Artificial Intelligence

By

Dr. Manoj Kumar

University School of Automation and Robotics

GGSIP University, East Campus, Delhi, India

What is machine learning?

"Learning is any process by which a system improves performance from experience."
- Herbert Simon

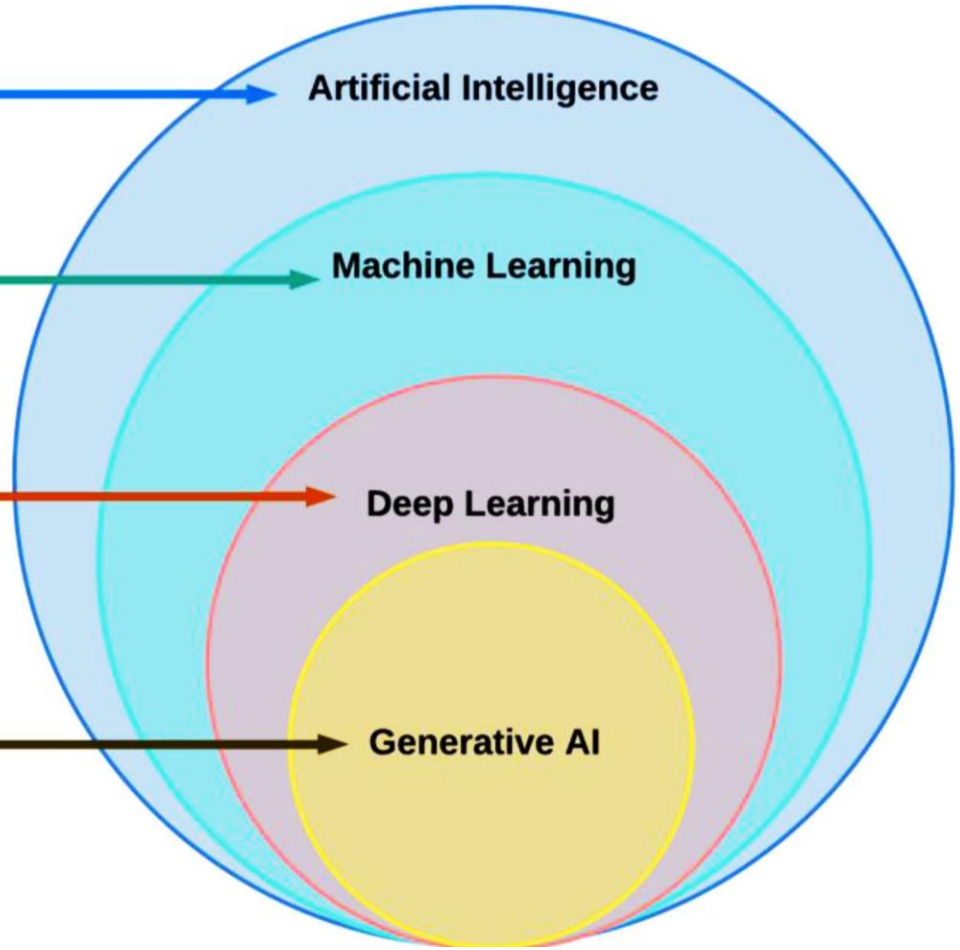
- A branch of artificial intelligence, concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data.
- As intelligence requires knowledge, it is necessary for the computers to acquire knowledge.
- Machine learning refers to a system capable of the autonomous acquisition and integration of knowledge

Artificial Intelligence
AI involves techniques that equip computers to emulate human behavior, enabling them to learn, make decisions, recognize patterns, and solve complex problems in a manner akin to human intelligence.

Machine Learning
ML is a subset of AI, uses advanced algorithms to detect patterns in large data sets, allowing machines to learn and adapt. ML algorithms use supervised or unsupervised learning methods.

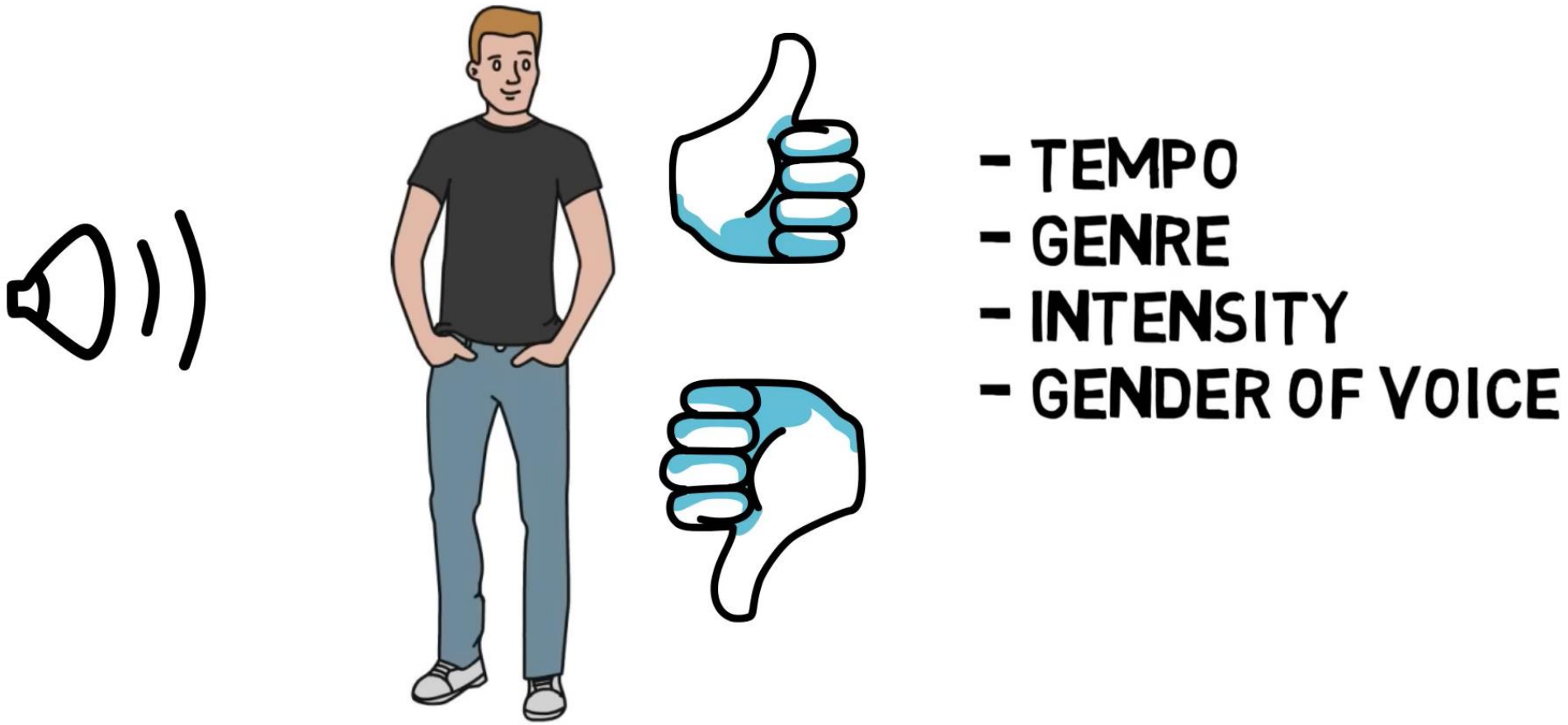
Deep Learning
DL is a subset of ML which uses neural networks for in-depth data processing and analytical tasks. DL leverages multiple layers of artificial neural networks to extract high-level features from raw input data, simulating the way human brains perceive and understand the world.

Generative AI
Generative AI is a subset of DL models that generates content like text, images, or code based on provided input. Trained on vast data sets, these models detect patterns and create outputs without explicit instruction, using a mix of supervised and unsupervised learning.

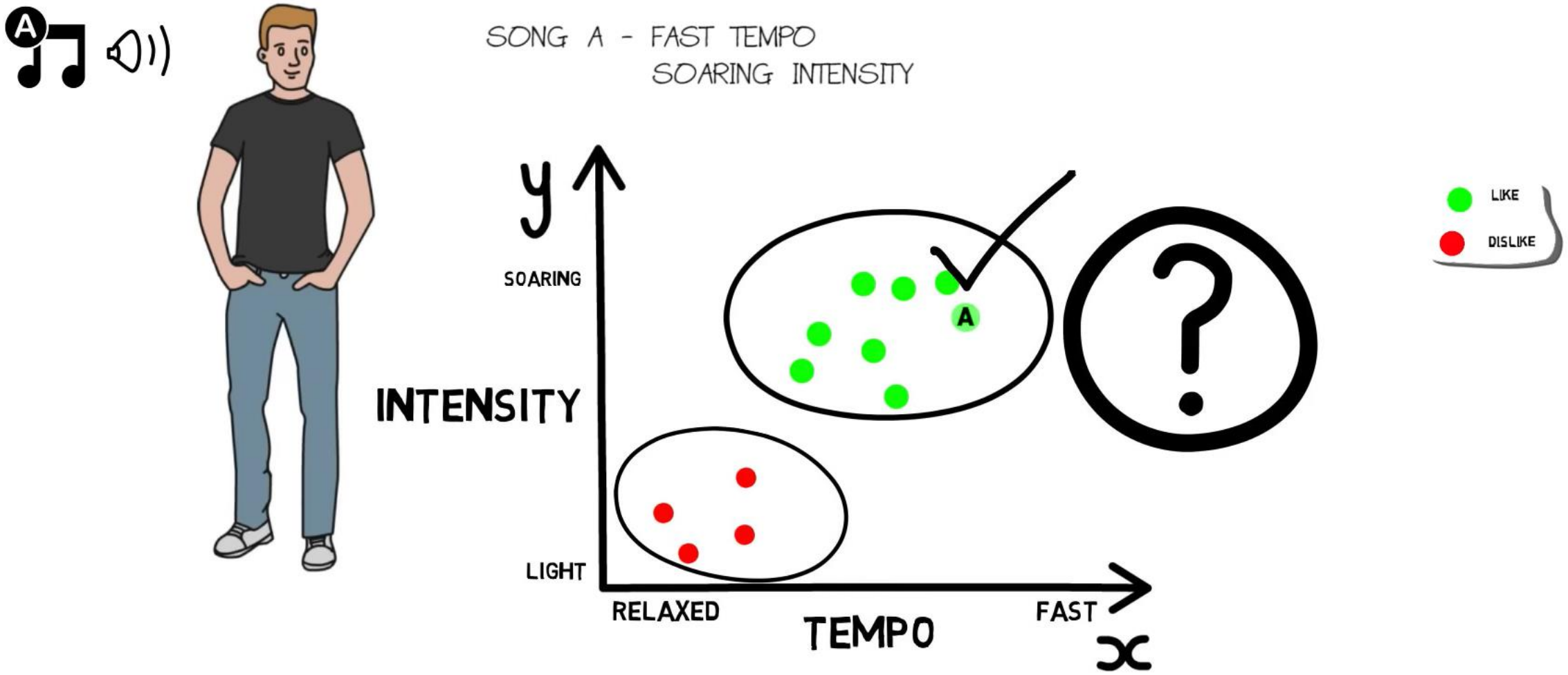


What is machine learning?

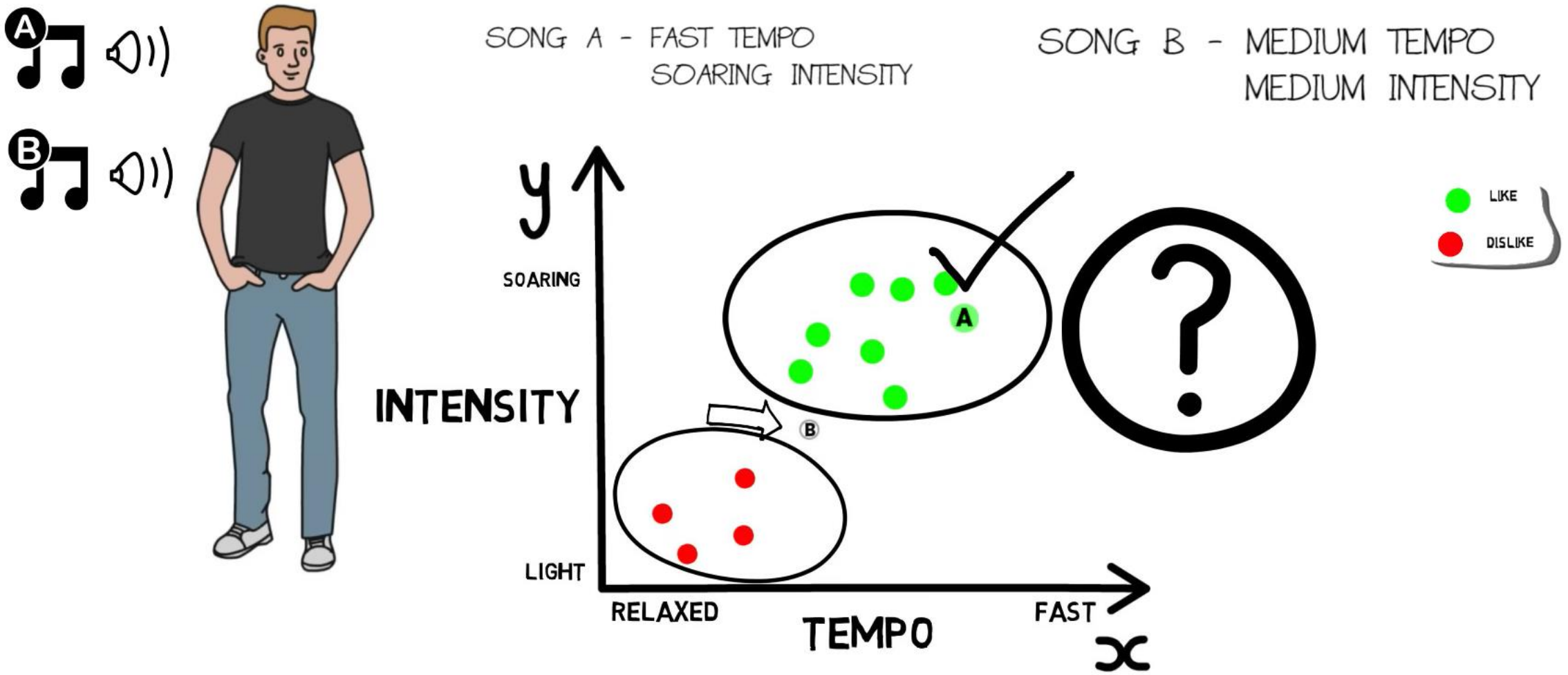
- **What if human can train the machines to learn from the past data and do what humans can do and much faster?**



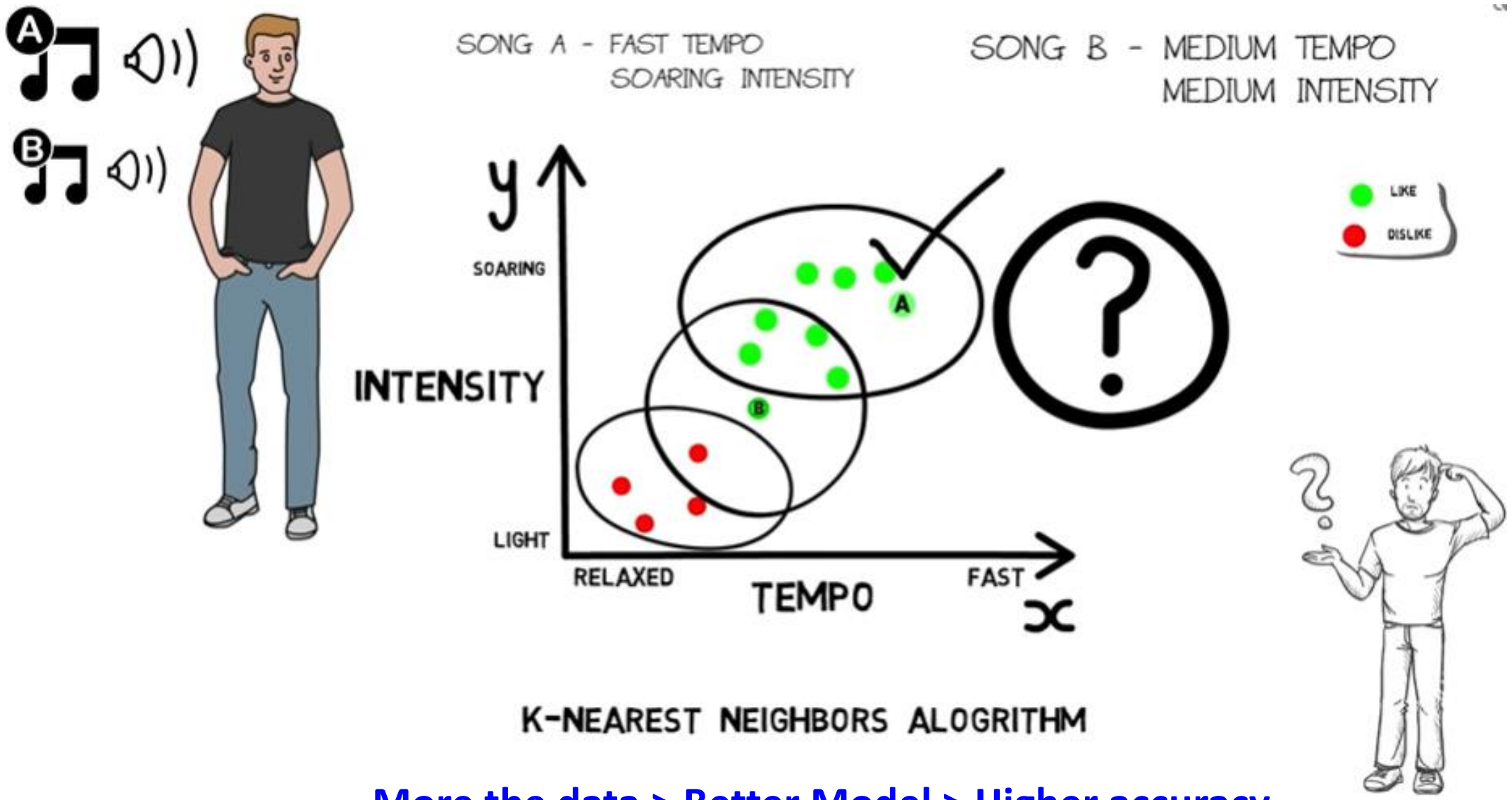
What is machine learning?



What is machine learning?



What is machine learning?



More the data > Better Model > Higher accuracy

What is machine learning?

- The name machine learning was coined in 1959 by Arthur Samuel Tom M. Mitchell provided a widely quoted, more formal definition of the algorithms studied in the machine learning:

An agent/computer program is said to learn from experience with respect to some class of tasks (T), and a *performance measure (P)*, if its [the learner's: agent/computer program] performance at tasks in the class, as measured by P, improves with experience (E). "A well-defined learning task is given by <P,T,E>.

It could be diagnosing the illness of patient: the number of patients didn't have adverse reaction of medicine you gave
Writing exams : The number of marks you got.

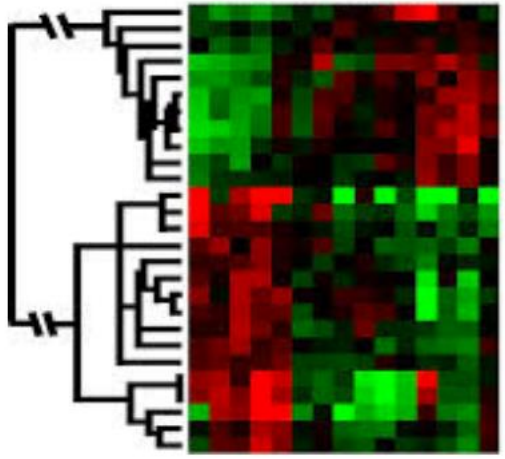
When you learning to improve your performance based on experience is known as: Inductive Learning

When Do We Use Machine Learning?

ML is used when:

- Human expertise does not exist (navigating on Mars)
- Humans can't explain their expertise (speech recognition)
- Models must be customized (personalized medicine)
- Models are based on huge amounts of data (genomics)

No human experts
industrial/manufacturing control.
mass spectrometer analysis, drug design, astronomic discovery.
Black-box human expertise
face/handwriting/speech recognition.
driving a car, flying a plane.



Rapidly changing phenomena credit
scoring, financial modeling.
diagnosis, fraud detection.
Need for customization/personalization
personalized news reader.
movie/book recommendation.

Defining the Learning Task

Improve on task T, with respect to performance metric P, based on experience E

T: Playing checkers
P: Percentage of games won against an arbitrary opponent
E: Playing practice games against itself

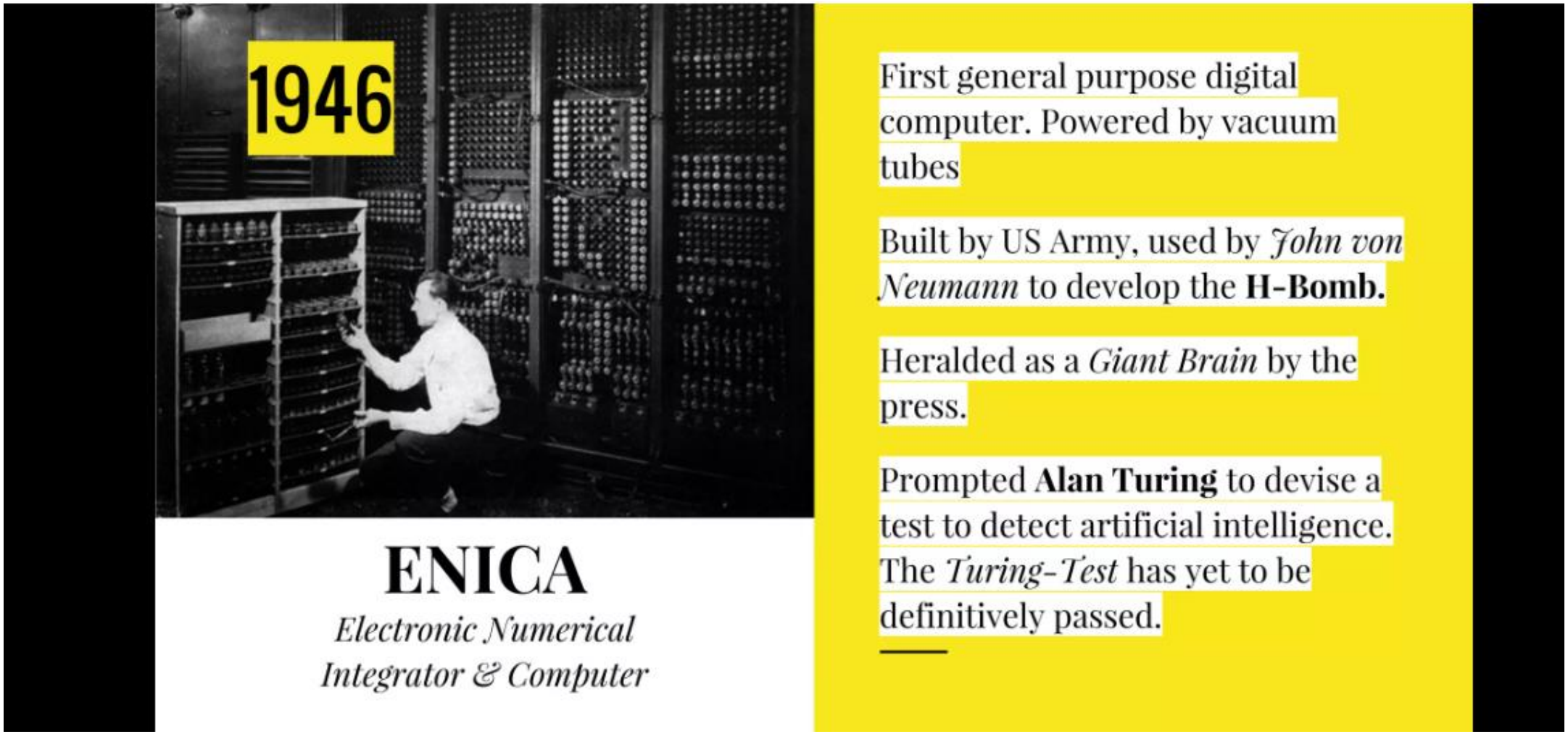
T: Recognizing hand-written words
P: Percentage of words correctly classified
E: Database of human-labeled images of handwritten words

T: Driving on four-lane highways using vision sensors
P: Average distance traveled before a human-judged error
E: A sequence of images and steering commands recorded while observing a human driver.

T: Categorize email messages as spam or legitimate.
P: Percentage of email messages correctly classified.
E: Database of emails, some with human-given labels

History of machine Learning

- **Machine learning is a field of computer science that deals with the development of algorithms that can learn from data.**



History of machine Learning

- **Machine learning is a field of computer science that deals with the development of algorithms that can learn from data.**

1943: The First Neutral Network with Electric Circuit

The first neutral network with electric circuit was developed by Warren McCulloch and Walter Pitts in 1943. The goal of the network was to solve a problem that had been posed by John von Neumann and others: how could computers be made to communicate with each other?

This early model showed that it was possible for two computers to communicate without any human interaction. This event is important because it paved the way for machine learning development.

History of machine Learning

1950: Turing Test

The Turing Test is a test of artificial intelligence proposed by mathematician Alan Turing. It involves determining whether a machine can act like a human, or if humans can't tell the difference between human and machine given answers.

The goal of the test is to determine whether machines can think intelligently and demonstrate some form of emotional capability. It does not matter whether the answer is true or false but whether it is considered human or not by the questioner. There have been several attempts to create an AI that passes the Turing Test, but no machine has yet successfully done so.

The Turing Test has been criticized because it measures how much a machine can imitate a human rather than proving their true intelligence.

History of machine Learning

1952: Computer Checkers

Arthur Samuel was a pioneer in machine learning and is credited with creating the first computer program to play championship-level checkers. His program, which he developed in 1952, used a technique called **alpha-beta pruning** to measure the chances of winning a game. This method is still widely used in games today. In addition, Samuel also developed the **minimax algorithm**, which is a technique for minimizing losses in games.

History of machine Learning

1957: Frank Rosenblatt – The Perceptron

Frank Rosenblatt was a psychologist who is most famous for his work on machine learning. In 1957, he developed the perceptron, which is a machine learning algorithm. The Perceptron was one of the first algorithms to use artificial neural networks, widely used in machine learning.

It was designed to improve the accuracy of computer predictions. The goal of the Perceptron was to learn from data by adjusting its parameters until it reached an optimal solution. Perceptron's purpose was to make it easier for computers to learn from data and to improve upon previous methods that had limited success.

History of machine Learning

1967: The Nearest Neighbor Algorithm

The Nearest Neighbor Algorithm was developed as a way to automatically identify patterns within large datasets. The goal of this algorithm is to find similarities between two items and determine which one is closer to the pattern found in the other item. This can be used for things like finding relationships between different pieces of data or predicting future events based on past events.

In 1967, Cover and Hart published an article on “Nearest neighbor pattern classification.” It is a method of inductive logic used in machine learning to classify an input object into one of two categories. The pattern classifies the same items that are classified in the same categories as its nearest neighbors. This method is used to classify objects with a number of attributes, many of which are categorical or numerical and may have overlapping values.

History of machine Learning

1974: The Backpropagation

Backpropagation was initially designed to help neural networks learn how to recognize patterns. However, it has also been used in other areas of machine learning, such as boosting performance and generalizing from data sets to new instances. The goal of backpropagation is to improve the accuracy of a model by adjusting its weights so that it can more accurately predict future outputs.

Paul Werbos laid the foundation for this approach to machine learning in his dissertation in 1974, which is included in the book "**The Roots of Backpropagation**".

History of machine Learning

1979: The Stanford Cart

The Stanford Cart is a remote-controlled robot that can move independently in space. It was first developed in the 1960s and reached an important milestone in its development in 1979. The purpose of **the Stanford Cart** is to avoid obstacles and reach a specific destination: In 1979, “The Cart” succeeded for the first time in traversing a room filled with chairs in 5 hours without human intervention.

History of machine Learning

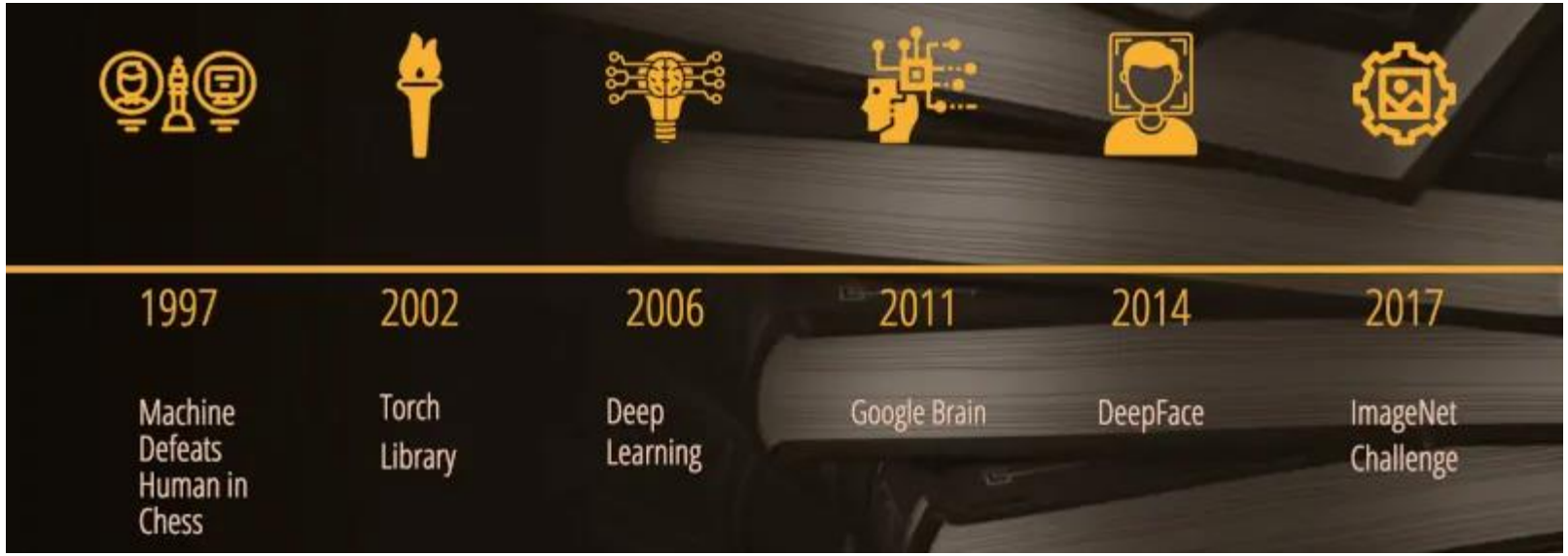
The AI Winter in the History of Machine Learning

AI has seen a number of highs and lows over the years. The low point for AI was known as the AI winter, which happened in the **late 70s to the 90s**. During this time, research funding dried up and many projects were shut down due to their lack of success. It has been described as a series of hype cycles that have led to disappointment and disillusionment among developers, researchers, users, and media.

History of machine Learning

The Rise of Machine Learning in History

The rise of machine learning in the 21th century is a result of **Moore's Law** and its exponential growth. When computing power was becoming more affordable, it became possible to train AI algorithms using more data, which resulted in an increase of the accuracy and efficiency of these algorithms.

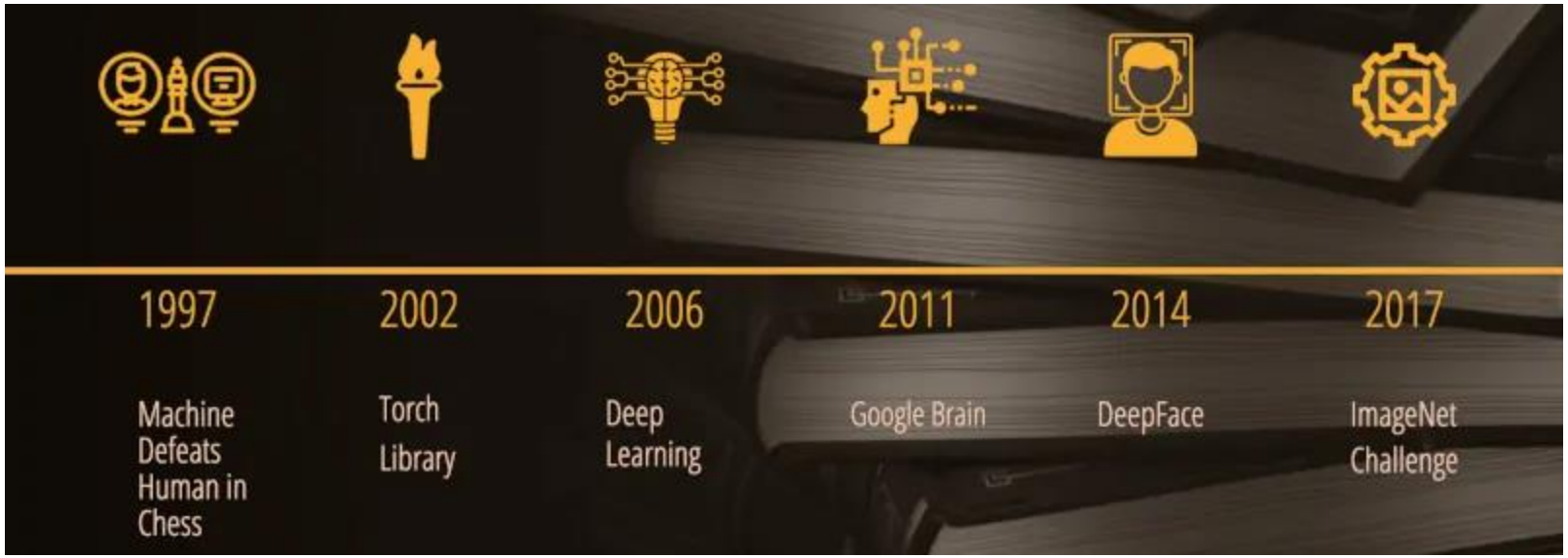


1997: A Machine Defeats a Man in Chess

In 1997, the IBM supercomputer **Deep Blue defeated chess grandmaster Garry Kasparov** in a match. It was the first time a machine had beaten an expert player at chess and it caused great concern for humans in the chess community. This was a landmark event as it showed that AI systems could surpass human understanding in complex tasks.

This marked a magical turning point in machine learning because the world now knew that mankind had created its own opponent- an artificial intelligence that could learn and evolve on its own.

History of machine Learning

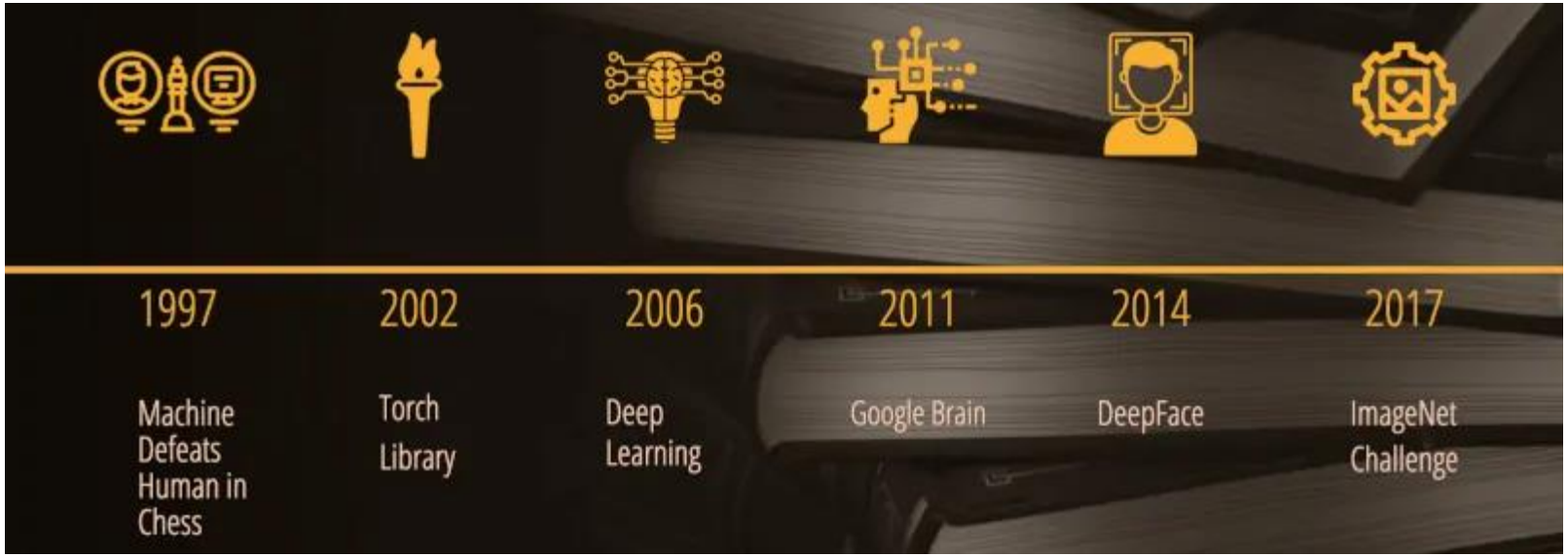


2002: Software Library Torch

Torch is a software library for machine learning and data science. Torch was created by **Geoffrey Hinton**, Pedro Domingos, and Andrew Ng to develop the first large-scale free machine learning platform. In 2002, the founders of Torch created it as an alternative to other libraries because they believed that their specific needs were not met by other libraries. As of 2018, it has over 1 million downloads on Github and is one of the most popular machine learning libraries available today.

Keep in mind: No longer in active development, however, PyTorch can be used, which is based on the Torch Library.

History of machine Learning

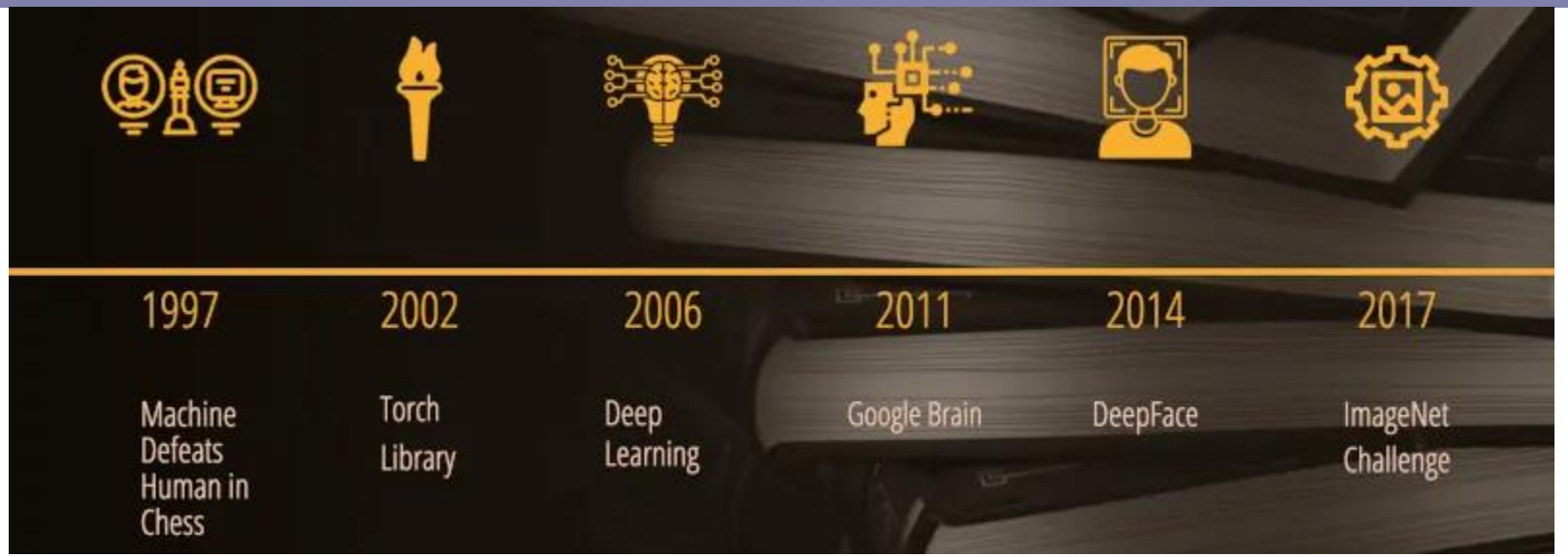


2006: Geoffrey Hinton, the father of Deep Learning

In 2006, Geoffrey Hinton published his “A Fast Learning Algorithm for Deep Belief Nets.” **This paper was the birth of deep learning.** He showed that by using a deep belief network, a computer could be trained to recognize patterns in images.

Hinton’s paper described the first deep learning algorithm that can achieve human-level performance on difficult and complex pattern recognition tasks.

History of machine Learning

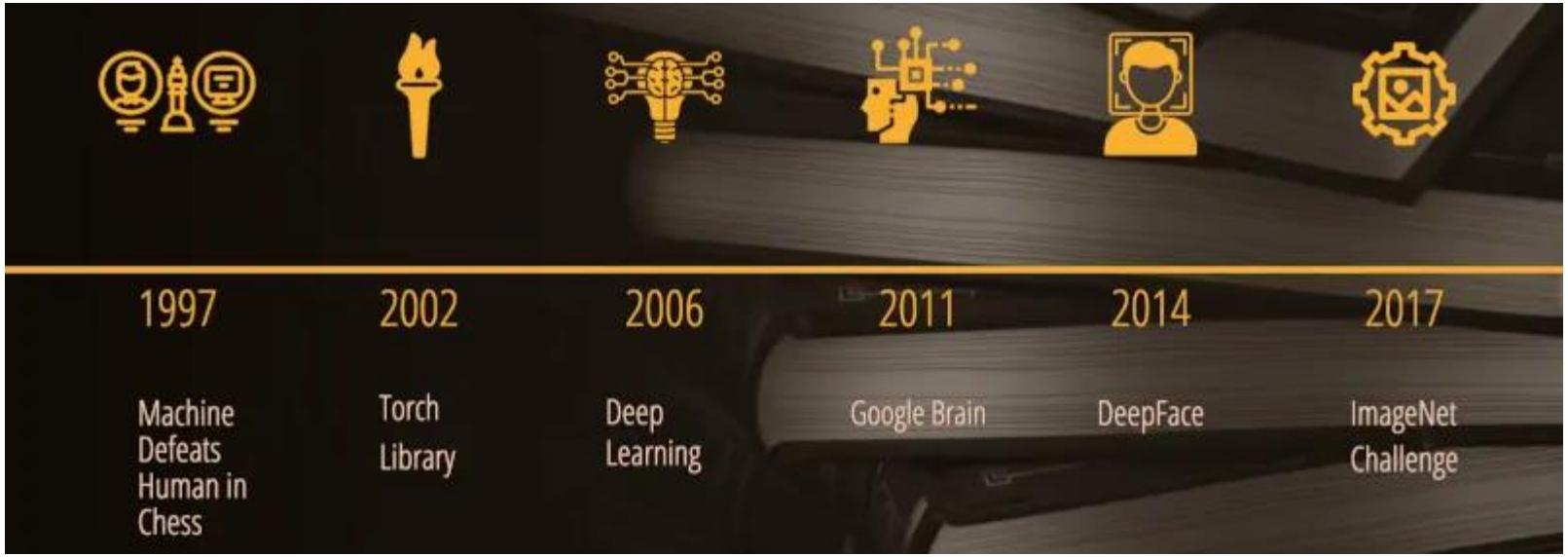


2011: Google Brain

Google Brain is a research group of Google devoted to artificial intelligence and machine learning. The group was founded in 2011 by Google X and is located in Mountain View, California. The team works closely with other AI research groups within Google such as the DeepMind group that has developed AlphaGo, an AI that defeated the world champion at Go. Their goal is to build machines that can learn from data, understand language, answer questions in natural language, and have common sense reasoning.

The group is, as of 2021, led by Geoffrey Hinton, Jeff Dean and Zoubin Ghahramani and focuses on deep learning, a model of artificial neural networks that is capable to learn complex patterns from data automatically without being explicitly programmed.

History of machine Learning

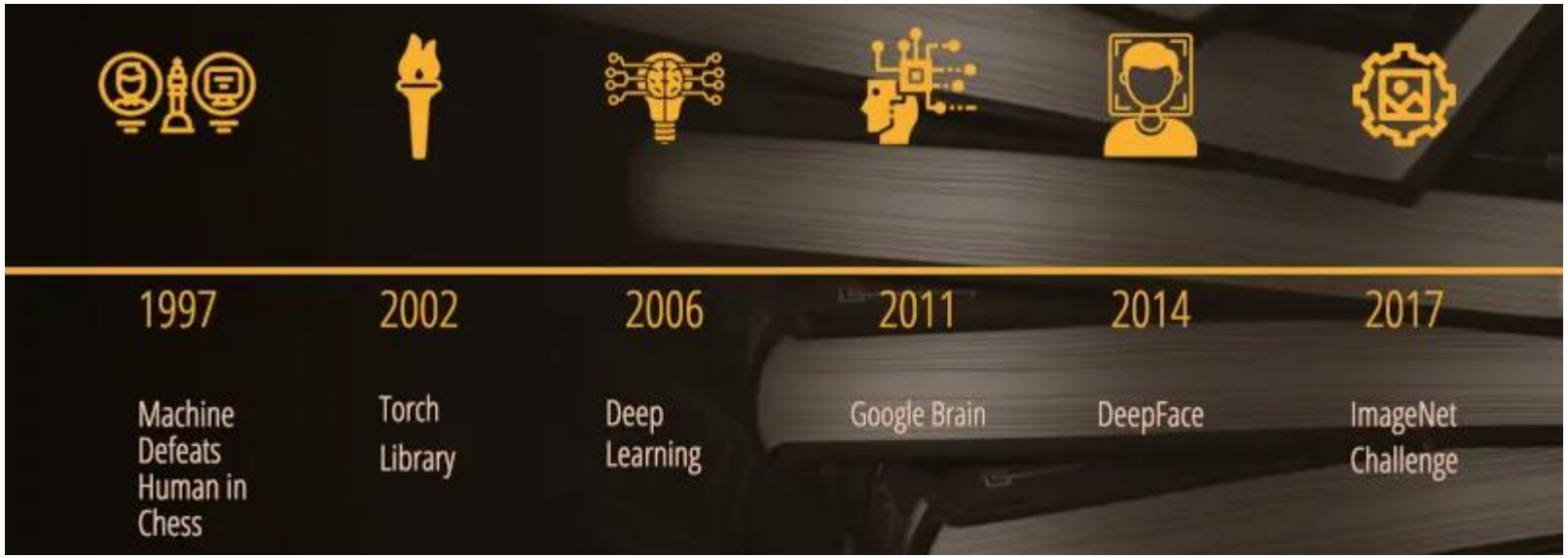


2014: DeepFace

DeepFace is a deep learning algorithm which was originally developed in 2014 and is part of the company “Meta”. The project received significant media attention after it outperformed human performance on the well-known “Faces in the Wild” test.

DeepFace is based on a deep neural network, which consists of many layers of artificial neurons and weights that connect each layer to its neighboring ones. The algorithm takes as input a training data set of photographs, with each photo annotated with the identity and age of its subject. The team has been very successful in recent years and published many papers on their research results. They have also trained several deep neural networks that have achieved significant success in pattern recognition and machine learning tasks.

History of machine Learning



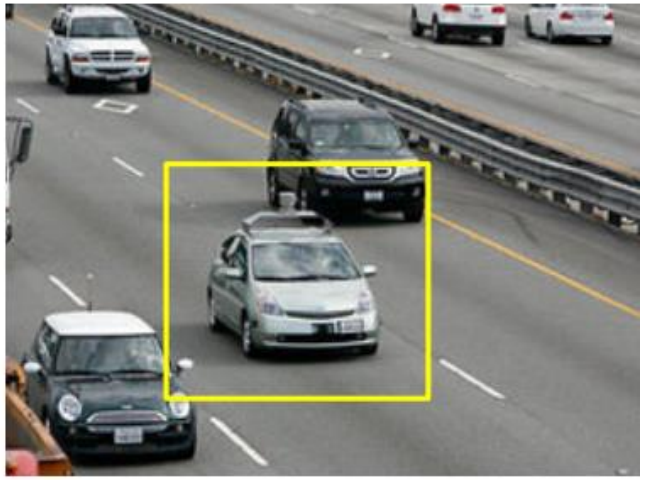
2017: ImageNet Challenge – Milestone in the History of Machine Learning

The **ImageNet Challenge** is a competition in computer vision that has been running since 2010. The challenge focuses on the abilities of programs to process patterns in images and recognize objects with varying degrees of detail.

In 2017, a milestone was reached. **29 out of 38 teams achieved 95%** accuracy with their computer vision models. The improvement in image recognition is immense.

State of the art Applications of machine Learning

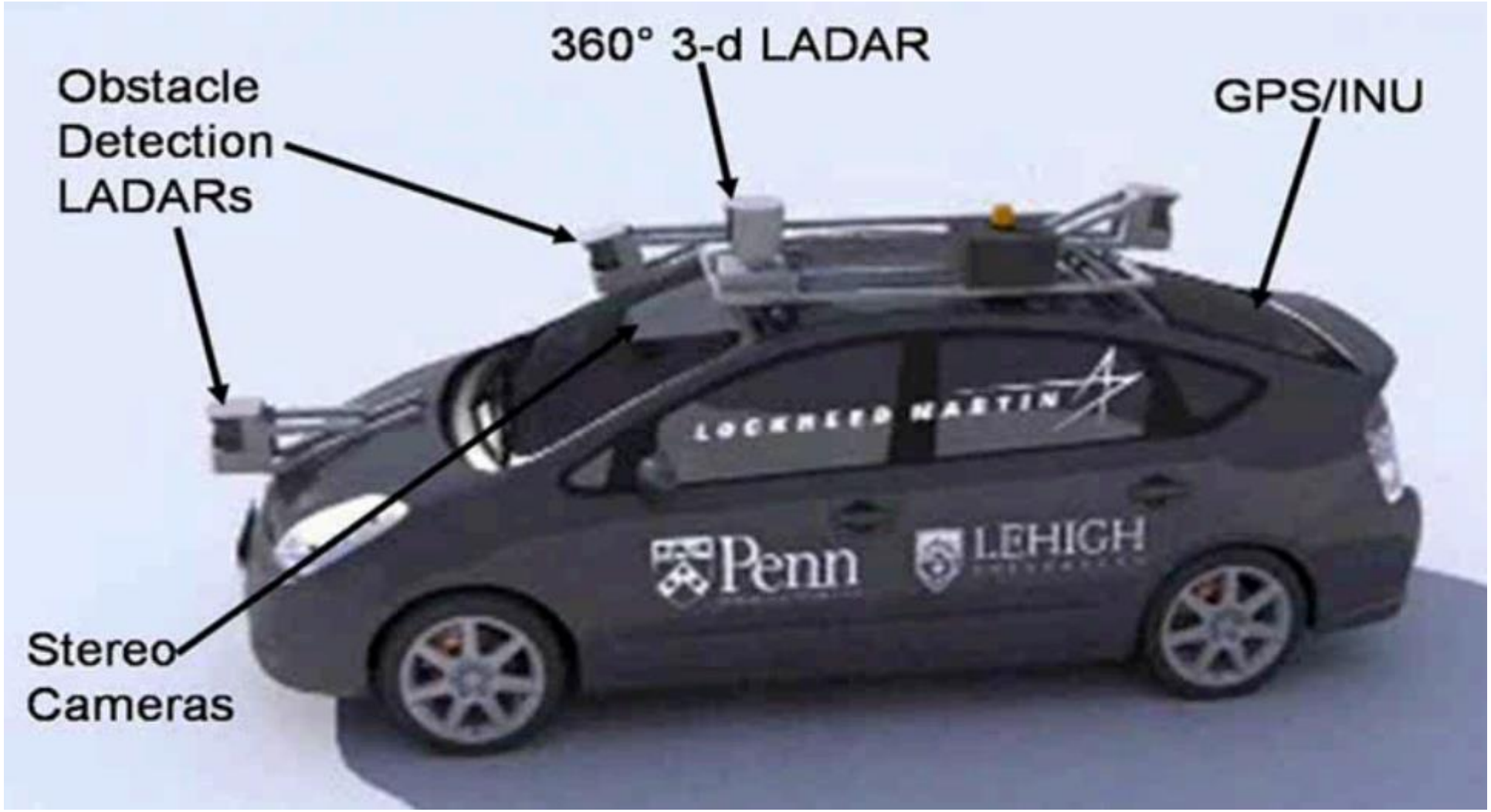
Autonomous Cars



- Nevada made it legal for autonomous cars to drive on roads in June 2011
- As of 2013, four states (Nevada, Florida, California, and Michigan) have legalized autonomous cars

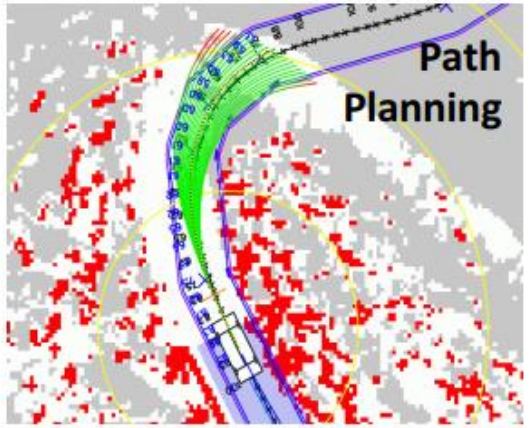
Penn's Autonomous Car →
(Ben Franklin Racing Team)

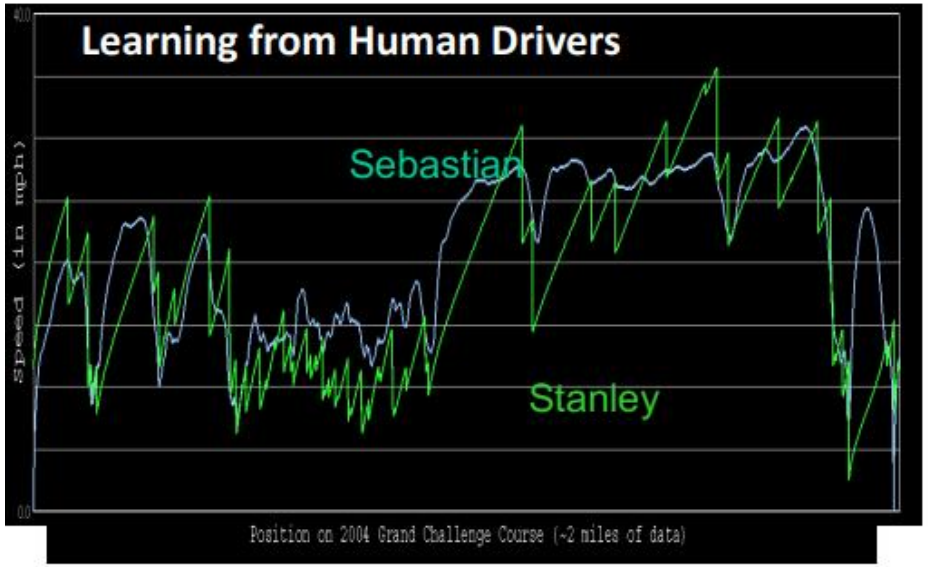
State of the art Applications of machine Learning



State of the art Applications of machine Learning

Autonomous Car Technology





Images and movies taken from Sebastian Thrun's multimedia website.

State of the art Applications of machine Learning

Deep Learning in the Headlines

BUSINESS NEWS

MIT Technology Review

Is Google Cornering the Market on Deep Learning?

A cutting-edge corner of science is being wooed by Silicon Valley, to the dismay of some academics.

By Antonio Regalado on January 29, 2014

How much are a dozen deep-learning researchers worth? Apparently, more than $400 million.

This week, Google reportedly paid that much to acquire DeepMind Technologies, a startup based in

BloombergBusinessweek
Technology

Acquisitions

The Race to Buy the Human Brains Behind Deep Learning Machines

By Ashlee Vance | January 27, 2014

intelligence projects. "DeepMind is bona fide in terms of its research capabilities and depth," says Peter Lee, who heads Microsoft Research.

According to Lee, Microsoft, Facebook (FB), and Google find themselves in a battle for deep learning talent. Microsoft has gone from four full-time deep learning experts to 70 in the past three years. "We would have more if the talent was there to

WIRED GEAR SCIENCE ENTERTAINMENT BUSINESS SECURITY DESIG

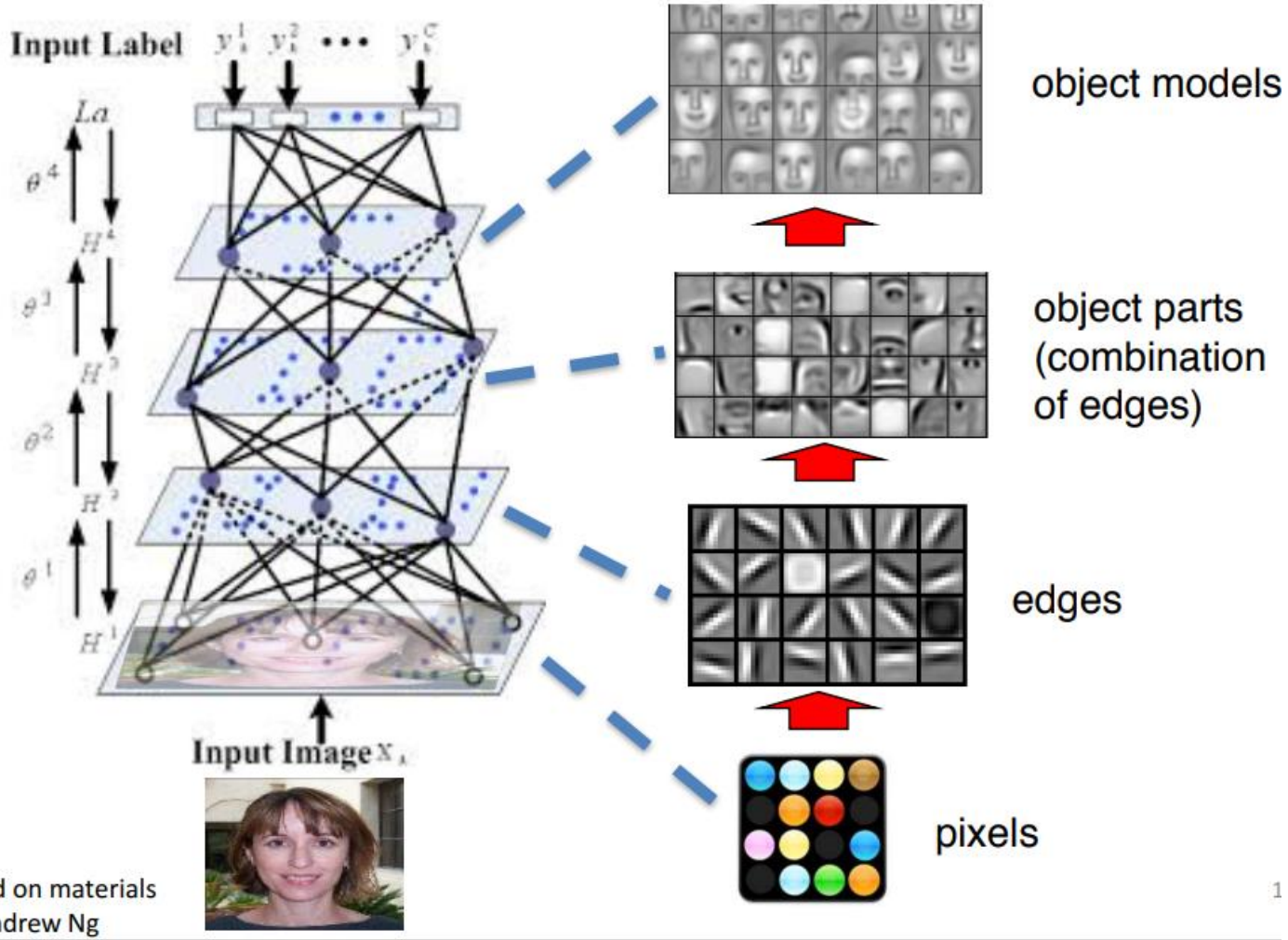
INNOVATION INSIGHTS | community content | featured

Deep Learning's Role in the Age of Robots

BY JULIAN GREEN, JETPAC 05.02.14 2:56 PM

State of the art Applications of machine Learning

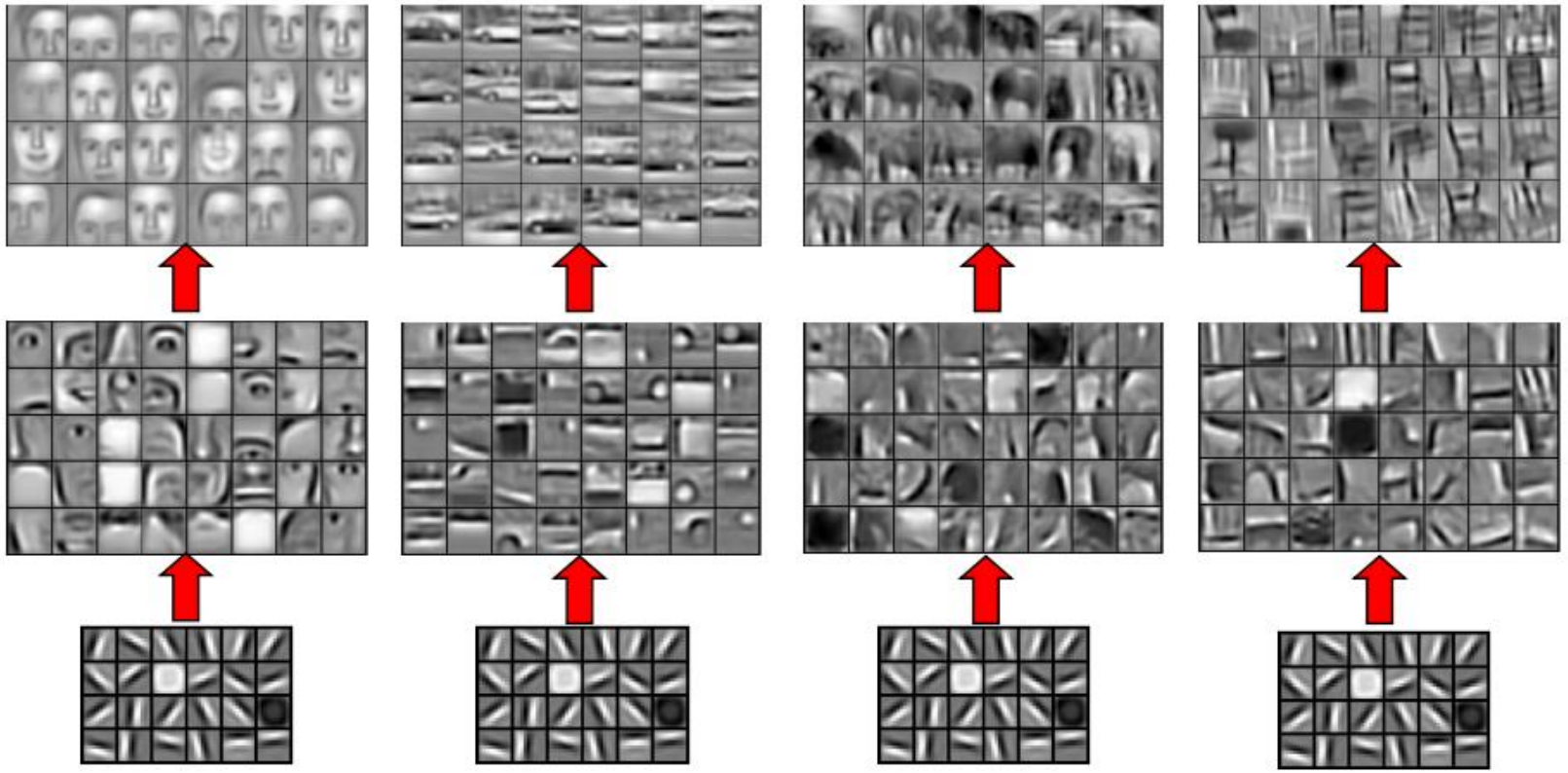
Deep Belief Net on Face Images



Based on materials by Andrew Ng

State of the art Applications of machine Learning

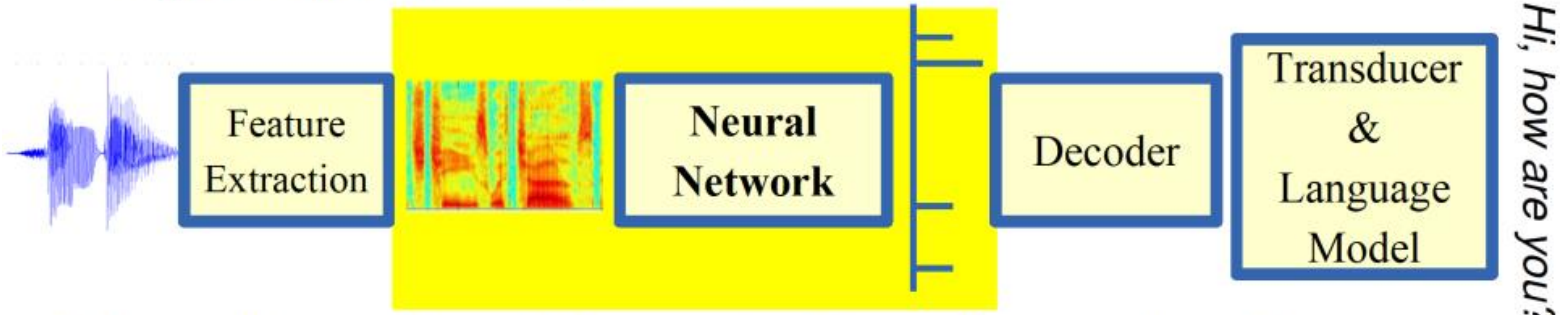
Learning of Object Parts



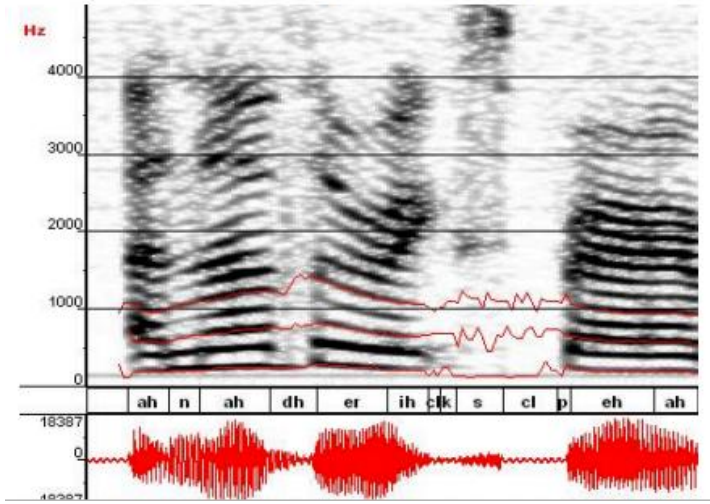
State of the art Applications of machine Learning

Machine Learning in Automatic Speech Recognition

A Typical Speech Recognition System



ML used to predict of phone states from the sound spectrogram



Deep learning has state-of-the-art results

# Hidden Layers	1	2	4	8	10	12
Word Error Rate %	16.0	12.8	11.4	10.9	11.0	11.1

Baseline GMM performance = 15.4%

[Zeiler et al. "On rectified linear units for speech recognition" ICASSP 2013]

History of Machine Learning Extension

18th Century

Development of statistical methods

Several key concepts in machine learning are derived from probability theory and statistics. The roots of these date back to the 18th Century. For example, in 1763 Thomas Bayes set out a mathematical theorem for probability – which came to be known as Bayes Theorem – that remains a central concept in some modern approaches to machine learning.

1950

1950 The Turing Test

Papers by Alan Turing through the 1940s grappled with the idea of machine intelligence. In 1950, he posed the question "can machines think?", and suggested a test for machine intelligence – subsequently known as the Turing Test – in which a machine might be called intelligent, if its responses to questions could convince a person that it was human.

1952 Machines that can play checkers

An early learning machine was created in 1952 by the researcher Arthur Samuel, which was able to learn to play checkers, using annotated guides by human experts and games it played against itself to learn to distinguish good moves from bad.

1960

1956 The Dartmouth Workshop

The birth of the term 'artificial intelligence' is generally credited to computer scientist John McCarthy, who, alongside key figures in the field including Marvin Minsky, Nathaniel Rochester, and Claude Shannon, brought together leading researchers at a workshop to consider the development of the field in 1956.

History of Machine Learning Extension

1970

1957 The Perceptron

Frank Rosenblatt's perceptron was an early attempt at creating a neural network, using a rotary resistor (potentiometer) driven by an electric motor. This machine could take an input – the pixels of an image, say – and create an output, such as a label.

1980

1973 The Lighthill report and the AI winter

By the 1970s, it was clear that progress in the field was not as fast as had been expected. A report commissioned by the UK Science Research Council – the Lighthill report – noted that "in no part of the field have the discoveries made so far produced the major impact that was then promised". This assessment, coupled with slow progress in the field, contributed to a loss of confidence and a drop in resources for AI research.

1986 *Parallel Distributed Processing* volumes and neural network models

In 1986, David Rumelhart, James McClelland, and the PDP Research Group published *Parallel Distributed Processing*, a two-volume set of work which advanced the use of neural network models for machine learning.

1990

History of Machine Learning Extension

1990

1992 Playing backgammon
In 1992, researcher Gerald Tesauro created an artificial neural network-based program, which was able to play backgammon, and which could match the abilities of top human players.

1997 Deep Blue beats the reigning world champion at chess
In 1997 Deep Blue became the first computer chess-playing system to beat a reigning world chess champion. Deep Blue exploited the increased computing power available in the 1990s to perform large-scale searches of potential moves – it could reportedly process over 200 million moves per second – then select the best one.

2000

2012 ImageNet Classification and advances in computer vision
2012 saw the publication of a highly influential paper by Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. This paper described a model that had been used to win an annual image recognition competition, and dramatically reduced the error rate in image recognition systems.

2010

2011 Watson beats the two human Jeopardy! champions
In 2011, IBM's Watson won a game of the US quiz show Jeopardy against two of its champions.

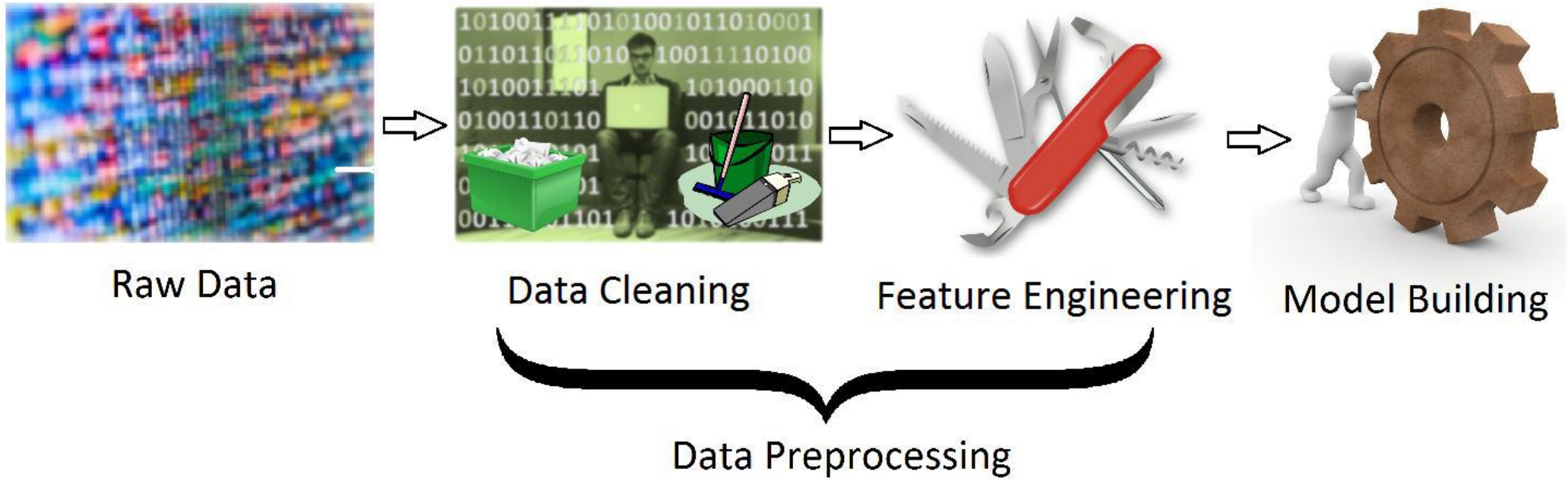
2017 Learning to play poker
Libratus, a system built by researchers at Carnegie Mellon University, defeated four top players at no-limit Texas Hold 'Em after 20 days of play in 2017. Researchers at the University of Alberta reported similar success with their system, Deepstack.

2016 AlphaGo beats the world champion at Go
In 2016, AlphaGo – a system created by researchers at Google DeepMind to play the ancient Chinese game of Go – won four out of five matches against Lee Sedol, who has been the world's top Go player for over a decade.

2020

Data in Machine Learning

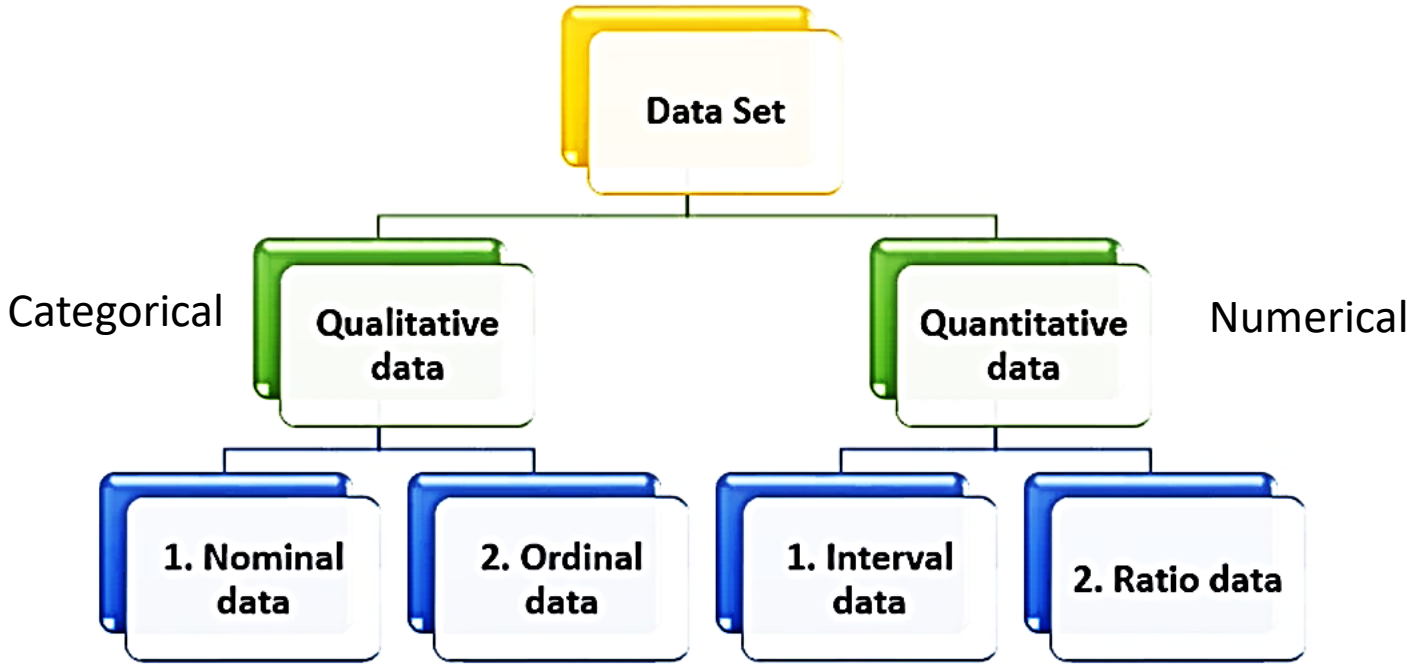
- **Machine learning depends largely on test data.**
- **A large amount of data is required for ML.**



Types of Data in ML

Basic Types Of Data In Machine Learning

- Data Set
- Qualitative data
- 1. Nominal data
- 2. Ordinal data
- Quantitative data
- 1. Interval data
- 2. Ratio data



Types of Data in ML

Data Set

- A data set is a collection of related information or records.
- The information may be on some entity or some subject area.
- Example-
- 1. A data set on students in which each record consists of information about a specific student.

Student data set:

Roll Number	Name	Gender	Age
129/011	Mihir Karmarkar	M	14
129/012	Geeta Iyer	F	15
129/013	Chanda Bose	F	14
129/014	Sreenu Subramanian	M	14
129/015	Pallav Gupta	M	16
129/016	Gajanan Sharma	M	15

(Attributes of the data set)

Types of Data in ML

Data Set...

- 2. Another data set on student performance which has records providing performance, i.e. marks on the individual subjects.

Student performance data set:

Roll Number	Maths	Science	Percentage
129/011	89	45	89.33%
129/012	89	47	90.67%
129/013	68	29	64.67%
129/014	83	38	80.67%
129/015	57	23	53.33%
129/016	78	35	75.33%

Types of Data in ML

Data Set...

- Each row of a data set is called a record.
- Each data set also has multiple attributes, each of which gives information on a specific characteristic.
- For example, in the data set on students, there are four attributes namely Roll Number, Name, Gender, and Age,
- Attributes can also be termed as feature, variable, dimension or field.
- Both the data sets, Student and Student Performance, are having four features or dimensions;
- hence they are told to have four dimensional data space.

Student data set:

Roll Number	Name	Gender	Age
129/011	Mihir Karmarkar	M	14
129/012	Geeta Iyer	F	15
129/013	Chanda Bose	F	14
129/014	Sreenu Subramanian	M	14
129/015	Pallav Gupta	M	16
129/016	Gajanan Sharma	M	15

Student performance data set:

Roll Number	Maths	Science	Percentage
129/011	89	45	89.33%
129/012	89	47	90.67%
129/013	68	29	64.67%
129/014	83	38	80.67%
129/015	57	23	53.33%
129/016	78	35	75.33%

Types of Data in ML

Data Set...

- each row has specific values for each of the four attributes or features.
- Value of an attribute, vary from record to record.
- For example, if we refer to the first two records in the Student data set, the value of attributes Name, Gender, and Age are different.

Roll Number	Name	Gender	Age
129/011	Mihir Karmarkar	M	14
129/012	Geeta Iyer	F	15

Types of Data in ML

Types of Data

- Data can broadly be divided into following two types:
- 1. Qualitative data
- 2. Quantitative data

Types of Data in ML

Qualitative data

- **Qualitative data** provides information about the quality of an object or information which cannot be measured.
- For example, if we consider the quality of performance of students in terms of 'Good', 'Average', and 'Poor',
- it falls under the category of qualitative data.
- Also, name or roll number of students are information that cannot be measured using some scale of measurement.
- Qualitative data is also called **categorical data**.
- Qualitative data - two types:
- 1. Nominal data
- 2. Ordinal data

Types of Data in ML

Qualitative data - Nominal data

- **Nominal data** is one which has no numeric value, but a named value.
- It is used for assigning named values to attributes.
- Nominal values cannot be quantified.
- Examples of nominal data are
- 1. Blood group: A, B, O, AB, etc.
- 2. Nationality: Indian, American, British, etc.
- 3. Gender: Male, Female, Other

4. Colour: Red, Blue etc.

Types of Data in ML

Qualitative data - Nominal data

- mathematical operations such as addition, subtraction, multiplication, etc. and statistical functions such as mean, variance, etc. cannot be performed on nominal data.
- a basic count is possible.
- The mode is possible, i.e. most frequently occurring value, can be identified for nominal data.

Types of Data in ML

Qualitative data - Ordinal data,

- **Ordinal data**, naturally ordered.
- This means ordinal data assigns named values to attributes
- They can be arranged in a sequence of increasing or decreasing value
- Hence comparison is possible here.
  - 1. Customer satisfaction: 'Very Happy', 'Happy', 'Unhappy', etc.
  - 2. Grades: A, B, C, etc.
  - 3. Hardness of Metal: 'Very Hard', 'Hard', 'Soft', etc.
- Like nominal data, basic counting is possible for ordinal data.
- Hence, the mode and median can be identified.
- But Mean can not be calculated.

The data is arranged in an order so we can find its median value.

Types of Data in ML

Quantitative data

- **Quantitative data** relates to information about the quantity of an object – hence it can be measured.
- For example, if we consider the attribute 'marks', it can be measured using a scale of measurement.
- Quantitative data is also termed as numeric data.
- There are two types of quantitative data:
- 1. Interval data
- 2. Ratio data

Types of Data in ML

Quantitative data - Interval data

- **Interval data** is numeric data – identify the order and difference between values
- Example - Celsius temperature - the difference between each value remains the same
- For example, the difference between 12°C and 18°C degrees is measurable and is 6°C
- Other examples include date, time, etc.
- For interval data, mathematical operations such as addition and subtraction are possible.
- For that reason, for interval data, the central tendency can be measured by mean, median, or mode.
- Standard deviation can also be calculated.

Types of Data in ML

Quantitative data - Interval data

- Interval data do not have a 'true zero' value.
- For example, there is nothing called '0 temperature' or 'no temperature'.
- Hence, only addition and subtraction applies for interval data.
- The ratio cannot be applied.
- This means, we can say a temperature of 40°C is equal to the temperature of 20°C + temperature of 20°C.
- However, we cannot say the temperature of 40°C means it is twice as hot as in temperature of 20°C.

Types of Data in ML

Quantitative data - Ratio data

- **Ratio data** represents numeric data for which exact value can be measured.
- Absolute zero is available for ratio data.
- Also, these variables can be added, subtracted, multiplied, or divided.
- The central tendency can be measured by mean, median, or mode and also standard deviation.
- Examples of ratio data include height, weight, age, salary, etc.

Types of Data in ML

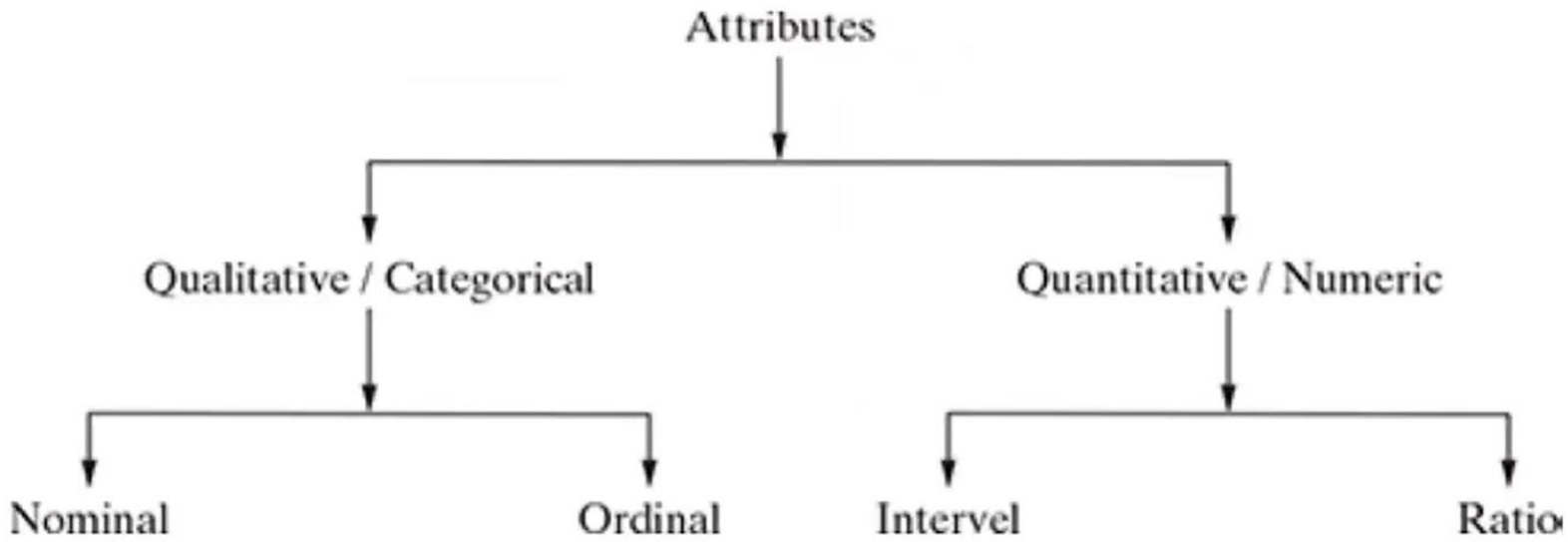
Quantitative data - Time-Series Data

Time-series data in machine learning falls under the category of **quantitative data**. This is because time-series data is numerical and can be measured or counted. It typically consists of real numbers that represent a certain variable over different points in time. This type of data is often used in forecasting models, trend analysis, and other statistical methods within machine learning.

Types of Data in ML

Data types based on Attributes.

- A summarized view of different types of data that we may find in a typical machine learning problem.



Data Attributes in ML

Data Attributes

- Attributes can also be categorized into types based on a number of values that can be assigned.
- The types of attributes based on factor.

1. Discrete Attributes
2. Nominal attributes
3. Numeric Attributes
4. Binary Attributes
5. Continuous Attributes

Data Attributes in ML

Data Attributes…

- Discrete attributes can assume a finite or countably infinite number of values.
- Nominal attributes such as roll number, street number, pin code, etc. can have a finite number of values
- Numeric attributes such as count, rank of students, etc. can have countably infinite values.
- binary attribute, a special type of discrete attribute which can assume two values only is called
- Examples of binary attribute include male/ female, positive/negative, yes/no, etc.
- Continuous attributes can assume any possible value which is a real number.
- Examples of continuous attribute include length, height, weight, price, etc.

Machine learning Activities

Machine Learning Activities

- The first step in machine learning activity starts with data.
- Supervised Machine Learning, has a labelled training data set followed by test data which is not labelled.
- Unsupervised Machine Learning, there is no labelled data (no training) but to find patterns in the input data.
- A thorough review and exploration of the data is needed to understand
  - **the type of the data,**
  - **the quality of the data and**
  - **relationship between the different data elements.**
- Based on that, multiple pre-processing activities to be done on the input data before the machine learning activities.

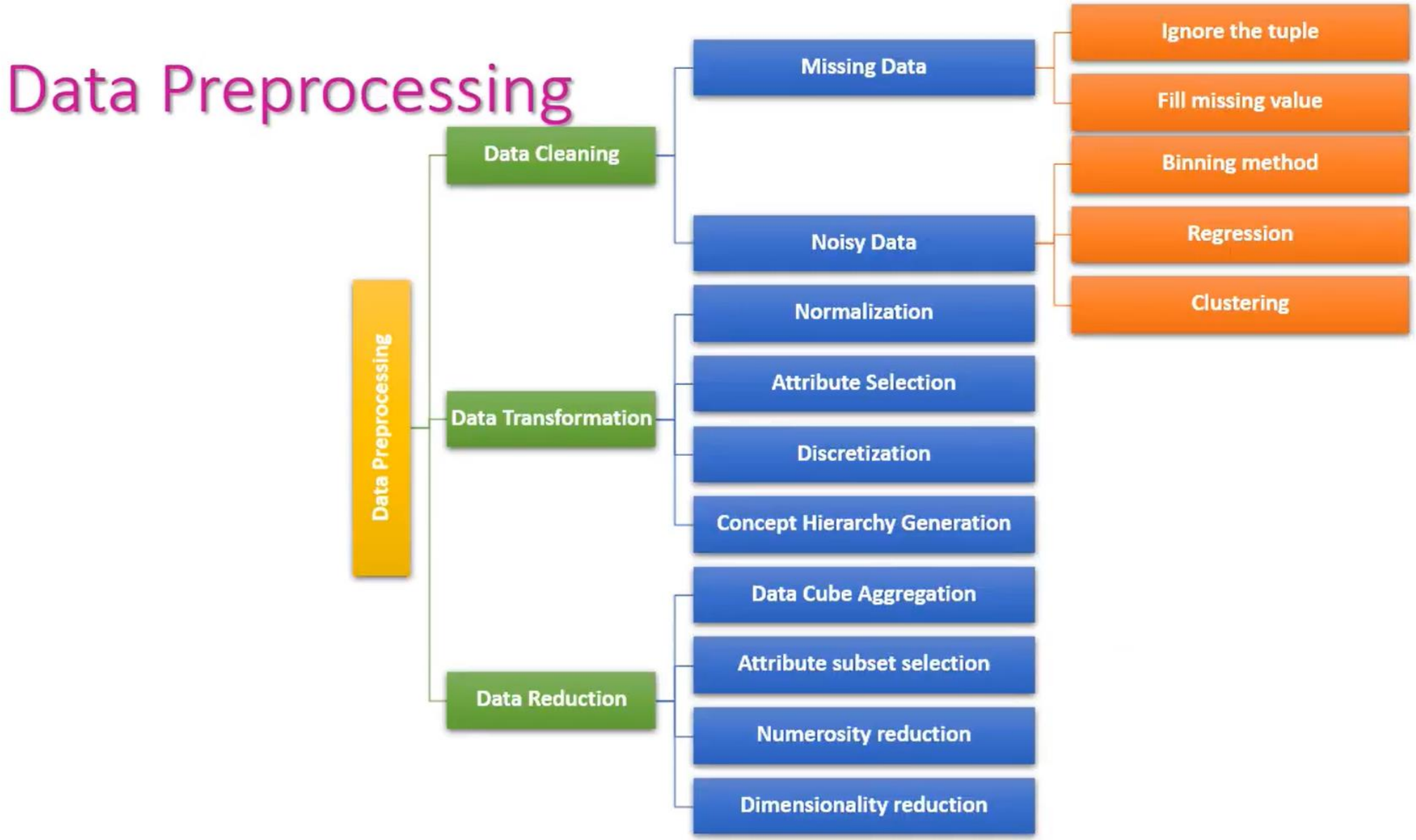
Machine learning Activities

Machine Learning Activities – Preparation activities

- Following are the typical **preparation** activities done once the input data comes into the machine learning system:

1. Understand the type of data in the given input data set.
2. Explore the data to understand the nature and quality.
3. Explore the relationships amongst the data elements, e.g. inter-feature relationship.
4. Find potential issues in data.
5. Do the necessary remediation, e.g. include missing data values, etc., if needed.
6. Apply pre-processing steps, as necessary.

Data Preprocessing in ML



Data Preprocessing in ML

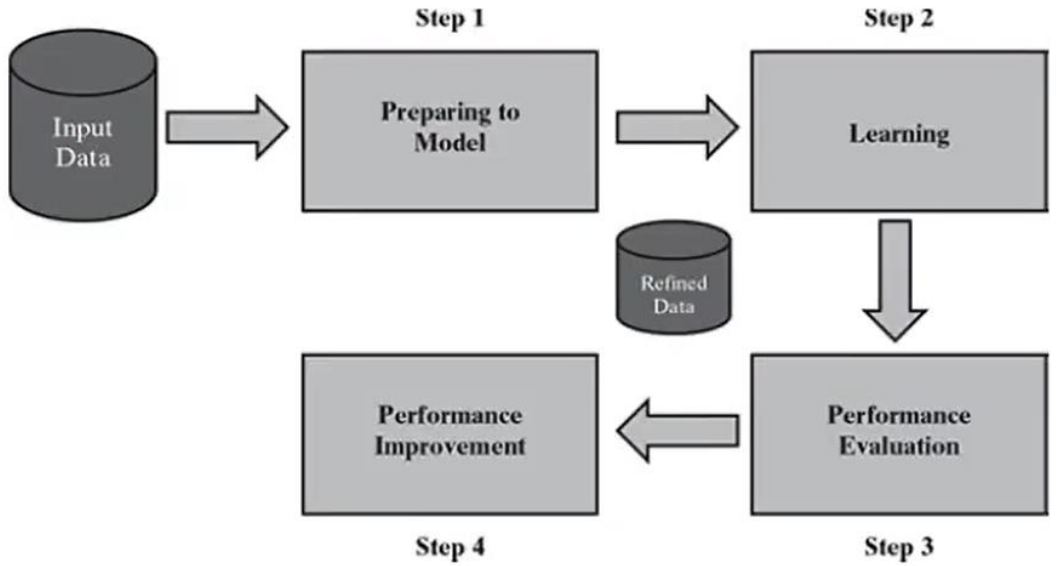
Machine Learning Activities – Preparation activities…

- Once the data is prepared for modelling, then the learning tasks starts with the following activities.
  1. The input data is first divided into two parts – the training data and the test data (called holdout). This step is applicable for supervised learning only.
  2. Consider different models or learning algorithms for selection.
  3. Train the model based on the training data for supervised learning problem and apply to unknown data.
  4. Directly apply the chosen unsupervised model on the input data for unsupervised learning problem.

Data Preprocessing in ML

Detailed process of Machine Learning

- After the model is selected,
- trained (for supervised learning), and applied on input data,
- the performance of the model is evaluated.
- Based on options available, specific actions can be taken to improve the performance of the model, if possible.



Data Preprocessing in ML

Summary of Steps and Activities Involved

Step #	Step Name	Activities Involved
Step 1	Preparing to Model	• Understand the type of data in the given input data set • Explore the data to understand data quality • Explore the relationships amongst the data elements, e.g. inter-feature relationship • Find potential issues in data • Remediate data, if needed • Apply following pre-processing steps, as necessary: ✓ Dimensionality reduction ✓ Feature subset selection
Step 2	Learning	• Data partitioning/holdout • Model selection • Cross-validation
Step 3	Performance evaluation	• Examine the model performance, e.g. confusion matrix in case of classification • Visualize performance trade-offs using ROC curves
Step 4	Performance improvement	• Tuning the model • Ensembling • Bagging • Boosting

What is data ?

Data are raw facts that have not been processed to explain their meaning.

There are three types of Data

1. Structured data
2. Semi-structured data
3. Unstructured data

What is structured data?

Structured data is data whose elements are addressable for effective analysis. It has been organized into a formatted repository that is typically a database. It concerns all data which can be stored in database SQL in a table with rows and columns. They have relational keys and can easily be mapped into pre-designed fields. Today, those data are most processed in the development and simplest way to manage information. *Example:* Relational data.

Stored in tabular format

Clearly defined

Data is stored in a pre-defined data model

FOR EXAMPLE

Excel files

SQL databases

The rows and columns are related to each other

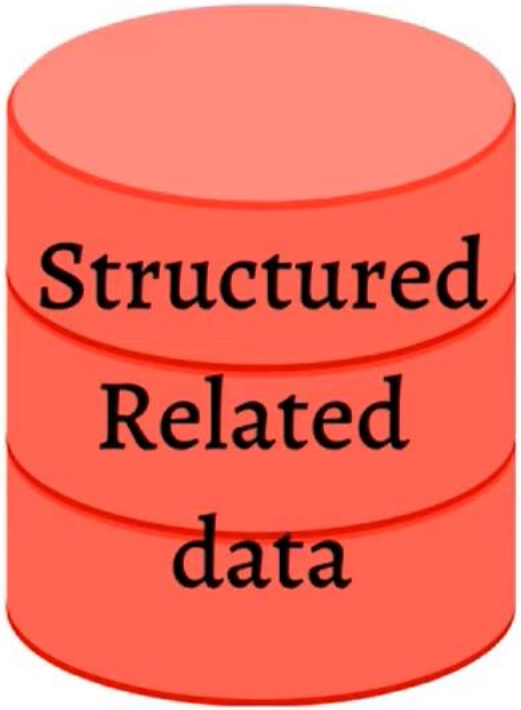
ID	NAME	ADDRESS	PHONE NO

Proper view and understanding of data

What is structured data?

AN EXAMPLE OF EMIRATES AIRLINES FROM DUBAI TO PARIS

Data



DATABASE

MANAGE

DATA IN THE RELATED FORMAT

DATABASE MANAGEMENT SYSTEM

RDBMS

What is Unstructured Data?

Unstructured data is a data which is not organized in a predefined manner or does not have a predefined data model, thus it is not a good fit for a mainstream relational database. So for Unstructured data, there are alternative platforms for storing and managing, it is increasingly prevalent in IT systems and is used by organizations in a variety of business intelligence and analytics applications. *Example*: Word, PDF, Text, Media logs.

What is Unstructured Data?

No pre-defined structure

No data model

Data is irregular and ambiguous

Easiest to extract data

Hence 80-90% of data is unstructured

It is a combination of text, numbers, audio, video, images, messages, social media posts, etc

What is Semi-structured data

Semi-structured data is a type of data that is not purely structured, but also not completely unstructured. It contains some level of organization or structure, but does not conform to a rigid schema or data model, and may contain elements that are not easily categorized or classified.

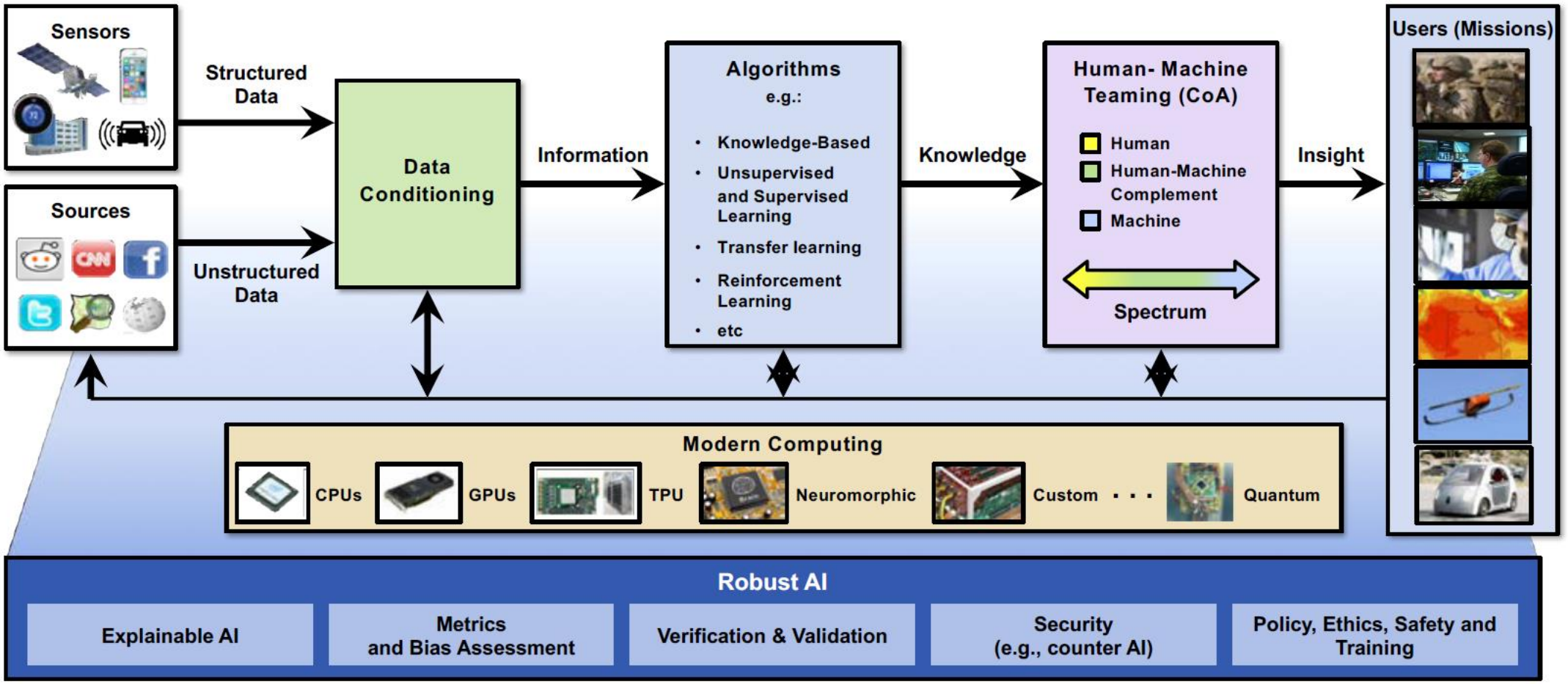
Advantages of Semi-structured Data:

- The data is not constrained by a fixed schema
- Flexible i.e Schema can be easily changed.
- Data is portable
- It is possible to view structured data as semi-structured data
- Its supports users who can not express their need in SQL
- It can deal easily with the heterogeneity of sources.
- Flexibility: Semi-structured data provides more flexibility in terms of data storage and management, as it can accommodate data that does not fit into a strict, predefined schema. This makes it easier to incorporate new types of data into an existing database or data processing pipeline.

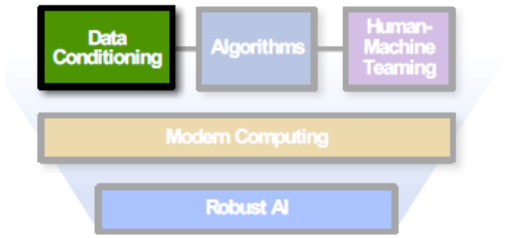
Comparison

Properties	Structured data	Semi-structured data	Unstructured data
Technology	It is based on Relational database table	It is based on XML/RDF(Resource Description Framework).	It is based on character and binary data
Transaction management	Matured transaction and various concurrency techniques	Transaction is adapted from DBMS not matured	No transaction management and no concurrency
Version management	Versioning over tuples,row,tables	Versioning over tuples or graph is possible	Versioned as a whole
Flexibility	It is schema dependent and less flexible	It is more flexible than structured data but less flexible than unstructured data	It is more flexible and there is absence of schema
Scalability	It is very difficult to scale DB schema	It’s scaling is simpler than structured data	It is more scalable.
Robustness	Very robust	New technology, not very spread	—
Query performance	Structured query allow complex joining	Queries over anonymous nodes are possible	Only textual queries are possible

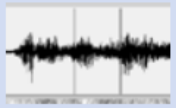
AI Canonical Architecture



Unstructured and Structured Data



Structured Data Types


Speech


Sensors


Network Logs

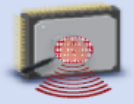

Metadata

Unstructured Data Types


Social Media

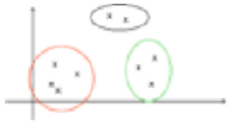
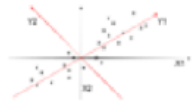

Human Behavior


Reports


Side Channel

Data Conditioning/Storage Technologies

- Data to Information -

Technologies	Capabilities Provided
Infrastructure/Databases  	• Indexing/Organization/Structure • Domain Specific Languages • High Performance Data Access • Declarative Interfaces
Data Curation  	• Unsupervised machine learning • Dimensionality Reduction • Clustering/Pattern Recognition • Outlier Detection
Data Labeling 	• Initial data exploration • Highlight missing or incomplete data • Reorient sensors/recapture data • Look for errors/biases in collection

Often takes up 80+% of overall AI/ML development work

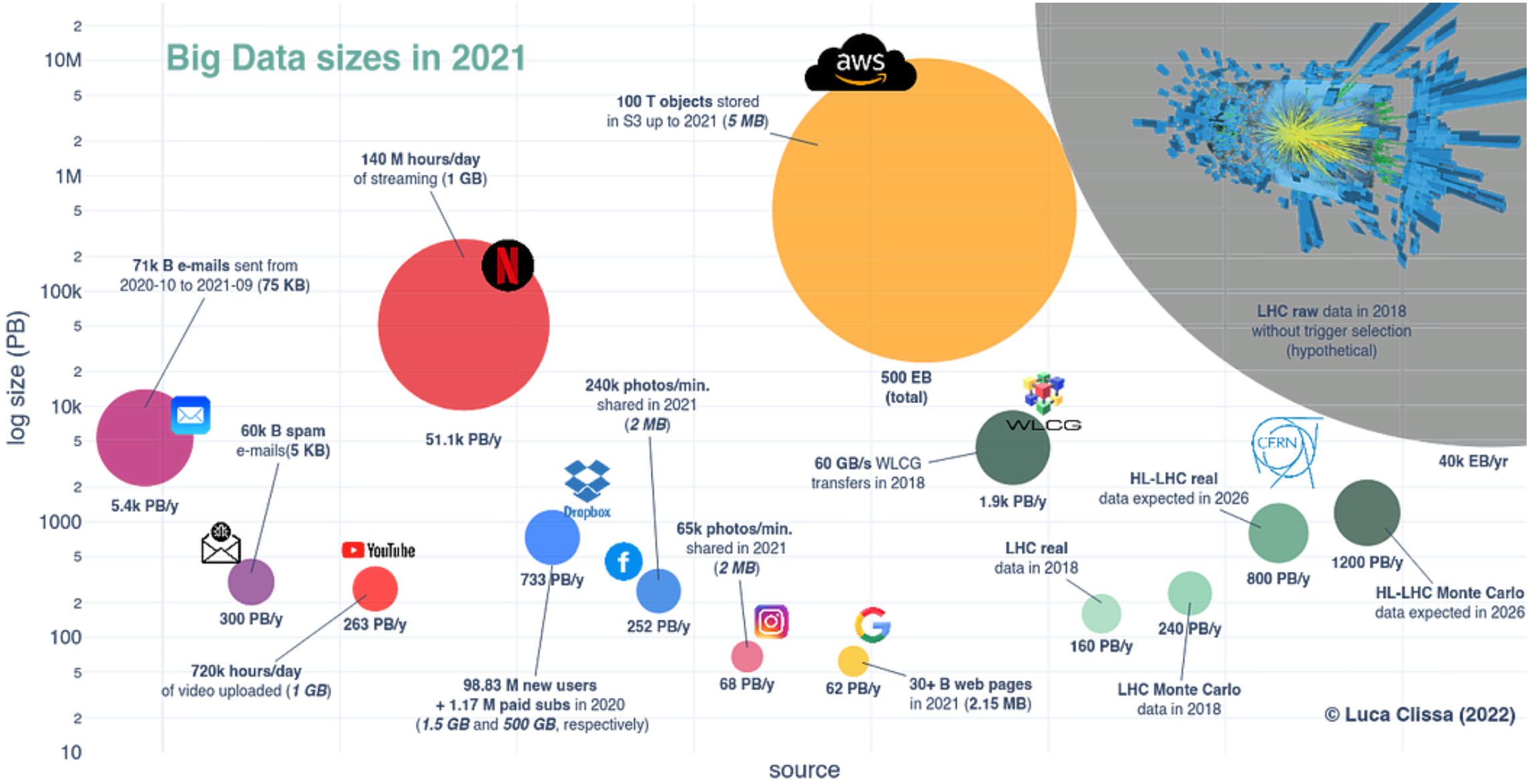
BIG DATA

Data which are very large in size is called Big Data. Normally we work on data of size MB(WordDoc ,Excel) or maximum GB(Movies, Codes) but data in Peta bytes i.e. 10^15 byte size is called Big Data. It is stated that almost 90% of today's data has been generated in the past 3 years.

Sources of Big Data

- **Social networking sites:** Facebook, Google, LinkedIn all these sites generates huge amount of data on a day to day basis as they have billions of users worldwide.
- **E-commerce site:** Sites like Amazon, Flipkart, Alibaba generates huge amount of logs from which users buying trends can be traced.
- **Weather Station:** All the weather station and satellite gives very huge data which are stored and manipulated to forecast weather.
- **Telecom company:** Telecom giants like Airtel, Vodafone study the user trends and accordingly publish their plans and for this they store the data of its million users.
- **Share Market:** Stock exchange across the world generates huge amount of data through its daily transaction.

BIG DATA


Big Data sizes in 2021
log size (PB)
source
10M
1M
100k
10k
1000
100
10
100 T objects stored in S3 up to 2021 (5 MB)
500 EB (total)
140 M hours/day of streaming (1 GB)
51.1k PB/y
71k B e-mails sent from 2020-10 to 2021-09 (75 KB)
5.4k PB/y
60k B spam e-mails(5 KB)
300 PB/y
720k hours/day of video uploaded (1 GB)
263 PB/y
240k photos/min. shared in 2021 (2 MB)
252 PB/y
65k photos/min. shared in 2021 (2 MB)
68 PB/y
98.83 M new users + 1.17 M paid subs in 2020 (1.5 GB and 500 GB, respectively)
733 PB/y
60 GB/s WLCG transfers in 2018
1.9k PB/y
62 PB/y
30+ B web pages in 2021 (2.15 MB)
LHC raw data in 2018 without trigger selection (hypothetical)
40k EB/yr
HL-LHC real data expected in 2026
800 PB/y
1200 PB/y
HL-LHC Monte Carlo data expected in 2026
LHC real data in 2018
160 PB/y
240 PB/y
LHC Monte Carlo data in 2018
© Luca Clissa (2022)

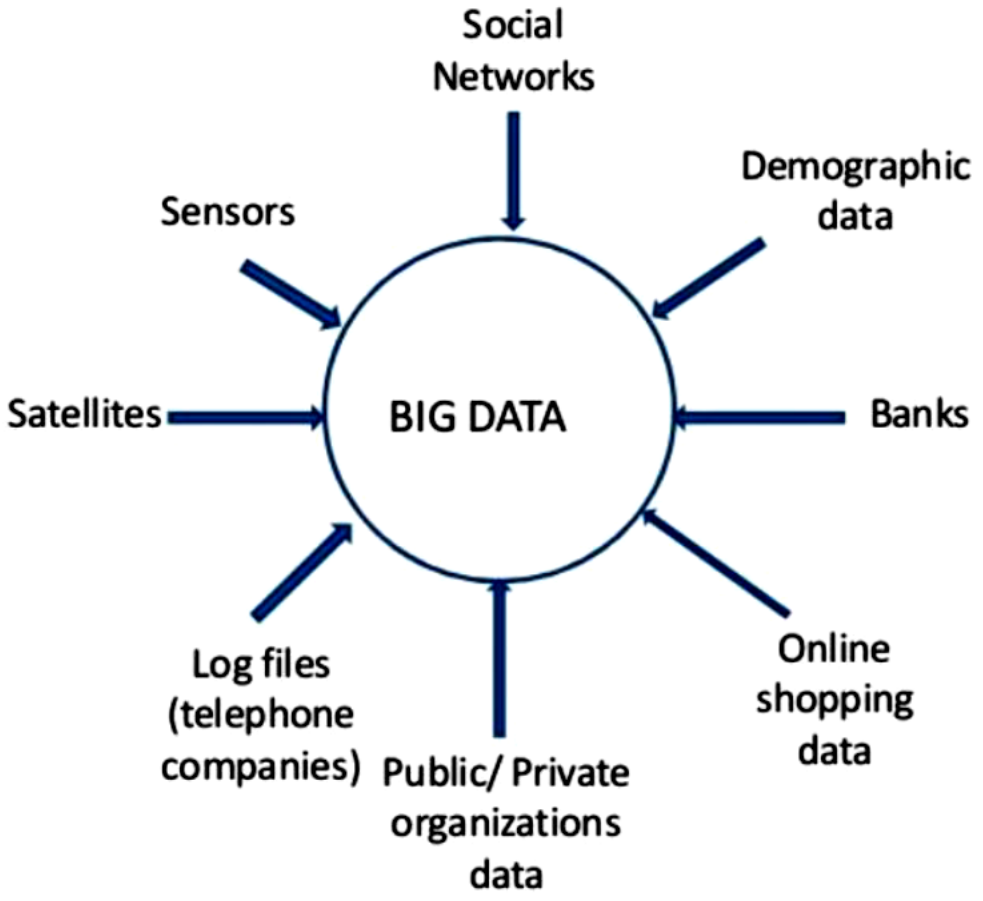
BIG DATA

BIG DATA

As the name implies, Big Data is huge data. In fact, very huge of the order of Petabyte (1PB=1000TB) or exabyte (1EB=1000PB)

Examples:

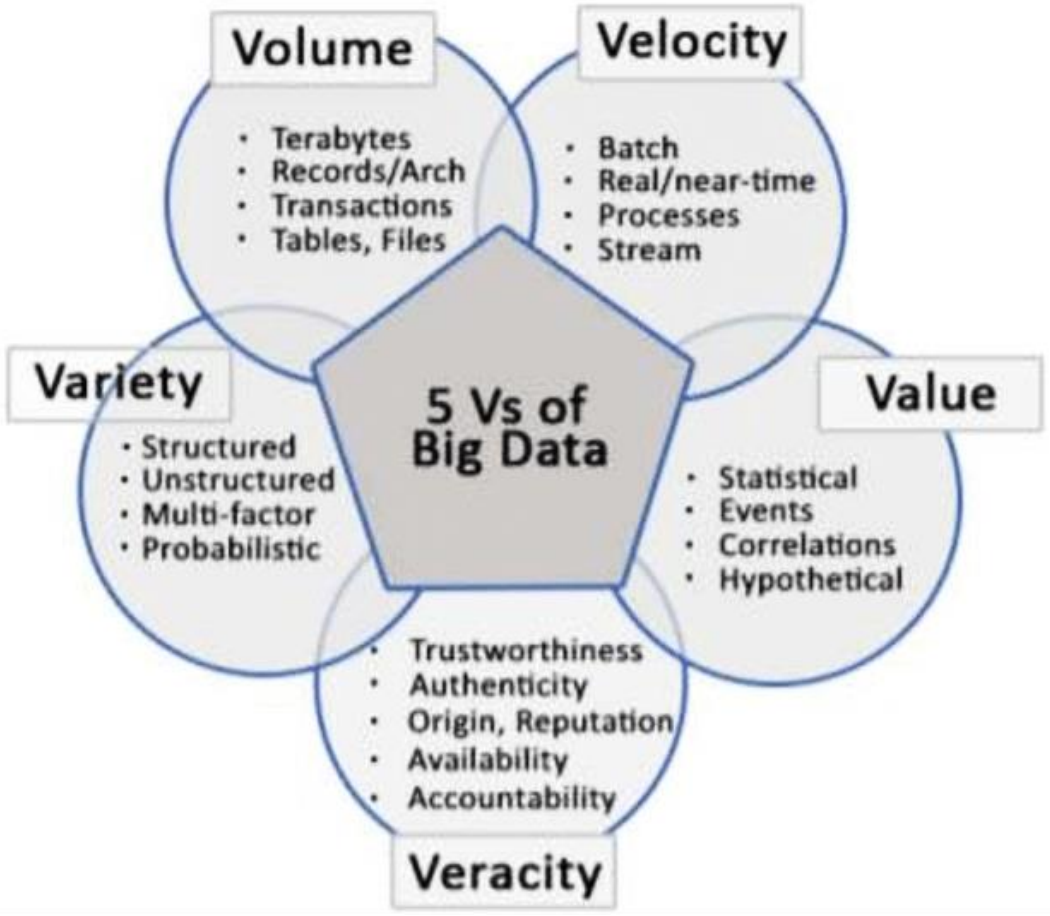
1. **Tweets from Twitter**-1000s/ second.
2. **Facebook**- 1000 of comments, 293,000 statuses, 1,36,000 photos uploaded/minute.
3. **Walmart, a departmental store chain**-one million customer transactions/ hour



Big data Sources

BIG DATA Characteristics

•**Volume:** the size and amounts of big data that companies manage and analyze
•**Value:** the most important "V" from the perspective of the business, the value of big data usually comes from insight discovery and pattern recognition that lead to more effective operations, stronger customer relationships and other clear and quantifiable business benefits
•**Variety:** the diversity and range of different data types, including unstructured data, semi-structured data and raw data
•**Velocity:** the speed at which companies receive, store and manage data – e.g., the specific number of social media posts or search queries received within a day, hour or other unit of time
•**Veracity:** the "truth" or accuracy of data and information assets, which often determines executive-level confidence
The additional characteristic of variability can also be considered:
•**Variability:** the changing nature of the data companies seek to capture, manage and analyze – e.g., in sentiment or text analytics, changes in the meaning of key words or phrases



BIG DATA

- Big data and machine learning are like a dynamic duo, each having its strengths.
- They're not rivals; instead, they work well together.
- When you use them together, amazing things can happen.
- Think of big data's 5Vs (volume, velocity, variety, veracity, and value) - machine learning helps handle them and make accurate predictions.
- On the flip side, when creating machine learning models, big data plays a role by giving top-notch data and enhancing learning methods through analytics.
- It's like they bring out the best in each other!

- There is no secret that almost all organizations, such as Google, Amazon, IBM, Netflix, etc., have already discovered the power of big data analytics enhanced by machine learning.

- Machine Learning is a very crucial technology, and with big data, it has become more powerful for data collection, data analysis, and data integration.
- All big organizations use machine learning algorithms for running their business properly.

BIG DATA

We can apply machine learning algorithms to every element of Big data operation, including:

- Data Labeling and Segmentation
- Data Analytics
- Scenario Simulation

In machine learning algorithms, we need multiple varieties of data for training a machine and predicting accurate results.
However, sometimes it becomes difficult to manage these bulkified data.
So, it becomes a challenge to manage and analyze Big Data.
Further, this unstructured data is useless until it is well interpreted.
Thus, to use information, there is a need for talent, algorithms, and computing infrastructure.
Machine Learning enables machines or systems to learn from past experience and use data received from big data, and predict accurate results.
Hence, this leads to generating improved quality business operations and building better customer relationship management.
Big Data helps machine learning by providing a variety of data so machines can learn more or multiple samples or training data.
In such ways, businesses can accomplish their dreams and get the benefit of big data using ML algorithms. However, for using the combination of ML and big data, companies need skilled data scientists.

How Does Big Data Analytics Work?

Companies need to work around analytics applications, partner with data scientists and engage with other data analysts to extract relevant and valid insights from big data. In addition, they must have an enhanced understanding of all available data. Finally, the analytics team also needs to clarify what they want to extract from the data.

The team needs to take care of :

- Cleansing,
- Profiling,
- Transformation,
- Validation of data sets.

These are some of the most important initial steps taken in data analysis.

Once all the big data has been prepared and gathered for interpretation, a combination of advanced data science and analytics disciplines is applied through different machine learning tools.

This will help to generate results that lead to businesses growth and development.

Leveraging Machine Learning

Leveraging Machine Learning:

•It involves understanding how to effectively use machine learning algorithms and techniques to solve real-world problems. It covers the application of machine learning in various domains, such as healthcare, finance, marketing, and more. How to choose appropriate machine learning models, preprocess data, train models, and evaluate their performance? Additionally, it may include discussions on the ethical considerations and challenges associated with leveraging machine learning in different industries.

Algorithms and Techniques:

1.Supervised Learning Algorithms:

1. **Regression:** Predicting a continuous outcome.
2. **Classification:** Assigning labels to data points (e.g., spam or not spam).

2.Unsupervised Learning Algorithms:

1. **Clustering:** Grouping similar data points together (e.g., customer segmentation).
2. **Dimensionality Reduction:** Reducing the number of features while preserving essential information.

3.Semi-Supervised and Self-Supervised Learning:

1. Learning from a combination of labeled and unlabeled data.
2. Learning from data without explicit labels, using the inherent structure of the data.

4.Reinforcement Learning:

1. Teaching models to make decisions by trial and error, receiving feedback in the form of rewards or penalties.

5.Neural Networks and Deep Learning:

1. Applying deep neural networks for complex tasks like image recognition, natural language processing, and speech recognition.

BIG DATA Real-World Problems

Real-World Problems:

1.Healthcare:

1. Predicting patient outcomes based on medical records.
2. Personalized treatment recommendations.

2.Finance:

1. Credit scoring and risk assessment.
2. Fraud detection in financial transactions.

3.Marketing:

1. Customer segmentation and targeted advertising.
2. Predicting customer churn and recommending retention strategies.

4.Manufacturing:

1. Predictive maintenance to reduce equipment downtime.
2. Quality control in production processes.

5.E-commerce:

1. Recommender systems for personalized product recommendations.
2. Fraud detection in online transactions.

BIG DATA

1.Autonomous Vehicles:

1. Object detection and recognition for safe navigation.
2. Decision-making algorithms for route planning.

2.Natural Language Processing (NLP):

1. Sentiment analysis for customer reviews.
2. Language translation and chatbot applications.

3.Image and Video Analysis:

1. Facial recognition for security.
2. Object detection in video surveillance.

4.Climate Prediction:

1. Predicting weather patterns and climate trends.
2. Analyzing environmental data for sustainable practices.

5.Education:

1. Personalized learning plans based on student performance.
2. Early detection of learning difficulties.

What is analytics?

• "Analytics in general, involves the use of mathematical or scientific methods to generate insight from data"

Key components of that definition:

1. Mathematical/Scientific Methods

2. Insight from data [5]:
   - Four V's
   1. Volume: 40 Zetabytes (43 Trillion GB) by 2020
   2. Variety: Healthcare, Video, Text, Social Media
   3. Velocity: Speed of available data (18.9Billion network connections by 2016)
   4. Veracity: Data quality and uncertainty

Data Analytics

Data analytics is the process of examining, cleaning, transforming, and modeling data with the goal of discovering useful information, drawing conclusions, and supporting decision-making. It involves the use of various techniques and tools to analyze patterns, trends, and relationships within datasets.

Here are key aspects of data analytics:

1. **Data Collection:**
   - •Gathering relevant data from various sources, which can include databases, spreadsheets, sensors, social media, and more.
2. **Data Cleaning:**
   - •Ensuring that the collected data is accurate, complete, and free from errors or inconsistencies.
3. **Data Transformation:**
   - •Converting raw data into a format suitable for analysis. This may involve normalization, aggregation, or other preprocessing steps.
4. **Data Analysis:**
   - •Applying statistical and mathematical methods, as well as using tools and algorithms, to uncover patterns, correlations, and trends in the data.
5. **Data Visualization:**
   - •Representing the results of data analysis through charts, graphs, and other visualizations to make complex information more understandable.
6. **Descriptive Analytics:**
   - •Summarizing and interpreting historical data to understand what has happened. It involves the examination of past events and their characteristics.

BIG DATA

7.Predictive Analytics:

- •Using statistical algorithms and machine learning techniques to make predictions about future events based on historical data.

8.Prescriptive Analytics:

- •Recommending actions to optimize outcomes based on the insights gained from data analysis.

9.Business Intelligence:

- •Leveraging data to support business decision-making by providing actionable insights and relevant information.

10.Big Data Analytics:

- •Analyzing large and complex datasets, often characterized by the three Vs (volume, velocity, and variety), using specialized tools and technologies.

Data analytics plays a crucial role in various fields, including business, healthcare, finance, marketing, and science. It helps organizations make informed decisions, identify opportunities, and address challenges by extracting meaningful insights from data. The growing importance of data analytics has led to the development of specialized roles and tools to meet the increasing demand for skilled professionals in this field.

BIG DATA

1. **Descriptive analytics:**

   Here the information that is present in the data is obtained and summarized.
   It is primarily involved in **finding all the statistics** that describes the data.

Eg: How many buyers bought A.C. in the month of December previous years?

BIG DATA

2. Diagnostic/ Discovery Analytics:

This stage involves finding out the reason for the statistics determined in the previous analytics stage.

Otherwise it involves, **why that statistics** have happened?

Eg: Why there is an increase/ decrease in the sales of A.C.in the month of December?

BIG DATA

4. **Prescriptive Analytics:**

It involves **planning actions or making decisions** to improve the Business based on the predictive analytics .

Eg: **How much amount of material** should be procured to increase the production?

BIG DATA

Descriptive vs Predictive Analytics:

•Descriptive analytics involves analyzing historical data to understand what has happened in the past. It focuses on summarizing and interpreting data to gain insights into trends and patterns. This type of analytics is useful for reporting and data visualization.

•Predictive analytics, on the other hand, aims to forecast future outcomes based on historical data and statistical algorithms. It involves the use of machine learning models to make predictions and identify potential trends. Predictive analytics is valuable for making informed decisions and planning for the future.

Types of Big Data Technologies

Big Data Technology is mainly classified into two types:

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

BIG DATA

TICKET BOOKING

ONLINE SHOPING

SOCIAL MEDIA

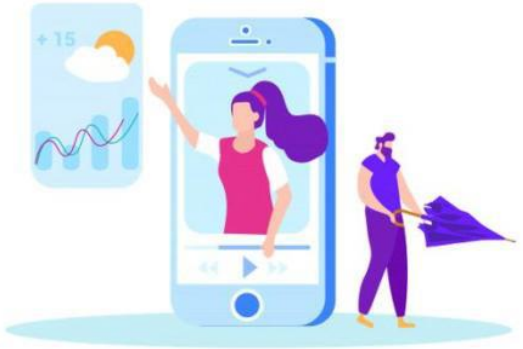
ORGANISATION

BIG DATA

STOCK MARKET

SPACE MISSION

WEATHER FORECAST



MEDICAL

BIG DATA

EVOLUTION OF DATA ANALYTICS

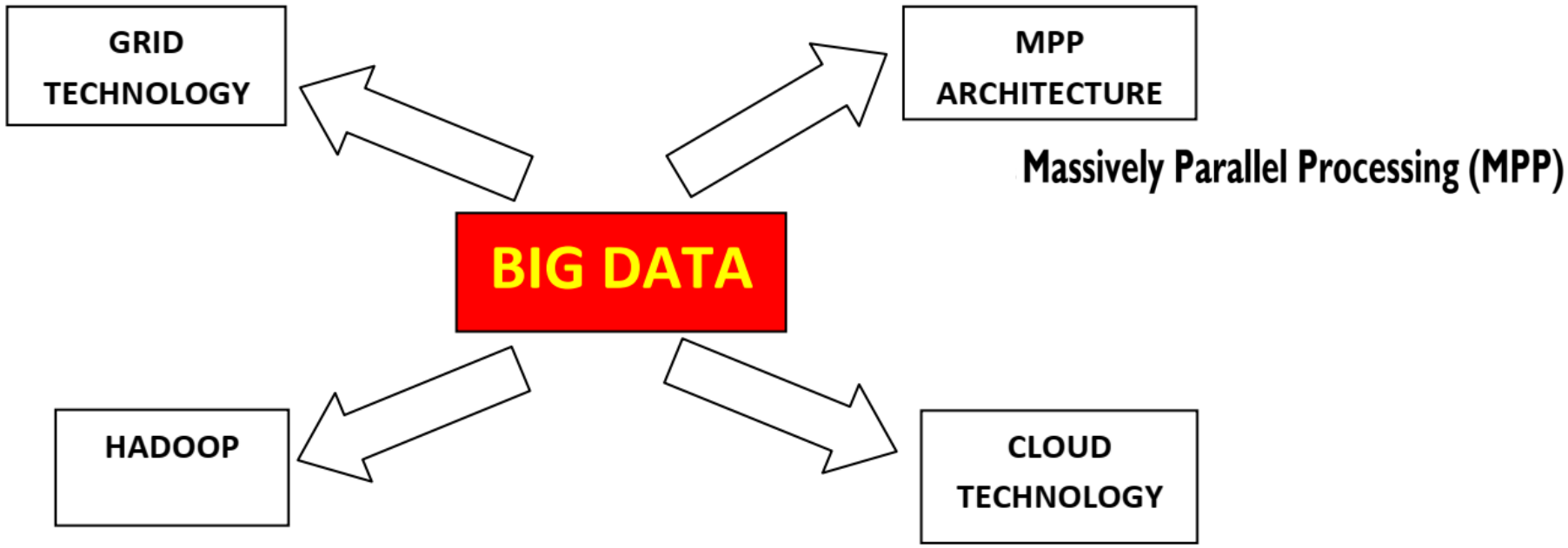


Figure 1. Big data technologies

BIG DATA

Modern Analytical tools