

DATA SCIENCE

(As per the Latest University Syllabus)

Dr. V. N. RAJAVARMAN

Professor and Dean, Parttime studies
Additional Dean, Computer Studies
Dr. M.G.R. Educational and Research Institute
(Deemed to be University), Chennai, India.

Mrs. R. GAYATHRI

Assistant Professor & Research Scholar
Dr. M.G.R. Educational and Research Institute
(Deemed to be University), Chennai, India.

Mrs. R. HEMAVATHI

Assistant Professor & Research Scholar
Dr. M.G.R. Educational and Research Institute
(Deemed to be University), Chennai, India.



www.magesticts.com

Data Science

(As per the Latest University Syllabus)

Authors:

Dr. V. N. Rajavarman

Mrs. R. Gayathri

Mrs. R. Hemavathi

@ All rights reserved with the publisher.

First Published: December 2022

ISBN 978-93-92090-09-7



ISBN: 978-93-92090-09-7

DOI: <https://doi.org/10.47716/MTS.B.978-93-92090-09-7>

Pages: 214 (Front pages 10 & Inner pages 204)

Price: 360/-

Publisher & Imprint:

Magestic Technology Solutions (P) Ltd.

Chennai, Tamil Nadu, India.

www.magesticts.com

E-Mail: info@magesticts.com

TITLE VERSO

Title of the Book:

Data Science

Author's Name:

Dr. V. N. Rajavarman

Mrs. R.Gayathri

Mrs R. Hemavathi

Published By:

Magestic Technology Solutions (P) Ltd.

Publisher's Address:

544, Anna Street, Kathivedu, Chennai 600 066

Tamil Nadu

Printer's Details:

Magestic Technology Solutions (P) Ltd.

Edition Details: First Edition

ISBN: 978-93-92090-09-7

Copyright: @ Magestic Technology Solutions (P) Ltd.

COPYRIGHT

Magestic Technology Solutions (P) Ltd

544, Anna Street, Kathirvedu
Chennai 600 066. Tamil Nadu. India

@ 2022, Magestic Technology Solutions (P) Ltd
Imprint Magestic Technology Solutions (P) Ltd

Printed on acid-free paper

International Standard Book Number (ISBN): 978-93-92090-09-7
(Paperback)

Digital Object Identifier (DOI): 10.47716/MTS.B. 978-93-92090-09-7.

This book provides information obtained from reliable and authoritative sources. The author and publisher have made reasonable attempts to publish accurate facts and information, but they cannot be held accountable for any content's accuracy or usage. The writers and publishers have endeavoured to track down the copyright holders of every content copied in this book and regret if permission to publish in this format was not acquired. Please notify us through email if any copyright-protected work has not been recognised so that we may make the necessary corrections in future reprints. No portion of this book may be reprinted, reproduced, transmitted, or used in any form by any electronic, mechanical, or other means, now known or hereafter developed, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without the publisher's written permission.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks and are used only for identification and explanation without intent to infringe.

Visit the Magestic Technology Solutions (P) Ltd

Web site at

<http://www.magesticts.com>

DEDICATION



Er. A. C. S. ARUN KUMAR

B.Tech (Hons), LMISTE, MIET (UK), LMCSI

President

**Dr. M. G. R. Educational Research Institute
Chennai, T.N. India**

We thank the Almighty God for giving us the wonderful opportunity to write this book.

We express our heartfelt thanks and gratitude to our honorable President Er. A. C. S. Arun Kumar, Dr. M. G. R. Educational Research Institute, Chennai for his incessant support and motivation for us to write this book.

We convey our sincere thanks to Management, Executives, Staff & Students, Dr.M.G.R Educational Research Institute, for their valuable support and encouragement.

We also thank our parents & family members, who made our work a success.

Authors

Dr. V. N. Rajavarman

Mrs. R.Gayathri

Mrs R. Hemavathi

This Page Intentionally Left Blank

PREFACE

In the field of study known as "data science," the goal is to glean useful information from massive volumes of data by using a wide variety of scientific approaches, algorithmic procedures, and other procedures. Discovering hidden patterns in the raw data is much easier with its assistance. As a result of developments in mathematical statistics, data analysis, and big data, a new field known as "data science" has come into existence.

The discipline of Data Science is an interdisciplinary one that enables one to derive information from either organised or unstructured data. The field of data science gives you the ability to turn an issue with your company into a research project and then turn that project into a solution for real-world problems. When it comes to the field of data science, we need some kind of programming language or instrument, such as Python. In spite of the fact that there are more tools for data science, such as R and SAS, the primary emphasis of this post will be on Python and how it may be advantageous for data science.

In recent years, Python has emerged as a dominant force in the world of programming languages. Its incorporation into data science, the internet of things, artificial intelligence, and other technological fields has contributed to the rise in its popularity.

Python is a popular choice for usage as a programming language for data science due to the fact that it provides access to a variety of powerful mathematical and statistical tools. Python is used by data scientists all around the globe, and this is a big reason why Python is used. If you've been paying attention to industry developments over the last few years, you've probably seen that Python has emerged as the dominant programming language, especially in the data science sector.

This book has five units that cover the whole of the university curriculum for Data Science. Beginning with an introduction, the courses address describing data, using Python for data handling, describing relationships, and using Python for data visualisations respectively. Students of Computer Science, Engineering and Technology, and Computer Applications from all of India's universities are the intended audience for this book, which was developed just for them.

- **Authors**

Dr. V. N. Rajavarman

Mrs. R. Gayathri

Mrs R. Hemavathi

SYLLABUS

Data Science

UNIT I INTRODUCTION

Need For Data Science – Benefits and Uses – Facets of Data – Data Science Process – Setting the Research Goal – Retrieving Data – Cleansing, Integrating, And Transforming Data – Exploratory Data Analysis – Build the Models – Presenting and Building Applications

UNIT II DESCRIBING DATA

Frequency Distributions – Outliers – Relative Frequency Distributions – Cumulative Frequency Distributions – Frequency Distributions for Nominal Data – Interpreting Distributions – Mode – Median – Mean – Averages for Qualitative and Ranked Data – Describing Variability – Range – Variance – Standard Deviation – Degrees of Freedom – Interquartile Range

UNIT III PYTHON FOR DATA HANDLING

Basics Of Numpy Arrays – Aggregations – Computations on Arrays – Comparisons, Masks, Boolean Logic – Fancy Indexing – Structured Arrays – Data Manipulation with Pandas – Data Indexing and Selection – Operating on Data – Missing Data – Hierarchical Indexing – Combining Datasets – Aggregation and Grouping – Pivot Tables

UNIT IV DESCRIBING RELATIONSHIPS

Correlation – Scatter Plots – Correlation Coefficient for Quantitative Data – Computational Formula for Correlation Coefficient – Regression – Regression Line – Least Squares Regression Line – Standard Error of Estimate – Interpretation of R^2 – Multiple Regression Equations – Regression Towards the Mean

UNIT V PYTHON FOR DATA VISUALIZATION

Visualization With Matplotlib – Line Plots – Scatter Plots – Visualizing Errors – Density and Contour Plots – Histograms, Binnings and Density – Three-Dimensional Plotting – Geographic Data – Data Analysis Using Pandas and Seaborn – Graph Plotting Using Plotly – Interactive Data Visualization Using Bokeh

TABLE OF CONTENTS

UNIT - I

Introduction	1
Need for data science	3
Benefits and uses	4
Facets of data	6
Data science process	11
Setting the research goal	18
Retrieving data	20
Cleansing, integrating, and transforming data	22
Exploratory data analysis	30
Build the models	35
Presenting and building applications	37

UNIT - II

Frequency Distributions	41
Outliers	43
Relative Frequency Distributions	49
Cumulative Frequency Distributions	50
Frequency Distributions for Nominal Data	52
Interpreting Distributions	57
Mode – Median – Mean	59
Averages for Qualitative and Ranked Data	61
Describing Variability	63
Range – Variance – Standard Deviation	64
Degrees of Freedom – Interquartile Range	66

UNIT - III

Basics of Numpy Arrays	75
Computations on Arrays	80
Comparisons, Masks, Boolean Logic	85
Fancy Indexing	93
Structured Arrays	94
Data Manipulation with Pandas	96

Data Indexing and Selection	101
Operating on Data	106
Missing Data	109
Hierarchical Indexing	112
Combining Datasets	114
Aggregation and Grouping	120
Pivot Tables	125

UNIT - IV

Correlation	131
Scatter Plots	134
Correlation Coefficient for Quantitative Data	136
Computational Formula for Correlation Coefficient	138
Regression	142
Regression Line	142
Least Squares Regression Line	143
Standard Error of Estimate	143
Interpretation of R^2	145
Multiple Regression Equations	149
Regression Towards the Mean	150

UNIT - V

Visualization With Matplotlib	155
Line Plots	160
Scatter Plots	165
Visualizing Errors	170
Density and Contour Plots	174
Histograms, Binnings and Density	178
Three-Dimensional Plotting	185
Geographic Data	187
Data Analysis Using Pandas and Seaborn	194
Graph Plotting Using Plotly	196
Interactive Data Visualization Using Bokeh	197

Bibliography

Bibliography & Webliography	201
-----------------------------	-----

This Page Intentionally Left Blank

UNIT - I
INTRODUCTION TO
DATA SCIENCE

This Page Intentionally Left Blank

NEED FOR DATA SCIENCE

The field of data science has two facets: first, it investigates and analyses vast quantities of data; second, it has branches in almost every discipline.

The data we deal with are not simple; they are complicated and organised in several levels. The fundamental components of data science include statistics, mathematics, and computer language.

Artificial intelligence encompasses the principles of all three disciplines and serves as the data science's machinery or brain. Data science utilises concepts, processes, algorithms, rules, and tools from all three subfields and functions as a cohesive mechanism to tackle the complex issues that develop in our environment.

What is the driving force behind data science?

As implied by its name, data science is centred on data. Data is a unit of knowledge-containing information, while data science employs mathematical algorithms, rules, and artificial intelligence to gather, refine, align, store, manipulate, and utilise data. The objective is to execute result-driven computations on the data to provide commercial and research-relevant insights. (Blei and Smyth, 2017)

Why is Data Science Needed?

Data plays an essential role in every aspect of a community, from the business sector to the health industry, from science to everyday life, from marketing to research. Technology has taken over our lives, and it is evolving at such a rate and with such variety that the operating methods used a few years ago are now obsolete. The same holds true for obstacles and issues. The difficulties and worries of the past about a certain topic, ailment, or deficiency may no longer exist due to their increased complexity.

Therefore, every area of research and study, as well as every company, need an updated set of operational systems and technology to meet the challenges of today and future and to find answers to unresolved problems.

BENEFITS AND USES

The field of data science has become an integral part of every industry today. By converting company data into assets, companies can increase revenue, cut costs, capture business opportunities, and enhance customer satisfaction, among other benefits.

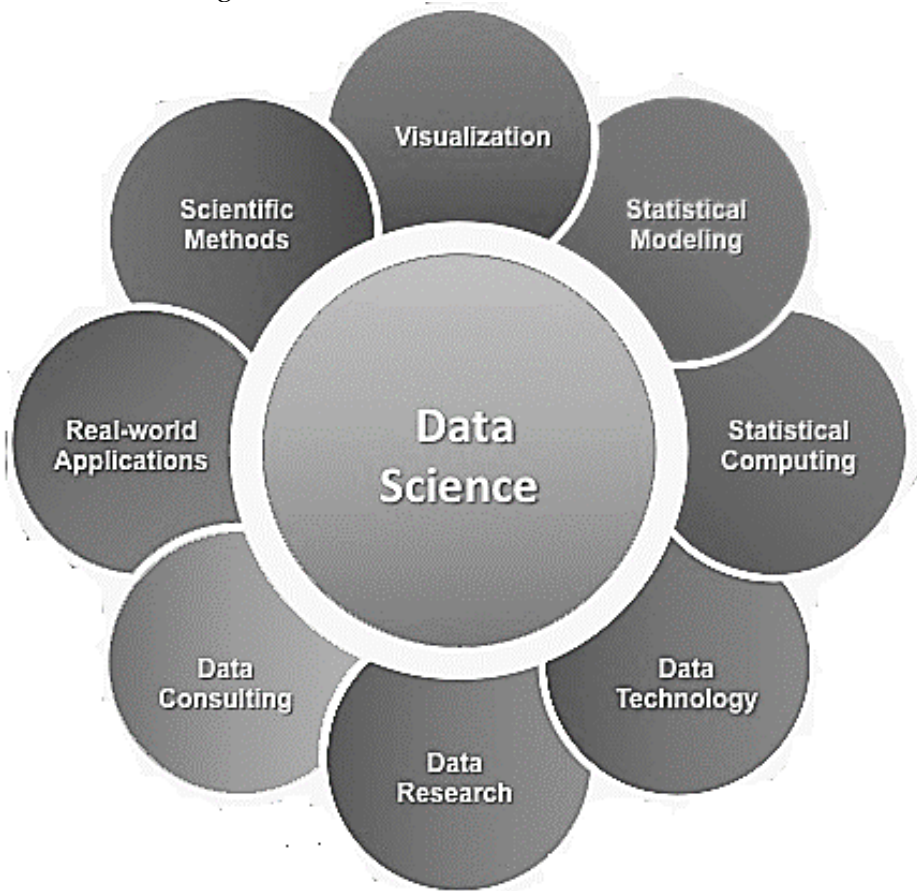


Fig. 1.1 Uses of Data Science

Data science is now one of the most contested subjects in industry. Companies have been applying data science strategies to enhance their businesses and improve consumer satisfaction as its popularity has increased over time. The field of data science aims to discover previously

unknown patterns in data, extract valuable information, and make business decisions by utilizing current technologies and methodologies.

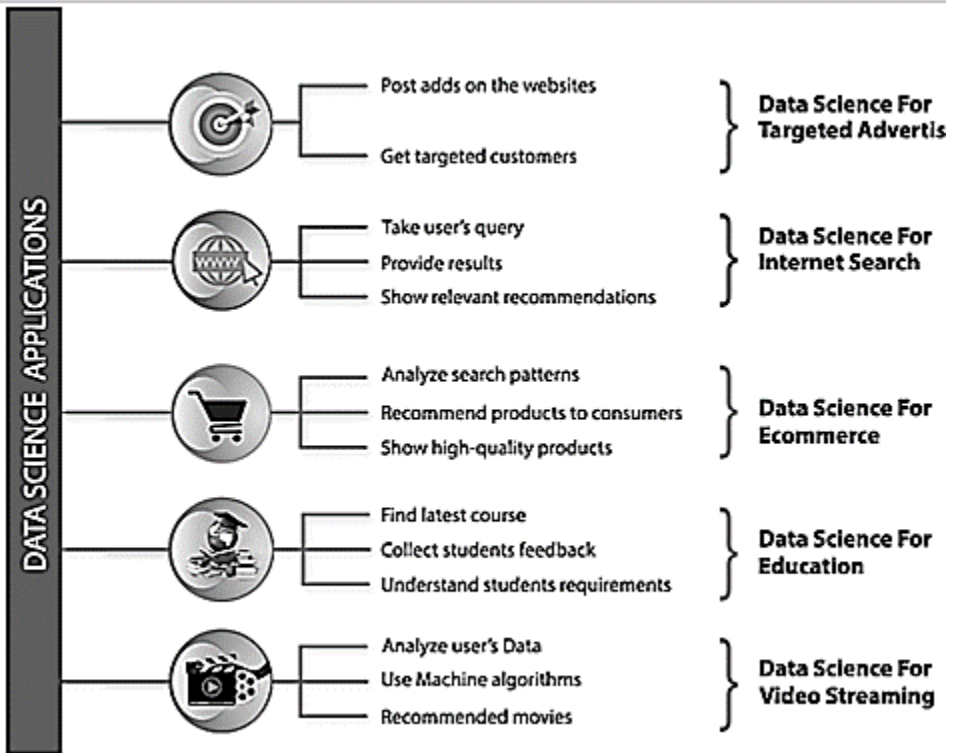


Fig. 1.2 Applications of Data Science

In the modern world, data is being generated at an alarming rate. There is a great deal of data created every second, whether it is from users of Facebook or any other social networking site, from calls that are made, or from other organizations. Consequently, the discipline of Data Science is enhanced in a variety of ways due to this large quantity of data. Several of the benefits are listed below:

Due to its high demand, it has created a significant number of employment opportunities in its many fields. Some examples of these positions include Data Scientist, Data Analyst, Research Analyst, Business Analyst, Analytics Manager, Big Data Engineer, etc.

Through data science, businesses can determine which items sell the best and when to supply them, ensuring that they are always available at the most appropriate time and location. In order to increase efficiency and revenues, the organisation makes choices more quickly and with greater quality.

As Data Scientists remain one of the most desirable occupations, they also enjoy high salaries. Data Scientists earn an average of \$106,000 a year, according to a Dice Salary Survey.

It has made it relatively simpler to sift data and search for the finest individuals for a business. Big Data and data mining have facilitated the recruiting teams' processing and selection of resumes, aptitude tests, and games.

FACETS OF DATA

Data Science and Big Data involve a wide variety of data types, each of which requires a unique set of tools. These are the principal types of data:

- Structured
- Unstructured
- Natural Language
- Machine-Generated
- Graph-based
- Audio, video, and photographs
- Streaming

Explore each of these fascinating data types.

Structured data

In structured data, information is contained within the fixed fields of a record and is determined by a data model. Structured data is frequently stored in tables within databases or in Excel documents. SQL, or Structured Query Language, is the primary method for managing and querying database data. Further, you may encounter complex data that is difficult to store in a relational database. Information that is organized hierarchically, such as a family tree, is an example.

Structured data is imposed on the world by people and technology; it does not exist naturally.

	A	B	C	D	E	F	G	H	I
	Sales								Total Sale
1	Representative	Location	Region	Customer	Order Date	Item	Quantity	Price	Amount
2	Sara Snyder	New York	East	Phyllis Johnston	2016-10-30	Things	1	17.83	17.83
3	Sara Snyder	New York	East	Kimberly Little	2016-05-23	Junk	3	12.42	37.26
4	Frances Warren	Massachusetts	East	Justin Dixon	2016-09-27	Widgets	4	53.35	213.40
5	Sara Snyder	Massachusetts	East	Shirley Rivera	2016-02-12	Junk	5	12.42	62.10
6	Diane Gonzalez	Oregon	West	Marilyn Franklin	2016-02-14	Things	8	17.83	142.64
7	Patrick Graham	Washington	West	Henry Sanders	2016-04-11	Widgets	4	53.35	213.40
8	Sara Snyder	Connecticut	East	Benjamin Phillips	2016-09-02	Junk	4	12.42	49.68
9	Frances Warren	New Jersey	East	Theresa Torres	2016-11-26	Junk	4	12.42	49.68
10	Patrick Graham	Oregon	West	Roger Bell	2016-07-13	Junk	10	12.42	124.20
11	Sara Snyder	New Jersey	East	Harold Matthews	2016-06-02	Junk	3	12.42	37.26
12	Frances Warren	New York	East	Roy Young	2016-06-02	Widgets	8	53.35	426.80
13	Sara Snyder	New York	East	Debra Allen	2016-02-20	Things	1	17.83	17.83
14	Randy Watson	Connecticut	East	Alan Dean	2016-06-07	Junk	7	12.42	86.94
15	Randy Watson	Massachusetts	East	Robin Matthews	2016-10-31	Stuff	5	16.32	81.60
16	Randy Watson	New York	East	Randy Burton	2016-03-13	Stuff	4	16.32	65.28
17	Patrick Graham	Washington	West	Terry Nguyen	2016-02-10	Widgets	10	53.35	533.50

Fig. 1.3 Example of Structured Data

Unstructured data

It is difficult to incorporate unstructured data into a data model since it is context-specific or variable. Email correspondence is one example of unstructured data. Despite the fact that email has organised features, such as the sender, subject, and body text, it is difficult to determine the number of individuals who have sent an email complaint regarding an employee since a person can be referred to in a variety of ways. This is further complicated by the tens of thousands of distinct languages and dialects that exist.

An email composed by a person also exemplifies natural language data perfectly.

```

weblogic.application.utils.StateMachineDriver.nextState(StateMachineDriver.java:26)
####<Dec 29, 2006 2:14:24 PM IST> <Notice> <Log Management> <svaidyan02> <xbusServer>
<[ACTIVE] ExecuteThread: 0' for queue: 'weblogic.kernel.Default (self-tuning)''> <<wls
kernel>> <<> <<1167381864275> <BEA-170027> <The server initialized the domain log
broadcaster successfully. Log messages will now be broadcasted to the domain log.>
####<Dec 29, 2006 2:14:24 PM IST> <Notice> <weblogicserver> <svaidyan02> <xbusServer> <Main
Thread> <<wls kernel>> <<> <<1167381864976> <BEA-000365> <server state changed to ADMIN>
####<Dec 29, 2006 2:14:24 PM IST> <Notice> <weblogicserver> <svaidyan02> <xbusServer> <Main
Thread> <<wls kernel>> <<> <<1167381864996> <BEA-000365> <server state changed to RESUMING>
####<Dec 29, 2006 2:14:28 PM IST> <Notice> <security> <svaidyan02> <xbusServer> <[STANDBY]
ExecuteThread: 5' for queue: 'weblogic.kernel.Default (self-tuning)''> <<wls kernel>> <<>
<<1167381868541> <BEA-090171> <Loading the identity certificate and private key stored under
the alias demoidentity from the jks keystore file
C:\bea2613a\WEBLOG-1\server\lib\demoIdentity.jks.>
####<Dec 29, 2006 2:14:29 PM IST> <Notice> <security> <svaidyan02> <xbusServer> <[STANDBY]
ExecuteThread: 5' for queue: 'weblogic.kernel.Default (self-tuning)''> <<wls kernel>> <<>
<<1167381869643> <BEA-090169> <Loading trusted certificates from the jks keystore file
C:\bea2613a\WEBLOG-1\server\lib\demoTrust.jks.>
####<Dec 29, 2006 2:14:29 PM IST> <Notice> <security> <svaidyan02> <xbusServer> <[STANDBY]
ExecuteThread: 5' for queue: 'weblogic.kernel.Default (self-tuning)''> <<wls kernel>> <<>
<<1167381869713> <BEA-090169> <Loading trusted certificates from the jks keystore file
C:\bea2613a\JROCKI-1\jre\lib\security\cacerts.>
####<Dec 29, 2006 2:15:32 PM IST> <Warning> <server> <svaidyan02> <xbusServer>
<-dynamicSSLListenerThread[DefaultSecure[1]]> <<wls kernel>> <<> <<1167381932743> <BEA-002611>
<hostname "svaidyan02.apac.bea.com", maps to multiple IP addresses: 192.168.1.5,
172.22.56.120>
####<Dec 29, 2006 2:15:32 PM IST> <Notice> <server> <svaidyan02> <xbusServer> <[STANDBY]
ExecuteThread: 5' for queue: 'weblogic.kernel.Default (self-tuning)''> <<wls kernel>> <<>
<<1167381932733> <BEA-002613> <channel "Default[2]" is now listening on 127.0.0.1:7021 for

```

Fig. 1.3 Example of Unstructured Data

Natural Language

The natural language is an unstructured data type that requires an understanding of certain data science methodologies and linguistics in order to be analyzed.

Despite the fact that the community of natural language processing has achieved success in the recognition of entities, topics, summarization, text completion, and sentiment analysis, models trained in one domain do not translate well to another.

Even cutting-edge tools are incapable of deciphering the meaning of every text. It should not come as a surprise, however, since people also struggle with natural language. It has inherent ambiguity. In this case, the whole notion of meaning is contested. Allow two persons to hear the same discussion. Will the same meaning be conveyed? The meaning of identical words might alter depending on whether they are said by a person who is sad or happy.

Machine-generated Data

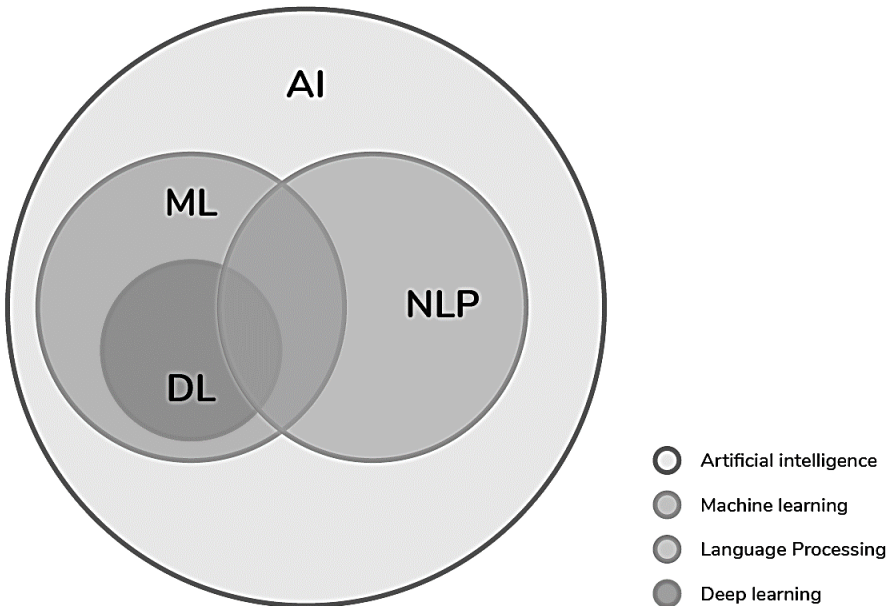


Fig. 1.3 An Illustration of Natural Language Processing

An example of machine-generated data is a set of data that is created automatically by a computer, process, program, or other machine without the involvement of a human. Machine-generated data will continue to be an important data resource.

Due to their enormous volume and velocity, Machine data analysis requires highly scalable technologies.

Examples include web server logs, call detail records, network event logs, and telemetry data.

For heavily linked or "networked" data, where the relationships between entities are important, this may not be the most effective strategy.

Graph Based Data

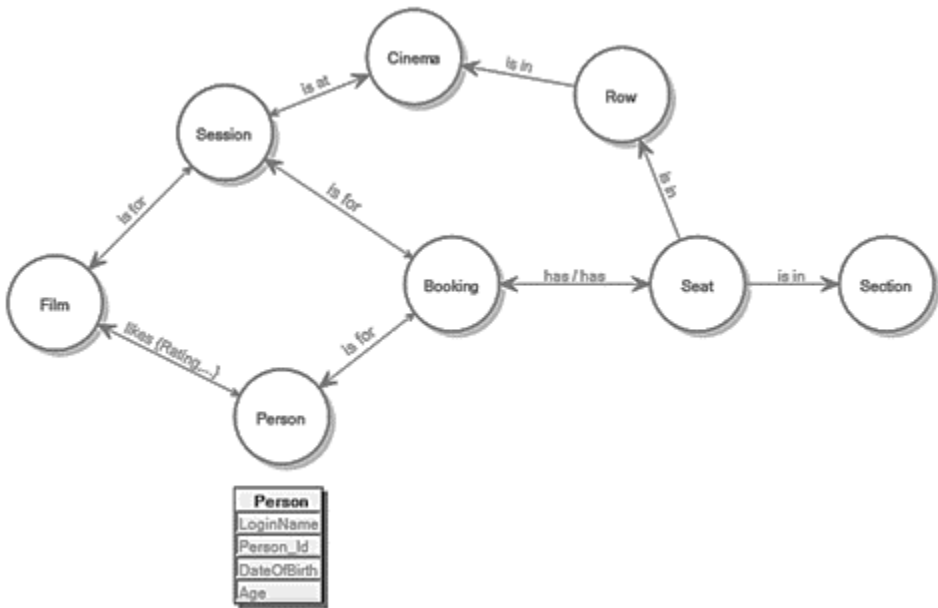


Fig. 1.4 An Illustration of Graph based Data

The phrase "graph data" may be ambiguous since any data can be shown in a graph. Graph theory refers to graphs in mathematics in this context. In graph theory, a graph is a mathematical structure that describes pair-wise interactions between objects. Data relating to the connection or adjacency of items is known as graph data or network data.

Nodes, edges, and attributes are utilised by graph structures to represent and store graphical data.

Graph-based data can be illustrated by friends in a social network.

It is natural to use graph-based data to describe social networks, and its structure allows us to calculate certain metrics, such as a person's influence or the shortest path between two individuals.

In graph databases, graph-based data is stored and searched using specific query languages like SPARQL.

The interpretation of additive and picture data can be more complex for a computer than the interpretation of graph data.

Audio, Visuals, and Photographs

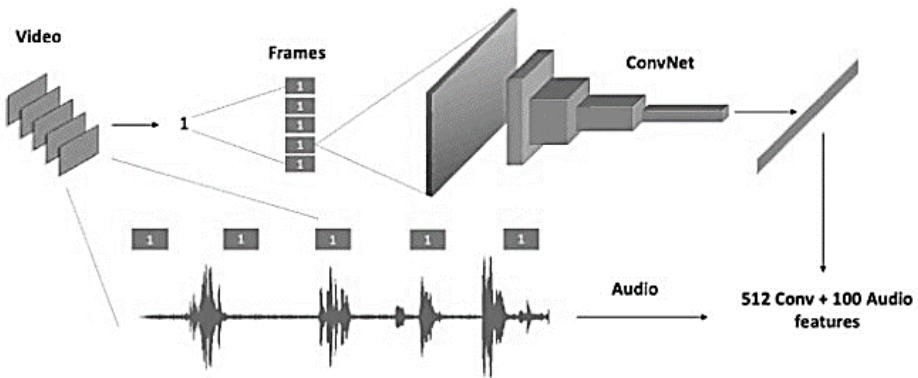


Fig. 1.5 An Illustration of Audio, Visuals and Photographs

A data scientist has unique problems when working with audio, picture, and video data. Recognizing objects in images is an example of a task that is simple for people but difficult for computers.

Multimedia data in the form of music, video, pictures, and sensor signals have become a fundamental aspect of daily living. In addition, by providing many data sources for quantitative and systematic evaluation, they have transformed product testing and evidence collecting.

Numerous libraries, programming languages, and integrated development environments (IDEs) are available, including:

MATLAB

Streaming Data

Streaming data has an additional characteristic, despite the fact that it can assume practically any of the preceding forms. As an event occurs, the data is not put into a data storage system in a single batch. Although it is not a

distinct sort of data, we will consider it as such because you will need to modify your process to accommodate it.

A few examples include Twitter's "What's trending," live sporting events, and the financial markets.

These are the seven most crucial features of Data Science...

DATA SCIENCE PROCESS

Though data scientists may disagree on the implications of a particular data set, practically all experts agree on the importance of following a disciplined data science process. There are several frameworks available, some of which are more appropriate for corporate use cases than for research use cases.

The following section will provide an overview of the most prominent data science process frameworks, which ones are best suited to each use case, and the core components of each framework.

What is the process of data science?

Data science is a methodical approach to solving data problems. The framework enables the formulation of a problem as a question, the determination of how to resolve it, and the presentation of the solution to stakeholders.

Data Science Evolution

A data science life cycle can be viewed as a synonym for a data science process. The terms describe a workflow process that begins with data collection and ends with the deployment of a model that will hopefully answer your questions. The steps consist of:

Defining Issue

The initial phase of the data science life cycle is problem understanding and formulation. This framework will assist you in developing a model that will positively benefit your business.

Gathering Data

The next step is to obtain the appropriate data set.

It is essential to acquire high-quality, focused data, as well as the means to collect them, in order to achieve significant outcomes.

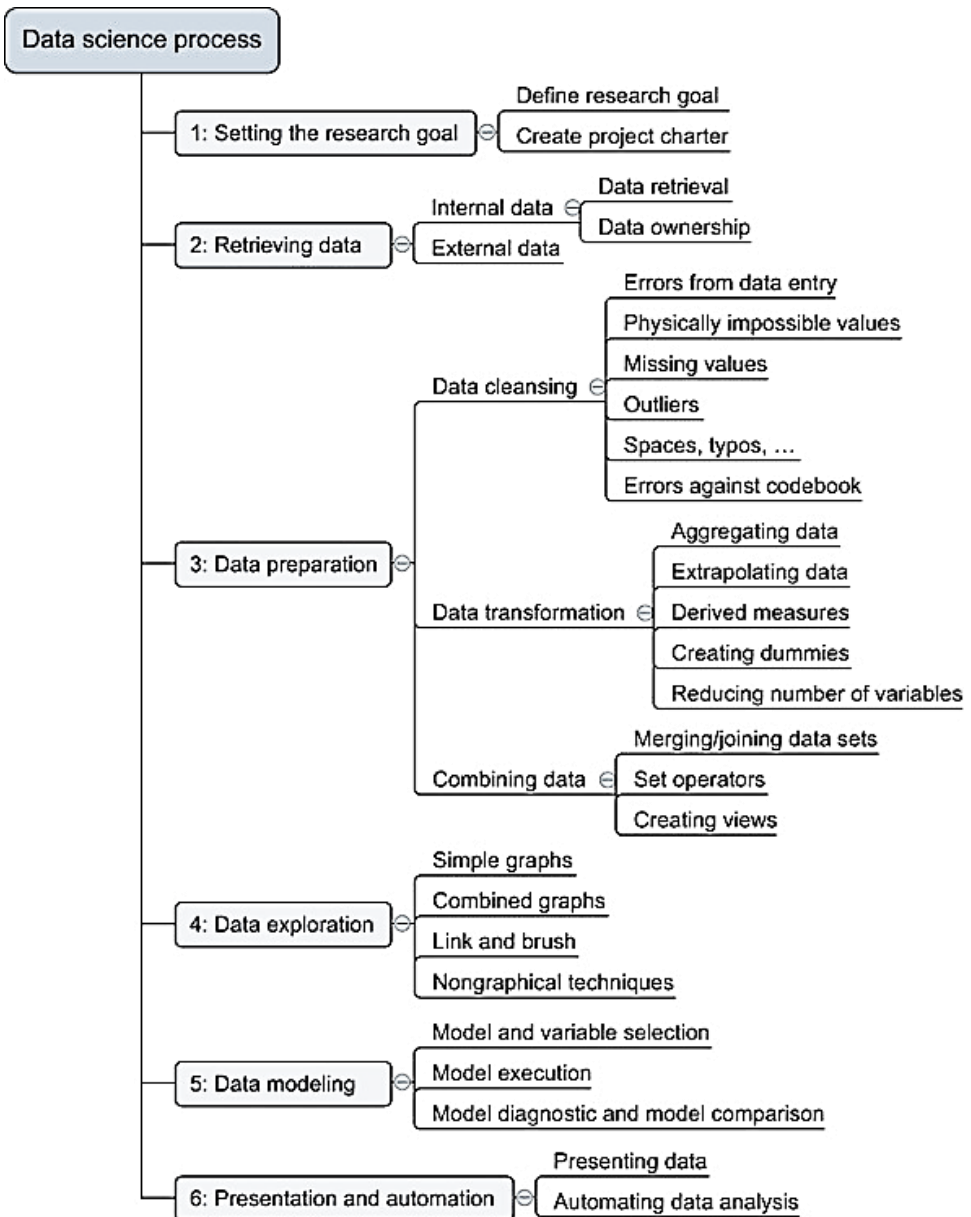


Fig. 1.6 An Illustration of Data Process

As most of the approximately 2.5 quintillion bytes of data produced every day are unstructured, you will probably need to convert the data into a useful format, such as CSV or JSON.

Cleaning Data science process: Cleaning Data

In the gathering phase, the majority of the data you acquire will be unstructured, irrelevant, and unfiltered. Precision and usefulness of your analysis will be highly dependent upon the quality of your data, as inaccurate data will yield inaccurate results. In order to clean data, it is necessary to eliminate duplicate and null values, corrupt data, inconsistent data types, incorrect entries, missing data, and poor formatting.

In order to develop good models, you must find and fix problems in your data.

Exploratory Data Analysis (EDA)

Having gathered a substantial amount of well-organized, high-quality data, you may now conduct exploratory data analysis (EDA). Effective EDA enables the discovery of important insights for the subsequent phase of the data science lifecycle.

Model Construction and Deployment

The next step is to perform the actual data modeling. In order to derive valuable insights and predictions, you will utilize machine learning, statistical models, and algorithms.

Presenting Your Results

As a final step, you will share your results with stakeholders. To do this, every data scientist must develop their visualisation abilities.

Stakeholders are primarily interested in what your results mean for their business; they are typically uninterested in the intricate process used to develop your model. Your results should be communicated in a manner that emphasizes their importance to the strategic planning and operations of your company.

Data Science Procedures and Structure

There are many data science process frameworks that you should be familiar with. Although they all strive to lead you through a productive process, different use cases are better suited for various approaches.

1. CRISP-DM

CRISP-DM stands for Cross Industry Standard Process for Data Mining. An industry-standard technique and process model that can be adapted and adjusted. It is also a tried-and-true way for directing data mining

initiatives. The CRISP-DM model has six data process life cycle phases. These are the six phases:

Understanding the data science procedure: Business Understanding

CRISP-DM begins with defining the business's objectives and bringing the data science project into focus. Clearly describing the objective should go beyond simply naming the measure you intend to change. Metrics cannot be altered without action, regardless of how extensive the analysis is.

Data scientists engage with stakeholders, subject matter experts, and other individuals who may provide insight into the issue at hand to gain a deeper understanding of the company. It may also be necessary for them to conduct preliminary research to determine how others have dealt with similar issues in the past. In the end, they will have a well-defined problem and a plan for addressing it.

2. Data Understanding

In CRISP-DM, the next phase is data comprehension. As a result of this step, you will be able to identify what data you own, where you might be able to obtain more, what your data consists of, and its quality. You will also determine the data collection technologies you will use as well as how your initial data will be collected. Next, you will specify the attributes of your initial data, including its format, number, records, and fields.

As soon as you have collected and documented your data, you can begin analyzing it. By posing data science questions that can be addressed by queries, visualizations, or reports, you can develop your initial hypothesis. By identifying any errors or missing values, you will conclude your data quality assessment.

3. Data Preparation data science process:

Typically, data preparation consumes the most time, and you may need to review it several times throughout the project.

In its raw form, data is usually worthless, as it often contains erroneous or absent properties, contradictory values, and outliers. Preparing your data eliminates these concerns and enhances its quality, allowing you to use it successfully in your modelling process.

The preparation of data involves several tasks that can be carried out in a variety of ways. The primary data preparation tasks are:

Data cleansing is the process of correcting incomplete or inaccurate data.

Data integration: integrating data from diverse sources

Transformation of data: formatting the data

Data reduction: the simplification of data.

Data discretization: the process of lowering the number of values to facilitate data management

Feature engineering is the process of choosing and modifying variables to improve machine learning performance.

4. Modelling

There are several data modelling possibilities. You will determine the optimal solution based on the business's objectives, the factors involved, and the available resources.

When choosing your modelling approach, you will generate two reports. The first will specify the modelling approach that will be employed. The second section will detail the modelling report's underlying assumptions, such as whether the model needs a certain sort of data distribution.

As soon as you have chosen a modelling approach, you will create tests to determine how well your model performs. Your deliverable for this phase will be your test design. In order to prevent overfitting, your training data and testing data may need to be separated into training data and testing data. In overfitting, a model fits one piece of data flawlessly but not another. You must avoid injecting any bias into your data during this phase.

As a next step, you will construct your model in accordance with the specific objectives of your organization. This will result in the delivery of three items:

- Parameter configurations
- A summary of the models
- Models on their own
- The final step of the modelling process is model evaluation. You will evaluate them from both a technical and business perspective. It is also possible that your project team will include subject matter specialists when examining your models.

Model evaluations summarize the results of your model review, including a rating of your models if you have created a number of them. You may now modify your parameters and run another round of simulations.

5. Evaluation data science process:

The evaluation phase involves analyzing the model in light of the objectives of your company. As a result, you will evaluate your work process, describe how your model will benefit the firm, summarize your results, and make any necessary adjustments.

As a final step, you will select your next course of action. Is your model ready for deployment? Does it require a fresh iteration or another project as a dependency?

6. Deployment

Despite the fact that deployment is the final step of the CRISP-DM process, it does not necessarily mean that your project has been completed. During the deployment phase, you will describe how you plan to implement the model and how the results will be provided. During the deployment phase, you'll also need to monitor the findings and maintain the model.

To complete your assignment, prepare a summary report and presentation. Evaluate the entire procedure to determine what worked and what could be improved.

7. OSEMN

OSEMN is not an iterative procedure, as is CRISP-DM. The technique is not used as frequently as CRISP-DM, however, in certain situations it is superior to CRISP-DM.

Because it excludes business-oriented phases, it is suitable for exploratory research projects and is frequently used by research institutes and public health agencies. OSEMN is less concerned with asking specific questions than it is with what the statistics have to say.

Obtain Data data science process:

As OSEMN is not a goal-driven initiative, the initial phase involves the collection of data. As with other frameworks, you may collect your data using a variety of technologies, such as web scraping, APIs, and SQL.

Scrub Data

No matter what you intend to do with your data, it must be cleansed before it can be utilized. It may be necessary to reformat even clean data in order to make it usable. Any data project must begin with this phase.

Explore Data

By exploring your data, you will be able to gain a deeper understanding of any initial trends and connections you observe. As of yet, you have not evaluated any hypotheses or made any predictions. The following techniques can be used to explore your data during this phase:

Command-line utilities

- Histograms to summarise data characteristics
- Pairwise histograms to identify correlations and emphasise outliers
- Dimensional reduction techniques
- Clustering to identify groups

Model Data

The accuracy of a model is the ultimate measure of its success. In general, the most predictive model is the most effective model. You may test the predicted accuracy of your model by seeing how it performs on new data. Machine-learning approaches utilising supervised or unsupervised algorithms are used to model data. Supervised learning use labelled datasets to train a model to produce correct predictions or classifications. Using algorithms, unsupervised learning clusters and analyses unlabeled datasets. It is possible to uncover hidden patterns and connections using unsupervised learning models.

It is important to note that both types of algorithms can be used in the OSEMN procedure, but unsupervised models are particularly valuable because they may help you discover patterns that you were not aware of when you began your project.

Interpret Results

It is the interpretation of the results of the OSEMN model that constitutes the conclusion. In this regard, it is important to emphasize that a model's predictive and interpretive abilities do not always match and may even be in conflict. The results of a highly predictive model may be difficult to comprehend, not simpler.

The predictive ability of a model depends on its ability to generalize, whereas the interpretive ability provides insight into a particular issue or topic. You may wish to select a model with a balance of predictive and interpretative skills, or you may prefer to prioritise one over the other. It depends on the project's objectives.

SETTING THE RESEARCH GOAL

To begin a project, it is important to understand the what, the why, and the how (figure 1.7). What is expected of you by the company? Why does management place such a high value on your research? Is the project a part of a larger strategic plan or a lone wolf project initiated by someone who recognized a potential opportunity? This phase aims to answer these three questions (what, why, how) so that everyone knows what to do and can agree on the best approach.

A clear research objective, a good understanding of the context, well-defined deliverables, and a plan of action with a timetable should be the results. This information is then best placed in a project charter. Depending on the project and the company, the length and formality of the letter may vary. The early phases of the project emphasize the importance of people skills and business acumen over technical expertise, which is why more senior personnel will often guide this phase.

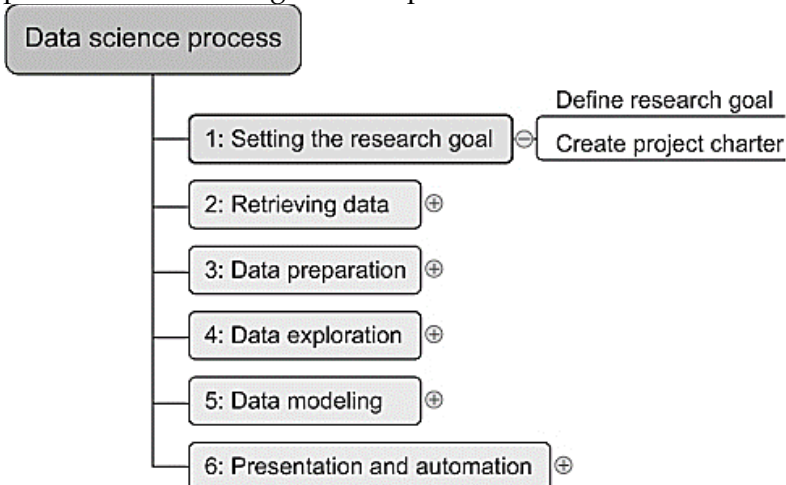


Fig. 1.7 An Illustration of setting the research goal

Spend time understanding the goals and context of your research

Your research goal should clearly and concisely state the purpose of your assignment. In order for a project to succeed, it is essential to understand the business goals and context. Ask questions and devise examples until you have a clear understanding of what the business expects. Assess how your project fits into the larger picture, understand how your research will impact the business, and determine how your results will be used. After spending months researching something, you have that one moment of brilliance and solve the problem, but when you report your findings back to the organization, everyone immediately realizes that you misunderstood their question. This is an important phase that should not be overlooked. In spite of their mathematical acumen and scientific brilliance, many data scientists fail to grasp business objectives and context.

Create a project charter

Once you have a clear understanding of the business problem, try to obtain a formal agreement on the deliverables. It is best to gather all of this information in a project charter. This would be a requirement for any significant project.

In order for a project charter to be successful, teamwork is necessary, and your input should include at least the following:

- A clear goal for the research
- Mission and context of the project
- Describe how you intend to conduct your analysis
- Resources you expect to use
- A demonstration of the feasibility of the project, or a proof of concept
- A measure of success and deliverables
- Here is a timeline
- This information can be used by your client to estimate the project costs, as well as the data and people required to ensure the success of your project.

RETRIEVING DATA

The subsequent phase in data science is data retrieval. Occasionally, you will be required to go into the field and create the data gathering procedure yourself, but this will be the exception. In many cases, businesses have already acquired and stored the data for you, and what they do not have may be purchased from third parties. It is not necessary to be hesitant to get data outside of your organization, as an increasing number of companies are making even high-quality data publicly available.

From basic text files to database tables, data can be stored in a variety of formats. Currently, the purpose of the project is to collect the necessary information. Even if you are able to accomplish this, data is often like a raw diamond that needs to be polished before it can be used effectively.

The first step should be to begin with information owned by the company.

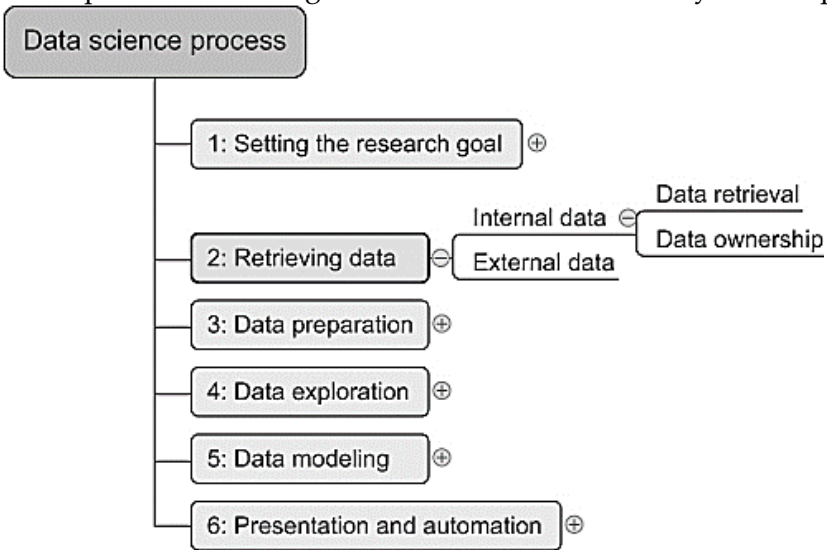


Fig. 1.8 An Illustration of Retrieving Data

Your first step should be to assess the relevance and quality of the readily available data within your organization. In most businesses, critical data is stored in a secure location, so the majority of the cleaning work has already been completed. It is possible to store this data in IT-managed databases, data marts, data warehouses, and data lakes. The primary purpose of a database is to store data, whereas the primary purpose of a data warehouse

is to retrieve and analyze data. A data mart is a subset of a data warehouse that serves a specific business unit. In contrast to data warehouses and data marts, data lakes store data in its natural or unprocessed state. It is possible, however, that your data is still stored in Excel files on the desktop of a subject matter expert.

Information can be difficult to locate, even within your own organization. As businesses expand, their data is dispersed across multiple locations. It is possible for employees to lose knowledge of the data as they change roles or leave the organization. There is a possibility that you will need Sherlock Holmes-like skills in order to find all the missing documentation and metadata.

Accessing data is another challenging task. Organizers recognize the importance and sensitivity of data, and often have policies in place to ensure that everyone has access only to the information they need. Chinese walls, which are physical and digital obstacles, are the result of these rules. For client data, these "walls" are required and well-regulated in the majority of nations. Also, this is a good reason; imagine if every employee of a credit card company had access to your spending habits. Accessing the data may require business politics and a considerable amount of time.

In the event that data is not available within your organization, you should look outside. The collection of important data is a specialty of many businesses. It is well known that Nielsen and GFK provide this service to the retail industry. You receive data from other companies in order to enhance their services and ecosystems. The same applies to Twitter, LinkedIn, and Facebook.

A growing number of governments and organizations are sharing their data with the world for free because they view data as a more valuable asset than oil. The quality of the data may vary depending on the institution that generates and manages it. Furthermore, they exchange information on the number of accidents and drug abuse in a particular region, as well as its demographics. In addition to enriching proprietary data, this data can also be used to practice data science skills at home.

Perform data quality checks now to avoid difficulties in the future.

You should expect to spend up to 80 percent of your project's time repairing and cleansing data. In the data science process, the first time you will examine the data is during retrieval. It is easy to notice most of the errors during the data collection phase, but if you are negligent, you will spend countless hours fixing data problems that you could have prevented if you had done your homework during the data collection phase.

Data exploration occurs during the steps of data import, data preparation, and data exploration. The difference lies in the objective and scope of the investigation. A data retrieval process involves comparing the data with data in the source document and determining whether the data types are appropriate. It should not take you too long to complete this process; you should finish when there is sufficient evidence that the data matches the information in the original document. During the preparation of data, a more thorough examination is conducted. The mistakes you discover now will also be present in the source document if you did a decent job during the previous phase. The focus should be on the substance of the variables: you want to eliminate typos and other data input errors, and standardize data across different datasets. In the exploratory stage, the focus shifts to what can be learned from the data. If the data are clean, you will now examine statistical features such as distributions, correlations, and outliers. These periods will be repeated frequently. An outlier may indicate a problem with the data input during the exploratory phase. As you have gained a better understanding of how data quality is improved throughout the process, we will now discuss the data preparation step in more detail.

CLEANSING, INTEGRATING AND TRANSFORMING DATA

In the data retrieval phase, you may have obtained "rough diamonds." It is now your responsibility to clean and prepare the data for use in the modeling and reporting phases. As a result, your models will perform better and you will spend less time trying to rectify unexpected outputs. The maxim "trash in, trash out" cannot be emphasized enough. Your model requires data to be in a particular format, so data transformation will always be required. Correcting data inaccuracies as early in the process as

feasible is a recommended practise. In an actual scenario, this may not always be possible, so your software will need to be adjusted.

The figure illustrates the most common steps used during data cleansing, integration, and transformation.

At the moment, this mind map appears somewhat abstract, but we will elaborate on each of these elements in the following sections. All of these acts share a great deal of similarity.

Cleansing data

Data cleaning is a subprocess of the data science process that focuses on eliminating inaccuracies from your data so that it is an accurate and consistent depiction of the processes from which it was derived.

The phrase "true and consistent representation" implies the existence of at least two forms of mistake.

Bibliography

This Page Intentionally Left Blank

BIBLIOGRAPHY

- Bisong, E. (2019).** *Building machine learning and deep learning models on Google cloud platform: A comprehensive guide for beginners*. Apress.
- Blum, A., Hopcroft, J., & Kannan, R. (2018).** *Foundations of Data Science (2018)*. URL: <https://www.cs.cornell.edu/jeh/book.pdf>.
- Cao, L. (2017).** Data science: a comprehensive overview. *ACM Computing Surveys (CSUR)*, 50(3), 1-42.
- De Brouwer, P. J. (2020).** *The Big R-Book: From Data Science to Learning Machines and Big Data*. John Wiley & Sons.
- Efron, B., & Hastie, T. (2021).** *Computer Age Statistical Inference, Student Edition: Algorithms, Evidence, and Data Science (Vol. 6)*. Cambridge University Press.
- Erl, T., Khattak, W., & Buhler, P. (2016).** *Big data fundamentals: concepts, drivers & techniques*. Prentice Hall Press.
- Maier-Hein, L., Eisenmann, M., Sarikaya, D., März, K., Collins, T., Malpani, A., ... & Speidel, S. (2022).** Surgical data science—from concepts toward clinical translation. *Medical image analysis*, 76, 102306.
- National Academies of Sciences, Engineering, and Medicine. (2018).** *Data science for undergraduates: Opportunities and options*. National Academies Press.
- Pierson, L. (2021).** *Data science for dummies*. John Wiley & Sons.
- Provost, F., & Fawcett, T. (2013).** Data science and its relationship to big data and data-driven decision making. *Big data*, 1(1), 51-59.
- Saura, J. R. (2021).** Using data sciences in digital marketing: Framework, methods, and performance metrics. *Journal of Innovation & Knowledge*, 6(2), 92-102.
- Sharda, R., Delen, D., & Turban, E. (2021).** *Analytics, data science, & artificial intelligence: Systems for decision support*. Pearson Education Limited.
- Selvan, C., & Balasundaram, S. R. (2021).** Data Analysis in Context-Based Statistical Modeling in Predictive Analytics. In *Handbook of Research on Engineering, Business, and Healthcare Applications of Data Science and Analytics* (pp. 96-114). IGI Global.
- Shwartz-Ziv, R., & Armon, A. (2022).** Tabular data: Deep learning is not all you need. *Information Fusion*, 81, 84-90.
- Singleton, A., & Arribas-Bel, D. (2021).** Geographic data science. *Geographical Analysis*, 53(1), 61-75.
- Szajowski, K. J. (2017).** On a book Algorithms for data science by Brian Steele, John Chandler and Swarn Reddy. *Mathematica Applicanda*, 45(2).
- Van Der Aalst, W. (2016).** *Process mining: data science in action (Vol. 2)*. Heidelberg: Springer.
- Van Der Aalst, W. (2016).** *Process mining: data science in action (Vol. 2)*. Heidelberg: Springer.

WEBLIOGRAPHY

- ConductScience.** (2020, January 20). Why do we need data science. Conduct Science. <https://conductscience.com/why-do-we-need-data-science/>
- Kumar, R. (2022, August 31). What is data science & advantages and disadvantages of data science what is data science. DevOpsSchool.com. <https://www.devopsschool.com/blog/what-is-data-science-advantages-and-disadvantages-of-data-science/>
- Saradalakshmi8074.** (2022, October 3). Data journalism. Medium. <https://medium.com/mlearning-ai/facets-of-data-science-54d7e448ae75>
- Data science process: A beginner's guide in plain English.** (2022, August 10). Springboard Blog. <https://www.springboard.com/blog/data-science/data-science-process>
- Software, E.** (2022, September 6). The importance of defining a research goal in a data science project. Expeed Software | Trustworthy Software Solutions. <https://expeed.com/blog-posts/the-importance-of-defining-a-research-goal-in-a-data-science-project/>
- Chapter 2. The data science process** · Introducing data science: Big data, machine learning, and more, using Python tools. (n.d.). liveBook · Manning. <https://livebook.manning.com/book/introducing-data-science/chapter-2/42>
- Zubair, M.** (2022, October 16). To increase data analysing power you must know frequency distribution (stat-04). Medium. <https://towardsdatascience.com/to-increase-data-analysing-power-you-must-know-frequency-distribution-afa438c3e7a4>
- Raj, A.** (2022, August 18). Outlier detection and treatment in data science. CloudyML. <https://www.cloudymml.com/blog/outlier-detection-and-treatment/>
- Vijayamohan, P.** (2022, September 19). Nominal data 101 - Definition, examples, analysis. SurveySparrow. <https://surveysparrow.com/blog/nominal-data/>