Data Mining

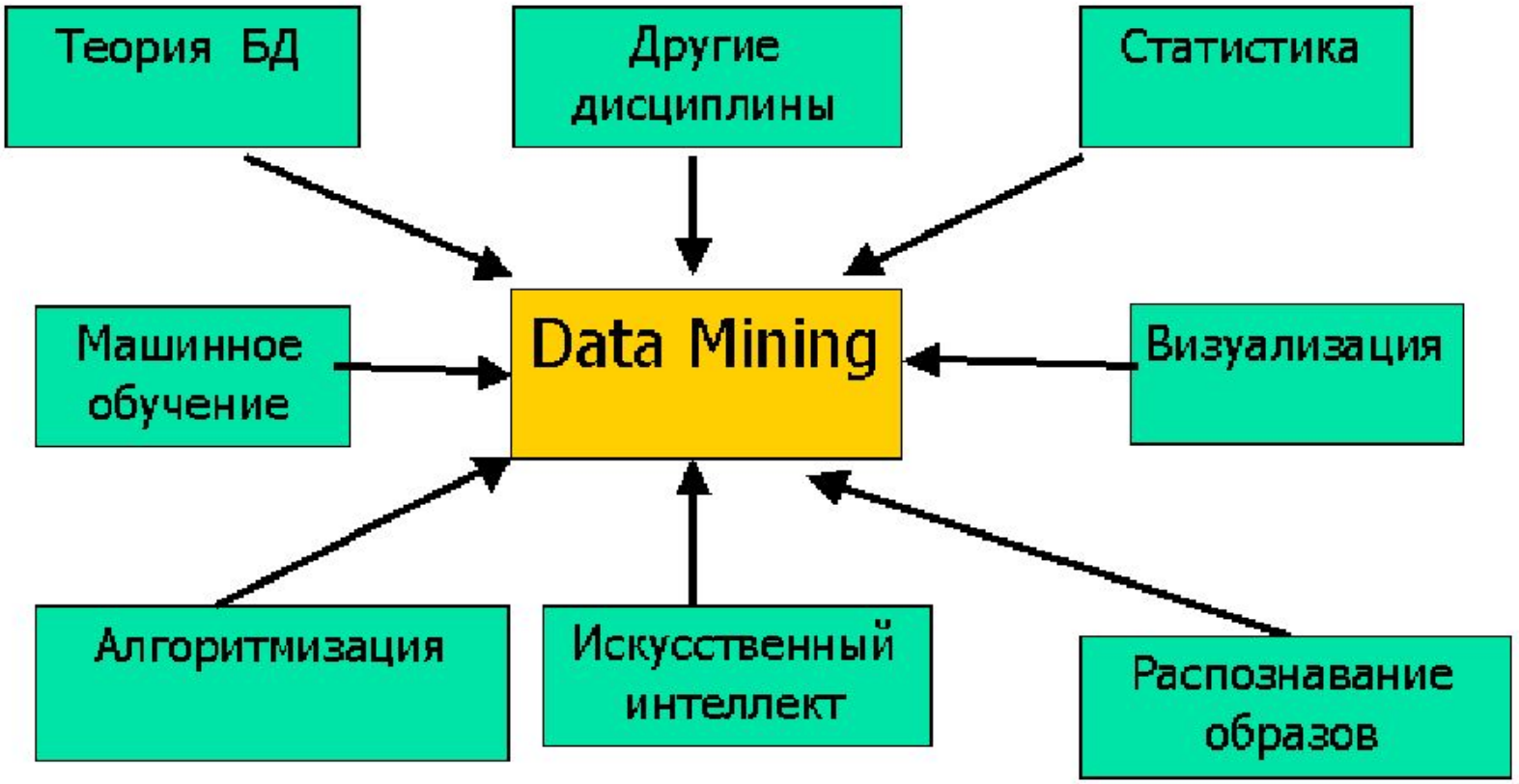
Что такое Data Mining?

- "За последние годы, когда, стремясь к повышению эффективности и прибыльности бизнеса, при создании БД все стали пользоваться средствами обработки цифровой информации, появился и побочный продукт этой активности - горы собранных данных: И вот все больше распространяется идея о том, что эти горы полны золота".
- Сегодня появились новые научные методы и специализированные инструменты, сделавшие горную промышленность намного более точной и производительной. Data Mining для данных развилась почти таким же способом. Старые методы, применявшиеся математиками и статистиками, отнимали много времени, чтобы в результате получить конструктивную и полезную информацию.
- Термин Data Mining получил свое название из двух понятий: поиска ценной информации в большой базе данных (data) и добычи горной руды (mining). Оба процесса требуют или просеивания огромного количества сырого материала, или разумного исследования и поиска искомых ценностей.
- Термин Data Mining часто переводится как добыча данных, извлечение информации, раскопка данных, интеллектуальный анализ данных, средства поиска закономерностей, извлечение знаний, анализ шаблонов, "извлечение зерен знаний из гор данных", раскопка знаний в базах данных, информационная проходка данных, "промывание" данных.

Что такое Data Mining?

- Data Mining - мультидисциплинарная область, возникшая и развивающаяся на базе таких наук как прикладная статистика, распознавание образов, искусственный интеллект, теория баз данных и др.,

Data Mining как мультидисциплинарная область



Понятие Статистики

- **Статистика** - это наука о методах сбора данных, их обработки и анализа для выявления закономерностей, присущих изучаемому явлению.
- **Статистика** является совокупностью методов планирования эксперимента, сбора данных, их представления и обобщения, а также анализа и получения выводов на основании этих данных.

Понятие Машинного обучения

- Единого определения машинного обучения на сегодняшний день нет.
- **Машинное обучение** можно охарактеризовать как процесс получения программой новых знаний. в 1996
- "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.

Понятие Data Mining

- 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.

- **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.

- Data Mining - это процесс выделения из данных неявной и неструктурированной информации и представления ее в виде, пригодном для использования.
- Data Mining - это процесс выделения, исследования и моделирования больших объемов данных для обнаружения неизвестных до этого структур (patterns) с целью достижения преимуществ в бизнесе (определение SAS Institute).
- Data Mining - это процесс, цель которого - обнаружить новые значимые корреляции, образцы и тенденции в результате просеивания большого объема хранимых данных с использованием методик распознавания образцов плюс применение статистических и математических методов (определение Gartner Group).

- Прежде чем использовать технологию Data Mining, необходимо тщательно проанализировать ее проблемы, ограничения и критические вопросы, с ней связанные, а также понять, чего эта технология не может.
- Data Mining не может заменить аналитика
- Технология не может дать ответы на те вопросы, которые не были заданы. Она не может заменить аналитика, а всего лишь дает ему мощный инструмент для облегчения и улучшения его работы.
- Сложность разработки и эксплуатации приложения Data Mining

  Поскольку данная технология является мультидисциплинарной областью, для разработки приложения, включающего Data Mining, необходимо задействовать специалистов из разных областей, а также обеспечит их качественное взаимодействие.

Отличия Data Mining от других методов анализа данных

- Традиционные методы анализа данных (статистические методы) и OLAP в основном ориентированы на проверку заранее сформулированных гипотез (verification-driven data mining) 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 такие 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 будет 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.

Данные

- **Что такое данные?**

  В широком понимании данные представляют собой факты, текст, графики, картинки, звуки, аналоговые или цифровые видео-сегменты.
- Данные могут быть получены в результате измерений, экспериментов, арифметических и логических операций.
- Данные должны быть представлены в форме, пригодной для хранения, передачи и обработки.
- Иными словами, данные - это необработанный материал, предоставляемый поставщиками данных и используемый потребителями для формирования информации на основе данных.

Набор данных и их атрибутов

- В таблице представлена двухмерная таблица, представляющая собой набор данных.

Код	Возраст	Семейное Положение	Доход	Класс
1	18	Single	125	1
2	22	Married	100	1
3	30	Single	70	1
4	32	Married	120	1
5	24	Divorced	95	2
6	25	Married	60	1
7	32	Divorced	220	1
8	19	Single	85	2
9	22	Married	75	1
10	40	Single	90	2

- По горизонтали таблицы располагаются атрибуты объекта или его признаки. По вертикали таблицы - объекты.
- Объект описывается как набор атрибутов.
- Объект также известен как запись, случай, пример, строка таблицы и т.д.
- Атрибут - свойство, характеризующее объект.
- Например: цвет глаз человека, температура воды и т.д.
- Атрибут также называют переменной, полем таблицы, измерением, характеристикой.

Базы данных. Основные положения

- **База данных (Database) - это особым образом организованные и хранимые в электронном виде данные.**
- **Схема данных - описание логической структуры данных, специфицированное на языке описания данных и обрабатываемое СУБД.**
- **СУБД (Database Management System, DBMS) представляет собой оболочку, с помощью которой при организации структуры таблиц и заполнения их данными получается та или иная база данных.**
- **Метаданные (Metadata)** - это данные о данных.

Методы и стадии Data Mining

- Основная особенность Data Mining - это сочетание широкого математического инструментария (от классического статистического анализа до новых кибернетических методов) и последних достижений в сфере информационных технологий. В технологии Data Mining гармонично объединились строго формализованные методы и методы неформального анализа, т.е. количественный и качественный анализ данных.

К методам и алгоритмам Data Mining относятся следующие:

- искусственные нейронные сети,
- деревья решений,
- символьные правила,
- методы ближайшего соседа и k-ближайшего соседа,
- линейная регрессия, корреляционно- регрессионный анализ; иерархические методы кластерного анализа,
- неиерархические методы кластерного анализа
- методы поиска ассоциативных правил, в том числе алгоритм Apriori;
- метод ограниченного перебора,
- эволюционное программирование и генетические алгоритмы,
- разнообразные методы визуализации данных и множество других методов.

Классификация стадий Data Mining

Data Mining может состоять из двух или трех стадий

- **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, следующую за

  стадией свободного поиска. Цель валидации - проверка достоверности найденных закономерностей.
- **Стадия 3**. Анализ исключений - стадия предназначена для выявления и объяснения аномалий, найденных в закономерностях.

Свободный поиск (Discovery)

Свободный поиск представлен такими **действиями**:

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

Могут быть найдены, например, такие закономерности

- "**Если** возраст < 20 лет и желаемый уровень вознаграждения > 700 условных единиц, то в 75% случаев соискатель ищет работу программиста"

- "**Если** возраст >35 лет и желаемый уровень вознаграждения > 1200 условных единиц, то в 90% случаев соискатель ищет руководящую работу".

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

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

2. Прогностическое моделирование (Predictive Modeling)

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).

В процессе прогностического моделирования решаются задачи классификации и прогнозирования.

Продолжая рассмотренный пример первой стадии, можем сделать следующий вывод.

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

Сравнение свободного поиска и прогностического моделирования с точки зрения логики

- Свободный поиск раскрывает общие закономерности. Он по своей природе индуктивен. Закономерности, полученные на этой стадии, формируются от частного к общему. В результате мы получаем некоторое общее знание о некотором классе объектов на основании исследования отдельных представителей этого класса.

- Прогностическое моделирование, напротив, дедуктивно. Закономерности, полученные на этой стадии, формируются от общего к частному и единичному. Здесь мы получаем новое знание о некотором объекте или же группе объектов на основании:

  · знания класса, к которому принадлежат исследуемые объекты;

  · знание общего правила, действующего в пределах данного класса объектов.

3. Анализ исключений (forensic analysis)

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.

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

  1, существует некоторое логическое объяснение, которое также может быть оформлено в виде правила.

  2, для оставшихся 10% - это ошибки исходных данных. В этом случае стадия анализа исключений может быть использована в качестве очистки данных

Классификация методов Data Mining

В 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.

Статистические методы Data mining

- 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.

Задачи Data Mining

Классификация (Classification)

- Краткое описание. Наиболее простая и распространенная задача Data Mining. В результате решения задачи классификации обнаруживаются признаки, которые характеризуют группы объектов исследуемого набора данных - классы; по этим признакам новый объект можно отнести к тому или иному классу.
- Методы решения. Для решения задачи классификации могут использоваться методы:
  - ближайшего соседа (Nearest Neighbor);
  - k-ближайшего соседа (k-Nearest Neighbor);
  - байесовские сети (Bayesian Networks);
  - индукция деревьев решений;
  - нейронные сети (neural networks).

Задачи Data Mining

Кластеризация (Clustering)

- 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.

Задачи Data Mining

Ассоциация (Associations)

- Краткое описание. В ходе решения задачи поиска ассоциативных правил отыскиваются закономерности между связанными событиями в наборе данных.
- Отличие ассоциации от двух предыдущих задач Data Mining: поиск закономерностей осуществляется не на основе свойств анализируемого объекта, а между несколькими событиями, которые происходят одновременно.
- Наиболее известный алгоритм решения задачи поиска ассоциативных правил – алгоритм Apriori.

Задачи Data Mining

Последовательность (Sequence), или последовательная ассоциация (sequential association)

- Краткое описание. Последовательность позволяет найти временные закономерности между транзакциями. Задача последовательности подобна ассоциации, но ее целью является установление закономерностей не между одновременно наступающими событиями, а между событиями, связанными во времени (т.е. происходящими с некоторым определенным интервалом во времени). Фактически, ассоциация является частным случаем последовательности с временным лагом, равным нулю.
- Правило последовательности: после события X через определенное время произойдет событие Y.
- Пример. После покупки квартиры жильцы в 60% случаев в течение двух недель приобретают холодильник, а в течение двух месяцев в 50% случаев приобретается телевизор. Решение данной задачи широко применяется в маркетинге и менеджменте, например, при управлении циклом работы с клиентом (Customer Lifecycle Management).

Задачи Data Mining

Прогнозирование (Forecasting)

- Краткое описание. В результате решения задачи прогнозирования на основе особенностей исторических данных оцениваются пропущенные или же будущие значения целевых численных показателей.
- Для решения таких задач широко применяются методы математической статистики, нейронные сети и др.

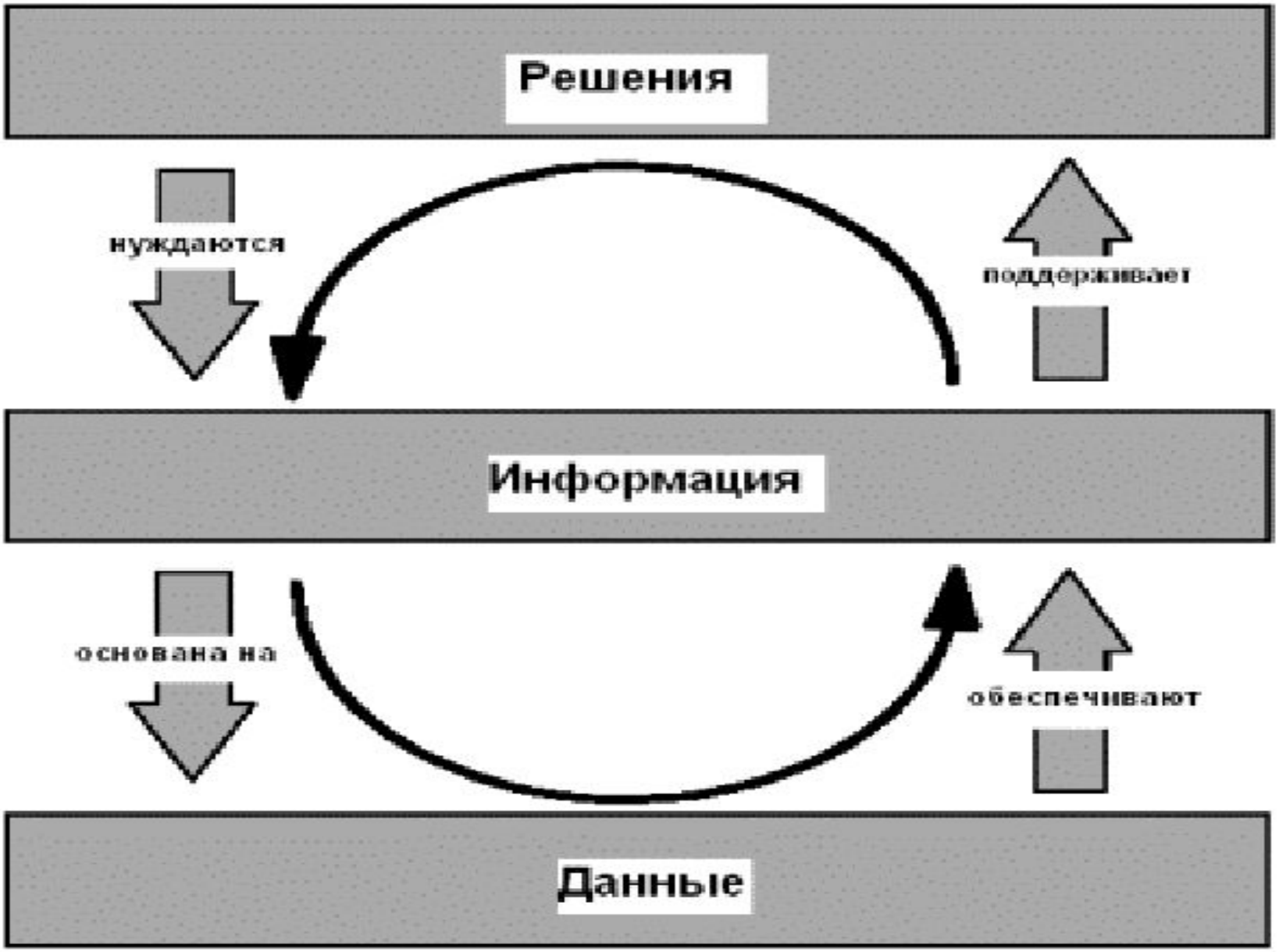
Задачи Data Mining

- Определение отклонений или выбросов (Deviation Detection), анализ отклонений или Выбросов
- Краткое описание. Цель решения данной задачи - обнаружение и анализ данных, наиболее отличающихся от общего множества данных, выявление так называемых нехарактерных шаблонов.

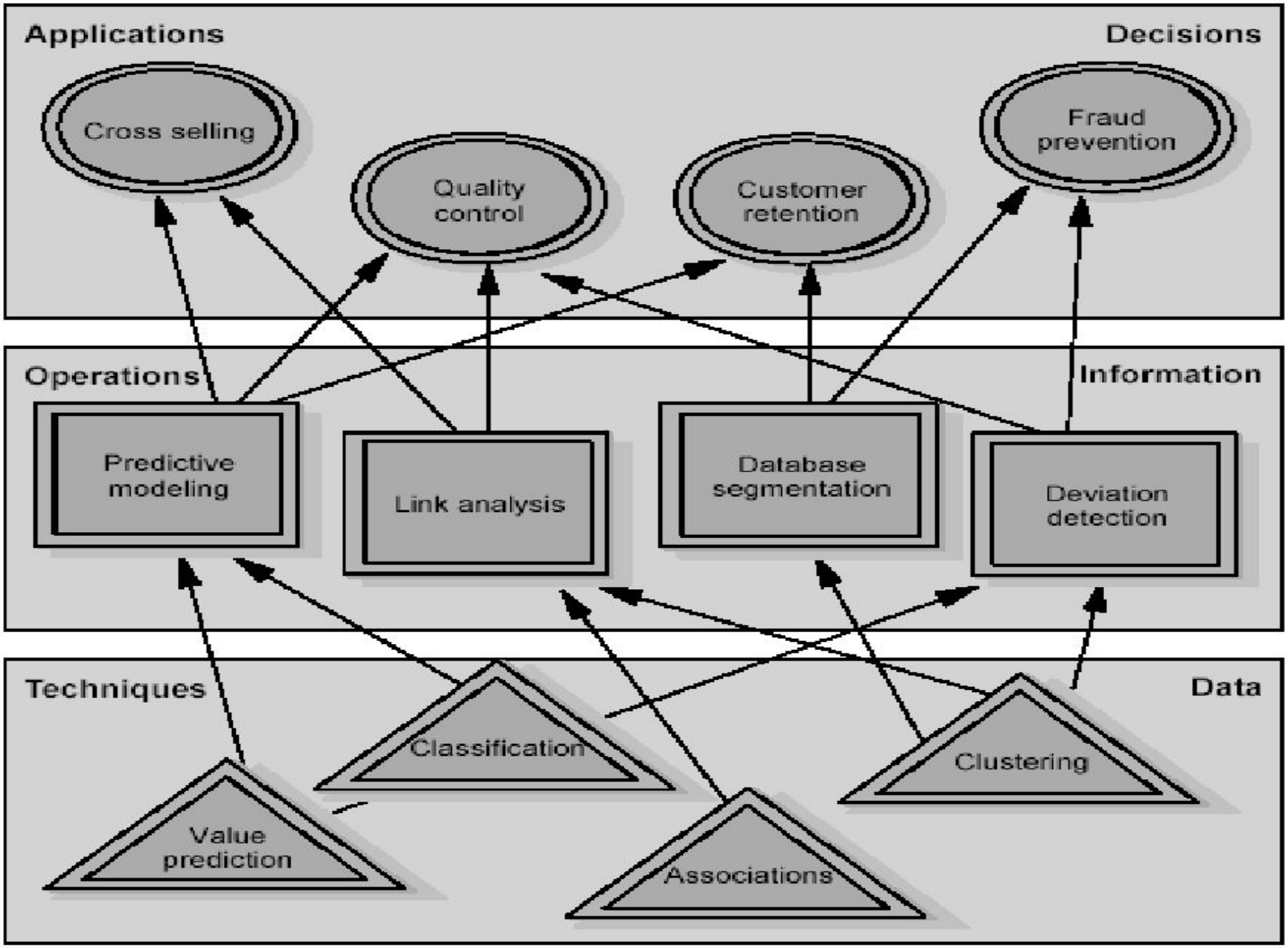
Задачи Data Mining

- Оценивание (Estimation). Задача оценивания сводится к предсказанию непрерывных значений признака.
- Анализ связей (Link Analysis) - задача нахождения зависимостей в наборе данных.
- Визуализация (Visualization, Graph Mining) Пример методов визуализации - представление данных в 2-D и 3-D измерениях.

Связь понятий



Задачи, действия, приложения



- Верхний - уровень приложений - является уровнем бизнеса (если мы имеем дело с задачей бизнеса), на нем менеджеры принимают решения. Приведенные примеры приложений: перекрестные продажи, контроль качества, удерживание клиентов.
- Средний - уровень действий - по своей сути является уровнем информации, именно на нем выполняются действия Data Mining; на рисунке приведены такие действия: прогностическое моделирование (было рассмотрено в предыдущей лекции), анализ связей, сегментация данных и другие.
- Нижний - уровень определения задачи Data Mining, которую необходимо решить применительно к данным, имеющимся в наличии; на рис, приведены задачи предсказания числовых значений, классификация, кластеризация, ассоциация.

- Рассмотрим задачу удержания клиентов (определения надежности клиентов фирмы).
- **Первый уровень**. Данные - база данных по клиентам. Есть данные о клиенте (возраст, пол, профессия, доход). Определенная часть клиентов, воспользовавшись продуктом фирмы, осталась ей верна; другие клиенты больше не приобретали продукты фирмы. На этом уровне мы определяем тип задачи - это задача классификации.
- На **втором уровне** определяем действие – прогностическое моделирование. С помощью прогностического моделирования мы с определенной долей уверенности можем отнести новый объект, в данном случае, нового клиента, к одному из известных классов - постоянный клиент, или это, скорее всего, его разовая покупка.
- На **третьем уровне** мы можем воспользоваться приложением для принятия решения. В результате приобретения знаний, фирма может существенно снизить расходы, например, на рекламу, зная заранее, каким из клиентов следует активно рассылать рекламные материалы.

Информация. *Свойства информации*

- · Полнота информации.
- Это свойство характеризует качество информации и определяет достаточность данных для принятия решений, т.е. информация должна содержать весь необходимый набор данных.
- · Достоверность информации.
- Информация может быть достоверной и недостоверной. В недостоверной информации присутствует информационный шум, и чем он выше, тем ниже достоверность информации.
- · Ценность информации.
- Ценность информации не может быть абстрактной. Информация должна быть полезной и ценной для определенной категории пользователей.
- · Адекватность информации.
- · Актуальность информации.
- Информация должна быть актуальной, т.е. не устаревшей. Это свойство информации характеризует степень соответствия информации настоящему моменту времени.
- · Ясность информации.
- Информация должна быть понятна тому кругу лиц, для которого она предназначена.
- · Доступность информации.
- Доступность характеризует меру возможности получить определенную информацию. На это свойство информации влияют одновременно доступность данных и доступность адекватных методов.
- · Субъективность информации.
- Информация носит субъективный характер, она определяется степенью восприятия субъекта (получателя информации).

Знания

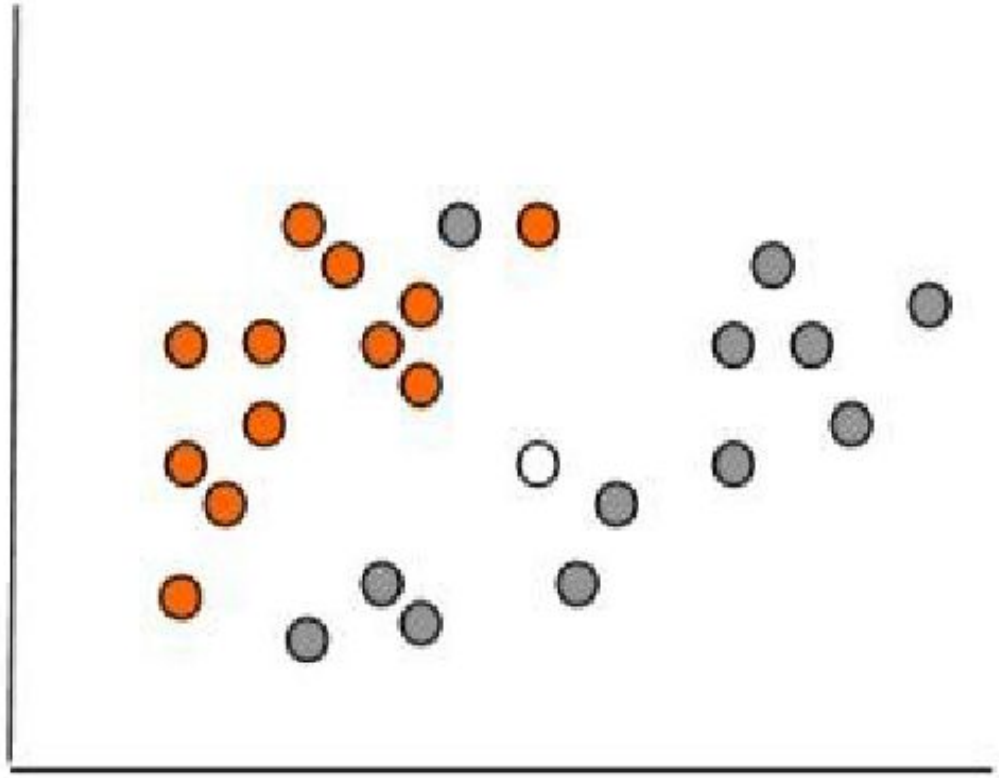
- Знания - совокупность фактов, закономерностей и эвристических правил, с помощью которых решается поставленная задача.
- Знания имеют определенные свойства, которые отличают их от информации .
- 1. **Структурированность**. Знания должны быть "разложены по полочкам".
- 2. **Удобство доступа и усвоения**. 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.

Задачи Data Mining. Классификация и кластеризация

- **Задача классификации**
- Классификация - упорядоченное по некоторому принципу множество объектов, которые имеют сходные классификационные признаки (одно или несколько свойств), выбранных для определения сходства или различия между этими объектами.
- Классификация требует соблюдения следующих правил:
- · в каждом акте деления необходимо применять только одно основание;
- · деление должно быть соразмерным, т.е. общий объем видовых понятий должен равняться объему делимого родового понятия;
- · члены деления должны взаимно исключать друг друга, их объемы не должны перекрещиваться;
- · деление должно быть последовательным.

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

- Рассмотрим задачу классификации на простом примере. Допустим, имеется база данных о клиентах туристического агентства с информацией о возрасте и доходе за месяц. Есть рекламный материал двух видов: более дорогой и комфортный отдых и более дешевый, молодежный отдых. Соответственно, определены два класса клиентов: класс 1 и класс 2.
- **Задача.** Определить, к какому классу принадлежит новый клиент и какой из двух видов рекламных материалов ему стоит отсылать.



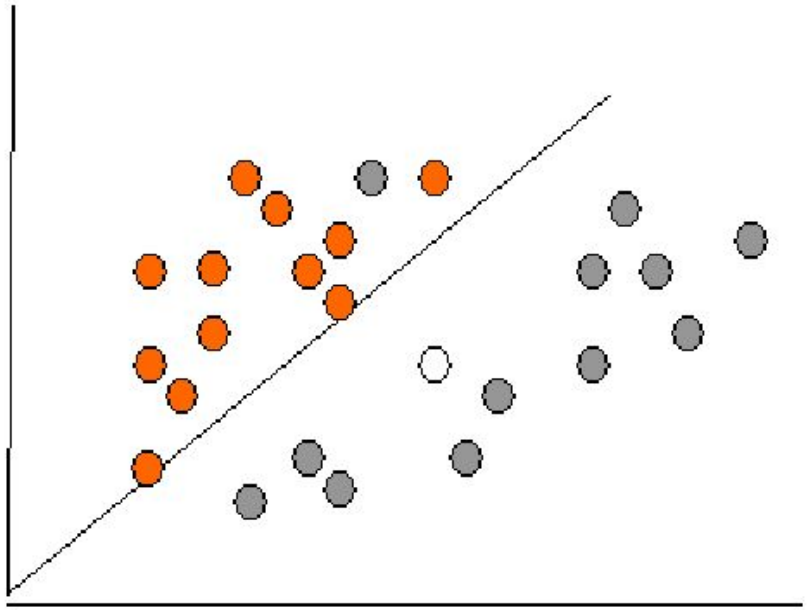
Процесс классификации состоит из двух этапов: конструирования модели и ее использования.

1. Конструирование модели: описание множества предопределенных классов.
   - о Каждый пример набора данных относится к одному предопределенному классу.
   - о На этом этапе используется обучающее множество, на нем происходит конструирование модели.
   - о Полученная модель представлена классификационными правилами, деревом решений или математической формулой.
2. Использование модели: классификация новых или неизвестных значений.
   - о Оценка правильности (точности) модели.
     1. Известные значения из тестового примера сравниваются с результатами использования полученной модели.
     2. Уровень точности - процент правильно классифицированных примеров в тестовом множестве.
     3. Тестовое множество, т.е. множество, на котором тестируется построенная модель, не должно зависеть от обучающего множества.
   - о Если точность модели допустима, возможно использование модели для классификации новых примеров, класс которых неизвестен.

Методы, применяемые для решения задач классификации

Для классификации используются различные методы.
Основные из них:

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



Решение задачи классификации методом линейной регрессии

Решение задачи классификации методом нейронных сетей

Задача кластеризации

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

Цель кластеризации - поиск существующих структур.

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

Кластер можно охарактеризовать как группу объектов, имеющих общие свойства.

Характеристиками кластера можно назвать два признака:

· внутренняя однородность;

· внешняя изолированность.

Задачи Data Mining. Прогнозирование и визуализация

- **Задача прогнозирования**

Прогнозирование направлено на определение тенденций динамики конкретного объекта или события на основе ретроспективных данных, т.е. анализа его состояния в прошлом и настоящем.

Таким образом, решение задачи прогнозирования требует некоторой обучающей выборки данных.

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

· выбор модели прогнозирования;

· анализ адекватности и точности построенного прогноза.

Точность прогноза

Точность прогноза характеризуется ошибкой прогноза.

Наиболее распространенные виды ошибок:

- **Средняя ошибка (СО).** Она вычисляется простым усреднением ошибок на каждом шаге.

Недостаток этого вида ошибки - положительные и отрицательные ошибки аннулируют друг друга.

- **Средняя абсолютная ошибка (САО)**. Она рассчитывается как среднее абсолютных ошибок. Если она равна нулю, то мы имеем совершенный прогноз. В сравнении со средней квадратической ошибкой, эта мера "не придает слишком большого значения" выбросам.
- **Сумма квадратов ошибок (SSE)**, среднеквадратическая ошибка. Она вычисляется как сумма (или среднее) квадратов ошибок. Это наиболее часто используемая оценка точности прогноза.

- **Относительная ошибка (ОО).** Предыдущие меры использовали действительные значения ошибок. Относительная ошибка выражает качество подгонки в терминах относительных ошибок.

Задача визуализации

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

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

Кратко роль визуализации можно описать такими ее возможностями:

· поддержка интерактивного и согласованного исследования;

· помощь в представлении результатов;

· использование глаз (зрения), чтобы создавать зрительные образы и осмысливать их.

Методы классификации и прогнозирования. Деревья решений

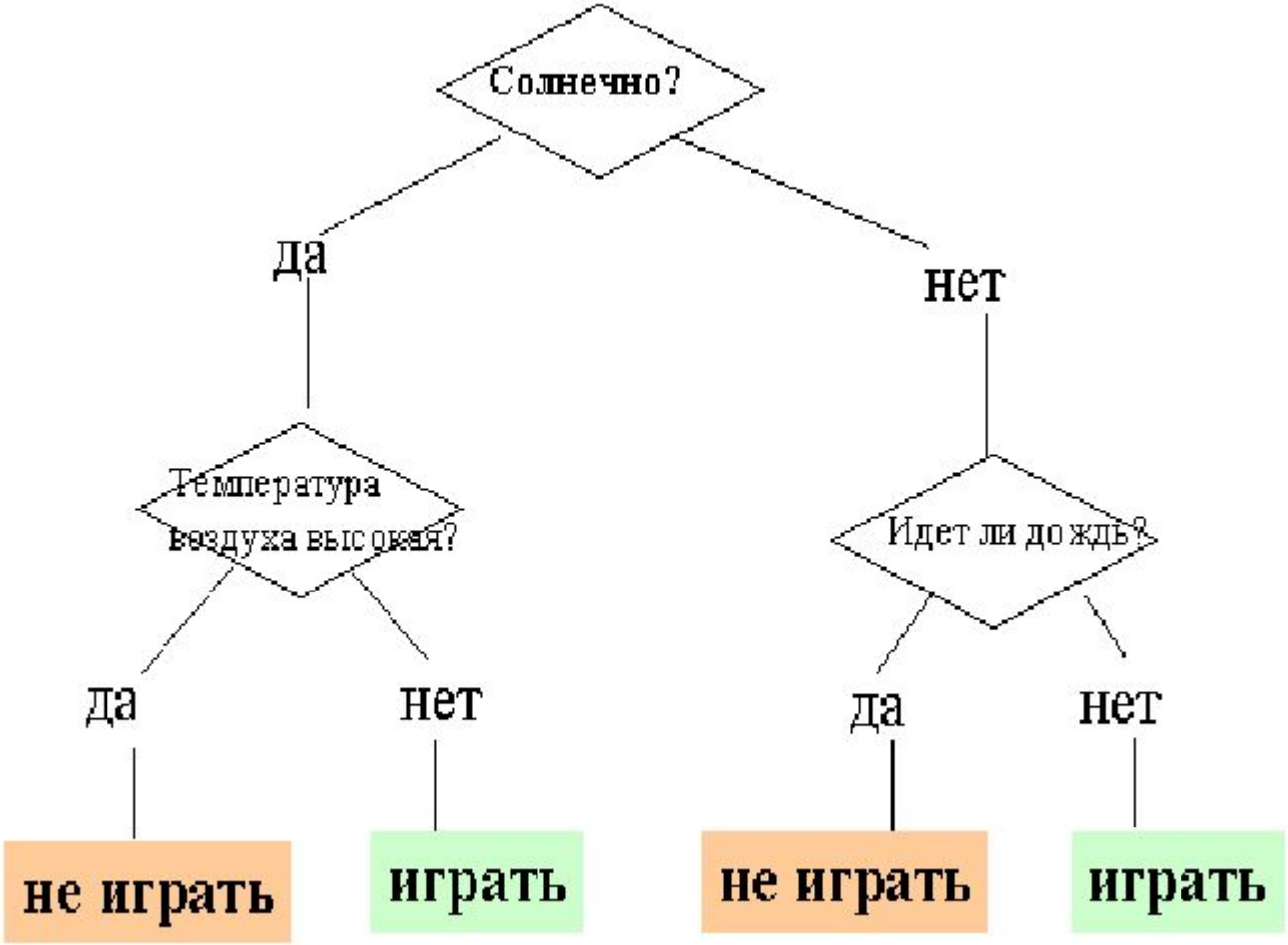
Метод деревьев решений (decision trees) является одним из наиболее популярных методов решения задач классификации и прогнозирования.

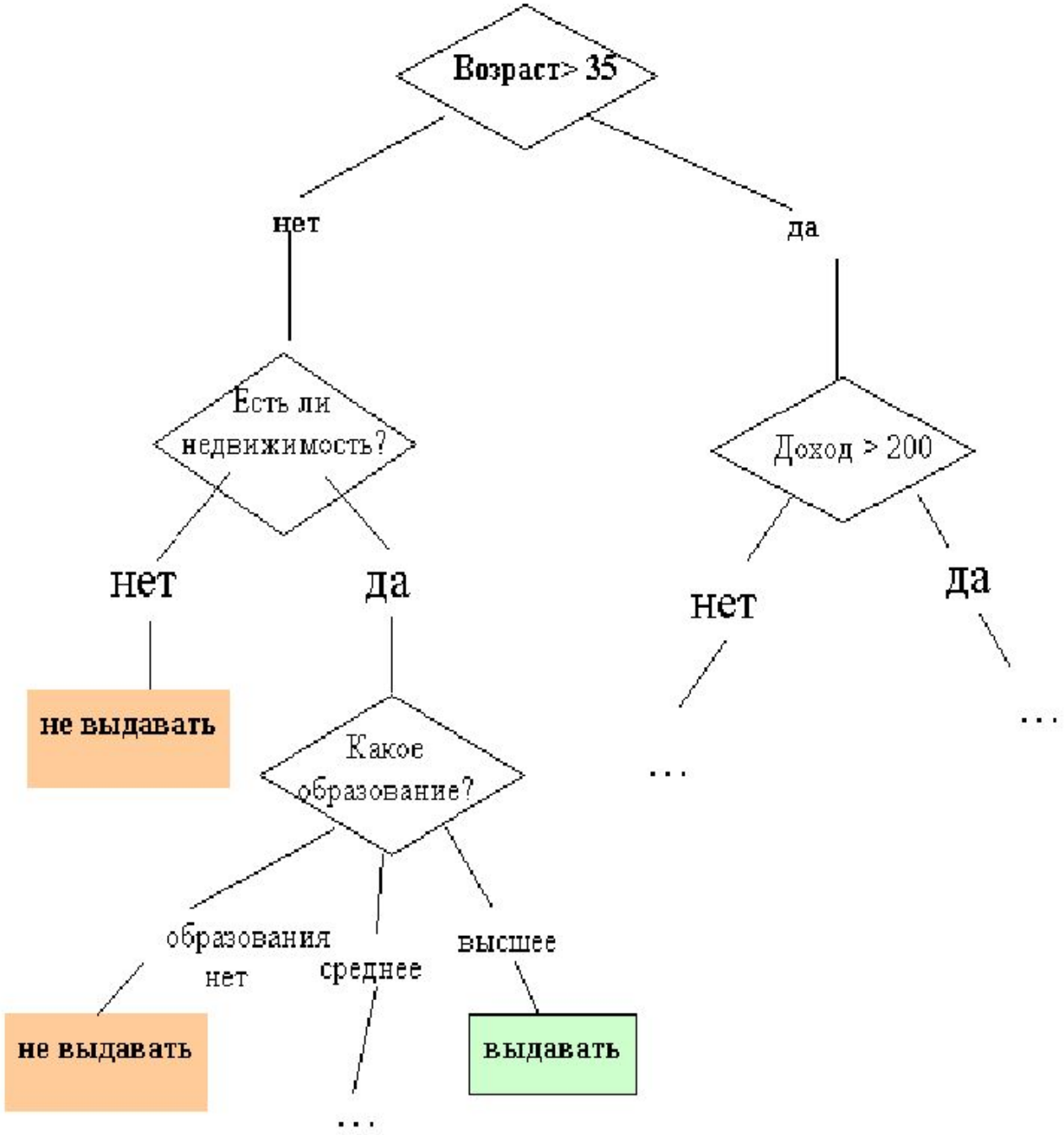
Иногда этот метод Data Mining также

называют деревьями решающих правил, деревьями классификации и регрессии.

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

Играть ли в гольф?





задача "Выдавать ли кредит клиенту?".

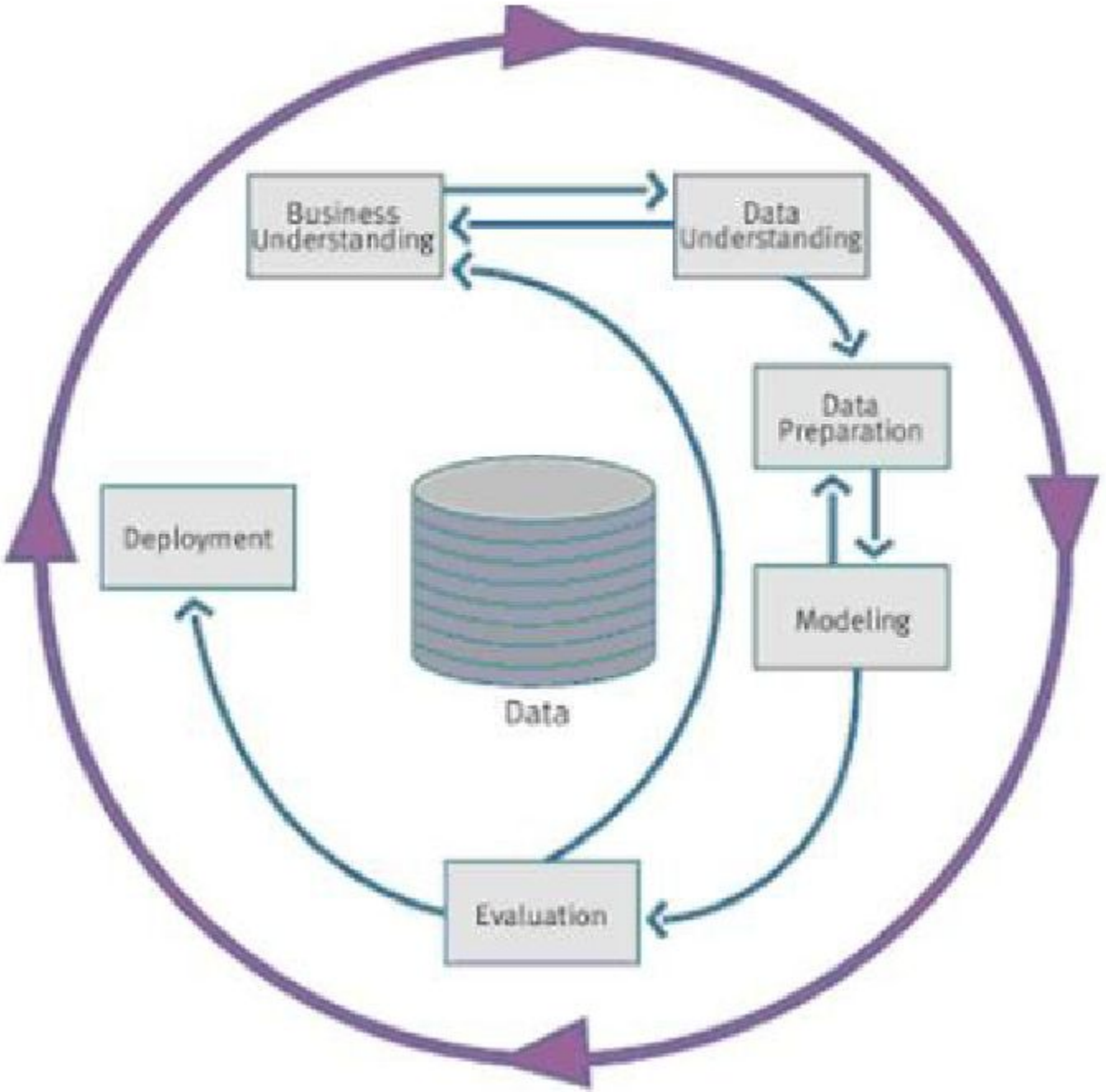
CRISP-DM методология

CRISP-DM [100] (The Cross Industrie Standard Process for Data Mining – Стандартный межотраслевой процесс Data Mining) является наиболее популярной и распространенной методологией. Членами консорциума CRISP-DM являются NCR, SPSS и DaimlerChrysler.

В соответствии со стандартом CRISP, **Data Mining является непрерывным процессом со многими циклами и обратными связями**.

Data Mining по стандарту CRISP-DM включает следующие фазы:

1. Осмысление бизнеса (Business understanding).
2. Осмысление данных (Data understanding).
3. Подготовка данных (Data preparation).
4. Моделирование (Modeling).
5. Оценка результатов (Evaluation).
6. Внедрение (Deployment).

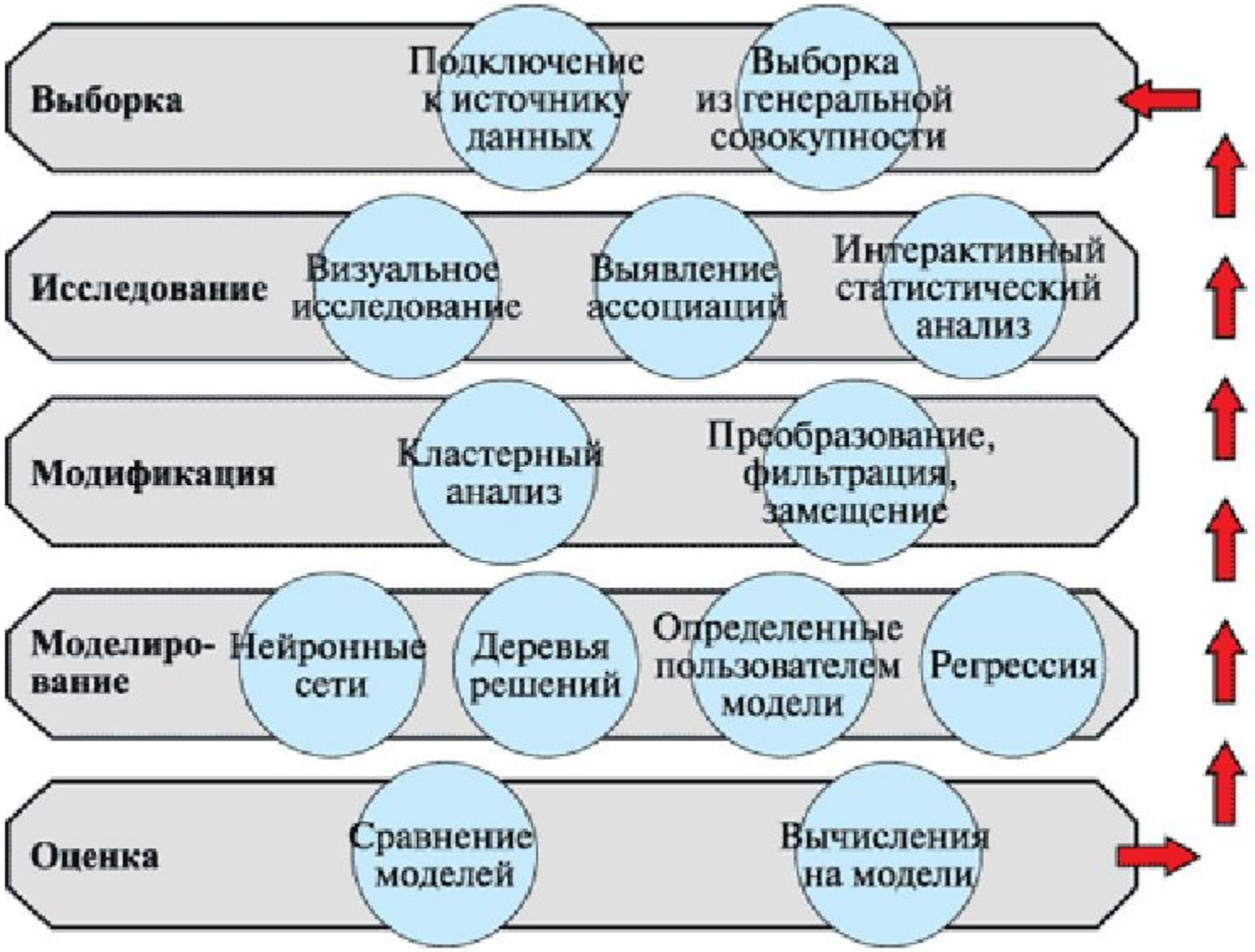

Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

SEMMA методология

SEMMA методология реализована в среде SAS Data Mining Solution (SAS) [102].

Ее аббревиатура образована от слов Sample ("Отбор данных", т.е. создание выборки), Explore ("Исследование отношений в данных"), Modify ("Модификация данных"), Model ("Моделирование взаимозависимостей"), Assess ("Оценка полученных моделей и результатов").

Методология разработки проекта Data Mining в соответствии с методологией SEMMA изображена на рис.


Выборка
Подключение к источнику данных
Выборка из генеральной совокупности
Исследование
Визуальное исследование
Выявление ассоциаций
Интерактивный статистический анализ
Модификация
Кластерный анализ
Преобразование, фильтрация, замещение
Моделиро-вание
Нейронные сети
Деревья решений
Определенные пользователем модели
Регрессия
Оценка
Сравнение моделей
Вычисления на модели