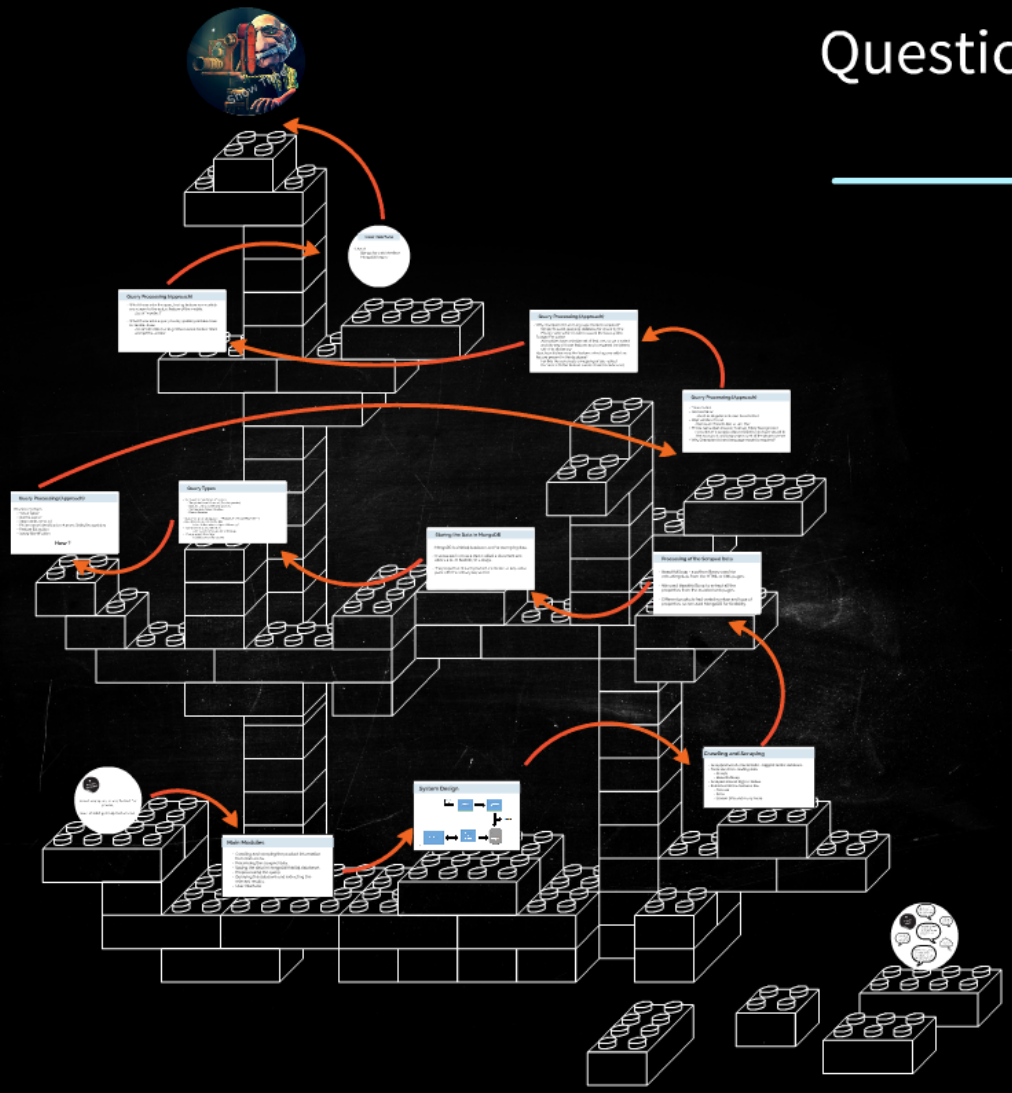
Question and Answering System for Mobiles

Ganesh
Satyam
Sonam
Vivek

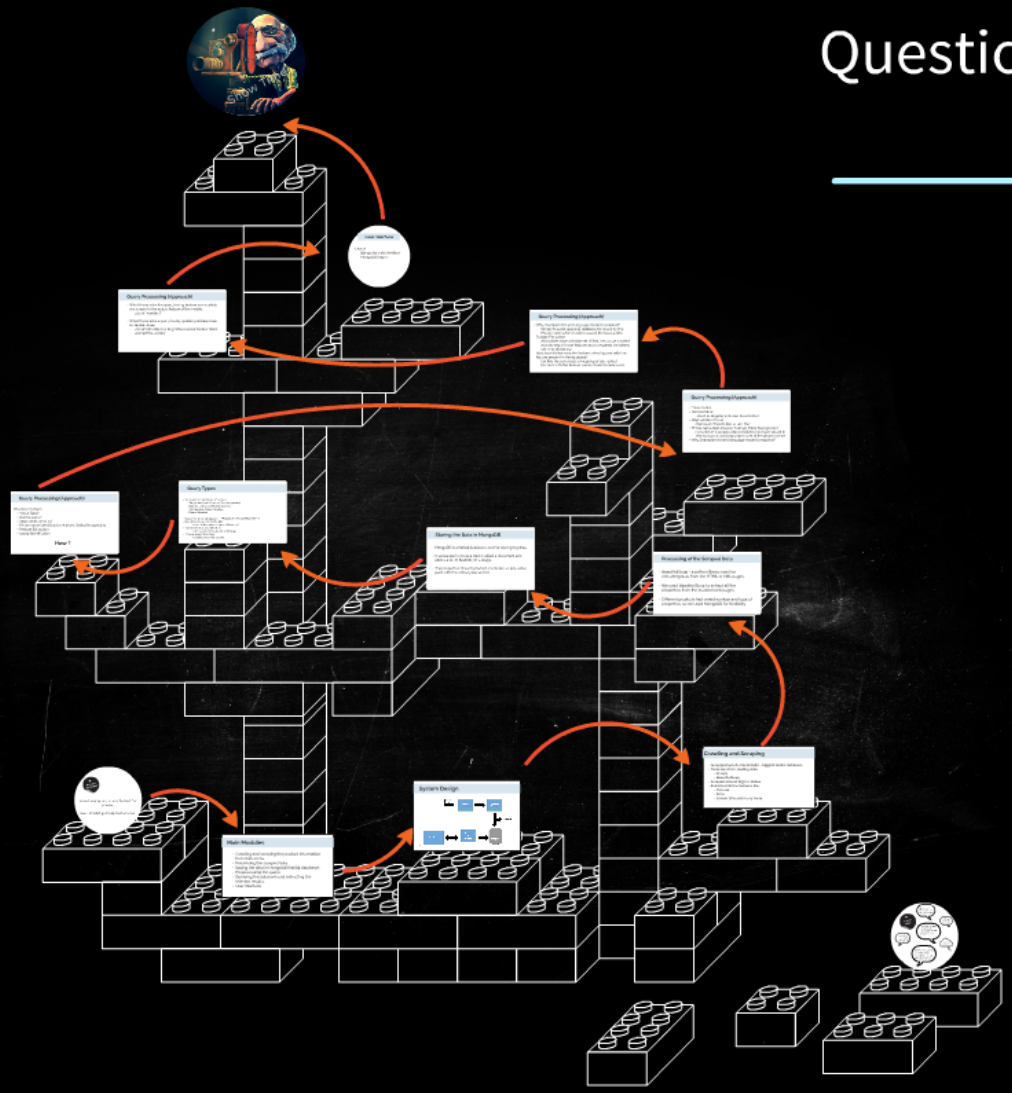


add logo here

Thank You

Question and Answering System for Mobiles

Ganesh
Satyam
Sonam
Vivek



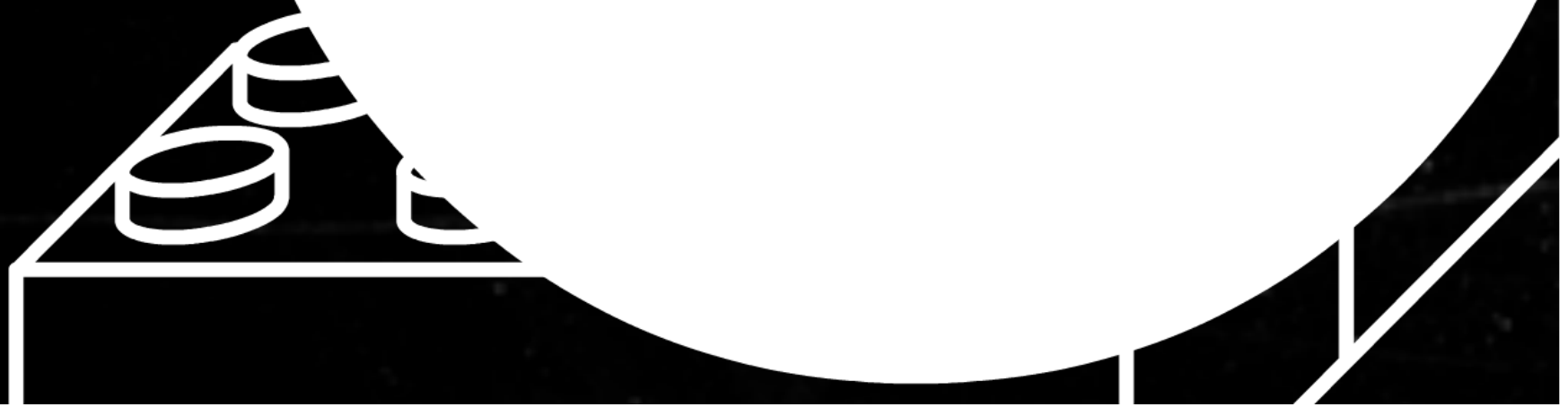
add logo here

Thank You

Problem Statement

Asked any query in any format for phone,

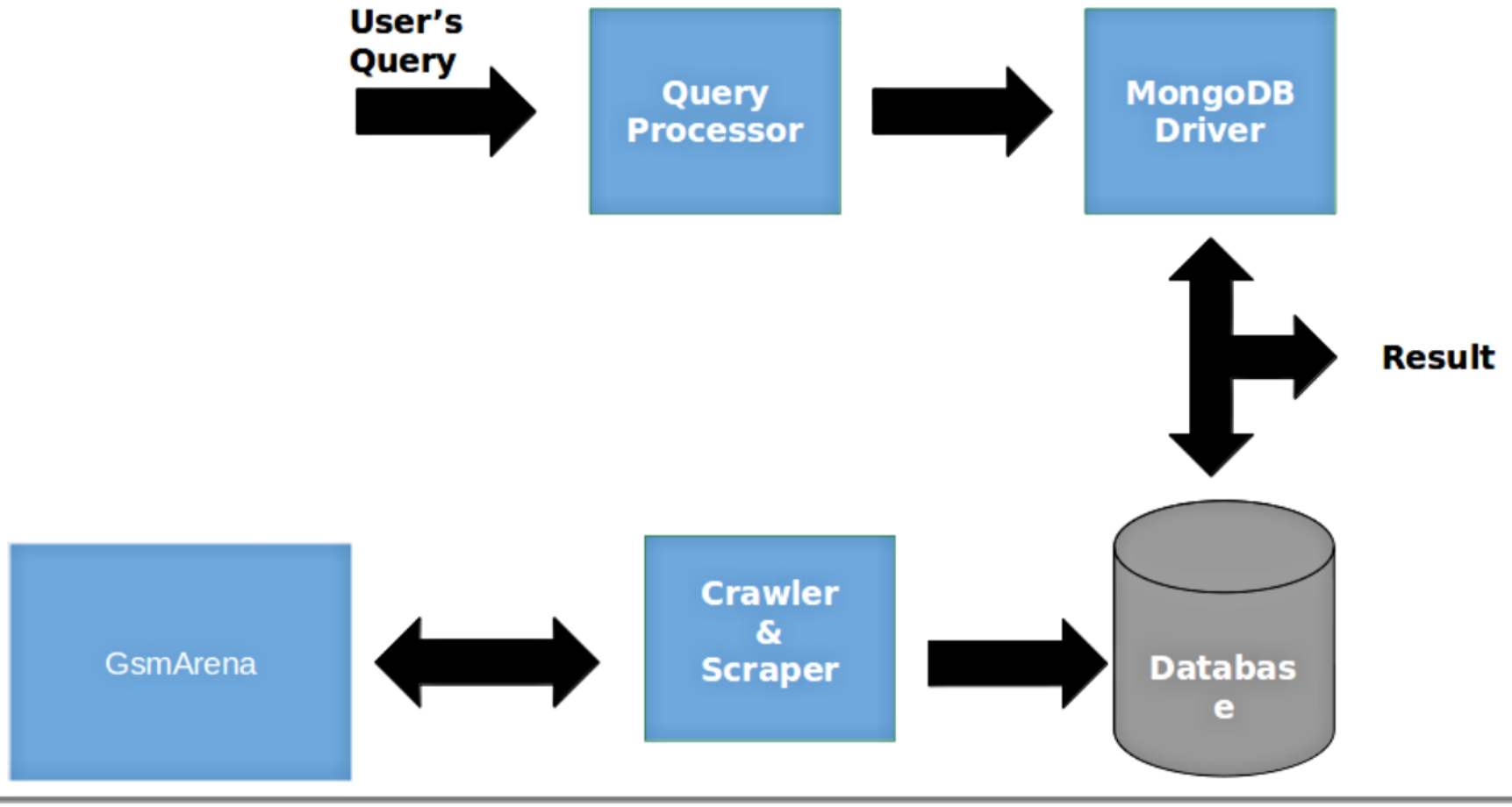
user should get required answer



Main Modules

- Crawling and scraping the product information from GsmArena.
- Processing the scraped data.
- Saving the data in MongoDB (NoSql database).
- Preprocessing the query.
- Querying the database and extracting the relevant results.
- User Interface

System Design



Crawling and Scraping

- Scrapped GsmArena website - biggest online database
- Tools used for crawling data
  - Scrapy
  - BeautifulSoup
- Scraped around 8530 mobiles.
- Extracted all the features like
  - Camera
  - Price
  - Screen Size and many more

Processing of the Scraped Data

- BeautifulSoup - a python library used for extracting data from the HTML or XML pages.
- We used BeautifulSoup to extract all the properties from the crawled web pages.
- Different products had varied number and type of properties, so we used MongoDB for flexibility.

Storing the Data in MongoDB

- MongoDB is a NoSql database used for storing big data.
- It stores each row as a JSON, called a document and allows a lot of flexibility in storage.
- The properties of each product are stored as key-value pairs with the primary key as Uid.

Query Types

- We handled four types of queries
  - Template Based Queries (Simple queries)
  - Natural Language Based Queries.
  - Comparison Based Queries.
  - Range Queries

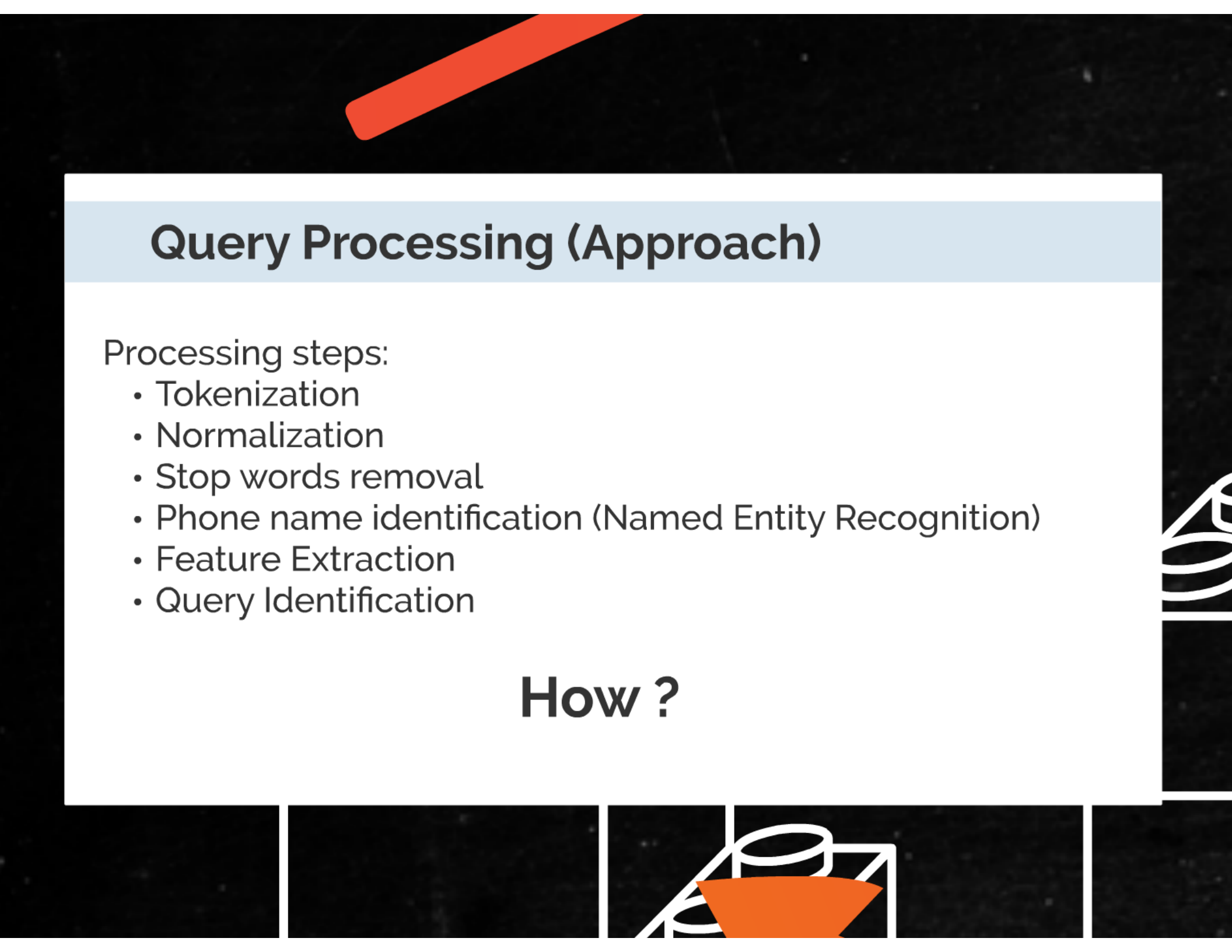
- Syntax of a simple query : [PRODUCT NAME], [PROPERTY]
- NLP based queries can be like :
  What is the price of Apple iPhone 5s?
- Comparison query structure :
  compare iphones 4s and iphone 5s
- Range query structure :
  mobiles under Rs 10,000

Query Processing (Approach)

Processing steps:

- Tokenization
- Normalization
- Stop words removal
- Phone name identification (Named Entity Recognition)
- Feature Extraction
- Query Identification

How ?



Query Processing (Approach)

- Tokenization
- Normalization
  plural to singular and case lowerization
- Stop words removal
  - Removal of words like 'a', 'an', 'the'
- Phone name identification (Named Entity Recognition)
  - Creation of champion lists of UID that we have stored in the mongodb and language model of the phone names
- Why Champion list and language model is required?

Query Processing (Approach)

- Why Champion list and language model is required?
  Simply to avoid querying database for checking the Phone name which in turn reduced the lookup time.

Feature Extraction

  All mobiles have a similar set of features, so we created a dictionary of those features and compared the tokens with this dictionary

Now, how did we map the feature asked by user with the feature present in the database?

  For this, we can create a mapping of user asked features with the feature names stored in database]

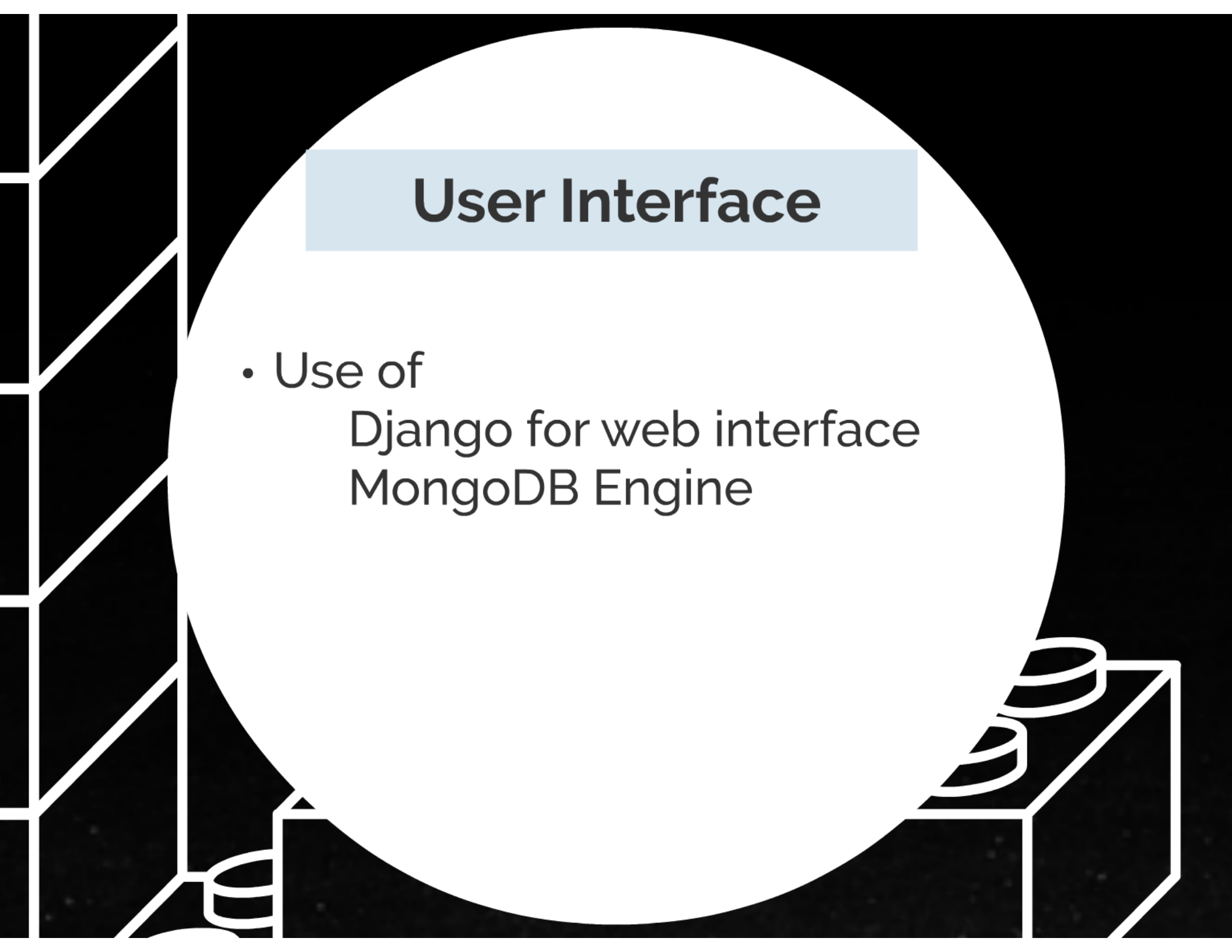
Query Processing (Approach)

- What if user asks the query having feature name which a synonym to the actual feature of the mobile.
  Use of "wordnet"

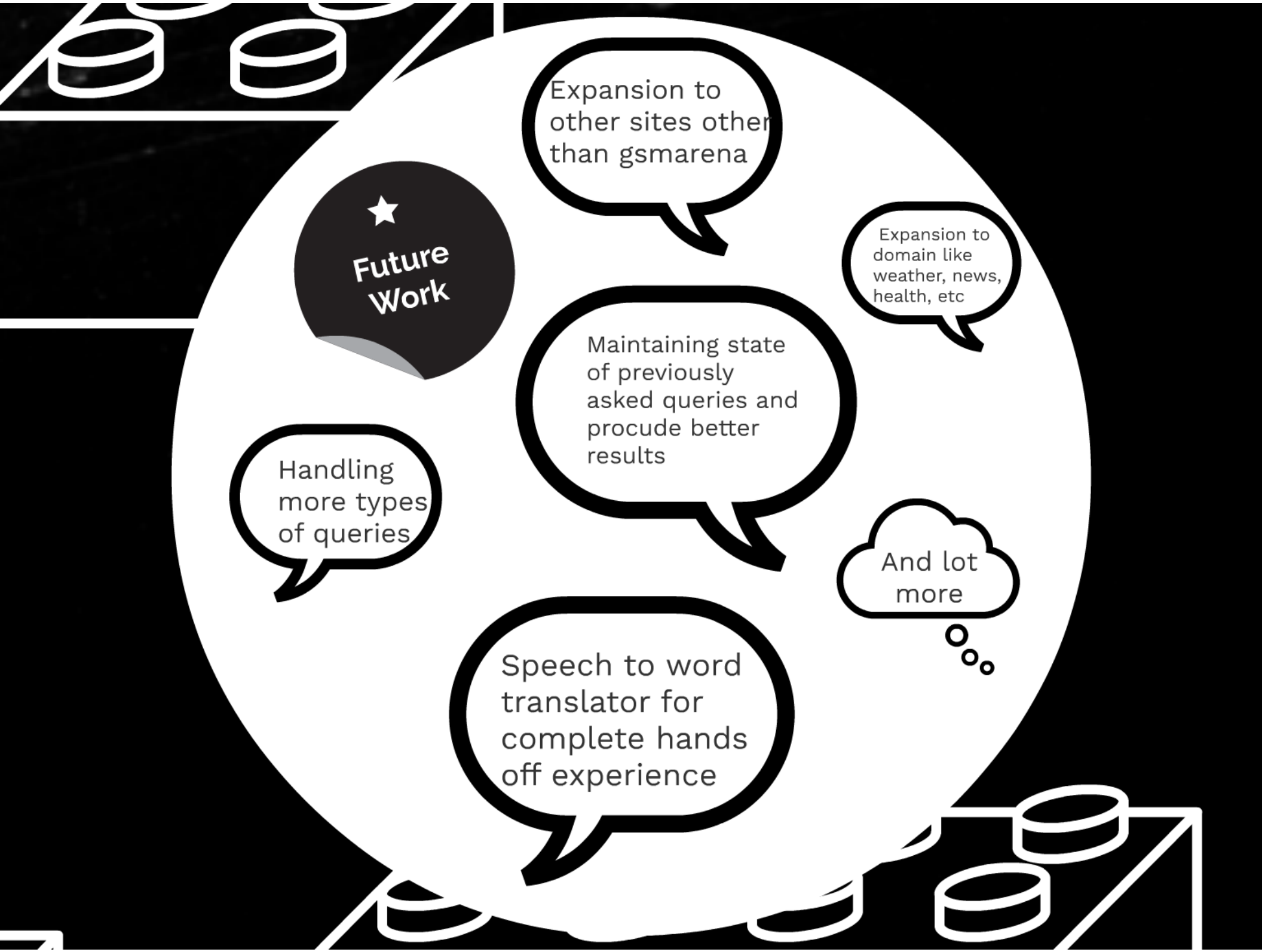
- What if user asks a query having spelling mistake, how to handle those.
  Use of edit distance to get the nearest feature word and get the answer

User Interface

- Use of
  Django for web interface
  MongoDB Engine




Show Time


Future Work
Expansion to other sites other than gsmarena
Expansion to domain like weather, news, health, etc
Maintaining state of previously asked queries and procude better results
Handling more types of queries
And lot more
Speech to word translator for complete hands off experience


Thank You
to be continued...