# Recommendation system for investor-fund interactions

**Нижний Новгород, 22 ноября 2019**

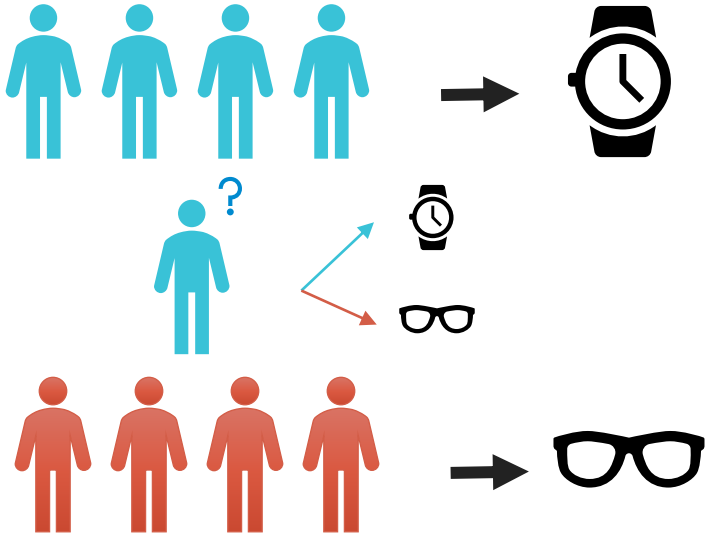
## ГРЕЧИХИН ИВАН

- EPAM Systems – 2018-н.в. (аналитик)
- Samsung-PDMI AI Center – 2018-н.в. (исследователь)
- НИУ ВШЭ – 2015-н.в. (преподаватель)
  - Курсы: Анализ данных, Теория вероятности
- ЛАТАС НИУ ВШЭ – 2012-н.в. (стажёр-исследователь)

# Содержание:

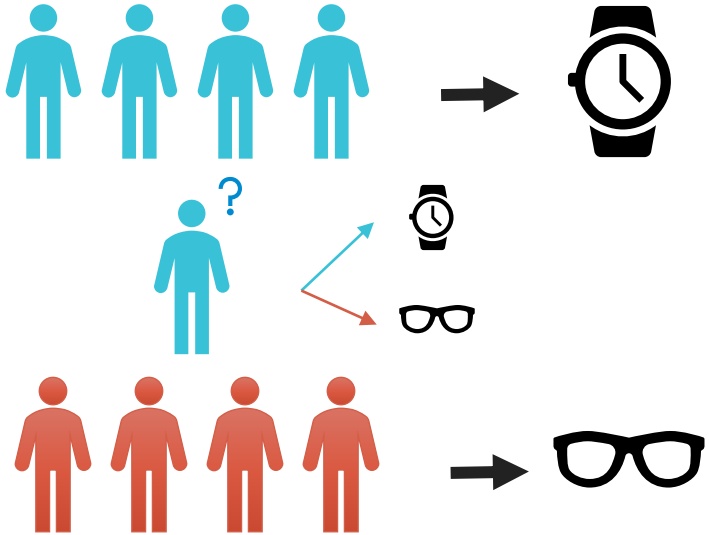
- Рекомендательные системы
- Проектная задача
- Простейшие подходы user-to-user/item-to-item
- Нейронный «велосипед»
- Матричная факторизация
- ...и что было между

# Рекомендательные системы



- Задача:
  - Рекомендовать товары/объекты похожие на те, что раньше понравились пользователю
  - Рекомендовать товары/объекты похожие на те, что могут быть необходимы/интересны пользователю
- Рекомендательная Система – задача классификации ***взаимодействий*** между двумя сущностями

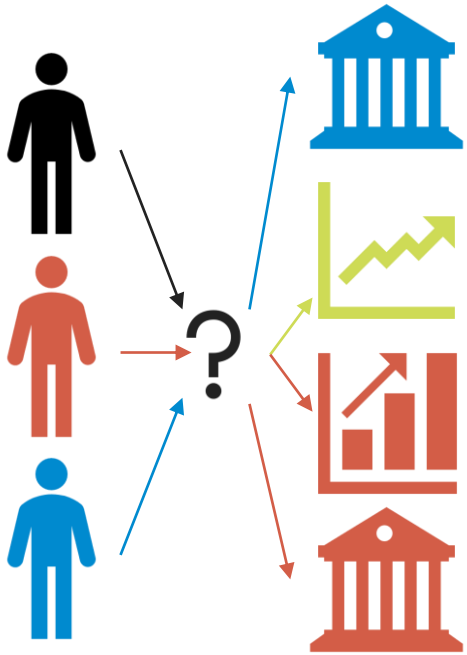
# Рекомендательные системы



- Примеры:
  - Рекомендации фильмов (Kinopoisk, IMDb)
  - Рекомендации в пользовательскую корзину («с этим товаром также берут..»)
  - Реклама (Google, Yandex)
- Бизнес-цель:
  - Повысить конверсию/объём продаж
  - Предложить интересный сервис

# Задача

- Создать рекомендательную систему инвестор-фонд, для крупной инвестиционной фирмы
    - **Инвесторы – «пользователи», фонды – «объекты».**
- Цель: для фонда получить отсортированный список инвесторов, от самого рекомендуемого до наименее рекомендуемого
    - **Получить цифровую оценку уверенности в рекомендации**
- Интересный момент: процесс получения задач и знаний от заказчика



# Данные

- Взаимодействия инвесторов и фондов:
  - Тип/регион инвестора
  - Результат взаимодействия
    - Оценка-статус от 1 до 11 (от «потерян» до «успех»)
    - Дата последнего изменения статуса
    - Сумма вложений (в случае 10-11 статуса)
  - Дополнительные параметры инвестора
- Список активных фондов
  - Необходим для выполнения целей, поставленных заказчиком
- Дополнительные данные
  - Предоставляются заказчиком из других источников (не собственная база данных)

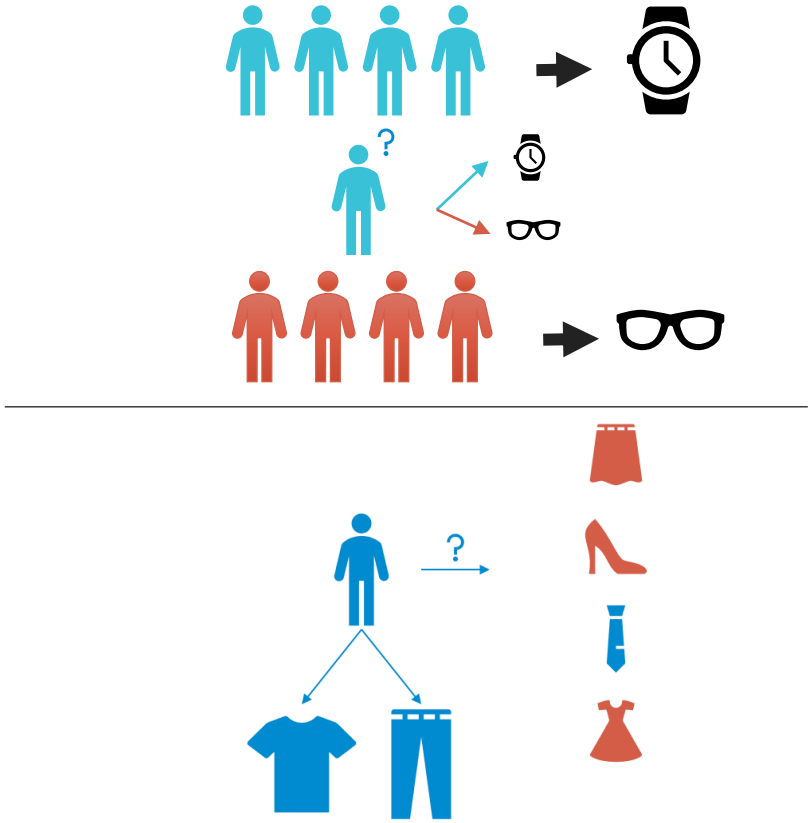
# Данные

| inv_id | fund_id | inv_type | inv_region | money | strategy | product | date | status |
|---|---|---|---|---|---|---|---|---|
| 24 | 601 | Pension Fund | Europe | 2500000 | 252 | 1346 | 22.11.10 | 10 |
| 26 | 603 | Hedge Fund | Americas | 0 | 252 | 1427 | 23.11.10 | 4 |

- У каждого фонда только одна стратегия и продукт
- Product ⊂ Strategy
- У каждого инвестора один регион и тип
- Деньги указаны только для статусов 10 и 11
- ~4000 инвестора, ~500 фондов, ~22000 взаимодействий

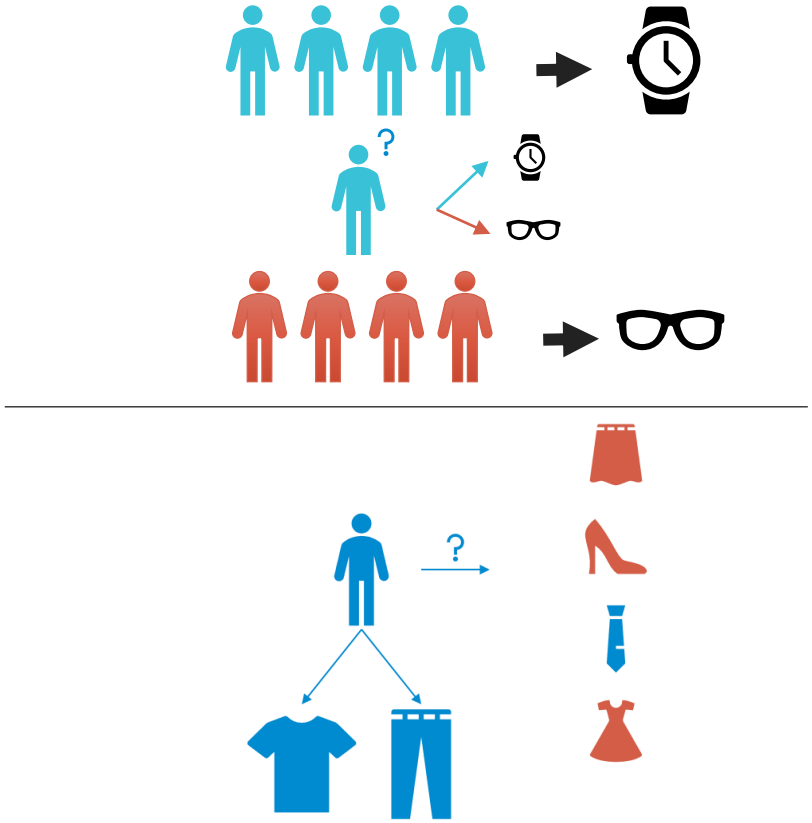
# Первый подход

- User-to-user
  - Рекомендация объектов, интересных схожим пользователям
- Item-to-item
  - Рекомендация пользователей согласно схожести интересных им объектов
- Схожесть оценивается с помощью евклидового / косинусного расстояния



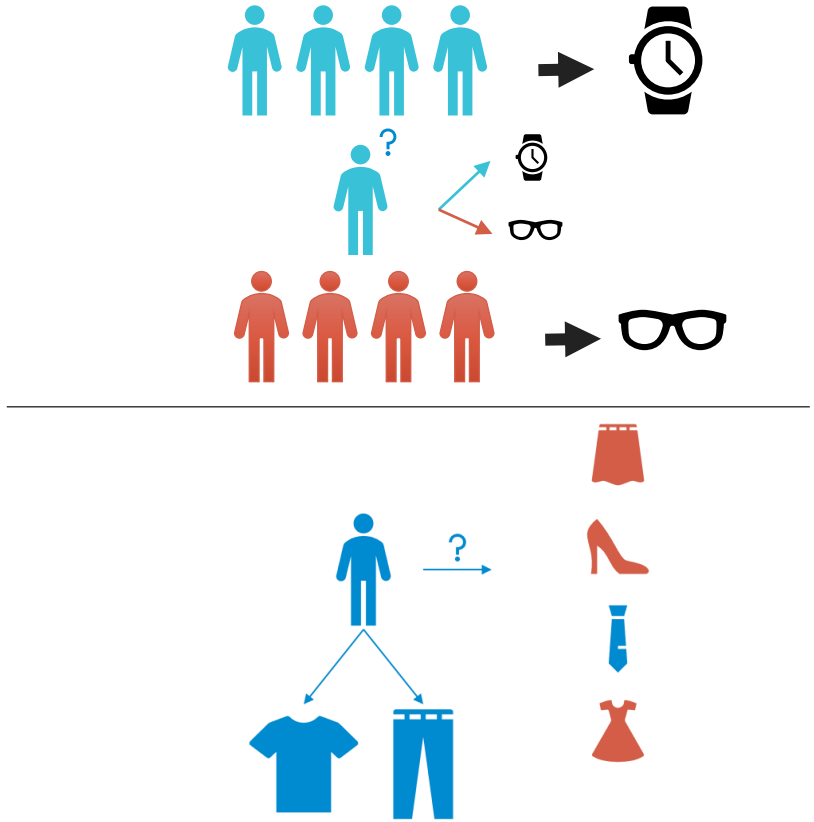
# Первый подход

- Процесс обучения:
  - 10% данных случайным образом помещаются в тестовую выборку
  - В тестовую выборку попадают только фонды/инвесторы, у которых больше 1 взаимодействия
  - Схожесть инвесторов/фондов исследуется по векторам взаимодействий
- Тест:
  - Сравнение полученных оценок с реальным результатом взаимодействия



# Первый подход

- Результат:
  - В среднем предсказанный статус/стадия отличаются на 1.4-1.5 балла (MAE).
  - Однако, всего 28% точно угаданных взаимодействий и больше 40% с разницей в 2 балла.
- Особенности:
  - Не для каждой пары инвестор-фонд возможно получить предсказание
  - Проблема малого количества данных для некоторых инвесторов/фондов
  - Не используем дополнительные данные – требование заказчика



# Итоги первого подхода

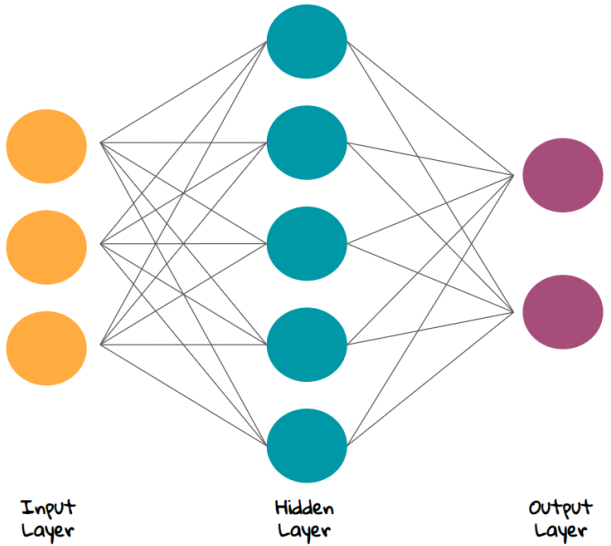
- Получен baseline – результат, на который можно ориентироваться
- Новые требования:
  - Необходимо получить рекомендации ***для нового фонда***, отсортированные по некоторой метрике (по ожидаемому/предсказанному статусу)
    - Таким образом, фонд становится по сути пользователем (однако мы не будем менять принятые обозначения)
    - При этом, нового фонда нет в исторических данных – известна только стратегия
  - Хотелось бы использовать дополнительные внешние данные
  - Хочется попробовать нейронные сети

# Дополнительные данные

- Другие источники информации по взаимодействиям инвесторов и фондов
- Заказчик приводил другие источники к своим стандартам
    - Соответствие id фондов и инвесторов в своей системе
    - Соответствие статусов/наполнения своей системе
- Сильное отличие по наполнению – содержащейся информации – и как её возможно использовать
    - Отсутствуют стратегии вложений
    - Сложный маппинг стратегий

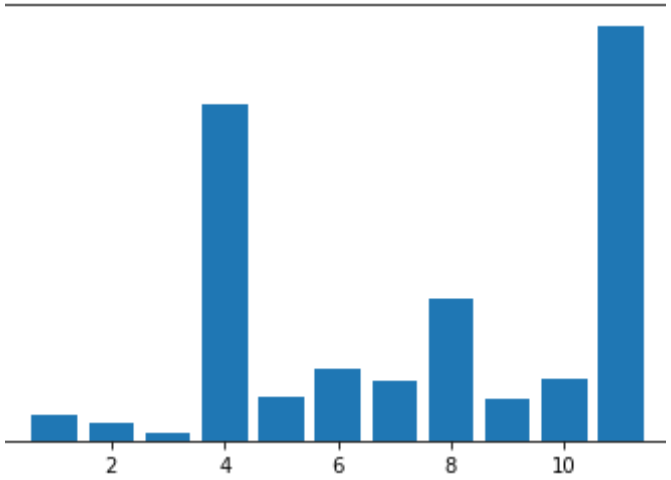
# Нейронная сеть

- Простая нейронная сеть для предсказания результата взаимодействия
- Плюсы:
  - Возможность использования внешних данных
    - Используем внешние данные, предоставленные заказчиком
    - Создаём новые признаки – например распределение взаимодействий инвестора по стратегиям
  - Возможность получения предсказания для любой пары инвестор-фонд



# Изменение концепции данных

- Переобучение: классы 4 и 11 значительно превосходят остальные числом
- Нелинейность: статус 2 не в 2 раза лучше статуса 1.
- Не работает: внешние данные не дали никакого прироста к качеству рекомендаций
- Переход к 0-1 implicit матрице: использование статуса сделки как успех/неуспех – теперь мы предсказываем число между 0 и 1, как некий confidence level в успешности рекомендации




$$\begin{bmatrix} 0 & 0 & 1 & & 0 \\ 0 & 1 & 0 & \cdots & 1 \\ 0 & 0 & 1 & & 1 \\ & \vdots & & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

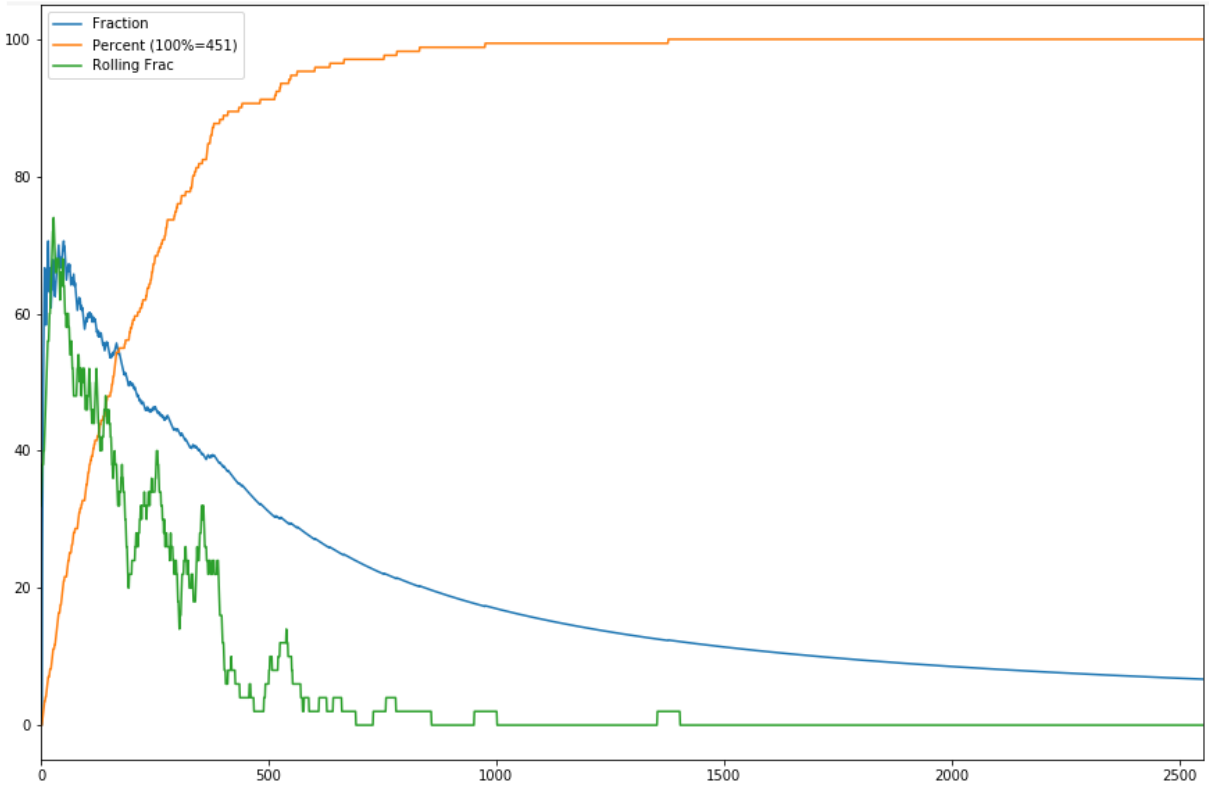
# Итоги нейронных сеток

- MAE для статусов был 1.2 – но переобучение
- Схожие проблемы по недостаточному проценту угаданных статусов-классов

- Была создана новая комплексная метрика оценки качества рекомендации...
  - ...которая показала, что на implicit данных всё равно переобучается

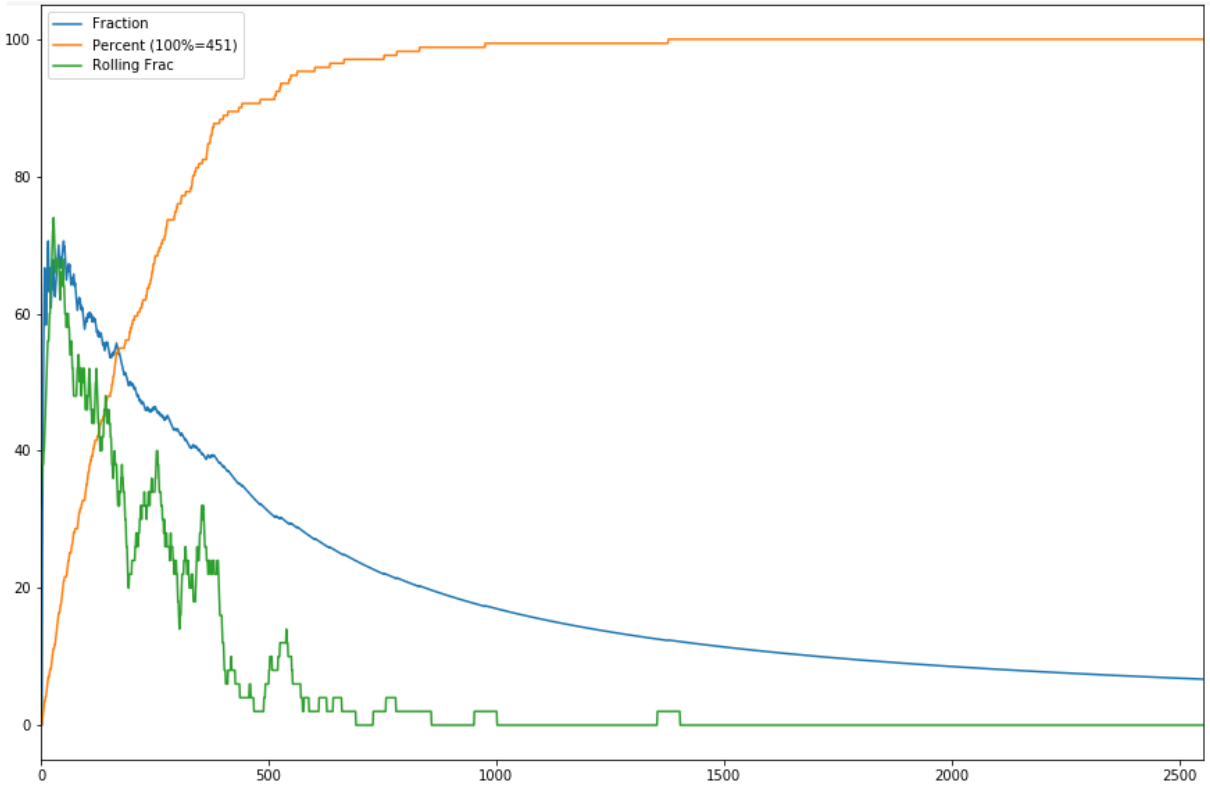
- Не сработало в итоге ничего
  - Использование 3х источников данных как параллельные треки в нейронке
  - Смешивание всех признаков
  - Уменьшение параметров до минимального (сужение сетки)
- Проблемы:
  - Слишком мало данных
  - Сетка просто выучивает значения, параметров больше чем данных

# Метрика



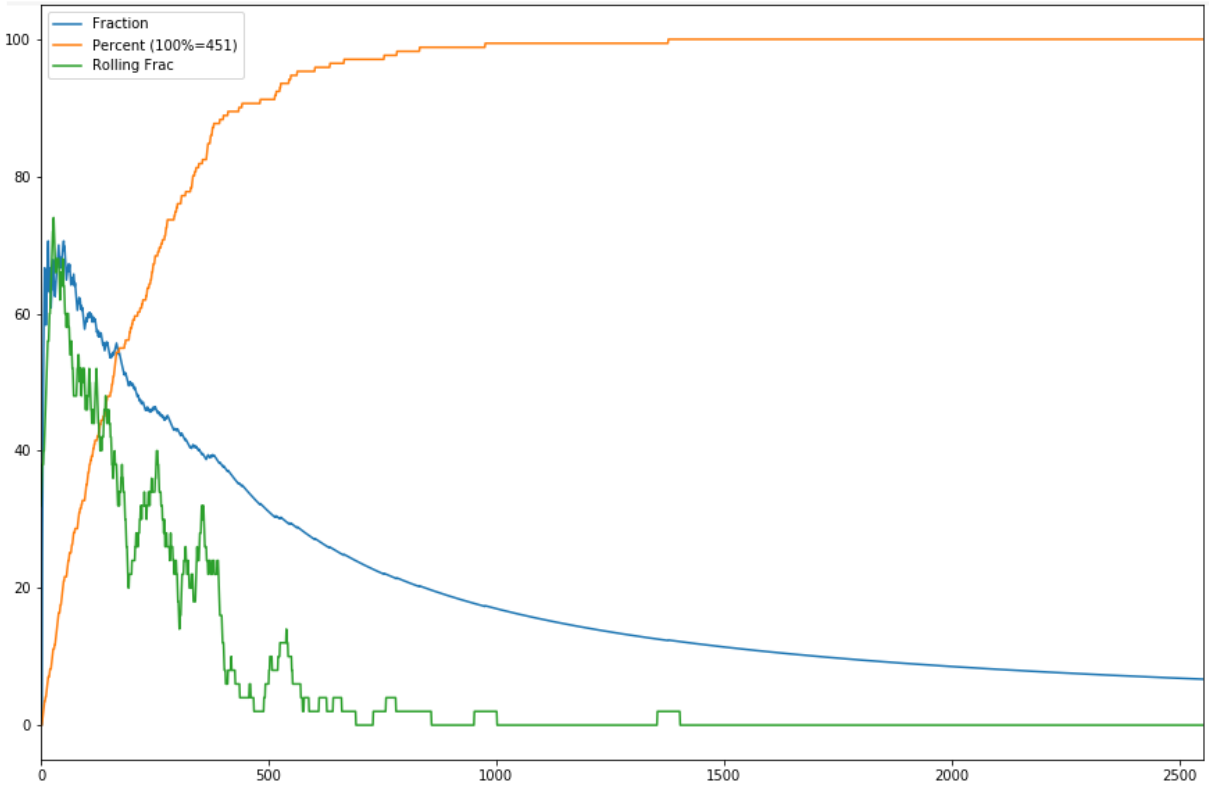

- Оценка качества происходит на тестируемом примере – фонде с большой историей
- Предполагается, что тестируемый фонд – новый (все его взаимодействия убираются из обучающей выборки)

# Метрика



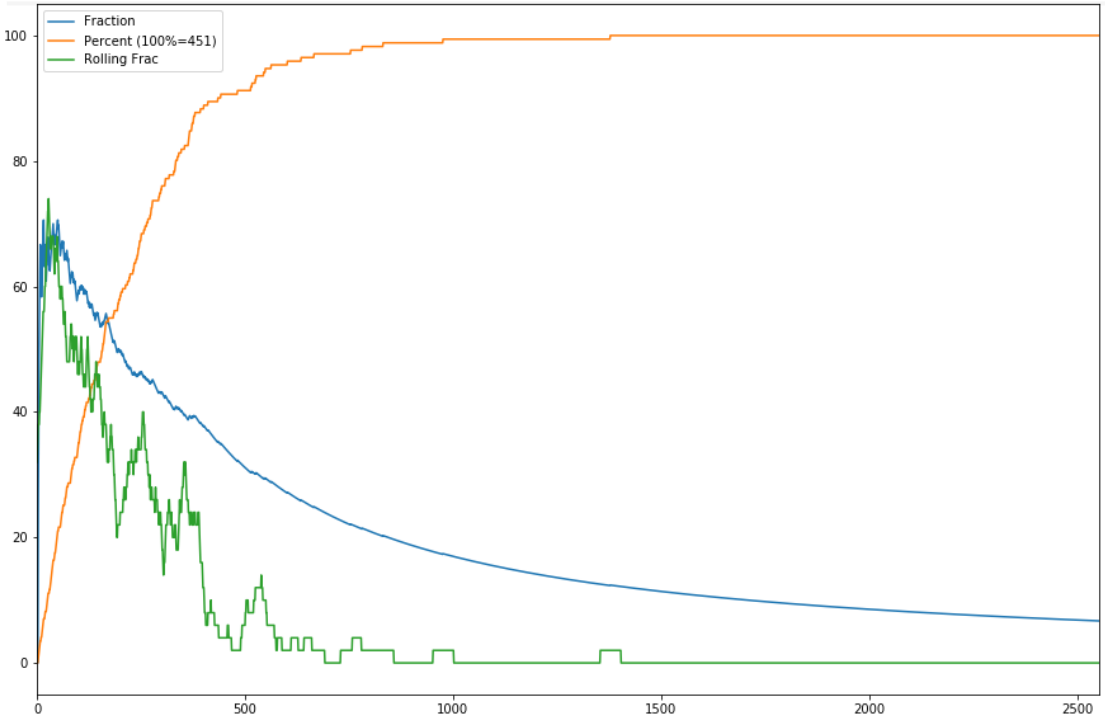

- Модель обучается на всех оставшихся данных
- Качество модели оценивается по количеству реальных взаимодействий попавших в топ рекоммендаций
- Учитываются все взаимодействия, не только успешные (то есть хотя бы минимальный интерес)

# Метрика




- Ось X – инвесторы. Предполагается, что они отсортированы по уровню уверенности в успешности рекомендации
- Ось Y – процент успешности

# Метрика




- Interactions to **x** – сколько инвесторов среди первых **x** реально взаимодействовали с текущим фондом
- Interactions **i to i+50** – сколько инвесторов на позициях от **i** до **i+50** реально взаимодействовали с текущим фондом

- $Blue(x) = \frac{interactions\ to\ x}{x} * 100$
- $Orange(x) = \frac{interactions\ to\ x}{total\_interactions} * 100$
- $Green(x, i) = \frac{interactions\ i\ to\ i{+}50}{50} * 100$

# Что дальше?

- Своя нейронная сеть не сработала
- Другие опробовать – долго и не факт, что сработает

- Совет коллеги – факторизационные машины
  - Легче попробовать
  - Сработало сразу неплохо и без переобучения (и даже лучше чем было)

Здесь должна быть картинка как в рекламе, как будто двое коллег результативно общаются

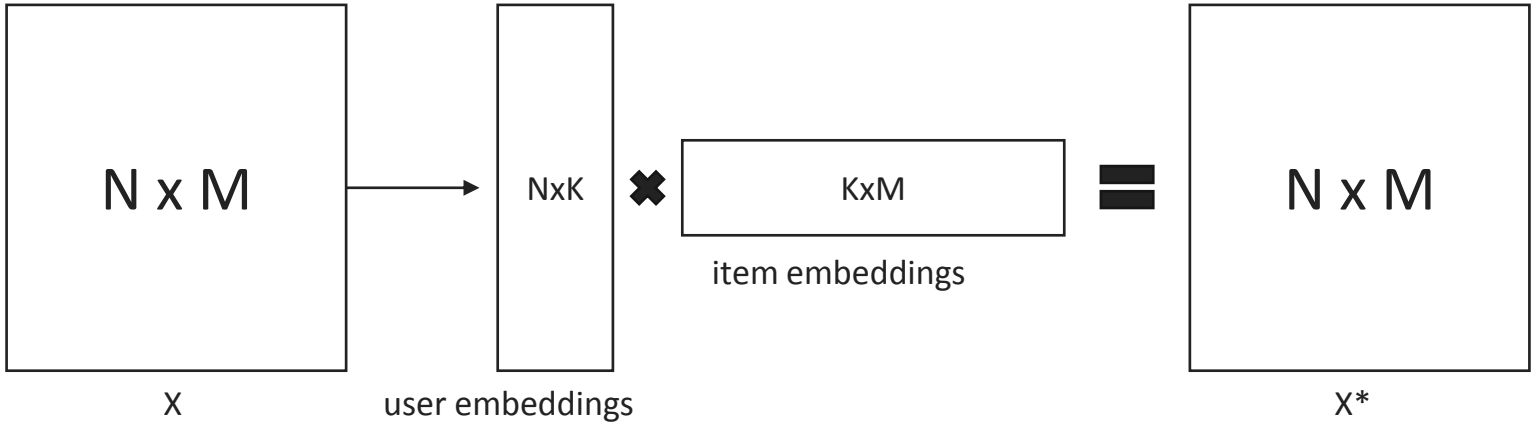
# Новые скрытые особенности в данных

| inv_id | fund_id | inv_type | inv_region | money | strategy | product | date | status |
|---|---|---|---|---|---|---|---|---|
| 24 | 601 | Pension Fund | Europe | 2500000 | 252 | 1346 | 22.11.10 | 10 |
| 26 | 603 | Hedge Fund | Americas | 0 | 252 | 1427 | 23.11.10 | 4 |

- Vehicle-funds – фонды-«заглушки» для фондов с филиалами
  - Можно выцепить по product столбцу
  - Данные филиалов дублируются в основном фонде (!)

# Матричная факторизация

- Разложение матрицы взаимодействий на вектора-эмбеддинги, которые характеризуют пользователей и объекты
- Вектора-эмбеддинги могут быть использованы в предыдущих подходах



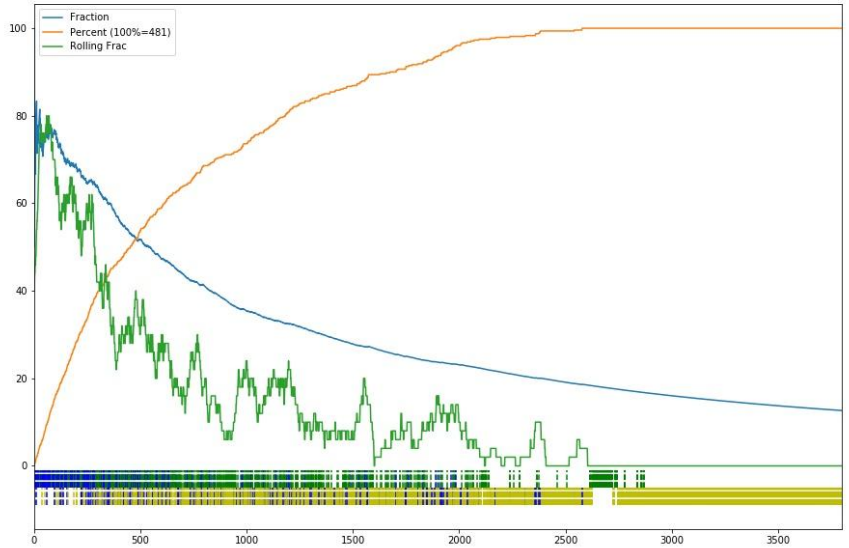

$$|X - X^*| \rightarrow min$$

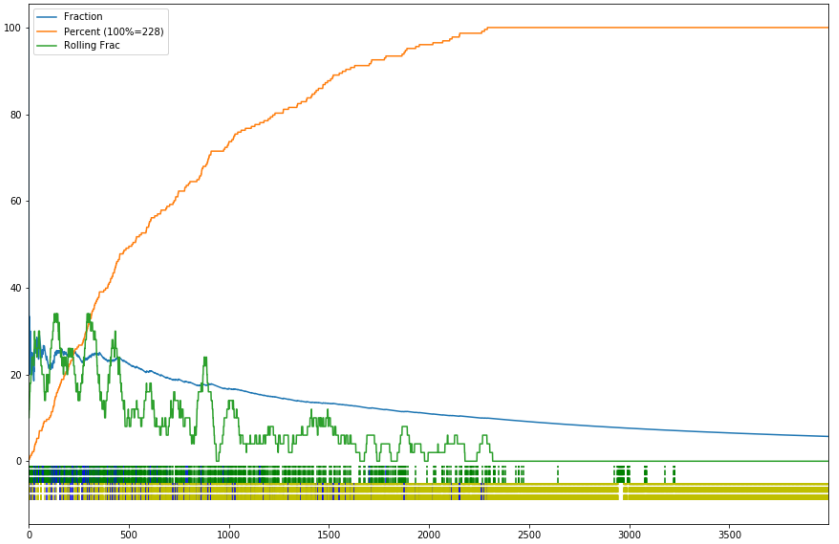
# Используемая модель

- LightFM
  - https://arxiv.org/abs/1507.08439 - **Metadata Embeddings for User and Item Cold-start Recommendations**
- Используем только данные от заказчика, без дополнительных внешних (они не улучшают результат) вместе с признаком: распределение стратегий по инвесторам
- Используемая метрика – график + площадь под оранжевой кривой (по сути ROC AUC)

- **Цель**: для фонда получить отсортированный список инвесторов, от самого рекомендуемого до наименее рекомендуемого
- Разложение матрицы взаимодействий и дополнительных признаков с помощью факторизации
- Получение предсказаний для всех фондов с той же стратегией
- Усреднение полученных предсказаний и сортировка полученных значений от большего к меньшему

# Примеры результатов




- В зависимости от стратегии, предсказания могут быть не очень хороши (это зависит от объёма выборки по стратегии и истории фонда)
- *полоски внизу – информационные для заказчика, их смысл и необходимость спорная

# Итог

- Система создана, сейчас создаётся API для того, чтобы модель крутилась на сервере
- В ходе проекта не раз менялась концепция и используемый метод...
  - ...по воле заказчика
  - ...по причинам неудачности модели
- Итоговый ROC AUC в среднем для модели – 0.8

- 100% есть методы, которые я не попробовал, как и варианты обработки данных, но не всегда была возможность выбирать между
  - Дать готовый результат или
  - Попробовать что-то новое (на месяц-два исследования)