# 1

# CHAPTER Big Data and Data Science

## 1.1 INFORMATION EXPLOSION

The biggest innovation of the second decade of twenty first century was the phenomenon of Expansion and movement of information from hands of elite society to hands of masses.

There are now more than **4 billion** people around the world using the internet. Well over half of the world's population is now online, with the latest data showing that nearly *a quarter of a billion* new users came online for the first time in 2017.

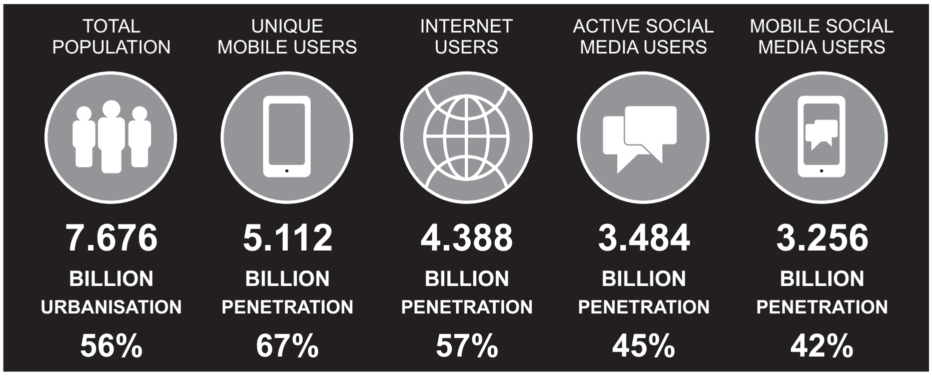

**Fig. 1.1** Display of Impact of Digitisation 2019 (reported by Hootsuite's web site)

- There are **5.11 billion** unique mobile users in the world today, up 100 million (2 percent) in the past year.
- There are **4.39 billion** internet users in 2019, an increase of 366 million (9 percent) versus January 2018.
- There are **3.48 billion** social media users in 2019, with the worldwide total growing by 288 million (9 percent) since this time last year.
- **3.26 billion** people use social media on mobile devices in January 2019, with a growth of 297 million new users representing a year-on-year increase of more than 10 percent.

***By 2020, there will be around 40 trillion gigabytes of data (40 zettabytes).***

Social media and *digitisation* has completely revolutionised the world. The various popular social media content providers are :

1. **Facebook** – 2.23 billion media access unit (MAUs). Facebook is the **biggest social media** site around, with more than two billion people using it every month. There

are more than 65 million businesses using Facebook Pages and more than six million advertisers actively promoting their business on Facebook, with over 2.271 billion active monthly users, generates the most amount of social data – users like over 4 million posts every minute – 4,166,667 to be exact, which adds up to 250 million posts per hour.

2. **U-Tube**- 1.9 billion MAUs. on video-sharing platform
3. **WhatsApp** – 1.5 billion MAUs. a messaging app used by people in over 180 countries. Initially, WhatsApp was only used by people to communicate with their family and friends. Gradually, people started communicating with businesses
4. **Messenger** – 1.3 billion MAUs, a messaging feature within Facebook
5. **WeChat** – 1.06 billion MAUs, a messaging app, just like WhatsApp and Messenger, into an all-in-one platform. Besides messaging and calling, users can now use WeChat to shop online and make payment offline, transfer money, make reservations, book taxis, and more.WeChat is most popular in China and some parts of Asia.
6. **Instagram** – 1 billion MAUs, Instagram is a photo and video sharing social media app. It allows you to share a wide range of content such as photos, videos, Stories, and live videos. It has also recently launched IGTV for longer-form videos.
7. **QQ** – 861 million MAUs, an instant messaging platform that is extremely popular among young Chinese. (It is used in 80 countries and also available in many other languages.) Besides its instant messaging features, it also enables users to decorate their avatars, watch movies, play online games, shop online, blog, and make payment
8. **Tumblr** – 642 million MUVs, Tumblr is a microblogging and social networking site for sharing text, photos, links, videos, audios, and more. People share a wide range of things on Tumblr from cat photos to art to fashion.
9. **Qzone –** 632 million MAUs. (app based in China, where users can upload multimedia, write blogs, play games, and decorate their own virtual spaces.
10. **Tik Tok** – 500 million MAUs (also known as Douyin in China) is a rising music video social network.
11. **Sina Weibo** – 392 million MAUs often known as Twitter for Chinese users (since Twitter is banned in China)
12. **Twitter** – 335 million MAUs
13. **Reddit** – 330 million MAUs also known as the front page of the Internet, is a platform where users can submit questions, links, and images, discuss them, and vote them up or down.
14. **Baidu Tieba** – 300 million MAUs, a Chinese online forum created by Baidu, the largest Chinese search engine in the world
15. **Linkedin** – 294 million MAUs,now more than just a resume and job search site. It has evolved into a professional social media site where industry experts share content, network with one another, and build their personal brand.
16. **Viber –** 260 million MAUs, social messaging apps such as WhatsApp and Messenger. It allows users to send messages and multimedia, call, share stickers and GIFs.
17. **Snapchat** – 255 million MAUs,focuses on sharing photos and short videos (as known as snaps) between friends. It made the Stories format popular, which eventually proliferated on other social media platforms like Instagram.

GlobalWebIndex reports that the average social media user now spends 2 hours and 16 minutes each day on social platforms. Most important fundamental concept to grasp, in understanding **the information explosion is** impact of **social media on** marketing strategies of big corporate houses and also the fact that people are worldover willingly participating and providing the information on internet without the differences of caste. creed, religion and nations. All this had been made possible **by big data sciences.**

## 1.2 BIG DATA

**Big Data** is a phrase used to mean a massive volume of both structured and unstructured **data** that is so **large,** it is difficult to process it using traditional database methods and software techniques. In most enterprise scenarios the volume of **data** is too **big** or it moves too fast or it exceeds current processing capacity.

The term has been in use since the 1990s, with some giving credit to John Mashey for coining or at least making it popular. Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time.

Although the term "big data" have been gaining momentum in current decade, the act of gathering and storing large amounts of information for eventual analysis is ages old. Humanity tried to gather the trusted data in form of Wikipedia. Currently, the English Wikipedia alone has over 5,904,484 articles of any length, and the combined Wikipedias for all other languages greatly exceed the English Wikipedia in size, giving more than 27 billion words in 40 million articles in 293 languages. The English Wikipedia alone has over 3.6 billion words, over 60 times as many as the next largest English-language encyclopedia,

The World Data Centre for Climate (WDCC) is the largest database in the world. The WDCC claims having data worth 220 terabytes readily accessible on the web including information on climate research and anticipated climatic trends, as well as 110 terabytes (or 24,500 DVD's) worth of climate simulation data.

If you do a Google search on big data (13.08.18), **over** 6,29,00,00,000 results search results are returned in 0.61 seconds in my PC, showing the importance of involvement of this technology. Big Data refers to technologies and initiatives that involve data that is too diverse, fast-changing or massive for conventional technologies, skills and infrastructure to address efficiently.

Over the past decade, major web companies like Google, Amazon and Facebook pioneered businesses built on monetizing massive data volumes. In the process, they invented new paradigms not only for extracting value from data, but also for managing data and compute resources from data center design, to hardware, to software, to application provisioning. Governments and even Google can detect and track the emergence of disease outbreaks via social media signals. Oil and gas companies can take the output of sensors in their drilling equipment to make more efficient and safer drilling decisions. In the same way that the mission to the moon spawned a wave of innovation across multiple industries, Big Data has pushed information technology a quantum leap forward.

### 1.2.1 How Big is the Canvas of Big Data?

Big data does not have always to be big (*i.e.*, data in range of peta/exabytes). Even 50 GB can be said as big data if the structure is too complex for a normal RDBMS to store. What is small data? Small data is simple data structures, e.g. numbers (be it monetary, integers, fractions or floating points), strings (names, description, types), dates, times, and all the data we used to know in the last 30 years of data warehousing history.

Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time. Big data "size" is a constantly moving target, as of 2012 ranging from a few dozen terabytes to many petabytes of data. Big data requires a set of techniques and technologies with new forms of integration to reveal insights from datasets that are diverse, complex, and of a massive scale.

The Information size is calculated as below (to be multiplied by 1.024 times, as actually 1 KB = 1024 = $2^8$) :

| | | |
|---|---|---|
| 1000 ($10^3$) KB | 1 kilobyte | |
| 1000 ($10^6$) MB | 1 megabyte | (1 million byte) |
| 1000 ($10^9$) GB | 1 gigabyte | (1 billion byte) |
| 1000 ($10^{12}$) TB | 1 terabyte | (1 trillion byte) |
| 1000 ($10^{15}$) PB | 1 petabyte | |
| 1000 ($10^{18}$) EB | 1 exabyte | |
| 1000 ($10^{21}$) ZB | 1 zettabyte | |
| 1000 ($10^{24}$) YB | 1 yottabyte | |

As per IBM 2.5 petabytes is Memory capacity of the human brain. A processor with a 64-bit address bus can address 18 exabytes of memory. Vedic Mathematics thought of $10^{60}$ as ultimate that in Sanskrit known as Mahogh. As per IDC, Digital universe is doubling in size every two years, and by 2020 the digital universe – the data we create and copy annually – will reach 44 zettabytes, or 44 trillion gigabytes.

Data sets are growing rapidly in part because they are increasingly gathered by cheap and numerous information-sensing mobile devices, aerial (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers and wireless sensor networks. Internet of Things (IoT) is making evry such things possible. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s; as of 2012, every day 2.5 exabytes ($2.5\times10^{18}$) of data are created. One question for large enterprises is determining who should own big data initiatives that affect the entire organization. IDC and EMC project that data will grow to 40 zettabytes by 2020, resulting in a 50-fold growth from the beginning of 2010.

### 1.2.2 Examples of Big Data

Artificial Intelligence (AI), mobile, social and Internet of Things (IoT) are driving data complexity, new forms and sources of data. Big data analytics is the use of advanced analytic techniques against very large, diverse data sets that include structured, semi-structured and unstructured data, from different sources, and in different sizes from terabytes to zettabytes.

Big data comes from sensors, devices, video/audio, networks, log files, transactional applications, web, and social media - much of it generated in real time and in a very large scale.

What we can do, however, is gain a sense of just how much data the average organization has to store and analyze today. Toward that end, here are some metrics that help put hard numbers on the scale of Big Data today:

- Analysts predict that by 2020, there will be 5,200 gigabytes of data on every person in the world.
- Amazon sells 600 items per second.
- On average, each person who uses email receives 88 emails per day and send 34. That adds up to more than 200 billion emails each day.

- Master Card processes 74 billion transactions per year.
- 2–40 exabytes of storage capacity will be needed by 2025 just for the human genomes
- *eBay.com* uses two data warehouses at 7.5 *peta bytes* and 40PB as well as a 40PB Hadoop cluster for search, consumer recommendations, and merchandising.
- *Amazon.com* handles millions of back-end operations every day, as well as queries from more than half a million third-party sellers. The core technology that keeps Amazon running is Linux-based and as of 2005 they had the world's three largest Linux databases, with capacities of 7.8 TB, 18.5 TB, and 24.7 TB.
- *Facebook* handles 50 billion photos from its user base.
- *Google* was handling roughly 100 billion searches per month as of August 2012.
- *Oracle NoSQL Database* has been tested to past the 1M ops/sec mark with 8 shards and proceeded to hit 1.2M ops/sec with 10 shards.

## 1.3 THE 5 V's OF BIG DATA

Big data is often characterised by the 5 **V's of Big Data** . Gartner was the first to put forward the concept of the term 3 V of big data which now stands modified as the five V's: Volume, *Velocity, Variety, Value, and Veracity*. The 5 V's, signifies the unique features of big data:

1. **Volume:** This signifies large amounts of data. Typically when people discuss big data volumes, they discuss peta-bytes. But in reality, most real-life big data implementations are still in the 10's to 100's of terabytes range which is still a lot of data.
2. **Velocity:** Velocity in the context of big data refers to the speed of data acquisition and processing. Big data technologies provide horsepower that accelerates these processes, thereby making data provisioning and usage faster, too.
3. **Variety:** This refers to the evolving types and growing sources of data, including semi-structured and unstructured data. An example of semi-structured data might be an e-mail message, where the date might be in a common, structured format, but the e-mail text itself is more unstructured. An example of unstructured data would be notes that a customer support representative might type in, free-form, about a customer's trouble ticket. Variety: big data draws from text, images, audio, video; plus it completes missing pieces through data fusion.

   Today's data is unstructured. In fact, 80% of all the world's data fits into this category, including photos, video sequences, social media updates, etc. New and innovative big data technology is now allowing structured and unstructured data to be harvested, stored, and used simultaneously.
4. **Value :** Talking about value means we are referring to the worth of the data being extracted. We can extract endless amounts of data but unless it can be turned into value it is useless. While there is a clear link between data and insights, this does not always mean there is value in Big Data. The most important part of embarking on a big data initiative is to understand the costs and benefits of collecting and analyzing the data to ensure that ultimately the data that is reaped can be monetized.
5. **Veracity :** Veracity is the quality or trustworthiness and accuracy of the data. For example, think about the data available on Wikipedia that is an amulgamation of data provided by millions of users worldwide. And think of the all the Twitter posts depending upon one's ego and alliance, having a lot of conflicts, hash tags, abbreviations, typos, etc., and the reliability and accuracy of all that content. Although Wikipedia makes

a lot of efforts to verify the data, it is often biased to western world. Gleaning loads and loads of data is of no use if the quality or trustworthiness is not accurate. Another good example of this relates to the use of Global Positioning System (GPS) data that depends upon the accuracy of measuring equipment. Often the GPS will "drift" off course as you peruse through an urban area. Satellite signals are lost as they bounce off tall buildings or other structures. When this happens, location data has to be fused with another data source like road data, or data from an accelerometer to provide accurate data.

## 1.4 BIG DATA HANDLING PROCESS: DATA MINING, DATA WARE HOUSING, DATA LAKES AND DATA MARTING

Big data is extracted or collected from various soft sources *i.e.*, mined and stored in data lakes and then sent to data warehouse and ultimately marketed to business market by and to various business houses and Government department with slight variations here and there.

### 1.4.1 Data Mining

Data mining is defined as a process of discovering hidden valuable knowledge by analyzing large amounts of data, which is stored in databases or data warehouse, using various data mining techniques such as machine learning, artificial intelligence(AI) and statistical.

Many organizations in various industries are taking advantages of data mining including manufacturing, marketing, chemical, aerospace… etc, to increase their business efficiency.

More extensive data mining techniques were needed to get resorted and resolved, partially because the size of the information is much larger and because the information tends to be more varied and extensive in its very nature and content. With large data sets, it is no longer enough to get relatively simple and straightforward statistics out of the system. With 30 or 40 million records of detailed customer information, knowing that two million of them live in one location is not enough. You want to know whether those two million are a particular age group and their average earnings so that you can target your customer needs better.

As per IBM these business-driven needs changed simple data retrieval and statistics into more complex data mining. The business problem drives an examination of the data that helps to build a model to describe the information that ultimately leads to the creation of the resulting report.

For example, IBM SPSS®,(we shall read this in chapter 13 on Machine Learning), which has its roots in statistical and survey analysis, can build effective predictive models by looking at past trends and building accurate forecasts. IBM InfoSphere® Warehouse provides data sourcing, pre-processing, mining, and analysis information in a single package, which allows you to take information from the source database straight to the final report output.

### 1.4.2 Data Lakes and Data Warehouses

Data lakes and Data warehouses are two different types of data storage *repository*, but with many differences. The data lake stores *raw structured* and *unstructured data* in whatever form the data source provides. It does not require prior knowledge of the analyses you think you want to perform.

The data warehouse integrates data from different sources and suits business reporting.

Data warehouse stores data in files or folders, in hierarchical manner where as data lake uses a flat architecture to store data.

Data warehouse is a core component of business intelligence, the data warehouse is a central repository of integrated data from one or more disparate sources, and it's used for reporting and data analysis. When the board makes a strategic decision on its future, or a call center agent reviews a customer's profile the data is typically being sourced from a data warehouse.

*Characteristics Features of Data warehouse concept*:

1. Holds multiple subject areas
2. Holds very detailed information
3. Works to integrate all data sources
4. Does not necessarily use a dimensional model but feeds dimensional models.

Data mart is a special category of data warehouse that often holds only one subject area-for example, *Finance*, or *Sales*. It may hold more summarized data and concentrates on integrating information from a given subject area or set of source systems.

The following are the reasons for creating a data mart

1. Easy access to frequently needed data
2. Creates collective view by a group of users
3. Improves end-user response time
4. Ease of creation
5. Lower cost than implementing a full data warehouse
6. Potential users are more clearly defined than in a full data warehouse
7. Contains only business essential data and is less cluttered.

## 1.5 DIFFERENCE BETWEEN DATA WAREHOUSING AND BIG DATA TECHNOLOGY

Big data uses Hadoop ecosystem to extract useful data from dynamic social and technical data present on Internet while data warehouse is a static repository of raw data first naturally collected in data lake, then sorted systematically and modelled into useful database.

The diagram below illustrates how the data warehouse and big data environments can come together in an integrated and very complementary way. In this scenario, the Hadoop

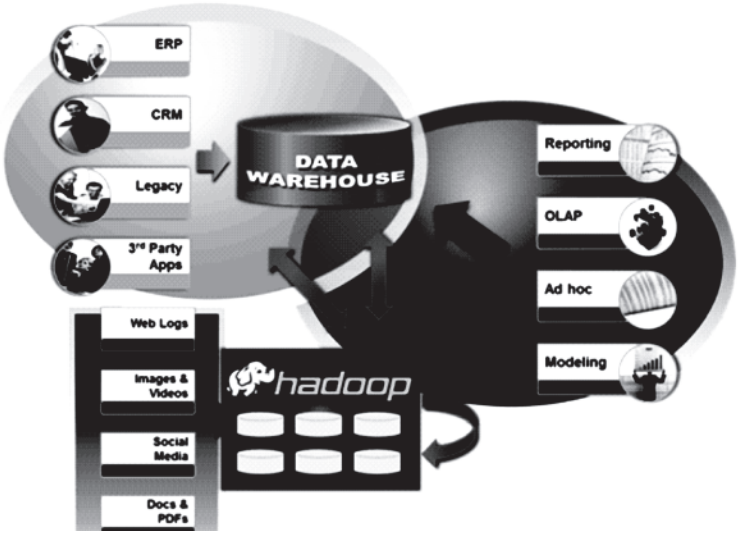

**Fig. 1.2** Difference Between Big data and Data Warehousing

system can perform quickly. For instance, a high-tech company might decide to pull data from its social networking site and combine it with data from the data warehouse to update a customer's social network circle of friends. The environment might also use Hadoop to quickly "score" that person's social influence. Then that data will be provisioned back to the data warehouse so that, say, a campaign manager can view that person's influence score and re-segment him (or her) accordingly.

The example here is one of many potential uses for Hadoop and the data warehouse working together. The point to make here is that each system is doing what it's best designed to do. In the case of Hadoop, it's processing large amounts of social networking data quickly and in parallel. In the case of the data warehouse, it's availing that data to business users, knowledge workers, or data scientists, who are using that data—and other data as well—to make business decisions.

While there are exceptions to every rule, big data and data warehouse technologies are optimized for different purposes. Again, the goal is to use these solutions for what they were designed to do. In other words, use the best tool for the job.

Analyzing big data allows analysts, researchers, and business users to make better and faster decisions using data that was previously inaccessible or unusable. Using advanced analytics techniques such as text analytics, machine learning, predictive analytics, data mining, statistics, and natural language processing, businesses can analyze previously untapped data sources independent or together with their existing enterprise data to gain new insights resulting in better and faster decisions.

| BUSINESS REQUIREMENT | BIG DATA | DATA WAREHOUSE |
|---|---|---|
| Discovery of unexplored business questions | ● | ● |
| Clean, consistent, high-quality data | ◒ | ● |
| Low latency, interactive reports, OLAP | ◒ | ● |
| Raw, unstructed data | ● | |
| Analysis of preliminary data | ● | |

**Fig. 1.3** Comparing Business Requirement for Big Data and Datawarehousing

## 1.6 BIG DATA TYPES

Big data is data sets that are so voluminous and complex that traditional data processing application software are inadequate to deal with them. Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating and information privacy. Lately, the term "big data" tends to refer to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set. "There is little doubt that the quantities of data now available are indeed large, but that's not the most relevant characteristic of this new data ecosystem."

Data sets grow rapidly, in part because they are increasingly gathered by cheap and numerous information-sensing Internet of things devices such as mobile devices, aerial (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers and wireless sensor networks.

Big Data philosophy encompasses unstructured, semi-structured and structured data, however the main focus is on unstructured data.[1] Big data 'size' is a constantly moving target, as of 2012 ranging from a few dozen terabytes to many petabytes of data.[2] Big data requires a set of techniques and technologies with new forms of integration to reveal *insights* from datasets that are diverse, complex, and of a massive scale.[3].

Following are examples of the three types:

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

### 1.6.1 Sources of Unstructured Big Data

Unstructured data is everywhere. In fact, most individuals and organizations conduct their lives around unstructured data. Just as with structured data, unstructured data is either machine generated or human generated.

Here are some examples of machine-generated unstructured data:

1. **Satellite images:** This includes weather data or the data that the government captures in its satellite surveillance imagery. Just think about Google Earth, and you get the picture.
2. **Scientific data:** This includes seismic imagery, atmospheric data, and high energy physics.
3. **Photographs and video:** This includes security, surveillance, and traffic video.
4. **Radar or sonar data:** This includes vehicular, meteorological, and oceanographic seismic profiles.

The following list shows a few examples of human-generated unstructured data:

1. **Text internal to your company:** Think of all the text within documents, logs, survey results, and e-mails. Enterprise information actually represents a large percent of the text information in the world today.
2. **Social media data:** This data is generated from the social media platforms such *as YouTube, Facebook, Twitter, LinkedIn, and Flickr.*
3. **Mobile data:** This includes data such as text messages and location information.
4. **Website content:** This comes from any site delivering unstructured content, like ***YouTube, Flickr, or Instagram, Wikipedia.***

Of the useful data, IDC estimates that in 2013 perhaps 5% was especially valuable, or "target rich". That percentage should more than double by 2020 as enterprises take advantage of new Big Data and analytics technologies and new data sources, and apply them to new parts of the organization.

### 1.6.2 Semi-structured Data

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

**Personal data stored in a XML file**

```
<rec><name>Amitabh Bajaj</name><sex>Male</sex><age>49</age></rec>
<rec><name>JyotsnAgarwal</name><sex>Female</sex><age>47</age></rec>
<rec><name>Anurag Jain</name><sex>Male</sex><age>44</age></rec>
<rec><name>Subrato Roy</name><sex>Male</sex><age>26</age></rec>
<rec><name>Jeremiah J.</name><sex>Male</sex><age>35</age></rec>
```

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

### 1.6.3 Meta Data

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

There are three main types of metadata:

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

### 1.6.4 Structured Data

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

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

| Employee_ID | Employer_Name | Gender | Department | Salary_In_lacs |
|---|---|---|---|---|
| 2365 | Rajesh Kulkarni | Male | Finance | 650000 |
| 3398 | Pratibha Joshi | Female | Admin | 650000 |
| 7465 | Shushi Roy | Male | Admin | 500000 |
| 7500 | Subhojit Das | Male | Finance | 500000 |
| 7699 | Priya Sane | Female | Finance | 550000 |

## 1.7 DATA SCIENCE

Data science, also known as *data-driven science*, is an interdisciplinary field of scientific methods, processes, and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining. We shall learn about data mining in short in this unit itself. In fact it is emerging as convergence of various knowledge domains for effective utilisations of various analysis methods for better output of experts in their activities (Fig. 1.4).

**Convergence of various knowledge domains**

| | |
|---|---|
| **Math and Theory** | • Statistics, Linear Algebra, Optimization, Time Series, etc. |
| **Applied Algorithms** | • Machine Learning, Data Structures, Parallel Algorithms, etc. |
| **Engineering and Technologies** | • Storage and computing platforms, statistical tools, etc. |
| **Domain Expertise** | • Test, Finance, Images, Econometrics etc. |
| **Art** | • Visualization, Infographics |
| **Best practices and hacks** | • Handle missed values in data, transform and represent data, etc. |

**Fig. 1.4** Data science as convergence of various knowledge domains

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

Data science is blossoming as "concept to unify statistics, data analysis and their related methods" in order to "understand and analyze actual phenomena" with big data.[3]. In its extended canvas (*i.e.*, while dealing with Big. data), data science employs techniques and theories

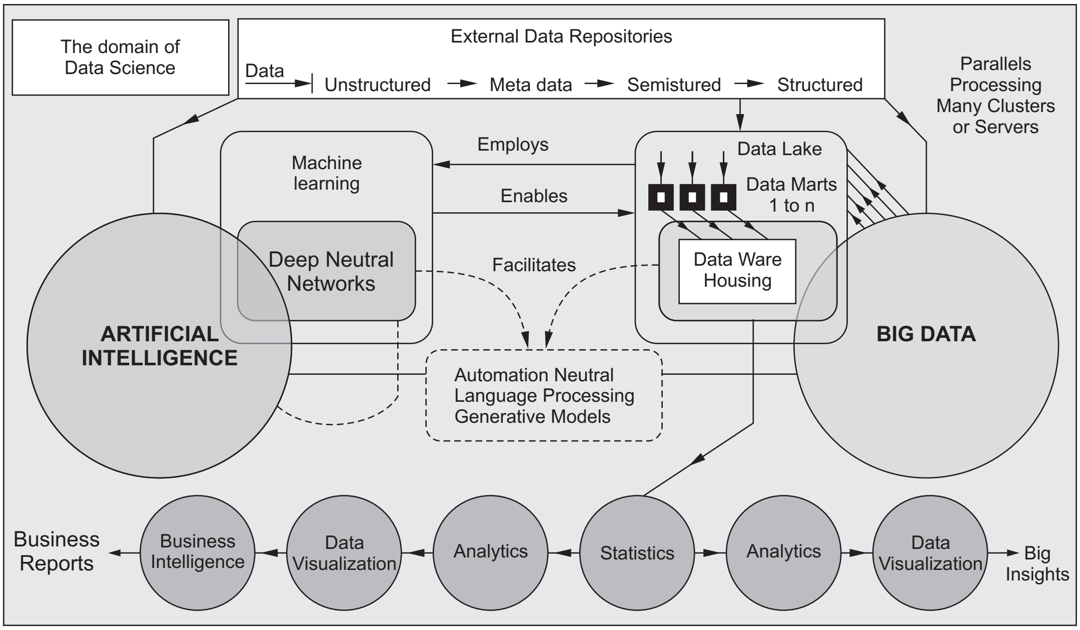

**Fig. 1.5** Broad Canvas of Data Science Dealing with Big Data

drawn from many fields within the broad areas of mathematics, statistics, information science, and computer science, in particular from the sub-domains of *machine learning, classification, cluster analysis, data lakes data mining and warehousing, databases,* and *visualization* (vide Fig. 1.5).We shall learn about dimensions of Big data in this unit very shortly.

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

## 1.8 BUSINESS INTELLIGENCE (BI)

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

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

Business intelligence can be used by enterprises to support a wide range of business decisions - ranging from operational to strategic. Basic operating decisions include product positioning or pricing. Strategic business decisions involve priorities, goals and directions at the broadest level.

Often BI applications use data gathered from a data warehouse (DW) or from a data mart.

## 1.9 DIFFERENCE BETWEEN BIG DATA & BUSINESS INTELLIGENCE

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

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

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

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

## 1.10 TERMS RELATED WITH DATA SCIENCE

The canvas of data science is so large that we need to understand many different terms for easy grasping of subject matter lateron**.**

1. **Data Wrangling:** The process of conversion of data, often through the use of scripting languages, to make it easier to work with is known as data *Wrangling* or *data munging*. If you have 900,000 birth year values of the format yyyy-mm-dd and 100,000 of the

format mm/dd/yyyy and you write a Perl script to convert the latter to look like the former so that you can use them all together, you're doing data wrangling. Discussions of data science often bemoan the high percentage of time that practitioners must spend doing data wrangling, the discussions then recommend the hiring of data engineers to address this Data engineers build massive reservoirs for big data. They develop, construct, test and maintain architectures such as databases and large-scale data processing systems. Once continuous pipelines are installed to – and from – these huge "pools" of filtered information, data scientists can pull relevant data sets for their analyses.

2. **Algorithm:** A series of repeatable steps for carrying out a certain type of task with data. As with data structures, people studying computer science learn about different algorithms and their suitability for various tasks. Specific data structures often play a role in how certain algorithms get implemented.
3. **Analytics:** Analytics is the discovery and communication of meaningful patterns in data. Especially valuable in areas rich with recorded information, analytics relies on the simultaneous application of statistics, computer programming and operations research to quantify performance. Analytics often favours data visualization to communicate insight.

   Firms may commonly apply analytics to business data, to describe, predict, and improve business performance. Specifically, areas within analytics include predictive analytics, enterprise decision management, retail analytics, store assortment and stock-keeping unit optimization, marketing optimization and marketing mix modelling, web analytics, sales force sizing and optimization, price and promotion modelling, predictive science, credit risk analysis, and fraud analytics. Since analytics can require extensive computation, the algorithms and software used for analytics harness the most current methods in computer science, statistics and mathematics.

   In a nutshell, analytics is the scientific process of transforming data into insight for making better decisions.
4. **Machine Learning:** Machine learning is a subfield of artificial intelligence (AI). The goal of machine learning generally is to understand the structure of data and fit that data into models that can be understood and utilized by people.

   Although machine learning is a field within computer science, it differs from traditional computational approaches. In traditional computing, algorithms are sets of explicitly programmed instructions used by computers to calculate or problem solve. Machine learning algorithms instead allow for computers to train on data inputs and use statistical analysis in order to output values that fall within a specific range. Because of this, machine learning facilitates computers in building models from sample data in order to automate decision-making processes based on data inputs.

   Machine learning refers to a broad class of methods that revolve around data modelling to :

   1. *Supervised learning* to algorithmically make predictions, and
   2. *Unsupervised learning* to algorithmically decipher patterns in data.

   We shall learn in chapter 7. that *Supervised Learning* includes predictive modelling techniques such as neural nets, classification and regression trees (decision trees), naive

Bayes, k-nearest neighbour, and support vector machines are generally included. One characteristic of these techniques is that the form of the resulting model is flexible, and adapts to the data. Statistical modelling methods that have highly structured model forms, such as linear regression, logistic regression and discriminant analysis are generally not considered part of machine learning. Unsupervised learning methods such as association rules and clustering are also considered part of machine learning.

5. **Web Analytics:** Statistical or machine learning methods applied to web data such as page views, hits, clicks, and conversions (sales), generally with a view to learning what web presentations are most effective in achieving the organizational goal (usually sales). This goal might be to sell products and services on a site, to serve and sell advertising space, to purchase advertising on other sites, or to collect contact information. Key challenges in web analytics are the volume and constant flow of data, and the navigational complexity and sometimes lengthy gaps that precede users' relevant web decisions.
6. **Business insight**: Business insight is a thought, fact, combination of facts, data and/or analysis of data that induces meaning and furthers understanding of a situation or issue that has the potential of benefiting the business or re-directing the thinking about that situation or issue which then in turn has the potential of benefiting the business.

## 1.11 GOALS OF DATA ANALYTICS

The goal of Data Analytics (big and small) is to get actionable *insights* resulting in smarter decisions and better business outcomes. How you architect business technologies and design data analytics processes to get valuable, actionable insights varies.

It is critical to design and build a data warehouse / business intelligence (BI) architecture that provides a flexible, multi-faceted analytical ecosystem, optimized for efficient ingestion and analysis of large and diverse datasets.

There are three types of data analysis:

1. Predictive (forecasting)
2. Descriptive (business intelligence and data mining)
3. Prescriptive (optimization and simulation)

The growing maturity of the concept more starkly delineates the difference between big data and Business Intelligence:

- Business Intelligence uses descriptive statistics with data with high information density to measure things, detect trends, etc.
- Big data uses inductive statistics and concepts from nonlinear system identification to infer laws (regressions, nonlinear relationships, and causal effects) from large sets of data with low information density to reveal relationships and dependencies, or to perform predictions of outcomes and behaviors.

## 1.12 PERSONNEL INVOLVED WITH DATA SCIENCE

1. **Data Scientist:** A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician. "Data science" is often used to define a (new) profession whose practitioners are capable in many or all the above areas; one often sees the term "*data scientist*" in job postings. While

"statistician" typically implies familiarity with research methods and the collection of data for studies, "*data scientist*" implies the ability to work with large volumes of data generated not by studies, but by ongoing organizational processes. Due to the complexity of dealing with large datasets and data flows, most of the day-to-day work of a data scientist lies in data pipeline challenges - storing relevant data, getting it into appropriate form for analysis, and managing the real-time implementation of models**.**

2. **Data Analyst:** Data analysts collect, process and perform statistical analyses of data. Their skills may not be as advanced as data scientists (e.g. they may not be able to create new algorithms), but their goals are the same – to discover how data can be used to answer questions and solve problems.
3. **Data Engineer:** A specialist in data wrangling. "Data engineers are the ones that take the messy data... and build the infrastructure for real, tangible analysis. They run ETL software, marry data sets, enrich and clean all that data that companies have been storing for years.

## 1.13 DATA SCIENCE vs. DATA ANALYSIS

It's very important to know that data science and Data analysis are little similar but, there so many differences between them. Let's check out the differences.

| | **Data Science** | **Data Analysis** |
|---|---|---|
| (*i*) | Providing strategic actionable insights into the world. | Providing operational observations into issues |
| (*ii*) | Mathematical, technical and strategic knowledge are mandatory. | Data analysis and visualization skills required |
| (*iii*) | Deal with big data | Not necessarily deal with big data |

## 1.14 THE DATA SCIENCE PROCESS (DSP)

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

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

The process may involve 7 clear cut steps for data analysis as shown in Fig. 1.6 :

**Step 1:** Frame or define the (business) problem

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

**Step 3:** Data preparation for process the data for analysis

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

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

**Step 6:** Evaluation

**Step 7 :** Visualisation and Communication of Results of the Analysis

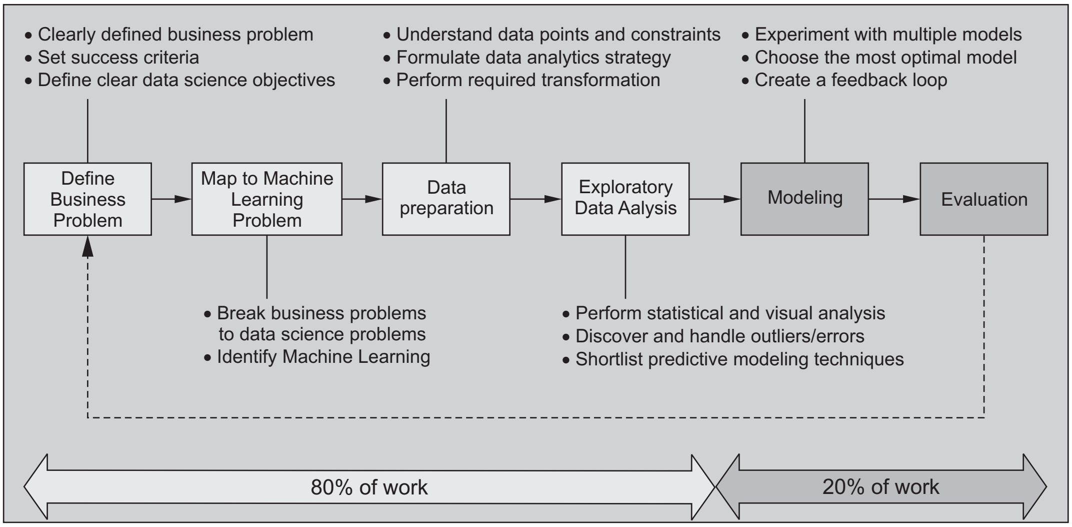
• Clearly defined business problem
• Set success criteria
• Define clear data science objectives
• Understand data points and constraints
• Formulate data analytics strategy
• Perform required transformation
• Experiment with multiple models
• Choose the most optimal model
• Create a feedback loop
Define Business Problem
Map to Machine Learning Problem
Data preparation
Exploratory Data Aalysis
Modeling
Evaluation
• Break business problems to data science problems
• Identify Machine Learning
• Perform statistical and visual analysis
• Discover and handle outliers/errors
• Shortlist predictive modeling techniques
80% of work
20% of work

**Fig. 1.6** Seven steps of the Data Science Process (DSP)

### Step 1: Frame or define the (business) problem

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

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

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

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

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

You should ask questions like the following:

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

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

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

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

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

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

**Step 3: Data Preparation for process the data for analysis**

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

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

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

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

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

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

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

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

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

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

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

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

### Step 6: Evaluation

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

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

### Step 7 : Visualisation and Communication of Results of the Analysis

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

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

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

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

## 1.16 DATA SCIENCE PROJECT'S LIFECYCLE

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

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

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

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

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

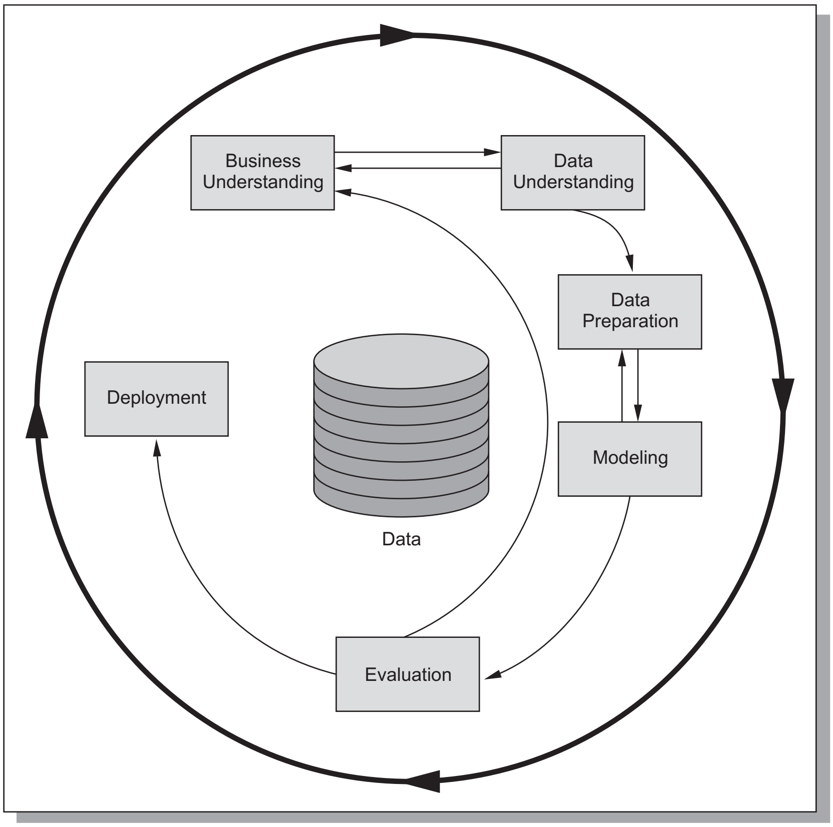

**Fig. 1.7** The 6 high-level phases of CRISP-DM suggested for the Data Science Projects

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

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

Fig. 1.8 provides a visual representation of the **Team Data Science Process lifecycle**.

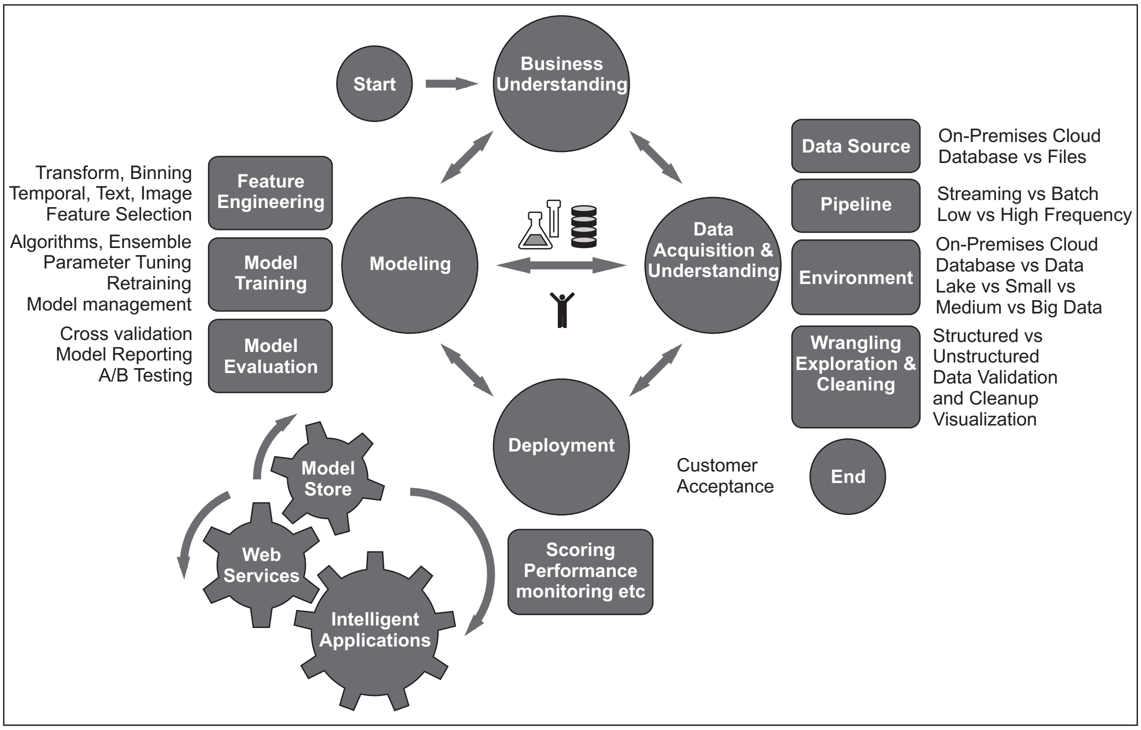
Start
Business Understanding
Data Source
On-Premises Cloud Database vs Files
Pipeline
Streaming vs Batch Low vs High Frequency
Environment
On-Premises Cloud Database vs Data Lake vs Small vs Medium vs Big Data
Wrangling Exploration & Cleaning
Structured vs Unstructured Data Validation and Cleanup Visualization
Data Acquisition & Understanding
Transform, Binning Temporal, Text, Image Feature Selection
Feature Engineering
Algorithms, Ensemble Parameter Tuning Retraining Model management
Model Training
Cross validation Model Reporting A/B Testing
Model Evaluation
Modeling
Deployment
Customer Acceptance
End
Scoring Performance monitoring etc
Model Store
Web Services
Intelligent Applications

**Fig. 1.8** Life Cycle of Data Science Process (Courtesy Microsoft Azure)

## 1.17 POPULAR DATA SCIENCE TOOLKITS

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

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

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

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

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

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

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

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

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

**We have used this language extensively in Chapter 6.**

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

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

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

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

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

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

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

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

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

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

6. **D3 Data Science Tools:** D3 is a javascript library for building interactive data visualizations within your browser. It allows data scientists to create rich visualizations with a high level of customizability. It's a great addition to your data science toolkit if you're looking to dynamically express your data insights.
7. **TensorFlow:** TensorFlow is the product of Google's Brain Team coming together for the purpose of advancing machine learning. It's a software library for numerical computation and built for everyone from students and researchers to hackers and innovators. It allows programmers to access the power of deep learning without needing to understand some of the complicated principles behind it, and ranks as one of the data science tools that helps make deep learning accessible for thousands of companies. TensorFlow is an open-source software library for dataflow programming across a range of tasks. It is a symbolic math library, and also used for machine learning applications such as neural networks. TensorFlow was developed by the Google Brain team for internal Google use. It was released under the Apache 2.0 open source license on November 9, 2015.
8. **Rstudio:** RStudio integrates with R as an IDE (Integrated Developmet Environment) to provide further functionality. RStudio combines a source code editor, build automation tools and a debugger.

## References

1. *Dedić, N.; Stanier, C. (2017). "Towards Differentiating Business Intelligence, Big Data, Data Analytics and Knowledge Discovery".* ***285****. Berlin; Heidelberg: Springer International Publishing. ISSN 1865-1356. OCLC 909580101.*
2. *Everts, Sarah (2016). "Information Overload". Distillations.* ***2*** *(2): 26–33. Retrieved 22 March 2018.*
3. *Ibrahim; Targio Hashem, Abaker; Yaqoob, Ibrar; Badrul Anuar, Nor; Mokhtar, Salimah; Gani, Abdullah; Ullah Khan, Samee (2015). "big data" on cloud computing: Review and open research issues". Information Systems.* ***47****: 98–115. doi:10.1016/j.is.2014.07.006.*
4. *Stewart Tansley; Kristin Michele Tolle (2009). The Fourth Paradigm: Data-intensive Scientific Discovery. Microsoft Research. ISBN 978-0-9825442-0-4.*
5. *Bell, G.; Hey, T.; Szalay, A. (2009). "COMPUTER SCIENCE: Beyond the Data Deluge". Science. 323 (5919): 1297–1298. doi:10.1126/science.1170411. ISSN 0036-8075*