Introduction to Data Analytics

Dr. Cahit Karakuş

Dr. Cahit Karakuş

- After graduating from Istanbul Technical University, Department of Electronics and Communication Engineering in 1984, I worked as a research assistant at the same university for 4 years. In my master's and doctoral theses, I have studied on microwave image processing and data analysis.
- I wanted to leave university and work in the private sector. I worked as a software and hardware engineer in the test engineering department of Nortel for 2 years.
- After completing my military service, I worked in the research and development department of Nortel for 9 years starting from 1991. I took part in the design of computer moduls in digital communication systems. Therefore, I know computer organization and microprocessor software languages well.
- Also, I have designed, implemented and tested directional microstrip antennas used in wireless communication systems. I have patents on over 10 antennas. Therefore, I am an expert in microwave detection and radar systems and electromagnetic weapons.
- In the 2000s, I left Nortel and worked as an expert engineer in a foreign company that installed wireless communication systems and satellite ground stations in various cities in Turkey and other countries.
- Between 2008 and 2014, I have provided expert consultancy services to senior management at Türk Telekom and own GSM company.
- I have been teaching at Esenyurt University, Department of Computer and Software Engineering since 2015. The courses I teach are: Digital Communication and Network Technologies, Computer Organization, Microprocessor, Quantum Computing, Data Analytics, Probability in Data Analytics, Signals and Systems, Artificial Intelligence, Machine Learning.
- I have also been teaching Master's Degree, Introduction to Data Analytics and Probability Theory in Data Analytics courses at Altınbaş University for 3 years.

The rules that will be applied during the midterm and final exams are given below

1- Attendance is mandatory.

2- Each student was given a project as a midterm exam. The student will do research on the subject; will make a two-page summary of the subject and will submit it to the instructor in both document and e-mail format within 2 weeks.

3- As a final exam, the student will research an application on the project he/she took in the midterm, prepare a presentation and explain it in class.

4- Two students will explain each week in the final exam explanations.

5- Presentations will start on December 9, 2024.

6- In project explanations, each student who does not come to the project explanation will receive -5 points. 3 students will ask questions in each presentation.

7- In the evaluation, Attendance: 15%; Presentation: 25%, Explanation: 45%; Answering questions: 15% will be evaluated.

Introduction

- Data is a central phenomenon in our digital information society. Data effects our production and economic systems and offers enormous potential to positively influence our behavior and environment in society, business and science.
- The main goals of the Intelligence Data Analysis research group are to automatically extract and access information in data-intensive environments, create economic value and new business opportunities, and develop new machine learning methods for smart data analysis.
- Intelligent data analytics methods are a central component in obtaining usable insights into complex data sets and their interrelated processes. Therefore, intelligent data analytics methods should be developed to obtain accurate information from data sets and develop a successful application. A user-centered data analytics process where data can be analyzed together with users and with the help of appropriate algorithms is a step in the right direction.
- Data-centric analytics solutions are designed and developed to analyze both large data sets and real-time data streams. The determining point here is not only the data, but also the user's needs and the capabilities of the technologies. Using the latest methods in machine learning, big data and data science, it is possible to develop effective analytics solutions for complex problems.

History

Awareness and Consciousness?

- I would like to start my lesson with the topic of awareness and consciousness raising.
- Consciousness is the continuity of the transfer of the functions of questioning, comparing, wondering and researching from generation to generation.
- Consciousness is perceiving the changes, visualizing them in the brain, making sense of them and waking up purified.
- Consciousness and awareness initiate change.
- Consciousness is the continuity of the skills and experiences gained through thinking.
- Awareness is a process that aims to raise people's consciousness in order to achieve a certain goal or objective.

Consciousness process that will create awareness

Awareness-raising processes:

- Wheel; horse-drawn carriage
- Bridges
- The turning power of wind and water: Sailboats, Propellers, Cogs, Mills, Clock towers
- Priests who learned to be self-sufficient…
- The turning power of water vapor (coal, iron, steel): Ships, Trains, Textile Machines
- Generation and transmission of electricity – Faraday
  - Lamp - Edison
  - Asynchronousmotor and wireless transmission – Tesla
- Message – Morse
- Telephone - Graham Bell

Awareness-raising processes:

- Atom, Subatomic particles – Quantum
- Transistor, Computer – Alan Turing
- Fiber cable
- Smart phones
- Autonomous mobile machines
- Wireless Network Solution
- The power of information: Quantum computing,
- Artificial Intelligence
- 5G – 6G Mobile Telecommunication

Deep Learning & AI in Context of Human History

Perspective:

- Universe created
  13.8 billion years ago
- Earth created
  4.54 billion years ago
- Modern humans
  300,000 years ago
- Civilization
  12,000 years ago
- Written record
  5,000 years ago

1700s and beyond: Industrial revolution, steam engine, mechanized factory systems, machine tools

Those who raise awareness at crossroads

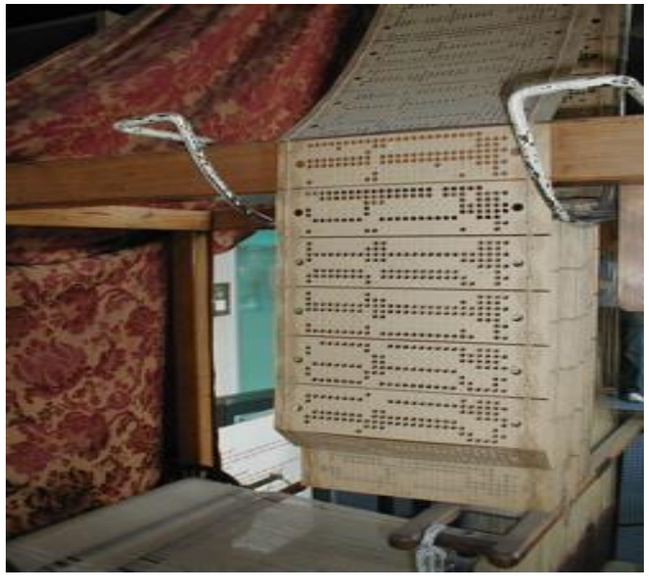
- Sumerians – Invention of writing: Symbols were processed into tablets.
- EL – Harazmi (780 – 850): Number system and algorithms
- Eb-Ül-İz El Cezeri (1136-1206): Mechanical robots were produced.
- Joseph Marie Jacquard (1752 – 1834): Mechanical computer and memory were produced.
- Michael Faraday (1791-1867): Electrical signal was discovered.
- Samuel Morse (1791 – 1872): Symbols were converted into electrical signals
- Alexander Graham Bell (1847 – 1922): Sound was converted into electrical signals.
- Nikola Tesla (1856, 1943, New York): Wireless signal was transmitted.
- Alan Turing (1912 – 1954): Computer operating logic
- John von Neumann (1903 – 1957): Computer was developed.
- John Forbes Nash (1928 – 2015): Mind Games
- Claude Shannon (1916 – 2001): Digitized information has become a physical quantity. Bit: 0/1. Symbols such as sound, text, and images are represented by bits, and bits are represented as a physical quantity by electrical signals.

Eb-Ül-İz El Cezeri's autonomous machines

- Eb-ul-iz El-Cezire, who made many technical and mechanical inventions such as water clocks, water robots, automatic thermos, was born in Cizre in 1136.
- Ebu'l İz El Cezeri (1136-1206), who worked on today's cybernetic and robot technologies in terms of world science history, presented these works in his work Kitab-ül Cami Beyn'el İlmi ve el Ameli'en Nafi fi Sinati'l Hiyel (Book Containing the Utilization of Mechanical Movements in Engineering) which he wrote for the Sultan of Artukoğulları.
- Although the original of Cezeri's book has not survived to the present day, ten copies of it are kept in different museums in Europe and five copies are kept in the Topkapı Palace and Süleymaniye libraries. The work known as Kitab-ül Hiyel consists of 6 sections.
- It is no coincidence that many scholars, especially Eb-ül-İz, were raised in Cizre during that period.
- Cizre during that period; It is a city that hosts different cultures and where scientific research is carried out along with religious sciences.

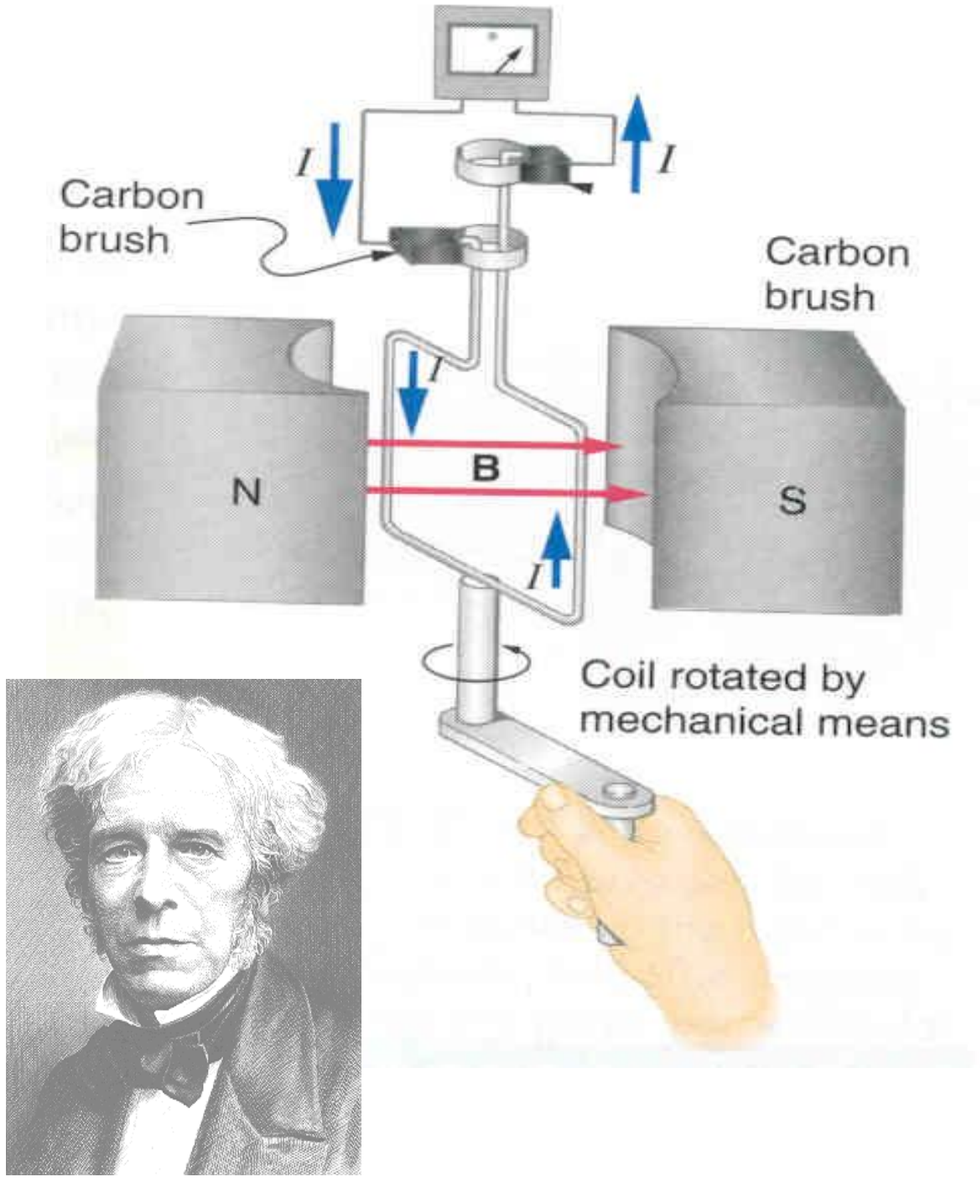
Joseph Marie Jacquard (1752 – 1834) Looms

- In the 18th century, the world's best silk weaving in Lyon had become a sectoral power. There were thousands of looms.
- Silk weaving, which included decorations and reliefs, involved very difficult work processes and was incredibly slow.
- In 1804, Joseph Marie Jacquard designed a device that was a miracle of creativity and created patterns and symbols with a very complex mechanism in silk weaving.
- Pictures, reliefs, and symbols were converted into punched cards as information. These looms were miracles of creativity. These punched cards decided which of the many threads would be used in which pattern, when and for how long.
- Joseph Jacquard, a French engineer, developed a weaving machine that performed the silk weaving process with punched cards in 1804. In fact, he produced a code stored in memories from the patterns and symbols in the binary number system and developed a computer system that weaved fabric according to this code.
- Spoken language could be symbolized with binary language. This was a very deep and forward-looking idea. Information could be converted into abstract symbols, stored and processed. Thus, the power of information was revealed. Information was transferred to punched cards. Symbols and patterns were converted into 0s and 1s and patterned fabrics were woven very quickly. These machines were the first computer-controlled machines that processed the first software codes. No electricity!!!



Discovery of the electrical signal: Michael Faraday (September 22, 1791 – August 25, 1867)

- In September of 1831, Michael Faraday (September 22, 1791 - August 25, 1867) made the discovery of Electromagnetic Induction.
- Cathode rays were discovered by Michael Faraday (1791-1867).
- In October of 1831, Faraday connected two wires to a disk and rotated the disk between opposite poles of a horseshoe magnet, creating an electric current flowing through the wire.
- Faraday discovered the electric current.



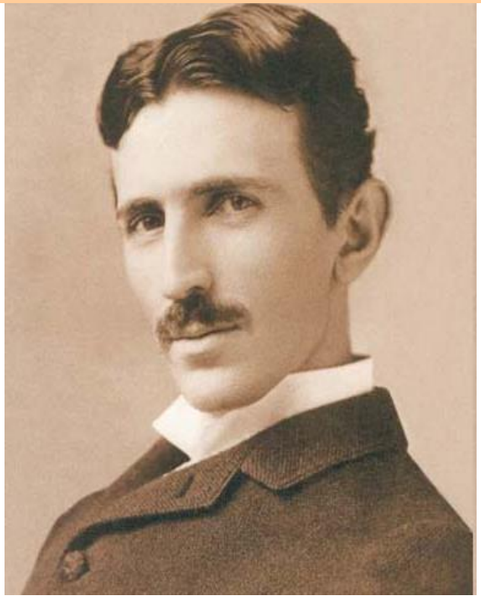
How will messages be transmitted via electrical signals over conductive wires?

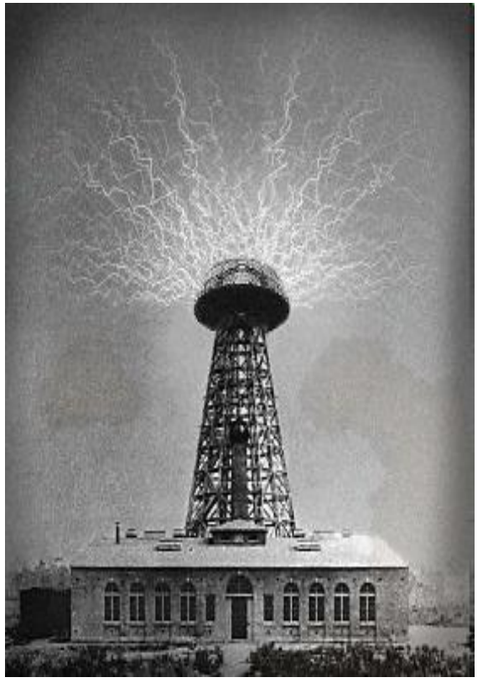
Samuel Morse (1791 – 1872)

- In the 19th century, there was an incredible development in the speed of information transfer. This development was the discovery of electricity.
- How could complex symbols be sent electrically with a simple signal?
- In 1840, the device developed by Samuel Morse (1791 - 1872) and his friend Alfred Vail used electric currents consisting of short and long pulses, showing and transmitting letters of the alphabet with dots and dashes called Morse code (Binary state).
- The telegraph showed that information could be converted and transmitted from one medium to another. The information in the human brain was shown with simple symbols. In the system called telegraph, information was converted into symbols.
- Information was combined with electricity. Information became an electrical signal. The telegraph network spread all over the world and the foundations of the modern information age were laid. Information could be transmitted very quickly to all parts of the world via cables.

Message could be transmitted via electromagnetic waves? Nikola Tesla (1856, 1943)

- Nikola Tesla (1856, 1943, New York). He was a Serbian inventor, electrical and mechanical engineer. He was the first inventor of alternating current systems.
- Tesla invented the AC Electric Motor.
- Signals were transmitted over electric wires. Can signals be transmitted through the air with electromagnetic waves now?
- Tesla's biggest invention: The Wardenclyffe Tower (1901–1917), also known as the Tesla Tower, was an early wireless telecommunications tower designed by Nikola Tesla to demonstrate commercial transatlantic wireless telephony, broadcasting, and power transmission without connecting cables. The main facility was not completed due to financial problems and was never fully operational.
- Marconi (1874 – 1937) made radio a commercial success by using and modifying the work of physicists and researchers before him, especially Tesla. He transmitted signals between the USA and Europe with the device he developed.





What happens in the human mind when it makes calculations?

Alan Turing (1912 – 1954)

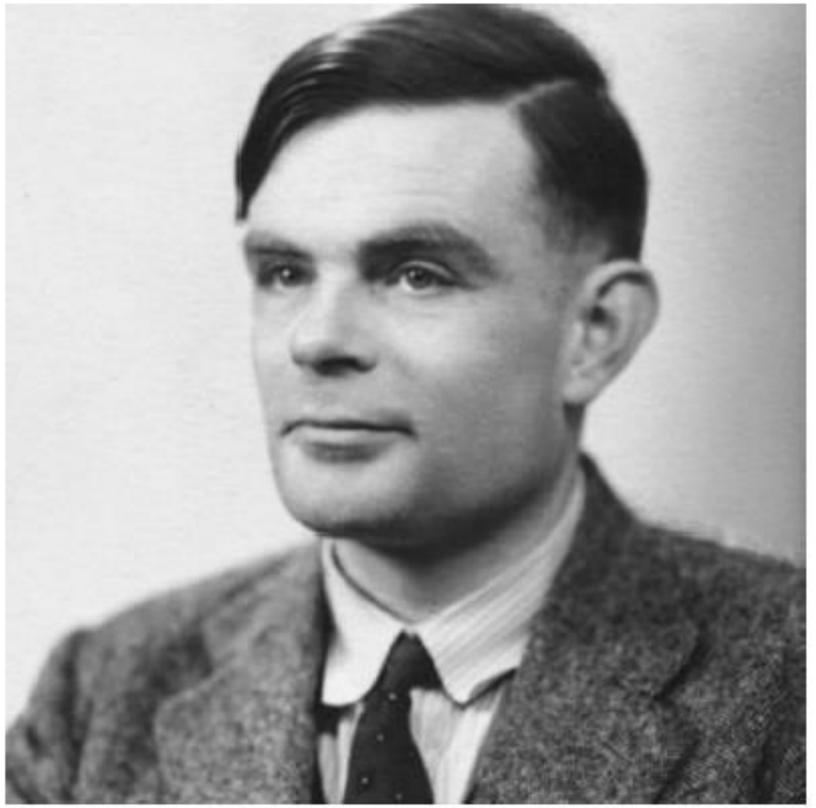
Alan Turing was the first person to create the mathematical basis of the computer. Turing was actually thinking about solving a mathematical problem. What would happen if problems in mathematics were solved by following a simple set of rules? This made him think about computers. Something unexpected happened and the computer emerged.

Turing's great idea was first published in the now legendary 36-page book "Application of Decision Problems in Computable Numbers", which he wrote in 1936 when he was 24 years old. Initially, he was interested in very abstract mathematics rather than practical calculation.

Turing asked a question:

- What happens in the mind of a person who is doing a calculation?
- What is vital for the person doing the calculation?
- What is the key function in the human brain in the calculation process?

Does Turing have the idea of a machine that processes and manipulates information? He noticed that certain rules are repeated in the human mind in calculation processes. He saw that all calculations are in binary dimension (0/1).



Dreams, mathematical foundations, and engineering in reality.

Alan Turing, 1951: “It seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers. They would be able to converse with each other to sharpen their wits. At some stage therefore, we should have to expect the machines to take control."

MIT Biomechatronics - Hugh Herr

- At 17, he became one of America's best climbers. However, while climbing a mountain, his legs were amputated because he was caught in a storm and froze. Now he has immortal legs! "It is the ability, not the disability, that counts", Andrew Carnegie.
- Hugh Herr develops bionic limbs that mimic the function of natural limbs.
- Herr is responsible for groundbreaking advances in bionic limbs that provide greater mobility and new hope to people with physical disabilities.
- Herr's team developed the first autonomous exoskeleton to reduce the metabolic cost of human walking. Herr's Biomechatronics group has developed knee prostheses and variable impedance ankle-foot orthoses for patients with foot drop, a gait pathology caused by stroke, cerebral palsy and multiple sclerosis, which are compatible with walking for transfemoral amputees.
- He also designed his own bionic limbs, the world's first bionic lower leg, called the BiOM Ankle System. As published in the 2012 Proceedings of the Royal Society, the BiOM Ankle System is the first leg prosthesis that provides biomechanical and physiological normalization, and has been clinically demonstrated to allow people with leg amputations to walk at normal speeds and metabolic levels, just like their legs. Biomechanics are different from ordinary prosthetics. The technological synthetic skins are connected to the main body, understand what is being done and move accordingly. In fact, they move so well that in trials with people without disabilities, it has been revealed that the support units work better than your biological limbs. Can you imagine, with the studies done, biomechanical limbs that are sensitive enough to make a dancer dance again can be created.

Can information be expressed as physically?

Claude Shannon (1916 – 2001)

- Claude Shannon (1916 – 2001) wrote a book called “Mathematical Theory of Communication” in 1948 and it is one of the most important scientific booklets of the twentieth century.
- Shannon found a way to measure and evaluate the amount of information in a message.
- He realized that the content of the information in a message was not related to its meaning. He had to give a unit of measurement to the information. He showed that a message to be transmitted could be measured when converted to binary number system.
- The message was a long series of ones and zeros. He realized that converting information to binary number system was a very powerful movement. Bit: 0/1 was defined. Bit is the smallest amount of information in the digital world.
- Information was converted to a measurable power, reality.

Mind Games

John Forbes Nash (1928 – 2015)

- Nobel Prize-winning American mathematician John Nash
- In 1959, Nash began to show obvious signs of mental disorder and spent several years in mental hospitals with a diagnosis of paranoid schizophrenia. After 1970, his condition began to improve, albeit slowly, and in the mid-1980s he was able to return to his academic career.

- On May 23, 2015, Nash and his wife Alicia Nash lost their lives in a traffic accident on a toll road while traveling by taxi.

- He solved the problems in game theory, invented by John von Neumann, and made them usable.

- Nash had developed emotionless, numerical formulas for the relationships that needed to be established between people due to interests and decisions that needed to be made.

- Nash Equilibrium: He developed a game theory in which each individual could develop a strategy against the best strategies of other individuals and in the end everyone would win. Each individual should try to do the best he could by seeing the actions of all the other individuals he was in a relationship with. The individual's choice depended on the choices of the other individuals.

Mobile Autonomous Machines

Industrial Revolutions

In order to notice change, it is necessary to be conscious. Inventions initiate changes and then transformations. While the power to continuously turn the wheel was sought, the steps of the industrial revolutions that developed after the invention of the steam engine are listed historically;

- The first industrial revolution began with the invention of the steam engine at the end of the 18th century and the first half of the 19th century. By 1781, Watt had developed his machine thoroughly, increased its performance and had also invented mechanical devices that skillfully converted the reciprocating motion of the piston into the rotational motion of a wheel. The basis of the Industrial Revolution was the use of wheels that started to rotate with steam in ships, trains and industry.
- In October 1831, Michael discovered electric current. Then, the second industrial revolution began with the discovery of electric and gasoline engines. Nikola Tesla (1856, 1943, New York). He was a Serbian inventor, electrical and mechanical engineer. He invented the electric motor that works with alternating current.
- French scientist Nicolas Leonard Sadi Carnot (1796 - 1832) discovered that heat engines could be operated by using the heat flow between heat and cold instead of steam engines. The pressure created by an explosion as a result of the combustion of a mixture of gasoline and air with a spark in the cylinder is converted into kinetic energy by the piston. Nikolaus Otto is a German mechanical engineer who invented the four-stroke compressed internal combustion engine in 1876.
- Max Planck discovered in 1900 that the photons that make up light spread in the form of energy packets, thus creating the first quantum theory. Einstein's Theory of Relativity, published in 1905 about the nature of very small and very high-speed substances (subatomic particles), led to new studies on electromagnetic waves and charged particles. After 1930, subatomic particles were discovered, nuclear power plants, jet aircraft and submarines were developed.
- In 1947, at Bell Laboratories, a team led by William Shockley, John Bardeen and Walter Brattain discovered the transistor, the lifeblood of electronic circuits.
- After the 1980s, developments in electronics, information and computer technologies initiated the information technology industrial revolution. In the early 21st century, information emerged as power. Algorithms and systems capable of super quantum calculations are being developed on the basis of artificial intelligence for mobile autonomous machines.

The Invention of the Computer

Computer science has its roots in two areas

- Maths
  - Leibniz's Dream (1600s), Can we find a universal language for mathematical algorithms that would allow us to describe and solve any problem?
  - George Boole (1800s), Introduces binary representation of computation. Computers use binary numbering for logic and arithmetic.
  - Alan Turing and the Turing machine (1930s), Theories developed on how to do computations by hand with paper and pencil.

- Engineering
  - Abacus – developed in the Middle East 5,000 years ago.
  - Pascaline – first mechanical calculator that used gears for calculation (1642).
  - Charles Babbage's Difference Engine – conceptual design that used hundreds of gears to calculate mathematical functions (1820s).
  - John von Neumann and the von Neumann machine (1940s), demonstrated how to build physical computers from electronic circuits.

The First Computer

- Alan Turing (1912 – 1954) was the first person to develop the working logic of the computer. A machine that processes and changes information!
- Turing was actually thinking about the solution of a mathematical problem.
- Something unexpected happened and the computer emerged. This machine changed the lives of almost everyone. Turing was interested in solving certain operations in mathematics by following a series of simple rules. In 1936, the literal meaning of the word computer was arithmetic calculations.
- Turing saw that all calculations were in binary. He focused on data and instructions that told him what to do with the data. Turing wanted to translate arithmetic operations into a language that machines could understand.
- Turing succeeded in this; he showed that when instructions consisting of 1s and 0s were given to the computer as commands on a tape, the machine would perform functions like the human brain. Tapes had become environments where information and commands were stored and processed. Today, pictures, music, writings, sounds, and images can all be processed by a single machine. All the processes we call programs and applications on a computer are nothing more than data on very long strips of 1 and 0. The 1s and 0s on the incredibly large strip can show you how a whole universe was created on the screen before your eyes. It showed that knowledge is power.

Transistor

- The transistor was discovered in 1947.
- A semiconductor circuit element that controls the flow of electrons (Electric current).
- Subatomic particles (Quantum Mechanics): Proton, Neutron, Electron, Photon
- Current is formed by the flow of electrons.
- The transistor memory element stores the bit (0/1) status on it. It makes a switch. Or it strengthens the signal.
- The transistor is the most widely used electronic circuit element in the world.
- The smallest basic electronic circuit element of the microprocessor is the transistor.
- The basic function cycles of the CPU occur at a dizzying speed due to the fact that transistors open and close millions and even billions of times per second.
- Today, they are produced in atomic structure.

Computer - Microprocessor

- Computer: A programmable machine that receives data as input in the binary number (bit: 0/1) system, stores and processes the data, and provides output data in a usable form.
  - Input: Information (Data)
  - Storage: Memories
  - Commands: Software
  - Data processing: Microprocessor
  - Data transfer: Transmitter, communication media and systems, receiver
  - Output: Information
- Microprocessor: The main component of the computer system. It is a program-controlled semiconductor device that receives commands from memory, decodes, processes, and produces output for memory or I/O units. It is called CPU (Central Processing Unit).
- Today's microprocessors: CPU, GPU, ALU (Quantum Computers or Quantum Computing), Signal Processing Units

Data Types

Symbols:

- Documents, Texts
- Graphics, Tables
- Numberic, Alphabetic, Alphanumeric
- İmages, Videos

Signals:

- Sound, Vibration
- Electrical
- Electromagnetic

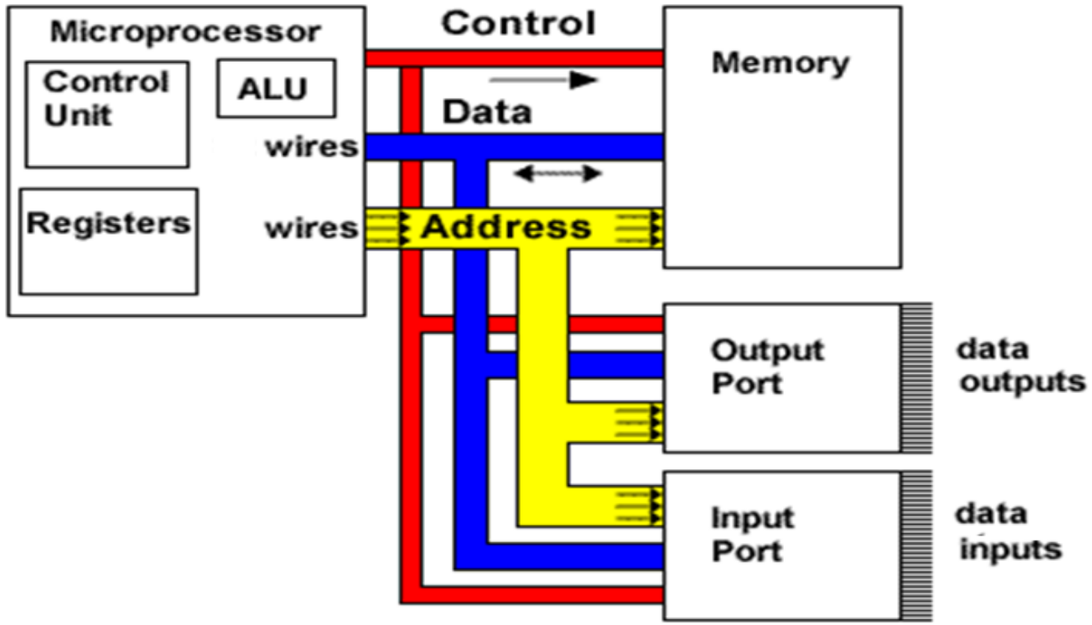
- When the speed of propagation of vibrating electromagnetic waves was calculated, the speed of light was revealed. Visible light, ultraviolet light and infrared light are electromagnetic waves of various wavelengths.
- Odor and particle emission
- Data comes to the computer from the outside world in different ways, in different sizes and also types: Numeric, integer, float, complex, character, string, binary, boolean. Inside the computer, all data is described by bits. Binary numbering system, bit: 0/1

Components of a Computer System

- The basic electronic circuit element in traditional computers is the transistor.
- Signals are represented by the binary number system as electrical.
- CPU – Microprocessor
- Memories (Main Memory: Ram (Write/Read, data is lost when the power is turned off), Rom (Read only memory, data is not lost when the power is turned off), CMOS (Battery)
- Clock & Timing
- I/O Peripheral Input/Output Units
- System Bus (Address, Data, Control): These are the lines that provide data communication where the memory cells or I/O units are selected by the CPU, data is written and read, and synchronization is provided.
- Software

Internal Structure of a Microprocessor

- System Bus (Address Bus, System Bus, Control Bus): All units (Memories, I/O) in and around the CPU are connected to the CPU via conductive wires or lines called "System Bus" and all communication is provided as electrical signals bit: 0/1.
- Arithmetic Logic Unit (ALU): performs all mathematical and logical operations, receives the instruction sequence from the CU.
- Control Unit (CU): Organizes the process flow, interprets commands and ensures that these commands are executed. CU controls the timing / sequencing of operations.
- Registers (Special Purpose Registers): Temporary registers to which the instructions to be processed are transferred.
- CPU Interconnection provides communication between the control unit, ALU and registers.



IoT- Internet of Things

- IoT (Internet of Things) are intelligent systems that communicate and cooperate with each other. They collect data from sensors and meters. They store the data they collect and produce information from the data. They communicate with each other and share data within an intelligent network structure to cooperate with each other.
- With the development of mobile (mobile) networks and the internet, it has become easier for smart objects to communicate with people, and people have the chance to observe and control them from anywhere, anytime.
- In the near future, the amount of data that will emerge thanks to smart objects will increase incredibly, and the analysis and processing of this large data will become difficult and complex. Autonomous software will play a strategic role.
- The privacy and security of data also emerge as an important issue.
- We will enter a period where mutual interaction will enter every object and different objects will move mobile for common purposes.
- In the meantime, how people will respond to this change physiologically and psychologically also presents itself as an important question.
- In driverless vehicles, people have handed over their life safety to objects.

IoT - Components

- Systems and machines
- Mobility
- Energy
- Communication systems
- Computer System – Embedded System
- Telemetry, transducers and sensors: Conversion of environmental changes into electrical signals
- Analog to Digital Converter (ADC – DAC; bit: 0/1)
- Actuator: Conversion of electrical signals into movement
- Intellegince Algorithms – Mathematical models

What is Artificial Intelligence?

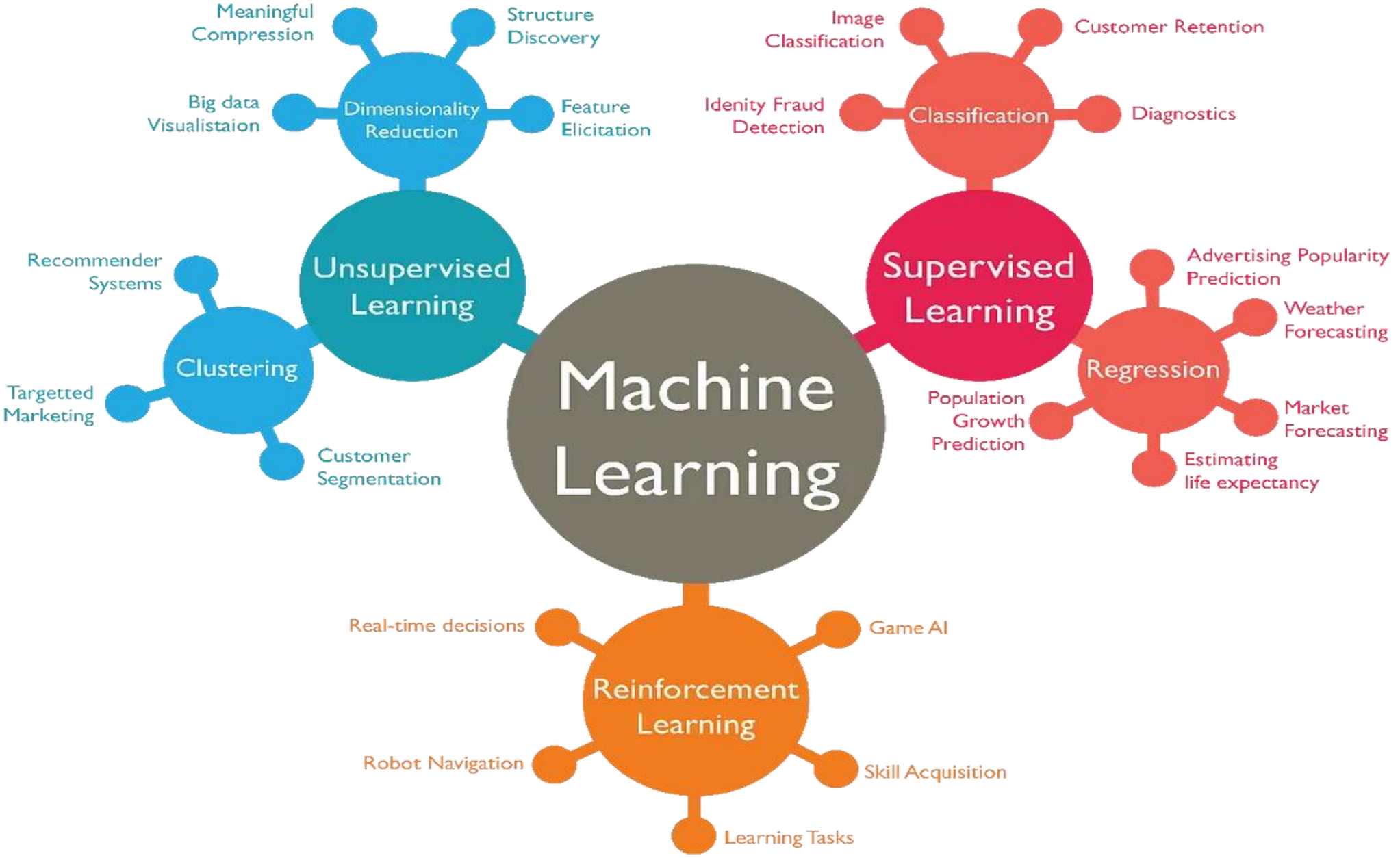
- Artificial: Instead of occurring naturally, it is brought about by human experience, knowledge, art and effort.
- Artificial Intelligence: The development of computer systems that exhibit human intelligence, and the representation of intelligent actions related to humans using computers.
- Artificial Intelligence is the development of computer systems and software that can perform tasks that require human intelligence.
- Learning from a dataset: The process by which a computer-controlled system or machine gains experience from a dataset and improves its performance from the experience it gains. Examples of these tasks include visual perception, speech recognition, decision making and translation between languages.

Artificial Intelligence

- The term “Artificial Intelligence” was coined in 1956 by John McCarthy of the Massachusetts Institute of Technology. It is a branch of computer science that aims to make computers behave like humans.
- Artificial Intelligence is the development of algorithms and mathematical models that represent data sets to develop machines that make decisions in real-life situations.
- Artificial Intelligence is programmed computers to hear, see and respond to sensory stimuli and systems that mimic human intelligence by trying to reproduce the types of physical connections between neurons (neural networks) in the human brain.
- Artificial intelligence is the development of computer-controlled autonomous robotic systems.

What is machine learning?

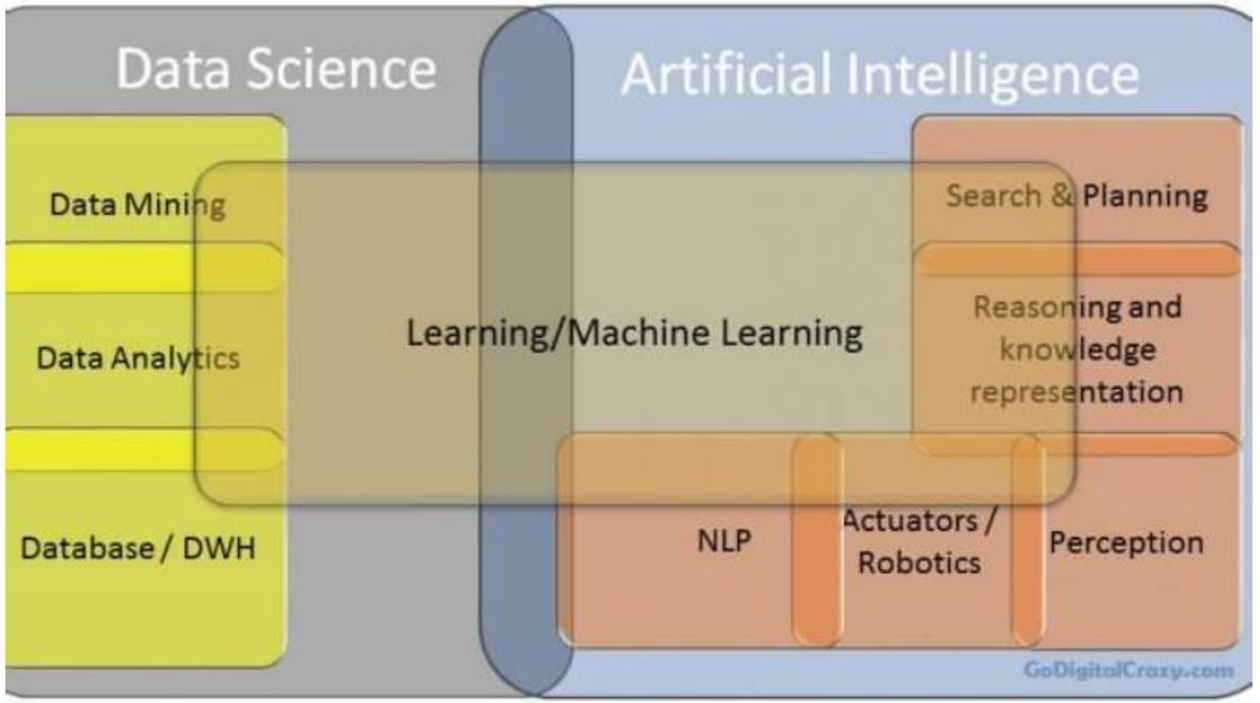
Machine learning is the study of intelligent algorithms that improve performance by learning from model that represented to a mass of data, rather than relying on rigidly coded rules and making predictions on new data.


Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Market Forecasting
Estimating life expectancy
Population Growth Prediction
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

What is deep learning?

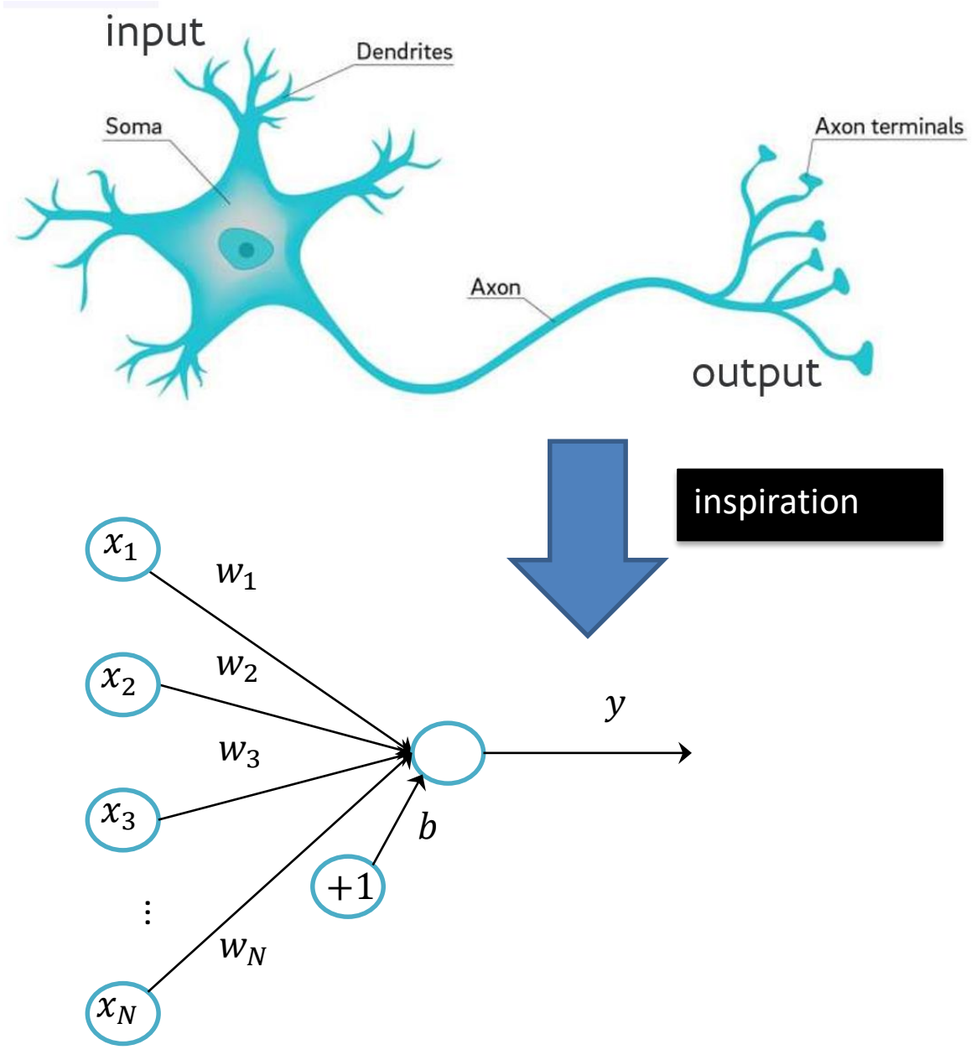
Deep learning is a subfield of machine learning that focuses on learning data representations over a range of variation in which increasingly meaningful features of the model are developed by learning from the dataset.

AI vs ML



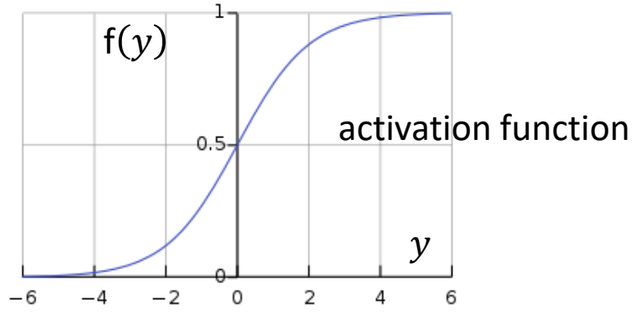
Data science combines statistical tools, methods, and technology to generate meaning from data. Artificial Intelligence takes this one step further and uses the data to solve cognitive problems commonly associated with human intelligence, such as learning, pattern recognition, and human-like expression.

From Neurons to Artificial Neural Networks



Artificial neural networks (ANN) are an information processing technology inspired by the information processing technique of the human brain. With ANN, the working style of the simple biological nervous system is imitated. Here wi coefficients, xi input values and b gives the tendency or trend parameter. (i=1,2,3, … , N)

$$y = \left(\sum_{i=1}^{N} w_i x_i + b \right)$$



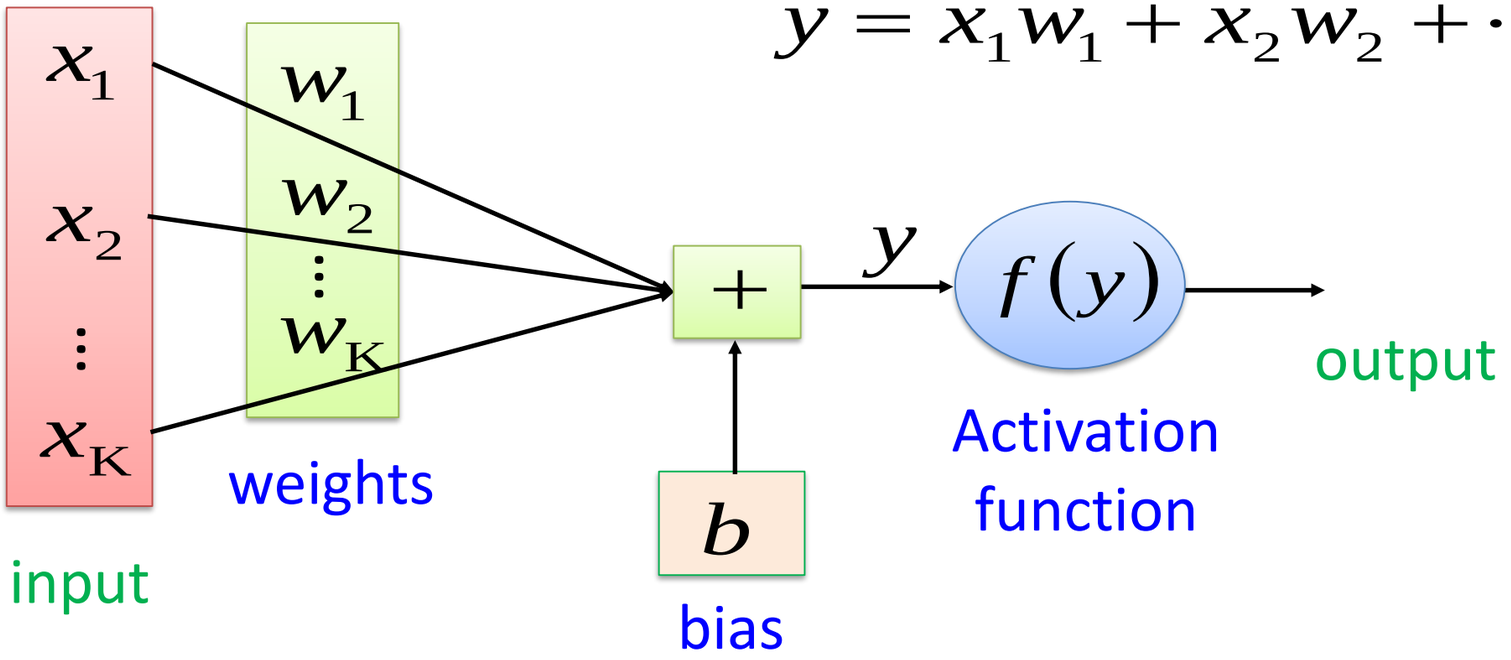
activation function

Artificial Neural Networks

- The learning process in humans occurs through the limitless connections established between biological neuron networks in the brain.
- Each new stimulus that comes to the brain creates new connections between neural networks, leading to the reorganization of the relationships between neurons.
- As a result, the constant repetition of a certain task strengthens the neurological connections related to this task, and learning occurs.
- Once learning has occurred, the brain takes these previously formed connections into account when evaluating its responses to new stimuli.
- Artificial neural networks, which form the basis of artificial intelligence, are based on a much simpler, narrow-scope imitation of the brain's biological functioning. Software programmed to learn attempts to make predictions using statistical data analysis.

Fundamentals of Artificial Neural Network

- NNs consist of hidden layers with neurons (i.e., computational units)
- A single neuron maps a set of inputs into an output number, or $f: R^K \rightarrow R$



$$y = x_1 w_1 + x_2 w_2 + \cdots + x_K w_K + b$$

$$f(\mathrm{y}) = \frac{1}{1 + \mathrm{e}^{-\mathrm{y}}}$$

$$e^{-10} \cong 0$$

What is a Quantum Computer?

- Quantum Computer (Electron, Photon)
  - A computer that uses quantum mechanical phenomena to perform operations on data through kuantum physical devices such as quantum superposition and entanglement.

- Quantum computing is a multidisciplinary field comprising aspects of computer science, physics, and mathematics that utilizes quantum mechanics to solve complex problems faster than on classical computers. The field of quantum computing includes hardware research and application development.

- Classical Computer (Binary – Bit:0/1)
  - A computer that can be made calculations entirely with classical mechanics, using electrical signals passing through electronic circuits and gates.

Quantum Computing - Machine Learning

- In recent years, data analytics applications and research on intelligent machines have re-emerged strongly.
- This renewed interest is partly due to the developments in classical computational methods and partly due to the enormous parallelism potential offered by Quantum Computing (QC) and related quantum technologies.
- These developments in computational methods, Machine Learning (ML), data-driven learning and quantum-supported computational methods have a strong potential to realize the demands of a fully intelligent communication network focused on services.
- In the emerging paradigm of increasing human-machine connectivity, a significant increase in the number of network nodes and data traffic is expected.
- Machine Learning (ML) and Quantum Computing (QC) methods will provide a new framework for efficient processing of large amounts of data, enabling Quantum ML (QML) technologies.

Computer Science

- Computer science, as a discipline, encompasses a range of topics, from the theoretical study of algorithms to the study of computation and the limits of computation, to the practical and theoretical implementation of computer systems in hardware and software.

- The four areas considered important in computer science are defined as follows:
  - Theory of computation (Applied Mathematics)
  - Algorithms (Clever algorithms) and data structures
  - Programming methodology and languages: C++, Python, Java Script, Matlab, Assembly
  - Computer organization and architecture

Data Science

- Data science is the science of preparing and analyzing raw data using computational mathematics, statistics, probability, and artificial intelligence techniques to draw conclusions.
- Data Science is the in-depth study of large amounts of data that involves extracting meaning from raw, structured and unstructured data. Extracting meaningful data from large amounts uses data processing algorithms and this processing can be done using statistical techniques and algorithms, scientific techniques, different technologies, etc. Data science uses various tools and techniques to extract meaningful data from raw data. Data Science is the Future of Artificial Intelligence.

Optimization

It is the process of finding a better or more suitable design example among possible design variations. It focuses on the sensitivity of the parameters. It uses statistical data analysis and computational mathematics methods. Tuning - It is the function of adjusting to the appropriate one.

In mathematical modeling, it refers to the process of systematically examining or solving a problem by selecting real or integer values in a defined range and placing them in the function in order to minimize or maximize a real function. While optimization algorithms minimize the loss in a learning set, machine learning is the best optimization method in the parameters used in the algorithm that minimizes the loss in unseen samples.

How to optimize scalability?

- Parametric scans are automated.
- Real-time parameter adjustments are performed using analytical derivatives.
- Performance characteristics are determined.
- Sensitivity and statistical analysis are performed on the optimized model.

Simulation: Creating a Simulated Model with a Mathematical Model

- Simulation is the process of creating a mathematical and algorithm-based digital environment that mimics the real world.
- Simulations are inspired by reality using mathematical models and algorithms. In this way, it is possible to experience different scenarios and predict their results.
- Simulations are used in many areas from education to the health sector, from engineering to games.
- Simulation is a software system that allows experimentation with logical and mathematical modeling based on computational mathematical applications in order to understand the structure and behavior of the real system.

Signals and Systems

Signals

- A signal is the variation of a physical, or non-physical, quantity with respect to one or more independent variable(s). Signals typically carry information that is somehow relevant for some purpose.
- Electrical signals: voltage as a function of time
- Acoustic signals: acoustic pressure as a function of time
- Speech is produced by creating fluctuations in acoustic pressure, which can be sensed by a microphone and converted into an electrical signal.
- Picture: brightness as a function of two spatial variables
- A camera senses the incoming light and records the light reflectivity as a function of space onto a magnetic film.
- Other examples: sequence of bases in a gene (biological signal), sequence of daily stock prices in the financial market, ...
- We will mostly refer to the independent variable as time (t), although it can be other things (such as space) depending on application.
- We consider two types of signals : continuous-time (CT) signals and discrete-time (DT) signals.
  - In continuous-time (CT) signals, the independent variable is continuous.
  - In discrete-time (DT) signals, the independent variable is discrete.

Signals

- Signal is the carrier of information. Information is given meaning by signal. If you measure, collect, store, process, transfer signals, you manage them. Also, if you classify, define them as coefficients of functions with clusters or regression, you interpret them, and direct them in making profits.
- We become conscious by discovering the universe and the information hidden in signals.
- Signals that carry information about the source they exist in provide extraordinary features to the environments they interact with.
- In electronics, a signal is an electric current or electromagnetic field used to carry data from one place to another.
- Signals spread far away in the form of waves. Waves carry signals. Acoustic signals, Seismic signals, Electrical signals, Electromagnetic signals, Heat, Vibration, movements of subatomic particles, gravity...
- In the near future, information will be carried and processed as signals formed by subatomic particles such as electrons and photons.
- The information carried is stored on the signal and gains meaning; information is written on a stone, a book. It is written into a memory or brain. What makes information powerful is that it can be carried and stored integrated with signals. Information processed on a clay tablet is stored for ages and stops time. The signal shows how information behaves according to the laws of physics. Humanity must learn that information is integrated with the physical world.

Signals

- A signal is a pattern of change that carries information.
- All mathematical functions are called signals.
- Signals are mathematically represented as a function of one or more independent variables A picture is the change in brightness as a function of two spatial variables (x and y).
- Signals that contain a single independent variable, usually called time, t, are considered.
- A signal is a real-valued or scalar-valued function of an independent variable t.

Signal Sources

- Electronics: Pulse Generator, Oscillator
- Electromagnetics: Antenna
- Heat, Light, Sound, Vibration, Gravitational forces
- Natural signal sources
- When electromagnetic radiation with high energy at very high frequencies is sent to a conductive plate, a current flows through the conductors. This event is called Photoelectric.
- Radioactive materials emit signals.
- Stars are signal sources.
- Types: Sine Wave, Ramp, Step, Chirp, Clock, Consant, linear, exponential, polynomial, trigonometric,..
- Noise, Signal to Noise
- Signals are affected by other signals as attenuation and interference while being transmitted (Interference)

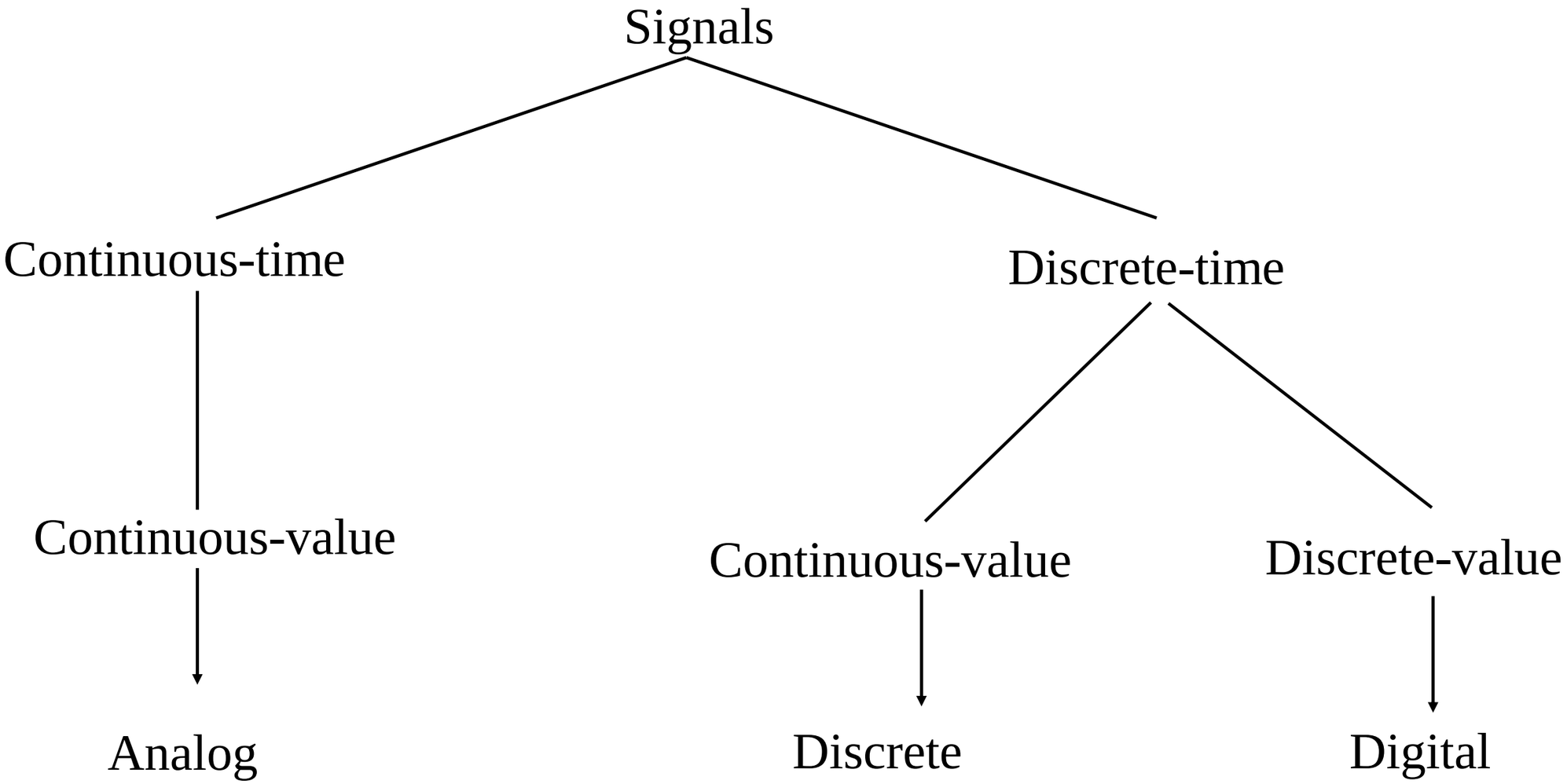
Error sources that corrupt signals

Errors: Intentional errors. Unnoticed systematic errors. Individual errors.

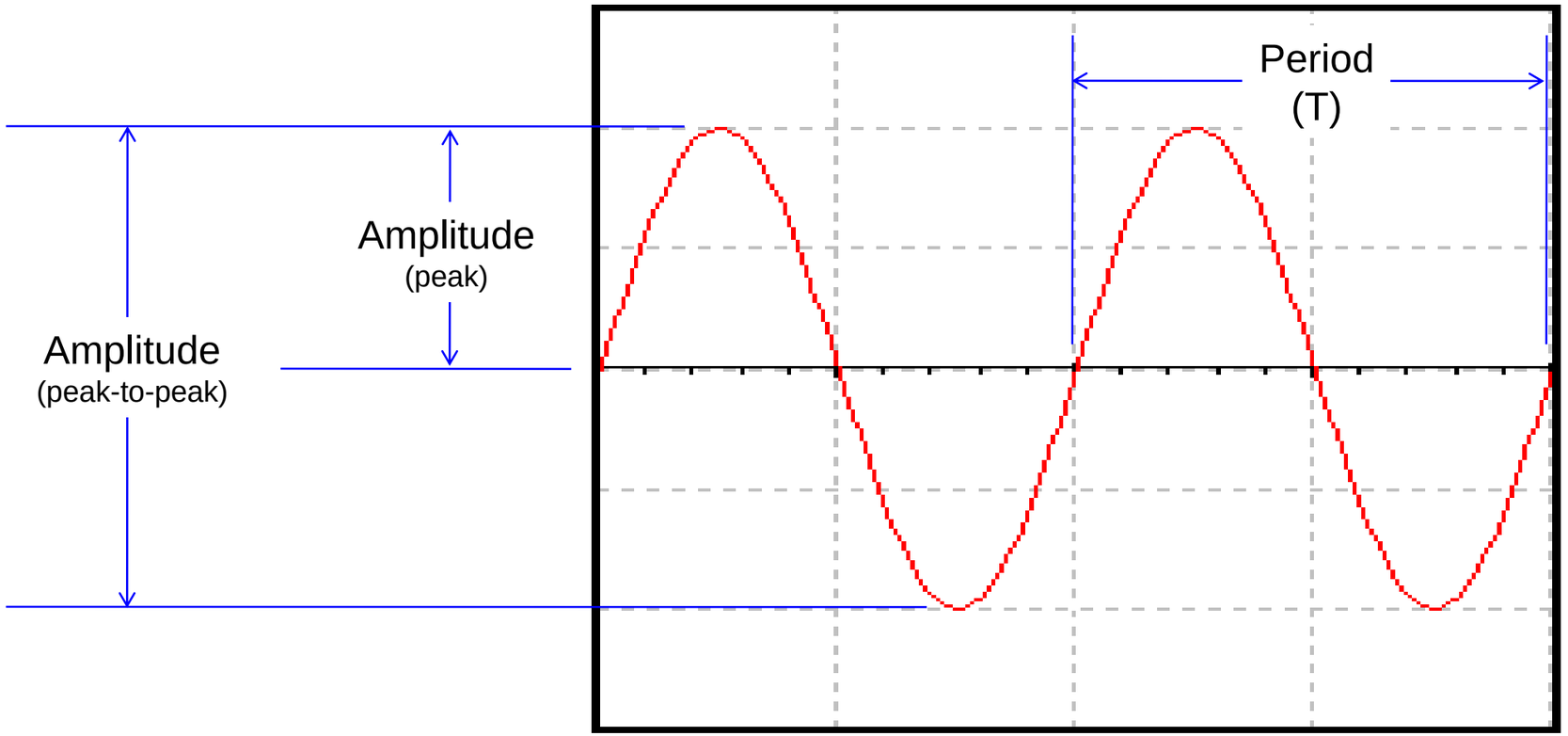
Software errors: mathematical modeling, algorithm, coding; incorrect data entry
Systematic error: Random, Measurement error, Sampling error.

- Noise
- Uncertainty
- Lost data
- Precision
- Variability
- Interference: noise, crosstalk, blocking
- Deviation
- Sizing, Normalization
- Attenuation

Signal Types



Parts of an Sinusoidal Signal

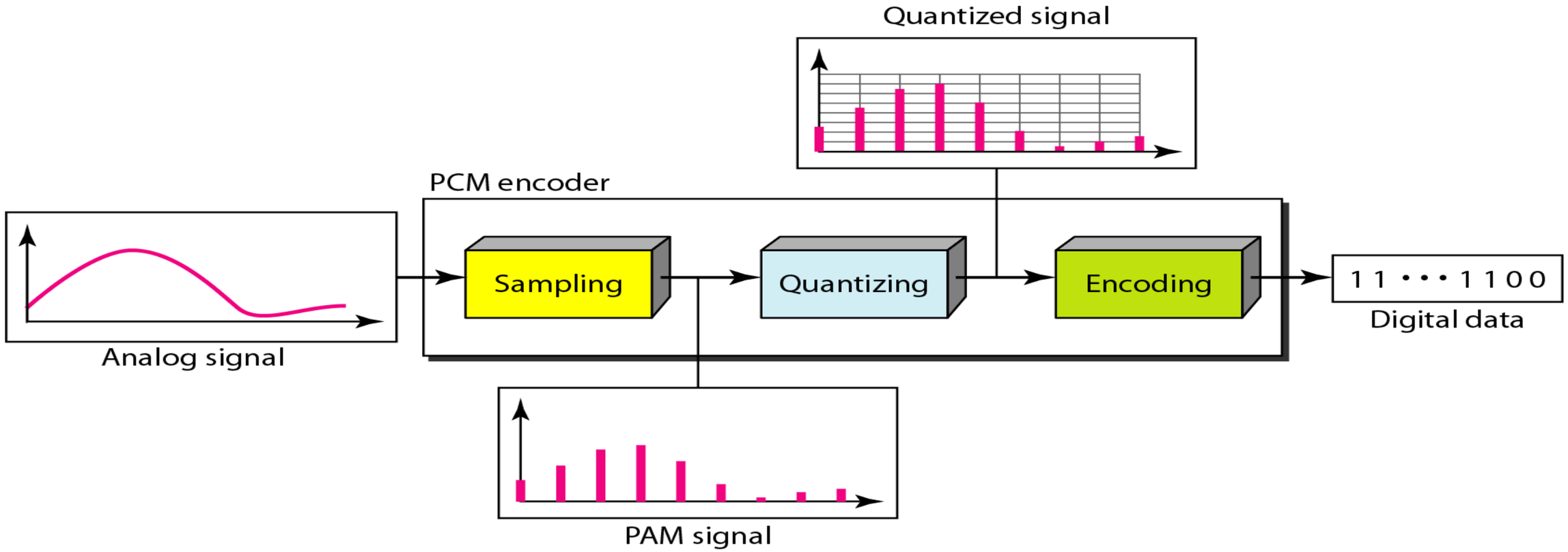


Frequency:

$$F = \frac{1}{T} Hz$$

Parts of an analog signal: amplitude, period, Phase & frequency. Analog sinyal çok sayıda farklı frekansları, fazları ve genlikleri olan sinüsoidal sinyalerin bileşmesinden oluşur.

Digitization of analog signal



Binary Numbering Systems - Bit (0/1)

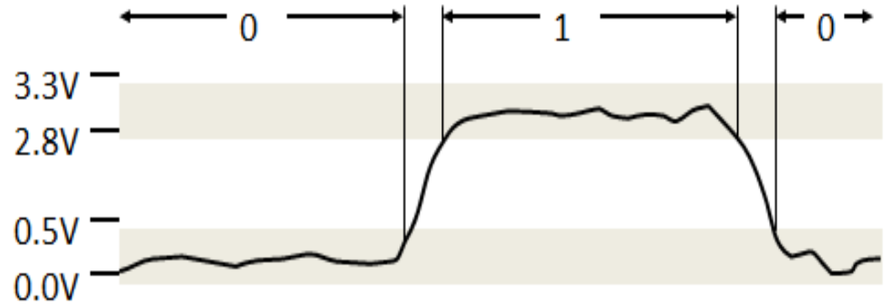
Binary signal, binary state signal: Data with two states (0/1).

- – off & on
- – Carried and stored with electrical signals.

low voltage & high voltage; 0v & 5v

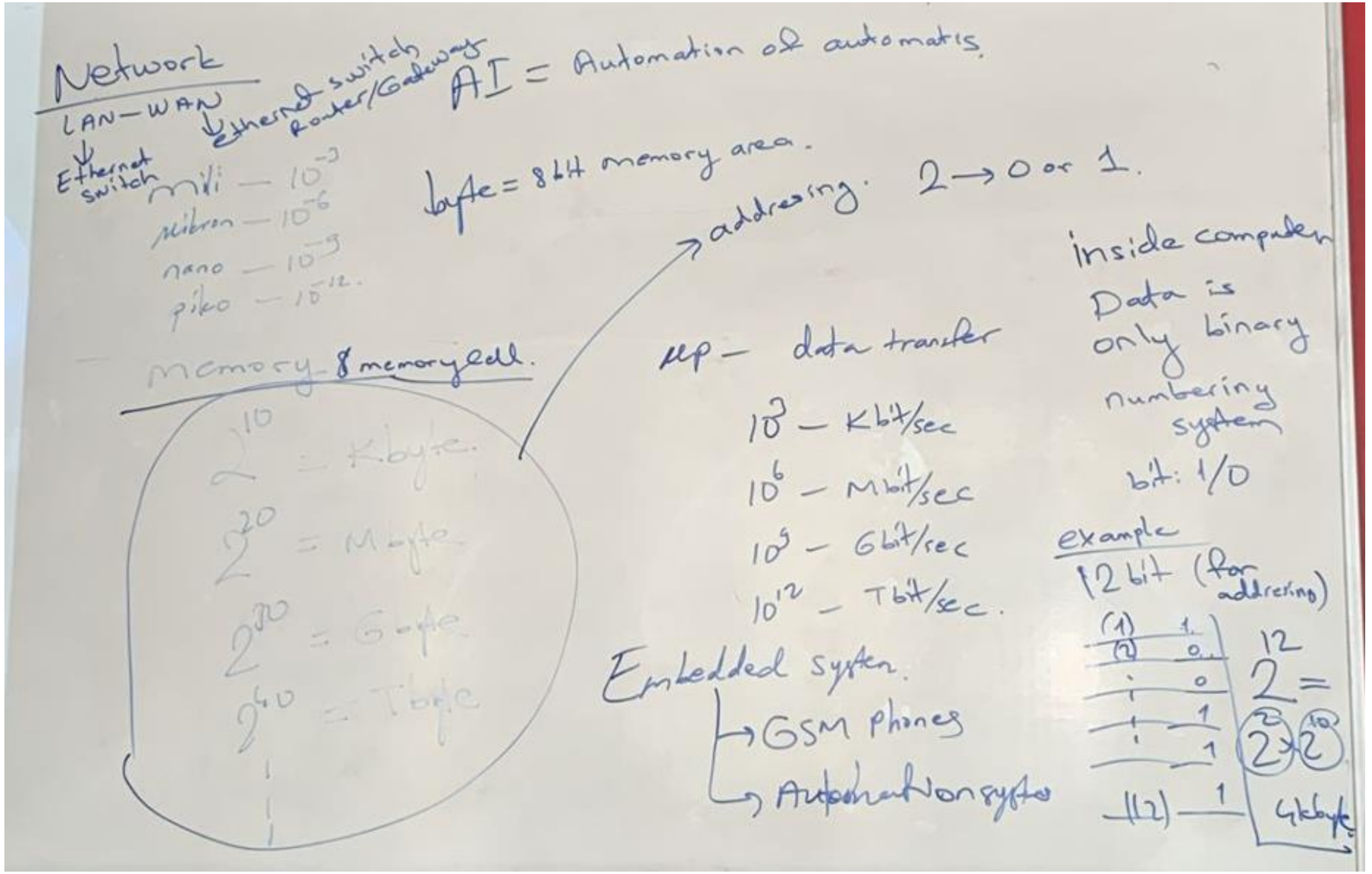
Bit: Not only a mathematical concept, but also has a physical world equivalent.

- The binary number system has a value of 0 or 1 and nothing else.
- A bit is the smallest unit of information in a computer



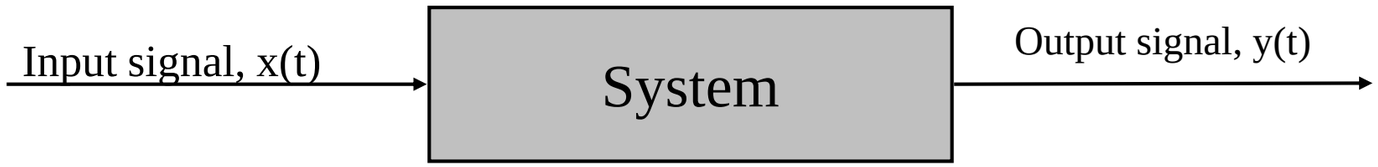
Basic Units in Digital Systems

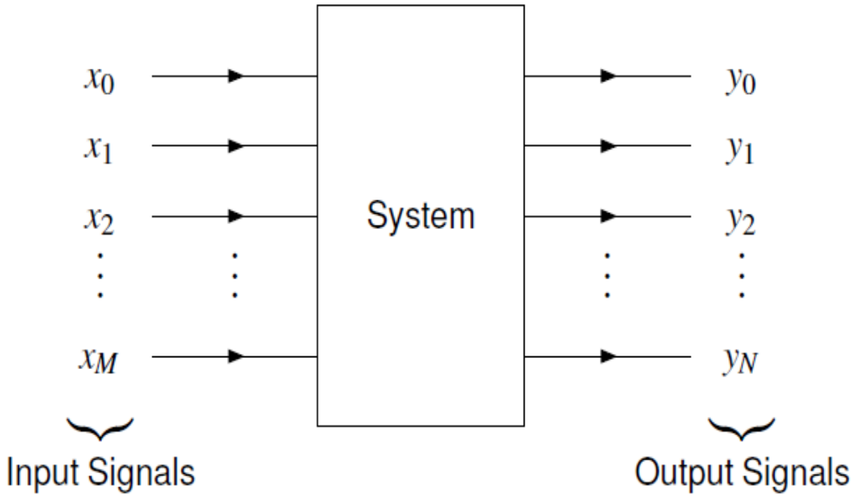
- The basic unit of information in computer systems is bit: (1/0)
- Bit: 0/1; bits are represented by electrical signals.
- Byte: Represents 8-bit data. Or indicates a 1-byte memory slot. Represents operations in base 2 in memories. Address Bus; selects memory and memory slot.
- Memory Size: Expressed with 2^n. Here n: is the number of address lines coming to the memory. If 2 lines come, the lines will be: 00, 01,10,11. In that case, 2^2=4 bytes
- Bit/sec: Represents the amount of data to be transferred or processed in one second. It is shown with exponential operations in base 10^n.
- Qubit: Represents the smallest data in quantum calculations.
- Electron: Qubits are represented by electrons.


Network
LAN – WAN
Ethernet switch
Router/Gateway
Ethernet switch
AI = Automation of automatis.
byte = 8bit memory area.
addressing: 2 → 0 or 1.
mili – 10^{-3}
mikron – 10^{-6}
nano – 10^{-9}
piko – 10^{-12}.
memory & memorycell.
2^{10} = Kbyte
2^{20} = Mbyte
2^{30} = Gbyte
2^{40} = Tbyte
up – data transfer
10^3 – Kbit/sec
10^6 – Mbit/sec
10^9 – Gbit/sec
10^{12} – Tbit/sec.
Embedded syst.
GSM phones
Autopilot Navsyst
inside computer
Data is only binary numbering system
bit: 1/0
example
12bit (for addressing)
12
2^{12} = 2*2
4kbyte

What is a System?

- Systems are units that process input signals and convert input signals to another signal to produce output signals in line with the purpose.
- Systems can be physical or hardware, as well as completely software. Software models are mathematical models.
- A system is a software and/or hardware unit that processes one or more input signals to produce one or more output signals.
- Systems provide the desired system response by converting one signal to another.
- Signals can be physical as well as numerical values. (Vector, Matrix, equation)
- A system is a physical integrity of components or parts that collect and process information to produce output or ,nputs.





System Components

- Computer Systems
- Sensors, Transducers
- Automation systems, Analog, Digital I/O; control and management software
- Actuators, units that convert electrical energy into motion
- Robot Arms: Sensor + Actuators
- Wireless communication systems
- Driver, interface modules and software
- AI software

Data

Data

- Data are unrelated, undigested, unprocessed facts or pieces of information. They are in forms devoid of any content.
- Data: All kinds of initial raw information obtained during research. They are symbols such as signals, pictures, images, shapes, numbers, texts and sounds transferred to the computer's memory.
- Data is a physical quantity that carries information.
- They do not carry interpretation but are ready to be processed. They are not effective in decision making.
- Information: Processed, organized, and meaningful data.
- Consciousness: It is to provide the mind with the ability to perceive, visualize and interpret.
- Understand: It is the continuity of understanding by questioning, grasping and feeling.
- Ability - Experience (Knowledge): It is to increase performance in decision making, estimation and searching for the truth.
- Wisdom: It is an evaluated understanding. It is effective in decision making and interpretation by questioning and estimation.

Data – Information – Knowledge - Wisdom

- Symbols (Signals, Pictures, Shapes, ...): They consist of information messages such as numbers, words, images, video and sound that are transferred to the computer's memory during the input phase. They are represented by signals that carry messages.
- Data: They are facts or pieces of information that have not gained meaning, have not been associated, have not been assimilated, and have not been processed. They are in forms devoid of any content. Sometimes they are physical events, observations. They do not carry comments, but they are ready to be processed. They are not effective in decision making.
- Information: What, who, when, where are the questions that need to be answered. Information is processed, organized, and meaningful data. Information is organized, meaningful and useful data. During the output phase, the information created is put into a presentation form with printed reports, graphics and visuals. The information is stored in the computer for future use.
- Knowledge (knowledgeable, Ability - Experience, Experience): It is to increase performance in decision making, estimation, and searching for the truth.
- Understand: is to become conscious by understanding, grasping, feeling. Wisdom: is an evaluated understanding. It is to decide and interpret by questioning, making predictions.

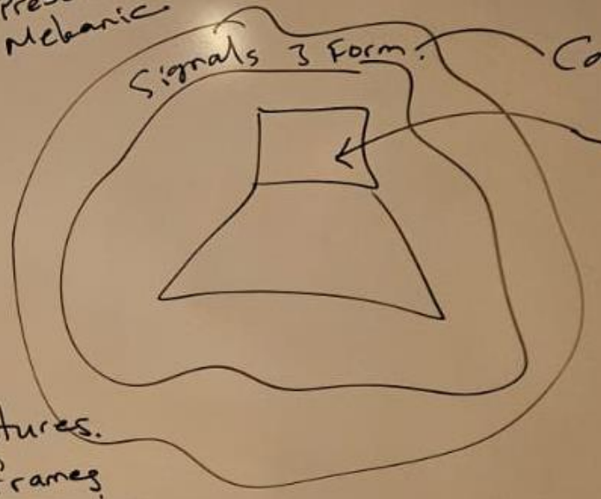
If an organization has no memory!

- *Stored information is a value that illuminates the past and future of organizations at the same time. If an organization does not have a memory, it does not have the value of existing.*
- *In the archiving of information produced in an electronic environment, the functions of content, structure, context, presentation and behavior should not be impaired in order to preserve the information without being corrupted or changed and to transfer it to the next generations.*
- *Inadequate archiving destroys accumulations and causes confusion. Organizations suffer great losses due to information and documents that cannot be found or are lost. The basic rule for the healthy operation of the system is that when information is needed in time, that information is found quickly. Therefore, all types of information should be classified and stored.*
- *If the right information is collected and archived at the right time, in the right place, from the right source, the position to be taken is also determined correctly; opportunities and dangers can be foreseen in advance and predicted before events occur.*
- *When the storage and classification of information is done properly, the corporate brain of the organization turns into an intelligence that learns to make autonomous decisions by combining and evaluating information.*

Data is 4 Form.

1- Symbols: Computer Klauye. has alot symbols: Number, Alphabet.
+, −, /, *
Others ( ), ;, /; ...

2- Signals: Acustic Signal (Voice) ↓ Pressure Mekanic
Digital signals
Analog "
Electrical "
Electromagnetic "
Heating "
Gravity "



Convert to
Binary Numbering System
Bit: 1/0
Symbols.
Electrial Signal.

3- Picture, Video
↳ has alot pictures. or frames

2-5 picture send to 1 second to your eyes.

Analog Signal has some parameters: Amplitude, Time frequency, phase.

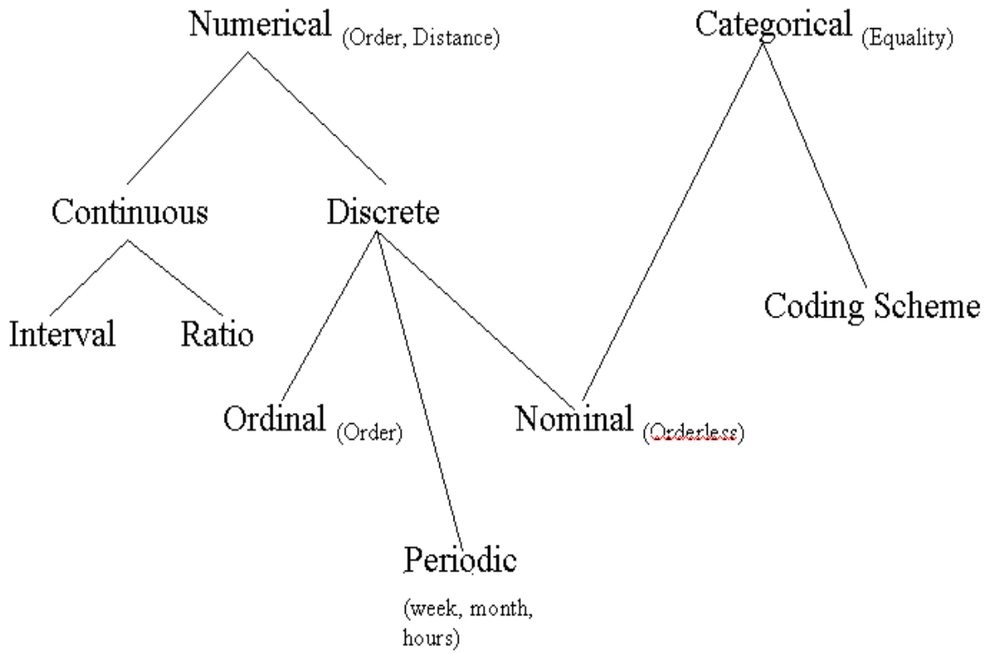
In an analog Signal Amplitude, frequency and phase change by timing.

$y(t) = at^2 + bt + c$

dependt t: independ variable
a, b, c: Constant.

Data Types and Forms

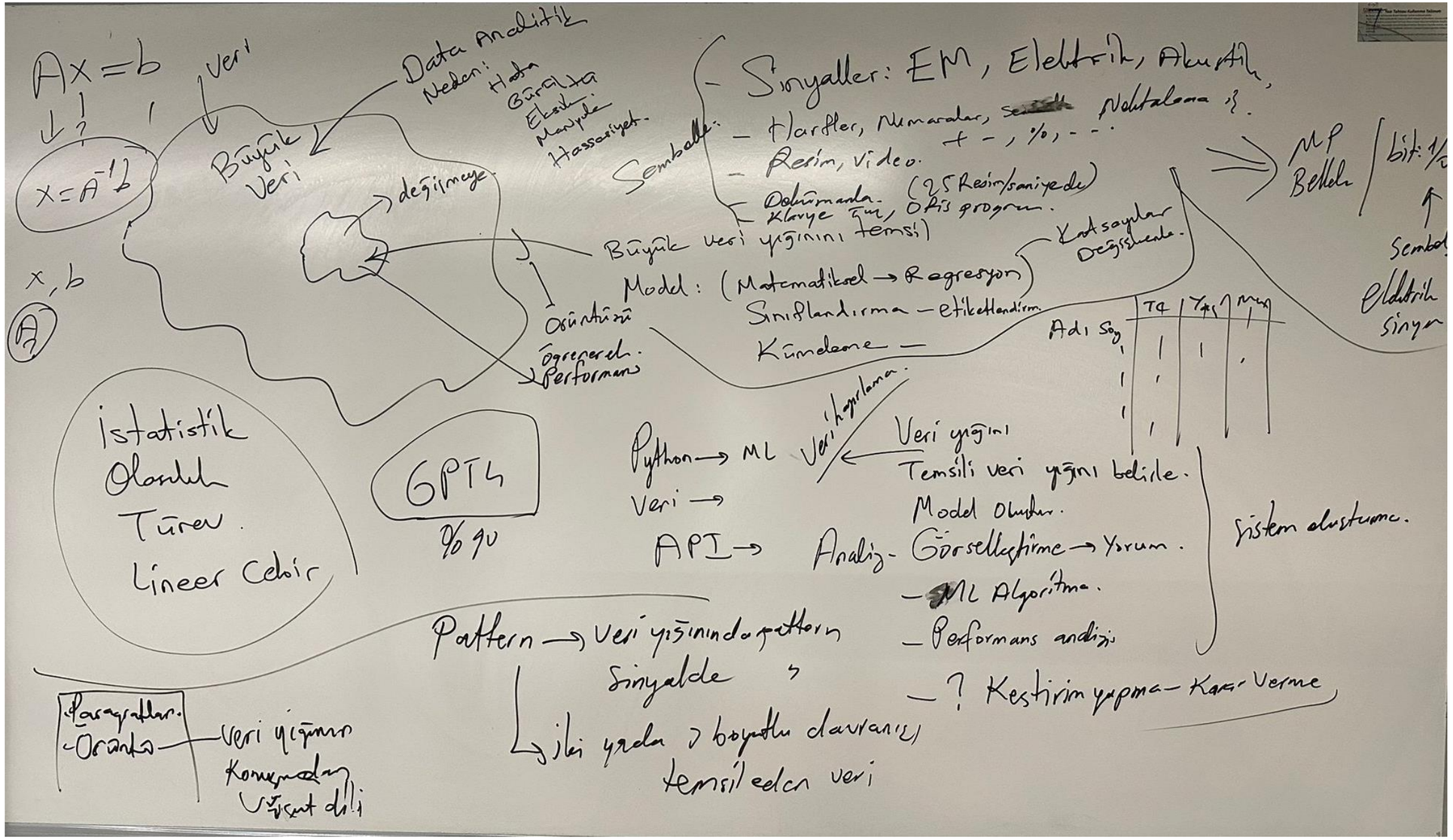
- Attribute-value data:
- Data types
  - numeric, categorical (see the hierarchy for its relationship)
  - static, dynamic (temporal)
- Other kinds of data
  - distributed data
  - text, Web, meta data
  - images, audio/video



Pattern

- In the world of data analysis, a pattern is a sequence of data points that show a shape, structure, algorithm, mathematical models or mathematical functions the intellegence system can easily recognize. Think about it this way: a pattern can be as simple as the regular rise and fall of daily temperatures throughout the year or as complex as the unpredictable fluctuations in stock market prices. The electromagnetic radiation pattern is a good example of determining the coverage area in wireless communication.
- Data patterns are very useful when they are drawn graphically. Data patterns commonly described in terms of features like center, spread, shape, and other unusual properties. Other special descriptive labels are symmetric, bell-shaped, skewed, etc.
- The behavioral model represented by the data stack.
- The behavioral model represented by the signal.
- The data representing the two or three-dimensional behavioral model.
- The body language of the data stack

BIG DATA


Ax = b
x = A⁻¹b
x, b
Veri
Data Analitik
Neden: Hata
Büyüklük
Eksik
Manyak
Hassasiyet.
Büyük Veri
değişmeye
Büyük veri yığınını temsil)
Model: (Matematiksel → Regresyon)
Sınıflandırma – etiketlendirm.
Kümeleme –
Örüntüyü
öğrenerek
Performans
Kat sayılar
Değişkenler
Semboller:
– Sinyaller: EM, Elektrik, Akustik,
– Harfler, Numaralar, Noktalama :) .
+ − , %, . . .
– Resim, Video.
– Dokümanlar (25 Resim/saniyede)
Klavye Fare, Ofis program.
MP
Bellek
bit: 1/0
Sembol
Elektrik
Sinyal
Adı Soy
TC
istatistik
Olasılık
Türev
Lineer Cebir
GPT4
%90
Python → ML
Veri →
API →
Veri Hazırlama
Veri yığını
Temsili veri yığını belirle.
Model Oluştur.
Analiz – Görselleştirme → Yorum.
– ML Algoritma.
– Performans analizi
– ? Kestirim yapma – Karar Verme
Sistem oluşturma.
Pattern → Veri yığınında pattern
Sinyalde
İki yada > boyutlu davranışı
temsil eden veri
Karşıtlar.
Örüntü
Veri yığının
Konuşmadan
Ustasını dili

What is Big Data?

- Big Data is a collection of data that is huge in volume, yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently. Big data is also a data but with huge size.
- Big Data are high volume, high velocity, or high-variety information assets that require pattern forms of processing to enable enhanced decision making, insight discovery, and process optimization.
- Big data is a collection of huge data sets that normal computing techniques cannot process.
- Technological advancement and the advent of new channels of communication (like social networking) and new, stronger devices have presented a challenge to industry players in the sense that they have to find other ways to handle the data.
- Big data is an all-inclusive term, representing the enormous volume of complex data sets that companies and governments generate in the present-day digital environment.
- Big data, typically measured in petabytes or terabytes, materializes from three major sources— transactional data, machine data, and social data

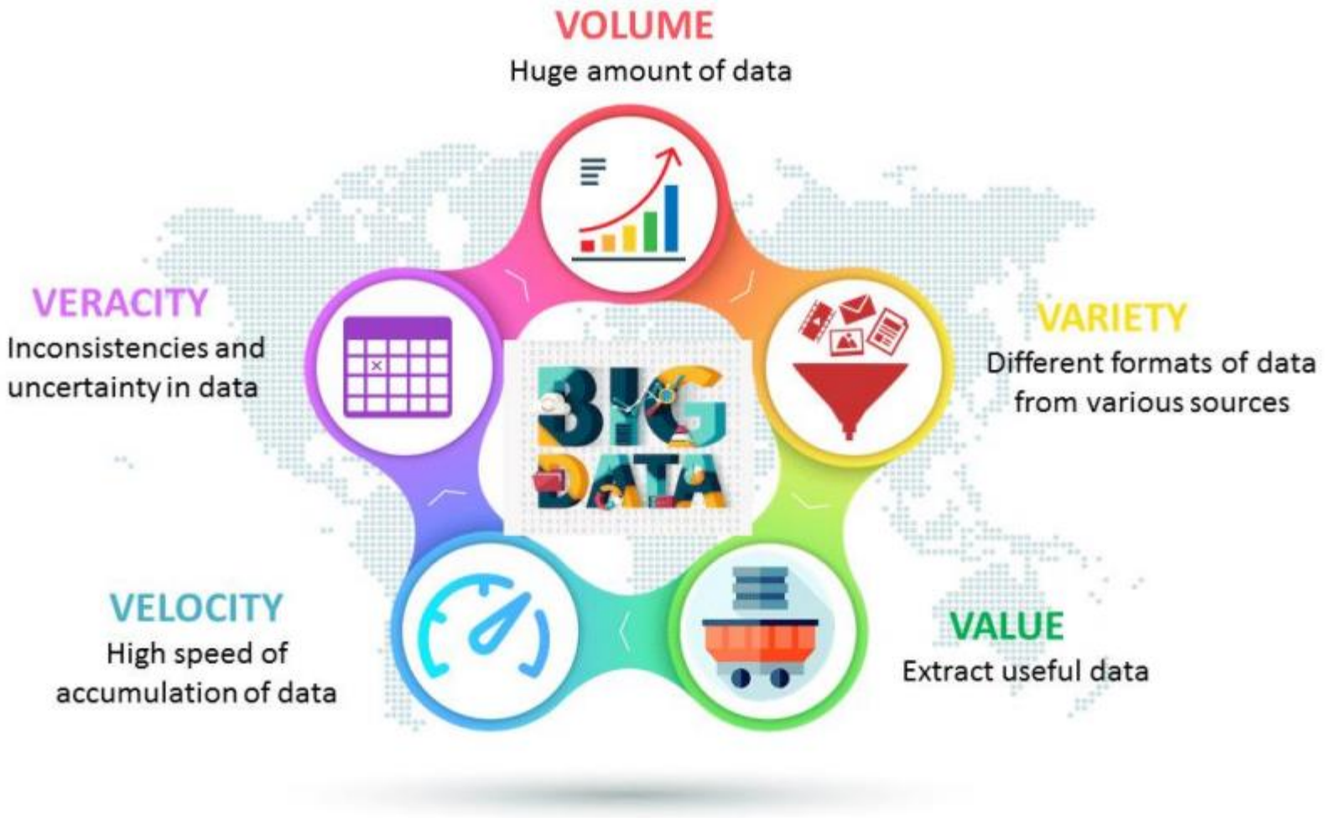
Types of Big-data

- Structured data 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: Database Management Systems(DBMS)
- Semi-structured data can be considered as another form of structured data. It inherits a few properties of structured sata, 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 an RDBMS (Relational Database Management Systems - Relational Database). Example: Comma Separated Values(CSV) File.
- Unstructured data is completely a different type of which neither has a structure nor obeys to follow the formal structural 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 Files, Images etc

Parameters that define big data

1) Volume (Hacim),
2) Velocity (Hız),
3) Variety (Çeşitlilik),
4) Verification (Doğrulama)
5) Value (Değer).
6) Veracity (Gerçeklik),
7) Volatility (Oynaklık)
8) Validity (Geçerlik)
9) Vulnerability (Hassaslık),
10) Variability (Değişkenlik),
11) Visualization (Görselleştirme).

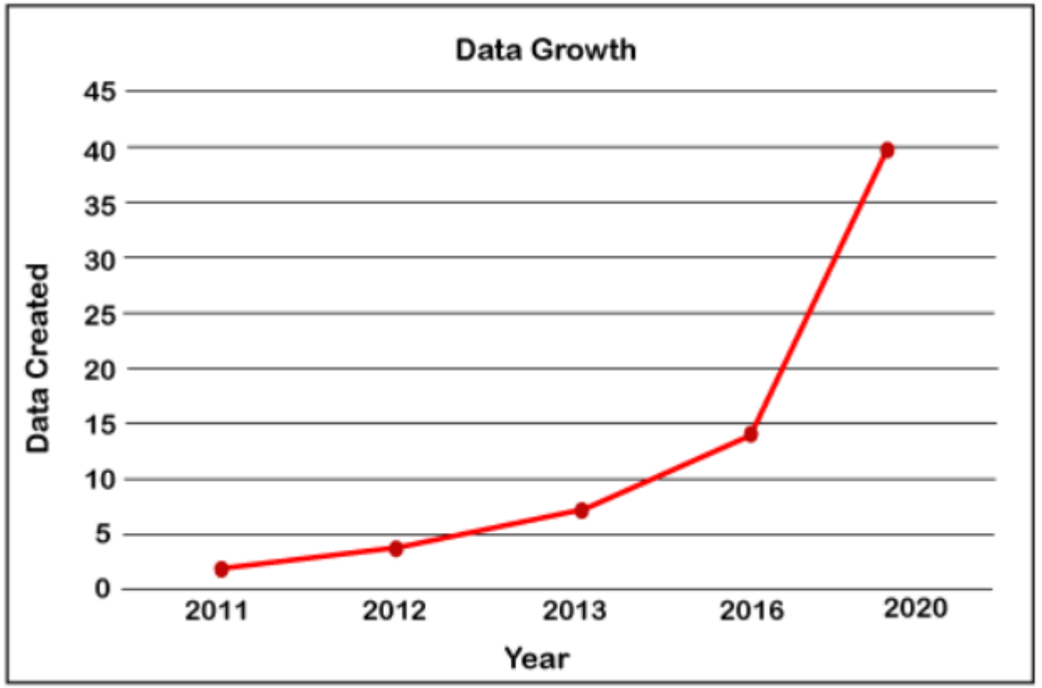
The Characteristics of Big Data -1



The Characteristics of Big Data -2

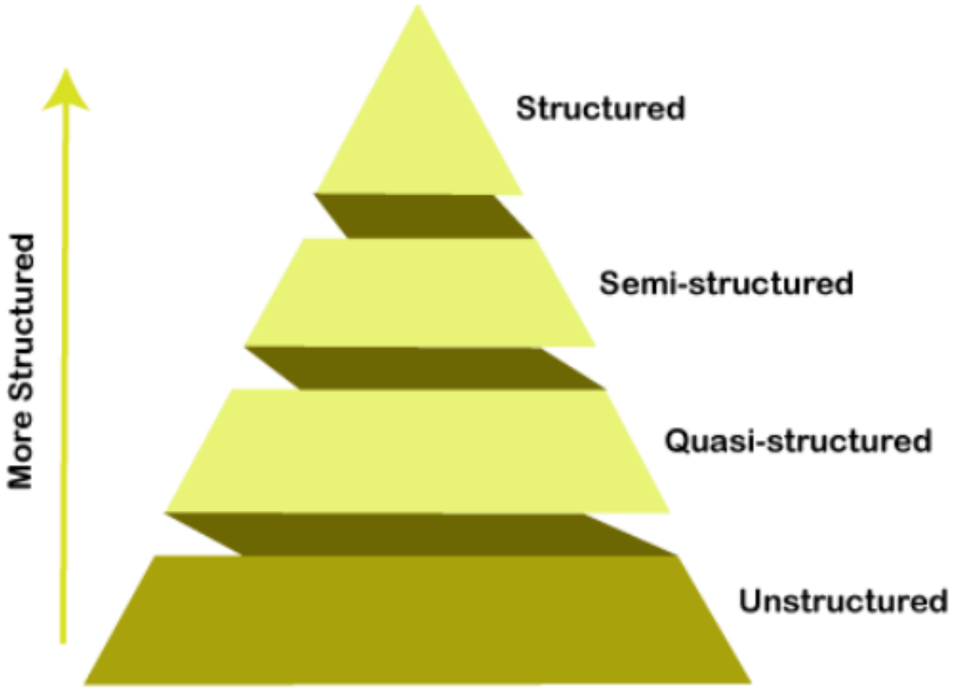
Volume

- Volume refers to the unimaginable amounts of information generated every second from social media, cell phones, cars, credit cards, M2M sensors, images, video, and whatnot.



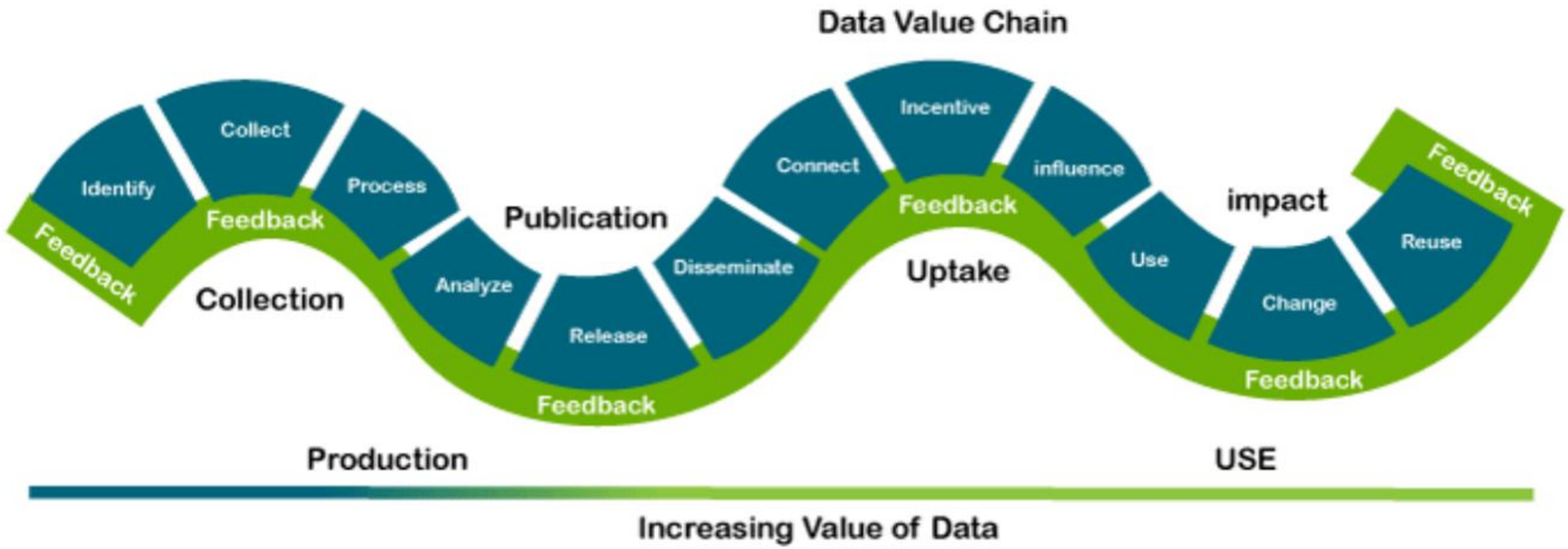
The Characteristics of Big Data -3

- **Variety:** Big data is generated in multiple varieties. Compared to the traditional data like phone numbers and addresses, the latest trend of data is in the form of photos, videos, and audios and many more, making about 80% of the data to be completely unstructured
- **Veracity:** 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.



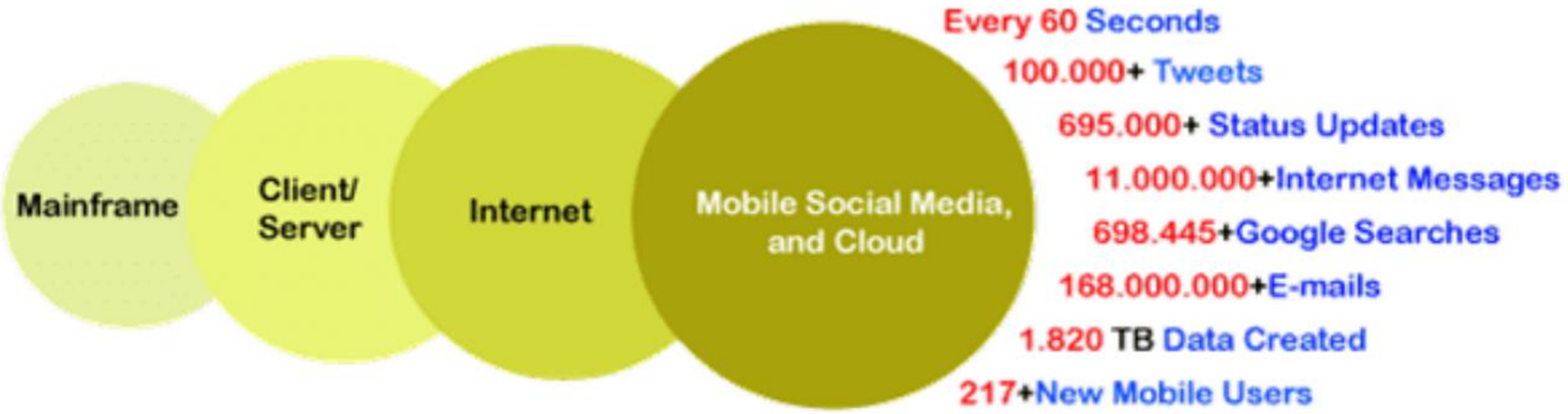
The Characteristics of Big Data -4

- Value is the major issue that we need to concentrate on. It is not just the amount of data that we store or process. It is actually the amount of valuable, reliable and trustworthy data that needs to be stored, processed, analyzed to find insights



The Characteristics of Big Data -5

- **Velocity:** Velocity plays a major role compared to the others. The major aspect of Big data is to provide data on demand and at a faster pace.



Applications of Big Data -1

- **Retail:** Leading online retail platforms are wholeheartedly deploying big data throughout a customer's purchase journey, to predict trends, forecast demands, optimize pricing, and identify customer behavioral patterns. Big data is helping retailers implement clear strategies that minimize risk and maximize profit.
- **Healthcare:** Big data is revolutionizing the healthcare industry, especially the way medical professionals in the past diagnosed and treated diseases. In recent times, effective analysis and processing of big data by machine learning algorithms provide significant advantages for the evaluation and assimilation of complex clinical data, which prevent deaths and improve the quality of life by enabling healthcare workers to detect early warning signs and symptoms.
- **Financial Services and Insurance:** The increased ability to analyze and process big data is dramatically impacting the financial services, banking, and insurance landscape. In addition to using big data for swift detection of fraudulent transactions, lowering risks, and supercharging marketing efforts, few companies are taking the applications to the next levels.
- **Government:** Cities worldwide are undergoing large-scale transformations to become "smart", through the use of data collected from various Internet of Things (IoT) sensors. Governments are leveraging this big data to ensure good governance via the efficient management of resources and assets, which increases urban mobility, improves solid waste management, and facilitates better delivery of public utility services.

Applications of Big Data -2

- **Manufacturing:** Advancements in robotics and automation technologies, modern-day manufacturers are becoming more and more data focused, heavily investing in automated factories that exploit big data to streamline production and lower operational costs. Top global manufacturers are also integrating sensors into their products, capturing big data to provide valuable insights on product performance and its usage.
- **Energy To combat** the rising costs of oil extraction and exploration difficulties because of economic and political turmoil, the energy industry is turning toward data-driven solutions to increase profitability. Big data is optimizing every process while cutting down energy waste from drilling to exploring new reserves, production, and distribution.
- **Logistics & Transportation:** State-of-the-art warehouses use digital cameras to capture stock level data, which, when fed into ML algorithms, facilitates intelligent inventory management with prediction capabilities that indicate when restocking is required. In the transportation industry, leading transport companies now promote the collection and analysis of vehicle telematics data, using big data to optimize routes, driving behavior, and maintenance.

Data Preparation

79

Basic components of artificial intelligence

- Technology:
  - Robot: Sensors, actuators
  - Computer systems: Memory, MP, Data processing speed
  - Machines (IoT)
- Data Analytics
  - Data collection
  - Database management
  - Data pre-preparation
  - Data analysis and interpretation
- Applications
  - End-user applications: Language detection, translation, discussion
  - Professional support – Assistance
  - Programming – Software
  - Algorithms – Mathematical models
- Minimizing technologies
  - Miniaturization of measurement, sensor and analysis devices
  - Smallness of computer systems
  - Material science in space (Manufacturing)
- Expertise - Team

Cleaning, transforming and integrating data

Once the data is captured, it must be cleaned and prepared for use with a visualization tool. Depending on the data you are using, this process can involve different steps, but data cleaning and transformation will almost always be the longest step in the process.

Cleaning steps of data:

- Both duplicate values and empty fields in the data table must be addressed.
- Variables or fields that are unnecessary for the research question are removed.
- Outliers or invalid data that could skew observable trends in the visualization are accounted for.
- Names and values are standardized for machine use.
- Spelling errors, mislabeling, misaligned columns, and other typos that could affect rendering are checked.
- This cleaned data can then be integrated with other datasets and other information that informs the visualization.

Recommended Tools for Data Cleaning

- OpenRefine is an open-source desktop program used for “data wrangling.” In other words, it can clean and manage data, export it, and crosswalk the format to a number of desktop and online sources. When working with spreadsheets or other data sets that may have inaccurate or missing information, OpenRefine can clean them without damaging or altering the original file. For information on how to install and use OpenRefine, visit the OpenRefine documentation & guide pages.

- Python is a programming language that can be used to clean data through a variety of packages. Utilizing the Natural Language Tool Kit or SpaCy, one can tokenize elements in Python for easy removal or replacement when cleaning data. For more information on Python, as well as a number of recommended books, please visit our Python Library Guide.

- R Studio is a desktop environment for the R language, which excels at language processing and data visualization. R Studio supports plug-ins like TidyText for a number of tokenizing and NLP processes. You can learn more about R and RStudio, as well as find a number of recommended books, in our R Library Guide & our RStudio Library Guide.

Data prep process

- **Collecting the data**: relevant data is gathered from various sources like data warehouses/lakes, operational systems, marketing platforms, etc. The data comes in all different formats and compositions. It is important to understand what data is relevant to what you are trying to achieve and whether it is accurate. This helps to ensure the other steps run smoother.
- **Profiling the data:** after you conclude what data is needed, you need to know what needs to be done to prepare the collected data after exploring. This includes studying the data to recognize the structure and content and ultimately providing a deeper insight into the data collected. This stage also is a precursor to the next stage and is a necessary step.
- **Cleaning the data:** This stage entails removing mistakes and discrepancies in order to refine its accuracy and quality. This could entail replacing null or missing numbers, fixing errors and typos, getting rid of duplicates, or addressing outliers. The strategy for cleaning data is determined by the kind of data and the particular needs of the study.
- **Transforming and integrating the data:** Once the data has been cleaned, it is necessary to convert it into a uniform format that can be used for analysis. This might involve procedures like encoding category data, standardizing numerical data, or developing new variables based on preexisting ones. To offer a comprehensive perspective of the information, data from other sources must also be merged.
- **Structuring the data:** depending on the type of visualization required, data is modeled in a certain way to reflect the end visually. Data may be organized or restructured in a variety of ways, depending on the ultimate result. This could entail constructing tables, establishing schemas, or modifying the data to accommodate certain analytical needs. For data visualization, there are multiple ways of structuring the data including long structuring, hierarchical, categorization, aggregation, and so on.

Data Visualization

84

Data Visualization

- Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. Additionally, it provides an excellent way for employees or business owners to present data to non-technical audiences without confusion.

- In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.

Why data visualization is important

- The importance of data visualization is simple: it helps people see, interact with, and better understand data. Whether simple or complex, the right visualization can bring everyone on the same page, regardless of their level of expertise.

- It’s hard to think of a professional industry that doesn’t benefit from making data more understandable. Every field benefits from understanding data—and so do fields in government, finance, marketing, history, consumer goods, service industries, education, sports, and so on.

- While we’ll always wax poetically about data visualization there are practical, real-life applications that are undeniable. And, since visualization is so prolific, it’s also one of the most useful professional skills to develop. The better you can convey your points visually, whether in a dashboard or a slide deck, the better you can leverage that information. The concept of the citizen data scientist is on the rise. Skill sets are changing to accommodate a data-driven world. It is increasingly valuable for professionals to be able to use data to make decisions and use visuals to tell stories of when data informs the who, what, when, where, and how.

- While traditional education typically draws a distinct line between creative storytelling and technical analysis, the modern professional world also values those who can cross between the two: data visualization sits right in the middle of analysis and visual storytelling.

Data visualization and big data

- As the “age of Big Data” kicks into high gear, visualization is an increasingly key tool to make sense of the trillions of rows of data generated every day. Data visualization helps to tell stories by curating data into a form easier to understand, highlighting the trends and outliers. A good visualization tells a story, removing the noise from data and highlighting useful information.

- However, it’s not simply as easy as just dressing up a graph to make it look better or slapping on the “info” part of an infographic. Effective data visualization is a delicate balancing act between form and function. The plainest graph could be too boring to catch any notice or it make tell a powerful point; the most stunning visualization could utterly fail at conveying the right message or it could speak volumes. The data and the visuals need to work together, and there’s an art to combining great analysis with great storytelling.

General Types of Visualizations

- Chart: Information presented in a tabular, graphical form with data displayed along two axes. Can be in the form of a graph, diagram, or map.
- Table: A set of figures displayed in rows and columns. Learn more.
- Graph: A diagram of points, lines, segments, curves, or areas that represents certain variables in comparison to each other, usually along two axes at a right angle.
- Geospatial: A visualization that shows data in map form using different shapes and colors to show the relationship between pieces of data and specific locations.
- Infographic: A combination of visuals and words that represent data. Usually uses charts or diagrams.
- Dashboards: A collection of visualizations and data displayed in one place to help with analyzing and presenting data.

Visual analysis

- Data Analytics Experts provide visual analytics to make sense of the unspoken.
- Extracting visual narratives from a mountain of data
- Images capture unspoken experiences
- If a picture is worth a thousand words, consider the billions of photos shared publicly by consumers online, capturing every conceivable consumer experience from multiple perspectives.
- New AI-powered methods are allowing us to organize and map such images to find useful patterns and see the world through consumers' eyes. Using cutting-edge deep learning techniques, a visual theme landscape emerges from the bottom up.

Data Analysis

90

Data Analysis -1

- Data analytics is the practice of working with data to gather useful information that can then be used to make informed decisions.
- Data analysts use data to solve problems. As such, the data analytics process typically goes through several iterative stages.

- Extracting meaning from data allows us to make better decisions. And we live in a time when we have more data at our fingertips than ever before. As such, companies are taking advantage of the benefits of data and turning to data analytics to find insights into further business objectives.
- The World Economic Forum’s Future of Jobs Report 2023 listed data analysts and scientists as some of the most in-demand jobs, alongside AI and machine learning experts and big data experts.
- Identify the business question you want to answer. What problem is the company trying to solve? What do you need to measure, and how will you measure it?
- Gather the raw datasets you’ll need to answer the question you’ve identified. Data collection can come from sensors, meters, I/O units, corporate memory, internal sources such as a company's customer relationship management (CRM) software, or secondary sources such as government records or social media application programming interfaces (APIs).

Data Analysis -2

- To prepare for analysis, the data is cleaned. This typically involves cleaning out duplicate and anomalous data, reconciling inconsistencies, standardizing the data structure and format, and dealing with white space and other syntax errors.

- The data is analyzed. By manipulating the data using a variety of data analysis techniques and tools, you can begin to find trends, correlations, outliers, and variations that tell a story. At this stage, you can use data mining to discover patterns in databases or data visualization software to help transform the data into an easy-to-understand graphical format.

- Interpret the results of your analysis to see how well the data answers your original question. What recommendations can you make based on the data? What are the limitations of your conclusions?

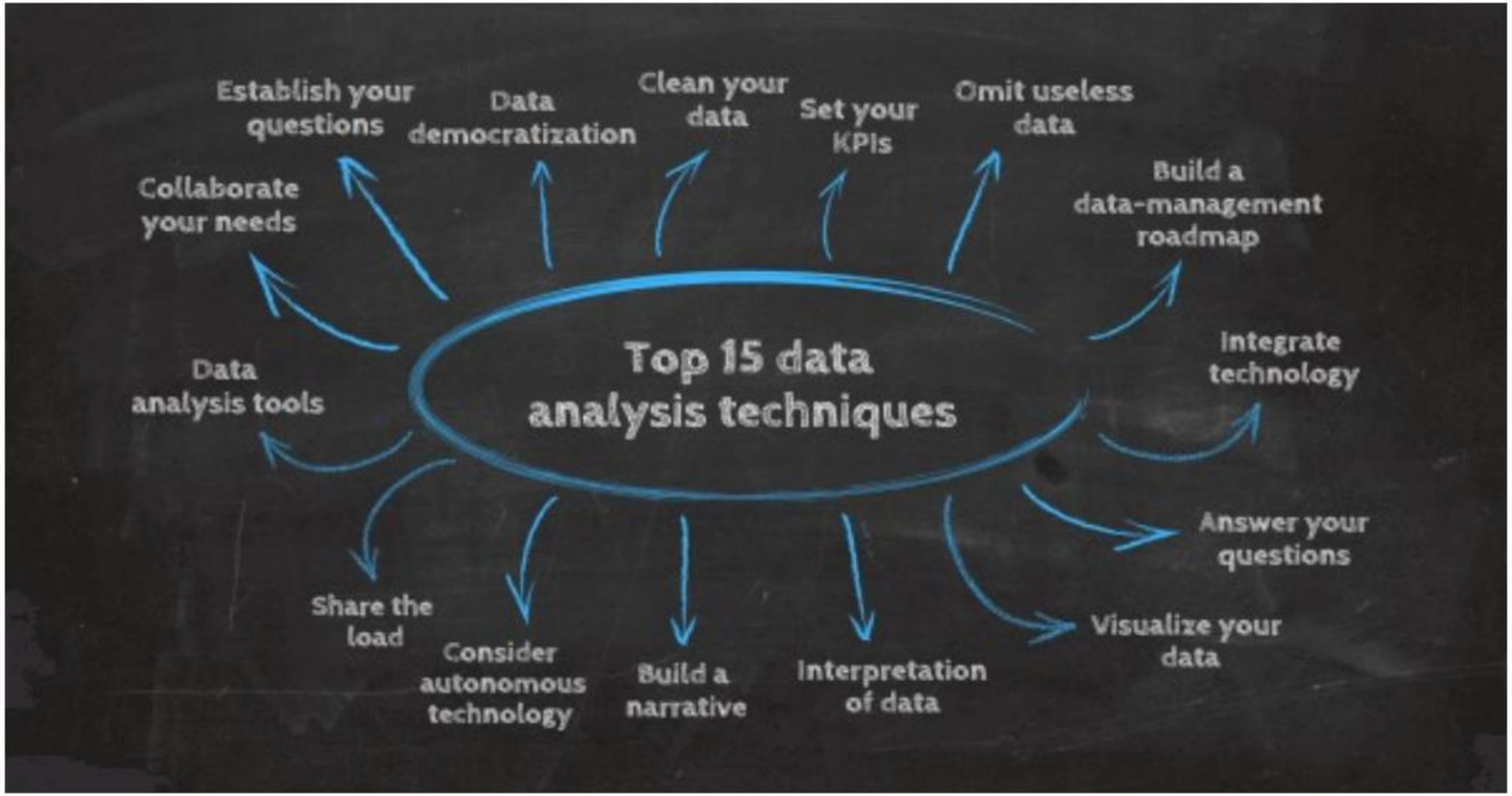
- Generative AI, Data Analysis, Querying Databases, Data Generation, Data Augmentation, Data Science, Python Programming, Pandas, Jupyter notebooks, Numpy, Dashboards and Charts, dash, Data Visualization, Matplotlib, Dashboard, SQL and RDBMS, Model Selection, Predictive Modeling, Microsoft Excel, IBM Cognos Analytics, Spreadsheet, Pivot Table, Cloud Databases, Relational Database Management System (RDBMS), SQL

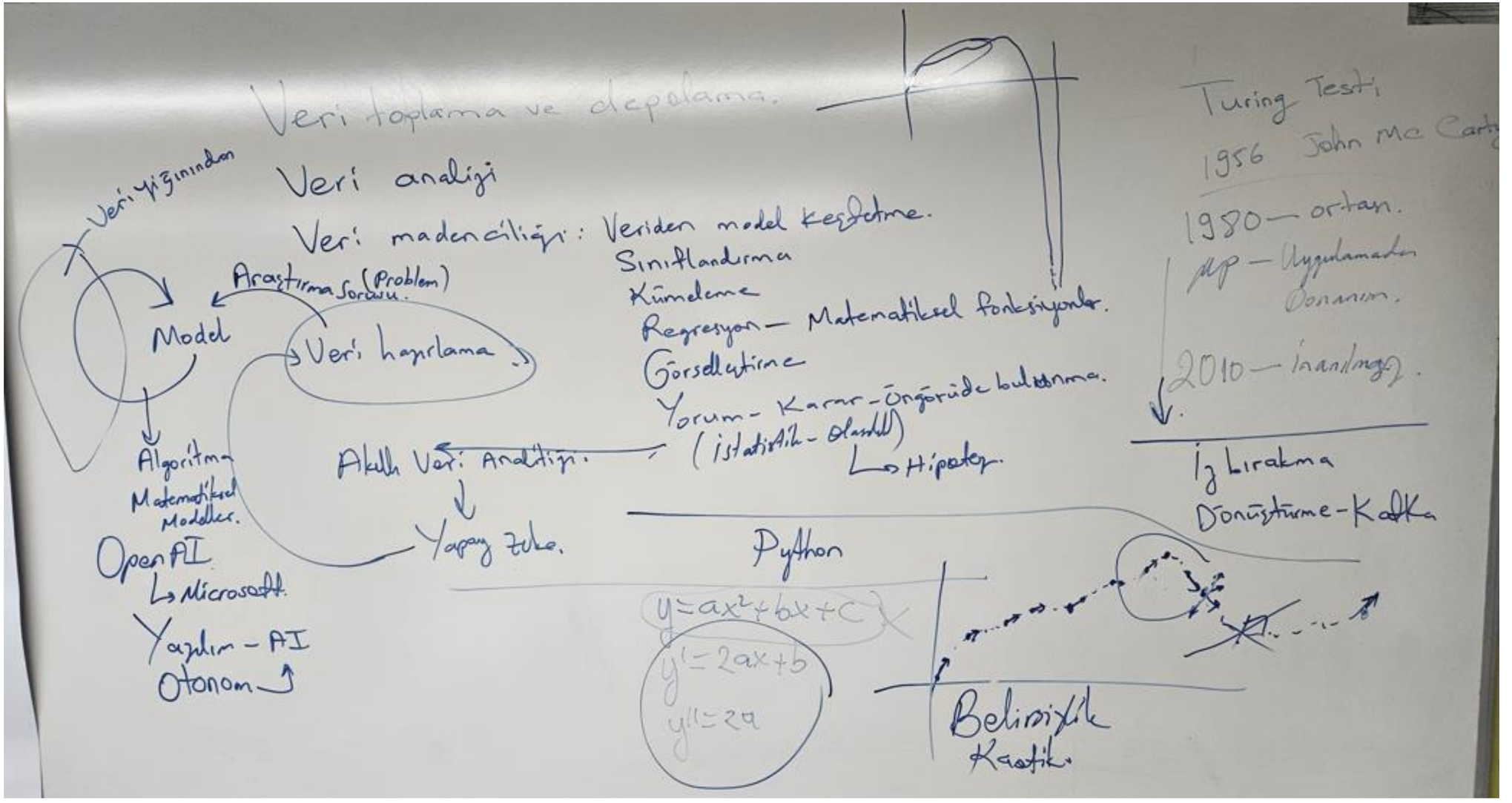
What is big data analysis?

- Big data refers to large data sets that are constantly being generated at high speeds and volumes, with a wide variety of types, including structured, unstructured, and semi-structured data. Big data is typically measured in terabytes or petabytes. One petabyte is equal to 1,000,000 gigabytes. To make sense of this, consider that a single HD movie contains about 4 gigabytes of data. One petabyte is equivalent to 250,000 movies. Large data sets can range in size from hundreds of petabytes to thousands or even millions of petabytes.
- Big data analytics is the process of finding patterns, trends, and relationships within large data sets. These complex analyses require specific tools and technologies, computing power, and data storage capacity to support that scale.

- Big data analytics follows five steps to analyze any large data set:
  - Data collection
  - Data storage
  - Data processing
  - Data cleaning
  - Data analysis

Data Analysis Techniques




Veri toplama ve depolama.
Veri analizi
Veri madenciliği: Veriden model keşfetme.
Sınıflandırma
Kümeleme
Regresyon — Matematiksel fonksiyonlar.
Görselleştirme
Yorum - Karar - Öngörüde bulunma.
(İstatistik - Olasılık)
Hipotez.
Veri yığınından
Araştırma Sorusu (Problem)
Model
Veri hazırlama
Algoritma
Matematiksel Modeller.
Akıllı Veri Analitiği.
Yapay Zeka.
OpenAI
Microsoft
Yazılım - AI
Otonom
Turing Testi
1956 John Mc Carth
1980 — ortası.
Donanım.
2010 — İnanılmaz.
İz Bırakma
Dönüştürme - Kafka
Python
$y = ax^2 + bx + c$
$y' = 2ax + b$
$y'' = 2a$
Belirsizlik
Kaotik.

Data Analysis

- Data analysis: This is the step where raw data is converted into actionable insights. The four types of data analysis are as follows:
- 1. Descriptive analysis: Data scientists analyze data to understand what happened in the data environment in the past or present. It is characterized by data visualizations or constructed explanations such as pie charts, bar charts, line graphs, tables.
- 2. Diagnostic analysis: Diagnostic analysis is the process of in-depth or detailed data analysis to understand why something happened. It is characterized by techniques such as drilling down into details, data exploration, data mining, and correlations. Each of these techniques uses multiple data operations and transformations to analyze the raw data.
- 3. Predictive analysis: Predictive analysis uses past data to make accurate predictions about future trends. It is characterized by techniques such as machine learning, forecasting, pattern matching, and predictive modeling. In each of these techniques, computers are trained to reverse engineer causal relationships in the data.
- 4. Prescriptive analysis: Prescriptive analysis takes predictive data to the next level. It not only predicts the likely outcome but also suggests the ideal response for that outcome. It can analyze the potential outcomes of different choices and recommend the best course of action. It is characterized by graphical analysis, simulation, complex event processing, neural networks, and recommendation infrastructures.

4 types of data analysis:

- Data can be used to answer questions and support decisions in many different ways. In order to determine the best way to analyze your data, it's helpful to have some knowledge of the different types of analysis most commonly used in this field.
- 1. Descriptive analysis: Descriptive analysis tells us what happened. This type of analysis helps describe or summarize quantitative data by providing statistics. For example, descriptive statistical analysis can show the distribution of sales among a group of employees and the average sales per employee.
- Descriptive analysis answers the question: "what happened?"
- 2. Diagnostic analysis: If descriptive analysis determines the "what," diagnostic analysis determines the "why." Let's say a descriptive analysis shows an unusual influx of patients at a hospital. Further examination of the data may reveal that many of these patients share symptoms of a particular virus. This diagnostic analysis can help you determine whether an infectious agent—the "why"—is causing the influx of patients. Diagnostic analysis answers the question: "why?"
- 3. Predictive analytics: So far, we've looked at types of analytics that examine the past and draw conclusions about the outcome. Predictive analytics uses data to create predictions about the future. Using predictive analytics, you might notice that a particular product sees its best sales in September and October each year, leading you to predict a similar high point for the upcoming year.
- Predictive analytics answers the question: "What could happen next?"
- 4. Prescriptive analytics: Prescriptive analytics takes all the insights gathered from the first three types of analytics and uses them to create recommendations for how a company should proceed. Using our previous example, this type of analytics might suggest a market plan to build on the success of high sales months and capitalize on new growth opportunities in slower months.
- How does big data analytics work?

Data Analysis Stages

- Analysis, Determining what all or a few of the components that make up a whole are, and what each one changes in the whole. Analysis, Analysis...
- The main purpose of collecting, processing and organizing data, and displaying it in the form of tables or graphs is to respond to queries and make predictions about the future.
- In order to obtain information about errors, instability, uncertainty and variables with statistical operations, measurement and comparison are made using probability calculation techniques, and predictions are made. Statistics is a science of uncertainty.
- Statisticians are not interested in the question "What is?", but in the questions "What could happen?" or "What is likely?"
- The collection, compilation, summarization, presentation, analysis of data and also the drawing of a valid conclusion are the main areas of interest of the branch of statistics.
- In perception, which initiates the ability to produce solutions to problems by learning, numerical measurements must be grouped, summarized and interpreted using various statistical techniques.

Data Analysis

- Data analysis is defined as a process of cleaning, transforming, and modeling data to discover useful information for business decision making.
- The purpose of Data Analysis is to extract useful information from data and taking the decision based upon the data analysis.
- Data Analysis – Types: There are several types of Data Analysis techniques that exist based on business and technology.
- However, the major Data Analysis methods are:
  - Text Analysis
  - Statistical Analysis
  - Diagnostic Analysis
  - Predictive Analysis
  - Prescriptive Analysis

Textual Analysis

- Text analytics is the process of using computer systems to read and understand human-written text for business insights. Text analytics software can independently classify, sort, and extract information from text to identify patterns, relationships, sentiment, and other actionable information. You can use text analytics to process multiple text-based sources, such as emails, documents, social media content, and product reviews, as efficiently and accurately as if they were written by a human.
- Sentiment analysis, or opinion mining, uses text analytics methods to understand the sentiment conveyed in a piece of text. You can perform sentiment analysis on product reviews, blogs, forums, and other online media sources to determine whether your customers are happy with the product or service they purchased. Sentiment analysis helps you discover new trends, track sentiment changes, and address PR issues. By using sentiment analysis and identifying specific keywords, you can track changes in customer opinions and identify the root cause of the problem.
- The basis of text analytics is training computer software to match words with specific meanings and understand the semantic context of unstructured data. This training is similar to how people associate words with objects, actions, and emotions when learning a new language.
- Text analytics software is based on deep learning and natural language processing principles.
- Steps of text analytics: Data collection, Data preparation, Text analysis (classification, Extraction), Visualization

Understanding a language

- Understanding a language is one of the biggest challenges in designing artificial intelligence. Languages are detailed and complex, and present many challenges for computers with fixed rules. Language can also be easily misinterpreted.
- Machines have communicated with us without any difficulty.
- But it is really very difficult for a computer to understand the message that is intended to be given in the language spoken.
- This problem is solved by natural language processing. We try to find answers to questions such as what are the meanings of some words in a sentence? What are the relationships between sentences? What is the connection of one word to another and why.
- Understanding language is a very easy area for us, but very difficult for computers.
- A computer can read millions of articles in seconds.
- It is very difficult for computers to understand the question in order to argue with humans. We understand the question very easily, but our memory capacity is not developed enough to give valid answers.
- As humans, we are so used to understanding language that we do not even realize how difficult it is. It is very easy for humans to establish arguments and connect things together.
- When we try to teach a computer a language, we can only use the building blocks.
- We can teach a computer system to understand parts of speech and identify concepts, and by doing so, we can make it see similarities between concepts and sentences.
- To do this, we need to teach the system to connect things together in clever ways.

Discussing with the computer

- When a computer is given a topic of discussion, the system first tries to understand the meaning of the topic.
- Then, it scans millions of articles to identify potential arguments that it can use to construct a defense.
- It uses unique language processing, machine learning, and reasoning techniques to understand the underlying themes of the discussion and to organize its argument effectively and persuasively.
- It constructs sentences and conveys its own thinking.
- Then, the computer system must listen to its opponent for a certain period of time.
- If the opponent puts forward complex arguments such as moral reasons, the system must somehow understand all of this and provide a proper response.
- After listening, it constructs a rebuttal argument. It puts forward its own argument on a different level.
- If the person in front of the computer system talks about the negatives that the arguments put forward will bring and then lists the steps to be taken to prevent the negatives.

Discussing with the computer

- The computer opposes the steps to be taken to solve the negativities. Or it puts forward different approaches. It starts to put forward different solution methodologies. If it needs to be accepted, it draws the broad perspective of the steps to be taken.
- Computer systems improve themselves a little more after each discussion. It learns more and more to identify the points that it needs to emphasize in order to claim that it is different from humans and right.
- Over time, it should be switched from searching to research. Because when you search for something, it is desired to reach the documents that are suitable for your search title. It helps to find the research topic by going one layer further. When you start researching the topic, the positive and negative aspects of the topic are also obtained.
- Incredible benefits can be obtained from discussion technologies in the future.
- An artificial intelligence system that can structure a convincing argument can change the way we make decisions as a society.

Data Preparation in Text Analysis

- Data preparation is an important part of text analysis. It involves structuring the raw text data into an acceptable format for analysis. Text analysis software automates this process and includes the following common natural language processing (NL) methods.
- Tokenization is the separation of raw text into multiple semantic parts. For example, the phrase businesses benefit from text analysis is broken down into tokens such as businesses, text, analysis, and benefits.
- In the method of tagging sentence elements, grammatical labels are assigned to the text that is broken down into tokens. For example, when this step is applied to the tokens mentioned above, businesses: noun; text: noun; analysis: noun; benefits: verb is the result.
- In the parsing process, meaningful connections are established between the words that are broken down into tokens and English grammar. It helps the text analysis software visualize the relationship between words.
- Stemming is a linguistic process in which words are reduced to their root (lemma) in the dictionary according to the meaning they are used in the sentence. For example, the dictionary form of the word visualization is visualization.
- Noise words are words that add little or no semantic context to a sentence, such as and, or, for. Depending on the usage scenario, the software can remove such words from structured text.

Statistical Analysis

- Statistics is the branch of science that uses probabilities as a basis to influence the possible outcomes of situations determined by numerical data during the collection, interpretation and validation of numerical data.
- Statistical analysis occurs when we collect and interpret data in order to identify patterns and trends.
- From the creation of the idea of what is to be investigated, to the definition of objectives, hypotheses, variables, to the collection, organization, examination, classification, tabulation and production of results for the analysis of data, it is important to know how to use different measurements and statistical models appropriately for analysis.
- By making a simpler interpretation through the analysis and categorization of a range of data from qualitative to quantitative data, we will be able to manipulate data and adjust situations in certain contexts with appropriate decision making. In general, statistics will help in:
  - Identifying unnoticed trends.
  - Adding objectivity to the decision-making process.
  - No need to make instinctive decisions.
  - Reducing operating costs.
  - Conducting market analysis.

How to perform a functional statistical analysis?

- A clear and realistic description of the available data is provided.
- Analyze how the data relates to the study subjects.
- A model is designed that considers and explains the relationship between the data and the study subjects.
- The model is evaluated to determine its validity.
- Scenarios and tests are evaluated using predictive analytics.

Types of Statistical Analysis

Five basic types of statistical analysis are used to simplify data sets.

- Descriptive Analysis.
- Inferential Analysis.
- Difference Analysis.
- Relationship Analysis.
- Prediction Analysis.

Statistical Analysis Techniques

- Design and Planning of the Study
- Descriptive Statistics.
- Confidence Interval Estimates.
- Hypothesis Tests Applied to Three or More Groups.
- Simple Regression.
- Correlation Analysis.
- Analysis of Covariance.

Insights (İçgörüler)

- Consumer insights: View a phenomenon or event through the eyes of consumers and identify strategy, innovation, and marketing opportunities
- Insert insights: Understand how people view and describe places to find gaps and opportunities for destination development and placemaking.
- Brand scans: Map how products and brands are used and delivered by their users, and feed into product development and branding.
- Trend spotting: Discover what trend setters think is cool from global hotspots without having to go there.

Different types of analysis

- There are many ways of categorizing data analysis – far more than can be described in this paper. One way is to categories it according to the type of data collected. (Note that many organizations, projects and programmers use a combination of different types of data analysis).
- Quantitative data analysis is used to analyze numbers rather than words. It can range from simple exercises to process and tabulate data through to very complicated processes designed to accurately measure quantitative changes with calculated degrees of precision.
- Qualitative data analysis, on the other hand, is used to analysis words – quotes, cases, transcripts, reports – and, sometimes, images. Qualitative methods rely on rules and processes which are very different from those of quantitative methods.
- Some analysis methodologies are designed to translate qualitative data into quantitative information through rating or scaling exercises. This involves developing ratings or scales based on qualitative analysis, and then processing them through quantitative methods.
- Participatory data analysis can involve quantitative or qualitative data analysis, and is often treated as a separate case. This is because participatory data analysis follows different rules, and is usually based on stakeholders' sense making and consensus rather than rigorously applied methods. The purpose of participatory analysis may also be quite different –encouraging stakeholders to analyze their own situations rather than coming to a conclusion based on an external viewpoint.

Another way of categorising data analysis

- Descriptive data analysis is only concerned with processing and summarising data. This is often true offinancial or administrative data analysis.
- Theory driven data analysis is used to test theories of change, assumptions or hypotheses. The aim is to analyse data to see if it confirms (or not) the theory or hypothesis.
- Data or narrative driven analysis involves letting patterns emerge from data, and then developing theories afterwards.

Steps in Data Analysis

The process involved in data analysis involves several steps:

- 1. Determine the data requirements or how the data will be grouped. Data can be broken down by age, demographics, income, or gender. Data values can be numeric or broken down by category.
- 2. Collect the data. This can be done through a variety of sources, such as computers, online sources, cameras, environmental sources, or personnel.
- 3. Once collected, organize the data so that it can be analyzed. This can happen in a spreadsheet or other software format that can capture statistical data.
- 4. Clean the data before it can be analyzed. This is done by cleaning the data and making sure there are no duplicates, errors, or omissions. This step helps correct errors before the data goes to a data analyst for analysis.

Intelligent Data Analysis (IDA)

What is Data Intelligence?

- Data intelligence refers to the practice of using artificial intelligence and machine learning tools to analyze and transform massive datasets into intelligent data insights, which can then be used to improve services and investments. The application of data intelligence tools and techniques can help decision makers develop a better understanding of collected information with the goal of developing better business processes.

- The five major components of data driven intelligence are descriptive data, prescriptive data, diagnostic data, decisive data, and predictive data. These disciplines focus on understanding data, developing alternative knowledge, resolving issues, and analyzing historical data to predict future trends. Some industries with the greatest need for data intelligence include cybersecurity, finance, health, insurance, and law enforcement. Intelligent data capture technology is a valuable application in these industries for transforming print documents or images into meaningful data.

- Intelligent data is a core component of big data and business intelligence. Intelligent data processing provides a strong data foundation, restructuring and enhancing big datasets that AI uses; cleanses and transforms data into information that is valuable and relevant to business performance; enables businesses to identify patterns, make informed decisions, and adapt to new information; and incorporate advanced analytics techniques to enhance visualized prescriptive and predictive analytics.

Intelligent Data Analysis delivers expert-based knowledge…

- Intelligent Data Analysis (IDA) is an interdisciplinary study that is concerned with the extraction of useful knowledge from data, drawing techniques from a variety of fields, such as artificial intelligence, high-performance computing, pattern recognition, and statistics. Data intelligence platforms and data intelligence solutions are available from data intelligence companies such as Data Visualization Intelligence, Strategic Data Intelligence, Global Data Intelligence.

- Intelligent data analysis refers to the use of analysis, classification, conversion, extraction organization, and reasoning methods to extract useful knowledge from data. This data analytics intelligence process generally consists of the data preparation stage, the data mining stage, and the result validation and explanation stage.
-
- Data preparation involves the integration of required data into a dataset that will be used for data mining; data mining involves examining large databases in order to generate new information; result validation involves the verification of patterns produced by data mining algorithms; and result explanation involves the intuitive communication of results.

IDA or …

- Data mining
- Obtaining knowledge from data
- Rule discovery based on genetic algorithm
- Knowledge discovery
- Learning classifier system
- Machine learning etc.

Knowledge acquisition …

- *The process of extracting, analyzing, transforming, classifying, organizing, and integrating information and representing that information in a form that can be used in a computer system.*

Data Analytics

118

How will you discover the rules hidden in the data?

Data Analytics Definition

- Data analytics is a discipline that focuses on the tools and techniques used to collect, organize, and store data to analyze it and extract insights from it.
- It comprises the processes, tools and techniques of data analysis and management, including the collection, organization, and storage of data.
- The aim of data analytics is to apply statistical analysis and technologies on data to find trends and solve problems.
- Data analytics has become increasingly important in the enterprise as a means for analyzing and shaping business processes and improving decision-making and business results.
- Data analytics draws from a range of disciplines — including computer programming, mathematics, and statistics — to perform analysis on data in an effort to describe, predict, and improve performance.
- To ensure robust analysis, data analytics teams leverage a range of data management techniques, including data mining, data cleansing, data transformation, data modeling, and more.

What is Data Analytics?

- It is the development of mathematical models that produce predictions and solutions.
- It is the science of analyzing raw data to obtain results from a mass of data. It is to obtain autonomous decision-making processes from mathematical models and algorithms.
- It is to reveal strategically important trends and measurements that may be lost or overlooked in the mass of information.
- It is the simulation and optimization of the performance of processes.
- It consists of analysts and machine learning engineers.
- It is the development of mathematical models that produce predictions and solutions.
- It is to obtain autonomous decision-making processes from mathematical models and algorithms that work on raw data.
- It is the use of statistical models and forecasting techniques to understand the future. Predictive Analytics: "What might happen?"
- It is the use of optimization and simulation algorithms to answer questions and make recommendations about possible outcomes. Predictive Analytics: "What should we do?"

Data Analytics Expert-1

- Compare and match multiple data sources and access powerful machine learning tools in one place.
- Trigger complex operations with simple user interactions and iterate rapidly using visual analytics.
- Leverage fast, distributed, fully parallelized computing at scale in the cloud to optimize datasets of any size.
- Analytics professionals
- Market researchers and competitive intelligence professionals use analytics to extract patterns and insights from social media, news, and other text sources.
- Data processors, analysts, text analysts, and academic researchers use analytics capabilities to accelerate preprocessing, data wrangling, natural language processing, text enrichment, and annotation.
- Engineers and data scientists use analytics to train text classifiers and embed automated natural language processing workflows into their algorithms and applications.

Data Analytics Expert-2

- Programming skills: Knowledge of programming languages such as Matlab, R, and Python is a must for any data analyst.
- Statistics, probability, and applied and computational mathematics: Descriptive and inferential statistics and experimental designs are a must for data scientists.
- Knowledge of machine learning algorithms and software applications.
- Data processing skills: Ability to map, classify, cluster, and transform raw data into information in a more appropriate way.
- Has acquired communication and data visualization skills.
- Has experiential data intuition: It is essential for a professional to think like a data analyst.

Concepts in Data Analytics

- Theory: These are hypotheses that have not been proven to be true, nor have they been refuted, and are constantly supported by new studies. A theory usually consists of two elements. a) Axiom b) Hypothesis
- Hypothesis: These are claims, assumptions or propositions that are put forward regarding the research topic; the truth of which needs to be investigated. Hypotheses are basic ideas, assumptions and propositions that guide research. They are solutions that are expected to lead to the correct result of a problem.
- Axiom: These are general principles that have not been tested and whose validity and accuracy are accepted.
- Finding: These are the results obtained by processing the raw data obtained.
- Fact: A proposition or an expected action that has been proven to be true.
- Event: These are the cases that constitute facts. For example; It is a fact that it will rain, and it is an event that it will happen on Tuesday.
- Induction: This is called deriving general propositions from individual facts, finding general principles and laws.

Expertise in Data Analytics

- Programming skills: Knowledge of programming languages such as Matlab, R, and Python, etc. is essential for any data analyst.
- Statistics, probability, and applied and computational mathematics: Descriptive and inferential statistics and experimental designs are a must for data scientists.
- Machine learning algorithms and software applications
- Data processing skills: Ability to map, classify, cluster, and transform raw data into information in a more appropriate way.
- Communication and data visualization skills
- Data intuition experiential ability: It is essential for a professional to think like a data analyst.

Research topics include:

- Exploratory Data Analysis
- Time Series Analysis
- Machine Learning and Deep Learning
- Anomaly detection
- Computer Vision
- Natural Language Processing
- Large Language Models

Why data analytics?

- Observing by gathering information.
- Responding to a question with numbers.
- Commenting on the current situation.
- Making an estimation or prediction about what is happening.
- Making a prediction for the future.

- In making a decision, the common sense developed for problem solving is used to choose one of the alternatives, while choosing among the solution alternatives for the problems. Making a choice, called decision making, is mostly related to problem solving. Therefore, making a decision is taking a risk.

Why Data Analytics?

- Staying ahead in a rapidly changing world
- A data-driven approach to scanning the contextual environment
- Keeping up with developments in technology, science, business, society, and the regulatory landscape is a daunting yet critical task for any future-focused organization.
- Separating signal from noise in a flood of information is notoriously difficult.
- Data Analytics professionals make it easier for competitive intelligence (CI) professionals to extract insights from a wealth of publicly available information, such as patents, scientific articles, policy documents, and news.
- Text analytics is a powerful way to extract useful information about the business environment from text sources.
- The technology landscape through patent analysis
- Analyze cutting-edge scientific research and identify researchers and institutions through mining academic articles
- Accelerate text coding by iteratively training an AI-based curation loop to find all relevant sections
- Match narratives and themes in text data

Data Analytics Matters

- Data analytics is important because it helps businesses optimize their performance. Applying it to a business model means that companies can help reduce costs by identifying more efficient ways to do business and storing large amounts of data.

- A company can also use data analytics to make better business decisions and help analyze customer trends and satisfaction, which can lead to new and better products and services.

- Data analytics forms the basis of many quality control systems in the financial world, including the ever-popular Six Sigma program. If you can’t measure something right — whether it’s your weight or the number of defects per million on a production line — it’s nearly impossible to optimize it.

- If you can’t measure something right, you can’t manage it

Qualitative analysis in data analytics

- Using AI to accelerate qualitative research
- Mapping narratives and themes in text data
- Rapidly finding clusters of meaning in text through a bottom-up text landscape
- Accelerating text coding by iteratively training an AI-based curation loop to find all relevant passages
- With some meanings and concepts that are only perceptible to humans, qualitative research is here to stay. However, AI-powered analytics can structure and radically accelerate the qualitative research process.
- Self-service text analytics with all the capabilities and data sources you need in one place.
- A complete suite of powerful and accurate text analytics to power your solution

Big Data and Machine Intelligence?

Big Data

- Big data analysis is the analysis of data by developing new techniques and algorithms to manage the scale, variety and complexity in the data mass, to create awareness by obtaining value and information.
- Volume: scale of data
- Velocity: analysis of streaming data
- Variety: numerous forms of data
- Veracity: mitigating uncertainty of data
- Accuracy: reducing data uncertainty

Machine Learning

These are algorithms that continuously improve performance by learning interactions, changes and deviations in both data and people to make autonomous decisions and predictions.

Machine Intelligence

Machine Intelligence is an autonomous entity that can observe its environment, navigate and make decisions like humans.

Machine Learning and Data Analytics

I. Machine learning and data analysis tasks

II. Classification
  - Classification tasks
  - Building a classifier
  - Evaluating a classifier

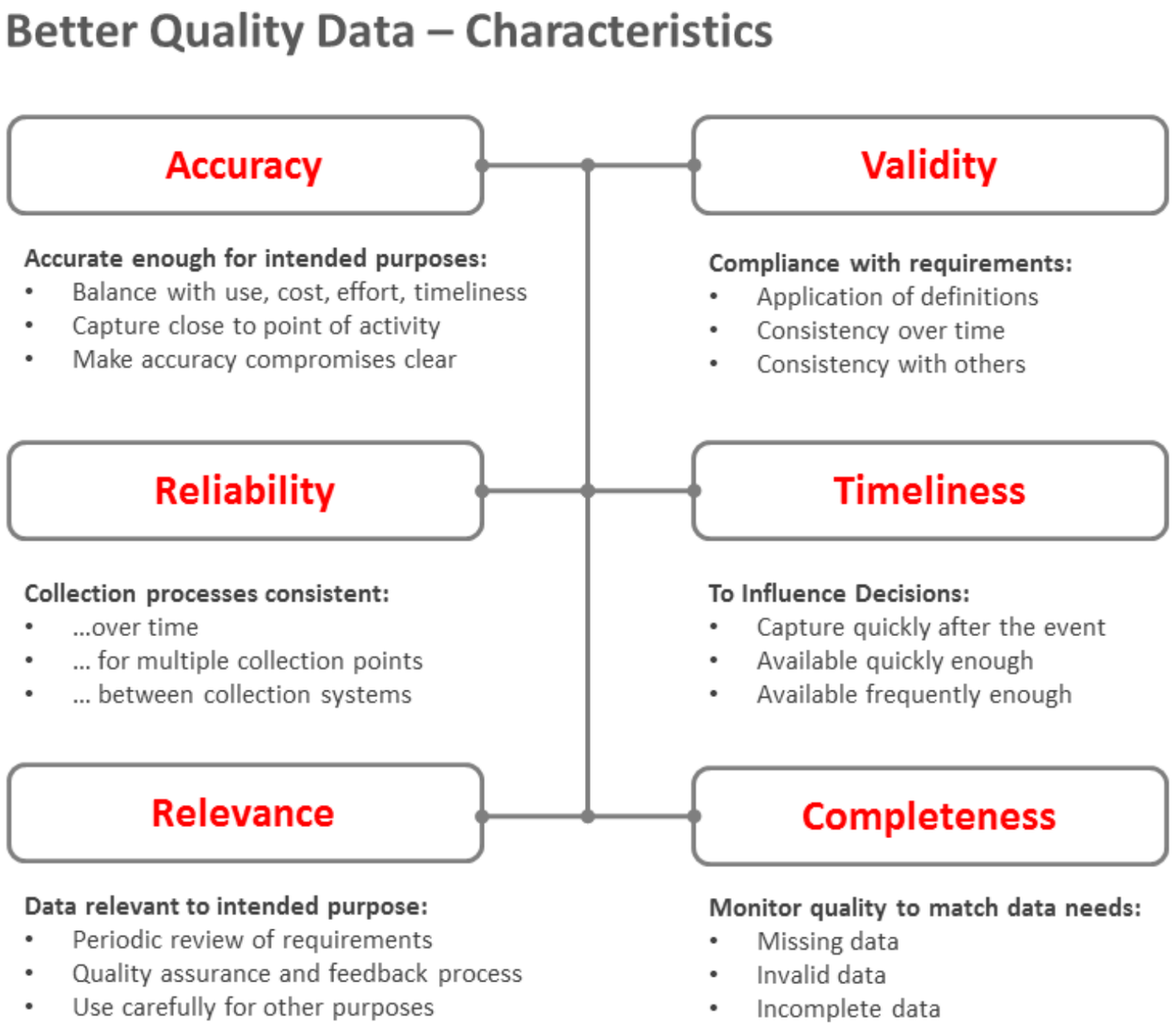
III. Pattern learning and clustering
  - Pattern detection, Quantum Radar
  - Pattern learning and pattern discovery
  - Clustering, K-means clustering

IV. Causal discovery
  - Correlation
  - Causation
  - Causal models
    - Bayesian networks
    - Markov networks

V. Simulation and mathematical modeling

VI. Practical use of machine learning and data analysis

Better Quality Data



Accuracy: Doğruluk
Validity: Geçerlilik
Reliability: Güvenilirlik
Relevance: Uygunluk
Timeliness: Yerindelik, zamanında
Completeness: Bütünlük, eksiksizlik
Compromise: taviz, ödün
İnfluence: etki
Assurance: sigorta, güven, güvence

What is Behavior Analysis?

- Behavioral analytics is a field of data analysis that focuses on providing insight into people's actions, often as they relate to online decision making.
- Behavioral analytics is used in e-commerce, gaming, social media, and other applications to identify opportunities to optimize for specific business outcomes.
- Behavioral analytics includes demographic and geographic data, but it also goes deeper by profiling a user's past activity and pulling in additional available data.
- Behavioral analytics is used to track user preferences and serve or direct that user to targeted content.
- It is often used to direct potential customers to specific products or ads.
- Some consider the systems put in place to collect data to be harmful and intrusive, concerned that everything they do is being tracked and observed.

Behavior Analytics

- Behavioral analytics is based on hard data. It uses volumes of raw data that people use when they are on social media, in gaming apps, in marketing, on retail sites or in apps. This data is collected and analyzed, and then used as the basis for making certain decisions, including how to determine future trends or business activities, including ad placement.

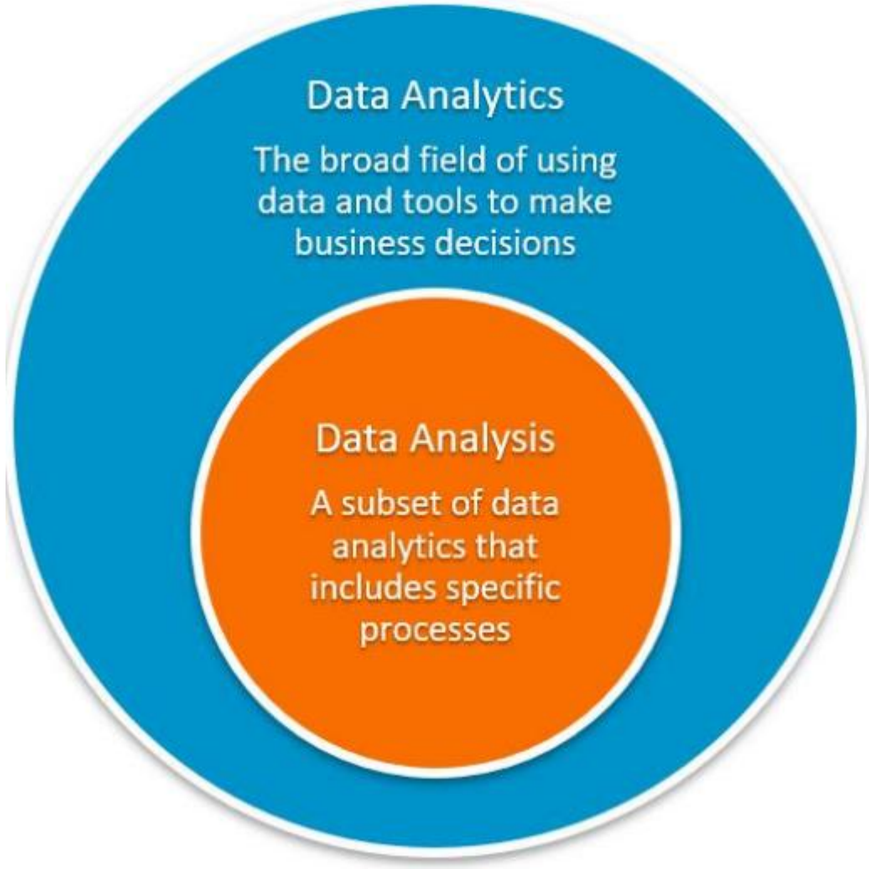
- However, there is a lot of uncertainty about the nature of the insights it provides. For example, online advertisers use behavioral analytics to help them prepare the right offer at the right time. This is often done using the user's demographic data, any past search or social information, and a local market to place the user into a larger group, sometimes called a cohort or demographic. The user is then presented with ads or offers that match the ads and offers that have the highest success rate with that group.

- Behavioral analytics can support a number of different hypotheses, so the process of elimination comes from experimentation and evaluation. Businesses often want to increase conversions, so if a change makes things worse, that hypothesis can be discarded in favor of a different hypothesis or not made at all.

- Behavioral analytics is often used to inform A or B testing, where one variable is changed at a time. As behavioral analytics deepens and the technology to test multiple changes in real time improves, companies are getting much better at targeting customers.

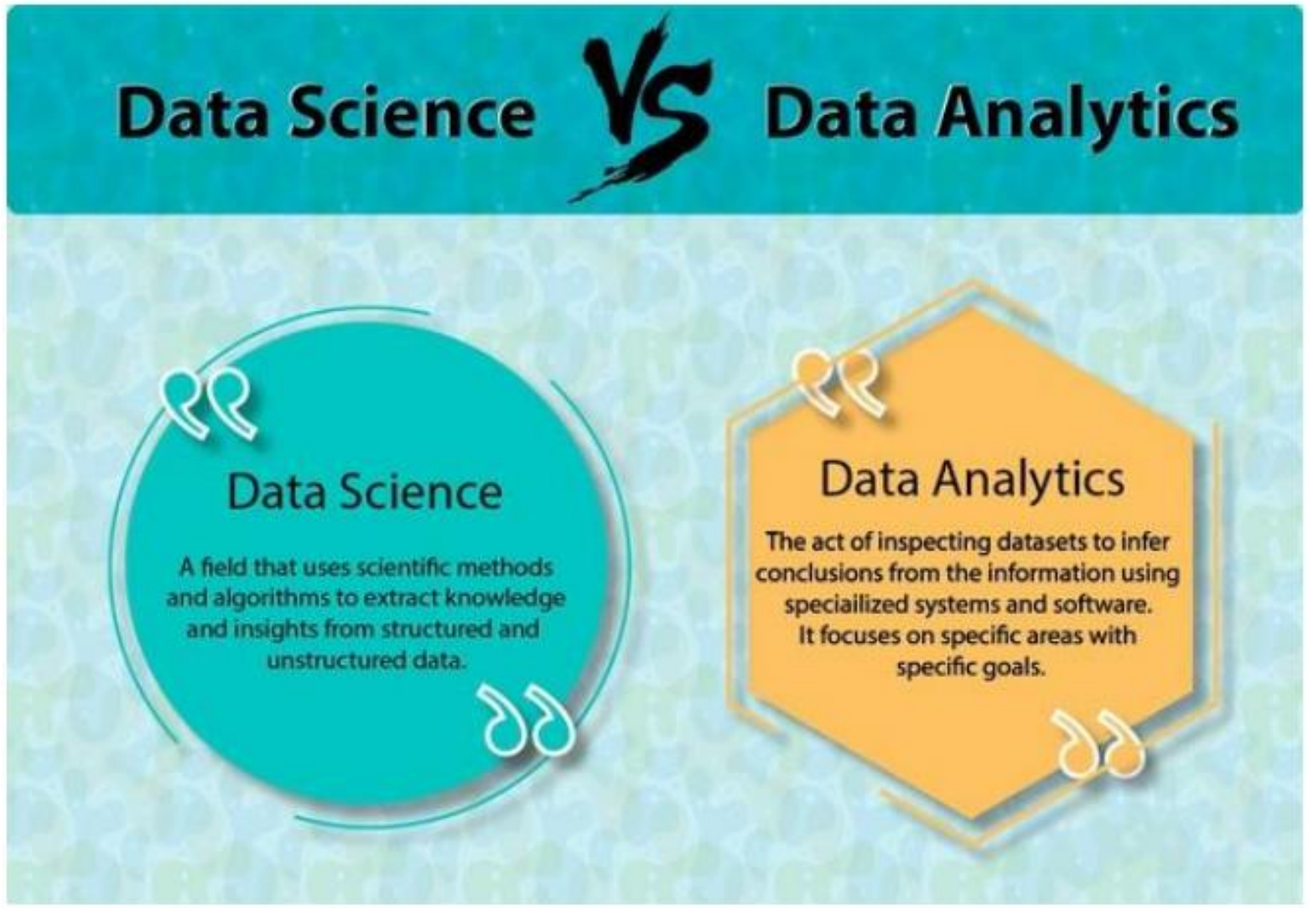
Data Analytics - Data Analysis

- While the terms data analytics and data analysis are frequently used interchangeably, data analysis is a subset of data analytics concerned with examining, cleansing, transforming, and modeling data to derive conclusions.
- Data analytics includes the tools and techniques used to perform data analysis.



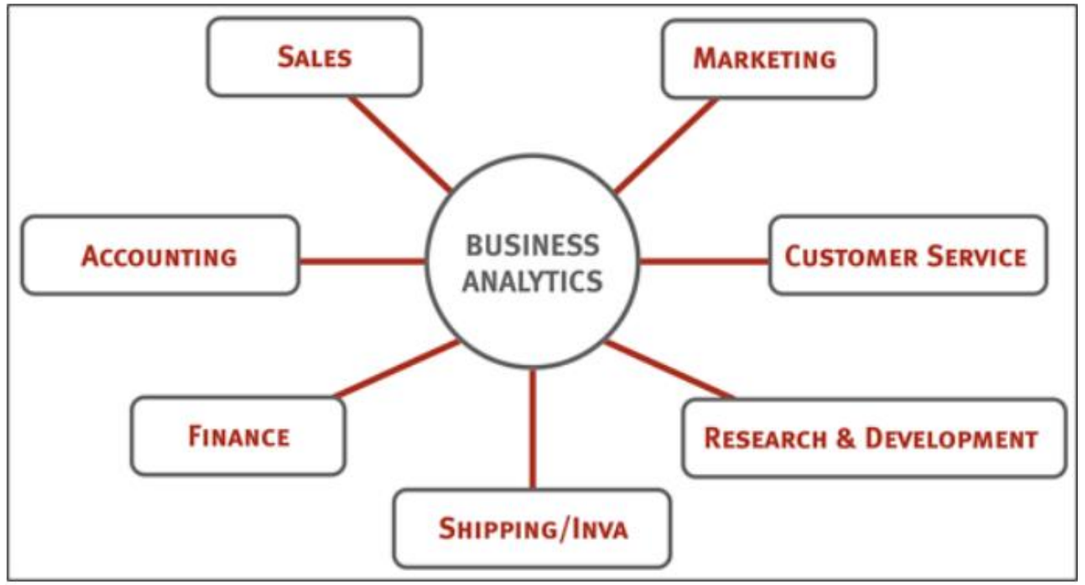
Data Analytics - Data Science

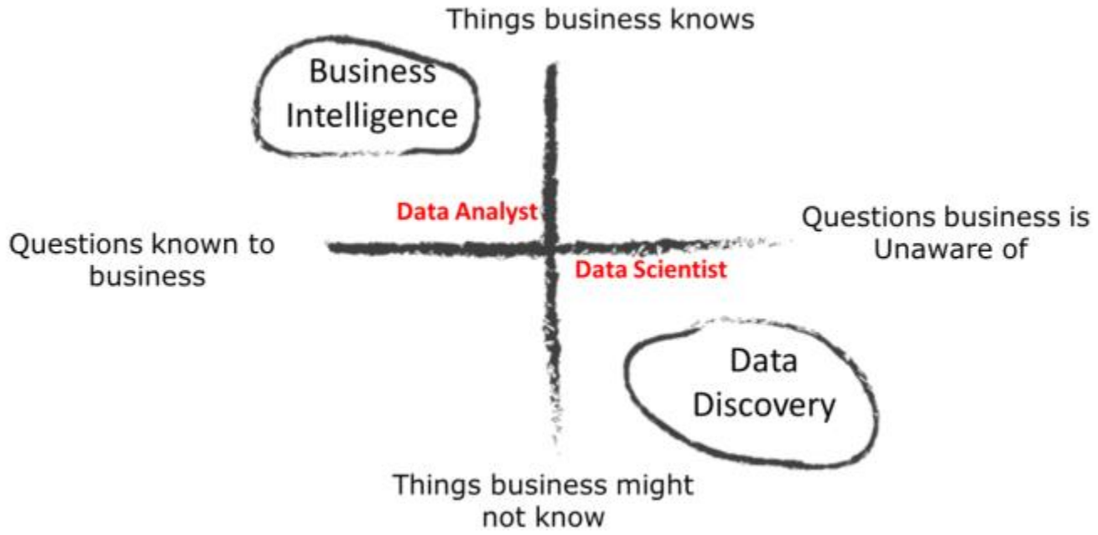
- Data analytics and data science are closely related.
- Data analytics is a component of data science, used to understand what an organization’s data looks like.
- Generally, the output of data analytics are reports and visualizations.
- Data science takes the output of analytics to study and solve problems.
- The difference between data analytics and data science is often seen as one of timescale.
- Data analytics describes the current or historical state of reality, whereas data science uses that data to predict and/or understand the future.



Data Analytics – Business Analytics

- Business analytics is another subset of data analytics.
- Business analytics uses data analytics techniques, including data mining, statistical analysis, and predictive modeling, to drive better business decisions.
- Gartner defines business analytics as “solutions used to build analysis models and simulations to create scenarios, understand realities, and predict future states.”





Journey from data to wisdom -1

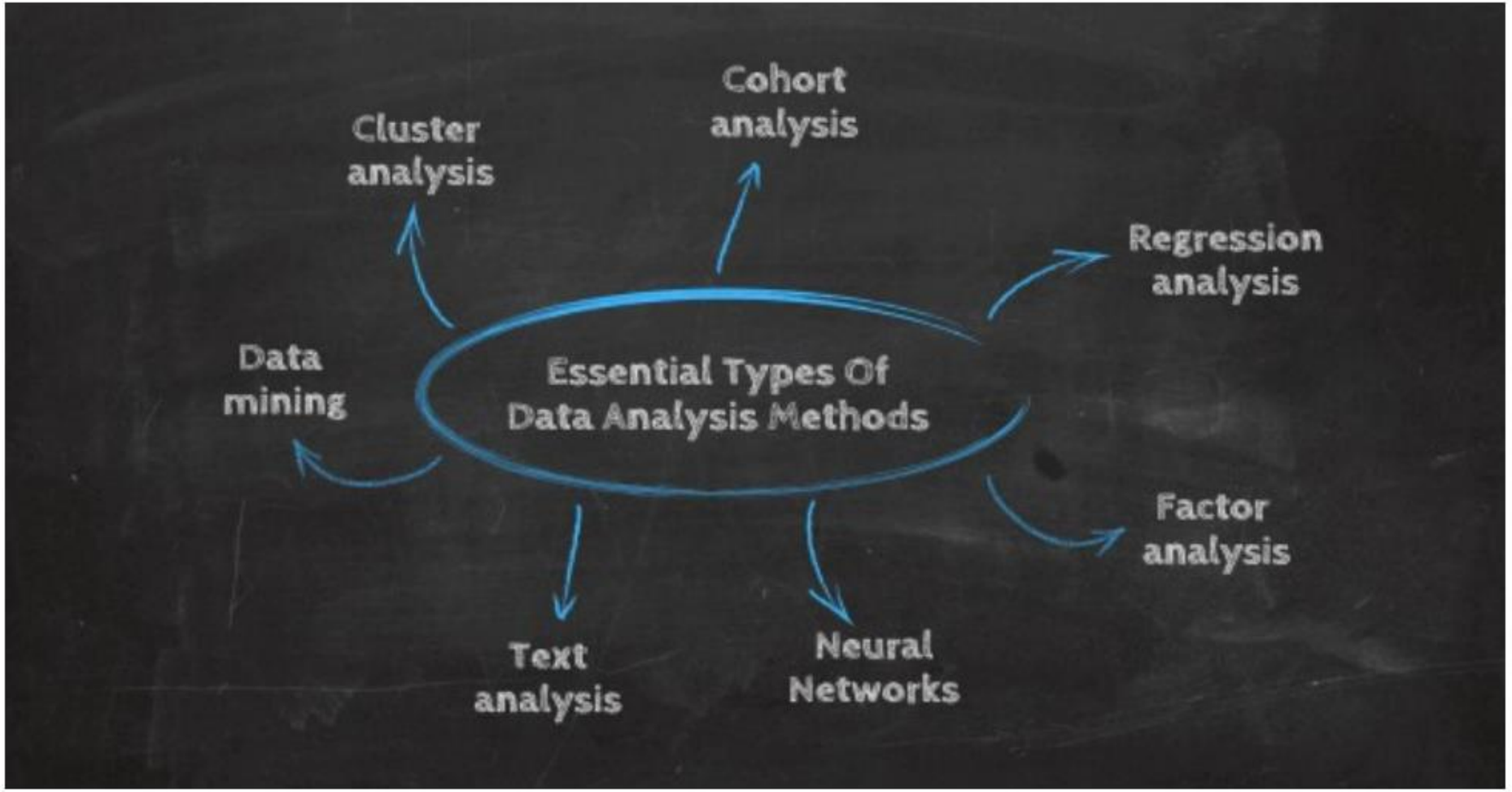
- The organism that goes hunting for information like searching for food can perceive the traces of its prey remotely by classifying the changes.
- At this stage, the number of uncertainties is quite high. While classifying the changes, the intelligence that is aimed at research should be developed to realize that it needs more information to increase accuracy and to notice the missing information.
- In order to find what it is looking for in a larger pile, it should learn to be a team to see if other hunters have seen what it is looking for beforehand.
- Being able to be a team allows the development of the intelligence focused on problem solving. Organisms that wander through the piles of information together develop the intelligence aimed at process management by learning to divide the work while sharing the information they find.
- When they start planning to divide the work in hunting a larger target, they develop the participatory intelligence aimed at problem solving.
- In task sharing, organisms start to act like organs by becoming experts in being successful in the task they undertake and doing the best. Thus, organs are created by sharing the functions of the work aimed at problem solving.

Journey from data to wisdom -2

- The integrity, that is, the body, is formed by the organs acting together. In order for the organs that form the body to feel each other, share tasks, monitor and manage, they need to form a leader brain.
- It should not be forgotten that on the path to success, being a team and teams feeling and perceiving each other very well is possible with a goal-oriented participatory mind.
- The basic rule of being successful in seizing opportunities and being different or finding difference is that when they learn to achieve excellence as a team, the mind that realizes the power of quality is developed. Organisms that feel that they need to be different not only in seizing threats but also in seizing opportunities should be able to focus on a single point at the same time in order to be a team.
- Organisms that quickly collect their own information, integrate it and form a body, start to predict the reactions they will give by monitoring the behaviors of the sensors that are the source of the information. In order to quickly perceive the difference and intensity of the reaction, it is necessary to establish a control mechanism similar to a neural network.

Data Analytics Methods And Techniques

Data analysis methods

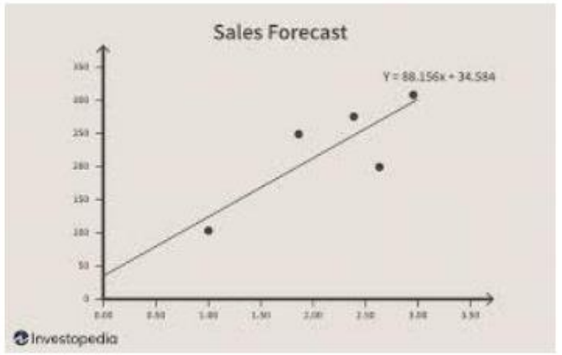


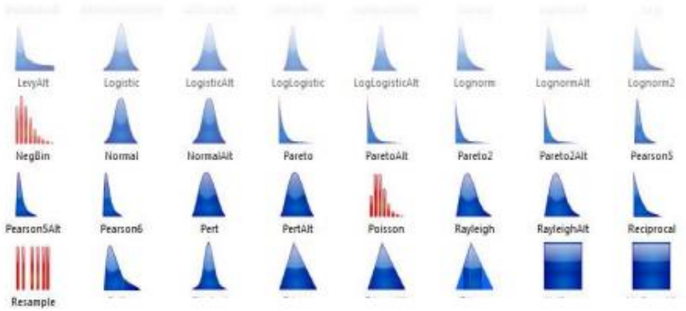
Data Analytics Methods and Techniques -1

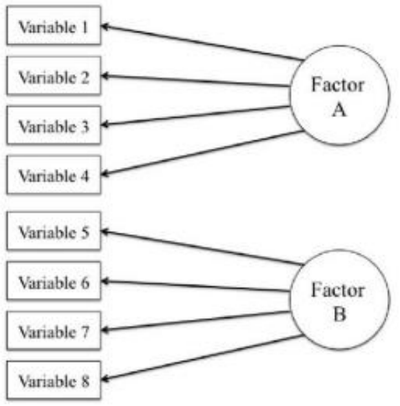
1) **Regression analysis:** Regression analysis is a set of statistical processes used to estimate the relationships between variables to determine how changes to one or more variables might affect another. For example, how might social media spending affect sales?

2) **Monte Carlo simulation:** “Monte Carlo simulations are used to model the probability of different outcomes in a process that cannot easily be predicted due to the intervention of random variables.” It is frequently used for risk analysis.

3) **Factor analysis:** Factor analysis is a statistical method for taking a massive data set and reducing it to a smaller, more manageable one. This has the added benefit of often uncovering hidden patterns. In a business setting, factor analysis is often used to explore things like customer loyalty





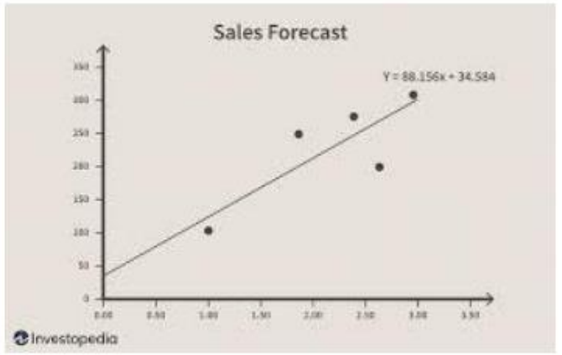


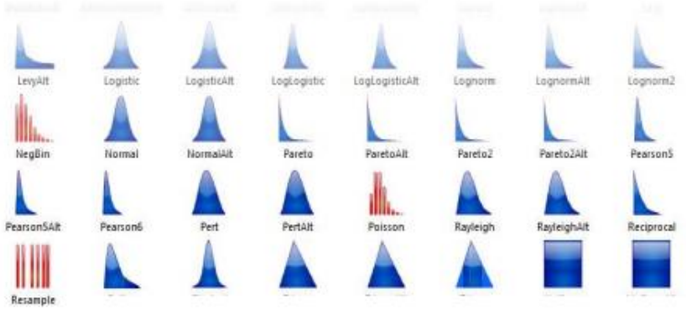
Data Analytics Methods and Techniques -2

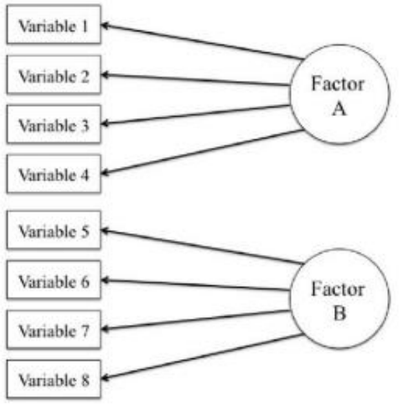
4) **Cohort analysis:** Cohort analysis is used to break a dataset down into groups that share common characteristics, or cohorts, for analysis. This is often used to understand customer segments.

5) **Cluster analysis:** Cluster analysis as “a class of techniques that are used to classify objects or cases into relative groups called clusters.” It can be used to reveal structures in data — insurance firms might use cluster analysis to investigate why certain locations are associated with particular insurance claims, for instance.

6) **Time series analysis**: Time series analysis as “a statistical technique that deals with time series data, or trend analysis. Time series data means that data is in a series of particular time periods or intervals. Time series analysis can be used to identify trends and cycles over time, e.g., weekly sales numbers. It is frequently used for economic and sales forecasting.

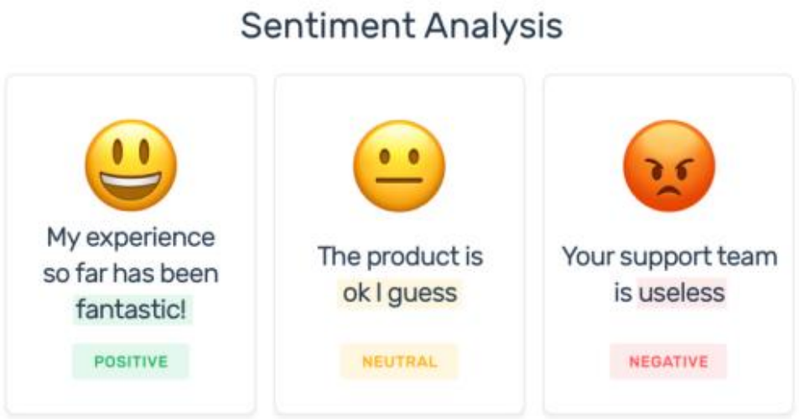






Data Analytics Methods and Techniques -3

7) **Sentiment analysis:** Sentiment analysis uses tools such as natural language processing text analysis, computational linguistics, and so on, to understand the feelings expressed in the data. While the previous six methods seek to analyze quantitative data (data that can be measured), sentiment analysis seeks to interpret and classify qualitative data by organizing it into themes. It is often used to understand how customers feel about a brand, product, or service.



Regression Analysis

- The regression analysis uses historical data to understand how a dependent variable's value is affected when one (linear regression) or more independent variables (multiple regression) change or stay the same.
- By understanding each variable's relationship and how they developed in the past, you can anticipate possible outcomes and make better business decisions in the future.

Factor Analysis

- The factor analysis, also called “dimension reduction,” is a type of data analysis used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors.
- The aim here is to uncover independent latent variables, an ideal analysis method for streamlining specific data segments.

Cohort Analysis

- This type of data analysis method uses historical data to examine and compare a determined segment of users' behavior, which can then be grouped with others with similar characteristics.
- By using this data analysis methodology, it's possible to gain a wealth of insight into consumer needs or a firm understanding of a broader target group.
- Cohort analysis can be really useful to perform analysis in marketing as it will allow you to understand the impact of your campaigns on specific groups of customers.

Cluster Analysis

- The action of grouping a set of data elements in a way that said elements are more similar (in a particular sense) to each other than to those in other groups – hence the term ‘cluster.’
- Since there is no target variable when clustering, the method is often used to find hidden patterns in the data. The approach is also used to provide additional context to a trend or dataset.

Neural Networks

- The neural network forms the basis for the intelligent algorithms of machine learning.
- It is a form of data-driven analytics that attempts, with minimal intervention, to understand how the human brain would process insights and predict values.
- Neural networks learn from each and every data transaction, meaning that they evolve and advance over time.

Data Mining

- A method of analysis that is the umbrella term for engineering metrics and insights for additional value, direction, and context.
- By using exploratory statistical evaluation, data mining aims to identify dependencies, relations, data patterns, and trends to generate and advanced knowledge.
- When considering how to analyze data, adopting a data mining mindset is essential to success - as such, it’s an area that is worth exploring in greater detail.

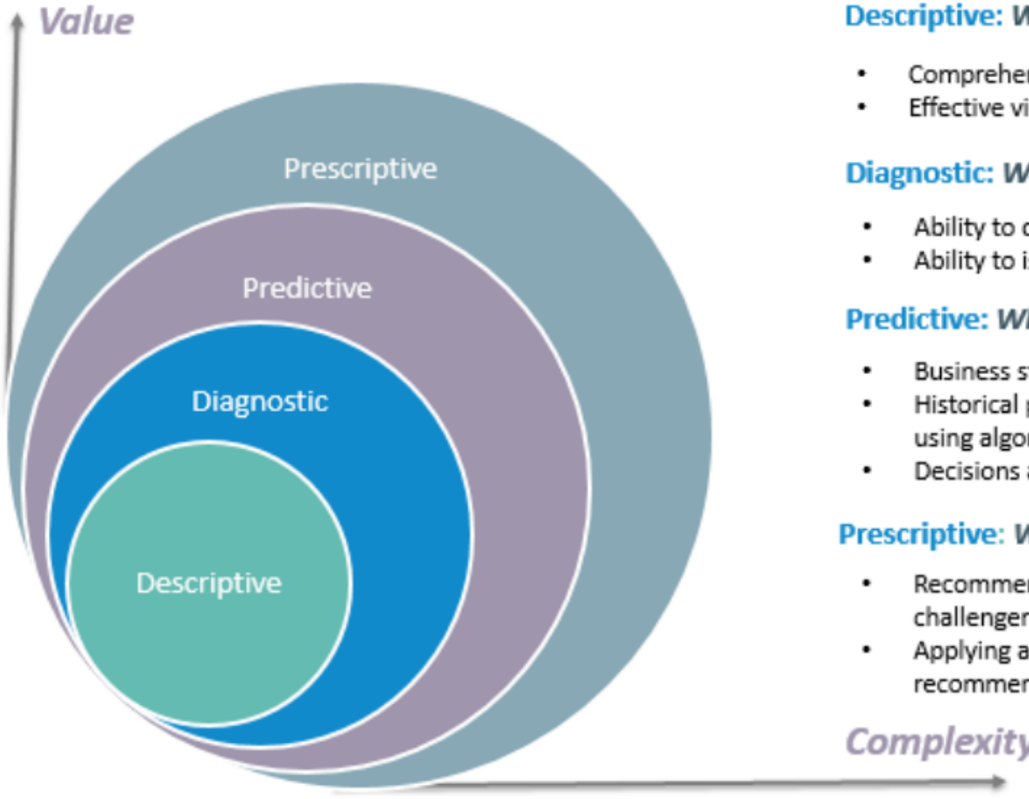
Text Analysis

- Text analysis, also known in the industry as text mining, is the process of taking large sets of textual data and arranging it in a way that makes it easier to manage.
- By working through this cleansing process in stringent detail, you will be able to extract the data that is truly relevant to your business and use it to develop actionable insights that will propel you forward.

Types of Data Analytics

TYPES OF DATA ANALYTICS

4 types of Data Analytics



What is the data telling you?

Descriptive: ***What's happening in my business?***

- Comprehensive, accurate and live data
- Effective visualisation

Diagnostic: ***Why is it happening?***

- Ability to drill down to the root-cause
- Ability to isolate all confounding information

Predictive: ***What's likely to happen?***

- Business strategies have remained fairly consistent over time
- Historical patterns being used to predict specific outcomes using algorithms
- Decisions are automated using algorithms and technology

Prescriptive: ***What do I need to do?***

- Recommended actions and strategies based on champion / challenger testing strategy outcomes
- Applying advanced analytical techniques to make specific recommendations

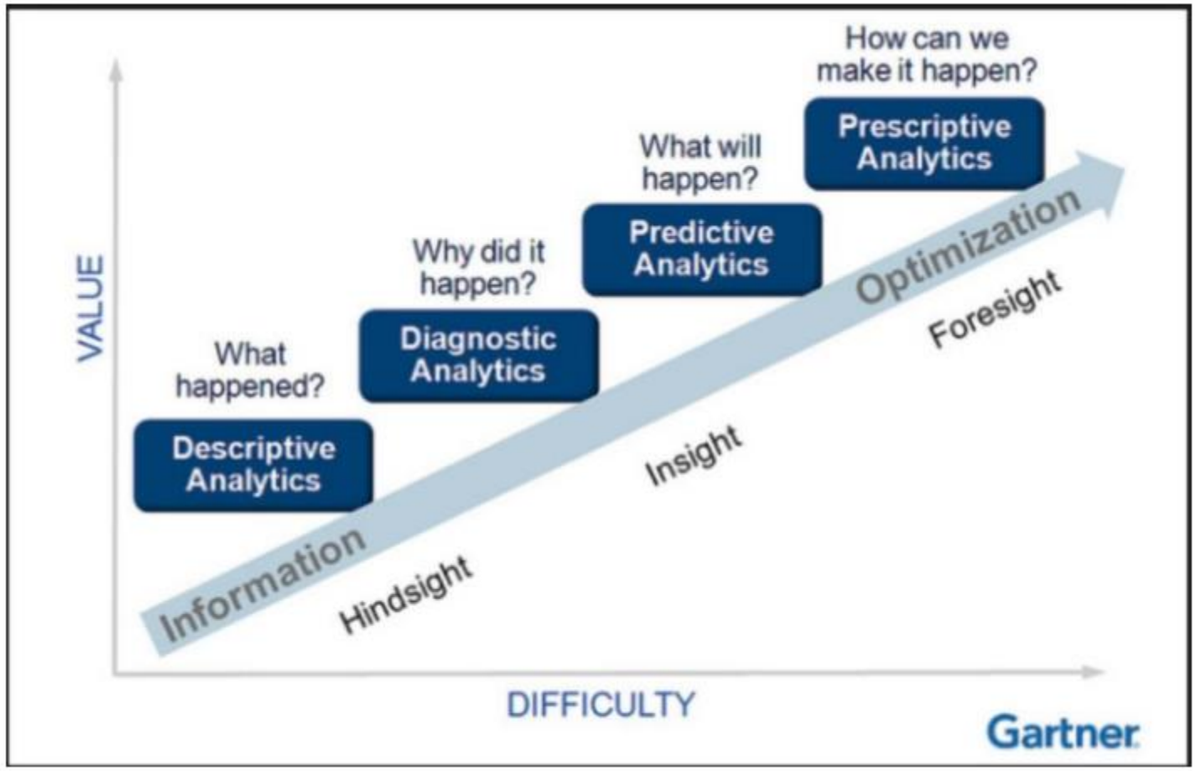
Type of data analytics
Descriptive – What happened?
Diagnostic – Why did it happen?
Predictive – What will happen?
Prescriptive – How can we achieve it?

Descriptive Analytics

- Descriptive analytics helps answer questions about what happened. These techniques summarize large datasets to describe outcomes to stakeholders.
- By developing key performance indicators (KPIs,) these strategies can help track successes or failures. Metrics such as return on investment (ROI) are used in many industries.
- Specialized metrics are developed to track performance in specific industries. This process requires the collection of relevant data, processing of the data, data analysis and data visualization. This process provides essential insight into past performance.

Descriptive Analytics

- **Descriptive Analytics:** Gaining insight into the past. Descriptive analytics, or statistics, summarizes raw data and makes it human-interpretable. They are analytics that describe the past. Whether the past was a minute ago or a year ago, descriptive analytics is useful because it allows us to learn from past behaviors and understand how they might affect future outcomes.
- **Descriptive analytics:** What has happened and what is happening right now? Descriptive analytics uses historical and current data from multiple sources to describe the present state by identifying trends and patterns. In business analytics, this is the purview of business intelligence (BI).



Predictive (Receteli) Analytics

- Predictive analytics helps answer questions about what will happen in the future. These techniques use historical data to identify trends and determine if they are likely to recur.
- Predictive analytical tools provide valuable insight into what may happen in the future and its techniques include a variety of statistical and machine learning techniques, such as: neural networks, decision trees, and regression.
- Predictive Analytics: Understanding the future. Predictive analytics is rooted in the ability to “guess” what might happen. These analytics are about understanding the future. Predictive analytics provide companies with actionable insights based on data. Predictive analytics provide estimates of the probability of a future outcome. It’s important to remember that no statistical algorithm can “predict” the future with 100% certainty. Companies use these statistics to predict what might happen in the future. This is because predictive analytics is based on probabilities.
- Predictive analytics: What is likely to happen in the future? Predictive analytics applies techniques such as statistical modeling, forecasting, and machine learning to the output of descriptive and diagnostic analytics to make predictions about future outcomes. Predictive analytics is often considered a type of “advanced analytics,” and frequently depends on machine learning and/or deep learning.

Prescriptive (Öngörücü) Analytics

- Prescriptive analytics helps answer questions about what should be done. By using insights from predictive analytics, data-driven decisions can be made.
- This allows businesses to make informed decisions in the face of uncertainty. Prescriptive analytics techniques rely on machine learning strategies that can find patterns in large datasets.
- By analyzing past decisions and events, the likelihood of different outcomes can be estimated.
- Prescriptive analytics suggests a course of action. If the average of five weather patterns measures the probability of a hot summer, we should add an evening shift to the brewery and rent an additional tank to increase production.
- Prescriptive analytics: What do we need to do? Prescriptive analytics is a type of advanced analytics that involves the application of testing and other techniques to recommend specific solutions that will deliver desired outcomes. In business, predictive analytics uses machine learning, business rules, and algorithms.

Diagnostic Analytics

- Diagnostic analytics focuses more on why something is happening. This involves a wider range of data inputs and some hypothesizing. Did the weather affect beer sales? Did this latest marketing campaign affect sales? (Problem, Diagnosis, Diagnosis, Treatment).
- Diagnostic analytics: Why is it happening? Diagnostic analytics uses data (often generated via descriptive analytics) to discover the factors or reasons for past performance.
- Diagnostic analytics helps answer questions about why things happened. These techniques supplement more basic descriptive analytics.
- They take the findings from descriptive analytics and dig deeper to find the cause. The performance indicators are further investigated to discover why they got better or worse. This generally occurs in three steps:
  - Identify anomalies in the data. These may be unexpected changes in a metric or a particular market.
  - Data that is related to these anomalies is collected.
  - Statistical techniques are used to find relationships and trends that explain these anomalies

BIG DATA Analytics

Big Data Analytcs

- Big data analytics describes the process of uncovering trends, patterns, and correlations in large amounts of raw data to help make data-informed decisions.
- Big data analytics is the use of advanced analytic techniques against very large, diverse data sets that include structured, semi-structured and unstructured data, from different sources, and in different sizes from terabytes to zettabytes.
- Analysis of big data allows analysts, researchers and business users to make better and faster decisions using data that was previously inaccessible or unusable.
- Businesses can use advanced analytics techniques such as text analytics, machine learning, predictive analytics, data mining, statistics and natural language processing to gain new insights from previously untapped data sources independently or together with existing

Big Data is Problematic

- Big data is not analyzed, it is not used in smart data analytics applications. Models are not produced.
- Because it is noisy, contains incomplete, incorrect, manipulated data. Since it is very large, it contains excessive sensitivity.
- A small dataset representing big data is taken. A model representing the behavior of the big dataset is created by training. The accuracy of the model is tested. Its performance is increased, experience and talent are gained. Wisdom is created by raising awareness.

How Big Data Analytics Works -1

- Big data analytics refers to collecting, processing, cleaning, and analyzing large datasets to help organizations operationalize their big data.
- 1. 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
- 2. 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 -2

- 3. 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.
- 4. 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.

Data Lake

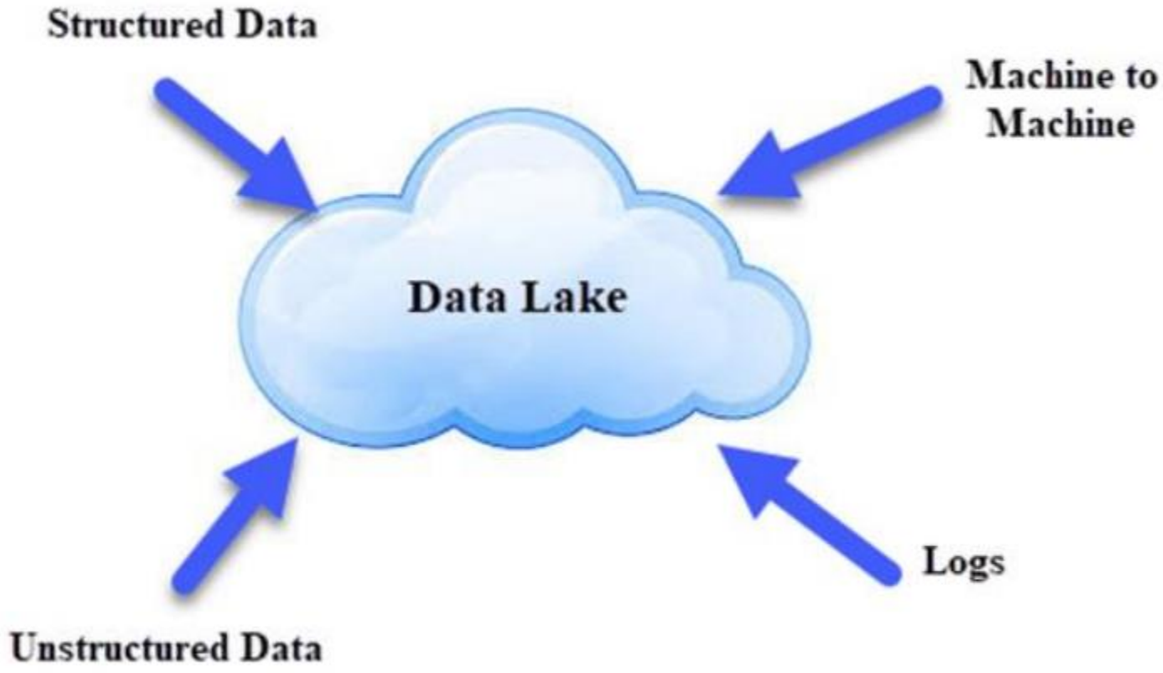
- Data Mining is the storage of large amounts of information designed for querying and analysis and is the process of transforming data into information.
- A Data Lake is a data repository that can store large amounts of structured, semi-structured and unstructured data. It is a place where you can store any type of data in its native format without a fixed limit and offers large amounts of data for increased analytical performance and local integration.
- A Data Lake is like a large water catchment area, much like a real lake and river. Just like in a lake, you have multiple streams coming in; similarly, a data lake has structured, unstructured, machine-to-machine, real-time data flowing in.
- A Data Lake stores all data regardless of its source and structure, while a Data Warehouse stores data in quantitative metrics along with its attributes (key features).
- A Data Lake is a storage repository that stores large amounts of structured, semi-structured and unstructured data, while a Data Warehouse is a blend of technologies and components that allow for strategic use of data.
- Data Lake defines the attribute after the data is stored, while Data Warehouse defines the attribute before the data is stored.
- Data Lake uses ELT(Extract Load Transform) process, while Data Warehouse uses ETL(Extract Transform Load) process.
- Data Lake is ideal for those who want in-depth analysis, while Data Warehouse is ideal for operations users.

What is Data Lake? -1

- A Data Lake is a storage repository that can store large amount of structured, semi-structured, and unstructured data.
- It is a place to store every type of data in its native format with no fixed limits on account size or file.
- It offers high data quantity to increase analytic performance and native integration.
- Data Lake is like a large container which is very similar to real lake and rivers.
- Just like in a lake you have multiple tributaries coming in, a data lake has structured data, unstructured data, machine to machine, logs flowing through in real-time

What is Data Lake? -2

- The Data Lake democratizes data and is a cost-effective way to store all data of an organization for later processing.
- Research Analyst can focus on finding meaning patterns in data and not data itself.
- Unlike a hierarchal Dataware house where data is stored in Files and Folder, Data lake has a flat architecture.
- Every data elements in a Data Lake is given a unique identifier and tagged with a set of metadata information.



Data collection - Data storage

- Depending on the level of complexity, data can be moved to storage areas such as cloud data warehouses or data lakes. When needed, business intelligence tools can access this data.

- Data Mining is the storage of large amounts of information designed for querying and analysis and is the process of transforming data into information. A Data Lake is a data pool that can store large amounts of structured, semi-structured and unstructured data. It is a place where you can store any type of data in its native format without a fixed limit and offers large amounts of data for increased analytical performance and local integration.

- A Data Lake is like a large water collection area, much like real lakes and rivers. Just like in a lake, you have multiple streams coming in; similarly, a data lake has structured, unstructured, machine-to-machine, real-time data flowing.
- A Data Lake stores all data regardless of its source and structure, while a Data Warehouse stores data in quantitative metrics along with its attributes (basic features).
- Data Lake is a storage pool that stores large structured, semi-structured and unstructured data, while Data Warehouse is a blend of technologies and components that allow strategic use of data.
- Data Lake defines the attribute after the data is stored, while Data Warehouse defines the attribute before the data is stored.
- Data Lake uses the ELT (Extract Load Transform) process, while Data Warehouse uses the ETL (Extract Transform Load) process.
- Data Lake is ideal for those who want in-depth analysis, while Data Warehouse is ideal for operations users.
- Data Lake uses the ELT (Extract Load Transform) process, while Data Warehouse uses the ETL (Extract Transform Load) process.
- This involves identifying data sources and collecting data from them. Data collection follows the ETL or ELT processes.

ETL - ELT

- ETL - Extract Transform Load:
- In ETL, the generated data is first transformed into a standard format and then loaded into storage.
- ELT - Extract Load Transform:
- In ELT, the data is first loaded into storage and then transformed into the required format.
- Comparison of data lakes and data warehouses
- A data warehouse is a database optimized for analyzing relational data from transaction-based systems and business applications. The data structure and schema are predefined for optimization for fast search and reporting. The data is cleansed, enriched, and transformed to function as a “single source of truth” that users can trust. Examples of data include customer profiles and product information.
- A data lake is different as it can store both structured and unstructured data without any detailed processing. The structure of the data or schema is not defined when the data is captured. This means that you can store all your data without careful design, and this functionality is especially useful when it is not known how the data will be used in the future. Data examples include social media content, IoT device data, and non-relational data from mobile applications.

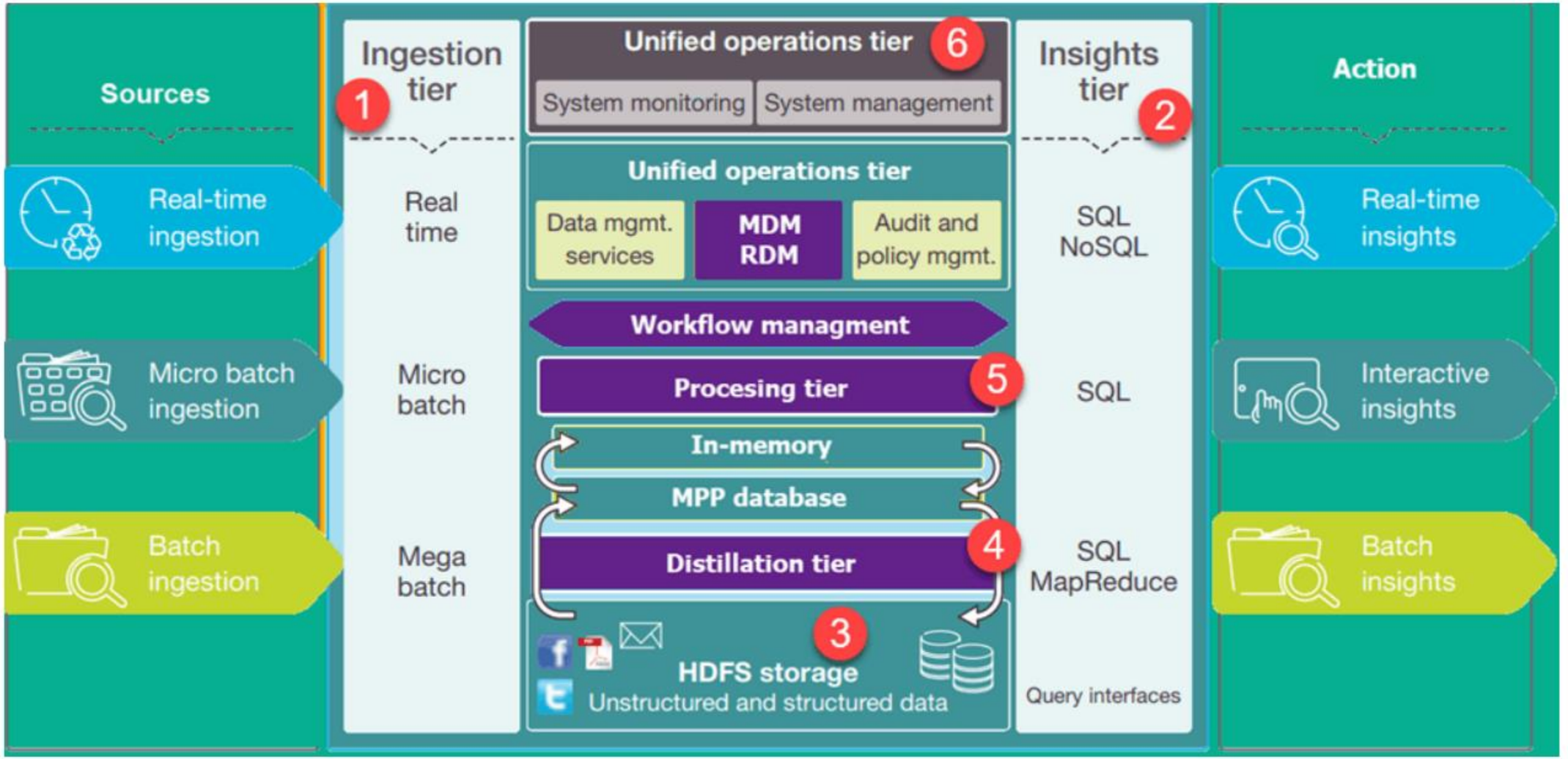
Reasons for using Data Lake

The main objective of building a data lake is to offer an unrefined view of data to data scientists.

Reasons for using Data Lake are:

- With the onset of storage engines like Hadoop storing disparate information has become easy. There is no need to model data into an enterprise-wide schema with a Data Lake.
- With the increase in data volume, data quality, and metadata, the quality of analyses also increases.
- Data Lake offers business Agility
- Machine Learning and Artificial Intelligence can be used to make profitable predictions.
- It offers a competitive advantage to the implementing organization.
- There is no data silo structure. Data Lake gives 360 degrees view of customers and makes analysis more robust.

Data Lake Architecture -1



Data Lake Architecture -2

The figure shows the architecture of a Business Data Lake. The lower levels represent data that is mostly at rest while the upper levels show real-time transactional data. This data flow through the system with no or little latency. Following are important tiers in Data Lake Architecture:

1. Ingestion Tier: The tiers on the left side depict the data sources. The data could be loaded into the data lake in batches or in real-time
2. Insights Tier: The tiers on the right represent the research side where insights from the system are used. SQL, NoSQL queries, or even excel could be used for data analysis.
3. HDFS is a cost-effective solution for both structured and unstructured data. It is a landing zone for all data that is at rest in the system.
4. Distillation tier takes data from the storage tire and converts it to structured data for easier analysis.
5. Processing tier run analytical algorithms and users queries with varying real time, interactive, batch to generate structured data for easier analysis.
6. Unified operations tier governs system management and monitoring. It includes auditing and proficiency management, data management, workflow management.

Key Data Lake Concepts -1

The Basic Data Lake concepts that need to be understood to fully understand the Data Lake Architecture are listed below.

Key Data Lake Concepts -2

- Data Ingestion allows connectors to get data from a different data sources and load into the Data lake. Data Ingestion supports:
  - All types of Structured, Semi-Structured, and Unstructured data.
  - Multiple ingestions like Batch, Real-Time, One-time load.
  - Many types of data sources like Databases, Webservers, Emails, IoT, and FTP.
- Data storage should be scalable, offers cost-effective storage and allow fast access to data exploration. It should support various data formats.
- Data governance is a process of managing availability, usability, security, and integrity of data used in an organization.
- Security needs to be implemented in every layer of the Data lake. It starts with Storage, Unearthing, and Consumption. The basic need is to stop access for unauthorized users. It should support different tools to access data with easy to navigate GUI and Dashboards. Authentication, Accounting, Authorization and Data Protection are some important features of data lake security.
- Data Quality: Data quality is an essential component of Data Lake architecture. Data is used to exact business value. Extracting insights from poor quality data will lead to poor quality insights.

Key Data Lake Concepts -3

- Data Discovery is another important stage before you can begin preparing data or analysis. In this stage, tagging technique is used to express the data understanding, by organizing and interpreting the data ingested in the Data lake.
- Data Auditing:Two major Data auditing tasks are tracking changes to the key dataset.
  - Tracking changes to important dataset elements
  - Captures how/ when/ and who changes to these elements.
  - Data auditing helps to evaluate risk and compliance.
- Data Lineage: This component deals with data's origins. It mainly deals with where it movers over time and what happens to it. It eases errors corrections in a data analytics process from origin to destination.
- Data Exploration: It is the beginning stage of data analysis. It helps to identify right dataset is vital before starting Data Exploration. All given components need to work together to play an important part in Data lake building easily evolve and explore the environment.

Maturity Stages of Data Lake

- Stage 1: Handle and ingest data at scale: This first stage of Data Maturity Involves improving the ability to transform and analyze data. Here, business owners need to find the tools according to their skillset for obtaining more data and build analytical applications.
- Stage 2: Building the analytical muscle: This is a second stage which involves improving the ability to transform and analyze data. In this stage, companies use the tool which is most appropriate to their skillset. They start acquiring more data and building applications. Here, capabilities of the enterprise data warehouse and data lake are used together.
- Stage 3: EDW and Data Lake work in unison:This step involves getting data and analytics into the hands of as many people as possible. In this stage, the data lake and the enterprise data warehouse start to work in a union. Both playing their part in analytics
- Stage 4: Enterprise capability in the lake: In this maturity stage of the data lake, enterprise capabilities are added to the Data Lake. Adoption of information governance, information lifecycle management capabilities, and Metadata management. However, very few organizations can reach this level of maturity, but this tally will increase in the future.

Best Practices for Data Lake Implementation

- Architectural components, their interaction and identified products should support native data types
- Design of Data Lake should be driven by what is available instead of what is required. The schema and data requirement is not defined until it is queried
- Design should be guided by disposable components integrated with service API.
- Data discovery, ingestion, storage, administration, quality, transformation, and visualization should be managed independently.
- The Data Lake architecture should be tailored to a specific industry. It should ensure that capabilities necessary for that domain are an inherent part of the design
- Faster on-boarding of newly discovered data sources is important
- Data Lake helps customized management to extract maximum value
- The Data Lake should support existing enterprise data management techniques and methods

Usage Notes

- A lot of slides are adopted from the presentations and documents published on internet by experts who know the subject very well.
- I would like to thank who prepared slides and documents.
- Also, these slides are made publicly available on the web for anyone to use
- If you choose to use them, I ask that you alert me of any mistakes which were made and allow me the option of incorporating such changes (with an acknowledgment) in my set of slides.

Sincerely,

Dr. Cahit Karakuş

cahitkarakus@gmail.com