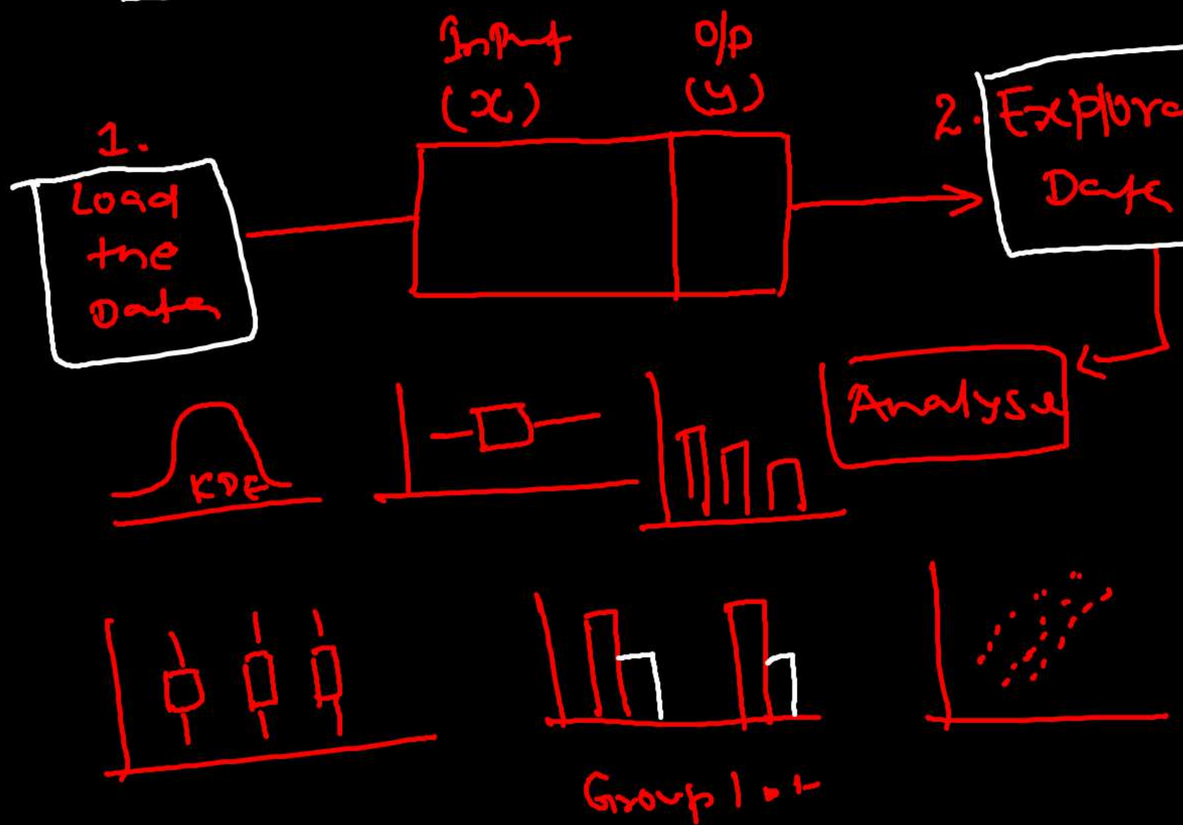
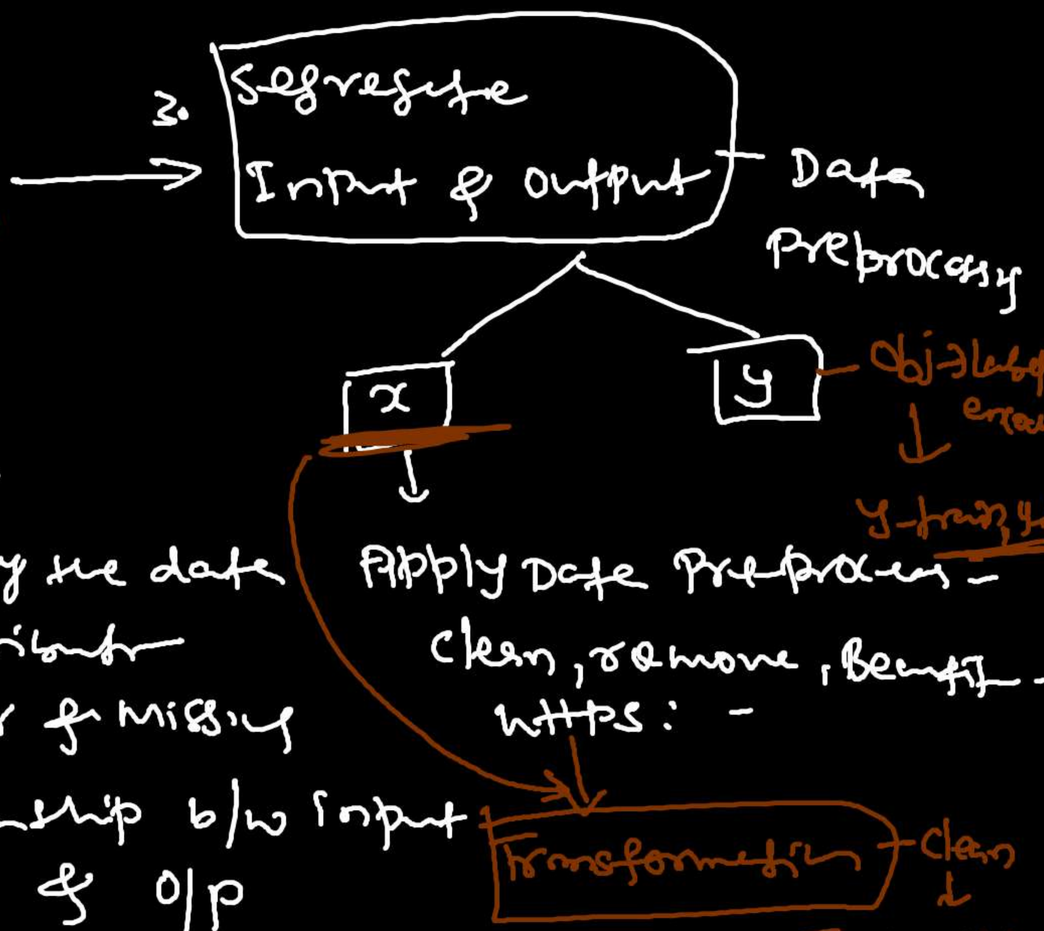
Introduction To NLP

Model Building pipeline



Objective

1. Identify the data distribution
2. outlier & missing
3. Relationship b/w input & o/p



* Learning Phase

Build the model

X_train_transformed → Input
→ Output

Algorithm → Model

→ Testing Phase

(a) Predict on the test data

X_test_transformed → (Model) → y_test_pred

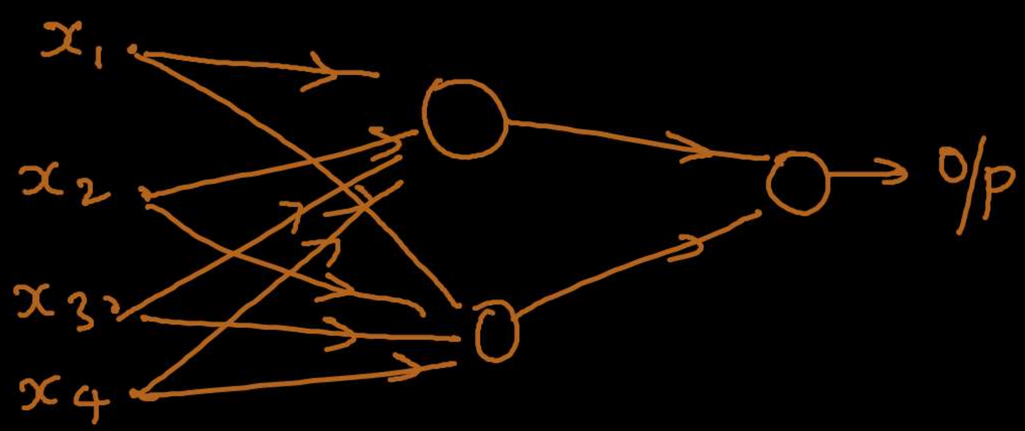
(b) Evaluation

y_train Vs y_train_Pred
y_test Vs y_test_Pred
} → Score < high bias X
high variance X

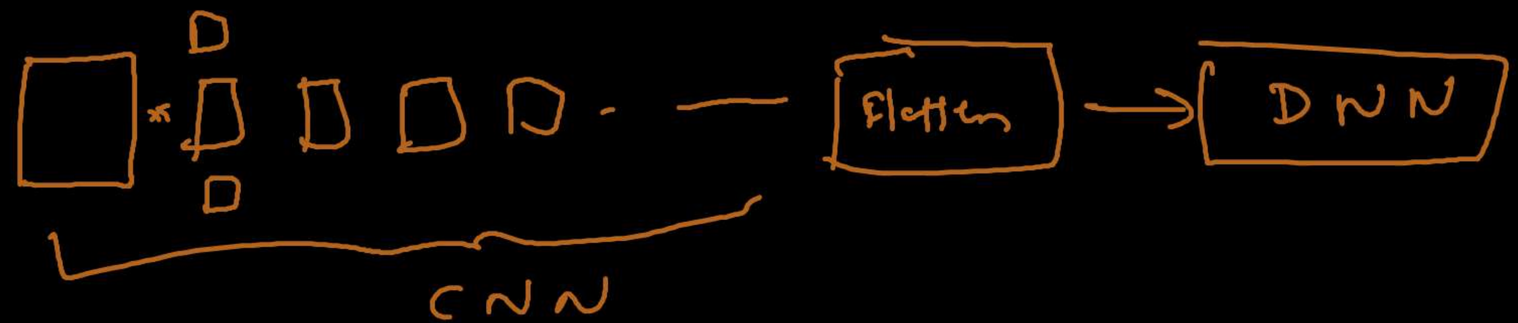
NLP in Deep Learning Prospective

① ANN → MLP / DNN → Tabular Data - Table < Rows, columns - Label Data

House size	Age	No. of rooms	Location	Price ?
x_1				y



(2) CNN & Computer vision → Image classification, Object Detection



(3) Sequential Data (Text Data)

- (i) RNN
- (ii) LSTM RNN
- (iii) GRU RNN
- (iv) Encoder-Decoder
- (v) Seq-2-Seq

Time series forecasting

NLP — Text Analytics / Text Generation

(vi) Transformer (Attention is all you need)

Adv :- GenAI

- → BERT
- → Gemini
- → Llama 3.1
- → LLM models

eg :-
Chat Bot App → Q & A
Language Translation
Text correct / Gener...

Interview Purpose

* What is NLP ?

→ NLP is the branch of AI that involves computers to understand, interpret & generate human language.

NLP
Interpret

NLU
understand

NLG
Generate

→ There are wide range of task in NLP :—

① Text classification :- Spam detection, Sentiment Analysis. Adult content filtering etc.

② Information Extraction → Named entity Recognition (NER), parts of speech tagging

NER eg :- Apple is looking to launch my macbook for $1500.
ORG Product Price

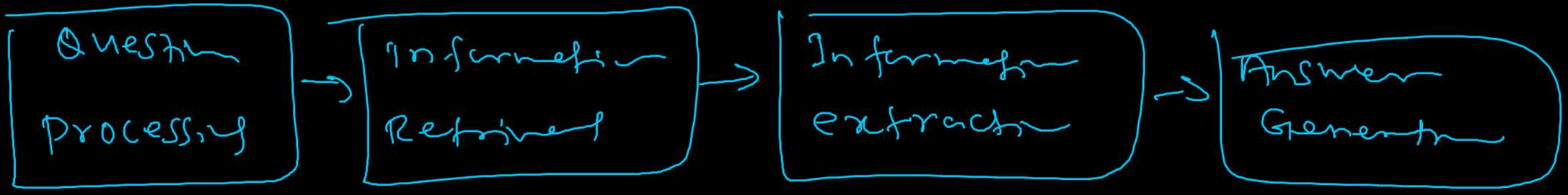
eg :- She runs fast.
POS tagging → Pronoun verb Adverb
Sub Action Describe

③ Information Retrival :- Search & Retrieval
→ Documents, website, https, Websh

④ Text Summarization → News feed creation, case report Summarization in a big law firm

(5) Machine Translation :- Translating one language to another

(6) Question & Answering :-



DNN & CNN

Q:- Why is NLP Hard?

1. Ambiguity :- ie. uncertainty of meaning

eg.- The chicken is ready to eat.

eg.- The car hit the pole while it was moving.

2. Complexity of representation

like -> Poems, sarcasm, Phrases etc.

eg.- You have a football game tomorrow, break a leg!

eg -> Yeah, right, because that worked so well last time.

Also called as

Text Representation (AKA vectorization or feature extraction)

1. Bag of words - BOW → captures word freq/count
2. Term Frequency - Inverse Document Frequency (TF-IDF) → capture the word 'importance'

Text to vect / Char to Num → ML Eg

Phase 1

Basic Language representation

non-numeric repres. → to → Numeric representation

3. Word2vec (word to vector) → 2013 (Google)
4. Glove (Global vector) → 2015 - (Stanford)
5. Fast Text → 2016 - (Facebook)

→ Distributed Language representation - 2009 - 2017

Capture the word Semantic meaning → During DL era

2017

Phase II

Phase 3

6. ELMo → 2017 (University of Washington)
7. Transformer (Attention all you need)
8. BERT → 2018 — Hugging face
9. GPT-4o → 2024 (Open AI)
10. LLAMA-3 (API) 8.1 (July) — (Meta)
11. Gemini - 1.5 - 2024 (Google)
12. Phi - 1.5 - Microsoft (2024)

Dumy Gen AI
Eg
+ ChatBot
+ RAG
+ Fine-tuning
+ Prompt engineering
+ Stable defusion

↳ Contextual Language Representation (LLM) / Gen AI

Terminology

(a) **Document** :- It is a single piece of text. It can be a sentence, paragraph, email, message, article review etc _ _

(b) **corpus** :- collection of documents

(c) **Vocabulary** :- set of all unique words that appear in a corpus.

(d) **Vectorization** → It's a technique to convert text data into numerical vectors

(e) **Document Vectors** → Numerical representation of documents

(f) Document Term matrix :—

— columns tex

Able	ability	Always
1	0.1	0.1
0.1	0.2	0.5
0.2	0.3	0.8
0.3	1	1.3
1.2	2	1.9
2.3	3	2.7

Rows

* Text Cl

① Removing

② converting

③ Removing

④ converting to root term

Able able

Ability ability

Dog dog

Cat Cat