**Definition of Data**

the quantities, characters, or symbols on which operations are performed by a computer, which may be stored and transmitted.

Another computer centric definition: a collection of facts (numbers, words, measurements, observations, etc) that has been translated into a form that computers can process

**What is Data collection in context of AI?**

Data collection is the process of gathering and measuring information from countless different sources. In order to use the data we collect to develop practical artificial intelligence (AI) and machine learning solutions, it must be collected and stored in a way that makes sense for the business problem at hand

**What is a DataSet?**

A data set is a collection of data. In the case of tabular data, a data set corresponds to one or more database tables, where every column of a table represents a particular variable, and each row corresponds to a given record of the data set in question.

**What is supervised learning? What data is provided in supervised learning?**

Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labeled training data consisting of a set of training examples.

*Is the below example supervised learning is used to predict cost of house based on area & number of bedrooms. Here area, no. of bedrooms & cost all form dataset & are provided to the machine learning algorithm as training set. Here mapping is A → B is represented A (area & no. of bedrooms) → B (cost of house)*

|  |  |  |
| --- | --- | --- |
| Input | Output | Application |
| Email | Spam? (0/1) | Spam filtering |
| Audio | Text transcript | Speech recognition |
| English | Hindi | Machine translation |
| Add, user info | Click? (0/1) | Online advertising |
| Image, radar info | Position of other cars | Self-driving car |
| Image of a product | Defect? (0/1) | Quality inspection |

**How is Data collected or acquired?**

* Manual Labelling
* Observing Behavior- humans, machines, responses
* Download from website/get from partners

**Importance of Data**

Data is important for any business to understand, predict and make necessary changes in the working. Data is important for computer to process for a reasonable information.

****

****

**Problems with Data**

* Data is at times bit over hyped -there is hardly anything as perfect dataset
* Data can be misused- AI can’t always work just with huge data
* Data is messy
	+ Garbage In Garbage Out (GIGO)
	+ Data Problems
		- Incorrect Label
		- Missing values
* Threat of Data theft

**Multiple type of Data**

**One way of looking at data is structured & unstructured data**

* Structured Data
	+ Flat files
	+ Relational database
* Unstructured Data
	+ Image
	+ Audio
	+ text
* The techniques for dealing with unstructured data are little bit different than the techniques for dealing with structured data. But AI techniques can work very well for both of these types of data

**Another way of looking at data can be:**

* Personal Data- is anything that is specific to you. It covers your demographics, your location, your email address and other identifying factors.
* Transactional data- is anything that requires an action to collect. You might click on an ad, make a purchase, visit a certain web page, etc. Pretty much every website you visit collects transactional data of some kind, either through [Google Analytics](http://www.google.com/analytics/), another 3rd party system or their own internal data capture system.
* Web data- is a collective term which refers to any type of data you might pull from the internet, whether to study for research purposes or otherwise. That might be data on what your competitors are selling, published government data, football scores, etc.
* Sensor Data- Sensor data is produced by objects and is often referred to as the [Internet of Things](http://www.forbes.com/sites/jacobmorgan/2014/05/13/simple-explanation-internet-things-that-anyone-can-understand/). It covers everything from your smartwatch measuring your heart rate to a building with external sensors that measure the weather.

**Data Storage**

Data storage refers to the use of recording media to retain data using computers or other devices.

**Forms of Data Storage**

The most prevalent forms of data storage are file storage, block storage and object storage, with each being ideal for different purposes.

* File Storage- when the data is stored in form of files.
* Block Storage- In computing, a block, sometimes called a physical record, is a sequence of bytes or bits, usually containing some whole number of records, having a maximum length; a block size. Data stored by blocks is known as Block Storage.
* Object storage is a computer data storage architecture that manages data as objects, as opposed to other storage architectures like file systems which manages data as a file hierarchy, and block storage which manages data as blocks within sectors and tracks.

**Types of Data Storage**

Data storage essentially means that files and documents are recorded digitally and saved in a storage system for future use. Storage systems may rely on electromagnetic, optical or other media to preserve and restore the data if needed. Data storage makes it easy to back up files for safekeeping and quick recovery in the event of an unexpected computing crash or cyberattack. Data storage can occur on physical hard drives, disk drives, USB drives or virtually in the cloud.

**Direct Attached Storage (DAS)**

* Direct attached storage (DAS) includes types of data storage that are physically connected to your computer. This storage is generally accessible to only a single machine. Some common devices in this category include:
* Hard Drives
* Solid-State Drives (SSD)
* CD/DVD Drives
* Flash Drives
* DAS solutions are great for creating local backups and can be more affordable than NAS solutions, but sharing data between machines can be cumbersome.

**Network Attached Storage (NAS)**

* Network attached storage (NAS) allows for multiple machines to share storage over a network. This is accomplished with multiple hard drives or other storage devices in a RAID configuration. One of the key benefits of NAS is the ability to centralize data and improve collaboration. Data can be easily shared among connected machines, and permission levels can be set to control access. While NAS solutions tend to be more costly than DAS solutions, they are still very affordable as storage technology has advanced significantly.
* Redundant Array of Independent/Inexpensive Disks (RAID) is a technology that allows storing data across multiple hard drives. The purpose of RAID is to achieve data redundancy to reduce data loss and, in a lot of cases, improve performance.

**Cloud & Hybrid Data Storage**

Besides LAN or WAN, the data can also be stored over internet in form of Cloud Storage or Hybrid Data Storage

* Complete cloud-based or online storage solutions offer virtual data storage and convenient access to your materials from anywhere, not just a local computer or external hard disk. Reliability tends to be on point, but organizations need to consider a cloud storage security strategy before implementing.
* Hybrid Model is the combination of cloud and on-premises data storage. Budget-friendly and flexible, hybrid cloud storage offers a secure and compliant option that helps to assure business continuity. This type of data storage accommodates frequent backups and long-term archives as well as future scaling and always-on availability. The combination of cloud and on-premises storage adds a layer of safety to ensure data is protected and available, and storage space could potentially be unlimited.

**Benefits of Efficient Data Storage**

* Reliable data preservation
* Data continuity and accessibility
* Quicker and easier data recovery
* Flexible price points and capacity options
* Effective security for protected files

**The Stages of data processing**

Data processing occurs when data is collected and translated into usable information. Usually performed by a data scientist or team of data scientists, it is important for data processing to be done correctly as not to negatively affect the end product, or data output. Data processing starts with data in its raw form and converts it into a more readable format (graphs, documents, etc.), giving it the form and context necessary to be interpreted by computers and utilized by employees throughout an organization.

**Six stages of data processing**

**1. Data collection**

Collecting data is the first step in data processing. Data is pulled from available sources, including data lakes and data warehouses. It is important that the data sources available are trustworthy and well-built so the data collected (and later used as information) is of the highest possible quality.

**2. Data preparation**

Once the data is collected, it then enters the data preparation stage. Data preparation, often referred to as “pre-processing” is the stage at which raw data is cleaned up and organized for the following stage of data processing. During preparation, raw data is diligently checked for any errors. The purpose of this step is to eliminate bad data (redundant, incomplete, or incorrect data) and begin to create high-quality data for the best business intelligence.

**3. Data input**

The clean data is then entered into its destination (perhaps a CRM like Salesforce or a data warehouse like Redshift), and translated into a language that it can understand. Data input is the first stage in which raw data begins to take the form of usable information.

**4. Processing**

During this stage, the data inputted to the computer in the previous stage is actually processed for interpretation. Processing is done using machine learning algorithms, though the process itself may vary slightly depending on the source of data being processed (data lakes, social networks, connected devices etc.) and its intended use (examining advertising patterns, medical diagnosis from connected devices, determining customer needs, etc.).

**5. Data output/interpretation**

The output/interpretation stage is the stage at which data is finally usable to non-data scientists. It is translated, readable, and often in the form of graphs, videos, images, plain text, etc.). Members of the company or institution can now begin to self-serve the data for their own data analytics projects.

**6. Data storage**

The final stage of data processing is storage. After all of the data is processed, it is then stored for future use. While some information may be put to use immediately, much of it will serve a purpose later on. Plus, properly stored data is a necessity for compliance with data protection legislation like GDPR. When data is properly stored, it can be quickly and easily accessed by members of the organization when needed.

**What is 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.
* Data visualization is one of the steps in analyzing data and presenting it to users.

In his 1983 book *The Visual Display of Quantitative Information*, [Edward Tufte](https://en.wikipedia.org/wiki/Edward_Tufte) defines 'graphical displays' and principles for effective graphical display in the following passage: "Excellence in statistical graphics consists of complex ideas communicated with clarity, precision and efficiency. Graphical displays should:

* show the data
* induce the viewer to think about the substance rather than about methodology, graphic design, the technology of graphic production or something else
* avoid distorting what the data has to say
* present many numbers in a small space
* make large data sets coherent.
* encourage the eye to compare different pieces of data.
* reveal the data at several levels of detail, from a broad overview to the fine structure.
* serve a reasonably clear purpose: description, exploration, tabulation or decoration
* be closely integrated with the statistical and verbal descriptions of a data set.

**Regression Analysis**

In statistical modeling, regression analysis is a set of statistical processes for estimating the relationships between a dependent variable and one or more independent variables. **Regression analysis** is a reliable method of identifying which variables have impact on a topic of interest.

Main types of Regression Analysis

**Linear and Logistic regressions**

* In statistics, linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables.
* In statistics, the logistic model is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick.

**Data Prediction/ Predictive Analytics**

Predictive analytics encompasses a variety of statistical techniques from data mining, predictive modelling, and machine learning, that analyze current and historical facts to make predictions about future or otherwise unknown events.

We may invoke 5 steps for data prediction.

1. Define the business result you want to achieve
2. Collect relevant data from all available sources
3. Improve the quality of data using data cleaning techniques
4. Choose predictive analytics solutions or build your own models to test the data
5. Evaluate and validate the predictive model to ensure robustness

**Data Classification**

**Data classification** is broadly defined as the process of organizing data by relevant categories so that it may be used and protected more efficiently. On a basic level, the classification process makes data easier to locate and retrieve. Data classification is of particular importance when it comes to risk management, compliance, and data security.

Data classification involves tagging data to make it easily searchable and trackable. It also eliminates multiple duplications of data, which can reduce storage and backup costs while speeding up the search process. Though the classification process may sound highly technical, it is a topic that should be understood by organization’s leadership.

**REASONS FOR DATA CLASSIFICATION**

Data classification has improved significantly over time. Today, the technology is used for a variety of purposes, often in support of data security initiatives. But data may be classified for a number of reasons, including ease of access, maintaining regulatory compliance, and to meet various other business or personal objectives.

**TYPES OF DATA CLASSIFICATION**

There are **three main types of data classification** that are considered industry standards:

* **Content**-based classification inspects and interprets files looking for sensitive information
* **Context**-based classification looks at application, location, or creator among other variables as indirect indicators of sensitive information
* **User**-based classification depends on a manual, end-user selection of each document. User-based classification relies on user knowledge and discretion at creation, edit, review, or dissemination to flag sensitive documents.

**Data Clustering**

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. Data **Clustering** is a process which partitions a given **data** set into homogeneous groups based on given features such that similar objects are kept in a group whereas dissimilar objects are in different groups.

**Difference between classification & clustering**

Although both techniques have certain similarities, the **difference** lies in the fact that **classification** uses predefined classes in which objects are assigned, while **clustering** identifies similarities between objects, which it groups according to those characteristics in common and which differentiate them from each other.

**Recommender Systems**

A **recommender system**, or a **recommendation system** is a subclass of information filtering system that seeks to predict the "rating" or "preference" a user would give to an item. They are primarily used in commercial applications. There are various algorithms used in recommender systems.

Recommender systems are utilized in a variety of areas and are most commonly recognized as playlist generators for video and music services, product recommenders for online stores, or content recommenders for social media platforms and open web content recommenders.