HIGH QUALITY, PERSONALIZED

RECOMMENDATIONS

ARE THE HOLY GRAIL FOR EVERY ONLINE STORE

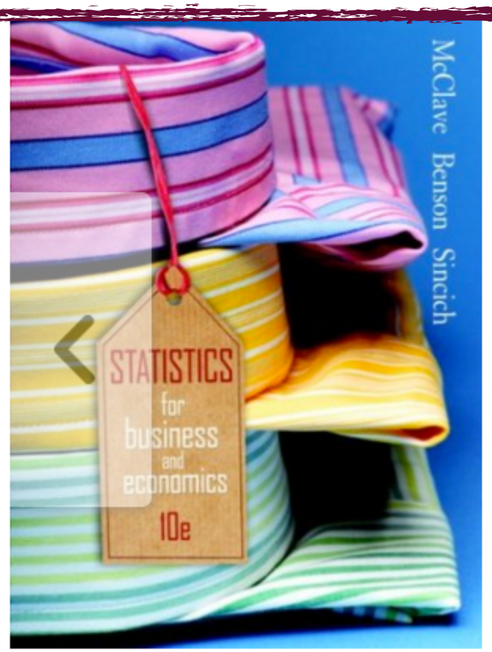
AMAZON

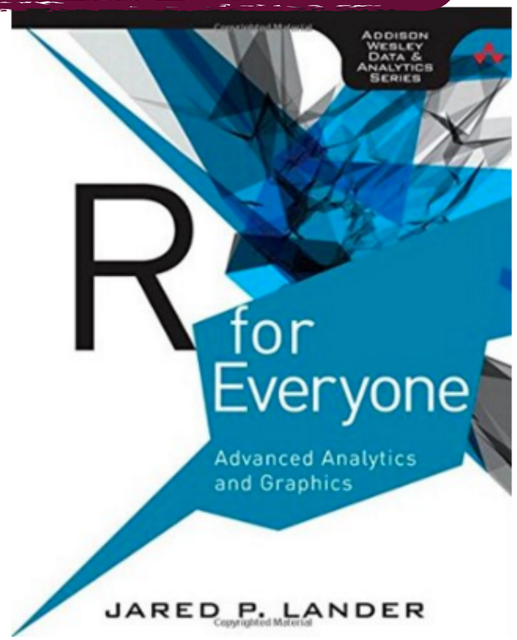
Related to Items You've Viewed See more

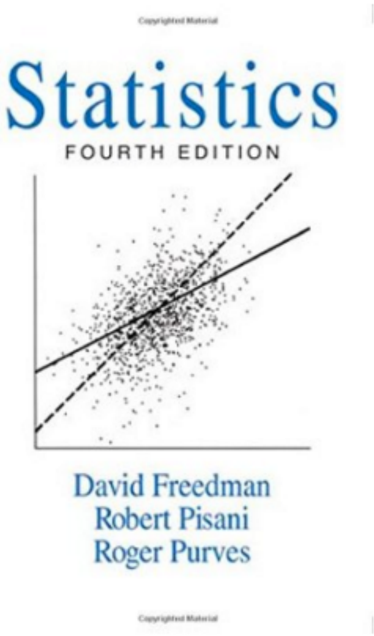
Inspired by Your Browsing History See more

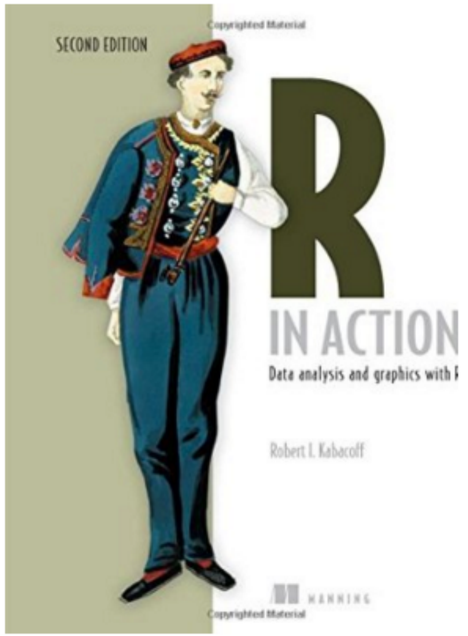
Related to Items You've Viewed

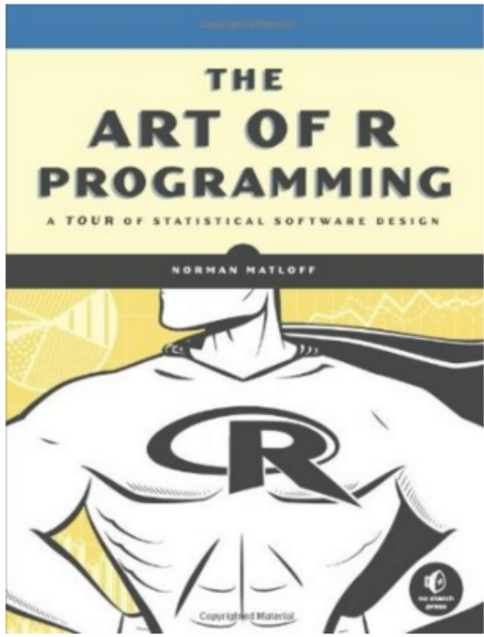
See more





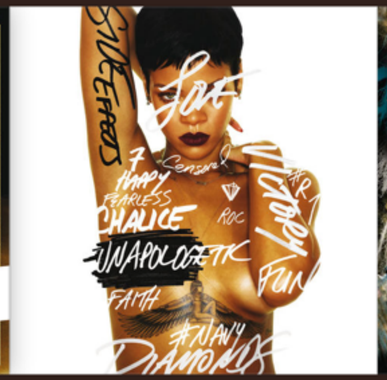






Inspired by Your Browsing History

See more



Zodiac: Pisces

Pisces are unstoppably creative, making for a serious and varied artistic legacy.

Playlist By Apple Music Hip-Hop/Rap

Intro to Alejandro Sanz

This Spanish singer has stolen hearts across the Latin world.

Playlist By Apple Music

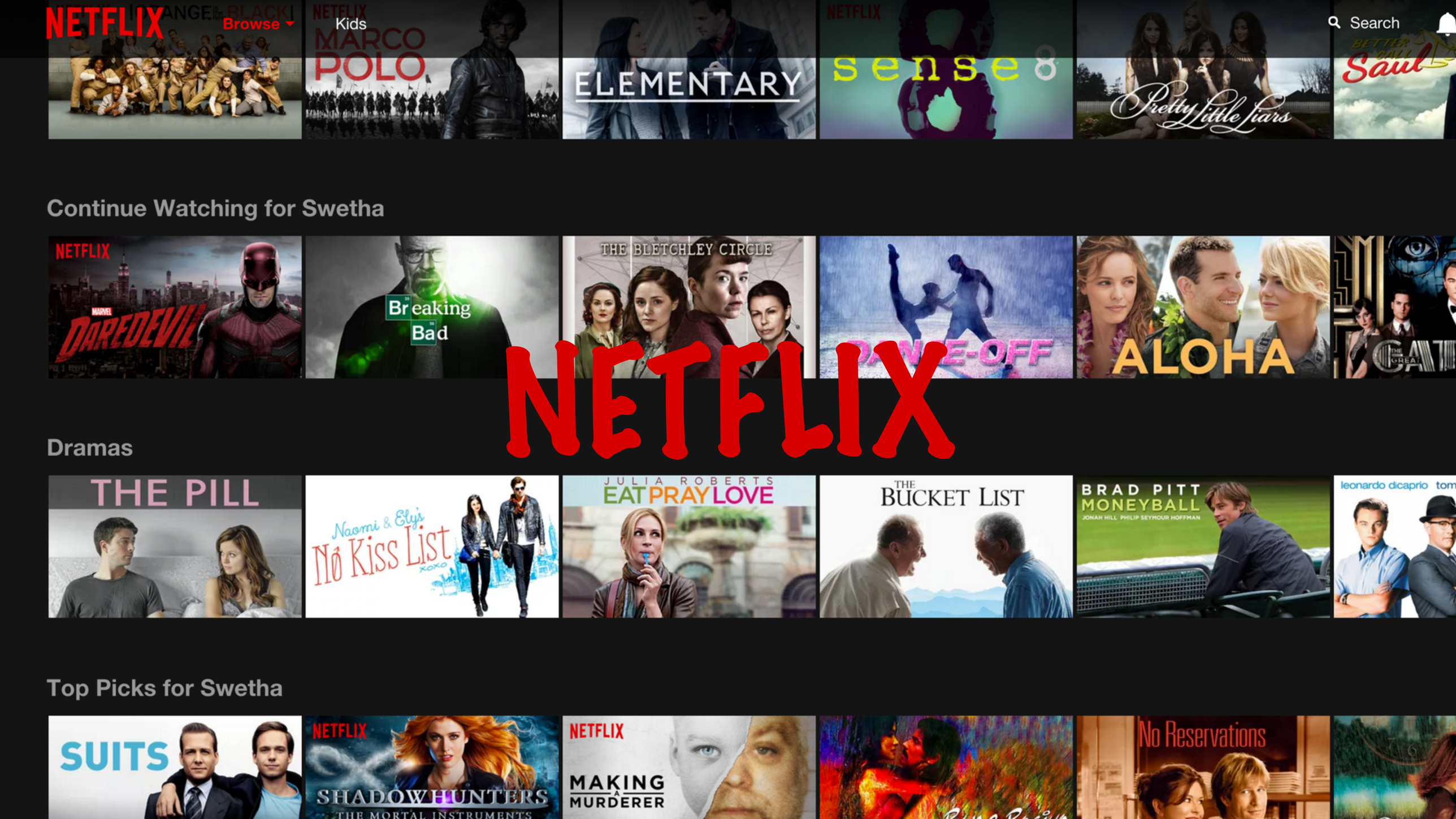

Freaks Only

Put on your freakum dress.

Playlist By Apple Music R&B

Accessing iTunes Store...

My Music Playlists For You New Radio Connect


NETFLIX
Browse
Kids
Search
MARCO POLO
ELEMENTARY
sense8
Pretty Little Liars
Saul
Continue Watching for Swetha
DAREDEVIL
Breaking Bad
THE BLETCHLEY CIRCLE
ALOHA
NETFLIX
Dramas
THE PILL
Naomi & Ely's No Kiss List
JULIA ROBERTS
EAT PRAY LOVE
THE BUCKET LIST
BRAD PITT
MONEYBALL
JONAH HILL PHILIP SEYMOUR HOFFMAN
leonardo dicaprio
Top Picks for Swetha
SUITS
SHADOWHUNTERS
THE MORTAL INSTRUMENTS
MAKING A MURDERER
No Reservations

Top Picks for Swetha

HIGH QUALITY, PERSONALIZED

RECOMMENDATIONS

ARE THE HOLY GRAIL FOR EVERY ONLINE STORE

HIGH QUALITY, PERSONALIZED RECOMMENDATIONS ARE THE HOLY GRAIL FOR EVERY ONLINE STORE

WHY?

HIGH QUALITY, PERSONALIZED **RECOMMENDATIONS**
ARE THE HOLY GRAIL FOR EVERY ONLINE STORE

WHY?

UNLIKE OFFLINE STORES

ONLINE STORES HAVE NO SALES PEOPLE

ONLINE STORES HAVE A **HUGE** NUMBER OF PRODUCTS

NETFLIX	AMAZON	ITUNES
THOUSANDS OF MOVIES	MILLIONS OF BOOKS	TENS OF THOUSANDS OF SONGS

HIGH QUALITY, PERSONALIZED RECOMMENDATIONS
ARE THE HOLY GRAIL FOR EVERY ONLINE STORE

WHY?

UNLIKE OFFLINE STORES

ONLINE STORES HAVE NO SALES PEOPLE
ONLINE STORES HAVE A HUGE NUMBER OF PRODUCTS

USERS ON THE OTHER HAND

HAVE LIMITED TIME AND PATIENCE
ARE NOT SURE WHAT THEY ARE LOOKING FOR

HIGH QUALITY, PERSONALIZED RECOMMENDATIONS ARE THE HOLY GRAIL FOR EVERY ONLINE STORE

UNLIKE OFFLINE STORES

ONLINE STORES HAVE NO SALES PEOPLE

ONLINE STORES HAVE A HUGE NUMBER OF PRODUCTS

WHY?

USERS ON THE OTHER HAND

HAVE LIMITED TIME AND PATIENCE

ARE NOT SURE WHAT THEY ARE LOOKING FOR

RECOMMENDATIONS HELP USERS

NAVIGATE THE MAZE OF ONLINE STORES

FIND WHAT THEY ARE LOOKING FOR

FIND THINGS THEY MIGHT LIKE, BUT DIDN'T KNOW OF

RECOMMENDATIONS HELP USERS
NAVIGATE THE MAZE OF ONLINE STORES
FIND WHAT THEY ARE LOOKING FOR
FIND THINGS THEY MIGHT LIKE, BUT DIDN'T KNOW OF

RECOMMENDATIONS HELP ONLINE STORES
SOLVE THE PROBLEM OF DISCOVERY

HOW?

RECOMMENDATIONS HELP USERS

NAVIGATE THE MAZE OF ONLINE STORES

FIND WHAT THEY ARE LOOKING FOR

FIND THINGS THEY MIGHT LIKE, BUT DIDN'T KNOW OF

RECOMMENDATIONS HELP ONLINE STORES

SOLVE THE PROBLEM OF DISCOVERY

HOW?

ONLINE STORES HAVE DATA

WHAT USERS BOUGHT

WHAT USERS CLICKED

WHAT USERS BROWSED

WHAT USERS RATED

RECOMMENDATIONS HELP USERS

NAVIGATE THE MAZE OF ONLINE STORES

FIND WHAT THEY ARE LOOKING FOR

FIND THINGS THEY MIGHT LIKE, BUT DIDN'T KNOW OF

RECOMMENDATIONS HELP ONLINE STORES

SOLVE THE PROBLEM OF DISCOVERY

HOW?

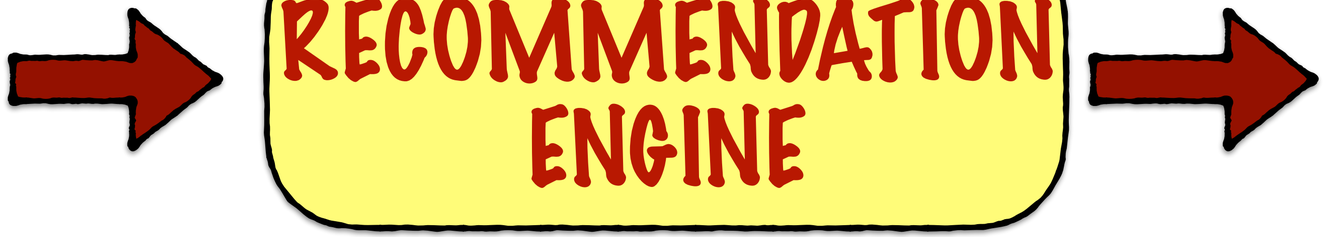
ONLINE STORES HAVE DATA

WHAT USERS BOUGHT

WHAT USERS BROWSED

WHAT USERS CLICKED

WHAT USERS RATED



TOP PICKS FOR YOU!!

IF YOU LIKE THIS, YOU'LL LOVE THAT!

IF YOU BUY THIS, YOU'LL NEED THAT!

RECOMMENDATION ENGINE

FILTER RELEVANT PRODUCTS

PREDICT WHAT RATING THE USER WOULD GIVE A PRODUCT

PREDICT WHETHER A USER WOULD BUY A PRODUCT

RANK PRODUCTS BASED ON THEIR RELEVANCE TO THE USER

SUBTASKS

TASKS PERFORMED BY RECOMMENDATION ENGINES

FILTER RELEVANT PRODUCTS

"SIMILAR" TO THE ONES THE USER "LIKED"

"LIKED" BY "SIMILAR" USERS

PURCHASED ALONG WITH THE ONES THE USER "LIKED"

BEFORE WE GO AHEAD, LET'S JUST CLARIFY -

HOW DO WE KNOW THAT A USER LIKES A PRODUCT?

THESE ARE PRODUCTS THE USER

PURCHASED

CLICKED ON

ADDED TO CART

RATED HIGHLY

SOMETIMES, STORES ALSO ASK FOR EXPLICIT INPUT FROM USERS

(IF THE USER RATES SOMETHING LOW, THAT'S IMPORTANT DATA TOO!)

FILTER RELEVANT PRODUCTS

"SIMILAR" TO THE ONES THE USER "LIKED"

JANANI LOOKED FOR A DSLR CAMERA FROM NIKON

LET'S SHOW HER OTHER DSLR CAMERAS FROM NIKON

"LIKED" BY "SIMILAR" USERS

VITTHAL LIKES BOOKS ABOUT TECHNOLOGY

LET'S SHOW HIM THE BESTSELLERS IN TECHNOLOGY

PURCHASED ALONG WITH THE ONES THE USER "LIKED"

FILTER RELEVANT PRODUCTS

CONTENT-BASED FILTERING

"SIMILAR" TO THE ONES THE USER "LIKED"

"LIKED" BY "SIMILAR" USERS

PURCHASED ALONG WITH THE ONES THE USER "LIKED"

FIND PRODUCTS BY MATCHING

ATTRIBUTES

DESCRIPTIVE CHARACTERISTICS

CONTENT

FILTER RELEVANT PRODUCTS

CONTENT-BASED FILTERING

"COLLABORATIVE FILTERING"

PURCHASED ALONG WITH THE ONES THE USER "LIKED"

VITTHAL LIKED "STEVE JOBS" AND "OUTLIERS"

JANANI ALSO LIKED "STEVE JOBS" AND "OUTLIERS"

JANANI LIKES "ZERO TO ONE"

LET'S SHOW VITTHAL "ZERO TO ONE"

FILTER RELEVANT PRODUCTS

CONTENT-BASED FILTERING

COLLABORATIVE FILTERING

"PURCHASED ALONG WITH THE ONES THE USER LIKED"

ASSOCIATION RULES

NAVDEEP BOUGHT A DSLR CAMERA

90% OF FOLKS WHO BOUGHT A DSLR ALSO BOUGHT RECHARGEABLE BATTERIES

LET'S SHOW NAVDEEP RECHARGEABLE BATTERIES

RECOMMENDATION ENGINES NORMALLY USE ONE OR MORE OF THESE TECHNIQUES

CONTENT-BASED FILTERING

COLLABORATIVE FILTERING

ASSOCIATION RULES

CONTENT-BASED FILTERING

IS NORMALLY USED WITH TEXT DOCUMENTS (BOOKS, ARTICLES ETC)

PRODUCTS ARE REPRESENTED IN TERMS OF DESCRIPTORS OR ATTRIBUTES

GENRE

(SCIENCE FICTION/ COMEDY/ DRAMA)

WORDS USED

AUTHOR

A USER PROFILE IS CREATED USING THE SAME TERMS

BASED ON USER'S HISTORY

BY ASKING THE USER WHAT THEY LIKE

FIND PRODUCTS WHICH MATCH THE USER'S PROFILE

CONTENT-BASED FILTERING

PRODUCTS ARE REPRESENTED IN TERMS OF DESCRIPTORS OR ATTRIBUTES

COMMERCIAL

USERS ARE REPRESENTED USING THE SAME DESCRIPTORS

JANANI CLICKED ON 10 MOVIES
7 COMMERCIAL, 9 DRAMA

DRAMA

COMEDY

ARTY

CONTENT-BASED FILTERING

PRODUCTS ARE REPRESENTED IN TERMS
OF DESCRIPTORS OR ATTRIBUTES

THIS IS THE KEY CHALLENGE IN CONTENT BASED FILTERING

WHAT ATTRIBUTES SHOULD WE USE?
HOW ARE THESE GENERATED?

TYPICALLY, YOU'LL NEED MANUAL
DATA COLLECTION

FOR TEXT DOCUMENTS, YOU COULD USE NLP
TO GENERATE DESCRIPTORS

RECOMMENDATION ENGINES THAT USE CONTENT-BASE
FILTERING ALONE ARE LESS COMMON

CONTENT-BASED FILTERING

THE MOST SUCCESSFUL EXAMPLE OF CONTENT-BASED FILTERING IS

THE MUSIC GENOME PROJECT

OWNED BY PANDORA RADIO

EVERY SONG IS REPRESENTED BY A VECTOR OF 450 "GENES"

TRAINED MUSICAL ANALYSTS SCORE EACH SONG ON THESE 450 ATTRIBUTES

THE PROCESS TAKES 20-30 MINS PER SONG!!!

CONTENT-BASED FILTERING

THE MOST SUCCESSFUL EXAMPLE OF CONTENT-BASED FILTERING IS

THE MUSIC GENOME PROJECT

OWNED BY PANDORA RADIO

EVERY SONG IS REPRESENTED BY A VECTOR OF 450 "GENES"

THE RADIO THEN KEEPS PLAYING SONGS THAT MATCH THE USER'S PREFERENCES

RECOMMENDATION ENGINES NORMALLY USE ONE OR MORE OF THESE TECHNIQUES

CONTENT-BASED FILTERING

COLLABORATIVE FILTERING

ASSOCIATION RULES

RECOMMENDATION ENGINES NORMALLY USE ONE OR MORE OF THESE TECHNIQUES

CONTENT-BASED FILTERING

COLLABORATIVE FILTERING

ASSOCIATION RULES

COLLABORATIVE FILTERING

WHAT IF YOU COULD RECOMMEND PRODUCTS WITHOUT KNOWING ANYTHING ABOUT THE PRODUCTS THEMSELVES?

UNLIKE CONTENT-BASED FILTERING, COLLABORATIVE FILTERING DOESN'T REQUIRE ANY PRODUCT DESCRIPTION DATA AT ALL!

HOW DOES THAT WORK?

COLLABORATIVE FILTERING

HOW DOES THAT WORK?

HOW DO YOU NORMALLY FIND

A MOVIE TO WATCH?

A RESTAURANT TO GO TO?

AN ARTIST TO CHECK OUT?

A BOOK TO READ?

SOMEONE WHO LIKES

[redacted]

COLLABORATIVE FILTERING

HOW DOES THAT WORK?

THE BASIC PREMISE IS THAT

IF 2 USERS HAVE THE SAME OPINION
ABOUT A BUNCH OF PRODUCTS

THEY ARE LIKELY TO HAVE THE SAME
OPINION ABOUT OTHER PRODUCTS TOO

THE BASIC PREMISE IS THAT

IF 2 USERS HAVE THE SAME OPINION ABOUT A BUNCH OF PRODUCTS

THEY ARE LIKELY TO HAVE THE SAME OPINION ABOUT OTHER PRODUCTS TOO

COLLABORATIVE FILTERING IS A GENERAL TERM

FOR ANY ALGORITHM THAT RELIES ONLY ON USER BEHAVIOR (HISTORY, RATINGS, SIMILAR USERS ETC)

THE ALGORITHM NORMALLY PREDICTS USERS' RATINGS FOR PRODUCTS THEY HAVEN'T YET RATED

COLLABORATIVE FILTERING IS A GENERAL TERM

FOR ANY ALGORITHM THAT RELIES ONLY ON USER BEHAVIOR (HISTORY, RATINGS, SIMILAR USERS ETC)

THE ALGORITHM NORMALLY PREDICTS USERS' RATINGS FOR PRODUCTS THEY HAVEN'T YET RATED

IT CAN ALSO PREDICT WHETHER A USER WILL BUY A PRODUCT

RATING HERE IS A GENERAL TERM

IT CAN MEAN THAT A USER LIKES A PRODUCT

EXPLICIT RATING

NETFLIX ASKS USERS TO RATE A MOVIE ONCE THEY HAVE WATCHED IT

IMPLICIT RATING

CLICKS, # PURCHASES, # SEARCHES

COLLABORATIVE FILTERING IS A GENERAL TERM

FOR ANY ALGORITHM THAT RELIES ONLY ON USER BEHAVIOR (HISTORY, RATINGS, SIMILAR USERS ETC)

THE ALGORITHM NORMALLY PREDICTS USERS' RATINGS FOR PRODUCTS THEY HAVEN'T YET RATED

THERE ARE MANY MANY DIFFERENT ALGORITHMS TO PERFORM COLLABORATIVE FILTERING

2 POPULAR TECHNIQUES ARE

NEAREST NEIGHBOUR BASED METHODS

LATENT FACTOR BASED METHODS

NEAREST NEIGHBOUR BASED METHODS

ACTIVE USER

THE OBJECTIVE IS TO PREDICT A USER'S RATING FOR A PRODUCT THEY HAVEN'T RATED YET

FIND THE K-NEAREST NEIGHBOURS OF THAT USER AND TAKE A WEIGHTED AVERAGE OF THEIR RATING

K-NEAREST NEIGHBOURS

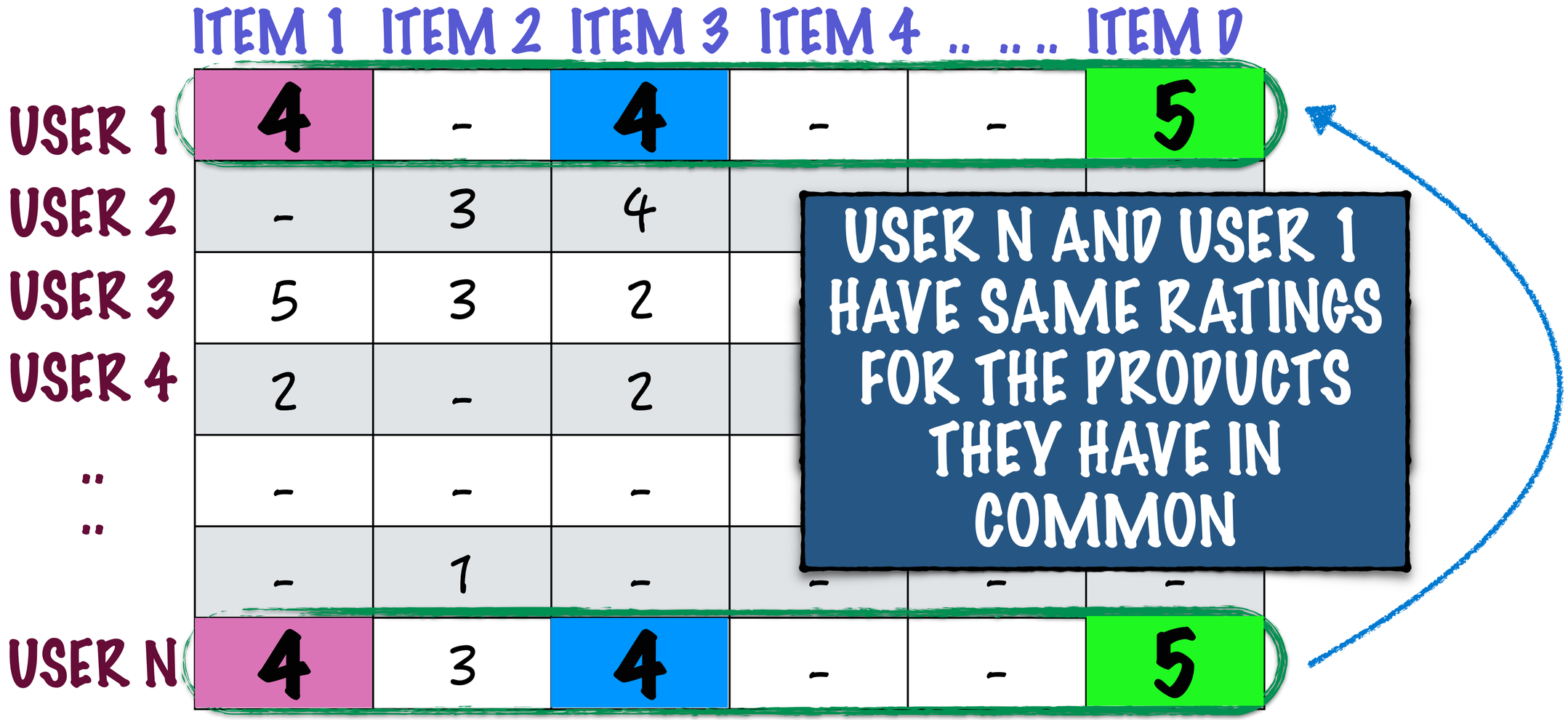
USERS WHO GIVE THE SAME RATINGS FOR A BUNCH OF PRODUCTS ARE "SIMILAR" TO EACH OTHER

THE K-NEAREST NEIGHBOURS ARE THE USERS WHO ARE "MOST SIMILAR" TO THE ACTIVE USER

REPRESENT USERS BY THEIR RATINGS FOR DIFFERENT PRODUCTS

K-NEAREST NEIGHBOURS

REPRESENT USERS BY THEIR RATINGS FOR DIFFERENT PRODUCTS



K-NEAREST NEIGHBOURS

REPRESENT USERS BY THEIR RATINGS FOR DIFFERENT PRODUCTS

	ITEM 1	ITEM 2	ITEM 3	ITEM 4	..	.. ..	ITEM D
USER 1	4	-	4	-	-	-	
USER 2	-	3	4	-	-	-	
USER 3	5	3	2	-	-	5	
USER 4	2	-	2	-	-	4	
.. ..	-	-	-	4	-	-	
	-	1	-	-	-	-	
USER N	4	3	4	-	-	5	

NEAREST NEIGHBOURS ARE FOUND USING A

SIMILARITY (OR) DISTANCE METRIC

EUCLIDEAN DISTANCE

COSINE SIMILARITY

PEARSON CORRELATION

THIS SAME METRIC CAN BE USED FOR THE WEIGHT OF EACH NEIGHBOUR IN THE PREDICTED RATING

K-NEAREST NEIGHBOURS

NEAREST NEIGHBOURS ARE FOUND USING A

SIMILARITY (OR) DISTANCE METRIC

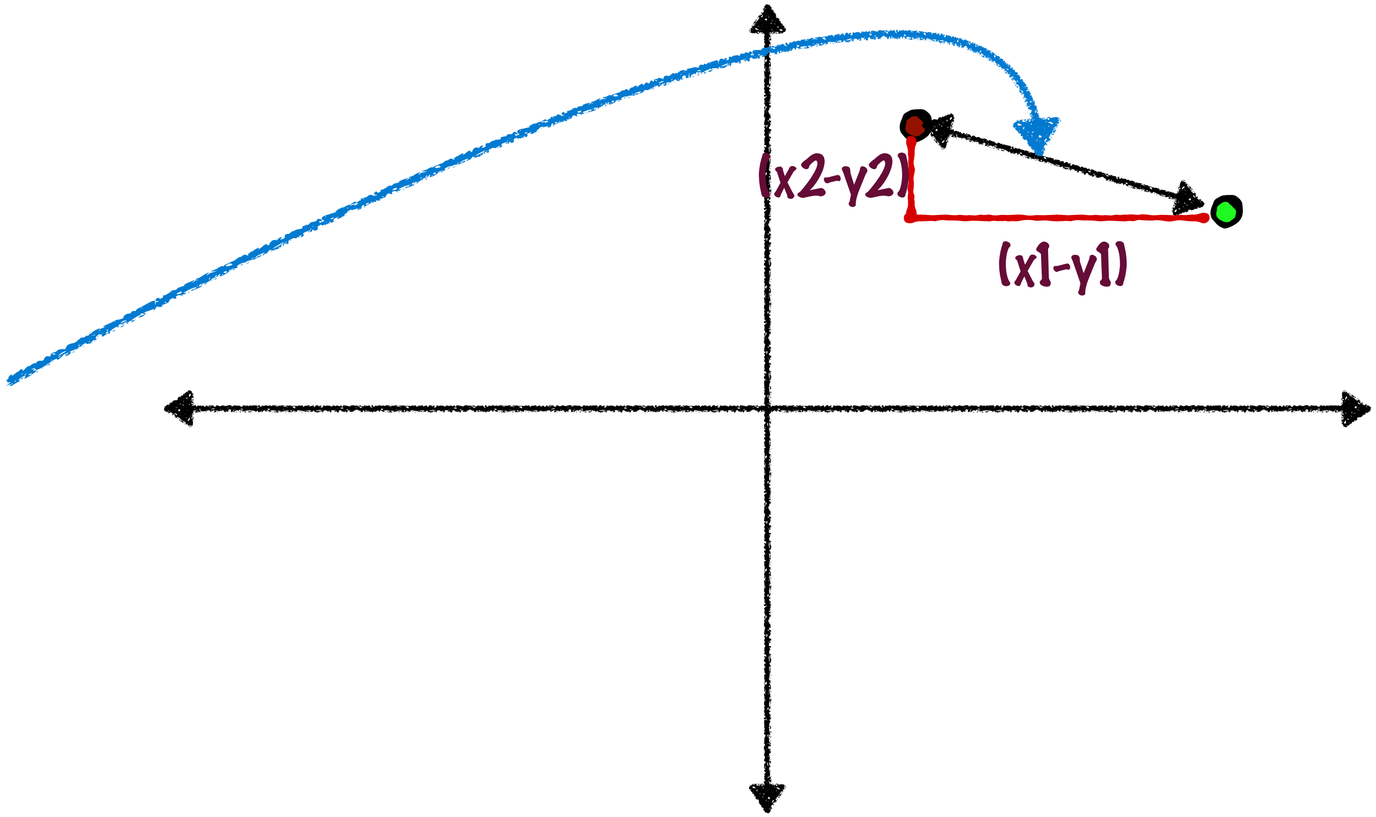
EUCLIDEAN DISTANCE

COSINE SIMILARITY

PEARSON CORRELATION

USER X (x1,x2...xn)
USER Y (y1,y2...yn)

$$d = \sqrt{\sum_{i=1}^{n} \left(x_i - y_i\right)^2}$$



K-NEAREST NEIGHBOURS

USER X (x1,x2...xn)
USER Y (y1,y2...yn)

NEAREST NEIGHBOURS ARE FOUND USING A

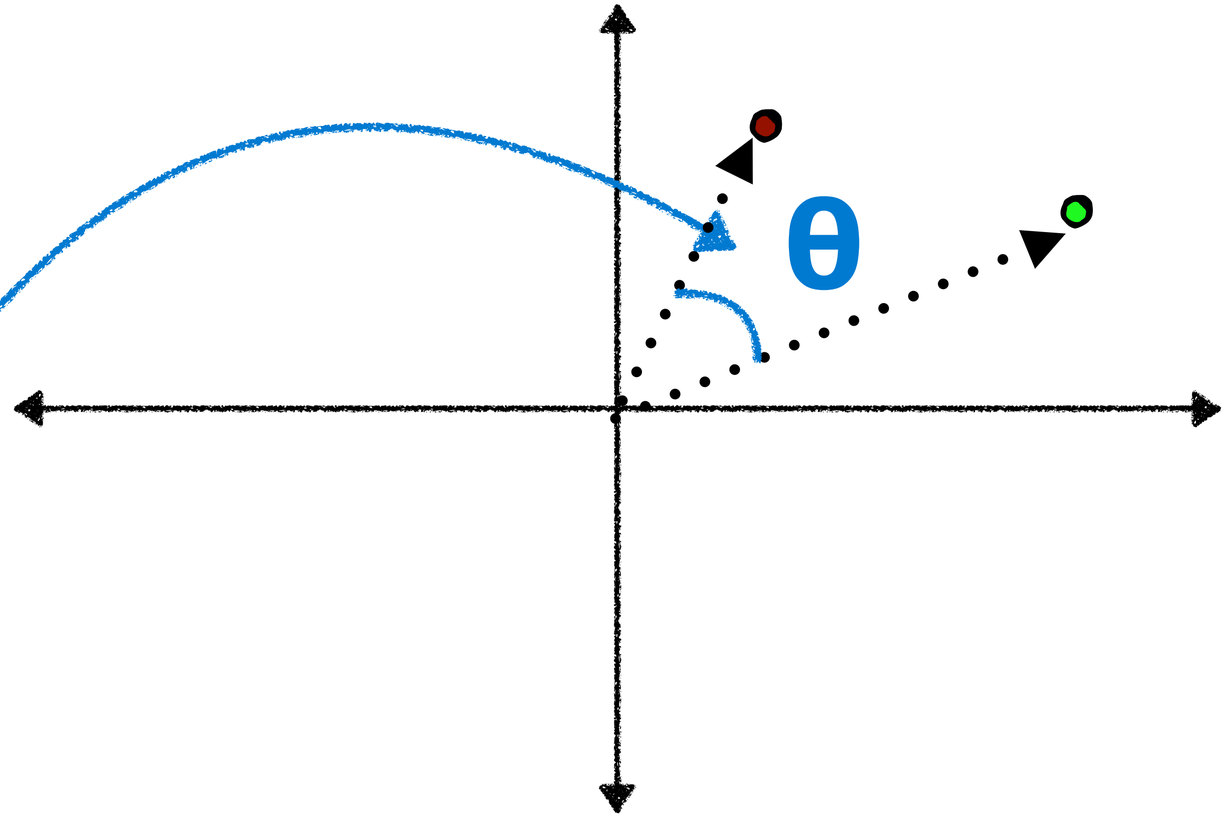
SIMILARITY (OR) DISTANCE METRIC

EUCLIDEAN DISTANCE

COSINE SIMILARITY

PEARSON CORRELATION

$$CosSim(x, y) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}} = \frac{\langle x, y \rangle}{||x||\ ||y||}$$



K-NEAREST NEIGHBOURS

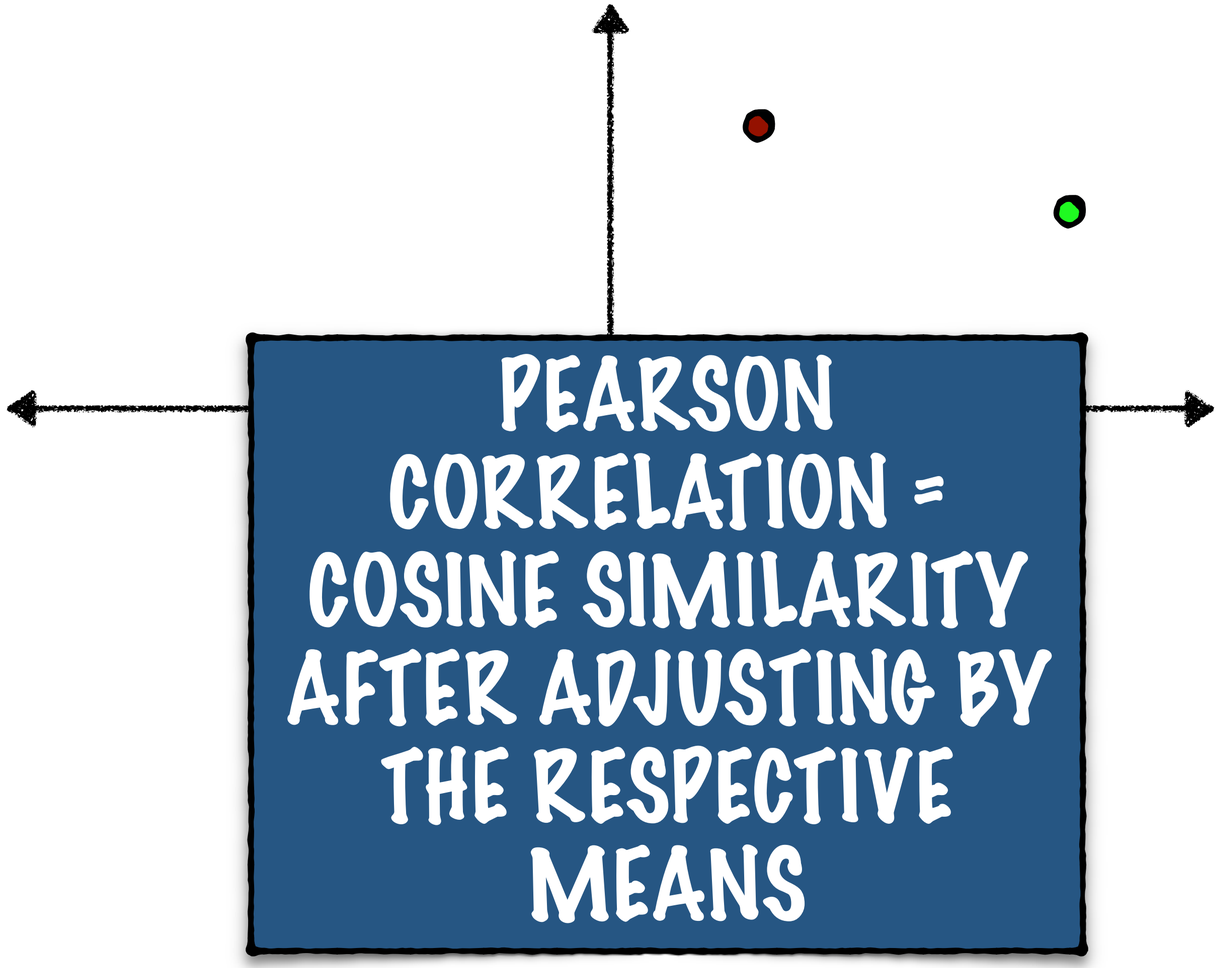
NEAREST NEIGHBOURS ARE FOUND USING A

SIMILARITY (OR) DISTANCE METRIC

EUCLIDEAN DISTANCE

COSINE SIMILARITY

PEARSON CORRELATION



K-NEAREST NEIGHBOURS

USER X (x1,x2...xn)
USER Y (y1,y2...yn)

NEAREST NEIGHBOURS ARE FOUND USING A

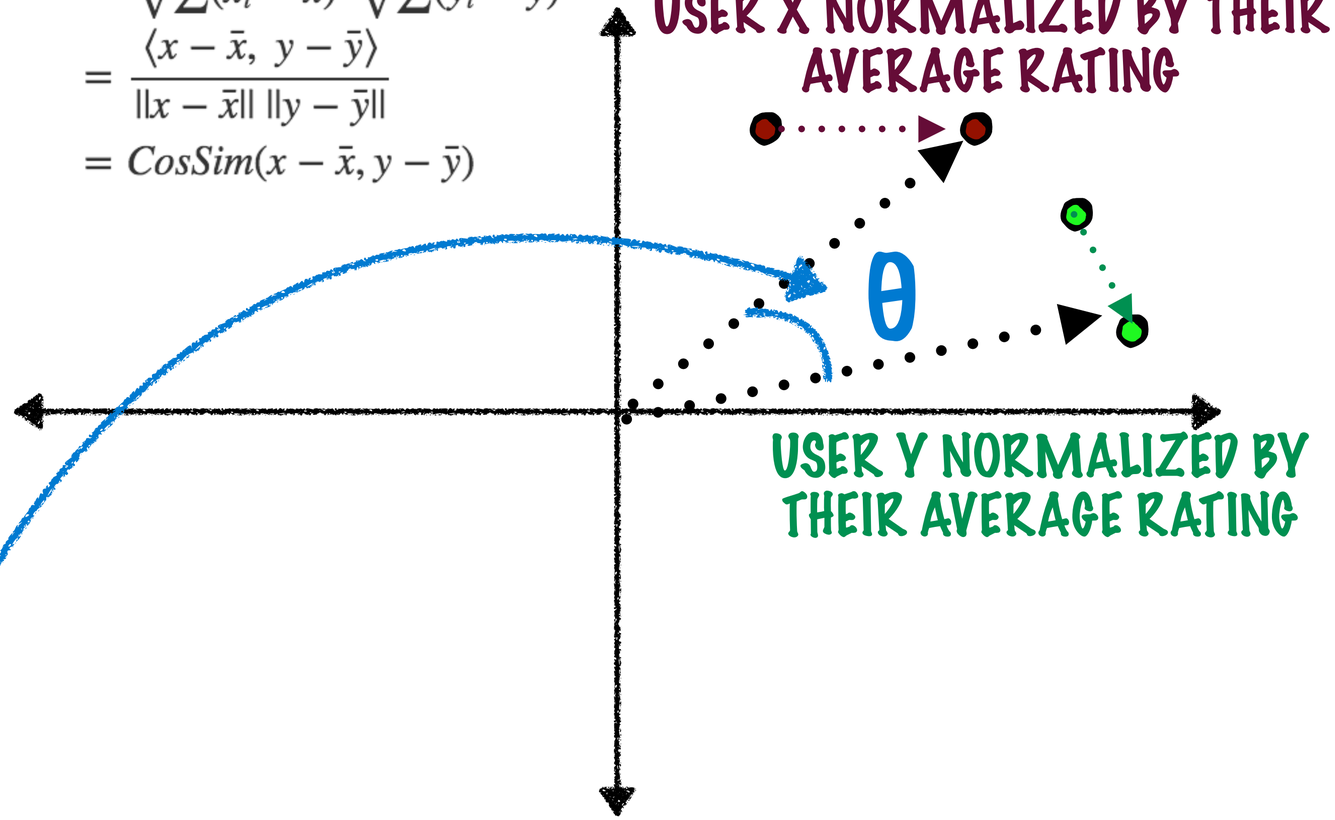
SIMILARITY (OR) DISTANCE METRIC

EUCLIDEAN DISTANCE

COSINE SIMILARITY

PEARSON CORRELATION

$$Corr(x,y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}}$$
$$= \frac{\langle x - \bar{x},\ y - \bar{y} \rangle}{\|x - \bar{x}\|\,\|y - \bar{y}\|}$$
$$= CosSim(x - \bar{x}, y - \bar{y})$$



K-NEAREST NEIGHBOURS

NEAREST NEIGHBOURS ARE FOUND USING A

SIMILARITY (OR) DISTANCE METRIC

EUCLIDEAN DISTANCE

COSINE SIMILARITY

PEARSON CORRELATION

PREDICTED RATING OF ACTIVE USER a FOR PRODUCT i

$$P_{a,i} = \bar{r}_a + \frac{\sum_{u \in U} (r_{u,i} - \bar{r}_u) \cdot w_{a,u}}{\sum_{u \in U} |w_{a,u}|}$$

START WITH THE AVERAGE RATING OF THE ACTIVE USER a FOR ANY PRODUCT

K-NEAREST NEIGHBOURS

NEAREST NEIGHBOURS ARE FOUND USING A

SIMILARITY (OR) DISTANCE METRIC

EUCLIDEAN DISTANCE

COSINE SIMILARITY

PEARSON CORRELATION

PREDICTED RATING OF ACTIVE USER a FOR PRODUCT i

$$P_{a,i} = \bar{r}_a + \frac{\sum_{u \in U} (r_{u,i} - \bar{r}_u) \cdot w_{a,u}}{\sum_{u \in U} |w_{a,u}|}$$

FOR EACH NEIGHBOR (U IS THE SET OF NEAREST NEIGHBOURS OF ACTIVE USER a)

K-NEAREST NEIGHBOURS

NEAREST NEIGHBOURS ARE FOUND USING A

SIMILARITY (OR) DISTANCE METRIC

EUCLIDEAN DISTANCE

COSINE SIMILARITY

PEARSON CORRELATION

PREDICTED RATING OF ACTIVE USER a FOR PRODUCT i

$$P_{a,i} = \bar{r}_a + \frac{\sum_{u \in U} (r_{u,i} - \bar{r}_u) \cdot w_{a,u}}{\sum_{u \in U} |w_{a,u}|}$$

ADD THE RATING OF A USER u FOR PRODUCT i

K-NEAREST NEIGHBOURS

NEAREST NEIGHBOURS ARE FOUND USING A

SIMILARITY (OR) DISTANCE METRIC

EUCLIDEAN DISTANCE

COSINE SIMILARITY

PEARSON CORRELATION

PREDICTED RATING OF ACTIVE USER a FOR PRODUCT i

$$P_{a,i} = \bar{r}_a + \frac{\sum_{u \in U} (r_{u,i} - \bar{r}_u) \cdot w_{a,u}}{\sum_{u \in U} |w_{a,u}|}$$

BUT ADJUST THE NEIGHBOUR'S RATING BY THEIR AVERAGE RATING

K-NEAREST NEIGHBOURS

NEAREST NEIGHBOURS ARE FOUND USING A

SIMILARITY (OR) DISTANCE METRIC

EUCLIDEAN DISTANCE

COSINE SIMILARITY

PEARSON CORRELATION

PREDICTED RATING OF ACTIVE USER a FOR PRODUCT i

$$P_{a,i} = \bar{r}_a + \frac{\sum_{u \in U} (r_{u,i} - \bar{r}_u) \cdot w_{a,u}}{\sum_{u \in U} |w_{a,u}|}$$

AND THE SIMILARITY BETWEEN USER u AND ACTIVE USER a

HOW DO YOU FIND THE TOP PICKS FOR A USER?

NETFLIX

HOW DO YOU FIND THE TOP PICKS FOR A USER?

PREDICT THE RATINGS FOR PRODUCTS THE USER HAS NOT YET BOUGHT (OR SEEN)

PICK THE TOP N RATED PRODUCTS

NEAREST NEIGHBOUR BASED METHODS

ARE ALSO CALLED

MEMORY BASED METHODS

THEY USUALLY INVOLVE IN-MEMORY CALCULATIONS ON THE ENTIRE (OR A LARGE PART) OF THE USER RATINGS DATABASE

AMAZON AND BARNES-AND-NOBLE NOTABLY USE MEMORY BASED COLLABORATIVE FILTERING TECHNIQUES

NEAREST NEIGHBOUR BASED METHODS

ARE ALSO CALLED

MEMORY BASED METHODS

THEY USUALLY INVOLVE IN-MEMORY CALCULATIONS ON THE ENTIRE (OR A LARGE PART) OF THE USER RATINGS DATABASE

NEIGHBOUR BASED METHODS CAN BE BASED ON ITEMS RATHER THAN USERS

FIND THE WEIGHTED AVERAGE RATING OF THE K-NEAREST NEIGHBOURS OF THE ITEM THAT THE USER HAS RATED

COLLABORATIVE FILTERING IS A GENERAL TERM

FOR ANY ALGORITHM THAT RELIES ONLY ON USER BEHAVIOR (HISTORY, RATINGS, SIMILAR USERS ETC)

THE ALGORITHM NORMALLY PREDICTS USERS' RATINGS FOR PRODUCTS THEY HAVEN'T YET RATED

THERE ARE MANY MANY DIFFERENT ALGORITHMS TO PERFORM COLLABORATIVE FILTERING

2 POPULAR TECHNIQUES ARE

NEAREST NEIGHBOUR BASED METHODS

LATENT FACTOR BASED METHODS

COLLABORATIVE FILTERING IS A GENERAL TERM

FOR ANY ALGORITHM THAT RELIES ONLY ON USER BEHAVIOR (HISTORY, RATINGS, SIMILAR USERS ETC)

THE ALGORITHM NORMALLY PREDICTS USERS' RATINGS FOR PRODUCTS THEY HAVEN'T YET RATED

THERE ARE MANY MANY DIFFERENT ALGORITHMS TO PERFORM COLLABORATIVE FILTERING

2 POPULAR TECHNIQUES ARE

NEAREST NEIGHBOUR BASED METHODS

LATENT FACTOR BASED METHODS

LATENT FACTOR BASED METHODS

IDENTIFY HIDDEN FACTORS THAT INFLUENCE A USER'S RATING

THIS IS ANALOGOUS TO CONTENT-BASED FILTERING EXCEPT THAT THE FACTORS ARE IDENTIFIED BY THE LEARNING ALGORITHM

SOMETIMES THE FACTORS MIGHT TURN OUT TO HAVE MEANING (LIKE GENRE OR BOX-OFFICE POPULARITY)

OTHER TIMES, THEY MIGHT BE ABSTRACT FACTORS WITH NO REAL LIFE MEANING

RECAP

CONTENT-BASED FILTERING

CONTENT-BASED FILTERING

PRODUCTS ARE REPRESENTED IN TERMS OF DESCRIPTORS OR ATTRIBUTES

RECAP

COMMERCIAL

USERS ARE REPRESENTED USING THE SAME DESCRIPTORS

JANANI CLICKED ON 10 MOVIES
7 COMMERCIAL, 9 DRAMA

DRAMA

COMEDY

ARTY

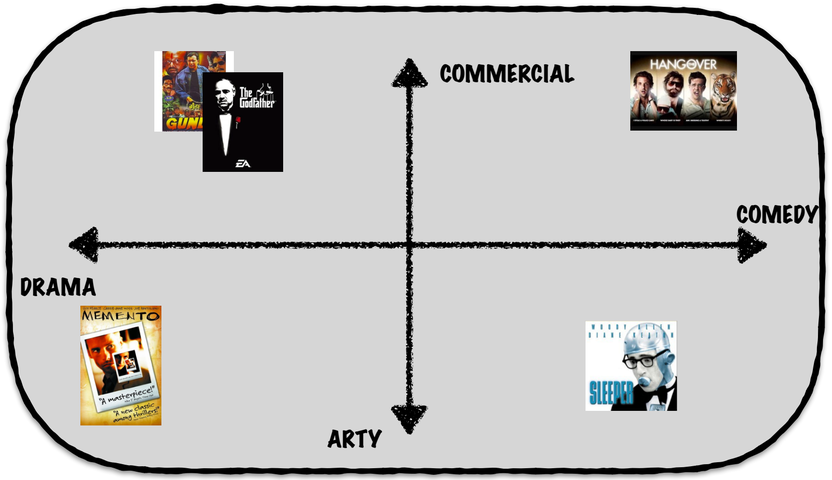
CONTENT-BASED FILTERING

USER HISTORY

+

PRODUCT DESCRIPTIONS REQUIRED

EXPLICITLY DEFINED FACTORS



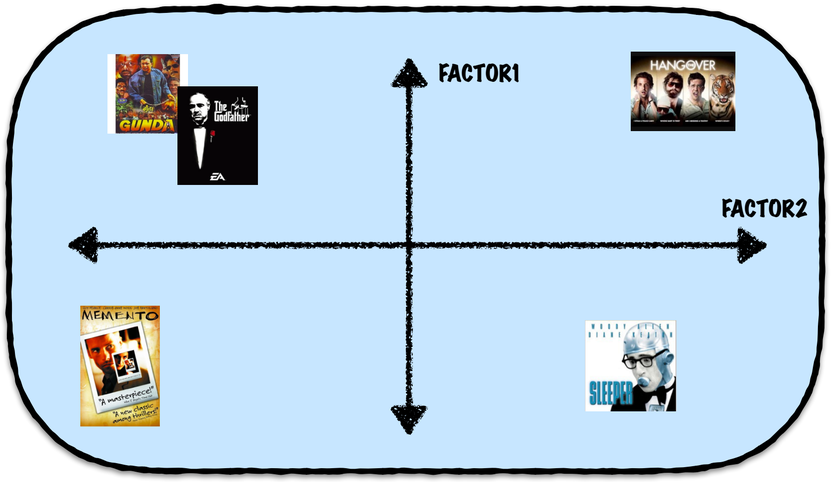
PREDICTED RATING

LATENT FACTOR COLLABORATIVE FILTERING

USER HISTORY (ONLY)

NO PRODUCT DATA REQUIRED

HIDDEN FACTORS IDENTIFIED



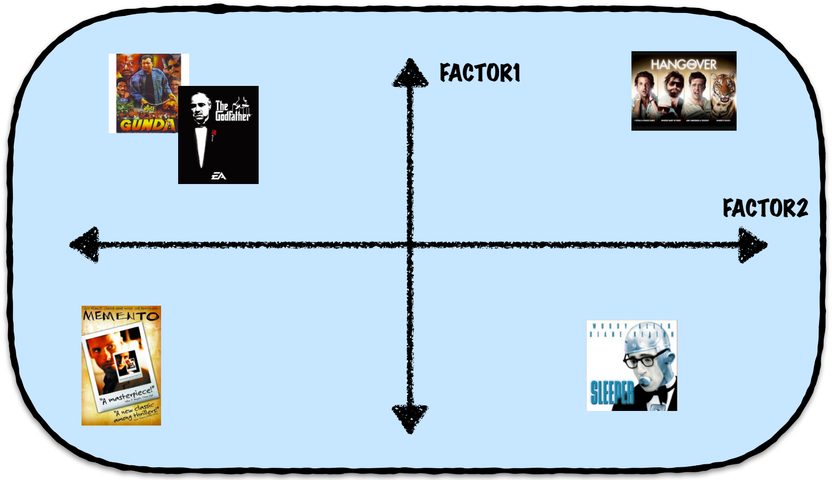
PREDICTED RATING

LATENT FACTOR COLLABORATIVE FILTERING

USER HISTORY (ONLY)

NO PRODUCT DATA REQUIRED

HIDDEN FACTORS IDENTIFIED



PREDICTED RATING

HOW?

REPRESENT USERS BY THEIR RATINGS FOR DIFFERENT PRODUCTS

	ITEM 1	ITEM 2	ITEM 3	ITEM 4	.. ..	ITEM D
USER 1	4	-	4	-	-	-
USER 2	-	3	4	-	-	-
USER 3	5	3	2	-	-	5
USER 4	2	-	2	-	-	4
.. ..	-	-	-	4	-	-
	-	1	-	-	-	-
USER N	4	3	4	-	-	5

	F1	F2	F3
USER 1			
USER 2			
USER 3			
USER 4			
.. ..			
USER N			

*

	ITEM 1	ITEM 2	ITEM 3	ITEM 4	.. ..	ITEM D
F1						
F2						
F3						

LATENT FACTOR COLLABORATIVE FILTERING

R

USER-ITEM RATING MATRIX

	ITEM 1	ITEM 2	ITEM 3	ITEM 4	..	.. ..	ITEM D
USER 1	4	-	4	-	-		-
USER 2	-	3	4	-	-		-
USER 3	5	3	2	-	-		5
USER 4	2	-	2	-	-		4
.. ..	-	-	-	4	-		-
	-	1	-	-	-		-
USER N	4	3	4	-	-		5

DECOMPOSED INTO

P

	F1	F2	F3
USER 1			
USER 2			
USER 3			
USER 4			
.. ..			
USER N			

USER-FACTOR MATRIX

EACH ROW IS A USER DESCRIBED BY THE HIDDEN FACTORS

*

Q

	ITEM 1	ITEM 2	ITEM 3	ITEM 4	..	.. ..	ITEM D
F1							
F2							
F3							

PRODUCT-FACTOR MATRIX

EACH COLUMN IS A PRODUCT DESCRIBED BY THE HIDDEN FACTORS

THIS METHOD IS CALLED

MATRIX FACTORIZATION

IT WAS INVENTED AND POPULARIZED BY THE NETFLIX PRIZE WINNERS

NOWADAYS, MANY RECOMMENDATION SYSTEMS USE SOME OR OTHER FORM OF MATRIX FACTORIZATION

LATENT FACTOR COLLABORATIVE FILTERING

R

USER-ITEM RATING MATRIX

	ITEM 1	ITEM 2	ITEM 3	ITEM 4	..	.. ..	ITEM D
USER 1	4	-	4	-	-		-
USER 2	-	3	4	-	-		-
USER 3	5	3	2	-	-		5
USER 4	2	-	2	-	-		4
.. ..	-	-	-	4	-		-
	-	1	-	-	-		-
USER N	4	3	4	-	-		5

DECOMPOSED INTO

P

	F1	F2	F3
USER 1			
USER 2			
USER 3			
USER 4			
.. ..			
USER N			

USER-FACTOR MATRIX

*

Q

	ITEM 1	ITEM 2	ITEM 3	ITEM 4	..	.. ..	ITEM D
F1							
F2							
F3							

PRODUCT-FACTOR MATRIX

THE OBJECTIVE OF

MATRIX FACTORIZATION

IS TO DECOMPOSE EACH USER RATING INTO A USER-FACTOR VECTOR AND A PRODUCT-FACTOR VECTOR

THIS IS ANALOGOUS TO WHAT HAPPENS IN

SINGULAR VALUE DECOMPOSITION OR PRINCIPAL COMPONENT ANALYSIS

LATENT FACTOR COLLABORATIVE FILTERING

THE OBJECTIVE OF

MATRIX FACTORIZATION

IS TO DECOMPOSE EACH USER RATING INTO A USER-FACTOR VECTOR AND A PRODUCT-FACTOR VECTOR

ALSO, SVD OF HUGE MATRICES IS USUALLY A PAIN!

IF THAT WERE THE CASE, THERE WOULD BE NO RECOMMENDATION PROBLEM TO SOLVE! :)

THIS IS ANALOGOUS TO WHAT HAPPENS IN

SINGULAR VALUE DECOMPOSITION OR PRINCIPAL COMPONENT ANALYSIS

HOWEVER, THESE TECHNIQUES WOULD ONLY MAKE SENSE

IF YOU KNEW ALL THE RATINGS FOR ALL THE USERS FOR ALL PRODUCTS

(IF THERE WERE NO MISSING VALUES IN THE USER ITEM RATING MATRIX)

AND YOU WERE JUST LOOKING TO FIND LATENT OR HIDDEN FACTORS TO EXPLAIN THE RATINGS

LATENT FACTOR COLLABORATIVE FILTERING

R

USER-ITEM RATING MATRIX

	ITEM 1	ITEM 2	ITEM 3	ITEM 4	..	.. ..	ITEM D
USER 1	4	-	4	-	-		-
USER 2	-	3	4	-	-		-
USER 3	5	3	2	-	-		5
USER 4	2	-	2	-	-		4
.. ..	-	-	-	4	-		-
	-	1	-	-	-		-
USER N	4	3	4	-	-		5

DECOMPOSED INTO

P

	F1	F2	F3
USER 1			
USER 2			
USER 3			
USER 4			
.. ..			
USER N			

USER-FACTOR MATRIX

*

Q

	ITEM 1	ITEM 2	ITEM 3	ITEM 4	..	.. ..	ITEM D
F1							
F2							
F3							

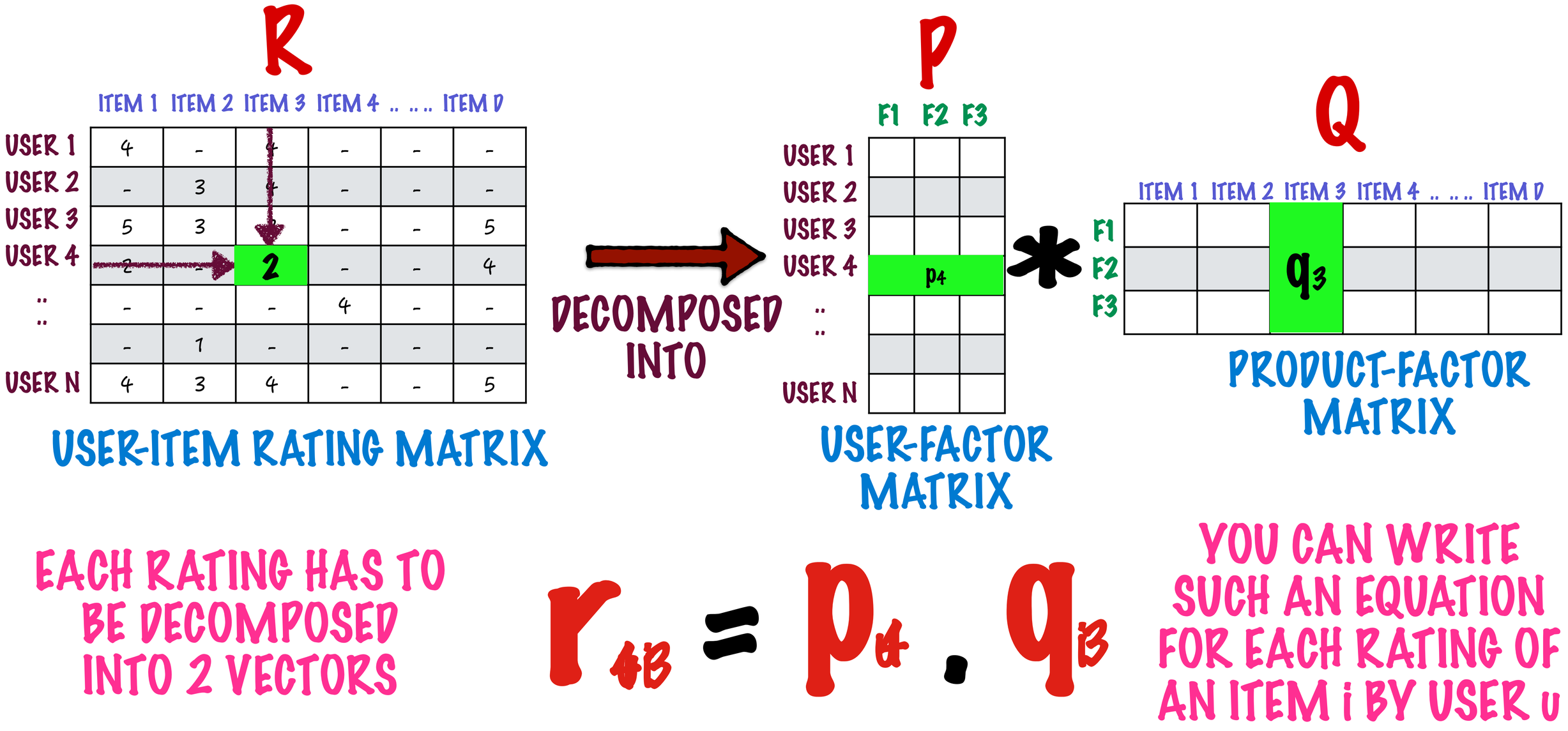
PRODUCT-FACTOR MATRIX

SO THE QUESTION IS

HOW DO YOU PERFORM MATRIX FACTORIZATION WHEN THE RATING MATRIX HAS SO MANY MISSING VALUES?

WE ONLY SOLVE FOR THE RATINGS WHICH ARE AVAILABLE

LATENT FACTOR COLLABORATIVE FILTERING

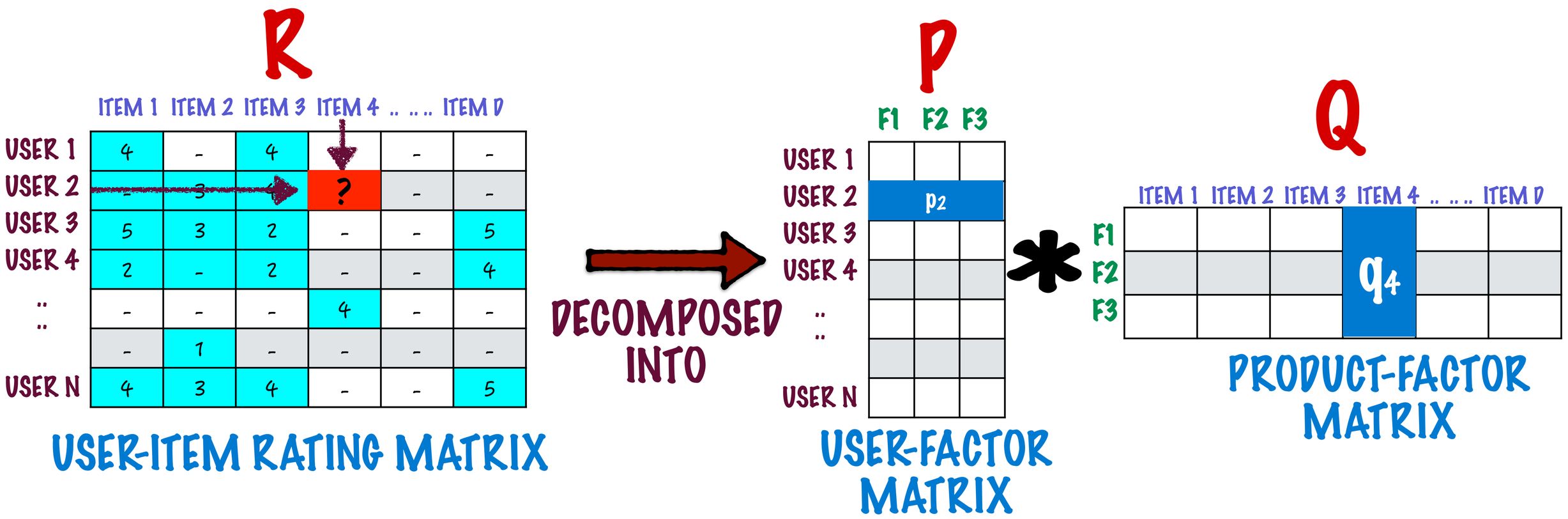


EACH RATING HAS TO BE DECOMPOSED INTO 2 VECTORS

$$r_{4B} = p_4 \cdot q_B$$

YOU CAN WRITE SUCH AN EQUATION FOR EACH RATING OF AN ITEM i BY USER u

LATENT FACTOR COLLABORATIVE FILTERING



SOLVE THIS SET OF EQUATIONS FOR THE SET OF RATINGS WHICH EXIST (TRAINING SET)

$$r_{ui} = p_u \cdot q_i$$

USE THE RESULTING P'S AND Q'S TO FIND THE RATING OF ANY USER FOR ANY PRODUCT

LATENT FACTOR COLLABORATIVE FILTERING

SOLVE THIS SET OF EQUATIONS FOR THE SET OF RATINGS WHICH EXIST (TRAINING SET)

$$r_{ui} = p_u . q_i$$

THIS CAN BE NICELY SET UP AS AN OPTIMIZATION PROBLEM

FIND THE SET OF FACTOR VECTORS

pu (FOR EACH USER u) AND

qi (FOR EACH ITEM i)

WHICH MINIMIZE THE ERROR ON THE TRAINING SET

$$\min_{q^*,p^*} \sum_{(u,i)\in\kappa} (r_{ui} - q_i^T p_u)^2 + \lambda(\| q_i \|^2 + \| p_u \|^2)$$

ERROR FOR ONE RATING IN THE TRAINING SET

ADD A REGULARIZATION TERM

WHAT IF WE FIND TOO MANY HIDDEN FACTORS - IE OVERFIT THE TRAINING SET?

PENALIZE MODELS WITH HIGHER NUMBER OF FACTORS

LATENT FACTOR COLLABORATIVE FILTERING

$$\min_{q^*,p^*} \sum_{(u,i)\in\kappa} (r_{ui} - q_i^T p_u)^2 + \lambda(\| q_i \|^2 + \| p_u \|^2)$$

FIND THE SET OF FACTOR VECTORS

pu (FOR EACH USER u) AND

qi (FOR EACH ITEM i)

WHICH MINIMIZE THE ERROR ON THE TRAINING SET

STOCHASTIC GRADIENT DESCENT

ALTERNATING LEAST SQUARES

ARE TWO POPULAR METHODS TO SOLVE MATRIX FACTORIZATION FOR RECOMMENDATIONS

LATENT FACTOR COLLABORATIVE FILTERING

FIND THE SET OF FACTOR VECTORS
p_u (FOR EACH USER u) AND
q_i (FOR EACH ITEM i)

$$\min_{q^*,p^*} \sum_{(u,i)\in\kappa} (r_{ui} - q_i^T p_u)^2 + \lambda(\| q_i \|^2 + \| p_u \|^2)$$

WHICH MINIMIZE THE ERROR ON THE TRAINING SET

STOCHASTIC GRADIENT DESCENT

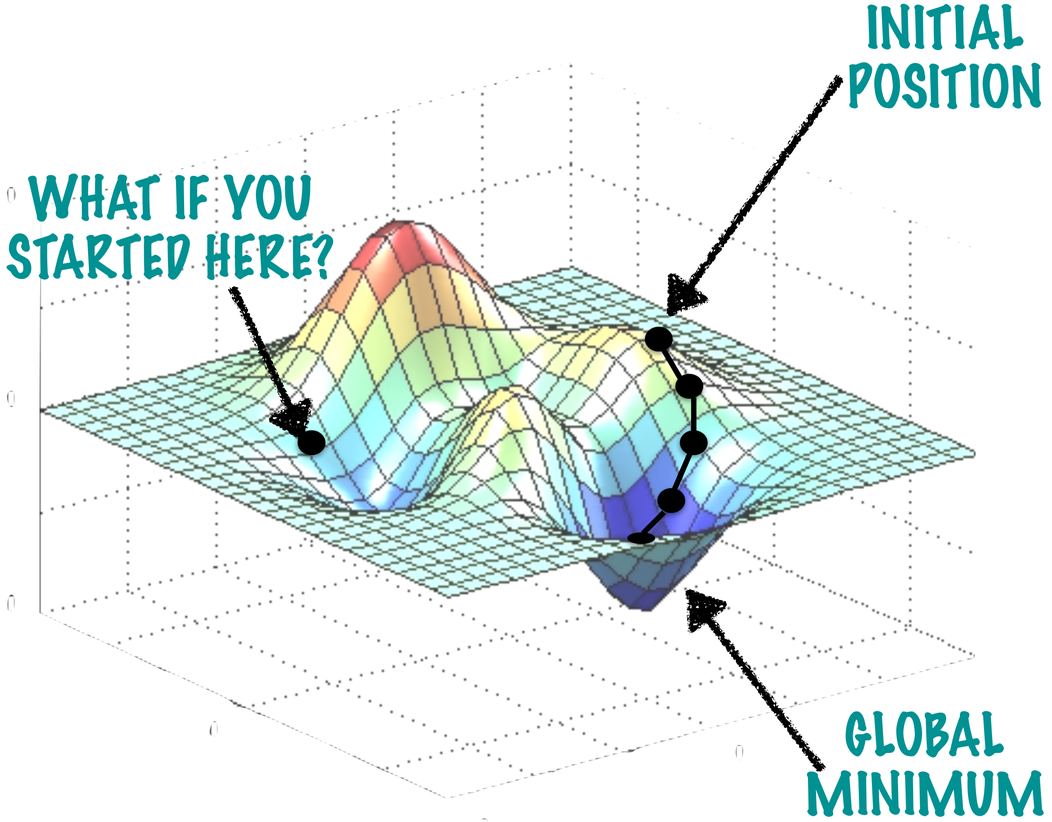
LET'S CONCENTRATE ON 1 RATING r_{ui}

1. INITIALIZE SOME VALUES OF p_u AND q_i

2. FIND THE CURRENT VALUE OF THE ERROR FUNCTION

3. FIND THE SLOPE AT THE CURRENT POINT AND MOVE SLIGHTLY DOWNWARDS IN THAT DIRECTION

4. REPEAT UNTIL YOU REACH A MINIMUM



GRADIENT DESCENT DOESN'T GUARANTEE THE GLOBAL MINIMUM

LATENT FACTOR COLLABORATIVE FILTERING

FIND THE SET OF FACTOR VECTORS
p_u (FOR EACH USER u) AND
q_i (FOR EACH ITEM i)

$$\min_{q^*,p^*} \sum_{(u,i)\in K} (r_{ui} - q_i^T p_u)^2 + \lambda(\| q_i \|^2 + \| p_u \|^2)$$

WHICH MINIMIZE THE ERROR ON THE TRAINING SET

STOCHASTIC GRADIENT DESCENT

LET'S CONCENTRATE ON 1 RATING r_{ui}

1. INITIALIZE SOME VALUES OF p_u AND q_i

2. FIND THE CURRENT VALUE OF THE ERROR FUNCTION $e_{ui} = (r_{ui} - q_i^T p_u)^2 + \lambda(\| q_i \|^2 + \| p_u \|^2)$

3. FIND THE SLOPE AT THE CURRENT POINT AND MOVE SLIGHTLY DOWNWARDS IN THAT DIRECTION

4. REPEAT UNTIL YOU REACH A MINIMUM

$$q_i \leftarrow q_i - \gamma \cdot \frac{\partial e_{ui}}{\partial q_i}$$

$$p_u \leftarrow p_u - \gamma \cdot \frac{\partial e_{ui}}{\partial p_u}$$

THE SIZE OF THE MOVEMENT

THE DIRECTION OF THE MOVEMENT

LATENT FACTOR COLLABORATIVE FILTERING

$$\min_{q^*,p^*} \sum_{(u,i)\in K} (r_{ui} - q_i^T p_u)^2 + \lambda(\| q_i \|^2 + \| p_u \|^2)$$

FIND THE SET OF FACTOR VECTORS

pu (FOR EACH USER u) AND

qi (FOR EACH ITEM i)

WHICH MINIMIZE THE ERROR ON THE TRAINING SET

STOCHASTIC GRADIENT DESCENT

ALTERNATING LEAST SQUARES

ARE TWO POPULAR METHODS TO SOLVE MATRIX FACTORIZATION FOR RECOMMENDATIONS

LATENT FACTOR COLLABORATIVE FILTERING

$$\min_{q^*,p^*} \sum_{(u,i)\in K} (r_{ui} - q_i^T p_u)^2 + \lambda(\| q_i \|^2 + \| p_u \|^2)$$

FIND THE SET OF FACTOR VECTORS

pu (FOR EACH USER u) AND

qi (FOR EACH ITEM i)

WHICH MINIMIZE THE ERROR ON THE TRAINING SET

STOCHASTIC GRADIENT DESCENT

ALTERNATING LEAST SQUARES

ARE TWO POPULAR METHODS TO SOLVE MATRIX FACTORIZATION FOR RECOMMENDATIONS

LATENT FACTOR COLLABORATIVE FILTERING

FIND THE SET OF FACTOR VECTORS p_u (FOR EACH USER u) AND q_i (FOR EACH ITEM i)

$$\min_{q^*,p^*} \sum_{(u,i)\in\kappa} (r_{ui} - q_i^T p_u)^2 + \lambda(\| q_i \|^2 + \| p_u \|^2)$$

WHICH MINIMIZE THE ERROR ON THE TRAINING SET

ALTERNATING LEAST SQUARES

LET'S CONCENTRATE ON 1 RATING r_{ui}

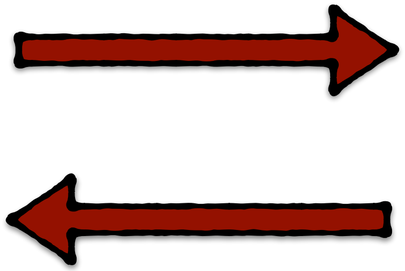
THE EQUATION TO SOLVE IS

THERE ARE 2 VARIABLES IE p_u AND q_i

$$(r_{ui} - q_i^T p_u)^2 + \lambda(\| q_i \|^2 + \| p_u \|^2) = 0$$

WHAT IF WE FIXED THE VALUE OF p_u ?

WE ARE LEFT WITH A QUADRATIC EQUATION FOR THE VARIABLE q_i



WHAT IF WE FIXED THE VALUE OF q_i ?

WE ARE LEFT WITH A QUADRATIC EQUATION FOR THE VARIABLE p_u

REPEAT UNTIL THE VALUES OF p_u AND q_i CONVERGE

LATENT FACTOR COLLABORATIVE FILTERING

$$\min_{q^*,p^*} \sum_{(u,i)\in K} (r_{ui} - q_i^T p_u)^2 + \lambda(\| q_i \|^2 + \| p_u \|^2)$$

FIND THE SET OF FACTOR VECTORS
p_u (FOR EACH USER u) AND
q_i (FOR EACH ITEM i)

WHICH MINIMIZE THE ERROR
ON THE TRAINING SET

NORMALIZING FOR

USER BIASES

TEMPORAL EFFECTS

STOCHASTIC GRADIENT DESCENT

ALTERNATING LEAST SQUARES

ARE TWO POPULAR METHODS TO SOLVE
MATRIX FACTORIZATION FOR
RECOMMENDATIONS

SOME OF THE WAYS TO IMPROVE
MATRIX FACTORIZATION EVEN
FURTHER

COLLABORATIVE FILTERING IS A GENERAL TERM

FOR ANY ALGORITHM THAT RELIES ONLY ON USER BEHAVIOR (HISTORY, RATINGS, SIMILAR USERS ETC)

THE ALGORITHM NORMALLY PREDICTS USERS' RATINGS FOR PRODUCTS THEY HAVEN'T YET RATED

THERE ARE MANY MANY DIFFERENT ALGORITHMS TO PERFORM COLLABORATIVE FILTERING

2 POPULAR TECHNIQUES ARE

NEAREST NEIGHBOUR BASED METHODS

LATENT FACTOR BASED METHODS

THERE ARE A FEW COMMON CHALLENGES WITH

COLLABORATIVE FILTERING

COLD START

HOW DO YOU DEAL WITH NEW PRODUCTS OR USERS WITH NO HISTORY?

SYNONYMY

HOW DO YOU DEAL WITH PRODUCTS WHICH ARE BASICALLY THE SAME BUT WITH DIFFERENT NAMES?

GRAY SHEEP

ARE THERE SOME KIND OF USERS THAT COLLABORATIVE FILTERING WON'T WORK FOR?

DATA SPARSITY

WHAT DO YOU DO WHEN YOUR RATINGS DATA IS VERY SPARSE?

SHILLING ATTACKS

HOW DO YOU DEAL WITH USERS THAT ARE TRYING TO GAME THE SYSTEM?

THERE ARE A FEW COMMON CHALLENGES WITH

COLLABORATIVE FILTERING

COLD START

COLLABORATIVE FILTERING RELIES ON USER HISTORY
WHAT DO WE DO WITH PRODUCTS, USERS WITH NO HISTORY?

CONTENT-BOOSTED COLLABORATIVE FILTERING

YOU'LL NEED TO USE A COMBINATION OF CONTENT BASED FILTERING AND COLLABORATIVE FILTERING

USE PRODUCT ATTRIBUTES AND DEMOGRAPHICS DATA TO AUGMENT COLLABORATIVE FILTERING

DATA SPARSITY

SHILLING ATTACKS

GRAY SHEEP

SYNONYMY

THERE ARE A FEW COMMON CHALLENGES WITH

COLLABORATIVE FILTERING

DATA SPARSITY

ONLINE STORES NORMALLY HAVE

A HUGE NUMBER OF PRODUCTS

A HUGE NUMBER OF USERS

VERY FEW PRODUCTS ARE RATED BY MULTIPLE USERS

VERY FEW USERS RATE MULTIPLE PRODUCTS

DIMENSIONALITY REDUCTION

CAN HELP REMOVE UNIMPORTANT DIMENSIONS AND REDUCE THE SPARSITY OF THE USER-ITEM RATING MATRIX

COLD START

SHILLING ATTACKS

GRAY SHEEP

SYNONYMY

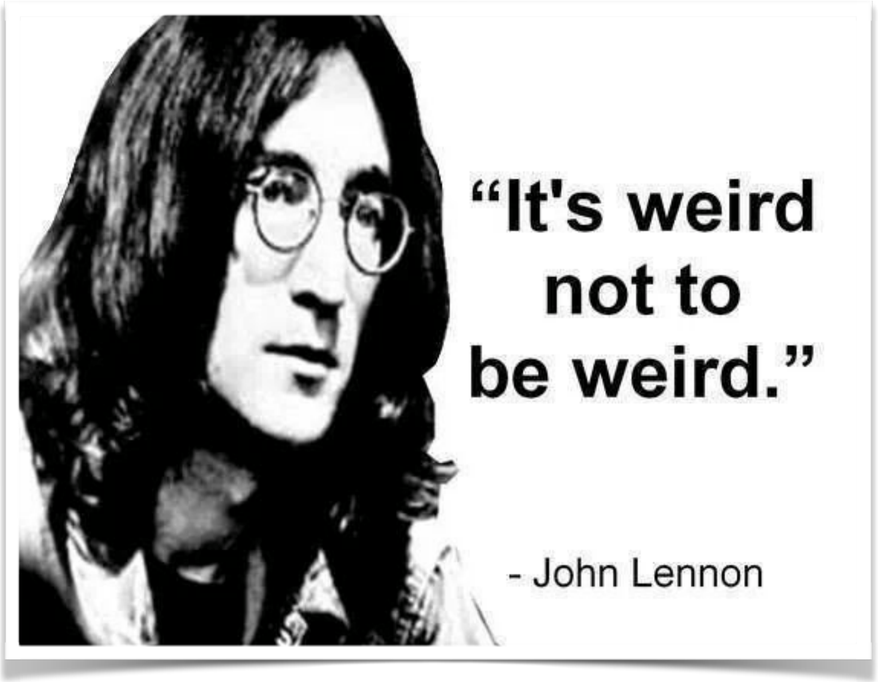
THERE ARE A FEW COMMON CHALLENGES WITH

COLLABORATIVE FILTERING

GRAY SHEEP

EVERY ONCE IN A WHILE, SOMEONE SPECIAL COMES ALONG :)

A GRAY SHEEP IS SOMEONE WHOSE OPINION IS NOT CONSISTENT



PURE COLLABORATIVE FILTERING JUST DOESN'T WORK FOR THESE USERS

COLD START

SHILLING ATTACKS

SYNONYMY

DATA SPARSITY

THERE ARE A FEW COMMON CHALLENGES WITH

COLLABORATIVE FILTERING

SYNONYMY

SOME PRODUCTS ARE PRACTICALLY THE SAME

DIFFERENT EDITIONS OF A BOOK

AN E-BOOK VS A PHYSICAL COPY

NORMALLY STORES WOULD HAVE DIFFERENT PRODUCT CODES FOR EACH OF THESE

SINCE COLLABORATIVE FILTERING DOESN'T USE ANY PRODUCT DESCRIPTIONS, IT MIGHT MISS OUT ON THIS INFORMATION

LATENT FACTOR BASED COLLABORATIVE FILTERING TECHNIQUES HAVE BEEN SEEN TO IDENTIFY SYNONYMS VERY WELL

COLD START

SHILLING ATTACKS

GRAY SHEEP

DATA SPARSITY

THERE ARE A FEW COMMON CHALLENGES WITH

COLLABORATIVE FILTERING

SHILLING ATTACKS

AN AUTHOR MIGHT GIVE TONS OF POSITIVE FAKE RATINGS FOR THEIR OWN CONTENT

AND TONS OF NEGATIVE RATINGS FOR OTHER PEOPLE'S CONTENT

TAKING PRECAUTIONS AGAINST THESE CAN MAKE THE RECOMMENDATION SYSTEM MORE ROBUST

COLD START

SYNONYMY

GRAY SHEEP

DATA SPARSITY

THERE ARE A FEW COMMON CHALLENGES WITH

COLLABORATIVE FILTERING

COLD START

HOW DO YOU DEAL WITH NEW PRODUCTS OR USERS WITH NO HISTORY?

SYNONYMY

HOW DO YOU DEAL WITH PRODUCTS WHICH ARE BASICALLY THE SAME BUT WITH DIFFERENT NAMES?

GRAY SHEEP

ARE THERE SOME KIND OF USERS THAT COLLABORATIVE FILTERING WON'T WORK FOR?

DATA SPARSITY

WHAT DO YOU DO WHEN YOUR RATINGS DATA IS VERY SPARSE?

SHILLING ATTACKS

HOW DO YOU DEAL WITH USERS THAT ARE TRYING TO GAME THE SYSTEM?

RECOMMENDATION ENGINES NORMALLY USE ONE OR MORE OF THESE TECHNIQUES

CONTENT-BASED FILTERING

COLLABORATIVE FILTERING

ASSOCIATION RULES

RECOMMENDATION ENGINES NORMALLY USE ONE OR MORE OF THESE TECHNIQUES

CONTENT-BASED FILTERING

COLLABORATIVE FILTERING

ASSOCIATION RULES

ASSOCIATION RULES

ARE NORMALLY USED FOR A FANCY TASK CALLED

MARKET BASKET ANALYSIS

IN OTHER WORDS..

WHAT ITEMS ARE NORMALLY
BOUGHT AT THE SAME TIME?

OR, BOUGHT BY THE SAME USER
WITHIN A SHORT PERIOD OF TIME?

ASSOCIATION RULE LEARNING

LET'S SAY YOU WORK AT AN ECOMMERCE COMPANY AS A CATEGORY MANAGER

YOU ARE IN CHARGE OF SELLING MOBILE ACCESSORIES - THINGS LIKE CELLPHONE CASES, CHARGERS ETC

YOUR JOB IS TO SELL A LOT OF STUFF, AND AT PRICES AS HIGH AS POSSIBLE, AND [illegible]SSIBLE ON MAR[illegible]

RECAP

WHAT IF YOU COULD FIGURE OUT, SOMEHOW, THAT FOLKS WHO BOUGHT ADAPTERS AND EARPLUGS WERE MORE LIKELY TO BUY CELLPHONE CHARGERS -

THAT INFORMATION COULD REALLY HELP - YOU COULD PERHAPS "BUNDLE" ADAPTERS AND CELLPHONE CHARGERS, OR DISPLAY PROMOTIONAL PRICING, OR OFFER QUANTITY DISCOUNTS

{Adapter, Earmuffs} -> {Cellphone Charger}

IDENTIFYING RULES OF THIS SORT IS EXACTLY WHAT ASSOCIATION RULE LEARNING IS ALL ABOUT

ASSOCIATION RULES

ARE NORMALLY USED FOR A FANCY TASK CALLED

MARKET BASKET ANALYSIS

IN OTHER WORDS..

WHAT ITEMS ARE NORMALLY
BOUGHT AT THE SAME TIME?

OR, BOUGHT BY THE SAME USER
WITHIN A SHORT PERIOD OF TIME?

THE APRIORI ALGORITHM

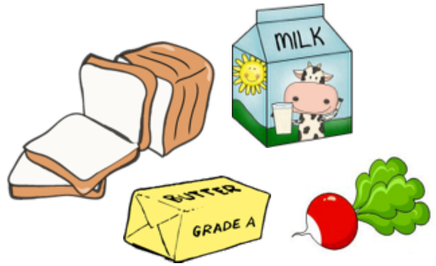
IS THE MOST WELL KNOWN TECHNIQUE FOR
MINING ASSOCIATION RULES

THE APRIORI ALGORITHM

HELPS US MINE BASKET DATA FOR RULES

A BASKET COULD BE
ITEMS BOUGHT IN 1 TRANSACTION (OR)
ITEMS BOUGHT BY A USER OVER A SHORT PERIOD OF TIME

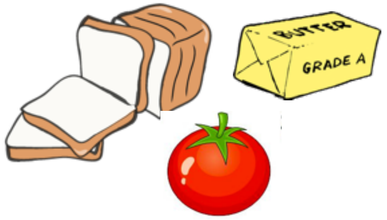
BASKET 1



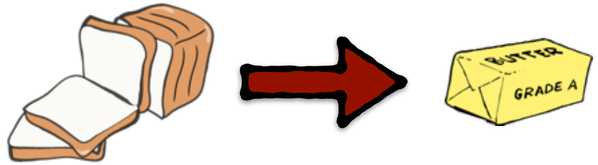
BASKET 3

BASKET 2

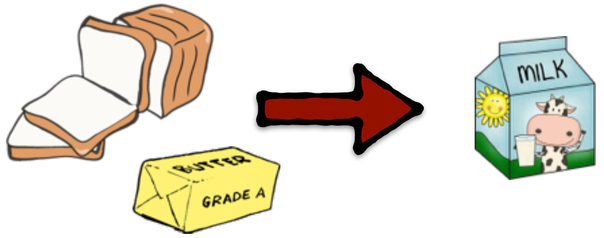
BASKET 4



BASKET 5



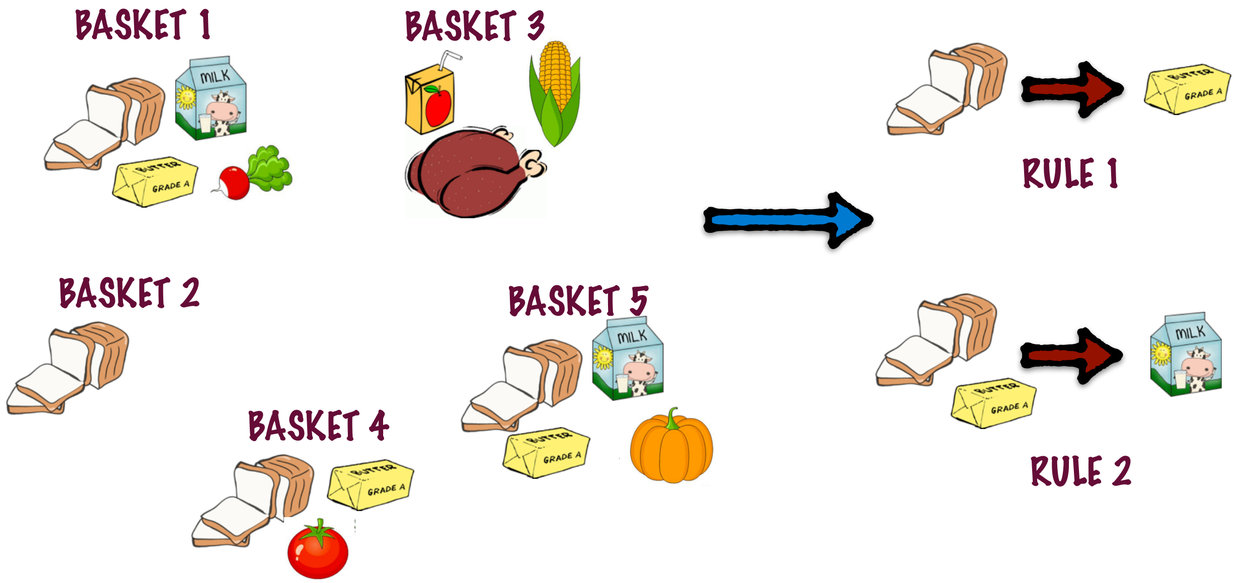
RULE 1



RULE 2

THE APRIORI ALGORITHM

HELPS US MINE BASKET DATA FOR RULES



EACH RULE HAS TO SATISFY A MINIMUM

SUPPORT THE PROPORTION OF BASKETS WHICH CONTAIN THE ITEMS ON ONE SIDE OF THE RULE

AND

CONFIDENCE

SUPPORT SPECIFIES THE MINIMUM NUMBER OF SAMPLES REQUIRED FOR THE RULE TO BE STATISTICALLY SIGNIFICANT

IF THE ITEM HAS BEEN SOLD TOO FEW TIMES, THERE ISN'T ENOUGH DATA TO SUPPORT THE RULE

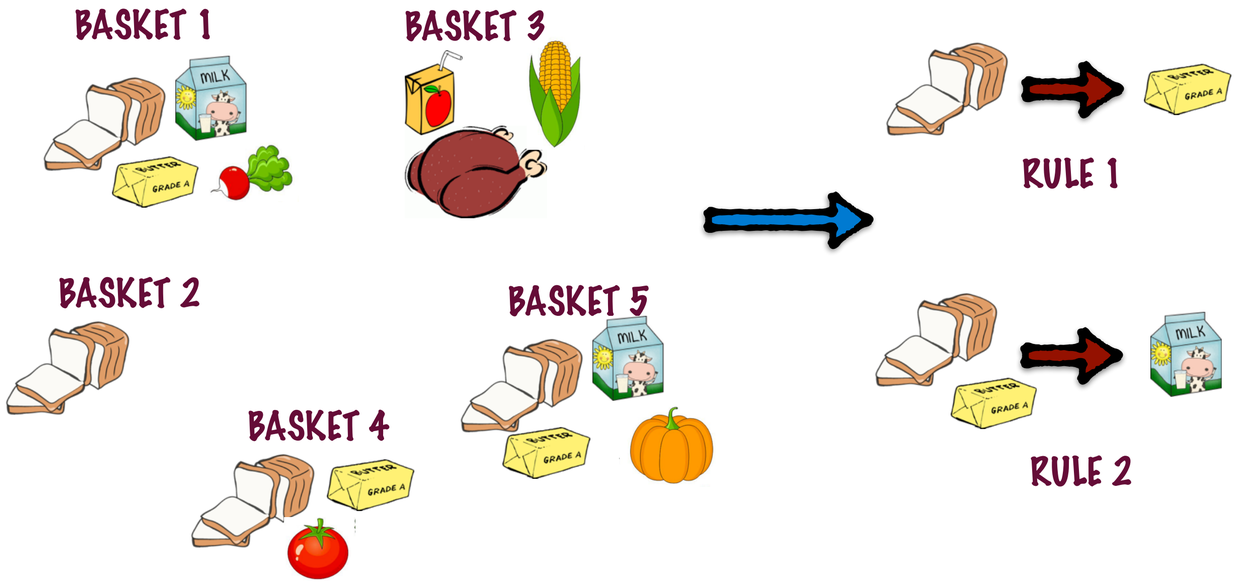
LET'S TAKE RULE 1

SUPP() = PROPORTION OF ALL TRANSACTIONS THAT CONTAIN  = 4/5 = 80%

SUPP() = PROPORTION OF ALL TRANSACTIONS THAT CONTAIN  = 3/5 = 60%

THE APRIORI ALGORITHM

HELPS US MINE BASKET DATA FOR RULES



EACH RULE HAS TO SATISFY A MINIMUM

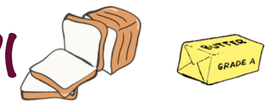
SUPPORT THE PROPORTION OF BASKETS WHICH CONTAIN THE ITEMS ON ONE SIDE OF THE RULE

AND

CONFIDENCE THE PROPORTION OF BASKETS WHICH CONTAIN THE ITEMS ON BOTH SIDES OF THE RULE AS COMPARED TO LEFT SIDE ALONE

LET'S TAKE RULE 1

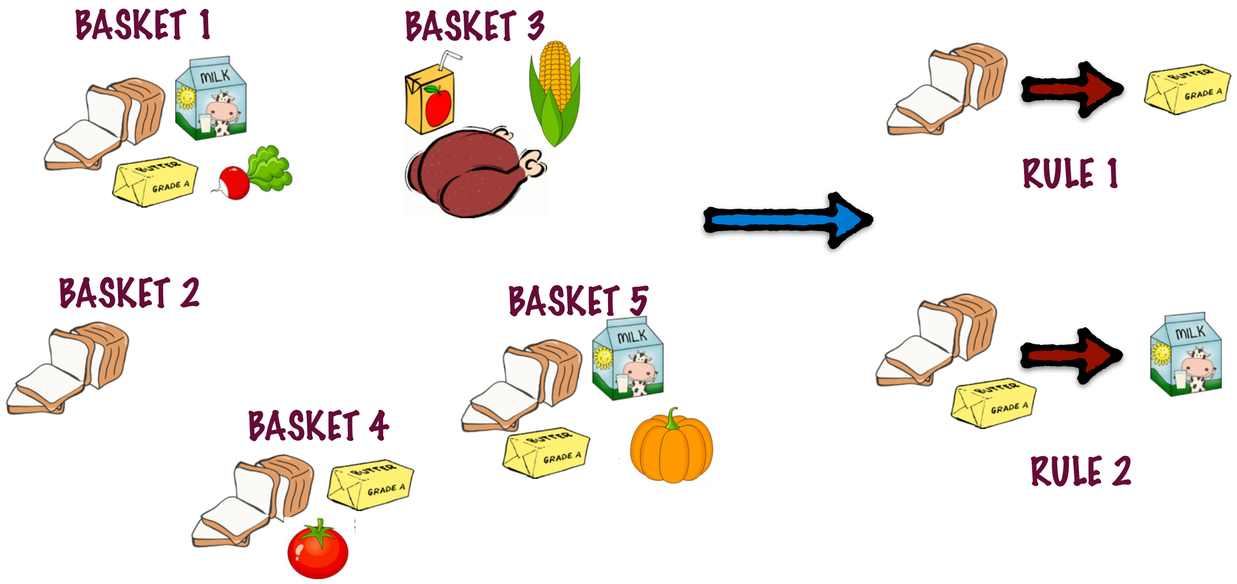
CONF() = HOW MANY TIMES WERE  AND  BOUGHT TOGETHER INSTEAD OF SEPARATELY

$$= \frac{\text{SUPP}(\ \text{  }\ )}{\text{SUPP}(\ \text{  }\ )} = (3/5)/(4/5) = 75\%$$

CONFIDENCE TELLS US HOW STRONG THE ASSOCIATION IS

THE APRIORI ALGORITHM

HELPS US MINE BASKET DATA FOR RULES



EACH RULE HAS TO SATISFY A MINIMUM

SUPPORT THE PROPORTION OF BASKETS WHICH CONTAIN THE ITEMS ON ONE SIDE OF THE RULE

AND

CONFIDENCE THE PROPORTION OF BASKETS WHICH CONTAIN THE ITEMS ON BOTH SIDES OF THE RULE AS COMPARED TO LEFT SIDE ALONE

THE GOAL IS TO FIND ALL RULES THAT SATISFY THE SUPPORT AND CONFIDENCE CONSTRAINTS

APRIORI DOES THIS IN AN EFFICIENT WAY WITHOUT GENERATING ALL POSSIBLE RULES

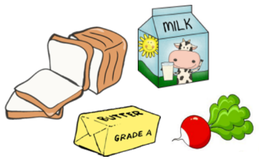
IT STARTS BY GENERATING 2 ITEM RULES, THEN 3 ITEM RULES AND SO ON

AT EACH STEP, IT WILL DROP THE ITEMS WHICH DON'T HAVE ENOUGH SUPPORT

THE APRIORI ALGORITHM

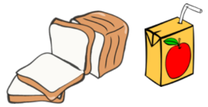
HELPS US MINE BASKET DATA FOR RULES

BASKET 1

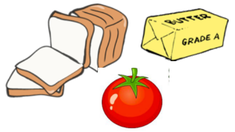


BASKET 3

BASKET 2

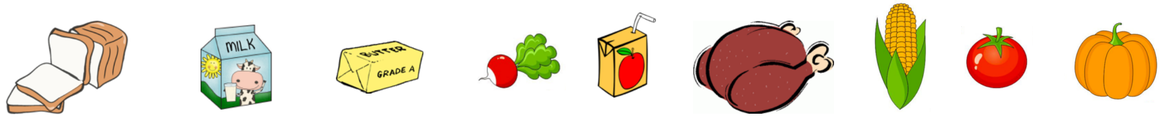


BASKET 4



BASKET 5

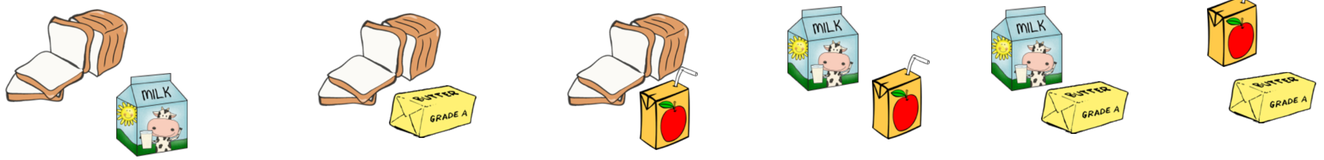
1. FIND ALL POSSIBLE SINGLE ITEM SETS



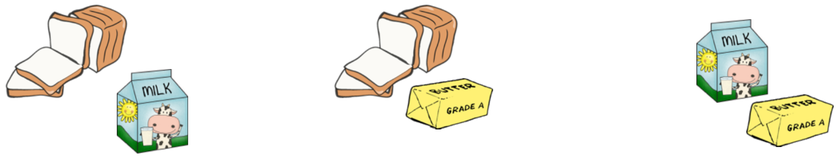
2. FILTER ONLY FOR THOSE ITEM SETS WITH THE MINIMUM SUPPORT

LET'S SAY MINIMUM SUPPORT=2/5

3. FIND ALL POSSIBLE 2 ITEM SETS FROM THIS SET



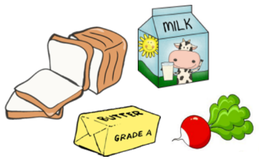
4. FILTER ONLY FOR THOSE ITEM SETS WITH THE MINIMUM SUPPORT



THE APRIORI ALGORITHM

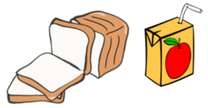
HELPS US MINE BASKET DATA FOR RULES

BASKET 1



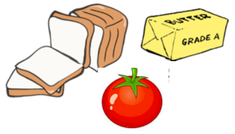
BASKET 3

BASKET 2

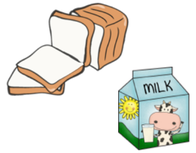


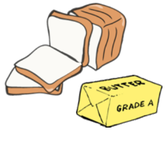
BASKET 5

BASKET 4



4. FILTER ONLY FOR THOSE ITEM SETS WITH THE MINIMUM SUPPORT





5. NOW FIND ALL POSSIBLE RULES

50%

100%

75%

75%

100%

50%

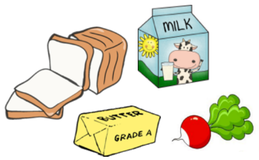
6. FILTER ONLY FOR THOSE RULES WITH THE MINIMUM CONFIDENCE

LET'S SAY MINIMUM CONFIDENCE=75%

THE APRIORI ALGORITHM

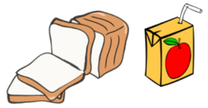
HELPS US MINE BASKET DATA FOR RULES

BASKET 1

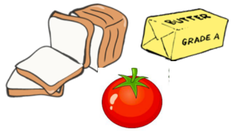


BASKET 3

BASKET 2



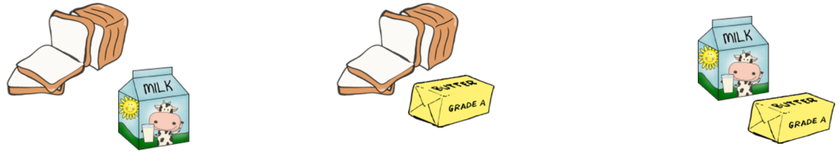
BASKET 4



BASKET 5

ADD THESE RULES TO THE LIST OF "LEARNT" RULES AND MOVE ON

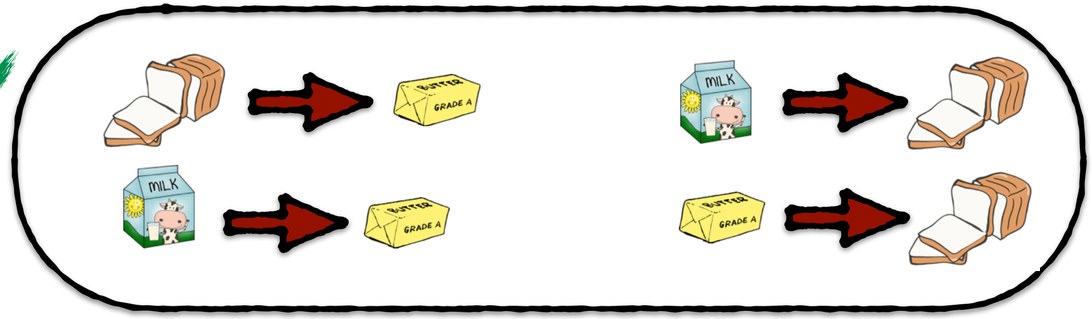
4. FILTER ONLY FOR THOSE ITEM SETS WITH THE MINIMUM SUPPORT



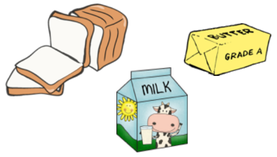
5. NOW FIND ALL POSSIBLE RULES

6. FILTER ONLY FOR THOSE RULES WITH THE MINIMUM CONFIDENCE

LET'S SAY MINIMUM CONFIDENCE=75%



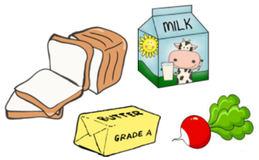
7. NOW FIND ALL 3 ITEM SETS FROM REMAINING ITEMS IN STEP 4



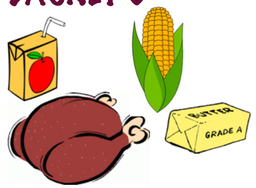
THE APRIORI ALGORITHM

HELPS US MINE BASKET DATA FOR RULES

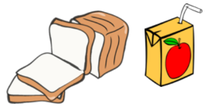
BASKET 1



BASKET 3

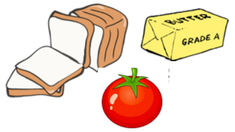


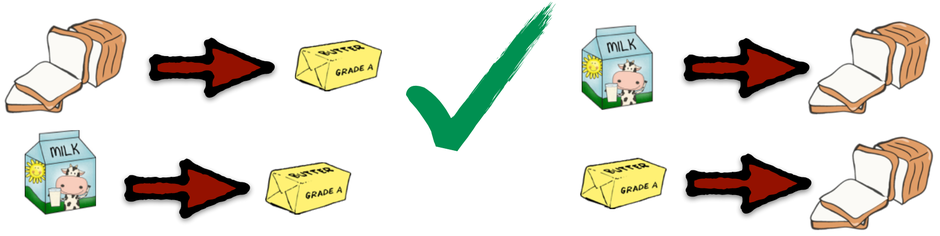
BASKET 2



BASKET 5

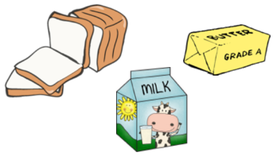
BASKET 4



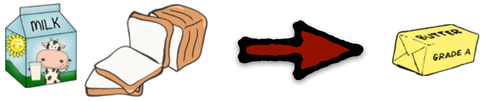


7. NOW FIND ALL 3 ITEM SETS FROM REMAINING ITEMS IN STEP 4

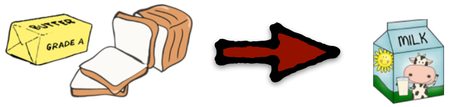
8. FILTER FOR THOSE WHERE MINIMUM SUPPORT IS SATISFIED (2/5)



9. NOW FIND ALL POSSIBLE RULES



100%



66%

100%

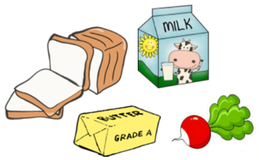
10. FILTER ONLY FOR THOSE RULES WITH THE MINIMUM CONFIDENCE

(MINIMUM CONFIDENCE=75%)

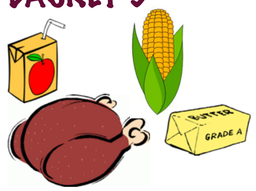
THE APRIORI ALGORITHM

HELPS US MINE BASKET DATA FOR RULES

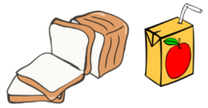
BASKET 1



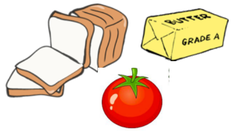
BASKET 3



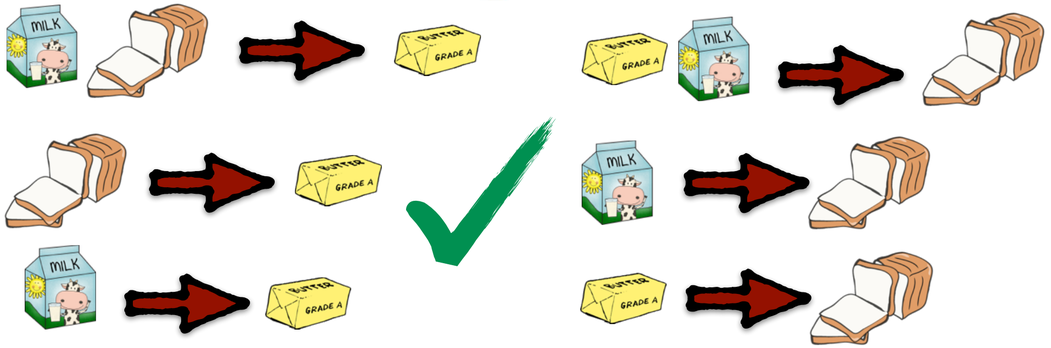
BASKET 2



BASKET 4

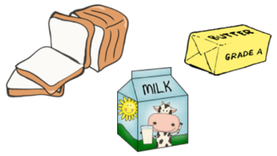


BASKET 5



7. NOW FIND ALL 3 ITEM SETS FROM REMAINING ITEMS IN STEP 4

8. FILTER FOR THOSE WHERE MINIMUM SUPPORT IS SATISFIED (2/5)



9. NOW FIND ALL POSSIBLE RULES

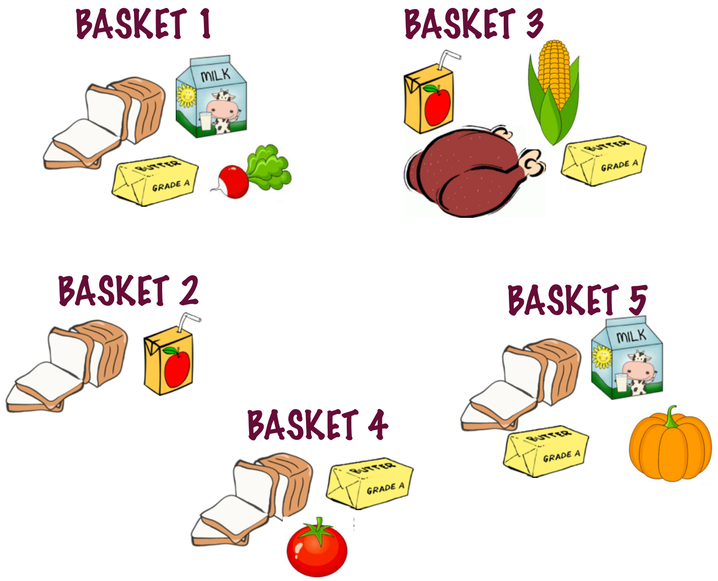
10. FILTER ONLY FOR THOSE RULES WITH THE MINIMUM CONFIDENCE

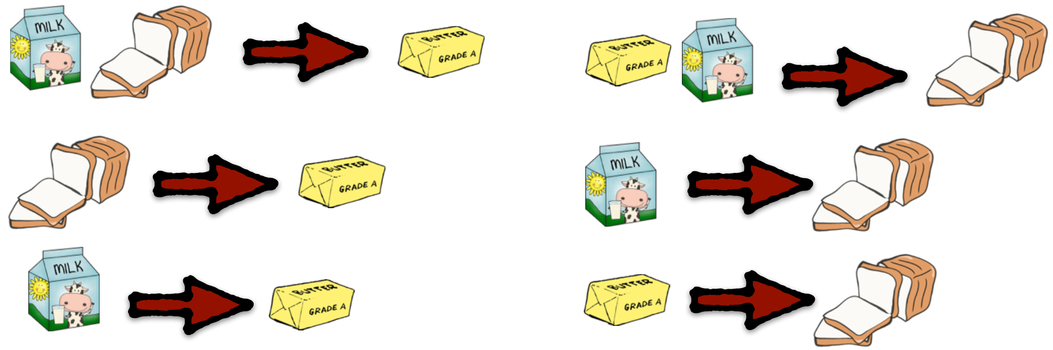
(MINIMUM CONFIDENCE=75%)

STOP WHEN YOU CAN'T MAKE ANY BIGGER ITEM SETS

THE APRIORI ALGORITHM

HELPS US MINE BASKET DATA FOR RULES





THE APRIORI ALGORITHM IS VERY EFFICIENT AT FINDING ASSOCIATION RULES

ONE KEY CHALLENGE IS WHEN THERE ARE TOO FEW ITEM SETS WHICH SATISFY THE MINIMUM SUPPORT

THIS MIGHT LEAD TO MISSING OUT ON STRONG ASSOCIATIONS BECAUSE OF THE LACK OF SUPPORT

RECOMMENDATION ENGINES NORMALLY USE ONE OR MORE OF THESE TECHNIQUES

CONTENT-BASED FILTERING

COLLABORATIVE FILTERING

ASSOCIATION RULES