

# Business Data Analytics

MTAT.03.319

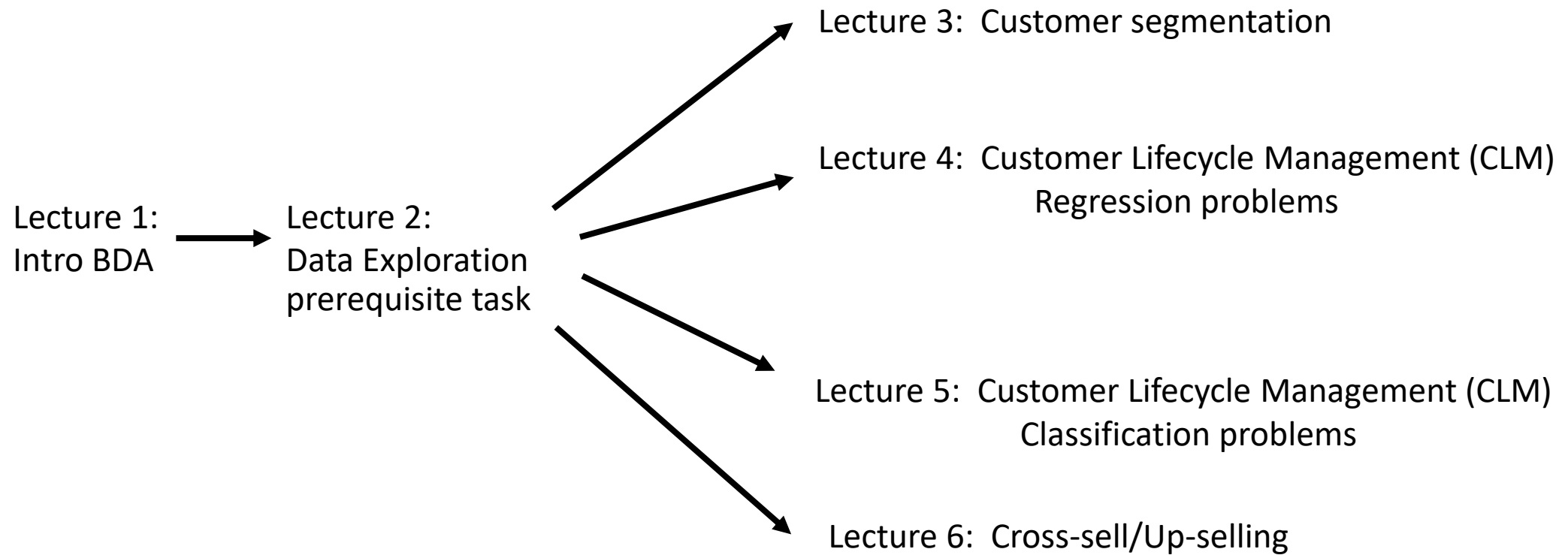


Lecture 6

Rajesh Sharma

<https://css.cs.ut.ee/>

# Till now and today !



# Cross Selling & Upselling

How to sell more ?

Here is a simple but powerful  
rule - always give people more  
than they expect to get.

Nelson Boswell

# Tips

- You already know about following tips
  - *“The cost of acquiring a new customer is often around 4 times more expensive than it is to sell to an existing customer. “*
- So better to sell it to existing customers.
- But how ?
  - *The most successful business practices to achieve this are by **up-selling** and **cross-selling**.*

# Fast Food Seller

Would you like potatoes  
with that ? 😊



**Cross Selling**

# Cross Selling: Amazon Shopping

← → ↻ Secure | <https://www.amazon.com/BLU-Advance-Unlocked-Smartphone-Black/dp/B071GVJ3B1/ref:>

NEW & INTERESTING FINDS ON AMAZON EXPLORE

amazon


Cell Phones & Accessories phone

Departments - Your Amazon.com Black Friday Deals Week Gift Cards Registry Sell Help

Cell Phones & Accessories Center Phones Unlocked Phones Prime Exclusive Phones Accessories Cases Wearable Technology Best Sellers Deals Track-In All Electronics

**\$10 & Under with FREE Shipping**

Back to search results for "phone"



Roll over image to zoom in


**BLU**  
**BLU Advance A5 -Unlocked Dual Sim Smartphone -Black**  
★★★★☆ 85 customer reviews | 78 answered questions  
Price: **\$49.99**

**In Stock.**  
This item ships to **Esteska, Want it Friday, Nov. 24?** Order within **6 hrs 39 mins** and choose **AmazonGlobal Priority Shipping** at c  
Ships from and sold by Amazon.com. Gift-wrap available.

Offer Type: **Advance A5**

Advance 5.2 Advance A4 **Advance A5** Advance A5 PRO Advance A6


Color: **Black**



- 5.0" Capacitive touchscreen display: 5MP Main Camera with flash and 2MP front Camera with flash
- 8GB Internal memory 512MB RAM Micro SD up to 64GB
- MediaTek 1.3 GHz Dual core 6570 processor with Mali-400 GPU
- GSM Quad band 850/900/1700/1900/2100 US compatibility Nationwide on all GSM Networks including AT&T, T-Mobile, Cric
- Purchase any qualifying 2017 BLU Smartphone for your chance to win a trip of a lifetime! See below for more details

Compare with similar items

**Used & new (\$)** from \$42.49 & FREE shipping. Details



Under \$100

## Frequently bought together



Total price: **\$57.78**

[Add both to Cart](#)

[Add both to List](#)

These items are shipped from and sold by different sellers. Show details

**This item:** BLU Advance A5 -Unlocked Dual Sim Smartphone -Black **\$49.99**

BLU Advance 5.0 case, Mady PU Leather Wallet Case for BLU Advance 5.0 Phone - Black **\$7.79**

# Cross Selling: Definition

- To sell related or complementary products to a new or existing customer.

<https://www.investopedia.com/terms/c/cross-sell.asp>

- Cross-selling is a sales technique used to get a customer to spend more by purchasing a product that's related to what's being bought already.

Source: <https://www.shopify.com/encyclopedia/cross-selling>

# UpSelling

**Listen free or subscribe to Spotify Premium.**

**Spotify Free**  
**\$0.00** /month

---

- ✓ Shuffle play
- ✓ Ad free
- ✓ Unlimited skips
- ✓ Listen offline
- ✓ Play any track
- ✓ High quality audio

---

**GET FREE**

**Spotify Premium**  
**\$9.99** / month  
Start your 30 day free trial\*

---

- ✓ Shuffle play
- ✓ Ad free
- ✓ Unlimited skips
- ✓ Listen offline
- ✓ Play any track
- ✓ High quality audio

---

**GET PREMIUM**



# Up-Selling

<

Fri 12/19/14  
From 1,269.82 \$

Sat 12/20/14  
From 1,269.82 \$

Sun 12/21/14  
From 902.82 \$

Mon 12/22/14  
From 872.82 \$

Tue 12/23/14  
From 872.82 \$

>

Sort by

Number of stops

Compare fares

Premium Economy Basic

Premium Economy Basic Plus

Premium Economy Flex

Business Basic

✈️

08:05 - 10:35

FRA - CHI 0 Stops 09h 30min

LH9152

i

- sold out -

- sold out -

- sold out -

☐ 3,692.82 \$

🕒

10:40 - 13:05

FRA - CHI 0 Stops 09h 25min

747-8 LH430

i

☒ 872.82 \$

☐ 1,089.82 \$

☐ 2,237.82 \$

☐ 3,692.82 \$

All fares include:

🍴

🧳

Rebooking fee

Refund

Mileage accrual \*

Mileage upgrade \*\*

✈️

€

✈️

✕

miles

100%

miles

✕

✈️

€

✈️

€

miles

100%

miles

✕

✈️

✓

✈️

✓

miles

150%

miles

↗

✈️

€

✈️

✕

miles

150%

miles

✕

# Cross and Up Selling: Definition

- Up-selling: is a sales technique where a seller induces the customer to purchase more expensive items, upgrades or other add-ons in an attempt to make a more profitable sale.

Source: <https://en.wikipedia.org/wiki/Upselling>

- Cross-selling: To sell related or complementary products to a new or existing customer.

# Tips For Cross/Up selling?

## **Cross Selling**

- Peers Also Bought
- Incentives
- Discounted Second Buy
- Build A Relationship And Then Ask

## **Up Selling**

- Sell the benefits of the up-sell
- Keep The Up-Sell Below 25% Of The Original Order

# Return of Cross/Up Selling Strategy?

- Amazon reportedly attributes as much as 35 percent of its sales to cross-selling through its options on every product page
  - “customers who bought this item also bought” and
  - “frequently bought together”.



# Cross Selling | Up Selling

## Who?

- Identify the customer or a cluster for a better approach.
- Present relevant offers based on his buying history and/or social-demographics characteristics.

## What?

- Identify the products or services which best fit the buying situation.
- Constantly analyze buying behavior in order to identify new trends (predictive models)

## When, Where?

- Identify the best moment during the buying flow to offer another product or service.
- Respect the users main objective

## How?

- Identify the best position on the screen
- Identify the best model (text based, txt+img, advertising, radiobuttons, checkboxes, etc)

# Customer Lifecycle Management



# How to solve this problem ?

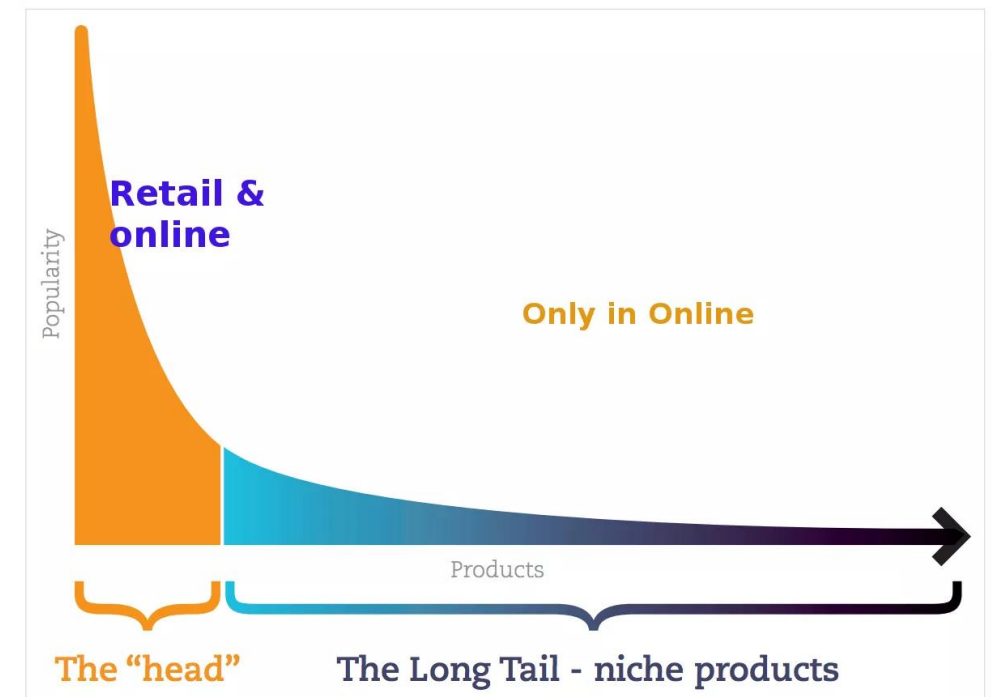


- Question: What products to recommend to whom?
- Solution: Recommendation Systems
- What products ?
  - Popularity
  - Market Basket Analysis
  - Collaborative filtering
- What products to whom ?
  - Collaborative filtering



# What products to recommend to whom? : Recommender Systems

Goal of a Recommender System:  
Identify products most relevant to the user (Eg. Top n offers).





# Recommendation Examples in Online Markets

## Platforms

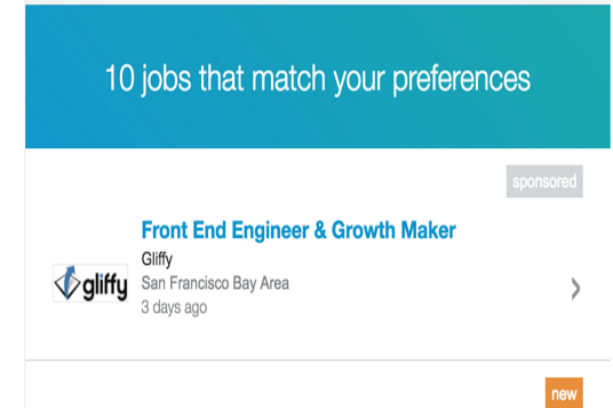


## Recommendations

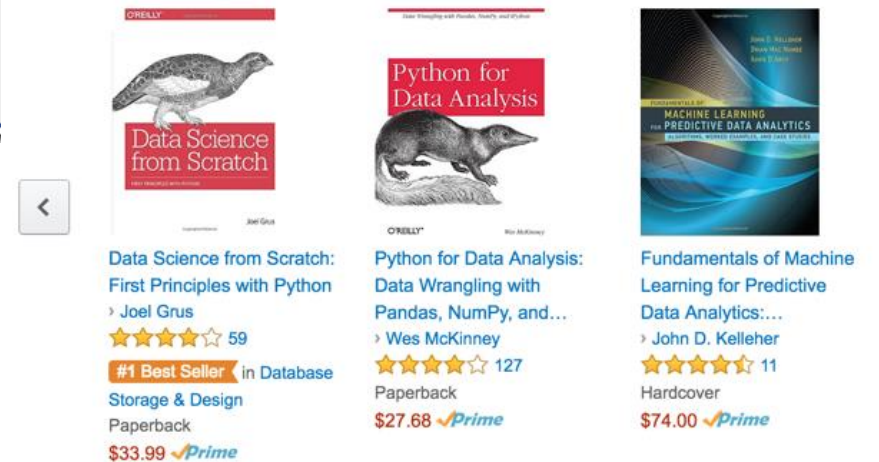
### Friends



## Jobs



### Customers Who Bought This Item Also Bought



## Books

# Solution 1: Popularity based Recommender System

Recommend items viewed/purchased by most people

Recommendations: Ranked list of items by their purchase count

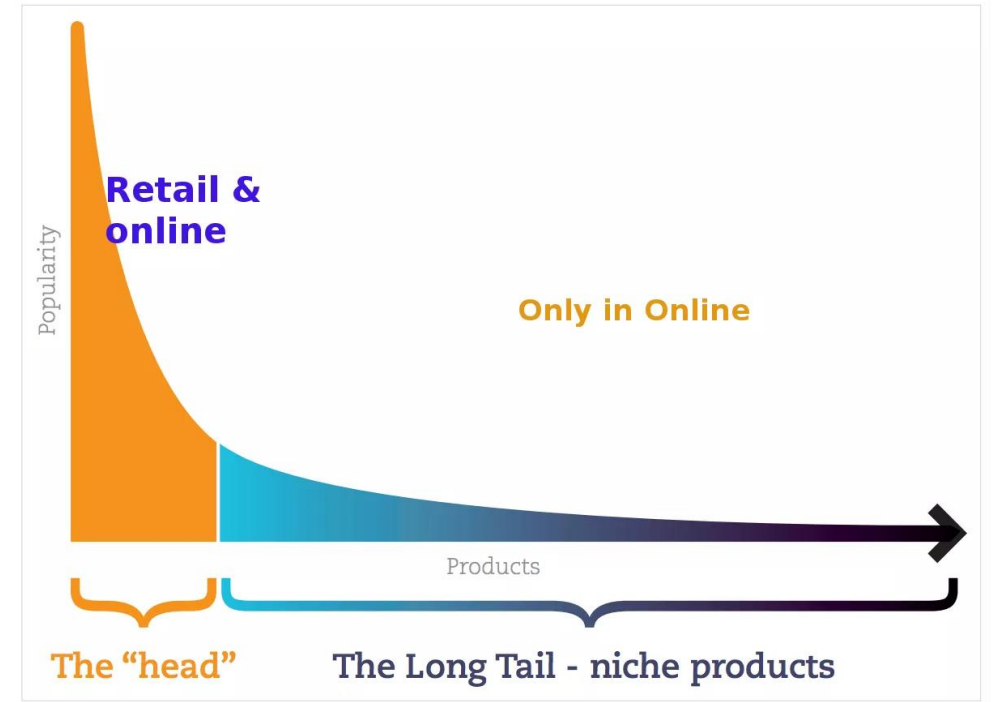
The screenshot shows the Google News homepage. At the top is the Google logo and a search bar. Below it, the 'News' section is visible with a 'U.S. edition' dropdown. The 'Most Popular' section features a large article titled 'Republican Convention Day 3: Trump Makes an Entrance' from the New York Times, dated 3 hours ago. The article text states: 'Day three of the Republican National Convention will feature expressions of support from several of the men who Donald J. Trump vanquished in the primaries.' Below the main article are three smaller video thumbnails from Newsy, CNN, and Voice of America. A 'See realtime coverage' button is also present. The 'Top Stories' section on the left lists categories like 'News near you', 'Suggested for you', 'World', 'U.S.', 'Elections', 'Business', 'Technology', 'Entertainment', 'Sports', and 'Science'.

The screenshot shows the Kohl's website. At the top is the Kohl's logo and a search bar. Below it, there are promotional banners for 'JUMPING BEANS' clothing and 'BATH TOWELS AND BATH RUGS'. A 'SHOP ALL JUMPING BEANS CLOTHING' link is visible. Below the banners, a 'See more recommendations' link is present. The 'POPULAR PRODUCTS' section displays five items with their sale prices and original prices:

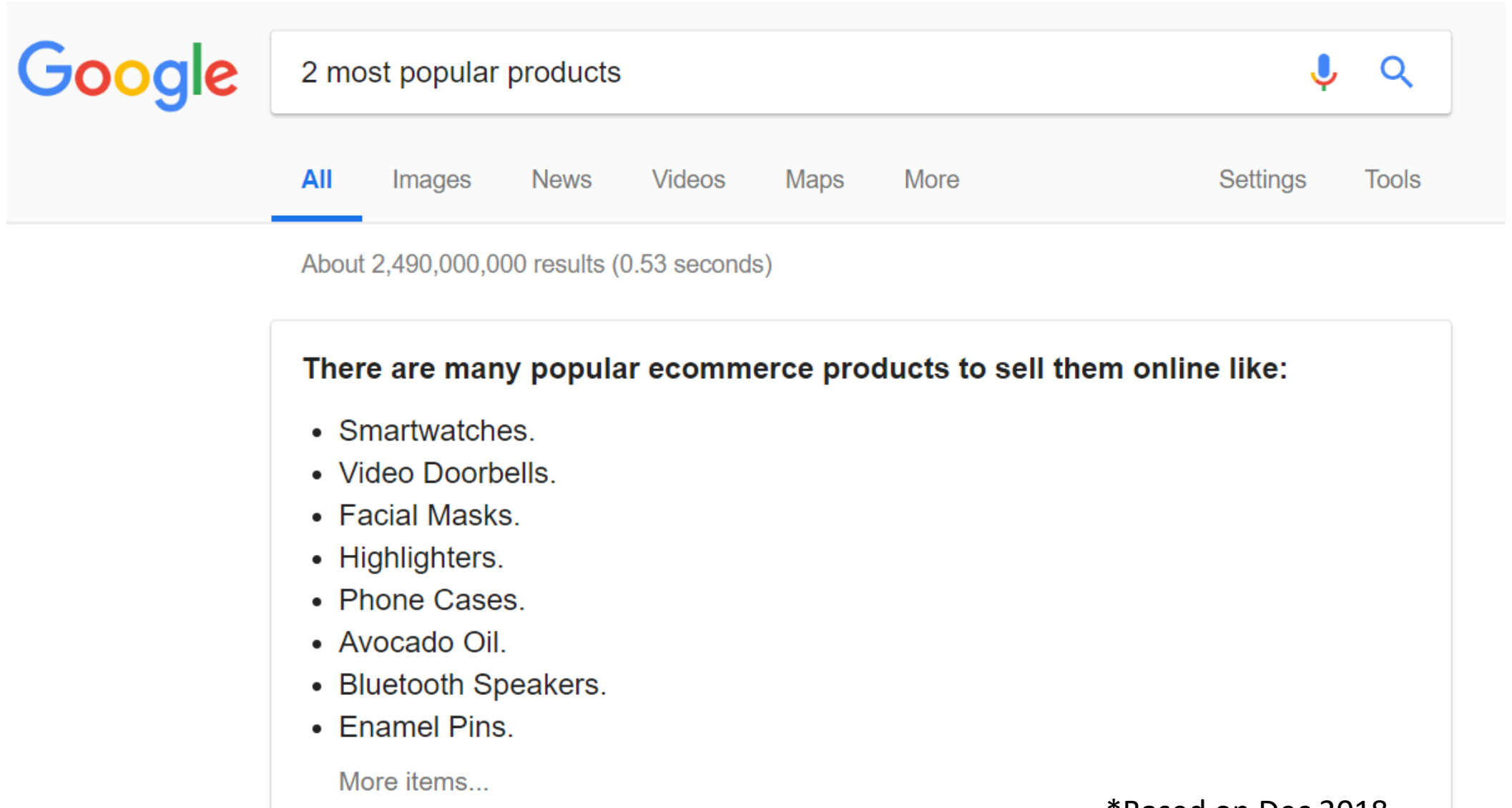
| Product                                      | Sale Price | Original Price | Rating | Count  |
|--|------------|----------------|--------|--------|
| Women's SONOMA Goods for Life™ Essential...  | \$4.99     | \$14.00        | ★★★★★  | (56)   |
| Plus Size SONOMA Goods for Life™...          | \$8.99     | \$18.00        | ★★★★☆  | (91)   |
| Women's SONOMA Goods for Life™ Juliette...   | \$17.99    | \$30.00        | ★★★★☆  | (40)   |
| The Big One® Microfiber Pillow               | \$4.99     | \$11.99        | ★★★★☆  | (1040) |
| Women's Croft & Barrow® Essential V-Neck Tee | \$6.99     | \$16.00        | ★★★★☆  | (94)   |

# Popularity follows almost Pareto Law

- 20% (highly valued customers) of customers bring 80% of profit
- 20% of products bring 80% of the profit
  - **But what about rest of the 80% products ?**  
Popularity based techniques are not helpful



# Popular products\* !



The image is a screenshot of a Google search interface. At the top left is the Google logo. To its right is a search bar containing the text "2 most popular products". To the right of the search bar are icons for voice search and a magnifying glass. Below the search bar is a horizontal menu with tabs: "All" (which is underlined in blue), "Images", "News", "Videos", "Maps", "More", "Settings", and "Tools". Below the menu, it says "About 2,490,000,000 results (0.53 seconds)". The main content area is a white box with a thin border. Inside, it starts with the text "There are many popular ecommerce products to sell them online like:". Below this is a bulleted list of eight items: Smartwatches, Video Doorbells, Facial Masks, Highlighters, Phone Cases, Avocado Oil, Bluetooth Speakers, and Enamel Pins. At the bottom of the list is the text "More items...".

Google

2 most popular products

All Images News Videos Maps More Settings Tools

About 2,490,000,000 results (0.53 seconds)

**There are many popular ecommerce products to sell them online like:**

- Smartwatches.
- Video Doorbells.
- Facial Masks.
- Highlighters.
- Phone Cases.
- Avocado Oil.
- Bluetooth Speakers.
- Enamel Pins.

More items...

\*Based on Dec 2018



Popularity is safe  
but



what about association  
among the products you  
recommend ?

?



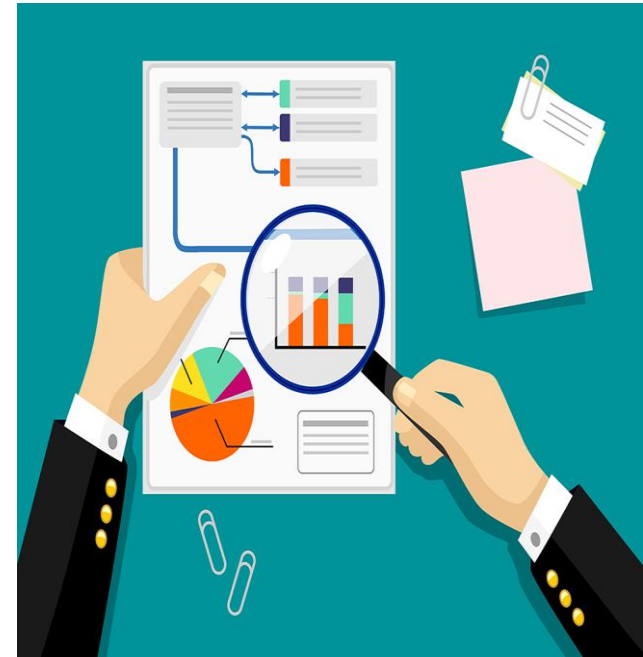
# Solution 2: Market Basket Analysis



Market



Basket



Analysis

MBA

# MBA put sense while recommending products



# Market Basket Analysis (MBA)

- Technique/Algorithm to identify the association rules from the data
- Input
  - List of purchases by customers over different visits
- Output
  - What items purchased together



# MBA: Terminologies

- **Items:** Objects that we are identifying associations between.
- Examples:
  - In a supermarket, each item is a product.
  - For a publisher, each item might be an article, a blog post, a video etc.
- A group of items is an item set.
  - $\{i_1, i_2, i_3 \dots, i_k\}$

# MBA: Terminologies Cont..

- **Items:** Objects that we are identifying associations between.
- **Transactions:** Transactions are instances of groups of items co-occurring together.
- Examples:
  - For an online retailer, a transaction is, generally, a group of items bought together
  - For a publisher, a transaction might be the group of articles read in a single visit to the website.
  - **NOTE:** It is up to the analyst to define over what period to measure a transaction.
- For each transaction, we have an item set.
  - $t_n = \{i_1, i_2, i_3, \dots, i_k\}$

# MBA: Terminologies Cont..

- **Items:** Objects that we are identifying associations between.
- **Transactions:** Transactions are instances of groups of items co-occurring together.
- **Rules:** are statements of the form
  - $\{i_1, i_2, i_3, \dots\} \Rightarrow \{i_k\}$
  - if you have the items in item set (on the left hand side (LHS) of the rule i.e.  $\{i_1, i_2, \dots\}$ ), then it is likely that a visitor will be interested in the item on the right hand side (RHS) i.e.  $\{i_k\}$ .
  - In the example above, rule would be:
    - $\{\text{flour, sugar}\} \Rightarrow \{\text{eggs}\}$

# MBA: Terminologies Cont..

- **Items:** Objects that we are identifying associations between.
- **Transactions:** Transactions are instances of groups of items co-occurring together.
- **Rules:** are statements of the form
  - $\{i_1, i_2, i_3, \dots\} \Rightarrow \{i_k\}$
  - if you have the items in item set (on the left hand side (LHS)) then the visitor will be interested in the item on the right hand side (RHS)
  - In the example above, rule would be:
    - $\{\text{flour, sugar}\} \Rightarrow \{\text{eggs}\}$



This is what MBA finds

# MBA: Terminologies Cont..

- **Items:** Objects that we are identifying associations between.
- **Transactions:** Transactions are instances of groups of items co-occurring together.
- **Rules:** Find associated items for sale
- **Output of MBA ?**

Market basket analysis is generally a set of rules, that we can then exploit to make business decisions (related to marketing or product placement, for example).

Association rules -> Generates rules  
Example: (X -> Y)

Market Basket -> Assigns business outcome to those rules  
Example: X,Y could be sold together

# What MBA tries to find ?

- MBA investigates if association between A and B (that is  $A \rightarrow B$ ) is
  - Random
  - Or there is some statistical basis of it
- Question: Can we come up with some quantitative metric for the above investigation ?
  - Answer: Lift
- Question: How to calculate Lift ?
- Answer: By using:
  - Support
  - Confidence
  - Expected Confidence

# MBA: Terminologies Cont..

- It works basically on following concepts
  - Support:  $\#(\text{co-occurrence of A and B})/T$ 
    - What is co-occurrence of two items name A and B
    - $(\text{Co-occurrence of A and B}) / \text{Total Transactions}$ .
  - Confidence:  $\#(A \text{ and } B) / \#(A)$ :
    - The proportion of transactions which contain A and also contain B.
    - How confident we are that B is present in presence of A.
    - Ratio of Support of (A and B), and Support of A.
  - Expected Confidence:
    - How confident we are that B is present in absence of A (or do not care about A).
    - $\# \text{ transactions where B is present} / \text{Total transactions}$

# MBA: Terminologies Cont..

- Lift:
  - Ratio of Confidence and Expected Confidence
  - Ratio of (B in presence of A) and (B in absence of A)
  - Explains the change in probability of B over “presence of A” and “absence of A”
  - Lift  $\leq 1$ 
    - A has no impact on B
  - Lift  $> 1$ 
    - Relationship between A and B is significant
    - Larger the lift ratio, the more significant the association.



# Retail Case Study

| Possible shopping Baskets (T) |  |
|-------------------------------|--|
| Transaction 1                 | Beer, Diaper, Chips, Aspirin             |
| Transaction 2                 | Diaper, Beer, Chips, Lotion, Juice, Milk |
| Transaction 3                 | Soda, Chips, Milk                        |
| Transaction 4                 | Soup, Beer, Diaper, Milk, Icecream       |
| Transaction 5                 | Soda, Coffee, Milk, Bread                |
| Transaction 6                 | Beer , Chips                             |

# Retail Case Study

## Possible shopping Baskets (T)

|               |  |
|---------------|--|
| Transaction 1 | Beer, Diaper, Chips, Aspirin             |
| Transaction 2 | Diaper, Beer, Chips, Lotion, Juice, Milk |
| Transaction 3 | Soda, Chips, Milk                        |
| Transaction 4 | Soup, Beer, Diaper, Milk, Icecream       |
| Transaction 5 | Soda, Coffee, Milk, Bread                |
| Transaction 6 | Beer , Chips                             |

## Frequent Items (based on Ms = 30)

(Beer, Diaper) : with support 50 %

(Beer, Chips) : with support 50 %

Support = (Co-occurrence of A and B)/ Total Transactions.

# Retail Case Study

## Possible shopping Baskets (T)

|               |  |
|---------------|--|
| Transaction 1 | Beer, Diaper, Chips, Aspirin             |
| Transaction 2 | Diaper, Beer, Chips, Lotion, Juice, Milk |
| Transaction 3 | Soda, Chips, Milk                        |
| Transaction 4 | Soup, Beer, Diaper, Milk, Icecream       |
| Transaction 5 | Soda, Coffee, Milk, Bread                |
| Transaction 6 | Beer , Chips                             |

## Frequent Items (based on Ms = 30)

(Beer, Diaper) : with support 50 %

(Beer, Chips) : with support 50 %

Support = (Co-occurrence of A and B)/ Total Transactions.

A is Beer and B is either Diaper or Chips

Confidence:  $\#(A \text{ and } B) / \#(A)$

Expected Confidence: # transactions where B is present / Total transactions

Lift = Confidence / Expected Confidence

Rule 1 Beer -> Diaper

Confidence =  $3/4$  , Expected Confidence =  $3/6$

Lift =  $(3/4) / (3/6) = 1.5$

Rule 2 Beer -> Chips

Confidence =  $3/4$ , Expected Confidence =  $4/6$

Lift =  $(3/4) / (4/6) = 1.1$

# Let us summarize about the **problem**\*

- Generate set of rules that link two or more products together.
- Each of these rules should have a lift greater than one.
- Also, we are interested in the support and confidence of those rules:
  - Higher confidence rules are ones where there is a higher probability of items on the RHS being part of the transaction given the presence of items on the LHS.
- Recommendations based on these rules to drive a higher response rate.
  - We're also better off **actioning** rules with higher support first, as these will be applicable to a wider range of instances.



\*Problem: How to find rules which can help us in finding patterns ?

# MBA is through association rules

- We have to generate rules
- 3 Types
  - Actionable Rules: On which you can take action.
  - Trivial Rules: Interesting and need to do more research on.
  - Inexplicable Rules: Complex or uncomprehensible (does not make sense)

Association rules -> Generates rules

Example:  $X \rightarrow Y$

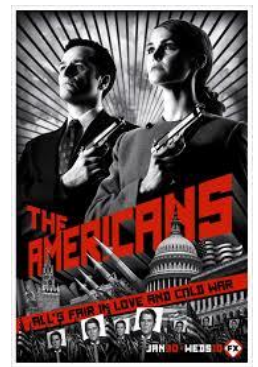
Market Basket -> Assigns business outcome to those rules

Example: X,Y could be sold together

# How to make Personalized recommendation



Users who liked “Love Simon” also liked following movies



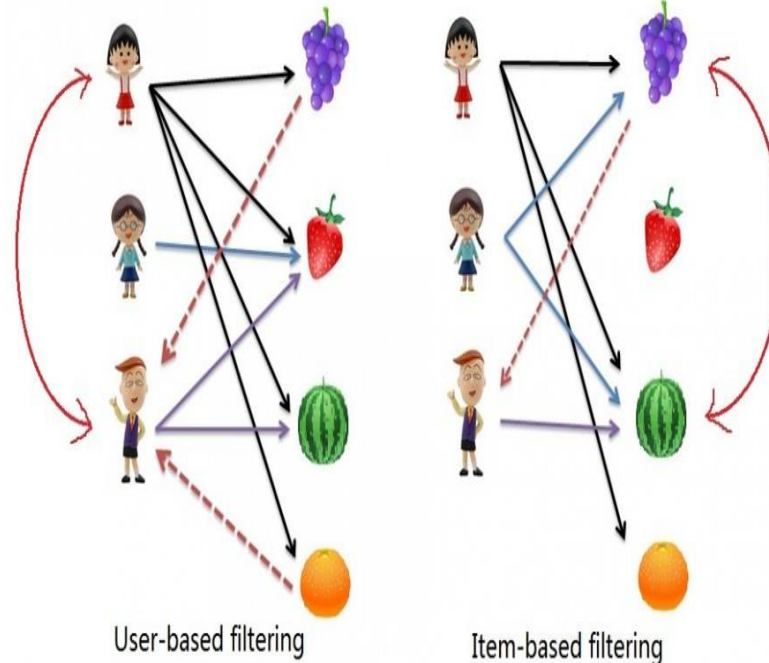
# Solution 3: Collaborative Filtering (CF)

## User-based

Find users who have a similar taste of products as the current user.

Similarity is based upon similarity in users' purchasing behaviour.

“User x is similar to user y because both purchased items A, B and C.”



## Item-based

Recommend items that are similar to the items the user bought.

Similarity is based upon co-occurrence of purchases.

“Items A and B were purchased by user x, so they are similar.”

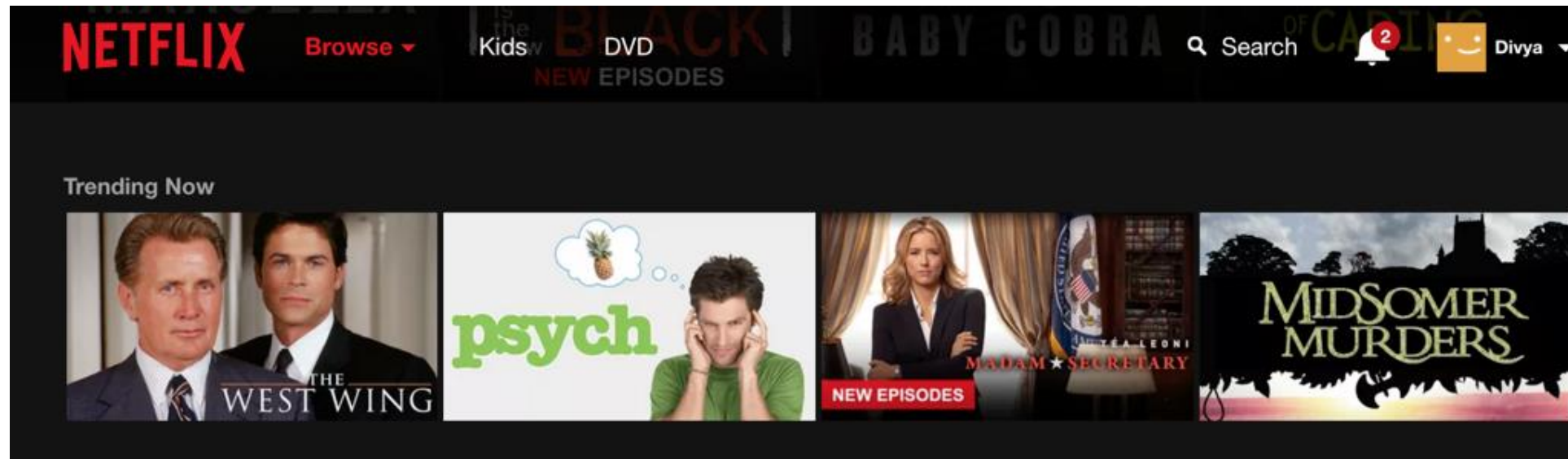


# Netflix Challenge !

DVD Rental  
Utilizing the Inventory



Grand prize  
of US\$1,000,000  
September 21, 2009





# User – User Collaborative Filtering

# Similar Users

|       |   | Movies |     |     |    |     |     |     |
|-------|---|--------|-----|-----|----|-----|-----|-----|
| Users |   | HP1    | HP2 | HP3 | TW | SW1 | SW2 | SW3 |
|       | A | 4      |     |     | 5  | 1   |     |     |
|       | B | 5      | 5   | 4   |    |     |     |     |
|       | C |        |     |     | 2  | 4   | 5   |     |
|       | D |        | 3   |     |    |     |     | 3   |

- Consider users  $x$  and  $y$  with rating vectors  $r_x$  and  $r_y$
- We need similarity metric  $\text{Sim}(x,y)$
- Capture the intuition that  $\text{Sim}(A,B) > \text{Sim}(A,C)$

# Similar Users: Jaccard Similarity

|       |   | Movies |     |     |    |     |     |     |
|-------|---|--------|-----|-----|----|-----|-----|-----|
| Users |   | HP1    | HP2 | HP3 | TW | SW1 | SW2 | SW3 |
|       | A | 4      |     |     | 5  | 1   |     |     |
|       | B | 5      | 5   | 4   |    |     |     |     |
|       | C |        |     |     | 2  | 4   | 5   |     |
|       | D |        | 3   |     |    |     |     | 3   |

- Jaccard similarity(A,B) =  $\frac{r_A \cap r_B}{r_A \cup r_B}$
- Jaccard distance =  $1 - \frac{r_A \cap r_B}{r_A \cup r_B}$
- Sim (A,B) = 1/5 ; Sim (A,C) = 2/4
  - Sim(A,B) < Sim(A,C) : Ignores the rating values

# Similar Users: Cosine Similarity

| Users | Movies |     |     |    |     |     |     |
|-------|--------|-----|-----|----|-----|-----|-----|
|       | HP1    | HP2 | HP3 | TW | SW1 | SW2 | SW3 |
|       | A      | 4   | 0   | 0  | 5   | 1   | 0   |
|       | B      | 5   | 5   | 4  | 0   | 0   | 0   |
|       | C      | 0   | 0   | 0  | 2   | 4   | 5   |
|       | D      | 0   | 3   | 0  | 0   | 0   | 3   |

**NOTE:** Fill empty values by 0

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|_2 \|\mathbf{B}\|_2} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$

$$\begin{aligned} \text{Cosine similarity(A,B)} &= \frac{4*5 + 0*5 + 0*4 + 5*0 + 1*0 + 0*0 + 0*0}{\text{Sqrt}(4^2 + 0^2 + 0^2 + 5^2 + 1^2 + 0^2 + 0^2) * \text{Sqrt}(5^2 + 5^2 + 4^2 + 0^2 + 0^2 + 0^2 + 0^2)} \\ &= 0.38 \end{aligned}$$

# Similar Users: Cosine Similarity

| Users | Movies |     |     |    |     |     |     |
|-------|--------|-----|-----|----|-----|-----|-----|
|       | HP1    | HP2 | HP3 | TW | SW1 | SW2 | SW3 |
|       | A      | 4   | 0   | 0  | 5   | 1   | 0   |
|       | B      | 5   | 5   | 4  | 0   | 0   | 0   |
|       | C      | 0   | 0   | 0  | 2   | 4   | 5   |
|       | D      | 0   | 3   | 0  | 0   | 0   | 3   |

**NOTE:** Fill empty values by 0

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|_2 \|\mathbf{B}\|_2} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$

- Cosine similarity(A,B) = Cos( $r_A$ ,  $r_B$ )
- -1 : dissimilar, 0: orthogonal; +1: similar
- Sim (A,B) = 0.38 ; Sim (A,C) = 0.32
  - Sim(A,B) > Sim(A,C) : but not much

**Problem: Treat missing values as negative**

# Similar Users: Centered Cosine

Normalized ratings by subtracting the row mean

Movies

Users

|   | HP1 | HP2 | HP3 | TW | SW1 | SW2 | SW3 | Avg. Rat |
|---|-----|-----|-----|----|-----|-----|-----|----------|
| A | 4   |     |     | 5  | 1   |     |     | 10/3     |
| B | 5   | 5   | 4   |    |     |     |     | 14/3     |
| C |     |     |     | 2  | 4   | 5   |     | 11/3     |
| D |     | 3   |     |    |     |     | 3   | 6/2 = 3  |

|   | HP1 | HP2 | HP3  | TW   | SW1  | SW2 | SW3 |
|---|-----|-----|------|------|------|-----|-----|
| A | 2/3 | 0   | 0    | 5/3  | -7/3 | 0   | 0   |
| B | 1/3 | 1/3 | -2/3 | 0    | 0    | 0   | 0   |
| C | 0   | 0   | 0    | -5/3 | 1/3  | 4/3 | 0   |
| D | 0   | 0   | 0    | 0    | 0    | 0   | 0   |

In each row, original value – Avg. Rat

Each row addition = 0

Ratings are centered around 0.

+ : users liked it

- : users did not like it

# Similar Users: Centered Cosine (2)

| Users | Movies |     |     |      |      |      |     |
|-------|--------|-----|-----|------|------|------|-----|
|       | HP1    | HP2 | HP3 | TW   | SW1  | SW2  | SW3 |
|       | A      | 2/3 | 0   | 0    | 5/3  | -7/3 | 0   |
|       | B      | 1/3 | 1/3 | -2/3 | 0    | 0    | 0   |
|       | C      | 0   | 0   | 0    | -5/3 | 1/3  | 4/3 |
|       | D      | 0   | 0   | 0    | 0    | 0    | 0   |

Also known as  
pearson correlation.

- $\text{Sim}(A,B) = 0.09$  ;  $\text{Sim}(A,C) = -0.56$ 
  - $\text{Sim}(A,B) \gg \text{Sim}(A,C)$  : but not much
- Captures intuition better
  - Missing ratings treated as “average”
  - Handles “tougher raters” and “easy raters”

# Rating Predictions

- Goal: Prediction for user  $X$  and item  $i$
- What we need:
  - Let  $r_X$  be the rating for the user  $X$ .
  - Let  $N$  be the set of  $k$  users most similar to  $X$ , who have rated item  $i$ .
- Option 1:  $r_{xi} = \frac{1}{k} \sum_{y \in N} r_{yi}$  (Average) For a neighbor  $y$  in  $(\in)$  the set  $N$
- Option2:  $r_{xi} = \frac{\sum_{y \in N} s_{xy} r_{yi}}{\sum_{y \in N} s_{xy}}$  (Weighted Average)  $s$  is the similarity of the user  $x$  and its neighbor  $y$



# Item – Item Collaborative Filtering

# Item – Item Collaborative Rating

- For item  $i$ , find other similar items.
- Estimate rating for item  $i$  based on ratings for similar items
- Can use some similarity metrics and prediction functions as in user-user model.

- $$r_{xi} = \frac{\sum_{j \in N(i:x)} s_{ij} r_{xj}}{\sum_{j \in N(i:x)} s_{ij}}$$

$s_{ij}$  : similarity of items  $i$  and  $j$

$r_{xi}$  : ratings of item  $i$  by the user  $x$

$N(i:x)$  : set of items similar to  $i$  , rated by user  $x$ .

# Item – Item Collaborative Filtering

|        | users |   |   |   |   |   |   |   |   |    |    |    |
|--------|-------|---|---|---|---|---|---|---|---|----|----|----|
|        | 1     | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| movies | 1     |   | 3 |   | ? | 5 |   |   | 5 |    | 4  |    |
|        | 2     |   | 5 | 4 |   |   | 4 |   |   | 2  | 1  | 3  |
|        | 3     | 2 | 4 |   | 2 |   | 3 |   | 4 | 3  | 5  |    |
|        | 4     |   | 2 | 4 | 5 |   |   | 4 |   |    | 2  |    |
|        | 5     |   | 4 | 3 | 4 | 2 |   |   |   |    | 2  | 5  |
|        | 6     | 1 | 3 |   | 3 |   |   | 2 |   |    | 4  |    |



- estimate rating of movie 1 by user 5

Ratings are between 1 to 5

Empty boxes: unknown rating

? : Estimate the rating of movie 1 by the user 5

Neighborhood (N) = 2

Select 2 movies similar to “movie 1” and rated by user 5.

|   | users |   |   |   |   |   |   |   |   |    |    |    |          |
|---|-------|---|---|---|---|---|---|---|---|----|----|----|----------|
|   | 1     | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | sim(1,m) |
| 1 | 1     |   | 3 |   | ? | 5 |   |   | 5 |    | 4  |    | 1.00     |
| 2 |       |   | 5 | 4 |   |   | 4 |   |   | 2  | 1  | 3  | -0.18    |
| 3 | 2     | 4 |   | 1 | 2 |   | 3 |   | 4 | 3  | 5  |    | 0.41     |
| 4 |       | 2 | 4 |   | 5 |   |   | 4 |   |    | 2  |    | -0.10    |
| 5 |       |   | 4 | 3 | 4 | 2 |   |   |   |    | 2  | 5  | -0.31    |
| 6 | 1     |   | 3 |   | 3 |   |   | 2 |   |    | 4  |    | 0.59     |

Remember N = 2

Select 2 movies similar to “movie 1” and rated by user 5.

$$r_{xi} = \frac{\sum_{j \in N(i:x)} s_{ij} r_{xj}}{\sum_{j \in N(i:x)} s_{ij}}$$

Sim = Pearson Coeff.

1) Subtract mean rating  $m_i$  from each movie i.

1)  $m_1 = (1+3+5+5+4)/5 = 3.6$

2) Row 1 = (-2.6, 0, -0.6, 0, 0, 1.4, 0, 0, 1.4, 0, 0.4, 0)

2) Compute Cosine similarities between rows

Weighted Average =  $(0.41*2 + 0.59*3)/(0.41+0.59)$

= 2.6

# User to User Vs Item to Item

- Item-Item outperforms User-User
- Users are more complex than Items
  - Sparse: Users have limited interests (in buying)
  - Not all users can have likes/interests about all the items
- Items are simple: example: limited genres.
- Item similarity makes more sense than Users similarity

# Evaluation

Diagram illustrating a user-movie rating matrix. The vertical axis is labeled **users** and the horizontal axis is labeled **movies**. The matrix contains numerical ratings for 10 users across 6 movies.

|   |   |   |   |   |   |
|---|---|---|---|---|---|
| 1 | 3 | 4 |   |   |   |
|   | 3 | 5 |   |   | 5 |
|   |   | 4 | 5 |   | 5 |
|   |   | 3 |   |   |   |
|   |   | 3 |   |   |   |
| 2 |   |   | 2 |   | 2 |
|   |   |   |   | 5 |   |
|   | 2 | 1 |   |   | 1 |
|   | 3 |   |   | 3 |   |
| 1 |   |   |   |   |   |

Diagram illustrating a user-movie rating matrix, similar to the one on the left, but with a portion of the data highlighted as **Test Data**. The vertical axis is labeled **users** and the horizontal axis is labeled **movies**. The matrix contains numerical ratings for 10 users across 6 movies. The last three columns (columns 4, 5, and 6) are shaded gray and contain question marks, indicating unknown or test data. An arrow points to this shaded region with the label **Test Data**.

|   |   |   |   |   |   |
|---|---|---|---|---|---|
| 1 | 3 | 4 |   |   |   |
|   | 3 | 5 |   |   | 5 |
|   |   | 4 | 5 |   | 5 |
|   |   | 3 |   |   |   |
|   |   | 3 |   |   |   |
| 2 |   |   | ? | ? | ? |
|   |   |   | ? | ? | ? |
|   | 2 | 1 | ? | ? | ? |
|   | 3 |   | ? | ? | ? |
| 1 |   |   | ? | ? | ? |

# If you analyse it as regression problem

- MAE
- Mean Square Error
- Root Mean Square Error

# If you analyse it as regression problem

- Problems of RMSE/MAE etc
  - Prediction diversity
  - Prediction context
  - Order of predictions
- Alternative: Precision at top-k
  - Percentage of predictions in the user's top-k withheld ratings



# How can we find similar items ?

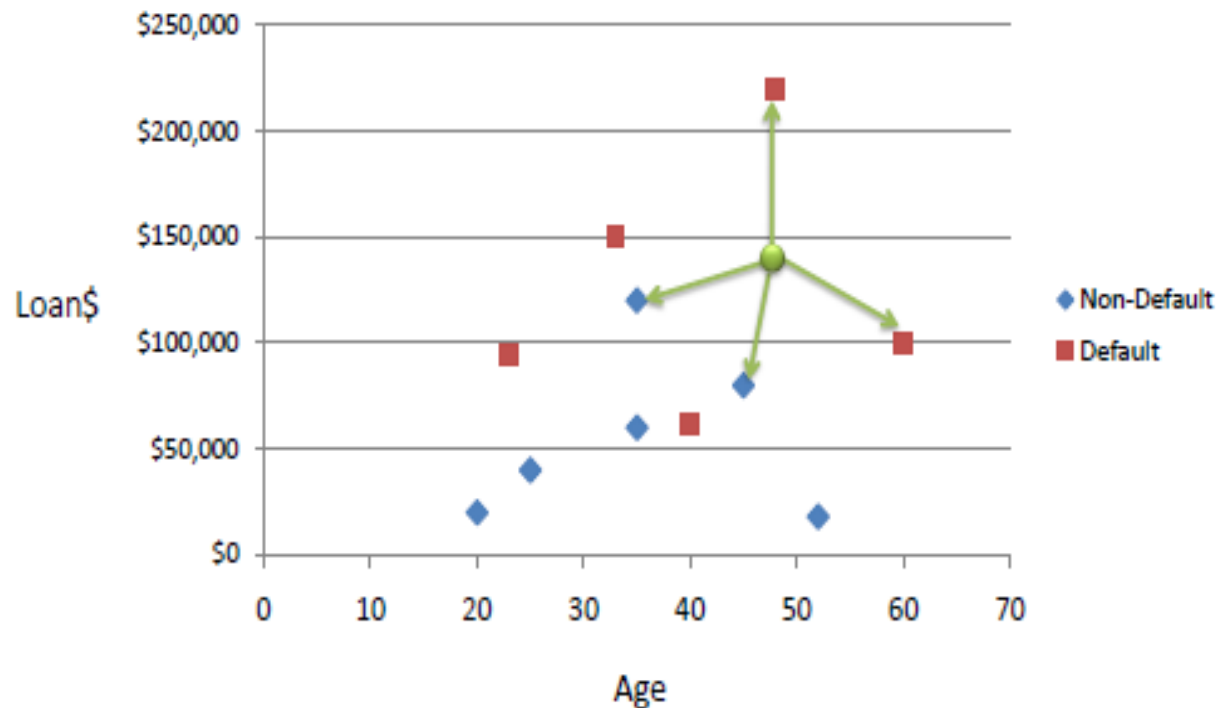
- You can use algorithm like KNN (K-Nearest Neighbor)
- It is a classification algorithm
- But basically you can use the idea of this algorithm for finding similar users ..

# Loan Default Problem using KNN

Loan and Age are two input parameters/features

Red and Blue are the training data points

Green is the test data point



Q1 : How to classify (or predict) about the gender of the **green** data ?

## Algorithm

- 1) Find K data points which are nearest to the green dot.
- 2) Assign (classify) the color (class) of the majority (among K neighbors) to the green (or the new data point)

Q2: How to find identify which K neighbors are nearest among all the neighbors ?

Answer: Use Distance metric

# Distance Metrics revisited

## Categorical

- Jaccard Distance: Ratio of Intersection/Union

- Hamming Distance

$$D_H = \sum_{i=1}^k |x_i - y_i|$$

$$x = y \Rightarrow D = 0$$

$$x \neq y \Rightarrow D = 1$$

| X    | Y      | Distance |
|------|--------|----------|
| Male | Male   | 0        |
| Male | Female | 1        |

## Continuous

Euclidean

$$\sqrt{\sum_{i=1}^k (x_i - y_i)^2}$$

Manhattan

$$\sum_{i=1}^k |x_i - y_i|$$

**NOTE:** Distance is (1- similarity)

# How to Pick the value of K ?

- Inspect your data to choose the value of K: Visualize your data
- Historically, the optimal value of K=3 to 10
- Large K, reduces the noise but no guarantee
- If K is odd : Avoid ties
- If K is even: Flip a coin (randomly assign) or do not assign the category
- Set aside a data from your training data to determine a good value of K

# K-NN

| Age | Loan      | Default | Distance |                          |
|-----|-----------|---------|----------|--------------------------|
| 25  | \$40,000  | N       | 102000   |                          |
| 35  | \$60,000  | N       | 82000    |                          |
| 45  | \$80,000  | N       | 62000    |                          |
| 20  | \$20,000  | N       | 122000   |                          |
| 35  | \$120,000 | N       | 22000    | <input type="checkbox"/> |
| 52  | \$18,000  | N       | 124000   |                          |
| 23  | \$95,000  | Y       | 47000    |                          |
| 40  | \$62,000  | Y       | 80000    | <input type="checkbox"/> |
| 60  | \$100,000 | Y       | 42000    |                          |
| 48  | \$220,000 | Y       | 78000    |                          |
| 33  | \$150,000 | Y       | 8000     | 1                        |
|     |           |         |          |                          |
| 48  | \$142,000 | ?       |          |                          |

Euclidean Distance

$$D = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

We can now use the training set to classify an unknown case (Age=48 and Loan=\$142,000) using Euclidean distance.

$$D = \text{Sqrt}[(48-33)^2 + (142000-150000)^2]$$

$$D = 8000.01$$

If K=1 then the nearest neighbor is the last case in the training set with Default=Y.

# K-NN

| Age | Loan      | Default | Distance |   |
|-----|-----------|---------|----------|---|
| 25  | \$40,000  | N       | 102000   |   |
| 35  | \$60,000  | N       | 82000    |   |
| 45  | \$80,000  | N       | 62000    |   |
| 20  | \$20,000  | N       | 122000   |   |
| 35  | \$120,000 | N       | 22000    | 2 |
| 52  | \$18,000  | N       | 124000   |   |
| 23  | \$95,000  | Y       | 47000    |   |
| 40  | \$62,000  | Y       | 80000    |   |
| 60  | \$100,000 | Y       | 42000    | 3 |
| 48  | \$220,000 | Y       | 78000    |   |
| 33  | \$150,000 | Y       | 8000     | 1 |
|     |           |         |          |   |
| 48  | \$142,000 | ?       |          |   |

Euclidean Distance

$$D = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

We can now use the training set to classify an unknown case (Age=48 and Loan=\$142,000) using Euclidean distance.

With K=3, there are two Default=Y and one Default=N out of three closest neighbors. The prediction for the unknown case is again Default=Y.

# Standardized Distance

| Age   | Loan | Default | Distance |
|-------|------|---------|----------|
| 0.125 | 0.11 | N       | 0.7652   |
| 0.375 | 0.21 | N       | 0.5200   |
| 0.625 | 0.31 | N       | 0.3160   |
| 0     | 0.01 | N       | 0.9245   |
| 0.375 | 0.50 | N       | 0.3428   |
| 0.8   | 0.00 | N       | 0.6220   |
| 0.075 | 0.38 | Y       | 0.6669   |
| 0.5   | 0.22 | Y       | 0.4437   |
| 1     | 0.41 | Y       | 0.3650   |
| 0.7   | 1.00 | Y       | 0.3861   |
| 0.325 | 0.65 | Y       | 0.3771   |
| 0.7   | 0.61 | ?       |          |

Standardized Variable

$$X_s = \frac{X - Min}{Max - Min}$$

If K = 1, then green dot = N

Remember: without non  
standardized, answer is Y

# Comparison of MBA & CF

## **Market Basket Analysis**

- Association Rule Mining
- Lacks the personalized approach
- Clustering problem
- Unsupervised approach
- No labeled data is provided
- Scalable
- No serendipity
- Mostly look for popular items

## **Collaborative Filtering**

- User-user or Item-Item Filtering
- Can be used for personalized recomm.
- More of a regression problem
- UnSupervised approach
- Labels (ratings etc) are provided.
- Computationally expensive
- Serendipity possible
- Looks at products in the long tail
- Cons: Cold start, Sparisity, First Rater, popularity bias



# Summary

- What is cross and up selling.
- Techniques
  - Popularity based
  - Market Basket Analysis
  - Collaborative Filtering
    - User-User CF
    - Item-Item CF
- K-NN algorithm