Topic for the class:Data to Data Science - understanding data
Unit _1 : Title-Data Evolution
Date & Time : 4.7.24 10.00 AM – 10.50 AM

Dr. Bhramaramba Ravi

Professor

Department of Computer Science and Engineering

GITAM School of Technology (GST)

Visakhapatnam – 530045

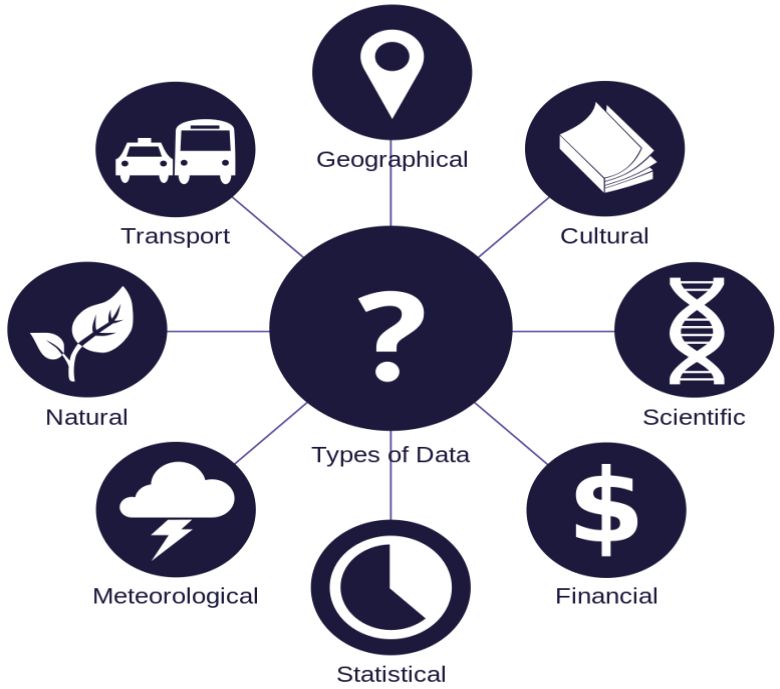
Email: **bravi@gitam.edu**

Unit1 syllabus

- **UNIT 1 Data Evolution 9 hours, P – 2 hours Data Evolution**: Data to Data Science – Understanding data: Introduction – Type of Data, Data Evolution – Data Sources. Preparing and gathering data and knowledge - Philosophies of data science - data all around us: the virtual wilderness - Data wrangling
-
- : from capture to domestication - Data science in a big data world - Benefits and uses of data science and big data - facets of data.
- https://www.coursera.org/learn/intro-analyticthinking-datascience-datamining?specialization=data-science- fundamentals

Definition and Use of Data

- Data are discrete or continuous values conveying information.
- Data can be abstract ideas or concrete measurements.
- Data are used in scientific research, economics, and human organizational activities.



Concept of Data

- Data are the smallest units of factual information.
- Thematically connected data is viewed as information.
- Contextually connected pieces of information are described as data insights or intelligence.
- The stock of insights and intelligence resulting from the synthesis of data into information is described as knowledge.

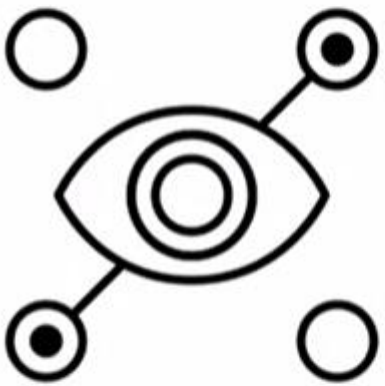
Definition of Data (COMPUTER SCIENCE)

- Data is any sequence of symbols, with datum being a single symbol of data.
- Data requires interpretation to become information.
- Digital data is represented using the binary number system of ones and zeros.

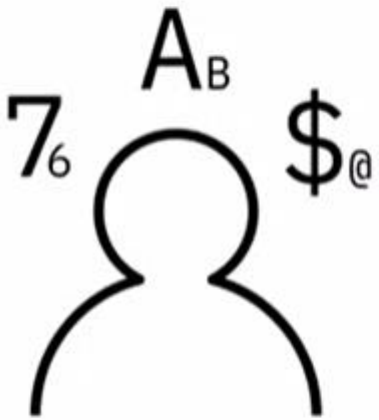
States of Data

- Data exists in three states: data at rest, data in transit, and data in use.
- Data within a computer moves as parallel data, while data moving to or from a computer moves as serial data.
- Data representing quantities, characters, or symbols are stored and recorded on various recording media and transmitted in the form of digital signals.

What is Data?



Facts
Observations
Perceptions



Numbers
Characters
Symbols

Images

Storage of Data (Data Structures)

- Physical computer memory elements consist of an address and a byte/word of data storage.

- Digital data are often stored in relational databases and can be represented as abstract key/value pairs.

- Data can be organized in various types of data structures, including arrays, graphs, and objects.

Storage of Data

- To store data bytes in a file, they must be serialized in a file format.

- Executable files contain programs, all other files are also data files.

- The line between program and data can become blurry.

META DATA and its CHARACTERISTICS

DATA about DATA is called META DATA

- Metadata helps translate data to information.
- Data relating to physical events or processes has a temporal component.
- Computers follow a sequence of instructions given in the form of data.
- A single datum is a value stored at a specific location, allowing computer programs to operate on other computer programs by manipulating their programmatic data.

Collection and Analysis of Data

• Data are collected using techniques like measurement, observation, query, or analysis.

• Field data are collected in an uncontrolled in-situ environment.

• Experimental data are generated in a controlled scientific experiment.

• Data are analyzed using techniques such as calculation, reasoning, discussion, presentation, visualization, or other forms of post-analysis.

The Evolution of Personal Computing

1940s to 1989 – Data Warehousing and Personal Desktop Computers

• The world's first programmable computer, ENIAC, was developed by the U.S. army during World War 2.

• IBM released the first transistorized computer, TRADIC, in the early 1960s, allowing data centers to branch out of the military.

• The first personal desktop computer with a Graphical User Interface (GUI) was Lisa, released by Apple Computers in 1983.

• Companies like Apple, Microsoft, and IBM released a wide range of personal desktop computers, leading to widespread personal computer use.

1989 to 1999 – Emergence of the World Wide Web

• Between 1989 and 1993, Sir Tim Berners-Lee created the fundamental technologies for the World Wide Web.

• The decision to make the underlying code for these web technologies free led to a massive explosion in data access and sharing.

2000s to 2010s – Controlling Data Volume, Social Media and Cloud Computing

• Companies like Amazon, eBay, and Google generated large amounts of web traffic and unstructured data.
• AWS launched in 2002, offering a range of cloud infrastructure services, attracting customers like Dropbox, Netflix, and Reddit.
• Social media platforms like MySpace, Facebook, and Twitter spread unstructured data, leading to the creation of Hadoop and NoSQL database queries.

Data To Big Data and Data Science Evolution

Early Beginnings (1960s-1970s):
• Rooted in the world of statistics, with early computers enabling data analysis and visualization.

Emergence of Data Mining (1980s-1990s):
• Researchers developed algorithms to uncover meaningful patterns and insights within vast datasets.
• Worked on classification, clustering, and association rule mining.

Growth of Data Warehousing (1990s):
• Data warehousing technologies allowed organizations to centralize and manage large data volumes.

Rise of Machine Learning (1990s-Present):

• Advances in algorithms and techniques paved the way for predictive modeling and data-driven decision-making.

Big Data Era (2000s-Present):
• The 2000s saw the explosive growth of data due to the internet, social media, and sensors.
• Technologies like Hadoop and MapReduce were developed to process and analyze these massive datasets.

Data Science as a Discipline (2000s-Present):
• The term "data science" gained popularity in the early 2000s.

Data Science in Industry (2010s-Present):

• Data science became indispensable across various industries.

Tools and Frameworks (2010s-Present):

• The open-source ecosystem for data science tools and frameworks expanded rapidly.

Ethical and Regulatory Concerns (2010s-Present):

• Concerns about data privacy, algorithmic bias, and ethical considerations gained prominence.

Artificial Intelligence and Data Science Integration (2020s-Present):

• Data science has become tightly intertwined with AI, with machine learning and deep learning playing central roles.

Big Data

- Big data refers to very large quantities of data

- This large amount of data is of structured, semi-structured, and unstructured data.

- It arrives at a higher volume, faster rate, in a wider variety of file formats, and from a wider variety of sources.

- The term was officially coined by NASA researchers in 1997 to describe processing and visualizing vast amounts of data from supercomputers.

- Doug Laney's 2001 paper established three primary components of big data: Volume (size of data), Velocity (speed of data growth), and Variety (number of data types and sources).

Big Data and Processing

- Traditional data analysis methods and computing are difficult to work with large datasets.

- The new field of data science uses machine learning and AI methods for efficient applications of analytic methods to big data.

“Big Data refers to the dynamic, large and disparate volumes of data being created by people, tools, and machines.

- It requires new, innovative, and scalable technology to collect, host, and analytically process the vast amount of data gathered in order to derive real-time business insights that relate to consumers, risk, profit, performance, productivity management, and enhanced shareholder value.”
- There is no one definition of Big Data, but there are certain elements that are common across the different definitions, such as velocity, volume, variety, veracity, and value. These are the V's of Big Data.

Velocity is the speed at which data accumulates.

- Data is being generated extremely fast, in a process that never stops.
- Near or real-time streaming, local, and cloud-based technologies can process information very

quickly.

Example for Velocity: Every 60 seconds, hours of footage are uploaded to YouTube which is generating data.

- Think about how quickly data accumulates over hours, days, and years.

Volume is the scale of the data, or the increase in the amount of data stored.

- Drivers of volume are the increase in data sources, higher resolution sensors, and scalable infrastructure.

Example for Volume: The world population is approximately seven billion people and the vast majority are now using digital devices; mobile phones, desktop and laptop computers, wearable devices, and so on.

These devices all generate, capture, and store data -- approximately 2.5 quintillion bytes every day.

Variety is the diversity of the data.

- Structured data fits neatly into rows and columns, in relational databases while unstructured data is not organized in a pre-defined way, like Tweets, blog posts, pictures, numbers, and video.
- Variety also reflects that data comes from different sources, machines, people, and processes, both internal and external to organizations.
- Drivers are mobile technologies, social media, wearable technologies, geo technologies, video, and many, many more.

Example for Variety: Let's think about the different types of data; text, pictures, film, sound, health data from wearable devices, and many different types of data from devices connected to the Internet of Things.

Veracity is the quality and origin of data, and its conformity to facts and accuracy.

- Attributes include consistency, completeness, integrity, and ambiguity.
- Drivers include cost and the need for traceability.
- With the large amount of data available, the debate rages on about the accuracy of data in the digital age. Is the information real, or is it false?
- Example for Veracity: 80% of data is considered to be unstructured and we must devise ways to produce reliable and accurate insights.
- The data must be categorized, analyzed, and visualized.

Value is our ability and need to turn data into value.

- Value isn't just profit.
- It may have medical or social benefits, as well as customer, employee, or personal satisfaction.
- The main reason that people invest time to understand Big Data is to derive value from it.