СБЕР

# Что такое MLOps и как это работает на примере Сбера

Дмитрий Бугайченко

# Немного о себе

**Дмитрий Бугайченко**
**CDS B2C** Сбер

**12** лет в DM

**22** года в IT

**СПбГУ**
Кандидат физико-математических наук

2008

**Mail.ru**

- Рекомендательные системы музыки, сообществ, новостей, товаров, …
- Облачная МЛ платформа
- Платформа А/Б тестирования

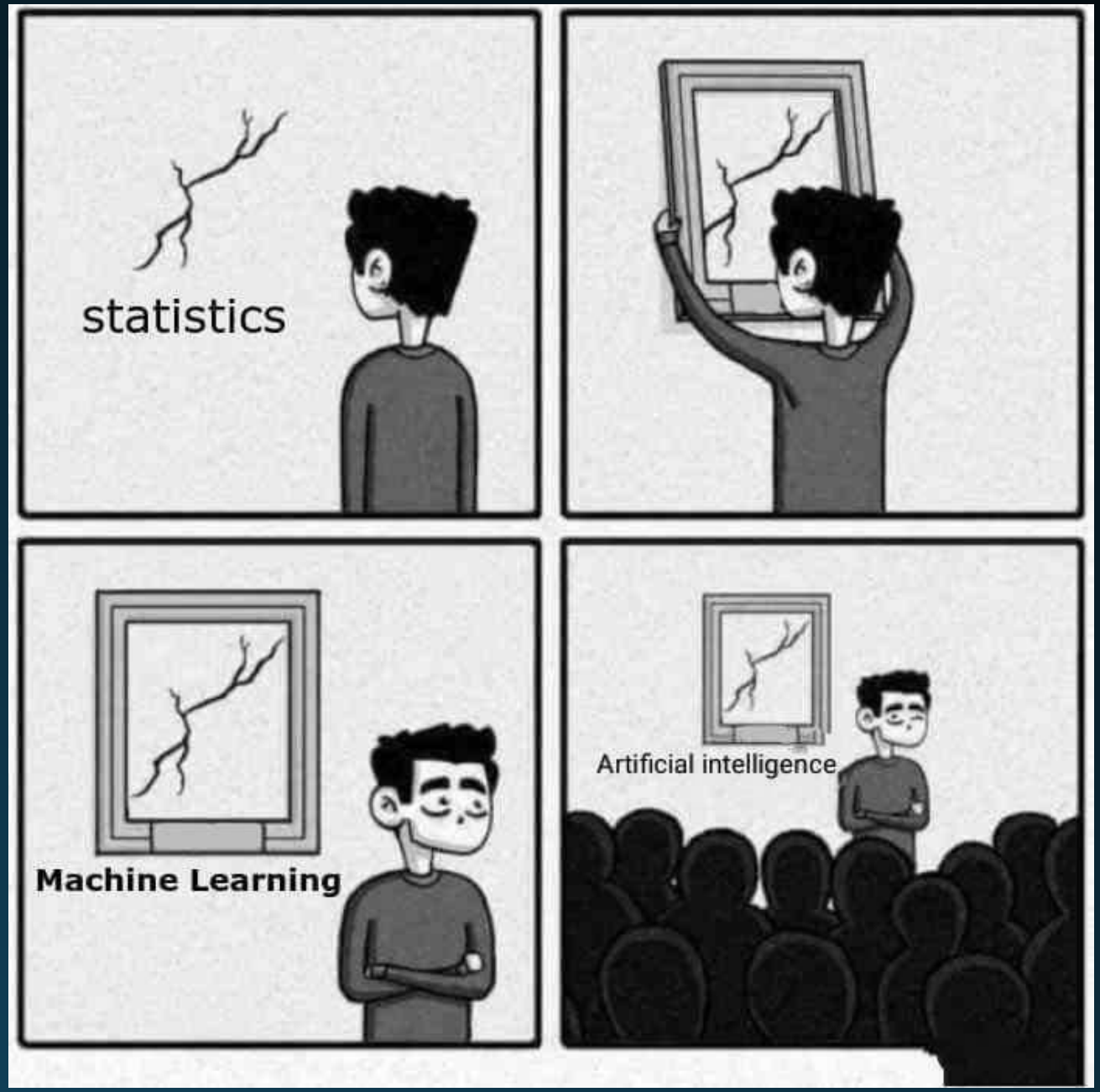
2011

**Сбер**

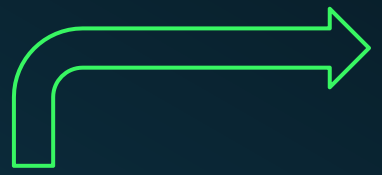
- Платформа рекомендательных систем – генерация рекомендаций в режиме онлайн
- Инструменты разработки МЛ моделей в распределенной среде
- Платформа проактивных клиенто-центричных коммуникаций

2019

# Что такое «искусственный интеллект»?

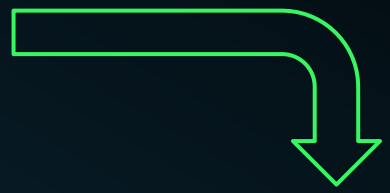

statistics
Machine Learning
Artificial intelligence

# Разберемся в основных понятиях



## Machine learning

Класс методов поиска зависимостей, основанный на подборе параметров математической модели

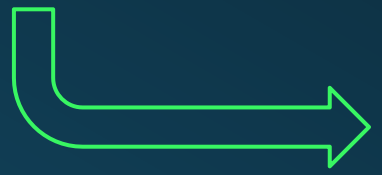


## Data Mining

Поиск неочевидных зависимостей в данных для реализации практических целей

## Artificial Intelligence

Принятие решений на базе найденных зависимостей



## Big Data

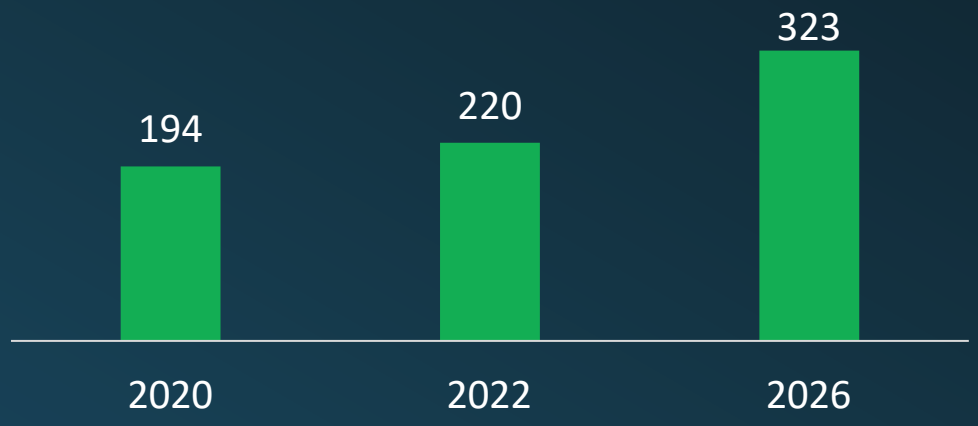
Набор технологий и подходов для обработки данных, не вмещающихся в память одной машины

# AI – искусство, наука или ремесло?

# В современном мире AI решения стали неотъемлемой частью бизнеса, что требует ремесленной тиражируемости

## Мир

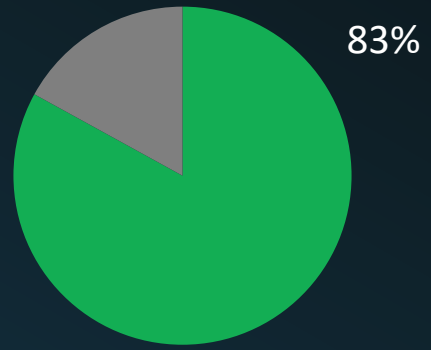
Объем мирового рынка больших данных, млрд долл.




- Глобальный рынок данных растет вместе с объемом информации. В 2018 году пользователи и предприятия создали 33 ЗБ информации, к 2025 году это показатель составит 175 ЗБ.

## Бизнес

83% компаний заявляют, что AI является главным приоритетом в их бизнес-планах.

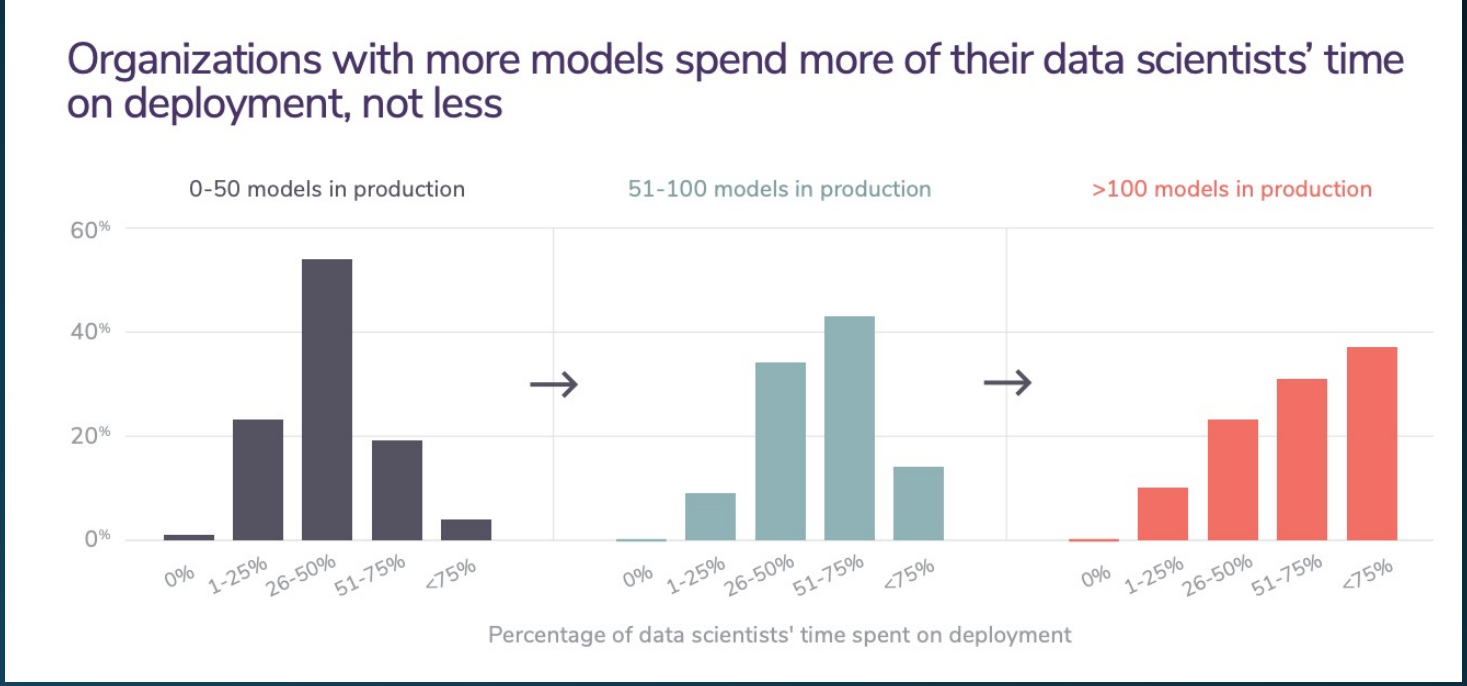



- Netflix ежегодно зарабатывает 1 миллиард долларов на автоматизированных персонализированных рекомендациях.
- К 2025 году в сфере искусственного интеллекта будут работать 97 миллионов человек

https://explodingtopics.com/blog/ai-statistics#ai-stats
https://www.globenewswire.com
https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

# Но при внедрении AI решений компании, особенно крупные, сталкиваются с существенными сложностями

Чем больше у компании моделей,
тем дольше они идут в прод ☹




Только 50% пилотов доходят до прода

В среднем на это уходит 9 месяцев

https://mikenashtech.com/2021/12/enterprise-trends-in-machine-learning-report-analysis/

https://www.gartner.com/en/newsroom/press-releases/2021-09-07-gartner-identifies-four-trends-driving-near-term-artificial-intelligence-innovation

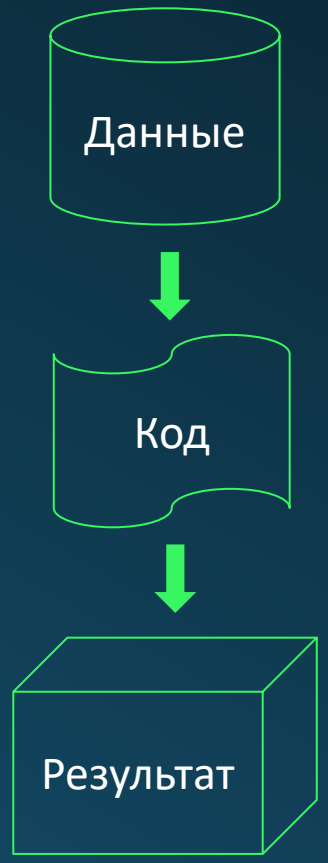

MLOps (Machine Learning Operations) - это практика, объединяющая различные методы и инструменты для управления жизненным циклом машинного обучения, начиная с разработки моделей и заканчивая их развертыванием и мониторингом в боевых условиях. MLOps включает в себя автоматизацию процессов разработки и развертывания моделей, контроль версий, тестирование и непрерывную интеграцию и развертывание (CI/CD) моделей. Также в рамках MLOps проводится мониторинг работы моделей в реальном времени и сбор обратной связи для улучшения моделей. Основная цель MLOps - создать систему, которая позволяет быстро и эффективно разрабатывать, тестировать и развертывать модели машинного обучения, управлять их жизненным циклом и обеспечивать их работоспособность в реальном времени в продакшене.

(c) ChatGPT

# Что отличает AI от IT?

# В IT код преобразует данные в результат, в AI из данных рождается «код» модели

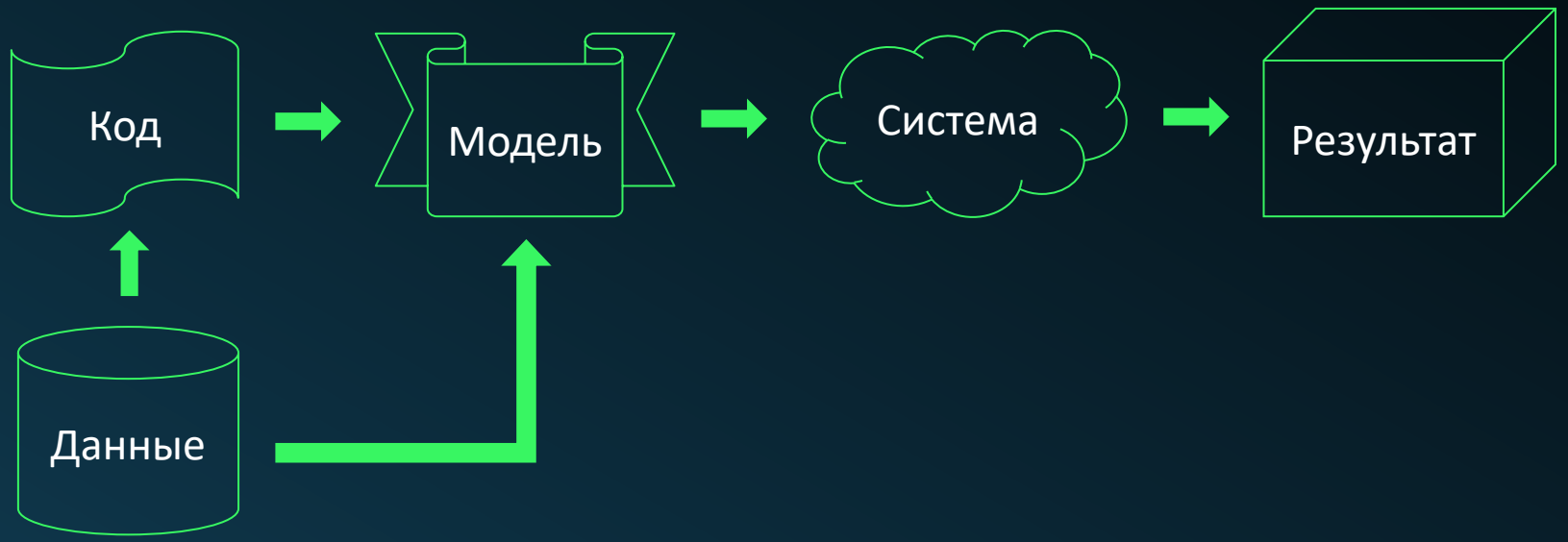
В ИТ код детерминировано преобразует входные данные в результат



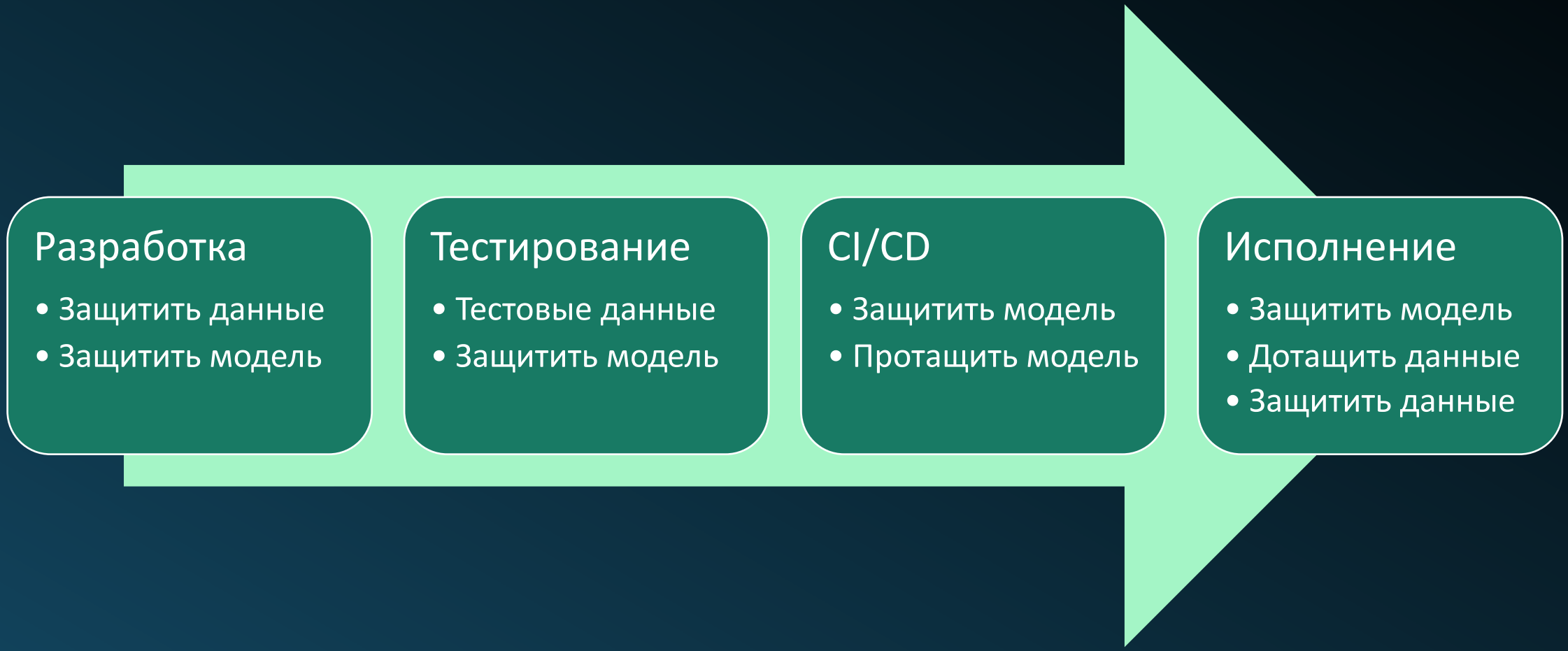

1. В AI код из данных получает модель

2. Модель на данных строит прогноз

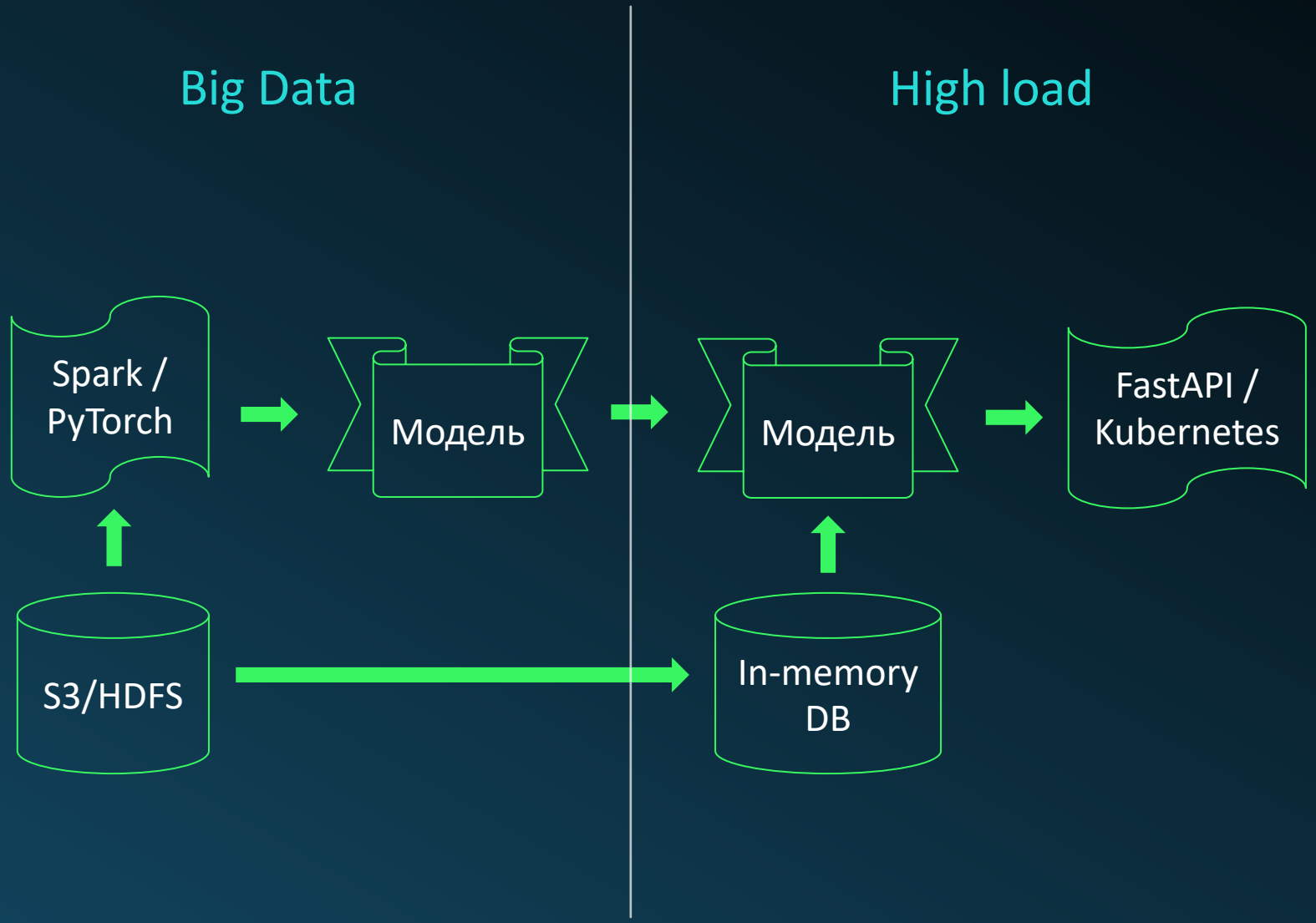
3. Модель интегрируется с системой, принимающей решение, которая приносит результат



# Данные – ценный актив, в процессе разработки, внедрения и исполнения модели и данные, и модель, нуждаются в защите



# Модели рождаются в озере больших данных, но многим из них придется жить в high load + low latency мире



# Качество кода зависит от разработчика, качество модели от многих факторов

- Качество и количество входных данных
- Качество разметки
- Алгоритм машинного обучения
- Гиперпараметры алгоритма
- Функция потерь
- Random seed
- …

Качество может меняться от запуска к запуску даже если вы ничего не меняли.

Качество деградирует со временем ☹

# Даже если данные доступны, не всякое их использование будет законным

DS: У нас есть данные, почему я не могу их использовать?!

DPO: 152 ФЗ

Требования 152 ФЗ:

- Согласие
- Состав данных
- Цели использования

# Как разрабатывают AI в Сбере?

# Защита данных – изолированный контур «фабрики данных»

**Офисная сеть**

- Контур «фабрики данных» физически изолирован от внешнего мира


**Контур Фабрики Данных**

Кластер источник 1

Кластер источник N

BitBucket

Супермаркет данных

**Кластер команды**

Jupyter Node

Spark Node 1

Spark Node N

АРМ Фабрики данных

GPU container

Командное пространство

Datalab

# Защита данных – изолированный контур «фабрики данных»

**Офисная сеть**

- Единое окно для получение и отслеживания доступа
- Инструмент поставки

D-people

APM Alpha

BitBucket

BAPM Фабрики данных

**Контур Фабрики Данных**

Кластер источник 1

Кластер источник N

Супермаркет данных

**Кластер команды**

Jupyter Node

Spark Node 1

Spark Node N

GPU container

Datalab

Командное пространство

# Защита данных – изолированный контур «фабрики данных»

**Офисная сеть**

- Совместная работа в рамках командных пространств


АРМ Alpha

**Контур Фабрики Данных**

Кластер источник 1

Кластер источник N

BitBucket

Супермаркет данных

АРМ Фабрики данных

**Кластер команды**

Jupyter Node

Spark Node 1

Spark Node N

GPU container

Командное пространство

Datalab

# Защита данных – изолированный контур «фабрики данных»

**Офисная сеть**

- Доступ в контур через «виртуальное рабочее место» из офисной сети

**Контур Фабрики Данных**

D-people

APM Alpha

BitBucket

BAPM Фабрики данных

Кластер источник 1

Кластер источник N

Супермаркет данных

**Кластер команды**

Jupyter Node

Spark Node 1

Spark Node N

GPU container

Datalab

Командное пространство

# DS-ы в фабрике данных глазами разработчиков в едином сегменте разработки и тестирования

# Тестирование моделей – математическая валидация

**Разработка кода**

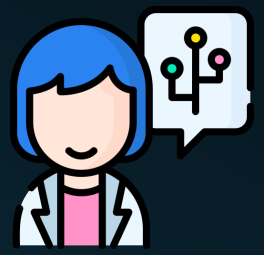
Программист

- Дебагер
- Юнит тесты
- Профилировщик

Тестировщик

- Тест кейсы
- Автотесты
- Нагрузочные тесты
- Деструктивные тесты
- …

**Разработка модели**



Data Scientist

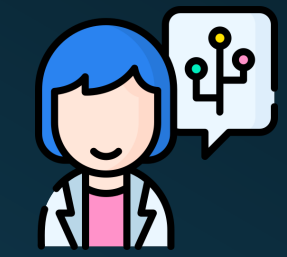
- Построение модели
- Кросс-валидация

Валидатор

- Кросс-валидация
- Альтернативное моделирование
- Тест стабильности
- Важность признаков

# Тестирование моделей – IT валидация

**Разработка модели**



Data Scientist

- Построение модели
- Кросс-валидация

Валидатор

- Кросс-валидация
- Альтернативное моделирование
- Тест стабильности
- Важность признаков

**Внедрение модели**

ML Инженер

- Адаптация модели для среды исполнения
- Интеграции
- Юнит тесты
- Автоматические self-тесты

IT Валидатор

- Идентичность данных в средах обучения и исполнения
- Соответствие математической модели

# Тестирование моделей – интеграционно-функциональное тестирование

ML Инженер

**Внедрение модели**

- Адаптация модели для среды исполнения
- Интеграции
- Юнит тесты
- Автоматические self-тесты

Тестировщик

- Тест кейсы
- Автотесты
- Нагрузочные тесты
- Деструктивные тесты
- …

**Ключевые сложности**

- Тестирование интеграций идет не в изолированном контуре, а в контуре разработки!
- Нет доступа к реальным данным – делаем синтетику или анонимизированный сэмпл
- Может не быть доступа к весам модели – используем мок
- Может не хватить вычислительной мощности – заменяем упрощенной версией

# Один день из жизни ML инженера

# Среды исполнения – защитить модель от среды и среду от модели

### Пакетные модели

- ПИМ – Пакетное Исполнение Моделей
- High Security Spark – изоляция на базе Linux Containers
- Доступ на базе ресурсно-ролевой модели

### NRT/онлайн модели

- MEF – Model Execution Framework
- Модель – контейнер в Pod-е
- Доступ во вне через набор «контроллеров»

### Дистрибуция модели

- ML Storage
- Инструмент на базе DVC + S3
- Контролирует доступ к весам
- Проверяет состояние quality gates

# Feature store – инструмент синхронизации данных обучения и исполнения


1. На основе базовых витрин данных формируются витрины признаков
2. Витрины признаков переносятся в in-memory хранилище для быстрого доступа
3. При необходимости, признаки в in-memory обновляются в потоковом режиме
4. Для вычисления модели используется информация из запроса + данные in-memory
5. Все что использовалось при вычислении логируется в аудит
6. При обучении используем витрины признаков и логи аудита

# Автомониторинг – непрерывный контроль качества моделей

**Технический**

- Пакетные модели
  - Готовы ли данные в источниках
  - Произведено ли вычисление модели
- NRT/онлайн модели
  - Получены ли пакетные обновления данных
  - Работает ли потоковый расчет
  - Получается ли запросы модель
  - Выполняется ли SLA на скорость и количество ошибок
- Инструменты
  - Prometheus + Grafana
  - ELK
  - Дежурная смена

**Математический**

- Стабильны ли распределения признаков
- Стабильны ли распределения прогнозов
- Находятся ли модельные метрики в допустимых диапазонах
- Какой прогноз динамики качества работы
- Инструменты
  - Sber.DS

# Фокусы ML Ops 2023

## Cloud-Native разработка и внедрение

- Запуск “Spark на Kubernetes” для обучения и исполнения
- Больше ресурсов при лучшей утилизации
- Меньше изменений кода при опромышливании
- Прозрачная интеграция Big Data и GPU вычислений

## Удаленная работа для D-people

- Отдельный контур для удаленной работы с данными
- Автоматическая анонимизация данных при передаче из основного контура в удаленный
- CI/CD для поставки результатов в основной контур

## Практика Continuous Improvements

- Измеряем Lead Time на основе данных из систем производственного процесса
- Ищем узкие места
- Используем лучшие практики IT разработки для сокращения процесса

# Спасибо за внимание!