Прикладной анализ данных

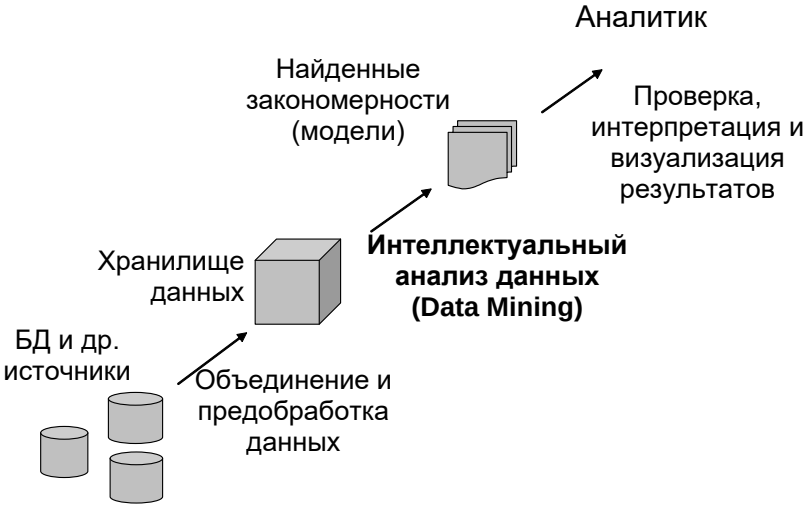
СОДЕРЖАНИЕ КУРСА

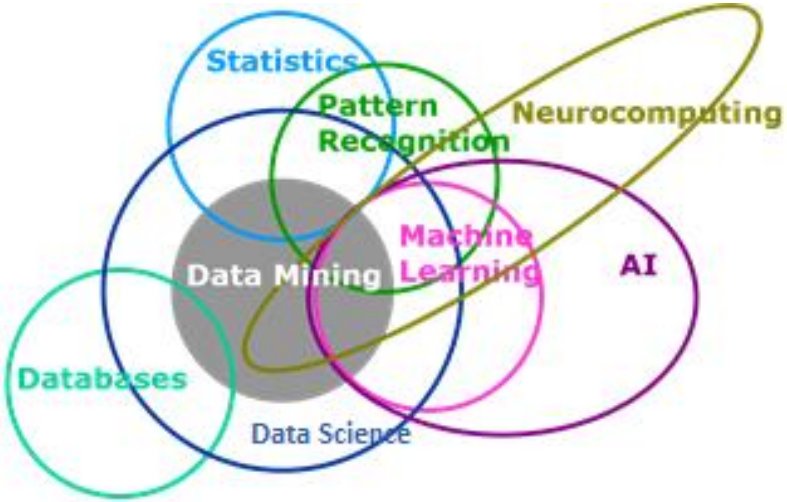
1. Введение, задачи анализа данных, обзор архитектуры SAS Viya
2. Методы интеллектуального анализа данных на платформе SAS Viya
   - разведочный анализ, выявление скрытых закономерностей на основе «обучения без учителя» (1 лекция)
   - построение, оценка и применение моделей прогнозирования, регрессии (1 лекция)
   - методы на основе деревьев решений и их бустинг и бэгинг ансамблей (2 лекции)
   - автотьюнинг гиперпараметров и моделенезависимая визуализация (1 лекция)
   - машины опорных векторов, введение в нейросети (1 лекция)
   - глубокое обучение, CNN, RNN (2 лекции).
3. Инструменты и методы анализа временных рядов (3 лекции)

ЛЕКЦИИ и ПРАКТИЧЕСКИЕ ЗАДАНИЯ!!!

Итог = практические задания + посещаемость + экзамен

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ (DATA MINING)





Системы *интеллектуального анализа данных* (ИАД) – класс программных систем поддержки принятия решений, задачей которых является поиск *скрытых, ранее неизвестных, содержательных и потенциально полезных* закономерностей в *больших объемах разнородных, сложно структурированных* данных.

Han J., Kamber M. Data Mining: Concepts and Techniques // Morgan Kaufmann, 2000

ЗАДАЧИ ИАД = ТИПЫ ВЫЯВЛЯЕМЫХ ЗАКОНОМЕРНОСТЕЙ

- Классификация («Обучение с учителем»)
  - Отнесение объектов к заранее определенным категориям
- Ранжирование («Обучение с учителем»)
  - Оценка степени соответствия объектов одной или более заранее определенным категориям
- Прогнозирование («Обучение с учителем»)
  - На основании известных значений атрибутов анализируемого объекта определяются значения неизвестных атрибутов
- Ассоциации («Обучение без учителя»)
  - Выявление зависимостей между атрибутами в виде правил или аналитических зависимостей, выявление скрытых свойств объектов
- Кластеризация («Обучение без учителя»)
  - Выделение компактных подгрупп «похожих» объектов
- Выявление исключений («Обучение с учителем и без»)
  - Поиск объектов, которые своими характеристиками значительно отличаются от остальных

ПРОЦЕСС ИАД (1)

- Анализ предметной области:
  - выявление и формулировка необходимых априорных знаний о предметной области, целей анализа, задач приложения, сценариев использования
- Формирование и подготовка данных для анализа:
  - поиск (или выбор) «сырых» данных, возможно, реализация подсистемы сбора (консолидации)
  - предобработка данных (нормализация, дискретизация, обработка пропущенных значений, удаление артефактов, проверка консистентности)
  - уменьшение размерности, выбор значимых характеристик, расчет интегральных показателей и инвариантов
- Определение типа решаемой задачи анализа:
  - классификация, прогнозирование, кластеризация, поиск исключений, ассоциативный анализ и т.д.

ПРОЦЕСС ИАД (2)

- Выбор (или разработка) алгоритма анализа:
  - определение ограничений и требований к алгоритму по точности, размеру, интерпретируемости, скорости построения и применения получаемых моделей, по типу исходных данных
- Непосредственно «Data mining»:
  - применение выбранного алгоритма анализа для поиска закономерностей выбранного типа и построение моделей
- Проверка моделей и представление результатов анализа:
  - визуализация, преобразование, удаление избыточности, оценка точности, достоверности моделей и т.д.
- Применение построенных моделей:
  - Descriptive data mining - информирование аналитика, «описательные» модели, основная цель – визуализация
  - Predictive data mining – прогнозирование неизвестных значений или характеристик в «новых» данных с помощью построенных моделей , основная цель – прогноз

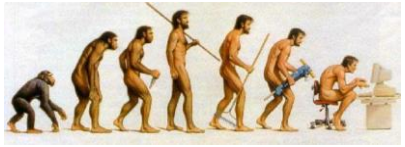
БОЛЬШИЕ ДАННЫЕ

В научной среде термин используется с 1990х

(2008) «Как могут повлиять на будущее науки технологии, открывающие возможности работы с большими объёмами данных?», Клиффорд Линч (редактору журнала Nature)

(2011) «Big Data: The next frontier for innovation, competition and productivity», McKinsey Global Institute

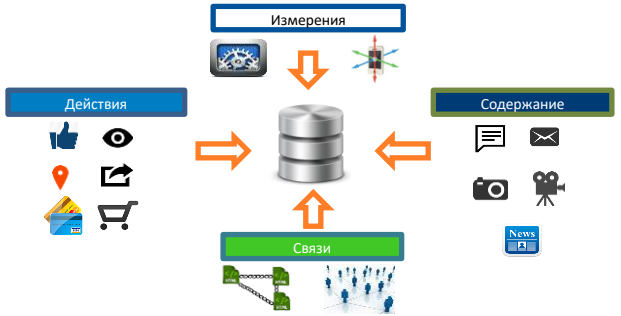
(2015) – термин Data Science

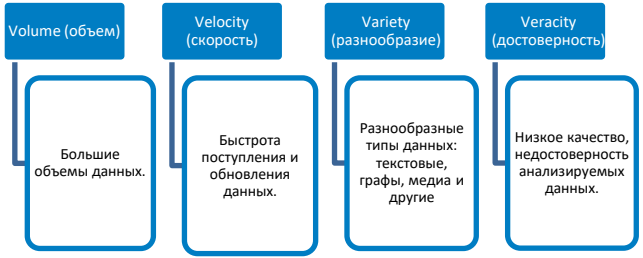


Начало цивилизации 2003

5 экзабайт

20+ экзабайт в сутки!





КТО ВИНОВАТ И ЧТО ДЕЛАТЬ С БОЛЬШИМИ ДАННЫМИ?

Виноваты жесткие диски:

50ГБ/сек

1x

1ГБ/сек
(10Гб/сек)

50x

500МБ/сек

100x

166x

0,3ГБ/сек

100МБ/сек

500x

Что делать?

Вертикальное масштабирование:

- дорого, технологически ограниченно
- НО относительно легко переносить аналитические алгоритмы

Горизонтальное масштабирование:

- дешево, потенциально технологически неограниченно
- НО сложно переносить аналитические алгоритмы

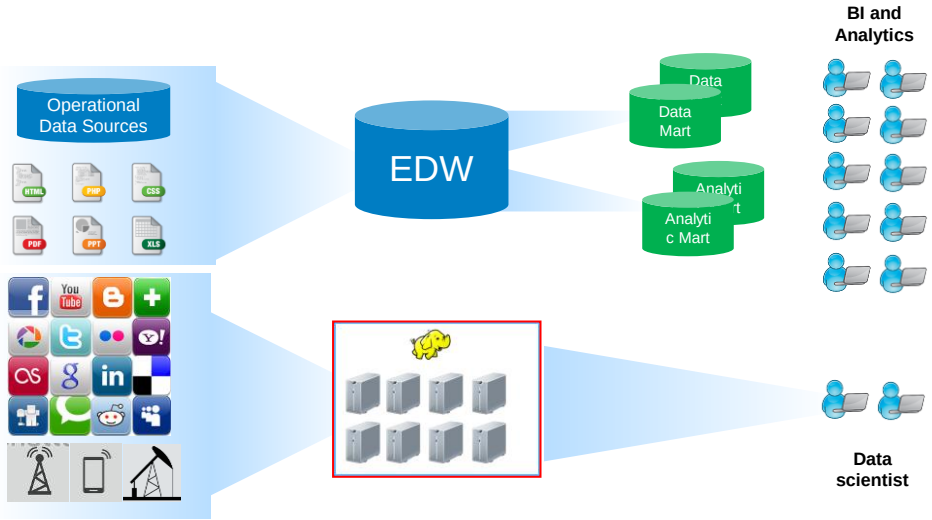
Индустрия выбирает MPP, а «математики» к этому не готовы

ОТЛИЧИЕ АНАЛИТИКИ БОЛЬШИХ ДАННЫХ ОТ ТРАДИЦИОННОЙ

Кто такой **Data Scientist**?

«три в одном»:

- Аналитик **прикладник-** понимает предметную область, в которой строит модель
- **Математик** - владеет методами прикладной статистики и ИИ
- **Программист** - может писать код для эффективной обработки больших объемов сложно структурированных данных



КОМПАНИЯ SAS

ПОЧЕМУ ПОЛЕЗНО ИЗУЧИТЬ ПЛАТФОРМУ SAS?

- Более 45 лет на рынке (с 1976 г.)
- Более 15 000 сотрудников в 400 офисах SAS в 56 странах
- Клиенты SAS - более 80 тысяч организаций в 140 странах мира.
- Более 90 компаний из top 100 списка «FORTUNE Global 500®».
- SAS занимает более 30% рынка аналитического ПО

Figure 1. Magic Quadrant for Data Science and Machine Learning Platforms

Source: Gartner (February 2020)

Инвестиции в R&D
> 30 %

sas THE POWER TO KNOW.

КОМПАНИЯ SAS В РОССИИ И СНГ

- **В России и странах СНГ компания SAS начала работу в 1996 году**
- **Полный спектр решений и услуг в области бизнес-аналитики:**
  - Консалтинг, внедрение, обучение, техническая поддержка
  - Более 300 сотрудников и стажеры
- **Крупнейшие клиенты SAS в России и СНГ:**
  - Все ведущие банки, включая топ 10 крупнейших российских банков (Альфа-банк, ЮниКредит банк, Райффайзенбанк, Ситибанк, GE Consumer Finance, Банк «Возрождение», Банк «Тинькофф Кредитные Системы», Райффайзен и др.)
  - Многие ведущие транспортные компании, включая РЖД и «Аэрофлот»
  - Крупнейшие компании из телекоммуникационного и топливно-энергетического сектора
  - Государственные организации: ЦБ РФ, ФТС, Налоговый Комитет Республики Казахстан и другие

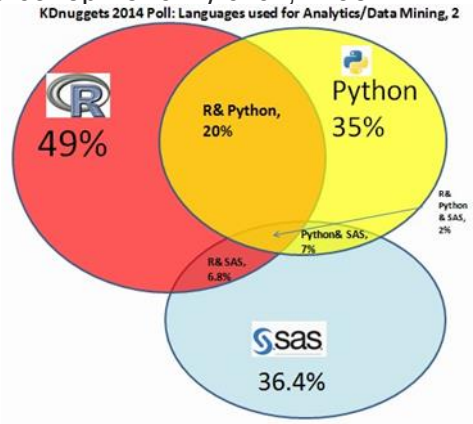
АНАЛИТИЧЕСКАЯ ПЛАТФОРМА SAS

КОМПАНИЯ SAS | ЦЕЛЕВЫЕ ОТРАСЛИ

- Автомобилестроение
- **Банковский сектор**
- Финансовые рынки
- Игорный бизнес
- **Телеком**
- Потребительские товары
- Безопасность
- Госсектор
- Здравоохранение
- Страхование
- Высшее и среднее образование
- Гостиничный бизнес
- Биология
- Фармакология
- СМИ
- Нефть и газ
- **Розничная торговля**
- СМБ
- Спорт
- **Транспорт**
- ЖКХ
- Производство

SAS ANALYTICS | ДОСТОИНСТВА

- **Исходные данные**. Обработка больших объемов данных сложной структуры из разных источников.
- **Глубина**. Реализованы самые современные методы анализа, которые постоянно совершенствуются, чтобы соответствовать самым последним достижениям.
- **Широта**. Совокупность методов:
  - Статистический анализ, визуализация и интеллектуальный анализа данных
  - Временные ряды, прогноз, эконометрика
  - Контроль и улучшение качества
  - Исследование операций
  - Имитационное моделирование
  - Анализ текстовых данных



- **Открытость**.Поддержка множества парадигм, основаных на многих дисциплинах, чтобы наилучшим образом формулировать и решать аналитические задачи.
- **Наглядность**. Поддерживается много графических методов визуального исследования данных, поиска взаимосвязей и неочевидных зависимостей для улучшения поддержки принятия решений.
- **Воспроизводилось**. Генерируемый код удовлетворяет большинству корпоративных и государственных требований к воспроизводимости и верифицируемости.
- **Независимость**. Работает на многих платформах.

SAS® Viya™

Источники данных

- **In-Stream**
- **In-Hadoop**
- **In-Database**

Parallel & Serial, Pub / Sub, Web Services, MQs

In-Memory Runtime Engine

Cloud Analytics Services (CAS)

Microservices

UAA, Data Source Mgmt, Folders, etc..., BI GUIs, CAS Mgmt, Log, Analytics GUIs, Data Mgmt GUIs, Query Gen, Env Mgr, Model Mgmt, Audit

Решения

- Клиентская аналитика
- Аналитика
- Управление рисками
- Бизнес-визуализация
- Безопасность и противодействие мошенничеству
- Управление данными

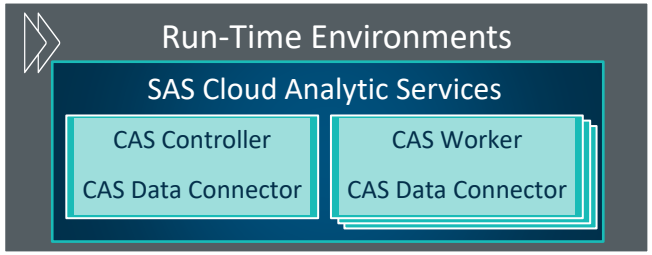
Программные Интерфейсы

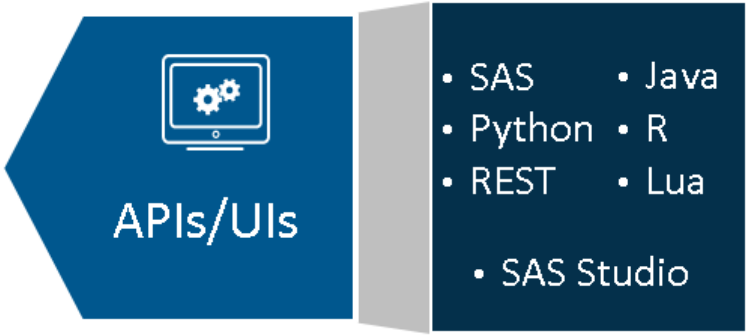
REST API, python, Java, R, Lua

CLOUDFOUNDRY, docker, Ci, ANSIBLE, rpm

amazon web services, Azure, Google Cloud Platform, openstack, vmware, ORACLE CLOUD, sas

CAS АРХИТЕКТУРА





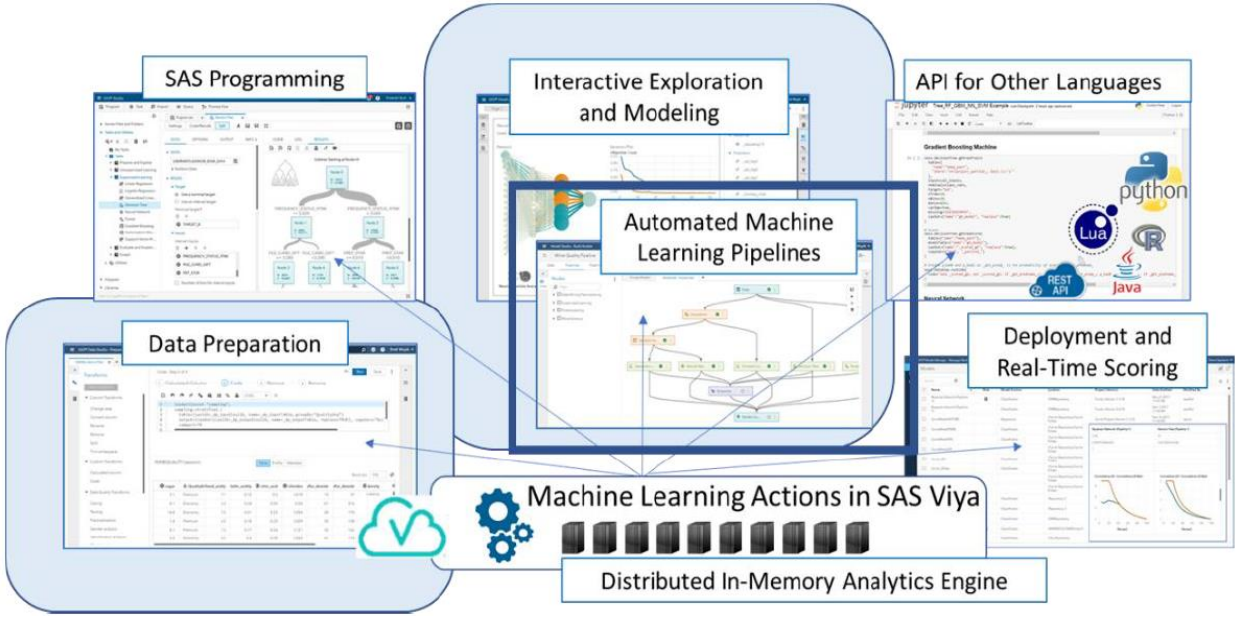
ВОЗМОЖНОСТИ SAS VIYA

- SAS® Model Manager
- SAS® Decision Manager
- SAS® Event Stream Processing
- SAS® Scoring Accelerators

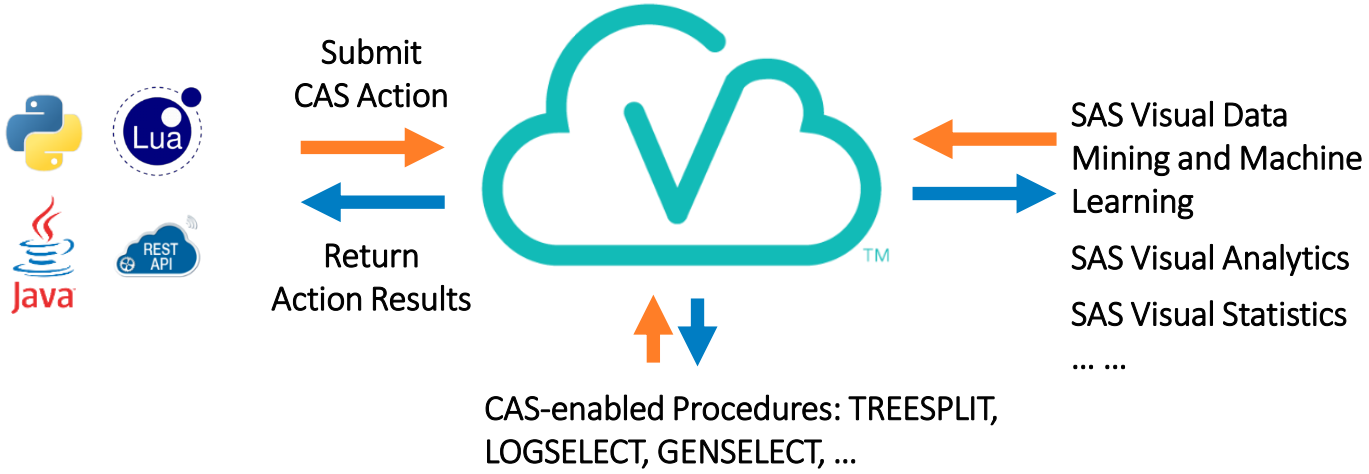
- SAS® In-Database Technologies
- SAS® Data Preparation
- SAS® Data Quality

- SAS® Visual Forecasting
- SAS® Econometrics
- SAS® Visual Investigator
- SAS® Visual Analytics
- SAS® Optimization
- SAS® Visual Text Analytics
- SAS® Visual Data Mining and Machine Learning
- SAS® Visual Statistics

SAS VISUAL DATA MINING AND MACHINE LEARNING

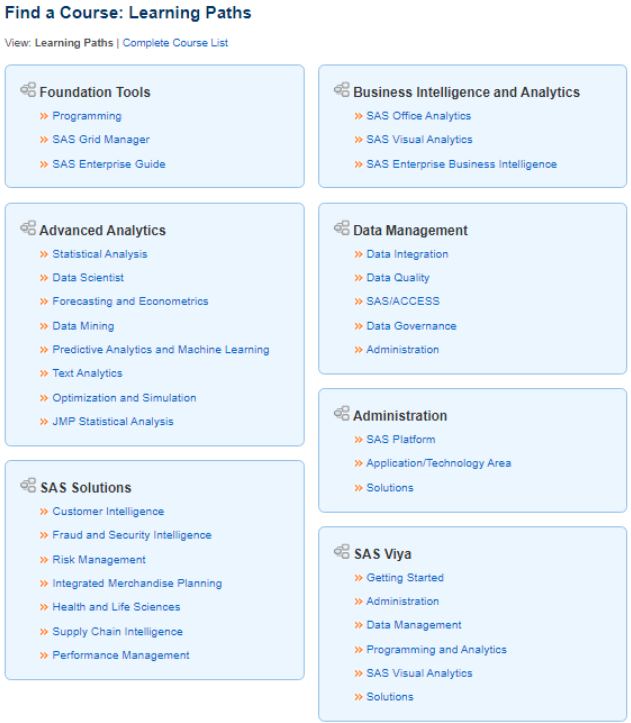


SAS VIYA – РАЗНЫЕ ИНТЕРФЕЙСЫ, ОДИН РЕЗУЛЬТАТ



Направление изучения SAS

Специализация



- https://support.sas.com/training/us/paths/