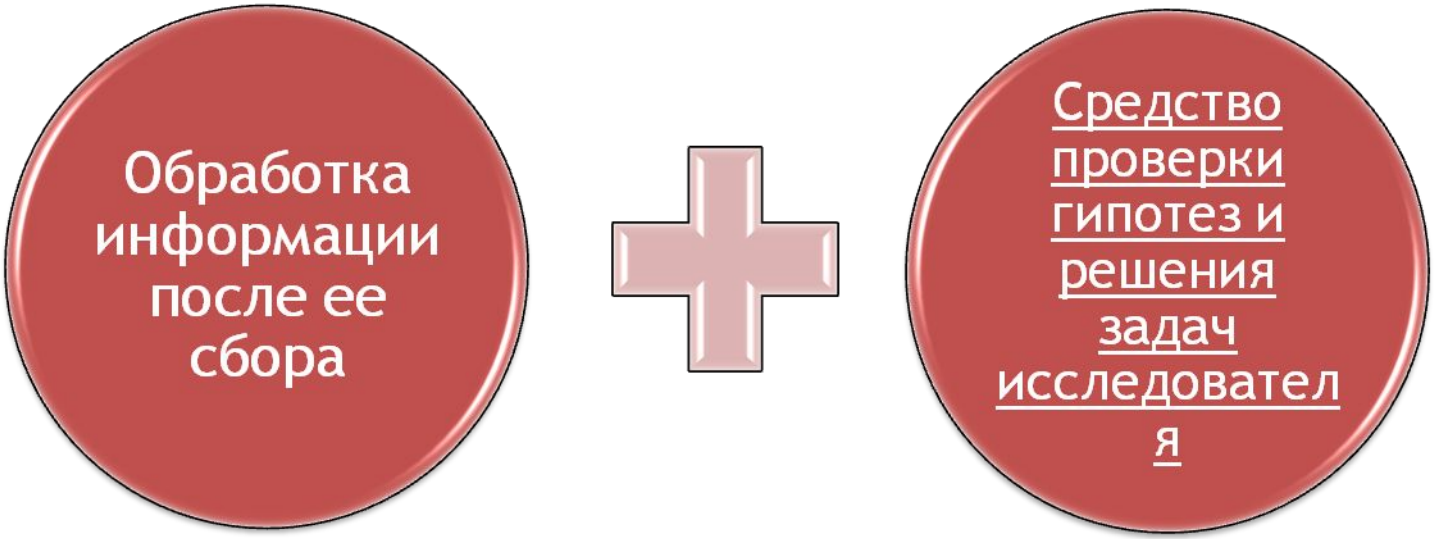
# Современные компьютерные технологии сбора, хранения, обработки, анализа и передачи информации для решения профессиональных задач

«Информационные технологии в науке и производстве», тема 3

Д.т.н., доц. Ханова А.А.

# Анализ данных

- исследования, связанные с обсчетом многомерной системы данных, имеющей множество параметров




- анализ данных тесно связан с моделированием

## *Пример*

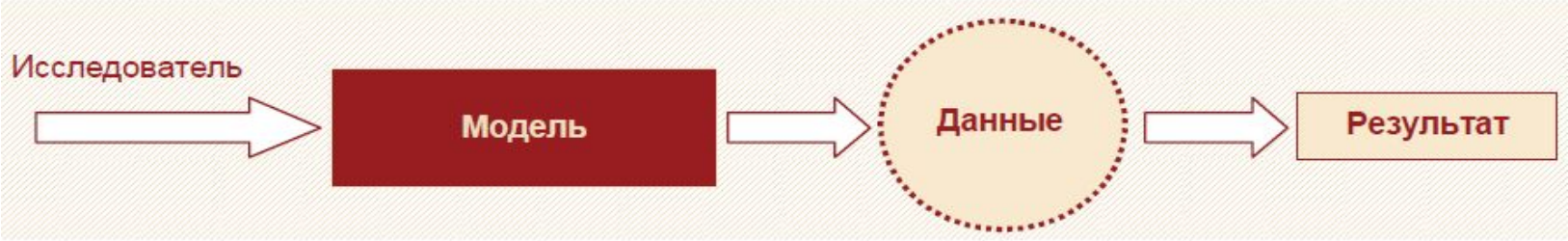
*Рассмотрим экономическую систему. Величина ожидаемого спроса* s *на будущий месяц* (t + 1) *рассчитывается на основе формулы* s(t + 1) = [s(t) + s(t – 1) + s(t – 2)] / 3, *то есть как среднее от продаж за предыдущие три месяца. Это простейшая математическая* модель прогноза продаж. *При построении этой модели были приняты следующие гипотезы.*

*Во-первых, годовая сезонность в продажах отсутствует.*

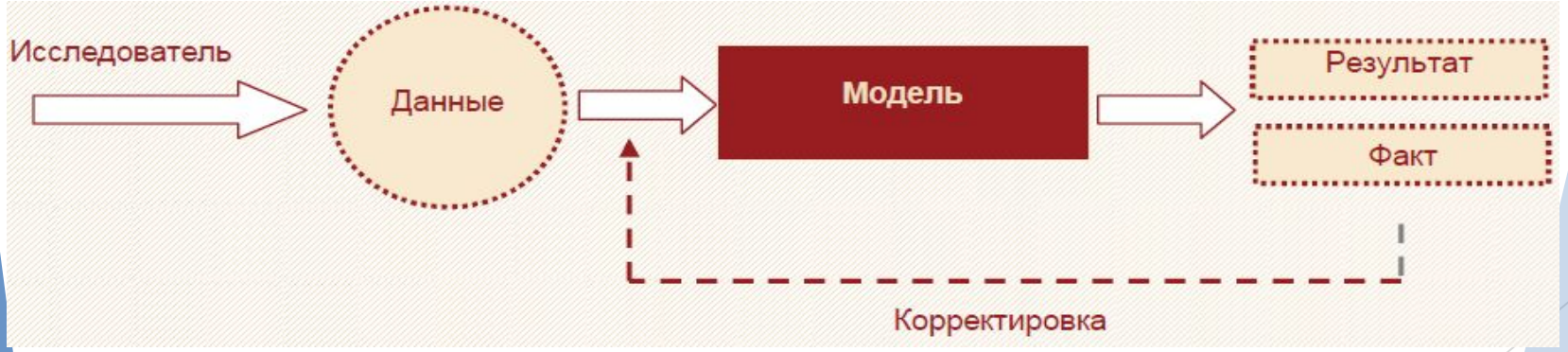
*Во-вторых, на величину продаж не влияют никакие внешние факторы: действия конкурентов, макроэкономическая ситуация и т. д.*

# Подходы к моделированию

- Аналитический




- При аналитическом подходе мы пытаемся подобрать существующую аналитическую модель таким образом, чтобы она адекватно отражала реальность
- Информационный




- при информационном подходе отправной точкой являются данные, характеризующие исследуемый объект, и модель «подстраивается» под действительность.

## Пример

*В банковском риск-менеджменте широко известна модель Дюрана для расчета рейтинга кредитоспособности заемщика, которая получила распространение в 40–50-е гг. XX в. На основе собственного опыта Дюран разработал балльную модель для оценки заемщика по совокупности его имущественных и социальных параметров (возраст, пол, профессия и т. д.). Преодолев некоторый порог, заемщик считался кредитоспособным. Эта модель представляет собой аналитическую зависимость* y = f(*X*), где y – рейтинг, *X* – набор признаков заемщика.

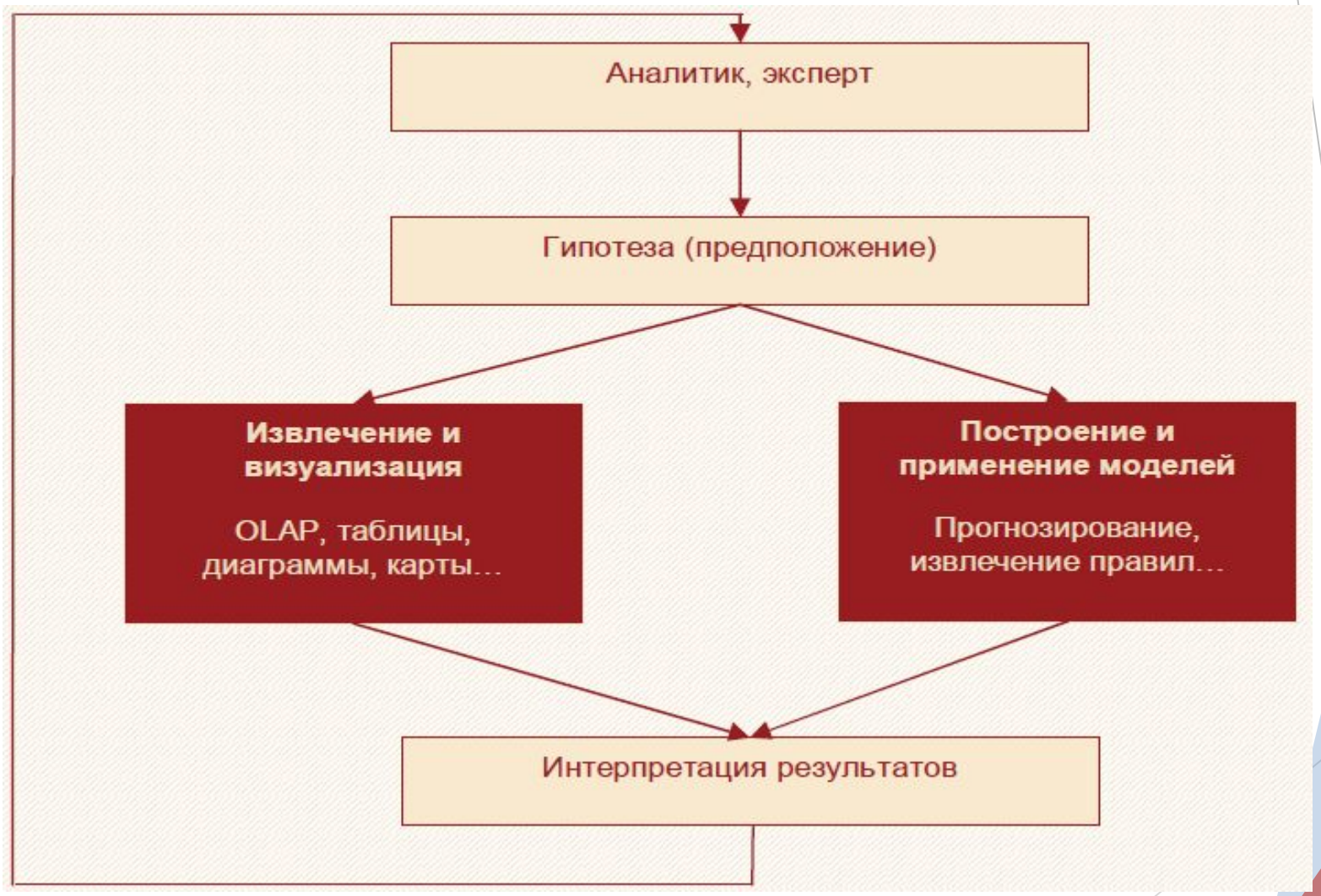
*Если перед современным российским банком встанет задача рассчитать рейтинг заемщика, банк может воспользоваться моделью Дюрана. Однако будет ли адекватной для современной российской действительности модель, разработанная в середине прошлого века на Западе? Естественно, не будет, так как она не учитывает связи между характеристиками российских заемщиков (возраст, образование, доход и т.д.) и дефолтностью по кредитам. Если же банк возьмет собственные данные по кредитным историям и на их основе построит модель, рассчитывающую рейтинг клиента, то, вполне вероятно, она окажется работоспособной.*

*В первом случае, когда мы брали модель Дюрана, мы использовали* аналитический подход. *Во втором —* информационный; *для построения модели нам понадобились данные — кредитные истории заемщиков банка.*

# Процесс анализа

- *Эксперт — специалист в предметной области, профессионал, который за годы обучения и практической деятельности научился эффективно решать задачи, относящиеся к конкретной предметной области*
- *Гипотеза - предположение о влиянии какого-либо фактора или группы факторов на результат.*
- *Аналитик — специалист в области анализа и моделирования:*
  - *владеет инструментальными и программными средствами анализа данных,*
  - *систематизирует данные, проводит опрос мнений экспертов,*
  - *координирует действий всех участников проекта по анализу данных.*

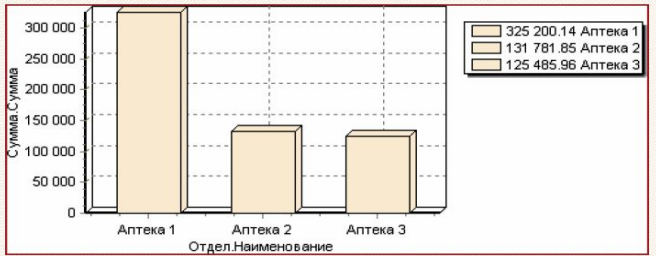
# Общая схема анализа

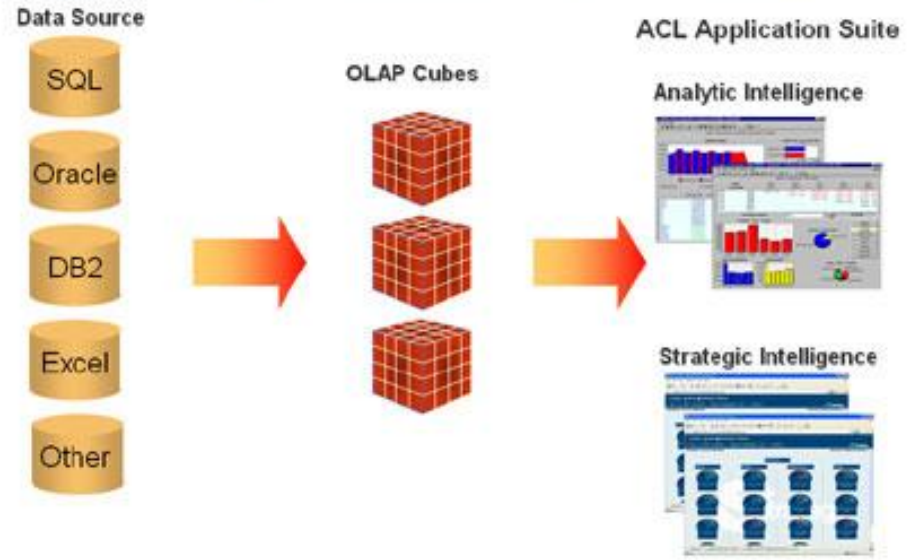


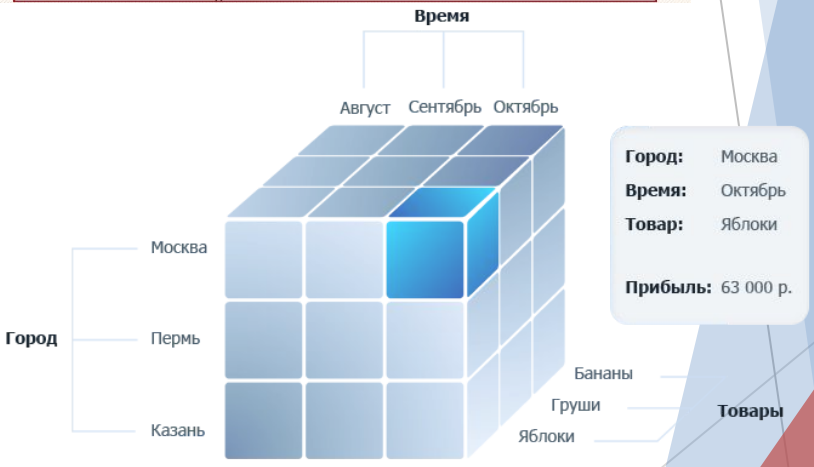
# Извлечение и визуализация данных

- Способы визуализации:
  - многомерные кубы;
  - таблицы;
  - диаграммы, гистограммы;
  - карты, проекции, срезы и т.п.

| Дата | Отдел.Наименование | Группа.Наименование | Количество | Су |
|---|---|---|---|---|
| 18.06.2004 | Аптека 1 | Антисептики и дезинфицирующие средства | 3 | |
| 18.06.2004 | Аптека 1 | Витамины и витаминоподобные средства | 3 | |
| 18.06.2004 | Аптека 1 | Иммуномодуляторы | 2 | |
| 18.06.2004 | Аптека 1 | Местные анестетики | 1 | |
| 18.06.2004 | Аптека 2 | Антисептики и дезинфицирующие средства | 1 | |
| 18.06.2004 | Аптека 2 | Витамины и витаминоподобные средства | 1 | |
| 19.06.2004 | Аптека 1 | Антисептики и дезинфицирующие средства | 4 | |
| 19.06.2004 | Аптека 1 | Витамины и витаминоподобные средства | 3 | |
| 19.06.2004 | Аптека 1 | Желчегонные средства и препараты желчи | 1 | |
| 19.06.2004 | Аптека 1 | Местные анестетики | 1 | |
| 19.06.2004 | Аптека 1 | Микро- и макроэлементы | 1 | |
| 19.06.2004 | Аптека 1 | Общетонизирующие средства и адаптогены | 1 | |

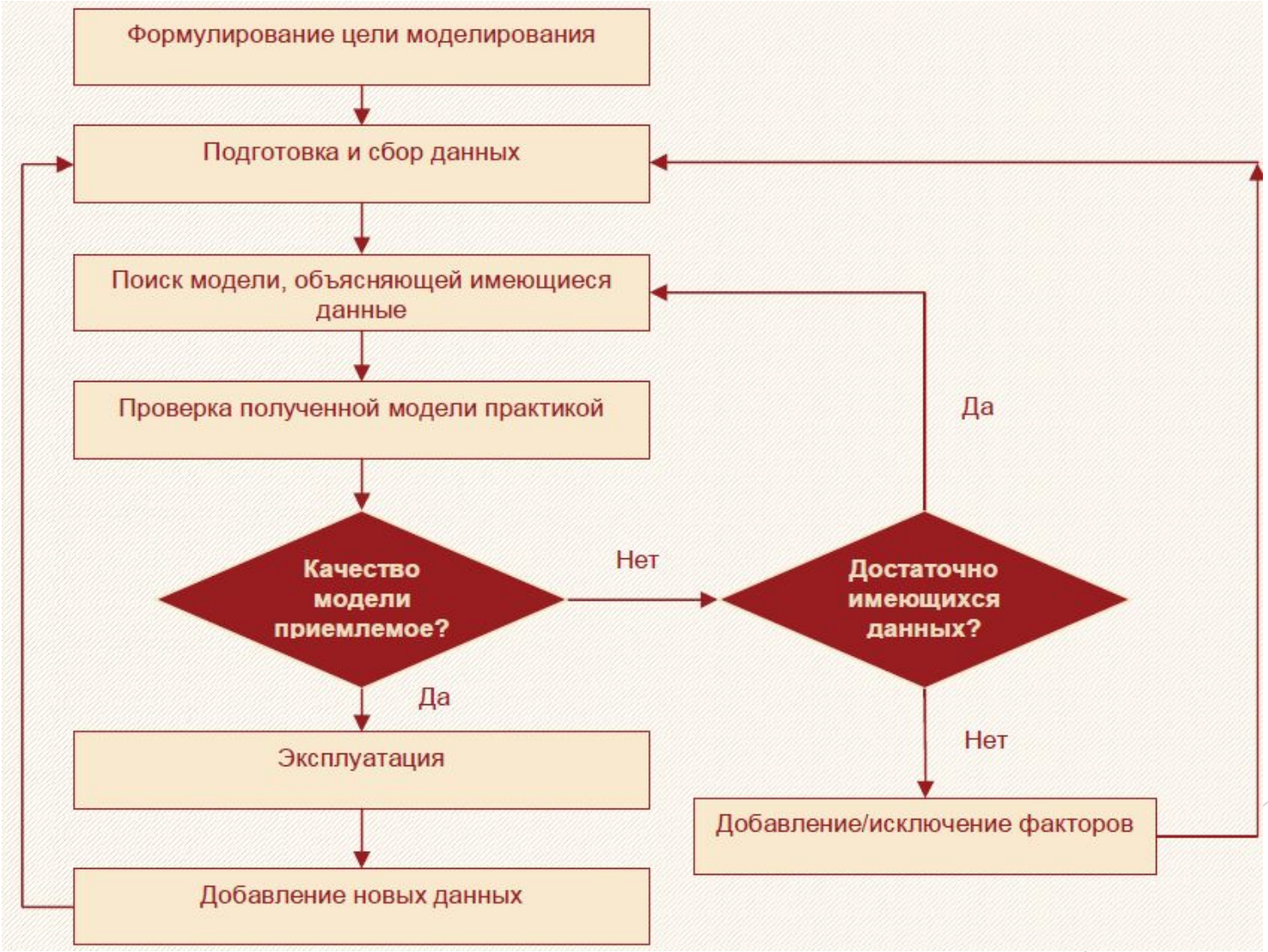


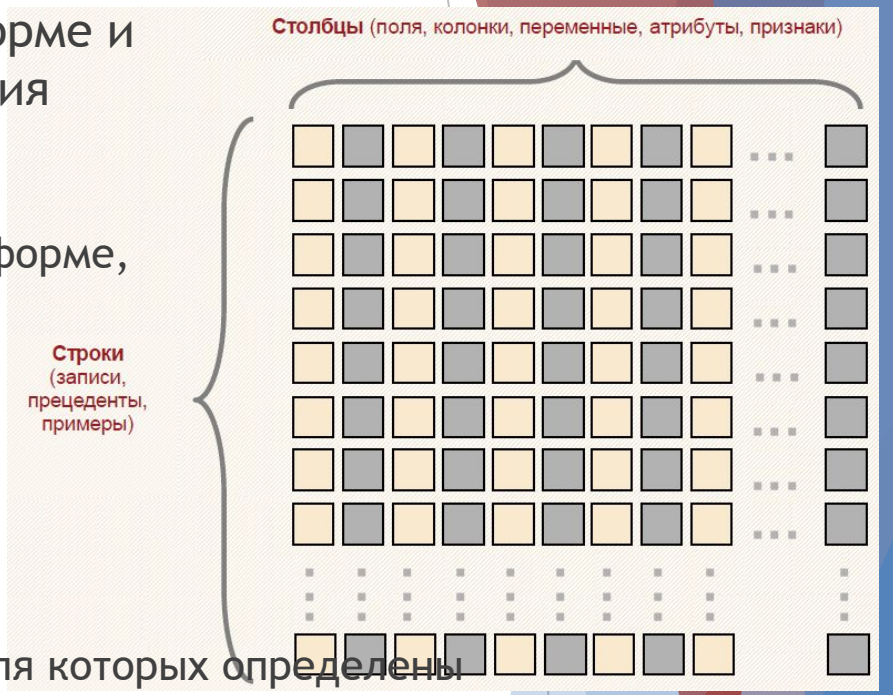

- - люди не могут обнаруживать сложные и нетривиальные зависимости, невозможно отделить знания от эксперта и тиражировать знания

# Этапы моделирования



# Формы представления данных

- Данные – сведения, характеризующие систему, явление, процесс или объект, представленные в определенной форме и предназначенные для дальнейшего использования
- По степени структурированности:
  - Неструктурированные данные - произвольные по форме, включают тексты и графику, мультимедиа (видео, речь, аудио).
  - Структурированные данные отражают отдельные факты предметной области. Это основная форма представления сведений в базах данных.
  - Слабоструктурированные данные — это данные, для которых определены некоторые правила и форматы, но в самом общем виде. Например, строка с адресом, строка в прайс-листе, ФИО и т. п.




390045 г. Рязань, ул. Ленина, д. 45 корп. 1

| Поле | Значение |
|---|---|
| Индекс | 390045 |
| Город | Рязань |
| Улица | Ленина |
| Дом | 45 |
| Корпус | 1 |

# Типы структурированных данных

- целый (количество товара, код товара и т. п.);
- вещественный (цена, скидка и т. п.);
- строковый (фамилия, наименование, адрес, пол, образование и т.п.);
- логический;
- дата/время.

# Виды структурированных данных

Непрерывные данные — данные, значения которых могут принимать какое угодно значение в некотором интервале. Над непрерывными данными можно производить арифметические операции сложения, вычитания, умножения, деления, и они имеют смысл.

Дискретные данные — значения признака, общее число которых конечно либо бесконечно, но может быть подсчитано при помощи натуральных чисел от одного до бесконечности. С дискретными данными не могут быть произведены никакие арифметические действия, либо они не имеют смысла.

| Тип данных | Вид данных | |
|---|---|---|
| | Непрерывный | Дискретный |
| Целый | + | + |
| Вещественный | + | + |
| Строковый | | + |
| Логический | | + |
| Дата/время | + | + |

Соответствие между типами и видами данных

# Представления наборов данных

- Упорядоченный набор данных - каждому столбцу соответствует один фактор, а в каждую строку заносятся упорядоченные по какому-либо признаку (например, время) события с интервалом периода между строками.

| Дата | Количество | Сумма |
|---|---|---|
| 01.01.2004 | 4 | 283.31 |
| 01.01.2004 | 1 | 72.48 |
| 01.01.2004 | 1 | 173.32 |
| 02.01.2004 | 6 | 294.84 |
| 02.01.2004 | 2 | 405.76 |
| 02.01.2004 | 12 | 303.13 |
| 02.01.2004 | 1 | 210.5 |
| 03.01.2004 | 6 | 521.16 |
| 03.01.2004 | 3 | 156.96 |

| Глубина | BK | DS |
|---|---|---|
| 887.9 | 8.85 | 0.218 |
| 888.1 | 9.627 | 0.216 |
| 888.3 | 14.584 | 0.217 |
| 888.5 | 21.647 | 0.215 |
| 888.7 | 17.172 | 0.216 |
| 888.9 | 6.118 | 0.215 |
| 889.1 | 2.886 | 0.217 |
| 889.3 | 2.506 | 0.219 |

- Неупорядоченный набор - каждому столбцу соответствует фактор, а в каждую строку заносится пример (ситуация, прецедент), упорядоченность строк не требуется.

| Номер | Банк | Реутеры | Филиалы | Город | Собственные активы |
|---|---|---|---|---|---|
| 2 | Внешторгбанк | - | 32 | Москва | 23236327 |
| 3 | Газпромбанк | GZPM | 27 | Москва | 9255041 |
| 4 | ООО "Международный Промышленный банк" | TIBP | 4 | Москва | 26409116 |
| 5 | Международный Московский Банк | IMBX | 1 | Москва | 1176462 |
| 6 | ОАО "АЛЬФА-БАНК" | ALFM | 17 | Москва | 12446938 |
| 7 | ОАО "ПСБ" | ICSP | 44 | Санкт-Петербург | 1275859 |
| 8 | Банк Москвы | - | 34 | Москва | 3335734 |
| 9 | АКБ "РОСБАНК" (ОАО) | - | 13 | Москва | 4691449 |
| 10 | АКБ "ДИБ" | DIBM | 0 | Москва | 2616993 |

- Транзакционные **данные -** любые связанные объекты или действия.

# Подготовка данных к анализу

- **Особенности данных, накопленных в организациях**
  - *Данные редко накапливаются специально для решения задач анализа.*
  - *Данные, как правило, содержат ошибки, аномалии, противоречия и пропуски.*
  - *С точки зрения анализа объемы хранимых данных очень велики.*
- **Принципы формализации данных**
  1. Абстрагироваться от существующих ИС и имеющихся в наличии данных.
  2. Описать все факторы, потенциально влияющие на анализируемый процесс/объект.
  3. Экспертно оценить значимость каждого фактора.
  4. Определить способ представления информации — число, дата, да/нет, категория (то есть тип данных).
  5. Собрать все легкодоступные факторы.
  6. Обязательно собрать наиболее значимые, с точки зрения экспертов, факторы.
  7. Оценить сложность и стоимость сбора средних и наименее важных по значимости факторов.

# Решение задачи прогнозирования спроса

- При создании таблицы 1 следуют принципам 1-3 формализации данных.

Таблица 1

| Показатель | Экспертная оценка значимости (1–100) |
|---|---|
| Сезон | 100 |
| День недели | 80 |
| Объем продаж за предыдущие недели | 100 |
| Объем продаж за аналогичный период прошлого года | 95 |
| Рекламная кампания | 60 |
| Маркетинговые мероприятия | 40 |
| Качество продукции | 30 |
| Рейтинг бренда | 25 |
| Отклонение цены от среднерыночной | 60 |
| Наличие данного товара у конкурентов | 15 |

Далее необходимо определить способ представления данных и оценить стоимость их сбора. К таблице добавятся еще два столбца (таблица 2). И уже после этого можно принимать решение о том, какие факторы включать в анализ, а какими пренебречь. Очевидно, что все легкодоступные показатели с высокой экспертной значимостью требуется включать в рассмотрение. А фактором Качество продукции, например, можно пренебречь: по мнению экспертов, он малозначим, а стоимость его сбора велика

Таблица 2

| Показатель | Экспертная оценка значимости (1–100) | Способ представления | Экспертная оценка сложности получения |
|---|---|---|---|
| Сезон | 100 | Число | низкая |
| День недели | 80 | Дата | низкая |
| Объем продаж за предыдущие недели | 100 | Число | низкая |
| Объем продаж за аналогичный период прошлого года | 95 | Число | низкая |
| Рекламная кампания | 60 | Число | средняя |
| Маркетинговый бюджет | 40 | Число | средняя |
| Качество продукции | 30 | Строка (плохое/хорошее/отличное) | высокая |
| Рейтинг бренда | 25 | Строка (известный/малоизвестный и т. д.) | средняя |
| Отклонение цены от среднерыночной | 60 | Число | средняя |
| Наличие данного товара у конкурентов | 15 | Логическое (да/нет) | средняя |

# Методы сбора данных

1. Получение из учетных систем.
2. Получение данных из косвенных источников информации.
3. Использование открытых источников (статистические сборники, отчеты корпораций, опубликованные результаты маркетинговых исследований и пр.).
4. Приобретение аналитических отчетов у специализированных компаний
5. Проведение собственных маркетинговых исследований и аналогичных мероприятий по сбору данных.
6. Ввод данных вручную.

соизмерение затрат с результатами

представление в структурированном виде (MS Excel, DBase, текстовые файлы с разделителями или в набор таблиц в любой реляционной СУБД )

унифицированное представление данных

# Информативность данных

- **неинформативные признаки:**
  - признаки, содержащие *только одно значение (а)*;
  - признаки, содержащие *в основном одно значение (б)*;
  - признаки с *уникальными значениями (в)*;
  - признаки, *между которыми имеет место сильная корреляция, — в этом случае для* анализа можно взять один столбец (*г*).

| Признак |
|---|
| 1 |
| 1 |
| 1 |
| 1 |
| 1 |
| 1 |
| 1 |
| 1 |
| 1 |
| 1 |
| 1 |
| 1 |

*а)*

| Признак |
|---|
| 1 |
| 1 |
| 1 |
| 1 |
| 1 |
| 0 |
| 1 |
| 1 |
| 1 |
| 1 |
| 1 |
| 1 |

*б)*

| № паспорта |
|---|
| 0936-866096 |
| 8355-512943 |
| 8017-098471 |
| 2762-945535 |
| 0459-997701 |
| 6291-817248 |
| 0094-883508 |
| 6385-082612 |
| 9290-732300 |
| 7022-736158 |
| 3127-709332 |
| 4179-171975 |

*в)*

| Пол | Gender |
|---|---|
| Жен | 0 |
| Жен | 0 |
| Жен | 0 |
| Муж | 1 |
| Муж | 1 |
| Жен | 0 |
| Жен | 0 |
| Муж | 1 |
| Муж | 1 |
| Жен | 0 |
| Жен | 0 |
| Муж | 1 |

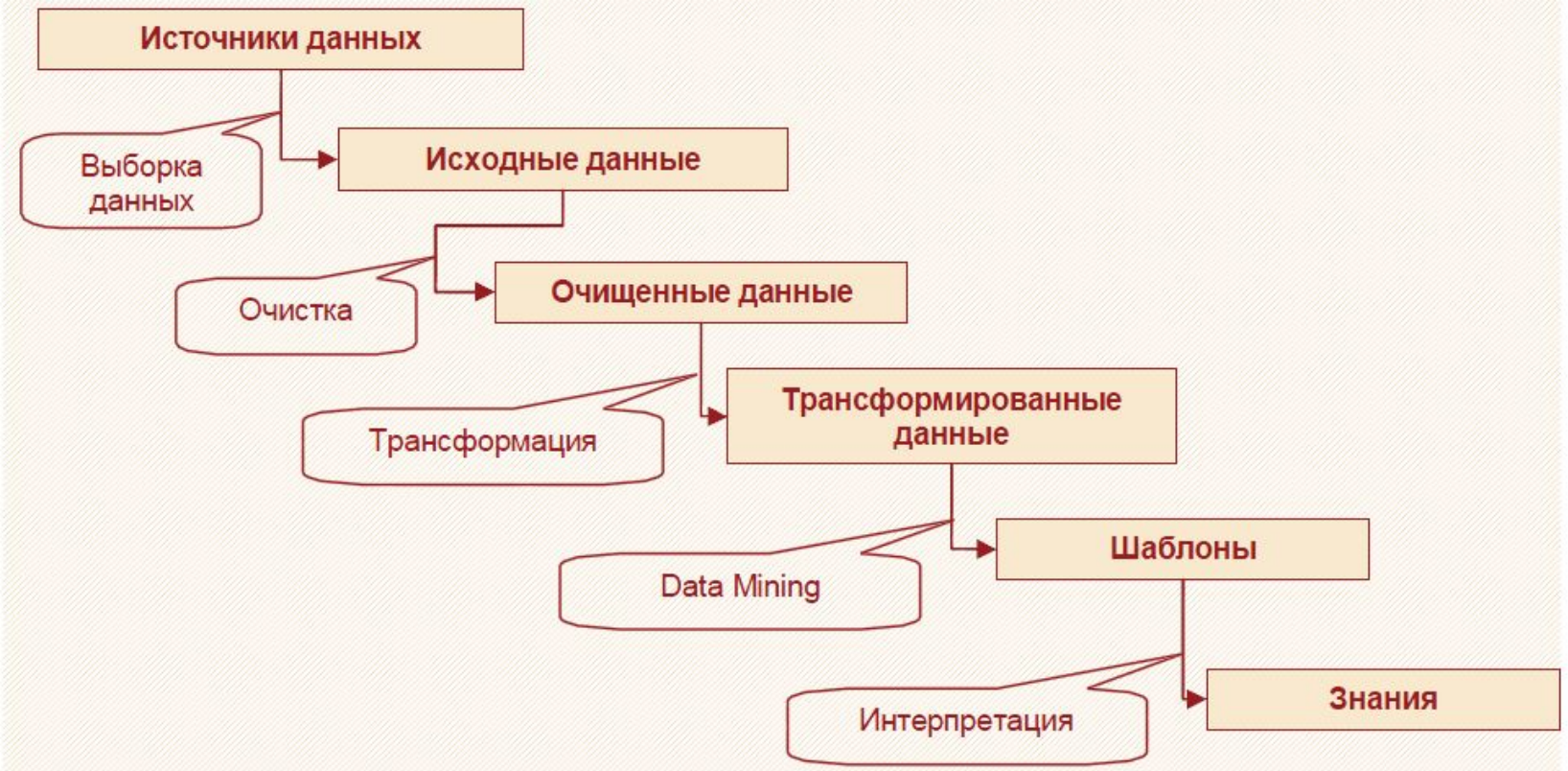
*г)*

# Требования к данным

- **Для временных рядов, которые относятся к упорядоченным данным.**
  - Если для моделируемого бизнес-процесса (например, продажи) характерна сезонность/цикличность, то необходимо иметь данные хотя бы за один полный сезон/цикл с возможностью варьирования интервалов (понедельное, помесячное и т. д.).
  - Максимальный горизонт прогнозирования зависит от объема данных:
    - данные за 1,5 года — прогноз возможен максимум на 1 месяц;
    - данные за 2–3 года — на 2 месяца.
- **Для неупорядоченных данных :**
  - Количество примеров (прецедентов) должно быть значительно больше количества факторов.
  - Желательно, чтобы данные покрывали как можно больше ситуаций реального процесса.
  - Пропорции различных примеров (прецедентов) должны примерно соответствовать реальному процессу.
- **Для Транзакционных данных.**
  - Анализ транзакций целесообразно производить на большом объеме данных, иначе могут быть выявлены статистически необоснованные правила. Алгоритмы поиска ассоциативных связей способны быстро перерабатывать огромные массивы данных.
  - 300–500 объектов — не менее 10 тыс. транзакций;
  - 500–1000 объектов — более 300 тыс. транзакций.

# Методика извлечения знаний

- Knowledge Discovery in Databases — процесс получения из данных знаний в виде зависимостей, правил, моделей, обычно состоящий из таких этапов, как выборка данных, их очистка и трансформация, моделирование и интерпретация полученных результатов.

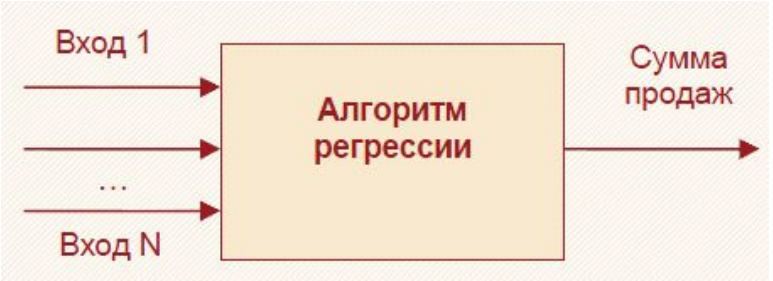


# Data Mining

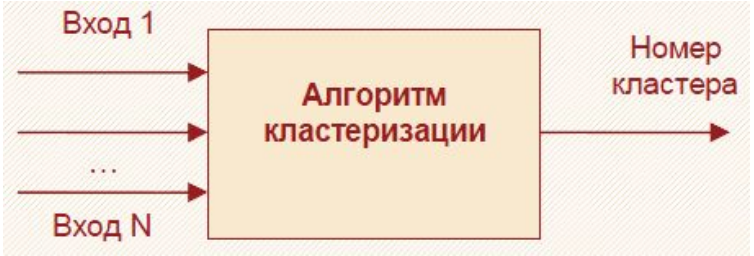
- обнаружение в «сырых» данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.
- Кассы задач:

1. Классификация – установление зависимости *дискретной выходной переменной от* входных переменных.
2. Регрессия – установление зависимости *непрерывной выходной переменной от* входных переменных.
3. Кластеризация – группировка объектов (наблюдений, событий) на основе данных, описывающих свойства объектов. Объекты внутри кластера должны быть похожими друг на друга и отличаться от других, которые вошли в другие кластеры.
4. Ассоциация – выявление закономерностей между связанными событиями. Примером такой закономерности служит правило, указывающее, что из события *X следует* событие *Y*. *Такие правила называются ассоциативными*.
5. Анализ отклонений (deviation detection).
6. Связей (link analysis).
7. Отбор значимых признаков (feature selection).

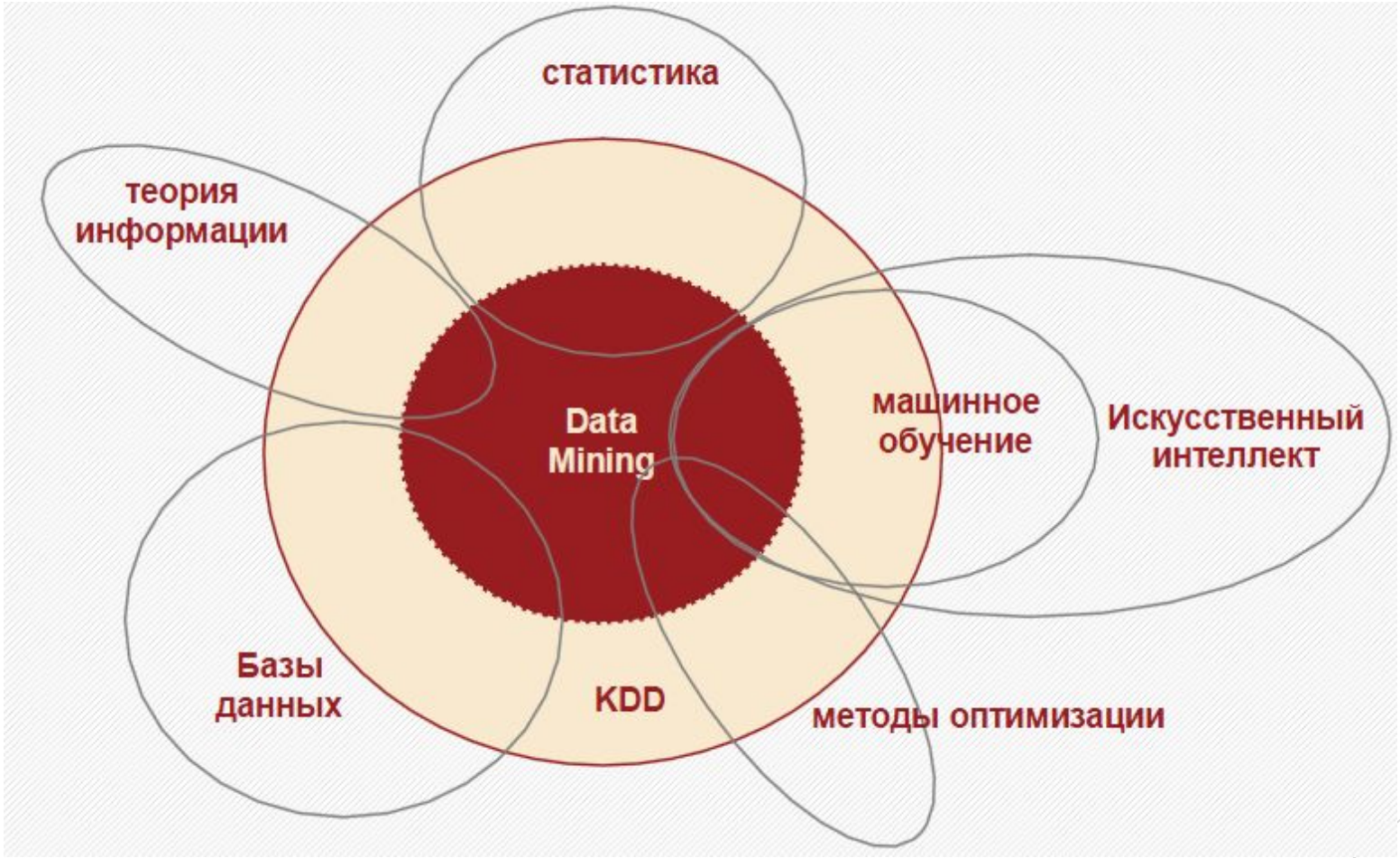

- Отнесение нового товара к той или иной товарной группе, клиента к какой-либо категории
- При кредитовании – по каким-то признакам к одной из групп риска




- Зависимость между суммой продаж, и факторами, влияющими на нее, (предыдущие объемы продаж, изменение курсов валют, активность конкурентов)
- При кредитовании физических лиц вероятность возврата кредита зависит от личных характеристик человека, сферы его деятельности, наличия имущества




- При достаточно большом количестве клиентов становится трудно подходить к каждому индивидуально, поэтому их удобно объединять в группы — сегменты с однородными признаками. Например по сфере деятельности, по географическому расположению. После кластеризации можно узнать, какие сегменты наиболее активны, какие приносят наибольшую прибыль, выделить характерные для них признаки. Эффективность работы с клиентами повышается благодаря учету их персональных предпочтений.

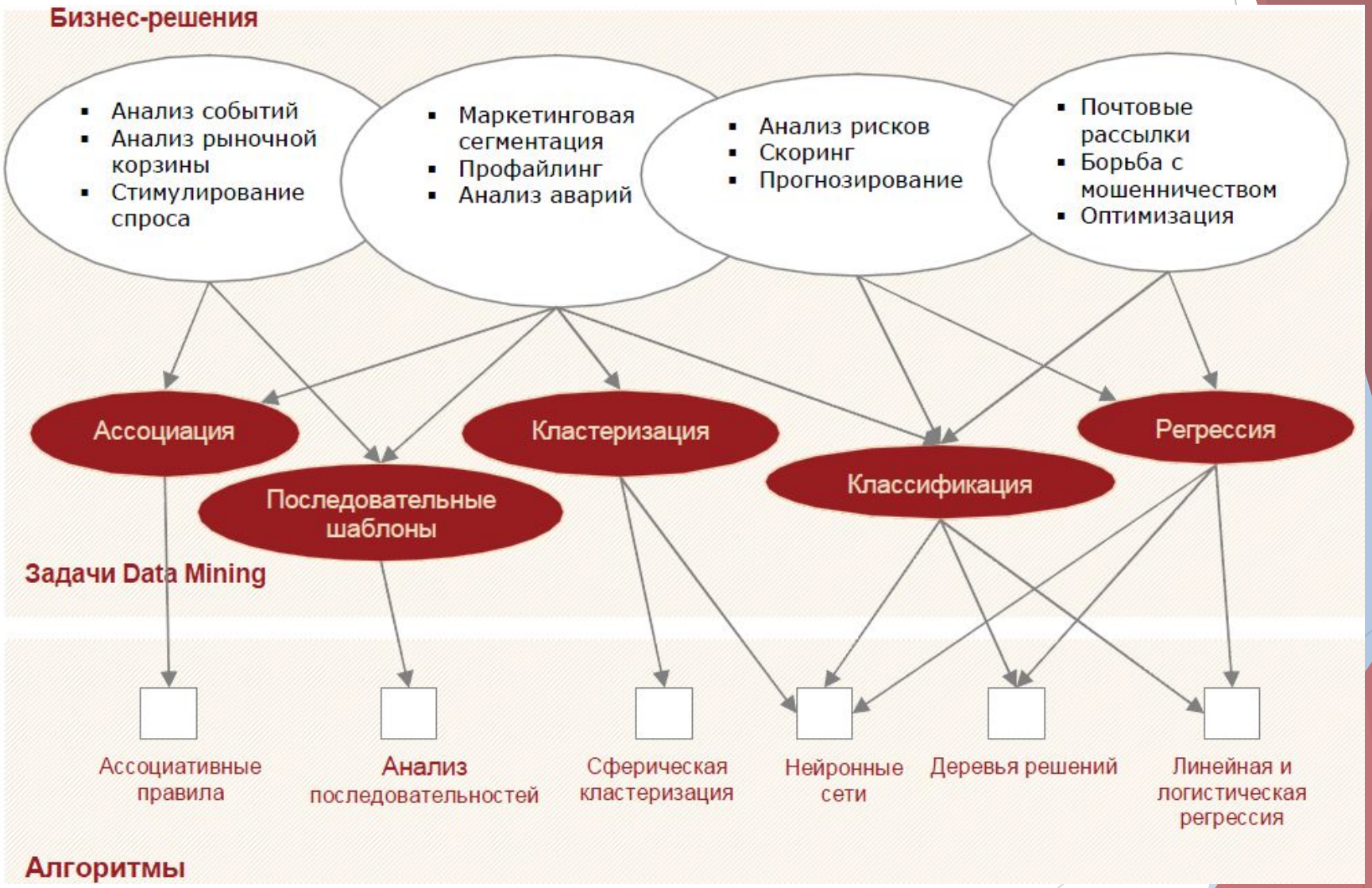

статистика
теория информации
Data Mining
машинное обучение
Искусственный интеллект
Базы данных
KDD
методы оптимизации

# Машинное обучение

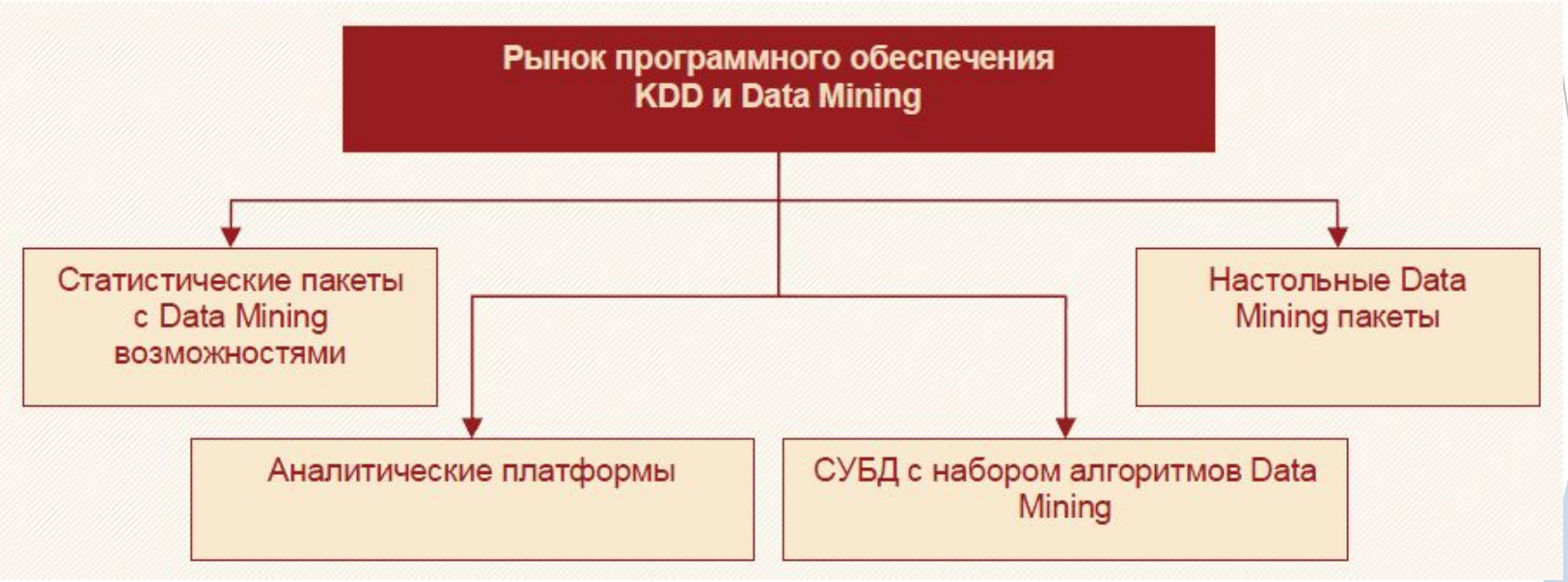
- Машинное обучение (machine learning) — обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться на данных.
- Имеется множество *объектов (ситуаций) и* множество возможных *ответов (откликов, реакций)*.
- *Между ответами и объектами существует* некоторая зависимость, но она неизвестна. Известна только конечная совокупность *прецедентов — пар вида «объект — ответ», — называемая* обучающей выборкой.
- ***На основе*** этих данных требуется обнаружить зависимость, то есть построить модель, способную для любого объекта выдать достаточно точный ответ. Чтобы измерить точность ответов, вводится критерий качества.

# Причины распространения KDD и Data Mining

1. **Развитие технологий автоматизированной обработки информации** создало основу для учета сколь угодно большого количества факторов и достаточного объема данных.
2. **Возникла острая нехватка высококвалифицированных специалистов в области** статистики и анализа данных. Поэтому потребовались технологии обработки и анализа, доступные для специалистов любого профиля за счет применения методов визуализации и самообучающихся алгоритмов.
3. **Возникла объективная потребность в тиражировании знаний.** Полученные в процессе KDD и Data Mining результаты являются формализованным описанием некоего процесса, а следовательно, поддаются автоматической обработке и повторному использованию на новых данных.
4. **На рынке появились программные продукты, поддерживающие технологии KDD и Data Mining**, – аналитические платформы. С их помощью можно создавать полноценные аналитические решения и быстро получать первые результаты.


Бизнес-решения
Анализ событий
Анализ рыночной корзины
Стимулирование спроса
Маркетинговая сегментация
Профайлинг
Анализ аварий
Анализ рисков
Скоринг
Прогнозирование
Почтовые рассылки
Борьба с мошенничеством
Оптимизация
Ассоциация
Последовательные шаблоны
Кластеризация
Классификация
Регрессия
Задачи Data Mining
Ассоциативные правила
Анализ последовательностей
Сферическая кластеризация
Нейронные сети
Деревья решений
Линейная и логистическая регрессия
Алгоритмы

# Программное обеспечение в области анализа данных

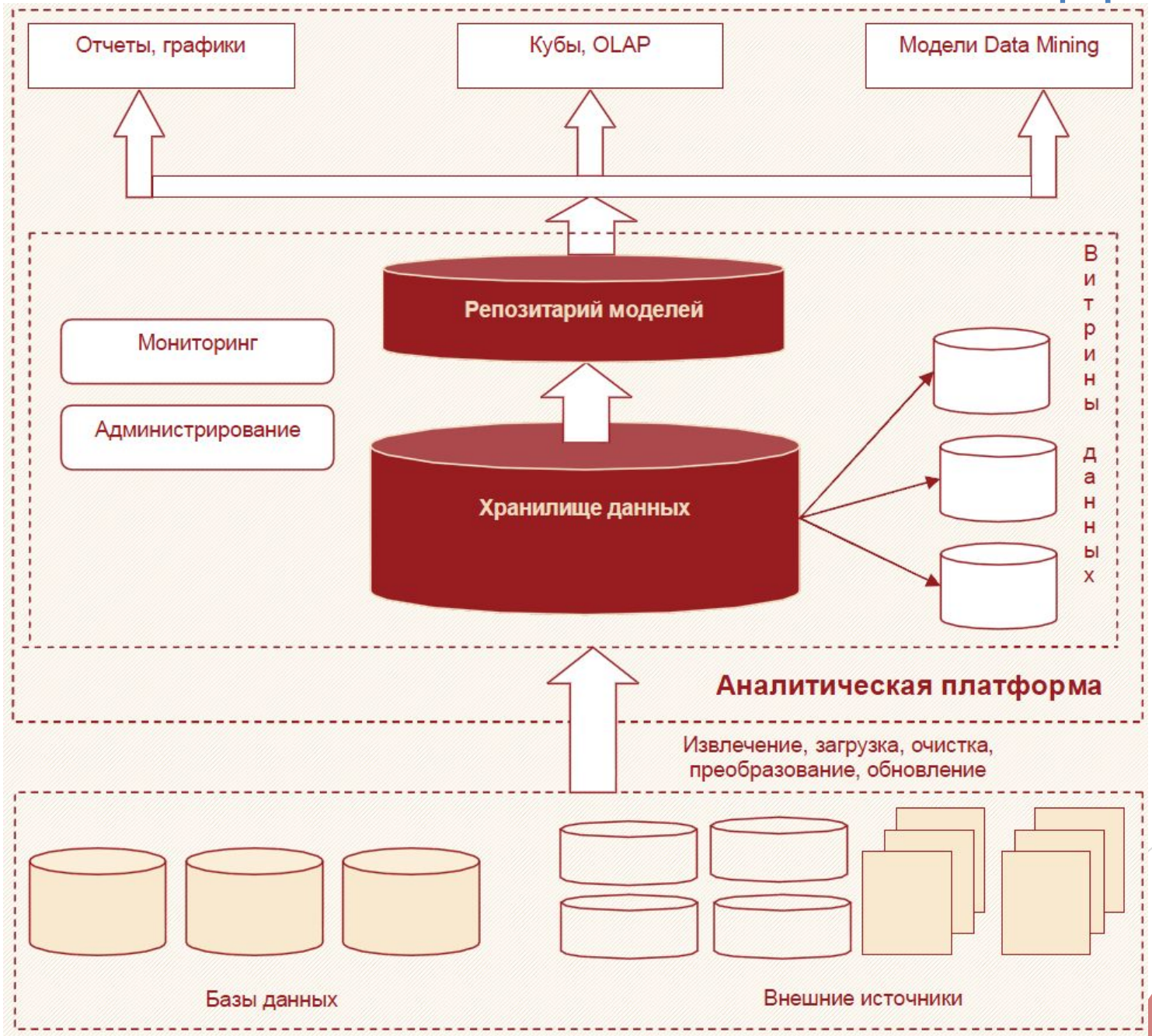


# *Статистические пакеты с возможностями Data Mining и настольные Data Mining пакеты*

1. слабая интеграция с промышленными источниками данных;
2. бедные средства очистки, предобработки и трансформации данных;
3. отсутствие гибких возможностей консолидации информации, например, в специализированном хранилище данных;
4. конвейерная (поточная) обработка новых данных затруднительна или реализуется встроенными языками программирования и требует высокой квалификации;
5. из-за использования пакетов на локальных рабочих станциях обработка больших объемов данных затруднена.

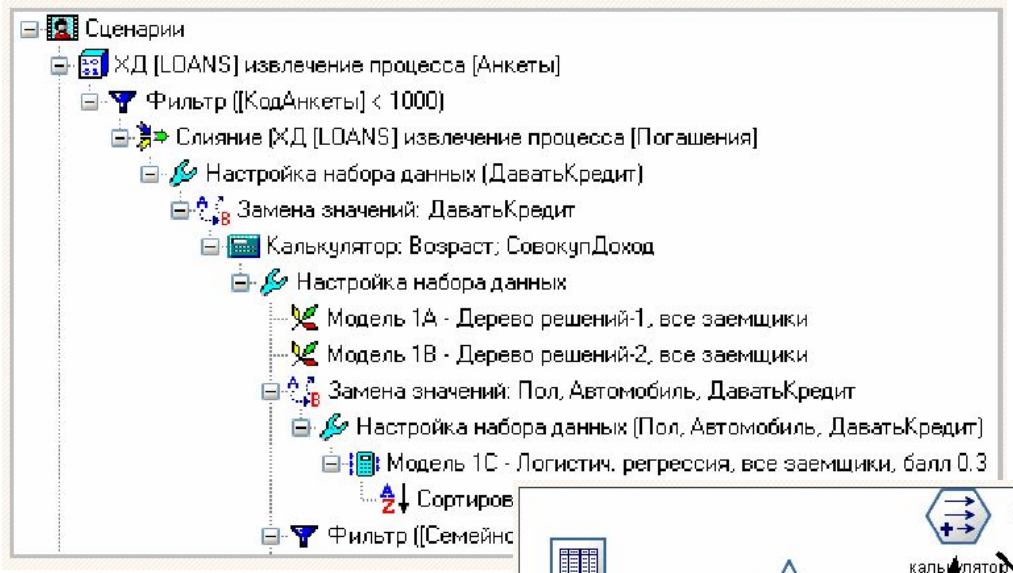
- СУБД с элементами Data Mining:
  - высокая производительность;
  - алгоритмы анализа данных по максимуму используют преимущества СУБД;
  - жесткая привязка всех технологий анализа к одной СУБД;
  - сложность в создании прикладных решений, поскольку работа с СУБД ориентирована на программистов и администраторов баз данных.
- 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,
  - возможности конвейерной» обработки новых данных.

# Типовая схема системы на базе аналитической платформы



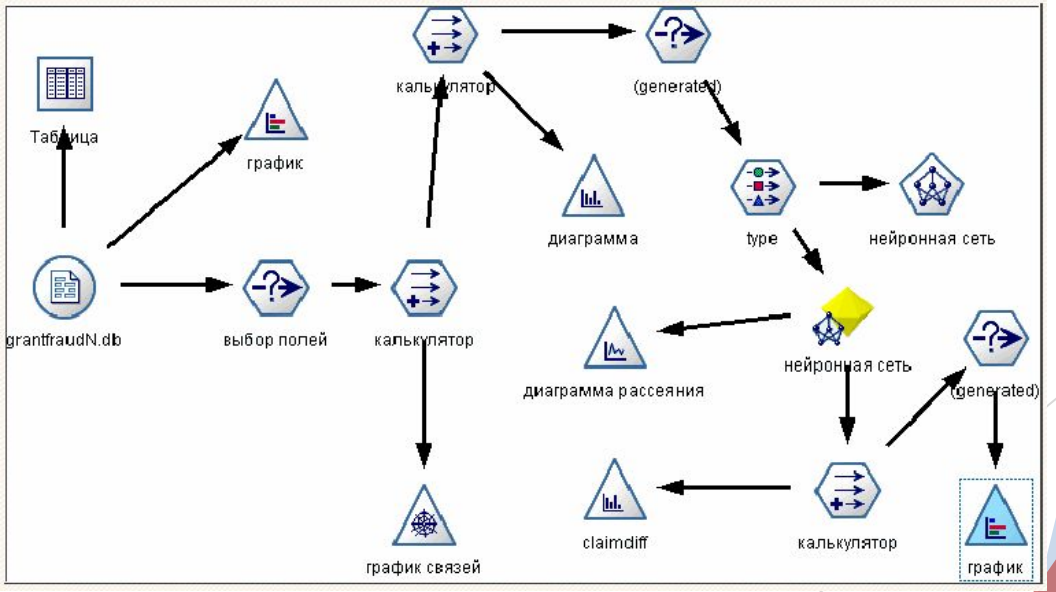
# Языки визуального моделирования

- !важно освободить аналитика от необходимости углубленного понимания сложных математических алгоритмов.




Формы представления диаграмм:
в виде дерева

и в виде графа



# Общие особенности языков моделирования в аналитических платформах

1. Базовым узлом, с которого начинается диаграмма, является узел импорта, поскольку в аналитических платформах обычно отсутствуют средства для ручного ввода данных; предполагается, что данные уже имеются в каких-либо источниках.
2. Графическое изображение, соответствующее какому-либо узлу, несет в себе большой семантический смысл. Оно помогает аналитику различать узлы по функциям и определять их активность (часто еще не выполненный узел обозначается иконкой серого цвета, а выполненный — цветной).
3. Диаграмма описывает формализованную последовательность действий над данными, и эти действия можно повторить на совершенно других данных, предварительно настроив соответствие колонок.