Лекция 1

Задачи интеллектуального анализа данных. Data Mining. Информация и знания. Методы и стадии Data Mining

Информация

Информация (лат. informatio) - это сведения, воспринимаемые человеком или специальными устройствами как отражение фактов материального или духовного мира в процессе коммуникации (*ГОСТ 7.0 – 99*).

в математике (кибернетике) - количественная мера устранения неопределенности (энтропия), мера организации системы;

в теории *информации* - раздел кибернетики, изучающий количественные закономерности, которые связаны со сбором, передачей, преобразованием и вычислением *информации*.

Информация

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

Операции: восприятие, передача, преобразование, хранение и использование.

Свойства информации

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

Понятие *информации* следует рассматривать только **при наличии источника и получателя** *информации*, а также канала связи между ними.

Данные

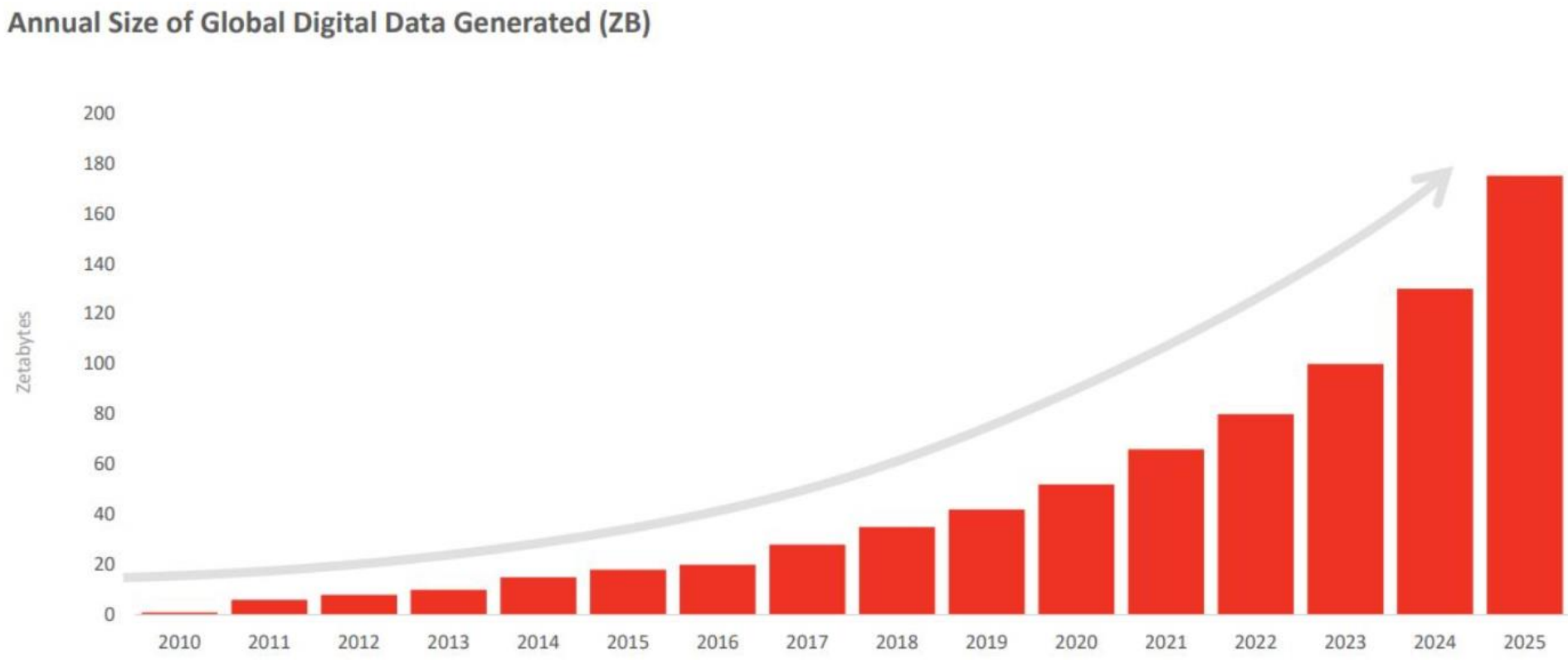
Данные могут представлять собой факты, понятия или команды, представленные в формализованном виде, позволяющем осуществить их передачу, интерпретацию или обработку [*Обработка данных. Словарь. Основные термины. – 1992*].

Данные — поддающееся многократной интерпретации представление информации в формализованном виде, пригодном для передачи, связи, или обработки [*определение по ISO/IEC 2382-1:1993*].

Данные

По прогнозам экспертов, к 2025 году общий объём информации в мире достигнет 163 зеттабайт (Збайт, равен 10^{21} байт).

Это будет соответствовать примерно десятикратному росту по сравнению с 2016-м.



Требования, предъявляемые к информации

- ✓ Динамический характер *информации*. Информация существует только в момент взаимодействия данных и методов, т.е. в момент информационного процесса. Остальное время она пребывает в состоянии данных.
- ✓ Адекватность используемых методов. *Информация* возникает и существует в момент диалектического взаимодействия объективных данных и субъективных методов.

Адекватность информации, соответствие ее содержания образу отображаемого объекта, может выражаться в трех формах:

- *синтаксической*;
- *семантической*;
- *прагматической*.

Требования, предъявляемые к информации

Синтаксическая адекватность связана с воспроизведением формальноструктурных характеристик отражения, абстрагирование от смысловых и потребительских параметров.

На синтаксическом уровне учитываются: *тип носителя, способ представления, скорость передачи и обработки, формат кодов, надежность и точность преобразования* и т.п.

При этом информация *инвариантна* по отношению к энергетическим и пространственно-временным свойствам своего носителя. Одна и та же информация может существовать в различных кодах.

Требования, предъявляемые к информации

Семантическая форма обеспечивает формирование понятий и представлений, выявление смысла, содержания информации.

Количество семантической информации в сообщении является величиной относительной: одно и то же сообщение может иметь смысловое содержание для компетентного пользователя и быть бессмысленным (семантическим шумом) для пользователя некомпетентного.

Прагматический аспект рассмотрения информации связан с ее *ценностью, полезностью, практическим использованием* для достижения целей деятельности системы.

Знания

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

По определению Денхема Грэя:

«Знания - это абсолютное использование информации и данных, совместно с потенциалом практического опыта людей, способностями, идеями, интуицией, убежденностью и мотивациями».

Свойства:

- **Структурированность**.
- **Удобство доступа и усвоения**.
- **Лаконичность**.
- **Непротиворечивость**.
- **Процедуры обработки**.

Одно из главных свойств знаний - возможность их передачи другим и способность делать выводы на их основе.

Сопоставление и сравнение понятий

- ***Информация***, в отличие от данных, имеет смысл.
- Понятия "*информация*" и "*знания*", с философской точки зрения, являются понятиями более высокого уровня, чем "*данные*", которое возникло относительно недавно.
- Понятие " *информации* " непосредственно связано с сущностью процессов внутри информационной системы, тогда так понятие "*знание*" скорее ориентировано на качество процессов.
- Понятие "*знание*" тесно связано с процессом *принятия решений.*
- Это части одного потока: у истока его находятся **данные**, в процессе передачи которых возникает ***информация***, и в результате использования *информации*, при определенных условиях, возникают ***знания*.**

Знания – это факты и правила, формализующие опыт специалистов в конкретной предметной области и позволяющие давать ответы (решения), которые не содержатся в исходной информации в явном виде.

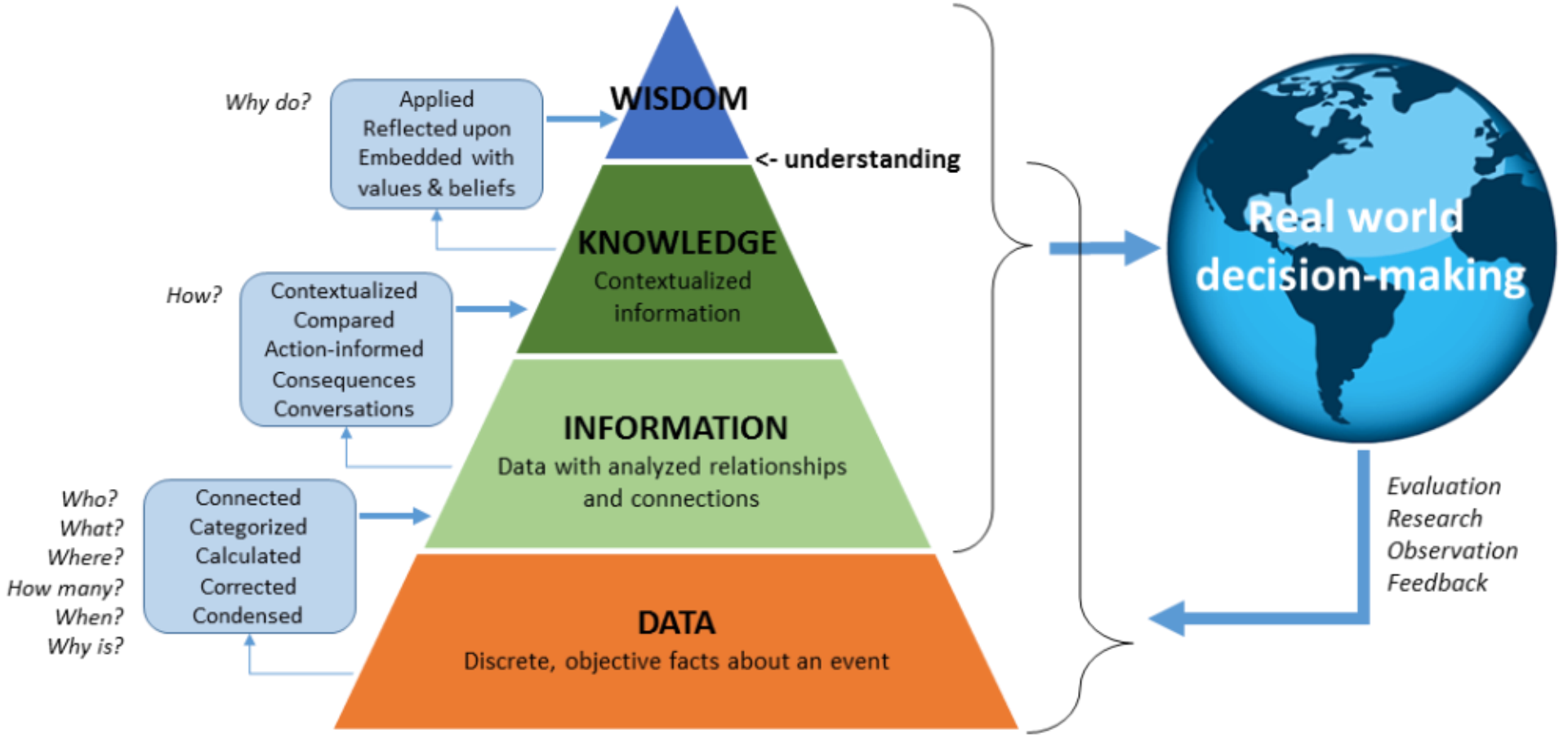
Главная задача анализа данных – преобразование данных в знание, т.е. в особый вид доступной для человеческого понимания информации.

Данные → Информация → Знания → «Глубокое понимание»

Модель DIKW

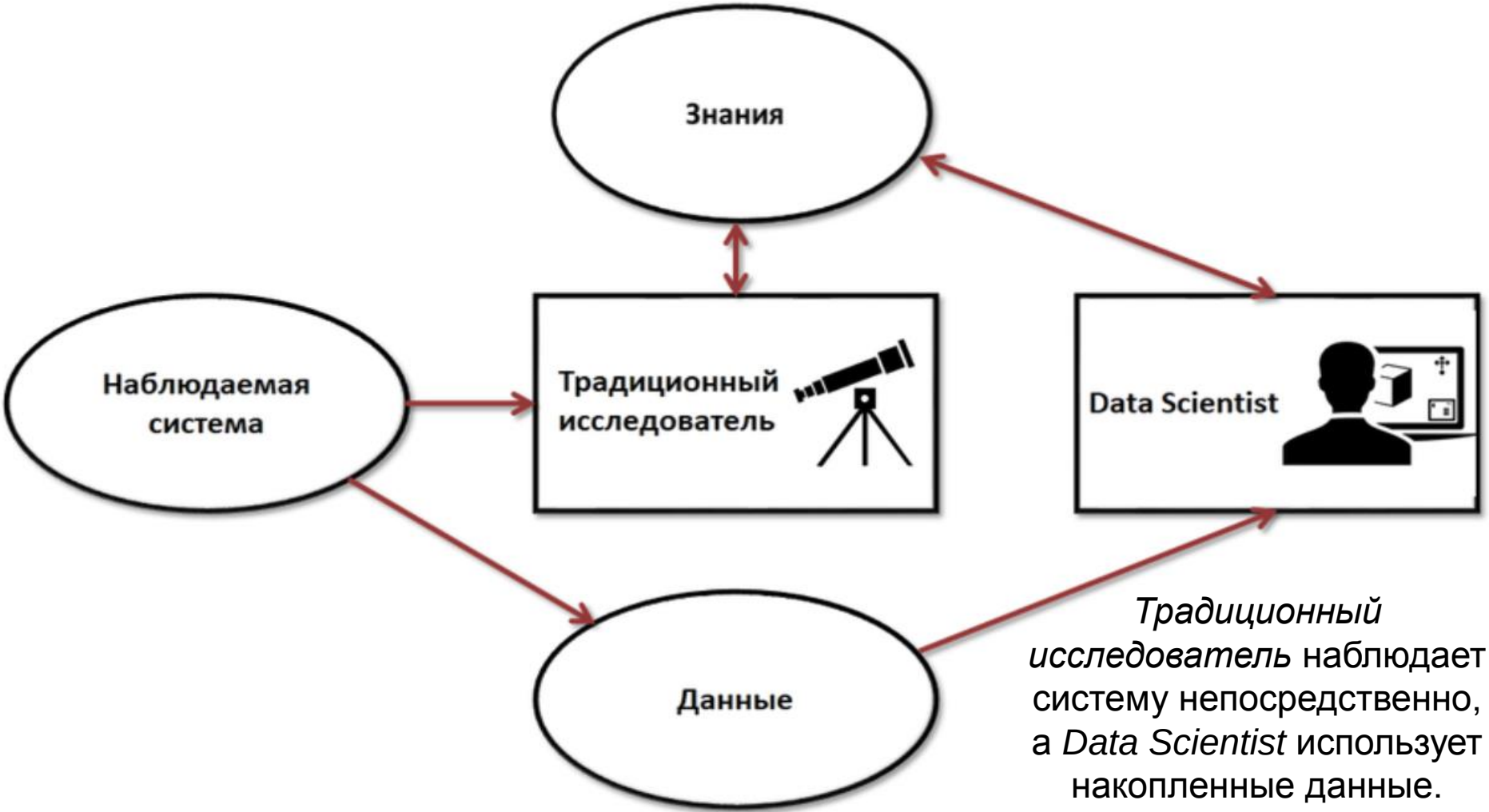
оставалась основой для исследований в области, которую называют «Управлением знаниями» (Knowledge Management, KM)

Knowledge Management → Data Science



Данные → Информация → Знания → «Глубокое понимание»

Data Science (в середине нулевых годов XXI века) определили как дисциплину, объединяющую в себе различные направления статистики, добычу данных (data mining), машинное обучение и применение баз данных для решения сложных задач, связанных с обработкой данных.



Традиционный исследователь наблюдает систему непосредственно, а *Data Scientist* использует накопленные данные.

Обобщенное иерархическое представление методологий обработки данных при принятии управленческих решений

Данные → Информация → Знания → «Глубокое понимание»

Анализ данных (АД) – это система подходов и методов, ориентированная на выявление механизма порождения представленных данных в рамках имеющейся априорной модели этого механизма.

Современные технологии анализа данных – новая парадигма процесса исследования данных, основанная на принципах, предложенных Джоном Тьюки:

- Анализ – это способ существования данных. Его материальная основа – системы «человек – машина».
- Принцип многократного возвращения к одним и тем же данным.
- Принцип множественности возможных моделей.
- Принцип варьирования предпосылок с рассмотрением последствий такого варьирования.
- Принцип множественности результатов и выбора на основе неформальных процедур принятия решений.
- Принцип полного использования эндогенной информации и максимального учета информации экзогенной.

История возникновения

Предпосылки:

- законы больших чисел для конечных выборок не выполняются;
- характеристики центральной тенденции (средняя арифметическая, мода, медиана) часто не являются характеристиками совокупности и приводят к операциям над фиктивными величинами (типа средней температуры больных по больнице, среднего дохода рабочих и миллионеров);
- закон распределения нельзя достоверно определить по выборочным данным;
- вероятность как характеристика неопределенности часто вводится необоснованно;
- сумма воздействия ненаблюдаемых и неконтролируемых факторов может привести к структурным изменениям в наблюдаемой системе, которые приведут к изменению априорных условий моделирования и т. д.
- «проклятие размерности» при анализе сложных систем, предполагающем исследование всей системы

История возникновения

Дж.Тьюки в 60-е годы предложил *разведочный анализ данных* (РАД; Exploratory data analysis), основанный на использовании методов многомерной статистики. РАД предполагает изучение не только вероятностной, но и геометрической природы данных.

РАД — анализ основных свойств данных, нахождение в них общих закономерностей, распределений и аномалий, построение начальных моделей.

Цели РАД:

- максимальное "проникновение" в данные
- выявление основных структур
- выбор наиболее важных переменных
- обнаружение отклонений и аномалий
- проверка основных гипотез
- разработка начальных моделей

Основные инструменты РАД:

- анализ вероятностных распределений переменных
- построение и анализ корреляционных матриц
- факторный анализ
- дискриминантный анализ
- многомерное шкалирование и др.

Дальнейшее развитие

1994 г. известный математик Лотфи Заде сформулировал принцип «мягких вычислений» - Soft Computing (терпимость к нечёткости и частичной истинности используемых данных для достижения интерпретируемости, гибкости и низкой стоимости решений)

Появление в середине 90-х годов XX века нового направления в науке - Data Mining (добыча данных), или иначе: интеллектуальный анализ данных.

- Идеология интеллектуального анализа данных (методы Data Mining) появилась на стыке ***прикладной статистики, искусственного интеллекта, баз данных*** и т. д.

Фактически рождению нового направления в анализе данных способствовало **появление компьютеров и совершенствование технологий записи и хранения данных**.

Понятие Data Mining

понятие *Data Mining* переводится на русский язык при помощи этих же трех понятий: как добыча ***данных***, извлечение ***информации***, раскопка ***знаний***.

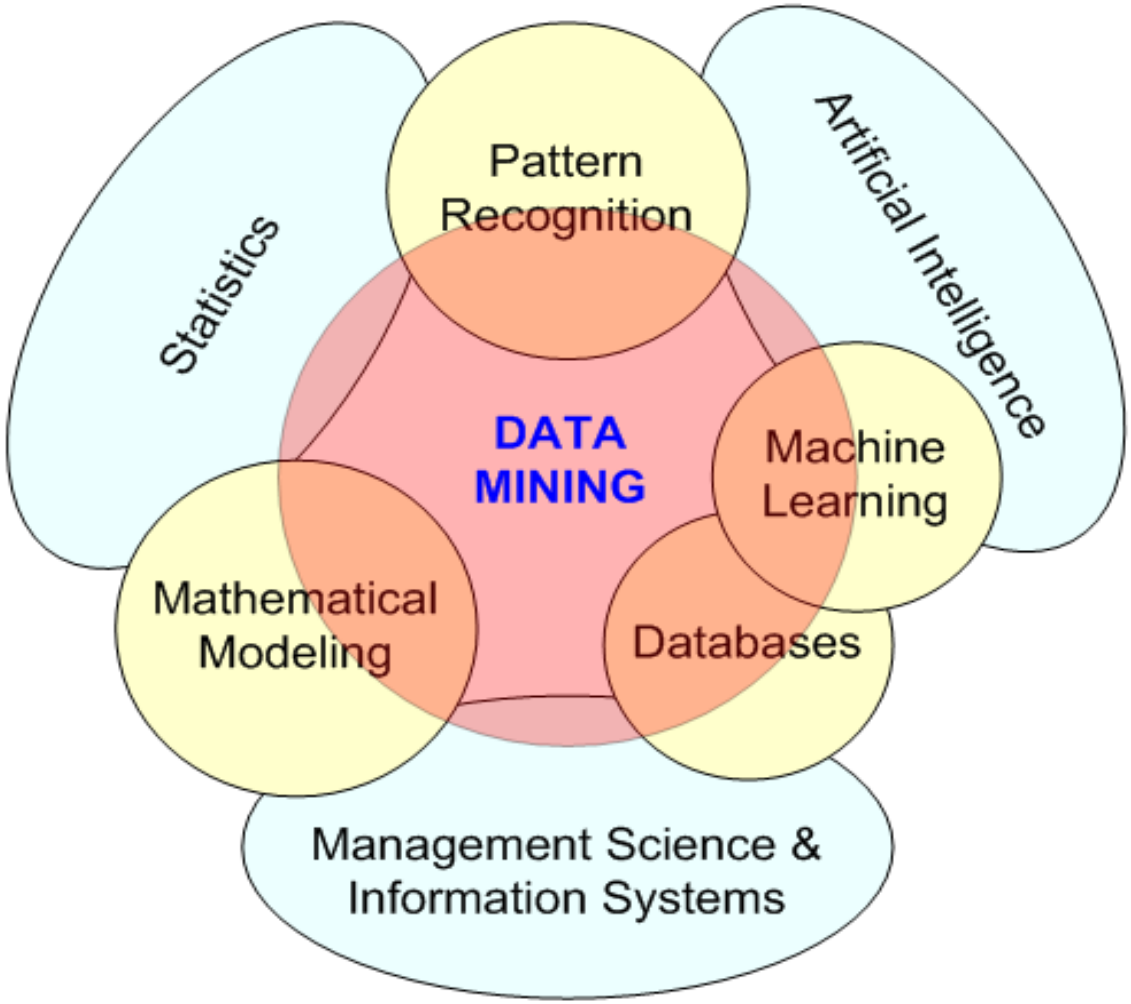
Data Mining - это процесс поддержки принятия решений, основанный на поиске в данных скрытых закономерностей (шаблонов информации).

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 технологии Data Mining можно охарактеризовать так: это технология, которая предназначена для поиска в больших объемах данных ***неочевидных, объективных и полезных на практике*** закономерностей.

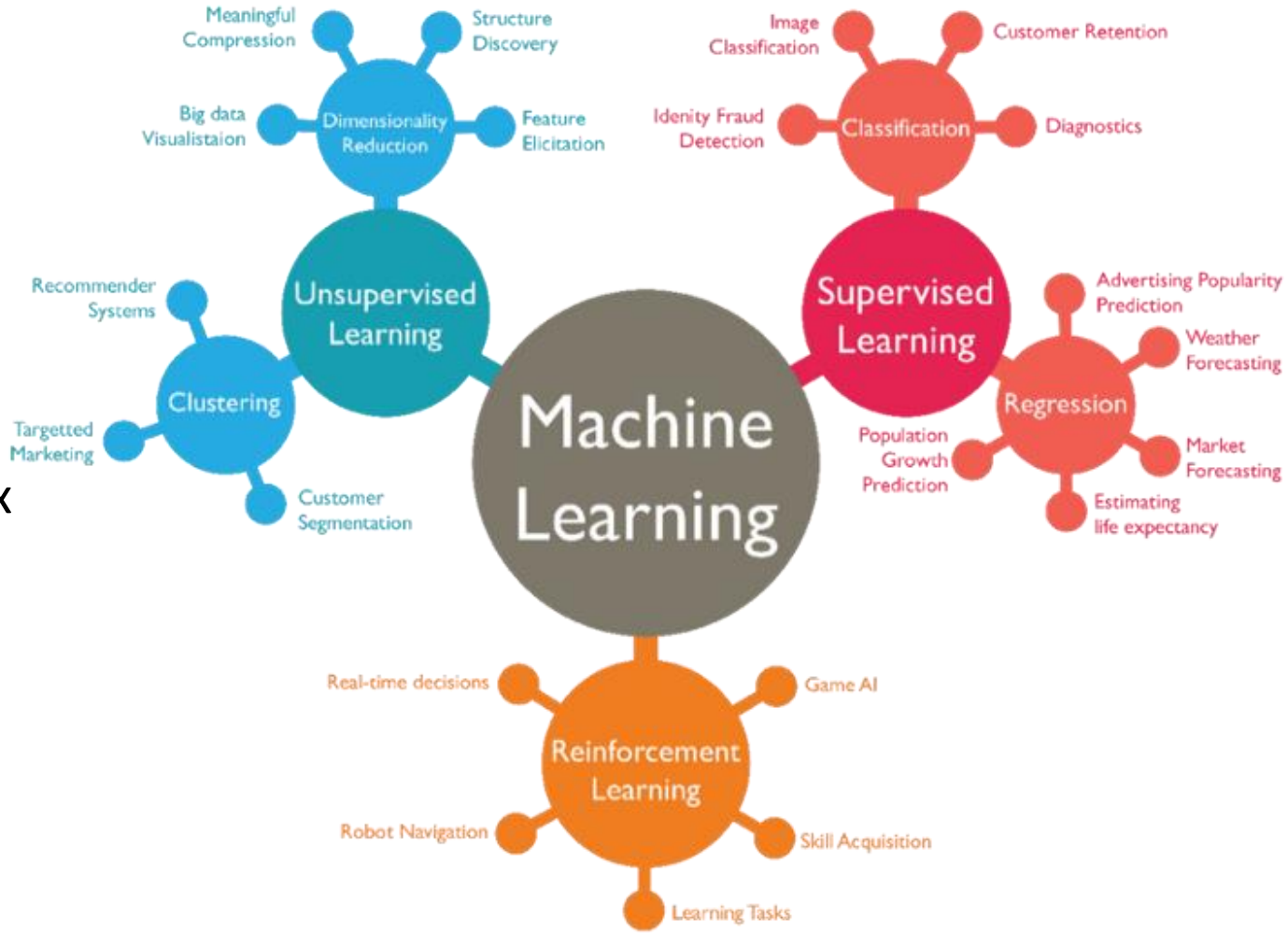
Понятие Data Mining

- **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.



Понятие Машинного обучения

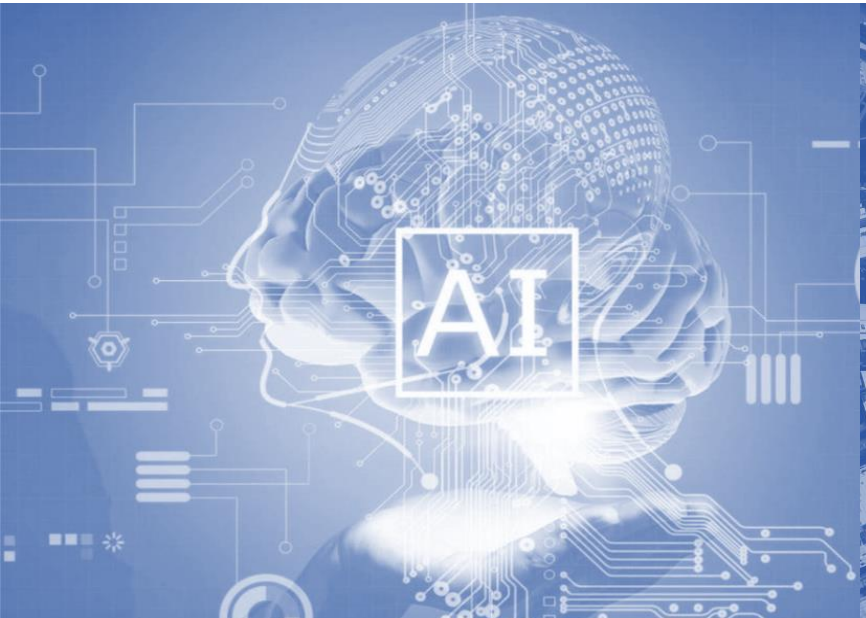
- Единого определения машинного обучения на сегодняшний день нет.
- **Машинное обучение** можно охарактеризовать как процесс получения программой новых знаний (1996 г.)
- «**Машинное обучение** - это наука, которая изучает компьютерные алгоритмы, автоматически улучшающиеся во время работы» /Митчелл/

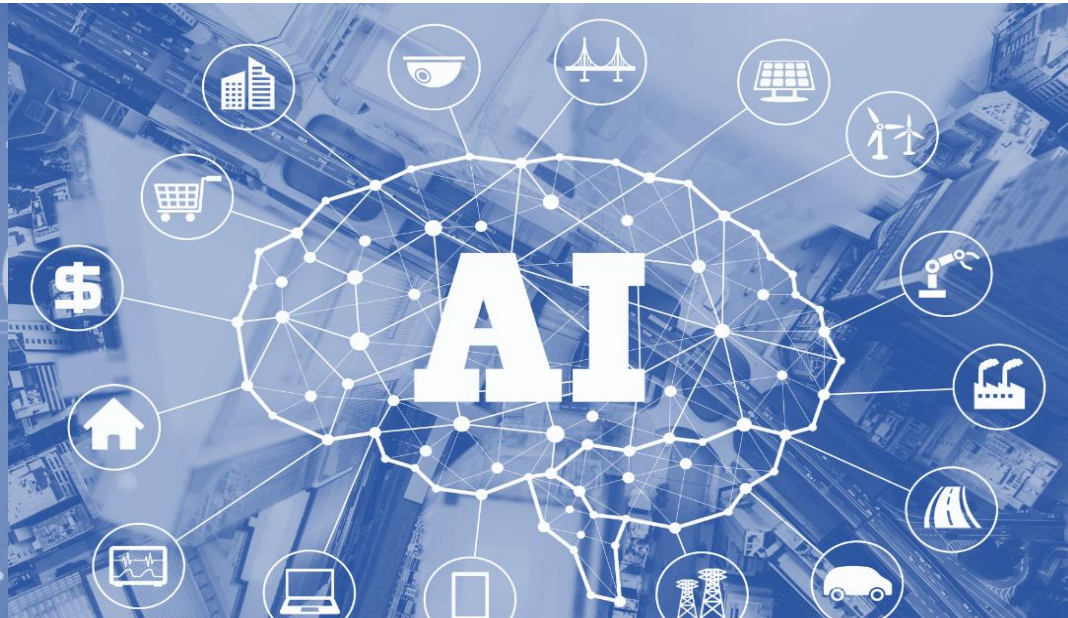


Одним из наиболее популярных примеров алгоритма машинного обучения являются нейронные сети.

Понятие Искусственного интеллекта

- **Искусственный интеллект** - научное направление, в рамках которого ставятся и решаются задачи аппаратного или программного моделирования видов человеческой деятельности, традиционно считающихся интеллектуальными.
- ***Искусственный интеллект*** (ИИ, artificial intelligence) — это общее понятие, описывающее «способность вычислительной машины моделировать процесс мышления за счет выполнения функций, которые обычно связывают с человеческим интеллектом»: построение и использование экспертных систем, логический вывод, понимание естественных языков, зрительное и слуховое восприятие [ГОСТ 15971 – 90. Системы обработки данных. Термины и определения].





Отличия Data Mining от других методов анализа данных

- *Традиционные методы* анализа данных (статистические методы и OLAP в основном *ориентированы на проверку заранее сформулированных гипотез* и на "грубый" разведочный анализ, составляющий основу оперативной аналитической обработки данных (OnLine Analytical Processing, OLAP), в то время как **одно из основных положений Data Mining - поиск неочевидных закономерностей**.

- **Инструменты Data Mining могут находить неочевидные закономерности самостоятельно** и также **самостоятельно строить гипотезы о взаимосвязях**. Поскольку именно формулировка гипотезы относительно зависимостей является самой сложной задачей, преимущество Data Mining по сравнению с другими методами анализа является очевидным.

- OLAP больше подходит для понимания ретроспективных данных **Data Mining опирается на ретроспективные данные** для получения ответов на вопросы о будущем.

Перспективы технологии Data Mining

выделение типов предметных областей с соответствующими им эвристиками, формализация которых облегчит решение соответствующих задач Data Mining, относящихся к этим областям;

создание формальных языков и логических средств, с помощью которых будут формализованы рассуждения и автоматизация которых станет инструментом решения задач Data Mining в конкретных предметных областях;

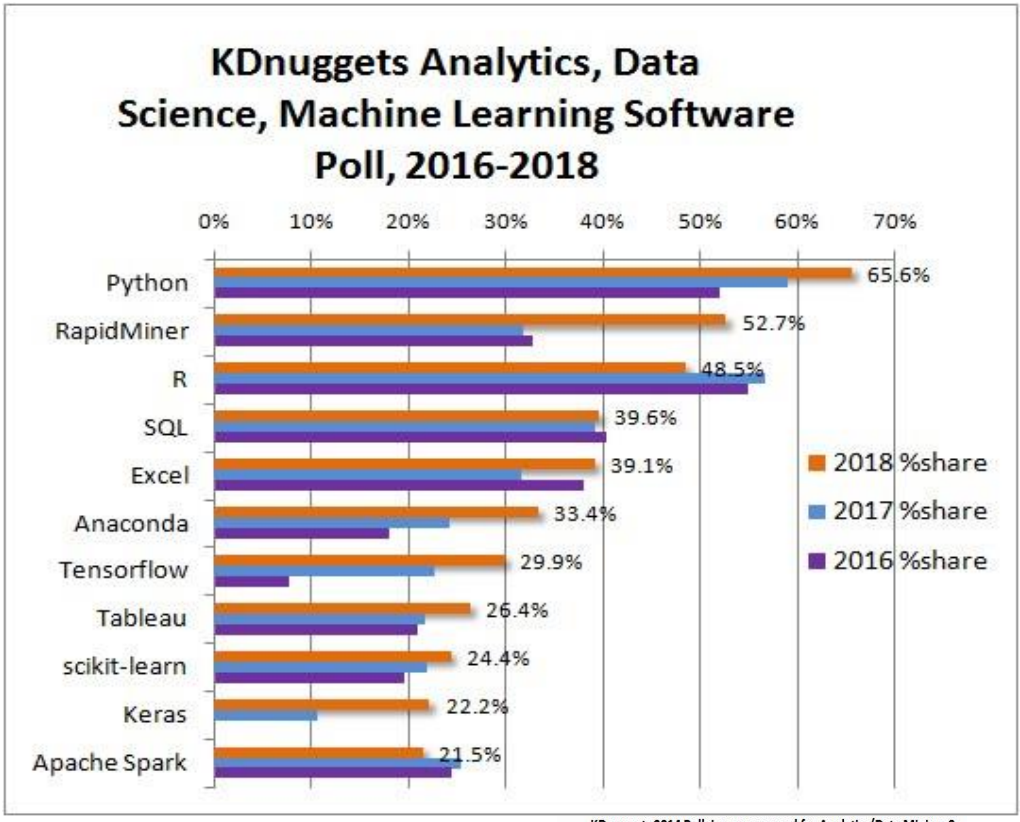
создание методов Data Mining, **способных** не только извлекать из данных закономерности, но и **формировать некие теории, опирающиеся на эмпирические данные**;

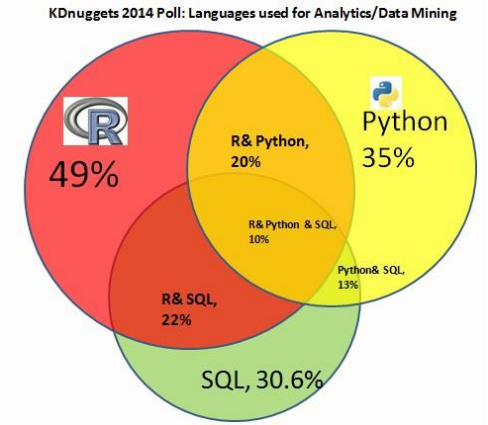
преодоление существенного отставания возможностей инструментальных средств Data Mining от теоретических достижений в этой области.

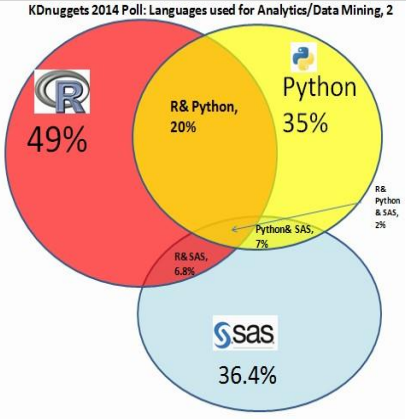
Инструментальные средства ИАД, Data Mining

интеллектуальный *анализ данных* может проводиться с помощью программных продуктов следующих классов:

- специализированных "коробочных" программных продуктов для интеллектуального анализа;
- математических пакетов;
- электронных таблиц (и различного рода надстроек над ними);
- средств интегрированных в системы управления базами данных (СУБД);
- других программных продуктов.







ОБЩИЕ ТИПЫ ЗАКОНОМЕРНОСТЕЙ ПРИ АНАЛИЗЕ ДАННЫХ

- Как правило, выделяют пять стандартных типов **закономерностей**, которые позволяют относить используемые методы к методам *Data Mining*:

1. Ассоциация.
2. Последовательность.
3. Классы.
4. Кластеры.
5. Временные ряды.

1. Ассоциация (англ. *Association*) имеет место в случае, если несколько событий связаны друг с другом.

2. В случае если несколько событий связаны друг с другом во времени, то имеет место тип зависимости, именуемый ***последовательностью*** (англ. *Sequential Patterns*).

3. Закономерность *классы* (англ. *Classes*) появляется в случае, если имеется несколько заранее сформированных классов (групп, типов) объектов. Отнесение нового объекта к какому-либо из существующих классов выполняется путем ***классификации***.

4. Закономерность *кластеры* (англ. *Clusters*) отличается тем, что классы (группы, типы) заранее не заданы, а их количество и состав определяются автоматически в результате процедуры ***кластеризации***.

5. Хранимая ретроспективная информация позволяет определить еще одну закономерность, *заключающуюся* в поиске существующих *временных рядов* (англ. *Time Series*) и ***прогнозировании*** динамики значений в них на будущие

ЗАКОНОМЕРНОСТИ В *DATA MINING*

- ***Ассоциация*** – это выделение различных типов связей между событиями: корреляционные связи, *if-then* правила и т.п.
- ***Последовательность*** – это ассоциация между событиями, сдвинутыми во времени.
- С помощью ***классификации*** выявляются признаки, характеризующие группу, к которой принадлежит тот или иной объект. Это делается посредством анализа уже классифицированных объектов и формулирования некоторого набора правил.
- ***Кластеризация*** отличается от классификации тем, что сами группы заранее не заданы. С помощью кластеризации средства Data Mining самостоятельно выделяют различные однородные группы данных.
- Основой для всевозможных систем ***прогнозирования*** служит историческая информация, хранящаяся в БД в виде временных рядов. Если удается построить найти шаблоны, адекватно отражающие динамику поведения целевых показателей, есть вероятность, что с их помощью можно предсказать и поведение системы в будущем.

КЛАССИФИКАЦИЯ (CLASSIFICATION)

В Data Mining ***задачу классификации*** рассматривают как задачу определения значения одного из параметров анализируемого объекта на основании значений других параметров. Определяемый параметр часто называют зависимой переменной, а параметры, участвующие в его определении - независимыми переменными.

Клиент банка: «кредитоспособен» и «некредитоспособен»;

Фильтр электронной почты: «спам», «не спам»

Распознавание цифр: от 0 до 9.

Если значениями независимых и зависимой переменных являются действительные числа, то задача называется ***задачей регрессии***.

Пример задачи регрессии - задача определения суммы кредита, которая может быть выдана банком клиенту.

Задача классификации и регрессии решается в два этапа:

1) выделяется обучающая выборка, в нее входят объекты, для которых известны значения как независимых, так и зависимых переменных.

На основании обучающей выборки строится модель определения значения зависимой переменной - функция классификации или регрессии.

2) построенную модель применяют к анализируемым объектам (к объектам с неопределенным значением зависимой переменной).

КЛАСТЕРИЗАЦИЯ (CLUSTERING)

Кластеризация является логическим продолжением идеи классификации.

Особенность кластеризации - классы объектов изначально не предопределены.

Результатом кластеризации является *разбиение* объектов на группы "похожих" объектов, называемых кластерами (cluster).

Пример метода решения задачи кластеризации: обучение "без учителя" особого вида нейронных сетей - *самоорганизующихся карт* Кохонена.

Ассоциация (Associations)

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

Отличие ***ассоциации***: *поиск* закономерностей осуществляется не на основе свойств анализируемого объекта, а между несколькими событиями, которые происходят одновременно.

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

Последовательность (Sequence)

Последовательность позволяет найти временные закономерности между транзакциями.

Задача *последовательности* подобна *ассоциации*, но ее целью является установление закономерностей не между одновременно наступающими событиями, а между событиями, связанными во времени.

Последовательность определяется высокой вероятностью цепочки связанных во времени событий.

А*ссоциация* является частным случаем *последовательности* с временным шагом, равным нулю.

Прогнозирование (Forecasting)

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

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

Определение отклонений или выбросов (Deviation Detection)

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

Оценивание (Estimation)

Задача *оценивания* сводится к предсказанию непрерывных значений признака.

Анализ связей (Link Analysis)

Задача нахождения зависимостей в наборе данных.

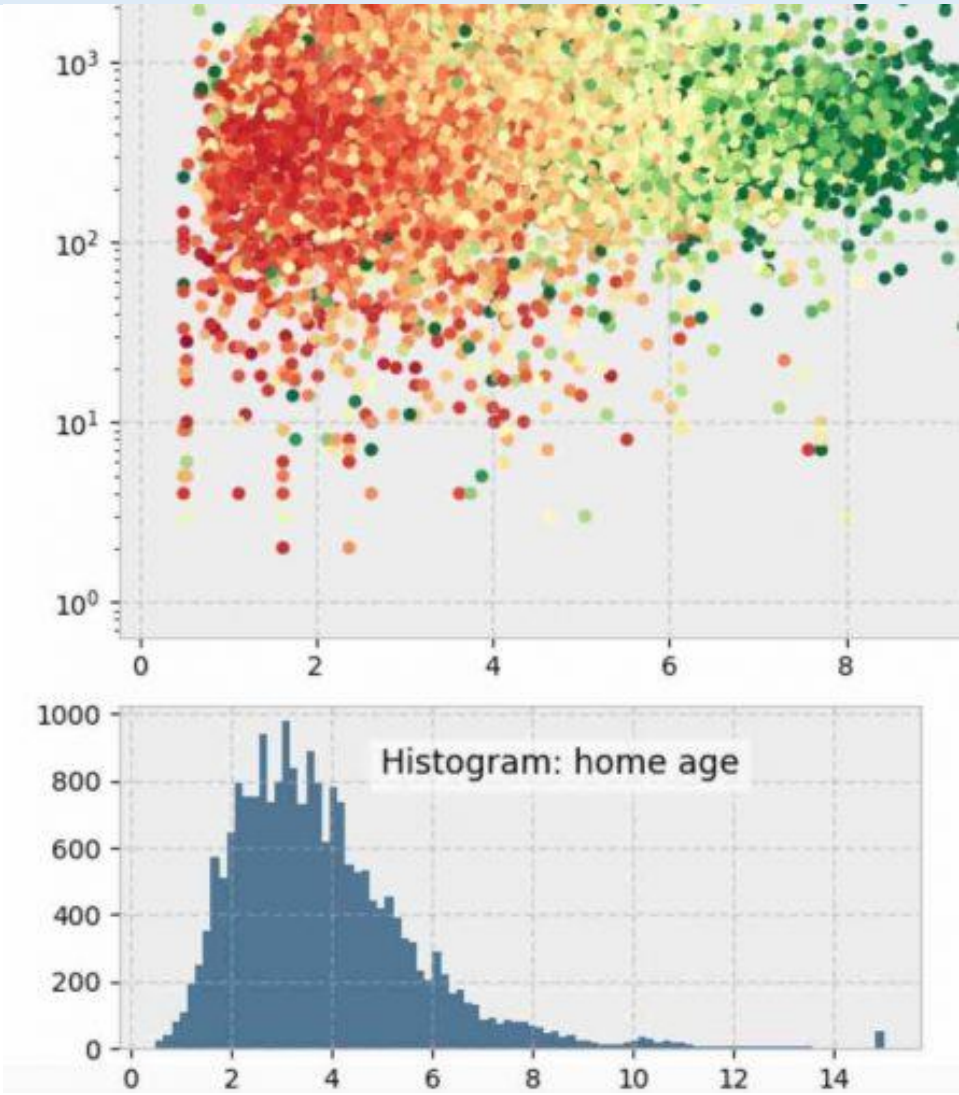
Визуализация (Visualization, Graph Mining)

В результате *визуализации*

создается графический образ анализируемых данных.

Для решения задачи *визуализации* используются графические методы, показывающие наличие закономерностей в данных.

Пример методов *визуализации* *представление* данных в 2D и *3D* измерениях.



Классификация задач Data Mining

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

***обучение с учителем (*Supervised learning - обучение с учителем –** задача анализа данных решается в несколько этапов:

- строится модель анализируемых данных – классификатор;
- классификатор подвергается обучению (проверяется качество его работы, и, если оно неудовлетворительное, происходит дополнительное обучение классификатора)
- продолжается пока не будет достигнут требуемый уровень качества или не станет ясно, что выбранный алгоритм не работает корректно с данными, либо же сами данные не имеют структуры, которую можно выявить***);***

***обучение без учителя (*Unsupervised learning - обучение без учителя –**

объединяет задачи, выявляющие описательные модели. Достоинство таких задач - возможность их решения без каких либо предварительных знаний об анализируемых данных***);***

- другие.

Категория *обучение с учителем*: *классификация*, *регрессия*, *прогнозирование*.

Категория *обучение без учителя* - задача кластеризации.

Описательные и предсказательные задачи

Описательные (descriptive) задачи предназначены для улучшения понимания анализируемых данных.

К такому виду задач относятся кластеризация и поиск ассоциативных правил

Предсказательные (predictive) задачи. Решение разбивается на два этапа:

1) на основании набора данных с известными результатами строится модель;

2) полученная модель используется для предсказания результатов на основании новых наборов данных (требование максимальной точности).

К данному виду задач относят задачи классификации и регрессии, задача поиска ассоциативных правил, если результаты ее решения могут быть использованы для предсказания появления некоторых событий.

Связь понятий

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.

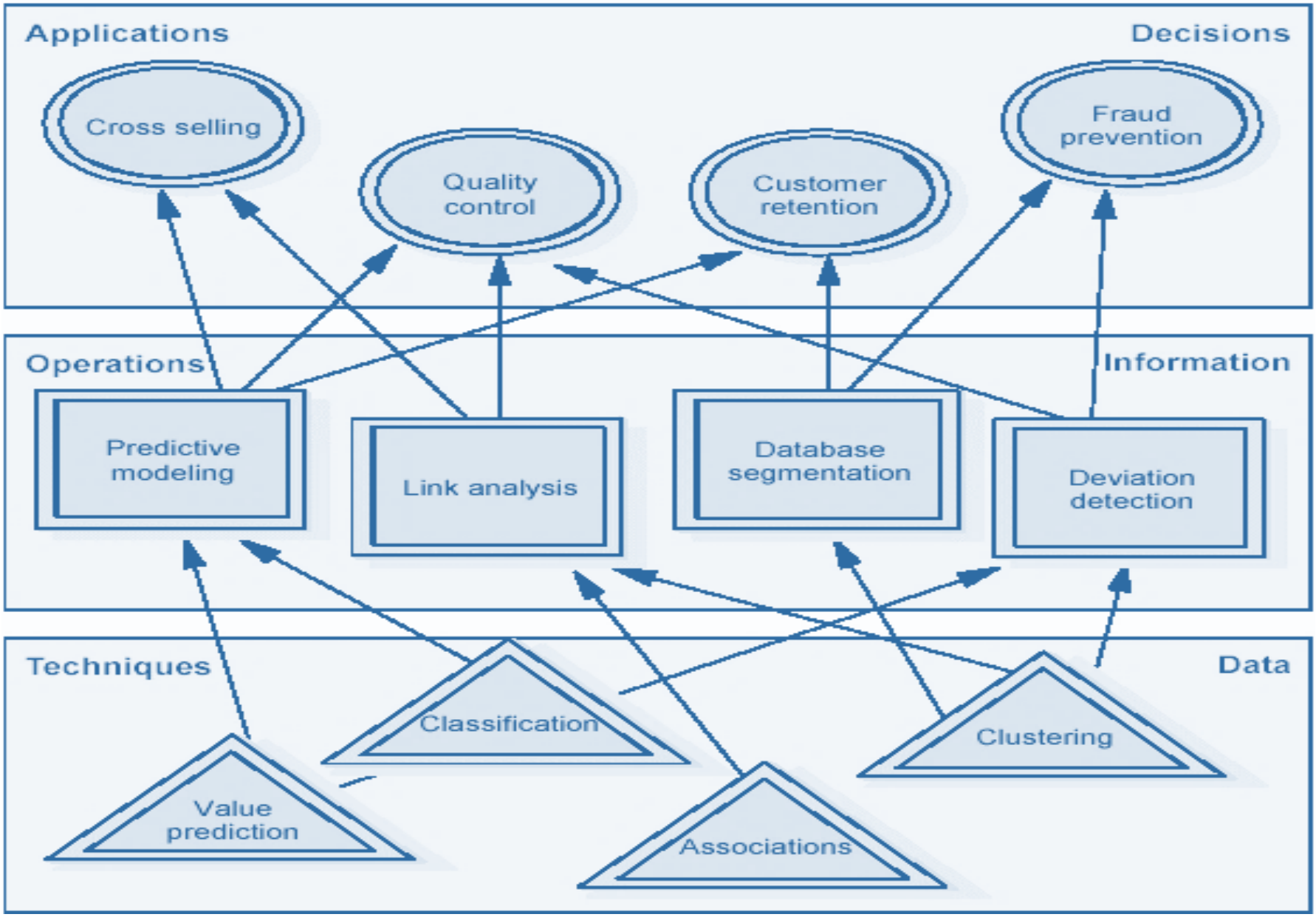
Два потока:

1) ДАННЫЕ → ИНФОРМАЦИЯ → ЗНАНИЯ И РЕШЕНИЯ

2) ЗАДАЧИ → ДЕЙСТВИЯ И МЕТОДЫ РЕШЕНИЯ → ПРИЛОЖЕНИЯ

Эти потоки являются "двумя сторонами одной медали"

ОТ ЗАДАЧИ К ПРИЛОЖЕНИЮ



От задачи к приложению

- **Верхний - уровень приложений** - является уровнем бизнеса, на нем менеджеры принимают решения.

Приведенные примеры приложений: перекрестные продажи, *контроль* качества, удерживание клиентов.

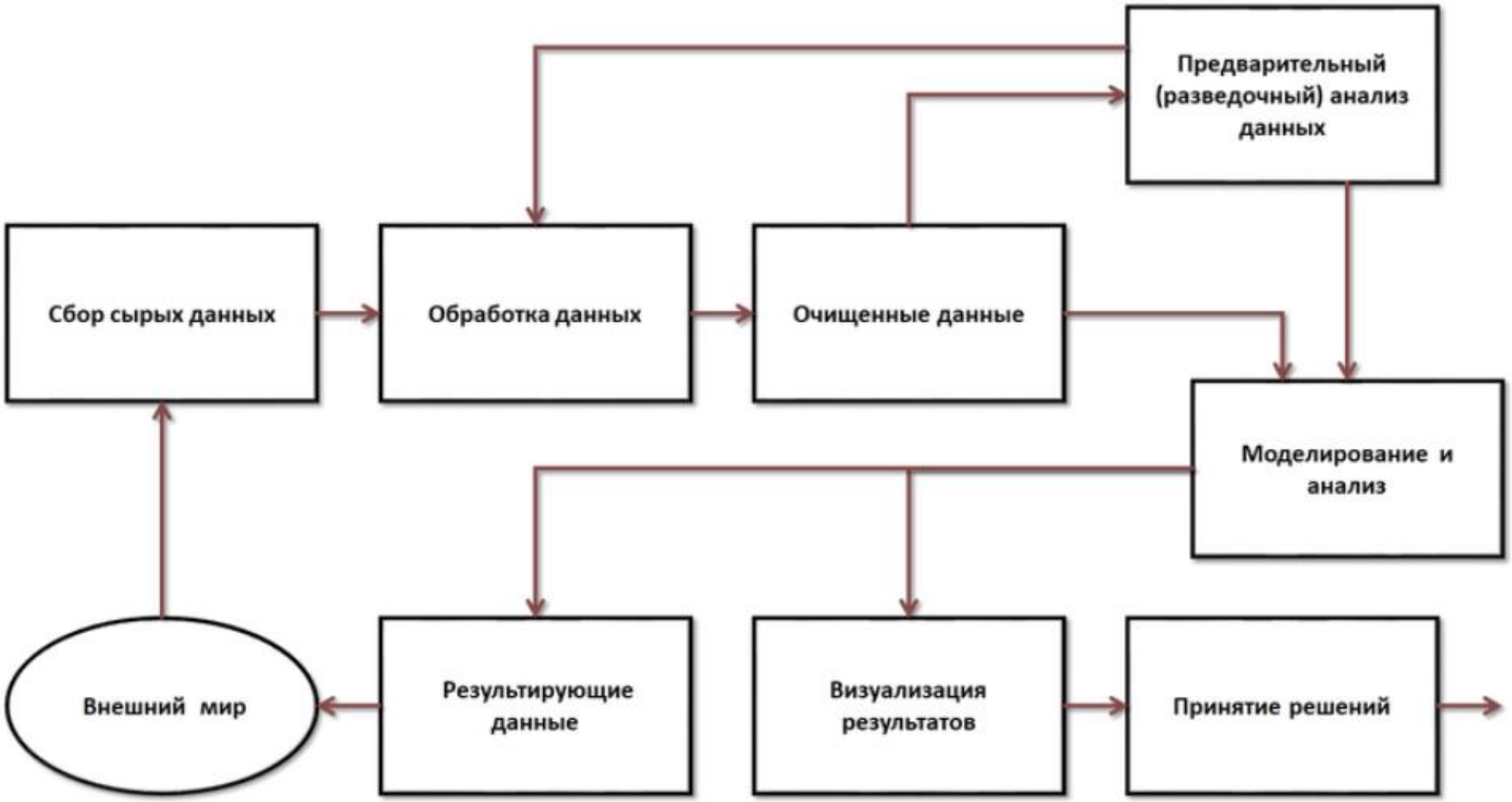
- **Средний - уровень действий** - уровень *информации*, именно на нем выполняются действия *Data* Mining;

На рисунке действия: *прогностическое моделирование*, *анализ связей*, *сегментация* данных и другие.

- **Нижний - уровень определения *задачи*** *Data Mining*, которую необходимо решить применительно к данным, имеющимся в наличии;

Приведены задачи предсказания числовых значений, *классификация*, *кластеризация*, *ассоциация*.

Технологический цикл Data Science



Технологический цикл Data Science

- **Формулировка проблемы**
- **Сбор сырых данных**
- **Data wrangling** — это подготовка сырых данных для выполнения последующей аналитики над ними, преобразование сырых данных, хранящихся в любых произвольных форматах, в требуемые для аналитических приложений.
- **Предварительный анализ данных**, выявление общих тенденций и свойств.
- **Выбор инструментов для глубокого анализа** данных (R, Python, SQL, математические пакеты, библиотеки).
- **Создание модели данных** и проверка ее на соответствие реальным данным.
- **В зависимости от задачи** выполнение статистического анализа, использование машинного обучения или рекурсивного анализа.
- **Сравнение результатов**, полученных разными методами.
- **Визуализация результатов**.
- **Интерпретация данных** и оформление полученной информации для передачи лицам, принимающим решения.

Выводы

- для получения ценных знаний необходимы качественные процедуры обработки.
- Процесс перехода от данных к *знаниям* занимает много времени и стоит дорого.
- Технология *Data* Mining с её мощными и разнообразными алгоритмами является инструментом, при помощи которого, продвигаясь вверх *по информационной пирамиде*, мы можем получать действительно качественные и ценные *знания*.

Основная особенность *Data Mining*

- это **сочетание** широкого математического инструментария (от классического статистического анализа до новых кибернетических *методов*);

- в технологии *Data Mining* гармонично объединились строго формализованные *методы* и *методы* неформального анализа,

т.е. **количественный и качественный *анализ* данных**.

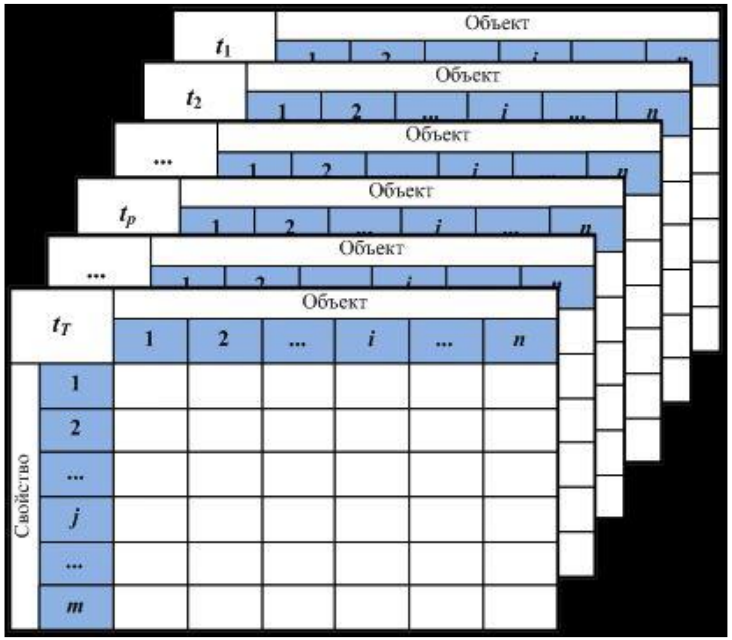
Основная особенность *Data Mining*

- Аналитик имеет дело и с документами, и с табличными значениями, которые также принято называть фактографическими.

- Под единичным фактом принято понимать описание некоторого события. В формализованном виде для этого применяется следующая запись:

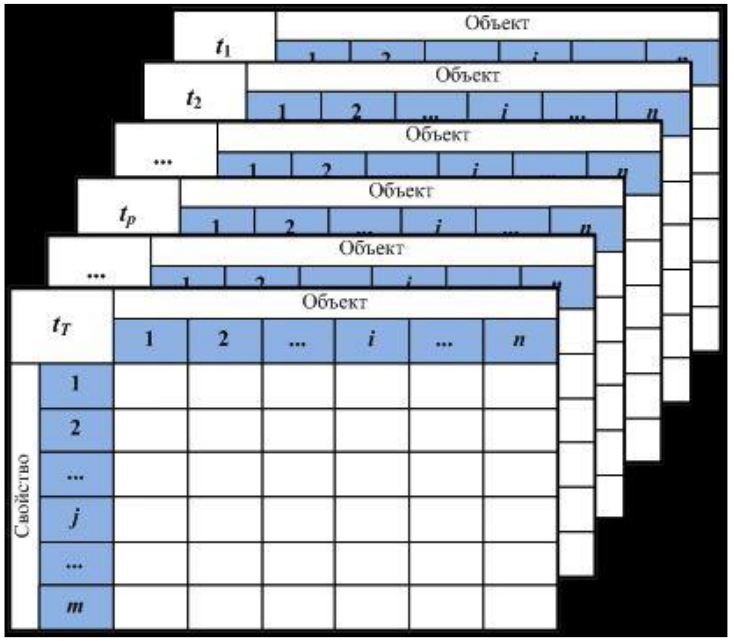
$$E_k = \{a_j, t, x_1, x_2, \ldots, x_m\}$$

- где *aj* – идентификатор (имя) объекта, *t* – время измерения, *xi* – значение ***i***-й характеристики объекта.
- Рассматривая любой документ как множество высказываний можно гомоморфно отобразить его на множество фактов. Иначе говоря, из любого документа можно выделить некоторые факты. Именно они являются исходным сырьем для последующего анализа.



Основная особенность *Data Mining*

- Обычно объекты предварительно упорядочиваются по некоторому признаку, как правило, представляющему
- собой одну из характеристик, которой обладают исследуемые объекты.
- Фактографические данные, т.е. данные, непосредственно относящиеся к заданной предметной области, удобно представлять в табличном виде: строки $a1$, $a2$, …, an отражают информацию о самих исследуемых объектах (различаемых, как правило, по естественному (имени) или условному идентификатору), а столбцы $x1$, $x2$, …, xm – информацию о значениях характеристик этих объектов.
- При необходимости учета временного фактора таких таблиц должно быть несколько: по одной на каждый отсчет времени.



Процесс *Data Mining*

СВОБОДНЫЙ ПОИСК →

ПРОГНОСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ →

→ АНАЛИЗ ИСКЛЮЧЕНИЙ

Стадия 1. Выявление *закономерностей* (***свободный поиск***).
Стадия 2. Использование выявленных *закономерностей* для предсказания неизвестных значений (***прогностическое моделирование***).
Стадия 3. ***Анализ исключений*** - стадия предназначена для выявления и объяснения аномалий, найденных в *закономерностях*.

Классификация методов Data Mining

Технологические методы Data Mining

Статистические методы Data mining

Кибернетические методы Data Mining

Классификация технологических методов Data Mining

Все ***методы Data Mining*** подразделяются на две большие группы **по принципу работы с исходными обучающими данными**.

В этой классификации верхний уровень определяется на основании того, сохраняются ли данные после *Data Mining* либо они дистиллируются для последующего использования.

Технологические методы Data Mining

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* - могут возникнуть сложности анализа сверхбольших баз данных.

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

Технологические методы Data Mining

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.

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

Статистические методы Data mining

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

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

Кибернетические методы Data Mining

- ***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**.

Классификация по задачам Data Mining.

- 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 и другие.

Классификация по задачам Data Mining.

- 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.

Свойства методов Data Mining

Среди основных свойств и характеристик *методов Data Mining* рассматривают следующие:

- **точность**,
- ***масштабируемость***,
- **интерпретируемость**,
- **проверяемость**,
- **трудоемкость**,
- **гибкость**,
- **быстрота** и
- **популярность**.

Выводы

- Каждый из *методов* имеет свои сильные и слабые стороны.
- Но **ни один метод**, какой бы не была его оценка с точки зрения присущих ему характеристик, **не может** обеспечить решение **всего спектра** задач *Data Mining*.
- Большинство инструментов *Data Mining*, **реализуют сразу несколько *методов***, например, деревья решений, индукцию правил и визуализацию, или же нейронные сети, *самоорганизующиеся карты* Кохонена и визуализацию.