АНАЛИТИКА КАК ПРОЦЕСС

- СОСТАВЛЯЮЩИЕ СОВРЕМЕННОЙ АНАЛИТИКИ
- ПРИНЦИПЫ АНАЛИЗА ДАННЫХ
  - ИЗВЛЕЧЕНИЕ И ВИЗУАЛИЗАЦИЯ ДАННЫХ
  - ПОСТРОЕНИЕ И ИСПОЛЬЗОВАНИЕ МОДЕЛЕЙ
- МЕТОДОЛОГИЯ CRISP-DM

СОСТАВЛЯЮЩИЕ СОВРЕМЕННОЙ АНАЛИТИКИ

Современная аналитика данных опирается на четыре составляющие: **эксперт**, **гипотеза**, **аналитик** и **руководитель проекта**. Дадим им определения.

СОСТАВЛЯЮЩИЕ СОВРЕМЕННОЙ АНАЛИТИКИ

Эксперт – это специалист в предметной области, профессионал, который за годы обучения и практической деятельности научился эффективно решать задачи, относящиеся к конкретной предметной области.

Эксперт является ключевой фигурой в процессе анализа. По-настоящему эффективные аналитические решения можно получить не на основе одних лишь компьютерных программ, а в результате сочетания лучшего из того, что может человек и компьютер. Эксперт выдвигает **гипотезы (предположения)** и для проверки их достоверности либо просматривает некие выборки различными способами, либо строит те или иные модели.

СОСТАВЛЯЮЩИЕ СОВРЕМЕННОЙ АНАЛИТИКИ

Гипотезой в аналитике данных часто выступает предположение о влиянии какого-либо фактора или группы факторов на результат. К примеру, во время оценки кредитоспособности потенциального заемщика можно предположить, что на его кредитоспособность влияют социально-экономические характеристики: возраст, образование, семейное положение и др.

СОСТАВЛЯЮЩИЕ СОВРЕМЕННОЙ АНАЛИТИКИ

Аналитик – это специалист в области анализа и моделирования. Аналитик на достаточном уровне владеет какими-либо инструментальными и программными средствами аналитики данных, например методами машинного обучения и Data Mining.

Аналитик играет роль «мостика» между экспертами, то есть является связующим звеном между специалистами разных уровней и областей. Он собирает у экспертов различные гипотезы, выдвигает требования к данным, проверяет гипотезы и вместе с экспертами анализирует полученные результаты. Аналитик должен обладать системными знаниями, так как помимо задач анализа на его плечи часто ложатся технические вопросы, связанные с базами данных, интеграцией с источниками данных, тестированием и производительностью. Для технических задач, связанных с обеспечением надежной инфраструктуры для данных, в последнее десятилетние появилась отдельная профессия: **инженер по данным** или **дата-инженер**.

Главным лицом в процессе анализа данных считается аналитик, т. к. он тесно сотрудничает со всеми участниками проекта. В крупных проектах участвуют, как правило, несколько экспертов, аналитиков и руководитель проекта.

СОСТАВЛЯЮЩИЕ СОВРЕМЕННОЙ АНАЛИТИКИ

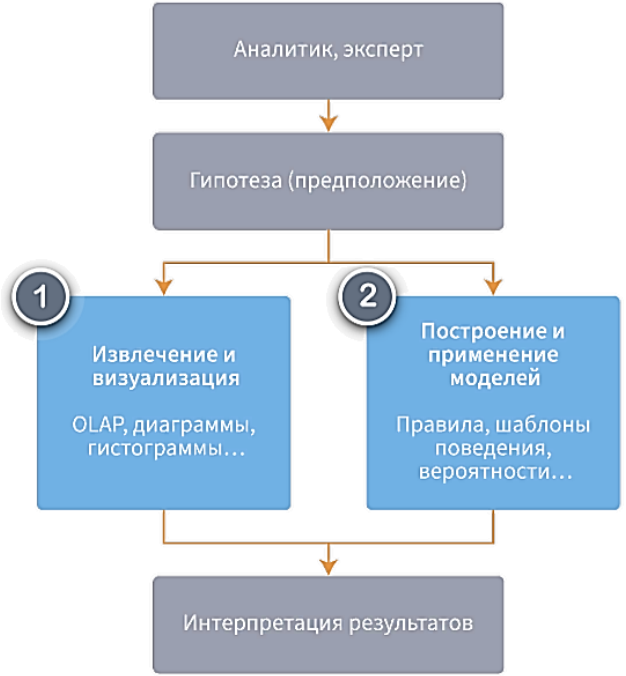
В обязанности **руководителя проекта** входят функции координации действий всех участников проекта, решение спорных вопросов, планирование и контроль сроков проекта.

ПРИНЦИПЫ АНАЛИЗА ДАННЫХ

Современная аналитика данных делит методы решения задач на две основные группы:

- *извлечение и визуализация данных*;
- *построение и использование моделей*.

Стоит отметить, что группе задач «Извлечение и визуализация данных» по сути соответствует этап ***разведочного анализа данных*** согласно концепции Дж. Тьюки, а группа «Построение и использование моделей» – это этап ***подтверждающего анализа***.



ИЗВЛЕЧЕНИЕ И ВИЗУАЛИЗАЦИЯ ДАННЫХ

Чтобы получить новые знания об исследуемом объекте или явлении, не обязательно строить сложные модели. Часто достаточно «посмотреть» на данные в нужном виде, чтобы сделать определенные выводы или выдвинуть предположение о характере зависимостей в системе, получить ответ на интересующий вопрос. Это помогает сделать **визуализация**.

В случае визуализации аналитик некоторым образом формулирует запрос к информационной системе, извлекает нужную информацию из различных источников и просматривает полученные результаты. На их основе он делает выводы, которые и являются результатом анализа.

Существует множество способов визуализации данных. Несомненными достоинствами визуализации являются относительная простота создания и введения в эксплуатацию подобных систем и возможность их применения практически в любой сфере деятельности. Кроме того, в этом случае по максимуму используются знания эксперта в предметной области и его способность принимать во внимание многие трудно формализуемые факторы, влияющие на бизнес.

ПОСТРОЕНИЕ И ИСПОЛЬЗОВАНИЕ МОДЕЛЕЙ

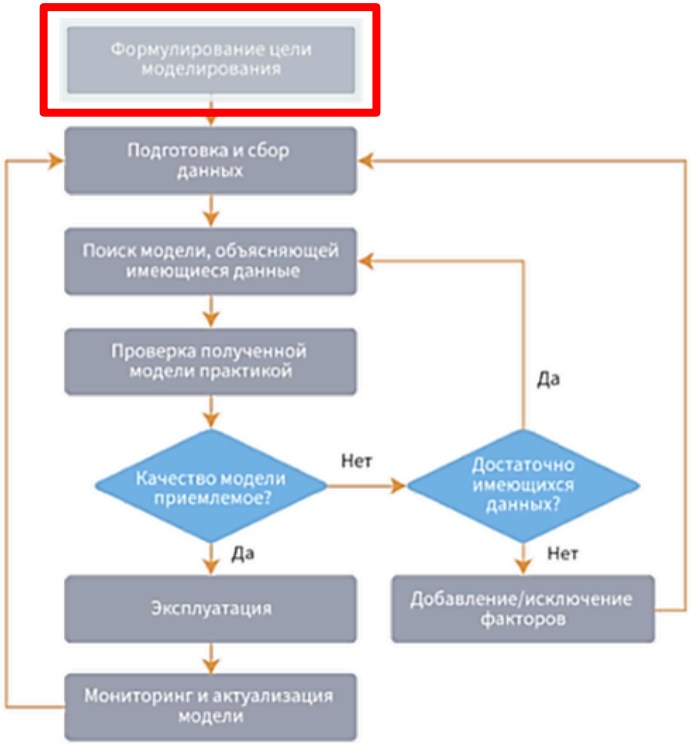
Построение моделей – это универсальный способ изучения окружающего мира, позволяющий обнаруживать зависимости, прогнозировать, разбивать на группы и решать множество других важных задач. Но самое главное: ***полученные таким образом знания можно тиражировать***. **Тиражирование знаний** – это совокупность инструментальных средств для создания моделей, которые обеспечивают конечным пользователям возможность использовать результаты моделирования для принятия решений, без необходимости понимания методик, при помощи которых эти результаты получены.

Рассмотрим подробнее элементы схемы процесса построения моделей.

ПОСТРОЕНИЕ И ИСПОЛЬЗОВАНИЕ МОДЕЛЕЙ

Формулирование цели и моделирование. При построении модели следует отталкиваться от задачи, которую можно рассматривать как получение ответа на интересующий заказчика вопрос. Например, в розничной торговле к таким вопросам относятся следующие:

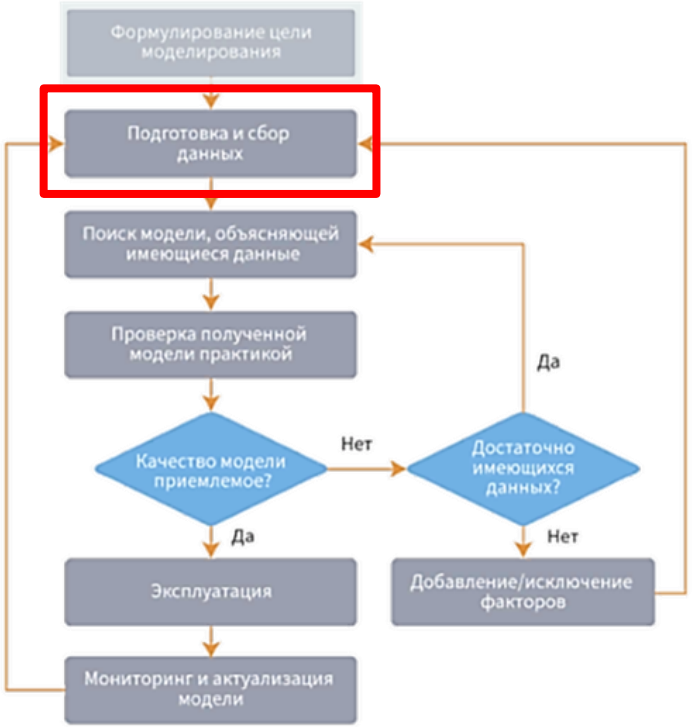
- *Сколько товара будет продано в следующем периоде?*
- *Какие клиенты откликаются на акции?*
- *Какие товары продаются или заказываются вместе?*



ПОСТРОЕНИЕ И ИСПОЛЬЗОВАНИЕ МОДЕЛЕЙ

Подготовка и сбор данных.

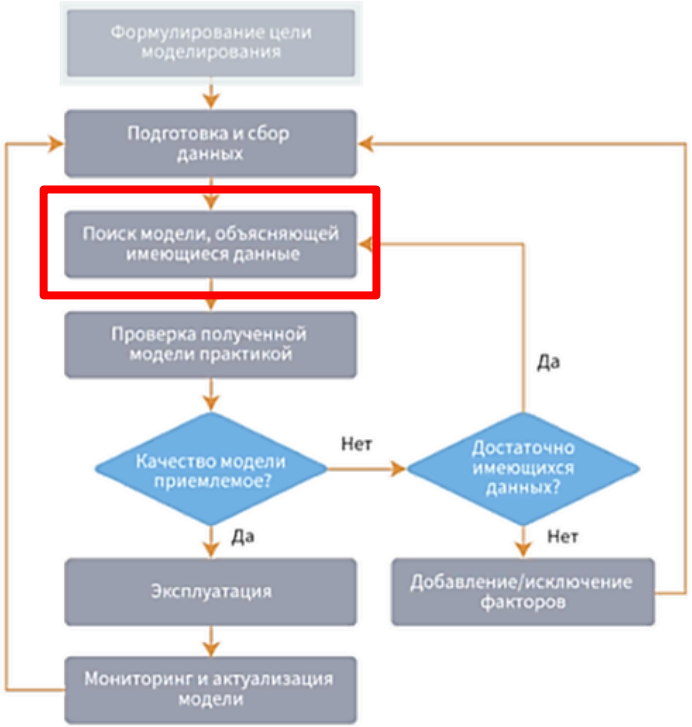
Аналитика данных опирается на использование подготовленных и систематизированных данных. Как правило, данные операции являются отдельными трудоемкими задачами.



ПОСТРОЕНИЕ И ИСПОЛЬЗОВАНИЕ МОДЕЛЕЙ

Поиск модели.

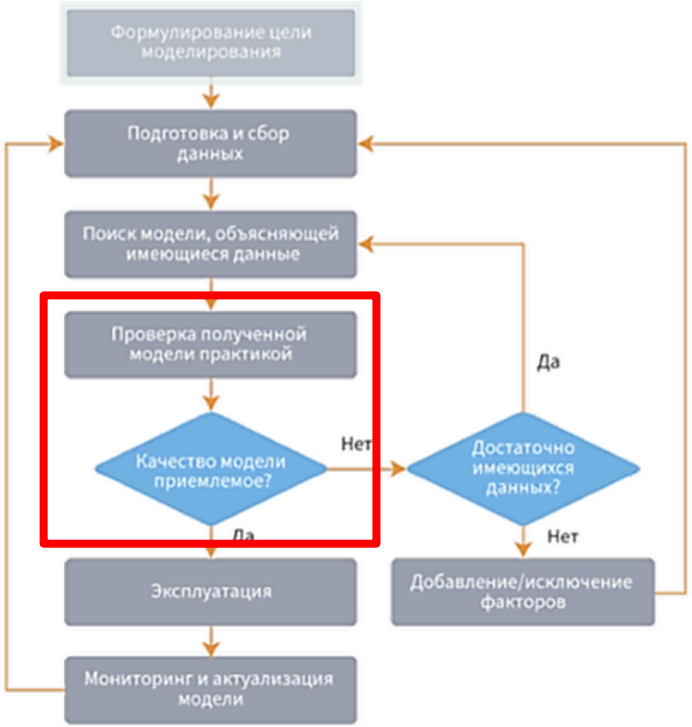
После сбора и систематизации данных переходят к поиску модели, которая объясняла бы имеющиеся данные, позволила бы добиться эмпирически обоснованных ответов на интересующие вопросы. В аналитике данных предпочтение отдается самообучающимся алгоритмам, машинному обучению, методам Data Mining, а в ряде случаев бывает достаточно статистических методов.



ПОСТРОЕНИЕ И ИСПОЛЬЗОВАНИЕ МОДЕЛЕЙ

Проверка модели.

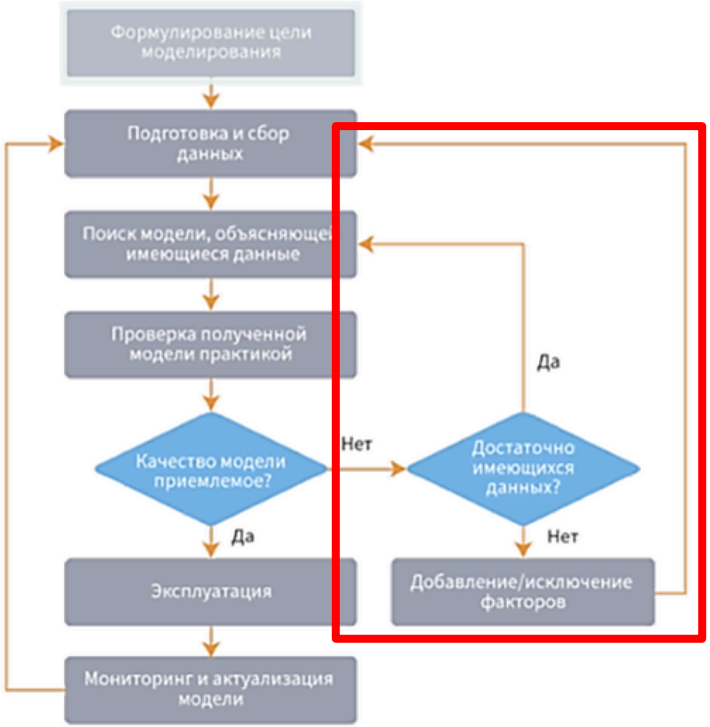
Если построенная модель показывает приемлемые результаты на практике (например, в опытной эксплуатации), ее запускают в промышленную эксплуатацию. Например, внедряют **скоринговую оценку** в принятие решений о выдаче кредита или займа. Важно, чтобы модель улучшала существующий бизнес-процесс компании. Как правило, опытная эксплуатация делается в режиме АВ-теста: одна часть объектов обрабатываются по «старой» схеме, а вторая – на основе модели. Результаты сравниваются, в том числе с применением методов статистической обработки данных эксперимента.



ПОСТРОЕНИЕ И ИСПОЛЬЗОВАНИЕ МОДЕЛЕЙ

Если качество модели неудовлетворительное, то процесс построения модели повторяется. Кроме того, любые модели подвержены «старению» и нуждаются в актуализации, поэтому можно сказать, что аналитика данных – это **непрерывный процесс**.

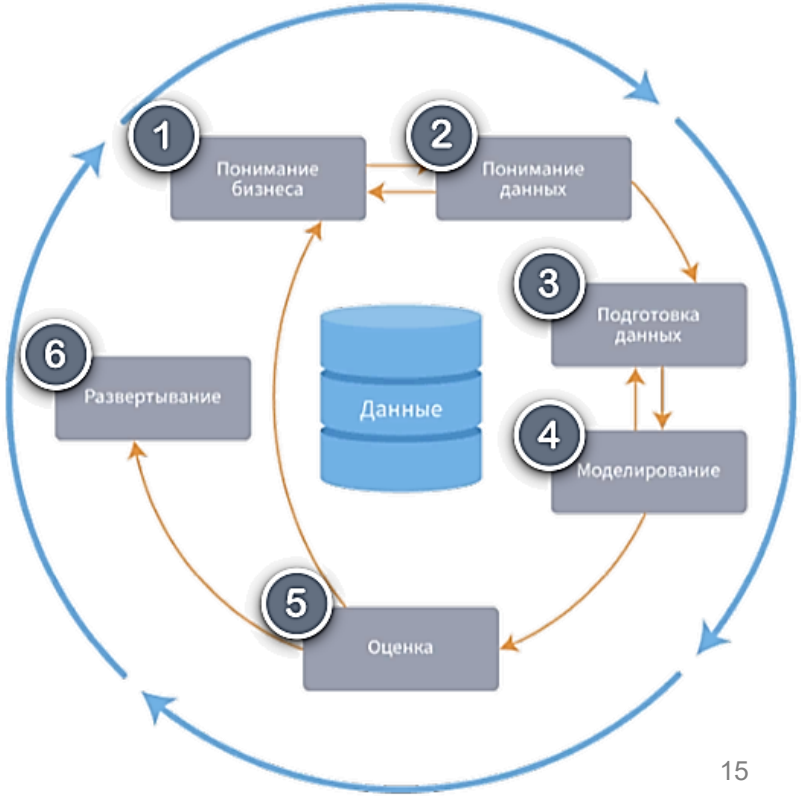
Стоит отметить, что рассмотренные подходы могут комбинироваться. Например, визуализация данных наводит аналитика на некоторые идеи, которые он пробует проверить при помощи различных моделей, а к полученным результатам снова применяются методы визуализации. Механизмы визуализации и построения моделей дополняют друг друга.



МЕТОДОЛОГИЯ CRISP-DM

Рассматривая аналитику данных как процесс, нельзя не упомянуть межотраслевой стандарт **CRISP-DM** *(англ.: CRoss Industry Standard Process for Data Mining)*. По сути это популярная методология ведения проектов в аналитике данных, особенно если в них используются многофакторные модели, построенные на принципах машинного обучения.

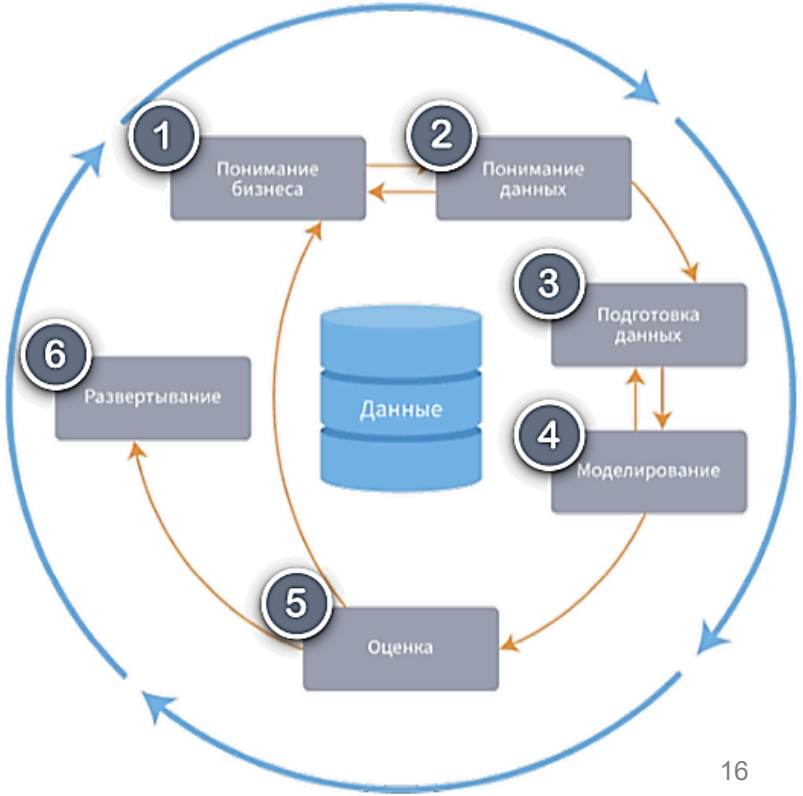
CRISP-DM был разработан в конце 1996 года четырьмя организациями: ISL (поглощена SPSS Inc.), NCR Corporation, Daimler-Benz и OHRA.



МЕТОДОЛОГИЯ CRISP-DM

На рисунке приведена **модель жизненного цикла проекта по аналитике данных CRISP-DM**. Сам проект состоит из шести этапов. Последовательность этапов не является строгой. Стрелки указывают наиболее важные и частые зависимости между этапами.

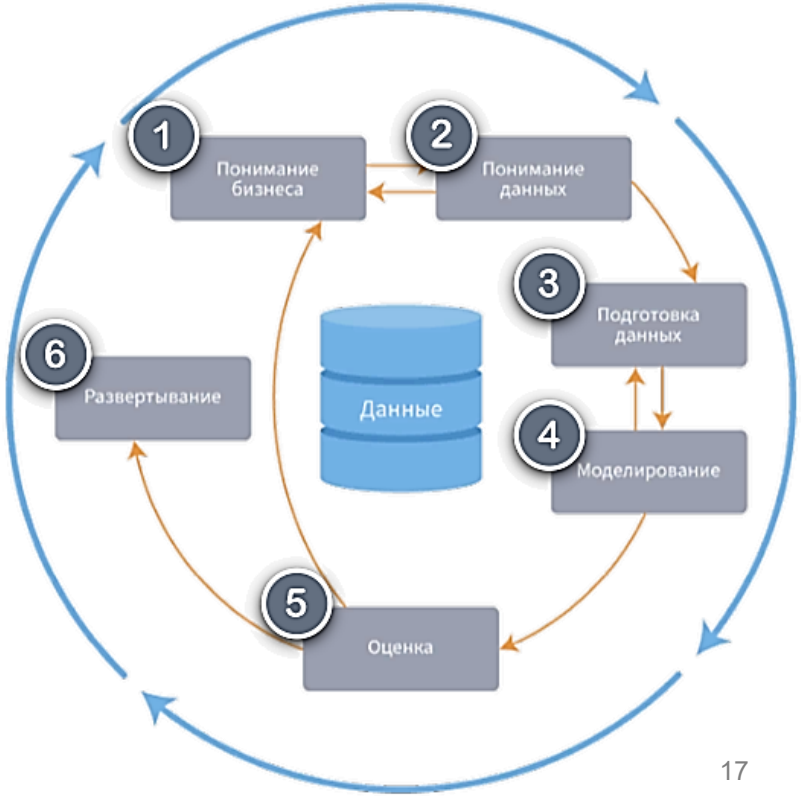
Внешний круг указывает на цикличность процесса аналитики данных, который продолжается и после развертывания проекта. Рассмотрим этапы CRISP-DM более подробно.



МЕТОДОЛОГИЯ CRISP-DM

Понимание бизнеса.

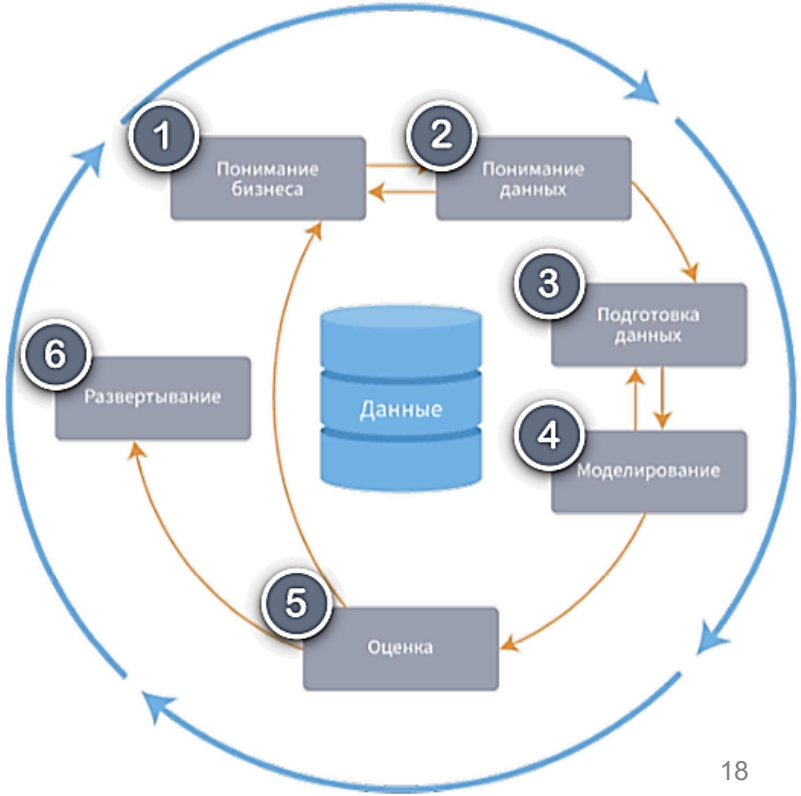
Первый этап посвящен определению целей проекта и требований к результату с точки зрения бизнеса. Далее необходимо сформулировать их на языке аналитики данных, а также разработать предварительный план проекта.



МЕТОДОЛОГИЯ CRISP-DM

Понимание данных.

Этап начинается с первоначального сбора данных, визуализации и разведочного анализа, выявления проблем с качеством данных. Цель этапа – *понять структуру данных, обнаружить интересные подмножества для формирования и последующей проверки гипотез.*

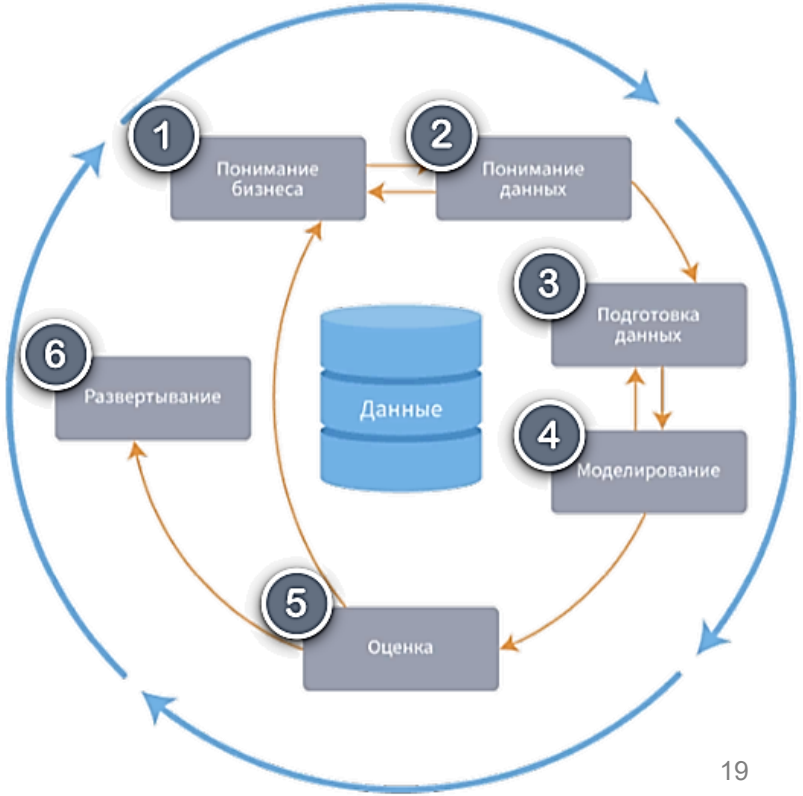


МЕТОДОЛОГИЯ CRISP-DM

Подготовка данных.

Этап ставит целью получить итоговый набор данных, которые будут использоваться при моделировании, из исходных первичных источников. Процедуры подготовки данных могут выполняться много раз. Они включают в себя:

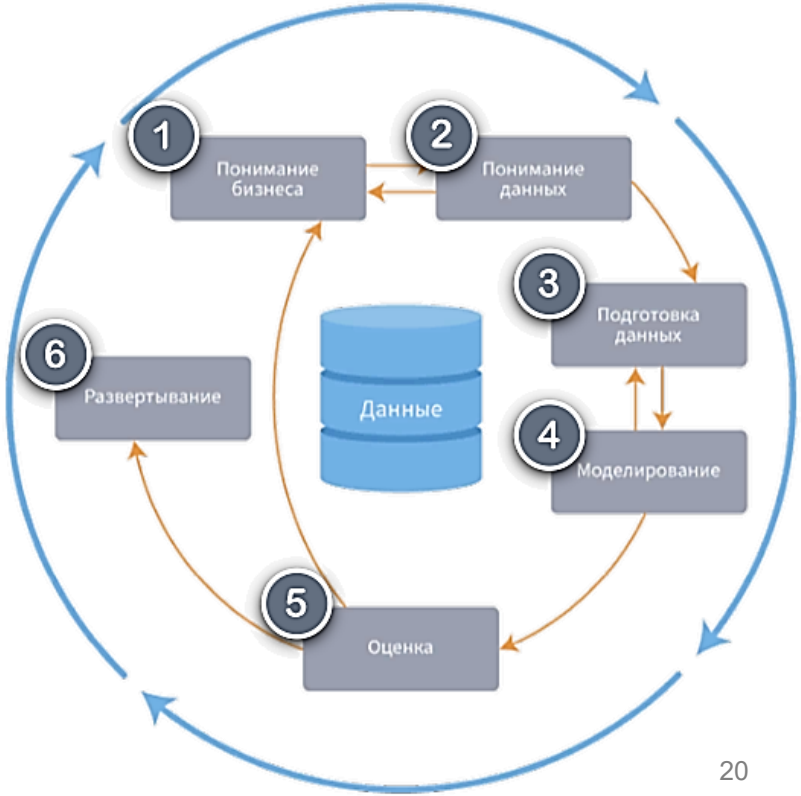
- отбор таблиц, записей и атрибутов;
- очистку данных;
- получение производных данных;
- объединение данных;
- перевод данных в нужный формат.



МЕТОДОЛОГИЯ CRISP-DM

Моделирование.

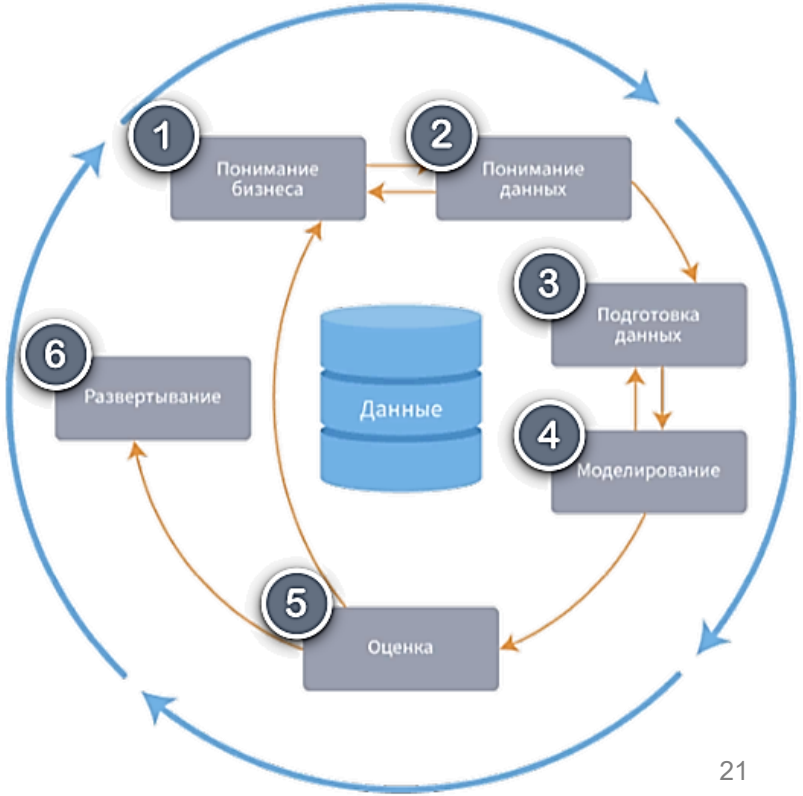
На этом этапе идет выбор методов и алгоритмов моделирования, строятся модели, а их параметры настраиваются на оптимальные значения. Как правило, для решения любой задачи анализа данных существует несколько подходов. Некоторые подходы накладывают особые требования на представление данных. Таким образом, часто бывает нужен возврат на шаг назад к фазе подготовки данных.



МЕТОДОЛОГИЯ CRISP-DM

Оценка.

На этом этапе проекта уже построена модель и получены количественные оценки ее качества. Перед тем, как внедрять эту модель, необходимо убедиться, что основная бизнес-цель проекта достигнута. Возможно, придется какие-то вопросы рассмотреть более детально. В конце этапа принимается решение по использованию результатов анализа данных на практике.

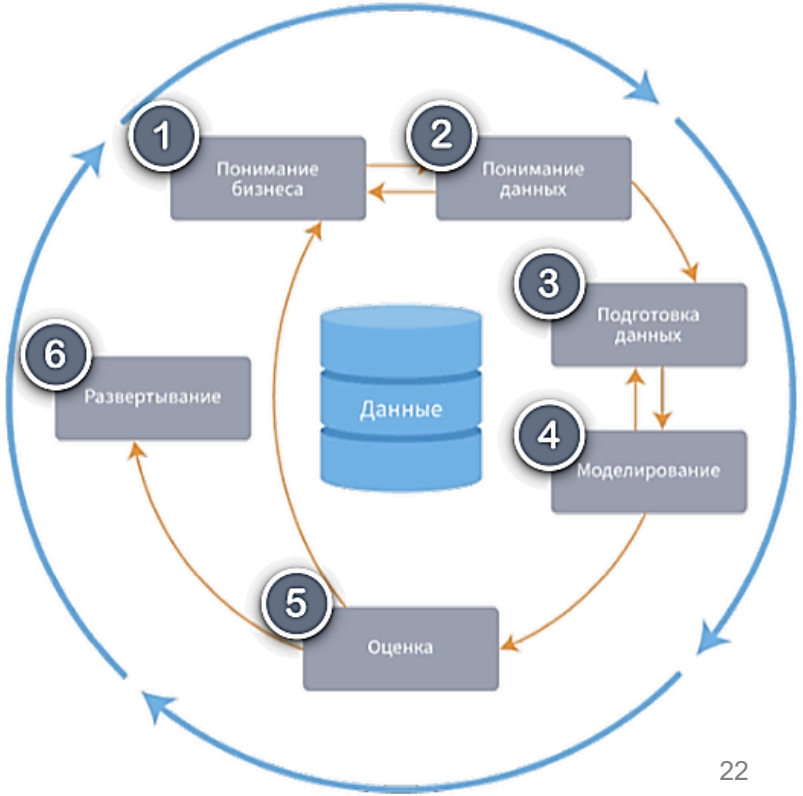


МЕТОДОЛОГИЯ CRISP-DM

Развертывание.

В зависимости от требований этот этап может быть **простым**, например, составление финального отчета, или **сложным**, например, встраивание модели в бизнес-процесс. Обычно развертывание – это забота клиента. Но важно дать понять клиенту, что ему нужно сделать для того, чтобы начать использовать полученные модели:

- запланировать развертывание;
- запланировать поддержку проекта;
- подготовить документацию;
- провести аудит проекта.



МЕТОДОЛОГИЯ CRISP-DM

Основными преимуществами методологии CRISP-DM являются:

- Модель процесса CRISP-DM универсальна и подходит для внедрения проектов по аналитике в любых отраслях.
- Нет привязки к конкретным программных продуктам или инструментам.
- Методология близка по духу к технологии извлечения знаний из баз данных – Knowledge Discovery in Databases (KDD).