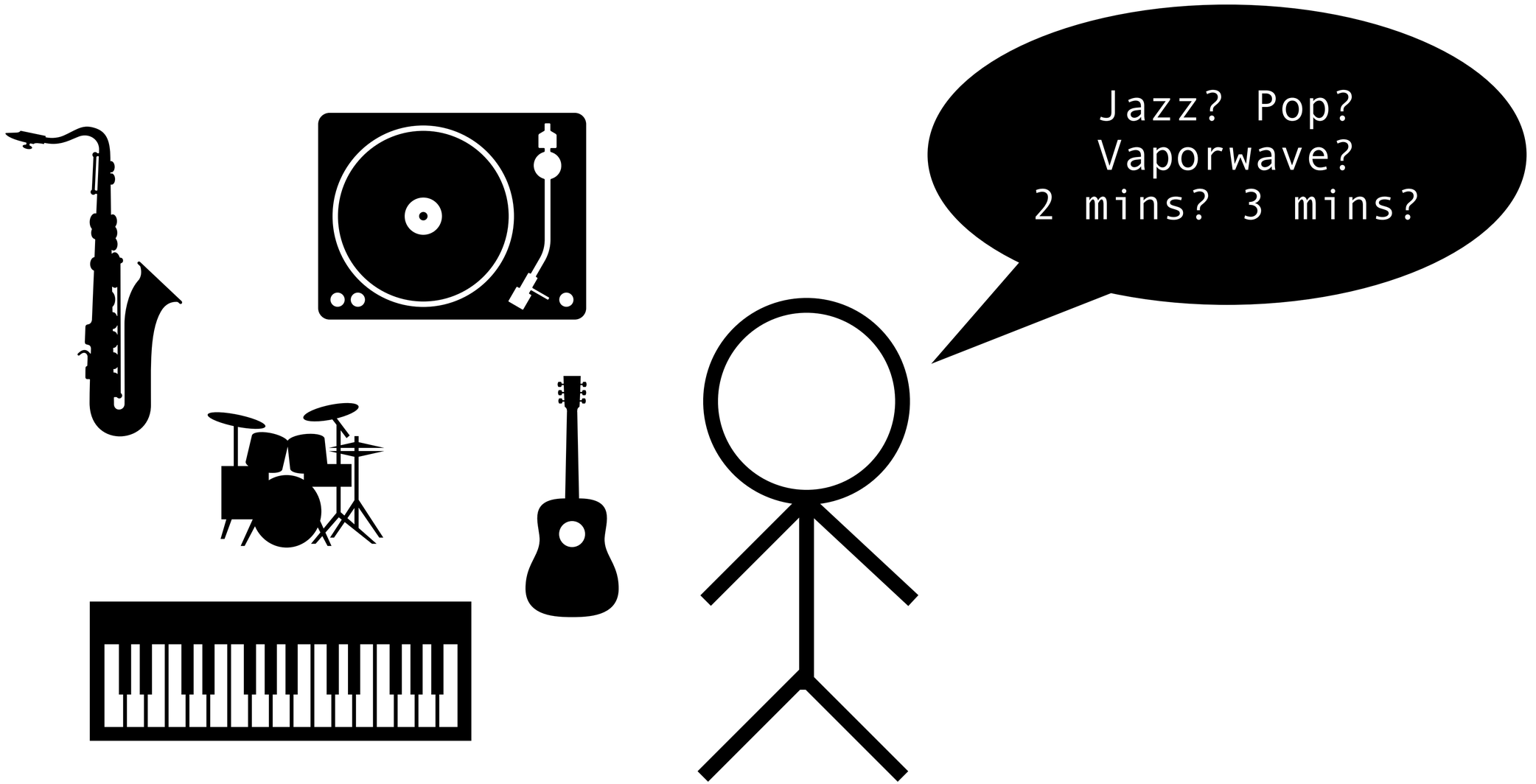
Music Popularity Prediction

CSYE7200 - Final Project

Our team

3 members!

- **Yiqing Huang**（ Jackie / 黄以清)
- **Qinyun Lin**（ Niro / 林沁昀)
- **Zhilue Wang**（ Harry / 王之略)



(Music producer who is trying to write next song)

Title
Artist
Duration
Energy
Loudness
...

Music Features

Music Popularity Prediction

Yes! 🥳

Your song will be popular!

Or

No! 😭

Most people won't like your song

Methodology

Big data

- Whole application is a big-data application
- 2 loops
  - **Training**: Ingest -> Feature extraction -> Machine learning
  - **Inference**: Ingest -> Feature extraction -> Model inference -> Return

Methodology

Machine learning

- Spark built-in ML library
- Algorithms planning to be used:
  - Logistic regression
  - Random forest
  - Bag of Words

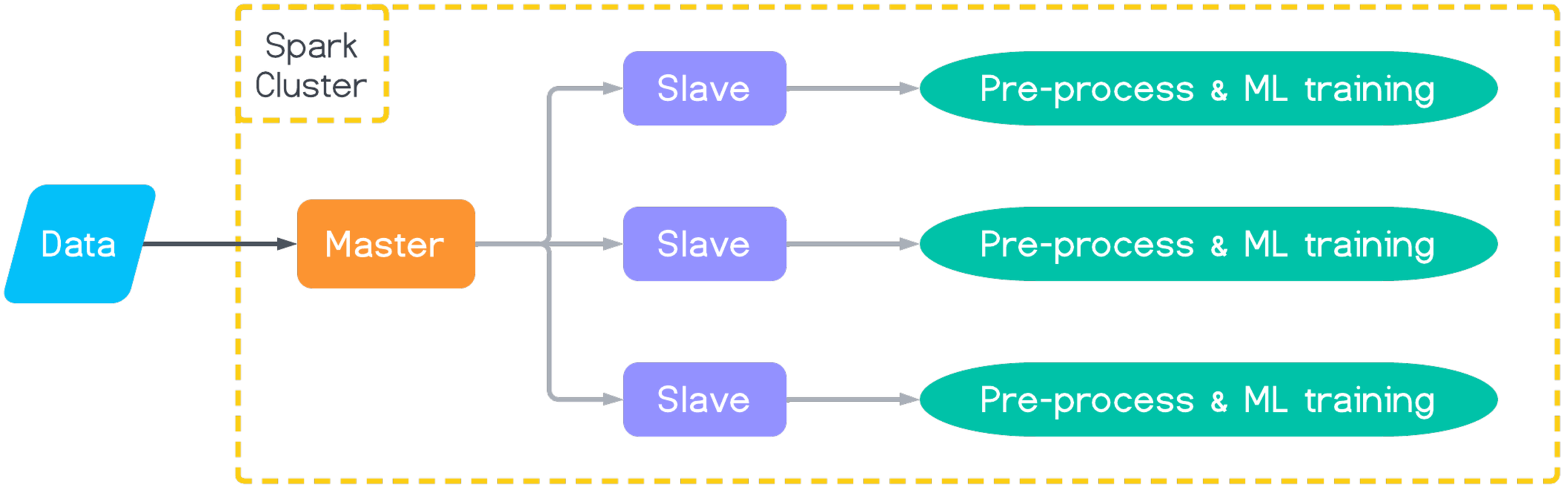
Methodology

Cloud

- Deploy Spark cluster on AWS
- To leverage the power of parallel computing
- Expose an API

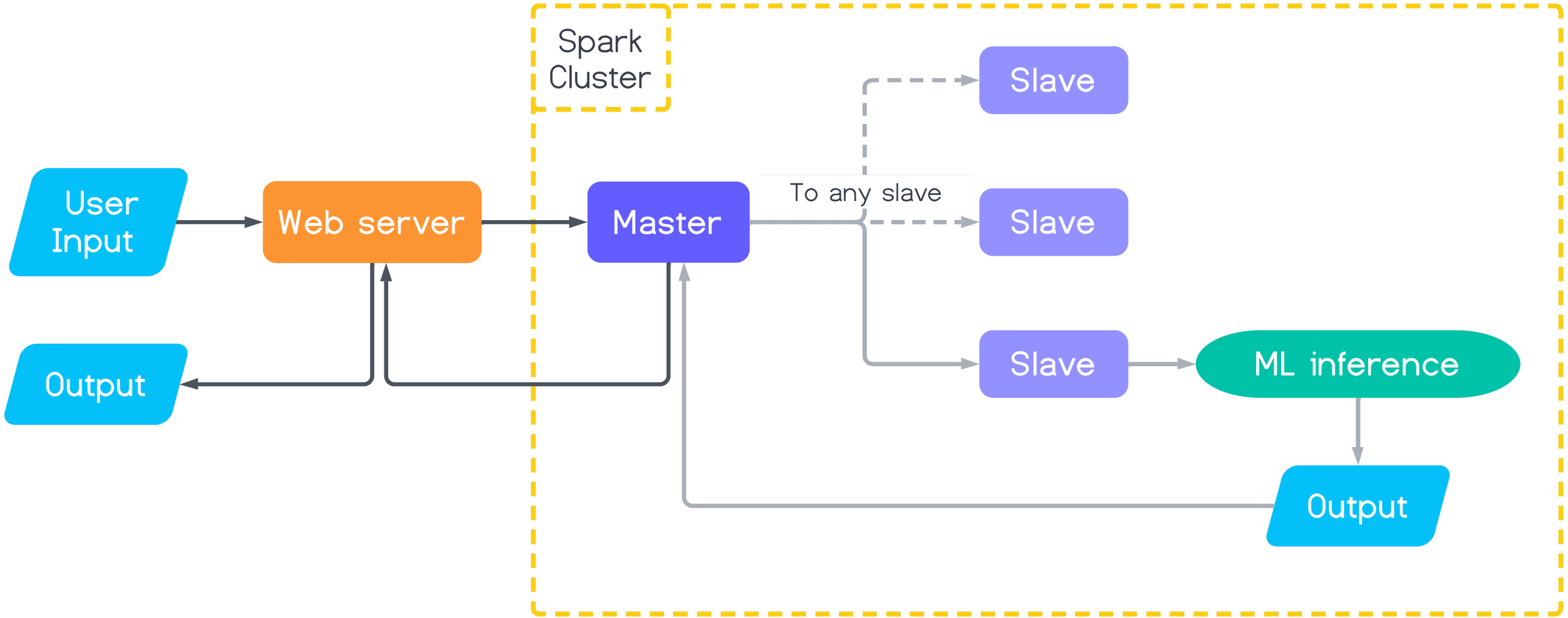
Methodology

Training pipeline



Methodology

Inference pipeline



Date sources

- Million Song Dataset
  - http://millionsongdataset.com/
- ~1,000,000 rows of data
- Has “song_hotttnesss” attribute for popularity

Some attributes

end_of_fade_in
start_of_fade_out
loudness
tempo
title
year
song_hotttness
...

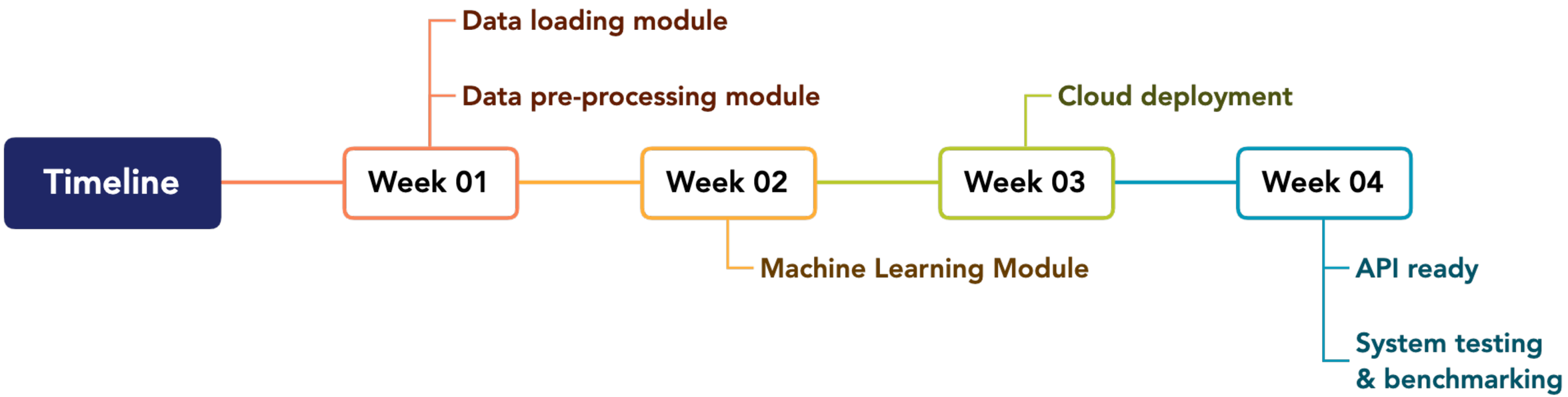
Use cases

- User calls the API with music features, and receives prediction value.
- User calls the API and system validates user input, informs user if there is any error.
- User calls the API and system processes the input data, runs ML model on it and returns the prediction value.

Acceptance criteria

- User queries responding time:
  - Single record: <5s
  - Batch records: <1s per record (on average)(batch has >5 records)
- Model predicting time: <4s
- Precision & recall of the ML model: >60%

Milestones


Timeline
Week 01
Data loading module
Data pre-processing module
Week 02
Machine Learning Module
Week 03
Cloud deployment
Week 04
API ready
System testing
& benchmarking

Scala in our project

- Planning to write all codes in Scala
  - Data loading and pre-processing — Scala & Spark
  - Machine learning — Spark build-in ML library
  - API web service — Scala Play! Framework

- Our repo: https://github.com/NiftyMule/csye7200-bigdata-project

Our goal

- Learn how to:
  - Process & load data in Spark and Scala
  - Train a machine learning model and use it for prediction
  - Deploy Spark cluster to cloud environment
  - Implement a simple web server in Scala
- Apply these knowledges to build a real-world application!

Q & A