ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ
МАТЕРИАЛЫ К ЛЕКЦИИ 1

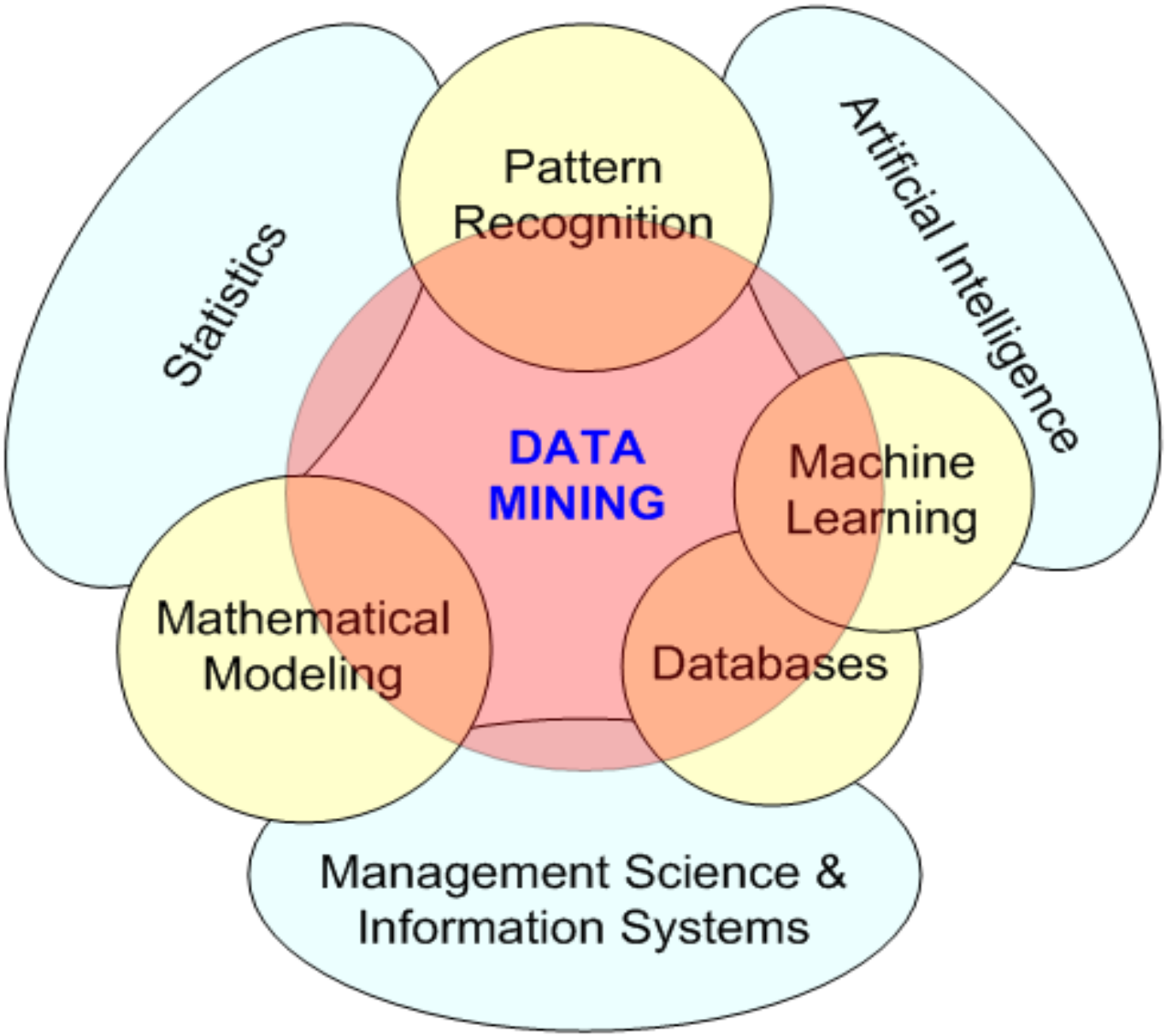
Введение в курс
Методы анализа данных
Data Mining

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/

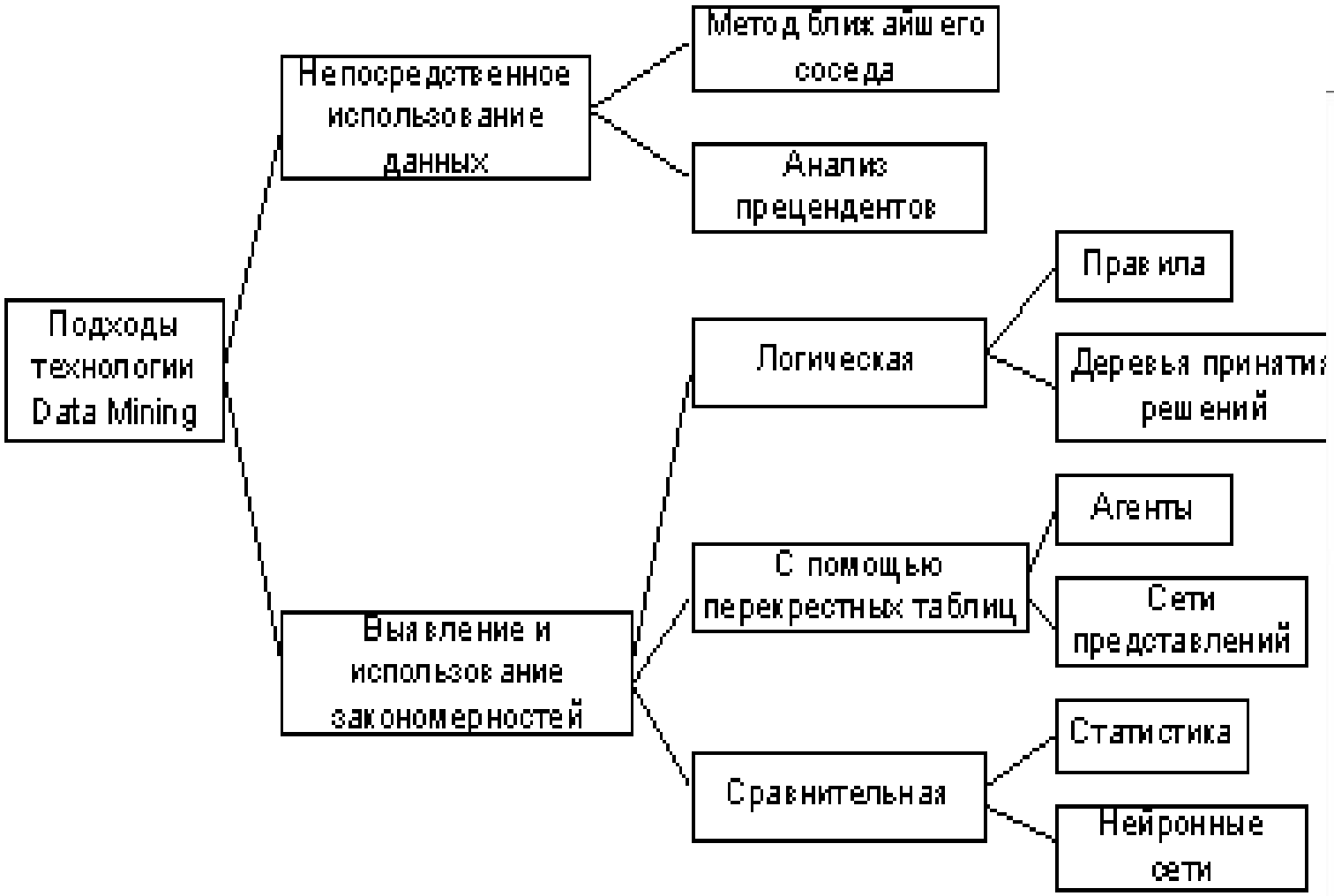
Связь с другими дисциплинами



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.

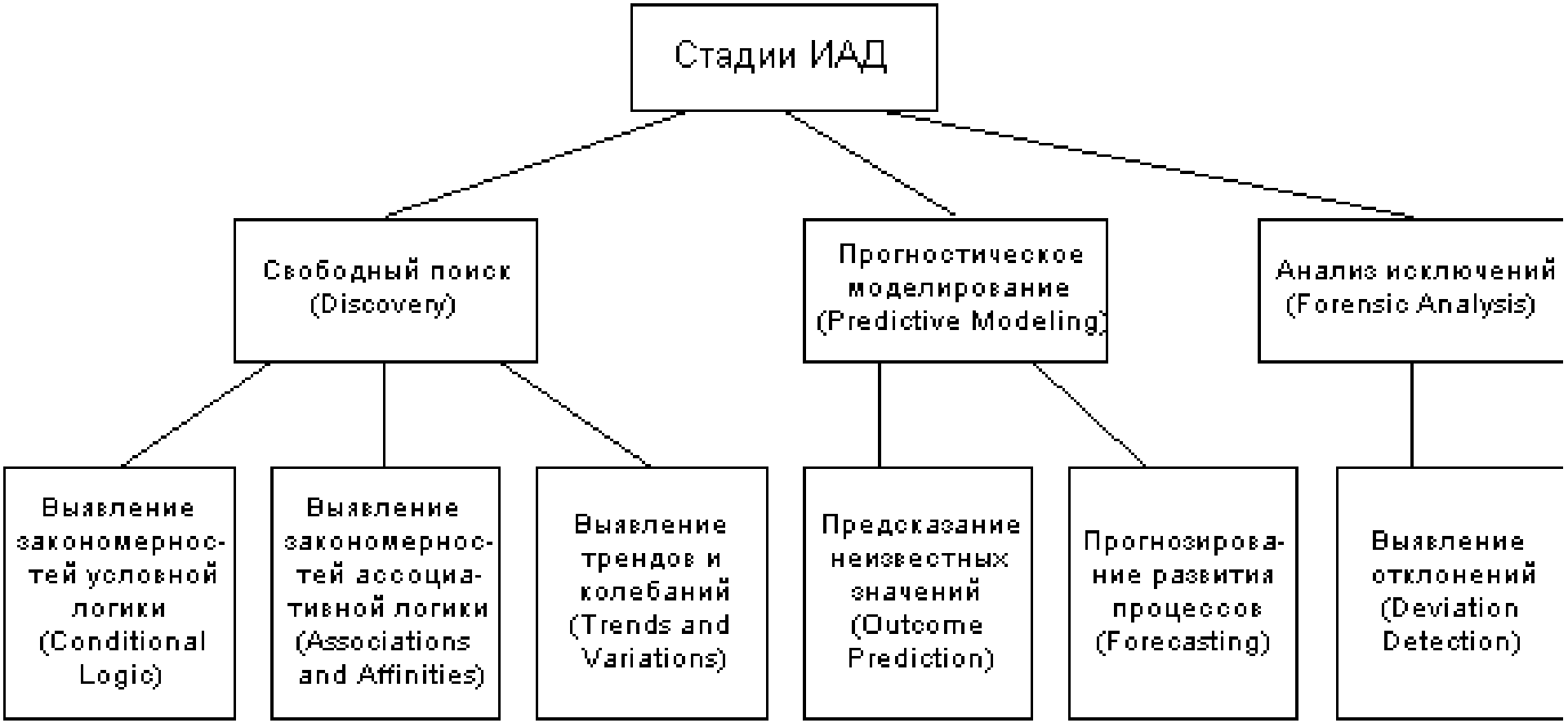
Подходы технологии ИАД



Математическая Статистика:

1. Оценивание параметров Распределения
2. Дисперсионный Анализ
3. Корреляционно-Регрессионный Анализ
4. Анализ временных Рядов
5. Многомерный Анализ:
   - кластерный
   - дискриминантный
   - факторный
   - мет. главных компонент и др.

Стадии интеллектуального анализа данных




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

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

Задача классификации сводится к определению класса объекта по его характеристикам. Множество классов известно заранее.

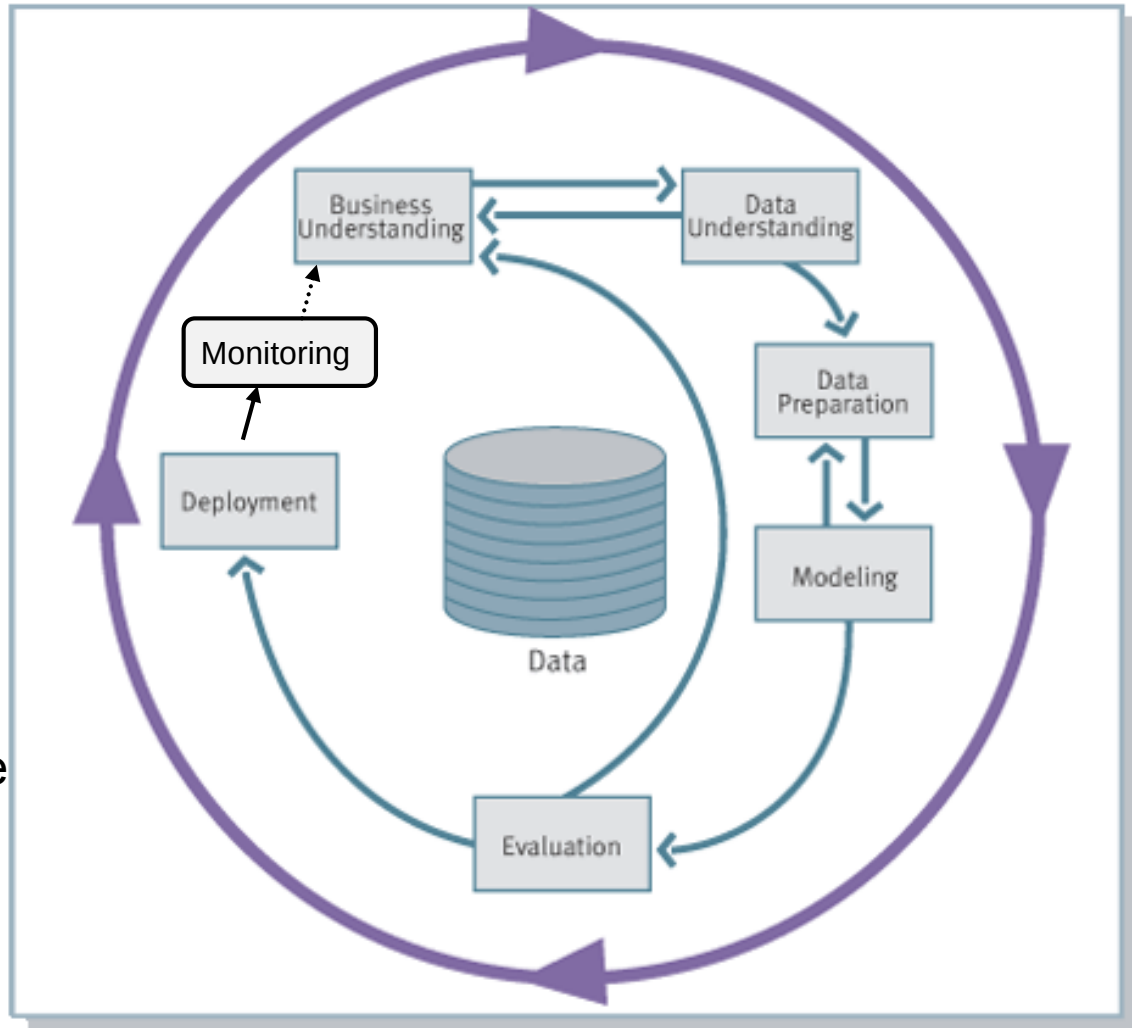
Задача регрессии подобно задаче классификации позволяет определить по известным характеристикам объекта значение некоторого параметра из множества действительных чисел.

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.

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 относительно плохо работает на примерах, не участвовавших в обучении (на примерах из тестовой выборки).
Это связано с тем, что при построении модели («в процессе обучения») в обучающей выборке обнаруживаются некоторые случайные закономерности, которые отсутствуют 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

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

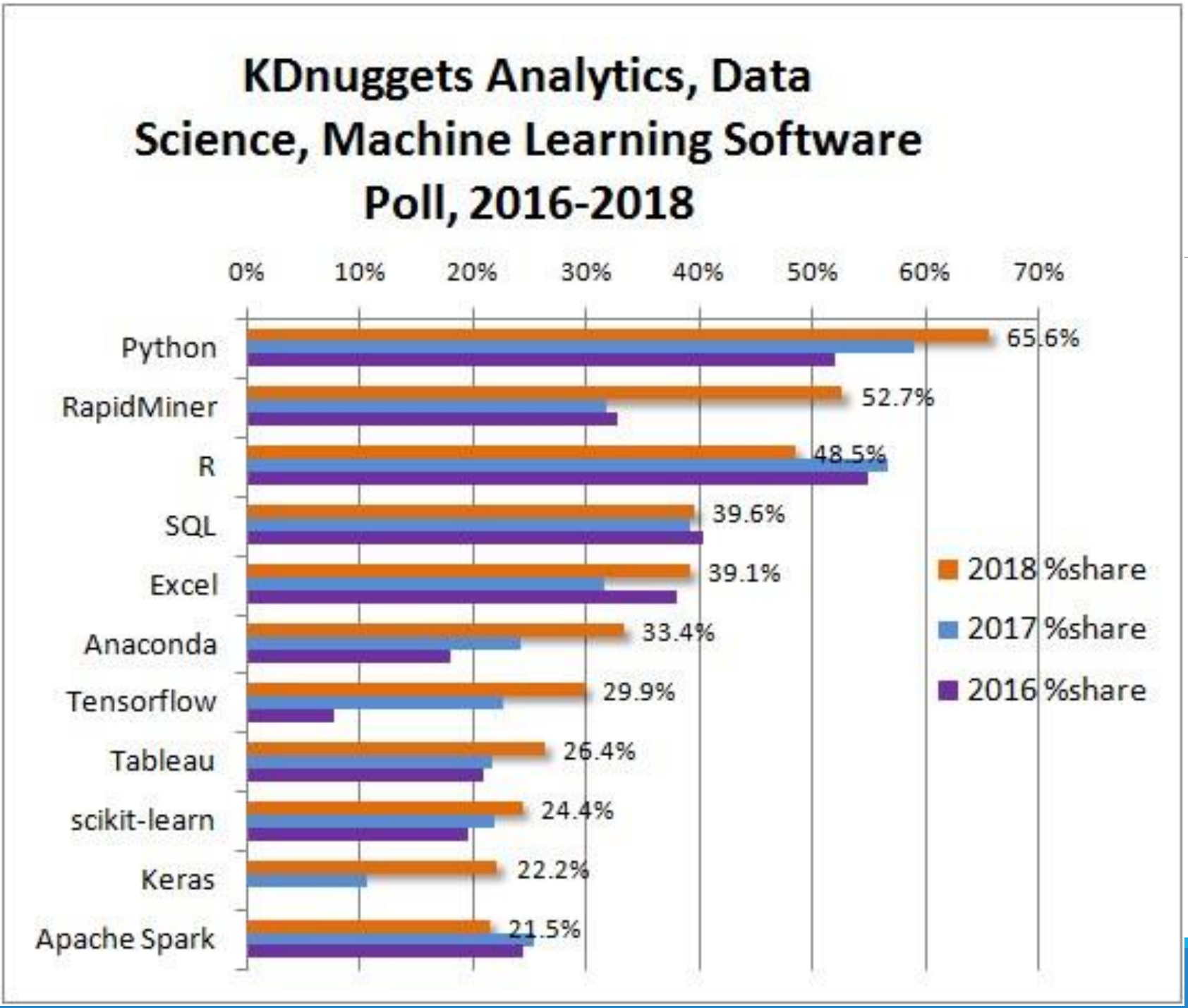
Итоговые модели - выявление ограничений на данные анализируемого массива.

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

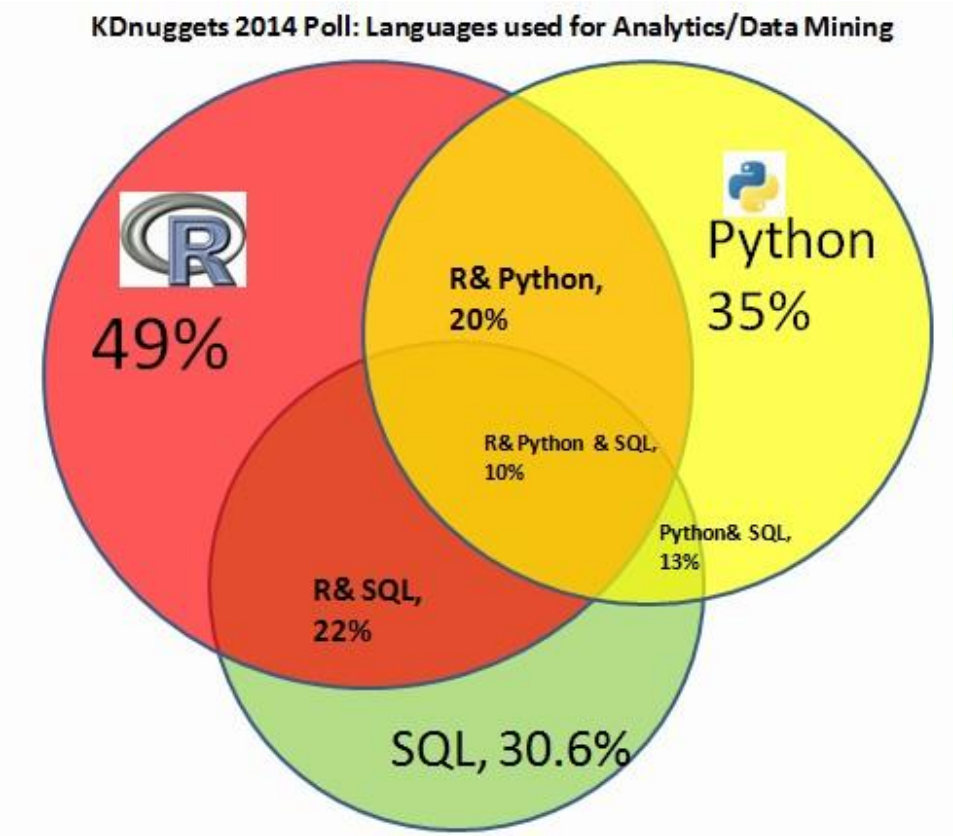
Построение итоговых моделей заключается в нахождении каких либо фактов, которые верны для всех или почти всех записей в изучаемой выборке данных, но которые достаточно редко встречались бы во всем мыслимом многообразии записей;

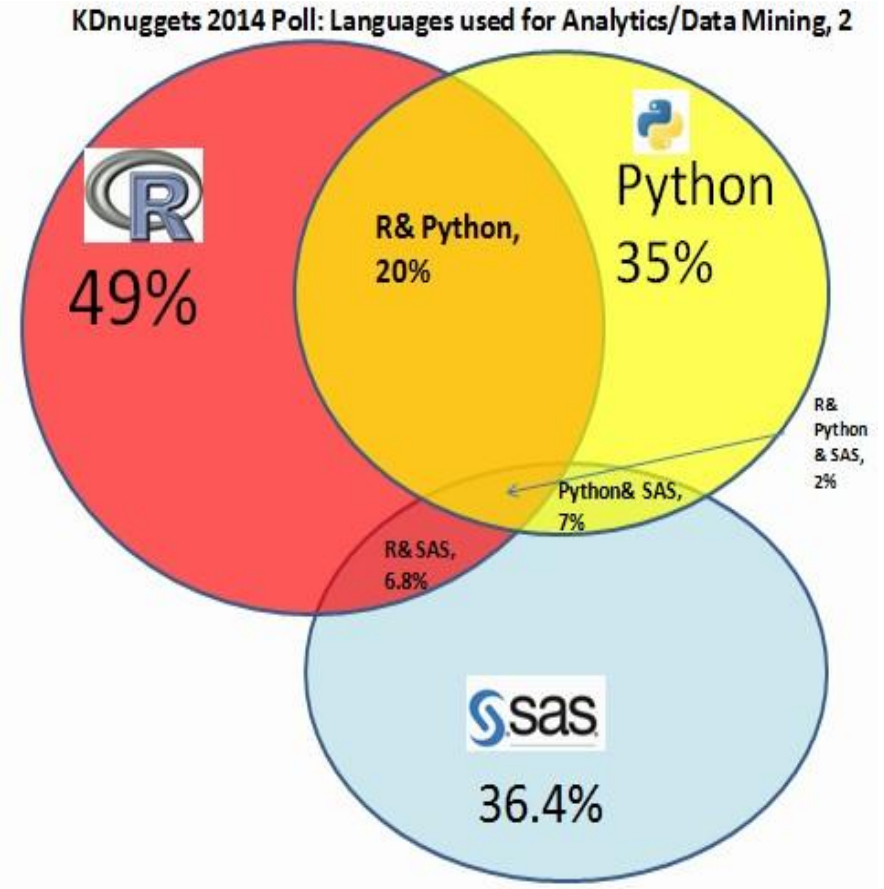
ассоциативные модели - выявление закономерностей между связанными событиями.

Методы исследований. Обработка и анализ данных


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/