Data Acquisition and Learning Aspects in AI

- Data acquisition is the process of sampling signals that measure real-world physical conditions and converting the resulting samples into digital numeric values that a computer can manipulate.
- Data acquisition systems (DAS or DAQ) convert physical conditions of analog waveforms into digital values for further storage, analysis, and processing.
- In simple words, Data Acquisition is composed of two words: Data and Acquisition, where data is the raw facts and figures, which could be structured and unstructured and acquisition means acquiring data for the given task at hand.
- Data acquisition meaning is to collect data from relevant sources before it can be stored, cleaned, preprocessed, and used for further mechanisms. It is the process of retrieving relevant business information, transforming the data into the required business form, and loading it into the designated system.

The data acquisition involves:

- **Collection and Integration of the data:** The data is extracted from various sources and also the data is usually available at different places so the multiple data needs to be combined to be used. The data acquired is typically in raw format and not suitable for immediate consumption and analysis. This calls for future processes such as:
  - **Formatting:** Prepare or organize the datasets as per the analysis requirements.
  - **Labeling:** After gathering data, it is required to label the data. One such instance is in an application factory, one would want to label the images of the components if the components are defective or not. In another case, if constructing

a knowledge base by extracting information from the web then would need to label that it is implicitly assumed to be true. At times, it is needed to manually label the data.

The Data Acquisition Process

The process of data acquisition involves searching for the datasets that can be used to train the Machine Learning models. Having said that, it is not simple. There are various approaches to acquiring data, here have bucketed into three main segments such as:

1. Data Discovery
2. Data Augmentation
3. Data Generation

Each of these has further sub-processes depending upon their functionality.

- **Data Discovery:**

The first approach to acquiring data is Data discovery. It is a key step when indexing, sharing, and searching for new datasets available on the web and incorporating data lakes. It can be broken into two steps: Searching and Sharing. Firstly, the data must be labeled or indexed and published for sharing using many available collaborative systems for this purpose.

2. **Data Augmentation:**

The next approach for data acquisition is Data augmentation. Augment means to make something greater by adding to it, so here in the context of data acquisition, we are essentially enriching the existing data by adding more external data. In Deep and Machine learning, using pre-trained models and embeddings is common to increase the features to train on.

3. **Data Generation:**

As the name suggests, the data is generated. If we do not have enough and any external data is not available, the option is to generate the datasets manually or automatically. Crowdsourcing is the standard technique for manual construction of the data where people are assigned tasks to collect the required data to form the generated dataset. There are automatic techniques available as well to generate synthetic datasets. Also, the data generation method can be seen as data

augmentation when there is data available however it has missing values that need to be imputed.

Learning Aspects in AI

- Computational learning theory (CoLT) is a branch of AI concerned with using mathematical methods or the design applied to computer learning programs.
- It involves using mathematical frameworks for the purpose of quantifying learning tasks and algorithms.
- It seeks to use the tools of theoretical computer science to quantify learning problems. This includes characterizing the difficulty of learning specific tasks.
- Computational learning theory can be considered to be an extension of statistical learning theory or SLT for short, that makes use of formal methods for the purpose of quantifying learning algorithms.
    - Computational Learning Theory (CoLT): Formal study of learning tasks.
    - Statistical Learning Theory (SLT): Formal study of learning algorithms.

Learning

1. **Supervised Learning**: Training models on labeled data to predict outcomes.
2. **Unsupervised Learning**: Finding hidden patterns in unlabeled data.
3. **Reinforcement Learning**: Learning through trial and error to achieve long-term goals.