1. **Информация и информационное общество**

Три существенных аспекта информации.

• **Гносеологический аспект** — понимание информации, обладающей свойством атрибута материи, как продукта ее отражения.

• **Функциональный аспект**. Информация — это обозначение содержания, полученного из внешнего мира в процессе нашего приспособления к нему и приспосабливания к нему наших органов чувств. Процесс получения и использования информации является процессом нашего приспособления к случайностям внешней среды и нашей жизнедеятельности в этой среде».

• **Деятельностный аспект** — информация рассматривается как обмен сведениями между людьми, человеком и автоматом, автоматом и автоматом, обмен сигналами в растительном и животном мире, передача признаков от клетки к клетке.

**Информация** — это совокупность сведений, необходимых для активного воздействия на управляемую систему с целью ее оптимизации; • набор узкоспециальных данных, продуцированных в огромных количествах во всех сферах деятельности общества; • необходимый резерв и ресурс социально-экономического развития общества, подобный другим ресурсам — трудовым, материальным.

Характерными **чертами информации** являются следующие: 1) это наиболее важный ресурс современного. 2) информация вызывает к жизни новые производства. 3) информация является товаром, причем продавец информации ее не теряет после продажи. 4) информация придает дополнительную ценность другим ресурсам, в частности трудовым.

С информацией всегда связывают три понятия: • **источник информации** — тот элемент окружающего мира (объект, процесс, явление, событие), сведения о котором являются объектом преобразования; • **потребитель информации** — тот элемент окружающего мира, который использует информацию (для выработки поведения, принятия решения, управления или обучения); • **сигнал** — материальный носитель, который фиксирует информацию для переноса ее от источника к потребителю. В данном случае сигнал носит электронный характер

**Общество**, в котором большинство трудоспособного населения занято производством, хранением, переработкой и реализацией информации, принято называть **информационным**. **Отличительными признаками** информационного общества являются следующие. **1**. Важнейшими продуктами социальной деятельности становятся информационные технологии, услуги и знания. Основными аспектами их использования являются актуализация, защита, обеспечение целостности информации. **2**. За счет появления новых информационных технологий удается решать новые задачи, справиться с которыми ранее было либо невозможно, либо экономически нецелесообразно. **3**. Каждый гражданин и учреждение в любое время могут получить какую угодно информацию, необходимую для их жизни и деятельности. **4**. Существует вся необходимая инфраструктура для информационных технологий — вычислительная техника, средства телекоммуникации и связи, программные продукты, базы данных и знаний коллективного использования и др.

**Информационная среда** — это техническая материальная база, готовность и умение общества использовать информационную среду.

**2) Использование математики для записи и обработки информации**

**Определением** называется выражение, в котором разъясняется смысл нового понятия. **Теорема** есть утверждение, справедливость которого устанавливается путем некоторого рассуждения, называемого **доказательством**. **Аксиомой** называется истина, принимаемая без доказательства. Непосредственный вывод из аксиомы или теоремы называется **следствием**. Основным методом построения современной математики является аксиоматический метод. Аксиоматическое построение того или иного конкретного раздела математики осуществляется следующим образом: 1) отбираются первичные термины; 2) выделяются некоторые первичные утверждения — аксиомы; 3) все новые понятия, вводимые в данной теории, должны быть определены через первичные термины или через ранее определенные понятия и соотношения; все новые утверждения теории (термины) должны быть доказаны на основе первичных терминов или аксиом путем дедукции. **Дедукция** — способ рассуждения, посредством которых из общих посылок с необходимостью следует заключение частного характера.

**3) Математическое моделирование как один из основных методов познания**

**Модель** — это искусственно созданный объект, дающий упрощенное представление о реальном объекте, процессе или явлении, отражающий существенные стороны изучаемого объекта с точки зрения цели моделирования.

**Моделирование** — это построение моделей, предназначенных для изучения и исследования объектов, процессов или явлений. Объект, для которого создается модель, называют оригиналом или прототипом. **Моделирование** — исследование явлений, процессов или систем объектов путем построения и изучения их моделей.

**Признаки**, по которым классифицируются модели: • цель использования (модели учебные, опытные, имитационные, игровые, научно-технические); • область знаний (биологические, экономические, исторические, социологические и т.д.); • фактор времени (динамические и статические); • способ представления

**Информационная модель** — это совокупность информации об объекте, описывающая свойства и состояние объекта, процесса или явления, а также связи и отношения с окружающим миром. Строится только на информации

Образные модели, знаковые информационные модели, устественные языки, языки математики и алгебры логики

**4. Математические средства представления информации в виде знаковых информационных моделей**

**Формулы**. Информация показана в свернутом виде. Использование формул позволяет: 1) дать описание объекта в свернутом виде 2) отразить причинно-следственные связи физического явления, 3) передать свой-ва объекта 4) предсказать свой-ва и поведение моделируемого объекта за пределами видимых наблюдений

**Таблицы**. Самый простой способ систематизировать данные. **Подлежащее** таблицы – объект статистического изучения. **Сказуемое** таблицы – статистические показатели характеризующие изучаемый объект. Три вида таблиц: простые (справочные материал), групповые(статистическая совокупность разбивается на отдельные группы), комбинационные(комбинированные)(группы и подгруппы)

**Графики**. В виде ломанной линии, радиальные диаграммы (замкнутые, спиральные), столбчатый, круговой, ленточный, z-образный.

**5) Математическая обработка информации**

**6) Теоретико-множественный подход в обработки информации**

**7) Понятие статистки**

**Статистика** — отрасль знаний, в которой излагаются общие вопросы сбора, измерения и анализа массовых статистических (количественных или качественных) данных. Разрабатывает специальную методологию исследования и обработки материалов: массовые статистические наблюдения, метод группировок, средних величин, индексов, балансовый метод, метод графических изображений и другие методы анализа статистических данных.

**Статистический анализ** - применение принципов возможности к данным исследований для того, чтобы узнать, насколько вероятно, что данные получены случайно. ***Статистический анализ***исследуемого явления или процесса всегда опирается на *исходные статистические данные*. Выводы статистического анализа составляют существенный компонент *системы поддержки принятия стратегических решения*. Форма и содержание исходных статистических данных зависят от конечных прикладных целей исследования и используемых источников (см. ниже п.3.). В частности, конечные прикладные цели статистического анализа *механизма функционирования фирм*(предприятий) и связанных с этих задач *прогнозирования*обусловливают состав и структуру показателей (так называемое *фазовое пространство*), наблюдение за которыми и образует массив исходных статистических данных.

**8) Машинное обучение**

**Машинное обучение** — обширный подраздел [искусственного интеллекта](http://www.machinelearning.ru/wiki/index.php?title=%D0%98%D1%81%D0%BA%D1%83%D1%81%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%BD%D1%8B%D0%B9_%D0%B8%D0%BD%D1%82%D0%B5%D0%BB%D0%BB%D0%B5%D0%BA%D1%82), изучающий методы построения [алгоритмов](http://www.machinelearning.ru/wiki/index.php?title=%D0%90%D0%BB%D0%B3%D0%BE%D1%80%D0%B8%D1%82%D0%BC), способных обучаться. Различают два типа обучения. *Обучение по прецедентам*, или *индуктивное обучение*, основано на выявлении общих закономерностей выборкой данных. *Дедуктивное обучение* предполагает формализацию знаний экспертов и их перенос в компьютер в виде [базы знаний](http://www.machinelearning.ru/wiki/index.php?title=%D0%91%D0%B0%D0%B7%D0%B0_%D0%B7%D0%BD%D0%B0%D0%BD%D0%B8%D0%B9&action=edit). Дедуктивное обучение принято относить к области [экспертных систем](http://www.machinelearning.ru/wiki/index.php?title=%D0%AD%D0%BA%D1%81%D0%BF%D0%B5%D1%80%D1%82%D0%BD%D0%B0%D1%8F_%D1%81%D0%B8%D1%81%D1%82%D0%B5%D0%BC%D0%B0), поэтому термины *машинное обучение* и *обучение по прецедентам* можно считать синонимами.

**9) Понятие искусственного интеллекта**

**Искусственный интеллект** - это возможность решения задач, которые до сих пор не удавалось решить человеку, машинным способом с помощью программных средств. Основной проблемой искусственного интеллекта является разработка методов представления и обработки знаний. Выделяются следующие направления искусственного интеллекта: 1. экспертные системы; 2. нейронные сети; 3. естественно-языковые системы; 4. эволюционные методы и генетические алгоритмы; 5. нечеткие множества; 6. системы извлечения знаний.

Так, одна из классификаций выделяет два подхода к разработке ИИ:

* нисходящий, семиотический — создание символьных систем, моделирующих высокоуровневые [психические процессы](http://www.psychologos.ru/articles/view/psihicheskie_processy): мышление, рассуждение, речь, эмоции, творчество и т. д.;
* восходящий, биологический — изучение нейронных сетей и эволюционные вычисления, моделирующих интеллектуальное поведение на основе более мелких «неинтеллектуальных» элементов.

**10) Системы поддержки принятия решений (СППР). Классификация СППР**

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

Предназначена для поддержки многокритериальных решений в сложной информационной среде. Многокритериальность - результаты принимаемых решений оцениваются по совокупности многих показателей, рассматриваемых одновременно. Информационная сложность определяется необходимостью учета большого объема данных.

**Две основные задачи**: выбор наилучшего решения из множества возможных (оптимизация), упорядочение возможных решений по предпочтительности (ранжирование).

**Классификации СППР**

**По взаимодействию с пользователем** выделяют три вида СППР: пассивные помогают в процессе принятия решений, но не могут выдвинуть конкретного предложения; активные непосредственно участвуют в разработке правильного решения; кооперативные предполагают взаимодействие СППР с пользователем. Выдвинутое системой предложение пользователь может доработать, усовершенствовать, а затем отправить обратно в систему для проверки. После этого предложение вновь представляется пользователю, и так до тех пор, пока он не одобрит решение.

**По способу поддержки различают**: модельно-ориентированные СППР, используют в работе доступ к статистическим, финансовым или иным моделям; СППР, основанные на коммуникациях, поддерживают работу двух и более пользователей, занимающихся общей задачей; СППР, ориентированные на данные, имеют доступ к временным рядам организации. Они используют в работе не только внутренние, но и внешние данные; СППР, ориентированные на документы, манипулируют неструктурированной информацией, заключенной в различных электронных форматах; СППР, ориентированные на знания, предоставляют специализированные решения проблем, основанные на фактах.

**По сфере использования выделяют**: общесистемные, настольные СППР.

**11) Технологии БД**

**БД**- совокупность взаимосвязанных, хранящихся вместе сведениях о различных сущностях одной предметной, обеспечивающая наличие такой минимальной избыточности, которая допускает их использование оптимальным образом для одного или нескольких приложений, или пользователей;

**СУБД**- программные средства, предназначенные для управления БД.

**Модель данных**- является фундаментом технологий баз данных; на ней базируется конкретная СУБД. Модель описывает набор понятий и признаков, которыми должна обладать конкретная СУБД и управляемые ими базы данных.

**Парадигма**- это пространство идей и законы движения в этом пространстве. В рамках парадигмы определены аксиомы, на которых выстраивается своя логика. Решения, вырабатываемые в рамках парадигмы, непротиворечивы и логичны.

**Интеграция неоднородных информационных ресурсов**. Информационная неоднородность ресурсов заключается в разнообразии понятий, словарей; отображаемых реальных объектов; правил, определяющих адекватность моделируемых объектов реальности; видов данных, способов их сбора и обработки; интерфейсов пользователей и т.д.

Реализационная неоднородность источников проявляется в использовании разнообразных компьютерных платформ, средств управления базами данных, моделей данных и знаний, средств программирования, операционных систем, и т.п.

Традиционные системы баз данных, используемые в информационных системах для сопровождения бизнес - процессов поддерживают большие объемы информации с помощью технологий оперативная обработка транзакций – OLTP. В OLTP-технологии обрабатывается детализированные данные, главные свойства данных здесь, их полнота и актуальность.

Для поддержки принятия решений нужны другие технологии. Необходимо объединять данные из различных источников (как из корпоративной информационной системы, так и из внешней среды), накапливать данные, делая их срезы во времени. Анализ таких данных позволяет оценивать состояние и динамику развития организации, делать обоснованные прогнозы и принимать обоснованные решения. Программные продукты, необходимые для обеспечения управленческих решений, должны обеспечивать хранение больших объемов данных, эффективный доступ к ним, а так же располагать развитыми средствами анализа данных и представления результатов в удобной для специалистов и руководства форме.

**OLAP-технология, оперативная аналитическая обработка**- информационная технология, которая предоставляет руководителям различного уровня возможность получения необходимой информации для принятия управленческих, финансовых и кадровых решений.

**Хранилище данных** - обеспечивает накопление с течением времени данные для содействия в принятии решений. Хранилище это данных репозиторий (склад) информации содержащий объединенные, проверенные данные, отражающие работа организации за длительный период. Объемы данных в хранилищах данных в несколько раз превосходят объемы данных в OLTP-системах.

**12. Data mining**

Data Mining – это процесс обнаружения в "сырых" данных ранее неизвестных нетривиальных практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

**Суть и цель** *Data Mining*: это технология, которая предназначена для поиска в больших объемах *данных* неочевидных, объективных и полезных на практике *закономерностей*.

**Неочевидных** - это значит, что найденные *закономерности* не обнаруживаются стандартными методами обработки информации или экспертным путем.

**Объективных** - это значит, что обнаруженные *закономерности* будут полностью соответствовать действительности, в отличие от экспертного мнения, которое всегда является субъективным.

**Практически полезных** - это значит, что выводы имеют конкретное *значение*, которому можно найти практическое применение.

***Знания*** - совокупность сведений, которая образует целостное описание, соответствующее некоторому уровню осведомленности об описываемом вопросе, предмете, проблеме и т.д.

В основу технологии *Data Mining* положена концепция ***шаблонов*** (patterns), которые представляют собой *закономерности*, свойственные под выборками *данных*, кои могут быть выражены в форме, понятной человеку.

"Mining" по-английски означает "добыча полезных ископаемых", а *поиск* *закономерностей* в огромном количестве *данных* действительно сродни этому процессу.

Цель поиска *закономерностей* - *представление* *данных* в виде, отражающем искомые процессы. Построение моделей прогнозирования также является целью поиска закономерностей.

**13. Data mining как технология обработки информации**

Несмотря на большое количество разнообразных бизнес-задач почти все они решаются по единой методике. Эта методика называется Knowledge Discovery in Databases. Она описывает не конкретный алгоритм или математический аппарат, а последовательность действий, которую необходимо выполнить для построения модели (извлечения знания). Данная методика не зависит от предметной области, это набор атомарных операций, комбинируя которые можно получить нужное решение. Рабочий цикл Data Mining Мониторинг качества Построение моделей Очистка данных Сбор данных Формирование гипотез Передача в эксплуатацию лучших моделей Проверка адекватности текущей модели

1. **Данные и знания**

**Данные** - это отдельные факты, характеризующие объекты, процессы и явления в предметной области, а также их свойства.

**Знания** - совокупность сведений, образующих целостное описание, соответствующее некоторому уровню осведомленности об описываемом вопросе, предмете, проблеме и т.д. Знания - это выявленные закономерности в предметной области (принципы, связи, законы), позволяющие решать задачи в этой области.

**База знаний**, наравне с базой данных, - необходимая составляющая программного комплекса ИИ. Машины, реализующие алгоритмы ИИ, называются машинами, основанными на знаниях, а подраздел теории ИИ, связанный с построением экспертных систем, - инженерией знаний.

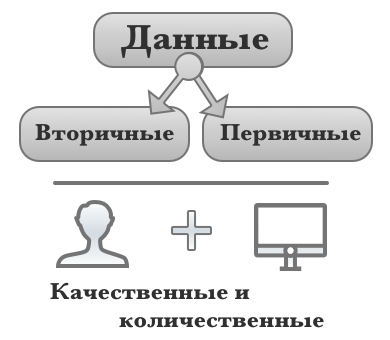
**Знания бывают классифицированы по следующим категориям**: **поверхностные** - знания о видимых взаимосвязях между отдельными событиями и фактами в предметной области; **глубинные** - абстракции, аналогии, схемы, отображающие структуру и процессы в предметной области.

Вместе с тем, знания можно разделить на следующие виды: **процедурные**: знания, отвечающие на вопрос ʼʼКак решать поставленную задачу? ʼʼ; эти знания хранятся в памяти интеллектуальной системы в виде описаний процедур, с помощью которых их можно получить. В таком виде обычно описывается информация о предметной области, характеризующая способы решения задач в этой области, а также различные инструкции, методики и т.п.

**декларативные**: знания, не содержащие в явном виде процедуры решения задач; которые записаны в памяти так, что они непосредственно доступны для использования после обращения к соответствующему полю памяти. В таком виде обычно записывается информация о свойствах предметной области фактах, имеющих в ней место и т.п. информация.

1. Виды данных

Данные представляют собой определенным образом полученные и зафиксированные наблюдения относительно окружающей действительности.  
Исходя из того, для каких целей были собраны данные, их можно разделить на первичные и вторичные. Первичными являются такие данные, которые собраны для решения целей данного исследования. Соответственно, вторичными являются данные, собранные в рамках другого исследования, то есть для решения целей, отличных от целей данного исследования. Принято начинать с анализа вторичных данных, а уже в случае их недостаточности переходить к сбору первичных.



Поскольку вторичные данные собираются в рамках других исследовательских проектов, всегда существует проблема их качества. Критериями его оценки являются следующие:

|  |  |
| --- | --- |
| Критерии | Ключевые факторы |
| Процедура и методика сбора | Метод сбора данных, процент ответов, способ формирования выборки, размер выборки, объем и логика анкеты, полевой этап, анализ данных. |
| Ошибки и точность | Сравнение данных, полученных из разных источников. |
| Своевременность | Временной промежуток между сбором данных и их публикацией, частота обновления (если исследование предполагает несколько этапов). |
| Степерь соответствия цели исследования |  |
| Содержание данных | Ключевые переменные, шкалы измерения, исследуемые связи. |
| Надежность | Компетентность, достоверность, репутация и надежность источника информации. |

Если же говорить о сущностных характеристиках данных, то принято выделять качественные и количественные данные. Можем ли мы редуцировать данные к форме чисел или они могут быть представлены только с помощью слов? Важно проводить различие между этими двумя типами данных, поскольку они определяют способы их получения, фиксации и анализа.  
Большое количество научной информации фиксируется в виде чисел. Природа чисел разрешает манипулировать соответствующими данными с помощью методов математической статистики. Такого рода данные являются количественными. Главная проблема сбора количественных данных - разработка точных измерительных инструментов в виде анкетных вопросов, шкал или тестов. Такие измерительные инструменты должны тщательно проверяться на валидность и надежность. В зависимости от особенностей измеряемых свойств и/или точности самого измерения, количественные данные можно получить с помощью одной из трех измерительных шкал – номинальной, порядковой и метрической.  
Вместе с тем, существует важная информация, которую нельзя редуцировать к форме чисел. Мысли, чувства, идеи и традиции людей нуждаются в их словесном выражении. Соответствующие данные называют качественными. Вербальные концепты и взаимосвязи между ними являются менее точными, чем числа и соответствующие связи. Это делает качественные исследования более зависимыми от особенностей определения значения слов, разработки понятий и определения взаимосвязей между ними. В отличии от количественных исследований, в качественных не существует общепринятых образцов анализа соответствующих данных, что требует рефлексивности ученого (внимания к возможным субъективным смещениям).

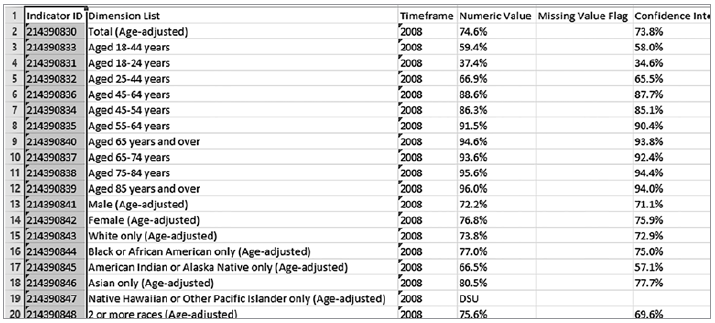
По способу получения, данные можно разделить на четыре типа:

* **Наблюдение** Фиксация случаев, ситуаций или событий, известных из собственного опыта, в том числе с использованием специальных средств (например, камеры, диктофона, микроскопа и т.д.).
* **Участие** Данные получаются благодаря опыту, который может рассматриваться как интенсивная форма наблюдения (например, опыт обучения управлению автомобилем сообщит такие нюансы относительно авто, которые невозможно получить, лишь наблюдая за ним извне).
* **Измерение** Фиксация величины или количества какого-либо параметра (например, демографическая статистика, измерение физических величин и т.д.).
* **Интерегация** Данные получаются посредством вопросов к людям (информация относительно убеждений людей, их мотивации и т.д.).

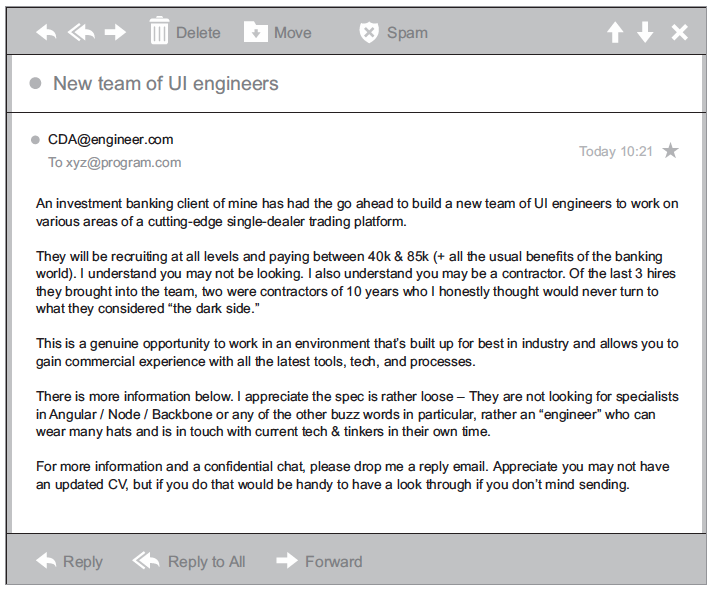
В **data science** и области больших данных выделяют много разных типов данных, для каждого из которых требуются свои инструменты и методы. Основные категории данных перечислены ниже.

* Структурированные.
* Неструктурированные.
* На естественном языке.
* Машинные.
* Графовые.
* Аудио, видео и графика.
* Потоковые.

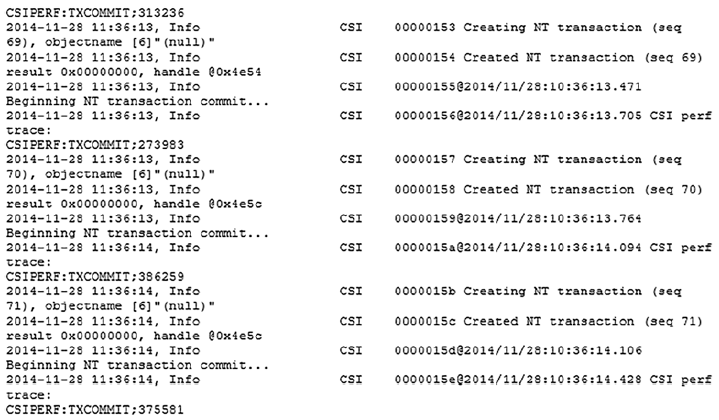
Все эти типы данных представляют интерес, и их стоит рассмотреть подробнее. **Структурированные данные** зависят от модели данных и хранятся в фиксированном поле внутри записи. Соответственно, структурированные данные часто бывает удобно хранить в таблицах, в базах данных или файлах Excel:



SQL (Structured Query Language, язык структурированных запросов) является основным средством управления и обращения с запросами к данным, хранящимся в базах данных. Также иногда встречаются структурированные данные, которые достаточно трудно сохранить в традиционной реляционной базе данных (один из примеров — иерархические данные, например генеалогическое дерево).  
Впрочем, мир не состоит из структурированных данных; просто это представление удобно для человека и машин. Чаще реальные данные хранятся в неструктурированном виде. **Неструктурированные данные** трудно подогнать под конкретную модель данных, потому что их содержимое зависит от контекста или имеет переменный характер. Один из примеров неструктурированных данных — обычные сообщения электронной почты:



Хотя сообщение содержит структурированные элементы (отправитель, заголовок, тело), одни и те же задачи могут решаться множеством разных способов, например, существует бесчисленное количество вариантов упоминания конкретного человека в сообщениях. Проблема дополнительно усложняется существованием тысяч языков и диалектов.  
Сообщение электронной почты, написанное человеком (наподобие показанного выше), также является идеальным примером данных на естественном языке.  
**Данные на естественном языке** составляют особую разновидность неструктурированных данных; обработка таких данных достаточно сложна, потому что она требует знания как лингвистики, так и специальных методов data science.  
Сообщество обработки данных на естественном языке добилось успеха в области распознавания сущностей, распознавания тематических областей, обобщения, завершения текста и анализа эмоциональной окраски, но модели, адаптированные для одной предметной области, плохо обобщаются для других областей. Даже самые современные методы не смогут расшифровать смысл произвольного фрагмента текста. И этот факт вряд ли кого-то удивит: у людей также возникают проблемы с восприятием естественного языка. Он неоднозначен по своей природе. Сама концепция смысла выглядит спорно. Два человека слушают один разговор; вынесут ли они одинаковый смысл из него? Даже смысл отдельных слов может изменяться в зависимости от настроения говорящего.  
К **машинным данным** относится информация, автоматически генерируемая компьютером, процессом, приложением или устройством без вмешательства человека. Машинные данные становятся одним из основных источников информации, и ситуация вряд ли изменится. Wikibon предсказывает, что рыночная стоимость промышленного Интернета (термин, предложенный компанией Frost&Sullivan для обозначения совокупности сложного физического оборудования с сетевыми датчиками и программным обеспечением) к 2020 году составит приблизительно 540 миллиардов долларов. По оценкам IDC (International Data Corporation), количество узлов сети к 2020 году в 26 раз превысит численность населения. Эта сеть часто называется Интернетом вещей.  
Анализ машинных данных из-за их громадных объемов и скоростей сильно зависит от инструментов с высокой масштабируемостью. К примерам машинных данных относятся журналы веб-серверов, записи детализации звонков, журналы сетевых событий и телеметрии:



Машинные данные на рисунке выше хорошо укладываются в структуру классической базы данных. Это не лучший формат для данных с высокой степенью связности или «сетевых» данных, в которых достаточно значимую роль играют отношения между сущностями.  
Термин **«графовые данные»** может сбить с толку, потому что любые данные могут быть представлены в виде графа. Под «графом» в данном случае имеется в виду понятие графа из математической теории графов — математическая структура для моделирования попарных отношений между объектами. Вкратце, в графовых, или сетевых, данных особое внимание уделяется связям или смежности объектов. Графовые структуры данных используют узлы, ребра и свойства для представления и хранения графических данных. Графовые данные естественным образом подходят для представления социальных сетей, а их структура позволяет вычислять такие специфические метрики, как влияние участников и кратчайший путь между двумя людьми.  
Примеры графовых данных встречаются на многих веб-сайтах социальных сетей. Например, в LinkedIn можно увидеть, кого вы знаете в той или иной компании. Ваш список читателей в Твиттере также является примером графовых данных. Сила и мощь связанных данных проявляется при анализе нескольких перекрывающихся графов, построенных на одних и тех же узлах. Например, представьте, что ребра обозначают «друзей» на Facebook. А теперь возьмем другой граф с теми же людьми, но связывающий коллег по бизнесу через LinkedIn, и третий граф, основанный на интересе к фильмам на Netflix. Наложение этих трех графов позволит получить ответы на многие интересные вопросы.  
Для хранения графовых данных используются графовые базы данных, а для построения запросов к ним — такие специализированные языки запросов, как SPARQL.  
Работа с графовыми данными создает специфические проблемы, причем для компьютера эта задача становится еще сложнее.  
**Аудио, видео и графика** — типы данных, ставящие непростые задачи перед специалистом data science. Задачи, тривиальные с точки зрения человека (например, распознавание объекта на картинке), оказываются сложными для компьютера. В 2014 году компания MLBAM (Major League Baseball Advanced Media) объявила, что объем записываемых видеоматериалов для одного бейсбольного матча будет увеличен приблизительно до 7 Тбайт с целью проведения оперативного анализа. Высокоскоростные камеры на стадионах записывают движения мяча и спортсменов для того, например, чтобы вычислять в реальном времени траекторию движения защитника.  
Недавно компании DeepMind удалось создать алгоритм, который способен обучаться играть в видеоигры. Алгоритм получает на входе содержимое экрана и учится интерпретировать эти данные в сложном процессе глубокого обучения. Это замечательное достижение, и компания Google приобрела DeepMind для разработки искусственного интеллекта. Алгоритм обучения получает данные, генерируемые компьютерной игрой, т. е. потоковые данные.  
**Потоковые данные** могут принимать почти любую из перечисленных форм, однако у них имеется одно дополнительное свойство. Данные поступают в систему при возникновении некоторых событий, а не загружаются в хранилище данных большими массивами. И хотя формально они не являются отдельной разновидностью данных, мы выделяем их в особую категорию, потому что вам придется приспособить свой рабочий процесс для работы с потоковой информацией.  
Примерами потоковых данных могут служить раздел «Что происходит?» в Твиттере, прямые трансляции спортивных и музыкальных мероприятий и данные биржевых котировок.

1. **Набор данных и их атрибутов**

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

### Набор данных и их атрибутов

В таблице 2.1 представлена двухмерная *таблица*, представляющая собой набор *данных*.

***Объект*** описывается как набор атрибутов.

*Объект* также известен как *запись*, случай, пример, строка таблицы и т.д.

***Атрибут*** - свойство, характеризующее *объект*.

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

В результате перехода от общих категорий к конкретным величинам, получается набор переменных изучаемого понятия.

**Переменная** - свойство или характеристика, общая для всех изучаемых *объектов*, проявление которой может изменяться от *объекта* к *объекту*.

**Значение** переменной является проявлением признака.

**Генеральная совокупность**  - вся совокупность изучаемых *объектов*, интересующая исследователя.

***Выборка*** (sample) - часть генеральной совокупности, определенным способом отобранная с целью исследования и получения выводов о свойствах и характеристиках генеральной совокупности.

**Параметры** - числовые характеристики генеральной совокупности.

**Статистики** - числовые характеристики *выборки*.

Часто исследования основываются на *гипотезах*. *Гипотезы* проверяются с помощью *данных*.

***Гипотеза*** - частично обоснованная *закономерность* знаний, служащая либо для связи между различными эмпирическими фактами, либо для объяснения факта или группы фактов.

1. **Измерения и шкалы**

**Измерение** - процесс присвоения чисел характеристикам изучаемых *объектов* согласно определенному правилу.

В процессе подготовки *данных* измеряется не сам *объект*, а его характеристики.

***Шкала*** - правило, в соответствии с которым объектам присваиваются числа.

Переменные могут являться **числовыми** данными либо **символьными**.

Числовые данные, в свою *очередь*, могут быть дискретными и непрерывными.

**Дискретные данные** являются значениями признака, общее число которых конечно либо бесконечно, но может быть подсчитано при помощи натуральных чисел от одного до бесконечности.

**Непрерывные данные** - данные, значения которых могут принимать какое угодно *значение* в некотором интервале. Измерение непрерывных *данных* предполагает большую *точность*.

**Существует пять типов*****шкал* измерений**: номинальная, порядковая, интервальная, относительная и дихотомическая.

**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.

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

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

Эта *шкала* позволяет находить разницу между двумя величинами

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

**Дихотомическая шкала -** шкала, содержащая только две категории.

1. **Типы наборов данных**

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

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

#### Графические данные

Примеры графических *данных*: WWW-данные; молекулярные структуры; графы; карты.

С помощью карт, например, можно отследить изменения *объектов* во времени и пространстве, определить характер их распределения на плоскости или в пространстве.

Пример карты, являющейся картой Кохонена

#### Химические данные

Химические данные представляют собой особый тип *данных*. Пример таких *данных*: C6H6

1. **Форматы хранения данных**

Возможны четыре аспекта работы с данными: *определение* *данных*, *вычисление*, манипулирование и обработка (сбор, передача и др.).

При манипулировании данными используется структура *данных* типа "*файл*". Файлы могут иметь различные форматы.

Как уже было отмечено ранее, большинство инструментов *Data* Mining позволяют импортировать *данные* из различных источников, а также экспортировать результирующие данные в различные форматы.

Данные для экспериментов удобно хранить в каком-то одном формате.

В некоторых инструментах *Data* Mining эти процедуры называются импорт/экспорт *данных*, другие позволяют напрямую открывать различные источники *данных* и сохранять результаты *Data* Mining в одном из предложенных форматов.

Наибольшее число опрошенных (23%) предпочитают хранить данные в формате той базы *данных*, которую они используют. В формате *Text*, *CSV* - 18%, *по* 14% опрошенных хранят *данные* в формате *Text*, *space* or *tab separated* и *SAS*; в формате *Excel* - 9%, SPSS - 8%, S-Plus/R - 4%, Weka ARFF - 6%, в других форматах инструментов *Data* Mining - 2%.

Как видим из результатов опроса, наиболее распространенным форматом хранения *данных* для *Data* Mining выступают базы *данных*.

1. **Базы данных. Основные положения**

**БД**- совокупность взаимосвязанных, хранящихся вместе сведениях о различных сущностях одной предметной, обеспечивающая наличие такой минимальной избыточности, которая допускает их использование оптимальным образом для одного или нескольких приложений, или пользователей;

*Данные* организованы неким конкретным способом, способным облегчить их *поиск* и *доступ* к ним для одного или нескольких приложений. Также такая организация *данных* предусматривает наличие минимальной избыточности *данных*.

Базы *данных* являются одной из разновидностей информационных технологий, а также формой хранения *данных*.

Целью создания баз *данных* является построение такой системы *данных*, которая бы не зависела от программного обеспечения, применяемых технических средств и физического расположения *данных* в ЭВМ. Построение такой системы *данных* должно обеспечивать непротиворечивую и целостную информацию. При проектировании базы *данных* предполагается многоцелевое ее использование.

База *данных* в простейшем случае представляется в виде системы двумерных таблиц.

**Схема данных -** описание логической структуры данных, специфицированное на языке описания данных и обрабатываемое СУБД.

**Схема пользователя** - зафиксированный для конкретного пользователя один вариант порядка полей таблицы.

1. **СУБД**

**СУБД** - это программное обеспечение, контролирующее организацию, хранение, целостность, внесение изменений, чтение и безопасность информации в базе *данных*.

Система управления **реляционными базами данных** - это СУБД, основанная на реляционной модели *данных*.

В реляционной модели *данных* любое представление *данных* сводится к совокупности реляционных таблиц (двумерных таблиц особого типа). Системы управления реляционными базами *данных* используются для построения хранилищ *данных*.

К базам *данных*, а также к СУБД предъявляются такие требования: Высокое быстродействие; Простота обновления данных; Независимость данных; Возможность многопользовательского использования данных; Безопасность данных; стандартизация построения и эксплуатации БД (фактически СУБД); адекватность отображения *данных* соответствующей предметной области; дружелюбный интерфейс пользователя.

**Высокое быстродействие** предусматривает малое время отклика, т.е. малый промежуток времени от момента запроса к базе *данных* до момента реального получения *данных*.

**Независимость данных** - это возможность изменения логической и физической структуры базы *данных* без изменения представлений пользователей.

Независимость *данных* обеспечивает минимальные изменения структуры базы *данных* при изменениях стратегии доступа к данным и структуры самих исходных *данных*. Эти изменения должны быть предусмотрены на этапах концептуального и *логического проектирования* базы *данных* с обеспечением минимальных изменений на этапе физического ее проектирования.

**Безопасность данных** - это защита *данных* от преднамеренного или непреднамеренного нарушения секретности, искажения или разрушения. Безопасность включает два компонента: целостность и защиту *данных* от несанкционированного доступа.

**Целостность данных** - устойчивость хранимых *данных* к разрушению и уничтожению, связанным с неисправностями технических средств, системными ошибками и ошибочными действиями пользователей.

Целостность *данных* - точность и валидность *данных*. Целостность *данных* предполагает: отсутствие неточно введенных *данных*, защиту от ошибок при обновлении баз *данных*; невозможность удаления (или *каскадное удаление*) связанных *данных* разных таблиц; сохранность *данных* при сбоях техники (возможность восстановления *данных*) и др.

**Защита данных** от несанкционированного доступа предполагает ограничение доступа к определенным данным базы и достигается введением мер безопасности: разграничение прав доступа к данным различных пользователей в зависимости от выполняемых ими функций и/или должностных обязанностей; введением защиты в виде паролей; использованием представлений, т.е. таблиц, которые являются производными от исходных и предназначены для работы конкретных пользователей для решения конкретных задач.

**Стандартизация** обеспечивает преемственность поколений конкретной СУБД, упрощает взаимодействие баз *данных* одного поколения СУБД с одинаковыми и различными моделями *данных*.

СУБД отвечает за обработку запросов к базе *данных* и получение ответа. Способы хранения *данных* могут быть различными: модель *данных* может быть как реляционной, так и многомерной, сетевой или иерархической.

1. **Классификация видов данных**

**Реляционные данные** - это данные из реляционных баз (таблиц).

**Многомерные данные** - это данные, представленные в кубах *OLAP*.

**Измерение (dimension) или ось** - в многомерных *данных* - это собрание *данных* одного и того же типа, что позволяет структурировать многомерную базу *данных*.

*По* критерию постоянства своих значений в ходе решения задачи данные могут быть:

* переменными;
* постоянными;
* условно-постоянными.

**Переменные данные** - это такие данные, которые изменяют свои значения в процессе решения задачи.

**Постоянные данные** - это такие данные, которые сохраняют свои значения в процессе решения задачи (математические *константы*, *координаты* неподвижных *объектов*) и не зависят от внешних факторов.

**Условно-постоянные данные** - это такие данные, которые могут иногда изменять свои значения, но эти изменения не зависят от процесса решения задачи, а определяются внешними факторами.

Данные, в зависимости от тех функций, которые они выполняют, могут быть **справочными**, **оперативными**, **архивными**.

Следует различать данные за период и точечные данные. Эти различия важны при проектировании системы сбора информации, а также в процессе измерений.

* Данные за период;
* Точечные данные.

**Данные за период** характеризуют некоторый период времени. Примером *данных* за период могут быть: *прибыль* предприятия за месяц, средняя температура за месяц.

**Точечные данные** представляют *значение* некоторой переменной в конкретный момент времени. Пример точечных *данных*: *остаток* на счете на первое число месяца, температура в восемь часов утра.

Данные бывают первичными и вторичными. **Вторичные данные** - это данные, которые являются результатом определенных вычислений, примененных к **первичным данным**. Вторичные данные, как правило, приводят к ускоренному получению ответа на *запрос* пользователя за счет увеличения объема хранимой информации.

1. **Метаданные**

**Метаданные (Metadata)** - 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*, 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.

1. **Принципы обработки и анализа данных**

В основе обработки и анализа данных лежат математические методы, которые в большинстве своем являются неизменными уже в течение многих десятилетий. Соответственно неизменны и общие принципы, и последовательность действий при обработке данных. Однако технология обработки данных меняется, и существенно. В первую очередь это связано с совершенствованием технических средств проведения вычислений (листок бумаги, логарифмическая линейка, калькулятор, компьютер).

Компьютерный анализ медицинских данных предполагает некоторое математическое преобразование данных с помощью определенных программных средств. Необходимо иметь представление, как о математических методах обработки данных, так и о соответствующих программных средствах.

Как уже отмечалось, математические, статистические методы, реально применяемые на практике, за последние 30 лет существенно не изменились. Однако, благодаря использованию компьютеров, значительно расширился круг применяемых методов, и, соответственно, возникла необходимость в овладении основами этих методов медицинскими работниками.

Напротив, соответствующее программное обеспечение за это время изменилось очень сильно. Средства обработки данных появились одновременно с первыми вычислительными машинами - раньше операционных систем, редакторов 1 электронных таблиц. Со сменой поколений ЭВМ и даже чаще менялись поколения программных средств обработки данных, И, если возможности первых ЭВМ по анализу данных не превосходили возможности современных средних калькуляторов, то в 70-е годы появились пакеты, содержащие практически все те математические методы обработки, которые включены и в современные пакеты (88Р, ВМОР и др.). дальнейшее развитие пакетов обработки данных шло по пути совершенствования технологии обработки и анализа данных.).

## 5.2 Современная технология анализа данных

Совершенствование технических средств приводит к изменению относительной трудоемкости различных этапов процесса обработки анализа, что также ведет к изменению технологии обработки данных вообще и медицинских, в частности. В те сравнительно недавние времена, когда обработка данных осуществлялась вручную, самым трудоемким процессом был этап собственно статистических вычислений, расчетов по различным формулам. На этом этапе было сосредоточено внимание специалистов, предлагались различные упрощенные варианты расчетов, более простые методы, специально приспособленные для ручного счета и т. д.

Затем с появлением первых компьютерных пакетов технология основывалась на принципе командной строки и требовала довольно приличных знаний статистики и владения компьютером на уровне программиста.

Далее развитие пошло по пути использования меню и готовых процедур, что резко снизило требования как к знанию статистики, так и к владению компьютером.

И, наконец, в последнее время продолжилось улучшение интерфейса с пользователем, активнее используется графический подход, важное значение приобретает визуализация данных, что еще больше облегчает обработку данных неспециалисту.

В настоящее время, благодаря использованию компьютеров, вычислительный этап стал наименее трудоемким. Облегчились и другие этапы обработки данных. На первое место по относительной трудоемкости вышли другие этапы: освоение статистического пакета, этап подготовки данных к анализу, этап предварительного анализа данных и этап интерпретации результатов. Все в целом привело к изменению технологии обработки и анализа данных. При этом для применения основных методов обработки данных от исполнителя требуется лишь выполнение определенных статистических правил и грамотное использование пакета. Врачу не нужно углубляться в сложность математических определений, а следует понять, для чего и как эти методы используются. Реализации такого подхода и будет посвящено дальнейшее изложение.

На практике для врача в настоящее время обработка и анализ данных сводится к решению следующих задач:

1) получение представления об основных статистических методах,

2) освоение пакета,

3) собственно анализ данных.

Если ограничиваться минимальными представлениями о статистических методах (см. п. 2.2), то освоение пакета анализа является одним из наиболее трудоемких-эТаков обработки данных. -

Собственно, анализ данных с использованием статистического пакета (работа с пакетом, собственно технология анализа данных) включает следующие разделы.

1. Планирование исследования

2. Подготовка данных к анализу

3. Предварительный (разведочный) анализ данных

4. Выбор метода анализа и его реализация

5. Интерпретация результатов

6. Представление результатов

*Планирование исследования*. Наиболее предпочтительным случаем является такой, когда еще до проведения исследования уже существует определенная ясность о предполагаемых к использованию в дальнейшем методах обработки данных. В этом случае обычно удается спланировать исследование с учетом последующей обработки данных и избежать ситуаций, когда оказывается, что какие-то наблюдения были лишними, а каких-то не хватает для реализации выбранных методов анализа.

К сожалению, на практике на начальных этапах исследования часто еще нет полной ясности о методах обработки результатов исследований. Поэтому, следует представлять себе наиболее часто используемые методы обработки медицинских данных и требования к исходному материалу, предъявляемые ими. Для первоначального выбора метода обработки может быть использована таблица.

*Подготовка данных к анализу* — это крайне важный, зачастую недооцениваемый этап работы. Обычно он включает: ввод данных, предварительное преобразование данных, визуализацию данных с целью формирования представления об исследуемом материале. В настоящее время практически отпадает необходимость в предварительных структурировании, построении необходимых выборок, ранжировании и т. д. Все эти задачи в современных пакетах автоматизированы и выполняются непосредственно при реализации выбранного метода анализа. На этом этапе остаются только необходимые преобразования данных и их визуализация, которые тоже существенно облегчены. Важное значение приобретает *предварительный анализ данных* (или разведочный анализ данных). На этом этапе формируются представления о типе анализируемых данных, когда выясняется структура, определяются зависимости между данными, производится их группировка и, если это не было ясно с самого начала, осуществляется предварительный выбор методов анализа. В простейших случаях обработка данных может ограничиваться этапом предварительного анализа.

*Выбор и реализация метода анализа* в связи с многообразием методов может оказаться нетривиальной задачей. Однако в современных пакетах введенные данные достаточно просто обработать с использованием различных процедур, а затем можно выбрать метод, дающий наилучшие результаты.

*Интерпретация результатов анализа* часто вызывает затруднения у исследователей-медиков в связи с ограниченностью знаний в области статистики. Поэтому к этому этапу следует относиться особенно внимательно и, по возможности, быть предельно точными в следовании указаниям руководств. Это же относится и к практическим рекомендациям и выводам, которые делаются на основе результатов статистического анализа.

*Представление результатов* является одним из важнейших компонентов качества применения статистических методов. Поэтому пол нота и уровень описания, как самого анализа, так и его результатов, наглядность их представления не должны снижать общий уровень выполненной работы, что особенно важно при оформлении диссертаций.

1. **Структурированные данные**

Структурированные типы характеризуются множественностью образующих этот тип элементов, т.е. имеют несколько компонентов. Каждый компонент, в свою очередь может принадлежать структурированному типу, т.е. допускается вложенность типов.

**Массивы** представляют собой формальное объединение нескольких однотипных объектов (чисел, символов, строк и т.п.), рассматриваемое как единое целое. Все компоненты массива – это данные одного типа.

**Строки**– это массив символов, но количество символов в строке может меняться. Строка трактуется как цепочка символов произвольной длины. Максимальное количество символов не более 255. Каждый символ в строке имеет свой индекс (номер).

**Запись** – это структура данных, состоящая из фиксированного числа компонентов, называемых полями записи. В отличие от массива, компоненты записи (поля) могут быть различного типа. Записи позволяют объединять значения различных типов.

**Множества** – это наборы однотипных, логически связанных друг с другом объектов. Количество элементов, входящих в множество может меняться от 0 до 256. Именно непостоянством своих элементов множества отличаются от массивов и записей.

**Файл** – именованная область внешней памяти. Файл содержит компоненты одного типа, кроме файлов (т.е. нельзя создать «файл файлов»). Длина файла не оговаривается и ограничивается только ёмкостью устройств внешней памяти.

Более подробно со структурированными типами ознакомимся при дальнейшем изучении языка.

## Указатель (ссылочный тип)

Содержит адрес байта памяти, в котором находится значение данных определённого типа. Этот тип называют также ссылочным.

## Константы

**Константой** называют величину, значение которой не меняется в процессе выполнения программы.

* **Числовые**константы служат для записи чисел. Различают следующие их виды:

1. **Подготовка данных к анализу и обработки**

Обработка и анализ маркетинговой информации представляют собой серьезный этап статистической обработки полученных результатов. От того, насколько правильно и качественно будет произведена обработка данных, зависит достоверность принимаемых выводов и последующих практических решений. Анализ собранной информации представляет следующий этап маркетингового исследования, т.е. извлечение из совокупности полученных данных наиболее важных сведений и результатов. Исследователь сводит полученные данные в таблицы. На основе этих таблиц выводят или рассчитывают показатели эмпирического распределения, сравнивают их со стандартными. Затем исследователь решает, какие методы статистики использовать. Для получения дополнительных сведений применяют современные статистические методики и модели.

Обработка и анализ маркетинговой информации включает: предварительные этапы (редактирование, кодирование, табулирование и представление табулированных данных); оценку различий (проверка согласия, проверка Колмогорова — Смирнова, анализ средних выборки); выбор методов исследования (простой регрессионный и корреляционный анализ, множественный регрессионный анализ).

**1 этап. Проверка**

**2 этап. Редактирование данных.**

**3 этап. Кодирование данных.**

**4 этап. Преобразование и табулирование данных.**

1. Технологии KDD и Data mining (DM)

Knowledge discovery in databases (дословно, «обнаружение знаний в базах данных») - аналитический процесс исследования человеком большого объема информации с привлечением средств автоматизированного исследования данных с целью обнаружения скрытых в данных структур или зависимостей. Предполагается полное или частичное отсутствие априорных представлений о характере скрытых структур и зависимостей. KDD включает предварительное осмысление и неполную формулировку задачи (в терминах целевых переменных), преобразование данных к доступному для автоматизированного анализа формату и их предварительную обработку, обнаружение средствами автоматического исследования данных (data mining) скрытых структур или зависимостей, апробация обнаруженных моделей на новых, не использовавшихся для построения моделей данных и интерпретация человеком обнаруженных моделей.

Data mining (дословно, «разработка данных») - исследование и обнаружение “машиной” (алгоритмами, средствами искусственного интеллекта) в сырых данных скрытых структур или зависимостей, которые ранее не были известны, нетривиальны, практически полезны, доступны для интерпретации человеком.

Направленная на решение этих проблем - это технология knowledge discovery in databases (KDD). KDD - это область, впитавшая в себя последние достижения искусственного интеллекта, численных математических методов, статистики и эвристических подходов. Цель технологии - нахождение моделей и отношений, скрытых в базе данных, таких моделей, которые не могут быть найдены обычными методами. KDD позволяет увидеть такие взаимоотношения между данными, которые прежде даже не приходили в голову исследователю, а применение которых может способствовать увеличению эффективности работы и исследований, а в нашем конкретном случае позволит выявлять пакеты, являющиеся прямым следствием посягательством на вычислительные ресурсы.

1. Аналитические платформы

В отличие от СУБД с набором алгоритмов Data Mining, аналитические платформы изначально ориентированы на анализ данных и предназначены для создания готовых решений.

**Аналитическая платформа** — специализированное программное решение (или набор решений), которое содержит в себе все инструменты для извлечения закономерностей из сырых данных: средства **консолидации** информации в едином источнике (хранилище данных), **извлечения**, **преобразования**, **трансформации данных**, **алгоритмы Data Mining**, **средства визуализации** и **распространения результатов** среди пользователей, а также **возможности конвейерной обработки новых данных**.

В аналитической платформе, как правило, всегда присутствуют гибкие и развитые средства консолидации (объединения), включающие богатые механизмы интеграции с промышленными источниками данных, инструменты очистки и преобразования структурированных данных и их последующее хранение в едином источнике в многомерном хранилище данных. Модели, описывающие выявленные закономерности, также хранятся в специальном источнике данных — **репозитории моделей**.

Вообще говоря, приведенную на рис. систему можно построить с использованием нескольких программных решений, например, разделить функции извлечения/загрузки, OLAP-отчетности, хранилища данных, Data Mining между различным программным обеспечением. Но чтобы эти отдельные компоненты превратились в полноценную аналитическую систему, **необходимо произвести интеграцию между ними на уровне обмена данными, а еще лучше — метаданными.**

**Понятие «консолидация данных», основные задачи консолидации, основные подходы к организации хранения данных**

Прежде чем приступать к анализу данных, необходимо выполнить ряд процедур, цель которых — доведение данных до приемлемого уровня качества и информативности, а также организовать их интегрированное хранение в структурах, обеспечивающих их целостность, непротиворечивость, высокую скорость и гибкость выполнения аналитических запросов.

**Консолидация** — комплекс методов и процедур, направленных на извлечение данных из различных источников, обеспечение необходимого уровня их информативности и качества, преобразование в единый формат, в котором они могут быть загружены в хранилище данных или аналитическую систему.

1. Алгоритмы Data mining

Интеллектуальный анализ данных включает указанные ниже типы алгоритмов.

* **Алгоритмы классификации** осуществляют прогнозирование одной или нескольких дискретных переменных на основе других атрибутов в наборе данных.
* **Регрессивные алгоритмы** осуществляют прогнозирование одной или нескольких непрерывных числовых переменных, например прибыли или убытков, на основе других атрибутов в наборе данных.
* **Алгоритмы сегментации** делят данные на группы или кластеры элементов, имеющих схожие свойства.
* **Алгоритмы взаимосвязей** осуществляют поиск корреляции между различными атрибутами в наборе данных. Наиболее частым применением этого типа алгоритма является создание правил взаимосвязи, которые могут использоваться для анализа потребительской корзины.
* **Алгоритмы анализа последовательностей** обобщают часто встречающиеся в данных последовательности, такие как серия переходов по веб-сайту или событий, зарегистрированных в журнале перед ремонтом оборудования.

Однако ничто не заставляет пользователя ограничиваться одним алгоритмом в своих решениях. Опытные аналитики часто используют один алгоритм для выявления наиболее эффективных входных данных (то есть переменных), после чего применяют другой алгоритм для прогнозирования определенного результата на основе этих данных. SQL Server Интеллектуальный анализ данных позволяет на базе одной структуры интеллектуального анализа построить много моделей таким образом, что в рамках одного решения для интеллектуального анализа данных можно было использовать алгоритм кластеризации, модель дерева решений, а также модель упрощенного алгоритма Байеса для получения разных представлений данных. В одном решении также можно использовать несколько алгоритмов для выполнения отдельных задач. Например, с помощью регрессии можно получать финансовые прогнозы, а с помощью алгоритма нейронной сети выполнять анализ факторов, влияющих на прогнозы.

1. Консолидация данных

***Консолидация*** — это объединение данных из одной или нескольких областей данных и вывод их в виде таблицы в итоговом листе.

Существует два основных способа консолидации данных.

* **Консолидация по расположению.** Этот метод применяется, если данные из различных источников упорядочены одинаково и в них используются одни и те же подписи столбцов и строк (например, при наличии нескольких листов расходов, созданных из одного и того же шаблона).
* **Консолидация по категории.** Этот метод применим, если данные из различных источников упорядочены по-разному, но и в них используются одни и те же подписи столбцов и строк (например, при наличии листов с данными о запасах для каждого месяца, которые имеют одинаковый макет, но содержат разные элементы или разное число элементов).

**Консолидация данных по расположению** используется, если консолидируемые данные размещены в одном и том же порядке, т.е. таблицы идентичны.

**Консолидация данных по категориям** используется, если данные исходных областей не упорядочены, но имеют одни и те же заголовки. Технология этой консолидации совпадает с технологией консолидации данных по значению

1. Трансформация данных

Анализируемая информация, представленная в виде набора данных, имеет определенный формат. Для анализа различных аспектов информации может потребоваться изменение ее формата, или *трансформация*. *Трансформация данных* состоит из трех этапов, выполняемых в строгой последовательности (каждый из которых, однако, может быть пропущен).

***Квантование значений***

При выполнении этой операции осуществляется разбиение диапазона числовых значений на указанное количество интервалов определенным методом и замена каждого обрабатываемого значения на число, связанное с интервалом, к которому оно относится, либо на метку интервала. Интервалы разбиения включают в себя нижнюю границу, но не включают верхнюю, кроме последнего интервала, который включает в себя обе границы. Результатом преобразования может быть: номер интервала (от нуля до значения, на единицу меньшего количества интервалов), значение нижней или верхней границы интервала разбиения, среднее значение интервала разбиения, метка интервала.

*Квантование* может быть осуществлено интервальным или квантильным методом.

Интервальное *квантование* подразумевает разбиение диапазона значений на указанное количество значений равной длины. Например, если значения в поле попадают в диапазон от 0 до 10, то при интервальном *квантовании* на 10 интервалов мы получим отрезки от 0 до 1, от 1 до 2 и т.д. При этом 0 будет относиться к первому интервалу, 1 - ко второму, а 9 и 10 - к десятому.

Квантильное *квантование* подразумевает разбиение диапазона значений на равновероятные интервалы, то есть на интервалы, содержащие равное (или, по крайней мере, примерно равное) количество значений. Нарушение равенства возможно только тогда, когда значения, попадающие на границу интервала, встречаются в наборе данных несколько раз. В этом случае все они относятся к одному определенному интервалу и могут вызвать "перевес" в его сторону.

**Табличная замена значений**

В результате выполнения этой операции производится замена значений по таблице подстановки, которая содержит пары, состоящие из исходного и выходного значения. Например, 0 - "красный", 1 - "зеленый", 2 - "синий". Или "зима" - "январь", "весна" - "апрель", "лето" - "июль", "осень" - "октябрь". Для каждого значения исходного набора данных ищется соответствие среди исходных значений таблицы подстановки. Если соответствие найдено, то значение меняется на соответствующее выходное значение из таблицы подстановки. Если значение не найдено в таблице, оно может быть либо заменено значением, указанным для замены "по умолчанию", либо оставлено без изменений (если такое значение не указано).

" *Скользящее окно* "

При решении некоторых задач, например, при прогнозировании временных рядов с помощью нейросети, требуется подавать на вход анализатора значения несколько смежных отсчетов из исходного набора данных. Такой метод отбора данных называется ***скользящим окном*** (*окно* - поскольку выделяется только некоторый непрерывный участок данных, *скользящее* - поскольку это окно "перемещается" по всему набору). При этом эффективность реализации заметно повышается, если не выбирать данные каждый раз из нескольких последовательных записей, а последовательно расположить данные, относящиеся к конкретной позиции окна, в одной записи.

**Преобразование даты**

Разбиение даты необходимо для анализа всевозможных показателей за определенный период (день, неделя, месяц, квартал, год). Суть разбиения заключается в том, что на основе столбца с информацией о дате формируется другой столбец, в котором указывается, к какому заданному интервалу времени принадлежит строка данных. Тип интервала задается аналитиком, исходя из того, что он хочет получить, - данные за год, квартал, месяц, неделю, день или сразу по всем интервалам.

*Группировка*

Трудно делать какие-либо выводы по данным каждой записи в отдельности. Аналитику для принятия решения часто необходима сводная информация. Совокупные данные намного более информативны, тем более если их можно получить в разных разрезах. В *Deductor Studio* предусмотрен инструмент, реализующий сбор сводной информации, - " *Группировка* ". ***Группировка*** позволяет объединять записи по полям-измерениям, агрегируя данные в полях-фактах для дальнейшего анализа.

*Разгруппировка*

*Группировка* используется для объединения фактов по каким-либо измерениям. При этом под объединением понимается применение некоторой функции агрегации. Если в исходном наборе данных присутствовали какие-либо другие измерения, то теряется информация о значениях фактов в разрезе этих измерений. Алгоритм *разгруппировки* позволяет восстановить эти факты, но их значения восстанавливаются не точно, а пропорционально вкладу в сгруппированные значения.

**Комплексная предобработка**

Термин "предобработка" можно трактовать шире, а именно, как процесс предварительного экспресс-анализа данных. Например, как оценить, является ли фактор значимым или нет, все ли факторы учтены для объяснения поведения результирующей величины и так далее. Для этих целей используются такие алгоритмы как корреляционный анализ, факторный анализ, метод главных компонент, регрессионный анализ. Подобный анализ в *Deductor Studio* называется комплексной предобработкой, в рамках которой осуществляется понижение размерности входных данных и/или устранение незначащих факторов.

**Понижение размерности пространства факторов**

Понижение размерности необходимо в случаях, когда входные факторы коррелированы друг с другом, т.е. взаимозависимы. Имеется возможность пересчитать их в другую систему координат, выделяя при этом главные компоненты. Понижение размерности получается путем отбрасывания компонент, в наименьшей степени объясняющих дисперсию результирующих значений (при этом предполагается, что исходные факторы полностью объясняют дисперсию результирующих факторов).

Требуется указать порог значимости, задающий дисперсию результата. Значение порога значимости может изменяться от 0 до 1.

**Устранение незначащих факторов**

Устранение незначащих факторов основано на поиске таких значений, которые в наименьшей степени коррелированы (взаимосвязаны) с выходным результатом. Такие факторы могут быть исключены из результирующего набора данных практически без потери полезной информации. Критерием принятия решения об исключении является порог значимости. Если корреляция (степень взаимозависимости) между входным и выходным факторами меньше порога значимости, то соответствующий фактор отбрасывается как незначащий.

1. Визуализация данных

На любом этапе обработки можно визуализировать данные. Система самостоятельно определяет, каким способом она может это сделать, например, если будет обучена нейронная сеть, то помимо таблиц и диаграмм можно просмотреть граф нейросети. Пользователю необходимо выбрать нужный вариант из списка доступных и настроить несколько параметров.

Возможные способы визуализации данных:

**1***Таблица*. Стандартное табличное представление с возможностью фильтрации данных, сортировки и быстрого расчета статистики (онлайн статистика).

**2***Статистика*. Статистические показатели выборки.

**3***Диаграмма*. График изменения любого показателя. Имеется возможность выбора различных вариантов диаграмм: столбчатые, линейные, круговые и прочее.

**4***Многомерная диаграмма*. Отображает данные в многомерном виде – поверхность или топографическим способом.

**5***Диаграмма размещения*– показывает объекты, размещенные в пространстве.

**6***Гистограмма*. График разброса показателей. Гистограмма предназначена для визуальной оценки распределения данных. Распределение данных оказывает значительное влияние на процесс построения модели. Кроме того, по гистограмме можно судить о величине отклонений различной степени (гистограмма распределения ошибок).

**7***Куб*. Многомерное представление данных. Любые данные, используемые в программе, можно посмотреть в виде кросс-таблицы и кросс-диаграммы. Пользователю доступен весь набор механизмов манипуляции многомерными данными – группировка, фильтрация, сортировка, произвольное размещение измерений, детализация, выбор любого способа агрегации, отображение в абсолютных числах и в процентах.

**8***Дубликаты и противоречия*. Специальный визуализатор, сделанный на основе таблицы, для более удобного отображения результатов поиска дубликатов и противоречий.

**9***Матрица корреляции*. Отображает зависимость (корреляцию) между входными и выходными полями обработчика «Корреляционный анализ».

**10***Граф нейросети*. Визуальное отображение обученной нейросети. Отображается структура нейронной сети и значения весов.

**11***Дерево решений*. Отображение дерева решений, полученного при помощи соответствующего алгоритма. Имеется возможность посмотреть детальную информацию по любому узлу и фильтровать попавшие в него данные.

**12***Правила*. Отображают в текстовом виде правила, полученные при помощи алгоритма построения деревьев решений или поиска ассоциаций. Такого рода информация легко интерпретируется человеком.

**13***Значимость атрибутов*. Отображается степень влияния каждого входного атрибута на результат построения дерева решений. Параметр значимость тем выше, чем больше вклад вносит конкретный входной атрибут при классификации выходного поля. Фактически данный визуализатор показывает степень нелинейной зависимости между выходным и входными полями.

**14***Карта Кохонена*. Отображение самоорганизующихся карт, построенных при помощи соответствующего алгоритма. Широкие возможности настройки – выбор количества кластеров, фильтрация по узлу/кластеру, выбор отображаемых полей. Мощный и гибкий механизм отображения кластеризованных данных.

**15***Профили кластеров*. Статистическая информация по кластерам, которые получаются на выходе обработчиков «Карта Кохонена» и «Кластеризацияk-means»

**16***ROC-анализ*. Доступен после обработчика «Логистическая регрессия». Позволяет оценить прогностическую силу модели, рассчитать оптимальный порог отсечения, проанализировать модель на чувствительность и специфичность.

**17***Коэффициенты регрессии*. Отображает коэффициенты, рассчитанные при помощи алгоритма «Линейная регрессия» и «Логистическая регрессия».

**18***Популярные наборы*. Отображение наиболее часто встречающихся в ассоциативных правилах множеств в виде списка с возможностью фильтрации и сортировки.

**19***Дерево правил*. Отображение в иерархическом виде (в виде дерева) ассоциативных правил. Содержит всегда два уровня. На первом – условие, на втором – следствие правила (или наоборот).

**20***Что-если*. Таблица и диаграмма. Позволяют «прогонять» через построенную модель любые интересующие пользователя данные и оценить влияние того или иного фактора на результат. Активно используется для решения задач оптимизации;

**21***Обучающий набор*. Выборка, используемая для построения модели. Выделяются цветом данные, попавшие в обучающее и тестовое множество с возможностью фильтрации. Необходима для понимания того, какие записи и каким образом использовались при построении модели.

**22***Диаграмма прогноза*. Применяется после использования метода обработки – Прогнозирование. Прогнозные значения на диаграмме выделяются цветом;

**23***Таблица сопряженности*. Предназначена для оценки результатов классификации вне зависимости от используемой модели. Таблица сопряженности отображает результаты сравнения категориальных значений исходного выходного столбца и категориальных значений рассчитанного выходного столбца. Используется для оценки качества классификаторов.

**24***Диаграмма рассеяния*. График отклонения значений, прогнозируемых при помощи модели, от реальных. Может быть построена только для непрерывных величин и только после использования механизмов построения модели, например, нейросети, линейной регрессии или пользовательской модели. Используется для визуальной оценки качества построенной модели.

**25***Диаграмма сезонных индексов*. Показывает тренд и сезонные индексы после применения обработчика «Декомпозиция временного ряда».

**26***Сведения*. Текстовое описание параметров импорта/обработки/экспорта в дереве сценариев обработки.

**27***Метаданные*– визуализатор для отображения метаданных хранилища данных.

**28***Навигатор*– навигатор по объектам базы данных.

Практически все механизмы визуализации поддерживают экспорт результатов в таблицы (MS Excel, MS Word, HTML, текстовые файлы…) или в графические файлы (GIF, BMP, EMF…).

Настроенные визуализаторы могут быть вынесены на панель **Отчеты**. Таким образом, конечный пользователь сможет просто получить и просмотреть необходимый результат, не задумываясь, каким способом он был получен.

1. Препроцессинг данных: очистка и предобработка

Процедура подготовки данных к анализу в процессе которой они приводятся в соответствие с требованиями, определяемыми спецификой решаемой задачи.

Предобработка данных включает два направления: очистку и оптимизацию. Очистка производится с целью исключения факторов, снижающих качество данных и мешающих работе аналитических алгоритмов. Она включает обработку дубликатов, противоречий и фиктивных значений, восстановление и заполнение пропусков, сглаживание и очистку данных от шума, подавление и редактирование аномальных значений. Кроме этого, в процессе очистки восстанавливаются нарушения структуры, полноты и целостности данных, преобразуются некорректные форматы.

Оптимизация данных, как элемент предобработки, включает снижение размерности входных данных, выявление и исключение незначащих признаков. Основное отличие оптимизации от очистки в том, что факторы, устраняемые в процессе очистки, существенно снижают точность решения задачи или делают работу аналитических алгоритмов невозможной. Проблемы, решаемые при оптимизации, адоптируют данные к конкретной задаче и повышают эффективность их анализа.

Предобработка данных является важнейшим этапом аналитического процесса, и ее элементы выполняются на всех его шагах, начиная от OLTP-систем и заканчивая аналитическим приложением.

1. Классификация стадий DM

#### Классификация стадий Data Mining

*Data Mining* может состоять из двух [8] или трех стадий [9]:

**Стадия 1**. Выявление *закономерностей* ( ***свободный поиск*** ).

**Стадия 2**. Использование выявленных *закономерностей* для предсказания неизвестных значений ( ***прогностическое моделирование*** ).

В дополнение к этим стадиям иногда вводят стадию валидации [10], 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.

**Стадия 3**. ***Анализ исключений*** - стадия предназначена для выявления и объяснения аномалий, найденных в *закономерностях*.

Итак, процесс *Data Mining* может быть представлен рядом таких последовательных стадий [11]:

**СВОБОДНЫЙ ПОИСК (в том числе ВАЛИДАЦИЯ) ->**

**-> ПРОГНОСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ ->**

**-> АНАЛИЗ ИСКЛЮЧЕНИЙ**

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

*Cвободный поиск* представлен такими **действиями**:

* выявление *закономерностей* условной логики (conditional logic);
* выявление *закономерностей* ассоциативной логики (associations and *affinities*);
* выявление трендов и колебаний (*trends* and *variations*).

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 (*outcome* prediction);
* прогнозирование развития процессов (*forecasting*).

На третьей стадии *Data Mining* анализируются исключения или аномалии, выявленные в найденных *закономерностях*.

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

1. **Классификация методов DM**

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.

1. 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.

2. 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.

Логические *методы*, или *методы* логической индукции, включают: нечеткие запросы и анализы; символьные правила; деревья решений; генетические *алгоритмы*.

*Методы* этой группы являются, пожалуй, наиболее интерпретируемыми - они оформляют найденные *закономерности*, в большинстве случаев, в достаточно прозрачном виде с точки зрения пользователя. Полученные правила могут включать непрерывные и дискретные переменные. Следует заметить, что деревья решений могут быть легко преобразованы в наборы символьных правил путем генерации одного правила по пути от корня дерева до его *терминальной вершины*. Деревья решений и правила фактически являются разными способами решения одной задачи и отличаются лишь по своим возможностям. Кроме того, реализация правил осуществляется более медленными *алгоритмами*, чем индукция деревьев решений.

*Методы* кросс-табуляции: агенты, баесовские (доверительные) сети, кросс-табличная визуализация. Последний метод не совсем отвечает одному из свойств *Data Mining* - самостоятельному поиску *закономерностей* аналитической системой. Однако, предоставление информации в виде кросс-таблиц обеспечивает реализацию основной задачи *Data Mining* - поиск шаблонов, поэтому этот *метод* можно также считать одним из *методов* *Data Mining* [13].

*Методы* на основе уравнений.

*Методы* этой группы выражают выявленные *закономерности* в виде математических выражений - уравнений. Следовательно, они могут работать лишь с численными переменными, и переменные других типов должны быть закодированы соответствующим образом. Это несколько ограничивает применение *методов* данной группы, тем не менее они широко используются при решении различных задач, особенно задач прогнозирования.

Основные *методы* данной группы: статистические *методы* и нейронные сети

Статистические *методы* наиболее часто применяются для решения задач прогнозирования. Существует множество *методов*статистического анализа данных, среди них, например, корреляционно-регрессионный анализ, корреляция рядов динамики, выявление тенденций динамических рядов, гармонический анализ.

Другая классификация разделяет все многообразие *методов* *Data Mining* на две группы: статистические и кибернетические *методы*. Эта схема разделения основана на различных подходах к обучению математических моделей [14].

Следует отметить, что существует два подхода отнесения статистических *методов* к *Data Mining*. Первый из них противопоставляет статистические *методы* и *Data Mining*, его сторонники считают классические статистические *методы* отдельным направлением анализа данных. Согласно второму подходу, статистические *методы* анализа являются частью математического инструментария *Data Mining*. Большинство авторитетных источников придерживается второго подхода [5, 14].

В этой классификации различают две группы *методов*:

* статистические *методы*, основанные на использовании усредненного накопленного опыта, который отражен в ретроспективных данных;
* кибернетические *методы*, включающие множество разнородных математических подходов.

Недостаток такой классификации: и статистические, и кибернетические *алгоритмы* тем или иным образом опираются на сопоставление статистического опыта с результатами мониторинга текущей ситуации.

Преимуществом такой классификации является ее удобство для интерпретации - она используется при описании математических средств современного подхода к *извлечению знаний* из массивов исходных наблюдений (оперативных и ретроспективных), т.е. в задачах *Data Mining*.

Рассмотрим подробнее представленные выше группы.

**Статистические методы Data mining**

В [14] эти *методы* представляют собой четыре взаимосвязанных раздела:

* предварительный анализ природы статистических данных (проверка гипотез стационарности, нормальности, независимости, однородности, оценка вида функции распределения, ее параметров и т.п.);
* выявление связей и *закономерностей* (линейный и нелинейный регрессионный анализ, корреляционный анализ и др.);
* многомерный статистический анализ (линейный и нелинейный дискриминантный анализ, кластерный анализ, компонентный анализ, *факторный анализ* и др.);
* *динамические модели* и прогноз на основе временных рядов.

Арсенал статистических *методов* *Data Mining* классифицирован на четыре группы *методов*:

1. Дескриптивный анализ и описание исходных данных.
2. *Анализ связей* (корреляционный и регрессионный анализ, *факторный анализ*, *дисперсионный анализ*).
3. Многомерный статистический анализ (компонентный анализ, дискриминантный анализ, многомерный регрессионный анализ, канонические корреляции и др.).
4. Анализ временных рядов (*динамические модели* и прогнозирование).

**Кибернетические методы Data Mining**

Второе направление *Data Mining* - это множество подходов, объединенных идеей компьютерной математики и использования теории искусственного интеллекта.

К этой группе относятся такие *методы*:

* *искусственные нейронные сети* (распознавание, кластеризация, прогноз);
* *эволюционное программирование* (в т.ч. *алгоритмы* метода группового учета аргументов);
* генетические *алгоритмы* (оптимизация);
* *ассоциативная память* (поиск аналогов, прототипов);
* нечеткая логика;
* деревья решений;
* системы обработки экспертных знаний.

Методы *Data Mining* также можно классифицировать по задачам *Data Mining*.

В соответствии с такой классификацией выделяем две группы. Первая из них - это подразделение *методов* *Data Mining* на решающие задачи сегментации (т.е. задачи классификации и кластеризации) и задачи прогнозирования.

В соответствии со второй классификацией по задачам *методы* *Data Mining* могут быть направлены на получение описательных и прогнозирующих результатов.

Описательные *методы* служат для нахождения шаблонов или образцов, описывающих данные, которые поддаются интерпретации с точки зрения аналитика.

К методам, направленным на получение описательных результатов, относятся итеративные *методы* кластерного анализа, в том числе: *алгоритм* k-средних, k-медианы, иерархические *методы* кластерного анализа, *самоорганизующиеся карты* Кохонена, *методы* кросс-табличной визуализации, различные *методы* визуализации и другие.

Прогнозирующие *методы* используют значения одних переменных для предсказания/прогнозирования неизвестных (пропущенных) или будущих значений других (целевых) переменных.

К методам, направленным на получение прогнозирующих результатов, относятся такие *методы*: нейронные сети, деревья решений, линейная регрессия, метод ближайшего соседа, метод *опорных векторов* и др.

1. **Задачи DM. Информация и знания**

### Информация

*Информация* - любые, неизвестные ранее сведения о каком-либо событии, сущности, процессе и т.п., являющиеся объектом некоторых операций, для которых существует содержательная *интерпретация*.

Под операциями здесь подразумевается восприятие, передача, преобразование, хранение и использование. Для восприятия *информации* необходима некоторая воспринимающая система, которая может интерпретировать ее, преобразовывать, определять соответствие определенным правилам и т.п. Таким образом, понятие *информации* следует рассматривать только при наличии источника и получателя *информации*, а также канала связи между ними.

#### Свойства информации

* Полнота *информации*.
* Достоверность *информации*.
* Ценность *информации*.
* Адекватность *информации*.
* Актуальность *информации*.
* Ясность *информации*.
* Доступность *информации*.
* Субъективность *информации*.

### Знания

***Знания*** - совокупность фактов, закономерностей и эвристических правил, с помощью которых решается поставленная задача.

Все чаще истинные *знания* образуются на основе распределенных взаимосвязей разнородной *информации* [19]. Когда *информация* собрана и передана для получения явно неопределенного заранее результата, то вы получаете *знания*. Сама *по* себе *информация* в чистом виде бессмысленна. Отсюда следует *вывод*, что *информация* - это чье-то тактическое *знание*, передаваемое в виде символов и при помощи каких-либо прикладных средств.

*По* определению Денхема Грэя, " *знания* - это абсолютное использование *информации* и данных, совместно с потенциалом практического опыта людей, способностями, идеями, интуицией, убежденностью и мотивациями".

*Знания* имеют определенные свойства, которые отличают их от *информации* [20].

1. **Структурированность**. *Знания* должны быть "разложены по полочкам".
2. **Удобство доступа и усвоения**. Для человека - это способность быстро понять и запомнить или, наоборот, вспомнить; для компьютерных знаний - средства доступа к *знаниям*.
3. **Лаконичность**. Лаконичность позволяет быстро осваивать и перерабатывать *знания* и повышает "коэффициент полезного содержания". В данный список лаконичность была добавлена из-за всем известной проблемы шума и мусорных документов, характерной именно для компьютерной *информации* - Internet и *электронного документооборота*.
4. **Непротиворечивость**. *Знания* не должны противоречить друг другу.
5. **Процедуры обработки**. *Знания* нужны для того, чтобы их использовать. Одно из главных свойств знаний - возможность их передачи другим и способность делать выводы на их основе. Для этого должны существовать процедуры обработки знаний. Способность делать выводы означает для машины наличие процедур обработки и вывода и подготовленность структур данных для такой обработки, т.е. наличие специальных форматов знаний.

1. **Задачи DM. Классификация**

#### Классификация задач 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. Эти типы моделей будут подробно описаны в лекции, посвященной процессу Data Mining.

В соответствии с этой классификацией, *задачи Data Mining* представлены группами *описательных* и *прогнозирующих задач*.

В результате решения *описательных (descriptive) задач* аналитик получает шаблоны, описывающие данные, которые поддаются интерпретации.

Эти задачи описывают общую концепцию анализируемых данных, определяют информативные, итоговые, отличительные особенности данных. Концепция *описательных задач* подразумевает характеристику и сравнение наборов данных.

Характеристика набора данных обеспечивает краткое и сжатое описание некоторого набора данных.

Сравнение обеспечивает сравнительное описание двух или более наборов данных.

***Прогнозирующие*** (predictive) основываются на анализе данных, создании модели, предсказании тенденций или свойств новых или неизвестных данных.

Достаточно близким к вышеупомянутой классификации является подразделение *задач Data Mining* на следующие: исследования и открытия, прогнозирования и классификации, объяснения и описания.

**Автоматическое исследование и открытие** ( *свободный поиск* )

Пример задачи: обнаружение новых сегментов рынка.

Для решения данного класса задач используются методы кластерного анализа.

*прогнозирование* и *классификация*

Пример задачи: предсказание роста объемов продаж на основе текущих значений.

Методы: регрессия, нейронные сети, *генетические алгоритмы*, деревья решений.

Задачи классификации и прогнозирования составляют группу так называемого индуктивного моделирования, в результате которого обеспечивается изучение анализируемого объекта или системы. В процессе решения этих задач на основе набора данных разрабатывается общая модель или гипотеза.

**Объяснение и описание**

Пример задачи: характеристика клиентов по демографическим данным и историям покупок.

Методы: деревья решения, системы правил, правила *ассоциации*, *анализ связей*.

Если доход клиента больше, чем 50 условных единиц, и его возраст - более 30 лет, тогда класс клиента - первый.

В интерпретации обобщенной модели аналитик получает новое знание. Группировка объектов происходит на основе их сходства.

#### Связь понятий

Итак, в предыдущей лекции нами были рассмотрены методы Data Mining и действия, выполняемые в рамках стадий Data Mining. Только что мы рассмотрели основные *задачи Data Mining*.

Напомним, что главная ценность Data Mining - это практическая направленность данной технологии, путь от сырых данных к конкретному знанию, от постановки задачи к готовому приложению, при поддержке которого можно принимать решения.

Многочисленность понятий, которые объединились в Data Mining, а также разнообразие методов, поддерживающих данную технологию, начинающему аналитику могут напомнить мозаику, части которой мало связаны между собой.

Как же мы можем связать в одно целое задачи, методы, действия, закономерности, приложения, данные, информацию, решения?

Рассмотрим два потока:

1. **ДАННЫЕ - ИНФОРМАЦИЯ - ЗНАНИЯ И РЕШЕНИЯ**
2. **ЗАДАЧИ - ДЕЙСТВИЯ И МЕТОДЫ РЕШЕНИЯ - ПРИЛОЖЕНИЯ**

Эти потоки являются "двумя сторонами одной медали", отображением одного процесса, результатом которого должно быть знание и принятие решения.

1. **Задачи DM. 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: обучение "без учителя" особого вида нейронных сетей - *самоорганизующихся карт* Кохонена

1. **Задачи DM. Статистические методы**

Основные *методы* данной группы: статистические *методы* и нейронные сети

Статистические *методы* наиболее часто применяются для решения задач прогнозирования. Существует множество *методов*статистического анализа данных, среди них, например, корреляционно-регрессионный анализ, корреляция рядов динамики, выявление тенденций динамических рядов, гармонический анализ.

Другая классификация разделяет все многообразие *методов* *Data Mining* на две группы: статистические и кибернетические *методы*. Эта схема разделения основана на различных подходах к обучению математических моделей [14].

Следует отметить, что существует два подхода отнесения статистических *методов* к *Data Mining*. Первый из них противопоставляет статистические *методы* и *Data Mining*, его сторонники считают классические статистические *методы* отдельным направлением анализа данных. Согласно второму подходу, статистические *методы* анализа являются частью математического инструментария *Data Mining*. Большинство авторитетных источников придерживается второго подхода [5, 14].

В этой классификации различают две группы *методов*:

* статистические *методы*, основанные на использовании усредненного накопленного опыта, который отражен в ретроспективных данных;
* кибернетические *методы*, включающие множество разнородных математических подходов.

Недостаток такой классификации: и статистические, и кибернетические *алгоритмы* тем или иным образом опираются на сопоставление статистического опыта с результатами мониторинга текущей ситуации.

Преимуществом такой классификации является ее удобство для интерпретации - она используется при описании математических средств современного подхода к *извлечению знаний* из массивов исходных наблюдений (оперативных и ретроспективных), т.е. в задачах *Data Mining*.

Рассмотрим подробнее представленные выше группы.

**Статистические методы Data mining**

В [14] эти *методы* представляют собой четыре взаимосвязанных раздела:

* предварительный анализ природы статистических данных (проверка гипотез стационарности, нормальности, независимости, однородности, оценка вида функции распределения, ее параметров и т.п.);
* выявление связей и *закономерностей* (линейный и нелинейный регрессионный анализ, корреляционный анализ и др.);
* многомерный статистический анализ (линейный и нелинейный дискриминантный анализ, кластерный анализ, компонентный анализ, *факторный анализ* и др.);
* *динамические модели* и прогноз на основе временных рядов.

Арсенал статистических *методов* *Data Mining* классифицирован на четыре группы *методов*:

1. Дескриптивный анализ и описание исходных данных.
2. *Анализ связей* (корреляционный и регрессионный анализ, *факторный анализ*, *дисперсионный анализ*).
3. Многомерный статистический анализ (компонентный анализ, дискриминантный анализ, многомерный регрессионный анализ, канонические корреляции и др.).
4. Анализ временных рядов (*динамические модели* и прогнозирование).
5. Задачи DM. Регрессия и регрессионный анализ
6. Метод наименьших квадратов в построении функциональных зависимостей
7. Метод группового учета аргументов
8. Задачи DM. Анализ и прогнозирование временных рядов
9. Процесс DM. Начальные этапы
10. Процесс DM. Очистка данных
11. Процесс DM. Построение и использование модели
12. Инструменты DM и инновационных технологий обработки информации
13. Методы классификации и прогнозирования.
14. Основные положения теории искусственных нейронных сетей. Понятие искусственной нейронной сети (НС). Наиболее
15. Биологический и искусственный нейрон. Сходство и различие.
16. Структура и свойства искусственного нейрона. Понятие, нейрона, виды функций активации. Общий принцип работы.
17. Классификация НС и их свойства. Виды нейронов. Топология НС. Однослойные и многослойные НС.
18. Слоистая топология НС: монотонная, сети без обратных связей (сети прямого распространения), сети с обратными связи. Классификация
19. Обучение и самообучение НС: понятие «обучение», процесс обучения, алгоритмы обучения. Обучение без учителя.
20. Этапы нейросетевого проектирования.
21. Нейронные сети встречного распространения: назначение, основные свойства, достоинства, недостатки.
22. Персептронные НС: свойства, структура, достоинства и недостатки. Обучение персептрона.
23. Нейронная сеть Хопфилда: структура, назначение, основные достинства и недостатки.
24. Самоорганизующиеся НС Кохонена: общая характеристика, задачи, решаемые данной сетью.
25. Предварительный подбор архитектуры НС.
26. Классификация НС по структуре нейронов.
27. Понятие искусственной нейронной сети (НС). Наиболее распространенные задачи, решаемые НС.