The Data Deluge

# Introduction

In the present context, all the activities that we see around us have been mostly operating with the use of technology. The use of technological equipment has increased by a large number with every person having at least one device with them. It is told that as of April 2021, about 4.72 billion people have access to the internet. With this much of people being engaged with the internet and technology, we can predict the number of data that we have been generating. Likewise, we can predict the number of information and patterns that are hidden in that volume of data. All of this available data can be collected, preprocessed, evaluated and analysed further in order to boost the business productivity. Huge organizations have invested large sums of money for this sole purpose. They hire highly skilled manpower and buy the latest technological devices so that they have an upper hand in this data-driven world.

# 1. Rapid Growth in Data Acquisition

## Introduction

The huge increase in data is due to the increase in the number of people getting access to the internet and other technological devices. The latest mobile phones, watches, smart vehicles, CCTV cameras and other devices are built in such a way that it tracks all of our daily basis activities. For technological devices, such information and activities are taken as data. All these data are digitized so that it can be analysed and stored. And this particular process can be called data acquisition which has been correspondingly increasing with the growth of data.

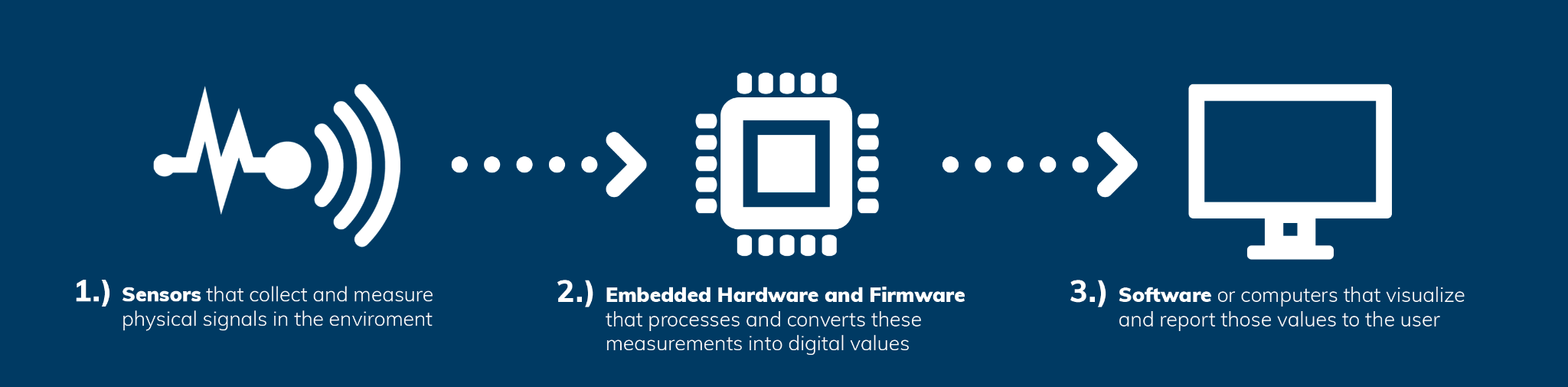


Fig: - Data Acquisition process

## Reason for its rise

We all might be well aware of IoT (Internet of Things). The IoT can be referred to as a system which is a network of physical objects that are embedded with software and sensors. And with use of those software and sensors the data are exchanged with other devices over the internet.

IoT has surely made a positive change in the life of millions of people. From entertainment to education and security, it has been used in every possible sector.

The different applications that we use in our smart phone stores information like the places we have visited, the items that we are searching for in the market, the types of movies we like, the taste of music we prefer the most and so on. Whereas from smart watches it can track our health related information. The other type of devices that are in huge rise are virtual assistants like Amazon Echo, Alexa and Google Home. These devices have made the life of people very easy going.

## Importance of Data Acquisition

If we are only generating data and not doing anything with them, then it would be a huge loss for everybody. Some of the importance of data acquisition are: -

### Quick analysis of problems

Real-time data are those information which are sent instantly after collection. When we receive data in real time, then we can instantly visualize them, process them and perform an instant quick analysis. So, even if we face a problem unexpectedly, then it can be quickly analysed and solved.

### Help in decision making

When making a decision in an organization, many factors need to be taken in consideration. There can be many events and actions that can affect the performance of a business. So, we have more data, then we can extract more information from them. Such information can be vital while making important business decisions.

### Invention of new idea

The world is always in need of new innovative inventions that can help to bring a positive change in the lives of people. Collection of different types of data, processing and analysing them can help the people to get better knowledge and even expand their vision. This can help the people to create a new idea that can bring positive change in the society.

# 2. Extracting knowledge from data

## Introduction

People and machines are producing data at a very higher rate than ever before. The volume and variety of data which are being produced bring more challenges in identifying useful information from them. With the huge volume of data available on the internet, it becomes very important that we utilize them for our better goods. Extracting information from the data is always important as it can lead us to new knowledge that could benefit us in many ways. Some of the areas in which knowledge extraction plays an important role are: -

* Business Intelligence
* Financial investigation
* Health care research
* Scientific research

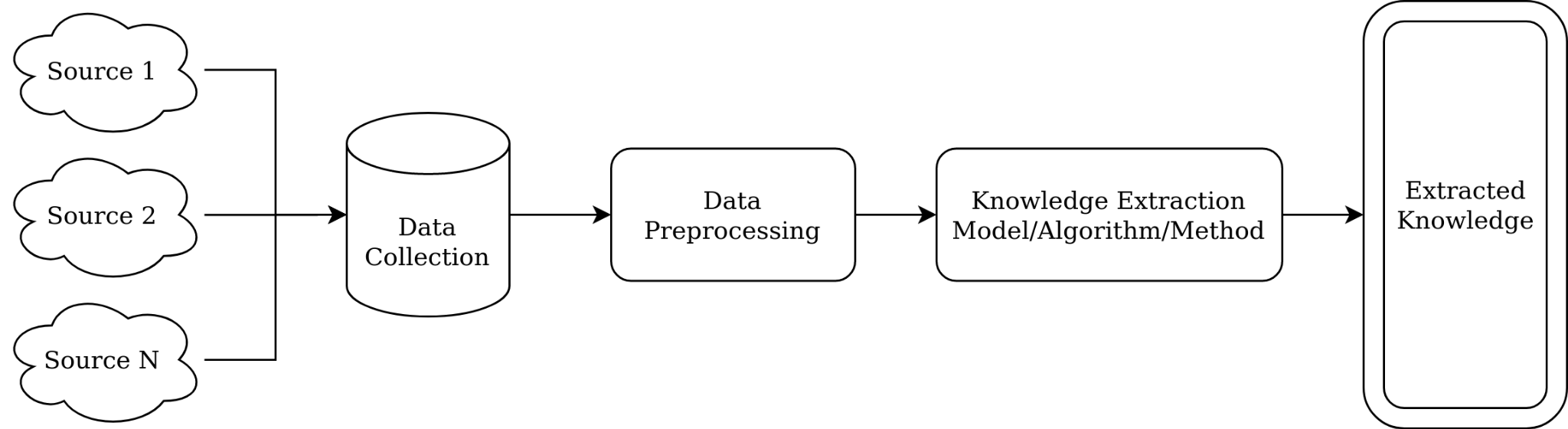


Fig: - An overview of Knowledge Extraction

Some of the commonly used technique for extraction knowledge are: -

### Data Mining

Data Mining is a technique used for discovering hidden information or patterns in our data. The information that is received using this technique can be further used to build machine learning models. In Data mining, we explore and analyze huge blocks of information and try to extract the hidden patterns or trends which can be used for different tasks like fraud detection, database marketing or filtering spam mails.

### Information Retrieval

Information Extraction can be defined as a technique used to find important information that exists within a document. A document can be either structured, semi-structured or even unstructured. Like in Data mining, information retrieval does not help us in identifying hidden patterns in a data. It only helps us to retrieve information from them.

## Challenges and Limitations

Though we know the importance of knowledge extraction, it is not a task which can be easily performed. It has its own challenges and limitations which has to be monitored when working. The data being in heterogeneous and stored in different formats (structured, semi-structured and unstructured), it is quite a challenging task to extract valuable information from them. Some of them have been discussed below.

### Diverse Data

Real world data are heterogeneous ones. The data which we collect, comes from multiple sources. We can collect data from textual documents, videos, images and even sounds. Because of this, they can be of different formats. The data can be time series data, multimedia data, complex data, spatial data, temporal data and so on. With the diversity in data, it raises complexity when dealing with them.

### Suitable Algorithm or Method

There is not an absolute method or algorithm that can be implemented in every data for extracting information. For different formats of data, different approaches are to be implemented. And also depending upon the type of data we want to extract, different algorithms have to be applied. So, it is a challenging task to create new algorithms every time for every new requirement.

### Noisy Data

Noisy data can be referred to as unwanted data. When we collect a huge volume of data from our different sources, it is obvious that it might contain some unwanted data. It is very important for us to identify and remove such data as it will only lead us to unwanted errors.

Handling noisy data is always a challenging task. It has to be performed before fitting the data in the algorithm. So, data preprocessing is a step which has to be performed in a serious manner. Some of the steps which can be used for removing noisy data are: -

* Binning

This approach is applied mostly on sorted data. The sorted data is divided into multiple segments having equal size and each segment is processed independently. The noisy data can be replaced by bin mean, bin median or bin boundary.

* Clustering

In clustering, we create a group of similar data. All the data that falls under a similar category are grouped under its own category. Since the outliers (noisy data) does not fall under any group, we can figure them out and exclude them for further process.

* Regression

By applying a regression function to our data, we can polish our data further. If we are working with only two features then we implement linear regression. And if we are working with multiple features (more than 2) then we use the multiple linear regression to find the best line that fits all of our data.

* PCA (Principal Component Analysis)

PCA is a mathematical approach through which we can reduce our large dimension data into a smaller dimension data without losing its quality.

# 3. Analysis of Data

Analysis of data is an important task as it provides us better knowledge about our data and helps us to make an effective data-driven decision. While making a data-driven decision, we use the different metrics, facts and figures that are stored and hidden in the data. Since the data consists of information related to business, market and customers, data-driven results surely helps to make better decisions. To make a better decision using the data, first we need to understand how to analyse the data that we have. Due to the diversity of data, analysing them is a very demanding task.

Analysis of data can be performed by using some tools like Pandas, Python, R, Java, SQL, Matlab, Power BI and Tableau. Some of the commonly used techniques for data analysis are: -

### Text Analysis

Text analysis is an approach for extracting out facts and information from a textual document. This technique can also be referred to as information extraction, text analytics or text mining. From this technique, we can receive a chunk of textual information in a structured or unstructured format.

### Statistical Analysis

Statistical analysis involves the process of collecting, analysing, interpreting, presenting and modeling the data. As we use past data, this technique helps us to figure out “what happened” previously. Several statistical operations like mean, median are applied, which gives us a descriptive analysis of the data.

### Diagnostic Analysis

Diagnostic analysis helps us to find out the reason of “why did it happened” by detecting the behaviour or pattern within the data. This technique helps us to give us a reason for a certain incident to be occured. So, in the near future, if some similar incidents are to be occurred then we would easily identify its cause and come up with a suitable solution as fast as possible.

### Predictive Analysis

Predictive analysis is about predicting the events that are most likely to happen in the near future. For such analysis, we use the past data and feed them in the Deep Learning (DL) or Machine Learning (ML) models or other algorithms. Then, we use our current data and predict the possible event that will occur. This approach can be helped in fraud detection, optimizing the marketing strategies, improving the operations and reducing risks.

### Prescriptive Analysis

In perspective analysis, we combine the information that we have gained from all the previous techniques and determine the best action to be taken for the current problem.

# 4. Ethics

Ethics can be explained as a standard of what is right and wrong for a human to do in a certain event or scenario. When we are in a certain environment with a certain authority or power, then we are bound by a set of obligations that we ought to perform. We cannot go beyond them and even do the things we like to do. Similarly, when working with data, it also has its own set of ethics which we have to follow.

## Ethics in Data

Data as we all know is information which can be of a person, event or anything. And when dealing with data, we not only deal with our personal data but also of other people too which could be sensitive to them. So, there are some certain things that we should consider before handling such data.

### User Consent when collecting data

When we work on collecting data, we do it to achieve a certain goal or objective. There is a certain purpose for us when we collect information about other people. So, whatever the goal is that we are trying to achieve, we have to make sure that we have the consent of the user i.e. the user is well aware of our reason to collect the data. If we are collecting information through a survey then it has to be clearly noted in the top or if we are scraping data from the internet then we have to ask permission from the owner of the website. There might be some information that people wish to share and some which they do not. Such things have to be made clear to them beforehand so that we would not have to face any trouble later on.

### Data Privacy

One of the major risks that comes when working with data is its security. The information that we have collected from the people can contain sensitive information about them which they do not want to release in the public. And there are many ways through which the privacy of the data can be compromised. The data can be leaked from within the organization or it can be stolen from outside with the use of malware softwares. So, we have to make sure that the people who are handling the data do not share it with the outside world. And also only the applications approved by the organizations are used for processing and analysing tasks.

### Data Transparency

Data transparency is also one of the major topics that has been headlining from the past few years. With the huge volume of data available everywhere on the internet, we do not know how much of it is true and false. It has become quite an easy job to influence people on the internet by spreading false rumors. So, it has become very essential to assure that the data that is being shared are accurate and coming from the official sources.

## Ethics in Machine Learning

Machine Learning (ML) is a subset of Artificial Intelligence (AI) that has been on a huge rise in the current time period. With the access of huge volumes of data, enhancement in the hardwares and invention of new algorithms, Machine Learning has become the talk of the town. In the Machine Learning (ML) field also, there are some ethics that have to be considered.

### Algorithm

Algorithms can be defined as a set of rules (problem-solving operation) that is followed for achieving a certain result. In Machine Learning, it is the algorithms that do the hard and complex job of finding the patterns and valuable information from the data. Making such algorithms public and easily accessible to the people is surely a positive move. The more we share our knowledge, the more new ideas and algorithms can be created. So, if we are able to create a new algorithm which works best for a certain scenario, then we should surely share it with everyone.

### Result

The Machine Learning models are created to help people overcome the human errors, automate the repetitive tasks and perform the tasks which are beyond the human capacity. So, it becomes very important that the evaluation of the result is done in a proper way before releasing it to the outside world. The factors like true positive, false positive, true negative and false negative should be evaluated thoroughly. Sometimes, in order to increase the accuracy of the model, its testing is done on the dataset upon which it has been trained. Such things have to be clearly avoided.

# Conclusion

Since the world has been gaining progress in the technological sector, the huge increase in volume of data was inevitable. More the people are getting access to the latest information and technologies, the more data is being generated and collected. And due to the rise in data, the level of difficulty to utilize them properly for the better benefit of the people has also risen. This is because of the diversity in the types of data. Collecting the data, cleaning and preprocessing them properly and analysing and getting valuable information from them has become a major goal for every business organization. And doing all these tasks by staying inside the ethical boundary is surely a challenge for everyone.

So, as long as we see the increase in volume of data, all the above mentioned things will keep on increasing in parallel. With the correct and ethical methods taken in to collect, store, preprocess, analyse and extract the data, we will surely be able to make the best of this technological rise.

# References

DataReportal – Global Digital Insights. 2021. *Digital Around the World — DataReportal – Global Digital Insights*. [online] Available at: <https://datareportal.com/global-digital-overview>

Intellisense Systems, Inc. 2021. *Data Acquisition | See How We Do It at Intellisense Systems Inc.*. [online] Available at: <https://www.intellisenseinc.com/innovation-lab/data-acquisition>

Oracle.com. 2021. *What is the Internet of Things (IoT)?*. [online] Available at: <https://www.oracle.com/internet-of-things/what-is-iot/>

Investopedia. 2021. *Data Mining: How Companies Use Data to Find Useful Patterns and Trends*. [online] Available at: <https://www.investopedia.com/terms/d/datamining.asp>

Ontotext. 2021. *What is Information Extraction? | Ontotext Fundamentals*. [online] Available at: <https://www.ontotext.com/knowledgehub/fundamentals/information-extraction/>

GeeksforGeeks. 2021. *Data Preprocessing in Data Mining - GeeksforGeeks*. [online] Available at: <https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/>

Guru99.com. 2021. *What is Data Analysis? Research | Types | Methods | Techniques*. [online] Available at: <https://www.guru99.com/what-is-data-analysis.html>

Medium. 2021. *Ethics in Machine Learning*. [online] Available at: <https://towardsdatascience.com/ethics-in-machine-learning-9fa5b1aadc12>