UNIT-1
DATA SCIENCE CONCEPT

1.1 What is Data Science?

Data science is the study of data. It involves developing methods of recording, storing, and analyzing data to effectively extract useful information. It is the process of deriving knowledge and insights from a huge and diverse set of data through organizing, processing and analysing the data. The goal of data science is to gain insights and knowledge from any type of data — both structured and unstructured. Data science is related to computer science, but is a separate field. Computer science involves creating programs and algorithms to record and process data, while data science covers any type of data analysis, which may or may not use computers. It involves many different disciplines like mathematical and statistical modelling, extracting data from its source and applying data visualization techniques. Data Science is a blend of various tools,

algorithms and machine learning principles with the goal to discover hidden patterns from the raw data. It is more closely related to the mathematics field of Statistics, which includes the collection, organization, analysis, and presentation of data. Data science continues to evolve as one of the most promising and in-demand career paths for skilled professionals.

1.2 The Data Science Life Cycle

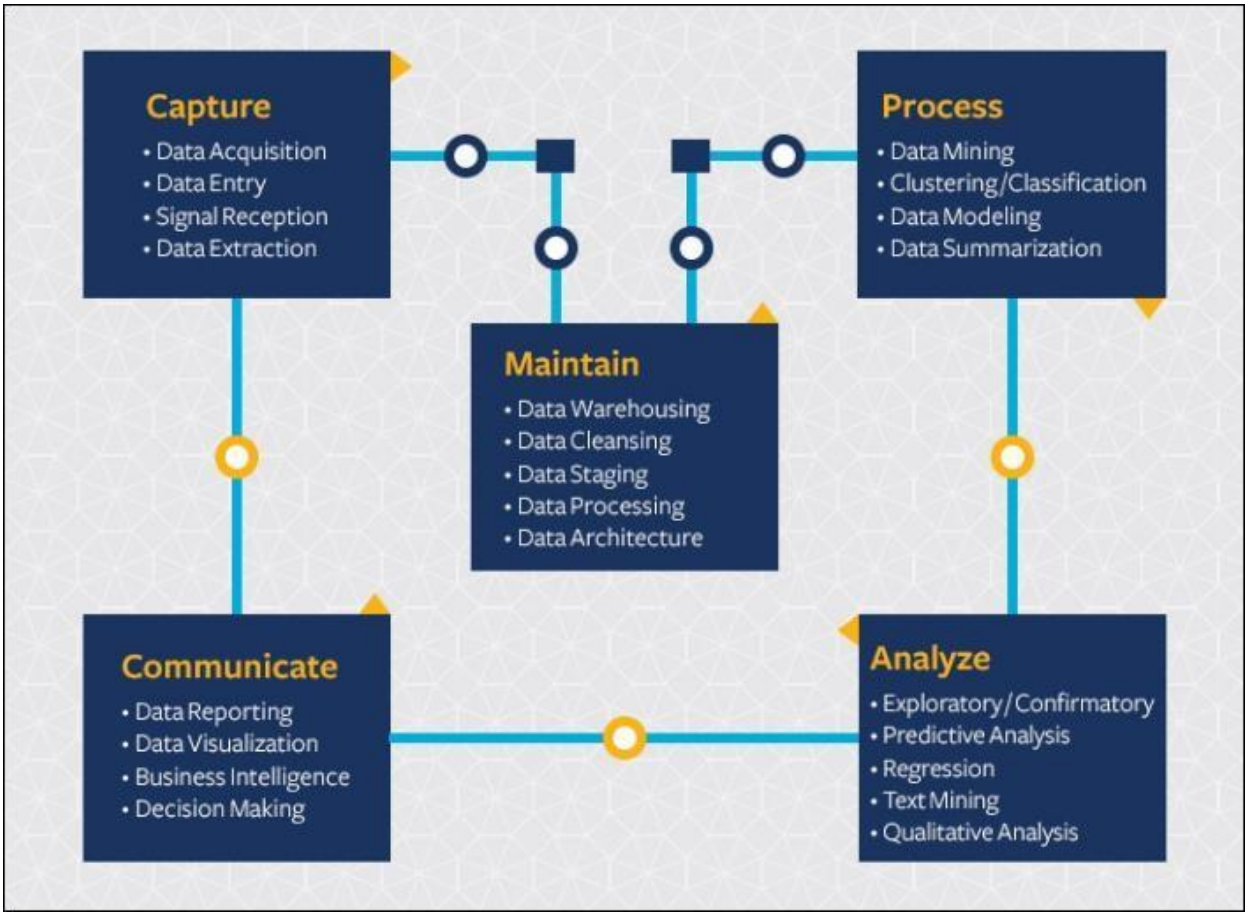

Fig. 1.1 Lifecycle of Data Science

Effective data scientists are able to identify relevant questions, collect data from a multitude of different data sources, organize the information, translate results into solutions, and communicate their findings in a way that positively affects business decisions. These skills are required in almost all industries, causing skilled data scientists to be increasingly valuable to companies.

1.3 Data

Anything that is recorded is data. Observations and facts are data. Anecdotes and opinions are also data, of a different kind. Data can be numbers, like the record of daily weather, or daily sales. Data can be alphanumeric, such as the names of employees and customers.

1. Data could come from any number of sources. It could come from operational records inside an organization, and it can come from records compiled by the industry bodies and government agencies. Data could come from individuals telling stories from memory and from people's

interaction in social contexts. Data could come from machines reporting their own status or from logs of web usage.

2. Data can come in many ways. It may come as paper reports. It may come as a file stored on a computer. It may be words spoken over the phone. It may be e-mail or chat on the Internet. It may come as movies and songs in DVDs, and so on.

3. There is also data about data. It is called metadata. For example, people regularly upload videos on YouTube. The format of the video file (whether it was a high-def file or lower resolution) is metadata. The information about the time of uploading is metadata. The account from which it was uploaded is also metadata. The record of downloads of the video is also metadata.

1.4 Types of Data and Variables:

Data can be of different types. There are two types of variables that can be found in your data – numerical and categorical. Numerical data can be divided into continuous or discrete values.

1.4.1 Numerical data

Numerical data is the information that is measurable, and it is, of course, data represented as numbers and not words or text. Continuous numbers are numbers that don't have a logical end to them. Examples include variables that represent money or height. Discrete numbers are the opposite; they have a logical end to them. These may be defines as:

- Data may have discrete numeric values defined in a certain range, with the assumption of equal distance between the values. Customer satisfaction score may be ranked on a 10-point scale with 1 being lowest and 10 being highest. This requires the respondent to carefully calibrate the entire range as objectively as possible and place his own measurement in that scale. This is called interval (equal intervals) data.

- The highest level of numeric data is ratio data which can take on any numeric value. The weights and heights of all employees would be exact numeric values. The price of a shirt will also take any numeric value. It is called ratio (any fraction) data.

1.4.2 Categorical data

Categorical data can be broken down into nominal and ordinal values. Ordinal values are values that have a set order to them. Nominal values are the opposite of ordinal values, and they represent values with no set order to them. These may be defined as:

- Data could be an unordered collection of values. For example, a retailer sells shirts of red, blue, and green colors. There is no intrinsic ordering among these color values. One can hardly argue that any one color is higher or lower than the other. This is called nominal (means names) data. For example:

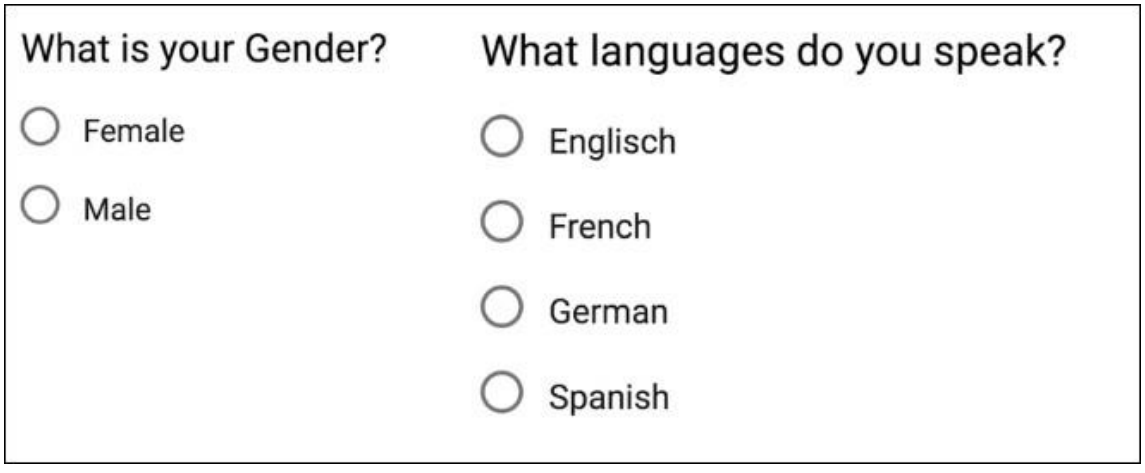

Fig. 1.2 Example of Nominal Data

- Data could be ordered values like small, medium and large. For example, the sizes of shirts could be extra-small, small, medium, and large. There is clarity that medium is bigger than small, and large is bigger than medium. But the differences may not be equal. This is called ordinal (ordered) data. For example:

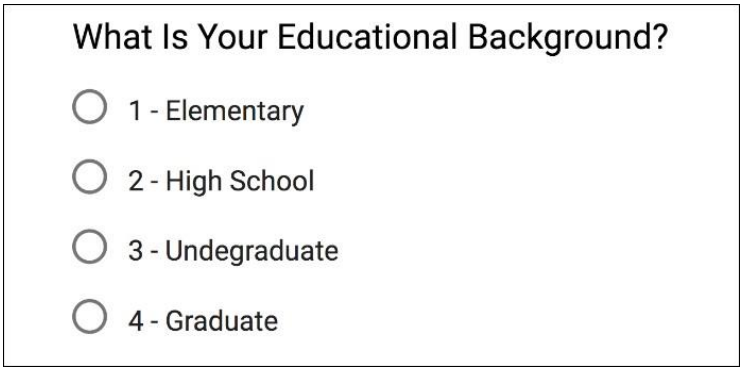

Fig. 1.3 Example of Ordinal Data

In addition to ordinal and nominal values, there is a special type of categorical data called binary. Binary data types only have two values – yes or no. This can be represented in different ways such as "True" and "False" or 1 and 0. Binary data is used heavily for classification machine learning models. Examples of binary variables can include whether a person has stopped their subscription service or not, or if a person bought a car or not.

1.4.3 Binary Large Objects

There is another kind of data that does not lend itself to much mathematical analysis, at least not directly. Such data needs to be first structured and then analyzed. This includes data like audio, video, and graphs files, often called BLOBs (Binary Large Objects). These kinds of data lend themselves to different forms of analysis and mining. Songs can be described as happy or sad, fast-paced or slow, and so on. They may contain sentiment and intention, but these are not quantitatively precise.

The precision of analysis increases as data becomes more numeric. Ratio data could be subjected to rigorous mathematical analysis. For example, precise weather data about temperature, pressure, and humidity can be used to create rigorous mathematical models that can accurately predict future weather. Data may be publicly available and sharable, or it may be marked private. Traditionally, the law allows the right to privacy concerning one's personal data. There is a big debate on whether the personal data shared on social media conversations is private or can be used for commercial purposes.

1.4.4 Variables:

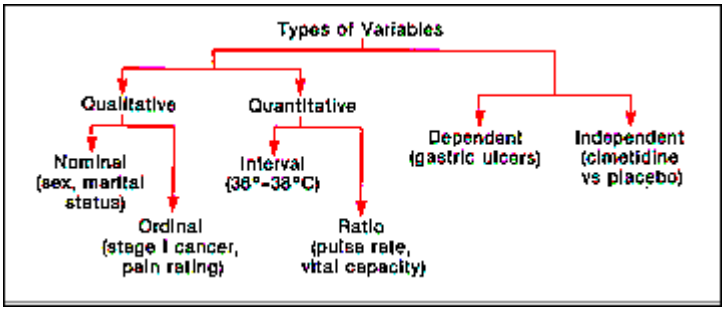

Fig. 1.4 Categorization of Variables

Independent Variables are those things which are changed in the experiment and the **Dependent variable** changes as a result of change in independent variable. For example: If the amount of water received by a plant changes, the height of the plant will change. In this statement, the water is the independent variable, because it is being changed and height of the plant is the dependent variable because it is affected by the water.

1.5 Database

A database is a modelled collection of data that is accessible in many ways. A data model can be designed to integrate the operational data of the organization. The data model abstracts the key entities involved in an action and their relationships. Most databases today follow the relational data model and its variants. Each data modelling technique imposes rigorous rules and constraints to ensure the integrity and consistency of data over time.

Take the example of a sales organization. A data model for managing customer orders will involve data about customers, orders, products, and their interrelationships. The relationship between the customers and orders would be such that one customer can place many orders, but one order is associated with one and only one customer. It is called a one-to-many relationship. The relationship between orders and products is a little more complex. One order may contain many products. And one product may be contained in many different orders. This is called a many-to-many relationship. Different types of relationships can be modelled in a database.

Databases have grown tremendously over time. They have grown in complexity in terms of number of the objects and their properties being recorded. They have also grown in the quantity of data being stored. A decade ago, a terabyte-sized database was considered big. Today databases

are in petabytes and exabytes. Video and other media files have greatly contributed to the growth of databases. E-commerce and other web-based activities also generate huge amounts of data. Data generated through social media has also generated large databases. The e-mail archives, including attached documents 20 of organizations, are in similar large sizes.

Many database management software systems (DBMSs) are available to help store and manage this data. These include commercial systems, such as Oracle and DB2 system. There are also open-source, free DBMS, such as MySQL. These DBMSs help process and store millions of transactions worth of data every second.

1.6 Data Analytics

The data analytics process has some key components that are needed for any initiative. By combining these components, a successful data analytics initiative will provide a clear picture of where you are, where you have been and where you should go. Data analytics is a broad field. There are four primary types of data analytics: descriptive, diagnostic, predictive and prescriptive analytics. Each type has a different goal and a different place in the data analysis process. These are also the primary data analytics applications in business.

- Descriptive analytics helps answer questions about **what happened**. These techniques summarize large datasets to describe outcomes to stakeholders. By developing key performance indicators (KPIs,) these strategies can help track successes or failures. This process requires the collection of relevant data, processing of the data, data analysis and data visualization. This process provides essential insight into past performance.

- Diagnostic analytics helps answer questions about **why things happened**. These techniques supplement more basic descriptive analytics. They take the findings from descriptive analytics and dig deeper to find the cause. The performance indicators are further investigated to discover why they got better or worse. This generally occurs in three steps:

    - Identify anomalies in the data.
    - Data that is related to these anomalies is collected.
    - Statistical techniques are used to find relationships and trends that explain these anomalies.

- Predictive analytics helps answer questions about **what will happen in the future**. These techniques use historical data to identify trends and determine if they are likely to recur. Predictive analytical tools provide valuable insight into what may happen in the future and its techniques include a variety of statistical and machine learning techniques, such as: neural networks, decision trees, and regression.
- Prescriptive analytics helps answer questions about **what should be done**. By using insights from predictive analytics, data-driven decisions can be made. This allows businesses to make informed decisions in the face of uncertainty. Prescriptive analytics techniques rely on machine learning strategies that can find patterns in large datasets. By analyzing past decisions and events, the likelihood of different outcomes can be estimated.

1.7 Wholeness of Data Analytics

Business is the act of doing something productive to serve someone's needs, and thus earn a living and make the world a better place. Business activities are recorded on paper or using electronic media, and then these records become data. There is more data from customers' responses and on the industry as a whole. All this data can be analyzed and mined using special tools and techniques to generate patterns and intelligence, which reflect how the business is functioning. These ideas can then be fed back into the business so that it can evolve to become more effective and efficient in serving customer needs and the cycle continues on.

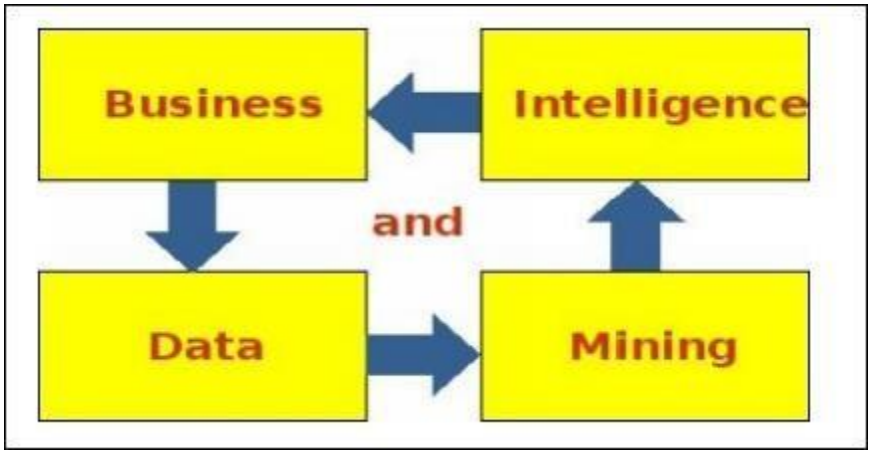

Fig. 1.5 Business Intelligence and Data Mining Cycle

1.7.1 Business Intelligence

Any business organization needs to continually monitor its business environment and its own performance, and then rapidly adjust its future plans. This includes monitoring the industry, the competitors, the suppliers, and the customers. The organization needs to also develop a balanced scorecard to track its own health and vitality. Executives typically determine what they want to track based on their key performance Indexes (KPIs) or key result areas (KRAs). Customized reports need to be designed to deliver the required information to every executive. These reports can be converted into customized dashboards that deliver the information rapidly and in easy-to grasp formats.

Business intelligence is a broad set of information technology (IT) solutions that includes tools for gathering, analyzing, and reporting information to the users about performance of the organization and its environment. These IT solutions are among the most highly prioritized solutions for investment.

Consider a retail business chain that sells many kinds of goods and services around the world, online and in physical stores. It generates data about sales, purchases, and expenses from multiple locations and time frames. Analyzing this data could help identify fast-selling items, regional-selling items, seasonal items, fast-growing customer segments, and so on. It might also help generate ideas about what products sell together, which people tend to buy which products, and so

on. These insights and intelligence can help design better promotion plans, product bundles, and store layouts, which in turn lead to a better-performing business.

The vice president of sales of a retail company would want to track the sales to date against monthly targets, the performance of each store and product category, and the top store managers that month. The vice president of finance would be interested in tracking daily revenue, expense, and cash flows by store; comparing them against plans; measuring cost of capital; and so on.

1.7.2 Pattern Recognition

A pattern is a design or model that helps grasp something. Patterns help connect things that may not appear to be connected. Patterns help cut through complexity and reveal simpler understandable trends. Patterns can be as definitive as hard scientific rules, like the rule that the sun always rises in the east. They can also be simple generalizations, such as the Pareto principle, which states that 80 percent of effects come from 20 percent of the causes.

A perfect pattern or model is one that:

(a) accurately describes a situation
(b) is broadly applicable
(c) can be described in a simple manner.

Very often, all three qualities are not achievable in a single model, and one has to settle for two of three qualities in the model. Patterns can be temporal, which is something that regularly occurs over time. Patterns can also be spatial, such as things being organized in a certain way. Patterns can be functional, in that doing certain things leads to certain effects. Good patterns are often symmetric. They echo basic structures and patterns that we are already aware of.

A temporal rule would be that “some people are always late,” no matter what the occasion or time. Some people may be aware of this pattern and some may not be. A spatial pattern, following the 80–20 rule, could be that the top 20 percent of customers lead to 80 percent of the business. Or 20 percent of products generate 80 percent of the business. Or 80 percent of incoming customer service calls are related to just 20 percent of the products. A functional pattern may involve test-taking skills. Some students perform well on essay-type questions. Others do well in multiple-choice questions. Yet other students excel in doing hands-on projects, or in oral presentations.

1.8 Data Processing Chain

Data is the new natural resource. Implicit in this statement is the recognition of hidden value in data. Data lies at the heart of business intelligence. There is a sequence of steps to be followed to benefit from the data in a systematic way. Data can be modeled and stored in a database. Relevant data can be extracted from the operational data stores according to certain reporting and analyzing purposes, and stored in a data warehouse. The data from the warehouse can be combined with other sources of data, and mined using data mining techniques to generate new insights. The insights need to be visualized and communicated to the right audience in real time for competitive advantage. Figure 1.6 explains the progression of data processing activities. The rest of this chapter will cover these five elements in the data processing chain.

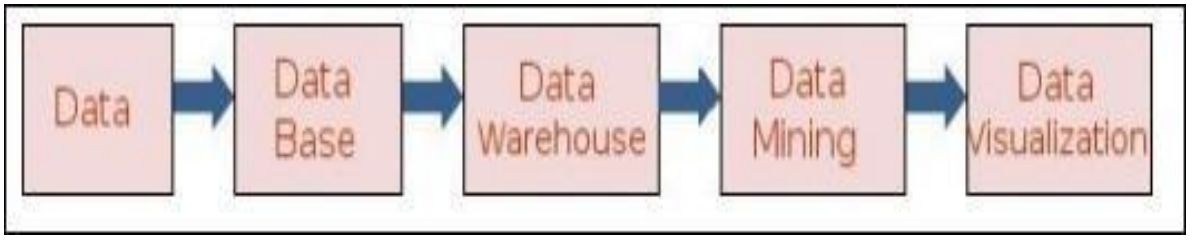

Fig. 1.6 Data Processing Chain

1.8.1 Data: (Refer Topic 1.3 and 1.4)

Datafication is a new term that means that almost every phenomenon is now being observed and stored. More devices are connected to the Internet. More people are constantly connected to "the grid," by their phone network or the Internet, and so on. Every click on the web, and every movement of the mobile devices, is being recorded. Machines are generating data. The "Internet of things" is growing faster than the Internet of people. All of this is generating an exponentially growing volume of data, at high velocity. Kryder's law predicts that the density and capability of hard drive storage media will double every 18 months. As storage costs keep coming down at a rapid rate, there is a greater incentive to record and store more events and activities at a higher resolution. Data is getting stored in more detailed resolution, and many more variables are being captured and stored.

1.8.2 Database (Refer Topic 1.5)

Here is a simple database of the sales of movies worldwide for a retail organization. It shows sales transactions of movies over three quarters. Using such a file, data can be added, accessed, and updated as needed.

Movies Transaction Database

Order #	Date Sold	Product Name	Location	Amount in dollars
1	April 2018	Monty Python	US	9
2	May 2018	Gone with the Wind	US	15
3	June 2018	Monty Python	India	9
4	June 2018	Monty Python	UK	12
5	July 2018	Matrix	US	12
6	July 2018	Monty Python	US	12

7	July 2018	Gone with the Wind	US	15
8	August 2018	Matrix	US	12
9	September 2018	Matrix	India	12
10	September 2018	Monty Python	US	9
11	September 2018	Gone with the Wind	US	15
12	September 2018	Monty Python	India	9
13	November 2018	Gone with the Wind	US	15
14	December 2018	Monty Python	US	9
15	December 2018	Monty Python	US	9

Table 1.1 Example of a Database

1.8.3 Data Ware House:

Data Warehouse A data warehouse is an organized store of data from all over the organization, specially designed to help make management decisions. Data can be extracted from operational database to answer a particular set of queries. This data, combined with other data, can be rolled up to a consistent granularity and uploaded to a separate data store called the data warehouse. Therefore, the data warehouse is a simpler version of the operational data base, with the purpose of addressing reporting and decision-making needs only. The data in the warehouse cumulatively grows as more operational data becomes available and is extracted and appended to the data warehouse. Unlike in the operational database, the data values in the warehouse are not updated.

To create a simple data warehouse for the movies sales data, assume a simple objective of tracking sales of movies and making decisions about managing inventory. In creating this data warehouse, all the sales transaction data will be extracted from the operational data files. The data will be rolled up for all combinations of time period and product number. Thus, there will be one row for every combination of time period and product. The resulting data warehouse will look like the table as follows:

Row #	**Qtr. Sold**	**Product Name**	**Amount in Dollars**
1	Q2	Gone with the Wind	15
2	Q2	Monty Python	30

3	Q3	Gone with the Wind	30
4	Q3	Matrix	36
5	Q3	Monty Python	30
6	Q4	Gone with the Wind	15
7	Q4	Monty Python	18

Table 1.2 Example of a Data Warehouse

The data in the data warehouse is at much less detail than the transaction database. The data warehouse could have been designed at a lower or higher level of detail, or granularity. If the data warehouse were designed on a monthly level, instead of a quarterly level, there would be many more rows of data. When the number of transactions approaches millions and higher, with dozens of attributes in each transaction, the data warehouse can be large and rich with potential insights. One can then mine the data (slice and dice) in many different ways and discover unique meaningful patterns. Aggregating the data helps improve the speed of analysis. A separate data warehouse allows analysis to go on separately in parallel, without burdening the operational database systems. For Example:

Function	**Database**	**Data Warehouse**
Purpose	Data stored in databases can be used for many purposes including day-to-day operations.	Data stored in DW is cleansed data useful for reporting and analysis.
Granularity	Highly granular data including all activity and transaction details.	Lower granularity data; rolled up to certain key dimensions of interest.
Complexity	Highly complex with dozens or hundreds of data files, linked through common data fields.	Typically organized around a large fact tables, and many lookup tables.
Size	Database grows with growing volumes of activity and transactions. Old completed transactions are deleted to reduce size.	Grows as data from operational databases is rolled-up and appended every day. Data is retained for long-term trend analysis.
Architectural Choices	Relational, and object-oriented, databases	Star schema, or Snowflake schema
Data Access Mechanism	Primarily through high level languages such as SQL. Traditional programming access DB through	Accessed through SQL; SQL output is forwarded to reporting tools and data visualization tools

	Open Database Connectivity (ODBC) interfaces

Table 1.3 Comparing Database Systems with Data Warehouse Systems

1.8.4 Data Mining

Data Mining is the art and science of discovering useful innovative patterns from data. There is a wide variety of patterns that can be found in the data. There are many techniques, simple or complex, that help with finding patterns. In this example, a simple data analysis technique can be applied to the data in the data warehouse above. A simple cross-tabulation of results by quarter and products will reveal some easily visible patterns.

Movies Sales by Quarters – Cross-tabulation (amount in dollars)

Qtr/Product	Gone With the Wind	Matrix	Monty Python	Total Sales Amount
Q2	15	0	30	45
Q3	30	36	30	96
Q4	15	0	18	33
Total Sales Amount	60	36	78	174

Table 1.4 Example of a Data Mining

Based on the cross-tabulation above, one can readily answer some product sales questions, like:

1. What is the best selling movie by revenue? – Monty Python.
2. What is the best quarter by revenue this year? – Q3
3. Any other patterns? – Matrix movie sells only in Q3 (seasonal item).

These simple insights can help plan marketing promotions and manage inventory of various movies. If a cross tabulation was designed to include customer location data, one could answer other questions, such as

1. What is the best-selling geography? – US
2. What is the worst selling geography? – UK
3. Any other patterns? – Monty Python sells globally, while Gone with the Wind sells only in the US.

If the data mining was done at the monthly level of data, it would be easy to miss the seasonality of the movies. However, one would have observed that September is the highest selling month.

The previous example shows that many differences and patterns can be noticed by analyzing data in different ways. However, some insights are more important than others. The value of the insight depends upon the problem being solved. The insight that there are more sales of a product in a

certain quarter helps a manager plan what products to focus on. In this case, the store manager should stock up on Matrix in Quarter 3 (Q3). Similarly, knowing which quarter has the highest overall sales allows for different resource decisions in that quarter. In this case, if Q3 is bringing more than half of total sales, this requires greater attention on the e-commerce website in the third quarter.

Data mining should be done to solve high-priority, high-value problems. Much effort is required to gather data, clean and organize it, mine it with many techniques, interpret the results, and find the right insight. It is important that there be a large expected payoff from finding the insight. One should select the right data (and ignore the rest), organize it into a nice and imaginative framework that brings relevant data together, and then apply data mining techniques to deduce the right insight.

A retail company may use data mining techniques to determine which new product categories to add to which of their stores; how to increase sales of existing products; which new locations to open stores in; how to segment the customers for more effective communication; and so on.

Data can be analyzed at multiple levels of granularity and could lead to a large number of interesting combinations of data and interesting patterns. Some of the patterns may be more meaningful than the others. Such highly granular data is often used, especially in finance and high-tech areas, so that one can gain even the slightest edge over the competition.

Here are brief descriptions of some of the most important data mining techniques used to generate insights from data.

Decision Trees: They help classify populations into classes. It is said that 70% of all data mining work is about classification solutions; and that 70% of all classification work uses decision trees. Thus, decision trees are the most popular and important data mining technique. There are many popular algorithms to make decision trees. They differ in terms of their mechanisms and each technique work well for different situations. It is possible to try multiple decision-tree algorithms on a data set and compare the predictive accuracy of each tree.

Regression: This is a well-understood technique from the field of statistics. The goal is to find a best fitting curve through the many data points. The best fitting curve is that which minimizes the (error) distance between the actual data points and the values predicted by the curve. Regression models can be projected into the future for prediction and forecasting purposes.

Artificial Neural Networks: Originating in the field of artificial intelligence and machine learning, ANNs are multi-layer non-linear information processing models that learn from past data and predict future values. These models predict well, leading to their popularity. The model's parameters may not be very intuitive. Thus, neural networks are opaque like a black-box. These systems also require a large amount of past data to adequate train the system.

Cluster analysis: This is an important data mining technique for dividing and conquering large data sets. The data set is divided into a certain number of clusters, by discerning similarities and

dissimilarities within the data. There is no one right answer for the number of clusters in the data. The user needs to make a decision by looking at how well the number of clusters chosen fit the data. This is most commonly used for market segmentation. Unlike decision trees and regression, there is no one right answer for cluster analysis.

Association Rule Mining: Also called Market Basket Analysis when used in retail industry, these techniques look for associations between data values. An analysis of items frequently found together in a market basket can help cross-sell products, and also create product bundles.

1.8.5 Data Visualization

As data and insights grow in number, a new requirement is the ability of the executives and decision makers to absorb this information in real time. There is a limit to human comprehension and visualization capacity. That is a good reason to prioritize and manage with fewer but key variables that relate directly to the Key Result Areas (KRAs) of a role.

Here are few considerations when presenting using data:

1. Present the conclusions and not just report the data.
2. Choose wisely from a palette of graphs to suit the data.
3. Organize the results to make the central point stand out.
4. Ensure that the visuals accurately reflect the numbers. Inappropriate visuals can create misinterpretations and misunderstandings.
5. Make the presentation unique, imaginative and memorable.

Executive dashboards are designed to provide information on select few variables for every executive. They use graphs, dials, and lists to show the status of important parameters. These dashboards also have a drill-down capability to enable a root-cause analysis of exception situations (Fig. 1.7)

Fig. 1.7 Sample Executive Dashboards

Data visualization has been an interesting problem across the disciplines. Many dimensions of data can be effectively displayed on a two-dimensional surface to give a rich and more insightful description of the totality of the story.

1.9 Data Distribution

Having a sound statistical background can be greatly beneficial in the daily life of a Data Scientist. Every time we start exploring a new dataset, we need to first do an Exploratory Data Analysis (EDA) in order to get a feeling of what are the main characteristics of certain features.

If we are able to understand that if any pattern is present in the data distribution, we can then tailor-made our Machine Learning models to best fit our case study. In this way, we will be able to get a better result in less time. In fact, some Machine Learning models are designed to work best under some distribution assumptions. Therefore, knowing with which distributions we are working with can help us to identify which models are best to use.

Every time we are working with a dataset, our dataset represent a **sample** from a **population**. Using this sample, we can then try to understand it's main patterns so that we can use it to make predictions on the whole population (even though we never had the opportunity to examine the whole population). Let's imagine we want to predict the price of a house given a certain set of features. We might be able to find online a dataset with all the house prices of San Francisco (our sample) and after performing some statistical analysis, we might be able to make quite accurate predictions of the house price in any other city in the USA (our population).

Datasets are composed of two main types of data: **Numerical** (eg. integers, floats),and **Categorical** (eg. names, laptops brands). Numerical data can additionally be divided into othertwo categories: **Discrete** and **Continue**.

Discrete data can take only certain values (eg. number of students in a school) while continuous data can take any real or fractional value (eg. the concepts of height and weights). From discrete random variables, it is possible to calculate **Probability Mass Functions**, while from continuous random variables can be derived **Probability Density Functions**.

Probability Mass Functions gives the probability that a variable can be equal to a certain value, instead, the values of Probability Density Functions are not itself probabilities because they need first to be integrated over the given range. There exist many different probability distributions(consider Fig. 1.8), out of which commonly used distributions in Data Science are explained as follows:

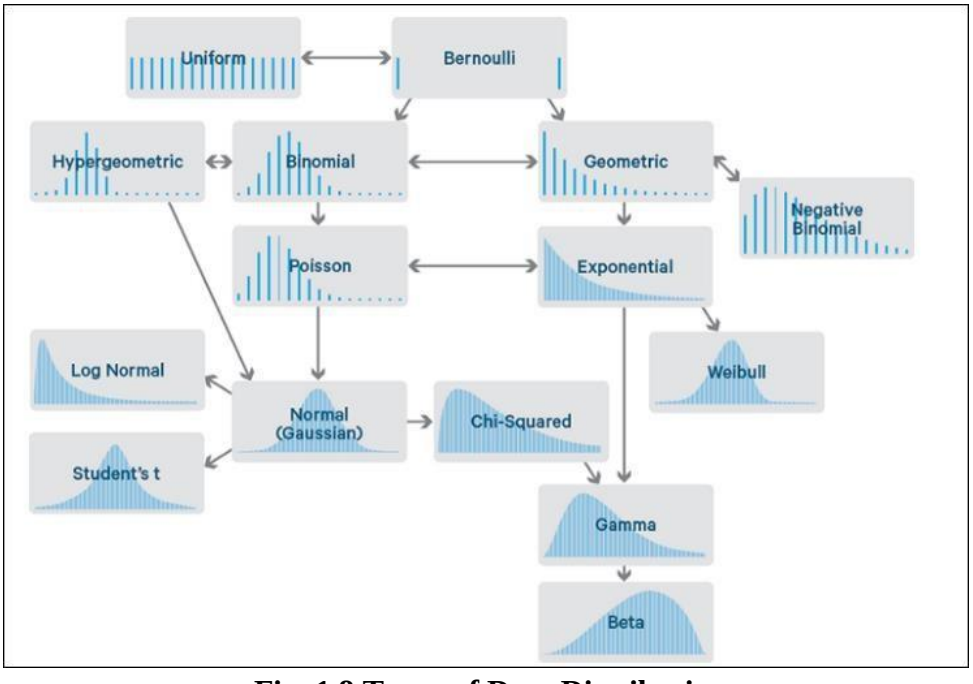

Fig. 1.8 Types of Data Distribution

1.9.1 Bernoulli Distribution

The Bernoulli distribution is one of the easiest distributions to understand and can be used as a starting point to derive more complex distributions. This distribution has only two possible outcomes and a single trial. A simple example can be a single toss of a biased/unbiased coin. In this example, the probability that the outcome might be heads can be considered equal to **p** and **(1 - p)** for tails (the probabilities of mutually exclusive events that encompass all possible outcomes needs to sum up to one).

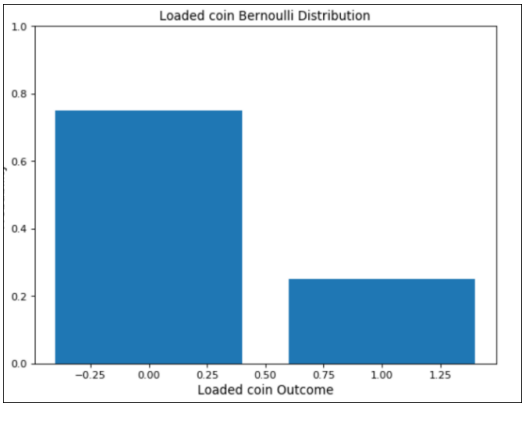

Fig. 1.9 Bernoulli distribution biased coin

1.9.2. Uniform Distribution

The Uniform Distribution can be easily derived from the Bernoulli Distribution. In this case, a possibly unlimited number of outcomes are allowed and all the events hold the same probability

to take place. As an example, imagine the roll of a fair dice. In this case, there are multiple possible events with each of them having the same probability to happen.

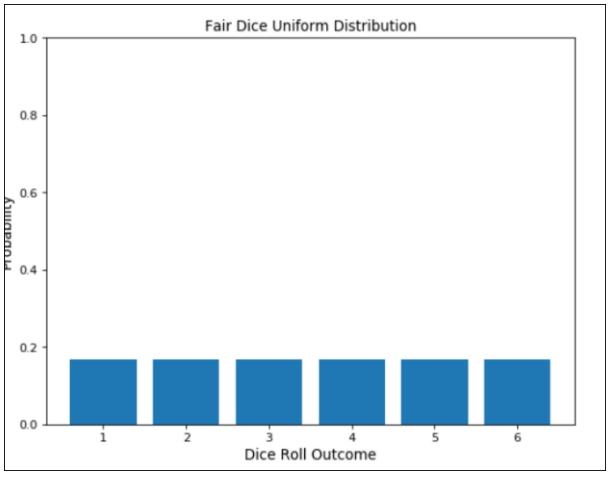

Fig. 1.10 Fair Dice Roll Uniform distribution

1.9.3 Binomial Distribution

The Binomial Distribution can instead be thought as the sum of outcomes of an event following a Bernoulli distribution. The Binomial Distribution is therefore used in binary outcome events and the probability of success and failure is the same in all the successive trials. This distribution takes two parameters as inputs: the number of times an event takes place and the probability assigned to one of the two classes.

A simple example of a Binomial Distribution in action can be the toss of a biased/unbiased coin repeated a certain amount of times. Varying the amount of bias will change the way the distribution will look like (Figure 1.11).

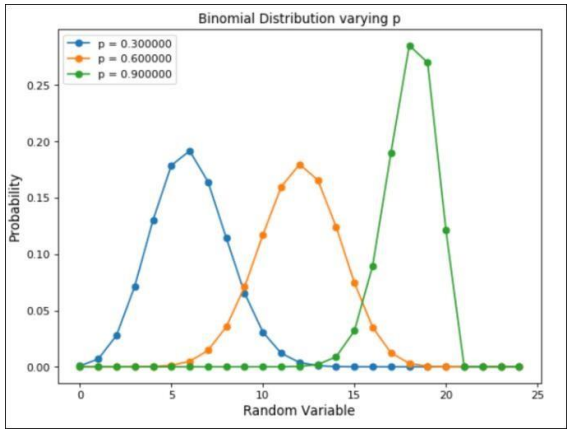

Fig. 1.11 Binomial Distribution varying event occurrence probability

The main characteristics of a Binomial Distribution are:

- Given multiple trials, each of them is independent of each other (the outcome of one trial doesn't affect another one).

- Each trial can lead to just two possible results (eg. winning or losing), which have probabilities p and (1 - p).

If we are given the probability of success (p) and the number of trials (n), we can then be able to calculate the probability of success (x) within these n trials using the formula below

$$P\left(X = x\right) = \frac{n!}{x!(n-x)!}\, p^x \left(1 - p\right)^{(n-x)}$$

1.9.4 Normal (Gaussian) Distribution

The Normal Distribution is one of the most used distributions in Data Science. Many common phenomena that take place in our daily life follows Normal Distributions such as: the income distribution in the economy, student's average reports, the average height in populations, etc… In addition to this, the sum of small random variables also turns out to usually follow a normal distribution (Central Limit Theorem).

"In probability theory, the **central limit theorem** (**CLT**) establishes that, in some situations, when independent random variables are added, their properly normalized sum tends toward a normal distribution even if the original variables themselves are not normally distributed."

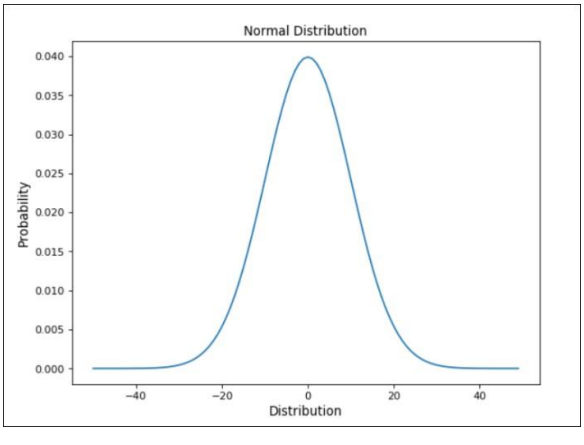

Fig. 1.12 Gaussian Distribution

Some of the characteristics which can help us to recognise a normal distribution are:

- The curve is symmetric at the centre. Therefore mean, mode and median are all equal to the same value, making distribute all the values symmetrically around the mean.
- The area under the distribution curve is equal to 1 (all the probabilities must sum up to 1).

When using Normal Distributions, the distribution mean and standard deviation plays a really important role. If we know their values, we can then easily find out the probability of predicting exact values by just examining the probability distribution (Figure 1.13). In fact, thanks to the distribution properties, 68% of the data lies within one standard deviation of the mean, 95% within two standard deviations of the mean and 99.7% within three standard deviations of the mean:

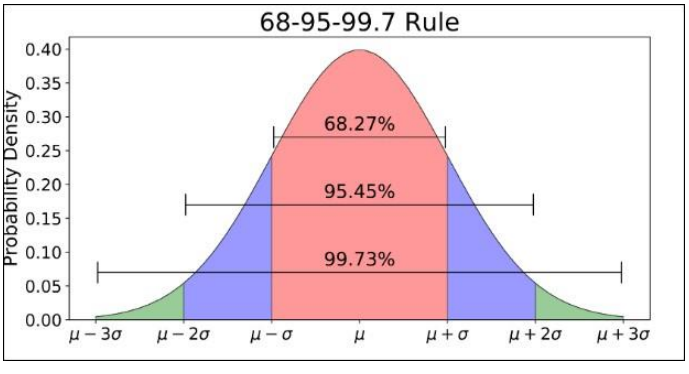

Fig. 1.13 Normal Distribution 68-95-99.7 Rule

Many Machine Learning models are designed to work best-using data that follow a Normal Distribution. Some examples are:

- Gaussian Naive Bayes Classifier
- Linear Discriminant Analysis
- Quadratic Discriminant Analysis
- Least Squares based regression models

Additionally, it is also possible in some cases to transform not-normal data into a normal form by applying transformations such as logarithms and square roots.

1.9.5 Poisson Distribution

Poisson Distributions are commonly used to find the probability that an event might happen or not knowing how often it usually occurs. Additionally, Poisson Distributions can also be used to predict how many times an event might occur in a given time period. For example: Poisson Distributions are frequently used by insurance companies to conduct risk analysis (eg. predict the number of car crash accidents within a predefined time span) to decide car insurance pricing.

When working with Poisson Distributions, we can be confident of the average time between the occurrence of different events, but the precise moment an event might take place is randomly spaced in time.
A Poisson Distribution can be modelled using the following formula where **λ** represents the expected number of events which can take place in a period.

$$P(X) = \frac{\lambda^x e^{-\lambda}}{X!}$$

The main characteristics which describe Poisson Processes are:

1. The events are independent of each other (if an event happens, this does not alter the probability that another event can take place).
2. An event can take place any number of times (within the defined time period).
3. Two events can't take place simultaneously.
4. The average rate between events occurrence is constant.

In Figure 1.14, is shown how varying the expected number of events which can take place in a period (λ) can change a Poisson Distribution.

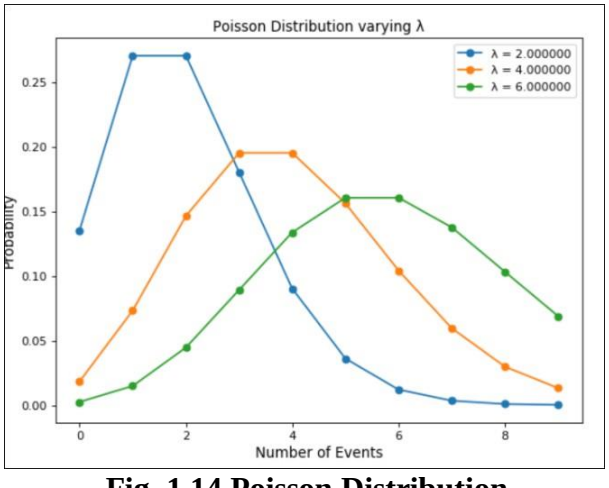

Fig. 1.14 Poisson Distribution

1.9.6 Exponential Distribution

As an example, let's imagine we work at a restaurant and we want to predict what is going to be the time interval between different customers coming to the restaurant. Using an Exponential Distribution for this type of problem, could be the perfect place where to start.
Another common application of Exponential distributions is survival analysis (eg. expected life of a device/machine). Exponential distributions are regulated by a parameter λ. The greater the value of λ and the faster the exponential curve is going to decade (Figure 1.15).

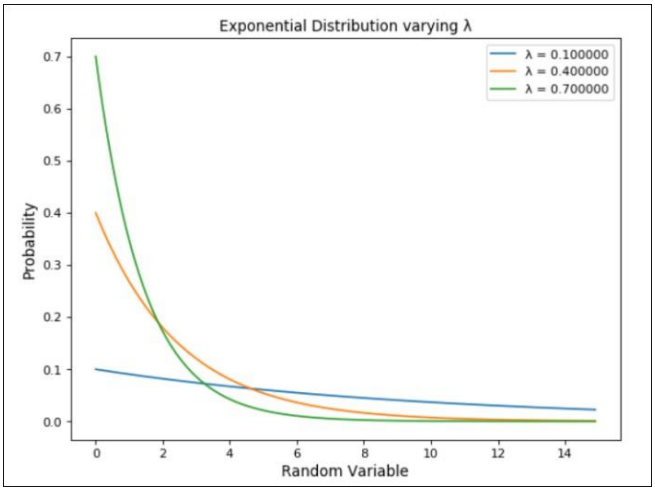

Fig. 1.14 Exponential Distribution

1.10 Advices for New Data Scientists

It is intended primarily for data scientists embedded in product teams, but many of the tips can be generalized to any new hire in a tech role.

1. Prioritization,
2. Estimating how long tasks will take,
3. How to get your questions answered, and
4. Communicating & sharing your work.

1.10.1 Prioritization

To work well with PMs, a new data scientist needs some understanding of what they do. In the beginning, pay close attention to how the work is being prioritized: there's no better way to learn and understand what is important to them — and the business. One should always have an ongoing document in which he log his work. Having an overview of what's on his plate not only allows him to easily see how he spend his time across various types of work (big bets, small bets, ad-hoc work, infrastructure, etc) but also makes having performance conversations much easier. If he has a sense that much of the work serves short-term goals or projects, he will have the data to substantiate it. If this is the case, he should communicate it to the manager as part of his job to make sure the time is well spent. This practice also provides stakeholders and business partners with visibility into his workflow.

1.10.2 Estimating how long tasks will take

As a new data scientist, it is crucial to understand the differences between:

- How long he want something to take (how desirable is the task relative to your other priorities and interests),
- How long it should take, (how feasible is it due to tooling, infrastructure, logging, etc), and
- How long it actually takes to just do it, given desire and reality.

PMs are not data scientists and it is not their job to evaluate different analytic approaches — it's the data scientist. As mentioned above, it is their job to provide a framework for delivering good product. The best PMs should have frameworks that are clear and consistent, particularly when it comes to data products: to them, perfect and opaque are (always) the enemy of the good. If one solution takes twice as long as another, is quite complicated to implement, or is a black box, one need to be very clear and convincing about why it should be preferred. And his convincing should rarely, if ever, include words like 'AUC' or 'gradient descent'. Always focus on business impact and characterize the various data products/solutions you build in those terms!

1.10.3 How to get your questions answered

New Data scientist should describe in great detail what is he doing, and ask for very specific advice: 'How do I transform the data in this particular way?', 'How do I use [this specific tool] to do [this very specific thing]'. He totally gets the impulse here: He is demonstrating that he is trying, and has invested in a solution. But what he may have missed is that the solution he has already honed in on is likely one of many. When he seeks advice on only a particular implementation, he has narrowed the path forward. When seeking help (from anyone, really) always start with the goal; this opens him up to a wider range of inputs.

1.10.4 Communicating & Sharing your work

If he works on an embedded product team, communication is one of, if not the most important aspects of his job. The most powerful advice for junior data scientists in these regards is the importance of communicating at different altitudes. For most communications with those outside of the data org, it's not the Appendix that they are interested in; it's the TL;DR (*Too Long; Don't Read*). In practice, this often means that it is not his job to tell business partners how much work he did or how hard it was or what the various model evaluation measures were — save these discussions for the manager and peers. If a PM asks a question about the users, answer it as simply as possible, within reason. Do not hide the response in a maze of technical details- he will lose people this way! If they have questions (which they always should) they will follow up.
The more he works with someone, the more he will be able to anticipate their follow-ups. But do not assume that they are as interested in the path he took to get there. The Business Partners need to deliver product and he needs to help them get there.
Be sure to get them to specify and document those deadlines when the project is committed. If he is getting close to a deadline and know he will miss it, communicate it proactively. This is a sign of maturity, not failure.

1.11 Introduction to Cloud

Cloud computing is an Internet-based computing in which the shared pool of resources are available over a broad network access, these resources can be provisioned or released with minimum management efforts and service provider interaction.

There are four types of cloud:
1. Public cloud
2. Private cloud
3. Hybrid cloud

4. Community cloud

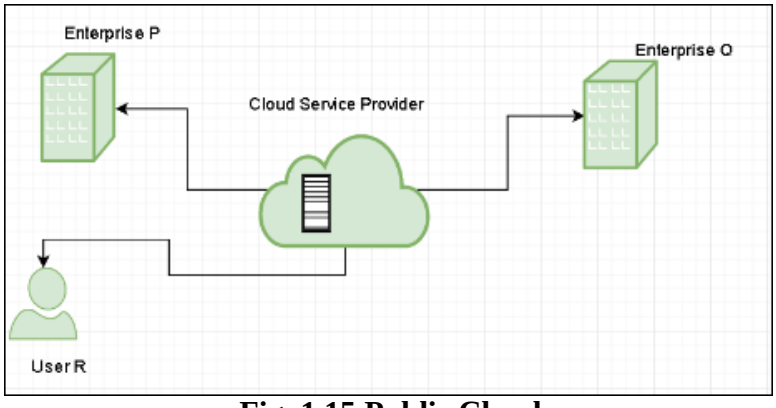

Fig. 1.15 Public Cloud

1.11.1Public-Cloud

Public cloud are managed by third parties which provide cloud services over the internet to public, these services are available as pay-as-you-go billing mode. They offer solutions for minimizing IT infrastructure costs and act as a good option for handling peak loads on the local infrastructure. They are a goto option for small enterprises, which are ableto start their businesses without large upfront investments by completely relying on public infrastructure for their IT needs. A fundamental characteristic of public clouds is multitenancy. Apublic cloud is meant to serve multiple users, not a single customer. A user requires a virtual computing environment that is separated, and most likely isolated, from other users.

1.11.2Private-Cloud

Private clouds are distributed systems that work on a private infrastructure and providing the users with dynamic provisioning of computing resources. Instead of a pay-as-you-go model as in public clouds, there could be other schemes in that take into account the usage of the cloud and proportionally billing the different departments or sections of an enterprise.

The advantages of using a private cloud are:

1. **Customer information protection**: In private cloud security concerns are less since customer data and other sensitive information does not flow out of a private infrastructure.

2. **Infrastructure ensuring SLAs**: Private cloud provides specific operations such as appropriate clustering, data replication, system monitoring and maintenance, and disaster recovery, and other uptime services.

3. **Compliance with standard procedures and operations**: Specific procedures have to be put in place when deploying and executing applications according to third-party compliance standards.This is not possible in case of public cloud.

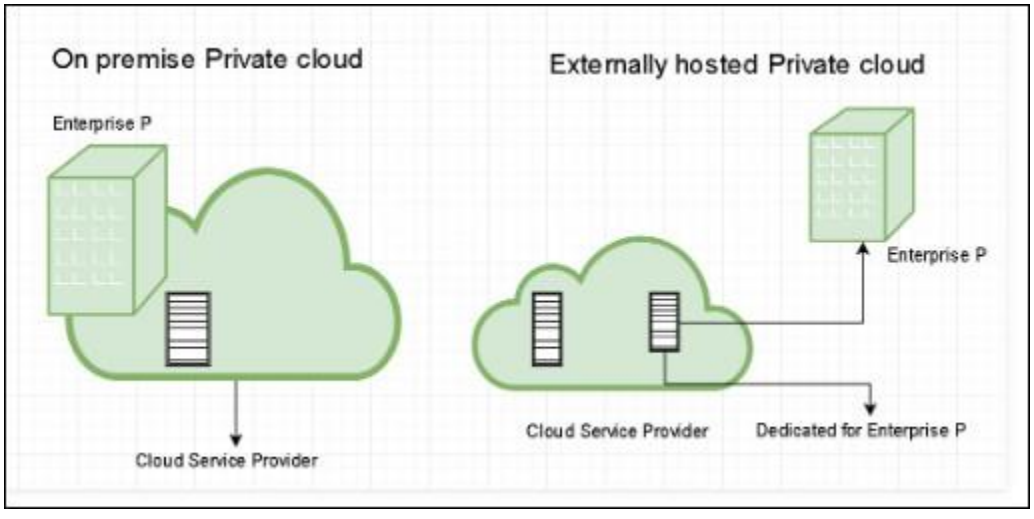

Fig. 1.16 Private Cloud

1.11.3Hybrid-Cloud

Hybrid cloud is a heterogeneous distributed system resulted by combining facilities of public cloud and private cloud. For this reason they are also called heterogeneous clouds. A major drawback of private deployments is the inability to scale on demand and to efficiently address peak loads. Here public clouds are needed. Hence, a hybrid cloud takes advantages of both public and private cloud.

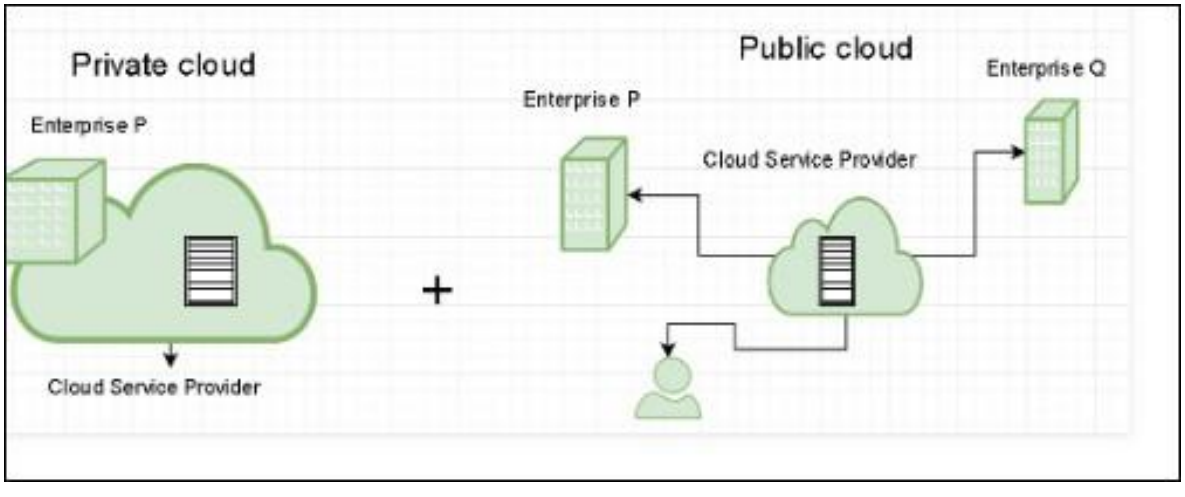

Fig. 1.17 Hybrid Cloud

1.11.4Community-Cloud

Community clouds are distributed systems created by integrating the services of different clouds to address the specific needs of an industry, a community, or a business sector.In community cloud, the infrastructure is shared between organization which has shared concernsor tasks. The cloud may be managed by an organization or a third party.

Sectors that use community clouds are:

1. **Media industry:** Media companies are looking for quick, simple, low-cost way for increasing efficiency of content generation. Most media productions involve an extended ecosystem of partners. In particular, the creation of digital content is the outcome of a collaborative process that includes movement of large data, massive compute-intensive rendering tasks, and complex workflow executions.
2. **Healthcare industry**: In healthcare industry community clouds are used to share information and knowledge on the global level with sensitive data in the private infrastructure.
3. **Energy and core industry**: In these sectors, the community cloud is used to cluster set of solution which collectively addresses management, deployment, and orchestration of services and operations.
4. **Scientific research**: In this organization with common interests of science share large distributed infrastructure for scientific computing.

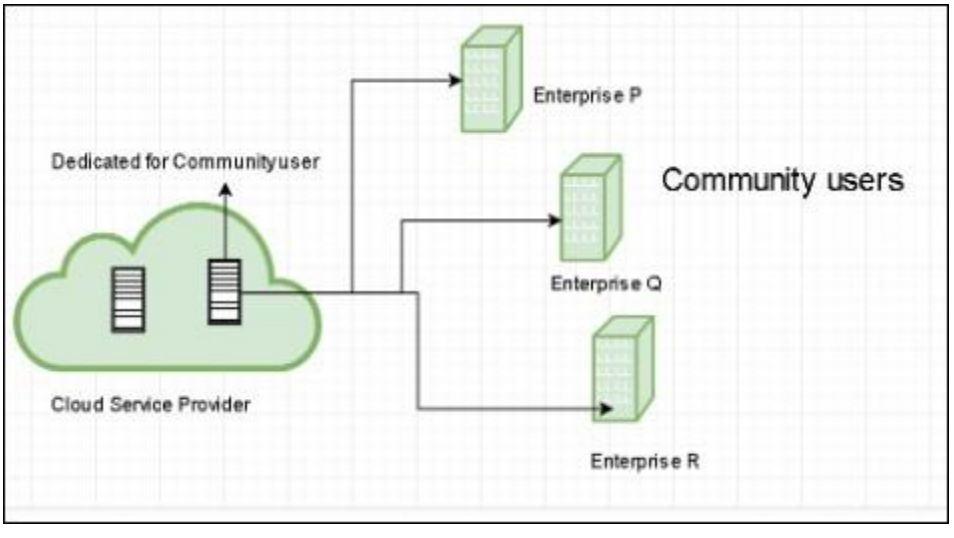

Fig. 1.18 Community Cloud

1.12 Cloud Based Services

Cloud Computing can be defined as the practice of using a network of remote servers hosted on the Internet to store, manage, and process data, rather than a local server or a personal computer. Companies offering these computing services are called cloud providers and typically charge for cloud computing services based on usage.

Types of Cloud Services

Most cloud computing services fall into three broad categories:

1. Software as a service (Saas)
2. Platform as a service (PaaS)
3. Infrastructure as a service (IaaS)
4. Anything as a service (XaaS)

These are sometimes called the **cloud computing stack**, because they are build on top of one another. Knowing what they are and how they are different, makes it easier to accomplish your goals.

1.12.1. Software as a Service

Software-as-a-Service (SaaS) is a way of delivering services and applications over the Internet. Instead of installing and maintaining software, we simply access it via the Internet, freeing ourselves from the complex software and hardware management.It removes the need to install and run applications on our own computers or in the data centers eliminating the expenses of hardware as well as software maintenance. SaaS provides a complete software solution which you purchase on a **pay-as-you-go** basis from a cloud service provider.Most SaaS applications can be run directly from a web browser without any downloads or installations required.The SaaS applications are sometimes called **Web-based software, on-demand software, or hosted software.**

Advantages of SaaS :

1. **Cost Effective :** Pay only for what you use.

2. **Reduced time :** Users can run most SaaS apps directly from their web browser without needing to download and install any software.This reduces the time spent in installation and configuration, and can reduce the issues that can get in the way of the software deployment.

3. **Accessibility :** We can Access app data from anywhere.

4. **Automatic updates :** Rather than purchasing new software, customers rely on a SaaS provider to automatically perform the updates.

5. **Scalability :** It allows the users to access the services and features on demand.

1.12.2. Platform as a service

PaaS is a category of cloud computing that provides a platform and environment to allow developers to build applications and services over the internet. PaaS services are hosted in the cloud and accessed by users simply via their web browser. A PaaS provider hosts the hardware and software on its own infrastructure. As a result, PaaS frees users from having to install in-house hardware and software to develop or run a new application.Thus, the development and deployment of the application takes place **independent of the hardware**. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly configuration settings for the application-hosting environment.

Advantages of PaaS :

1. **Simple and convenient for users :** It provides much of the infrastructure and other IT services, which users can access anywhere via a web browser.

2. **Cost Effective :** It charges for the services provided on a per-use basis thus eliminating the expenses one may have for on-premises hardware and software.

3. **Efficiently managing the lifecycle :** It is designed to support the complete web application lifecycle: building, testing, deploying, managing and updating.

4. **Efficiency :** It allows for higher-level programming with reduced complexity thus, the overall development of the application can be more effective.
The various companies providing Platform as a service are Amazon Web services, Salesforce, Windows Azure, Google App Engine, cloud Bess and IBM smart cloud.

1.12.3. Infrastructure as a Service

Infrastructure as a service (IaaS) is a service model that delivers computer infrastructure on an outsourced basis to support various operations. Typically IaaS is a service where infrastructure is provided as an outsource to enterprises such as networking equipments, devices, database and web servers. **Infrastructure as a service (IaaS) is also known as Hardware as a service (HaaS).**IaaS customers pay on a per-use basis, typically by the hour, week or month. Some providers also charge customers based on the amount of virtual machine space they use.It simply provides the underlying operating systems, security, networking, and servers for developing such applications, services, and for deploying development tools, databases, etc.

Advantages of IaaS :

1. **Cost Effective :** Eliminates capital expense and reduces ongoing cost and IaaS customers pay on a per use basis, typically by the hour, week or month.

2. **Website hosting :** Running websites using IaaS can be less expensive than traditional web hosting.

3. **Security :** The IaaS Cloud Provider may provide better security than your existing software.

4. **Maintainence :** There is no need to manage the underlying data center or the introduction of new releases of the development or underlying software. This is all handled by the IaaS Cloud Provider.
The various companies providing Infrastructure as a service are Amazon web services, Bluestack, IBM, Openstack, Rackspace and Vmware.

1.12.4. Anything as a Service

Most of the cloud service providers now a days offer anything as a service that is a compilation of all of the above services including some additional services.

Advantages of XaaS :
All of the above advantages

1.13 Cloud Computing Architecture:

Cloud computing architecture refers to the components and sub components required for cloud computing. These component typically refer to:

1. Front end(fat client, thin client)
2. Back end platforms(servers,storage)
3. Cloud based delivery and a network(Internet, Intranet, Inter cloud).

1.14 Hosting a cloud:

There are three layers in cloud computing. Companies use these layers based on the service they provide.

- Infrastructure
- Platform
- Application

1.14.1 Benefits of Cloud Hosting

1. **Scalability:** With Cloud hosting, it is easy to grow and shrink the number and size of servers based on the need. This is done by either increasing or decreasing the resources in the cloud.This ability to alter plans due to fluctuation in business size and needs is a superb benefit of cloud computing especially when experiencing a sudden growth in demand.

2. **Instant**: Whatever you want is instantly available in the cloud.

3. **Save Money**: An advantage of cloud computing is the reduction in hardware cost. Instead of purchasing in-house equipment, hardware needs are left to the vendor. For companies that are growing rapidly, new hardware can be a large, expensive, and inconvenience. Cloud computing alleviates these issues because resources can be acquired quickly and easily. Even better, the cost of repairing or replacing equipment is passed to the vendors. Along with purchase cost, off-site hardware cuts internal power costs and saves space. Large data centers can take up precious office space and produce a large amount of heat. Moving to cloud applications or storage can help maximize space and significantly cut energy expenditures.

4. **Reliability**: Rather than being hosted on one single instances of a physical server, hosting is delivered on a virtual partition which draws its resource, such as disk space, from an extensive network of underlying physical servers. If one server goes offline it will have no effect on availability, as the virtual servers will continue to pull resource from the remaining network of servers.

5. **Physical Security:** The underlying physical servers are still housed within data centres and so benefit from the security measures that those facilities implement to prevent people accessing or disrupting them on-site.

1.15 Artificial Intelligence

Artificial Intelligence: Artificial intelligence is the study of how make computers to do things which people do better at the moment. It refers to the intelligence controlled by a computer machine.

What Is Artificial Intelligence (AI)?

Artificial intelligence (AI) refers to the simulation of human intelligence in machines that are programmed to think like humans and mimic their actions. The term may also be applied to any machine that exhibits traits associated with a human mind such as learning and problem-solving. The ideal characteristic of artificial intelligence is its ability to rationalize and take actions that have the best chance of achieving a specific goal.

Understanding Artificial Intelligence

When most people hear the term artificial intelligence, the first thing they usually think of is robots. That's because big-budget films and novels weave stories about human-like machines that wreak havoc on Earth. But nothing could be further from the truth.

Artificial intelligence is based on the principle that human intelligence can be defined in a way that a machine can easily mimic it and execute tasks, from the most simple to those that are even more complex. The goals of artificial intelligence include learning, reasoning, and perception.

As technology advances, previous benchmarks that defined artificial intelligence become outdated. For example, machines that calculate basic functions or recognize text through optimal character recognition are no longer considered to embody artificial intelligence, since this function is now taken for granted as an inherent computer function.

AI is continuously evolving to benefit many different industries. Machines are wired using a cross-disciplinary approach based in mathematics, computer science, linguistics, psychology, and more.

One View of AI is

- About designing systems that are as intelligent as humans.
- Computers can be acquired with abilities nearly equal to human intelligence.
- How system arrives at a conclusion or reasoning behind selection of actions.
- How system acts and performs not so much on reasoning process.

Why Artificial Intelligence?

- Making mistakes on real-time can be costly and dangerous.
- Time-constraints may limit the extent of learning in real world.

The AI Problem

There are some of the problems contained within AI.

1. Game Playing and theorem proving share the property that people who do them well are considered to be displaying intelligence.

2. Another important foray into AI is focused on Common sense Reasoning. It includes reasoning about physical objects and their relationships to each other, as well as reasoning about actions and other consequences.

3. To investigate this sort of reasoning Nowell Shaw and Simon built the General Problem Solver (GPS) which they applied to several common sense tasks as well as the problem of Performing symbolic manipulations of logical expressions. But no attempt was made to create a program with a large amount of knowledge about a particular problem domain. Only quite simple tasks were selected.

4. The following figure (Fig. 1.19) showing some of the tasks that are the targets of work in AI:

Perception of the world around us is crucial to our survival. Animals with much less intelligence than people are capable of more sophisticated visual perception. Perception tasks are difficult because they involve analog signals. A person who knows how to perform tasks from several of the categories shown in figure learns the necessary skills in standard order.

First perceptual, linguistic and common sense skills are learned. Later expert skills such as engineering, medicine or finance are acquired.

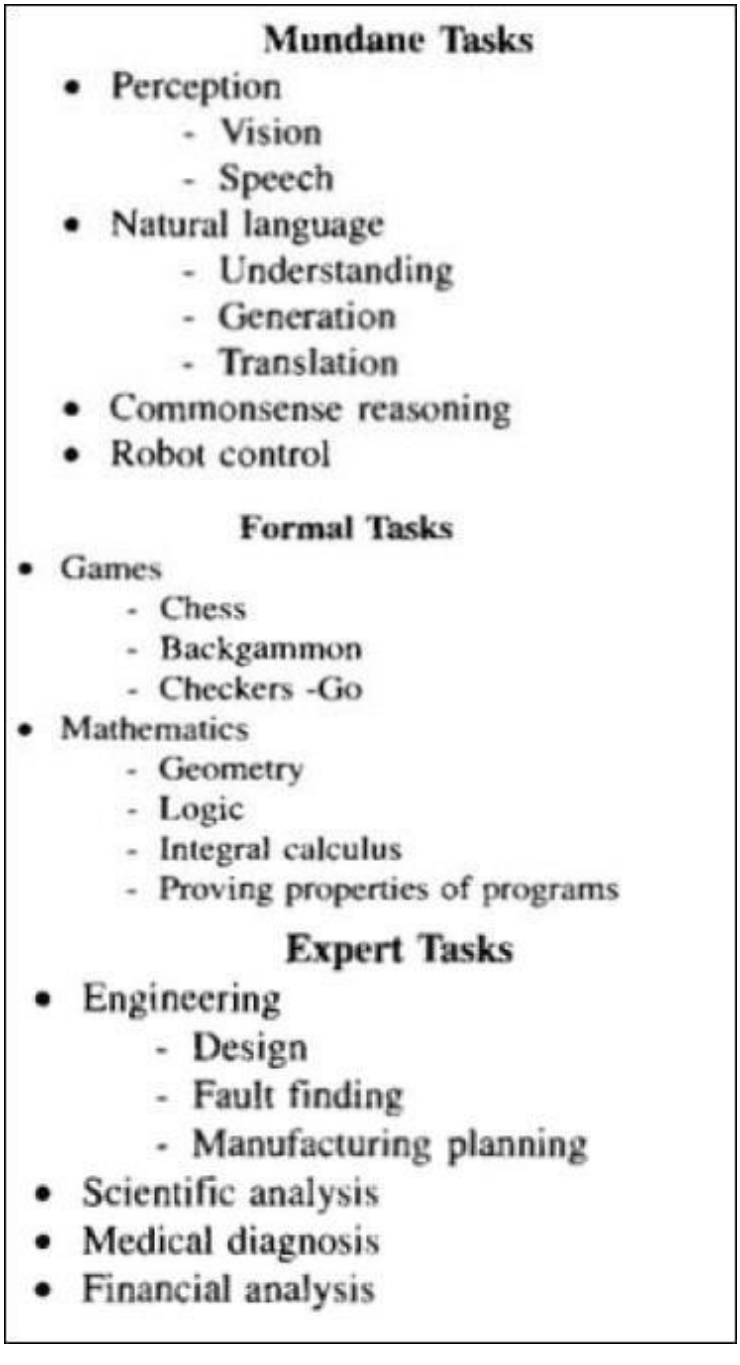

Fig. 1.19 Task domains of AI

1.16 Machine Learning

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. **Machine learning focuses on the development of computer programs** that can access data and use it learn for themselves. **Machine Learning** is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across.

The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. **The primary aim is to allow the computers learn automatically** without human intervention or assistance and adjust actions accordingly.

Machine learning algorithms are often categorized as supervised or unsupervised.

- **Supervised machine learning algorithms** can apply what has been learned in the past to new data using labelled examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly.

- In contrast, **unsupervised machine learning algorithms** are used when the information used to train is neither classified nor labelled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabelled data. The system doesn't figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabelled data.

- **Semi-supervised machine learning algorithms** fall somewhere in between supervised and unsupervised learning, since they use both labelled and unlabelled data for training – typically a small amount of labelled data and a large amount of unlabelled data. The systems that use this method are able to considerably improve learning accuracy. Usually, semi-supervised learning is chosen when the acquired labelled data requires skilled and relevant resources in order to train it / learn from it. Otherwise, acquiring unlabeled data generally doesn't require additional resources.

- **Reinforcement machine learning algorithms** is a learning method that interacts with its environment by producing actions and discovers errors or rewards. Trial and error search and delayed reward are the most relevant characteristics of reinforcement learning. This method allows machines and software agents to automatically determine the ideal behavior within a specific context in order to maximize its performance. Simple reward feedback is

required for the agent to learn which action is best; this is known as the reinforcement signal.

Machine learning enables analysis of massive quantities of data. While it generally delivers faster, more accurate results in order to identify profitable opportunities or dangerous risks, it may also require additional time and resources to train it properly. Combining machine learning with AI and cognitive technologies can make it even more effective in processing large volumes of information.

1.17 Learning Approaches:

What is learning for a machine?

A machine is said to be learning from past Experiences (data feed in) with respect to some class of Tasks, if it's Performance in a given Task improves with the Experience.

1.17.1 Supervised learning

Supervised learning as the name indicates the presence of a supervisor as a teacher. Basically supervised learning is a learning in which we teach or train the machine using data which is well labelled that means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples (data) so that supervised learning algorithm analyses the training data (set of training examples) and produces a correct outcome from labelled data.

For instance, suppose you are given a basket filled with different kinds of fruits. Now the first step is to train the machine with all different fruits one by one like this:

- If shape of object is rounded and depression at top having colour **Red** then it will be labelled as –Apple.
- If shape of object is long curving cylinder having colour **Green-Yellow** then it will be labelled as –Banana.

Now suppose after training the data, you have given a new separate fruit say Banana from basket and asked to identify it.

Since the machine has already learned the things from previous data and this time have to use it wisely. It will first classify the fruit with its shape and colour and would confirm the fruit name as BANANA and put it in Banana category. Thus the machine learns the things from training data (basket containing fruits) and then apply the knowledge to test data (new fruit).

1.17.1.1 Types of Supervised Learning

Supervised learning is when the model is getting trained on a labelled dataset. **Labelled** dataset is one which has both input and output parameters.

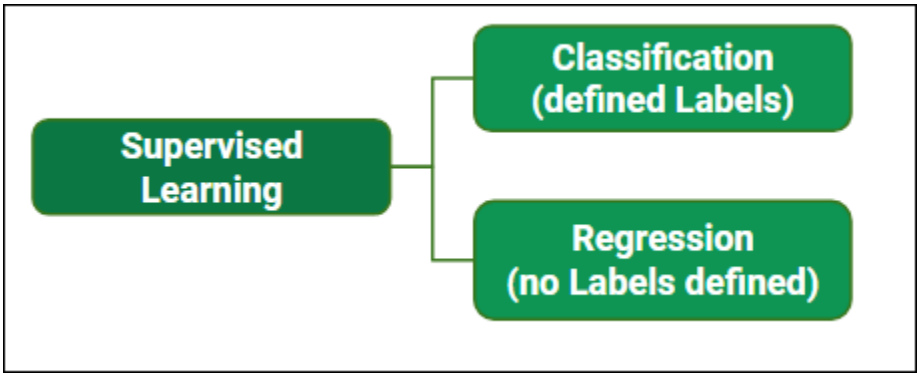

Fig. 1.20 Types of Supervised Learning

Supervised learning classified into two categories of algorithms:

1. **Classification:** A classification problem is when the output variable is a category, such as "Red" or "blue" or "disease" and "no disease". It is a Supervised Learning task where output is having defined labels(discrete value). For example in below Figure A, Output – Purchased has defined labels i.e. 0 or 1 ; 1 means the customer will purchase and 0 means that customer won't purchase.

   The goal here is to predict discrete values belonging to a particular class and evaluate on the basis of accuracy. It can be either binary or multi class classification. In binary classification, model predicts either 0 or 1 ; yes or no but in case of multi class classification, model predicts more than one class. Example: Gmail classifies mails in more than one class like social, promotions, updates, and forum.

User ID	Gender	Age	Salary	Purchased
15624510	Male	19	19000	0
15810944	Male	35	20000	1
15668575	Female	26	43000	0
15603246	Female	27	57000	0
15804002	Male	19	76000	1
15728773	Male	27	58000	1
15598044	Female	27	84000	0
15694829	Female	32	150000	1
15600575	Male	25	33000	1
15727311	Female	35	65000	0
15570769	Female	26	80000	1
15606274	Female	26	52000	0
15746139	Male	20	86000	1
15704987	Male	32	18000	0
15628972	Male	18	82000	0
15697686	Male	29	80000	0
15733883	Male	47	25000	1

Figure A: CLASSIFICATION

Both the above figures have labelled data set –

- **Figure A:** It is a dataset of a shopping store which is useful in predicting whether a customer will purchase a particular product under consideration or not based on his/ her gender, age and salary.

  Input : Gender, Age, Salary
  Output : Purchased i.e. 0 or 1 ; 1 means yes the customer will purchase and 0 means that customer won't purchase it.

- **Figure B:** It is a Meteorological dataset which serves the purpose of predicting wind speed based on different parameters.

  Input : Dew Point, Temperature, Pressure, Relative Humidity, Wind Direction
  Output : Wind Speed

2. **Regression:** A regression problem is when the output variable is a real value, such as "dollars" or "weight". It is a Supervised Learning task where output is having continuous value.

   Example in above Figure B, Output – Wind Speed is not having any discrete value but is continuous in the particular range. The goal here is to predict a value as much closer to actual output value as our model can and then evaluation is done by calculating error value. The smaller the error the greater the accuracy of our regression model.

Temperature	Pressure	Relative Humidity	Wind Direction	Wind Speed
10.69261758	986.882019	54.19337313	195.7150879	3.278597116
13.59184184	987.8729248	48.0648859	189.2951202	2.909167767
17.70494885	988.1119385	39.11965597	192.9273834	2.973036289
20.95430404	987.8500366	30.66273218	202.0752869	2.965289593
22.9278274	987.2833862	26.06723423	210.6589203	2.798230886
24.04233986	986.2907104	23.46918024	221.1188507	2.627005816
24.41475295	985.2338867	22.25082295	233.7911987	2.448749781
23.93361956	984.8914795	22.35178837	244.3504333	2.454271793
22.68800023	984.8461304	23.7538641	253.0864716	2.418341875
20.56425726	984.8380737	27.07867944	264.5071106	2.318677425
17.76400389	985.4262085	33.54900114	280.7827454	2.343950987
11.25680746	988.9386597	53.74139903	68.15406036	1.650191426
14.37810685	989.6819458	40.70884681	72.62069702	1.553469896
18.45114201	990.2960205	30.85038484	71.70604706	1.005017161
22.54895853	989.9562988	22.81738811	44.66042709	0.264133632
24.23155922	988.796875	19.74790765	318.3214111	0.329656571

Figure B: REGRESSION

1.17.1.2 Example of Supervised Learning Algorithms:

- Linear Regression
- Nearest Neighbour
- Gaussian Naive Bayes
- Decision Trees
- Support Vector Machine (SVM)
- Random Forest

1.17.2 Unsupervised Learning

Unsupervised learning is the training of machine using information that is neither classified nor labelled and allowing the algorithm to act on that information without guidance. Here the task of machine is to group unsorted information according to similarities, patterns and differences without any prior training of data.

Unlike supervised learning, no teacher is provided that means no training will be given to the machine. Therefore machine is restricted to find the hidden structure in unlabelled data by our-self. For instance, suppose it is given an image having both dogs and cats which have not seen ever.

Thus the machine has no idea about the features of dogs and cat so we can't categorize it in dogs and cats. But it can categorize them according to their similarities, patterns, and differences i.e., we can easily categorize the above picture into two parts. First: first may contain all pictures having dogs in it and second part may contain all pictures having cats in it. Here we didn't learn anything before, means no training data or examples.

1.17.2.1 Types of Unsupervised learning:

- **Clustering:** Clustering is an important concept when it comes to unsupervised learning. It mainly deals with finding a structure or pattern in a collection of uncategorized data. Clustering algorithms will process your data and find natural clusters (groups) if they exist in the data. A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behaviour.

- **Association:** An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.

1.18 Applications in the Real World

Data is the necessity of industries and therefore, Data Science has a large number of applications.

1. Banking

Banking is one of the biggest applications of Data Science. Big Data and Data Science have enabled banks to keep up with the competition. With Data Science, banks can manage their resources efficiently; furthermore, banks can make smarter decisions through fraud detection, management of customer data, risk modelling, real-time predictive analytics, customer segmentation, etc.

Banks also assess the customer lifetime value that allows them to monitor the number of customers that they have. It provides them with several predictions that the business bank will derive through their customers. In case of fraud detection, banks allow the companies to detect frauds that involve a credit card, insurance, and accounting. Banks are also able to analyse investment patterns and cycles of customers and suggest you several offers that suit you accordingly.

Furthermore, banks have the ability to **risk modelling through data science** through which they can assess their overall performance. With Data Science, banks are able to tailor personalized marketing that suits the needs of their clients. In real-time and predictive analytics, banks

use **machine learning algorithms** to improve their analytics strategy. Furthermore, banks use real-time analytics to understand underlying problems that impede their performance.

2. Finance

Data Science has played a key role in automating various financial tasks. Just like how banks have automated risk analytics, finance industries have also used data science for this task. Financial industries need to automate risk analytics in order to carry out strategic decisions for the company. Using machine learning, they identify, monitor and prioritize the risks. These machine learning algorithms enhance cost efficiency and model sustainability through training on the massively available customer data. Similarly, financial institutions use machine learning for predictive analytics. It allows the companies to predict customer lifetime value and their stock market moves. Data Science also plays a key role in algorithmic trading. Through rigorous analysis of data, financial institutions are able to make data-driven decisions. It is also playing an important role in making the customer experiences better for the users. Through extensive analysis of client experience and modification of preferences, financial institutions are able to create a personalized relationship with their customers.

This is further boosted by the real-time analytics of customers which increases the personalization. Through various customer sentiment analysis techniques and machine learning algorithms, we can boost the social media interaction, boost their feedback and analyze customer reviews. Also, the additional machine learning techniques like **natural language processing** and data mining have contributed to the transformation of information for smarter governance that helps to increase the profitability of businesses.

3. Manufacturing

In the 21st century, Data Scientists are the new factory workers. That means that data scientists have acquired a key position in the manufacturing industries. Data Science is being extensively used in manufacturing industries for optimizing production, reducing costs and boosting the profits. Furthermore, with the addition of technologies like the **Internet of Things (IoT)**, data science has enabled the companies to predict potential problems, monitor systems and analyze the continuous stream of data.

Furthermore, with data science, industries can monitor their energy costs and can also optimize their production hours.

With a thorough analysis of customer reviews, data scientists can help the industries to make better decisions and improve the quality of their products. Another important aspect of data science in industries is Automation. With the help of historical and real-time data, industries are able to develop autonomous systems that are helpful in boosting the production of manufacturing lines. It has taken away the redundant jobs and introduced powerful machines that use machine learning technologies like **reinforcement learning**.

4. Transport

Another important application of data science is transport. In the transportation sector, Data Science is actively making its mark in making safer driving environments for the drivers. It is also

playing a key role in optimizing vehicle performance and adding greater autonomy to the drivers. Furthermore, in the transport sector, Data Science has actively increased its manifold withthe **introduction of self-driving cars**.

Through extensive analysis of fuel consumption patterns, driver behaviour and active vehicle monitoring, data science has created a strong foothold in the transport industry. The self-driving cars the most trending topics in the world today. With the introduction of autonomy to vehicles through reinforcement learning, vehicle manufacturers are able to create intelligent automobiles. Furthermore, industries can create better logistical routes with the help of data science. Using a variety of variables like consumer profile, location, economic indicators, and logistics, vendors can optimize delivery routes and provide a proper allocation of resources.

Also, various transportation companies like **Uber is using data science** for price optimization and providing better experiences to their customers. Using powerful **predictive tools**, they accurately predict the price based on parameters like a weather pattern, availability of transport, customers, etc.

5. Healthcare

In the health-care industry, data science is making great leaps. The various industries in health-care making use of data science are

Medical Image Analysis
Genetics and Genomics
Drug Discovery
Predictive Modeling for Diagnosis
Health bots or virtual assistants

i. Medical Image Analysis

In the medical image analysis, data science has created a strong sphere of influence for analyzing medical images such as X-rays, MRIs, CT-Scans, etc. Previously, doctors and medical examiners would have to manually search for clues in the medical images. However, with the advancements in computing technologies and surge in data, it is possible to create machines that can automatically detect flaws in the imagery. Data Scientists have created powerful **image recognition tools** that allow doctors to have an in-depth understanding of complex medical imagery.

ii. Genomic Data Science

Genomic Data Science applies the statistical techniques to genomic sequences, allowing the bioinformaticians and geneticists to understand the defects in genetic structures. It is also helpful in classifying diseases that are genetic in nature. With data science, we can analyze how genes react to varying kinds of medicines. Also, several big data technologies like **MapReduce** have significantly reduced the processing time for genome sequencing.

iii. Drug Discovery

Another important field making use of data science is drug discovery. In drug discovery, new candidate medicines are formulated. Drug Discovery is a tedious and often complex process. Data

Science can help us to simplify this process and provide us with an early insight into the success rate of the newly discovered drug. With Machine Learning, we can also analyze several combinations of drugs and their effect on different gene structure to predict the outcome.

iv. Predictive Modeling for Diagnosis
With the advancements in predictive modeling, data scientists can help to predict the outcome of disease given the historical data of the patients. Data Science has enabled practitioners to analyzethe data, make correlations between the variables of the data and also provide insights to doctors and medical practitioners.

v. Natural Language Processing
Natural Language Processing is a technology of data science that is focused on the analysis of textual information. Using NLP, we can create intelligent bots that answer to user queries. The application of this can be extended to the healthcare sector where we can create bots that answer questions of patients and provide them with proper diagnostic guidelines.

6. E-Commerce
E-commerce and retail industries have been hugely benefitted by data science. Some of the waysin which data science has transformed the e-commerce industries are-
For identifying a potential customer base, data science is being heavily utilized.Usage of predictive analytics for forecasting the goods and services.
Data Science is also used for identifying styles of popular products and predicting their trends.With data science, companies are optimizing their pricing structures for their consumers.

Data Science is also being heavily used in collaborative filtering, where it forms the backbone of advanced recommendation system. Using this technique, the e-commerce platforms are able to provide insights to the customers based on their historical purchases and purchases made by peopleof the same style. These type of hybrid recommendation systems, consisting of both collaborative and content-based filtering are helping the industries to provide better services to their customers.

Also, companies are making use of sentiment analysis to analyze the feedbacks provided by the customers. This makes use of natural language processing to analyze texts and online surveys. Fraud Detection, which is the central role of machine learning in industries, is tailored for findingfraud merchants and frauds in wire-transfers.

UNIT -2
Introduction to Data Science Tools

2.1 A day in the life of a Data Scientist

Data science is a multidimensional field that uses scientific methods, tools, and algorithms to extract knowledge and insights from structured and unstructured data.

In reality, a Data Scientist does so much more than just studying the data. It's true that his work revolves around the data but apart from that it also involves a number of other processes based on data. This multidisciplinary field involves the systematic blend of scientific and statistical methods, processes, algorithm development and technologies to extract meaningful information from data.

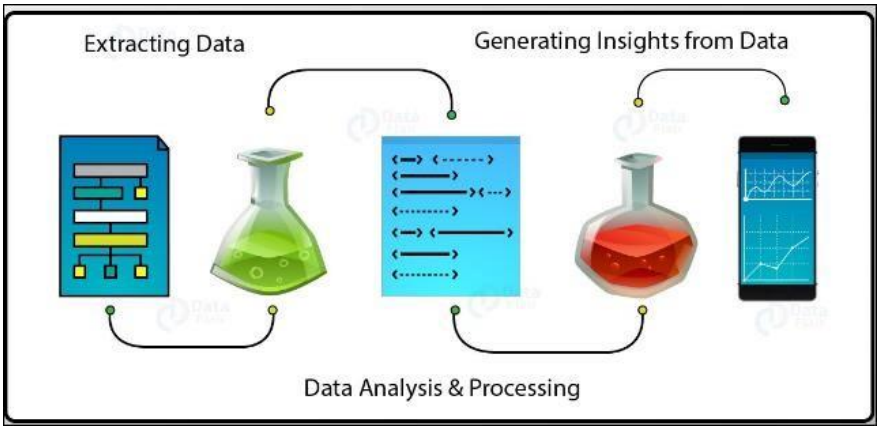

Fig. 2.1 Work of a Data Scientist

The work of a data scientist on a day-to-day basis:

- Ask questions to frame the business problem
- Get relevant data for analysis of the problem
- Explore the data to make error corrections
- Model the data for in-depth analysis
- Communicate the results of the analysis

Data Science is a vast field based on multiple fields. Day to day work is not a cake walk but knowledge of the underlying fields is needed. But, as it is rightly said that there is no such thing as free lunch, getting Data Science skills also comes with a cost.

It is further explained as follows:

1. **Understand the problem:** We start by understanding the problem, and this usually looks like sitting down with the person who owns this feature, who built this feature, and asking them: what does it do, what are you concerned about, how could this break the user experience, and what do you think this is going to do for your users? How would you characterize success, what does a happy user look like, how do we win? So, we sit down in this meeting and it is tried to build an understanding of what this feature is and then how we're going to measure it.

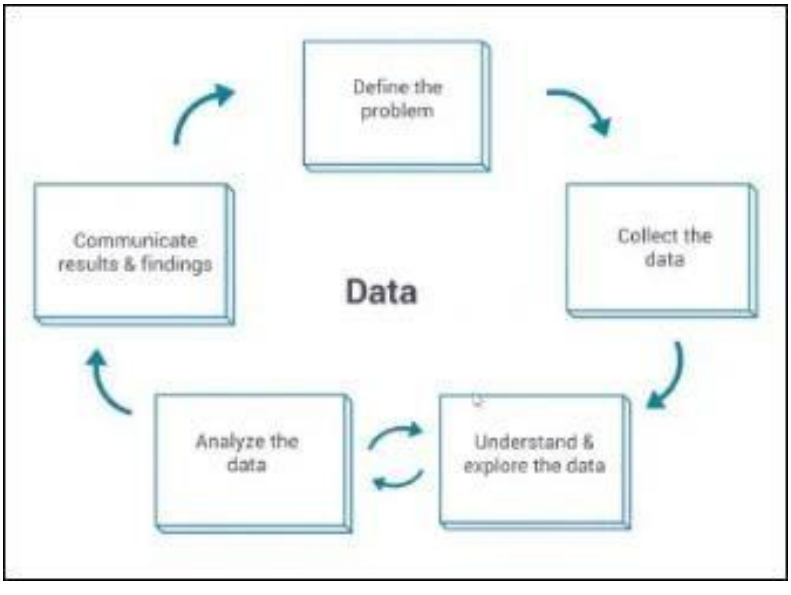

Fig. 2.2 Data Science Process

2. **Collecting data:** Once it is understood why they're coming to me—are they trying to figure out if this feature is good? Are they trying to figure out if this feature's good enough to launch (which is kind of a different question)? Are they trying to figure out if it causes pages to crash—maybe a narrower question? Once I understand what they're trying to figure out, I can help them collect data.

3. **Understanding the data**: To summarize, we understand the problem, we've decided that we're going to run an experiment, we launch an experiment, and we collect some data. Moving on, once we have some data, we need to understand it. This process, like I said before, looks a lot like an oscillation.

4. **Communicating the results**: Once you understand the data, you need to work on communicating the results so that other people can act on them. So this can be, you know, going into every meeting that people have about this feature and explaining why it's good or bad. Usually, it's more effective to produce some report and make a launch/no launch assertion. So you say: this is good because of these reasons, this is what we observed and this is how we measured.

2.2. Data Science Tools and Technologies

Data Science includes obtaining the value from data. It is all about understanding the data and processing it to extract the value out of it. Data Scientists are the data professionals who can organize and analyze the huge amount of data.

The functions that data scientists perform include identifying relevant questions, collecting data from different data sources, data organization, transforming data to the solution, and communicating these findings for better business decisions.

Python and R are the most popular languages among data scientists.

Tools for those who don't have programming knowledge	Tools for programmers
Rapid Miner	Python
Data Robot	R
Trifacta	SOL
IBM Watson Studio	Tableau
Amazon Lex	TensorFlow
	NoSQL
	Hadoop

Fig. 2.3 Types of Data Science Tools

RapidMiner:

RapidMiner is a tool for the complete life-cycle of prediction modeling. It has all the functionalities for data preparation, model building, validation, and deployment. It provides a GUI to connect the predefined blocks.

Features:

- RapidMiner Studio is for data preparation, visualization, and statistical modeling.
- RapidMiner Server provides central repositories.
- RapidMiner Radoop is for implementing big-data analytics functionalities.
- RapidMiner Cloud is a cloud-based repository.

Data Robot:

Data Robot is the platform for automated machine learning. It can be used by data scientists, executives, software engineers, and IT professionals.

Features:

- It provides an easy deployment process.
- It has a Python SDK and APIs.
- It allows parallel processing.
- Model Optimization.

Apache Hadoop:

Apache Hadoop is an open source framework. Simple programming models that are created using Apache Hadoop, can perform distributed processing of large data sets across computer clusters.

Features:

- It is a scalable platform.
- Failures can be detected and handled at the application layer.
- It has many modules like Hadoop Common, HDFS, Hadoop Map Reduce, Hadoop Ozone, and Hadoop YARN.

Trifacta:

Trifacta provides three products for data wrangling and data preparation. It can be used by individuals, teams, and organizations.

Features:

- Trifacta Wrangler will help you in exploring, transforming, cleaning, and joining the desktop files together.
- Trifacta Wrangler Pro is an advanced self-service platform for data preparation.
- Trifacta Wrangler Enterprise is for empowering the analyst team.

Alteryx:

Alteryx provides a platform to discover, prep, and analyze the data. It will also help you to find deeper insights by deploying and sharing the analytics at scale.

Features:

- It provides the features to discover the data and collaborate across the organization.
- It has functionalities to prepare and analyze the model.
- The platform will allow you to centrally manage users, workflows, and data assets.
- It will allow you to embed R, Python, and Alteryx models into your processes.

KNIME:

KNIME for data scientists will help them in blending tools and data types. It is an open source platform. It will allow you to use the tools of your choice and expand them with additional capabilities.

Features:

- It is very useful for the repetitive and time-consuming aspects.
- Experiments and expands to Apache Spark and Big data.
- It can work with many data sources and different types of platforms.

Excel:
Excel can be used as a tool for data science. It is easy to use tool for non-technical persons. It is good for analyzing data.

Features:

- It has good features for organizing and summarizing the data.
- It will allow you to sort and filter the data.
- It has conditional formatting features.

Matlab:
Matlab provides you the solution for analyzing data, developing algorithms, and for creating models. It can be used for data analytics and wireless communications.

Features:

- Matlab has interactive apps which will show you the working of different algorithms on your data.
- It has the ability to scale.
- Matlab algorithms can be directly converted to C/C++, HDL, and CUDA code.

Java:

Java is an object-oriented programming language. The compiled Java code can be run on any Java supported platform without recompiling it. Java is simple, object-oriented, architecture-neutral, platform-independent, portable, multi-threaded, and secure.

Features:
As features, we will see why Java is used for data science:

- Java provides a good number of tools and libraries that are useful for machine learning and data science.
- Java 8 with Lambdas: With this, You can develop large data science projects.
- Scala provides the support to data science.

Python:
Python is a high-level programming language and provides a large standard library. It has the features of object-oriented, functional, procedural, dynamic type, and automatic memory management.

Features:

- It is used by data scientists as it provides a good number of useful packages to download for free.
- Python is extensible.
- It provides free data analysis libraries.

2.2.1 Additional Data Science Tools

R
R is a programming language and can be used on a UNIX platform, Windows, and Mac OS.

SQL
This domain-specific language is used for managing the data from RDBMS through programming.

Tableau
Tableau can be used by individuals as well as teams and organizations. It can work with any database. It is easy to use because of its drag-and-drop functionality.

Cloud DataFlow
Cloud DataFlow is for stream and batch processing of data. It is a fully-managed service. It can transform and enrich the data in the stream and batch mode.

Kubernetes
Kubernetes provides an open source tool. It is used to automate the deployment, scale, and manage containerized applications.

2.3 R and Python

2.3.1 Introducing R

Ross Ihaka and Robert Gentleman created the open-source language R in 1995 as an implementation of the S programming language. The purpose was to develop a language that focused on delivering a better and more user-friendly way to do data analysis, statistics and graphical models. At first, R was primarily used in academics and research, but lately the enterprise world is discovering R as well. This makes R one of the fastest growing statistical languages in the corporate world.

One of the main strengths of R is its huge community that provides support through mailing lists, user-contributed documentation and a very active Stack Overflow group. There is also CRAN, a huge repository of curated R packages to which users can easily contribute. These packages are a collection of R functions and data that makes it easy to immediately get access to the latest techniques and functionalities without needing to develop everything from scratch. To end, if you’re an experienced programmer, you probably won't have a hard time to get up to speed with R. As a beginner, however, you might find yourself struggling with the steep learning curve. Luckily, there are many great learning resources you can consult nowadays.

2.3.2 Introducing Python

Python was created by Guido Van Rossem in 1991 and emphasizes productivity and code readability. Programmers that want to delve into data analysis or apply statistical techniques are some of the main users of Python for statistical purposes. The closer you get to working in an engineering environment, the more likely it is you might prefer Python. It's a flexible language that is great to do something novel, and given its focus on readability and simplicity, its learning curve is relatively low.

Similar to R, Python has packages as well. PyPi is the Python Package index and consists of libraries to which users can contribute. Just like R, Python has a great community but it is a bit more scattered, since it's a general purpose language. Nevertheless, Python for data science is rapidly claiming a more dominant position in the Python universe: the expectations are growing and more innovative data science applications will see their origin here.

2.4 R v/s Python

For individual data scientists, some common points to consider:

- Python is a great general programming language, with many libraries dedicated to data science.
- Many (if not most) general introductory programming courses start teaching with Python now.
- Python is the go-to language for many ETL and Machine Learning workflows.
- Many (if not most) introductory courses to statistics and data science teach R now.
- R has become the world's largest repository of statistical knowledge with reference implementations for thousands, if not tens of thousands, of algorithms that have been vetted by experts. The documentation for many R packages includes links to the primary literature on the subject.
- R has a very low barrier to entry for doing exploratory analysis, and converting that work into a great report, dashboard, or API.
- R with RStudio is often considered the best place to do exploratory data analysis.

For organizations with Data Science teams, some additional points to keep in mind:

- For some organizations, Python is easier to deploy, integrate and scale than R, because Python tooling already exists within the organization. On the other hand, we at RStudio have worked with thousands of data teams successfully solving these problems with our open-sourceand professional products, including in multi-language environments.

- R has a great community of supportive data scientists from diverse backgrounds. For example, R-Ladies is a global organization dedicated to promoting gender diversity in the R Community.
- Most interfaces for novel machine learning tools are first written and supported in Python, while many new methods in statistics are first written in R.
- Trying to enforce one language to the exclusion of the other, perhaps out of vague fears of complexity or costs to support both, risks excluding a huge potential pool of Data Scientist candidates either way.
- Advice on building Data Science teams often stresses the importance of having a diverse team bringing a variety of viewpoints and complementary skills to the table, to make it more likely to efficiently find the "best" solution for a given problem. In this vein, R users tend to come from a much more diverse range of domain expertise (ecology, economics, psychology, bioinformatics, policy analysis, etc.).

Thus, the focus on "R or Python?" risks missing the advantages that having both can bring to individual data scientists and data science teams.

DATA SCIENCE	Python	R
Description	King of data science programming languages	Golden child of Data Science
Purpose	It is a general-purpose language which is known for its simple syntax and compatibility with different operating systems.	It is an open source programming language and very beneficial for statistical computing. It operates smoothly on Linux, Windows, and Mac.
Features	1. Broadness 2 .Efficient 3. Extensible 4. Can be mastered easily	1. Open source 2. All in one analysis toolkit 3. Robust 4. Powerful package ecosystem
Libraries	1. NUMPY/SCIPY 2. MATPLOTLIB 3. PANDAS	1. CARET 2. STRINGR 3. GGPLOT2
Popular Applications	1. Dropbox is written in Python programming language which is now close to 170 million users. 2. Various plugins of python are created in Python.	1. R programming is widely used in the industries which use data driven decision support and statistical data analysis. 2. Zillow uses R programming to promote prices.

Fig. 2.4 R v/s Python

2.5 Data Science in Business

Taking an analytical approach based on numbers, facts, and statistics can supply a reasonable solution that might not have seemed apparent at first. Because of insights that data science can offer, more and more businesses are utilizing the power of data science to make evidence-based decisions, promote employee training, and understand their customers. Investing in a data science expert or data science technology can begin to add value to your business in these five meaningful ways:

1. Better decision making with quantifiable evidence
2. Improving the relevance of your product
3. Recruiting the best talent
4. Training staff
5. Finding your target audience

Business intelligence (BI) is an umbrella term that includes a variety of IT applications that are used to analyze an organization's data and communicate the information to relevant users.

BI includes a variety of software tools and techniques to provide the managers with the information and insights needed to run the business. Information can be provided about the current state of affairs with the capability to drill down into details, and also insights about emerging patterns which lead to projections into the future. BI tools include data warehousing, online analytical processing, social media analytics, reporting, dashboards, querying, and data mining.

A spreadsheet tool, such as Microsoft Excel, can act as an easy but effective BI tool by itself. Data can be downloaded and stored in the spreadsheet, then analyzed to produce insights, then presented in the form of graphs and tables. This system offers limited automation using macros and other features. The analytical features include basic statistical and financial functions. Pivot tables help do sophisticated what-if analysis. Add-on modules can be installed to enable moderately sophisticated statistical analysis.

As data grows and exceeds our capacity to make sense of it, the tools need to evolve, and so should the imagination of the BI specialist. "Data Scientist" has been called as the hottest job of this decade.

BI tools are required in almost all industries and functions. The nature of the information and the speed of action may be different across businesses, but every manager today needs access to BI tools to have up-to-date metrics about business performance. Businesses need to embed new insights into their operating processes to ensure that their activities continue to evolve with more efficient practices. The following are some areas of applications of BI and data mining.

Customer Relationship Management

A business exists to serve a customer. A happy customer becomes a repeat customer. A business should understand the needs and sentiments of the customer, sell more of its offerings to the existing customers, and also, expand the pool of customers it serves. BI applications can impact many aspects of marketing.

1. Maximize the return on marketing campaigns: Understanding the customer's pain points from data-based analysis can ensure that the marketing messages are fine-tuned to better resonate with customers.

2. Improve customer retention (churn analysis): It is more difficult and expensive to win new customers than it is to retain existing customers. Scoring each customer on their likelihood to quit, can help the business design effective interventions, such as discounts or free services, to retain profitable customers in a cost-effective manner.

3. Maximize customer value (cross-, up-selling): Every contact with the customer should be seen as an opportunity to gauge their current needs. Offering a customer new products and solutions based on those imputed needs can help increase revenue per customer. Even a customer complaint can be seen as an opportunity to wow the customer. Using the knowledge of the customer's history and value, the business can choose to sell a premium service to the customer.

4. Identify and delight highly-valued customers. By segmenting the customers, the best customers can be identified. They can be proactively contacted, and delighted, with greater attention and better service. Loyalty programs can be managed more effectively.

5. Manage brand image. A business can create a listening post to listen to social media chatter about itself. It can then do sentiment analysis of the text to understand the nature of comments, and respond appropriately to the prospects and customers.

Healthcare and Wellness

Health care is one of the biggest sectors in advanced economies. Evidence based medicine is the newest trend in data-based health care management. BI applications can help apply the most effective diagnoses and prescriptions for various ailments. They can also help manage public health issues, and reduce waste and fraud.

1. Diagnose disease in patients: Diagnosing the cause of a medical condition is the critical first step in a medical engagement. Accurately diagnosing cases of cancer or diabetes can be a matter of life and death for the patient. In addition to the patient's own current situation, many other factors can be considered, including the patient's health history, medication history, family's history, and other environmental factors. This makes diagnosis as much of an art form as it is science. Systems, such as IBM Watson, absorb all the medical research to date and make probabilistic diagnoses in the form of a decision tree, along with a full explanation for their recommendations. These systems take away most of the guess work done by doctors in diagnosing ailments.
2. Treatment effectiveness: The prescription of medication and treatment is also a difficult choice out of so many possibilities. For example, there are more than 100 medications for hypertension (high blood pressure) alone. There are also interactions in terms of which drugs work well with others and which drugs do not. Decision trees can help doctors learn about and prescribe more effective treatments. Thus, the patients could recover their health faster with a lower risk of complications and cost.
3. Wellness management: This includes keeping track of patient health records, analyzing customer health trends and proactively advising them to take any needed precautions.

4. Manage fraud and abuse: Some medical practitioners have unfortunately been found to conduct unnecessary tests, and/or overbill the government and health insurance companies. Exception reporting systems can identify such providers and action can be taken against them.
5. Public health management: The management of public health is one of the important responsibilities of any government. By using effective forecasting tools and techniques, governments can better predict the onset of disease in certain areas in real time. They can thus be better prepared to fight the diseases. Google has been known to predict the movement of certain diseases by tracking the search terms (like flu, vaccine) used in different parts of the world.

Education

As higher education becomes more expensive and competitive, it becomes a great user of data-based decision-making. There is a strong need for efficiency, increasing revenue, and improving the quality of student experience at all levels of education.

1. Student Enrollment (Recruitment and Retention): Marketing to new potential students requires schools to develop profiles of the students that are most likely to attend. Schools can develop models of what kinds of students are attracted to the school, and then reach out to those students. The students at risk of not returning can be flagged, and corrective measures can be taken in time.

2. Course offerings: Schools can use the class enrolment data to develop models of which new courses are likely to be more popular with students. This can help increase class size, reduce costs, and improve student satisfaction.

3. Fund-raising from Alumni and other donors: Schools can develop predictive models of which alumni are most likely to pledge financial support to the school. Schools can create a profile for alumni more likely to pledge donations to the school. This could lead to a reduction in the cost of mailings and other forms of outreach to alumni.

Banking

Banks make loans and offer credit cards to millions of customers. They are most interested in improving the quality of loans and reducing bad debts. They also want to retain more good customers, and sell more services to them.

1. Automate the loan application process: Decision models can be generated from past data that predict the likelihood of a loan proving successful. These can be inserted in business processes to automate the financial loan approval process.

2. Detect fraudulent transactions: Billions of financial transactions happen around the world every day. Exception-seeking models can identify patterns of fraudulent transactions. For example, if money is being transferred to an unrelated account for the first time, it could be a fraudulent transaction.

3. Maximize customer value (cross-, up-selling). Selling more products and services to existing customers is often the easiest way to increase revenue. A checking account customer in good standing could be offered home, auto, or educational loans on more favorable terms than other customers, and thus, the value generated from that customer could be increased.

4. Optimize cash reserves with forecasting. Banks have to maintain certain liquidity to meet the needs of depositors who may like to withdraw money. Using past data and trend analysis, banks can forecast how much to keep and invest the rest to earn interest.

Financial Services

Stock brokerages are an intensive user of BI systems. Fortunes can be made or lost based on access to accurate and timely information.

1. Predict changes in bond and stock prices: Forecasting the price of stocks and bonds is a favorite pastime of financial experts as well as lay people. Stock transaction data from the past, along with other variables, can be used to predict future price patterns. This can help traders develop longterm trading strategies.

2. Assess the effect of events on market movements. Decision models using decision trees can be created to assess the impact of events on changes in market volume and prices. Monetary policy changes (such as Federal Reserve interest rate change) or geopolitical changes (such as war in a part of the world) can be factored into the predictive model to help take action with greater confidence and less risk.

3. Identify and prevent fraudulent activities in trading: There have unfortunately been many cases of insider trading, leading to many prominent financial industry stalwarts going to jail. Fraud detection models seek out-of-the-ordinary activities, and help identify and flag fraudulent activity patterns.

Insurance

This industry is a prolific user of prediction models in pricing insurance proposals and managing losses from claims against insured assets.

1. Forecast claim costs for better business planning: When natural disasters, such as hurricanes and earthquakes strike, loss of life and property occurs. By using the best available data to model the likelihood (or risk) of such events happening, the insurer can plan for losses and manage resources and profits effectively.

2. Determine optimal rate plans: Pricing an insurance rate plan requires covering the potential losses and making a profit. Insurers use actuary tables to project life spans and disease tables to project mortality rates, and thus price themselves competitively yet profitably.

3. Optimize marketing to specific customers: By micro-segmenting potential customers, a data-savvy insurer can cherry pick the best customers and leave the less profitable customers to its

competitors. Progressive Insurance is a US-based company that is known to actively use data mining to cherry pick customers and increase its profitability. 4. Identify and prevent fraudulent claim activities. Patterns can be identified as to where and what kinds of fraud are more likely to occur. Decision-tree-based models can be used to identify and flag fraudulent claims.

Manufacturing

Manufacturing operations are complex systems with inter-related subsystems. From machines working right, to workers having the right skills, to the right components arriving with the right quality at the right time, to money to source the components, many things have to go right. Toyota's famous lean manufacturing company works on just-in-time inventory systems to optimize investments in inventory and to improve flexibility in their product-mix.

1. Discover novel patterns to improve product quality: Quality of a product can also be tracked, and this data can be used to create a predictive model of product quality deteriorating. Many companies, such as automobile companies, have to recall their products if they have found defects that have a public safety implication. Data mining can help with root cause analysis that can be used to identify sources of errors and help improve product quality in the future.

2. Predict/prevent machinery failures: Statistically, all equipment is likely to break down at some point in time. Predicting which machine is likely to shut down is a complex process. Decision models to forecast machinery failures could be constructed using past data. Preventive maintenance can be planned, and manufacturing capacity can be adjusted, to account for such maintenance activities.

2.6 Tips for recruiting data science people

The ideal candidate has skills in each of three broad buckets: math/statistics, databases/programming, and business. My hiring process is designed around probing the candidate in each area to see where they fall.

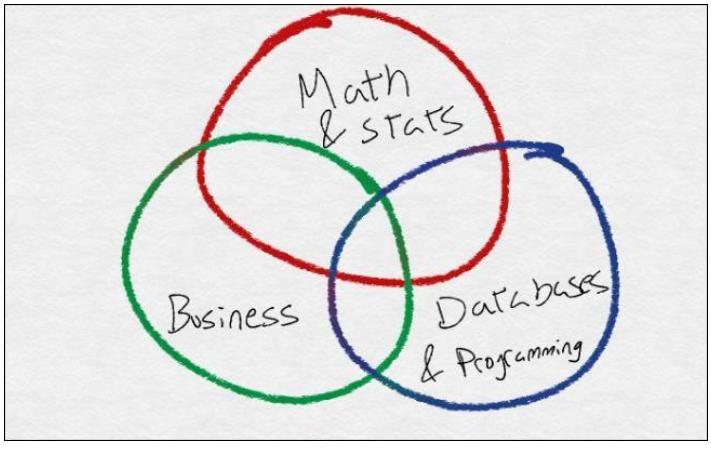

Fig. 2.5 Skills for an Ideal Data Scientist

Math and statistics

This includes basic statistics (example: linear regression — what it is and when it works well), and model building (ex: training versus test data, different models of training like cross validation, what does "boosting" even mean). If the candidate is sufficiently experienced in this area, they should be able to list the different models they've used and the different type of problems they've worked on.

Though someone with a degree in statistics or data science should pass this portion of assessment with flying colors, someone with a related degree (math, computer science, or economics) may not have the machine learning component of this skill. If they don't have machine learning knowledge, then having a related degree suggests they can pick the skill on the job. One of my best hires had his Master's degree in *fishing science*.

Databases and programming

In its most basic form, data science is the art of taking existing data and processing it in a meaningful fashion. That means you need to be able to (1) pull the data from a source and (2) process it into an insight. The ideal candidate should have the technical knowledge to do both of these steps. Counter to intuition, the tools to get data and process it are not the same.

To pull the data, the candidate must understand relational databases. Since data is stored in relational databases, and these are queried using SQL, the candidate must know SQL. Sometimes a candidate doesn't know SQL but conceptually knows how to join tables and aggregate them. If this is the case they can pick up SQL on the job. If they have experience in storing data in other ways like NoSQL then that's a plus, but not something I expect. Only knowing how to read data from flat files isn't enough.

Once you have data, it needs to be used. That *may* mean the complex act of building a machine learning model, but it *certainly* always means creating a visualization of the material. The candidate needs to be able to do this, which requires writing code. If a candidate knows R, Python, or MATLAB they are good to go on day one. If they know a language more common in software development like Java then they can easily pick up a statistics-focused language. There are GUI-based tools to do this work, but if the candidate has only used GUIs then their skill set is too limited to do the wide array of work I expect from a data scientist.

If they only have experience in Excel then they do not satisfy this requirement. While Excel can create visualizations, that's pretty much all it can do and it doesn't even do that quickly. Having only used Excel shows the candidate hasn't thought about what can be done outside of Excel — or worse, they decided they'd rather stick to only using what they already know.

Someone with a computer science degree satisfies this by definition. A data analyst often only works in Excel and hasn't connected directly to their data. Someone in business intelligence may be able to query data, but they lack the programming tools to meaningfully manipulate data. Therefore most data analysts and people from business intelligence do not satisfy this criterion.

Business

Let's face it: the ability to understand a business environment and work within it is just as important of an ability as statistics and programming. The whole idea of data science is using technical skills to create real world, practical insights, so you need to be able to understand how the real world works and what insights people need. The candidate needs to be able to:

- understand a problem a person or department is having within a company,
- translate that into a problem that data science can solve,
- solve it (using their math/stats or databases/programming skills), and
- convert that solution into an insight that someone who doesn't know anything about data science can use.

75% of those steps are centered around business fundamentals. For example, if a company is having trouble with their promotional emails, the candidate may consider segmenting the email recipients using a clustering algorithm and giving each segment a personalized email. If their approach involved using a *k*-means algorithm with 7 clusters, the candidate should be able to explain why they chose 7 to someone who doesn't know what a clustering algorithm is. Depending on their level, they should also understand how often to check in with a client or project stakeholder, be able to ask questions to a data-owner if they don't understand what's in the data, and make a well-formatted PowerPoint. Someone with several years of experience working at a company usually satisfies this requirement.

2.7 Final Deliverable

Before one jumps into analytics and working with data, one has to ask the first and most important questions. When all the entire data analysis is done, we have some answers and at this point, we need to ask what format we are presenting the answers to the stakeholders. You have to ask how your solution will be presented which is what is called the Final Deliverable. It could be a report, a powerpoint presentation, or a snippet of code. It could be in any form but the very basic question that is needed before analyzing is what format will the final deliverables be? Imagine you didn't do this and you embark on the analytics and later discovered your solution is not properly formatted, this means going back and starting all over again, this is why it's really important to know the format before starting to analyze your code.

2.7.1 Format of Deliverables

For every budding data scientist, you have to make the story telling a compelling one to those you are presenting to. The quality of the deliverables depends on how finely refined your questions are else the solution or result you find will suffer. So yes, it can be in any format but one that's spectacular, fascinating and solution oriented. Personally, I will also suggest your kind of audience should play a part on how your presentation looks.

2.7.2 Second step of getting the Final Deliverable

Before we begin to analyze data, we need to look into resources available, resources in terms of data sets, survey, questionnaires and available reports. Based on a quick scan of the resources, we would then be able to give a good summary of what we can come up it which is essential before analyzing. The second step is to know what answers are possible given what resources are available.

2.7.3 Third step in the process the actual Analytical work

As a data scientist, you have to be efficient, proficient and should not be the one busy re-inventing the wheel. Given that we have the internet and these online resources available, it is very

IMPORTANT that we do not re-invent the wheel and write same code somebody else has written and made available. I'll personally like to call this ESME(Eliminating Stupid Mental Effort). So the next step as a Data Scientist after you have gotten the questions and answers is to decide the appropriate tool or algorithm or method to check if it's readily available for use, if not, try to check if others have tried to attempt what you want to do in order to avoid repeating others.
The Google Scholar is actually great for making research on papers, journals, publications and also relevant reports from industry leaders and at the same time with blog aggregators which is also popular among data scientist where you can search for algorithms, code or tools.

2.7.4 Last step in the process of getting the final deliverable?

At this step, we should know the tools to use and so the first step to your analytics should be very basic tabulations and see how distributions are, how frequent data set exists and trends are by department, age, education, etc. At the same time, while you are developing these basic tabulations, cross tabulations, also try to generate graphics, scatter plots or bar charts. Why? Because you will see that different data sets with different distributions may give you the same average value or the same standard deviation. So the first process is to start with simple process and tool to develop a level of familiarity with the dataset. The more you are familiar, the more you will be quick to apply advanced algorithms and tools.

2.7.5 Presenting the final deliverable

After analyzing, tabulating and now you've gotten your model. The next thing is to write a report and after writing, you want to make a final check; if you've done a good or compelling story, etc. Below are the seven major points to consider or ask yourself?

- Have you told your readers at the very outset what they might gain by reading your report? i. what benefits they stand to gain from your report? what is the question? what is the answer? and how they can gain from it?
- The second question is have you made the aim of your work clear? the purpose?
- The significance of the contribution. i.e the significance of your findings and it's effectiveness.
- Have you set your work in the appropriate context?
- Have you addressed the question of practicality and usefulness? Your solution/findings have to be useful, you have to explain how your solution is solving problems.
- Have you identified future developments?
- Have you structured your report in a clear and logical fashion?

2.7.6 End to end scenario demonstrating the process and getting the final deliverable and its presentation

The very first step is to think of the format we are presenting the data and the second is to take look at the question and refine it to see if this study is being measured for girls only drop-outs, boys only drop outs, primary school drop-outs, high school drop-outs or maybe Universtiy drop outs.

The second phase; what answers can be generated based on the available resources. Remember you're in a developing country where authentic data may or may not be available. After the second stage of being able to predict the kinds of answer you can derive from the available resources, you move on to making research for already existing solutions and see how others have found answers to this. You need to look at other answers that might have been provided by other people within Nigeria or other developing countries to see how you can apply that method to yours.

The fourth step is analytics, this is where you need to start with basic regulations and graphics.

The final stage is to put your model together and presenting it in the respective format. Afterward, you can ask the questions; what is my story here? what is the most compelling story? what are the most significant answers I found? what is the significance of my findings to the rate of school drop-outs in this country? how can I help my firm with this information? how can I help my country with this result? how does it come to solve and reduce the rate of school drop outs in my country? All these questions are powerful enough to make your analytical work efficient, effective, solution-oriented and accepted.

2.8 Report Structure:

Structure of a Data Analysis Report

A data analysis report is somewhat different from other types of professional writing that you may have done or seen, or will learn about in the future. It is related to but not the same as:

The overall structure of a data analysis report is simple:

1. Introduction
2. Body
3. Conclusion(s)/Discussion
4. Appendix/Appendices

The data analysis report is written for several different audiences at the same time:

- Primary Audience: A primary collaborator or client
- Secondary Audience: An Executive Person
- Secondary Audience: A Technical Supervisor

The data analysis report has two very important features:

- It is organized in a way that makes it easy for different audiences to skim/fish through it to find the topics and the level of detail that are of interest to them.

- The writing is as invisible / unremarkable as possible, so that the content of the analysis is what the reader remembers, not distracting quirks or tics in the writing. Examples of distractions include:
  - Extra sentences, overly formal or flowery prose, or at the other extreme overly casual or overly brief prose.
  - Grammatical and spelling errors.
  - Placing the data analysis in too broad or too narrow a context for the questions of interest to your primary audience.
  - Focusing on process rather than reporting procedures and outcomes.
  - Getting bogged down in technical details, rather than presenting what is necessary to properly understand your conclusions on substantive questions of interest to the primary audience.

The data analysis report isn't quite like a research paper or term paper in a class, nor like a research article in a journal. It is meant, primarily, to start an organized conversation between you and your client/collaborator. In that sense it is a kind of "internal" communication, sort of like an extended memo. On the other hand it also has an "external" life, informing a boss or supervisor what you've been doing.

Now let's consider the basic outline of the data analysis report in more detail:

1. **Introduction.** Good features for the Introduction include:
   - Summary of the study and data, as well as any relevant substantive context, background, or framing issues.
   - The "big questions" answered by your data analyses, and summaries of your conclusions about these questions.
   - Brief outline of remainder of paper.

   The above is a pretty good order to present this material in as well.

2. **Body.** The body can be organized in several ways. Here are two that often work well:
   - Traditional. Divide the body up into several sections at the same level as the Introduction, with names like:
     - Data
     - Methods
     - Analysis
     - Results

This format is very familiar to those who have written psych research papers. It often works well for a data analysis paper as well, though one problem with it is that the Methods section often sounds like a bit of a stretch: In a psych research paper the Methods section describes what you did to get your data. In a data analysis paper, you should describe the analyses that you performed. Without the results as well, this can be pretty sterile sounding, so I often merge these "methods" pieces into the "Analysis" section when I write.

3. **Question-oriented.** In this format there is a single Body section, usually called

"Analysis", and then there is a subsection for each question raised in the introduction, usually taken in the same order as in the introduction (general to specific, decreasing order of importance, etc.). Within each subsection, statistical method, analyses, and conclusion would be described (for each question). Forexample:

Analysis

2.1 Success Rate
Methods
Analysis
Conclusions

2.2 Time to Relapse

Methods

Analysis

Conclusions

2.2 E*ff***ect of Gender**

Methods

Analysis

Conclusions

2.3 Hospital E*ff*ects
Methods
Analysis
Conclusions

Other organizational formats are possible too. Whatever the format, it is useful to provide one or two well-chosen tables or graphs per question in the body of the report, for two reasons: First, graphical and tabular displays can convey your points more efficiently than words; and second, your "skimming" audiences will be more likely to have their eye caught by an interesting graph or table than by running text. However, too much graphical/tabular material will break up the flow of the text and become distracting; so extras should be moved to the Appendix.

3. **Conclusion(s)/Discussion.** The conclusion should reprise the questions and conclusions of the introduction, perhaps augmented by some additional observations or detailsgleaned from the analysis section. New questions, future work, etc., can also be raised here.

4. **Appendix/Appendices.** One or more appendices are the place to out details and ancillary materials. These might include such items as

    - Technical descriptions of (unusual) statistical procedures

- Detailed tables or computer output
- Figures that were not central to the arguments presented in the body of the report
- Computer code used to obtain results.

In all cases, and especially in the case of computer code, it is a good idea to add some text sentences as comments or annotations, to make it easier for the uninitiated reader to follow what you are doing.

It is often difficult to find the right balance between what to put in the appendix and what to put in the body of the paper. Generally you should put just enough in the body to make the point, and refer the reader to specific sections or page numbers in the appendix for additional graphs, tables and other details.

2.9 Big Data Analytics - Core Deliverables

As mentioned in the big data life cycle, the data products that result from developing a big data product are in most of the cases some of the following −

- **Machine learning implementation** − This could be a classification algorithm, a regression model or a segmentation model.
- **Recommender system** − The objective is to develop a system that recommends choices based on user behavior. **Netflix** is the characteristic example of this data product, where based on the ratings of users, other movies are recommended.
- **Dashboard** − Business normally needs tools to visualize aggregated data. A dashboard is a graphical mechanism to make this data accessible.
- **Ad-Hoc analysis** − Normally business areas have questions, hypotheses or myths that can be answered doing ad-hoc analysis with data.

2.10 Data Science Careers

Data science experts are needed in virtually every job sector—not just in technology. In fact, the five biggest tech companies—Google, Amazon, Apple, Microsoft, and Facebook—only employ one half of one percent of U.S. employees. However—in order to break into these high-paying, in-demand roles—an advanced education is generally required.

"Data scientists are highly educated–88 percent have at least a master's degree and 46 percent have PhDs–and while there are notable exceptions, a very strong educational background is usually required to develop the depth of knowledge necessary to be a data scientist," reports KDnuggets, a leading site on Big Data.

Here are some of the leading data science careers you can break into with an advanced degree.

2.10.1 Business Intelligence (BI) Developer

Typical Job Requirements: BI developers design and develop strategies to assist business users in quickly finding the information they need to make better business decisions. Extremely data-savvy, they use BI tools or develop custom BI analytic applications to facilitate the end-users' understanding of their systems

Notable Companies: DollarShave Club, Discover, and Liberty Mutual

2.10.2 Data Architect

Typical Job Requirements: Ensure data solutions are built for performance and design analytics applications for multiple platforms.

Notable Companies: IBM, eBay, AAA Club Alliance, T-Mobile

2.10.3 Applications Architect

Typical Job Requirements: Track the behavior of applications used within a business and how they interact with each other and with users.

Notable Companies: UPS, Humana, Dow Jones, Oracle

2.10.4 Infrastructure Architect

Typical Job Requirements: Oversee that all business systems are working at optimally and can support the development of new technologies and system requirements. A similar job title is Cloud Infrastructure Architect, which oversees a company's cloud computing strategy.

Notable Companies: Abbott Labs, Hewlett-Packard, Dell, Ford Motor Company

2.10.5 Enterprise Architect

Typical Job Requirements: According to Techopedia, an enterprise architect, "Works closely with stakeholders, including management and subject matter experts (SME), to develop a view of an organization's strategy, information, processes and IT assets."

Notable Companies: Cisco, Boeing, Lockheed Martin, Microsoft

2.10.6 Data Scientist

Typical Job Requirements: Find, clean, and organize data for companies. Data scientists will need to be able to analyze large amounts of complex raw and processed information to find patterns that will benefit an organization and help drive strategic business decisions. Compared to data analysts, data scientists are much more technical.

Notable Companies: Facebook, Capital One, Airbnb, Twitter

2.10.7 Data Analyst

Typical Job Requirements: Transform and manipulate large data sets to suit the desired analysis for companies. For many companies, this role can also include tracking web analytics and analyzing A/B testing.

Notable Companies: Walmart, Gap, Bank of America, Kohler

2.10.8 Data Engineer

Typical Job Requirements: Perform batch processing or real-time processing on gathered and stored data. Make data readable for data scientists.

Notable Companies: Spotify, Verizon, General Motors, Shutterfly

2.10.9 Machine Learning Scientist

Typical Job Requirements: Research new data approaches and algorithms.

Notable Companies: Apple, The Johns Hopkins Hospital, Expedia, Tinder

2.10.10 Machine Learning Engineer

Typical Job Requirements: Create data funnels and deliver software solutions.

Notable Companies: Nike, Dropbox, LinkedIn, Uber

2.10.11 Statistician

Typical Job Requirements: Interpret, analyze, and report statistical information, such as formulas and data for business purposes.

Notable Companies: U.S. Census Bureau, Google, PayPal, U.S. Department of Agriculture

2.11 Data Scientists Are in Constant Demand

Dr. Schedlbauer concludes that while some data science work will likely be automated within the next 10 years, "there is a clear need for professionals who understand a business need, can devise a data-oriented solution, and then implement that solution."

Data science experts are needed in almost every field, from government security to dating apps. Millions of businesses and government departments rely on big data to succeed and better serve their customers. Data science careers are in high demand and this trend will not be slowing down any time soon, if ever.

2.12 Data Science Case Studies

Here are the most famous Data Science Case Studies that will brief you how Data Science is used in different sectors.

1. Data Science in Pharmaceutical Industries

With the enhancement in data analytics and cloud-driven technologies, it is now easier to analyze vast datasets of patient information. In Pharmaceutical Industries, **Artificial Intelligence and Data Science** have revolutionized oncology. With new pharmaceutical products emerging every day, it is difficult for the physicians to keep themselves updated on the treatment products. Moreover, more generic diagnostic treatment options find it difficult to tap into a complex

competitive market. However, with the advancements in analytics and through the processing of parallel pipelined statistical models, it is now easier for pharmaceutical industries to have a competitive edge over the market.

With various statistical models like Markov Chains, it is now possible to predict the likelihood of doctors prescribing medicines based on their interaction with the brand. Similarly, reinforcement learning is starting to establish itself in the realm of digital marketing. It is used to recognize the patterns of digital engagement of physicians and their prescriptions. The main motive of this data science case study is to share the issues faced and how data science provides solutions for that.

2. Predictive Modeling for Maintaining Oil and Gas Supply

Crude oil and gas industries face a major problem of equipment failures which usually occurs due to the inefficiency of oil wells and their performance at a subpar level. With the adoption of a successful strategy that advocates for predictive maintenance, the well operators can be alerted of crucial stages for shutdown as well as can be notified of maintenance periods. This will lead to a boost in oil production and prevent further loss.

Data Scientists can apply Predictive Maintenance Strategy to use data in order to optimize high-value machinery for manufacturing and refining oil products. With the telemetry data extracted through sensors, a steady stream of historical data can be used to train our machine learning model. This machine learning model will predict the failing of machine parts and will notify the operators of timely maintenance in order to avert oil losses. A Data Scientist assigned with the development of PdM strategy will help to avoid hazards and will predict machine failures, prompting the operators to take precautionary steps.

3. Data Science in BioTech

The human gene is composed of four building blocks – A, T, C and G. Our looks and characteristics are determined by the three billion permutations of these four building blocks. While there are genetic defects and defects acquired during lifestyle, the consequences of it can lead to chronic diseases. Identifying such defects at an early stage can help the doctors and diagnostic teams to take preventive measures.

Helix is one of the genome analysis companies that provide customers with their genomic details. Also, several medicines tailored for specific genetic designs have become increasingly popular due to the advent of new computational methodologies. Due to the explosion in data, we can understand complex genomic sequences and analyze them on a large scale. Data Scientists can use contemporary computing power to handle large datasets and understand patterns of genomic sequences to identify defects and provide insights to physicians and researchers. Furthermore, with the usage of wearable devices, data scientists can use the relationship between the genetic characteristics and the medical visits to develop a predictive modeling system.

4. Data Science in Education

Data Science has also changed the way in which students interact with teachers and evaluate theirperformance. Instructors can use data science to analyze the feedback received from the students and use it to improve their teaching. Data Science can be used to create predictive modeling that can predict the drop-out rate of students based on their performance and inform the instructors to take necessary precautions.

IBM analytics has created a project for schools to evaluate student's performance based on their performance. Universities are using data to avoid retention supplement the performance of their students. For example, the University of Florida makes use of IBM Cognos Analytics to keep trackof student performance and make necessary predictions. Also, MOOCs and online education platforms are using data science to keep track of the students, to automate the assignment evaluation and to better the course based on student feedback.

Summary

So, these were the most viewed Data Science Case studies that are provided by Data Science experts. Data Science has created a strong foothold in several industries. There are many more case studies that prove that data science has boosted the performance of industries and has madethem smarter and more efficient.

Data Science has not only accelerated the performance of companies but has also made itpossible for them to manage & sustain their performance with ease.

UNIT -3

Big Data

The term Big Data refers to all the data that is being generated across the globe at an unprecedented rate. This data could be either structured or unstructured. Today's business enterprises owe a huge part of their success to an economy that is firmly knowledge-oriented.

Data drives the modern organizations of the world and hence making sense of this data and unraveling the various patterns and revealing unseen connections within the vast sea of data becomes critical and a hugely rewarding endeavor indeed. There is a need to convert Big Data into Business Intelligence that enterprises can readily deploy. Better data leads to better decision making and an improved way to strategize for organizations regardless of their **size**, **geography, market share, customer segmentation** and such other categorizations. Hadoop is the platform of choice for working with extremely large volumes of data.

Big data is an umbrella term for a collection of data sets so large and complex that it becomes difficult to process them using traditional data management tools. There has been increasing democratization of the process of content creation and sharing over the Internet, using social media applications. The combination of cloud-based storage, social media applications, and mobile access devices is helping crystallize the big data phenomenon. The leading management consulting firm, McKinsey & Co. created a flutter when it published a report in 2011 showing a huge impact of such big data on business and other organizations. They also reported that there will be millions of new jobs in the next decade, related to the use of big data in many industries.

Big data can be used to discover new insights from a 360-degree view of a situation that can allow for a complete new perspective on situations, new models of reality, and potentially new types of solutions. It can help spot business trends and opportunities. For example, Google is able to predict the spread of a disease by tracking the use of search terms related to the symptoms of the disease

over the globe in real time. Big Data can help determine the quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions. Big Data is enabling evidence-based medicine, and many other innovations.

Data has become the new natural resource. Organizations have a choice in how to engage with this exponentially growing volume, variety and velocity of data. They can choose to be buried under the avalanche, or they can choose to use it for competitive advantage. Challenges in big data include the entire range of operations from capture, curation, storage, search, sharing, analysis, and visualization. Big data is more valuable when analyzed as a whole. More and more information is derivable from analysis of a single large set of related data, as compared to separate smaller sets. However, special tools and skills are needed to manage such extremely large data sets.

Big Data has certain characteristics and hence is defined using 4Vs namely:

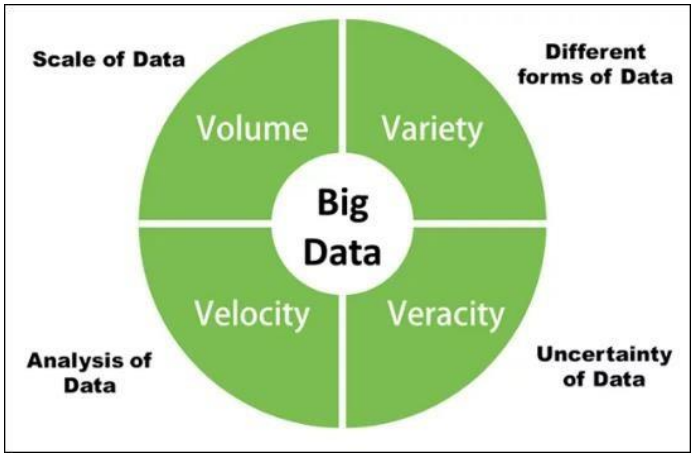

Fig. 3.1 Characteristics of Big Data

Volume: the amount of data that businesses can collect is really enormous and hence the volume of the data becomes a critical factor in Big Data analytics.

Velocity: the rate at which new data is being generated all thanks to our dependence on the internet, sensors, machine-to-machine data is also important to parse Big Data in a timely manner.

Variety:the data that is generated is completely heterogeneous in the sense that it could be in various formats like video, text, database, numeric, sensor data and so on and hence understanding the type of Big Data is a key factor to unlocking its value.

Veracity: knowing whether the data that is available is coming from a credible source is of utmost importance before deciphering and implementing Big Data for business needs.

Sources of Data: There are several sources of data, including some new ones. Data from outside the organization may be incomplete, and of a different quality and accuracy.

1. **Social Media:** All activities on the web and social media are considered stores and are accessible. Email was the first major source of new data. Google searches, Facebook posts, Tweets, Youtube videos, and blogs enable people to generate data for one another.

2. **Organizations:** Business organizations and government are a major source of data. ERP systems, e-Commerce systems, user-generated content, web-access logs, and many other sources of data generate valuable data for organizations.

3. **Machines:** The Internet of things is evolving. Many machines are connected to the web and autonomously generate data that is untouched by humans. RFID tags and telematics are two major applications that generate enormous amounts of data. Connected devices such as phones and refrigerators generate data about their location and status.

4. **Metadata:** There is enormous data about data itself. Web crawlers and web-bots scan the web to capture new webpages, their html structure, and their metadata. This data is used by many applications, including web search engines.

The data also includes varied quality of data. It depends upon the purpose of collecting the data, and how carefully it has been collected and curated. Data from within the organization is likely to be of a higher quality. Publicly available data would include some trustworthy data such as from the government.

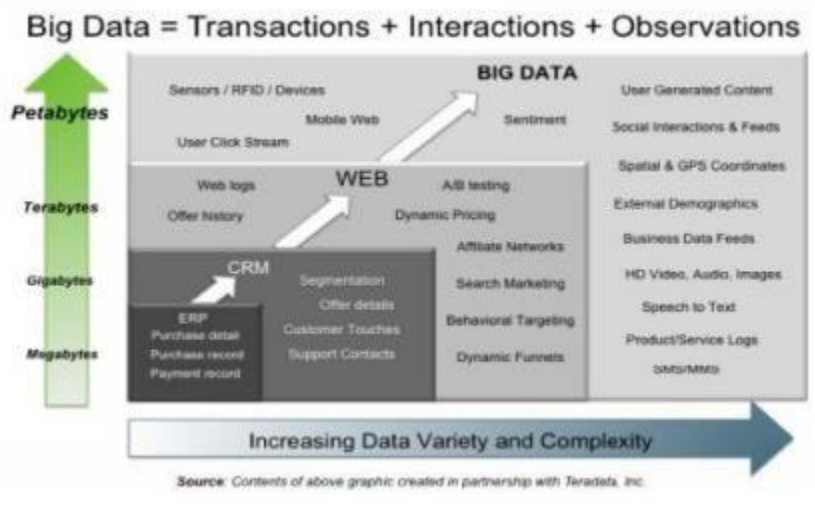

Fig. 3.2 Sources of Big Data

Big Data Landscape

Big data can be understood at many levels (Figure below). At the highest level are business applications to suit particular industries or to suit business intelligence for executives. A unique concept of "data as a service" is also possible for particular industries. At the next level, there are infrastructure elements for broad cross-industry applications, such as analytics and structured databases. This also includes offering this infrastructure as a service with some operational management services built in. At the core, big data is about technologies and standards to store and manipulate the large fast streams of data, and make them available for rapid data-based decision making.

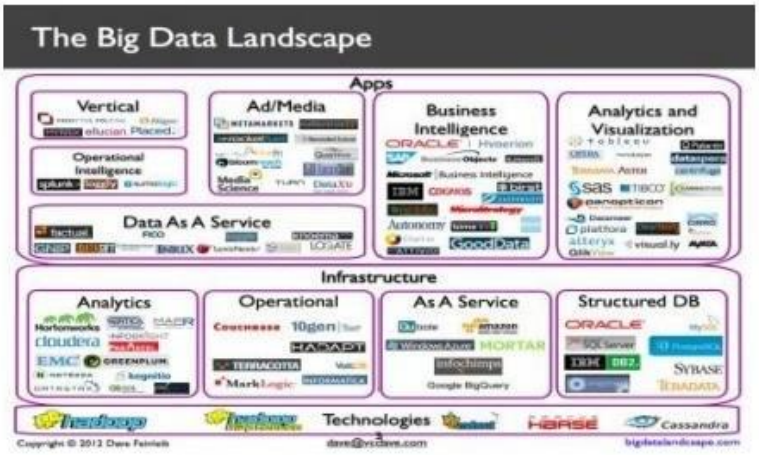

Fig. 3.3 The Big Data Landscape

3.1 Types of Big Data

Big Data could be found in three forms:

1. Structured
2. Unstructured
3. Semi-structured

3.2.1 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 has 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, nowadays, we are foreseeing issues when a size of such data grows to a huge extent, typical sizes are being in the rage of multiple zettabytes.

Examples of Structured Data

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

Employee_ID	Employee_Name	Gender	Department	Salary_In_lacs
2365	Rajesh 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

Fig. 3.4 Employee Table

3.2.2 Unstructured

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. A typical example of unstructured data is a heterogeneous data source containing a combination of simple text files, images, videos etc. Now 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.

Examples of Un-structured Data

The output returned by 'Google Search'

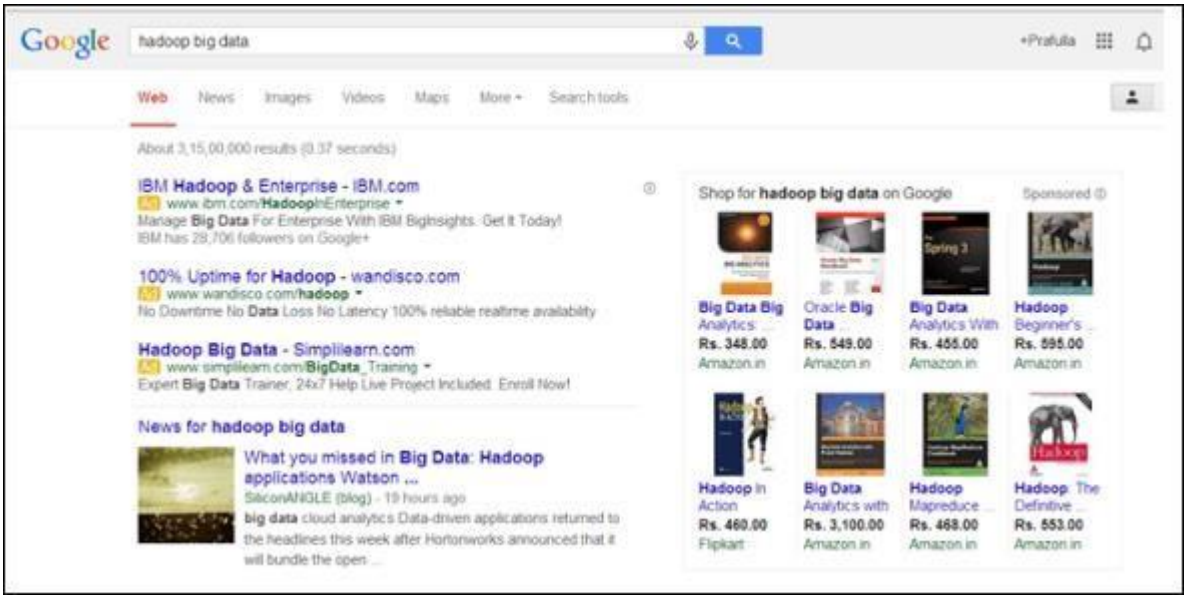

Fig. 3.5 Employee Table

3.2.3 Semi-structured

Semi-structured data can contain both the forms of data. We can see semi-structured data as a structured 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 an XML file.

Examples Of Semi-structured Data:

```
<rec><name>Prashant Rao</name><sex>Male</sex><age>35</age></rec>
<rec><name>Seema R.</name><sex>Female</sex><age>41</age></rec>
<rec><name>Satish Mane</name><sex>Male</sex><age>29</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>
```

Fig. 3.6 Personal data stored in an XML file-

3.2 Benefits of Big Data Processing

Ability to process Big Data brings in multiple benefits, such as-

- Businesses can utilize outside intelligence while taking decisions

Access to social data from search engines and sites like facebook, twitter are enabling organizations to fine tune their business strategies.

- Improved customer service

Traditional customer feedback systems are getting replaced by new systems designed with Big Data technologies. In these new systems, Big Data and natural language processing technologies are being used to read and evaluate consumer responses.

- Early identification of risk to the product/services, if any
- Better operational efficiency

Big Data technologies can be used for creating a staging area or landing zone for new data before identifying what data should be moved to the data warehouse. In addition, such integration of Big Data technologies and data warehouse helps an organization to offload infrequently accessed data.

3.3 Big Data Technology

Big Data Technology can be defined as a Software-Utility that is designed to **Analyse**, **Process** and **Extract** the information from an extremely complex and large data sets which the **Traditional Data Processing Software** could never deal with.

We need Big Data Processing Technologies to Analyse this huge amount of Real-time data and come up with Conclusions and Predictions to reduce the risks in the future.

Big Data Technologies New tools and techniques have arisen in the last 10-20 years to handle this large and still growing data. There are technologies for storing and accessing this data.

1. Non-relational data structures: Big data is stored using non-traditional data structures. Large non-relational databases like Hadoop have emerged as a leading data management platform for

big data. In Hadoop's Distributed File System (HDFS), data is stored as 'key and data-value' combinations. Google BigFile is another prominent technology. NoSQL is emerging as a popular language to access and manage non-relational databases. There is a matching Data Warehousing system called Hive along with its own PigSQL language. The opensource stack of programming languages (such as Pig) and other tools help make Hadoop a powerful and popular tool.

2. Massively parallel computing: Given the size of data, it is useful to divide and conquer the problem quickly using multiple processors simultaneously. Parallel processing allows for the data to be processed by multiple machines so that results can be achieved sooner. MapReduce algorithm, originally generated at Google for doing searches faster, has emerged as a popular parallel processing mechanism. The original problem is divided into smaller problems, which are then mapped to multiple processors that can operate in parallel. The outputs of these processors are passed to an output processor that reduces the output to a single stream, which is then sent to the end user. Figure 13.4 shows an example of a Map-Reduce algorithm.

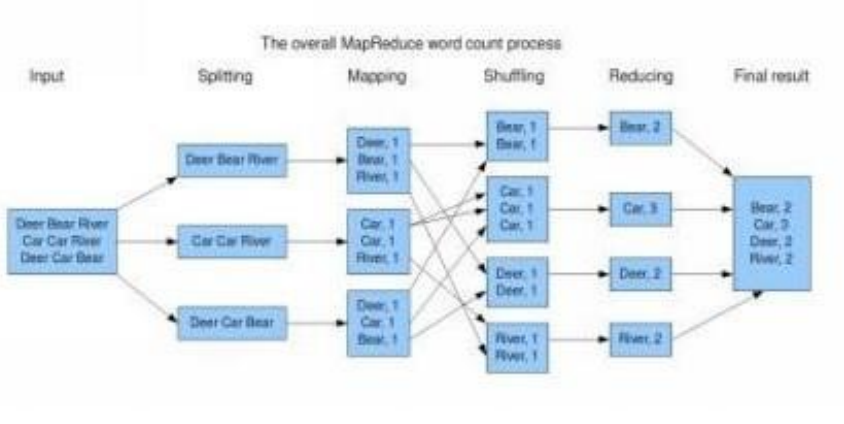

Fig.3.7 Map Reduce Algorithm

Unstructured Information Management Architecture (UIMA).
This is one of elements in the "secret sauce" behind IBM' Watson's system that reads massive amounts of data, and organizes for just-in-time processing. Watson beat the Jeopardy (quiz program) champion in 2011 and is now used for many business applications, like diagnosis, in health care situations. Natural language processing is another capability that helps extend the power of big data technologies.

Big Data Technology is mainly classified into two types:

1. **Operational Big Data Technologies**
2. **Analytical Big Data Technologies**

Firstly, The Operational Big Data is all about the normal day to day data that we generate. This could be the **Online Transactions, Social Media,** or the data from a **Particular Organisation** etc.

You can even consider this to be a kind of Raw Data which is used to feed the **Analytical Big Data Technologies.**

A few examples of **Operational Big Data Technologies** are as follows:

Fig.3.7 Operational Big Data Technologies

- Online ticket bookings, which includes your Rail tickets, Flight tickets, movie tickets etc.
- Online shopping which is your Amazon, Flipkart, Walmart, Snap deal and many more.
- Data from social media sites like Facebook, Instagram, what's app and a lot more.
- The employee details of any Multinational Company.

So, with this let us move into the Analytical Big Data Technologies.

Analytical Big Data is like the advanced version of Big Data Technologies. It is a little complex than the Operational Big Data. In short, Analytical big data is where the actual performance part comes into the picture and the crucial real-time business decisions are made by analyzing the Operational Big Data.

Few examples of Analytical Big Data Technologies are as follows:

Fig.3.7 Operational Big Data Technologies

- Stock marketing
- Carrying out the Space missions where every single bit of information is crucial.
- Weather forecast information.
- Medical fields where a particular patients health status can be monitored.

3.4.1 Top Big Data Technologies

Top big data technologies are divided into **4** fields which are classified as follows:

- **Data Storage**
- **Data Mining**
- **Data Analytics**
- **Data Visualization**

Data Storage

Hadoop

Hadoop Framework was designed to store and process data in a Distributed Data Processing Environment with commodity hardware with a simple programming model. It can Store and Analyse the data present in different machines with High Speeds and Low Costs.

Developed by: Apache Software Foundation in the year 2011 10th of Dec.
Written in: JAVA
Current stable version: Hadoop 3.11

MongoDB

The NoSQL Document Databases like MongoDB, offer a direct alternative to the rigid schema used in Relational Databases. This allows MongoDB to offer Flexibility while handling a wide variety of Datatypes at large volumes and across Distributed Architectures.

Developed by: MongoDB in the year 2009 11th of Feb
Written in: C++, Go, JavaScript, Python
Current stable version: MongoDB 4.0.10

RainStor

Rain Stor is a software company that developed a Database Management System of the same name designed to Manage and Analyse Big Data for large enterprises. It uses Deduplication Techniques to organize the process of storing large amounts of data for reference.

Developed by: RainStor Software company in the year 2004.
Works like: SQL
Current stable version: RainStor 5.5

Hunk lets you access data in remote Hadoop Clusters through virtual indexes and lets you use the Splunk Search Processing Language to analyse your data. With Hunk, you can Report and Visualize large amounts from your Hadoop and NoSQL data sources.

Developed by: Splunk INC in the year 2013.
Written in: JAVA
Current stable version: Splunk Hunk 6.2

Data Mining

Presto

Presto is an open source **Distributed SQL Query Engine** for running **Interactive Analytic Queries** against data sources of all sizes ranging from Gigabytes to Petabytes. Presto allows querying data in Hive, Cassandra, **Relational Databases** and **Proprietary Data Stores.**

Developed by: Apache Foundation in the year 2013.
Written in: JAVA
Current stable version: Presto 0.22

RapidMiner

RapidMiner is a Centralized solution that features a very powerful and robust Graphical User Interface that enables users to Create, Deliver, and maintain Predictive Analytics. It allows creating very Advanced Workflows, Scripting support in several languages.

Developed by: RapidMiner in the year 2001
Written in: JAVA
Current stable version: RapidMiner 9.2

Elasticsearch

Elasticsearch is a Search Engine based on the Lucene Library. It provides a Distributed, MultiTenant-capable, Full-Text Search Engine with an HTTP Web Interface and Schema-free JSON documents.

Developed by: Elastic NV in the year 2012.
Written in: JAVA
Current stable version: ElasticSearch 7.1

Data Analytics

Kafka

Apache Kafka is a Distributed Streaming platform. A streaming platform has Three Key Capabilities that are as follows:

- Publisher
- Subscriber
- Consumer

This is similar to a Message Queue or an Enterprise Messaging System.

Developed by: Apache Software Foundation in the year 2011

Written in: Scala, JAVA
Current stable version: Apache Kafka 2.2.0

Splunk

Splunk captures, Indexes, and correlates Real-time data in a Searchable Repository from which it can generate Graphs, Reports, Alerts, Dashboards, and Data Visualizations. It is also used for Application Management, Security and Compliance, as well as Business and Web Analytics.

Developed by: Splunk INC in the year 2014 6th May
Written in: AJAX, C++, Python, XML
Current stable version: Splunk 7.3

KNIME

KNIME allows users to visually create Data Flows, Selectively execute some or All Analysis steps, and Inspect the Results, Models, and Interactive views. KNIME is written in Java and based on Eclipse and makes use of its Extension mechanism to add Plugins providing Additional Functionality.

Developed by: KNIME in the year 2008
Written in: JAVA
Current stable version: KNIME 3.7.2

R

R is a Programming Language and free software environment for Statistical Computing and Graphics. The R language is widely used among Statisticians and Data Miners fordeveloping Statistical Software and majorly in Data Analysis.

Developed by: R-Foundation in the year 2000 29th Feb
Written in: Fortran
Current stable version: R-3.6.0

BlockChain

BlockChain is used in essential functions such as payment, escrow, and title can also reduce fraud, increase financial privacy, speed up transactions, and internationalize markets.

BlockChain can be used for achieving the following in a Business Network Environment:

- Shared Ledger: Here we can append the Distributed System of records across a Business network.
- Smart Contract: Business terms are embedded in the transaction Database and Executed with transactions.
- Privacy: Ensuring appropriate Visibility, Transactions are Secure, Authenticated and Verifiable
- Consensus: All parties in a Business network agree to network verified transactions.

Developed by: Bitcoin
Written in: JavaScript, C++, Python
Current stable version: Blockchain 4.0

3.4 Data Visualization

Data Visualization is the art and science of making data easy to understand and consume, for the end user. Ideal visualization shows the right amount of data, in the right order, in the right visual form, to convey the high priority information. The right visualization requires an understanding of the consumer's needs, nature of the data, and the many tools and techniques available to present data. The right visualization arises from a complete understanding of the totality of the situation. One should use visuals to tell a true, complete and fast-paced story.

Data visualization is the last step in the data life cycle. This is where the data is processed for presentation in an easy-to-consume manner to the right audience for the right purpose. The data should be converted into a language and format that is best preferred and understood by the consumer of data. The presentation should aim to highlight the insights from the data in an actionable manner. If the data is presented in too much detail, then the consumer of that data might lose interest and the insight.

Excellence in Visualization

Data can be presented in the form of rectangular tables, or it can be presented in colorful graphs of various types. “Small, non-comparative, highly-labeled data sets usually belong in tables” – (Ed Tufte, 2001, p 33). However, as the amount of data grows, graphs are preferable. Graphics help give shape to data. Tufte, a pioneering expert on data visualization, presents the following objectives for graphical excellence:

1. Show, and even reveal, the data: The data should tell a story, especially a story hidden in large masses of data. However, reveal the data in context, so the story is correctly told.

2. Induce the viewer to think of the substance of the data: The format of the graph should be so natural to the data, that it hides itself and lets data shine.

3. Avoid distorting what the data have to say: Statistics can be used to lie. In the name of simplifying, some crucial context could be removed leading to distorted communication.

4. Make large data sets coherent: By giving shape to data, visualizations can help bring the data together to tell a comprehensive story.

5. Encourage the eyes to compare different pieces of data: Organize the chart in ways the eyes would naturally move to derive insights from the graph.

6. Reveal the data at several levels of detail: Graphs leads to insights, which raise further curiosity, and thus presentations should help get to the root cause.

7. Serve a reasonably clear purpose – informing or decision-making.

8. Closely integrate with the statistical and verbal descriptions of the dataset: There should be no separation of charts and text in presentation. Each mode should tell a complete story. Intersperse text with the map/graphic to highlight the main insights.

Context is important in interpreting graphics. Perception of the chart is as important as the actual charts. Do not ignore the intelligence or the biases of the reader. Keep the template consistent, and only show variations in data. There can be many excuses for graphical distortion. E.g. “we are just approximating.” Quality of information transmission comes prior to aesthetics of chart. Leaving out the contextual data can be misleading.

A lot of graphics are published because they serve a particular cause or a point of view. It is particularly important when in a for-profit or politically 96 contested environments. Many related dimensions can be folded into a graph. The more the dimensions that are represented in a graph, the richer and more useful the chart become. The data visualizer should understand the client’s objects and present the data for accurate perception of the totality of the situation.

Types of Charts

There are many kinds of data as seen in the caselet above. Time series data is the most popular form of data. It helps reveal patterns over time. However, data could be organized around alphabetical list of things, such as countries or products or salespeople. Figure below shows some of the popular chart types and their usage.

1. Line graph. This is a basic and most popular type of displaying information. It shows data as a series of points connected by straight line segments. If mining with time-series data, time is usually shown on the x-axis. Multiple variables can be represented on the same scale on y-axis to compare of the line graphs of all the variables.

2. Scatter plot: This is another very basic and useful graphic form. It helps reveal the relationship between two variables. In the above caselet, it shows two dimensions: Life Expectancy and Fertility Rate. Unlike in a line graph, there are no line segments connecting the points.

3. Bar graph: A bar graph shows thin colorful rectangular bars with their lengths being proportional to the values represented. The bars can be plotted vertically or horizontally. The bar graphs use a lot of more ink than the line graph and should be used when line graphs are inadequate.

4. Stacked Bar graphs: These are a particular method of doing bar graphs. Values of multiple variables are stacked one on top of the other to tell an interesting story. Bars can also be normalized such as the total height of every bar is equal, so it can show the relative composition of each bar.

5. Histograms: These are like bar graphs, except that they are useful in showing data frequencies or data values on classes (or ranges) of a numerical variable.

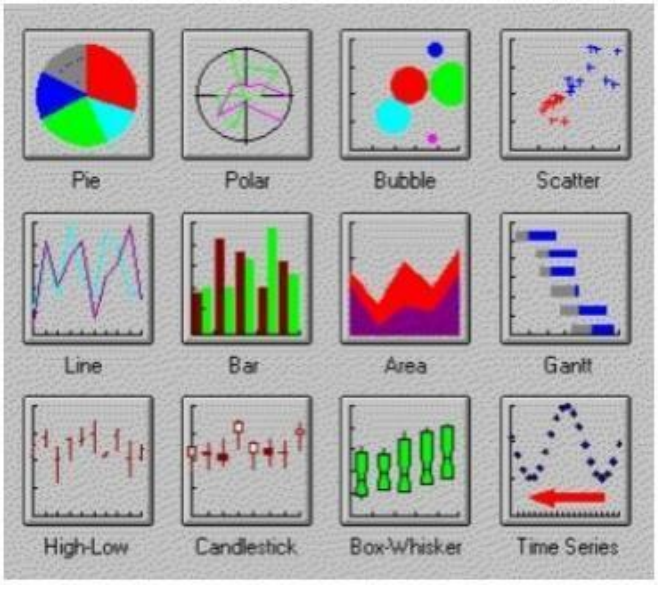

Fig. Types of Charts

6. Pie charts: These are very popular to show the distribution of a variable, such as sales by region. The size of a slice is representative of the relative strengths of each value.

7. Box charts: These are special form of charts to show the distribution of variables. The box shows the middle half of the values, while whiskers on both sides extend to the extreme values in either direction.

8. Bubble Graph: This is an interesting way of displaying multiple dimensions in one chart. It is a variant of a scatter plot with many data points marked on two dimensions. Now imagine that each data point on the graph is a bubble (or a circle) … the size of the circle and the color fill in the circle could represent two additional dimensions.

9. Dials: These are charts like the speed dial in the car, that shows whether the variable value (such as sales number) is in the low range, medium range, or high range. These ranges could be colored red, yellow and gree to give an instant view of the data.

10. Geographical Data maps : These are particularly useful maps to denote statistics. Figure below shows a tweet density map of the US. It shows where the tweets emerge from in the US.

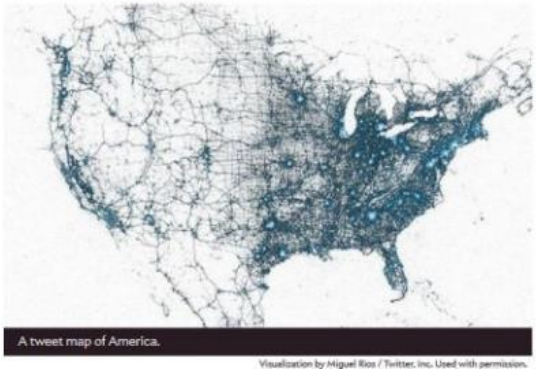

Fig. Geographical Data Maps

11. Pictographs: One can use pictures to represent data. E.g. Figure below shows the number of litres of water needed to produce one pound of each of the products, where images are used to show the product for easy reference. Each droplet of water also represents 50 litres of water.

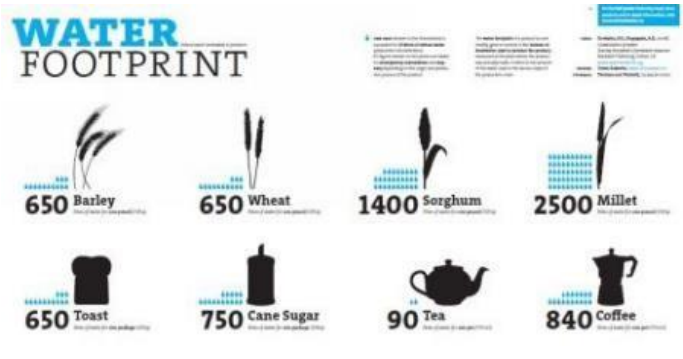

Fig. Pictograph

Visualization Example

To demonstrate how each of the visualization tools could be used, imagine an executive for a company who wants to analyze the sales performance of his division. Table below shows the important raw sales data for the current year, alphabetically sorted by Product names.

Product	Revenue	Orders	SalesPers
AA	9731	131	23
BB	355	43	8
CC	992	32	6
DD	125	31	4
EE	933	30	7
FF	676	35	6
GG	1411	128	13
HH	5116	132	38
JJ	215	7	2
KK	3833	122	50
LL	1348	15	7
MM	1201	28	13

Table: Raw Data Performance

To reveal some meaningful pattern, a good first step would be to sort the table by Product revenue, with highest revenue first. We could total up the values of Revenue, Orders, and Sales persons for all products. We can also add some important ratios to the right of the table below.

Product	Revenue	Orders	SalesPers	*Rev/Order*	*Rev/SalesP*	*Orders/SalesP*
AA	9731	131	23	*74.3*	*423.1*	*5.7*
HH	5116	132	38	*38.8*	*134.6*	*3.5*
KK	3833	122	50	*31.4*	*76.7*	*2.4*
GG	1411	128	13	*11.0*	*108.5*	*9.8*
LL	1348	15	7	*89.9*	*192.6*	*2.1*
MM	1201	28	13	*42.9*	*92.4*	*2.2*
CC	992	32	6	*31.0*	*165.3*	*5.3*
EE	933	30	7	*31.1*	*133.3*	*4.3*
FF	676	35	6	*19.3*	*112.7*	*5.8*
BB	355	43	8	*8.3*	*44.4*	*5.4*
JJ	215	7	2	*30.7*	*107.5*	*3.5*
DD	125	31	4	*4.0*	*31.3*	*7.8*
Total	**25936**	**734**	**177**	*35.3*	*146.5*	*4.1*

Table: Sorted data, with additional ratios

There are too many numbers on this table to visualize any trends in them. The numbers are in different scales so plotting them on the same chart would not be easy. E.g. the Revenue numbers are in thousands while the SalesPers numbers and Orders/SalesPers are in the single or double digit.

One could start by visualizing the revenue as a pie-chart. The revenue proportion drops significantly from the first product to the next. (Figure below). It is interesting to note that the top 3 products produce almost 75% of the revenue.

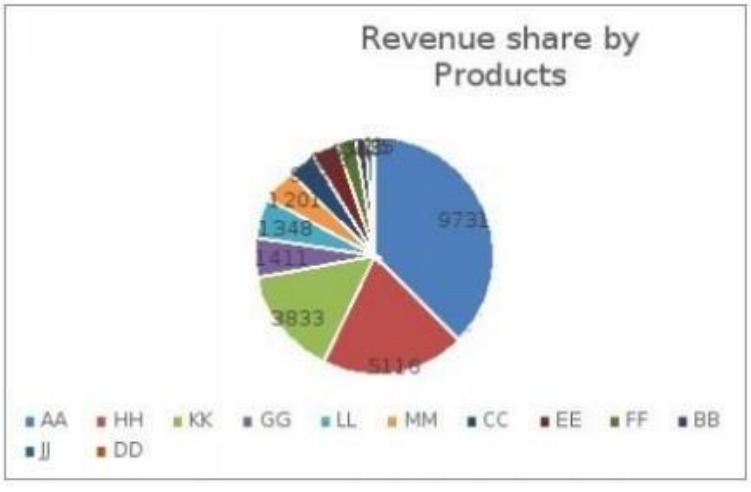

Fig. Revenue share by Products

The number of orders for each product can be plotted as a bar graph (Figure below). This shows that while the revenue is widely different for the top four products, they have approximately the same number of orders.

Fig. Order by Products

Therefore, the orders data could be investigated further to see order patterns. Suppose additional data is made available for Orders by their size. Suppose the orders are chunked into 4 sizes: Tiny, Small, Medium, and Large. Additional data is shown in the table below.

Product	Total Orders	Tiny	Small	Medium	Large
AA	131	5	44	70	12
HH	132	38	60	30	4
KK	122	20	50	44	8
GG	128	52	70	6	0
LL	15	2	3	5	5
MM	28	8	12	6	2
CC	32	5	17	10	0
EE	30	6	14	10	0
FF	35	10	22	3	0
BB	43	18	25	0	0
JJ	7	4	2	1	0
DD	31	21	10	0	0
Total	734	189	329	185	31

Table: Additional data on order sizes

Figure below is a stacked bar graph that shows the percentage of Orders by size for each product. This chart brings a different set of insights. It shows that the product HH has a larger proportion of tiny orders. The products at the far right have a large number of tiny orders and very few large orders.

Fig. Product Orders by Order Size

Visualization Example phase -2

The executive wants to understand the productivity of salespersons. This analysis could be done both in terms of the number of orders, or revenue, per salesperson. There could be two separate graphs, one for the number of orders per salesperson, and the other for the revenue per salesperson. However, an interesting way is to plot both measures on the same graph to give a more complete picture. This can be done even when the two data have different scales. The data is here resorted by number of orders per salesperson.

Figure below shows two line graphs superimposed upon each other. One line shows the revenue per salesperson, while the other shows the number of orders per salesperson. It shows that the highest productivity of 5.3 orders per sales person, down to 2.1 orders per salesperson. The second line, the blue line shows the revenue per sales person for each for the products. The revenue per salesperson is highest at 630, while it is lowest at just 30. And thus additional layers of data visualization can go on for this data set.

Fig. Salesperson productivity by product

Tips for Data Visualization

To help the client in understanding the situation, the following considerations are important:

1. Fetch appropriate and correct data for analysis. This requires some understanding of the domain of the client and what is important for the client. E.g. in a business setting, one may need to understand the many measure of profitability and productivity.

2. Sort the data in the most appropriate manner. It could be sorted by numerical variables, or alphabetically by name.

3. Choose appropriate method to present the data. The data could be presented as a table, or it could be presented as any of the graph types.

4. The data set could be pruned to include only the more significant elements. More data is not necessarily better, unless it makes the most significant impact on the situation.

5. The visualization could show additional dimension for reference such as the expectations or targets with which to compare the results.

6. The numerical data may need to be binned into a few categories. E.g. the orders per person were plotted as actual values, while the order sizes were binned into 4 categorical choices.

7. High-level visualization could be backed by more detailed analysis. For the most significant results, a drill-down may be required.

8. There may be need to present additional textual information to tell the whole story. For example, one may require notes to explain some extraordinary results.

Tableau

Tableau is a Powerful and Fastest growing Data Visualization tool used in the Business Intelligence Industry. Data analysis is very fast with Tableau and the Visualizations created are in the form of Dashboards and Worksheets.

Developed by: TableAU 2013 May 17th
Written in: JAVA, C++, Python, C
Current stable version: TableAU 8.2

Plotly

Mainly used to make creating Graphs faster and more efficient. API libraries for Python, R, MATLAB, Node.js, Julia, and Arduino and a REST API. Plotly can also be used tostyle Interactive Graphs with Jupyter notebook.

Developed by: Plotly in the year 2012
Written in: JavaScript

Current stable version: Plotly 1.47.4

3.5 Emerging Big Data Technologies

TensorFlow

TensorFlow has a Comprehensive, Flexible Ecosystem of tools, Libraries and Community resources that lets Researchers push the state-of-the-art in Machine Learning and Developers can easily build and deploy Machine Learning powered applications.

Developed by: Google Brain Team in the year 2019
Written in: Python, C++, CUDA
Current stable version: TensorFlow 2.0 beta

Beam

Apache Beam provides a Portable API layer for building sophisticated Parallel-Data Processing Pipelines that may be executed across a diversity of Execution Engines or Runners.

Developed by: Apache Software Foundation in the year 2016 June 15th
Written in: JAVA, Python
Current stable version: Apache Beam 0.1.0 incubating.

Docker

Docker is a tool designed to make it easier to Create, Deploy, and Run applications by using **Containers.** Containers allow a developer to Package up an application with all of the partsit needs, such as Libraries and other Dependencies, and Ship it all out as One Package.

Developed by: Docker INC in the year 2003 13th of March.
Written in: Go
Current stable version: Docker 18.09

Airflow

Apache Airflow is a WorkFlow Automation and Scheduling System that can be used to author and manage Data Pipelines. Airflow uses workflows made of Directed Acyclic Graphs (DAGs) of tasks. Defining Workflows in code provides Easier Maintenance, Testing and Versioning.

Developed by: Apache Software Foundation on May 15th 2019
Written in: Python
Current stable version: Apache AirFlow 1.10.3

Kubernetes

Kubernetes is a Vendor-Agnostic Cluster and Container Management tool, Open Sourced by Google in 2014. It provides a platform for Automation, Deployment, Scaling, and Operations of Application Containers across Clusters of Hosts.

Developed by: Cloud Native Computing Foundation in the year 2015

21st of JulyWritten in: Go

Current stable version: Kubernetes 1.14

UNIT -4
Data Science and Ethical Issues

4.1 Data Protection/Privacy

Background

While it may well be that only companies will survive that rigorously exploit (big) data, one should not forget that data science and data exploitation must not lead to an **infringement of privacy rights**. Data protection and privacy are protected by the Swiss constitution as fundamental constitutional rights. Data protection laws are meant to specify the constitutional rights of privacy. Those data protection laws also have to be taken into accoun**t in the field of data science.**

Swiss data protection law is mainly set forth in the Federal Act on Data Protection of June 19, 1992 (DPA), and the Swiss Federal Ordinance to the Federal Act on Data Protection of June 14, 1993 (DPO). In the EU, the General Data **Regulation (GDPR)** has entered into force and will apply as of May 25, 2018. The Swiss DPA is currently under revision and it is expected that it will be strongly influenced by the GDPR, in particular because cross-border dat**a transfers are daily business.**

4.1.1 Personal Data:

Swiss data protection laws only deal with the processing of personal data. Obviously, not all data is personal data. Under Swiss law, personal data is defined as "all information relating to an identified or identifiable person" (Article 3 let. a DPA). A person is considered to be identifiable if identification is possible without undue efforts and one has to expect that this will possibly be done.

While this definition seems clear, there is a large spectrum between data that is clearly connected to an identifiable person and data that cannot in any way be re-identified.

De-identification of data generally is used to denominate a process of "removing or obscuring any personally identifiable information from individual records in a way that minimizes the risk of unintended disclosure of the identity of individuals and **information about them".** Therefore, de-identified data may theoretically still be linked to individuals, for example, using a code, algorithm, or pseudonym.

The definition of "pseudonymization" in the GDPR is somewhat different: "'pseudonymisation' means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organizational measures to ensure that the personal data are not attributed to an identified or identifiable na**tural person"**.

Anonymization on the other hand is a process of data de-identification leading to data where individual records cannot be linked back to an individual as they do not include the required translation variables to do so. Consequently, anonymized data, as it is often used in data science, is generally not subject to the DPA. De-identification may also be sufficient to exclude data from the scope of DPA, if the re-identification is not possible without undue efforts or if one does not have to expect that this will possibly be done.

However, data scientists should be aware that the process of anonymization or de-identification of data, which currently constitutes personal data, does, in itself, constitute the processing of personal data and, thus, is subject to the DPA. Only the result of the anonymization (and possibly of the de-identification) is no longer relevant from a perspective of data protection laws.

Also, there is no guarantee that de-identification and/or anonymization completely precludes re-identification of the data subject. On the contrary: in particular in connection with Big Data, if there is a large amount of data, re-identification of the data subject becomes more likely and possible (Baeriswyl 2013, p. 15).7 Once such re-identification becomes possible "without undue efforts" and one has to expect that this will possibly be done, the data becomes personal data, and the DPA applies. Consequently, if one has sufficient data to conduct such re-identification, one will have to comply with the DPA (Weber 2014, p. 20). So, while the process of re-identification itself constitutes a data processing that is relevant under the Swiss DPA, one has to be aware that the DPA becomes applicable already at an earlier stage, that is, once re-identification is "possible without undue efforts and one has to expect that this will possibly be **done"**. If personal data is generated by accident, nevertheless, Swiss data protection laws would apply. Finally, even if the data scientist provides de-identified and/or anonymized data to third parties, data protection laws will have to be complied with, if the data scientist has to expect that re-individualization will take place**.** This is an issue that should be further analyzed in joint research activities conducted by IT specialists and legal scholars.

Thus, the boundary between personal data and other data is somewhat vague, in particular because of the technical developments; data that cannot be re-individualized today may well become related to an identifiable person tomorrow, and, thus, become personal data .

Consequently, even anonymization or de-individualization of the respective data does not completely exclude that data protection laws will be applicable to the activities of a data scientist. This is true irrespective of whether the data is used only internally in a data product or whether it is visible also externally and irrespective of the effect of the data product on the data subject concerned (e.g., whether you use the data for personalized pricing or to achieve better usability of **software for the data subject).**

4.1.2 Privacy by Design

Sometimes, legal developments are outpaced by technological developments. Data protection laws try to address this issue by provisions concerning "privacy by design"—the GDPR as well as the draft for a revision of the DPA. Privacy by design is an approach which takes privacy into account already in the phase of designing a product or a data analysis.

The idea of including this principle into the relevant laws is that law and technology should complement each other and that technologies, which already take privacy into account, are necessary to help implement data protection laws (Legislative Message DPA 2017, p. 7029). Technology may be used to enhance data security and, at the same time, the level of protection of personal data.

In the draft to the revised DPA, the principle requires that technical and organizational measures have to be set up in order for the data processing to meet the data protection regulations. It has to be considered from as early as the planning stage. The purpose is to achieve that systems for data processing are engineered (from a technological and organizational perspective) from the beginning in a way that they comply with data protection principles**.**

While this is rather vague, there are already certain reports and principles that can be used when trying to determine what "privacy by design" requires. Some guidance can be found, for example, in the following "7 foundational principles" of privacy by design

1. Proactive not reactive, preventive not **remedial:** The privacy by design approach aims to identify, anticipate, and prevent privacy invasive events before they arise. It does not wait for privacy risks to materialize, nor does it offer remedies in case a privacy breach occurs.

2. Privacy as the default**:** The default settings deliver the maximum degree of privacy. No action is required by the individual in order to protect their privacy.

3. Privacy embedded into design**:** Privacy is integral to the system, without diminishing functionality. It becomes an essential component of the core functionality being delivered.

4. Full functionality—positive-sum, not zero-sum **:** Privacy by design accommodates all legitimate interests and objectives in a positive-sum "win-win" manner. It avoids the pretense of false dichotomies, such as privacy vs. security, demonstrating that it is possible to have both.

5. End-to-end-security—full lifecycle protection**:** Privacy must be protected by strong security measures throughout the entire lifecycle of the data involved; **from the cradle to the grave.**

6. Visibility and transparency—keep it open**:** The data subject is made fully aware of the personal data being collected, and of the purpose(s). Moreover, the component parts and operations remain visible and transparent.

7. Respect for user privacy—keep it user-centric Privacy measures are consciously designed around the interests and needs of individual users.

In addition, the European Union Agency for Network and Information Security has addressed the issue in its report "Privacy and Data Protection by Design—from policy to engineering," which tries to bridge the gap between the law and the available technologies. It can also provide further insight into this issue and is certainly a good reference for data scientists.

4.1.3 Privacy by Default

While "privacy by default" is listed as one of the "7 foundational principles" of privacy by design above, this principle is also explicitly mentioned in the GDPR as well as the draft for a revision of the DPA.

The respective legal provisions require that it is ensured by suitable settings that by default only such personal data are processed that are required for the respective purpose of the processing. The "default setting" is the setting that is automatically given or applied to a software application, computer program, or device, if not altered or customized by the user.

In other words, the respective data processing should—as a default—be as privacy friendly as possible, except if the data subject changes the default settings (Legislative Message DPA 2017, p. 7030), for example, to obtain additional functionalities. Such settings have to enable the data subject to make its own choices concerning privacy to a certain extent.

4.1.4 Automated Individual Decisions

Another provision in data protection law which could substantially affect the activities of data scientists concerns "automated individual decisions." The GDPR as well as the draft for a revision of the DPA restrict automated individual decision making under certain circumstances. The GDPR states that the "data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her". The draft for a revision of the DPA provides that a data subject has to be informed "of a decision which is taken exclusively on the basis of an automated processing, including profiling, which has legal effects on the data subject or affects him significantly".

This may lead to substantial difficulties in data science, in particular in cases where individual decisions are taken by algorithms. However, the GDPR only covers decisions based "solely" on automated processing while the term used in the draft for a revised DPA is "exclusively." So how should the term "solely" in the GDPR or "exclusively" in the draft to the revised DPA be interpreted?

While one could argue that it is already sufficient if a human was included at the end of the process to formally "make the decision," this would defy the purpose of the legal provisions. Rather, it should only be considered that the decision is not based solely or exclusively on automated processing, if a person "actively exercises any real influence on the outcome of a particular decision-**making process"** and actively assesses the result of the automated processing before this person takes the decision also formally.

There are, however, also exceptions to this requirement. One important exception is that the provisions will not apply if the automated process was based on the data subject's explicit consent. According to the GDPR, the data subjects must be provided with information not only of the existence of such automated decision making, but also of the logic involved and the significance

and envisaged consequences of such processing for the data subject, which also is a necessary foundation for a valid consent.

However, explaining (and understanding) what goes on, for example, in a neural network in terms of a generated outcome (i.e., why is this case decided that way?) is a difficult task, even for an expert (cf. Dong et al. 2017; Stock and Cisse 2017). It will be substantially more difficult to try and explain such issues (or other algorithms) to an average data subject. In particular, if one cannot easily trace the precise path of a neural network to a final answer, the description of automated decisions is open to interpretation. This difficulty may also affect the issue of validity of a data subject's consent, since such consent not only has to be freely given, specifically and unambiguously, but also has to be made on an "informed" basis (Article 4 (11) GDPR). And even in cases of valid consent, the data subjects will still have to be informed and the data subjects will have the

(1) right to obtain human intervention;

(2) right to express their point of view;

(3) right to obtain an explanation of the decision **reached**

(4) right to challenge **the decision (Recital 71 GDPR)**

Since algorithms are an important means of governing data streams, assessments of how an automated decision will affect the data subject may have to be made on a regular basis. However, this seems to be an impossible reality, should automated decisions become the norm (Naudits 2016).

4.2 Security of Data Science

In this section, we discuss challenges and solution approaches related to the security of data science methods and applications. Since any data product needs an infrastructure to run on, a piece of software that implements it, data that fuels it, and customers that feel comfortable using it, we provide a brief overview and references to more in-depth material on (1) infrastructure security, (2) software security, (3) data protection, and (4) data anonymization. Furthermore, we discuss (5) examples of attacks on machine learning algorithms.

Infrastructure Security

Infrastructure security is concerned with securing information systems against physical and virtual intruders, insider threats, and technical failures of the infrastructure itself. As a consequence, some of the more important building blocks to secure an infrastructure are access control, encryption of data at rest and in transit, vulnerability scanning and patching, security monitoring, network segmentation, firewalls, anomaly detection, server hardening, and (endpoint) security policies. Resources such as the NIST special publications series (National Institute of Standards and Technology 2017) or the CIS top 20 security controls (Center for Internet Security 2017) provide

guidance and (some) practical advice. However, getting all of this right is far from easy and failing might carry a hefty price tag.

In 2007, for example, Sony was accused of having some serious security vulnerabilities. In an interview, Sony's senior vice president of information security stated: "It's a valid business decision to accept the risk of a security breach. I will not invest $10 million to avoid a possible $1 million loss" (Holmes 2007). The data theft and outage of the PlayStation network in 2011 cost Sony $171 million (Schreier 2011). The Sony Pictures hack in 2014 (Risk Based Security 2014), where personal information of employees were stolen, cost Sony $35 million. Nevertheless, as Sony stated, it is indeed a valid business decision to limit investments in security. But such decisions should be made in full awareness of the value of the assets that are at stake, especially in light of the fact that massive amounts of user accounts or data can pose a very attractive target for cyber criminals: they could steal or destroy it and then ask for a ransom to restore it, they might sell it on the black market, or misuse it to perform other crimes.

The fact that many companies have failed to secure their infrastructure can be considered an anecdotal proof that this is a complex task and should not be done without involving security experts. This is even more true when big data systems are involved, since they might require the implementation of new use-case or product specific security measures (Moreno et al. 2016). For a checklist of what should be considered when building and securing big data systems, check out the top 100 best B. Tellenbach et al. practices in big data security and privacy (Cloud Security Alliance 2016). However, note that many of the best practices also apply to "normal" information systems.

Fortunately, data scientists do rarely have to build and secure an infrastructure from scratch. However, they often have to select, configure, and deploy base technologies and products such as MongoDB, Elasticsearch, or Apache Spark. It is therefore important that data scientists are aware of the security of these products. What are the security mechanisms they offer? Are they secure by default? Can they be configured to be secure or is there a need for additional security measures and tools? Recent events have demonstrated that this is often not the case.

In January 2017, 30,000 MongoDB instances were compromised (Pauli 2017b) because they were configured to accept unauthenticated remote connections. The underlying problem was that MongoDB versions before 2.6.0. were insecure by default. When installed, the installer did not force the user to define a password for the database admin account, and the database service listened on all network interfaces for incoming connections, not only the local one. This problem was well known and documented (Matherly 2015), but apparently, many operators of such instances didn't know or didn't care. Just one week later, the same hackers started to attack more than 35,000 Elastic search instances with ransomware (Pauli 2017a). Most of these instances were located on Amazon Web Services (AWS) and provided full read and write access without requiring authentication.

It is important to keep in mind that many of these new technologies are designed to facilitate easy experimentation and exploration, and not to provide enterprise grade security by default. The examples mentioned in the previous paragraph are certainly not the only ones that illustrate this problem. A broader study in the area of NoSQL databases revealed that many products and technologies do not support fundamental security features such as database encryption and secure

communication (Sahafizadeh and Nematbakhsh 2015). The general advice here is that before setting up such a technology or product, it is important to check the security features it offers and to verify whether the default configuration is secure enough. If problems are identified, they should be fixed before the product is used.

Software Security

Software security sets the focus on the methodologies of how applications can be implemented and protected so that they do not have or expose any vulnerabilities. To achieve this, traditional software development life cycle (SDLC) models (Waterfall, Iterative, Agile, etc.) must integrate activities to help discover and reduce vulnerabilities early and effectively and refrain from the common practice to perform security-related activities only toward the end of the SDLC as part of testing. A secure SDLC (SSDLC) ensures that security assurance activities such as security requirements, defining the security architecture, code reviews, and penetration tests, are an integral part of the entire development process.

An important aspect of implementing an SSDLC is to know the threats and how relevant they are for a specific product. This allows prioritizing the activities in the SSDLC. For data products, injection attacks and components that are insecure by default are among the biggest threats. Many data products are based on immature cutting-edge technology. They process data from untrusted sources including data from IoT devices, data from public data sources such as Twitter, and various kinds of user input, to control and use the data product.

For instance, if the code assembles SQL queries by concatenating user input and instructions for the database, this can turn out badly. As an example, consider the following line of code where a SELECT query is built and where userinput is provided by the user:

String query = *"SELECT name, description from Product WHERE name LIKE '%" + userinput + "%'";*

If the user (an attacker in this case) specifies the following data as userinput,

' UNION SELECT username, password FROM User—

then the following query is built:

SELECT name, description from Product WHERE name LIKE '%' UNION SELECT username, password FROM User--%'

This query is syntactically correct (note that—is used in SQL for comments, which means that the part—%' will be ignored by the database system) and will not only return all products, but also all usernames and password that are stored in table User.

The solution to this so-called SQL injection problem seems simple: input data must be sanitized so that if the data contains SQL commands, it is just interpreted as textual data and not as a

potentially harmful SQL command. Another safeguard to protect from SQL injection is to use only minimal access rights for the technical database user that executes the query. This cannot completely prevent SQL injection, but in case of a vulnerability, it serves as a damage control mechanism to make sure that the amount of data that can be accessed by the attacker is limited.

However, although the mechanisms to prevent SQL injection vulnerabilities are well known, history shows that they are not used consistently in practice—even if incidents related to SQL injection regularly make it into the headlines of mass media. For instance, in 2008, SQL injection was used to steal more than 134 million credit card data records from Heartland Payment Systems (Krebs 2009). Three years later, Global Payment Systems faced the same problem and lost about $92.2 million during the incident (Krebs 2012). Even now, the problem is still around. In 2016, data of 55 million voters were stolen from Comelec, the Philippines Commission on Elections (Estopace 2016), and an SQL injection vulnerability might also have played an important role in the incident of the Panama Papers (Burgees and Temperton 2016), where 11.5 million financial documents about offshore entities were leaked.

Clearly, SQL might not see widespread use in big data systems. New technologies such as NoSQL databases are far more prominent. However, their security history does not look much better, as a recent paper demonstrated similar issues with injection attacks as SQL (Ron et al. 2016).

One reason why it is difficult to get rid of such vulnerabilities is that preventive measures have to be considered by the developers and integrated into the code. If they are not aware of such risks and security is not a crucial part of the SDLC they are employing, it is very likely that vulnerabilities creep into the code because countermeasures are missing completely or are implemented incorrectly. There exists also no magic bullet in the sense of tools or formal proofs that can easily verify whether a piece of software is secure, although there exist tools that can detect some vulnerabilities. A good overview in this context is provided in (Software Testing Help 2017).

In general, the following steps help to address common software security problems when building a (software) product:

- Make sure that third party technology or products used are as mature as possible.
- Make sure that third party technology or products used offer a broad spectrum of security features and access controls options.
- Make sure that you have an SSDLC in place.

A good starting point to learn more about how to develop secure software are the SSDLC models of Microsoft (Microsoft 2017b) and the Open Web Application Security Project OWASP (OWASP 2017a). For more specific advice on what to consider when developing web services and web applications, OWASP (2017b) or Li and Xue (2014) offer well-suited sources. OWASP (2017b) lists the top 10 (web-) application security risks and provides technical guidance on how to test for them and how to avoid them. Five important takeaways from there are that developers should check their web applications and services for the following problems: • Incorrect or lack of input validation and data sanitation so that an attacker can trick an interpreter or query engine to do things that were not intended.

• Incorrect implementation of authentication and session management.

• Exposure of sensitive data because of problems like

(1) insufficient or missing data encryption at rest and in motion,
(2) password stores that do not use strong adaptive and salted hashing functions with a work factor (e.g., PBKDF21 or bcrypt2), or data leakage in log files.

• Incorrect implementation of the mechanisms to restrict what an authenticated user is allowed to do. For example, checks whether a user has the right permissions to execute an action might be done for all actions that a user can trigger via URL 1 https://tools.ietf.org/html/rfc2898#page-9 2 https://www.usenix.org/legacy/events/usenix99/provos/provos_html/node5.html 15 Security of Data Science and Data Science for Security 271 entries that are exposed in the web-interface—but not for actions that could be triggered by accessing portions of a website that are not exposed by such entries (forceful browsing).

• Use of insecure configurations as a result of insecure default configurations, incomplete or ad hoc configurations, outdated configurations, open cloud storage, misconfigured HTTP headers, verbose error messages containing sensitive information, or other root causes.

Data Protection

A core activity in data science is the processing of (large amounts of) data. For most processing tasks, the data must be available in unencrypted form. This has two main drawbacks. The first one is that when security measures such as access control fail, attackers can easily steal the data and make use of any information it contains. To make this more difficult, the data should always be stored in encrypted form. This way, the attacker must steal the data when it is being processed or manage to steal the keys used to encrypt it.

The second drawback is that the vast amount of processing power available in data centers around the world cannot be exploited if the data contains confidential information or is subject to data protection laws prohibiting the processing by (foreign) third parties. For such cases, it would have to be possible to do the processing in the encrypted space. Searchable encryption and homomorphic encryption (Prasanna and Akki 2015) offer interesting properties with this regard.

Searchable encryption (SE) introduced by Song et al. (2000) [see Bösch et al. (2014)] for an overview of different approaches) can be divided into many different subgroups. The core logic mostly consists of building an encrypted keyword search index on the client side. A search is then performed using trapdoor functions. A trapdoor function is a function that is easy to compute in one direction, but that is difficult to compute in the inverse direction unless one knows a secret. The most basic algorithms allow only queries with a single keyword and have performance issues when new data is added. If data is frequently modified, removed, or added, dynamic data search algorithms are required. Fuzzy-keyword search extends the algorithm to tolerate (some) spelling mistakes. There are also methods that support multiple keywords per query. SE offers methods to perform ranked search, for example, by taking the access history of a user and the access frequency into account. Although some research prototypes have been developed and partly also made available for general use and experimentation [e.g., Popa et al. (2011)], several limitations must be overcome before SE can be used widely in practice. One of these limitations is that algorithms based on secret (symmetric) key cryptography usually require a key exchange over a secured

channel and offer only limited search capabilities compared to traditional search engines. Another one is that public key cryptography-based approaches are insufficient for modern big data systems because of substantial computational overhead.

Homomorphic encryption (HE) addresses the challenge to perform general computations on encrypted data. HE allows performing simple operations such as additions, multiplications, or quadratic formulas on ciphertext. It generates an encrypted result, which when decrypted, delivers the same result as if the operations were performed on the plaintext. This offers the ability to run calculations on untrusted devices without giving up on data privacy. Craig Gentry (2009) described the first Fully Homomorphic Encryption (FHE) scheme. This scheme allows performing any desirable function on encrypted data. Unfortunately, FHE is currently far away from practical use, as it increases memory consumption and processing times of even basic operations by about 6–7 orders of magnitude (Brown 2017). Therefore, Somewhat Homomorphic Encryption (SwHE) techniques are proposed. Compared to FHE, they provide better efficiency but do not support all operations [see, e.g., Gentry et al. (2012)]. On the implementation side, there are some HE research prototypes available such as by Halevi (2017). However, given the current state of HE technology, it is expected that several years of further research are required before HE is ready for productive use.

Privacy Preservation/Data Anonymization

In many cases, data science analyzes data of human individuals, for instance, health data. Due to legal and ethical obligations, such data should be anonymized to make sure the privacy of the individuals is protected. Data anonymization basically means that any data record in the data set should not be easily linkable to a particular individual. Obvious solutions include stripping the real name or the detailed address of individuals from the records, but experience teaches that this is usually not enough to truly anonymize the data.

For instance, in 2006, Netflix started an open competition with the goal to find algorithms that allow predicting user ratings for films. As a basis, Netflix provided a large data set of user ratings as training data, where both users and movies were replaced by numerical IDs. By correlating this data with ratings from the Internet Movie Database, two researchers demonstrated that it is possible to de-anonymize users (Narayanan and Shmatikov 2008). Another example is the Personal Genome Project, where researchers managed to de-anonymize about 90% of all participants (Sweeney et al. 2013). Their basic approach was to link information in the data records (birth date, gender, and ZIP code) with purchased voter registration lists and other publicly available information.

To overcome these issues, a more scientific approach toward anonymization is required. The question is the following: Is it possible to modify data such that the privacy of the participants is fully protected without losing the essence of the data and therefore its utility? In this context, “privacy protection” means that an attacker should not be able to learn any additional information about the individuals than what is directly provided by the data records, even when this data is correlated with other information. Past and more recent research activities have provided several approaches that can help to achieve this, including generalization (Sweeney 1997) and suppression (Cox 1980), k-anonymity (Samarati and Sweeney 1998), and differential privacy (Dwork 2006). Each method has its advantages and drawbacks.

Suppression is a basic form of trying to achieve anonymity by either deleting attributes or substituting them with other values. Generalization describes the approach to blur data byreplacing specific values with categories or ranges of values. An attribute containing the age of a person is then translated to a range, so 33 may result in 30–39. Combining these methods can leadto k-anonymity, which means that each record cannot be distinguished from at least k-1 other records when considering the personally identifying information in the records.

As an example, assume that a data set includes data records of individual. Each record includes gender, age range, and disease from which the person is suffering. Assume there are three records with gender female and age range 50–59. This basically corresponds to 3-anonymity, as these three records cannot be distinguished from one another based on the attributes gender and age range. k-anonymity also has its limitations, especially if the diversity of the non-anonymized attributes is low. In the previous example, let us assume that the disease is heart-related in all three cases. This implies that if an attacker knows that Angela, who is 55 years old, is included in the data set, then he directly knows that she is suffering from heart-related health problems, as all female persons between 50 and 59 in the data set are suffering from it.

The basic idea of differential privacy is that the actual values of the attributes of any single record in the data set should only have a very limited effect on the outcome of any analysis performed on the data. If this is the case, an attacker, when querying the data set, cannot learn anything about a specific individual in the data set as the received outcome is possibly independent of the actual attributes of this individual. This is basically achieved by adding some noise to the result before it is presented to the analyst. For example, let us assume that there are 100 records of 100 persons in a data set and the attacker knows of 99 persons whether they have a heart-related disease or not (we assume that 33 of them have such an issue), but he doesn't know this of the remaining person, which we name Alice. If the attacker performs the query "how many persons have a heart-related disease," then he directly knows Alice's condition: If the query returns 33, Alice has no heart-related problem, if it returns 34, Alice has a heart-related issue. When using differential privacy, the query would not return the actual value, but it would distort it a little bit, that is, the query would return a value in the neighbourhood of 33 or 34, such as 30, 32, or 35. What's important is that the returned value does not indicate whether the true value is 33 or 34, which implies the attacker cannot learn anything about Alice's condition.

Obviously, any data anonymization method has its price as it has a negative impact on the quality of the data and the precision of the results when doing data analysis. Suppressing and generalizing data removes information, which means that the results of any analysis performed on the data will become less precise. And in the case of differential privacy, we usually get results that are "close to the correct result," but that usually do not correspond to the exact result. But this is the price of 274 B. Tellenbach et al. protecting the privacy of the involved individuals and this also implies that in practice, it is important to carefully balance the privacy of the individuals and the required precision of the analyses. A concise overview about anonymization methods is given by Selvi and Pushpa (2015). Detailed information about privacy preserving data publishing and corresponding research can be found in the survey by Fung et al. (2010).

4.3 A look back at Data Science

Statistics, and the use of statistical models, are deeply rooted within the field of Data Science. Data Science started with statistics, and has evolved to include concepts/practices such as Artificial Intelligence, Machine Learning, and the Internet of Things, to name a few. As more and more data has become available, first by way of recorded shopping behaviors and trends, businesses have been collecting and storing it in ever greater amounts. With growth of the Internet, the Internet of Things, and the exponential growth of data volumes available to enterprises, there has been a flood of new information or Big Data. Once the doors were opened by businesses seeking to increase profits and drive better decision making, the use of Big Data started being applied to other fields, such as medicine, engineering, and social sciences.

A functional Data Scientist, as opposed to a general statistician, has a good understanding of software architecture and understands multiple programming languages. The Data Scientistdefines the problem, identifies the key sources of information, and designs the framework for collecting and screening the needed data. Software is typically responsible for collecting, processing, and modeling the data. They use the principles of Data Science, and all the related sub-fields and practices encompassed within Data Science, to gain deeper insight into the data assets under review.

There are many different dates and timelines that can be used to trace the slow growth of Data Science and its current impact on the Data Management industry, some of the more significant ones are outlined below.

In 1962, John Tukey wrote about a shift in the world of statistics, saying, "… as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt…I have come to feel that my central interest is in data analysis…" Tukey is referring to the merging of statistics and computers, at a time when statistical results were presented in hours, rather than the days or weeks it would take if done by hand.

In 1974, Peter Naur authored the *Concise Survey of Computer Methods*, using the term "Data Science," repeatedly. Naur presented his own convoluted definition of the new concept:

"The science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences."

In 1977, The IASC, also known as the International Association for Statistical Computing was formed. The first phrase of their mission statement reads, "It is the mission of the IASC to link traditional statistical methodology, modern computer technology, and the knowledge of domain experts in order to convert data into information and knowledge."

In 1977, Tukey wrote a second paper, titled *Exploratory Data Analysis*, arguing the importance of using data in selecting "which" hypotheses to test, and that confirmatory data analysis and exploratory data analysis should work hand-in-hand.

In 1989, the Knowledge Discovery in Databases, which would mature into the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, organized its first workshop.

In 1994, Business Week ran the cover story, *Database Marketing*, revealing the ominous news companies had started gathering large amounts of personal information, with plans to start strange new marketing campaigns. The flood of data was, at best, confusing to company managers, who were trying to decide what to do with so much disconnected information.

In 1999, Jacob Zahavi pointed out the need for new tools to handle the massive amounts of information available to businesses, in Mining Data for Nuggets of Knowledge. He wrote:

"Scalability is a huge issue in data mining… Conventional statistical methods work well with small data sets. Today's databases, however, can involve millions of rows and scores of columns of data… Another technical challenge is developing models that can do a better job analyzing data, detecting non-linear relationships and interaction between elements… Special data mining tools may have to be developed to address web-site decisions."

In 2001, Software-as-a-Service (SaaS) was created. This was the pre-cursor to using Cloud-based applications.

In 2001, William S. Cleveland laid out plans for training Data Scientists to meet the needs of the future. He presented an action plan titled, *Data Science: An Action Plan for Expanding the Technical Areas of the field of Statistics*. It described how to increase the technical experience and range of data analysts and specified six areas of study for university departments. It promoted developing specific resources for research in each of the six areas. His plan also applies to government and corporate research.

In 2002, the International Council for Science: Committee on Data for Science and Technology began publishing the *Data Science Journal*, a publication focused on issues such as the description of data systems, their publication on the internet, applications and legal issues.

In 2006, Hadoop 0.1.0, an open-source, non-relational database, was released. Hadoop was based on Nutch, another open-source database.

In 2008, the title, "Data Scientist" became a buzzword, and eventually a part of the language. DJ Patil and Jeff Hammerbacher, of LinkedIn and Facebook, are given credit for initiating its use as a buzzword.

In 2009, the term NoSQL was reintroduced (a variation had been used since 1998) by Johan Oskarsson, when he organized a discussion on "open-source, non-relational databases".

In 2011, job listings for Data Scientists increased by 15,000%. There was also an increase in seminars and conferences devoted specifically to Data Science and Big Data. Data Science had proven itself to be a source of profits and had become a part of corporate culture.

In 2011, James Dixon, CTO of Pentaho promoted the concept of Data Lakes, rather than Data Warehouses. Dixon stated the difference between a Data Warehouse and a Data Lake is that the Data Warehouse pre-categorizes the data at the point of entry, wasting time and energy, while a Data Lake accepts the information using a non-relational database (NoSQL) and does not categorize the data, but simply stores it.

In 2013, IBM shared statistics showing 90% of the data in the world had been created within the last two years.

In 2015, using Deep Learning techniques, Google's speech recognition, Google Voice, experienced a dramatic performance jump of 49 percent.

In 2015, Bloomberg's Jack Clark, wrote that it had been a landmark year for Artificial Intelligence (AI). Within Google, the total of software projects using AI increased from "sporadic usage" to more than 2,700 projects over the year.

In the past ten years, Data Science has quietly grown to include businesses and organizations world-wide. It is now being used by governments, geneticists, engineers, and even astronomers. During its evolution, Data Science's use of Big Data was not simply a "scaling up" of the data, but included shifting to new systems for processing data and the ways data gets studied and analyzed.

Data Science has become an important part of business and academic research. Technically, this includes machine translation, robotics, speech recognition, the digital economy, and search engines. In terms of research areas, Data Science has expanded to include the biological sciences, health care, medical informatics, the humanities, and social sciences. Data Science now influences economics, governments, and business and finance.

One result of the Data Science revolution has been a gradual shift to writing more and more conservative programming. It has been discovered Data Scientists can put too much time and energy into developing unnecessarily complex algorithms, when simpler ones work more effectively. As a consequence, dramatic "innovative" changes happen less and less often. Many Data Scientists now think wholesale revisions are simply too risky, and instead try to break ideas into smaller parts. Each part gets tested, and is then cautiously phased into the data flow.

4.4 Next Generation Data Science:

Data scientists do exceptionally complex work. Their productivity depends on having comprehensive, easy access to analytics tools, data and other assets for sustaining productivity in intricate, collaborative environments. A typical day in the working life of a data science professional may involve navigating the challenges of any or all of the following tasks:

- **Sourcing:** Acquire data from diverse data lakes, big data clusters, cloud data services and more
- **Preparing:** Discover, acquire, aggregate, curate, pipeline, model and visualize complex, multistructured data
- **Modeling:** Tap into libraries of algorithms and models for statistical exploration, data mining, predictive analytics, machine learning, natural-language processing and interactive visualization among other functions

- **Developing:** Prototype and program data applications to be executed within in-memory, streaming, cloud and other runtime environments
- **Governing:** Secure, manage, track, audit and archive data, algorithms, models, metadata and other assets throughout their lifecycles

Team dynamics

The typical data science team involves a multifaceted interplay of roles, functions and workflows. Each of the principal roles has to handle its own set of complexities.

Data scientists

As statistical modelers, data scientists drive each step in the lifecycle of data-driven applications. They bring a holistic view to solving problems that involve complex data, algorithms and statistical models. They engage in highly challenging collaborative arrangements that involve data engineers, developers, business analysts and others to ensure delivery of desired outcomes on business projects with data science at their core. Data scientists need a complex set of technical and business skills and knowledge.

Benjamin Skrainka, principal data scientist at Galvanize, said "A data scientist needs expertise within multiple disciplines. You need to be good at databases. You need some knowledge of software engineering. You need to know some machine learning. And you need to know some statistics."

Data engineers

Data professionals tasked to manage the process of gathering, organizing, cleansing and integrating data are data engineers, who then ensure that data flows smoothly throughout the data science lifecycle. Data engineers implement and optimize the systems and processes required by other data science professionals and other stakeholders. They also work with front-end developers when moving data science projects into production. And their roles are beginning to blend into those of data scientists.

Jason Hill, senior big data engineer and scientist at CA Technologies, says "the way that we handle it is to have everybody on one team. Traditionally, an engineer wrote the code, and a data scientist developed the algorithm—within silos in their own areas. Now they know each other's role and work together. We have data scientists that can write code and work with the engineer to develop algorithms."

Developers

As builders who craft analytics applications that incorporate algorithmic models developed by data scientists, developers enable business analysts and other users to realize preferred business outcomes from data science assets in their day-to-day work. Andy Gants, principal data scientist at Spare5, says that the ability of data scientists to collaborate with "a software development department [for] iterating within terms of implementing analyses into specific software features" is essential. "Learning how to perform with the software development department," says Gants, "proved to be quite a challenge."

Business analysts

The knowledge workers on the team are business analysts who use self-service analytics tools to develop predictive analysis, machine learning and other data-driven models without coding and without having to request assistance from data scientists. Data scientists themselves need business analysis skills to do their job effectively.

According to Cliff Click, CTO at Neurensic, "Data scientists need a good blend of domain knowledge and business expertise. They need to be extremely inquisitive and relentless at figuring out how to solve a particular problem. That means digging into different approaches and alternatives—not just building models and running algorithms, but also interpreting the results to drive new business opportunities."

Open collaboration

Data science teams deriving benefit from an integrated development environment (IDE) to boost their collaborative output makes perfect sense. IDEs provide a comprehensive, extensible, open framework for accessing tools, data and other resources needed to build, test and deploy executable assets into production environments.

As data science moves into the inner circle of next-generation developer competencies, data scientists can expect the emergence of an industry-standard IDE that is equivalent to the industry-standard Eclipse framework that IT management professionals have been using for years.

IBM Data Science Experience (DSX) is the open IDE for team data science. DSX, which carries forward the functionality of IBM's Data Scientist Workbench, offers several productivity features.

Open analytics workbench

DSX is an interactive, cloud-based, scalable and secure visual workbench for consolidating open-source tools, languages and libraries. It provides unified access to open-source tools and libraries—including Apache Spark, R, Python and Scala—as well as solutions from IBM, IBM **partners such as RStudio and H20.ai, and others through an extensible architecture.**

Simplified data and model management

DSX provides built-in connectivity to diverse data sources and simplified data ingestion, refinement, curating and analysis capabilities while integrating IBM DataWorks for data integration and wrangling. Follow-on releases are expected to deliver productivity features such as data shaping, Spark pipeline deployment, SPSS analytics algorithms, automated modeling and data preparation, model management and deployment, advanced visualizations, text analytics, and geospatial analytics and integration with IBM Watson Analytics.

Team collaboration and learning environment

DSX provides a team collaboration environment within which data science professionals can connect with one another and rapidly deploy high-quality applications into production environments. It enables team members to access project dashboards and learning resources; fork and share projects; exchange development assets such as data sets, models, projects, tutorials and

Jupyter notebooks; and deliver results. Follow-on releases are planned to include comments, userprofiles, data science competitions, Zeppelin notebooks and real-time collaboration.