Recommender Systems

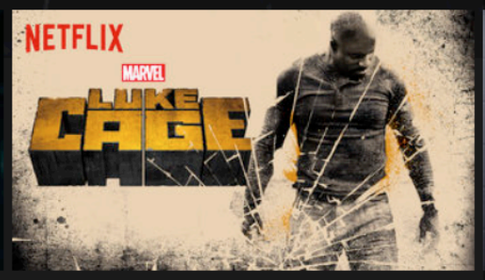
Ashwin Malshé

"There is an extensive class of Web applications that involve predicting user responses to options. Such a facility is called a recommendation system."

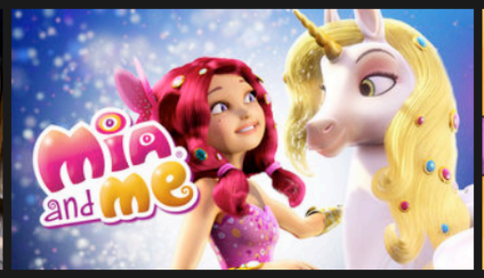
– Jute Leskovec, Stanford University,
Anand Rajaraman, Milliway Labs,
Jeffrey D. Ullman, Stanford University

Netflix

Because you watched Stranger Things



Trending Now

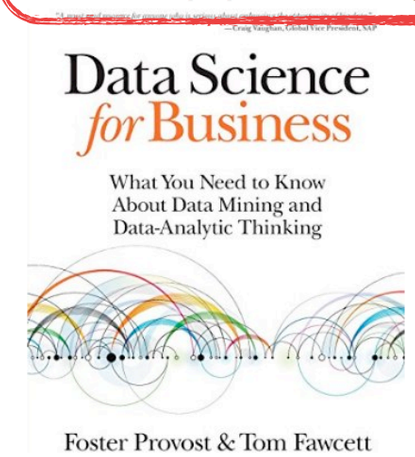


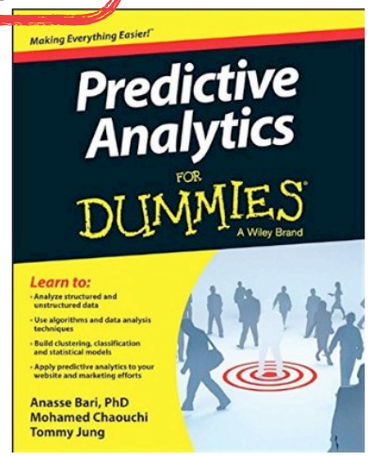
Because you watched Forrest Gump

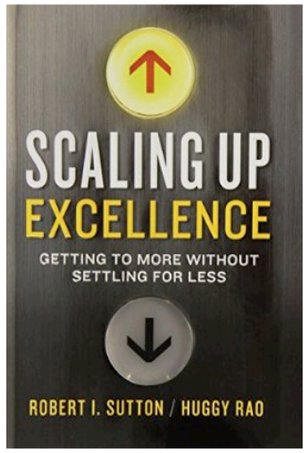
Amazon

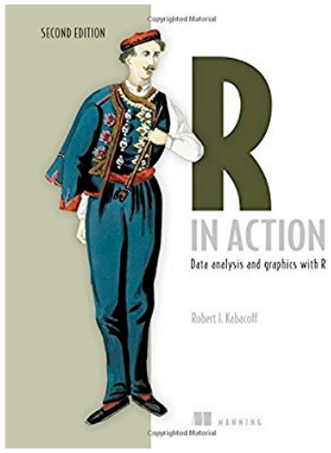
Inspired by your shopping trends

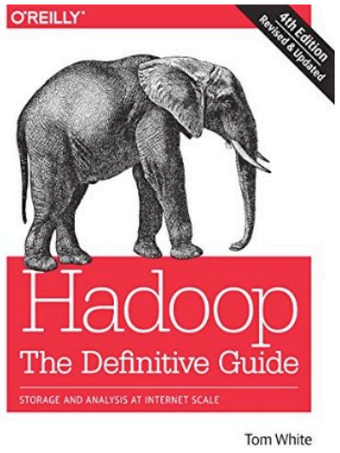
More top picks for you

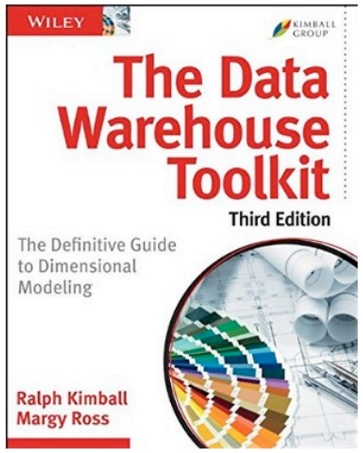


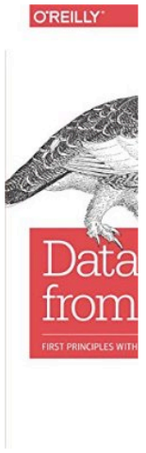




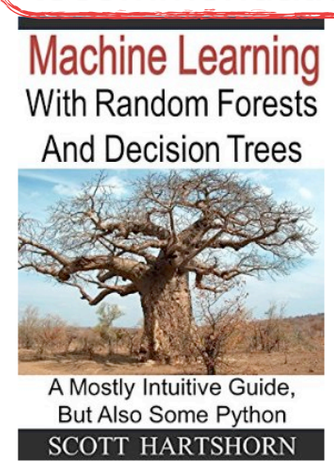


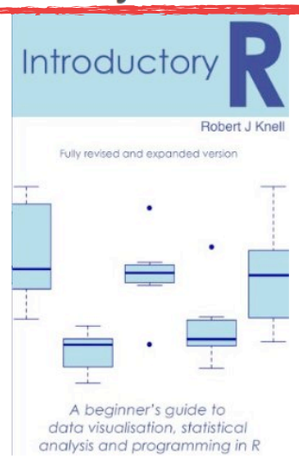


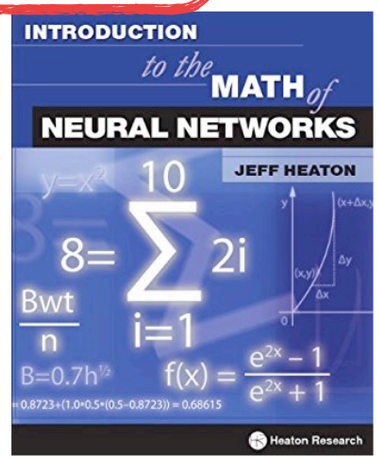


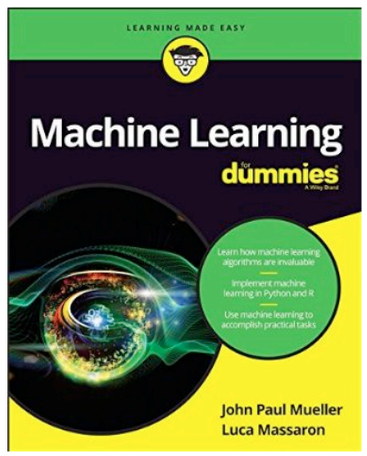


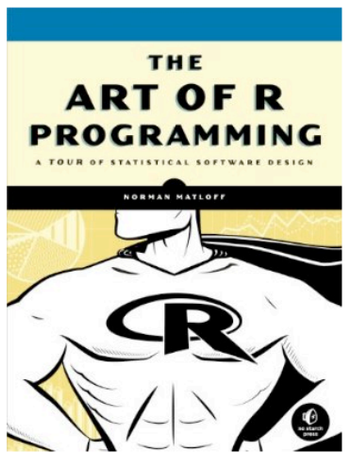
Recommendations for you in Kindle Store

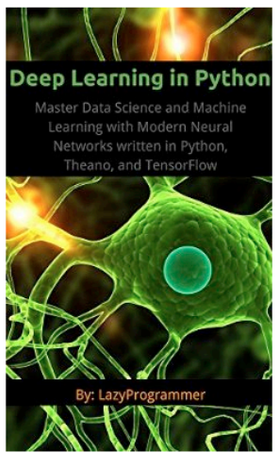






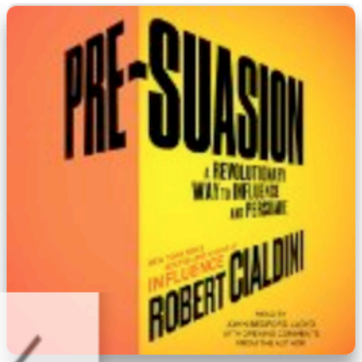


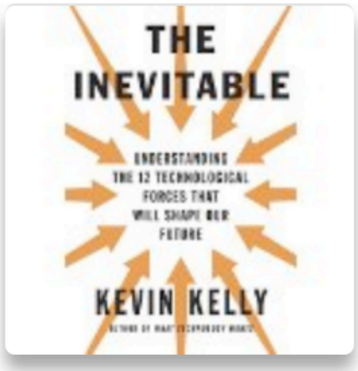


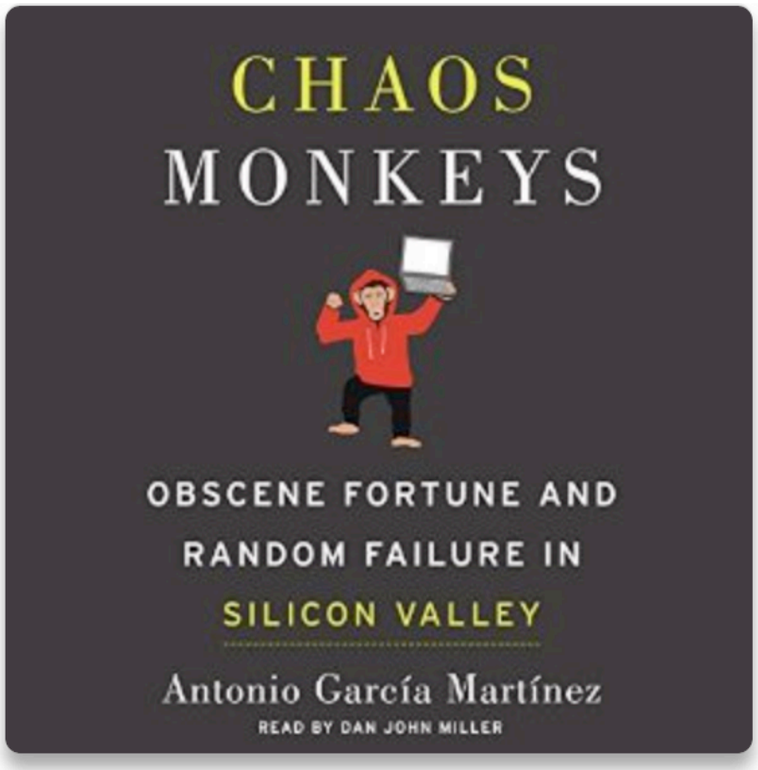


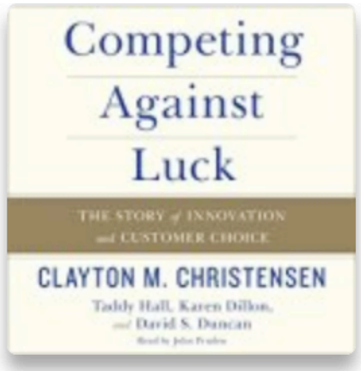
Audible

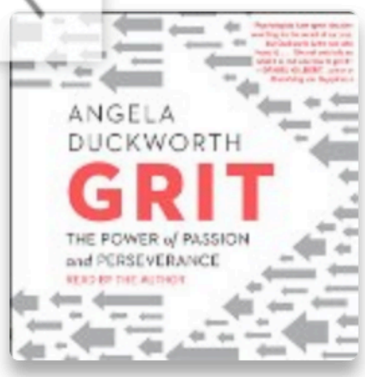
Based on Your Past Purchases, We Think You'll Enjoy

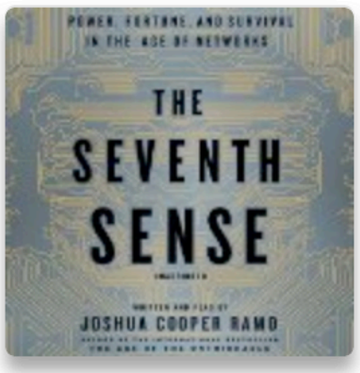


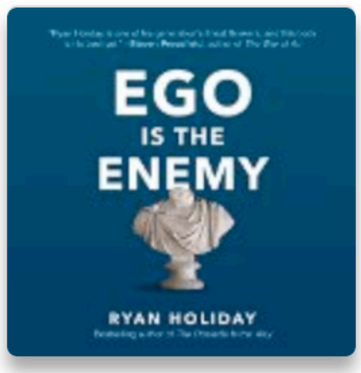


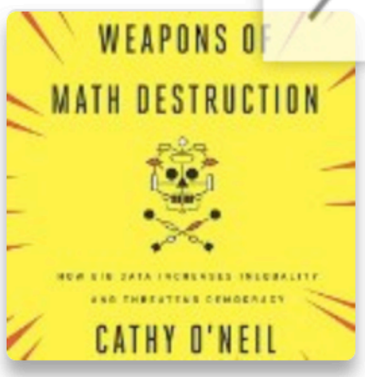




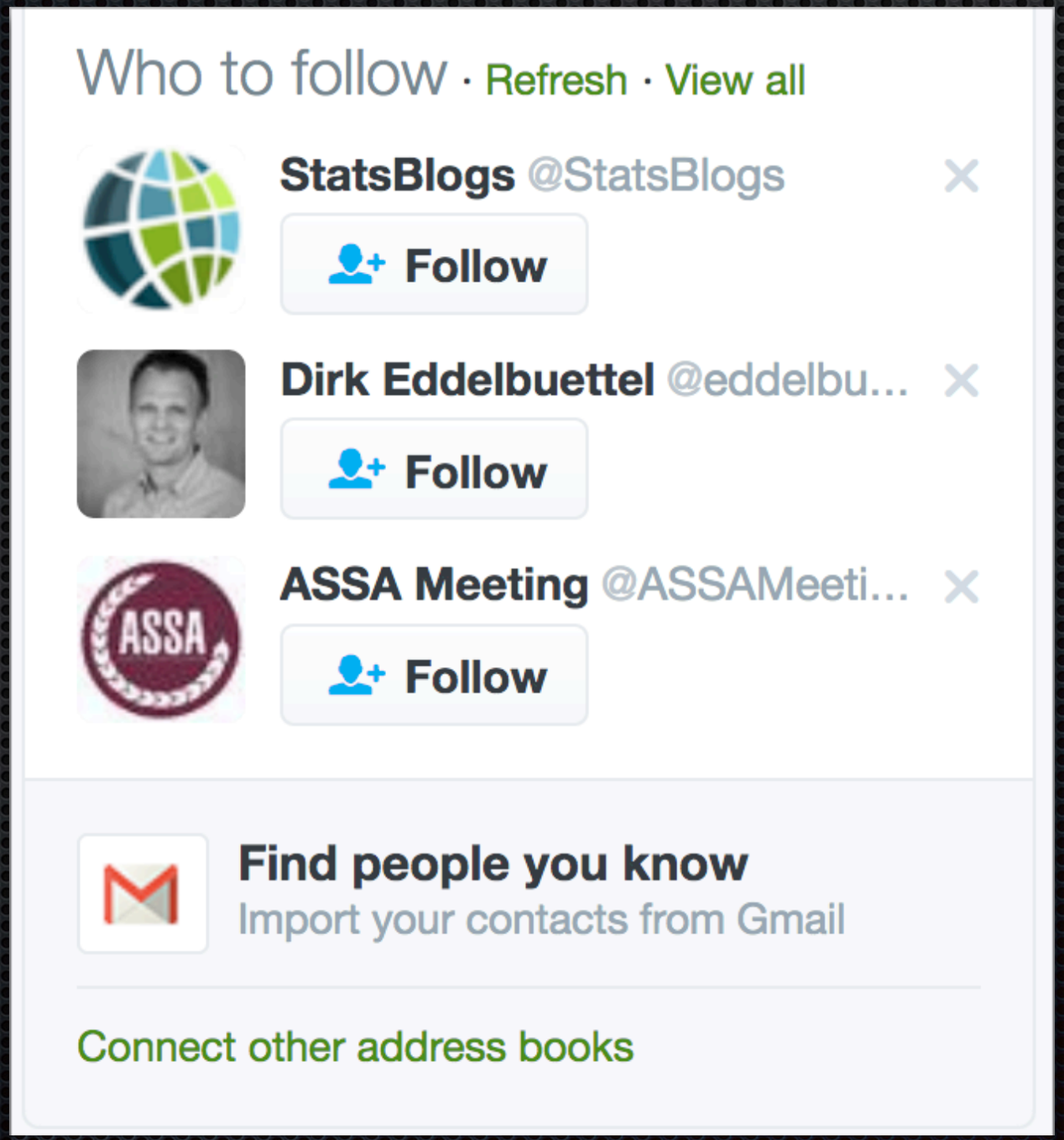




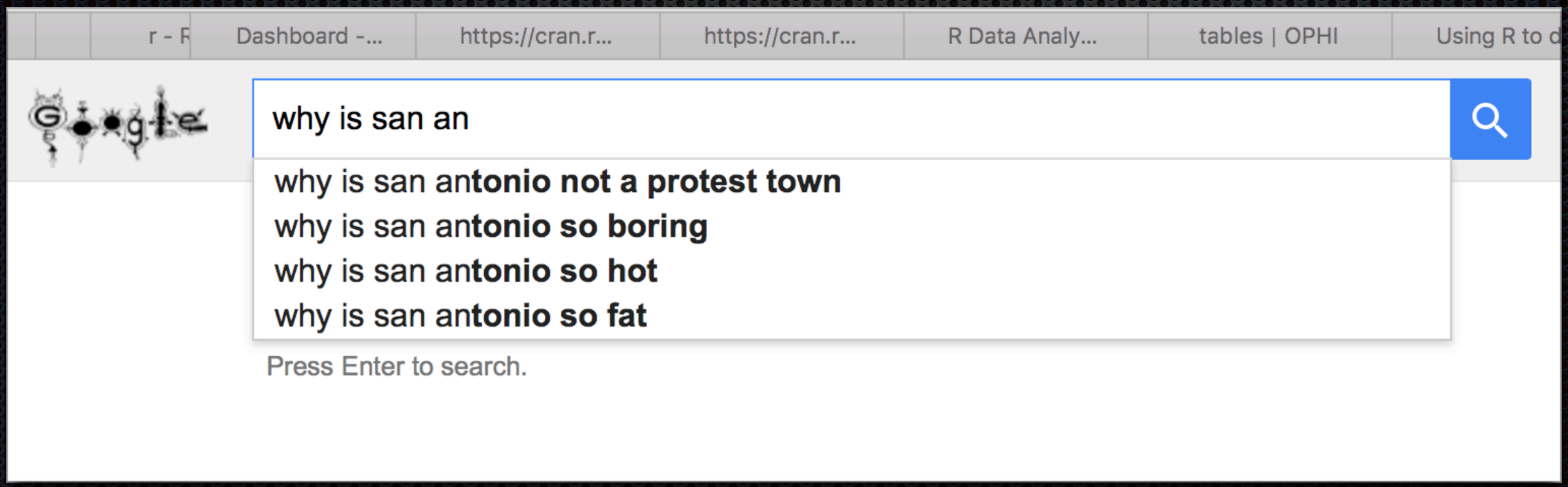




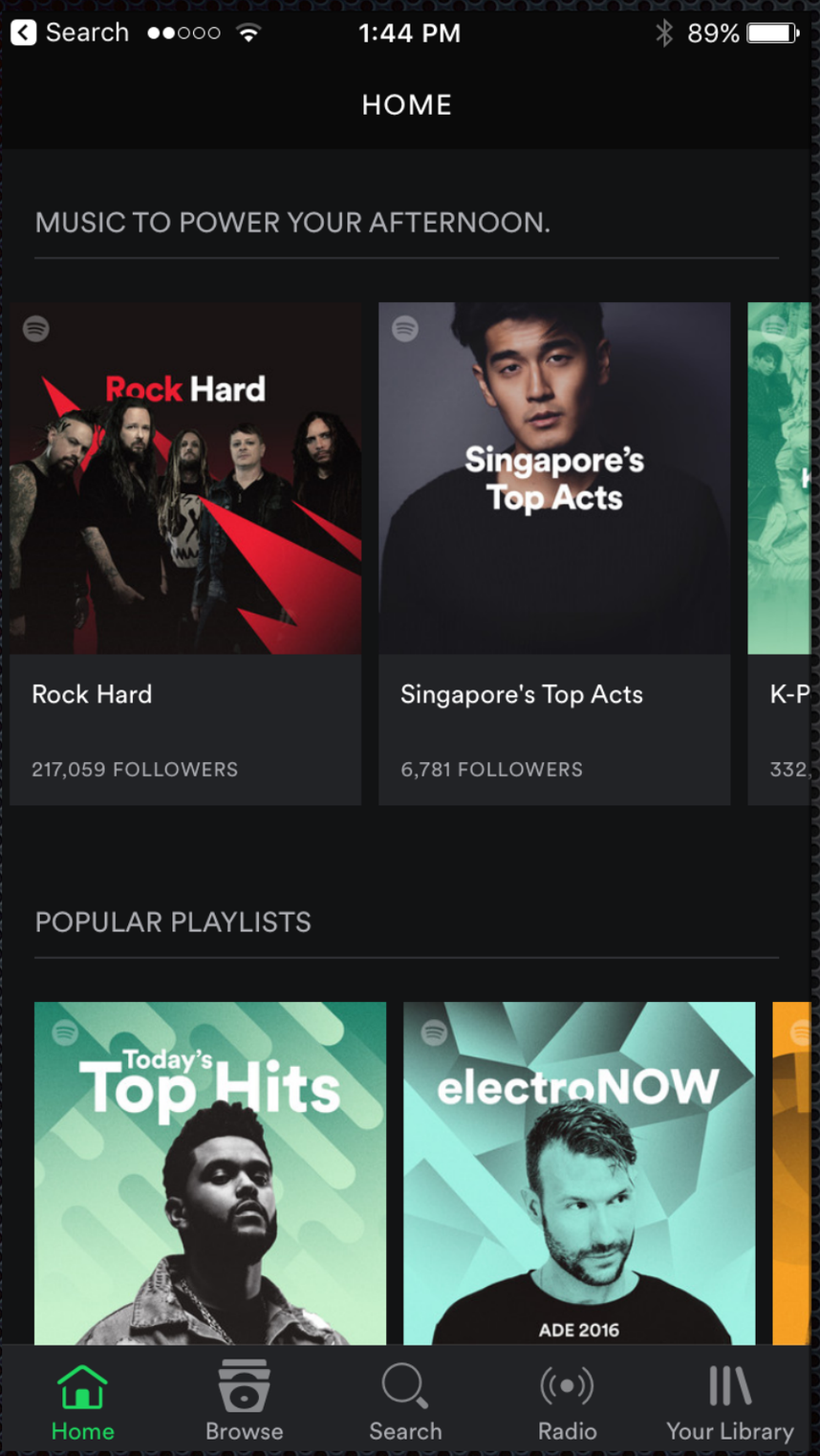
Twitter

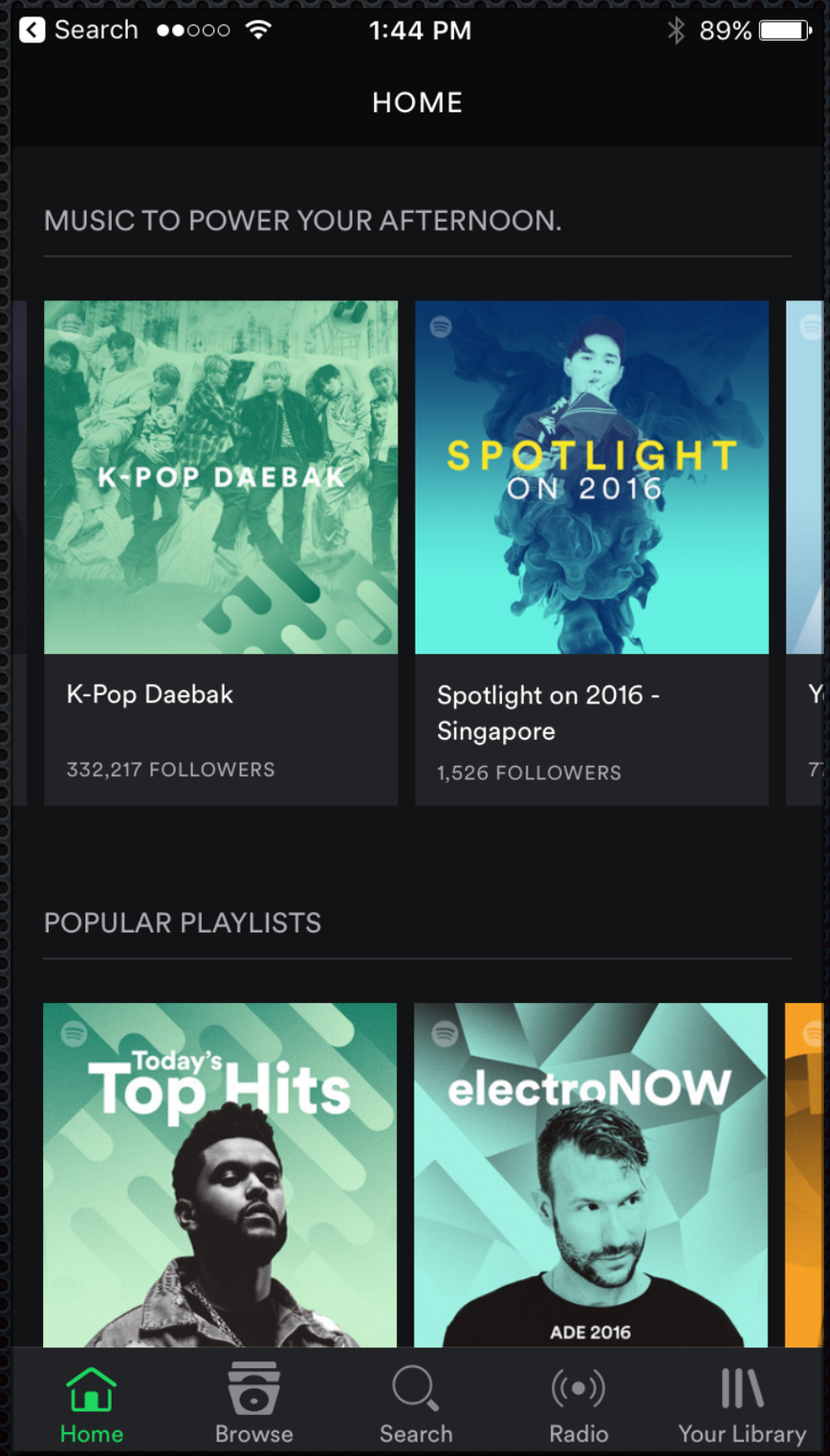

Who to follow · Refresh · View all
StatsBlogs @StatsBlogs
Follow
Dirk Eddelbuettel @eddelbu...
Follow
ASSA Meeting @ASSAMeeti...
Follow
Find people you know
Import your contacts from Gmail
Connect other address books

Google


r - R
Dashboard -...
https://cran.r...
https://cran.r...
R Data Analy...
tables | OPHI
Using R to d
why is san an
why is san antonio not a protest town
why is san antonio so boring
why is san antonio so hot
why is san antonio so fat
Press Enter to search.

Spotify


Search
1:44 PM
89%
HOME
MUSIC TO POWER YOUR AFTERNOON.
Rock Hard
Singapore's Top Acts
Rock Hard
217,059 FOLLOWERS
Singapore's Top Acts
6,781 FOLLOWERS
POPULAR PLAYLISTS
Today's Top Hits
electroNOW
ADE 2016
Home
Browse
Search
Radio
Your Library


Search
1:44 PM
89%
HOME
MUSIC TO POWER YOUR AFTERNOON.
K-POP DAEBAK
SPOTLIGHT ON 2016
K-Pop Daebak
332,217 FOLLOWERS
Spotlight on 2016 - Singapore
1,526 FOLLOWERS
POPULAR PLAYLISTS
Today's Top Hits
electroNOW
ADE 2016
Home
Browse
Search
Radio
Your Library

Technologies

- Content-based systems
  - Examine properties of the items recommended
  - Similarity of items is determined by measuring the similarity in their properties
- Collaborative filtering
  - Recommend items based on similarity measures between users and/or items
  - Similarity of items is determined by the similarity of the ratings of those items by the users who have rated both items

Our Focus

- Our focus is on collaborative filtering
  - The process of identifying similar users and recommending what similar users like is called collaborative filtering
- Content-based systems can be designed based on the item and user attributes. We have done these types of models before
  - e.g., Logistic regression, SVM

Mechanics

Utility Matrix

users/ items	i_1	i_2	i_3	…	i_n
u_1	r_{11}	r_{12}	r_{13}	…	r_{1n}
u_2	r_{21}	r_{22}	r_{23}	…	r_{2n}
u_3	r_{31}	r_{32}	r_{33}	…	r_{3n}
…	…	…	…	…	…
u_m	r_{m1}	r_{m2}	r_{m3}	…	r_{mn}

Utility Matrix is Sparse

users/ items	i_1	i_2	i_3	…	i_n
u_1	r_{11}		r_{13}	…	r_{1n}
u_2		r_{22}		…	
u_3	r_{31}			…	r_{3n}
…	…	…	…	…	…
u_m			r_{m3}	…	

Similarity Measures

- Jaccard distance
- Cosine similarity
- Pearson correlation

Example: Netflix Ratings

	M1	M2	M3	M4	M5	M6	M7
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

Jaccard Distance

- Applicable to yes/no ratings
- Ignores the ordinal nature of the rating
  - A and B have in common rated only one movie, M1. Collectively they have rated 5 movies.
    - This leads to Jaccard similarity = 1/5 and Jaccard distance = 4/5
  - A and C have in common rated 2 out of 4 movies so Jaccard similarity is 2/4 and Jaccard distance is 2/4

Cosine Distance

$$\cos\theta = \frac{A \cdot B}{\|A\| \times \|B\|}$$

Using this formula, cosine similarity between A and B is 0.38 and A and C is 0.322. Thus, A is more similar to B than C.

Pearson Correlation

- This is centered cosine similarity
- Center all the ratings row-wise
- Get cosine similarities

Types of Collaborative Filtering

- User-based collaborative filtering (UBCF)
  - UBCF is a memory-based algorithm which tries to mimic word-of-mouth by analyzing rating data from many individuals. The assumption is that users with similar preferences will rate items similarly.
- Item-based collaborative filtering (IBCF)
  - IBCF is a model-based approach, which produces recommendations based on the relationship between items inferred from the rating matrix

Clustering

- We can reduce the number of items and users by clustering
- Usually clustering will happen in sequence
  - Cluster items first then cluster users
  - Keep repeating until we have a reasonable number of clusters