Thursday, May 8, 2025 12:04 PM

2 kinds of statistics:

1. Descriptive statistics
2. Inferential statistics

- Measures of central tendency: mean, median , mode
- Measures of dispersion: range, variance, standard deviation (sd measures spread of values)

Usually Standard deviation is used to express the data

T-test

Thursday, May 8, 2025 12:14 PM

- Used when sample size is small and population standard deviation is unknown.
- One-sample t-test: compare sample mean to population mean
- Two-sample t-test: compare means of two independent samples

Correlation coefficient

Thursday, May 8, 2025 12:15 PM

- Measures strength and direction of relationship between two variables.
- Range: -1(perfect negative) to +1(perfect positive), 0 (no correlation)
- If you want to calculate correlation coefficient, both variables must be continuous

Chi-square test

Thursday, May 8, 2025   12:17 PM

2 varibles must me categorical

ANOVA test

Thursday, May 8, 2025 12:17 PM

- Analysis of variance compares means across multiple groups
- One-way ANOVA: one independent variable
- Example: test average exam score among 3 different teaching methods
- One variable must be categorical and other continuous

z-test

Thursday, May 8, 2025 12:18 PM

- Used when population variance is known and sample size is large.
- Example: test whether the average weight of sample weight of a sample differs from population mean.
- If we know population mean, we can follow z-test.
- If we don't know population mean, we can follow t-test

Central limit theorem

Thursday, May 8, 2025 12:23 PM

- CLT states that the distribution of sample means approaches a normal distribution as sample size increases.
- Example: averaging 100 samples of height, regardless of original distribution, forms a bell curve.
- Irrespective of what distribution it follows, if we take a sample of the population it follows normal distribution

Hypothesis testing

Thursday, May 8, 2025 12:27 PM

- A statistical method to determine the likelihood that a given hypothesis is true.
- null hypothesis: (H0): no effect or difference.
- Alternative hypothesis (H1): there is an effect or difference.
- Use p-value to decide whether to reject H0
- The column which has $p<0.05$, we consider that variable to be contributing in target variable

Feature engineering

Thursday, May 8, 2025 12:58 PM

1. Filter methods:
   Use statistical techniques to score features
   Example: chi-square test, ANOVA, correlation

2. Wrapper methods:
   Use predictive models to evaluate the feature subsets
   Examples: recursive feature elimination, forward selection

3. Embedded methods:
   Feature selection occurs during model training
   Example: LASSO, decision trees, random forest.

- Logistic regression should follow binomial distribution
- Parametric algorithms have mathematical formula and makes assumptions on the distributions of data
- For non-parametric algorithms, they don't have formula and don't make any assumptions on data distribution
- Linear and logistic regression are parametric algorithms, rest are non-parametric algorithms
- If we want to build a linear regression model, the data should be comply with linear regression assumptions
- There are 5/6 assumptions

Thursday, May 8, 2025 1:25 PM

Average of loss function from all the samples is called cost function

Evaluation metrices classification

1. Accuracy: (TP/TN)/total
2. Precision: TP(TP+FP)
3. Recall: TP(TP+FN)
4. F1 score: 2*(precision*recall)/(precision + recall)
5. ROC-AUC: trade-off between TPR and FPR

No target variable in unsupervised learning

1. Null value treatment
2. Handling categorical variable
3. Removing outliers depends on situation
4. No feature selection
5. No class imbalance
6. Feature scaling

clustering

Friday, May 9, 2025 12:22 PM

Clustering groups similar data points together
Eg: customer segmentation, document grouping
Common algo: k-means, Dbscan, hierarchical clustering

k-means clustering

Friday, May 9, 2025 12:26 PM

1. First we randomly initialise k points, called means or cluster centroids.
2. We repeat the process for a given number of iterations and at the end, we have our clusters.

Initialise k means with random values:

1. For a give number of iterations:
    a. Iterate through items
    b. Find the mean closet to the item by calculating the euclidean distance of the item with each of the means
    c. Assign items to mean

K values comes from:

1. Domain knowledge
2. Elbow method

The k value becomes the number of cluster centroids, ie, k number of clusters.
Each centroid calculates its nearest points and assigns that particular data point to into it's cluster

Hierarchical clustering

Friday, May 9, 2025 1:27 PM

Each data point starts as its own clusters.
The algo works as follows:

1. It calculates the distance between all pairs of clusters and merges the two closest clusters into a new cluster.
2. The distance between the new cluster and all other clusters are recalculated.
3. Steps 2 and 3 are repeated iteratively until all data points belong to a single cluster or until a certain stopping criteria is met.
4. The result is a dendrogram that visually represents the hierarchical relationship between data points. Based on the dendrogram, one can choose a suitable number of the clusters by cutting the tree at a specific

Anomaly detection

Friday, May 9, 2025 3:24 PM

- Identifies rare/unusual patterns that differ significantly from the norm.
- Example: fraud detection, network intrusive detection
- Common algo: isolation forest, one-class
- If we go with supervised learning, it comes under classification
- Anomalies are minor like 1%.
- Most of the entries are normal.

Class balancing will make both anomalies are normal data equal.
Anomalies will increase upto 30% or more in testing data if we build a classsification model. This happens

In case of banking domain, FP causes chaos.

Isolation forest

Friday, May 9, 2025 3:30 PM

The isolation forest algorithm works by identifying instances that are isolated or stand out from the rest of the data.

Isolation principle: the algorithm leverages the isolation principle, which stands that anomalies are often few in number and different from normal instances. As a result anomalies are expected to be easier to isolate.

Random partitioning: by randomly selecting features and splitting values, the algorithm creates a tree structure. Anomalies, requiring fewer splits to be isolated, end up closer to the root of the tree.

Anomaly score calculations: the anomaly score for each instance is determined by the average path length needed to isolate it across multiple trees. Shorter paths indicate that an instance is isolated more quickly suggesting it is likely an anomaly.

Threshold for anomalies: a threshold is set to distinguish between normal and anomalous instances. Instances with score above the threshold are classified as anomalies.

By focusing on the ease of isolation, the isolation forest algorithm efficiently detects anomalies in a dataset, marking it particularly useful for applications such as fraud detection, network security , or any scenario where identifying rare and unusual events is essential.

Overfitting and underfitting

Monday, May 12, 2025 1:04 PM

Overfitting:

Model works well on training data
False positive: type 1 error
False negative: type 2 error

Ways to reduce overfitting:

- Use simpler models (fewer parameters)
- Apply regularization (L1, L2)
- Use dropout in neural networks: LSTM, GRU, RNN
- Prune decision trees
- Cross-validation to monitor performance:
- Early stopping during training
- Increase training data
- We can apply this in case of linear regression
- L1- ridge, L2- lasso
- We add absolute coefficient value in error term in L1
- We add squared absolute coefficient value to error term in L2

Ways to reduce under fitting:

- Use more complex models (add layers/parameters)
- Add meaningful features
- Reduce regularization
- Train longer (more epochs)
- Improve data quality (data preprocessing)
- Try non-linear models eg decision trees, ensembles

F1 score: we are taking harmonic mean because precision and recall are rates.

ROC curve

Monday, May 12, 2025 1:53 PM

ROC = receiver operating characteristics curve
Plot: TPR (recall) vs FPR (false positive rate)
Shows performance across thresholds
Ideal curve hugs the top-left corner
Diagonal = random guessing

AUC score

Monday, May 12, 2025 1:58 PM

AUC = area under the ROC curve
Represents the degree of separability
AUC = 1 perfect prediction
AUC = 0.5 no discrimination
AUC < 0.5: less than random

Forecasting

Monday, May 12, 2025 3:09 PM

- Forecasting predicts future values based on historical data trends
- Target variable is continuous
- Example: stock market prediction, Sales forecasting
- Common models: ARIMA, SARIMA, prophets, LSTM (deep learning)
- The time column indicates how data is collected
- The way the data is collected, the same way data its forecasted

How does it differ from regression? we don't mention any time in regression.

Time series analysis:

- **Uni variate analysis:** we forecast only one variable
- **Multi variate analysis:** we predict more than one variable (all continuous variables)

No inference data
Data must be sorted data wise, ascending order

PreProcessing

Monday, May 12, 2025 3:34 PM

2 preprocessing steps only for forecasting:

1. trend: non repetitive pattern in a longer period of time.
2. Seasonality: repetitive pattern with fixed interval of time.

If we want to build seasonal algorithms, we have to remove trend and seasonality, then the data becomes stationary.

In case of deep learning, no need to follow these preprocessing steps, ie no need to remove trend and seasonality

Important points

Tuesday, May 13, 2025 12:16 PM

We have to perform 4 types of pipeline in classification, regression, clustering, anomaly detection, forecasting- no task is fulfilled until these are fulfilled.

- Training pipeline

- Inference/testing pipeline: data only has input variables, no target variable. Preprocessing steps in training pipeline are applicable to inference pipeline

- Retraining pipeline: model is retrained (base data patterns and testing data patterns are different)

- Deployment pipeline:

Data drift: refers to change in statistical properties of the input data over time, impacting model performance

Model drift: statistical properties of target variable may vary over the period of time.

Before predictions we must calculate data drift and model drift to decide if we must proceed with predictions or retrain.

Data Science

Tuesday, May 6, 2025 2:53 AM

Data Science life cycle:

1. Understanding the **business problem**
2. Preparing the data
3. Exploratory Data Analysis(EDA)
4. Modeling the data
5. Evaluating the model
6. Deploying the model

Data from clients:

1. In the form of API's
2. FLAT files
3. Credentials of Database

ML and DL algorithms work on structured data

Algorithms

Tuesday, May 6, 2025 3:00 AM

Machine Learning = python + pandas + numpy + Seaborn + Matplotlib, plotly + D-tale, Pandas-Profiling + scikit_learn + pickle + joblib

200-250 columns suitable for ml, after that dl will deal

ML algorithm

Tuesday, May 6, 2025 4:57 PM

1. Random forest
2. XGBoost
3. Gradient Boosting
4. Support Vector machine (SVM)
5. Decision Tree
6. K-nearest neighbour (KNN)
7. Logistic regression
8. Linear regression
9. Naïve Bayes

3 kinds of challenges can be solved using these algorithms:

1. Regression
2. Classification
3. forecasting

DL algorithm

Tuesday, May 6, 2025 4:57 PM

1. RNN (recurrent neural network)
2. LSTM (long short term memory)
3. GRU (gated recurrent neural network)
4. BiLSTM (long short term memory)

We can solve:

1. Forecasting
2. Machine translation
3. Regression and classification

DL catches limited number of words.

Types of Data

Tuesday, May 6, 2025 2:57 AM

Unstructured data
Semi-structured data
Structured data

Preparation of data(stage 2):
Unstructured and semi-structured data must be converted into structured data

Data cleaning/processing

1. Pandas
2. NumPy

Data visualisation

1. Matplotlib
2. Seaborn
3. Plotly

Auto-EDA libraries: D-tale, pandas profiling (quick action)

EDA

Tuesday, May 6, 2025 3:10 AM

1. Analyzing data column wise (statistical and referential analysis)
2. Mean median mode (calculate metric)
3. If insights can fulfill client requirements
4. Insights can be drawn using visual diagrams

Data PreProcessing

Tuesday, May 6, 2025 6:29 AM

Categorical variables: should be in int

*Nominal: no order (yes/no)

*ordinal: they have order (has ranking)

Continuous variables: int, float

Apart from this statistical methods, we have simple imputers and KNN imputer from scikit-learn

Dealing with missing values (1st preprocessing step)

Wednesday, May 7, 2025 5:00 PM

1. **Imputation of values(null/NaN):**
   1. Check how many values in each column are null.
   2. Check for datatypes of the columns which have null values.
   3. If the missing value belongs to **continuous variable**
      - If it follows **normal distribution**, we impute the missing value with **mean** of that column.
      - If it **doesn't follow ND**, try to impute missing value with **median** of the column
   4. If it is a **categorical variable,** nominal or ordinal, we try to impute the missing value with **mode**

Categorical variable treatment (2nd preprocessing step)

Tuesday, May 6, 2025 6:55 AM

- If a particular categorical belongs to nominal variable, we use **one-hot encoding:**
  - **pd.get_dummies()** from pandas library
  - one-hot encoder from scikit learn (production purpose)
- If it belong to ordinal variable use label encoder form scikit-learn

Outliers statement (3rd preprocessing step)

Wednesday, May 7, 2025 12:01 PM

- Outliers treatment is only **applicable to continuous variable.**
- Abnormal values, behavior and observation.
- There are many ways to treat these outliers.
- If continuous variable follows ND we use **jetscale** for identification of outliers.
- If a particular column doesn't follow ND we use IQR method (interquartile range).
- Z score value lies between -3 and 3. Rest all are considered outliers

ML Algorithms

Linear regression, logistic regression, KNN- building any of these models requires removing the outliers. These are distance based algorithms.

Gradient boosting, random forest, decision tree, Adaboost, XGBoost- building these models doesn't require removing the outliers. They have the capacity to deal with outliers

DL algorithms doesn't require removing outliers

Unsupervised technique: DBscan technique to remove outliers

Scaling (4th preprocessing step)

Wednesday, May 7, 2025 12:58 PM

- In order to make the training efficient.
- Scaling is only applicable to continuous variable.
- Scaling is only applicable to input variables, and not target variables.
- When columns have different ranges of values the column with higher scale or range dominates over the one with the lower range.
- Hence it becomes difficult to decide on the column that is required to be considered.
- scaling will be applied to all the continuous variables. It will to scaled to -1 to +1
- When the **target variable is continuous, it is regression.**
- If it is categorical, it is classification.
- If it follows nd, no need to scale
- When scaling is performed, mean is 1 and sd is 0

Distance based algorithms need feature scaling

Robust, standard and min max for scaling

Class Imbalance (5th preprocessing step)

Wednesday, May 7, 2025 1:25 PM

- Only applicable for target variables, if it is categorical.
- 2 of the class imbalance methods are over sampling and under sampling.
- Over sampling is always preferrable over under sampling

No need to perform class imbalance for regression.

Smot technique doesn't generate duplicate values

Important points

Wednesday, May 7, 2025 3:51 PM

- In training we use both input and target variables
- Inference pipeline, prediction pipeline, testing pipeline, we only use input variables.

GridSearchCV and RandomizedSearchCV, these 2 techniques for finding the best combination of hyperparameters

For anomaly detection it is better to implement unsupervised learning algorithms like:

1. isolation forest
2. local outlier factor
3. one class svm

Difference b/w **forecasting** and **regression:**

- Forecasting has 2 additional preprocessing steps: trend and seasonality
- In regression, we never talk about time.
- In forecasting we predict continuous variable, but we have to mention time dependency.
- **In regression, we have time column, but is it not considered a feature

1. Forecasting algorithms
2. LSTM (Long short term memory)
3. RNN
4. GRU
5. XGBoost
6. AdaBoost
7. Gradient boosting
8. SVM

9. Decision tree

For clustering 3 algorithms are most widely used:

1. Kmeans clustering
2. Hierarchical clustering
3. DBscan

Final PreProcessing

Thursday, May 8, 2025 4:43 PM

1. Finding negative values
2. Finding duplicate values
3. Null value treatment
4. Categorical variable treatment
5. Handling outliers
6. feature selection
7. Class balance (not for regression, only classification)
8. Feature scaling

Thursday, May 15, 2025 12:17 PM

system prompt is used to control LLM
by using system prompt, we are controlling the LLM response
it can be controlled upto a certain level
best 2 prompting: chain of thoughts, tree of thoughts
if the above 2 are not working, go w hybrid prompting

If we are using a reasoning model, there is no need for any chain of thoughts or tree of thoughts

1. persona based prompting

What is Generative AI

Wednesday, May 14, 2025 12:46 PM

Can build any application on any kind of data (structures, unstructured, semi-structured)

Comes under NLP (natural language processing)

Multimodal in GenAI

Wednesday, May 14, 2025 12:46 PM

Definition: a multimodal model can process and generate content across multiple data types like text, images, audio, video, code

Capabilities:

- Input: image + question -> output answer
- Input: audio -> output: text (transcription)
- Input: text -> output: image (text to image)
- Input: image -> output: code

Multilingual in GenAI

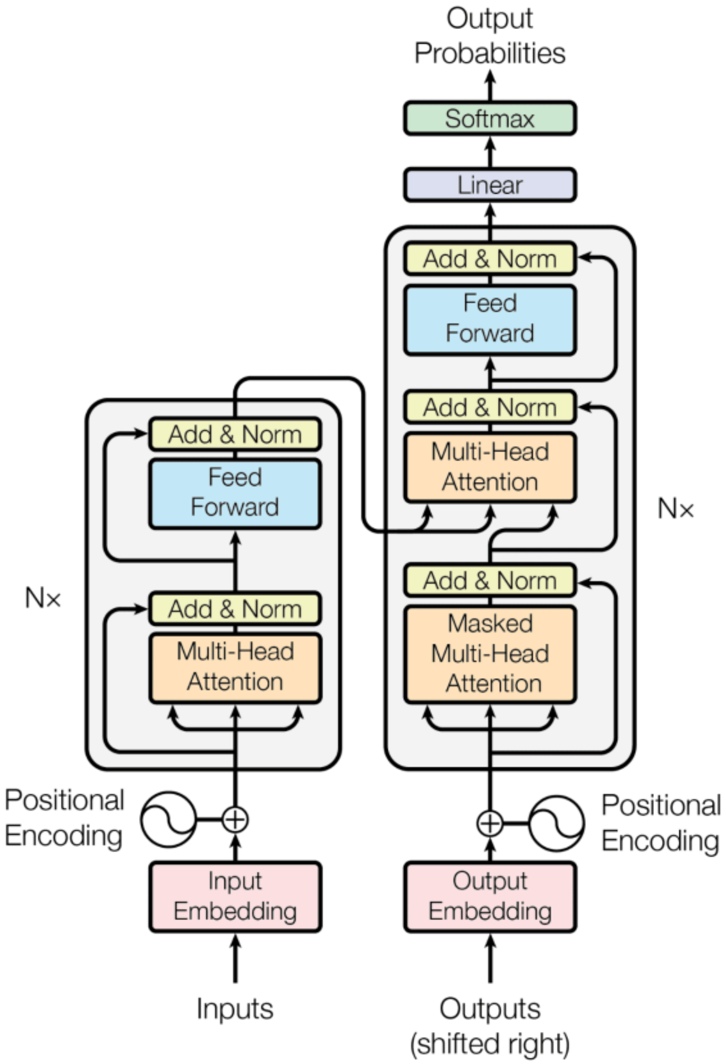
Wednesday, May 14, 2025 12:50 PM

Definition: a multilingual can understand and generate text in multiple languages.

Capabilities:

- Translate between languages
- Answer questions in the same or different language
- Generate responses in the user's native language

LLM's: large language models Building block: transformer architecture



Atleast 16GB of RAM, processor: more than 2.4GHz

Chatgpt models

Wednesday, May 14, 2025 1:02 PM

Coder model
Reasoning model
Question/answering model
Research model

The more parameter model we choose, the more is the accuracy
Always choose high billion parameter model

Ollama: open source framework which allows users to run large language models (LLMs) directly on their local machine

System prompt: we instruct the llm to respond to user query in a better way

Processing steps for text data

Thursday, May 15, 2025 12:17 PM

1. Lower casing
2. Removing numbers
3. Remove punctuations
4. Stemming: it cuts the word
5. Lamatization: language specific, if you give any word it will be converted into root word

Tokenization:
Embeding: convert text into numerical representations

Libraries for nlp text preprocessing:
NLTK
Scapy
Text Blob

RAG

Monday, May 19, 2025 12:08 PM

RAG: retrieval-augumented generation

- Used to make custom chatbots

Vector search: supports keyword and vector search
Mongo DB can be used in vector db

Traditional db support only keyword search

Based on embedding model context length we can decide no of chunks formed.

Any genAI application must have vector DB
Chroma DB
While using any open source db, make it telemetry disabled.

The query retrieved the related chunks

Introduction

Thursday, May 22, 2025 12:25 PM

Systematic way of implementation

Project: specialized to an individual/group of individual
Product: for everyone

Scrum -> framework under agile

Our working agreement:
Focus, openness, respect, courage, commitment
Main concept: self-organizing

- Came into existence in early 2000s
- 'Agile' term was coined in 2001 to describe **flexible nature of software developed** in **iterative stages** and become a blanket term for the new methodologies

Traditional

- Only one project manager and leader

Agile

- Self-organizing team, one person becomes the scrum master, servant leader facilitator.

The Agile Manifesto

Individuals and interactions	over	Process and tools
Working projects	over	Comprehensive documentation
Customer collaboration	over	Contract negotiation
Responding to change	over	Following a plan

The Agile Methodology

Thursday, May 22, 2025 1:10 PM

- Agile methodology: based on **iterative and incremental approach**
- Does not build an entire system at once, rather develops incrementally

The Agile Methodology- Scrum Framework

- The Team or Scrum Team: **product owner, scrum master, development team**
- **Development Team**: practice heads, architects, BA, project lead, team lead, Sr. Developers, Jr. Developers, UI/UX developer, QA lead, Sr. QA engineer, QA engineer, SQA
  (QA- quality analyst)

Sprint-Scrum Framework

Sprint:

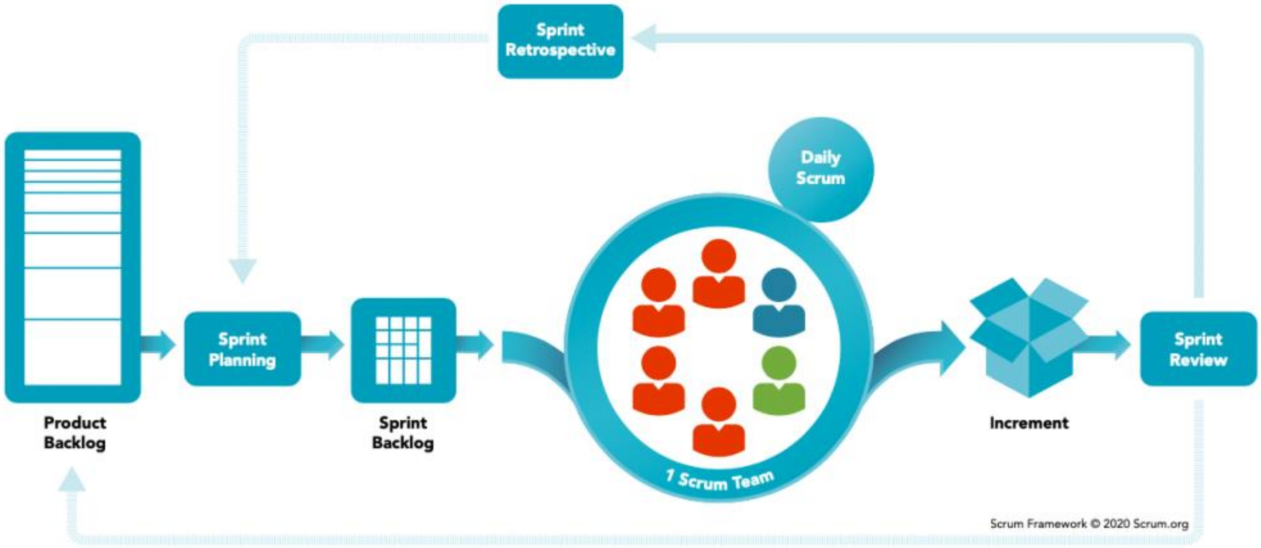
- Time box | fixed duration | fixed start and end time
- No gaps/overloads | same duration
- (Preparing, implementation, review, retrospect) - fixed length

Scrum Events - Scrum Framework

Scrum Events (4)

1. Sprint Planning
2. Daily Scrum(everyday)
3. Sprint Review (has to be there)
4. Sprint Retrospective

SCRUM FRAMEWORK



Scrum Framework © 2020 Scrum.org

Scrum Events

Thursday, May 22, 2025 1:15 PM

1. **Daily Scrum** (daily)

   - 15 mins for 2 weeks sprint and 30 mins for 4 weeks sprint (development team + scrum master + optional PO)
   - What did you do yesterday?
   - What will you do today?
   - Are there any impediments in your way?

2. **Sprint Planning** (at the end beginning of the sprint)

   - 4 hours (dev team + optional SM) for 2 weeks sprint.
   - Tasks to be done, technical concepts etc. for proper detailed planning.

3. **Sprint Review*** (at the end of the sprint and before sprint retrospective)

   - 2 hours (PO + stake holders (on PO invitation) + dev team + SM) for 2 weeks sprint.

4. **Sprint Retrospective** (at the end of sprint- after sprint review)

   - 1.5 hours (Scrum team with PO optional) for 2 wees sprint.

Important Points

Thursday, May 22, 2025 3:41 PM

Development Team

- Self-organizing team
- Cross functional team (t-shaped people instead of specialized people).

Cross-Functional Collaboration

1. Build Trust
2. Eliminate fear
3. Drive goals
4. Promote accountability
5. Team delivery

Virtual collaboration

Scrum Master Responsibilities

- Coaching scrum | process owner
- Replacing and changing resources
- Mentor
- Change agent
- Problem solver
- Resolve conflicts

POC format: user story format, acceptance criteria

Definition of Done (DOD)

1. UI/UX
2. Document mandatory & validations
3. Test case writing
4. Test case review

1. API's development
2. Development & unit testing
3. Code review

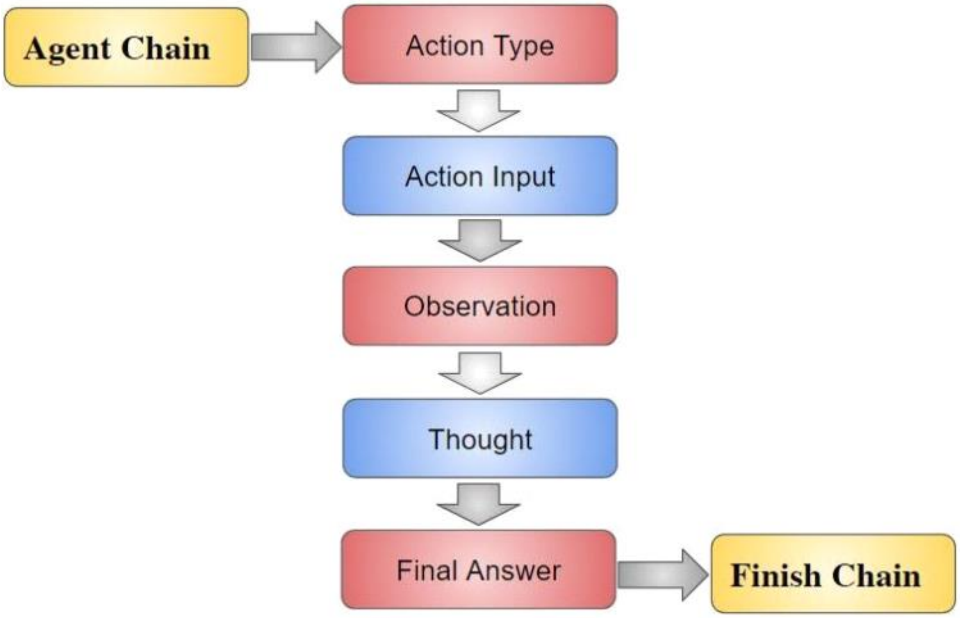
1. Test case execution
2. SQA - Review

LIM- Large Image Models

Peft- parameter efficient fine tuning
Rlhf: reinforcement learning with human feedback

We have to fine-tune embedding model as well.

Contextual memory: allows system to recall relevant info based on current task.



Some frameworks for agentic applications

1. LangChain: LLM ☑ Tools ✕ Memory ✕

- System contains components, how components interact with each other. At high level it is called HLD. At low level-LLD

Project Flow:

Business requirements -> functional requirements (making ourself) -> HLD (high level design) -> LLD (low level design) -> code

Visual only in HLD not LLD

Non-Functional Requirements:

1. Scaling
   Local system config: 16gb ram. If we perform stress testing, produces throughput value and latency value.

2. Security
   Data encryption pov -> data at rest, data at transmit (TLS)
   Application pov -> strong authentication and authorization.
   High availability -> its downtime should be less
   Reliability ->

Search if the same application is available and check what features are present.
List out tools required and compare.

CrewAI

Monday, May 26, 2025 6:13 PM

- Open-source AI-agent framework built from scratch. It offers 2 approaches to build an AI-agent application.
- CrewAI crews:

- V-Soft Consulting Group, Inc. (header and footer).
- Entire document in same font (time new roman).
- Document name: project name

Monday, June 2, 2025 12:40 PM

Case 1
If the client wants us to implement a functional requirements:

1. First check the time required and the budget to do so. Also check if there is any existing product with such functionalities available.
2. If no such product is available, do a research.
3. We need to make a clear MOM (minutes of meeting) and share with the people involved.
4. Also make a POC.

Frontend APIs

1. Registration API (name, age, email, gender, phone number)
2. Login API
3. Logout API
4. Feedback API
5. Password reset API

Gather atleast 15 functional requirements

- 2 dashboards: one client, one vendors (service providers)