Ведение DS-проектов

Лекция 1

Алексей Кузьмин
Директор
разработки
ДомКлик.ру

Работаю в ДомКлик.ру с 2016 года
Руковожу направлением Data Science и работы с данными
До этого работал в компании ABBYY, где занимался распознаванием языков со сложной письменностью
Окончил мехмат МГУ

О чем поговорим?

Сегодня на лекции

01 **Data Driven и аналитика**
Что это такое и почему нужно

02 **Участники процесса работы с данными**
Одними DS’ами сыт не будешь

03 **Как собирают команды для работы с данными**
Откуда они изначально берутся и как могут быть организованы

04 **Процесс исследования данных**
Как он выглядит?

Машинное обучение и аналитика в компании

Зачем оно?

5%

Настолько более продуктивны компании,
использующие data-driven подход

Brynjolfsson, et al 2011. Strength in numbers: how does data-driven decision making affect firm performance?

$13

Столько приносит аналитика на каждый вложенный доллар

Nucleus Research. 2014. http://bit.ly/XQFDbv

Что такое data-driven?

Что не делает компанию data-driven?

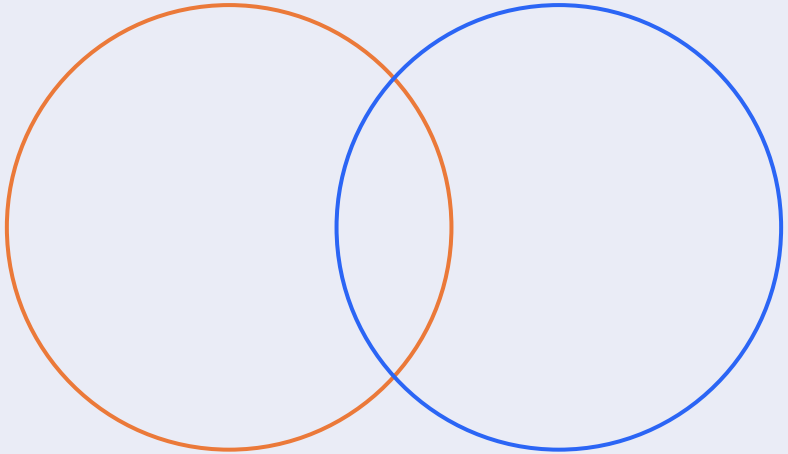
- Множество отчетов
- Множество дашбордов
- Множество моделей

Аналитика

Помощь в принятии решений на основе данных

Аналитика

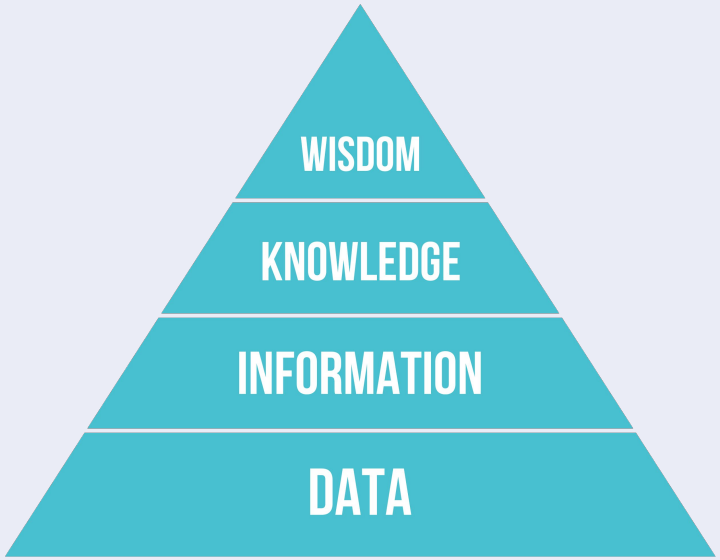
Задача аналитика
Найти 70% наиболее ценной информации / инсайтов



Задача менеджера
управлять рисками и принимать решение

DIKW

- В основании находится уровень данных.
- Информация добавляет контекст.
- Знание добавляет «как» (механизм использования)
- Мудрость добавляет «когда» (условия использования)

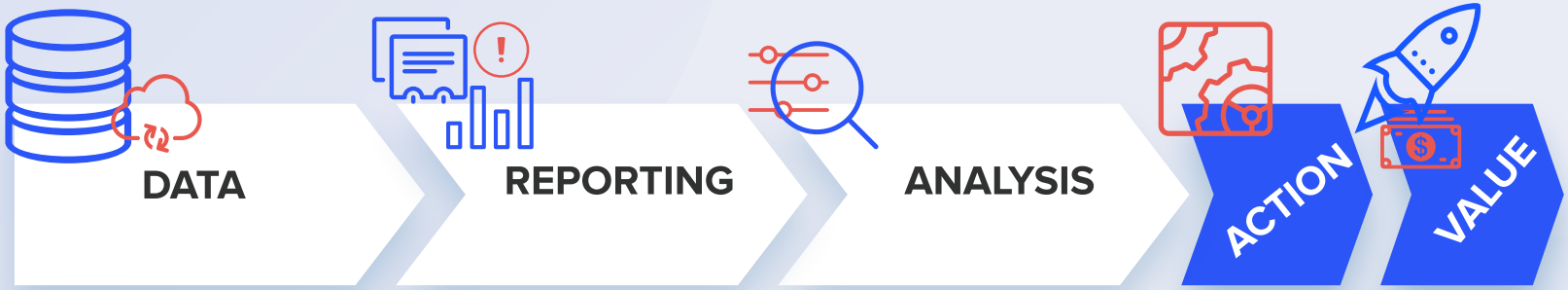


Data-Driven - значит иметь аналитику

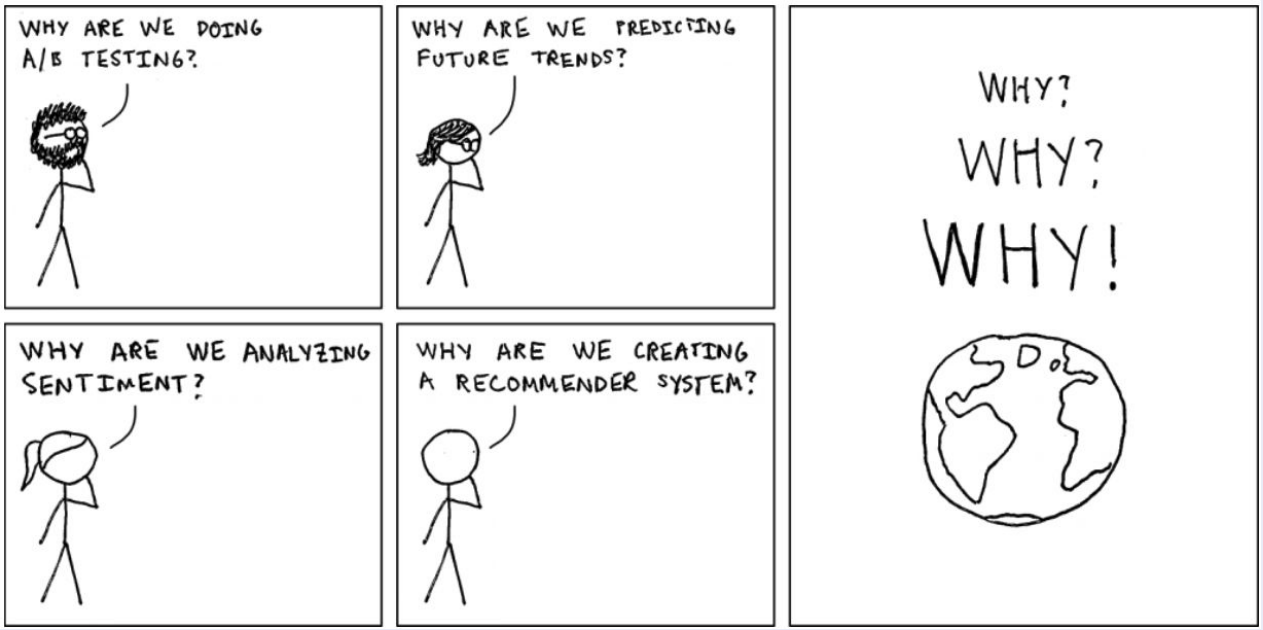
Отчетность	**Аналитика**
Описательная	Предиктивная
Что?	Почему?
“Смотрит назад”	“Смотрит вперед”
Задает вопросы	Отвечает на вопросы
Данные -> Информация	Данные + Информация -> Инсайты
Отчеты, Дашборды	Открытия, Модели
Нет контекста	История

Уровни аналитики

Analytics Value Chain



Аналитика - это понимание “зачем”



Для того, чтобы быть data-driven, должна быть культура работы с данными

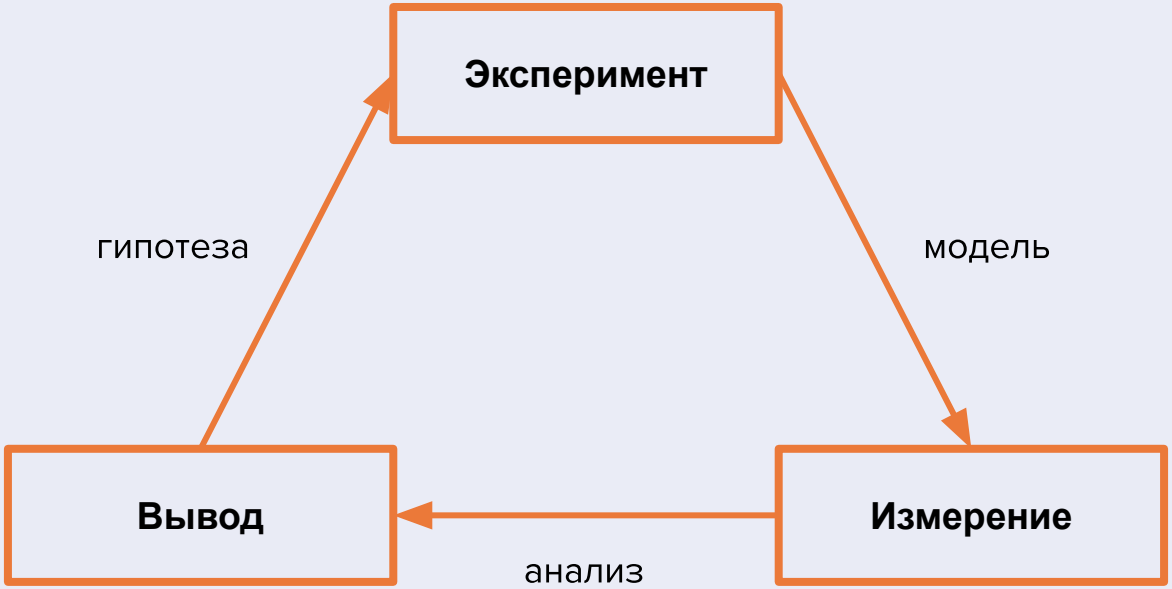
Из чего она складывается?

Data-driven - Тестирование

Инновационность достигается через регулярное тестирование. Аналитика дает гипотезы, которые нужно уметь проверять

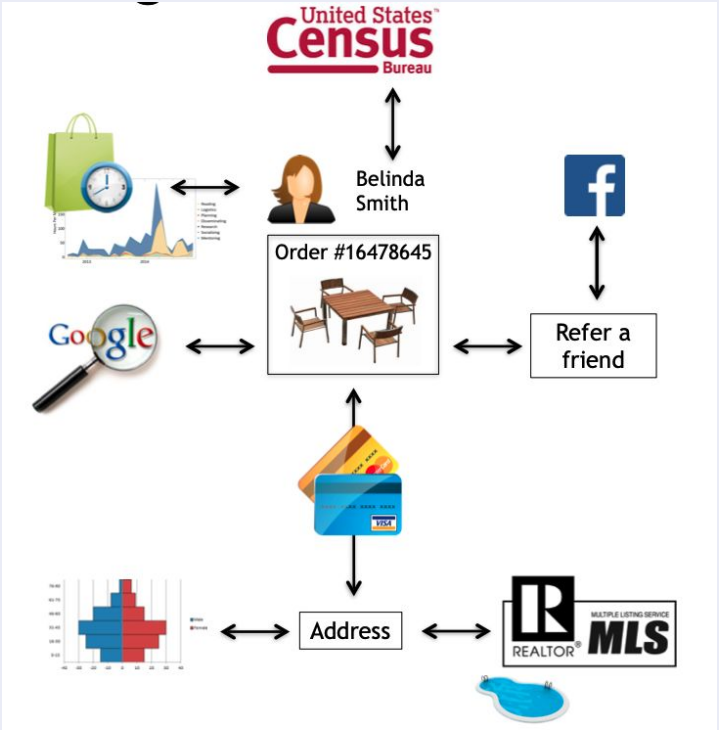
Важно, что мы принимаем решение на основе измеримых тестов, а не на основе мнения “самого опытного эксперта”

Data-driven - Итерации



Data-driven - Открытость

- Не должно быть сокрытия или утаивания части данных.
- Данные должны работать все вместе на единый анализируемый контекст.



Data-driven - Качество данных

- Некорректные данные ведут к некорректным решениям и выводам
- Более того, некорректные данные ведут к разным цифрам в разных срезах
- Качество подразумевает “единый источник правды”

Data-driven - сначала цель

- Сначала понимаем что хотим сделать
- Понимаем какую цель хотим достичь
- Определяем метрики успеха/неудачи
- Потом делаем исследование

На всякий случай

- Слепо следовать за данными - тоже плохая идея =)
- Не забывайте включать “голову”

https://www.youtube.com/watch?v=a2QIH2uz3p8

YouTube

Girls Crash into Lake following Bad GPS directions

CrushingBastards

Subscribe 744

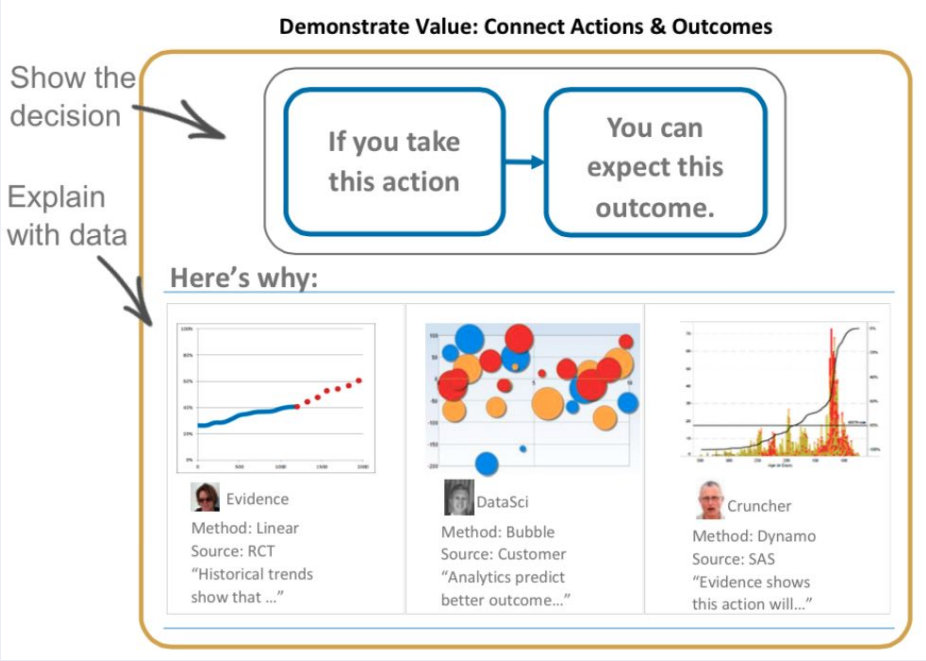
165,605

Аналитический продукт

Результат работы аналитика:

- Это **бизнес-польза**, которую можно получить сделав определенное действие
- Оно может быть оформлено в виде отчета или веб-сервиса
- **Но!** Оно **бесполезно**, пока оно не используется и не приносит пользу

Ну и конечно “аналитика” - это не только выводы, но еще и “упаковка”



Продавайте свой продукт

Команда для работы с данными

DS - это круто, но сам по себе он не принесет value

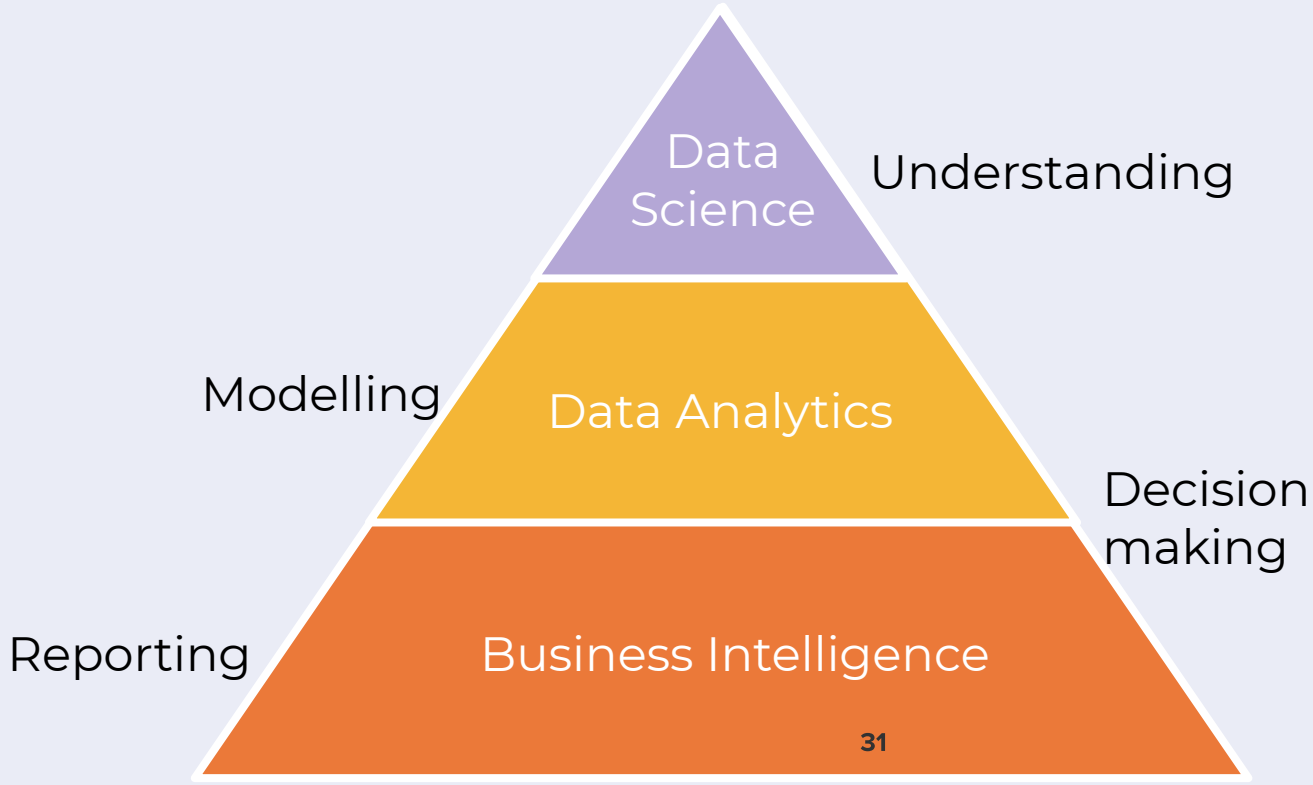
Для эффективной работы с данными помимо построения моделей и DS-аналитики нужно уметь:

- Доставать данные из различных бд, причем делать это на регулярной основе
- Разрабатывать веб-сервисы и прикладное ПО для применения моделей
- Делать отчетность для показа заказчику
- Собирать и приоритезировать потребности, чтобы обеспечивать актуальный бэклог
- Оценивать эффект от исследований/моделей
- ...

Роли в команде

DS специалисты	Аналитики	BI специалисты
Разработчики	Data Engineer-ы	Специалисты по БД
Владелец продукта	Заказчики	...

РАЗНЫЕ РОЛИ СПЕЦИАЛИСТОВ ПО РАБОТЕ С ДАННЫМИ



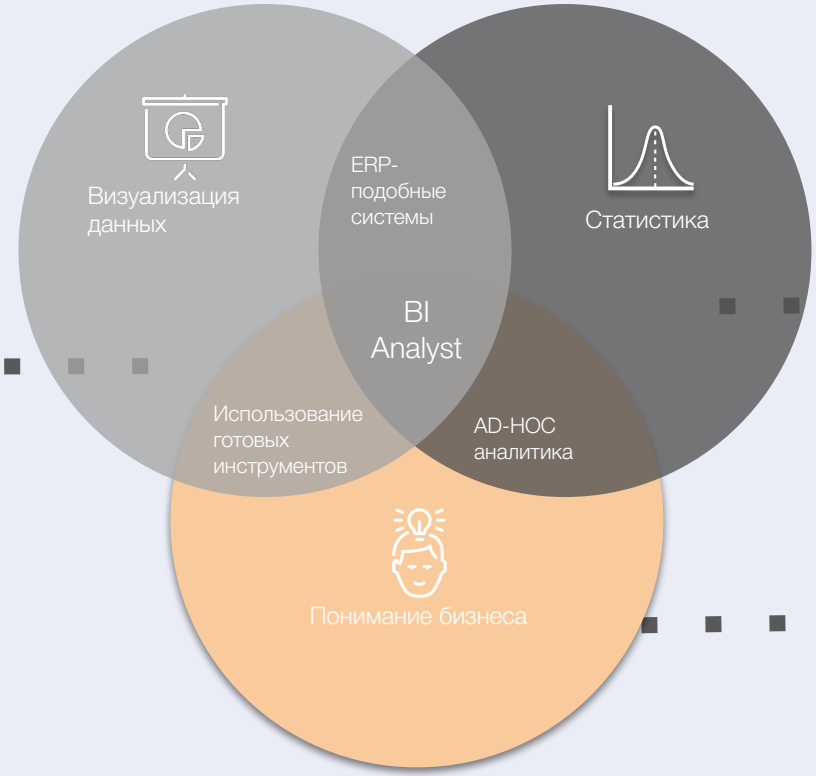
BI-Аналитик

Преобразует данные в доступную для лиц, принимающих решение, информацию в форме отчетов и dashbord’ов

- Сбор бизнес-данных. опросы, отчётность и тд
- Интерпретация большого количества данных. акцент лишь на ключевых факторах эффективности
- Моделирование исхода различных вариантов действий
- Отслеживание результатов принятия решений

BI Аналитик

создание дашбордов и выдвижение требований бизнеса к аналитической инфраструктуре



способность проверить значимость результата

диагностика текущего состояния бизнеса

Аналитик данных

- Составление, валидация, оценка метрик
- Пониманием взаимосвязи разных метрик
- Проведение экспериментов, АБ-тесты
- Прогнозирование
- Рекомендации бизнесу

Аналитик данных

способность создавать аналитические решения и использовать готовые инструменты

способность проверить значимость результата

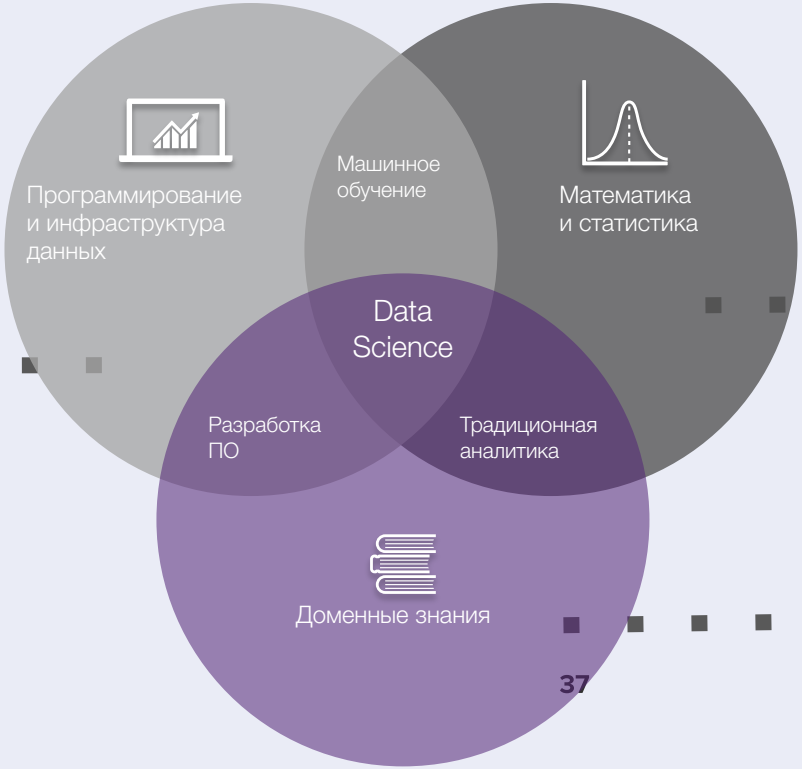
визуализация и представление результатов исследований

DS - специалист

- Извлечение важной информации и инсайтов
- Построение и валидация моделей
- Иногда: подготовка отчетов
- Создание готовых приложений,
  позволяющие решать те или иные предиктивные задачи

Data Scientist

быстро проверять гипотезы, тестировать модели и использовать готовые решения



задавать правильные сложные вопросы

понимать боли бизнеса и говорить с ним на одном языке

Разработчик

Если результат работы - это сервис для работы в режиме реального времени, то для него нужно:

- Разработать API
- Провести интеграции
- Настроить получение требуемых данных
- Встроить модель
- Протестировать
- Поставить заказчику

Data Engineer

ETL-специалист:

извлечение / преобразование/ загрузка данных.

- Сбор данных источников (эксель, БД, 1с, ...)
- Структурирование данных
- Подготовка выгрузок

Data Engineer

+ нереляционные источники данных
+ работа с большими данными
+ понимание потребностей DS
+ знание языков программирования
+ ...

Специалист по БД

Специалист-разработчик БД:

- Вопросы оптимального и надежного хранения данных
- Обеспечения быстрого и удобного к ним доступа

Архитектор баз данных

менеджер с глубоким понимаем БД и IT вообще. Главная задача: разработка понятной и масштабируемой БД/ХД

- Выбор технологии для хранения данных
- Создание и оптимизация запросов
- Составление план разработки и ТЗ для подчиненных
- Проектирование и оптимизация БД
- Контроль безопасности БД

Остальные роли

Владелец продукта - отвечает за то, что проекты в команде востребованы и нужны

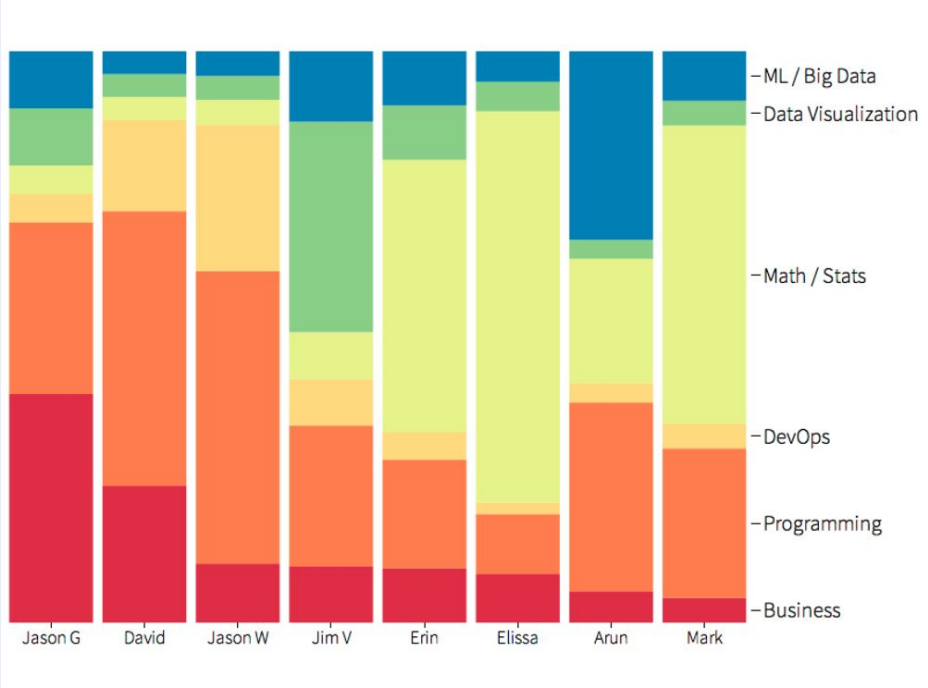
Бизнес-Заказчик - “человек” с потребностью / болью которую мы хотим решить

Менеджер продукта - человек, который следит за движением по роад мапу

Как понять кто нужен

- В каком состоянии у вас данные?
- Какие проблемы решит появление специалиста?
- Какие перед ним будут стоять задачи?
- Какой продукт вы ждете на выходе?
- Какими компетенции для этого нужны?

Роли в команде должны дополнять друг-друга



Поддержка руководства критична для успеха аналитического процесса

Executive Sponsorship Is So Vital To Analytical Competition...

Tom Davenport
Competing on Analytics

Data-Driven CEO

Основные направления деятельности Data-driven CEO:

- Стратегическое планирование на основе данных
- Понимание аналитики
- Технологическая осведомленность

CDO

- Содействие принятию решений на основе данных для поддержки ключевых инициатив компании
- Проверка, что компания собирает правильные данные
- контроль и продвижение аналитики по всей компании

Как собрать и организовать команду

Создание команды

4 способа:

- Трансформация
- Создание с нуля
- Как сервис
- Краудсорсинг

Трансформация

Преобразование и реорганизация с минимальным изменением текущей организационной структуры

• Отрасли, требующие глубокого знания предметной области (такие как генетика и секвенирование ДНК)

• Старые компании, которые хотят внедрить науку о данных в свой бизнес

• Компании, которые хотят обогатить собственные наборы навыков

Создание с нуля

Начинающие компании

- Компании, которые хотят ...
  - уделять больше внимания аналитике данных
  - Запустить новые ds-продукты
- Компании, где данные являются продуктом
- Глубокое знание предмета менее критично для аналитики

Как сервис

Когда привлекать DSaaS:

- Предпочтительно не менять существующую организационную структуру
- Когда создание или преобразование не являются важными для выживания компании

Учитывать уровни обслуживания (SLA) при определении того, привлекать ли внутренние ресурсы или внешних поставщиков

Краудсорс

Когда:

- Проблема «открыта» по природе

- Готовы принять мнения от распределенных и разнообразных групп людей

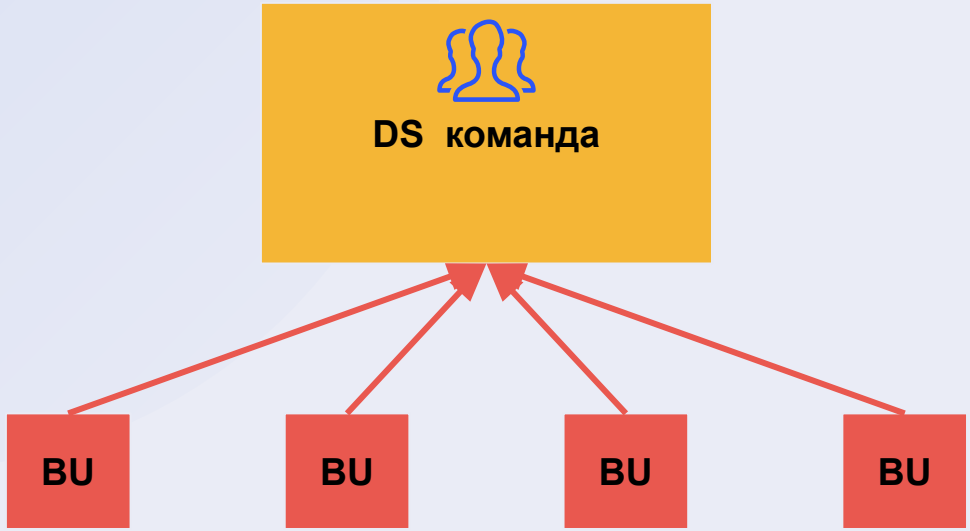
- Существует резервный план на случай «массового отказа»

Примеры: Википедия, Kaggle

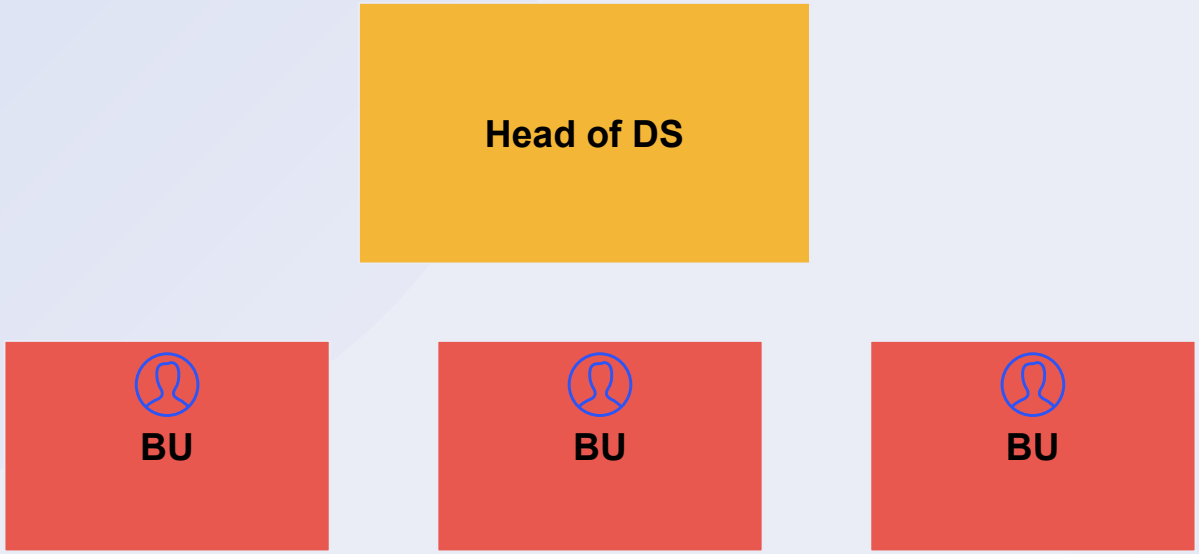
Сравнение

	Трансформация	Создание	Аутсорс	Краудсорс
Плюсы	Сильное знание предметной области • Знание бизнес-процессов • Новые таланты повышают уровень команды	Контроль над навыками • Больше гибкости • Высокое качество обслуживание	Возможность масштабирования по требованию • Можно получить лучший результат, чем внутри компании • Учиться у внешних экспертов	Мудрость толпы Разнообразные перспективы Более низкая стоимость Быстрые результаты
Минусы	Риск гомогенного мышления • Некоторые члены команды могут сопротивляться изменениям	Найм и передача знаний занимают много времени • Время, необходимое для поиска и найма правильных членов команды	Поставщик может не понять уникальные процессы компании • Трудно вернуть экспертизу на места • Снижение качества обслуживания с течением времени	Нет SLA; результат не гарантирован • Сложно разработать «открытую» задачу

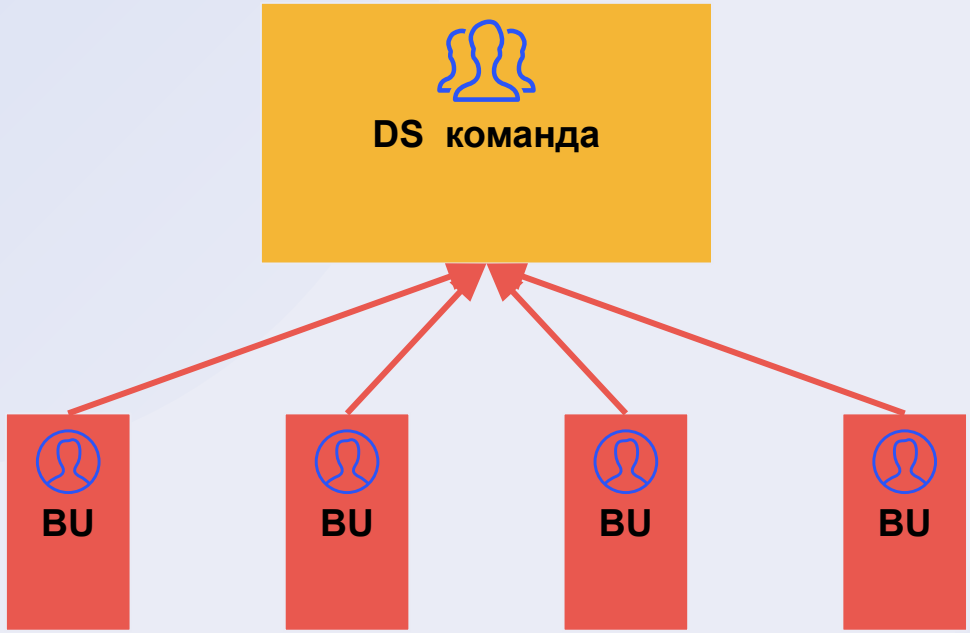
Организационная модель - Централизованная



Организационная модель - Децентрализованная

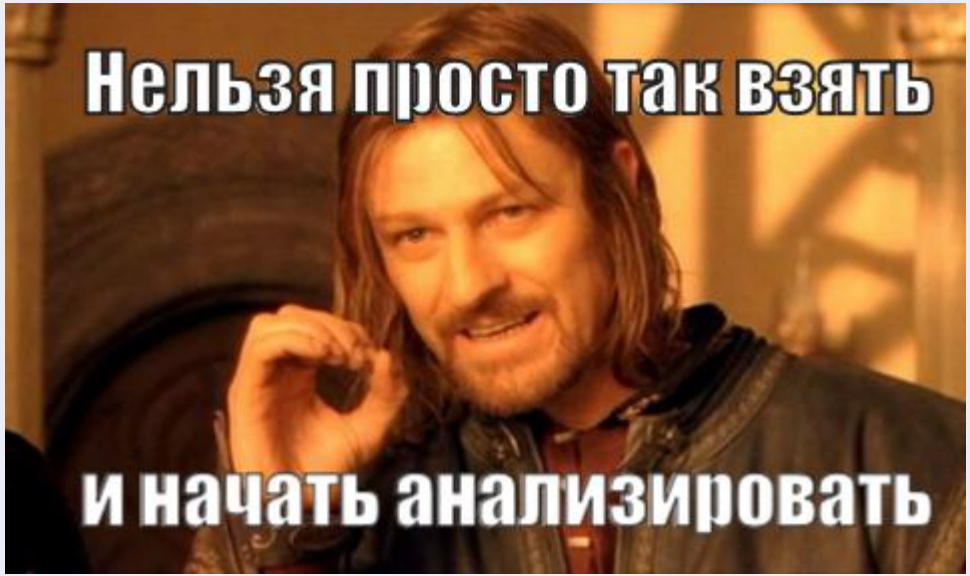


Организационная модель - Гибридная

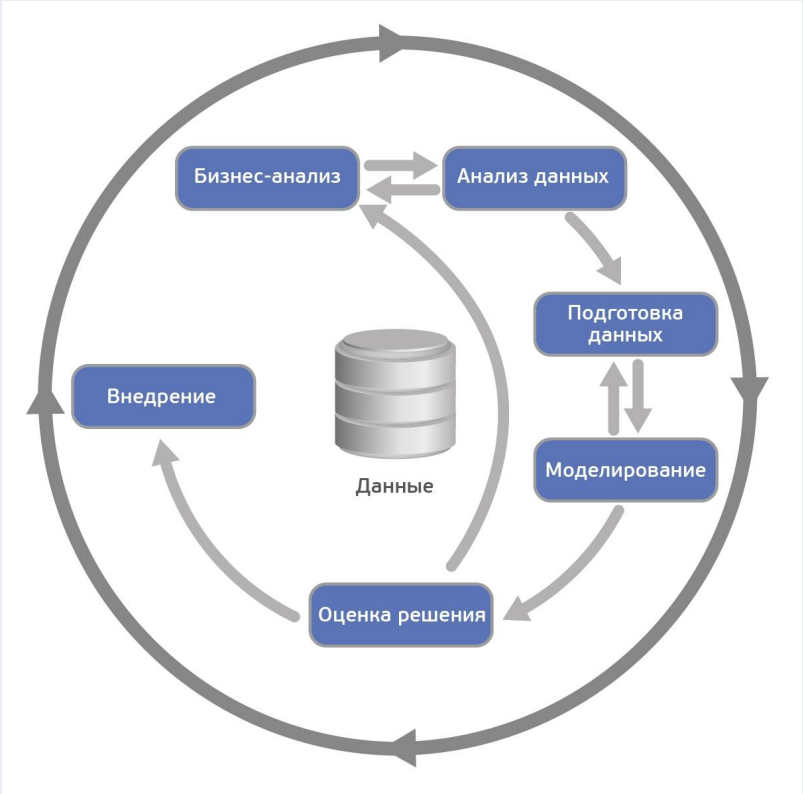


Как запустить процесс в команде?

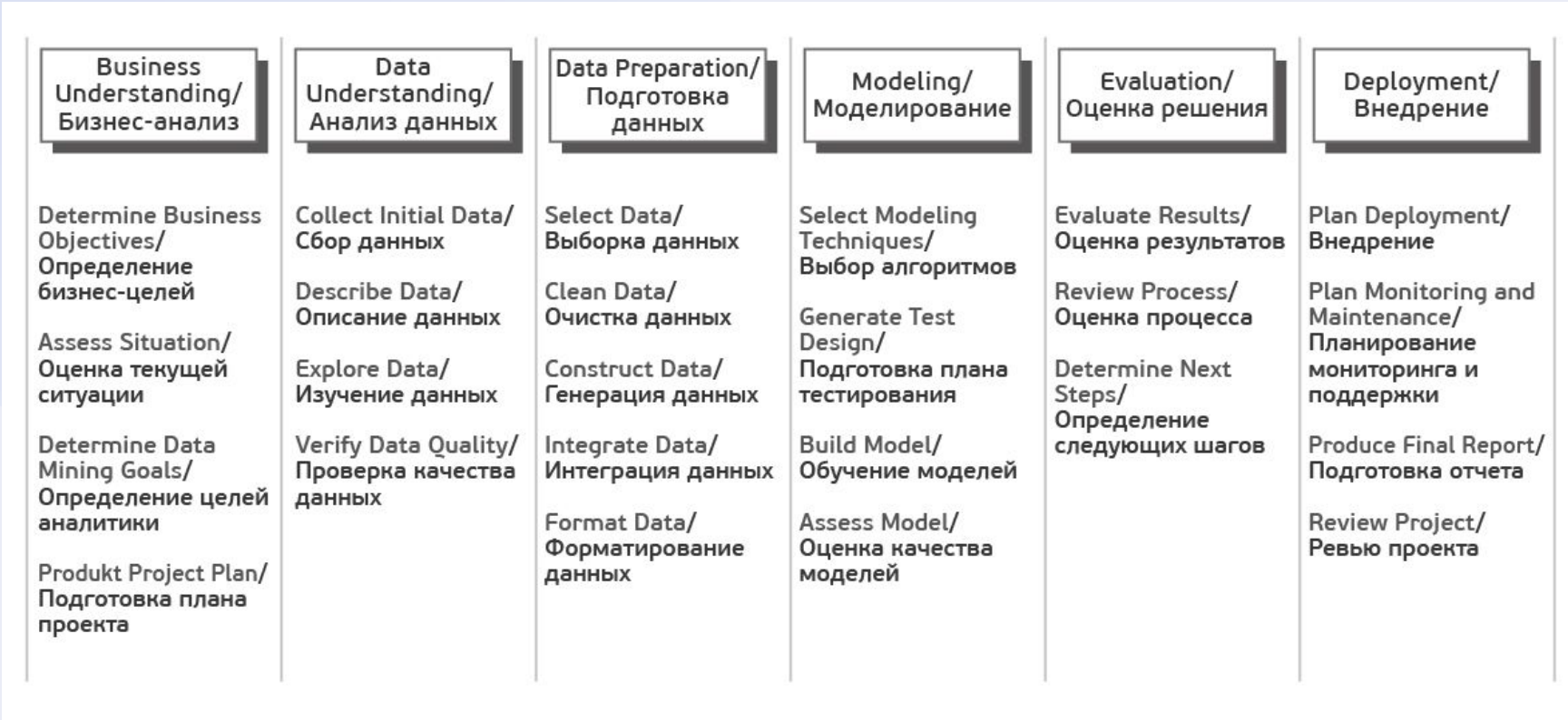
Процесс работы с данными



Crisp DM


Бизнес-анализ
Анализ данных
Подготовка данных
Моделирование
Оценка решения
Внедрение
Данные

CRISP-DM


Business Understanding/ Бизнес-анализ
Determine Business Objectives/ Определение бизнес-целей
Assess Situation/ Оценка текущей ситуации
Determine Data Mining Goals/ Определение целей аналитики
Produkt Project Plan/ Подготовка плана проекта
Data Understanding/ Анализ данных
Collect Initial Data/ Сбор данных
Describe Data/ Описание данных
Explore Data/ Изучение данных
Verify Data Quality/ Проверка качества данных
Data Preparation/ Подготовка данных
Select Data/ Выборка данных
Clean Data/ Очистка данных
Construct Data/ Генерация данных
Integrate Data/ Интеграция данных
Format Data/ Форматирование данных
Modeling/ Моделирование
Select Modeling Techniques/ Выбор алгоритмов
Generate Test Design/ Подготовка плана тестирования
Build Model/ Обучение моделей
Assess Model/ Оценка качества моделей
Evaluation/ Оценка решения
Evaluate Results/ Оценка результатов
Review Process/ Оценка процесса
Determine Next Steps/ Определение следующих шагов
Deployment/ Внедрение
Plan Deployment/ Внедрение
Plan Monitoring and Maintenance/ Планирование мониторинга и поддержки
Produce Final Report/ Подготовка отчета
Review Project/ Ревью проекта

Итоги

Итоги

1. Узнали что такое Data-Driven компании. Какие плюсы они имеют и что нужно сделать, чтобы стать полностью data-driven

2. Узнали, что на одних DS’ах далеко не уедешь. **Нужны и другие участники** аналитического процесса, которые будут помогать доставлять value заказчику.

3. Узнали способы организации команды для работы с данными, а также затронули Crisp DM - процесс организации исследований в области данных

Домашнее задание

Кейс

Возьмите за основу продукт компании “Нетология” - **профессию Data Scientist**.

- Как бы Вы организовали аналитику по данному продукту?
- Сформулируйте 2-3 ключевые метрики продукта
- Какие данные Вам могут помочь в поиске инсайтов?
- Опишите команду для работы с данными для данного продукта
    - Какие роли Вам будут нужны?
    - Чем они будут заниматься?

Спасибо за внимание