Yandex Money

Дата-инженеры и машинное обучение

Евгений Виноградов

SECR
Software Engineering Conference Russia
14-15 ноября, 2019. Санкт-Петербург

Кто занимается DS-проектами?

- Data Scientist
- Data Ingest
- Data Steward
- Data Dev Ops
- Information Security

Кто занимается DS-проектами?

- Data Scientist
- Data Ingest
- Data Steward
- Data Dev Ops
- Information Security

Что такое Data Science-проект?

Что такое Data Science-проект?

Что такое Data Science-проект?

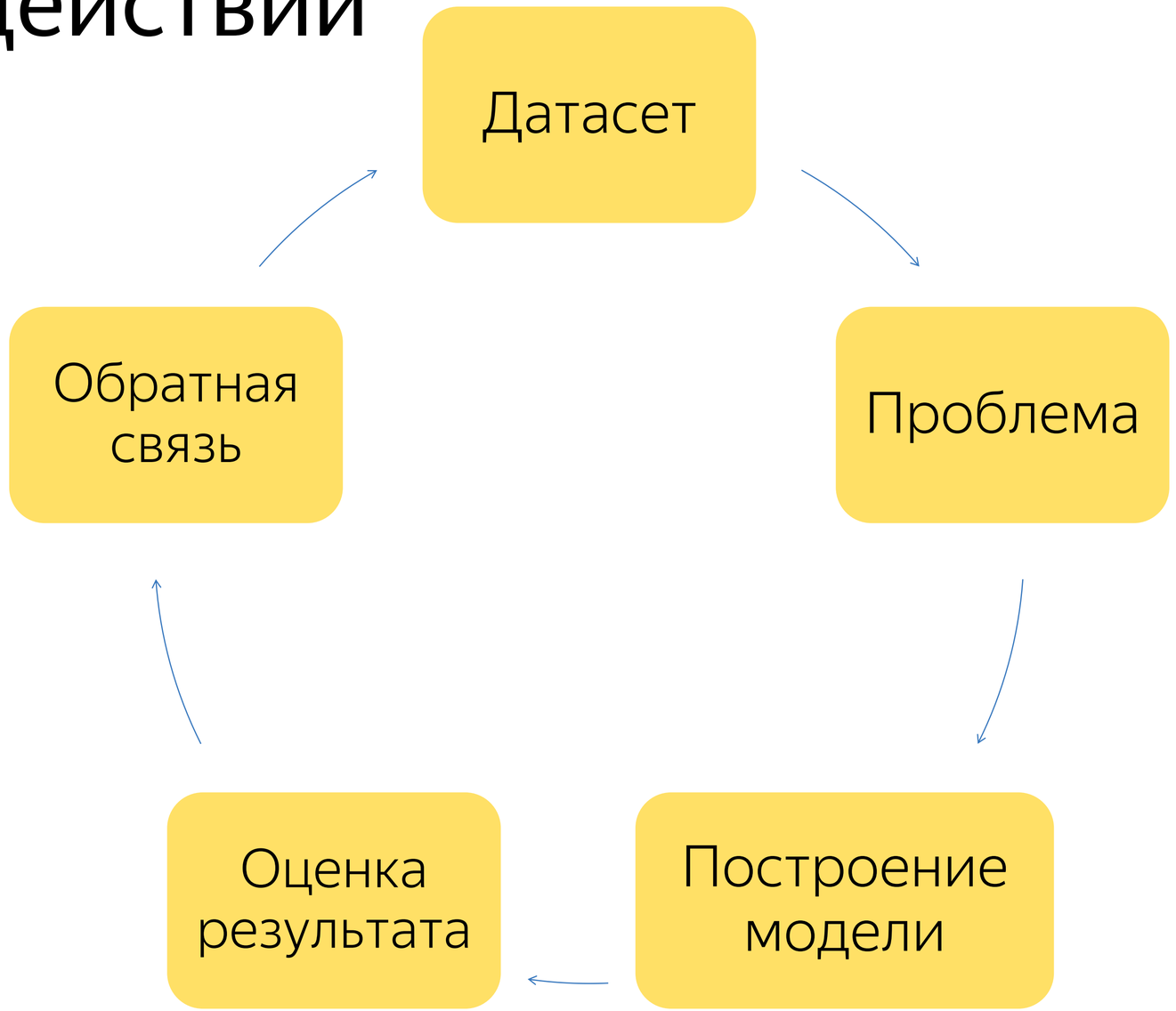
- Мы хотим автоматически определять что-нибудь (аварии и простои)
- И никто до нас не сделал устраивающего нас решения (или мы о нем не знаем)

Что такое Data Science-проект?

Порядок действий

- Датасет
- Проблема
- Построение модели
- Оценка результата
- Обратная связь

Порядок действий



Датасет

Кто-то что-то об этом знает

- Авторизация
- Сбор реквизитов
- Авторизация карты
- Зачисление денег в систему
- Перевод на счет получателя
- Уведомление получателя
- Клиринг и т.д.

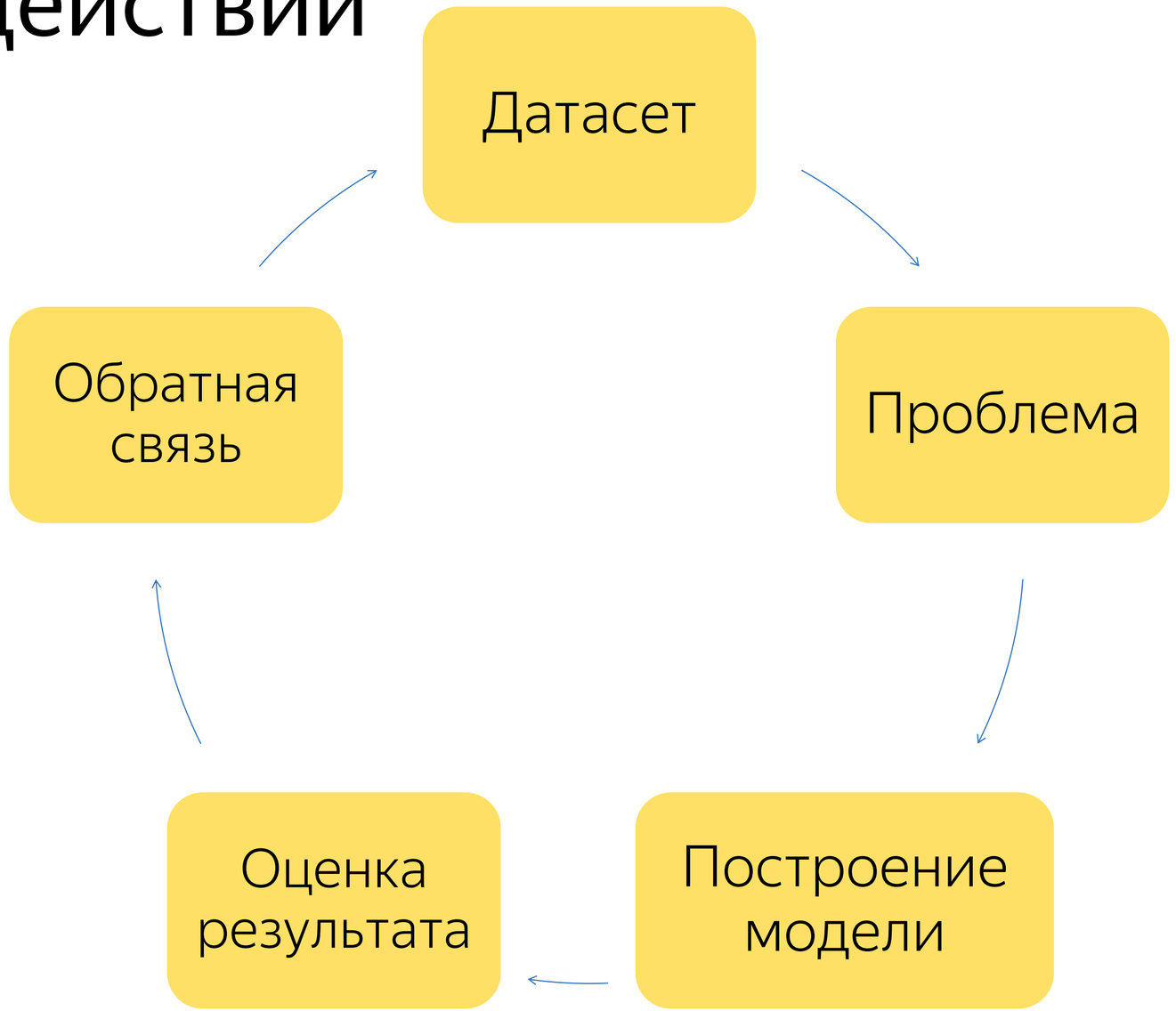
Кто-то что-то об этом знает

- Авторизация
- Сбор реквизитов
- Авторизация карты
- Зачисление денег в систему
- Перевод на счет получателя
- Уведомление получателя
- Клиринг и т.д.

Составление датасета

- Интеграция
- Разметка
- Очистка
- Расчет фич
- Актуализация

Порядок действий



Датасет – 2

- Новый расчет фич
- (или Realtime-расчет фич)

Feature Engineering

Создание признаков

- Каждый час берем с временным окном, увеличиваем окно пока не наберем достаточное число платежей
- Выбираем как далеко мы готовы смотреть в прошлое, и минимальное количество дней для подсчета статистики
- Накопив данных за минимально достаточное число дней, начинаем расчеты в рамках срезов:
  - аномальность значения: если значение полученное на текущий час сильно отклоняется от медианы предыдущих дней, то считаем значение аномальным
  - по неаномальным значениям считаем стандартное отклонение
  - рассчитываем критическое значение задержки: текущее значение + стандартное отклонение умноженное на выбранное значение лямбда
- В реальном времени сравниваем время с последнего платежа и критическую задержку

Модель

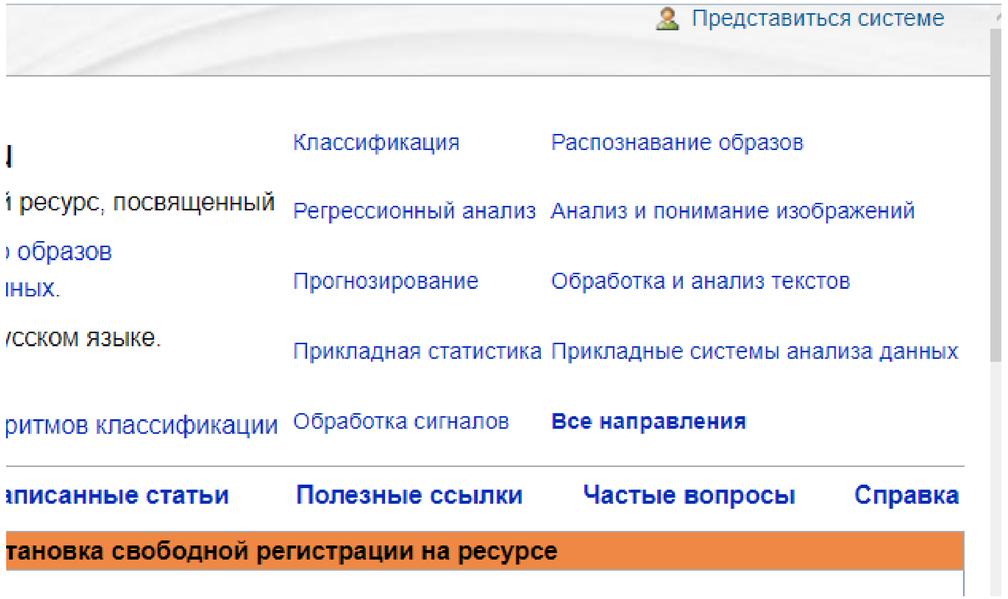
Как выбрать модель?

Применить экспертные знания

Как выбрать модель?

Применить экспертные знания

Посмотреть http://www.machinelearning.ru/



Как выбрать модель?

Применить экспертные знания

Посмотреть http://www.machinelearning.ru/

Почитать, кто чего недавно в этой области делал

- А если все спрыгнут с крыши, ты тоже спрыгнешь?

- Ну, вам же никто не мешает говорить фразу, которую все говорят!

- А если все спрыгнут с крыши, ты тоже спрыгнешь?

- Ну, вам же никто не мешает говорить фразу, которую все говорят!

- А если все спрыгнут с крыши, ты тоже спрыгнешь?

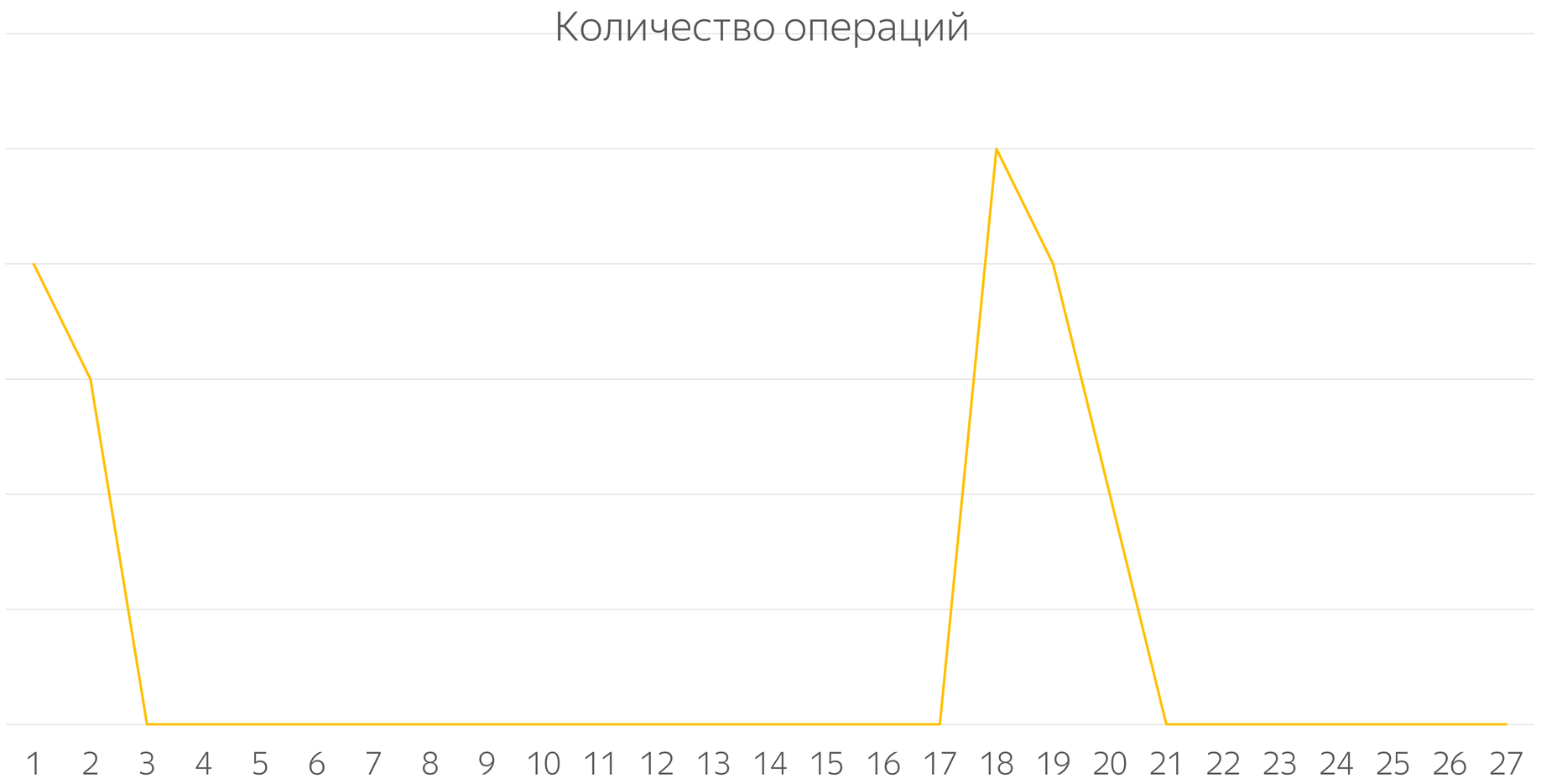
[изобретено машинное обучение]

- А если все спрыгнут с крыши, ты тоже спрыгнешь?

[изобретено машинное обучение]

- Да!

И тут меняется профиль



ML Engineer

- Переобучение моделей as a service
- CICD для моделей

Технологии

На каких технологиях все это работает?

Доставка событий:

Kafka, BatchAPI, RabbitMQ, …

Управление ETL:

SSIS, SQL Server Agent, python, cron, Jenkins-jobs,…

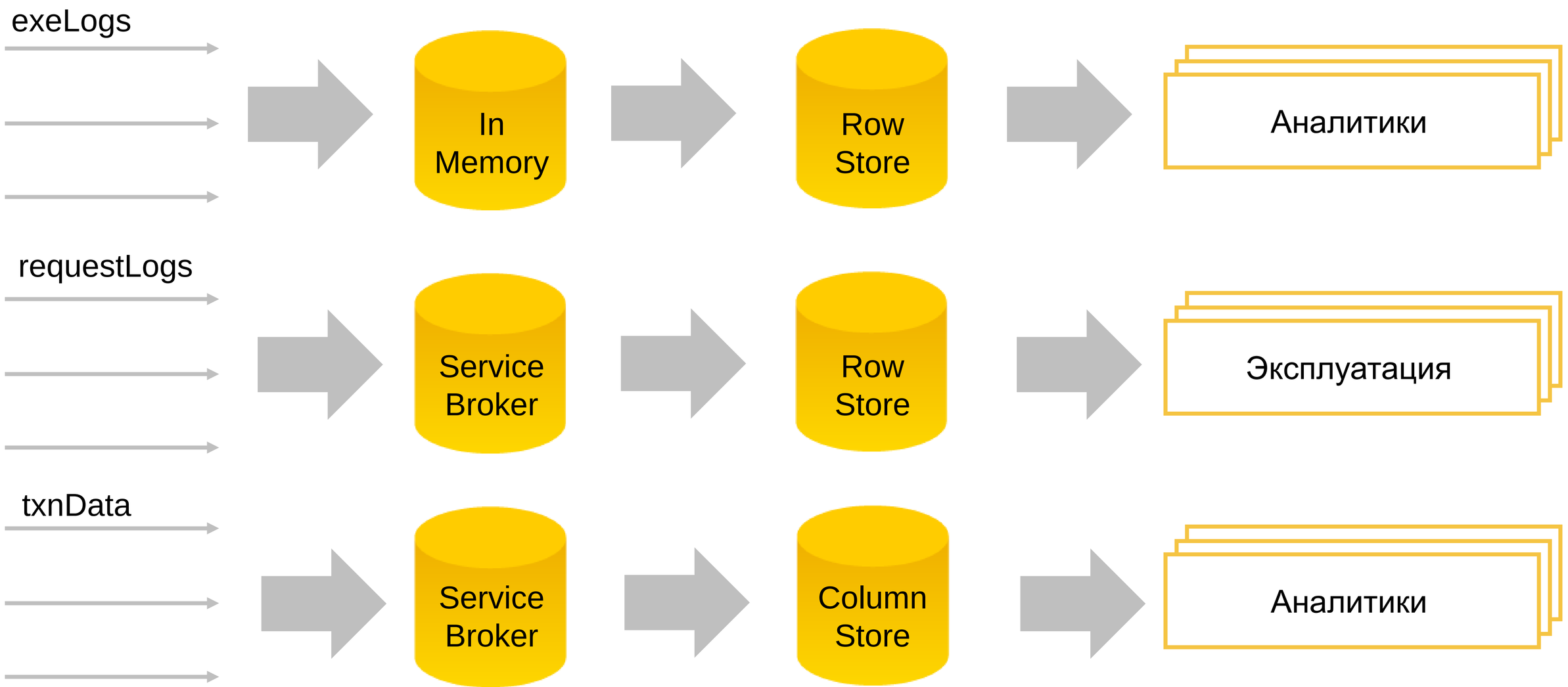
Репортинг:

PowerBI/SSRS, кубы, Tableau, d3js, …

ХД:

MSSQL (columnstore/rowstore), clickhouse, graphite, hdfs, …

RT Storage


exeLogs
In Memory
Row Store
Аналитики
requestLogs
Service Broker
Row Store
Эксплуатация
txnData
Service Broker
Column Store
Аналитики

Академический подход

- Математика
- Kaggle
- Coursera
- Тестовый проект
- Сбор данных
- Реальный проект

Бери и делай

- Найти задачу
- Coursera/Machinelearning.ru
- Сбор данных
- Coursera/Machinelearning.ru
- Решение
- Снова сбор данных

Итого

- В реальных проектах ML-проектах Data Engineer играет одну из наиболее важных ролей
- Бэкграунд разработчика сильно помогает
- Возможности по решению проблем часто шире, чем у дата-саентиста

Спасибо!

Евгений Виноградов

Руководитель отдела разработки
хранилищ данных

jonny@yamoney.ru