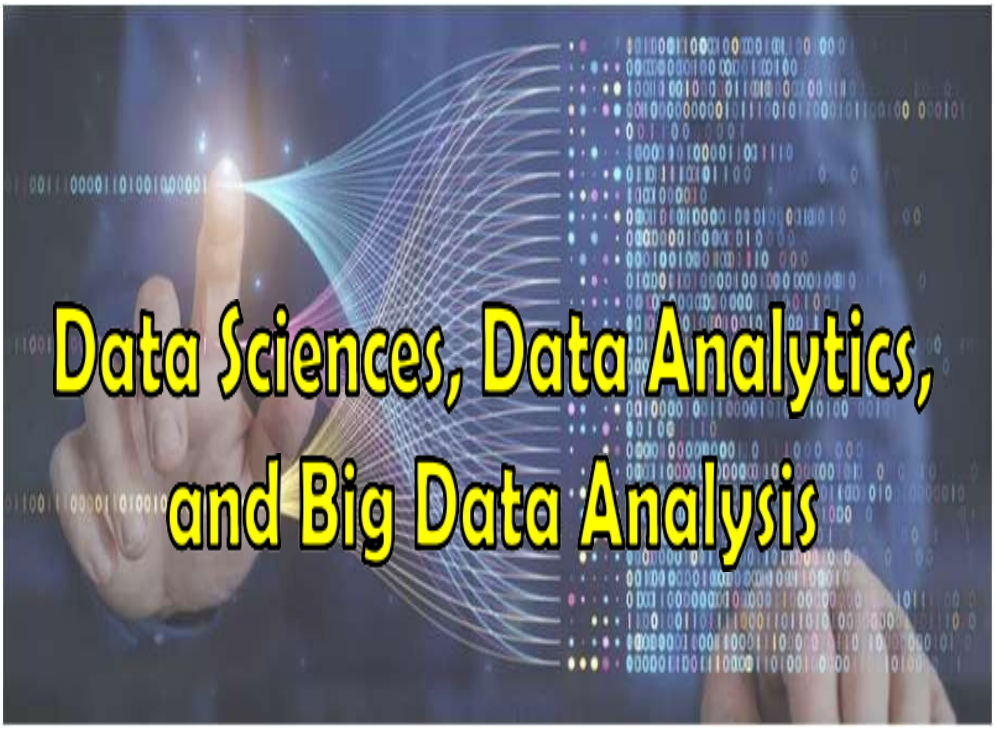


22nd -25th April, 2025

4 Session Hands On Training

C-Tea, Mysuru

Agenda: Day 1

- **Foundation Module**
  - Data Science : Broad Spectrum
  - ML, DL & RL
  - Programming Languages for DS and their significance
  - Classification of ML models
  - Supervised, Unsupervised, and Reinforcement Learning
  - Types of Data Sources
  - Data Processing Challenges
  - Data Structures for Handling Structured Data
  - Important Libraries – numpy, panda, scipy.stats, seaborn
  - Data Preparation steps and processing

Day 1

FOUNDATION MODULE

1. DATA SCIENCES – A BROAD SPECTRUM

Data Sciences

- **Data Science** is about unlocking the potential of data to drive innovation, solve problems, and create business value.
- It's a dynamic and ever-evolving field that requires a blend of technical skills, creativity, and ethical awareness.

Data Sciences

- **Data Science** is a multi disciplinary and dynamic field that combines statistics, computer science, and domain expertise to extract valuable insights and knowledge from data.
- Interdisciplinary in nature
- Compute intensive
- Algorithmic & Math centric
- Demands Continuous Learning

DATA SCIENCES & DATA ANALYTICS

Data Science

- **Data Science**:
  - **Scope**: Broad field encompassing various techniques to extract knowledge and insights from raw data.
  - **Methods**: Involves machine learning, statistical analysis, data mining, and more.
  - **Applications**: Predictive modeling, natural language processing, image recognition, etc.
  - **Goal**: To create predictive and prescriptive models to answer complex questions and solve problems.

9

9

Data Analytics

- **Data Analytics**:
  - **Scope**: Subset of data science focused primarily on analyzing / examining data to gain insights.
    - Analysis – know the present state from data
    - Analytics – includes estimations of future trends
  - **Methods**: Uses descriptive statistics, data visualization, and exploratory data analysis.
  - **Applications**: Business intelligence, performance metrics, trend analysis, etc.
  - **Goal**: To understand current and historical data and make informed decisions based on it.

10

10

Bid Data Analysis

- **Big Data Analytics**:
  - **Scope**: Focuses specifically on analyzing **very large** and complex datasets (big data).
  - **Methods**: Utilizes advanced tools and technologies that support distributed computing, such as Hadoop, Spark, and NoSQL databases.
  - **Applications**: Real-time data processing/analysis, recommendation systems, fraud detection, etc.
  - **Goal**: To handle and analyze vast amounts of data that traditional methods cannot efficiently process.

11

11

What is Data Analytics?

Data Analysis / Analytics is the use of:

data,

information technology,

statistical analysis,

quantitative methods, and

mathematical or computer-based models

to gain improved insight about business operations and make better, fact-based strategic decisions.

12

12

Analysis / Analytics

- **Data analysis**
- The process of evaluating data to extract meaning and generate insights. Data analysis focuses on exploring data in its raw form and is used to understand what has happened in the past.
- Descriptive
- Interpreting, collecting, transforming, and visualizing data to discover valuable insights that drive smarter and effective decisions related to the business.

13

13

Analysis / Analytics

- **Data analytics**
- The process of using data and analytical tools to find new insights and make predictions.
- Predictive
- Data analytics has a broader scope, and encompasses the entire data lifecycle, from collection to presentation. The goal of data analytics is to help organizations make informed decisions.

14

14

How of Data Analysis

- Python – Many Libraries (called Modules) that have abundant number of functions
- R – A powerful statistical software an data analysis tool
- Excel – 475 functions are available
- And many more ..
- *Data Analysis functions are easy to use, yet powerful mechanisms to work on your data to decipher hidden patterns!*

15

15

TYPES OF ANALYSIS

16

16

Types of Analysis

DESCRIPTIVE ANALYSIS

PREDICTIVE ANALYSIS

DIAGNOSTIC ANALYSIS

PRESCRIPTIVE ANALYSIS

EXPLORATORY DATA ANALYSIS

INFERENTIAL ANALYSIS

QUALITATIVE ANALYSIS

17

17

Types of Analysis, contd

- **Descriptive Analysis**
  - **Purpose**: Summarizes and describes the main features of a dataset.
  - **Examples**: Mean, median, mode, standard deviation, and data visualizations like bar charts and histograms.
  - **Use Case**: Understanding the overall distribution and basic characteristics of the data.

18

18

Types of Analysis, contd

- **Predictive Analysis**
  - **Purpose**: Uses historical data to make predictions about future events.
  - **Examples**: Regression analysis, time series forecasting, and machine learning models.
  - **Use Case**: Forecasting sales, predicting customer behavior, and risk assessment.

19

Types of Analysis, contd

- **Diagnostic Analysis**
  - **Purpose**: Investigates the reasons behind past outcomes or events.
  - **Examples**: Root cause analysis, drill-down analysis, and correlation analysis.
  - **Use Case**: Identifying the causes of specific trends or anomalies in the data.

20

Types of Analysis, contd

- **Prescriptive Analysis**
  - **Purpose**: Provides recommendations for decision-making based on data analysis.
  - **Examples**: Optimization models, simulation, and decision trees.
  - **Use Case**: Determining the best course of action, such as inventory management and supply chain optimization.

21

Types of Analysis, contd

- **Exploratory Data Analysis (EDA)**
  - **Purpose**: Uncovers patterns, trends, and relationships in the data without specific hypotheses.
  - **Examples**: Data visualization, summary statistics, and clustering.
  - **Use Case**: Gaining initial insights and forming hypotheses for further analysis.

22

Types of Analysis, contd

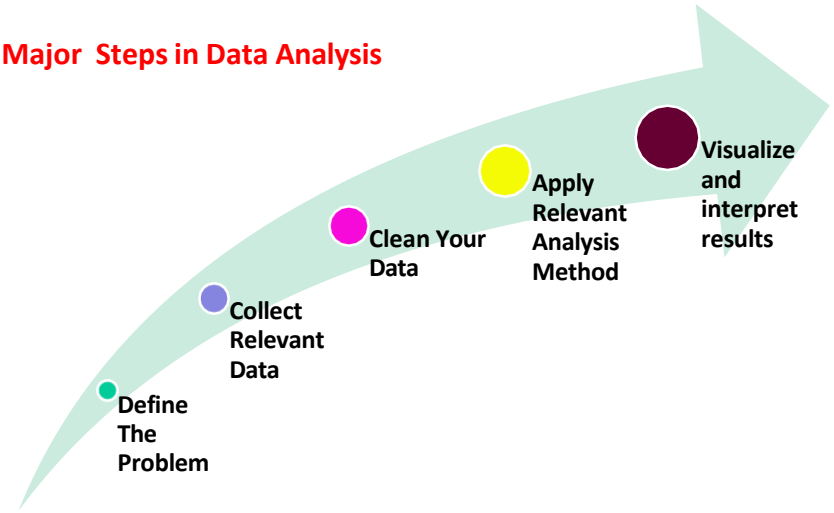
- **Inferential Analysis**
  - **Purpose**: Makes inferences about a population based on a sample of data.
  - **Examples**: Hypothesis testing, confidence intervals, and regression analysis.
  - **Use Case**: Generalizing findings from a sample to a larger population.

Types of Analysis, contd

- **Qualitative Analysis**
  - **Purpose**: Analyzes non-numeric data to understand concepts, opinions, or experiences.
  - **Examples**: Content analysis, thematic analysis, and narrative analysis.
  - **Use Case**: Analyzing open-ended survey responses, interviews, and social media content.

Data Analysis Pipeline

Major Steps in Data Analysis



25

25

2. MACHINE LEARNING: ML, DL & RL

26

26

ML, DL and RL

- **Machine Learning (ML), Deep Learning (DL), and Reinforcement Learning (RL)** constitute the foundational pillars of AI, enabling advancements in areas such as natural language processing, image recognition, and autonomous decision-making

- What is AI ?

27

Artificial Intelligence

- Ability to perform tasks normally requiring human intelligence, such as **visual** perception, **speech** recognition, **decision-making**, and **translation** between languages.
- Ability of a computer program or a machine to **think** and **learn**
- Ability to correctly interpret external data, to learn from such data, and to use those learning to achieve specific goals and tasks through flexible adaptation
- Ability to mimic human **cognition**
- A program that can **sense**, **reason**, **act** and **adapt**

28

AI Types

- **Narrow AI (weak AI)**:
  - AI systems that are designed and trained for a specific task or a narrow range of tasks.
  - Examples: Voice assistants (like Siri or Alexa), customer churn prediction, recommendation systems, image recognition software.
  - Capabilities: Can perform predefined tasks efficiently but lacks general intelligence and cannot perform tasks outside their specialization.

29

AI Types

- **General AI (strong AI)**:
  - AI systems that possess the ability to understand, learn, and apply knowledge across a wide range of tasks, similar to human's cognitive abilities.
  - Examples: Hypothetical at present; no such AI exists yet.
  - Capabilities: Could potentially perform any intellectual task that a human can do, including reasoning, problem-solving, and adapting to new situations.

30

AI Types

- **Generative AI**:
  - focuses on creating models capable of generating new content, such as text, images, music, and even videos.
    - Language Models – GPT 4,
    - Interactions – ChatGPT
    - Mixed Mode Transformers -Dall-E etc
  - These models learn from vast amounts of data and use that knowledge to produce original and often impressive outputs.

31

AI Types

- **Agentic AI**:
  - Artificial Intelligence systems that possess the ability to act autonomously, make decisions, and perform complex tasks without human intervention.
  - Use sophisticated reasoning, iterative planning, and real-time adaptation to solve multi-step problems and execute tasks
  - can interact with other software and systems to perform tasks such as processing transactions, managing bookings, or optimizing supply chains

New Arrival!

32

What is Human Learning?

- To learn: *to get knowledge of by study, experience, or being taught*
- *Learning guitar from a supervisor(trainer)*
- *After learning, he plays guitar*

33

33

Machine Learning

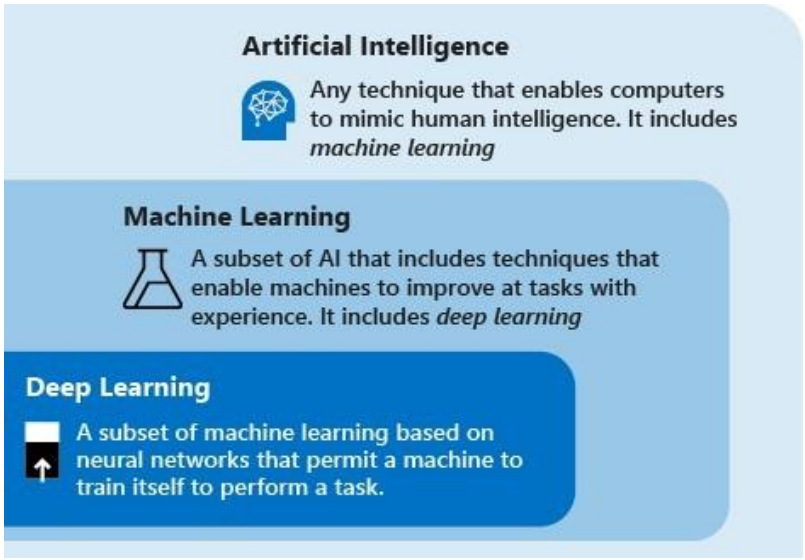
- A Robot is being trained by computer

- The trained robot plays itself

34

34

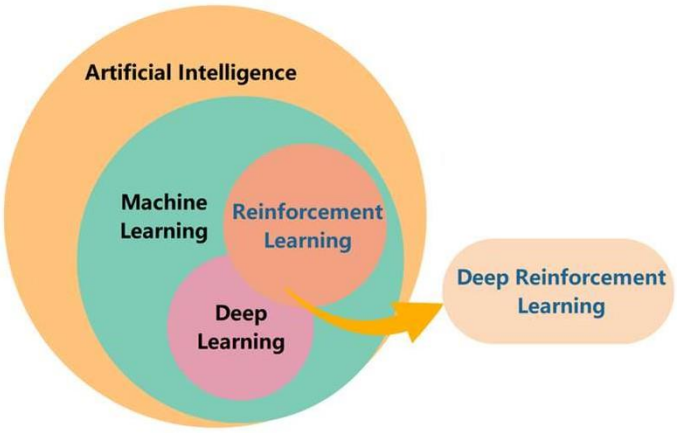
AI vs. ML vs. DL



35

35

AI, ML, DL and RL



36

36

AI - ML - DL - RL

AI is the effort to automate human intellectuals tasks by machine. Machines mimic Human intelligence.

Machine learning is to acquire knowledge from features of data that enables machines to improve at tasks with experience.

Deep Learning is to learn from data in successive layers of low level features to higher level features in deep neural network architecture .

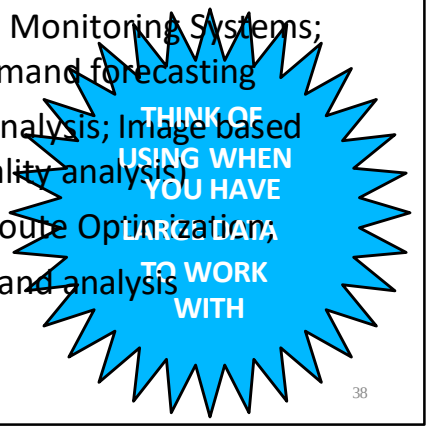
Reinforcement Learning involves learning through experience by using an agent that learns in an uncertain environment to achieve a goal by trial and error.

37

37

Some Example Application Areas

- **Sales**: Demand Forecasting; Discount offerings; Product bundles, Customer segmentation
- **Production**: Prediction of Downtime; Productivity Analysis; Anomaly detection; Monitoring Systems; Inventory Management & demand forecasting
- **Quality Control**: Inspection Analysis; Image based fault identification; (weld quality analysis)
- **Logistics** & **Transportation**: Route Optimization;
- **Procurement**: Vendor rating and analysis
- **Conversational**: ChatBots;



38

38

Financial Services

- Customer Service
- Backoffice Processes
- Customer profiling for Loan processing
- Fraudulent transaction detection
- Potential referrals customers from social media activity
- Debt collection assistance
- AI Chatbot

39

Manufacturing

- Predictive Maintenance
- Collaborative Robots (aka cobots)
- Correlation of process parameters to detect anomalies in real time (use Operational data from Sensors and IT data from enterprise systems) to reduce unplanned downtime
- Warehouse management
- Product demand prediction for production planning

40

Sales & Marketing

- Identify potential markets
  - Customers
  - Areas
  - Products
  - Price model
- Campaign management
- Target Customer identification
- Recommendation systems
- Best price bid (dynamic pricing based on real time supply, demand and other factors: ex Uber/ Spicejet)

41

41

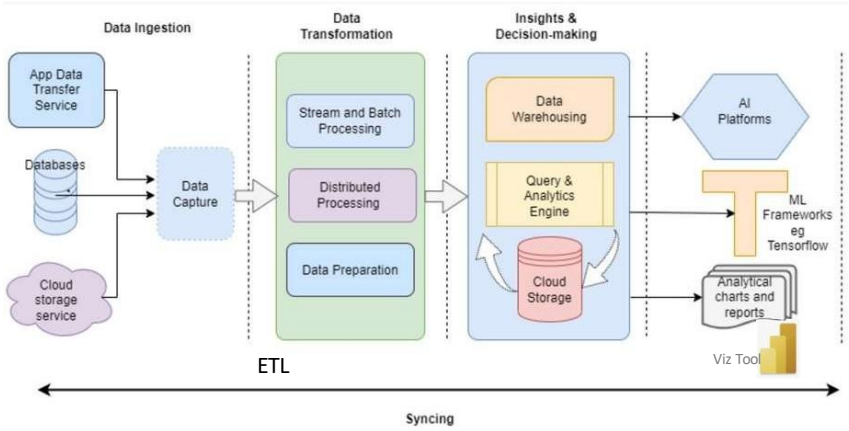
Procurement

- Spend Classification
- Vendor Matching
- Capturing supplier data from other media (social media channels etc) for new supplier identification and vendor classification
- Anomaly detection (track sudden changes in rates / delays in shipment
- Bid Management (automatic classification of bid data)
- Supply chain risk analysis
- NLP in contract management

42

42

The DA-BI Pipeline Technologies @ Work



MACHINE LEARNING

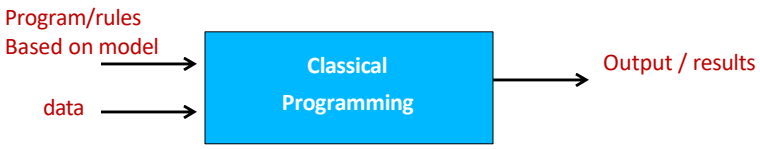
Machine Learning

- Machine Learning (Mitchell 1997)
  - Learn from past experiences (training)
  - Improve the performances of intelligent programs
- **Definition**
  - **A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at the tasks improves with the experiences**

Machine Leaning

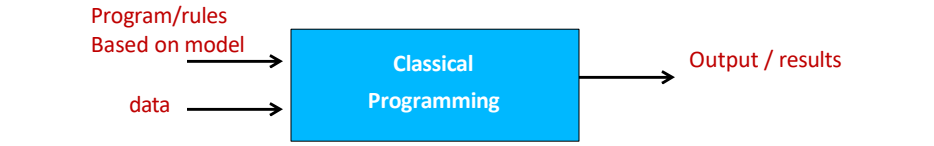
- A field of AI that enables computers to learn from data and improve their performance **without explicit programming**.
- It uses algorithms to identify patterns in the data, creates a model, make predictions, and automates decision-making.
- Learns from past experiences (training)
- Improves the performance of intelligent programs

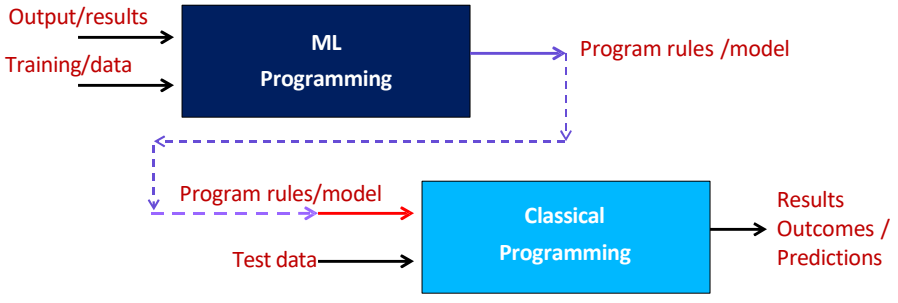
Classical Programming vs ML Programming



47

Classical Programming vs ML Programming





48

48

How these work?

- Access to data (more the merrier)
- Identify ‘features’ that define the behaviour
- Learn how these features relate to, or influence, or drive the outcomes
- Use this learning to handle outcomes on new data

49

Example

Identify the component / inspect the component

50

Example

Identify the component / inspect the component

51

51

The Machine Learning Framework

- Task: Given an image, identify the object
- Apply a prediction function to a **feature representation** of the image and get the desired output:

f() = “gear”

f() = “bearing”

f() = “bolt”

Ex: Estimate what feature values will uniquely identify the gear, or the bearing, or the bolt

52

52

The Machine Learning Framework

- The feature representation: [minimum set of input attributes]
- Identification rules are derived from these features
- Study of a few 'typical' samples gives the 'model' below:

features

Is Round	Is Elongated	Has teeth	Has spheres	Has chain	Has groove	Class
y	n	y	n	y	n	Gear
y	n	n	y	n	n	Bearing
y	y	n	n	n	Y	Bolt
y	n	y	n	n	y	Gear

Need more input samples for better understanding of the features and hence the model

53

53

Machine Learning Framework

- What about these inputs?

54

54

The Machine Learning Framework

- Task: Given an image, identify the object
- Apply a prediction function to a feature representation of the image to get the desired output:

f() = “gear ”, “bearing”

f() = “bearing”, ”gear”

f() = “bolt”

55

The Machine Learning Framework

- The feature representation after learning from more seen cases:

Is Round	Is Elongated	Has teeth	Has spheres	Has chain	Has groove	Class
y	n	y	n	y	n	**Gear**
y	n	n	y	n	n	**Bearing**
y	n	y	y	n	n	**Gear**
y	n	y	y	y	n	**Bearing**
y	n	y	y	n	n	**Bearing**
y	n	y	n	n	n	**Gear**
y	y	n	n	n	Y	**Bolt**
n	y	y	y	n	x	**Bolt**

56

Feature Extraction

- How do we get the features?
- How do we measure / quantify?

  A. Directly measure using equipment

  B. If cannot be measured directly, extract them using some processing / transformation methods.

  – Ex: If you have an image (obtained from a digital camera) then you can use image processing techniques to estimate the features and also quantify them.

57

1. Training Phase

- The model learns from a **training dataset**, which contains input data and correct output labels (in supervised learning).
- The model adjusts its parameters to minimize errors and improve prediction accuracy.
  – Example: A spam filter learns from labeled emails (spam vs. not spam).

58

2. Testing Phase

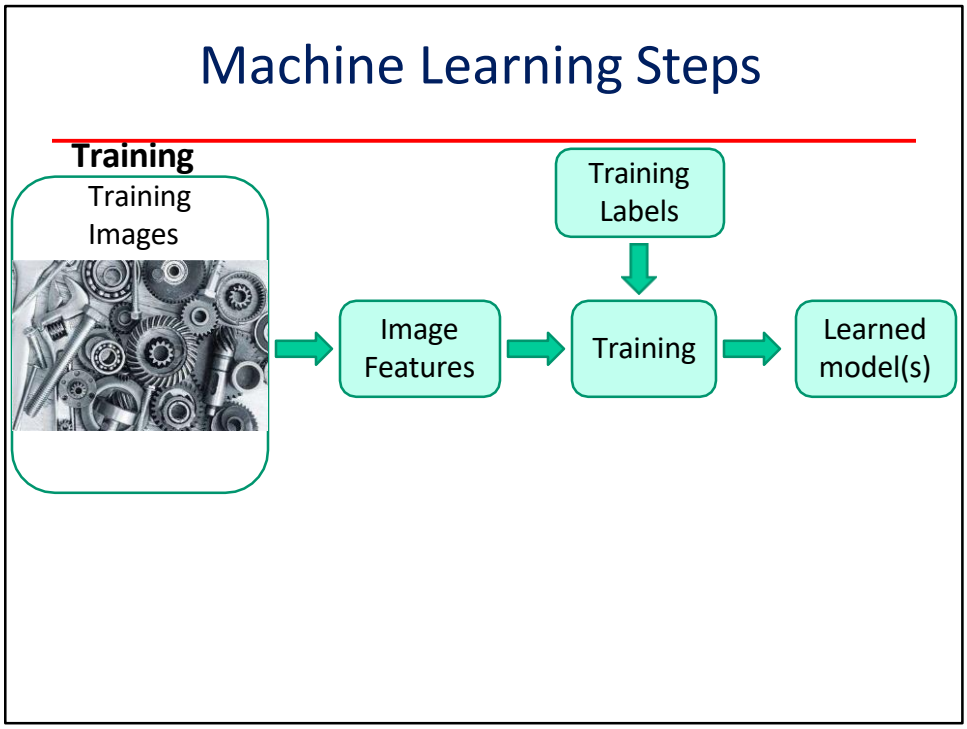
- After training, the model is tested on a **separate dataset** (testing set) that it has never seen before.
- This helps evaluate how well the model generalizes to new data.
  - Example: The spam filter is tested on new emails to check if it correctly identifies spam.
- Periodically we evaluate the effectiveness of the prediction and adjust the parameters for improved performance.

59

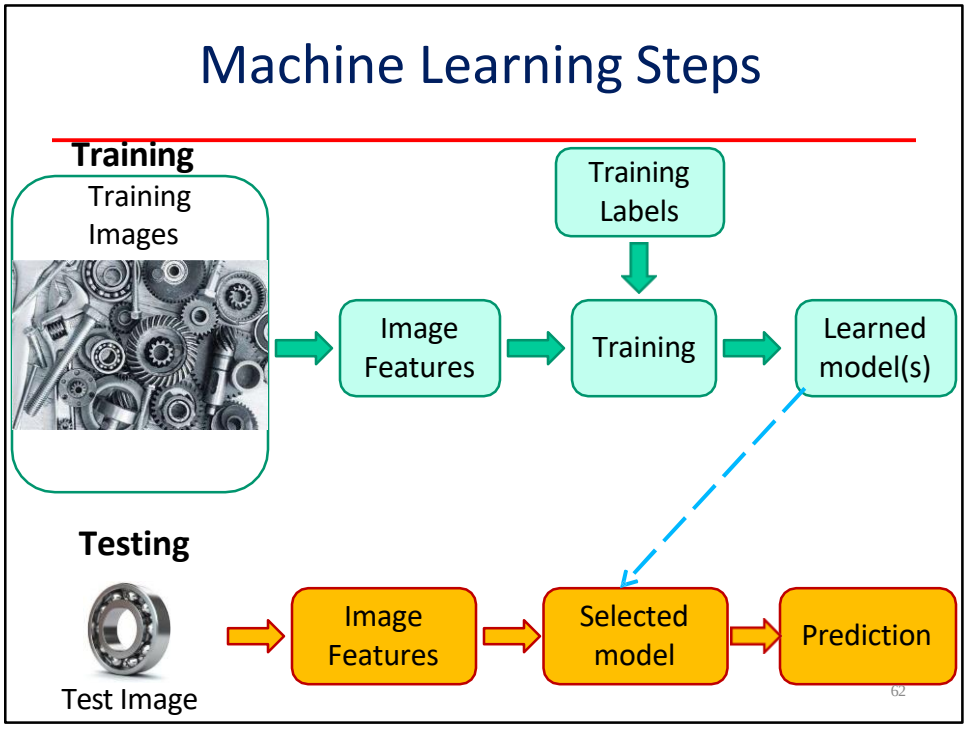
3. Validation Phase

- Sometimes, a **validation data set** is used to fine-tune hyperparameters before final testing.
- It prevents overfitting and ensures the model performs well on unseen data.
- Usually, a small part of the training data (seen and labelled) is used as the validation data.

60



61



62

The Machine Learning framework

$$y = f(\mathbf{x})$$

output — prediction function — Image feature

Training: given a *training set* of labeled examples $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$, estimate the prediction function f by minimizing the prediction error on the training set

- **Testing:** apply f to a never before seen *test example* $\mathbf{x}$ and output the predicted value $y = f(\mathbf{x})$

63

3. PROGRAMMING LANGUAGES & TOOLS FOR DATA SCIENCES / ML

64

Programming Languages for Data Sciences

- Programming Languages
  - Python
  - R
  - Java
  - SQL
  - Matlab
  - Scala

65

65

Programming Languages for Data Sciences

- Desktop Data Analysis Tools
  - MS Excel
- Data Visualization Tools (BI Tools)
  - MS Power Bi
  - Tableau
  - Google Charts / Google Analytics
  - Qlik Sense
  - Zoho Analytics
  - Plotly
  - Matplotlib/ Seaborn Libraries for Python

66

66

Software / Tools for Data Analytics

- Alteryx
- DBMiner 2.0 (Enterprise)
- Explorer from NXG
- IBM SPSS
- KnowledgeMiner
- Mathematica from Wolfram
- Pentaho – opensource BI
- RapidMiner
- SAS Enterprise Miner
- Teradata Analytics (TD Warehouse)
- TIBCO

67

Software for Data Analytics OpenSource

- Alteryx Project Edition (free version)
- KNIME
- ML-Flex
- Orange
- Weka
- Apache Spark for Big Data

68

Software for Data Cleansing (ETL)

- Ab Initio
- AMADRA
- Optimus – Python framework
- WinPure
- Zoho DataPrep

69

69

Software tools for Time Series Analysis

- EidoSearch
- KnowledgeMiner

70

70

Our Scope

- Python
  - With Supporting Modules for DA & ML
- MS Excel
  - With Data Analysis Toolpack
  - With add-ins
    - Power query
    - Power pivot

71

71

4. CLASSIFICATION OF ML MODELS

72

72

Classification of ML Models

- Machine learning (ML) models can be classified into different categories based on their learning approach and functionality.
  1. Based on Learning Approach
  2. Based on Functionality

73

73

Classification based on Learning Approach

- Supervised Learning – Models learn from labeled data.
  - Examples: Linear Regression, Decision Trees, Random Forest, Support Vector Machines (SVM).
- Unsupervised Learning – Models find patterns in unlabeled data.
  - Examples: K-Means Clustering, Principal Component Analysis (PCA), Autoencoders.
- Semi-Supervised Learning – Uses a mix of labeled and unlabeled data.
  - Example: Self-training models.
- Reinforcement Learning – Models learn by interacting with an environment and receiving rewards.
  - Examples: Q-Learning, Deep Q Networks (DQN), Proximal Policy Optimization (PPO).

74

74

Classification based on Functionality

- Classification Models – Predict discrete categories (e.g., spam detection).
  - Examples: Logistic Regression, Naïve Bayes, Decision Tree, K-NN, Neural Networks.
- Regression Models – Predict continuous values (e.g., stock prices).
  - Examples: Linear Regression, Polynomial Regression.
- Clustering Models – Group similar data points (e.g., customer segmentation).
  - Examples: K-Means, DBSCAN.
- Generative Models – Create new data similar to existing data.
  - Examples: Generative Adversarial Networks (GANs), Variational Autoencoders (VAE).

MODEL: THE BACKBONE FOR ML

Model

- An abstract view of the system's behaviour
- Mathematical representation of the system's functional behaviour, represented as a transfer function

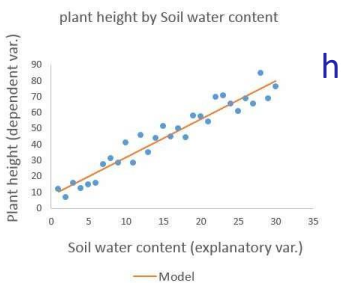
77

77

Model
ex: growth model of a plant

- Representation of a certain phenomenon

- Capture observations and algorithm prepares the model



h=b1*w+b0

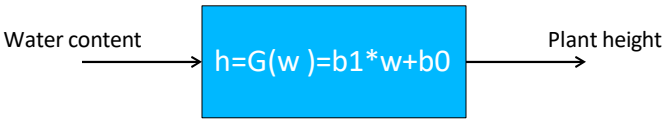
linear model

78

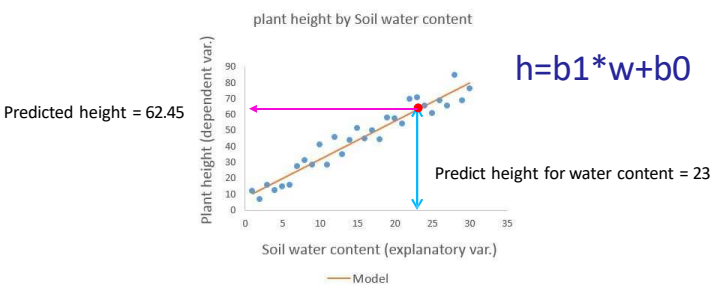
78

Model
ex: growth model of a plant

- Representation of a certain phenomenon



- Use the model to predict

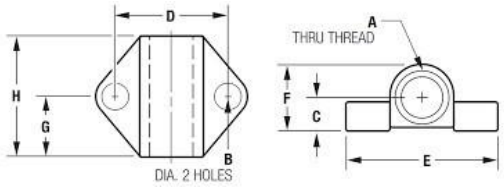


79

79

Case Study: Component Testing

- Consider the following component



- Drawing illustrates…
- There are 8 parameters (attributes or dimensions) which should be measured and verified to lie within limits
- Outcome: **REJECT**, **ACCEPT**, **REWORK**

80

80

Case Study: Component Testing

- Assume out of these 8, only 2 parameters are important (Principal Components) ex: H and D
- **Expert Design Engineer** devised the **rule(s)** which define the design limits are as follows
- **ACCEPT:**
  - H: 75mm ± 1% **and**
  - D: 50mm ± 1%
- **REWORK:**
  - H: greater than 75.75mm **or**
  - D: greater than 50.5mm
- **REJECT:** Otherwise

Rule Based Approach

- Rule Base:

  If (74.25 <= H <= 75.75) && (49.5 <= D <= 50.5)
  then **ACCEPT**
  else if (H>75.75 || D> 50.5)
  then **REWORK**
  else **REJECT**

Model Based Approach

- Assume an **expert** measures the principal parameters and **decides** the 'class'.
- These decisions of the expert is recorded and is used to understand the expert's working (called system behaviour/model)
- Apply this model on new data to check the outcomes
- Validate these outcomes with an expert and refine behaviour

Model Based Approach

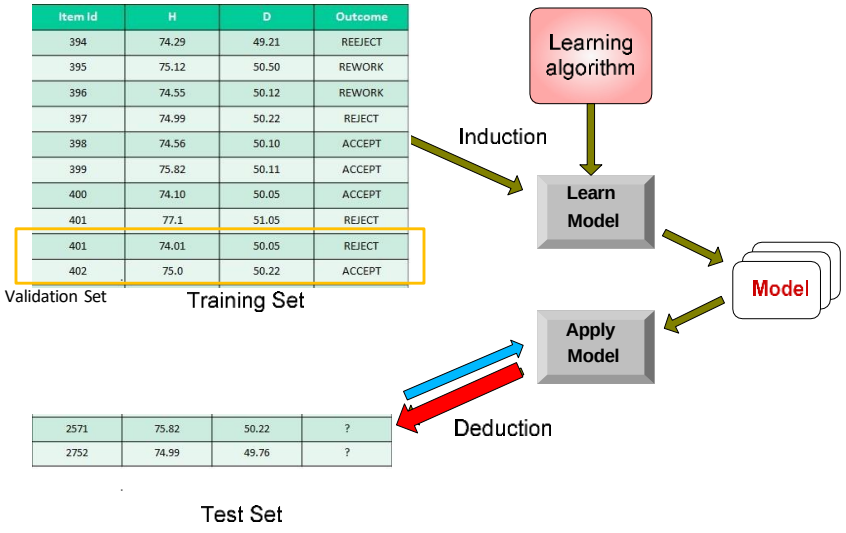
- Examine the training data and prepare a decision model.
- Use this model to identify the outcome for new unseen data

Seen data- training data

Item Id	H	D	Outcome
394	74.29	49.21	REEJECT
395	75.12	50.50	REWORK
396	74.55	50.12	REWORK
397	74.99	50.22	REJECT
398	74.56	50.10	ACCEPT
399	75.82	50.11	ACCEPT
400	74.10	50.05	ACCEPT
401	77.1	51.05	REJECT
401	74.01	50.05	REJECT
402	75.0	50.22	ACCEPT
..			
..			
2571	75.82	50.22	?
2752	74.99	49.76	?

Unseen data / test data

Model Based Approach



Popular Models in ML

- Decision Tree
- Bayesian Classifier
- KNN classifier
- Kmeans Clustering
- NN
- SVM
- Etc..

5.LEARNING

87

87

Learning..

- In machine learning, **learning** is the process by which a model improves its performance on a given task over time.
- This is achieved through repeated exposure to data and the application of algorithms that enable the model to recognize patterns, make predictions, and make decisions.
- These decisions are used to refine the model to ensure a more accurate model realization.

88

88

Learning Types

Main types are

- Supervised Learning
- Un supervised Learning
- Semi Supervised Learning
- Reinforced Learning

89

89

Supervised Learning

- **Supervised Learning**: Here, the model is trained on labeled data, which means the input data is paired with the correct output. The goal is for the model to learn the mapping between inputs and outputs so it can predict the output for new, unseen inputs.

90

90

Unsupervised Learning

- **Unsupervised Learning**: In this approach, the model is given data without explicit instructions on what to do with it. The model tries to find hidden patterns and structures in the data, such as clustering similar data points together or reducing data dimensionality.

91

Semi Supervised Learning

- **Semi-Supervised Learning** : Combination of supervised and unsupervised learning techniques. Use trained human input (supervisor) periodically to
  - Refine the training
  - Assess the efficiency and correctness of training
  - Learn newer concepts

92

Reinforced Learning

- **Reinforcement Learning**: This type of learning involves training an agent to make a sequence of decisions by rewarding it for desirable actions and penalizing it for undesirable ones. Over time, the agent learns to maximize its rewards by making better decisions.

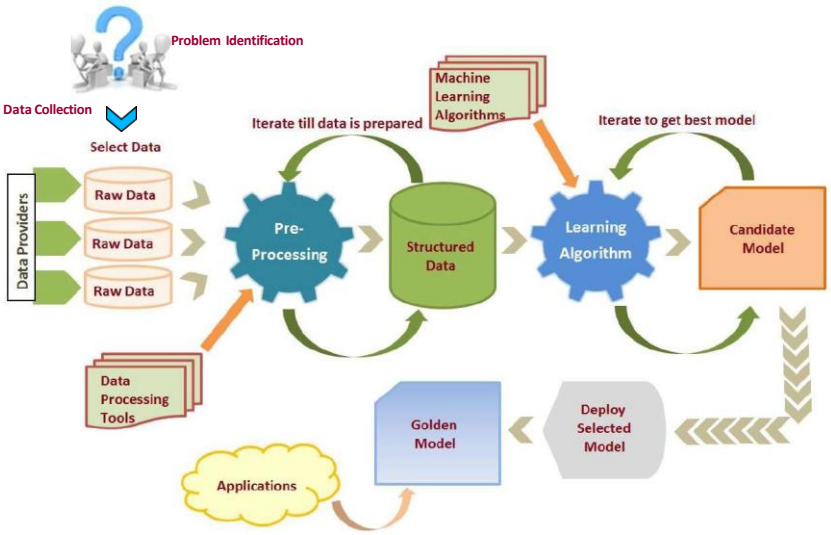
93

Machine Learning Pipeline

1. Understand the Problem Statement
2. Define the system with input and output variables
3. Assume a model Y = f (X1, X2, X3). The task of machine Learning is to find the function f( ).
4. Collect historical data on input and output variables
5. Preprocess the data to enhance the quality
6. Split the data into Training and Testing data
7. Training data are used for estimating f ( ) using machine learning algorithm
8. Validation data are used for validation of the above result and improvement of performance .
9. The system model in the form of f( ) is now used for prediction of output for future inputs (called the testing data) .
10. This is called supervised learning

94

Workflow of a ML project



95

95

6. TYPES OF DATA SOURCES

96

96

Types of Data

1. Structured Data

- Structured data is organized and stored in a tabular format, such as rows and columns. This type of data is common in databases and spreadsheets.
- **Examples**: Sales records, customer information, financial transactions.
- **Usage**: Useful in supervised learning tasks like regression and classification

97

97

Types of Data

2. Unstructured Data

- Unstructured data lacks a predefined format, making it more challenging to process.
- **Examples**: Text documents, e-mails, pdf, images, videos, audio files, social media posts etc.
- **Usage**: Found in applications like image recognition, natural language processing, and speech-to-text systems.

98

98

Types of Data

3. Semi-Structured Data

- Semi-structured data is partially organized data that lies between structured and unstructured data form. It has organizational elements but does not fit neatly into a tabular format.
- **Examples**: JSON files, XML files, and NoSQL databases, e-mails, social media posts, web pages etc.
- **Usage**: Often used in web scraping, API responses, and social media analysis

99

99

Representative Types of Data

- **Numerical Data**: Features measured in numbers (e.g., age, income, sales_amount, hours_worked).
- **Text Data:** Features that are represented as text (strings) which can be processed. (e.g., city_name, department, item_name)
- **Categorical Data**: Represents categories or labels (e.g., gender, fruit type).
- **Ordinal Data**: Categorical data with an inherent order (e.g., clothing sizes: Small, Medium, Large).
- **Special Data:** Data that has a pre-defined representation, like date, location (lat-long) etc

100

100

Types based on labels

- **Labeled Data**: Includes input feature variables and corresponding target output labels. Example: Features like “age” and “income” with a label like “loan approval status.”
- **Unlabeled Data**: Contains only input feature variables without any target labels. Example: Image features without annotations.

101

101

Data Sources…

- **Primary** : The data which is Raw, original, and extracted directly from the official sources is known as primary data.
  - Interviews
  - Surveys
  - Observations at an activity
  - Experimental
  - Simulated

102

102

Data Sources…

- **Secondary**: Secondary data is the data which has already been collected, processed and reused again for some valid purpose.
  - Internal Sources
  - External Sources like 3rd party or government publications etc
  - Sensor data
  - Satellite Data
  - Web data
  - Social media

103

7. DATA PROCESSING CHALLENGES

104

Main Challenges

- **Data quality:** One of the biggest issues with using data in machine learning is ensuring that the data is accurate, complete, and representative of the problem domain. Low-quality data can result in inaccurate or biased models.
- **Data quantity:** In some cases, there may not be enough data available to train an accurate machine learning model. This is especially true for complex problems that require a large amount of data to accurately capture all the relevant patterns and relationships.
- **Bias and fairness:** Machine learning models can sometimes perpetuate bias and discrimination if the training data is biased or unrepresentative. This can lead to unfair outcomes for certain groups of people, such as minorities or women.
- **Overfitting and underfitting:** Overfitting occurs when a model is too complex and fits the training data too closely, resulting in poor generalization to new data. Underfitting occurs when a model is too simple and does not capture all the relevant patterns in the data.
- **Privacy and security:** Machine learning models can sometimes be used to infer sensitive information about individuals or organizations, raising concerns about privacy and security.

Data Related Challenges

- Biggest challenges are
  - Data availability
  - Data quality
  - Data labeling
  - Bias and fiarness

Data Availability

- Models need large datasets to learn effectively, but data may be scarce due to the rarity of events, high collection costs, or data-sharing restrictions
  - Employ data augmentation techniques and synthetic data generation mechanisms
- Data privacy and security concerns limit data availability
  - Regulatory aspects limit availability of data

107

107

Data Quality

- The data collected may not always be suited for analysis; it's often noisy, incomplete, and inconsistent. Data could be coming from heterogeneous sources.
- Noisy data has irrelevant information, incomplete data has missing values, and inconsistent data arises from discrepancies in format or values.
  - Implement robust data cleaning and preprocessing techniques. Use data quality tools and data validation techniques to catch and fix data quality issues early.

108

108

Data Labelling

- Supervised learning requires labeled data.
- Acquiring labeled data can be difficult, especially in fields requiring specialized knowledge, such as medical imaging or natural language processing.
- Manual labeling is slow, costly, and prone to error.
  - Using active learning and semi-supervised learning instead can reduce the need of manual labeling.

Bias and Fairness

- Bias in data refers to systematic errors or distortions or un representation in datasets that lead to unfair or inaccurate outcomes in machine learning models. It can arise from historical biases, sampling errors, or misrepresentation of certain groups.
  - For example, if a hiring algorithm is trained mostly on data from male applicants, it might unfairly favor men over women
- Fairness in data aims to ensure that AI models make equitable and unbiased decisions.

Data Shaping

- Transforming & Re structuring data to suit the modelling requirements
- Enhance the quality of data
- Eliminate / reduce errors in data

- Preprocess data to make it usable for Analysis

111

111

Typical Transformations

112

112

8. DATA STRUCTURES USED IN DA / ML

113

113

Structured Data Needs Data Structures

- **Structured data** is organized in a predefined format, making it easy to store, retrieve, and analyze. Various **data structures** help efficiently handle structured data.
- Structuring data allows
  - Systematic access to search and retrieve data
  - Establish efficient algorithms for data access operations
  - Specifies standard methods and understanding of the data representation
  - Can use features of programming languages to optimize operations

114

114

Python Data Structures

- Lists
- Tuples
- Sets
- Dictionaries

115

115

Derived Data Structures

- Numpy Arrays
- Stack
- Queues
- Linked Lists
- Trees
- Heap
- Graph
- Pandas Data Frames
- Panda Series
- Matrix
- String

116

116

Stored Data Structures

- Files
  - Sequential text files
  - Comma separated files
  - Tabular spreadsheets
  - html
  - Xml
  - Json and other formats of file storage
- Databases
  - SQL tables (records)
  - Columnar tables (key-value pair data)

9. IMPORTANT LIBRARIES IN PYTHON TO SUPPORT DA & ML

Python Libraries for Data Analytics & ML

Python Libraries for DA & ML

- Python supports many modules that are effective for data analysis / machine learning applications
- Popular are
  - Numpy
  - Pandas
  - Matplotlib & Seaborn
  - SciPy
  - Scikit-learn
  - nltk

Frameworks

- Popular frameworks are
  - TensorFlow
  - Keras
  - PyTorch

10. DATA PREPARATION STEPS & PRE PROCESSING

Data Preparation Steps

- Data preparation is a crucial step in any **machine learning (ML) project**, ensuring that the dataset is clean, structured, and ready for training.
  - Also called data cleansing, data refinement, data scrubbing, data enrichment, data wrangling

1. Data Collection

- Gather relevant data from various sources like databases, APIs, or files and ingest into the workspace.

123

Data Preparation Steps

2. Data Cleaning

- Remove missing values, duplicates, and inconsistencies to improve data quality**.**
  - Perform range checks, data validity checks, eliminate unwanted data etc.

3. Data Transformation

- Convert raw data into a usable format, including normalization, scaling, and encoding categorical variables**.**

124

Data Preparation Steps

4. Feature Engineering

- Create new features or modify existing ones to improve model performance**.**

5. Data Splitting

- Divide the dataset into training, validation, and testing sets to evaluate model accuracy**.**

6. Data Augmentation (Optional)

- Enhance the dataset by generating synthetic data or applying transformations**.**

Data Preparation Steps

7. Data Integration

- Combine multiple datasets to create a unified dataset for better insights.

8. Data Validation

- Ensure data consistency and correctness before feeding it into the ML model.

127