

Национальный исследовательский ядерный университет «МИФИ»

Факультет бизнес-информатики и управления
комплексными системами

Кафедра экономики и менеджмента
в промышленности (№ 71)

*Математические и инструментальные методы обработки
статистической информации*

ЛЕКЦИЯ 1

ЗАДАЧИ И СТАНДАРТЫ АНАЛИЗА ДАННЫХ

Киреев В.С.,

к.т.н., доцент

v.kireev@inbox.ru

Москва,
2017

Предпосылки к использованию интеллектуального анализа данных

- Данные имеют неограниченный объем
- Данные являются разнородными (количественными, качественными, текстовыми)
- Результаты должны быть конкретны и понятны
- Инструменты для обработки сырых данных должны быть просты в использовании

Парадокс:

Чем больше данных, тем меньше знаний

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.

Пирамида знаний



Применение интеллектуального анализа данных

- **Реклама и продвижение товара**

- Какова эффективность рекламы?

- **Перекрестные продажи**

- Какие продукты покупатель готов дополнительно приобрести?

- **Обнаружение мошенничества**

- Правильные ли сведения были поданы?

- **Удержание клиента**

- Какие клиенты готовы разорвать договор?

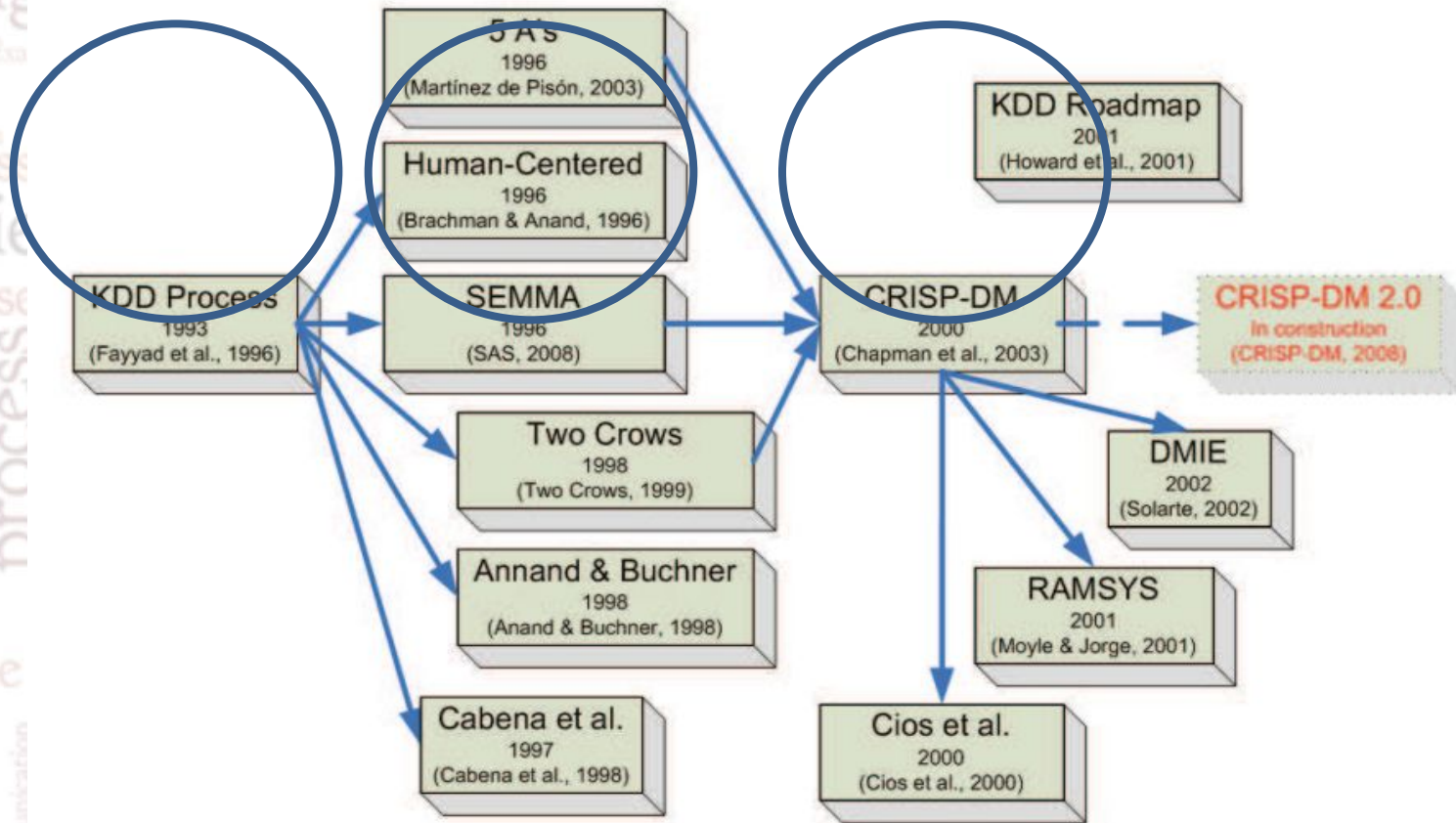
- **Управление рисками**

- Выдавать ли кредит данному заёмщику?

- **Сегментирование потребителей**

- Выдавать ли кредит данному заёмщику?

Развитие методологий анализа данных



Методология KDD

Несмотря на разнообразие бизнес-задач почти все они могут решаться по единой методике. Эта методика, зародившаяся в 1989 г., получила название **Knowledge Discovery in Databases** — извлечение знаний из баз данных. Она описывает не конкретный алгоритм или математический аппарат, а последовательность действий, которую необходимо выполнить для обнаружения полезного знания.

Методика не зависит от предметной области; это набор атомарных операций, комбинируя которые, можно получить нужное решение.

KDD включает в себя этапы подготовки данных, выбора информативных признаков, очистки, построения моделей, постобработки и интерпретации полученных результатов.

Этапы процесса анализа данных по методологии KDD



Методология KDD.

Выборка данных.

Первым шагом в анализе является получение исходной выборки. На основе отобранных данных строятся модели. Здесь требуется активное участие экспертов для выдвижения гипотез и отбора факторов, влияющих на анализируемый процесс. Желательно, чтобы данные были уже собраны и консолидированы. Крайне необходимы удобные механизмы подготовки выборки: запросы, фильтрация данных и сэмплинг. Чаще всего в качестве источника рекомендуется использовать специализированное хранилище данных, консолидирующее всю необходимую для анализа информацию.

Методология KDD.

Очистка данных.

Реальные данные для анализа редко бывают хорошего качества. Необходимость в предварительной обработке 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 данных относятся: заполнение пропусков, подавление аномальных значений, сглаживание, исключение дубликатов и противоречий и пр.

Методология KDD.

Трансформация данных.

Этот шаг необходим для тех методов, при использовании которых исходные данные должны быть представлены в каком-то определенном виде. Дело в том, что различные алгоритмы анализа требуют специальным образом подготовленных данных. Например, для прогнозирования необходимо преобразовать временной ряд при помощи скользящего окна или вычислить агрегированные показатели. К задачам трансформации данных относятся: скользящее окно, приведение типов, выделение временных интервалов, квантование, сортировка, группировка и пр.

Методология KDD. Data Mining.

Термин Data Mining дословно переводится как «добыча данных» или «раскопка данных» и имеет в англоязычной среде несколько определений. Data Mining — обнаружение в «сырых» данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности. Зависимости и шаблоны, найденные в процессе применения методов Data Mining, должны быть нетривиальными и ранее неизвестными, например, сведения о средних продажах таковыми не являются. Знания должны описывать новые связи между свойствами, предсказывать значения одних признаков на основе других.

Методология KDD. Интерпертация данных.

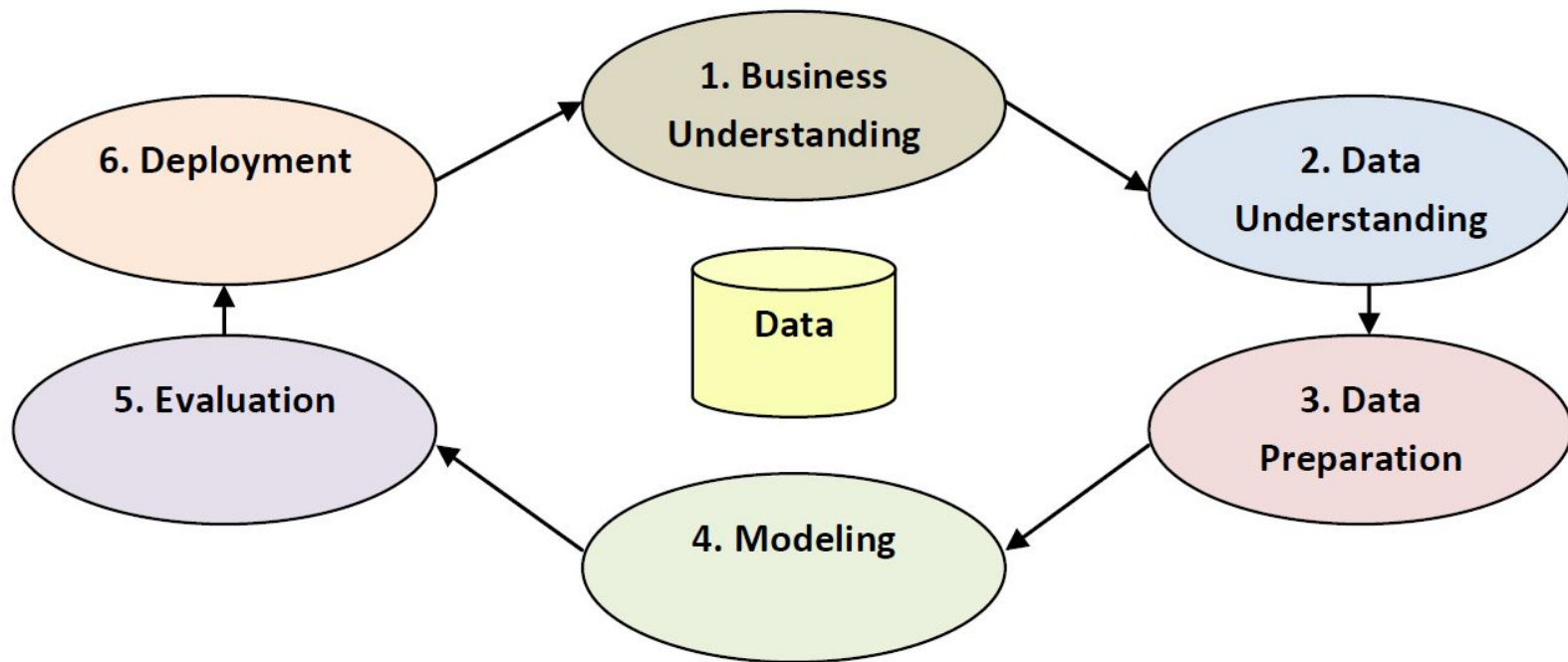
В случае, когда извлеченные зависимости и шаблоны непрозрачны для пользователя, должны существовать методы постобработки, позволяющие привести их к интерпретируемому виду. Для оценки качества полученной модели нужно использовать как формальные методы, так и знания аналитика. Именно аналитик может сказать, насколько применима полученная модель к реальным данным. Построенные модели являются, по сути, формализованными знаниями эксперта, а следовательно, их можно тиражировать. Найденные знания должны быть применимы и к новым данным с некоторой степенью достоверности.

Стандарт CRISP-DM

Хотя корни сбора данных могут быть прослежены до конца 1980-х, в течение большинства 1990-х, область была все еще в ее младенчестве. Интеллектуальный анализ данных все еще определялся и совершенствовался. Это было, в основном, свободное скопление моделей данных, аналитических алгоритмов и специальной продукции. В 1999 несколько больших компаний включая производителя автомобилей Daimler-Benz, страховую компанию OHRA, разработчика аппаратного и программного обеспечения NCR Corp. и разработчика статистического программного обеспечения SPSS, Inc. начали сотрудничать, чтобы формализовать и стандартизировать подход к сбору данных. Результатом их работы был кросс-индустриальный стандарт глубинного анализа данных (**CRISP-DM**, the **CR**oss-Industry **S**tandard **P**rocess for **D**ata **M**ining).

Хотя у участников создания **CRISP-DM**, конечно, были имущественные права в определенных инструментах программного и аппаратного обеспечения, процесс был разработан независимым от любого определенного инструмента или вида данных.

Этапы процесса анализа данных по стандарту **CRISP-DM**



Процессы понимания бизнеса

Определить бизнес цели

Оценить ситуацию

Определить цели анализа данных

Составить план проекта

Процессы понимания данных



Собрать исходные данные

Описать данные

Исследовать данные

Проверить качество данных

Процессы подготовки данных



Отобрать данные
Очистить данные
Сделать производные данные
Объединить данные
Привести данные в нужный формат

Процессы моделирования

Выбрать методику моделирования

Сделать тесты для модели

Построить модель

Оценить модель

Процессы оценки

Оценить результаты

Сделать ревью процесса

Определить следующие шаги

Процессы развёртывания



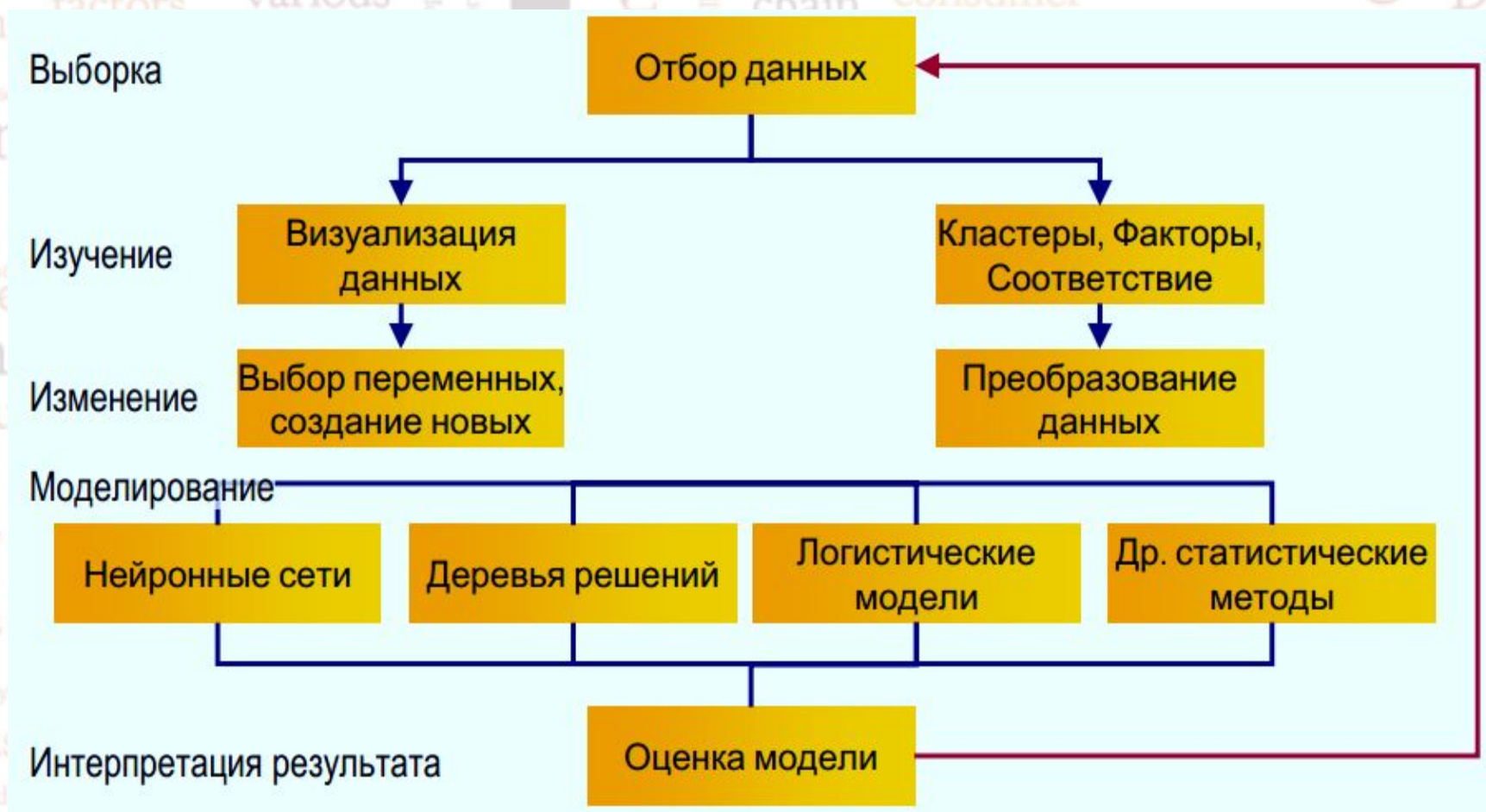
Запланировать развёртывание
Запланировать поддержку и мониторинг развернутого решения
Сделать финальный отчет
Сделать ревью проекта

Методология SEMMA

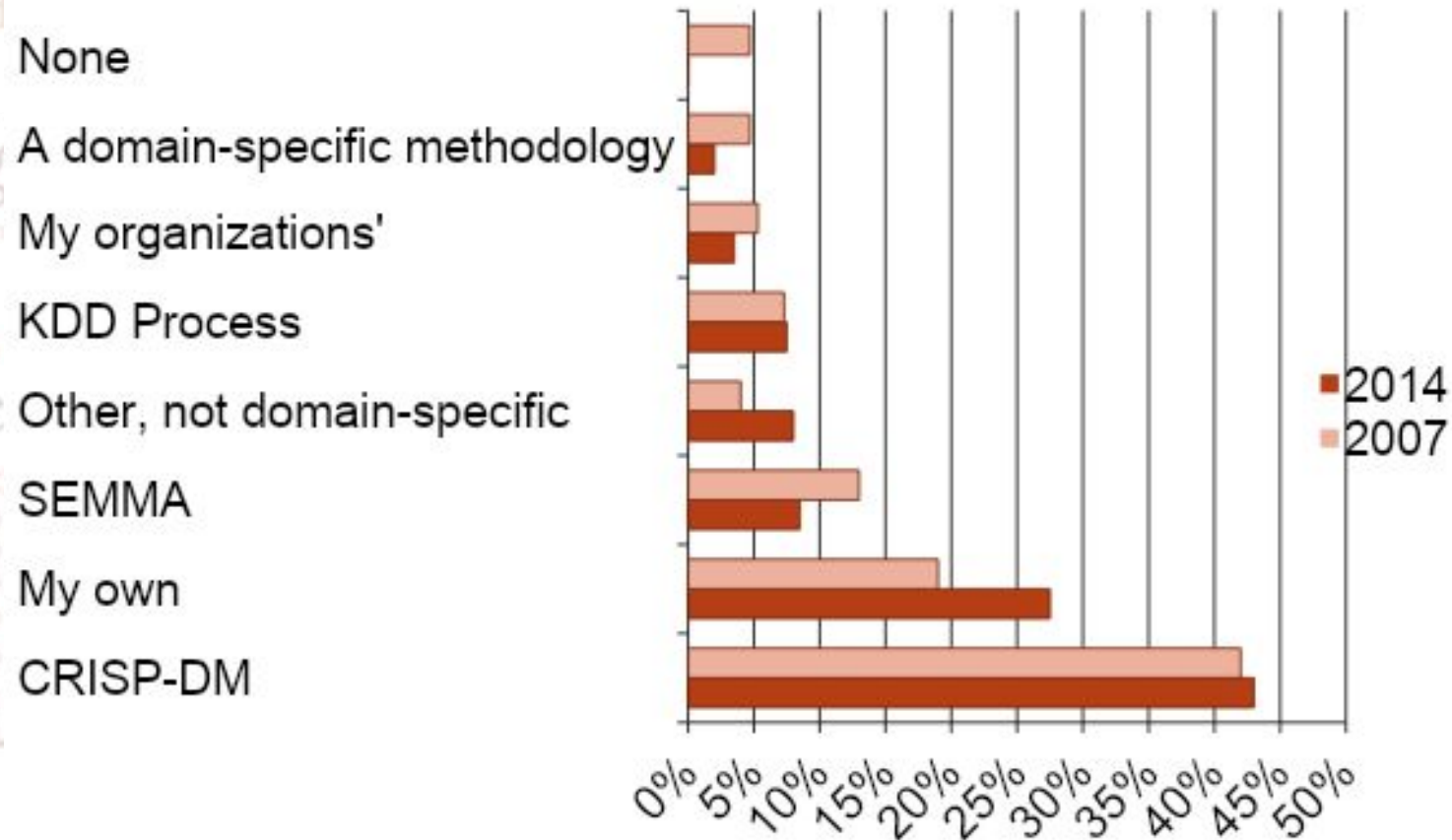
Методология SEMMA (аббревиатура, образованная от слов Sample, Explore, Modify, Model, Assess) заключается в поэтапном выполнении следующих процедур: выборки репрезентативных данных из общего массива, их исследовании, выявлении закономерностей и аномалий в данных, преобразовании и модификации данных (например, добавление новой информации или уменьшение количества анализируемых показателей), моделирование взаимосвязей между переменными (например, с помощью кластерного анализа, поиска ассоциаций, регрессии, нейронных сетей, деревьев решений и статистических методов), оценки полученных результатов моделирования.

Подход SEMMA подразумевает, что все процессы выполняются в рамках гибкой оболочки, поддерживающей выполнение всех необходимых работ по обработке и анализу данных. Благодаря диаграммам процессов обработки данных, подход SEMMA упрощает применение методов статистического исследования и визуализации, позволяет выбирать и преобразовывать наиболее значимые переменные, создавать модели с этими переменными, чтобы предсказать результаты, подтвердить точность модели и подготовить модель к развертыванию.

Этапы процесса анализа данных по методологии SEMMA



Использование различных методологий в анализе данных



<http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analyti-cs-data-mining-data-science-projects.html>

Типы задач анализа данных



Подготовка данных по CRISP-DM



Основные понятия

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

Значение переменной является проявлением признака

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

Генеральная совокупность - вся совокупность изучаемых объектов, интересующая исследователя

Параметры - числовые характеристики генеральной совокупности

Статистики - числовые характеристики выборки

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

Измерение - процесс присвоения чисел характеристикам изучаемых объектов согласно определенному правилу (шкале)

Шкалы измерений

Дихотомическая

$(\neq), (=)$

Номинальная

Порядковая

$(>), (<)$

Интервальная

$(+), (-)$

Относительная

$(\times), (\div)$

Примеры шкал измерений

Дихотомическая переменная

Пол ('Мужчины', 'Женщины')

Номинальная переменная

Город ('Москва', 'Санкт-Петербург', 'Казань')

Порядковая переменная

Доход ('Менее 15 тыс. руб.', 'От 15 до 25 тыс. руб.', 'Свыше 35 тыс. руб.')

Интервальная переменная

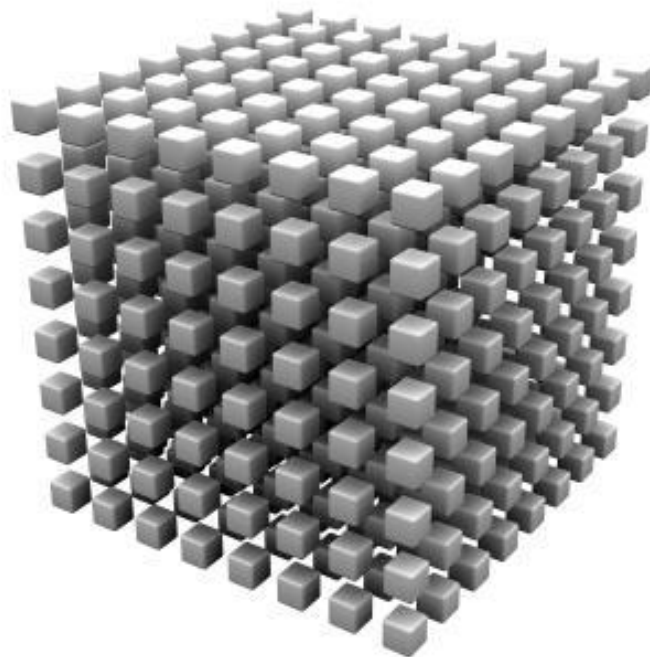
Баллы отношения к сервису компании (1,2,3,4,5)

Относительная переменная (количественная)

Возраст (18, 19, 20..., 65, ...)

Типовой вид исходных данных

ПАРАМЕТРЫ (АТТРИБУТЫ, СВОЙСТВА, ХАРАКТЕРИСТИКИ...)



ОБЪЕКТ
Ы

1
2
3

V1

V2

0,5

1,2

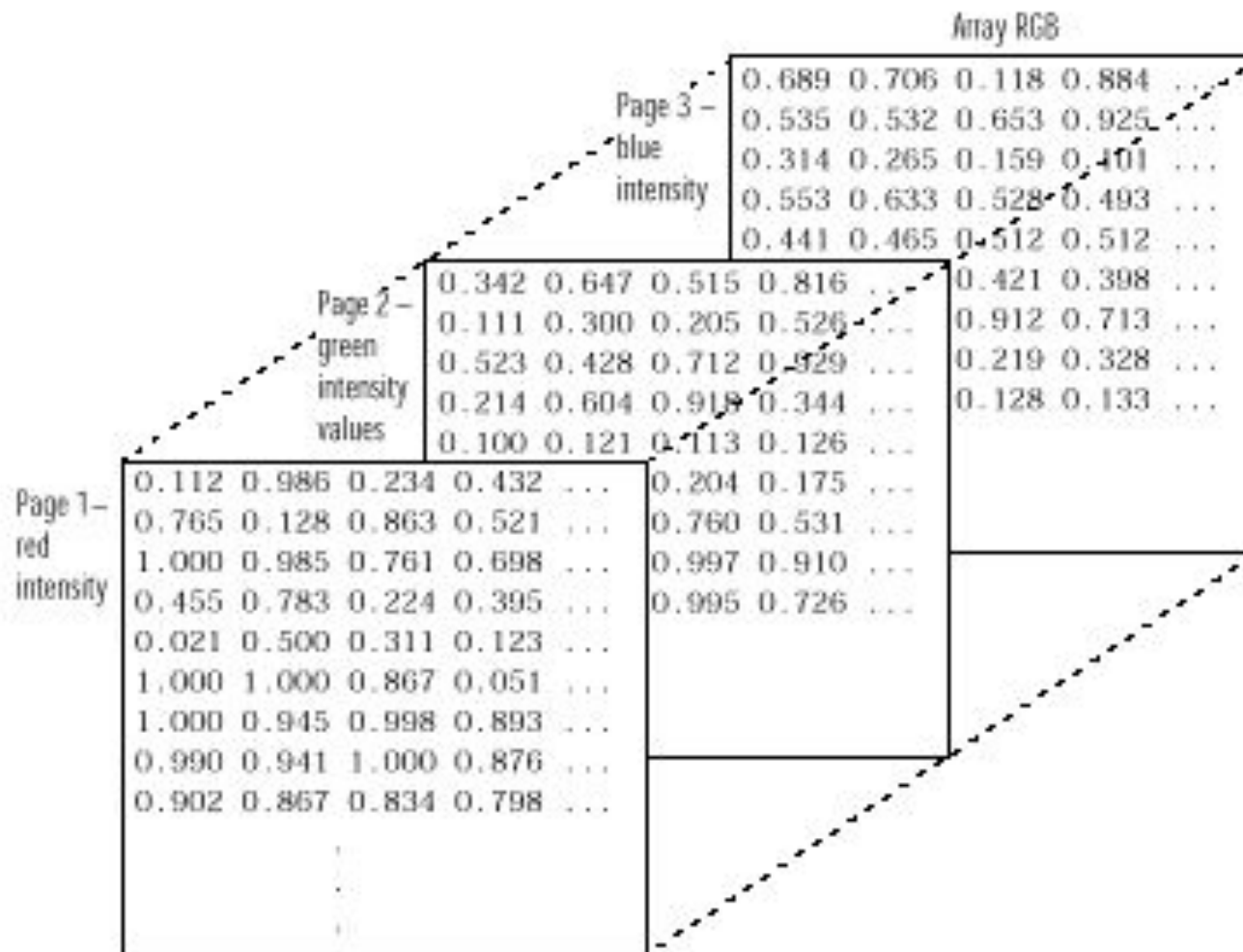
0,3

1,5

0,4

2,1

Представление изображений в формате RGB



Понятие очистки данных

Очистка данных – процедура корректировки данных, которые в каком-либо смысле не удовлетворяют определённым критериям качества, то есть содержат нарушения структуры данных, противоречия, пропуски, дубликаты, неправильные форматы и т.д.

Качество данных

Данные высокого качества

Данные содержащие критические ошибки

- невозможность загрузки в хранилище данных

Данные содержащие некритические ошибки

- аномальные значения
- пропуски
- дубликаты
- противоречия

Понятие обогащения данных

Обогащение данных — процесс насыщения данных новой информацией, которая позволяет сделать их более ценными и значимыми с точки зрения решения той или иной аналитической задачи.

Внешнее обогащение предполагает привлечение дополнительной информации из внешних источников.

Внутреннее обогащение предполагает повышение информативности и значимости данных за счёт изменения и реорганизации.

Восстановление пропущенных значений

Метод исключения некомплектных объектов

Методы с заполнением

Метод исключения некомплектных объектов

При отсутствии у ряда объектов значений каких-либо переменных некомплектные объекты удаляются из анализа. Подход легко реализуется и может быть удовлетворительным при малом числе пропусков. Однако иногда он приводит к серьезным смещениям и обычно не очень эффективен. Главный недостаток такого подхода обусловлен потерей информации при исключении неполных наблюдений.

Методы с заполнением

Заполнение средними.

Заполнение с пристрастным подбором

- Подстановка с подбором внутри группы
- Подбор ближайшего соседа

Заполнение с помощью регрессии

Методы взвешивания

Методы моделирования с помощью функции максимального правдоподобия

Понятие трансформации данных

Трансформация данных – комплекс методов и алгоритмов, направленных на оптимизацию представления и форматов данных с точки зрения решаемых задач и целей анализа. Трансформация данных не ставит целью изменить информационное содержание данных. Её задача представить эту информацию в таком виде, чтобы она могла быть использована наиболее эффективно.

Методы трансформации данных

Преобразование упорядоченных данных

Квантование

Сортировка

Слияние

Группировка и разгруппировка

Настройка набора данных

Табличная подстановка значений

Вычисляемые (производные) значения

Нормализация

Квантование

Квантование – процедура преобразования данных, состоящая из 2-х шагов. На первом шаге диапазон значений переменной разбивается на заданное число интервалов, каждому из которых присваивается некоторый номер (уровень квантования). На втором шаге каждое значение заменяется номером интервала квантования.

Квантование

Равномерное (однородное) квантование

Неравномерное (неоднородное) квантование

Равномерное квантование

Равномерное (однородное) квантование – преобразование, при котором диапазон значений переменной разбивается на интервалы одинаковой длины. Имеет смысл, если значения распределены равномерно по всему диапазону значений.

Гистограмма



Неравномерное квантование

Неравномерное (однородное) квантование – преобразование, при котором диапазон значений переменной разбивается на интервалы различной длины (асимметричные). Имеет смысл, если в значениях нет пропусков или сгустков.



Слияние

Внутреннее соединение

Внешнее соединение

Объединение

Полное внешнее соединение

Внутреннее соединение

Исходная таблица

Связываемая таблица

Внешнее соединение

Исходная таблица

Связываемая таблица

Исходная таблица

Связываемая таблица

Объединение

Исходная таблица

Связываемая таблица

Полное внешнее соединение

Исходная таблица

Связываемая таблица

Табличная подстановка значений (кодирование)

Преобразование в уникальные числовые коды

Двоичное кодирование

Кодирование с помощью дополнительной информации