NETFLIX

Netflix movie recommendation engine

Team Bob

date 2019-11

Yanke Chen
Heeyeon Lee
Jacob
Jean M
Danial

CONTENTS

1

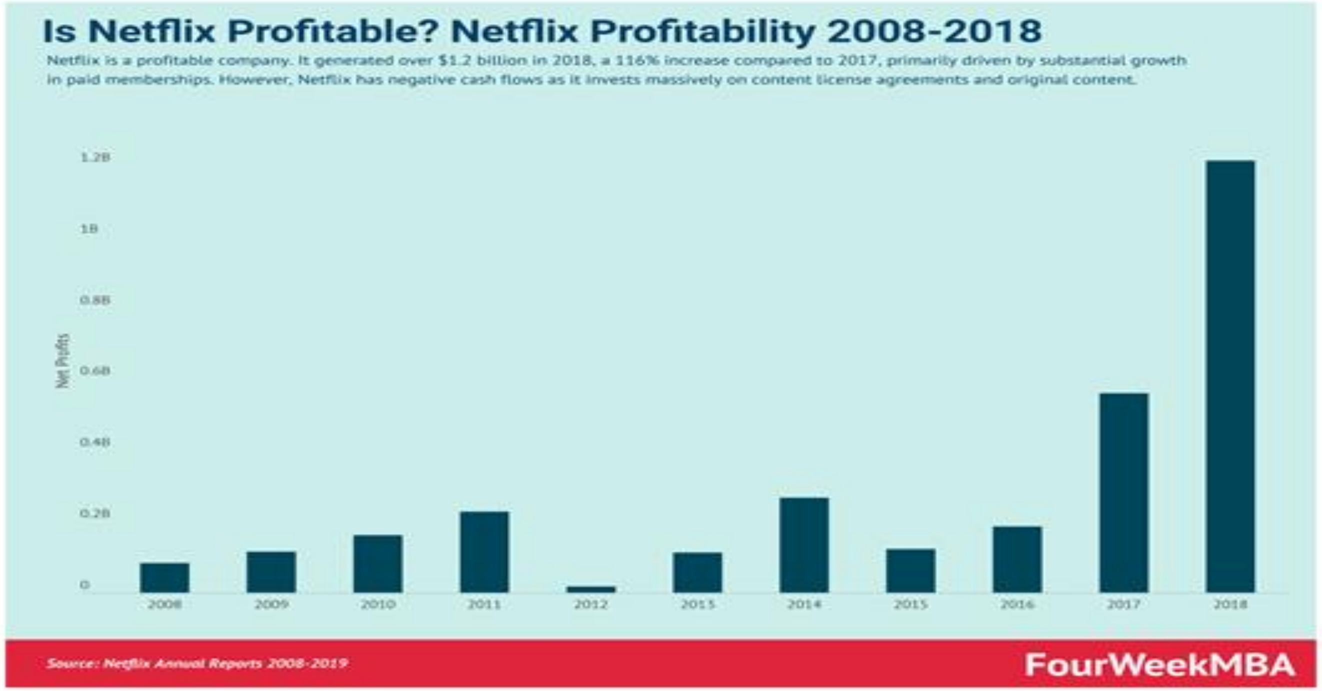
Part 1: Introduction

- Netflix is the biggest movie and TV streaming services in recent times

- Netflix enables subscribers to watch Movies, Documentaries, TV shows and more on an extensive variety of Internet-associated gadgets

- Netflix has 130 million worldwide streaming subscribers

- Netflix is a profitable company. It generated over $1.2 billion in 2018, a 116% increase compared to 2017



- In 2018 revenues drove profitability as they increased by 35%.

NETFLIX

- The essential source of income for Netflix business model is memberships.

- What makes Netflix leading in the Market is its personal recommendation system

- Netflix believes it could lose $1 billion or more every year from subscribers quitting its service if it weren't for its personalized recommendation engine.

NETFLIX

90 Seconds or Bust!

"Consumer research suggests that a typical Netflix member loses interest after perhaps 60 to 90 seconds of choosing, "

"The user either finds something of interest or the risk of the user abandoning our service increases substantially."

FACTS

- Netflix has negative cash flows as it invests massively on content license agreements and original content.
  –$1.7 billion for 2018

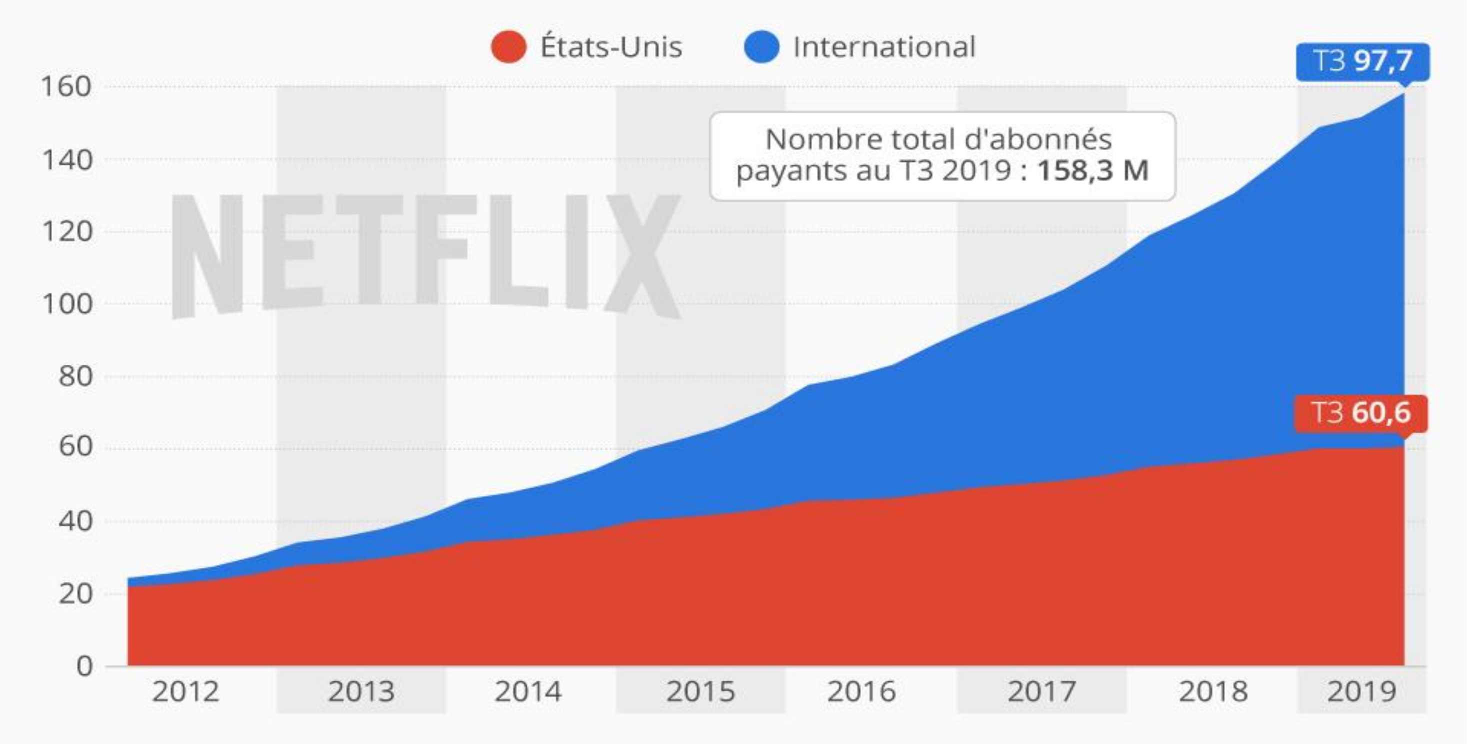
- With many competitors close behind, success isn’t guaranteed.

- Netflix had suffered its first loss of US subscribers and had failed to add its target of 5 million international subscribers in the first half of 2019.

2 Part 2: Market analysis & Business concept

Positioning

NETFLIX



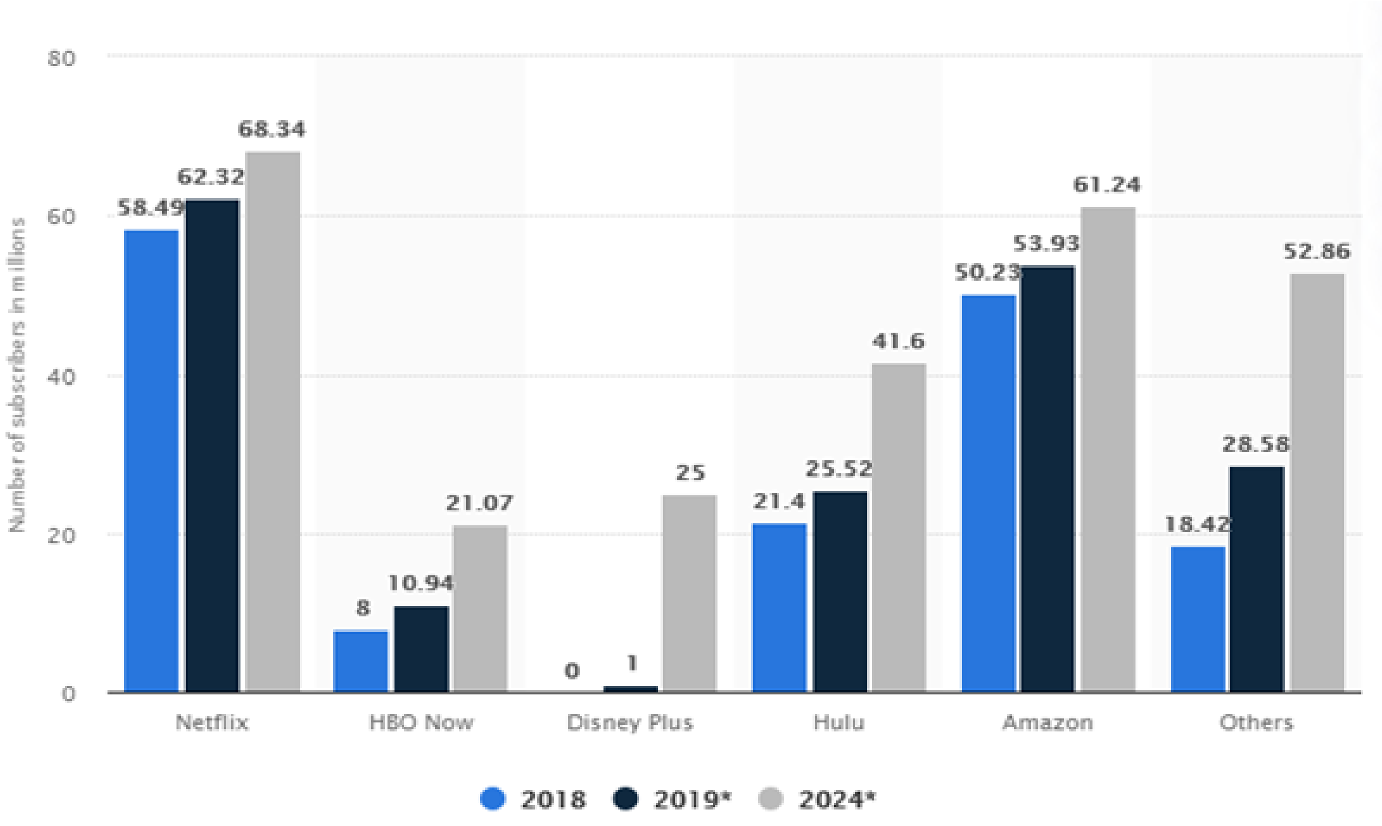
- **61%: international (190 contries)**
- **38%: number of subscribers in the USA**

TOTAL NUMBER OF NETFLIX SUBSCRIBERS IN THE WORLD (IN MILLIONS)
Source : Statista

Competitive Analysis

NETFLIX

Number of **subcribers**



Subscribers to SVOD services in the U.S. 2018-2024
Source: Statista

Competitive Analysis

NETFLIX

Total revenue generated by subscritions

	Subscribers (millions)	Average monthly cost (1)	Annual cost (3)	Annual subscription revenues (millions) (3)
Netflix	60.2	$12.66 (2)	$151.92	$7,646.6 (4)
Hulu	26.8	$8.99 (2)	$107.88	$2,891.2
Amazon Prime Video	26.0	$8.99	$107.88	$2,804.9
HBO Now	5.0	$14.99	$179.88	$899.4
CBS All Access	4.0	$7.99 (2)	$95.88	$383.5
Showtime	4.0	$10.99	$131.88	$527.5
Starz	3.0	$8.99	$107.88	$323.6
Sling TV	2.4	$30.00 (2)	$360.00	$871.2
Hulu with Live TV	2.0	$44.99	$539.88	$1,079.8
DirecTV Now	1.5	$52.50 (2)	$630.00	$945.0
YouTube Premium	1.5	$11.99	$143.88	$215.8
YouTube TV	1.0	$49.99	$599.88	$599.9
PlayStation Vue	0.8	$57.50 (2)	$690.00	$517.5
fuboTV	0.3	$66.24 (2)	$794.88	$198.7
Total	-	-	-	**$19,904.7**

Note: excludes advertising revenues; (1) cost does not take into account trial periods, introductory offers, special promotions or pricing plans for longer than one month; (2) average cost of different pricing plans; (3) eMarketer calculations excluding Netflix; (4) company reports
Source: company reports; eMarketer calculations, May 2, 2019

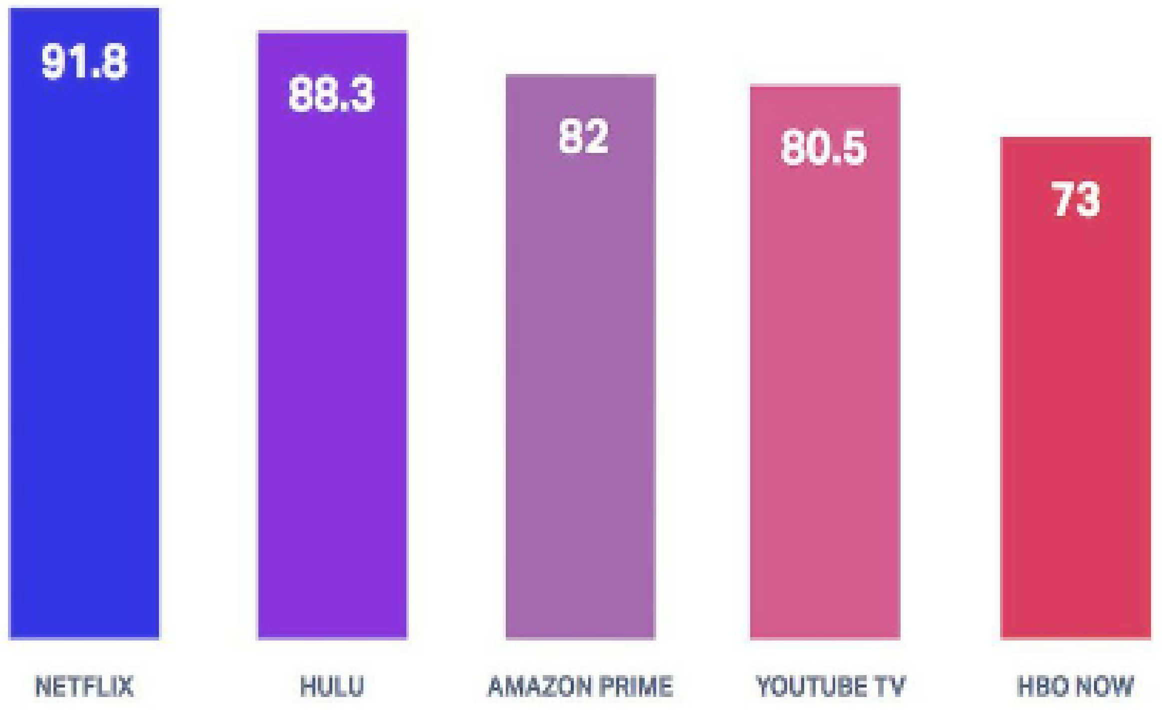
US SUBSCRIPTION VIDEO SERVICES REVENUE ESTIMATES 2018
Source: emarket

- **Netflix is the platform that made the most revenue in 2018 thanks to its high number of subscribers.**

Competitive Analysis

NETFLIX

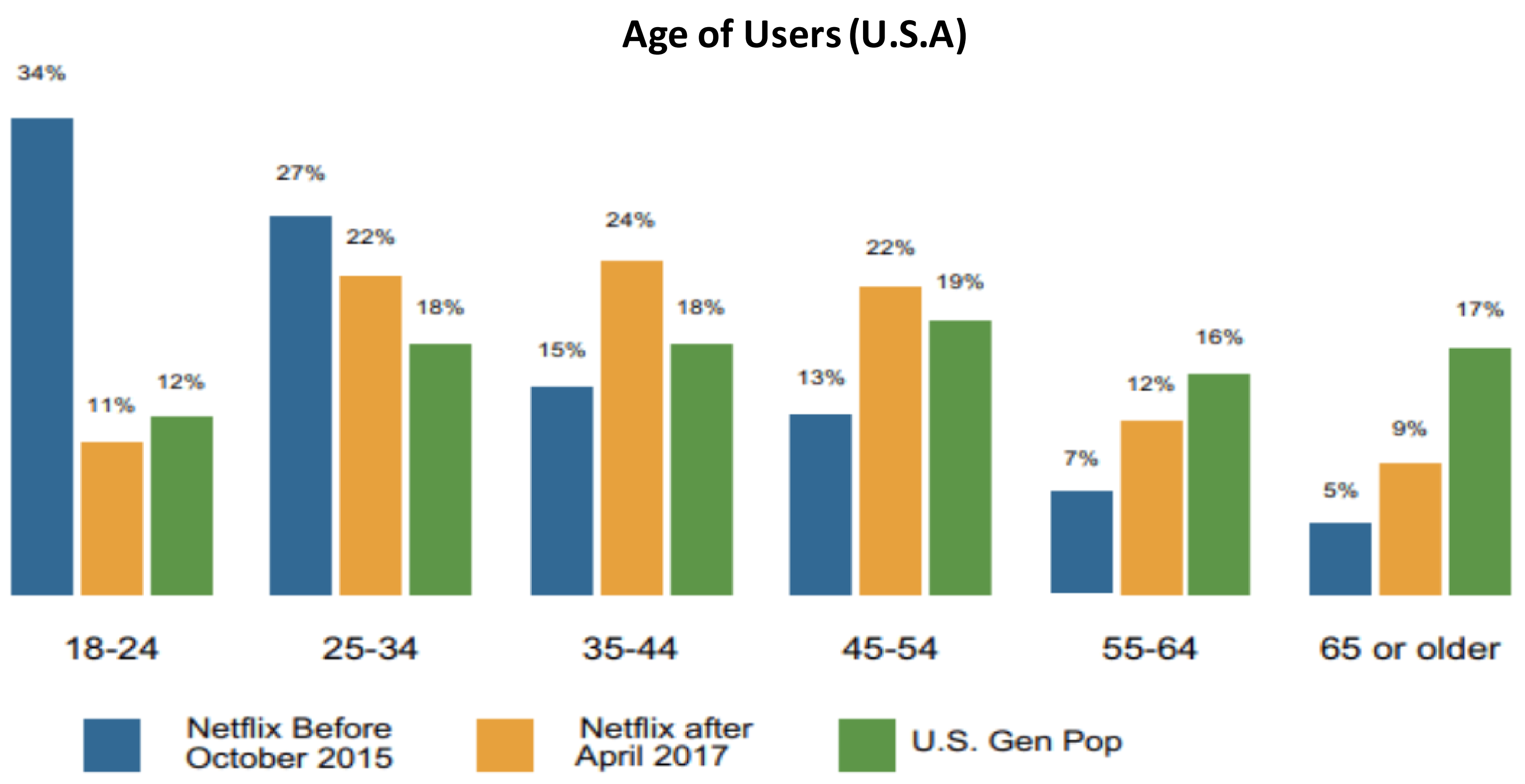
The ease of use



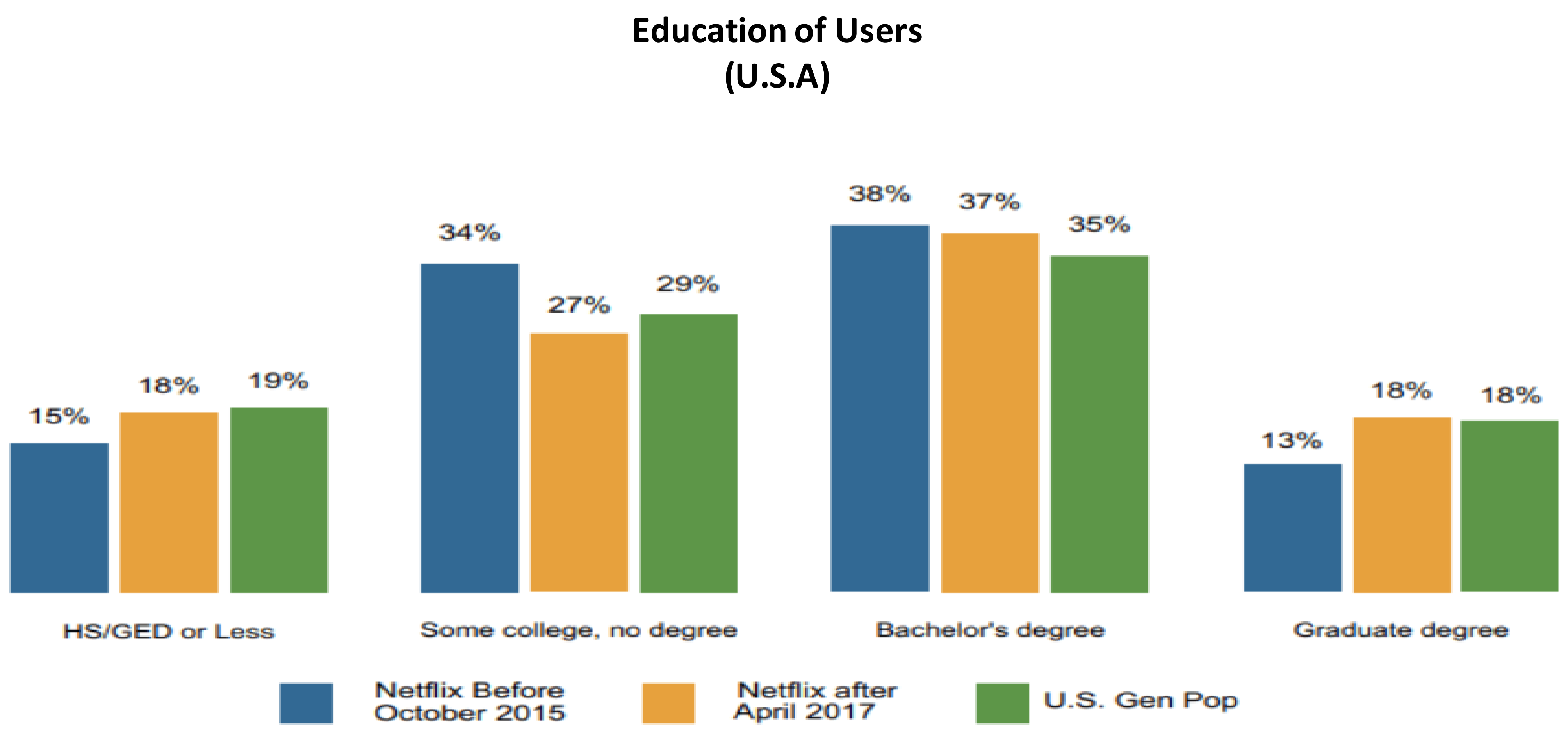
Ease of use for each streaming movies platform
Source: Usertesting

- Ease of use here means the speed at which information on the platform is accessed and the credibility of recommendation systems of diferrents platform.

Target Audience

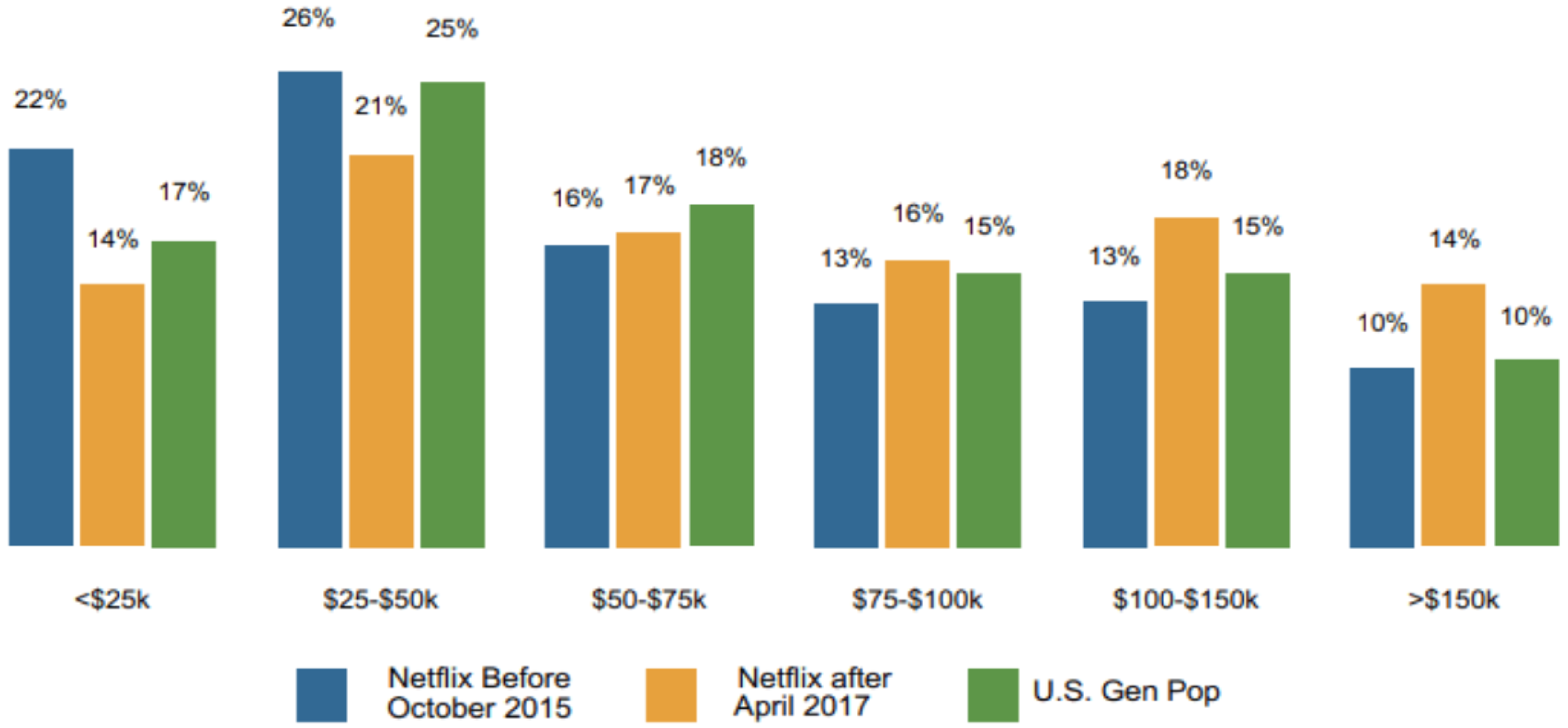


Target Audience



Target Audience

Average Income (U.S.A)



Type of users

How often do you watch streaming content on Netflix?

> All respondents in segment Blog - Streaming Video - Only two in my account
> Weighted according to U.S. Census figures for gender and age, 18 and older

A few times a week	1,966	59%
A few times a month	703	21%
A few times a year	303	9%
I don't use this service	307	9%
I don't use this service yet, but I'm plannin...	65	2%

Margin +/- 3%   3,344 responses from 02/06/2014 to 02/26/2018

- Time savers/Bingers
- Movie buffs
- Value seekers

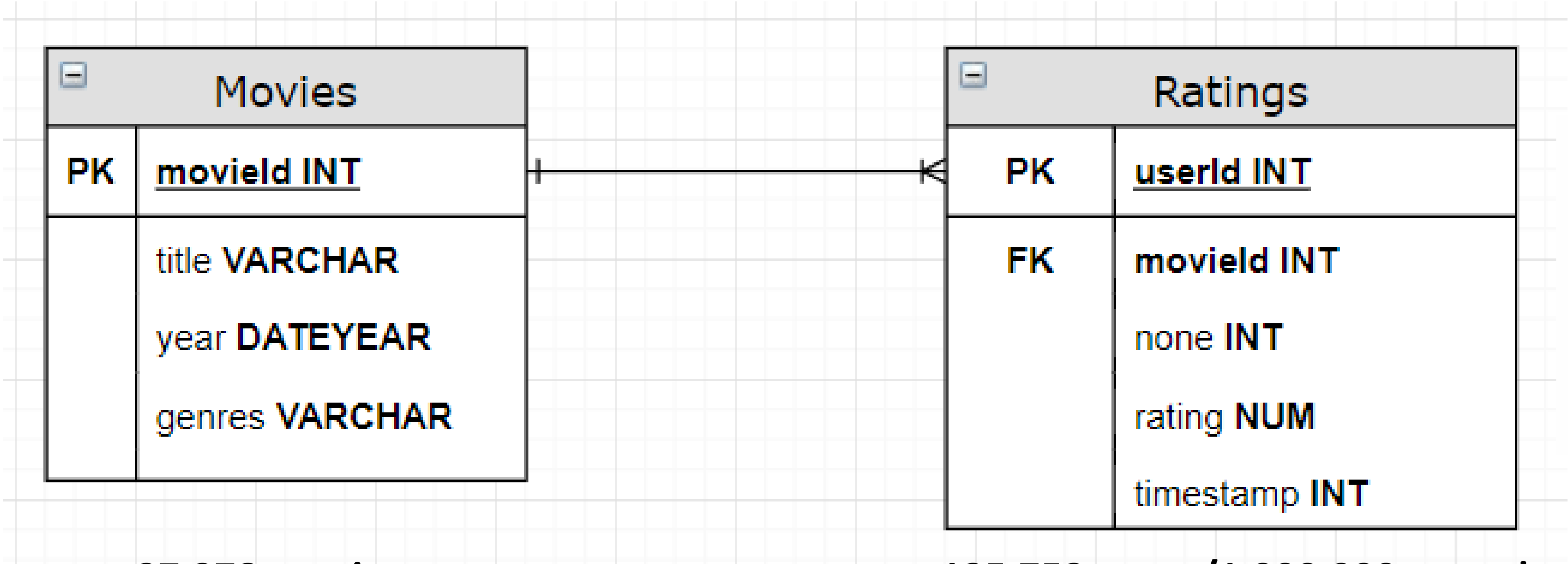
Value Proposition

NETFLIX

- Legal access to a huge movie and tv database with “best” personalized suggestion algorithm and no ads.
- Supported on wide range of devices.
- Releases new and exclusive series as full seasons rather than episode by episode.

3 Part 3: Descriptive Analysis & Data Manipulation

Data Source



27,278 movies

125,752 users/1,000,000 records

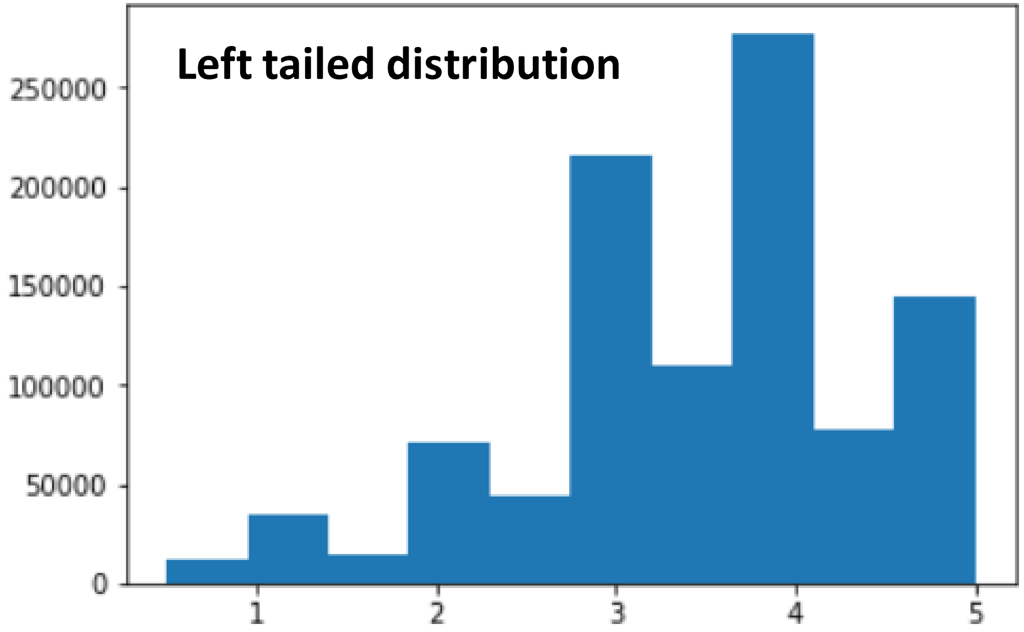
Entity Relationship Diagram of data set

Ratings file

Correlation analysis of ratings file

		userId	movieId	rating	timestamp
	1				
userId	0.999997	1			
movieId	-0.00056	-0.00059	1		
rating	0.001196	0.001192	0.003272	1	
timestamp	-0.00199	-0.00202	0.46037	0.000925	1

Ratings file



ratings average : 3.524071

```
In [39]: rating.describe()
```

Out[39]:

	userId	movieId	rating
count	1000000.000000	1000000.000000	1000000.000000
mean	68984.369969	9077.110015	3.524071
std	40030.083547	19840.493052	1.052555
min	1.000000	1.000000	0.500000
25%	34317.000000	903.000000	3.000000
50%	69064.000000	2174.000000	3.500000
75%	103508.000000	4798.000000	4.000000
max	138493.000000	131254.000000	5.000000

Movies file

NETFLIX

movieId	movies.title	Average of rating	Standard deviation of rating	userId
651	Superweib, Das (1996)	5.00	0.00	1
1462	Unforgotten: Twenty-Five Years After Willowbrook (1996)	5.00	0.00	1
1787	Paralyzing Fear: The Story of Polio in America, A (1998)	5.00	0.00	1
2223	Farmer's Wife, The (1928)	5.00	0.00	1
2564	Empty Mirror, The (1996)	5.00	0.00	1
2584	Foolish (1999)	5.00	0.00	1
2934	Amor brujo, El (Love Bewitched, A) (1986)	5.00	0.00	1
3575	Defying Gravity (1997)	5.00	0.00	1
3642	In Old California (1942)	5.00	0.00	1
3976	Stardom (2000)	5.00	0.00	1
4101	Dogs in Space (1987)	5.00	0.00	1
4601	Happy Together (1989)	5.00	0.00	1
4851	Things Behind the Sun (2001)	5.00	0.00	1
4895	Wash, The (2001)	5.00	0.00	1
5190	Inside Moves (1980)	5.00	0.00	1
5931	Britannia Hospital (1982)	5.00	0.00	1
6098	Marathon Family, The (Maratonci Trce Pocasni Krug) (1982)	5.00	0.00	1
6258	Dames du Bois de Boulogne, Les (Ladies of the Bois de Boulogne, The) (Ladies of the Park) (1945)	5.00	0.00	2
7200	Court-Martial of Billy Mitchell, The (1955)	5.00	0.00	1
7583	In This Our Life (1942)	5.00	0.00	1
7623	If You Only Knew (2000)	5.00	0.00	1
7912	Finder's Fee (2001)	5.00	0.00	2
8485	Samsara (2001)	5.00	0.00	1
8544	Now You See Him, Now You Don't (1972)	5.00	0.00	1
8619	Sister My Sister (1994)	5.00	0.00	1
8630	Such a Long Journey (1998)	5.00	0.00	1
8890	Alligator People, The (1959)	5.00	0.00	1
26073	Human Condition III, The (Ningen no joken III) (1961)	5.00	0.00	1
Total		**3.52**	**1.05**	**1000000**

5 star rated movies

movieId	movies.title	Average of rating	Standard deviation of rating	userId
296	Pulp Fiction (1994)	4.16	0.98	3420
356	Forrest Gump (1994)	4.02	0.97	3359
318	Shawshank Redemption, The (1994)	4.46	0.69	3218
593	Silence of the Lambs, The (1991)	4.17	0.84	3200
480	Jurassic Park (1993)	3.66	0.92	2906
260	Star Wars: Episode IV - A New Hope (1977)	4.21	0.90	2624
110	Braveheart (1995)	4.04	0.98	2607
589	Terminator 2: Judgment Day (1991)	3.96	0.89	2538
2571	Matrix, The (1999)	4.18	0.87	2514
1	Toy Story (1995)	3.92	0.89	2493
457	Fugitive, The (1993)	3.97	0.77	2484
150	Apollo 13 (1995)	3.87	0.88	2472
527	Schindler's List (1993)	4.30	0.83	2422
50	Usual Suspects, The (1995)	4.34	0.76	2400
1210	Star Wars: Episode VI - Return of the Jedi (1983)	4.00	0.91	2400
780	Independence Day (a.k.a. ID4) (1996)	3.36	1.07	2379
1196	Star Wars: Episode V - The Empire Strikes Back (1980)	4.20	0.87	2312
592	Batman (1989)	3.40	0.85	2309
47	Seven (a.k.a. Se7en) (1995)	4.07	0.86	2255
2858	American Beauty (1999)	4.17	0.84	2230
377	Speed (1994)	3.47	0.91	2191
858	Godfather, The (1972)	4.39	0.83	2181
380	True Lies (1994)	3.49	0.94	2179
590	Dances with Wolves (1990)	3.73	0.95	2175
1198	Raiders of the Lost Ark (Indiana Jones and the Raiders of the Lost Ark) (1981)	4.25	0.77	2168
32	Twelve Monkeys (a.k.a. 12 Monkeys) (1995)	3.89	0.86	2167
1270	Back to the Future (1985)	3.95	0.81	2146
608	Fargo (1996)	4.14	0.91	2138
588	Aladdin (1992)	3.67	0.92	2108
2959	Fight Club (1999)	4.24	0.87	2019
Total		**3.52**	**1.05**	**1000000**

Top viewed movies

Movies rated by more than 50 peoples = 3185

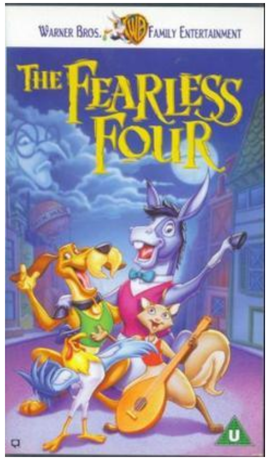
Movies rated by less than 50 peoples = 12,078

Percentage : 26.37%

Text mining

Not assigned value

Not assigned value in movie file



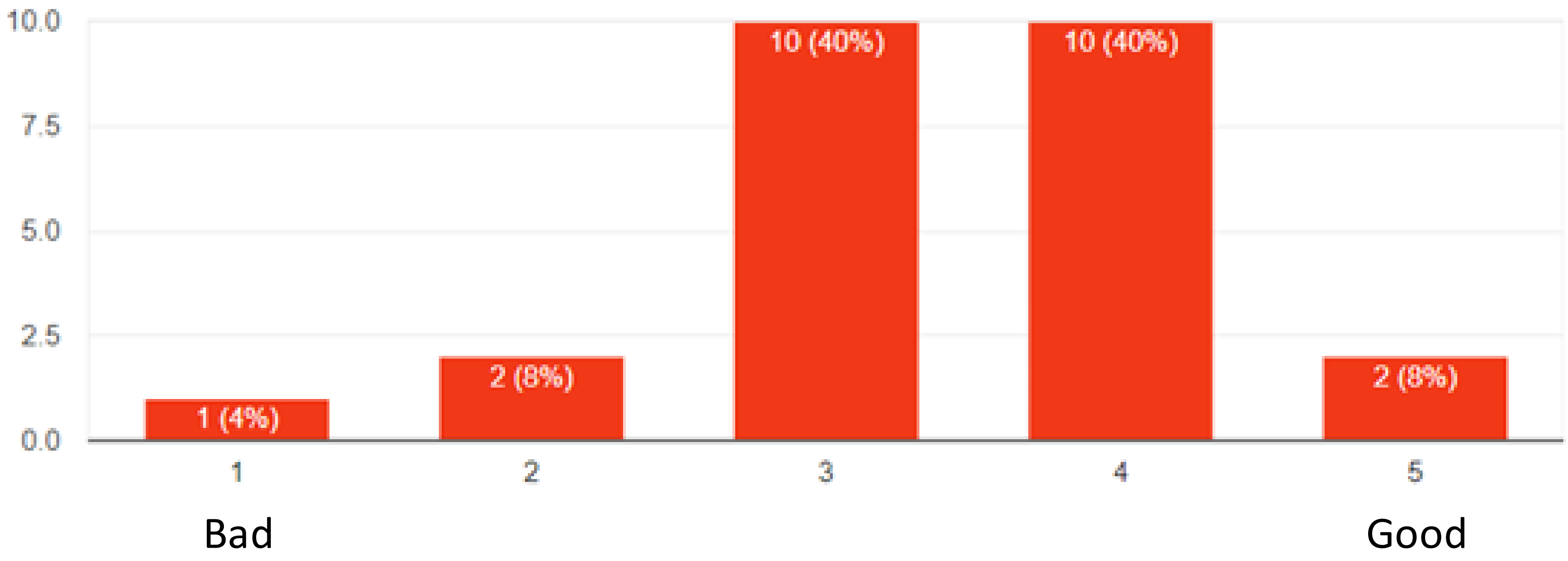
A number of movie which is not assigned genres : 246

- Title with the special characters
Ex)DÃ©sirÃ©, Beck - Ã
ga fÃ¶r Ã¶ga, Lilla JÃ¶nssonligan pÃ¥ styva linan, Tokyo FiancÃ©e

Customer behavior

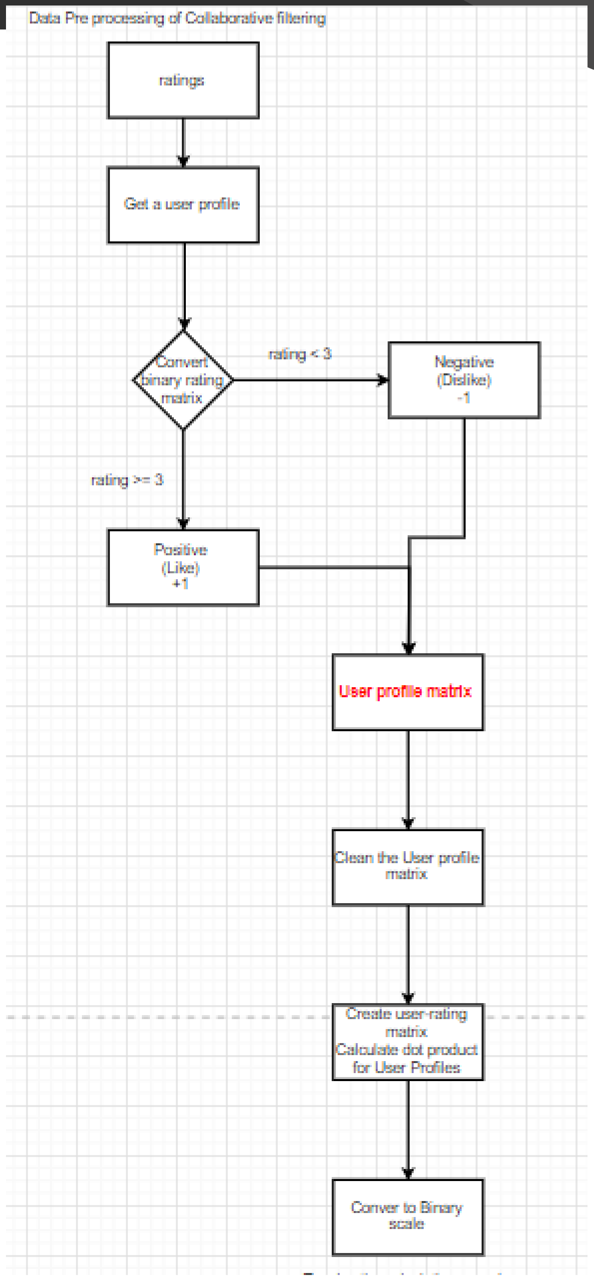
NETFLIX

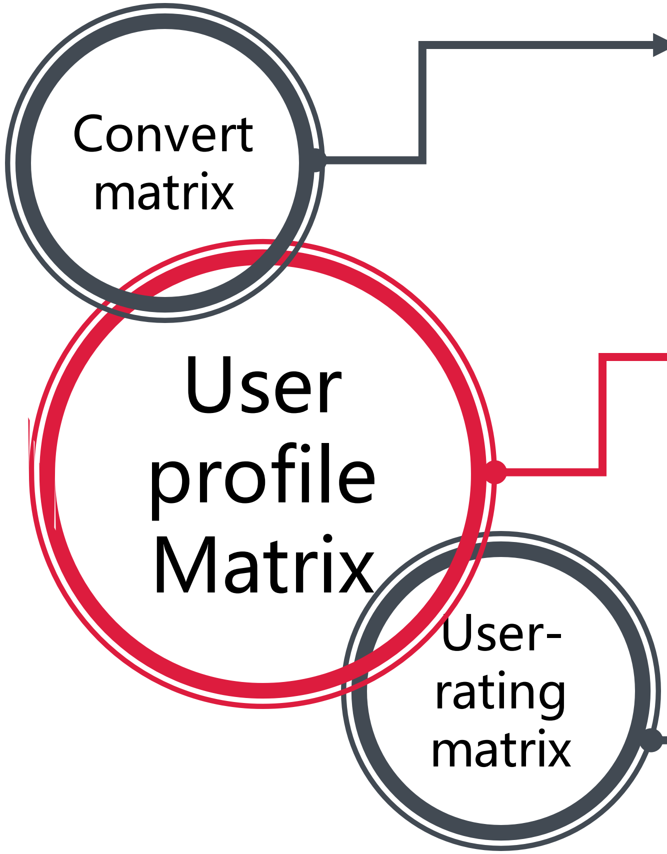
How likely are you to watch a movie with a recommendation of 3?



Data pre-processing

NETFLIX





Left tailed distribution & Customer behavior

Rating < 3 : Dislike -1

Rating >= 3 : Like +1

Not assigned value : none 0

	userId	movieId	rating
1	1	6	-1
2	1	22	-1
3	1	32	-1
4	1	50	1
5	1	110	1

Convert to binary scale to raise the calculating speed

	1	2	3	4	5	6	7	8	9	10
1	0	0	0	0	0	0	1	0	0	1
2	0	1	0	0	0	0	1	0	0	0
3	0	0	1	0	0	0	0	0	-1	0

4

Part 4: Algorithm

Challenges

Challenges still existing	descrption	Solution ideas
Data sparsity	In Netflix database, there are more than 48000 users and thousands items, but only 1% observation.	PCA- Dimensionality reduction;
Cold start	New users and new movies have no historical records	Hybrid
synonymy	Same genre/type in different word	Text mining, semantic analysis
Gray sheep	Someone has very special taste of movies	Hybrid
Shilling attack	Rating based on the subjective willingness (AntiSpam problem)	Hybrid
others	Privacy, noise, culture behaviour...	

Brainstorming idea

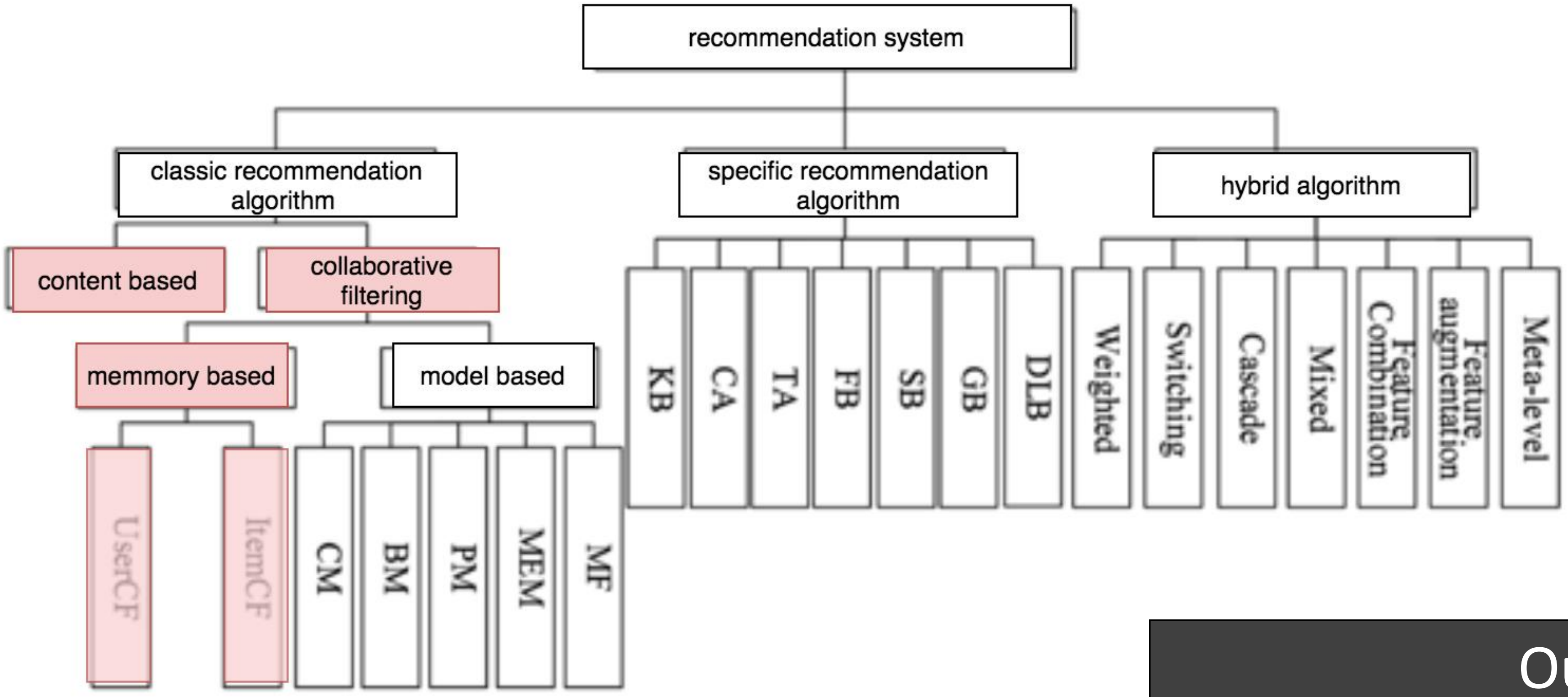


Figure 1. The classification of recommendation algorithms

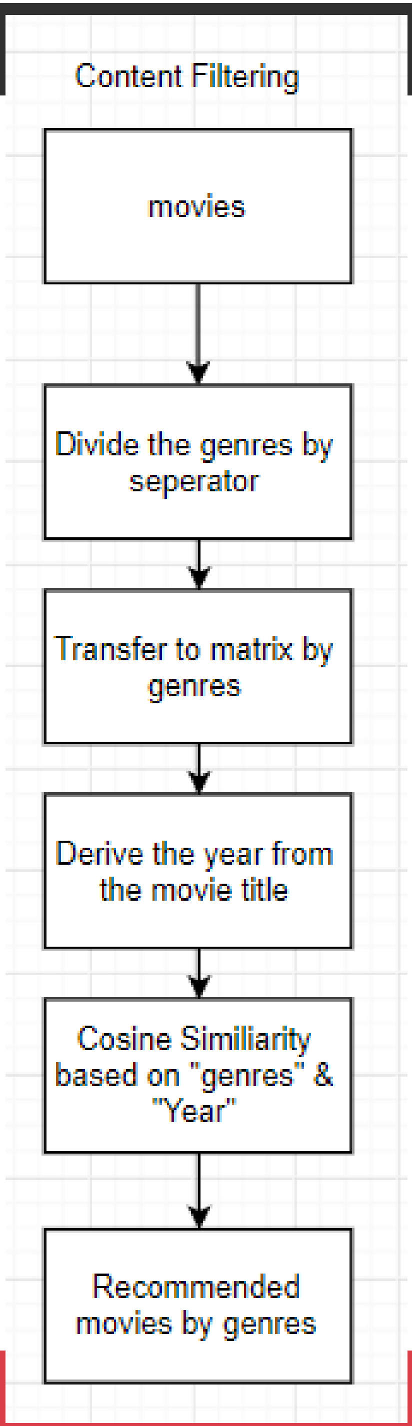
Our goal:

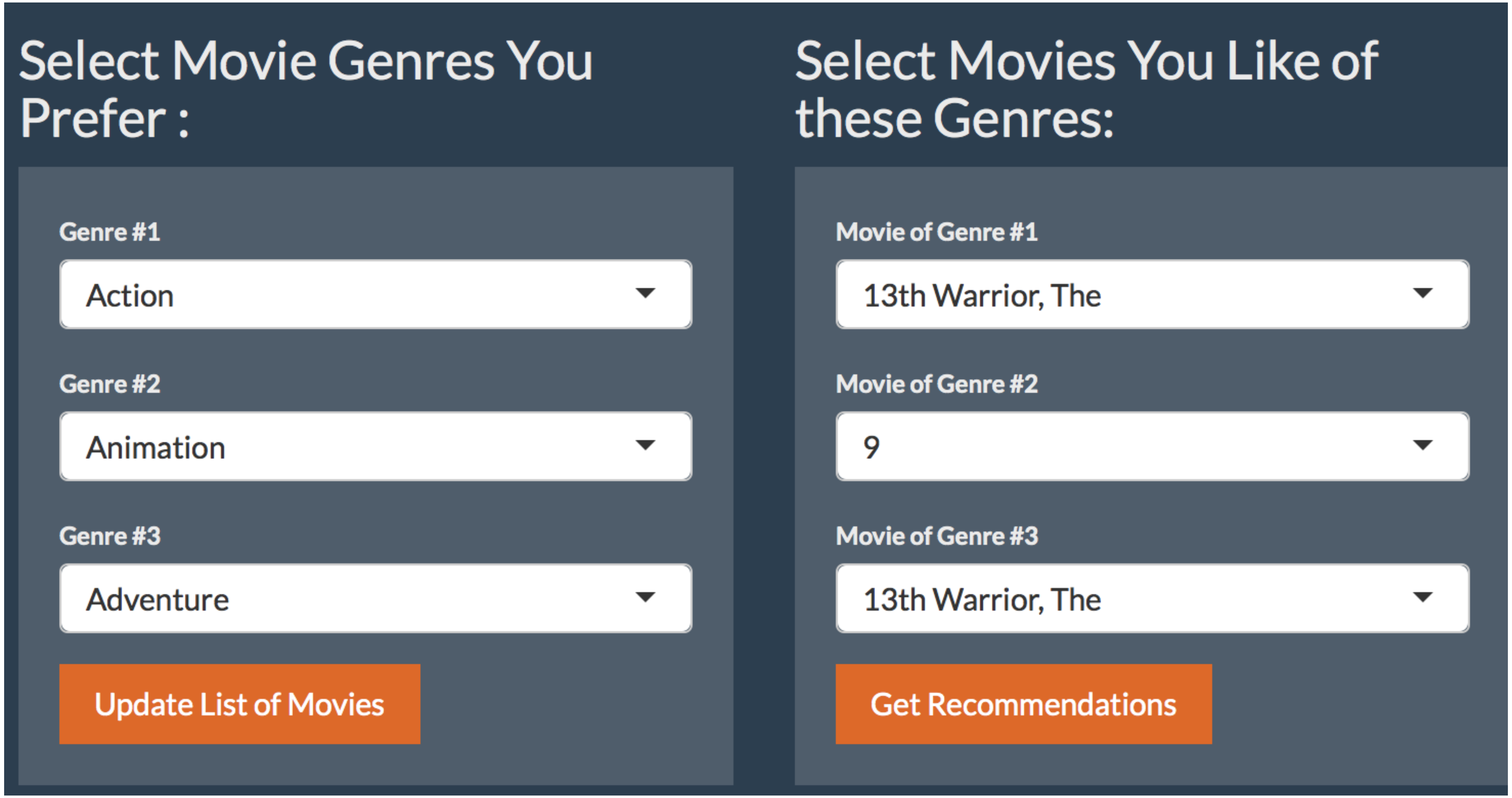
Hybrid -> Feature Augmentation

Methodology

	Content filtering	Collaborative filtering
advantages	- Fast to calculate - easy to interpret	- Based on user profile records, the results are more reliable - could get a result more satisfy user's preference, more individuality
disadvantages	- Data Sparsity problem - Hard to solve complicated problems - Recommended result is substitute good, rather than complementary good	- Cold start problem - Computational cost - Prediction quality is too much rely on historical data profile.

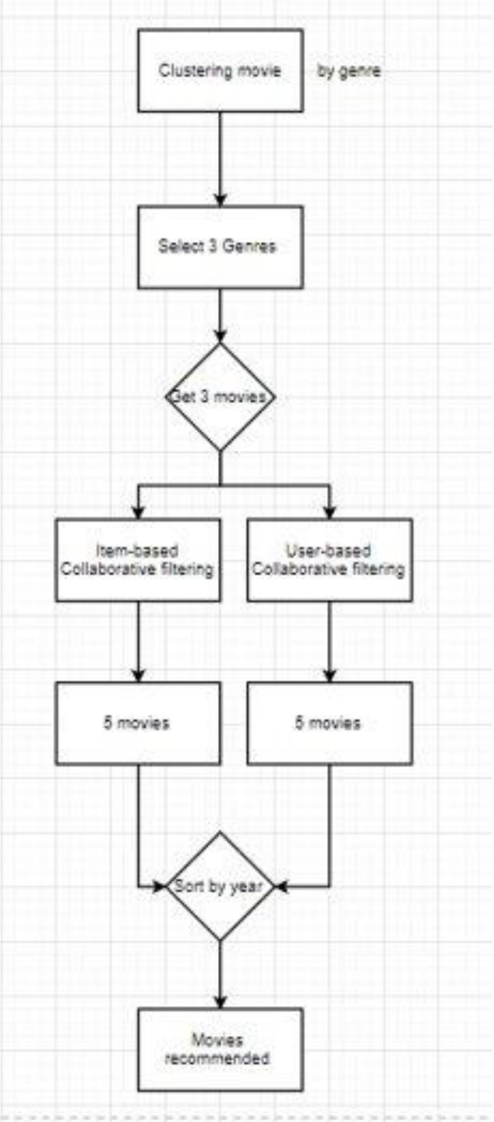
Content based





Collaborative filtering

NETFLIX



The output from content based is also as input of the collaborative filtering

Evaluation- single CF

NETFLIX

Test_train_split: 30%/70%

RMSE	IBCF-Cosine	IBCF-Euclidean
train	3.47	3.38
test	3.46	3.37

RMSE	UBCF-Cosine	UBCF-Euclidean
train	3.16	3.12
test	3.39	3.39

1. bigger dataset, more accurate
2. comparing two single collabrative filtering, user based is better

Test_train_split: 20%/80%

RMSE	IBCF-Cosine	IBCF-Euclidean
train	3.45	3.34
test	3.42	3.34

RMSE	UBCF-Cosine	UBCF-Euclidean
train	3.09	3.19
test	3.43	3.33

Evaluation – Teambob

Exact time cost

RMSE: 1.102

```
UBCF run fold/sample [model time/prediction time]
         1  [0.003sec/417.027sec]
         2  [0.002sec/443.91sec]
         3  [0.002sec/425.7sec]
         4  [0.003sec/455.162sec]
         5  [0.003sec/443.457sec]
```

"KPI" – Teambob

Recommendation system criteria index	Rating score (now -> future)
accuracy	
diversity	
novelty	
trust	
User satisfaction	
serendip	
robustness	
Real-time recommended	
Business goal	

Users' Satisfaction & Learn Users' Pattern

5 Part 5: Conclusion

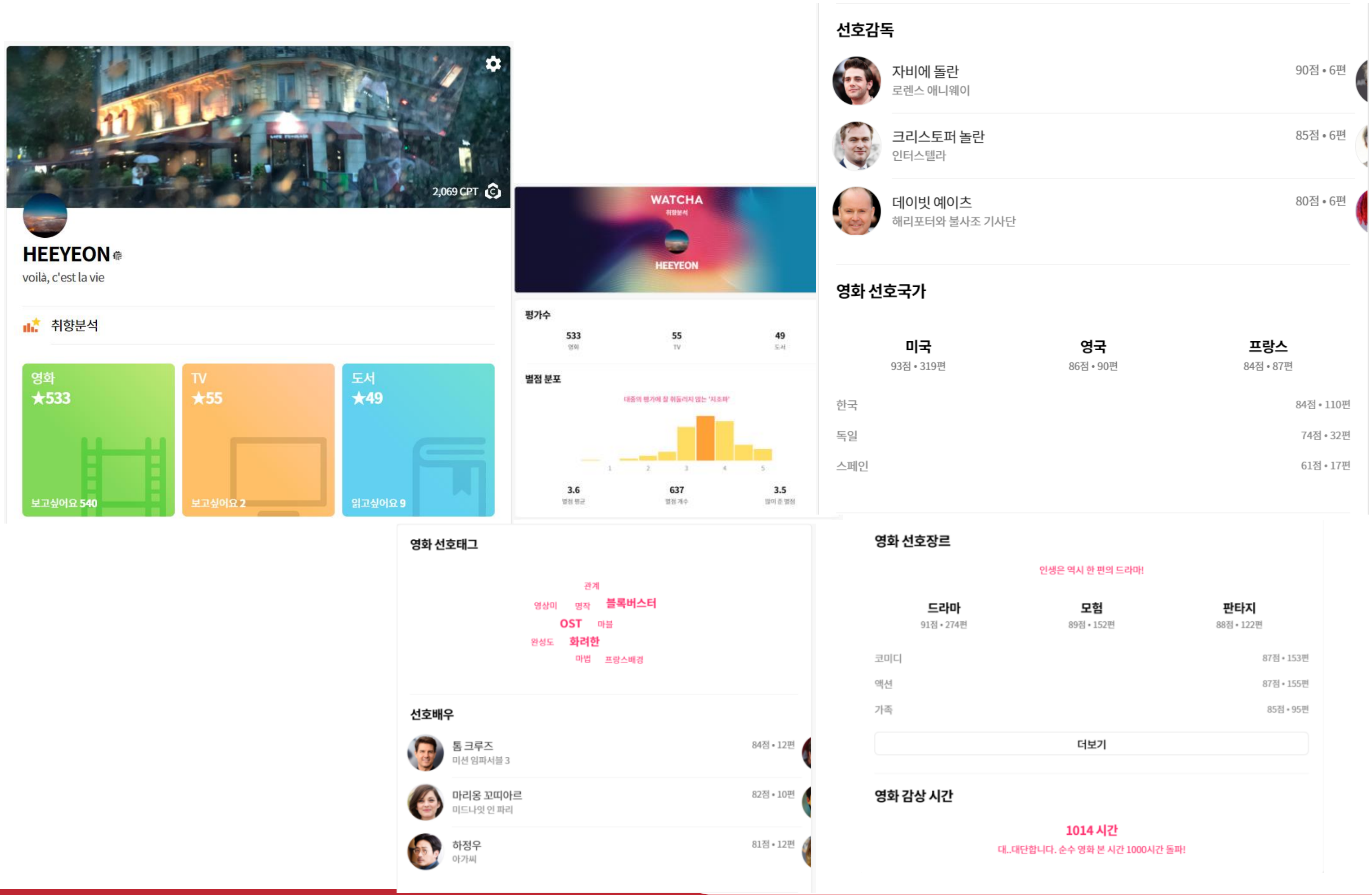
Business model Analysis

NETFLIX

- Netflix data mines not only to make good recommendations to users but also
- to uncover what type of shows they should produce in the future based on
- what is popular

Business model Analysis

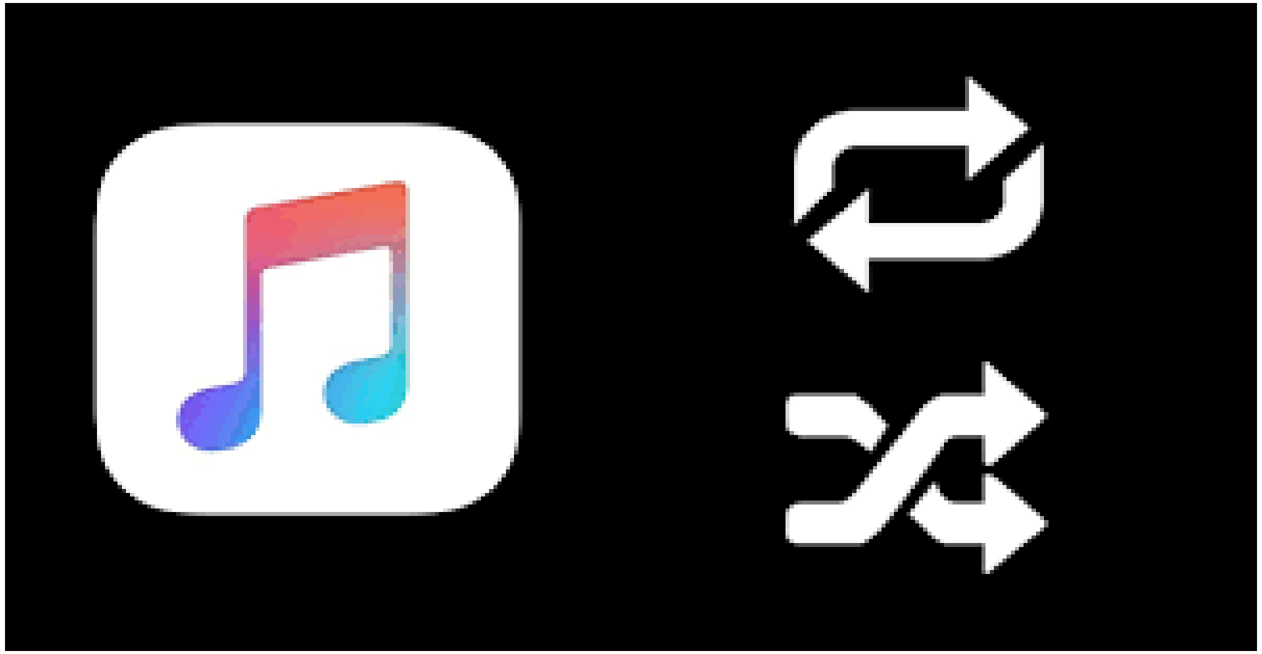
- Netflix data mines not only to make good recommendations to users but also
- to uncover what type of shows they should produce in the future based on
- what is popular



Recommendation

NETFLIX

- Better Rating System
- Shuffle Recommendation
- Age Expansion



NETFLIX

THANK YOU !

Q & A

Evaluation criteria

- Understanding of business context (identification of key points, understanding of sector) Mastery & Pertinence of statistical analysis (descriptive, modeling);
- Link between data / analysis / business recommendations (logical flow, pertinence, completeness);
- Business concepts (thoroughness, pertinence);
- Thoroughness and quality of business recommendations (pertinence, relevance, professionalism, originality);
- Understanding of big data (mastery of big data terms, confidence when using big data concepts);
- Visuals (professionalism, slides support well the main arguments of the presentation, appropriate content);
- Delivery (clear and logical organization, effective introduction and conclusion, creativity, transition between speakers, oral communication skills, eye contact);
- Q&A session (ability to answer questions);
- Report: Quality of data analysis

Challenges

- **How we gonna deal with our data size**
- H/W processor
- We didn't realize our idea to code (because of the reason that we don't have much variables)