КОНСПЕКТЫ ЛЕКЦИЙ СО СЛАЙДОВ

# ЛЕКЦИЯ № 1

# Технологии интеллектуального анализа данных

*Если достаточно долго мучить данные, они признаются [в чем угодно]. Рональд Коуз нобелевский лауреат по экономике*

*Самая большая проблема, с которой сталкиваются компании, пытающиеся внедрять инновации и трансформироваться, — корпоративная культура по типу «мы всегда так делали».*

*Анализ — преобразование данных в выводы, на основе которых будут приниматься решения и строиться действия с помощью людей, процессов и технологий.*

Информационная деятельность – деятельность человека, связанная с процессами:

* Получения информации;
* Преобразования информации;
* Накопления информации;
* Передачи информации;

Информационная технология (ИТ) – процесс, использующий совокупность средств и методов сбора, обработки и передачи данных (первичной информации) для получения информации нового качества о состоянии объекта или явления.

ИТ характеризуется на каждом этапе инструментарием, видами коммуникаций и целями.

## Различные классификации информационных технологий

1. «Ручная» (до второй половины XIXв.)
2. «Механическая» (с конца XIXв.)
3. «Электрическая» (40 – 60-е гг. XXв.)
4. «Электронная» (с начала 70-х гг. XXв.)
5. «Компьютерная» (с середины 60-х гг. XXв.)

## Носители информации

* Внемашинные
* Машиночитаемые
* Электронные

## Способы обработки информации

* Неавтоматизированные
* Автоматизированные
* Автоматические

## Информация, данные, знания

Изображение выглядит как текст, снимок экрана, Шрифт, линия

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, доска

Автоматически созданное описание

## Классификация информационных технологий

* По виду представления обрабатываемой информации
* В зависимости от класса решаемых задач
* По степени охвата задач управления

## Обработка информации

Обработка информации состоит в получении одних «информационных объектов» из других «информационных объектов» путем выполнения некоторых алгоритмов и является одной из основных операций, осуществляемых над информацией, и главным средством увеличения ее объема и разнообразия.

## Технология обработки информации

Взаимосвязанные действия, выполняемые в строго определённой последовательности с момента возникновения информации до получения заданных результатов.

## Виды обработки информации с точки зрения реализации технических СОД:

* Последовательные
* Параллельные
* Конвейерные

## Основные процедуры обработки данных

К процедурам обработки данных относятся: создание данных, модификация данных, поиск информации, принятие решений, создание отчетов, создание документов, повышение безопасности данных.

При обработке данных обращают внимание на их качество. Выделяют чистые и грязные данные. Грязные данные отличает наличие обработки, дополнительной, не связанных с первоначальными данными, информации, недостаток первичных данных.

Цели:

1. Сбор информации (различные формы)
2. Определение существенной информации (на данный момент)
3. Предоставление существенной информации (в нужном виде)

Задачи:

* Сбор данных (Оценка качества данных)
* Ввод данных в ИС (Ручной, Автоматический, Контроль и исправление ошибок ввода)
* Накопление данных
* Обеспечение доступа к данным (Поиск, Контроль и защита)
* Передача и обмен данными (Интеграция, Обеспечение надежности)
* Представление данных (Визуализация (текст, графика, таблицы и тд), Форматы представления)

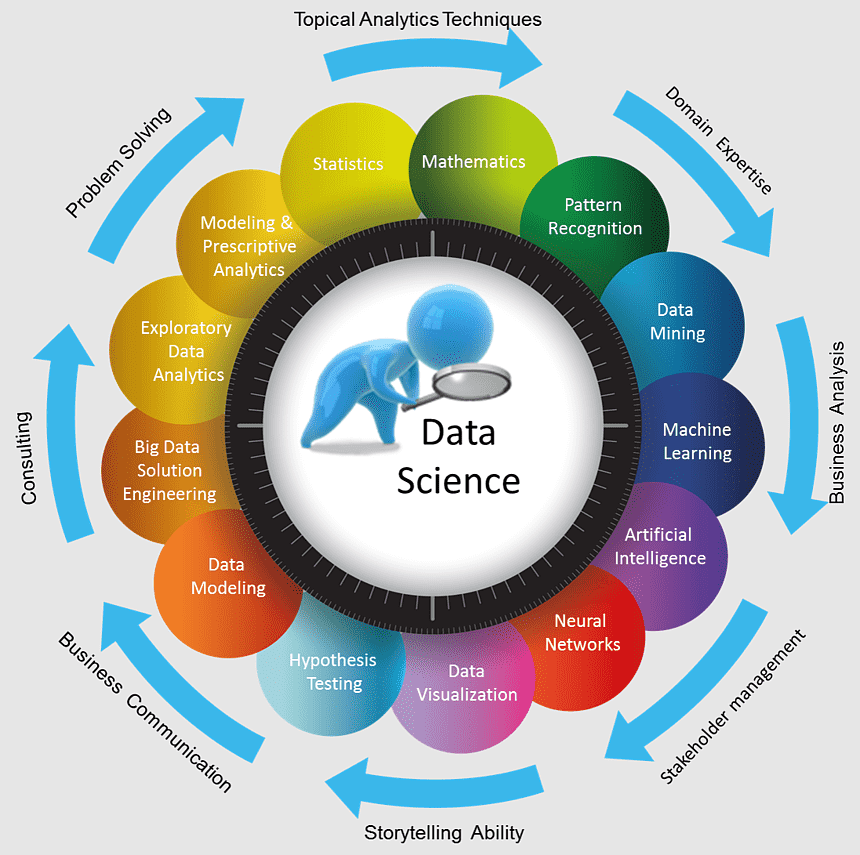
# ЛЕКЦИЯ № 2

# Business Intelligence

Data scientist – это специалист, владеющий тремя группами навыков:

1. IT-грамотность – программирование, придумывание и решение алгоритмических задач, владение софтом
2. Математические и статистические знания
3. Содержательный опыт в какой-то области – понимание бизнес-запросов своей организации или задач своей отрасли науки

В этой профессии нужно уметь строить гипотезы и ставить вопросы. А затем еще и переводить с языка данных на язык бизнеса, выстраивать совместную работу аналитиков, разработчиков и тех, кто занимается собственно развитием бизнеса.



## ИАД. BI

**Интеллектуальный анализ данных** – это обработка информации и выявление в ней моделей и тенденций, которые помогают принимать решения.

**Business Intelligence** – программные средства, функционирующие в рамках предприятия и обеспечивающие функции доступа и анализа информации, которая находится в хранилище данных, а также обеспечивающие принятие правильных и обоснованных управленческих решений.

## Процесс принятия решения

* Принятие решений в условиях определенности
* Принятие решений в условиях риска
* Принятие решений в условиях неопределенности
* Принятие решений в условиях многокритериальности

## Процесс выработки решения на основе первичных данных

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

* Обобщающий анализ
* Прогнозирование
* Ситуационное моделирование

## Business Intelligence

* Средства построения хранилищ данных (data warehousing. ХД)
* Системы оперативной аналитической обработки (OLAP)
* Информационно-аналитические системы (Enterprise Information Systems, EIS)
* Средства интеллектуального анализа данных (data mining)
* Инструменты для выполнения запросов и построения отчетов (query and reporting tools)

## Хранилища данных

Data warehouse – предметно-ориентированная информационная база данных, специально разработанная и предназначенная для подготовки отчётов и бизнес-анализа с целью поддержки принятия решений в организации. Строится на базе систем управления базами данных и систем поддержки принятия решений. Данные, поступающие в хранилище данных, как правило, доступны только для чтения.

## OLTP

OLTP (online transaction processing - оперативная обработка транзакций) — система обработки транзакций в реальном времени. Способ организации бд, при котором система работает с небольшими по размерам транзакциями, но идущими большим потоком, чтения.

Особенности использования:

* Строго нормализованные модели данных
* При возникновении ошибки транзакция должна целиком откатиться и вернуть систему к состоянию, которое было до начала транзакции
* Обработка данных в реальном времени (с малой задержкой)

## OLAP

OLAP (online analytical processing, интерактивная аналитическая обработка) - это технология комплексного многомерного анализа данных.

## Хранилища данных



## ИАС

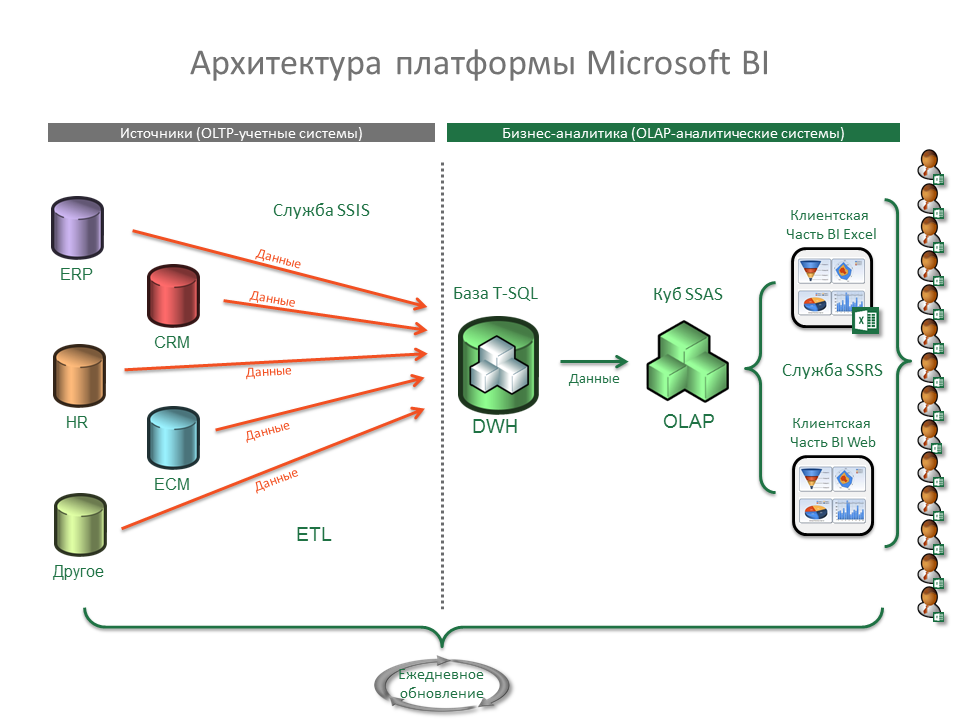
**Информационными системами руководителя**

* статистика запросов

**Динамические из поддержки решений**

* сфера детализированных данных
* сфера агрегированных показателей
* сфера закономерностей





Изображение выглядит как текст, снимок экрана, Устройство отображения, Прямоугольник

Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, Прямоугольник, дисплей

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, дисплей, компьютер

Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, Прямоугольник, дисплей

Автоматически созданное описание

Изображение выглядит как текст, СМИ, снимок экрана, Проекционный экран

Автоматически созданное описание

# ЛЕКЦИЯ № 3

# ИАД. KDD. DM.

## Развитие технологий БД

1960- е Сбор данных, разработка БД, сетевые СУБД

1970-е Реляционная модель данных, реляционные СУБД

1980-е Реляционные СУБД. Новые модели данных: иерархические, объетно-ориентированные, дедуктивные, сетевые. Предметно-ориентированные БД.

1990 е Мультимедиа БД, интернет БД, Data Mining, хранилища данных

2000 е Глобальные информационные системы, управление потоками данных, извлечение знаний, Data Mining web-технологии (SOA (Сервис-ориентированная архитектура), XML интеграция данных).

## ИАД

Интеллектуальный анализ данных

«извпечение зерен знаний из пор данных"

Knowledge Discovery in Databases (KDD)

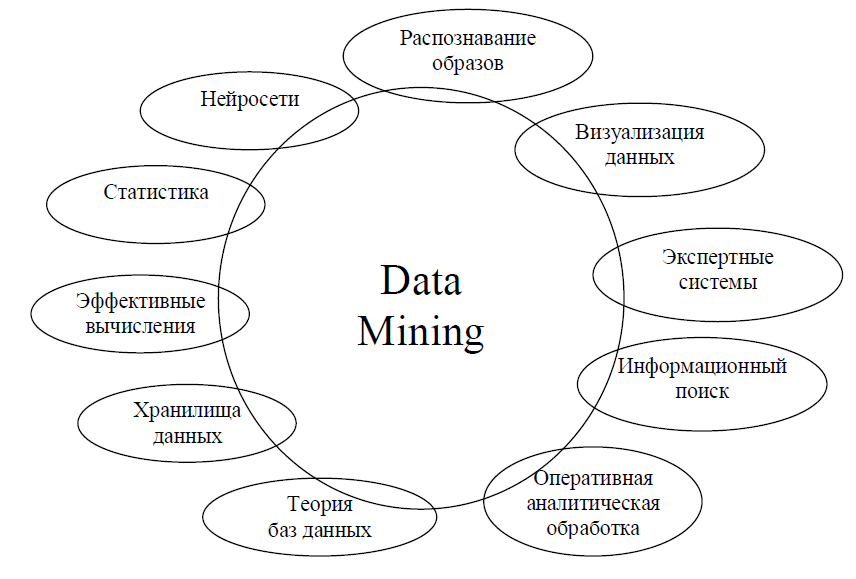
Data Mining (DM) – мультидисциплинарная область.

## Современные требования к интеллектуальной переработке сырых данных

* Данные имеют неограниченный объем
* Данные являются разнородными (количественными качественными, текстовыми)
* Результаты должны быть конкретны и понятны
* Инструменты для обработки сырых должны быть просты в использовании

Изображение выглядит как текст, презентация, СМИ, Проекционный экран

Автоматически созданное описание



## Формулировки задач при OLAP и Data Mining

Изображение выглядит как текст, снимок экрана, Шрифт, Красочность

Автоматически созданное описание

## Приложения Data Mining

**Массовый продукт для бизнес-приложений**

* Розничная торговля
* Страхование
* CRM
* Телекоммуникации
* Социология

**Инструмент для проведения уникальных исследований**

* Генетика
* Медицина
* Химия
* Исследование космоса



## Задачи, решаемые методами Data Mining (методы text mining)

* Классификация:
* Кластеризация:
* Прогнозирование;
* Ассоциация
  + Предсказуемость
  + Распространенность
  + Ожидаемая предсказуемость
* Последовательные шаблоны
* Анализ отклонений.

## ЛЕКЦИЯ № 4

## Классификация стадий и методов ИАД

  
Свободный поиск

Логические правила типа «ЕСЛИ …, ТО …»

ЕСЛИ Профессия=«Программист», ТО Возраст<=30 в 61% случаев;

ЕСЛИ Профессия=«Программист», ТО Возраст<=60 в 98% случаев.

## Индукция

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

## Прогностическое моделирование

Зная, Что Иванов - программист, можно быть

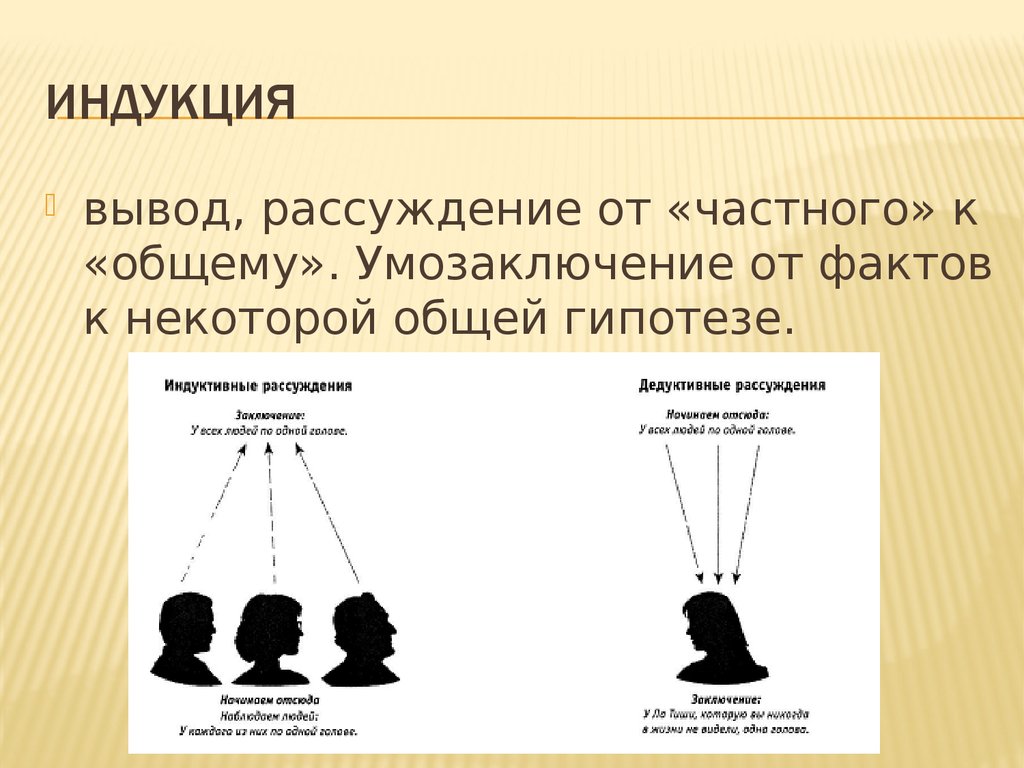
на 61% уверенным, что его возраст <=30 годам,

и на 98% - что он <=60 годам.

## Стадии ИАД

Свободный поиск - индуктивен;

Прогнозирование - дедуктивно.



## Методы ИАД

Изображение выглядит как текст, снимок экрана, дисплей, Устройство отображения

Автоматически созданное описание

Стадии ИАД:

* Непосредственное использование обучающих данных
  + Рассуждения на основе анализа прецедентов
* Выявление и использование формализованных закономерностей
  + Методы кросс табуляции
  + Методы логической индукции
  + Методы вывода уравнений

## Рассуждения на основе аналогичных случаев

Алгоритмы типа Lazy-Learning

(класс алгоритмов ленивого обучения, т.е обучение основано на запоминании обучающей выборки)

* ближайшего соседа (NearestNeighbor)
* k-ближайшего соседа (k-NearestNeighbor)

## Методы кросс-табуляции

* Кросс-табличная визуализация
* Байесовские сети

## Методы логической индукции

* Деревья решений
* Индукция правил

## Методы вывода уравнений

* Статистика
  + Корреляционно-регрессионный анализ
  + Ряды данных
  + Нелинейная регрессия
* Нейронные сети

## Исходные данные

* Обучающее множество (training set) - множество, которое включает данные, использующиеся для обучения (конструирования) модели.
* Тестовое (test set) множество используется для проверки работоспособности модели.

## Деревья решений

«ЕСЛИ… ТО…»

* Задача классификации - целевая переменная принимает дискретные значения.
* Задача прогнозирования - целевая переменная принимает непрерывные значения.

Изображение выглядит как текст, доска, компьютер, снимок экрана

Автоматически созданное описание

# ЛЕКЦИЯ № 5

# Классификация vs кластеризация

## Правила классификации

* в каждом акте деления необходимо применять только одно основание;
* деление должно быть соразмерным, те общий объем видовых понятий должен равняться объему депимого родового понятия;
* члены деления должны взаимно исключать друг друга, их объемы не должны перекрещиваться;
* деление должно быть последовательным.

## Классификация

* Естественная
* Искусственная
* Простая (дихотомия)
* Сложная

## Классификация

* Стратегия обучения с учителем
* Бинарная (по зависимой переменной) или множественная
* Одномерная (один признак) или многомерная (много признаков)

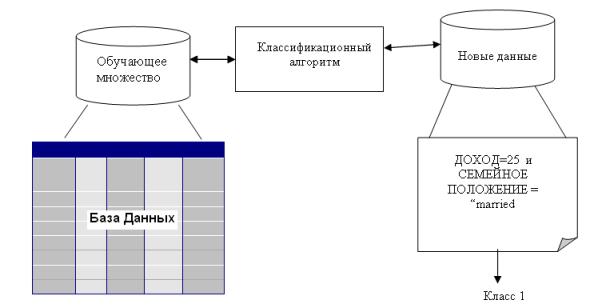
Задача классификации - предсказание категориальной зависимой переменной ( т.е. зависимой переменной, являющейся категорией) на основе выборки непрерывных и/или категориальных переменных

## Конструирование модели

* Каждый пример набора данных относится к одному предопределенному классу
* На этом этапе используется обучающее множество, на нем происходит конструирование модели
* Полученная модель представлена классификационными правилами, деревом решений или математической формулой

## Использование модели

* Оценка правильности (точности) модели.
  + Известные значения из тестового примера сравниваются с результатами использования полученной модели
  + Уровень точности – процент правильно классифицированных примеров в тестовом множестве
  + Тестовое множество, т.е. множество, на котором тестируется посторонняя модель, не должна зависеть и обучающего множества
* Если точность модели допустима, возможно использование модели для классификации новых примеров, класс которых неизвестен.

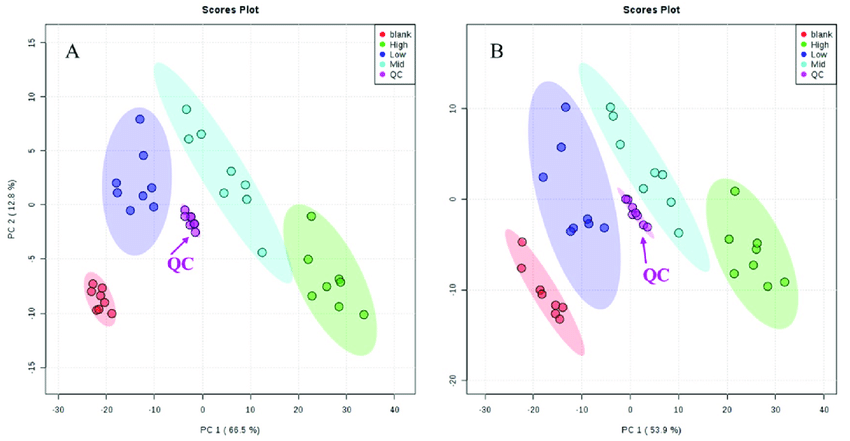


## Решение задачи классификации методом деревьев решений

Изображение выглядит как линия, текст, График, диаграмма

Автоматически созданное описание

## Решение задачи классификации методом нейронных сетей



## Оценка точности классификации

Изображение выглядит как текст, презентация, в помещении, Человеческое лицо

Автоматически созданное описание

## Кластеризация

* Стратегия обучения без учителя
* Описательная процедура
* Разведочный анализ

Кластеризация предназначена для разбиения совокупности объектов на однородные группы (кластеры или классы). Если данные выборки представить как точки в признаковом пространстве, то задача кластеризации сводится к определению «сгущений точек»

Характеристики кластера:

* внутренняя однородность,
* внешняя изолированность

## Алгоритмы кластеризации

* Алгоритмы, основанные на разделении данных в т.ч. итеративные:
  + Разделение объектов на k кластеров
  + Итеративное перераспределение объектов для улучшения кластеризации
* Иерархические алгоритмы:
  + Агломерация, каждый объект первоначально является кластером, кластеры соединяясь друг с другом, формируют большой кластер и тд
* Методы, основанные на концентрации объектов:
  + Основаны на возможности соединения объектов
  + Игнорируют шумы нахождение кластеров произвольной формы
* Грид-методы
  + Квантование объектов в грид-структуры
* Модельные методы
  + Использование модели для нахождения кластеров, наиболее соответствующих данных

## Оценка качества кластеризации

* Установление контрольных точек и проверка на полученных кластерах
* определение стабильности кластеризации путем добавления в модель новых переменных
* создание и сравнение кластеров с использованием различных методов. Разные методы кластеризации могут создавать разные кластеры, и это является нормальным явлением. Однако создание схожих кластеров различными методами указывает на правильность кластеризации

## Кластер

**Центр** кластера — это среднее геометрическое место точек в пространстве переменных

**Радиус** кластера – это максимальное расстояние точек от центра кластера

**Спорный** объект — это объект, который по мере сходства может быть отнесен к нескольким кластерам

## Кластеризация – предположения

* Рассматриваемые признаки объекта в принципе допускают желательное разбиение совокупности объектов на кластеры
* Правильность выбора масштаба или единиц измерения признаков
* **Стандартизация** / нормализация переменных

# ЛЕКЦИЯ № 6

# Прогнозирование.  Визуализация

## Прогнозирование - задача Data Mining

прогнозирование - установление функциональной зависимости между зависимыми и независимыми переменными.

## Задача прогнозирования рынков

Прогноз продаж товаров (например, с целью определения нормы товарного запаса)

Прогнозирование продаж товаров, оказывающих влияние друг на друга

Прогноз продаж в зависимости от внешних факторов

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

## Прогнозирование / Классификация

Сходства:

Обучающая выборка

Различия:

К - класс зависимой переменной

П - числовые значения зависимой переменной

## Прогнозирование и временные ряды

**Временной ряд** - последовательность наблюдаемых значений какого-либо признака, упорядоченных в неслучайные моменты времени.

* Члены временного ряда, в отличие от элементов спутанной выборки, не являются статистически независимыми
* Члены временного ряда не являются одинаково распределенными.

## Анализ временного ряда

* Определение природы ряда
* Прогнозирование будущих значений ряда

## Тренд, сезонность, цикл

Тренд - неслучайная функция, которая формируется под действием общих или долговременных тенденций, влияющих на временной ряд

* прогнозирование **односерийных** рядов,
* прогнозирование **мультисерийных,** или взаимовлияющих, рядов.

**Свойство сезонности** означает, что через примерно равные промежутки времени форма кривой, которая описывает поведение зависимой переменной, повторяет свои характерные очертания.

Отличия циклической компоненты от сезонной:

* Продолжительность цикла как правило, больше. чем один сезонный период,
* Циклы, в отличие от сезонных периодов, не имают определенной продолжительности.
* Что нужно прогнозировать?
* В каких временных элементах (параметрах)?
* **период** прогнозирования
* **горизонт** прогнозирования
* **интервал** прогнозирования
* С какой точностью прогноза?

## Виды прогнозов

* Краткосрочный прогноз
* Среднесрочный
* Долгосрочный прогноз

## Факторы влияния

* Анализируемый процесс относительно стабилен во времени изменения происходят медленно процесс не зависит от внешних факторов.
* Анализируемый процесс нестабилен и очень сильно зависит от внешних факторов.

## Методы прогнозирования

* Нейронные сети
* Линейная регрессия

## Задача визуализации

**Визуализация** — это инструментарий, который позволяет увидеть конечный результат вычислений, организовать управление вычислительным процессом и даже вернуться назад к исходным данным, чтобы определить наиболее рациональное направление дальнейшего движения.

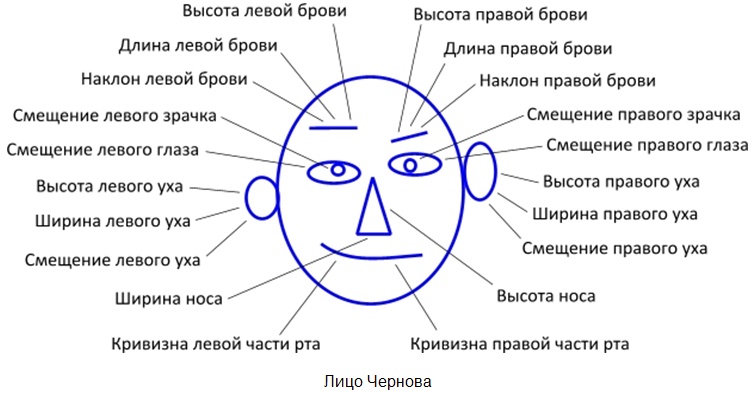
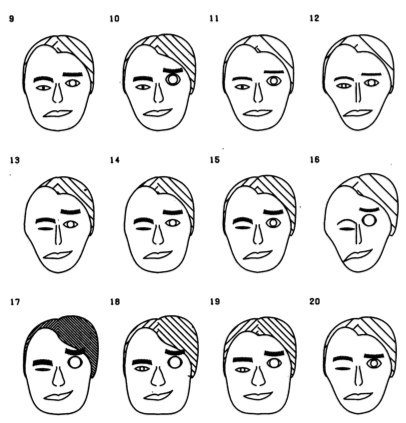
## Визуализация инструментов Data Mining

* Для деревьев решений это визуализатор дерева решений, список правил.
* Для нейронных сетей в зависимости от инструмента это может быть топология сети, график изменения величины ошибки, демонстрирующий процесс обучения.
* Для линейной регрессии в качестве визуализатора выступает пиния регрессии.
* Для кластеризации диаграммы, диаграммы рассеивания.
* Диаграммы и графики рассеивания часто используются для оценки качества работы того или иного метода.

## Методы визуализации

* Представление данных в двух и трех намерениях
* Представление данных в четырех и более измерениях
  + «лица Чернова»
  + лепестковые диаграммы.

## Лица Чернова - многофакторный анализ

## Визуализация данных

* Графики и диаграммы
* Инфорграфика и схемы
* Презентация и **анализ данных**
* Интерактивный сторителлинг
* Бизнес аналитика и дашборды
* Научная и медицинская визуализация
* Карты и картограммы

## Инструменты для визуализации данных

Изображение выглядит как текст, в помещении, Прямоугольник, презентация

Автоматически созданное описание

# Лекция 7 – Интеграция данных

**Уровни интеграции данных**

* Интеграция данных на **физическом** уровне (конверсия данных из различных источников в требуемый единый формат их физического представления)
* Интеграция данных на **логическом** уровне (доступ к данным, содержащимся в различных источниках, в терминах единой глобальной схемы свойств данных)
* Интеграция данных на **семантическом** уровне (единое представление данных с учетом их семантических свойств в контексте единой онтологии предметной области)

**Сложности интеграции данных**

* **Физический** уровень – разные форматы
* **Логический** уровень – разные модели/схемы данных
* **Семантический** уровень – разные онтологии/понятийная система

**Проблемы интеграции данных**

* **Разнородность** (гетерогенность)
* **Автономность**
* **Распределенность**

**Характеристики интеграции данных**

* *приложения* – это решения, созданные поставщиком в соответствии с требованиями клиентов, которые используют один или более продуктов интеграции данных
* *продукты* – это готовые коммерческие решения, поддерживающие одну или более технологий интеграции данных
* *технологии* реализуют один или более методов интеграции данных
* *методы* – это подходы к интеграции данных, независимые от технологий

***Методы интеграции (объединения) данных: Консолидация, Федерализация, Распространение, Сервисный подход (SOA), Семантическая интеграция***

**Консолидация:**

*ETL процессы (Extraction, Transformation, Load) 60-80% времени*

* Извлечение и очистка данных
* Трансформация данных
* Загрузка данных

**Интеграция**: синтаксис/семантика

**CDM – общая модель данных**

**Универсальные форматы обмена данных** (XML Conver.)

# Лекция 8 – Технология поиска информации Text Miming

## Информационный поиск (Information Retrieval)

**Информационный поиск** – один из основных информационных процессов.

**Информационная потребность** – это некий набор данных, необходимый пользователю для того, чтобы больше узнать об интересующей его предметной области.

**Релевантность** – устанавливаемое при информационном поиске соответствие содержания документа информационному запросу или поискового образа документа поисковому предписанию.

**Поисковый образ документа** – это выраженное в терминах формализованного информационно-поискового языка основное смысловое содержание этого документа.

**Text Mining – глубинный анализ текстов**

* извлечение из текста характерных элементов или признаков, которые могут использоваться в качестве ключевых слов, метаданных, аннотаций
* отнесение документов к некоторым категориям из заданной схемы их систематизации.

*«Процесс обнаружения в сырых данных ранее неизвестных нетривиальных практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.» Пятецкий-Шапро*

**Контент-анализ**

* Контент-анализ – это методика объективного качественного и систематического изучения содержания средств коммуникации
* Контент-анализ – это систематическая числовая обработка, оценка и интерпретация формы и содержания информационного источника
* Контент-анализ – это качественно-количественный метод изучения документов, который характеризуется объективностью выводов и строгостью процедуры и заключается в квантифициированной обработке текста с дальнейшей интерпретацией результатов.
* Контент-анализ состоит в нахождении в тексте определенных содержательных понятий (единиц анализа), выявлении частоты их появления и соотношения с содержанием всего документа

**Контент-анализ**

* качественный: основан на самом факте присутствия или отсутствия в тексте одной или нескольких характеристик содержания
* количественный: частота появления в документах определенных характеристик содержания (понятий, феноменов)

**Элементы Text Mining:** классификация; кластеризация; прогнозирование; нахождение исключений; поиск связанных признаков; извлечение фактов, понятий; реферирование; ответ на запросы; тематическое индексирование; поиск по ключевым словам.

**Извлечение понятий**

* извлечение слов или словосочетаний, важных для описания содержания текста. Это могут быть списки терминов предметной области, персон, организаций, географических названий и др.
* прослеживание связей между извлеченными понятиями
* извлечение сущностей, распознавание фактов и событий.

Изображение выглядит как текст, диаграмма, План, Шрифт

Автоматически созданное описание

*семантико-лингвистические методы*

**Автоматическое реферирование**

составление коротких изложений материалов, аннотаций или дайджестов, т.е. извлечение наиболее важных сведений из одного или нескольких документов и генерации на их основе лаконичных отчетов.

* квазиреферирование
* краткое изложение содержания первичных документов

**Квазиреферирование**

* статистические методы, основанные на оценке информативности разных элементов текста по частоте появления, которая служит основным критерием информативности слов, предложений или фраз
* позиционные методы, которые опираются на предположение о том, что информативность элемента текста зависит от его позиции в документе
* индикаторные методы, основанные на оценке элементов текста, исходя из наличия в них специальных слов и словосочетаний – маркеров важности, которые характеризуются их содержательной значимостью.

**Краткое изложение**

* первая стадия – сведение исходной текстовой информации к заданному числу фрагментов единиц значения, которыми являются категории, последовательности и темы.
* на второй стадии производится поиск регулярных связей между единицами значения
* третья стадия – формирование выводов и обобщений. На этой стадии создается структурная аннотация, представляющая содержание текста в виде совокупности концептуально связанных смысловых единиц.

**Автоматическое реферирование**

* удаление малозначащих смысловых единиц
* сокращения смысловых единиц – замена их основной лексический единицей, выражающей основной смысл
* гибридный способ, заключающийся в уточнение реферата с помощью статистических методов, с использованием семантических классов, особенностей контекста и синонимических связей.

Семантические методы:

* Синтаксический разбор
* Семантические структуры

**Выявление дублирования информации**

* Дубль
* Почти дубль

**Выявление новых событий**

* первому рассматриваемому документу ставится в соответствие первый кластер. Каждый кластер представляется вектором термов (ключевых слов), входящих в документ этого кластера
* каждый следующий документ сравнивается с **центроидами** существующих кластеров
* если документ достаточно близок к некоторому кластеру, то он предписывается этому кластеру, после чего происходит пересчет соответствующего центроида
* если документ не близок к существующим кластерам, то происходит формирование нового кластера, которому приписывается данный документ
* временной диапазон рассматриваемых документов принято называть «окном наблюдения». Кластеры, все документы которых выходят за пределы окна наблюдений, выносятся за рамки рассмотрения.

**TF-IDF**

***TF*** *- term frequency – частность термина, которая измеряет, насколько часто термин встречается в документе*

***IDF*** *– inverse document frequency – это обратная частность документов*

*статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса. Вес некоторого слова пропорционален количеству употребления этого слова в документе, и обратно пропорционален частоте употребления слова в других документах коллекции.*

*Мера TF-IDF часто используется в задачах анализа текстов и информационного поиска, например, как один из критериев релевантности документа поисковому запросу, при расчете меры близости документов при кластеризации.*

**Выявление новых событий**

* минимальное время, прошедшее с момента публикации
* минимизация веса термов, входящих в документ, по частному словарю, сформированному на основании анализа большого массива опубликованных документов
* максимизация суммарного веса термов, входящих в документ, по плюс-словарю (содержащему важные для содержания новостей слова типа «теракт», «конфликт», «сенсация» и т.п.)
* учет ранга в «авторитетности» источника (как правило определяемый экспертами).

**ИАТ**

* управление документооборотом компании
* e-commerce
* бренд менеджмент
* маркетинг
* конкурентная разведка
* управление опытом клиента
* информационная безопасность
* виртуальные ассистенты и др.

**Корпоративный поиск**

* нет статистики по поисковым запросам
* полнота важнее точности
* персонализация
* доступность и свежесть поискового индекса

Ответы на вопросы для подготовки к экзамену

# 1. Информация

**Информация** - сведения об объектах и явлениях окружающей среды, их параметрах, свойствах и состоянии, которые воспринимают информационные системы (живые организмы, управляющие машины и др.) в процессе жизнедеятельности и работы. Применительно к обработке данных на компьютерах - произвольная последовательность символов, несущих смысловую нагрузку.

Качество информации можно определить как совокупность свойств, определяющих возможность ее использования для удовлетворения определенных в соответствие с ее назначением потребностей.

# Свойства информации:

1. **Дуализм** - характеризует ее двойственность. С одной стороны, информация объективна в силу объективности данных, с другой - субъективна, в силу субъективности применяемых методов. (Субъективность, объективность);
2. **Полнота** - характеризует степень достаточности данных для принятия решения или создания новых данных на основе имеющихся. (Неполный набор данных, избыточный набор данных);
3. **Достоверность** - характеризуется степенью соответствия информации реальному объекту с необходимой точностью. (Доверительная вероятность необходимой точности);
4. **Точность** - определяется степенью близости получаемой информации к реальному состоянию объекта, процесса, явления и т.п. (Формальная точность, реальная точность, максимальная точность, необходимая точность);
5. **Доступность** - возможность получения информации при необходимости. Обеспечивается выполнением соответствующих процедур ее получения и преобразования. (Доступность данных и методов);
6. **Актуальность** - степень соответствия информации текущему моменту времени. A(t)=Z(t)/Z(t0), где Z(t) - ценность информации в момент времени 1;
7. **Своевременность** - означает ее поступление не позже заранее назначенного момента времени, согласованного с временем решения поставленной задачи;
8. **Устойчивость** - отражает ее способность реагировать на изменения исходных данных без нарушения необходимой точности;
9. **Репрезентативность** - связана с правильностью ее отбора и формирования в целях адекватного отражения свойств объекта;
10. **Адекватность** - выражает степень соответствия, создаваемого с помощью информации образа реальному объекту, процессу, явлению. (Синтаксическая, семантическая, прагматическая).

Два вида адекватности:

1. **Синтаксическая адекватность** отображает формально-структурные характеристики информации и не затрагивает ее смыслового содержания. Учитываются:

* тип носителя;
* способ представления информации;
* скорость передачи и обработки;
* размеры кодов представления информации;
* надежность и точность преобразования этих кодов и т.п.

**Синтаксическая мера информации** оперирует с обезличенной информацией, не выражающей смыслового отношения к объекту.

1. **Семантическая адекватность** определяет степень соответствия образа объекта и самого объекта. Предполагает учет смыслового содержания информации. На этом уровне анализируются те сведения, которые отражает информация, рассматриваются смысловые связи.

**Семантическая мера информации** выявляется на фоне способности пользователя принимать поступившее сообщение.

**Информационная деятельность** - деятельность человека, связанная с процессами:

* получения информации;
* преобразования информации;
* накопления информации;
* передачи информации.

**Информационная технология (ИТ)** - процесс, использующий совокупность средств и методов сбора, обработки и передачи данных (первичной информации) для получения информации нового качества о состоянии объекта, процесса или явления.

# Носители информации:

* внемашинные (книги и т.д.);
* машиночитаемые (жесткий диск, оптический диск и т.д.);
* электронные (то, что не существует без ЭВМ - оперативная память, облачные данные и т.д.).

# Способы обработки информации:

* неавтоматизированные;
* автоматизированные (человек участвует для оптимизации?);
* автоматические.

# Классификация информационных технологий:

* по виду представления обрабатываемой информации;
* в зависимости от класса решаемых задач;
* по степени охвата задач управления.

**Обработка информации** состоит в получении одних «информационных объектов» из других «информационных объектов» путем выполнения некоторых алгоритмов и является одной из основных операций, осуществляемых над информацией, и главным средством увеличения ее объема и разнообразия.

**Технология обработки информации** - взаимосвязанные действия, выполняемые в строго определенной последовательности с момента возникновения информации до получения заданных результатов.

# Основные процедуры обработки информации:



**Цели:**

1. Сбор информации (различные формы);
2. Определение существенной информации (на данный момент);
3. Предоставление существенной информации (в нужном виде).

# Данные

Данные относятся к способу представления, хранения и элементарным операциям обработки информации. Прежде всего, данные - это основа информации. Образно говоря, **данные** - это текст в некотором алфавите, а информация - это рассказ, сообщение, сведения, имеющие определенный смысл.

Одна из главных особенностей данных состоит в том, что их становится невероятно много. В современных условиях массового применения компьютеров источников данных гигантское количество.

Существуют четыре аспекта работы с данными:

* определение данных;
* вычисление данных;
* манипулирование данными и их обработка;
* управление данными (администрирование данных).

# Знания

**Знание** - проверенные общественной практикой полезные сведения, которые могут многократно использоваться людьми для решения тех или иных задач.

Обычно понятие Знания трактуется с двух позиций:

1. Совокупность сведений и фактов в какой-либо области;
2. Постижение действительности сознанием в виде науки.

# Понятие интеллектуального анализа данных (ИАД)

**Интеллектуальный анализ данных** - это обработка информации и выявление в ней моделей и тенденций, которые помогают принимать решения.

# Data Mining

**Data Mining** - мультидисциплинарная область, изучающая процесс нахождения новых, действительных и потенциально полезных знаний в базах данных, возникшая и развивающаяся на базе нескольких наук.

Некоторые дисциплины, на стыке которых появилась технология Data Mining:

* **статистика** - это наука о методах сбора данных, их обработки и анализа для выявления закономерностей, присущих изучаемому явлению;
* **машинное обучение** - это наука, которая изучает компьютерные алгоритмы, автоматически улучшающиеся во время работы;

Одним из наиболее популярных примеров алгоритма машинного обучения являются нейронные сети;

* **искусственный интеллект** - использование компьютеров и систем для имитации мыслительного процесса человека с целью решения задач и принятия решений.

# Business Intelligence

**Business Intelligence** - программные средства, функционирующие в рамках предприятия и обеспечивающие функции доступа и анализа информации, которая находится в хранилище данных, а также обеспечивающие принятие правильных и обоснованных управленческих решений.

# Business Intelligence:

* средства построения хранилищ данных (data warehousing, ХД);
* системы оперативной аналитической обработки (OLAP);
* информационно-аналитические системы (Enterprise Information Systems, EIS);
* средства интеллектуального анализа данных (data mining);
* инструменты для выполнения запросов и построения отчетов (query and reporting tools).

# Пирамида управления и связь с BI

В современном мире успех компании на рынке напрямую зависит от того, как быстро менеджмент компании может распознать изменения динамики рынка и насколько своевременно может отреагировать на них с целью увеличения прибыли, исходя из существующих реалий рынка. Менеджеры компании должны отслеживать тенденции рынка, идентифицировать конкурентов и угрозы, оценивать риски, преобразовывать стратегию компании, оценивать свои ресурсы и т.д. Информация является необходимым производственным ресурсом для принятия эффективных управленческих решений.

Компании накопили значительные объемы данных и имеют доступ к еще большим объемам внешних данных. Менеджерам необходимо, чтобы эта информация была преобразована, предварительно обработана и соответствующим образом организована для быстрого доступа, анализа и принятия решений. Такой подход к данным есть, с одной стороны, создание конкурентного преимущества, а с другой стороны - требование к публикации данных для менеджеров компании. Публикация данных для менеджеров, обеспечивающая быстрый доступ к данным, выполнение анализа данных и информационную поддержку процесса принятия решений, является основной целью систем бизнес-аналитики. Бизнес-аналитика помогает компании создавать знания из всей доступной информации для принятия эффективных управленческих решений и превращения этих решений в действие.

Таким образом, ключевую роль в управлении организацией в целом и ее отдельными производственными функциями играет информация. Данные, которые доступны менеджерам и аналитикам непосредственно из корпоративных информационных систем, не унифицированы, разрозненны и в общем случае не готовы для анализа. Системы деловой осведомленности или бизнес-аналитики являются тем классом информационных систем, который позволяет превратить данные корпоративных информационных систем и данные из внешних источников в полезные для бизнеса информацию и знания, используемые в управлении, на основе которых можно принимать решения.

Информационным фундаментом для бизнес-анализа и систем бизнес-аналитики является хранилище данных. Основное требование к хранилищу данных системы бизнес-анализа состоит в том, чтобы обеспечить структурированную и организованную

для решения задач бизнеса информационную среду. Как правило, такую среду лаконично представляют в виде пирамиды управления.



Пирамида управления формируется из нескольких уровней:

* **уровень оперативной информации**. На этом уровне ИТ обеспечивают работу с данными на уровне бизнес-процедур компании. Данные в автоматизированных системах являются хорошо структурированными и детальными. С этими данными работают специалисты компании: бухгалтеры, менеджеры продаж, плановики и т.д.
* **уровень тактической информации**. На этом уровне ИТ обеспечивают интеграцию данных на уровне бизнес-процессов оперативного управления производством в рамках подразделений компании. С этими данными работают руководители подразделений компании при выполнении ежедневных производственных заданий.
* **уровень стратегической информации**. На этом уровне ИТ обеспечивают интеграцию данных на уровне бизнес-процессов по направлениям хозяйственной деятельности компании. С этими данными работают аналитики и руководители высшего звена компании, которые готовят стратегические решения развития и деятельности компании на рынке.
* **уровень принятия решений**. На этом уровне ИТ обеспечивают интеграцию и агрегацию данных на уровне бизнес-процессов компании для руководителей высшего

звена компании. Этот уровень обеспечивает информационную поддержку принятия решений.

# Задачи систем поддержки принятия решений

**Система поддержки принятия решений** - компьютерная автоматизированная система, целью которой является помощь людям, принимающим решение в сложных условиях для полного и объективного анализа предметной деятельности.

СППР возникли в результате слияния управленческих информационных систем и систем управления базами данных.

Система поддержки принятия решений предназначена для поддержки многокритериальных решений в сложной информационной среде. При этом под многокритериальностью понимается тот факт, что результаты принимаемых решений оцениваются не по одному, а по совокупности многих показателей (критериев) рассматриваемых одновременно. Информационная сложность определяется необходимостью учета большого объема данных, обработка которых без помощи современной вычислительной техники практически невыполнима. В этих условиях число возможных решений, как правило, весьма велико, и выбор наилучшего из них "на глаз", без всестороннего анализа может приводить к грубым ошибкам.

Система поддержки решений СППР решает две основные задачи:

* выбор наилучшего решения из множества возможных (оптимизация);
* упорядочение возможных решений по предпочтительности (ранжирование).

В обеих задачах первым и наиболее принципиальным моментом является выбор совокупности критериев, на основе которых в дальнейшем будут оцениваться и сопоставляться возможные решения (будем называть их также альтернативами).

# Процесс принятия решения:

* принятие решений в условиях определенности;
* принятие решений в условиях риска;
* принятие решений в условиях неопределенности;
* принятие решений в условиях многокритериальности.

# Понятие модели предметной области

**Модель предметной области** - система, имитирующая структуру или функционирование исследуемой предметной области и отвечающая основному требованию - быть адекватной этой области.

Предварительное моделирование предметной области позволяет сократить время и сроки проведения проектировочных работ и получить более эффективный и качественный проект. Без проведения моделирования предметной области велика вероятность допущения большого количества ошибок в решении стратегических вопросов, приводящих к экономическим потерям и высоким затратам на последующее перепроектирование системы. Вследствие этого все современные технологии проектирования ИС основываются на использовании методологии моделирования предметной области.

К моделям предметных областей предъявляются следующие требования:

* формализация, обеспечивающая однозначное описание структуры предметной области;
* понятность для заказчиков и разработчиков на основе применения графических средств отображения модели;
* реализуемость, подразумевающая наличие средств физической реализации модели предметной области в ИС;
* обеспечение оценки эффективности реализации модели предметной области на основе определенных методов и вычисляемых показателей.

# Понятие информационного объекта

**Информационный объект** - это описание некоторой сущности (реального объекта, явления, процесса, события) в виде совокупности логически связанных реквизитов (информационных элементов). Такими сущностями для информационных объектов могут служить: цех, склад, материал, вуз, студент, сдача экзаменов и т.д.

Информационный объект определенного реквизитного состава и структуры образует класс (тип), которому присваивается уникальное имя, например Студент, Сессия, Стипендия.

Информационный объект может иметь множество реализаций - экземпляров. Например, каждый экземпляр объекта Ученик представляет конкретного ученика.

Экземпляр характеризуется совокупностью конкретных значений реквизитов и должен однозначно идентифицироваться значением ключа информационного объекта, который может состоять из одного или нескольких ключевых реквизитов. Таким образом, реквизиты подразделяются на описательные и ключевые.

# Понятие информационно-логической модели

**Информационно-логическая модель** - ориентированная на человека и не зависимая от типа СУБД модель предметной области, определяющая совокупности информационных объектов, их атрибутов и отношений между объектами, динамику изменений предметной области, а также характер информационных потребностей пользователей.

Основные подходы к созданию инфологической модели предметной области:

* **функциональный подход** к проектированию БД. Этот метод реализует принцип "от задач" и применяется тогда, когда известны функции некоторой группы лиц и/или комплекса задач, для обслуживания информационных потребностей которых создается рассматриваемая БД.
* **предметный подход** к проектированию БД. Предметный подход к проектированию БД применяется в тех случаях, когда у разработчиков есть четкое представление о самой предметной области и о том, какую именно информацию они хотели бы хранить в БД, а структура запросов не определена или определена не полностью. Тогда основное внимание уделяется исследованию ПО и наиболее адекватному ее отображению в БД с учетом самого широкого спектра информационных запросов к ней.
* проектирование с использованием метода "**сущность-связь**". Метод "сущность-связь" является комбинацией двух предыдущих и обладает достоинствами обоих. Этап инфологического проектирования начинается с моделирования предметной области. Проектировщик разбивает ее на ряд локальных областей, каждая из которых (в идеале) включает в себя информацию, достаточную для обеспечения запросов отдельной группы будущих пользователей или решения отдельной задачи (подзадачи). Каждое локальное представление моделируется отдельно, затем они объединяются.

# OLTP-системы

**Транзакция -** минимальная логически осмысленная операция, которая имеет смысл и может быть совершена только полностью.

**OLTP** (Online Transaction Processing - оперативная обработка транзакций) - система обработки транзакций в реальном времени. Способ организации БД, при котором система работает с небольшими по размерам транзакциями, но идущими большим потоком, чтения.

* строго нормализованные модели данных;
* при возникновении ошибки транзакция должна целиком откатится и вернуть систему к состоянию, которое было до начала транзакции;
* обработка данных в реальном времени (с малой задержкой).

# Концепция хранилища данных

**Data Warehouse** (хранилища данных) - предметно-ориентированная информационная база данных, специально разработанная и предназначенная для подготовки отчетов и бизнес-анализа с целью поддержки принятия решений в организации. Строится на базе систем управления базами данных и систем поддержки принятия решений. Данные, поступающие в хранилище данных, как правило, доступны только для чтения.

# Организация ХД

Схемы **«звезда»** и **«снежинка»** - это два способа структурировать хранилище данных.

Схема типа «звезда» имеет централизованное хранилище данных, которое хранится в таблице фактов. Схема разбивает таблицу фактов на ряд денормализованных таблиц измерений. Таблица фактов содержит агрегированные данные, которые будут использоваться для составления отчетов, а таблица измерений описывает хранимые данные.

Денормализованные проекты менее сложны, потому что данные сгруппированы. Таблица фактов использует только одну ссылку для присоединения к каждой таблице измерений. Более простая конструкция звездообразной схемы значительно упрощает написание сложных запросов.

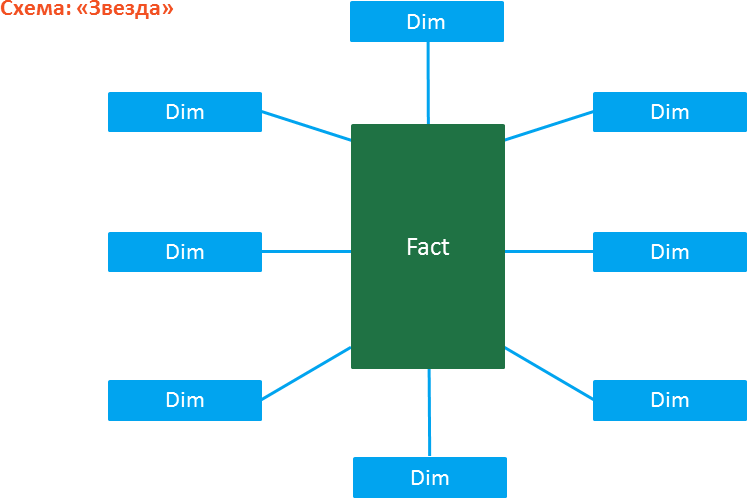
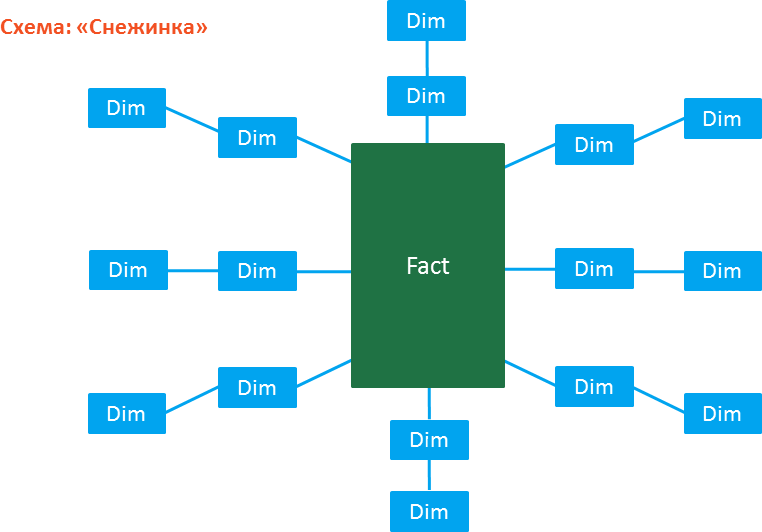


Схема типа «снежинка» отличается тем, что использует нормализованные данные. Нормализация означает эффективную организацию данных так, чтобы все зависимости данных были определены, и каждая таблица содержала минимум избыточности. Таким образом, отдельные таблицы измерений разветвляются на отдельные таблицы измерений.

Схема «снежинки» использует меньше дискового пространства и лучше сохраняет целостность данных. Основным недостатком является сложность запросов, необходимых для доступа к данным - каждый запрос должен пройти несколько соединений таблиц, чтобы получить соответствующие данные.

# Данные и атрибуты

Данные представляют собой факты, текст, графики, картинки, звуки, аналоговые или цифровые видео-сегменты.

Данные могут быть получены в результате измерений, экспериментов, арифметических и логических операций.

Данные должны быть представлены в форме, пригодной для хранения, передачи и обработки.

Иными словами, **данные** - это необработанный материал, предоставляемый поставщиками данных и используемый потребителями для формирования информации на основе данных.

**Атрибут** - свойство, характеризующее объект. Например: цвет глаз человека, температура воды и т.д

Атрибут также называют переменной, полем таблицы, измерением, характеристикой.

# Шкала измерений: номинальная

**Номинальная шкала** - шкала, содержащая только категории; данные в ней не могут упорядочиваться, с ними не могут быть произведены никакие арифметические действия.

Номинальная шкала состоит из названий, категорий, имен для классификации и сортировки объектов или наблюдений по некоторому признаку.

Пример такой шкалы: профессии, город проживания, семейное положение. Для этой шкалы применимы только такие операции: равно (=), не равно (≠).

# Шкала измерений: порядковая

**Порядковая шкала** - шкала, в которой числа присваивают объектам для обозначения относительной позиции объектов, но не величины различий между ними.

Шкала измерений дает возможность ранжировать значения переменных. Измерения же в порядковой шкале содержат информацию только о порядке следования величин, но не позволяют сказать "насколько одна величина больше другой", или "насколько она меньше другой".

Пример такой шкалы: место (1, 2, 3-е), которое команда получила на соревнованиях, номер студента в рейтинге успеваемости (1-й, 23-й, и т.д.), при этом неизвестно, насколько один студент успешней другого, известен лишь его номер в рейтинге.

Для этой шкалы применимы только такие операции: равно (=), не равно (≠), больше (>), меньше (<).

# Шкала измерений: интервальная

**Интервальная шкала** - шкала, разности между значениями которой могут быть вычислены, однако их отношения не имеют смысла.

Эта шкала позволяет находить разницу между двумя величинами, обладает свойствами номинальной и порядковой шкал, а также позволяет определить количественное изменение признака.

Пример такой шкалы: температура воды в море утром - 19 градусов, вечером - 24, т.е. вечерняя на 5 градусов выше, но нельзя сказать, что она в 1,26 раз выше.

Номинальная и порядковая шкалы являются дискретными, а интервальная шкала

- непрерывной, она позволяет осуществлять точные измерения признака и производить арифметические операции сложения, вычитания, умножения, деления.

Для этой шкалы применимы только такие операции: равно (=), не равно (≠), больше (>), меньше (<), операции сложения (+) и вычитания (-).

# Шкала измерений: относительная

**Относительная шкала** - шкала, в которой есть определенная точка отсчета и возможны отношения между значениями шкалы.

Пример такой шкалы: вес новорожденного ребенка (4 кг и 3 кг). Первый в 1,33 раза тяжелее. Цена на картофель в супермаркете выше в 1,2 раза, чем цена на базаре.

Относительные и интервальные шкалы являются числовыми.

Для этой шкалы применимы только такие операции: равно (=), не равно (≠), больше (>), меньше (<), операции сложения (+) и вычитания (-), умножения (\*) и деления (/).

# Шкала измерений: дихотомическая

**Дихотомическая шкала** - шкала, содержащая только две категории. Пример такой шкалы: пол (мужской и женский).

# Табличные данные

**Табличные данные** - данные, состоящие из записей, каждая из которых состоит из фиксированного набора атрибутов.

# Транзакционные данные

**Транзакционные данные** представляют собой особый тип данных, где каждая запись, являющаяся транзакцией, включает набор значений.

# Метаданные

**Метаданные** - это данные о данных.

В состав метаданных могут входить: каталоги, справочники, реестры. Метаданные содержат сведения о составе данных, содержании, статусе,

происхождении, местонахождении, качестве, форматах и формах представления, условиях доступа, приобретения и использования, авторских, имущественных и смежных с ними правах на данные и др.

Метаданные - важное понятие в управлении хранилищем данных.

Метаданные, применяемые при управлении хранилищем, содержат информацию, необходимую для его настройки и использования. Различают бизнес-метаданные и оперативные метаданные.

**Бизнес-метаданные** содержат бизнес-термины и определения, принадлежность данных и правила оплаты услуг хранилища.

**Оперативные метаданные** - это информация, собранная во время работы хранилища данных:

* происхождение перенесенных и преобразованных данных;
* статус использования данных (активные, архивированные или удаленные);
* данные мониторинга, такие как статистика использования, сообщения об ошибках и т.д.

Метаданные хранилища обычно размещаются в репозитории. Это позволяет использовать метаданные совместно различным инструментам, а также процессам при проектировании, установке, эксплуатации и администрировании хранилища.

# Многомерная модель данных

Метод **многомерного моделирования** дает абстрактную модель предметной области, используя следующие основные понятия: показатели или метрики, факты и измерения.

В многомерной модели данные представляются в виде гиперкубов, используя которые, можно получать различные срезы при аналитической обработке данных. Оси гиперкуба содержат параметры, а ячейки - зависящие от них агрегатные данные. Вдоль каждой оси данные могут быть организованы в виде иерархии, представляющей различные уровни их детализации. Благодаря такой модели данных пользователи могут формулировать сложные запросы, генерировать отчеты, получать подмножества данных.

# OLAP-системы

**OLAP (online analytical processing, интерактивная аналитическая обработка)**

* это технология комплексного многомерного анализа данных.

Основное назначение OLAP - поддержка аналитической деятельности, произвольных (не регламентированных) запросов пользователей-аналитиков.

Главная идея данной системы заключается в построении многомерных таблиц, которые могут быть доступны для запросов пользователей. Эти многомерные таблицы или так называемые многомерные кубы строятся на основе исходных и агрегированных данных. И исходные, и агрегированные данные для многомерных таблиц могут храниться как в реляционных, так и в многомерных базах данных. Взаимодействуя с OLAP-системой, пользователь может осуществлять гибкий просмотр информации, получать различные срезы данных, выполнять аналитические операции детализации, свертки, сквозного распределения, сравнения во времени. Вся работа с OLAP-системой происходит в терминах предметной области

Задачами OLAP систем является обобщение, агрегация, гиперкубическое представление информации и многомерный анализ. Это могут быть многомерные СУБД или же реляционные базы с предварительной агрегацией данных.

# Концептуальное многомерное представление данных

**MOLAP** - многомерное концептуальное представление. Являет собой множественную систему, состоящую из нескольких независимых измерений, вдоль которых могут быть проанализированы определенные совокупности данных. Одновременный анализ по нескольким измерениям определяется как многомерный анализ.

Исходные и агрегатные данные хранятся в многомерной базе данных. Хранение данных в многомерных структурах позволяет манипулировать данными как многомерным массивом, благодаря чему скорость вычисления агрегатных значений одинакова для любого из измерений. Однако в этом случае многомерная база данных оказывается избыточной, так как многомерные данные полностью содержат исходные реляционные данные.

# Практическое применение ИАД

**Интернет-технологии**. В системах электронного бизнеса, где особую важность имеют вопросы привлечения и удержания клиентов, технологии Data Mining применяются для построения рекомендательных систем интернет-магазинов и для решения проблемы персонализации посетителей сайтов. Рекомендации товаров и услуг, построенные на основе закономерностей в покупках клиентов, обладают огромной убеждающей силой. Статистика показывает, что почти каждый посетитель магазина Amazon не упускает возможности посмотреть на то, что же купили "Те, кто купил эту книгу, также купили ...". Персонализация клиентов или, другими словами, автоматическое распознание принадлежности клиента к определенной целевой аудитории позволяет компании проводить более гибкую маркетинговую политику. Поскольку в электронной коммерции деньги и платежные системы тоже электронные, то важной задачей становится обеспечение безопасности при операциях с пластиковыми карточками. Data Mining позволяет обнаруживать случаи мошенничества.

**Торговля**. Для успешного продвижения товаров всегда важно знать, что и как продается, а также кто является потребителем. Исчерпывающий ответ на первый вопрос дают такие средства Data Mining, как анализ рыночных корзин и сиквенциальный анализ. Зная связи между покупками и временные закономерности, можно

оптимальным образом регулировать предложение. С другой стороны, маркетинг имеет возможность непосредственно управлять спросом, но для этого необходимо знать как можно больше о потребителях - целевой аудитории маркетинга. Data Mining позволяет решать задачи выделения групп потребителей со схожими стереотипами поведения, т. е. сегментировать рынок. Для этого можно применять такие технологии Data Mining, как кластеризация и классификация. Сиквенциальный анализ помогает торговым предприятиям принимать решения о создании товарных запасов. Он дает ответы на вопросы типа "Если сегодня покупатель приобрел видеокамеру, то через какое время он вероятнее всего купит новые батарейки и пленку?".

**Медицина**. Примером использования Data Mining в медицине может служить построение диагностической системы или исследование эффективности хирургического вмешательства. Известно много экспертных систем для постановки медицинских диагнозов. Они построены главным образом на основе правил, описывающих сочетания различных симптомов отдельных заболеваний. С помощью таких правил узнают не только, чем болен пациент, но и как нужно его лечить. Правила помогают выбирать средства медикаментозного воздействия, определять показания/противопоказания, ориентироваться в лечебных процедурах, создавать условия наиболее эффективного лечения, предсказывать исходы назначенного курса лечения и т.п. Технологии Data Mining позволяют обнаруживать в медицинских данных шаблоны, составляющие основу указанных правил.

**Банковское дело**. Примером использования Data Mining в банковском деле может служить решение проблемы о возможной некредитоспособности клиентов банка. Если образ клиента в сознании банковского служащего соответствует его представлению о кредитоспособном клиенте, то кредит выдавать можно, иначе - отказать. По схожей схеме, но более продуктивно и полностью автоматически работают установленные в тысячах американских банков системы поддержки принятия решений со встроенной функциональностью Data Mining. Лишенные субъективной предвзятости, они опираются в своей работе только на историческую базу данных банка, где записывается детальная информация о каждом клиенте и, в конечном итоге, факт его кредитоспособности.

# Основные этапы интеллектуального анализа и обработки данных

* + отбор;
  + очистка;
  + обогащение;
  + кодирование;
  + извлечение данных.

# Отбор данных

**Отбор данных** - это выбор наблюдений по определенным критериям.

# Очистка данных

**Очистка данных** - процесс выявления и исправления ошибок, несоответствий данных с целью улучшения их качества. Некорректная, дублирующаяся или утраченная информация может стать причиной неадекватной статистики и неверных выводов в контексте бизнеса. Поэтому очистка данных является обязательной процедурой

# Обогащение данных

**Обогащение данных** - процесс насыщения данных новой информацией, которая позволяет сделать их более ценными и значимыми с точки зрения решения той или иной задачи анализа.

Существует два основных метода обогащения данных - внешнее и внутреннее.

**Внешнее** обогащение предполагает привлечение дополнительной информации из источников, которые находятся вне информационной системы предприятия. Таковыми могут быть другие предприятия и организации, финансово-кредитные учреждения, банки, страховые компании, государственные налоговые и статистические службы, органы государственного и муниципального управления, различные службы социальной сферы - миграционная служба, органы труда и занятости, система здравоохранения, пенсионные фонды и т.д.

Источником информации для обогащения данных могут быть любые организации, которые в процессе своей деятельности собирают, структурируют и хранят сведения, связанные с их деятельностью. Внешнее обогащение особенно

эффективно при проведении стратегического анализа, когда необходимо учесть все факторы и тенденции, действующие в бизнес-окружении предприятия.

**Внутренне** обогащение не предполагает привлечения какой-либо внешней информации. Оно обычно связано с получением и включением в набор данных полезной информации, которая отсутствует в явном виде, но может быть тем или иным способом получена с помощью манипуляций с имеющимися данными. Затем эта информация встраивается в виде новых полей или даже таблиц в хранилище данных и может быть использована для дальнейшего анализа. Для обогащения данных могут использоваться и результаты их анализа.

Например, при анализе клиентской базы оператора мобильной связи могут использоваться признаки, содержащие продолжительность разговоров в ночное, утреннее, дневное и вечернее время. Возможно, будет полезным обогатить данные полем, в котором будет среднесуточное потребление трафика клиентом. Поскольку среднее значение вычисляется только с использованием собственных данных, такое обогащение является внутренним.

# Кодирование данных

**Кодирование данных** - процесс построения данных из элементов конечного множества по установленным правилам.

# Извлечение знаний

**Извлечение знаний** - процесс взаимодействия инженера по знаниям с источником знаний (экспертом), в результате которого становятся явными процесс рассуждений специалистов при принятии решения и структура их представлений о предметной области.

# Индукция. Дедукция

В интеллектуальном анализе данных обсуждают два основных подхода извлечения практически полезных знаний - **дедуктивный** (на основе некоторой априори сформулированной гипотезы, от общего к частному) и **индуктивный** (на основе известных паттернов, от частного к общему).

Дедуктивный подход к исследованию данных предполагает наличие некоторой сформулированной гипотезы, подтверждение или опровержение которой после анализа данных позволяет получить некоторые частные сведения.

Индуктивный подход к исследованию данных позволяет сформулировать (скорректировать существующую) гипотезу и найти с ее помощью новые пути аналитических решений.

Для поиска значимых закономерностей порой требуется совместное попеременное использование индуктивного и дедуктивного подходов, при котором формируется такая среда, в которой модели не нужно быть исключительно статической или эмпирической. Вместо этого, модель непрерывно тестируется, модифицируется и улучшается до тех пор, пока не будет достаточно усовершенствована.

# Стадии ИАД

Data Mining может состоять из двух или трех стадий:

# Выявление закономерностей (свободный поиск, индуктивно).

На стадии свободного поиска осуществляется исследование набора данных с целью поиска скрытых закономерностей. Предварительные гипотезы относительно вида закономерностей здесь не определяются.

Закономерность - существенная и постоянно повторяющаяся взаимосвязь, определяющая этапы и формы процесса становления, развития различных явлений или процессов.

Система Data Mining на этой стадии определяет шаблоны, для получения которых в системах OLAP, например, аналитику необходимо обдумывать и создавать множество запросов. Здесь же аналитик освобождается от такой работы - шаблоны ищет за него система. Особенно полезно применение данного подхода в сверхбольших базах данных, где уловить закономерность путем создания запросов достаточно сложно, для этого требуется перепробовать множество разнообразных вариантов.

Свободный поиск представлен такими действиями:

* + выявление закономерностей условной логики;
  + выявление закономерностей ассоциативной;
  + выявление трендов и колебаний.

Допустим, имеется база данных кадрового агентства с данными о профессии, стаже, возрасте и желаемом уровне вознаграждения. В случае самостоятельного задания запросов аналитик может получить приблизительно такие результаты: средний желаемый уровень вознаграждения специалистов в возрасте от 25 до 35 лет равен 1200 условных единиц. В случае свободного поиска система сама ищет закономерности, необходимо лишь задать целевую переменную. В результате поиска закономерностей система сформирует набор логических правил "если ..., то ...".

Могут быть найдены, например, такие закономерности "Если возраст < 20 лет и желаемый уровень вознаграждения > 700 условных единиц, то в 75% случаев соискатель ищет работу программиста" или "Если возраст >35 лет и желаемый уровень вознаграждения > 1200 условных единиц, то в 90% случаев соискатель ищет руководящую работу". Целевой переменной в описанных правилах выступает профессия.

При задании другой целевой переменной, например, возраста, получаем такие правила: "Если соискатель ищет руководящую работу и его стаж > 15 лет, то возраст соискателя > 35 лет в 65 % случаев".

Описанные действия, в рамках стадии свободного поиска, выполняются при помощи:

* + индукции правил условной логики (задачи классификации и кластеризации, описание в компактной форме близких или схожих групп объектов);
  + индукции правил ассоциативной логики (задачи ассоциации и последовательности и извлекаемая при их помощи информация);
  + определения трендов и колебаний (исходный этап задачи прогнозирования).

На стадии свободного поиска также должна осуществляться валидация закономерностей, т.е. проверка их достоверности на части данных, которые не принимали участие в формировании закономерностей. Такой прием разделения данных на обучающее и проверочное множество часто используется в методах нейронных сетей и деревьев решений и будет описан в соответствующих лекциях.

# Использование выявленных закономерностей для предсказания неизвестных значений (прогностическое моделирование, дедуктивно).

Вторая стадия Data Mining - прогностическое моделирование - использует результаты работы первой стадии. Здесь обнаруженные закономерности используются непосредственно для прогнозирования.

Прогностическое моделирование включает такие действия:

* + предсказание неизвестных значений;
  + прогнозирование развития процессов.

В процессе прогностического моделирования решаются задачи классификации и прогнозирования.

При решении задачи классификации результаты работы первой стадии (индукции правил) используются для отнесения нового объекта, с определенной уверенностью, к одному из известных, предопределенных классов на основании известных значений.

При решении задачи прогнозирования результаты первой стадии (определение тренда или колебаний) используются для предсказания неизвестных (пропущенных или же будущих) значений целевой переменной (переменных).

Продолжая рассмотренный пример первой стадии, можем сделать следующий вывод.

Зная, что соискатель ищет руководящую работу и его стаж > 15 лет, на 65 % можно быть уверенным в том, что возраст соискателя > 35 лет. Или же, если возраст соискателя > 35 лет и желаемый уровень вознаграждения > 1200 условных единиц, на 90% можно быть уверенным в том, что соискатель ищет руководящую работу.

В дополнение к этим стадиям иногда вводят стадию валидации, следующую за стадией свободного поиска. Цель валидации - проверка достоверности найденных закономерностей. Однако, мы будем считать валидацию частью первой стадии, поскольку в реализации многих методов, в частности, нейронных сетей и деревьев решений, предусмотрено деление общего множества данных на обучающее и проверочное, и последнее позволяет проверять достоверность полученных результатов.

# Анализ исключений - стадия предназначена для выявления и объяснения аномалий, найденных в закономерностях.

Действие, выполняемое на этой стадии, - выявление отклонений (deviation detection). Для выявления отклонений необходимо определить норму, которая рассчитывается на стадии свободного поиска.

Вернемся к одному из примеров, рассмотренному выше.

Найдено правило "Если возраст > 35 лет и желаемый уровень вознаграждения > 1200 условных единиц, то в 90 % случаев соискатель ищет руководящую работу". Возникает вопрос - к чему отнести оставшиеся 10 % случаев?

Здесь возможно два варианта. Первый из них - существует некоторое логическое объяснение, которое также может быть оформлено в виде правила. Второй вариант для оставшихся 10% - это ошибки исходных данных. В этом случае стадия анализа исключений может быть использована в качестве очистки данных

# 36. Методы ИАД

Технологические методы Data Mining: все методы Data Mining подразделяются на две большие группы по принципу работы с исходными обучающими данными. В этой классификации верхний уровень определяется на основании того, сохраняются ли данные после Data Mining либо они дистиллируются для последующего использования.

# Непосредственное использование данных, или сохранение данных.

В этом случае исходные данные хранятся в явном детализированном виде и непосредственно используются на стадиях прогностического моделирования и/или анализа исключений. Проблема этой группы методов - при их использовании могут возникнуть сложности анализа сверхбольших баз данных.

Методы этой группы: кластерный анализ, метод ближайшего соседа, метод k- ближайшего соседа, рассуждение по аналогии.

# Выявление и использование формализованных закономерностей, или дистилляция шаблонов.

При технологии дистилляции шаблонов один образец (шаблон) информации извлекается из исходных данных и преобразуется в некие формальные конструкции, вид которых зависит от используемого метода Data Mining. Этот процесс выполняется на стадии свободного поиска, у первой же группы методов данная стадия в принципе отсутствует. На стадиях прогностического моделирования и анализа исключений используются результаты стадии свободного поиска, они значительно компактнее

самих баз данных. Напомним, что конструкции этих моделей могут быть трактуемыми аналитиком либо нетрактуемыми ("черными ящиками").

Методы этой группы: логические методы; методы визуализации; методы кросс- табуляции; методы, основанные на уравнениях.

# Классификация задач ИАД (Data Mining)

* + **классификация;**

**Краткое описание**. Наиболее простая и распространенная задача Data Mining. В результате решения задачи классификации обнаруживаются признаки, которые характеризуют группы объектов исследуемого набора данных - классы; по этим признакам новый объект можно отнести к тому или иному классу.

**Методы решения**. Для решения задачи классификации могут использоваться методы:

o ближайшего соседа;

o k-ближайшего соседа;

o байесовские сети;

o индукция деревьев решений;

o нейронные сети.

* + **кластеризация;**

**Краткое описание**. Кластеризация является логическим продолжением идеи классификации. Это задача более сложная, особенность кластеризации заключается в том, что классы объектов изначально не предопределены. Результатом кластеризации является разбиение объектов на группы.

**Пример метода** решения задачи кластеризации:

o обучение "без учителя" особого вида нейронных сетей - самоорганизующихся карт Кохонена.

* + **прогнозирование;**

**Краткое описание**. В результате решения задачи прогнозирования на основе особенностей исторических данных оцениваются пропущенные или же будущие значения целевых численных показателей.

Для решения таких задач широко применяются методы математической статистики, нейронные сети и др.

* + **ассоциация;**

**Краткое описание**. В ходе решения задачи поиска ассоциативных правил отыскиваются закономерности между связанными событиями в наборе данных.

Отличие ассоциации от двух предыдущих задач Data Mining: поиск закономерностей осуществляется не на основе свойств анализируемого объекта, а между несколькими событиями, которые происходят одновременно.

Наиболее **известный алгоритм** решения задачи поиска ассоциативных правил - алгоритм Apriori.

o предсказуемость;

o распространенность;

o ожидаемая предсказуемость.

* + **последовательные шаблоны;**

**Краткое описание**. Последовательность позволяет найти временные закономерности между транзакциями. Задача последовательности подобна ассоциации, но ее целью является установление закономерностей не между одновременно наступающими событиями, а между событиями, связанными во времени (т.е. происходящими с некоторым определенным интервалом во времени). Другими словами, последовательность определяется высокой вероятностью цепочки связанных во времени событий. Фактически, ассоциация является частным случаем последовательности с временным лагом, равным нулю. Эту задачу Data Mining также называют задачей нахождения последовательных шаблонов.

Правило последовательности: после события X через определенное время произойдет событие Y.

**Пример**. После покупки квартиры жильцы в 60% случаев в течение двух недель приобретают холодильник, а в течение двух месяцев в 50% случаев приобретается телевизор. Решение данной задачи широко применяется в маркетинге и менеджменте, например, при управлении циклом работы с клиентом.

* + **анализ отклонений**.

**Краткое описание**. Цель решения данной задачи - обнаружение и анализ данных, наиболее отличающихся от общего множества данных, выявление так называемых нехарактерных шаблонов.

# Стратегии Data Mining

Согласно классификации по стратегиям, задачи Data Mining подразделяются на следующие группы:

* + обучение с учителем;
  + обучение без учителя;
  + другие.

Категория обучение с учителем представлена следующими задачами Data Mining: классификация, оценка, прогнозирование.

Категория обучение без учителя представлена задачей кластеризации.

В категорию другие входят задачи, не включенные в предыдущие две стратегии.

Задачи Data Mining, в зависимости от используемых моделей, могут быть дескриптивными и прогнозирующими.

# Обучение с учителем

Категория обучение с учителем представлена следующими задачами Data Mining: классификация, оценка, прогнозирование.

При обучении с учителем для каждого обучающего входного примера требуется знание правильного ответа или функции оценки качества ответа. Такое обучение называют управляемым. Нейронной сети предъявляются значения входных и выходных сигналов, а она по определенному алгоритму подстраивает веса синаптических связей. В процессе обучения производится корректировка весов сети по результатам сравнения фактических выходных значений с входными, известными заранее.

# Обучение без учителя

Категория обучение без учителя представлена задачей кластеризации.

При обучении без учителя раскрывается внутренняя структура данных или корреляции между образцами в наборе данных. Выходы нейронной сети формируются самостоятельно, а веса изменяются по алгоритму, учитывающему только входные и

производные от них сигналы. Это обучение называют также неуправляемым. В результате такого обучения объекты или примеры распределяются по категориям, сами категории и их количество могут быть заранее не известны.

# Задача классификации

Классификация является наиболее простой и одновременно наиболее часто решаемой задачей Data Mining. Ввиду распространенности задач классификации необходимо четкое понимания сути этого понятия.

Приведем несколько определений.

**Классификация** - системное распределение изучаемых предметов, явлений, процессов по родам, видам, типам, по каким-либо существенным признакам для удобства их исследования; группировка исходных понятий и расположение их в определенном порядке, отражающем степень этого сходства.

**Классификация** - упорядоченное по некоторому принципу множество объектов, которые имеют сходные классификационные признаки (одно или несколько свойств), выбранных для определения сходства или различия между этими объектами.

Классификация требует соблюдения следующих правил:

* + в каждом акте деления необходимо применять только одно основание;
  + деление должно быть соразмерным, т.е. общий объем видовых понятий должен равняться объему делимого родового понятия;
  + члены деления должны взаимно исключать друг друга, их объемы не должны перекрещиваться;
  + деление должно быть последовательным. Различают:
  + вспомогательную (искусственную) классификацию, которая производится по внешнему признаку и служит для придания множеству предметов (процессов, явлений) нужного порядка;
  + естественную классификацию, которая производится по существенным признакам, характеризующим внутреннюю общность предметов и явлений. Она является результатом и важным средством научного исследования, т.к. предполагает и закрепляет результаты изучения закономерностей классифицируемых объектов.

В зависимости от выбранных признаков, их сочетания и процедуры деления понятий классификация может быть:

* + **простой** - деление родового понятия только по признаку и только один раз до раскрытия всех видов. Примером такой классификации является дихотомия, при которой членами деления бывают только два понятия, каждое из которых является противоречащим другому (т.е. соблюдается принцип: "А и не А");
  + **сложной** - применяется для деления одного понятия по разным основаниям и синтеза таких простых делений в единое целое. Примером такой классификации является периодическая система химических элементов.

Под классификацией будем понимать отнесение объектов (наблюдений, событий) к одному из заранее известных классов.

**Классификация** - это закономерность, позволяющая делать вывод относительно определения характеристик конкретной группы. Таким образом, для проведения классификации должны присутствовать признаки, характеризующие группу, к которой принадлежит то или иное событие или объект (обычно при этом на основании анализа уже классифицированных событий формулируются некие правила).

Классификация относится к стратегии обучения с учителем, которое также именуют контролируемым или управляемым обучением.

**Задачей классификации** часто называют предсказание категориальной зависимой переменной (т.е. зависимой переменной, являющейся категорией) на основе выборки непрерывных и/или категориальных переменных.

Классификация может быть одномерной (по одному признаку) и многомерной (по двум и более признакам).

# Методы классификации

Для классификации используются различные методы. Основные из них:

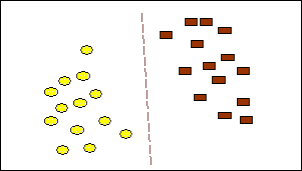
* + классификация с помощью деревьев решений;
  + байесовская (наивная) классификация;
  + классификация при помощи искусственных нейронных сетей;
  + классификация методом опорных векторов;
  + статистические методы, в частности, линейная регрессия;
  + классификация при помощи метода ближайшего соседа;
  + классификация CBR-методом;
  + классификация при помощи генетических алгоритмов.

# Метод опорных векторов

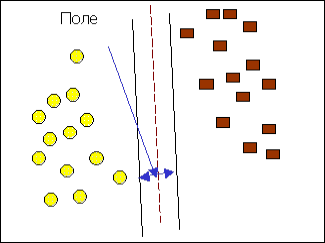
Метод опорных векторов относится к группе граничных методов. Она определяет классы при помощи границ областей.

При помощи данного метода решаются задачи бинарной классификации.

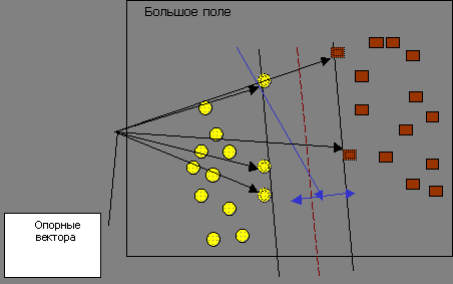
В основе метода лежит понятие плоскостей решений. Плоскость решения разделяет объекты с разной классовой принадлежностью.

На рисунке ниже приведен пример, в котором участвуют объекты двух типов. Разделяющая линия задает границу, справа от которой - все объекты типа brown (коричневый), а слева - типа yellow (желтый). Новый объект, попадающий направо, классифицируется как объект класса brown или - как объект класса yellow, если он расположился по левую сторону от разделяющей прямой. В этом случае каждый объект характеризуется двумя измерениями.

Цель метода опорных векторов - найти плоскость, разделяющую два множества объектов; такая плоскость показана на рисунке ниже. На этом рисунке множество образцов поделено на два класса: желтые объекты принадлежат классу А, коричневые - классу В.



Метод отыскивает образцы, находящиеся на границах между двумя классами, т.е. опорные вектора; они изображены на рисунке ниже.



Опорными векторами называются объекты множества, лежащие на границах областей.

Классификация считается хорошей, если область между границами пуста.

# Метод «ближайшего соседа»

**Метод «ближайшего соседа»** относится к классу методов, работа которых основывается на хранении данных в памяти для сравнения с новыми элементами. При появлении новой записи для прогнозирования находятся отклонения между этой записью и подобными наборами данных, и наиболее подобная (или ближний сосед) идентифицируется.

Например, при рассмотрении нового клиента банка, его атрибуты сравниваются со всеми существующими клиентами данного банка (доход, возраст и т.д.). Множество

«ближайших соседей» потенциального клиента банка выбирается на основании ближайшего значения дохода, возраста и т.д.

При таком подходе используется термин "k-ближайший сосед". Термин означает, что выбирается k "верхних" (ближайших) соседей для их рассмотрения в качестве множества "ближайших соседей". Поскольку не всегда удобно хранить все данные, иногда хранится только множество "типичных" случаев. В таком случае используемый метод называют рассуждением по аналогии (Case Based Reasoning, CBR), рассуждением на основе аналогичных случаев, рассуждением по прецедентам.

Преимущества метода

* + простота использования полученных результатов;
  + решения не уникальны для конкретной ситуации, возможно их использование для других случаев;
  + целью поиска является не гарантированно верное решение, а лучшее из возможных.

Недостатки метода "ближайшего соседа":

* + данный метод не создает каких-либо моделей или правил, обобщающих предыдущий опыт, - в выборе решения они основываются на всем массиве доступных исторических данных, поэтому невозможно сказать, на каком основании строятся ответы;
  + существует сложность выбора меры "близости" (метрики). От этой меры главным образом зависит объем множества записей, которые нужно хранить в памяти для достижения удовлетворительной классификации или прогноза. Также существует высокая зависимость результатов классификации от выбранной метрики;
  + при использовании метода возникает необходимость полного перебора обучающей выборки при распознавании, следствие этого - вычислительная трудоемкость;
  + типичные задачи данного метода - это задачи небольшой размерности по количеству классов и переменных.

# Прецедент

**Прецедент** - это описание ситуации в сочетании с подробным указанием действий, предпринимаемых в данной ситуации.

Подход, основанный на прецедентах, условно можно поделить на следующие этапы:

* + сбор подробной информации о поставленной задаче;
  + сопоставление этой информации с деталями прецедентов, хранящихся в базе, для выявления аналогичных случаев;
  + выбор прецедента, наиболее близкого к текущей проблеме, из базы прецедентов;
  + адаптация выбранного решения к текущей проблеме, если это необходимо;
  + проверка корректности каждого вновь полученного решения;
  + занесение детальной информации о новом прецеденте в базу прецедентов.

Таким образом, вывод, основанный на прецедентах, представляет собой такой метод анализа данных, который делает заключения относительно данной ситуации по результатам поиска аналогий, хранящихся в базе прецедентов.

Данный метод по своей сути относится к категории "обучение без учителя", т.е. является "самообучающейся" технологией, благодаря чему рабочие характеристики каждой базы прецедентов с течением времени и накоплением примеров улучшаются. Разработка баз прецедентов по конкретной предметной области происходит на естественном для человека языке, следовательно, может быть выполнена наиболее опытными сотрудниками компании - экспертами или аналитиками, работающими в данной предметной области.

Однако это не означает, что CBR-системы самостоятельно могут принимать решения. Последнее всегда остается за человеком, данный метод лишь предлагает возможные варианты решения и указывает на самый "разумный" с ее точки зрения.

# Деревья решений

**Метод деревьев** решений является одним из наиболее популярных методов решения задач классификации и прогнозирования. При помощи данного метода решаются задачи классификации и прогнозирования

Если зависимая, т.е. целевая переменная принимает дискретные значения, при помощи метода дерева решений решается задача классификации.

Если же зависимая переменная принимает непрерывные значения, то дерево решений устанавливает зависимость этой переменной от независимых переменных, т.е. решает задачу численного прогнозирования.

В наиболее простом виде **дерево решений** - это способ представления правил в иерархической, последовательной структуре. Основа такой структуры - ответы "Да" или "Нет" на ряд вопросов.

Бинарные деревья являются самым простым, частным случаем деревьев решений. В остальных случаях, ответов и, соответственно, ветвей дерева, выходящих из его внутреннего узла, может быть больше двух.

Качество построенного дерева решения весьма зависит от правильного выбора критерия расщепления. Над разработкой и усовершенствованием критериев работают многие исследователи.

Метод деревьев решений часто называют "наивным" подходом. Но благодаря целому ряду преимуществ, данный метод является одним из наиболее популярных для решения задач классификации.

Преимущества деревьев решений:

**Интуитивность деревьев решений**. Классификационная модель, представленная в виде дерева решений, является интуитивной и упрощает понимание решаемой задачи. Результат работы алгоритмов конструирования деревьев решений, в отличие, например, от нейронных сетей, представляющих собой "черные ящики", легко интерпретируется пользователем. Это свойство деревьев решений не только важно при отнесении к определенному классу нового объекта, но и полезно при интерпретации модели классификации в целом. Дерево решений позволяет понять и объяснить, почему конкретный объект относится к тому или иному классу.

Деревья решений дают возможность извлекать правила из базы данных на **естественном языке**. Пример правила: Если Возраст > 35 и Доход > 200, то выдать кредит.

Деревья решений позволяют создавать классификационные модели в тех областях, где аналитику достаточно сложно формализовать знания.

Алгоритм конструирования дерева решений **не требует от пользователя выбора входных атрибутов** (независимых переменных). На вход алгоритма можно подавать все существующие атрибуты, алгоритм сам выберет наиболее значимые среди них, и только они будут использованы для построения дерева. В сравнении, например, с нейронными сетями, это значительно облегчает пользователю работу, поскольку в нейронных сетях выбор количества входных атрибутов существенно влияет на время обучения.

**Точность** моделей, созданных при помощи деревьев решений, сопоставима с другими методами построения классификационных моделей (статистические методы, нейронные сети).

Разработан ряд **масштабируемых алгоритмов**, которые могут быть использованы для построения деревьев решения на сверхбольших базах данных;

масштабируемость здесь означает, что с ростом числа примеров или записей базы данных время, затрачиваемое на обучение, т.е. построение деревьев решений, растет линейно.

**Быстрый процесс обучения**. На построение классификационных моделей при помощи алгоритмов конструирования деревьев решений требуется значительно меньше времени, чем, например, на обучение нейронных сетей.

# Регрессионный анализ

**Регрессионный анализ** позволяет обнаружить лишь зависимости, а не связи, лежащие в основе этих зависимостей. Регрессионный анализ дает возможность оценить степень связи между переменными путем вычисления предполагаемого значения переменной на основании нескольких известных значений.

Этапы регрессионного анализа:

1. формулировка задачи. На этом этапе формируются предварительные гипотезы о зависимости исследуемых явлений;
2. определение зависимых и независимых (объясняющих) переменных;
3. сбор статистических данных. Данные должны быть собраны для каждой из переменных, включенных в регрессионную модель;
4. формулировка гипотезы о форме связи (простая или множественная, линейная или нелинейная);
5. определение функции регрессии (заключается в расчете численных значений параметров уравнения регрессии);
6. оценка точности регрессионного анализа;
7. интерпретация полученных результатов. Полученные результаты регрессионного анализа сравниваются с предварительными гипотезами. Оценивается корректность и правдоподобие полученных результатов;
8. предсказание неизвестных значений зависимой переменной.

При помощи регрессионного анализа возможно решение задачи прогнозирования и классификации. Прогнозные значения вычисляются путем подстановки в уравнение регрессии параметров значений объясняющих переменных. Решение задачи классификации осуществляется таким образом: линия регрессии делит все множество

объектов на два класса, и та часть множества, где значение функции больше нуля, принадлежит к одному классу, а та, где оно меньше нуля, - к другому классу.

Задачи регрессионного анализа:

* + установление формы зависимости;
  + определение функции регрессии;
  + оценка неизвестных значений зависимой переменной.

Характер и форма зависимости между переменными могут образовывать следующие разновидности регрессии:

* + положительная линейная регрессия (выражается в равномерном росте функции);
  + положительная равноускоренно возрастающая регрессия;
  + положительная равнозамедленно возрастающая регрессия;
  + отрицательная линейная регрессия (выражается в равномерном падении функции);
  + отрицательная равноускоренно убывающая регрессия;
  + отрицательная равнозамедленно убывающая регрессия.

# Задача поиска ассоциативных правил

Целью поиска ассоциативных правил является нахождение закономерностей между связанными событиями в базах данных.

Очень часто покупатели приобретают не один товар, а несколько. В большинстве случаев между этими товарами существует взаимосвязь. Так, например, покупатель, приобретающий макаронные изделия, скорее всего, захочет приобрести также кетчуп. Эта информация может быть использована для размещения товара на прилавках.

# Алгоритм Apriori

**Алгоритм Apriori** - алгоритм поиска ассоциативных правил.

Работа данного алгоритма состоит из нескольких этапов, каждый из этапов состоит из следующих шагов:

формирование кандидатов этап, на котором алгоритм, сканируя базу данных, создает множество i-элементных кандидатов (i - номер этапа). На этом этапе поддержка кандидатов не рассчитывается;

* + подсчет кандидатов - этап, на котором вычисляется поддержка каждого i- элементного кандидата. Здесь же осуществляется отсечение кандидатов, поддержка которых меньше минимума, установленного пользователем. Оставшиеся i-элементные наборы называем часто встречающимися.

# Задача прогнозирования

**Прогнозирование** - установление функциональной зависимости между зависимыми и независимыми переменными.

Задачи прогнозирования решаются в самых разнообразных областях человеческой деятельности, таких как наука, экономика, производство и множество других сфер. Прогнозирование является важным элементом организации управления как отдельными хозяйствующими субъектами, так и экономики в целом.

Задача прогнозирования, пожалуй, может считаться одной из наиболее сложных задач Data Mining, она требует тщательного исследования исходного набора данных и методов, подходящих для анализа.

**Прогнозирование** в широком понимании этого слова, определяется как опережающее отражение будущего. Целью прогнозирования является предсказание будущих событий.

**Прогнозирование** является одной из задач Data Mining и одновременно одним из ключевых моментов при принятии решений.

**Прогнозирование** направлено на определение тенденций динамики конкретного объекта или события на основе ретроспективных данных, т.е. анализа его состояния в прошлом и настоящем. Таким образом, решение задачи прогнозирования требует некоторой обучающей выборки данных.

# Задача прогнозирования рынков

* + прогноз продаж товаров (например, с целью определения нормы товарного запаса);
  + прогнозирование продаж товаров, оказывающих влияние друг на друга;
  + прогноз продаж в зависимости от внешних факторов.

В самых общих чертах решение задачи прогнозирования сводится к решению таких подзадач:

* + выбор модели прогнозирования;
  + анализ адекватности и точности построенного прогноза.

# Задача кластеризации

Задача кластеризации сходна с задачей классификации, является ее логическим продолжением, но ее отличие в том, что классы изучаемого набора данных заранее не предопределены.

Кластеризация предназначена для разбиения совокупности объектов на однородные группы (кластеры или классы). Если данные выборки представить как точки в признаковом пространстве, то задача кластеризации сводится к определению "сгущений точек".

Цель кластеризации - поиск существующих структур.

**Задачи кластерного анализа** можно объединить в следующие группы:

1. Разработка типологии или классификации.
2. Исследование полезных концептуальных схем группирования объектов.
3. Представление гипотез на основе исследования данных.
4. Проверка гипотез или исследований для определения, действительно ли типы (группы), выделенные тем или иным способом, присутствуют в имеющихся данных.

Кластеризация является описательной процедурой, она не делает никаких статистических выводов, но дает возможность провести разведочный анализ и изучить "структуру данных".

Кластер можно охарактеризовать как группу объектов, имеющих общие свойства. Характеристики кластера:

* + внутренняя однородность;
  + внешняя изолированность.

# Методы кластеризации

Методы кластерного анализа можно разделить на две группы:

* + иерархические;
  + неиерархические.

Каждая из групп включает множество подходов и алгоритмов.

# Иерархические агломеративные методы

Эта группа методов характеризуется последовательным объединением исходных элементов и соответствующим уменьшением числа кластеров.

В начале работы алгоритма все объекты являются отдельными кластерами. На первом шаге наиболее похожие объекты объединяются в кластер. На последующих шагах объединение продолжается до тех пор, пока все объекты не будут составлять один кластер.

# Иерархические дивизимные (делимые) методы

Эти методы являются логической противоположностью агломеративным методам. В начале работы алгоритма все объекты принадлежат одному кластеру, который на последующих шагах делится на меньшие кластеры, в результате образуется последовательность расщепляющих групп.

Иерархические методы кластеризации различаются правилами построения кластеров. В качестве правил выступают критерии, которые используются при решении вопроса о "схожести" объектов при их объединении в группу (агломеративные методы) либо разделения на группы (дивизимные методы).

Иерархические методы кластерного анализа используются при небольших объемах наборов данных.

Преимуществом иерархических методов кластеризации является их наглядность.

# Дендрограммы

Иерархические алгоритмы связаны с построением дендрограмм, которые являются результатом иерархического кластерного анализа. Дендрограмма описывает близость отдельных точек и кластеров друг к другу, представляет в графическом виде последовательность объединения (разделения) кластеров.

**Дендрограмма** - древовидная диаграмма, содержащая n уровней, каждый из которых соответствует одному из шагов процесса последовательного укрупнения кластеров.

Дендрограмму также называют древовидной схемой, деревом объединения кластеров, деревом иерархической структуры

Дендрограмма представляет собой вложенную группировку объектов, которая изменяется на различных уровнях иерархии.

# Алгоритм k-средних

Наиболее распространен среди неиерархических методов алгоритм k-средних. В отличие от иерархических методов, которые не требуют предварительных предположений относительно числа кластеров, для возможности использования этого метода необходимо иметь гипотезу о наиболее вероятном количестве кластеров.

Алгоритм k-средних строит k кластеров, расположенных на возможно больших расстояниях друг от друга. Основной тип задач, которые решает алгоритм k-средних, - наличие предположений (гипотез) относительно числа кластеров, при этом они должны быть различны настолько, насколько это возможно. Выбор числа k может базироваться на результатах предшествующих исследований, теоретических соображениях или интуиции.

Общая идея алгоритма: заданное фиксированное число k кластеров наблюдения сопоставляются кластерам так, что средние в кластере (для всех переменных) максимально возможно отличаются друг от друга.

Описание алгоритма:

# Первоначальное распределение объектов по кластерам.

Выбирается число k, и на первом шаге эти точки считаются "центрами" кластеров. Каждому кластеру соответствует один центр.

Выбор начальных центроидов может осуществляться следующим образом:

* + выбор k-наблюдений для максимизации начального расстояния;
  + случайный выбор k-наблюдений;
  + выбор первых k-наблюдений.

В результате каждый объект назначен определенному кластеру.

# Итеративный процесс.

Вычисляются центры кластеров, которыми затем и далее считаются покоординатные средние кластеров. Объекты опять перераспределяются.

Процесс вычисления центров и перераспределения объектов продолжается до тех пор, пока не выполнено одно из условий:

* + кластерные центры стабилизировались, т.е. все наблюдения принадлежат кластеру, которому принадлежали до текущей итерации;
  + число итераций равно максимальному числу итераций.

# Прогнозирование временных рядов

Основой для прогнозирования служит историческая информация, хранящаяся в базе данных в виде временных рядов.

**Временной ряд** - последовательность наблюдаемых значений какое-либо признака, упорядоченных в неслучайные моменты времени.

* + члены временного ряда, в отличие от элементов случайной выборки, не являются статистически независимыми;
  + члены временного ряда не являются одинаково распределенными.

Приведем два принципиальных отличия временного ряда от простой последовательности наблюдений:

* + Члены временного ряда, в отличие от элементов случайной выборки, не являются статистически независимыми.
  + Члены временного ряда не являются одинаково распределенными

Отличием анализа временных рядов от анализа случайных выборок является предположение о равных промежутках времени между наблюдениями и их хронологический порядок. Привязка наблюдений ко времени играет здесь ключевую роль, тогда как при анализе случайной выборки она не имеет никакого значения.

Информация, накопленная в разнообразных базах данных предприятия, является временными рядами, если она расположена в хронологическом порядке и произведена в последовательные моменты времени.

Анализ временного ряда осуществляется с целью:

* + определения природы ряда;
  + прогнозирования будущих значений ряда.

# Тренд, сезонность, цикл

Основными составляющими временного ряда являются тренд и сезонная компонента. Составляющие этих рядов могут представлять собой либо тренд, либо сезонную компоненту.

**Тренд** - неслучайная функция, которая формируется под действием общих или долговременных тенденций, влияющих на временной ряд.

Существует большое разнообразие постановок задач прогнозирования, которое можно подразделить на две группы:

* + прогнозирование односерийных рядов;
  + прогнозирование мультисерийных, или взаимовлияющих, рядов.

Кроме деления на классы по односерийности и многосерийности, ряды также бывают сезонными и несезонными.

**Свойство сезонности** означает, что через примерно равные промежутки времени форма кривой, которая описывает проведение зависимой переменной, повторяет свои характерные очертания.

**Циклическая** компонента временного ряда описывает относительно длительные периоды подъема и спада. Она состоит из циклов, которые меняются по амплитуде и протяжённости.

Отличия циклической компоненты от сезонной:

* + продолжительность цикла, как правило, больше, чем один сезонный период.
  + циклы, в отличие от сезонных периодов, не имеют определенной продолжительности.

# Период, горизонт, интервал прогнозирования

**Период прогнозирования** - основная единица времени, на которую делается прогноз.

**Горизонт прогнозирования** - это число периодов в будущем, которые покрывает прогноз.

**Интервал прогнозирования** - частота, с которой делается новый прогноз. Интервал прогнозирования может совпадать с периодом прогнозирования.

# Виды прогнозов

* + **краткосрочный прогноз** представляет собой прогноз на несколько шагов вперед, т.е. осуществляется построение прогноза не более чем на 3% от объема наблюдений или на 1-3 шага вперед.;
  + **среднесрочный прогноз** - это прогноз на 3-5% от объема наблюдений, но не более 7-12 шагов вперед; также под этим типом прогноза понимают прогноз на один или половину сезонного цикла. Для построения краткосрочных и среднесрочных прогнозов вполне подходят статистические методы.;
  + **долгосрочный прогноз** - это прогноз более чем на 5% от объема наблюдений. При построении данного типа прогнозов статистические методы практически не используются, кроме случаев очень "хороших" рядов, для которых прогноз можно просто "нарисовать".

# Виды визуализации результатов анализа данных

**Визуализация** - это инструментарий, который позволяет увидеть конечный результат вычислений, организовать управление вычислительным процессом и даже вернуться назад к исходным данным, чтобы определить наиболее рациональное направление дальнейшего движения.

Визуализации данных может быть представлена в виде: графиков, схем, гистограмм, диаграмм и т.д.

Кратко роль визуализации можно описать такими ее возможностями:

* + поддержка интерактивного и согласованного исследования;
  + помощь в представлении результатов;
  + использование глаз (зрения), чтобы создавать зрительные образы и осмысливать

их.

Методы визуализации, в зависимости от количества используемых измерений,

принято классифицировать на две группы

# представление данных в одном, двух и трех измерениях:

К этой группе методов относятся хорошо известные способы отображения информации, которые доступны для восприятия человеческим воображением.

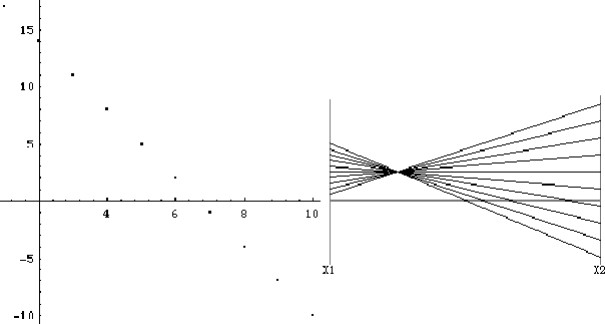
Практически любой современный инструмент Data Mining включает способы визуального представления из этой группы.

# представление данных в четырех и более измерениях:

Представления информации в четырехмерном и более измерениях недоступны для человеческого восприятия. Однако разработаны специальные методы для возможности отображения и восприятия человеком такой информации.

Наиболее известные способы многомерного представления информации:

# параллельные координаты:

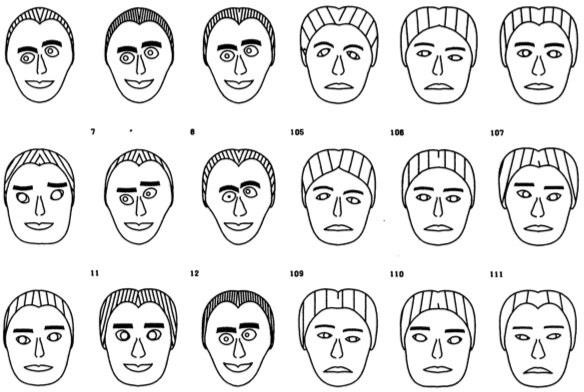
В параллельных координатах переменные кодируются по горизонтали, вертикальная линия определяет значение переменной.

# «лица Чернова»:

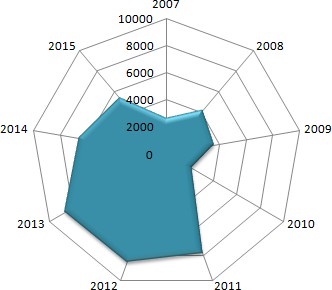
Основная идея представления информации в "лицах Чернова" состоит в кодировании значений различных переменных в характеристиках или чертах человеческого лица.

Для каждого наблюдения рисуется отдельное "лицо". На каждом "лице" относительные значения переменных представлены как формы и размеры отдельных черт лица (например, длина и ширина носа, размер глаз, размер зрачка, угол между бровями).

Анализ информации при помощи такого способа отображения основан на способности человека интуитивно находить сходства и различия в чертах лица.



# лепестковые диаграммы:

Лепестковые диаграммы используются для сравнения статистических значений нескольких рядов данных. Данный вид диаграммы отражается в виде криволинейного многоугольника с векторными линиями, выходящими из общего центра. Данные векторные линии являются осями координат по каждой их категорий.

# Уровни интеграции данных

* + интеграция данных **на физическом** уровне (конверсия данных из различных источников в требуемый единый формат их физического предоставления);
  + интеграция данных на **логическом** уровне (доступ к данным, содержащимся в различных источниках, в терминах единой глобальной схемы свойств данных);
  + интеграция данных на **семантическом** уровне (единое представление данных с учетом их семантических свойств в контексте единой онтологии предметной области).

# ETL

Средства **ETL (Extraction, Transformation, Loading)** обеспечивают возможность сложных преобразований и большей части технологического процесса преобразования и очистки данных. Общей проблемой средств ETL являются ограниченные за счет собственных API и форматов метаданных возможности взаимодействия, усложняющие совместное использование различных средств.

* + извлечение и очистка данных;
  + трансформация данных;
  + загрузка данных.

# Проблемы интеграции данных Сложности интеграции данных:

* + **физический** уровень - разные форматы;
  + **логический** уровень - разные модели/схемы данных;
  + **семантический** уровень - разные онтологии/понятийная система.

# Проблемы интеграции данных:

* + разнородность (гетерогенность);
  + автономность;
  + распределенность.

# Методы интеграции/объединения данных

* + консолидация;
  + федерализация;
  + распространение;
  + SOA;
  + семантическая интеграция.

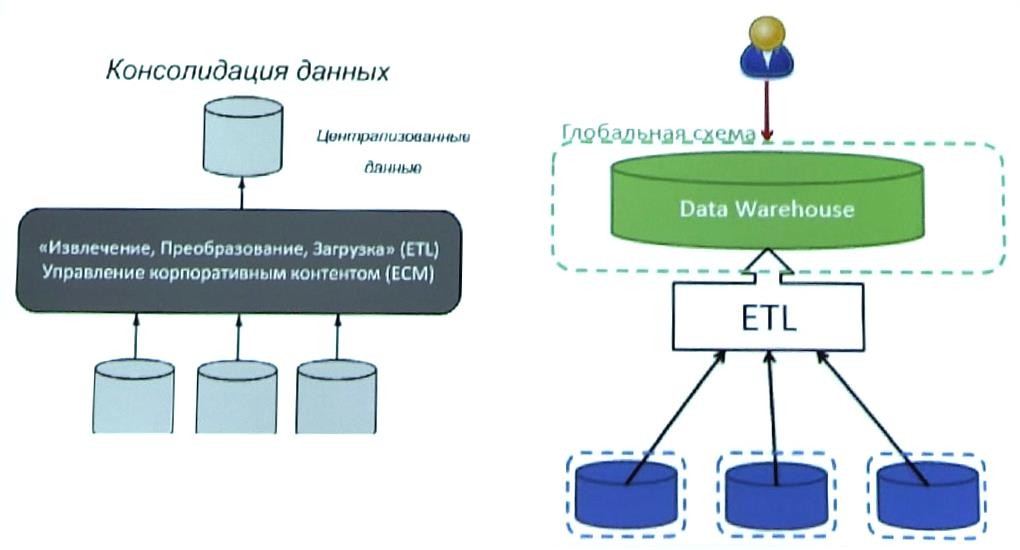
# Консолидация

**Консолидация** - сбор данных из нескольких первичных систем и их объединение в одном постоянном месте хранения. Хранилище данных позволяет хранить огромные массивы данных и может быть использовано в качестве базы для построения различных управленческих систем и для подготовки разных форм отчетности.

В случае консолидации данные извлекаются из источников, и помещаются в Хранилище данных. Процесс заполнения Хранилища состоит из трех фаз - извлечение, преобразование, загрузка (Extract, Transformation, Loading - ETL). Во многих случаях именно **ETL** понимают под термином «интеграция данных». Еще одна распространенная технология консолидации данных - управление содержанием корпорации (enterprise content management - ECM). Большинство решений **ECM** направлены на консолидацию и управление неструктурированными данными, такими как документы, отчеты и web-страницы.

Консолидация - **однонаправленный процесс**, то есть данные из нескольких источников сливаются в Хранилище, но не распространяются из него обратно в распределенную систему. Часто консолидированные данные служат основой для приложений бизнес-аналитики, OLAP-приложений.

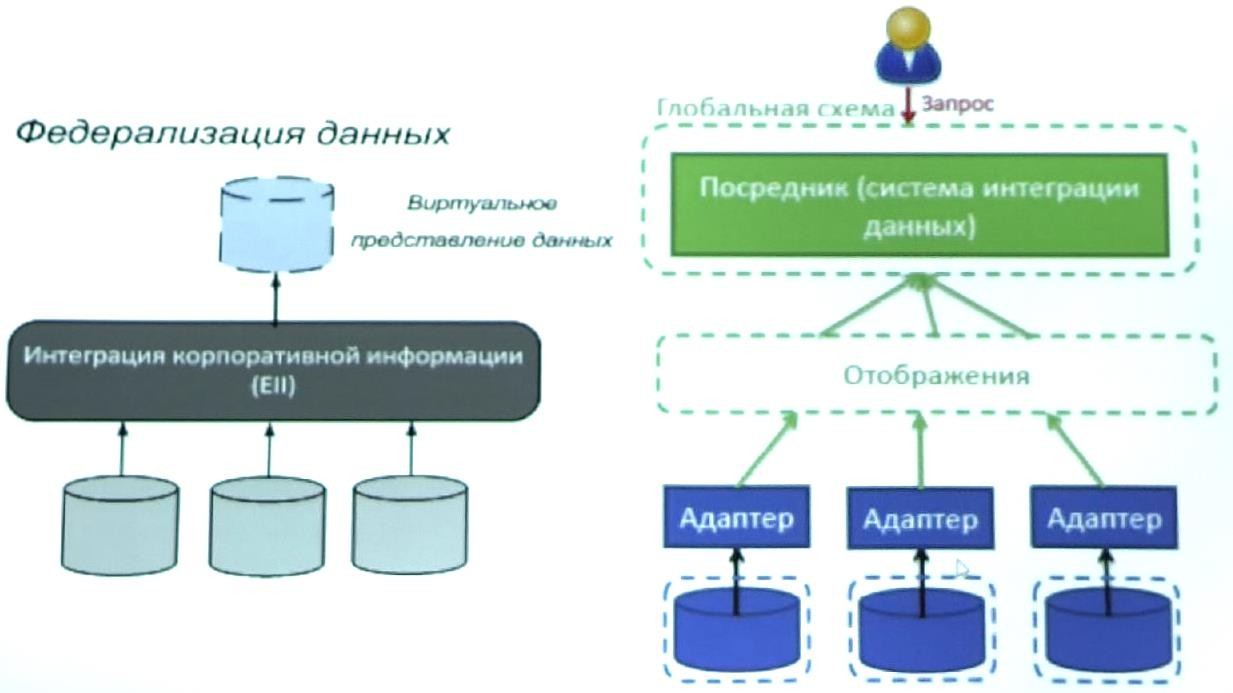
При использовании этого метода обычно существует некоторая задержка между моментом обновления информации в первичных системах и временем, когда данные изменения появляются в конечном месте хранения. Конечные места хранения данных, содержащие данные с большими временами отставания (например, более одного дня), создаются с помощью пакетных приложений интеграции данных, которые извлекают данные из первичных систем с определенными, заранее заданными интервалами. Конечные места хранения данных с небольшим отставанием обновляются с помощью оперативных приложений интеграции данных, которые постоянно отслеживают и передают изменения данных из первичных систем в конечные места хранения.



# Федерализация

**Федерализация** - это обеспечение единой виртуальной картины одного или нескольких источников исходных данных. Для реализации данного подхода используется технология интеграции корпоративной информации (Enterprise Information Integration), которая позволяет извлекать данные из различных источников данных, объединять их и представлять аналитику в режиме реального времени.

Отображение данных из источника в общую модель выполняется при каждом запросе специальной оболочкой. Для этого необходима интерпретация запроса к отдельным источникам и последующее отображение полученных данных в единую модель.



# Распространение

**Распространение** - данные физически копируются из одного места в другое, пока не попадут в некоторую целевую систему. Такие приложения обычно работают в оперативном режиме и производят перемещение данных к местам назначения, т. е. зависят от определенных событий. Обновления в первичной системе могут передаваться в конечную систему синхронно или асинхронно, причем метод распространения гарантирует доставку данных в систему назначения, что является ключевым признак распространения данных.



# Сервисно-ориентированная архитектура

**SOA** - это прикладная архитектура, в которой все функции определены как независимые сервисы с вызываемыми интерфейсами. Обращение к этим сервисам в определенной последовательности позволяет реализовать тот или иной бизнес-процесс.

Под **сервис-ориентированной архитектурой** понимается подход к проектированию прикладных информационных систем, который руководствуется следующими принципами:

* + явное отделение бизнес-логики прикладной системы от логики презентации информации;
  + реализация бизнес-логики прикладной системы в виде некоторого количества программных модулей (сервисов), которые доступны извне (пользователям и другим модулям), чаще всего в режиме "запрос-ответ", через четко определенные формальные интерфейсы доступа;
  + при этом "потребитель услуги", который может быть прикладной системой или другим сервисом, имеет возможность вызвать сервис через интерфейсы, используя соответствующие коммуникационные механизмы.

В целом, SOA представляет собой модель взаимодействия компонент, которая связывает различные функциональные модули приложений (сервисы) между собой с

помощью четко определяемых интерфейсов. Заметим, что одним из известных интеграционных шаблонов как раз и является "Сервис". Интерфейсы сами по себе не зависят от используемых аппаратных платформ, операционных систем или языков программирования, используемых для разработки этих приложений. Это позволяет отдельным сервисам взаимодействовать между собой одним и тем же стандартным, но в то же время универсальным способом. Такая особенность использования интерфейса, независимого от окружения и платформы, получила название модели "слабой связи". Ее очевидным преимуществом является повышенная гибкость и адаптируемость, поскольку замена или модернизация одной из компонент системы не сказывается на остальных.

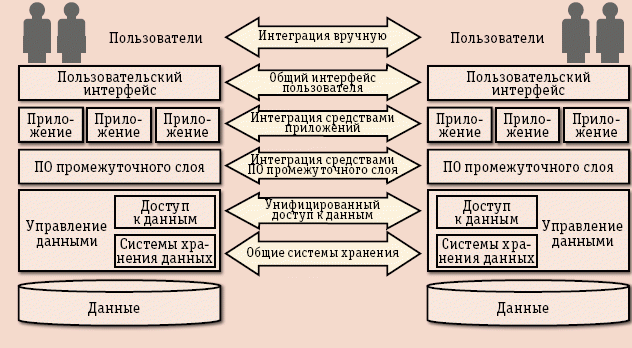
Эта модель состоит из следующих основных компонент:

* + **презентационный уровень** описывает интерфейсные сервисы для взаимодействия пользователей с информационной системой, включая корпоративные и публичные порталы, доступ с мобильных устройств, а также различные преобразования информации при взаимодействии с внешними системами и устройствами;
  + на уровне **бизнес-сервисов** формируются модели и осуществляется управление выполнением бизнес-процессов предприятия с использованием специализированных средств (типа BPEL), а также координация автоматизированных и "ручных" операций;
  + **интеграционные сервисы** обеспечивают взаимодействие между приложениями, которое может быть реализовано, в частности, с использованием средств обмена сообщениями или в рамках единой среды исполнения, такой как сервер приложений J2EE;
  + **сервисы уровня данных** реализуют средства извлечения и повторного использования данных из СУБД и приложений. Явное выделение такого уровня позволяет изолировать вышестоящие компоненты архитектуры от изменений в технологиях (например, смены вендора или версии продукта), а также обеспечить единый унифицированный подход к выполнению операций с данными;
  + **уровень инфраструктуры**, **приложений и СУБД** является как бы основой для всей структуры, и именно здесь концентрируются основные инвестиции в ИТ.



# Классификация синтаксической интеграции данных

**Синтаксис** - обеспечивает единообразие доступа к данным, но ни о какой интеграции информации предприятия или о создании «одной правды» речи быть не может.



* + **общие системы хранения**. Осуществляется за счет слияния данных из разных систем хранения данных в одну общую. Сегодня мы бы объединили эти два уровня в один и назвали бы его виртуализацией систем хранения;
  + **унифицированный доступ к данным**. На этом уровне осуществляется логическая интеграция данных, различные приложения получают единообразное видение физически распределенных данных. Такая виртуализация данных имеет свои несомненные достоинства, но гомогенизация данных в процессе работы с ними требует значительных ресурсов;
  + **интеграция средствами ПО промежуточного слоя**. ПО этого слоя играет посредническую роль, его составляющие способны к выполнению отдельных предписанных им функций, в полном объеме интеграционная задача решается во взаимодействии с приложениями;
  + **интеграция средствами приложени**й. Обеспечивает доступ к разным источникам данных и возвращает пользователю обобщенные результаты. Сложность интеграции на этом уровне объясняется большим разнообразием интерфейсов и форматов данных;
  + **общий пользовательский интерфейс**. Дает возможность единообразного доступа к данным, например, с помощью браузера, но при этом данные остаются неинтегрированными и неоднородными;
  + **интеграция вручную**. Пользователь сам объединяет данные, применяя различные типы интерфейсов и языки запросов.

# Универсальные форматы обмена данными

**XML** изначально придумали для более удобного хранения и передачи данных, в том числе через Интернет.

У него есть ряд преимуществ, которые позволяют успешно справляться с этой задачей. Во-первых, он легко читается и человеком, и компьютером. Во-вторых, поскольку данные хранятся в простом текстовом формате, при их передаче с одного компьютера на другой не возникнет никаких проблем с совместимостью.

Важно понимать, что XML - это не исполняемый код, а язык описания данных. После того, как ты описал данные с помощью XML, тебе нужно написать код (например, на Java), который сможет эти данные отправить/принять/обработать.

# Семантическая интеграция

**Семантика** - подразумевает, что информационные системы обмениваются данными в такой форме, которая полностью абстрагирована от внутренней структуры каждой из них.

Семантическая интеграция основывается на знании и учете природы данных. Разумеется, хранение данных вместе с метаданными создает дополнительные сложности, но обеспечивает большее удобство работы.

# Онтологии

**Онтология** - формальная спецификация разделяемой концептуальной модели

# область применения

* + - Semantic Web

o управление знаниями

o обучение

o экспертные системы

o обработка естественного языка

o интеграция пользовательских приложений

# роль

o коммуникация между людьми

o автоматизированные рассуждения

o способ представления знаний

o интеграция разнородной информации

# владелец

o индивидуум

o группа

o организация

o нация (государство)

o человечество

# тип

o по уровню обобщения

* онтологии представления
* общие онтологии
* онтологии предметной области
* онтологии приложений

o по глубине проработки

* каталог
* глоссарий
* тезаурус
* коллекция таксономий
* коллекция фреймов
* набор логических ограничений

o по предметной области

* организация (предприятие)
* раздел науки
* отрасль промышленности
* материальные и нематериальные объекты

# язык

o формальный

o формализованный

o неформальный

# Анализ текстовой информации

**Интеллектуальный анализ информации (Text Mining)** - направление в искусственном интеллекте, целью которого является получение информации из коллекций текстовых документов, основываясь на применении эффективных в практическом плане методов машинного обучения и обработки естественного языка. Название «интеллектуальный анализ текстов» перекликается с понятием

«интеллектуальный анализ данных», что выражает схожесть их целей, подходов к переработке информации и сфер применения; разница проявляется лишь в конечных

методах, а также в том, что ИАД имеет дело с хранилищами и базами данных, а не электронными библиотеками и корпусами текстов.

# Задачи Text Mining

* + извлечение из текста характерных элементов или признаков, которые могут использоваться в качестве ключевых слов, метаданных, аннотации;
  + отнесение документов к некоторым категориям из заданной схемы их систематизации.

# Стратегии поиска информации

**Информационный поиск** - один из основных информационных процессов

**Информационный поиск** - отыскание информации слабо структурированного типа, отвечающей информационной потребности, среди большого объема информации. Информации в данном случае присуще то, что она обычно представлена в виде текстовых документов и хранится в электронном виде.

**Поисковый образ документа** - это выраженное в терминах формализованного информационно-поискового языка основное смысловое содержание этого документа.

# Полнотекстовый поиск

**Полнотекстовый поиск** - автоматизированный поиск документов, при котором поиск ведётся не по именам документов, а по их содержимому, всему или существенной части.

# Поиск по метаданным

**Описательные метаданные** - это дополнительные данные, которые помогают изучать содержимое книг и файлов для их дальнейшего сбора или группирования. Например, описательные метаданные научной статьи будут включать такие сведения, как заголовок или имя автора публикации, аннотацию, используемые в работе ключевые слова и т.д.

Как правило, объем метаданных гораздо меньше объема самих данных. Поиск по метаданным существенно проще, чем поиск по содержанию.

# Оценка эффективности поиска

Для оценки эффективности информационных систем, содержащих текстовую информацию, очень часто используют семантические показатели, основанные на оценке релевантности между документами и запросами.

**Релевантность** - устанавливаемое при информационном поиске соответствие содержания документа информационному запросу или поискового образа документа поисковому предписанию.

**Структурирование** - потеря полезной информации.

**Информационная потребность** - это некий набор данных, необходимый пользователю для того, чтобы больше узнать об интересующей его предметной области.

# Релевантность

**Релевантность** - устанавливаемое при информационном поиске соответствие содержания документа информационному запросу или поискового образа документа поисковому предписанию.

Как правило, в любой поисковой системе по запросу выдается несколько найденных документов. Многие из них могут повествовать не о том. И наоборот, некоторые важные, релевантные, документы могут быть пропущены при поиске. Ясно, что количество и тех и других определяет качество поиска, которое можно определить достаточно точно. Основными понятиями в мире поисковых средств являются точность выдачи (информационный шум) и полнота выдачи (потери информации).

# Хэширование

**Хэширование** представляет собой преобразование любого объема информации в уникальный набор символов, который присущ только этому массиву входящей информации. Этот набор символов и будет называться хэшем.

У хэш-функции есть несколько обязательных свойств:

* + хэш всегда уникален для каждого массива информации. Однако иногда случаются так называемые коллизии, когда для разных входных блоков информации вычисляются одинаковые хэш-коды;
  + при самом незначительном изменении входной информации ее хэш полностью меняется;
  + хэш-функция необратима и не позволяет восстанавливать исходный массив информации из символьной строки. Это можно сделать, только перебрав все возможные варианты, что при бесконечном количестве информации требует много времени и денег;
  + хэширование позволяет достаточно быстро вычислить нужный хэш для достаточно большого объема информации;
  + алгоритм работы хэш-функции, как правило, делается открытым, чтобы при необходимости можно было оценить ее стойкость к восстановлению начальных данных по выдаваемому хэшу;
  + хэш-функция должна уметь приводить любой объем данных к числу заданной длины.

# Задачи Web Mining

Web Mining можно перевести как "добыча данных в Web". Способность определять интересы и предпочтения каждого посетителя, наблюдая за его поведением, является серьезным и критичным преимуществом конкурентной борьбы на рынке электронной коммерции.

Системы Web Mining могут ответить на многие вопросы, например, кто из посетителей является потенциальным клиентом Web-магазина, какая группа клиентов Web-магазина приносит наибольший доход, каковы интересы определенного посетителя или группы посетителей.

**Задачи** Web Mining:

* + предварительная обработка данных для Web Mining;
  + обнаружение шаблонов и открытие знаний с использованием ассоциативных правил, временных последовательностей, классификации и кластеризации;
  + анализ полученного знания.

# SEO

**SEO (Search Engine Optimization)** - это технология раскрутки сайта в поисковой выдаче с целью получения трафика. В основе работы лежит знание алгоритмов ранжирования ресурсов в выдаче и улучшение критериев сайта с целью выведения его в топ результатов поиска по требуемым ключевым фразам.

Принцип работы выглядит очень просто: пользователь вводит интересующий его запрос, а поисковая система строит список ссылок на ресурсы в определенном порядке. После этого потенциальный клиент знакомится с кратким описанием ресурса и переходит на него, если это его заинтересовало. Чем выше в результатах находится такая ссылка, тем вероятнее переход, поскольку пользователь обычно не заходит дальше первой страницы выдачи.

# Инструментальные средства анализа данных

Наиболее популярная группа инструментов содержит следующие категории:

* + наборы инструментов;
  + классификация данных;
  + кластеризация и сегментация;
  + инструменты статистического анализа;
  + анализ текстов (Text Mining), извлечение отклонений;
  + инструменты визуализации.

# Тестовые вопросы

**Предполагается, что тут выбраны только правильные варианты ответов, неправильные я не перечисляю, чтобы не забивать голову**

1. **Знания** — это факты и правила, сохраняющиеся в памяти людей и влияющие на их убеждения.
2. Что такое **знания** - вид информации, отражающей опыт и восприятие человеком окружающего мира.
3. Выберите временной период "ручной" информационной технологии - до второй половины XIX в.
4. Носители информации бывают - внемашинные, электронные, машинные.
5. Процесс ETL - это Extraction, Transformation, Loading.
6. **Релевантность** - это устанавливаемое при информационном поиске соответствие содержания документа информационному запросу или поискового образа документа поисковому предписанию. ТАКЖЕ ЭТО степень соответствия найденного документа или набора документов информационным нуждам пользователя.
7. **Количественный контент-анализ** основан на частоте появления в документах определенных характеристик содержания.
8. **Качественный контент-анализ** основан на самом факте присутствия или отсутствия в тексте одной или нескольких характеристик содержания.
9. **Горизонт прогнозирования** — это число периодов в будущем, которые покрывают прогноз.
10. **Информационная система** - совокупность средств, методов и организационных ресурсов (человеческие, технические, финансовые и т. д.), используемых для хранения, обработки и выдачи информации в интересах достижения поставленной цели.
11. **Информационная система** – система, предназначенная для хранения, поиска и обработки информации, и соответствующие организационные ресурсы (человеческие, технические, финансовые и тд), которые обеспечивают и распространяют информацию.
12. **Информационной системой** называется комплекс, включающий вычислительное и коммуникационное оборудование, программное обеспечение, лингвистические средства и информационные ресурсы, a также системный персонал и обеспечивающий поддержку динамической информационной модели некоторой части реального мира для удовлетворения информационных потребностей пользователей.
13. **Информационная система** есть совокупность технического, программного и организационного обеспечения, a также персонала, предназначенная для формирования семантически/синтаксически верных фактов. Неверно.
14. **SOA** — это прикладная архитектура, в которой все функции определены как независимые сервисы с вызываемыми интерфейсами. Обращение к этим сервисам в определенной последовательности позволяет реализовать тот или иной бизнес-процесс.

Отметьте все верные высказывания:

* + SOA — это компонентная модель, в которой разные функциональные единицы приложений, называемые сервисами, взаимодействуют по сети посредством интерфейсов
  + Сервисы в системах, построенных на SOA, могут быть реализованы в независимости от других служб системы, необходимо только знание интерфейса используемых сервисов.
  + Каждый, составляющий информационную систему сервис реализует отдельную бизнес-функцию, которая является логически обособленной, повторяющейся задачей, являющейся составной частью бизнес-процесса предприятия.
  + Все сервисы независимы друг от друга. Они выполняют определенные действия по запросам, полученным от других сервисов, и возвращают результаты. Все детали этого полностью скрыты: в концепции SOA сервисы - это "черные ящики".

1. **Таблицы измерений** содержат редко изменяемые либо неизменяемые данные. Верно.
2. **Таблица измерений** содержит метрики и факты o бизнес-процессе. Верно.
3. **Таблица фактов** содержит измерения, метрики и факты o бизнес-процессе. Верно. (Неверно)
4. **Таблицы фактов** содержат неизменяемые либо редко изменяемые данные. Неверно.
5. **Уровни интеграции данных**: логический, физический, семантический.
6. **Тренд** - неслучайная функция, которая формируется под действием общих или долговременных тенденций, влияющих на временной ряд.
7. **Период** - основная единица времени на который делается прогноз.
8. **Прогнозирование** - это не метод Text Mining. Неверно.
9. **Кластеризация** - это метод Text Mining. Верно.
10. **Классификация** - это метод Text Mining. Верно.
11. **Кластеризация** - группировка объектов на основе данных, характеризующих сущность этих объектов.
12. Выберите существующие методы интеграции: распределение, федерализация, консолидация.
13. **Консолидация** как метод интеграции данных - при использовании этого метода данные физически извлекаются из нескольких первичных систем и интегрируются в одно постоянное место хранения.
14. **Распространение** как метод интеграции данных - при использовании этого метода данные физически копируются из одного места в другое, пока не попадут в некоторую целевую систему.
15. **Федерализация** как метод интеграции данных - при использовании этого метода данные не консолидируются физически, а хранятся в своих источниках и становятся доступными только при выполнении соответствующего запроса.
16. Различие задач **классификации** и **прогнозирования** состоит в том, что в первой задаче предсказываются числовые значения зависимой переменной, пропущенные или неизвестные (относящиеся к будущему), а во второй - класс зависимой переменной. Нет.
17. Различие задач **классификации** и **прогнозирования** состоит в том, что в первой задаче предсказывается класс зависимой переменной, а во второй - числовые значения зависимой переменной, пропущенные или неизвестные (относящиеся к будущему). Да.
18. **База данных** - совокупность данных, организованных в соответствии c концептуальной структурой, описывающей характеристики этих данных и взаимоотношения между ними, которая поддерживает одну или более областей применения. ТАКЖЕ ЭТО совместно используемый набор логически связанных данных (и описание этих данных), предназначенный для удовлетворения информационных потребностей.
19. Выберите правильные свойства **цикличности**
    * циклы, в отличие от периодов сезонных, не имеют определенной продолжительности.
    * продолжительность цикла, как правило, больше, чем один сезон.
20. **Хранилища данных** – это информационная система, в которой хранятся исторические и коммутативные данные из одного или нескольких источников. Верно.
21. **Data Warehouse** - предметно-ориентированная информационная база данных.
22. **Хранилище данных** предназначено для анализа, отчетности, интеграции данных транзакций из разных источников. Верно.
23. **Business Intelligence** - обозначение компьютерных методов и инструментов для организаций, обеспечивающих перевод транзакционной деловой информации в человеческую форму, пригодную для бизнес-анализа. Верно.
24. **Business Intelligence** - обозначение компьютерных методов и инструментов для организаций, обеспечивающих перевод отчетов и прочей консолидированной деловой информации в транзакционную форму. Неверно.
25. **OLTP** - это система обработки транзакций в реальном времени. Способ организации базы данных, при котором система работает в небольшими по размеру транзакциями, но идущими большим потоком чтения.
26. **Аналитические системы (OLTP)** - структура базы данных ориентирована на быструю аналитику любых объемов данных. Данные в таких базах хранятся уже в подготовленном для аналитики виде. Это позволяет эффективно в реальном времени получать отчеты за любой период в любой детализации. Неверно.
27. **OLAP** - технология комплексного многомерного анализа данных.
28. **Аналитические системы (OLAP)** - структура базы данных ориентирована на быструю аналитику любых объемов данных. Данные в таких базах хранятся уже в подготовленном для аналитики виде. Это позволяет эффективно в реальном времени получать отчеты за любой период в любой детализации. Верно.
29. **Агрегация данных в OLAP-кубе** — это обобщение (суммирование) значений атрибутов (измерений) ИЛИ обобщение (суммирование) значения мер (правильный ответ неизвестен, но я бы выбрала первое)
30. Может ли существовать иерархия Год-Месяц-День? Да.
31. На какой стадии ИАД происходит выявление отклонений - анализ исключений.
32. Объект, который по мере сходства может быть отнесен к нескольким кластерам - спорный объект.
33. Какие виды контент-анализа существуют - качественный, количественный.
34. Выберете несуществующий способ обработки информации - циклический.
35. Алгоритмами кластеризации не являются - анализ временного ряда.
36. Какие из вариантов являются методами интеллектуального анализа данных

* методы вывода уравнений; методы кросс-табуляции.

1. Что не относится к методам ИАД - прогностическое моделирование.
2. Какая задача не решается методами Data Mining? Трансформация.
3. В каких условиях не может проходить процесс принятия решений? В условиях предопределенности
4. В каком временном периоде была «Электрическая» информационная технология? (40 – 60-е гг. XXв.)
5. Задачи, решаемые методами Data Mining

* Классификация
* Кластеризация
* Прогнозирование
* Ассоциация

1. Что НЕ является элементом text mining? Изложение, Кросс-табуляция
2. Элементы Text Mining: классификация; кластеризация; прогнозирование; нахождение исключений; поиск связанных признаков; извлечение фактов, понятий; реферирование; ответ на запросы; тематическое индексирование; поиск по ключевым словам.
3. Чем НЕ характеризуется OLTP? При возникновении ошибки транзакция так или иначе будет считаться выполненной, откаты системы не предусматриваются
4. Что из перечисленного не относится к свойствам информации? Предопределенность.
5. Что НЕ входит в такую стадию ИАД, как прогностическое моделирование? Выявление закономерностей, Выявление отклонений.
6. Какой бывает классификация? Естественной, искусственной
7. Выберите вариант ответа, в котором информационные технологии перечислены в хронологическом порядке по мере их возникновения: «Ручная» «Механическая» «Электрическая» «Электронная» «Компьютерная»
8. К какому понятию это относится: способ организации БД, при котором система работает с большими по размерам транзакциями, но идущими небольшим потоком и при этом клиенту требуется от системы минимальное время оклика. OLAP
9. Выберите стадии ИАД: Анализ исключений, Прогностическое моделирование, Свободный поиск.
10. СППР расшифровывается как: Система поддержки принятия решений
11. Очистка данных (cleaning) — это процесс отсеивания несущественных или исправления ошибочных данных на основании статистических или экспертных методов.
12. Первый этап разработки информационной системы Исследование предметной области
13. Какая из задач, решаемых методами data mining подходит под описание: «в результате решения задачи обнаруживаются признаки которые характеризуют группы объектов исследуемого набора данных» Классификация
14. Цель Business intelligence – интерпретировать большое количество данных, заостряя внимание лишь на ключевых факторах эффективности, моделируя исход различных вариантов действий, отслеживая результаты принятия решений. Верно