INNOPOLIS UNIVERSITY

Зимняя школа для бакалавров

Иннополис
2025

AI-трансформация

Армен Бекларян, доцент лаборатории машинного обучения и представления данных УИ

ВВЕДЕНИЕ В ПРОДВИНУТУЮ АНАЛИТИКУ

Ключевые понятия в аналитике данных

Бизнес аналитика (BI) – общий термин, включающий в себя приложения, инфраструктуру и инструменты, а также лучшие практики, которые обеспечивают доступ к информации и ее анализ для улучшения и оптимизации решений и производительности. Платформы бизнес аналитики позволяют организациям создавать BI приложения, предоставляя возможности по трем категориям: анализ, предоставление информации (отчеты и дашборды), среду интеграции и разработки.

Продвинутая аналитика – автономное или полуавтономное изучение данных или контента с использованием сложных методов и инструментов, обычно выходящих за рамки традиционной бизнес аналитики (BI), для обнаружения более глубоких сведений, прогнозирования или выработки рекомендаций. К методам продвинутой аналитики относятся: машинное обучение, глубокое обучение, прогнозирование, интеллектуальный анализ данных текста, семантический анализ, сопоставление с образцом, анализ настроений, сетевой и кластерный анализ, многомерная статистика, анализ графов, моделирование, обработка сложных событий.

Ключевые понятия в аналитике данных

Искусственный интеллект (ИИ) – применение логических методов и методов продвинутой аналитики, включая машинное обучение, для интерпретации событий, поддержки и автоматизации решений, выполнения действий.

Искусственный интеллект (ИИ) – использование цифровых технологий для создания систем, способных выполнять задачи, которые, как принято считать, требуют человеческого интеллекта. ИИ не является новой технологией. Некоторые технологии ИИ существуют уже несколько десятилетий, но развитие компьютерных мощностей, доступность больших объемов данных и новое программное обеспечение привели к серьезному прорыву за короткий промежуток времени.

Искусственный интеллект (ИИ) – способность цифрового компьютера или робота, управляемого компьютером, выполнять задачи, которые обычно ассоциируются с разумными существами. Термин ИИ часто применяется к проекту разработки систем, наделенных интеллектуальными процессами, характерными для человека, такими как способность рассуждать, находить смысл, обобщать или учиться на прошлом опыте.

Ключевые понятия в генеративном ИИ

Обработка естественного языка (NLP) – технология, позволяющая превратить текст или аудиоречь в закодированную, структурированную информацию, основанную на соответствующей онтологии. Структурированные данные могут использоваться как просто для классификации документа, так и для идентификации выводов, процедур и участников.

Генеративный ИИ – метод ИИ, который изучает представление артефактов на основе данных и использует его для создания совершенно новых, уникальных артефактов, которые похожи на исходные данные, но не повторяют их. Генеративный ИИ может создавать совершенно новый контент включая текст, изображения, видео, аудио, структуры, компьютерный код, синтетические данные, рабочие процессы и модели физических объектов.

Семантический ИИ – технология, объединяющая машинное обучение и обработку естественного языка (NLP), позволяя программному обеспечению понимать речь или текст на уровне, близком к человеческому. При этом учитывается не только значение слов в исходном материале, но и контекст и намерения пользователя.

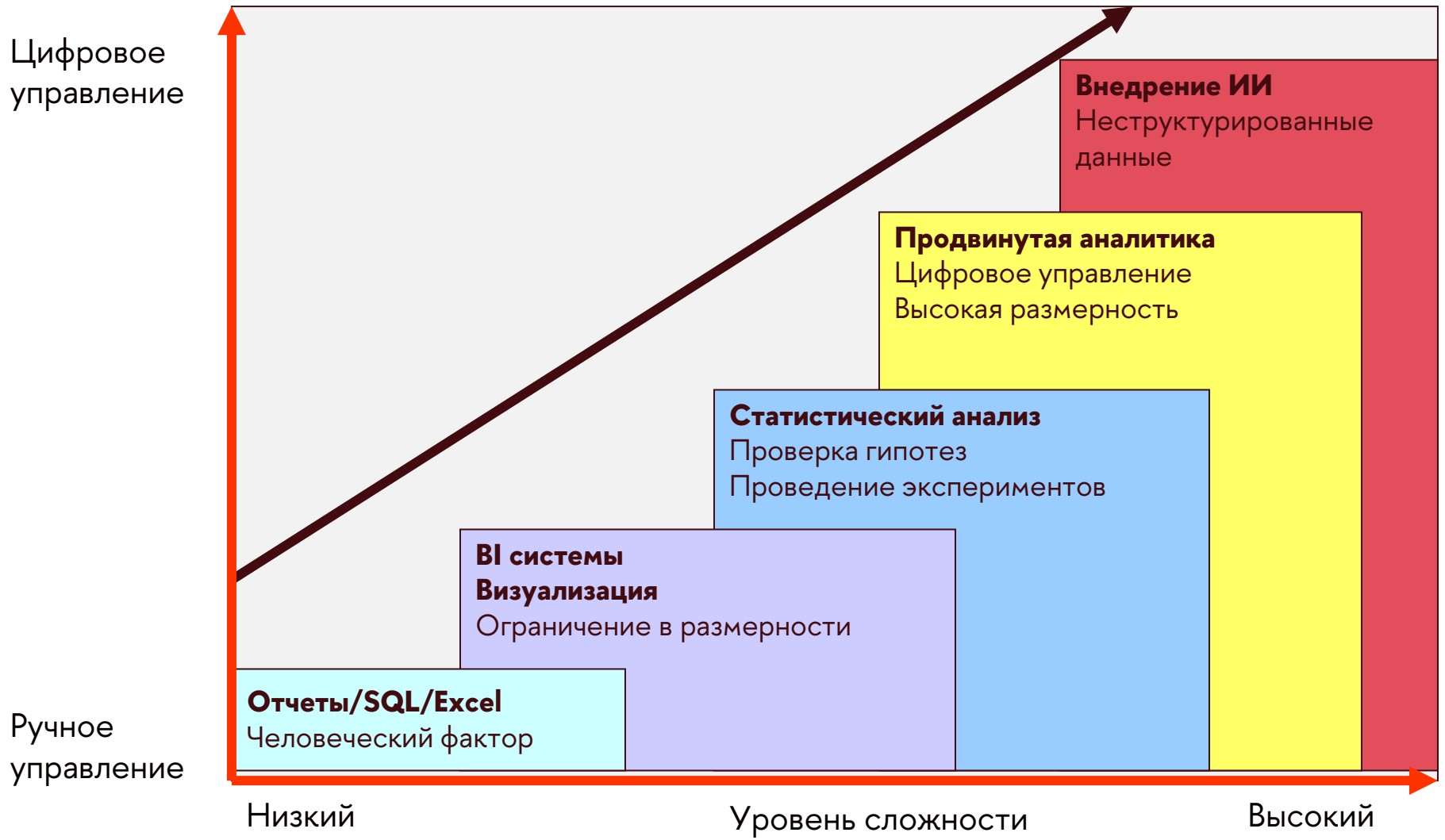
Ключевые понятия ИИ в промышленности

Цифровой двойник – цифровое представление объекта или системы реального мира. Реализация цифрового двойника это инкапсулированный программный объект или модель, которая отражает уникальный физический объект, процесс, организацию, человека или другую абстракцию. Данные из нескольких цифровых двойников могут быть агрегированы для получения комплексного представления о ряде реальных объектов, таких как электростанция или город, и связанных с ними процессов.

Мультиагентная система – тип системы искусственного интеллекта, состоящий из множества независимых но интерактивных) агентов, каждый из которых способен воспринимать окружающую среду и предпринимать действия. Агентами могут быть модели ИИ, программы, роботы и другие вычислительные объекты. Несколько агентов могут работать над достижением общей цели, которая выходит за рамки возможностей отдельных агентов, повышая их адаптивность и устойчивость.

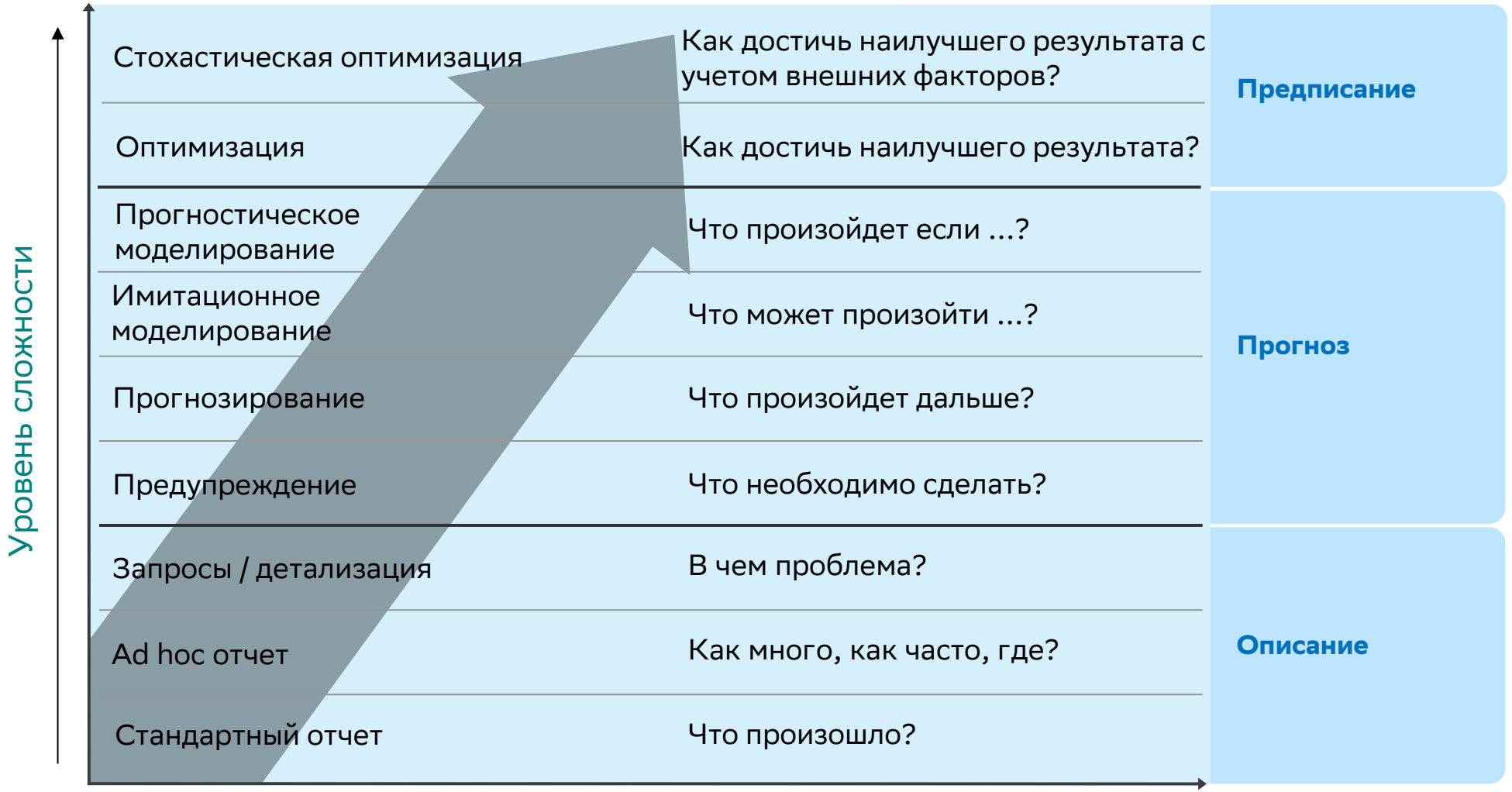
Инженерный ИИ – мультиагентная система типа Narrow AGI, сфокусированная на решении задач Системного Инжиниринга. ИИ, делающий работу инженера (желательно лучше, чем инженер).

Цифровое управление

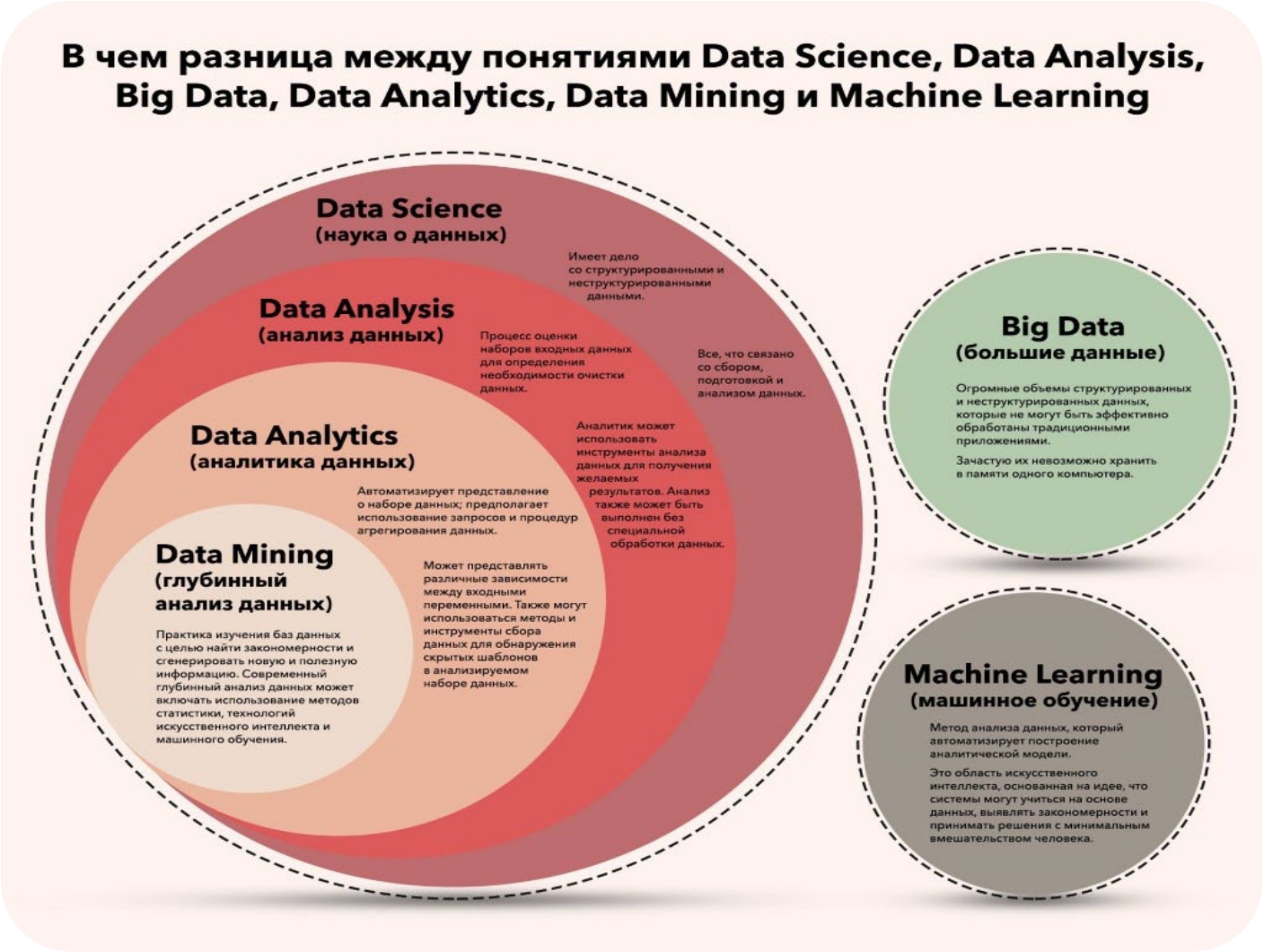


Цифровое управление – это целостный подход, который превращает информацию в знание, а знания в профит

Виды анализа



Виды анализа



Кто занимается анализом данных?

Data scientist

- Работа с данными
- Знание инструментов и методов
- Опыт решения задач

Менеджер

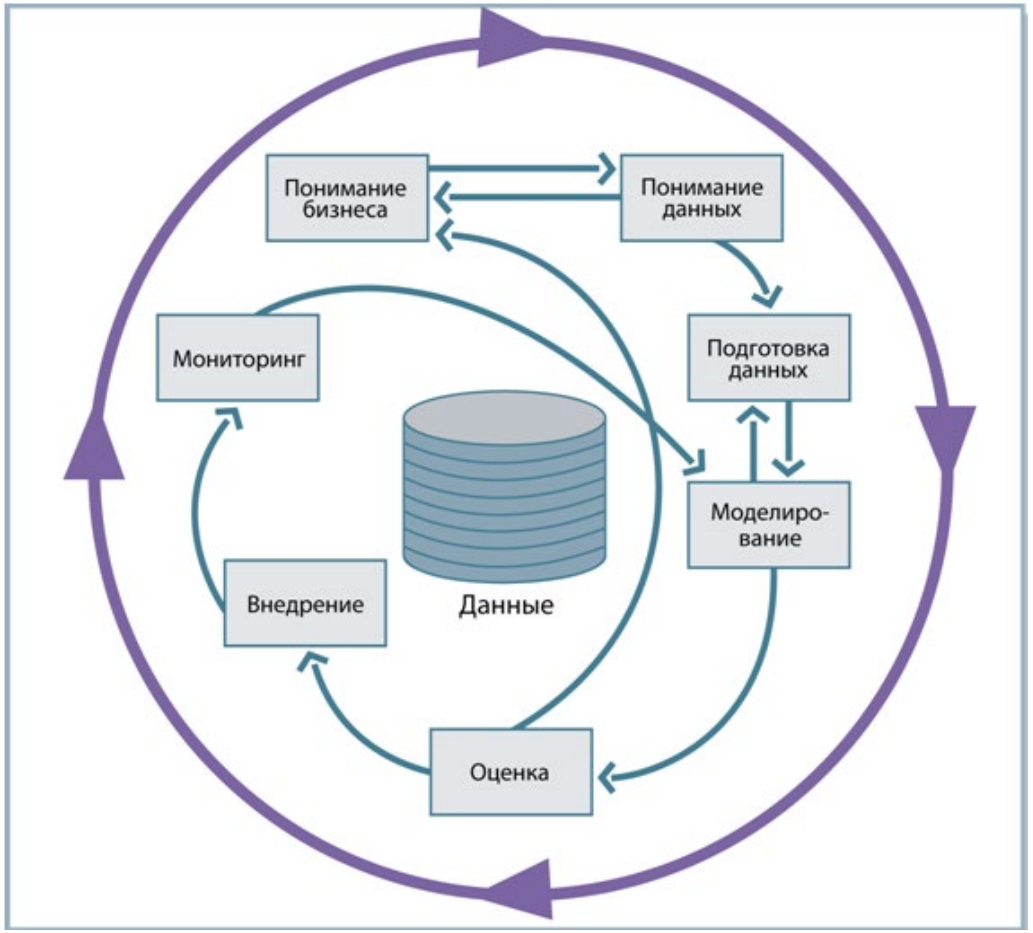
- Понимание, как работает машинное обучение
- Понимание узких мест, оценивание сроков

Заказчик

- Метрики качества
- Требования к данным
- Ограничения современных подходов

ПРОЕКТНАЯ МЕТОДОЛОГИЯ CRISP-DM/ML(Q)

Методология анализа данных



CRISP-DM

- **CR**oss-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining
- Описывает компоненты полного цикла проекта Data Mining
- Показывает итеративный характер интеллектуального анализа данных
- Не зависит от вендора и отрасли применения

Методология анализа данных

Понимание Бизнеса

Определить бизнес-цели
- Имеющийся задел
- Бизнес-цели
- Критерии успеха

Оценка текущей ситуации
- Инвентаризация ресурсов
- Требования, допущения и ограничения
- Риски и непредвиденные обстоятельства
- Терминология
- Затраты и дивиденды

Определить цель анализа данных
- Цели интеллектуального анализа данных
- Критерии успеха интеллектуального анализа данных

Составить план проекта
- План проекта
- Первоначальная оценка инструментов и методов

5%

Понимание Данных

Сбор исходных данных
- Отчет о сборе исходных данных

Описание данных
- Отчет с описанием данных

Исследование данных
- Отчет по исследованию данных

Проверка качества данных
- Отчет о качестве данных

15%

Подготовка Данных

Набор данных
- Описание набора данных

Отбор данных
- Обоснование включения / исключения

Чистка данных
- Отчет по очистке данных

Структурирование данных
- Описание атрибутов
- Описание целевых переменных

Интеграция данных
- Объединенные данные

Форматирование данных
- Переформатированные данные

40%

Моделирование

Выбор метод моделирования
- Техника моделирования
- Допущения моделирования

Создание дизайна теста
- Дизайн теста

Построение модели
- Настройка параметров модели
- Описание модели

Оценка модели
- Оценка модели
- Пересмотр настроек параметров

25%

Оценка

Оценка результатов
- Оценка результатов интеллектуального анализа данных
- Соответствие бизнес-критериям успеха
- Утверждение модели

Обзор процессов
- Обзор процессов

Определение следующих шагов
- Список возможных действий
- Итоговое решение

10%

Внедрение

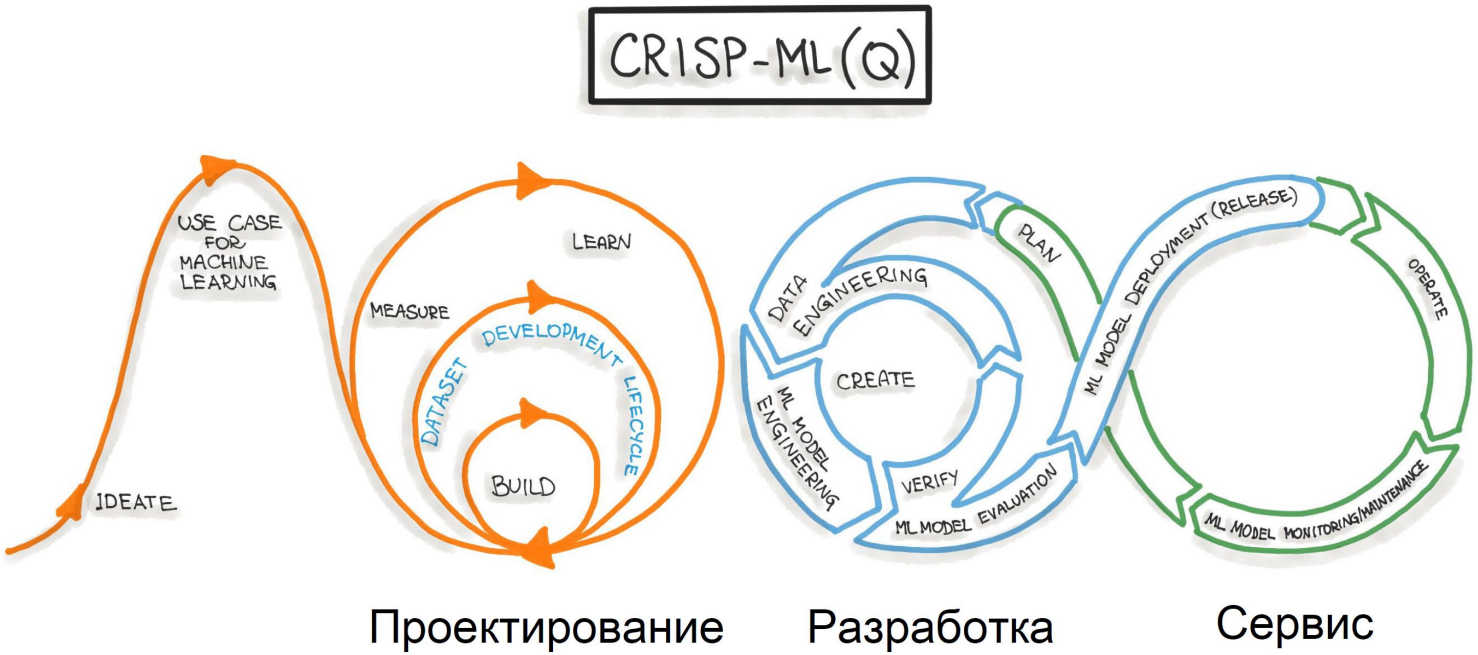
Подготовка итогового отчета
- Итоговый отчет
- Итоговая презентация

Обзор проекта
- Документация
- Полученный опыт

5%

Управление ИИ проектами

Стандартизация подходов и процессов позволяет унифицировать и масштабировать лучшие практики управления исследованиями и разработкой, в т.ч. распространяя их на другие домены. Подход к организации проектов машинного обучения получил название CRISP-ML(Q). Это аббревиатура от CRoss-Industry Standard Process model for the development of Machine Learning applications with Quality assurance methodology.

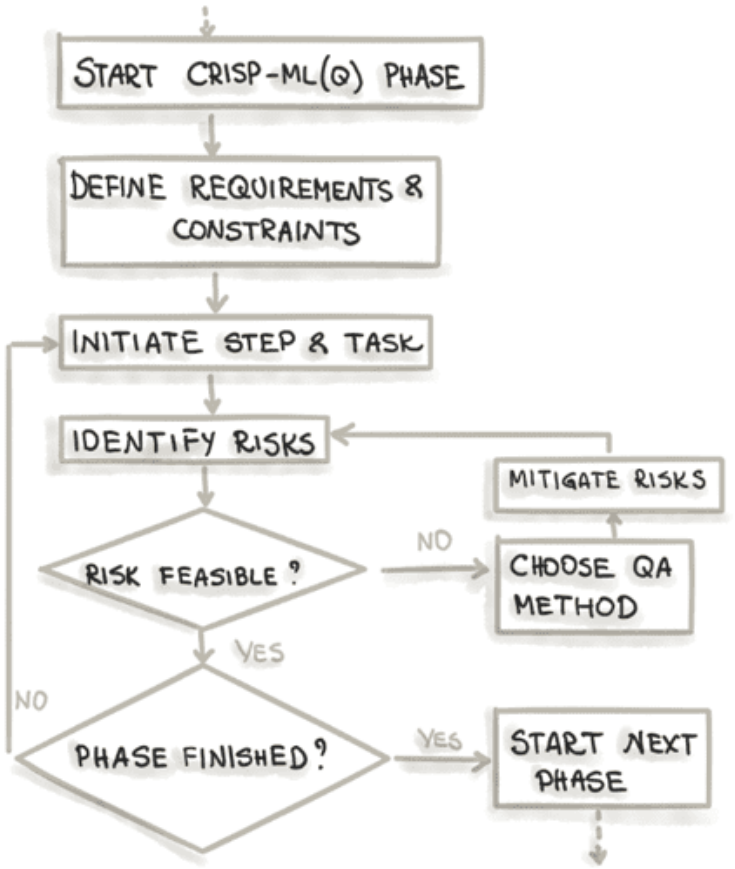


Модель CRISP-ML(Q)

Будучи основанной на CRISP-DM, модель процесса CRISP-ML(Q) описывает тоже шесть этапов:

- Понимание бизнеса и данных
- Инженерия данных (подготовка данных)
- Моделирование машинного обучения
- Обеспечение качества приложений машинного обучения
- Развертывание ML-модели
- Мониторинг и обслуживание ML-системы

Для каждого этапа модель CRISP-ML(Q) требует определения требований и ограничений, таких как производительность, требования к качеству данных, к устойчивости модели и пр. Также должны быть определены этапы создания экземпляра модели процесса, конкретные задачи, например, выбор алгоритма Machine Learning, обучение ML-модели. Особое внимание уделяется рискам, например, смещение данных, переобучение алгоритмов, отсутствие воспроизводимости и т.д. Для этого должны быть определены методы обеспечения качества, такие как перекрестная проверка, документирование процесса и результатов, логирование экспериментов и пр.

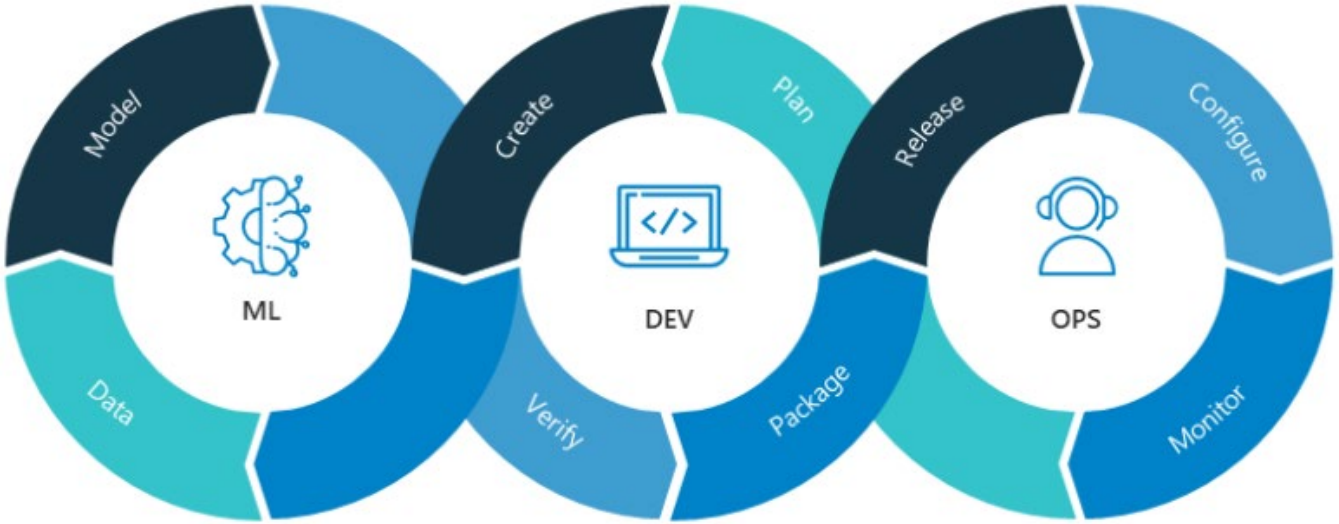


MLOps

Инструментально CRISP-ML(Q) поддерживается средствами **MLOps**-концепции, которая стремится устранить организационные и технические барьеры между разнопрофильными участниками процессов создания ML-систем, от подготовки данных до развертывания в производстве.

MLOps

это культура и набор практик комплексного и автоматизированного управления жизненным циклом систем машинного обучения, объединяющие их разработку (Development) и операции эксплуатационного сопровождения (Operations), в т.ч. интеграцию, тестирование, выпуск, развертывание и управление инфраструктурой. Можно сказать, что MLOps расширяет методологию CRISP-DM и реализует подход CRISP-ML(Q) с помощью Agile-подхода и технических инструментов автоматизированного выполнения операций с данными, ML-моделями, кодом и окружением.



6 фаз CRISP-ML(Q)

Фаза CRISP-ML(Q)	Задачи
Понимание бизнеса и данных	Определить бизнес-цели Преобразовать бизнес-цели в цели машинного обучения Собрать и проверить данные Оценить осуществимость проекта Подтвердить концепцию с помощью POC (Proof Of Concept)
Инженерия данных	Выбор фичей Выбор данных Балансировка классов Очистка данных (подавление шума) Разработка фичей Увеличение данных Стандартизация данных

6 фаз CRISP-ML(Q)

Фаза CRISP-ML(Q)	Задачи
Разработка моделей машинного обучения	Определить показатель качества модели Выбрать алгоритма машинного обучения (базовый выбор) Добавить специфику предметной области для специализации модели Обучить модель Сжатие модели Ансамблевое обучение Документировать ML-модели и эксперименты
Оценка модели машинного обучения	Проверить производительность модели Определить надежность Улучшить объяснимость модели Принять решение о развертывании в производстве Документировать этап оценки

6 фаз CRISP-ML(Q)

Фаза CRISP-ML(Q)	Задачи
Развертывание модели	Оценить модель в рабочем состоянии Обеспечить приемлемость и удобство использования Организовать управление моделью Выбрать стратегию развертывания и реализовать ее
Мониторинг и обслуживание модели	Обеспечить мониторинг эффективности и результативности предоставления прогнозов модели Сравнить результаты с ранее указанными критериями успеха (пороговыми значениями) Повторно обучить модель (при необходимости) Собрать новые данные Выполнить разметку новых точек данных Повторить задачи этапов моделирования и оценки, чтобы обеспечить непрерывность MLOps-процессов

Особенности CRISP-ML(Q)

Ключевые ошибки команд в проектах с машинным обучением:

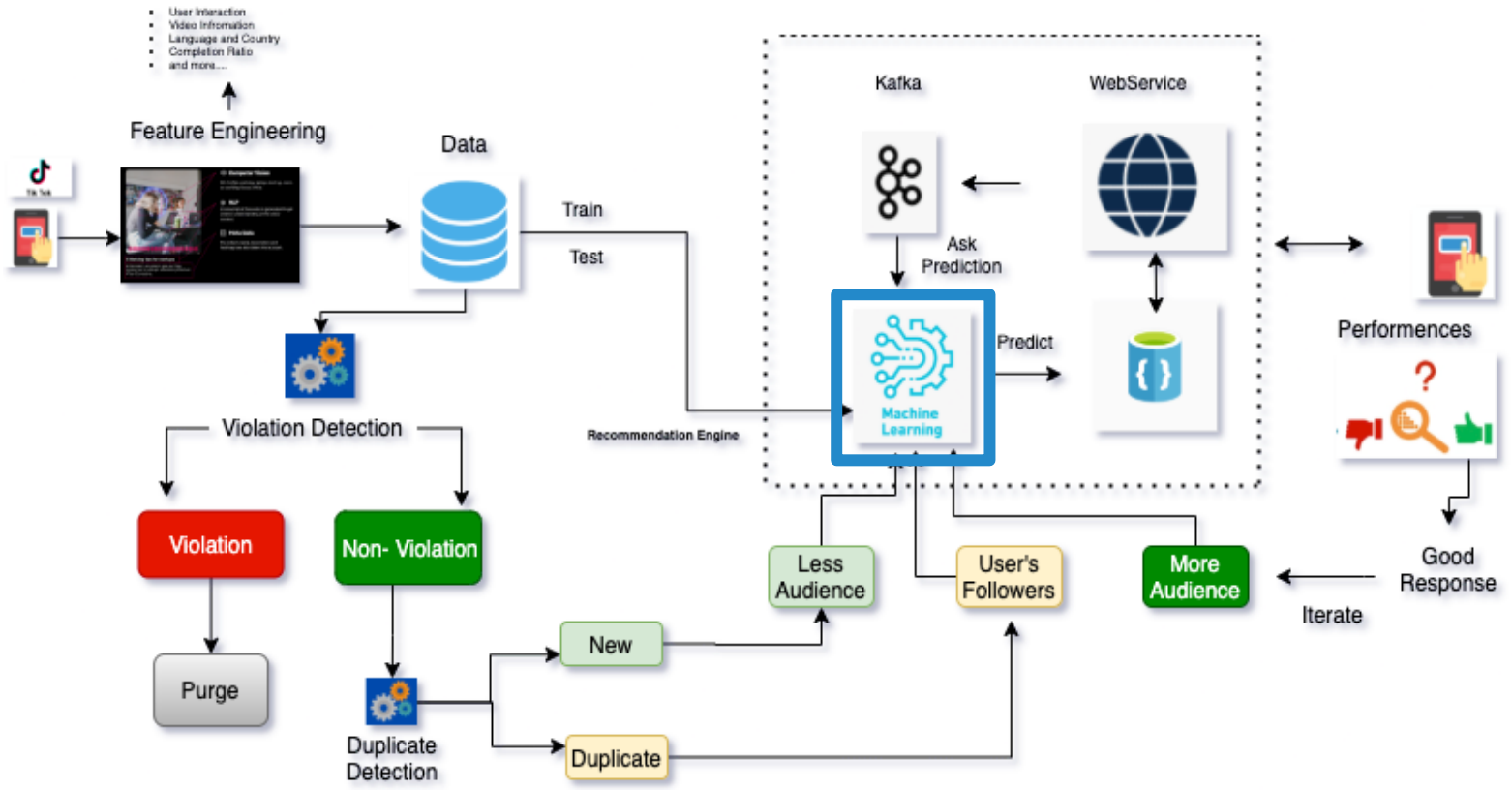
- Неверно определяют возможности для применения ML
- Не знают, как строить работу над ML-проектом
- Не проверяют риски на старте и впустую тратят много ресурсов
- Представители бизнеса не говорят на языке ML-специалистов
- ML-команда работает изолированно от остальных команд
- Команда увлеклась техническими экспериментами и упустила из внимания цели бизнеса

Решения:

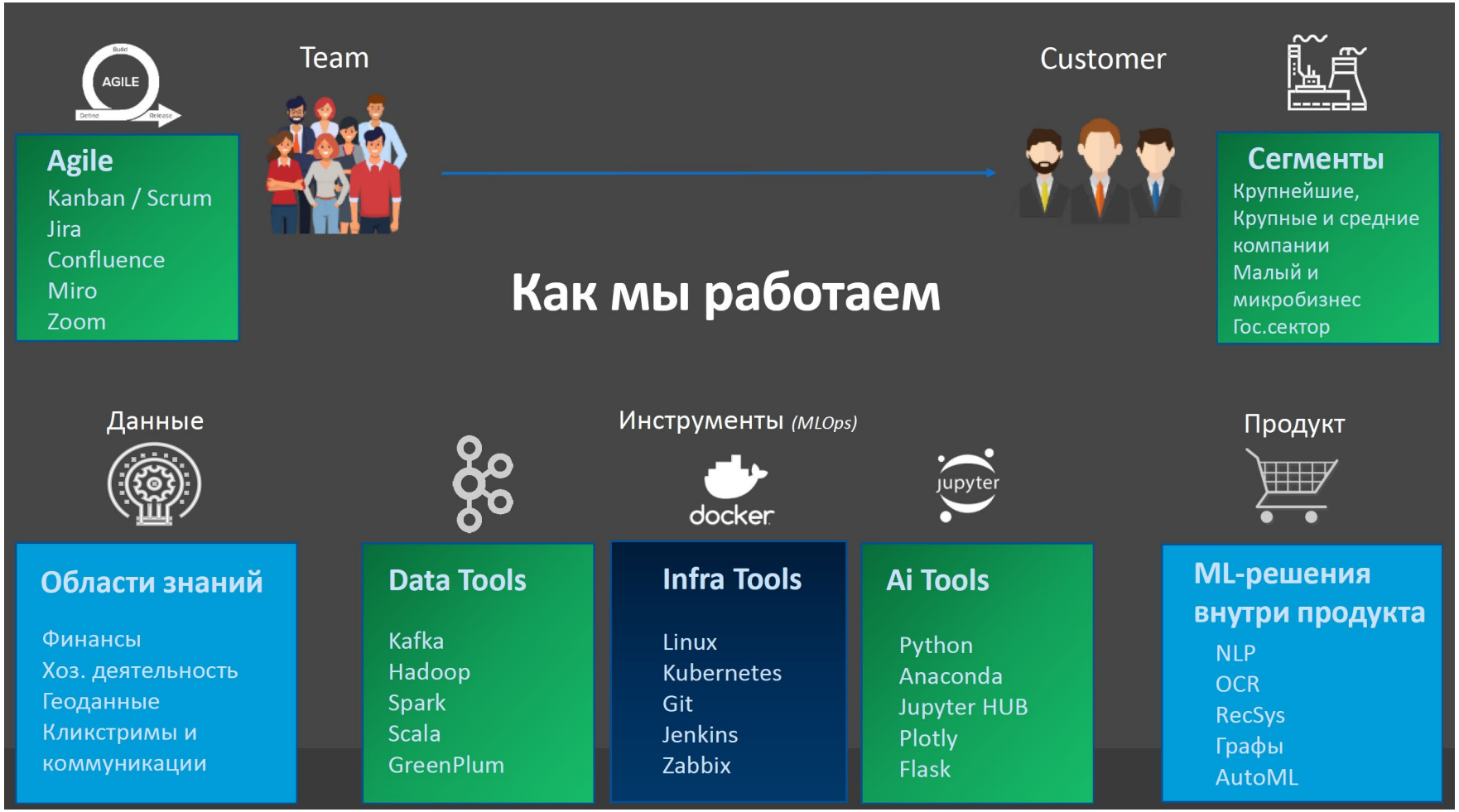
- Видеть возможности для использования ML в своих проектах
- Эффективно структурировать работу над ML-проектами
- Говорить на языке ML-специалистов
- Правильно выбирать приоритеты на каждом шаге проекта

Модель ≠ Продукт

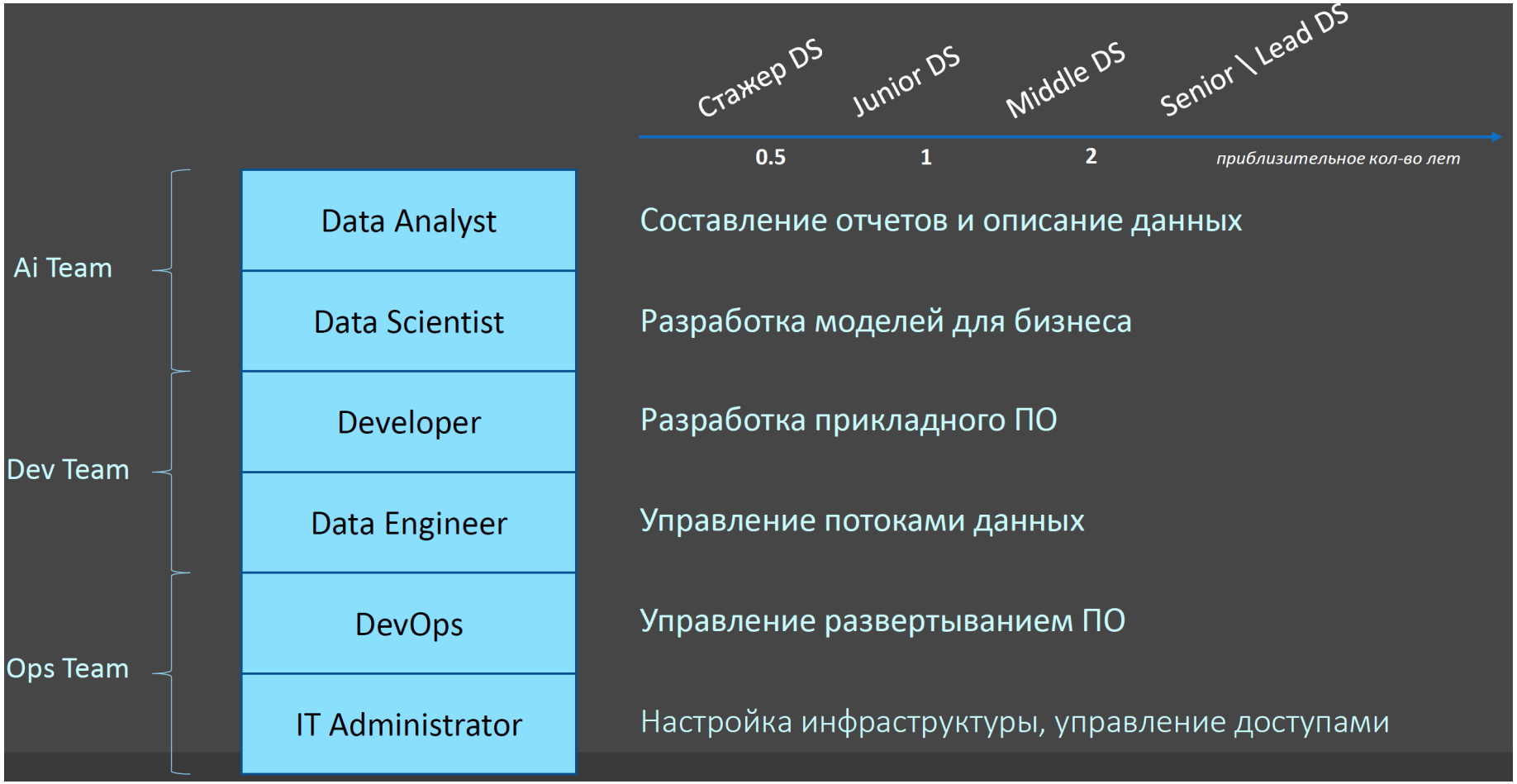
Модуль машинного обучения является только **частью** прикладного решения для клиента



Основные роли и карьерный путь D-people



Основные роли и карьерный путь D-people



УНИВЕРСИТЕТ
ИННОПОЛИС