Recommender Systems

How recommender systems work and shape our world

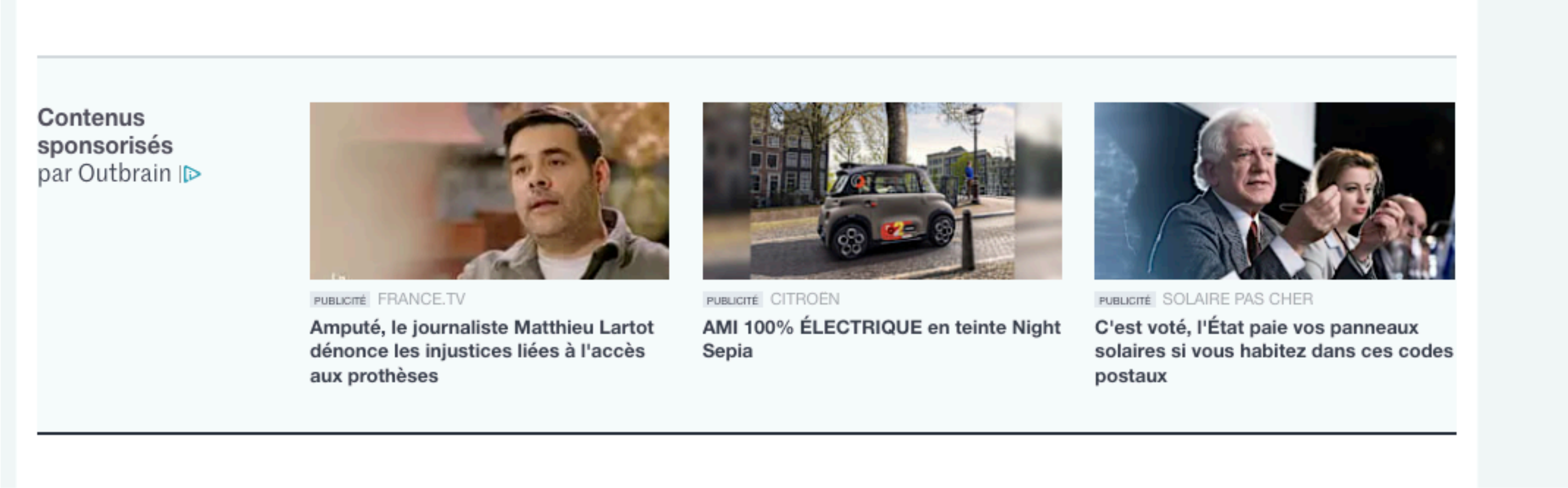
Guillaume Lagarde - Numerics - 20/01/2025

What does 'recommender systems' make you think of?

wooclap.com
code RGMYNH

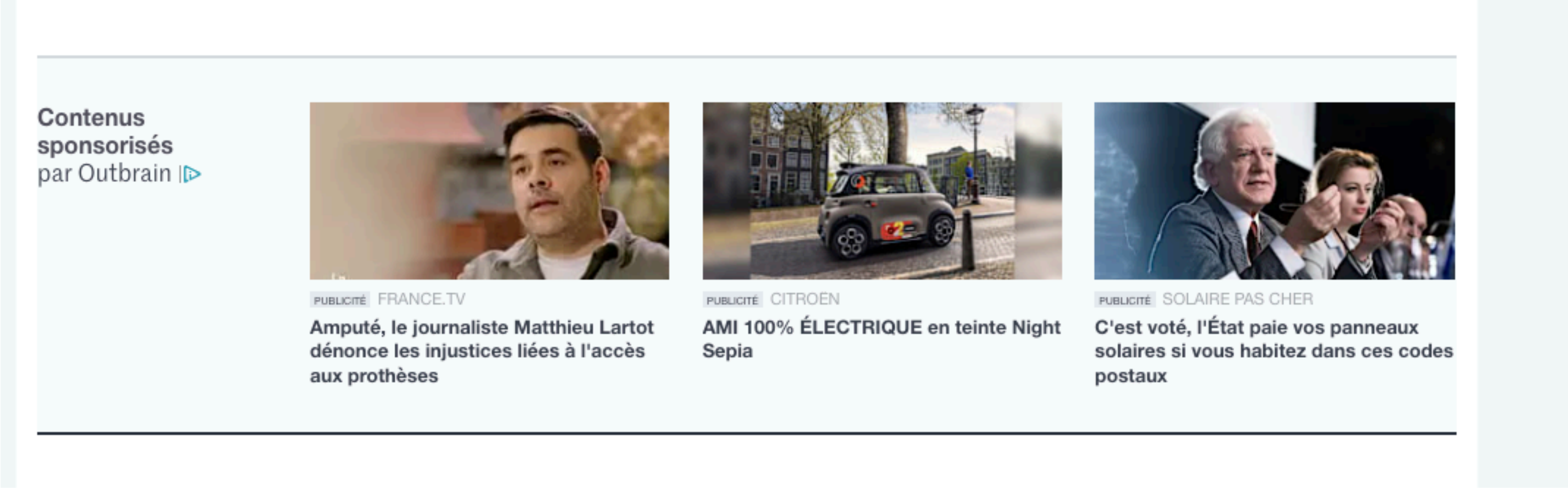
Everything You Wanted to Know About Internet Ads

(but never thought to ask)



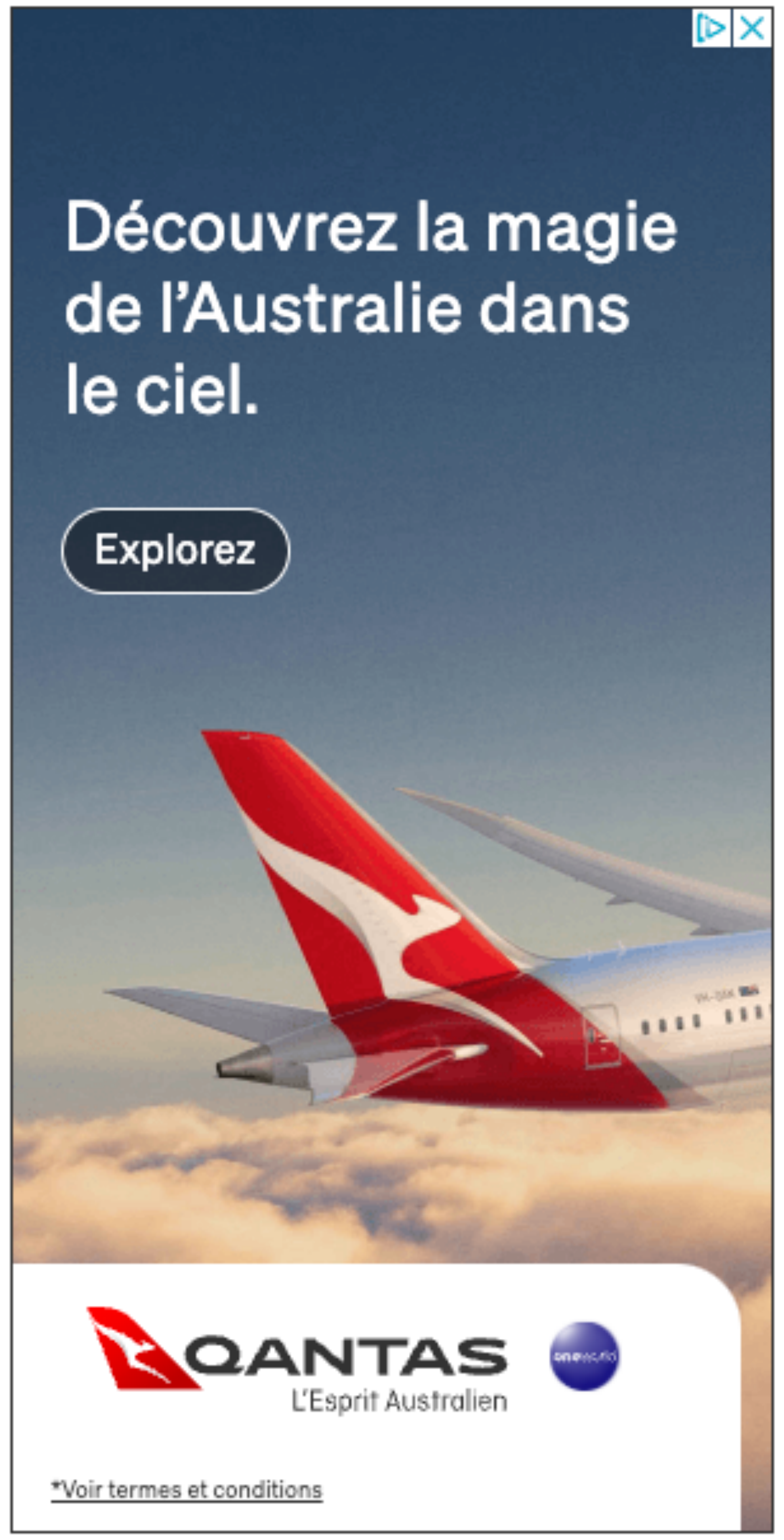
Everything You Wanted to Know About Internet Ads

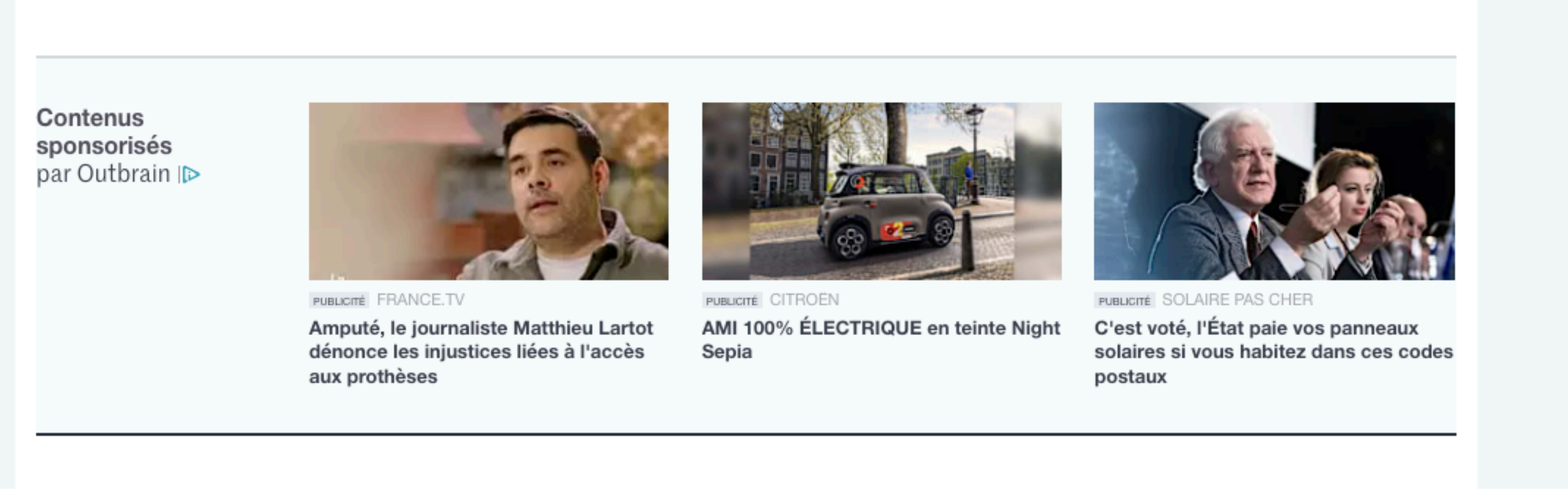
(but never thought to ask)

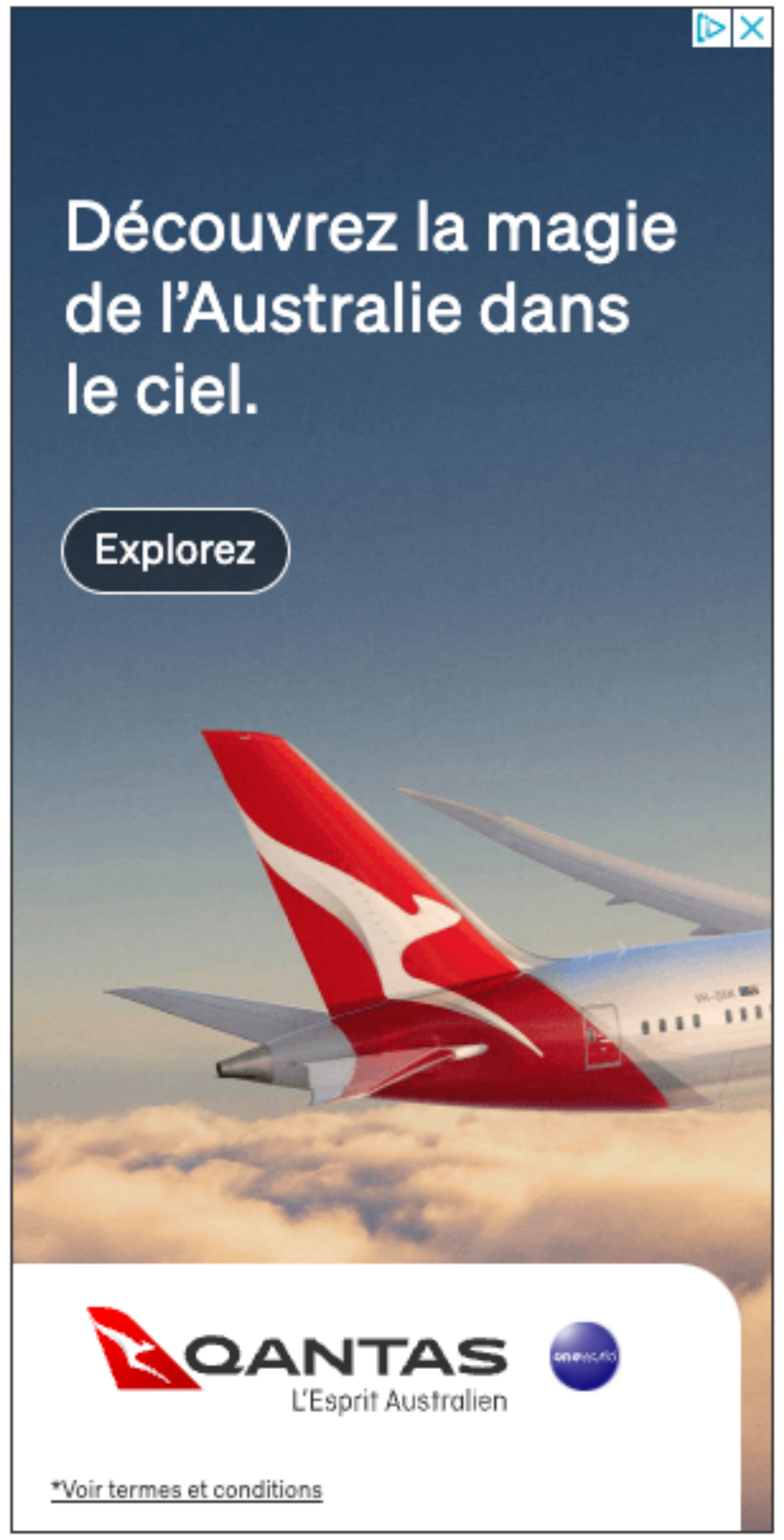


Everything You Wanted to Know About Internet Ads

(but never thought to ask)

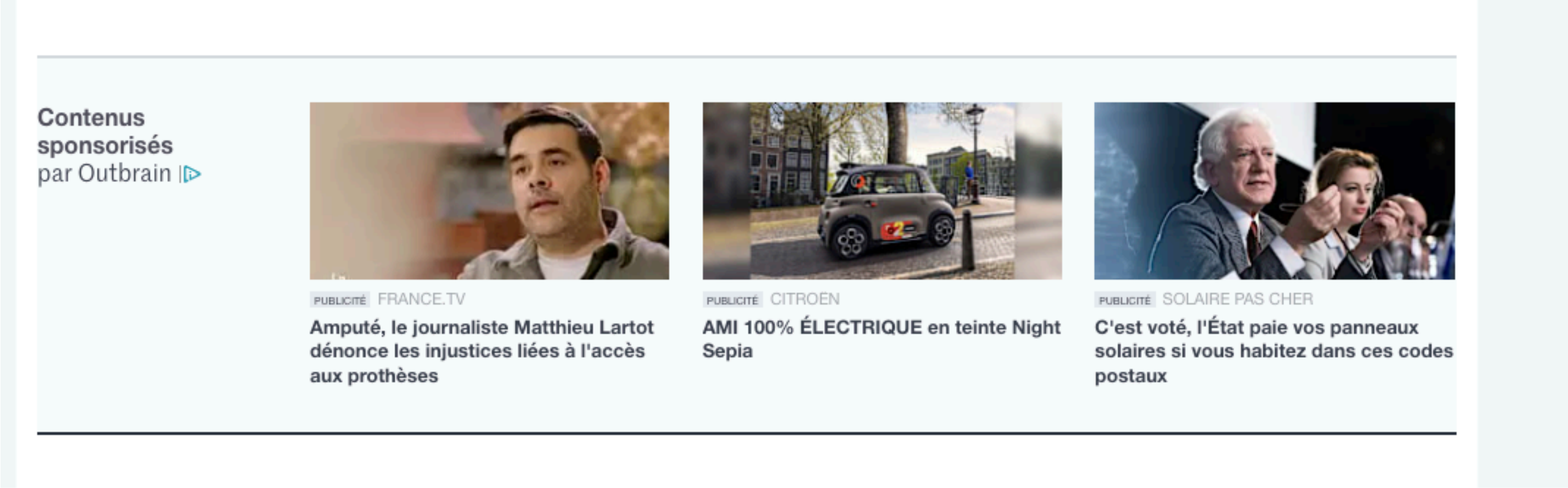


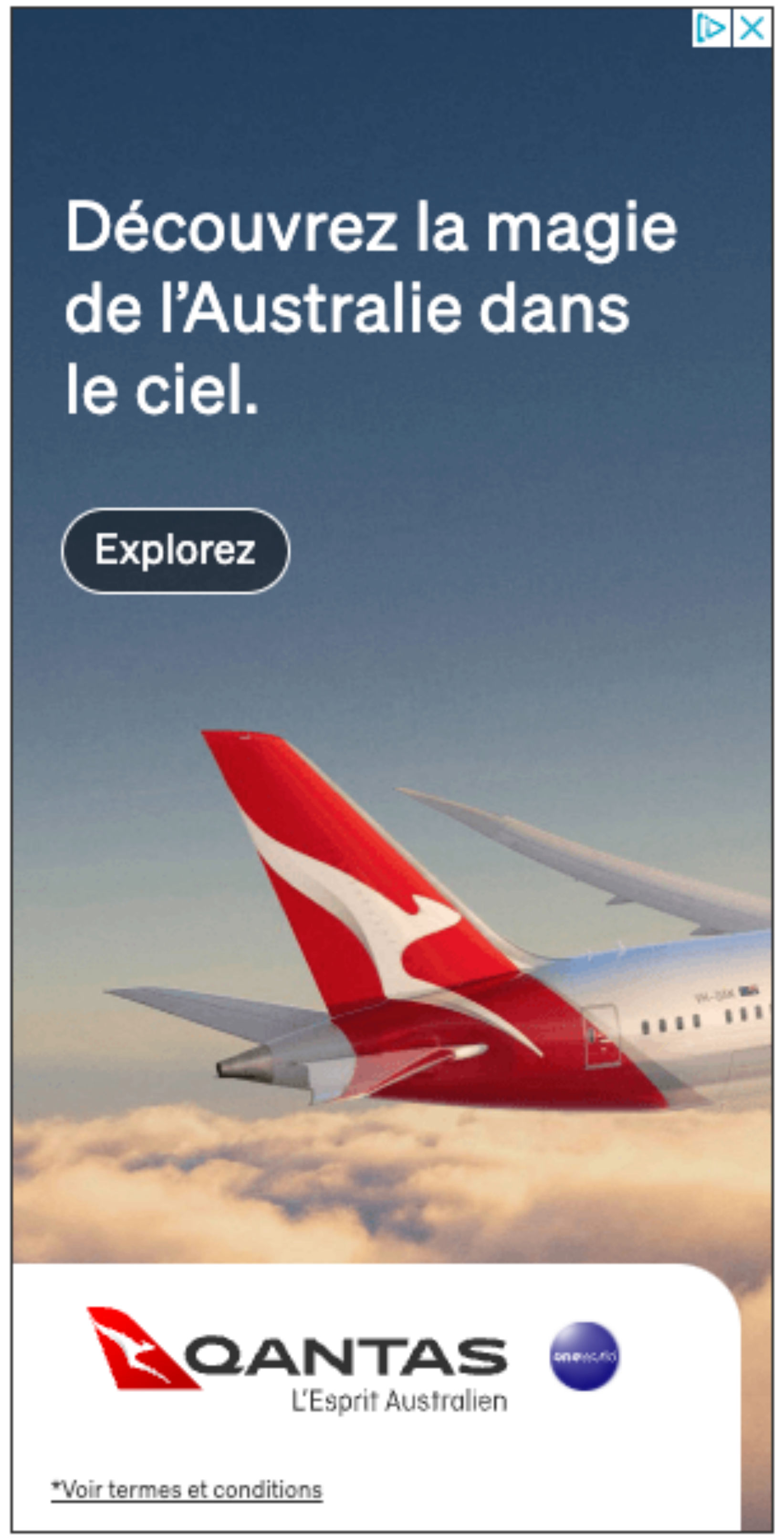




Everything You Wanted to Know About Internet Ads

(but never thought to ask)





Everything You Wanted to Know About Internet Ads

(but never thought to ask)

Blackboard

Important notions to keep in mind

- **Scalability / Performance.** Must deliver real-time decisions in milliseconds while processing billions of dynamic data points. Trying to find a needle in a shifting haystack.
- **Data.** Depend on vast amounts of information (user interactions, contextual data, etc.)
- **Goals.** The system needs a clear definition of "what makes an item interesting for a user". Loss functions
- **Impacts.** Algorithms at scale can significantly shape society.

Roadmap

1. Where and Why
2. How do they work.
3. Impact.

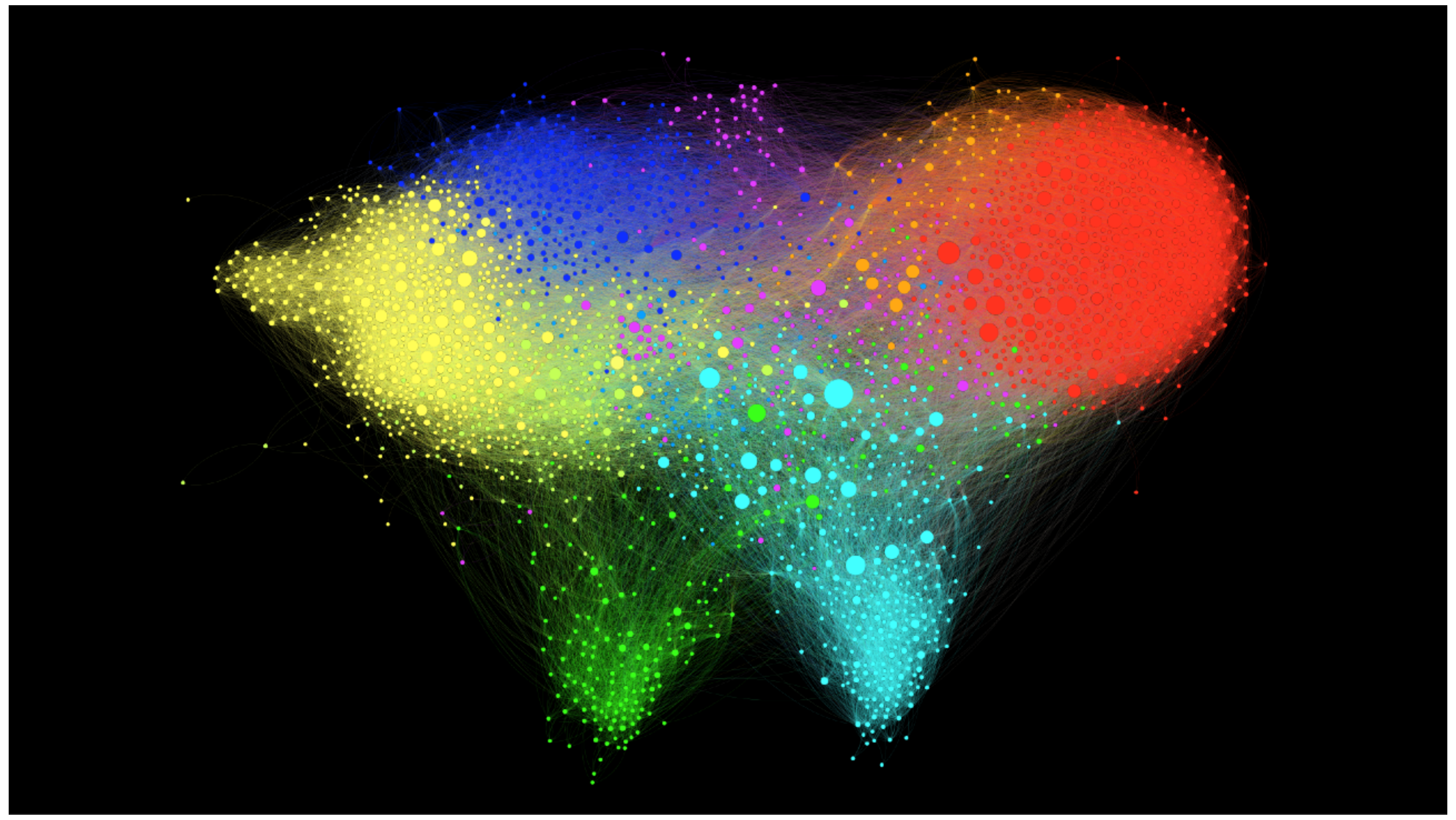
Where?

Literally everywhere.

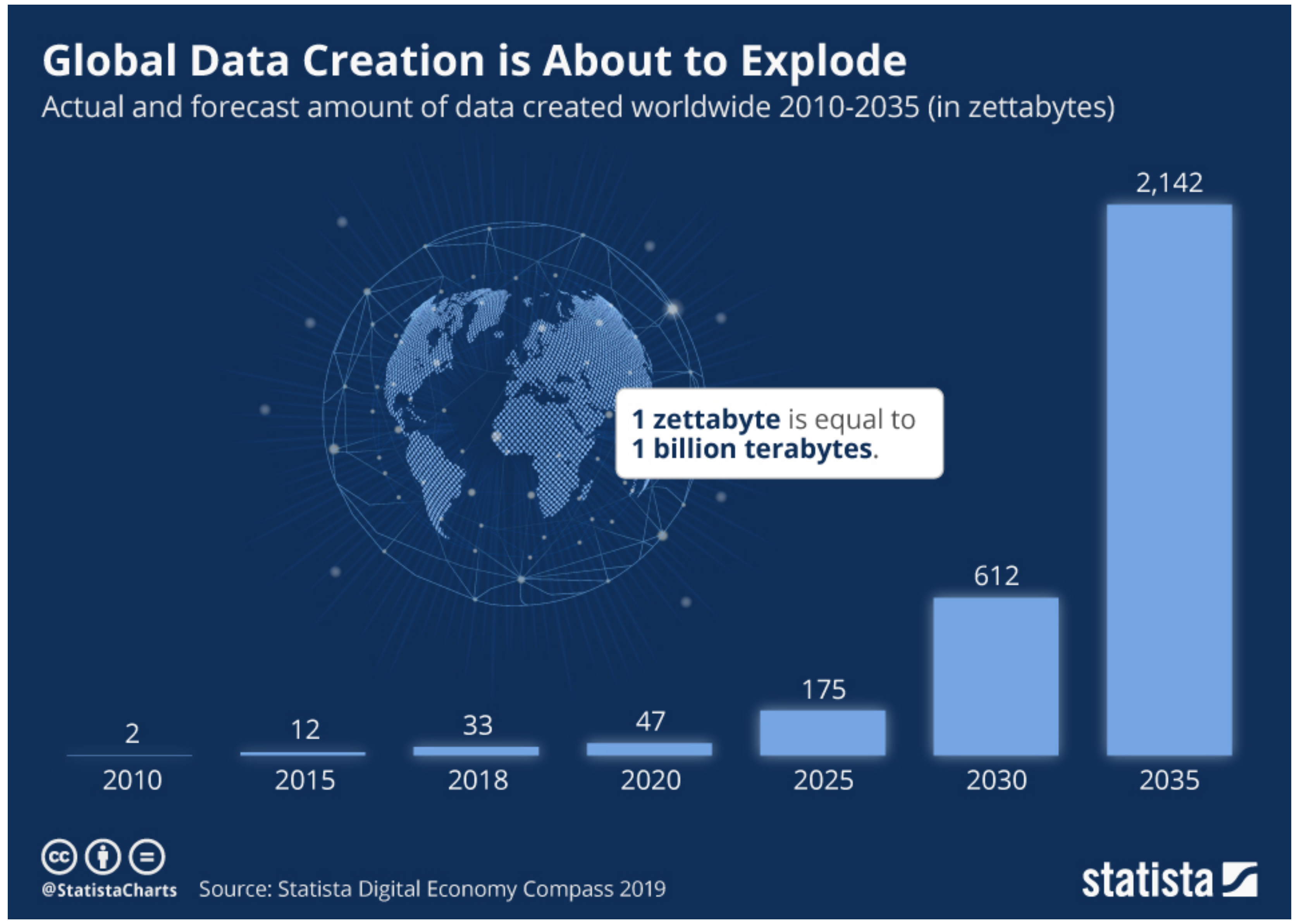
Why?

A solution to the modification of our social topology and the information overload problem.

Recommender systems were a new way to discover, process and extract information, at the individual level



Twitter world of Edouard Laurent. **Source** twitter-graph repo on GitHub

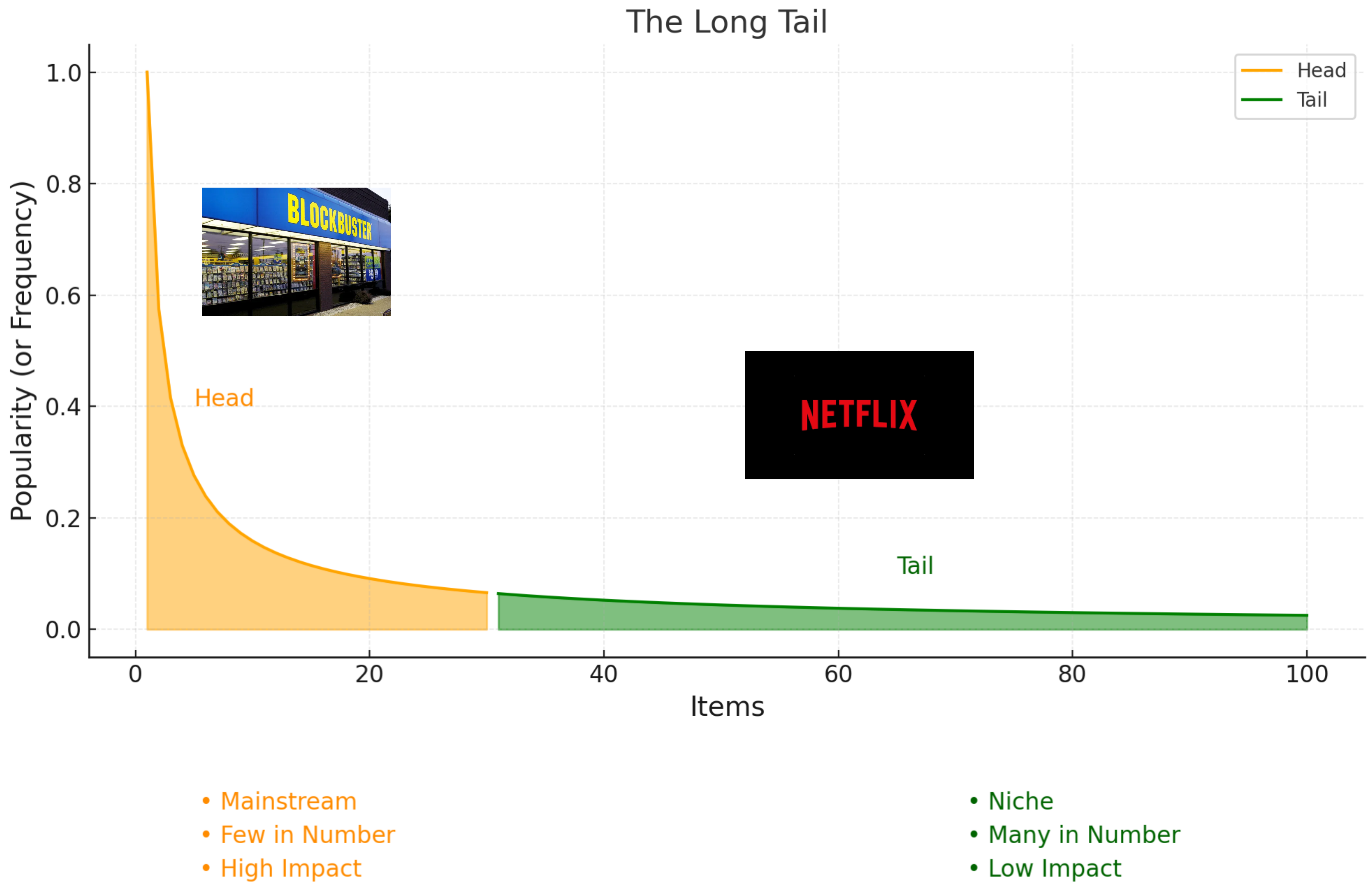


from

to

NETFLIX

Long-tail discovery.



The Head-Only Era

- Before the Digital Age (Pre-2000s)

The Shift to Long-Tail Discovery

- Netflix and Streaming Platforms (Post-2000s)

“First we build the tools,
then they build us”

Marshall McLuhan (Canadian philosopher, “father of media studies”)

Algorithmic choices + large scale

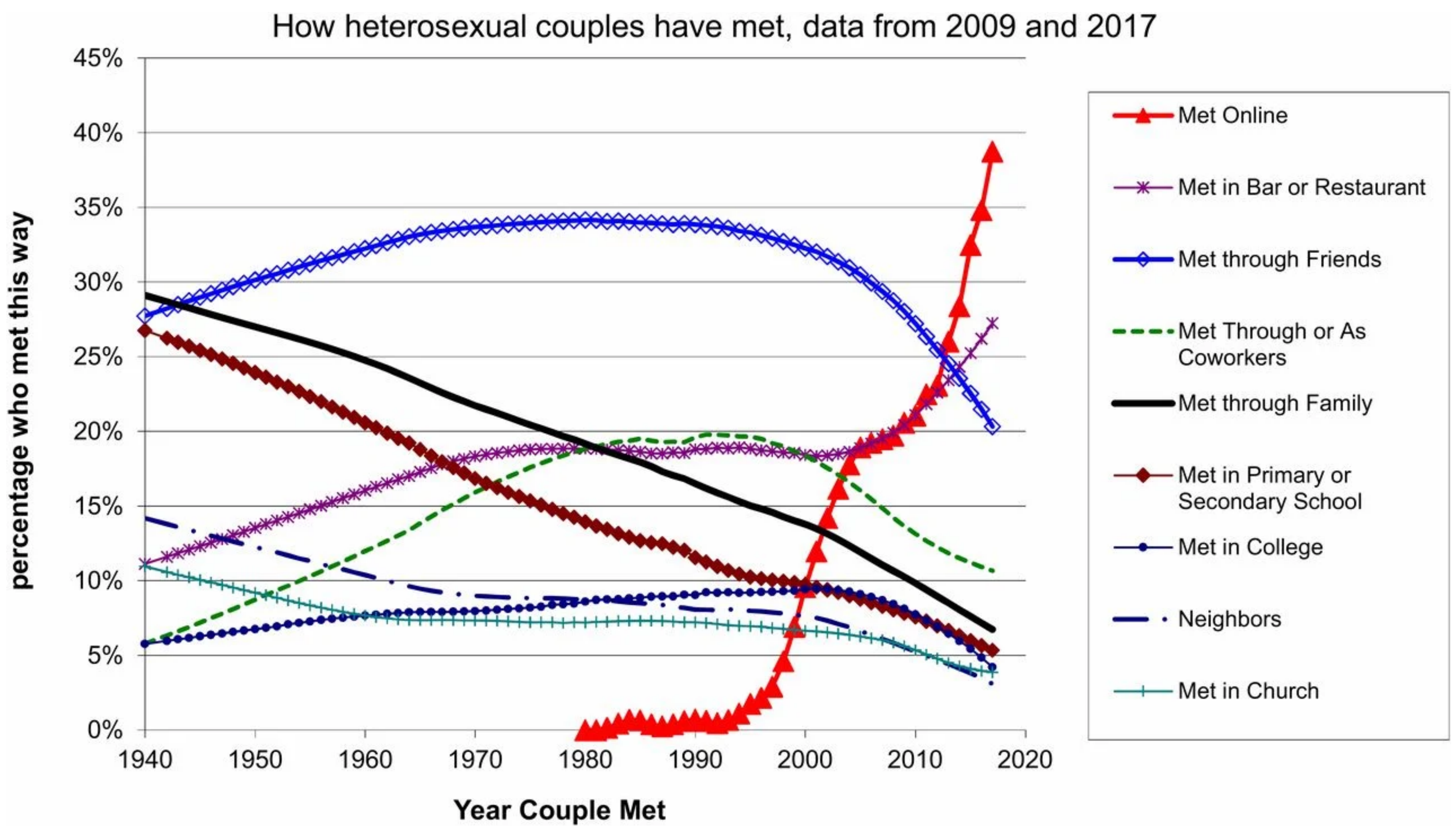
= Social consequences.

80%

of the series watched on Netflix are directly driven by its recommendation system

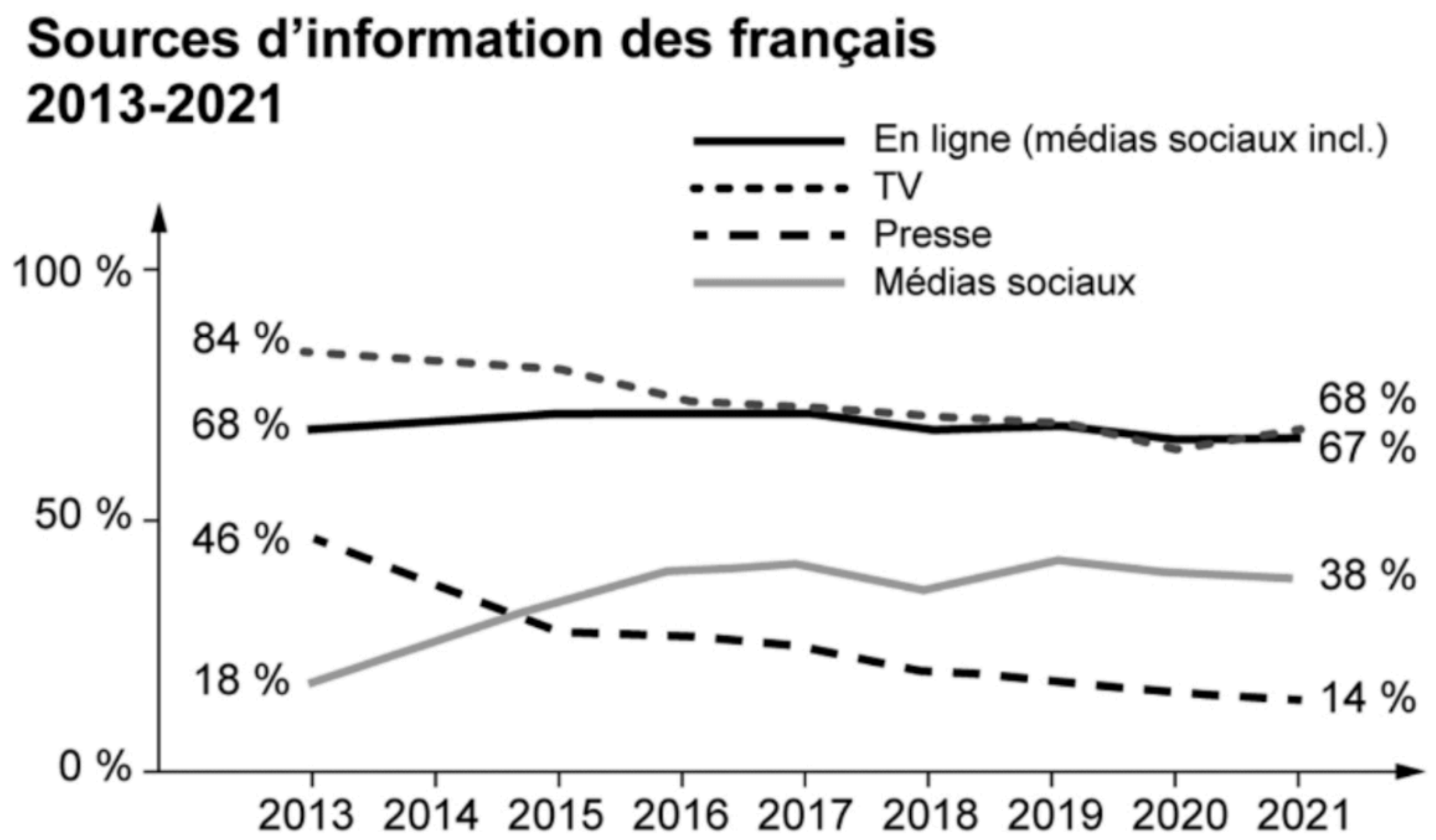
How Couples Meet and Stay Together (HCMST)

A dataset created through a Stanford survey across the U.S



Source. *Disintermediating your friends: How online dating in the United States displaces other ways of meeting*. Michael J. Rosenfeld, Reuben J. Thomas, and Sonia HausenAuthors

How French people find information



Évolution des sources d'information des Français entre 2013 et 2021. Sur cette période, la proportion de Français qui s'informent de l'actualité via les médias sociaux est passée de 18 % à 38 %. Source : Digital News Report 2021, Reuters Institute for the Study of Journalism.

Source. *Toxic data*, David Chavalarias

71%

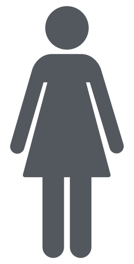
of 15-34 year olds use social networks as their primary source of information

How?

Two strategies

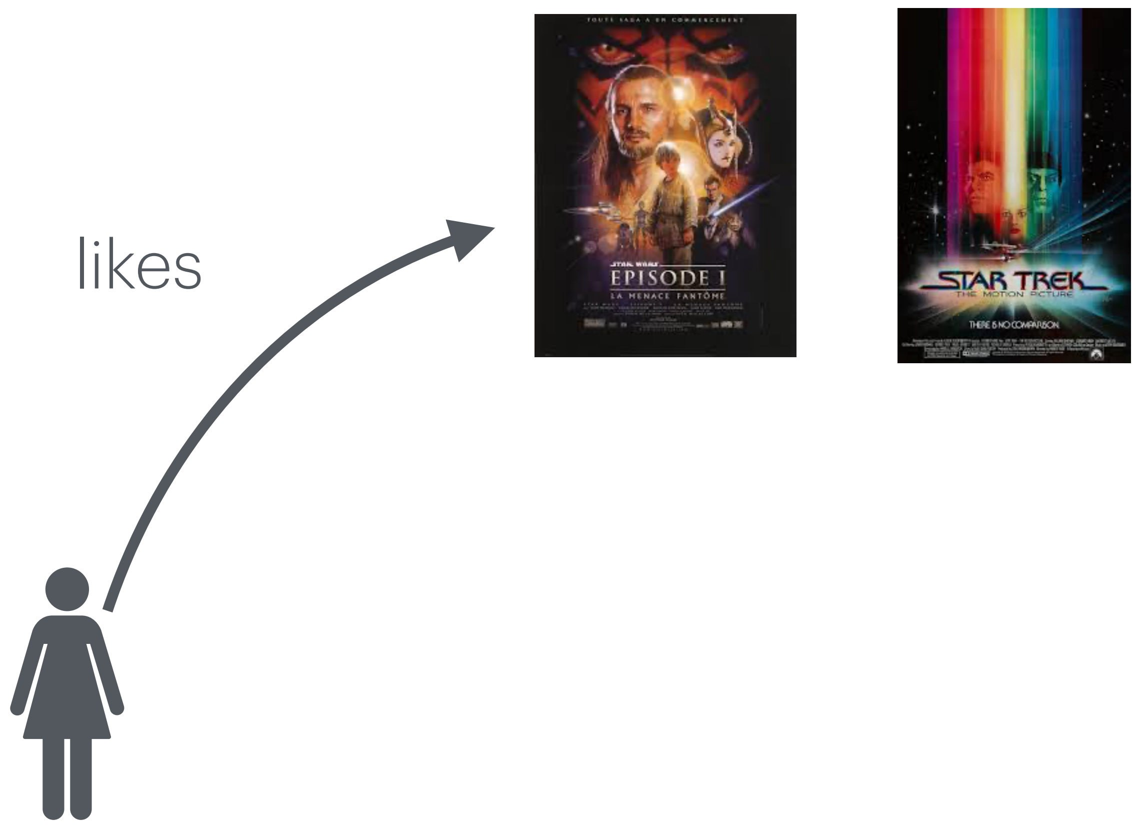
Content-based filtering

Collaborative filtering



Two strategies

Content-based filtering



Collaborative filtering

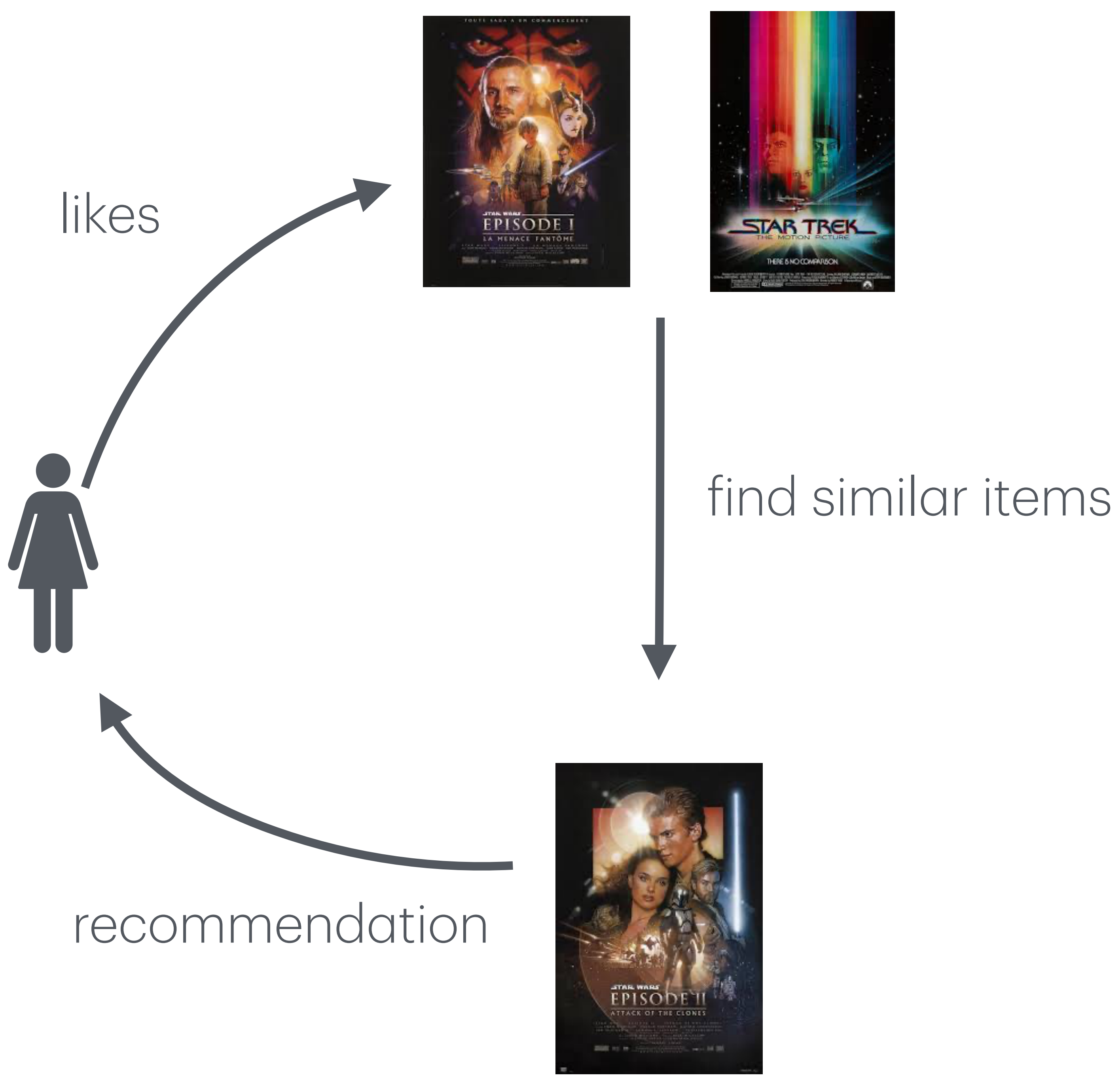
Two strategies

Content-based filtering

Collaborative filtering

Two strategies

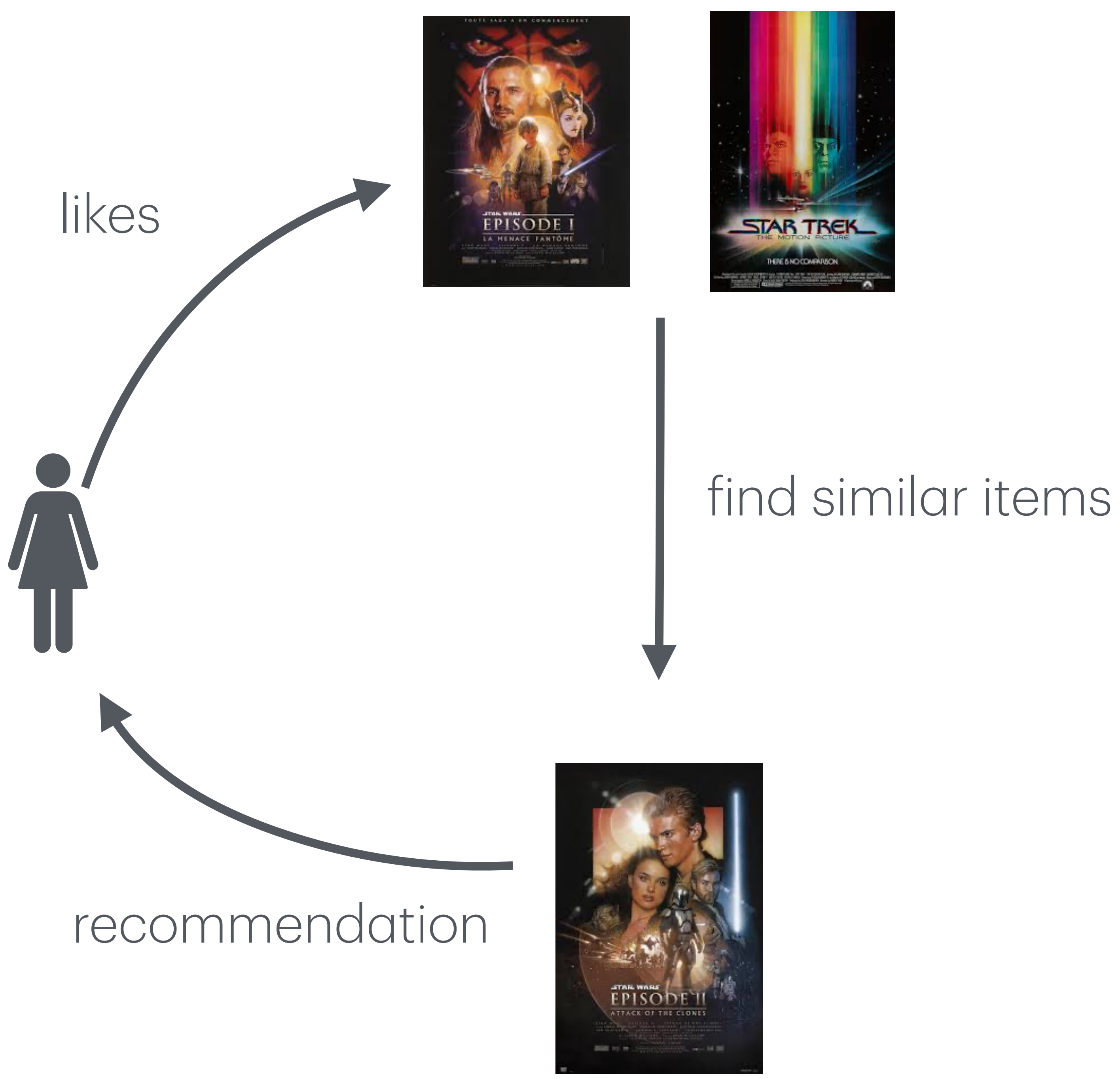
Content-based filtering



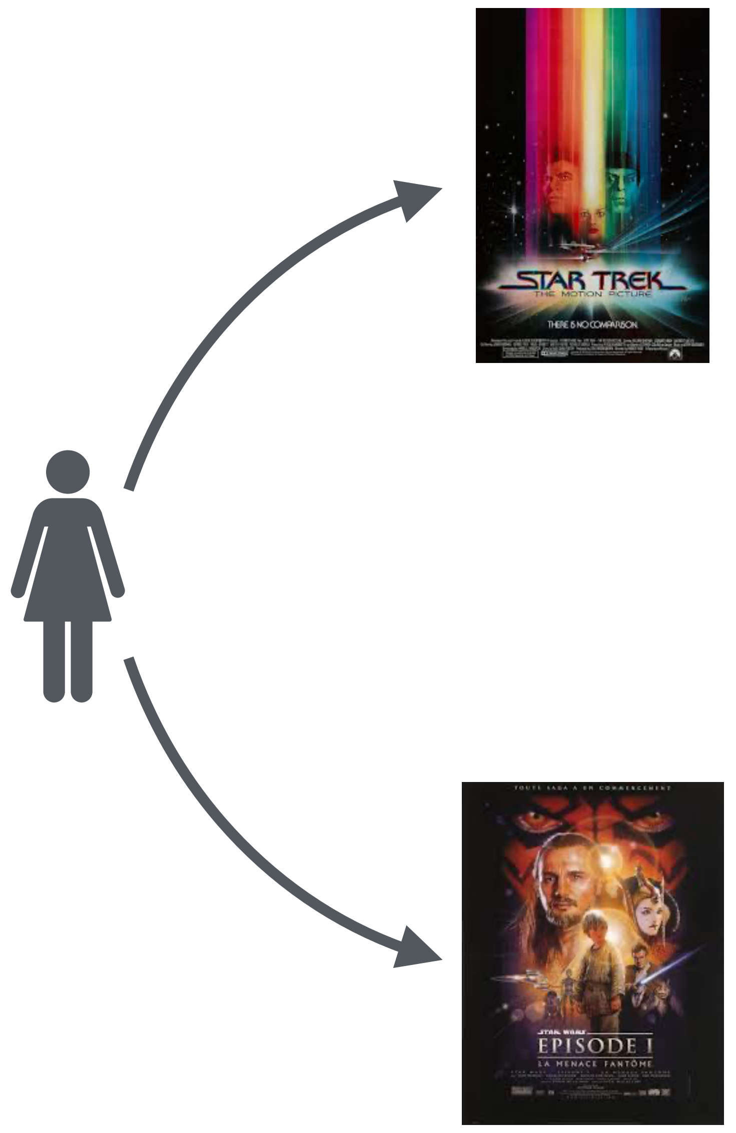
Collaborative filtering

Two strategies

Content-based filtering

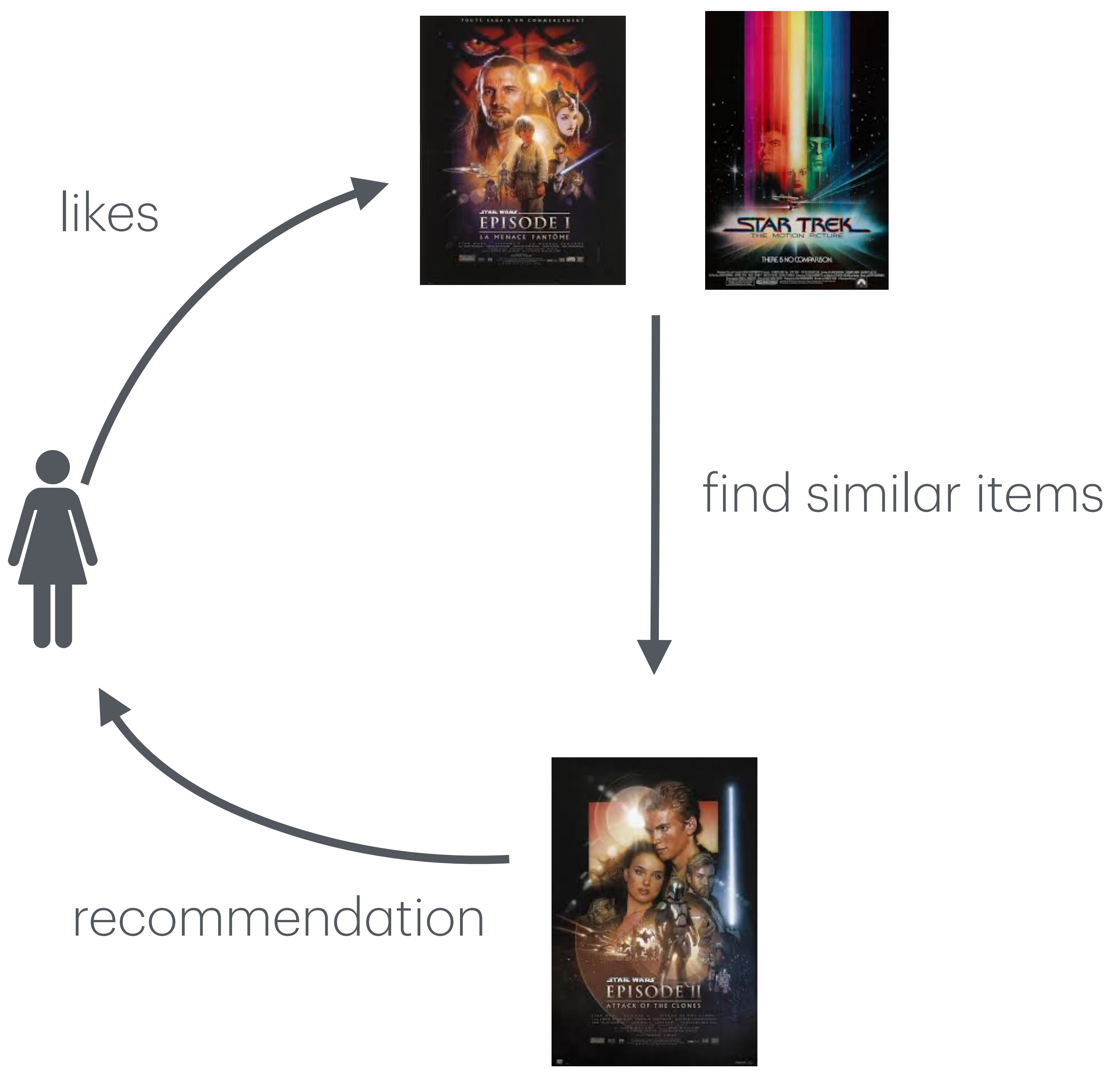


Collaborative filtering

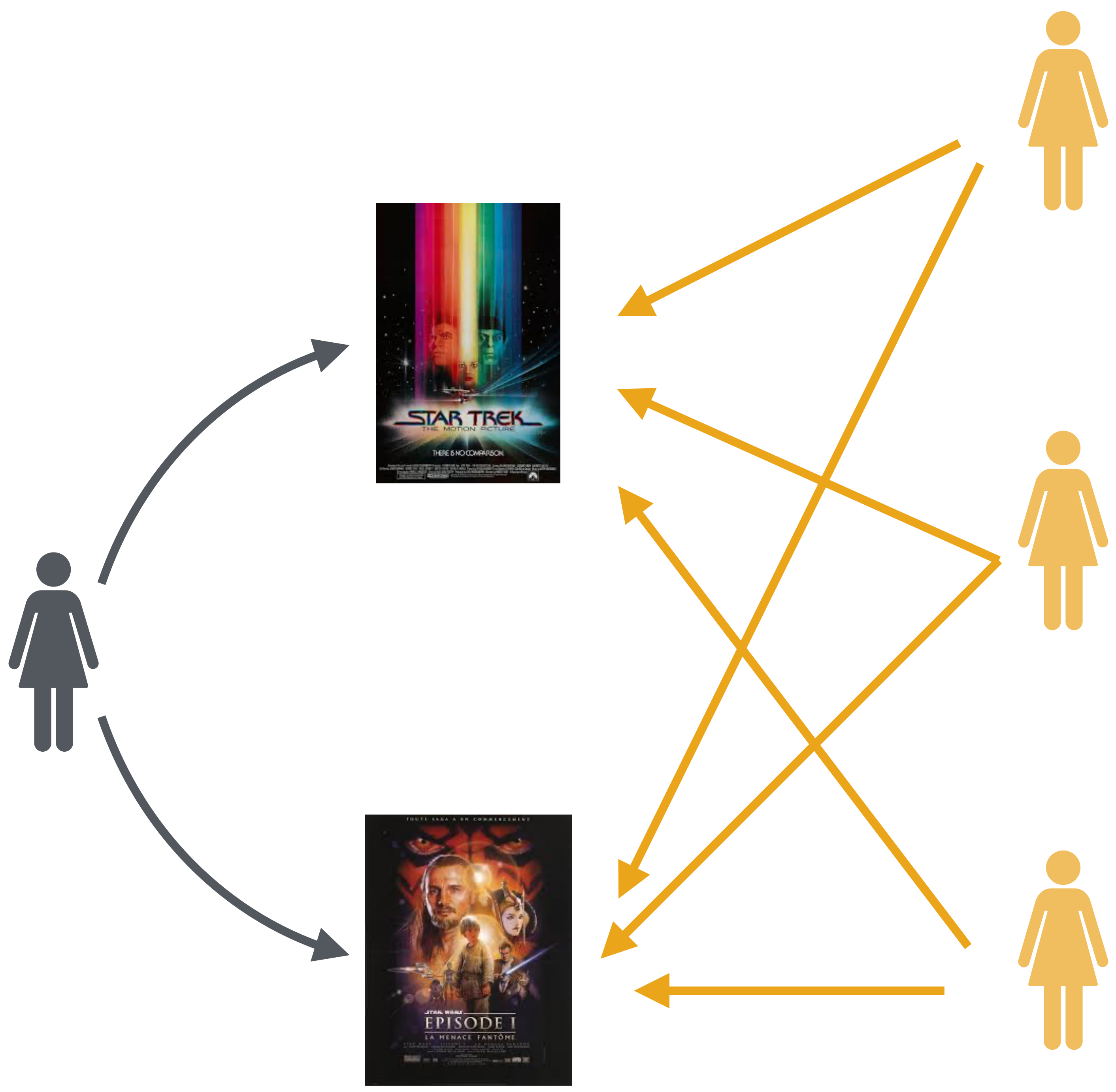


Two strategies

Content-based filtering

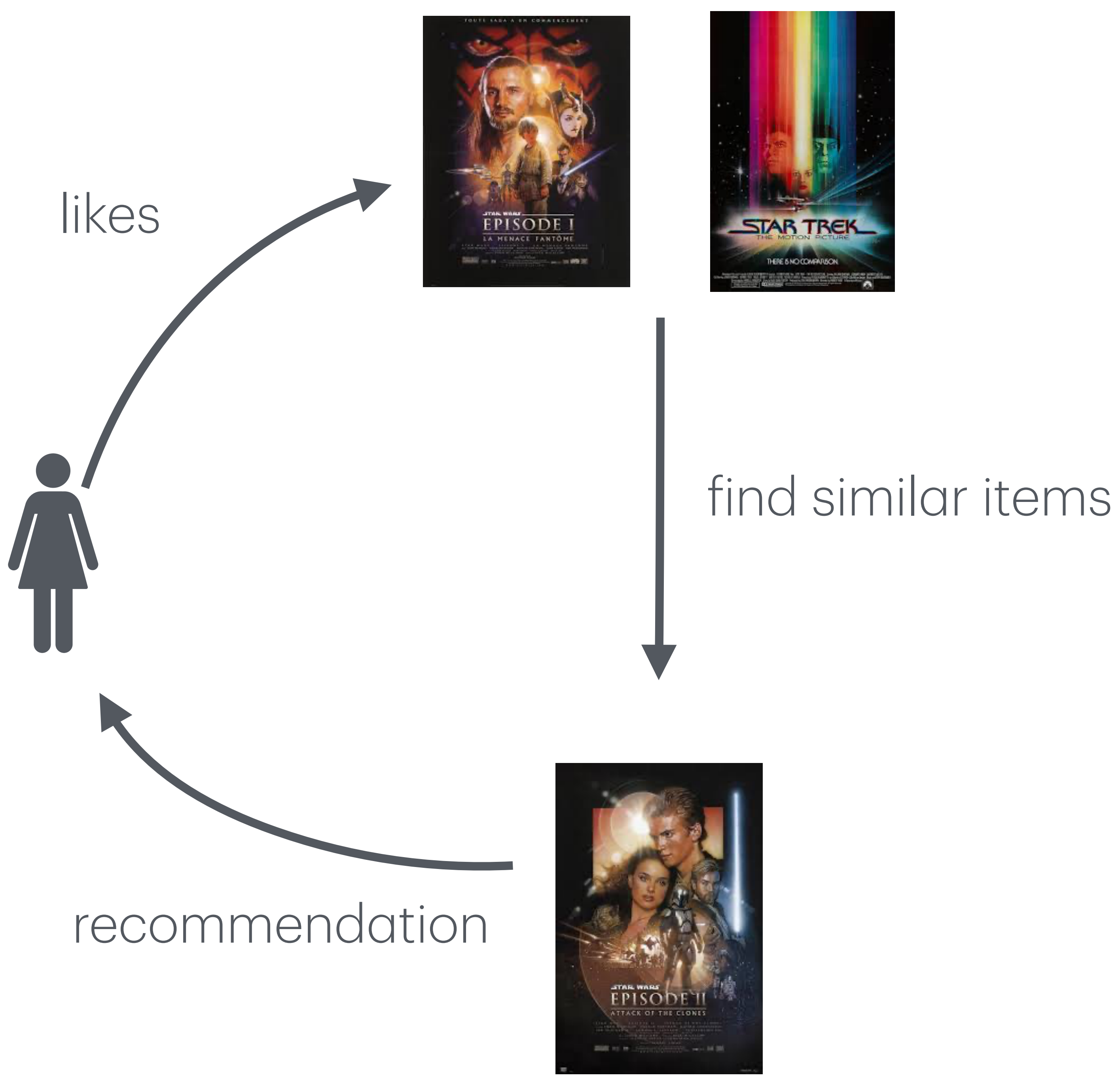


Collaborative filtering

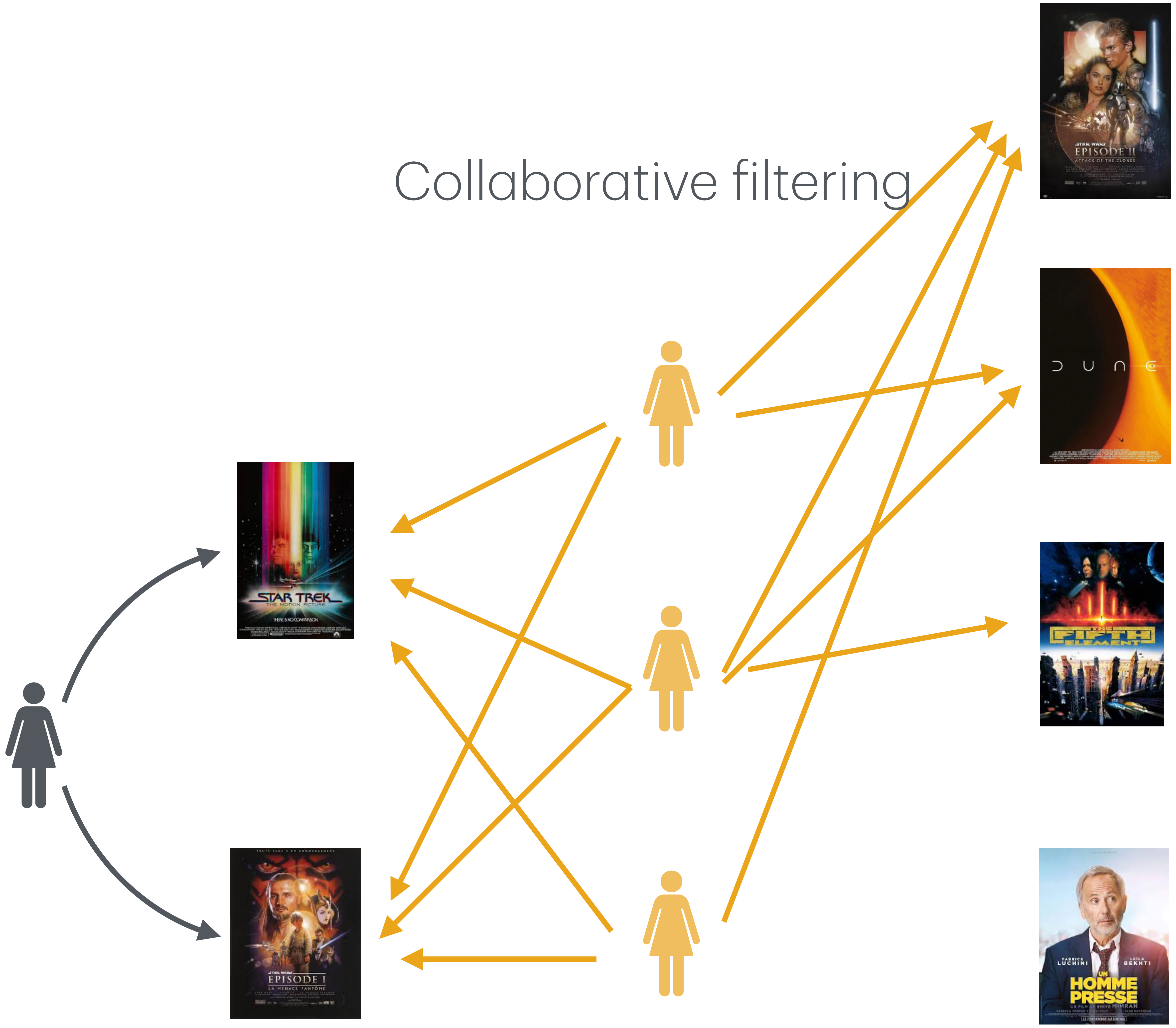


Two strategies

Content-based filtering



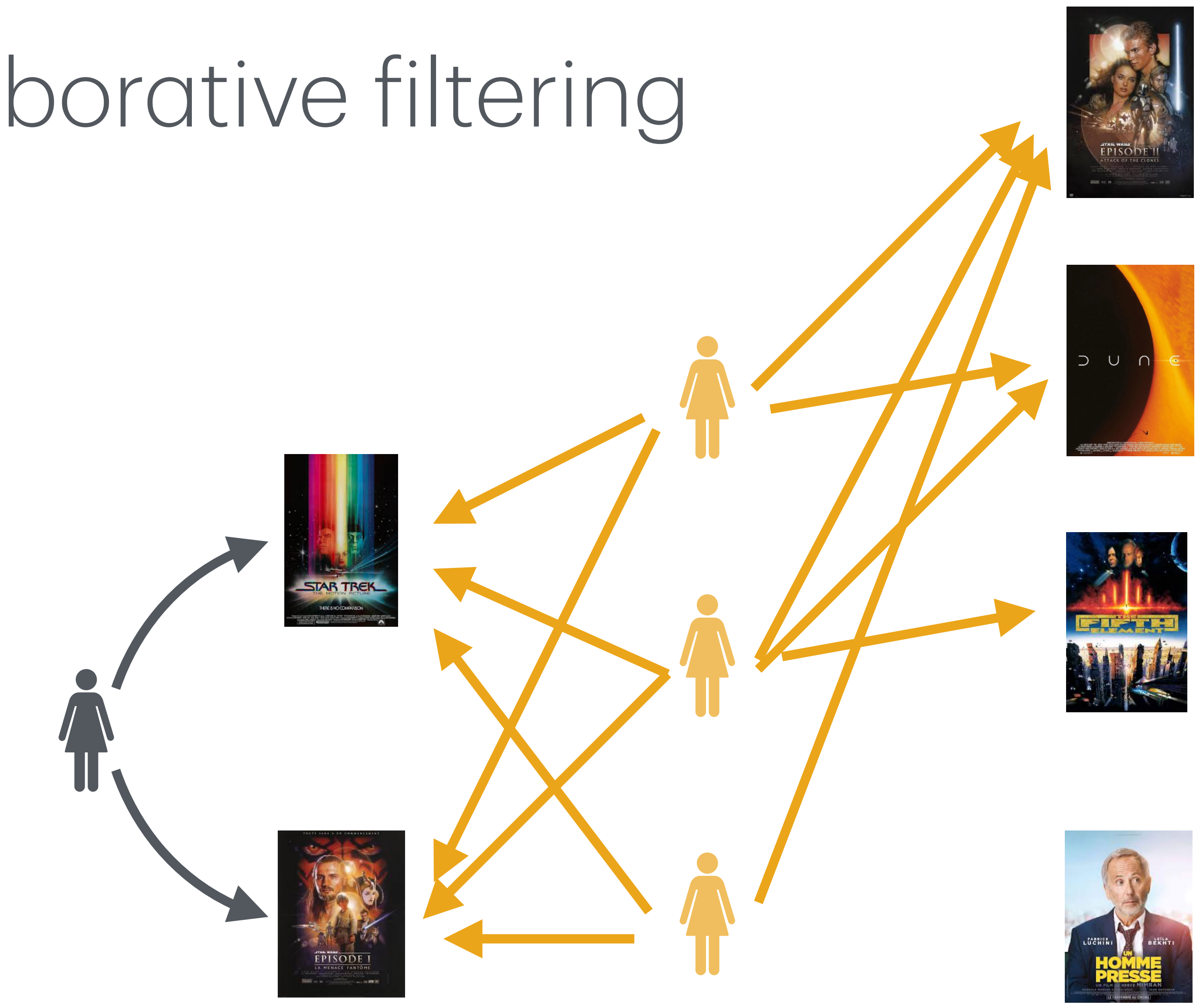
Collaborative filtering



Two strategies

Content-based filtering

Collaborative filtering



VS

- Understands “what’s inside the box” (features)
- Domain specific

- Uses interactions users / items
- Domain free

Tons of different ways to do these two methods.

- K nearest neighbours
- Pairwise logistic regression
- TF-IDF
- **Matrix factorization**
- Graph neural network
- **Deep learning and embeddings**
- Pointwise mutual information
- Pearson Correlation Coefficient
- Neural Collaborative Filtering
- Bayesian Personalized Ranking

etc...

First technique
Matrix Factorization

The Netflix Prize

Start date. 2006

Goal. Predict user rating for films. Improve by 10% the Netflix algorithm (*Cinematch*).

Data. 100,480,507 ratings from 480,189 users and 17,770 movies

Progress Prizes. $50,000

Grand Prize. $1,000,000

Example of training data

User Name	Movie	Rating
BradPittFan	The Shawshank Redemption	⭐⭐⭐⭐⭐
MerylStreep101	The Godfather	⭐⭐⭐
TomHanksLover	Forrest Gump	⭐⭐⭐⭐
BradPittFan	Titanic	⭐⭐
ScorseseBuff	Pulp Fiction	⭐⭐⭐⭐⭐
MerylStreep101	Titanic	⭐⭐⭐⭐

Root Mean Square Error (RMSE)

The formal goal

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(P_i - A_i\right)^2}$$

- N The total number of ratings in the test dataset.
- P_i The predicted rating for the i-th movie-user pair.
- A_i The actual rating for the i-th movie-user pair.

Baseline. RMSE from 0.9525 (the C*inematch Netflix algorithm*)

To win the Grand Prize. 10% improvement —> RMSE of 0.8572 to win the grand prize

2009, *BellKor's Pragmatic Chaos* wins - AT&T research engineers team

NETFLIX

Netflix Prize

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8588	9.84	2009-07-10 01:12:31
5	Vandelay Industries !	0.8591	9.81	2009-07-10 00:32:20
6	PragmaticTheory	0.8594	9.77	2009-06-24 12:06:56
7	BellKor in BigChaos	0.8601	9.70	2009-05-13 08:14:09
8	Dace_	0.8612	9.59	2009-07-24 17:18:43
9	Feeds2	0.8622	9.48	2009-07-12 13:11:51
10	BigChaos	0.8623	9.47	2009-04-07 12:33:59
11	Opera Solutions	0.8623	9.47	2009-07-24 00:34:07
12	BellKor	0.8624	9.46	2009-07-26 17:19:11
Progress Prize 2008 - RMSE = 0.8627 - Winning Team: BellKor in BigChaos				
13	xiangliang	0.8642	9.27	2009-07-15 14:53:22
14	Gravity	0.8643	9.26	2009-04-22 18:31:32
15	Ces	0.8651	9.18	2009-06-21 19:24:53
16	Invisible Ideas	0.8653	9.15	2009-07-15 15:53:04
17	Just a guy in a garage	0.8662	9.06	2009-05-24 10:02:54
18	J Dennis Su	0.8666	9.02	2009-03-07 17:16:17
19	Craig Carmichael	0.8666	9.02	2009-07-25 16:00:54
20	acmehill	0.8668	9.00	2009-03-21 16:20:50
Progress Prize 2007 - RMSE = 0.8723 - Winning Team: KorBell				
Cinematch score - RMSE = 0.9525				

MATRIX FACTORIZATION TECHNIQUES FOR RECOMMENDER SYSTEMS

Yehuda Koren, *Yahoo Research*
Robert Bell and Chris Volinsky, *AT&T Labs—Research*

As the Netflix Prize competition has demonstrated, matrix factorization models are superior to classic nearest-neighbor techniques for producing product recommendations, allowing the incorporation of additional information such as implicit feedback, temporal effects, and confidence levels.

Such systems are particularly useful for entertainment products such as movies, music, and TV shows. Many customers will view the same movie, and each customer is likely to view numerous different movies. Customers have proven willing to indicate their level of satisfaction with particular movies, so a huge volume of data is available about which movies appeal to which customers. Companies can analyze this data to recommend movies to particular customers.

DATASET

1. ("BradPittFan", "The Shawshank Redemption", "⭐⭐⭐⭐⭐")
2. ("MerylStreep101", "The Godfather", "⭐⭐⭐")
3. ("TomHanksLover", "Forrest Gump", "⭐⭐⭐⭐")
4. ("BradPittFan", "Titanic", "⭐⭐")
5. ("ScorseseBuff", "Pulp Fiction", "⭐⭐⭐⭐⭐")
6. ("MerylStreep101", "Titanic", "⭐⭐⭐⭐")
7. ...

DATASET

1. ("BradPittFan", "The Shawshank Redemption", “⭐⭐⭐⭐⭐”)
2. ("MerylStreep101", "The Godfather", "⭐⭐⭐")
3. ("TomHanksLover", "Forrest Gump", "⭐⭐⭐⭐")
4. ("BradPittFan", "Titanic", "⭐⭐")
5. ("ScorseseBuff", "Pulp Fiction", “⭐⭐⭐⭐⭐")
6. ("MerylStreep101", "Titanic", “⭐⭐⭐⭐")
7. ...

User-Item Interaction Matrix

	The Shawshank Redemption	**The Godfather**	**Forrest Gump**	**Titanic**	**Pulp Fiction**	**...**
BradPittFan	5	?	?	2	?	...
MerylStreep101	?	3	?	4	?	...
TomHanksLover	?	?	4	?	?	...
ScorseseBuff	?	?	?	?	5	...
...	...	...	...	...	...	...

User-Item Interaction Matrix

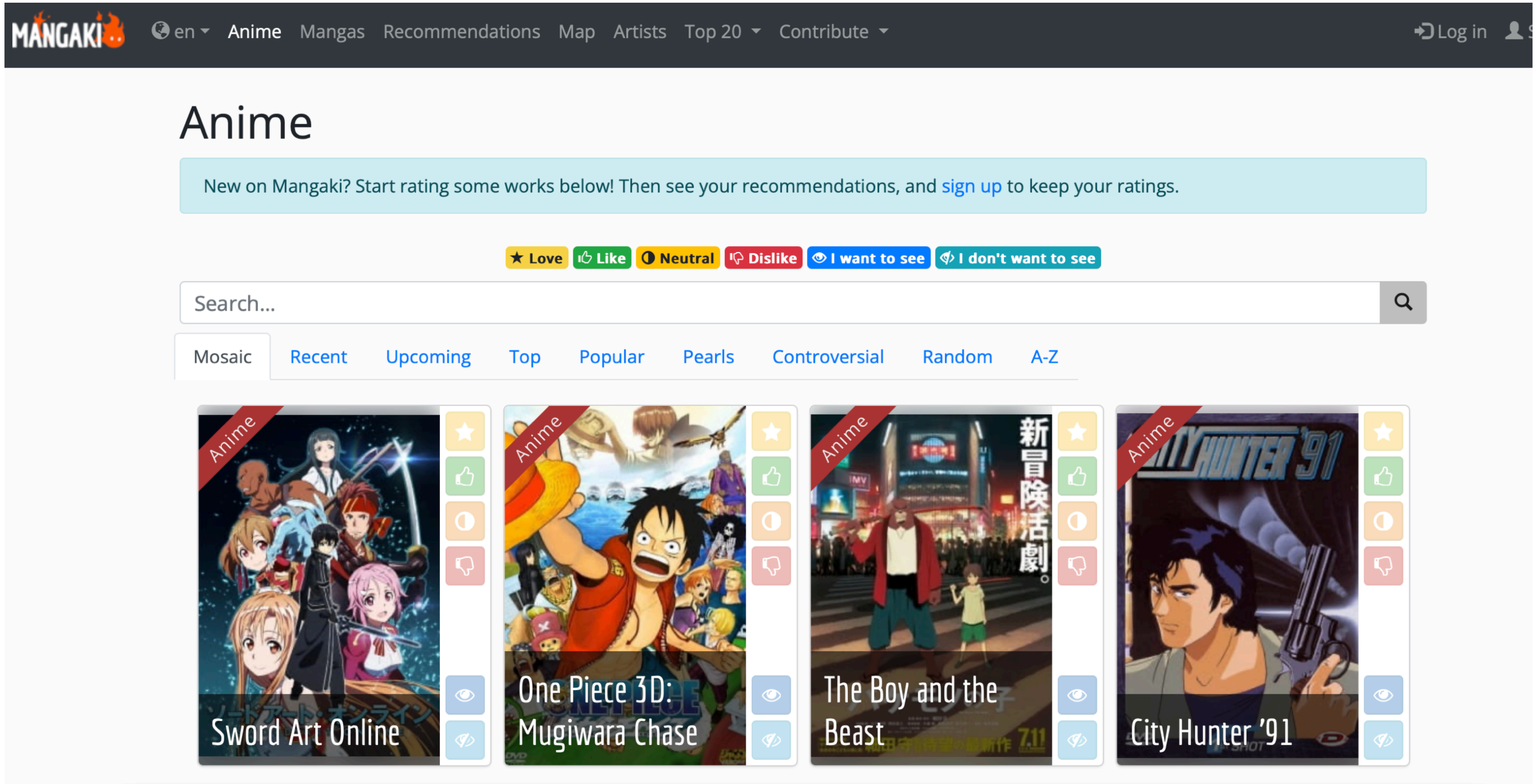
	The Shawshank Redemption	The Godfather	Forrest Gump	Titanic	Pulp Fiction	...
BradPittFan	5	?	?	2	?	...
MerylStreep101	?	3	?	4	?	...
TomHanksLover	?	?	4	?	?	...
ScorseseBuff	?	?	?	?	5	...
...	...	...	...	...	...	...

User-Item Interaction Matrix

	The Shawshank Redemption	The Godfather	Forrest Gump	Titanic	Pulp Fiction	...
BradPittFan	5	?	?	2	?	...
MerylStreep101	?	3	?	4	?	...
TomHanksLover	?	?	4	?	?	...
ScorseseBuff	?	?	?	?	5	...
...	...	...	...	...	...	...

Blackboard

Simple to implement.



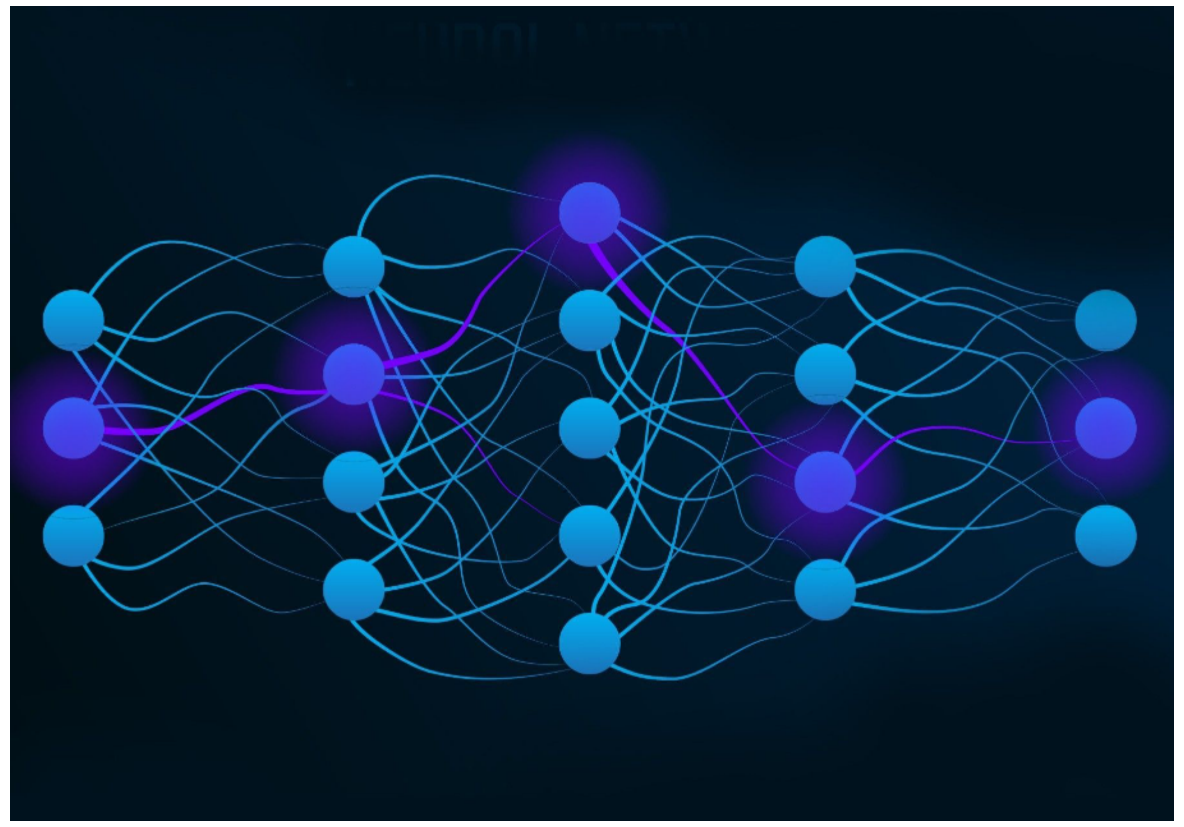
Mangaki, an anime and manga recommendation system.

Second technique.
Using Deep Learning and Embeddings

Architecture of a SOTA recommender system.

Using **deep learning** and **embeddings**.

Key idea. Everything can be a vector!



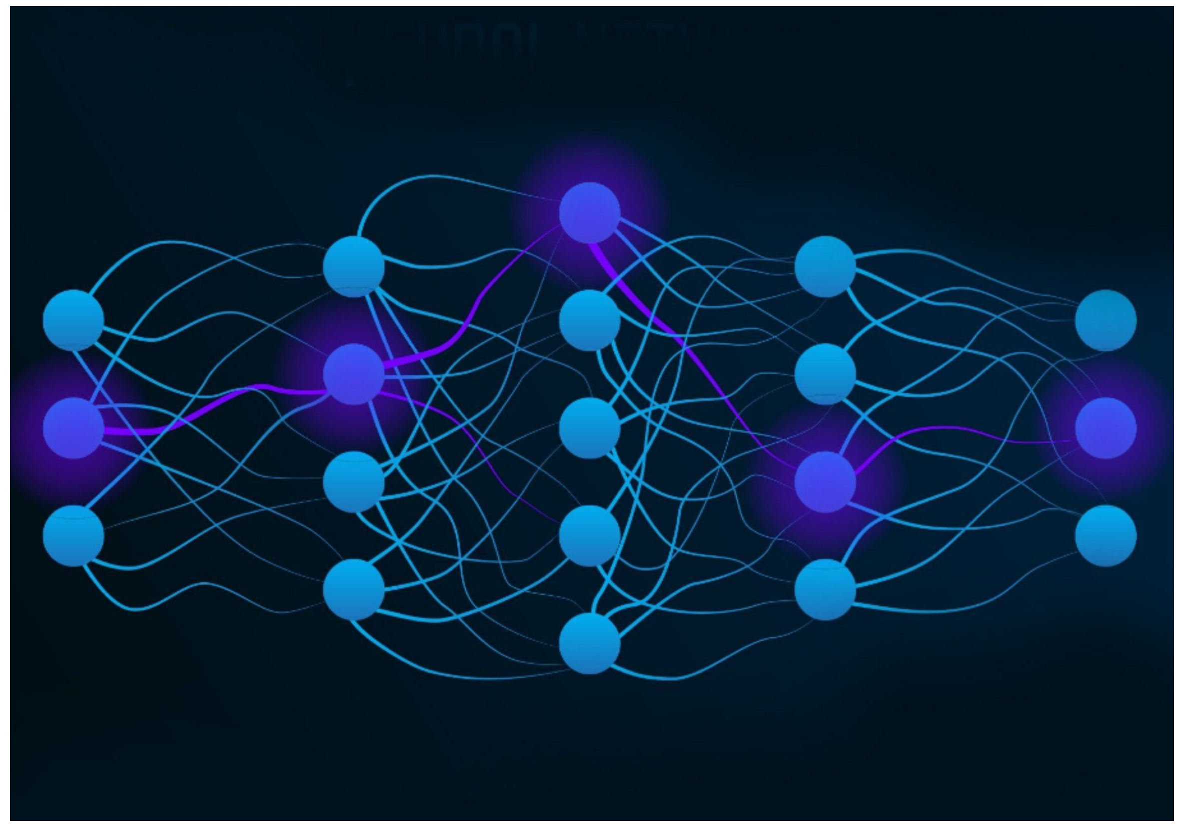
Source. The Scientist Magazine

[0.09; 0.9; 0.3; ...; 0.12]

Architecture of a SOTA recommender system.

Using **deep learning** and **embeddings**.

Key idea. Everything can be a vector!

“A cat is grumpy” →  → [0.11; 0.8; 0.27; ...; 0.1]

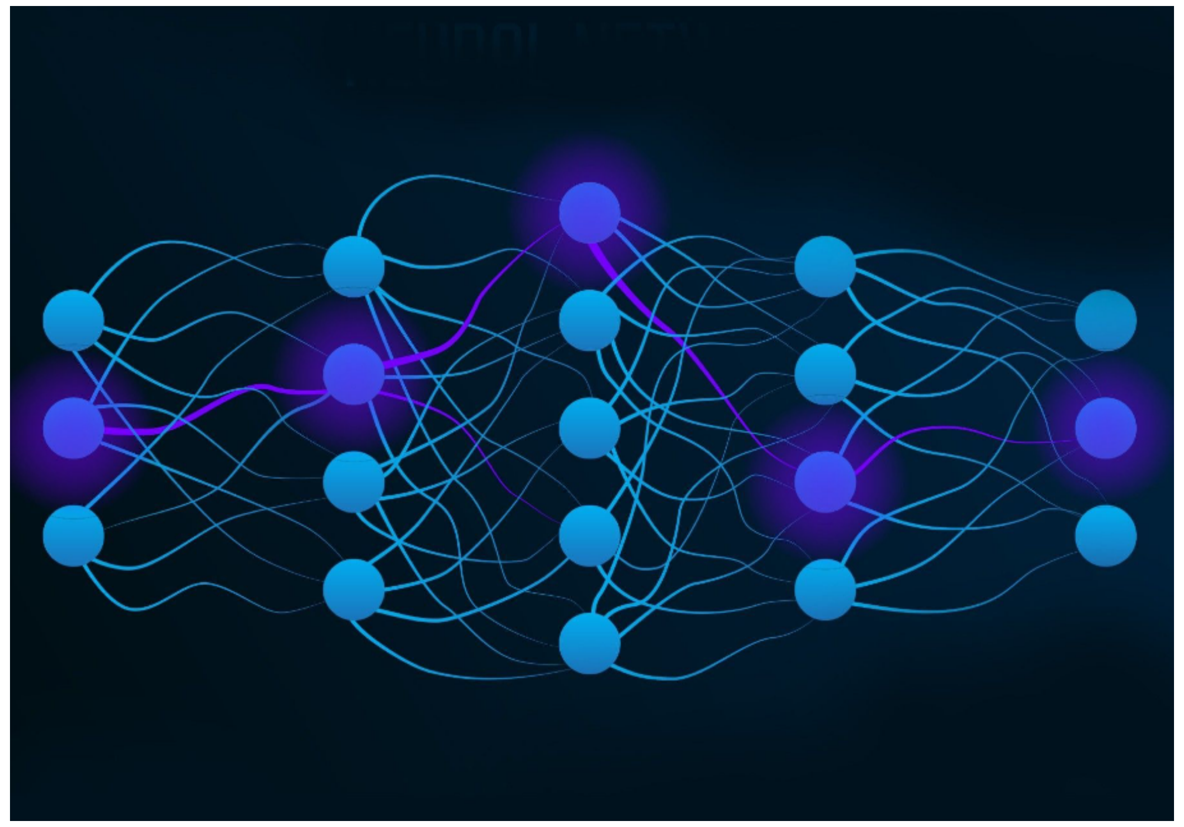
Source. The Scientist Magazine

Architecture of a SOTA recommender system.

Using **deep learning** and **embeddings**.

Key idea. Everything can be a vector!

Similar semantic, similar vectors!



[0.09; 0.9; 0.3; ...; 0.12]

"A cat is grumpy"

[0.11; 0.8; 0.27; ...; 0.1]

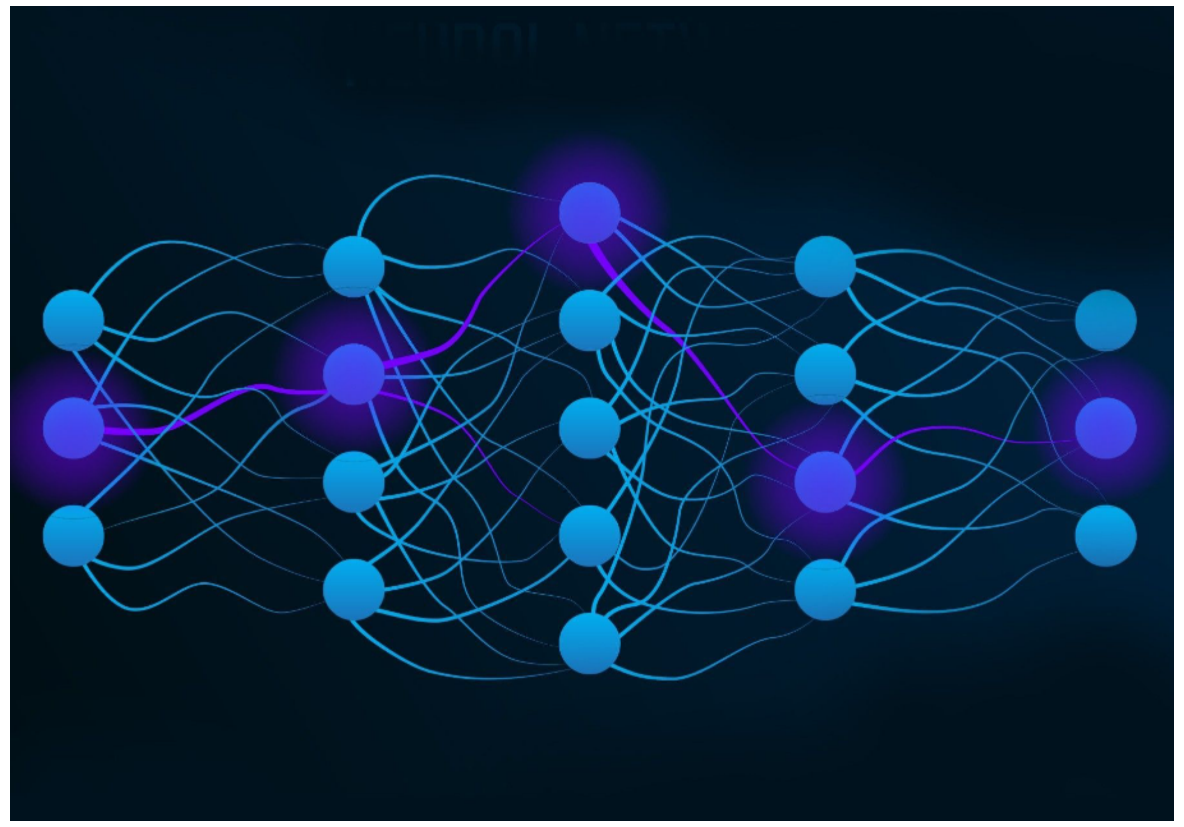
Source. The Scientist Magazine

Architecture of a SOTA recommender system.

Using **deep learning** and **embeddings**.

Key idea. Everything can be a vector!

"Someone playing the trumpet"



Source. The Scientist Magazine

Different semantic, different vectors!

[0.09; 0.9; 0.3; ...; 0.12]

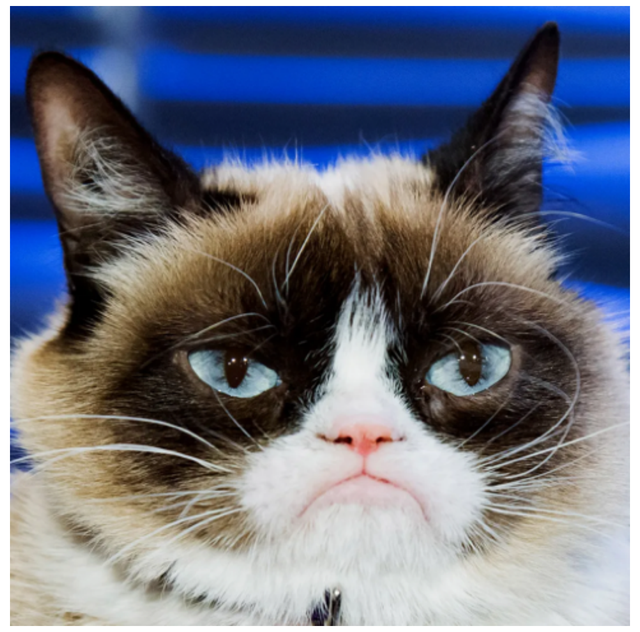
[-0.3; 0.3; 0.7; ...; -0.6]

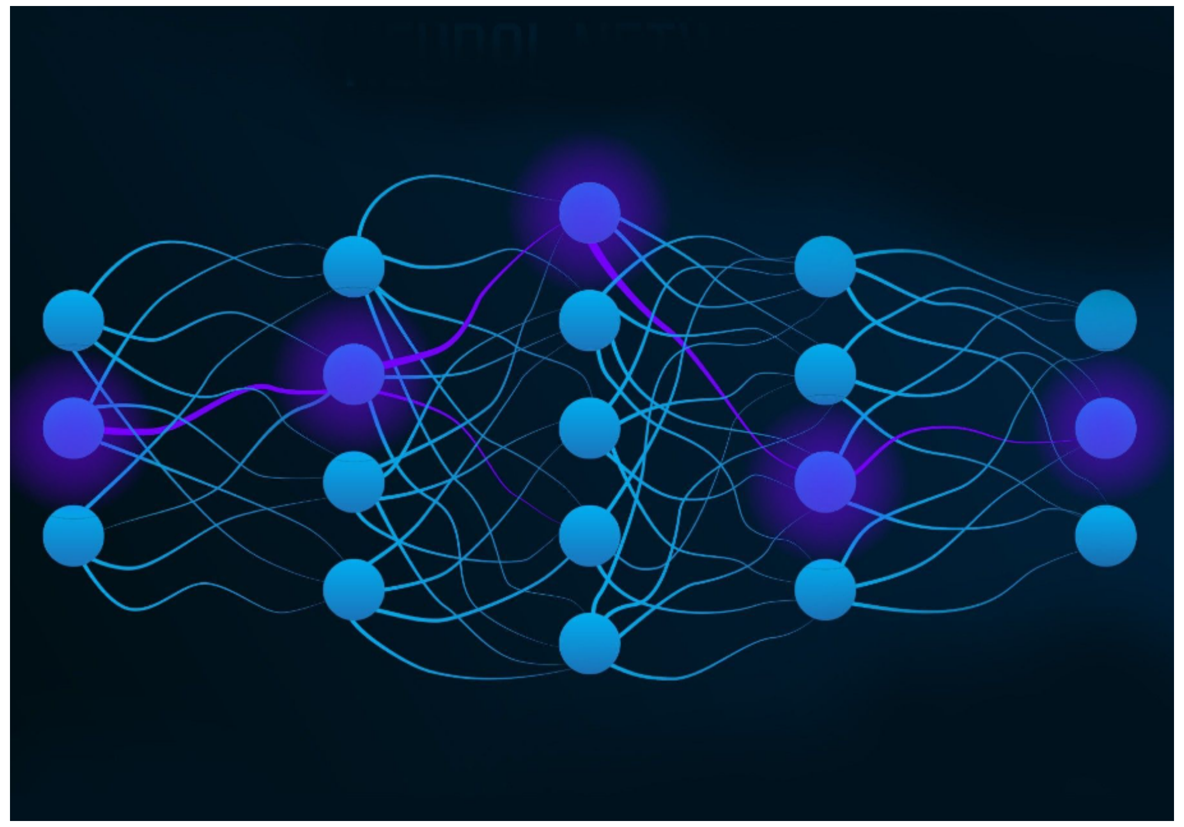
Architecture of a SOTA recommender system.

Using **deep learning** and **embeddings**.

Key idea. Everything can be a vector!

Similar semantic, similar vectors!





Source. The Scientist Magazine

[0.09; 0.9; 0.3; ...; 0.12]

[0.08; 0.85; 0.32; ...; 0.1]

Blackboard

How to ensure diversity?

What about diversity?

Suppose the recommendations (highest scores given by the system) are

1. Star Wars: A New Hope (1977)
2. Star Wars: The Empire Strikes Back (1980)
3. Star Wars: Return of the Jedi (1983)
4. Star Wars: The Phantom Menace (1999)
5. Star Wars: Attack of the Clones (2002)

All relevant but.... **BORING!**

What about diversity?

How to switch to

1. Star Wars: A New Hope (1977) (Highly relevant)
2. Dune (2021) (Epic science fiction, different franchise)
3. Star Wars: The Empire Strikes Back (1980) (Relevant sequel)
4. The Matrix (1999) (Diverse, groundbreaking sci-fi)
5. Interstellar (2014) (Modern sci-fi with emotional depth)

?

MMR - maximal marginal relevance.

S = list of movies already recommended

1. Star Wars: A New Hope
2. Dune
3. Star Wars: The Empire Strikes Back
4. **???**

MMR - maximal marginal relevance.

S = list of movies already recommended

1. Star Wars: A New Hope
2. Dune
3. Star Wars: The Empire Strikes Back
4. **???**

Naive idea

Next_Movie = argmax score(Movie)

Movie	Score
Star Wars: Return of the Jedi	**14**
Star Wars: The Phantom Menace	**13**
Star Wars: Attack of the Clones	**12**
The Matrix	**10**
Interstellar	**8**

MMR - maximal marginal relevance.

S = list of movies already recommended

1. Star Wars: A New Hope
2. Dune
3. Star Wars: The Empire Strikes Back

4. ???

MMR

Next_Movie = argmax score(Movie) - similarity(Movie, S)

Movie	Score
Star Wars: Return of the Jedi	**14 - 8 = 6**
Star Wars: The Phantom Menace	**13 - 8 = 5**
Star Wars: Attack of the Clones	**12 - 8 = 4**
The Matrix	**10 - 3 = 7**
Interstellar	**8 - 2 = 6**

Impacts

Mat Honan's experiment (2014)

- Andy Warhol "I think everybody should like everybody."
- "I like everything. Or at least I did, for 48 hours. Literally everything Facebook sent my way, I liked—even if I hated it."

Wikipedia:Getting to Philosophy

文A 6 languages

Project page Talk Read Edit View history Tools

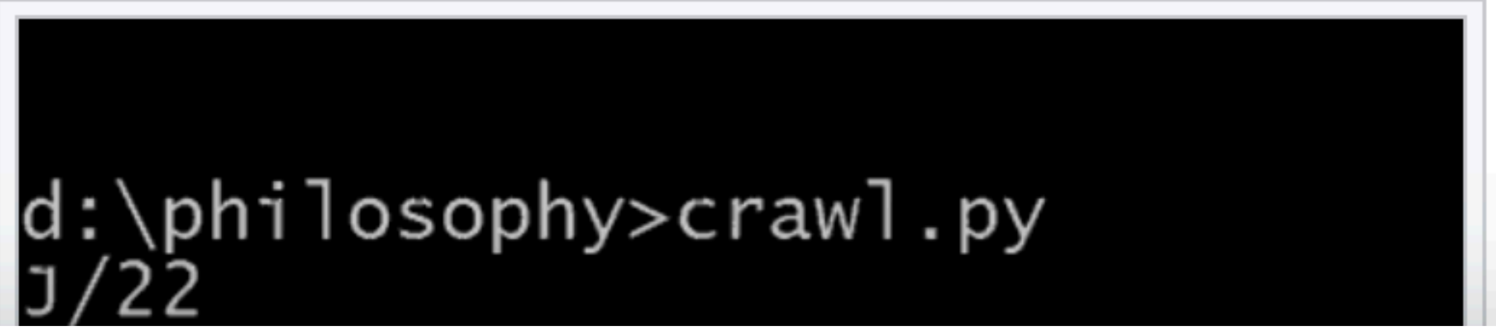
From Wikipedia, the free encyclopedia

> Please do not make edits to articles that are motivated by the concepts on this page. Such edits are purely disruptive and are generally quickly reverted, as they clearly do not improve the encyclopedia.

Shortcut
WP:GTP

Following the first hyperlink in the main text of an English Wikipedia article, and then repeating the process for subsequent articles, usually leads to the Philosophy article. In February 2016, this was true for 97% of all articles on Wikipedia[1] (including this one), an increase from 94.52% in 2011. The remaining articles lead to an article without any outgoing wikilinks, to pages that do not exist, or get stuck in loops.

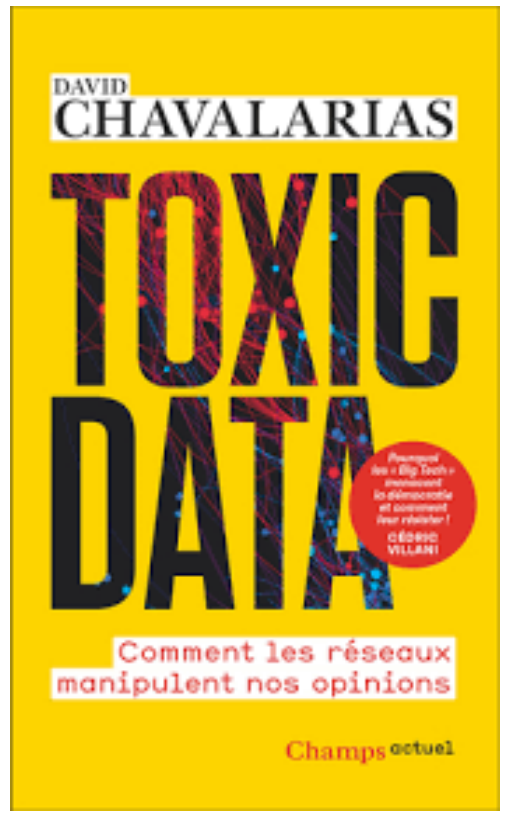
There have been some theories on this phenomenon, with the most prevalent being the tendency for Wikipedia pages



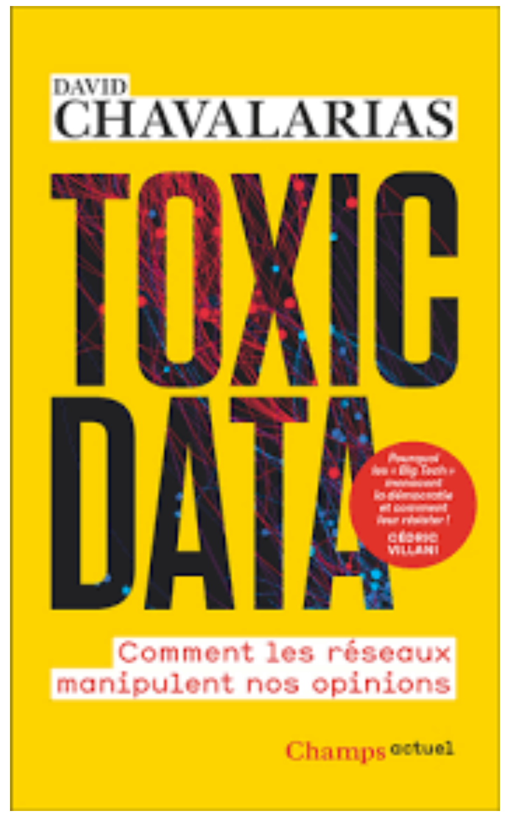

The Truman Show

Do you live in a parallel world?

Let's try to understand Mat Honan's experiment.

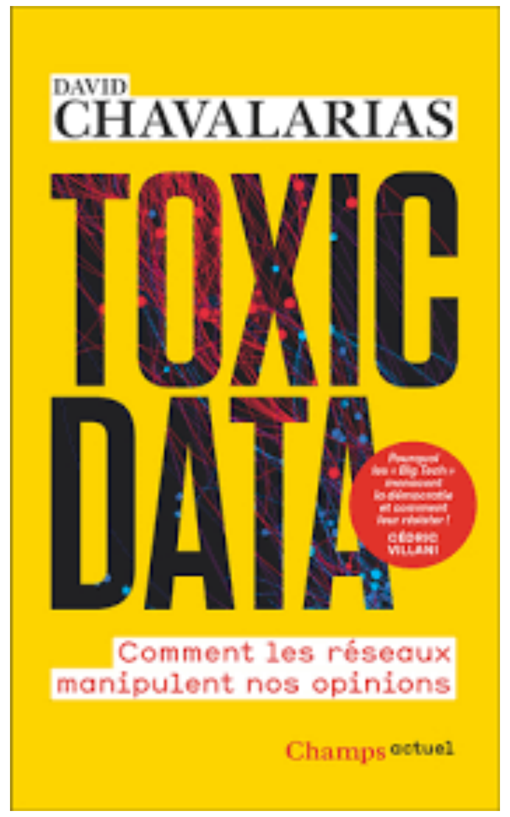


Let's try to understand Mat Honan's experiment.



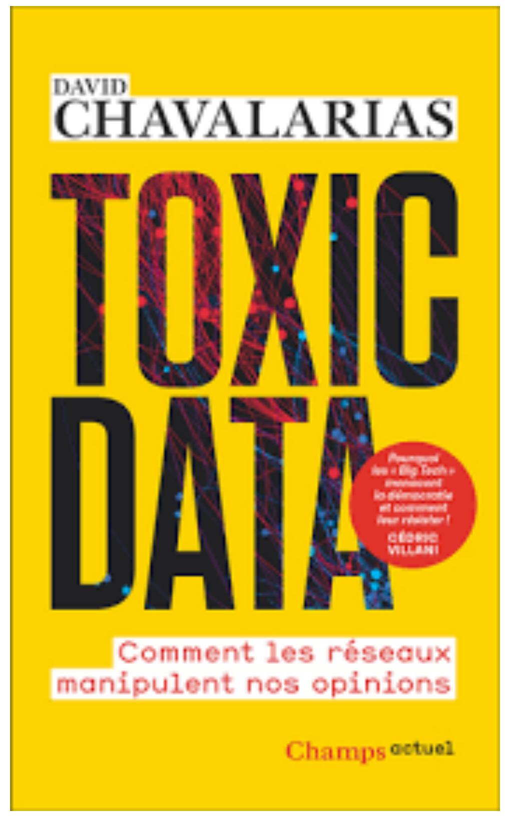
- ***Reinforcement.*** Based on an initial like, the system proposes more content aligned with that preference.

Let's try to understand Mat Honan's experiment.



- ***Reinforcement.*** Based on an initial like, the system proposes more content aligned with that preference.
- ***Algorithmic contagion.*** The algorithm influences the social environment by amplifying specific content (collaborative filtering as well)

Let's try to understand Mat Honan's experiment.

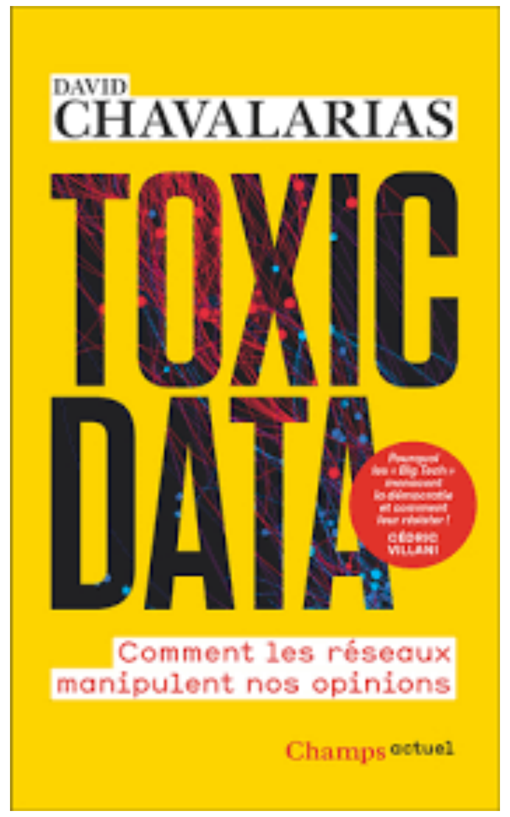


- ***Reinforcement.*** Based on an initial like, the system proposes more content aligned with that preference.

- ***Algorithmic contagion.*** The algorithm influences the social environment by amplifying specific content (collaborative filtering as well)

Convergence to an **Echo chamber** or **Bubble filter.** The information space is reduced, unique point of view biased toward initial interactions.

Let's try to understand Mat Honan's experiment.



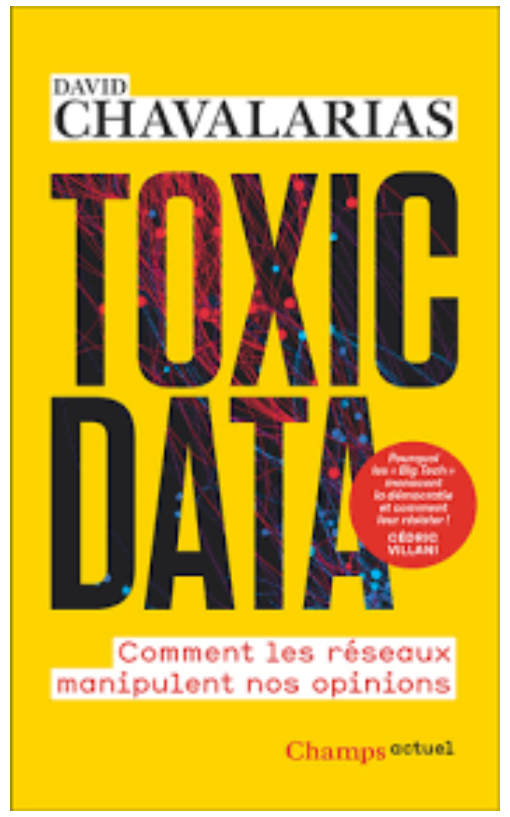
- ***Reinforcement.*** Based on an initial like, the system proposes more content aligned with that preference.

- ***Algorithmic contagion.*** The algorithm influences the social environment by amplifying specific content (collaborative filtering as well)

Convergence to an **Echo chamber** or **Bubble filter.** The information space is reduced, unique point of view biased toward initial interactions.

Coupled with psychological biases

Let's try to understand Mat Honan's experiment.



- ***Reinforcement.*** Based on an initial like, the system proposes more content aligned with that preference.

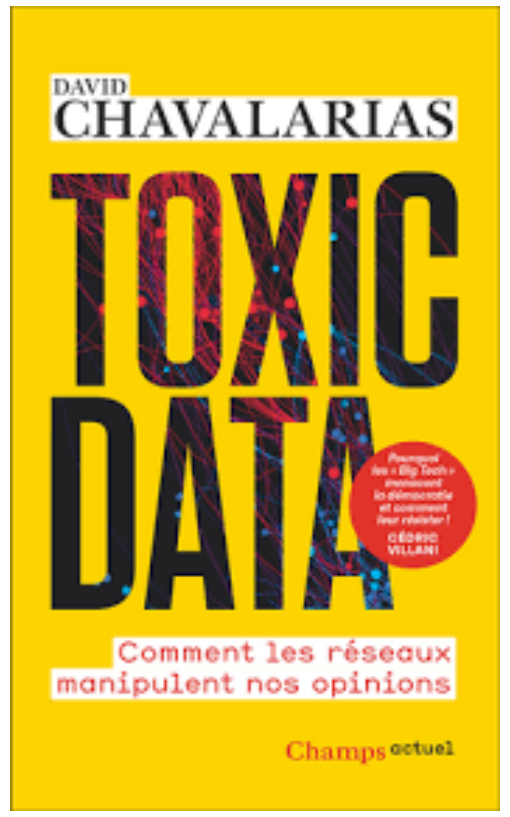
- ***Algorithmic contagion.*** The algorithm influences the social environment by amplifying specific content (collaborative filtering as well)

Convergence to an **Echo chamber** or **Bubble filter.** The information space is reduced, unique point of view biased toward initial interactions.

Coupled with psychological biases

- **Confirmation bias.** Humans tend to confirm their existing beliefs.

Let's try to understand Mat Honan's experiment.



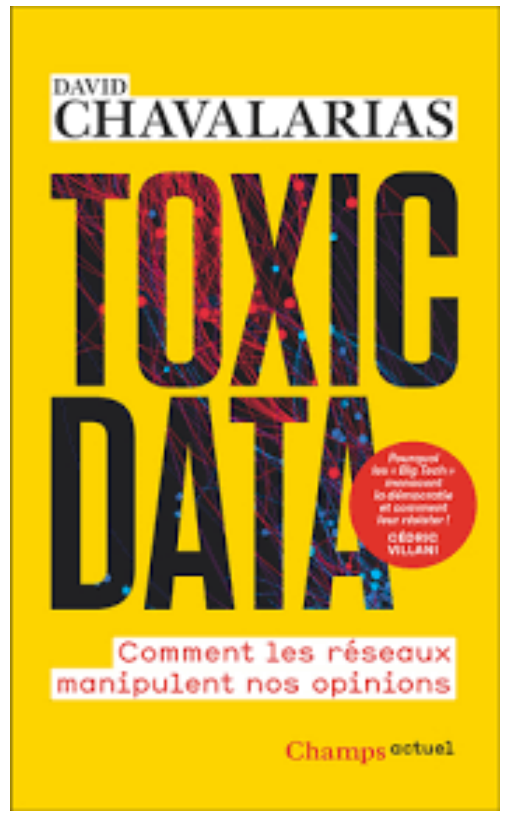
- ***Reinforcement.*** Based on an initial like, the system proposes more content aligned with that preference.
- ***Algorithmic contagion.*** The algorithm influences the social environment by amplifying specific content (collaborative filtering as well)

Convergence to an **Echo chamber** or **Bubble filter.** The information space is reduced, unique point of view biased toward initial interactions.

Coupled with psychological biases

- **Confirmation bias.** Humans tend to confirm their existing beliefs.
- **Negativity bias. Stronger than confirmation bias** (asymmetry in the way humans process data).

Let's try to understand Mat Honan's experiment.



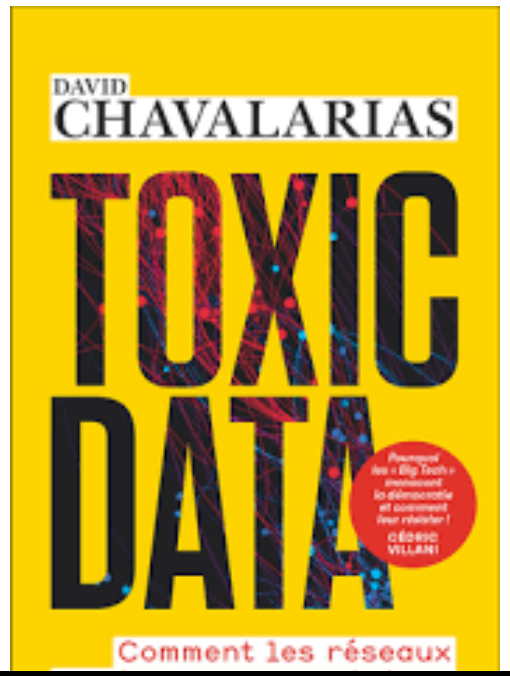
- ***Reinforcement.*** Based on an initial like, the system proposes more content aligned with that preference.
- ***Algorithmic contagion.*** The algorithm influences the social environment by amplifying specific content (collaborative filtering as well)

Convergence to an **Echo chamber** or **Bubble filter.** The information space is reduced, unique point of view biased toward initial interactions.

Coupled with psychological biases

- **Confirmation bias.** Humans tend to confirm their existing beliefs.
- **Negativity bias. Stronger than confirmation bias** (asymmetry in the way humans process data).
- —> Welcome to the Truman show! Parallel "fake" world.

Let's try to understand Mat Honan's experiment.



What if intentional amplifications added into the system? And how?

- ***Reinforcemen*** … preference.
- ***Algorithmic c*** … cific content (collaborative f

Convergence to … point of view biased toward ini

Coupled with psy

- **Confirmation**
- **Negativity bias. Stronger than confirmation bias** (asymmetry in the way humans process data).
- —> Welcome to the Truman show! Parallel "fake" world.

Conclusion

- Recommender systems are a new way to **navigate information** that create opportunities for niche content and small creators, such as independent artists and writers
- **Easier** to develop than ever
- The **scale** at which they are deployed can have an **impact**.
- Understanding **how they work** - and **how we function** - helps us avoid falling into some traps.
- Lots of **interesting research** to develop better recommendation algorithms.
- We should ask for **open-source recommender systems** (transparency, fairness, etc.).