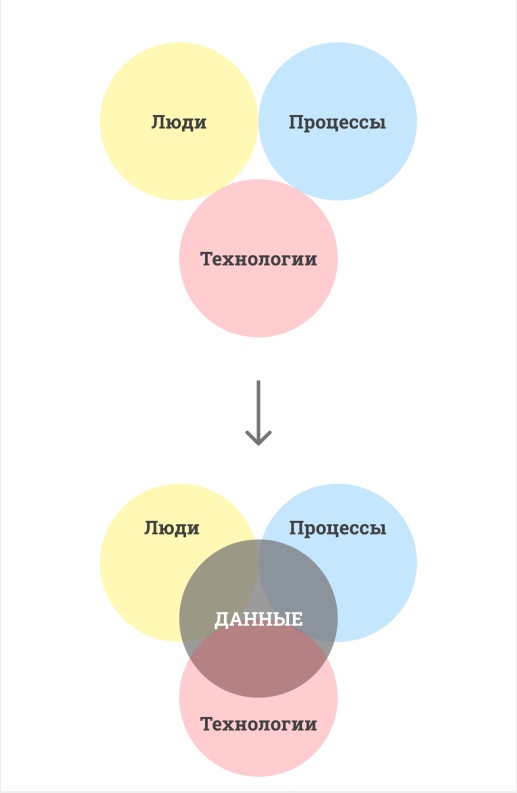
КОНСПЕКТ

1. Важные определения и термины

Данные — зафиксированная в цифровой форме информация о каких-то фактах, объектах или событиях.

Data Science (наука о данных) — это наука и практика анализа, обработки и представления данных.



Ещё 10 лет назад основными составляющими бизнеса считались три ключевых ресурса: сотрудники, процессы и технологии.

В современном бизнесе данные стали рассматриваться как **четвёртый и центральный из ресурсов**, так как они позволяют оптимизировать работу трёх остальных, например:

→ лучше привлекать новых клиентов, отслеживая по данным, какой канал привлечения клиентов наиболее выгоден;

→ улучшать и автоматизировать бизнес-процессы, например, внедряя чат-ботов, которые будут отвечать на самые распространённые вопросы клиентов;

→ принимать более обоснованные решения с учётом большого количества факторов,

→ а иногда и создавать совершенно новые бизнес-модели, например продавать данные другим компаниям как товар.

2. Направления науки о данных

Data Science можно разделить на три основных направления и соответствующих профессии:

Data Engineering **(инженерия данных)**	***Data Analytics*** **(анализ данных)**	***Machine Learning*** **(машинное обучение)**
Практика сбора, хранения и обработки данных. **Фокус:** обеспечение надёжности обработки данных. **Специалисты:** инженеры данных или дата-инженеры. Инженеры данных умеют загружать большие объёмы данных в базы, настраивать потоки данных между системами и делать так, чтобы расчёты производились быстро и с минимальными вычислительными ресурсами.	Практика исследования, использования и интерпретации данных для решения задач бизнеса, государства, науки и так далее. Это один из методов Data Science, при котором решения принимаются человеком. **Специалисты:** аналитики данных или дата-аналитики. Аналитики данных умеют выгружать нужные данные из базы, формировать из них понятные отчёты, ставить правильные вопросы и корректно отвечать на них.	Технологии создания самообучаемых алгоритмов. Это ещё один из методов *Data Science*, который используется для построения систем автоматического принятия решений, когда эти решения принимаются алгоритмами, а не людьми. В широком смысле под *Data Science* понимают все три этих направления, в узком смысле — только практику применения машинного обучения для решения практических задач.

3. Основные задачи машинного обучения

Задача	**Регрессия**	**Классификация**	**Кластеризация**
Вопрос	Сколько?	К какому классу относится объект?	Как разделить похожие объекты по группам?
Примеры	• **Прогнозирование продаж:** сколько помидоров продаст ларёк в следующем месяце? • **Предиктивное обслуживание:** через сколько сломается насос в котельной?	• **Прогнозирование оттока:** этот клиент перестанет покупать в следующем месяце или нет? • **Классификация изображений:** на картинке изображена кошка или собака? • **Фильтрация спама:** это письмо — спам или нет?	• **Сегментация клиентов:** какие группы похожих клиентов существуют? • **Детекция аномалий:** как выделить необычные (возможно, мошеннические) переводы по банковской карте?

4. Дополнительные понятия науки о данных

Искусственный интеллект — наука и технология создания интеллектуальных машин. Специалисты по работе с данными не очень любят этот термин, поскольку он часто используется в маркетинговых целях и формирует завышенные ожидания.

Big Data (большие данные) — набор технологий работы с данными большого объёма или разнообразия и/или поступающих очень быстро. Нельзя сказать, данные какого именно объёма считать большими, это зависит от ситуации. Для небольшой компании переписка 50 сотрудников в течение года — это уже большие данные. В Facebook такой же объём данных поступает за доли секунды. Понятие Big Data в большей степени относится к направлению инженерии данных.

Нейросети — это один из распространённых алгоритмов в машинном обучении, который часто используется для обработки текстов, изображений, аудио-, видео- и других сигналов. Кроме нейросетей в машинном обучении существует множество других алгоритмов.

5. Отрасли применения Data Science

Отрасль	Задачи
Банки и финансовые сервисы	Предотвращение мошенничества, кредитный скоринг, клиентская аналитика, алгоритмическая торговля, оптимизация маркетинга
Страхование	Предотвращение мошенничества, моделирование рисков, оптимизация ценообразования, клиентская аналитика, оптимизация маркетинга
Здравоохранение	Разработка лекарств, автоматизированная диагностика, рекомендации планов лечения, мониторинг состояния, оптимизация расписаний
Ритейл / Ecommerce	Прогнозирование спроса, оптимизация ассортимента, анализ покупательского поведения, рекомендации покупок, оптимизация ценообразования, оптимизация маркетинга, оптимизация логистики, распознавание событий на видео и аудио
Телекоммуникации	Персонализированные предложения, распределение ресурсов сети, предиктивное обслуживание, предотвращение мошенничества, геоаналитика, оптимизация ценообразования, мониторинг качества сервиса
Производство	Мониторинг оборудования, предиктивное обслуживание, оптимизация расписаний, оптимизация работы оборудования, распознавание дефектов, распознавание событий на видео и аудио
Рекламные сети	Таргетинг рекламы, прогнозирование отклика, оптимизация аукционов в реальном времени
Медиа- и стриминговые сервисы	Рекомендации контента и обложек, анализ отзывов, оптимизация рекламы, оптимизация расписаний показов, предиктивное моделирование для рекомендации создания контента
GameDev	Рекомендации и персонализация контента, прогнозирование времени игры и дохода от клиента, оптимизация ценообразования, предотвращение оттока
Маркетплейсы	Поисковое ранжирование, рекомендации предложений, оценка качества предложения, оценка адекватности цены, прогнозирование спроса и поставки, предотвращение мошенничества
Энергетическая сфера	Прогнозирование спроса, оптимизация ценообразования, мониторинг оборудования, предиктивное обслуживание
Спорт	Скоринг игроков, выбор оптимальных стратегий, оценка вклада отдельных игроков, оценка влияния погодных условий, прогнозирование исходов турниров
Образование	Выявление плагиата, оценка внимания, автоматизация оценки студенческих работ, удаленная аутентификация, оценка риска не закончить курс

Наука	Предсказание структуры белка, анализ генома, анализ спутниковых снимков, обработка результатов физических экспериментов, распространение информации в обществе
Социальные проекты	Распределение беженцев по местам временного пребывания, распределение воды в засушливых регионах, распределение социальной помощи (ссылка), оптимизация переобучения и трудоустройства (ссылка)

6. Уровни компетенций дата-сайентистов

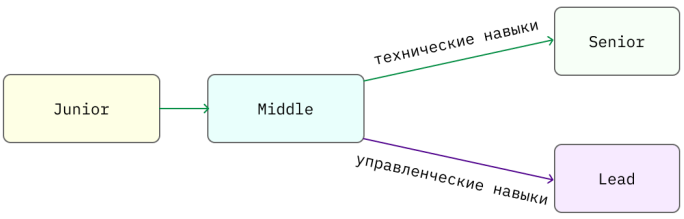
Как и во многих других IT-специальностях, в *Data Science* выделяется четыре уровня специалистов:

1. **Junior**. Специалист начального уровня, который может успешно выполнять задачи, время от времени прибегая к помощи со стороны коллег.

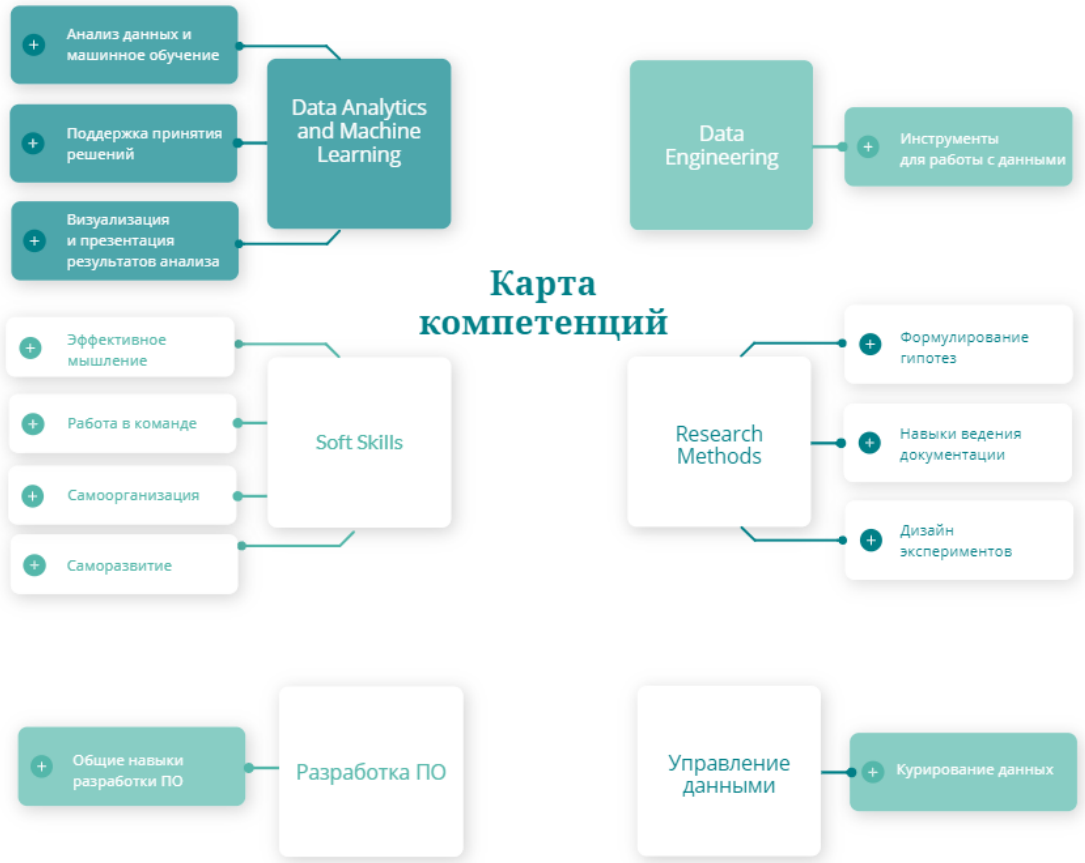
2. **Middle**. Специалист среднего уровня, который может самостоятельно реализовать поставленную задачу от начала и до конца.

3. **Senior**. Специалист с глубокими знаниями в определенной области *Data Science*, который имеет опыт разработки и внедрения целостных решений, способен предложить несколько способов выполнения задачи и знает, чем они хороши и чем плохи.

4. **Lead**. Специалист, который руководит командой *Data Science*, умеет переводить бизнес-задачи в технические, разделять задачи на составляющие и планировать работу команды, а также управлять взаимоотношениями в команде и решать проблемы взаимодействия со смежными командами.



7. Карта компетенций дата-сайентистов



→ **Анализ данных и машинное обучение**. Включает в себя умение выгружать, готовить и анализировать данные, создавать модели машинного обучения, презентовать результаты и использовать их для улучшения бизнес-процессов;

→ **Разработка ПО**. Содержит умение программировать, создавать и тестировать готовые программные продукты

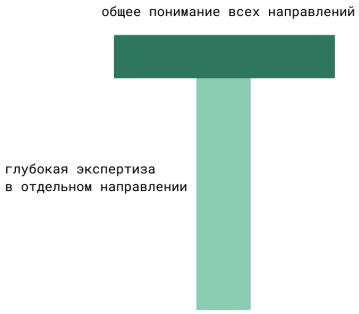
→ **Инженерия данных**. На уровне *Junior* сюда входит только умение работать и настраивать под себя инструменты для анализа данных, например *Jupyter Notebook*.

→ **Управление данными**. Здесь находится важное умение проверять данные, делать их разведывательный анализ.

→ **Методы исследований**. Включают в себя знание основ математической статистики, умение формулировать статистические гипотезы, делать корректные выборки, правильно интерпретировать и документировать результаты исследований и анализа данных.

→ **Soft Skills**. Включает в себя эффективное мышление (умение выделять главное, мыслить критически, но гибко и креативно), навыки работы в команде, самоорганизации и навыки саморазвития.

T-Shaped



В списке компетенций не перечислены конкретные алгоритмы и модели машинного обучения. Дело в том, что невозможно знать все модели очень глубоко.

Рекомендуем ориентироваться на подход под названием **T-shaped skills**. Он предполагает, что специалист очень хорошо разбирается минимум в одном из направлений Data Science (это могут быть модели классификации, регрессии, анализ временных рядов, анализ изображений и так далее), а остальные методы понимает на базовом уровне.

8. Методология CRISP-DM

Эта модель содержит все ключевые этапы работы. В большинстве проектов на этапах 1 и 6 необходимы специалисты уровня Senior или Lead.

Самые затратные по времени — этапы 2 и 3. Они могут занимать до 80 % времени всего проекта, особенно если используемые данные до этого никогда не анализировались.

CRISP-DM