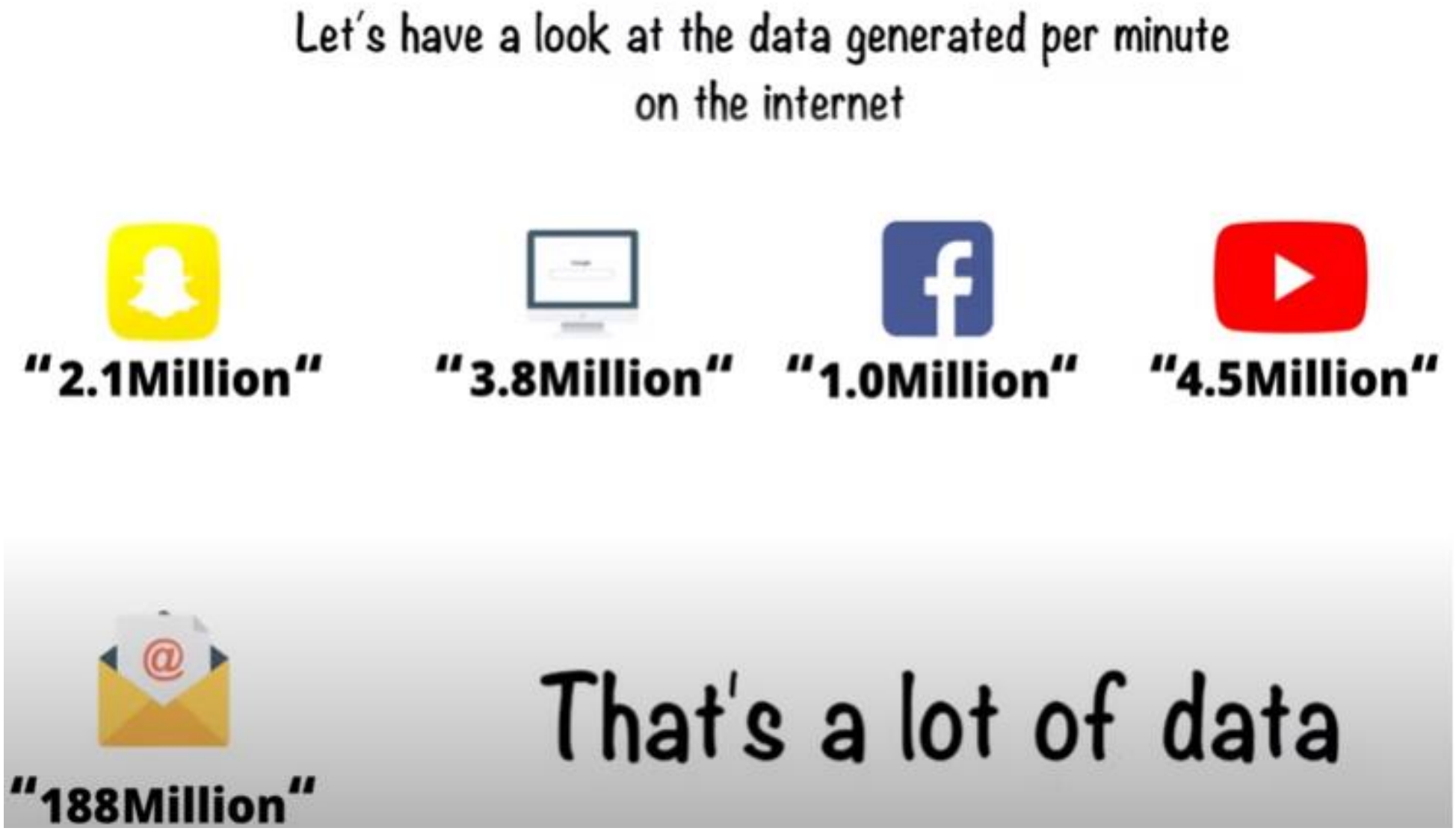
Big Data

Introduction to Big data

Introduction to Big data



Introduction to Big data

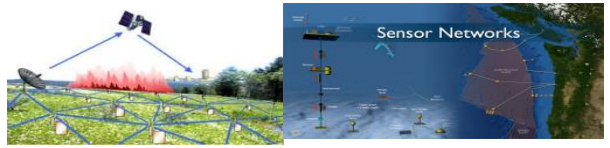
- Facebook: Facebook is collecting a huge amount of data. Every time whenever you are clicking a notification, visiting a page, uploading a photo, or checking out a friend's link, you're generating data for the company to track various records.
- Users shared 2.5 billion content items daily (status updates + wall posts + photos + videos + comments). 300 million photos are uploaded by users per day. 105 terabytes of data scanned via Hive, Facebook's Hadoop query language in every 30 minutes. 70,000 queries executed on these databases per day. 500+terabytes of new data ingested into the databases every day.

Digital World

Social media and networks
(all of us are generating data)

Mobile devices
(tracking all objects all the time)

Scientific instruments
(collecting all sorts of data)



Sensor technology and networks
(measuring all kinds of data)

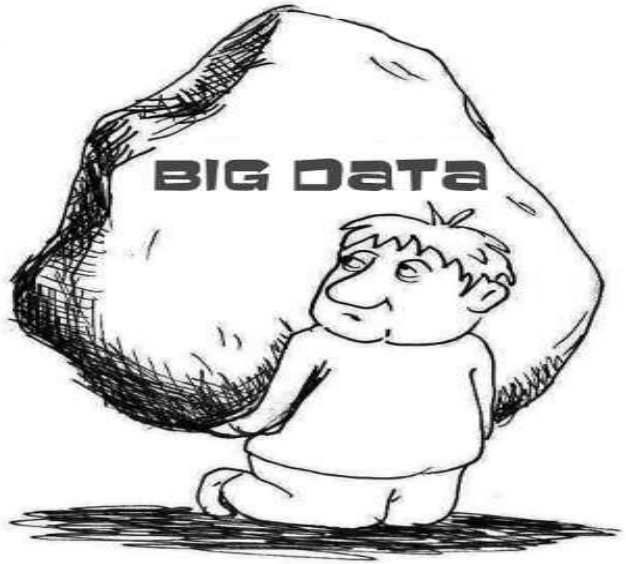
Introduction to Big data

- A massive amount of data that **cannot** be stored, processed, and analysed using the **traditional ways** (Why? => Storage capacity, Processing power)
- The term Big data refers to a huge volume of data that can not be stored processed by any traditional data storage or processing units, that is generated at a very large scale and it is being used to process and analyze in order to uncover insights.
- Big Data is a collection of data that is huge in volume, yet **growing exponentially** with time.

Introduction to Big data

Big Data is the amount of data just beyond technology's capability to store, manage and process efficiently.

"*Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making*"



"*Big data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the informa*tion"

History of big data

- Although the concept of big data itself is relatively new, the origins of large data sets go back to the 1960s and ‘70s when the world of data was just getting started with the first data centres and the development of the relational database.
- Around 2005, people began to realize just **how much data users generated** through Facebook, YouTube, and other online services. Hadoop (an open-source framework created specifically to store and analyse big data sets) was developed that same year. NoSQL also began to gain popularity during this time.

History of big data

- The development of open-source frameworks, such as Hadoop (and more recently, Spark) was essential for the growth of big data because they make big data easier to work with and **cheaper to store**. In the years since then, the volume of big data has skyrocketed. Users are still generating huge amounts of data—but it's not just humans who are doing it.
- With the advent of the Internet of Things (IoT), more objects and devices are connected to the internet, gathering data on customer usage patterns and product performance. The emergence of machine learning has produced still more data.

History of big data

- While big data has come far, its usefulness is only just beginning. Cloud computing has expanded big data possibilities even further. The cloud offers truly elastic scalability, where developers can simply spin up ad hoc clusters to test a subset of data. And **graph databases** are becoming increasingly important as well, with their ability to **display massive amounts of data** in a way that makes analytics fast and comprehensive.

Small Vs Big Data

Small / Traditional Data	Big Data
Mostly Structured	Structured, Unstructured and Semi-structured
Data store in MB, GB, TB	Data store in PB, EB
Data Increase Gradually	Data Increases Exponentially
Locally Present, Save in Centralized manner Example: Universities / Colleges data (Attendance, Library)	Globally Present, Distributed Example: Facebook, Google
Software: SQL Server, Oracle	Software: Hadoop, Spark, Big Query
To handle traffic - > Single node	To handle traffic -> Multinode cluster

Types Of Big Data

- **Structured**: Structured data owns a dedicated data model, is also has a **well defined structure**, it follows a consistent order and it is designed in such a way that it can be easily accessed and used by a person or a computer. Structured data is usually stored in well defined columns and also databases. **Example: DBMS**
- **Unstructured**: Different type od data which neither has a **structure** nor obeys to follow the **formal structure** rules of data models. It does not even have a consistent format and it found to be varying all the time. But rarely it may have information related to data and time.

  Example: audio, video, images etc.
- **Semi-Structured**: can be considered as another form of structure data. It inherits a few properties of structured data, but the major part of this kind of data **fails** to **have** a **definite structure** and also it does not obey the formal structure of data models such as RDBMS. **Example: CSV File**

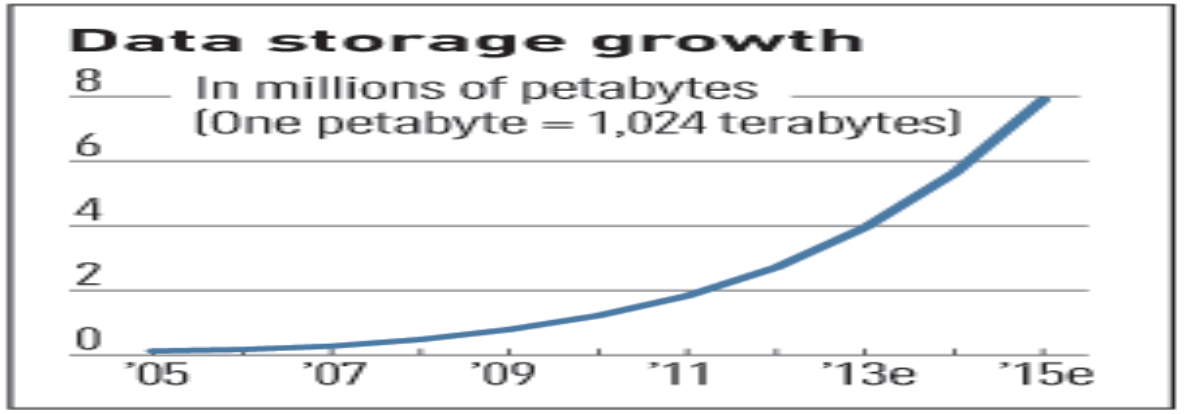
Characteristics Of Big Data

- Big data can be described by the following characteristics:

➢ Volume
➢ Variety
➢ Velocity
➢ Veracity
➢ Value

Characteristics Of Big Data

- **Volume –** The name Big Data itself is related to a **size** which is enormous. Size of data plays a very crucial role in determining value out of data. Also, whether a particular data can actually be considered as a Big Data or not, is dependent upon the volume of data. Hence, **'Volume'** is one characteristic which needs to be considered while dealing with Big Data solutions.
- Example: Facebook alone can generate about billion messages, 4.5 billion times that the “like” button is recorded and over 350 million new posts are uploaded each day.



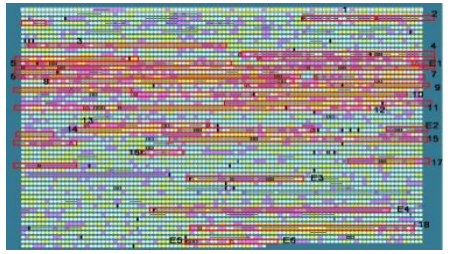
Characteristics Of Big Data

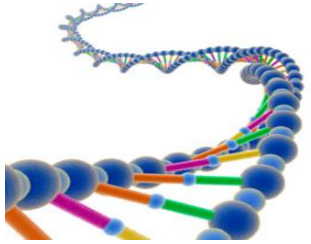
- **Variety** – The next aspect of Big Data is its variety.

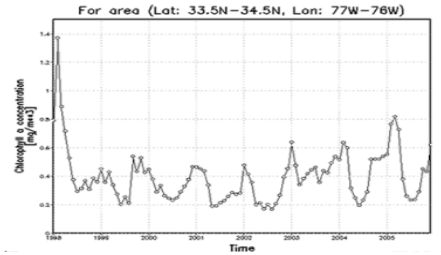
➢ Variety refers to **heterogeneous sources and the nature of data**, both structured and unstructured. During earlier days, spread sheets and databases were the only sources of data considered by most of the applications. Nowadays, data in the form of **emails, photos, videos, monitoring devices, PDFs, audio, etc**. are also being considered in the analysis applications. This variety of unstructured data poses certain issues for storage, mining and analyzing data.

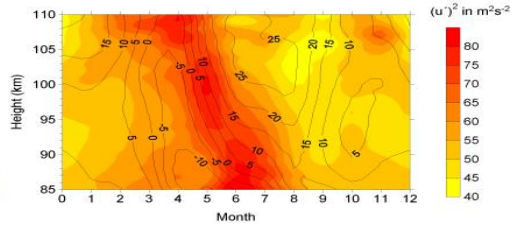
Characteristics Of Big Data

- Various formats, types, and structures
- Text, numerical, images, audio, video, sequences, time series, social media data, multi-dim arrays, etc…
- A single application can be generating/collecting many types of data









To extract knowledge➔ all these types of data need to linked together

Characteristics Of Big Data

- **Velocity** – The term 'velocity' refers to the **speed of generation of data**. How fast the data is generated and processed to meet the demands, determines real potential in the data.

➢ Big Data Velocity deals with the speed at which data flows in from sources like business processes, application logs, networks, and social media sites, sensors, Mobile devices, etc. The flow of data is massive and continuous.

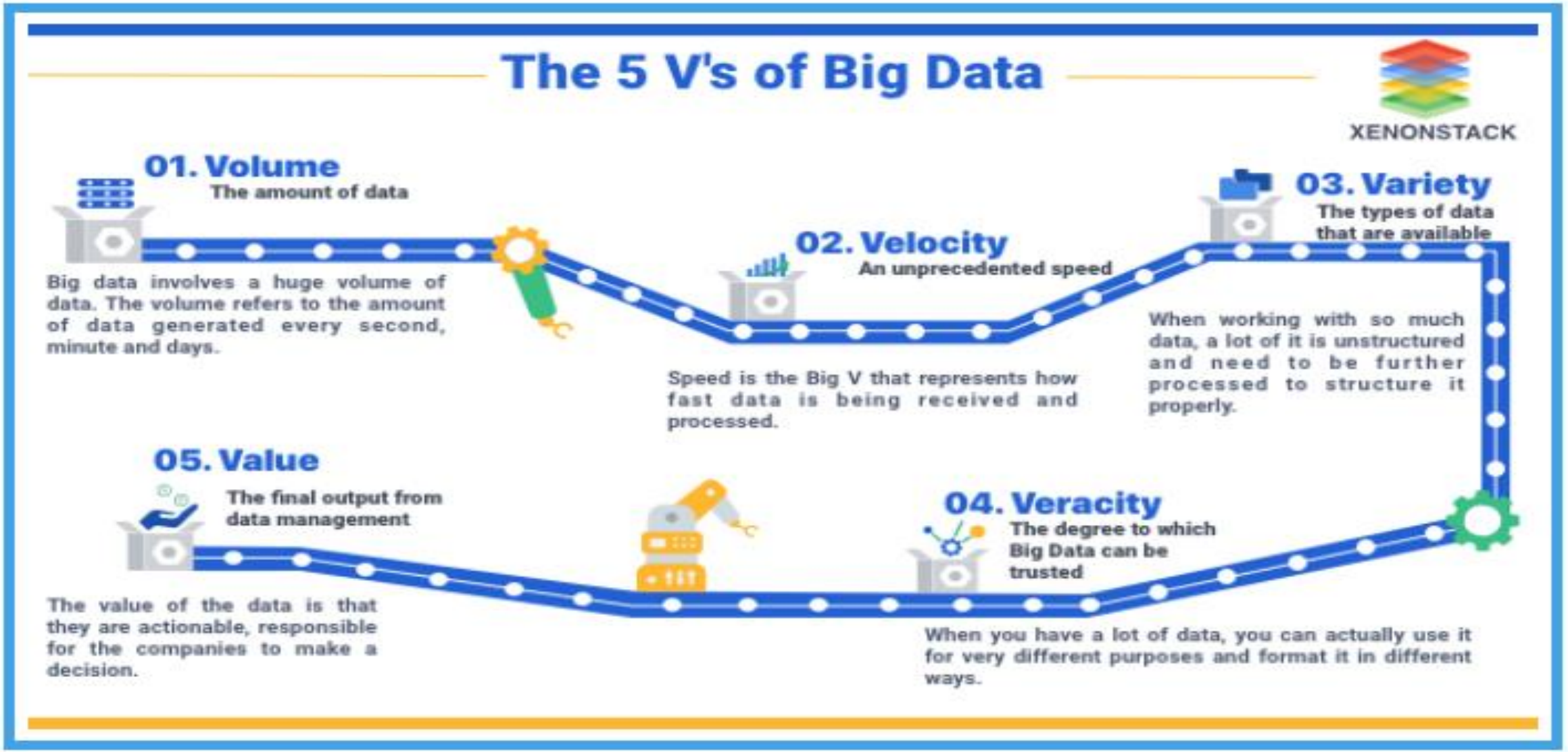
Characteristics Of Big Data

- Data is begin generated fast and need to be processed fast
- Online Data Analytics
- Late decisions ➔ missing opportunities
- **Examples**
  - **E-Promotions:** Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you
  - **Healthcare monitoring:** sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction

Characteristics Of Big Data

- **Veracity** – Veracity basically means the degree of reliability that the data has to offer. Since a major part of the data is **unstructured** and **irrelevant**, Big data needs to find an alternate way to filter them or to translate them out as the data is crucial in business developments
- **Value –** The value of the data is that they are actionable, responsible for the companies to make a decisions. Value is a major issue that we need to concentrate on. It is not just the amount of data that we store or process. It is the amount of valuable, reliable and trustworthy data that needs to stored, processed and analyzed to find insights.

Characteristics Of Big Data



Characteristics Of Big Data

- **Additional V's of Big Data**
- Other characteristics and properties are as follows:

➢ **Visualization** means collecting and analyzing a huge amount of information using Real time analytics to make it understandable and easy to read. Without this, it is impossible to maximize and leverage the raw information.

➢ **Validity**: It means how clean, accurate, and correct the information is to use. The benefit of analytics is only as good as its underlying information, so good data governance practices should be adopted to ensure consistent data quality, common definitions, and metadata.

Characteristics Of Big Data

- **Additional V's of Big Data**

- **Volatility**: How long is data valid and how long should it be stored. In this world of real time data we need to determine at what point is data no longer relevant to the current analysis.
- **Vulnerability**: A huge volume of data comes up with many new security concerns since there have been many big data breaches.
- **Variability**: Some data streams can have peaks and seasonality, periodicity. Managing a large amount of unstructured information is difficult and requires powerful processing techniques.

Applications Of Big Data

- **Entertainment**: Netflix and amazon use big data to make shows and movie recommendation to their users
- **Insurance**: uses big data to predict illness, accidents and price their products accordingly.
- **Driverless cars**: Google's driverless cars collect about one gigabyte data per second. These experiments require more and more data for their successful execution.
- **Education**: Opting for big data part technology as a learning tool instead of traditional lecture methods which enhance the learning of students as well as aided the teacher to track the performance better.
- **Automobiles**: Rolls Royce has embraced big data by fitting hundreds of sensors into its engines, Which record every tiny detail about their operation.

Applications Of Big Data

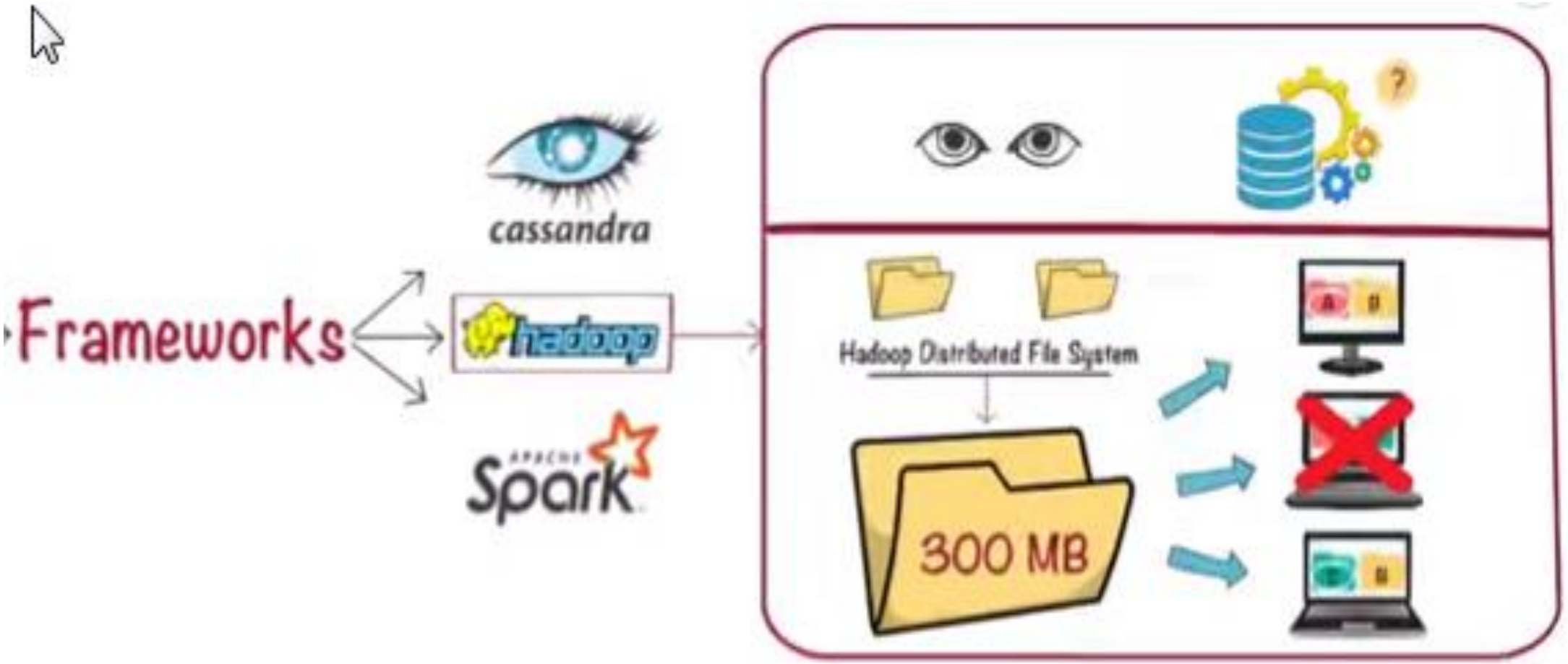
- **Travel and tourism**: Enabled us to predict requirements of travel facilities in many places improving the business through dynamic pricing and many more.
- **Financial and banking sector**: Big data analytics can aid banks in understanding customer behavior based on inputs received from the investments patterns, shopping trends motivation to invent and personal or financial backgrounds.
- **Healthcare Sector**: Medical professionals and healthcare persons are now capable to provide personalized healthcare services to individual patients.
- **Telecommunication and media sector**: Zettabytes of data getting generated every day and to handle such huge data we need big data technologies.
- **Politics**: To analyze patters and influence election results.

Benefits Of Big Data

- Better decision making
- Greater innovations
- Improvement in education sector
- Product price optimization
- Recommendation engines
- Life-Saving application in the healthcare industry
- Predictive analysis which can serve organizations from operational risk.
- Organizations grow business by analyzing customer needs.
- Enabled many multimedia platforms to share data Ex. YouTube, Instagram

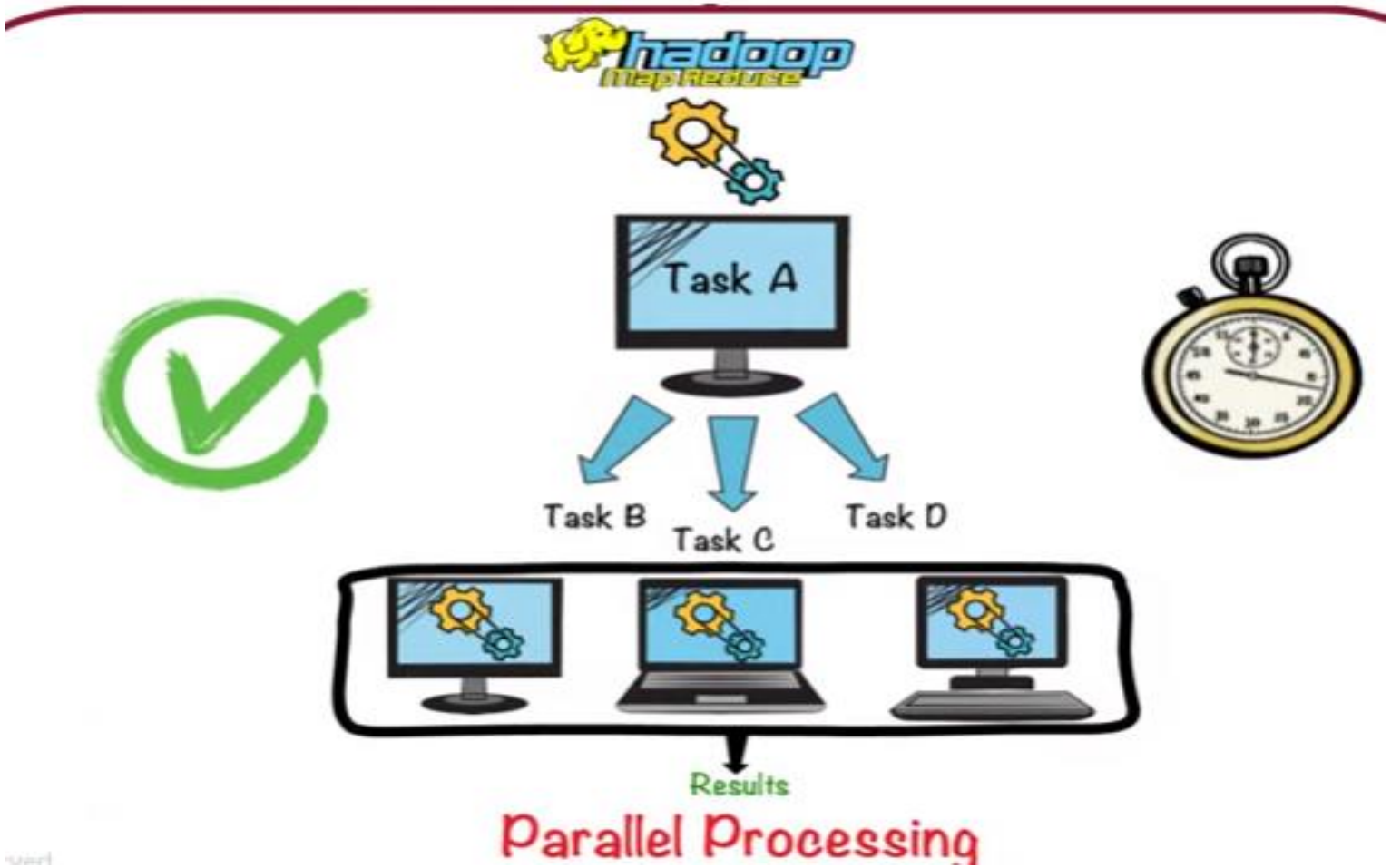
Frameworks

- **How do we store this big data?**



Frameworks

- **How do we process this big data?**



Big Data Frameworks/Tools

- Here is the list of top 10 big data tools –

- Apache Hadoop
- Apache Spark
- Kafka
- Flink
- Apache Storm
- Apache Cassandra
- MongoDB
- Tableau
- RapidMiner
- R Programming

Big Data Frameworks/Tools

- These big data tool not only helps you in **storing large data** but also helps in **processing the stored data in a faster way** and provides you better results and new ideas for the growth of your business.
- There are a vast number of Big Data tools available in the market. You just need to choose the right tool according to the requirements of your project.
- Remember, “If you choose the right tool and use it properly, you will create something extraordinary; If used wrong, it makes a mess.”

Big Data Case studies

- **Walmart**

- Walmart leverages Big Data and Data Mining to create personalized product recommendations for its customers. With the help of these two emerging technologies, Walmart can uncover valuable patterns showing the **most frequently bought products, most popular products**, and even the most **popular product bundles** (products that complement each other and are usually purchased together).
- Based on these insights, Walmart creates attractive and customized recommendations for individual users. By effectively implementing Data Mining techniques, the retail giant has successfully increased the conversion rates and improved its customer service substantially. Furthermore, Walmart uses Hadoop and NoSQL technologies to allow customers to access real-time data accumulated from disparate sources.

Big Data Case studies

- **Uber**

➢ Uber is one of the major cab service providers in the world. It leverages customer data to **track and identify the most popular and most used services by the users**. Once this data is collected, Uber uses data analytics to analyze the usage patterns of customers and determine which services should be given more emphasis and importance.

- Apart from this, Uber uses Big Data in another unique way. Uber closely studies the demand and supply of its services and changes the cab fares accordingly. It is the surge pricing mechanism that works something like this – suppose when you are in a hurry, and you have to book a cab from a crowded location, Uber will charge you double the normal amount!

Big Data Case studies

- **Starbucks**: Starbucks use big data to analyze the **preferences** of their customers to enhance and personalize their experience. They analyze their members **coffee buying habits** along with their **prefer drinks**.
- ➢ So even when people visit a new starbucks location that stores point-of-scale system is able to identify customer through their smartphone and give their buy star as their preferred order in addition based on ordering preferences their app will suggest the new products that the customers might be interested in trying.
- This is what we call big data analytics.

What is Big Data Analytics

- Big data analytics is largely used by companies to facilitate their growth and development. This majorly involves **applying** various **data mining algorithms** on the given set of data, which will then aid them in better **decision making**.

> *"Big data analytics examines large and different types of data to uncover hidden patterns, correlations and other insights"*

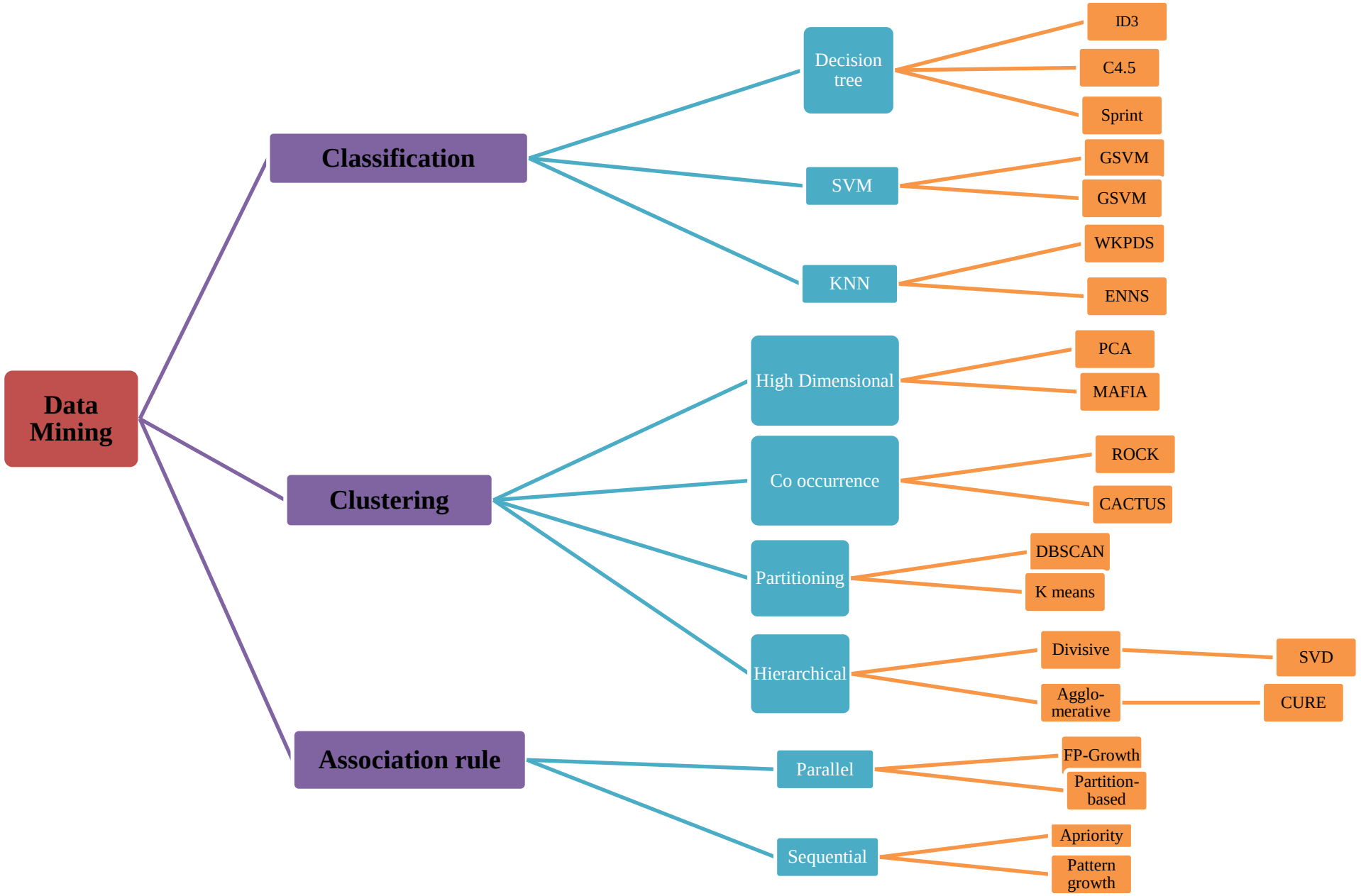


Figure: Taxonomy of Data mining algorithms

Need of Big Data Analytics

1. Making Smarter and more efficient organization

- Big data analytics is basically highly contributing to these factors and organizations are adopting this to basically lead them to faster decision making.
- Example: New York Police Department (NYPD)

- Big data and analytics helping the NYPD and the other large police departments to **anticipate and identify the criminal activity** before it occurs. so what they do is that they analyze the big data technology to geo locate and then analyze the historical patterns and they map these historical patterns with sporting events pea days, rain falls, traffic flows and federal holidays.

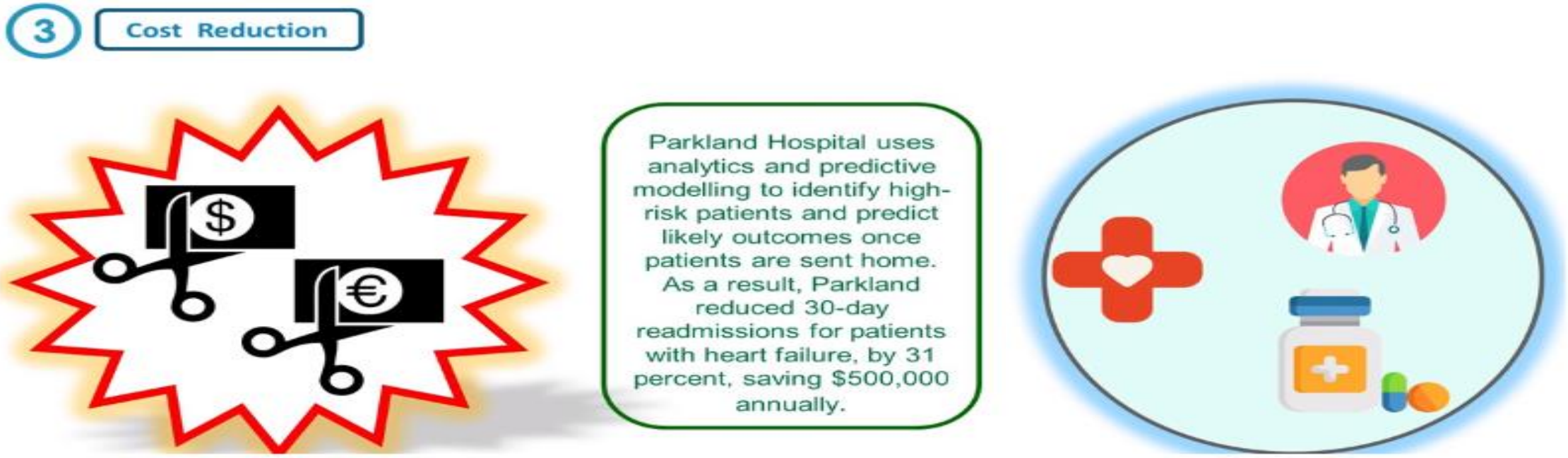
Need of Big Data Analytics

- Example: New York Police Department (NYPD)
- So essentially What the NYPD is doing that they utilizing these data patterns, scientific analytics, technological tools to do their job and they are ensuring that by using these different tools they are doing their job to the best of their ability.
- So by using a **big data and analytic strategy**; the NYPD is able to **identify** something called **crime hotspot** so basically where crime occurrence was more. so they were able to identify these hotspot and then from there they deployed their local officers so that they could reach there on time before it was actually committed.

Need of Big Data Analytics

- **Example:** Amazon uses customer click stream data and historical purchase data of more than 300 million customer and each user is shown customized results on customized web pages.

Need of Big Data Analytics



- Reduce cost by using Hadoop technology. Hadoop stores big data in distributed fashion so that we can process it parallel. So it reduces our cost a lot. So by using commodity hardware they are educing their cost significantly.

Need of Big Data Analytics

4 Next Generation Products

Big Data tools are used to operate Google's Self Driving Cars. The Toyota Prius is fitted with cameras, GPS as well as powerful computers and sensors to safely drive on the road without the intervention of human beings.

Netflix launched the seasons of its TV show House of Cards based on the user reviews, ratings and viewership.

NETFLIX

A smart yoga mat has sensors embedded in the mat will be able to provide feedback on your postures, score your practice, and even guide you through an at-home practice.

How big data analytics works

- Big data analytics refers to collecting, processing, cleaning, and analyzing large datasets to help organizations operationalize their big data.
- **Collect Data**

➢ Data collection looks different for every organization. With today's technology, organizations can gather both structured and unstructured data from a variety of sources — from **cloud storage to mobile applications to in-store IoT sensors and beyond**. Some data will be stored in data warehouses where business intelligence tools and solutions can access it easily. Raw or unstructured data that is too diverse or complex for a warehouse may be assigned metadata and stored in a data lake.

How big data analytics works

- **Process Data**

➢ Once data is collected and stored, it must be organized properly to get accurate results on analytical queries, especially when it's large and unstructured. Available data is growing exponentially, making data processing a challenge for organizations. One processing option is **batch processing**, which looks at large data blocks over time. Batch processing is useful when there is a longer turnaround time between collecting and analyzing data. **Stream processing** looks at small batches of data at once, shortening the delay time between collection and analysis for quicker decision-making. Stream processing is more complex and often more expensive.

How big data analytics works

- **Clean Data**

➢ Data big or small requires scrubbing to improve data quality and get stronger results; all data must be formatted correctly, and any duplicative or irrelevant data must be eliminated or accounted for. Dirty data can obscure and mislead, creating flawed insights.

How big data analytics works

- **Analyze Data**

➢ Getting big data into a usable state takes time. Once it’s ready, advanced analytics processes can turn big data into big insights. Some of these big data analysis methods include:

- **Data mining** sorts through large datasets to identify patterns and relationships by identifying anomalies and creating data clusters.
- **Predictive analytics** uses an organization’s historical data to make predictions about the future, identifying upcoming risks and opportunities.
- **Deep learning** imitates human learning patterns by using artificial intelligence and machine learning to layer algorithms and find patterns in the most complex and abstract data.

Aspects of Big data

Figure:Overview of different aspects of Big data [6]

The big challenges of big data

- Big data brings big benefits, but it also brings big challenges such new privacy and security concerns, accessibility for business users, and choosing the right solutions for your business needs. To capitalize on incoming data, organizations will have to address the following:
- **Making big data accessible**. Collecting and processing data becomes more difficult as the amount of data grows. Organizations must make data easy and convenient for data owners of all skill levels to use.
- **Maintaining quality data.** With so much data to maintain, organizations are spending more time than ever before scrubbing for duplicates, errors, absences, conflicts, and inconsistencies.

The big challenges of big data

- **Keeping data secure.** As the amount of data grows, so do privacy and security concerns. Organizations will need to strive for compliance and put tight data processes in place before they take advantage of big data.
- **Finding the right tools and platforms.** New technologies for processing and analyzing big data are developed all the time. Organizations must find the right technology to work within their established ecosystems and address their particular needs. Often, the right solution is also a flexible solution that can accommodate future infrastructure changes.

Thank you.