

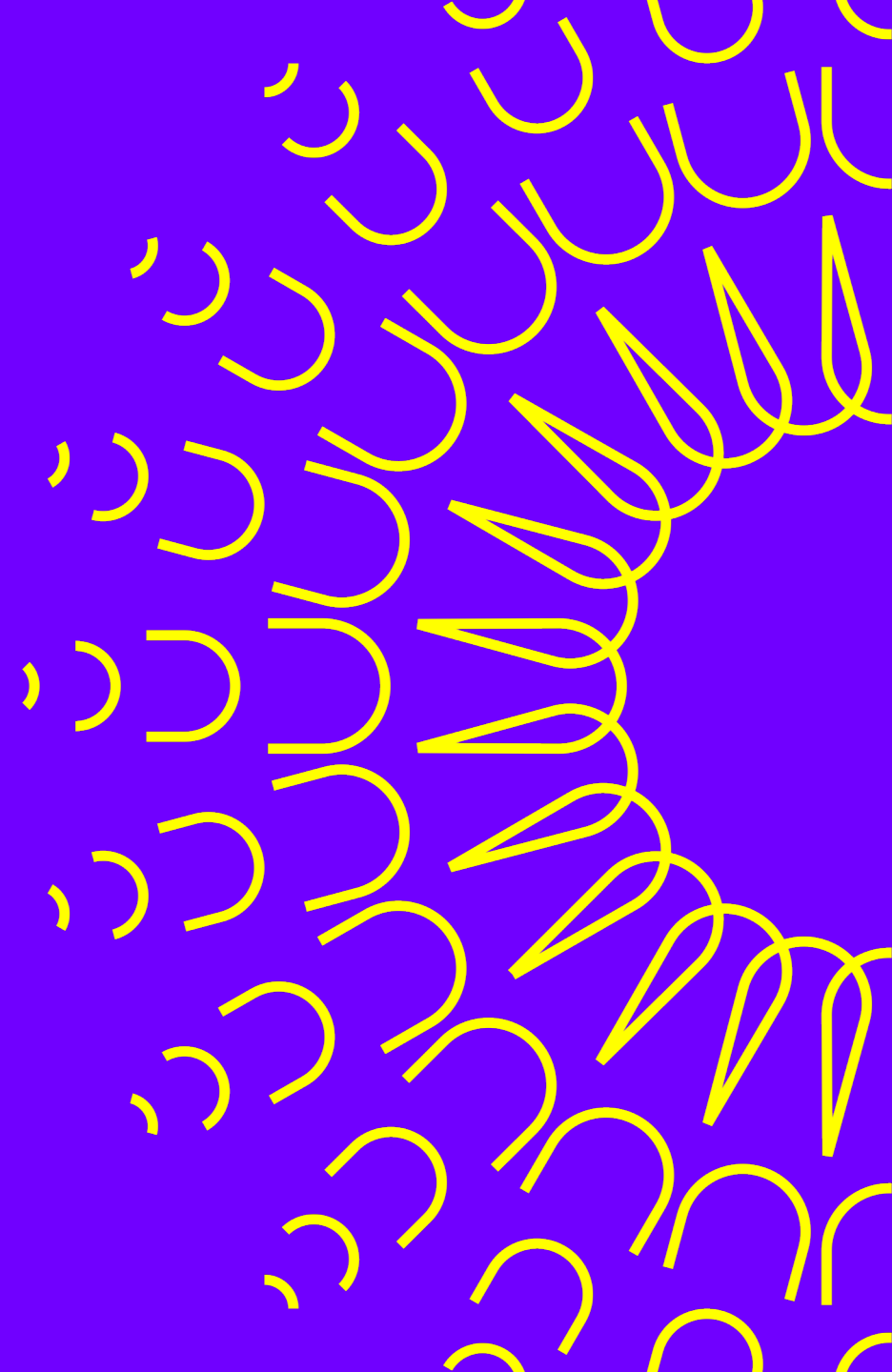


# Break, Learn, Refine – The Art of **Hypothesis- Driven Development** of ML-Powered Search

Andrey Kulagin

Head of Machine Learning @ Uzum Market

Haystack EU 2023



# Agenda

- Chapter 1: Uzum Market and its Search
- Chapter 2: Search complexity and how to handle it
- Chapter 3: Hypothesis driven development of ML-powered Search
  1. Right direction
  2. Fast iterations
  3. High chances of success
  4. Stable results

Chapter 1

# **UZUM MARKET AND ITS SEARCH**

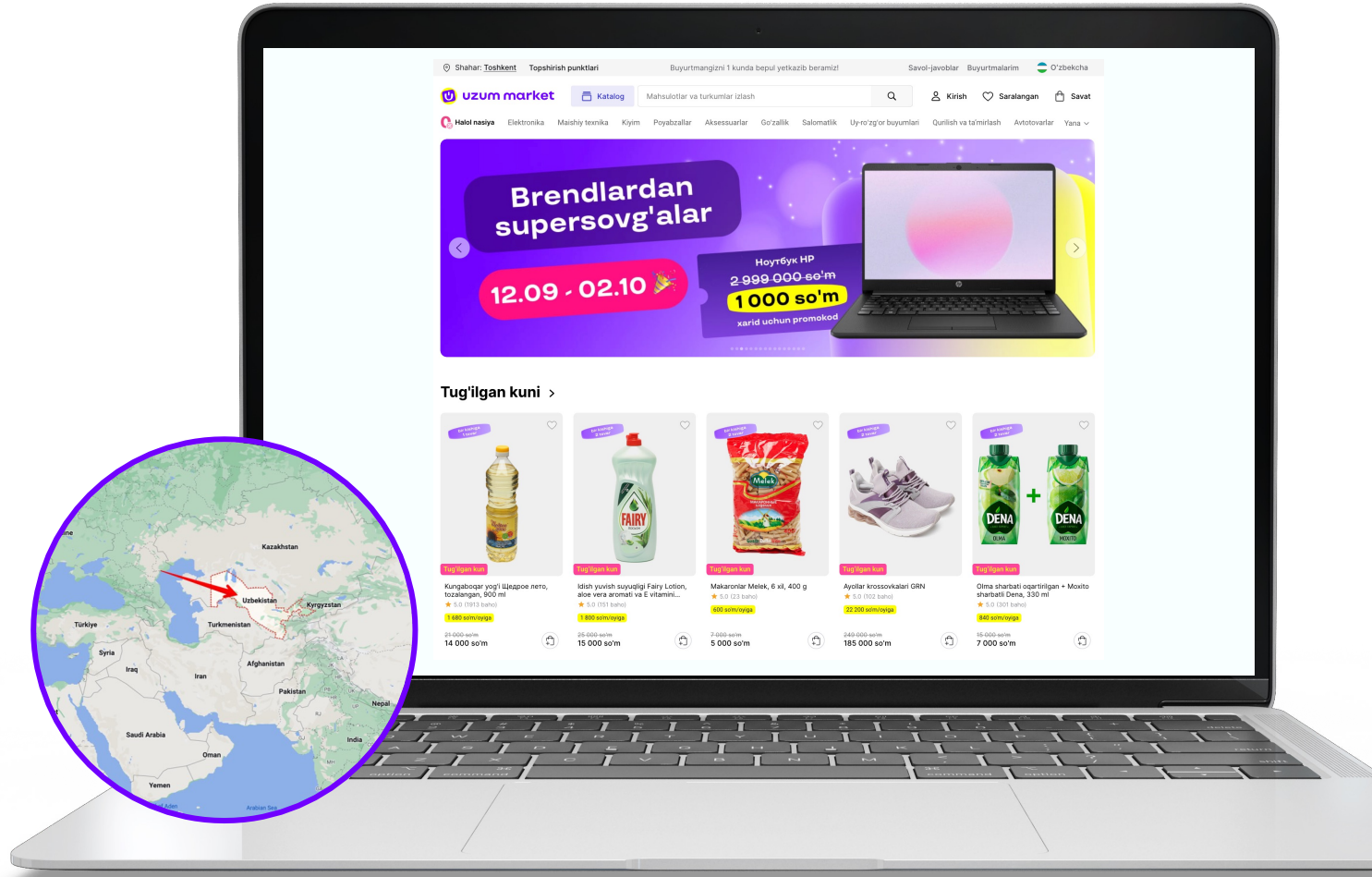
# UZUM MARKET

**> 600 000** items in the product catalog

**> 6000** sellers

**2 languages supported** Uzbek & Russian

**Sept 2022** 1<sup>st</sup> release



## So'rov bo'yicha qidiruv natijalari "iphone case"

Saralash

Ommabop ▾

## Turkumlar

Aksessuarlar

Elektronika

## Narx, baho

dan 7000

oldin 668000

## Rang

🟤 Qaymoqrang

🟡 Oq

🟠 Moviy

🟢 Sariq

🟣 Yashil

🟡 Tillarang

Yana 12

## Brend

 7saber ABC Apple Baseus Belkin Borofone

Yana 28



Texnika yarmarkasi

G'ilof Silicone Case iPhone 11 kamera himoyasi bilan (1 baho)

★ 4.8 (4 baho)

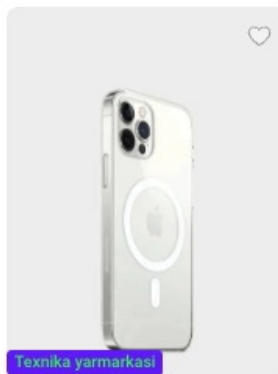
6 720 so'm/oyiga

70 000 so'm  
56 000 so'm

iPhone XR, 11, 12, 13, Pro, Pro Max, PLUS shaffof silikon uchun...

★ 4.8 (4 baho)

1 200 so'm/oyiga

65 000 so'm  
10 000 so'm

Texnika yarmarkasi

G'ilof Magsafe iPhone X, 11 Pro, 12 Pro, 12 Pro Max, 13 Pro, 13 Pro...

★ 4.8 (17 baho)

4 440 so'm/oyiga

66 000 so'm  
37 000 so'm

G'ilof iPhone 7,8,SE,XR,XS,11,12,13,14,Pro,Max,...

★ 5.0 (20 baho)

1 080 so'm/oyiga

49 000 so'm  
9 000 so'm

G'ilof shaffof iPhone 6s, 7, 8, SE, XR, XS, 11, 12, 13, 14, Pro, Max,...

★ 5.0 (115 baho)

1 080 so'm/oyiga

49 000 so'm  
9 000 so'm

Tug'ilgan kun Texnika yarmarkasi

G'ilof iPhone 6s, 7, 8, SE, XR, XS, 11, 12, 13, 14, Pro, Max, Plus, Min...

★ 4.9 (105 baho)

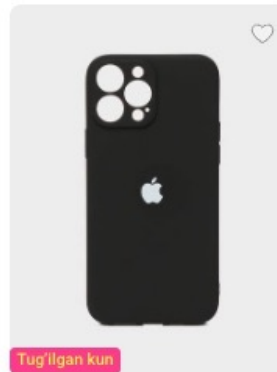
840 so'm/oyiga

45 000 so'm  
7 000 so'm

Case bilan karta cho'ntagi bilan iPhone XR, 11, 12, 13, 14, SE, Pro,...

★ 5.0 (23 baho)

1 080 so'm/oyiga

49 000 so'm  
9 000 so'm

Tug'ilgan kun

G'iloflar iPhone uchun, silikon

★ 5.0 (1 baho)

1 680 so'm/oyiga

35 000 so'm  
14 000 so'm

# WHY SEARCH IS IMPORTANT FOR THE MARKETPLACE?

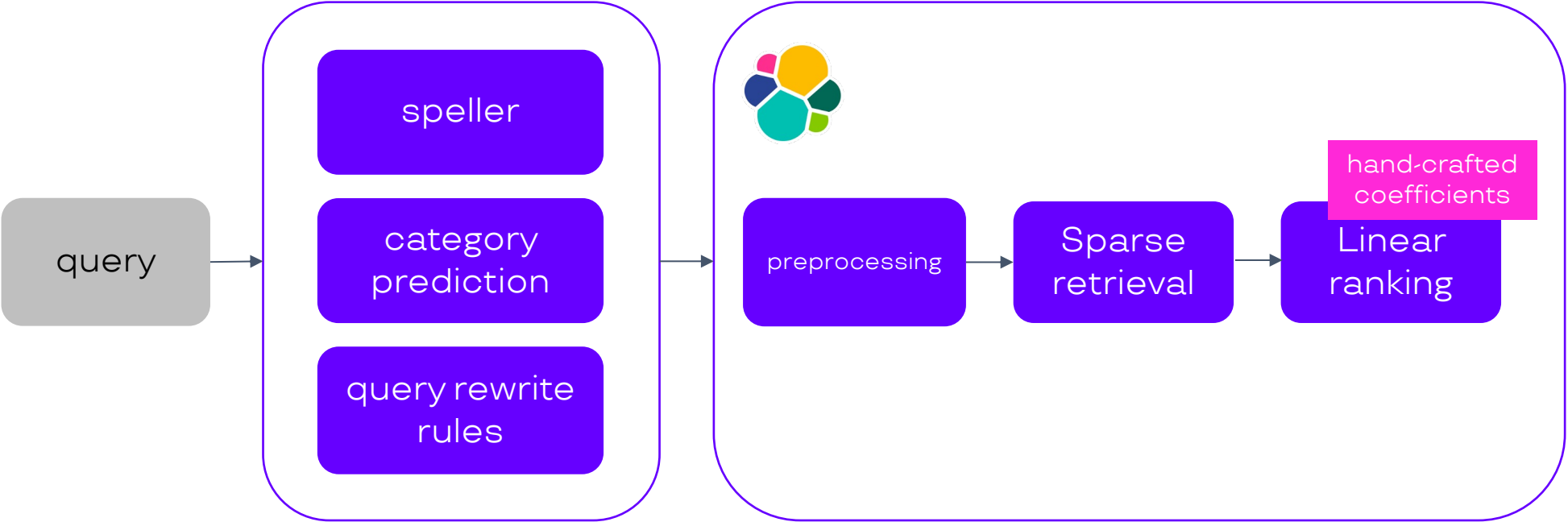
Искать товары и категории 

- The most common marketplace use-case is to purchase a particular item
- A good search system becomes essential for navigation when you have 100k+ of items (1000k+ in the future).

# SEARCH PIPELINE 1 YEAR AGO



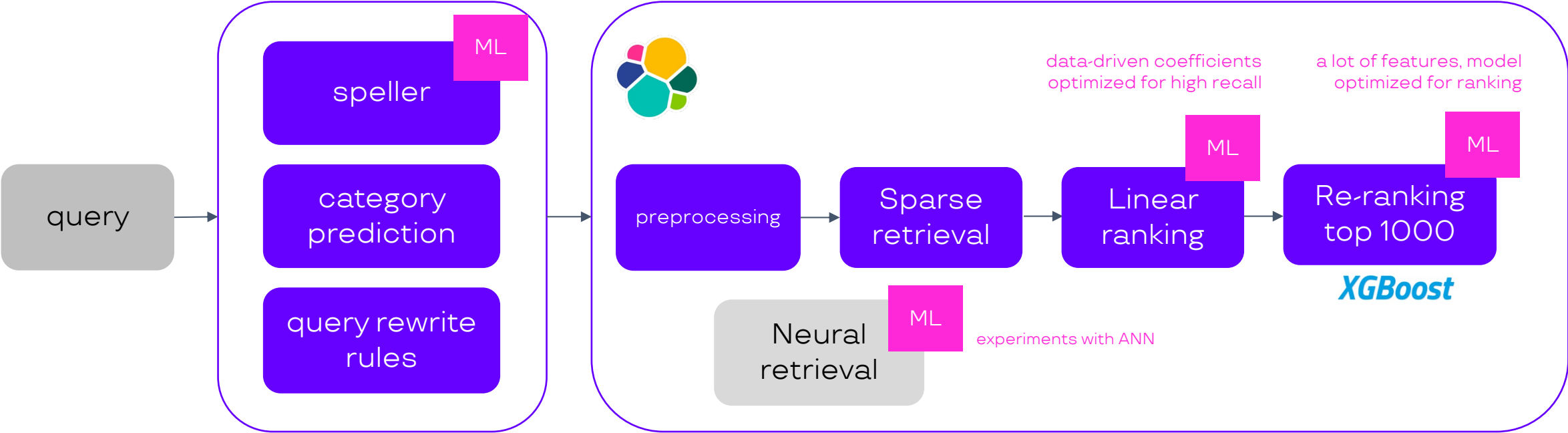
$$\text{bm25\_score} * A + \text{num\_orders} * B + \text{rating} * C$$



# SEARCH PIPELINE NOW



trained on our data





Chapter 2

**SEARCH COMPLEXITY  
AND HOW TO HANDLE IT**

## THE LONG WAY AHEAD

The search is broken!  
When will you fix it?

Just build the search like  
Google

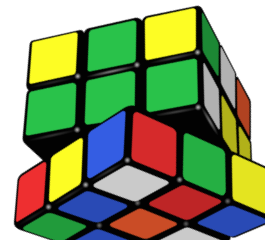
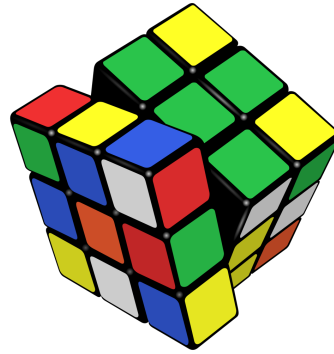
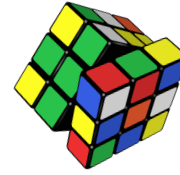
Why product X isn't  
shown for query Y?

I don't like the  
results!

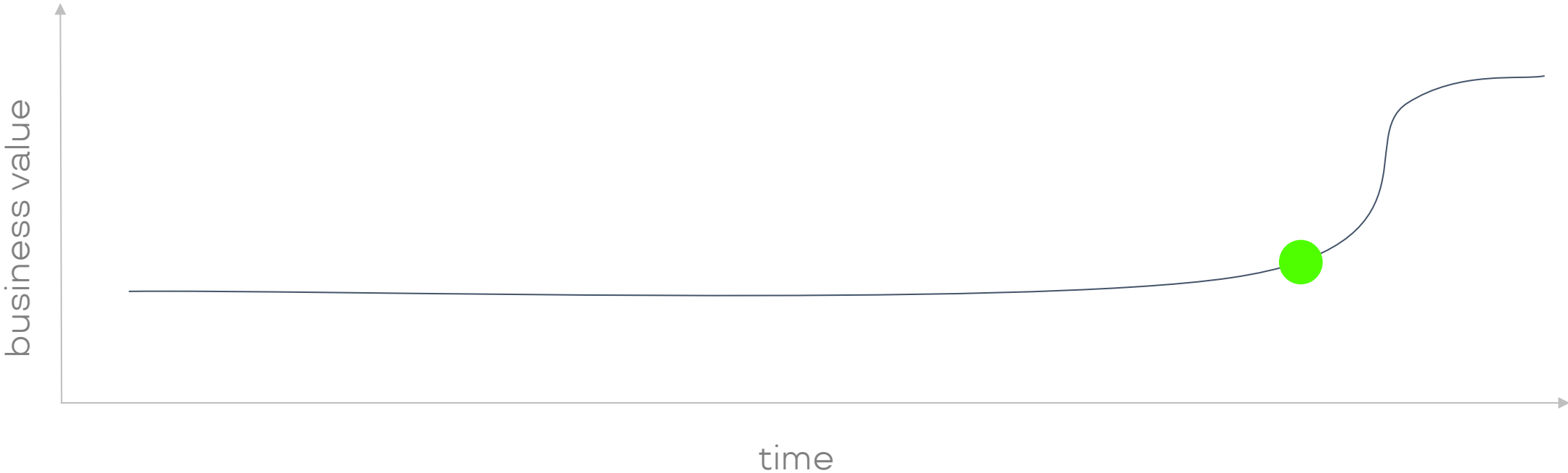
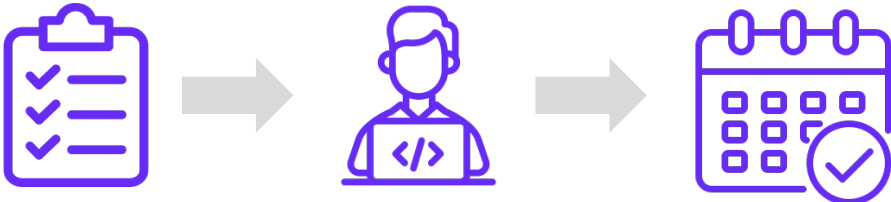


# HARD REALITY

- Extreme complexity and high uncertainty 🤯
- Hundreds of new sellers and thousands of new products every day 🆕
- Seasonality and new emerging trends 📈📉
- Balance between buyers and sellers interests ⚖️
- A lot of RnD 🔬

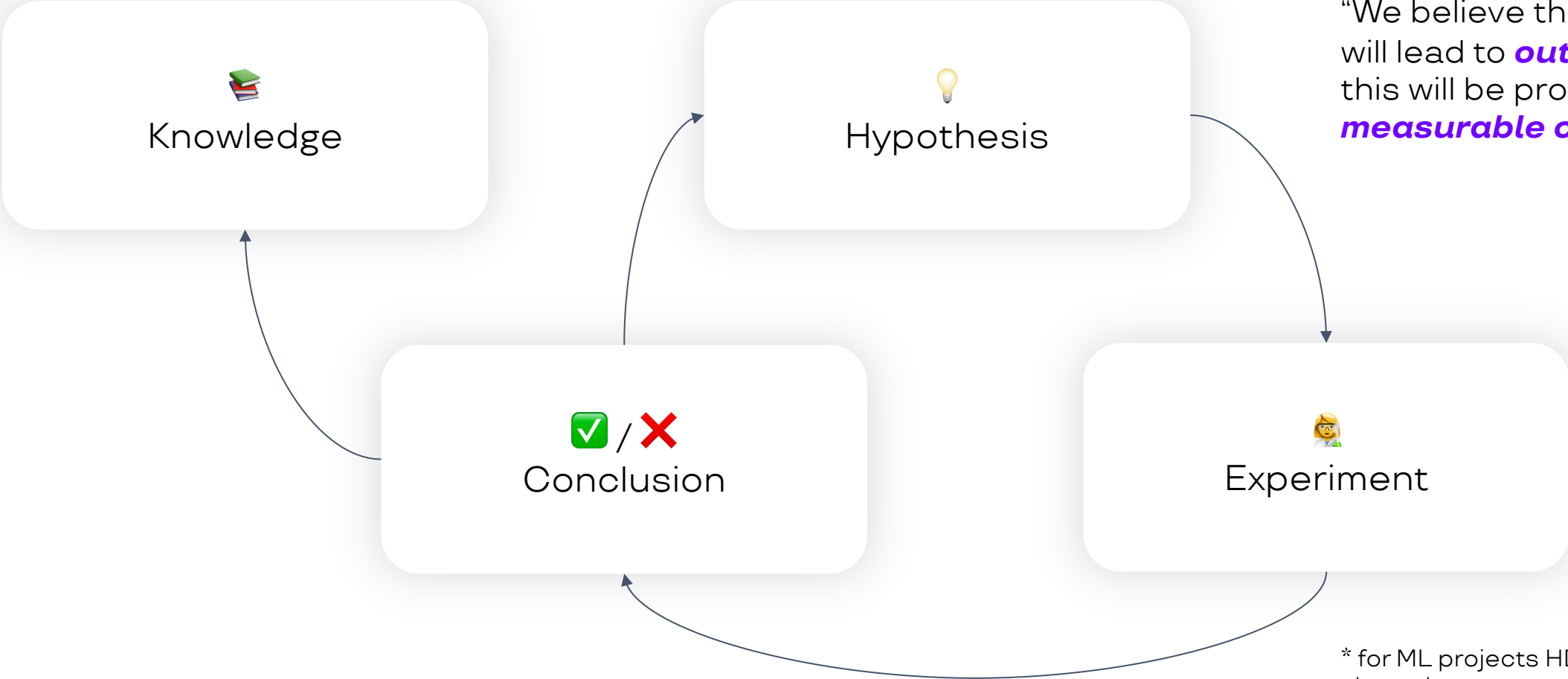


# REQUIREMENTS-DRIVEN DEVELOPMENT... ..is not going to work here



what if ●?

# HYPOTHESIS-DRIVEN DEVELOPMENT



“We believe that **change** will lead to **outcome** and this will be proven when **measurable condition**”.

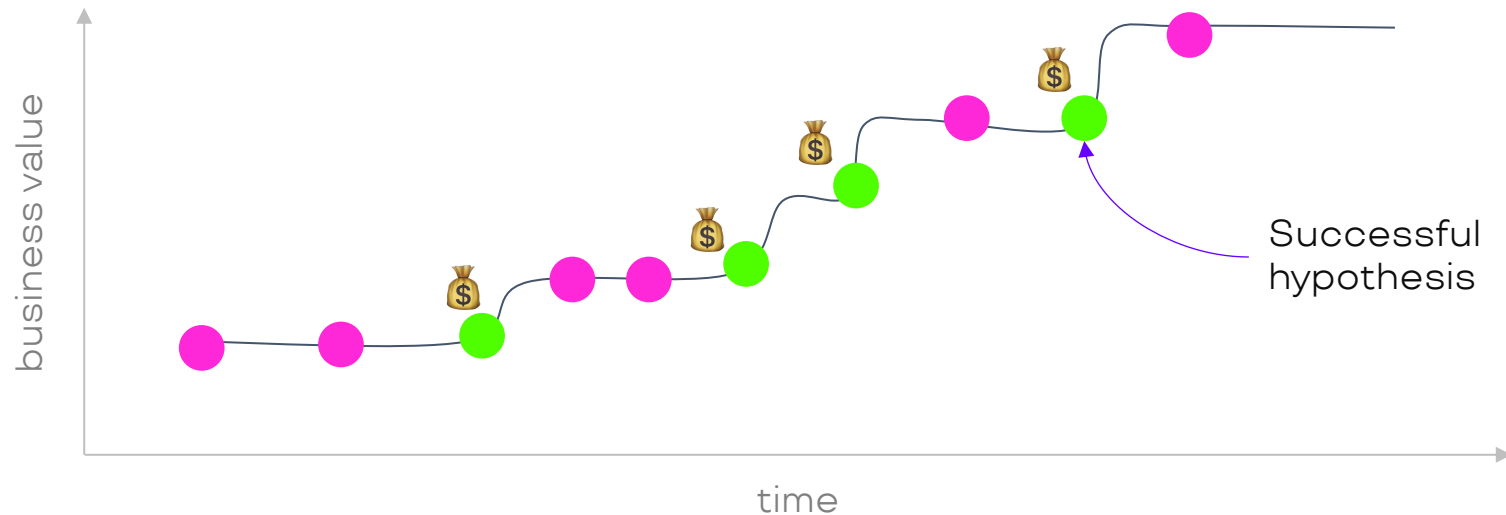
continuous, iterative process

\* for ML projects HDD flow will be shown later

# ADVANTAGES OF HDD

- Regular delivery of additional business value 💰
- Reduced time to market ⌚
- Decreased delivery risks
- Team management (dopamine) 😊

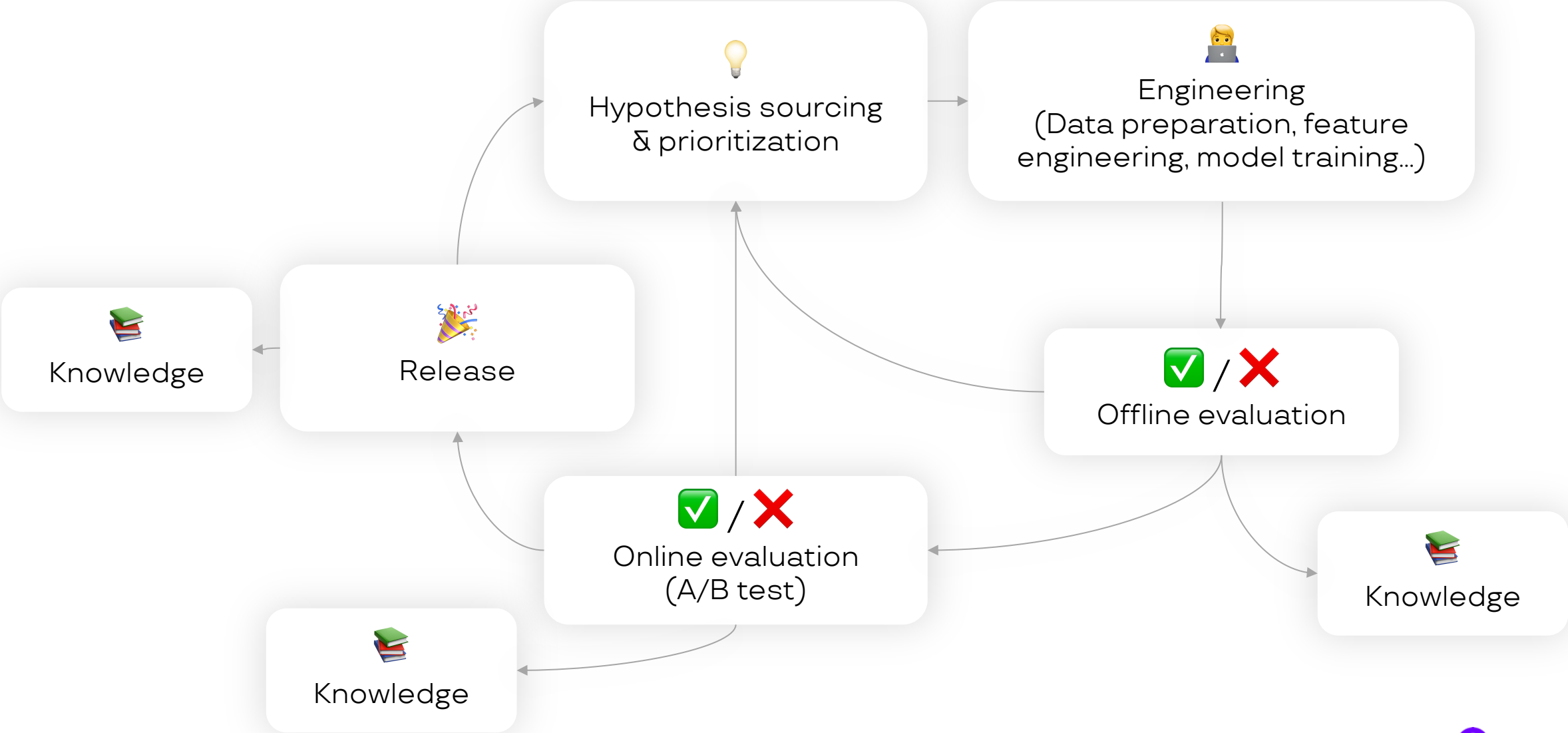
HOW TO  
MEASURE?



Chapter 3

**HYPOTHESIS DRIVEN  
DEVELOPMENT OF ML-POWERED  
SEARCH**

# HDD FOR ML PROJECTS







# 1. Right direction:

It's all about the correct metrics & evaluation procedures



Online evaluation  
(A/B test)



Offline evaluation

*business goals*

*product OKRs*

**WHAT IS THE BEST METRIC FOR  
SEARCH QUALITY?**

## THE BEST SEARCH QUALITY ESTIMATE 😊

**CTO@k** - how many “bad” results your CTO has found among their  $k$  random searches

\* where  $k$  directly correlates with their free time

# BUSINESS METRICS

- Global
  - ARPU, ARPPU, conversion to purchase, AOV, retention, LTV ...
- Search related
  - conversions from search to click, to add-to-cart, to order, to purchase
  - search ARPU / ARPPU
  - # empty queries
  - search abandonment rate
  - ...

can be used only as **online metrics**

we are especially interested in metrics which:

- are connected to the current business strategy
- can be tested during A/B (with adequate MDE / time)

# ONLINE METRICS: SEARCH-SESSION-WISE CONVERSIONS

1. **CR search2click (= 4/5)**
2. **CR search2atc (= 2/5)** - fraction of searches which resulted in at least one product from SERP added to cart
3. **CR search2purchase (= 1/5)** - fraction of searches, which resulted in at least one product from SERP purchased <<< **requires attribution modelling**

all of these are ratio-metrics

date, user, session_id	query	click	ATC	purchase
2022-01-01, Jane, 343g9n	"socks"	✓		
2022-01-01, Jane, 343g9n	"iphone"	✓	✓	✓
2022-01-01, Mark, s9g55n	"socks"			
2022-01-01, Mark, s9g55n	"sunflower oil"	✓	✓	
2022-01-01, Mark, s9g55n	"t shirt"	✓		

# ONLINE METRICS: GLOBAL DAILY METRICS, RELATED TO 💰

1. **ARPU\_daily (ARPPDAU)** (=  $\$30 / 5 = \$6$ ) Average Revenue Per Daily Active User
2. **cr2purchase\_daily** (=  $3/5$ ) - fraction of daily active users who made a purchase
3. **ARPPU\_daily (ARPPDAU)** (=  $\$30 / 3 = \$10$ ) - Average Revenue Per Paying Daily Active User (usually equal to AOV except cases when several orders are made on the same day)

$$\text{ARPU\_daily} = \text{cr2purchase\_daily} * \text{ARPPU\_daily}$$

↑  
responsible  
for total  
GMV

↑  
frequency

↑  
amount

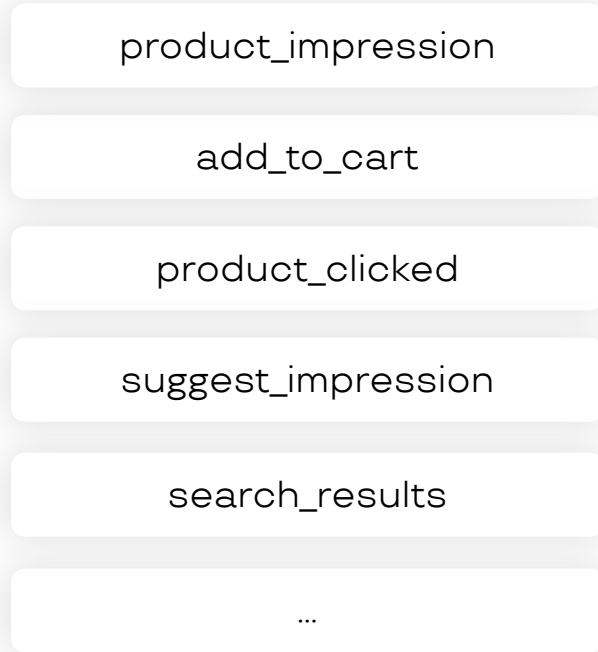
date, user	made a purchase	spend money
2022-01-01, Jane	✓	10 \$
2022-01-01, Mark		
2022-01-02, Jane	✓	15 \$
2022-01-02, Bob	✓	5 \$
2022-01-03, Alex		

## ONLINE METRICS: SEARCH DAILY METRICS, RELATED TO 💰

1.  $\text{ARPU\_daily\_search} = \text{attributed\_to\_search\_revenue} / \text{n\_search\_visitors}$
2.  $\text{cr2purchase\_daily\_search} = \text{n\_search\_buyers} / \text{n\_search\_visitors}$
3.  $\text{ARPPU\_daily\_search} = \text{attributed\_to\_search\_revenue} / \text{n\_search\_buyers}$



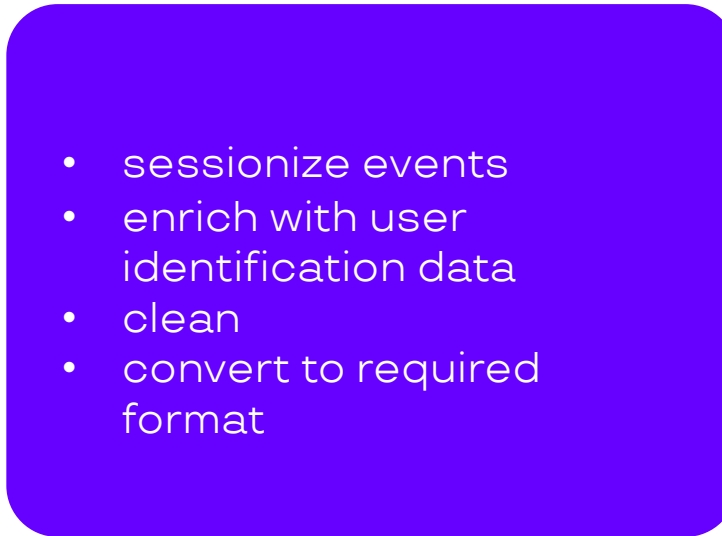
# CLICKSTREAM



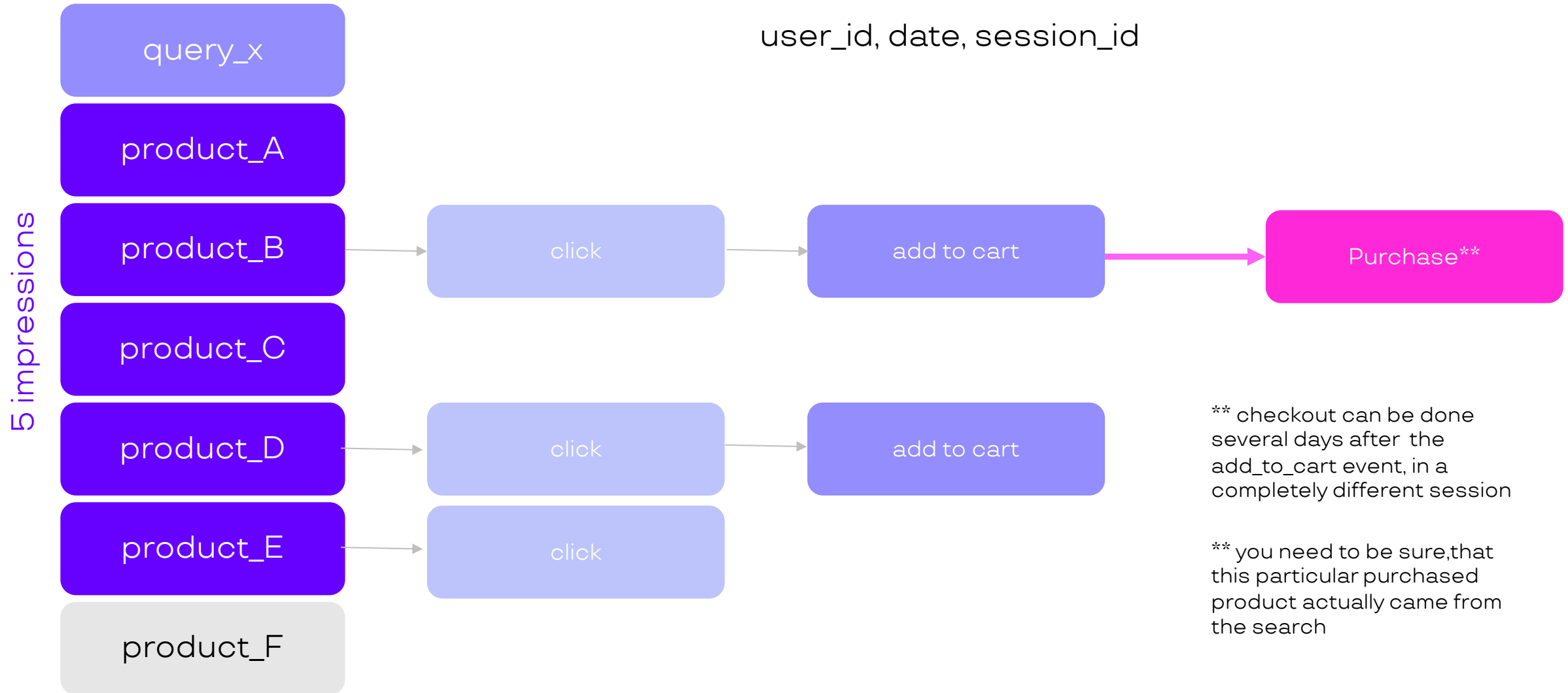
clients generate events



## Clickstream service



# ESSENTIAL DATA



# Search results, query="Sunflower oil"


Shahar: Toshkent Topshirish punktlari Buyurtmangizni 1 kunda bepul yetkazib beramiz! Savol-javoblar Bu

 **uzum market**

 Katalog

Kungaboqar yog'i



 Kirish



**Halol nasiya**

Elektronika

Maishiy texnika

Kiyim

Poyabzallar

Aksessuarlar

Go'zallik

Salomatlik

Uy-ro'zg'or buyumlari

Qurilish va ta'

Bosh sahifa / Barcha toifalar

## So'rov bo'yicha qidiruv natijalari "Kungaboqar yog'i"

Saralash

### Turkumlar

Oziq-ovqat mahsulotlari




### Narx, baho

dan 8000

oldin 200000

### Brend

- Milter
- Щедрое Лето
- Dardanel
- Groceries
- Iberica

 <p>bir kishiga 2 tovar</p>	 <p>bir kishiga 2 tovar</p>	 <p>bir kishiga 2 tovar</p>
<b>Hafta aksiyasi</b>	<b>Hafta aksiyasi</b>	<b>Hafta aksiyasi</b>
Kungaboqar yog'i Sofia, tozalangan va...	Kungaboqar yog'i Щедрое лето, tozalangan, 900 ml	Kungaboqar yog'i Oleyuna, 1 litr
★ 4.9 (1141 baho)	★ 5.0 (1757 baho)	★ 5.0 (467 baho)
<b>1 680 so'm/oyiga</b>	<b>1 680 so'm/oyiga</b>	<b>2 160 so'm/oyiga</b>
19 000 so'm <b>14 000 so'm</b>	21 000 so'm <b>14 000 so'm</b>	29 000 so'm <b>18 000 so'm</b>

# Main page, "Sale" collection

Shahar: **Toshkent** Topshirish punktlari Buyurtmangizni 1 kunda bepul yetkazib beramiz! Savol-javoblar Buyurtmalarim O'zbekcha

**uzum market** Katalog Mahsulotlar va turkumlar izlash Kirish Saralangan Savat

Halol nasiya Elektronika Maishiy texnika Kiyim Poyabzallar Aksessuarlar Go'zallik Salomatlik Uy-ro'zg'or buyumlari Qurilish va ta'mirlash Avtotovarlar Yana






## Birinchi navli Melek bug'doy uni, 1 kg

10 000 so'm **-40%**  
**6 000 so'm**


faqat 11-sentabrgacha

количество ограничено

### Arzon narxlar >

 <p>bir kishiga 2 tovar</p> <p><b>Hafta aksiyasi</b></p> <p>Kungaboqar yog'i Щедрое лето, tozalangan, 900 ml</p> <p>★ 5.0 (1757 baho)</p> <p><b>1 680 so'm/oyiga</b></p> <p>21 000 so'm 14 000 so'm</p>	 <p>bir kishiga 2 tovar</p> <p><b>Hafta aksiyasi</b></p> <p>Idishlarni yuvish uchun suyuqlik Fairy, limon, 450 ml</p> <p>★ 5.0 (636 baho)</p> <p><b>1 080 so'm/oyiga</b></p> <p>16 000 so'm 9 000 so'm</p>	 <p>bir kishiga 3 tovar</p> <p><b>Hafta aksiyasi</b></p> <p>Bolalar tagliklari Sleepy Natural Maxi 4, 7-14 kg, 30 dona</p> <p>★ 4.2 (10 baho)</p> <p><b>9 120 so'm/oyiga</b></p> <p>100 000 so'm 76 000 so'm</p>	 <p>bir kishiga 4 tovar</p> <p><b>Eksklyuziv Hafta aksiyasi</b></p> <p>Tozalash kukuni Oila tanlovi, limon xushbo'y hidi bilan, 400 g</p> <p>★ 4.7 (151 baho)</p> <p><b>600 so'm/oyiga</b></p> <p>9 000 so'm 5 000 so'm</p>	 <p>bir kishiga 2 tovar</p> <p><b>Hafta aksiyasi</b></p> <p>Kostriyulka Мечта Star</p> <p>★ 5.0 (96 baho)</p> <p><b>25 560 so'm/oyiga</b></p> <p>213 000 so'm</p>
--	--	--	---	---

## Other product's description page (PDP), "Similar products" collection



★ 5.0 (467 baho) 16000 ta buyurtma Istaklarga

### Kungaboqar yog'i Oleyna, 1 litr

Sotuvchi: [GROCERY](#)

Yetkazib berish: 1 kun, bepul

Miqdor:  Bor-yo'gi 4 dona qoldi

Narx: **18 000 so'm** ~~29 000 so'm~~ Hafta aksiyasi

Oyiga 2 160 so'mdan muddatli to'lov >


Savatga qo'shish Tugmani 1 bosishda xarid qilish

👤 Bu haftada 197 kishi sotib oldi

[Mahsulot tavsifi](#) [Sharhlar \(467\)](#)

Qattiq sifat nazorati va issiq suv, tabiiy loy, uzoq muddatli sovutish, filtrlash va yuqori haroratlar yordamida tabiiy tozalashning ko'p bosqichli tizimidan o'tgan tanlangan kungaboqar urug'laridan tayyorlangan.


### O'xshash mahsulotlar



Hafta aksiyasi

Kungaboqar yog'i Sofia, tozalangan va deodorizatsiyalangan, 1 litr


★ 4.9 (1141 baho)



Hafta aksiyasi

Kungaboqar yog'i Шедрое лето, tozalangan, 900 ml


★ 5.0 (1757 baho)



Hafta aksiyasi

Kungaboqar yog'i Oleyna, 2 litr

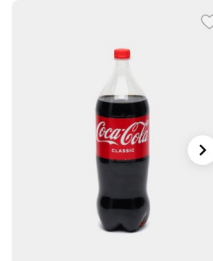
★ 5.0 (52 baho)



Hafta aksiyasi

Qahva Jacobs Monarch, 95 g

★ 5.0 (120 baho)

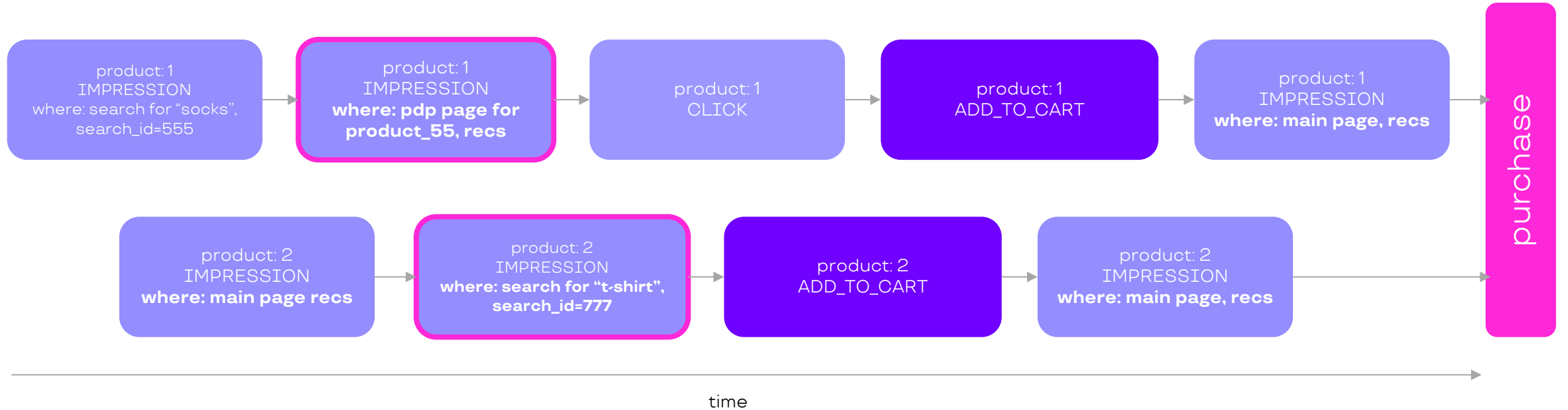


Hafta aksiyasi

Gazlangan ichimlik Coca-Cola Classic, 1.5 litr

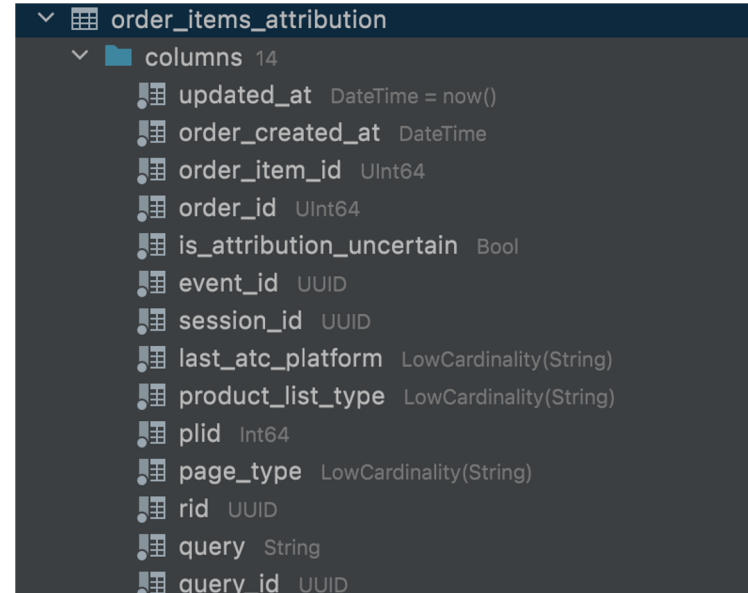
★ 5.0 (84 baho)

# LAST IMPRESSION BEFORE LAST ADD-TO-CARD



order_item_id	product	where found	additional_info	session_id
34963	1	recs on pdp	product_id=55	q5g67
34964	2	search	q="t-shirt", search_id=777	f9486

# ATTRIBUTION MODELING



A screenshot of a database schema for the 'order\_items\_attribution' table. The table has 14 columns. The columns are listed as follows:

Column Name	Data Type
updated_at	DateTime = now()
order_created_at	DateTime
order_item_id	UInt64
order_id	UInt64
is_attribution_uncertain	Bool
event_id	UUID
session_id	UUID
last_atc_platform	LowCardinality(String)
product_list_type	LowCardinality(String)
plid	Int64
page_type	LowCardinality(String)
rid	UUID
query	String
query_id	UUID

this table is a  and actively used not only in search team

# ATTRIBUTION MODELING

Congrats!

Now we can easily measure metrics like **cr\_search2purchase** and **ARPU\_daily\_search, ARPPU\_daily\_search**

In A/B tests too 🎉



# MY PERSONAL TOP OF MISTAKES RELATED TO A/B TESTS

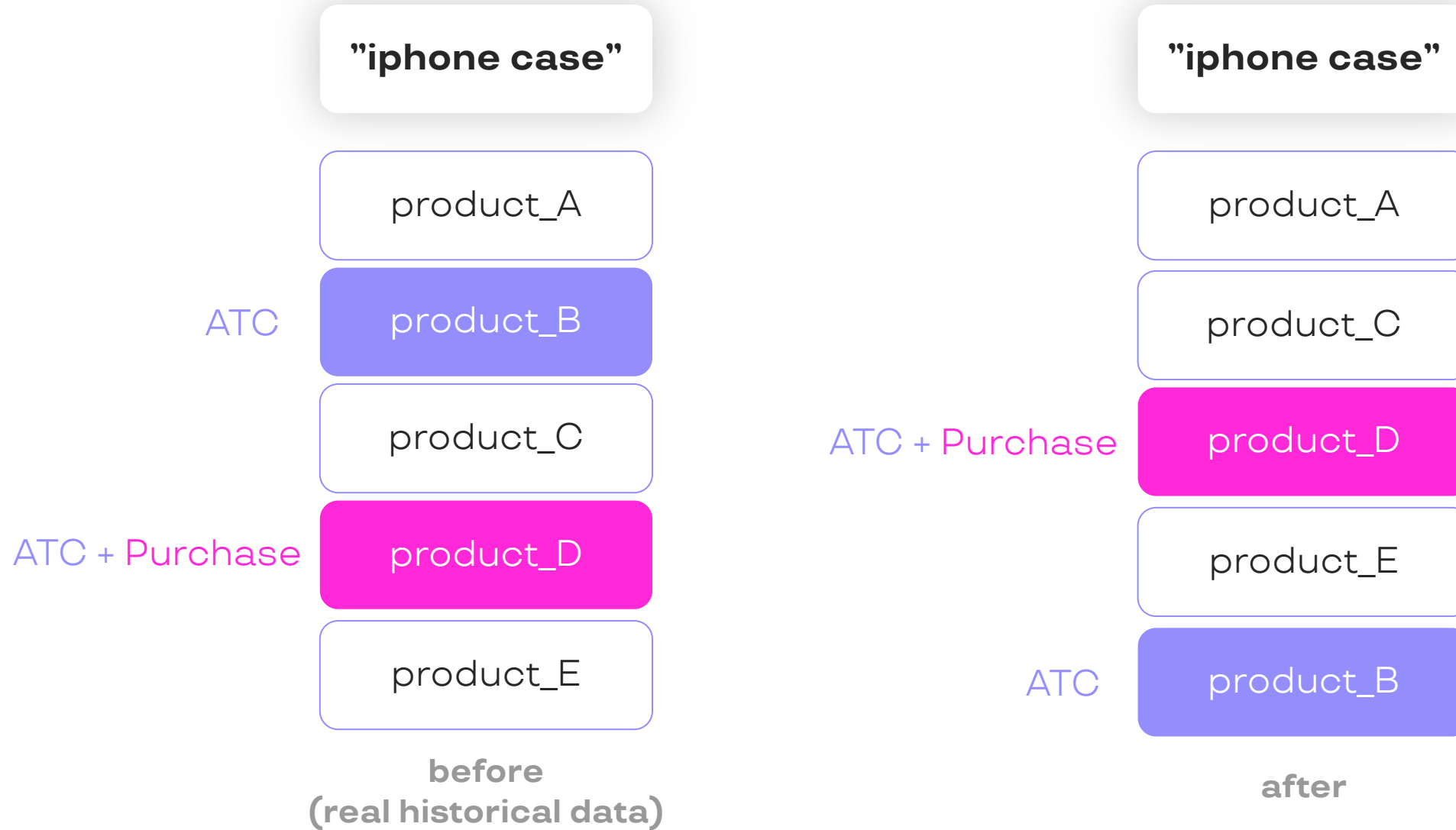
1. Apply T-test for ratio-metrics
  - Just use Delta-method: [Applying the Delta Method in Metric Analytics: A Practical Guide with Novel Ideas](#)
  - Or Linearization: [Approximations for Mean and Variance of a Ratio](#)
2. Run A/B without proper design
  - Calculate MDE & sample-size BEFORE test
  - Don't forget to handle multiple comparisons problem
  - Run A/A-test (simulation) for every new metric
3. Bugs in metric calculation
4. Run an experiment without events logged
5. Mess up group labels ("B" and "A") 😊
6. Forget to accurately document A/B results

# A/B TEST IS AN EXPENSIVE PROCEDURE

- Time for a/b itself
- Effort to prepare, conduct and analyse (engineers, analytics)
- User base is limited
- There is always a risk, that group B is worse

A lot of hypothesis can be checked preliminarily on historical data

# WE NEED PROXY METRICS FOR OFFLINE EVALUATION

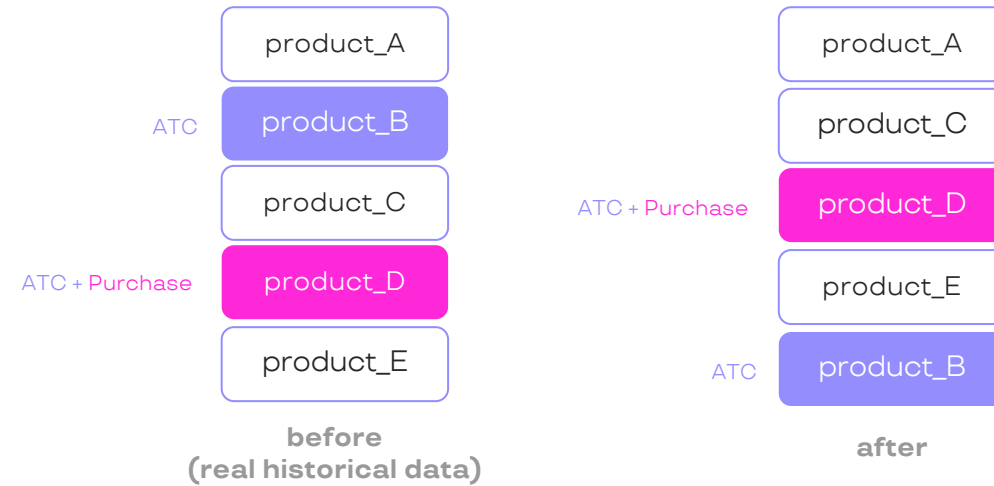


# RANKING METRICS

- NDCG@k - Normalized Discounted Cumulative Gain
  - 0 - impression, 1 - click, 2 - atc, 3 - order ?
  - 0 - impression, 1 - order?
- MAP@k - Mean Average Precision
- MRR - Mean Reciprocal Rank
- ERR - Expected Reciprocal Rank

Which is the best? How to choose “relevant” signal? How to determine k?

# RANKING METRICS



metric	relevance signal	Before	After	Change in %
NDCG	0 - impression, 1 - atc	0.65	0.54	-17%
NDCG	0 - impression, 1 - purchase	0.43	0.5	+16%
NDCG	0 - impression, 1 - atc, 2 - purchase	0.56	0.52	-6%
MRR	0 - impression, 1 - atc	0.50	0.33	-33%
MRR	0 - impression, 1 - purchase	0.25	0.33	+33%
mean first atc pos	0 - impression, 1 - atc	2	3	+50%
mean first order pos	0 - impression, 1 - purchase	4	3	-25%

**How to determine the best offline metric?**

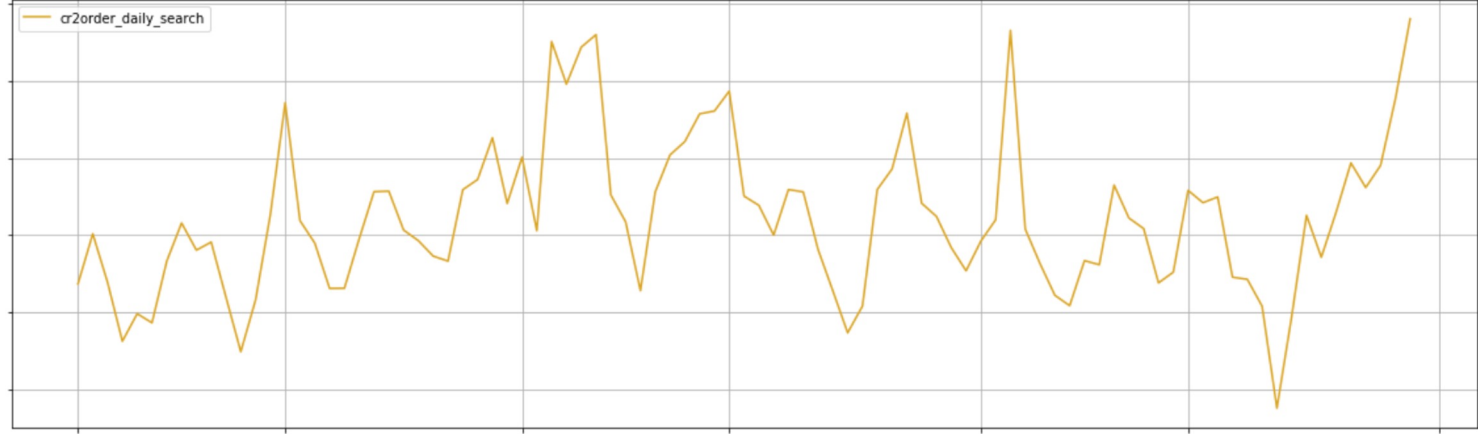
# THE PROPER APPROACH: from experiments history

<b>A/B #</b>	<b>Offline metric 1 uplift</b>	<b>Offline metric 2 uplift</b>	<b>Offline metric 3 uplift</b>	<b>Target online metric uplift</b>
001	+23%	+15%	-10%	-5%
002	-10%	-5%	+8%	+4%
003	+5%	-5%	+4%	+8%
004	+3%	-7%	-4%	-10%
...	...	...	...	...

Train a model that predicts

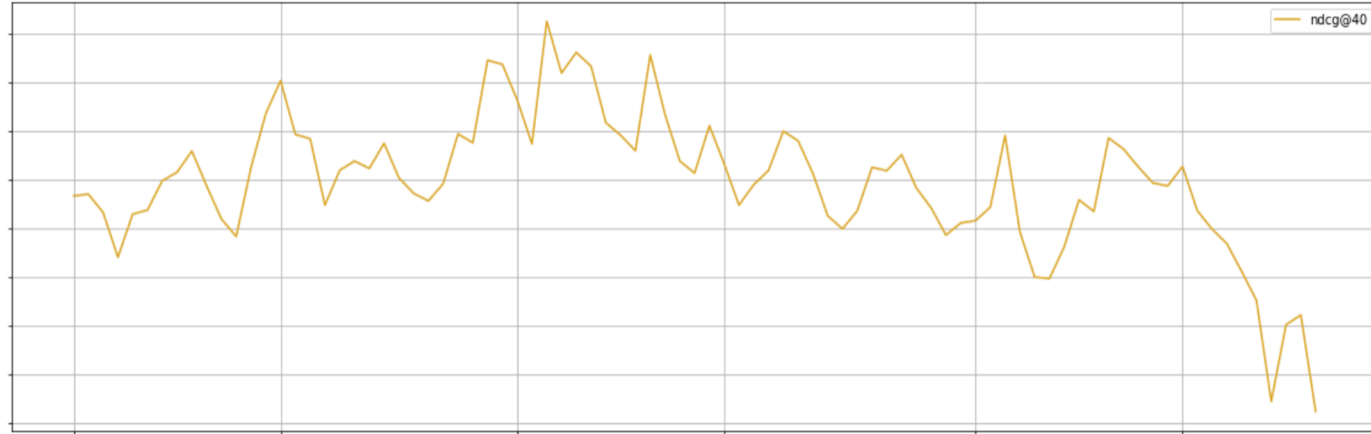
online metric

cr2order\_daily\_search



offline metric

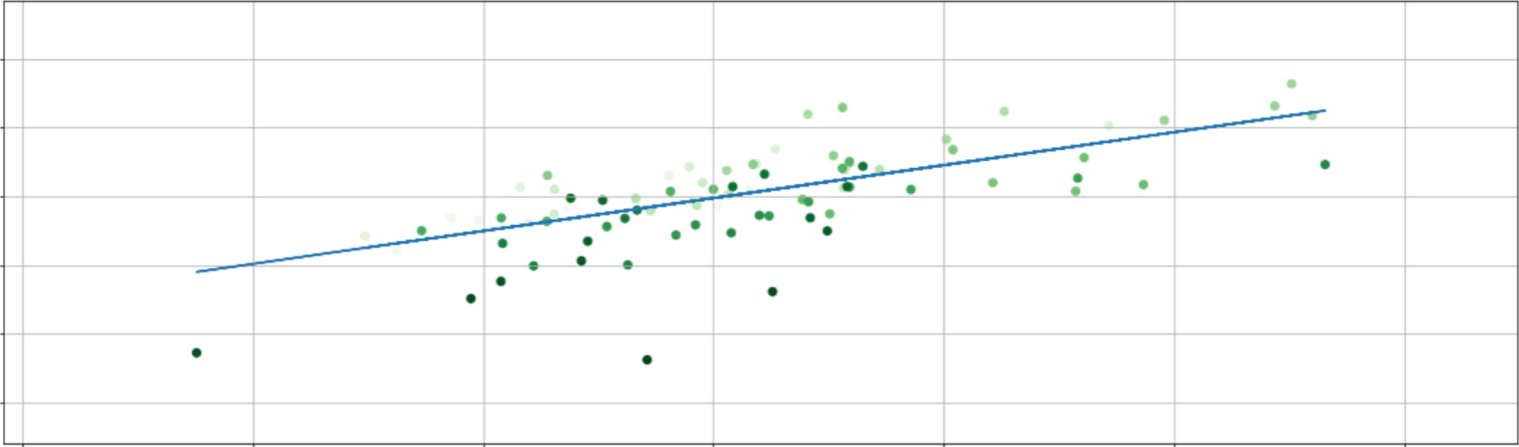
NDCG@40  
{purchase}





# Offline metric & online metric correlation

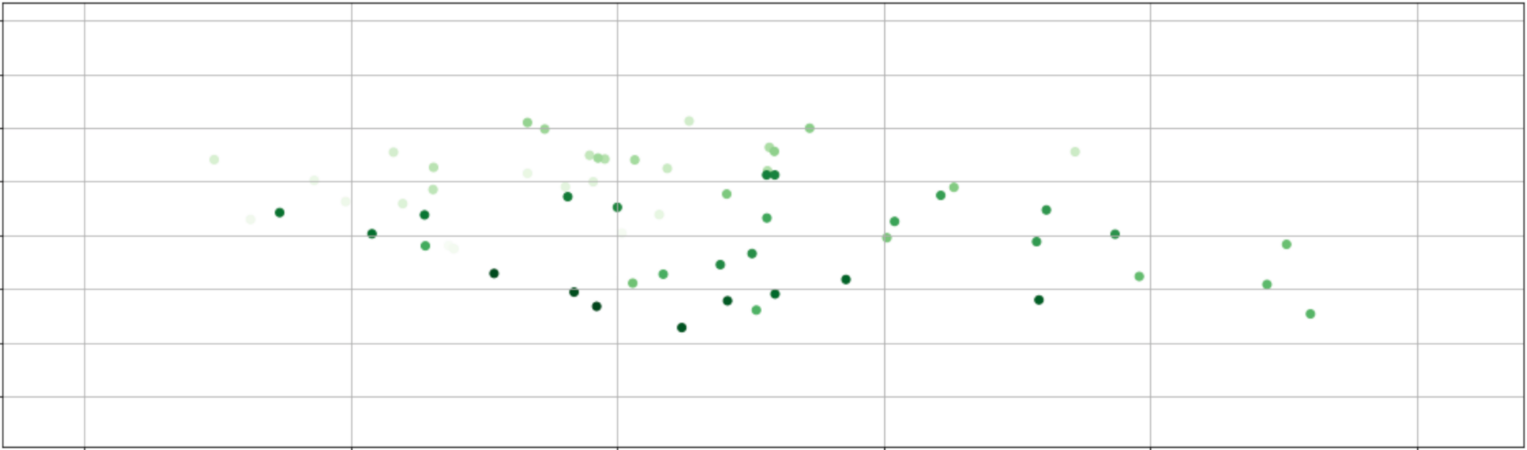
NDCG@40 {purchase}



pearson correlation coef = 0.70

cr2order\_daily\_search

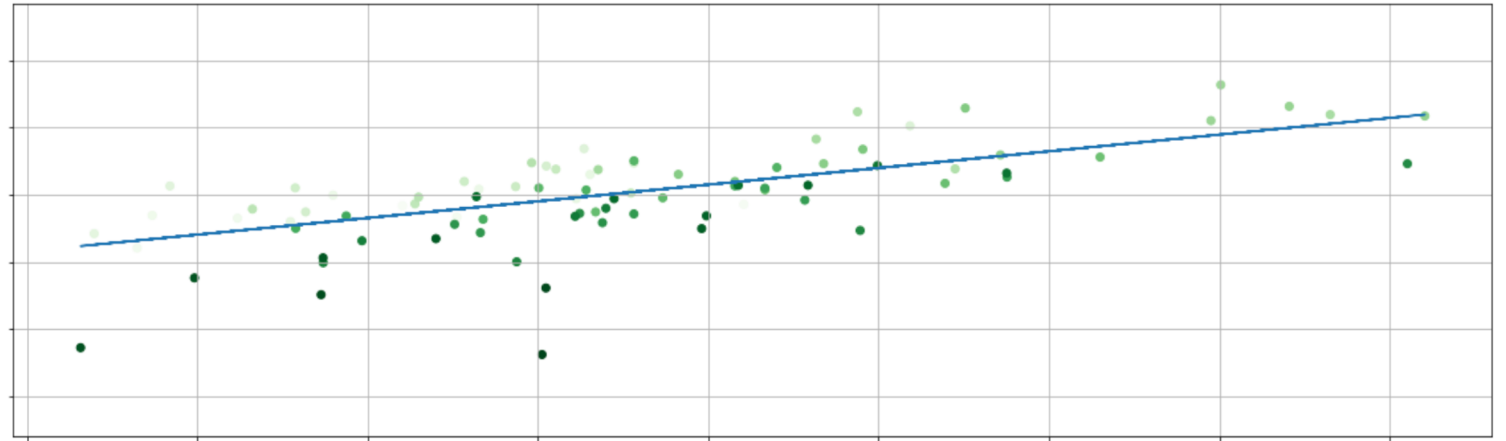
NDCG@40 {ATC}



pearson correlation coef = -0.28

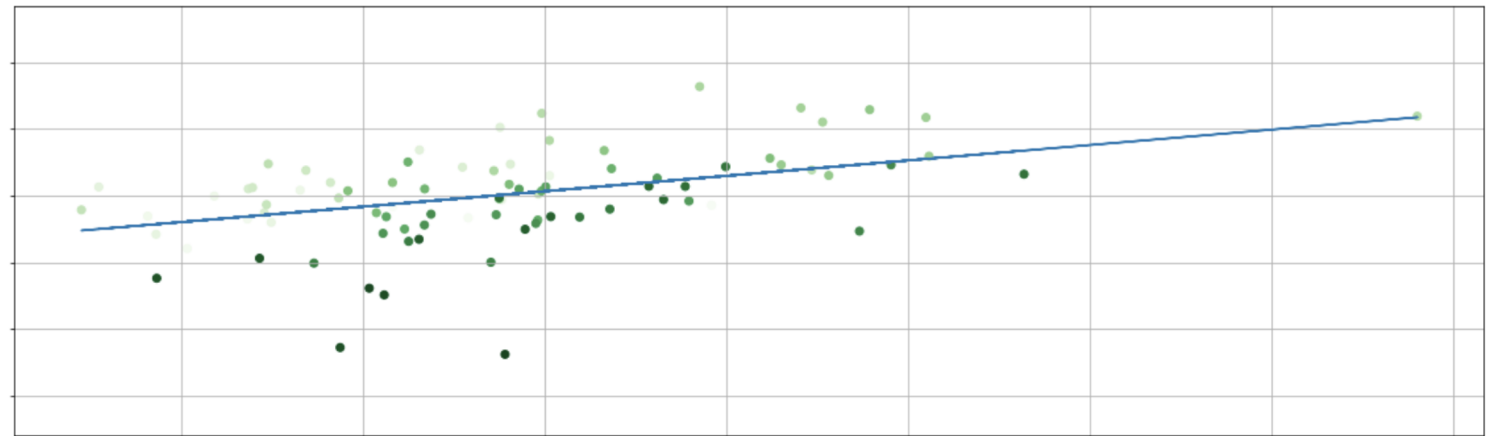
## ARPU daily search & NDCG@40{purchase}

pearson correlation coef = 0.69



## ARPPU daily search & NDCG@40{purchase}

pearson correlation coef = 0.49



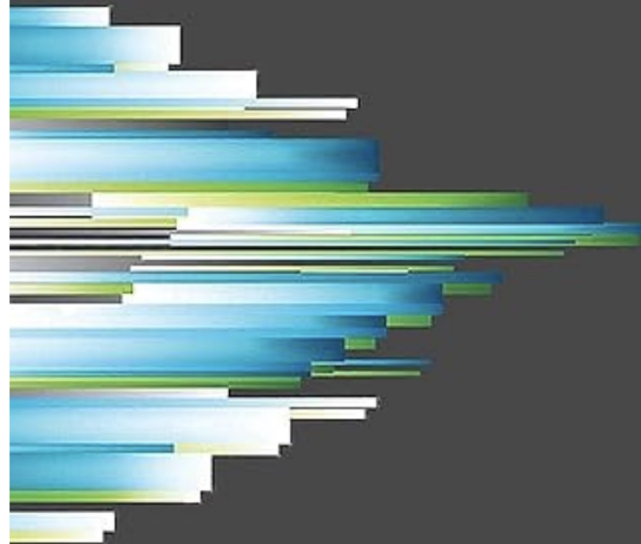
## 2. Fast iterations



Engineering  
(Data preparation, feature  
engineering, model  
training...)

THE SCIENCE OF DEVOPS  
**ACCELERATE**

Building and Scaling High Performing  
Technology Organizations



Nicole Forsgren, PhD  
Jez Humble *and* Gene Kim

# 3. High chances of success



Hypothesis sourcing  
& prioritization

# SOURCES FOR HYPOTHESES

## EXTERNAL

- Talk to other companies
- Participate in meetups / conferences
- Monitor new publications

## INTERNAL

- Analyse frequent problems
- Regularly organize brainstorming sessions with your engineers

# ANALYZE FREQUENT PROBLEMS

Every month we collect queries with the **lowest conversion rates** (with potential problems) **among top-10000 most frequent queries**. Then we ask dedicated assessors to specify what is the problem:

- typo
- incorrect keyboard layout
- transliteration
- incomplete query
- synonym
- the request is too specific
- ranking problem
- assortment problem

# ICE – IMPACT, CONFIDENCE, EASE

Feature category	Ease of data collection	Query-time usage in elasticsearch	Can be used in other projects	Expected impact on metrics	Is used by competitors
Price	2	2	1	2	1
Quality	2	1	1	0	0
Popularity	2	1	1	2	1
Popularity	2	1	1	2	1
Popularity	2	2	1	1	1
Tag	2	1	1	2	1
Price	2	1	1	2	1
Price	2	1	1	2	1
Popularity	2	1	1	1	1
Price	2	1	1	1	1
Rating	2	1	1	1	1
Rating	2	1	1	1	1
Rating	2	1	1	1	1
Popularity	1	1	1	2	1
Popularity	2	1	1	1	0
Popularity	1	1	1	2	0
Popularity	2	1	1	1	0
Quantity	2	1	1	1	0
Quantity	2	1	1	1	0
Price	2	0	0	2	0
Popularity	2	1	1	0	0
Rating	2	1	1	0	0
Rating	2	1	1	0	0
Quality	2	1	1	0	0
Quality	2	1	1	0	0
Tag	1	1	1	1	0
Tag	2	1	1	0	0
Popularity	2	-2	0	2	1



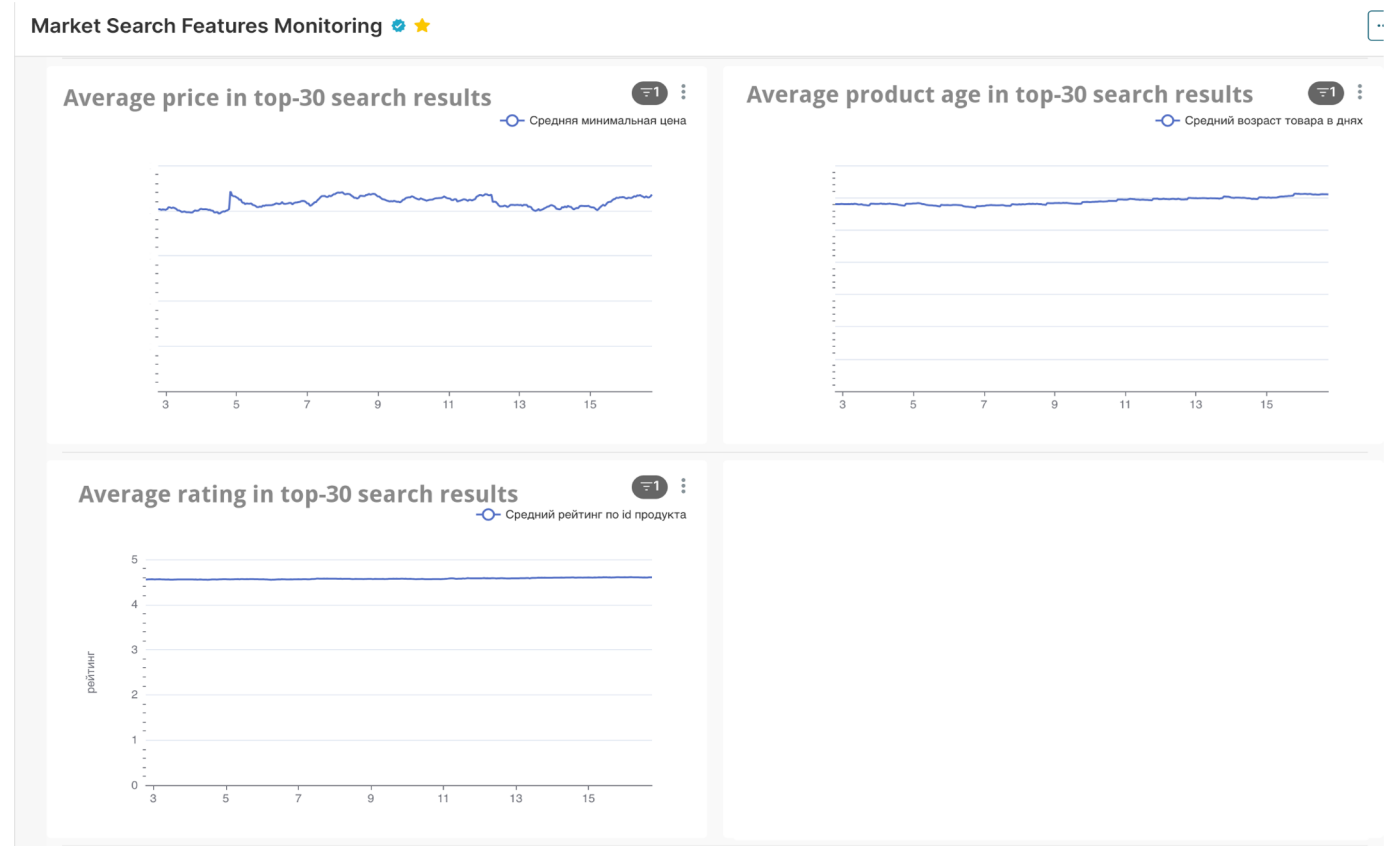
# 4. Stable results



Monitoring & Maintenance

# THE MORE YOU MONITOR, THE BETTER

- Online metrics
- Ranking features distributions
- Clickstream events quality
- Airflow jobs failures



## 🧠 A FEW WORDS ABOUT THE CULTURE...

🌟 Cultivate a failure-tolerant culture.

💡 Failure is part of the journey - embrace it.

🌱 Extract knowledge from failures.

🔍 Every hypothesis fuels the growth.



# Q&A time

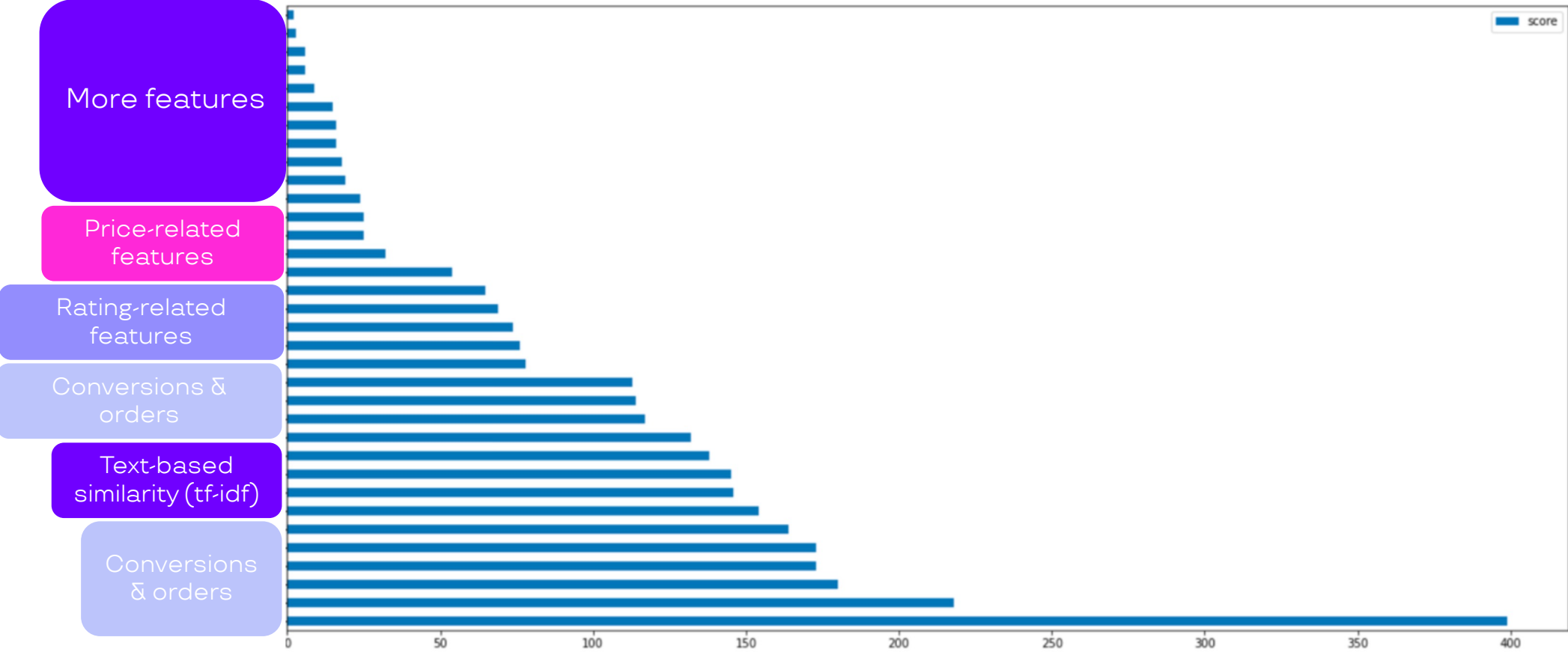
Andrey Kulagin

<https://www.linkedin.com/in/andkul/>

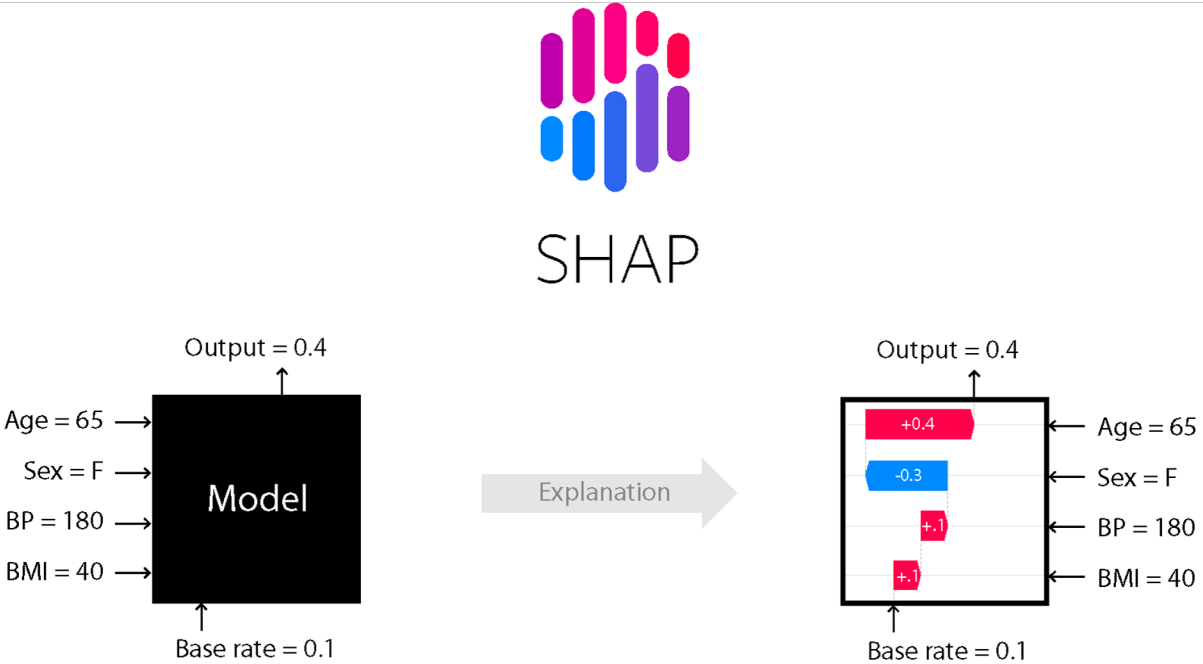
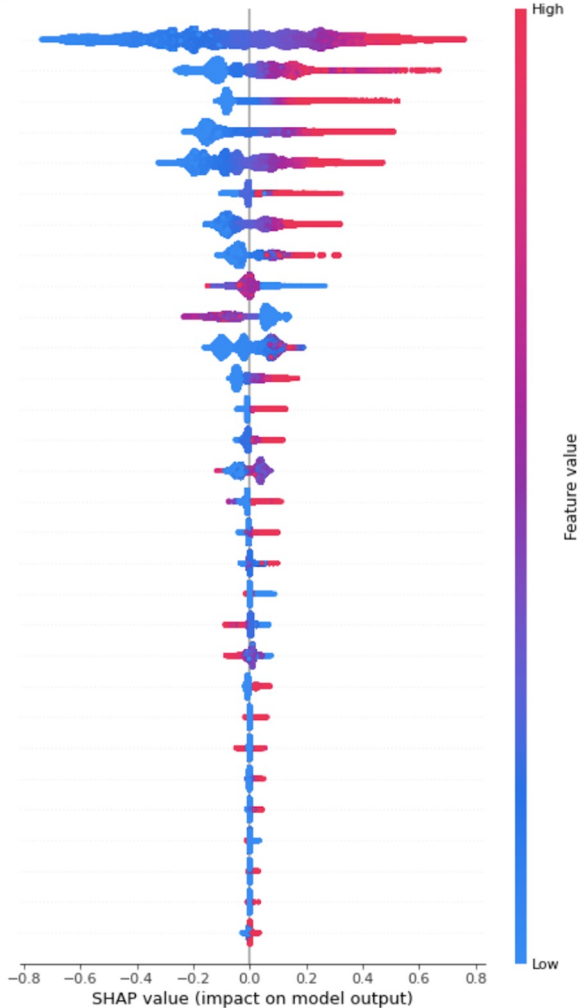
tg: @and\_kul



# FEATURE IMPORTANCES



# SHAP VALUES



<https://shap.readthedocs.io/en/latest/>