Интеллектуальный анализ данных

Лекция 1

Введение в курс
Методы анализа данных
Data Mining

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
- **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.

Разведочный анализ данных (РАД, Exploratory data analysis (EDA))

- — 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 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**.

Добыча данных - Data Mining

Data Mining - исследование и обнаружение "машиной" (алгоритмами, средствами искусственного интеллекта) в сырых данных скрытых знаний, которые ранее не были известны, нетривиальны, практически полезны, доступны для интерпретации человеком.

- Знания должны быть новые, ранее неизвестные.
- Знания должны быть нетривиальны.
- Знания должны быть практически полезны.
- Знания должны быть доступны для понимания человеку.

Термин введён Григорием Пятецким-Шапиро в 1989 году

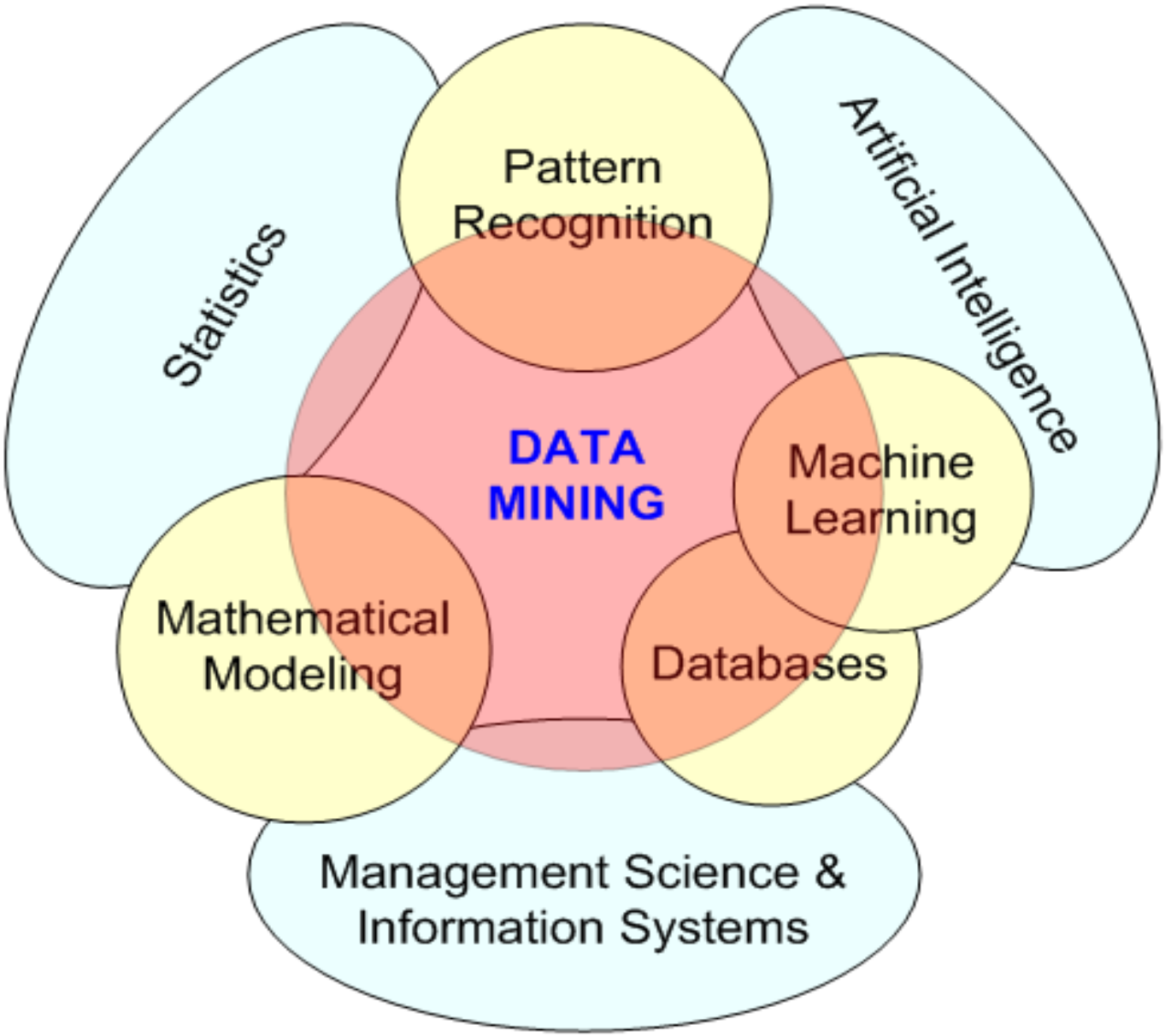
Важное отличие процедуры добычи данных от классического РАД: системы добычи данных в большей степени ориентированы на практическое приложение полученных результатов, чем на выяснение природы явления.

Knowledge Discovery in Database (KDD)

- Извлечение знаний из баз данных
- Описывает последовательность действий, которую необходимо выполнить для обнаружения полезного знания
- Не зависит от предметной области

Григорий Пятецкий-Шапиро
президент и главный редактор одного из первых сайтов (1994 г.) по Анализу данных «KDnuggets» http://www.kdnuggets.com/

Связь с другими дисциплинами



Классификация уровней информации

Уровень информации	Описание
Сырые данные (raw data)	Необработанные данные, получаемые в результате наблюдения за объектами и отображающие их состояние в конкретные моменты времени (например, данные о котировках акций за прошедший год, данные о ценах на рынке жилья, данные об абитуриентах, зачисленных на 1 курс)
Информация	Это либо: - сырые данные, но систематизированные, представленные в более компактном виде (например, результаты поиска – сведения об абитуриентах, поступивших в ИИТиУТС СевГУ в этом году); - обработанные данные, имеющие информационную ценность для пользователя (например, сводные статистические характеристики – средний балл абитуриентов, поступивших в ИИТиУТС СевГУ в этом году – его абсолютная величина и % по отношению к тому же показателю за предыдущий год)
Знания	Понятие «знания» включает: - скрытые взаимосвязи межу объектами (признаками объектов); - некоторое ноу-хау, алгоритмы, методы решения задач. Знания обладают практической ценностью.

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.

Этапы проведения интеллектуального анализа данных

- постановка задачи;
- подготовка данных;
- изучение данных;
- построение моделей;
- исследование и проверка моделей;
- развертывание и обновление моделей.


Источники данных
Выборка данных
Исходящие данные
Очистка
Очищенные данные
Трансформация
Трансформированные данные
Data Mining
Шаблоны
Интерпретация
Знания

Особенности проведения ИАД

- В ходе проведения интеллектуального анализа данных проводится исследование *множества* объектов (или вариантов).
- В большинстве случаев его можно представить в виде таблицы, каждая строка которой соответствует одному из вариантов, а в столбцах содержатся значения параметров, его характеризующих.
- Зависимая *переменная* - *параметр*, *значение* которого рассматриваем как зависящее от других параметров (независимых переменных).
- Именно эту зависимость и необходимо определить, используя методы интеллектуального анализа данных.

Задачи Data Mining

- Прогнозирование (Forecasting)
- Классификация (Classification)
- Кластеризация (Clustering)
- Ассоциации (Associations)
- Визуализация (Data Visualization)
- Обобщение (Summarization): Обнаружение отклонений; Оценка; Анализ/поиск связей.

Задачи Data Mining

Задачи Data Mining

- **Задача классификации** сводится к определению класса объекта по его характеристикам. Множество классов известно заранее.
- **Задача регрессии** подобно задаче классификации позволяет определить по известным характеристикам объекта значение некоторого параметра из множества действительных чисел.
- При **поиске ассоциативных правил** целью является нахождение частых зависимостей (или ассоциаций)
- **Задача кластеризации** заключается в поиске независимых групп (кластеров) и их характеристик во всем множестве анализируемых данных.

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.

Supervised и unsupervised learnig

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.

К этому типу задач относят задачи классификации и регрессии.

Unsupervised learning - обучение без учителя –

объединяет задачи, выявляющие описательные модели.

Достоинство таких задач - возможность их решения без каких либо предварительных знаний. об анализируемых данных. К этим задачам относятся кластеризация и поиск ассоциативных правил.

Задача классификации и регрессии

Постановка: требуется определить, к какому из известных классов относятся исследуемые объекты, т. е. классифицировать их.

Клиент банка: «кредитоспособен» и «некредитоспособен».

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

Распознавание цифр: от 0 до 9.

В Data Mining задачу классификации рассматривают как задачу определения значения одного из параметров анализируемого объекта на основании значений других параметров.

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.

Задача классификации и регрессии

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

Для получения максимально точной функции к обучающей выборке предъявляются следующие основные требования:

- количество объектов, входящих в выборку, должно быть достаточно большим. Чем больше объектов, тем точнее будет построенная на ее основе функция классификации или регрессии;
- в выборку должны входить объекты, представляющие все возможные классы в случае задачи классификации или всю область значений в случае задачи регрессии;
- для каждого класса в задаче классификации или для каждого интервала области значений в задаче регрессии выборка должна содержать достаточное количество объектов.

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

Задача поиска ассоциативных правил

Первоначально она решалась при анализе тенденций в поведении покупателей в супермаркетах (анализ рыночных корзин - Basket Analysis). При анализе этих данных интерес прежде всего представляет информация о том, какие товары покупаются вместе, в какой последовательности, какие категории потребителей какие товары предпочитают, в какие периоды времени и т. п.

В сфере обслуживания интерес представляет информация о том, какими услугами клиенты предпочитают пользоваться в совокупности.

В медицине - анализ сочетания симптомов и болезней.

Сиквенциальный анализ учитывает последовательность происходящих событий (телекоммуникационные компании, анализ аварий).

Задача кластеризации

Задача кластеризации состоит в разделении исследуемого множества объектов на группы "похожих" объектов, называемых кластерами (cluster).

Периодическая система элементов Д.И. Менделеева.

Сегментация в маркетинге. Критериями сегментации являются: географическое местоположение, социально-демографические характеристики, мотивы совершения покупки и т. п.

На основании результатов сегментации маркетолог может определить, например, такие характеристики сегментов рынка, как реальная и потенциальная емкость сегмента, группы потребителей, чьи потребности не удовлетворяются в полной мере ни одним производителем, работающим на данном сегменте рынка, и т. п.

Алгоритм интеллектуального анализа данных

- **Алгоритм интеллектуального анализа данных** представляет собой механизм, создающий модель интеллектуального анализа данных.
- Чтобы создать модель, *алгоритм* сначала анализирует набор данных, осуществляя *поиск* определенных закономерностей и трендов.
- *Алгоритм* использует результаты этого анализа для определения параметров модели интеллектуального анализа данных.
- Затем эти параметры применяются ко всему набору данных, чтобы выявить пригодные к использованию закономерности и получить подробную статистику.

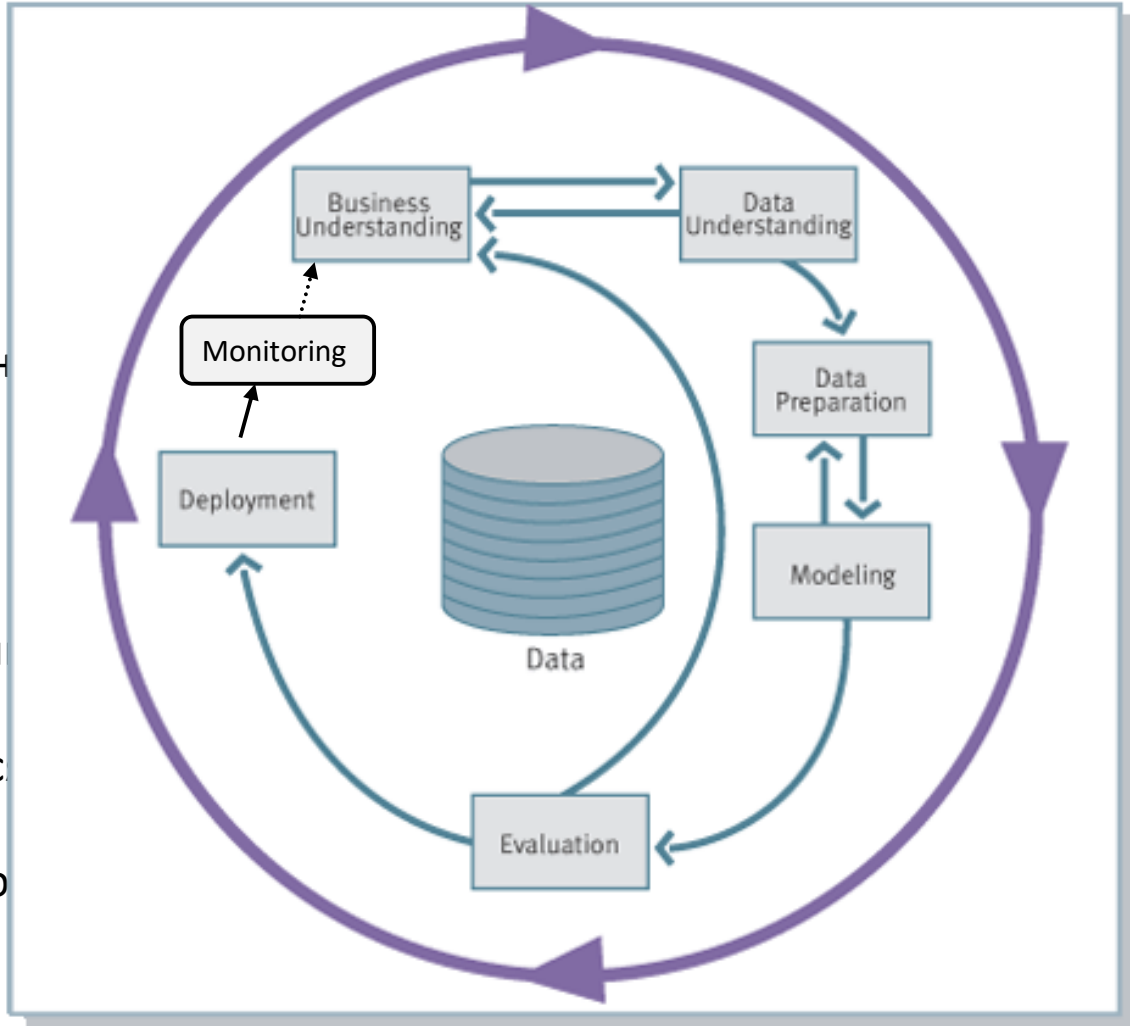
Примеры использования алгоритмов интеллектуального анализа

Задача и пример	Подходящие алгоритмы
Прогнозирование дискретного атрибута. Например, купит ли получатель целевой рассылки определенный продукт.	Алгоритм дерева принятия решений. Упрощенный алгоритм Байеса. Алгоритм кластеризации. Алгоритм нейронной сети.
Прогнозирование непрерывного атрибута. Например, прогноз продаж на следующий год.	Алгоритм дерева принятия решений. Алгоритм временных рядов.
Прогнозирование последовательности. Например, анализ маршрута перемещения по веб-узлу компании.	Алгоритм кластеризации последовательностей
Нахождение групп общих элементов в транзакциях. Например, использование анализа покупательского поведения для предложения дополнительных продуктов заказчику.	Алгоритм взаимосвязей Алгоритм дерева принятия решений
Нахождение групп схожих элементов. Например, разбиение демографических данных на группы для лучшего понимания связей между атрибутами.	Алгоритм кластеризации. Алгоритм кластеризации последовательностей.

Главное задание Data Mining:

- найти истинные закономерности и избежать *переобучения*

Переобучение (*overfitting*) в машинном обучении и статистике - явление, когда построенная модель хорошо объясняет примеры из обучающей выборки, но относительн плохо работает на примерах, не участвовавших в обучении (на примерах из тестовой выборки).
Это связано с тем, что при построени модели («в процессе обучения») в обучающей выборке обнаруживаютс некоторые случайные закономерности, которые отсутствую в генеральной совокупности.



Практическое применение Data Mining

Интернет-технологии

- персонализация посетителей Web-сайтов
- поиск случаев мошенничества с кредитными картами
- Web Mining: Web content mining и Web usage mining

Торговля

- анализ рыночных корзин и сиквенциональный анализ

Телекоммуникации

- анализ доходности и риска потери клиентов
- защита от мошенничества,
- выявление категорий клиентов с похожими стереотипами пользования услугами и разработка привлекательных наборов цен и услуг

Практическое применение Data Mining

Промышленное производство

- прогнозирование качества изделия в зависимости от замеряемых параметров технологического процесса.

Медицина и биология

- построение диагностической системы
- исследование эффективности хирургического вмешательства
- Биоинформатика – изучение генов, разработка новых лекарств

Банковское дело

- оценка кредитоспособности заемщика

Примеры применения интеллектуального анализа данных

	Информационные технологии	Торговля	Финансовая сфера
Классификация			Оценка кредитоспособности
Регрессия			Оценка допустимого кредитного лимита
Прогнозирование		Прогнозирование продаж	Прогнозирование цен акции
Кластеризации		Сегментация клиентов	Сегментация клиентов
Определения взаимосвязей		Анализ потребительской корзины	
Анализ последовательностей	Анализ переходов по страницам web-сайта		
Анализ отклонений	Обнаружение вторжений в информационные системы		Выявление мошенничества с банковскими картами

Наиболее известные алгоритмы интеллектуального анализа данных:

- алгоритм линейной регрессии - LinearRegression;
- алгоритм логистической регрессии - LogisticRegression.
- алгоритм *дерева принятия решений* - DecisionTrees;
- упрощенный алгоритм Байеса - NaiveBayes;
- алгоритм временных рядов - TimeSeries;
- алгоритм кластеризации - Clustering;
- алгоритм кластеризации последовательностей - SequenceClustering;
- алгоритм взаимосвязей - AssociationRules;
- нейронные сети -NeuralNetwork;

Модели Data Mining

Предсказательные модели

- модели классификации
- модели последовательностей

Описательные модели

- регрессионные модели
- модели кластеров
- модели исключений
- итоговые модели
- ассоциативные модели

Предсказательные модели

модели классификации описывают правила или набор правил, в соответствии с которыми можно отнести описание любого нового объекта к одному из классов

Такие правила строятся на основании информации о существующих объектах путем разбиения их на классы;

модели последовательностей описывают функции, позволяющие прогнозировать изменение непрерывных числовых параметров.

Они строятся на основании данных об изменении некоторого параметра за прошедший период времени.

Описательные модели

регрессионные модели описывают функциональные зависимости между зависимыми и независимыми показателями и переменными в понятной человеку форме.

описывают функциональную зависимость не только между непрерывными числовыми параметрами, но и между категориальными параметрами;

модели кластеров описывают группы (кластеры), на которые можно разделить объекты, данные о которых подвергаются анализу. Группируются объекты (наблюдения, события) на основе данных (свойств), описывающих сущность объектов.

объекты внутри кластера должны быть "похожими" друг на друга и отличаться от объектов, вошедших в другие кластеры.

Чем сильнее "похожи" объекты внутри кластера и чем больше отличий между кластерами, тем точнее кластеризация;

Описательные модели

Модели исключений описывают исключительные ситуации в записях (например, отдельных пациентов), которые резко отличаются чем либо от основного множества записей (группы больных).

Знание исключений может быть использовано двояким образом. Возможно, эти записи представляют собой случайный сбой, например ошибки операторов, вводивших данные в компьютер.

С другой стороны, отдельные исключительные записи могут представлять самостоятельный интерес для исследования, т. к. они могут указывать на некоторые редкие, но важные аномальные заболевания.

Описательные модели

Итоговые модели - выявление ограничений на данные анализируемого массива.

Например, при изучении выборки данных по пациентам не старше 30 лет, перенесшим инфаркт миокарда, обнаруживается, что все пациенты, описанные в этой выборке, либо курят более 5 пачек сигарет в день, либо имеют вес не ниже 95 кг

Построение итоговых моделей заключается в нахождении каких либо фактов, которые верны для всех или почти всех записей в изучаемой выборке данных, но которые достаточно редко встречались бы во всем мыслимом многообразии записей;

ассоциативные модели - выявление закономерностей между связанными событиями.

Методы исследований. Обработка и анализ данных

Количественный анализ:

это манипуляции с измеренными характеристиками изучаемого объекта

- направлен в основном на формализованное, «внешнее» изучение объекта, анализ его измеряемых признаков;
- основным итогом является упорядоченная совокупность «внешних», измеряемых признаков объекта;
- реализуется при помощи математико-статистических методов;
- доминирует аналитическая составляющая познания.

Качественный анализ

способ проникновения в сущность объекта путем выявления таких его свойств, измерить которые невозможно количественными методами, можно описать различными способами

- **доминирует** синтетическая составляющая познания;
- **направлен** на содержательное, внутреннее изучение объекта;
- **результатом** качественного анализа является интегрированное представление о множестве свойств объекта в форме классификаций и типологий.

Фазы процесса количественной обработки данных

1. Первичная обработка данных:
   - Табулирование,
   - Построение диаграмм,
   - Построение гистограмм,
   - Полигонов распределений,
   - Кривых распределений.

Нацелена на упорядочение информации об объекте и предмете изучения, полученной на полевом этапе исследования.

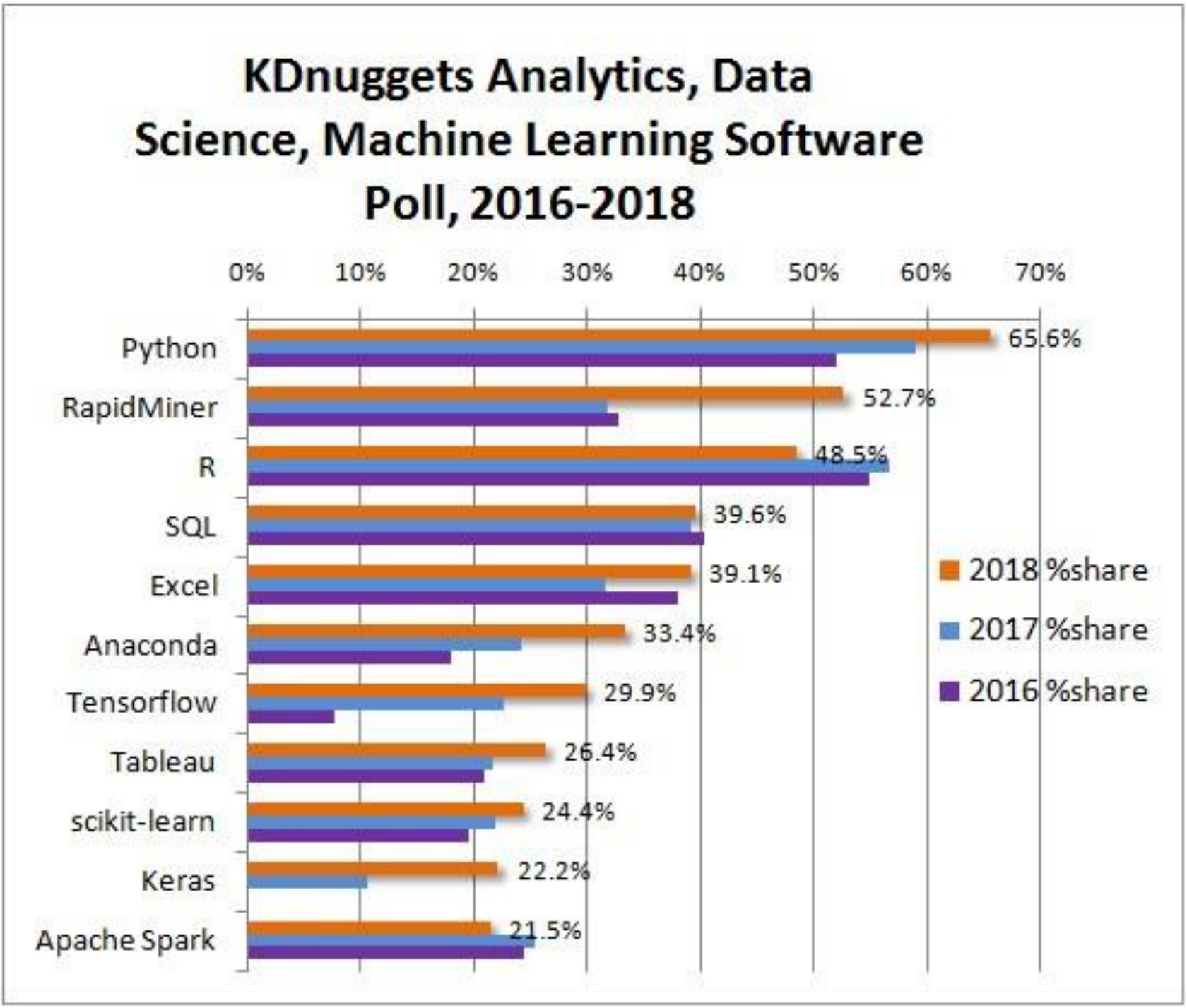
2. Вторичная обработка данных:
   - описательная статистика,
   - индуктивная статистика,
   - корреляционная статистика.

Вторичная обработка данных

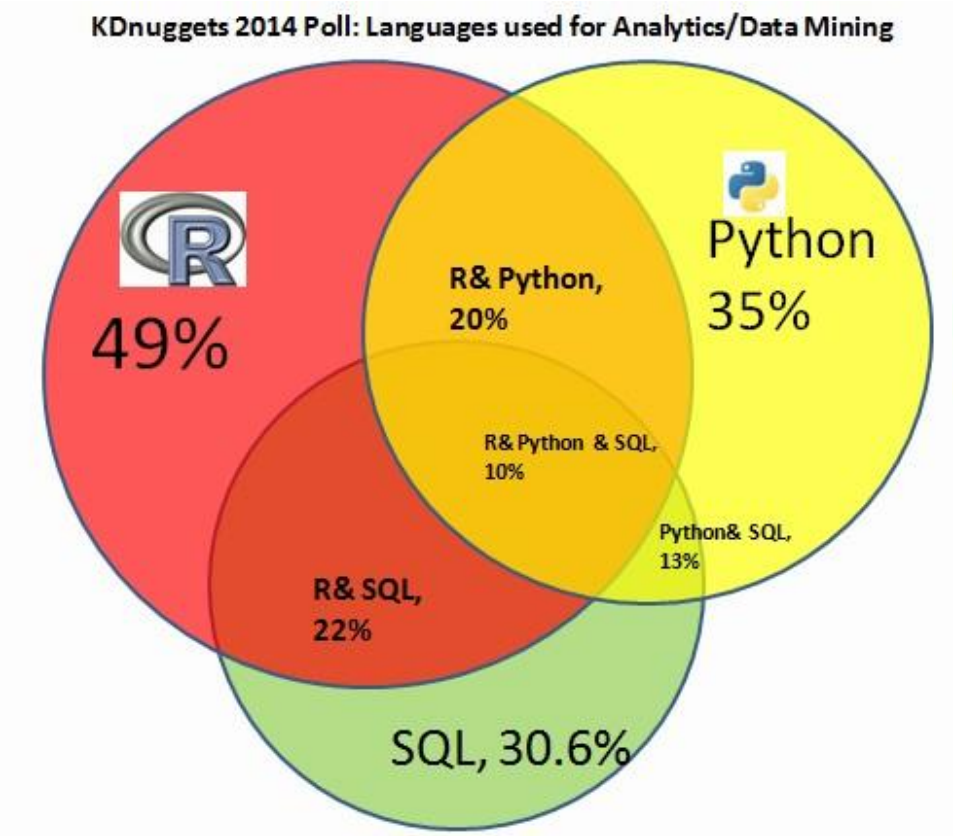
- Состоит в ***статистическом анализе*** итогов первичной обработки.
- ***Описательная статистика*** – позволяет охарактеризовать основные значения переменных.
- ***Индуктивная статистика -*** осуществляет проверку соответствия данных выборки всей популяции (проверка гипотез, метод Стьюдента, метод Хи-квадрат и др.)
- ***Корреляционная статистика –*** выявляет связи между явлениями (коэффициент Пирсона)

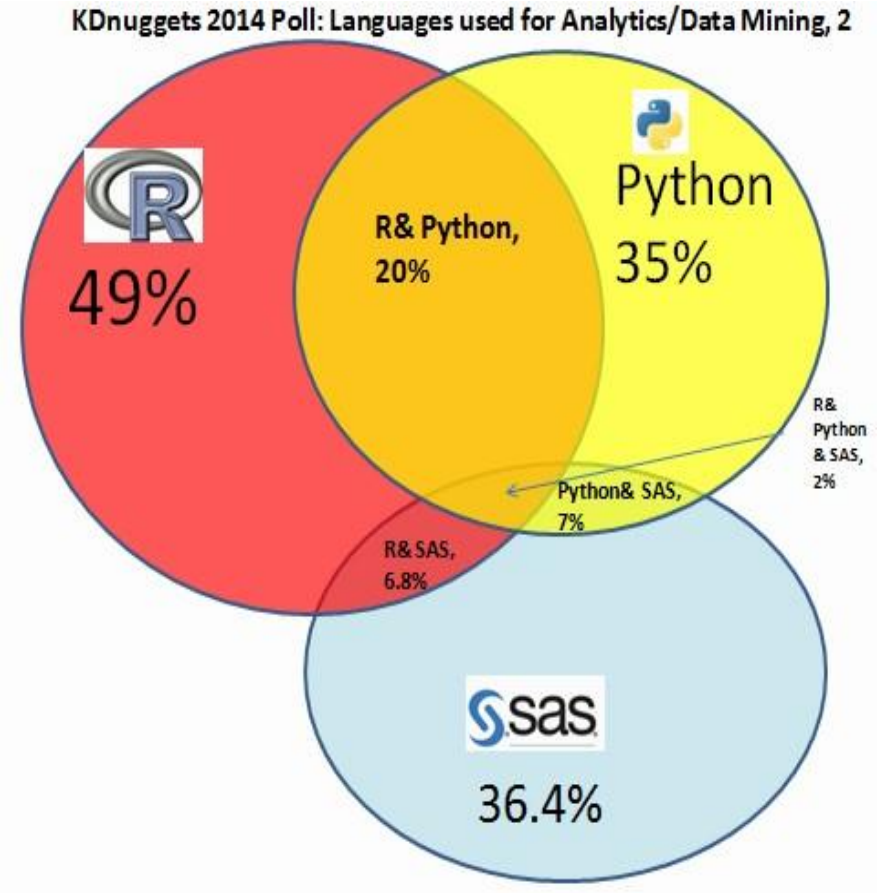
Приемы качественного анализа данных

- Кейс-стади (case-study) – исследование конкретной общности (случая).
- Этнографические исследования.
- Исторические исследования.
- Фокус-группа.
- Интервью (нарративное, полуструктурированное, биографическое, лейтмотивное, фокусированное)


KDnuggets Analytics, Data Science, Machine Learning Software Poll, 2016-2018
0%
10%
20%
30%
40%
50%
60%
70%
Python
RapidMiner
R
SQL
Excel
Anaconda
Tensorflow
Tableau
scikit-learn
Keras
Apache Spark
65.6%
52.7%
48.5%
39.6%
39.1%
33.4%
29.9%
26.4%
24.4%
22.2%
21.5%
2018 %share
2017 %share
2016 %share

Консолидация среди топ-4 языков





Основные понятия и сведения из истории создания пакета R

- Термин **R** используется в двух значениях:
- (интерпретируемый) язык программирования для статистической обработки данных
- программная среда вычислений.
- Год разработки **1993**.

Разработчики - сотрудники Оклендского университета (Новая Зеландия)

- Росс Айхэка (Ross Ihaka)
- Роберт Джентлмен (Robert Gentleman).
- Название языка и среды вычислений - первая буква имён разработчиков.
- Язык и среда **R** широко используются как статистическое программное обеспечение для анализа данных - *фактический стандарт программного обеспечения для статистической обработки информации*.
- В 2010 году **R** вошёл в список победителей конкурса журнала **InfoWorld** в номинации на *лучшее открытое программное обеспечение для разработки приложений*.

- **Полезные ссылки:**
- 1. «KDnuggets»: **https://www.kdnuggets.com/**
- 2. Ссылки для скачивания дистрибутива (R):

The Comprehensive R Archive Network - https://cran.r-project.org/
The R Project for Statistical Computing - https://www.r-project.org/

RStudio - https://www.rstudio.com/products/rstudio/download/#download

3. Роберт И. Кабаков. R в действии. Анализ и визуализация данных в программе R / пер. с англ. Полины А. Волковой. – М.: ДМК Пресс, 2014. – 588 с. - http://kek.ksu.ru/eos/WM/Kabacoff2014ru.pdf

4. DATA MINING FOR BUSINESS ANALYTICS. Concepts, Techniques, and Applications in R (Galit Shmueli et. al)
https://edu.kpfu.ru/pluginfile.php/278552/mod_resource/content/1/MachineLearningR__Brett_Lantz.pdf

5. Data Science Central - https://www.datasciencecentral.com/