

Межфакультетский курс “Большие Данные и машинное обучение” Вводная лекция

Миронов Андрей Михайлович

Кафедра Математической Теории Интеллектуальных Систем
Механико-математический факультет
Московский Государственный Университет имени М.В.Ломоносова

Москва, 12 сентября 2018 г.

Понятие Больших Данных (Big Data)

Термин Big Data (Большие Данные) появился в 2008 г. (статья Клиффорда Линча в журнале Nature).

Lynch C. Big data: how do your data grow? // Nature. 2008. Vol. 455. N 7209. P. 28-29.

Большими Данными принято называть огромные массивы информации со сложной неоднородной и неопределенной структурой.

Большие Данные (как научная дисциплина) – совокупность подходов, инструментов и методов обработки структурированных и неструктурированных данных

- огромных объемов, и
- значительного многообразия

для получения воспринимаемых человеком результатов.

Глобальный феномен Больших Данных

Большие Данные (Big Data) стали глобальным феноменом всей современной жизни.

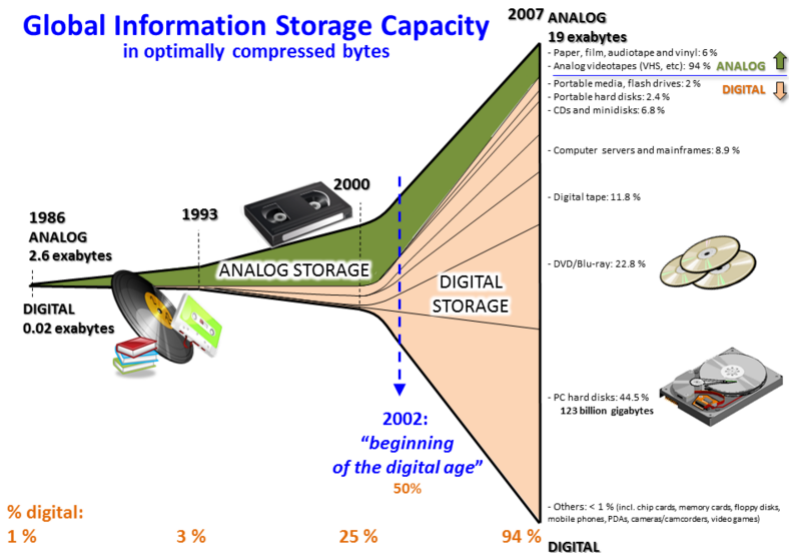
Цифровые технологии пронизали жизнь современного человека.

Объем данных о самых разных сторонах жизни растет в геометрической прогрессии, удваиваясь каждые 1,2 года, и одновременно растут возможности хранения информации.

Только в одном Twitter формируется почти 100 млн. коротких сообщений в сутки. Национальная администрация океана и атмосферы США формирует более 30 петабайт (10^{15}) данных в год.

Мировой потенциал для хранения, передачи и вычисления информации (динамика роста)

Global Information Storage Capacity in optimally compressed bytes



Мировой потенциал для хранения, передачи и вычисления информации (динамика роста)

2002 год явился переломным в изменении соотношения мирового объема аналоговых и цифровых данных в пользу последних, объем которых увеличивался в геометрической прогрессии (лавинообразно).

В 2007 году объем цифровых данных превысил объем аналоговых почти в 15 раз.

К 2020 году количество данных в мире достигнет 40 36 (zettabytes), что эквивалентно 5200 Гб на каждого жителя планеты.

Классификация объемов данных

- **Большие наборы данных:** от 1000 мегабайт (1 гигабайт) до сотен гигабайт (10^9)
- **Огромные наборы данных:** от 1000 гигабайт (1 терабайт) до нескольких терабайт
- **Big Data:** от нескольких терабайт до сотен терабайт
- **Extremely Big Data:** от 1000 до 10000 терабайт = от 1 до 10 петабайт (10^{15}).

Источники Больших Данных:

- социальные сети,
- мобильные устройства,
- данные с измерительных устройств,
- бизнес-информация, и т.д.

Основные сферы использования Больших Данных

- прогнозирование климата,
- создание новых материалов,
- расчет безопасности атомных электростанций
- моделирование газодинамического обтекания тел (самолетостроение, автомобилестроение)
- анализ данных геологоразведки,
- моделирование процесса нефтедобычи
- отслеживание состояния оборудования
- астрономия (наблюдение 40 миллиардов объектов)
- дистанционное зондирование земли, контроль окружающей среды (400 космических аппаратов, данные объемом 8 терабайт в день)

Основные сферы использования Больших Данных в экономике и банковской деятельности

- прогнозирование спроса и цены товара,
- клиентский сервис,
- управление рисками
- управление поставками товара, его хранением и продажами,
- оптимизация затрат,
- изучение кредитоспособности заемщика,
- операционная эффективность,

Основные сферы использования Больших Данных (продолжение)

- поисковые системы и веб-аналитика,
- биоинформатика и медицина:
 - ▶ изучение генома человека - 3300 петабайт (10^{15}),
 - ▶ предсказание медицинских рисков на основе данных об активности генов,
 - ▶ построение искусственных белков с заданной функциональностью (по команде выполнить некоторое действие)
 - ▶ компьютерное проектирование лекарств
- нейронаука (изучение мозга): альтернативные способы изучения мозга, новые теоретические модели кодирования информации, объединение гуманитарных и естественных наук

Основные сферы использования Больших Данных (продолжение)

- Большой адронный коллайдер (большая международная коллаборация в которой принимают участие физики МГУ).

В год датчики коллайдера дают 150 зеттабайт (10^{21} байтов) информации - это больше, чем от всех остальных источников данных в мире.

Именно там на основе методов анализа Больших Данных был открыт бозон Хиггса, за что в 2013 г. была присуждена Нобелевская премия.

Отличительные особенности Больших Данных

Основные характеристики, соответствие которым позволяет относить данные к категории больших:

- **Volume:** накопленная база данных (БД) является слишком большой для обработки и хранения традиционными способами
- **Velocity:** скорость накопления и обработки данных превышает возможности типичных БД по занесению, хранению, управлению и анализу информации
- **Variety:** большое разнообразие структурированных и неструктурированных данных.

Другие характеристики Больших Данных

Другие характеристики Больших Данных:

- **Veracity** (достоверность)
- **Variability** (переменчивость)
- **Value** (ценность)
- **Visualization** (визуализация)
- **Viability** (жизнеспособность)

Неадекватное использование Больших Данных

Некоторые эксперты сравнивают системы накопления, обработки и хранения Больших Данных с кораблем, тянущим за собой гигантскую сеть и собирающим все, что в нее попадает.

Неадекватное использование Больших Данных может привести к неадекватной сложности при их обработке и неадекватному пониманию реальности.

Например, достоверный список потенциальных покупателей товаров, демографическая информация, и т.п. может оказаться гораздо полезнее, чем огромная БД с данными о количестве посещений интернет-сайта компании.

Три эпохи работы с данными:

- 1 эпоха - классическая пара: эксперимент и теория.
Ручные вычисления.
- 2 эпоха - компьютерное моделирование
(например моделирование взрыва атомной и водородной бомбы).
- 3 эпоха - работа с Большими Данными.
Тут нужна новая математика (включающая математические модели гуманитарных явлений и процессов).

Data Science (наука о данных) - изучение проблем хранения, обработки и передачи больших данных.

Основные математические направления, связанные с Большими Данными

Математические направления, связанные с Большими Данными:

- хранение и оптимальное представление данных
- обработка и извлечение знаний из данных
- высокоскоростная передача и безопасность данных

- распределенные хранилища информации,
- облачные сервисы,
- модели и алгоритмы
 - ▶ автоматизированного контроля качества данных,
 - ▶ своевременного выявления злонамеренных действий и сбоев,
 - ▶ оперативного нахождения причин их возникновения
- оптимальные архитектуры организации данных
- механизмы защиты целостности и конфиденциальности данных,

- агрегирование и фильтрация данных,
- анализ зашумленности данных,
- кластеризация и структуризация данных
- статистические методы машинного обучения,
- нейросетевые алгоритмы машинного обучения,

- кодирование
- протоколы передачи данных
- верификация и анализ производительности протоколов,
- методы криптографической защиты информации при ее передаче в небезопасных сетях связи

Наука об анализе Больших Данных

Главное содержание науки о Больших Данных – построение методов трансформации Больших Данных в простые алгоритмы принятия решений.

Анализ Больших Данных – это междисциплинарная деятельность, которая требует универсальных специалистов. Для анализа Больших Данных нужны фундаментальные знания в

- математике,
- статистике,
- информатике и программировании,
- бизнесе,
- экономике, и других областях.

Нужна новая математика.

По поводу новой математики

За 1992-2012 годы:

- компьютеры ускорились примерно в 8000 раз
- благодаря развитию математических идей алгоритмы расчета ускорились почти в 500 000 раз

Совместный эффект математики и компьютеров - ускорение в 4 миллиарда раз.

Есть прогноз, что возможно ускорение в 450 миллиардов раз.

Задача, которая решалась на компьютере в 91 году 126 лет, в 2012 г. решается за 1 секунду.