Recommenders in the Wild Part I

Tutorial RecSys 2023.

Kim Falk, September 18th 2023

Kim Falk

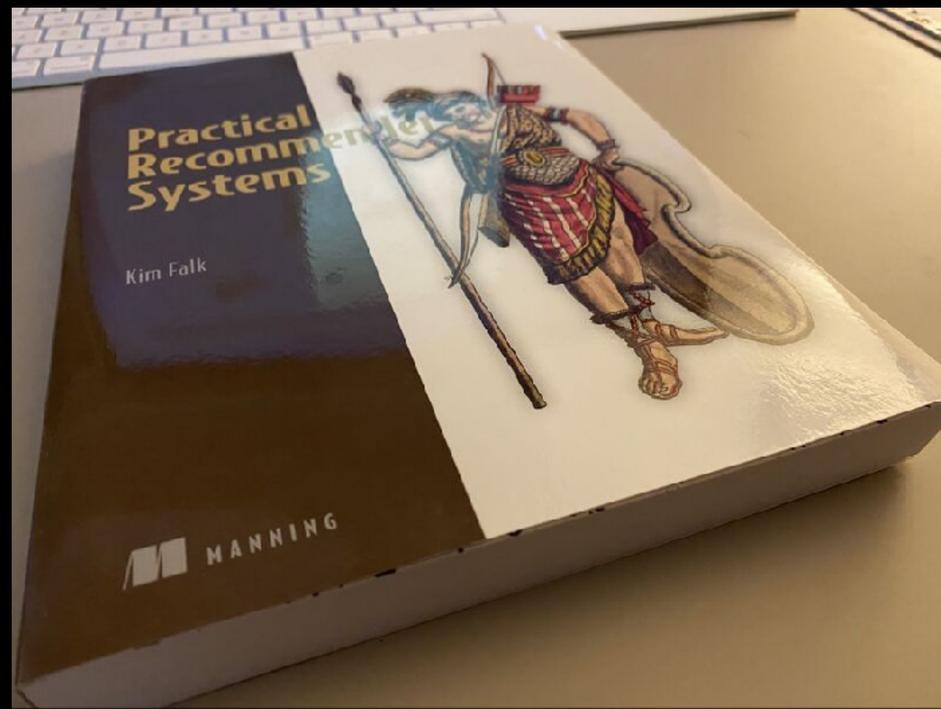
Computer scientist since 2003

Worked with big data and ML and recommenders since 2010

Author of Practical recommender systems

RecSys Industry Chair 2019 and 2020

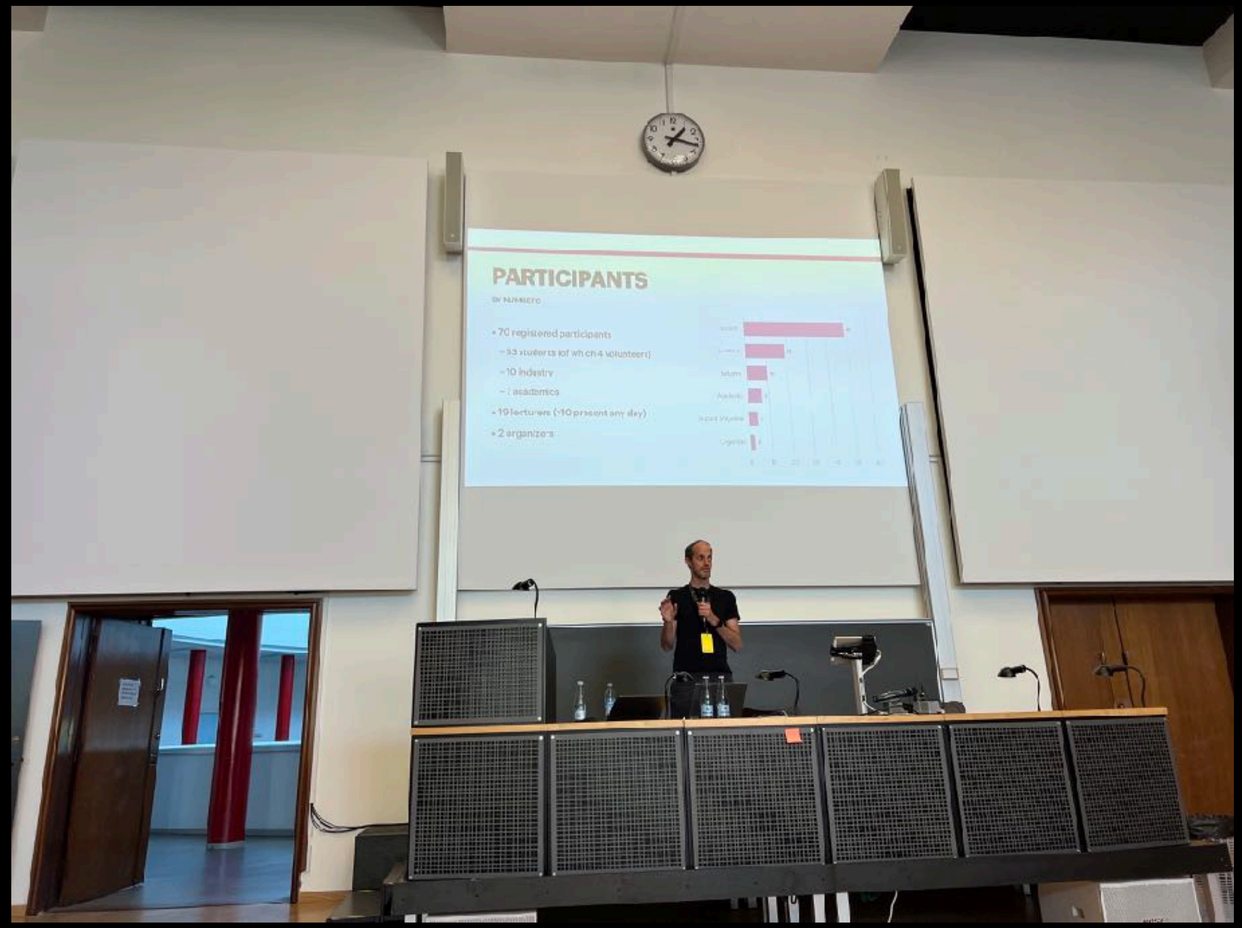
Principle Data Scientist at ViSENZE



Assumptions for this presentation

I am assuming that people listening are mostly people from academia. Non academics are welcome too

I also assume that you know what a recommender system is



The goal is to share my personal experience of working with recommender systems, trying to add new perspectives which are not commonly discussed at conferences.

Agenda

What is a recommender in the wild

Developing the recommender

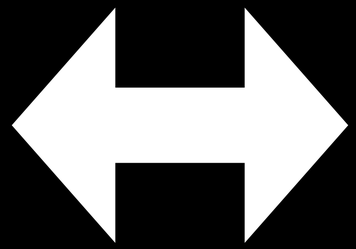
Beyond accuracy

Business considerations


What is
a Recommender in
the wild?

Whats a recommender in the wild

Recommender in the wild



Production system servicing recommendations to people

Recommenders are domain specific

Movielens-1M	Recall@10	0.2038	0.2273	**0.2354**	+15.5%	+3.56%
	NDCG@10	0.3396	0.3705	**0.4123**	+21.4%	+11.3%
	Recall@20	0.3025	0.3126	**0.3179**	+5.09%	+1.70%
	NDCG@20	0.3387	0.3566	**0.3949**	+16.6%	+10.7%
	Time per epoch	16.5s	30.5s	17.1s	-	-
Gowalla	Recall@10	0.1153	0.1242	**0.1296**	+12.4%	+4.35%
	NDCG@10	0.1258	0.1375	**0.1456**	+15.7%	+5.89%
	Recall@20	0.1674	0.1784	**0.1838**	+9.80%	3.03%
	NDCG@20	0.1404	0.1524	**0.1594**	+13.5%	+4.59%
	Time per epoch	28.3s	50.1s	28.7s	-	-
Yelp2018	Recall@10	0.0352	0.0397	**0.0411**	+16.8%	+3.53%
	NDCG@10	0.0446	0.0477	**0.0526**	+17.9%	+10.3%
	Recall@20	0.0591	0.0673	**0.0686**	+16.1%	+1.93%
	NDCG@20	0.0519	0.0571	**0.0608**	+17.1%	+6.48%
	Time per epoch	45.8s	88.6s	46.1s	-	-
Amazon-Book	Recall@10	0.0207	0.0241	**0.0267**	+29.0%	+10.8%
	NDCG@10	0.0223	0.0259	**0.0294**	+31.9%	+13.5%
	Recall@20	0.0366	0.0421	**0.0451**	+23.2%	+7.13%
	NDCG@20	0.0286	0.0329	**0.0363**	+26.9%	+10.3%
	Time per epoch	127.8s	230.1s	128.9s	-	-

Recommenders are domain specific

Why ?

Mostly due to

Streaming data

E-commerce

Other

Why ?

Mostly due to

Streaming data

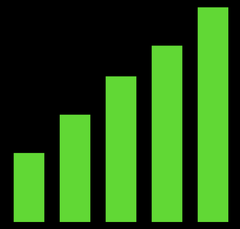
Video on demand | Music | Social Media

E-commerce

Online stores

Why ?

Mostly due to



Popularity and trends of content

Speed of feedback

Size of data

Recommenders are domain specific

One humans rubbish model is another humans gold model?


I will focus on
the E ce
I will focus on the (mostly)
E-commerce

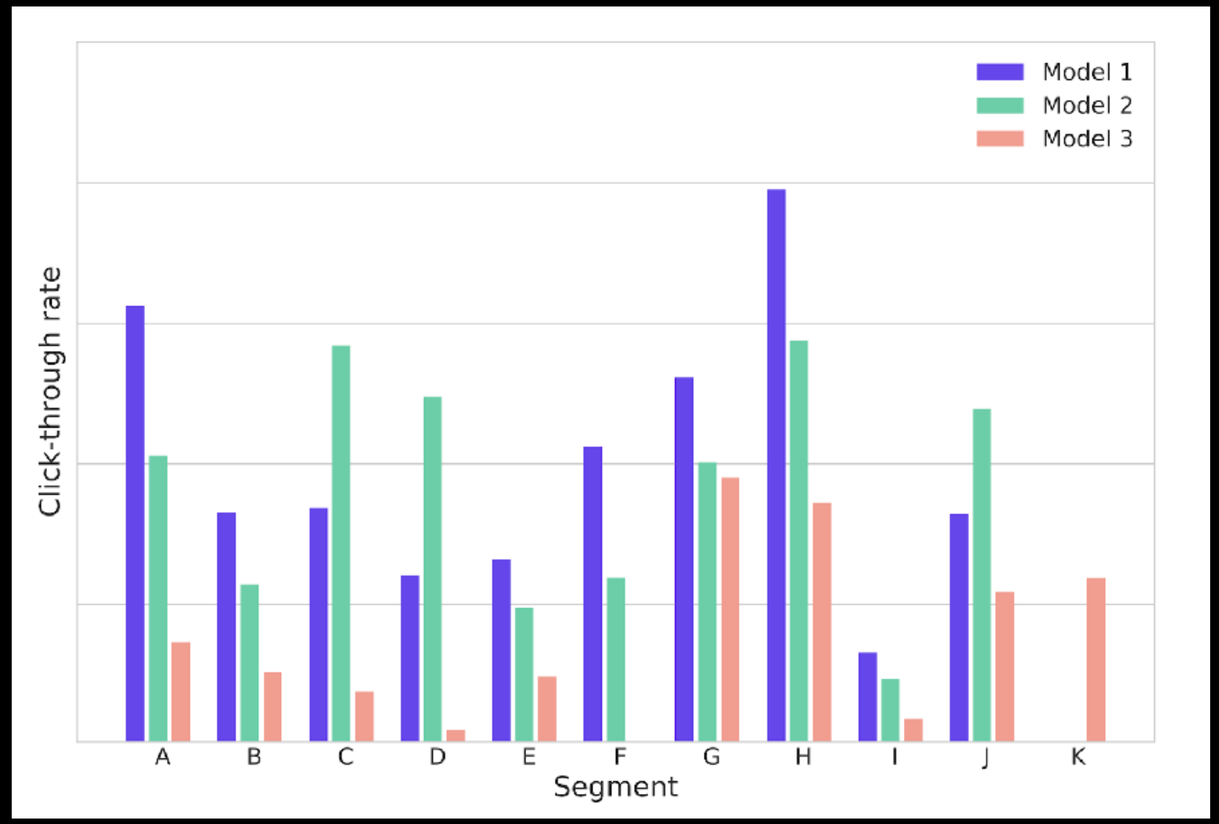
I will focus on the E ce

I will focus on the (mostly) E-commerce

But you can probably use the same considerations in scenarios

Recommender systems are even sub-domain specific

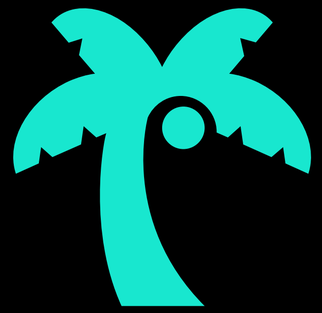
Different domains has different signals



K.Falk et all: Optimizing product recommendations for millions of merchants

Find content for users

Users



When I say “*Find content for users*” I really mean

Rerank lists of content

When I say “*Find content for users*” I really mean

Is there one optimal rank of content?

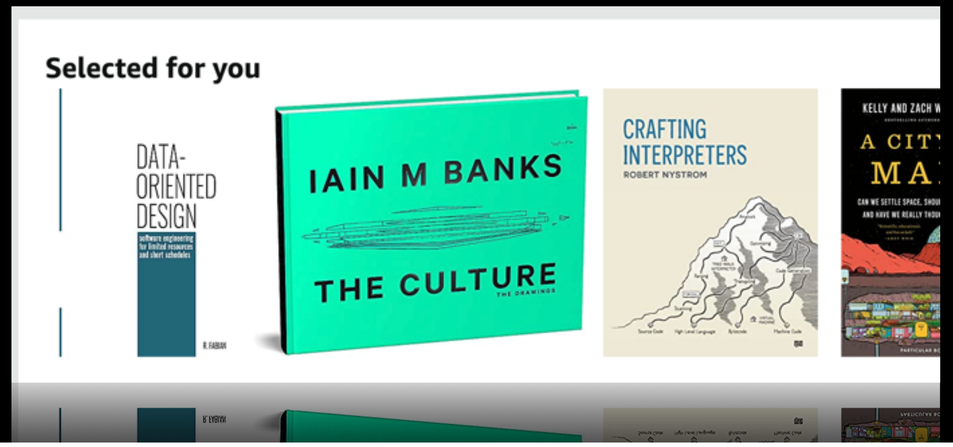
Find Content for Users

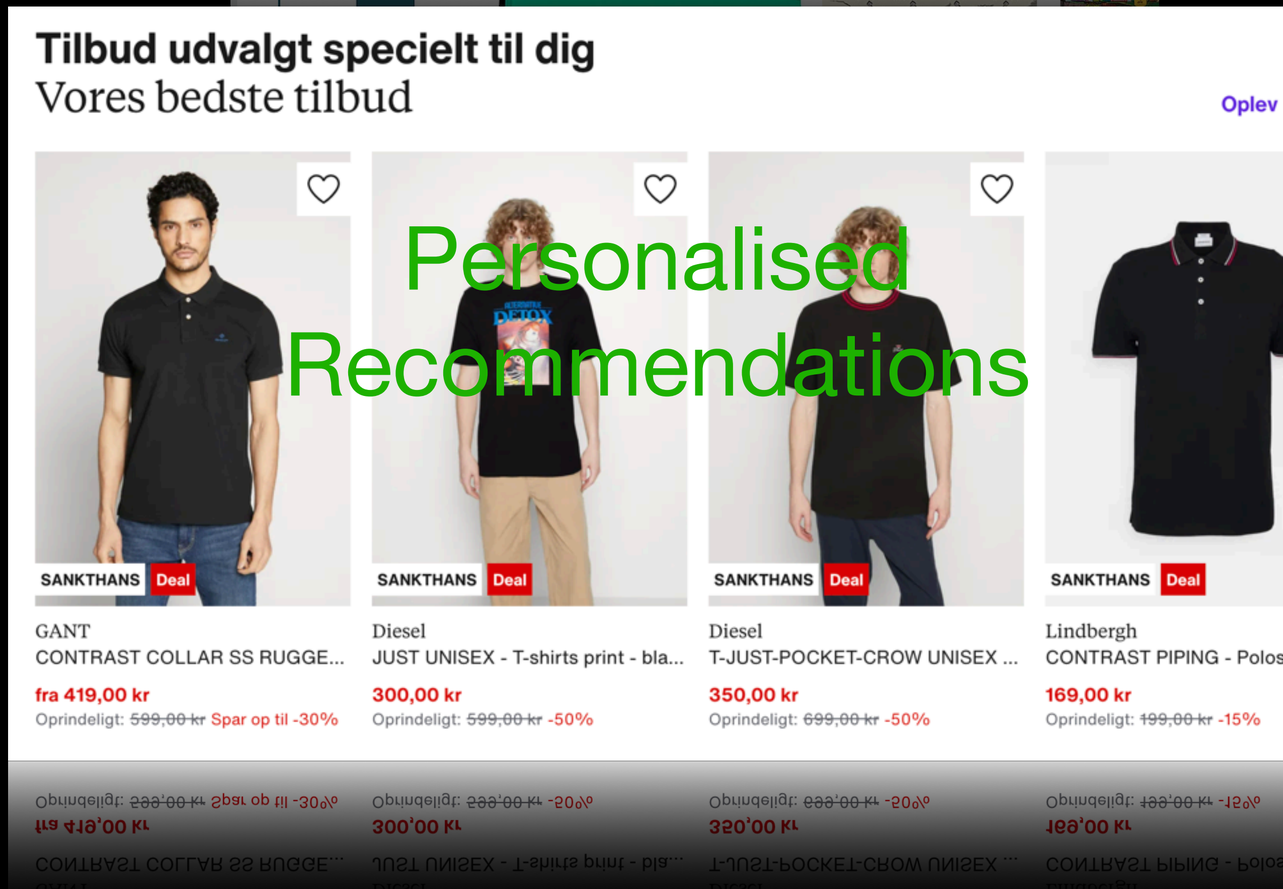
- What they need now
- What they didn’t know that they needed, but did
- Educate them on what else the catalogue contains.

Supports the current use case

Find Content for Users

- What they need now
- What they didn’t know that they needed, but did
- Educate them on what else the catalogue contains.
- Supports the current use case

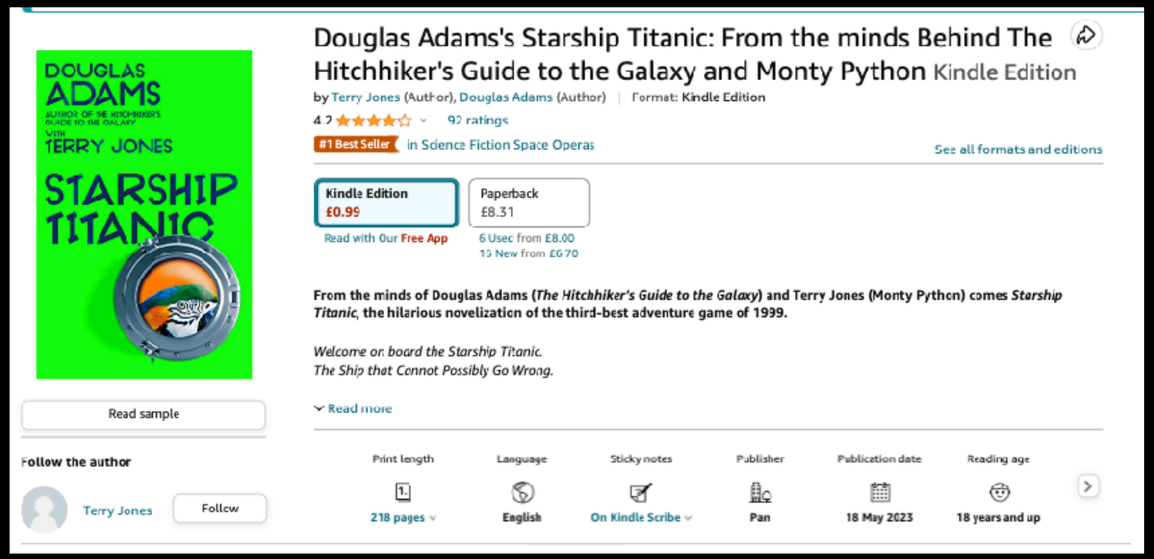




Personalised Recommendations

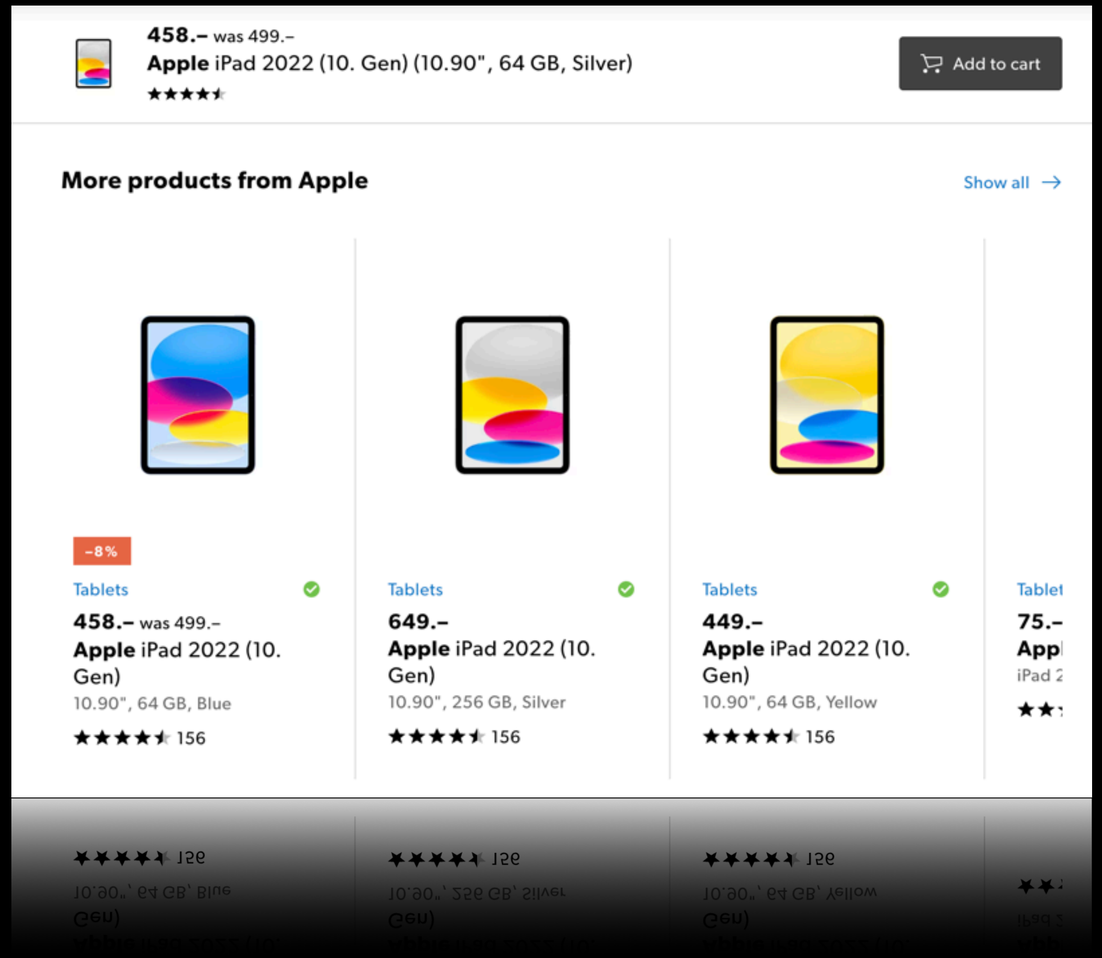
Find Content for Users

- What they need now
- **What they didn’t know that they needed, but did**
- Educate them on what else the catalogue contains.
- Supports the current use case



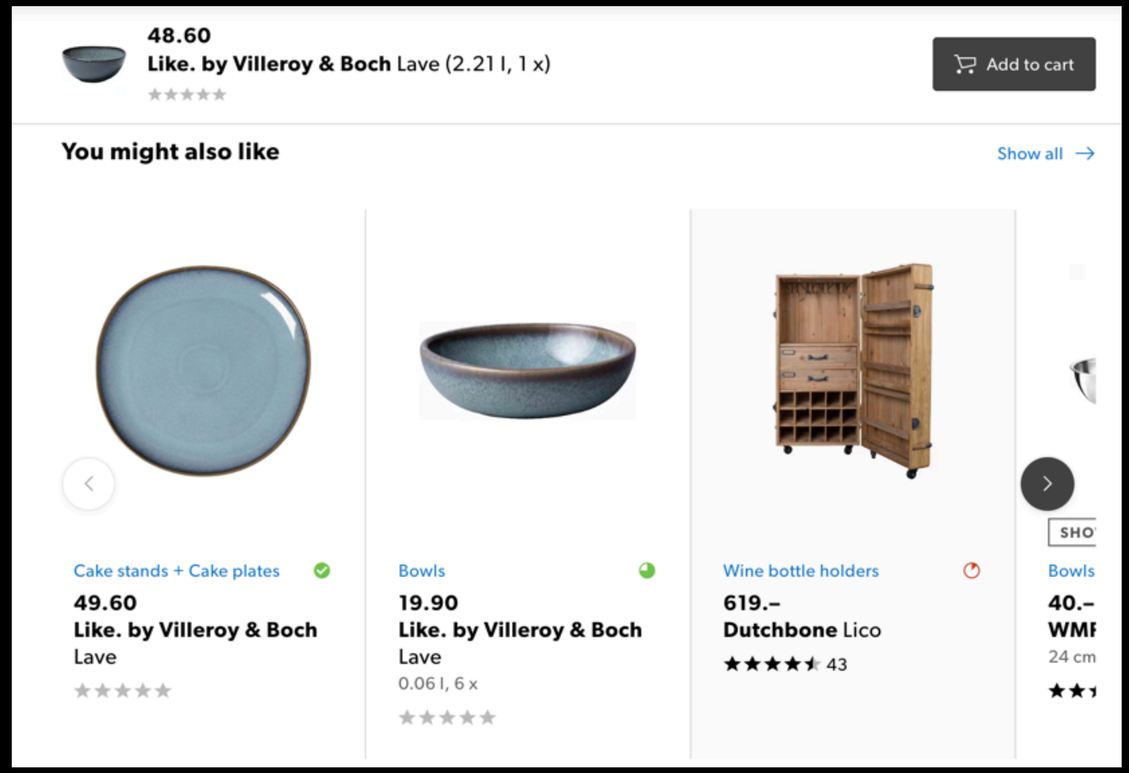
Find Content for Users

- What they need now
- What they didn’t know that they needed, but did
- **Educate them on what else the catalogue contains.**
- Supports the current use case



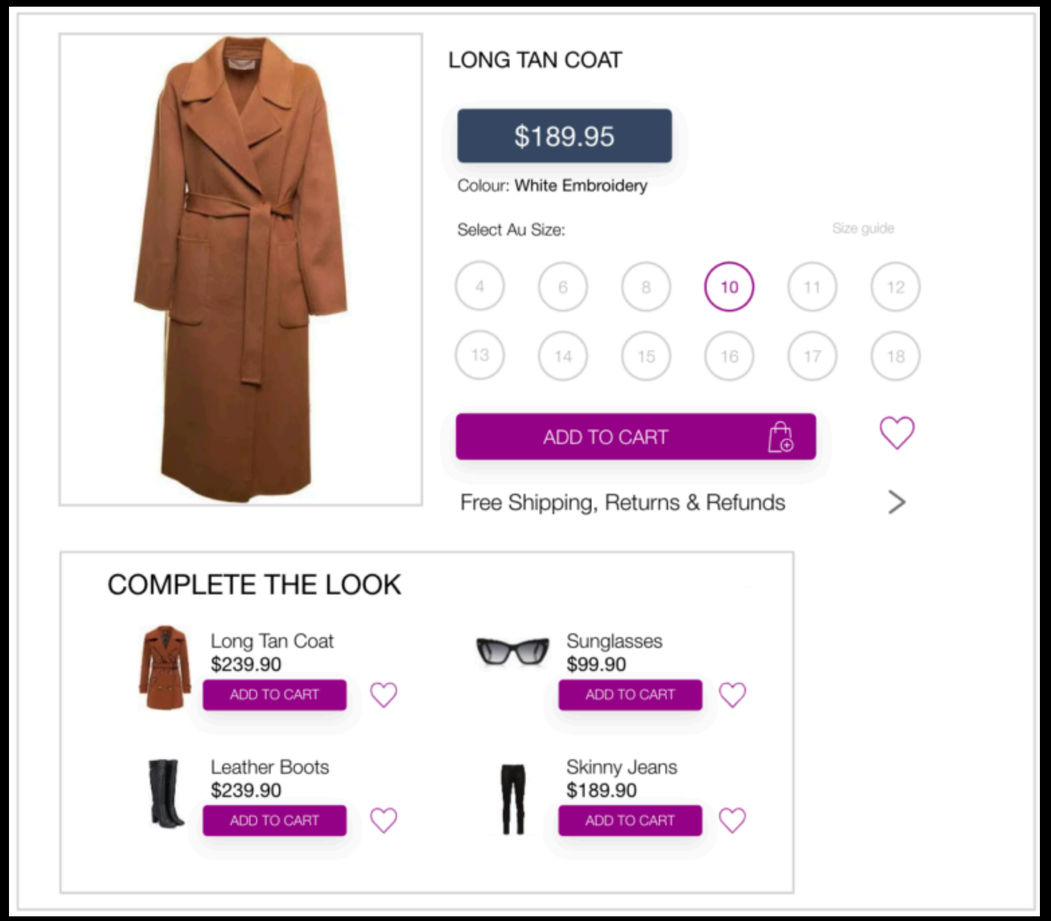
Find Content for Users

- What they need now
- What they didn’t know that they needed, but did
- **Educate them on what else the catalogue contains.**
- Supports the current use case



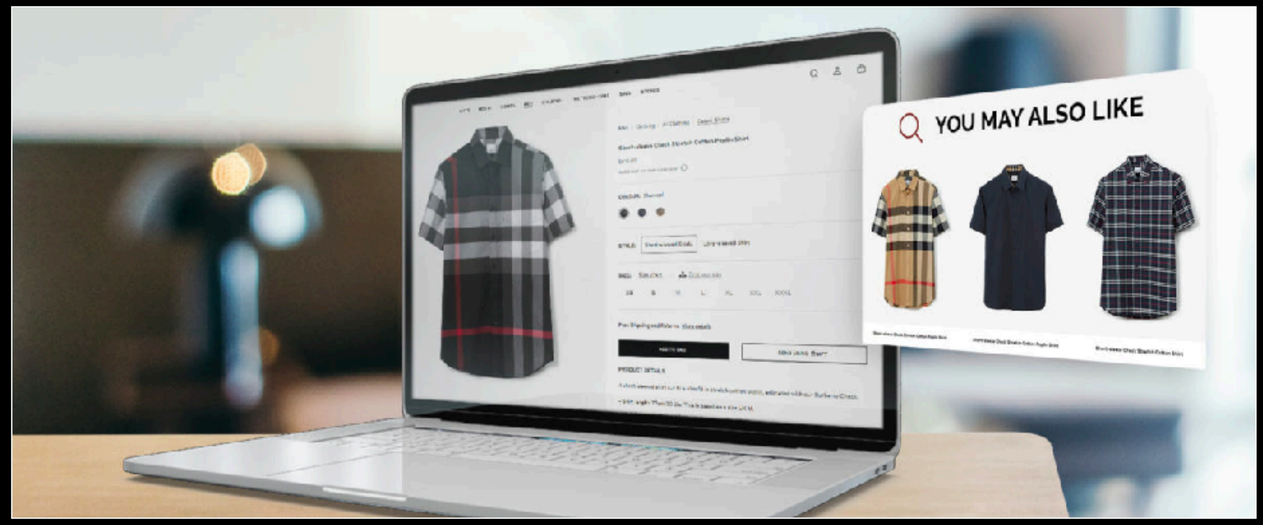
Find Content for Users

- What they need now
- What they didn’t know that they needed, but did
- **Educate them on what else the catalogue contains.**
- Supports the current use case



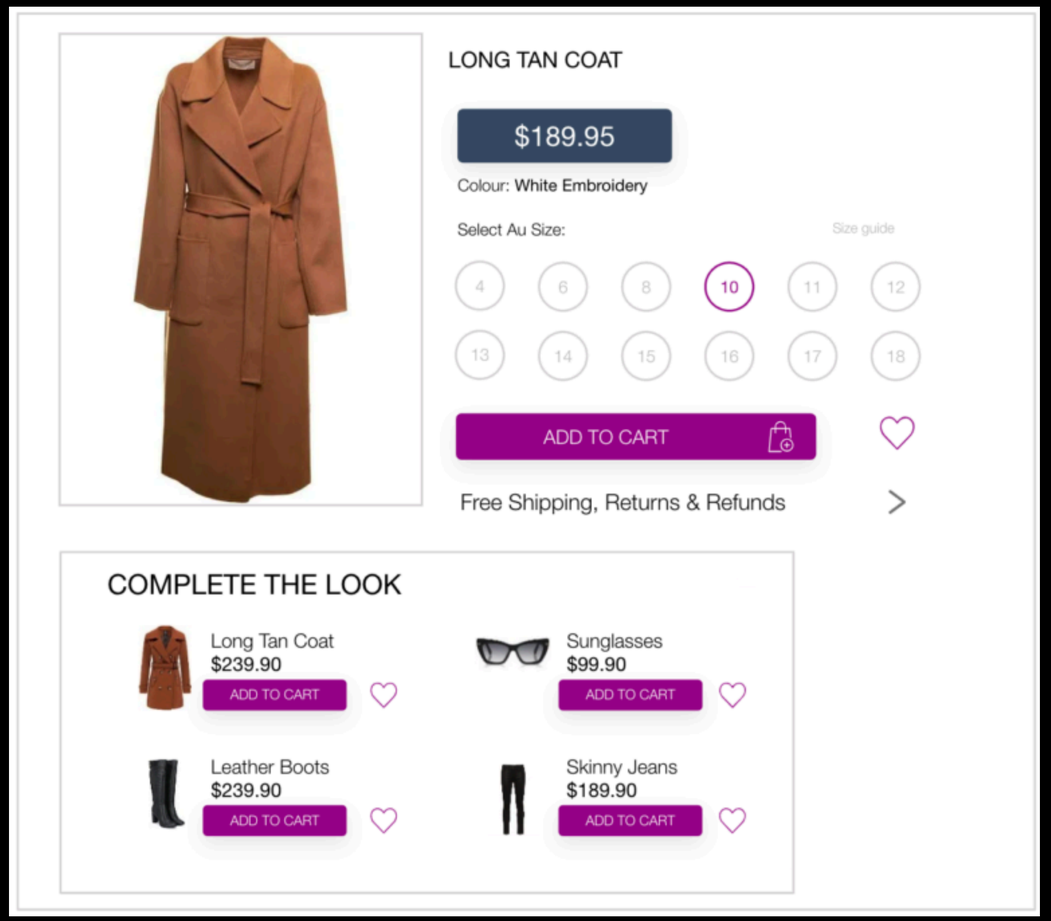
Find Content for Users

- What they need now
- What they didn’t know that they needed, but did
- Educate them on what else the catalogue contains.
- **Supports the current use case**



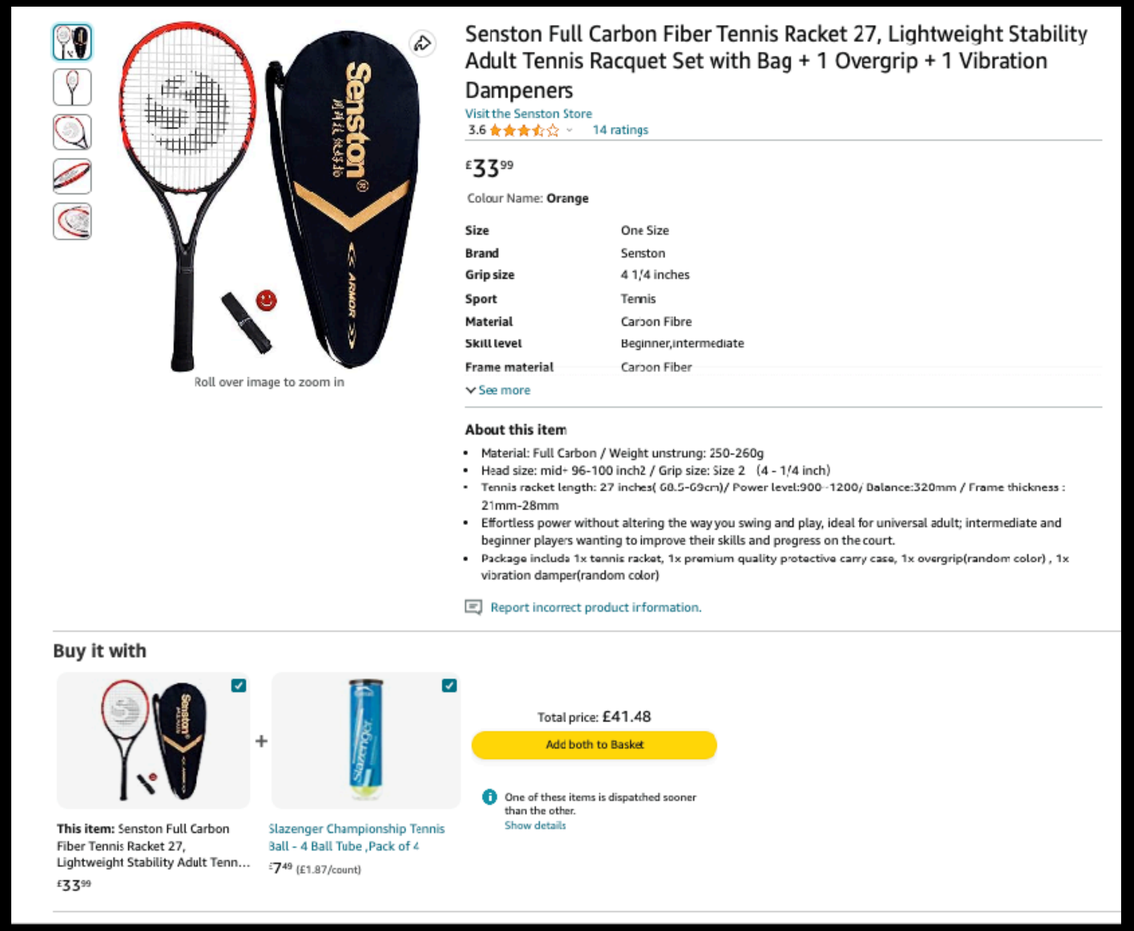
Find Content for Users

- What they need now
- What they didn’t know that they needed, but did
- Educate them on what else the catalogue contains.
- **Supports the current use case**



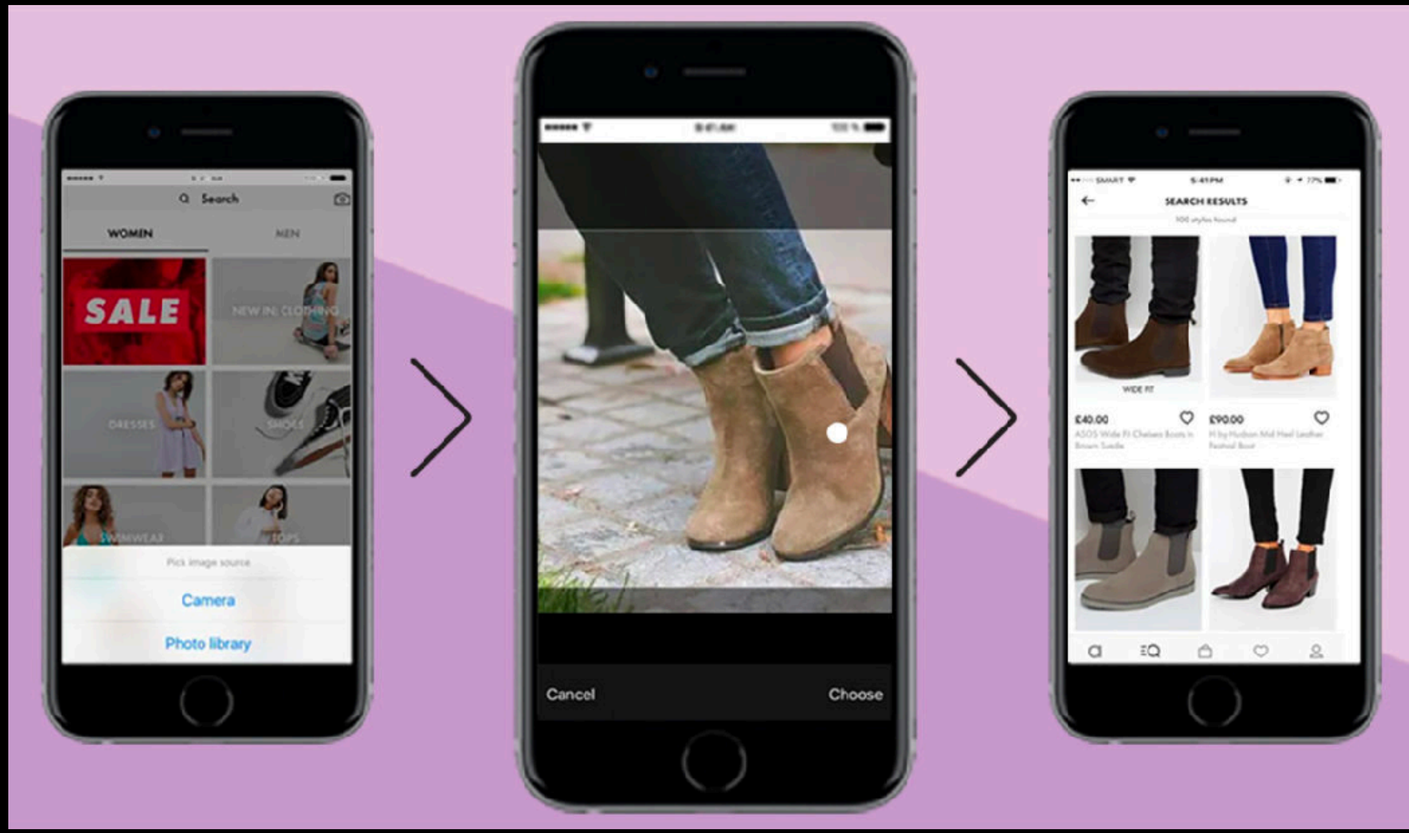
Find Content for Users

- What they need now
- What they didn’t know that they needed, but did
- Educate them on what else the catalogue contains.
- **Supports the current use case**



Find Content for Users

- What they need now
- What they didn’t know that they needed, but did
- Educate them on what else the catalogue contains.
- **Supports the current use case**

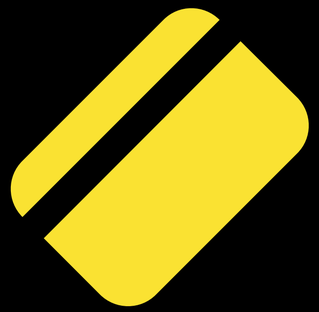


https://www.visenze.com/blog/2023/06/22/the-rise-of-visual-search-how-retailers-can-be-more-effective/

Use cases

Product details page (PDP)

Add to Cart



Check Out

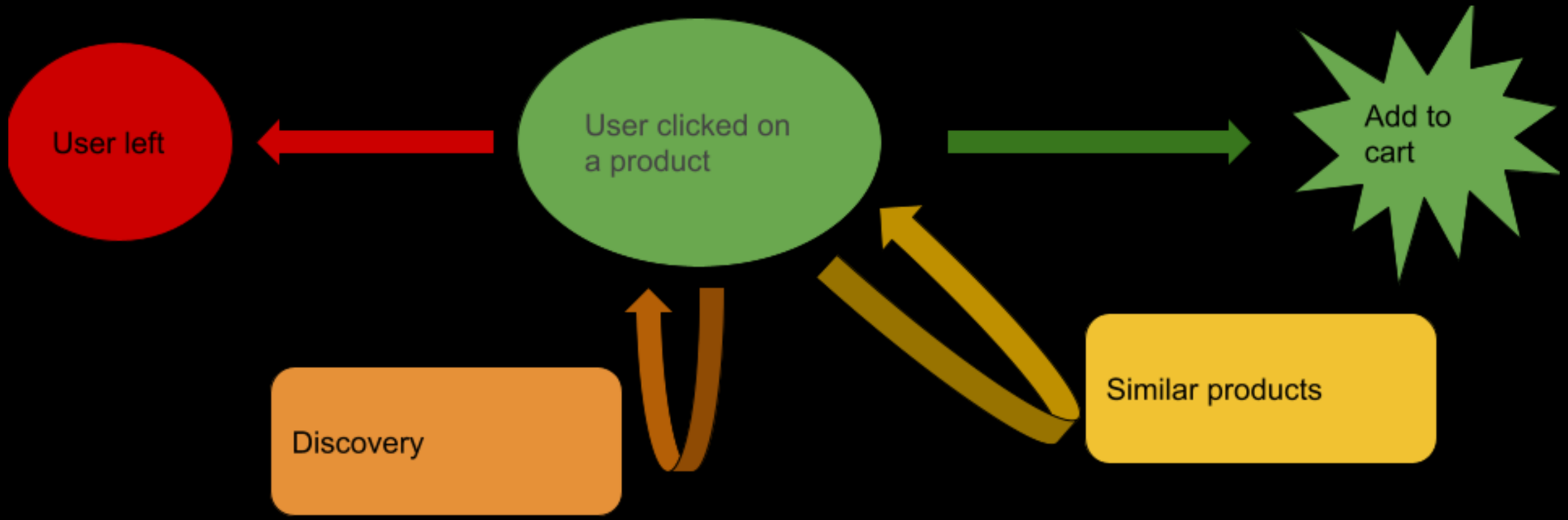
Post purchase

For you

User landed on a product details page

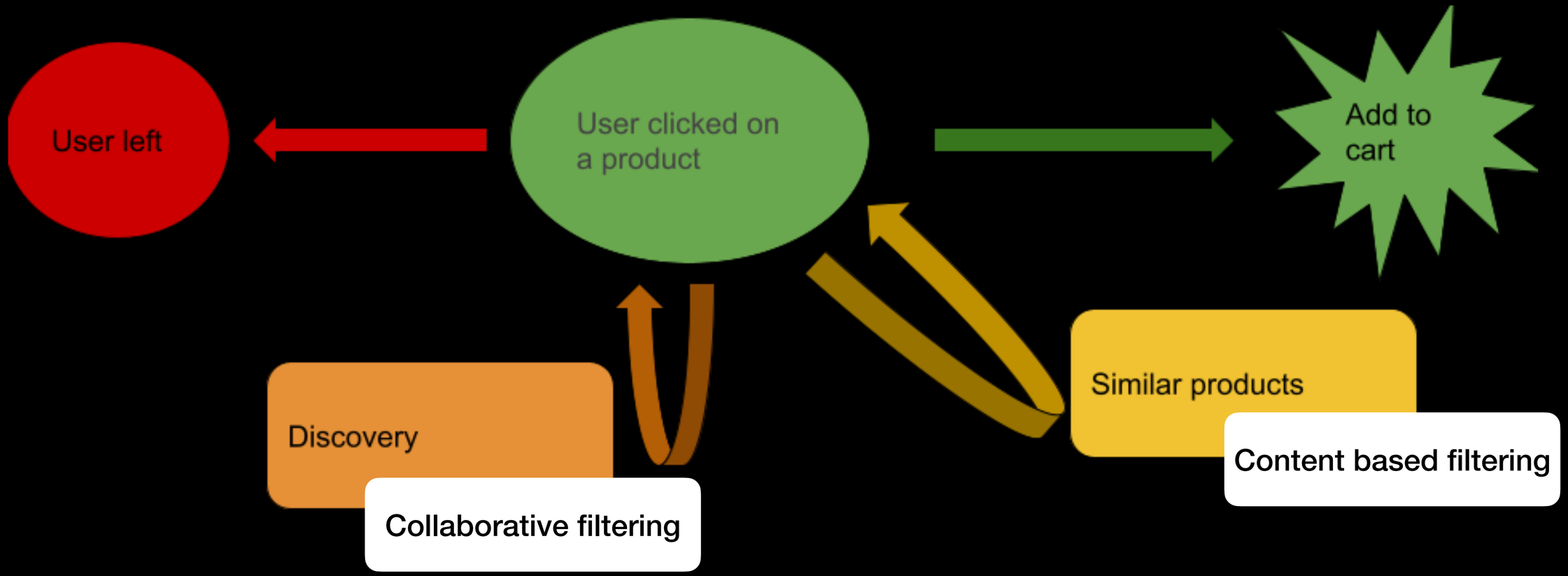
Product Detail Page

What do you want the user to do?



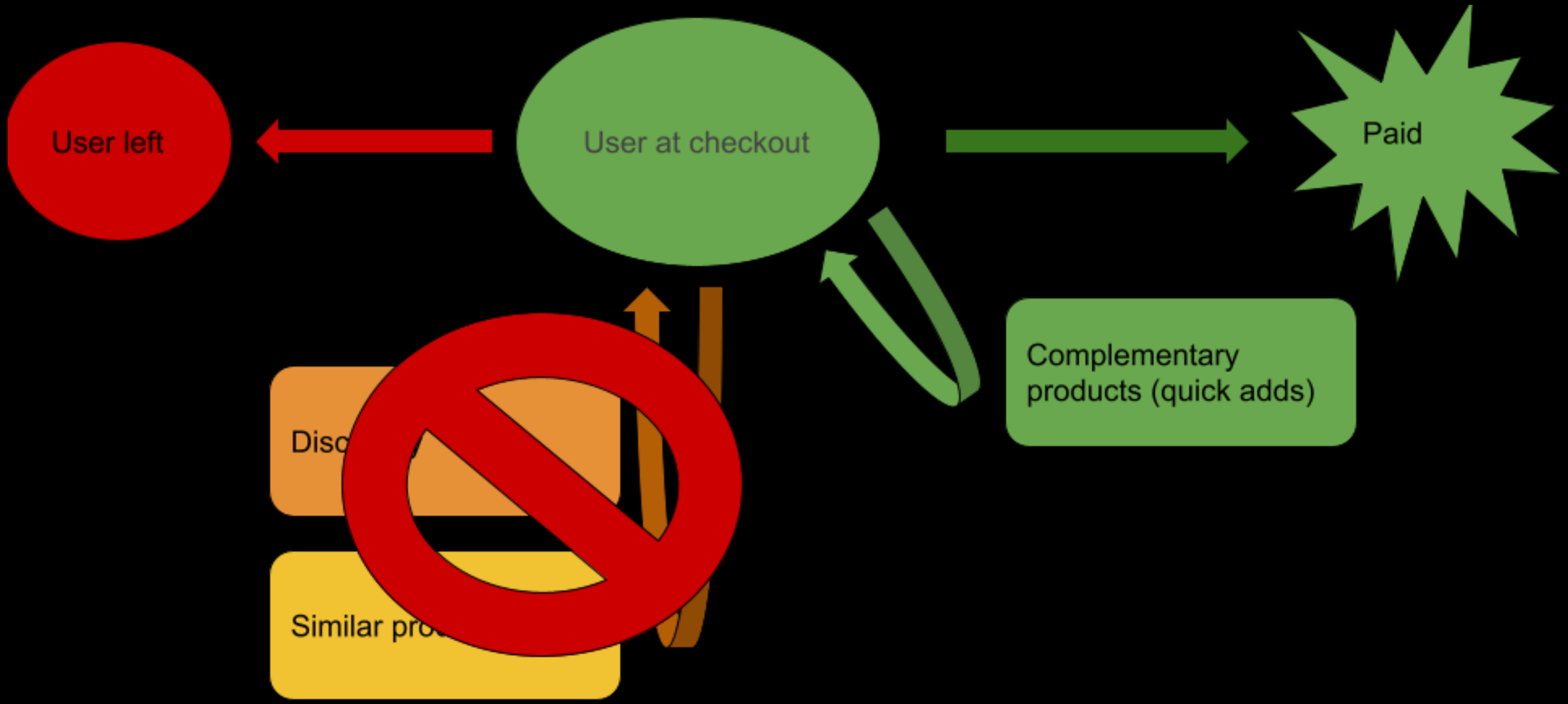
Product Detail Page

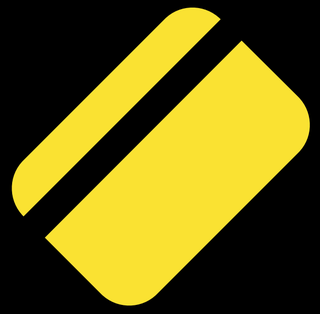
What do you want the user to do?



User added something to cart

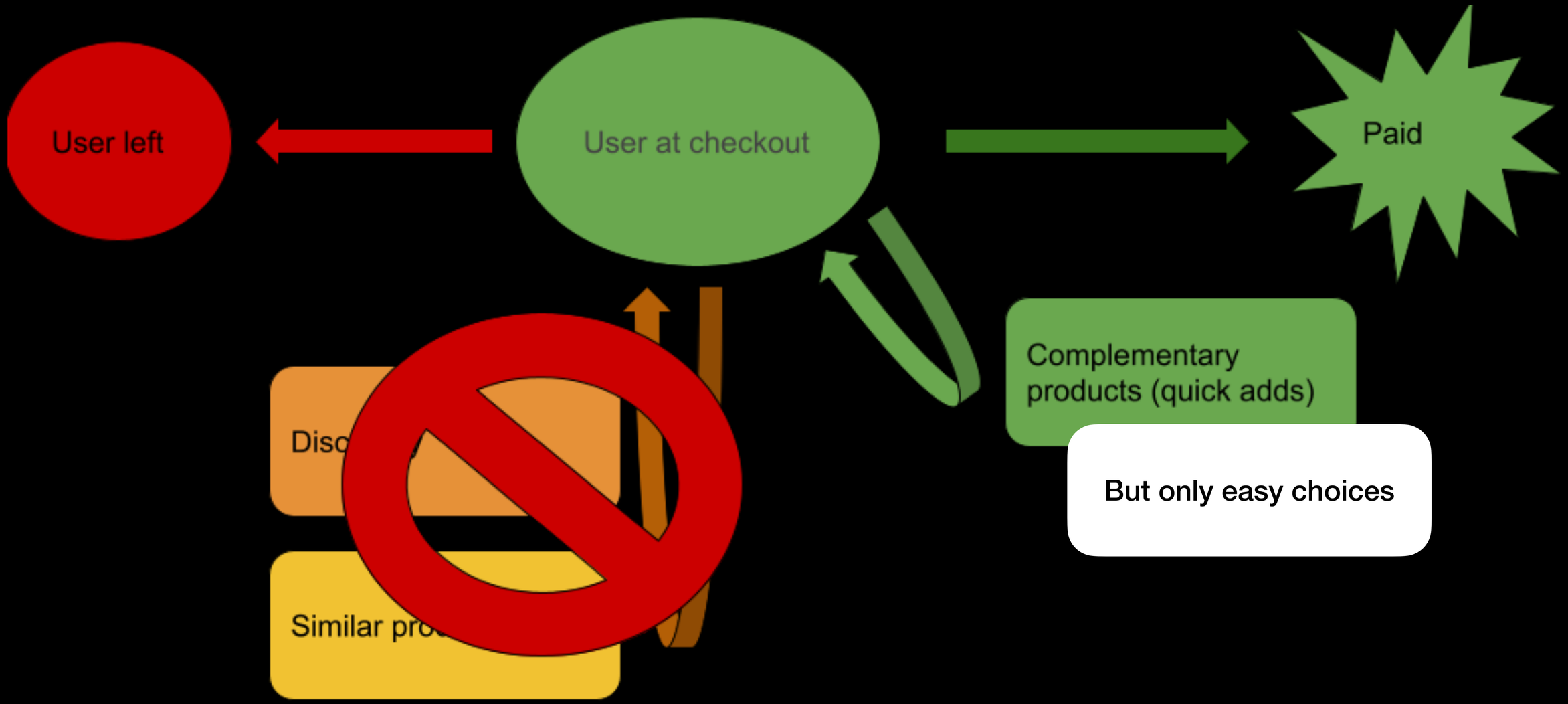
Add to cart recommendations



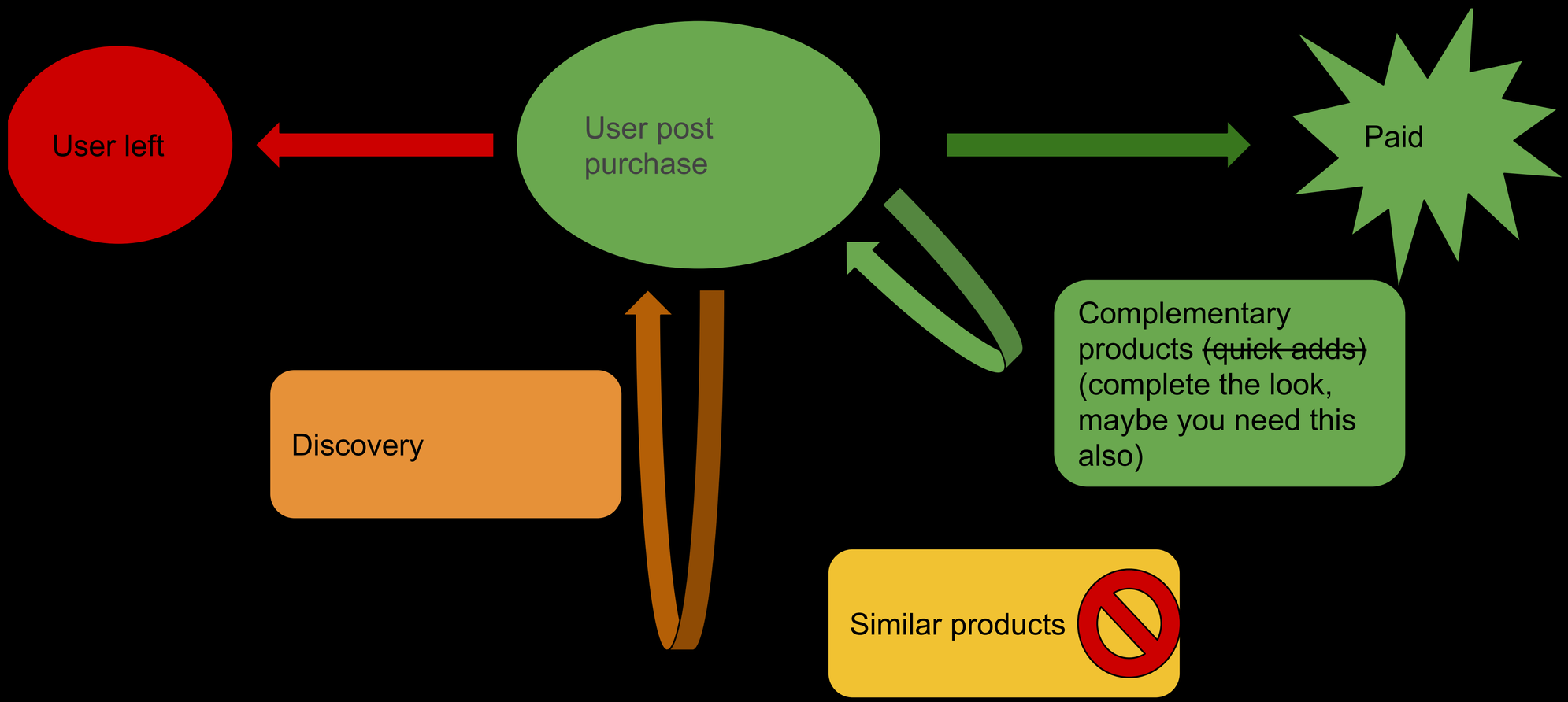


User is about to pay

Check out recommendations



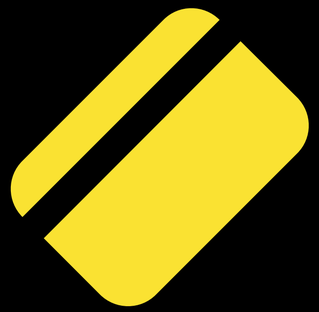
User bought something


User left
User post purchase
Paid
Complementary products ~~(quick adds)~~ (complete the look, maybe you need this also)
Discovery
Similar products

Use cases

Product details page (PDP)

Add to Cart

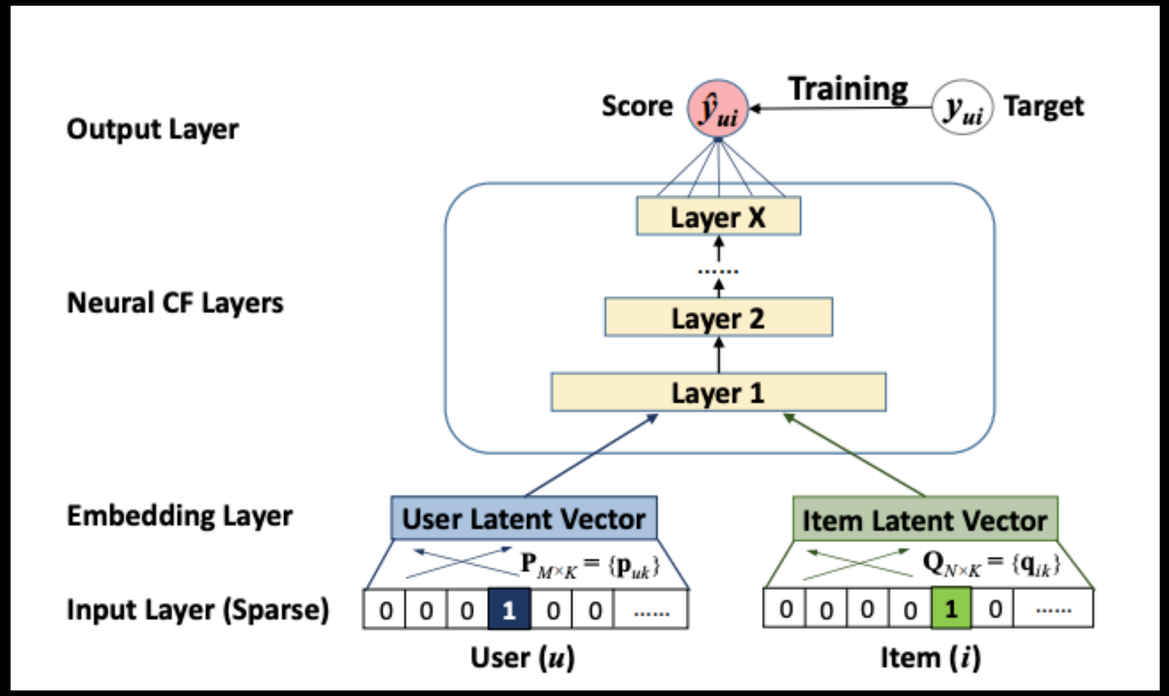


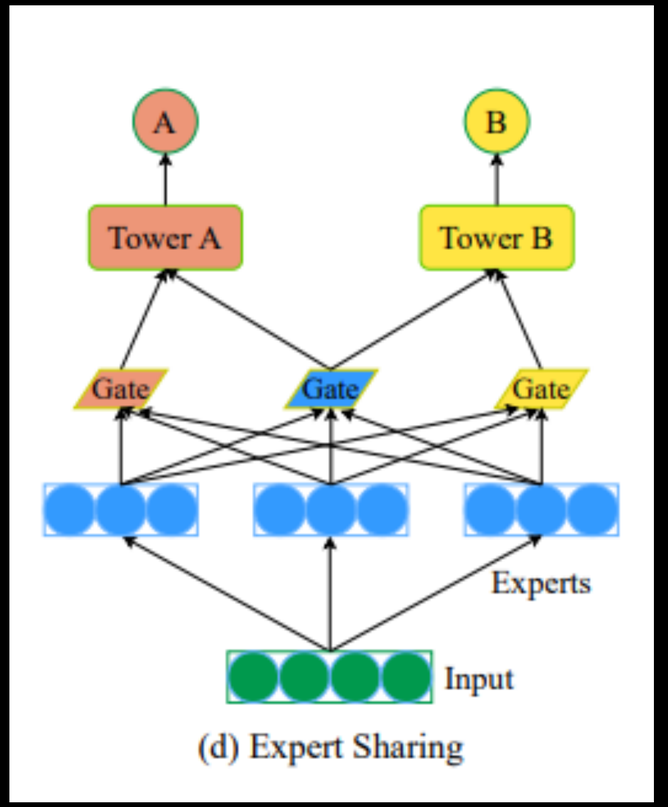
Check Out

Post purchase

For you

How to solve it?

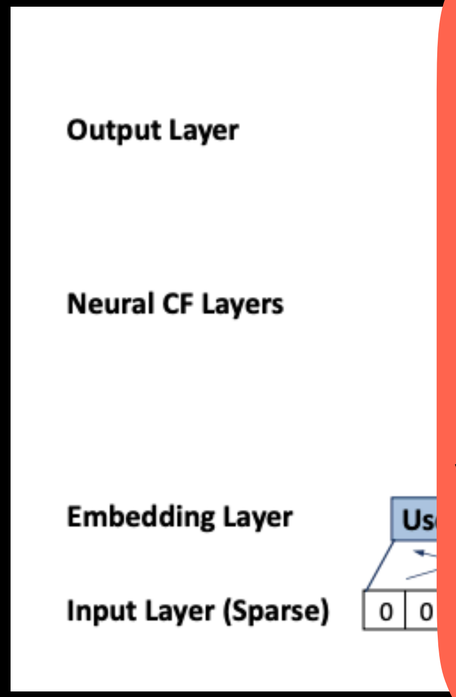




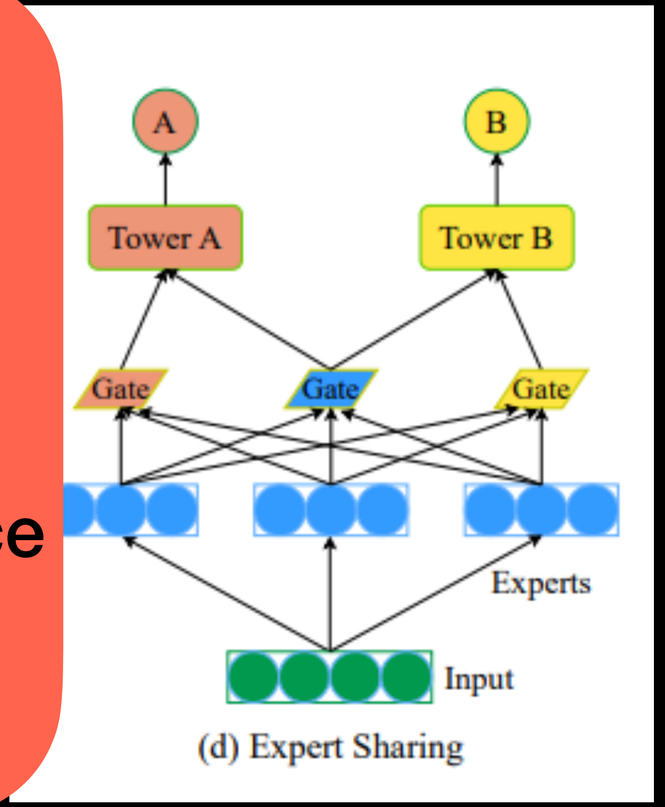
Multi-Task Deep Recommender Systems: A surveyhttps://arxiv.org/pdf/2302.03525.pdf

Neural Collaborative Filtering https://arxiv.org/pdf/1708.05031.pdf

How to solve it?



Start solving the problems individually, as simple as possible and then move on to more complex solutions when you understand the performance of the simpler ones



Multi-Task Deep Recommender Systems: A surveyhttps://arxiv.org/pdf/2302.03525.pdf

Neural Collaborative Filtering https://arxiv.org/pdf/1708.05031.pdf

How do we know we have solved it

Happy visitors ?

How do you measure that?

Time spend on site

Number of items consumed

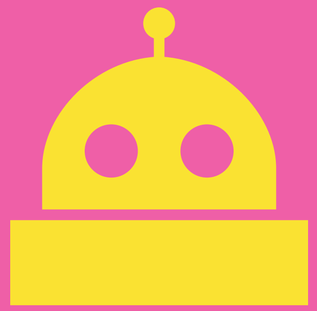
Increased business

How do you measure that?

Spend pr user

Items sold per user

We got a goal! lets start developing



AI is clever, but a good rule is that if you have a hard time explaining what it should do, then the AI will have a hard time doing it also

Personalisation and the Myth of Long Sessions

What is Personalization

In E-commerce - Customizing product features, pricing, or delivery options to cater to specific customer needs and preferences.

Often defined as:

Each user sees something different.

User 1:

CLASSIC BACKPACK 30L 599 KR.

STOP MICRO WASTE 219 KR.

TWILL BASEBALL CAP 249 KR.

User 2:

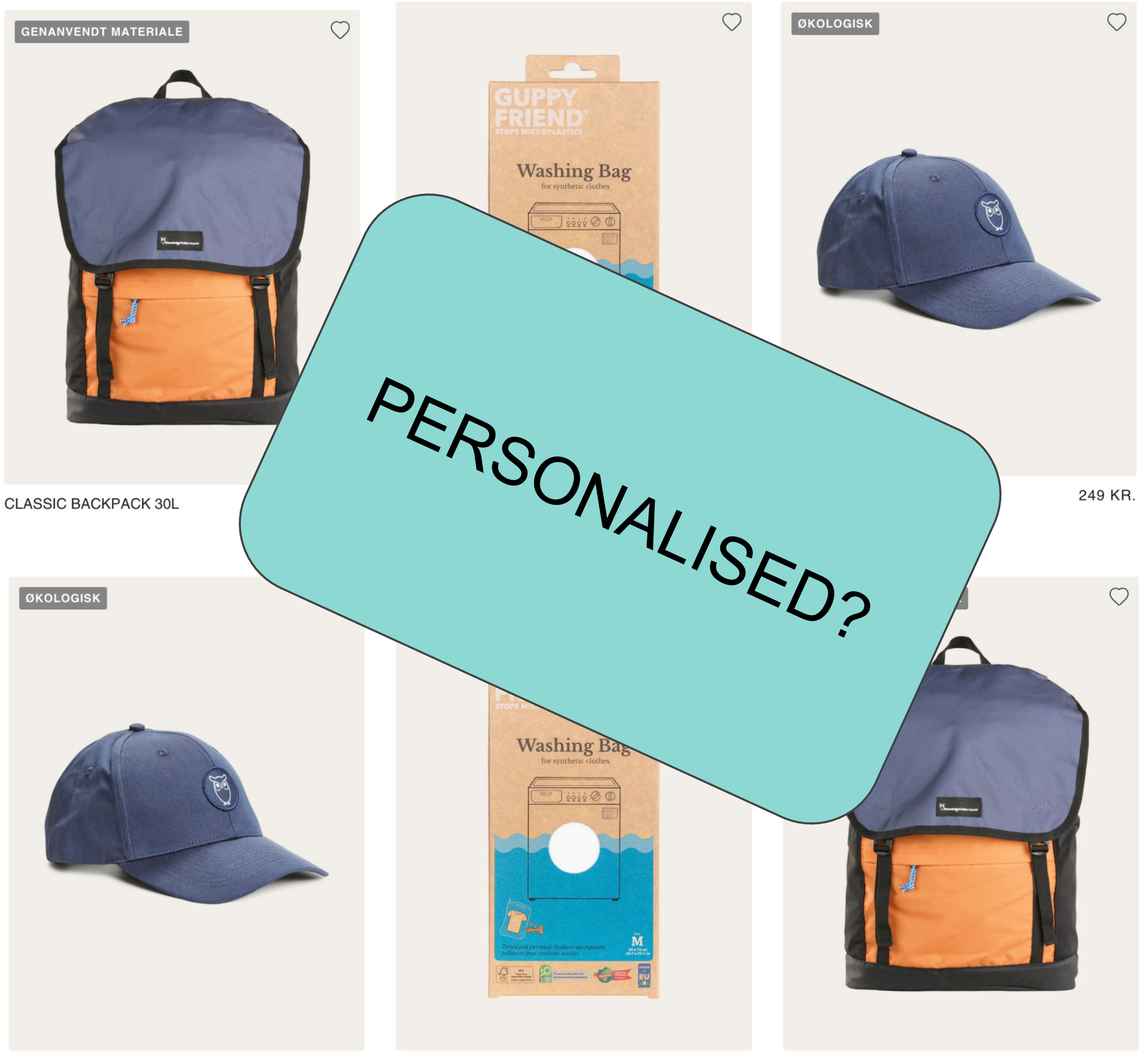
TWILL BASEBALL CAP 249 KR.

STOP MICRO WASTE 219 KR.

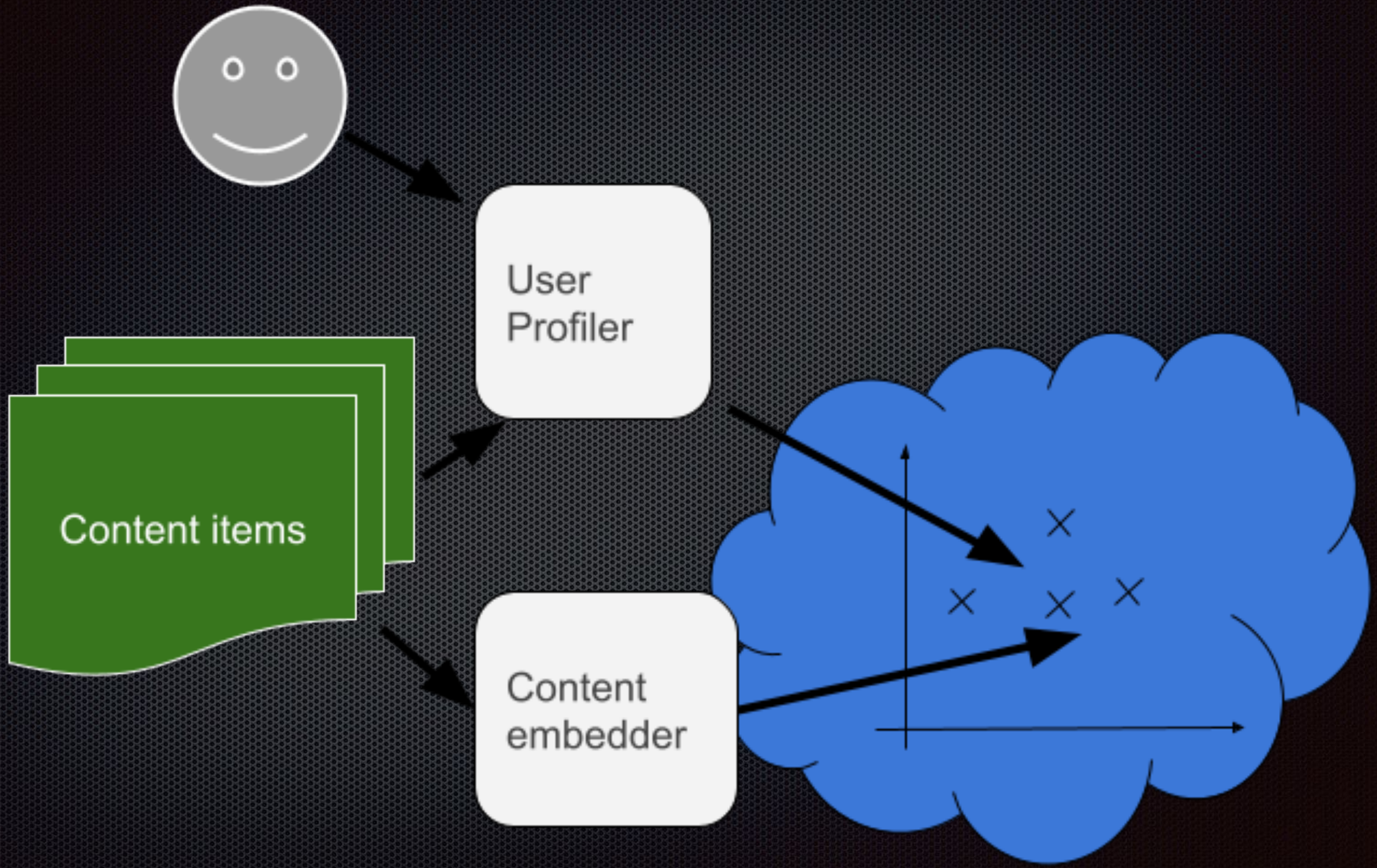
CLASSIC BACKPACK 30L 599 KR.

User 1:

User 2:


GENANVENDT MATERIALE
CLASSIC BACKPACK 30L
GUPPY FRIEND
STOPS MICROPLASTICS
Washing Bag
for synthetic clothes
ØKOLOGISK
249 KR.
PERSONALISED?
ØKOLOGISK
TWILL BASEBALL CAP
249 KR.
Washing Bag
for synthetic clothes
STOP MICRO WASTE
219 KR.
CLASSIC BACKPACK 30L
599 KR.

But what about the user?



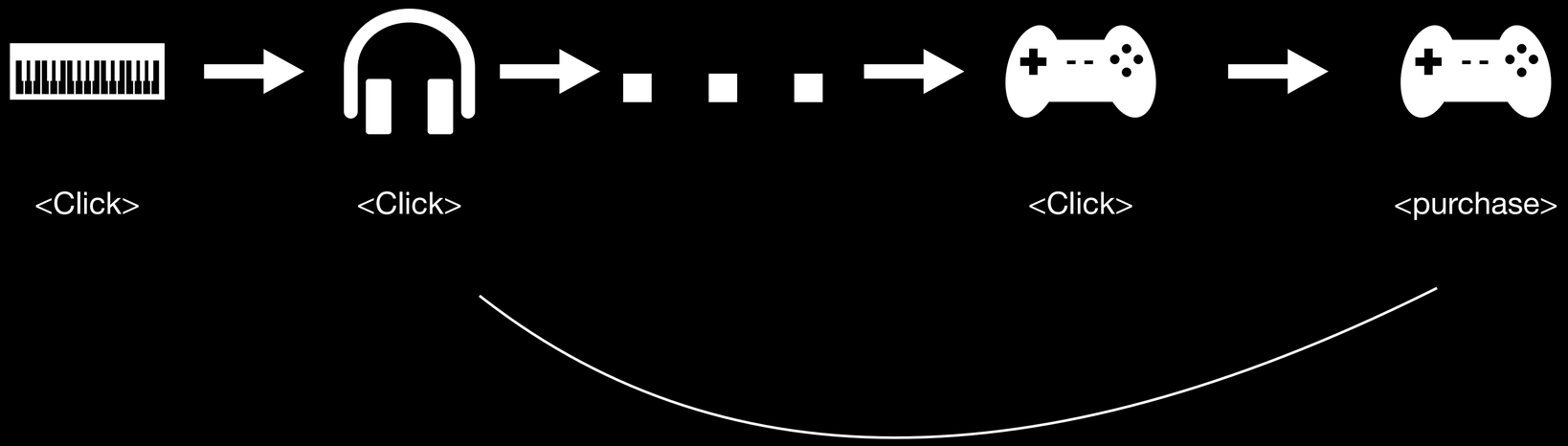
User behaviour modelling

A Survey on User Behavior Modeling in Recommender Systems https://arxiv.org/pdf/2302.11087.pdf

What is personalisation?

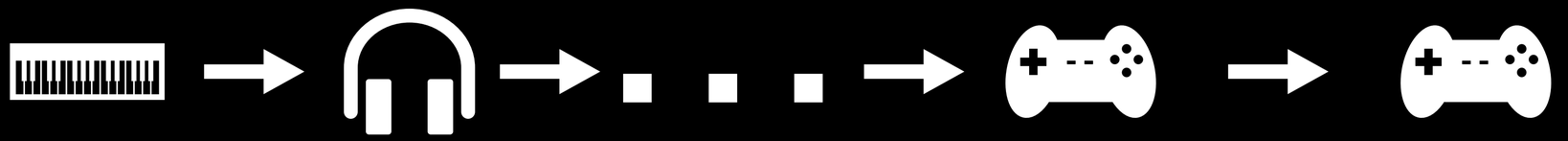
- Help user find the (right) content quicker
- Teach the user about products they need but didn’t know about

Help the user find the content quicker



Decrease time from landing page to purchase

Help the user find the content quicker

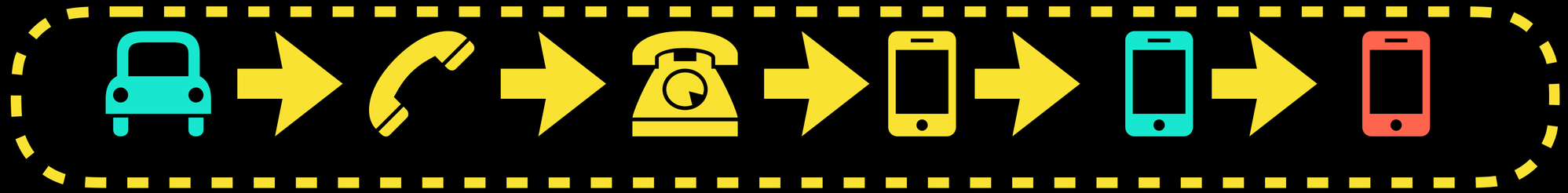


<Click> <Click> <Click> <purchase>

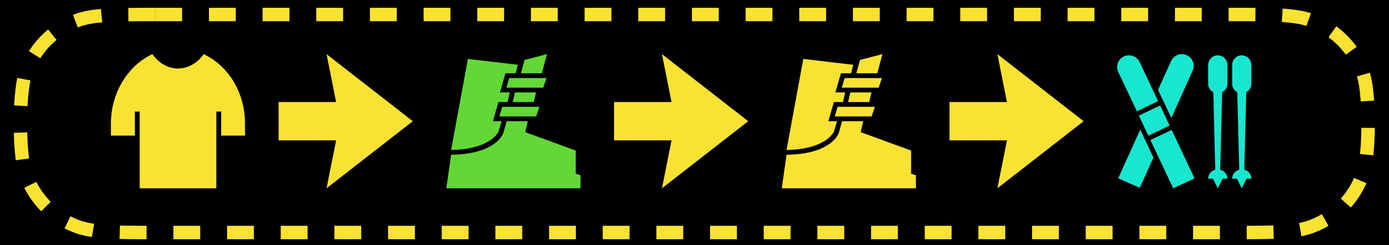
When can we personalise

Sessions

Session 1



Session 2



What do we want to learn?

Personal preferences

But what could they be in the domain you are looking in?

Color preferences

Style preferences

Gender preferences

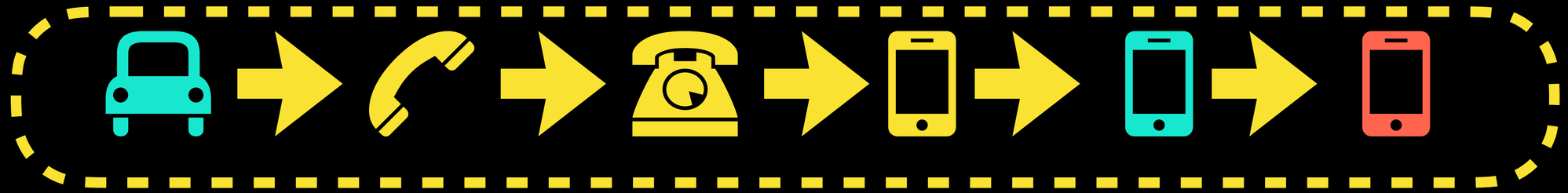
Historical data

But what could they be in the domain you are looking in?

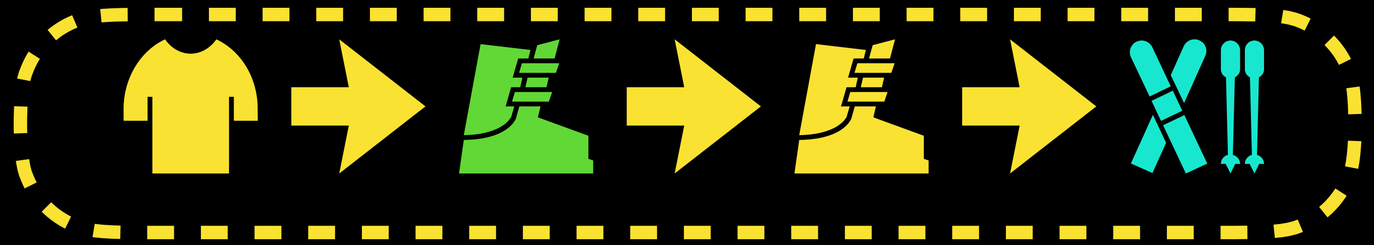
Don’t recommend what’s already consumed

What can I learn from session one that makes session two easier?

Session 1

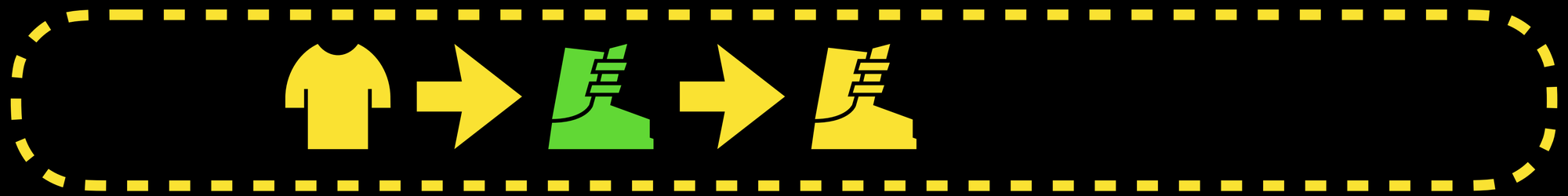


Session 2

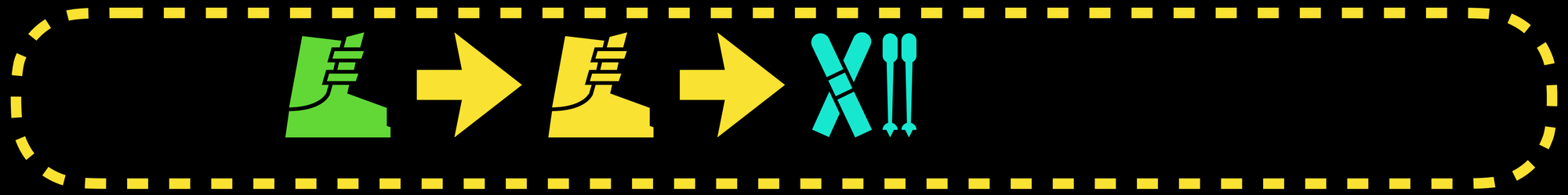


What can I learn from session one that makes session two easier?

Session 1

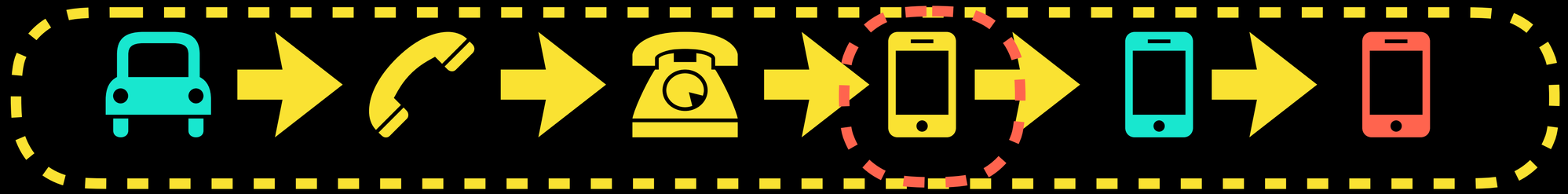


Session 2

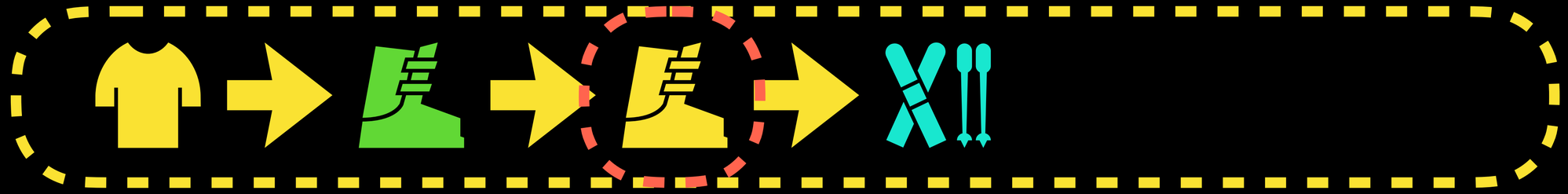


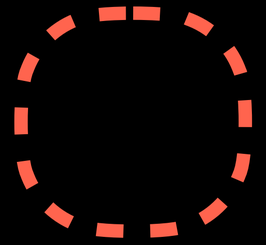
What can I learn from session one that makes session two easier?

Session 1



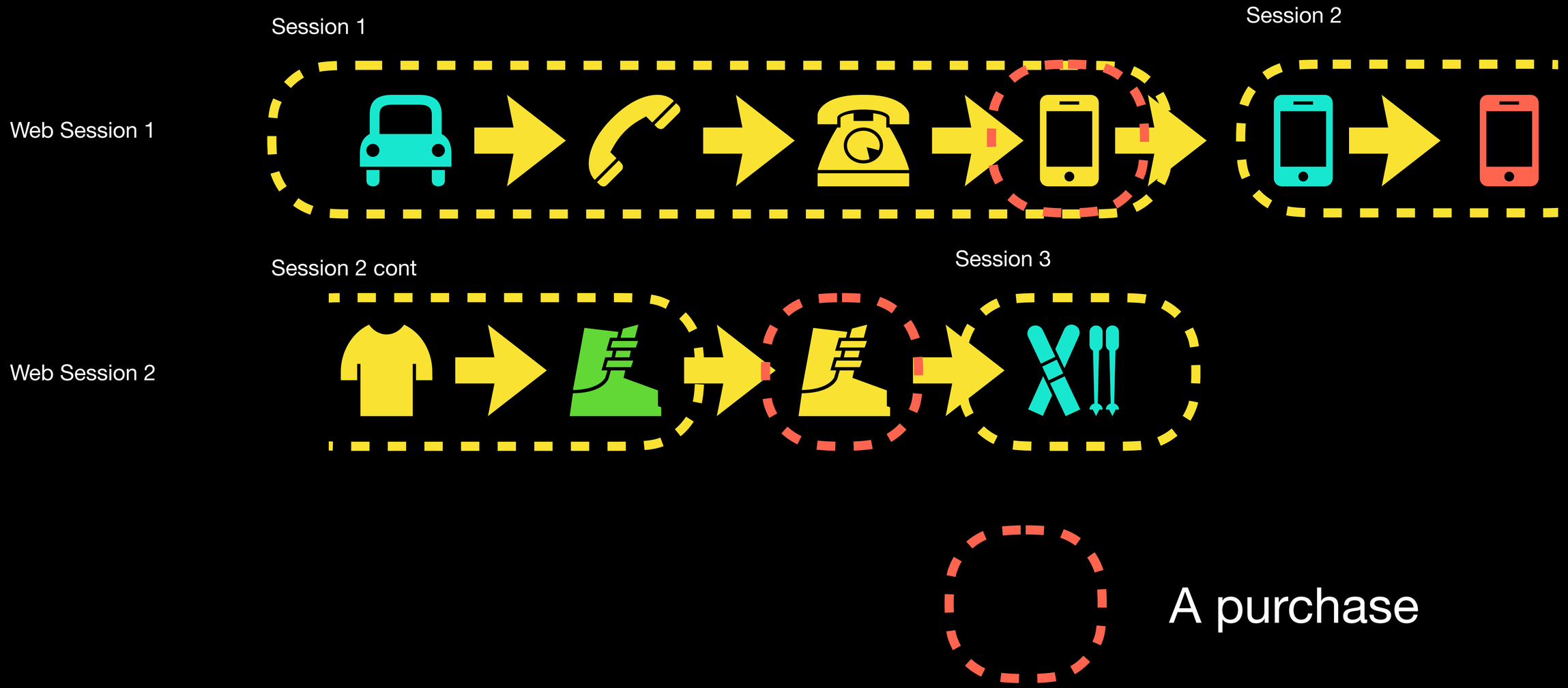
Session 2



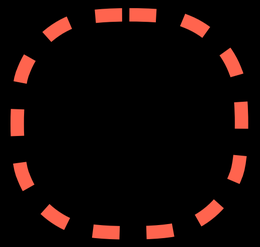


A purchase

What can I learn from session one that makes session two easier?



What can I learn from session one that makes session two easier?



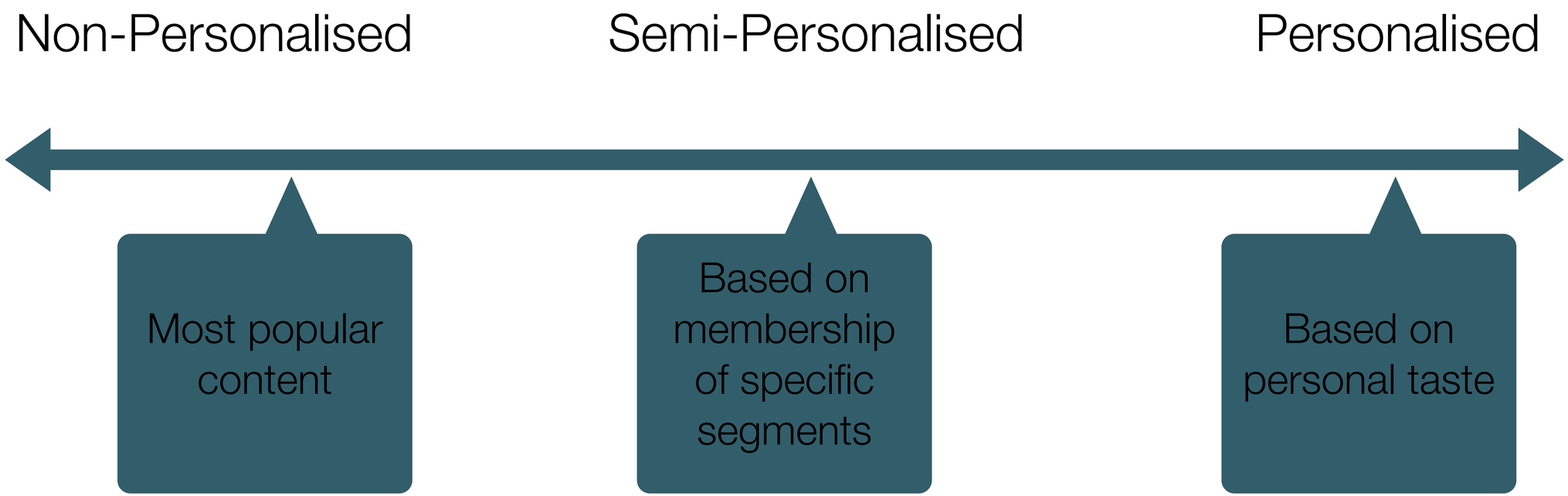
A purchase or a add to cart or checkout
finishes one logical session

What is the logical next thing to recommend after that?

Again!

No data means no Personalisation

Gracefully goes down in degrees of personalisation



Remember
Personalisation
is discrimination

Developing the Recommender

Often Software Engineers

Software Engineer Assumptions

- Its data so you don’t test using SE assumptions
- If the API responds as expected it works.

Data Science Assumptions:

- What’s testing again?
- Evaluation is beating SOTA

Building recommenders

Software Engineering

- Versioning of code
- Continuous integration

Data/ML engineering

- ML functional testing
- Experiment logging

Building recommenders

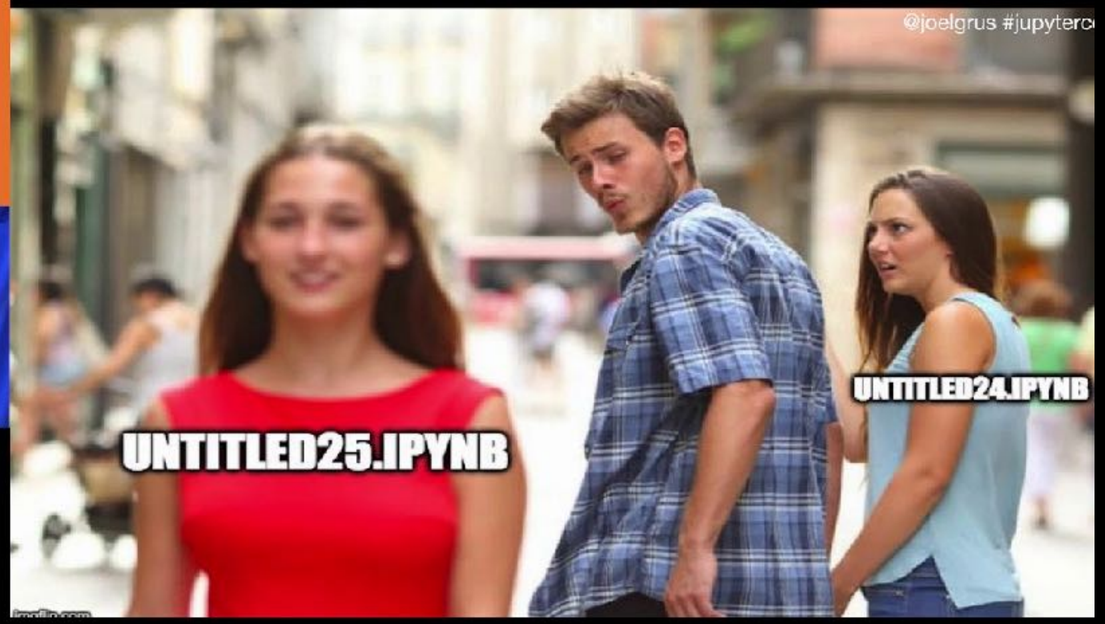
Software Engineering

- Versioning of code
- Continuous integration

Data/ML engineering

- ML functional testing
- Experiment logging

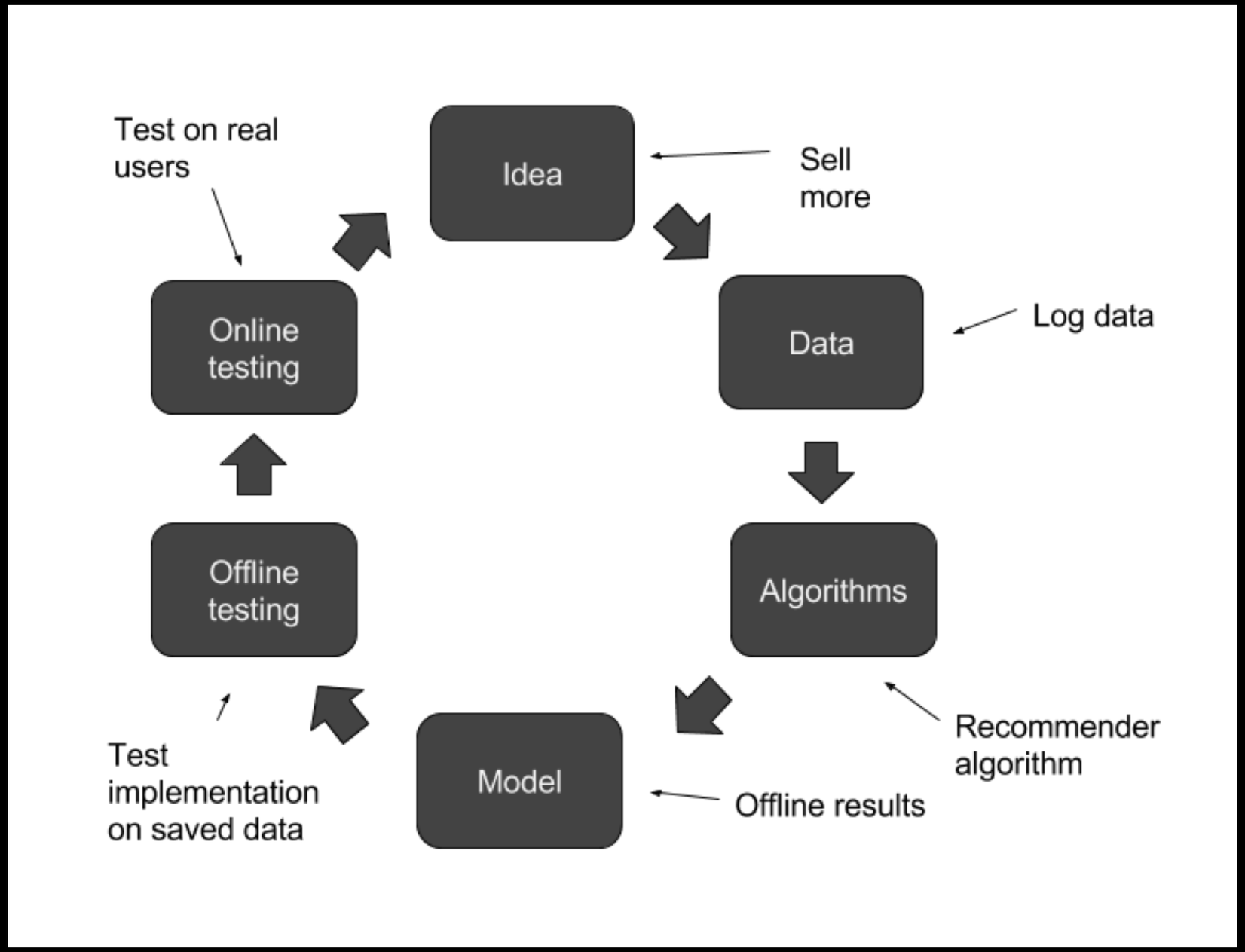
Expect to write testable code



Being scientific about it.

Start with the problem not the coolest article from #RecSys23

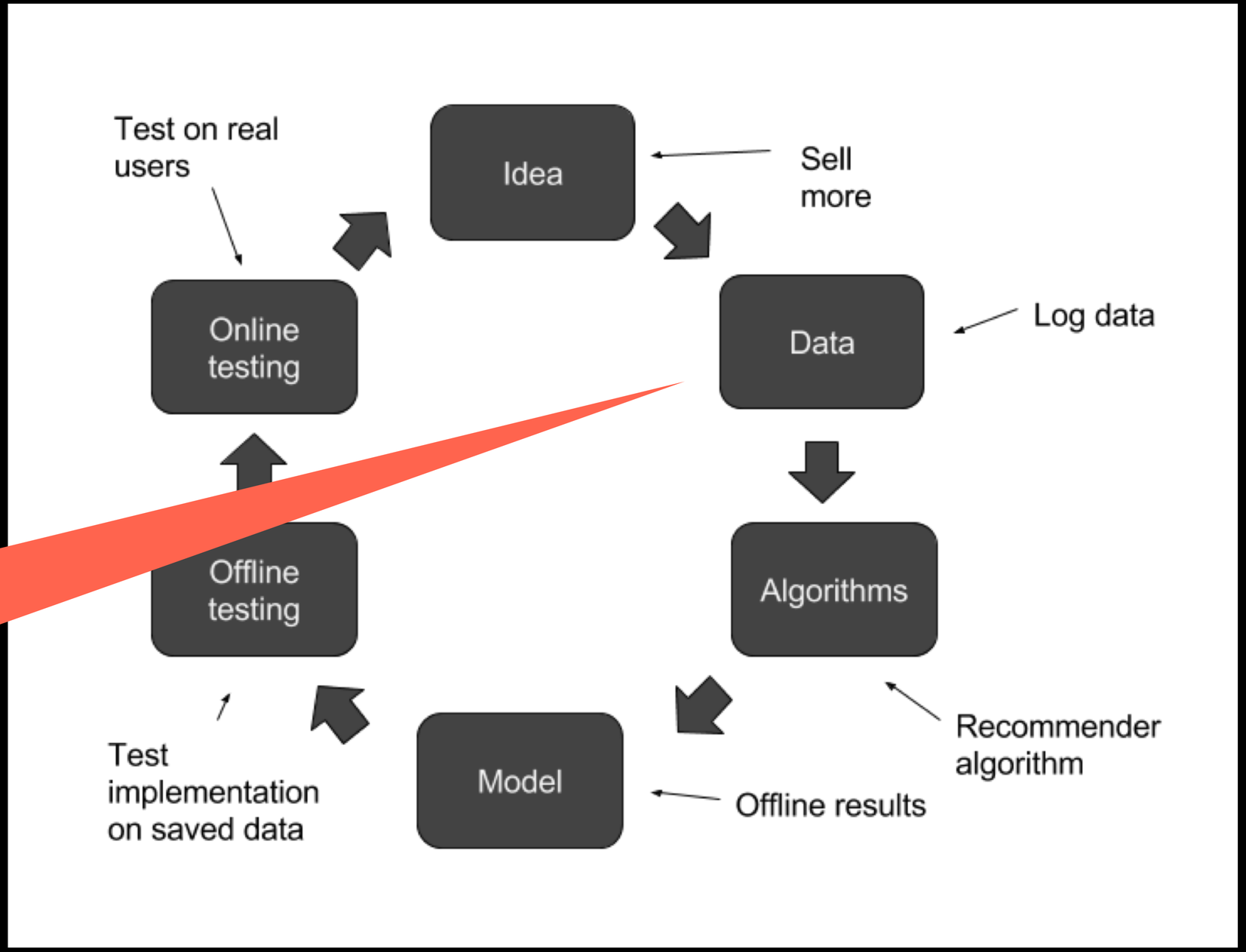
ML Engineering Cycle



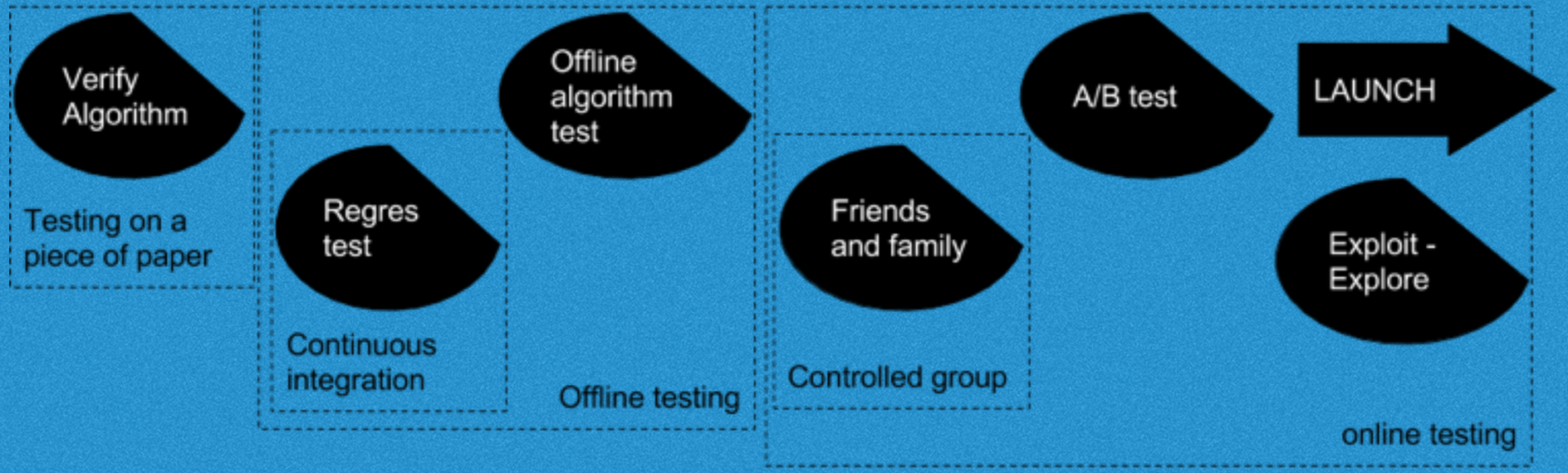
ML Engineering Cycle

Two silly questions, that is almost never asked:

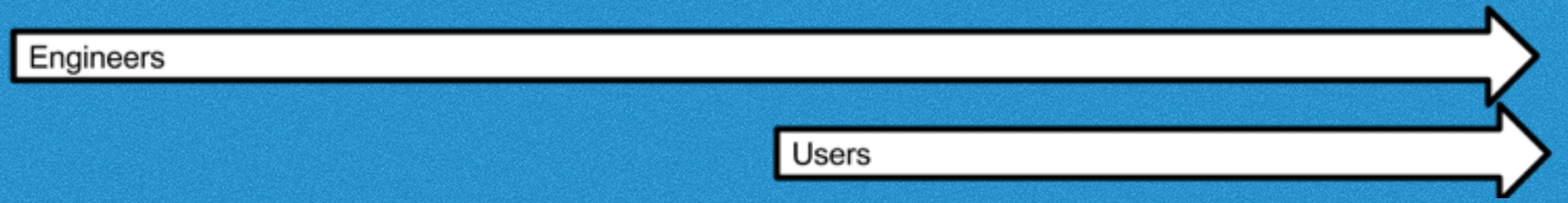
1) Is the data available at all?
2) Will it be accessible at runtime?



Recommender algorithm evaluation:



Involved:



Functional testing

- Should recommendation algorithms be tested?
- What aspects other recommendation algorithm would benefit most from testing?

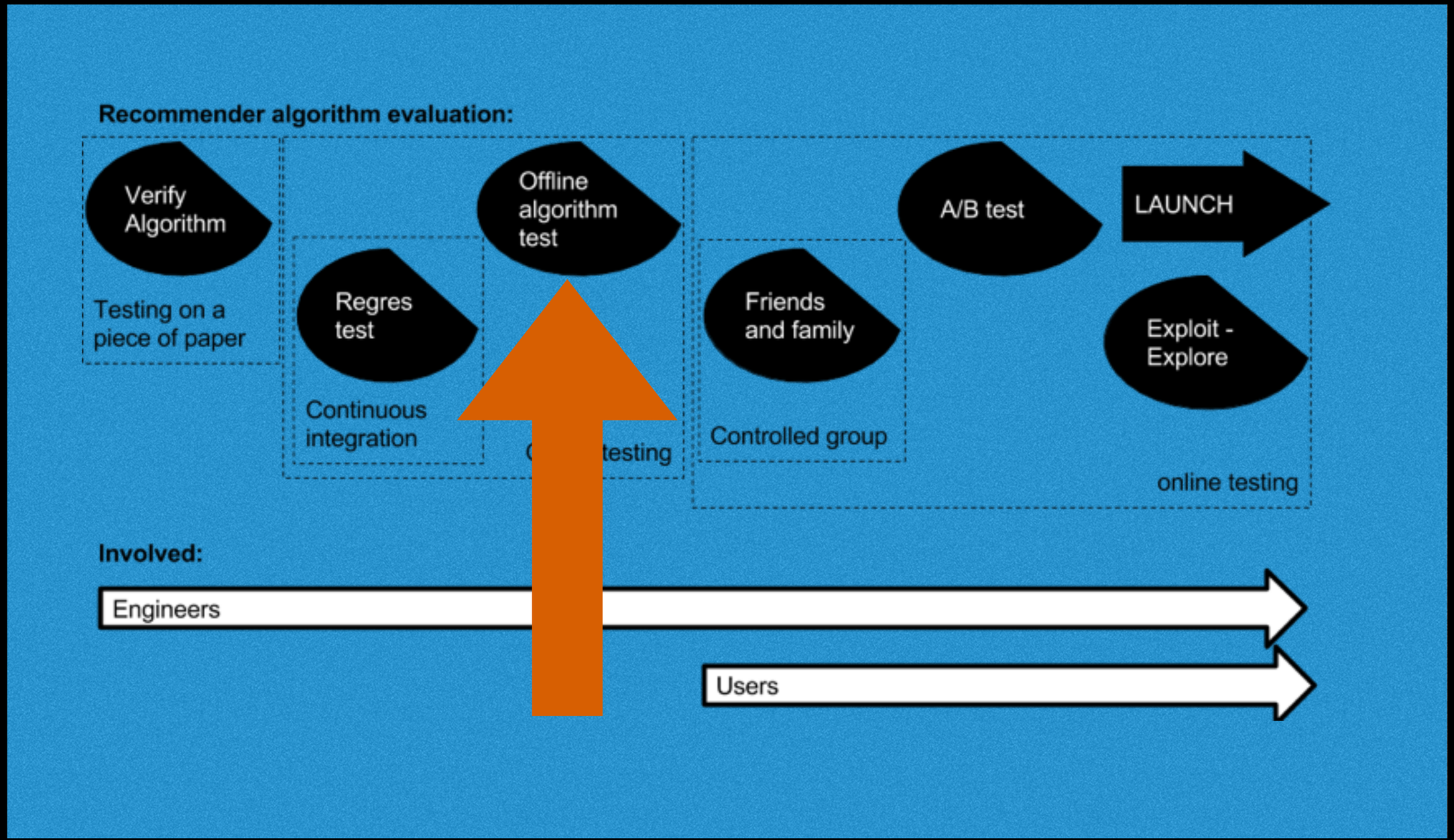
Abstract: Should Algorithm Evaluation Extend to Testing? We Think So.

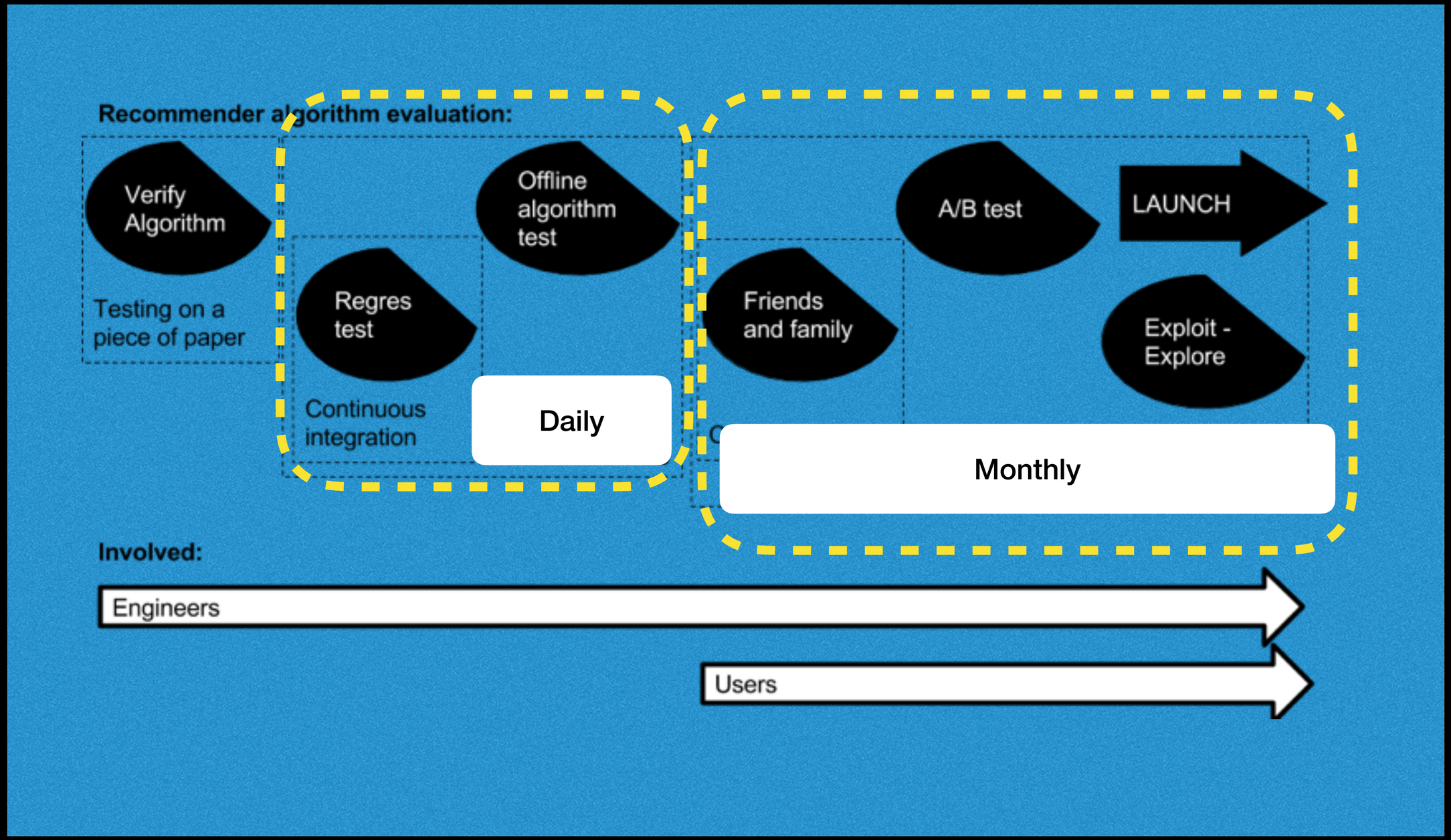
Lien Michiels[1,2,†], Robin Verachtert[1,2,†], Kim Falk[3] and Bart Goethals[1,2,4]

[1]*Froomle N.V., Belgium*
[2]*University of Antwerp, Antwerp, Belgium*
[3]*Shopify, Canada*
[4]*Monash University, Melbourne, Australia*

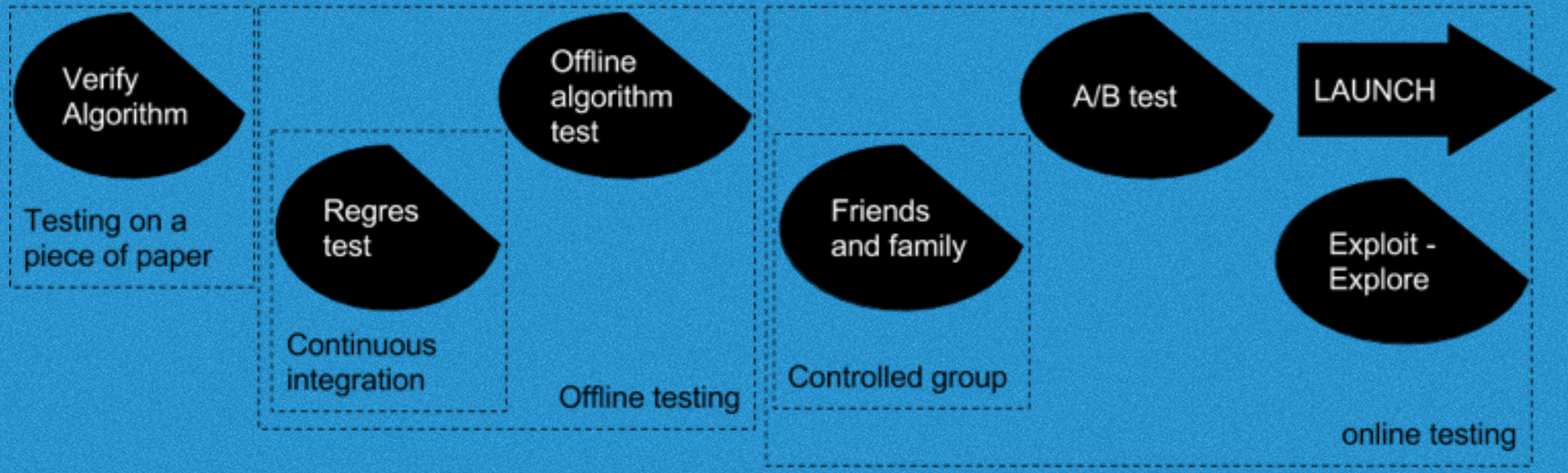
Abstract

Software engineers test virtually all of their code through unit, regression and integration te
contrast, data scientists and machine learning engineers often evaluate models based solely o
training or evaluation loss and task performance metrics such as accuracy, precision or recall. Whe
becomes 'algorithms', software best practices are often neglected. In our research, we found tha
publicly available algorithm implementations indeed are not tested beyond ranking performance
such as recall and normalized discounted cumulative gain. Applying software testing best pr
to algorithms can seem daunting (and unnecessary). However, software packages like scikit-lea
SpaCy have demonstrated that it definitely is possible to test (at least some aspects of) algorith
believe that algorithms should be tested. Without tests, you may just end up with dead code
gradients that do not update, or logical errors you failed to detect. The question then become
should we test algorithms? During the workshop, we would like to open up this discussion. We sta
an overview of software testing paradigms: from black-box to white-box testing, unit to regression
and more. We then present some examples of testing patterns we have applied to our recommen
algorithm implementations. At the end of the discussion, we hope to have answered some of the fo
questions:

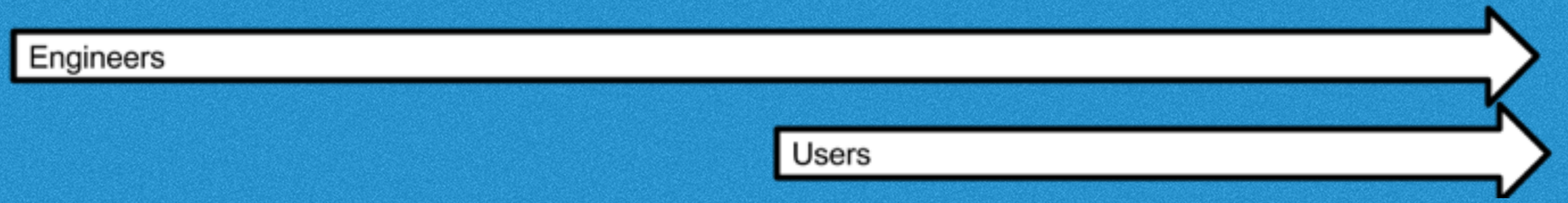

Recommender algorithm evaluation:
Verify Algorithm
Testing on a piece of paper
Regres test
Continuous integration
Offline algorithm test
testing
Friends and family
Controlled group
A/B test
LAUNCH
Exploit - Explore
online testing
Involved:
Engineers
Users


Recommender algorithm evaluation:
Verify Algorithm
Testing on a piece of paper
Regres test
Offline algorithm test
Continuous integration
Daily
Friends and family
A/B test
LAUNCH
Exploit - Explore
Monthly
Involved:
Engineers
Users

Recommender algorithm evaluation:

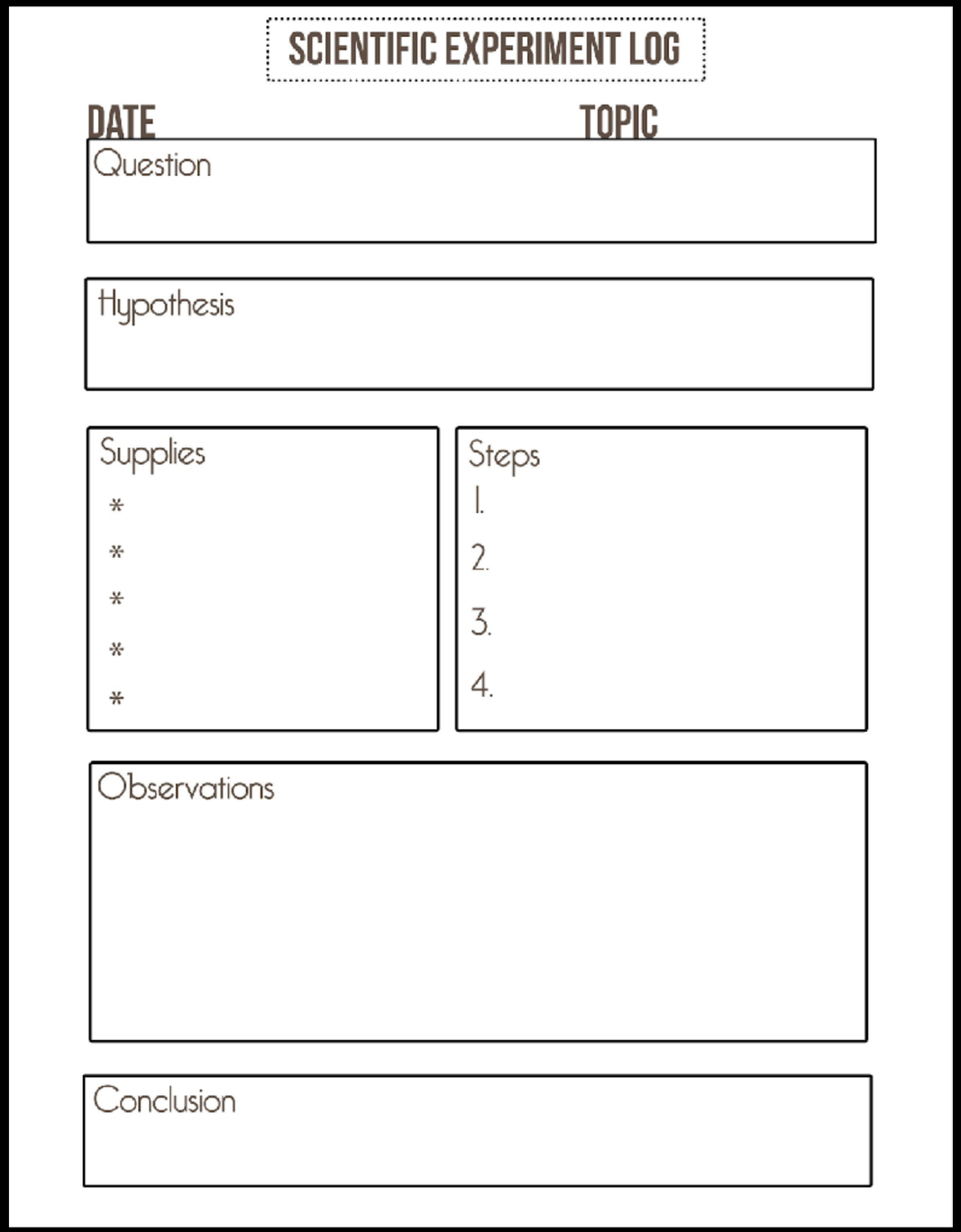


Involved:



Training a model

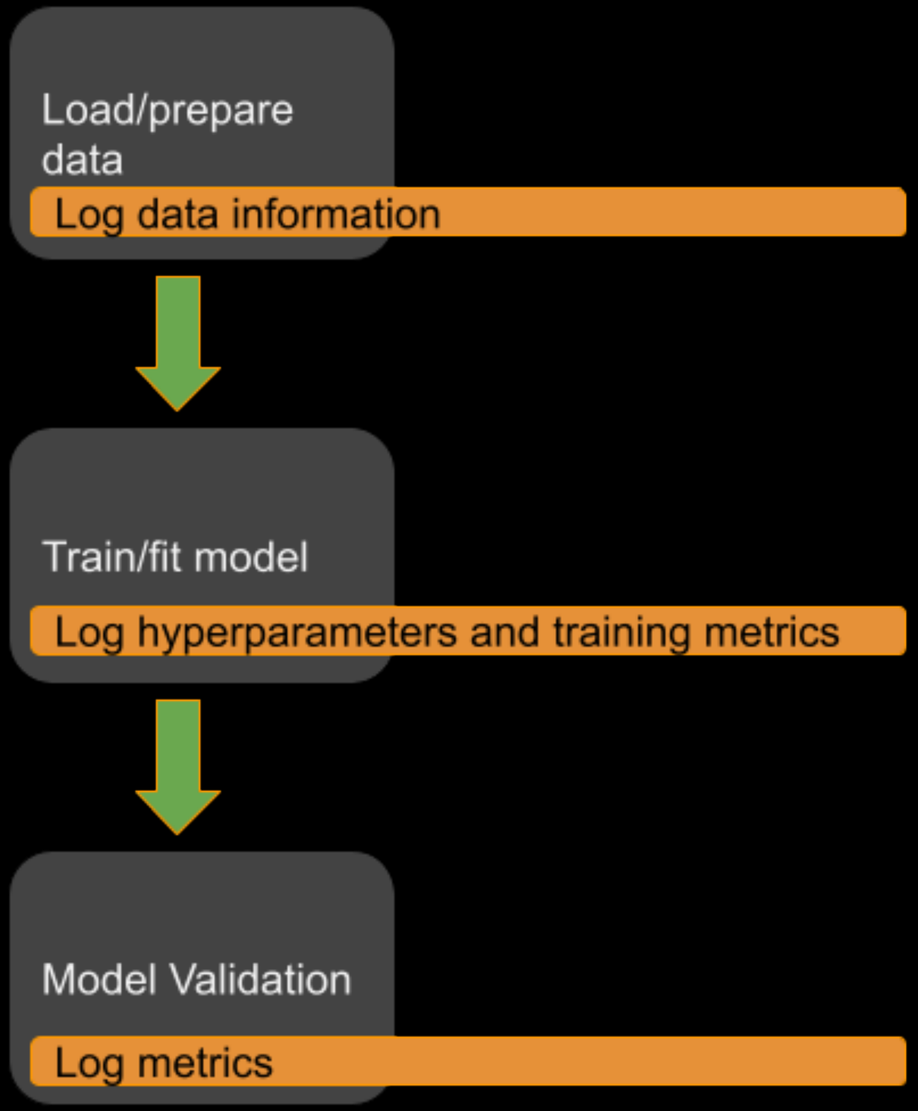
- Should be considered as a scientific experiment.
- Everything needed to rerun it should be saved.



	Date	User	Source	Version	Parameters		Metrics		
					alpha	l1_ratio	mae	r2	rmse
☐	2018-06-04 23:00:10	mlflow	train.py	05e956	1	1	0.649	0.04	0.862
☐	2018-06-04 23:00:10	mlflow	train.py	05e956	1	0.5	0.648	0.046	0.859
☐	2018-06-04 23:00:10	mlflow	train.py	05e956	1	0.2	0.628	0.125	0.823
☐	2018-06-04 23:00:09	mlflow	train.py	05e956	1	0	0.619	0.176	0.799
☐	2018-06-04 23:00:09	mlflow	train.py	05e956	0.5	1	0.648	0.046	0.859
☐	2018-06-04 23:00:09	mlflow	train.py	05e956	0.5	0.5	0.628	0.127	0.822
☐	2018-06-04 23:00:09	mlflow	train.py	05e956	0.5	0.2	0.621	0.171	0.801
☐	2018-06-04 23:00:09	mlflow	train.py	05e956	0.5	0	0.615	0.199	0.787
☐	2018-06-04 23:00:09	mlflow	train.py	05e956	0	1	0.578	0.288	0.742
☐	2018-06-04 23:00:09	mlflow	train.py	05e956	0	0.5	0.578	0.288	0.742
☐	2018-06-04 23:00:09	mlflow	train.py	05e956	0	0.2	0.578	0.288	0.742
☐	2018-06-04 23:00:08	mlflow	train.py	05e956	0	0	0.578	0.288	0.742

Evaluation Framework

- Each step should be logged.
- Create a experimentation log.

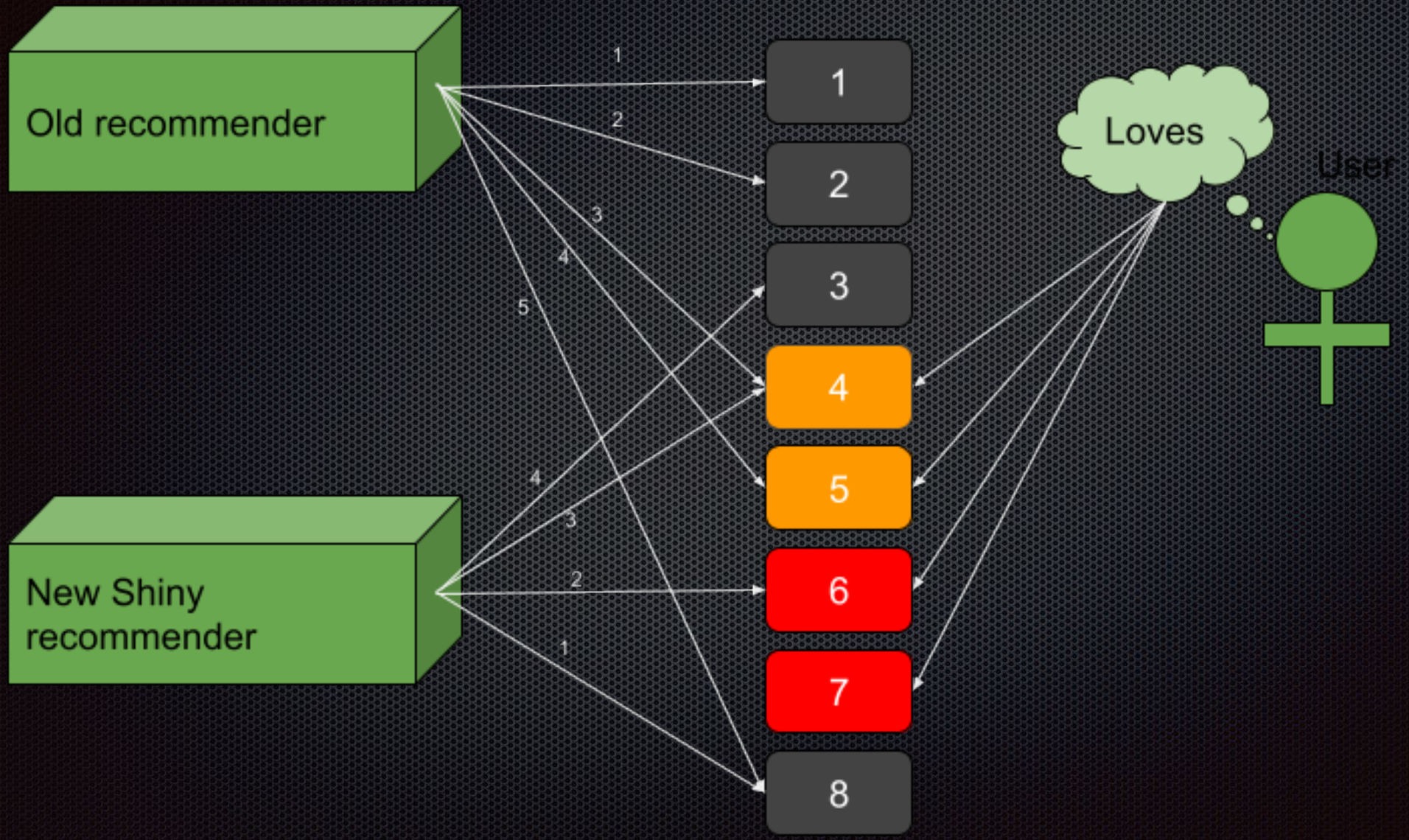


Remember to show Progress.
The managers doesn’t understand what you are doing

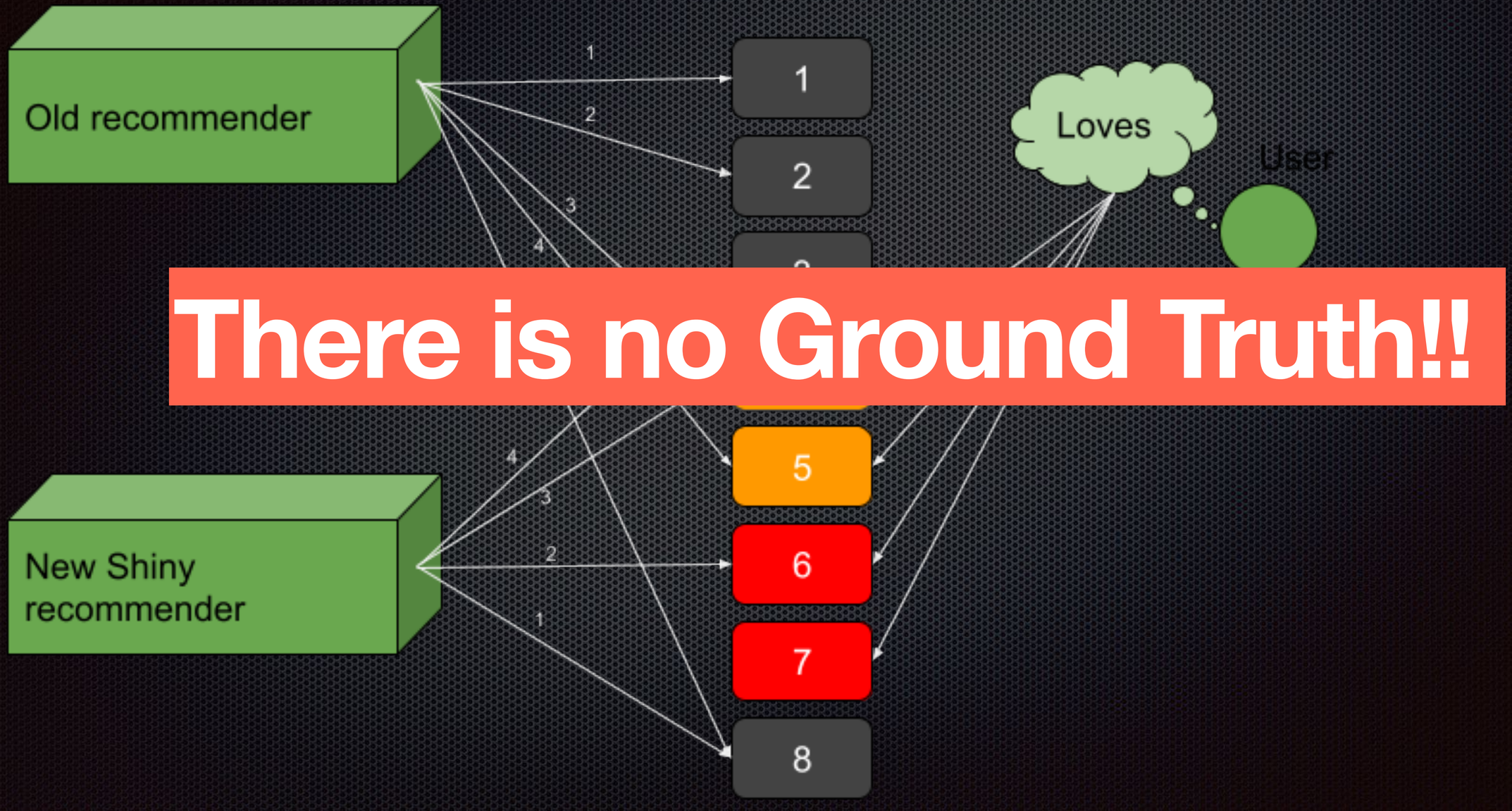
Constant dataset while developing

- When you are developing your model, keep as many things static as possible.
- Only update the dataset at intervals, when a dataset is updated your experiment log is basically invalid.
- However, when getting closer to production the dataset should be kept up to date.

Historical data



Historical data



Understand the data to create good recommendations

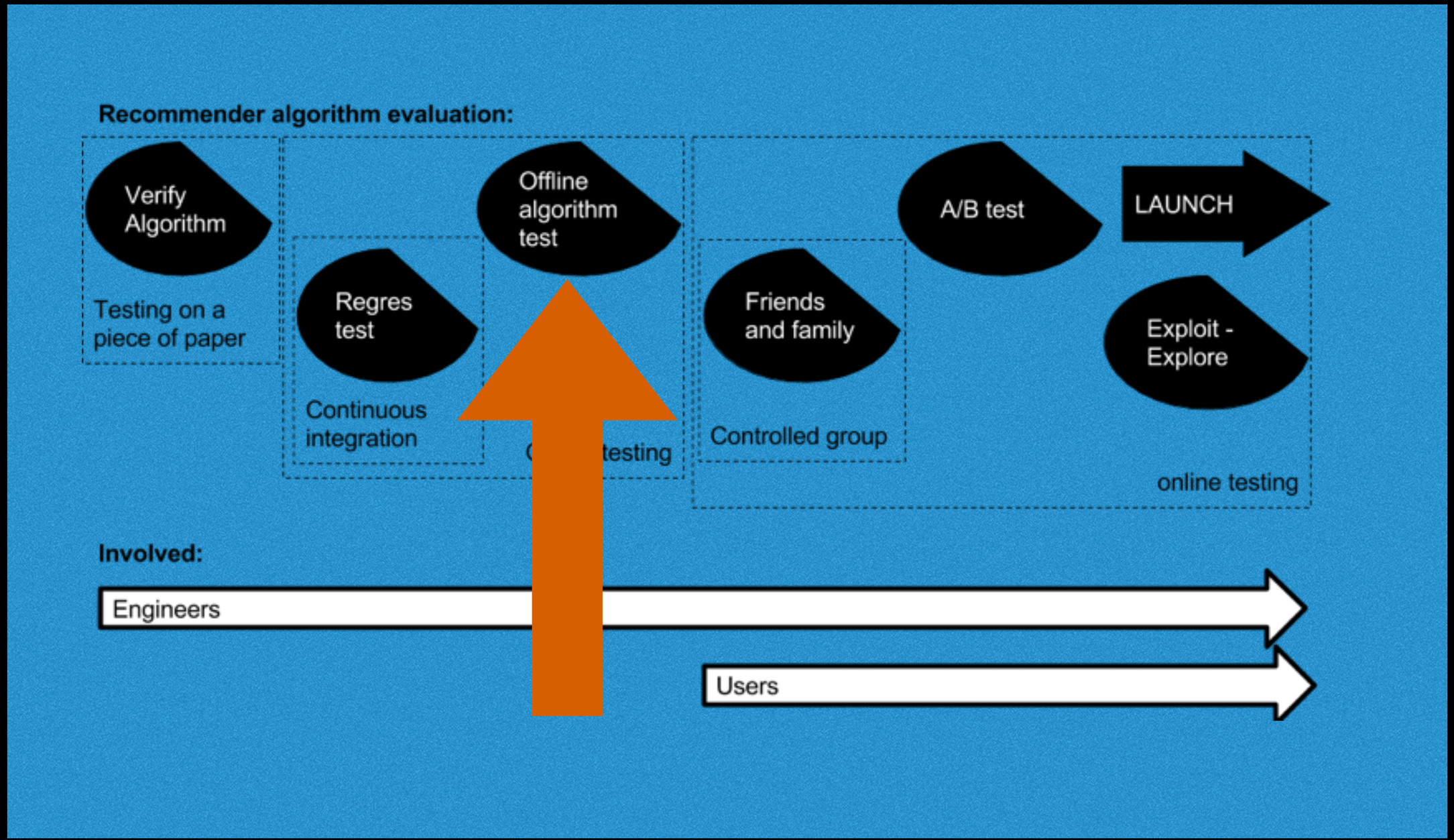
Training data Test data

Training data

Test data

=

Production data


Recommender algorithm evaluation:
Verify Algorithm
Testing on a piece of paper
Regres test
Continuous integration
Offline algorithm test
testing
Friends and family
Controlled group
A/B test
LAUNCH
Exploit - Explore
online testing
Involved:
Engineers
Users

Beyond accuracy – what metrics are interesting

Evaluation is not the goal.

“If you cannot measure it,
you cannot improve it.”
Lord Kelvin

Where do we evaluate

And why

Business

Improve performance
and minimise risk.

Academia

Show case
performance of new
research

Where do we evaluate

And why

Business

Improve performance and minimise risk.

Offline evaluation

Decision point of whether the new model should be A/B

Online offline evaluation

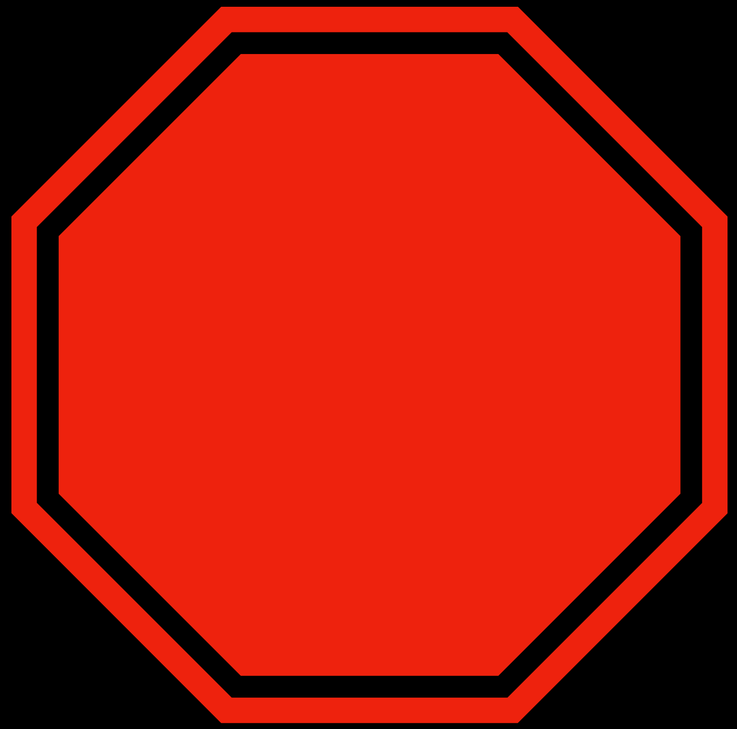
Monitoring of the performance the model.

Academia

Show case performance of new research

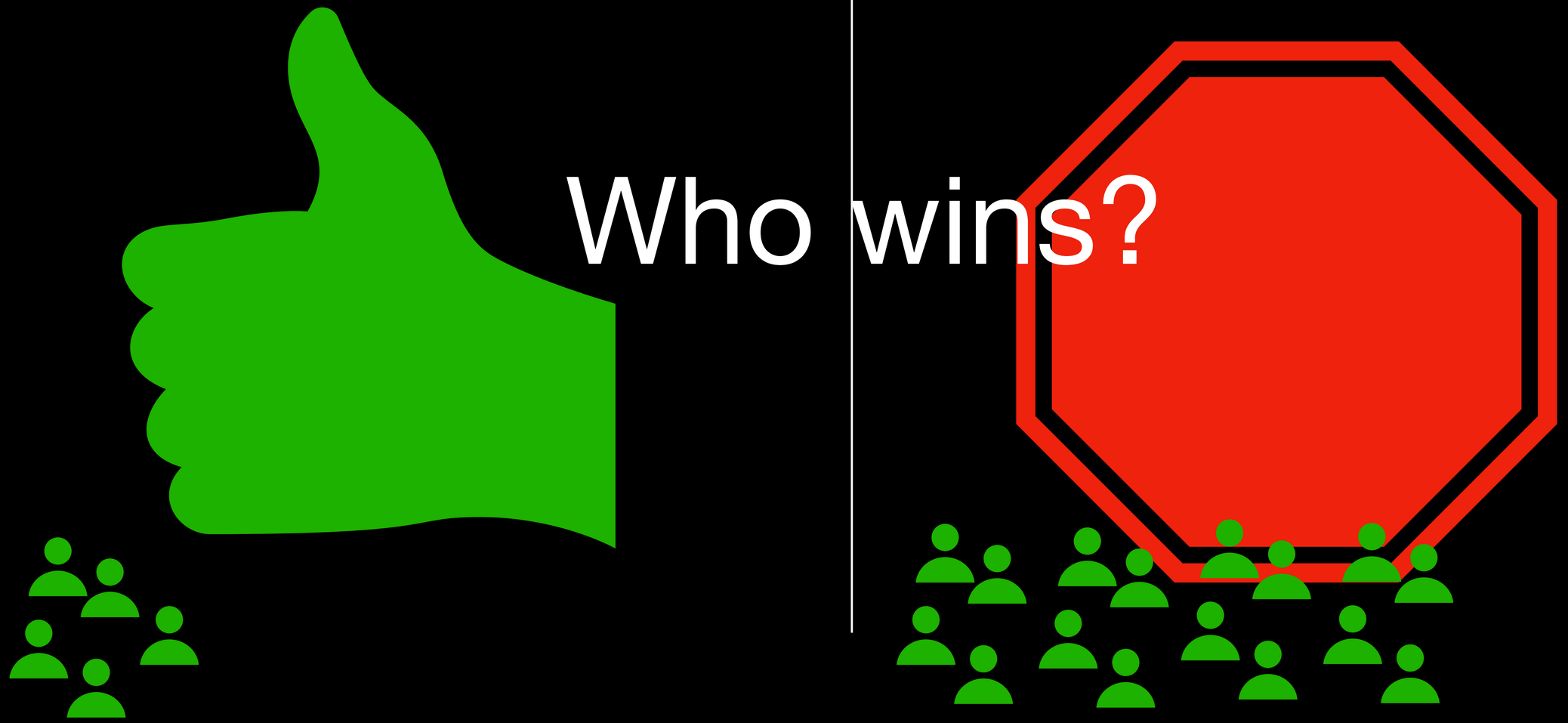
Business offline

A gateway



Business offline

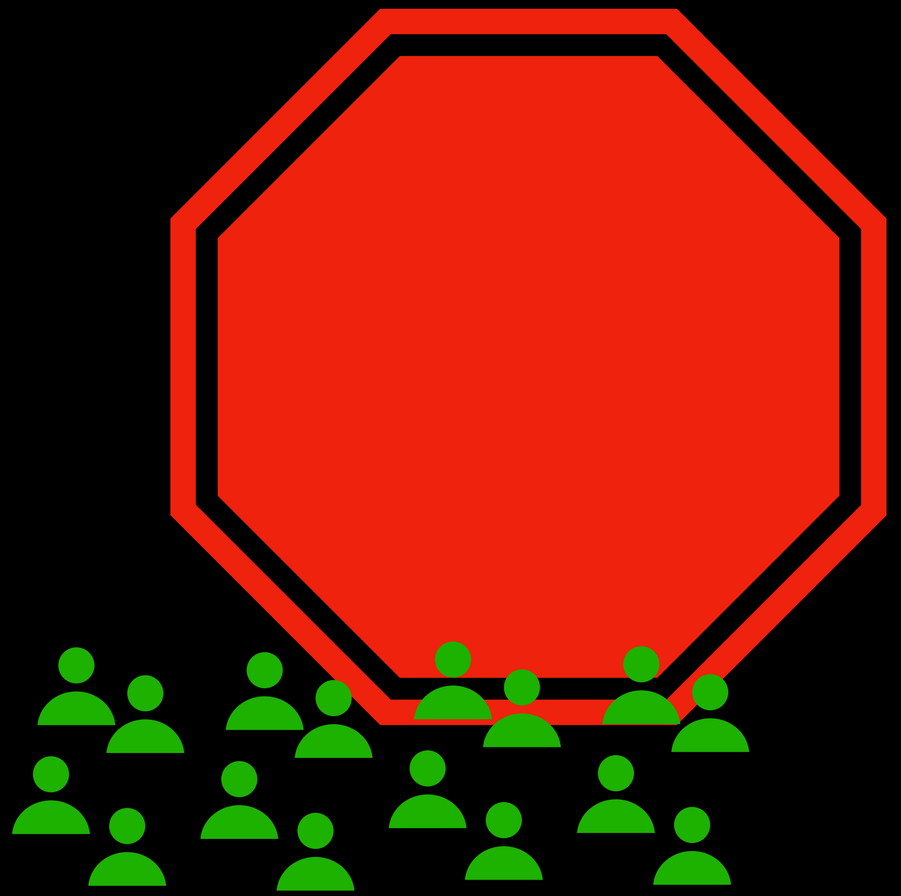
A gateway



The losers

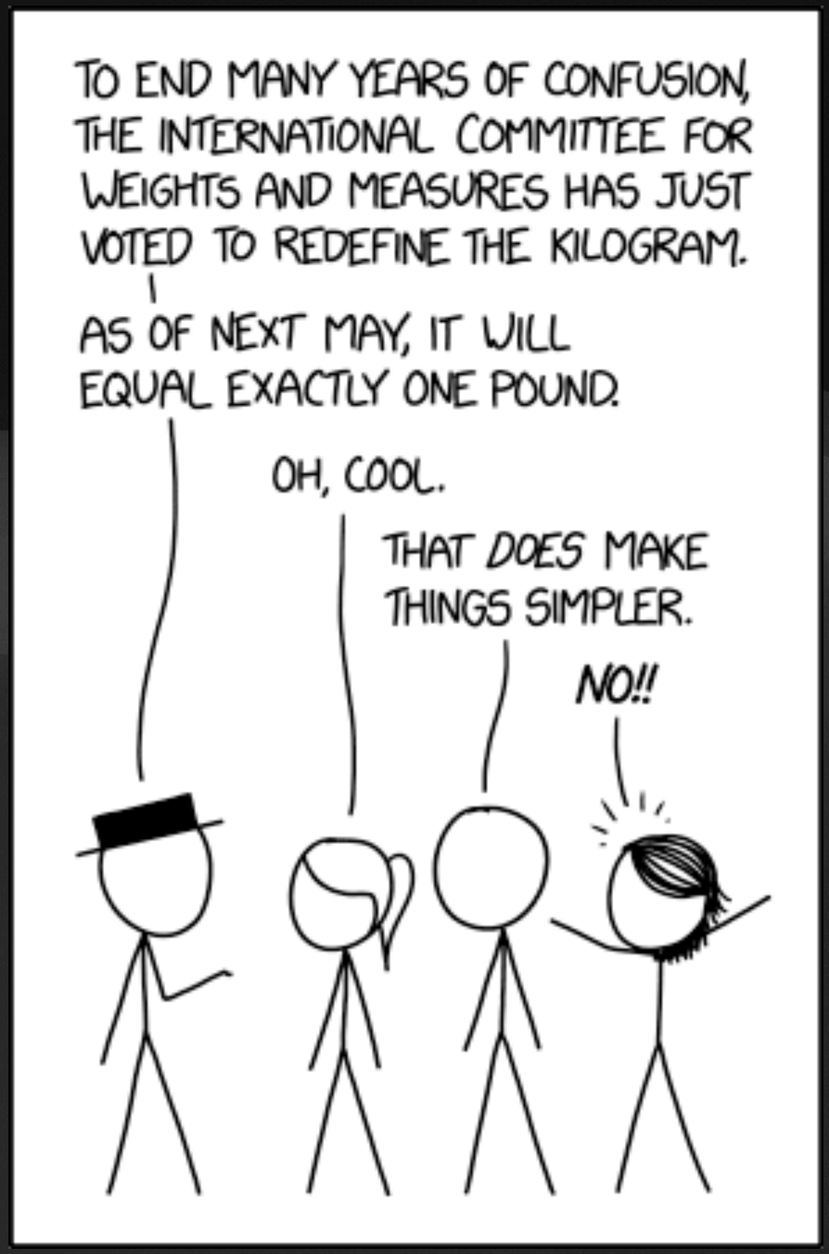
Is it accaptable

We need to ensure that this is acceptable

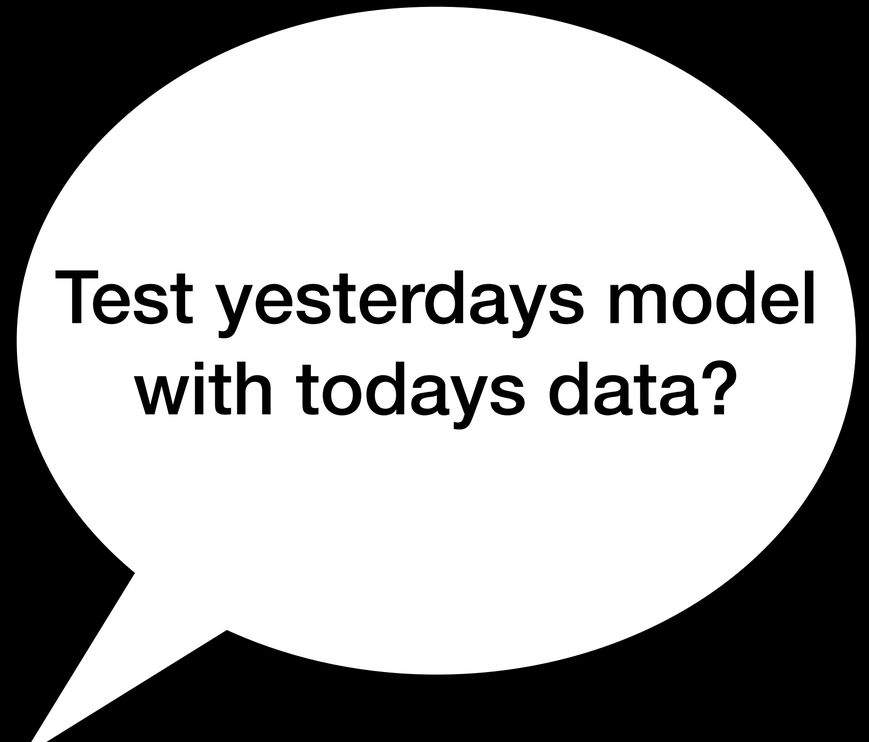


Do you know what metrics means?

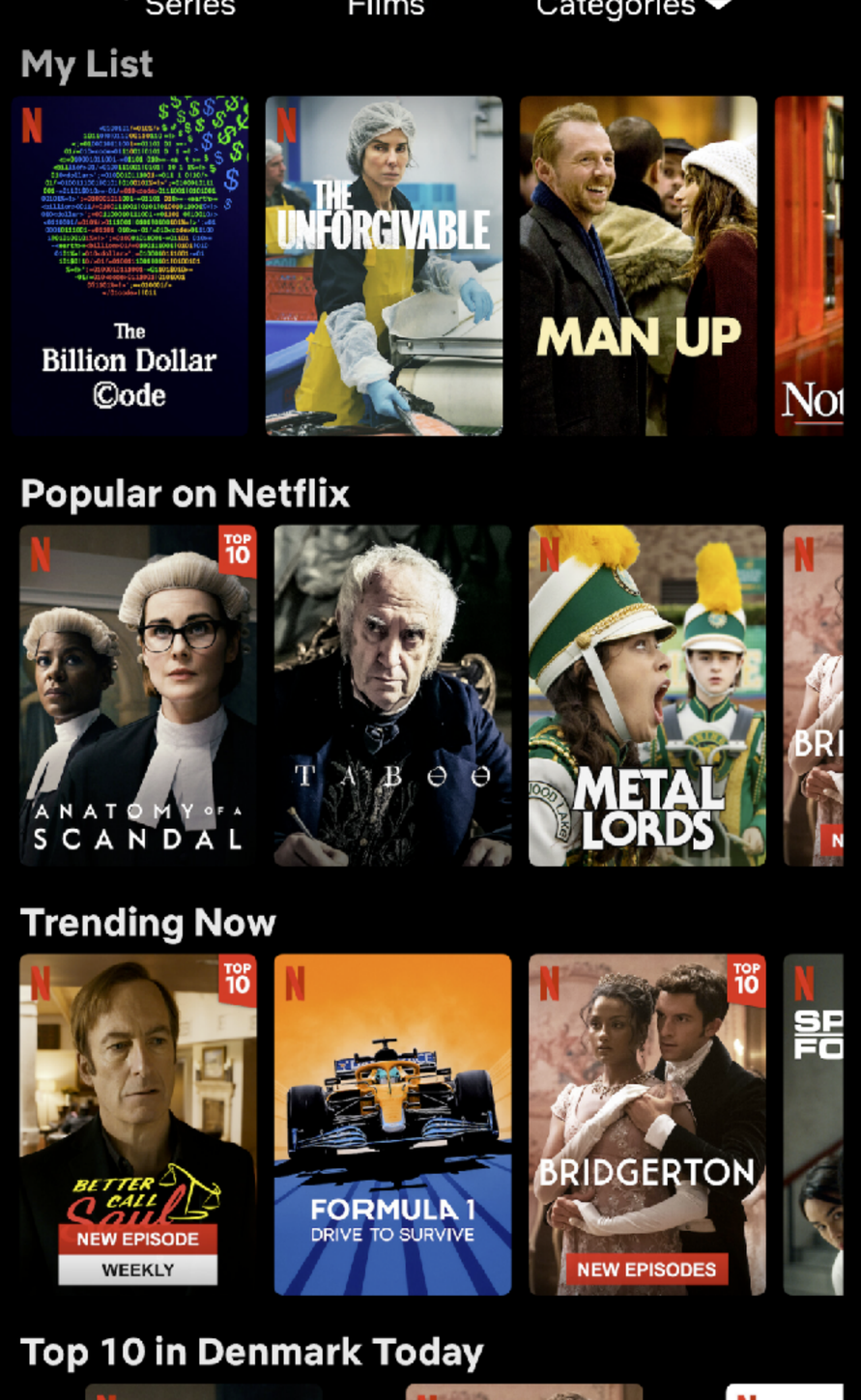
A new metric should be measured before changes are made



Test your metric on real data



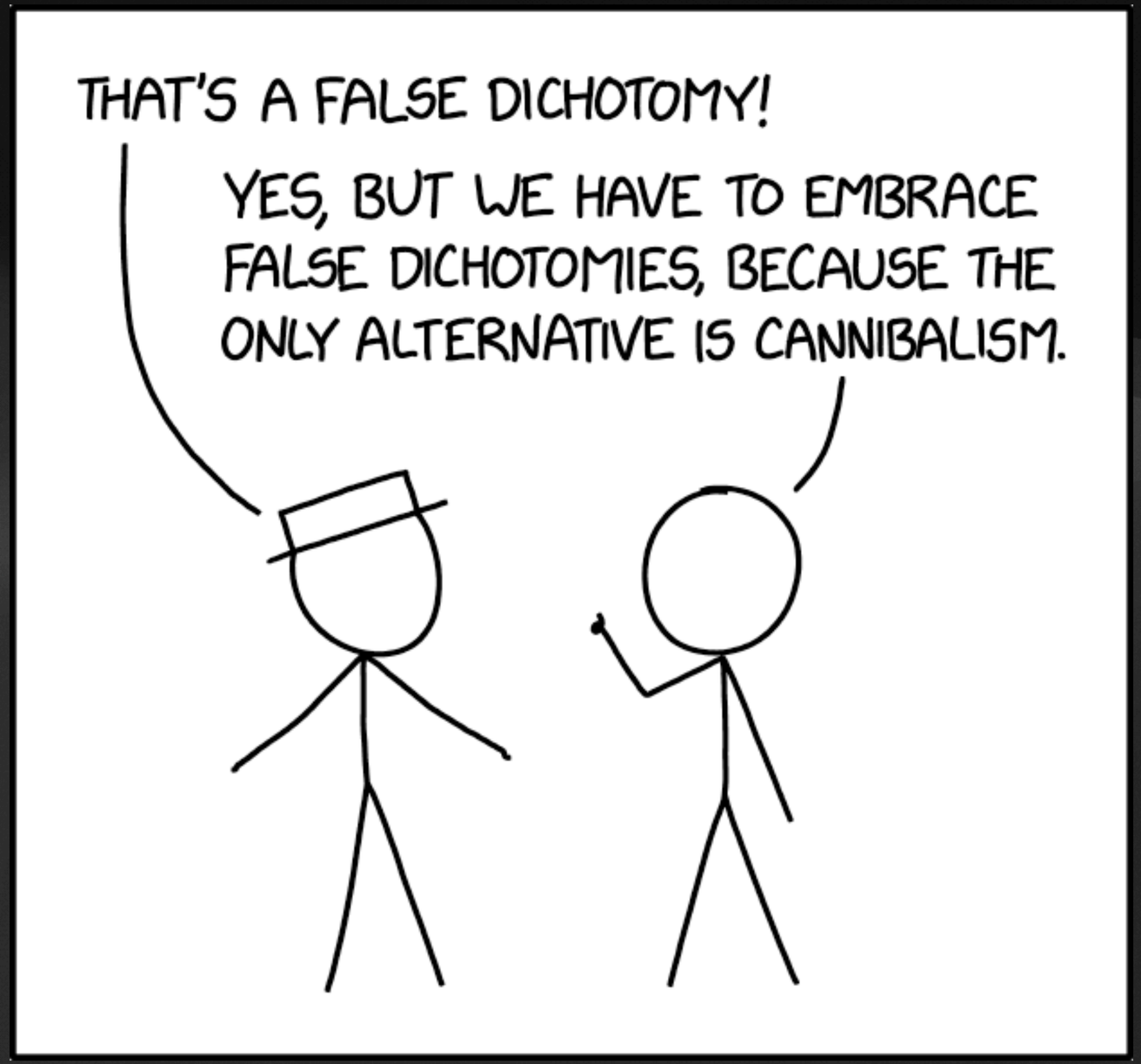
Consider the user and context



Did I view this row?

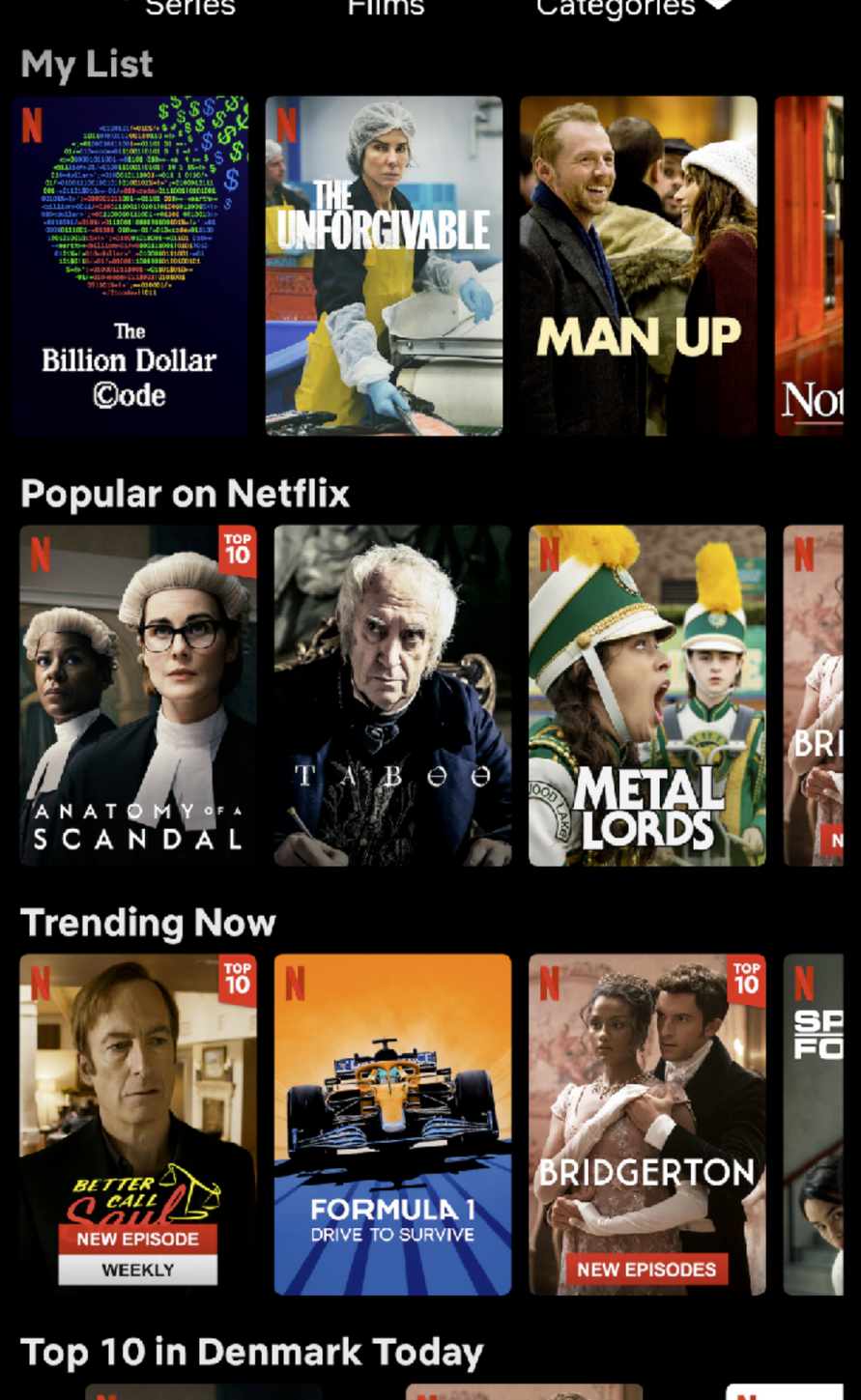
Cannibalisation

- The same content recommended.
- Change the customer journey.
- Position bias

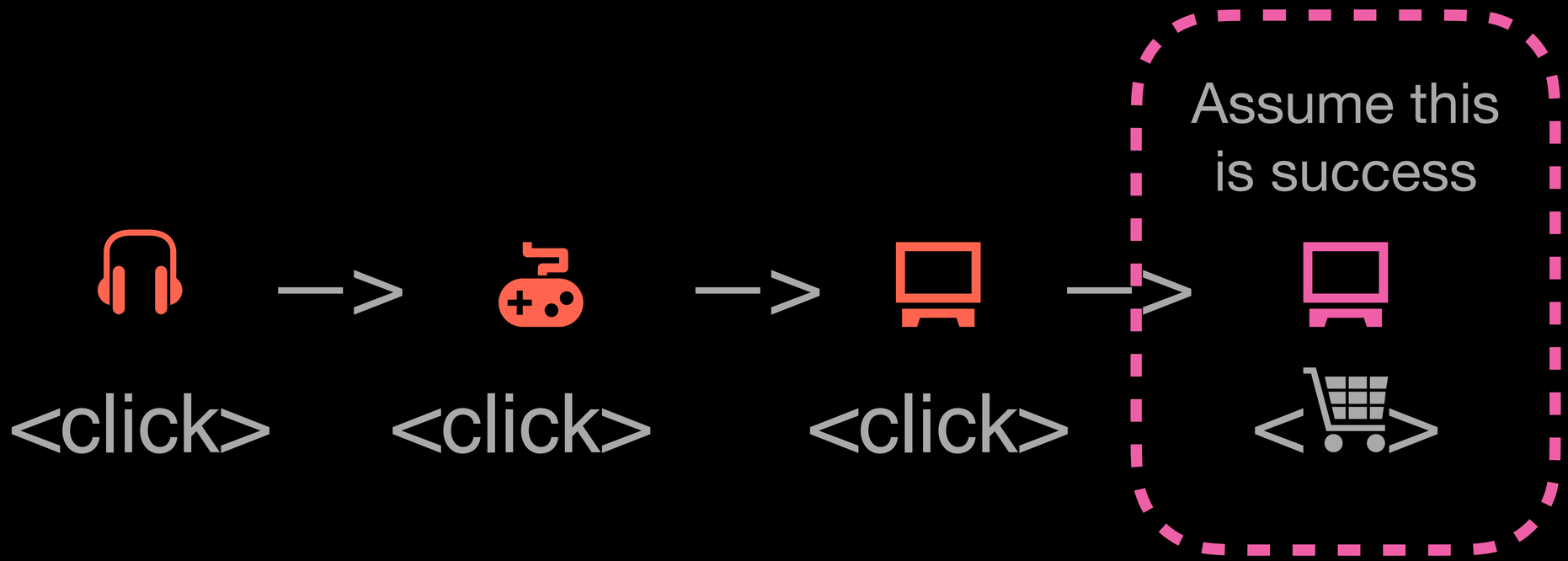


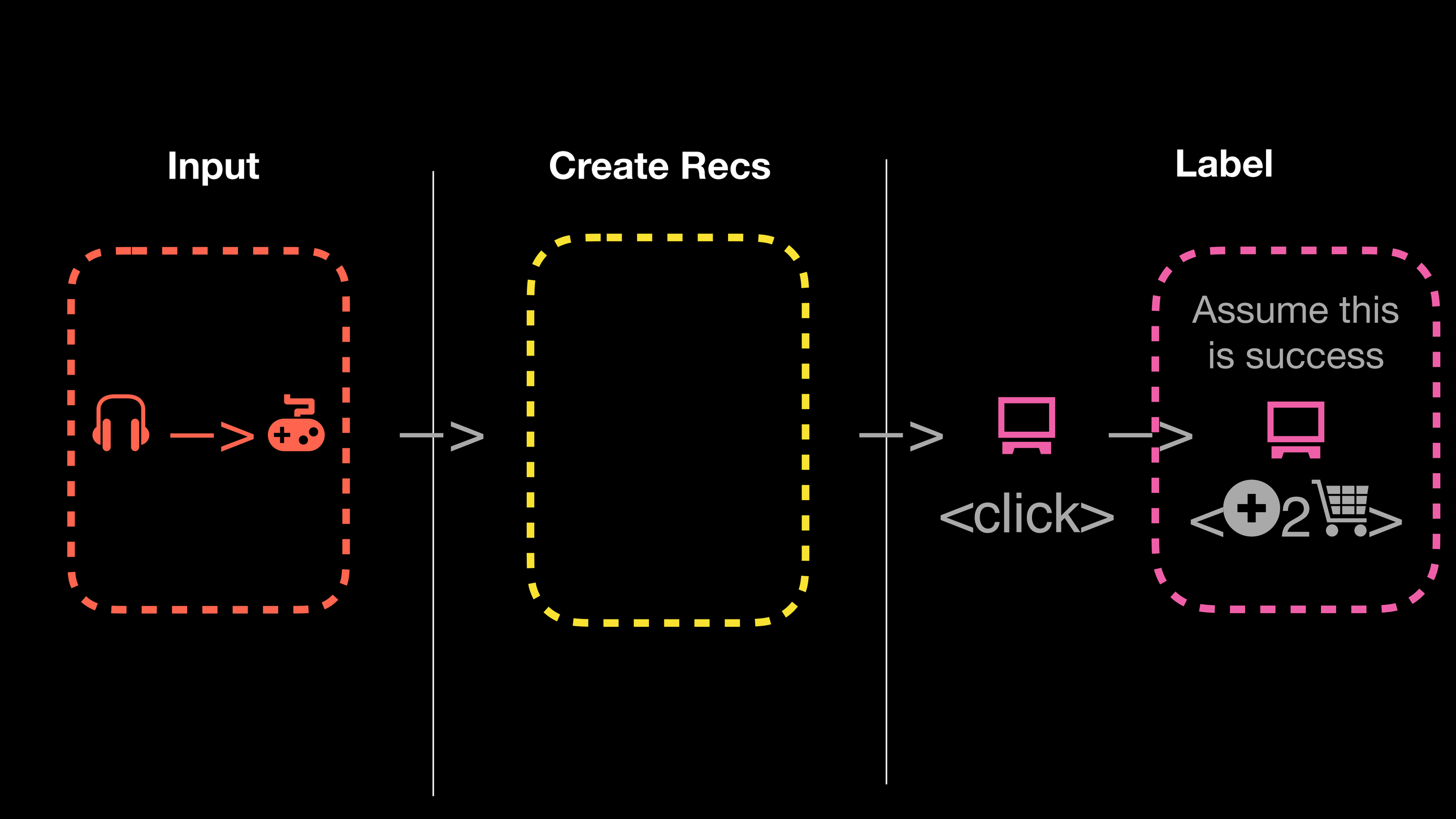
Position bias

This one gets more attention then the next one



Given a session

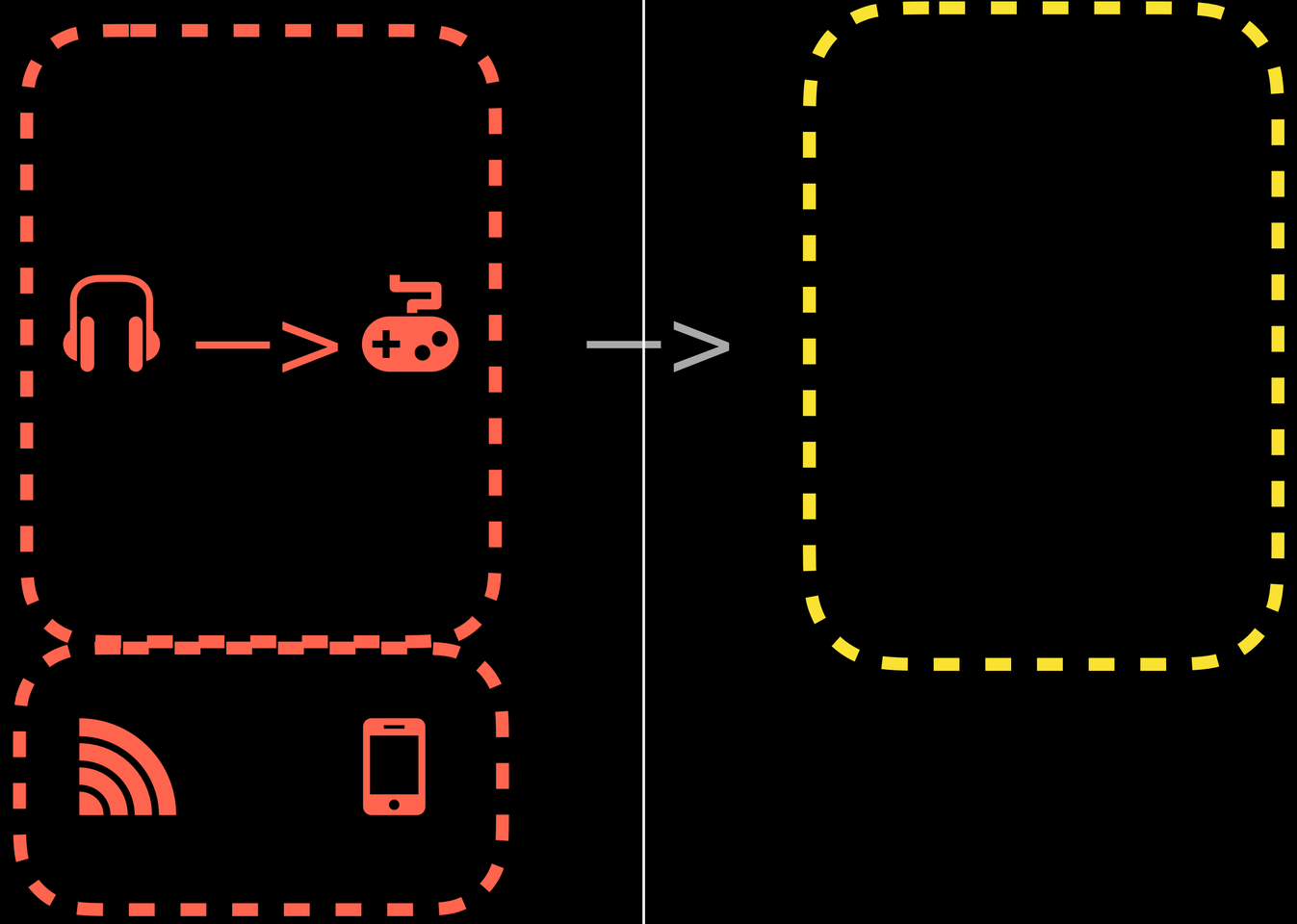


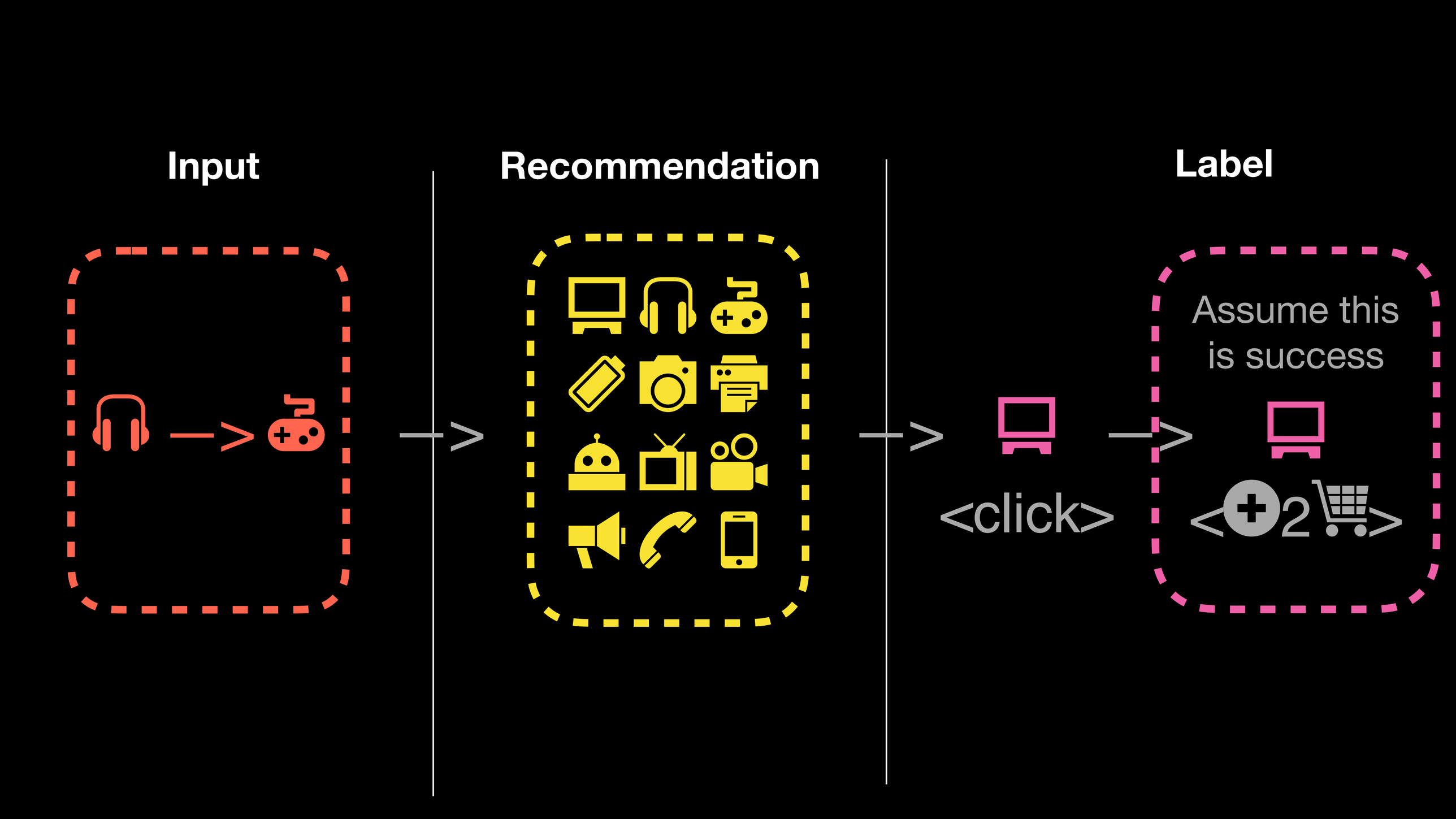

Input
Create Recs
Label
Assume this is success
<click>
2

Input

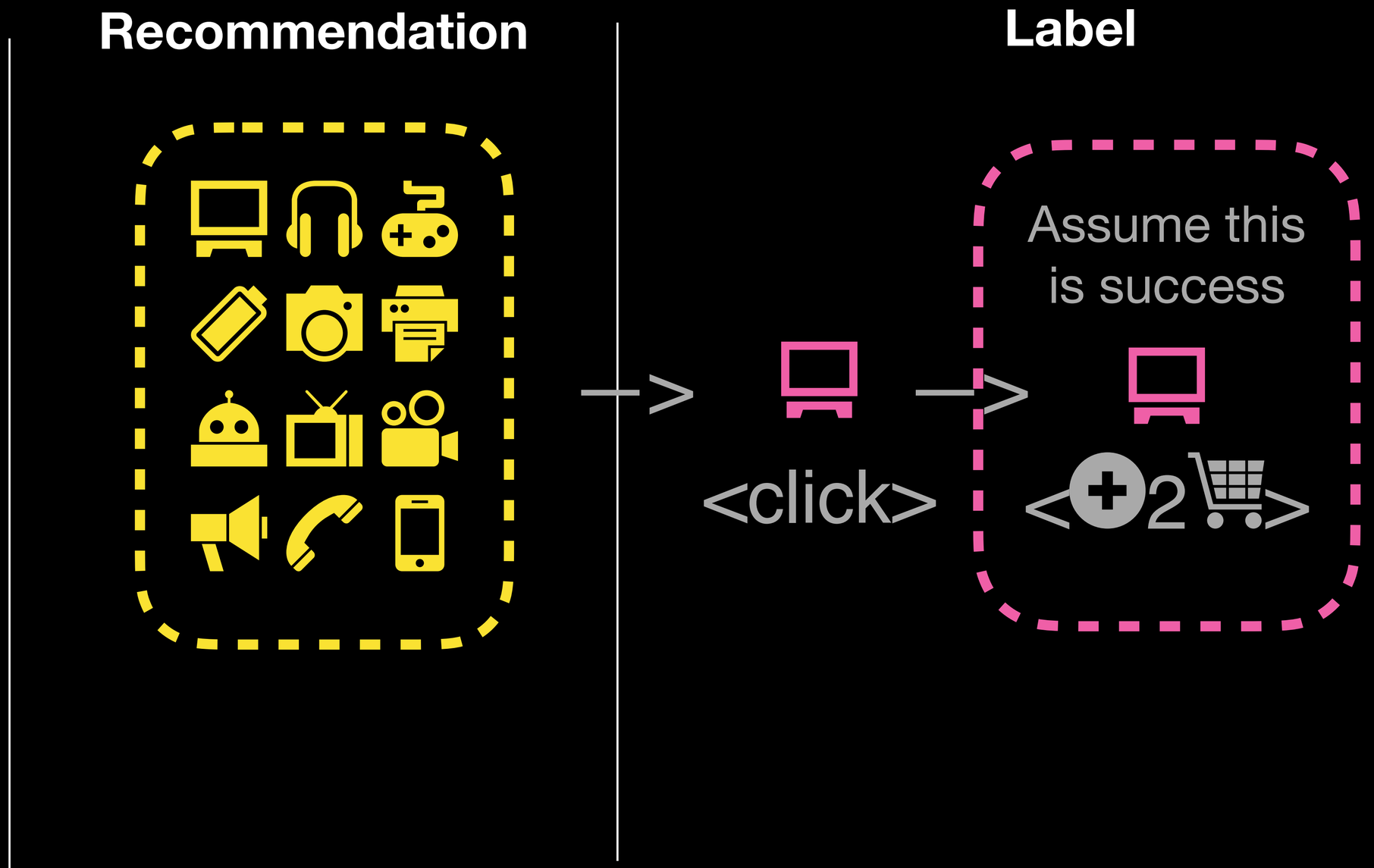
Create Recs

Hide this part




Input
Recommendation
Label
Assume this is success
<click>
<2>

Evaluation happens by comparing the recommendation with the historic outcome



To collect it over many sessions you should look at several different things

$$MAP = \frac{1}{|S|}\sum_{s \in S} AP(s)$$

Ranking metric

$$HitRate@PV = \frac{1}{|s|}\sum_{s \in S} Hit(s)$$

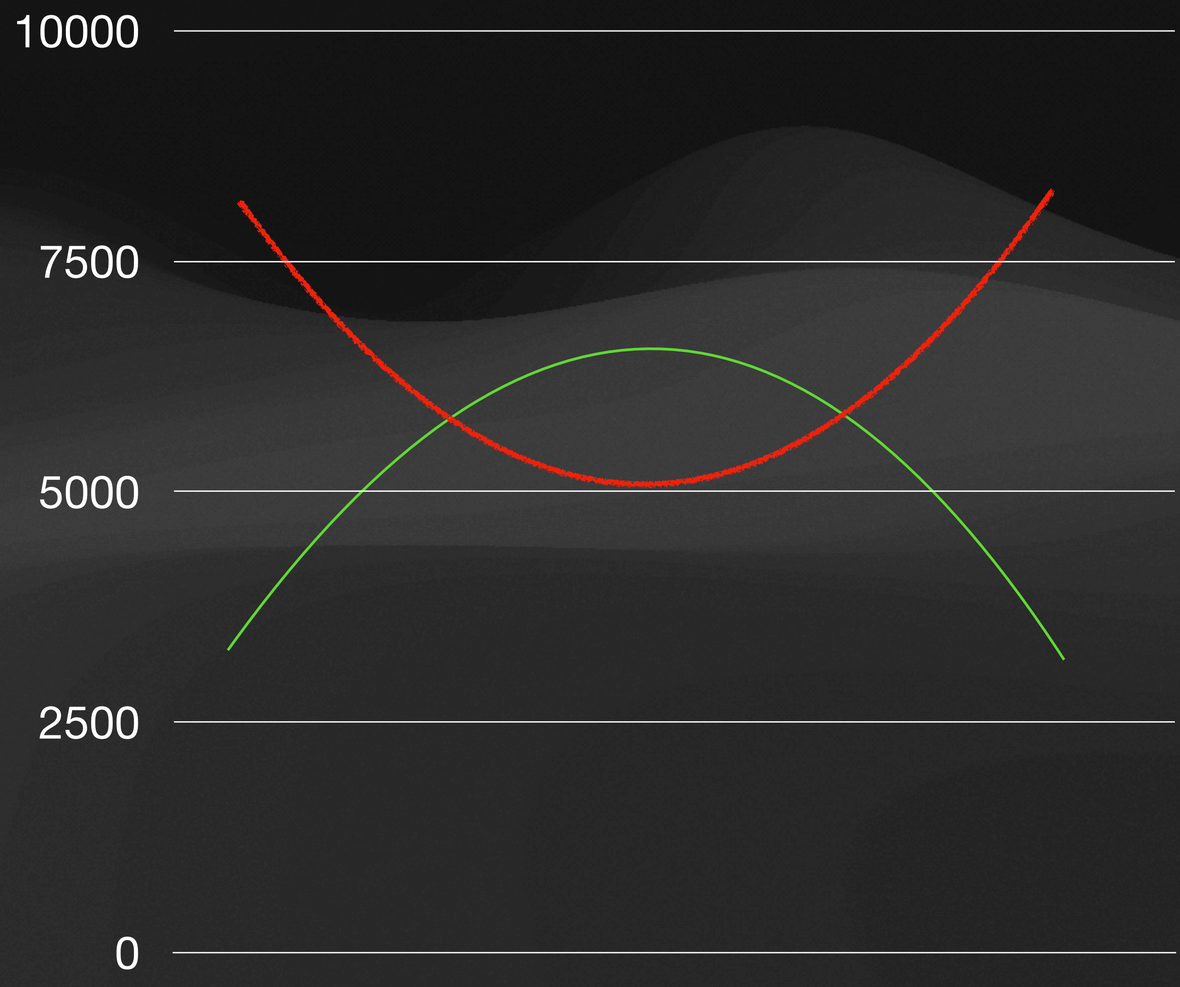
Hit metric

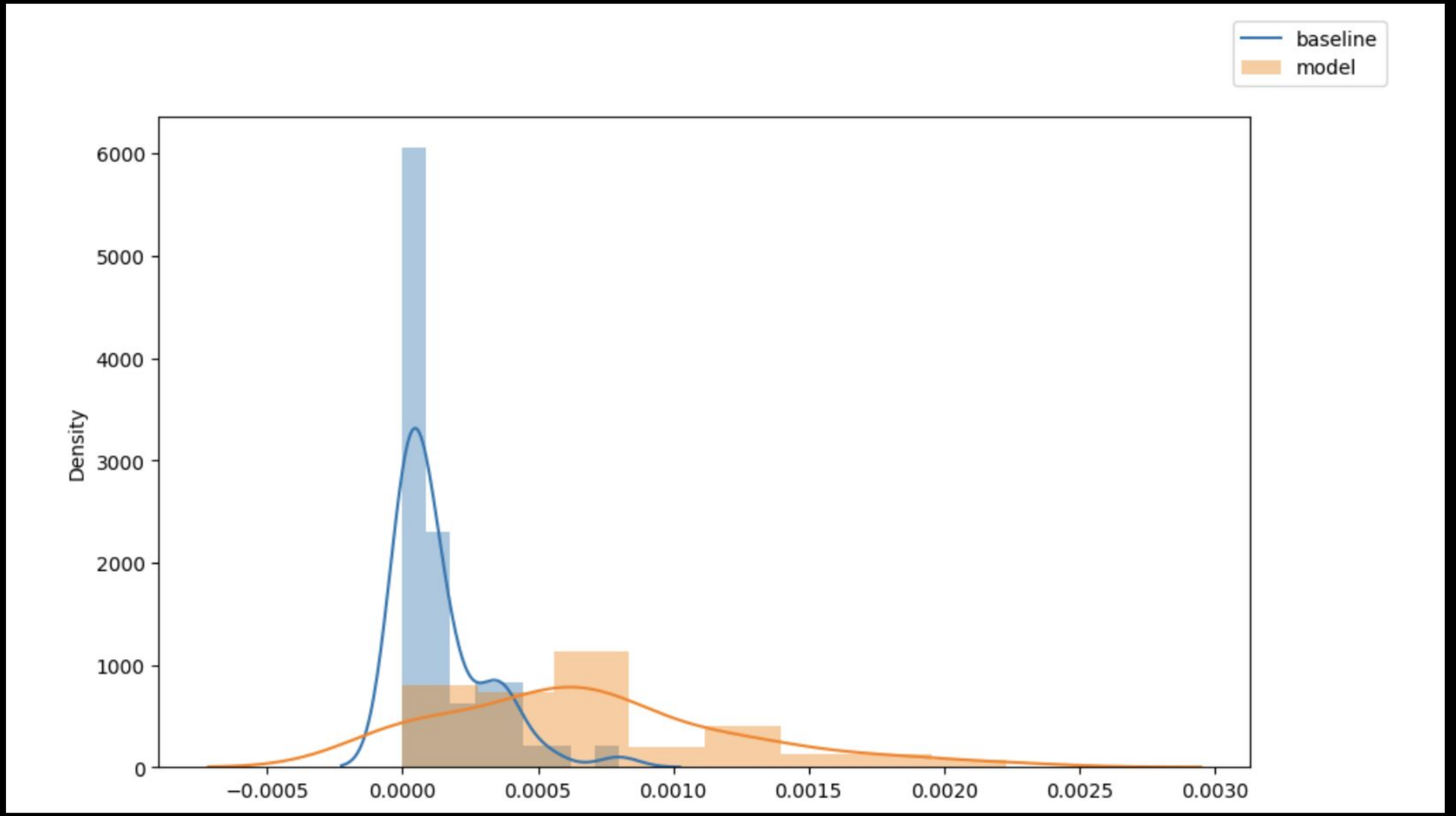
?

Diversity

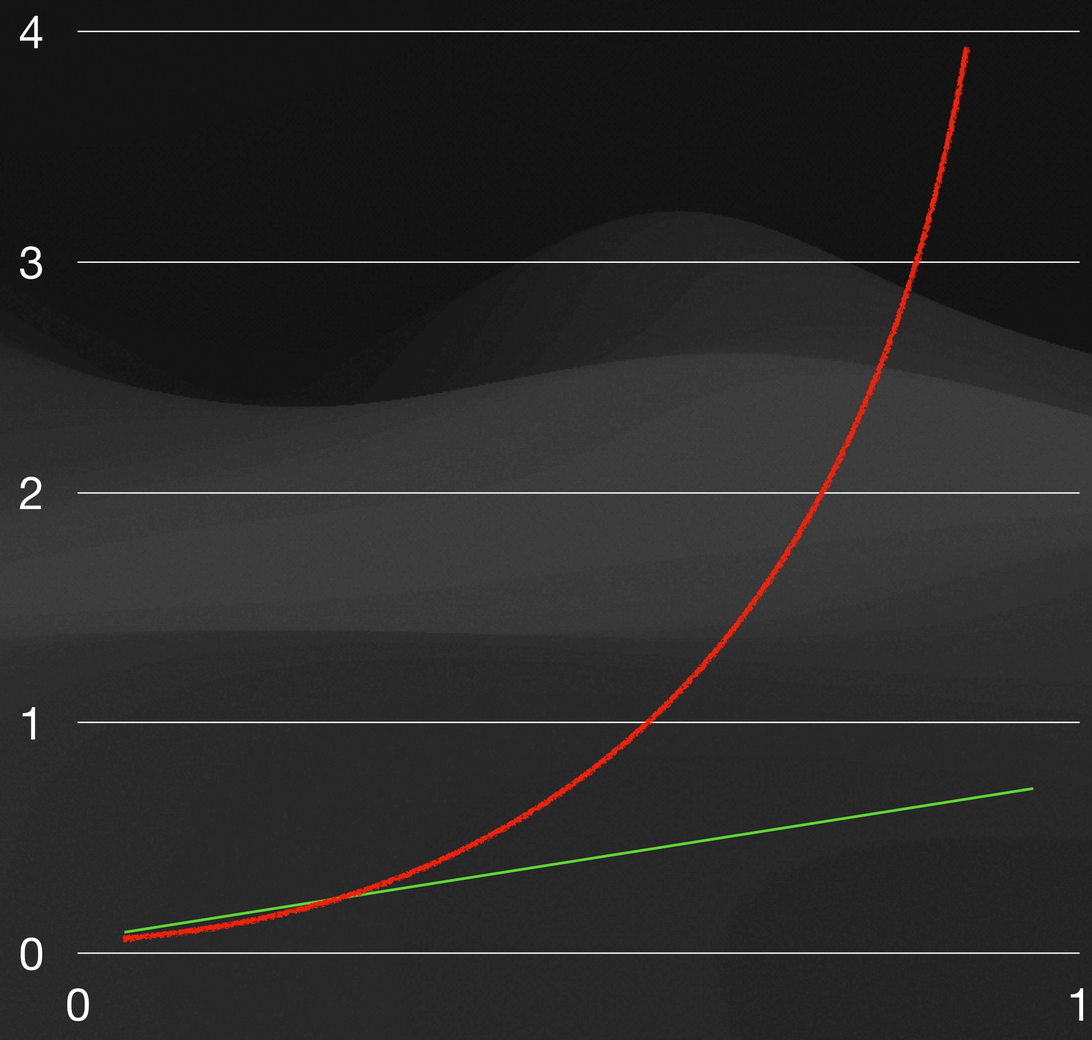
The one number often represents a distribution

- Which part to you want to improve?
- Left, centre or right side.



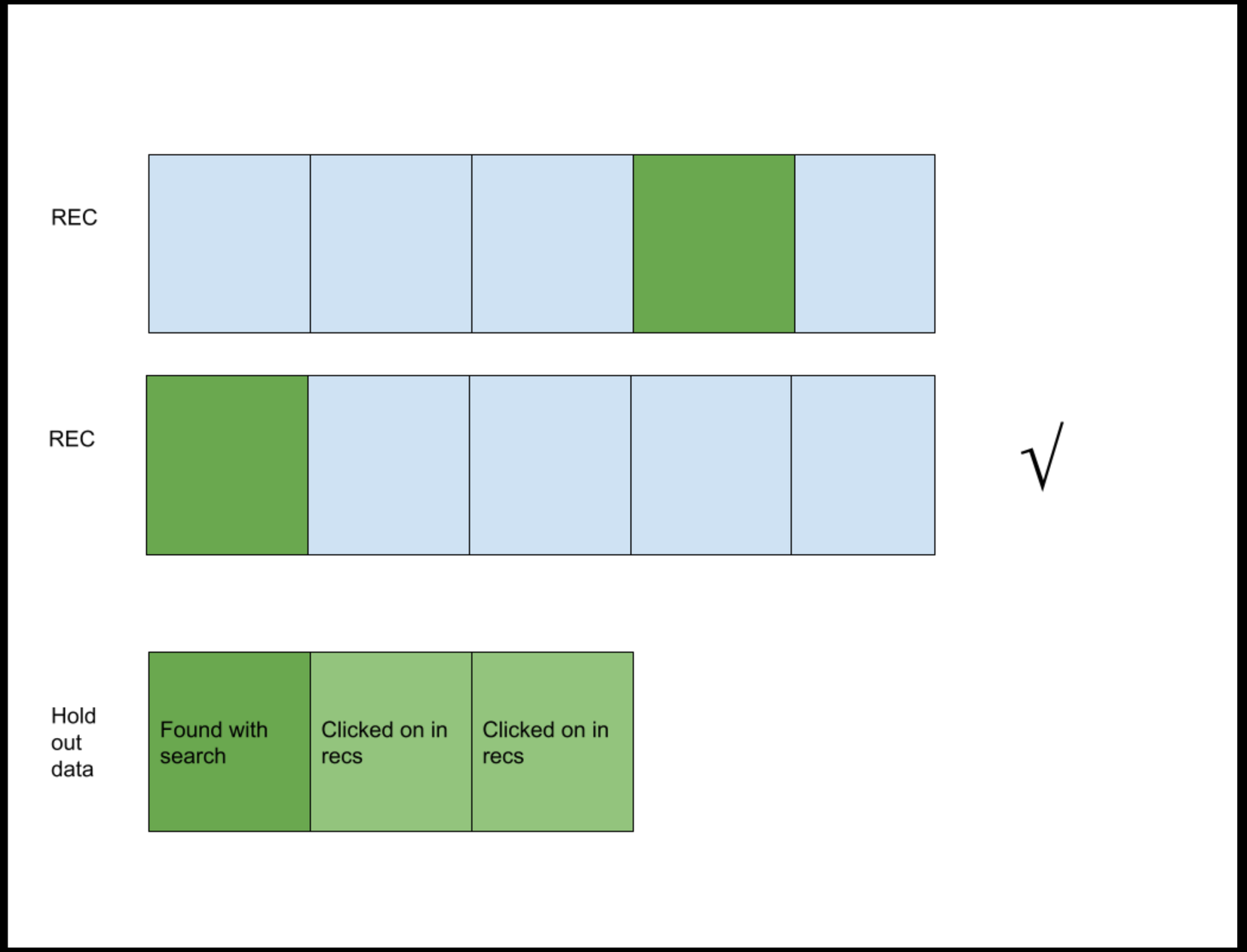

baseline
model
Density
6000
5000
4000
3000
2000
1000
0
−0.0005
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
0.0030

Understanding the metric



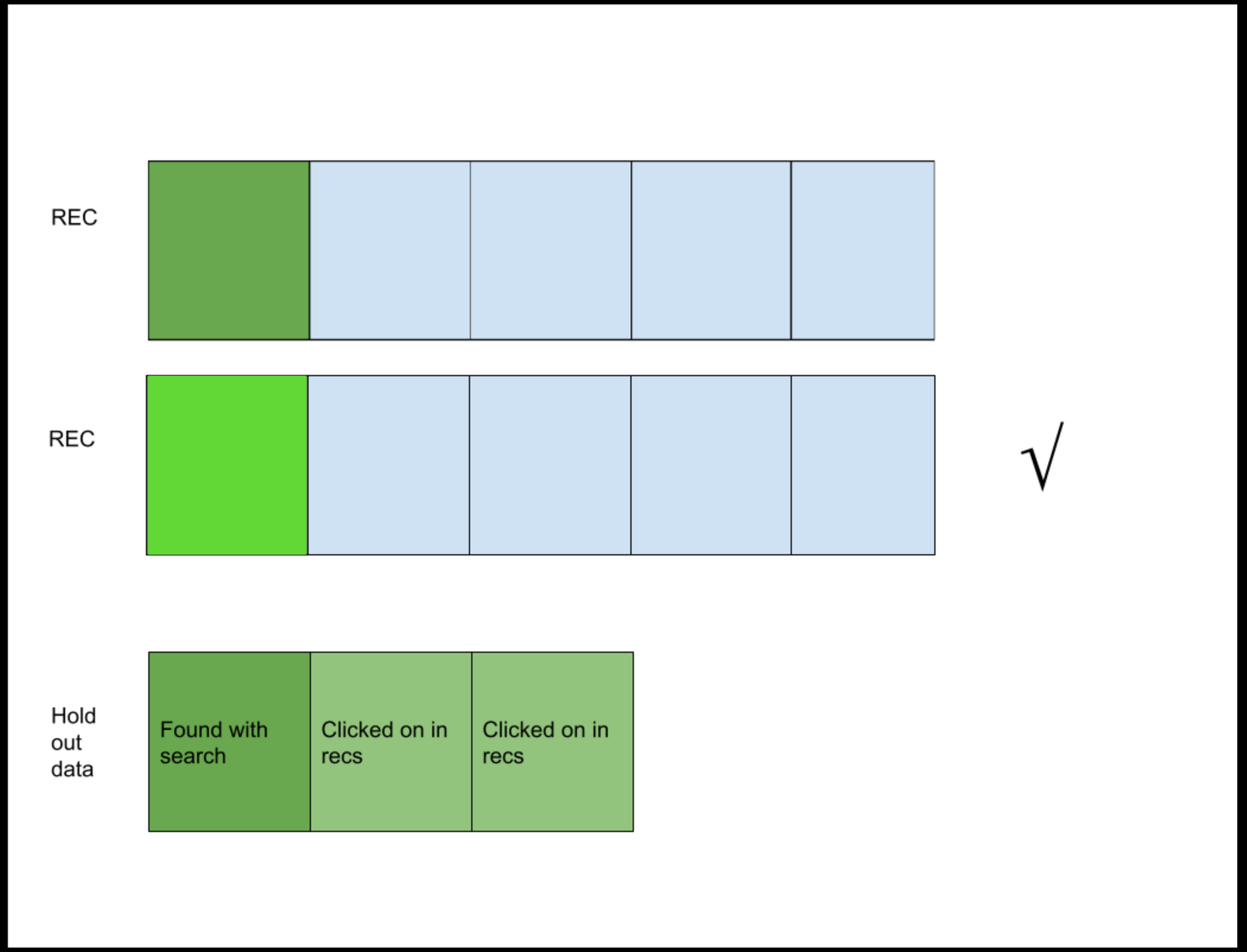
A few examples to consider

Which is better ?

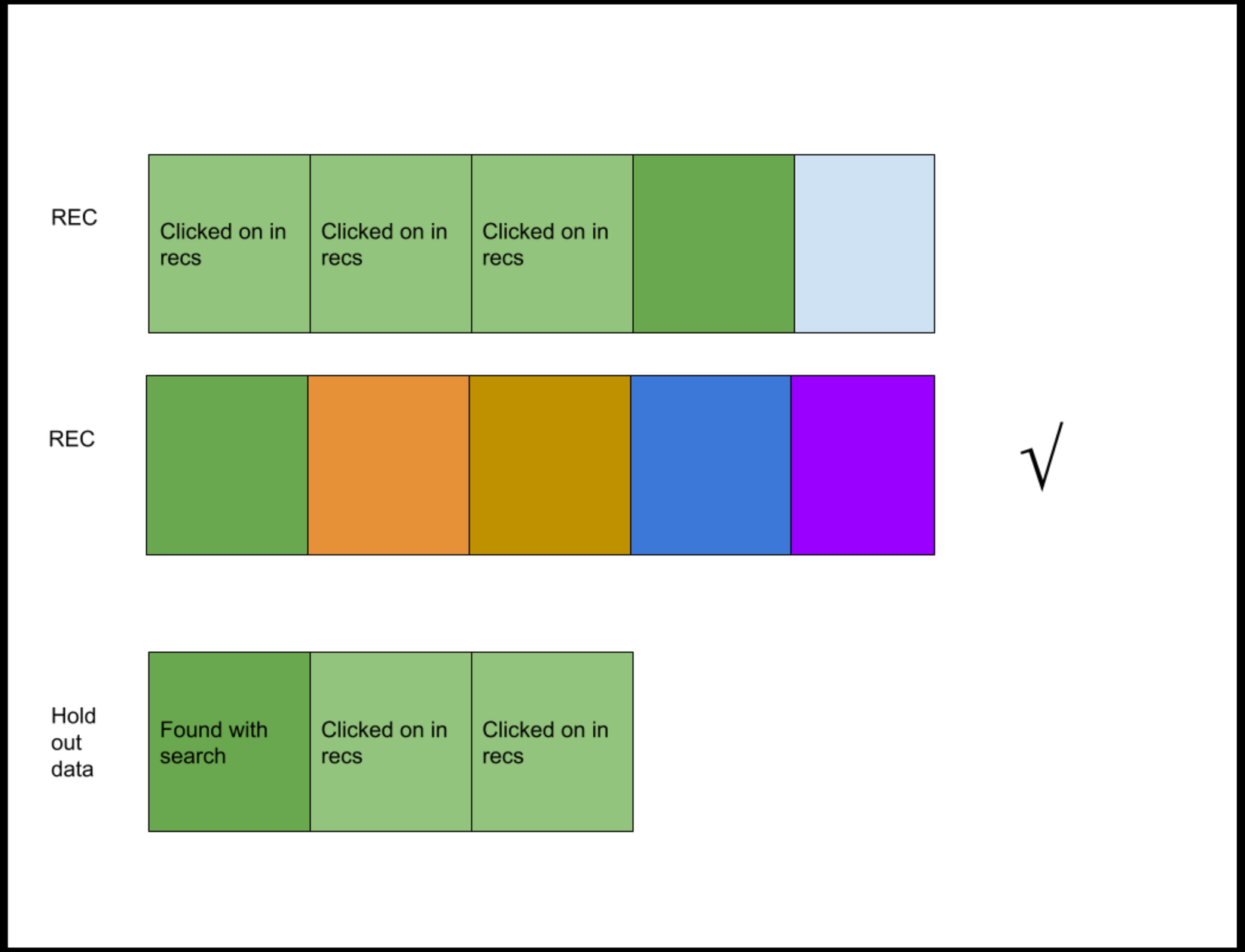


Which is better ?

Same position but different popularity.



Which is better ?



Watch out what you optimise for

Clicks vs time on site.

Ready of online testing?

If it doesn’t work figure out why

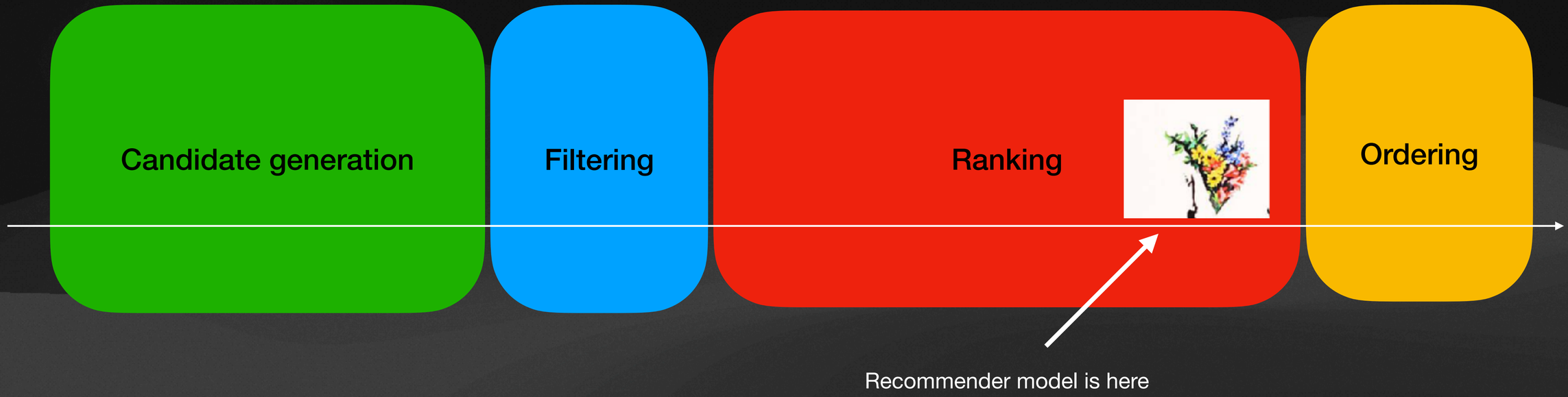
If the results looks good, assume its a mistake

Throwing Prototypes over the wall

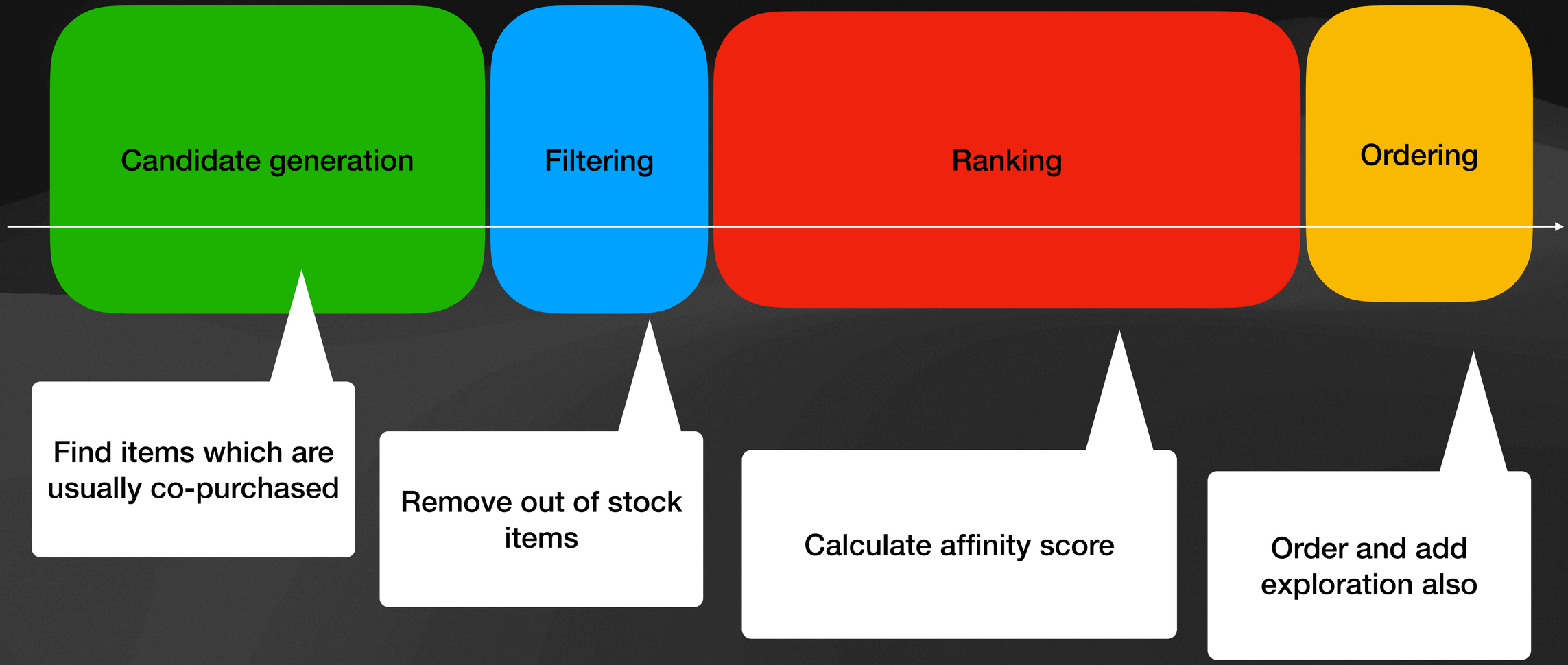
Is this the whole system?

Recommender model is here

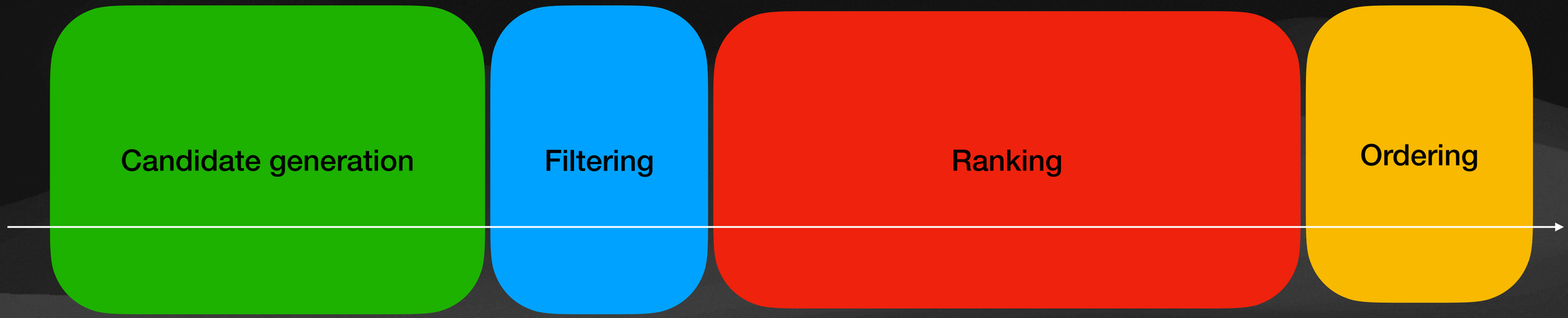
This is a Recommender System



This is a Recommender System



Evaluating the System

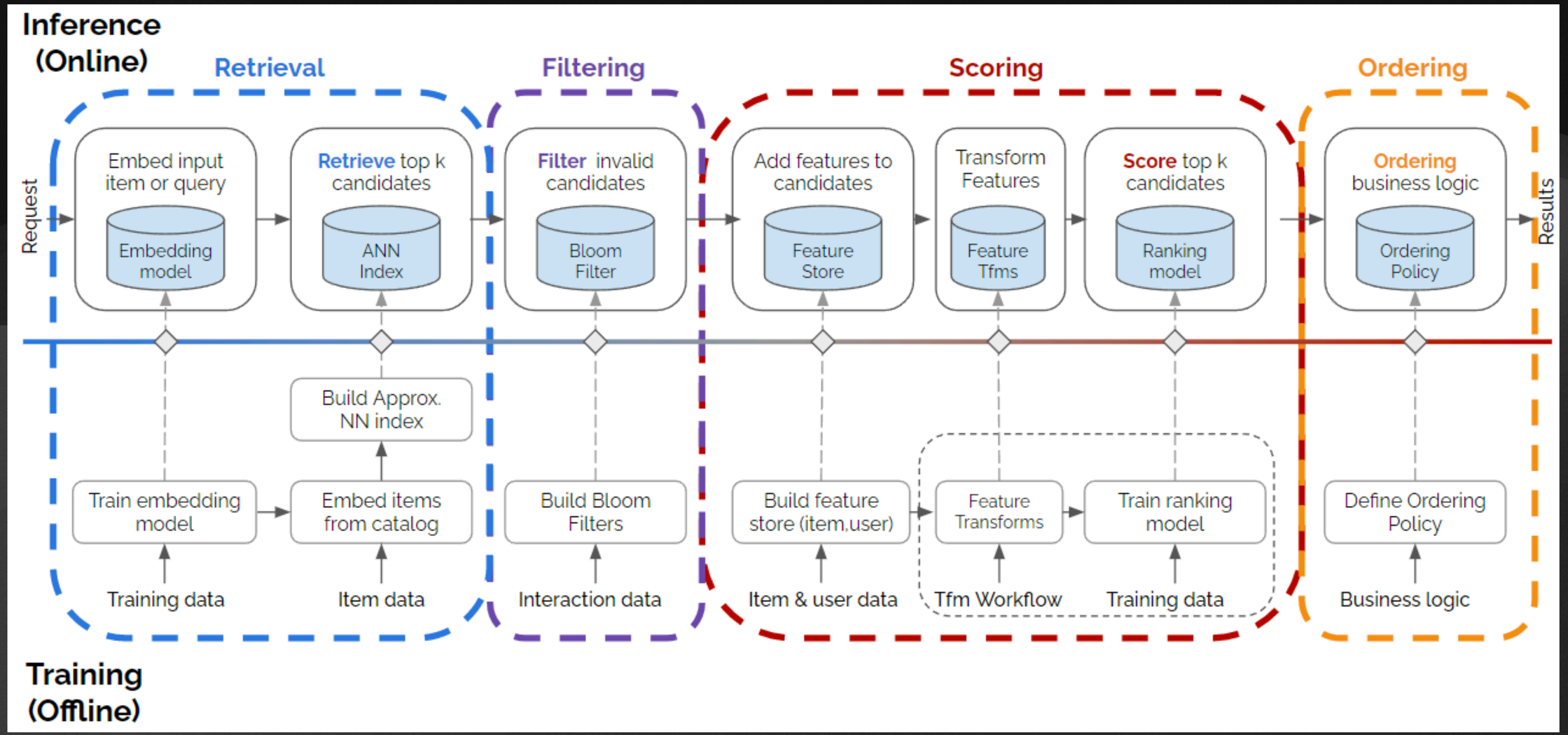


Optimise with recall, and then be too greedy

Optimise with “precision”

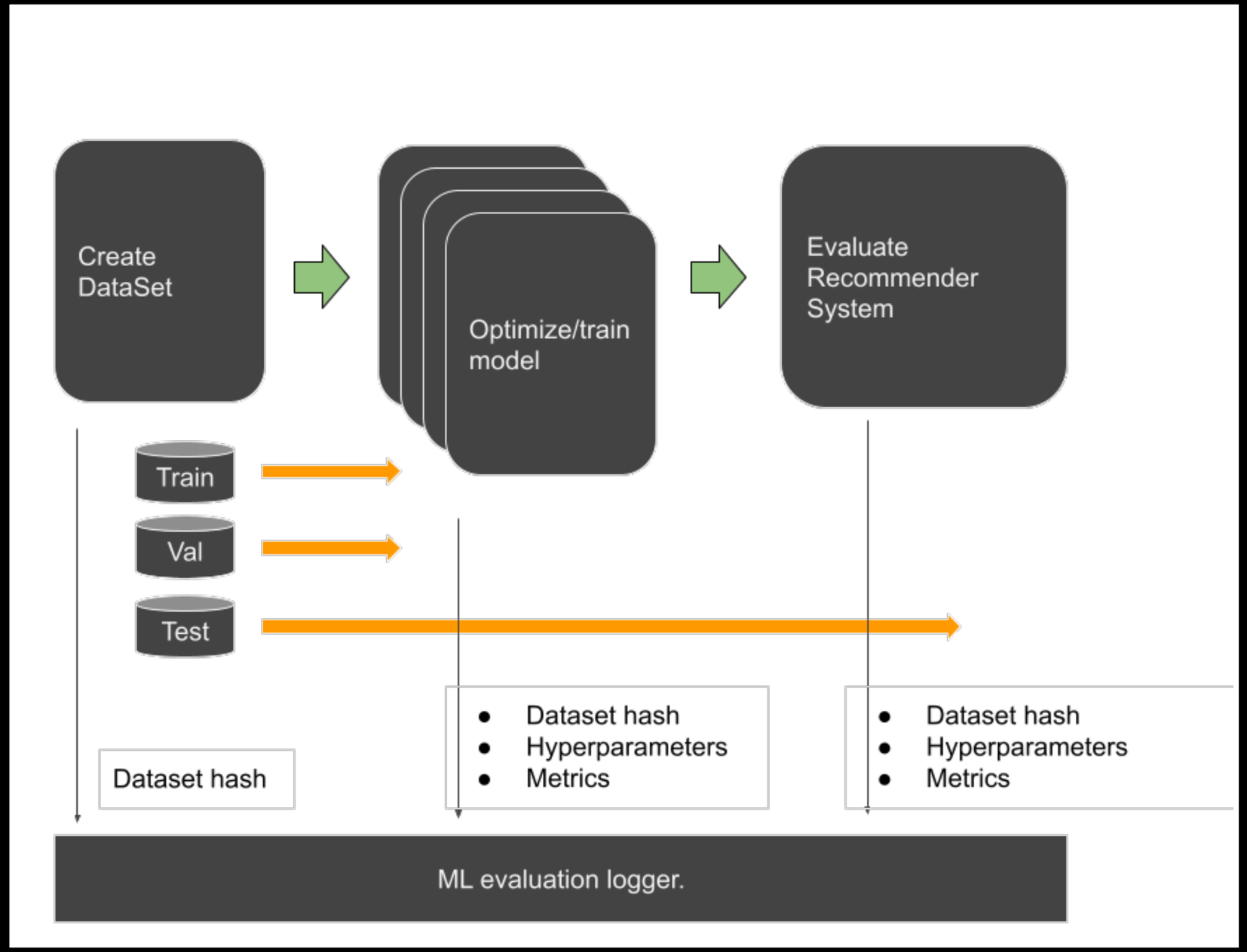
Can only ever rank the candidates

Even from Nvidia



https://medium.com/nvidia-merlin/recommender-systems-not-just-recommender-models-485c161c755e

ML development framework



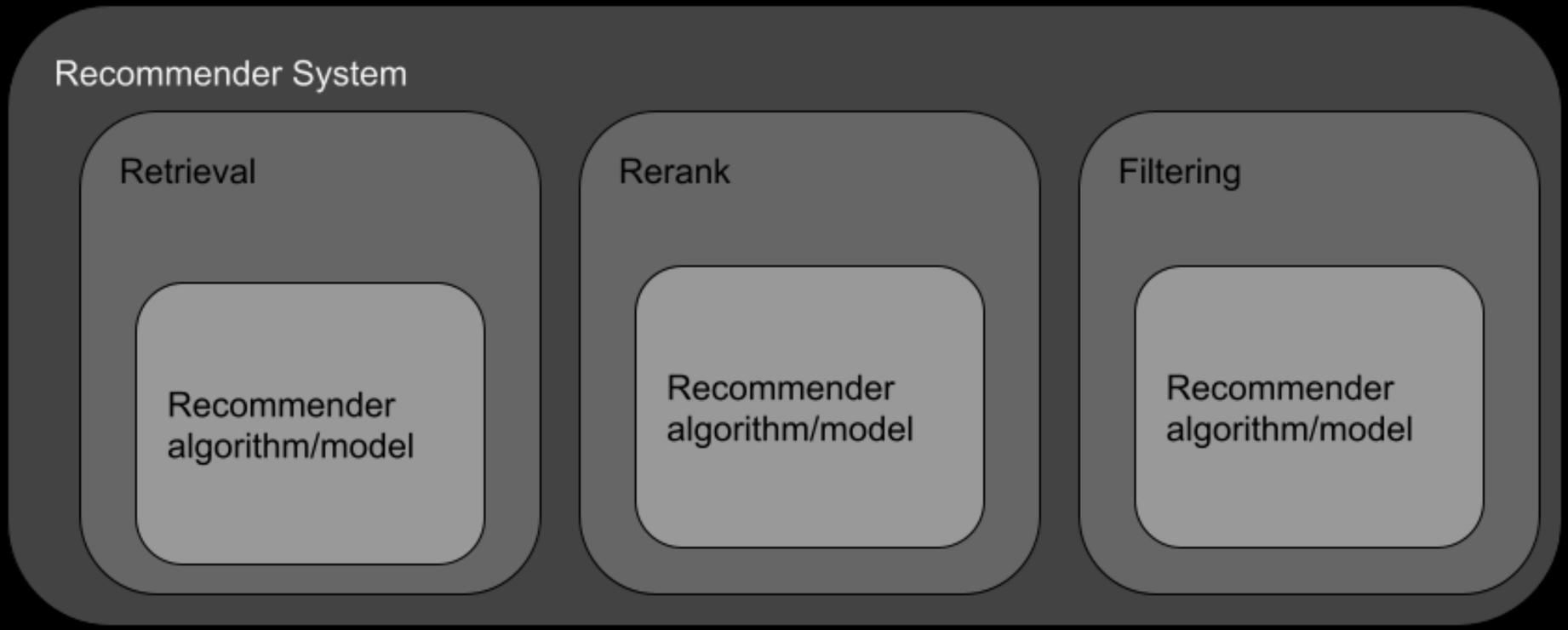
Some e-commerce examples

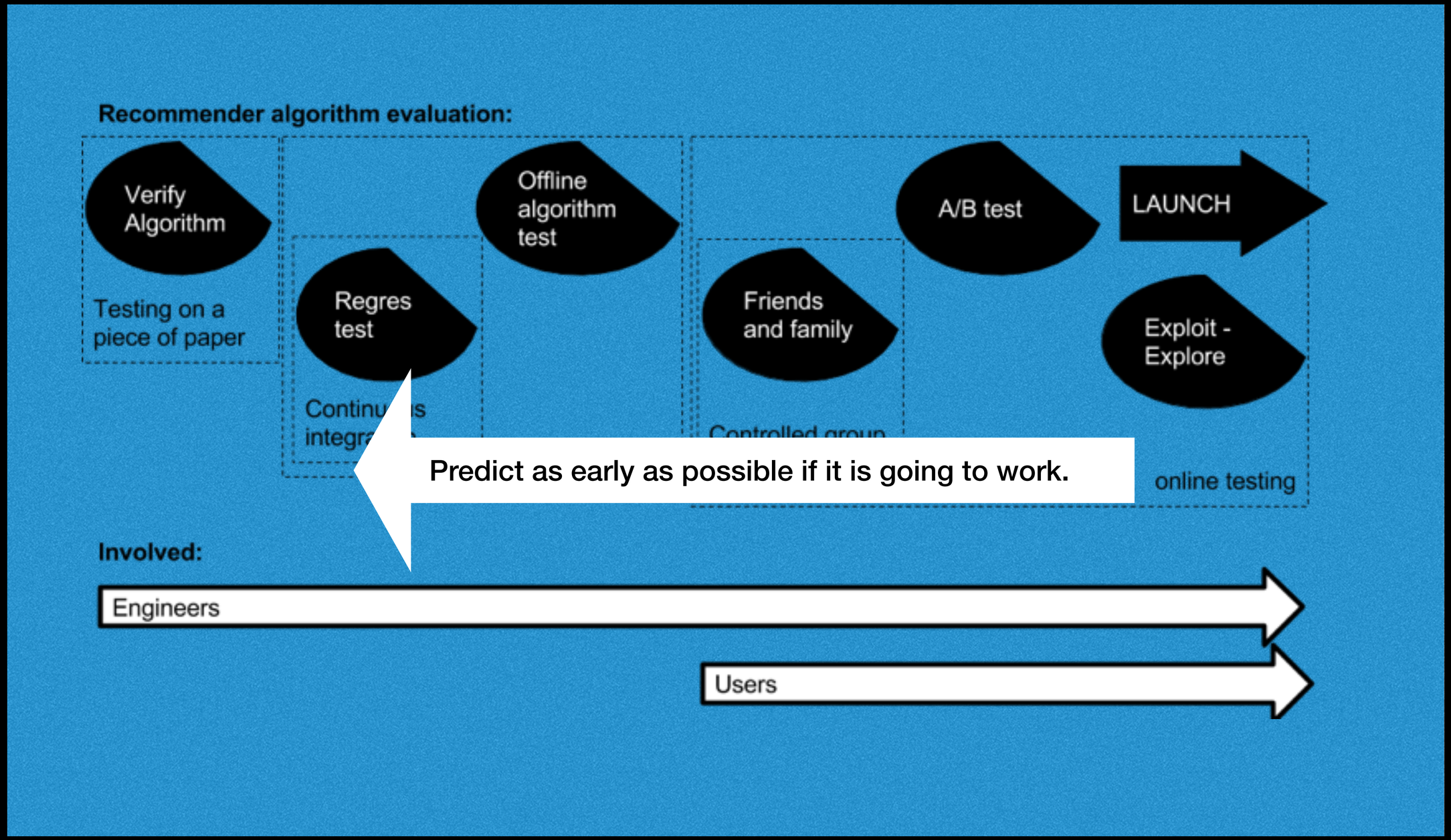
Some e-commerce examples

Some search example

Recs models & systems

But optimised for different things




Recommender algorithm evaluation:
Verify Algorithm
Testing on a piece of paper
Regres test
Offline algorithm test
Friends and family
A/B test
LAUNCH
Exploit - Explore
online testing
Predict as early as possible if it is going to work.
Involved:
Engineers
Users

We just want recommendations

No matter what you call the algorithm

We just want recommendations

No matter what you call the algorithm

Can we afford running the new model?

We just want recommendations

I’m worried

No matter what you call the algorithm

Can we afford running the new model?

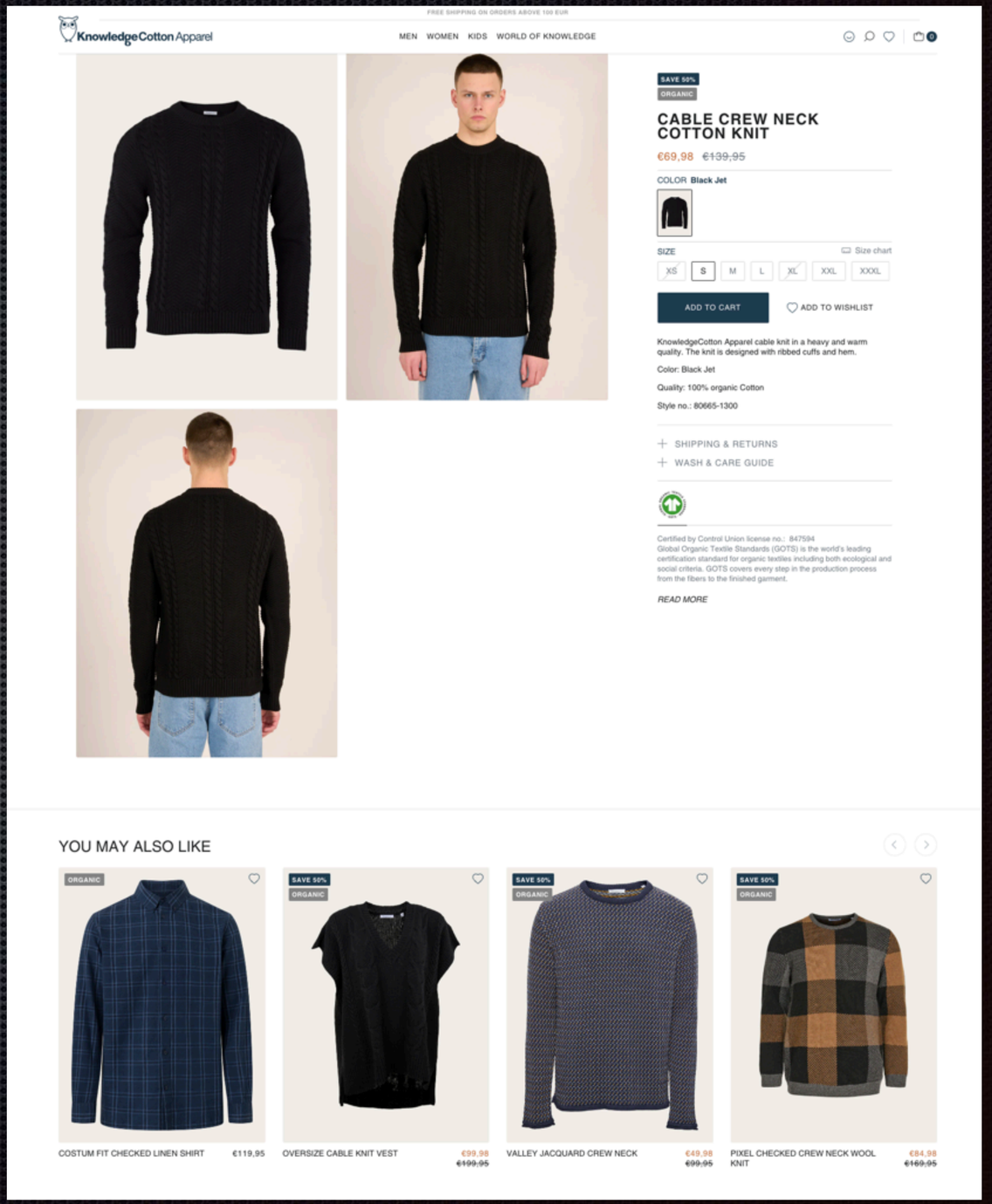
Most non-tech managers becomes worried if you refer to a research article.

Consider Mixing models

- Interleaving
- Casading (CF, CB, Pop)
- Using content to improve behavioural signal.

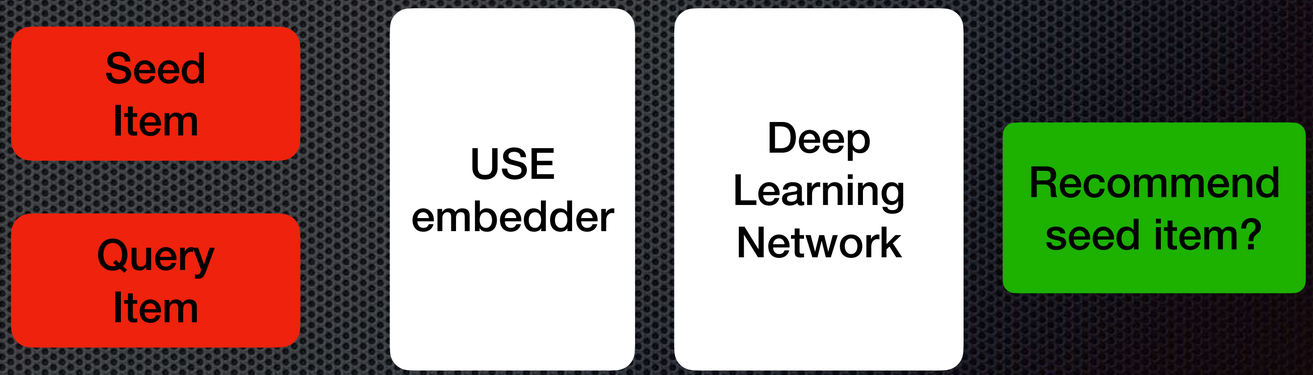
Simply interleaving

- Collaborative filtering first and third
- Content based second and fourth.



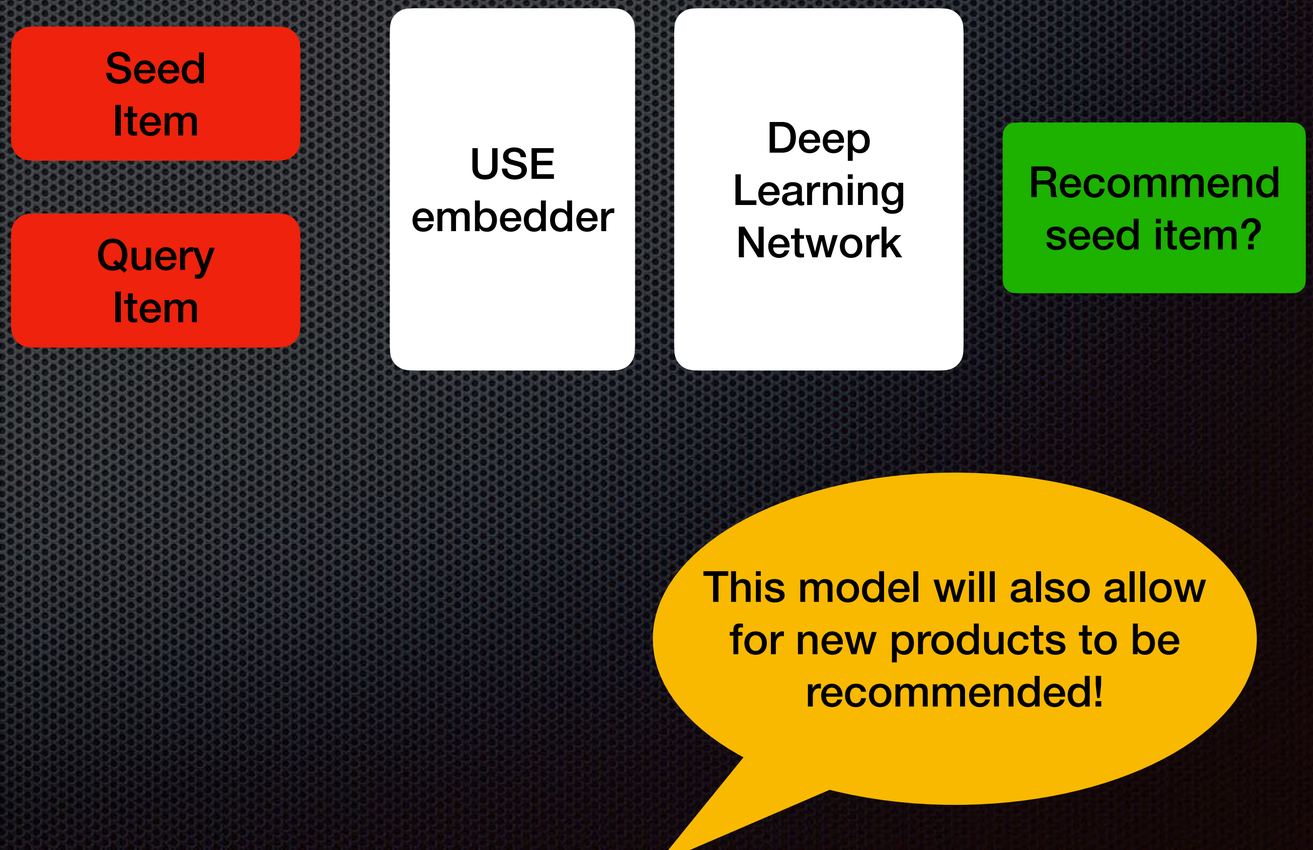
Content and behaviour

- Create embeddings of all items.
- Train a model using behavioural data to predict Probability that two items will be purchased together.



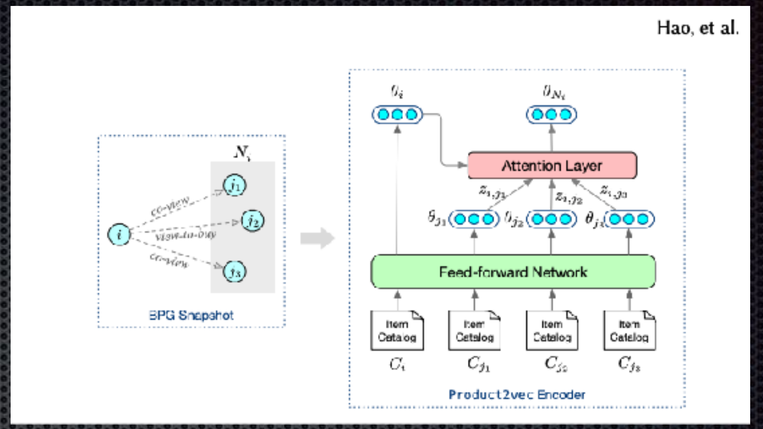
Content and behaviour

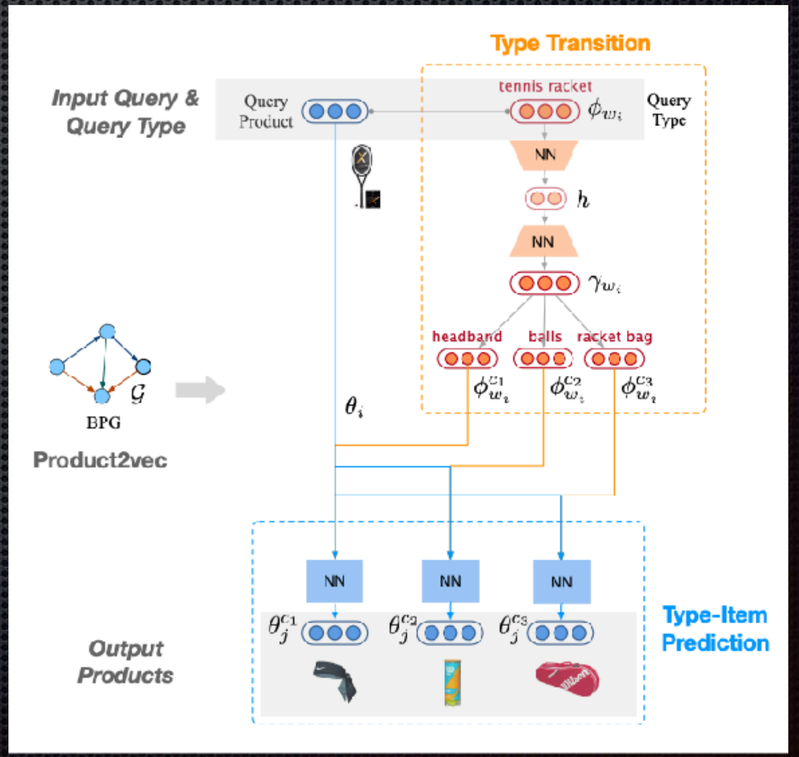
- Create embeddings of all items.
- Train a model using behavioural data to predict Probability that two items will be purchased together.



P-companion

- Graph Attention Network
- Process:
  - Create graph embeddings
  - Rec Product type to product type
  - Find best items in product type





https://assets.amazon.science/d5/16/3f7809974a899a11bacdadefdf24/p-companion-a-principled-framework-for-diversified-complementary-product-recommendation.pdf

Business considerations when deploying a recommender

Several aspects to consider.

Users

Business

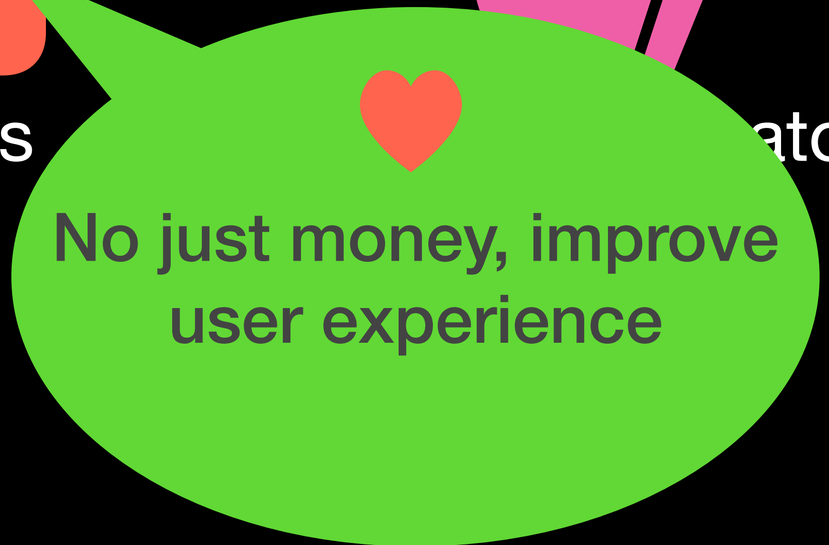
Content creator

Several aspects to consider.

Users

Business

ator



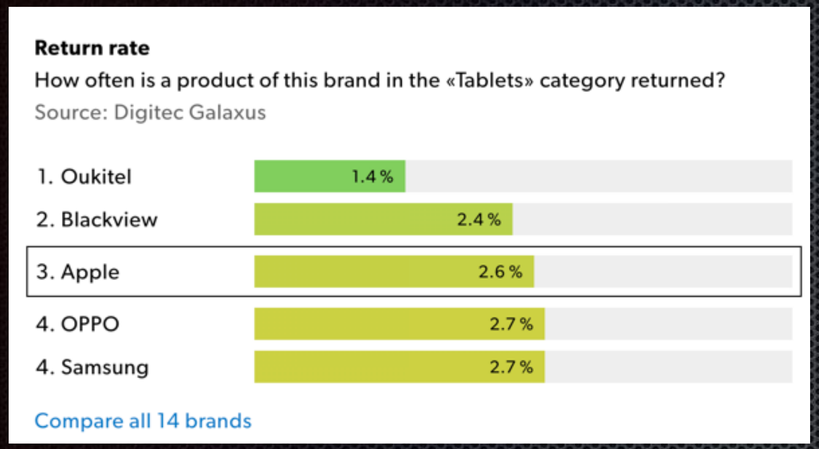
Expect to be asked to create recommendations for

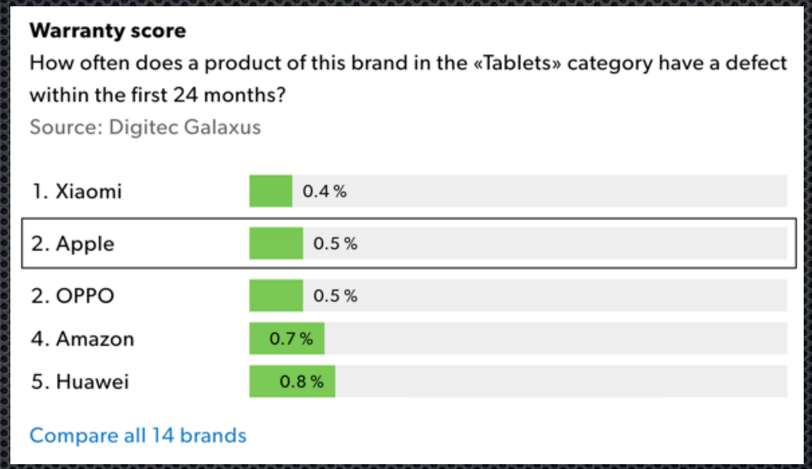
- Non logged in user.
- Returning customer
- session based
- Contextual
  - No click personalisation

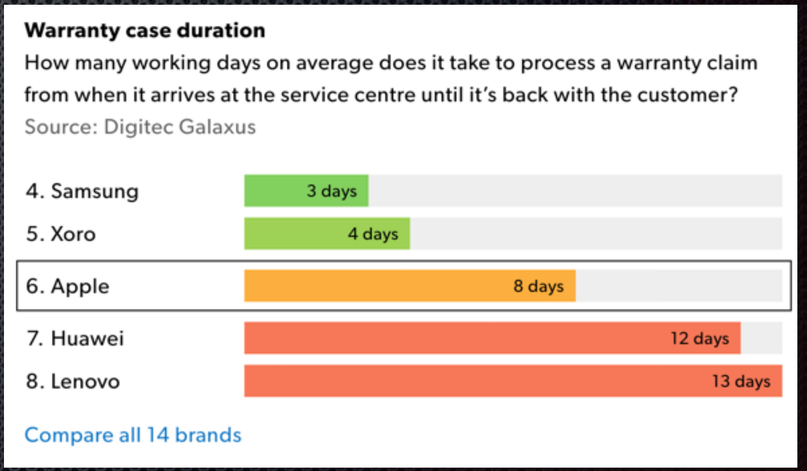
But does it make sense for the business

* Is the model fast enough
* Is it able to recommend the whole catalogue
* Is how much time and work does it cost to maintain it
* Is it any better than a popularity based model

After purchase





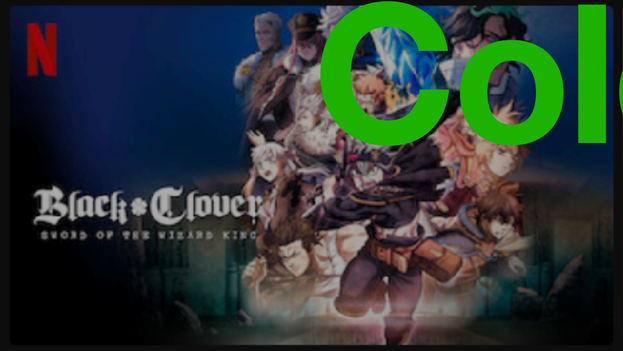


New on Netflix

(Ways not to solve)

Cold start problems

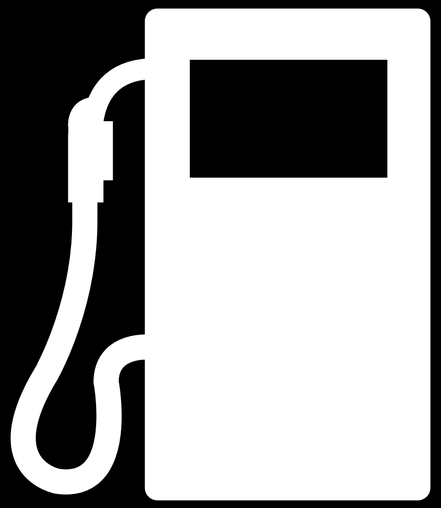
Worth the Wait



Coming this week

Let the recommender explore,
this will lower the KPIs short term,
but hopefully not long term

Don’t scare the customer away.



Data

Spend time on the meta data

Make sure you (meta)data is correct

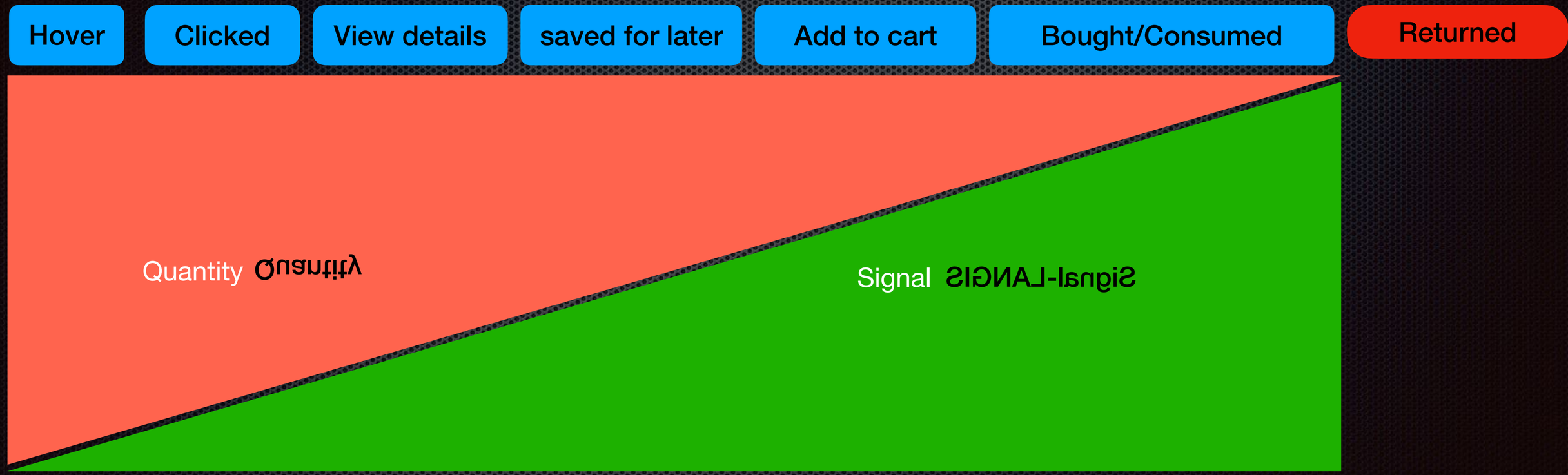
Make sure you (meta)data c

Try to avoid scenarios where your model correctly understands that a users prefers blue and then recommends red clothes because products are classified wrong.

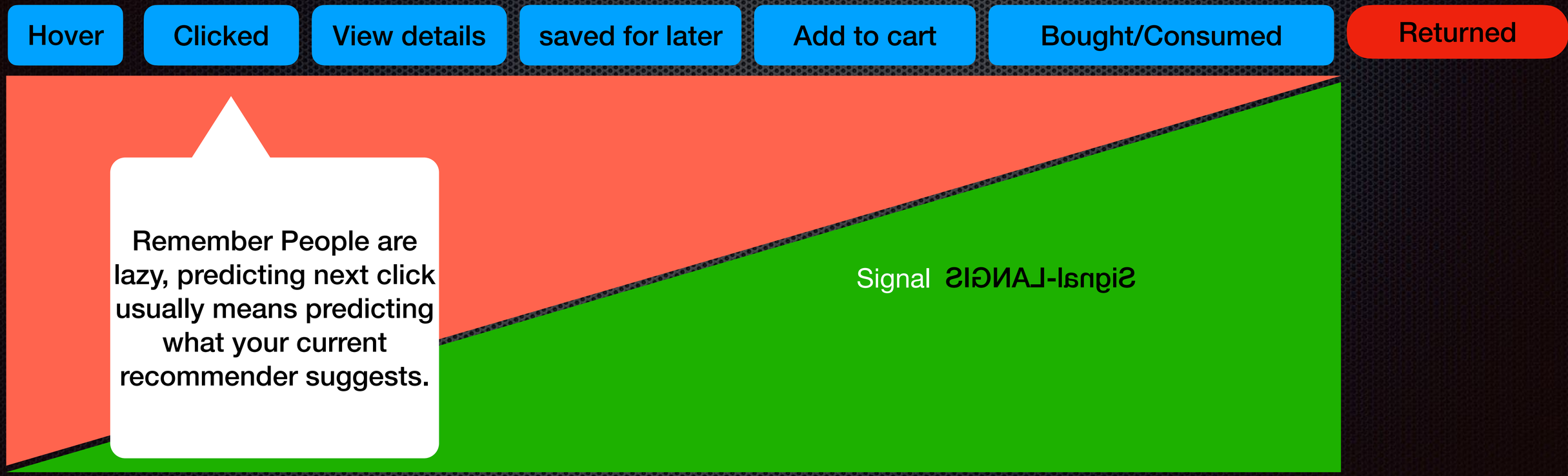
Its hard to detect faulty classifications

- Sampling the data and check for correctness.
- Compare different types of embeddings
  - Visual vs Text embeddings.
  - Text vs Behavioural embeddings.

Data amount vs signal



Data amount vs signal

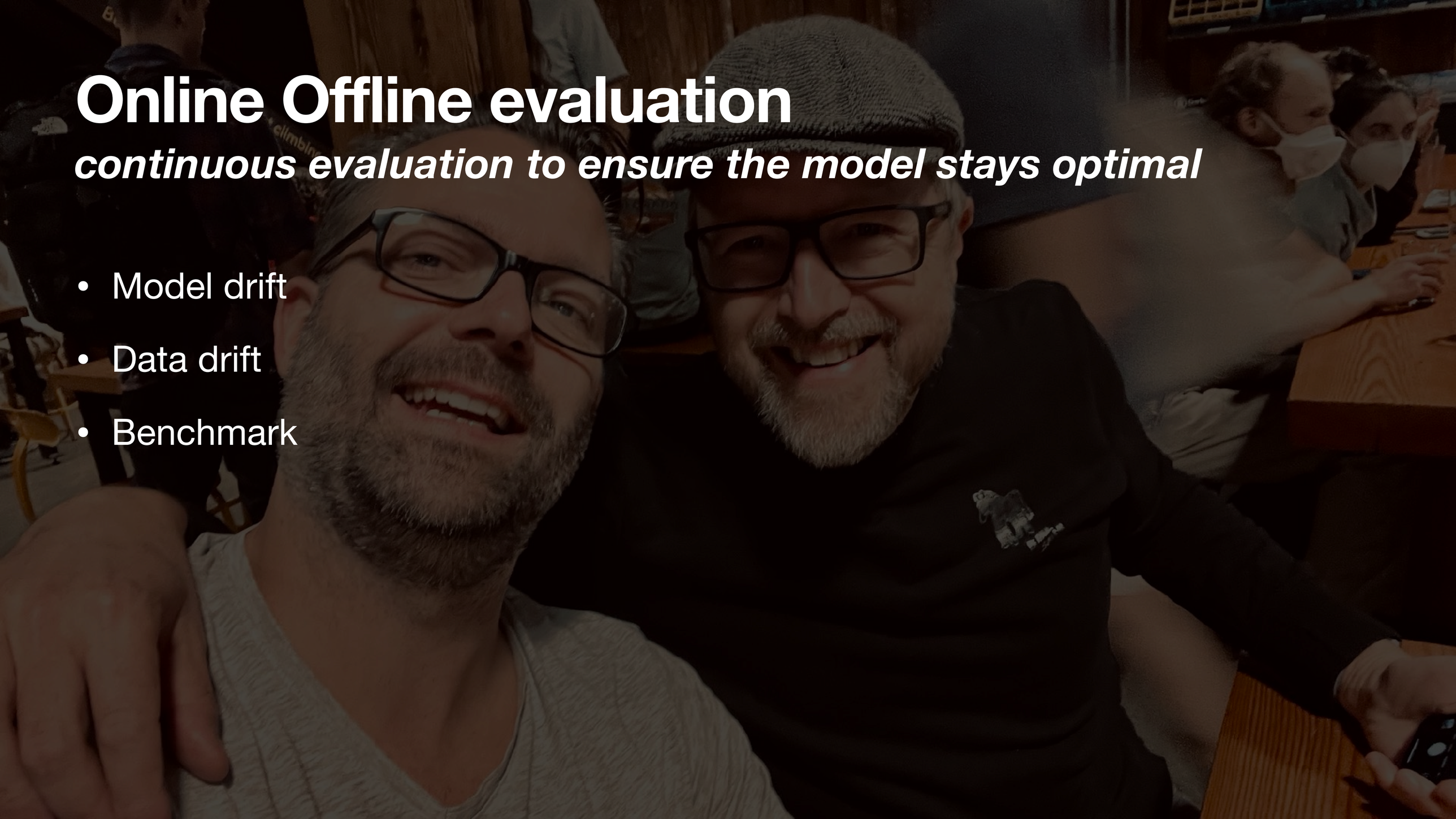


Data collection?

Bandit

Organic

RecSysOps



Online Offline evaluation

continuous evaluation to ensure the model stays optimal

- Model drift
- Data drift
- Benchmark


Questions ?