Машинное обучение

к.ф.-м.н. Михаил Игоревич Петровский
michael@cs.msu.ru

Задачи курса

- Познакомить слушателей с предметной областью:
  - дать основные определения и терминологию, обсудить прикладные задачи даже тем, кто не планирует в этой области специализироваться дальше
- Рассмотреть базовые методы машинного обучения для решения типовых задач:
  - Первый семестр – акцент на обучение с учителем
  - Второй семестр – обучение без учителя
  - Меньше теории, больше алгоритмов и понимания как их настраивать и использовать на практике
  - Нейросети, работа с данными, визуализация, а тем более операционализация моделей – отдельная история, в основном за рамками данного курса
- Дать практический опыт решения задач машинного обучения:
  - на Питоне с использованием стандартных библиотек
  - у третьего больше акцент на программирование, включая демо-примеры на лекциях

Лекция 1: Введение

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**»

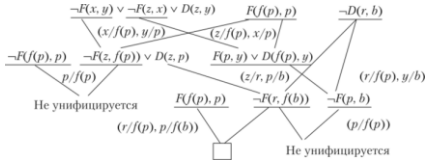
Почему «думать» и «делать» это разные области в ИИ?

«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

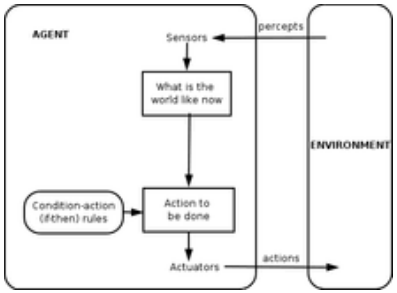


«Действовать» - взаимодействовать с окружающей средой (**интеллектуальный агент)**

- 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)

Всегда ли **нерационально** значит **плохо? Нет!** (история про джинна, проблема «здравого смысла»)

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** (пожалуй кроме трансформеров)

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 …» быстро сокращался
- Разработаны «универсальные» решатели (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
- 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 и связь с ML и Data Science

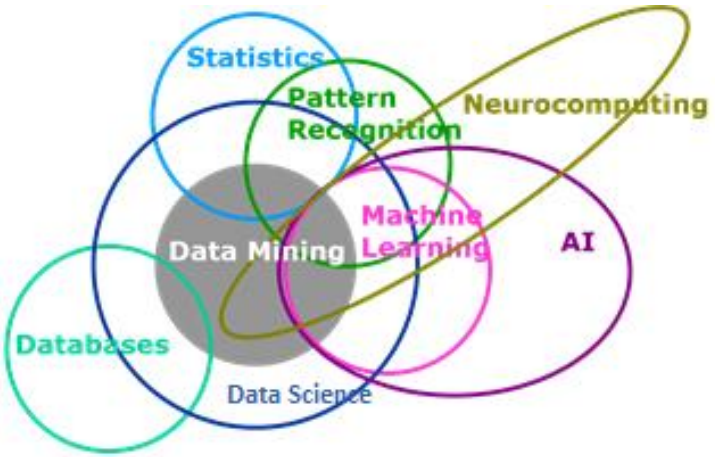
Относительный застой в теории - ничего принципиально нового уже больше 20 лет

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 (**неправильно!!!**) 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**.



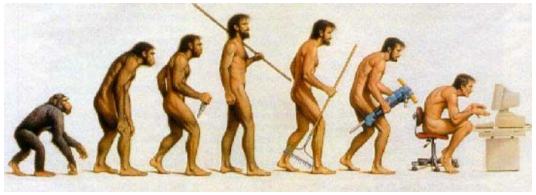
Большие данные

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

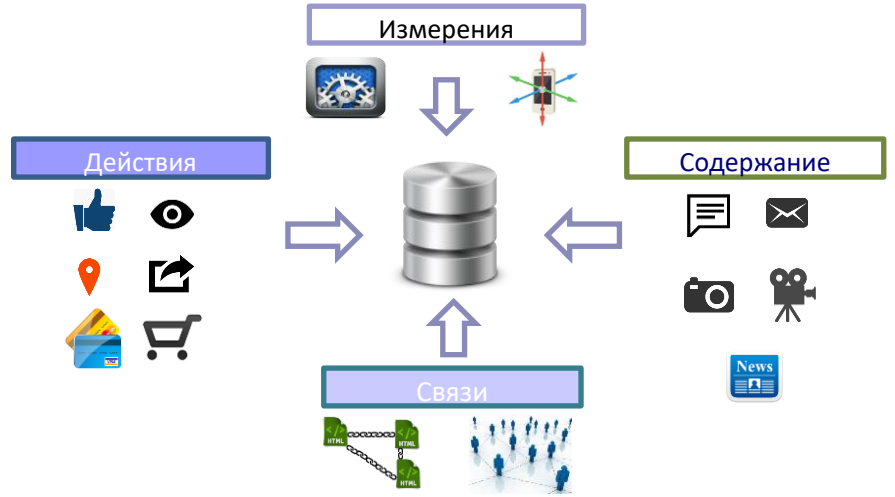
(2015) – термин Data Science



Начало цивилизации — 2003

5 экзабайт

20+ экзабайт в сутки!



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

Кто виноват и что делать с Большими данными?

50ГБ/сек

1x

ы жесткие

1ГБ/сек
(10Гб/сек)

50x

500МБ/сек

100x

166x

0,3ГБ/сек

100МБ/сек

500x

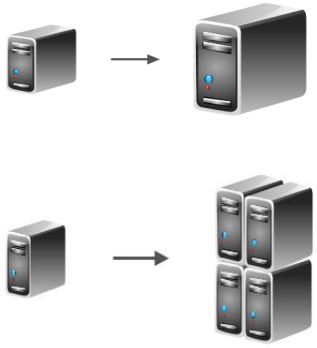
Что делать?

Вертикальное масштабирование:

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

Горизонтальное масштабирование:

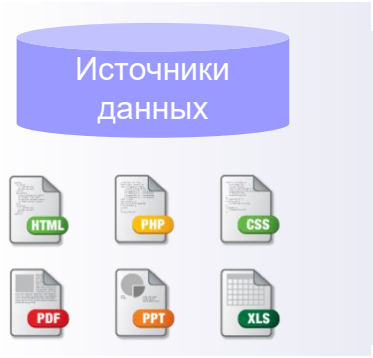
- дешево, потенциально технологически неограниченно
- НО сложно переносить аналитические алгоритмы



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

Роль человека в аналитике больших данных

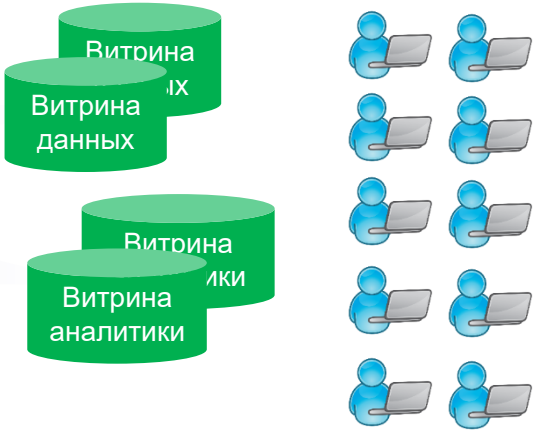
До эры больших данных:



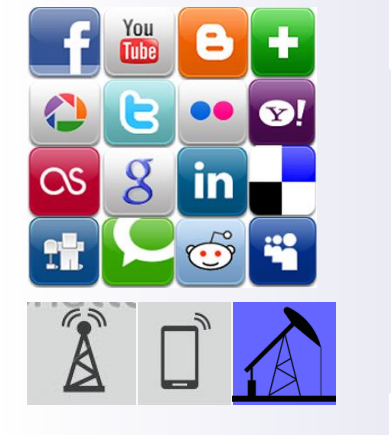
Корпоративное хранилище данных

Высококачественные надежные данные, последовательные, актуальные

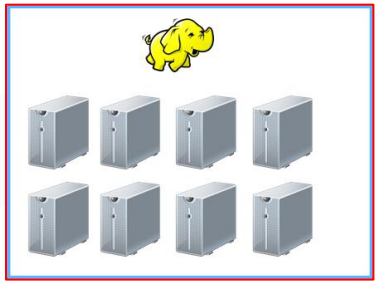
Традиционные аналитики



Сейчас:



Хранилище больших данных



Хранение “as is”

Data scientists:
Математики +
Программисты +
аналитики-прикладники

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
- Методы текстовой аналитики, включая выявления ключевых слов и скрытых тематик, аннотирования текстов, ответы на вопросы, чат-боты, обученные на больших корпусах (например, Wikipedia) работают все лучше, а используют лингвистику все меньше (или вообще не используют), например, 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

Современная индустрия ИИ и Больших данных

A heat map shows the relative ease of capturing the value potential across sectors

Top quintile (easiest to capture)
2nd quintile
3rd quintile
4th quintile
Bottom quintile (most difficult) to capture)
No data available

Cate-gories	Sectors	Overall ease of capture index[1]	Talent	IT intensity	Data-driven mind-set	Data availability
Goods	Manufacturing					
	Construction					
	Natural resources					
	Computer and electronic products					
	Real estate, rental, and leasing					
	Wholesale trade					
	Information					
Services	Transportation and warehousing					
	Retail trade					
	Administrative, support, waste management, and remediation services					
	Accommodation and food services					
	Other services (except public administration)					
	Arts, entertainment, and recreation					
	Finance and Insurance					
	Professional, scientific, and technical services					
	Management of companies and enterprises					
Regulated and public	Government					
	Educational services					
	Health care and social assistance					
	Utilities					

1 See appendix for detailed definitions and metrics used for each of the criteria.
SOURCE: McKinsey Global Institute analysis

Source: Gartner (March 2021)

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

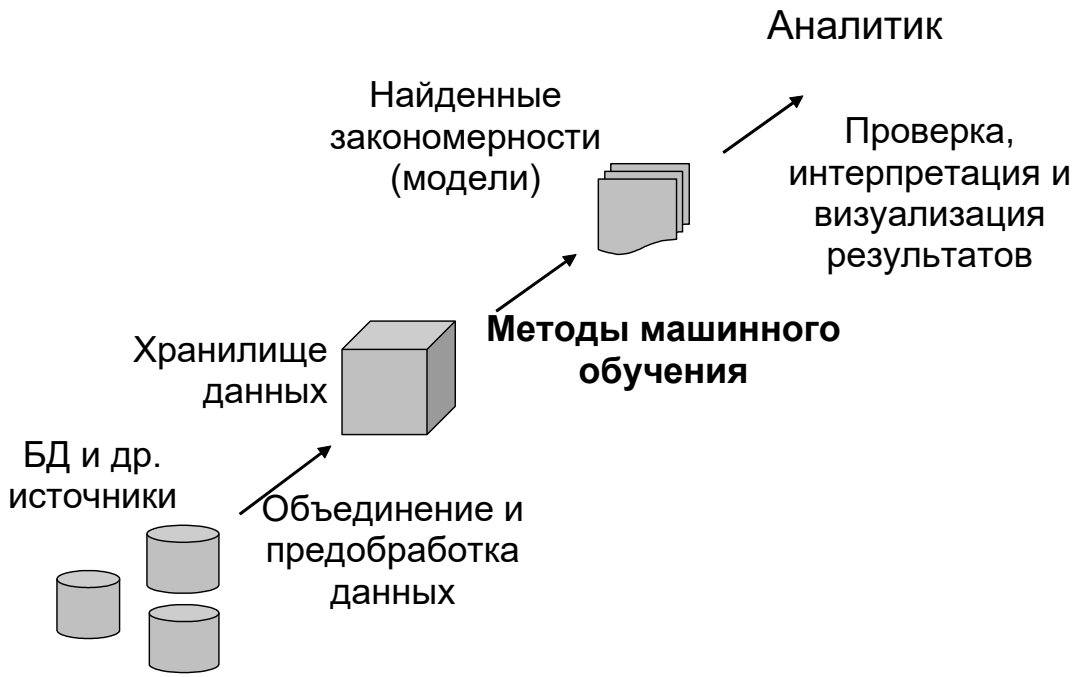
Эволюция технологий хранения и обработки данных

- … — 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

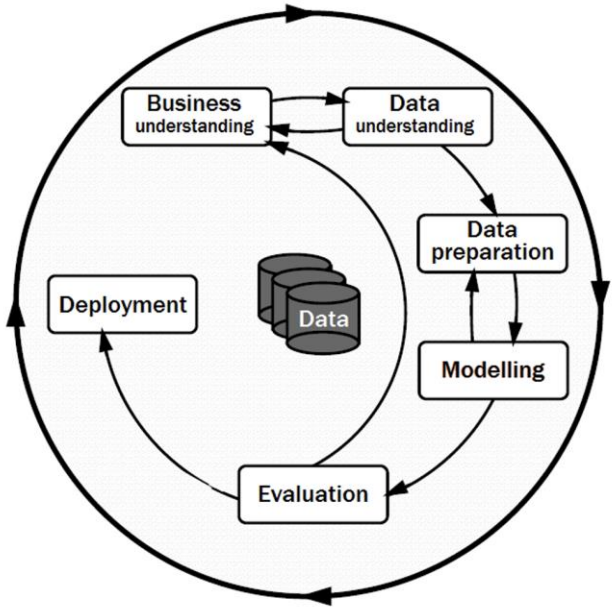
Актуальность и необходимость интеллектуального анализа данных

- 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



CRISP-DM: CRoss Industry Standard Process for Data Mining (1999)



Системы *интеллектуального анализа данных* – 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

Процесс интеллектуального анализа данных

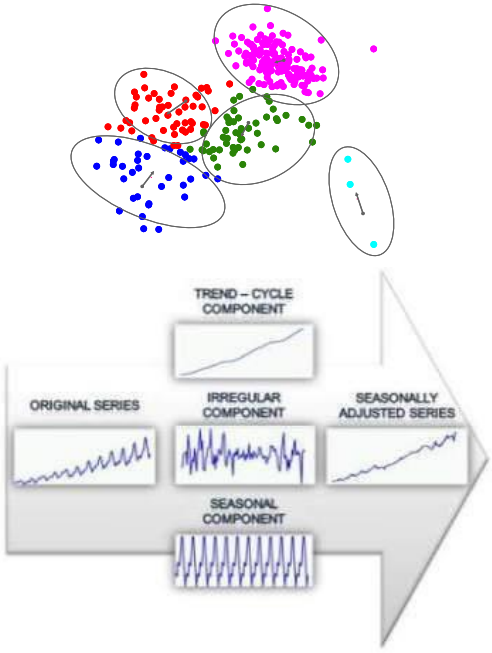
- Анализ предметной области:
  - выявление и формулировка необходимых априорных знаний о предметной области, целей анализа, задач приложения, сценариев использования
- Формирование и подготовка данных для анализа:
  - поиск (или выбор) «сырых» данных, возможно, реализация подсистемы сбора (консолидации)
  - 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 интеллектуального анализа данных

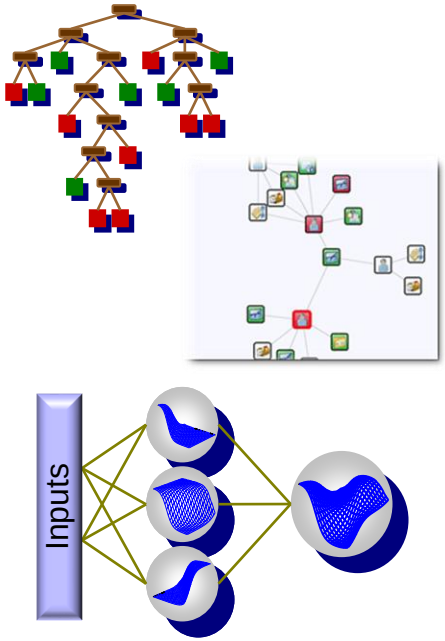
- Выбор (или разработка) алгоритма машинного обучения:
  - определение ограничений и требований к алгоритму по точности, размеру, интерпретируемости, скорости построения и применения получаемых моделей, по типу исходных данных
- Непосредственно построение моделей:
  - применение выбранного алгоритма 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

Жизненный цикл аналитических моделей

Выявление зависимостей



Построение моделей



Внедрение моделей

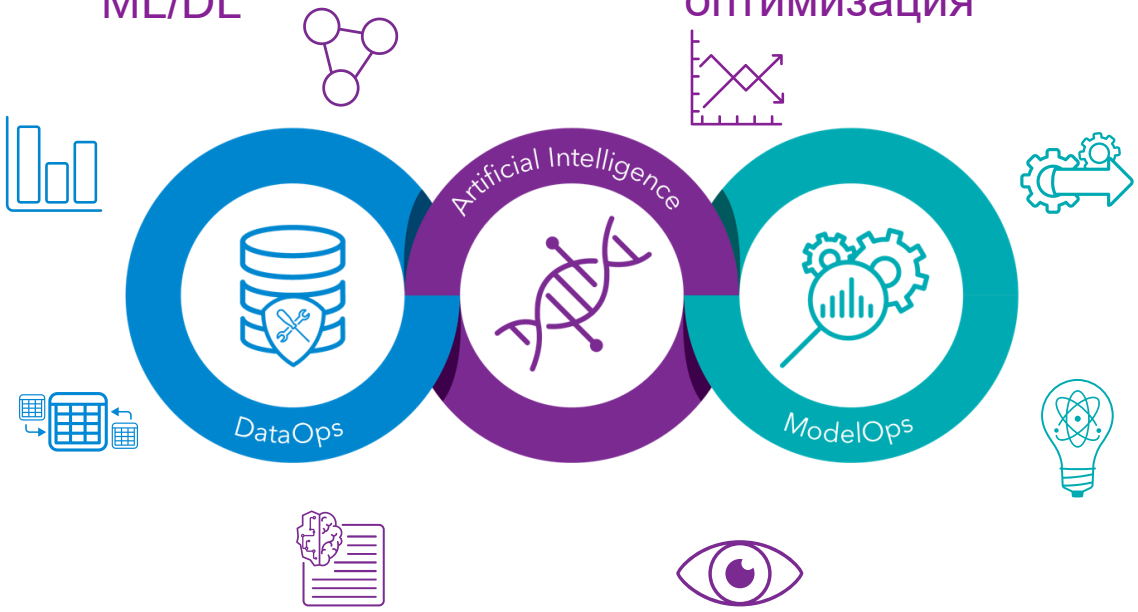
Современный подход к организации жизненного цикла аналитических моделей

Визуализация данных

Статистика, ML/DL

Прогнозирование и оптимизация

Внедрение моделей



Управление данными

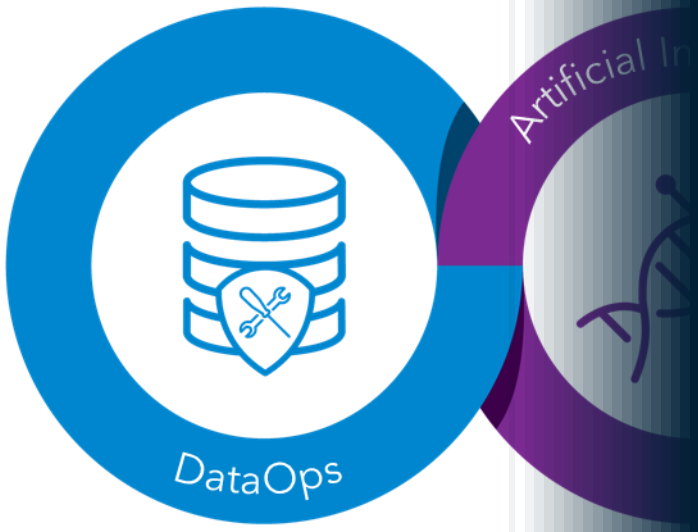
Принятие решений

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 (применение МО)

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 машинного обучения для решения реальных задач

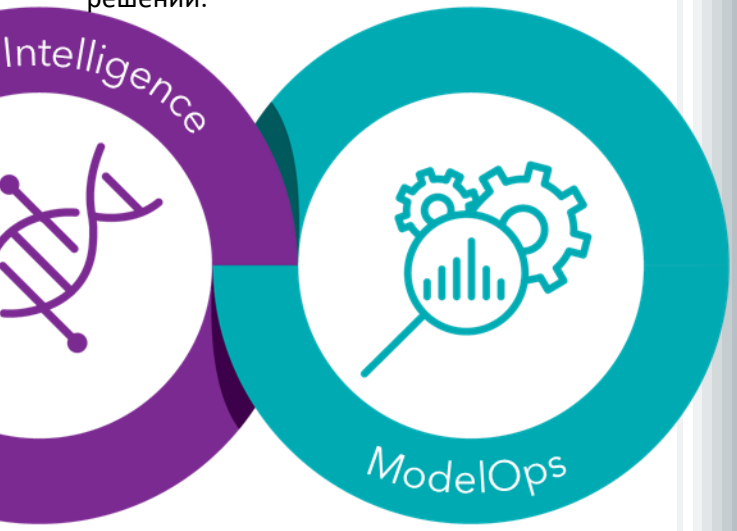
Автоматизация
Автоматизация рутинных задач по формированию признакового пространства и тьюнингу моделей

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

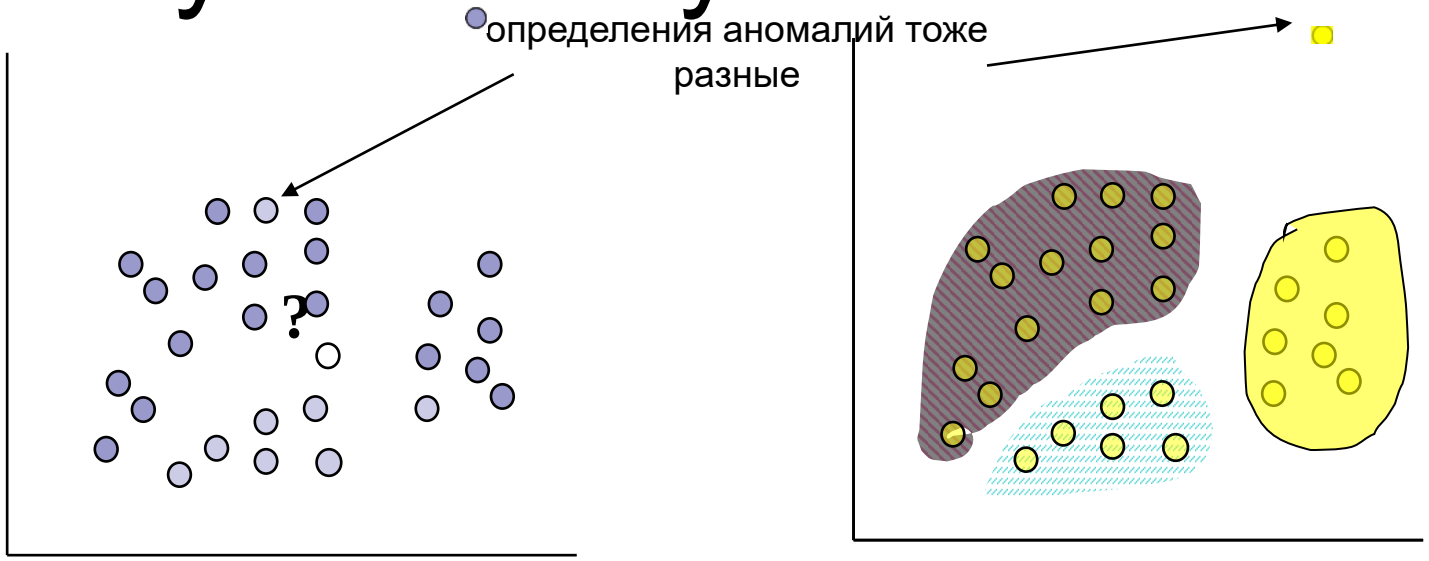
Управление
Подтверждение надежности, достоверности и безопасности решений на основе ИИ моделей

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

Основные типы исходных данных

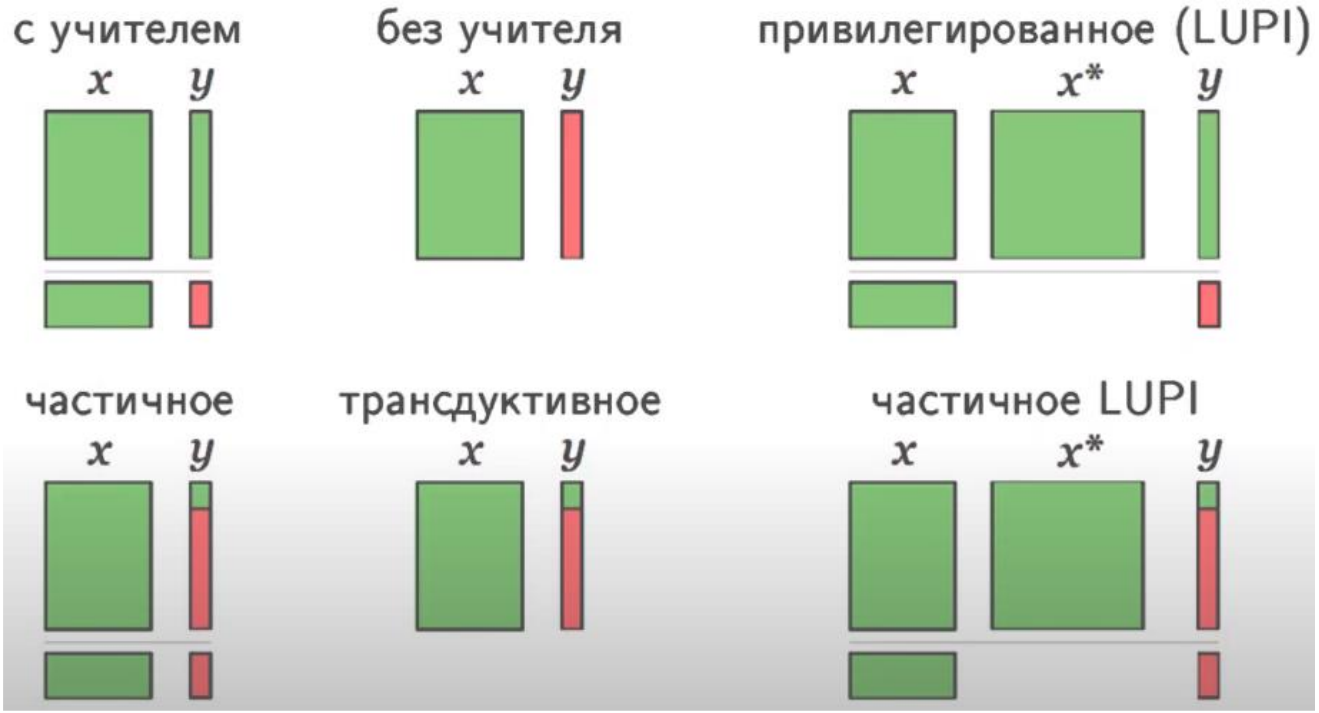
- Транзакционные
  - Объекты анализа – «события» различной структуры с числовыми и категориальными атрибутами и с временной меткой
- Табличные
  - Объекты анализа представлены в виде реляционных таблиц, возможно взаимосвязанных (заданно ER-схемой), имеют разнотипные атрибуты
- Временные ряды и числовые 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

Обучение с учителем и без



- «Размеченный» набор данных – 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, …)

Типы задач обучения в зависимости от доступности разметки



- Трансдуктивное обучение – тестовая выборка известна заранее
- Привилегированное обучение – часть признаков известна только на этапе обучения

Базовые задачи машинного обучения = 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:
  - Компьютерная безопасность
  - Производство - прогнозирование качества изделий
  - Распознавание образов

Ранжирование

- Дано:
  - «размеченный» тренировочный набор – для каждого объекта известен его класс или несколько не обязательно 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 - рубрикация документов
  - Кредитование - оценка рисков

Прогнозирование/регрессия

- 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

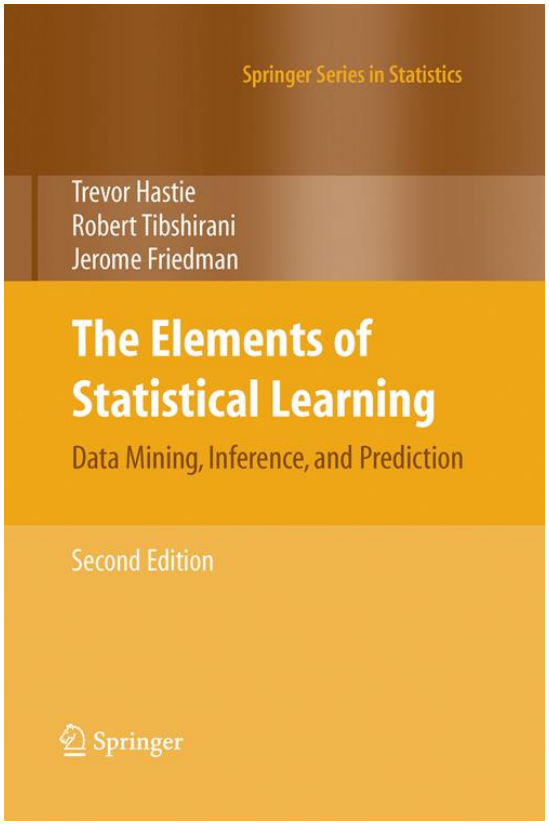
Кластеризация

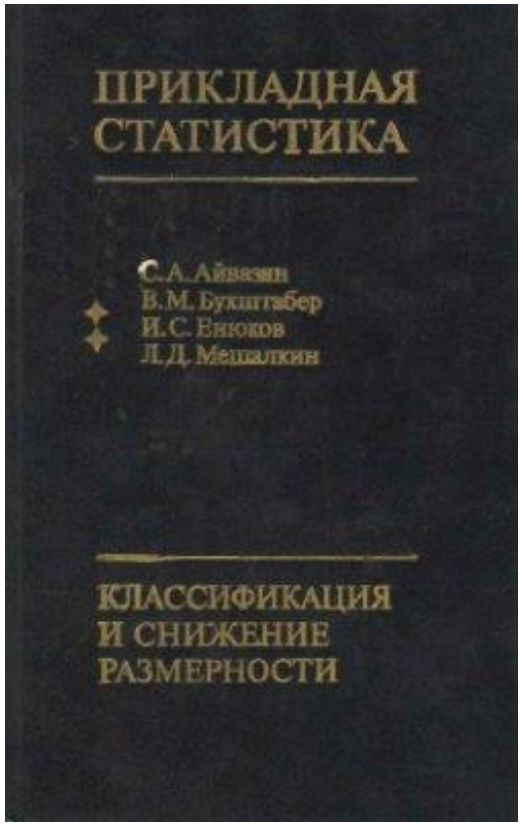
- Дано:
  - «не размеченный» тренировочный набор – для каждого объекта известны только значения его свойств (атрибутов)
- Цель:
  - Найти «непохожие» группы «похожих» объектов
- Приложения:
  - Маркетинг – сегментация клиентов, товаров и т.д.
  - Производство – выявление типовых состояний и ситуаций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

Литература и полезные материалы





github.com/MSU-ML-COURSE/ML-COURSE-22-23
http://www.machinelearning.ru/wiki/
http://www-stat.stanford.edu/~tibs/ElemStatLearn
https://t.me/+rfMxWm1fSbRmOTcy