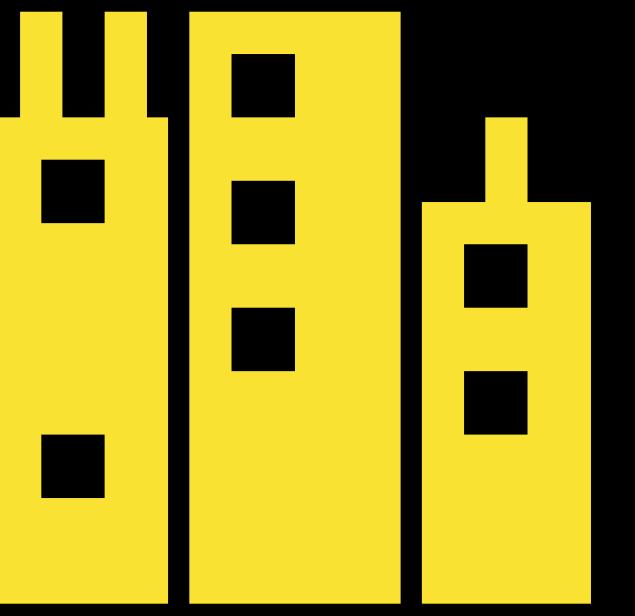


Recommenders in the Wild

Part I

Tutorial RecSys 2023.



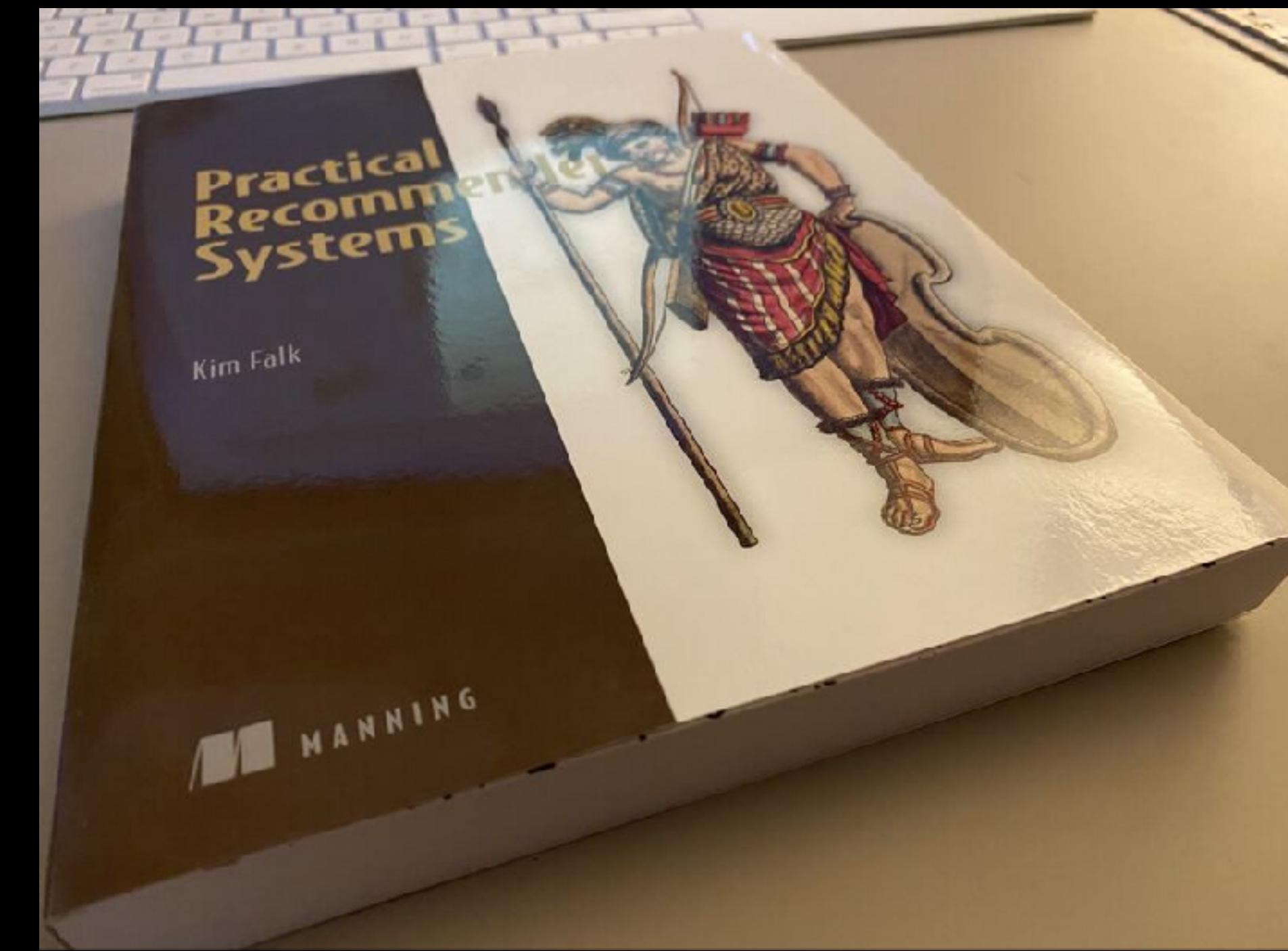
Kim Falk, September 18th 2023

Kim Falk

Computer scientist since 2003

- 💻 Worked with big data and ML and recommenders since 2010
- 🏗 Author of Practical recommender systems
- fx 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

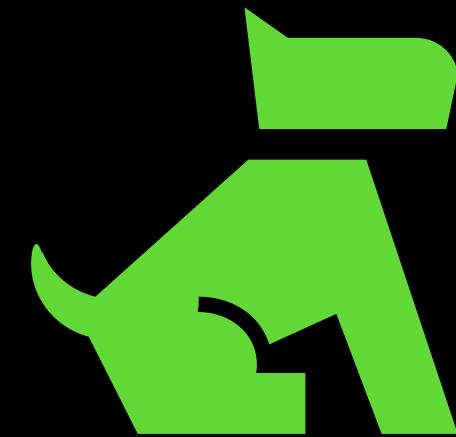
I also assume that you know what a recommender system is



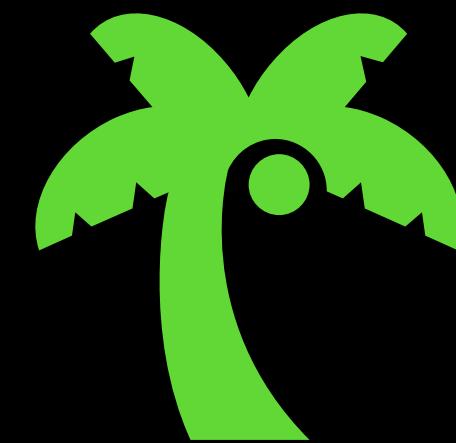
The goal is to share **my personal experience** of working with recommender systems, trying to add new perspectives which are not commonly discussed at conferences.

Agenda

What is a
recommender in the
wild



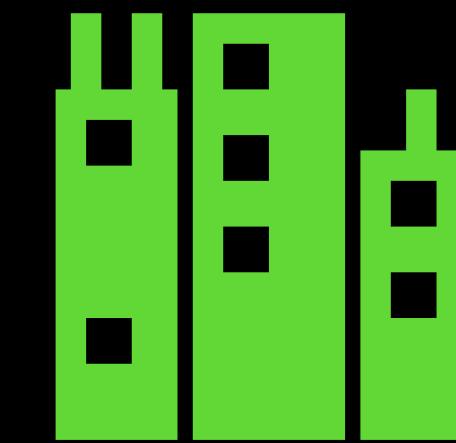
Developing the
recommender



Beyond accuracy



Business
considerations

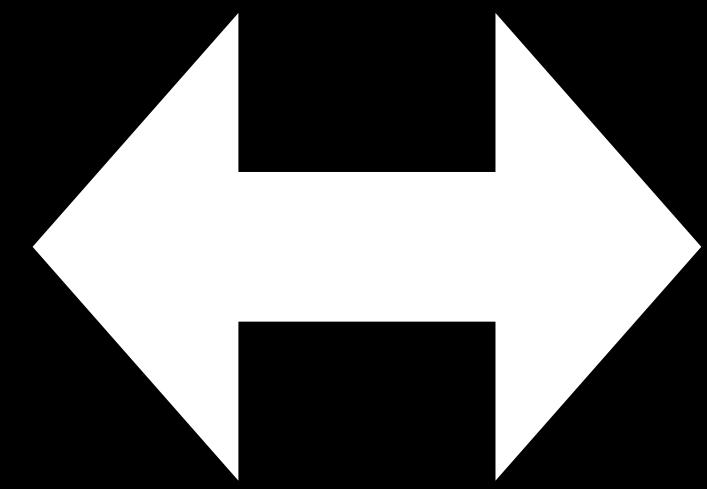


A close-up photograph of a squirrel's head and shoulders. The squirrel has a light brown or greyish-brown coat with darker stripes on its ears and back. It is looking directly at the camera with large, dark eyes. The background is filled with a dense carpet of fallen pink petals, likely from a cherry blossom tree, creating a soft, textured backdrop.

**What is
a Recommender in
the wild?**

Whats a recommender in the wild

**Recommender
in the wild**



**Production
system servicing
recommendations
to people**

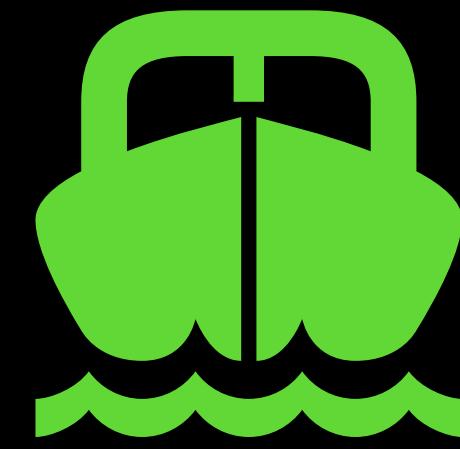
Recommenders
are domain specific

		Recall@10	0.2038	0.2273	0.2354	+15.5%	+3.56%
Movielens-1M	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	-	-	

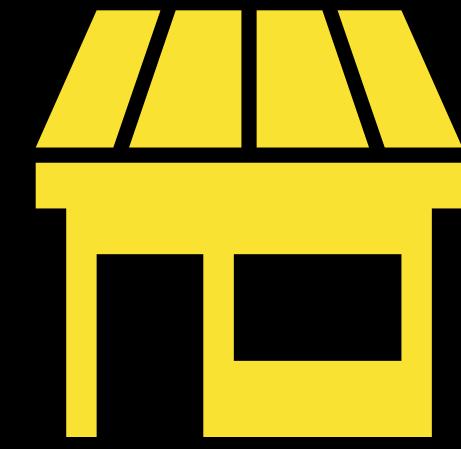
Recommenders
are domain specific

Why ?

Mostly due to



Streaming data



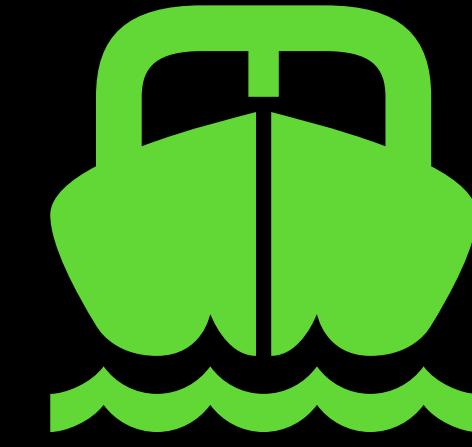
E-commerce



Other

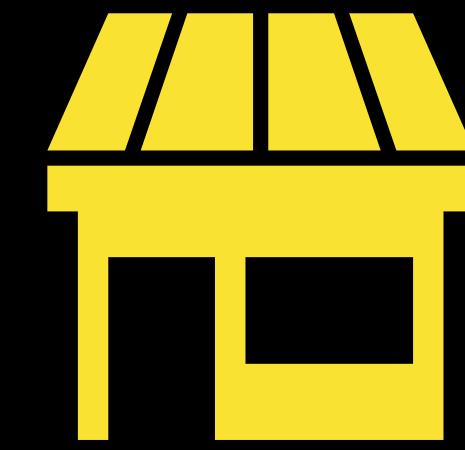
Why ?

Mostly due to



Streaming data

Video on demand | Music | Social Media

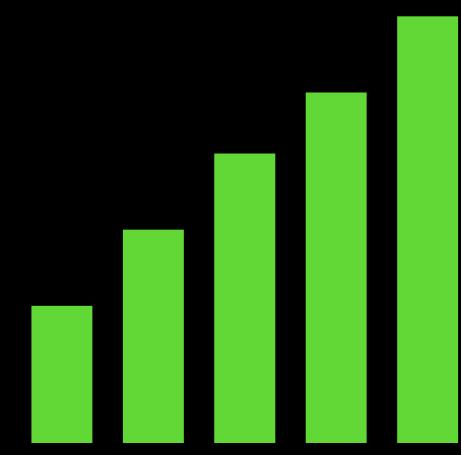


E-commerce

Online stores

Why ?

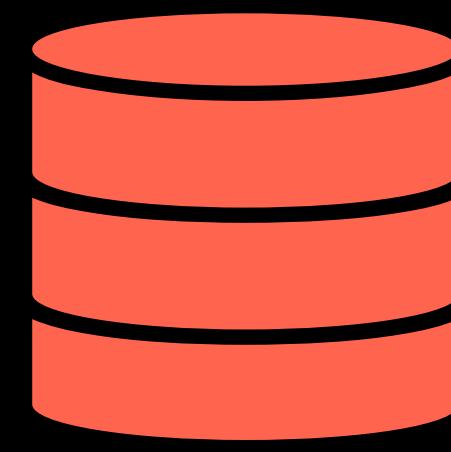
Mostly due to



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-commerce



I will focus on the (mostly)
E-commerce

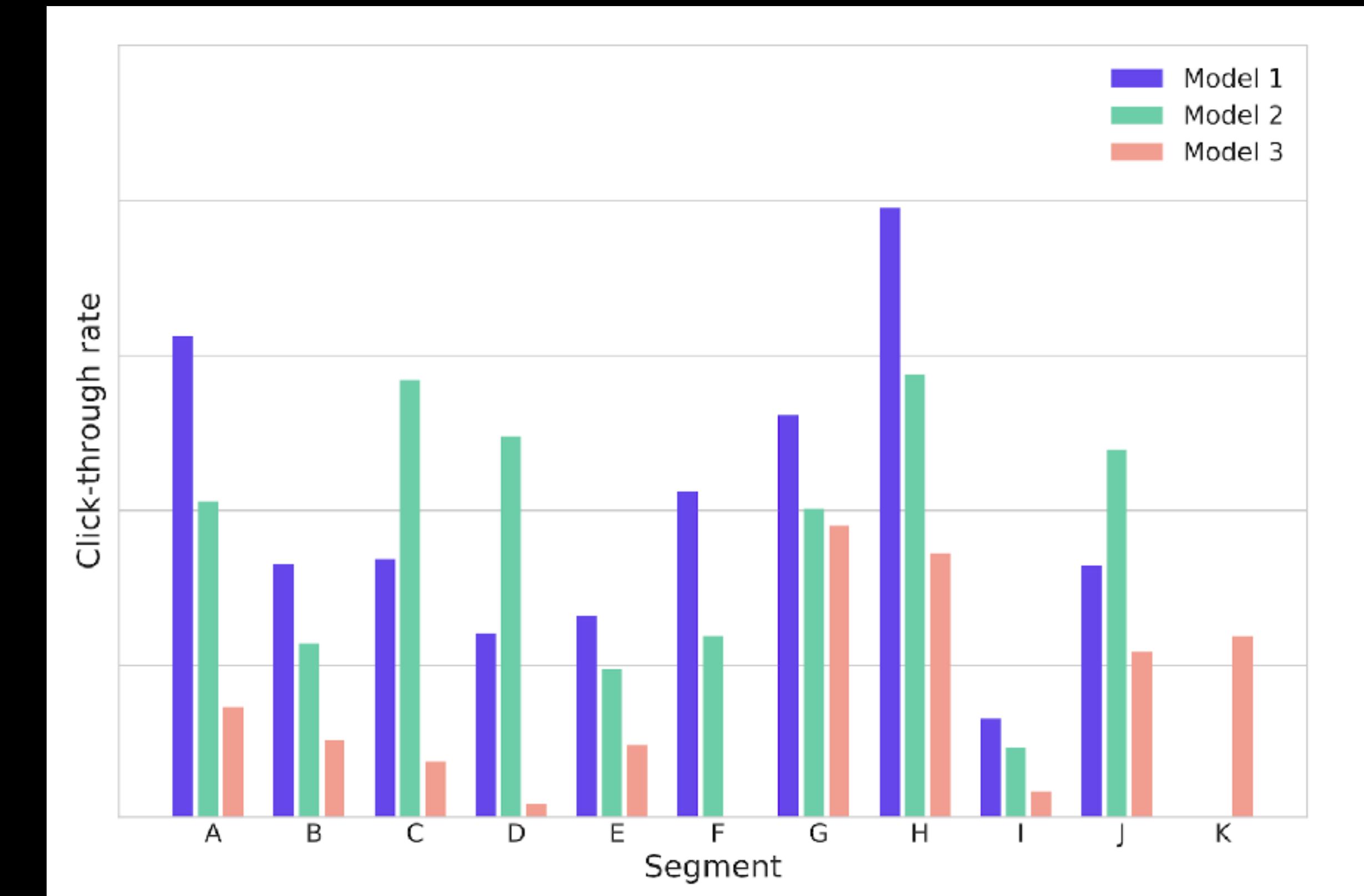
I will focus on
the E-commerce
experience

I will focus on the (mostly)
E-commerce

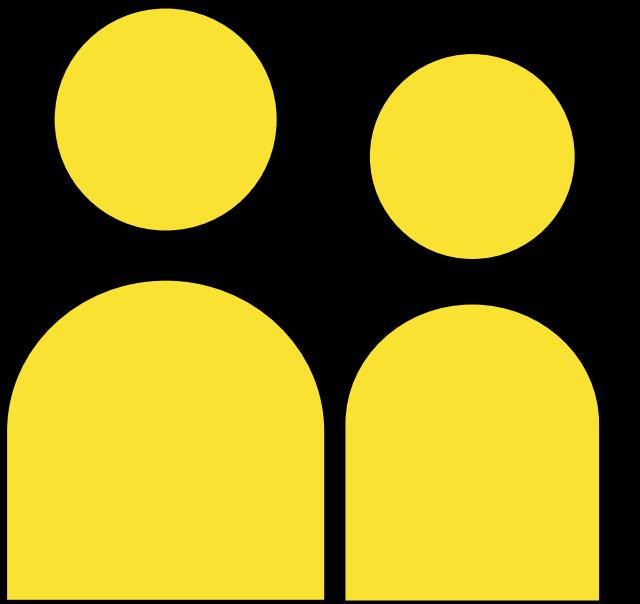
But you can
probably use the same
considerations in
scenarios

**Recommender
systems are even
sub-domain specific**

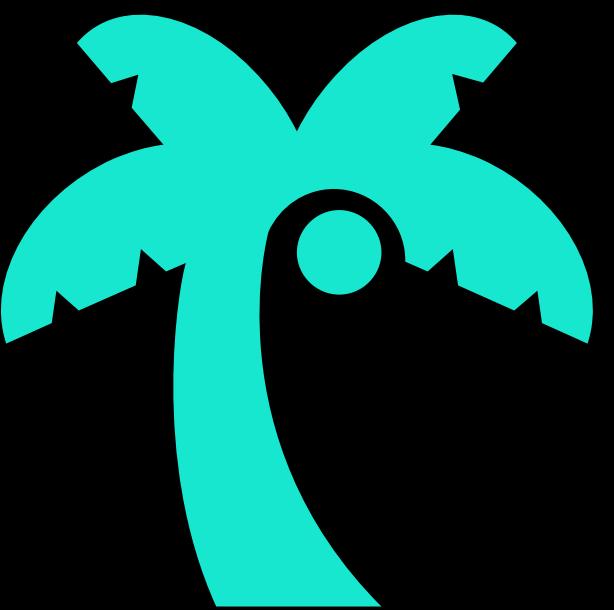
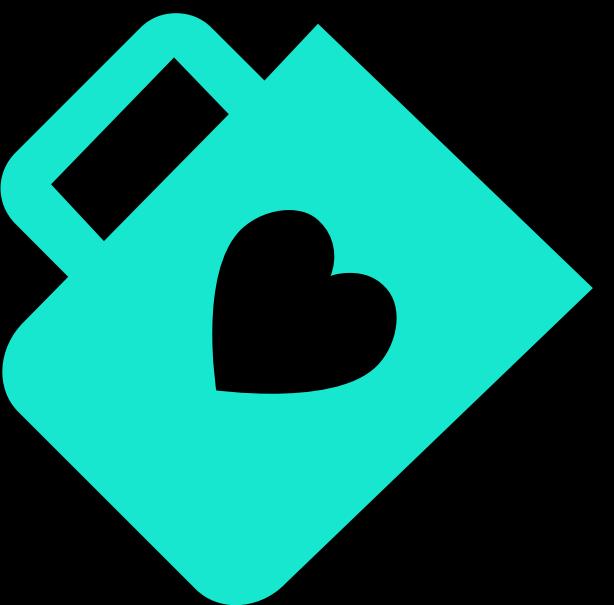
Different domains has different signals



Find content for
users



Users



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

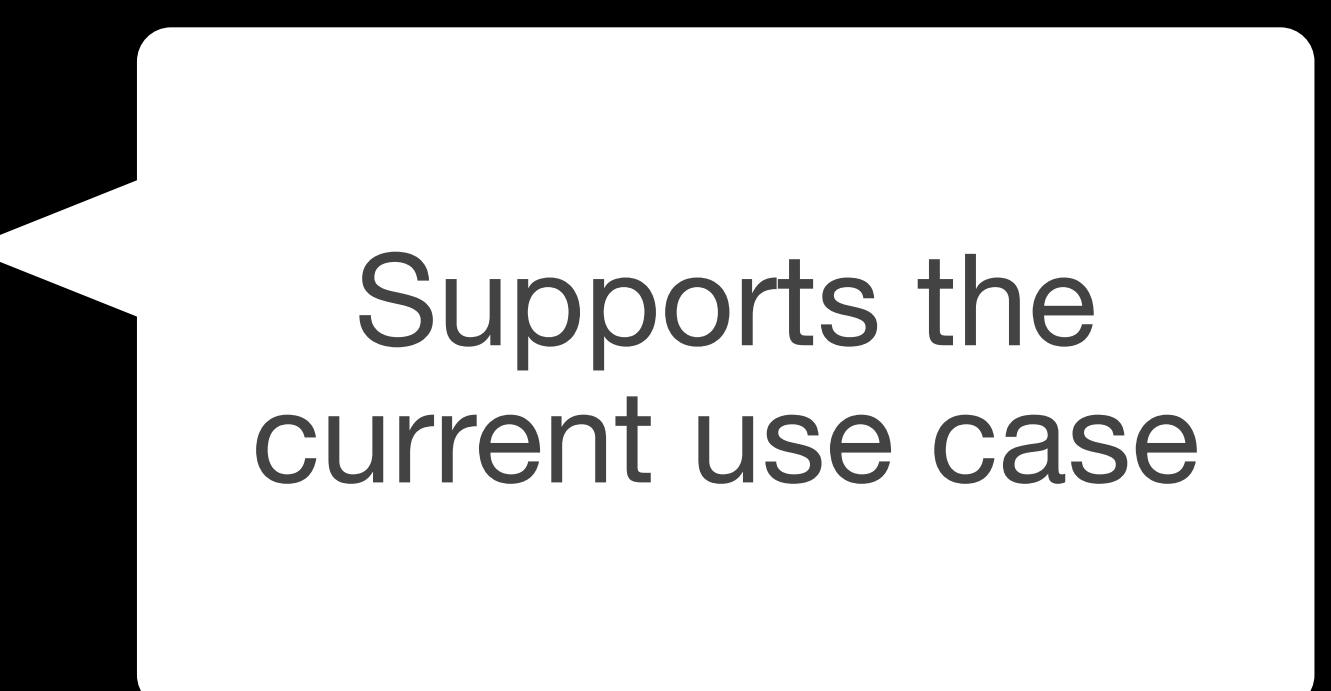
Rerank lists of
content

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

Is there one optimal
rank of content?

Find Content for Users

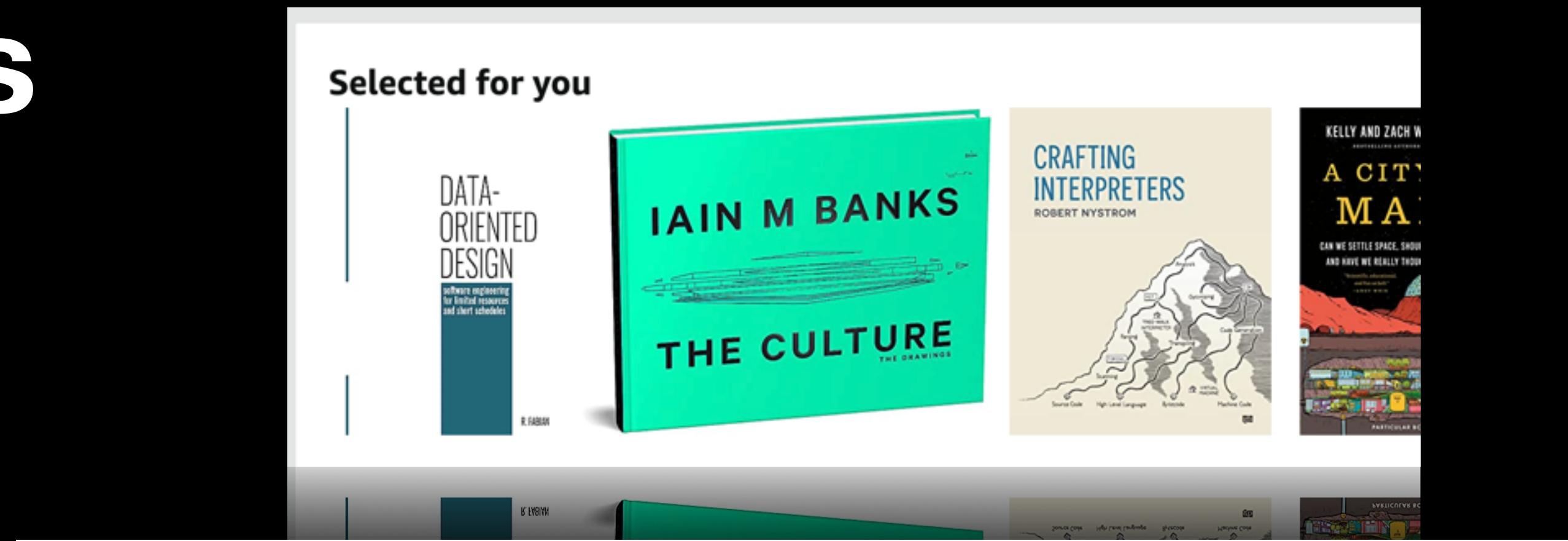
- What they need now
- What they didn't know that they needed, but did
- Educate them on what else the catalogue contains.



Supports the current use case

Find Content for Users

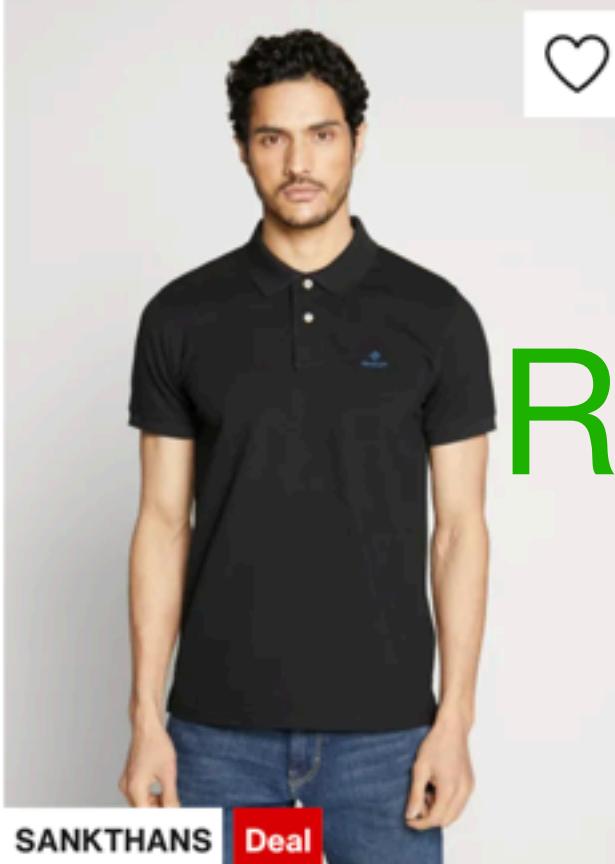
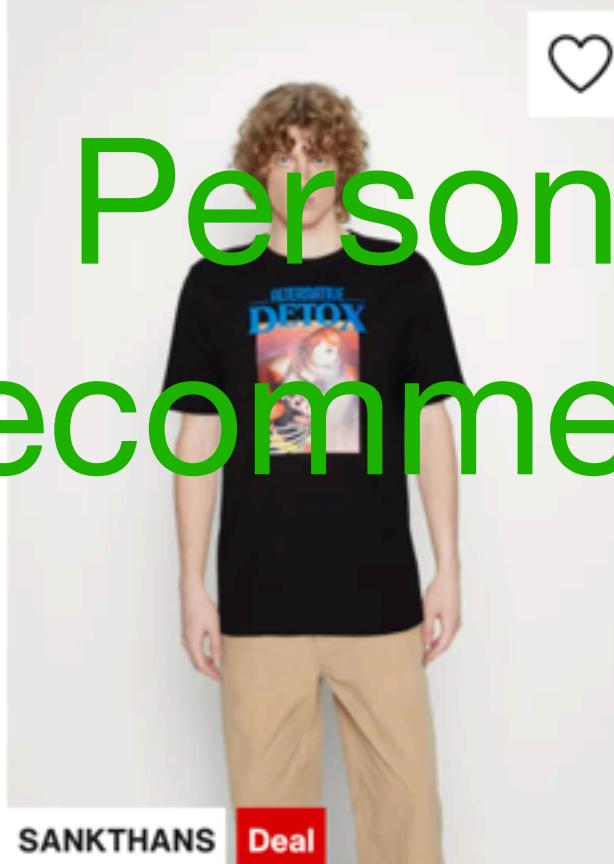
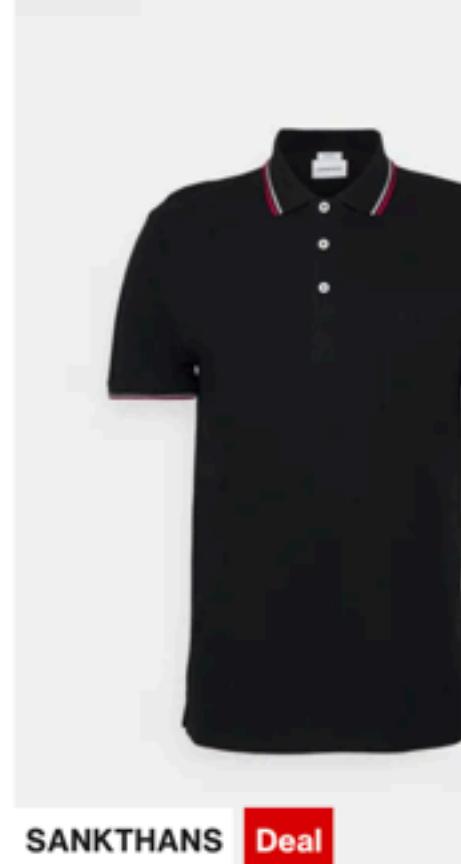
- What they need now
- What they didn't know that they needed, but did
- Educate them on what else the catalogue contains.
- Supports the current use case



Tilbud udvalgt specielt til dig
Vores bedste tilbud

Oplev

Personalised Recommendations

 SANKTHANS Deal GANT CONTRAST COLLAR SS RUGGE... fra 419,00 kr Oprindeligt: 599,00 kr Spar op til -30% 1K 00,00 kr	 SANKTHANS Deal Diesel JUST UNISEX - T-shirts print - bla... 300,00 kr Oprindeligt: 599,00 kr -50% 1K 00,00 kr	 SANKTHANS Deal Diesel T-JUST-POCKET-CROW UNISEX ... 350,00 kr Oprindeligt: 699,00 kr -50% 1K 00,00 kr	 SANKTHANS Deal Lindbergh CONTRAST PIPING - Polos 169,00 kr Oprindeligt: 199,00 kr -15% 1K 00,00 kr
--	--	--	---

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

Douglas Adams's Starship Titanic: From the minds Behind The Hitchhiker's Guide to the Galaxy and Monty Python Kindle Edition

by Terry Jones (Author), Douglas Adams (Author) | Format: Kindle Edition

4.2 ★★★★☆ 92 ratings

#1 Best Seller in Science Fiction Space Operas

See all formats and editions

Kindle Edition £0.99 Paperback £8.31

Read with Our Free App 6 Used from £8.00
15 New from £6.70

From the minds of Douglas Adams (*The Hitchhiker's Guide to the Galaxy*) and Terry Jones (*Monty Python*) comes *Starship Titanic*, the hilarious novelization of the third-best adventure game of 1999.

Welcome on board the *Starship Titanic*.
The Ship that Cannot Possibly Go Wrong.

Read more

Print length 1. 218 pages English On Kindle Scribe

Language Sticky notes

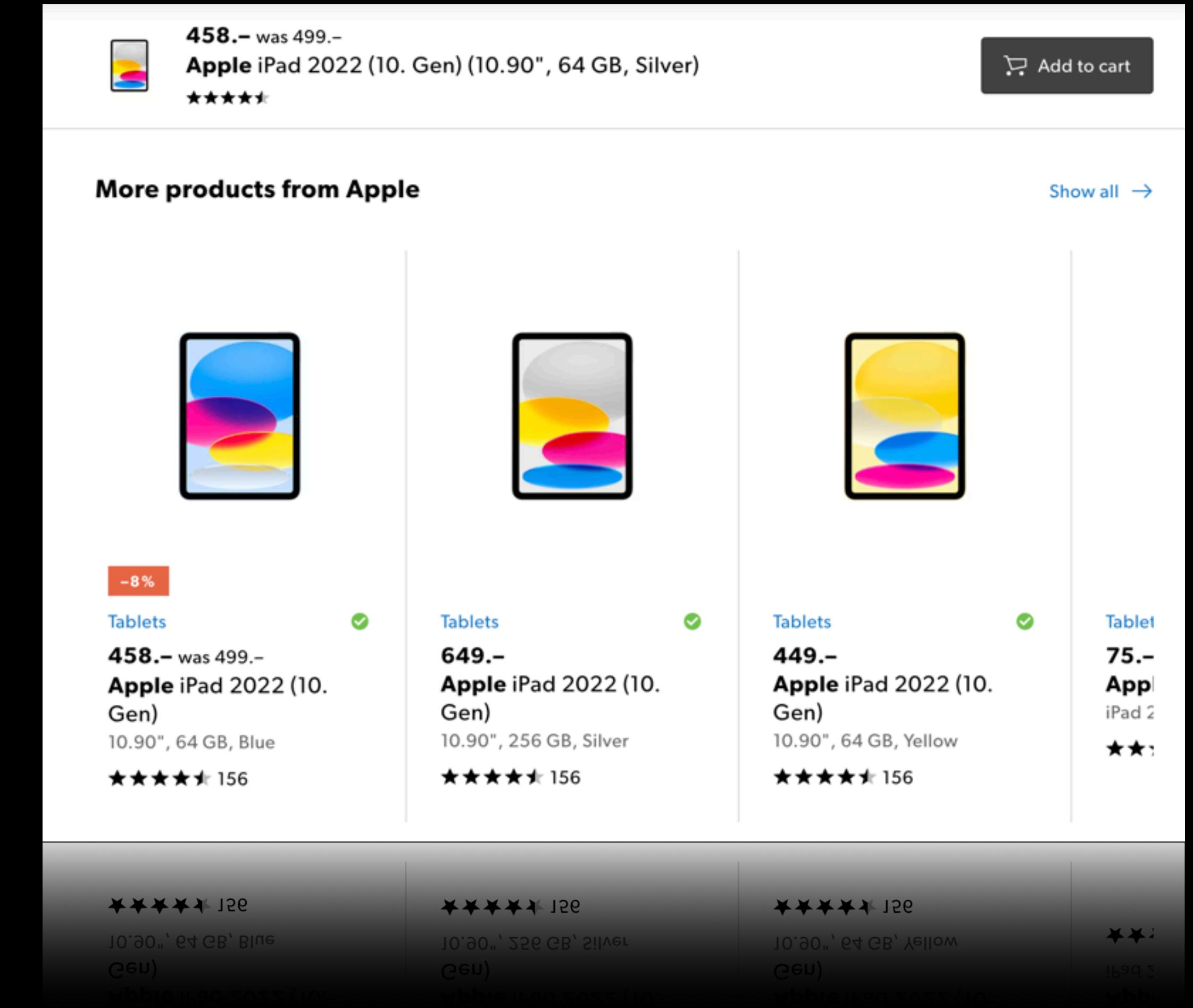
Publisher Pan

Publication date 18 May 2023

Reading age 18 years and up

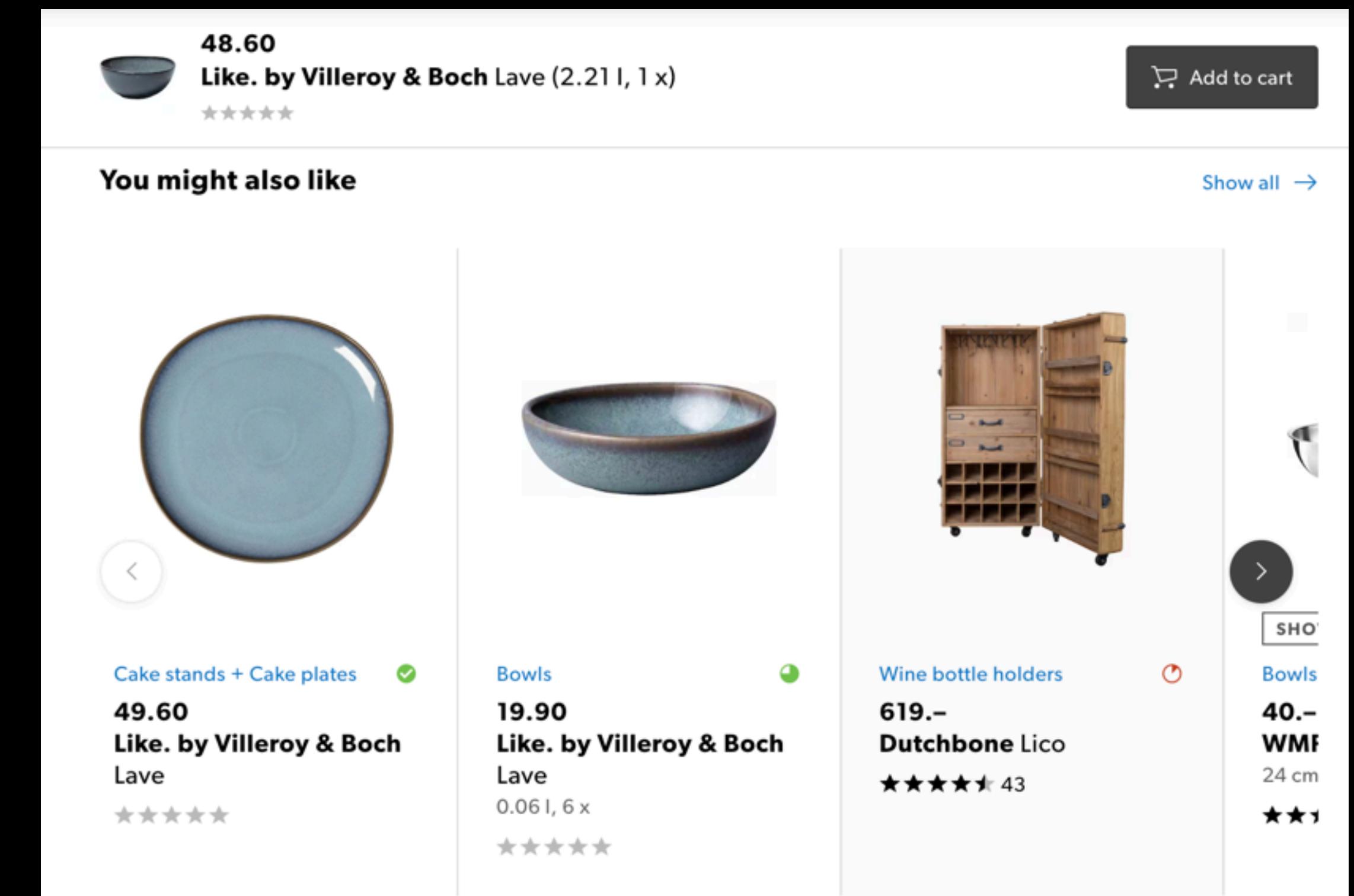
Find Content for Users

- 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



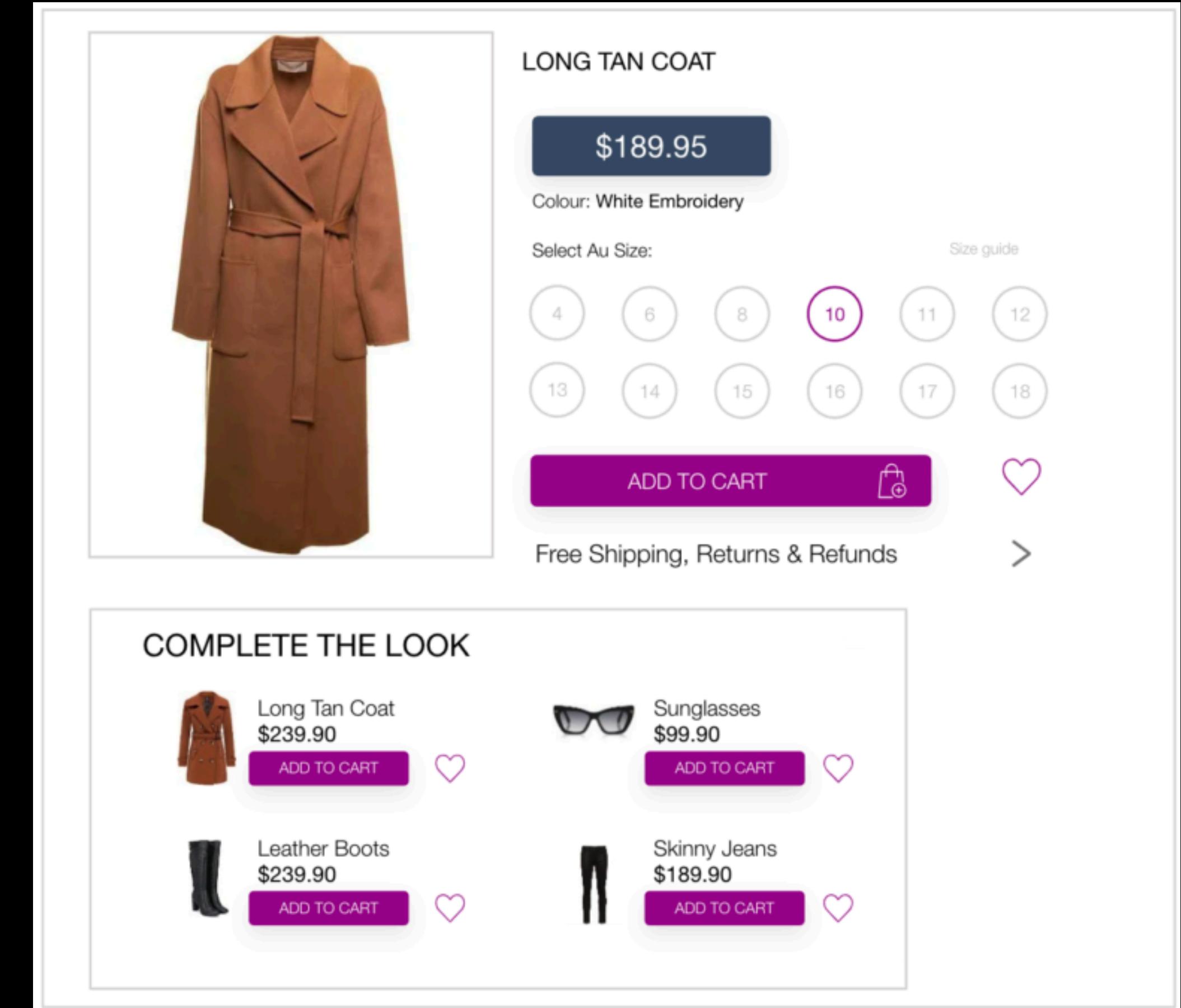
Find Content for Users

- 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



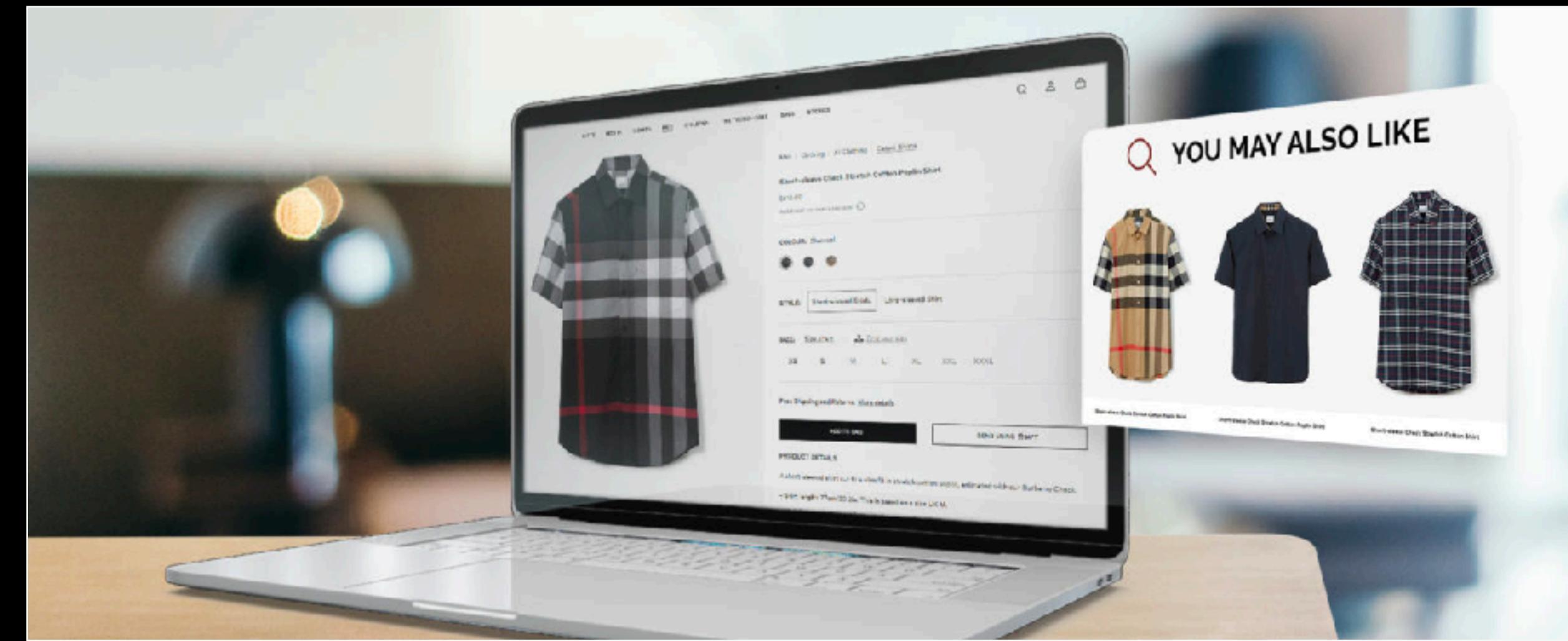
Find Content for Users

- 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



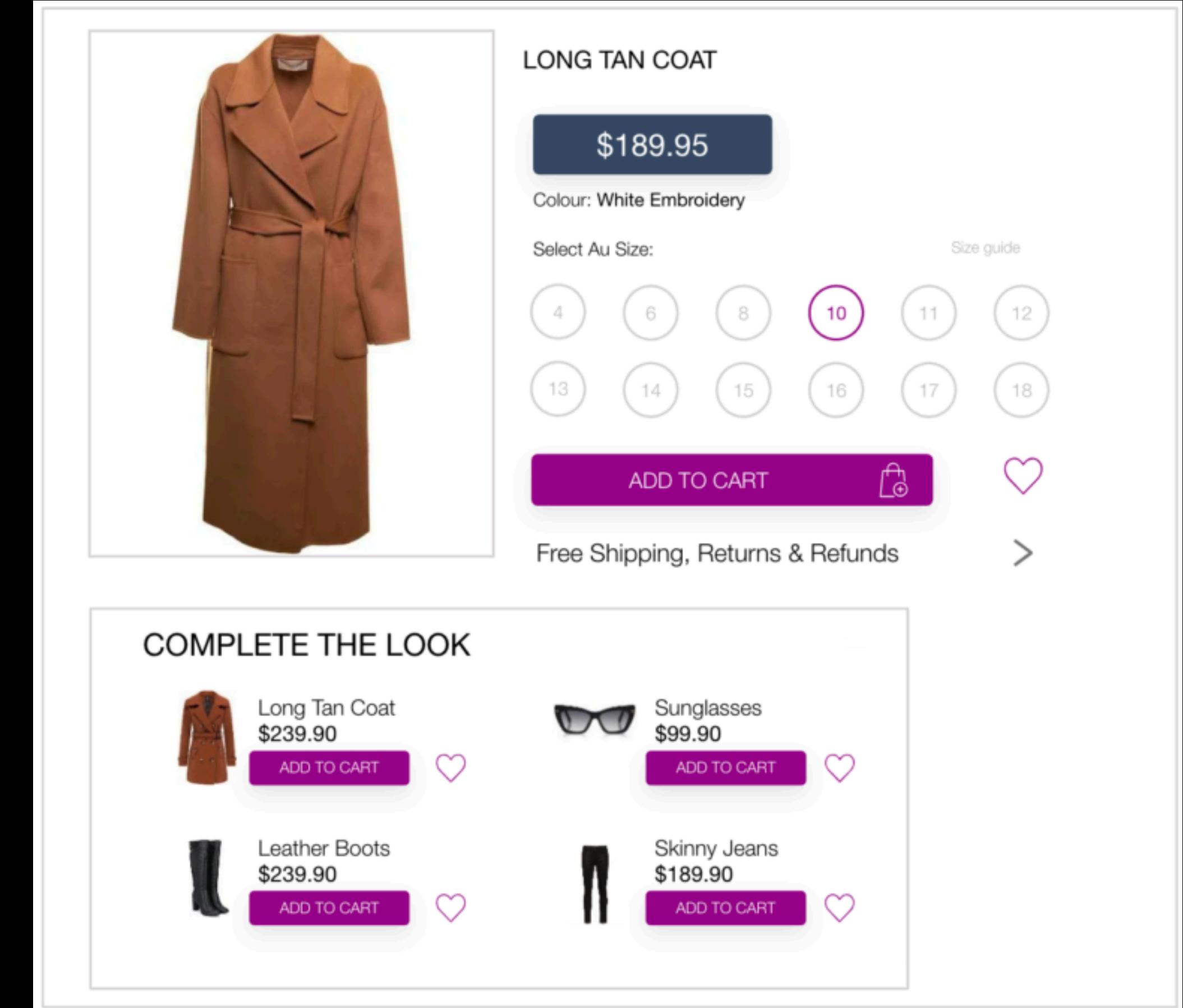
Find Content for Users

- 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**

The screenshot shows a product page for a Senston Full Carbon Fiber Tennis Racket. The main image displays the racket and its matching dark blue protective cover. To the left, there's a vertical menu with icons for search, categories, and filters. The product title is "Senston Full Carbon Fiber Tennis Racket 27, Lightweight Stability Adult Tennis Racquet Set with Bag + 1 Overgrip + 1 Vibration Dampeners". It has a rating of 3.6 stars from 14 reviews. The price is £33.99, and the color is Orange. Technical specifications include One Size, Senston brand, 4 1/4 inches grip size, Tennis sport, Carbon Fibre material, Beginner/Intermediate skill level, and Carbon Fiber frame material. A "See more" link is available. Below the main image, there's a section titled "About this item" with a bulleted list of features like full carbon material and weight range. A "Report incorrect product information" link is also present. At the bottom, there's a "Buy it with" section showing a racket and ball tube, with a total price of £41.48 and a "Add both to Basket" button. A note indicates one item is dispatched sooner than the other.

Senston Full Carbon Fiber Tennis Racket 27, Lightweight Stability Adult Tennis Racquet Set with Bag + 1 Overgrip + 1 Vibration Dampeners

Visit the Senston Store
3.6 ★★★★☆ 14 ratings

£33.99

Colour Name: Orange

Size	One Size
Brand	Senston
Grip size	4 1/4 inches
Sport	Tennis
Material	Carbon Fibre
Skill level	Beginner,Intermediate
Frame material	Carbon Fiber

Roll over image to zoom in

About this item:

- Material: Full Carbon / Weight unstrung: 250-260g
- Head size: mid- 96-100 inch² / Grip size: Size 2 (4 - 1/4 inch)
- Tennis racket length: 27 inches(68.5-69cm)/ Power level:900-1200/ Balance:320mm / Frame thickness : 21mm-28mm
- Effortless power without altering the way you swing and play, ideal for universal adult; intermediate and beginner players wanting to improve their skills and progress on the court.
- Package includes 1x tennis racket, 1x premium quality protective carry case, 1x overgrip(random color), 1x vibration damper(random color)

Report incorrect product information.

Buy it with

Total price: £41.48

Add both to Basket

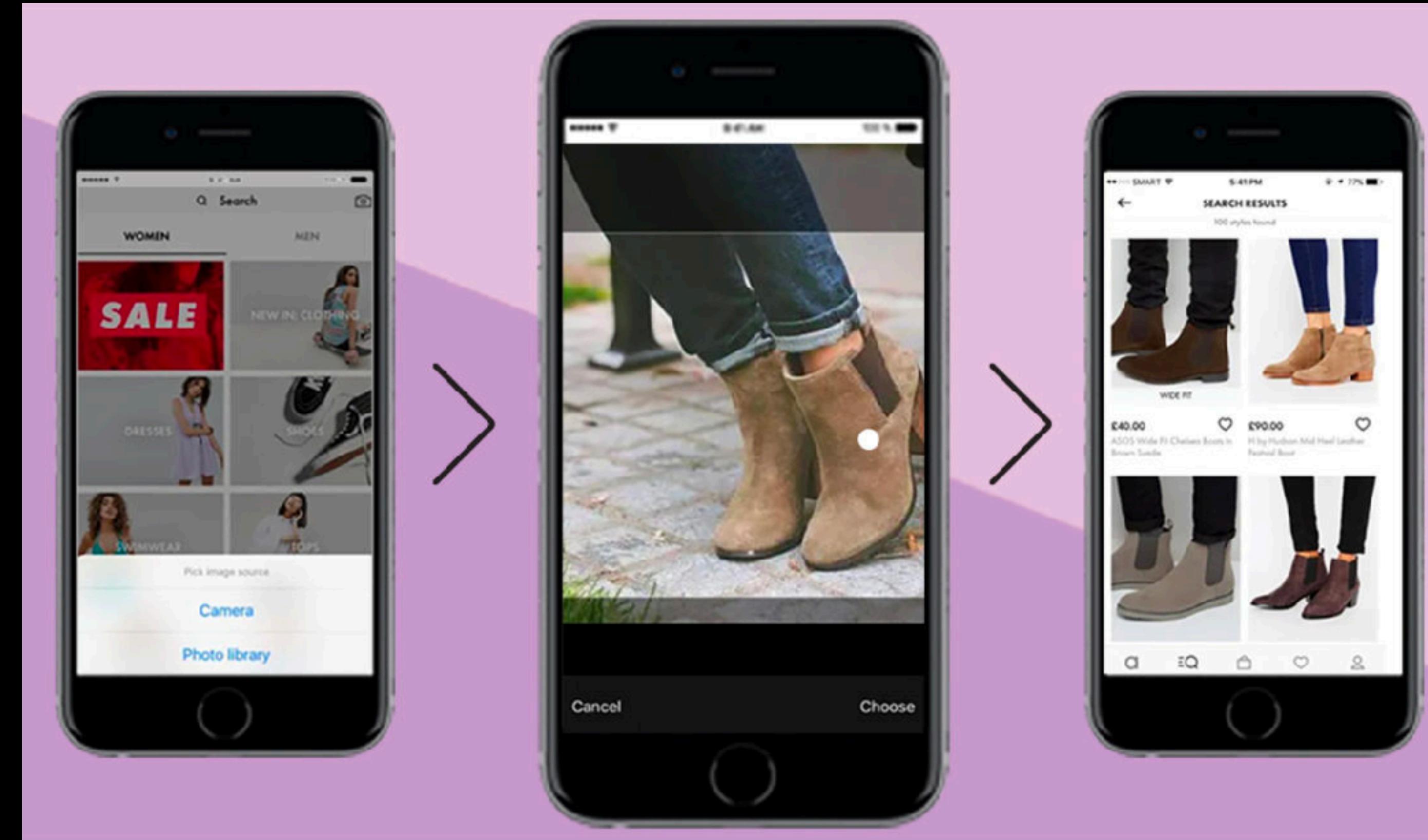
This item: Senston Full Carbon Fiber Tennis Racket 27, Lightweight Stability Adult Tenn... £33.99

Slazenger Championship Tennis Ball - 4 Ball Tube , Pack of 4 £7.49 (£1.87/count)

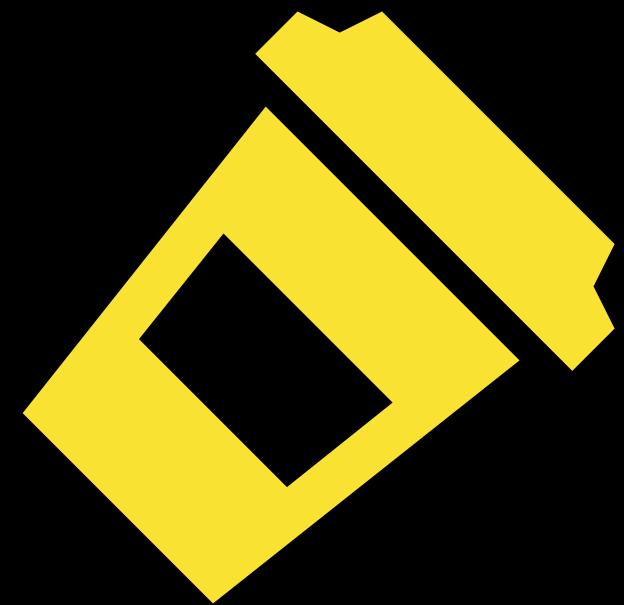
One of these items is dispatched sooner than the other.
Show details

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**



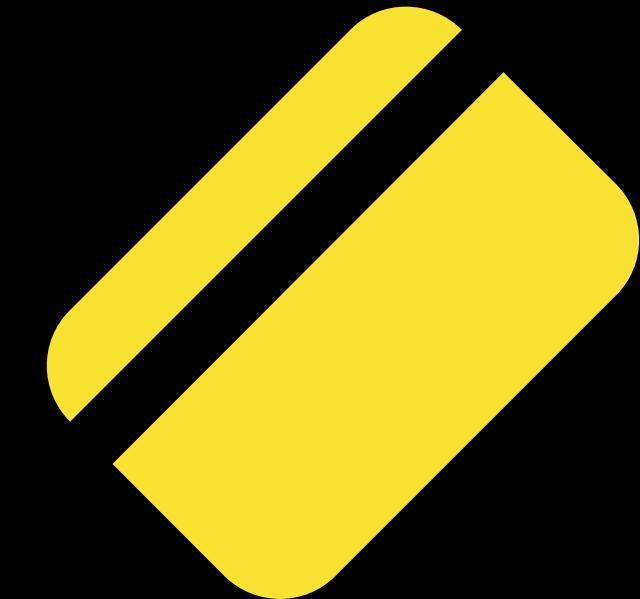
Use cases



Product details
page (PDP)



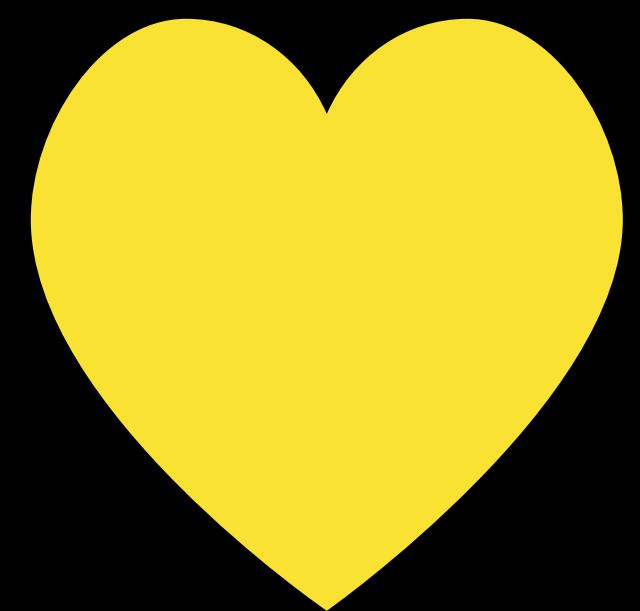
Add to
Cart



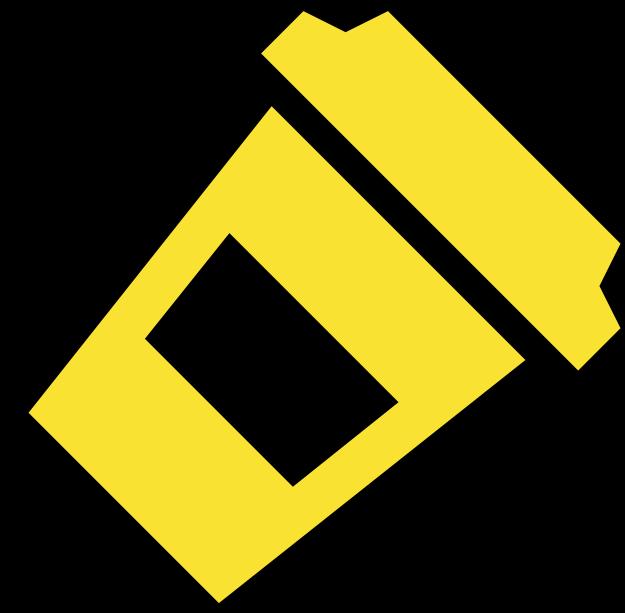
Check
Out



Post
purchase



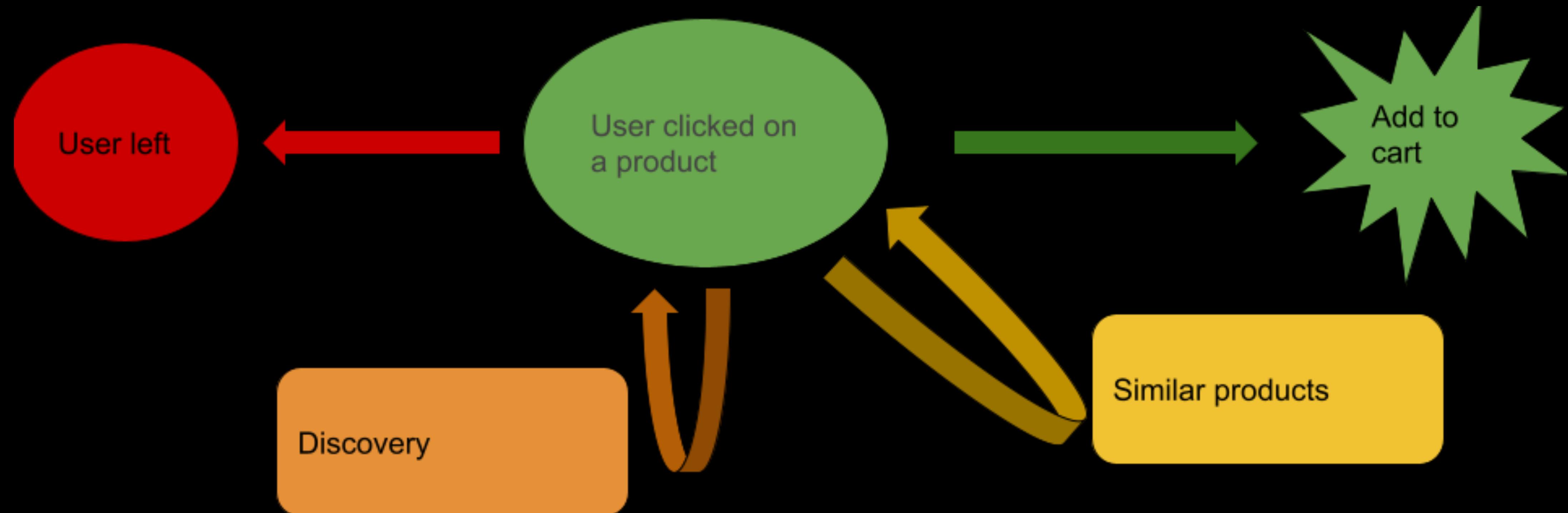
For you



User landed on a product details page

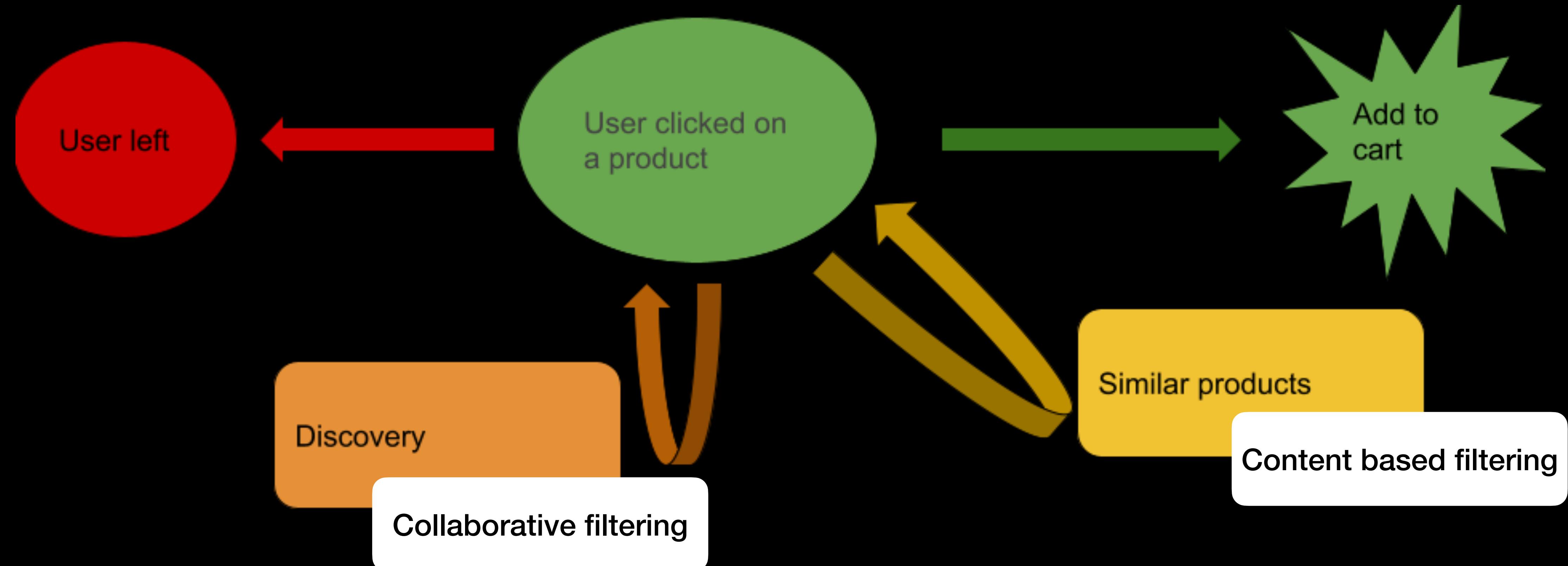
Product Detail Page

What do you want the user to do?



Product Detail Page

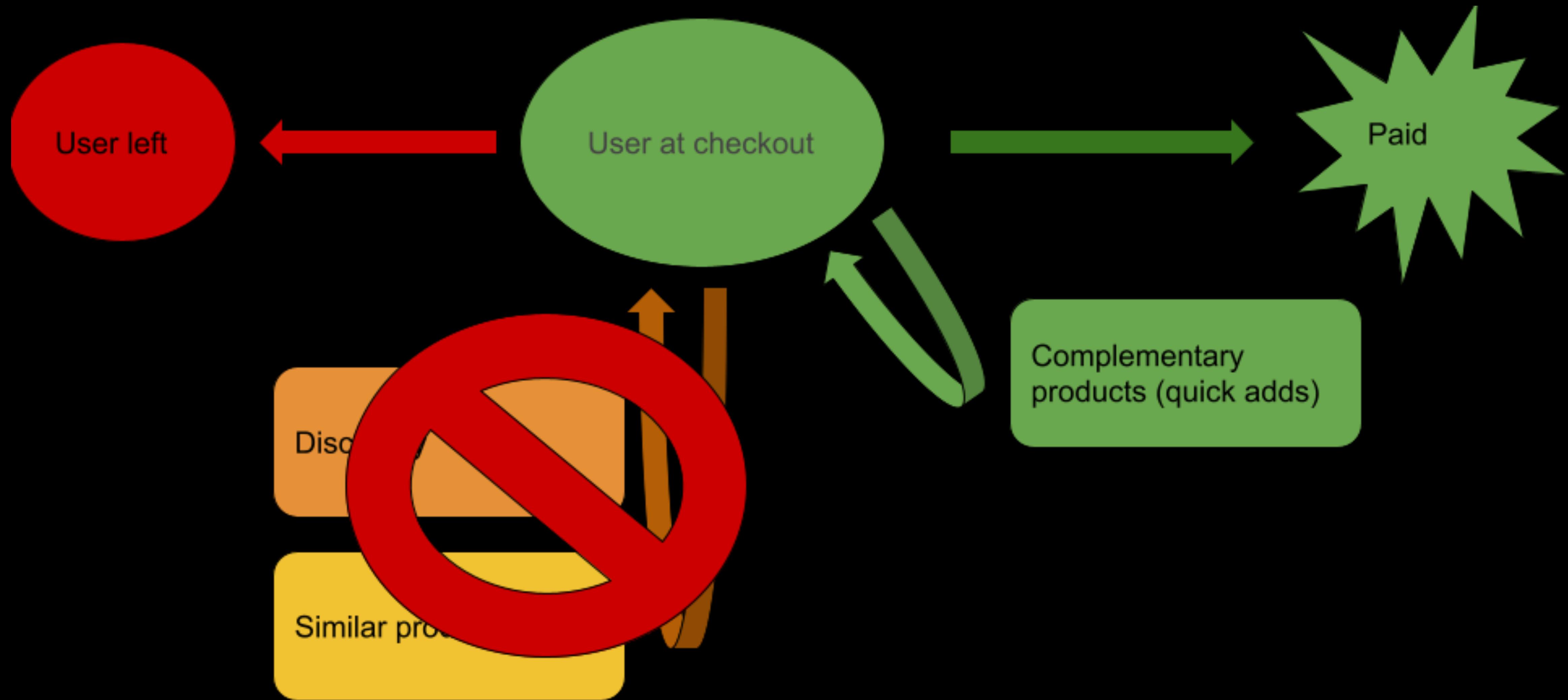
What do you want the user to do?

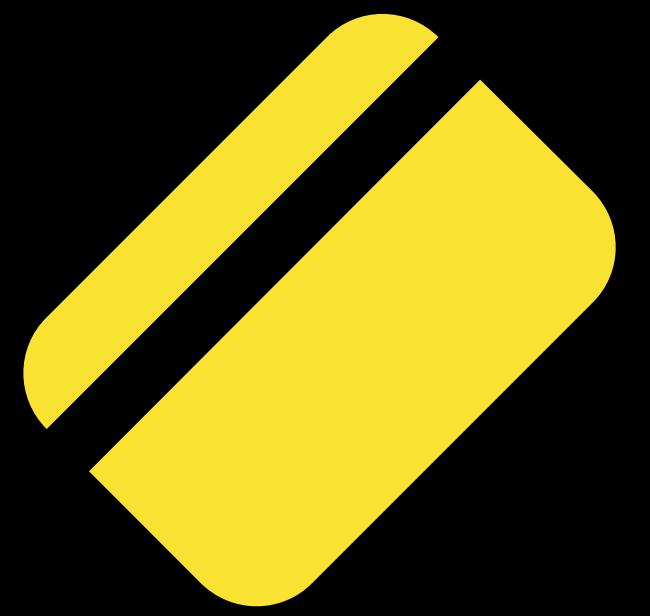




User added something to cart

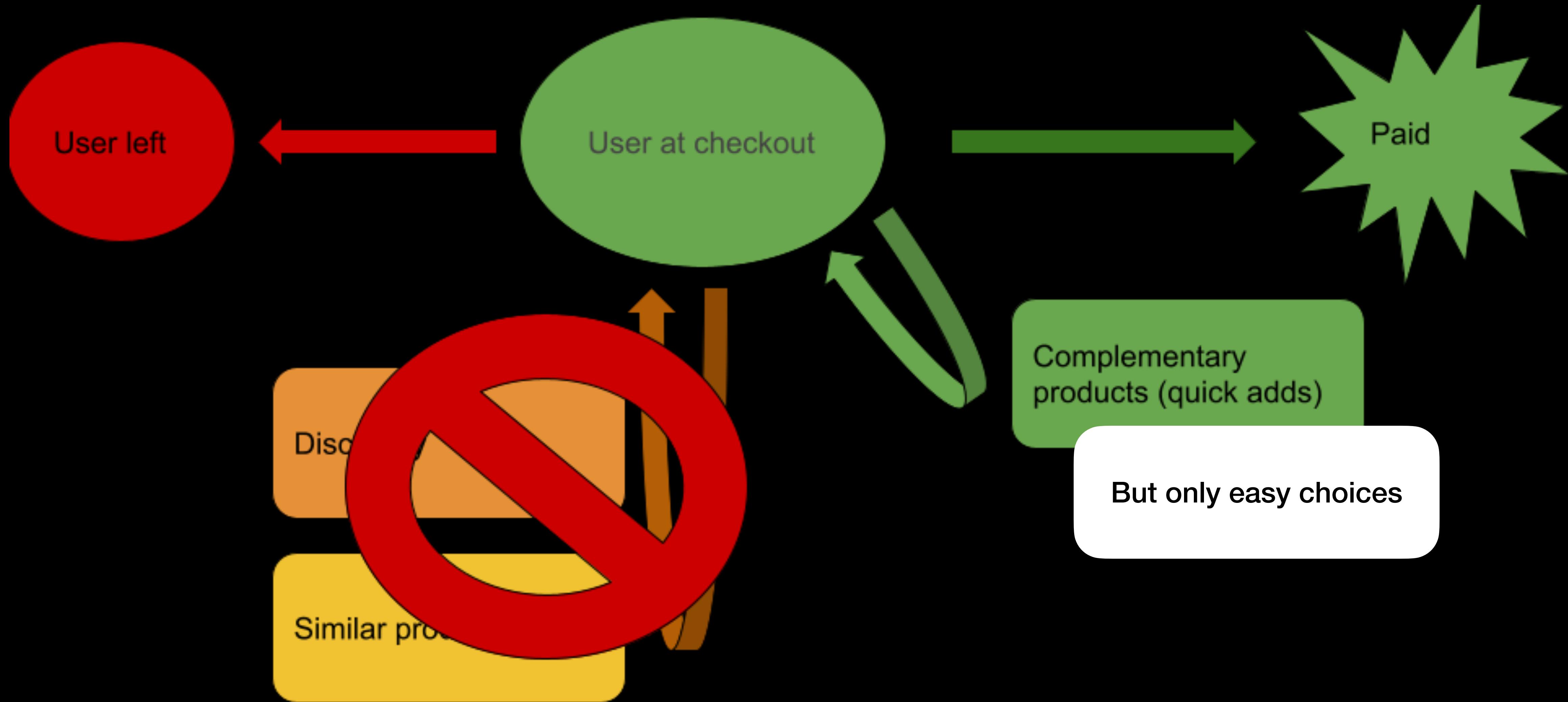
Add to cart recommendations





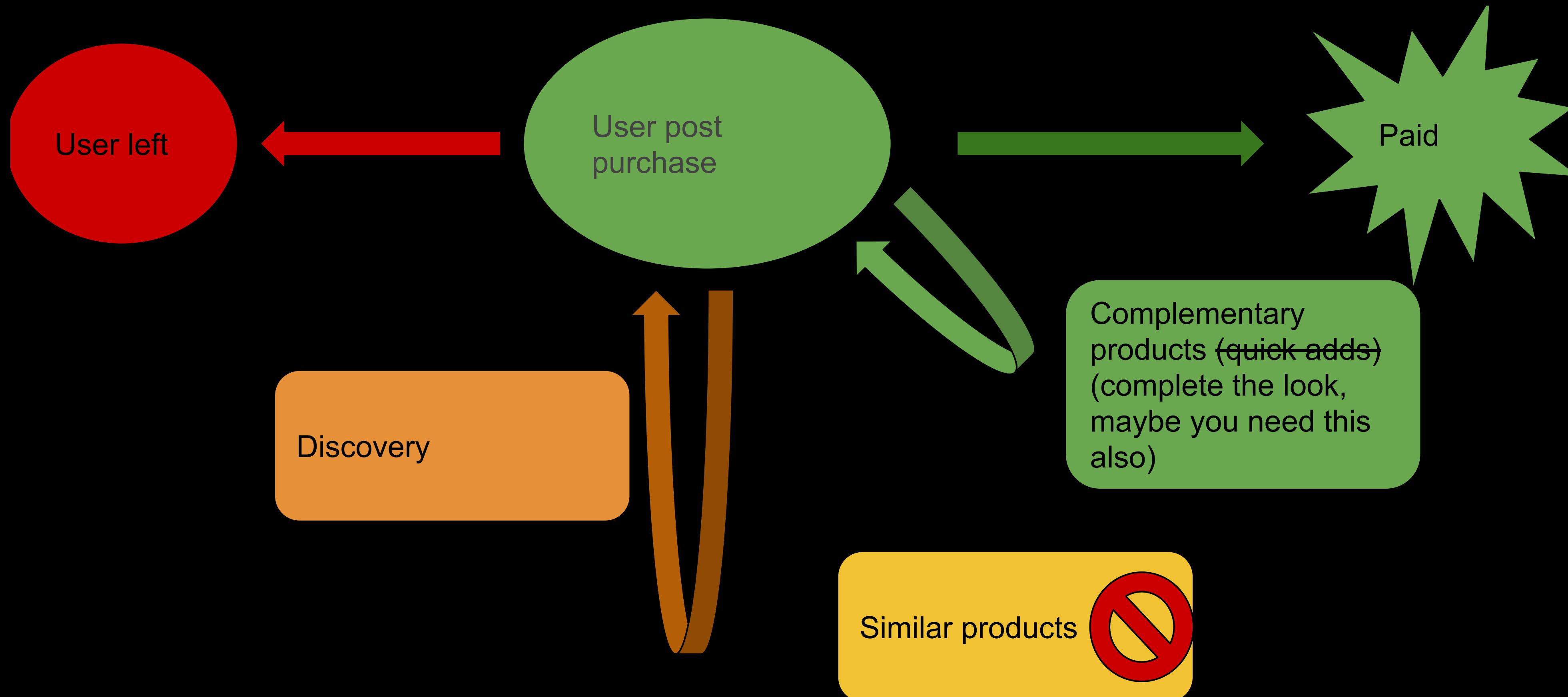
User is about to pay

Check out recommendations

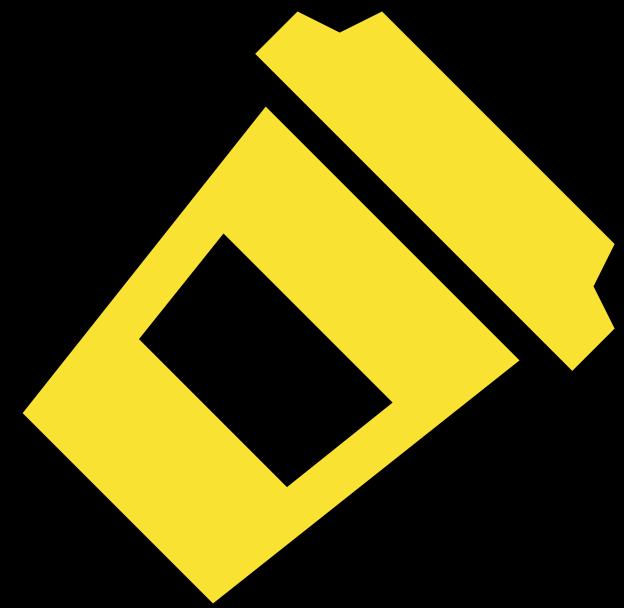




User bought something



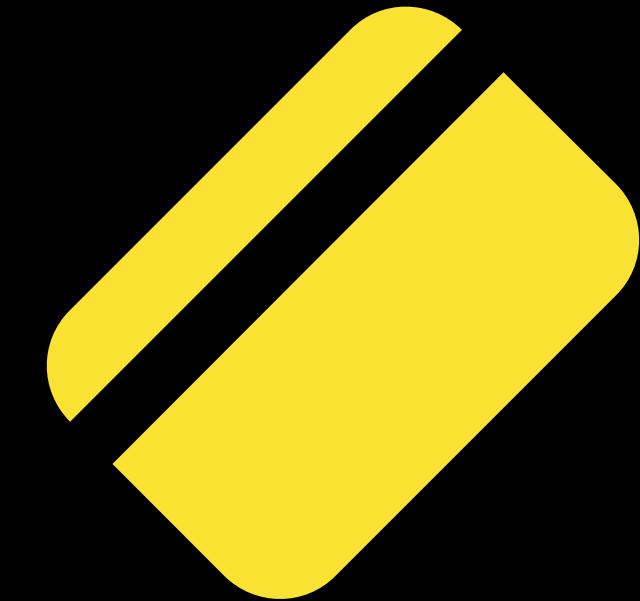
Use cases



Product details
page (PDP)



Add to
Cart



Check
Out

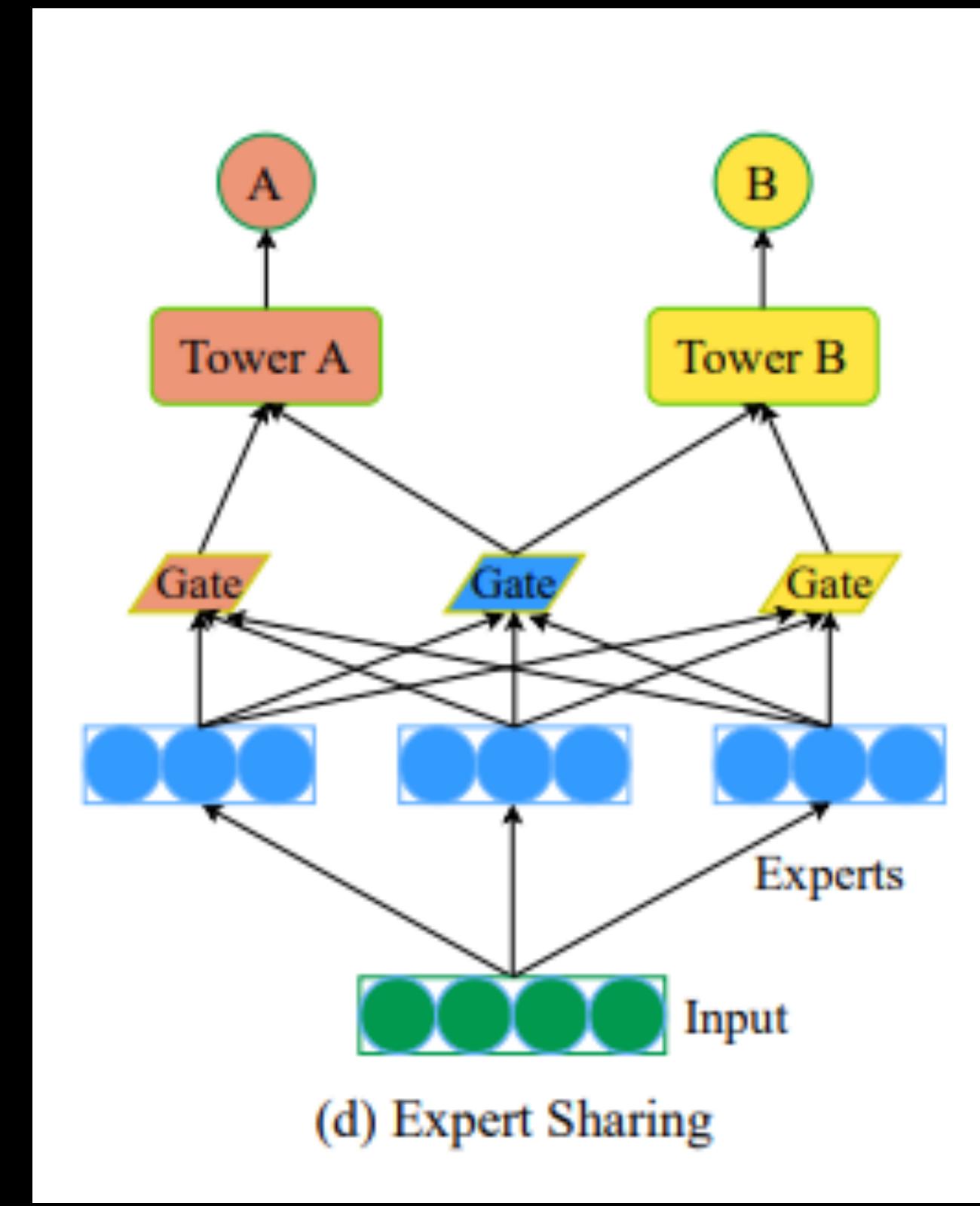
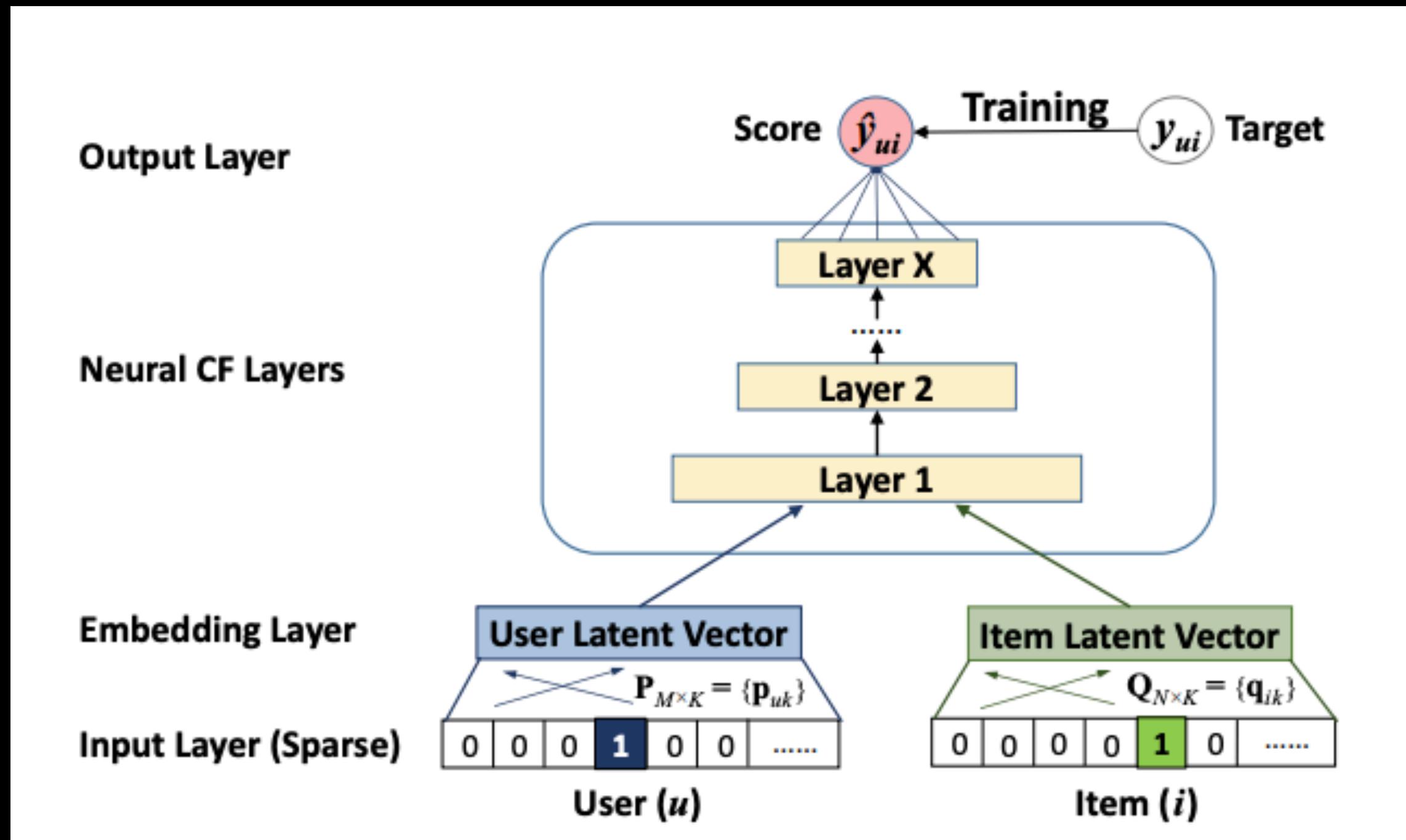


Post
purchase



For you

How to solve it?



How to solve it?

Output Layer

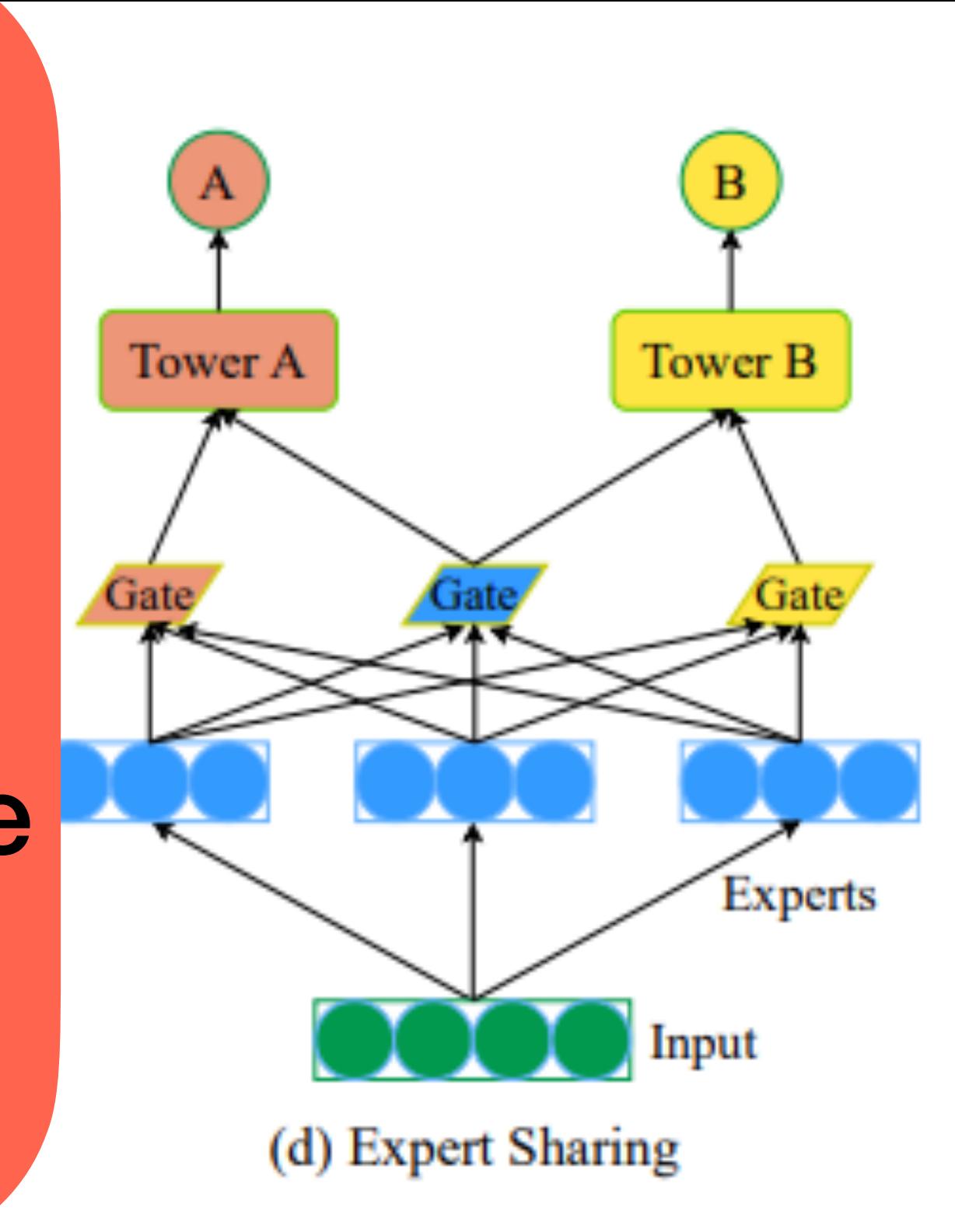
Neural CF Layers

Embedding Layer

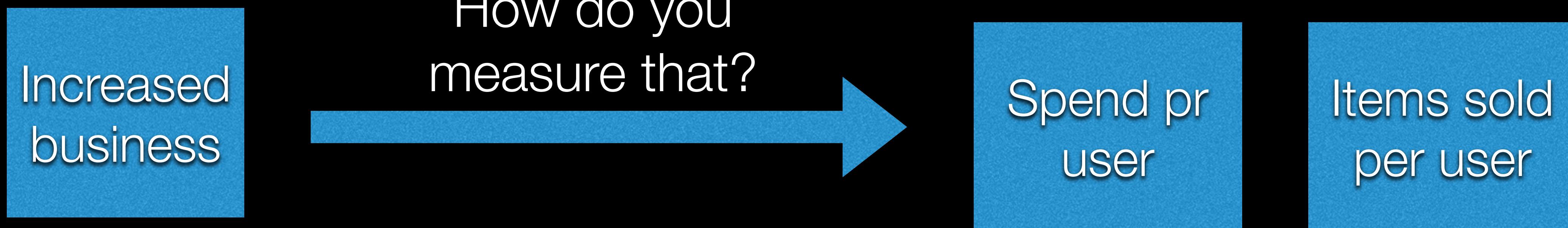
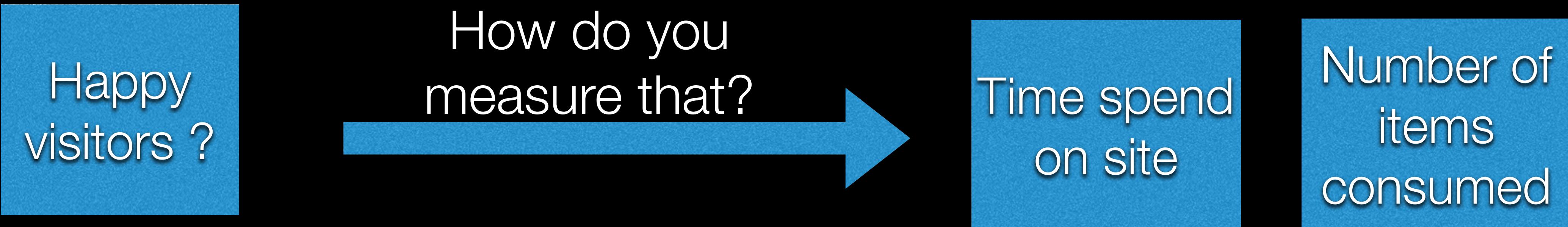
Input Layer (Sparse)



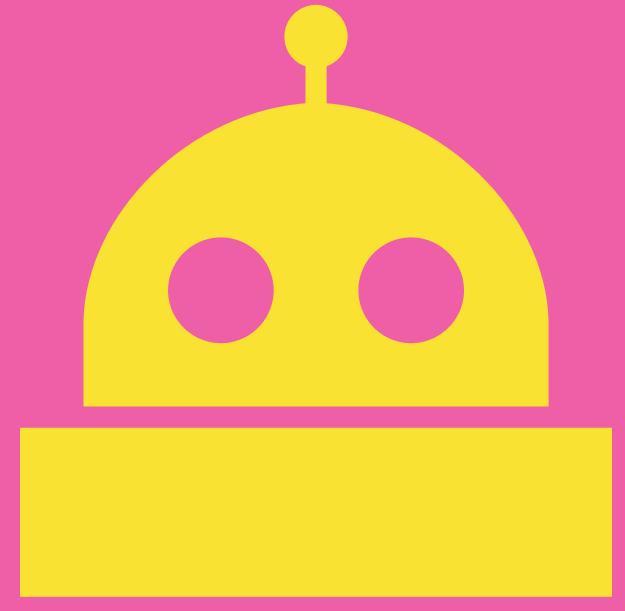
Start solving the problems individually, as simple as possible and then move on to more complex solutions when you understand the performance of the simpler ones



How do we know we have solved it



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

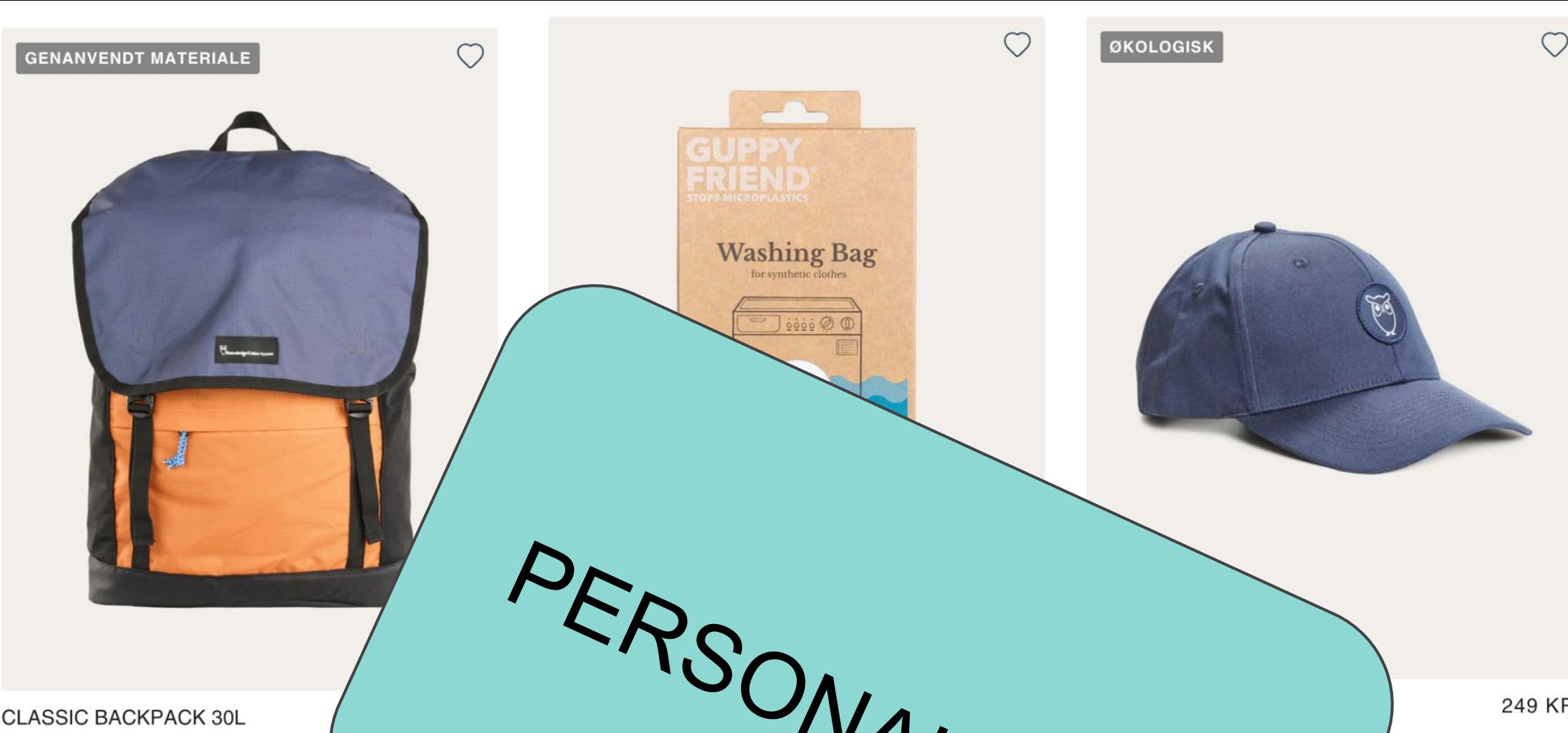
219 KR.



CLASSIC BACKPACK 30L

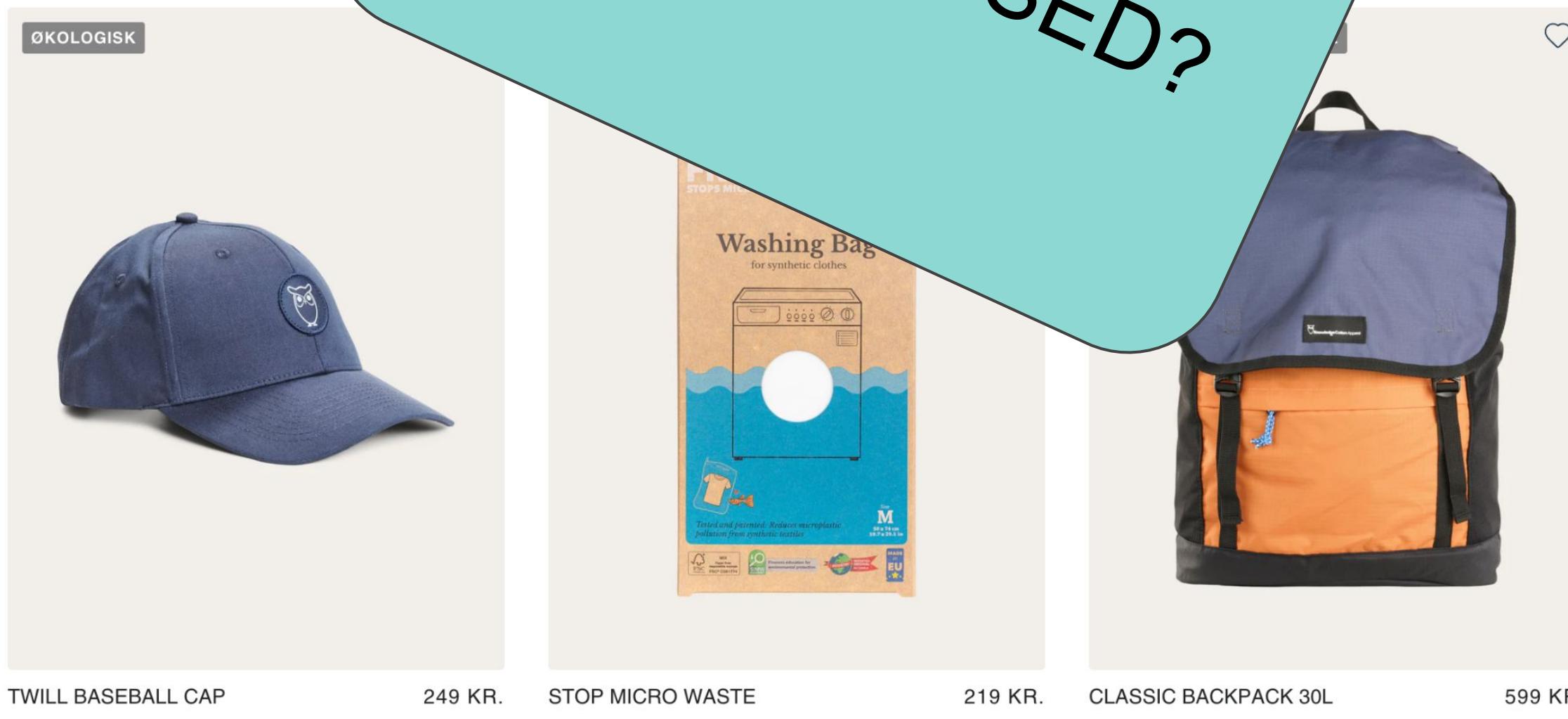
599 KR.

User 1:

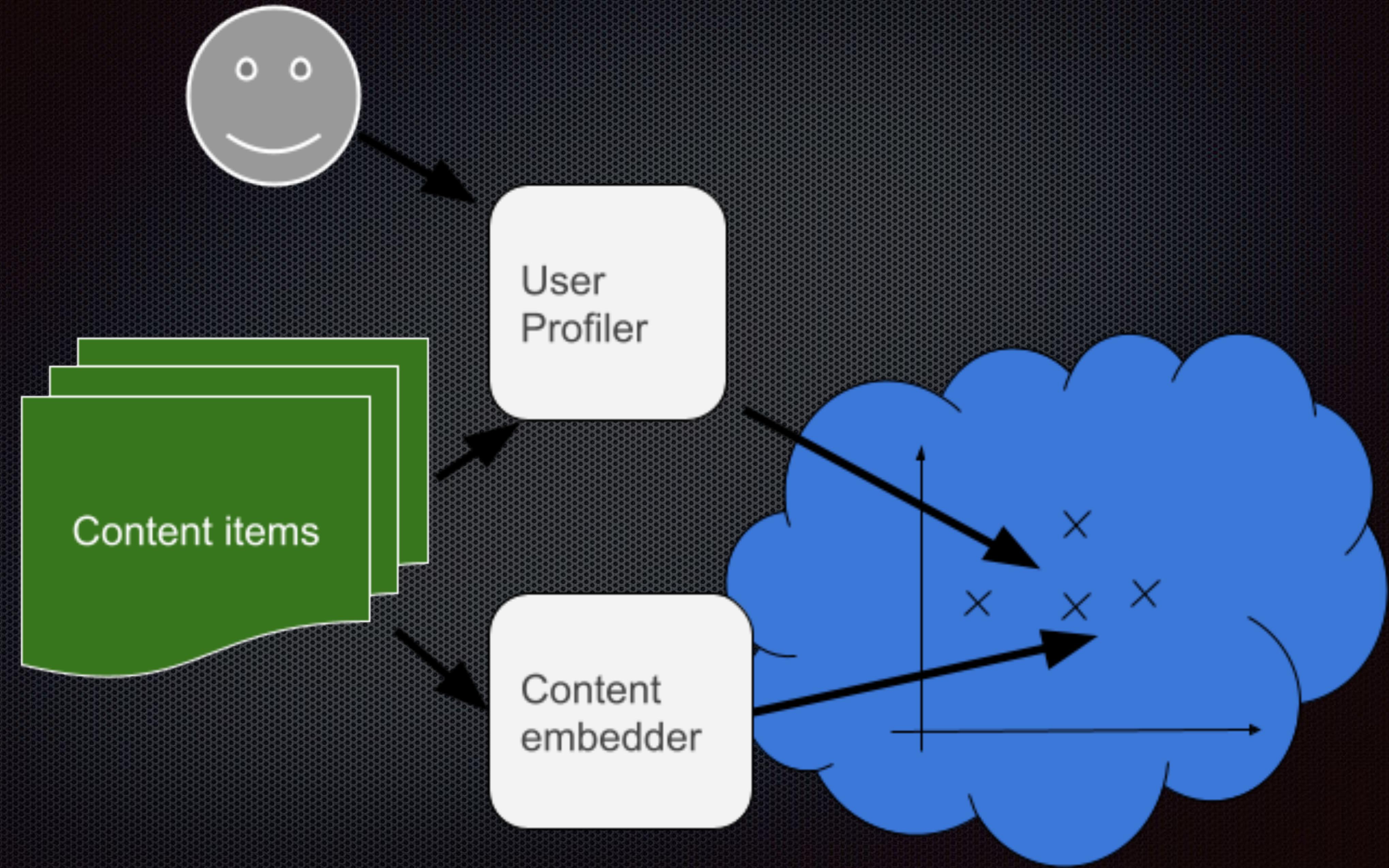


PERSONALISED?

User 2:



But what
about
the user?



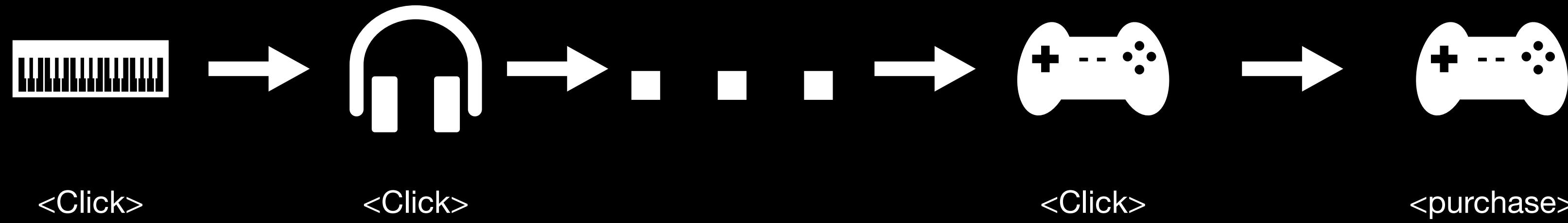
User behaviour modelling

What is personalisation?

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

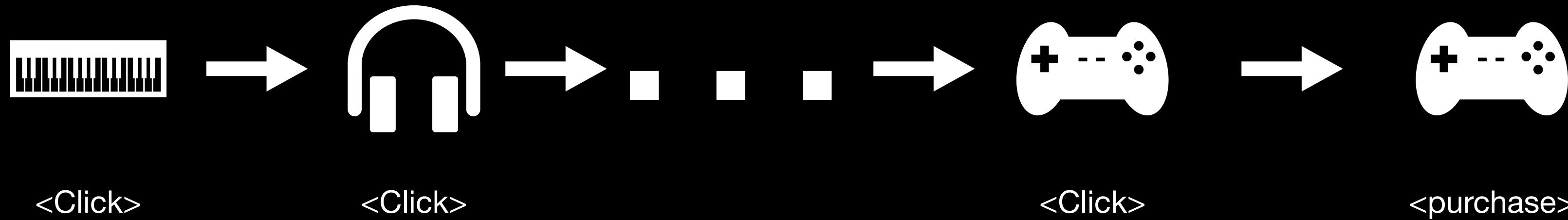


Help the user find the content quicker



Decrease time from landing page to purchase

Help the user find the content quicker

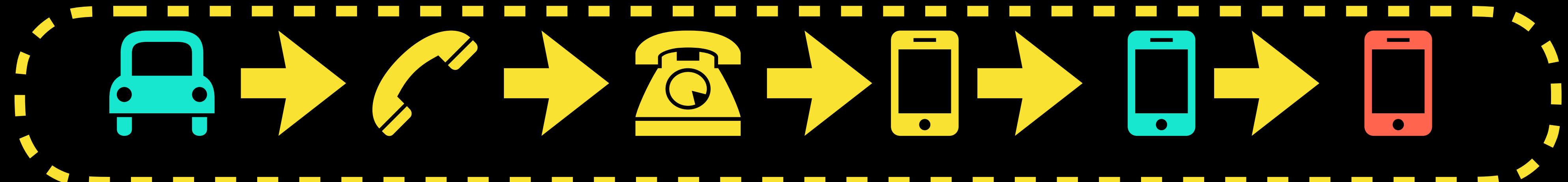


When can we personalise

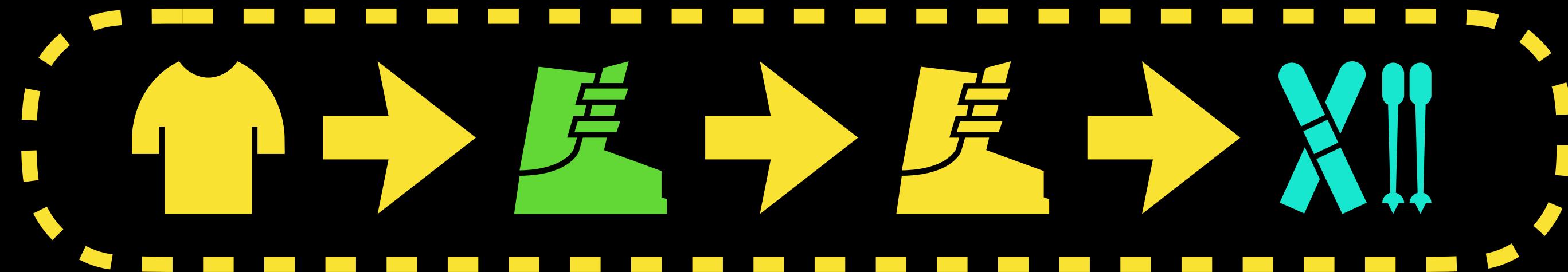


Sessions

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

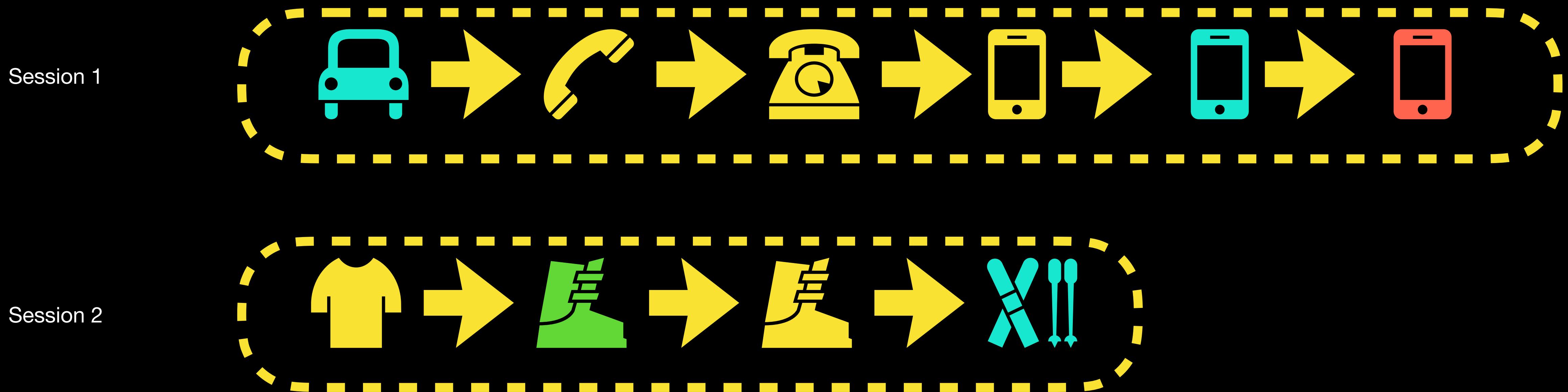
Gender preferences

Historical data

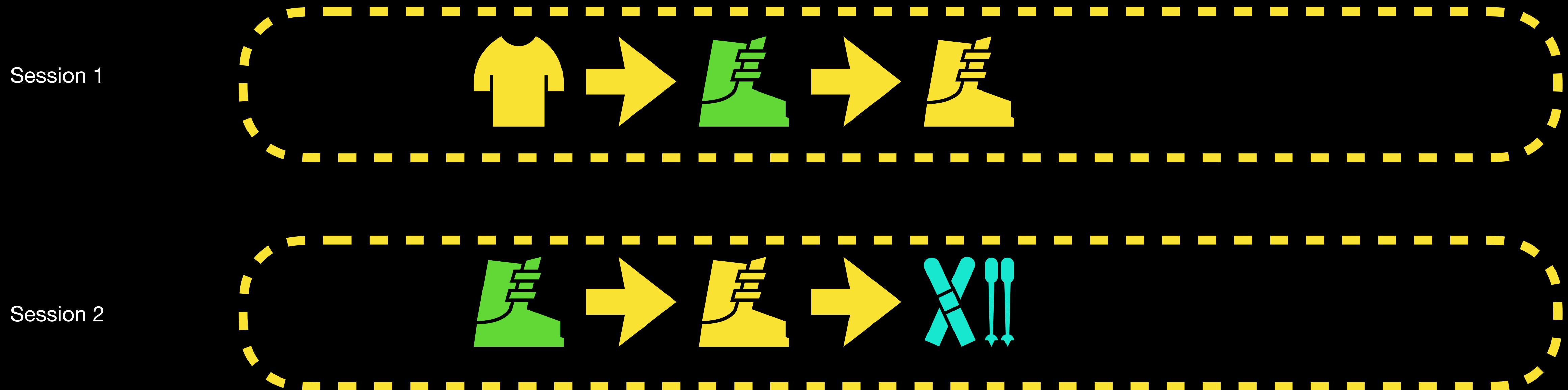
But what could they be in the domain you are looking in?

Don't recommend what's already consumed

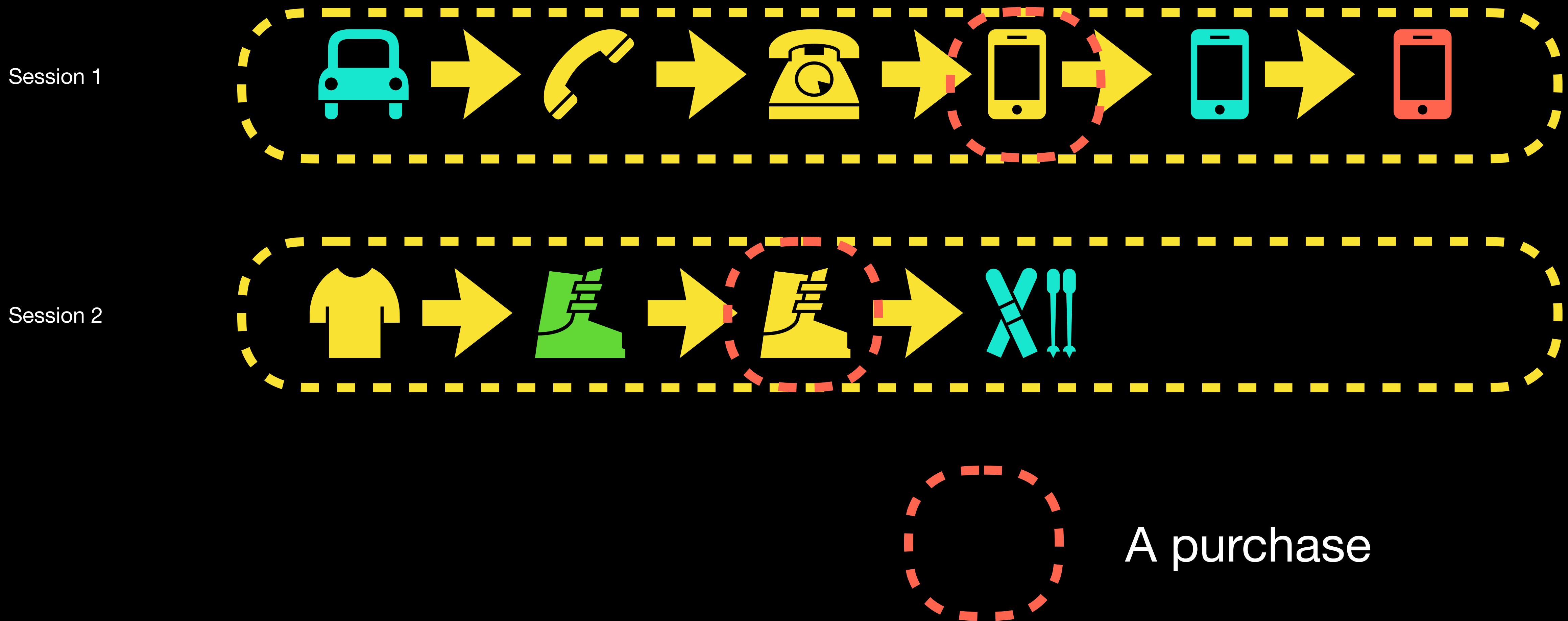
What can I learn from session one that makes session two easier?



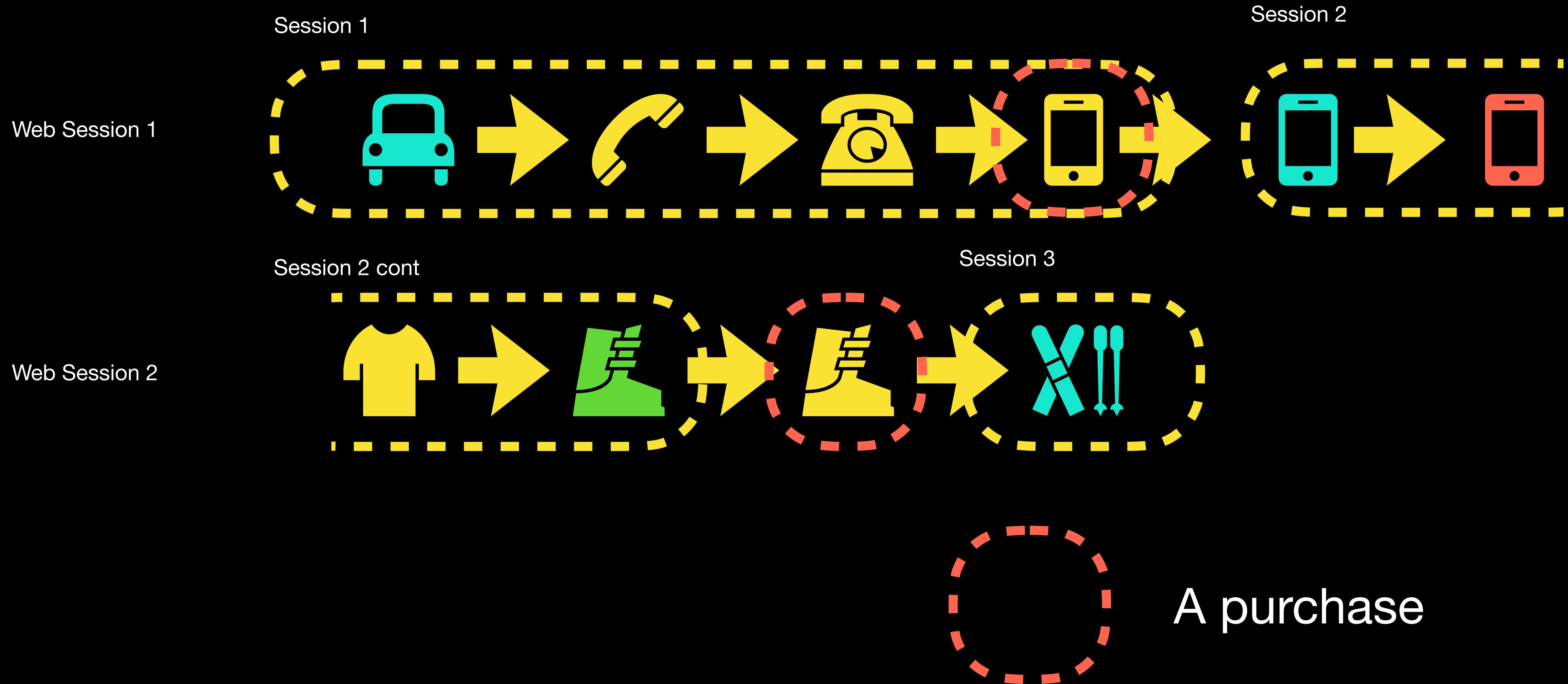
What can I learn from session one that makes session two easier?



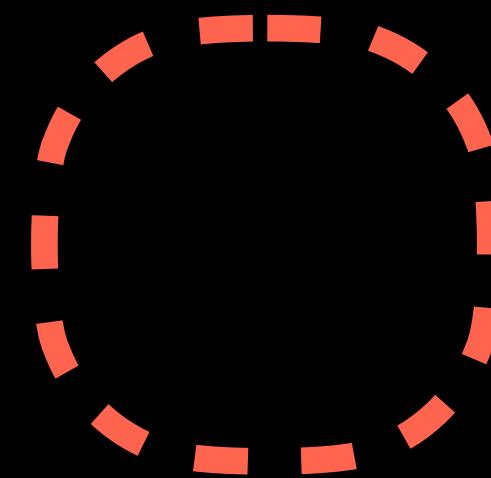
What can I learn from session one that makes session two easier?



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



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



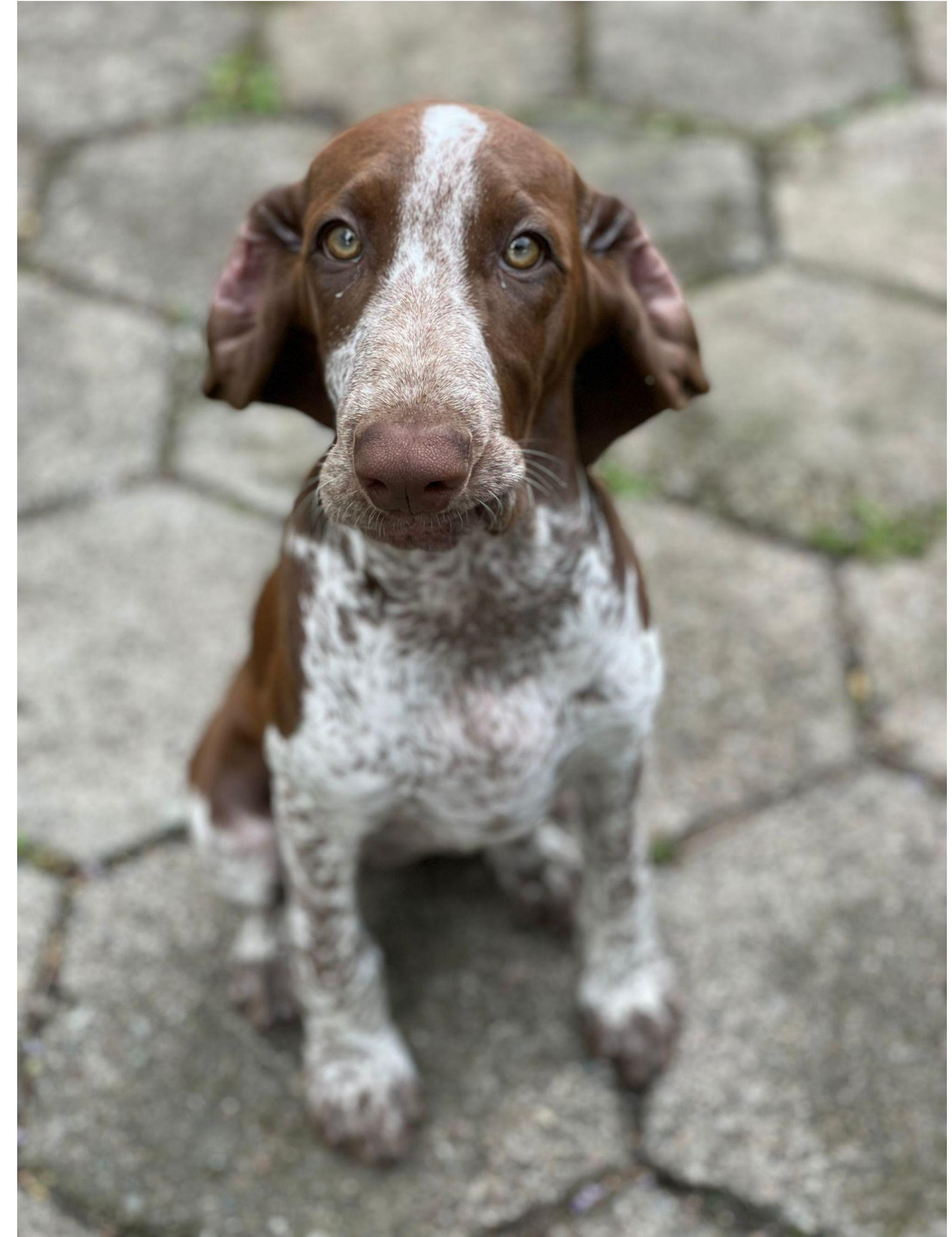
A purchase or a add to cart or checkout
finishes one logical session

What is the logical next thing to recommend after that?

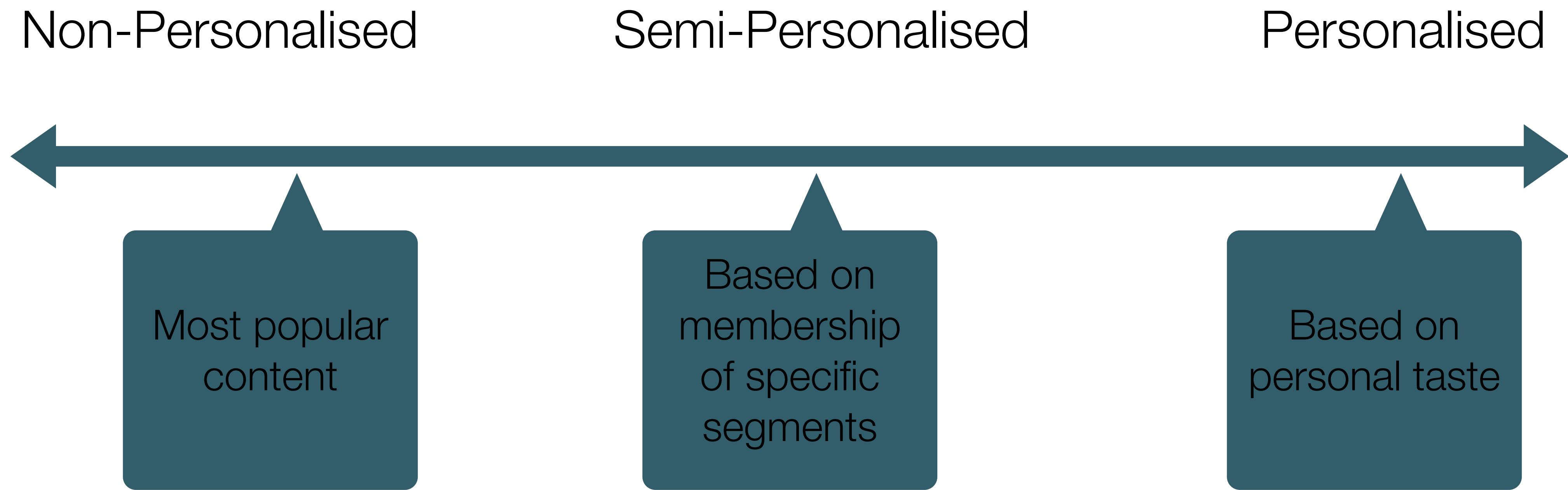


Again!

**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

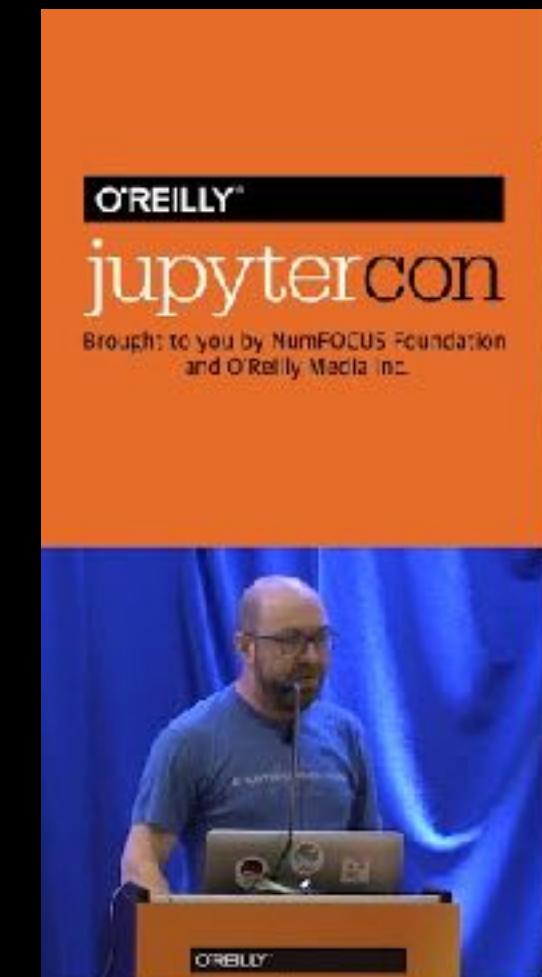
- Versioning of code
- Continuous integration



Data/ML engineering

- ML functional testing
- Experiment logging

Expect to write testable code



AUG 21-24, 2018
NEW YORK, NY

jupytercon.com
#JupyterCon



@joelgrus #jupyterc

UNTITLED25.IPYNB

UNTITLED24.IPYNB



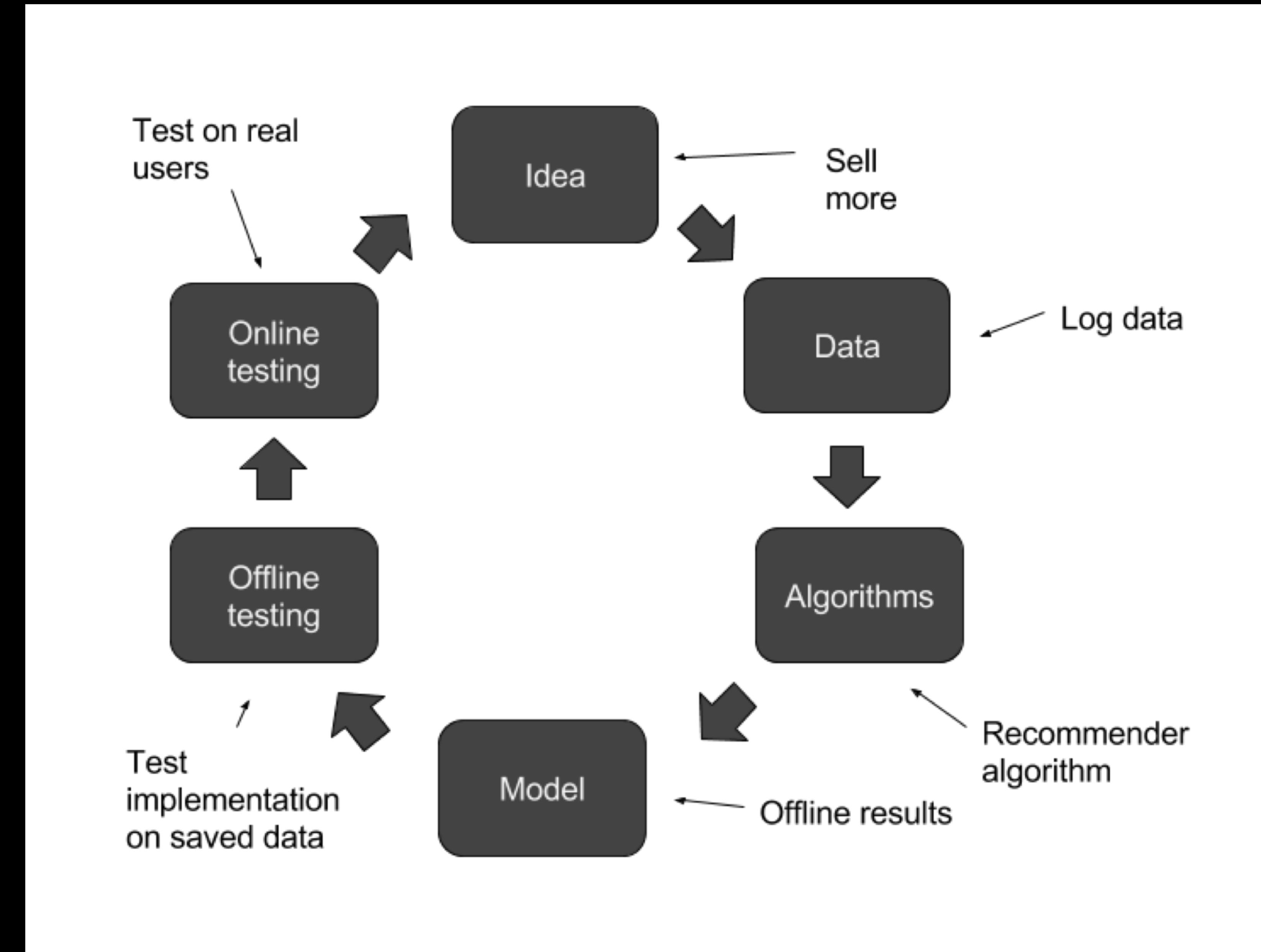
A brown and white basset hound dog is sitting on a dark-colored couch. The dog has long, drooping ears and is looking towards the right side of the frame. In the background, there's a window with a white frame, a potted plant on a shelf, and a decorative vase on a surface to the left.

How do we develop
recommender systems in
house?

Being scientific about it.

**Start with the
problem not the
coolest article
from #RecSys23**

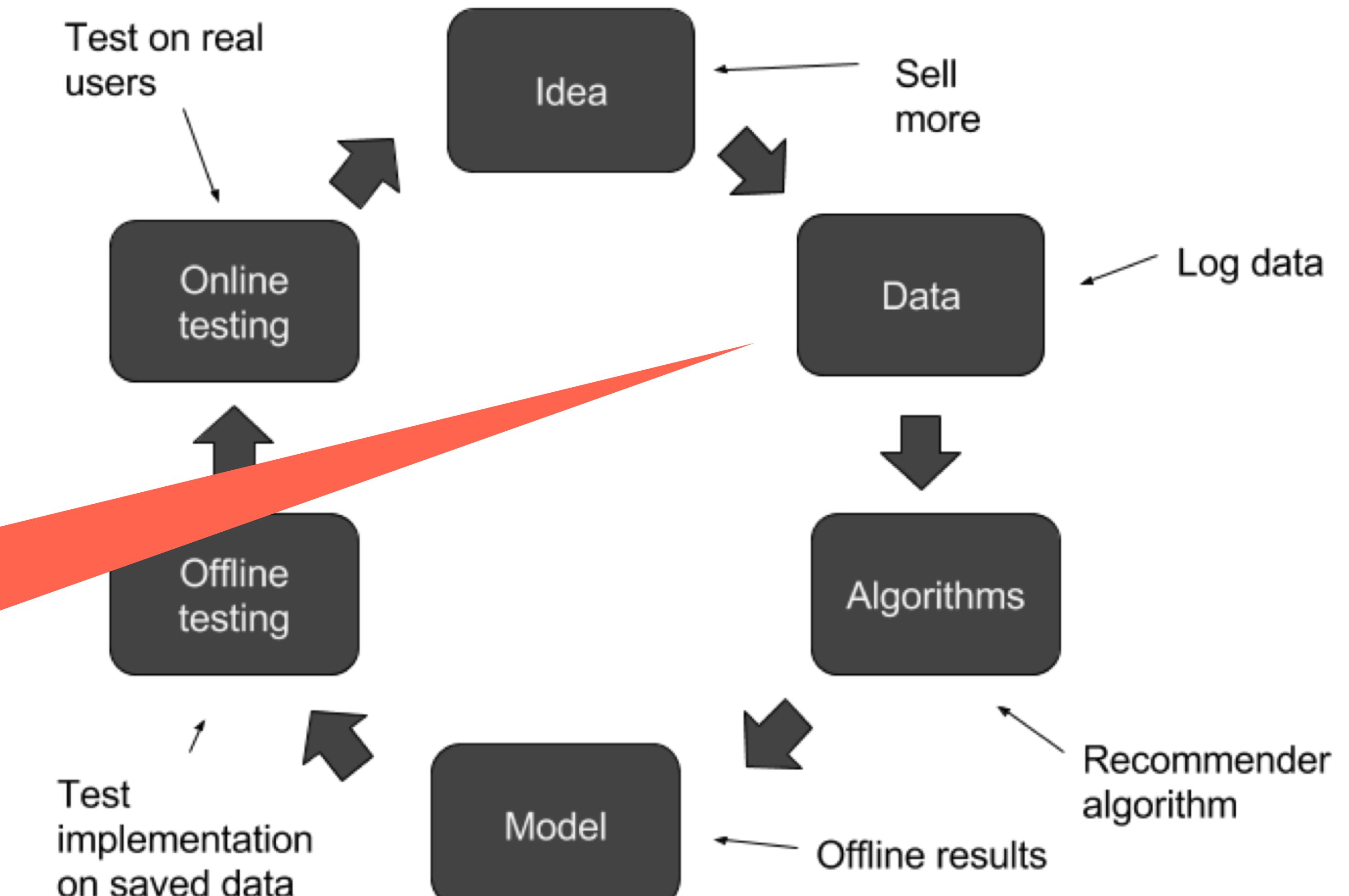
ML Engineering Cycle



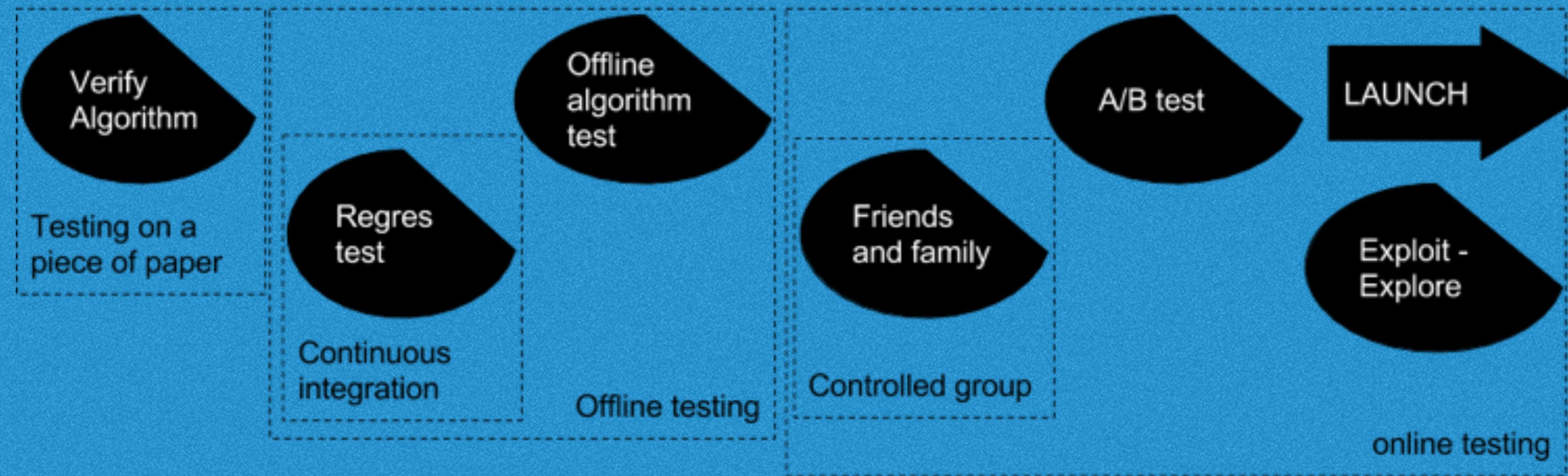
ML Engineering Cycle

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:

Engineers

Users

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 Falk³ and Bart Goethals^{1,2,4}

¹*Froomle N.V., Belgium*

²*University of Antwerp, Antwerp, Belgium*

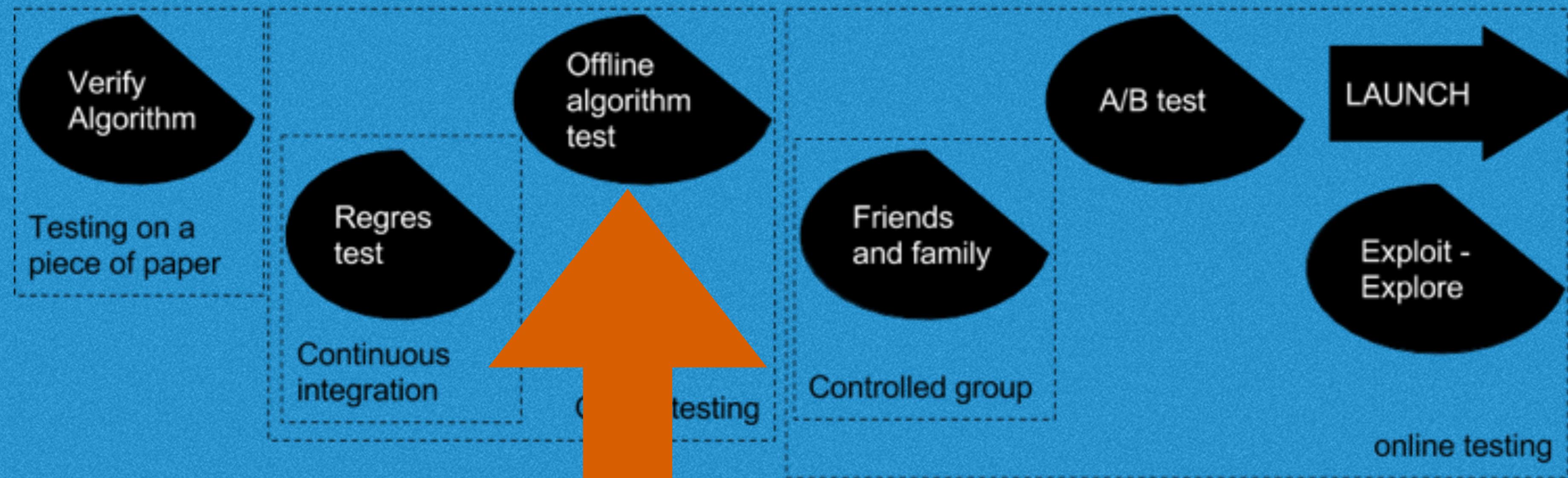
³*Shopify, Canada*

⁴*Monash University, Melbourne, Australia*

Abstract

Software engineers test virtually all of their code through unit, regression and integration tests. In contrast, data scientists and machine learning engineers often evaluate models based solely on training or evaluation loss and task performance metrics such as accuracy, precision or recall. When a model becomes ‘algorithms’, software best practices are often neglected. In our research, we found that publicly available algorithm implementations indeed are not tested beyond ranking performance metrics such as recall and normalized discounted cumulative gain. Applying software testing best practices to algorithms can seem daunting (and unnecessary). However, software packages like scikit-learn and SpaCy have demonstrated that it definitely is possible to test (at least some aspects of) algorithms. We 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 becomes: should we test algorithms? During the workshop, we would like to open up this discussion. We start with 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 recommendation algorithm implementations. At the end of the discussion, we hope to have answered some of the following questions:

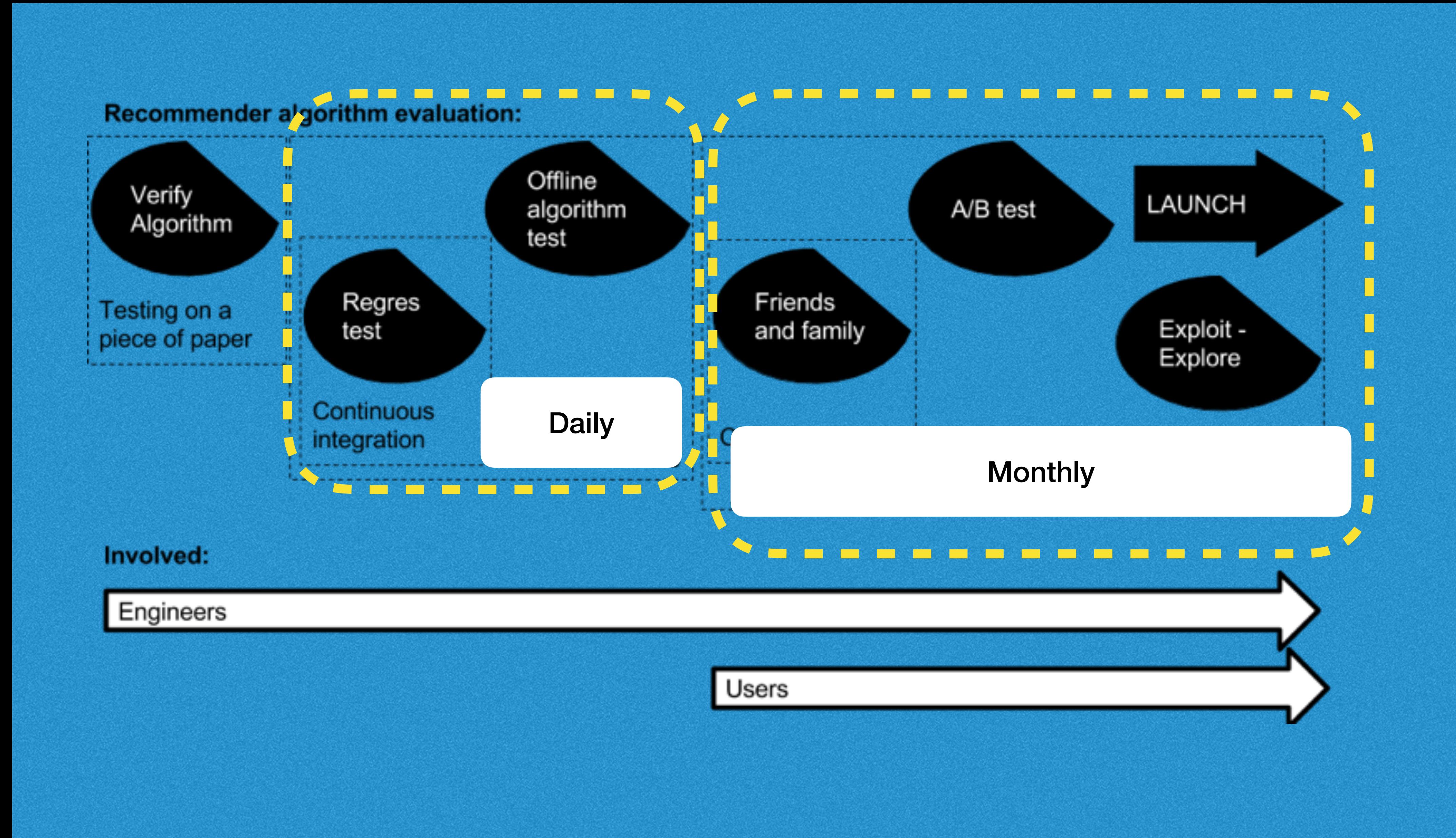
Recommender algorithm evaluation:



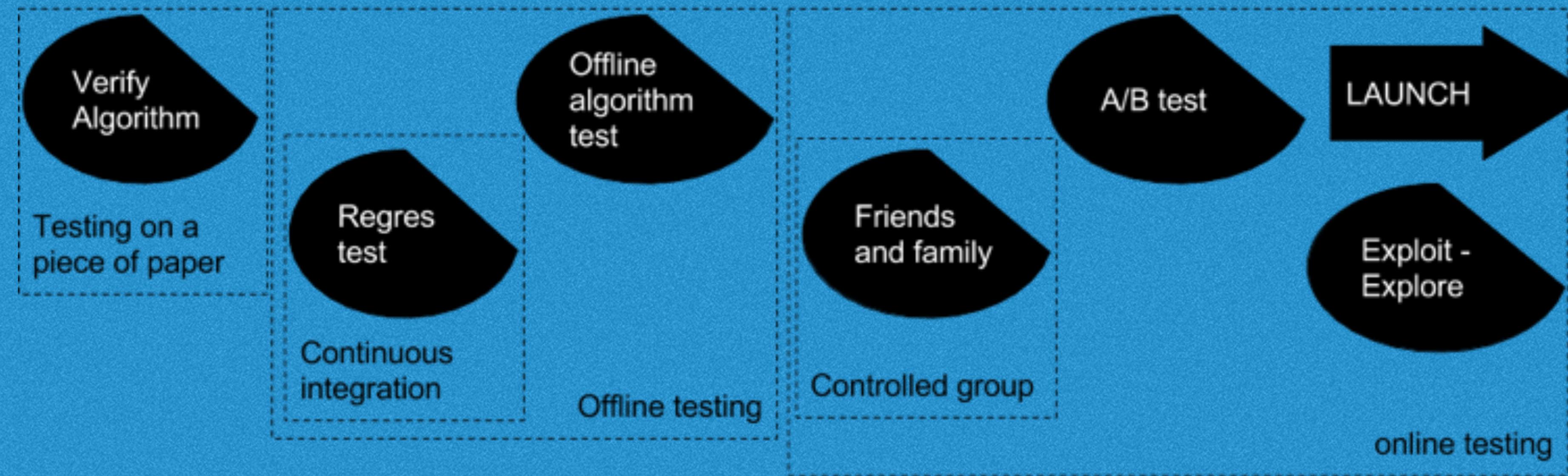
Involved:

Engineers

Users



Recommender algorithm evaluation:



Involved:

Engineers

Users

Training a model

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

SCIENTIFIC EXPERIMENT LOG

DATE

Question

TOPIC

Hypothesis

Supplies

*
*
*
*
*

Steps

1.
2.
3.
4.

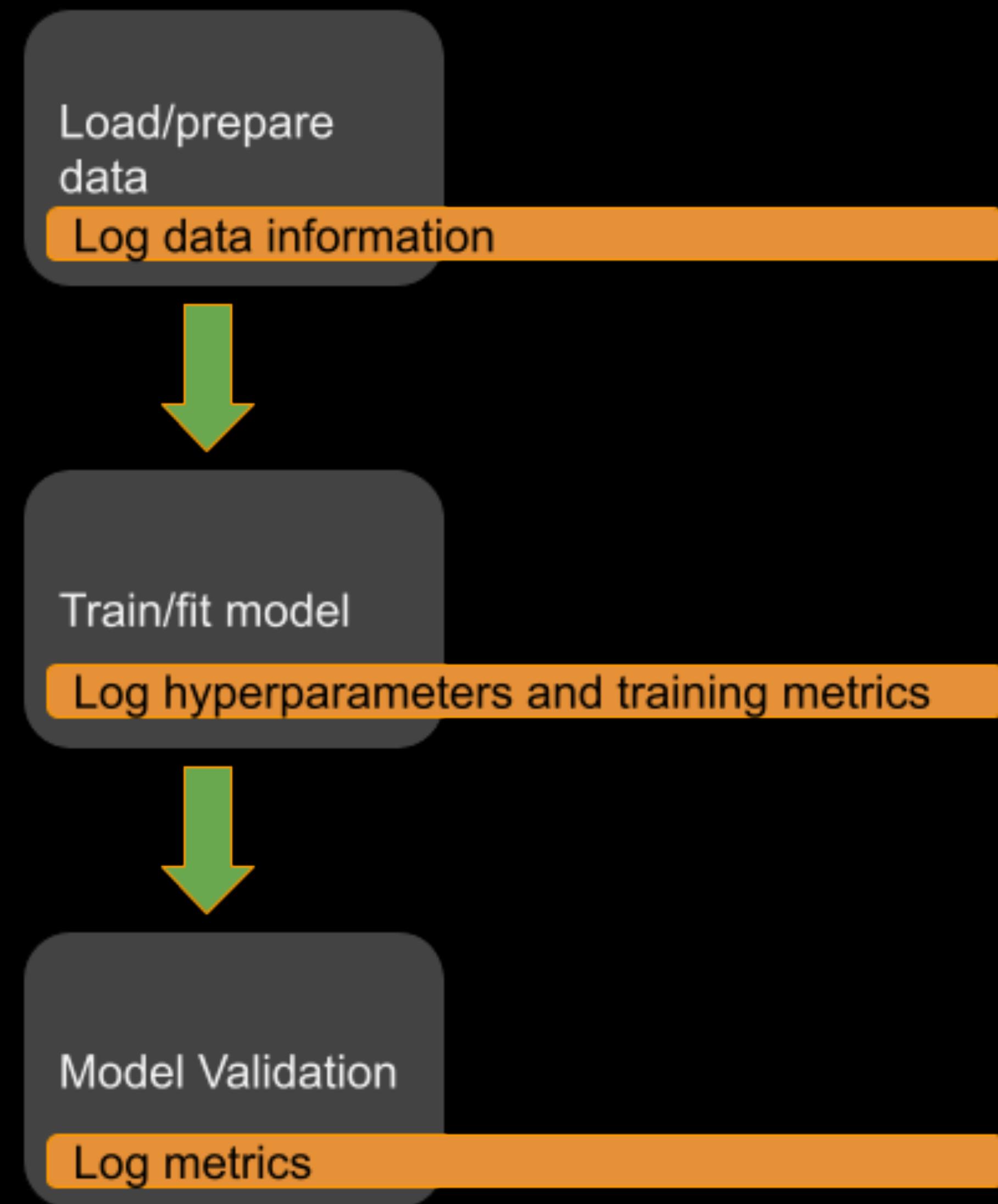
Observations

Conclusion

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

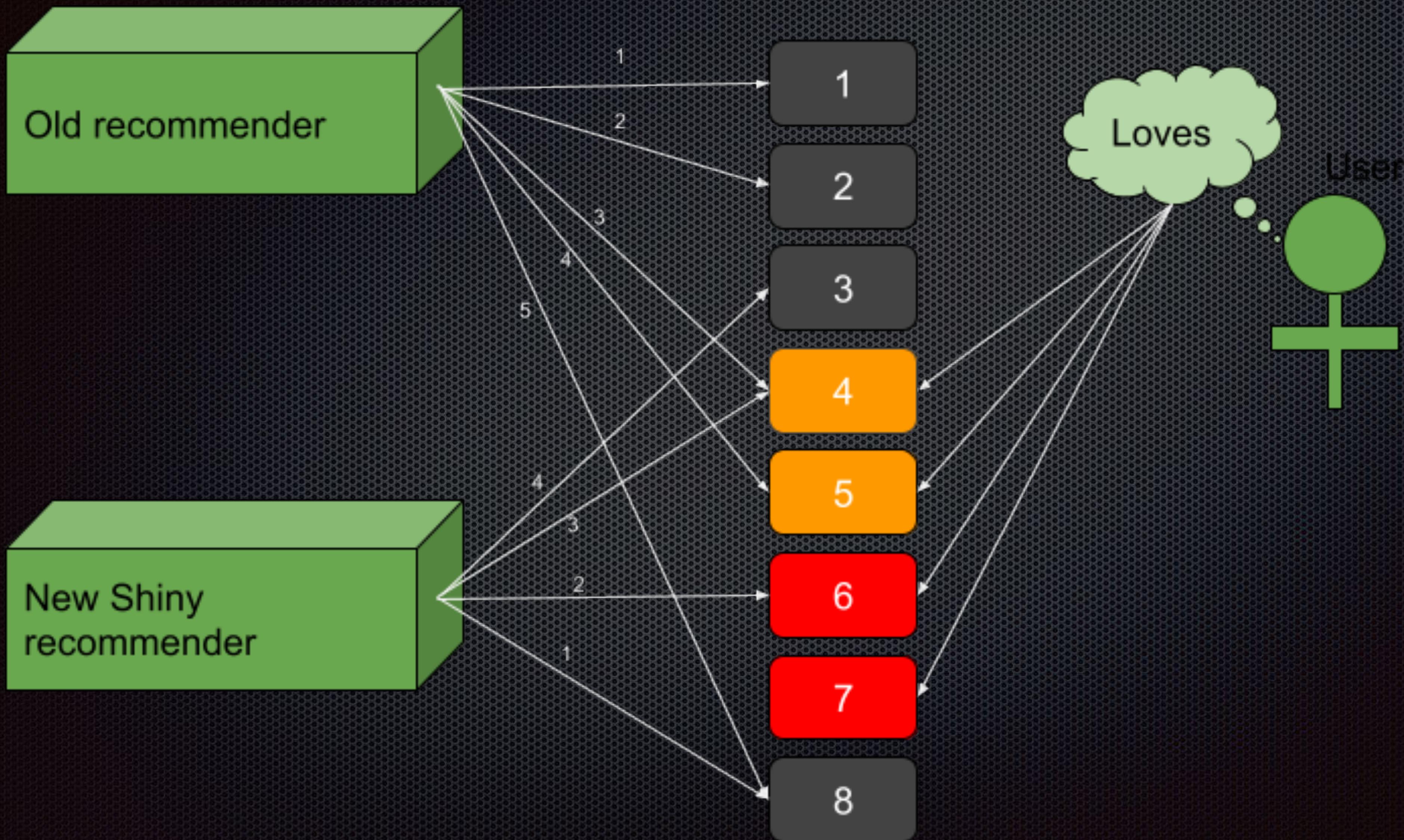
BUT I NEED TO
KEEP ALL OF THE
WORTHLESS EMPLOYEES
BECAUSE MY PAY IS
BASED ON HOW MANY
PEOPLE REPORT TO ME.



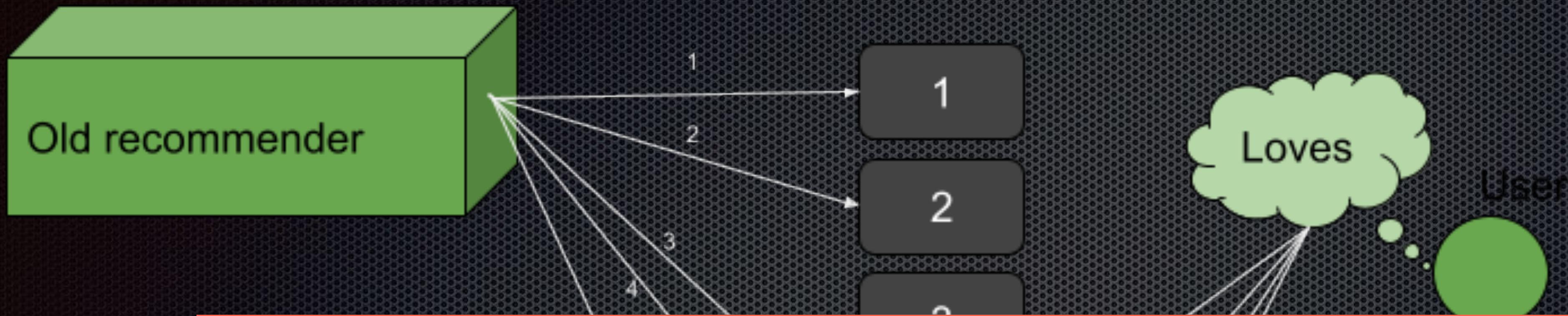
Constant dataset while developing

- 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.

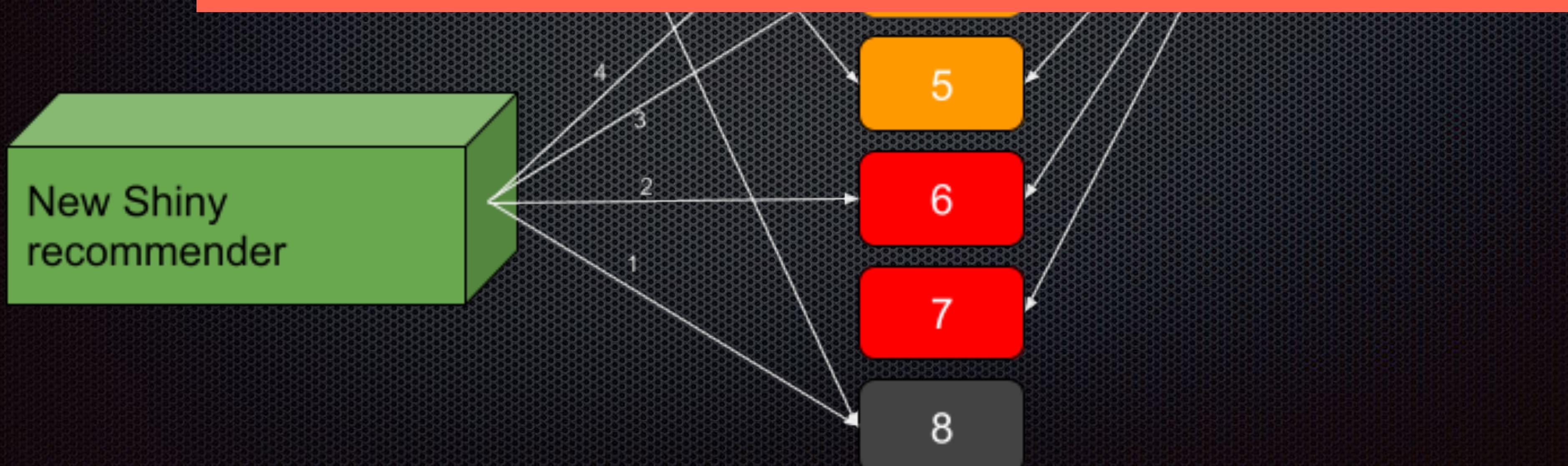
Historical data



Historical data



There is no Ground Truth!!

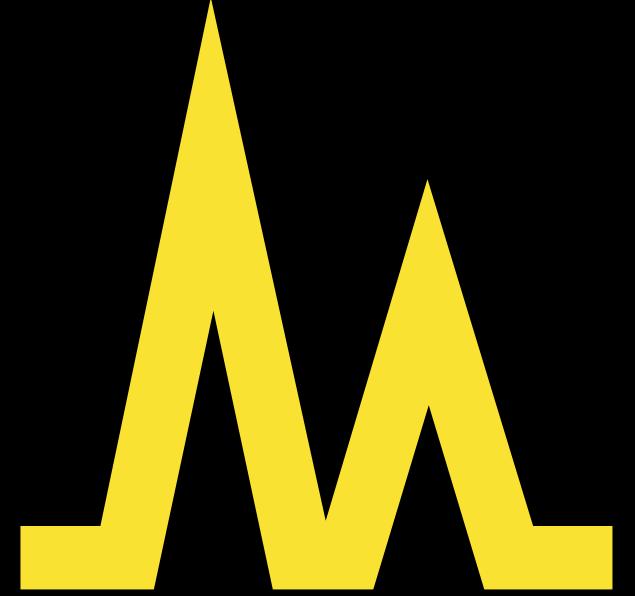


Understand the data to create good
recommendations

Training data Test data



=



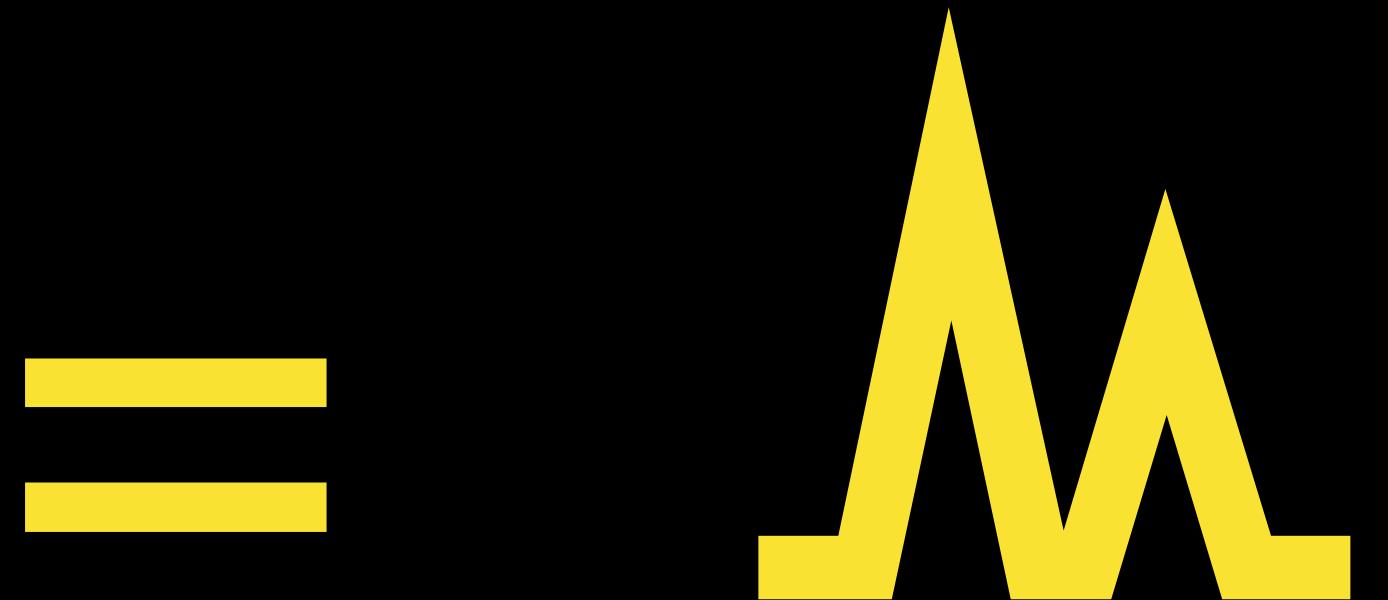
Training
data



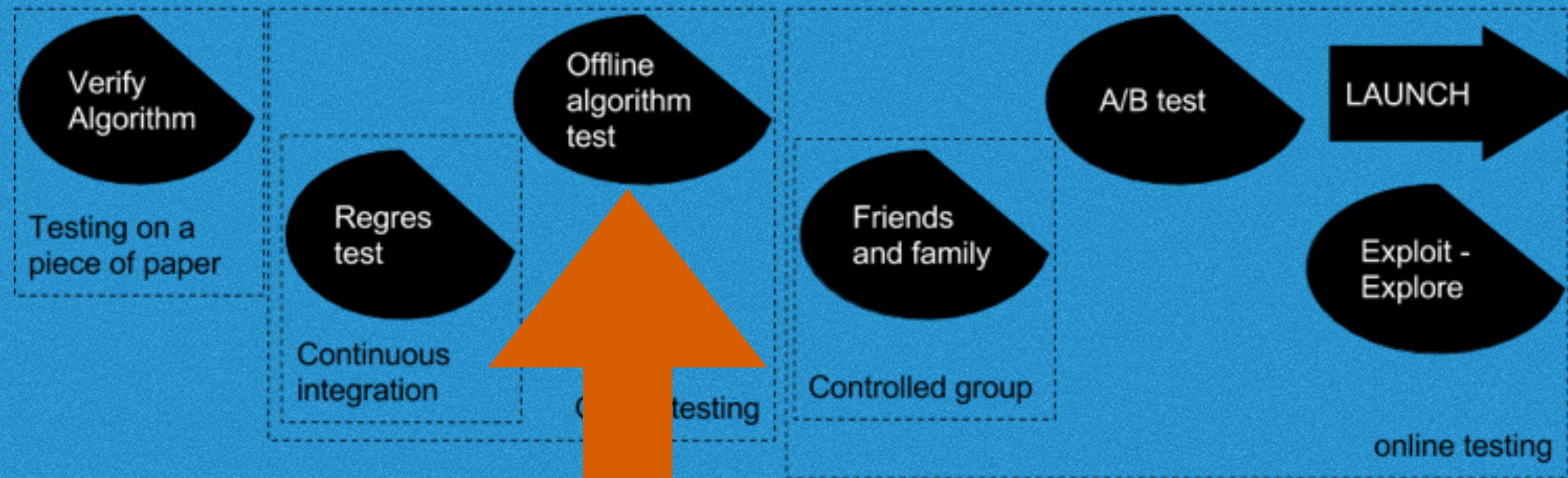
Test
data



Production
data



Recommender algorithm evaluation:



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

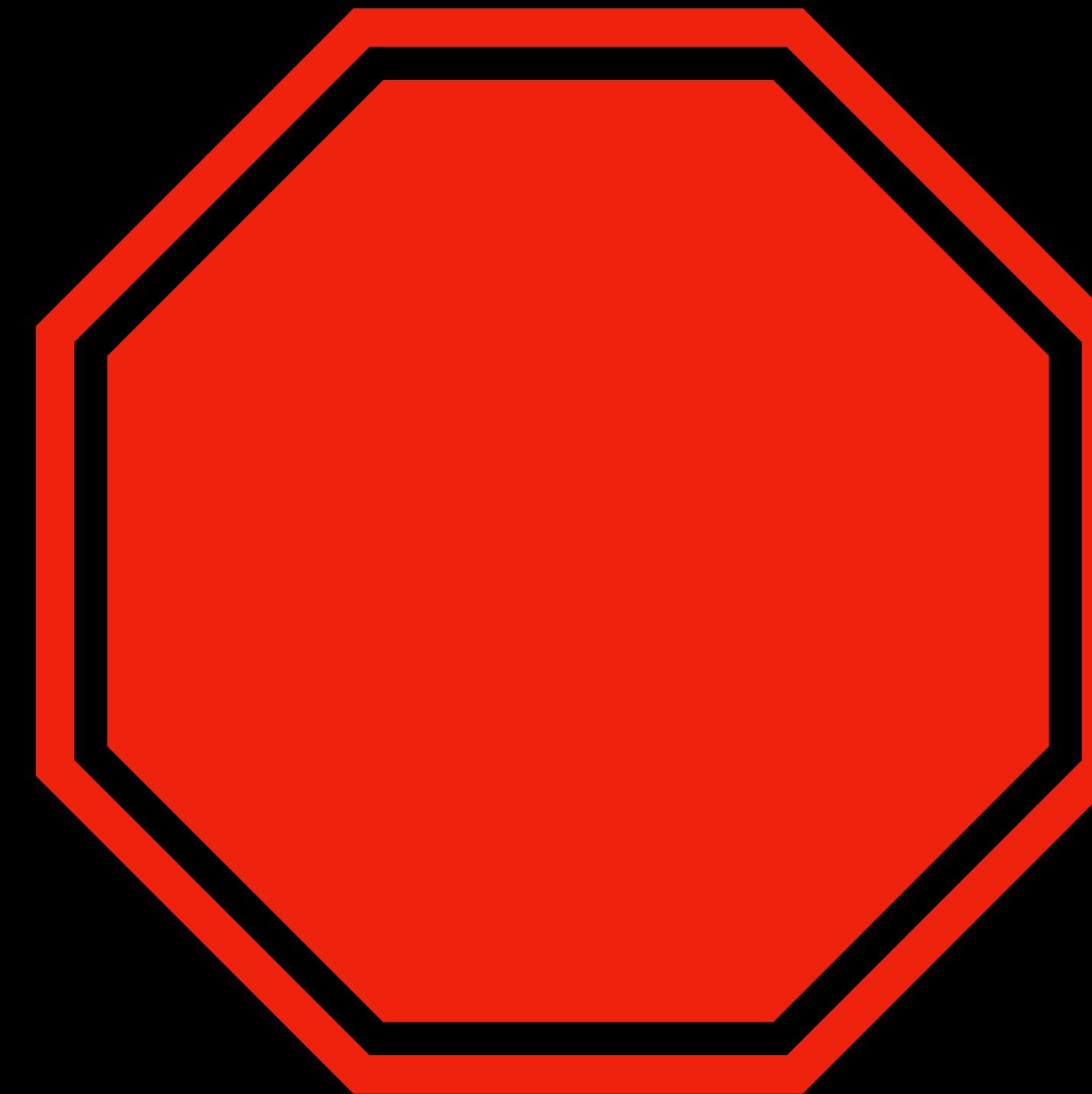
Monitoring of the
performance the
model.

Academia

Show case
performance of new
research

Business offline

A gateway

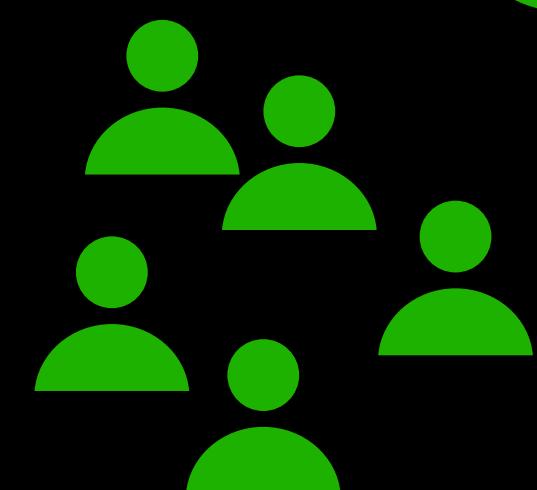


Business offline

A gateway



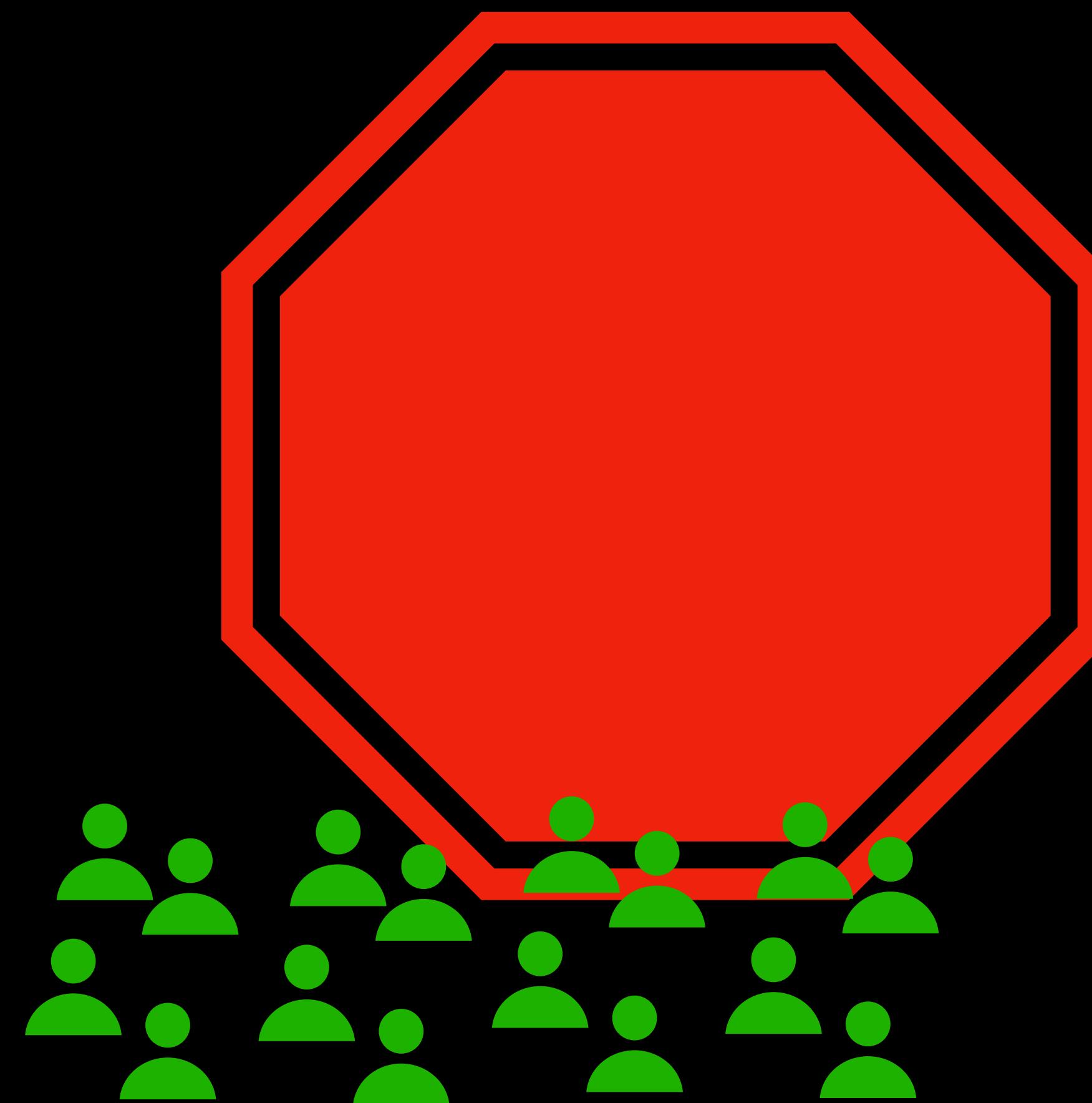
Who wins?



The losers

Is it acceptable

We need to ensure
that this is
acceptable



Do you know what metrics means?



A new metric
should be
measured before
changes are made

TO END MANY YEARS OF CONFUSION,
THE INTERNATIONAL COMMITTEE FOR
WEIGHTS AND MEASURES HAS JUST
VOTED TO REDEFINE THE KILOGRAM.

AS OF NEXT MAY, IT WILL
EQUAL EXACTLY ONE POUND.

OH, COOL.

THAT DOES MAKE
THINGS SIMPLER.

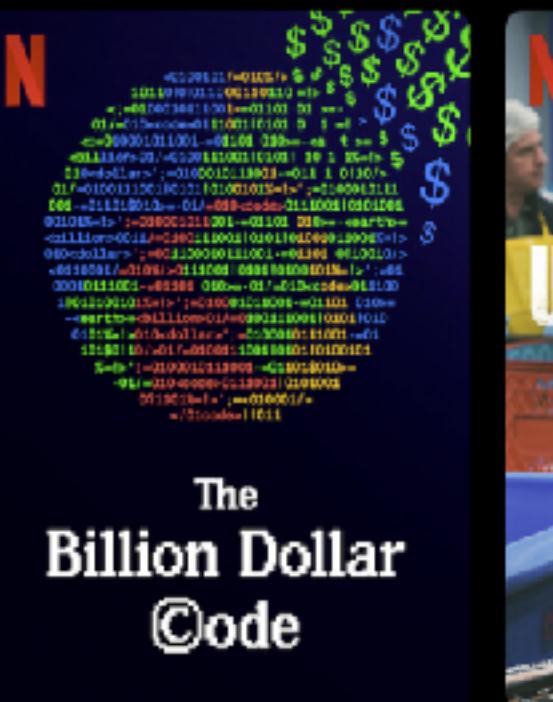
NO!!



Test your metric on real data

Test yesterdays model
with todays data?

My List



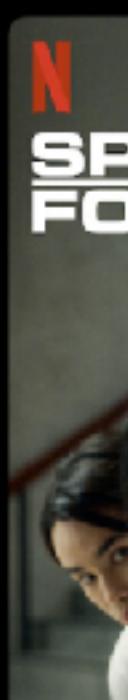
Consider the user and context

Did I view this row?

Popular on Netflix



Trending Now



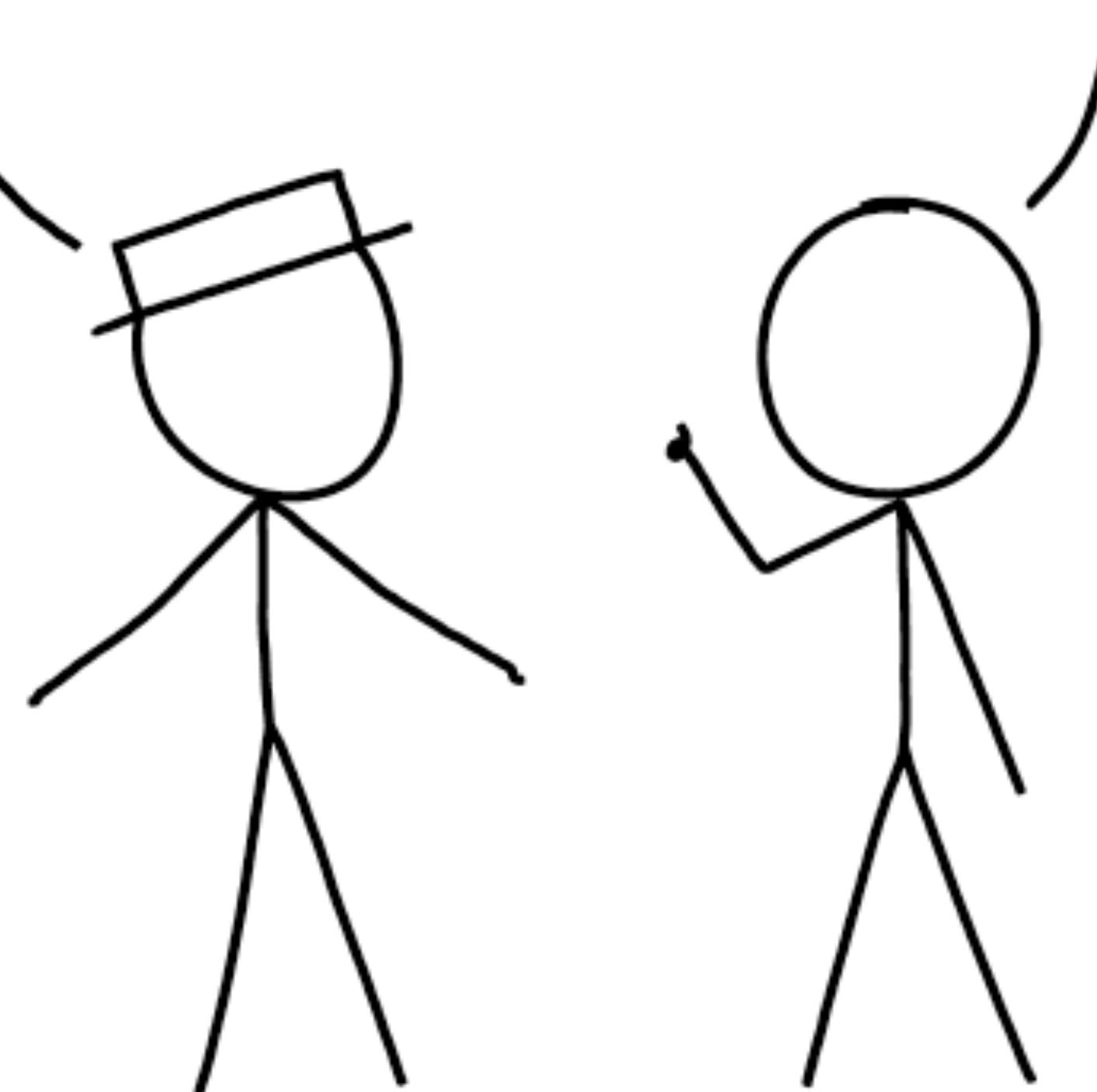
Top 10 in Denmark Today

Cannibalisation

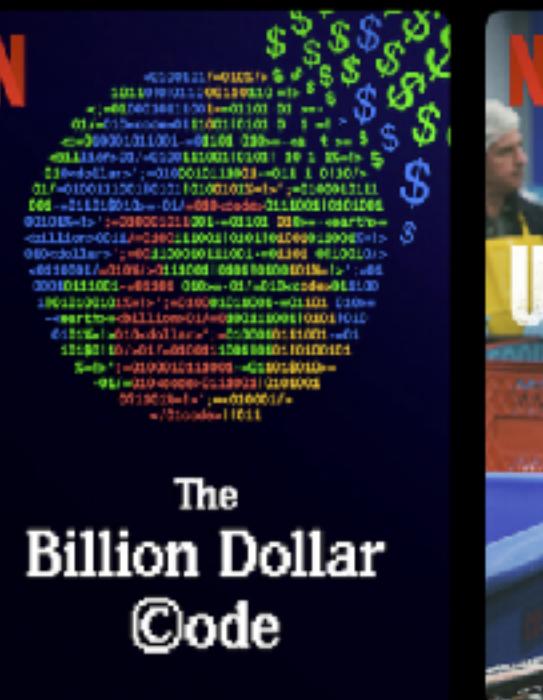
- The same content recommended.
- Change the customer journey.
- Position bias

THAT'S A FALSE DICHOTOMY!

YES, BUT WE HAVE TO EMBRACE FALSE DICHOTOMIES, BECAUSE THE ONLY ALTERNATIVE IS CANNIBALISM.



My List



This one gets more attention
then the next one

Position bias

Popular on Netflix



Trending Now

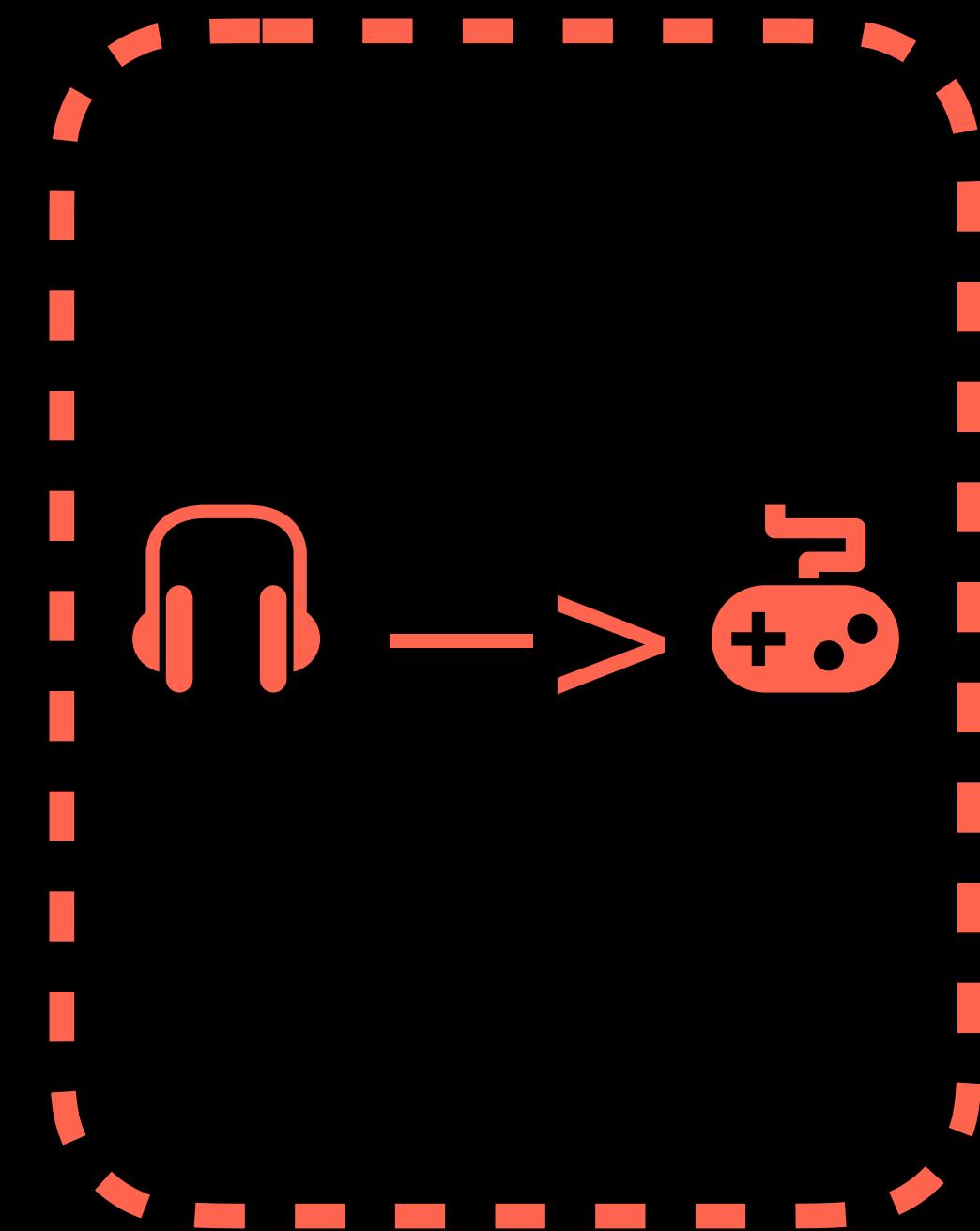


Top 10 in Denmark Today

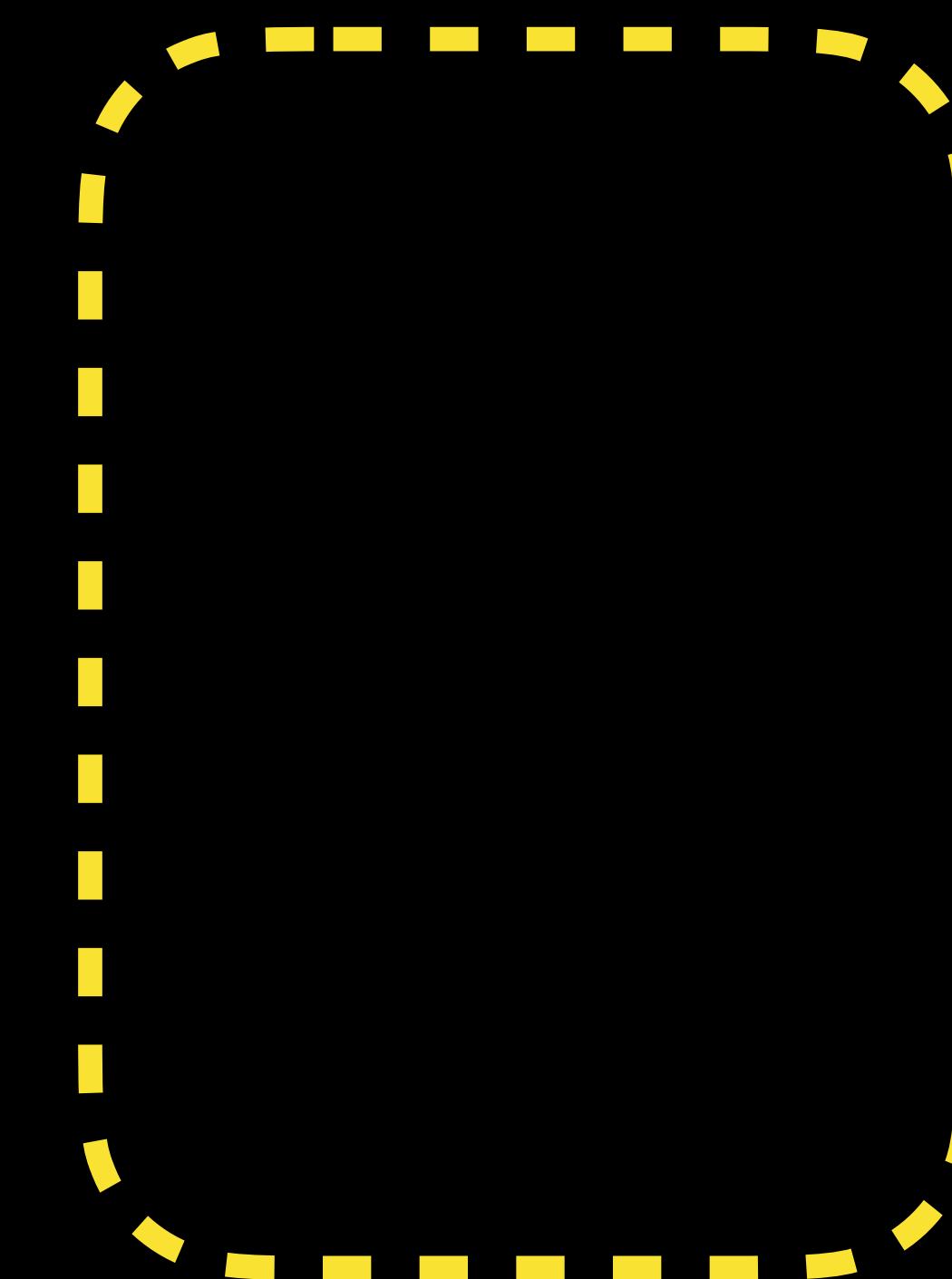
Given a session



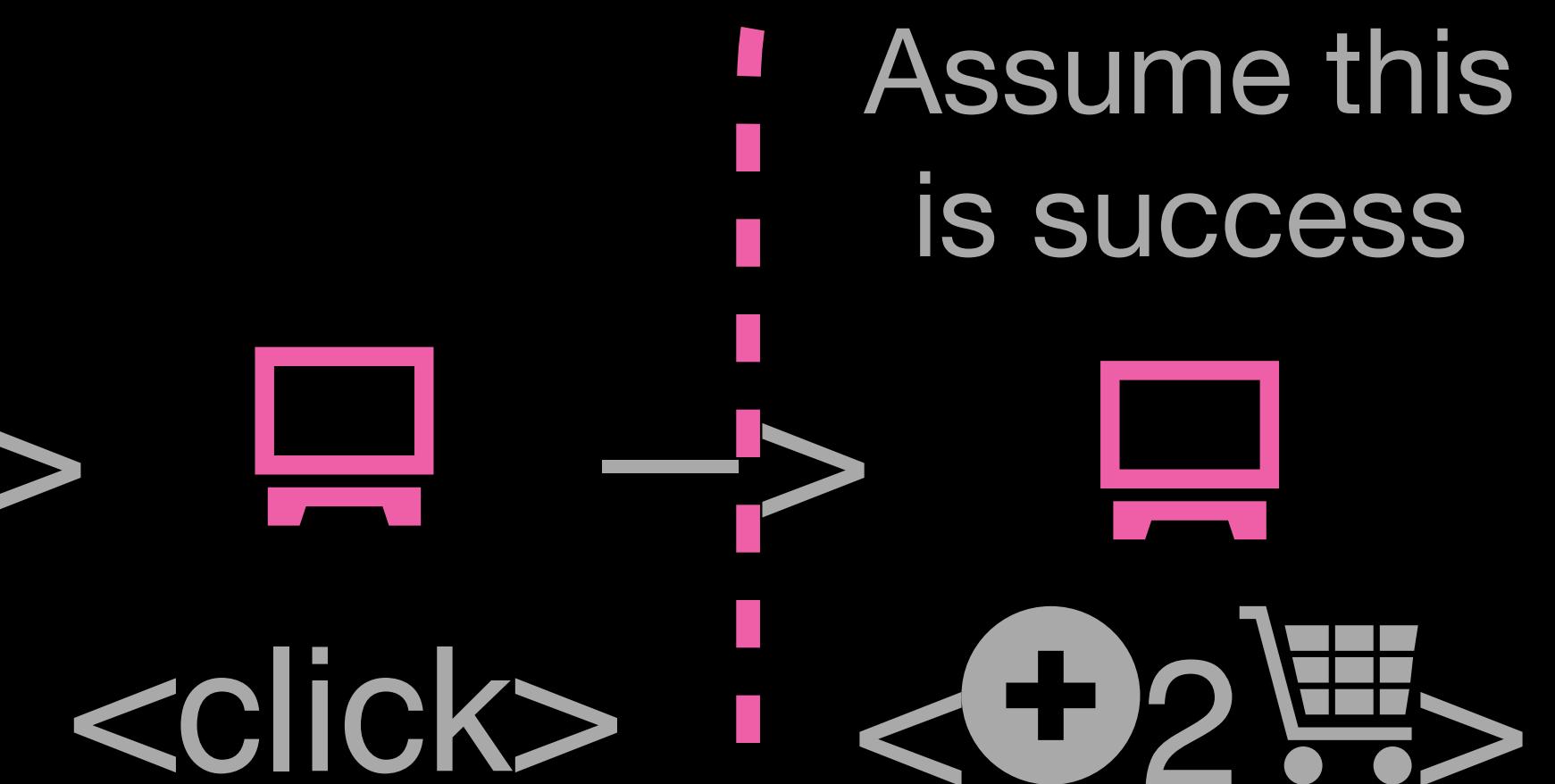
Input



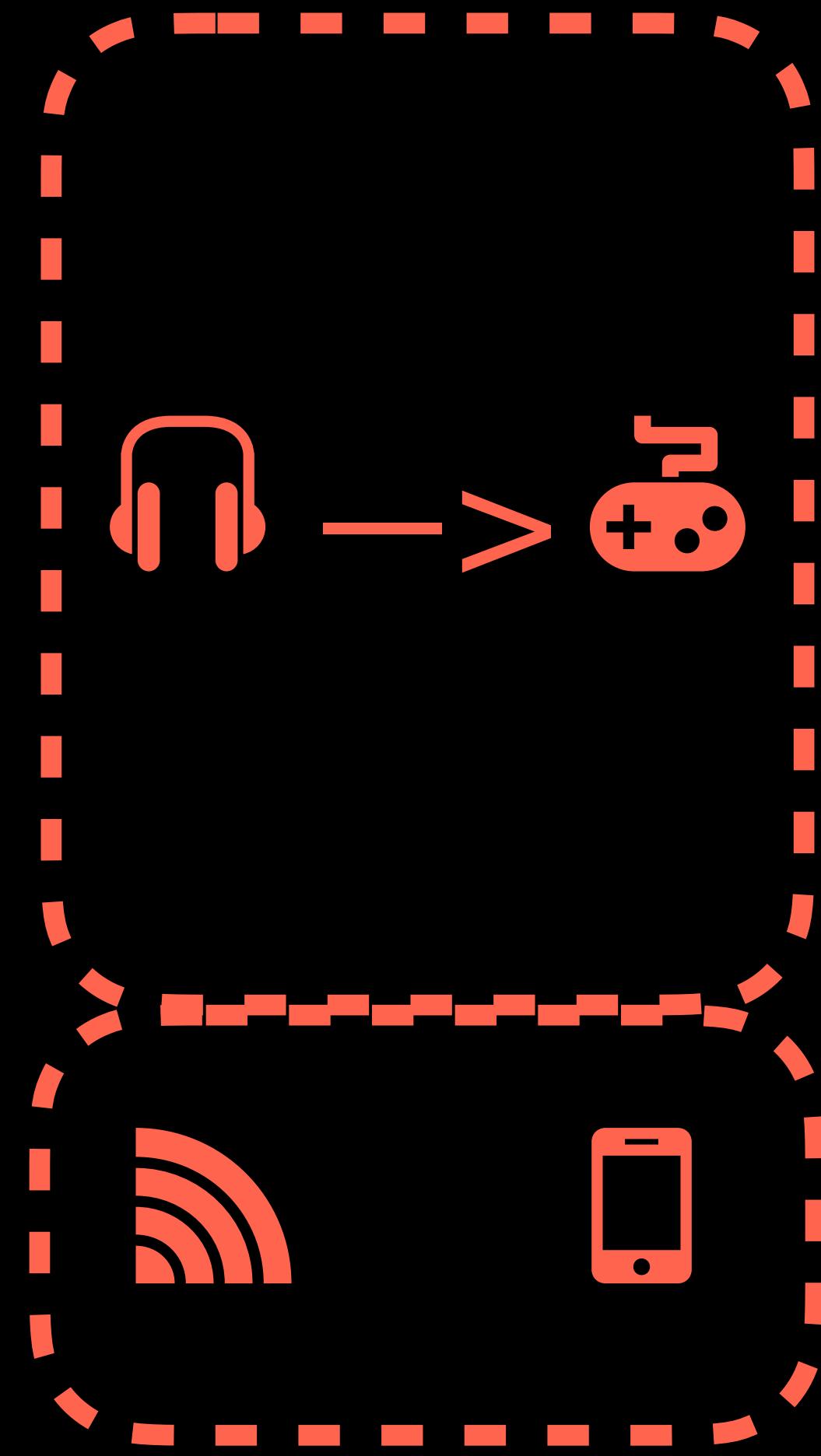
Create Recs



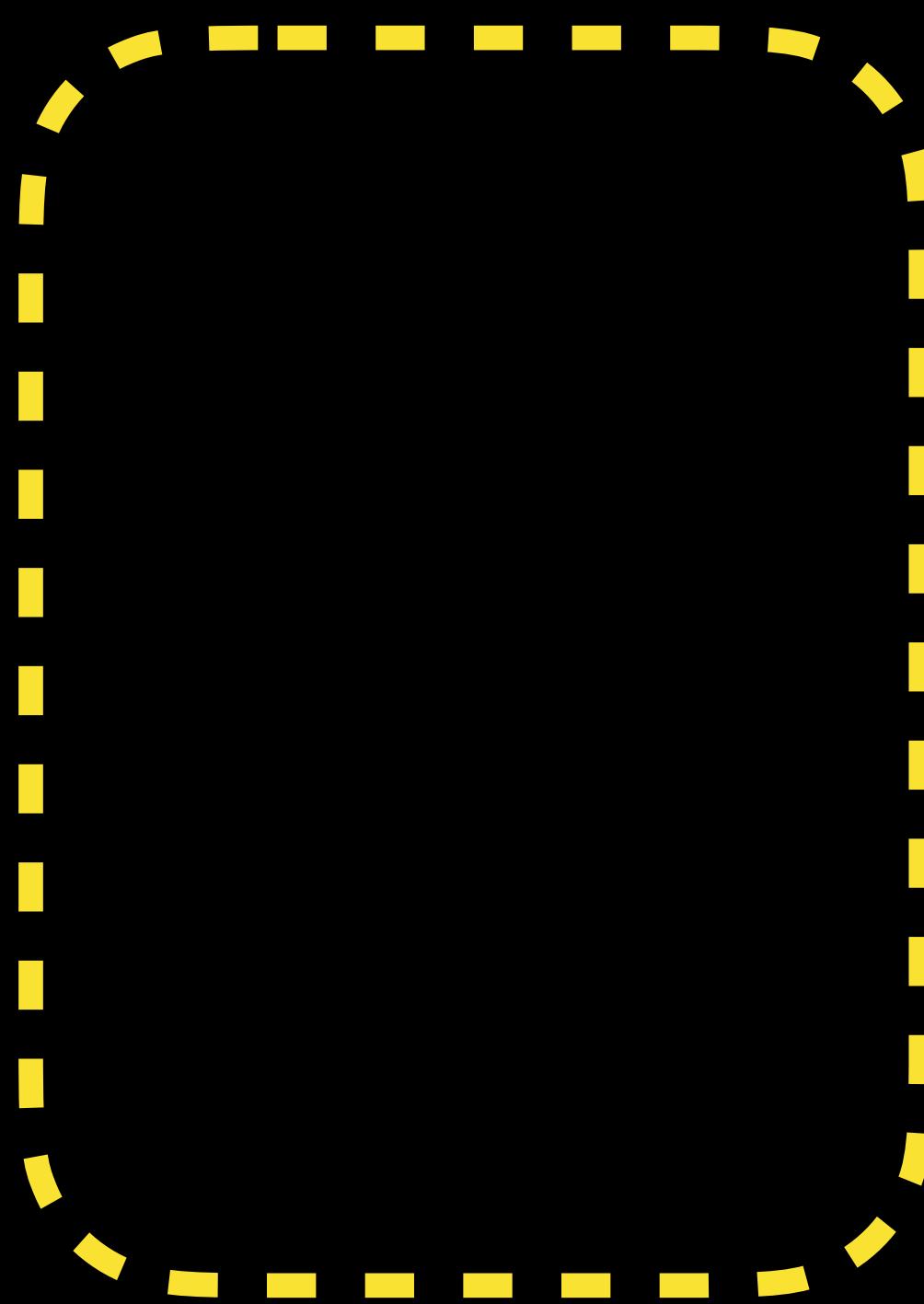
Label



Input

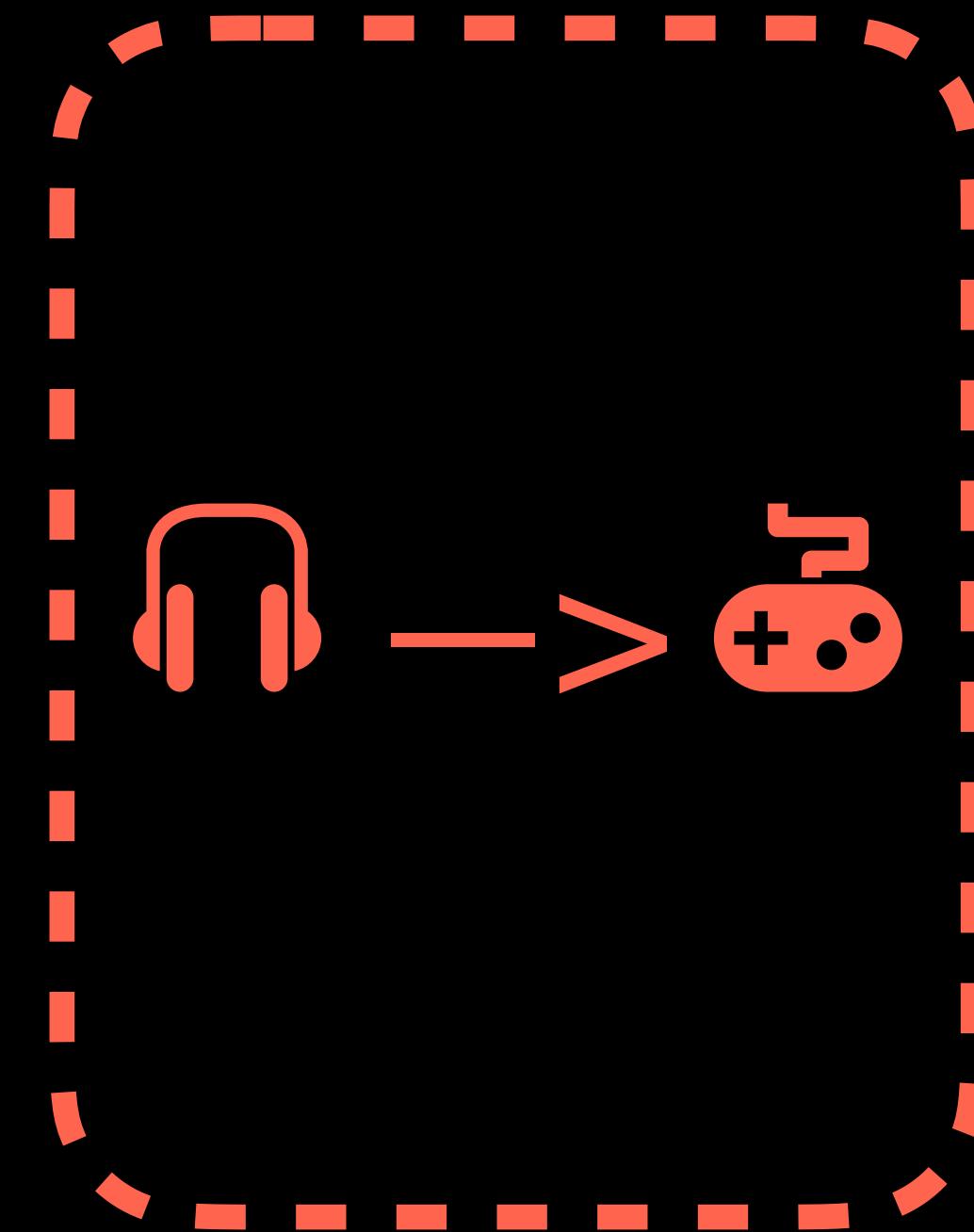


Create Recs

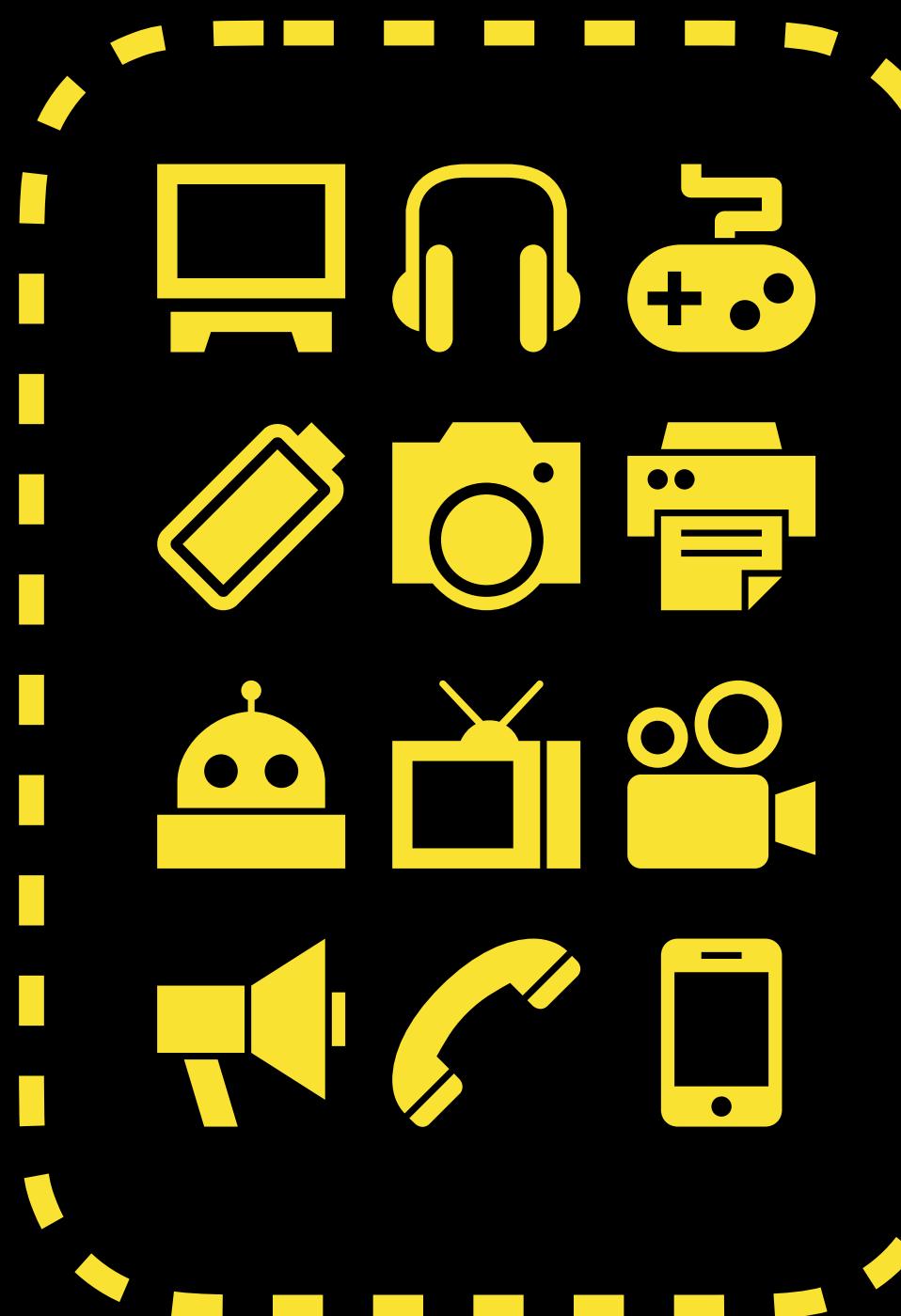


Hide this part

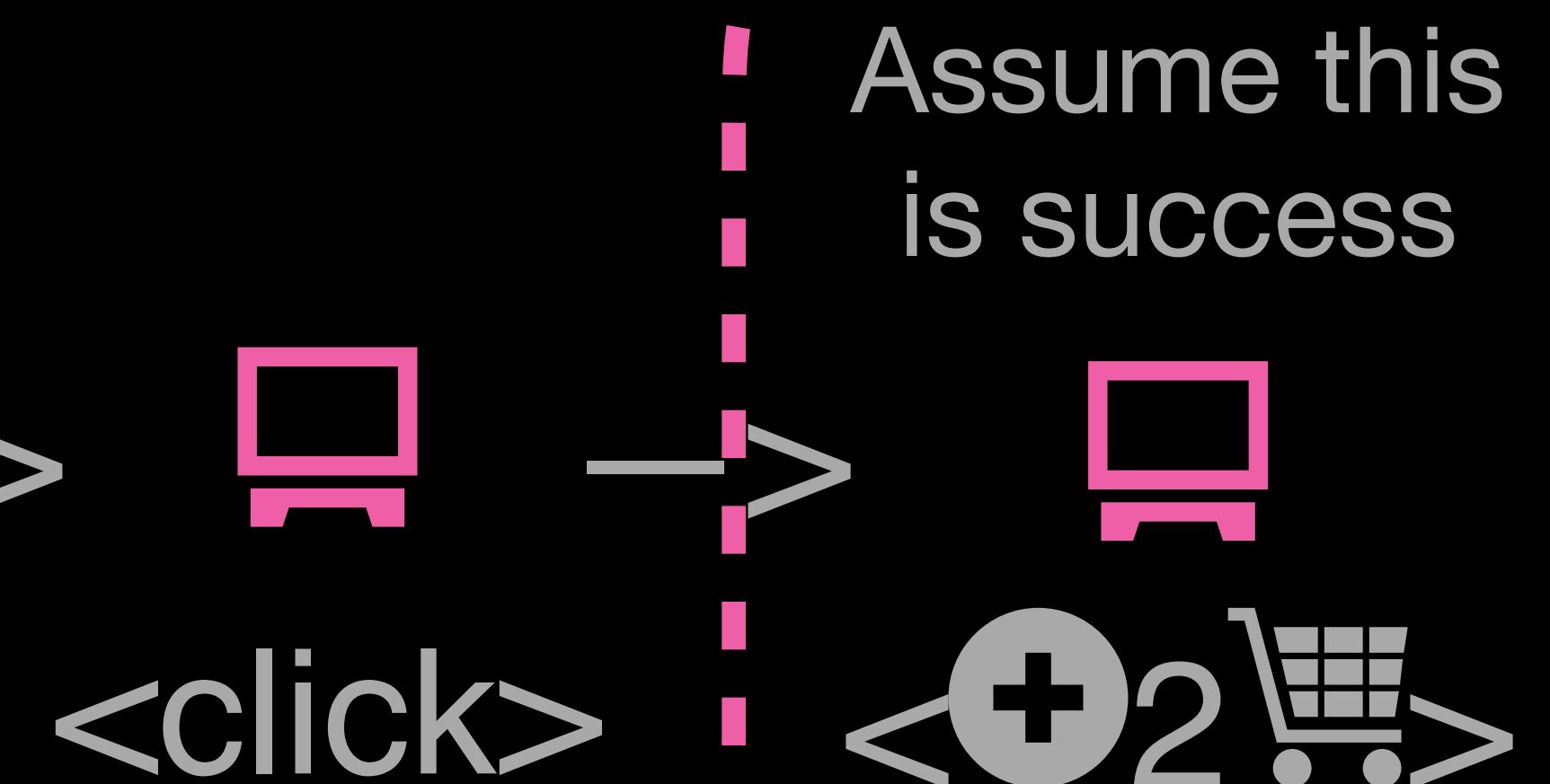
Input



Recommendation

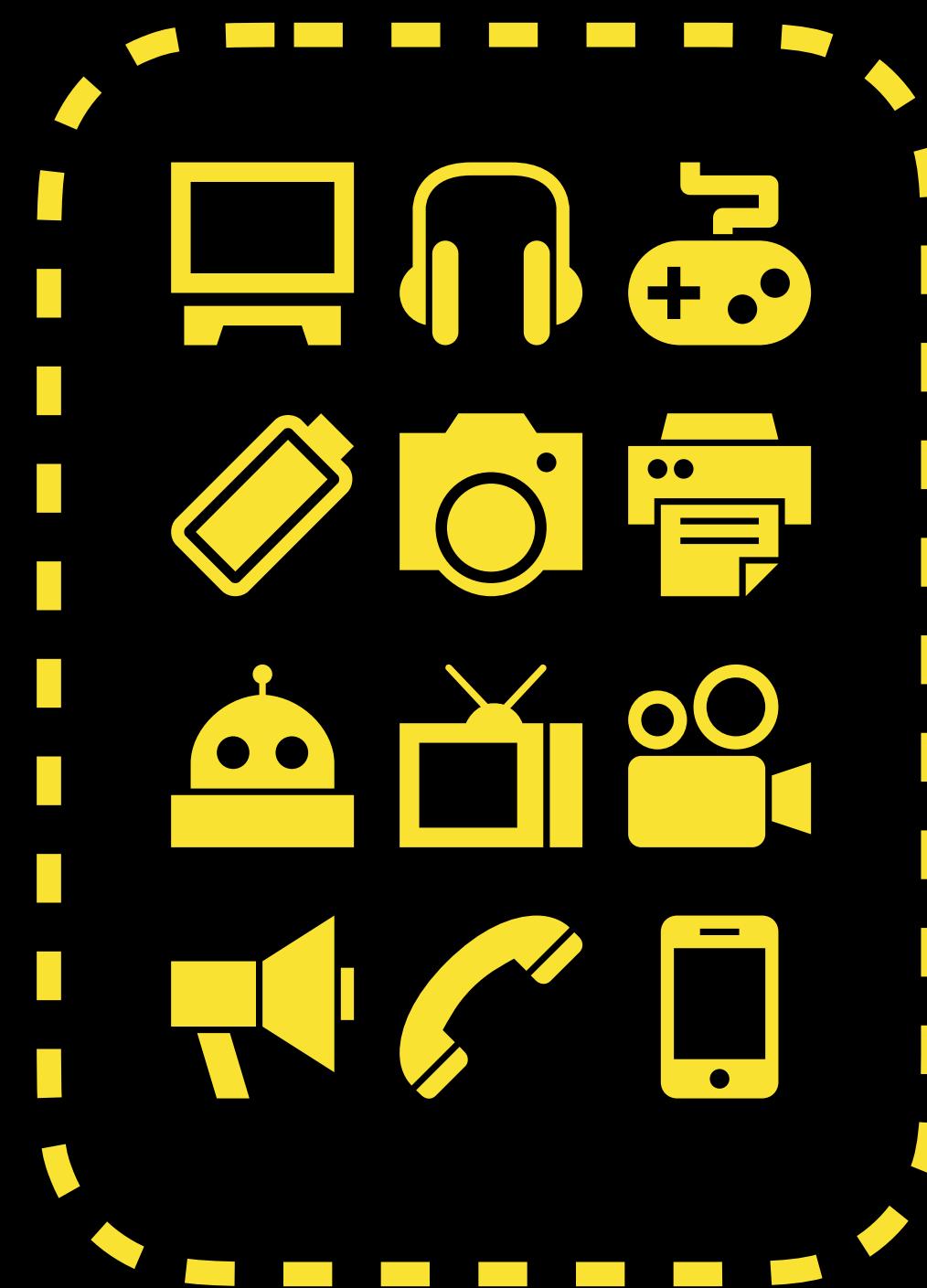


Label



**Evaluation
happens
by
comparing
the
recommen-
dation with
the historic
outcome**

Recommendation



→ **<click>**

Label

Assume this
is success





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

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

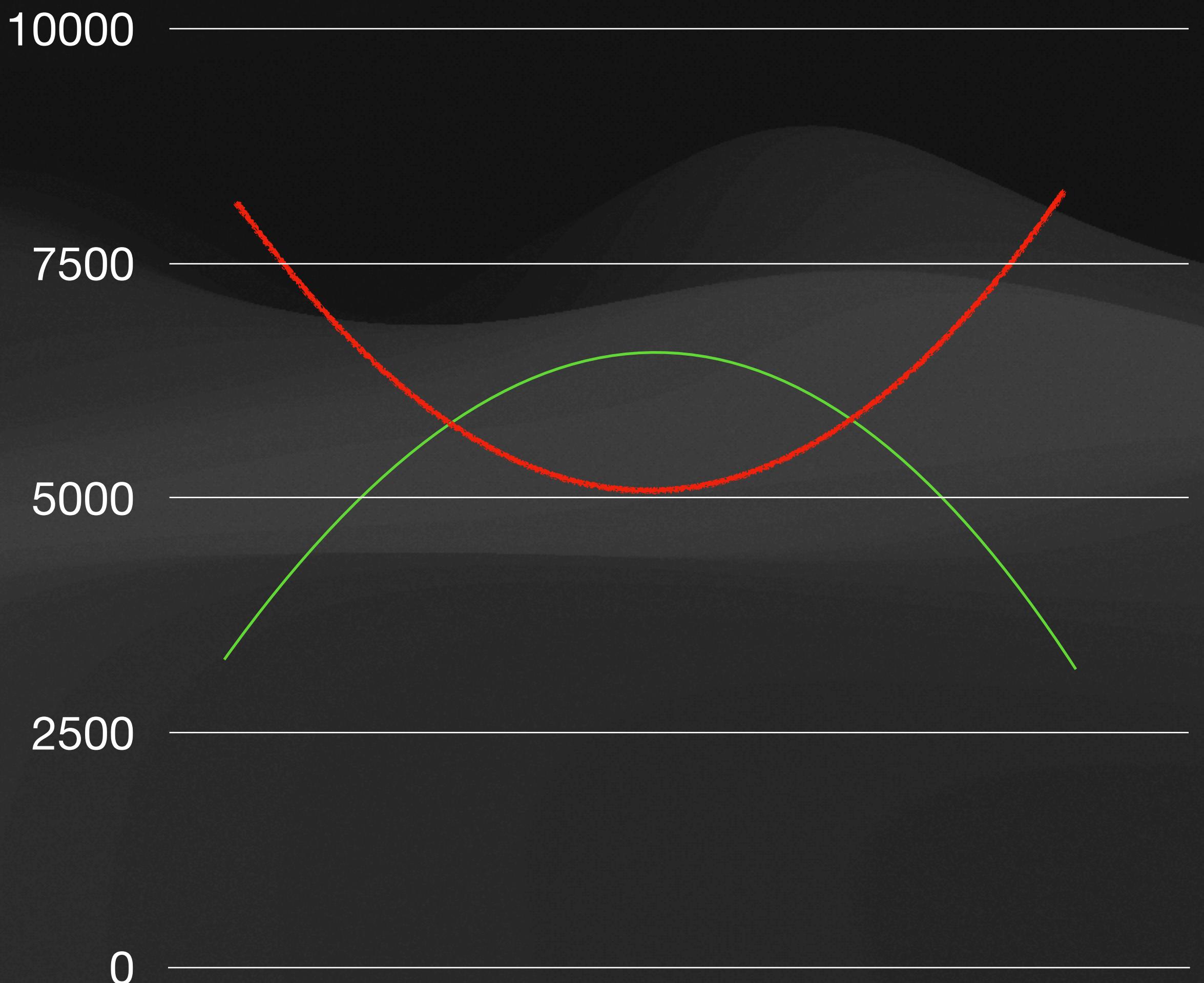
Hit metric

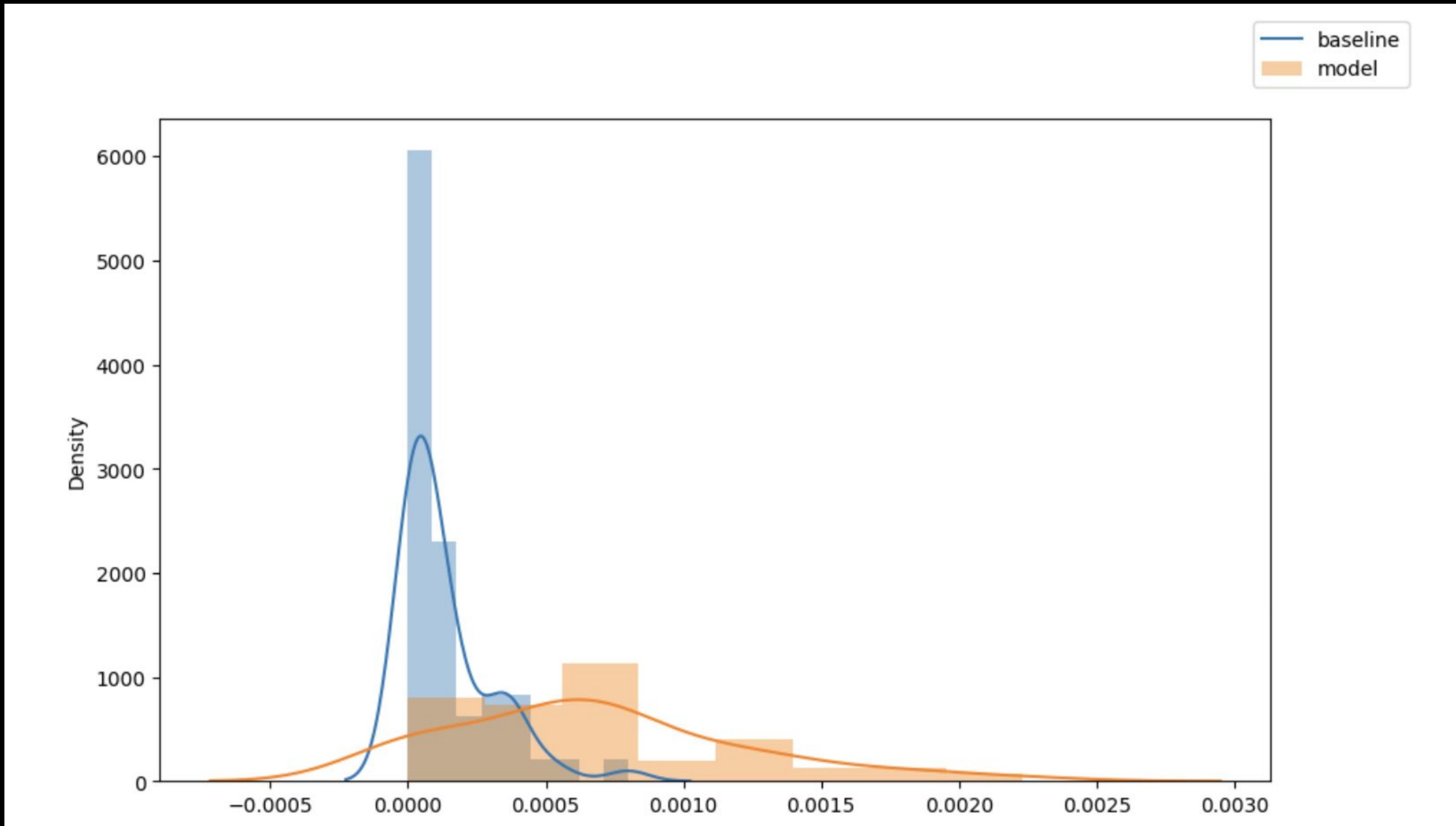
?

Diversity

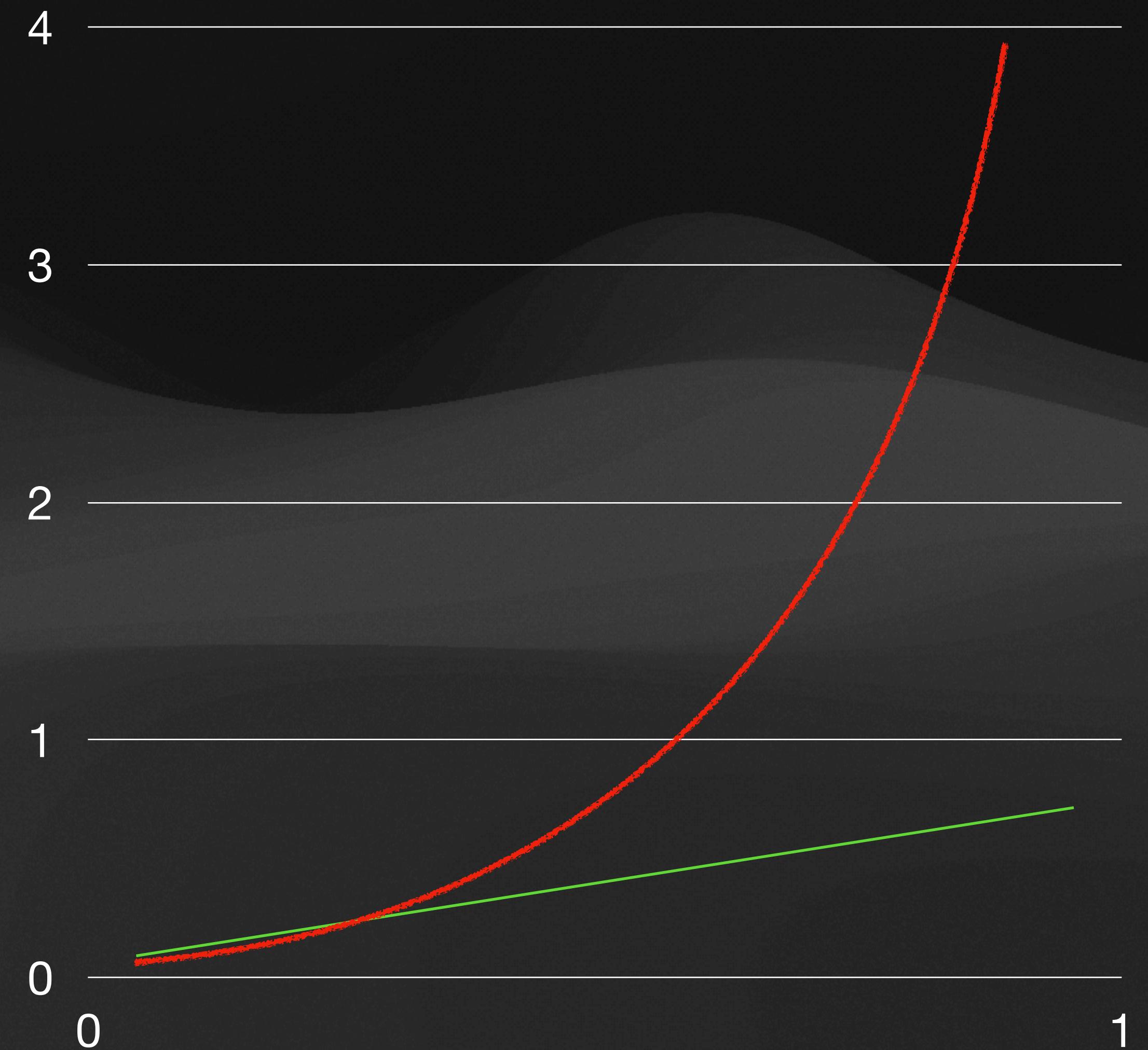
The one number often represents a distribution

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





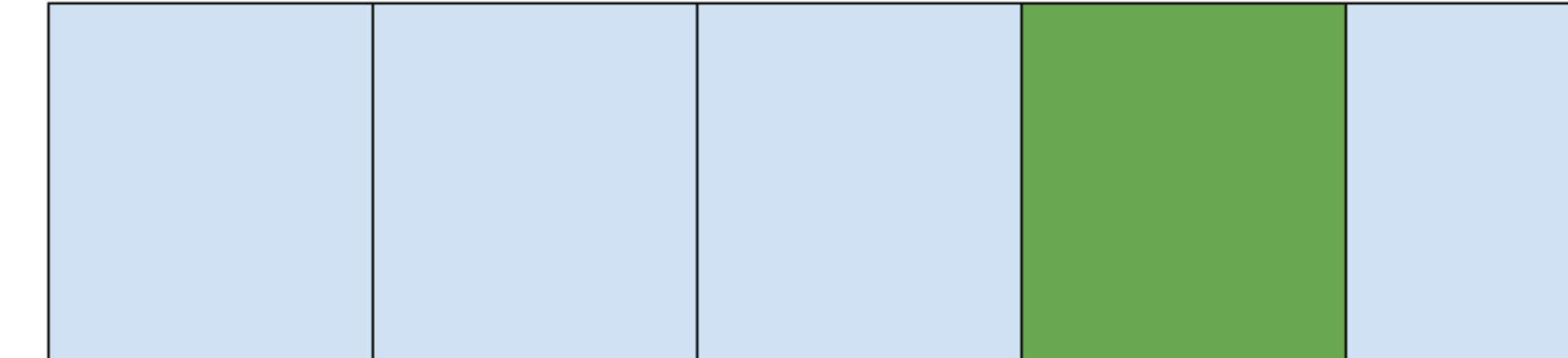
Understanding the metric



A few examples
to consider

Which is better ?

REC



REC



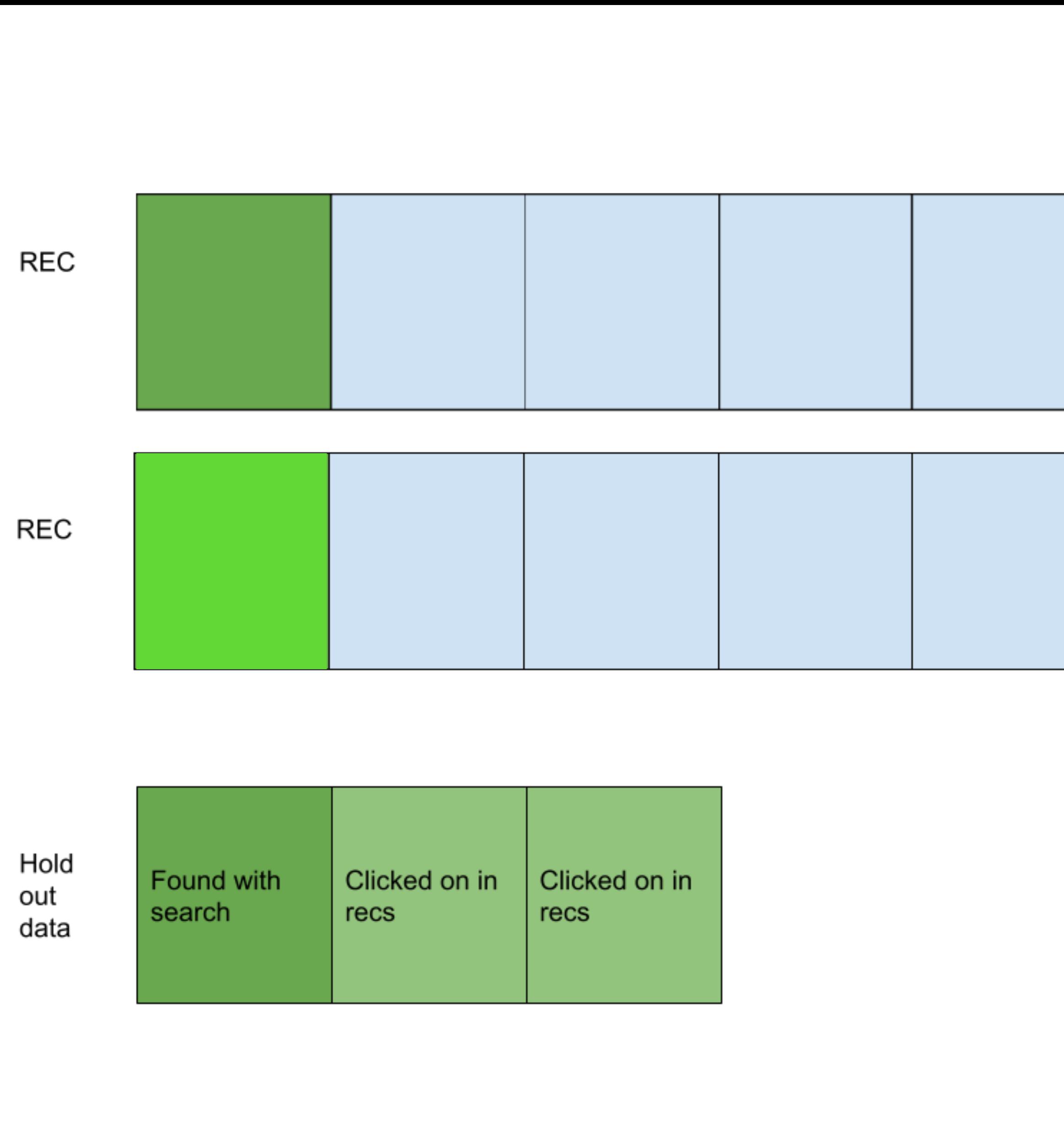
✓

Hold
out
data

Found with search	Clicked on in recs	Clicked on in recs
-------------------	--------------------	--------------------

Which is better ?

Same position but different popularity.



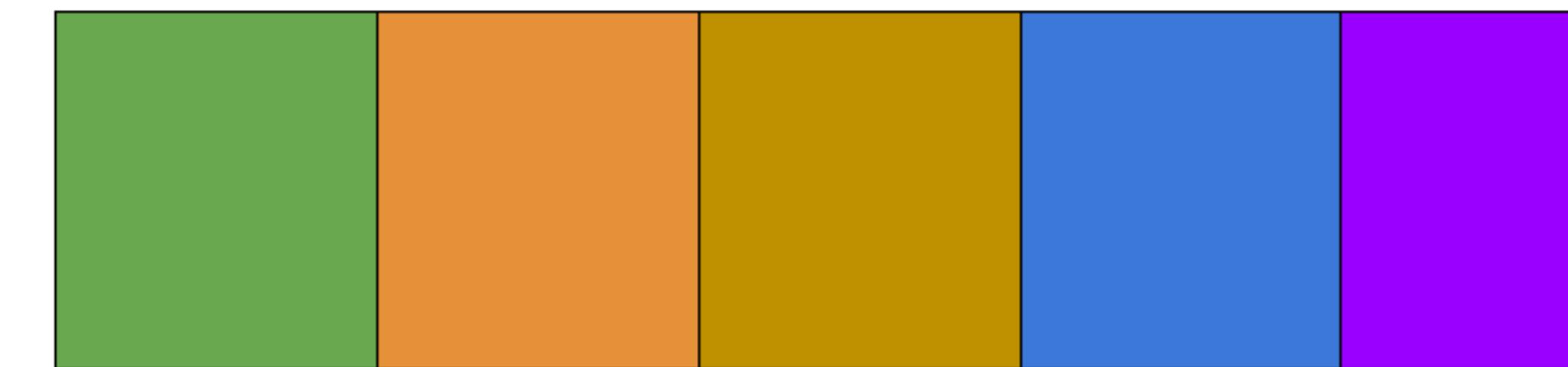
✓

Which is better ?

REC

Clicked on in recs	Clicked on in recs	Clicked on in recs		
-----------------------	-----------------------	-----------------------	--	--

REC



✓

Hold
out
data

Found with search	Clicked on in recs	Clicked on in recs
----------------------	-----------------------	-----------------------

Watch out what you optimise for

Clicks vs time on site.

The news feed includes the following headlines:

- Russisk aktiehandel åbnet
- 'Risikoen for kup vokser'
- Russer tjener på dansk megaprojekt
- Endelig er den her
- Det må ikke ende sådan
- 'Så bliver det svært'
- Køber villa for 27 mio.

Other visible elements include a sidebar for "nemlig.com" with a discount offer, a small video player, and several smaller images and captions related to current events.

The news feed includes the following headlines:

- Krig i Europa
- .. LIVE
350.000 civile fanget uden vand og el i vigtig havneby
- Formodet gerningsmand havde opført sig »mærkeligt«: Skoleelever øvede musical, da han angiveligt angreb med økse, hammer og kniv
- 18-årig anholdt: To kvinder dræbt på gymnasium i Malmö
- Folketinget sagde, at det tog magten tilbage over coronareglerne, men et år senere fortæller de nøgne tal en anden historie
- JA
- NEJ
- Amerikansk mediehus investerer millioner i dansk brillefirma
- Polen presser på for stop for handel med Rusland
- Mens bomberne faldt i titlenine, døde mere end 100 personer i en million-kilometer² et sted i Sydøstasien
- Da scenen blev syg, løb fanfaren Græsene ud for ord: »Det kunne ikke bare om overlevelsen»
- Professor i virologi: Jeg kan ikke hukse, at vi har set noget lignende her
- Med egne ord beskyldt for årsættelse af hukse: Tidligere byrådsmedlem elektroderet fra Enhedslisten

Other visible elements include a large image of a destroyed utility pole, a portrait of Volodymyr Zelensky, and a photo of two people in military uniforms.

**Ready of
online testing?**

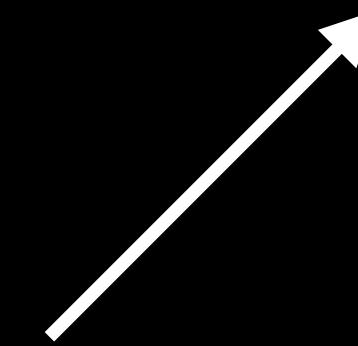
If it doesn't
work figure
out why

If the results
looks good,
assume its
a mistake

Throwing Prototypes over the wall

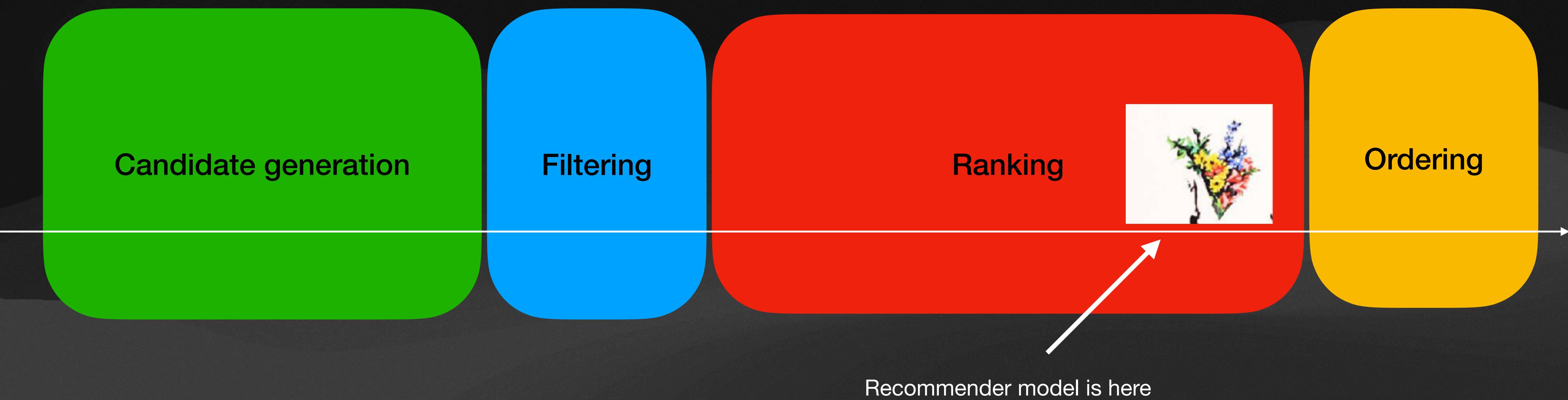


Is this the whole system?

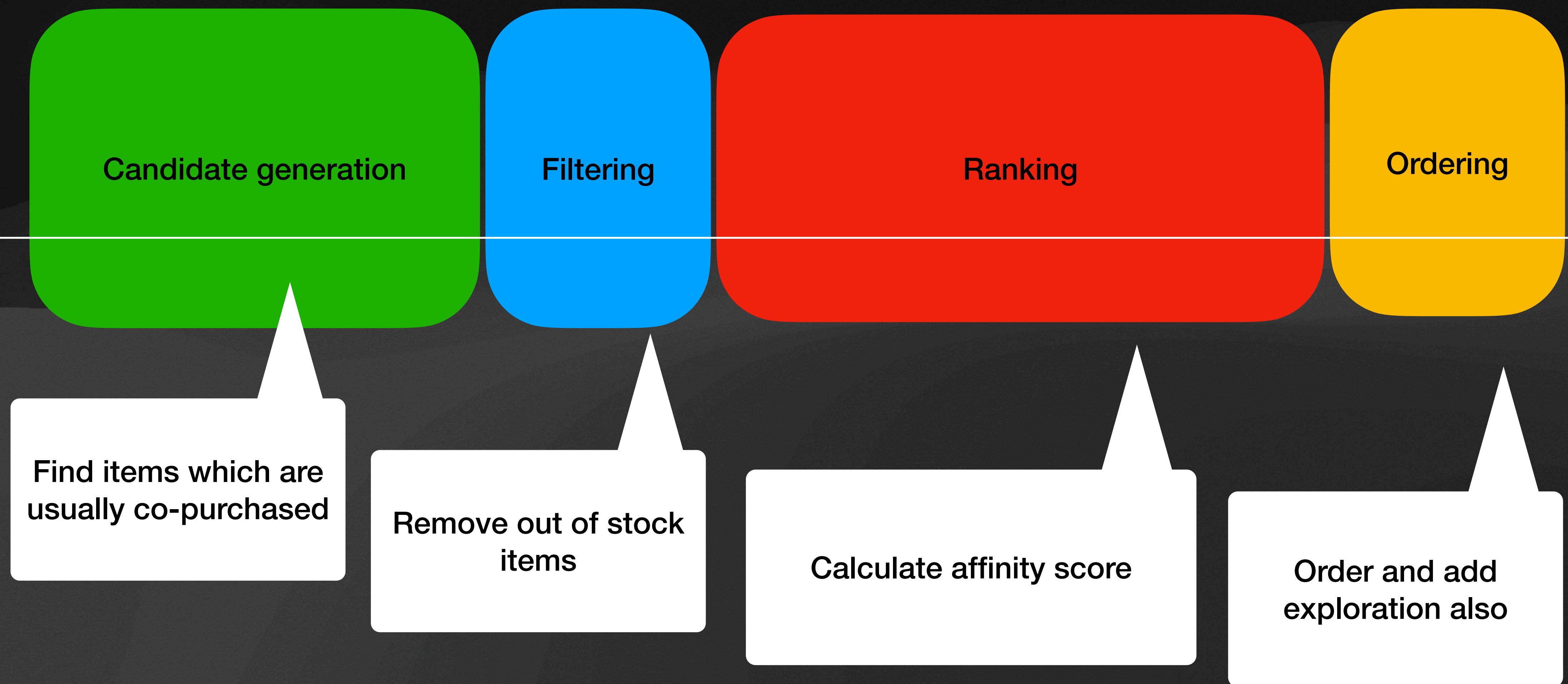


Recommender model is here

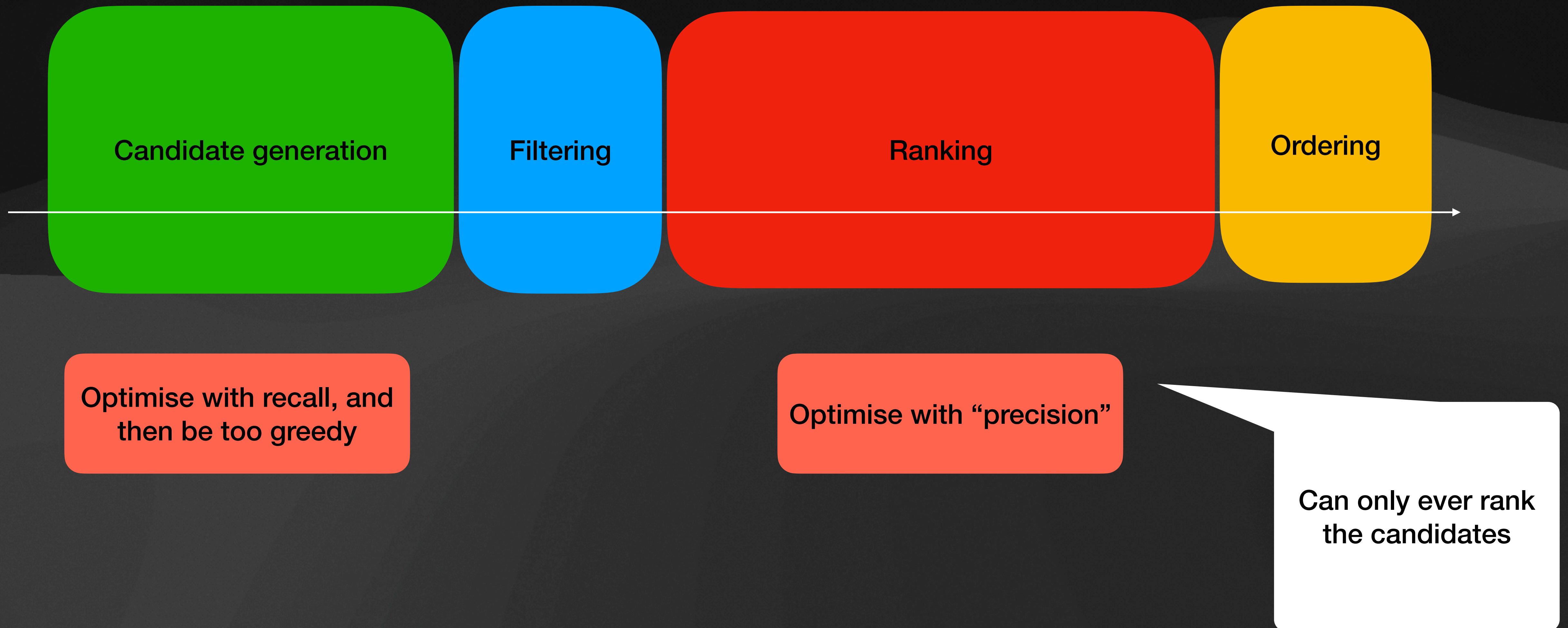
This is a Recommender System



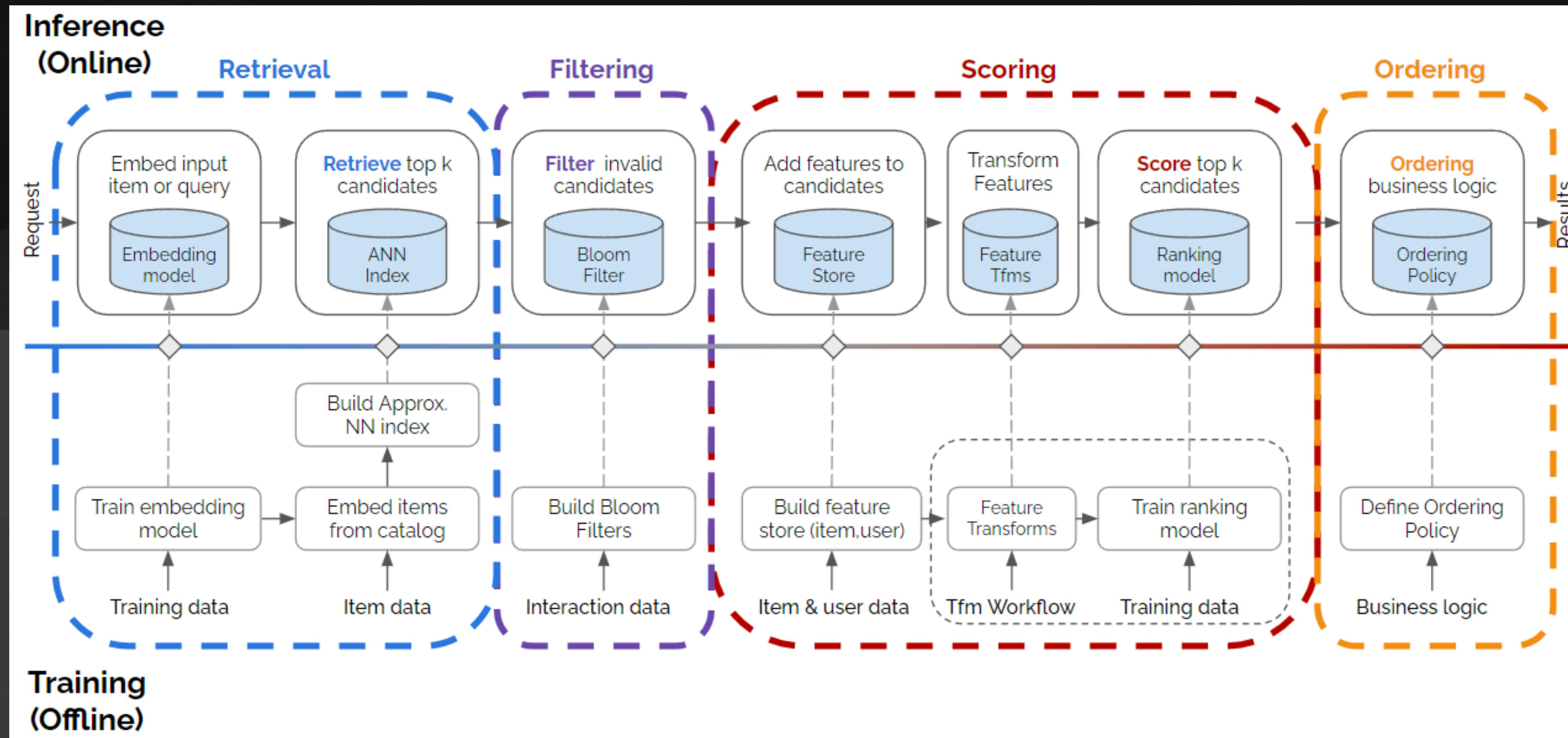
This is a Recommender System



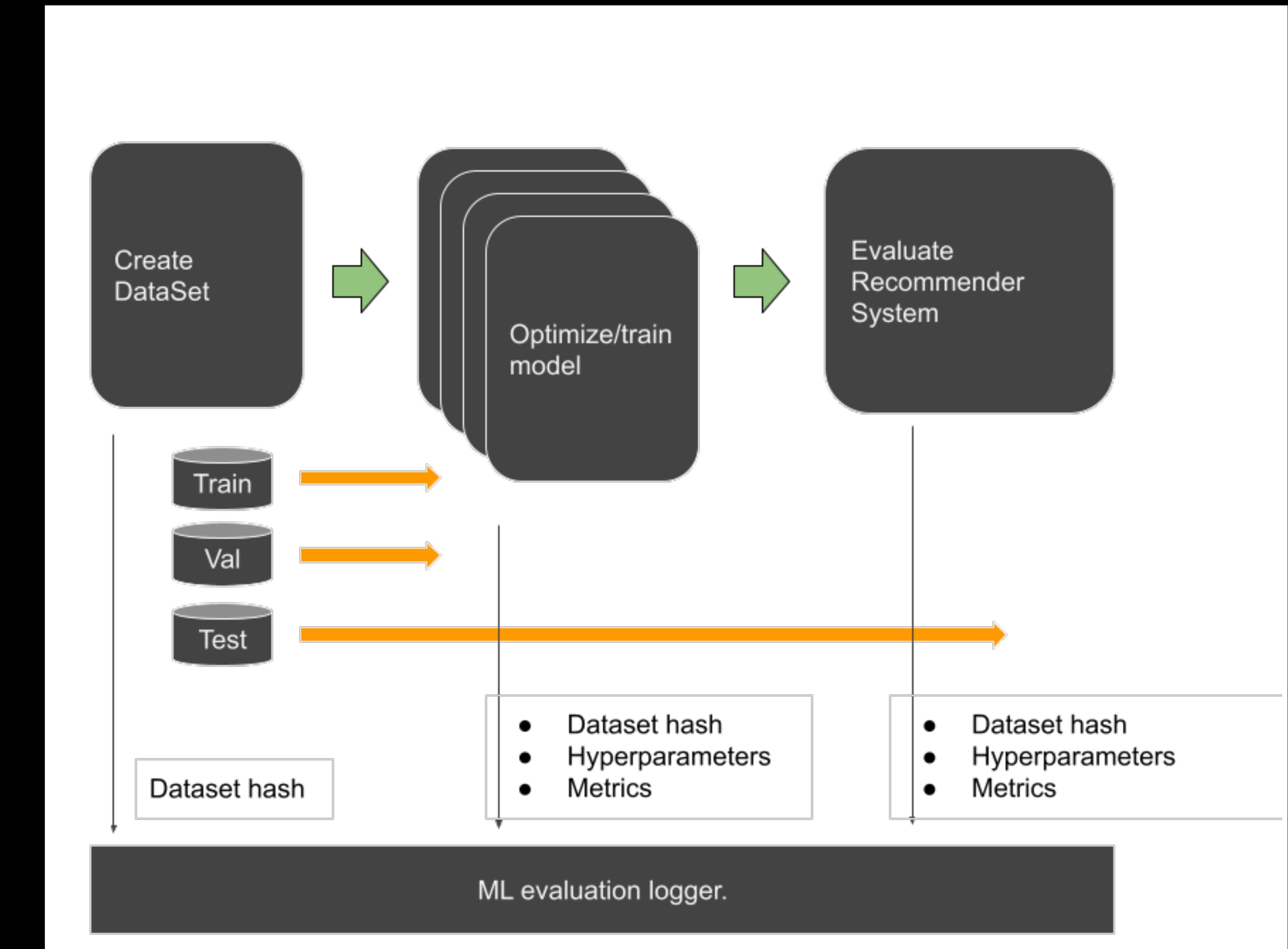
Evaluating the System



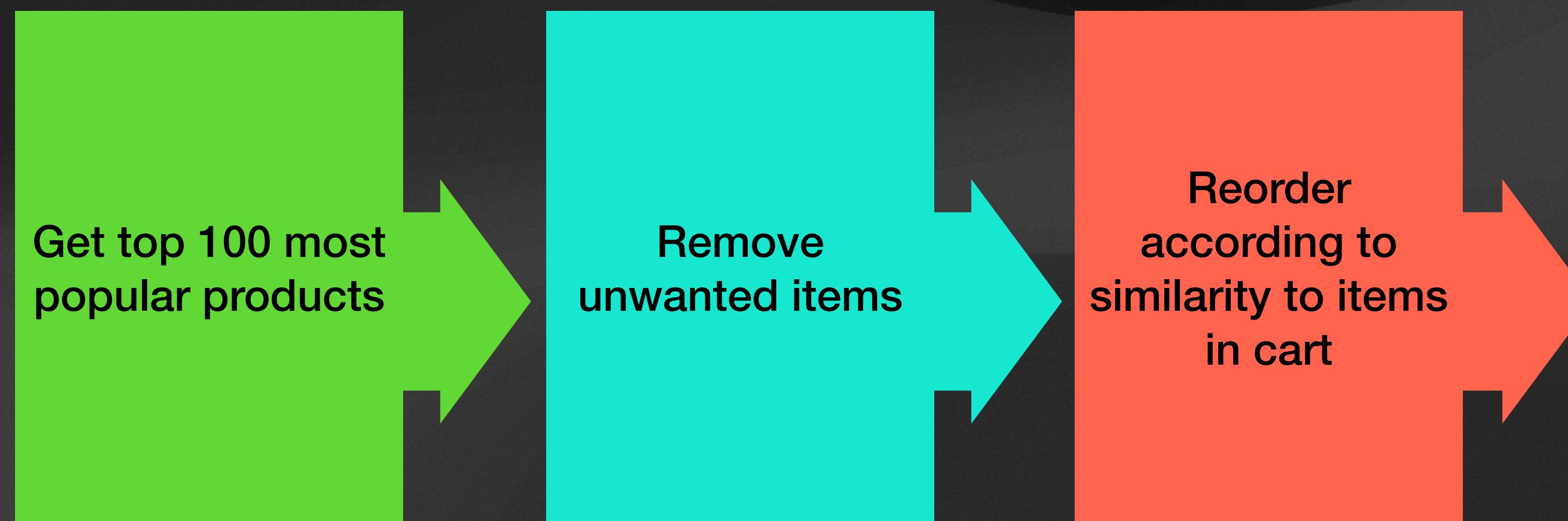
Even from Nvidia



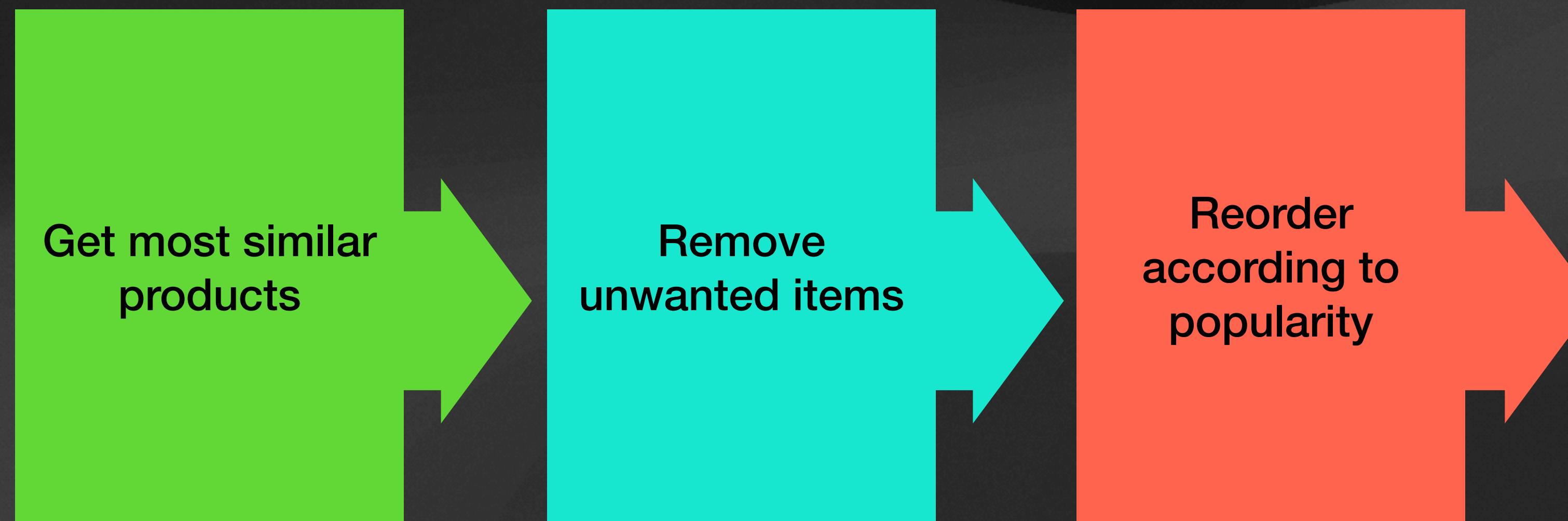
ML development framework



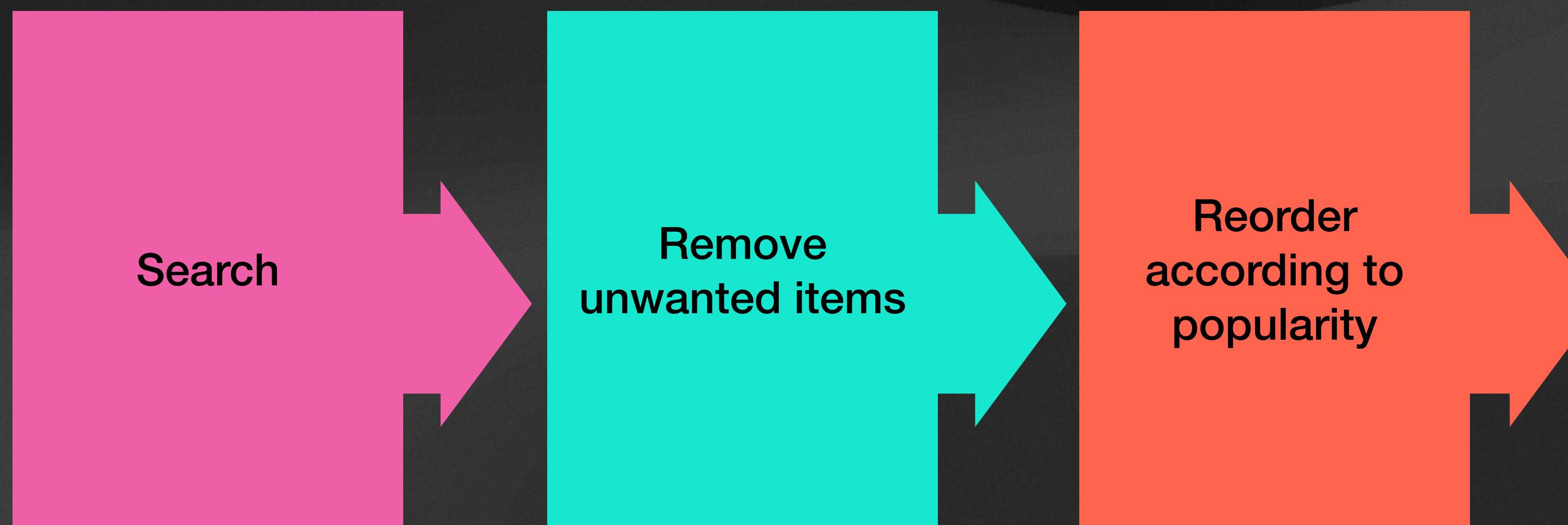
Some e-commerce examples



Some e-commerce examples

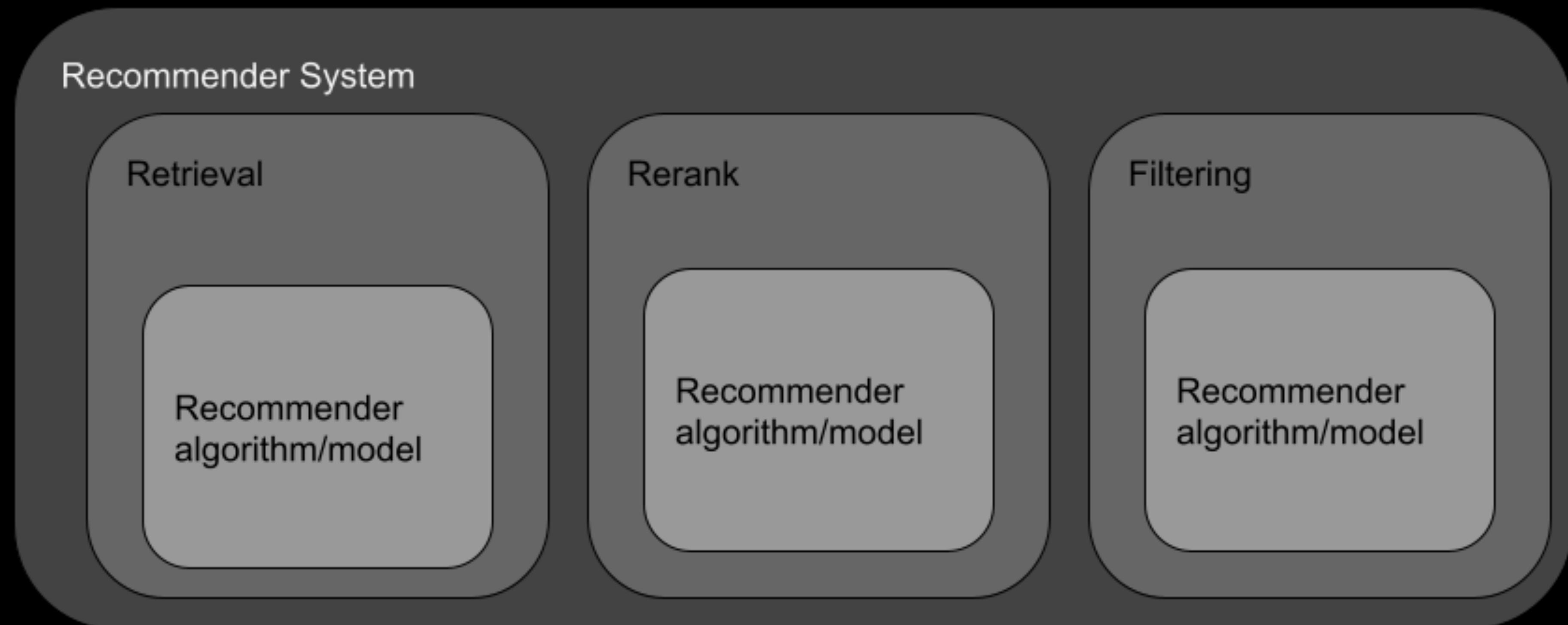


Some search example

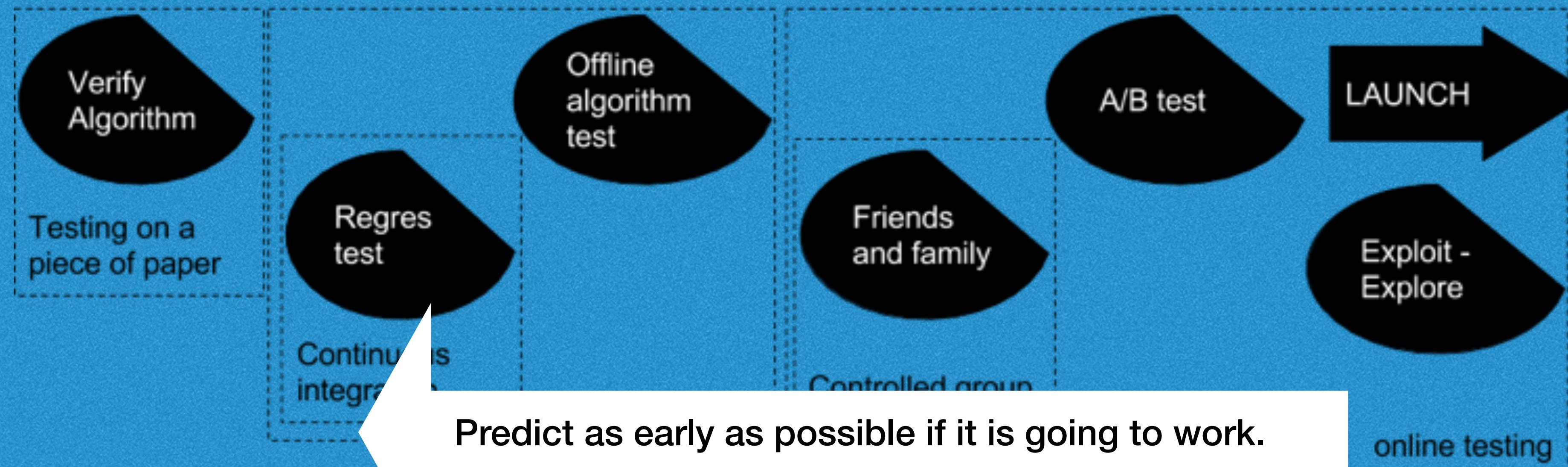


Recs models & systems

But optimised for
different things



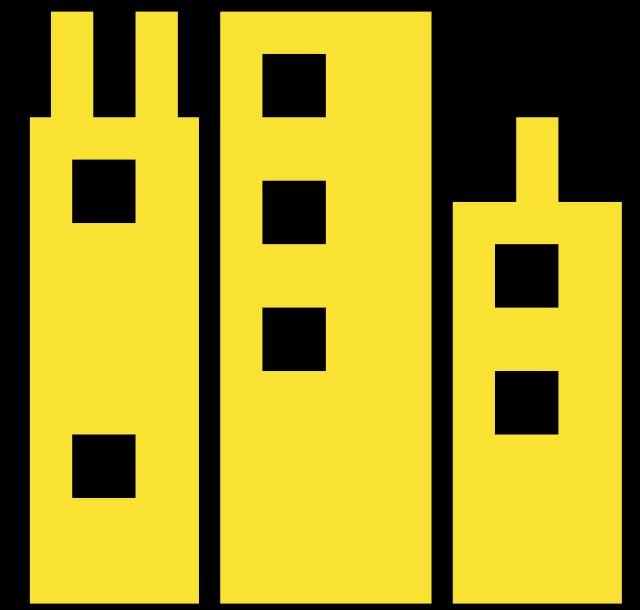
Recommender algorithm evaluation:



Involved:

Engineers

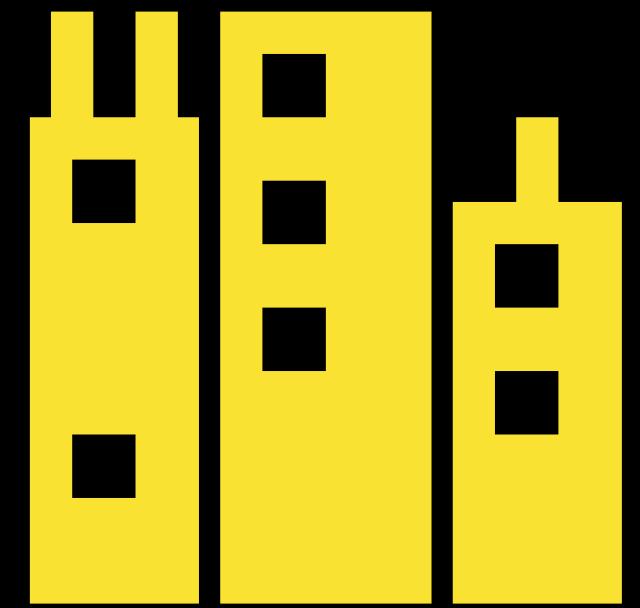
Users



We just want recommendations



No matter what you call
the algorithm

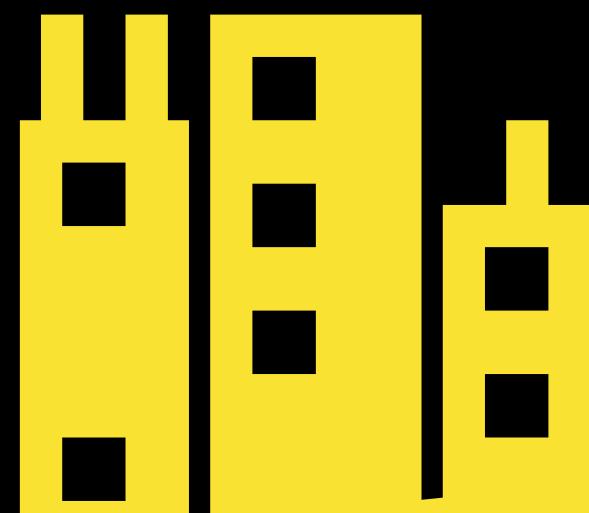


We just want recommendations



No matter what you call
the algorithm

Can we afford running
the new model?



We just want recommendations

I'm worried

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



No matter what you call
the algorithm

Can we afford running
the new model?

Consider Mixing models

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



Simply interleaving

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

FREE SHIPPING ON ORDERS ABOVE 100 EUR

MEN WOMEN KIDS WORLD OF KNOWLEDGE

SAVE 5%
ORGANIC

CABLE CREW NECK COTTON KNIT

€69,98 €139,95

COLOR Black Jet



SIZE

[ADD TO CART](#) [ADD TO WISHLIST](#)

KnowledgeCotton Apparel cable knit in a heavy and warm quality. The knit is designed with ribbed cuffs and hem.
Color: Black Jet
Quality: 100% organic Cotton
Style no.: 80665-1300

[+ SHIPPING & RETURNS](#) [+ WASH & CARE GUIDE](#)



Certified by Control Union license no.: 847594
Global Organic Textile Standards (GOTS) is the world's leading certification standard for organic textiles including both ecological and social criteria. GOTS covers every step in the production process from the fibers to the finished garment.

[READ MORE](#)

YOU MAY ALSO LIKE


COSTUM FIT CHECKED LINEN SHIRT €119,95

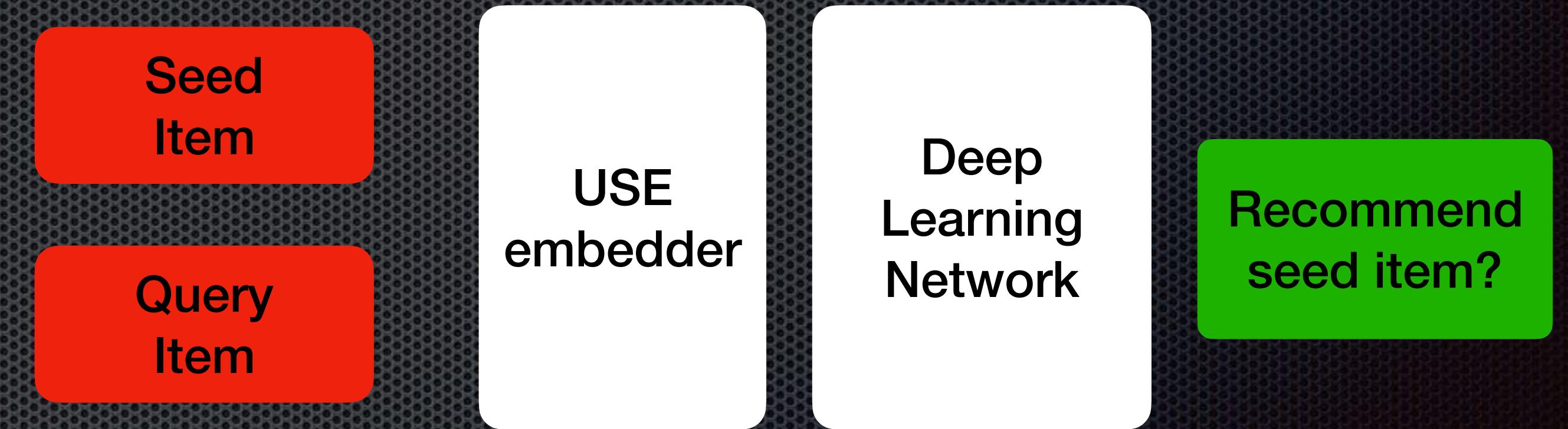

OVERSIZE CABLE KNIT VEST €199,95


VALLEY JACQUARD CREW NECK €49,98


PIXEL CHECKED CREW NECK WOOL KNIT €160,95

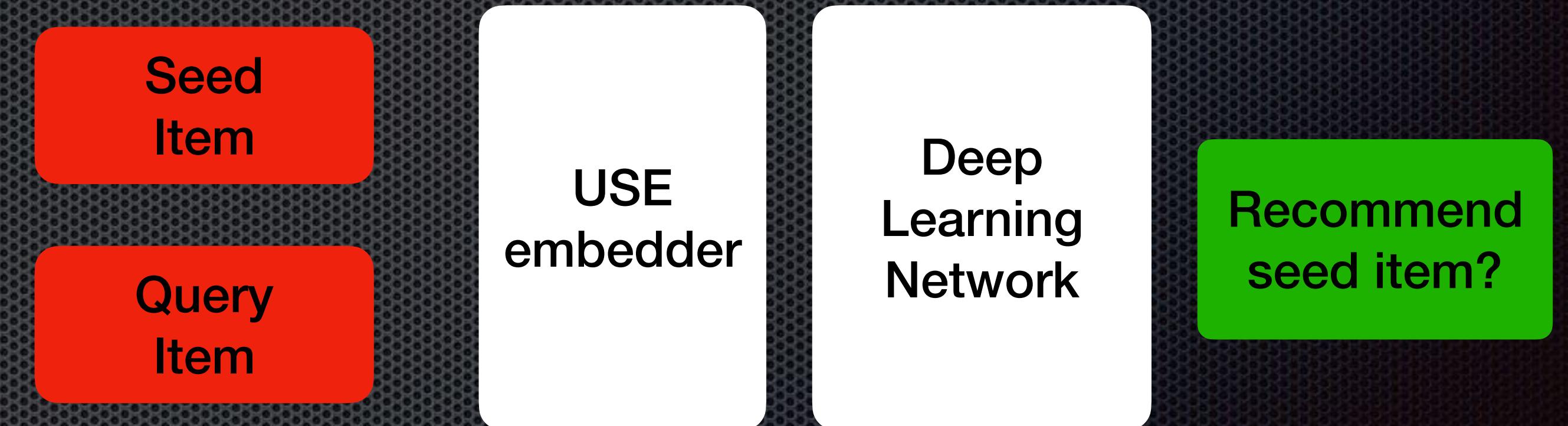
Content and behaviour

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



Content and behaviour

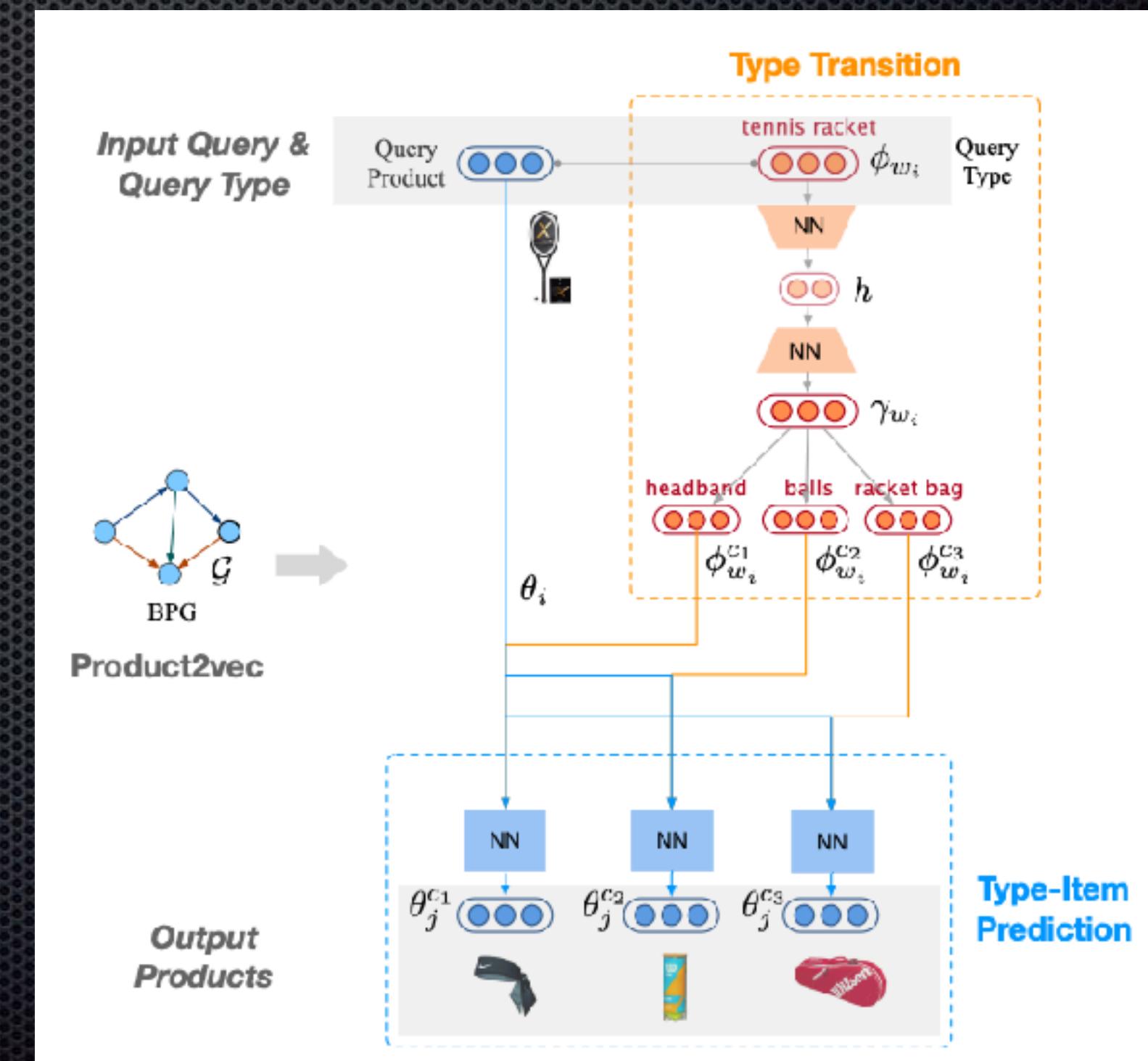
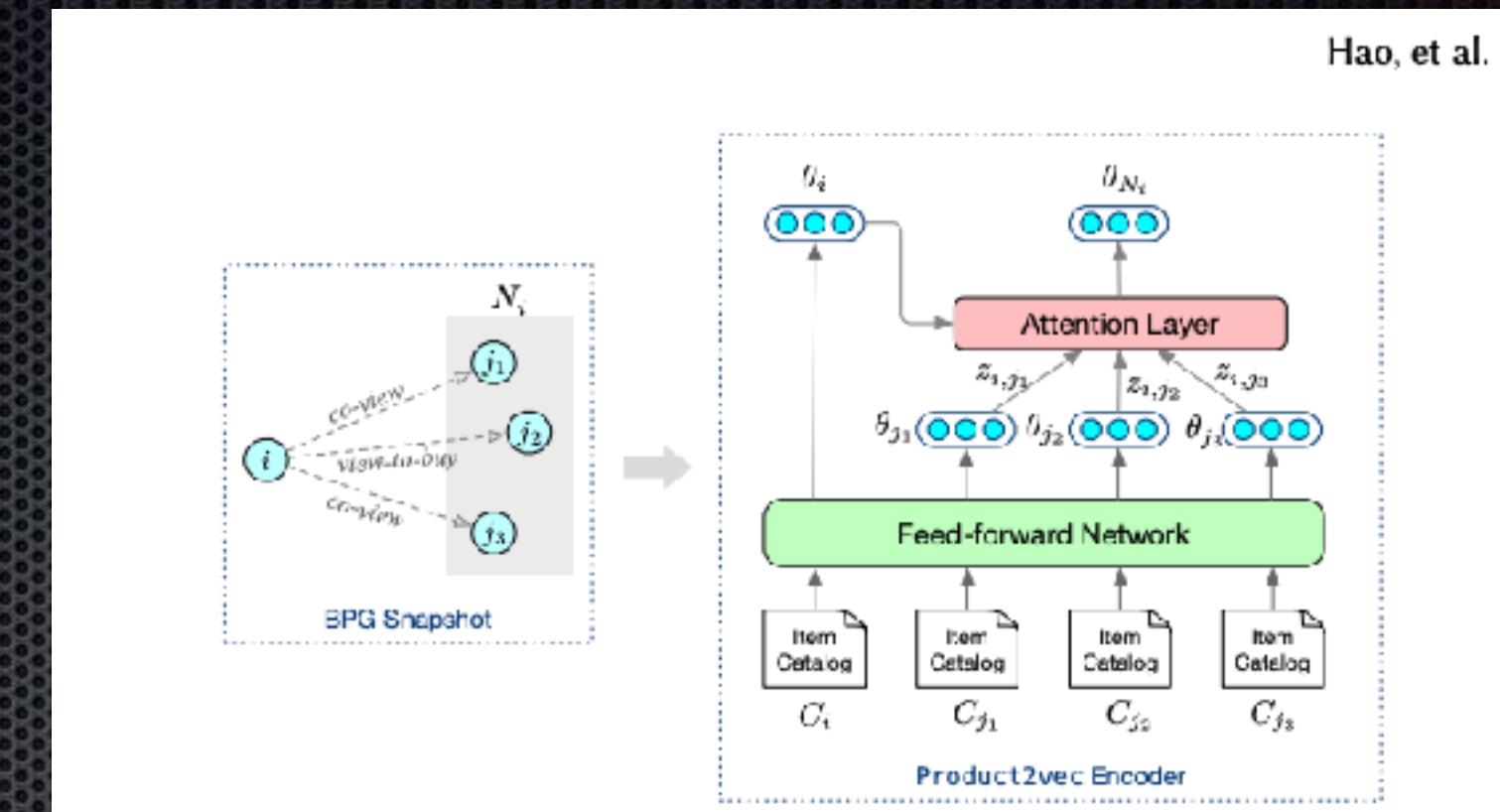
- Create embeddings of all items.
- Train a model using behavioural data to predict Probability that two items will be purchased together.



This model will also allow for new products to be recommended!

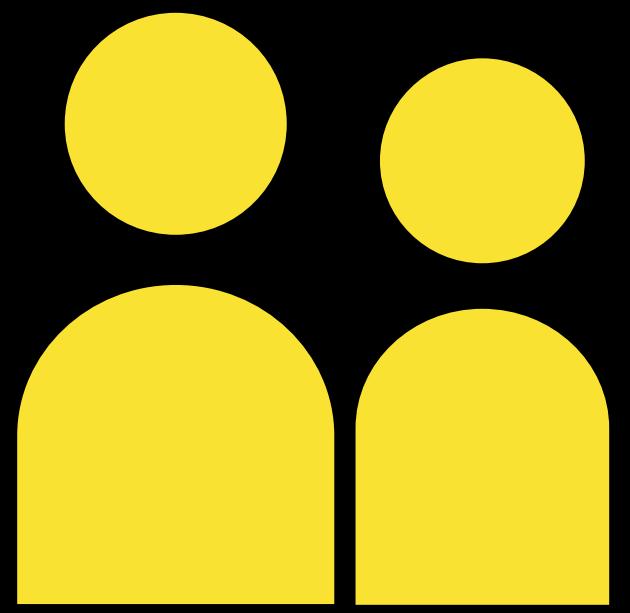
P-companion

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



Business considerations when deploying a recommender

Several aspects to consider.



Users



Business



Content creator

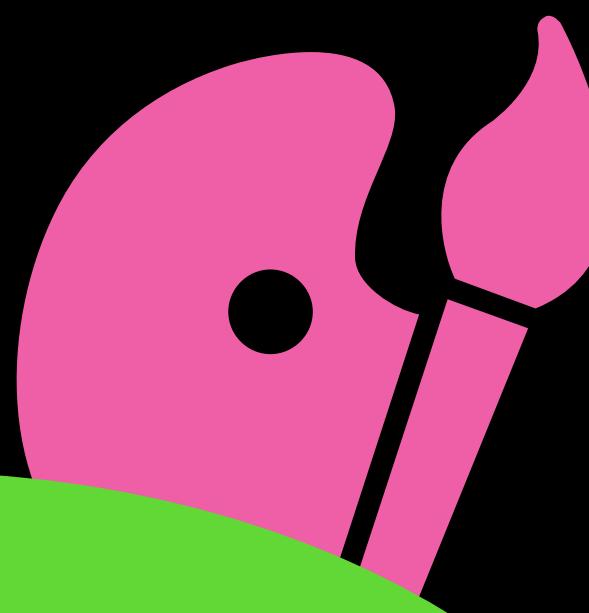
Several aspects to consider.



Users



Business



Creator

No just money, improve
user experience

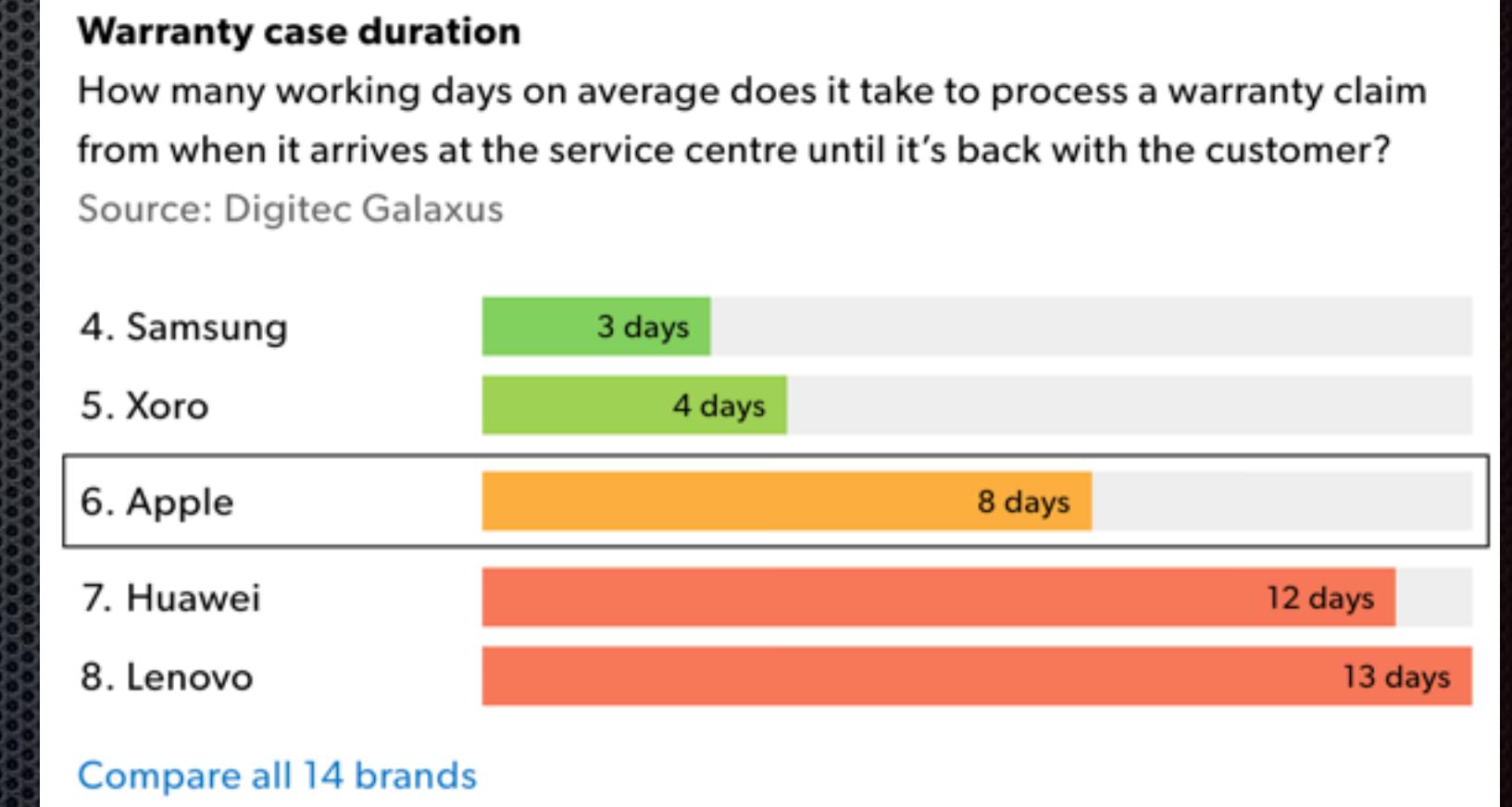
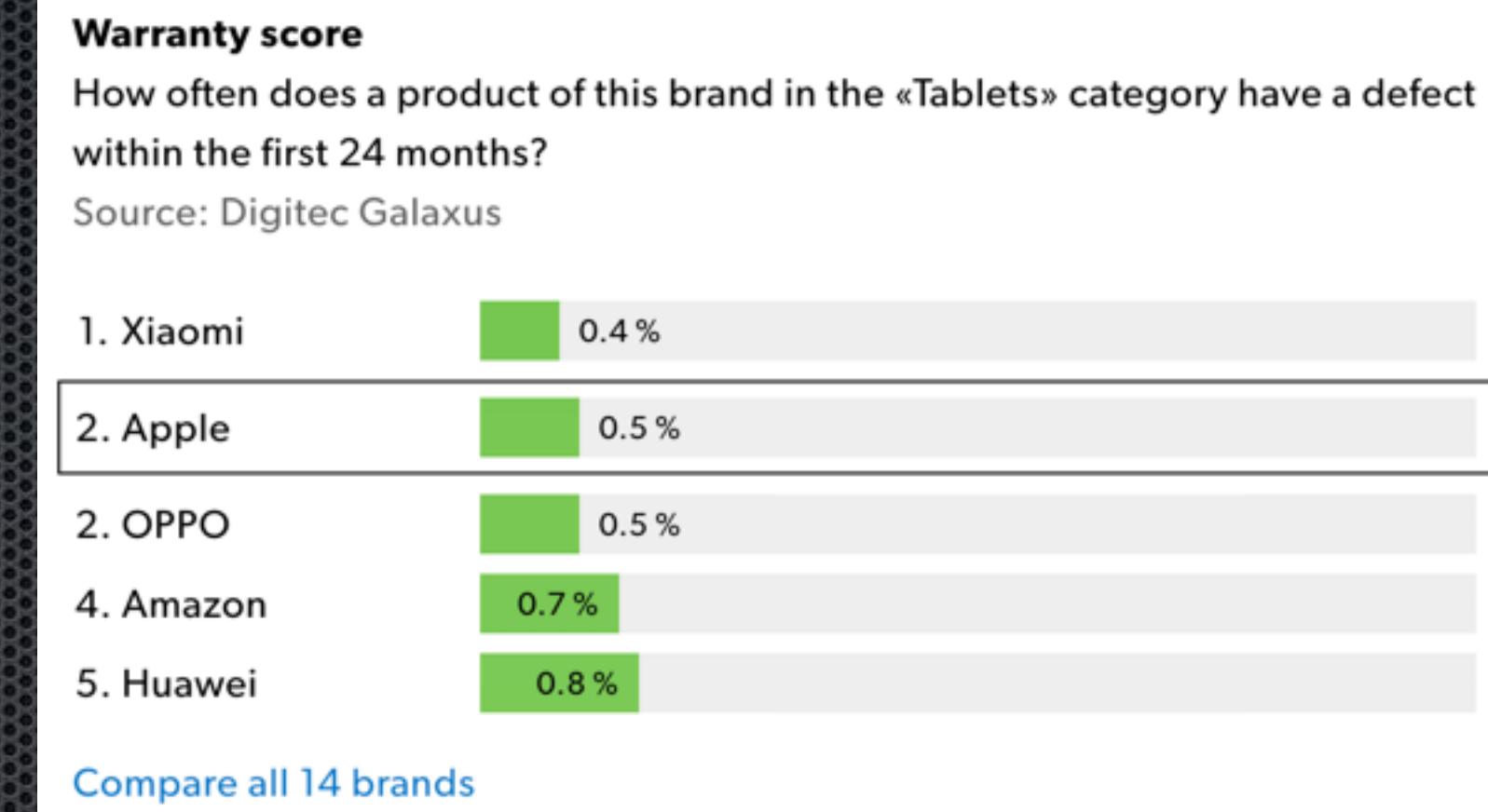
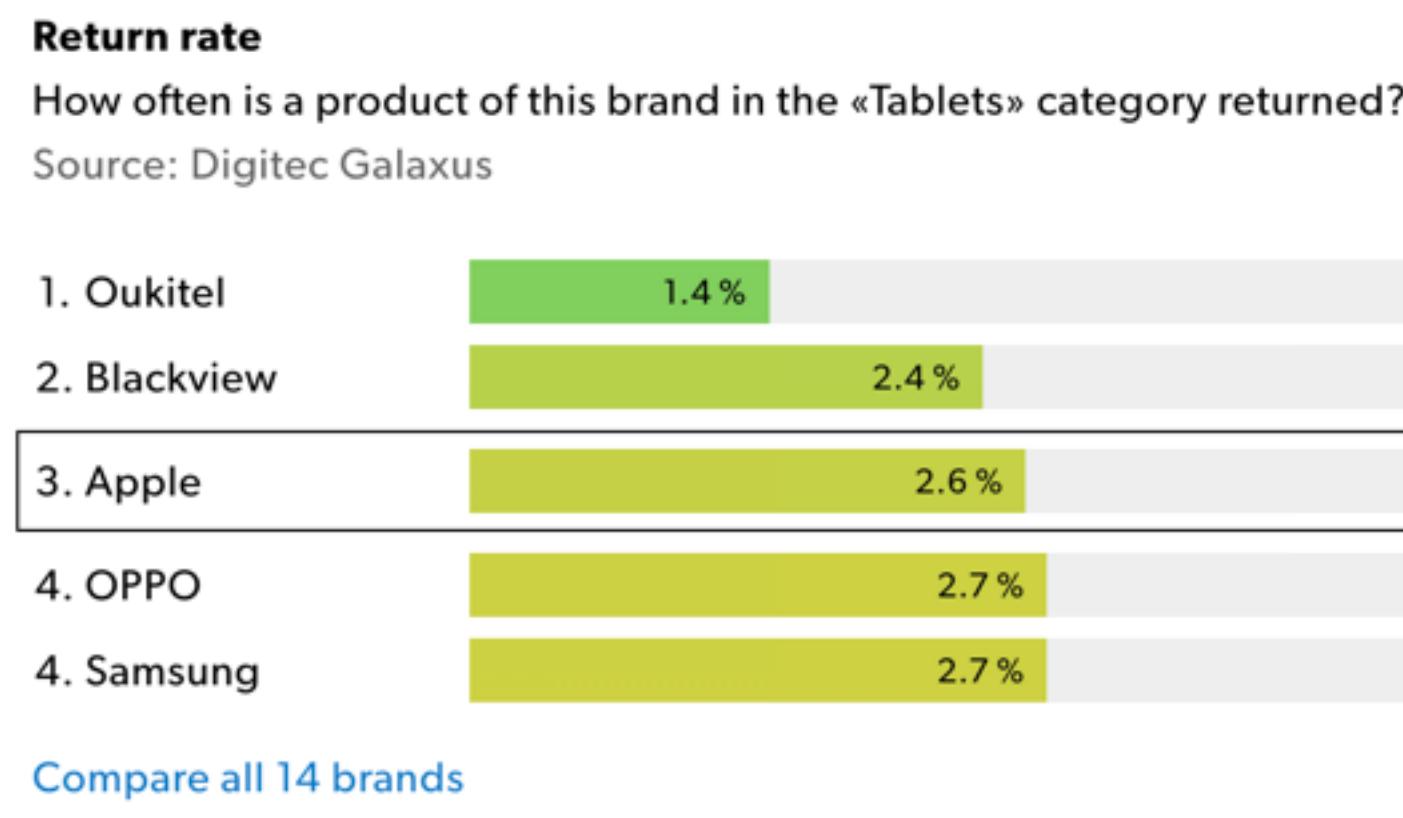
Expect to be asked to create recommendations for

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

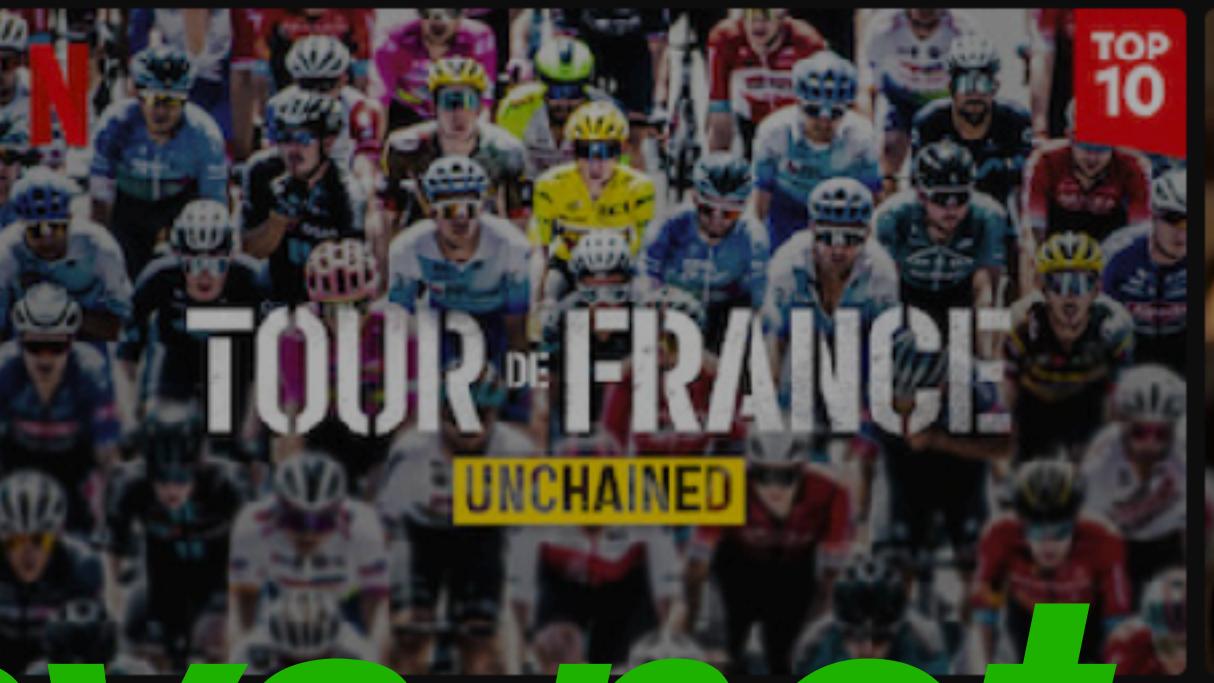
But does it make sense for the business

- * 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

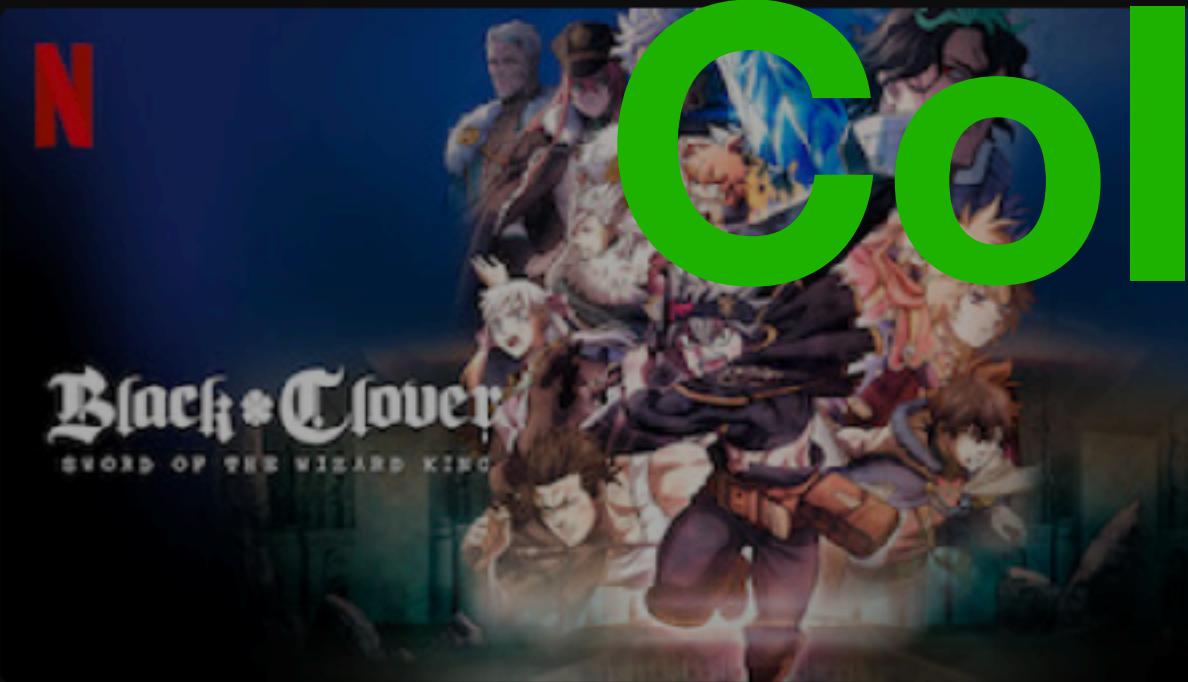


New on Netflix



(Ways not to solve) Cold start problems

Worth the Wait



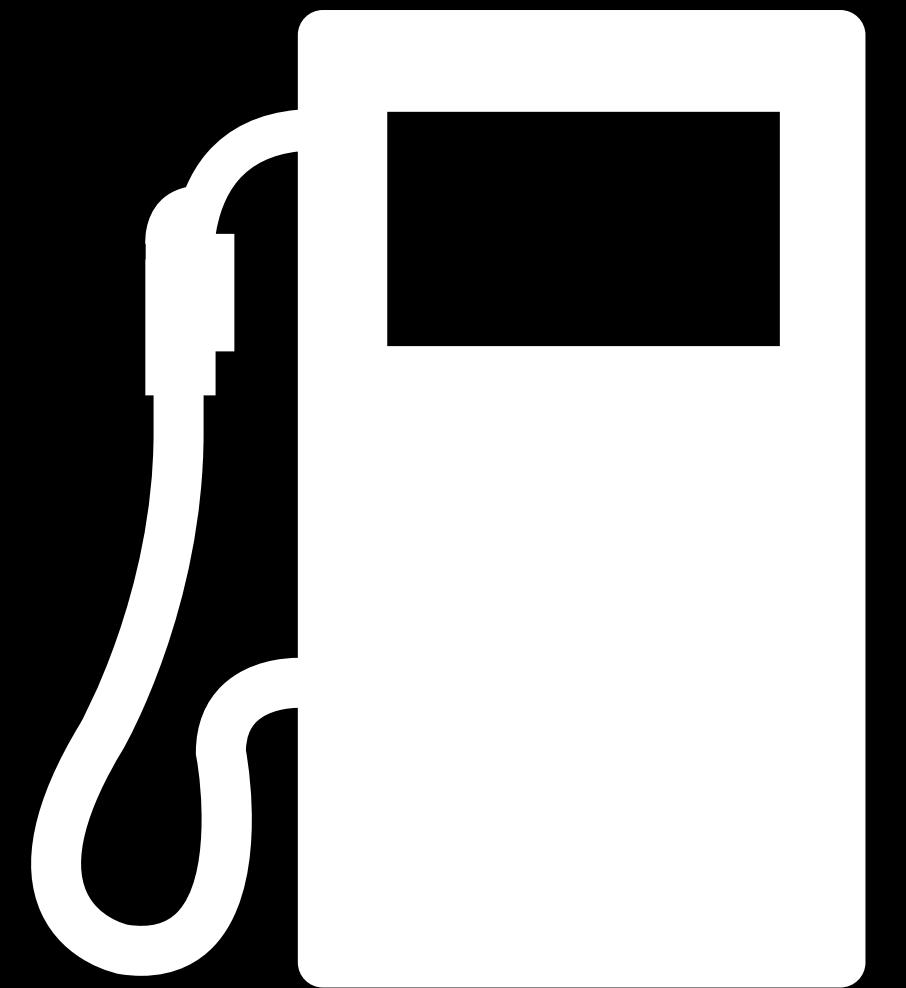
Coming this week



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 your
(meta)data is
correct

Make sure you (meta)data comes from

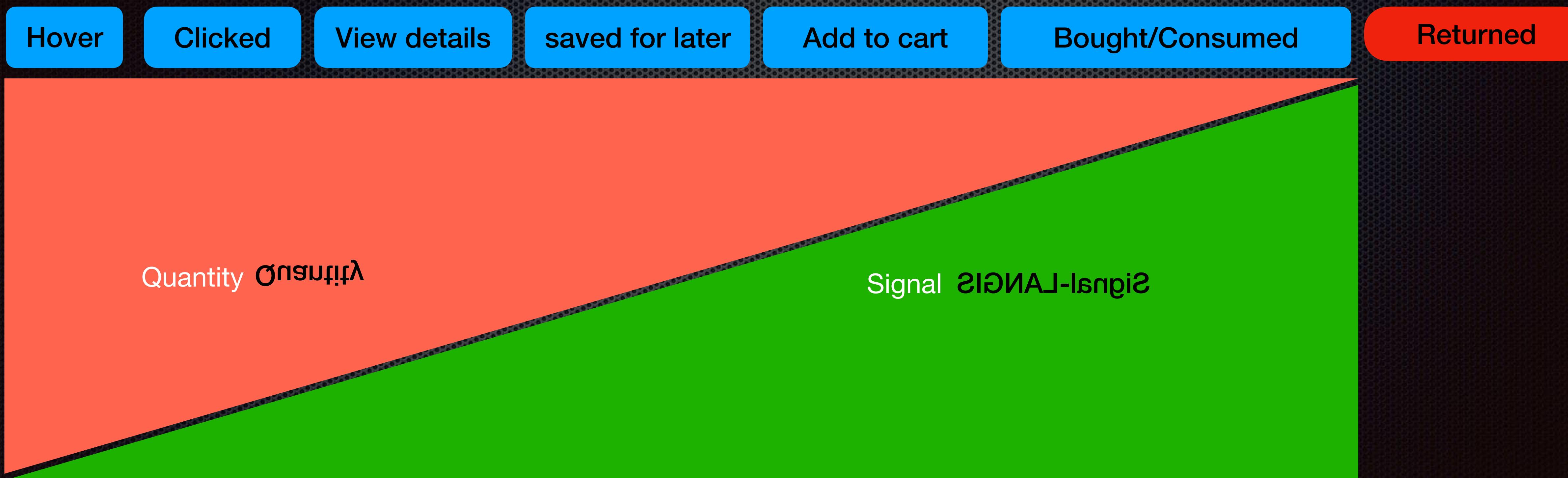


Try to avoid scenarios where your model correctly understands that a user prefers blue and then recommends red clothes because products are classified wrong.

Its hard to detect faulty classifications

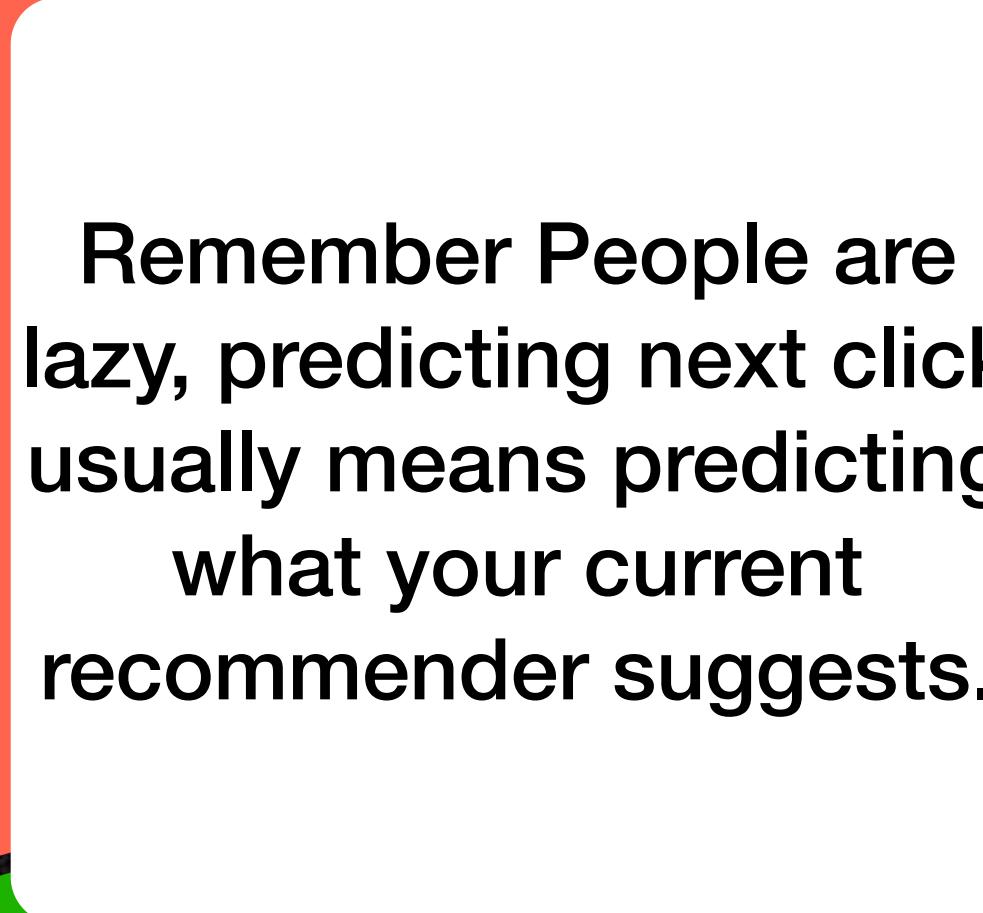
- Sampling the data and check for correctness.
- Compare different types of embeddings
 - Visual vs Text embeddings.
 - Text vs Behavioural embeddings.

Data amount vs signal



Data amount vs signal

Hover Clicked View details saved for later Add to cart Bought/Consumed Returned



Remember People are
lazy, predicting next click
usually means predicting
what your current
recommender suggests.

Signal **SiGNAL-İ-LANGİS**



Data collection?

Bandit

Organic

RecSysOps

Online Offline evaluation

continuous evaluation to ensure the model stays optimal

- Model drift
- Data drift
- Benchmark



Questions ?