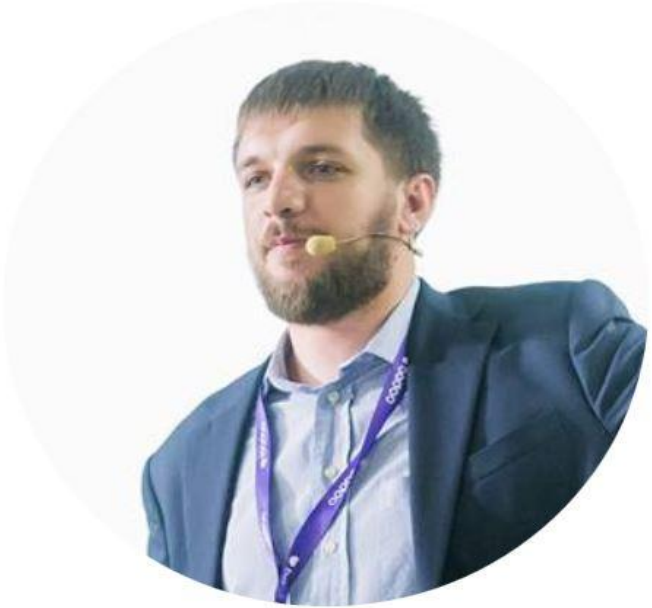
НЕТОЛОГИЯ
групп

Лекция 9

Организация команды для работы с данными



Алексей Кузьмин

Директор разработки; Data Scientist

ДомКлик.ру

aleksej.kyzmin@gmail.com

Работа с данными

Источники данных

Сбор данных
Лекция 8

SQL-БД
Лекция 1

NoSQL
Лекция 4

MapReduce
Лекция 5-7

Мотивация Big Data
Лекция 3

Data Science
Лекция 2

Отчеты
Лекция 1

Модели
Лекция 2

Люди и процессы
Лекция 9

Примеры кейсов
Лекция 10

О чём поговорим?

1. Crisp-DM
2. Организация команды

CRISP-DM

CRISP-DM

- Межотраслевой стандарт для процесса анализа данных
- Наиболее распространенная методология по исследованию данных

Основатели:

- DaimlerChrysler, SPSS и NCR

Первая версия - в 1999 году

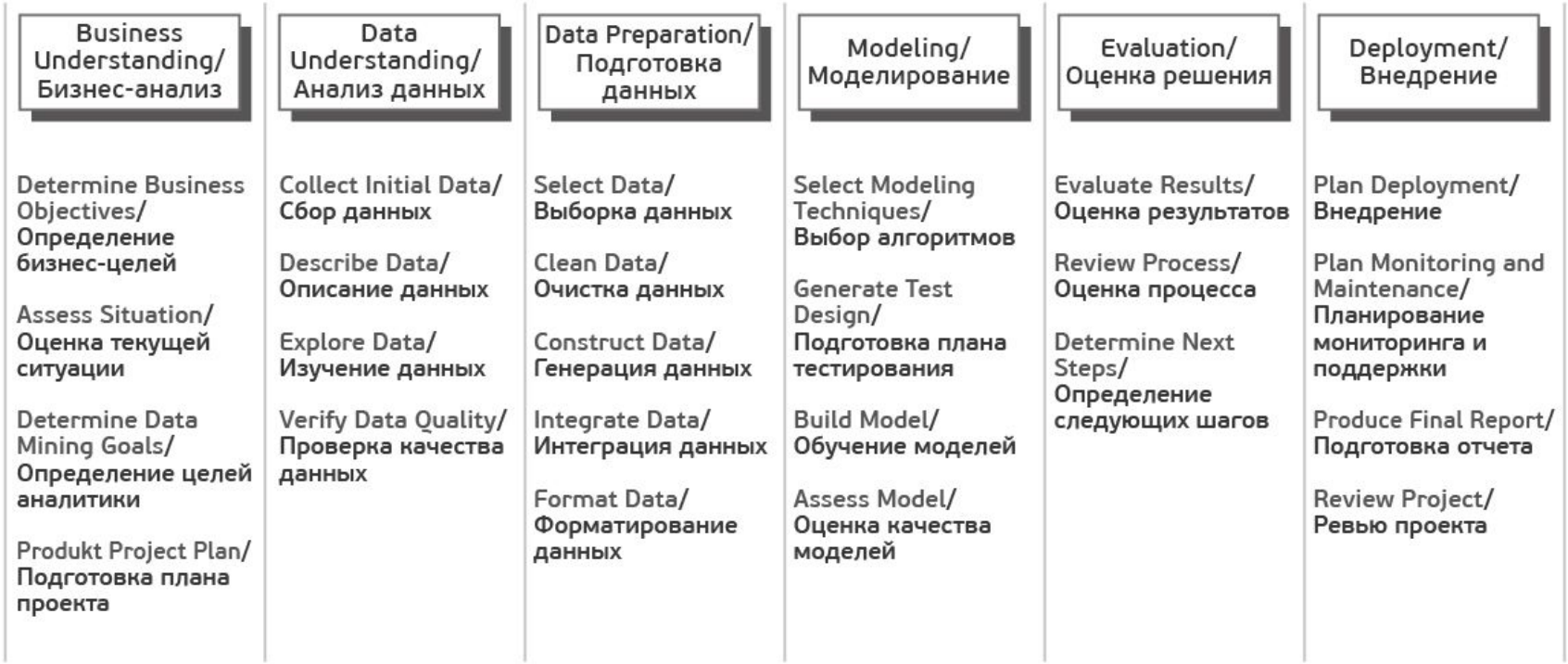
CRISP-DM

Цель:

- Разработать отраслевой, инструментальный и прикладной процесс для поиска знаний
- Определить задачи, выходы из этих задач, терминологию и характеристики майнинга

Фазы CRISP-DM

CRISP-DM


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

Business Understanding Phase

Понять бизнес-цели

- Текущий статус?
  - Понять бизнес-процессы
  - Выявить основную боль
- Определите критерии успеха
- Разработайте/изучите словарь терминов
- Проанализируйте затраты/прибыль

Business Understanding Phase

Текущая оценка систем

- Определите ключевых участников
  - Минимум: Спонсор и Ключевой пользователь
- В какой форме должен быть представлен результат?
- Интеграция результатов с существующим технологическим ландшафтом
- Изучить рыночные нормы и стандарты

Business Understanding Phase

Декомпозиция задач

- Разбейте цель на подзадачи
- Сопоставьте подзадачи с инструментарием анализа данных

Определить ограничения

- Ресурсы
- Законы, например “О персональных данных”

Составьте план проекта

- Перечислите предположения и факторы риска (технические / финансовые / бизнес / организационные)

Data Understanding Phase

Сбор данных

- Источники данных
  - Внутренние и внешние источники
  - Правила включения/исключения
  - Экспертиза в доменных знаниях и зависимость от нее
  - Проблемы доступа к данным
    - Юридические и технические
- Существуют ли проблемы с распределением данных между различными базами данных / устаревшими системами
  - Потенциальные нестыковки?

Data Understanding Phase

Описание данных

- Проблемы с качеством данных
  - требования к подготовке данных
- Вычисление базовых статистик

Data Understanding Phase

Исследование данных

- Простые одномерные графики данных / распределения
- Изучить взаимодействия атрибутов
- Проблемы качества данных
    - Пропущенные значения
        - Понять источник: пропущенные или нулевые значения
    - Странные распределения

Data Preparation Phase

Интеграция данных

- Объединение нескольких таблиц данных
- Агрегация данных

Выбор данных

- Выбор подмножества атрибутов
  - Обоснование включения / исключения
- Выборка данных
  - Наборы для обучения / проверки и тестирования

Data Preparation Phase

Преобразование данных

- Логирование
- Снижение размерности
- Нормализация / Дискретизация / бинаризация

Очистка данных

- Обработка пропущенных значений / выбросов

Построение данных (FE)

- Производные атрибуты

The Modelling Phase

Выбор модели

- Зависимости от предварительной обработки данных
  - Зависимость от атрибутов
  - Зависимость от типов данных и распределений
- Зависимости от
  - Типа проблемы анализа данных
  - Требований к выходу

The Modelling Phase

Разработать режим тестирования

- отбор тестовой выборки
- Убедитесь, что образцы имеют сходные характеристики и являются репрезентативными.

The Modelling Phase

Постройте модель

- Выберите начальные приближения параметров
- Исследуйте поведение модели
    - Анализ чувствительности к изменениям в данных (устойчивость)

The Modelling Phase

Оцените модель

- Остерегайтесь переобучения
- Исследуйте распределение ошибок
  - Определите сегменты данных, где модель менее эффективна
- Последовательно улучшайте параметры модели
  - Документируйте причины изменений

The Evaluation Phase

Свалидируйте модель

- Оценка результатов экспертами в области
- Оцените полезность результатов с точки зрения бизнеса
    - Определите контрольные группы
    - Рассчитайте метрики
    - Ожидаемая выгода (ROI - Return on Investment)

The Evaluation Phase

Определите следующие шаги

- Потенциал для експлуатация
- Архитектура внедрения
- Метрики для успеха внедрения

The Deployment Phase

Схема внедрения зависит от целей

- Презентация
- Интеграция с существующей ИТ-инфраструктурой
  - Автоматизированная предварительная обработка потоков данных в реальном времени
  - Интеграция со сторонними инструментами
- Генерация отчета
  - Online / Offline

Процесс развертывания / производства

Подготовить окончательный отчет по проекту

- Документируйте все

Команда

BI vs DS

DS - Основные задачи:

- Оптимизация, прогнозное моделирование, прогнозирование, статистический анализ
- Структурированные / неструктурированные данные, много типов источников, очень большие наборы данных

DS - основные вопросы:

- Что, если.....?
- Каков оптимальный сценарий для нашего бизнеса?
- Что будет дальше? Почему это происходит?

BI vs DS

BI - Основные задачи:

- Стандартные и специальные отчеты, информационные панели, оповещения, запросы, подробная информация по запросу
- Структурированные данные, традиционные источники

BI - основные вопросы:

- Что произошло в прошлом квартале?
- Сколько мы продали?
- В чем проблема? В каких ситуациях?

Companies Are Always Looking To Reinvent Themselves....But It’s A Mistake To Treat Data Science Teams Like Any Old Product Group.

To Build Teams That Create Great Data Products, You Have To Find People With The Skills And The Curiosity To Ask The Big Questions

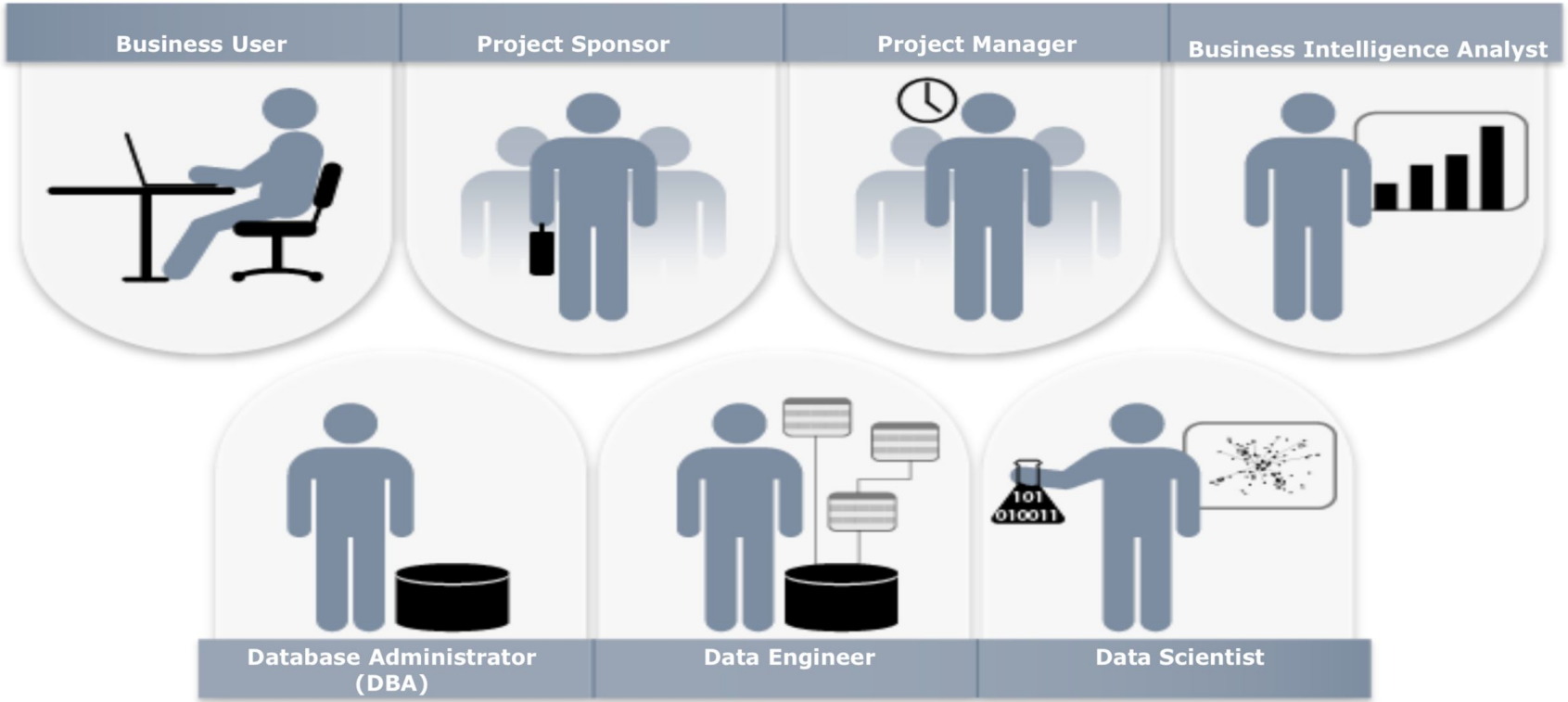
DJ Patil, Data Scientist in Residence at Greylock Partners

DATA SCIENTIST - THE SEXIEST JOB OF THE 21TH SENTURY

Thomas H. Davenport and D.J. Patil

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Основные роли



Должности и основной продукт

ETL Extract, Transform, Load - специалист — преобразование данных

Data Engineer — целостность и оптимальное хранение данных

Специалист Баз Данных — работа способность баз данных как софта

Архитектор Баз Данных/Хранилища данных — проектирование хранения данных

Аналитик — анализ метрик, экспериментов, прогнозы

Data science — продукт основанный на данных, рекомендательная система

BI-специалист — визуализация, dashbord

ETL-специлист

ETL
Extract/Transform/Load

извлечение / преобразование/ загрузка данных.

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

Data Engineer

- Вопросы оптимального и надежного хранения данных
- Обеспечения быстрого и удобного к ним доступа
- **Иногда: работы ETL**
  Extract/Transform/Load – извлечение/преобразование/загрузка

Архитектор баз данных

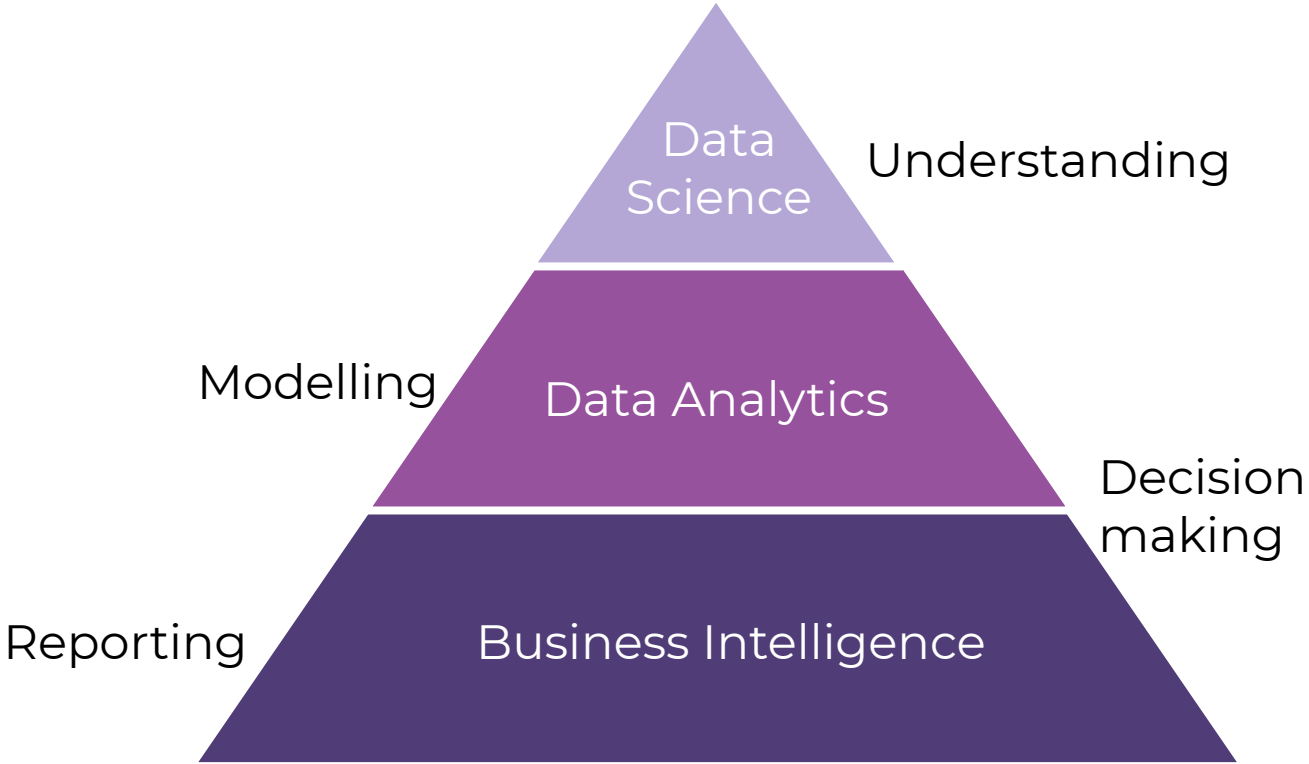
Архитектор баз данных — менеджер с глубоким понимаем БД и IT вообще

- 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

Специалист баз данных

- Проектирование БД
- Предоставление доступа
- Оптимизация работы БД
- Документация по БД
  пожалуйста, заставьте его это делать!

РАЗНЫЕ РОЛИ СПЕЦИАЛИСТОВ ПО РАБОТЕ С ДАННЫМИ



Data Scientist

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

Аналитик

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

Специалист по Business intelligence (BI)

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

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?

Где искать

- Slack, канал Open Data Science
- Соревнования Kaggle
- Собственное соревнование по Data Science:
  хакатон, олимпиада по программированию
- Конференции
- Рекомендации коллег
- Профессиональные хедхантеры

Как выбрать

- Честные вакансии
  не только творческие обязанности, но и рутина
- Акцент на практический опыт и проекты
  не только коммерческие, но и с хакатонов и конкурсов
- Интерес к причинами выбора тех или иных подходов
  и альтернативными вариантами решения задачи

Кого на самом деле не хватает на рынке?

Аналитики McKinsey еще в 2012 году предсказали громадный *дефицит специалистов* по данным, который только в США к 2018 году должен был составить от 140 до 190 тыс. человек.

Этот прогноз часто цитируют, но никто не обратил внимания на следующий абзац того же отчета, говорящий о том, что **будет не хватать 1,5 млн менеджеров, способных задавать аналитикам правильные вопросы**

Сколько платить

Data science	250 000–700 000
Аналитик	150 000–500 000
Data Engineer	200 000–400 000
BI-специалист	200 000–350 000
ETL	100 000–250 000

Сравните двух аналитиков данных

Traditional BI Analyst

ACME Healthcare

John

Sample Tasks

- Report Regional Sales For Last Quarter
- Perform Customer Feedback Surveys
- Identify Average Cost Per Supplier

Data Scientist

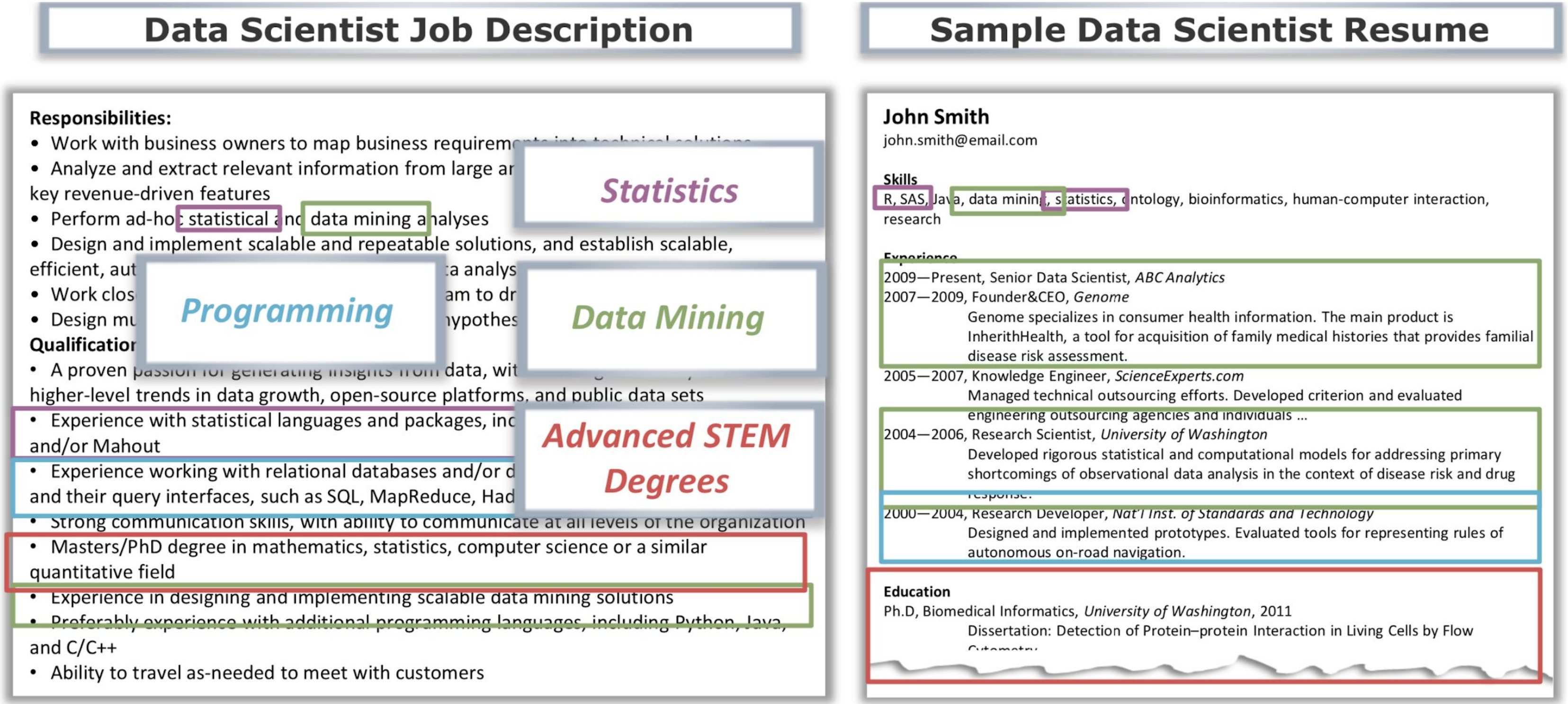
ACME Healthcare

Janet

Sample Tasks

- Predict Regional Sales For Next Quarter
- Discover Customer Opinions Via Social Media
- Identify Ways to Maximize Sales Campaign ROI

Резюме DS


Data Scientist Job Description
Sample Data Scientist Resume
Responsibilities:
• Work with business owners to map business requirem
• Analyze and extract relevant information from large ar
key revenue-driven features
• Perform ad-ho statistical and data mining analyses
• Design and implement scalable and repeatable solutions, and establish scalable,
efficient, aut
• Work clos
• Design mu
Qualification
• A proven
higher-level trends in data growth, open-source platforms, and public data sets
• Experience with statistical languages and packages, inc
and/or Mahout
• Experience working with relational databases and/or d
and their query interfaces, such as SQL, MapReduce, Had
• Strong communication skills, with ability to communicate at all levels of the organization
• Masters/PhD degree in mathematics, statistics, computer science or a similar
quantitative field
• Experience in designing and implementing scalable data mining solutions
and C/C++
• Ability to travel as-needed to meet with customers
Statistics
Programming
Data Mining
Advanced STEM
Degrees
John Smith
john.smith@email.com
Skills
R, SAS, Java, data mining, statistics, ontology, bioinformatics, human-computer interaction,
research
2009—Present, Senior Data Scientist, ABC Analytics
2007—2009, Founder&CEO, Genome
Genome specializes in consumer health information. The main product is
InherithHealth, a tool for acquisition of family medical histories that provides familial
disease risk assessment.
2005—2007, Knowledge Engineer, ScienceExperts.com
Managed technical outsourcing efforts. Developed criterion and evaluated
engineering outsourcing agencies and individuals ...
2004—2006, Research Scientist, University of Washington
Developed rigorous statistical and computational models for addressing primary
shortcomings of observational data analysis in the context of disease risk and drug
2000—2004, Research Developer, Nat'l Inst. of Standards and Technology
Designed and implemented prototypes. Evaluated tools for representing rules of
autonomous on-road navigation.
Education
Ph.D, Biomedical Informatics, University of Washington, 2011
Dissertation: Detection of Protein–protein Interaction in Living Cells by Flow

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

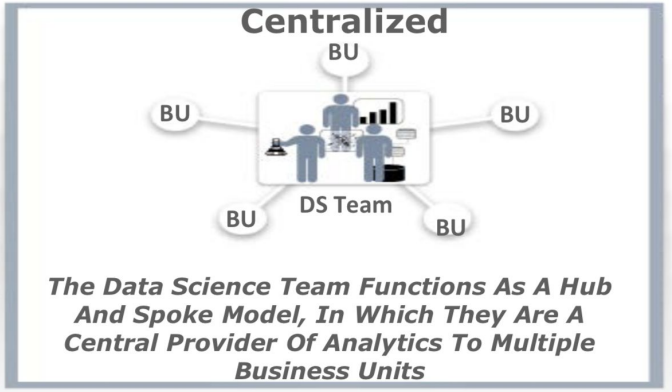
- Существует резервный план на случай «массового отказа»

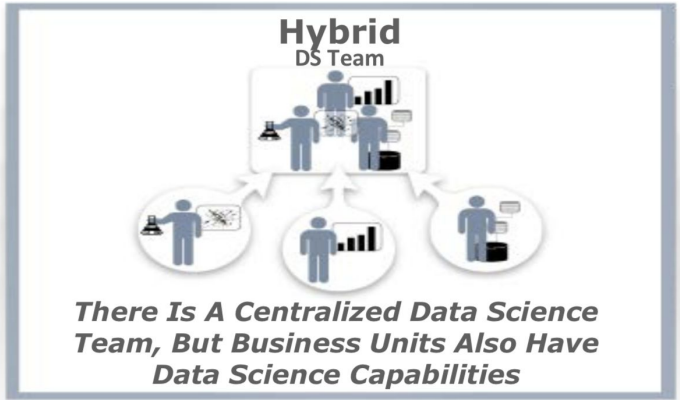
Примеры: Википедия, приз Netflix в размере 1 000 000 долларов

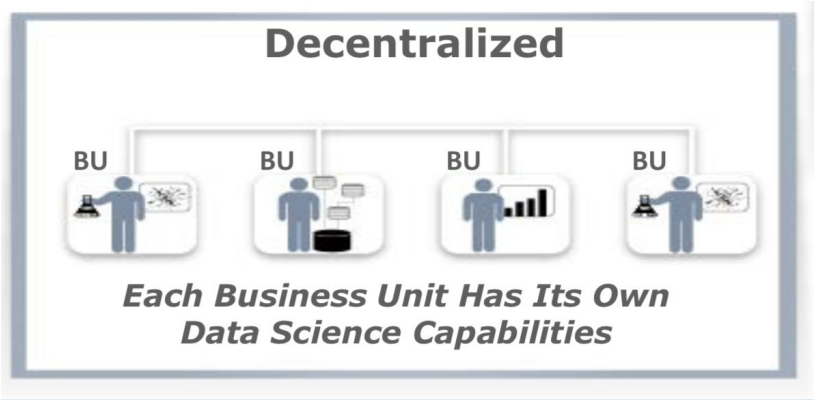
Сравнение

	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







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

Домашнее задание

Домашнее задание

Возьмите кейс из домашнего задания по разработке хранилища данных. Проработайте команду и ее организацию для работы с данными. Возьмите какую-нибудь задачу по кейсу и разложите ее по crisp-dm

Вопросы?

НЕТОЛОГИЯ
групп

Спасибо за внимание!

Алексей Кузьмин

aleksej.kyzmin@gmail.com