PERSONAL PROJECT

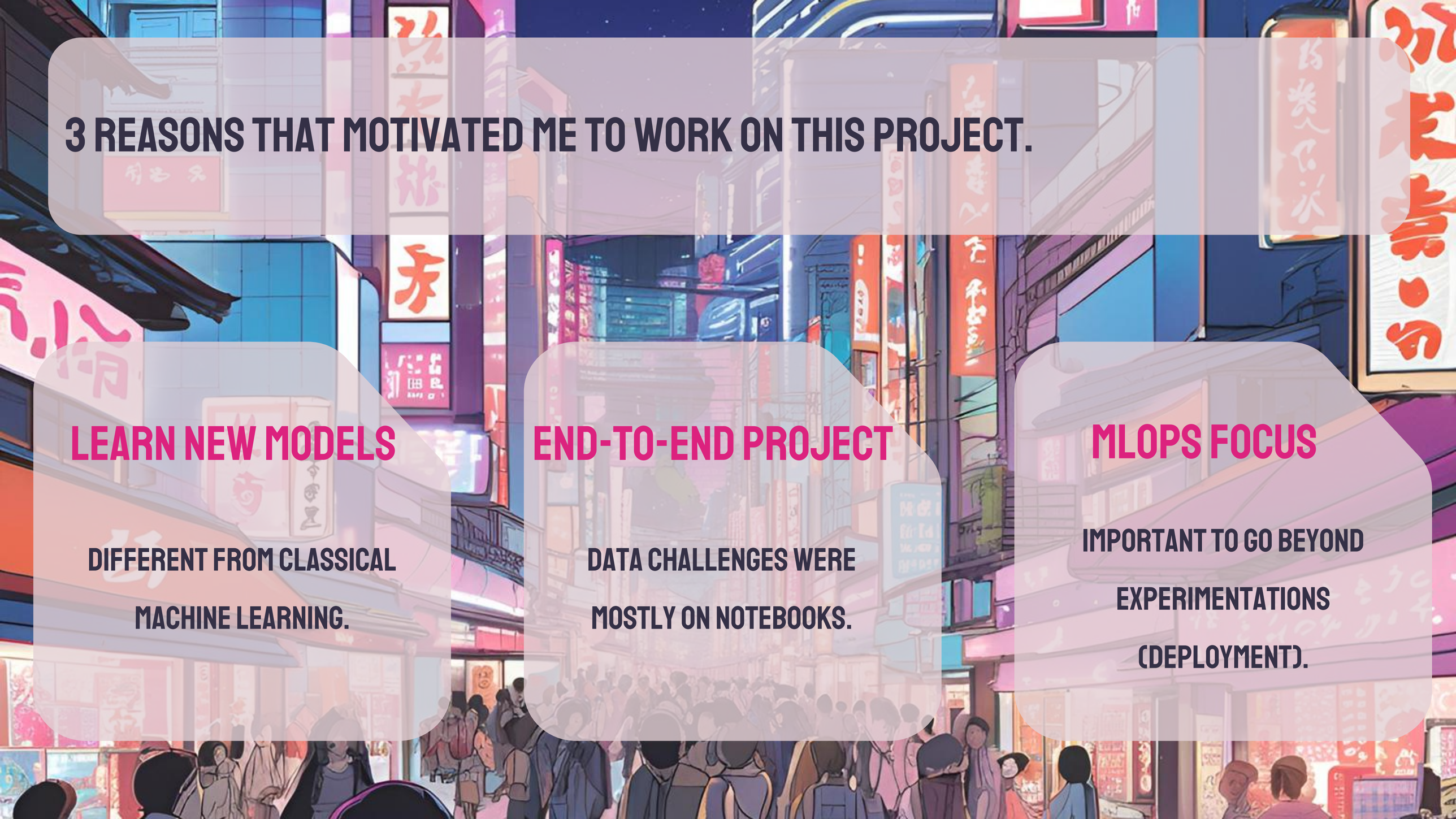
MANGA RECOMMENDER SYSTEM

TRUNG DAN PHAN

ÉCOLE POLYTECHNIQUE
UNIVERSITÉ PARIS-SACLAY

HEC PARIS

TABLE OF CONTENTS



3 REASONS THAT MOTIVATED ME TO WORK ON THIS PROJECT.

LEARN NEW MODELS

DIFFERENT FROM CLASSICAL MACHINE LEARNING.

END-TO-END PROJECT

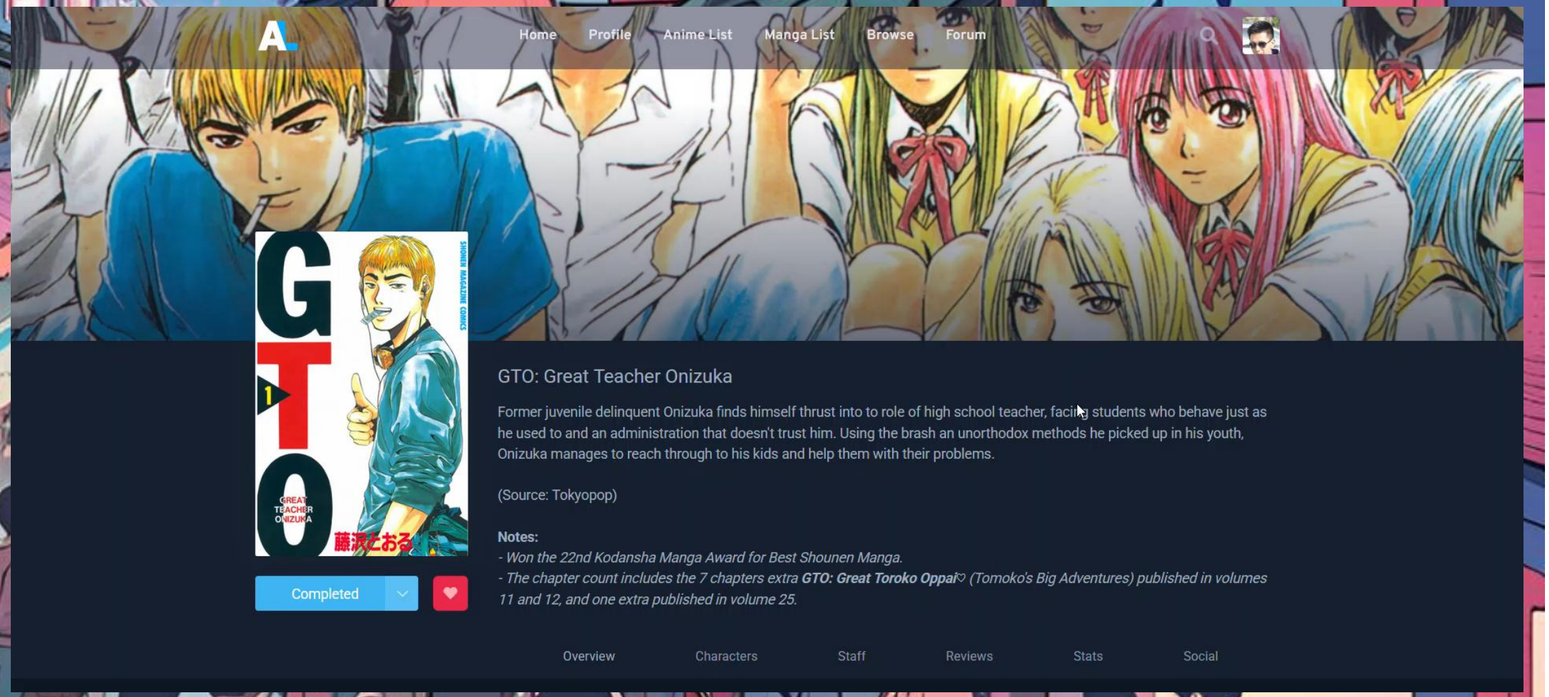
DATA CHALLENGES WERE MOSTLY ON NOTEBOOKS.

MLOPS FOCUS

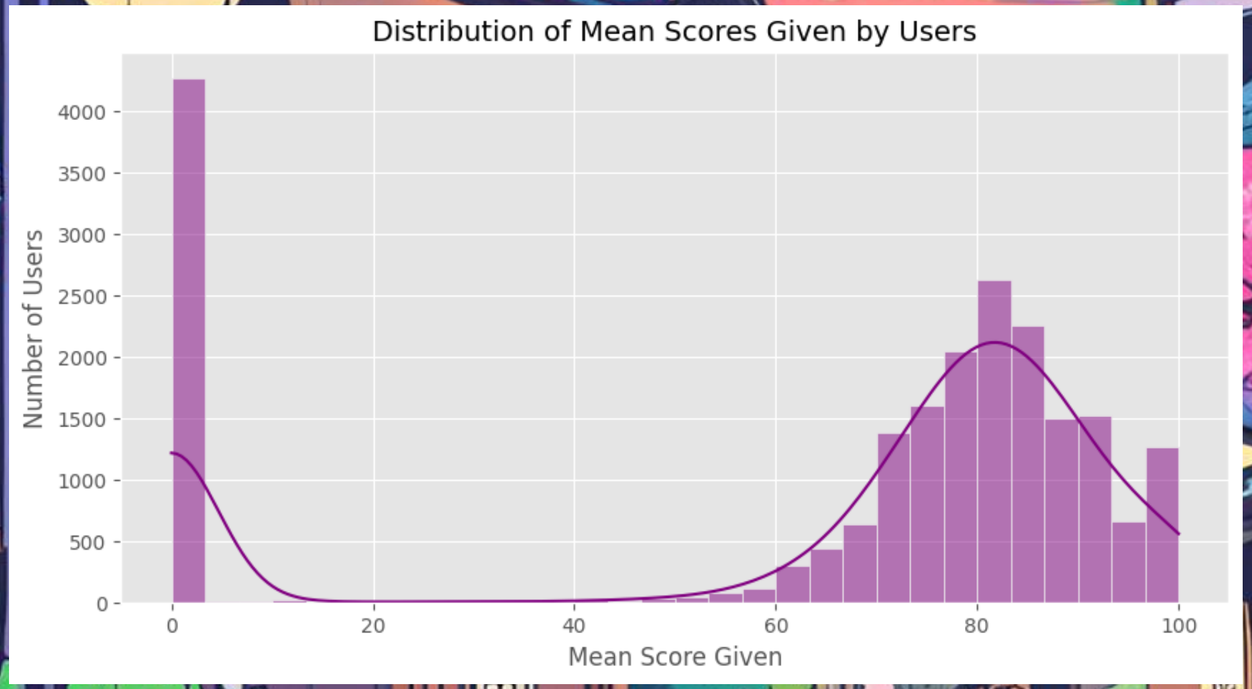
IMPORTANT TO GO BEYOND EXPERIMENTATIONS (DEPLOYMENT).

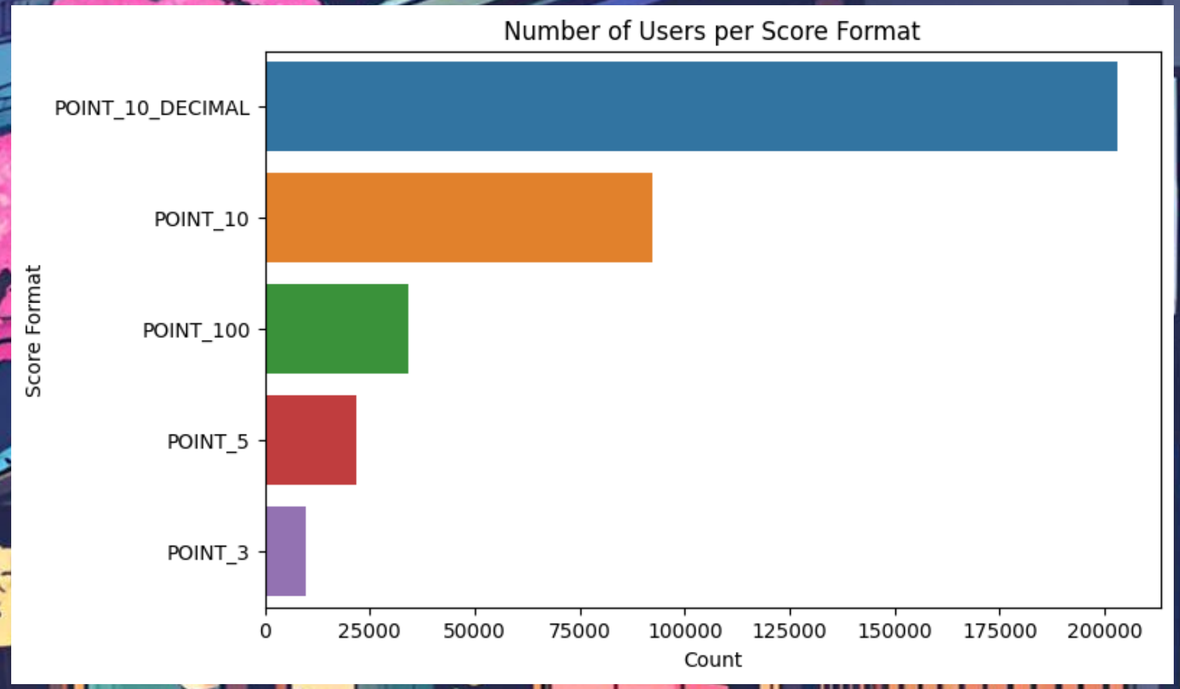
ANILIST

- Discover, Share, Track manga progress
- 3 Million Users & 100k Mangas available
- Fetched **10k Mangas** & Reading History of **8k Users** using API requests



ALL MANGA USERS DON'T ALWAYS RATE THE SAME WAY.





HOW TO EVALUATE A RECOMMENDER SYSTEM?

MEAN ABSOLUTE ERROR

Measures how far (on average) the predicted ratings are from the actual user ratings.

PRECISION@K SCORE

In the recommender's top 10 recommendations, how many were relevant* for the user

RECALL@K SCORE

Out of all the relevant* manga, how many were in the top 10 recommendations for the user

*A MANGA TITLE WITH A RATING ABOVE 7 IS CONSIDERED RELEVANT

EVALUATION METRICS

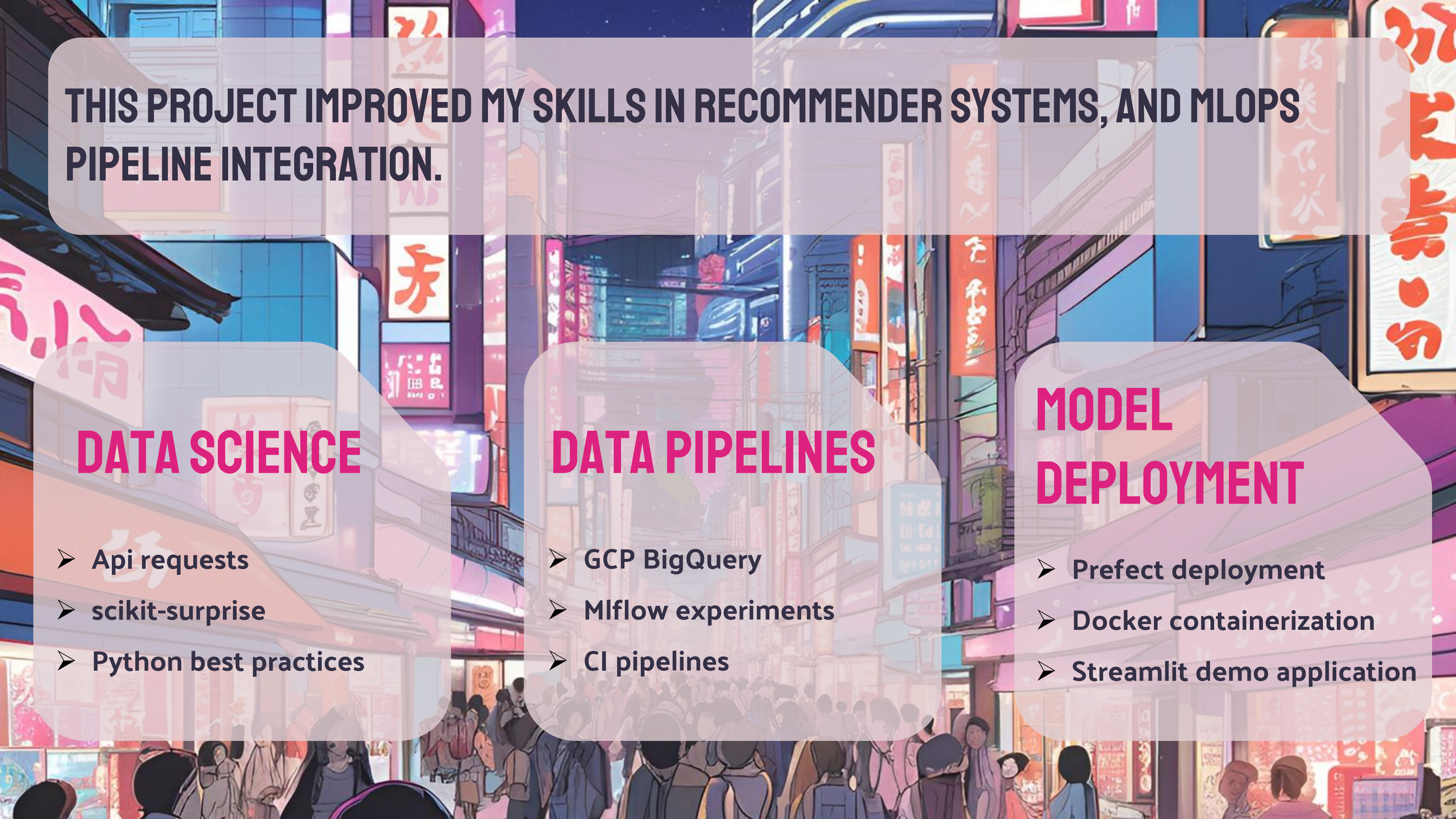
Algorithm Name	FCP	MAE	RMSE	Precision@10	Recall@10
Baseline Only	0.747	1.008	1.356	0.946	0.501
KNN Basic	0.745	1.106	1.522	0.955	0.498
KNN Baseline	0.783	0.885	1.219	0.958	0.506
SVD	0.881	**0.637**	**0.895**	**0.981**	**0.513**
KNN with Z-Score	0.771	0.928	1.272	0.949	0.497
KNN with Means	0.773	0.952	1.293	0.942	0.492
Co-clustering	0.720	1.015	1.378	0.943	0.497
Slope One	**0.886**	0.709	1.048	0.963	0.497
Non-negative Matrix Factorization	0.801	1.608	1.851	0.869	0.317
Random Normal Predictor	0.495	2.125	2.680	0.762	0.333

Table 1: Comparison of different recommendation algorithms based on FCP, MAE, RMSE, Precision@10, and Recall@10 scores on the test set. Bold values indicate the best results.

*A MANGA TITLE WITH A RATING ABOVE 7 IS CONSIDERED RELEVANT

APP DEMO

Github repository



THIS PROJECT IMPROVED MY SKILLS IN RECOMMENDER SYSTEMS, AND MLOPS PIPELINE INTEGRATION.

DATA SCIENCE

- Api requests
- scikit-surprise
- Python best practices

DATA PIPELINES

- GCP BigQuery
- Mlflow experiments
- CI pipelines

MODEL DEPLOYMENT

- Prefect deployment
- Docker containerization
- Streamlit demo application

SOME IMPROVEMENTS TO CONSIDER IF MORE TIME WAS GIVEN.

Implement Content-based Filtering for
Hybrid Recommender System

Improve application using more advanced Front-End tools
(FastAPI & React JS)

Automate the fetching of new data periodically for
larger and up-to-date database

APPENDIX – SVD ALGORITHM