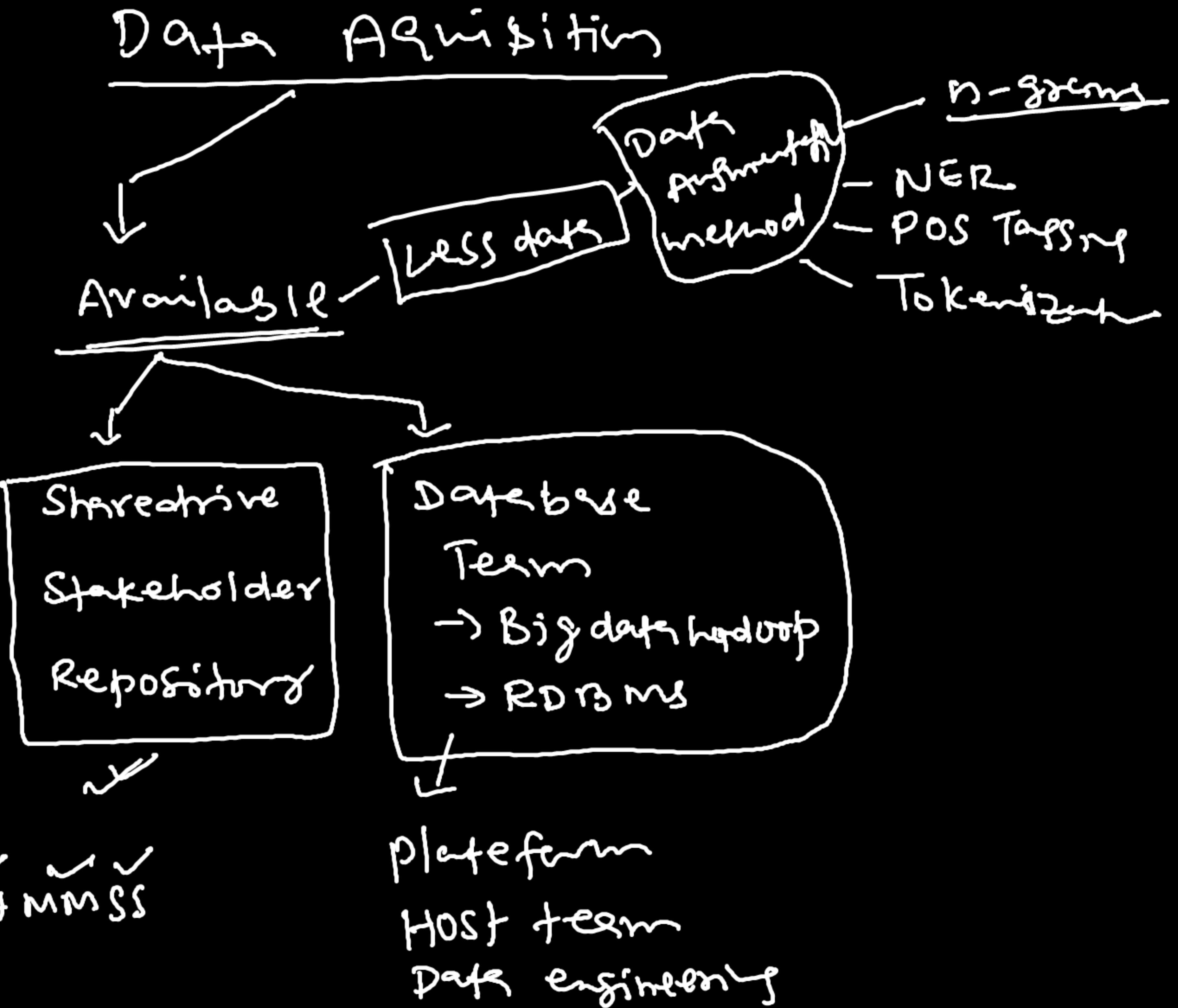


NLP - Pipeline

- ① Data Aquisition
- ② Text preprocessing
- ③ Feature engineering
- ④ Model building
- ⑤ Deployment

YYYYMMDD HHMMSS



Daten Aquisition

② Other Source

→ web scrapping

→ API - Rapid

→ State of the art

↗ PDF

✓ → Audio / Video

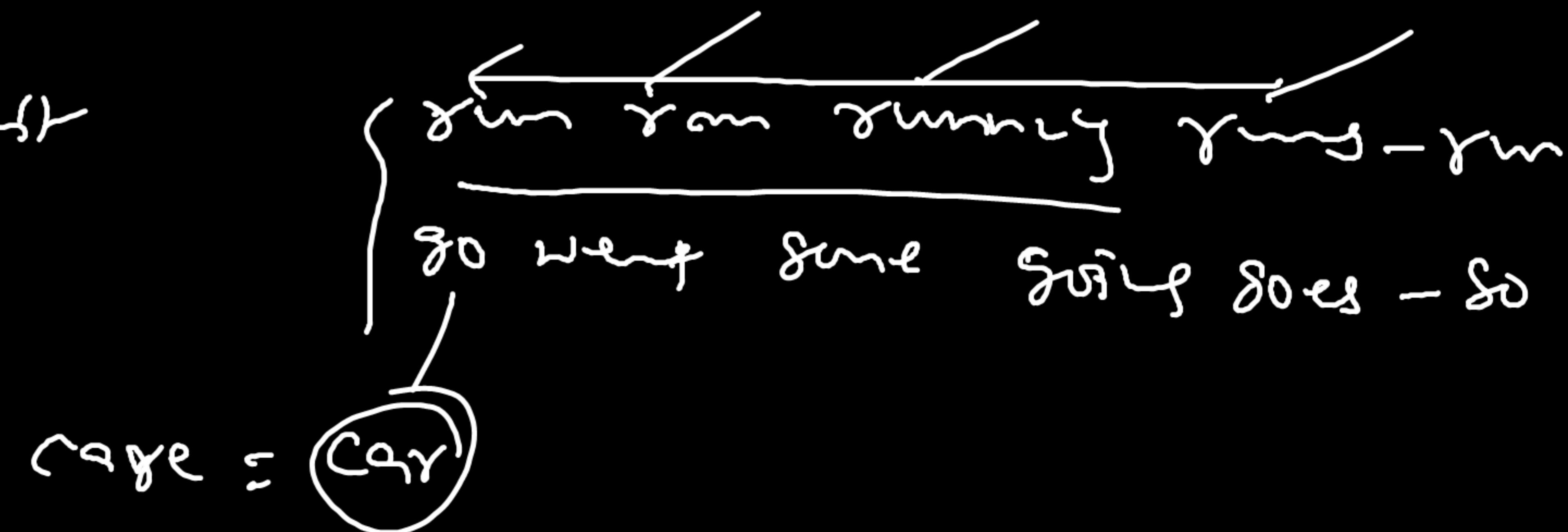
→ Public draft

→ SQL | JSON | SAS | R

③ Not available - Networking

Text Pre-processing

- Cleaning - html tags, https, emojis, Grammar correction, Spelling mistakes
- Tokenization - Word, sent, whitespace, punctuation, stop words
- Stemming / Lemmatization - , lower case, language detection
- Symbols, Number digit
- POS tagging
- NER



Feature engineering

Text → vector

Algorithm

ML

DL (RNN) LSTM GRU/seq2seq

Encoder-Decoder / Transformer

BERT / GPT - LLM etc

ML

Sparse very complex computability

OHE X

→ Bag of words (BOW)

→ N-grams (uni-grams, Bi-grams, Tri-grams, Tetra grams etc.)

→ Term Frequency - Inverse Document Frequency (TF-IDF)

DL

→ Word 2 vec {
 - Dense
 - Glove }
 Semantic meaning

happy → enjoy

Model building



- RF, XGB
- Naive Bayes theorem
- Logistic / DT / SVM / kNN etc

DL

RNN | LSTM | GRU

Transformer (Seq2Seq,

Encoder Deco-

BERT GPT
BERT
GPT
BERT

Deployment → chatbot, API, Flask, Heroku, AWS, Azure,
GCP etc