Формализация информации и Big Data

http//vikchas.ru

Тема 2. Big Date

Лекция 3 «Большие данные и перспективы их развития»

Часовских Виктор Петрович
д-р техн. наук, профессор кафедры ШИиКМ

ФГБОУ ВО «Уральский государственный экономический университет»

Екатеринбург 2023

Данные как объект управления

ДАННЫЕ

Большие массивы цифровых структурированных и неструктурированных данных

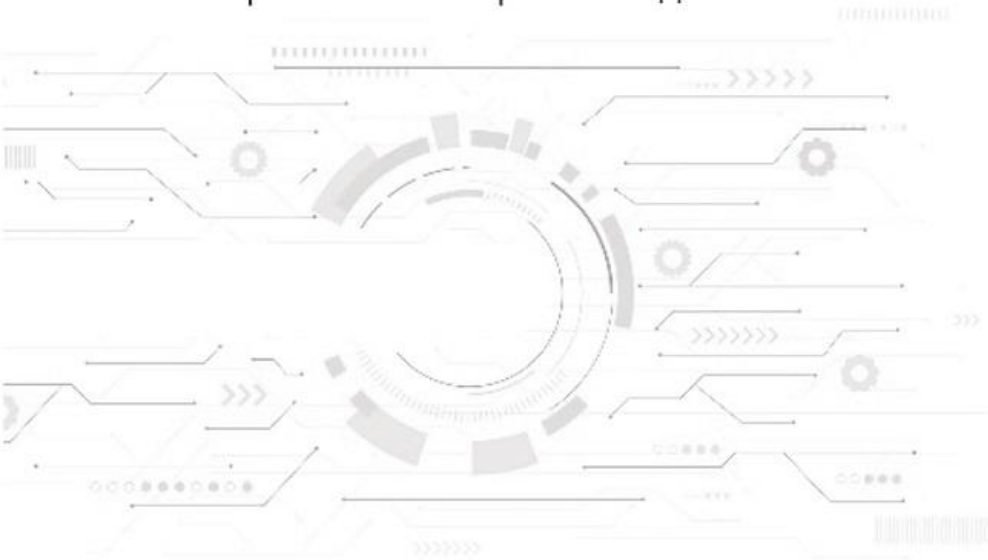
- Управление данными для принятия решений
- Управление данными как источником дохода
- Управление ассортиментом
- Управление качеством
- Управление данными как активом

Технологии больших данных как объект управления

ТЕХНОЛОГИИ

Возможность хранить и обрабатывать практически неограниченные объемы данных любой структуры.

Существенное снижение стоимости хранения и обработки данных



- Распределенное хранение и обработка
- Управление технологическим обеспечением
- Соответствие архитектуры задачам
- OpenSource
- Развитие инженерных компетенций

Аналитика и машинное обучение как объект управления

АНАЛИТИКА И МАШИННОЕ ОБУЧЕНИЕ

Выявление скрытых зависимостей на основе анализа всего объема данных

Новое качество результатов машинного обучения

- Монетизация направлений бизнеса
- Управление на основе процесса CRISP-DM*
- Развитие аналитических компетенций
- Внедрение аналитики в операционный процесс
- Замещение функций автоматическими системами

* Кросс-индустриальный стандарт исследования данных (Cross industry standard for data mining)

Технологический стек Сбера для работы с большими данными

ML Space

Машинное обучение и искусственный интеллект
Machine Learning (ML) and Artificial intelligence (AI)

SDP Analytics

Технологии и инструменты визуализации и анализа данных
Business Intelligence (BI)

Традиционные реляционные системы управления базами данных (RDBMS)	Массивно-параллельные системы управления базами данных (MPP DB)	Распределенные системы хранения и обработки данных любых форматов (HADOOP)	Системы распределенной обработки данных в оперативной памяти (InMemory)	Специализированные системы управления базами данных (GraphDB)
Pangolin	**SDP Greenplum**	**SDP Hadoop**	**Ignite SE**	**SDP Fast Graph**

SDP DataQuality

Технологии и инструменты интеграции и трансформации данных
Extract Transform Load (ETL)

SDP DataFlow

Люди как объект управления

D-PEOPLE

Создание и развитие эффективных команд

Формирование конкурентного преимущества на базе управления интеллектуальным капиталом

- Поиск и подбор
- Адаптация и обучение
- Мотивация и удержание
- Развитие и планирование карьеры
- Управление эффективностью работы

Эффективность работы специалистов по большим данным определяется квалификацией и инструментарием

Сервис работы с данными и моделями для Data Scientist с полным технологическим стеком

Заказ инфраструктуры	1 ЧАС ▾ 3 МЕСЯЦА
Предоставление доступа к данным	1 ЧАС ▾ 2 МЕСЯЦА
Предоставление данных по каталогу	1 ЧАС ▾ 2 МЕСЯЦА
Вывод моделей в промышленную эксплуатацию	4 ЧАСА ▾ 3 МЕСЯЦА
Установка ПО и библиотек	1 МИН ▾ 2 МЕСЯЦА

BIG DATA: данные, технологии, аналитика, люди. Резюме

- Пять ключевых элементов экосистемы больших данных:
  - Вертикальные решения и услуги
  - Технологические инструменты
  - Цифровая инфраструктура
  - Базовая инфраструктура
  - Генерация данных

- Большие данные – наборы данных, которые настолько объемны или сложны, что требуют специальных средств обработки. Это всегда оцифрованные данные, могут быть структурированными и не структурированными

- Данные как объект управления включают в себя следующие области: данные, технологии, аналитика и люди

- Технологии больших данных позволяют хранить и обрабатывать практически неограниченные объемы данных любой структуры

- Методы машинного обучения позволяют извлекать из массивов данных дополнительную ценность - выявлять скрытые зависимости, монетизировать данные, развивать аналитику на новом уровне

- Специалисты в области больших данных – важный элемент в организации работы с данными. Эффективность работы D-people зависит от уровня инфраструктуры работы с большими данными, квалификации сотрудников (постоянное обучение)

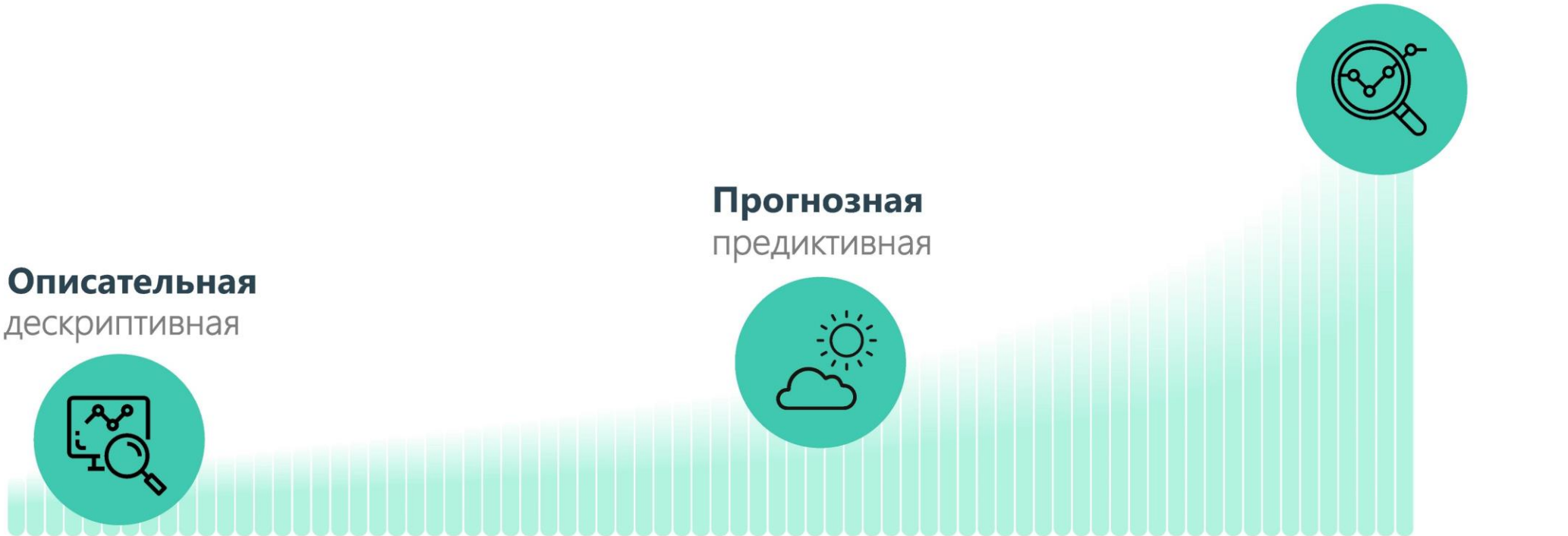
Аналитика больших данных и ML

Виды аналитики

Описательная
дескриптивная

Прогнозная
предиктивная

Предписывающая
прескриптивная



Путь развития аналитики в компании

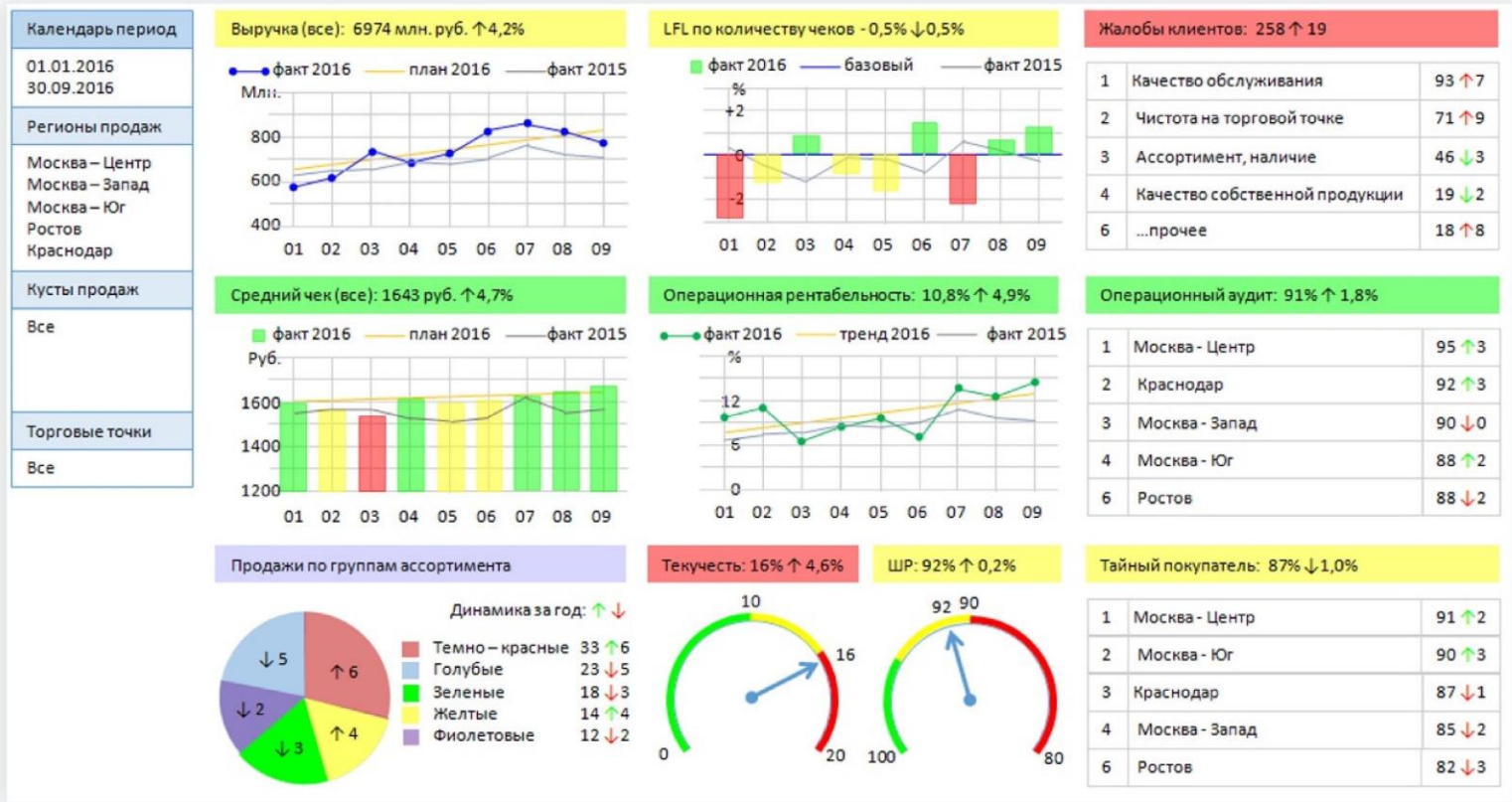
Описательная (дескриптивная) аналитика

Дескриптивная аналитика: **Что происходит?**

Результат: аналитический отчет, дашборд

Примеры:

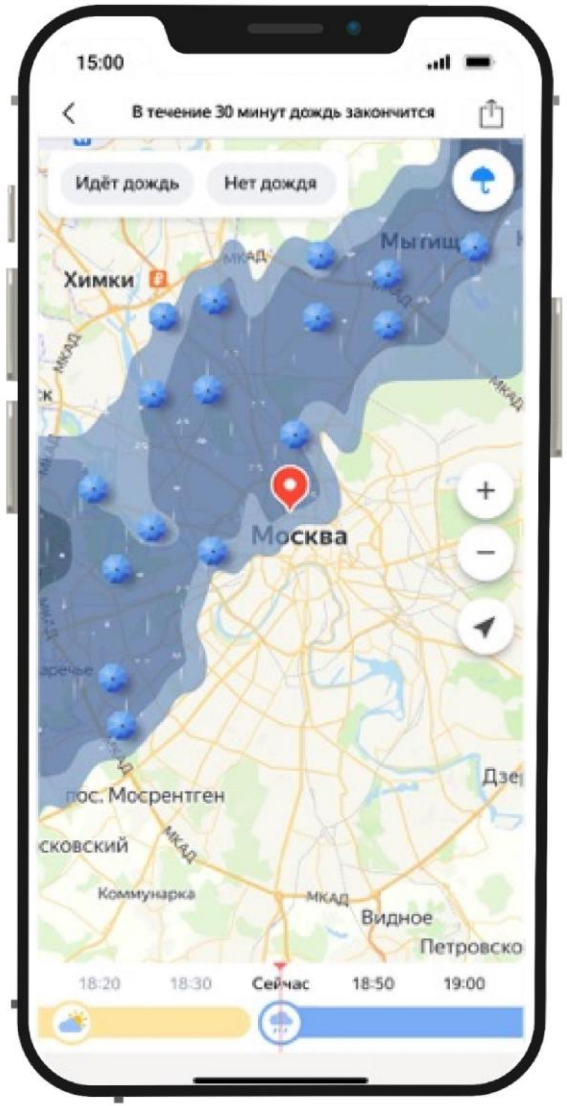
- Аналитический отчет
- Дашборд



Прогнозная (предиктивная) аналитика

Прогнозная аналитика: **Что может случиться?**

Результат: прогнозная модель



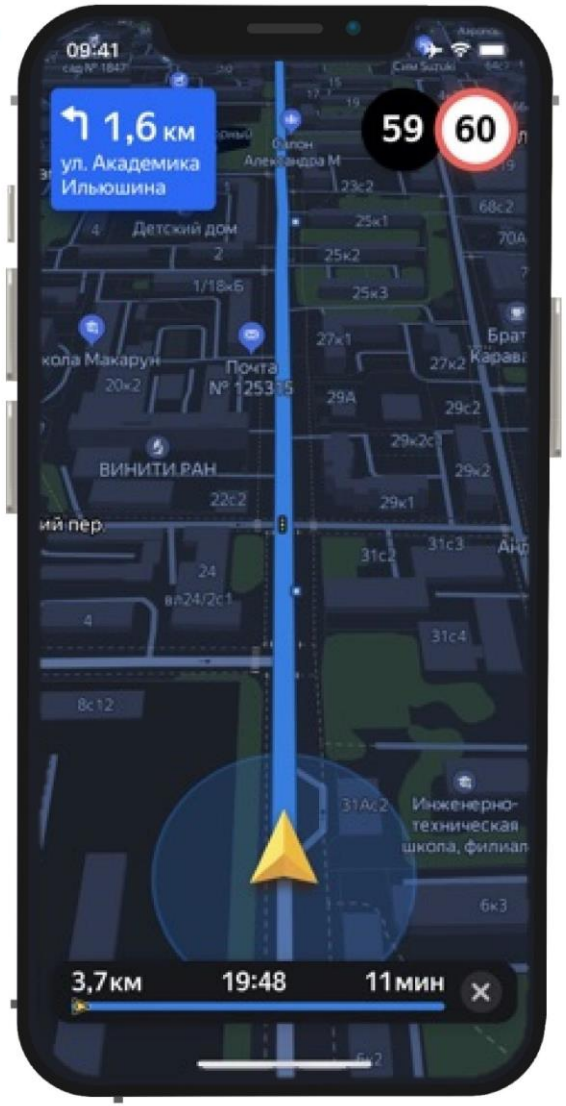
Примеры:

- Сколько устройств мы продадим в следующем месяце?
- Какими будут продажи перед Новым годом в этот раз в разрезе почтовых индексов?
- Сколько штук Продукта А вернут в следующем месяце?
- Какими будут доходы и прибыль в следующем квартале?
- Сколько сотрудников нам нужно будет нанять в следующем году?
- Прогноз погоды

Предписывающая (прескриптивная) аналитика

Предписывающая аналитика: **Что мы должны сделать?**

Результат: прогнозная модель с рекомендациями действий

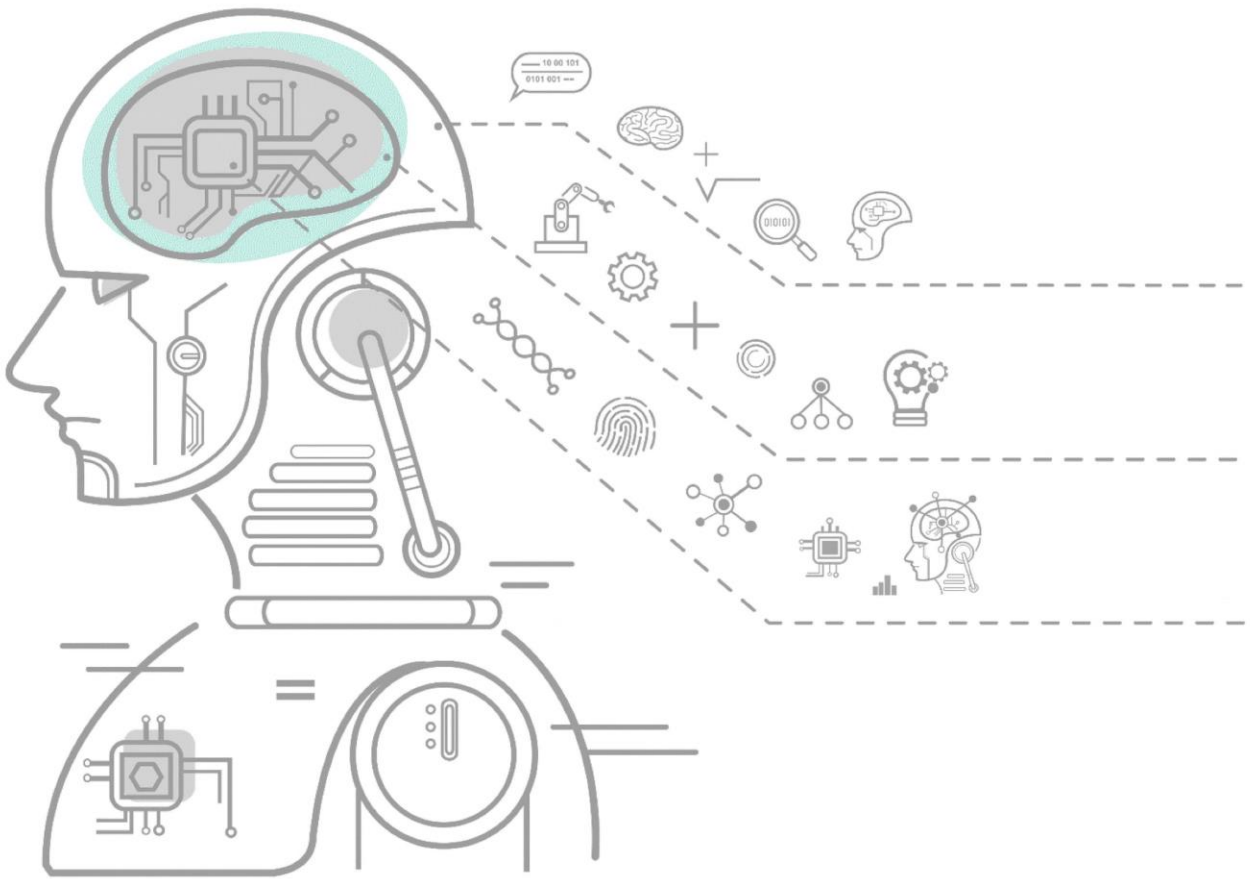


Примеры:

- Закажите 5 000 штук Элемента Б для поддержки продаж устройств в следующем месяце.
- Наймите X новых менеджеров по продажам в этих районах города, чтобы справиться с потоком заказов перед Новым годом.
- Увеличьте количество кандидатов на вакансию на 35%, чтобы достичь целей по найму персонала.
- Блокировка мошеннических транзакций
- Скоринговая модель, встроенная в процесс принятия решения

В основе предиктивной и прескриптивной аналитики лежит машинное обучение (ML)

Машинное обучение – изучает способы воспроизведения связей между событиями и результатом



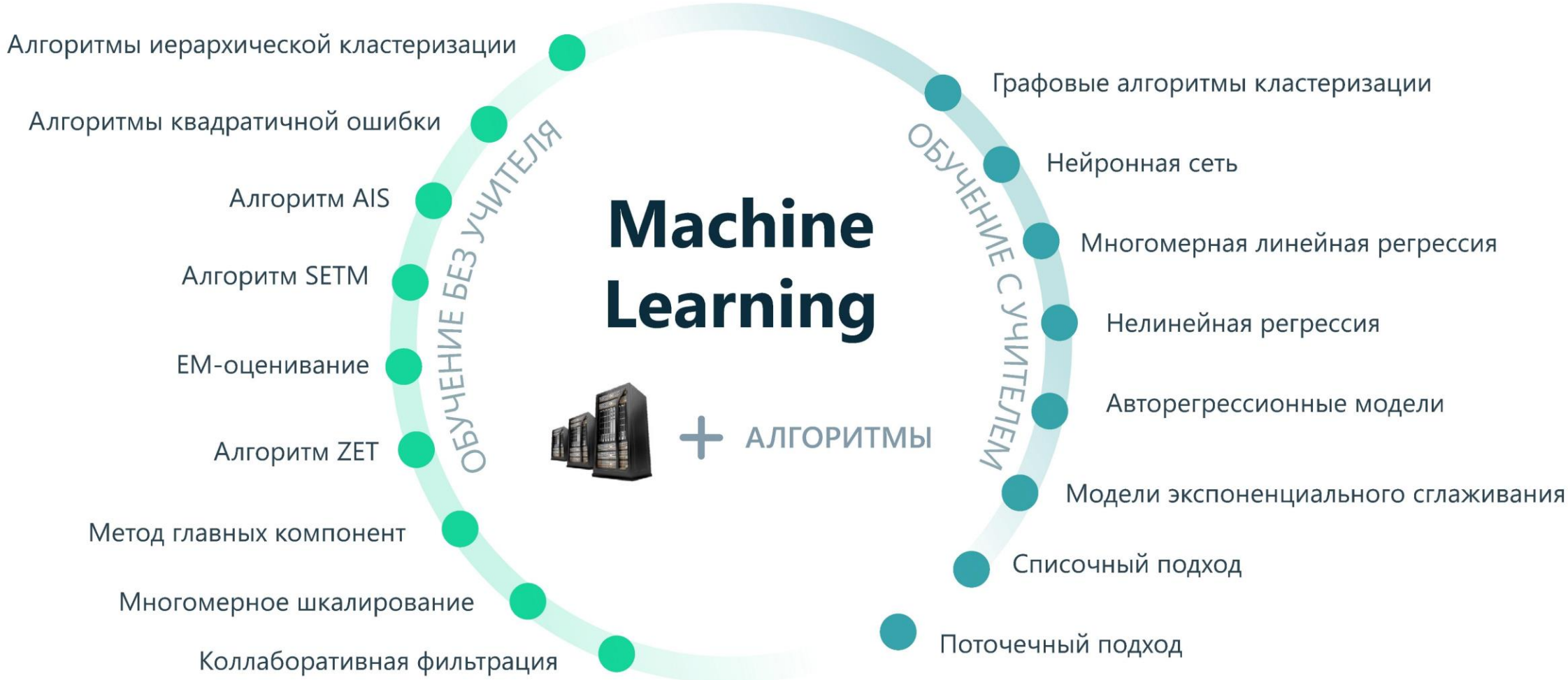
Машинное обучение – изучает способы воспроизведения связей между событиями и результатом

- Примеры спамовых и неспамовых сообщений
- Примеры хороших и плохих заемщиков
- Примеры мошеннических и нормальных транзакций
- Известные показатели прибыли по месяцам

В основе машинного обучения лежит процесс CRISP-DM

CRISP-DM (Cross Industry Standard for Data Mining) появился в 1999 году и представляет собой стандартную методологию и процесс исследования данных.

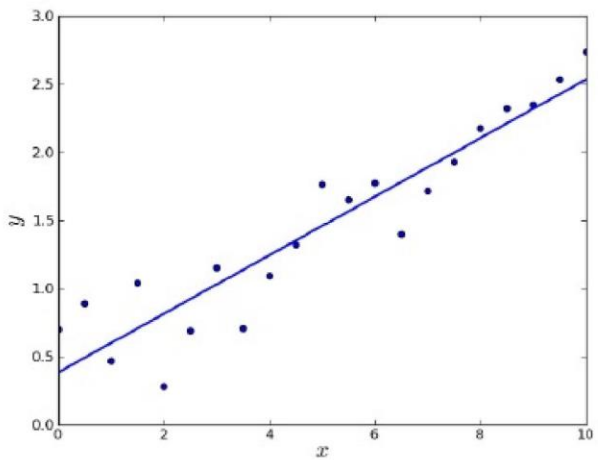
Машинное обучение – набор методов построения моделей, способных обучаться, и алгоритмов для их построения и обучения



Классы задач машинного обучения

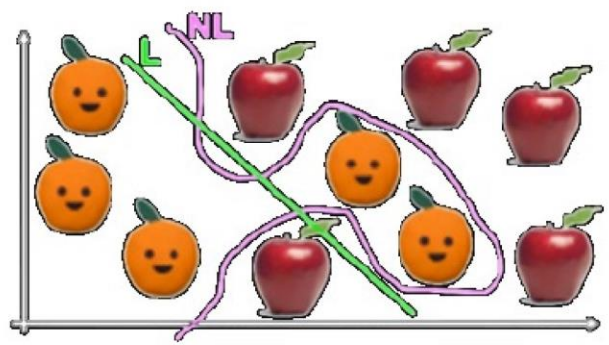
Регрессия

- Вещественные ответы
- Пример: предсказание роста по весу



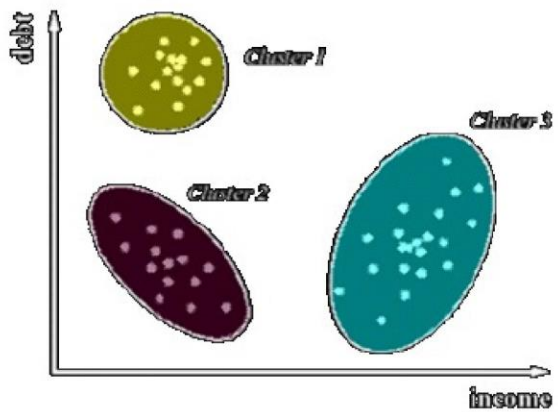
Классификация

- Конечное множество ответов
- Частный случай — бинарная классификация, два класса



Кластеризация

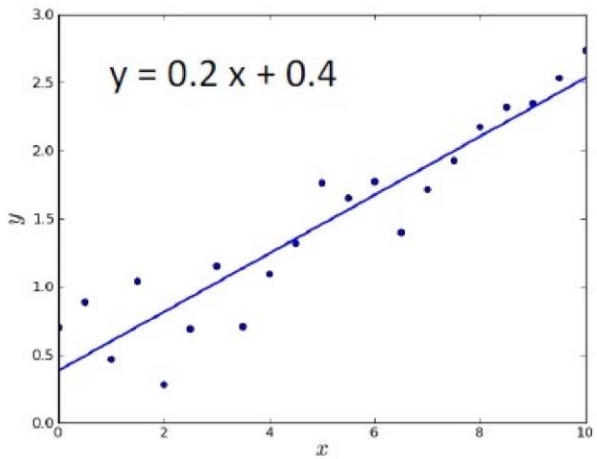
- Ответы отсутствуют
- Необходимо найти группы похожих объектов
- Пример: сегментация клиентов



Обучение алгоритма: классы моделей

Линейные модели

- Интерпретируемые
- Легко обучаются
- Невысокая сложность моделей

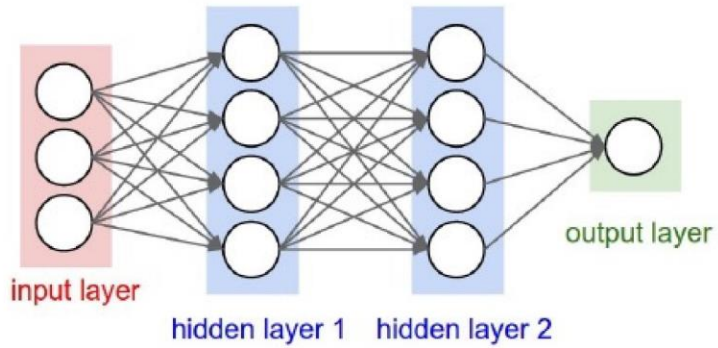


Решающие деревья

- Позволяют получать сложные модели
- Используются во многих задачах
- Технические сложности с обучением на больших данных

Нейронные сети

- Позволяют решать задачи для сложных данных с внутренней структурой, таких как изображения и тексты.
- Подходят для больших данных



Сравнение моделей машинного обучения

Класс моделей	Выразительная способность	Необходимый объем данных	Допустимое количество признаков	Типы данных
Линейные модели	Низкая	Низкий	Практически любое	Структурированные
Решающие деревья	Высокая	Высокий	Около 1000	Структурированные
Нейронные сети (deep learning)	Очень высокая	Очень высокий	Практически любое	Любые (включая изображения, звук, тексты)

Ключевые тренды в области больших данных и аналитики

Повышение доступности

- **Расширение пользовательского опыта**
  Преднастроенные панели данных (дэшборды) всё чаще заменяются на автоматизированные, динамические данные, адаптируемые к потребностям конечных пользователей
- Перевод больших данных и технологий анализа данных из традиционных центров обработки и облачных сред на физические активы

Повышение качества аналитики данных

- **Использования малых и широких данных для повышения точности аналитики**
  Анализ малых и широких данных позволяет организациям повысить точность аналитики, прогнозов и качество принятия решений
- Масштабирование использования технологий ИИ в корпоративных целях
- Построение и использование составной аналитики на базе множества различных компонентов данных, аналитики и решений в области ИИ
- Использование матрицы данных в качестве поддерживающей архитектуры составных данных и аналитики

Проникновение данных в операционные и стратегические процессы компаний

- **Данные и аналитика как одна из основных бизнес-функций**
  Значительно увеличивается влияние руководителей отделов данных и аналитики на цели и будущее компании
- Применение инженерной аналитики для формирования сетей принятия срочных решений в коммерческих организациях
- Масштабирование использования передовых практик DevOps[2] в системах принятия решений

Аналитика больших данных и ML. Резюме

- Аналитика может быть описательной, предсказательной и предписывающей. В основе предписывающей и предсказательной (прогнозной) аналитики лежит машинное обучение
- Машинное обучение – это набор методов построения моделей, способных обучаться, и алгоритмов для их построения и обучения
- В основе машинного обучения лежит процесс CRISP-DM (Cross-Industry Standard Process for Data Mining)
- Определены три класса задач машинного обучения (регрессия, классификация и кластеризация) и три класса моделей (линейные, решающие деревья и нейронные сети)
- Данные и аналитика становятся основной бизнес-функцией организации, проникая в операционные и стратегические процессы компаний

Благодарю за внимание!