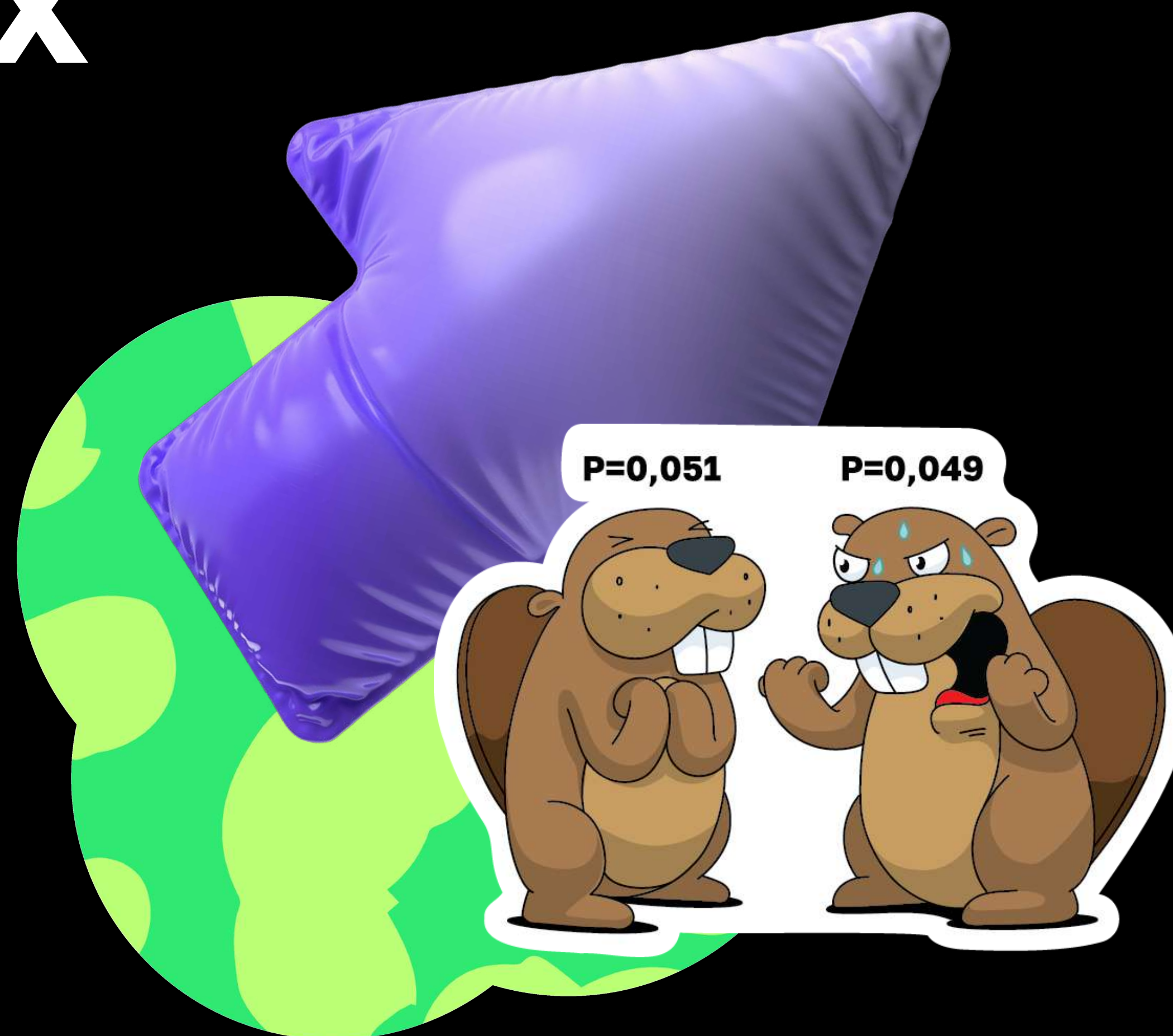




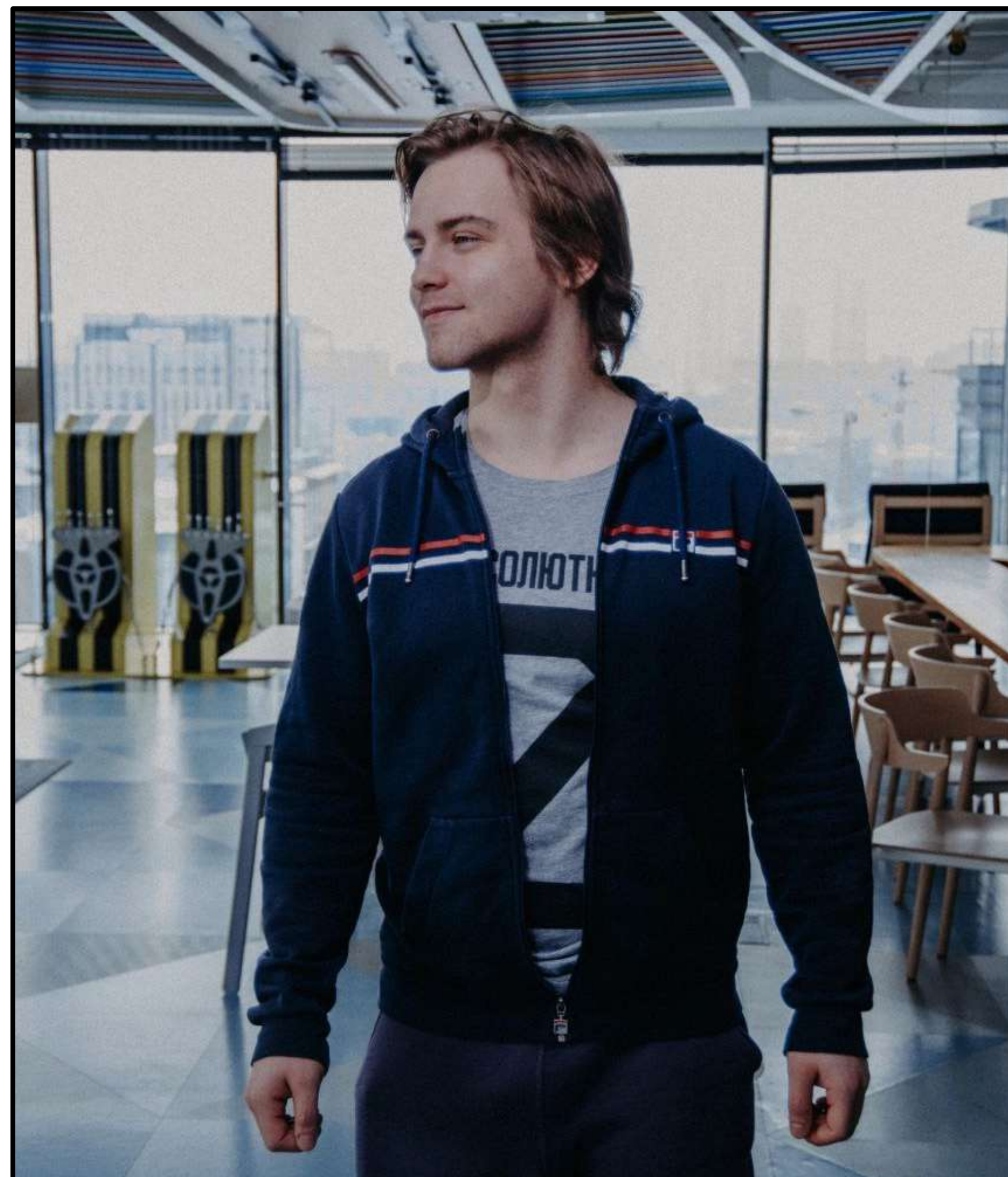
# От марковских моделей до SasRec





# Анатолий Мастрюков

- Работаю DS в рекомендациях 2 года
- Окончил МФТИ ФОПФ

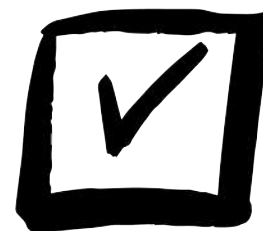




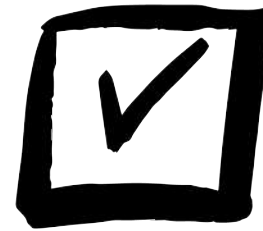
# Чем мы занимаемся?

Наша команда занимается рекомендациями  
во всем Авито

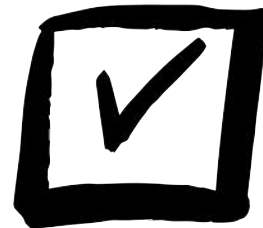
Основные продукты:



Рекомендации на главной



Похожие товары



Short videos

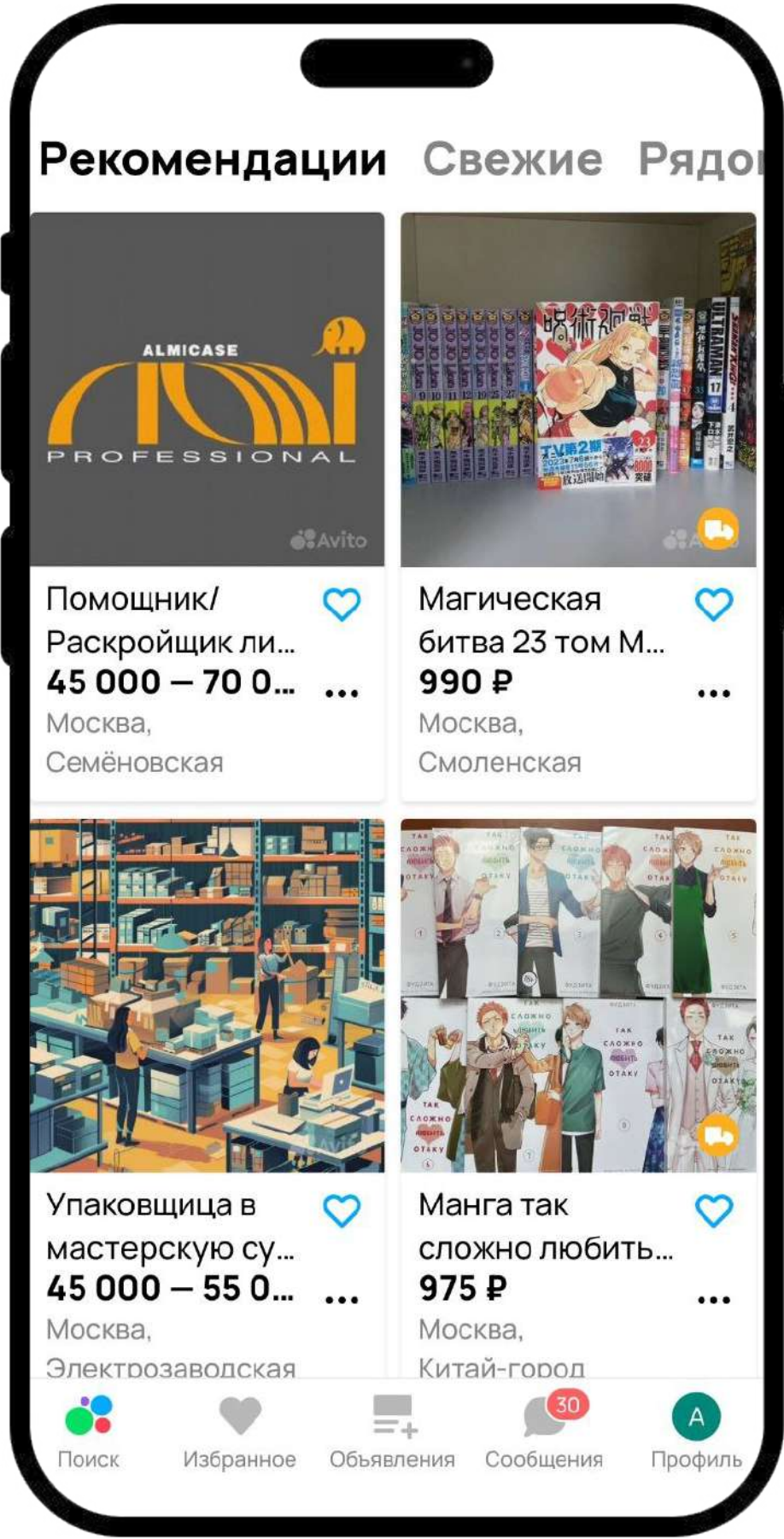


# Рекомендации на главной

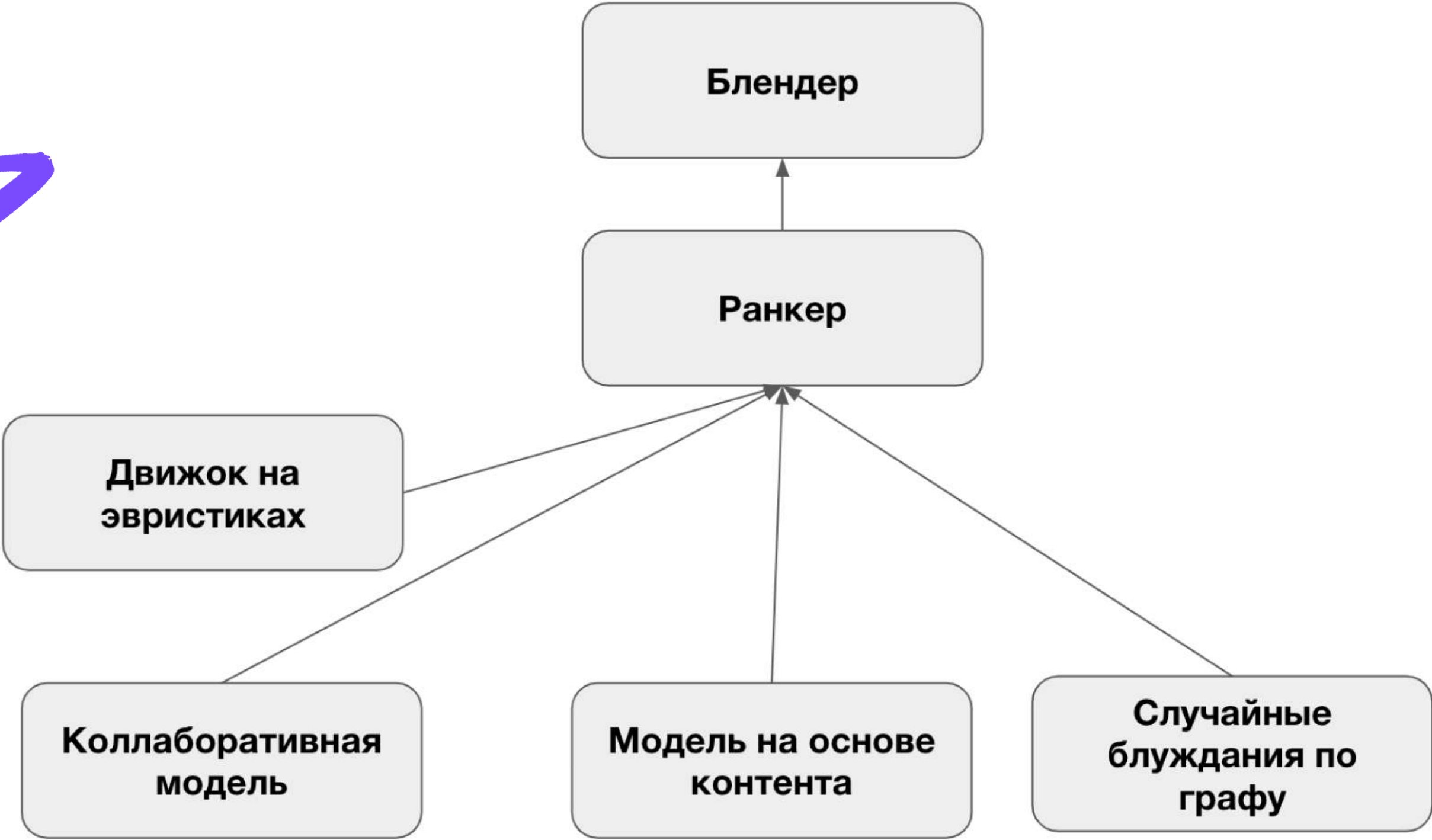
Основной продукт команды

28%

пользователей  
контактируют  
на главной



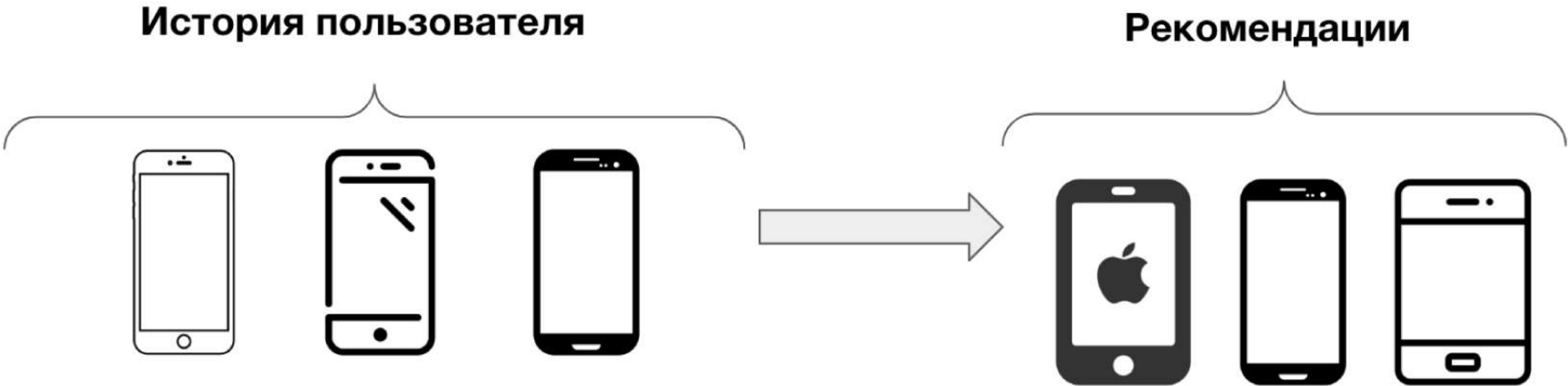
# Архитектура





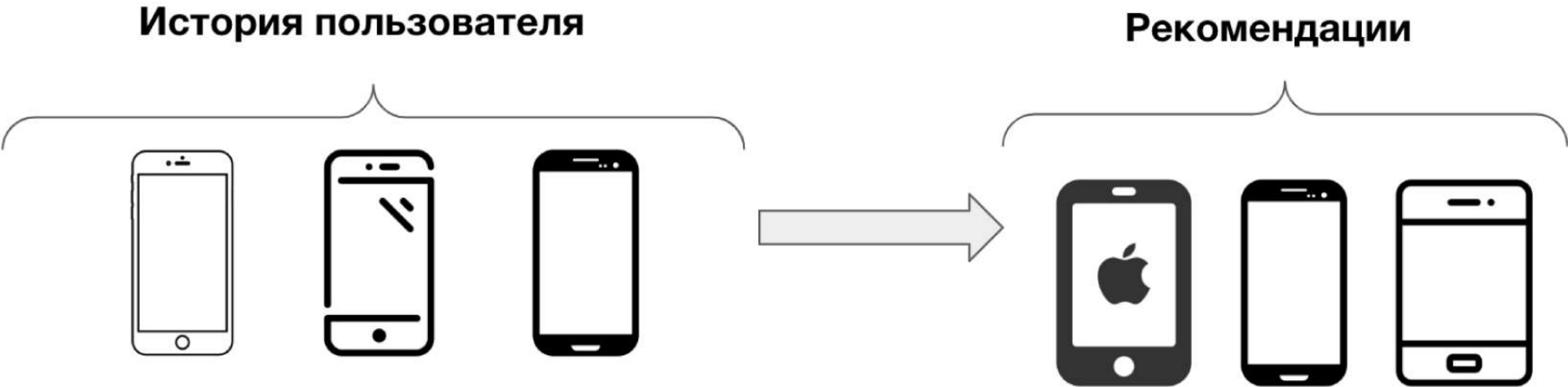
# Модели 1 уровня

Все модели создают рекомендации по категориям

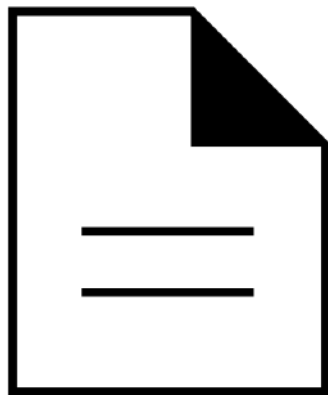


# Модели 1 уровня

Все модели создают рекомендации по категориям



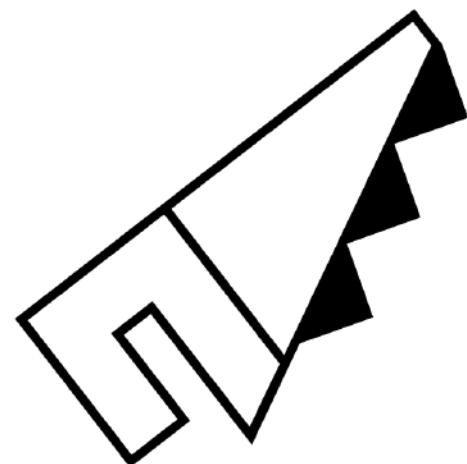
# Решение в лоб



Можно обучить гигантскую  
коллаборативную модельку



Но это тяжело: матрица **200kk** (объявления)  
x **50kk** (пользователи).



Потребуются значительные инфровые  
доработки

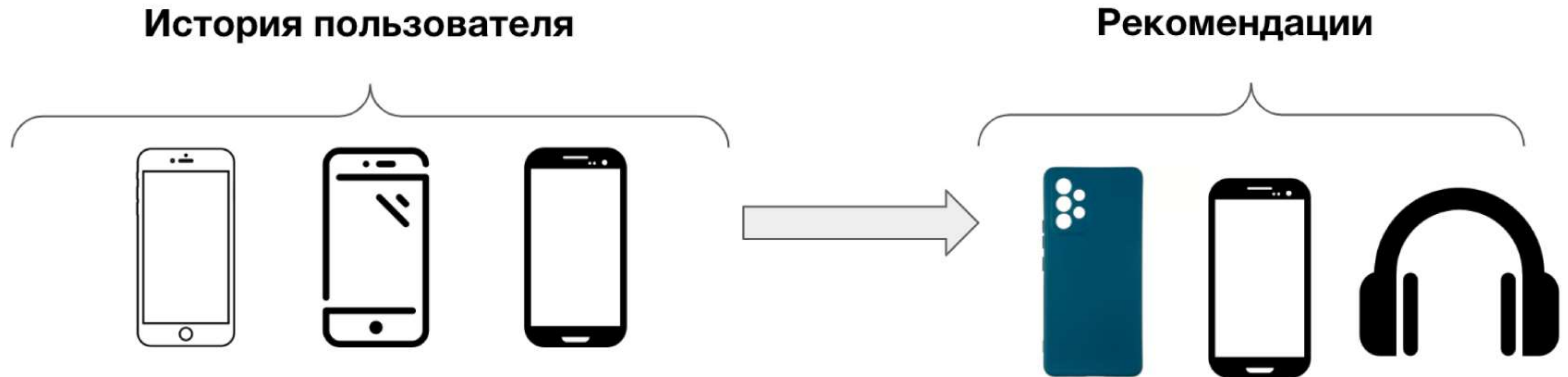


# Добавим легкую модель 1-ого уровня

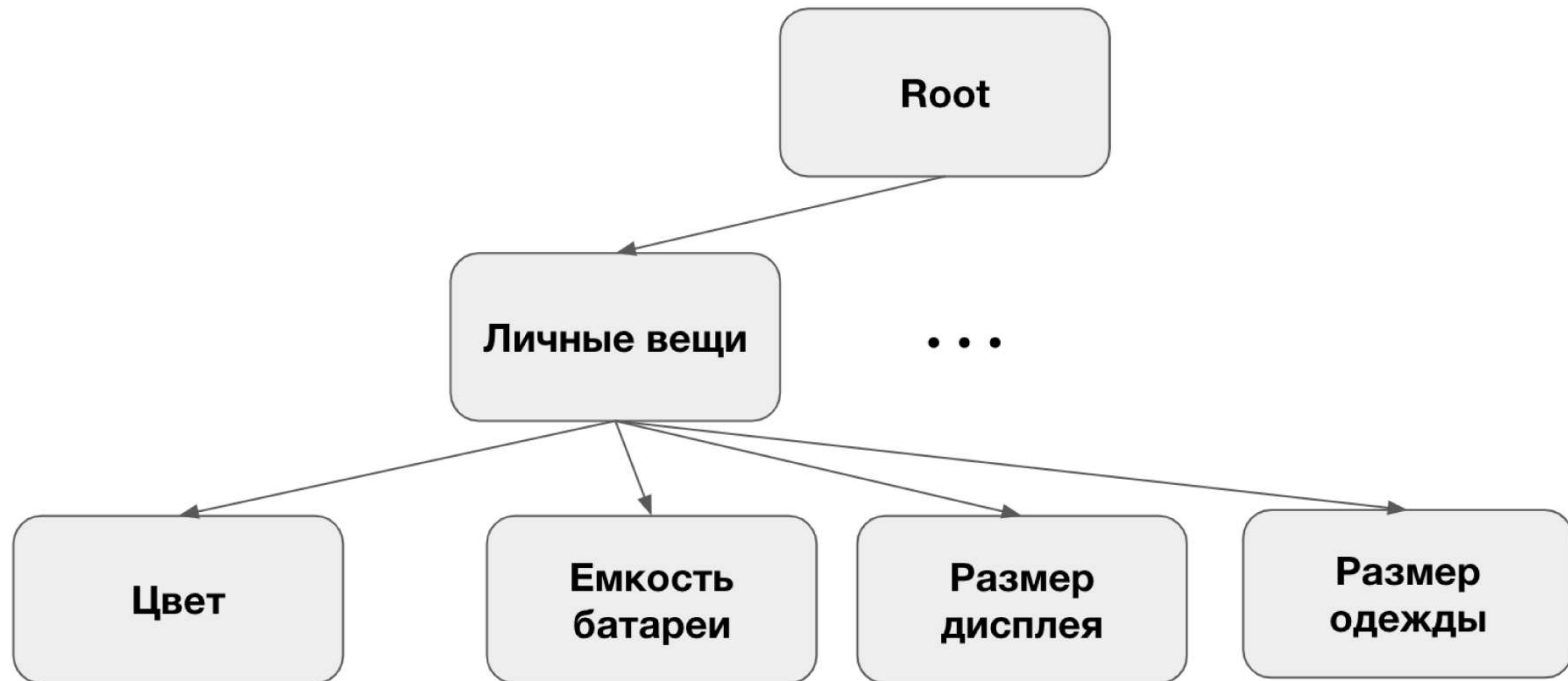


# В чем идея движка?

Хотим рекомендовать пользователю объявления из категорий которые ему могут быть интересны.

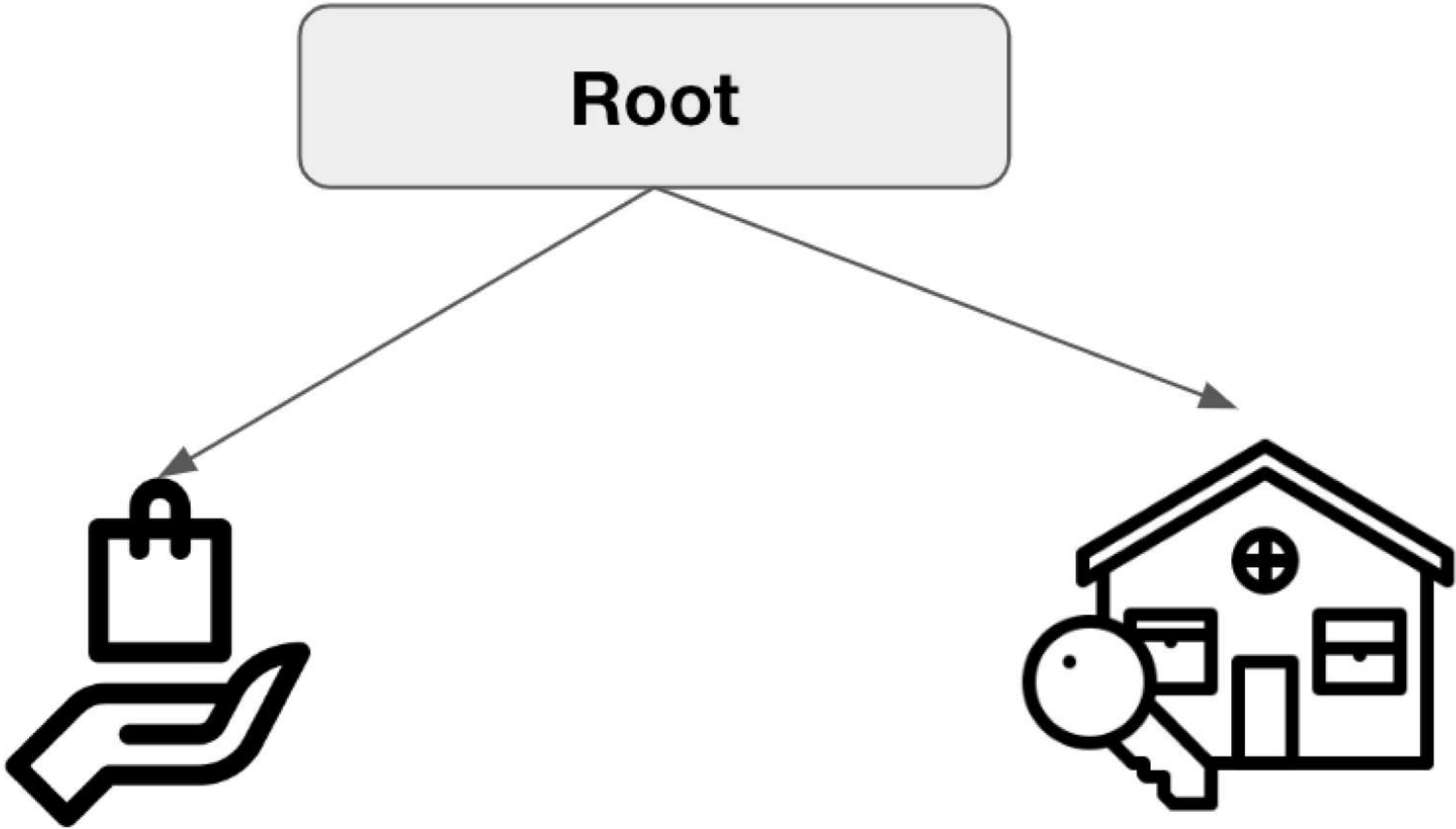


# А какие могут быть категории?



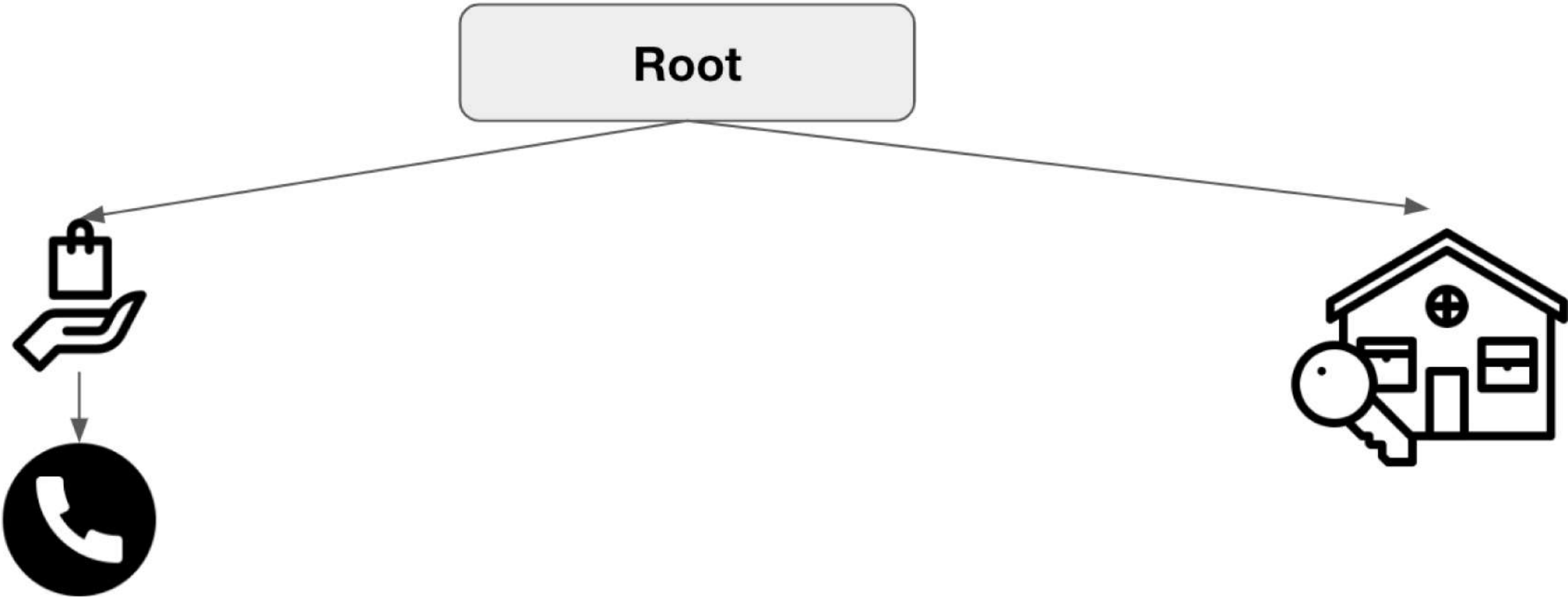


# Дерево категорий



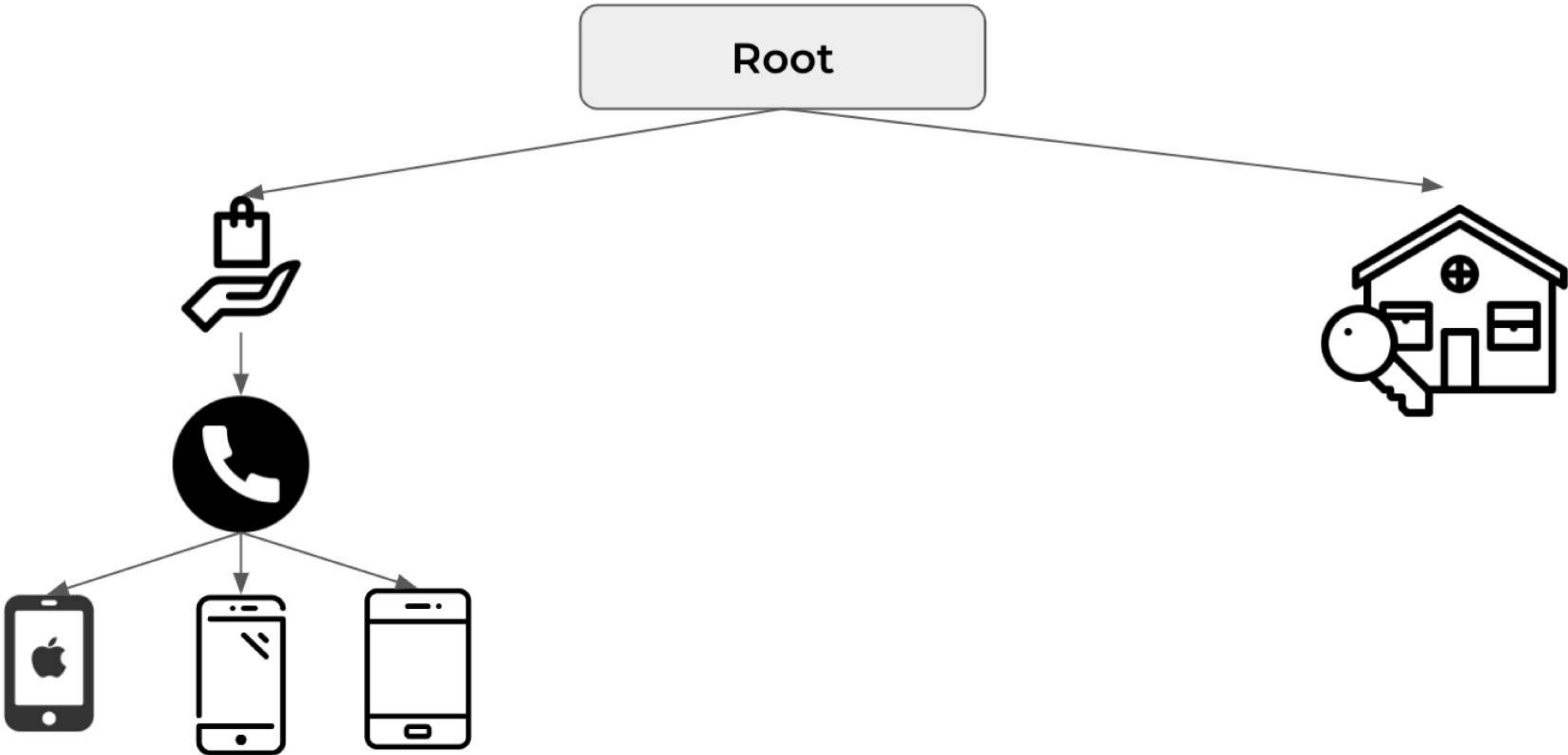
# Дерево категорий

Если показать объявления из категории «телефоны»,  
то какова вероятность, что пользователь найдет нужный?



# Дерево категорий

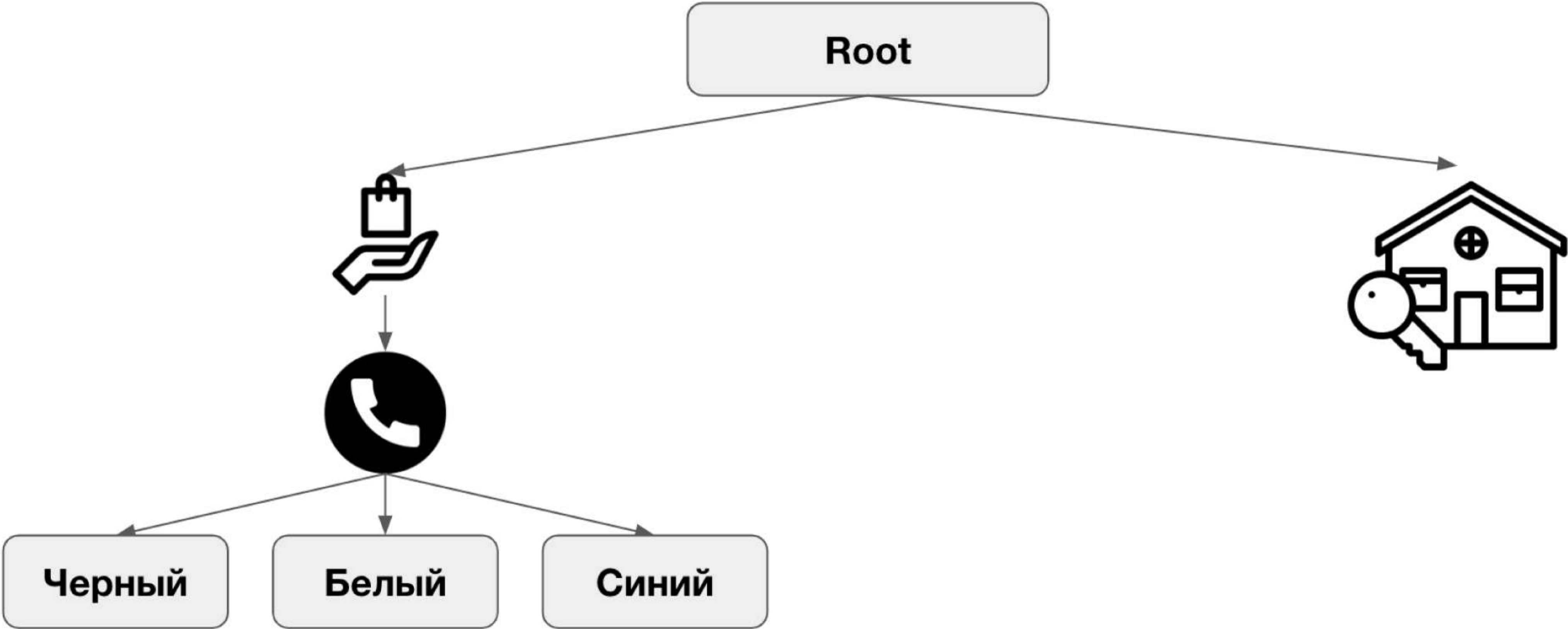
Можем продолжать разбивку дальше: цвет, память и т. д., но она должна быть осознанной.





# Дерево категорий

В случае телефона пользователю важнее технические характеристики, чем цвет.

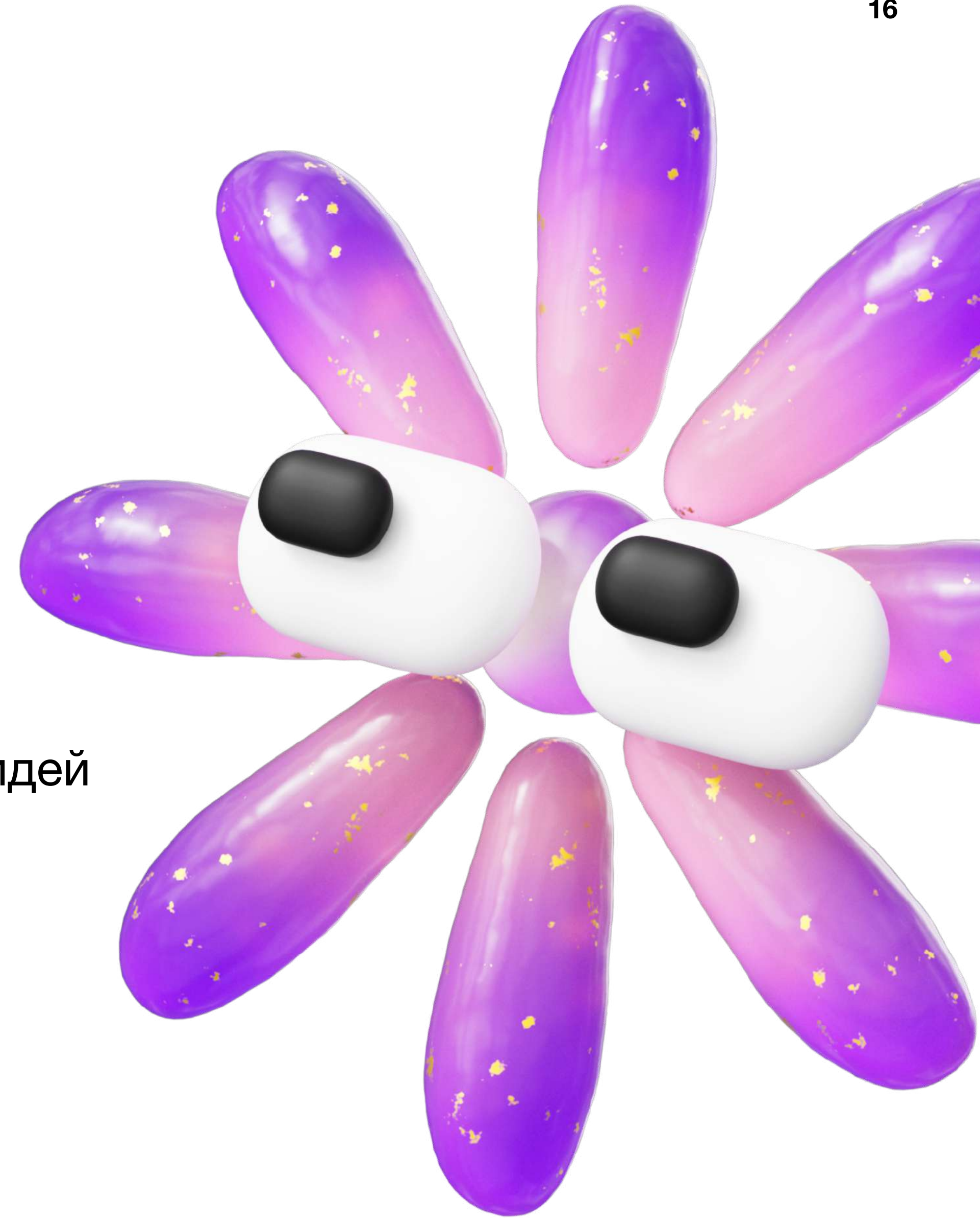


# Дерево категорий

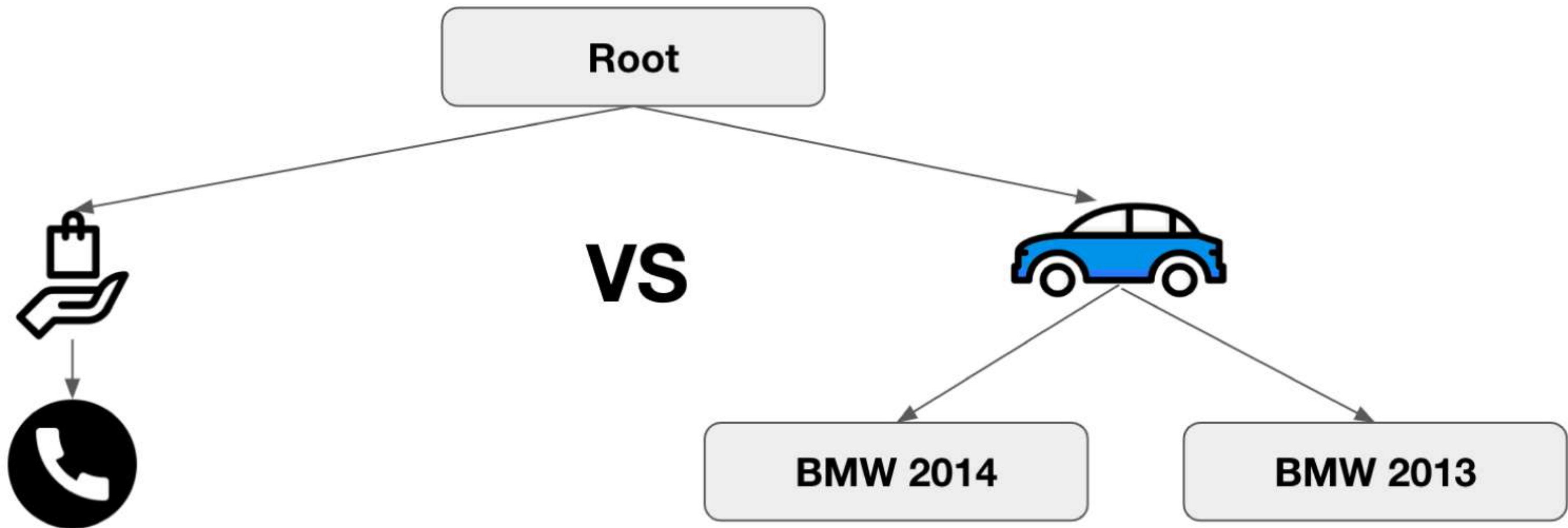
У нас уже есть такое дерево.

Свойства:

- Около 6k нод
- Может сильно меняться при появлении новых бизнес идей
- Придумано экспертно
- Некоторые отцы-ноды не сбалансированные



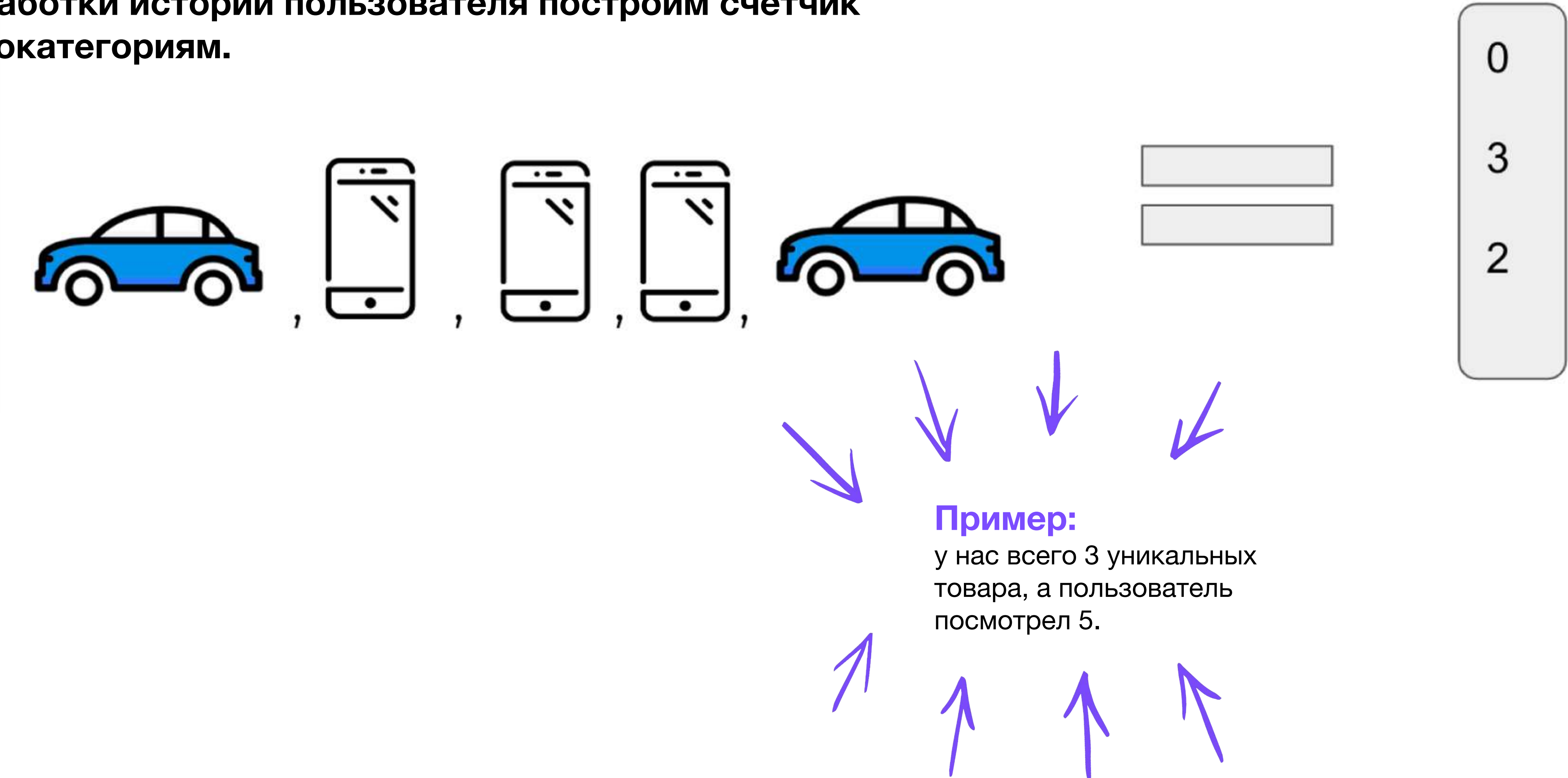
# Несбалансированность





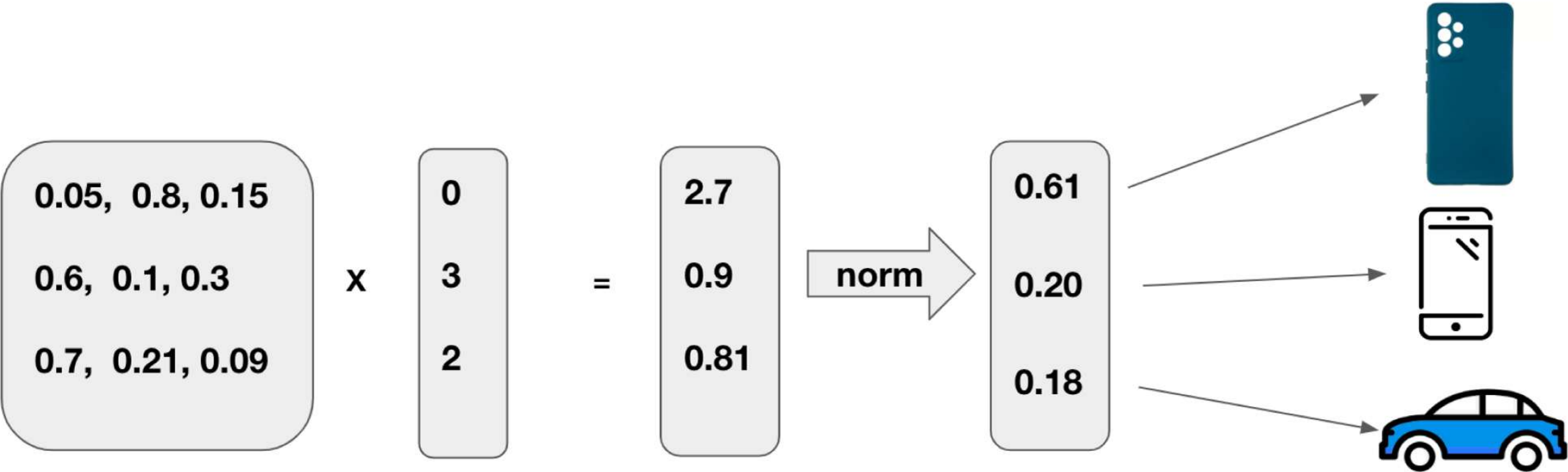
# Обработка истории

Для обработки истории пользователя построим счетчик по микрокатегориям.



# Базовая модель

В качестве модели возьмем матрицу весов (марковскую цепочку):  $R(|M| \times |M|) \rightarrow R|M|$   
Пересечение  $i$  столбца  $j$  строки - вероятность , что после просмотра  $i$  интересен  $j$ .



# Простой бейзлайн



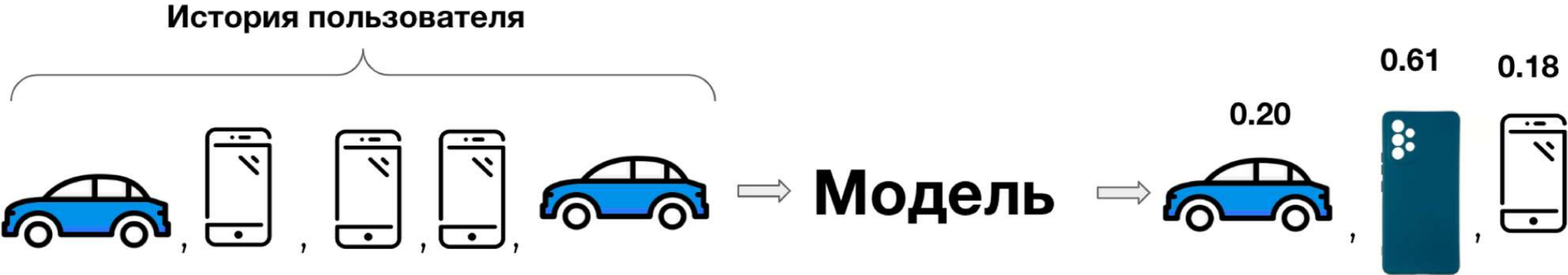


# Простой бейзлайн

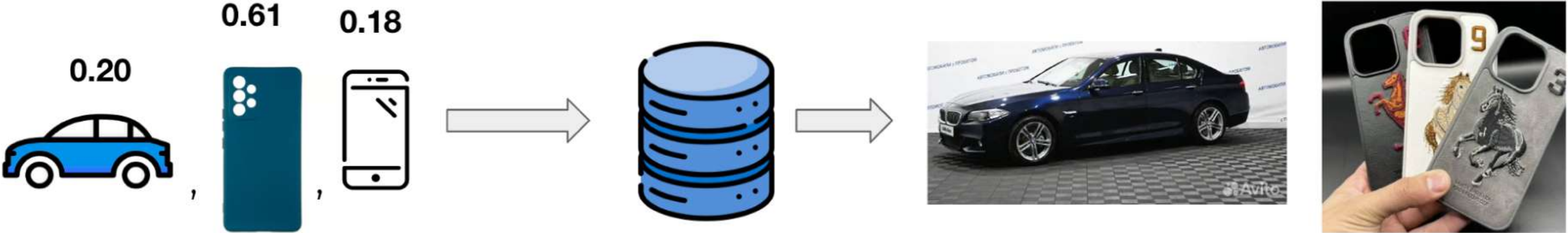
$$X = \begin{bmatrix} 0 \\ 3 \\ 2 \end{bmatrix} \quad Y = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix} \quad W = \begin{bmatrix} 0.1 & 0.3 & 0.0 \\ 0.2 & 0.4 & 0.1 \\ 0.1 & 0.1 & 0.3 \end{bmatrix}$$

$W @ X = Y_{\text{pred}}$   
 $\text{LossVector} = \log(Y_{\text{pred}}) * Y$   
 $\text{loss} = \text{LossVector.mean()}$

# В проде

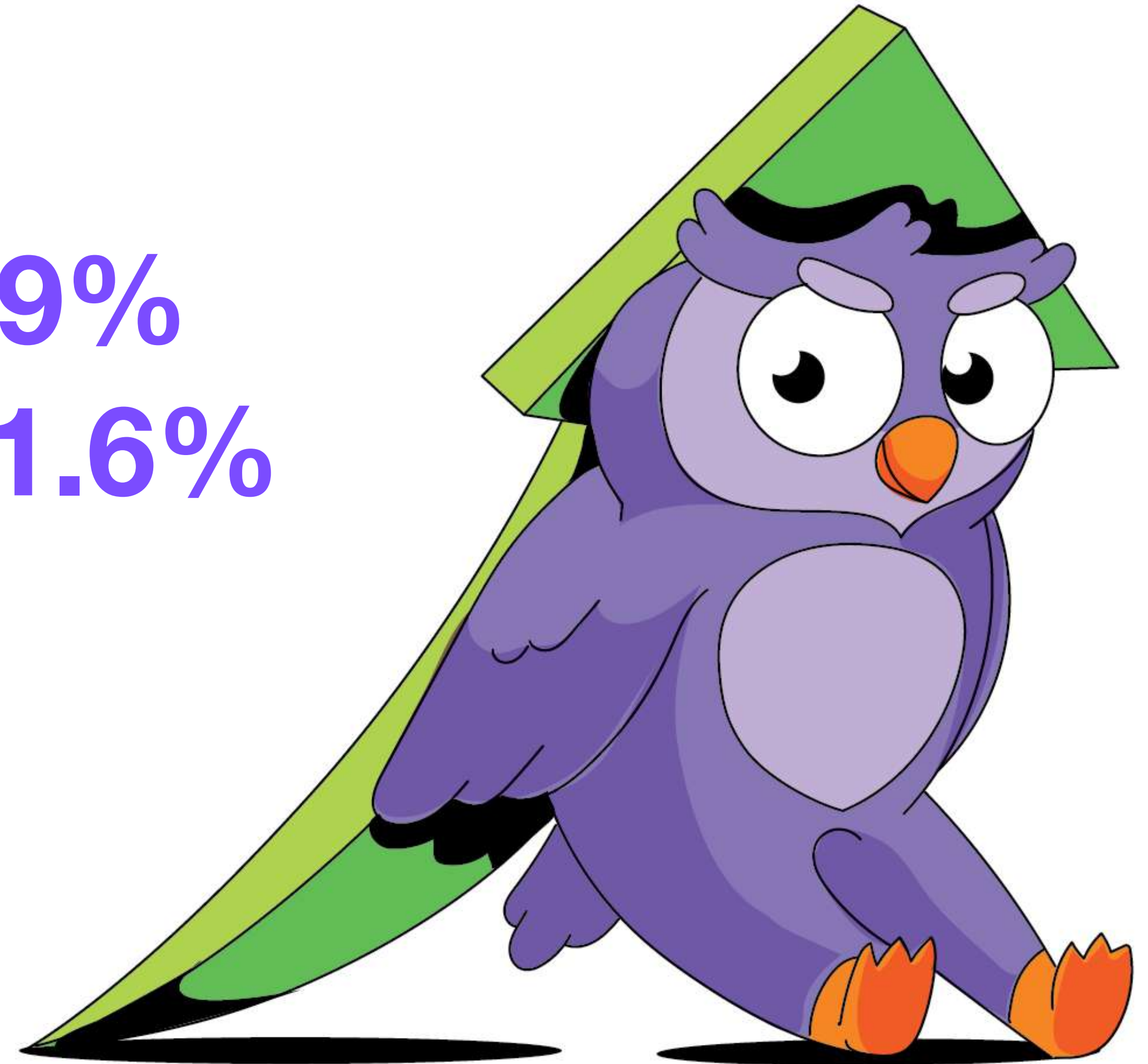


# В проде



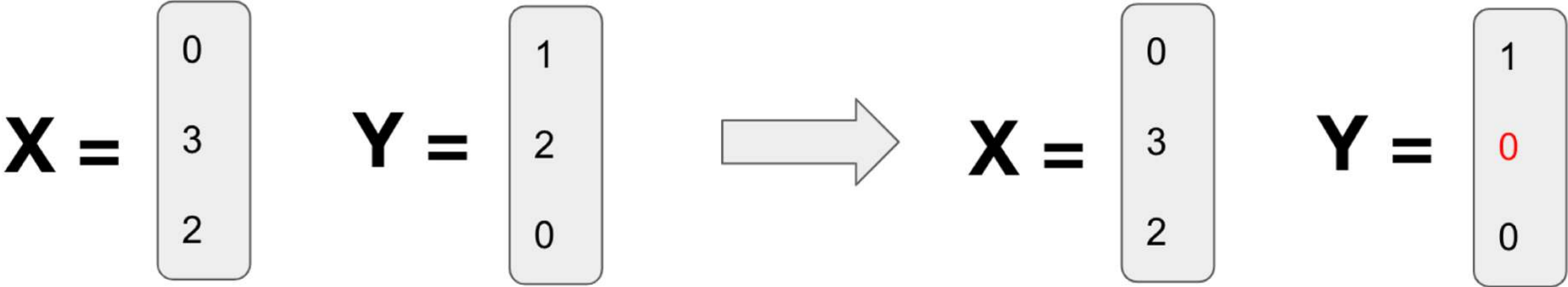
# Запустили в прод

- Пользователи с контактом с главной **+0.9%**
- Количество пользователей с контактами в новых категориях **+1.6%**
- Круто! Давайте улучшать бейзлайн





# Exclude Seen



Выросли пользователей с контактами **в новых категориях** **+0.6%**

# Переход на дерево поглубже

## Предложение:

давайте возьмем дерево поглубже.  
И такое тоже было до нас.



Получили прирост пользователей  
с контактом на главной **+0.5%**

Контакты с движка **+98%**

# Вывод по марковской модели



## Достоинства:



- ❑ Легко интерпретировать
- ❑ Хорошо работает с малочисленными товарами

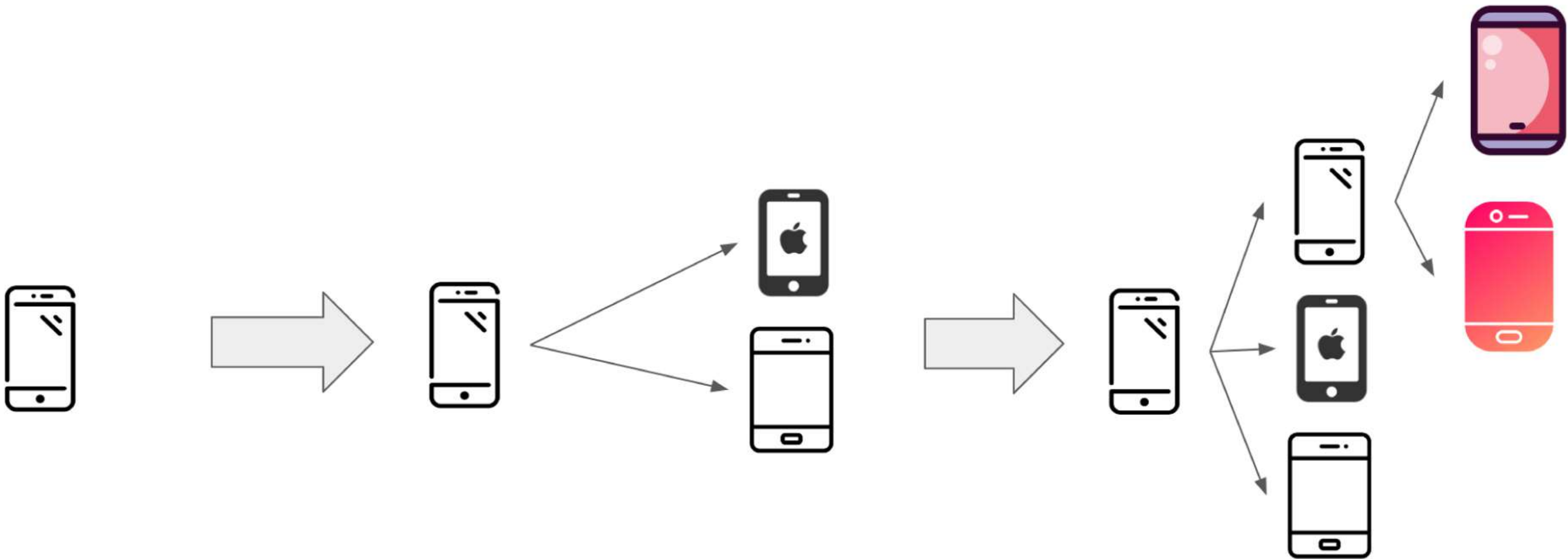
## Недостатки:



- ❑ Рекомендации бывают сильно тривиальными
- ❑ При большом количестве товаров матрица будет очень тяжелой



# Стратегия по развитию движка






# Рост модели






# Какие существуют современные подходы?

Статьи по отбору кандидатов за последние годы:

-  **gSASRec: Reducing Overconfidence in Sequential Recommendation Trained with Negative Sampling**
-  **Contrastive Learning for Sequential Recommendation**
-  **Intent Contrastive Learning for Sequential Recommendation**

# Даже приносят пользу в индустрии

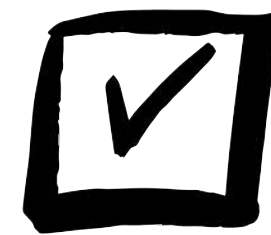
Статьи по отбору кандидатов за последние годы:

-  **PinnerSage: Multi-Modal User Embedding Framework for Recommendations at Pinterest**
-  **Behavior Sequence Transformer for E-commerce Recommendation in Alibaba**
-  **End-to-end Learnable Clustering for Intent Learning in Recommendation**

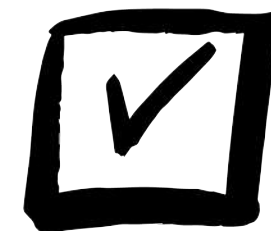


# Основные тренды

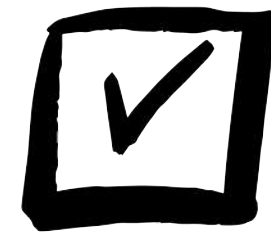
Огромное количество направлений для развития архитектуры.



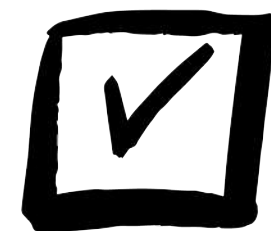
Несколько векторов пользователей для разных интенгов



Обработка контента айтемов



Self-supervised learning для последовательностей



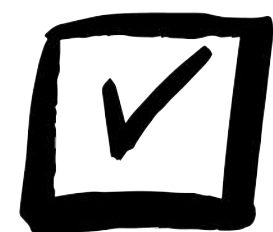
Long memory вектора



# Вишенка на торте



Везде используются **трансформеры**.



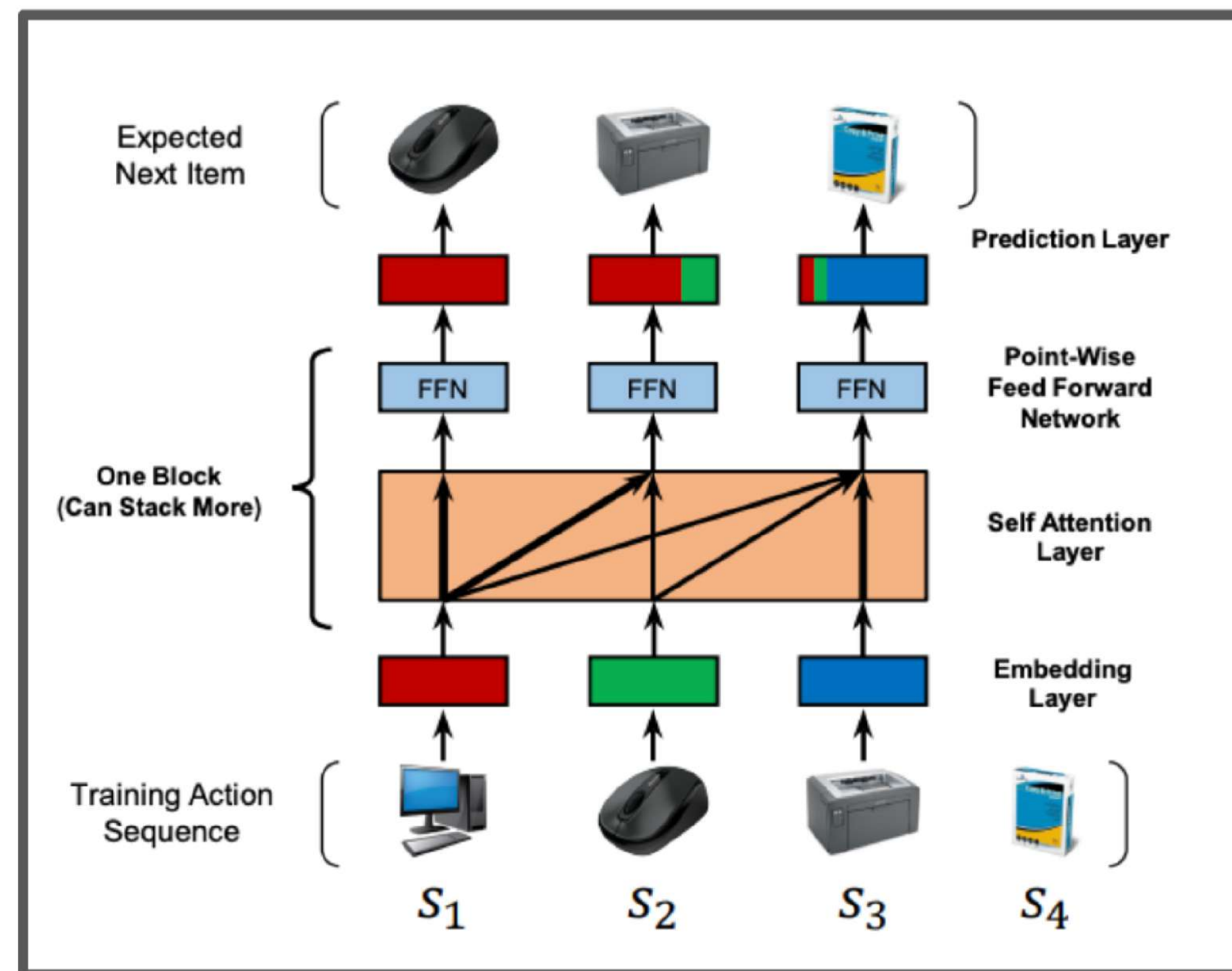
2 основные архитектуры: **bert4rec** и **sasrec**.





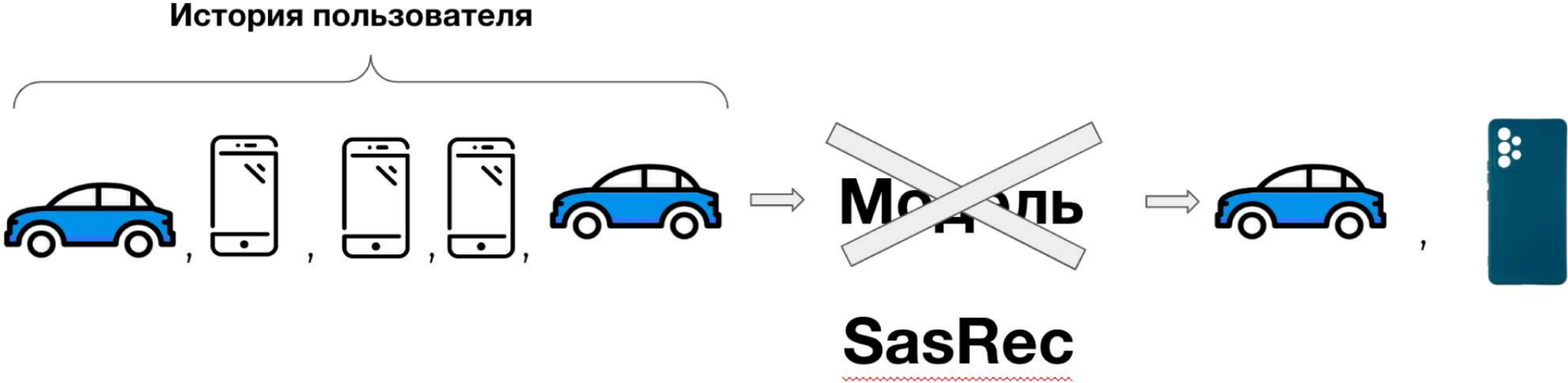
# Адаптируем архитектуру для нас

- За место объявлений будем подавать на вход тип товара
- Храним колонку категорий
- Можно подавать сырую историю



Self-Attentive Sequential Recommendation

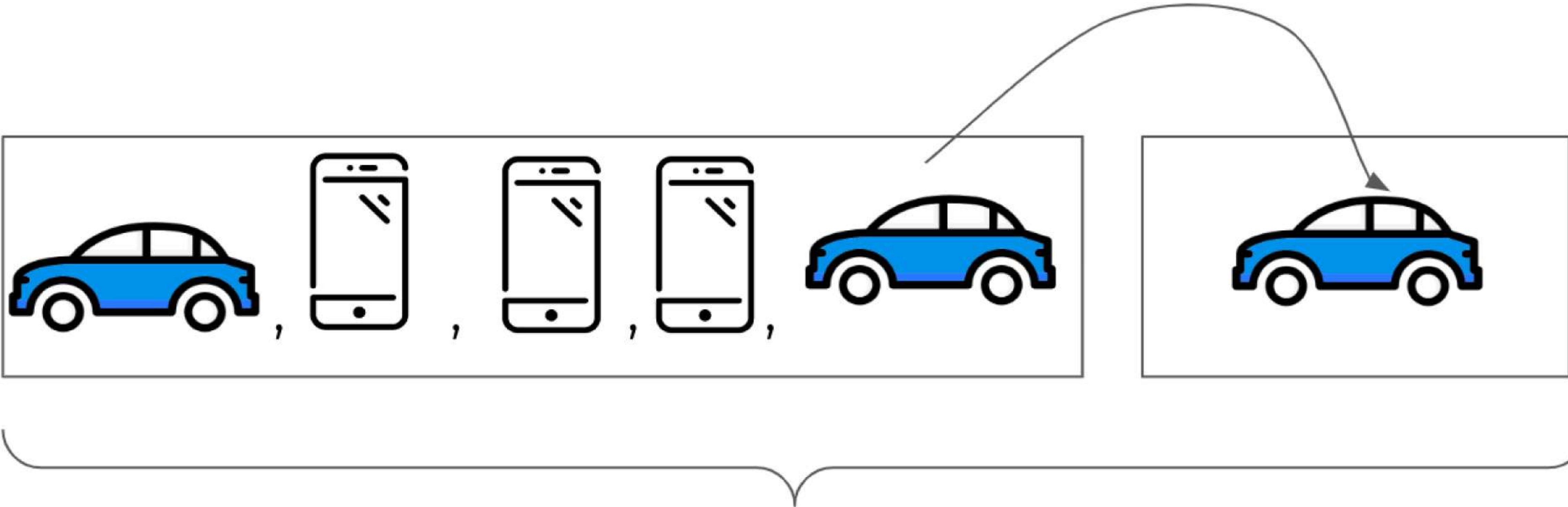
# Инференс модели не меняется





# Оффлайн приемка

Считаем метрики ранжирования для next item prediction.



# Сравним бейзлайны

Модель	$\Delta$ , Recall@10	$\Delta$ , NDCG@10	$\Delta$ , Latency	$\Delta$ , num params
markov chain	0%	0%	0%	0%
SasRec	+15%	+18%	-66%	-95%

# Против коллаборативной модели

Модель	$\Delta$ , Recall@10	$\Delta$ , NDCG@10	$\Delta$ , Latency	$\Delta$ , num params
ALS	0%	0%	0%	0%
<u>SasRec</u>	+3%	+4%	0%	-99%

# Перебор параметров

Пробовали перебор количества голов — статзначимых результатов не получили.

SasRec, num blocks	$\Delta$ , Recall@10	$\Delta$ , NDCG@10	$\Delta$ , Latency	Latency, ms
1	0%	0%	0%	8.4
2	+9%	+9%	+20%	10
3	+10%	+10%	+44%	12.1



# АБТ #1

## 01

Оставили старый ранкер,  
но подмешиваем кандидатов  
от новой модели.

## 02

В первые дни **-2%** precision@5 ,  
score ранкера **+0.3%**

**Вывод:**

Ранкер не готов к новым кандидатам



# АБТ #2

## 01

Собрали выдачу от ранкера  
и переобучили модель.

## 02

Конверсия движка **+17%** ,  
precision@5 **+1.6%**

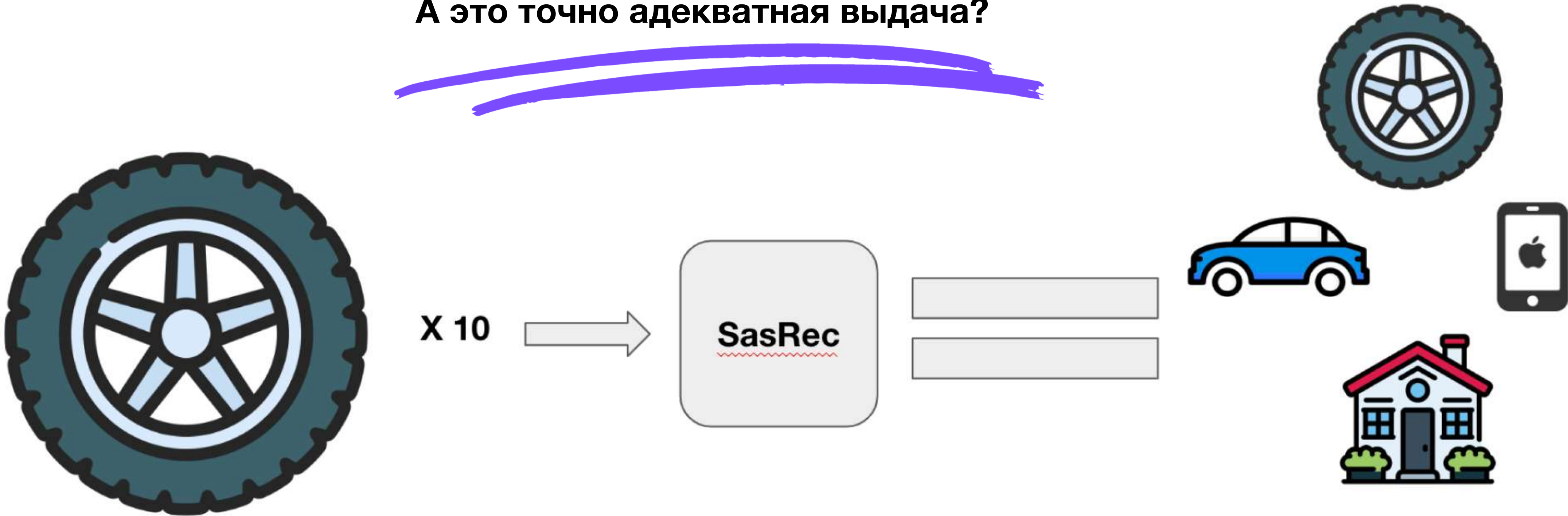
## 03

Но контакты в totale  
не прокрасились,  
подозрительный трафик  
в отдельных категориях.

# А что с quality анализом?

Проверяли модель в отдельных кейсах.

А это точно адекватная выдача?

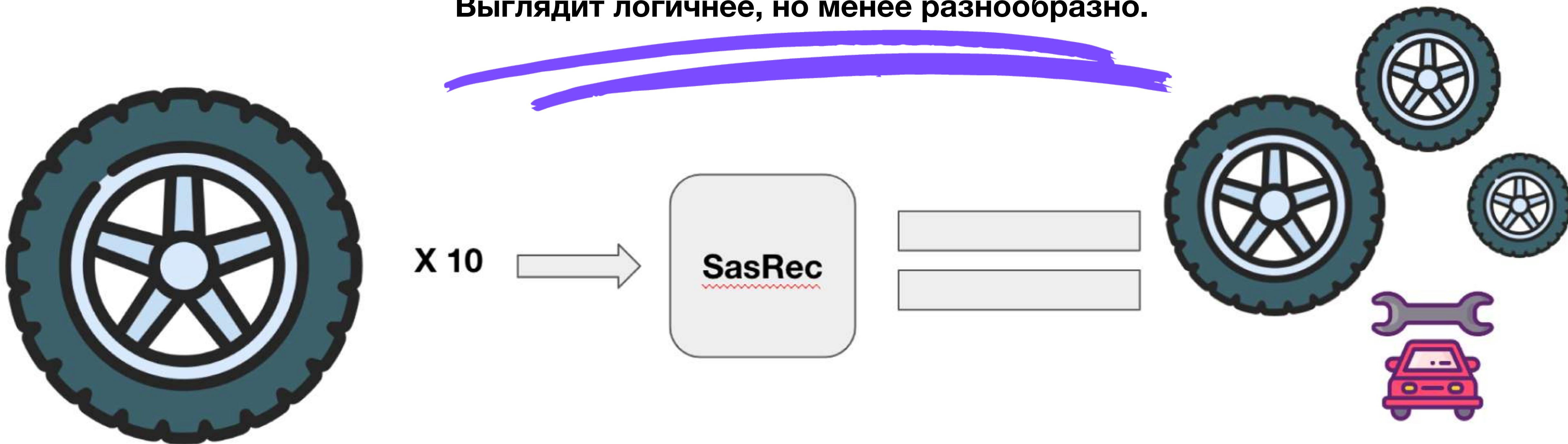




# А что с quality анализом?

Давайте обучать модель только на контактах, а не на всем кликстриме.

Выглядит логичнее, но менее разнообразно.





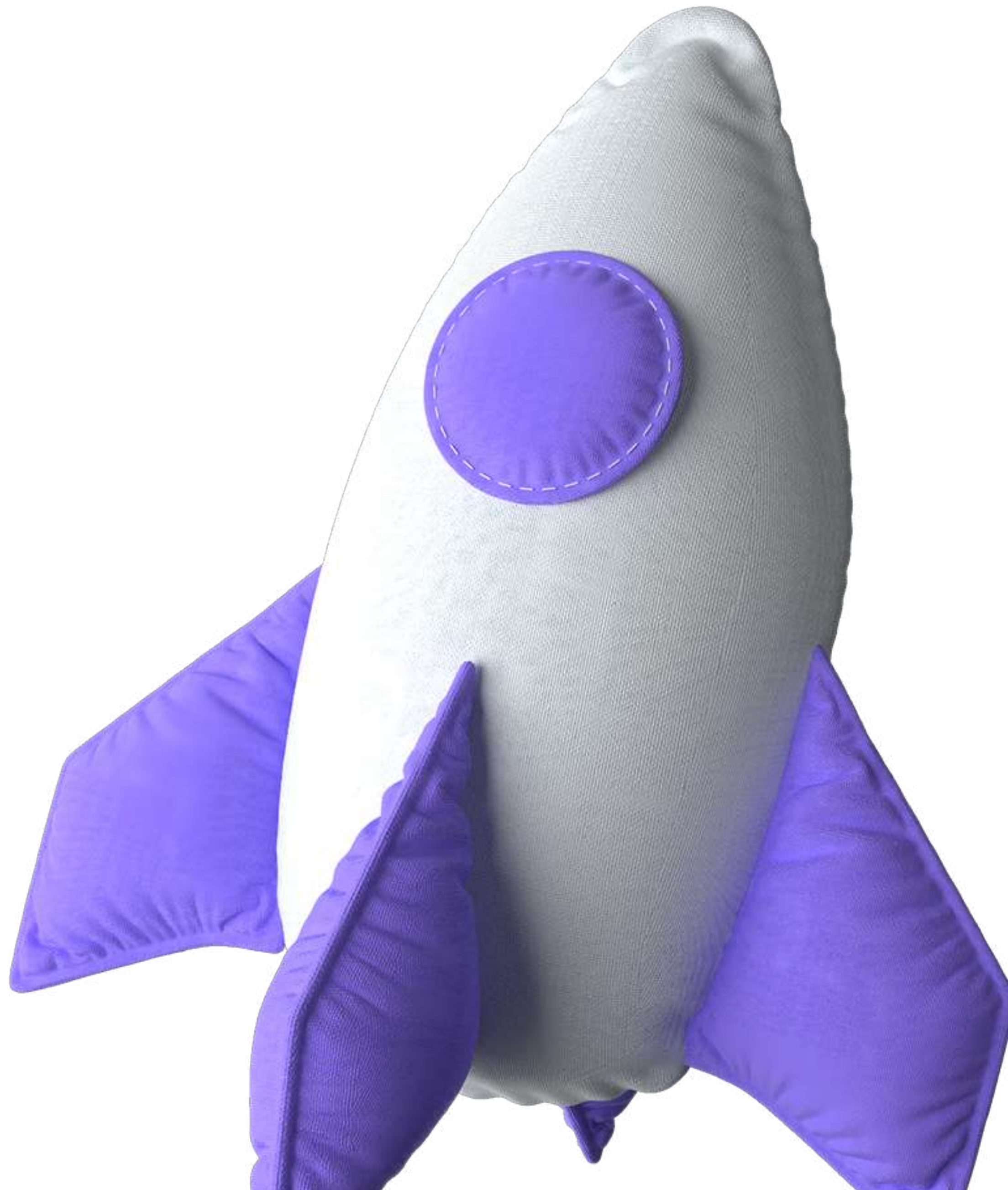
# АБТ #3

## 01

Запустили АБТ с модель  
переобученной на события  
контакта

## 02

Ждем результатов.



# Что не зашло:

## 01

Смотрим на скор ранкера в разных группах АБ - нет статзначимого различия

## 02

МЦ пробовали предсказывать айтемы в следующую сессию/день — хуже в онлайн, а лучше в оффлайне

## 03

Трансформеры «из коробки», готовые библиотеки

# Выводы :

## 01

Предпочтительнее  
использовать максимально  
кастомные модели.

## 02

Проверять quality ожидания  
от модели.

## 03

Удобно обкатать на простой  
модели пайплайн, а потом  
переходить к моделям  
потяжелее.



# Планы :

## 01

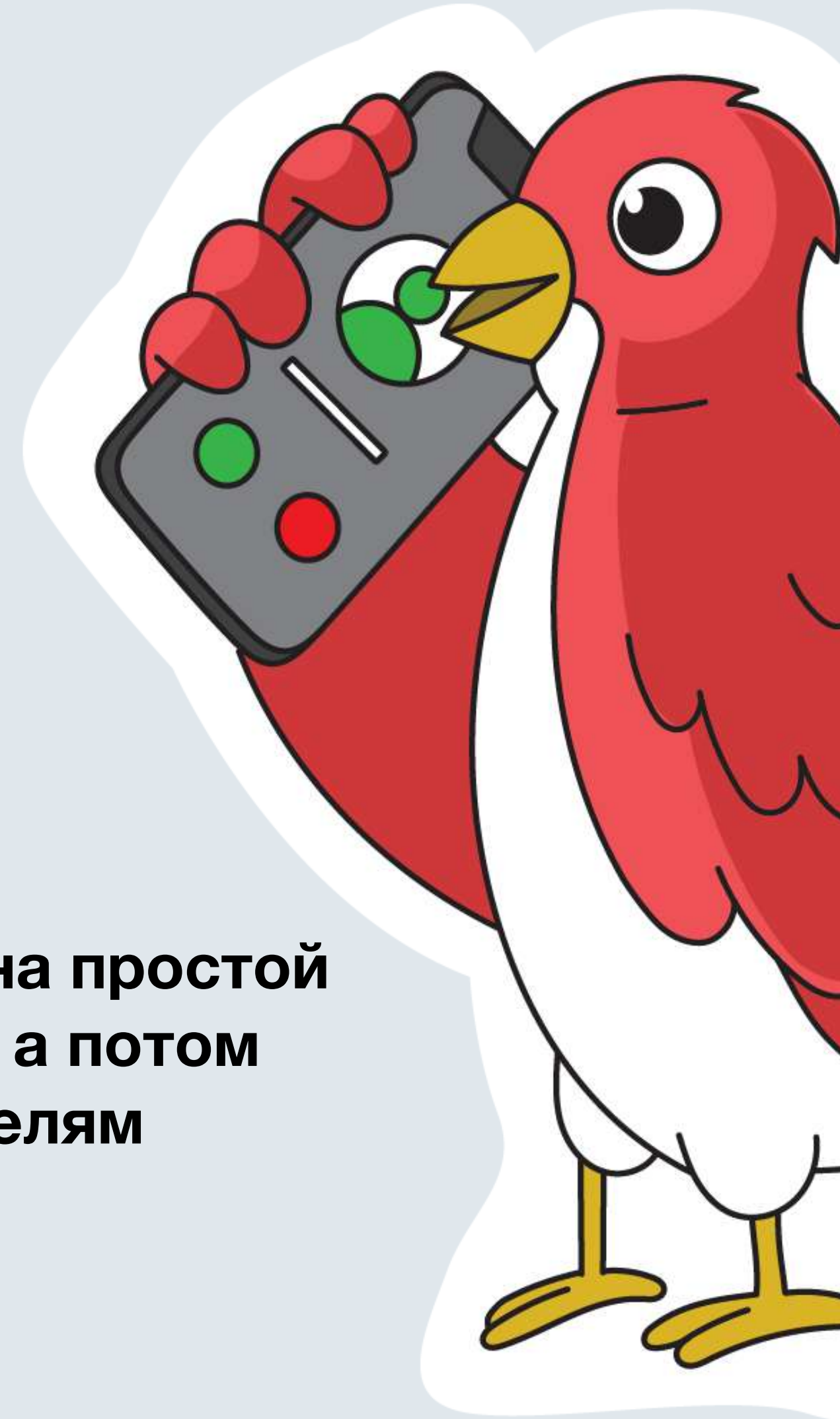
Предпочтительнее  
использовать максимально  
кастомные модели.

## 02

Проверять quality  
ожидания от модели.

## 03

Удобно обкатать на простой  
модели пайплайн, а потом  
переходить к моделям  
потяжелее.





**Спасибо за  
внимание!**

