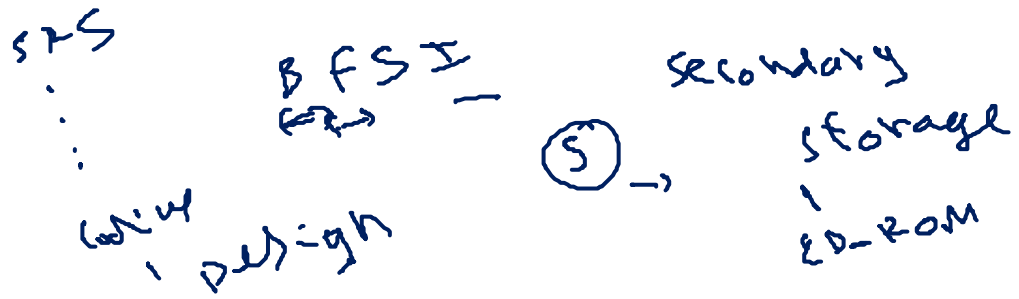
Stages of Data life cycle

Stages of Data life cycle

- The data life cycle is the sequence of stages that a particular unit of data goes through from its initial generation or capture to its eventual archival and/or deletion at the end of its useful life.
- The various phases of a typical data lifecycle are;

1. **Data Creation** - Data can be created in different forms like PDF, image, Word document or SQL database data and in any organisation, data is created in one of the 3 ways:
   - **Data Acquisition**: acquiring already existing data which has been produced outside the organisation. Converting real world signal to digital for display, storage.
   - **Data Entry**: manual entry of new data by personnel within the organisation.
   - **Data Capture**: capture of data generated by devices used in various processes in the organisation.

Stages of Data life cycle



2.Storage - Once data has been created, it needs to be stored and protected, with the appropriate level of security. A robust backup and recovery process should also be implemented to ensure retention of the data during the lifecycle.

3.Usage - During the usage phase of the data lifecycle, data is used to support activities in the organization. Data can be viewed, processed, modified and saved. An audit trail should be maintained for all critical data to ensure that all modifications to data are fully traceable. Data may also be made available to share with others outside the organization.

4.Archival - Data archival is the copying and removal of data from all active production environments where it is stored in case it is needed again, but where no maintenance or general usage occurs. If necessary, the data can be restored to an environment where it can be used.

5.Destruction - Data destruction or purging is the removal of every copy of a data item from an organization and is typically done from an archive storage location. The challenge of this phase is to ensure that the data has been properly destroyed. It is important to ensure before destroying data that the data items have exceeded their required regulatory retention period.

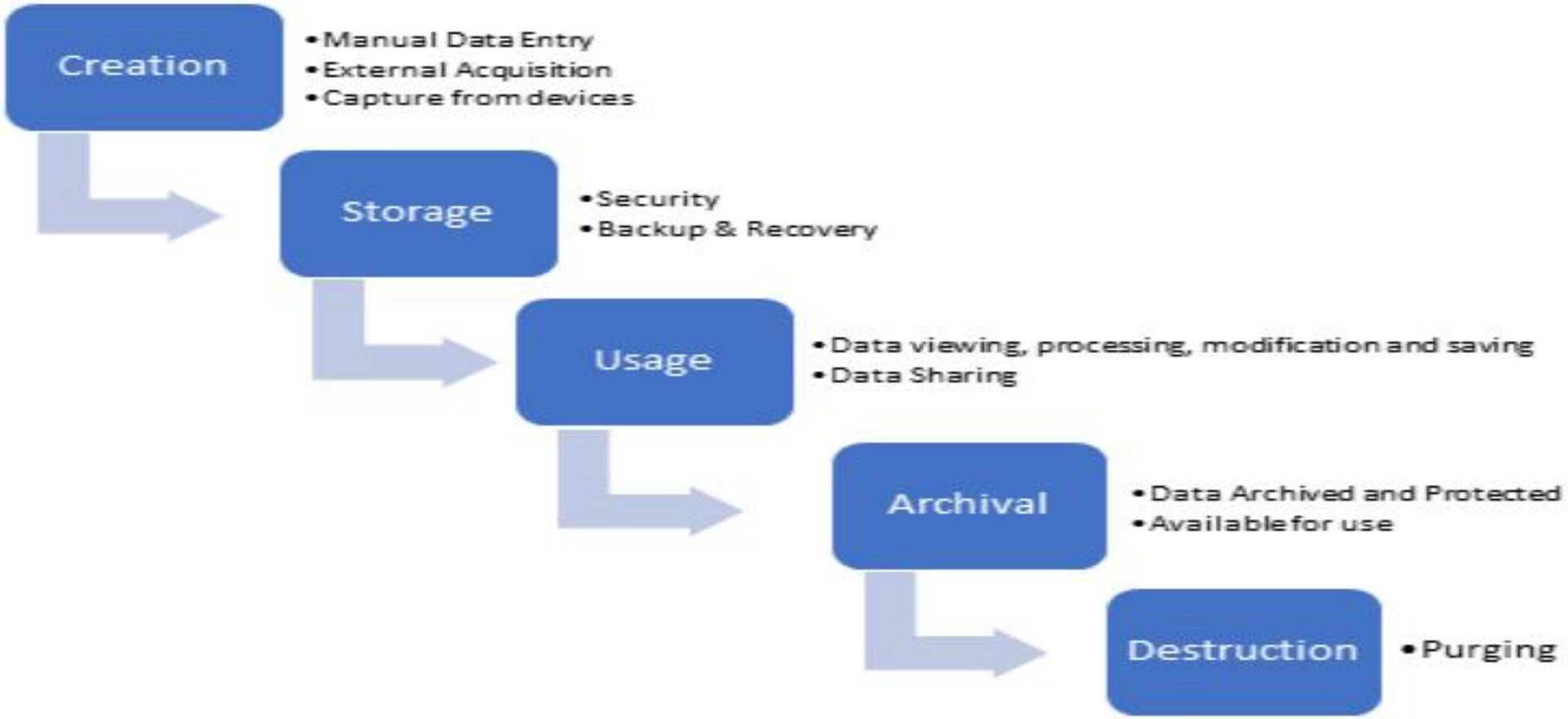
IoT data processing

- IoT Devices are able to collect data from their environment.
- **IoT cloud services to process and store IoT data**, or support real-time event processing from IoT devices.
- IoT protocols enable it to exchange data in an organized and significant manner.
- IoT data collection is the process of using sensors to track the conditions of physical things. Devices and technology connected over the Internet of Things (IoT) can monitor and measure data in real time.

IoT data processing

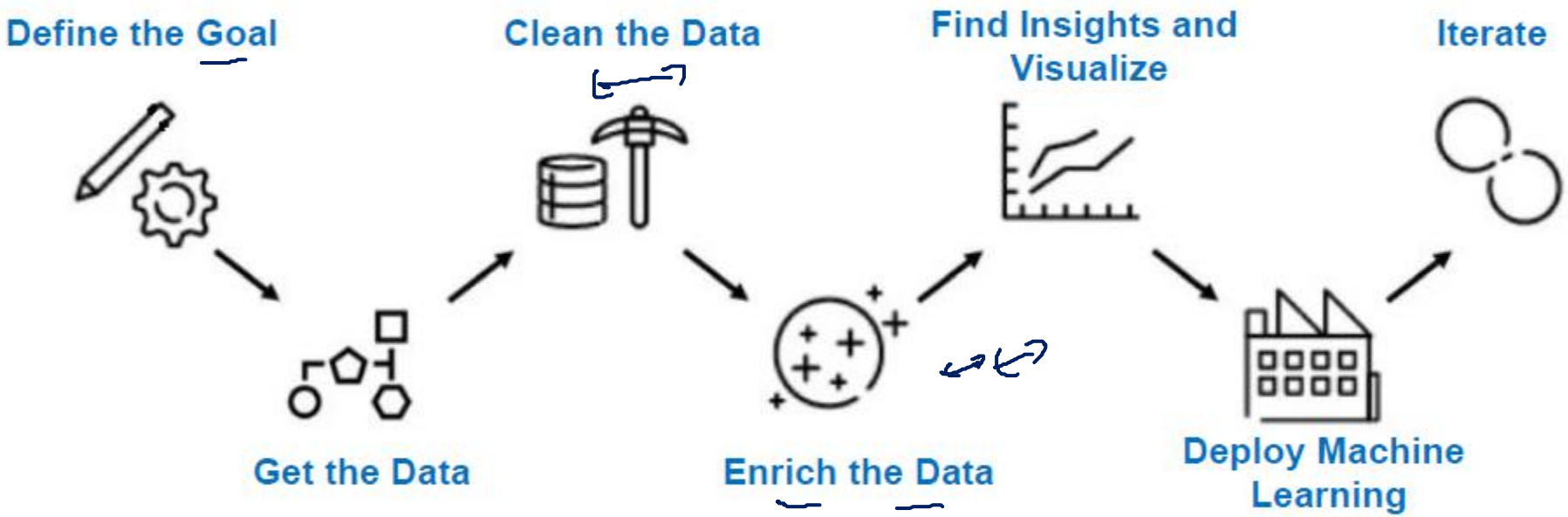
- Devices create data that is sent to the main application, consumed and used.
- Time-series data can be created as events take place around the device and then sent. This use of real-time information provides a complete record for each device.
- IoT devices share this data through an IoT gateway or platform or another IoT device. Data is typically sent to the cloud for storage and analysis.
- **Tapes, disks, flash storage**, hard drives, and cloud data archiving are all possible storage mediums.

Overview of the stages of Data life cycle



Steps of a data analytics project

Fundamental steps of a data analytics project plan(AI, Machine Learning and Big DATA)



These seven data science steps ensure realization of business value from each unique project and mitigate the risk of error.

❑ Define the Goal

- Understand the business or activity of the data project
- Identify a clear objectives of what you want to do with data

❑ Get your Data

- Once the objectives are identified, collect and aggregate the data from different sources

<u>Ways to collect the data</u>

- ❖ Query database(using technical skills like MySQL to process the data)
- ❖ Scrape from the websites
- ❖ Connect to Web APIs
- ❖ Obtain data directly from files (downloading it from Kaggle or existing corporate data which are stored in CSV)

❑ Explore and Clean the Data

Clean the Data:

- Process of organizing and tidying up the data, removing what is no longer needed, replacing what is missing and standardizing the format of the collected data.

Steps Involved

- Convert the data from one format to another and consolidate everything into one standardized format
- Filter the data
- Extracting and Replacing values(missing data sets or they could appear to be non-values, this is the time to replace them accordingly)
- Split, Merge and Extract data (For example, the place of origin, contains both “City” and “State”. Depending on the requirements, the data can be either merged or split)

Explore the Data

- Once the data is ready to be used, and before processing using AI and Machine Learning, the data need to be examined
  - figure out the business question and transform them into a data science question
  - inspect the data and its properties
  - compute descriptive statistics to extract features and test significant variables. Testing significant variables often is done with correlation. (e.g. exploring the risk of someone getting high blood pressure in relations to their height and weight)

Enrich the Dataset

- Process of enhancing, refining, and improving raw data.

❑ Build Helpful Visualizations

- Visualization is the best way to explore and communicate the findings and is the next phase of data analytics project.
- Graphs are also another way to enrich the dataset and develop more interesting features
- Data visualization to helps to identify significant patterns and trends of the data. A better picture of the data can be obtained through simple charts like line charts or bar charts further to understand the importance of the data.

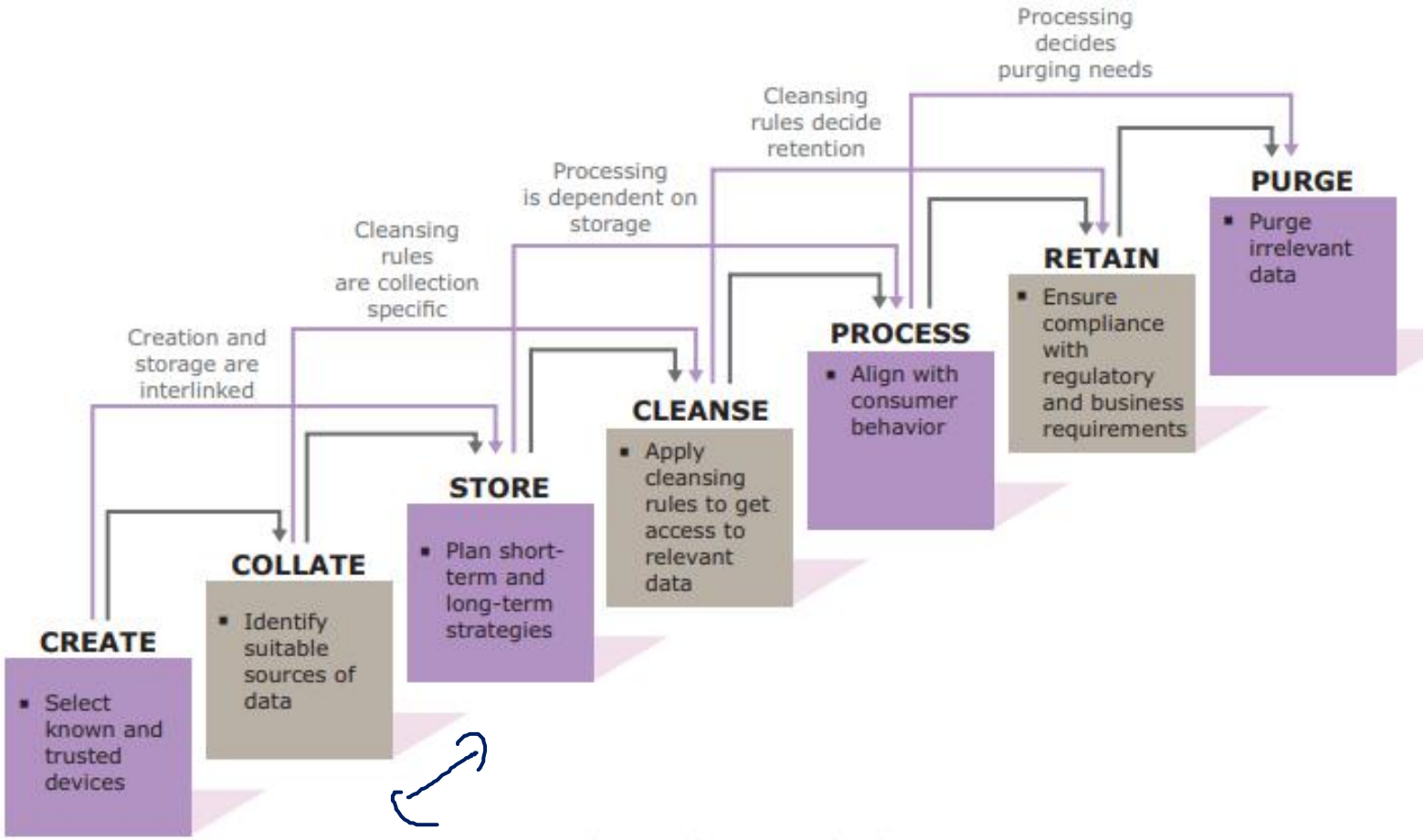
❑ Get Predictive

- Machine learning algorithms help in getting insights and predicting future trends
- Reduce the dimensionality of the data set and select the relevant ones which contribute to the prediction of results.

❑ Iterate

- To make the data useful and accurate, need to constantly reevaluate, retrain it, and develop new features

IoT Data lifecycle Management



IoT Data lifecycle

- Data creation: to be aligned with business case. Selection of right devices and end-user utilities is a major key. These measures help to sustain data creation to produce highly usable data. Generated data to be ready for processing, optimizing transformation and cleansing.
- Collation: Identify suitable sources to collate data relevant for business. Frequency, trigger points and mode of data collection to be defined for standardized process.
- Storage: Critical security. Storage mechanism to be scalable for accommodating more data, cost, availability and co-exist with enterprise. Big data are suited for storing IoT data.

IoT Data lifecycle

- Cleansing: proper rules for existing requirements and flexibility to handle future changes. Cleaning mechanism to be light weight to reduce overheads. Retention of critical parameters.
- Processing: business case and data characteristics should align with the modes of data collection and consumption. Optimized performance and cost.
- Example: data related to consumer behavior depends on period of operation.
- Retention: criteria for data retention, spanning duration, frequency. Identify relevant data and data elements periodically to ensure optimal use of data.

IoT Data lifecycle

- Archival: Regulatory requirements across industries. Suitable for the use-case, cost effective, easy accessible.
- Purging: strategy to prevent business risks. Purging to be done on archived data