Введение в методы интеллектуального анализа данных (Data Mining)

к.ф.-м.н. М.И. Петровский (michael@cs.msu.su), SAS Certified Data Scientist

лаборатория «Технологий программирования»
ВМиК МГУ им. М.В. Ломоносова

Задачи курса

- Познакомить с предметной областью:
  - дать основные определения и терминологию, обсудить прикладные задачи
- Рассмотреть основные задачи Data Mining:
  - и популярные алгоритмы на основе методов машинного обучения для их решения
  - меньше теории, больше алгоритмов и понимания как их настраивать и использовать
- Дать практический опыт решения задач Data mining:
  - Практические задания на Питоне:

Содержание курса (1/3)

1. Введение
2. Выявление структур в данных (обучения без учителя)
   - Поиск ассоциативных правил (алгоритмы apriori и fp-tree) и тематическое моделирование (методы главных компонент, неотрицательная матричная факторизация)
   - Кластеризация (иерархическая, метрическая, вероятностная)

Содержание курса (2/3)

3. Задача прогнозирования (обучение с учителем)
   - Виды задач прогнозирования, проблема переобучения, оценка и сравнение моделей, простейшие методы прогнозирования (kNN и Naïve Bayes)
   - Методы на основе деревьев решений и их ансамблей
   - Регрессионные модели (отбор и преобразование пространства признаков, регуляризация, обобщенные линейные модели)

Содержание курса (3/3)

3. Задача прогнозирования (обучение с учителем)
   - Нейронные сети (MLP, RBF, борьба с переобучением, SOM, задачи глубинного обучения)
   - Методы опорных векторов для задач классификации и регрессии
   - Моделе-независимая визуализация зависимостей
4. Выявление аномалий

ЛЕКЦИИ и ПРАКТИЧЕСКИЕ ЗАДАНИЯ!!!

Итог = практические задания + посещаемость + экзамен

Интуитивное определение ИИ

Искусственный интеллект – проблема определения термина

- **Нет общепризнанного** научного определения
- Сильный коммерческий **«хайп»**, смещающий акценты
- Часто термин ИИ **неправильно используется** в очень узком смысле, как машинное обучение, или даже нейросети, или даже глубокое обучение нейросетей
- Надо делать акцент на слово **«искусственный»**

Пример определения:

- «ИИ - междисциплинарная **область знаний**, занимающаяся исследованием и разработкой методов и артефактов (**устройств или программ**), которые способны **имитировать интеллектуальную** (разумную/рациональную) **деятельность** (мышление/принятие решение) **человека**»

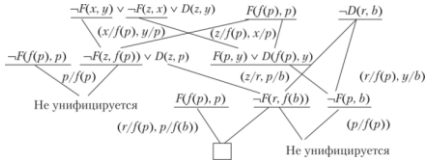
Почему «думать» и «делать» это разные области в ИИ?

«Думать» («мыслить») – оперировать знаниями

- Есть формальное **представление знаний** и интеллектуальная система, способная на их основе **генерировать** новые непротиворечивые знания или **проверять** утверждения, в том числе в условиях неопределенности
- Примеры задач ИИ из категории «думать»

  «рационально» – автоматическое доказательство теорем

  «как человек» – распознавание эмоций по фото или видео

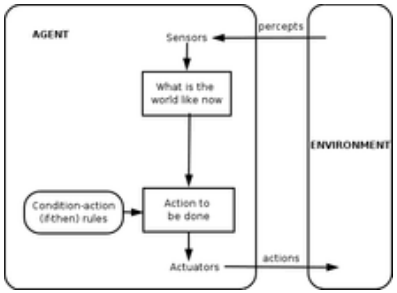


«Действовать» - взаимодействовать с окружающей средой (**интеллектуальный агент)**

- производит **действия**, получает отклик **среды**,
- самокорректируется (**учится**) с определенной **целью**
- Примеры задач ИИ из категории «действовать»

  «рационально» - беспилотный автомобиль

  «как человек» - чат-бот, голосовой помощник, игровой ИИ



Почему человек нерационален и плохо ли это?

Что значит «рационально»?

- Достижение заданной цели эффективным (а лучше оптимальным) непротиворечивым путем
- По сути – **задача оптимизации** (даже там, где это неочевидно, например, системы автоматических рассуждений не используют полный перебор вариантов)

Причины **нерациональности** человека:

- Недостаток информации
- Огромное пространство перебора при поиске решений (шахматы)
- Невозможность задать целевую функцию (помогает теория полезности)
- Биологические особенности работы мозга человека

Механизмы принятия решений человеком (все моделируются в ИИ):

- **Рефлексные** (не используют мозг, например, отдернуть обожженную руку)
- **Интуитивные/эмоциональные/спонтанные** (используют лимбическую систему, поощряются гормонально, приносят удовольствие) – «золотая жила» для ИИ (Эмоциональная экономика)
- **Рациональные** (работает неокортекс, ничего приятного, сильно устаешь, никто не любит думать)

Искусственный Интеллект

Общий ИИ (**AGI**)

- Философские и этические вопросы ИИ
- Футуристика
- Исследования принципов работы биологического интеллекта
- Вопросы создания универсального автономного интеллектуального агента («скайнеты» и прочие «матрицы»)

Большинство ученых считает, что в обозримом будущем в этой области **прогресс маловероятен:**

- нет работающих теорий, инструментов и проблема «общечеловеческого бэкграунда» или «здравого смысла» - ограниченность знаний любой интеллектуальной системы
- **Но** есть надежда на **Big Data**!

ИИ в узком смысле (**ANI**)

Не интересуется общими вопросами, а изучает и развивает инструменты и приложения ИИ:

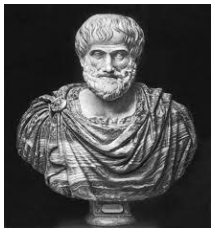
- Автоматические рассуждения
- Машинное обучение (сейчас ключевой инструмент)
- Поиск и оптимизация
- Человеко-машинное взаимодействие

«Дополненный» интеллект: **не AI, а IA** (Intelligence amplification) – не замена, а усиление

Бурное **развитие приложений** и алгоритмов из-за развития вычислительной техники

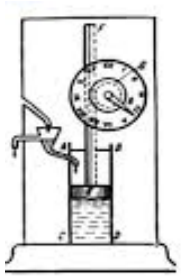
Но по сути застой в теории –последние фундаментальные результаты **20+ лет назад**

Предыстория ИИ (античность и средние века)



Какие проблемы волновали:

- **Можно ли получать новые знания «механически» без опыта, а на основе уже доступных знаний?**
- **Могут ли «думать» неживые системы?**

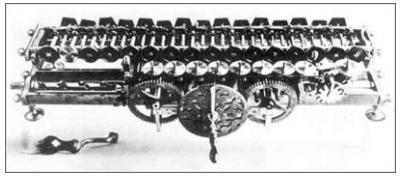


Античная Греция:

- Силлогизмы Аристотеля (4 в. до н.э.) – основы формальной логики и процедуры доказательства
- «отец пневматики» Ктесибий (3 в. до н.э.) – «умные» механизмы, включая водяные часы (календарь) с саморегулирующимся потоком воды

Первые вычислительные средства

- «концептуальные колеса» («механический пролог») Раймунд Луллий (14в.) + Томас Гоббс (17в.)
- «механический арифмометр» Леонардо да Винчи (15 в.) + Блез Паскаль (17 в.)
- «Концептуальный калькулятор» Готфрид Вильгельм Лейбниц (17 в.)



Предыстория ИИ (Новое время 17-19 вв.)

Какие проблемы волновали:

- **Понять природу появления знаний**
- **Развить аппарат формальной логики, в том числе с учетом неопределенности**
- **Сделать обратную связь в механизмах, чтобы они могли работать под собственным управлением**

Основные вехи:

- Дуализм Р. Декарта (17в.) – «мысль (душа) отдельно, материя отдельно». Альтернатива дуализму – **материализм** = свобода воли есть результат перебора решений в рамках некоторого материального процесса.
- Поиск источника знаний и принципов их формирования:

  Эмпиризм - Фрэнсис Бэкон (16в.), Принцип индукции - Дэвид Юм (17в.), Логический позитивизм – Венский кружок (19в.)
- **Томас Байес (18в.)** – правило обновления вероятностей с учетом новых фактов, Байесовский вывод в условиях неопределенности
- **Джордж Буль (19в.)** – логика высказываний
- **Готлоб Фреге (19в.)** – логика первого порядка = булева логика + отношения + высказывания

Саморегулируемые (**с обратной связью**) механизмы: термостат (17в.), регулятор паровой машины (19в.)

Предыстория ИИ (20в.) в психологии и лингвистике

Какие проблемы волновали:

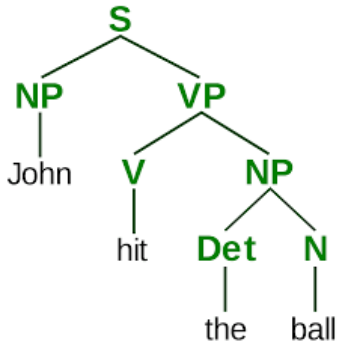
- **Как думают и действуют живые существа?**
- **Как речь связана с интеллектом?**

На рубеже 19 и 20 вв. господствовал бихейвиористический (поведенческий) подход:

- отрицал понятия «мысли», «знания», «мышление» как неизмеримые, а потому ненаучные
- оперировал только со стимулами и откликами
- более или менее работал на простых животных, почти не работал на людях и высших животных

Как ответы-опровержения, пытавшиеся закрыть пробелы «бихейвиористов», появились:

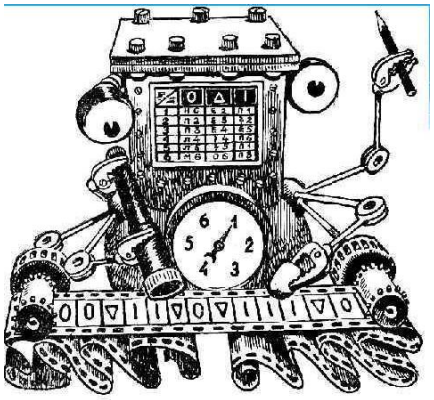
- **когнитивная психология** (1910), рассматривающая мозг (даже в процессе распознавания визуальных образов) как устройство обработки информации, работающее на «скрытых» логических правилах
- синтаксические модели естественных языков (1950е), **теория формальных грамматик** Хомского и позже **компьютерная лингвистика**



Предыстория ИИ (20в.) в математике

Проблемы **сложности** и **вычислимости**, связанные с ИИ:

- Доказано, что существует эффективная процедура проверки истинности любого высказывания в логике 1 порядка
- Теорема Гёделя о неполноте, в логике первого порядка нельзя выразить принцип мат. индукции, существуют функции от целых чисел, которые нельзя вычислить (нельзя доказать их истинность или ложность)
- Тезис вычислимости Черча-Тьюринга – любую вычислимую функцию можно вычислить с помощью машины Тьюринга, но есть невычислимые функции
- Экспоненциальная сложность и неразрешимость задач, NP-полнота и NP-трудность



Теория полезности (объективной и субъективной, численной и порядковой) – позволяет формализовать оптимизационную поисковую задачу, в том числе в приложениях ИИ

Теория игр – аппарат для принятия решений:

- важный с точки зрения ИИ «философский» результат – есть ситуации, когда рациональный интеллектуальный агент должен принимать случайные решения (смешанные стратегии)
- отдельная область в машинном обучении – обучение с подкреплением

Исследование операций и **Марковские процессы** принятия (последовательных) решений

Предыстория ИИ (20в.) в кибернетике и вычислительной технике

Алан Тьюринг (Англия):

- (1940) Heath Robinson – дешифратор для энигмы,
- (1943) Colossus – компьютер общего назначения, на лампах, но без программ

Конрад Цузе (Германия):

- (1943) Z-3 – с программами, языком, плавающей точкой, но на механических реле

Джон Атанасов (США)

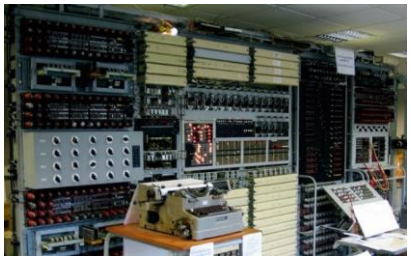
- (1942) ABC - без программ, с двоичной арифметикой

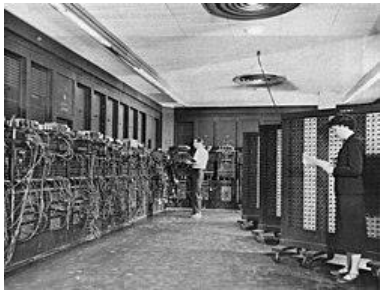
Джон Экерт, Джон Мочли (США)

- (1945) ENIAC

Норберт Винер – «отец» кибернетики, а вообще-то и ИИ, и оптимального управления (вместе с Беллманом и Понтрягиным):

- Управление (в том числе саморегуляция) как оптимизация некоторой целевой функции, возможно во времени (стохастическое управление)

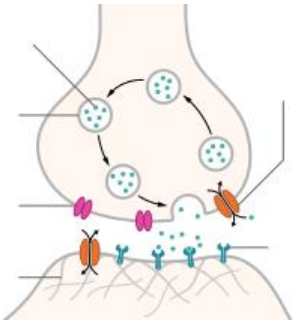




Предыстория ИИ (20в.) в неврологи

Результаты к рубежу 19 и 20 вв.:

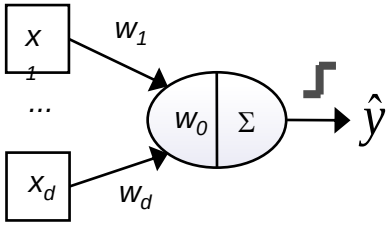
- Мозг – орган мышления, разные части отвечают за разные функции («спасибо» войнам и ранениям)
- Метод Гольджи (1873!!) – окрашивание клеток нервной ткани для наблюдения
- В мозге более 10^{11} специальных клеток (**нейронов**)
- Каждый нейрон соединен через синапсы (около 10^{14}) с другими нейронами
- Мозг может: обучаться, адаптироваться, распознавать образы, осознавать «себя», устойчив к шуму, травмам и ошибкам
- Нейрон имеет «входные» отростки (**дендриты**) и «выходные» (**аксоны**)
- Информация (сигнал, «нервный импульс») идет от дендритов к аксону через тело (ядро) клетки
- Аксоны соединяются с дендритами (других клеток) через **синоптический** переход (щель), в нем через нейромедиаторы электрический сигнал преобразуется в химический и наоборот



Искусственный нейрон

(1943) Мак-Каллок и Питс - Искусственная нейросеть

- **грубая модель биологического нейрона + логика высказываний + теория вычислений Тьюринга**
- сетью бинарных искусственных нейронов можно описать функции алгебры логики



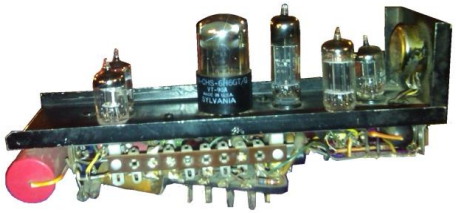
Дональд Хэбб (нейрофизиолог) предложил правило «обучения»

- $\Delta w_{ij} = \eta\, \hat{y}_i x_j$

$$\hat{y} = f\left(w_0 + \sum_{i=1}^{d} w_i x_i\right)$$

(1951) Аспиранты М. Минский и Д. Эдмондс собрали первый обучаемый нейро-компьютер Snarc:

- 40 нейронов, с обучением, на 3000 лампах + автопилот от бомбардировщика
- им не хотели присуждать Ph.D. – «это не математика! где теоремы?»
- но вступился фон Нейман - **«это математика будущего»**



Рождение ИИ и ранние успехи (1950е-1970е)

(1950) Краеугольная работа Тьюринга **«Computing Machinery and Intelligence»**:

- Тест Тьюринга, принципы машинного обучения, генетические и другие поисковые алгоритмы, обучение с подкреплением

(1956) **Дартмутский семинар** (2 месяца, 10 человек), итоги – «развод» с кибернетикой и теорией управления:

1. ИИ не математика, а информатика (без компьютера нельзя)
2. ИИ моделирует и изучает поведение и мышление человека (в том числе нерациональное)

Череда **успехов**:

- Изначальный список Тьюринга «машина никогда не сможет …» сокращался очень быстро
- Разработаны «универсальные» решатели (Logic Theorist, General Problem Solver, Prolog и др.)
- Разработан LISP, показал возможности символьного решения задач (в том числе математических)
- Усовершенствование методов обучения нейросетей (обратное распространение ошибки)
- Персептрон Розенблатта и теорема о его сходимости
- Прикладные успехи: экспертные системы в медицине, управлении и инженерии на основе сложных моделей представления знаний (типа фреймов), машинный перевод и распознавание образов

Зима ИИ (с 1960х до 80х)

«Зима ИИ» - сокращение финансирования и интереса общества, отток специалистов, коммерческий и научный провал многих проектов, оказалось, что многое **«без ИИ лучше и дешевле»** плюс **проблема здравого смысла** (common sense):

- Провал методов машинного перевода (с русского, кстати) и закрытие гос. финансирования, из-за проблемы **семантической неоднозначности**:

  «the spirit is willing, but the flesh is weak» $\xleftrightarrow[\text{на русский и обратно}]{}$ «the vodka is good, but the meat is rotten»

- **комбинаторный взрыв -** проблемы сложности вычислений в системах логического вывода и автоматических рассуждений (в принципе решит, но лет через 100)
- Провал идеи **«эволюции программ»** – самопрограммирующиеся программы по принципу генетических алгоритмов
- Принципиальные **ограничения персептронов** (например, задача XOR для однослойного), книга Минского и Пейперта с критикой ⇒ смерть Френка Розенблатта (возможно, покончил с собой)
- **Крах** рынка **LISP машин** – оказались хороши в науке, плохи в бизнес-приложениях
- Провал идеи «**компьютера 5 поколения»** – «интеллектуального компьютера», например на прологе
- **Неэффективность экспертных систем** на основе фреймворков и семантических сетей: сложно описывать, долго настраивать, низкая точность, противоречивость

Причины краха больших надежд

Основная причина – **изоляционизм** специалистов по ИИ от остальных компьютерных наук:

- Изначальная уверенность, что символьные вычисления, логические методы и формальные грамматики есть основа разумной деятельности и они решат все проблемы
- Оказалось, что «умение решать» математические задачи школьного уровня или проходить тест на IQ не делает умнее не только человека, но и компьютер
- Сложные модели представления знаний (фреймворки и семантические сети) не принесли существенной пользы в реальных задачах
- Машинное обучение не следует отделять от теории информации и прикладной статистики
- Рассуждения в условиях неопределенности нельзя изолировать от теории вероятности, байесовских методов принятия решений и других классических математических дисциплин
- Поиск в пространстве состояний на самом деле раздел классической оптимизации
- Автоматизированное формирование рассуждений не должно трактоваться как независимое от формальных логических методов

Стало понятно, что в будущем будут востребованы **гибридные интеллектуальные системы**:

- сочетающие в себе несколько методов ИИ или классические математические методы и ИИ, например машинное обучение + оптимальное управление

Оттепель ИИ (90е)

Многие **классические методы** успешно пережили «зиму», например:

- Экспертные системы в медицине, логистике, проектировании и других областях
- Интеллектуальное планирование и распределение ресурсов в задачах управления
- Системы нечеткого вывода в задачах управления механизмами (автоматические коробки передач)
- Обучение с подкреплением для обнаружения и разрешения конфликтов в воздушном движении
- Нейросети в задачах распознавания визуальных и звуковых образов
- Системы на основе поиска в пространстве состояний в компьютерных играх
- Робототехника

Рывок в методах **машинного обучения** и интеллектуального анализа данных:

- В 80х заново «переизобрели» все, что было в нейросетях 50х, включая разные формы Back Propagation
- Архитектуры Deep Learning (CNN, RNN, AE, LSTM, …) и методы их обучения (да, да, им более 20 лет)
- Бустинг слабых моделей и другие ансамбли
- Метод опорных векторов – «убийца нейросетей», который так и не смог их убить
- Скрытые Марковские модели и обучаемые сети Байеса

Бум ИИ в 21 веке, связь с ML и Data Science

Застой в теории - ничего принципиально нового уже больше 20 лет

Прорыв в практике, почему? **Вычислительная техника** стала мощной и дешевой!

- Дешево накапливать и хранить большие объемы данных
- Можно просчитывать сложные модели за разумное время
- Математика подстраивается под вычислительную технику

В бизнес-сообществе часто термин ИИ используют как синоним **Data Science** или **ML**

Машинное обучение подраздел ИИ, изучающий методы построения алгоритмов, способных обучаться на прецедентах для решения задач: прогнозирования (классификации, ранжирования, регрессии), поиска скрытых структур в данных (ассоциаций, корреляций, кластеризации), обнаружения аномалий.

Data Science (наука о данных) - раздел информатики, изучающий проблемы анализа, обработки (в том числе интеллектуальной) и представления данных в цифровой форме.

Тесно связано с понятием **больших данных**.

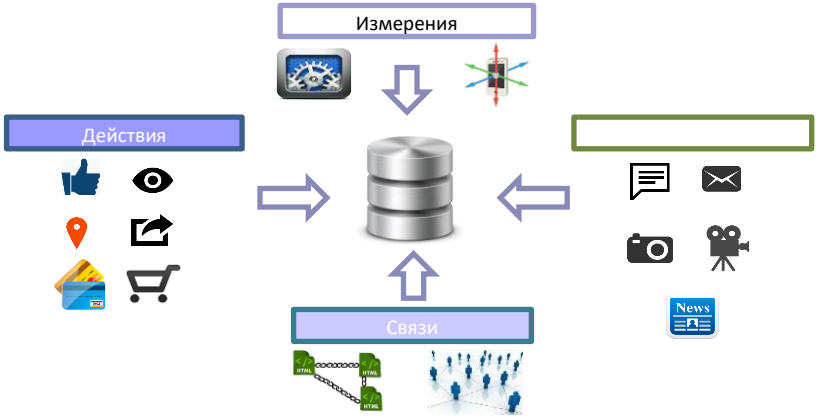
Большие данные

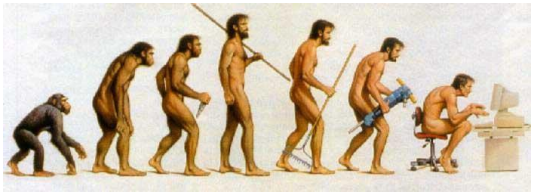
В научной среде термин используется с 1990х

(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!

Volume (объем)	Velocity (скорость)	Variety (разнообразие)	Veracity (достоверность)
Большие объемы данных.	Быстрота поступления и обновления данных.	Разнообразные типы данных: текстовые, графы, медиа и другие	Низкое качество, недостоверность анализируемых данных.

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

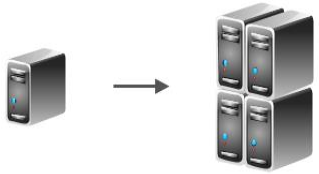
Что делать?

Вертикальное масштабирование:

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

Горизонтальное масштабирование:

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



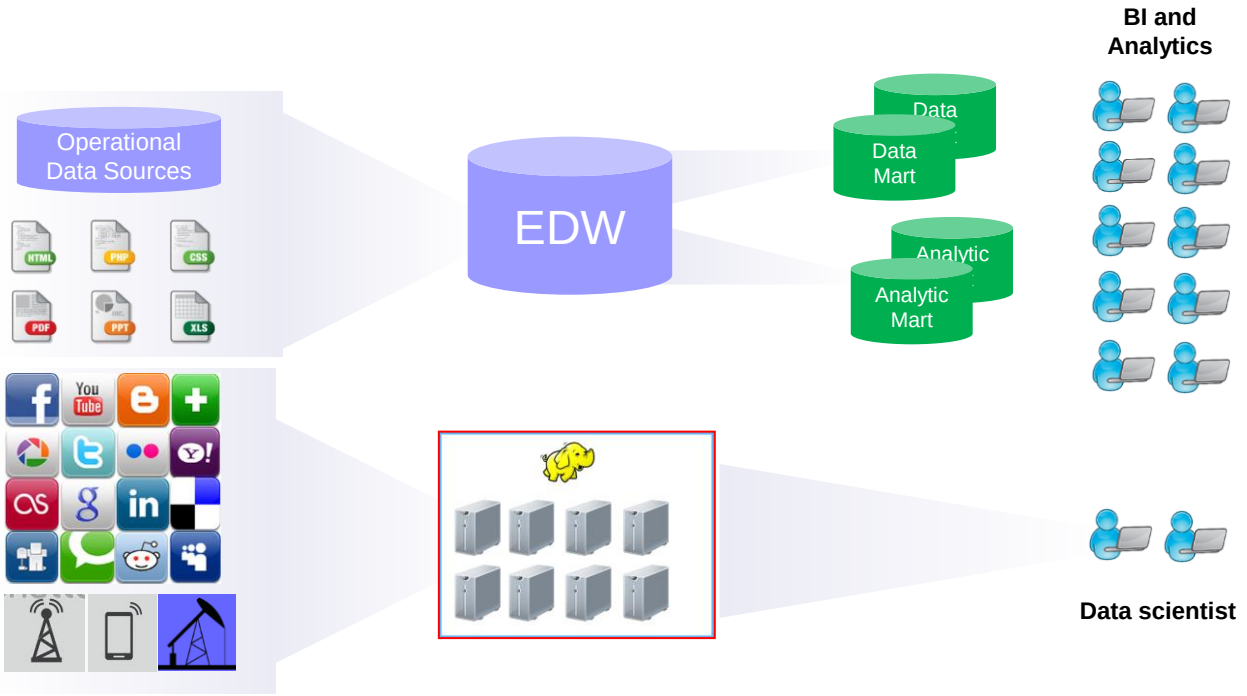
Индустрия выбирает MPP, а «математики» к этому не готовы

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



Успехи современного ИИ

Адаптируемый (с обучением) **ИИ** + **Большие данные** + **мощная вычислительная техника** = **заявка на AGI**

Еще 10 лет назад ученые были уверены, что все, что перечислено ниже, невозможно:

- Нейросети глубокого обучения распознают лица людей лучше чем сами люди
- Самообучающийся ИИ для игр (шахматы и го) обыгрывает любого человека, причем играет «по-человечески» (технически не всегда рационально), пример – претензии Каспарова к Deep Blue
- Алгоритмы выявления ключевых слов, аннотирования текстов, ответов на вопросы, обученные на больших корпусах (например, Wikipedia) работают все лучше, а используют лингвистику все меньше
- Многоязыковые переводчики – учатся на одном наборе пар языков и успешно переводят другие пары (Google Multilingual Neural Machine Translation), используют языково-независимое представление
- Беспилотные автомобили на реальных дорогах

ИИ в современной индустрии

Что делает компанию ИИ-компанией:

- Сформулированная обоснованная общекорпоративная **стратегия** внедрения и использования ИИ, которую поддерживает руководство
- Непрерывная работа по развитию **ИИ команды** и ее экспертизы (обучение, наем, мотивация, централизованное управление)
- Фреймворк и процедуры для **поддержки жизненного цикла ИИ** средств и моделей
- Ответственная и надежная **работа с данными** (сохранение, очистка, нормализация, валидация)

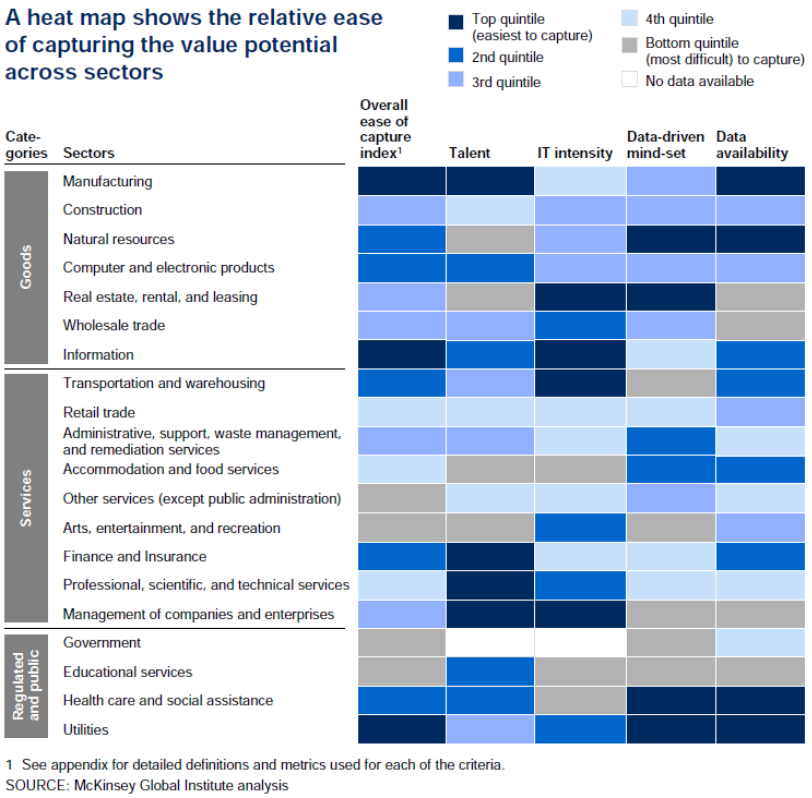
Что мешает внедрению ИИ в компании - **люди, люди** и **еще раз люди**:

- Руководство, не понимающее, что такое ИИ, но желающее его использовать
- От неаккуратности до скрытого саботажа рядовыми сотрудниками – в процессе внедрения ИИ если и не принесет прибыль, то «косяки» с хранением и обработкой данных, неверной отчётностью и неэффективным управлением «найдет» точно

Страх безработицы при внедрении ИИ (как и любых инноваций) сильно преувеличен:

- «безработица, порождаемая автоматизацией, не является более предметом предположений – она стала одной из животрепещущих проблем современного общества» (Норберт Винер, **70 лет назад**)

Современная индустрия ИИ и Больших данных



Gartner 2021 Magic Quadrant for Data Science and Machine Learning Platforms.

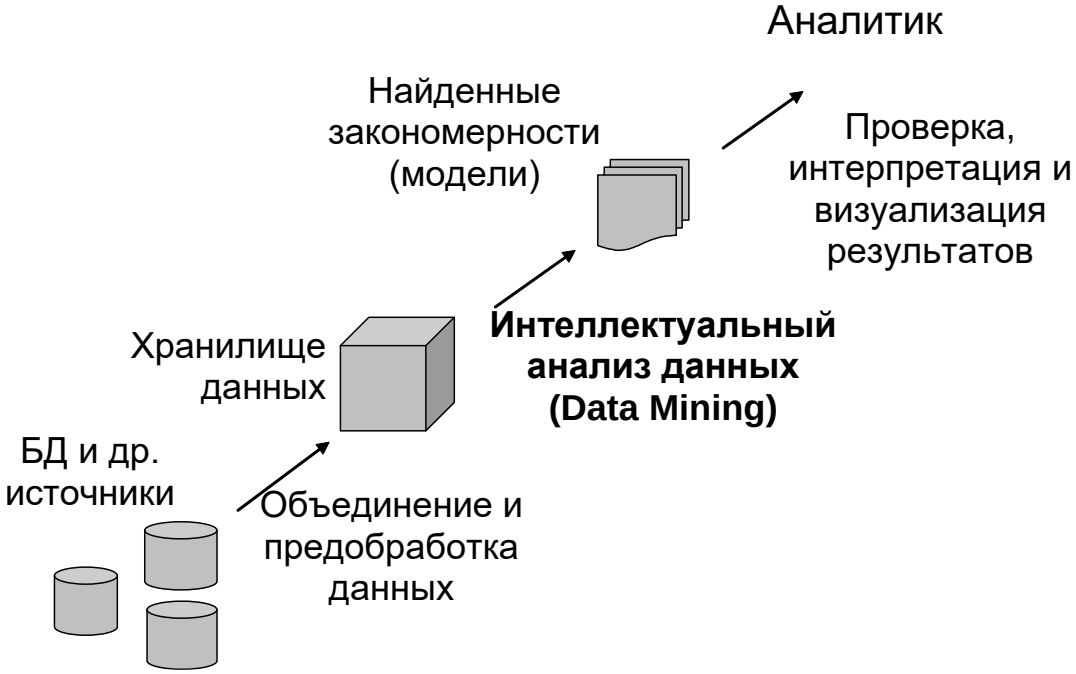
Эволюция технологий хранения и обработки данных

- … — 1960-е:
  - Файлы и файловые архивы
- 1960-е:
  - Первые СУБД, иерархические, сетевые и т.д.
- 1970-е:
  - Реляционная модель данных, реляционные СУБД
- 1980-е:
  - «Продвинутые» СУБД (объектно-реляционные и объектные, «расширенные» реляционные, дедуктивные и др.)
  - «Специализированные» СУБД (гео-,научные, инженерные и др.)
- 1990-е — …:
  - Мультимедийные БД, WWW, хранилища,
  - витрины данных,OLAP, Data Mining

Актуальность и необходимость интеллектуального анализа данных (ИАД)

- Проблема больших объемов («Data explosion»):
  - Средства автоматического сбора данных, повсеместное внедрение СУБД, электронный документооборот, WWW, мультимедийные архивы и т.д. приводят к росту объемов и усложнению структуры хранимой информации.
- Традиционные средства не справляются:
  - Информационный поиск и стат. анализ не везде помогают – много данных, сложная структура и нужно знать точно, что искать.
  - Вывод: много данных, но мало информации для аналитика.
- Необходимо:
  - Наличие программных средств автоматизированного анализа данных большого объема и сложной структуры.

Интеллектуальный анализ данных (Data Mining)



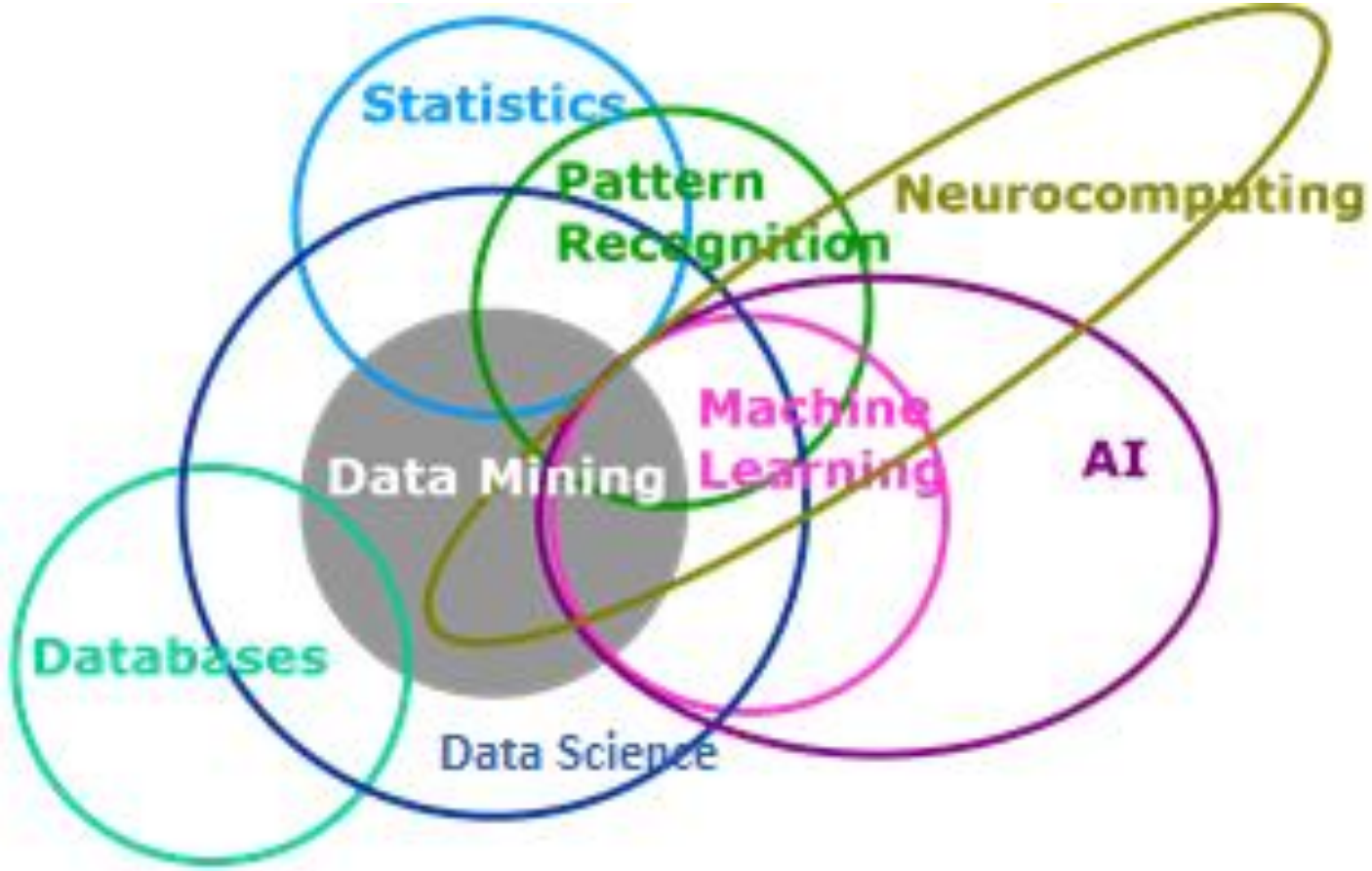
Системы *интеллектуального анализа данных* (ИАД) – класс программных систем поддержки принятия решений, задачей которых является поиск *скрытых, ранее неизвестных, содержательных и потенциально полезных* закономерностей в *больших объемах разнородных, сложно структурированных* данных.

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

Краткая история ИАД

- 1989 IJCAI Workshop on Knowledge Discovery in Databases (Piatetsky-Shapiro)
  - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 Workshops on Knowledge Discovery in Databases
  - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
  - Journal of Data Mining and Knowledge Discovery (1997)
- 1998 ACM SIGKDD, SIGKDD'1999-2001 conferences, and SIGKDD Explorations
- Другие конференции по data mining
  - PAKDD, PKDD, SIAM-Data Mining, (IEEE) ICDM, etc.

Место Data mining среди современных подходов анализа данных



Обратите внимание на пересечения областей!

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

Место ИАД в процессе поддержки принятия решений

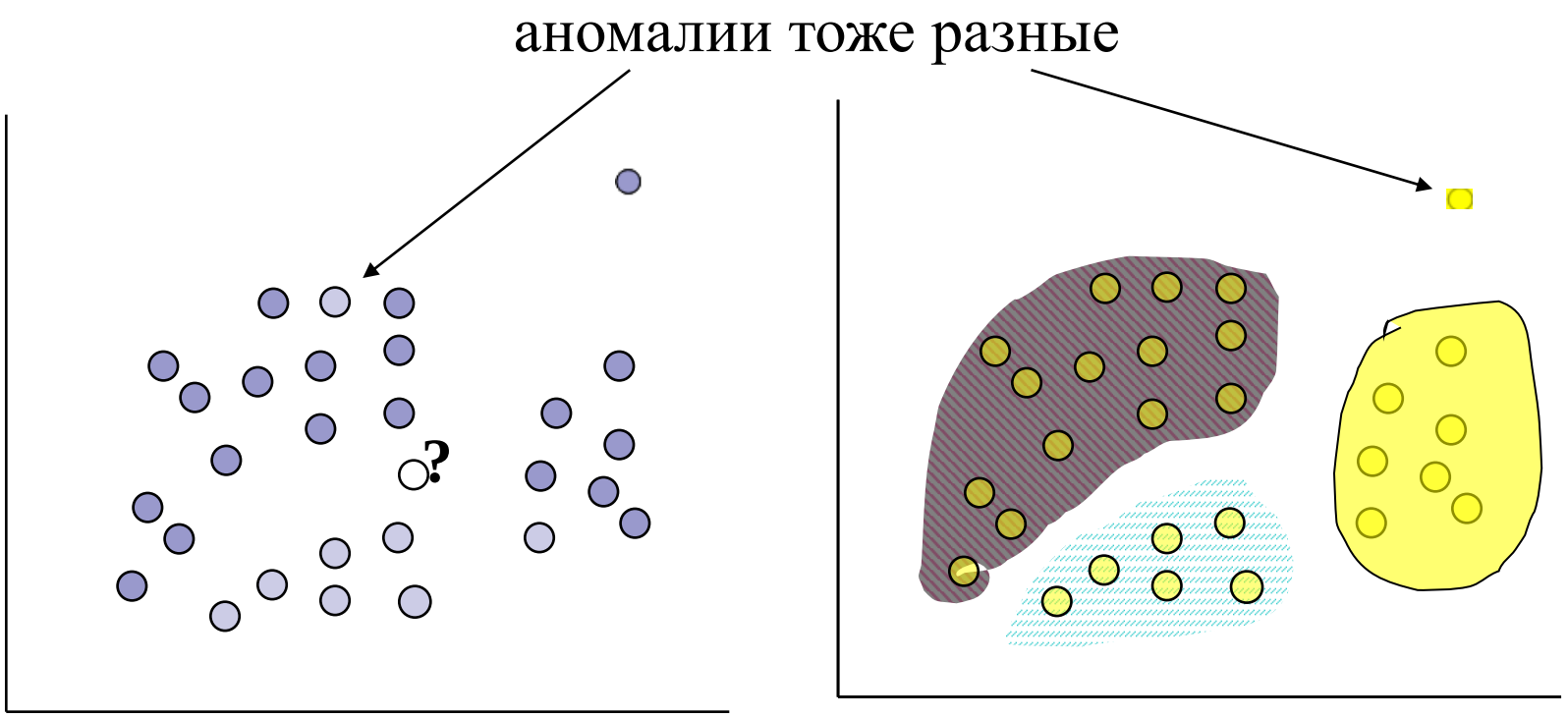
Основные типы исходных данных

- Транзакционные
  - Объекты анализа – «события» различной структуры с числовыми и категориальными атрибутами и с временной меткой
- Табличные
  - Объекты анализа представлены в виде реляционных таблиц, возможно взаимосвязанных (заданно ER-схемой), имеют разнотипные атрибуты
- Временные ряды и числовые данные большого объема
  - Обработка результатов наблюдений, научных экспериментов, характеристик технологических процессов
- Электронные тексты на естественном языке
  - анализ содержимого документов
- Графовые данные
  - Анализ взаимосвязей (SNA)
- Специализированные данные
  - Мультимедия, геоданные, ДНК, программный код и многое другое

Данные для анализа

- Объект анализа (или прецедент, или кейс, или наблюдение, …) задается набором признаков (или атрибутов, или свойств, …)
- Признаки по типам бывают:
  - Категориальные - нет расстояний, не задан порядок
  - Ординальные (порядковые) – нет расстояний
  - Числовые – есть расстояние
- «Размеченный» набор данных – для каждого объекта выделен один или более признаков, которые могут быть неизвестны и которые нужно предсказывать, тогда задача обучения «с учителем», иначе «без учителя» («неразмеченный» набор данных):
  - «Выходные» признаки - нужно предсказывать (они же отклики, или «зависимые переменные», или …)
  - «Входные» признаки, которые считаются всегда известными (они же входы, или «независимые переменные», или регрессоры, …)

Обучение «с учителем» и «без»



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

Классификация

- Дано:
  - «размеченный» тренировочный набор – для каждого объекта известен его класс
- Цель:
  - Построить классификатор – функцию или алгоритм, который в зависимости от свойств объекта предсказывает его класс
- Приложения в медицине:
  - Компьютерная безопасность
  - Производство- прогнозирование качества изделий
  - Распознавание образов

Ранжирование

- Дано:
  - «размеченный» тренировочный набор – для каждого объекта известен его класс или несколько не взаимоисключающих классов
- Цель:
  - Построить функцию или алгоритм ранжирования, который в зависимости от свойств объекта вычисляет степень его соответствия классам
  - Результат ранжирования: в рамках каждого класса можно упорядочить объекты по степени соответствия данному классу, и наоборот, в рамках каждого объекта можно упорядочить классы по степени соответствия данному объекту
- Приложения:
  - Документооборот и информационный поиск - рубрикация документов
  - Кредитование - оценка заемщика
  - Рекомендательные системы

Прогнозирование

- Дано:
  - «размеченный» тренировочный набор – для каждого объекта известно значение некой числовой величины, которое необходимо спрогнозировать
- Цель:
  - Построить функцию, которая в зависимости от свойств объекта предсказывает значение данной величины
- Приложения:
  - Финансы - прогноз курсов валют, цен на нефть и др., оценка ожидаемых доходов или убытков предприятия
  - Маркетинг – прогнозирование числа новых клиентов или убыли старых
  - Прогноз электропотребления

Поиска ассоциаций

- Дано:
  - «не размеченный» тренировочный набор – для каждого объекта известны только значения его свойств (атрибутов)
- Цель:
  - Найти зависимости между значениями атрибутов
  - Найти аналитические зависимости между атрибутами и выявить скрытые признаки и характеристики
- Приложения:
  - Маркетинг и рекомендательные системы - анализ зависимостей между покупаемыми товарами или услугами
  - Финансовый анализ – поиск зависимостей между значениями индексов и другими финансовыми параметрами
  - Латентно-семантический анализ текстов

Кластеризация

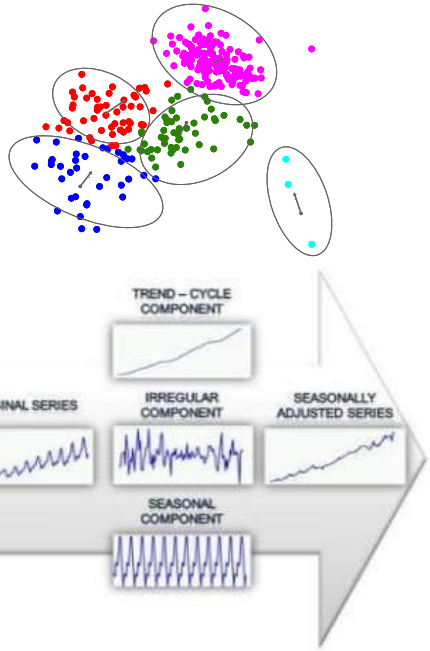
- Дано:
  - «не размеченный» тренировочный набор – для каждого объекта известны только значения его свойств (атрибутов)
- Цель:
  - Найти «непохожие» группы «похожих» объектов
- Приложения:
  - Маркетинг – сегментация клиентов, товаров и т.д.
  - Производство – выявление типовых состояний и ситуаций
  - Индексирование документов

Выявление исключений

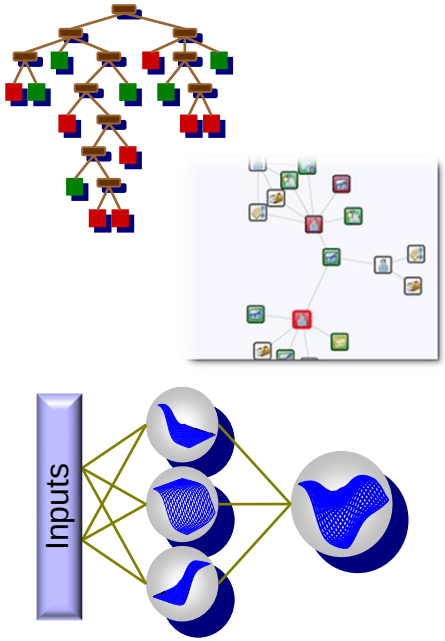
- Дано:
  - тренировочный набор («размеченный» или нет) – для каждого объекта известны значения его свойств
- Цель:
  - Построить модель и найти наиболее «не типичные» объекты
- Приложения:
  - Безопасность – подозрительные финансовые транзакции, звонки, люди, организации
  - Производство – выявление нештатных ситуаций
  - Медицина – диагностика

Жизненный цикл аналитических моделей

Выявление зависимостей



Построение моделей



Внедрение моделей

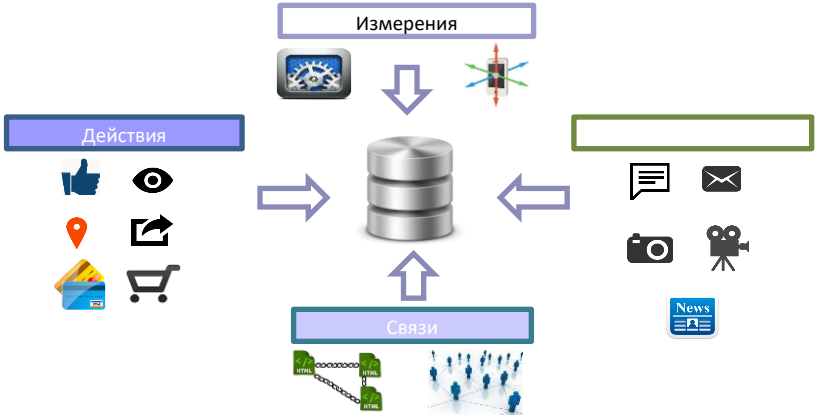
Большие данные

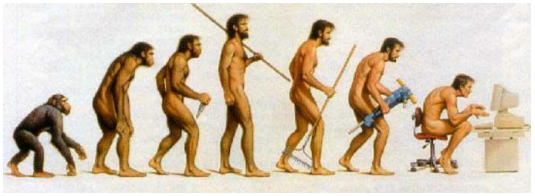
В научной среде термин используется с 1990х

(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!

Volume (объем)	Velocity (скорость)	Variety (разнообразие)	Veracity (достоверность)
Большие объемы данных.	Быстрота поступления и обновления данных.	Разнообразные типы данных: текстовые, графы, медиа и другие	Низкое качество, недостоверность анализируемых данных.

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

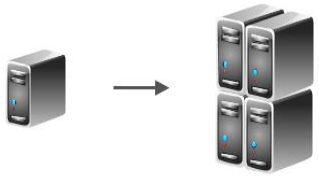
Что делать?

Вертикальное масштабирование:

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

Горизонтальное масштабирование:

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



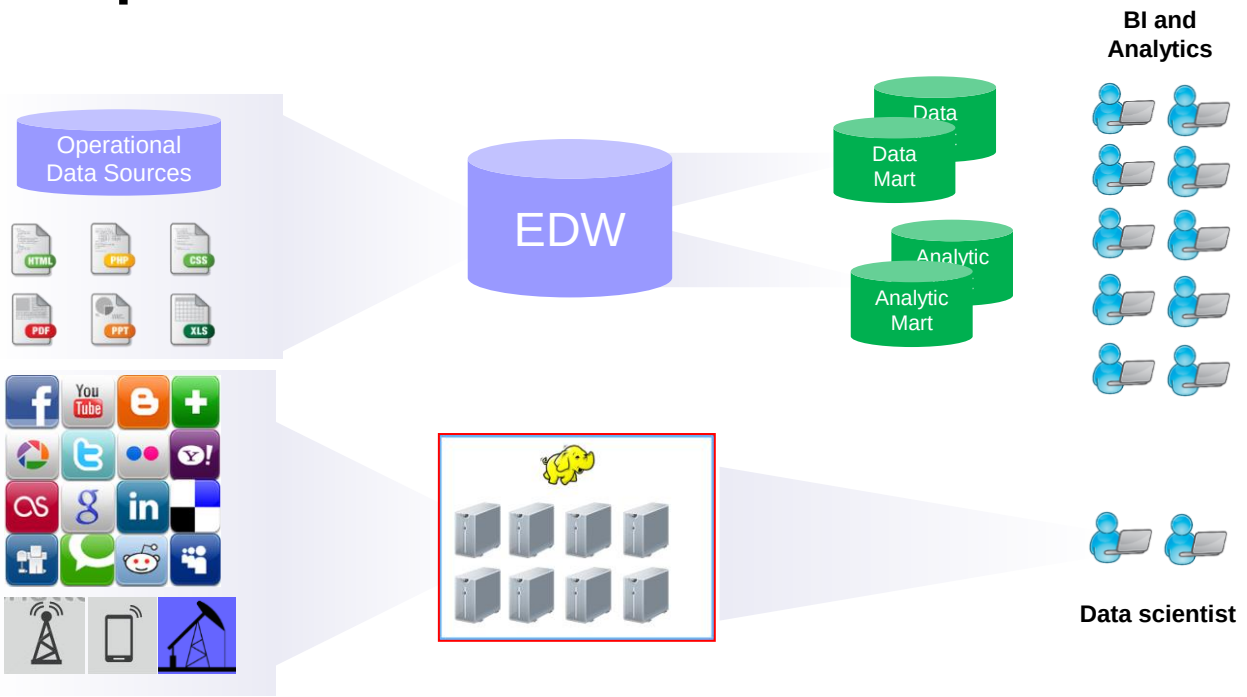
Индустрия выбирает MPP, а «математики» к этому не готовы

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

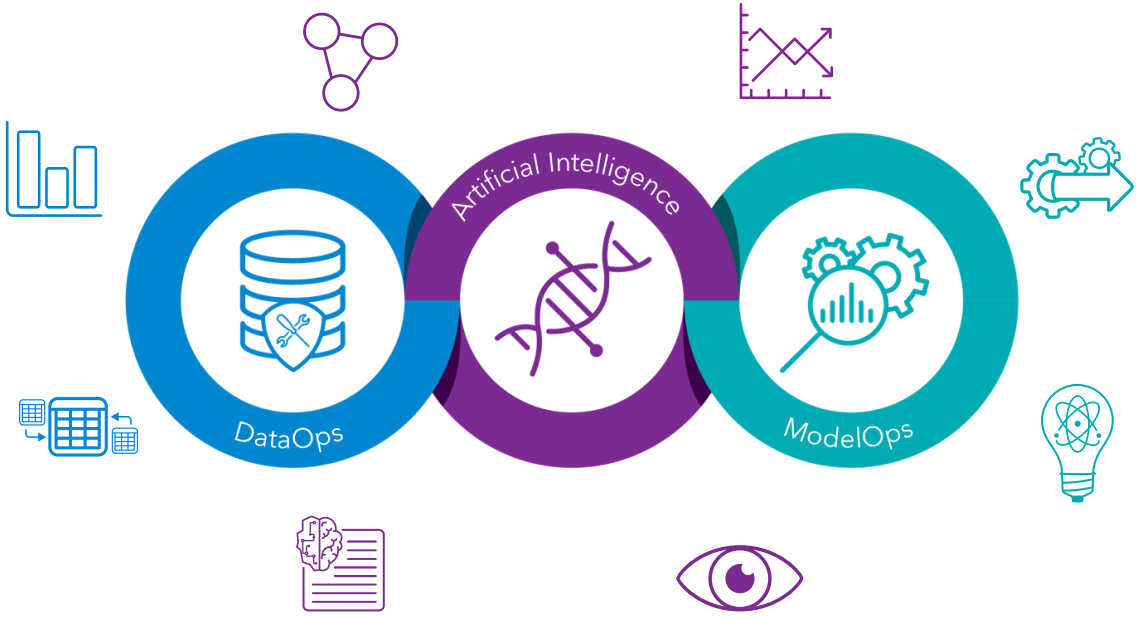


Современный подход к организации жизненного цикла аналитических моделей

Статистика, ML/DL

Прогнозирование и ортимизация

Визуализация данных



Внедрение моделей

Принятие решений

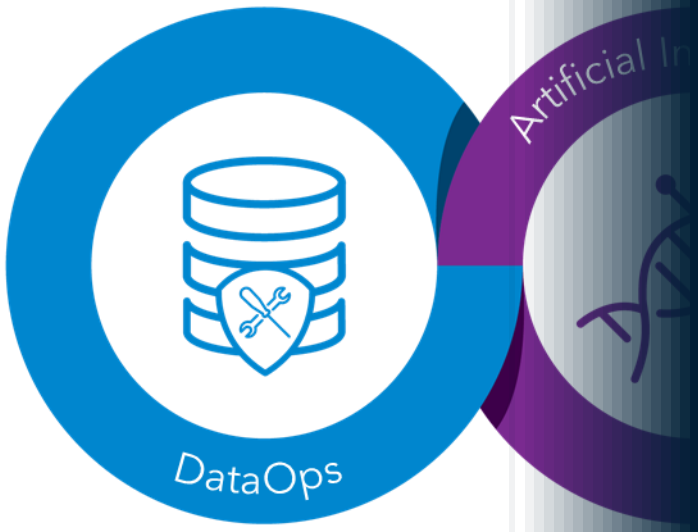
Управление данными

Обработка текстов на естественном языке

Компьютерное зрение

DATAOPS

Заимствуя методы Agile разработки программного обеспечения, DataOps обеспечивает гибкий подход к организации доступа к данным, управлению их качеством, и визуализации. Это обеспечивает большую надежность, адаптируемость, скорость и совместную работу в ваших усилиях по внедрению данных и аналитических рабочих процессов.



Доступ
Организация эффективного доступа к данным любого объема и структуры

Подготовка
Преобразование сырых данных в том числе с использованием AI

Визуализация
Выявление и наглядное представление основных зависимостей в данных

Управление
Построение хранилища очищенных и доверенных данных с учетом истории пополнения

Моделирование

Специалисты по обработке данных используют комбинацию методов для анализа данных и построения прогнозных моделей. Они используют статистику, машинное обучение, глубокое обучение, обработку естественного языка, компьютерное зрение, прогнозирование, оптимизацию и другие методы, чтобы решать реальные задачи.

Моделирование
Построение моделей с использованием различных методов ИИ для решения реальных задач

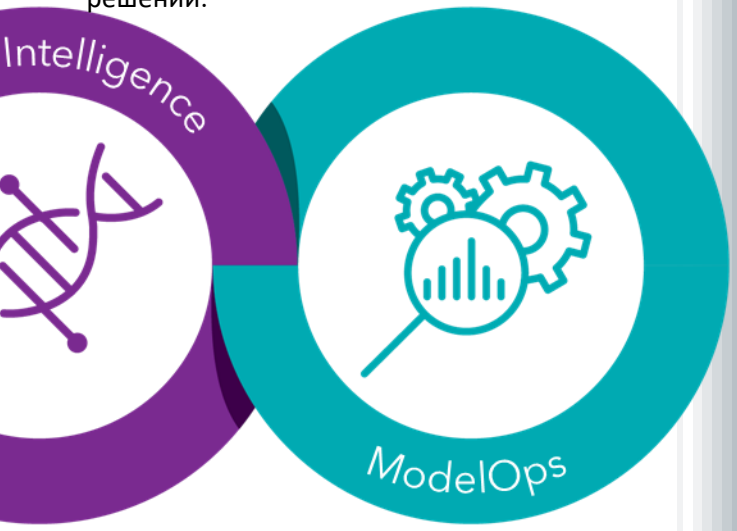
Автоматизация
Автоматизация рутинных задач по формированию признакового пространства и тьюнингу моделей

Взаимодействие
Групповая разработка моделей

Интеграция
Совмещение возможностей разных платформ

MODELOPS

ModelOps фокусируется на том, чтобы как можно быстрее получить модели ИИ через этапы проверки, тестирования и развертывания, обеспечивая при этом качественные результаты. Он также основан на постоянном мониторинге, дообучении и управлении моделями для обеспечения максимальной производительности и прозрачности решений.



Валидация
Объективная оценка качества моделей моделей

Внедрение
Внедрение моделей в операционные процессы и организация их мониторинга

Управление
Подтверждение надежности, достоверности и безопасности решений на основе ИИ моделей

Интеграция
Комбинация бизнес-правил и ИИ для принятие решений в режиме близком к реальному времени

Использование ИИ на разных стадиях

ИИ в DataOps

Оценка и исправление качества данных, обеспечение безопасности, жизненный цикл данных

ИИ в Моделировании

Построение, сравнение и оценка качества моделей, интерпретируемость моделей

ИИ в DataOps

Мониторинг производительности моделей, поддержка принятия надежных и достоверных решений

Отличия ИАД систем (1)

- Наличие «обучения»
  - база знаний формируются на основе анализируемых данных, а не экспертных знаний (в отличие от традиционных экспертных систем и систем информационного поиска)
  - структура модели и искомые зависимости заранее не известны (в отличие от статистических пакетов, ориентированных на расчет статистик, проверку гипотез и оценку параметров распределений)

Отличия ИАД систем (2)

- Наличие большого объема данных сложной структуры
  - зачастую скорость работы алгоритмов в ИАД важнее отклонений по точности (“quick and dirty solution”)
  - большинство алгоритмов работают с исходными данными в виде числовой матрицы признаков, сложная структура реальных объектов в ИАД приводит к необходимости решать задачу построения пространства характеристик и отображения в него свойств исходных объектов
  - перечисленные особенности отличают ИАД системы от традиционных систем машинного обучения, в которых, как правило, решается обратная задача – построение достоверной модели в условиях малой обучающей выборки

Отличия ИАД систем (3)

- Наличие человека - аналитика как оконечного потребителя результатов работы ИАД системы
  - в сценарии работы любой системы ИАД всегда присутствует аналитик, даже если полученная в результате модель далее используется для автоматической классификации
  - аналитик формирует тренировочные наборы, производит настройку алгоритмов, обучение и дообучение, анализирует полученные модели и принимает решения об их дальнейшем использовании
  - таким образом, системы автоматической классификации, кластеризации и распознавания образов, даже использующие возможность дообучения, не являются системами ИАД

Литература

http://www-stat.stanford.edu/~tibs/ElemStatLearn