CONTENTS

Chapter 1

INTRODUCTION TO DATA SCIENCE

☞ Learning Objectives and Outcomes

In this unit we will learn about

- ☞ Introduction to core concepts and technologies
- ☞ Familiarity with terminology related with Data Science
- ☞ Dealing with Data science process,
- ☞ Getting acquainted with various popular Data science toolkits
- ☞ Types of data dealt with in Data Science
- ☞ Familiarity with example applications

1.1 DATA SCIENCE

Data science, also known as *data-driven science*, is an interdisciplinary field of scientific methods, processes, and systems to extract knowledge or insights from data in various forms, either structured or unstructured,[1][2] similar to data mining. We shall learn about data mining in short in this unit itself. In fact it is emerging as convergence of various knowledge domains for effective utilisations of various analysis methods for better output of experts in their activities (vide Figure 1.1).

Convergence of various knowledge domains	
Math and Theory	• Statistics, Linear Algebra, Optimization, Time Series, etc.
Applied Algorithms	• Machine Learning, Data Structures, Parallel Algorithms, etc.
Engineering and Technologies	• Storage and computing platforms, statistical tools, etc.
Domain Expertise	• Text, Finance, Images, Economies etc.
Art	• Visualization, Infographics
Best practices and hacks	• Handle missed values in data, transform and represent data, etc.

Fig. 1.1: Data science as convergence of various knowledge domains

As such Data Science is one of the recent fields combining *big data, unstructured data* and *combination of statistics* and *analytics* and *business intelligence*. It is a new field that has emerged within the field of Data Management providing understanding of correlation between structured and unstructured data. More accurately, Data Science is the discipline of using quantitative methods from **statistics** and **mathematics** along with **technology** (computers and software) to develop algorithms designed to discover patterns, predict outcomes, and find optimal solutions to complex problems. Nowadays, data scientists are in great demand as they can transform unstructured data into actionable insights, helpful for businesses.

Data science is blossoming as "concept to unify statistics, data analysis and their related methods" in order to "understand and analyze actual phenomena" with big data.[3] In its extended canvas (i.e. while dealing with Big. data), data science employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, information science, and computer science, in particular from the sub-domains of *machine learning, classification, cluster analysis, data lakes data mining and warehousing, databases,* and *visualization* (vide Figure 1.2). We shall learn about dimensions of Big data in this unit very shortly.

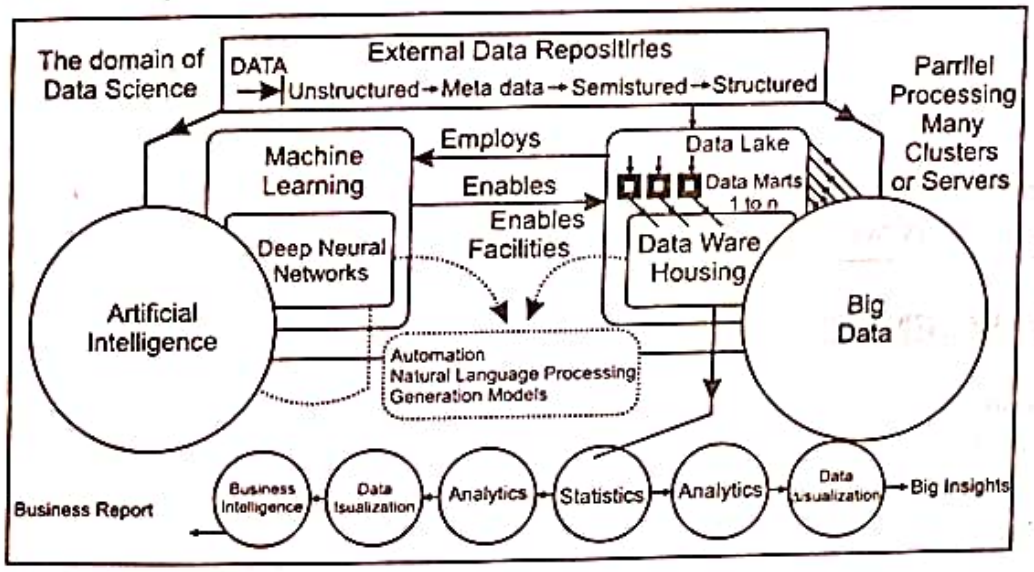

Fig. 1.2: Broad Canvas of Data Science Dealing with Big Data

Turing award winner Jim Gray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge.[4][5]

1.2 TERMINOLOGY RELATED WITH DATA SCIENCE

1.2.1 Big Data

As per Oxford English Dictionary , the definition of Big Data is "data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges." Actually Big data is a term applied to data sets whose size or type is beyond the ability of traditional relational databases to capture, manage, and process the data with low-latency.

Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time. For reader's convenience and as memory refresher, I am giving the numeric details of decimal number system below:

1000 (10^3) kB 1 kilobyte

1000 (10^6) MB 1 megabyte

1000 (10^9) GB 1 gigabyte

1000 (10^{12}) TB 1 terabyte

1000 (10^{15}) PB 1 petabyte

1000 (10^{18}) EB 1 exabyte

1000 (10^{21}) ZB 1 zettabyte

1000 (10^{24}) YB 1 yottabyte

Actually 1 KB = 1024 = 2^8 .We have to enhance each of above by this magnitude. A processor with a 64-bit address bus can address 16 exbibytes of memory, which is over 18 exabytes.

As per IDC, Digital universe is doubling in size every two years, and by 2020 the digital universe – the data we create and copy annually – will reach 44 zettabytes, or 44 trillion gigabytes. Between 2013

Because big data is simply larger than life itself, it can offer a detailed rendition of the user. The amount of data produced by users has surpassed the petabyte (10^{15}) level; it has clocked many zettabytes (10^{21}) of raw data or information and this figure is growing at a rapid range. In some years to come, the amount of data that is globally stored is expected to clock the yottabyte (10^{24}) level.

As per IDC, Digital universe is doubling in size every two years, and by 2020 the digital universe i.e. the data we create and copy annually will reach 44 zettabytes, or 44 trillion gigabytes. By 2025, there will be 163 zettabytes of data. Compare it with the human capacity of the human brain that is only not more than 2.5 petabytes as per IBM.

1.2.2 Business Intelligence (BI)

Business Intelligence is the technology which uses the transformed and loaded historical data to get or create the reports. It is a set of methodologies, process, theories that transform raw data into useful information to help companies make better decisions

BI is a process for analyzing data and presenting actionable information to help executives, managers and other corporate end users make informed business decisions and thus help in decision making. Common functions of business intelligence technologies include reporting, online analytical processing, analytics, data mining, process mining, complex event processing, business performance management, benchmarking, text mining, predictive analytics and prescriptive analytics.

Business intelligence can be used by enterprises to support a wide range of business decisions - ranging from operational to strategic. Basic operating decisions include product positioning or

pricing. Strategic business decisions involve priorities, goals and directions at the broadest level. Often BI applications use data gathered from a data warehouse (DW) or from a data mart.

1.2.3 Data Analytics

Data analytics and **analytics**, by contrast, are general terms used to describe the field and a comprehensive collection of associated methods. All these terms tend to be used for the application of analytic methods to data that large organizations generate or have available ("big data"). We shall learn in detail about application of data science in big data in unit 6 of this book.

Data analysts collect, process and perform statistical analyses of data. Their skills may not be as advanced as data scientists (e.g. they may not be able to create new algorithms), but their goals are the same – to discover how data can be used to answer questions and solve problems.

1.2.3.1 Difference between Big Data & Business Intelligence

The difference between Big Data & Business Intelligence is synonymous to fishing in the sea versus fishing in the lake. ... If you try to understand your business data that is structured and not of huge volume or variety or velocity then you make use of typical business intelligence tools & technologies.

Big Data collectively refers to the act of generating, capturing and usually processing enormous amounts of data on a continuing basis. Unlike business Intelligence, which encompasses only commercial activities, its domain is larger. The data is collected in data lakes and refined in data ware housing through data mining techniques. Refinement is done department wise first in datamarts and then in ware housing.

Business Intelligence collectively refers to software and systems that import data streams of any size and use them to generate informational displays that point towards specific decisions.

Big data is the technology which collects transforms the huge data which is in a unstructured manner. It takes help from Artificial intelligence techniques which demands unusually high rate processing of data. Figure 1.1 is an effort by author to mitigate differences.

1.2.4 Data Wrangling

The process of conversion of data, often through the use of scripting languages, to make it easier to work with is known as data *Wrangling* or *data munging*. If you have 900,000 birthYear values of the format yyyy-mm-dd and 100,000 of the format mm/dd/yyyy and you write a Perl script to convert the latter to look like the former so that you can use them all together, you're doing data wrangling. Discussions of data science often bemoan the high percentage of time that practitioners must spend doing data wrangling; the discussions then recommend the hiring of data engineers to address this

Data engineers build massive reservoirs for big data. They develop, construct, test and maintain architectures such as databases and large-scale data processing systems. Once continuous pipelines are installed to – and from – these huge "pools" of filtered information, data scientists can pull relevant data sets for their analyses.

1.2.5 Algorithm

A series of repeatable steps for carrying out a certain type of task with data. As with data structures, people studying computer science learn about different algorithms and their suitability for various tasks. Specific data structures often play a role in how certain algorithms get implemented.

1.2.6 Machine Learning

Analytics in which computers "learn" from data to produce models or rules that apply to those data and to other similar data. Predictive modelling techniques such as neural nets, classification and regression trees (decision trees), naive Bayes, k-nearest neighbour, and support vector machines are generally included. One characteristic of these techniques is that the form of the resulting model is flexible, and adapts to the data. Statistical modelling methods that have highly structured model forms, such as linear regression, logistic regression and discriminant analysis are generally not considered part of machine learning. Unsupervised learning methods such as association rules and clustering are also considered part of machine learning.

1.2.7 Web Analytics

Statistical or machine learning methods applied to web data such as page views, hits, clicks, and conversions (sales), generally with a view to learning what web presentations are most effective in achieving the organizational goal (usually sales). This goal might be to sell products and services on a site, to serve and sell advertising space, to purchase advertising on other sites, or to collect contact information. Key challenges in web analytics are the volume and constant flow of data, and the navigational complexity and sometimes lengthy gaps that precede users' relevant web decisions.

1.3 METHODS OF DATA REPOSITORY

Data repository is a somewhat general term used to refer to a destination designated for data storage.A data repository refers to an enterprise data storage entity (or sometimes entities) into which data has been specifically partitioned for an analytical or reporting purpose.

Data repositories may assume several different shapes like:

1. Data lakes ,
2. Data Marts
3. Data Ware Housing
4. Big Data and Hadoop and similar frameworks.

1.3.1 Data Lake

A data lake is a storage repository that holds a vast amount of raw data in its native format until it is needed and refined elsewhere. Data lake shares a data environment that comprises multiple

11. UML
12. ETL tools
13. Python, C/C++ Java, Perl
14. UNIX, Linux, Solaris and MS Windows
15. Hadoop and NoSQL databases
16. Machine learning
17. Data visualization

As always, this list is subject to changes in technology.

3. Expected Business Skills

1. Analytical Problem-Solving: Approaching high-level data challenges with a clear eye on what is important; employing the right approach/methods to make the maximum use of time and human resources.
2. Effective Communication: Carefully listening to management, data analysts and relevant staff to come up with the best data design; explaining complex concepts to non-technical colleagues.
3. Expert Management: Effectively directing and advising a team of data modelers, data engineers, database administrators and junior architects.
4. Industry Knowledge: Understanding the way your chosen industry functions and how data are collected, analyzed and utilized; maintaining flexibility in the face of big data developments.
5. Data Scientist
6. Data scientists are big data wranglers. They take an enormous mass of messy data points (unstructured and structured) and use their formidable skills in math, statistics and programming to clean, massage and organize them. Then they apply all their analytic powers – industry knowledge, contextual understanding, scepticism of existing assumptions – to uncover hidden solutions to business challenges.

1.5 TYPES OF DATA

Thus Data and Big Data includes huge volume, high velocity, and extensible variety of data. The data in it will be of three types.

1. Unstructured data : Word, PDF, Text, Media Logs.
2. Semi Structured data : XML data.
3. Meta Data :Data about data
4. Structured data : Relational data.

1.5.1 Unstructured Big Data

Any data with unknown form or the structure is classified as unstructured data. In addition to the size being huge, un-structured data poses multiple challenges in terms of its processing for deriving value out of it. Typical example of unstructured data is, a heterogeneous data source containing a

combination of simple text files, images, videos etc. Now a day organizations have wealth of data available with them but unfortunately they don't know how to derive value out of it since this data is in its raw form or unstructured format.

Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data. Over the period of time, talent in computer science have achieved greater success in developing techniques for working with such kind of data (where the format is well known in advance) and also deriving value out of it. However, now days, we are foreseeing issues when size of such data grows to a huge extent, typical sizes are being in the rage of multiple zettabyte.

An 'Employee' table in a database is an example of Structured Data

Employee_ID	Employee_Name	Gender	Department	Salary_in_lacs
2365	Rajsh Kulkarni	Male	Finance	650000
3398	Pratibha Joshi	Female	Admin	650000
7465	Shushil Roy	Male	Admin	500000
7500	Shubhojit Das	Male	Finance	500000
7699	Priya Sane	Female	Finance	550000

1.5.2 Semi-structured data

Semi-structured data can contain both the forms of data. We can see semi-structured data as a strcutured in form but it is actually not defined with e.g. a table definition in relational DBMS. Example of semi-structured data is a data represented in XML file.

Personal data stored in a XML file

<rec><name>Amitabh Bajaj</name><sex>Male</sex><age>45</age></rec>

<rec><name>Jyotsna Agarwal</name><sex>Female</sex><age>41</age></rec>

<rec><name>Anurag Jain</name><sex>Male</sex><age>39</age></rec>

<rec><name>Subrato Roy</name><sex>Male</sex><age>26</age></rec>

<rec><name>Jeremiah J.</name><sex>Male</sex><age>35</age></rec>

Web pages are generated in scripting of HTML which is also an example of **Semi-structured data.**

1.5.3 Meta Data

Metadata is defined as the data providing information about one or more aspects of the data; it is used to summarize basic information about data which can make tracking and working with specific data easier.[

There are three main types of metadata:

- **Descriptive metadata** describes a resource for purposes such as discovery and identification. It can include elements such as title, abstract, author, and keywords.

- **Structural metadata** indicates how compound objects are put together, for example, how pages are ordered to form chapters.
- **Administrative metadata** provides information to help manage a resource, such as when and how it was created, file type and other technical information, and who can access it. There are several subsets of administrative data; two that are sometimes listed as separate metadata types are:
  1. Rights management metadata, which deals with intellectual property rights, and
  2. Preservation metadata, which contains information needed to archive and preserve a resource.

Metadata Repository

Metadata repository is an integral part of a data warehouse system. It contains the following metadata –

- **Business metadata** – It contains the data ownership information, business definition, and changing policies.
- **Operational metadata** – It includes currency of data and data lineage. Currency of data refers to the data being active, archived, or purged. Lineage of data means history of data migrated and transformation applied on it.
- **Data for mapping from operational environment to data warehouse** – It metadata includes source databases and their contents, data extraction, data partition, cleaning, transformation rules, data refresh and purging rules.
- **The algorithms for summarization** – It includes dimension algorithms, data on granularity, aggregation, summarizing, etc.

1.5.4 Structured

Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data. Over the period of time, talent in computer science have achieved greater success in developing techniques for working with such kind of data (where the format is well known in advance) and also deriving value out of it. However, now days, we are foreseeing issues when size of such data grows to a huge extent, typical sizes are being in the zettabyte. rage of multiple

An 'Employee' table in a database is an example of Structured Data

Employee_ID	Employee_Name	Gender	Department	Salary_in_lacs
2365	Rajsh Kulkarni	Male	Finance	650000
3398	Pratibha Joshi	Female	Admin	650000
7465	Shushil Roy	Male	Admin	500000
7500	Shubhojit Das	Male	Finance	500000
7699	Priya Sane	Female	Finance	550000

1.6 THE DATA SCIENCE PROCESS (DSP)

DSP is an agile, iterative data science methodology to deliver predictive analytics solutions and intelligent applications efficiently. DSP helps improve team collaboration and learning. It contains a distillation of the best practices and structures from Microsoft and others in the industry that facilitate the successful implementation of data science initiatives. The goal is to help companies fully realize the benefits of their analytics program.

We provide a generic description of the process here that can be implemented with a variety of tools. A more detailed description of the project tasks and roles involved in the lifecycle of the process is provided in additional linked topics.

The process may involve 7 clear cut steps for data analysis as shown in Figure 1.5 :

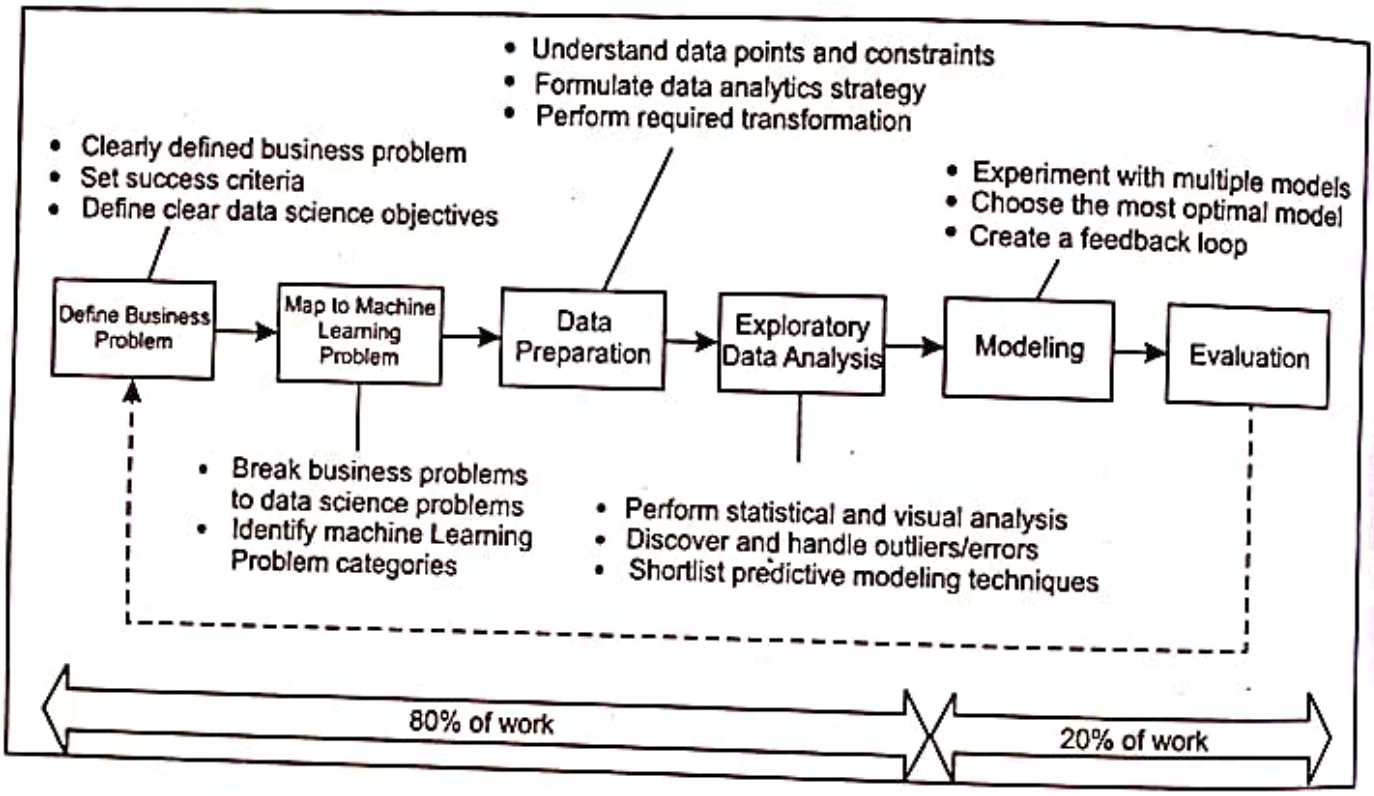

Fig. 1.5: Seven steps of the Data Science Process (DSP)

Step 1: Frame or define the (business) problem

Step 2: Collect the raw data needed for your problem (and map it to machine learning in case of Big data

Step 3: Data Preparation for process the data for analysis

Step 4: Explore the data (Exploratory Data Analysis (EDA)

Step 5: Perform in-depth analysis (Modelling) and producing prescriptive Business Insights

Step 6: Evaluation

Step 7: Visualisation and Communication of Results of the Analyse

Step 1: Frame or define the (business) problem [7]

The first thing you have to do before you solve a problem is to define exactly what it is. You need to be able to translate data questions into something actionable.

You'll often get ambiguous inputs from the people who have problems. You'll have to develop the intuition to turn scarce inputs into actionable outputs–and to ask the questions that nobody else is asking.

Say you're solving a problem for the VP Sales of your company. You should start by understanding their goals and the underlying why behind their data questions. Before you can start thinking of solutions, you'll want to work with them to clearly define the problem.

A great way to do this is to ask the right questions.

You should then figure out what the sales process looks like, and who the customers are. You need as much context as possible for your numbers to become insights.

You should ask questions like the following:

1. Who are the customers?
2. Why are they buying our product?
3. How do we predict if a customer is going to buy our product?
4. What is different from segments who are performing well and those that are performing below expectations?
5. How much money will we lose if we don't actively sell the product to these groups?

In response to your questions, the VP Sales might reveal that they want to understand why certain segments of customers have bought less than expected. Their end goal might be to determine whether to continue to invest in these segments, or de-prioritize them. You'll want to tailor your analysis to that problem, and unearth insights that can support either conclusion.

It's important that at the end of this stage, you have all of the information and context you need to solve this problem.

Step 2: Collect the raw data needed for your problem (and map it to machine learning in case of Big data

Once you've defined the problem, you'll need data to give you the insights needed to turn the problem around with a solution. This part of the process involves thinking through what data you'll need and finding ways to get that data, whether it's querying internal databases, or purchasing external datasets.

You might find out that your company stores all of their sales data in a CRM or a customer relationship management software platform. You can export the CRM data in a CSV file for further analysis. In case of big data, you have to adopt Machine Learning Process.

Step 3: Data Preparation for process the data for analysis

Now that you have all of the raw data, you'll need to process it before you can do any analysis. Oftentimes, data can be quite messy, especially if it hasn't been well-maintained. You'll see errors that will corrupt your analysis: values set to null though they really are zero, duplicate values, and missing values. It's up to you to go through and check your data to make sure you'll get accurate insights.

You'll want to check for the following common errors:

1. Missing values, perhaps customers without an initial contact date
2. Corrupted values, such as invalid entries
3. Timezone differences, perhaps your database doesn't take into account the different timezones of your users
4. Date range errors, perhaps you'll have dates that makes no sense, such as data registered from before sales started
5. You'll need to look through aggregates of your file rows and columns and sample some test values to see if your values make sense. If you detect something that doesn't make sense, you'll need to remove that data or replace it with a default value. You'll need to use your intuition here: if a customer doesn't have an initial contact date, does it make sense to say that there was NO initial contact date? Or do you have to hunt down the VP Sales and ask if anybody has data on the customer's missing initial contact dates?

Once you're done working with those questions and cleaning your data, you'll be ready for Exploratory Data Analysis (EDA).

Step 4: Explore the data (Exploratory Data Analysis (EDA)

When your data is clean, you'll should start playing with it!

The difficulty here isn't coming up with ideas to test, it's coming up with ideas that are likely to turn into insights. You'll have a fixed deadline for your data science project (your VP Sales is probably waiting on your analysis eagerly!), so you'll have to prioritize your questions. '

You'll have to look at some of the most interesting patterns that can help explain why sales are reduced for this group. You might notice that they don't tend to be very active on social media, with few of them having Twitter or Facebook accounts. You might also notice that most of them are older than your general audience. From that you can begin to trace patterns you can analyze more deeply.

Step 5: Perform in-depth analysis (Modelling) and producing prescriptive Business Insights

This step of the process is where you're going to have to apply your statistical, mathematical and technological knowledge and leverage all of the data science tools at your disposal to crunch the data and find every insight you can.

In this case, you might have to create a predictive model that compares your under performing group with your average customer. You might find out that the age and social media activity are significant factors in predicting who will buy the product.

If you'd asked a lot of the right questions while framing your problem, you might realize that the company has been concentrating heavily on social media marketing efforts, with messaging that is aimed at younger audiences. You would know that certain demographics prefer being reached by telephone rather than by social media. You begin to see how the way the product has been has been marketed is significantly affecting sales: maybe this problem group isn't a lost cause! A change in tactics from social media marketing to more in-person interactions could change everything for the better. This is something you'll have to flag to your VP Sales.

Step 6: Evaluation

You can now combine all of those qualitative insights with data from your quantitative analysis to craft a story that moves people to action.

It's important that the VP Sales understand why the insights you've uncovered are important. Ultimately, you've been called upon to create a solution throughout the data science process.

Step 7 : Visualisation and Communication of Results of the Analysis

Proper communication will mean the difference between action and inaction on your proposals.

You need to craft a compelling story here that ties your data with their knowledge. You start by explaining the reasons behind the underperformance of the older demographic. You tie that in with the answers your VP Sales gave you and the insights you've uncovered from the data. Then you move to concrete solutions that address the problem: we could shift some resources from social media to personal calls. You tie it all together into a narrative that solves the pain of your VP Sales: she now has clarity on how she can reclaim sales and hit her objectives.

Throughout the data science process, your day-to-day will vary significantly depending on where you are–and you will definitely receive tasks that fall outside of this standard process! You'll also often be juggling different projects all at once.

It's important to understand these steps if you want to systematically think about data science, and even more so if you're looking to start a career in data science.

1.5.7 Data Science Project's Lifecycle

The Team Data Science Process (TDSP) provides a lifecycle to structure the development of your data science projects. The lifecycle outlines the steps, from start to finish, that projects usually follow when they are executed.

If you are using another data science lifecycle, such as or your organization's own custom process, you can still use the task-based TDSP in the context of those development lifecycles. At a high level, these different methodologies have much in common.

This lifecycle has been designed for data science projects that ship as part of intelligent applications. These applications deploy machine learning or artificial intelligence models for predictive analytics. Exploratory data science projects or ad hoc analytics projects can also benefit from using this process. But in such cases some of the steps described may not be needed.+

CRISP-DM remains the top methodology for data mining projects.CRISP-DM was conceived around 1996.

The 6 high-level phases of CRISP-DM are still a good description for the analytics process, but the details and specifics need to be updated. CRISP-DM does not seem to be maintained and adapted to the challenges of Big Data and modern data science.

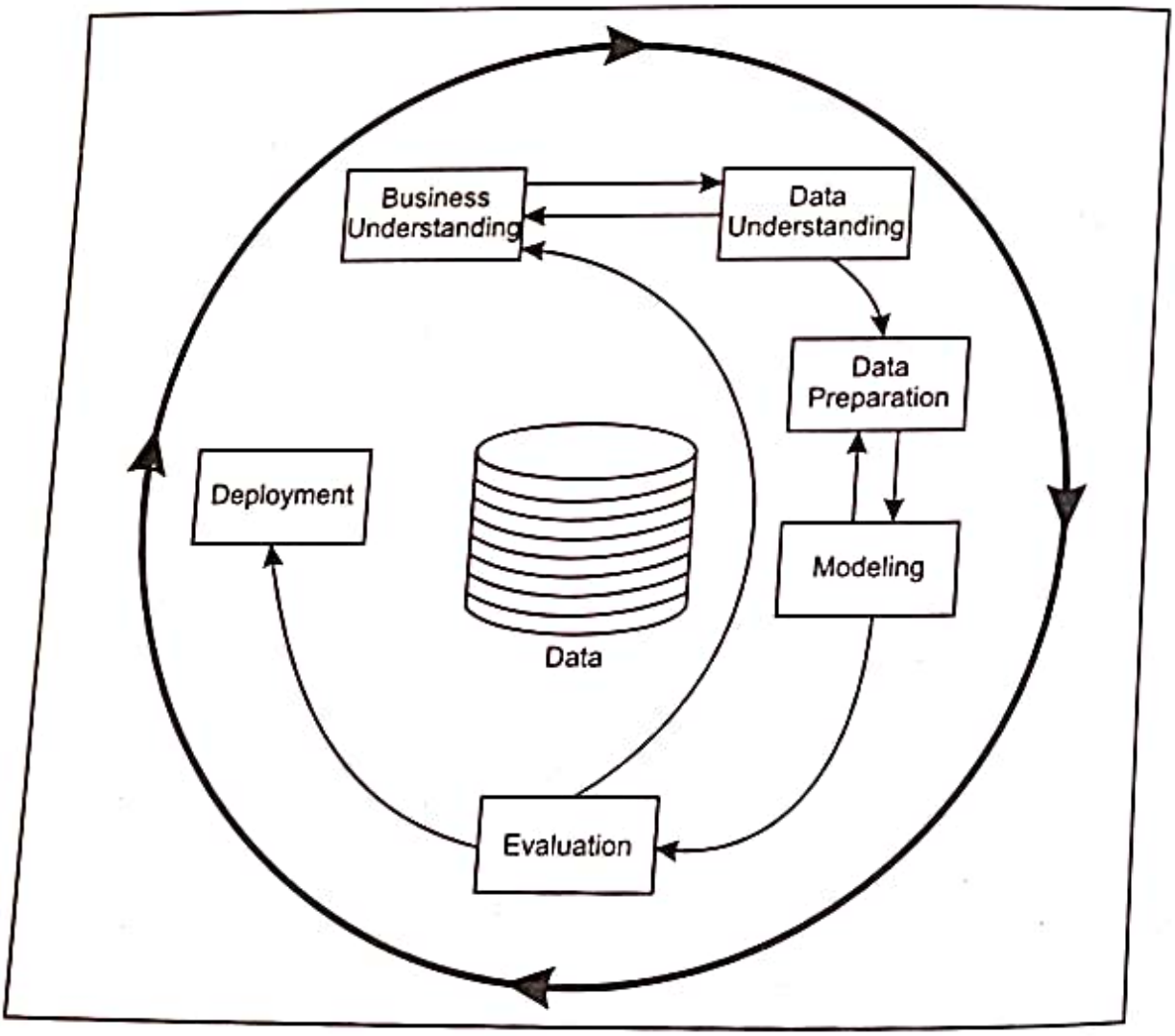

Fig. 1.6: The 6 high-level phases of CRISP-DM suggested for the Data Science Projects

The lifecycle outlines the major stages that projects typically execute, often iteratively:

- Business Understanding
- Data Acquisition and Understanding
- Modeling
- Deployment
- Customer Acceptance

Fig. 1.7 provides a visual representation of the Team Data Science Process lifecycle.

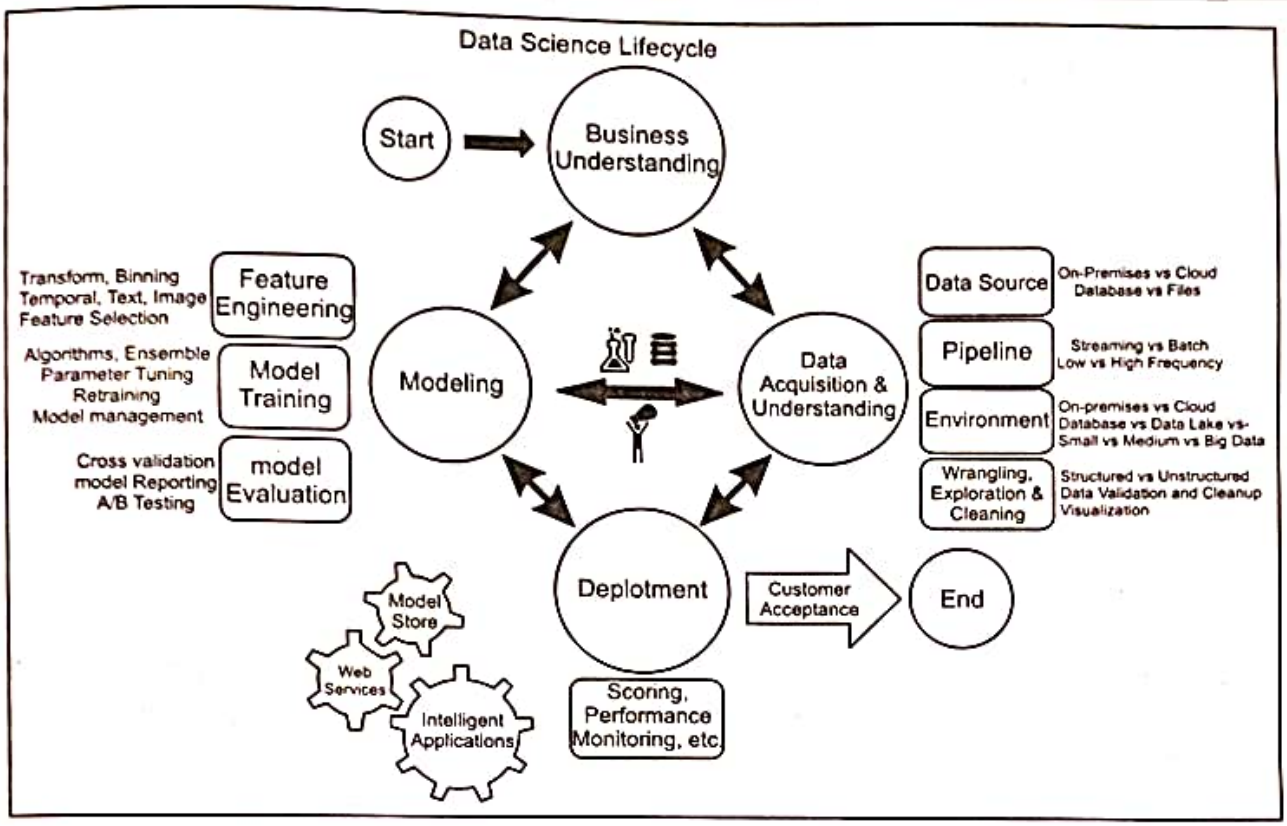

Fig. 1.7: Life Cycle of Data Science Process (Courtesy Microsoft Azure)

1.8 POPULAR DATA SCIENCE TOOLKITS

Tools are an important element of the data science field. The open source community has been contributing to the data science toolkit for years which has led to major advancements to the field. There has been debate in the data science community about the use of open source technology surpassing proprietary software offered by players such as IBM and Microsoft. In fact, many of the big enterprises have started to contribute to open source solutions so they can stay top of mind for users and the data science toolkit has increasingly become one dominated by open source tools.

Since there are a wide variety of open source tools available from data-mining platforms to programming languages, we put together a mix of technology that data scientists could add to their data science toolkit.

Naina Sethi of Spring Board[6] has described such tools in her article. Since there are a wide variety of open source tools available from data-mining platforms to programming languages, we put together a mix of technology that data scientists could add to their data science toolkit. All such tools listed here are free software.

1. **R Programming Language :** R is built by data Scientists for data scientist. R is a programming language used for data manipulation and graphics. Originating in 1995, this is a popular tool used among data scientists and analysts. It is the open source version of the S language widely used for research in statistics. According to data scientists, R is one of the easier languages to learn as there are numerous packages and guides available for users.

R has a steep learning curve and is generally built for stand alone systems. Although there are several packages to speed up the process.

If you are a beginner, I would strongly recommend downloading RStudio which is the de facto IDE for R.

Doing data analysis, building models, communicating results are the core strengths.

The major power of R is it's user community which offers extensive support and has developed the package base CRAN.

A few great packages for you to start exploring in R would be

1. ggplot2/ggvis – Data Visualization
2. dplyr (Data Munging and Wrangling)
3. data.table (Data Wrangling)
4. Caret: (Machine learning workbench)
5. reshape2: (Data Shaping)

We have used this language extensively in Unit 3.

2. **Python :** Python is another widely used language among data scientists, created by Dutch programmer Guido Van Rossem. It's a general-purpose programming language, focusing on readability and simplicity. If you are not a programmer but are looking to learn, this is a great language to start with. It's easier than other general-purpose languages and there are a number of tutorials available for non-programmers to learn. You can do all sorts of tasks such as sentiment analysis or time series analysis with Python, a very versatile general-purpose programming language. You can canvass open data sets and do things like sentiment analysis of Twitter accounts.

   Often the type of problem your solving has a bearing on the choice of language. If the nature of the problem at hand is to do thorough data analysis then I choose R, but If I need to write quick scripts to get things done, scrape the web then it is simpler to use Python.

   According to the data science survey conducted by O'Reilly almost 40% of the data scientists use Python to solve their problems. Python also has a great community of open source packages.

   The learning course about python is given in detail in unit 5 of this book.

3. **KNIME :** KNIME is a software company with headquarters in major tech hubs around the world. The company offers an open source analytics platform written in Java, used for data reporting, mining and predictive analysis. This base platform can be advanced with a suite of commercial extensions offered by the company, including collaboration, productivity and performance extensions.
4. **SQL :** Structured Query Language or SQL is a special-purpose programming language for data stored in relational databases. SQL is used for more basic data analysis and can perform tasks such as organizing and manipulating data or retrieving data from a database. Since SQL has been used by organizations for decades, there is a large SQL ecosystem in existence already which data scientists can tap into. Among data science tools, it ranks as one of the best at filtering and selecting through databases.
5. **Apache Hadoop and other Big data tools :** Apache Hadoop software library is a

framework, written in Java, for processing large and complex datasets. The base modules for the Apache Hadoop framework include Hadoop Common, Hadoop Distributed File System (HDFS), Hadoop Yarn and Hadoop MapReduce.

(*a*) **Apache Mahout:** Apache Mahout is an environment for building scalable machine learning algorithms. The algorithms are written on top of Hadoop. Mahout implements three major machine learning tasks: collaborative filtering, clustering and categorization.

(*b*) **Apache Spark:** Apache Spark is a cluster-computing framework for data analysis. It has been deployed in large organizations for its big data capabilities combined with speed and ease of use. It was originally developed at the University of California as Spark and later, the source code was donated to the Apache Foundation so that it could be free forever. It's often preferred to other big data tools due to its speed.

(*c*) **Impala:** Impala is the massive parallel processing (MPP) database for Apache Hadoop. It's used by data scientists and analysts allowing them to perform SQL queries for data stored in Apache Hadoop clusters.

(*d*) **Apache Storm:** Apache Storm is a computational platform for real-time analytics. It's often compared to Apache Spark and is known as a better streaming engine than Spark. It's written in the Clojure programming language and is known to be a simple, easy to use tool.

(*e*) **MongoDB:** MongoDB is a NoSQL database known for its scalability and high performance. It provides a powerful alternative to traditional databases and makes the integration of data in specific applications easier. It can be an integral part of the data science toolkit if you're looking to build large-scale web apps.

6. **D3 Data Science Tools :** D3 is a javascript library for building interactive data visualizations within your browser. It allows data scientists to create rich visualizations with a high level of customizability. It's a great addition to your data science toolkit if you're looking to dynamically express your data insights.

7. **Tensor Flow :** Tensor Flow is the product of Google's Brain Team coming together for the purpose of advancing machine learning. It's a software library for numerical computation and built for everyone from students and researchers to hackers and innovators. It allows programmers to access the power of deep learning without needing to understand some of the complicated principles behind it, and ranks as one of the data science tools that helps make deep learning accessible for thousands of companies. TensorFlow is an open-source software library for dataflow programming across a range of tasks. It is a symbolic math library, and also used for machine learning applications such as neural networks.

   TensorFlow was developed by the Google Brain team for internal Google use. It was released under the Apache 2.0 open source license on November 9, 2015.

   Keras is a deep learning library written in Python. It runs on TensorFlow allowing for fast experimentation. Keras was developed to make deep learning models easier and helping users treat their data intelligently in an efficient manner.

   We hope you've got some new data science tools for your data science toolkit in this article! Comment below if you can think of any more.

8. **Rstudio :** Rstudio integrates with R as an IDE (Integrated Developmet Environment) to provide further functionality. RStudio combines a source code editor, build automation tools and a debugger.

1.9 FAMILIARITY WITH EXAMPLE APPLICATIONS

1. **Airline Route Planning:** Airline Industry across the world is known to bear heavy losses. Except a few airline service providers, companies are struggling to maintain their occupancy ratio and operating profits. With high rise in air fuel prices and need to offer heavy discounts to customers has further made the situation worse. It wasn't for long when airlines companies started using data science to identify the strategic areas of improvements. Now using data science, the airline companies can:
   1. Predict flight delay
   2. Decide which class of airplanes to buy
   3. Whether to directly land at the destination, or take a halt in between (For example: A flight can have a direct route from New Delhi to New York. Alternatively, it can also choose to halt in any country.)
   4. Effectively drive customer loyalty programs

   Southwest Airlines, Alaska Airlines are among the top companies who've embraced data science to bring changes in their way of working.
2. **Fraud and Risk Detection:** One of the first applications of data science originated from Finance discipline. Companies were fed up of bad debts and losses every year. However, they had a lot of data which use to get collected during the initial paper work while sanctioning loans. They decided to bring in data science practices in order to rescue them out of losses. Over the years, banking companies learned to divide and conquer data via customer profiling, past expenditures and other essential variables to analyze the probabilities of risk and default. Moreover, it also helped them to push their banking products based on customer's purchasing power.
3. **Delivery logistics:** Who says data science has limited applications? Logistic companies like DHL, FedEx, UPS, Kuhne+Nagel have used data science to improve their operational efficiency. Using data science, these companies have discovered the best routes to ship, the best suited time to deliver, the best mode of transport to choose thus leading to cost efficiency, and many more to mention. Further more, the data that these companies generate using the GPS installed, provides them a lots of possibilities to explore using data science.
4. **Uber's Taxy Service:** Uber is a smartphone-app based taxi booking service which connects users who need to get somewhere with drivers willing to give them a ride. The service has been hugely controversial, due to regular taxi drivers claiming that it is destroying their livelihoods, and concerns over the lack of regulation of the company's drivers. This hasn't stopped it from also being hugely successful – since being launched to purely serve San Francisco in 2009, the service has been expanded to many major cities on every continent except for Antarctica.