



Hybrid Scalable Online Recommendations

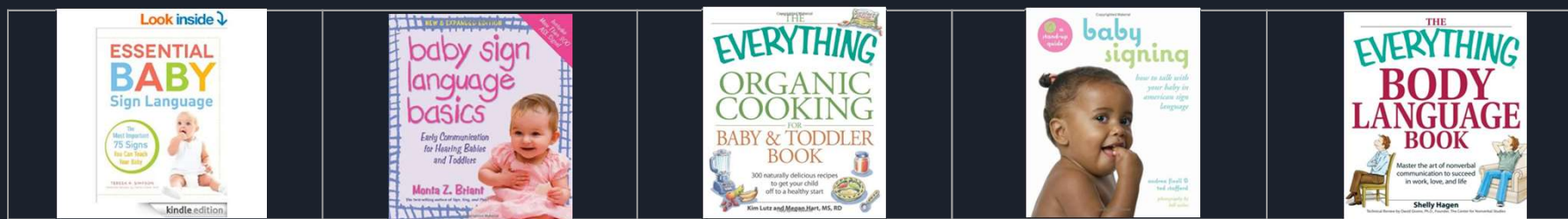
Diego Fanesi
Farooq Qaiser
Sai Chaitanya

Amazon

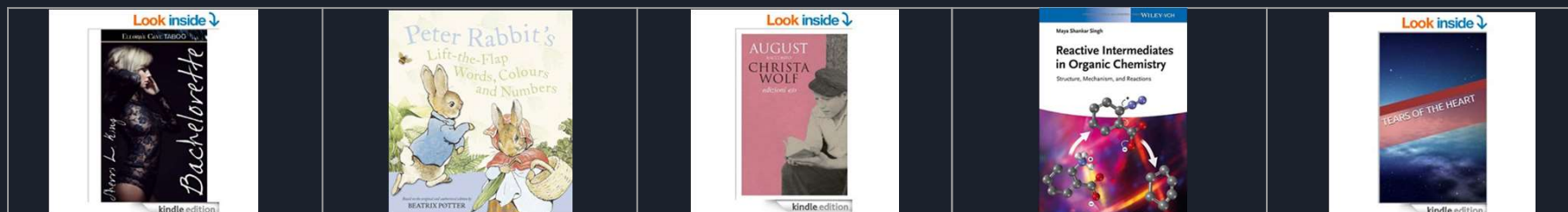


Title: The Everything Baby Sign Language Book
Price: \$9.99

Similar items to what you're viewing



Other items you might like



Problem	Experiments	Final Solution
Holistic recommendations	<ol style="list-style-type: none">1. Content Based Filtering2. Collaborative Filtering	Hybrid model
Scalable implementation	<ol style="list-style-type: none">1. JSON vs Parquet2. Local vs Cluster	<ol style="list-style-type: none">1. Amazon S3 + Parquet2. Amazon EC2 + Spark
Online capabilities	<ol style="list-style-type: none">1. N^2 approach2. LSH approach3. ALS approach4. SSGD approach	<ol style="list-style-type: none">1. Content Based Filtering using LSH2. Collaborative Filtering using SSGD

Content Based Recommender



How it works

User viewed
Item A



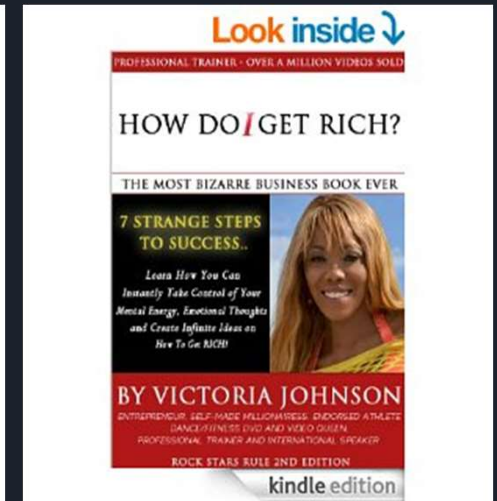
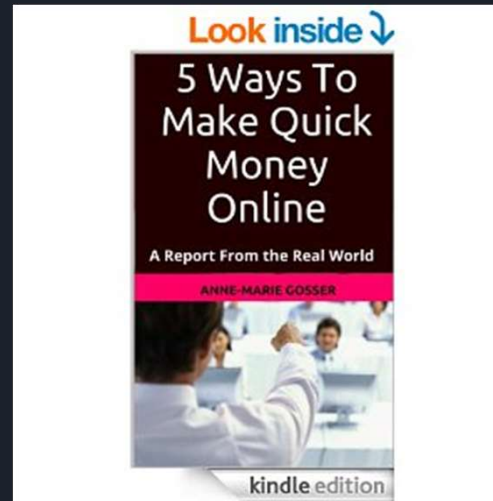
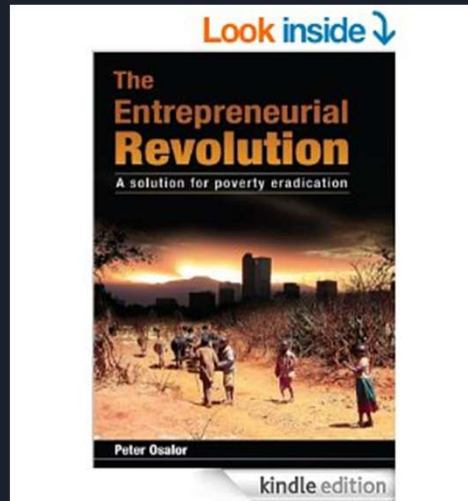
	A	B	C
A	1	0.75	0
B	0.75	1	0.5
C	0	0.5	1



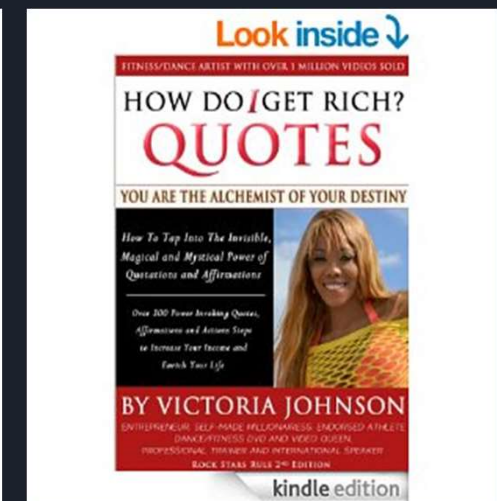
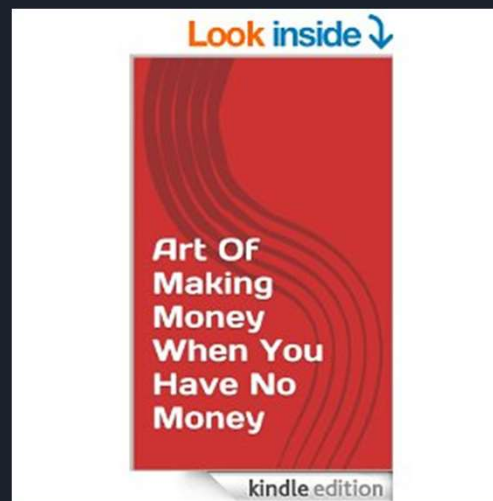
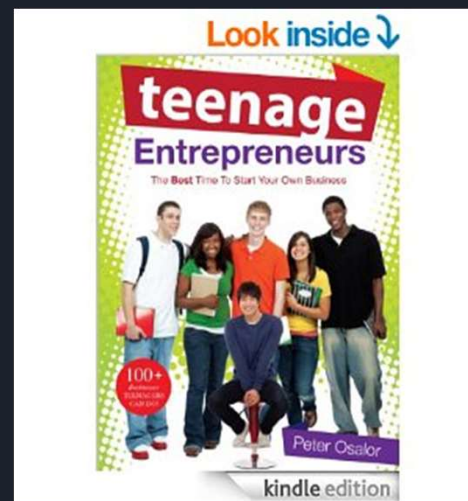
recommend
Item B

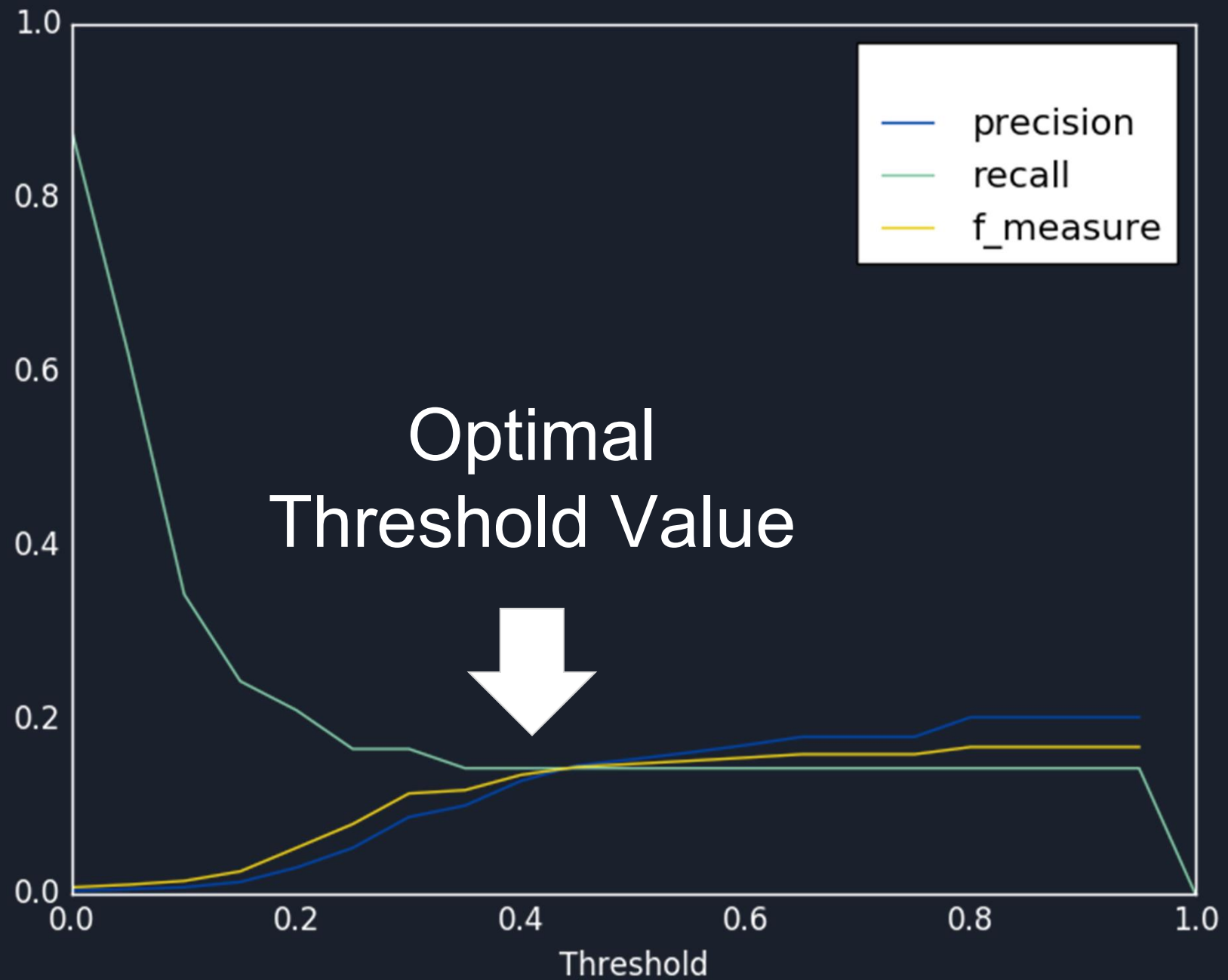
Selected Recommendations

Item



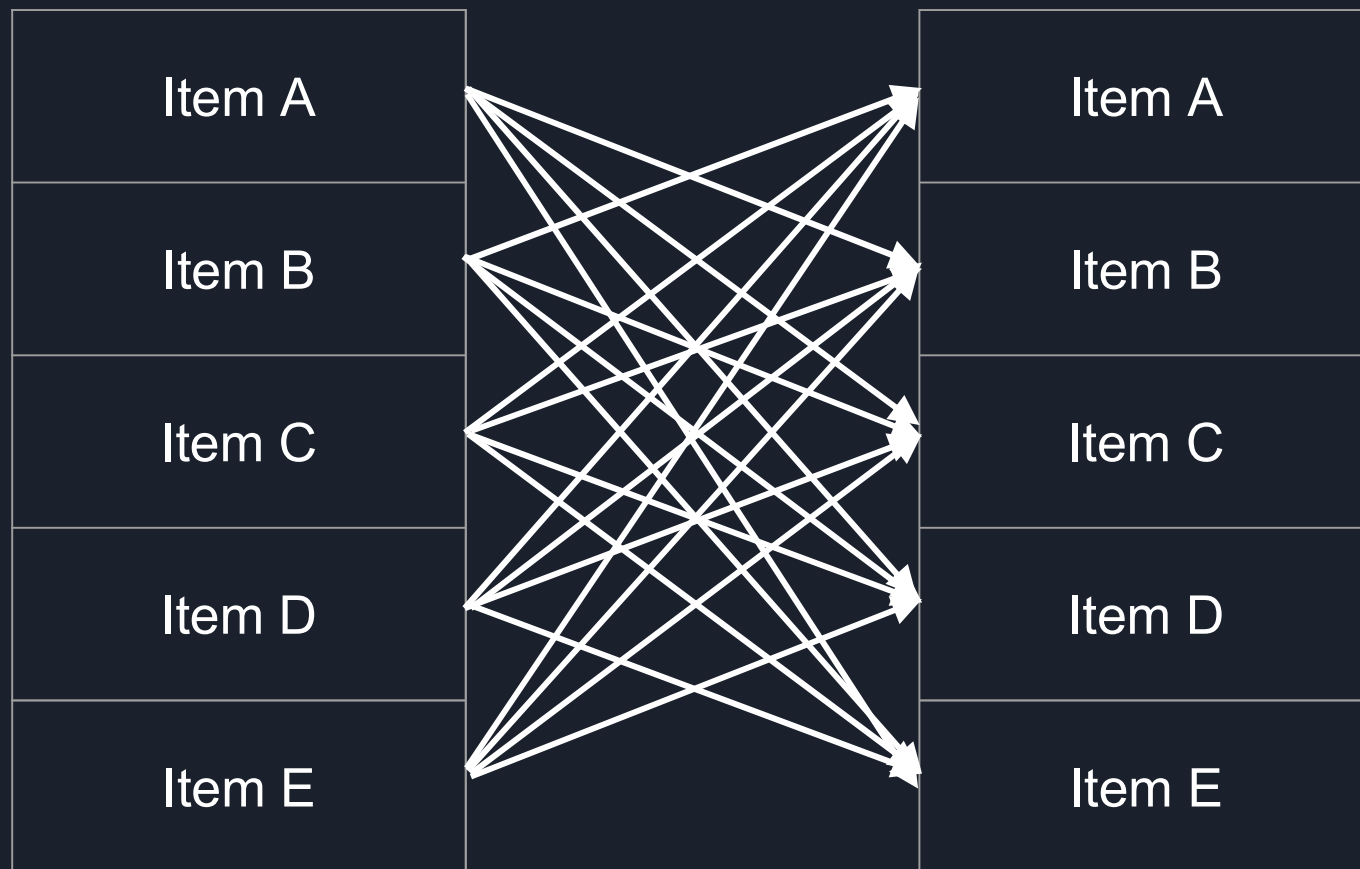
Similar Item





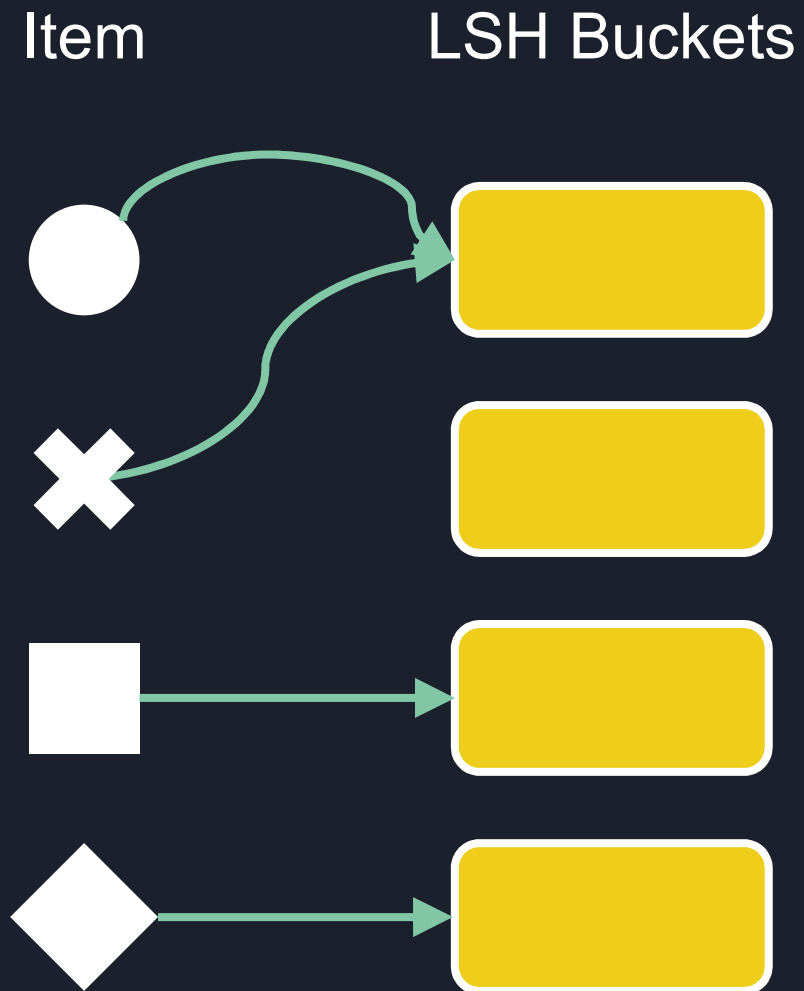
N^2 Approach

How it works



Locality Sensitive Hashing Approach

How it works










Collaborative Filtering Recommender



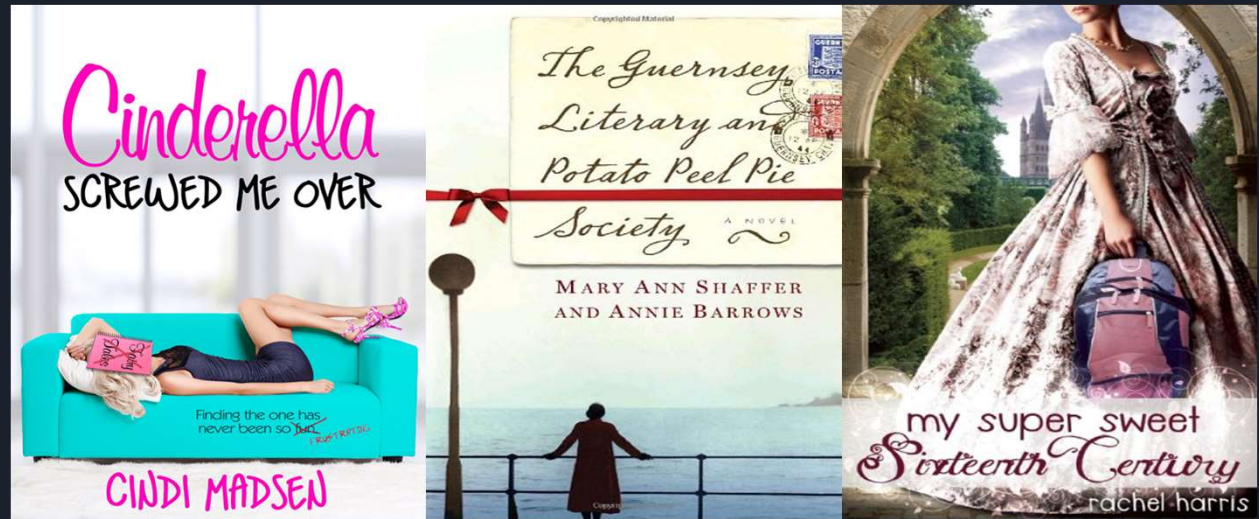
Collaborative Filtering

How it works

		Users		
		A	B	C
Items	1			
	2			
	3			
	4			
	5			

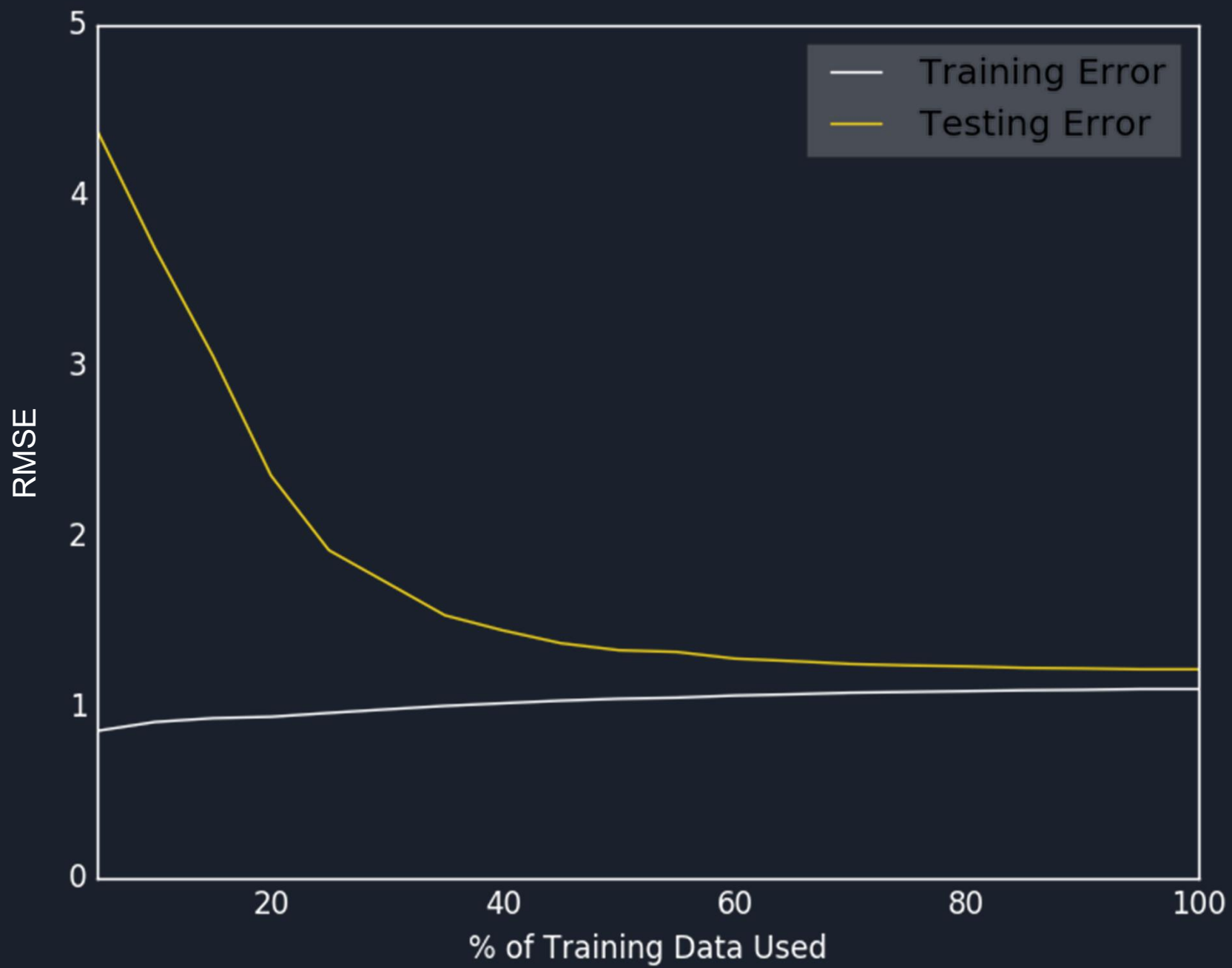
Selected Results

Items rated highly
by the user



Predictions from
the Collaborative
Filtering model





ALS

SSGD

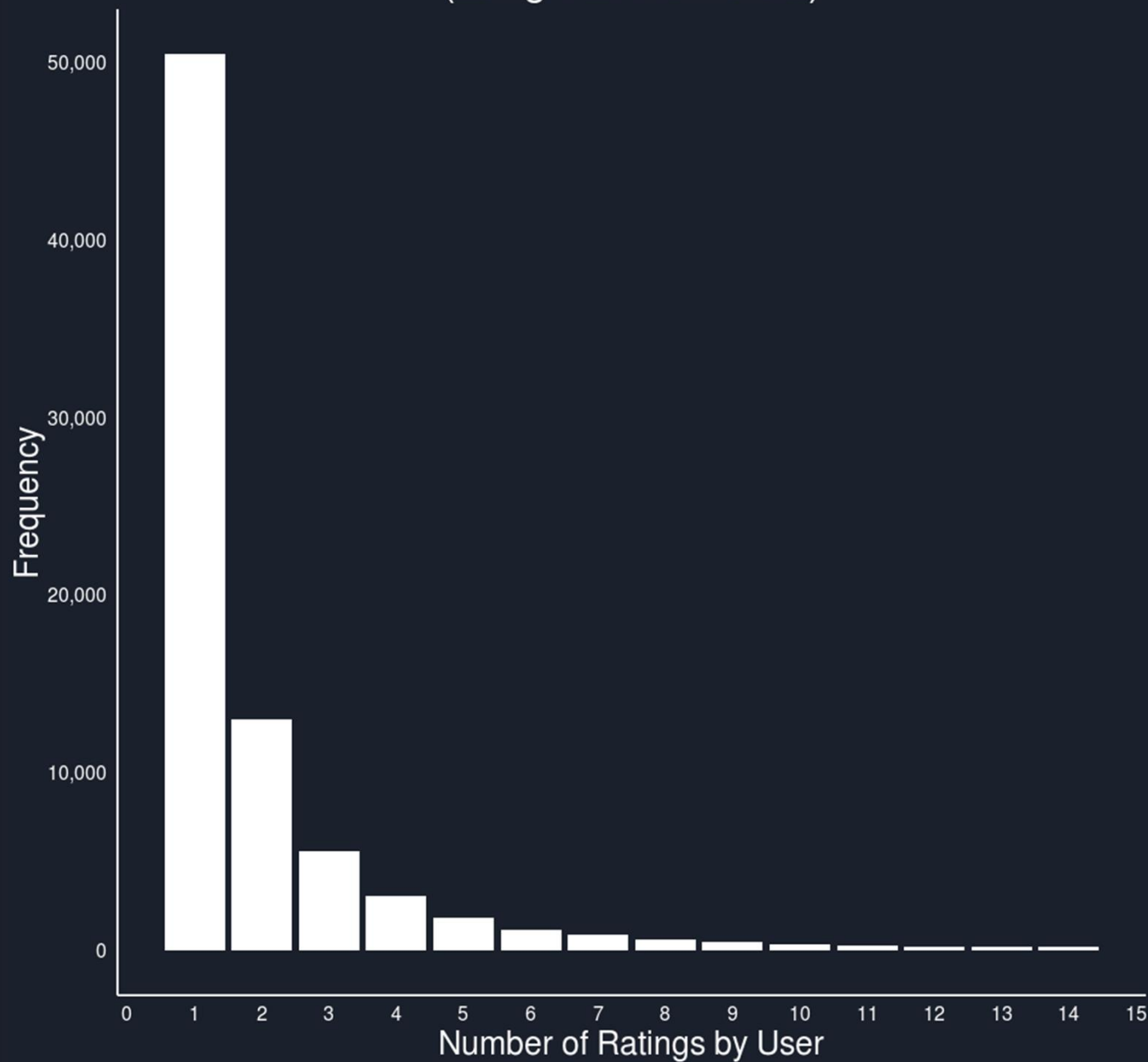
Alternating Least Squares

Stratified and Distributed
Stochastic Gradient Descent

Spark ML and MLlib

Custom Implementation

Histogram of Number of Ratings by Users
(Long Tail Removed)



Conclusions And Productionizing



Amazon



Title
Description
Price

Similar Items to what you're viewing

Content Based Filtering Recommender

Other items you might like

Collaborative Filtering Recommender

Conclusions

1. LSH > N2 for content based approach
2. SGD > ALS for collaborative filtering approach
3. Reviews > Ratings
4. CB is great as a content similarity tool
5. CF great as a content discovery tool

From PoC to Production

- Rewrite in scala to improve performances
- Improve the implementation of algorithms
- Implement Spark Streaming for online prediction and training
- Add autotests to make it more robust
- Tune up the models
- Semantic understanding for Content Based
- ngrams for Content Based

We would like to thank Julian McAuley for graciously sharing the full Amazon Books datasets that were used for this project.



Appendix

Supplemental Material

Storage

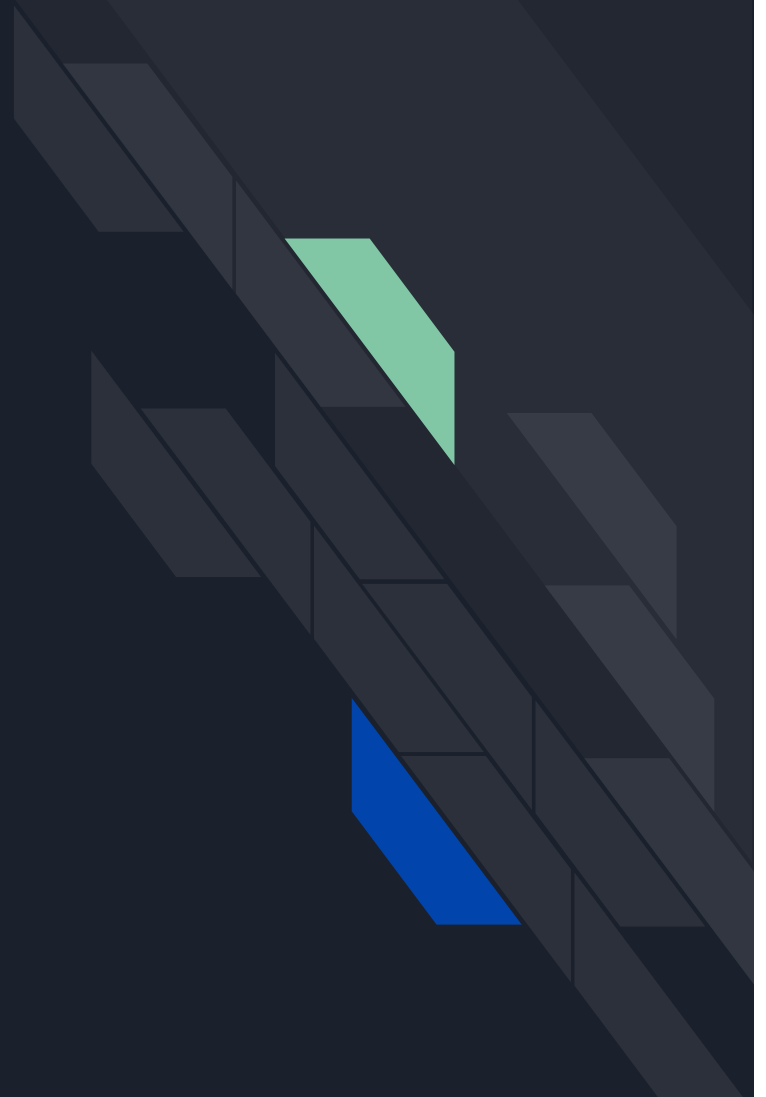


Computation



Exploratory Data Analysis

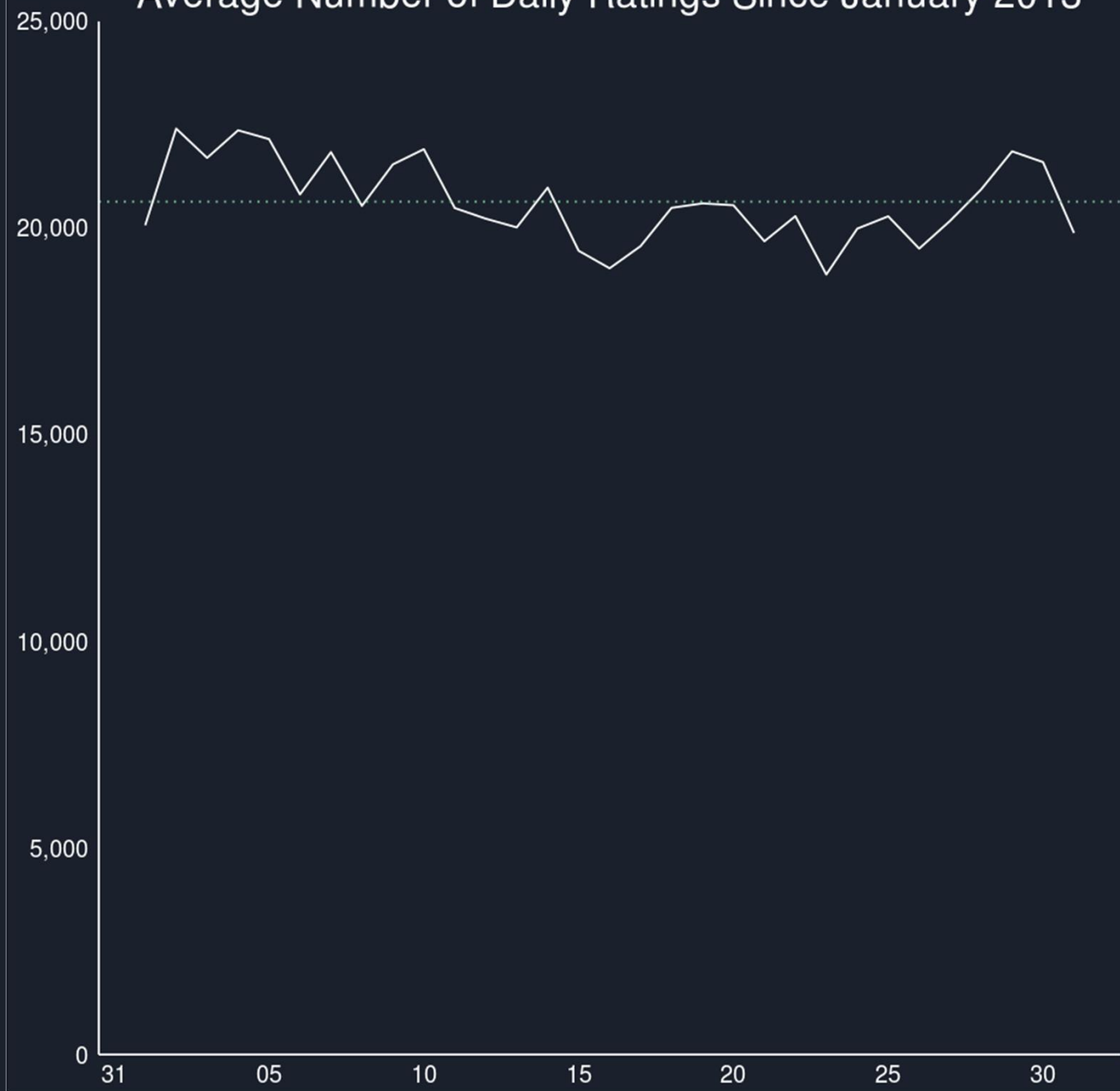
Supplemental Material



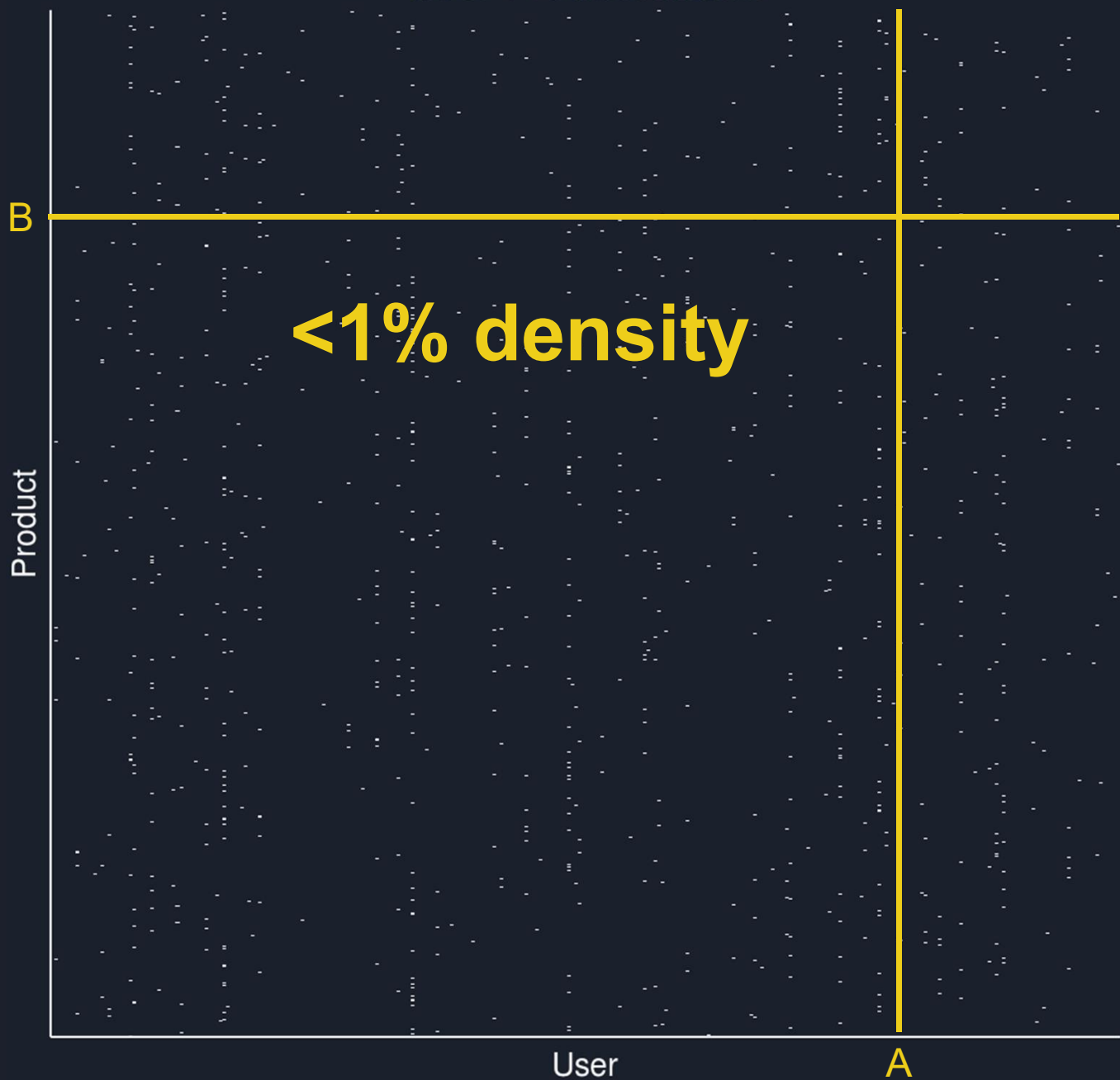
Books dataset

Users	8,026,459
Books	3,941,625
Ratings	22,507,206
Reviews	8,899,474

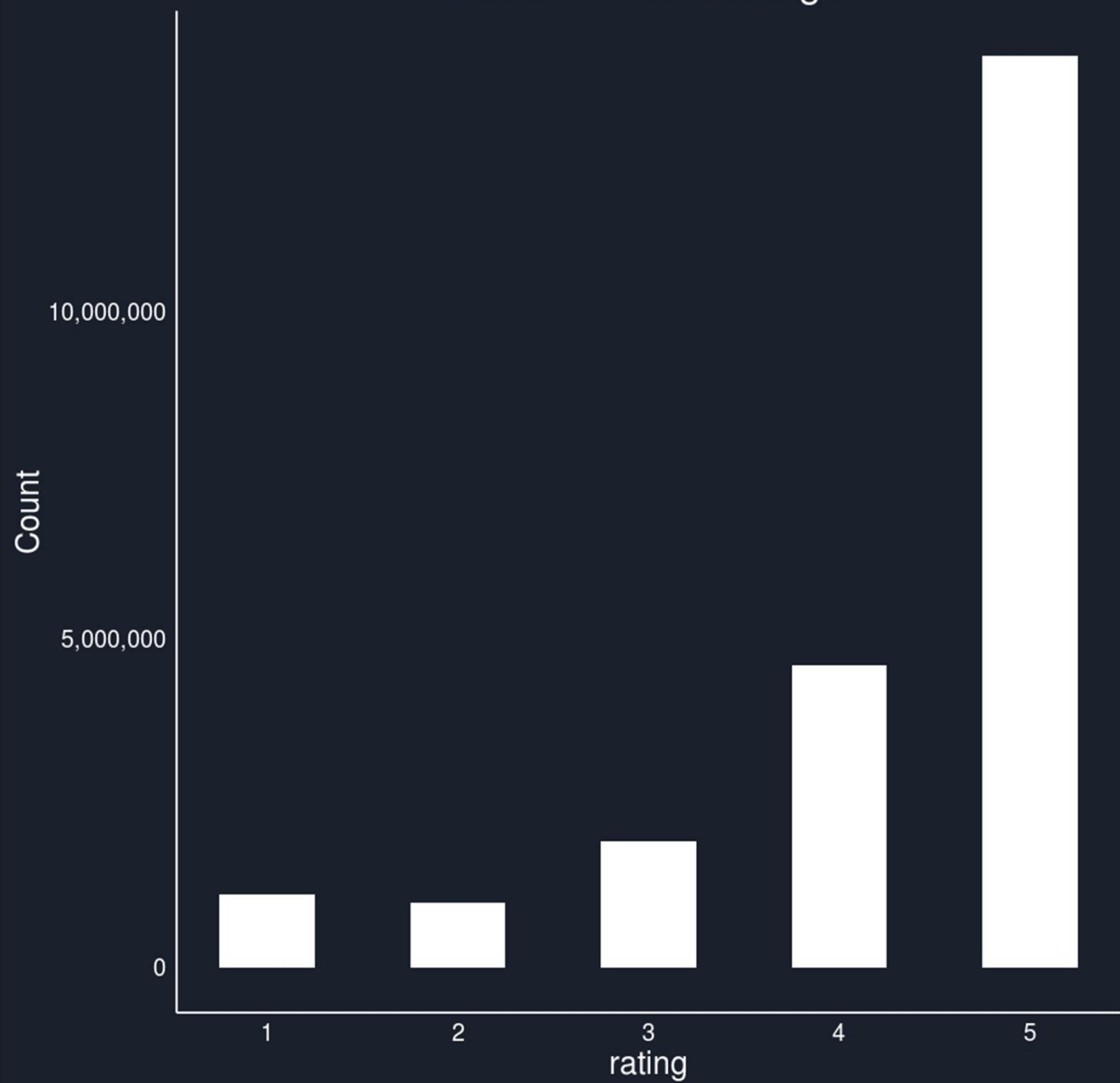
Average Number of Daily Ratings Since January 2013



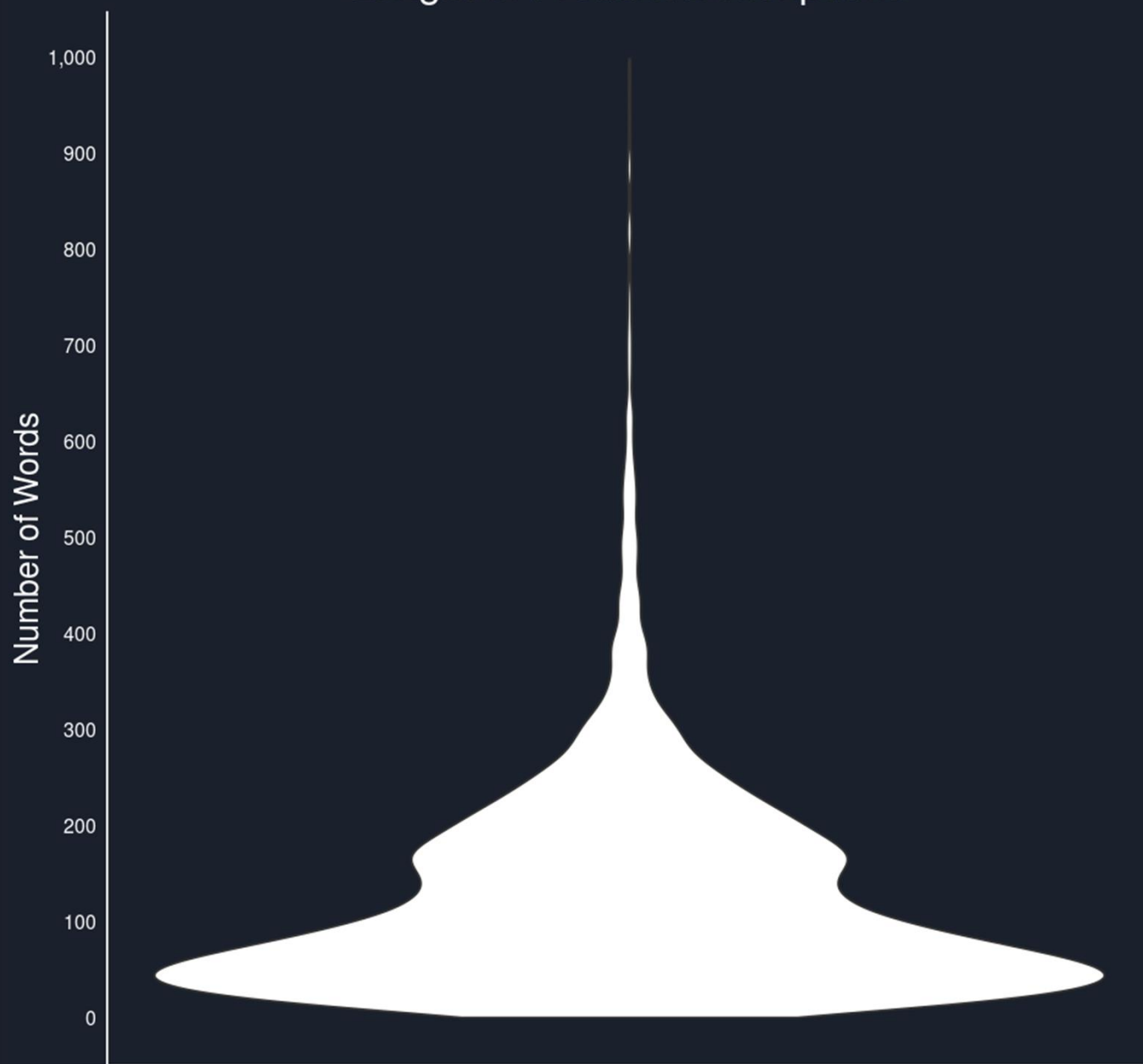
User-Product Matrix



Distribution of Ratings



Length of Product Descriptions



Content Based Recommender

Supplemental Material

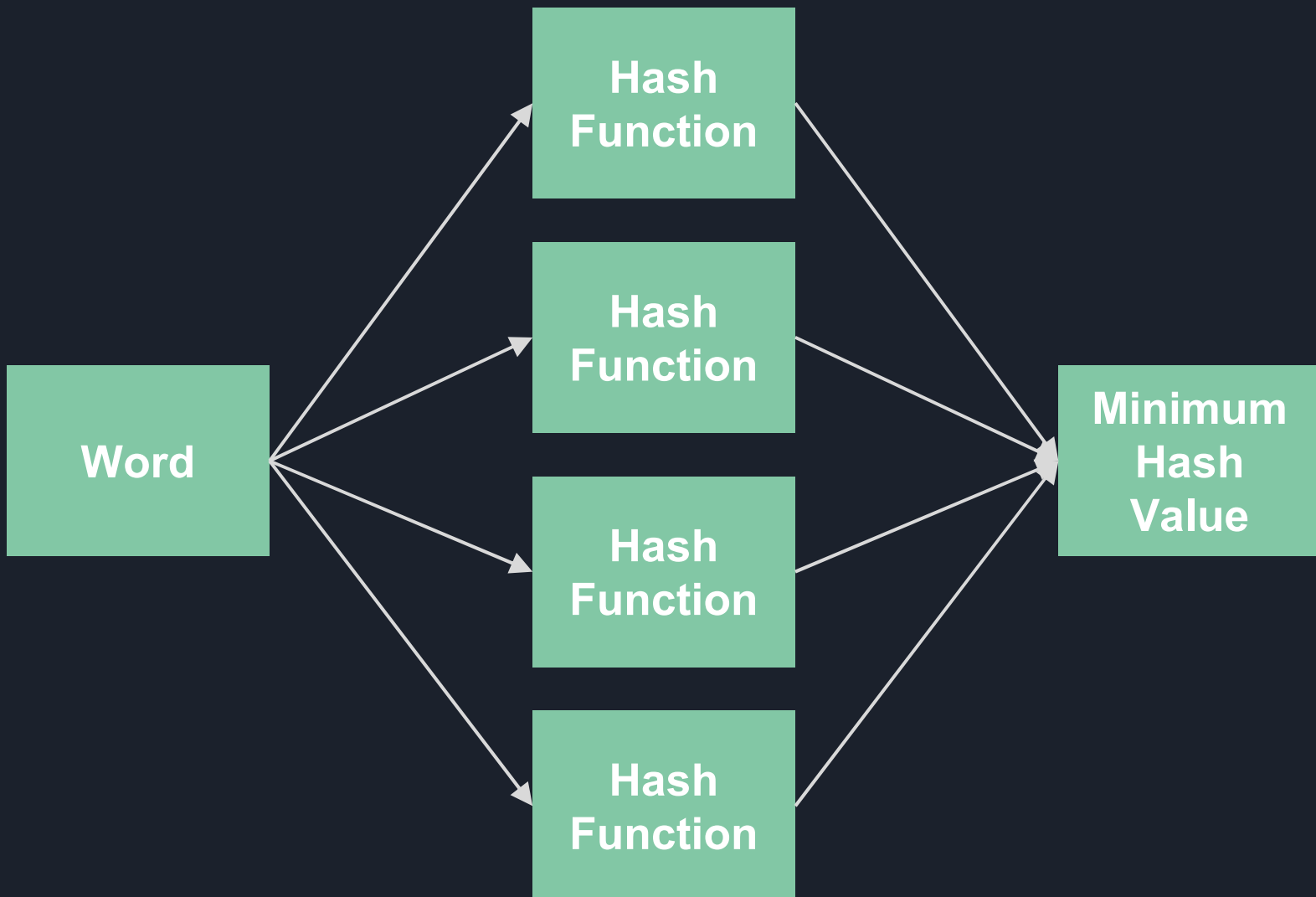


MinHash



LSH

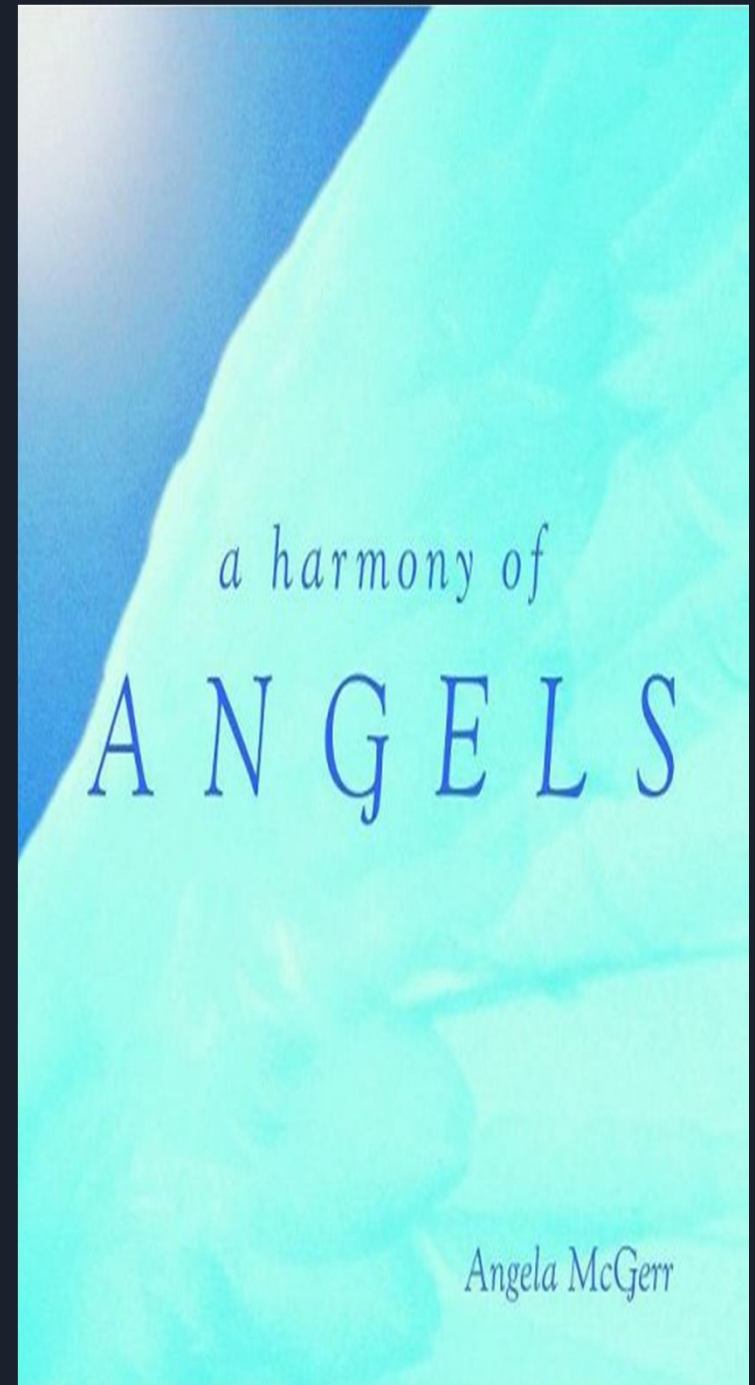
MinHash Function



Locality Sensitive Hashing

Item	Bucket 1	Bucket 2	Bucket 3	Bucket 4	Bucket 5
A	0	1	2	3	3
B	5	3	2	1	1
C	6	9	11	13	13
D	5	3	2	1	1
E	30	0	2	1	1

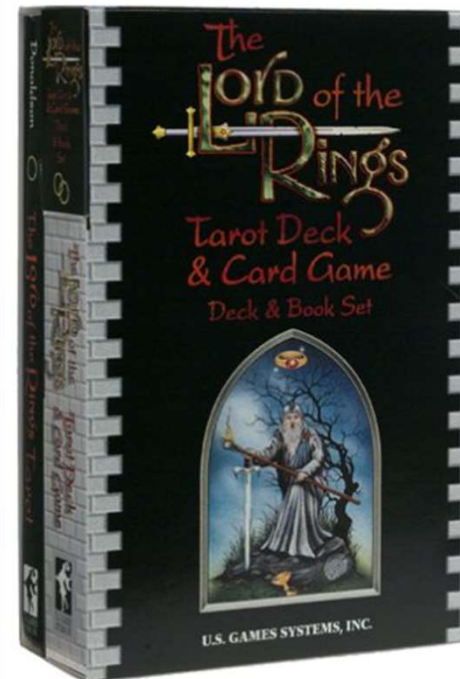
Product ID	682117
Title	A Harmony of Angels
Product Description	<p>"This is the loveliest book on angel work that I have ever seen." (Reader comment)</p> <p>"This breathtakingly beautiful book and card set brings you so many uses." (Reader comment)"</p> <p>I recommend it for any beginner on angels." (Reader comment)</p>
Price	None



Official NFL Pro Set Card Book

National Football League

Note: This is not the actual book cover



Noah's Angel Healing Energy

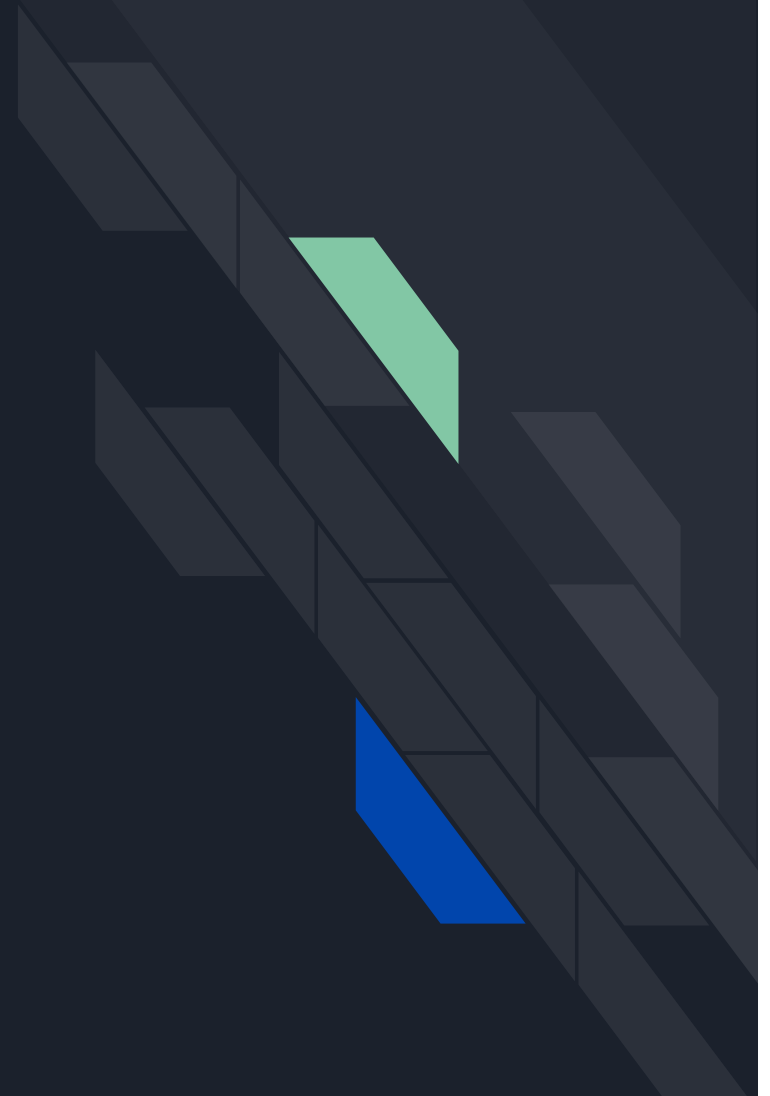


Jeanmarie Brenckle
Workbook for Students and Teachers
Copyrighted Material

N^2	LSH
Slow	Fast
Accurate	Approximate
Scales very poorly	Scales extremely well

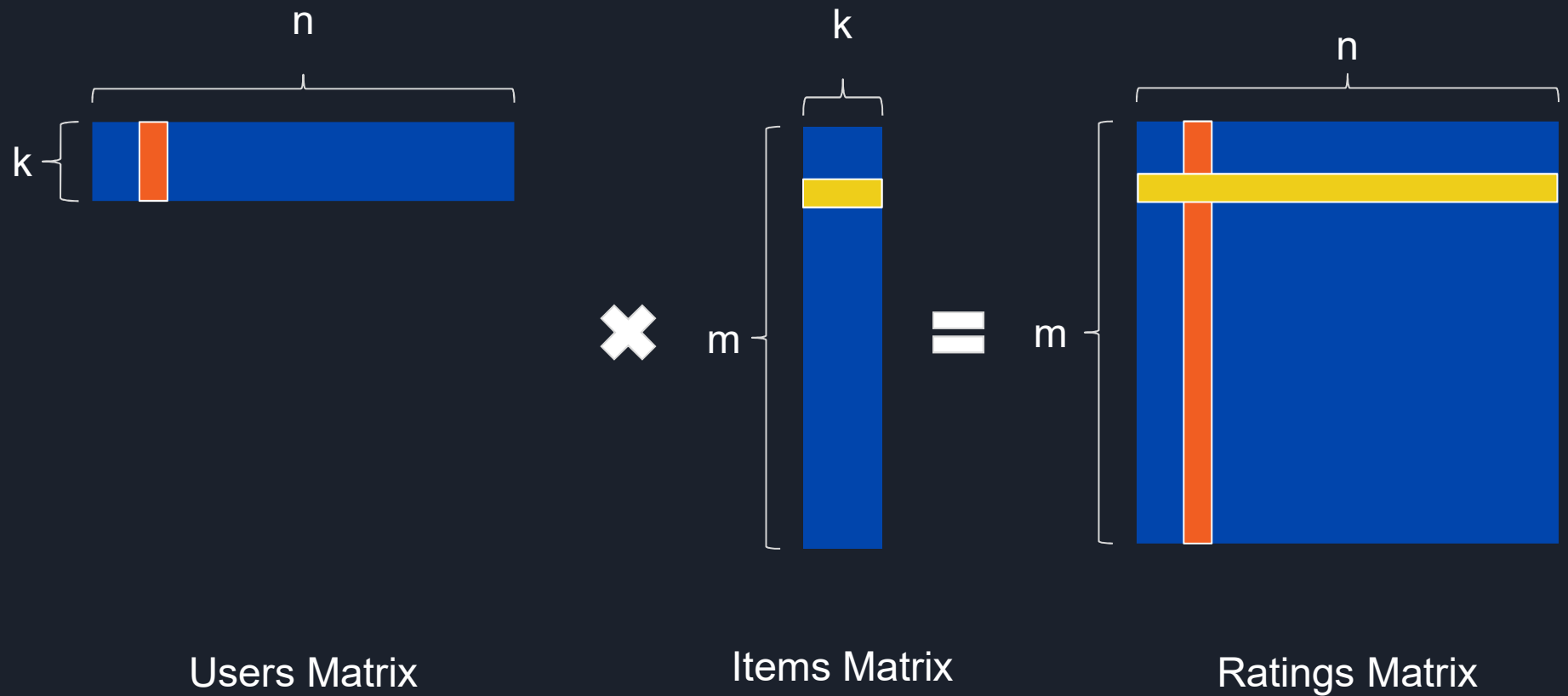
Collaborative Filtering Recommender

Supplemental Material



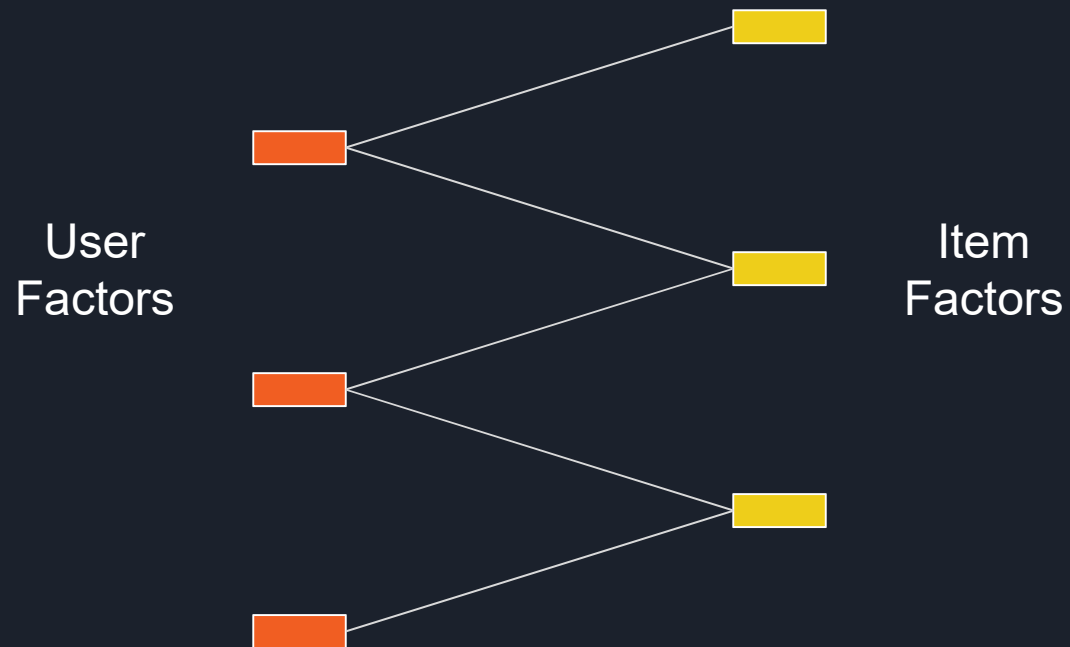
Low Rank Matrix Factorization

How it works



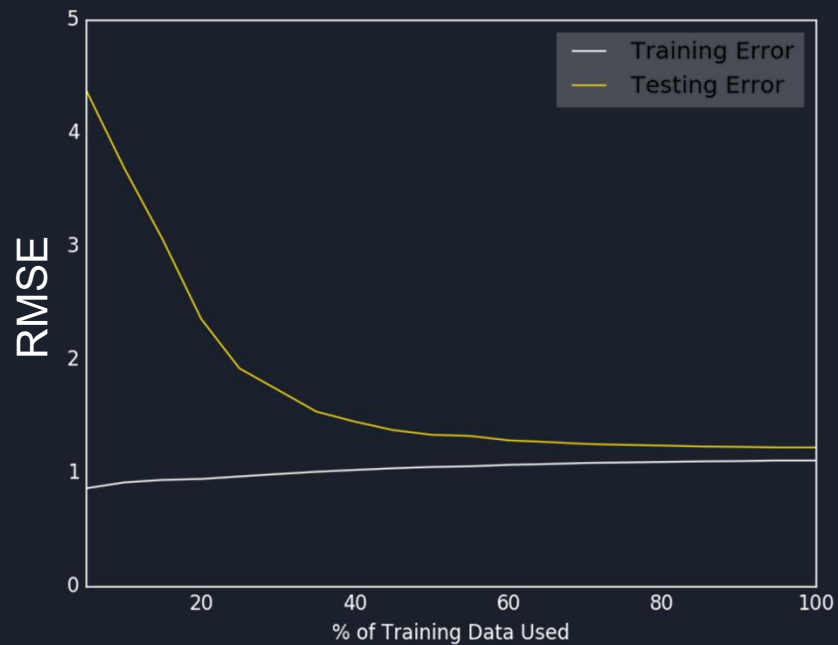
Alternating Least Squares

How it works

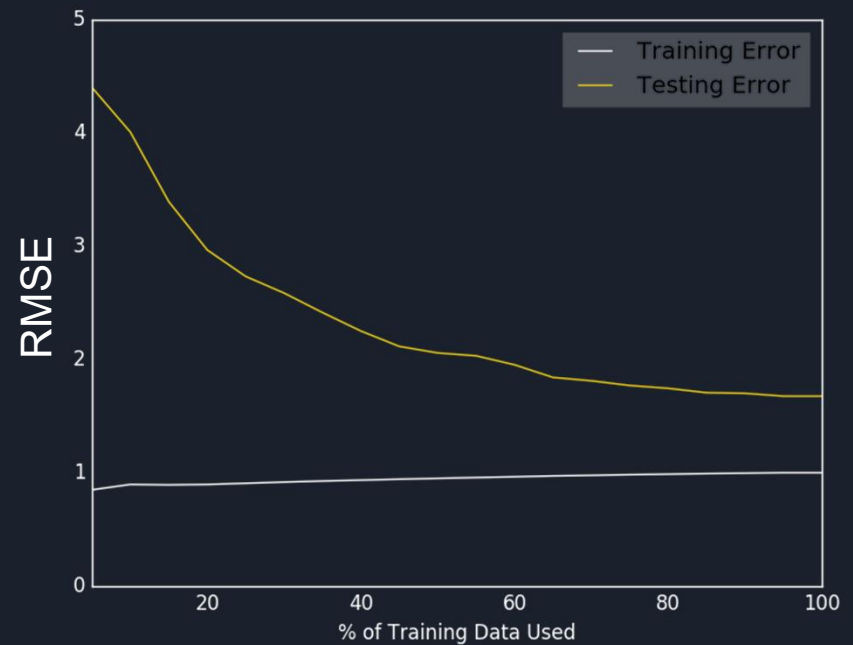


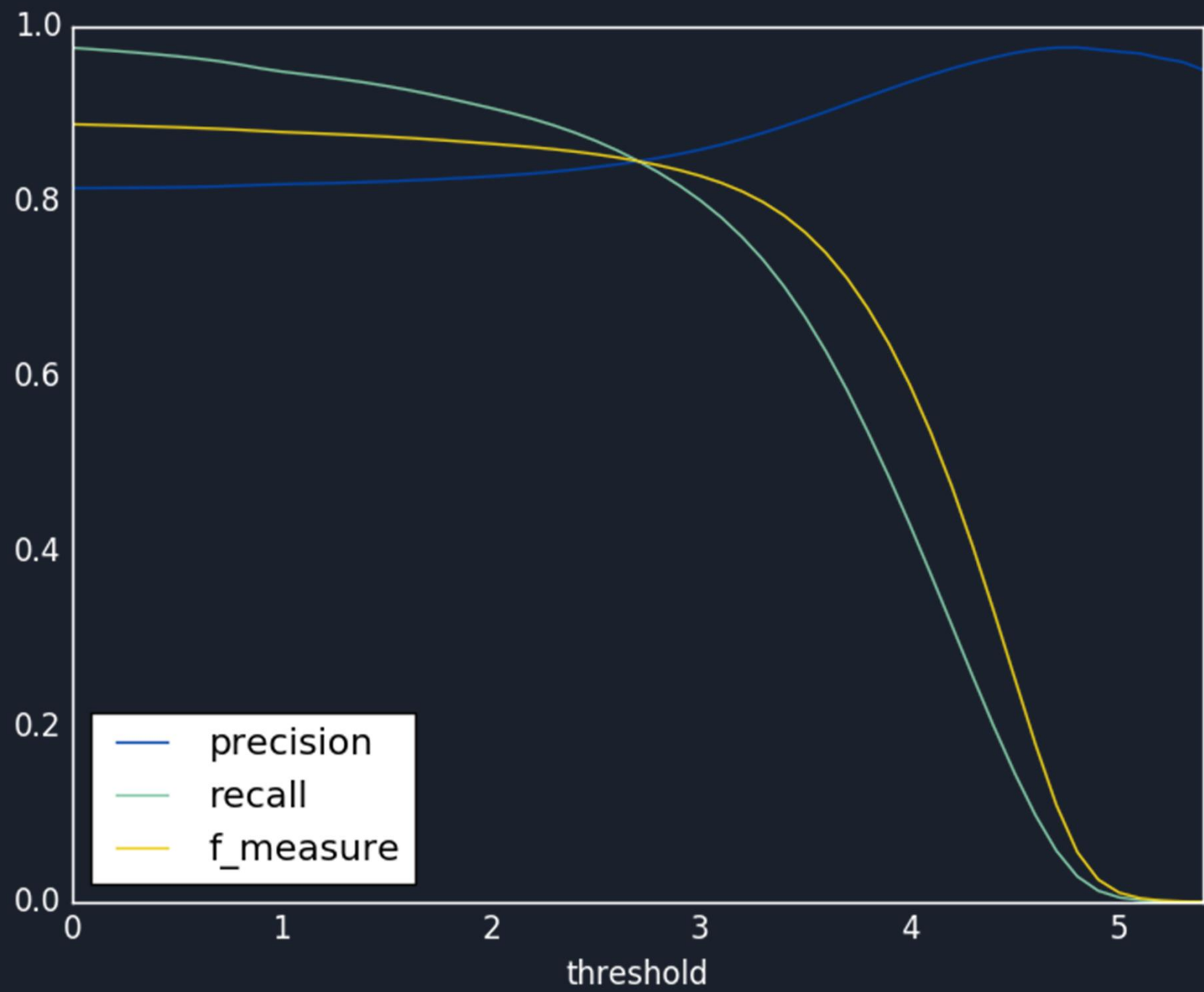
Collaborative Filtering Learning Curves

Reviews Only



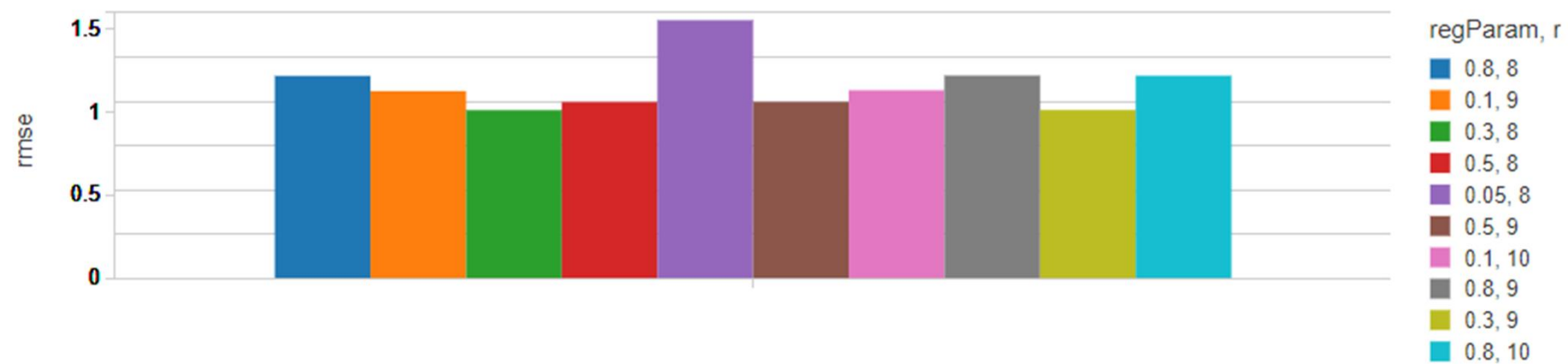
Ratings Only







Only showing the first ten series



Only showing the first ten series

Stochastic Gradient Descent

[Arberger \(2009\)](#) shows that SGD is generally faster and more accurate than ALS except in situations of extremely sparse data where ALS tends to perform better.

Stratified Stochastic Gradient Descent

How it works

