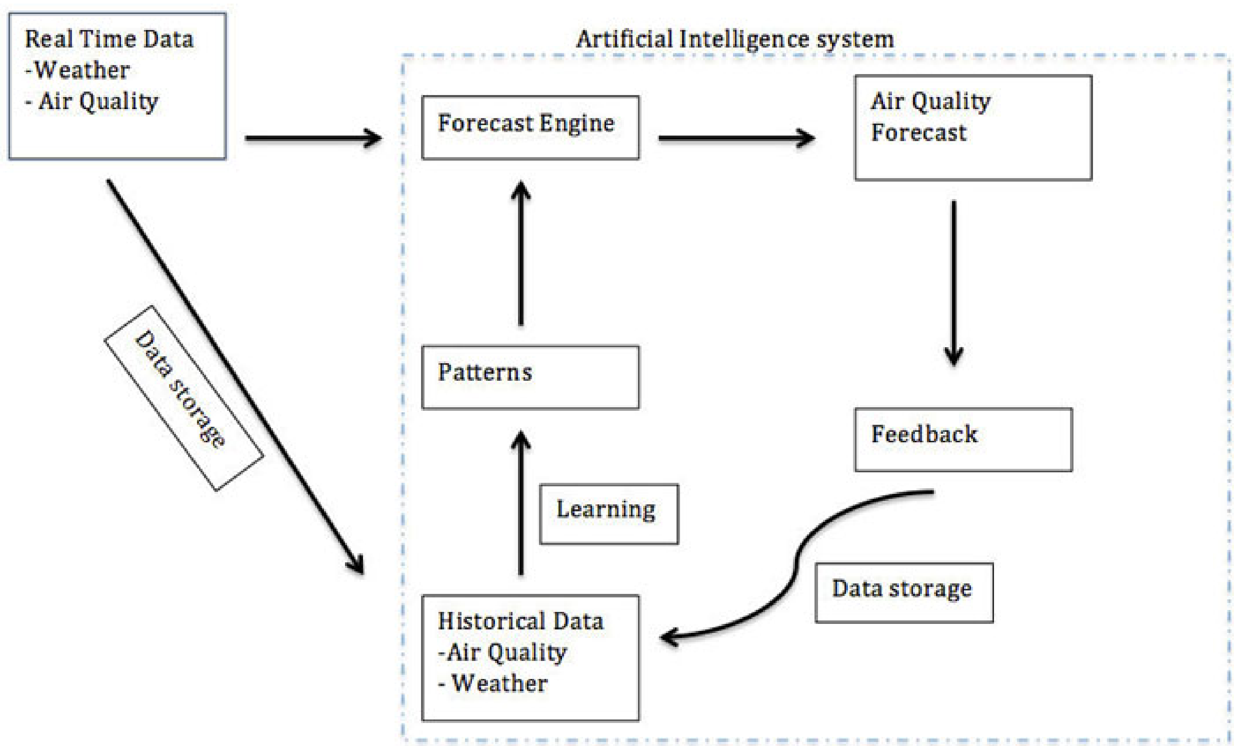


Figure 10.1: *AirVisual's Air quality forecast diagram as copied from IQAir website*

Education

AI, in its current state, is proving to be a colleague of teachers and professors by sharing their tedious workload and giving a very personalized learning experience to the students in addition to managing other administrative tasks. Current uses of AI in education include, but are not limited to, as follows:

Adaptive learning

- Enhanced student engagement by Gamification
- Learning management systems
- Skill development
- Learning languages

Administrative

- Staff scheduling and management
- Conveyance
- Accounting and finance
- Cybersecurity for online classes and tests
- Safety and security

Teacher's tools

- Students' behavior management
- Planning of lessons
- Parent-teacher interactions
- Preparation of tests
- Assessment of tests, both subjective and MCQs

A few real-world examples are:

- Content Technologies Inc. (CTI): **https://contenttechnologiesinc.com/**
- Brainly: **https://brainly.com/**

Public safety

AI can drastically help in predicting and preventing crime. AI plays a crucial role in managing crime scenes and detecting criminal behaviors and also leverages its capabilities of facial and image recognition.

Areas of public safety where AI levels the playing field are as follows:

Focussed deterrence

AI and related technologies are used to create predictive policing applications that use informational data to trace people's connections to gangs, sketch criminal histories, and also analyze social media posts for possible crime occurrence.

Investigating crimes

While AI and related technologies help predict and prevent any breach of public safety norms, they also play a significant role in successful convictions when a crime, accident, or disaster occurs.

Counter-terrorist threats

AI has proved to be a powerful ally in combatting terrorism. AI-powered tools can detect terrorist propaganda with accuracy in videos, and thus these can be prevented from going online. There are AI-powered tools that can also detect suspicious financial activities associated with financing terror activities.

Mitigating damage during natural disasters

AI has proved to be an effective way of identifying the occurrence of natural disasters and predicting their time, demography, and intensity. With this analysis, AI systems suggest actions be taken to mitigate the impact.

Crowd management

In 2019, India's *Kumbh Mela*- the largest religious gathering on earth- witnessed the use of artificial intelligence for better crowd management. With 1000 CCTV cameras over 3200 hectares of mela spread, artificial intelligence was used by the Integrated Command and Control Centre of the police, wherein the security personnel could assess the crowd size and see the visuals of crowd movement. This also enabled them to monitor anything that was suspicious.

Vehicles: Find optimal routes for an efficient and safe commute

AI-powered self-driving cars can navigate better routes, drive safely as compared to human counterparts, and also do real-traffic management in urban regions.

Agriculture: Maximizing food production and decreasing waste

The use of AI in the analysis of data from crops, soil, and weather can help in decision-making for optimizing water, fertilizers, and other resource usage. It can help in overall farm management, including the detection of pests and diseases. This can be achieved by analyzing images of the plants and monitoring livestock behavior. Even AI-powered robots and other automation save time by engaging in labor-intensive tasks.

Save the Bees

AI is being deployed in *The World Bee Project* to save bees from extinction due to the rapid decline in the global bee population. The decline has to be controlled for the good of our planet and the food supply. The World Bee Project has partnered with Oracle (providing cloud, database, analytics, and other solutions**). Internet of things** (**IoT**) based sensors are used to collect data from the hives, which is then uploaded to the cloud and analyzed to identify patterns and trends. This AI-based analysis and timely and advance interventions can help bees survive.

Accessibility for people with disabilities

There are various AI-powered applications that assist people with various disabilities. Examples include applications converting text into sign language for deaf students to learn. Application to identify visual disorders among children at an early age so that they can be rectified timely before causing blindness. Another application exists for assisting blind people in helping with the emotion on the person’s face they are talking by assessing the emotions and creating sounds accordingly.

Conservation of wildlife

Companies and universities are using AI to help conserve wildlife and protect endangered species. For example, a team analyzed 600 hours of sound to check the collision of birds with the power lines. Another application helps wildlife by detecting animal movement and poachers.

The uses are unlimited. Various aspects of AI, including image processing, unmanned aerial vehicles, and so on, are deployed for the cause.

Fight the World Hunger

AI-based applications can be used to analyze millions of data points to recommend perfect crops, maximize productivity, have the best quality seeds, and control other parameters. It can also forecast regions that have an increased risk of food shortages, possible draughts, and crop failure, along with any fluctuation in food prices that can impact negatively.

Principles for ethical AI

The more ubiquitous AI technology becomes, the more significant it becomes for its principled deployment and understanding of its ethical effects. When designing, developing, and deploying AI, organizations, businesses, and governments are expected to practice the highest ethical standards.

Figure 10.2 lists the principles for ethical AI:

Figure 10.2: *Principles of ethical AI*

Following is a brief on each of these principles:

Respect for freedom, autonomy and dignity of human beings

Human autonomy is not merely "functional autonomy," which is the capacity to operate independently without any external agents' control.

Human adults, based on their certain inherent cognitive and physical qualities, have the capacity to pursue and reasonably speculate on their values and goals. Having autonomy enables them to put to use these inbuilt capacities.

Self-determination in humans requires them to have adequate control over their instincts and impulses. Self-determination or self-rule is further subdivided into the autonomy of thought, of will, and of action.

There are certain prerequisites for humans to exercise their autonomy, examples of which are financial resources, access to complete information, and cultural resources.

Autonomy can be violated in many ways, say, when a person is manipulated, coerced, or deceived, resulting in being disrespectful to the person's autonomy. Others include limiting social and legal institutions, limiting access to economic and material resources, reducing the availability of cultural practices and resources, and more.

Let us see scenarios, not an exhaustive list, where AI can violate human autonomy:

- **Interference**: Physically preventing a human from doing any legitimate action. Such as a robot obstructing a person's path.
- **Coercion and threats**: Forcing or threatening a person to do certain tasks may be immoral or against the person's will or values.
- **Subterfuges**: Manipulating a person to achieve a certain outcome.
- **Nudging**: To prompt a person to take a certain decision or direction when the person has the predisposition to take up any.
- **Patronize and Paternalism**: AI deciding on behalf of humans for what is best for them
- **Heteronomy**: Humans are not using their own cognitive skills but delegating their own tasks to AI systems. Here tasks referred to here are those hindering autonomy.

Non-discrimination and fairness

AI-powered systems should be non-discriminating against humans, as individuals, as communities, or as groups. They must be accessible and fair to all.

Transparency and explainability

Transparency in AI systems allows human users to get access to the logic and decisions taken for the outcomes. Say, for example, a passenger in a self-driving car has the right to know why a certain detour or route was taken by the vehicle. This may reduce panic created, if any.

The AI systems, being explainable, are required to provide training data used, the kind of decision function, and the particular inputs for that decision.

Human oversight and accountability

Organizations or individuals who are involved in the life cycle of AI-powered systems, including but not limited to designing, developing, operating and deploying AI-powered systems, must be identifiable and take responsibility for such a system adhering to the regulations and policies while being accountable for the outcomes.

Human oversight of AI systems should be enabled, which may be achieved via governance mechanisms, such as the following approaches:

- **Human-in-the-loop** (**HITL**): refers to human intervention capability in every decision cycle of the AI system.
- **Human-on-the-loop** (**HOTL**): refers to human intervention capability during the design cycle, and monitoring of the operations of an AI system.
- **Human-in-command (HIC) approach**: refers to the human capability to supervise overall activities of the AI system such as its impact on society including economic, legal and ethical. Also, the ability to decide the use of the system in any specific situation.

Privacy, data protection and data security

AI systems are based on machine learning algorithms that learn from large data sets. Thus, it is imperative for such a system to maintain privacy rights and data protection ensuring security of the data.

Reliability and safety

AI systems, to be adopted for good in society, have to be trusted first. This means they should safely and reliably operate and have outcomes as per their intended purpose.

Inclusiveness and diversity

AI-powered ethical systems must not be biased and promote inclusiveness and diversity. They must help remove existing biases and imbalances by analyzing large data sets and giving appropriate suggestions.

Sustainability and environmental protection

AI-powered ethical systems must be designed with the aim to create a more sustainable world. These systems have the power to make the impossible possible. With the evolution of AI systems, difficult challenges related to sustaining and protecting the environment can be addressed especially by achieving a decrease in energy consumption, biodiversity protection, decrease in air and water pollution, environment friendly transportation network, optimizing manufacturing processes, improving harvest quality, forecast extreme weathers and more.

Human values

AI systems must be ethical and aligned with human-centred values by respecting their rights, diversity and autonomy.

Contestability

AI systems, when deviating from their intended purposes, must be allowed to be contested and challenged within appropriate timelines, thus minimizing any impact on humans and environment. Mechanism for continuous improvement and adaptation to changing scenarios must also be incorporated into ethical AI systems.

Wellbeing of humans, society, and environment

Overall, AI systems should be designed with an aim to improve human lives, their well-being, and bring benefit to individuals, societies, and the environment.

Types of bias (personal /cultural /societal)

Bias is a lack of objectivity, suggesting an illogical and unfair travesty, in judgment favoring or against a person or thing. Biases are usually discriminatory or prejudicial and are typically based on stereotypes, instead of logic, information, or experience.

Bias is uniquely a human trait. As such, biases are present within individuals, in cultures, and in society at large. At times, biases can be positive or helpful at large, however, in the majority of cases, biases are negative and damaging.

Personal

Personal biases are often based on triggers experienced in our routine life that contribute to creating these biases. Biases are also influenced and shaped by our beliefs, stories that stick with us, our values, educational background, family, friends, opinions, peers, and any other situation witnessed or learned about or have been part of. Biases often reinforce stereotypes and lead to incorrect judgments.

An individual can have two major types of biases. These can be explicit, which is conscious bias, or implicit, that is unconscious bias. Individuals can also experience cognitive biases arising due to systematic errors of thinking and misinterpretation of information.

Most common biases are based on following characteristics:

- Race
- Nationality or Ethnicity
- Gender and sexual orientation
- Religion
- Cult
- Socio-economic background
- Educational background

Being aware of your own biases can help towards personal well-being and professional success.

Social and cultural bias

A culture speaks for the beliefs and customs of a group, while society represents the people who share the same culture, that is, have same beliefs and follow same customs.

Bias and prejudices are apparently acts of social exclusion since they prevent the victims of such biases from making most of the wealth, knowledge, power and decision-making capacities of society at large. Social groups, households and individuals subject to such prejudice and discrimination become vulnerable and the impact can be long lived or engrained as a social norm.

Currently, women have low representation in high-skilled subjects like STEM. As per a global study, 35% of STEM students and only 29.3% of those working in scientific **research and development** (**R&D**) are women. This may be due to gender bias, if not at the entry level but certainly at job selection levels. There may be several biases that may have resulted in women occupying less than 50% of seats as students in this field.

Figure 10.3 describes the actual ideal scenario of outcome of technical interview of male and female candidates based on their technical skills:

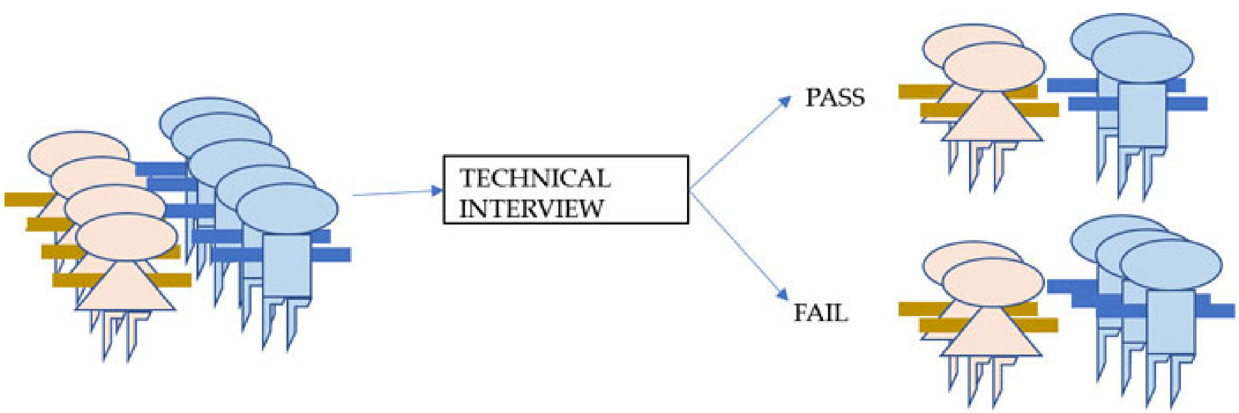


Figure 10.3: *Interview results with no biases*

Figure 10.4 describes gender bias while selection of technical employees by any organization:

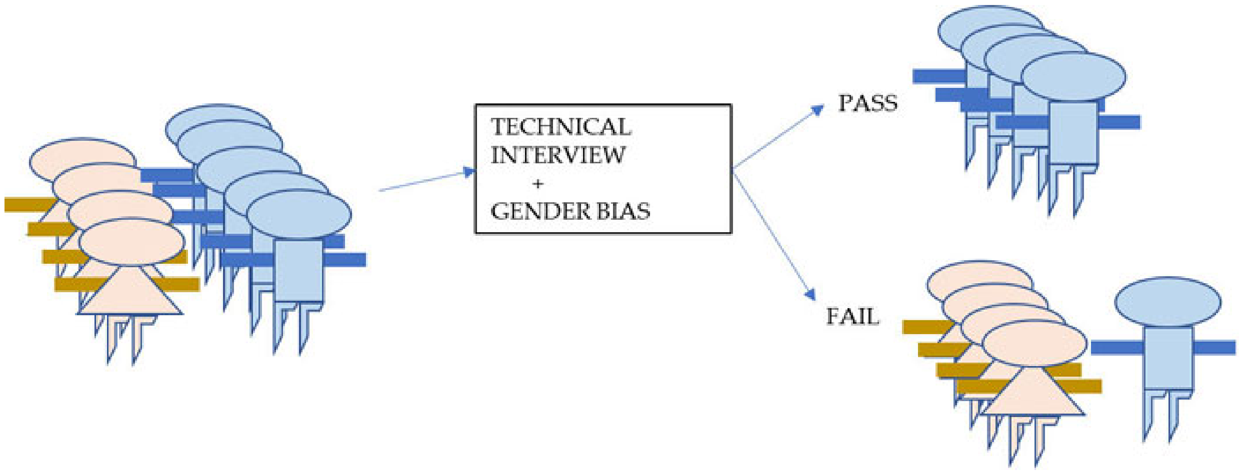


Figure 10.4: *Interview results with gender bias*

Some of the social and cultural biases are:

- Linguistic interpretation
- Ethical concepts of right and wrong
- Understanding of facts or evidence-based proof
- Intentional or unintentional ethnic or racial bias
- Religious beliefs or understanding
- Sexual attraction and mating

How bias influences AI-based decisions

AI has been evolving and being adopted at a fast pace. AI also has the potential to support humans in making unbiased decisions but only if AI is designed and modeled towards fairness in its results. We must also be aware, that if on one hand AI can help reduce biases, on the other, if modelled so, it also has the capacity to scale the biases and generate new ones.

Underlying data are often the source of bias that becomes part of AI-powered systems. There are three kinds of such data-based biases that can occur in AI-based solutions:

- Data Gathering Bias
- Data Analysis Bias
- Data Application Bias

[Figure 10.5](#) illustrates few of the biases structured under data gathering, analysis and application. These are just a handful of many biases that occur frequently, have a larger negative impact and can be prevented with effective counter measures.

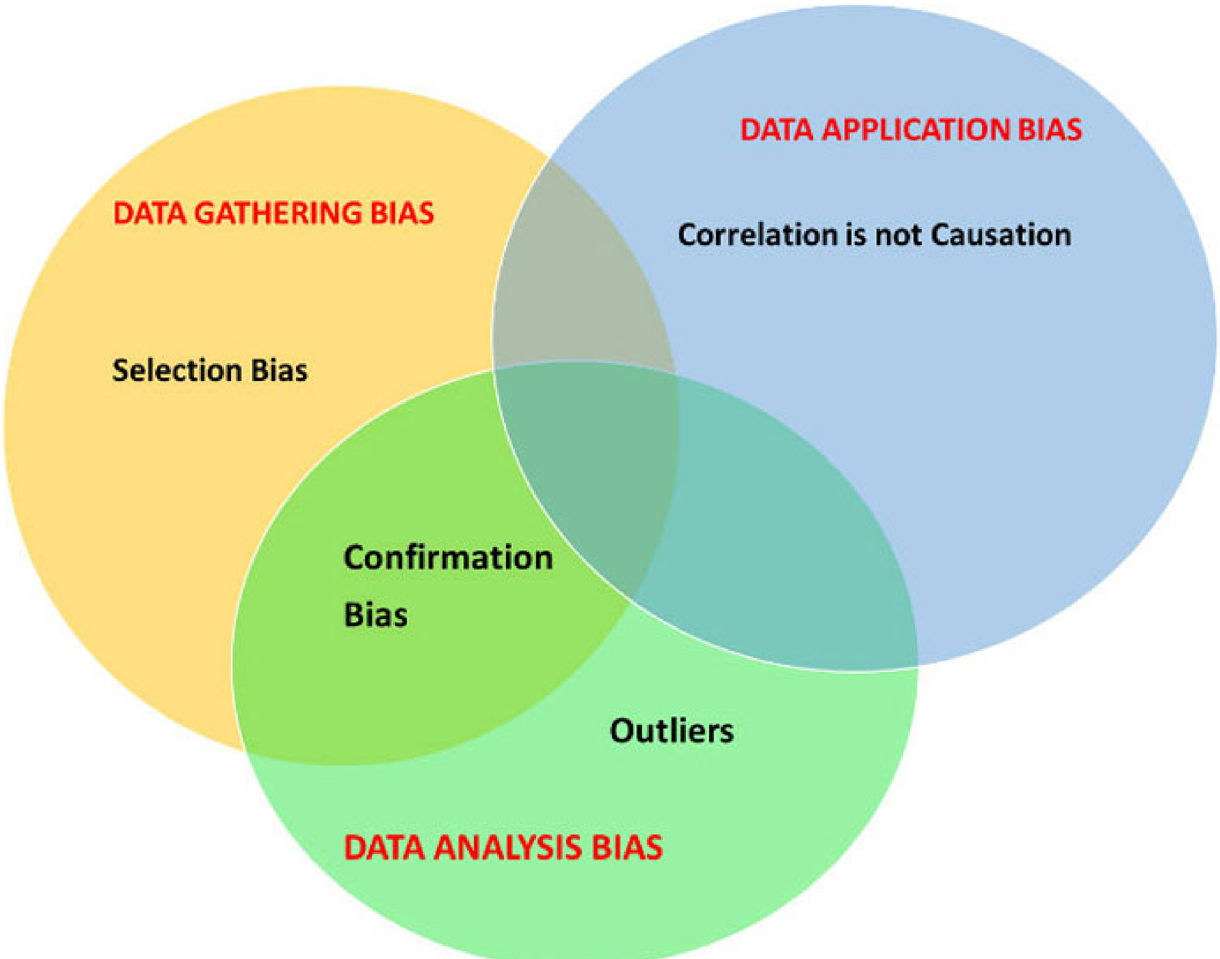


Figure 10.5: *Three Biases in AI-based solutions*

Data gathering bias

When the information gathered is not truly representative of the situation under investigation, there is bias in the data collected. This distortion can be prevented by carefully planning the sources used in the data collection process.

Selection bias

When the sample data collected doesn't represent the entire population of interest. This can happen, maybe unintentionally, especially when the data collected is the one that is conveniently available due to lack of budget or technical constraints, or maybe time pressure to gather enough sufficient data.

This, in turn, will not always give the intended results. The skewed data will result in gaps between expected results and reality.

It is, therefore, necessary to examine the sampling approach, and use randomization methods to select data and the criteria applied for inclusion and exclusion.

Confirmation bias

Confirmation bias occurs when the data sought confirms certain opinions but not all possibilities. An opinionated mindset or peer pressure, or a homogeneous team with specific social biases working on an AI project can contribute to such biases.

It is, therefore, necessary to seek data contradictory to the opinion and have a dedicated unbiased team for data gathering that have no constraints. Also, seeking a list of data or variables excluded from analysis help addresses remove such bias.

Data analysis bias

Outliers being neglected

When outliers are not acknowledged or simply eliminated, the expected results may differ from reality. This can happen when the extreme data sets go unchecked. Appropriate measures of central tendency can fix these.

Data application bias

Correlation is not causation: Causation Bias

Causation bias is when the misleading assumption is made that if two factors are correlated, one may be causing the other. This happens based on limited understanding and limited vision of the entire event. Enough data must be collected to differentiate between causation and correlation.

The following summarizes a few of the influences on decision-making due to bias in AI:

- Data collection, labeling, and analysis can result in biased algorithms that can perpetuate and step up existing social biases.
- AI systems can result in biased outcomes due to a lack of diversity in development and deployment.
- Over-reliance on AI decisions due to the presence of confirmation bias can reduce critical thinking.
- Biased data can produce prejudiced outcomes in areas such as employment and finance.
- AI decisions' interpretability and accountability can also be impacted due to biases.

Thus, it is pivotal to remove such biases in AI systems via regular audits, to promote diversity, and using transparent algorithms based on data representing the population of interest. Human discernment is still required to ensure fairness in AI-supported decision-making.

How data-driven decisions can be debiased

Data-driven decisions can be de-biased by taking appropriate actions, a few of which are listed as follows:

- Data sources' diversifying and elaborate representation can help reduce bias in data collection and labeling.
- Regular audits, re-training on new data sets for learning, and evaluations of AI systems need to be conducted to detect and address biases.
- One of the cores of AI systems, their algorithms must be designed with techniques that are fair and can counteract any bias in decision-making.
- AI development teams must be diverse and inclusive.
- Accountability in AI systems can ensure, to a large extent, that outcomes are close to reality.

- Transparency in algorithm logic and data sets can also help prevent a drift toward biased outcomes.
- Critical thinking and independent validation of AI decisions also ensure the prevention of the incorporation of biases.

Human judgment and oversight in AI decision-making processes incorporation is also the need of the hour in the current stage of the evolution of AI systems.

Conclusion

Biases come naturally in humans, and thus the data generated. This, in turn, makes the AI systems and their decision-making abilities biased too. The first step to address a bias is to accept its existence and then take preventive measures for removal and fairness in the decision-making. This, however, has to be a continuous process.

In the next chapter, we will be discussing a *Capstone Project.*

Multiple choice questions

List the social/personal biases in the following statements

1. **As per the August 2020 Fortune Global list, 13 women (2.6%), all of them white, were CEOs of Fortune Global 500 companies.**
   a. Ethnicity
   b. Gender
   c. Religion
   d. Education
2. **A 2020 analysis of over 1100 organizations by Mercer across the globe found the following data regarding women in leadership, that is, the representation of women decreasing as the levels progress:**
   - Executives: 23%
   - Senior managers: 29%
   - Managers: 37%
   - Professionals: 42%

- Support staff: 47%

    a. Social Background
    b. Gender
    c. Marital Status
    d. Education

3. **The neighborhood international school has all students having wealthy backgrounds and commute in luxury cars to the school. The neighborhood residents are mostly all white collared executives and high net-worth individuals**

    a. Socio-economic Background
    b. Gender
    c. Race
    d. Education

4. **The school teacher often uses phrases such as "Excellent work" to male students while telling female students "You can perform better" for similar homework and test results**

    a. Cultural
    b. Gender
    c. Race
    d. Educational

5. **Nationals from country A are extremely punctual and consider it of utmost importance. At the same time, nationals of country B ignore the delay from the planned schedule and do not prioritize it keeping it of the least importance. Once such nationals are supposed to attend the same meetings, there can be rude exchanges and lasting misunderstandings having a grave impact on businesses.**

    a. Cultural
    b. Nationality
    c. Organizational
    d. Educational

Answers

1. **a and b**
2. **a**
3. **a and d**
4. **b and d**
5. **a, b and c**

Questions

1. List down your own biases in terms of gender, relationships, teachers, brands, and others which you have.
2. Has your family, friends, or strangers displayed bias towards you or your actions? What did you do to correct it? Were your marks in the exams ever influenced by your own image?
3. List features of an application and possible framework of processes within it that will provide a solution to any of the biases you intend to remove from society.

Join our book's Discord space

Join the book's Discord Workspace for Latest updates, Offers, Tech happenings around the world, New Release and Sessions with the Authors:

https://discord.bpbonline.com

CHAPTER 11
Capstone Project

Introduction

As per the *Mckinsey report*, by 2030, Artificial Intelligence is expected to add $13 trillion to the world economy. AI systems are developed with the aim of achieving human intelligence levels and helping solve various problems with large and complex data. Thus, the challenge to understand a problem correctly in its entirety is such that an AI-powered system can predict and convert the problem into its understandable form and, therefore, be able to solve it.

Problem-solving refers to AI techniques such as forming efficient machine learning algorithms, applying best-suited statistical methodologies, performing root cause analysis, and executing them using modeling paradigms to solve the problem at hand. In many cases, there can be varied solutions to a specific problem, which can be achieved by applying different methodologies. While for some problems there are only unique solutions. As such, the nature of the given problem is the deciding factor for the algorithms, statistical methodologies, related visualizations, and data sets to be applied.

Structure

In this chapter, we will be discussing:

- Understanding the problem
- Decomposing the problem through the DT framework
- Analytic approach
- Data requirements
- Data collection
- Modeling approach
- How to validate the model quality
  - By test-train split
  - Introduce the concept of cross-validation

- Metrics of model quality by simple maths and examples from small datasets – scaled up to capstone project (Apply)
    - RMSE – Root Mean Squared Error
    - MSE – Mean Squared Error
    - MAPE – Mean Absolute Percent Error
- Introduction to commonly used algorithms and the science behind them
- Showcase through a compelling story

Understanding the problem

Artificial intelligence problems can be of three types:

- **Ignorable**: The solution steps can be ignored in such problem types. For example, solving a theorem. Say, a lemma that was to be proved as it was significant in later stages is found to be not that important, can still be helpful in solving the theorem but can be ignored, and another approach may be adopted to solve the theorem.
- **Recoverable**: The solution steps can be undone in such problem types. For example, solving a puzzle. In such cases, a wrong step can be undone to backtrack and then move forward with the right step to solve the problem.
- **Irrecoverable:** The solution steps cannot be undone in such problem types. For example, the game of chess. Once a move is made, both by you and the opponent, the steps cannot be undone.

Typically, all AI projects follow the following mentioned major steps:

1. Understanding the problem or problem scoping
2. Data acquisition or correct and reliable data gathering
3. Data exploration or data analysis, visualization, and data insights
4. Data Modeling: or creating models from data
5. Evaluation of project evaluation

Let us begin with best practices around the first step of problem scoping or "understanding the problem".

Machine learning algorithms are based on patterns and trends. As such, this becomes fundamental to embarking on the AI technology journey. It is, therefore, necessary for the problem to have an underlying pattern. The lack of any pattern disqualifies the problem of being a candidate to be solved using AI technology.

At a high level, machine learning is geared towards answering questions that are difficult to answer using human intellect. The following are basic questions within the umbrella of predictive analysis machine learning can answer for us:

- In case of sentiment analysis or email being spam, the answer we get is to query "**Which category?**" (*Classification*)
- In case of prediction of say, sales next month based on previous years, the answer we get is to query "**How much or how many?**" (*Regression*)
- In a hospital, 50% of all patients are diabetic, 40% have heart diseases while remaining 10% have other diseases. When a new patient arrives or seeking info on an existing patient, here the query is "**Which group?**" (*Clustering*)
- A bank call to report a fraudulent transaction on a credit card is a panic call. Here we get the answer to the query "**Is this unusual and weird?**" (*Anomaly Detection*)
- The basis on the purchasing pattern of the customer, the portal recommends certain products for further purchase. In this scenario, we get the answer to the query "**Which option to take?**" (*Recommendation*)

Identifying the right questions to ask is the first key step in problem-solving. This further helps in determining which model to use.

Project 11.1

Let us consider the following sectors to identify a problem to be solved using AI technology, followed by related questions and answers to express our understanding of the problem.

- Green vehicles
- Higher education
- E-commerce

Decomposing the problem through the DT framework

We have studied the Design Thinking framework in detail in *Chapter 6, Creative and Critical thinking*.

Let us revisit the framework that provides a systematic approach to finding solutions to the problems. Following are the five stages of design thinking: Empathize, Define, Ideate, Prototype, and Test. Design thinking is a non-linear process, as described in *Figure 11.1*:

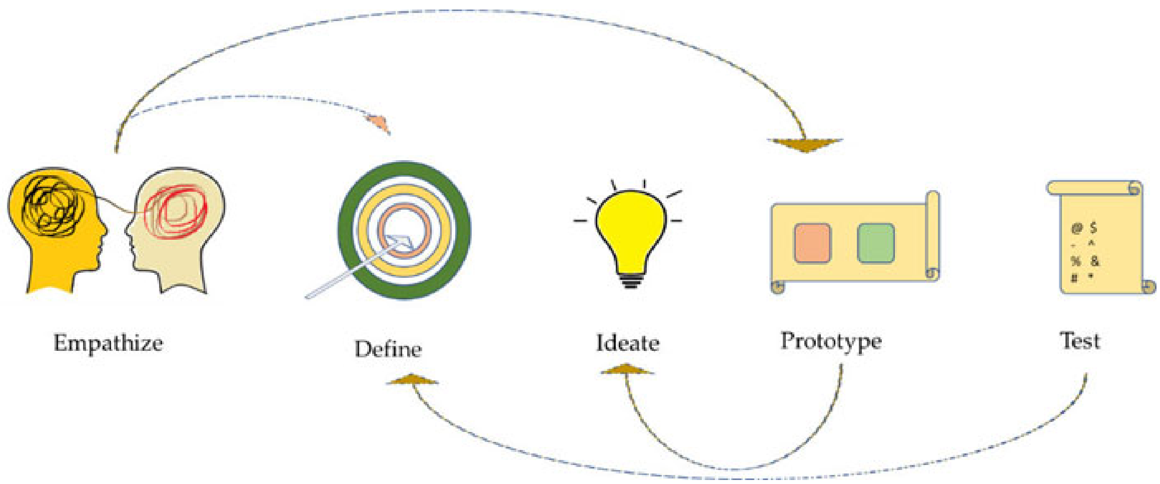


Figure 11.1: *Design thinking is a non-linear process*

Right questioning using a problem-solving method, "5W1H", a term for *Who*, *What*, *Where*, *When*, *Why,* and *How*, helps get clarity and make better decisions before taking any action. 5W1H is described in *Figure 11.2*:

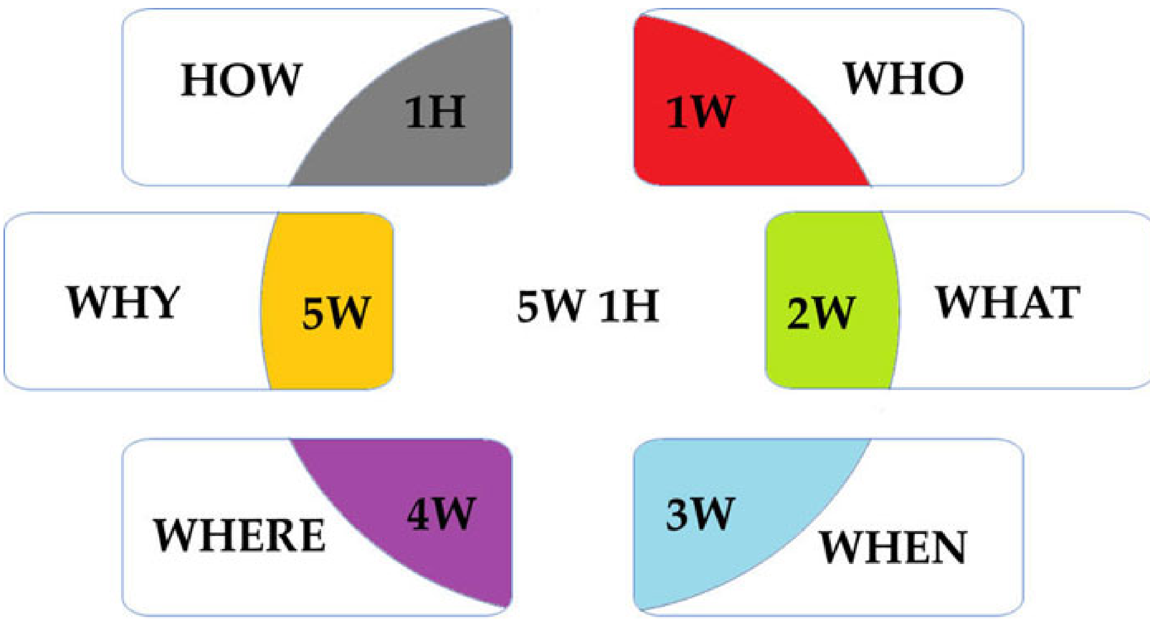


Figure 11.2: *5W1H, a problem-solving method*

The elements of 5W1H are a set of five questions as mentioned in the following:

- What is the problem?
- Where is the place for the problem?
- When is the time of the problem?
- Why is the problem?

- Who is affected by the problem?
- How can we solve the problem?

These are visualized in *Figure 11.3*:

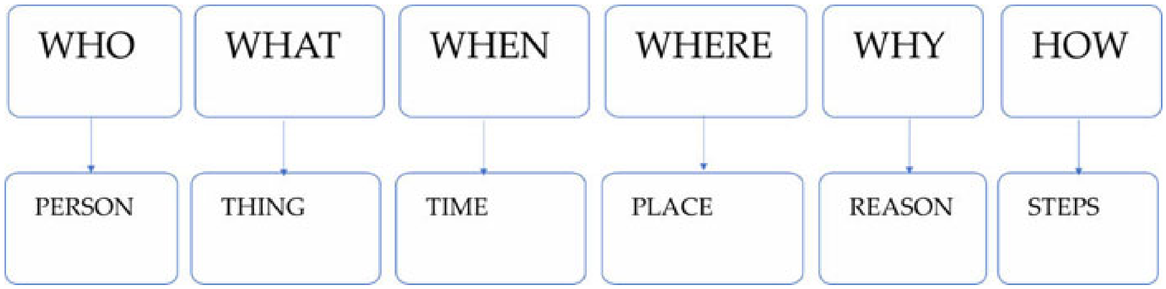


Figure 11.3: *5W1H questions and components*

Real-world problems have computational algorithms and processes that are complicated. Hence it is advisable to break down the problem into smaller units and code each unit separately.

The following are good practices during problem decomposing phase:

- **Avoid jumping to solutions:**

  Do not get distracted by thinking in terms of features and functionality but focus on understanding the fundamental problem and breaking it down accordingly.

- **Reframe problem statement:**

  Rephrase the problem statement in your own words leaving room for imagination, experimentation, and change. Clearly pen down the inputs and outputs and clarify each detail, leaving no room for ambiguity.

- **Make it manageable:**

  Apply the divide and conquer rule by breaking the problem down into logical pieces and further decomposing the complicated pieces into smaller tasks. Repeat until the tasks cannot be further divided. *Figure 11.5* describes the decomposition of a problem into smaller problem units:

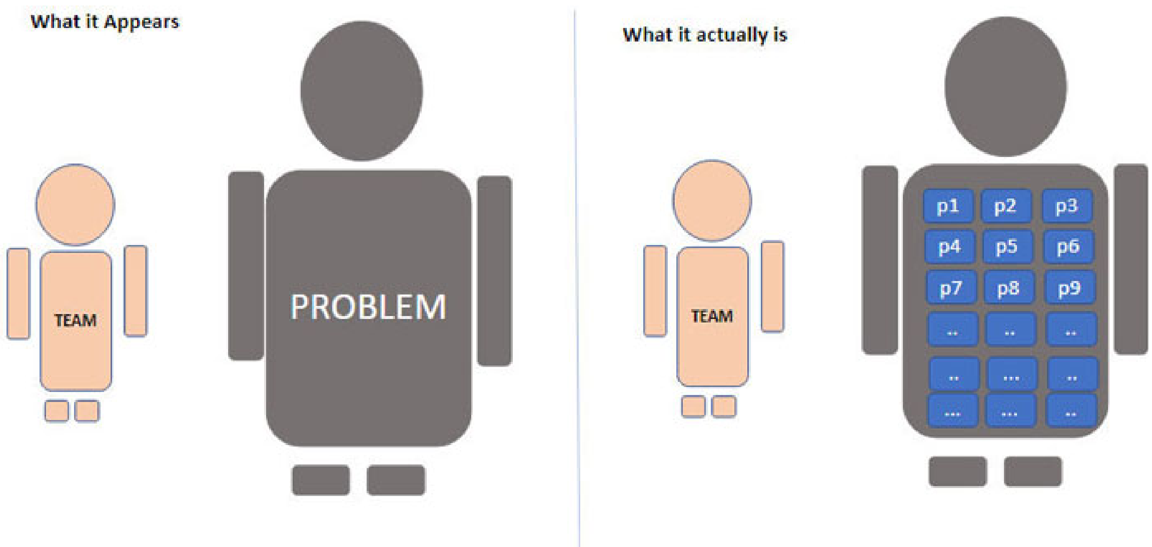


Figure 11.4: *A problem with how it actually is*

- Code each piece by following the design, development, and testing phases. Fix bugs/discrepancies encountered.

Project 11.2

Problem statement: India is an agriculture-intensive country, and it becomes very important for the economy to have good quality and large quantity agricultural produce. This also is important as for some crops, and the produce is once a year and the only time when farmers can earn for the entire year. Government departments ensure that the temperature and rains are predicted well in advance so that measures can be taken to save the crops.

Create a model based on weather conditions and past year data to predict the rains.

Analytic approach

Once the problem has been identified, comes the task of analytical approach that needs to be applied based on the model input and output data.

For example, in case of leasing apartments in the real estate sector, the house price is determined by the number of rooms, bathrooms, locality, the floor the apartment is located on, the size of the complex, number of balconies, access to basic infrastructure, and so on. Once the data has been collected, a regression technique can be applied to predict prices for any new apartments. *Figure 11.5* describes possible weather predictions based on past conditions:

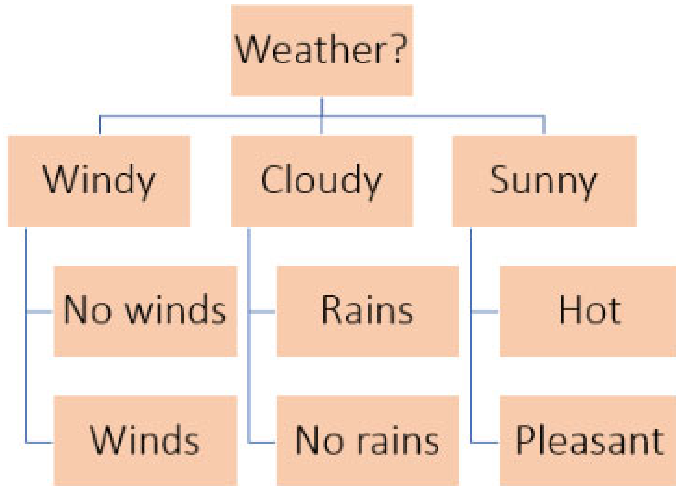


Figure 11.5: *Weather predictions*

If the problem needs classification of the entity, say biometric identity, speech recognition, predicting a customer will buy a product or not based on past purchasing trends, an email to be marked spam or not, a credit card transaction is to be marked fraud or normal, drug classification and more. Classification technique is to be applied in such discrete (0/1) scenarios. Spotify uses classification algorithms to recommend music its users may like, whether they have heard the songs before or not.

There can be problem statements which would require image processing methodologies. Say for example, use a car or vehicle camera to detect eyes of the driver and detect for drowsiness. This involves video analytics too. Such analytics can be used to predict if the driver is in a state to drive or raise an alert to take corrective action.

If it is textual data, say from social media posts, one can use a sentiment analysis approach to understand the underlying emotions, feedback or positive or negative responses. It can also be effectively used by product-based companies, to get feedback on their product and services and take active corrective measures, if need be.

For practical purposes, the project to be undertaken in the class must have reasonable volumes of data to be analyzed which is not in GBs.

Project 11.3

Create a model to predict if possible, ragging is in progress in a hostel.

Hint: use heat maps, image classification and video analytics or any other technique.

Data requirements

The data is imperative to any machine learning model. It is also most significant limiting factor that decides the performance of an ML model or how much it can grow. In other words, quality of data used to train, validate and test the model overshadows the efforts put towards model creation and its refinement when it comes to performance of the model.

Hence identification of the necessary data content, formats, and sources is necessary to expand the limits of the machine model. Data requirements are the stage for deciding the data for initial data collection to be fed to the algorithm we choose.

Data formats are important, say for example, for sentiment analysis data source is in text format, while the algorithm accepts the data in numeric format. Data sources should be reliable, authentic and ethical.

Project 11.4

Identify the data requirements for soil nutrition required for certain levels of wheat produce based on land size and type of wheat and other such parameters.

Data collection

The Data collection stage involves identifying the available data resources relevant to the problem domain. *Figure 11.6* recaps various data sources that can be used for data collection for the problem specific domain:

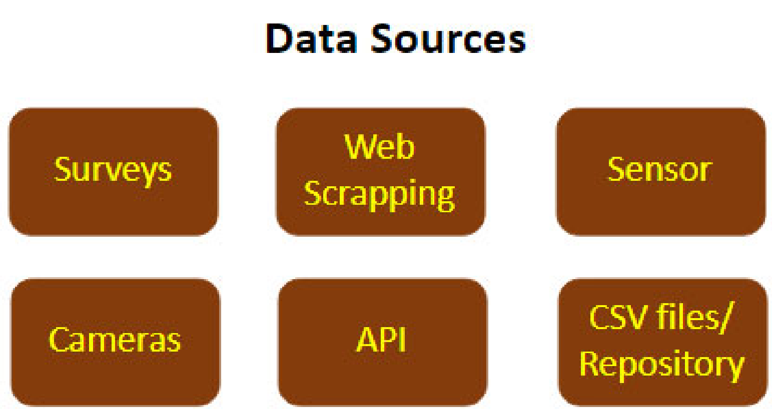


Figure 11.6: *Various Data Sources*

The data collected is further processed to identify data types in the data sets. It is also important to prepare the data in formats that would then be fed into the algorithms where data formats could be textual, numeric, or images and videos. Preparing them in the proper data frame to have appropriate columns, formats (say the text is in all capitals or small case or mixed) and likes.

For the capstone project, the following data sources, not an exhaustive list, provide data relevant to various problem domains:

- UC Irvine Machine Learning Repository **https://archive.ics.uci.edu/ml/datasets.php**
- Kaggle datasets **https://www.kaggle.com/datasets**
- Aws datasets **https://registry.opendata.aws/**
- India govt **https://data.gov.in/**
- World bank open data **https://data.worldbank.org/**
- Grouplens datasets **https://grouplens.org/datasets/**
- Google's Bigquery dataset **https://cloud.google.com/bigquery/public-data**
- Github repository **https://github.com/awesomedata/awesome-public-datasets**

Project 11.5

Use any of the data sources to collect data on cancer patients, their treatment, drugs and life span, and more such parameters for predicting the life of cancer patients depending on various parameters.

Modeling approach

The modeling approach is adopted at the stage in the life cycle of AI software, where the focus is on developing models. These models, based on analytical approaches, can be either descriptive or predictive.

A *Descriptive model*, a mathematical process, describes the realm it represents in a manner that can be interpreted by humans as well as computers. In other words, it details inter-relationships between real-world events and factors leading to these events. For example, a descriptive model of an enterprise comprises of vision, mission statements, as well as goals and strategies of the enterprise, among other things.

A *Predictive model* is a process that uses underlying patterns and relationships in data and probability to predict outcomes. For example, the predictive model may be used to determine whether the changes in product specifications will yield higher product sales or not. This is achieved once the model is fed historical data with already known outcomes. This process is iterative till the model is trained well enough to realistically predict the outcomes.

In this stage, hyperparameters are fine-tuned to calibrate the model to predict more accurately. Alternately, different algorithms can be chosen to ensure that the model parameters are chosen correctly.

The success of model creation depends on data processes, problem scoping and chosen analytical approach. The data gathered must be from the problem domain and provide an answer to the problem statement.

Models are a series of repeated processes with continuous refinements and tweaking to ensure realistic outcomes. The entire framework is geared toward understanding the problem statement, selecting the best suited analytic method, and data modeling and preparation.

How to validate the model quality

When building these models, ensure that the model not only works on the initially collected dataset but also for all similar types of data. That is, the model is not restricted to unique elements of the initial data subset.

Validating a model's quality is determining the accuracy with which the model responds in the real world with real data. Proving the validity of any model in its entirety is an unattainable goal.

The validity of a model cannot be established through any simple test, rather, it is an intuitive process through which the data scientist concludes the accuracy of the model based on available evidence. In other words, model validation is a set of processes and activities that examine its algorithms and relationships and are intended to verify model's performance in accordance with its intended purpose or system's definition.

Graphic animation can be used as a validation tool for models having complex processes.

For model validation, all possible data sets can never be available. *Figure 11.7* describes the subset of data (in a lighter shade) from entire data set that may be available for the validation process of the model:

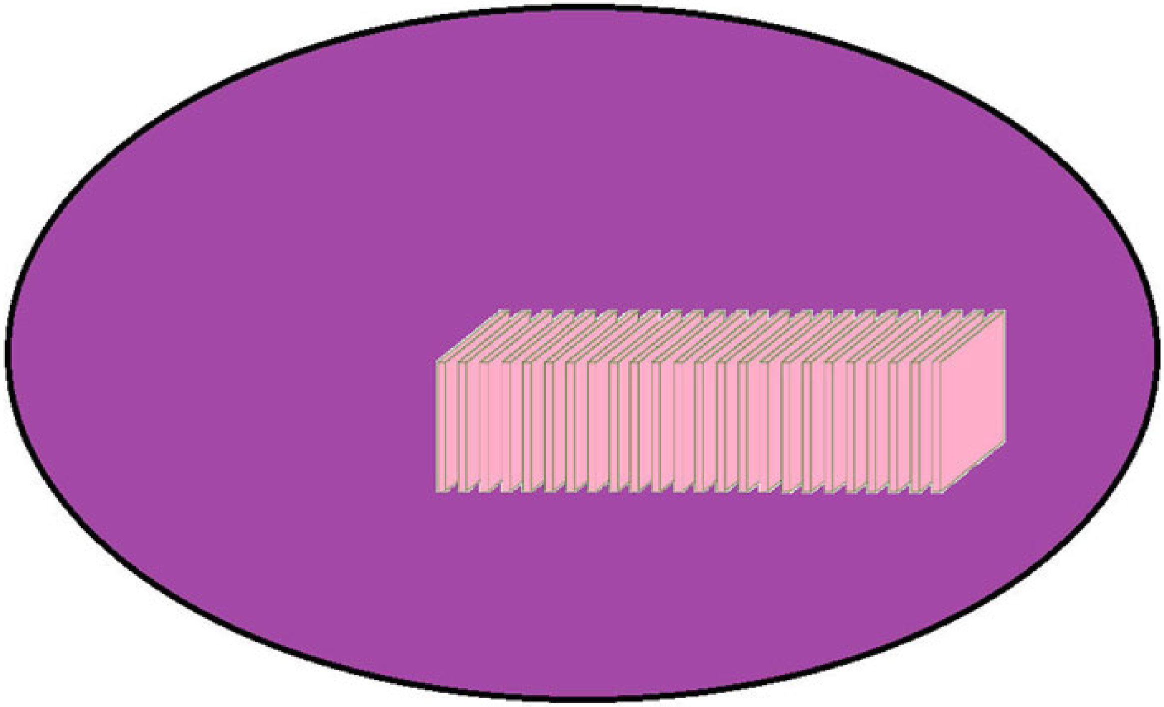


Figure 11.7: *All possible data sets and the subset available for the validation process*

Let us consider two of the model validation processes:

- Test-train split
- Cross-validation

The following sections summarize these two validation processes.

By test-train split

The train-test split is a performance evaluation technique for a machine learning algorithm.

Let us assume that the initial dataset available (from all possible datasets) is described in *Figure 11.8*:

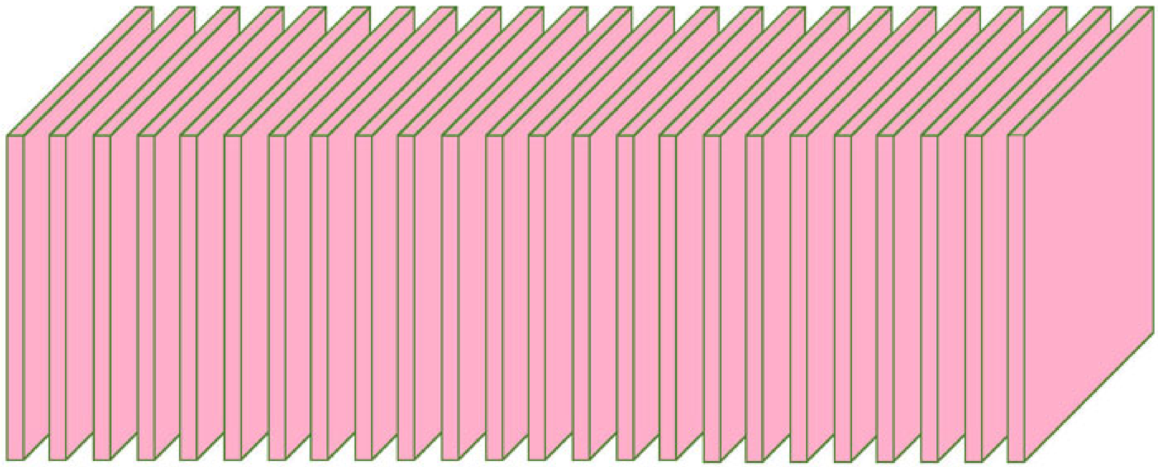


Figure 11.8: *Initial dataset*

The train-test split procedure divides the dataset into two subsets, as described in *Figure 11.9*:

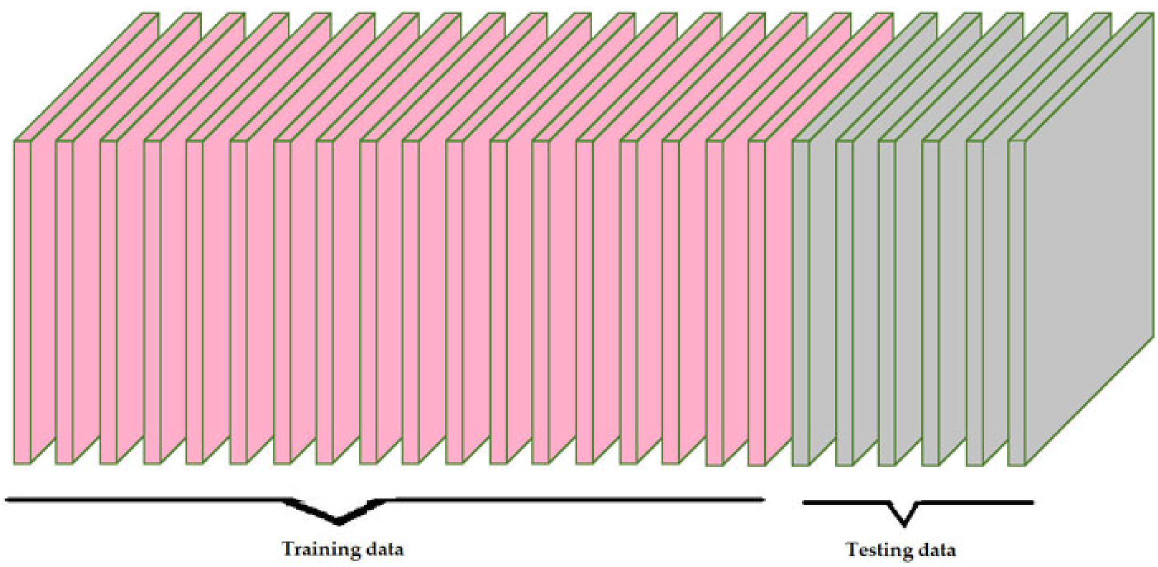


Figure 11.9: *Two data subsets – training and test*

The two subsets are referred to as

- Training
- Test

The training subset is used to fit the machine learning model and is referred to as the training dataset.

The test subset is provided as the input dataset to the machine learning model for predictions that are compared against the expected values. This data set is not used to train the model but only to test it.

The objective of this data split into two subsets is to gauge the performance of the machine learning model on the data (reference here is on test data) that was not used to fit it or train it.

The training data set is used to fit the model means to fit it on available data with known inputs and outputs. Post-training of the model, the test data is input to the model to make predictions where there are expected output or target values, but actual predicted output is unknown.

The train-test split methodology is best suited in cases of the availability of sufficiently large datasets.

We must also have answers to the following questions:

- How to split the data?
- What is optimal percentage of split between test and train subsets?

How to split the data

The data split must have sufficient representation of varied data between both training and testing. It shouldn't be that training set identifies to a unique set while the test set to another.

Optimal percentage for the split

The split is defined in terms of percentages. Though there is no optimal split percentage, it must meet the project's objectives. A project is limited not only by availability of data, technology, resources but by overall project costs as well. Hence the split percentages must be decided to keep the following considerations:

- Computational cost in model training.
- Computational cost in model evaluation.
- Training set characteristics.
- Test set characteristics.

Typically, common split percentages appear in the ratio of 80:20 or 67:33 or 50:50.

The train-test split methodology is used for supervised learning for both classification and regression problems. The methodology may not be best used with small datasets, due to the presence or absence of outliers and may limit evaluation outcomes, especially, the analysis of the model's extensibility.

Introduce concept of cross-validation

Cross-fold validation is more robust than a Train-test split. The essence of cross-fold validation remains the same as that of the train-test split methodology. In other words, cross-fold validation is applying train-test split, but several times. That is, the data is split multiple times such that every data point potentially falls within the testing or the training dataset.

While train-test split helps measure model quality on the basis of the test data, cross-validation extends the technique to model scoring or model validation. Cross-validation may take longer to run due to various splits in the data; it, however, gives a more reliable measure of the model's quality.

The cross-validation procedure

In cross-validation, different subsets of the data are used for the modeling process to get numerous measures of model quality. For example, the data can be divided into 5 pieces, each being 20% of the full dataset. Thus, there are 5 folds or experiments, as visualized in *Figure 11.10*:

Figure 11.10: *5 folds or experiments classified within total dataset*

Experiment n (from 1 to 5) can be run using the nth fold as test data, and the remaining data from the full set as training data. The results from one of the experiments with 20% of the test data are equivalent to results using the simple train-test split.

While evaluating results from all experiments put together, we have already used 100% of the data as test data at some point during any of the experiments.

The other way the experiments can be classified as in *Figure 11.11*:

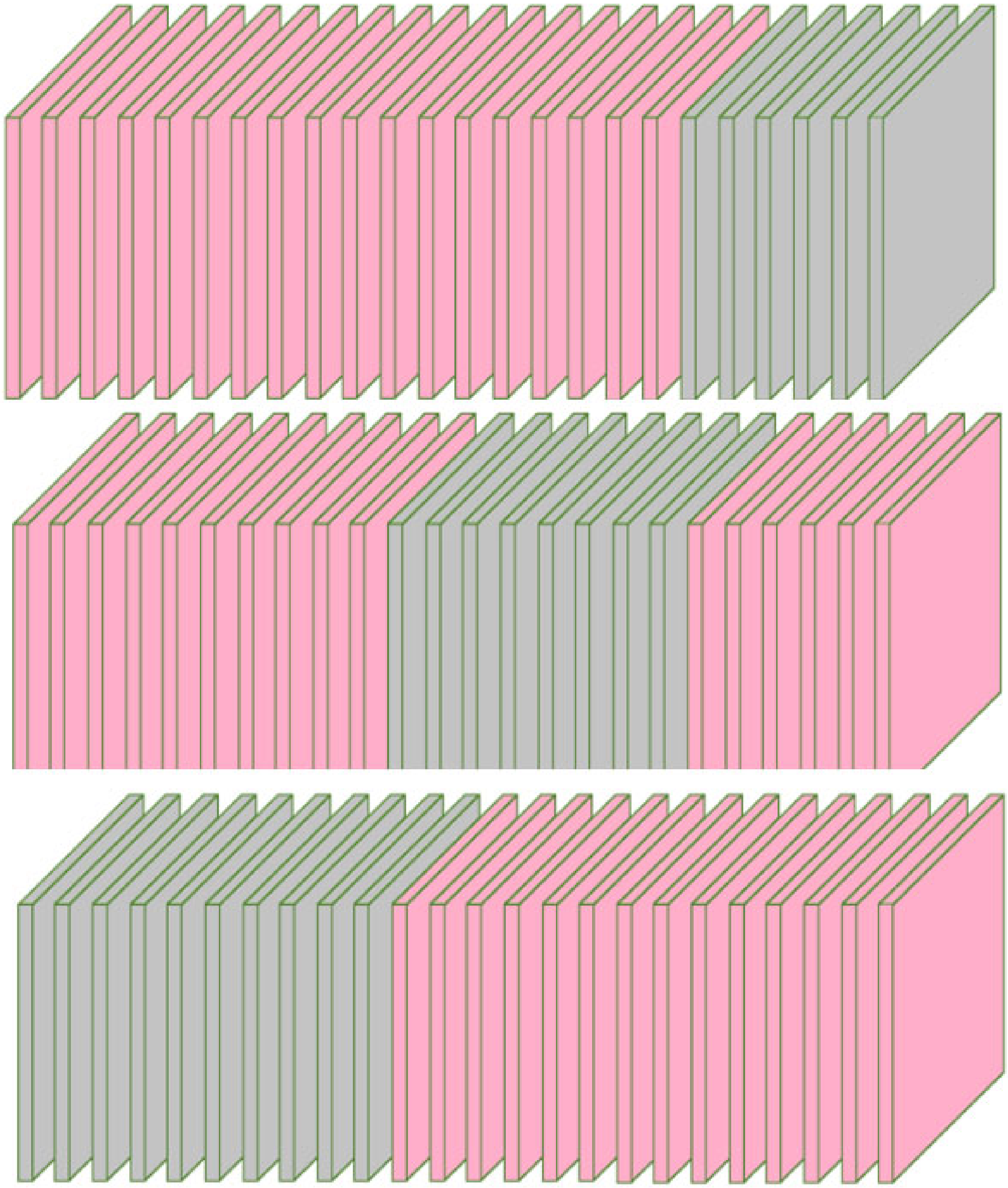


Figure 11.11: *Another set of experiments with different sets of tests and train*

Metrics of model quality by simple Maths

How do we know the following?

- The accuracy of predictions made by the model
- Performance of the model

Accuracy is a measure of true predicted values only, not specific for each label. Hence, the higher accuracy of the machine learning model doesn't always guarantee good performance on predicting a specific label. Therefore, overall model performance has more weightage than model accuracy.

For accuracy to be a useful measure, it is important to have same number of samples per class or label, otherwise, an imbalanced set of samples per label doesn’t make accuracy very useful.

Measuring the accuracy of predictions and model performance leads to the selection from multiple options of the following:

- **Data transforms**: Select one from the varied set of data transforms used for training the model.
- **Model**: Select from different machine learning models that are trained on the same data set.
- **Configurations (hyperparameters and parameters)**: Select from various configurations of the model when a model is trained on the same data set.

Performance metrics provide clarity on how accurate a set of predictions are, in turn, how good the model is. Thus, performance metrics make significant contributions as a required building block in the implementation of the machine learning algorithms from the start.

Performance of the model in terms of how good the model is in terms of predicting the expected outcome, is determined by the loss function. In other words, a loss function is a measure of model’s capability in terms of being able to predict the expected outcome or its capability to model the dataset.

All machine learning algorithms depend on minimizing or maximizing a function, termed “objective function” (a linear equation of the form C = ax + by). Here, minimizing or maximizing a function means the maximum possible value or the minimum possible value of that function.

“Loss functions” are the group of functions that are minimized. If the model’s predictions are totally off, the loss function’s output will be a higher number. In case the predictions of the model are good, output of the loss function will be a lower number.

Another optimization algorithm commonly used to train the machine learning models is Gradient descent which is used to gauge model’s accuracy with each iteration of parameter updates. The model continues to update its parameters to produce the smallest possible error, until the gradient descent function is close to or equal to zero.

The gradient points in the direction of the sharpest increase in the loss function. *Figure 11.12* describes how gradient descent algorithm uses the first derivative (the slope of the tangent line to the function at the point x) of the loss function to get to its minimum value:

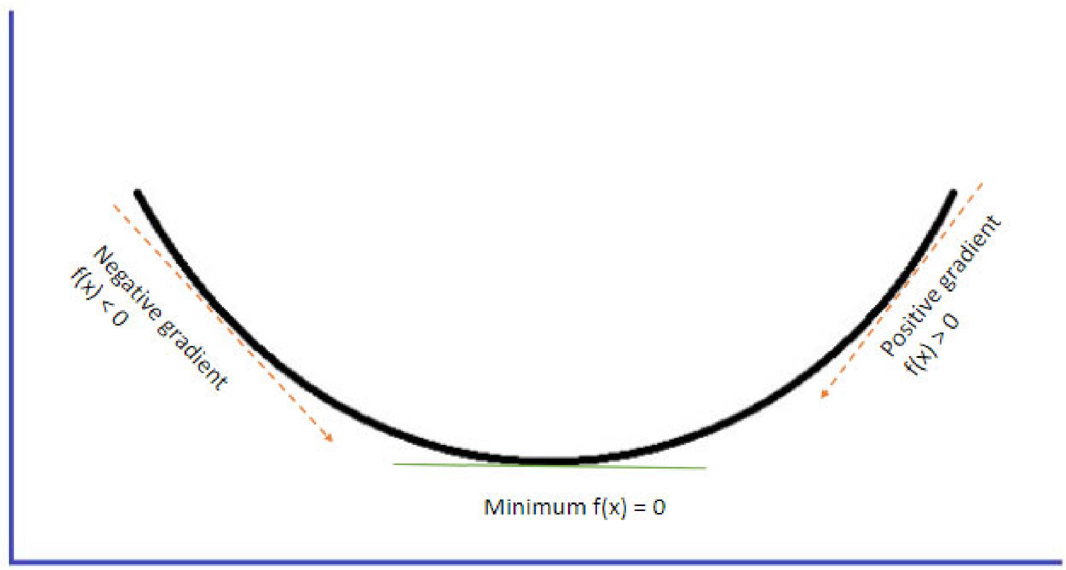


Figure 11.12: *Gradient descent algorithm using the first derivate of loss function to reach minimum value*

Loss functions are categorized into following two types, as also listed in *Figure 11.13*:

- Classification loss that predicts a label
- Regression loss that predicts a quantity

Classification Loss Functions	Regression Loss Functions
Log Loss	Mean Square Error/Quadratic Loss
Focal Loss	Mean Absolute Error
KL Divergence/Relative Entropy	Huber Loss/Smooth Mean Absolute Error
Hinge Loss	Quantile Loss
Exponential Loss	Log Cosh Loss

Figure 11.13: *Classification Loss and Regression Loss functions*

RMSE- Root Mean Squared Error

For a regression model the **Root Mean Squared Error** (**RMSE**) is one of the main model's performance indicators. RMSE is the measurement of the average difference between predicted values by the model and the actual values (true values). This helps in determining model accuracy in predicting the target values.

RMSE is a good measure of model's accuracy for:

- Comparing forecasting errors of different models
- Comparing model configurations for a particular variable

However, it being scale-dependent, it cannot be used to determine model accuracy with certain configurations for different variables. Scale-dependent refers to situations where the scale of measurement of the variables influences the relationship between the predicted variables and the outcome variable.

The formula to calculate RMSE is as follows:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}}$$

Where,

N is the number of data points.

RMSE being scale-dependent, is generally used over standardized data.

Large errors are relatively given high weightage in RMSE, since the errors are squared before they are averaged. This makes RMSE most useful in situations where large errors are specifically undesirable.

For cases where RMSE of test data is much higher than that of training data, it is likely that:

- Model has badly over fit the data
- Model tests well on sample, but has little accuracy when tested out of sample

Lower values of RMSE indicate better fit while values >1 means the model is not optimized well. This would indicate that feature change or hyperparameters tweaking is needed.

It is considered that RMSE values between 0.2 and 0.5 indicate that the model can predict the data accurately.

Graphically the values used in calculating RMSE can be represented as in *Figure 11.14*:

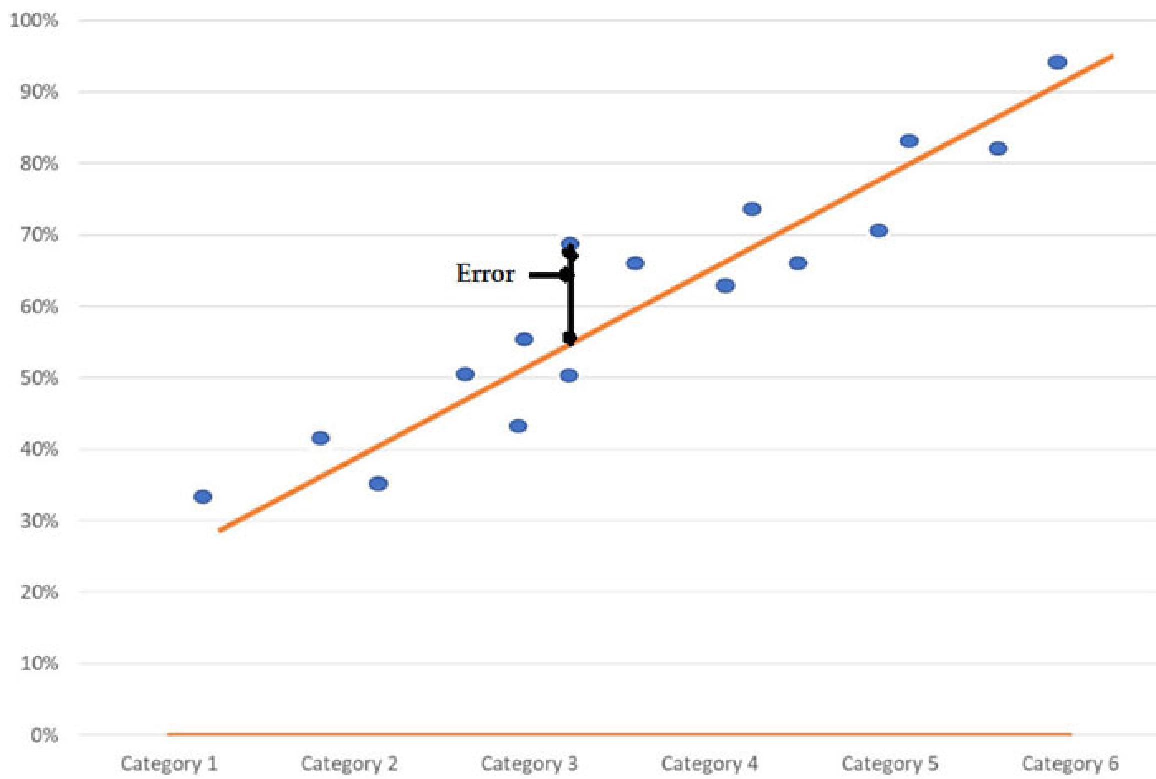


Figure 11.14: *Predicted, Actual and Residual values for RMSE calculation*

The scatter plot graph visualizes the following:

- The red line is the set of predicted values.
- The blue dots are the actual values.
- The black line represents the error for a single value (called residual), that is, the distance between the actual value and the predicted line for that data point. The residuals can be plotted for each point.

RMSE can be calculated as the square root of the average squared difference between the predicted and the actual values. Let us consider step wise approach as follows:

1. Calculate the residual (*residual = actual value - predicted value*).
2. Square each difference.
3. Sum up all the squared differences.
4. Divide the sum by the number of samples in the dataset.
5. The square root is taken to obtain the RMSE value.

RMSE is used in supervised learning applications, as RMSE is based on actual measurements at each predicted data point.

MSE – Mean squared error

Mean squared error is calculated by following steps:

1. Squaring the residual errors of each data point
2. Summing the squared errors
3. Dividing the sum by the total number of data points

The equation for MSE is as following:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Where,

n is the total number of observations

Y_i is the true data value at point i

$Y^\wedge{}_i$ is the predicted data value at point i

A low MSE value, that is, close to 0, is indicative of the model being a good fit for the data set. Conversely, a high MSE value far from 0 indicates that the model is not a good fit for the data set.

Why use mean squared error.

MSE is more sensitive towards outliers.

MSE is thus good to use if you believe that your target data, conditioned on the input, is normally distributed around a mean value, and when it's important to penalize outliers extra much.

A shortcoming in MSE is the fact that the unit of the variable is also squared, so if the model tries to forecast price in **Indian rupees** (**INR**), the MSE will produce a number with unit, currency squared or $(INR)^2$ which does not make sense in real world.

Project 11.6

Fine MSE and RMSE for following data set and analyze the model fitment.

X	Y (actual)	Y (estimated)
1	1	0.75
2	1.2	1.3
3	2	2.05
4	2.01	2.71
5	4	3.45

Table 11.1: *Sample data set 1*

MAPE – Mean Absolute Percent Error

One of the most common methods for determining the model's forecasting accuracy is MAPE. MAPE value can be calculated as follows:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|A_i - F_i|}{A_i} \times 100$$

MAPE = mean absolute percentage error

Where,

n is the total number of observations

A_i is the true data value at point i

F_i is the predicted data value at point i

MAPE is scale-independent and can thus be used to compare various forecast scenarios.

MAPE is easier to interpret and explain. For example, if the average difference between the forecasted value and the actual value is 1.5%, its MAPE value will be 1.5%.

A lower value of MAPE indicates that the model's forecasting accuracy is better. For example, a model having a MAPE value of 1% predicts more accurately than a model with a MAPE of value of 9%.

The MAPE, as a percentage and being scale-independent, doesn't make sense to calculate percentages of temperatures where ratios and divisions doesn't make sense. So one cannot use the MAPE to determine the accuracy of a temperature forecast.

Introduction to commonly used algorithms and the science behind them

Algorithms are computer programmed processes that are part of problem solving and are designed to perform a specific task. Algorithms can be used in simple arithmetic operations to complex artificial intelligence systems. Let us discuss some commonly used algorithms and the science behind them.

Machine learning algorithms are classified into four types:

Supervised learning algorithm

In supervised learning algorithms, dependent variables are predicted based on a set of independent variables. A relationship is established between input and output variables. The training is continued until certain accuracy levels are achieved.

The examples are as follows:

- Regression
- Decision Tree
- Random Forest
- KNN
- Logistic Regression

Unsupervised learning algorithm

Unsupervised learning algorithm is based on the classification of the data into number of clusters. These clusters have homogenous data within but are heterogenous from data in other clusters.

The examples are as follows:

- Apriori algorithm
- K-means

Semi-supervised algorithm

Text document classifier:

- This algorithm falls between supervised and unsupervised learning and trains the model with a combination of a small amount of labeled and a large amount of unlabeled data.

Reinforcement learning algorithm

In the reinforcement learning algorithms, the model trains itself continuously using trial and error. It learns from past data and outcomes.

The examples are as follows:

- Markov decision process

  A deep understanding of mathematical and computational concepts, including data structures, computational theory, and optimization techniques, is needed to understand the theoretical concepts behind these algorithms. These machine learning algorithms have their own strengths and weaknesses. Two main measures for the efficiency of an algorithm are time complexity and space complexity.

  Time complexity quantifies the amount of time required by an algorithm to solve a problem. Space complexity quantifies the amount of memory required by the algorithm. The goal of data scientists is to develop efficient algorithms that can perform successfully in the least amount of time and space possible.

Showcase through a compelling story

(Question, do we have to submit the story?)

In this section, we will learn about two real-world applications that are a few of the prominent examples to showcase machine learning algorithms.

AlphaGo

AlphaGo is an AI program created by Google's *DeepMind*. It is an application that is designed to play the ancient Chinese game of Go. In 2016, in a five-game match, AlphaGo defeated one of the best human players in the world, Lee Sedol.

Go, because of its complexity and the vast number of potential moves, is considered one of the most challenging games for AI to champion. The possible moves in Go are more than atoms in the universe, unlike in chess with limited moves.

AlphaGo uses a combination of deep neural networks and Monte Carlo tree search (a combination of classic tree search alongside machine learning principles of reinforcement learning) algorithms. Thousands of expert games of Go were used to train the neural networks to learn the best moves in different situations. The potential moves and the best next moves were determined using the Monte Carlo tree search algorithm using these networks.

Deep Patient

Deep Patient is an AI-powered system created by researchers at Mount Sinai Hospital in New York. It is used to predict patient health results based on medical records.

Deep Patient uses deep learning algorithms to analyze electronic medical records and identify the underlying patterns and relationships between various medical conditions, treatments, and outcomes that otherwise are difficult for human doctors to detect. It then can make predictions about the patient's health.

The medical team tested Deep Patient to predict the likelihood of patients developing various mental conditions, including schizophrenia, bipolar disorder, and other medical conditions such as diabetes. It was a success as the program can make predictions with a high degree of accuracy.

Project 11.7

Development of a Predictive Model for performance of student

Introduction: Being good at academics and improving scores in mathematical and computational science is key to becoming a data scientist or machine learning engineer. Let us then design a machine learning model that helps teachers to find out the parameters that are impacting a student's score.

Data collection: The data for this project will be collected from your own class or school. The database should contain information of more than 80 students including personal information, lifestyle factors, and various measurements related to studying patterns.

Methods: Pre-process the data to remove missing or inconsistent values. Select features using statistical analysis and feature selection techniques, say Correlation Coefficient. Split The data into training and testing sets to build and evaluate the predictive model.

Apply machine learning algorithms such as logistic regression, decision trees, random forests, and support vector machines. Compare the performance of each algorithm using performance and accuracy metrics. Select the best-performing algorithm as the final model.

Evaluation: The final model will be evaluated on the testing data to determine its performance in predicting the marks of the student in upcoming exams. The model will also be compared to other predictive models available to determine its overall accuracy and reliability.

Conclusion: This project will contribute to the development of a predictive model for giving attention to students in areas they need support for improving their performance, understanding, and scores. The model will be a valuable tool for teachers to analyze students' performance before it's too late in the year. Further research for model validation can be done to determine its generalizability and real-world applicability.

Conclusion

In this chapter, we learned to decompose the problem statement using a design thinking framework, validation of model quality, and performance analysis of the model with metrics. We revisited commonly used algorithms and applied our knowledge in showcasing a compelling story using all of these.

In the next chapter, we will learn about different aspects of a model and its lifecycle.

Multiple choice questions

1. **Which of the following cannot be used for currency variables?**
    a. RMSE
    b. MSE

    c. MAPE
    d. All of the above

2. **Cross-validation is inherently repeated train-test split methodology. Is this true or false?**
    a. True
    b. False

3. **What is not a data source?**
    a. Web Crawling
    b. Sensors
    c. APIs
    d. Process

Answers

1. **b**
2. **a**
3. **d**

Questions

1. What is cross-validation? Explain with an example.
2. What is MAPE? Explain with an example.
3. Explain the four types of algorithms and the concept behind them.

Join our book's Discord space

Join the book's Discord Workspace for Latest updates, Offers, Tech happenings around the world, New Release and Sessions with the Authors:

https://discord.bpbonline.com

CHAPTER 12

Model Lifecycle (Knowledge)

Introduction

We have learned the machine-learning models and understood the significance of associated data for proper outcomes. It is very important to learn about the data sets used to train these models before they are reliable for use in AI-based systems.

In this chapter, we will learn about the life cycle of a machine-learning model, the data sets to be used in training, testing, and validation along with related tools available off-the-shelf.

Structure

In this chapter, we will be discussing:

- Different aspects of the model
    - Train, test, validate,
    - What are hyperparameters
    - Commonly used platforms to build and run models (Introduction)
    - Recommended tools
    - Links to different platforms
        - Watson and others
- Lifecycle of an AI model
    - Build
    - Deploy
    - Retrain

Different aspects of model

Let us look at the different aspects of model in the following:

Train, test, and validate

For building a machine-learning model, the significance of the data set fed is undeniable. The original data set needs to partition if training a supervised learning model. The data set, therefore, can be partitioned into the following categories:

- Training dataset
- Validation dataset
- Test dataset

The same dataset shouldn't be used for testing and evaluation of any machine-learning model This is one of the key rules for making machine-learning models reliable, unbiased, and have their outcomes close to reality, more accurate or as desired.

The data, must, therefore, be split into training, validation, and test sets. It is important to establish that each of the subsets of original data contains the patterns or trends of the original data. It may otherwise lead to a model that is biased and is based on the patterns or trends in the subset data, that is, validation data.

The split is described in *Figure 12.1*:

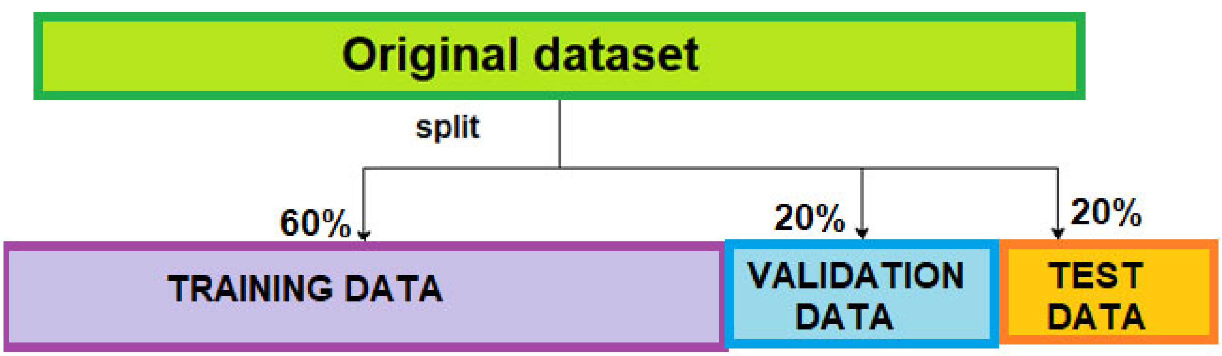


Figure 12.1: *Data split into test, validation, and training data*

Typically, the split proportion is often done as 60:20:20 (where 60% is for Training data, 20% is for Validation data, and the remaining 20% is for Test data). The ratio could also be 50:25:25. The split ratios depend on the size and type of dataset used.

Train with training data set

The set of data used for training a model is termed a training dataset as in *Figure 12.2*. This is also the largest dataset used in the machine-learning model. The model uses this data set to learn and understand the behavior based on underlying patterns and trends in the dataset. The training of the model, based on the training dataset is continuous.

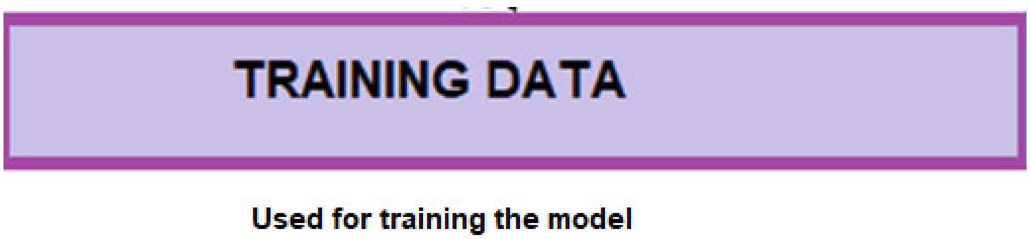


Figure 12.2: *Use of Training data*

The model continues to learn, during each iteration, when the same training data is fed to the model repeatedly.

For the model to be well trained, the training set should have a varied set of inputs covering all possible scenarios. This ensures that the model can predict reliably and even forecast any unseen data sample that may appear anytime in the future.

Validation dataset

The validation set is smaller than the learning set. The model doesn't learn from the validation set, but validates its performance and accuracy based on this set of data as in *Figure 12.3*:

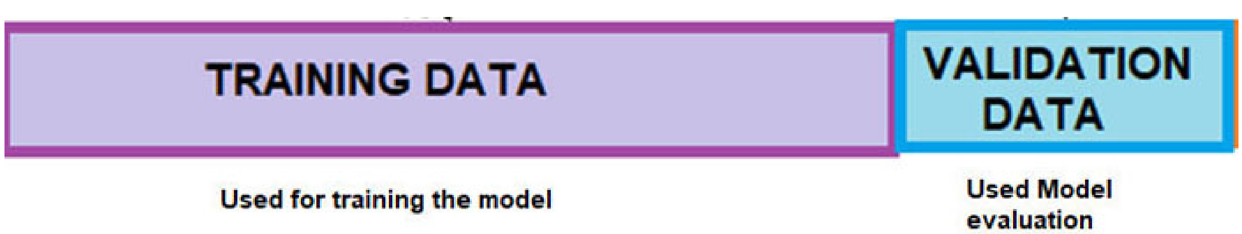


Figure 12.3: *Use of validation data*

The validation dataset is used to fine-tune the model hyperparameters by evaluating the performance of model with different hyperparameter values. Hyperparameters are explained in the following section of this chapter.

This dataset is also used to detect overfitting during the training stages, where overfitting is defined as the problem when random variations in samples unseen before are incorrectly classified as significant patterns. The process of validation is like an expert observer guiding whether the training is proceeding in the right direction to achieve the intended outcomes.

After each iteration of model is trained on training data, the model evaluation is performed on the validation set.

Test set

After being fed the training data set for learning and validation data sets for hyperparameters fine-tuning, the model is fed the test data for evaluating its final performance, as in *Figure 12.4*. Here, it is important to note that test data is final stage and is used only after the model has finished training and fine-tuning in the previous two stages.

Figure 12.4: *Use of test data*

It also gives the compares of different models' performance as compared against the other.

Test data provides an answer to the query "How effectively does the model perform" by providing unbiased performance metrics such as accuracy and precision. *Figure 12.5* describes the entire process flow in a diagram:

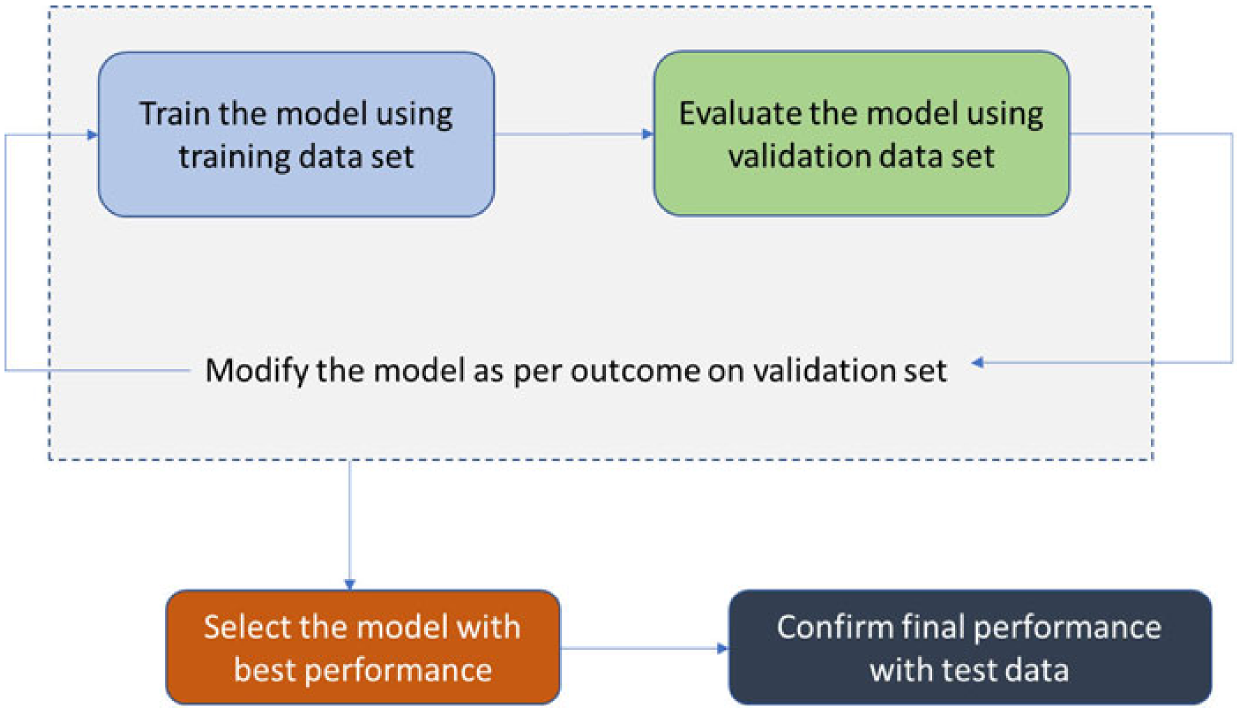


Figure 12.5: *Train, validate and test the process of a model*

What are hyperparameters?

There is a difference between machine-learning algorithms and machine-learning models. An algorithm is a piece of code written to get a task done or solve a problem and run on data, while machine-learning models are well-defined computations formed as outputs by algorithms comprising model data and a prediction algorithm.

Machine learning models have two different types of parameters:

Hyperparameters: The set of parameters external to the model that can be arbitrarily set before training starts.

Parameters: The set of parameters part of the model and are learned during the model training.

As such, to produce a desired model, the special values of machine-learning algorithms configured before the start of the training process are called hyperparameters.

Hyperparameters

Hyperparameters are used to configure algorithms to change the way the algorithms learn relationships in data, thus altering the produced model. In other words, changing hyperparameters changes the complexity of the model produced. That is using the exact same algorithm, a more complex model can be created that can learn complicated data relationships, or a simpler model can be created that can learn simpler data relationships.

Taking into consideration the fact that the hyperparameter values are set before the training of the model begins, hyperparameters are considered external to the model, also because during training the model cannot change the values of hyperparameters. In addition, the fact that the hyperparameters are not part of the model itself, there is no way for the model to know the hyperparameter values that were used to train it.

Let us consider an example of a hyperparameter. One of the basic examples is the depth of a decision tree. A tree, having just a few levels, or a shorter tree, is a simple model that learns basic relationships between inputs and outputs. Training such a model is, compared, simple and fast.

A decision tree having multiple levels, or a deeper tree is capable of learning much more complex data relationships and thus can handle difficult and complex problems. Training such a model takes longer and there is a risk of overfitting, that is fits, too tightly against its training data.

Figure 12.6 is to indicate the levels of a short tree and a deeper tree:

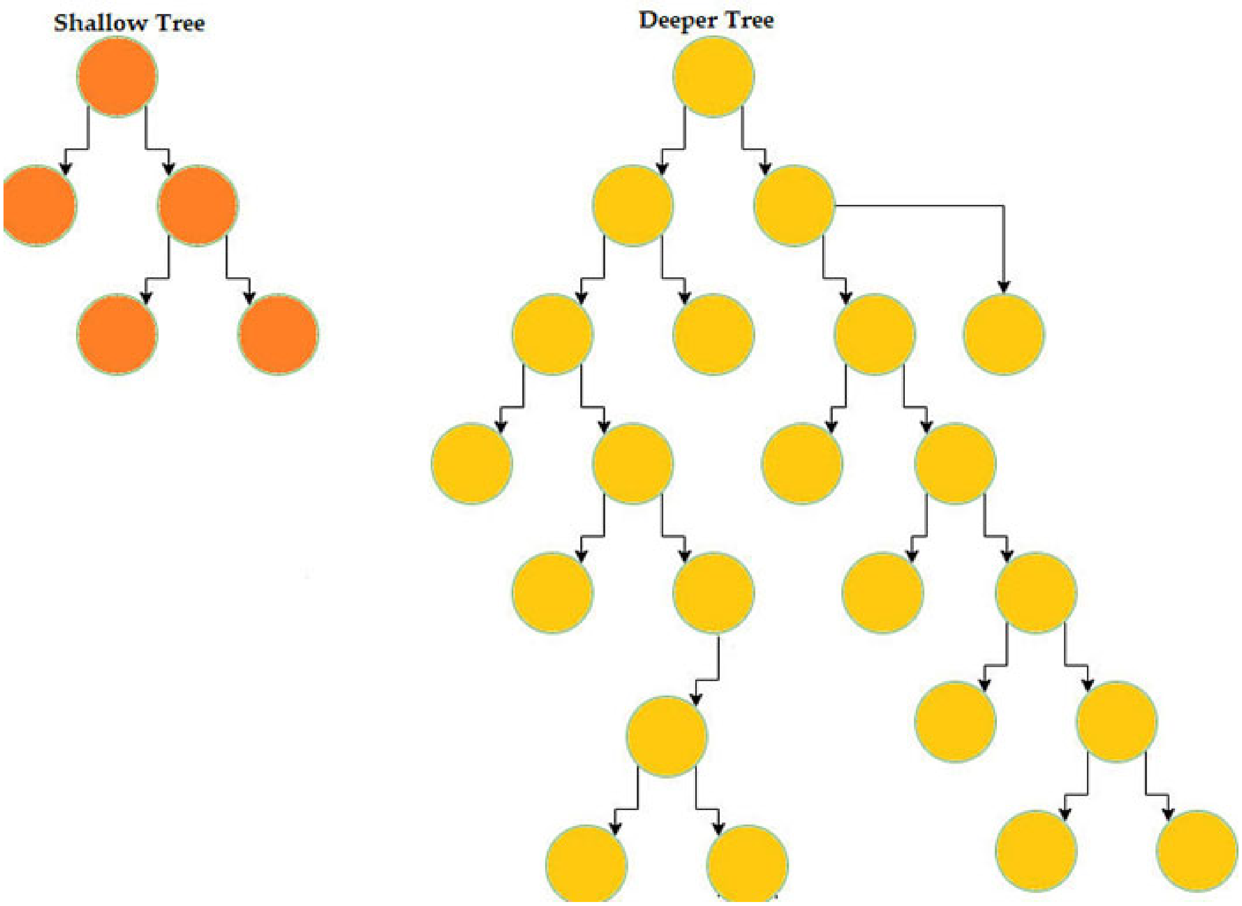


Figure 12.6: *Depth of a decision tree*

Following are some more common examples of hyperparameters especially for a neural network:

- Train-test dataset split ratio
- Activation function choice
- Number of
    - Hidden layers
    - Activation units in each layer
    - Iterations in training
    - Clusters in a clustering task

Commonly used platforms to build and run models (Introduction)

The process of developing machine-learning models and machine-learning-based AI applications in an efficient manner is termed a machine-learning lifecycle. Building and training models are long tedious processes, especially with the data collection and preparation for training and testing the model.

Thus, arise multiple tedious activities of managing data, code, model environments, and the machine-learning models themselves. These activities of deploying, monitoring, and retraining the models, become unmanageable with multiple models in production (in use by users).

A machine-learning platform automates and fastens the lifecycle of machine-learning models and associated predictive applications that operate on large data sets.

These platforms allow the team to build blocks and an environment to incorporate the solutions into products. Let's check out the common machine-learning model platforms:

Amazon SageMaker

Amazon SageMaker is a machine-learning lifecycle management platform that helps to build, train and manage machine-learning models. It also helps in deploying the model in a production environment. SageMaker comes with purpose-built tools that help in tasks such as abelling, data preparation, model training, monitoring and tuning, and much more.

Figure 12.7 taken from **https://aws.amazon.com/blogs/aws/sagemaker/**:

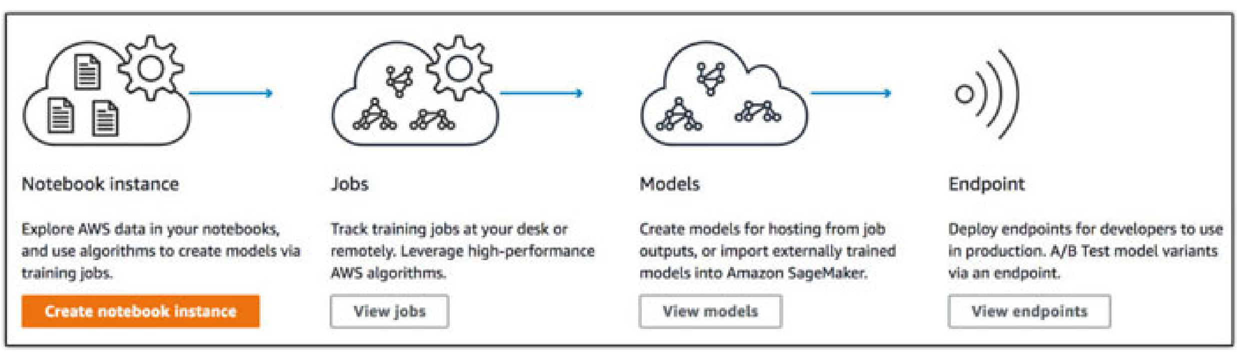


Figure 12.7: *Amazone SageMaker*

The end-to-end machine-learning platform accelerates the process of abelling, abelling, and deployment. Following are a few highlights about Sagemaker:

- For model training, Sagemaker includes many algorithms that help improve the accuracy, scale and speed of the model.
- For learning, Sagemaker includes both supervised and unsupervised ML algorithms.
- Its AutoML features automatically build, train and tune the best ML Model based on collected data.

Google Cloud AI Platform

Google Cloud AI platform, a managed service, is an end-to-end fully managed machine-learning lifecycle management platform. It has features that support faster and seamless management of services that make things easy for developers, data scientists and engineers. Let us study the features that perform machines learning model activities efficiently:

- Cloud storage helps with datasets preparation and storage with a built-in feature to label the data.
- Auto ML feature, with an easy-to-use UI, helps perform different tasks without writing any code.
- It supports many open-source frameworks.
- Auto ML features help with the deployment of the model while having features to perform real-time actions on the model.
- Provides end-to-end model management and monitoring providing ways to improve the model and data.

Google combined the two different tools, AI Platform and AutoML, into a single service which previously were offered as separate services on the google cloud platform. This unified machine-learning platform is called Vertex AI which is used to build, deploy and scale AI models.

Figure 12.8 describes Vertex AI. Image is sourced from **https://cloud.google.com/vertex-ai**:

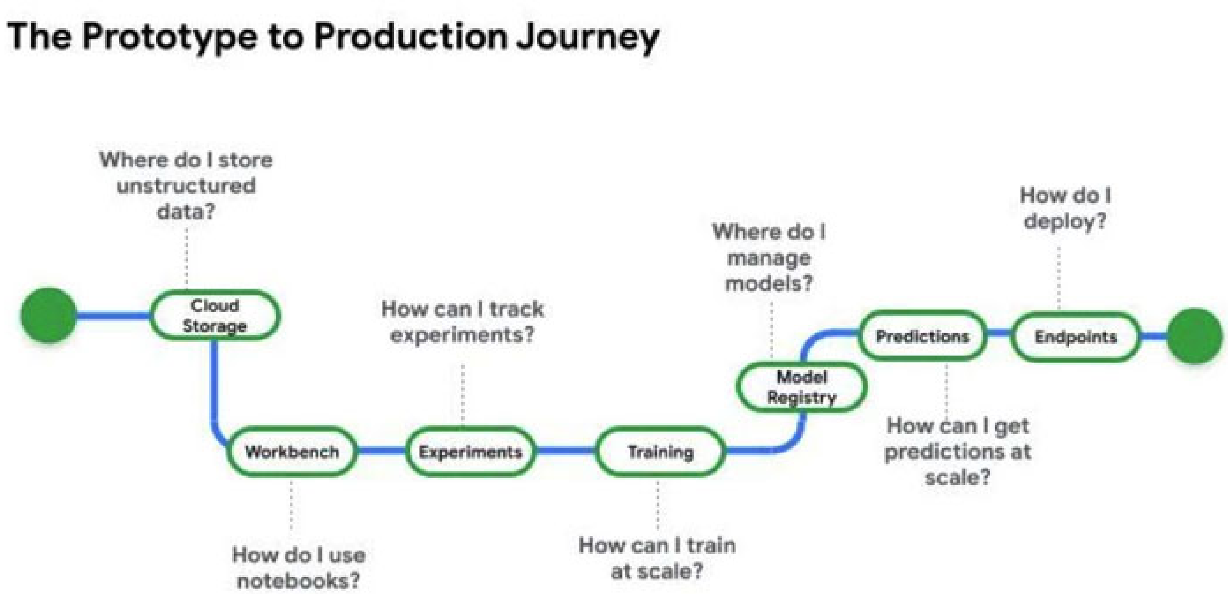


Figure 12.8: *Vertex AI*

Microsoft Azure

Azure ML by Microsoft is a cloud-based platform that is used for accelerating and managing the ML project lifecycle, that is, train, deploy, automate, manage, and monitor all machine-learning activities. It also supports both supervised and unsupervised learning.

The following are a few of the highlights of Azure ML:

- Code-free environment with drag and drop feature.
- Provides the option to train the model on the local machine or Azure ML Cloud.
- It has its own open-source platform for MLOps.
- Key features include, AutoML, data labeling, robust MLOps and different types of cloud support.

Figure 12.9 describes full MLOps of Microsoft Azure, copied from following link:

https://techcommunity.microsoft.com/t5/ai-machine-learning-blog/mlops-maturity-model-with-azure-machine-learning/ba-p/3520625

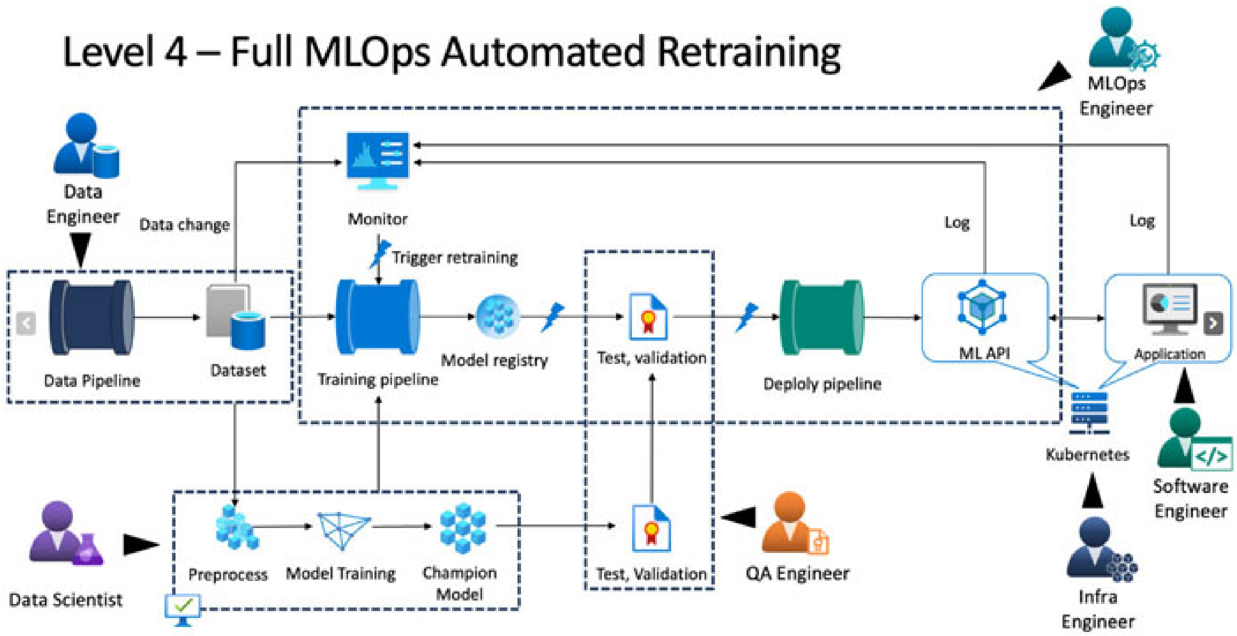


Figure 12.9: Microsoft Azure MLOps

Recommended tools

Few of the essential machine-learning capabilities that are needed in machine-learning platforms are face recognition, training and tuning. The following provides a summary of the best machine-learning software available off the shelf:

KNIME analytics platform

KNIME analytics platform is an online open-source machine-learning platform, that provides end-to-end data analysis, integration, and reporting.

TIBCO software

TIBCO is a data science platform. It supports the complete analytics lifecycle. It has the capability to integrate with many open source libraries.

Alteryx analytics

Alteryx, a data science platform, accelerates digital transformation. It offers self-service analytics including data accessibility and data science processes in a platform where data scientists can build models in a workflow.

SAS

SAS offerings include a robust suite of advanced analytics products. They also offer data science products.

H2O.ai

H2O.ai offerings include a wide range of AI and data science fully open-source platforms. It is a distributed in-memory ML platform with linear scalability specially used for deep learning.

Databricks unified analytics platform

It offers a cloud-based unified analytics platform. This platform is a combination of data engineering and data science.

DataRobot

DataRobot offerings include an enterprise AI machine-learning platform. This platform is used for fast and easy build, deploy and maintaining AI processes.

The platform includes independent but fully integrated tools such as Automated Machine-learning, Automated Time Series, and MLOps.

RapidMiner

RapidMiner's offerings include a data science platform for the entire AI lifecycle from data exploration, and preparation to model building, deployment, and operations.

Links to different platforms

- Watson: **https://www.ibm.com/in-en/watson**
- TensorFlow: **https://www.tensorflow.org/**
- Shogun: **https://github.com/shogun-toolbox**
- Apache Mahout: **https://mahout.apache.org//**
- Apache Spark-MLlib: **https://spark.apache.org/mllib/**

- Oryx 2: **http://oryx.io/**
- Pytorch: **https://pytorch.org/**
- RapidMiner: **https://rapidminer.com/**
- Weka: **https://www.weka.io/**
- Keras: **https://keras.io/**

Lifecycle of an AI model

We have learned that machine-learning models are well-defined computations formed as outputs by algorithms comprising model data and a prediction algorithm. Here the algorithms are programmed to solve a problem, while the model is trained rather than explicitly programmed.

Let us get introduced to two new terminologies: DevOps and MLOps. Both DevOps and MLOps software development strategies require collaboration between developers, operations, and data scientists. Here, DevOps focuses on application development while MLOps focuses on machine-learning.

A machine-learning model's lifecycle, from the training stage to deployment stage, must be repeatable, and the end-to-end process must be based on MLOps that is defined by three fundamental activities:

- Build
- Deploy
- Monitor

These three activities are followed by retraining.

The life cycle of an AI model is described in *Figure 12.10*:

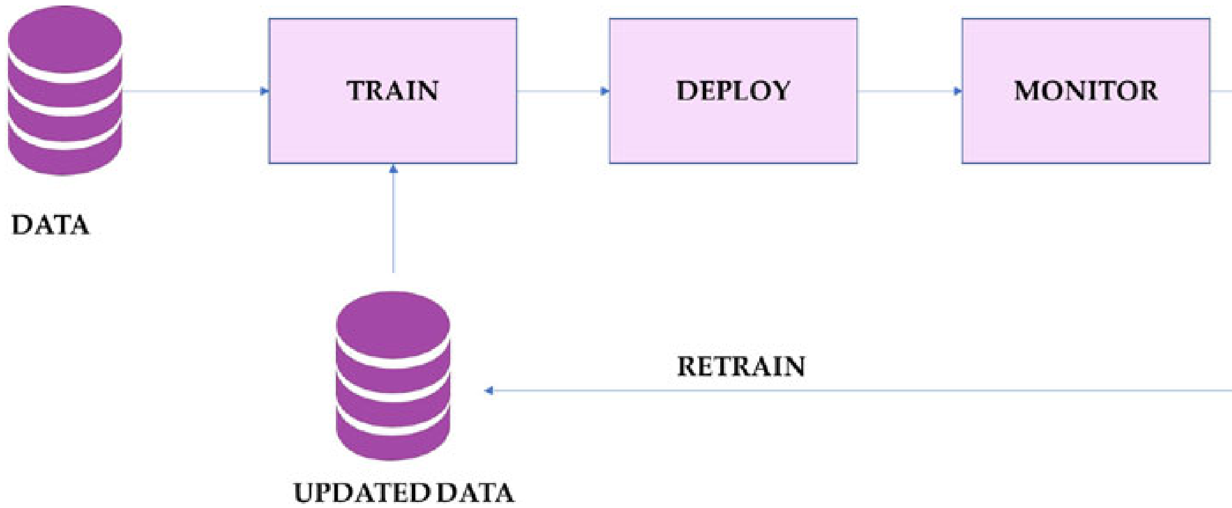


Figure 12.10: *AI Model Lifecycle*

Technically, the whole process of machine-learning model preparation has 8 steps. This framework represents the most basic way data scientists handle machine-learning. This is depicted in *Figure 12.11*:

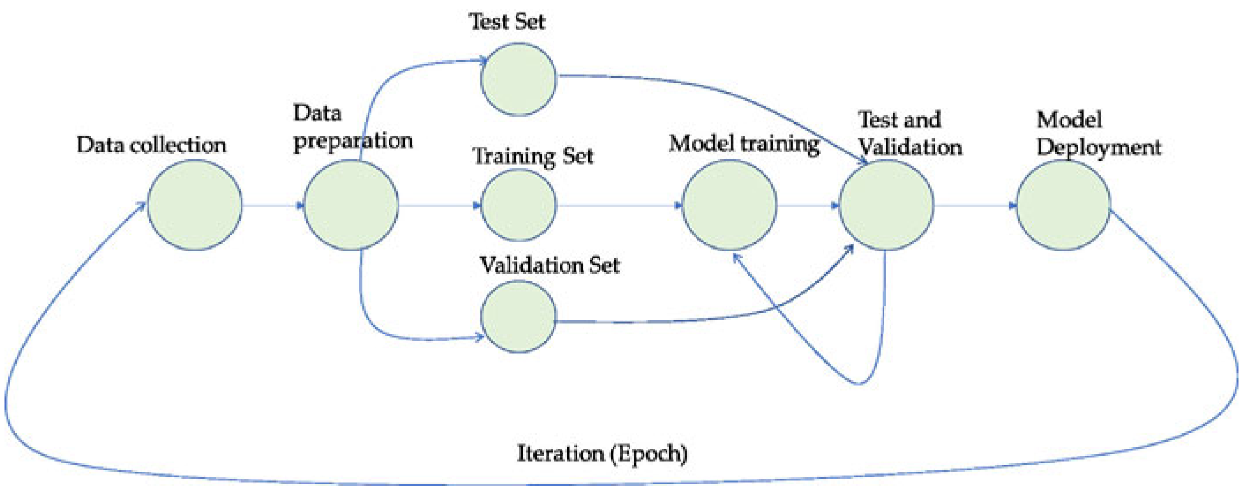


Figure 12.11: *AI Model Lifecycle technical stages*

Build

Model building has various sub-stages which include:

Data collection

AI model lifecycle starts with the collection of the required data that will be used for training.

François Chollet, a French AI researcher, is the creator of a deep-learning library and one of the contributors to the TensorFlow machine-learning framework, defines this step as "the problem definition."

Prepare the data

Data preparation means transforming the collected data by formatting, cleaning, labeling and enriching it.

Once data preparation step is through, the next step involves feature engineering which involves arriving at a set of features based on data attributes that have the most predictive power. Here, features are data values that are used both in training and production stages.

Choosing the algorithm

The algorithm forms the heart of any model and defines how a model will find patterns and trends in the data. Choice of the algorithm if one of the initial decisions that are done at the very beginning along with the data collection.

Train the model

Model training is the most significant part of the AI model lifecycle. We have already learned the process of data split and its importance in model training. The training can be based on methodologies like supervised learning, unsupervised learning or reinforcement learning. Model training enables it to make predictions on new data basis it is learning from historical data.

Test and validate model

Once the models are trained on the training data to ensure high forecasting accuracy, model then proceeds to the next step of testing against testing and validation data. These are iterative phases that are repeated unless the model reaches an acceptable prediction accuracy.

Deploy

The final stage of every software project is deploying the software in a production environment for the end user to use it. For machine-learning models, deployment would then mean that the end users can use it to get the predictions produced on live data.

Figure 12.12 describes how data scientists deploy their models as web services, such that they can be seamlessly used in various applications. The application's development team can access the web service endpoint or API to call from their applications in different ways.

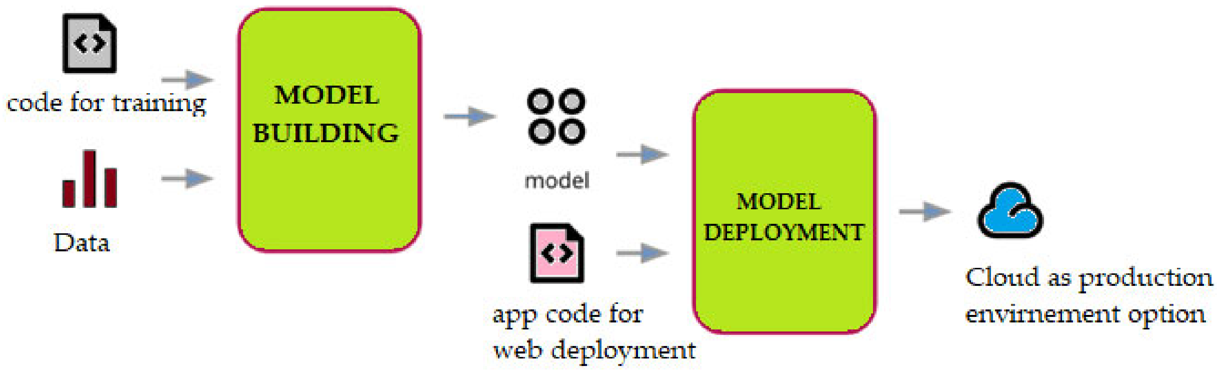


***Figure 12.12:** Model deployment stage*

Retrain

Machine-learning models require retraining enabling them to achieve high prediction capabilities, and accuracy and have desired outcomes close to reality. Additionally, possible scenarios where a model need to be retrained is addition of new input data, say, for example, addition of a new book in the e-commerce website. Retraining the model in such a scenario ensures its recommendation is up to date.

Retraining should be an automated process. Once initiated, the process includes pre-processing of data, model training, comparing outputs of new model to old model, and using predefined criteria to conclude whether to replace the old model.

Figure 12.13 describes the retraining of the model:

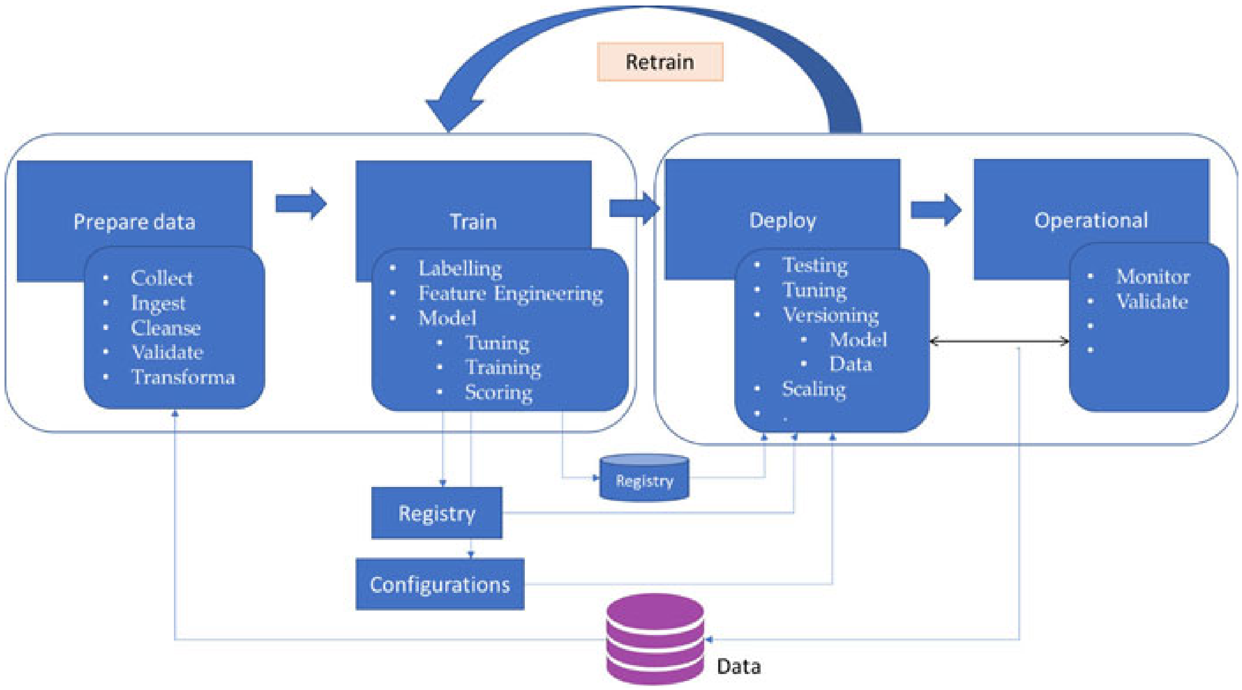


Figure 12.13: *Retrain the model*

Conclusion

By the end of this chapter, we have learned about machine-learning model life cycle management and related tools.

In the next chapter, we will be learning about storytelling through data.

Multiple choice questions

1. **A model is associated with three datasets. Which of the following is not one of these datasets?**
    a. Training
    b. Test
    c. Validation
    d. Labeling
2. **Which of the following are the three fundamental activities of machine learning operations?**
    a. Build

    b. Deploy
    c. Monitor
    d. Regression

3. **Addition of a new e-book to an e-commerce site would require which of the following model related activities:**
    a. Build
    b. Retrain
    c. Monitor
    d. Deploy

Answers

1. **d**
2. **a, b and c**
3. **b**

Questions

1. What are the stages of the AI life-cylce model?
2. Explain what hyperparameters are.
3. In what all scenarios are retraining a model required?

Join our book's Discord space

Join the book's Discord Workspace for Latest updates, Offers, Tech happenings around the world, New Release and Sessions with the Authors:

https://discord.bpbonline.com

CHAPTER 13

Storytelling Through Data

Introduction

"*Mama, today I had a fight with my best friend in my class*"! A student makes this remark to his mother. What do you think the mother's reaction would be? How many thoughts and queries would come to her? She would be eager to know the events that led to the fight incident and further details of the fight. Here the creative skills of the student to turn the events leading to the fight into a story brings life to the entire incident.

Let us consider the employment rate of men and women in a specific locality. This information can be depicted in the following two ways:

Simple statistics

- 58% of men and 28% of women are employed and work from the office.
- 13% of men and 12% of women are employed and work from home.
- 2% of men and 9% of women are unemployed.
- 19% of men and 2% of women are retired.
- 1% of men and 43% of women are homemakers.
- 4% of men and 3% of women are students.
- 3% of men and 3% of women are disabled.

In the form of visuals, say a pie graph as follows:

Figure 13.1 is a pie chart describing the employment status of men and women:

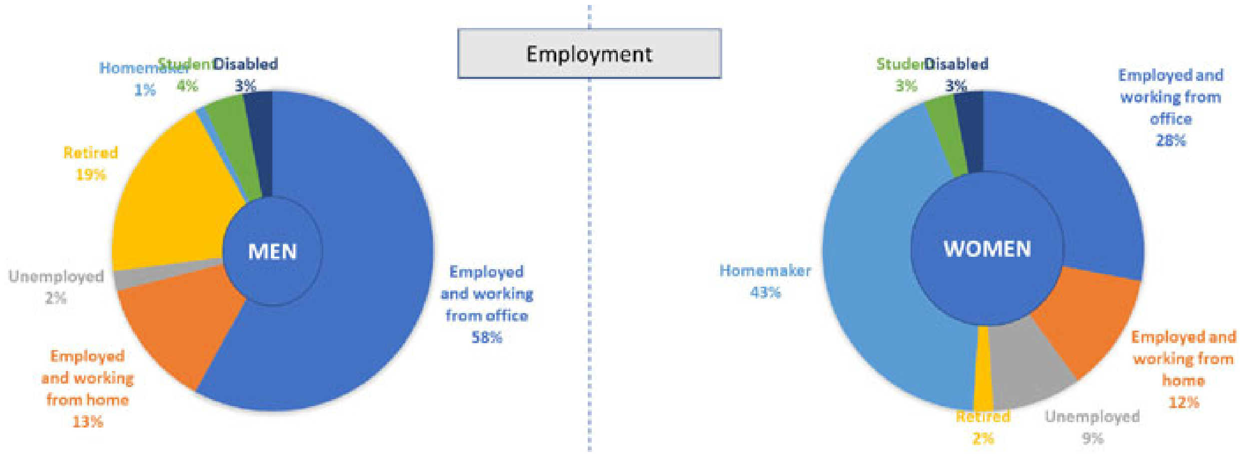


Figure 13.1: *Pie graph for the employment status of men and women*

Which of the statistical methods is easier to grasp and conveys a better story?

Structure

- The need for storytelling
    - Information processing and recalling stories
    - Why is storytelling important
    - Structure that story
- How to create stories
    - Begin with a pen-paper approach
    - Dig deeper to identify the sole purpose of your story
    - Use powerful headings
    - Design a roadmap
    - Conclude with brevity
- Ethics of storytelling
- Types of data and suitable charts
    - Text [Wordclouds]
    - Mixed [Facet Grids]
    - Numeric [Line Charts/ Bar Charts]
    - Stocks [Candlestick Charts]

- Geographic [Maps]

- Stories during the steps of predictive modeling
    - Data exploration
    - Feature visualizing
    - Model creation
    - Model comparisons

- Best practices of storytelling

The need for storytelling

A business manager remark, “the total sales are down as compared to last year,” is of hardly any use to the company executives unless it is backed by data substantiating the reason for the decline.

As such, the need for the story and, thereby, analyzing the situation becomes of utmost importance for corrective actions to be taken, if any. The data itself covey stories supported by visualization that complex mathematical operations need not be run on them.

The need for stories and visuals for data analysis is best explained with *Anscombe's Quartet*, representing four datasets as in *Figure 13.2:*

Anscombe's quartet

I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Figure 13.2: *Anscombe's Quartet with four datasets*

Figure 13.3 represents the statistical summary of the four datasets:

Property	Value	Accuracy
Mean of *x*	9	exact
Sample variance of *x* : s_x^2	11	exact
Mean of *y*	7.50	to 2 decimal places
Sample variance of *y* : s_y^2	4.125	±0.003
Correlation between *x* and *y*	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places

Figure 13.3: *Statistical summary of all four datasets*

Figure 13.4 represents the four datasets as visualized on graphs:

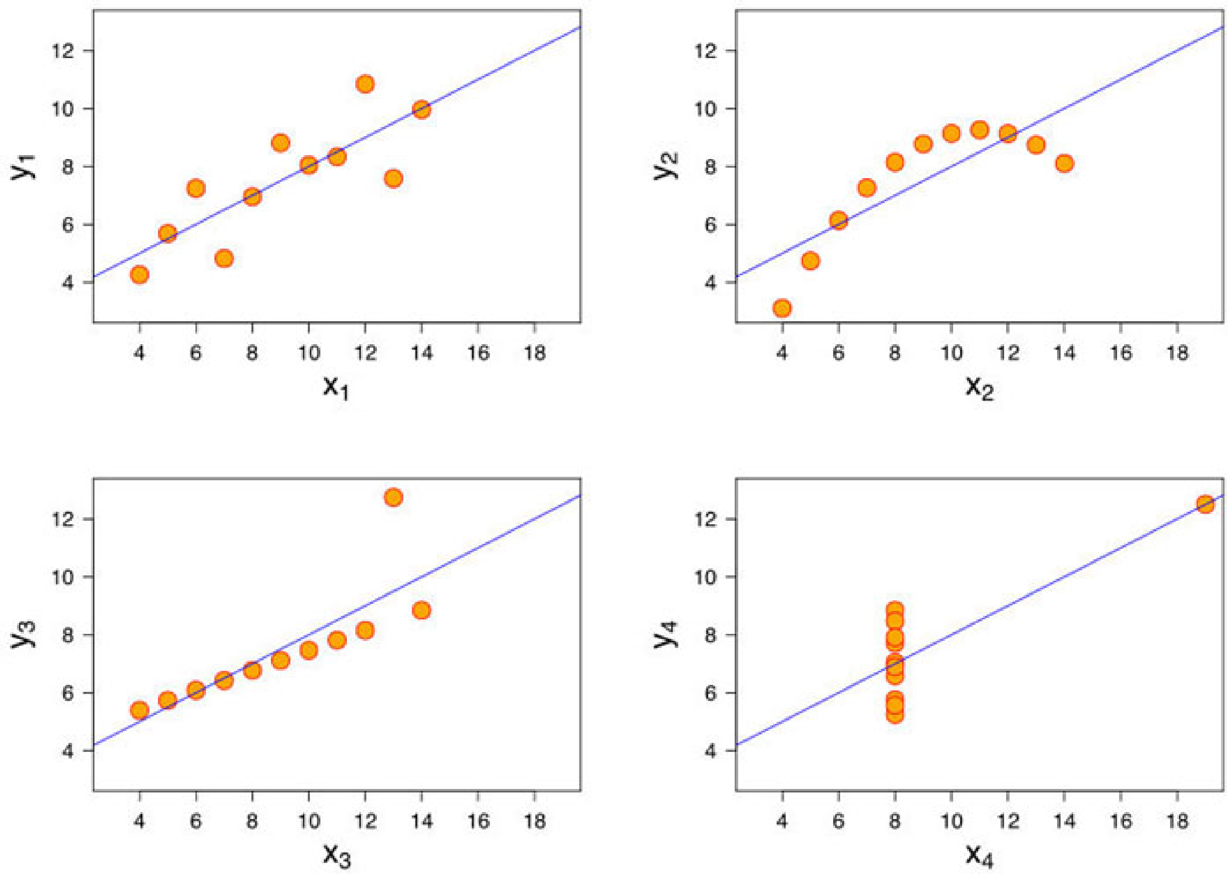


Figure 13.4: *Graphical representation of all four datasets*

The datasets that appeared seemingly identical based on their statistical summaries are actually very varied once they are graphed. This is the power of visualization, as it changed and added to the perspectives.

Refer to the Wikipedia page **https://en.wikipedia.org/wiki/Anscombe%27s_quartet** for further details on *Anscombe's Quartet*. *Figures 13.2*, *13.3*, and *13.4* are taken from the same Wikipedia page.

Information processing and recalling stories

Anscombe's Quartet datasets make it clear that information presentation is critical. It is also very important that the narrative set is also understood by the audience, whether agreed upon or not.

The utility of the narrative, as a powerful communication and thereby persuasive tool, supported by effective visual literature, is of no use if it fails to harness the attention of the audience. As such cognitive and emotional dimensions of the narrative add to the success of the presentation. Now, let us consider various aspects of the narrative flow and its impact on information processing, knowledge acquisition, and persuasion. Figure 13.5 describes the key ingredients in forming an impactful data story:

Figure 13.5: *Data Story and information processing*

Let us consider these ingredients and their significance in information processing as follows:

Attention and comprehension

For a seamless narrative flow, visual messages are more likely to stick to the end. Data must not only be aligned with the interest of the audience but also presented in statistical ways aimed to gain the *attention* of the audience. The audience must also be able to *comprehend* the presentation. It is very important to *prioritize style* over substance and align visuals with narratives. The audience's distractions must be avoided by keeping words and visuals *simple* but assertive. A simple story is more successful in making an impact than a complicated one.

Emotion and imagination

We intend to relate stories with our own existing experiences. To make an emotional connection with the audience and capture their imagination, the narration should include stories. Personal events and experiences, including as part of the narration, have a greater impact on the continued engagement of the audience. Stories regarding religion and other sensitive topics may be excluded to avoid provoking the audience.

Acceptance

It is very important for the presenter to assist the audience in processing the story. However, a good narration having conveyed the idea clearly does not necessarily would mean acceptance of the idea behind it. Narratives should be formed with the aim of overcoming audience resistance.

Learning

A thought-provoking narration with *learnings* and *key takeaways* adds to a successful presentation. Such narratives generate *continued interest* of the audience that would not like to miss any information to have any disconnect.

Inspiration

Addressing emotional and existential issues inspires the audience. Not only the audience but also the presenters' brains become more active when engaged in narrating stories. A good presenter can change the delivery style of the presentation to adjust to maintain audience's interest.

When a story is told, things change dramatically. As per researchers not only the language processing parts of brains are activated, but all other areas are used when experiencing or imagining the events of the story.

Actions

Stories, when broken down into the simplest forms prove to be a connection between cause and effect. Use narratives as a persuasive tool for the audience to act.

Why is storytelling important

Human beings communicate and motivate each other using techniques such as storytelling. Stories are powerful and magical that can invoke emotions, perceptions, ideas from completely different worlds. They have the power to teach us to hate or empathy and shape our thoughts in ways which any personal experience may be capable of. Eventful stories may leave a lasting impact on us and maybe permanently memorized. They are good for remembering content.

Organizations use storytelling techniques effectively for their internal processes involving product development or external communications such as marketing and sales.

Structure that story

Structure is the framework of a story, and the way it has been organized. In other words, it is the map of construction based on which the finished work has been outlined. The structure has significant elements like a plot, characters, setting, some uniqueness, a theme or maybe a mystery. The writing style and detailing are also key elements, while the most important element is the structure of the story.

An age-old narrative structure has following stages:

1. **Introduction**: This stage covers background of as-is or current situation, the character(s) and the hook.
2. **Incident (crisis or opportunity)**: This stage involves revealing supporting details that provide insights into main problem or conflict or opportunity, if any.
3. **The Aha moment (climax)**: The highest point of action in the story revealing central insight or major finding.
4. **Ending with the solution**: The end of the story involves putting forth potential solutions and recommendations.

Data stories are not secluded, from reaping benefits from the same age-old narrative structure. A data story follows the same outline beginning with the narration of current situation, proceeding by giving insights that lead up to the aha moment or climax and ending with potential solutions and recommendations. *Figure 13.6* describes the structure of the story.

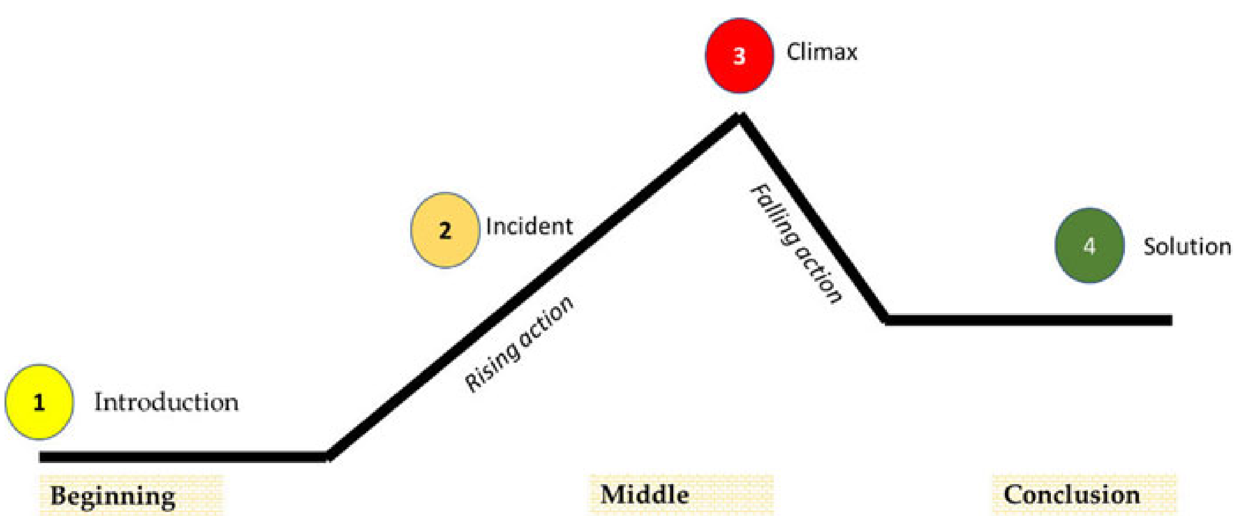


Figure 13.6: *structure of the story*

How to create stories

Having the framework or structure of the story handy, now lets us understand approach to creating meaningful stories and a good narrative flow.

Begin with a pen-paper approach

The first step towards creating visually engaging presentations is, as advisable, to write down the flow on a piece of paper. Scribbling the ideas and the flow before structuring the story will lead to more satisfying end products. The generated narrative flow may cause friction in the end product, since the context set requires relevant data, visuals and facts supporting it. It may also be decided types of graphs or visuals that may be used to convey the message and leave the audience with new learnings that may call for further action.

Dig deeper to identify the sole purpose of your story

Narrating a story with great presentation may miss the purpose unless the same has been identified and the narration revolves around it. Questions can be raised to self, such as:

"What message is being conveyed through the story", "Will the story help in decision-making", "Is the narration intriguing enough to capture interest of audience and keep them engaged". Prospects conveyed via passionate statements can excite the audience and the idea can be remembered for a long.

Use powerful headings

Powerful headlines for any kind of content (be it analysis, story, visuals, or questions) usually are crispy, specific, concise, and offer personal interests. These headings summarise and give a vision of what is expected from the content. These are written more from the audience's perspective and maintain engaging them.

Design a roadmap

The verbal roadmap of the story is designed by listing the key points that need to be conveyed to the audience. This list is a takeaway from the visuals, analysis and story that is presented. Supporting evidence like analogies, examples, facts must be added to substantiate each of the items on the list. The list, itself, can be concise to few items.

Conclude with brevity

The concluding statements and items should ideally be short and powerful. These may be the conclusions from the story, actions to be taken or the opportunities unearthed or problems to be addressed.

Ethics of storytelling

The power of storytelling comes with great responsibility. There is a thin line between "using a person's story" and empowering the person's voice via storytelling but "using or exploiting them". This is when arises the need for ethics in storytelling. The following lists of queries to ask and practices to adhere to while engaging in storytelling:

Queries

- **Each owns their story**: Do we have consent from the organization, individual or entity for their story to be told?
- **Informed consent**: Is the entity whose story is shared enthusiastically about the way the story is compiled and portrayed each time it is shared? This includes the right to withdraw.
- **Beneficiary**: Who will benefit from the story being shared?
- **Empowering**: Who will be empowered by the story?
- **Harm**: Is it reinforcement of any kind of harmful bias or social issue and stigma?
- **Life after story is shared**: What is journey of the entity whose story is shared and of those involved in the storytelling process as audience?

Practices

- Release power: when employing storytelling, it is vital to honor zealous consent by relinquishing power and following the needs and desires of the presenter.
- Power dynamics of presenter and audience: it's impossible to completely disassemble power and privilege, however, the goal remains to create a space where the presenter's sovereignty is honored and upheld largely and is not impacted by the positionalities of the audience.

Types of data and suitable charts

Different types of data use different visualization in form of graphs that can best convey their stories. Let us consider few common data types and their best-fit visualization.

Text [Wordclouds]

Worlcloud is best suited for textual data where larger text size either indicates the frequency of appearance of the word or the sentiment of the text where it appears.

Figure 13.7 describes a wordcloud from an executive job profile in a security agency describing entitlements, licenses, certifications and authority, and more. The word "entitle" is the most frequent and positive word.

Figure 13.7: *Wordcloud*

Mixed [Facet Grids]

Facet grids are used when data volumes are huge and it is nearly difficult to put all of it on a single graph. Also, putting on a single graph may fail to effectively reveal the patterns and relationships between variables.

In scenarios of huge data volumes, Facet grids or Trellis plots are used to visualize a subset of the data using multiple small graphs. Each visualization describes the relationship of certain variables under certain specific conditions. Here, the same plot type is used for all graphs making it easier for comparison.

For example, let us consider the following data as in *Table 13.1*:

Total Cheque amount (Euro)	Tip	Gender	Day	Time	Alcohol	Number of persons
83	5	Women	Monday	lunch	no	3
120	8	Men	Friday	dinner	yes	2
180	10	Women	Friday	dinner	yes	2
70	5	Women	Tuesday	lunch	no	2
78	5	Men	Wednesday	lunch	no	2
100	6	Men	Friday	dinner	yes	4

Table 13.1: *Total Cheque amount on different days and times of the day*

We can plot various graphs which can compare following:

Total Cheque amount at lunch or during dinner

Figure 13.8 describe the total cheque amount during lunch and during dinner as two separate graphs:

Figure 13.8: *Total cheque amount during lunch and during dinner as two separate graphs*

Tip amount at lunch or during dinner

The graphs in *Figure 13.9* describe the tip amount at lunch or during dinner:

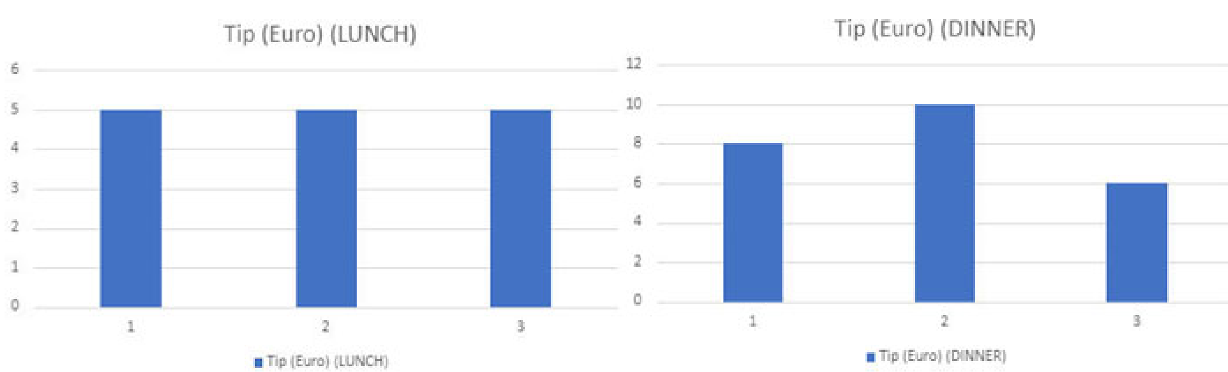


Figure 13.9: Tip during lunch and during dinner as two separate graphs

Total Cheque amount during dinner when alcohol is served and during lunch when it is not served

The graphs depicted in *Figure 13.10* describes the same:

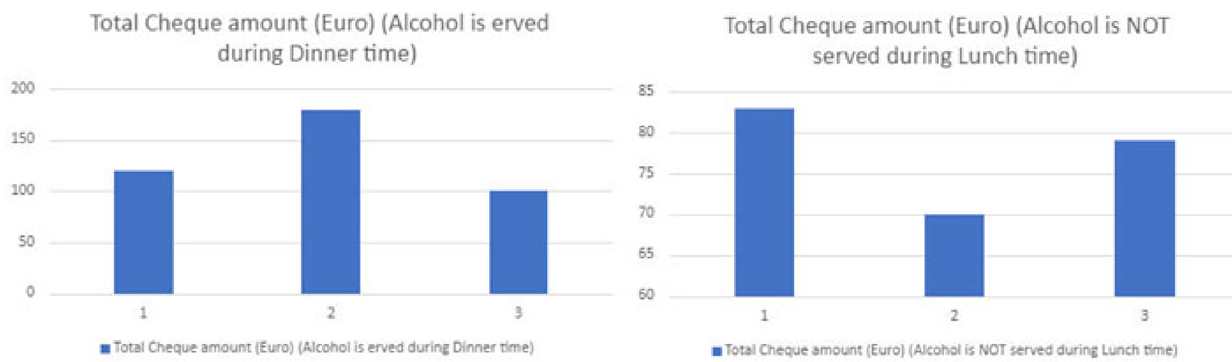


Figure 13.10: *Total Cheque amount during dinner when alcohol is served and during lunch when it is not served*

Facet grids can also be graphed in form of multivariate plots and following different colour coding as in *Figure 13.11*:

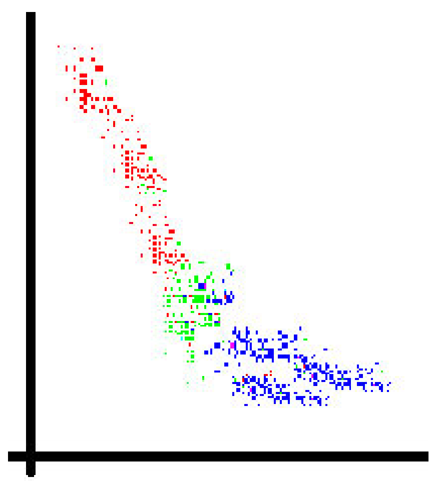


Figure 13.11: *Multivariate plot in different colors*

Numeric [Line Charts/ Bar Charts]

A line or a step chart is best suited for a numeric data. Even a bar chart can be used. *Figure 13.12* describes a step chart for gold prices each year (year 1 to year 7):

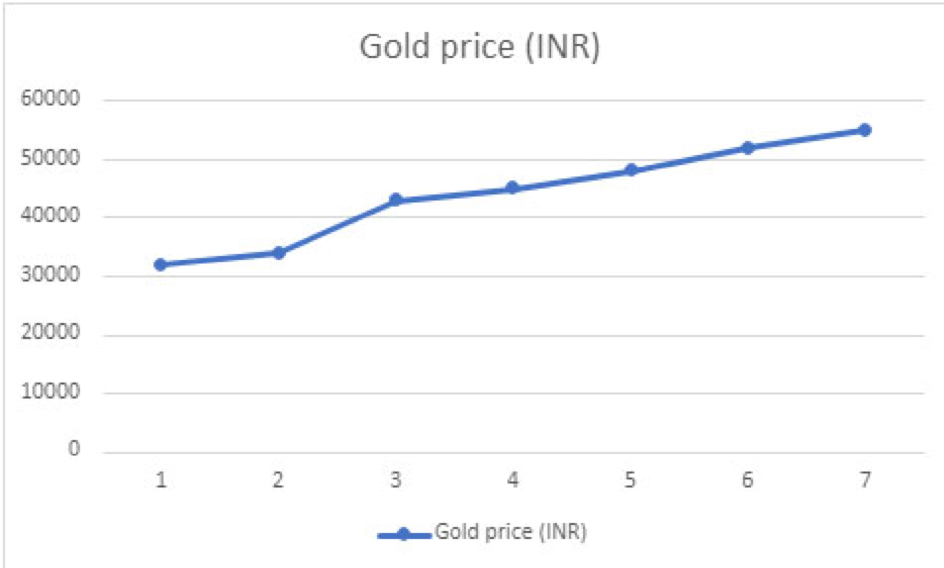


Figure 13.12: *Step chart for gold prices each year*

Stocks [Candlestick Charts]

Candlestick bars appear as candlesticks and are color coded for easier visuals. These are used to show the opening, closing, high, and low prices per day for the financial market over a period of time. *Figure 13.13* describes how to read candlesticks:

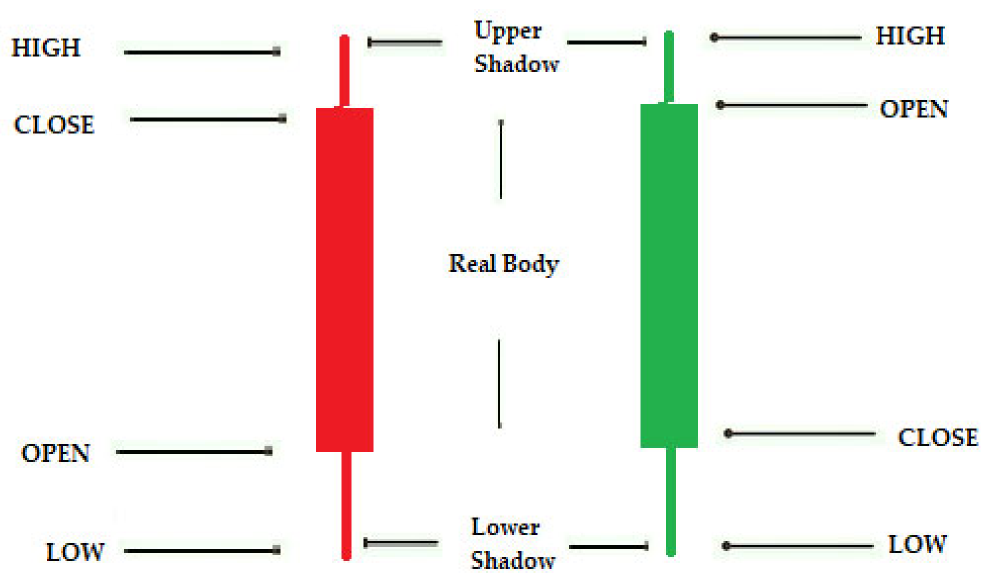


Figure 13.13: *Candlestick bars*

A candlestick chart used these candlesticks. It is a type of financial chart describing price action for any financial market as shown in *Figure 13.14*:

Figure 13.14: *The candlestick chart from investing.com for Netflix stock*

Geographic [Maps]

For data related to location, maps are best suited. *Figure 13.15 is* taken from thoughtco.com (**https://www.thoughtco.com/seismic-hazard-maps-of-the-world-1441205**) that describes the global seismic hazard map:

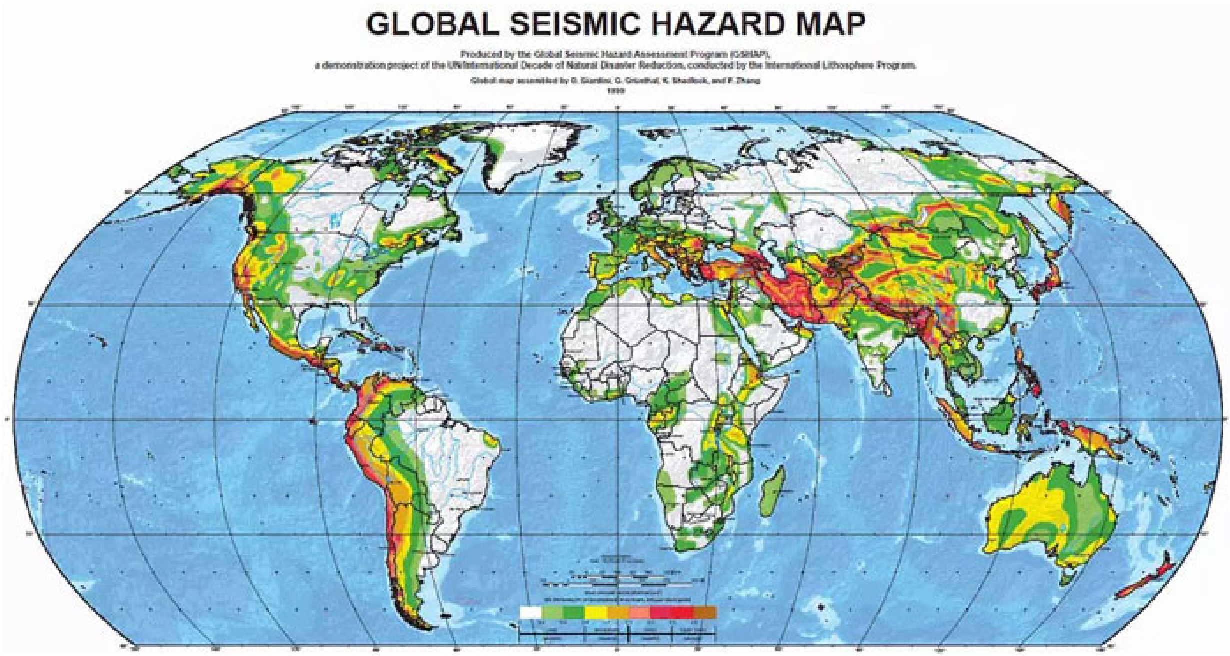


Figure 13.15: *Global seismic hazard map from thoughtco.com*

Stories during the steps of predictive modeling

Storytelling adds significantly at any stage of predictive modeling. Following section will take us through the basic steps of creating a model from the data and stories that can be told within these stages for better understanding.

Data exploration

Let us consider brewing time and caffeine content in the coffee while all other processes remain the same. *Table 13.2* describes the data collected:

Brewing time (in seconds for 8 ml)	Caffeine (in mg)
320	2
340	8
430	16
450	20
480	25
520	38
550	54

Table 13.2: Brewing time and caffeine content

The correlation between brewing time and caffeine quality can be found via Pearson's r value. However, this one parameter will not give much insight into the dataset relationship but only the fact that higher the brewing time, more the caffeine content. *Figure 13.16* is a visual representation of the dependency:

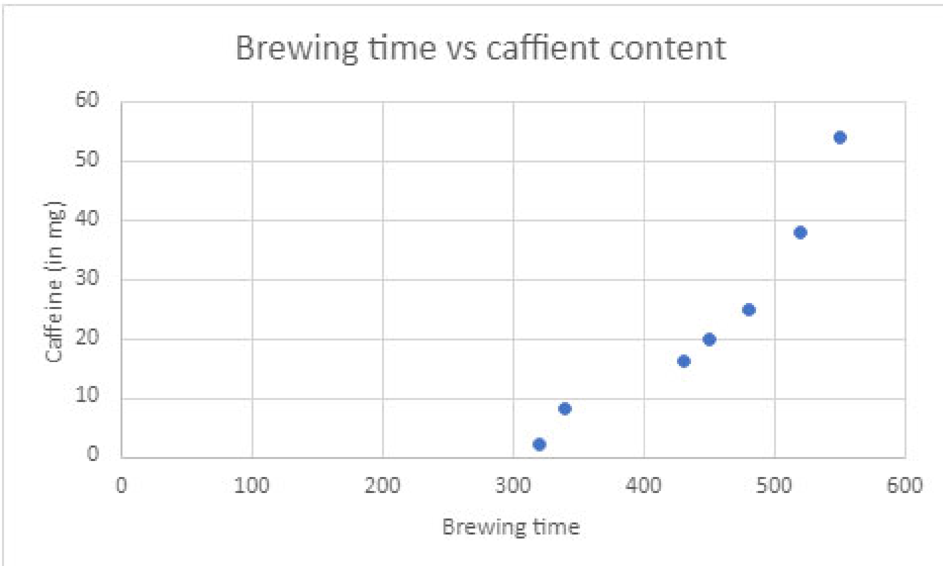


Figure 13.16: *Brewing time vs Caffeine content in 8 ml of the drink*

Feature visualizing

Feature visualization is a powerful tool for getting insights into how neural networks work. Visualization describes how well the model is predicting and how far away the predicted points are from the fitted line. The visuals are far more informative that any statistical data.

For example, the image in *Figure 13.17* is adapted from "Nature, 593, 95-100, Extended Data Figure 4". The figure describes **principal component analysis** (**PCA**). PCA is used to graph the most significant 2 or 3-dimensional projection such that minimal information is lost during the projection of a high-dimensional data set.

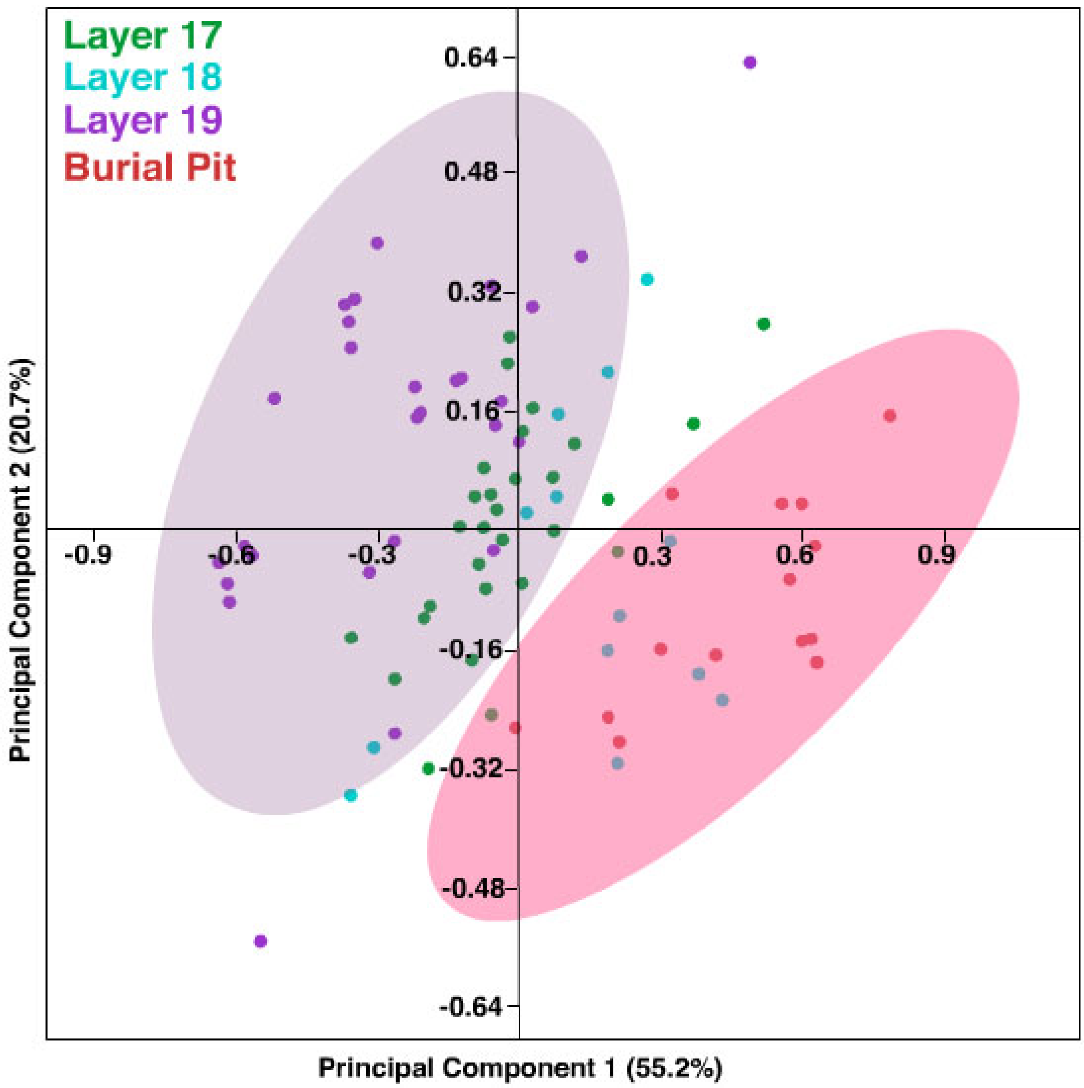


Figure 13.17: *Distinguishing nearby soil samples in a 78,000-year-old burial*

In *Figure 13.17*, 13-dimensional data from 80 soil samples are projected using two principal components. The projection shows samples from the burial pit (red dots) and samples (purple dots) from outside the pit at the same level (Layer 19) of the excavation.

Model creation

Visualization at the model creation stage is used for demonstrating and understanding how the data is being fitted. *Figure 13.18* describes the speed of an airplane based on two of the aerodynamic factors, that is, air thrust and air drag forces:

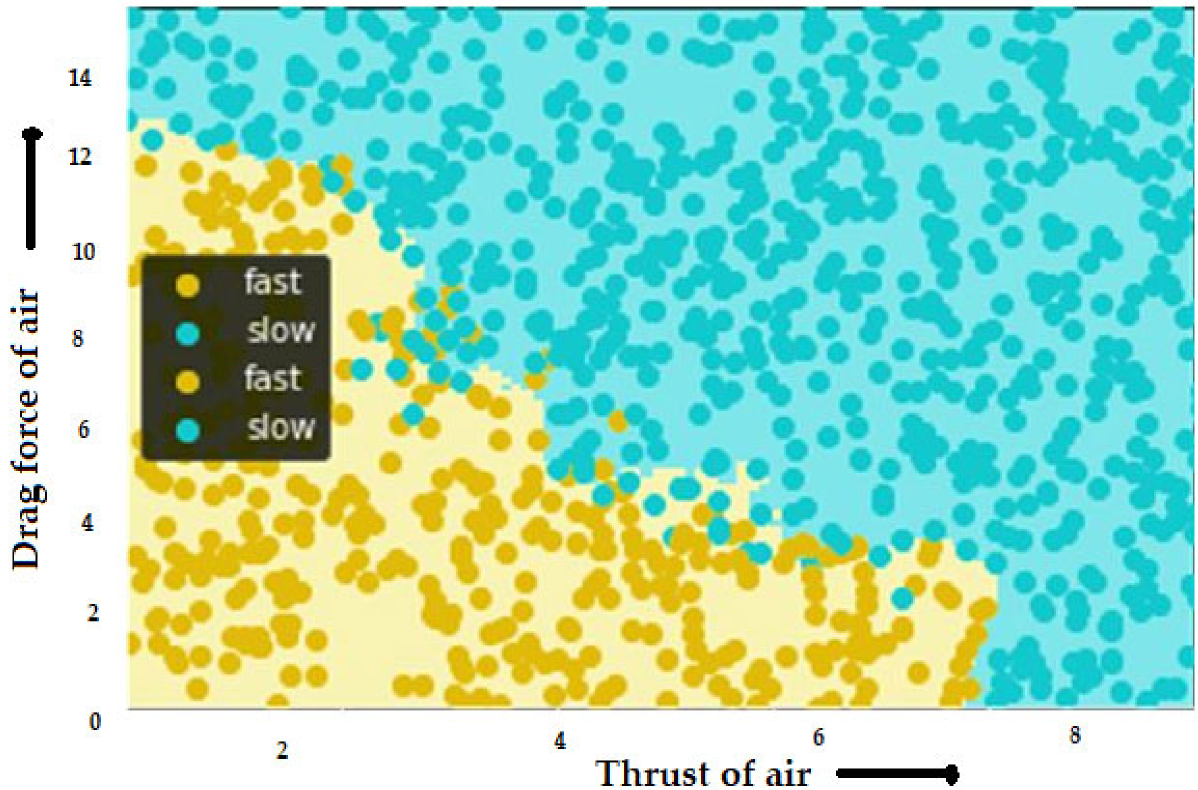


Figure 13.18: *speed of an aeroplane based on two of the aerodynamics factors - air thrust and air drag forces.*

The model predicts the speed of the plane to be maintained. The decision boundary of the aeroplane speed is clear from most of the data. However, the accuracy is < 100% at the boundary with the misclassified points being far from the decision boundary.

With this arises the need to compare algorithms and related techniques by analyzing data at the decision boundaries.

Model comparisons

In real-life applications it is nearly impossible to come across any dataset that is linearly separable. *Figure 13.19* describes an iris dataset based on the sepal length and petal length. The algorithm used is **support vector machine** (**SVM**) that used for both classification and regression challenges. SVMs can work with non-linear datasets. The trick is to find the decision boundary that clearly divides the data points. Functions called **Kernels** are used to achieve this. *Figure 13.19* is based on SVM with the linear kernel.

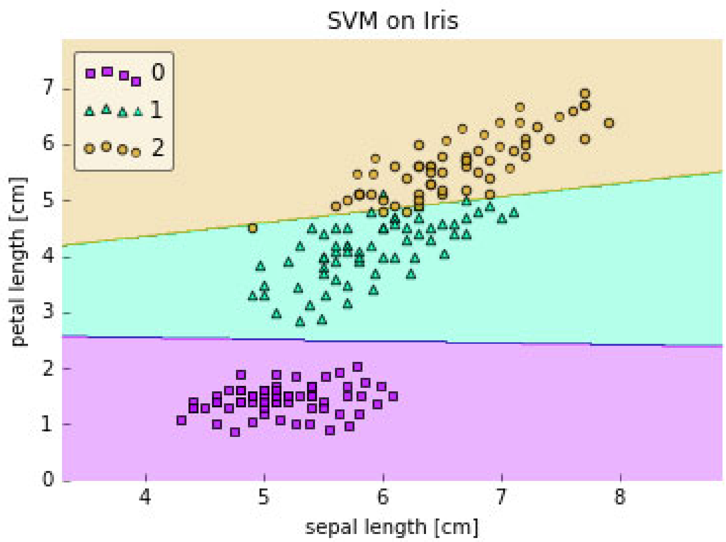


Figure 13.19: *SVM on iris for sepal length vs petal length using a linear kernel*

[*Figure 13.20*](#) If the dataset cannot be graphed using linear SVM, that is, the dataset is linearly inseparable, **radial basis function** (**RBF**) kernel may then be used to represent the dataset:

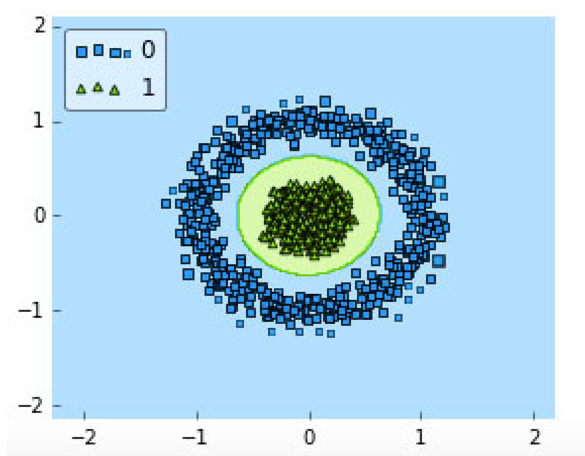


Figure 13.20: *SVM on iris for sepal length vs petal length using RBF kernel*

RBF kernel for SVM, as evident, becomes an obvious choice for very large and non-linear data where the linear kernel will not give accuracy due to linear inseparability.

Benefits of model visualization

Some of the benefits of model visualization are as follows:

- Model visualization helps understand how the algorithms are making decisions.
- Learning the features of a model helps with an understanding of its internal workings. This, in turn, helps in gaining the removal of bias and performance improvement.
- It helps in debugging models.

Best practices of storytelling

Towards the end of this chapter, let us learn best practices while creating a story. The audience tends to get distracted from verbal stories but they perceive visual data instantly if displayed well. Following list a few of the tips and suggestions that are best incorporated in visualization in storytelling:

- Labeled axes and graph headings are a must for instant grasping of visual data.
- Legends must be used wherever necessary.
- Light colors and shades that are soft to the eyes must be used instead of bright and blinding ones.
- Backgrounds that cause visual distractions and may cause a lowering of the readability of the actual significant visual related to the story must be avoided.
- Use connected lines and continuous points for visualization of statistical data based on time series.
- On two or three-dimensional axes, a point can be used to simultaneously encode quantitative values along the axes.

Conclusion

Storytelling is a soft skill and is key at varied stages of the machine learning model life cycle. The visualization not only helps in understanding the data but also helps find the best suitable model for the data.

The next chapter is the last chapter, and it describes the real-world artificial intelligence off-the-shelf applications in use.

Multiple choice questions

Let us now answer the following questions:

1. **Wordcloud is used to represent the frequency of occurrence of a word by-**
    a. Size of the word
    b. color of the word
    c. the horizontal or vertical placement of the word
    d. all of these
2. **Financial data use which of the following techniques for statistical representation?**
    a. FacetGrid
    b. Candlestick
    c. geographic maps
    d. none of these
3. **Stories can be told for explanation and representation of the following stages of data modeling.**
    a. Model creation
    b. model comparison
    c. data exploration
    d. all of these

Answers

1. **a**
2. **b**
3. **d**

Questions

1. Tell a story around marks in a particular subject, say mathematics, across years.

2. Write an essay and create a corresponding Wordcloud for it.
3. Summarize some of the best practices of storytelling.

Join our book's Discord space

Join the book's Discord Workspace for Latest updates, Offers, Tech happenings around the world, New Release and Sessions with the Authors:

https://discord.bpbonline.com

CHAPTER 14
AI Applications in Use in Real-World

Introduction

This chapter will take us through real-world AI applications and systems in use. Few may be in their beta versions, but soon will be available off-the-shelf with a complete list of features.

The companies have been categorized by the industry segment they cater to.

Structure

In this chapter, we will be discussing:

- Self-driving cars
    - Automobile companies
    - Games based on self-driving cars
- Industrial AI
- AI and artificial womb
- Chatbots
    - Travel chatbots
    - Multilingual chatbots
- AI-enabled voice
- AI-enabled video creator
- Content creators
- Storytelling
- Meeting assistant
- Schedule assistant
- Automated website browsing
- Text to image

- Personal lawyer
- Other AI tools
- AI tools database

Self-driving cars

AI tools in self-driving cars strive to make them autonomous and are also used in games involving driverless cars. In this section, we learn about both the categories of AI tools available in the market.

Automobile companies

Self-driving cars can now be seen through the streets of cities across the globe. Autonomous vehicle manufacturers aim to provide riders with a hands-free experience without compromising personal and public safety.

It's been forecasted that by 2040, there will be over 33 million driverless vehicles on the road. **Advanced driver assistance systems** (**ADAS**) deployed in these cars provide:

- Collision avoidance technologies
- Driver aids such as night vision
- Driver alertness
- Adaptive cruise control

A few of the related companies (including manufacturers, technology, and auto parts makers) are as follows:

Motional: **https://motional.com/**

A joint venture between automotive technology *Aptiv* and vehicle manufacturing *Hyundai Motor Group* developing driverless technology.

Magna International: **https://www.magna.com/**

This company provides the following tools and technologies:

- ADAS complete system capability
- Personal park assist
- Highway pilot
- Advanced trailering

- Driver monitoring system
- Occupant monitoring system
- Front integration panels
- Integration and validation of ADAS
- EE Architecture
- Invision™ Adaptive Driving Beam
- ClearView™ Digital Vision System

Cruise: **https://getcruise.com/**

Driverless cars to navigate through cities' roads safely.

Waymo: **https://waymo.com/**

Waymo began as the Google self-driving car project in 2009 (from their website).

Swift Navigation: **https://www.swiftnav.com/**

Provides precise positioning solutions for autonomous. *Figure 14.1* is from the company's website **https://www.swiftnav.com/automotive:**

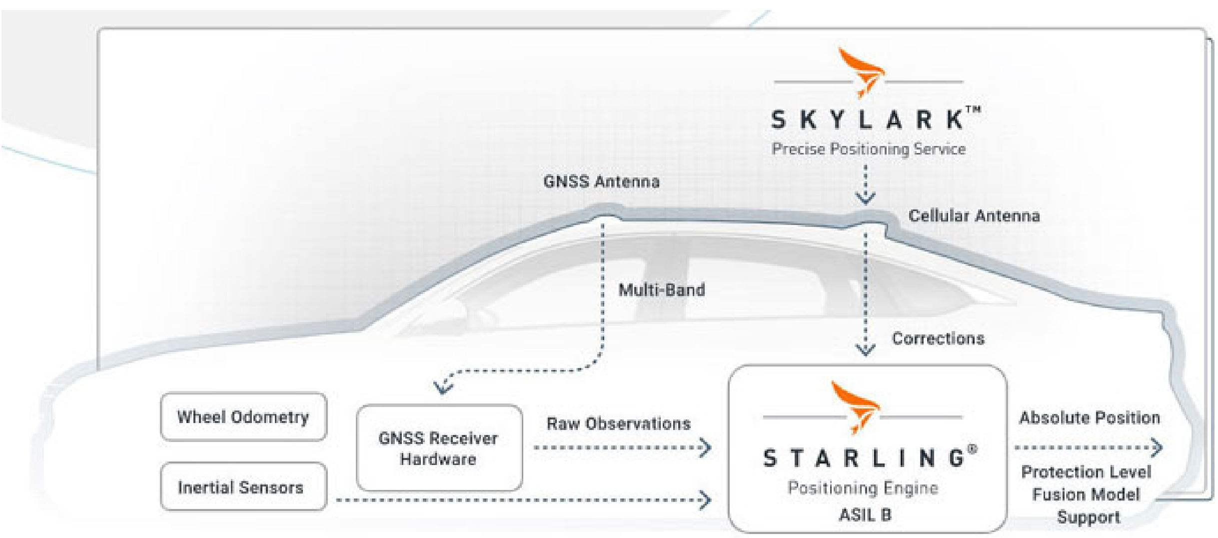


Figure 14.1: *Skylark precise positioning system*

Games based on self-driving cars

These companies are providing a deep neural network that learns to drive in video games. Also, there are companies training real-world cars with video games. Let us consider a few of them in this section.

AWS DeepRacer: **https://aws.amazon.com/deepracer/**

The AWS DeepRacer League is the world's first global autonomous racing league, open to all developers from anywhere in the world.

AWS DeepRacer is based on reinforcement learning and is modeled to make short-term decisions while optimizing its outcomes for a longer-term goal.

Moral Machine: **https://www.moralmachine.net/**

A platform for gathering a human perspective on moral decisions made by machine intelligence, such as self-driving cars.

Industrial AI

BOSCH: **https://www.bosch-ai.com/industrial-ai/**

Bosch uses deep learning techniques and applies AI to various manufacturing processes at Bosch plants, such as:

- Automated optical inspection
- Anomaly detection
- Root cause analysis
- Production scheduling

AI monitored artificial womb

Ectolife: https://www.youtube.com/watch?v=O2RIvJ1U7RE

The World's First Artificial Womb Facility is monitored by AI systems.

Press release: **https://drive.google.com/drive/u/0/folders/14hx7q9J6_imatRlSgVNAxeGc_Wk4Bro2**

Read more about related projects at **https://scienceandstuff.com/ectolife-artificial-wombs/**

Chatbots

AI assistant chatbots are in commercial use in various sectors to provide support to clients. A few of the industries and their chatbots are listed in this section.

Travel chatbots

IndiGo Chat Support: https://www.goindigo.in › support

IndiGo's chat assistant, Dottie is a conversational AI-powered assistant trained to help with frequently asked queries.

Multilingual chatbots

These chatbots help to converse with clients in multiple languages.

NativeChats: https://www.nativechats.com/

Helps overcome language barriers by empowering support users to communicate with the customers in their native language. Refer to the following figure:

Figure 14.2: *From the company's website*

AI-enabled voices

Several startups are using deep learning to simulate synthetic voices which appear like human voices. A few of them are listed in these sections:

Murf: https://murf.ai/

This AI-powered system turns text into a lifelike voice that is used in podcasts, presentations, and videos.

Lovo: https://lovo.ai/

Creates realistic voices for e-learning, marketing, animations, entertainers, explainers, and games.

Altered: **https://www.altered.ai/**

This application enables one to change voice to custom voices that are best suited to be used for professional voice performances.

AI-enabled video creator

These AI-powered systems create videos from the text. Companies with products in this space are listed in this section:

Synthesia: **https://www.synthesia.io/**

Synthesia is an AI video creation platform supported in 120 languages converting plain text to realistic video clips almost instantly.

Supercreator: https://www.supercreator.ai/

AI-powered application for creating short-form videos 10x faster.

Audio transcripts

AssemblyAI: https://www.assemblyai.com/

API exposes AI models for speech recognition, speaker detection, speech summarization, and more.

Content creators

LEX Page: **https://lex.page/**

A Free AI Tool, a word processor, helps to write better and faster and helps escape writer's block.

Craftly: **https://www.craftly.ai/**

A basic AI Copywriter tool with a large library of templates for all possible kinds of documents.

Draft: **https://draft.co/**

It's an AI tool that combines AI writers with real humans for fast, quality content on demand.

CopyAI: **https://www.copy.ai/**

An AI-powered copywriter to generate content. *Figure 14.3* is a snapshot from the website for the services offered for social media managers for a business:

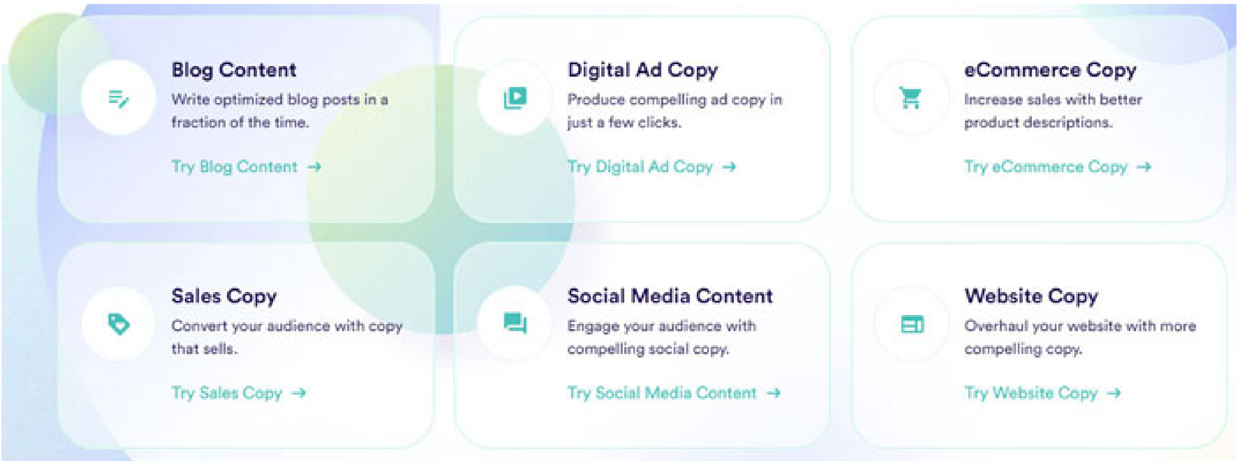


Figure 14.3: *A snapshot from the website*

Texta: https://texta.ai/

AI tool for writing blogs.

Smartwriter: https://www.smartwriter.ai/

AI tool for personalized cold emails that get 8x more replies.

Storytelling

There are AI-powered tools that use interactive storytelling to immerse crowds. This section covers a few such tools commercially available.

Charisma: https://charisma.ai/

Figure 14.4 is a snapshot from the website for the services offered by AI-powered digital humans:

Figure 14.4: A snapshot from the website charisma.ai

Tome: https://beta.tome.app/

Generative storytelling.

Meeting assistant

AI tools have made their way into the meetings too. One such tool is covered in this section:

Sembly: https://www.sembly.ai/

Sembly transcribes, takes meeting notes, and generates insights for your professional meetings.

Schedule assistant

AI tools can help in managing schedules. One such tool is as follows:

timelyAI: https://www.timelyai.com/

timelyAI helps organize the day's schedule via WhatsApp making it as easy as texting.

Automated website browsing

This section captures some unique details about AI-powered tools for automated browsing activities.

Browse: https://www.browse.ai/

Figure 14.5 is a snapshot from the website on the services being offered by *Browse:*

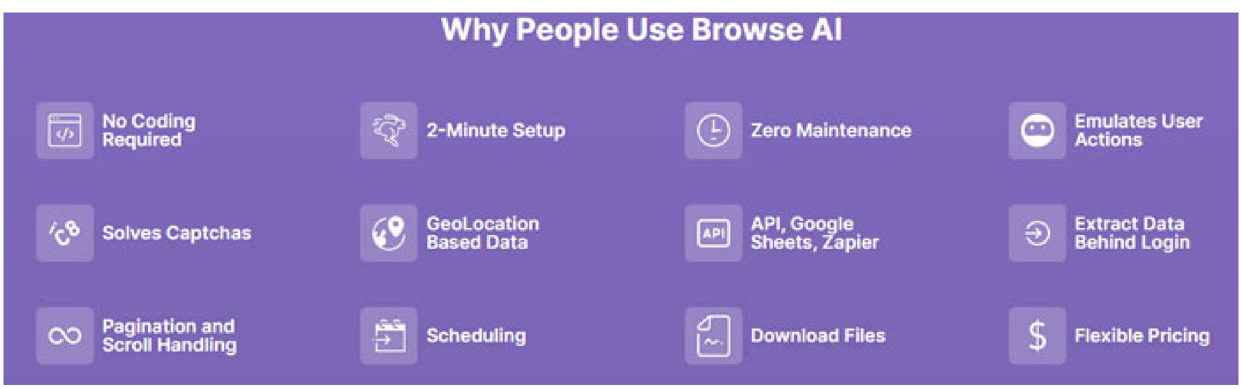


Figure 14.5: *A snapshot from the Browse website*

Figure 14.6 is a snapshot from the website on the supported websites and premium charges:

Which sites does Browse AI support?

Browse AI is the only intelligent web automation software that lets you record and run automations reliably on **any of the 1.8 billion websites out there**.

Some sites (like Twitter or LinkedIn) try to block any automated browsing activity. We have systems in place (such as rotating geolocated residential proxies and automated captcha solving) to avoid these blockers, but their cost is significant so they are marked as Premium.

Tasks on premium sites have a minimum credit cost between 2 and 10. A few examples are:

- Aeroplan.com
- Instagram.com
- LinkedIn.com
- Twitter.com

Figure 14.6: *A snapshot from the website for supported websites*

Text to image

AI tools can help in generating images from input text. This section lists a few of such tools as follows:

DALL.E -2: https://openai.com/dall-e-2/

- DALL-E2 provides better resolution and greater realism than DALL-E. DALL.E can be accessed at **https://labs.openai.com/** DALL-E is a GPT-3 AI for images and generates images as per text requests. This artificial intelligence tool creates original images and has started its beta testing.
- DALL.E-2 is more than just an image generator.

Figure 14.7 describes one of the images generated by the tool (as copied from the website **https://openai.com/dall-e-2/**):

Figure 14.7: *An astronaut riding a horse*

Flair: https://flair.ai/

AI tool for generating branded content. *Figure 14.8* is a snapshot from the website:

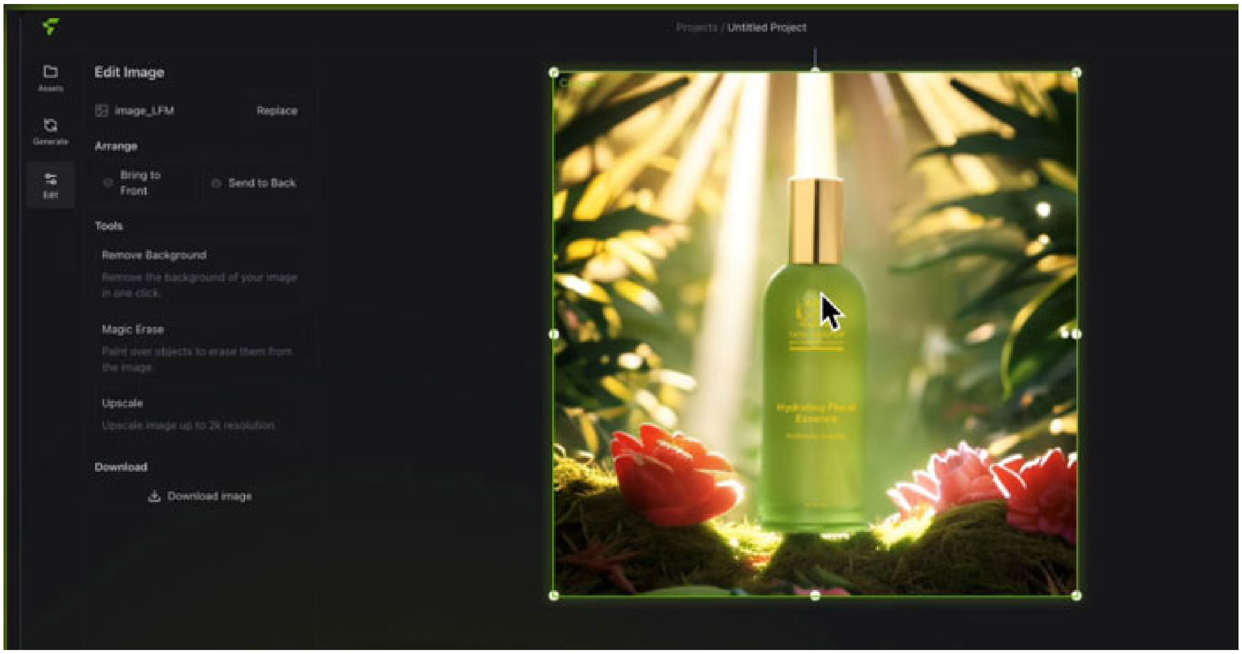


Figure 14.8: *A snapshot from the website.*

Personal lawyer

AI tools have also been designed to assist in courts of law. Let us know about one such AI tool in this section:

DoNotPay: https://donotpay.com/

The DoNotPay AI tool is the world's first robot lawyer. It helps fight corporations manage bureaucracy and files lawsuits at the press of a button.

Figure 14.9 is a snapshot from the website:

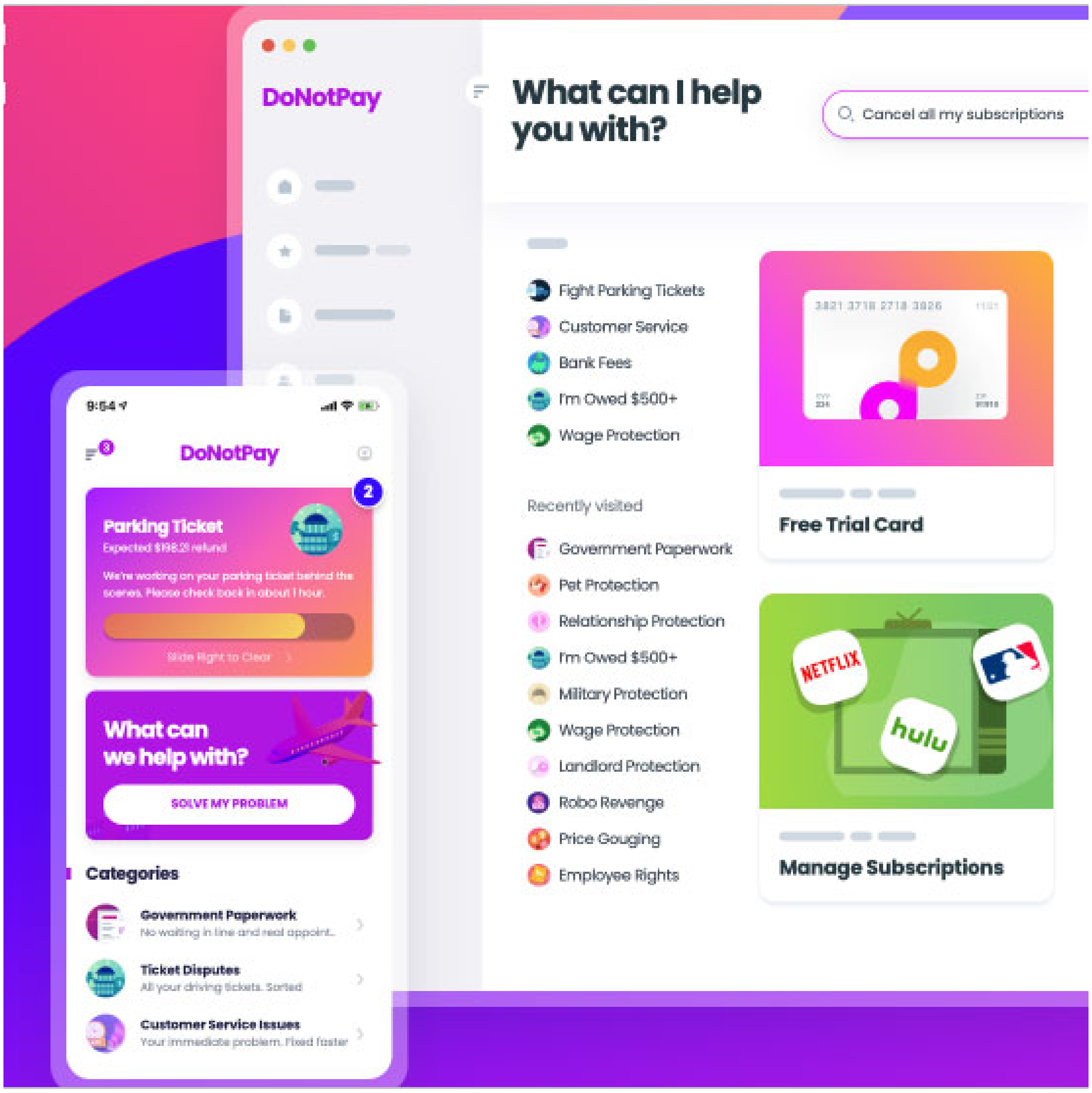


Figure 14.9: *A snapshot from the website describing service offerings*

Other AI tools

Some other AI tools have been listed as follows:

PrismaAI: https://prisma-ai.com/

Create AI Avatar by turning photos into art. This tool also enhances the photographs.

Krisp: https://krisp.ai/

Krisp's AI tool is specifically used in business calls. The tool helps remove background voices, noises, and echoes from the calls. *Figure 14.10* is a snapshot from the website on the features of the tools:

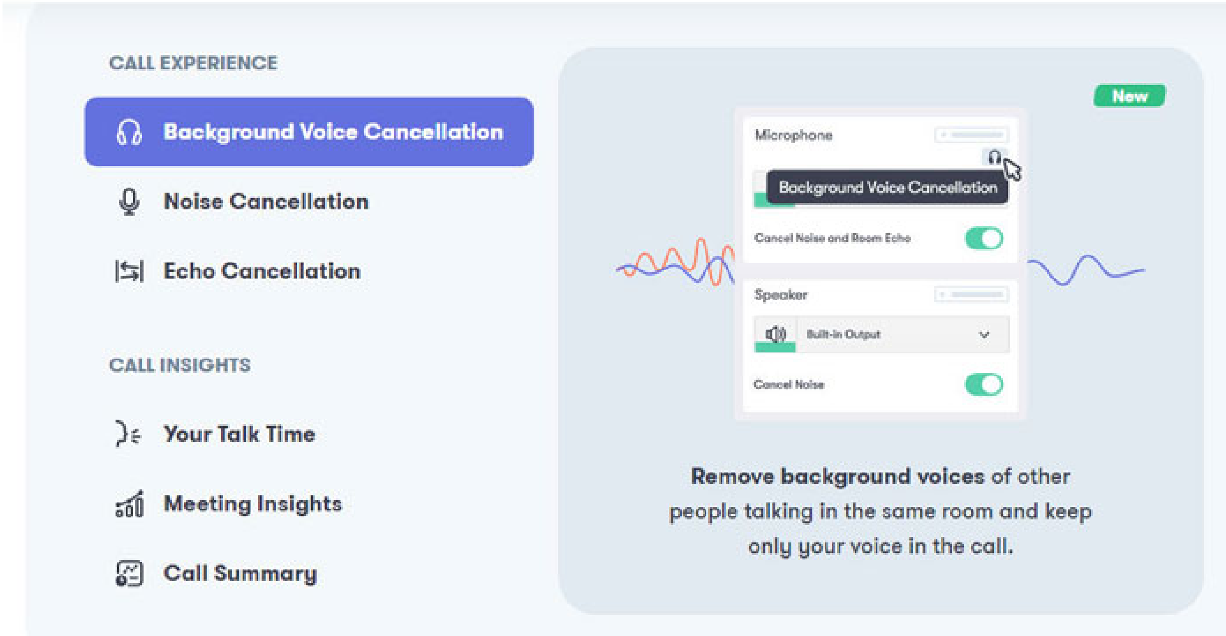


Figure 14.10: *A snapshot of the website on the features of the tools*

Beatoven: https://www.beatoven.ai/

This AI tool helps create royalty-free music.

Cleanvoice: https://cleanvoice.ai/

Automatically edit your podcast episodes. *Figure 14.11* is a snapshot of the website for various use cases supported by them too:

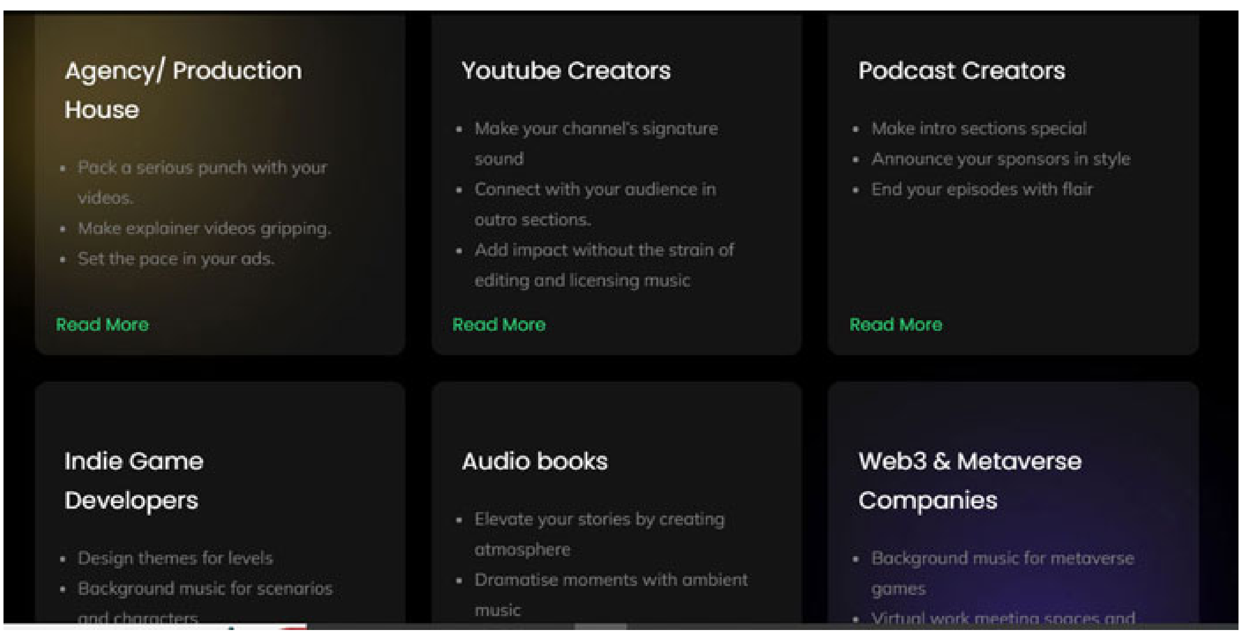


Figure 14.11: *A snapshot from the website for various use cases supported*

Podcastle: https://podcastle.ai/

This tool offers studio-quality recording. *Figure 14.12* is a snapshot from the website for various services offerings:

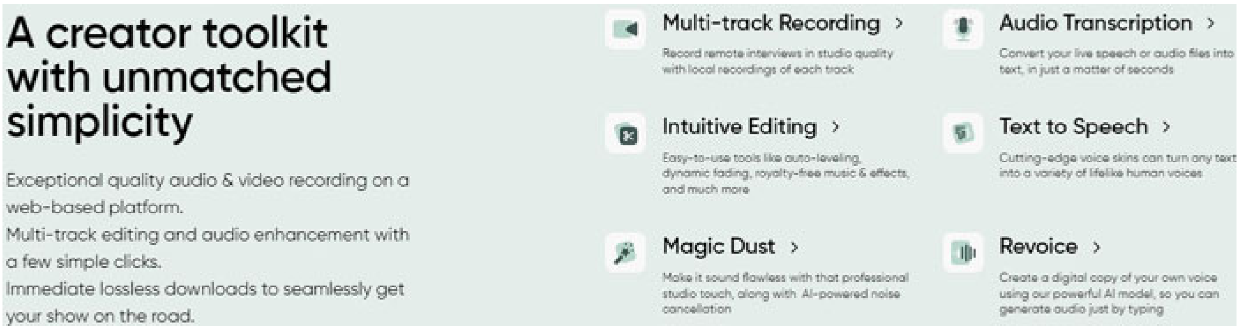


Figure 14.12: *A snapshot from the website for various services offerings*

Resumeworded: https://www.resumeworded.com/

This tool is designed to improve resume and LinkedIn profile. *Figure 14.13* is a snapshot from the website for various services offerings:

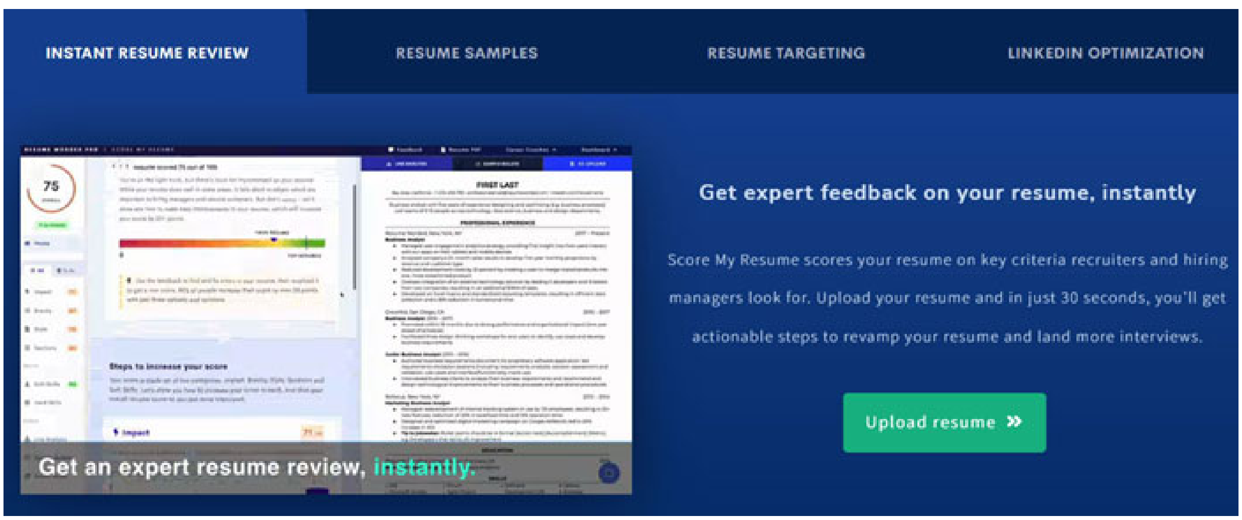


Figure 14.13: *A snapshot from the website for various services offerings*

CopyMonkey: http://copymonkey.ai/

This tool helps with the creation of Amazon listings. *Figure 14.14* is a snapshot from the website:

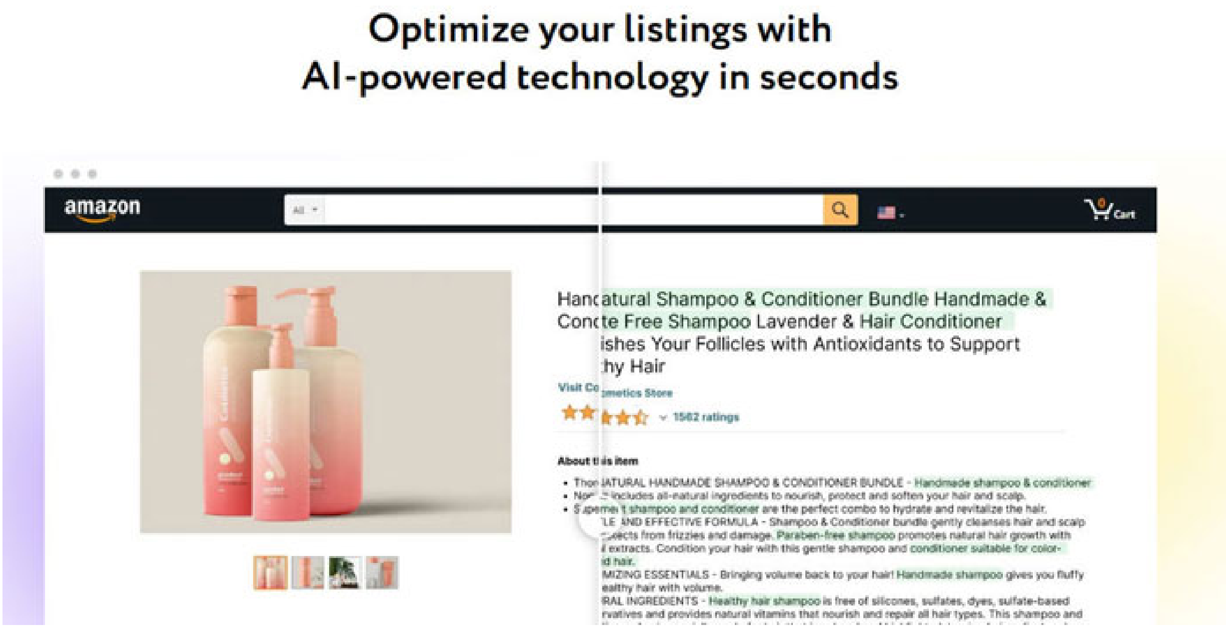


Figure 14.14: *A snapshot from the website*

Patterned: https://www.patterned.ai/

This AI tool helps to generate the exact patterns and designs needed. [*Figure 14.15*]() is a snapshot from the website:

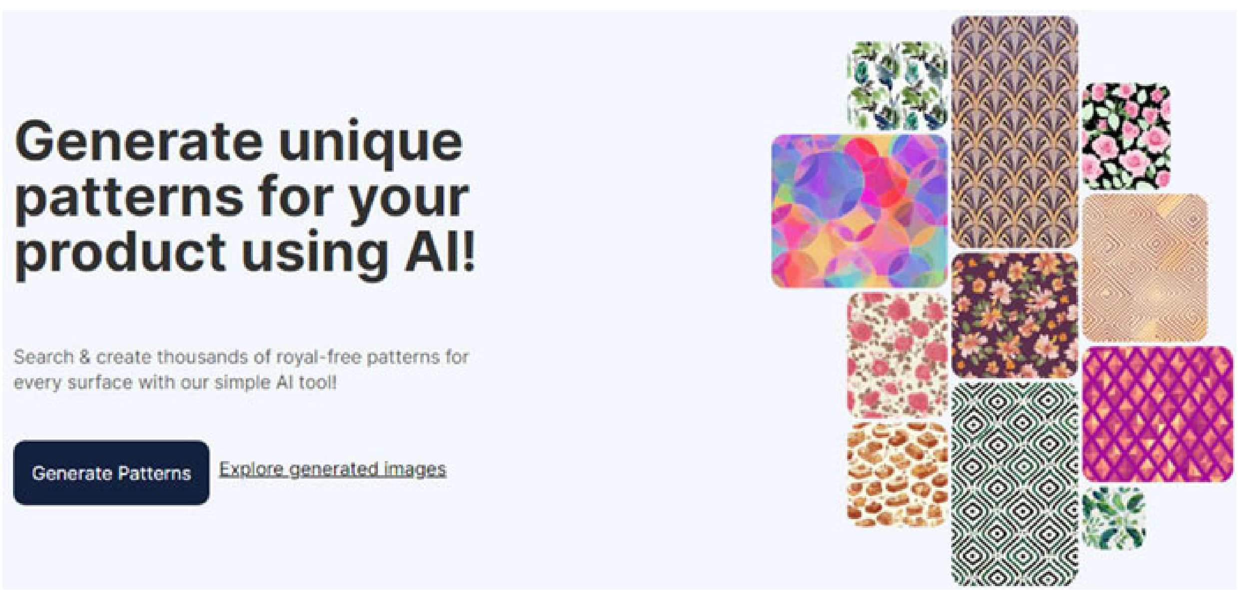


Figure 14.15: *A snapshot from the website*

Puzzle: https://www.puzzlelabs.ai/

These AI-powered tools help build a knowledge base for the team and customers. [*Figure 14.16*]() is a snapshot from the website:

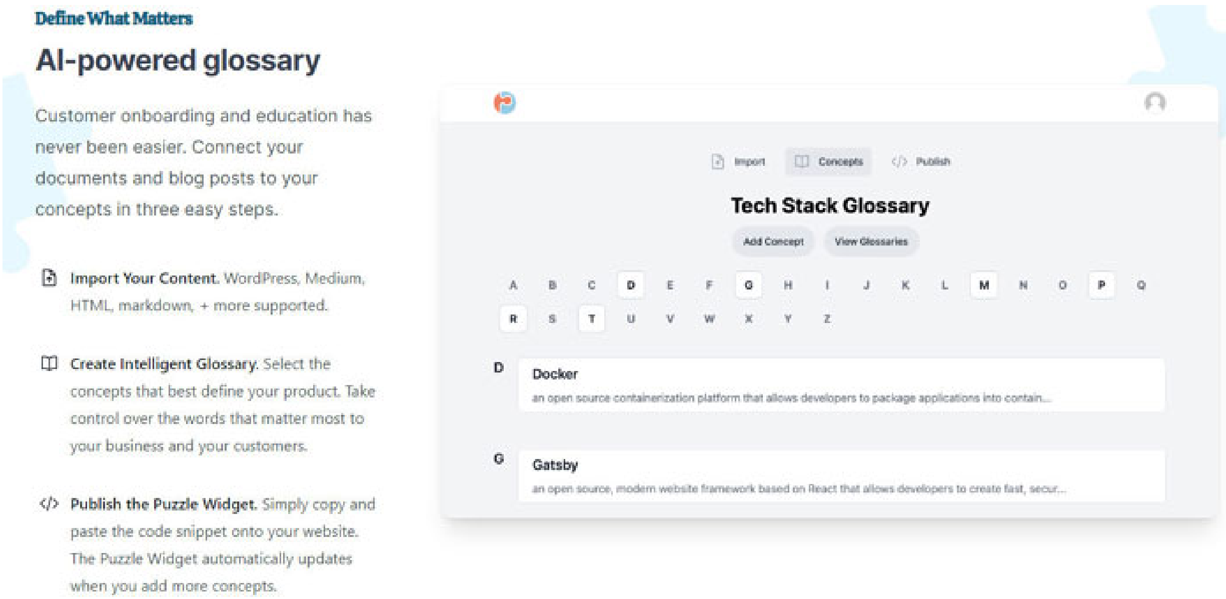


Figure 14.16: A snapshot from the website

PRAGMA: https://www.pragma.ai/

This AI-powered tool helps in the sales and support processes of any organization. *Figure 14.17* is a snapshot from the website:

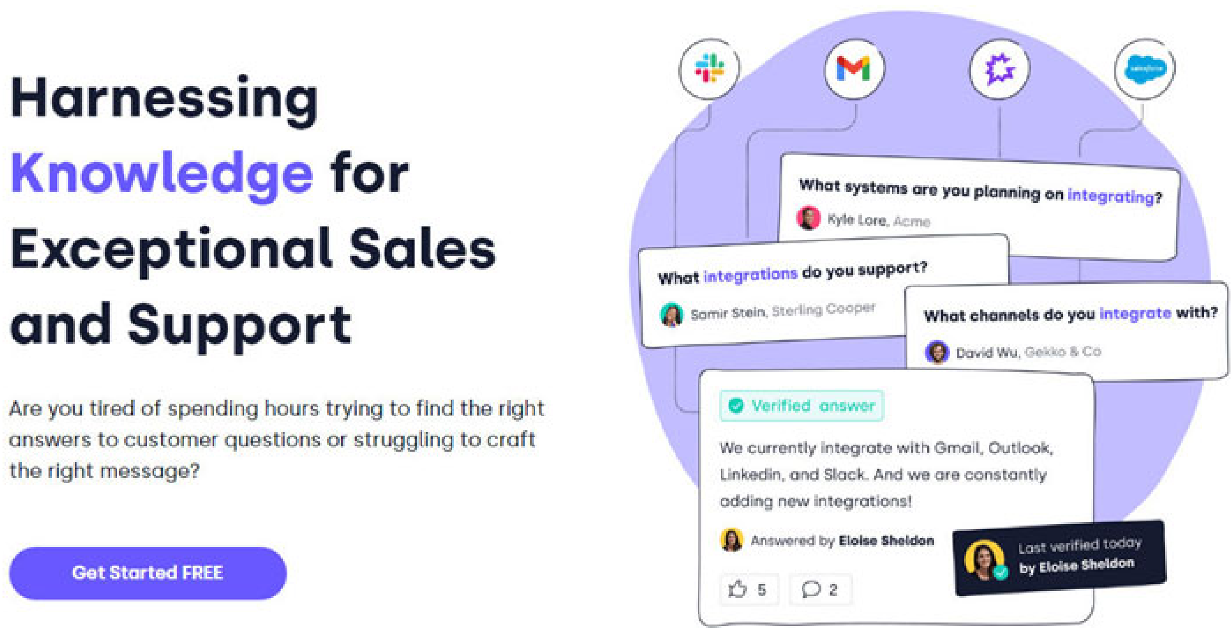


Figure 14.17: A snapshot from the website

[AI tools database]()

This section lists one of the websites that is a database of AI tools for all possible tasks, mentioned as follows:

theresanaiforthat: **https://theresanaiforthat.com/**

This website is updated daily and provides a comprehensive database of AIs available for every task. *Figure 14.18* is a snapshot from the website:

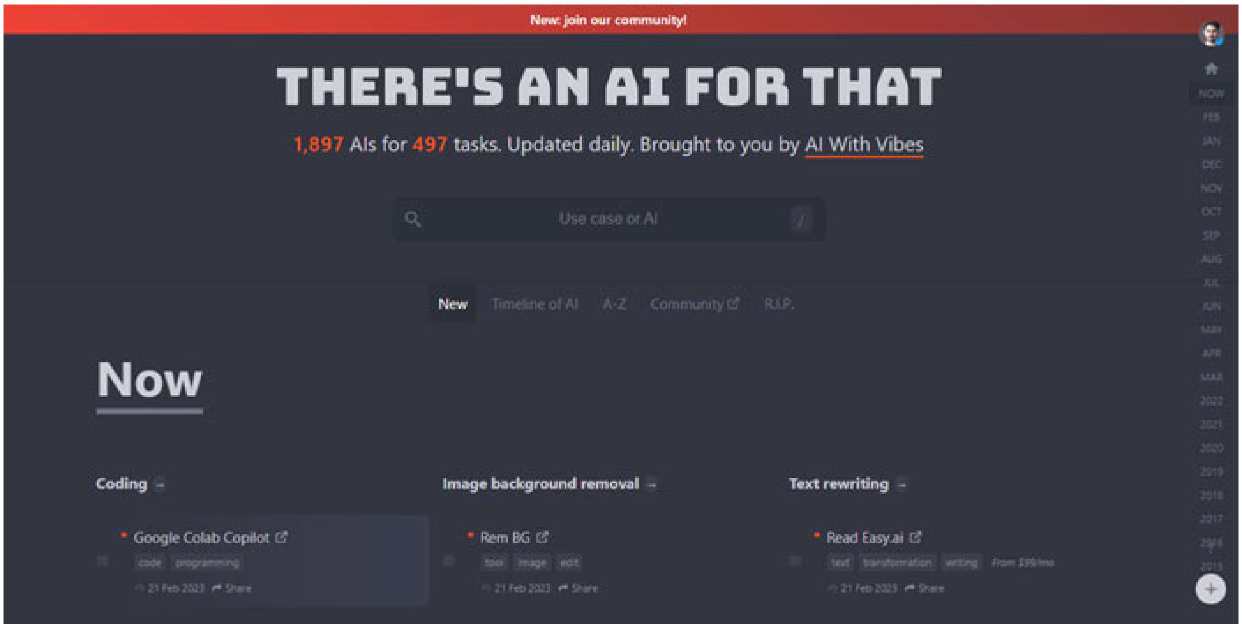


Figure 14.18: *A snapshot from the website*

Conclusion

The list of companies venturing into AI product development and deployment is endless. AI is making its way in each industry vertical. There are no code AI development platforms available, making it easy to develop AI tools. When are you developing yours?

Join our book's Discord space

Join the book's Discord Workspace for Latest updates, Offers, Tech happenings around the world, New Release and Sessions with the Authors:

https://discord.bpbonline.com

Index

Symbols

A

B

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

Z