

# Анализ данных

Интеллектуальный анализ данных Интеллектуальный анализ данных — это особый метод анализа данных, который фокусируется на моделировании и открытии данных, а не на их описании. Бизнес-аналитика Интеллектуальный анализ данных — это особый метод анализа данных, который фокусируется на моделировании и открытии данных, а не на их описании. Бизнес-аналитика охватывает анализ данных, который полагается на агрегацию. В

статистическом Интеллектуальный анализ данных — это особый метод анализа данных, который фокусируется на моделировании и открытии данных, а не на их описании. Бизнес-аналитика охватывает анализ данных, который полагается на агрегацию. В статистическом смысле некоторые разделяют анализ данных на описательную

статистику Интеллектуальный анализ данных — это особый метод анализа данных, который фокусируется на моделировании и открытии данных, а не на их описании. Бизнес-аналитика охватывает анализ данных, который полагается на агрегацию. В статистическом

Интеграция данных Интеграция данных это предшественник анализа данных, а сам анализ данных тесно связан с визуализацией данных Интеграция данных это предшественник анализа данных, а сам анализ данных тесно связан с визуализацией данных, и не на их распространением данных. Термин «Анализ данных» иногда используется как синоним в моделировании данных некоторые разделяют анализ данных на описательную статистику,

«[Интеллектуальный анализ данных](#)» Не следует путать с [Извлечение информации](#).

**Data Mining** ([рус.](#) добыча данных, интеллектуальный анализ данных, глубинный анализ данных) — собирательное название, используемое для обозначения совокупности методов обнаружения в [данных](#)) — собирательное название, используемое для обозначения совокупности методов обнаружения в данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности. Термин введён Григорием Пятецким-Шапиро в [1989 году](#).  
Английское словосочетание «*Data Mining*» пока не имеет устоявшегося перевода на русский язык. При передаче на русском языке используются следующие словосочетания: *просев информации, добыча данных, извлечение данных*, а также **интеллектуальный анализ данных**. Более полным и точным является словосочетание «*обнаружение знаний в базах данных*» ([англ.](#) *knowledge discovery in databases*, KDD).

Основу методов Data Mining составляют всевозможные методы классификации, моделирования и прогнозирования, основанные на применении [деревьев решений](#) Основу методов Data Mining составляют всевозможные методы классификации, моделирования и прогнозирования, основанные на применении деревьев решений, [искусственных нейронных сетей](#) Основу методов Data Mining составляют всевозможные методы классификации, моделирования и прогнозирования, основанные на применении деревьев решений, искусственных нейронных сетей, [генетических алгоритмов](#) Основу методов Data Mining составляют всевозможные методы классификации, моделирования и прогнозирования, основанные на применении деревьев решений, искусственных нейронных сетей, генетических алгоритмов, [эволюционного программирования](#) Основу методов Data Mining составляют всевозможные методы классификации, моделирования и прогнозирования, основанные на применении деревьев решений, искусственных нейронных сетей, генетических алгоритмов, эволюционного программирования, [ассоциативной памяти](#) Основу методов Data Mining составляют всевозможные методы классификации, моделирования и прогнозирования, основанные на применении деревьев решений, искусственных нейронных сетей, генетических алгоритмов, эволюционного программирования, ассоциативной памяти, [нечёткой логики](#). К методам Data Mining нередко относят *статистические методы* ([дескриптивный анализ](#) (дескриптивный анализ, [корреляционный](#) (дескриптивный анализ, корреляционный и [регрессионный анализ](#) (дескриптивный анализ, корреляционный и регрессионный анализ,

Методы Data Mining (или, что то же самое, Knowledge Discovery In Data, сокращённо, KDD) лежат на стыке [баз данных](#) Методы Data Mining (или, что то же самое, Knowledge Discovery In Data, сокращённо, KDD) лежат на стыке баз данных, [статистики](#) Методы Data Mining (или, что то же самое, Knowledge Discovery In Data, сокращённо, KDD) лежат на стыке баз данных, статистики и [искусственного интеллекта](#).

### **Исторический экскурс**

Область Data Mining началась с семинара (англ. workshop), проведённого Григорием Пятецким-Шапиро в 1989 году.

Ранее, работая в компании GTE Labs, [Григорий Пятецкий-Шапиро](#) заинтересовался вопросом: можно ли автоматически находить определённые правила, чтобы ускорить некоторые запросы к крупным базам данных. Тогда же было предложено два термина — Data Mining («добыча данных») и Knowledge Discovery In Data (который следует переводить как «открытие знаний в базах данных»).

В 1993 году вышла первая рассылка «Knowledge Discovery Nuggets», а в 1994 году был создан один из первых сайтов по Data Mining.

### **Постановка задачи**

Первоначально задача ставится следующим образом:

имеется достаточно крупная база данных;

предполагается, что в базе данных находятся некие «скрытые знания».

Необходимо разработать методы обнаружения знаний, скрытых в больших объёмах исходных «сырых» данных. В текущих условиях глобальной конкуренции именно найденные закономерности (знания) могут быть источником дополнительного конкурентного преимущества.

Что означает «скрытые знания»? Это должны быть обязательно знания:

ранее неизвестные — то есть такие знания, которые должны быть новыми (а не подтверждающими какие-то ранее полученные сведения);

нетривиальные — то есть такие, которые нельзя просто так увидеть (при непосредственном визуальном анализе данных или при вычислении простых статистических характеристик);

практически полезные — то есть такие знания, которые представляют ценность для исследователя или потребителя;

доступные для интерпретации — то есть такие знания, которые легко представить в наглядной для пользователя форме и легко объяснить в терминах предметной области.

Эти требования во многом определяют суть методов Data mining и то, в каком виде и в каком соотношении в технологии Data mining используются системы управления базами данных, статистические методы анализа и методы искусственного интеллекта.

## **Data mining и базы данных**

Методы Data mining могут быть применены как для работы с [большими данными](#), так и для обработки сравнительно малых объемов данных (полученных, например, по результатам отдельных экспериментов, либо при анализе данных о деятельности компании). В качестве критерия достаточного количества данных рассматривается как область исследования, так и применяемый алгоритм анализа.

Развитие технологий баз данных сначала привело к созданию специализированного языка — языка запросов к базам данных. Для реляционных баз данных — это язык [SQL](#). Развитие технологий баз данных сначала привело к созданию специализированного языка — языка запросов к базам данных. Для реляционных баз данных — это язык SQL, который предоставил широкие возможности для создания, изменения и извлечения хранимых данных. Затем возникла необходимость в получении аналитической информации (например, информации о деятельности предприятия за определённый период), и тут оказалось, что традиционные реляционные базы данных, хорошо приспособленные, например, для ведения оперативного учёта на предприятии, плохо приспособлены для проведения анализа. Это привело, в свою очередь, к созданию т. н. «[хранилищ данных](#)», сама структура которых наилучшим способом соответствует проведению всестороннего математического анализа.

## **Data mining и искусственный интеллект**

Знания, добываемые методами Data mining, принято представлять в виде *закономерностей (паттернов)*. В качестве таких выступают:

- ассоциативные правила;
- деревья решений;
- кластеры;
- математические функции.

Алгоритмы поиска таких закономерностей находятся на пересечении областей: Искусственный интеллект, Математическая статистика, Математическое программирование, Визуализация, OLAP.

Задачи, решаемые методами Data Mining, принято разделять на описательные ([англ. descriptive](#)) и предсказательные ([англ. predictive](#)).

В описательных задачах — это дать наглядное описание имеющихся скрытых закономерностей, в то время как в предсказательных задачах на первом плане стоит вопрос о предсказании для тех случаев, для которых данных ещё нет.

К описательным задачам относятся:

поиск ассоциативных правил или паттернов (образцов);  
группировка объектов, кластерный анализ;  
построение регрессионной модели.

К предсказательным задачам относятся:

классификация объектов (для заранее заданных классов);  
[регрессионный анализ](#) регрессионный анализ, анализ [временных рядов](#).

## Алгоритмы обучения

Для задач классификации характерно «обучение с учителем», при котором построение модели производится по выборке, содержащей входные и выходные векторы.

Для задач кластеризации и ассоциации применяется «обучение без учителя», при котором построение модели производится по выборке, в которой нет выходного параметра. Значение выходного параметра («относится к кластеру ...», «похож на вектор ...») подбирается автоматически в процессе обучения.

Для задач сокращения описания характерно *отсутствие разделения на входные и выходные векторы.*

## Ряд этапов решения задач методами Data Mining:

1. Постановка задачи анализа
2. Сбор данных
3. Подготовка данных (фильтрация, дополнение, кодирование)
4. Выбор модели или алгоритма анализа данных
5. Подбор параметров модели и алгоритма обучения
6. Обучение модели или автоматический поиск остальных параметров модели



**Топологический анализ данных** — новая область теоретических исследований для задач [анализа данных](#) — новая область теоретических исследований для задач анализа данных (Data mining) и [компьютерного зрения](#).

### **Основные вопросы:**

Как из низкоразмерных представлений получать структуры высоких размерностей?

Как дискретные единицы складываются в глобальные структуры?

Человеческий мозг легко строит представление об общей структуре по частным данным низких размерностей.

Ему не составляет труда получить трехмерную форму объекта по плоским изображениям в каждом глазу.

Создание общей структуры также производится при объединении [дискретных](#)

Создание общей структуры также производится при объединении дискретных во времени фрагментов в [непрерывный](#) образ. Так, например,

телевизионное изображение технически является массивом

отдельных точек воспринимается как единая сцена

- **В метод топологического анализа данных входят:**
- Замена набора элементов данных некоторым семейством КОМПЛЕКСОВ в соответствии с параметром близости.
- Анализ топологических комплексов с помощью алгебраической топологии, а конкретно новой теорией **устойчивых гомологий**.
- Перекодировка устойчивой гомологии набора данных в параметризованную версию чисел Бетти называемую **ШТРИХКОДОМ**.

## Облако точек

Данные часто представлены множеством точек в Евклидовом пространстве, форма которого отражает описываемый данными феномен.

Реальные трехмерные объекты могут представляться в виде **облака точек**. Лазером отмечают отдельные точки и их неструктурированный набор служит представлением объекта в компьютере. **Облаком точек** считается любой набор точек или проекций точек в более низкой размерности.

Разведочный анализ данных (РАД; Exploratory data analysis) употребляется, когда, с одной стороны, у исследователя имеется таблица многомерных данных, а с другой стороны, априорная информация о физическом (причинном) механизме генерации этих данных отсутствует или неполна. В этой ситуации РАД может оказать помощь в *компактном* и *понятном* исследователю описании структуры данных (например, в форме визуального представления этой структуры), отталкиваясь от которого он уже может «прицельно» поставить вопрос о более детальном исследовании данных с помощью того или иного раздела статистического анализа, обоснования полученной структуры данных с помощью аппарата проверки статистических гипотез, а также, возможно, сделать некоторые заключения и о причинной модели данных. Этот этап называется «подтверждающим анализом данных» (confirmatory data analysis). Иногда выявление структуры данных с помощью РАД может оказаться и завершающим этапом анализа. С другой стороны, ряд методов РАД можно рассматривать и как методы подготовки данных для последующей статистической обработки

**Школа анализа данных (ШАД)** — бесплатные двухгодичные очные вечерние курсы от компании «[Яндекс](#)» — бесплатные двухгодичные очные вечерние курсы от компании «Яндекс», открытые в [2007 году](#) с целью подготовки кадров в области обработки и анализа данных и применения информации в

Интернет-анализа данных, [компьютерных наук](#) Есть три отделения: анализа данных, компьютерных наук, и [больших данных](#) Есть три отделения: анализа данных, компьютерных наук, и больших данных; отделение [биоинформатики](#) является самостоятельной академической структурой.

Поступление на первые три отделения состоит из прохождения интерактивного теста, письменного экзамена и очного собеседования.

Ежегодно школа выпускает 81 человека по специальности «компьютерная наука».

Школа имеет филиалы в [Санкт-Петербурге](#) Школа

Среди преподавателей — российские и  
зарубежные специалисты:

[Борис Теодорович Поляк](#)

[Андрей Михайлович Райгородский](#)

[Алексей Яковлевич Червоненкис](#)

[Альберт Николаевич Ширяев](#)

# Анализ социологических данных

Основная цель анализа данных в социологии — выявление, подтверждение, корректировка статистических закономерностей.

В методологии анализа данных следует выделить следующие взаимосвязанные части:

Типы данных (данные, полученные посредством вопросников простой и сложной структуры; об использовании бюджета времени, текстовые данные разного вида).

*Приемы, подходы к сбору данных, к измерению*

(одномерное и многомерное шкалирование; формирование индексов; ранжирование).

*Восходящая стратегия анализа данных.* Логика Логика и

методы Логика и методы проверки описательных гипотез.

Поиск эмпирических закономерностей.

*Нисходящая стратегия анализа данных.*

Типологический анализ, факторный анализ, причинный

анализ данных

Понятие «анализ» на различных этапах исследования трактуется по-разному. Упрощенная схема социологического исследования, опирающегося на эмпирические данные.

Она состоит из трех элементов:

Концептуальная схема исследования (предмет, объект, цели, задачи, гипотезы исследования, понятийный аппарат исследования).

Методика сбора эмпирических данных (понятия и инструментарий исследования).

Методика обработки данных (формы представления



- На всех этих трех уровнях понятие «**анализ**» имеет различную трактовку.
- На последнем уровне анализ рассматривается как статистическая обработка информации, применение математического метода, вычисление индекса обобщенного показателя, полученного посредством использования логических операций, например, конъюнкция и дизъюнкция) и т. д.

- Под анализом могут пониматься различные логические схемы: логика решения задач разного класса, логика интерпретации эмпирических закономерностей.
- В целом любое социологическое исследование есть анализ фрагмента социальной реальности.

# Виды анализа по объектам управления

- **Функциональный анализ**
- Его объектом являются функции потребительных стоимостей, т.е. продуктов конкретного труда.
- **Технический анализ**
- Его предметом выступают причинно-следственные связи натуральных процессов деятельности, обеспечивающие формирование продуктов конкретного труда с заданными потребительскими свойствами (функциями).

- **Экономический анализ**
- Важным объектом управления и, следовательно, анализа как управляющей функции являются экономические процессы, которые в узком смысле слова выражают индивидуальные и общественно-необходимые затраты труда на создание потребительной стоимости в денежной форме или в показателях рабочего времени.

- **Социальный анализ**

- Сложным важным объектом управления и анализа являются социальные процессы, в которых выражается многогранность социальной сферы хозяйственной деятельности. К ним относятся: создание нормальных, отвечающих требованиям охраны здоровья трудящихся условий труда по чистоте воздуха, освещенности, температуре, шуму, вибрации и другим производственным факторам; обеспечение соответствующих социально-психологических и психофизиологических условий труда, вопросы адаптации вновь поступающих на работу; улучшение санитарно-бытовых условий на производстве и вне его, включая задачи лечебного, профилактического и оздоровительного характера; обеспечение необходимыми жилищно-бытовыми условиями, дошкольными детскими учреждениями; развертывание культурно-массовой и спортивно-массовой работы; развитие подсобного сельского хозяйства для улучшения питания работающих.

- **Экологический анализ (ЭКА)**
- Объектом ЭКА являются экологические процессы – взаимоотношения природы и общества, а его предметом – причинно-следственные связи во взаимоотношениях природы и общества, изменяющие их в лучшую или худшую сторону относительно жизни человека.

## **Виды анализа по взаимосвязанным объектам управления**

- **Функционально-экономический анализ**
- Объектом его выступают функции или свойства изделий и процессов, т.е. потребительная стоимость (ПС) и затраты живого и овеществленного труда (стоимость) на создание этих функций, а непосредственным предметом – причинно-следственные связи между потребительной стоимостью и стоимостью конкретных

# Технико-экономический анализ (ТЭА)

- Его объектом служат технические (натуральные) процессы создания потребительных стоимостей с заданными функциями и связанные с этими процессами затраты живого и овеществленного труда, а непосредственным его предметом – причинно-следственные связи технико-экономических процессов, формирующих соответственные результаты.



# Социально-экономический анализ (СЭА)

- Его объектом являются социальные процессы хозяйственной деятельности и связанные с ними затраты и экономия живого и овеществленного труда, а непосредственным предметом – причинно-следственные связи, определяющие результаты социально-экономического развития трудового коллектива.

# Экономико-экологический анализ (ЭЭКА)

- Объект ЭЭКА – экологические и экономические процессы, связанные с сохранением или улучшением взаимоотношений природы и общества с затратами труда на улучшение или сохранение баланса отношений человека и природы.
- Предметом ЭЭКА являются причинно-следственные связи, определяющие результаты взаимодействия экономических и экологических процессов и изменения результатов за рассматриваемый период.
- Цель ЭЭКА – сохранение нормального состояния взаимоотношений природы и человека или его улучшение с минимальными затратами материальных и трудовых ресурсов (в денежной форме).

# Маркетинговый анализ

- применяется для изучения внешней среды функционирования предприятия, рынков сырья и сбыта готовой продукции, ее конкурентоспособности, спроса и предложения, коммерческого риска, формирования ценовой политики, разработки тактики и стратегии маркетинговой деятельности.

# Вопросы для повторения

1. Как рассматривается понятие **«анализ»**?
2. Чем отличается Data Mining от анализа?
3. Какие существуют задачи, решаемые методами Data Mining и как они подразделяются?
4. Какие виды Data Mining Вам знакомы?
5. Как различаются виды анализа по объекту управления?
6. Какое сходство или различие по видам анализа наблюдается по