

Яндекс



Разбор ошибок при проектировании рекомендательной системы

Сергей Кузнецов

Tech Lead RecSys&Search Platform,
МТС Web Services

Привет!



Сергей Кузнецов

Tech Lead RecSys&Search Platform MWS

- Занимаюсь DS уже более 10 лет
- Делаем рекомендации для более чем 10 сервисов экосистемы МТС (KION, «Строки»...)
- Преподаю в ВШЭ и ИТМО

Что мы делаем?



Постановка
условий

Кодим/думаем

—
Обсуждаем
основные
рекомендатель-
ные модели

Смотрим, на что
было важно
обратить внимание

Подводим итоги

Принципы мастер-класса

- 01 Представляем, что мы делаем продакшен модели
- 02 Не столь важно успеть всё закодить, сколь важно успеть всё продумать
- 03 Гиперпараметры моделей в данном случае не важны

Скачиваем

<https://github.com/SerSmith/masterClassRecSysYa2025#>



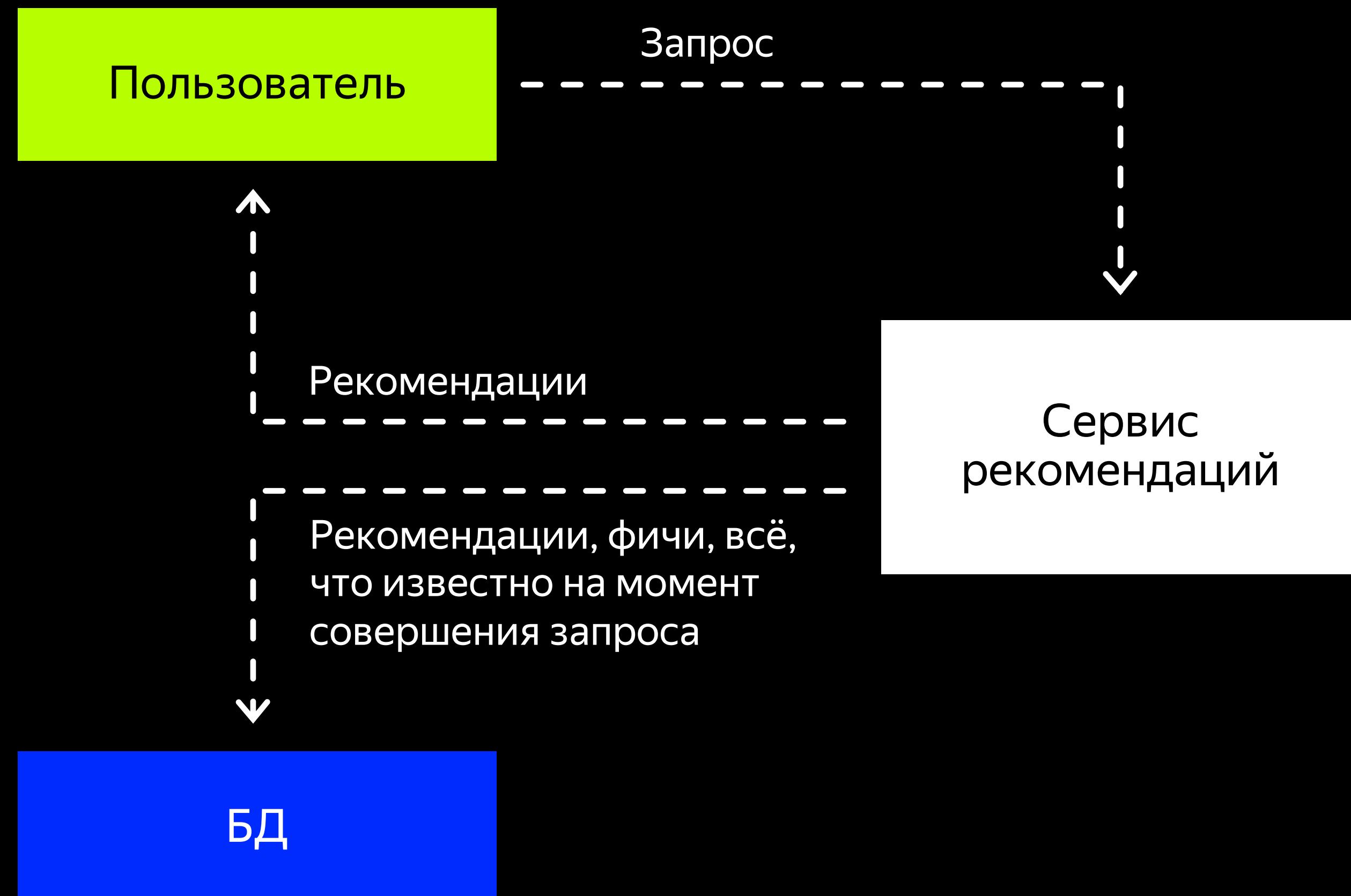
Задача

Сейчас полка ранжируется неперсонализированно на основе средней пользовательской оценки



Movie Title	Rating	Action
Мистер и миссис Смит	★ 8 / ►	ПОВТОР ЭФИРА
Перелом	★ 9,3 / ►	ПОВТОР ЭФИРА
Человек, который изменил всё	★ 8,2 / ►	ПОВТОР ЭФИРА
Не тот сын	★ 8 / ►	ПОВТОР ЭФИРА
Птичка на проводе	★ 8,3 / ►	ПОВТОР ЭФИРА
Идеальный шторм	★ 8,3 / ►	ПОВТОР ЭФИРА

Обучаемся на логах



Задача

Вам повезло — стажёр уже подготовил своё решение, вам предлагается посмотреть на него и подумать, что могло пойти не так



Устройство решения:

1. Обучаем ALS, EASE и SASRec
2. Стакаем их в модель бустинга

Данные интеракций



# user_id	# item_id	A\saction	A\sdatetime
5825	3.0	click	2025-03-28 07:43:51.136213735
4878	30.0	show	2025-01-21 23:02:41.540121919
1397	93.0	show	2025-01-28 09:26:58.670975772
1411	80.0	show	2025-01-06 04:04:18.822920341
5590	59.0	show	2025-02-13 20:07:52.951001645
9898	69.0	show	2025-02-19 08:37:58.351036063
8891	57.0	click	2025-03-06 09:58:55.841114114
7283	43.0	show	2025-03-13 19:42:07.261953402
7149	69.0	show	2025-01-20 15:07:37.656699670
7818	75.0	show	2025-01-21 09:37:45.843542689

Данные пользователей



#	user_id	sex_type
0		F
1		M
2		M
3		F
4		F
5		F
6		F
7		M
8		M
9		M

Данные объектов



# item_id	genre	# mean_users_score
0	Комедия	1.8727005942368125
1	Комедия	4.75357153204958
2	Триллер	3.659969709057026
3	Драма	2.993292420985183
4	Комедия	0.7800932022121826
5	Мелодрама	0.7799726016810132
6	Драма	0.2904180608409973
7	Драма	4.330880728874676
8	Фантастика	3.005575058716044
9	Мелодрама	3.540362888980227

ИТОГО

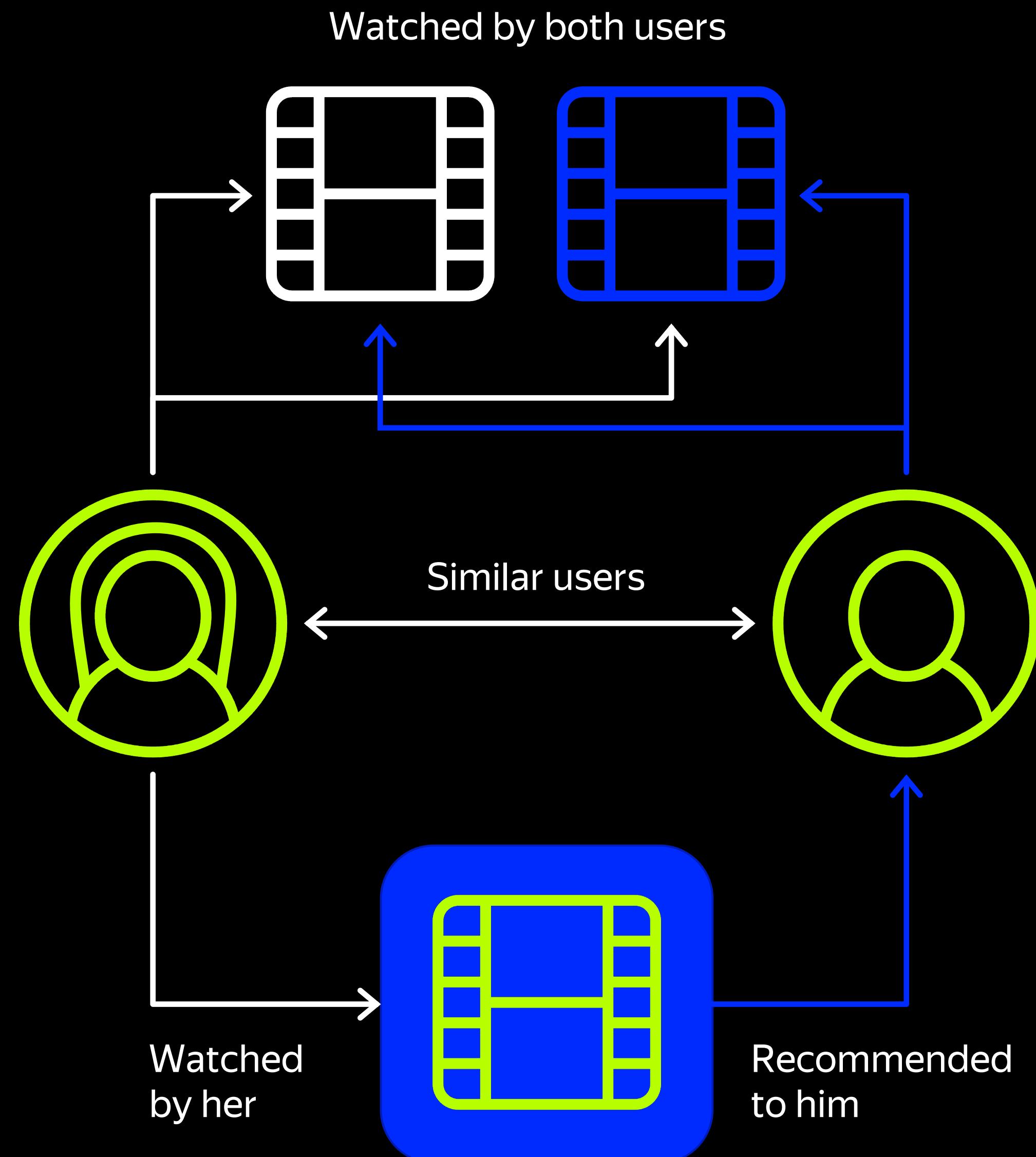
- Строим перанжирование, есть ноутбук с ошибками
- Ошибки идеологические, а не в синтаксисе
- Данные симулируют выдачу пользователю объектов по `mean_user_score`
- В данном примере гиперпараметры моделей не важны

Поехали!



Коллаборативный сигнал

- Алгоритмы нацелены на предсказание матрицы интеракций и обычно не используют дополнительную информацию
- Требуется регулярное переобучение
- Наиболее часто применяем модели ALS, EASE, SANSA...



ALS

INTERACTIONS

=

ITEMS

x

USERS

EASE

$$\|X - X \times B\|_{F^2} + \|B\|_{F^2} \rightarrow \min$$

$$s.t. diag(B) = 0$$

Контентная информация

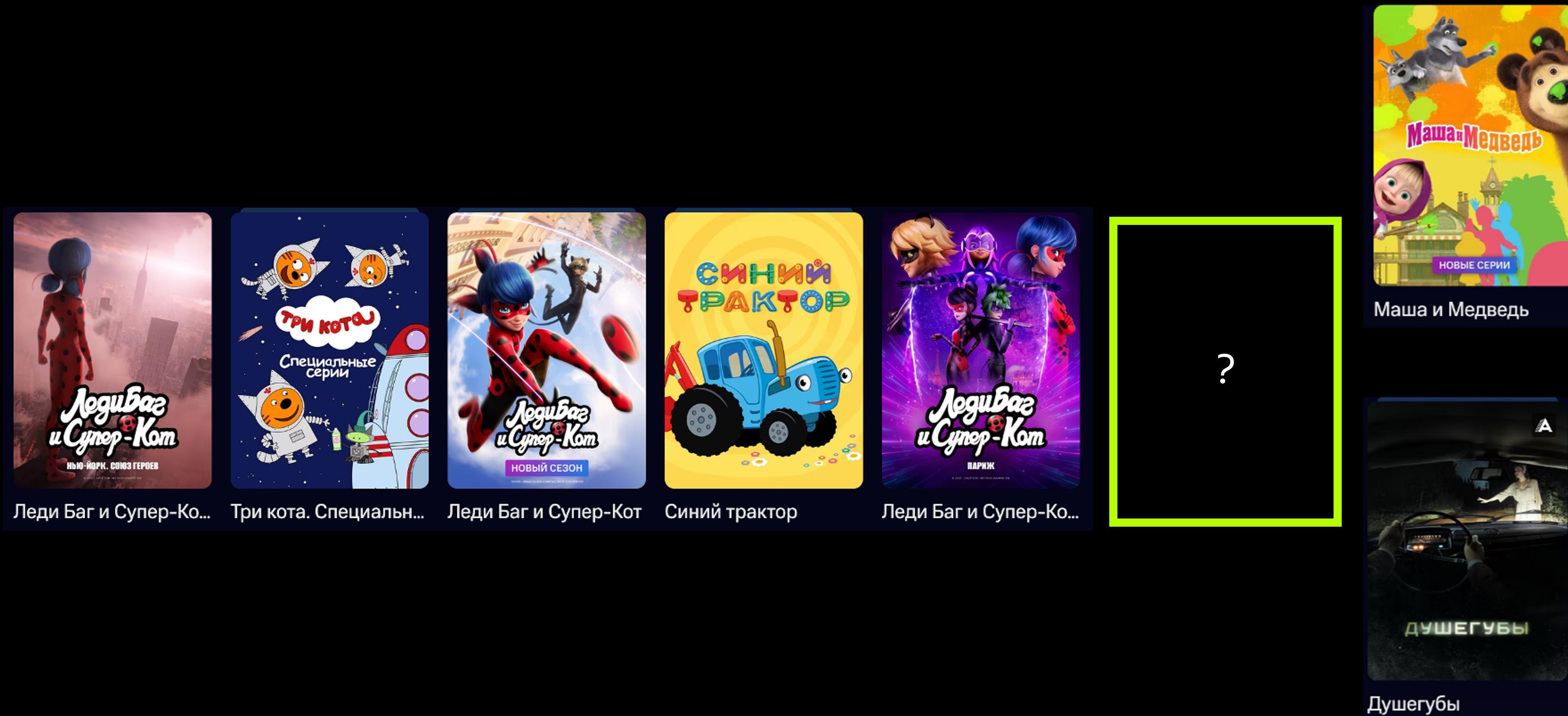


- Оценка
- Жанр
- Актёры
- Описание

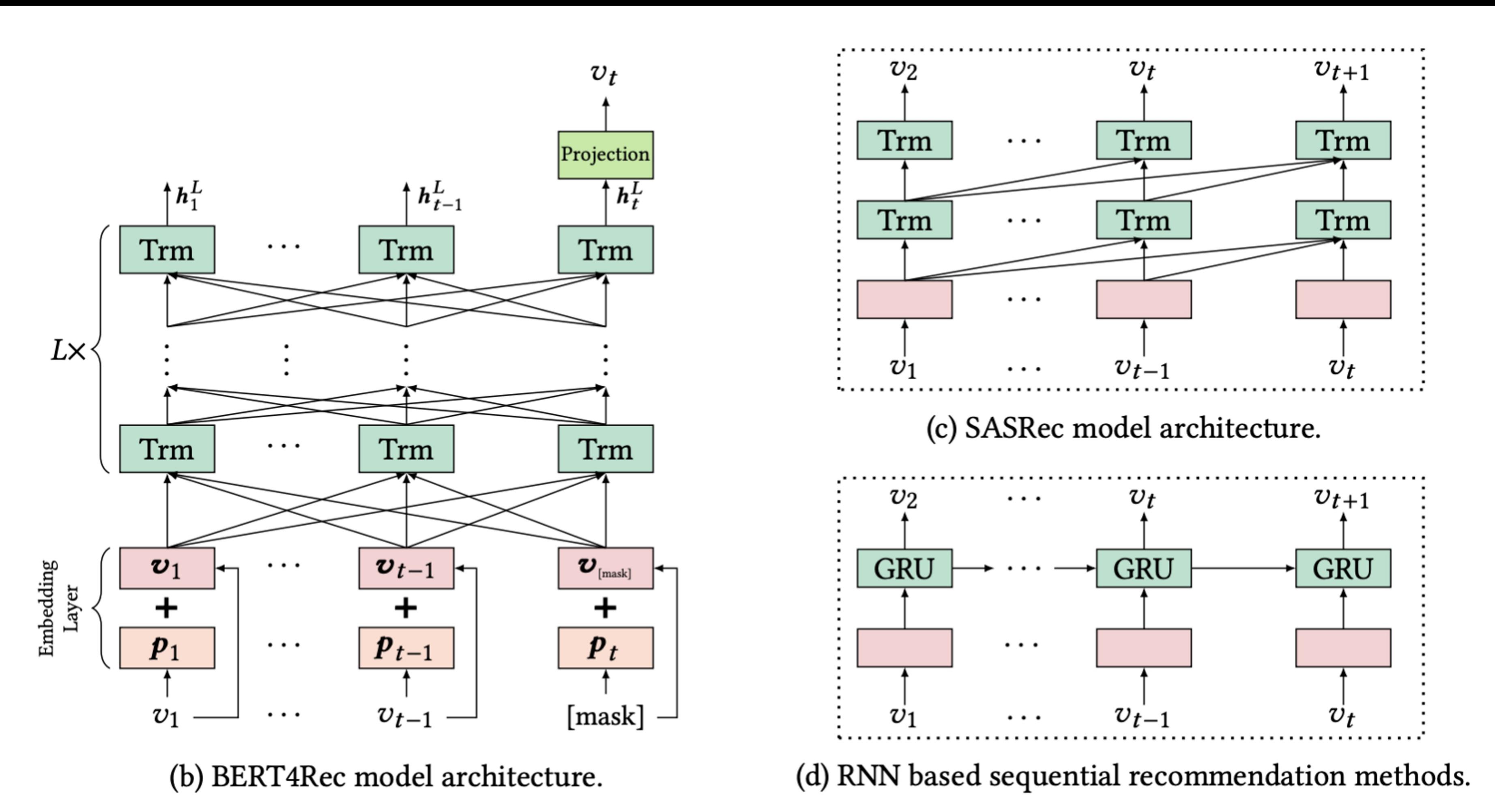


- Соц-дем
- Поведение в сервисе

Последовательность действий



Последовательность действий



Смотрим!



Настоящая формула

- mean_user_score
- Попадание в интересы пользователя
- Попадание в ЦА фильма
- Увеличение кликов со временем

При стакинге регулярно
переобучаемых моделей
важно подумать
о стабильности
распределения скоров

Проверка логичности поведения модели — важный шаг

Обучение на логах —
не гарантия
отсутствия лицов

Перед обучением
модели важно подумать
про Inductive Bias

Яндекс



Спасибо!
Вопросы?



Сергей Кузнецов

Tech Lead RecSys&Search
Platform, МТС Web Services