

«Базы данных»

МГУ им. Н.П. Огарева, 2019 год
(18 часов лекций + экзамен)

«Базы данных»

МГУ им. Н.П. Огарева, 2019 год
(18 часов лекций + экзамен)



Андрей Владимирович Попов

доцент кафедры фундаментальной
информатики

<https://andpop.ru>

Содержание

- Значение баз данных. История их развития.
- Основные теоретические понятия.
- Реляционные базы данных. SQL, нормализация, ORM.
- Базы данных NoSQL.

Лекция 1. Часть 1.

Базы данных: введение

Базы данных: введение

1. Значение и классификация данных.
2. Базы данных как часть технологий работы с данными.

Что такое данные?

Что такое данные?

Дословно: Data (англ. *данные*) = «данность», «факт».

ISO, 1996: Формы представления информации, с которыми имеют дело информационные системы и их пользователи.

ISO, 2015: Поддающееся многократной интерпретации представление информации в формализованном виде, пригодном для передачи, связи, или обработки.

Данные – факты, текст, графики, картинки, звуки, аналоговые или цифровые видео-сегменты, представленные в форме, пригодной для хранения, передачи и обработки.

Для чего нужны данные?

Для чего нужны данные?

Основная задача обработки данных:

- получение из данных **информации**
 - из информации – **знаний**
 - из знаний – **мудрости**

Для чего нужны данные?

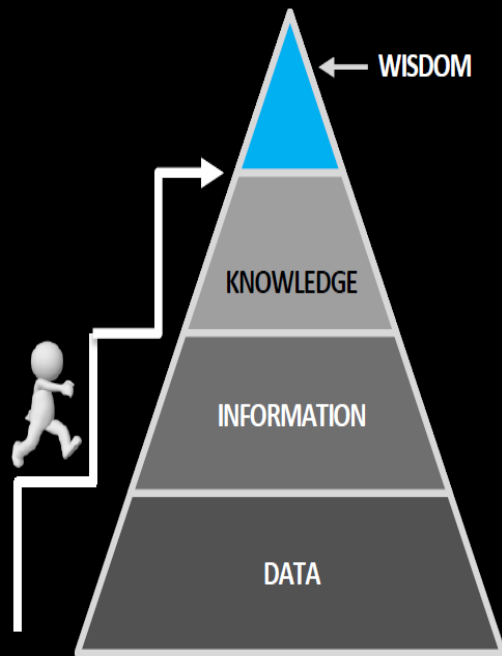
Основная задача обработки данных:

- получение из данных **информации**
 - из информации – **знаний**
 - из знаний – **мудрости**

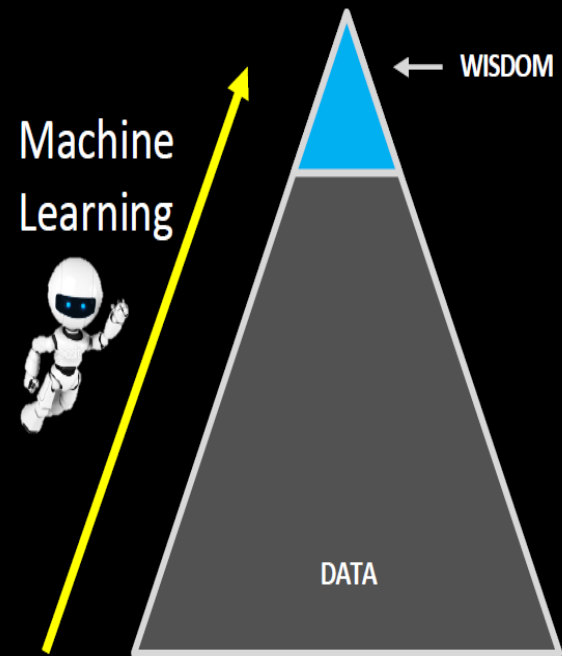
Знание – понимание того, ЧТО мы должны делать в определенный момент времени и КАК это нужно делать.

Мудрость – предвидение возможного развития событий.

До Big Data и Deep Learning



После Big Data и Deep Learning



С ростом объёма данных и появлением технологий Big Data, сократился путь от данных к мудрости за счет развития технологий машинного обучения

Где используются данные?

Процессы, использующие данные

Цель – замена человека машиной

- **Механизация** – замена физического труда человека машинным.

Процессы, использующие данные

Цель – замена человека машиной

- **Механизация** – замена физического труда человека машинным.
- **Автоматизация** – замена управленческого труда человека (машина реализует некоторую программу, алгоритм управления).

Процессы, использующие данные

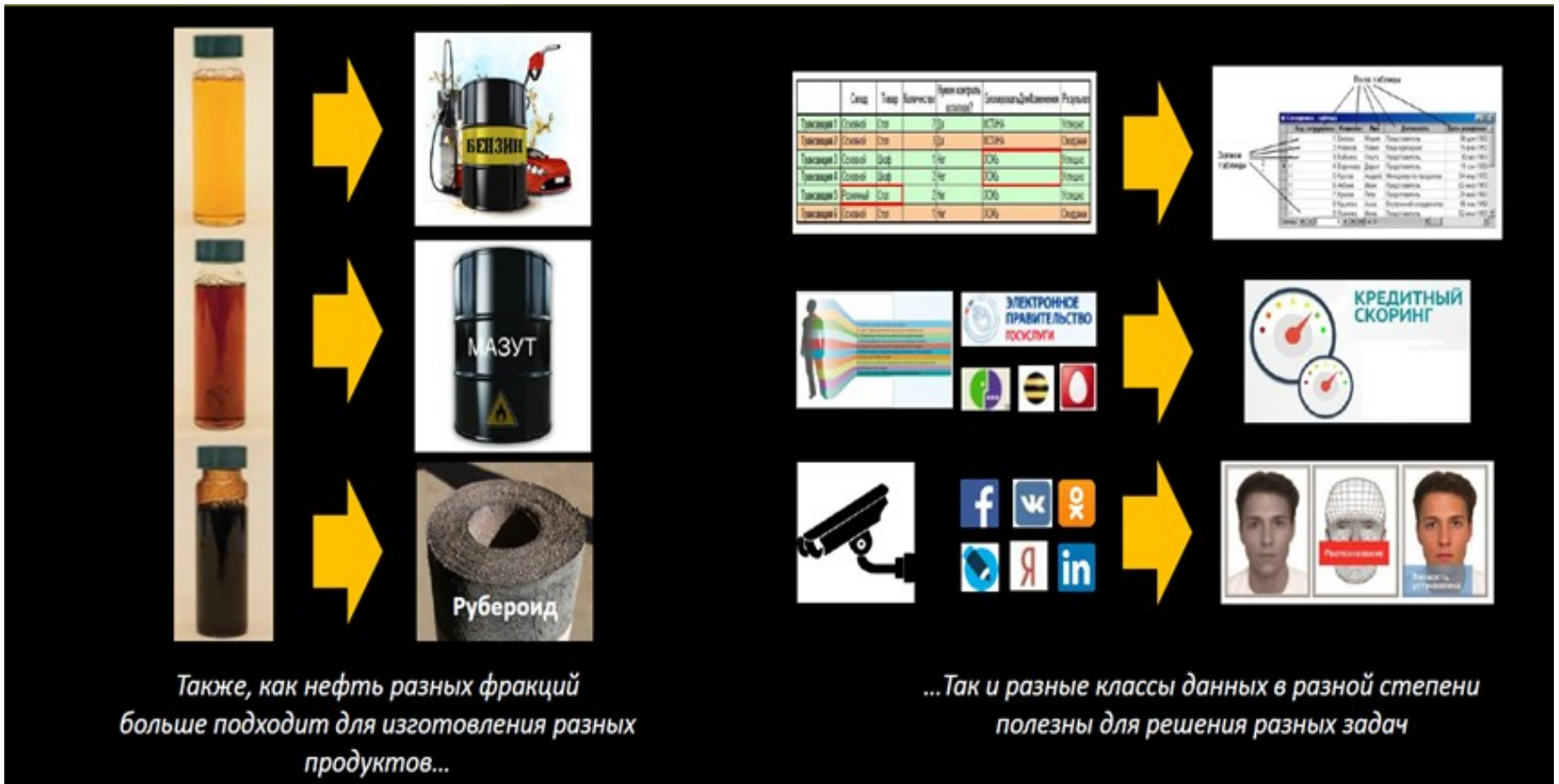
Цель – замена человека машиной

- **Механизация** – замена физического труда человека машинным.
- **Автоматизация** – замена управленческого труда человека (машина реализует некоторую программу, алгоритм управления).
- **Кибернетизация** – замена человека в сфере принятия решений.

Данные и **базы данных** используются в процессах автоматизации и кибернетизации.

Данные – новая нефть

Данные = сырье для создания ценности.
Разные данные – для разных задач.



**Какие задачи решаются с
помощью анализа данных?**

Задачи, решаемые с помощью анализа данных

Классические

- формирование отчетностей
- учетные системы
- описательная аналитика

Современные

- воспроизведение связей между событиями и результатом (machine learning)
- таргетированные коммуникации
- распознавание образов (лиц)
- классификация

Задачи, решаемые с помощью анализа данных

Классические

- формирование отчетностей
- учетные системы
- описательная аналитика

Современные

- воспроизведение связей между событиями и результатом (machine learning)
- таргетированные коммуникации
- распознавание образов (лиц)
- классификация

Данные – это важнейший актив, так как из них можно автоматически получать знание.

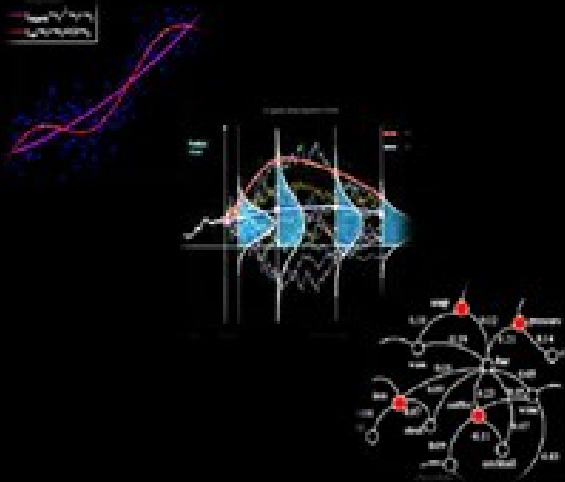
Знания из данных дают модели и аналитика

‘Интеллектуальный анализ данных (Data Mining) включает изучение больших объемов данных и поиска в них закономерностей с применением статистических методов, искусственного интеллекта, а также некоторых технологий управления базами данных’



Машинное обучение

Machine Learning – способы воспроизведения **связей** между событиями и результатом



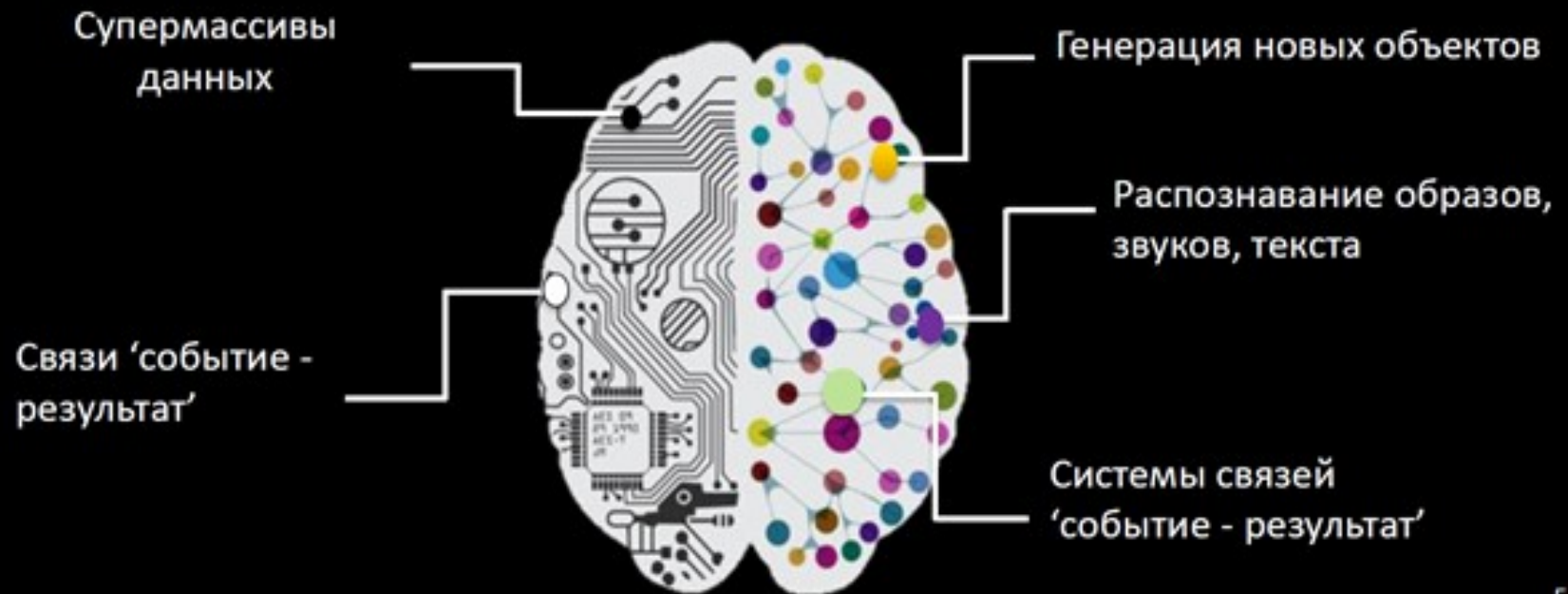
Алгоритмы

+

Технологии

Глубинное обучение

Deep Learning – способы воспроизведения **системы связей** между событиями и результатом, в том числе **скрытых связей**

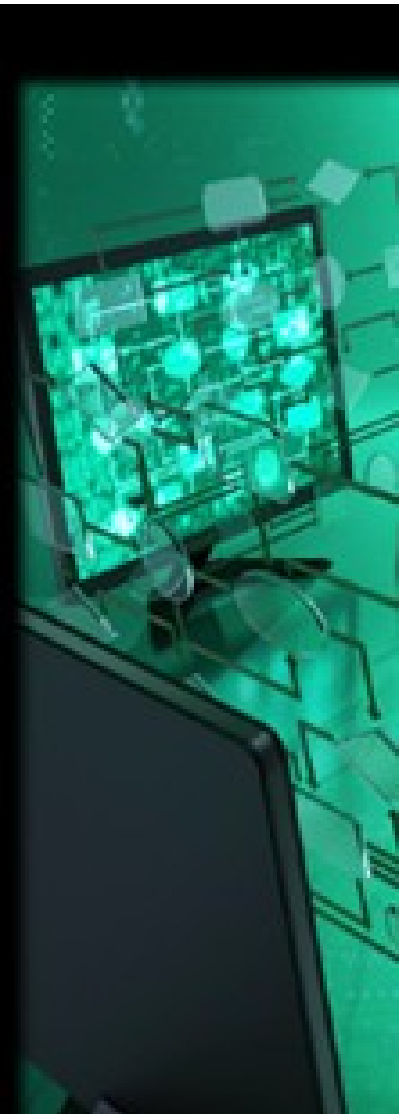


Данные без технологий - **бесполезны**

‘Сегодня углубленная аналитика применяется все чаще.

Рост вычислительных мощностей, улучшенная инфраструктура данных, появление новых алгоритмов создали для этого технические возможности, а с резким увеличением объемов данных их качественный детальный анализ превратился в средство обеспечения значительных конкурентных преимуществ’

Алан Ньюджент, Марсия Кауфман, Джудит Гурвиц, Ферн Халпер «Просто о больших данных»



Задачи технологий работы с данными



Базы данных и СУБД – часть технологий для работы с данными.

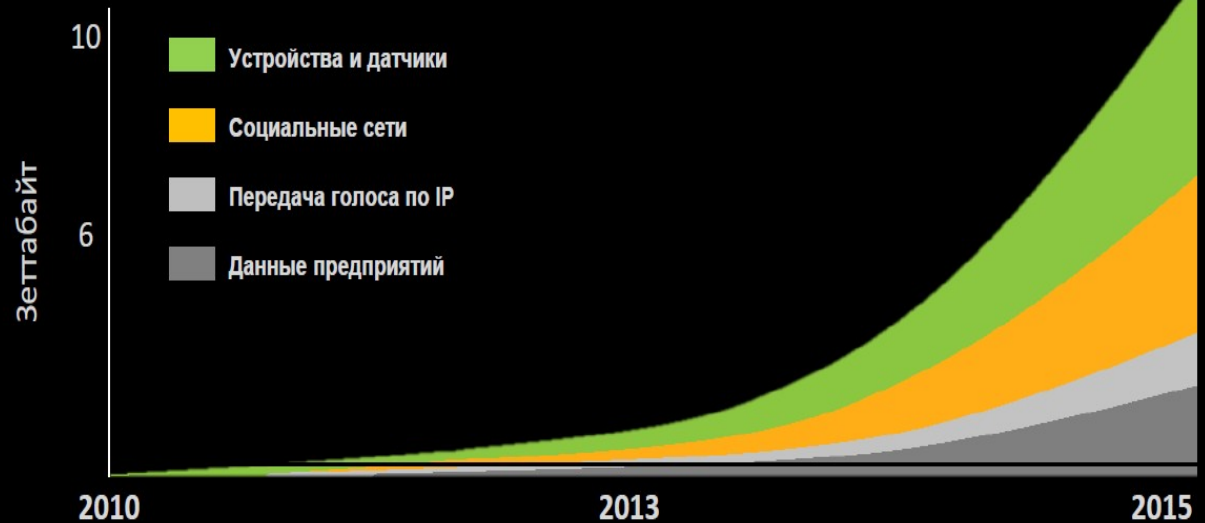
Используются на этапах **загрузки, сохранения и обработки данных.**

Откуда берутся данные?

Откуда берутся данные?

1

Развитие интернета, социальных сетей, мобильных устройств и M2M стимулировало огромный рост объемов и разнообразия типов данных



2

Каждый человек и каждое его действие создают данные

Финансовые операции



Мобильные устройства



Социальные сети



Государственные учреждения



Аудио и видео



«Умные» вещи



Классификация данных

Тип носителя	Цифровые	Аналоговые	Цифровой или аналоговый носитель данных
Место генерации	Внутренние	Внешние	Данные формируются внутри или вовне организации
Способ генерации	Машина	Человек	Данные создаются машиной или человеком
Объем ¹	Малый	Большой	Гб/Тб. Малые - пригодны к обработке на ПК. Большие требуют решений класса Big Data ¹
Доступность ²	Низкая	Высокая	Показатель доступности
Качество	Низкое	Высокое	Показатель качества
Структурированность	Низкая	Высокая	Простота классификации данных по атрибутам
Однородность	Низкая	Высокая	Число различных форматов данных
Связность	Низкая	Высокая	Возможность сопоставить данные между собой



¹ Малые данные – те, на которых можно анализировать на современном ПК (до 16 Гб) (Том Андерсон, Большие данные – нужны ли они в маркетинговых исследованиях)

² Одни и те же данные могут быть доступными для одного пользователя и не доступными для другого

**Какое значение имеют
базы данных?**

Обработка данных: от сырья к результату



У данных много общего с нефтью, но, в отличие от неё, данные – неисчерпаемый ресурс

Обработка данных: от сырья к результату



Базы данных – это инструменты, средства достижения конечного результата (месторождения, буровые установки, насосы и нефтеочистительные заводы).

Развитие технологий баз данных

Год	Событие	Данные	Анали-тика	Техно-логии
1881	Первая машина, работающая на перфокартах		+	+
1928	Создана магнитная лента	+		+
1956	Первые магнитные диски (IBM)	+		+
1963	Чарльз Бахман разработал первую СУБД Integrated Data Store			+
1965	Первый дата-центр в США	+		+
1970	Реляционная модель данных			+
1976	Компьютер стал использоваться в повседневных целях	+		
1991	Появился Интернет. Анализ и загрузка данных онлайн	+		+
1997	Google запустил поисковую систему			+
2001	Google создал файловую систему GFS			+
2005	Apache создал Hadoop			+
2005	Появляются промышленные NoSQL СУБД			+
2010	За два дня создается столько данных, сколько люди создали с начала цивилизации до 2003 года	+		+
2014	80% директоров компаний заявляют, что анализ Big Data – высший приоритет		+	