Machine Learning Capstone Project

Personalized Recommender System for Online Courses

Yutong He

2024/11/05

IBM Machine Learning

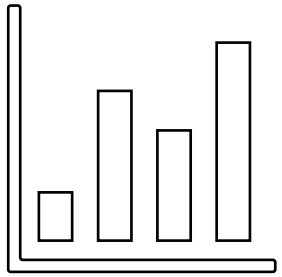
Outline

- Introduction and background
- Exploratory data analysis
- Content-based recommender system using unsupervised learning
- Collaborative-filtering based recommender system using supervised learning
- Conclusion
- Appendix

Introduction

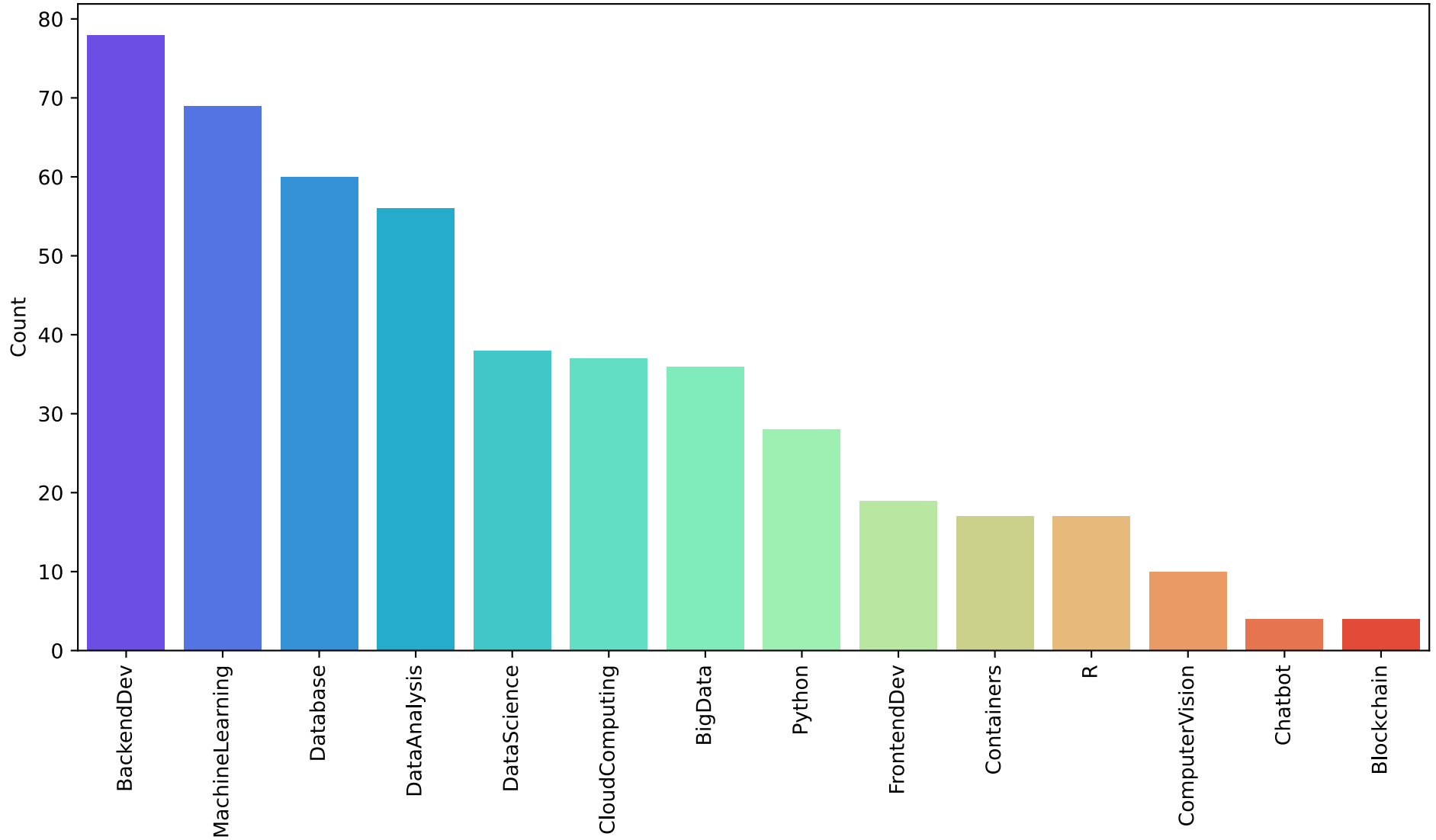
- Project background and context
  - Many platforms provide online courses for learners around the world. Since each learner has their own goals and preferences, it is beneficial for course providers to personalize recommendations to existing and prospective learners. This project explores different scientific methods to analyze data features and develop personalized recommender systems via machine learning (ML).
- Problem states and hypotheses
  - The problem of personalizing recommendations takes either the content-based or collaborative-filtering based approaches. In the case of online courses, the former approach provides recommendations based on the courses a learner has already enrolled. The latter approach comes up with recommendations based on what several other learners with similar preferences have enrolled in.
  - This project will also make use of both unsupervised and supervised ML techniques, where clustering similar features together is the main objective of the former technique, and training for a target variable is the goal of the latter.

Exploratory data analysis



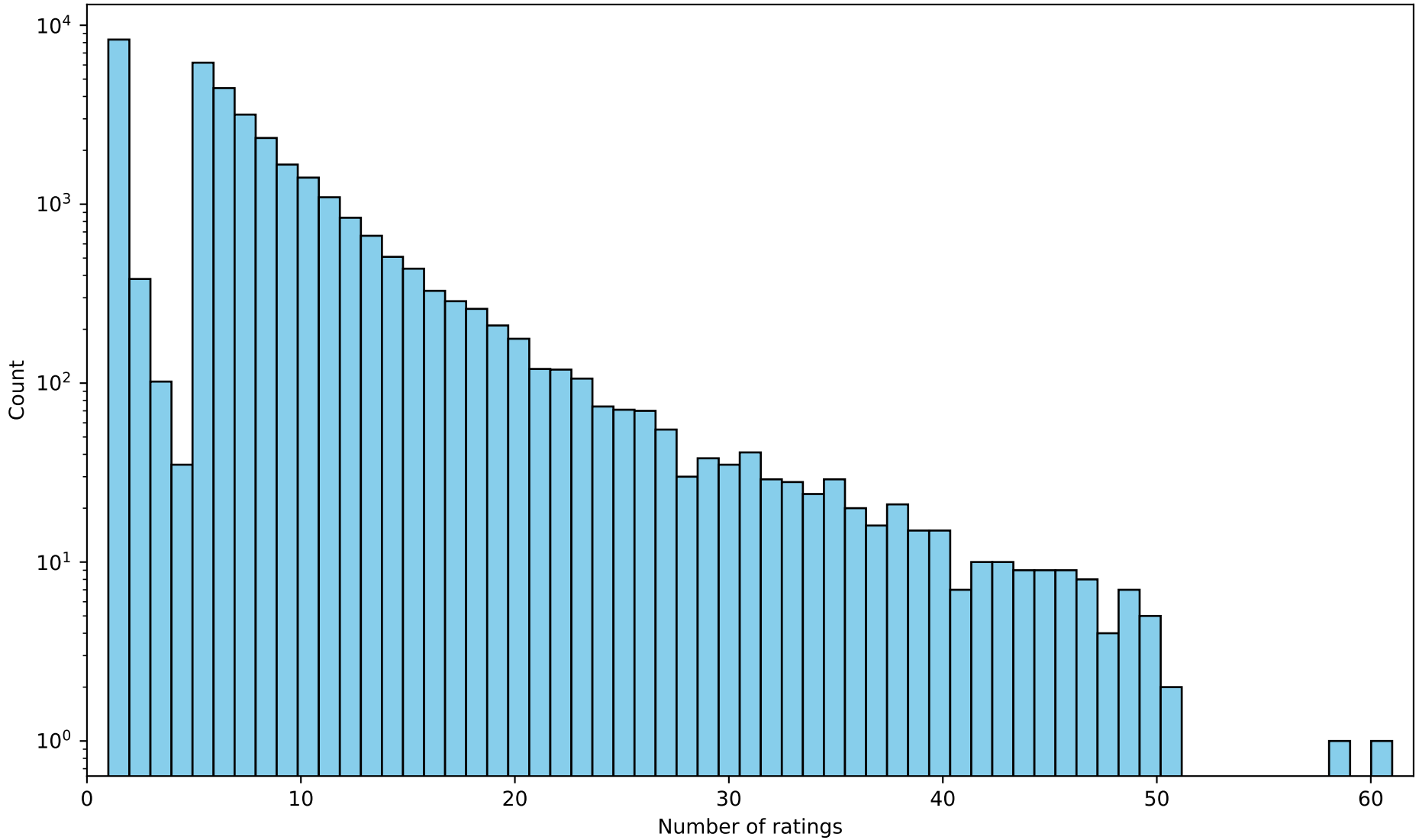
Course genre counts

The bar chart illustrates the course counts across genres, ranked from the most popular onwards.



Course enrollment/rating distribution

The histogram (semi-log) shows the rating number/enrollment distribution, i.e., how many users rated/enrolled in 1 course, how many in 2, 3 courses and so on.



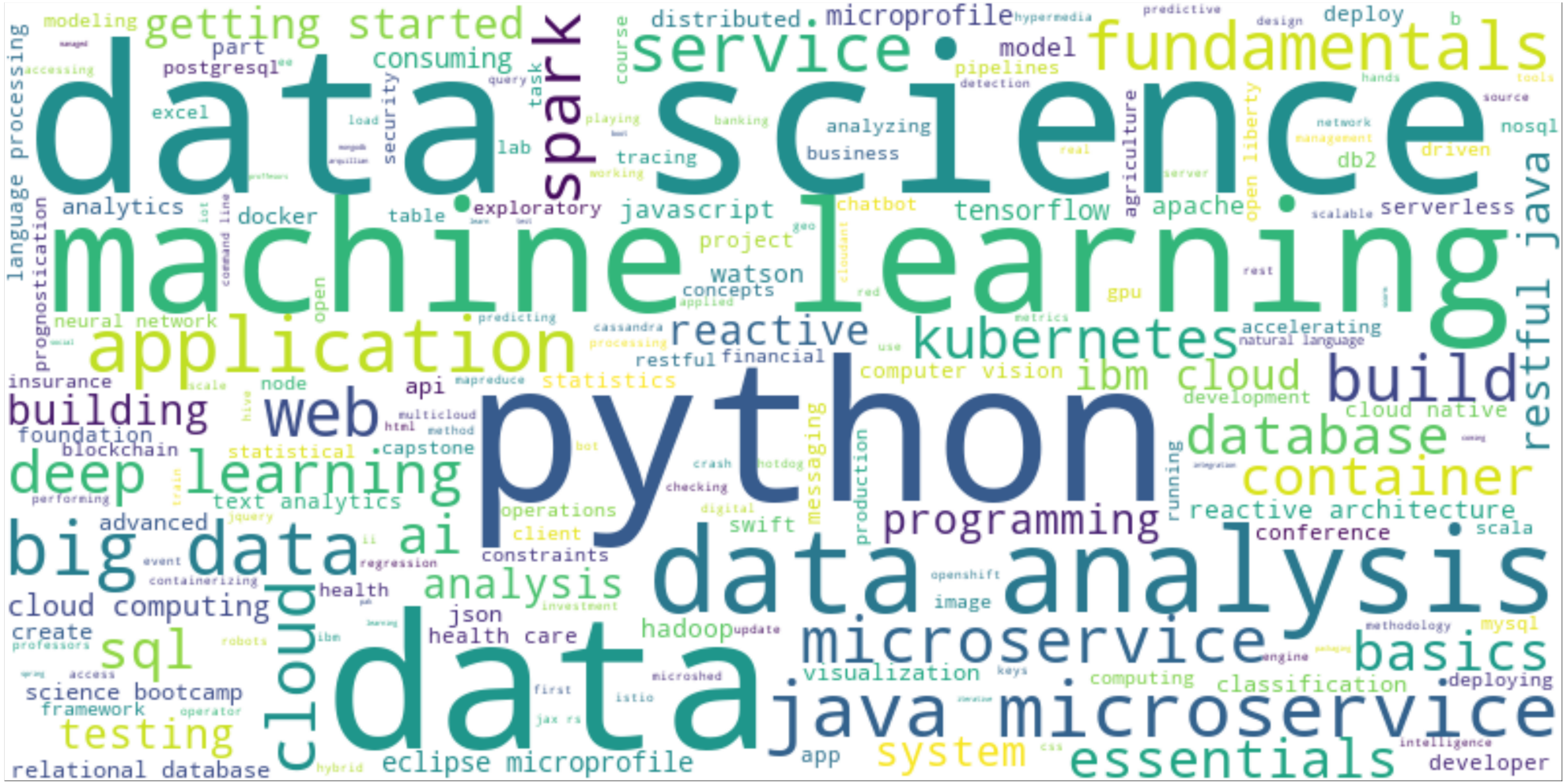
20 most popular courses

- The list of 20 most popular courses in terms of the number of ratings received.
- “Python for data science” course is the most popular, with 14936 ratings.
- All top 20 courses received at least 3624 ratings each.

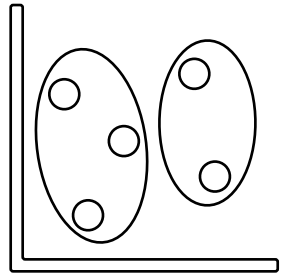
	TITLE	ratings
0	python for data science	14936
1	introduction to data science	14477
2	big data 101	13291
3	hadoop 101	10599
4	data analysis with python	8303
5	data science methodology	7719
6	machine learning with python	7644
7	spark fundamentals i	7551
8	data science hands on with open source tools	7199
9	blockchain essentials	6719
10	data visualization with python	6709
11	deep learning 101	6323
12	build your own chatbot	5512
13	r for data science	5237
14	statistics 101	5015
15	introduction to cloud	4983
16	docker essentials a developer introduction	4480
17	sql and relational databases 101	3697
18	mapreduce and yarn	3670
19	data privacy fundamentals	3624

Word cloud of course titles

The word cloud gives a visual understanding of popular course topics, such as python, data science and machine learning.

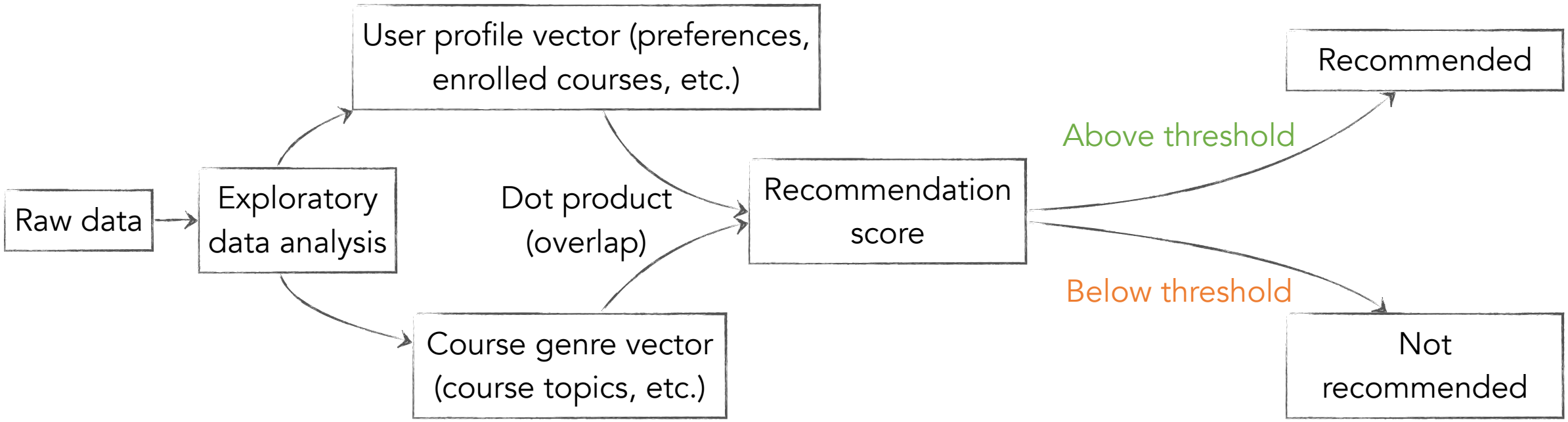


Content-based recommender system using unsupervised learning



Content-based recommender system using ***user profile and course genres***

Flowchart



Content-based recommender system using ***user profile and course genres***

Evaluation

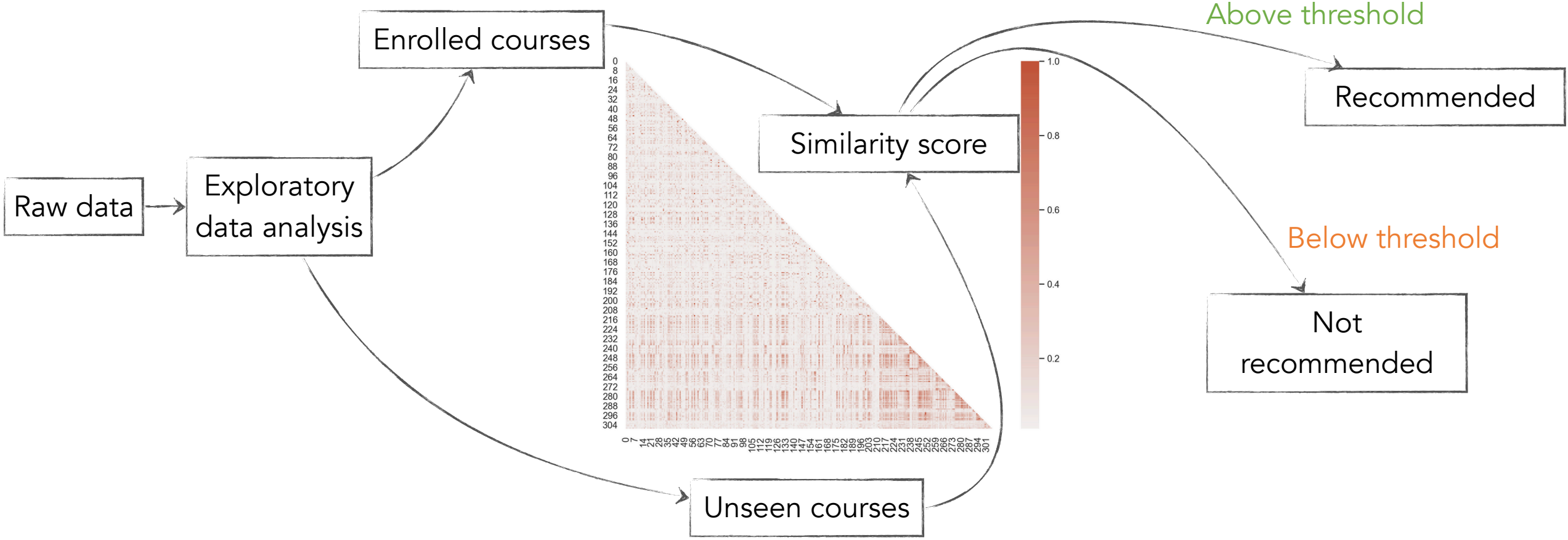
On average, **60.82** courses have been recommended per test user.

Top 10 most frequently recommended courses (all test users):

1. Text analytics at scale
2. Applied machine learning in Python
3. Introduction to data science in Python
4. Data science in insurance basic statistical analysis
5. Accelerating deep learning with GPU
6. SQL for data science
7. SQL for data science capstone project
8. Performing database operations in the Cloudant dashboard
9. Analyzing big data with SQL
10. Foundations for big data analysis with SQL

Content-based recommender system using ***course similarity***

Flowchart



Content-based recommender system using ***course similarity***

Evaluation

With a similarity score **0.6**, on average **1.07** courses are recommended per test user.

With a similarity score of **0.5**, on average **3.16** courses are recommended per test user.

Top 10 most frequently recommended courses (all test users):

1. Introduction to data science in Python
2. Watson analytics for social media
3. Data science with open data
4. Build your own chatbots
5. Data science fundamentals for data analysts
6. A crash course in data science
7. Deep learning with TensorFlow
8. Text analytics 101
9. Accelerating deep learning with GPUs
10. Introduction to cloud computing

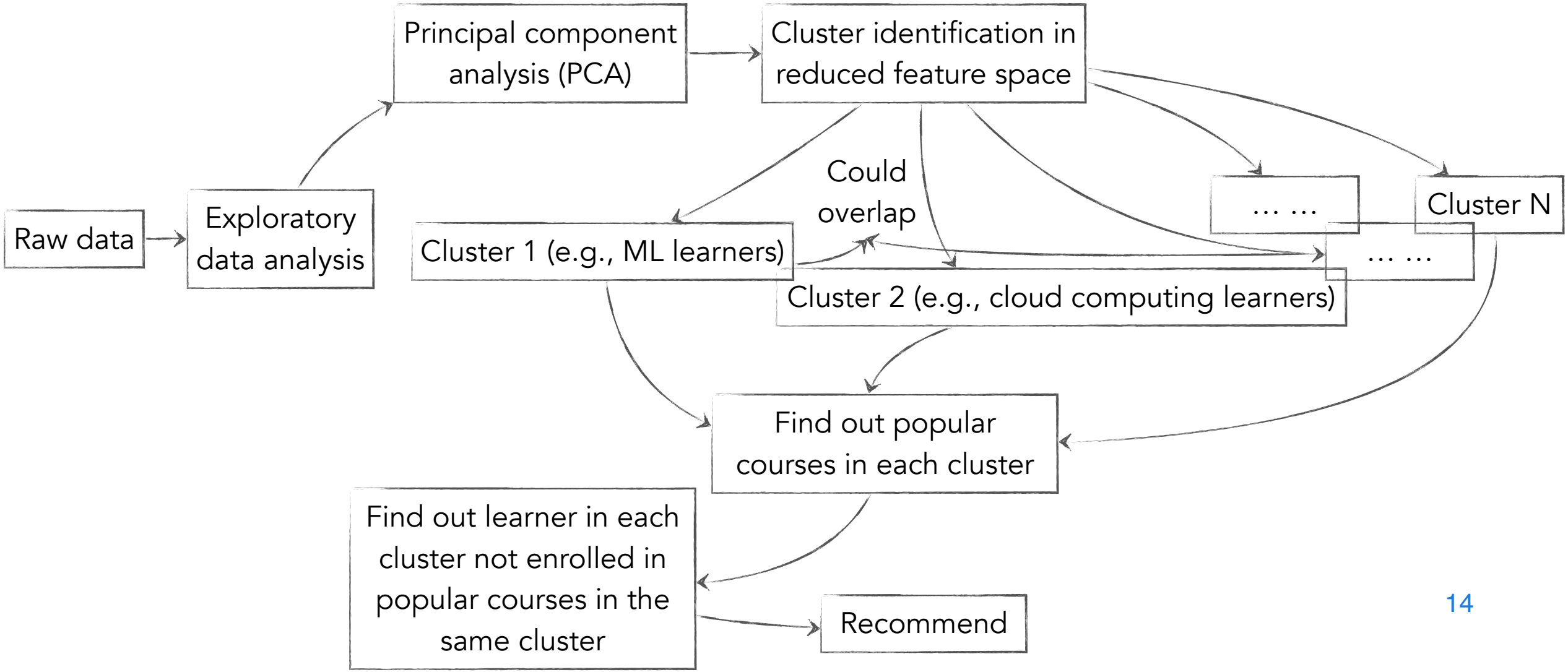
Top 10 most frequently recommended courses (all test users):

1. Data science bootcamp with Python
2. Introduction to data science in Python
3. Introduction to data analytics
4. Data science with open data
5. Big data modeling and management systems
6. Data analysis using R 101
7. Data analysis using Python
8. Data science fundamentals for data analysts
9. Process data from dirty to clean
10. SQL for data science

3 courses overlap with both similarity scores.

Content-based recommender system using ***user profile clustering***

Flowchart



Content-based recommender system using ***user profile clustering***

Evaluation

With a enrollment threshold of **100**, on average **33.53** courses are recommended per test user.

With a enrollment threshold of **200**, on average **23.15** courses are recommended per test user.

Top 10 most frequently recommended courses (all test users):

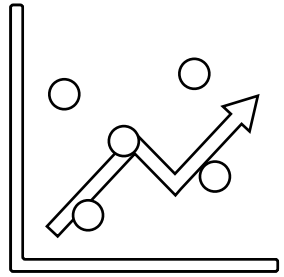
1. Watson analytics 101
2. Statistics 101
3. IBM cloud essentials
4. Docker essentials a developer introduction
5. Scala 101
6. Data privacy fundamentals
7. Introduction to cloud
8. SQL and relational databases 101
9. Digital analytics regression
10. R for data science

Top 10 most frequently recommended courses (all test users):

1. Statistics 101
2. IBM cloud essentials
3. Data science hands on with open source tools
4. Data science methodology
5. Deep learning 101
6. Data privacy fundamentals
7. Introduction to cloud
8. IBM cloud essentials v3
9. SQL and relational databases 101
10. R for data science

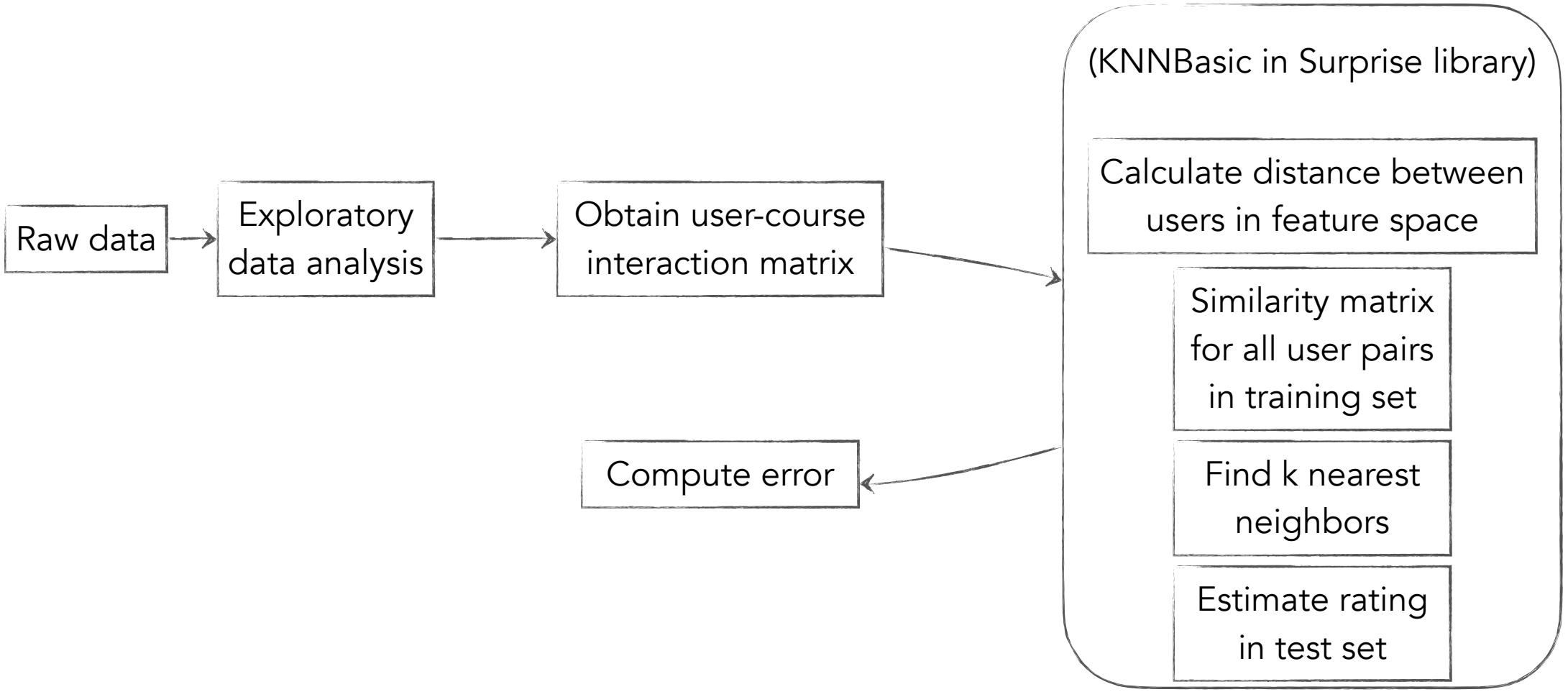
6 courses overlap with both enrollment thresholds.

Collaborative-filtering based recommender system using supervised learning



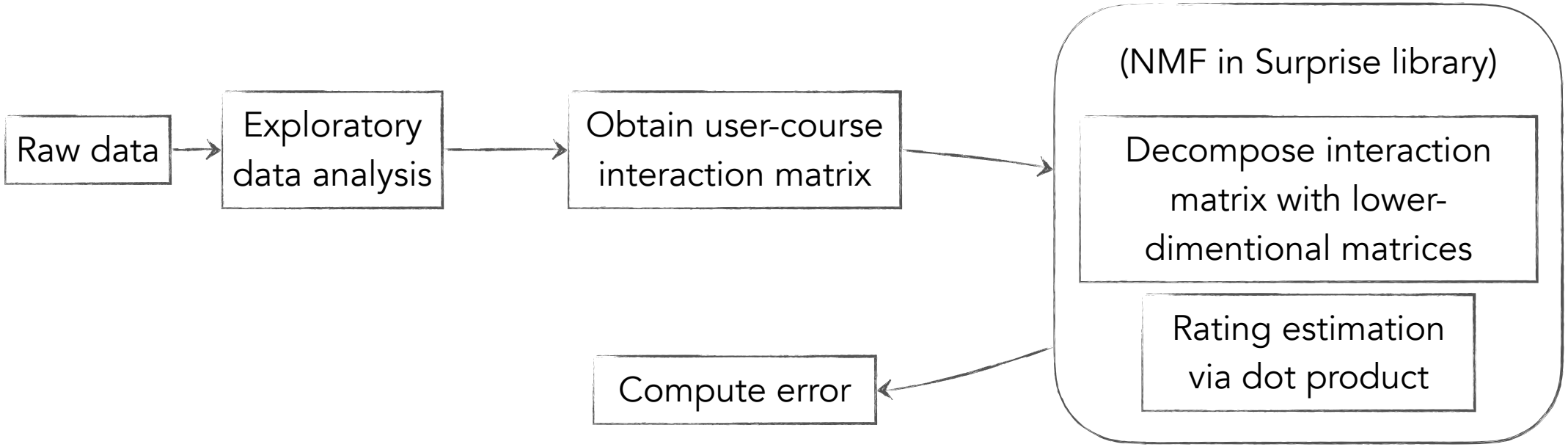
K-nearest neighbor (kNN)-based collaborative filtering

Flowchart



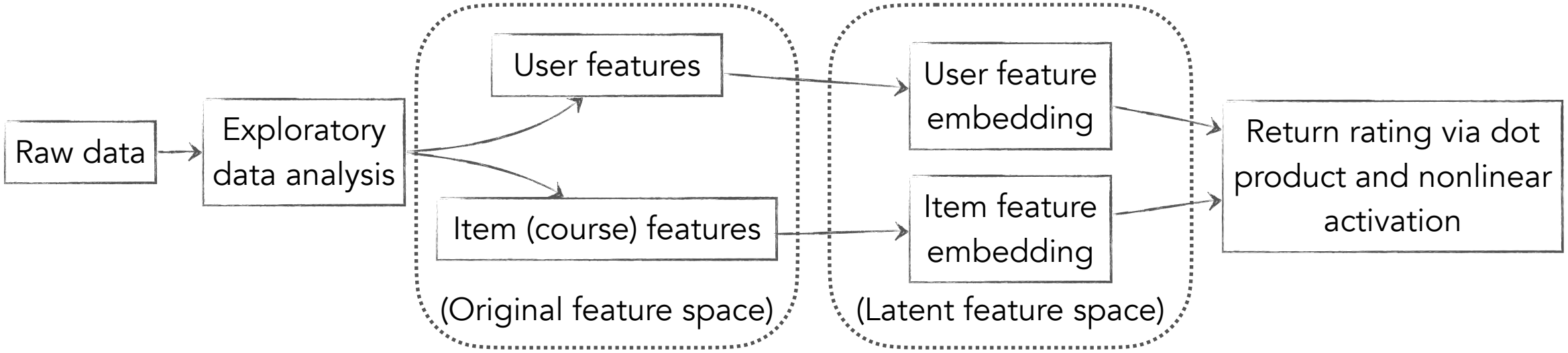
Non-negative matrix factorization (NMF)-based collaborative filtering

Flowchart



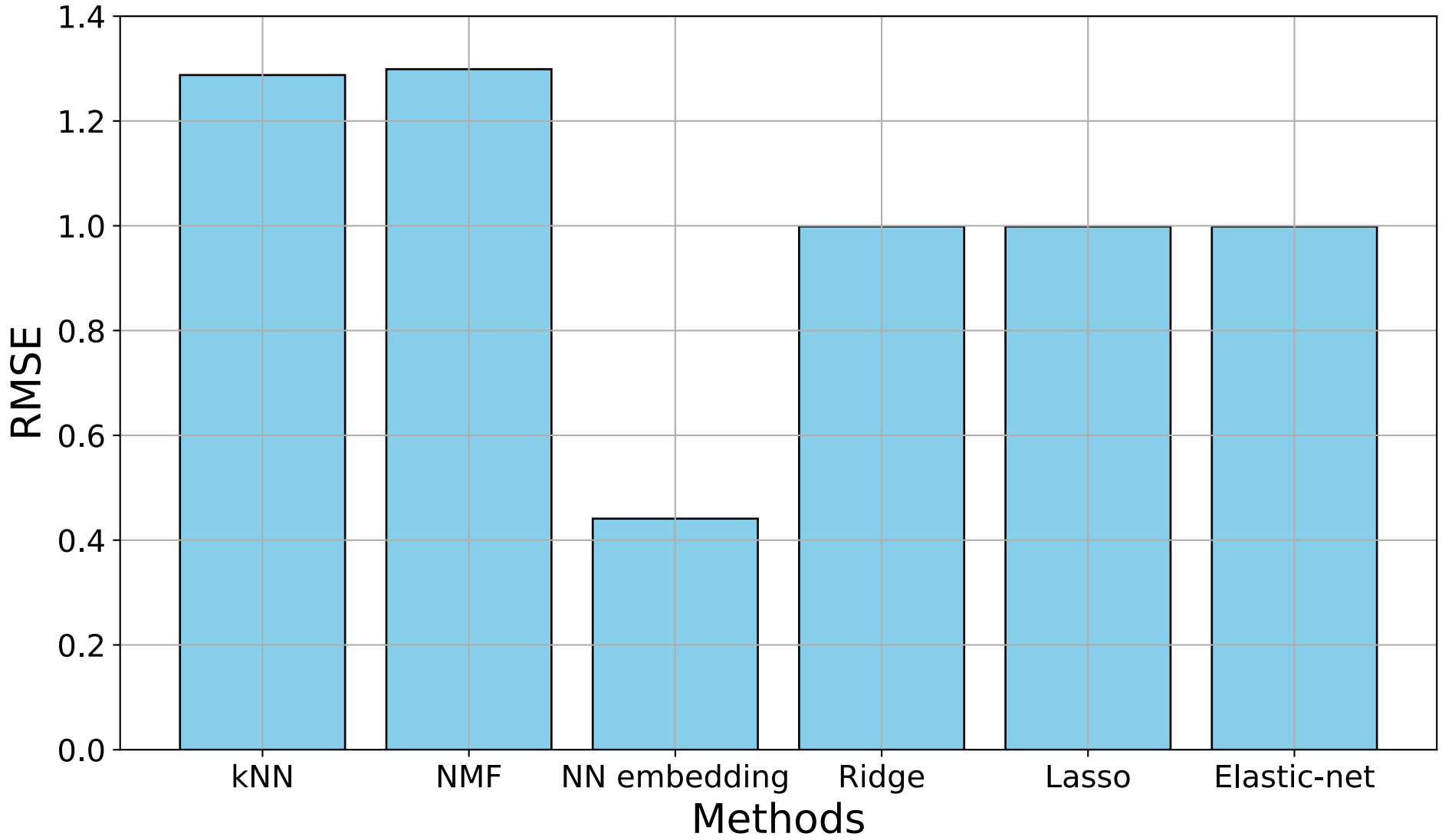
Neural network (NN) embedding-based collaborative filtering

Flowchart



Collaborative filtering algorithm evaluation

The bar chart compares the root mean square error (RMSE) of different collaborative-filtering based models using ML techniques: kNN, NMF, embedding, and linear regression (ridge, lasso, elastic-net).



Conclusions

- From exploratory data analysis:
    - Python and data science are among the most popular course topics.
- From content-based approach:
    - The number of recommended courses depends on user profile, course similarity threshold, and existing enrollment threshold.
    - Some popular courses remain on top regardless of the changes of variables mentioned above. These more or less align with the popular topics extracted from exploratory data analysis.
- From collaborative-filtering based approach:
    - NN embeddings are the best performing model in terms of the root-mean-square error.
    - Output of regression models depends very little on the choice of penalty functions.
    - kNN and NMF perform the worst out of the explored models.

Appendix

- Links to Jupyter Notebooks that produced the results in this presentation:
  - Exploratory data analysis: https://github.com/yutonghe96/ML_material/blob/main/capstone/6.03_EDA.ipynb.
  - User profile: https://github.com/yutonghe96/ML_material/blob/main/capstone/6.04_user_profile.ipynb.
  - Course similarity: https://github.com/yutonghe96/ML_material/blob/main/capstone/6.05_course_similarity.ipynb.
  - Unsupervised clustering: https://github.com/yutonghe96/ML_material/blob/main/capstone/6.06_clustering.ipynb.
  - Supervised kNN: https://github.com/yutonghe96/ML_material/blob/main/capstone/6.07_CF_KNN.ipynb.
  - Supervised NMF: https://github.com/yutonghe96/ML_material/blob/main/capstone/6.08_CF_NMF.ipynb.
  - NN embedding: https://github.com/yutonghe96/ML_material/blob/main/capstone/6.09_CF_NN_embedding.ipynb.
  - Regression: https://github.com/yutonghe96/ML_material/blob/main/capstone/6.10_CF_regression_embedding.ipynb.