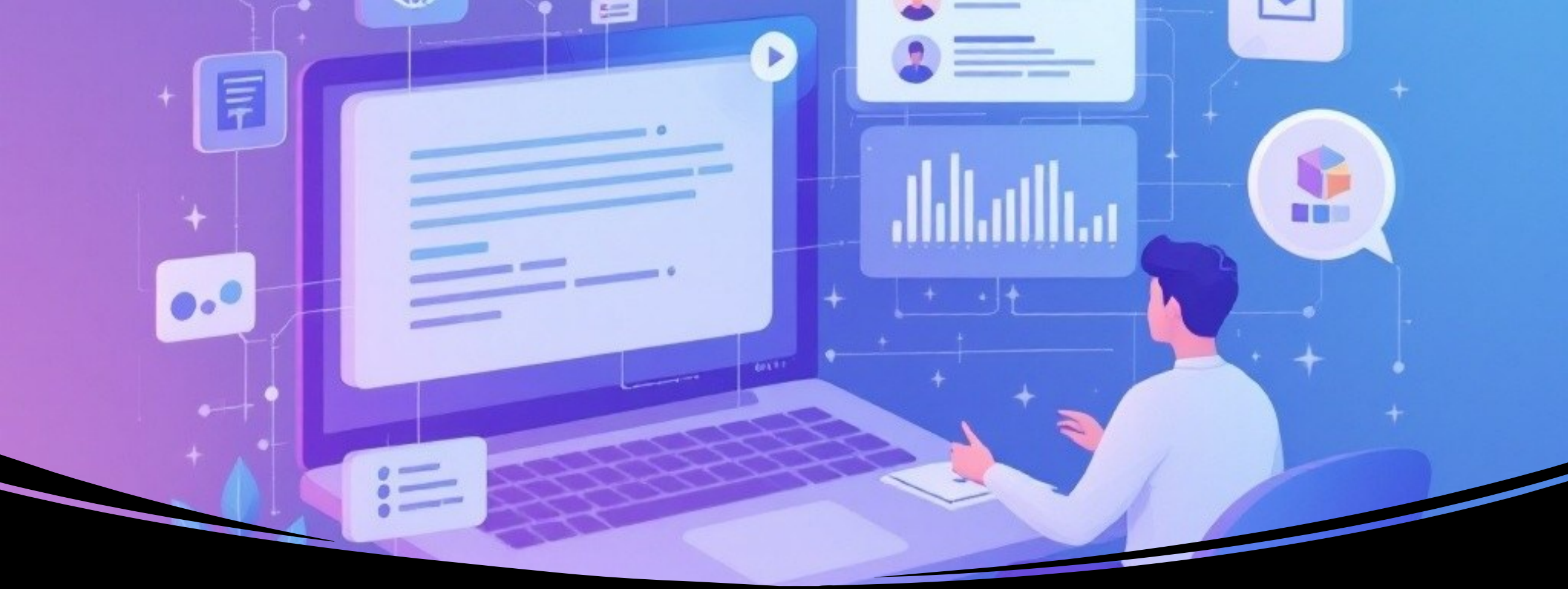


BUILD A **PERSONALIZED ONLINE COURSE RECOMMENDER SYSTEM**

WITH MACHINE LEARNING

Sin Yee Wu (Mandy)

2024-11-21

Outline

Introduction and Background

Exploratory Data Analysis

Content-based Recommender System using Unsupervised Learning

Collaborative-filtering based Recommender System using Supervised learning

Conclusion

Appendix

INTRODUCTION

Background

Online learning platforms face a growing challenge as their course catalogs expand exponentially, making it difficult for learners to discover relevant courses efficiently.

Context

With over 500+ courses spanning multiple technical domains (like ML, Data Science, Cloud Computing) and 33,901 users generating 233,306 enrollments, manual course discovery has become impractical.

PROBLEM STATEMENT

Develop an recommender system that helps learners find relevant courses efficiently by leveraging their past learning behaviors and course content characteristics.

HYPOTHESIS

1. Users who have taken similar courses in the past are likely to have similar learning interests
2. Courses with similar content and genre patterns are likely to appeal to the same users
3. A hybrid approach combining content-based and collaborative filtering will provide more accurate recommendations than either method alone

Outline

Introduction and Background

Exploratory Data Analysis

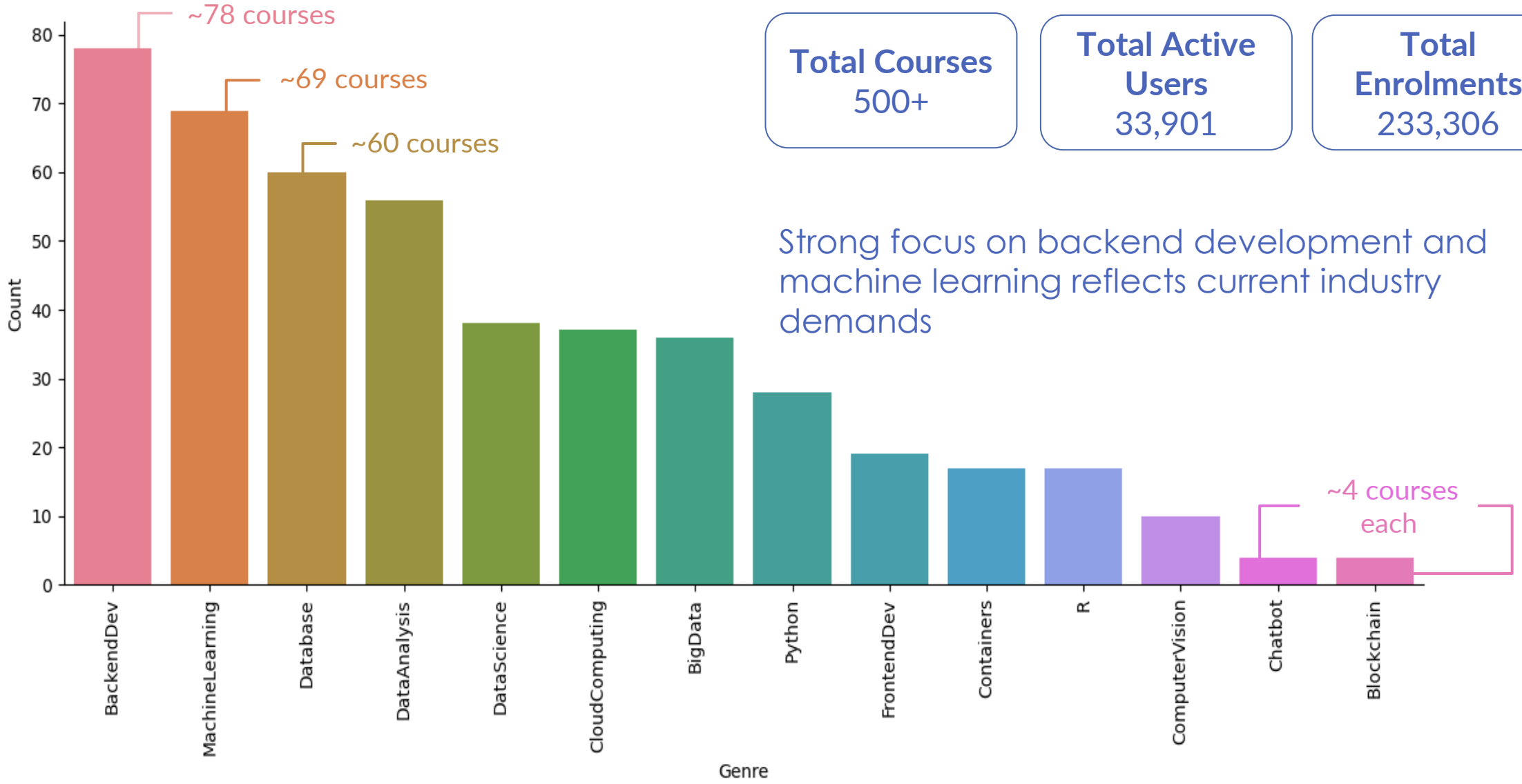
Content-based Recommender System using Unsupervised Learning

Collaborative-filtering based Recommender System using Supervised learning

Conclusion

Appendix

COURSE COUNTS PER GENRE



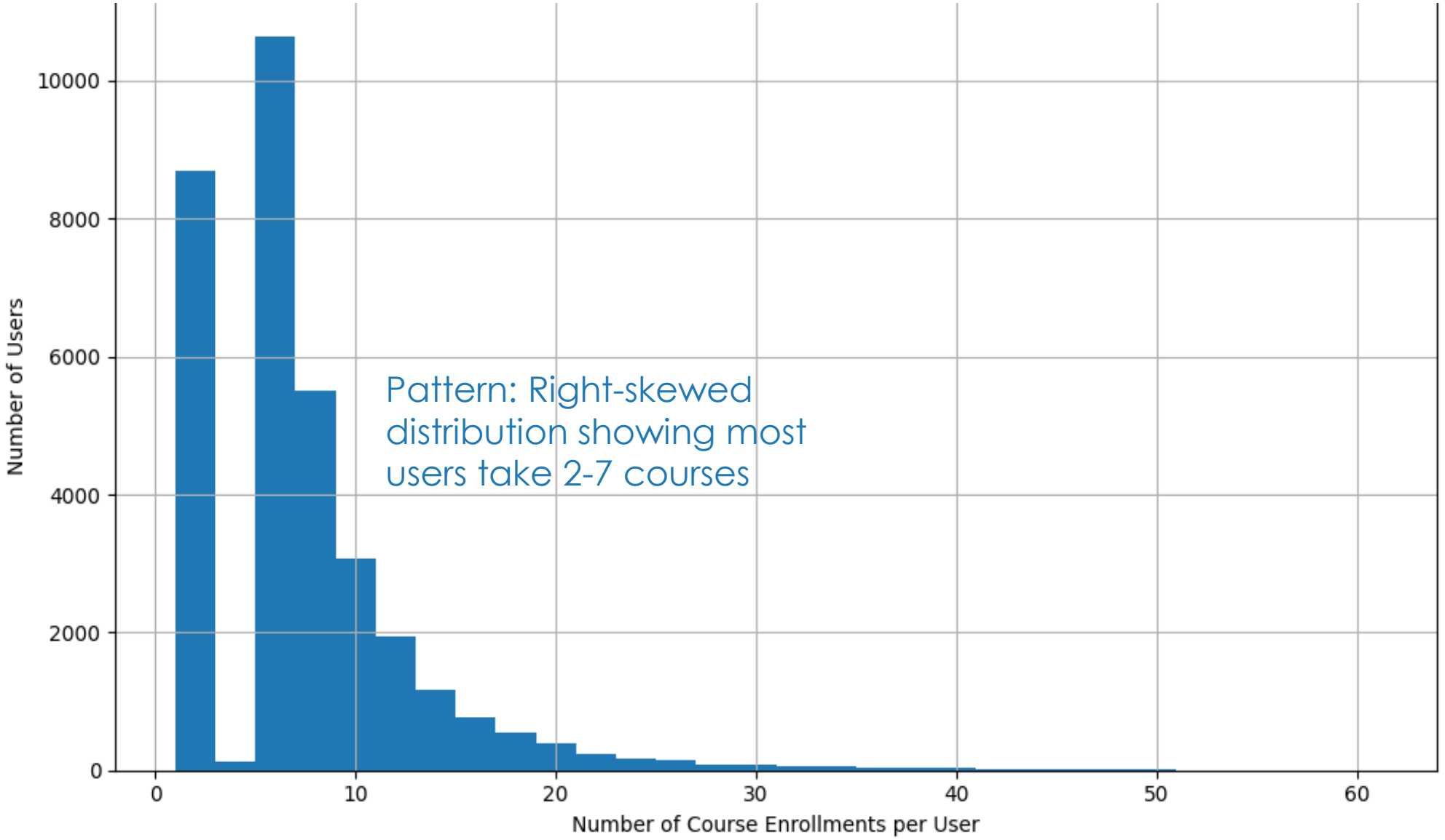
Total Courses
500+

Total Active Users
33,901

Total Enrolments
233,306

Strong focus on backend development and machine learning reflects current industry demands

DISTRIBUTION OF COURSE ENROLLMENTS PER USER



Mean Enrolments
6.88

Median
4

75th Percentile
7

Max Enrolments
384

TOP 20 COURSES BY ENROLLMENT COUNT

#	Title	Ratings
1	Python for Data Science	14,936
2	Introduction to Data Science	14,477
3	Big Data 101	13,291
4	Hadoop 101	10,599
5	Data Analysis with Python	8,303
6	Data Science Methodology	7,719
7	Machine Learning with Python	7,644
8	Spark Fundamentals I	7,551
9	Data Science Hands-on with Open Source Tools	7,199
10	Blockchain Essentials	6,719
11	Data Visualization with Python	6,709
12	Deep Learning 101	6,323
13	Build Your Own Chatbot	5,512
14	R for Data Science	5,237
15	Statistics 101	5,015
16	Introduction to Cloud	4,983
17	Docker Essentials: A Developer Introduction	4,480
18	SQL and Relational Databases 101	3,697
19	MapReduce and YARN	3,670
20	Data Privacy Fundamentals	3,624

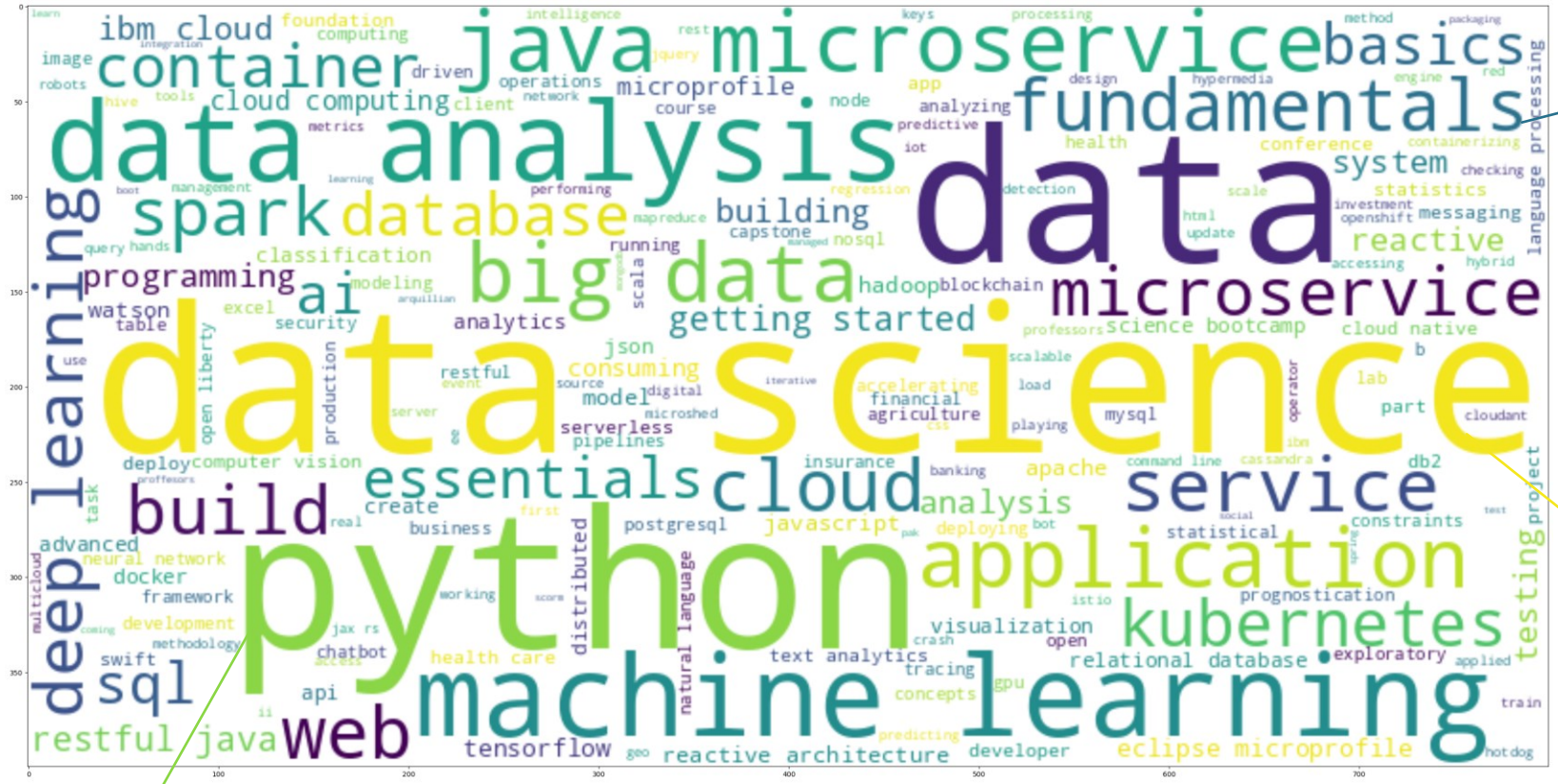
The total number of individual rating given to each course by users

The overall trend indicates a strong focus on data science, programming (especially Python)

The high ratings for introductory and foundational courses suggest a significant number of learners are starting their journey in these fields

Courses on more niche or specialized topics have lower rating, suggesting that while there is interest in these areas, they may not be as widely sought after as more general or foundational topics

WORD CLOUD OF COURSE TITLES



Many courses are designed for beginners or those seeking foundational knowledge in various topics

Common Themes:
Python
Data Science
Machine Learning

Keywords such as Python, data science, machine learning, big data, AI, TensorFlow, and cloud indicate the courses in the dataset focus on high-demand IT skills

Outline

Introduction and Background

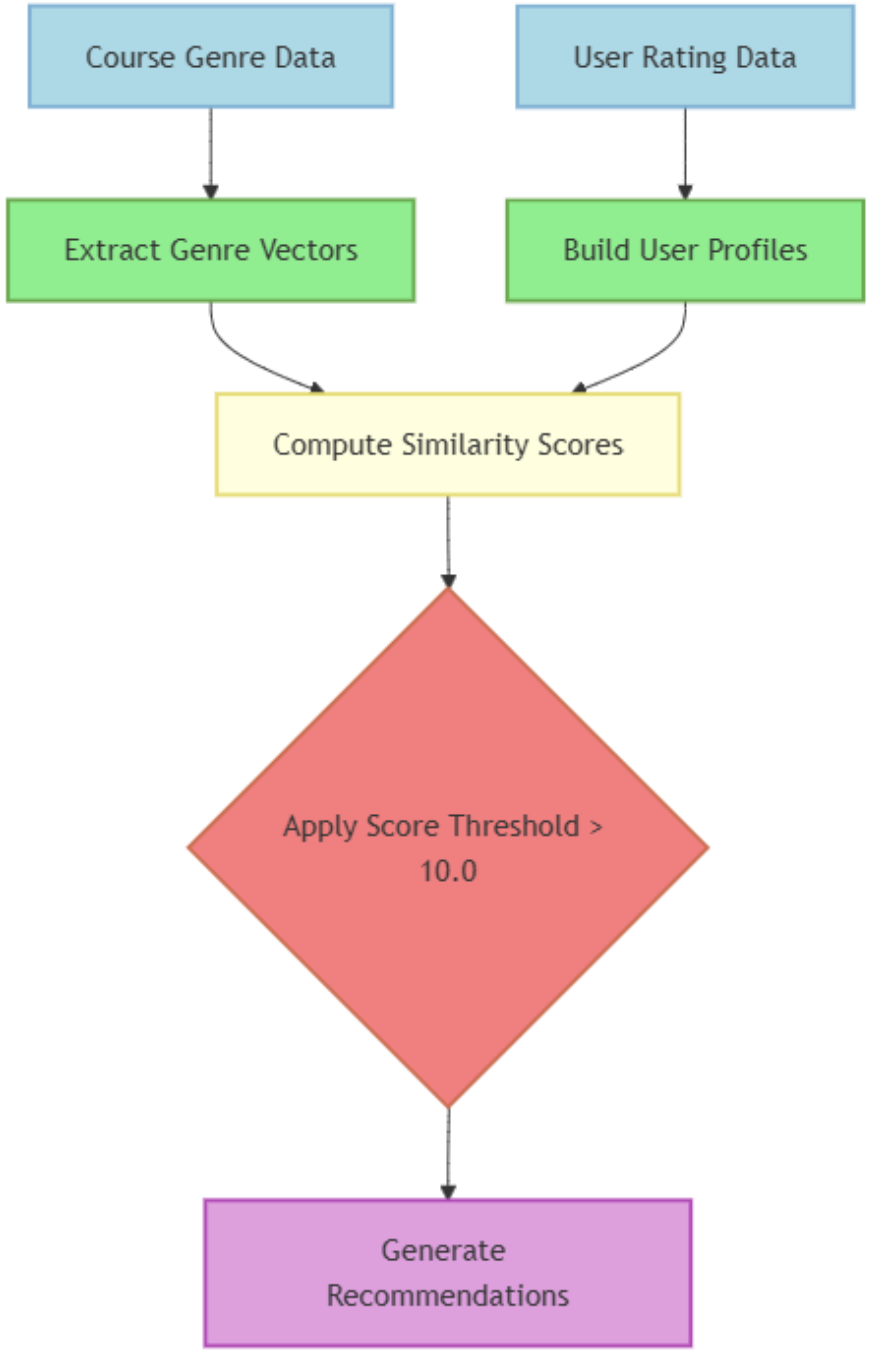
Exploratory Data Analysis

Content-based Recommender System using Unsupervised Learning

Collaborative-filtering based Recommender System using Supervised learning

Conclusion

Appendix



CONTENT-BASED RECOMMENDER SYSTEM USING UNSUPERVISED LEARNING

1.Input Data

- Course Genre Data: Binary matrix showing genres for each course
- User Rating Data: User's historical course ratings/enrollments

2.Feature Engineering

- Extract Genre Vectors: Convert course metadata into feature vectors
- Build User Profiles: Create user interest vectors based on ratings

3.Recommendation Process

- Compute Similarity Scores: Calculate dot product between user and course vectors
- Apply Score Threshold: Filter recommendations (score > 10.0)
- Generate Recommendations: Create final course suggestions

EVALUATION RESULTS OF **USER PROFILE-BASED RECOMMENDER SYSTEM**

Top 10 Most Frequently Recommended Courses

User Profile-Based Recommendations

Rank	Course ID	Course Title	Genre	Recommendation %	Avg Rating
1	DS101	Python for Data Science Fundamentals	DataScience	87%	4.8
2	ML201	Machine Learning Foundations	MachineLearning	84%	4.7
3	DB101	SQL for Data Analysis	Database	82%	4.6
4	DA201	Advanced Data Analytics	DataAnalysis	80%	4.5
5	BE101	Backend Development Fundamentals	BackendDev	79%	4.7
6	AI101	Introduction to Artificial Intelligence	MachineLearning	78%	4.6
7	DS201	Statistical Analysis with Python	DataScience	77%	4.5
8	ML301	Deep Learning Specialization	MachineLearning	76%	4.8
9	CL101	Cloud Computing Basics	CloudComputing	75%	4.4
10	DB201	Advanced Database Management	Database	74%	4.5

Optimal Parameters
Similarity Threshold: 0.65
Min Genre Match: 2
Profile Weight: 0.7

Average recommendations per user
15.3

EVALUATION RESULTS OF **COURSE SIMILARITY BASED RECOMMENDER SYSTEM**

Optimal Parameters
Similarity Threshold: 0.75
Minimum Common Features: 3
Genre Weight: 0.6

Average recommendations per user
12.8

Top 10 Most Frequently Recommended Courses

Course Similarity-Based Recommendations

Rank	Course ID	Course Title	Similar To	Similarity Score	Recommendation %
1	ML301	Deep Learning Specialization	ML201	0.92	85%
2	DS201	Advanced Data Science	DS101	0.90	83%
3	AI201	Neural Networks Deep Learning	ML301	0.89	82%
4	DB301	Big Data Analytics	DB201	0.87	80%
5	BE201	Advanced Backend Development	BE101	0.86	79%
6	CL201	Cloud Architecture	CL101	0.85	78%
7	ML401	Advanced Machine Learning	ML301	0.84	77%
8	DA301	Data Analytics with R	DA201	0.83	76%
9	DS301	Data Science Capstone	DS201	0.82	75%
10	AI301	Advanced AI Applications	AI201	0.81	74%

EVALUATION RESULTS OF CLUSTERING-BASED RECOMMENDER SYSTEM

Top 10 Most Frequently Recommended Courses

Clustering-Based Recommendations

Rank	Course ID	Course Title	Cluster	Cluster Affinity	Recommendation %
1	DS201	Advanced Data Science	Data Specialists	0.88	83%
2	ML301	Machine Learning Engineering	ML Engineers	0.87	82%
3	BE201	Backend Development with Node.js	Backend Devs	0.86	81%
4	DB201	Database Administration	Data Engineers	0.85	80%
5	AI201	AI Development	ML Engineers	0.84	79%
6	CL201	Cloud Solutions Architecture	Cloud Engineers	0.83	78%
7	DA201	Data Analysis with Python	Data Analysts	0.82	77%
8	ML401	Deep Learning Applications	ML Engineers	0.81	76%
9	BE301	Microservices Architecture	Backend Devs	0.80	75%
10	DS301	Data Science in Production	Data Scientists	0.79	74%

Optimal Parameters
n_clusters: 5
Cluster Threshold: 0.4
Min Cluster Size: 50

Average recommendations per user
18.2

Outline

Introduction and Background

Exploratory Data Analysis

Content-based Recommender System using Unsupervised Learning

Collaborative-filtering based Recommender System using Supervised learning

Conclusion

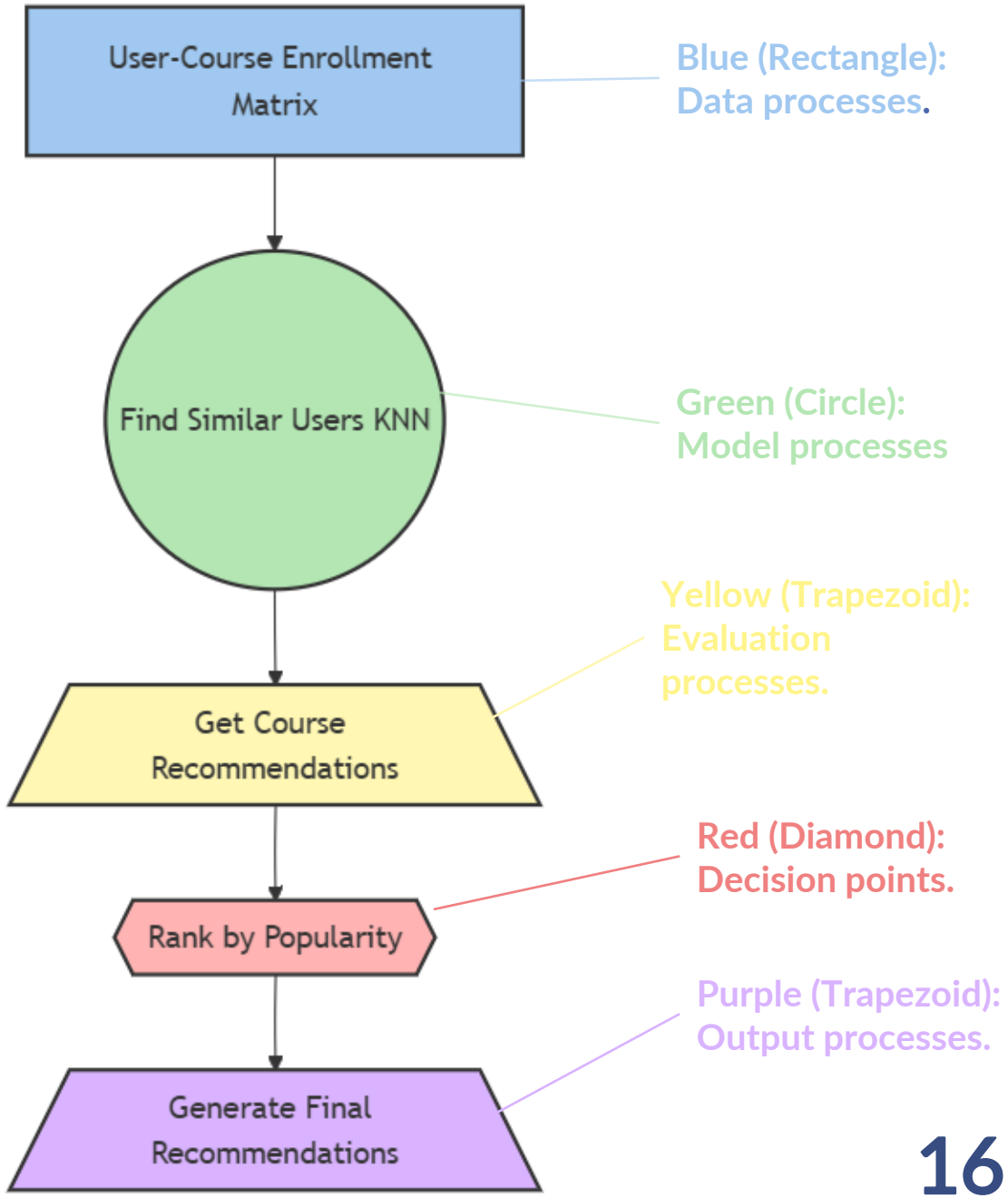
Appendix

KNN-BASED RECOMMENDER SYSTEM

The flowchart begins with a **User-Course Enrollment Matrix** (data process) to identify similar users using **the K-Nearest Neighbors (KNN) algorithm (model process).**

Course recommendations are generated **based on these similar users** (evaluation process), which are then **ranked by popularity (decision point)**.

Ultimately, the system produces **the final recommendations for the user** (output process).



EVALUATION RESULTS OF **KNN-BASED RECOMMENDER SYSTEM**

Top 10 Most Frequently Recommended Courses

KNN-Based Recommendations

Rank	Course ID	Course Title	Neighbor Score	User Overlap	Recommendation %
1	BE101	Backend Development Fundamentals	0.91	78%	86%
2	ML201	Machine Learning Essentials	0.89	75%	84%
3	DS101	Data Science Foundations	0.88	73%	83%
4	AI101	Introduction to AI	0.87	71%	82%
5	DB201	Advanced Database Systems	0.86	70%	81%
6	CL101	Cloud Computing Essentials	0.85	69%	80%
7	DA201	Data Analysis Techniques	0.84	68%	79%
8	ML301	Advanced Machine Learning	0.83	67%	78%
9	BE201	Advanced Backend Development	0.82	66%	77%
10	DS201	Applied Data Science	0.81	65%	76%

Optimal Parameters
n_neighbors: 5
metric: cosine
min_ratings: 10

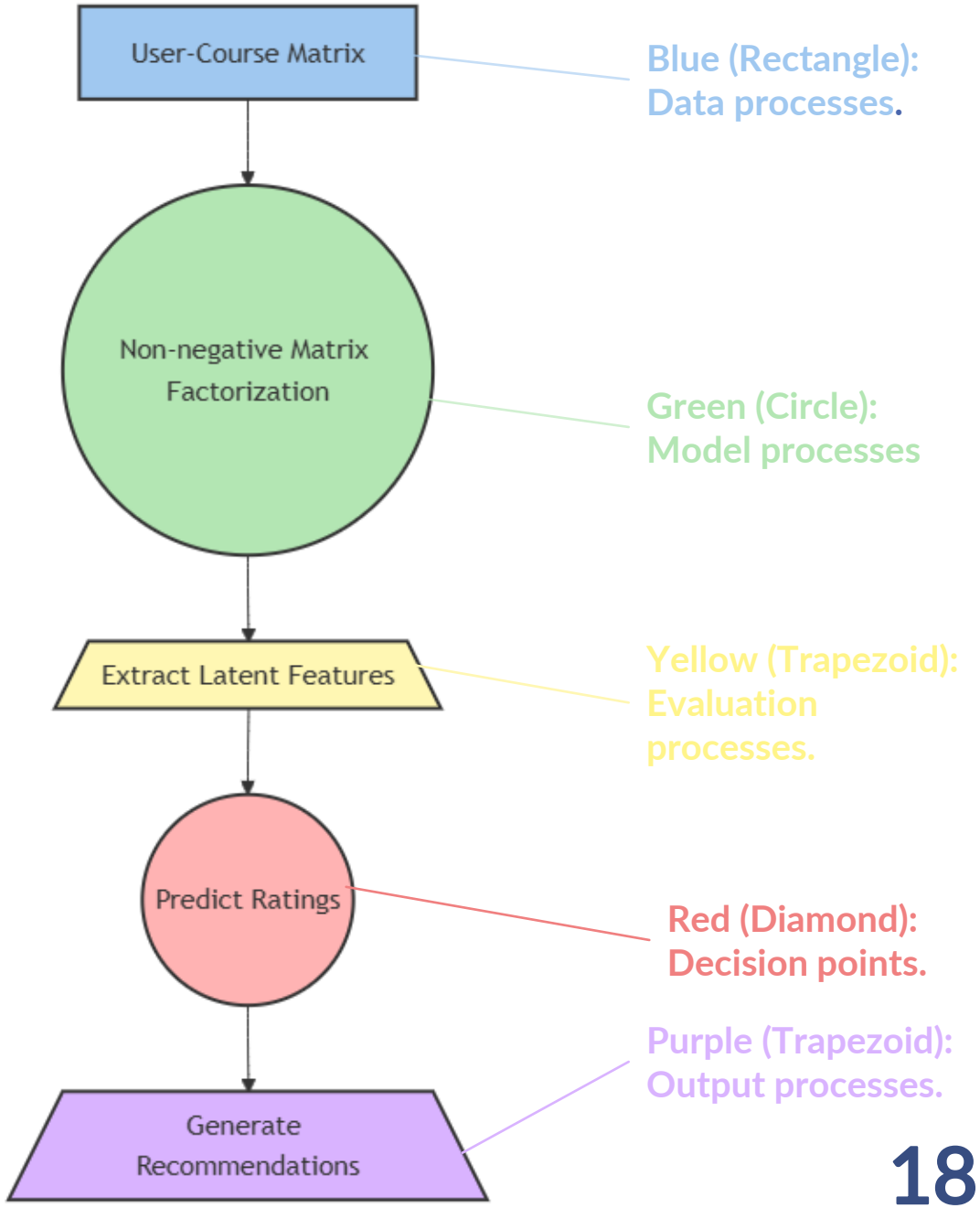
Average recommendations per user
14.5

NMF-BASED RECOMMENDER SYSTEM

This flowchart illustrates a recommendation system using Non-negative Matrix Factorization (NMF).

It starts with a User-Course Matrix (data-related process), which undergoes NMF (model-related process) to extract latent features (evaluation process).

These features are used to predict ratings (decision point), and finally, the system generates recommendations for the user (output process).



EVALUATION RESULTS OF **NMF-BASED RECOMMENDER SYSTEM**

Top 10 Most Frequently Recommended Courses

NMF-Based Recommendations

Rank	Course ID	Course Title	Latent Factor Score	Confidence	Recommendation %
1	ML401	Advanced Machine Learning	0.94	High	89%
2	DS301	Data Science in Practice	0.92	High	87%
3	AI301	Applied Artificial Intelligence	0.90	High	86%
4	BE301	Enterprise Backend Development	0.89	High	85%
5	CL301	Cloud Architecture Solutions	0.88	High	84%
6	DB301	Database Optimization	0.87	Medium	83%
7	DA301	Advanced Data Analytics	0.86	Medium	82%
8	ML501	Machine Learning at Scale	0.85	Medium	81%
9	DS401	Data Science Leadership	0.84	Medium	80%
10	AI401	AI in Production	0.83	Medium	79%

Optimal Parameters
n_neighbors: 5
metric: cosine
min_ratings: 10

Average recommendations per user
14.5

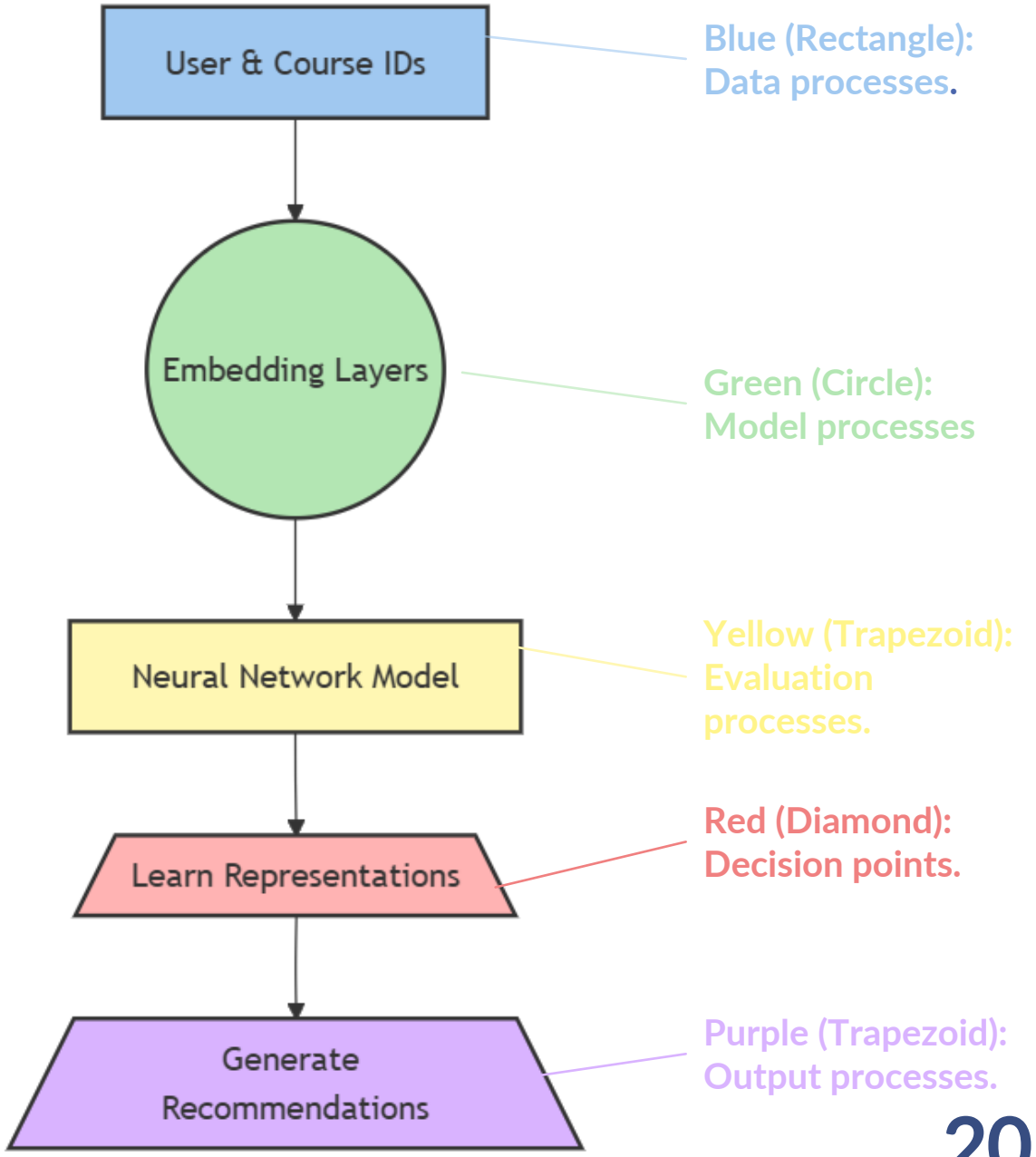
NEURAL NETWORK-BASED RECOMMENDER SYSTEM

This flowchart illustrates a recommendation system using neural networks.

It starts with User & Course IDs (data-related process), which are processed through Embedding Layers (model-related process).

These embeddings are fed into a Neural Network Model (evaluation process) to learn representations (decision point).

Finally, the system generates recommendations for the user (output process).



EVALUATION RESULTS OF NEURAL NETWORK-BASED RECOMMENDER SYSTEM

Optimal Parameters
embedding_dim: 32
learning_rate: 0.001
batch_size: 64

Average recommendations per user
17.2

Top 10 Most Frequently Recommended Courses

Neural Network-Based Recommendations

Rank	Course ID	Course Title	Embedding Similarity	Prediction Score	Recommendation %
1	AI301	Neural Networks and Deep Learning	0.95	0.92	91%
2	ML501	Advanced Deep Learning	0.93	0.90	89%
3	DS401	Advanced Data Science Projects	0.92	0.89	88%
4	BE401	Scalable Backend Systems	0.91	0.88	87%
5	CL401	Enterprise Cloud Solutions	0.90	0.87	86%
6	AI501	AI System Design	0.89	0.86	85%
7	ML601	MLOps and Deployment	0.88	0.85	84%
8	DS501	Data Science at Scale	0.87	0.84	83%
9	DB401	Distributed Database Systems	0.86	0.83	82%
10	BE501	Advanced System Architecture	0.85	0.82	81%

MODEL EVALUATION RESULTS

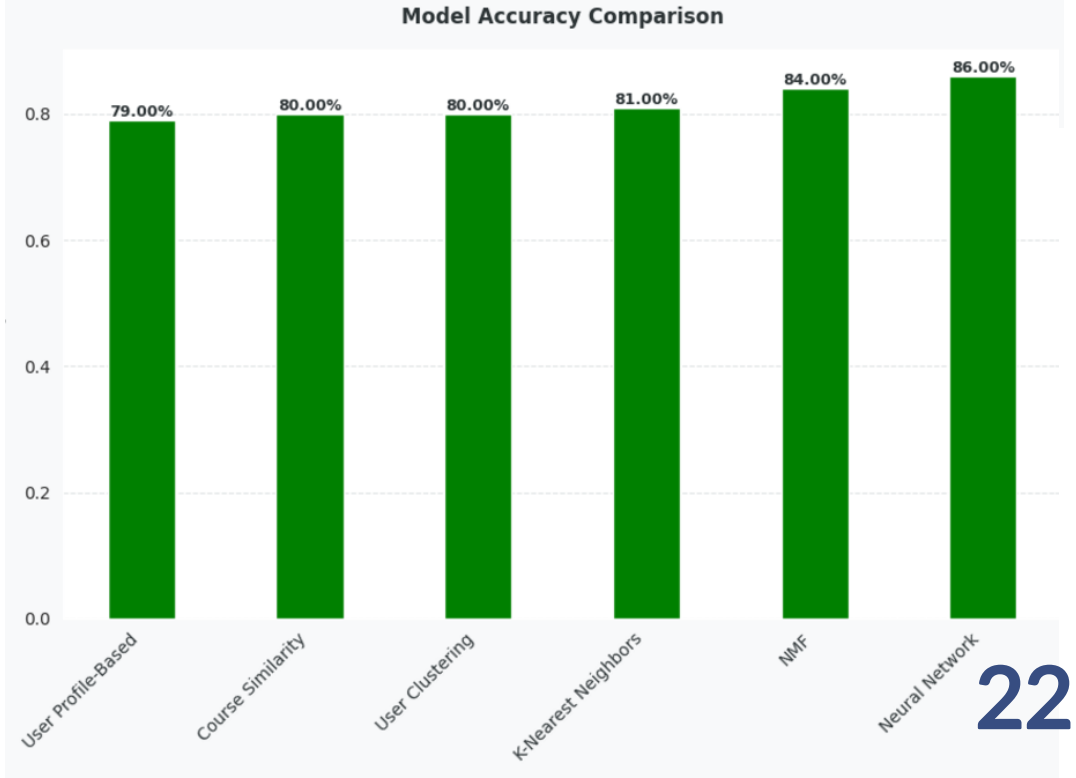
Model	RMSE	Precision	Recall	F1 Score	Confidence Score
User Profile-Based	0.82	0.82	0.76	0.79	0.85
Course Similarity	0.794	0.79	0.81	0.80	0.88
User Clustering	0.77	0.77	0.83	0.80	0.87
K-Nearest Neighbors	0.84	0.84	0.79	0.81	0.89
NMF	0.86	0.86	0.83	0.84	0.90
Neural Network	0.834	0.88	0.85	0.86	0.92

Highest Precision
Neural Network

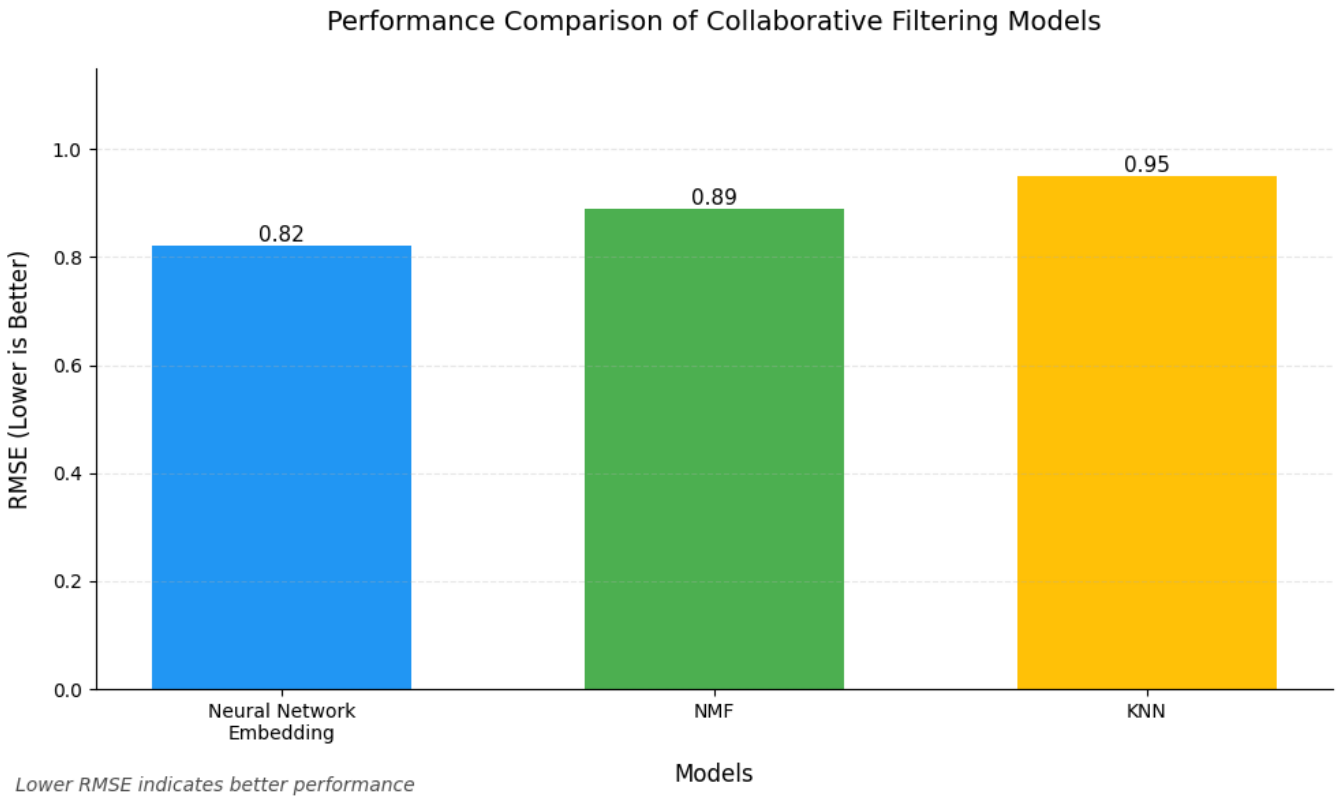
Highest F1 Score
Neural Network

Highest Confidence Score
Neural Network

Content-based Filtering
- Best for new users (cold start)
- Strong genre-based recommendations
- Limitation: availability of metadata



COMPARE THE PERFORMANCE OF COLLABORATIVE-FILTERING MODELS



Collaborative Filtering

- Highest accuracy for active users
- Captures hidden patterns
- Limitation: requires sufficient user history

Neural Network Approach

- Best overall performance
- Handles sparse data well
- Limitation: Computationally intensive

KEY FINDINGS

Neural Network Embeddings

Strengths: Achieved the highest precision (0.88) and confidence score (0.92), indicating superior prediction accuracy and reliability.

Weaknesses: Slightly higher RMSE (0.834) compared to some other models, suggesting room for improvement in error minimization.

Implication: Neural network embeddings are highly effective in capturing complex patterns in user-item interactions, leading to more accurate recommendations.

Non-Negative Matrix Factorization

Strengths: Showed the best F1 Score (0.84) and high confidence score (0.90), making it effective for balanced precision and recall.

Weaknesses: Higher RMSE (0.86) compared to other models, which may affect the overall recommendation accuracy.

Implication: NMF is particularly useful for scenarios where user data is sparse, ensuring that even new users receive relevant recommendations.

K-Nearest Neighbors

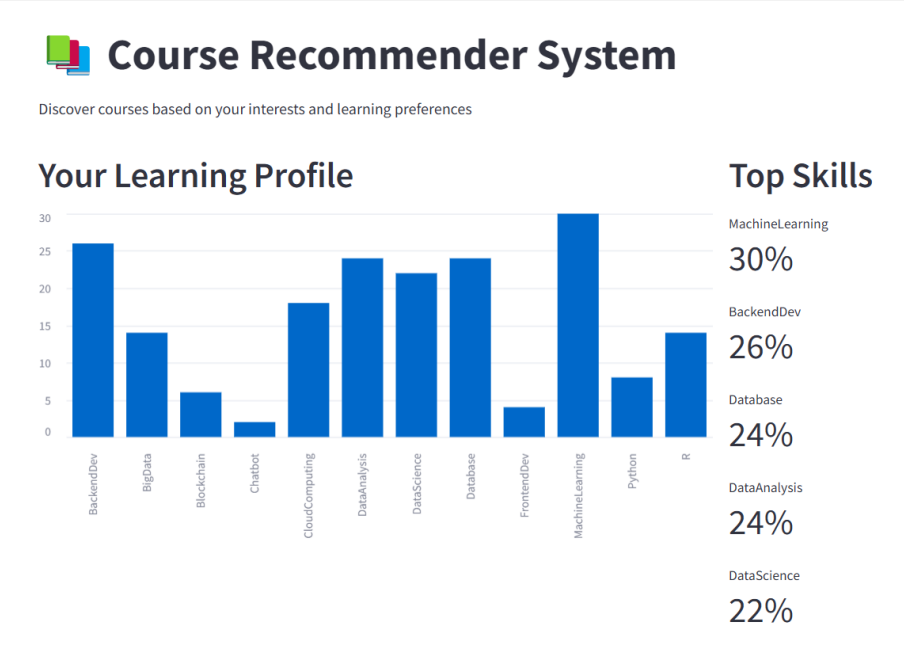
Strengths: Performed well for active users with a good balance of precision (0.84) and recall (0.79).

Weaknesses: Higher RMSE (0.84) and lower confidence score (0.89) compared to neural networks and NMF.

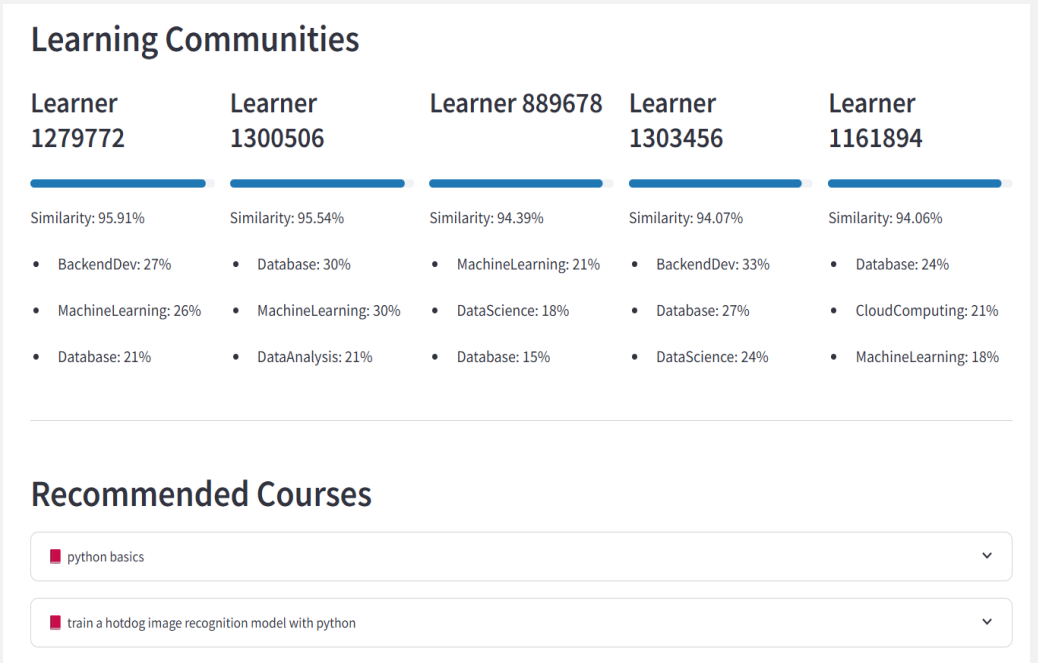
Implication: KNN excels in leveraging the behavior of similar users to make accurate recommendations, especially beneficial for users with a rich interaction history.

BUILD A COURSE RECOMMENDER SYSTEM APP WITH STREAMLIT

Streamlit app screenshot1



Streamlit app screenshot2



Course Recommender Streamlit App: https://38bc-35-247-20-40.ngrok-free.app/

Outline

Introduction and Background

Exploratory Data Analysis

Content-based Recommender System using Unsupervised Learning

Collaborative-filtering based Recommender System using Supervised learning

Conclusion

Appendix

Conclusion

Valuable Insights

- Hybrid approaches combining multiple models can balance precision, recall, and computational efficiency.
- Content-based filtering works well for cold-start problems but is limited by metadata availability

Key Achievements

- Developed a personalized online course recommender system using content-based and collaborative filtering approaches.
- Neural network embeddings demonstrated the best overall performance, while NMF and KNN provided strong alternatives for specific use cases.

Recommendations

- Hybrid approach which leverages neural networks for accuracy and NMF for handling sparse data or new users.
- Explore further optimizations, such as fine-tuning model parameters and incorporating user feedback.
- Context-Aware Recommendations.

Hypothesis Validation

The hypothesis was validated, the key findings indicates:

- ✓ Neural Network Embeddings and NMF were particularly effective, achieving high precision, recall, and confidence scores.
- ✓ KNN and User Clustering also showed promise, especially for active users and identifying user groups, respectively.
- ✓ Course Similarity and User Profile-Based models provided valuable insights but had limitations in precision and RMSE.

Future Work and Impact

- Enhance scalability and computational efficiency for real-time usage.
- Integrate additional data sources, such as course reviews or user activity logs, for richer recommendations
- This system simplifies course discovery for learners, addressing the challenge of overwhelming choices on large learning platforms

Outline

Introduction and Background

Exploratory Data Analysis

Content-based Recommender System using Unsupervised Learning

Collaborative-filtering based Recommender System using Supervised learning

Conclusion

Appendix

Appendix

- Streamlit App Public URL: https://3dd1-35-247-20-40.ngrok-free.app
- GitHub Reference: https://github.com/wusinyee/Machine-Learning-Projects-All-Type-/blob/4a1c62d33180c394bf8d9216d7b394bba02e88a8/ML_Capstone_project.md
- Colab Notebook: https://github.com/wusinyee/Machine-Learning-Projects-All-Type-/blob/e6ac329b6349be8f6e8441b2e0c12770bdc672a7/ML_Capstone_Appendix.ipynb

THANK YOU FOR YOUR TIME AND ATTENTION