Анализ бизнес информации – основные принципы

Последовательность работы

Способы анализа данных

Главным лицом в процессе анализа данных является **эксперт** – специалист в предметной области.

Несмотря на то, что существует большое количество аналитических задач, методы их решения можно поделить на 2 категории:

- Извлечение и визуализация данных
- Построение и использование моделей

Общая схема анализа

Визуализация данных

В случае визуализации эксперт формулирует некоторым образом запрос к системе, извлекает нужную информацию из различных источников и просматривает полученные результаты.

На основе имеющихся сведений он делает выводы, которые и являются результатом анализа. Существует множество способов визуализации данных:

- OLAP (кросс-таблицы и кросс-диаграммы)
- Таблицы, диаграммы, гистограммы
- Карты, проекции, срезы и прочие

Достоинства и недостатки визуализации

Достоинства:

- Простота создания
- Работа на данных малого объема и низкого качества
- Возможность использования экспертных знаний

Недостатки:

- Неспособность обрабатывать большие объемы
- Неспособность анализа сложных закономерностей
- Сильная зависимость от конкретного эксперта
- Отсутствие возможности тиражирования

Построение моделей

Построение моделей является универсальным способом изучения окружающего мира. Этот способ позволяет обнаруживать зависимости, прогнозировать, разбивать на группы и решать множество других интеллектуальных задач.

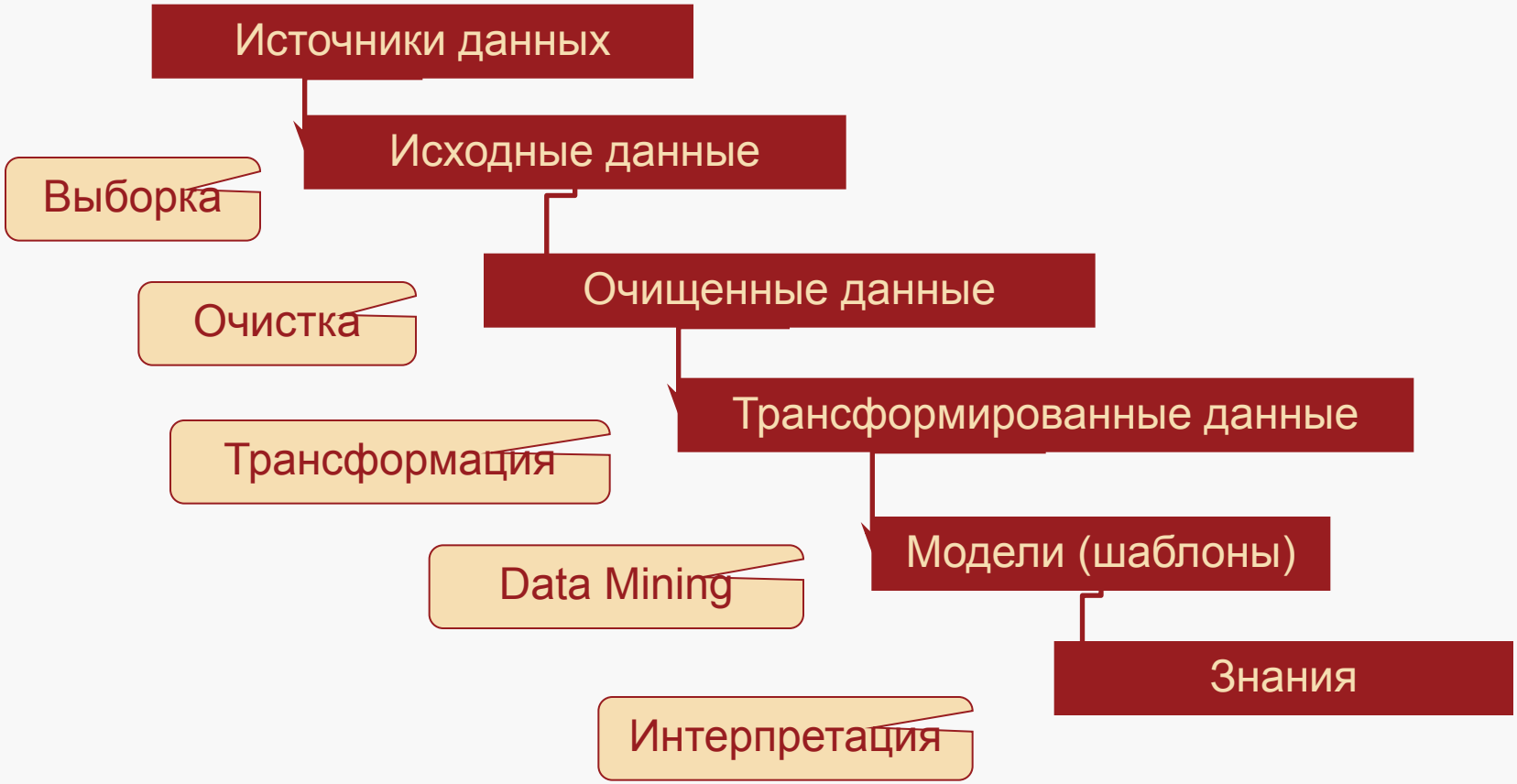
Но самое главное, что полученные таким образом знания можно **тиражировать**, т.е. построенную одним человеком модель могут применять другие без необходимости понимания методик, при помощи которых эти модели построены.

Методика извлечения знаний

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.

Knowledge Discovery in Databases



KDD – выборка данных

Первым шагом в анализе является получение исходной выборки. На основе этих данных и строятся модели. На этом шаге необходимо активное участие эксперта для выдвижения гипотез и отбора факторов, влияющих на анализируемый процесс. Желательно, чтобы данные были уже собраны и консолидированы. Крайне необходимо наличие удобных механизмов подготовки выборок.

Чаще всего в качестве источника рекомендуется использовать специализированное **хранилище данных**, агрегирующее всю необходимую для анализа информацию.

KDD – очистка данных

Реальные данные для анализа редко бывают хорошего качества. Необходимость предварительной обработки при анализе данных возникает независимо от того, какие технологии и алгоритмы используются. Более того, эта задача может представлять **самостоятельную ценность** в областях, не имеющих непосредственного отношения к анализу данных.

К задачам очистки относятся:

- Заполнение пропусков и редактирование аномалий
- Сглаживание, очистка от шумов
- Редактирование дубликатов и противоречий
- Устранение незначащих факторов

и прочее…

KDD – трансформация данных

Трансформация данных – последний этап перед, собственно, анализом. Различные алгоритмы анализа требуют специальным образом **подготовленные данные**, например, для прогнозирования необходимо преобразовать временной ряд при помощи скользящего окна.

Задачи трансформации данных:

- Скользящее окно
- Приведение типов
- Выделение временных интервалов
- Преобразование непрерывных значений в дискретные и наоборот
- Сортировка, группировка, агрегация

и прочее…

KDD – Data Mining

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

Информация, найденная в процессе применения методов Data Mining, должна быть нетривиальной и ранее неизвестной, например, средние продажи не являются таковыми. Знания должны описывать **новые связи** между свойствами, предсказывать значения одних признаков на основе других.

Data Mining – задачи

Задачи, решаемые методами Data Mining:

- **Классификация** – это отнесение объектов к одному из заранее известных классов.
- **Регрессия** – установление зависимости непрерывных выходных переменных от входных значений.
- **Кластеризация** – объекты внутри кластера должны быть «похожими» друг на друга и отличаться от объектов, вошедших в другие кластеры.
- **Ассоциация** – нахождение зависимости, что из события X следует событие Y.
- **Последовательность** – установление зависимостей между связанными во времени событиями.

Можно говорить еще и о задаче **анализа отклонений** – выявление наиболее нехарактерных шаблонов.

Data Mining – алгоритмы

Для решения вышеописанных задач используются различные методы и алгоритмы Data Mining. Ввиду того, что Data Mining развивался и развивается на стыке таких дисциплин, как статистика, теория информации, машинное обучение, теория баз данных, вполне закономерно, что большинство алгоритмов и методов Data Mining были разработаны на основе различных методов из этих дисциплин.

На сегодня наибольшее распространение получили **самообучающиеся методы** и **машинное обучение**.

KDD – интерпретация

В случае, когда извлеченные знания непрозрачны для пользователя, должны существовать методы постобработки, позволяющие привести их к интерпретируемому виду.

Для оценки качества полученной модели нужно использовать как **формальные методы оценки**, так и **знания эксперта**.

Полученные модели являются по сути формализованными знаниями эксперта, поэтому их можно **тиражировать**.

Достоинства и недостатки моделей

Достоинства:

- Возможность тиражирования знаний
- Обработка огромных объемов данных
- Обнаружение нетривиальных закономерностей
- Формализация процесса принятия решений

Недостатки:

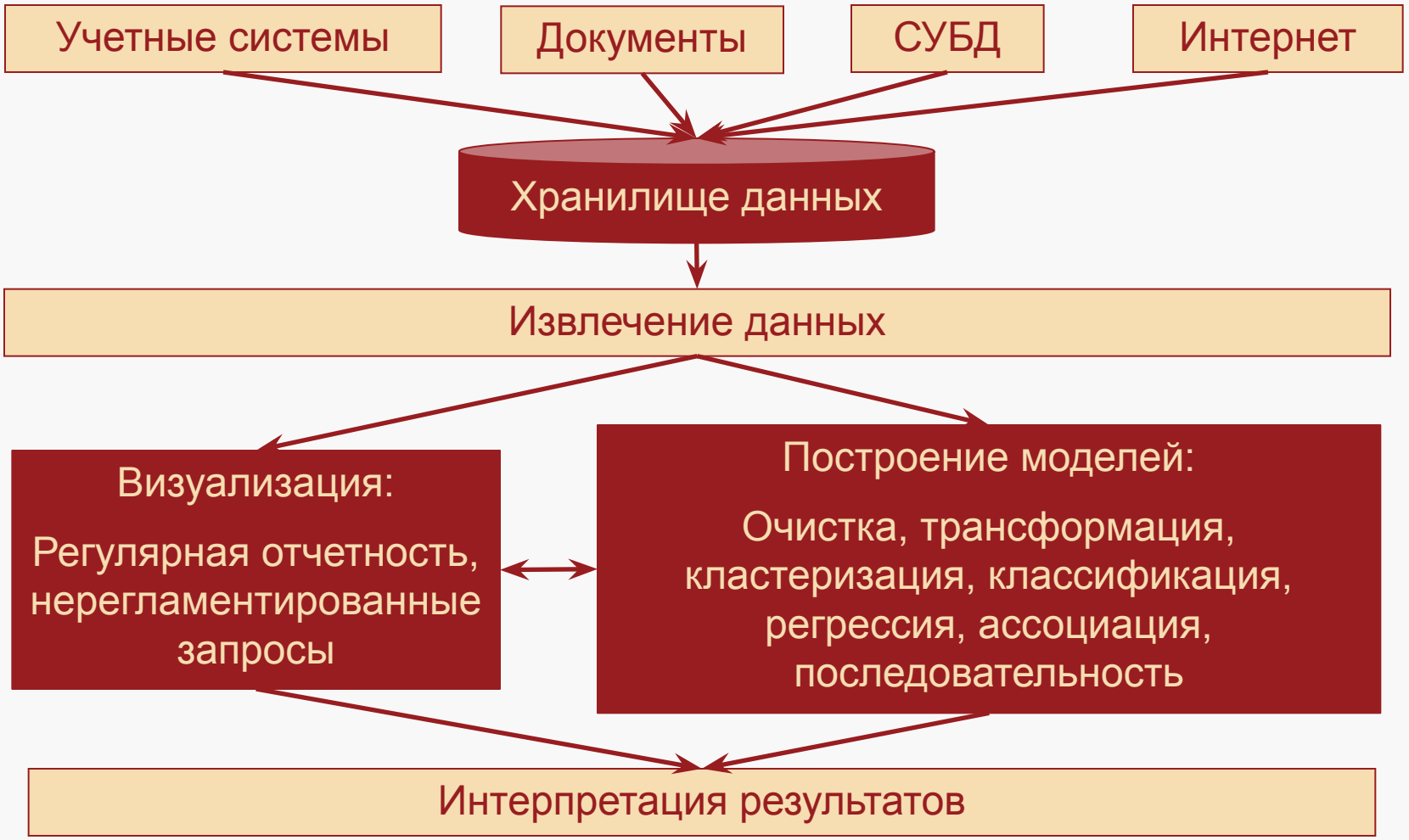
- Строгие требования к качеству и количеству данных
- Неспособность анализировать нестандартные случаи
- Высокие требования к знаниям эксперта

Аналитическая система

Наиболее оптимальной с точки зрения гибкости, возможностей и простоты использования является аналитическая система состоящая из хранилища данных, механизмов визуализации и методов построения моделей.

Подобная система позволяет **комбинировать подходы** к анализу данных. На стыке использования различных методов анализа получаются наиболее интересные результаты.

Схема аналитической системы



Решаемые бизнес-задачи

Подавляющее большинство бизнес-задач сводится к комбинированию описанных методов. Фактически, ранее были описаны базовые блоки, из которых собирается практически **любое бизнес-решение**:

- План-факторный анализ – визуализация данных
- Прогнозирование – задача регрессии
- Управление рисками – регрессия, кластеризация и классификация
- Стимулирование спроса – кластеризация, ассоциация
- Оценка эластичности спроса – регрессия
- Выявление предпочтений клиентов – последовательность, кластеризация…

Реализация в Deductor

Аналитическая платформа Deductor создавалась как система, реализующая описанную выше схему анализа. Она включает в себя хранилище данных и большой набор методов построения моделей.

Любые данные, полученные из хранилища данных, иного источника или в результате обработки, можно отобразить при помощи большого набора визуализаторов. **Универсальные методы анализа**, реализованные в Deductor, позволяют применять его для решения самого широкого спектра задач.

BaseGroup Labs

BaseGroup Labs – профессиональный поставщик Data Warehouse, OLAP, KDD, Data Mining решений и инструментов.

Web-сайт: www.basegroup.ru

Образование: edu.basegroup.ru

E-mail: info@basegroup.ru