

Основы искусственного интеллекта

Понимание процесса Data Science (сбор данных, анализ, моделирование).

Обзор популярных инструментов и фреймворков в Data Science.

Что такое Data Science?

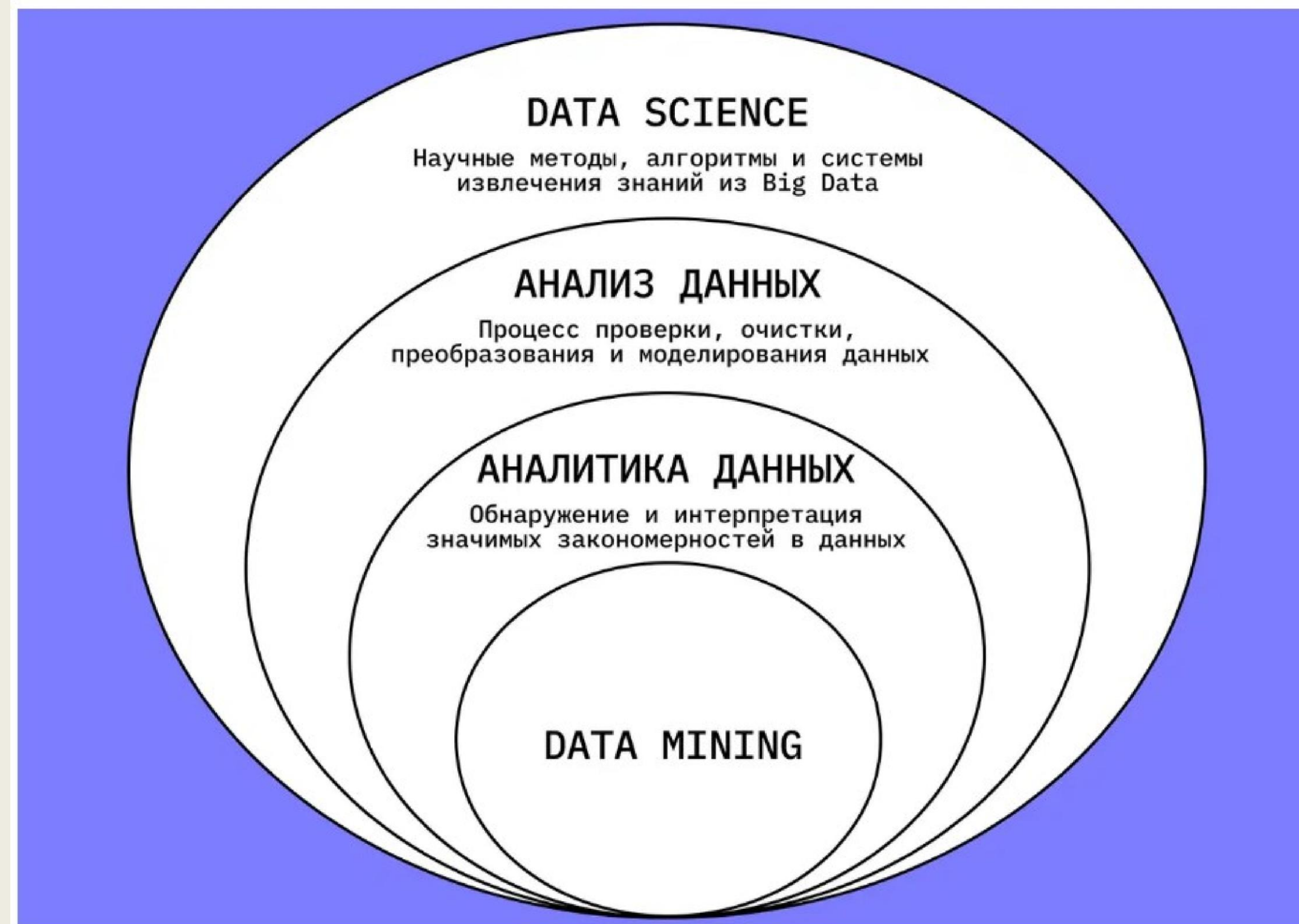
- **Data Science (наука о данных)** — это междисциплинарная область, которая использует методы статистики, машинного обучения, программирования и анализа данных для извлечения полезной информации и принятия решений.



Что такое Data Science?

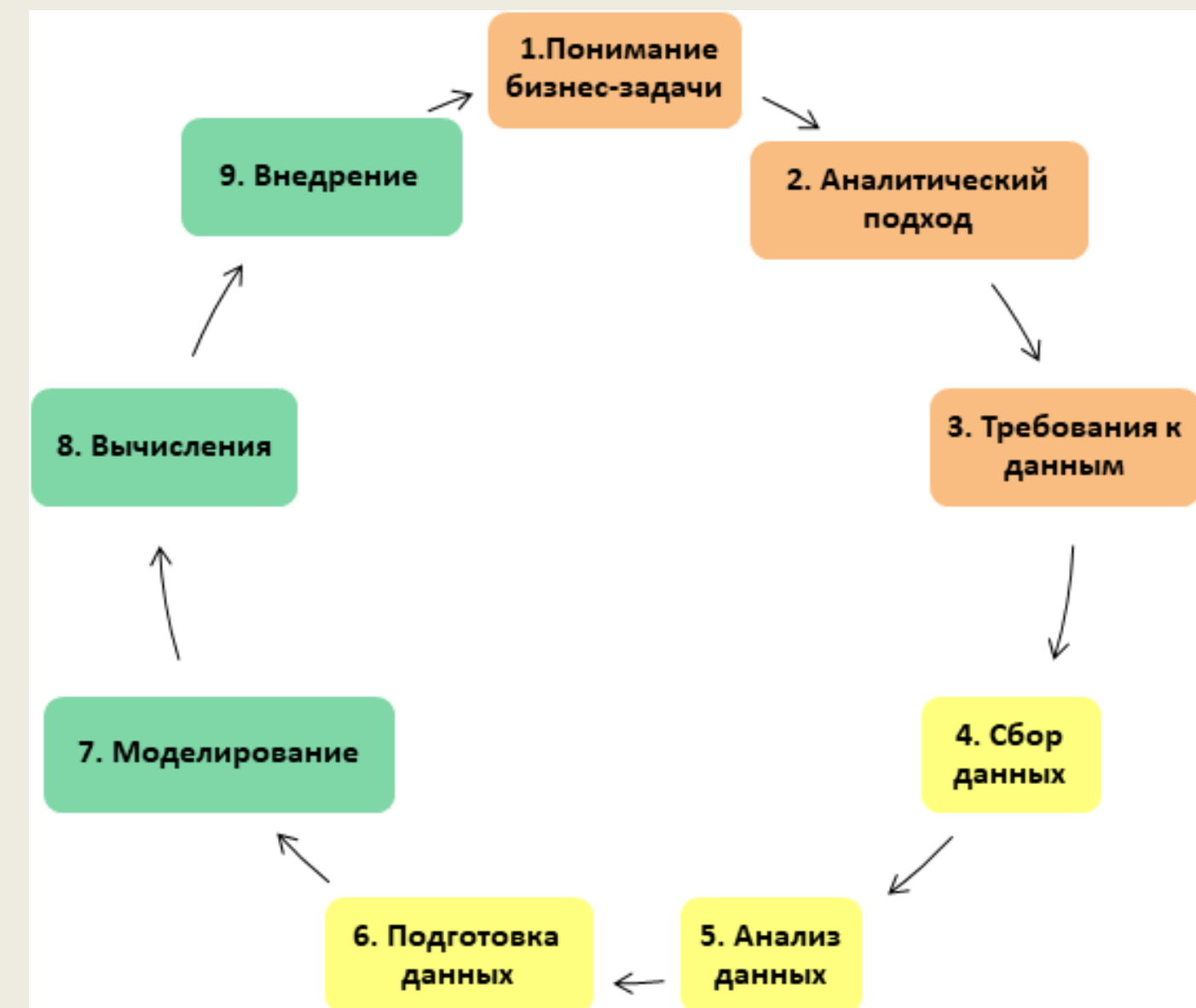


Что входит в Data Science?



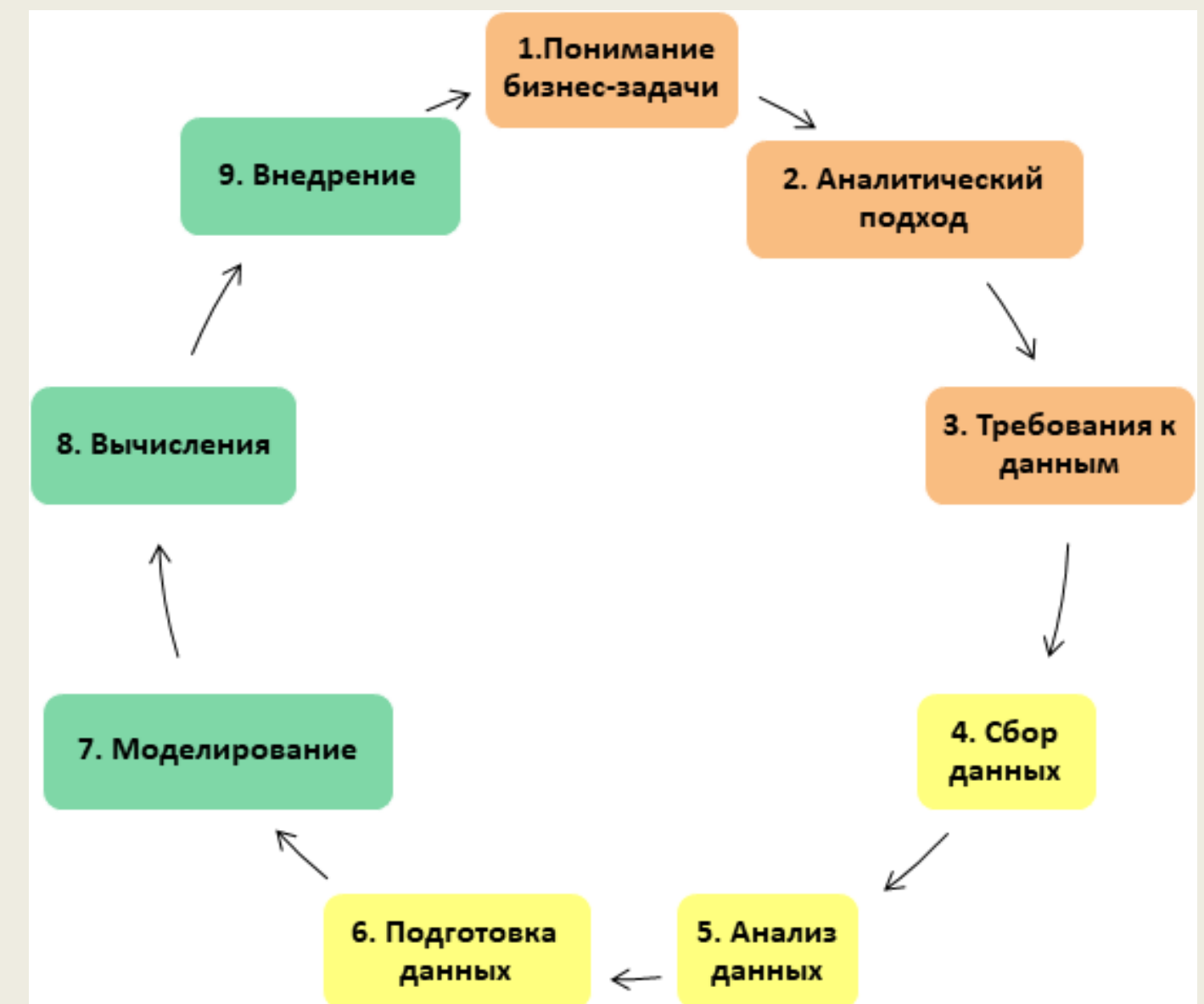
Основные этапы работы с данными

- 1. Сбор данных** - получение информации из различных источников (базы данных, веб-скрейпинг, API и т. д.).
- 2. Очистка данных** - обработка пропущенных значений, удаление дубликатов, исправление ошибок.
- 3. Анализ данных** - исследовательский анализ данных (EDA), визуализация и выявление закономерностей.



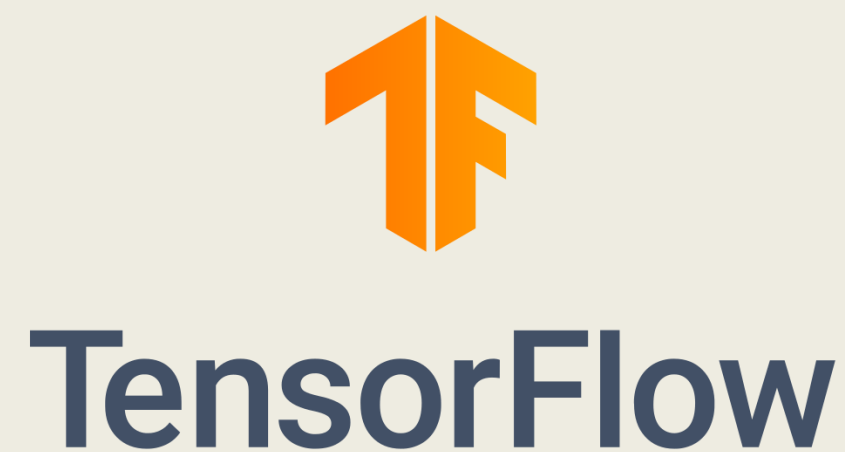
Основные этапы работы с данными

4. **Построение модели** - применение алгоритмов машинного обучения для создания предсказательной модели.
5. **Оценка и улучшение модели** - проверка точности, настройка гиперпараметров, кросс-валидация.
6. **Развертывание и интерпретация** - внедрение модели в бизнес-процессы, интерпретация результатов.



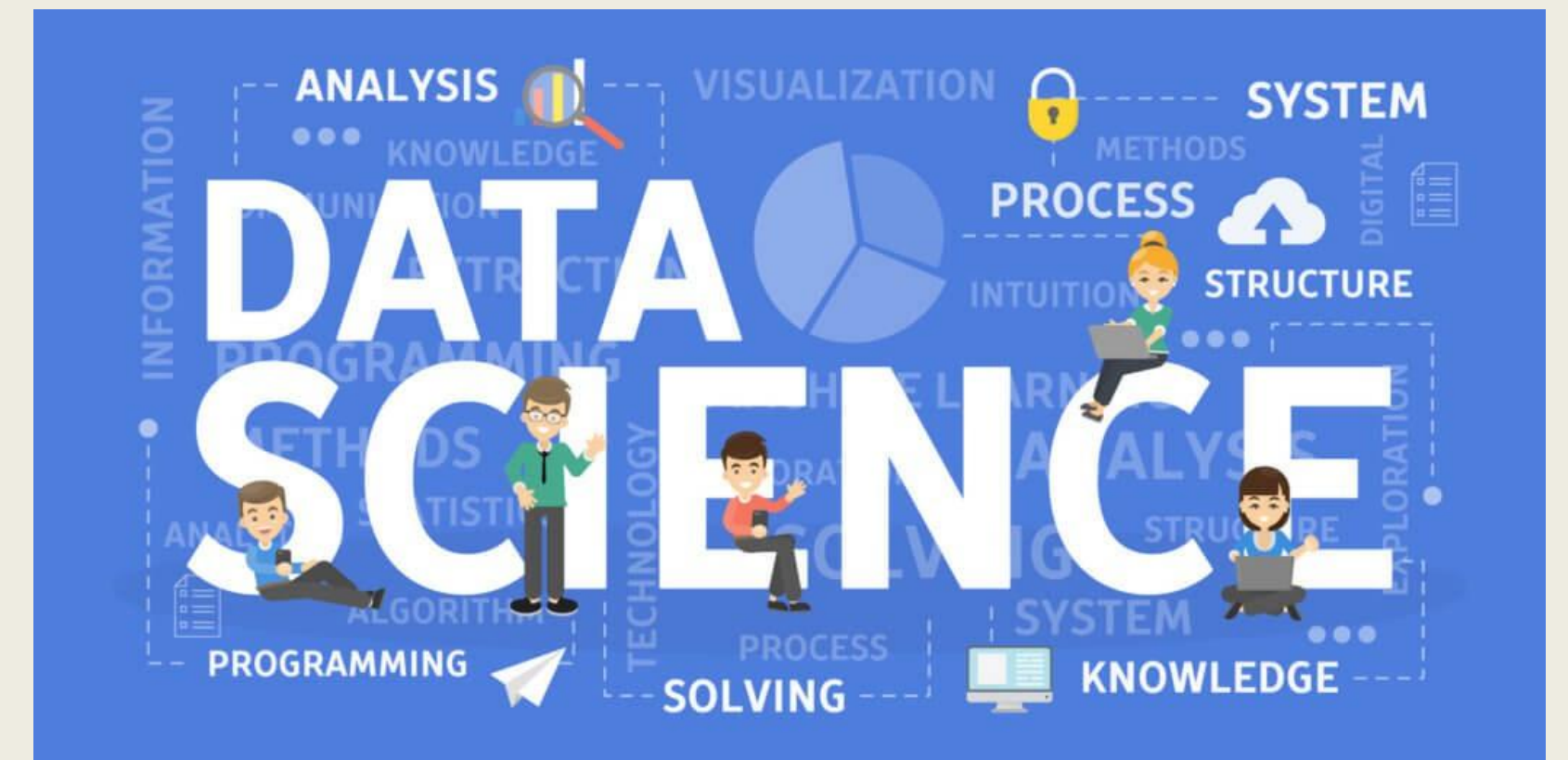
Основные инструменты и технологии

- Языки программирования: Python (NumPy, Pandas, Scikit-learn), R.
- Базы данных: SQL, MongoDB.
- Визуализация данных: Matplotlib, Seaborn, BI инструменты.
- Машинное обучение: PyTorch, TensorFlow, Scikit-learn.



Области применения Data Science

- Бизнес-аналитика и прогнозирование продаж.
- Финансовый анализ и алгоритмическая торговля.
- Обнаружение мошенничества и кибербезопасность.
- Обработка естественного языка (NLP) и чат-боты.
- Медицинская диагностика и биоинформатика.
- Рекомендательные системы (Netflix, Amazon, Spotify).



Карьерные перспективы в Data Science

- Data Scientist - специалист по анализу данных.
- Data Engineer - инженер по обработке данных.
- Machine Learning Engineer - разработчик моделей машинного обучения.
- Business Analyst - аналитик бизнес-процессов.

Карьерные перспективы в Data Science



ML - ИНЖЕНЕР



АНАЛИТИК ДАННЫХ



ИНЖЕНЕР ДАННЫХ



ML - ИССЛЕДОВАТЕЛЬ



АНАЛИТИК



РАЗРАБОТЧИК

Business Intelligence = BI = Бизнес – аналитика (rus) – это набор IT-технологий для сбора, хранения и анализа данных, позволяющих предоставлять пользователям достоверную аналитику в удобном формате, на основе которой можно принимать эффективные решения для управления бизнес- процессами компании.

Входные данные: разнородные отчеты

Отчет первого магазина

	A	B	C
1	Товар	Количество	Примечания
2	Бум Тул	100 шт.	Договор 100500
3	Кока-кола литровая	84 шт.	Закупить ещё
4	Кока-кола полу-литровая	12 шт.	-
5	Шетки зуб	145 шт.	-
6	Прочее	134 шт.	-

Отчет второго магазина

	Продукция	Кол-во	Комментарии
2	Бумага туалетная	105	Обсудить скидки
3	Соса-cola 1л	14	Нет
4	Соса-cola 0.5л	12	Нет
5	Зубные щётки	54	Нет
6	Другое	42	Разобрать

Отчет третьего магазина

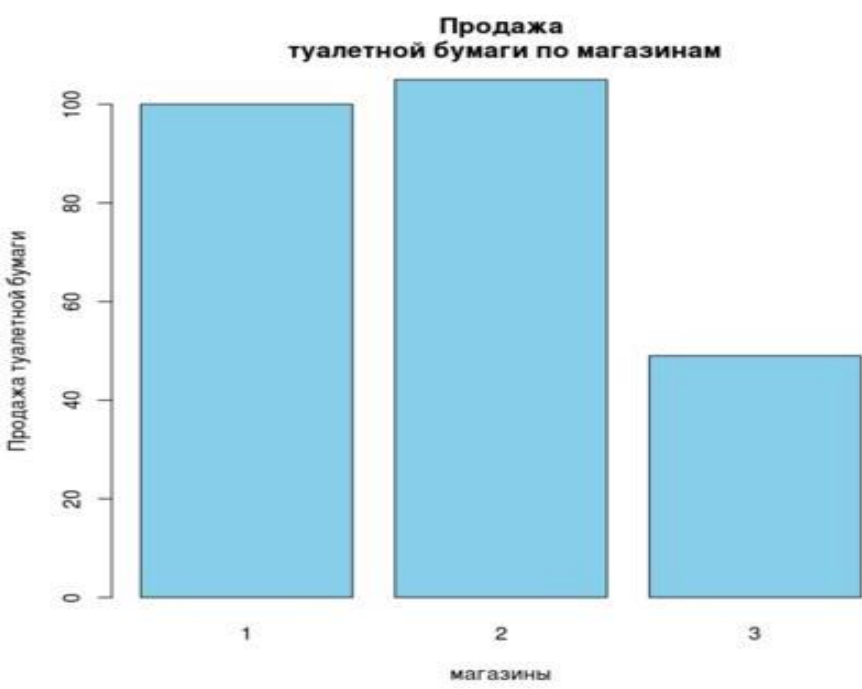
	Изделие	Продано	Примечание
2	Туалетная бумага	49 рулонов	Закупить двуслойные
3	Кола литровая	14 бутылок	Обсудить опт
4	Кола поллитра	12 бутылок	
5	Щётка зубная	54 упаковок	
6	Остальное	42 штук	

Выходные данные: общая таблица и визуализация

Сводная таблица

	A	B	C	D
1	Товар	Магазин 1	Магазин 2	Магазин 3
2	Туалетная бумага	100	105	49
3	Кола 1л	84	14	14
4	Кола 0.5 л	12	12	12
5	Зубные щётки	145	54	54
6	Прочее	134	42	42

Аналитика по сводным данным



Разница подходов

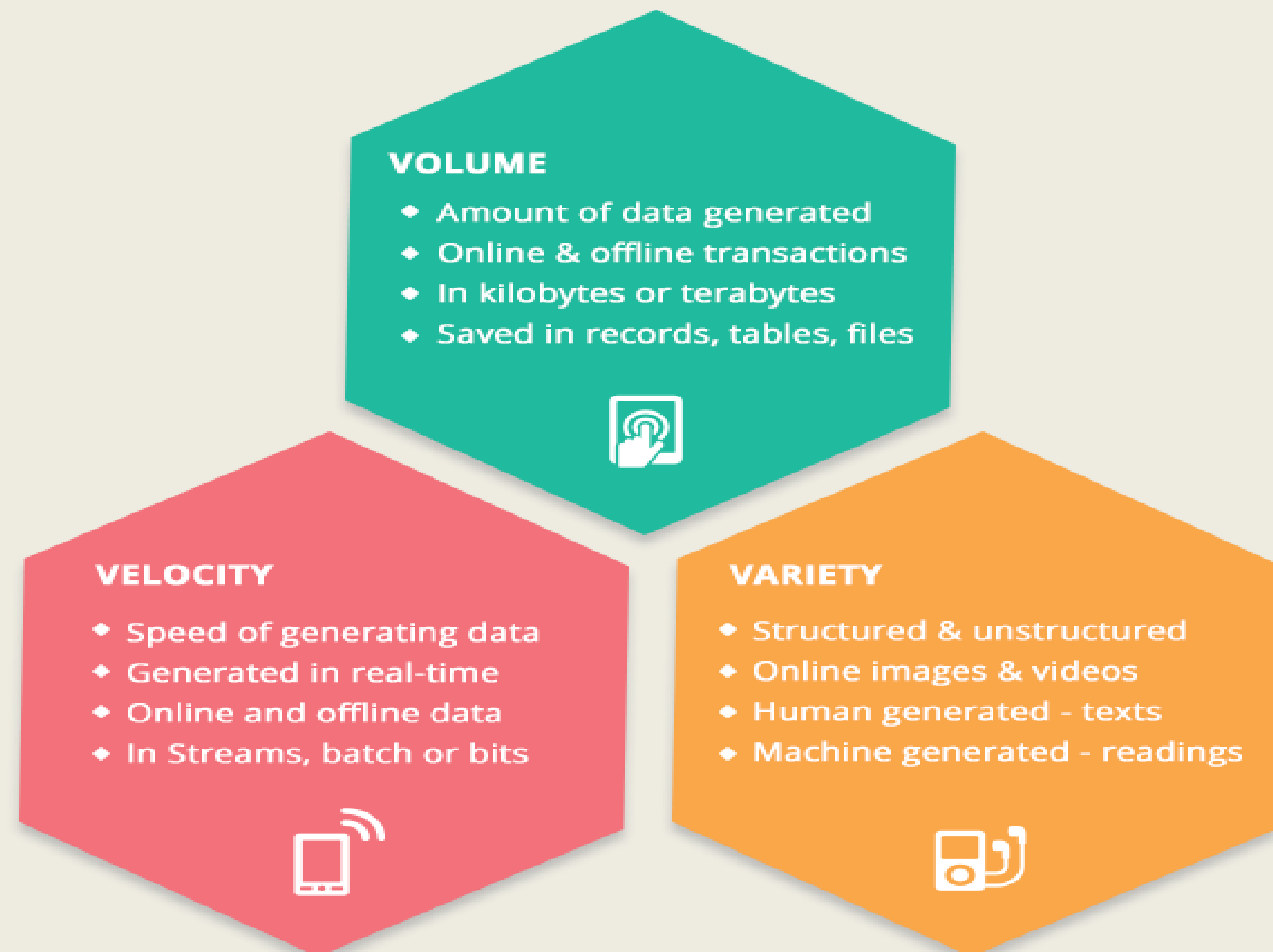
Традиционная аналитика	Big data аналитика
Постепенный анализ небольших пакетов данных	Обработка сразу всего массива доступных данных
Редакция и сортировка данных перед обработкой	Данные обрабатываются в их исходном виде
Старт с гипотезы и ее тестирования относительно данных	Поиск корреляций по всем данным до получения искомой информации
Данные собираются, обрабатываются, хранятся и лишь затем анализируются	Анализ и обработка больших данных в реальном времени, по мере поступления

Функции и задачи больших данных

Правило VVV — три признака или свойства, которыми большие данные должны обладать:

1. Volume – объем
2. Velocity – скорость
3. Variety - многообразие

THE 3Vs OF BIG DATA



Функции и задачи больших данных

Функция	Задача
Big Data — собственно массивы необработанных данных	Хранение и управление большими объемами постоянно обновляющейся информации
Data mining — процесс обработки и структуризации данных, этап аналитики для выявления закономерностей	Структурирование разнообразных сведений, поиск скрытых и неочевидных связей для приведения к единому знаменателю
Machine learning — процесс машинного обучения на основе обнаруженных связей в процессе анализа	Аналитика и прогнозирование на основе обработанной и структурированной информации

Источники больших данных

- Интернет — соцсети, блоги, СМИ, форумы, сайты, интернет вещей ([IoT](#)).
- Корпоративные данные — транзакционная деловая информация, архивы, базы данных.
- Показания устройств — датчиков, приборов, а также метеорологические данные, данные сотовой связи и т.д.



Принципы функционирования систем больших данных

- Горизонтальная масштабируемость
- Отказоустойчивость
- Локальность данных



Что можно сделать с помощью больших данных

Предсказать победителя Оскара!



Что можно сделать с помощью больших данных

Найти военную базу НАТО



Что можно сделать с помощью больших данных

Диагностировать беременность



Области применений технологий больших данных

Анализ данных опросов:

- Мониторинг общественного мнения и анализ социально-экономической ситуации
- Определение проблем, формирующих кризисную ситуацию
- Анализ реакции населения на внедрение различных региональных программ
- Анализ экономического положения и уровня жизни населения



Области применений технологий больших данных

Предвыборные исследования

- Анализ эффективности политической рекламы
- Анализ средств массовой информации
- Выявление наиболее эффективных средств влияния на мнения различных групп избирателей
- Диагностика предвыборной ситуации
- Анализ основных проблем избирателей



Области применений технологий больших данных

Общественная безопасность

- Анализ преступности
- Отслеживание уровня рецидивизма



Области применений технологий больших данных

Образование

- Планирование школьных округов
- Отслеживание успеваемости учащихся, выявление факторов способствующих повышению успеваемости
- Администрирование - контроль за уровнем выполнения обязательных программ и тестов.



Области применений технологий больших данных

Трудоустройство

- Анализ рынка труда - понимание состава и структуры рабочей силы
- Анализ заявлений о приеме на работу - разработка профилей претендентов.



Области применений технологий больших данных

Здравоохранение

- Отслеживание болезней и создание отчетов о случаях заболеваний
- Эпидемиология - выявление причин заболеваний и территории их распространения, а также контроль заболеваемости
- Медицинская помощь - определение профилей тех, кому часто требуется медицинская помощь
- Профилактика - выявление групп риска и необходимости медицинского вмешательства.



Области применений технологий больших данных

Окружающая среда

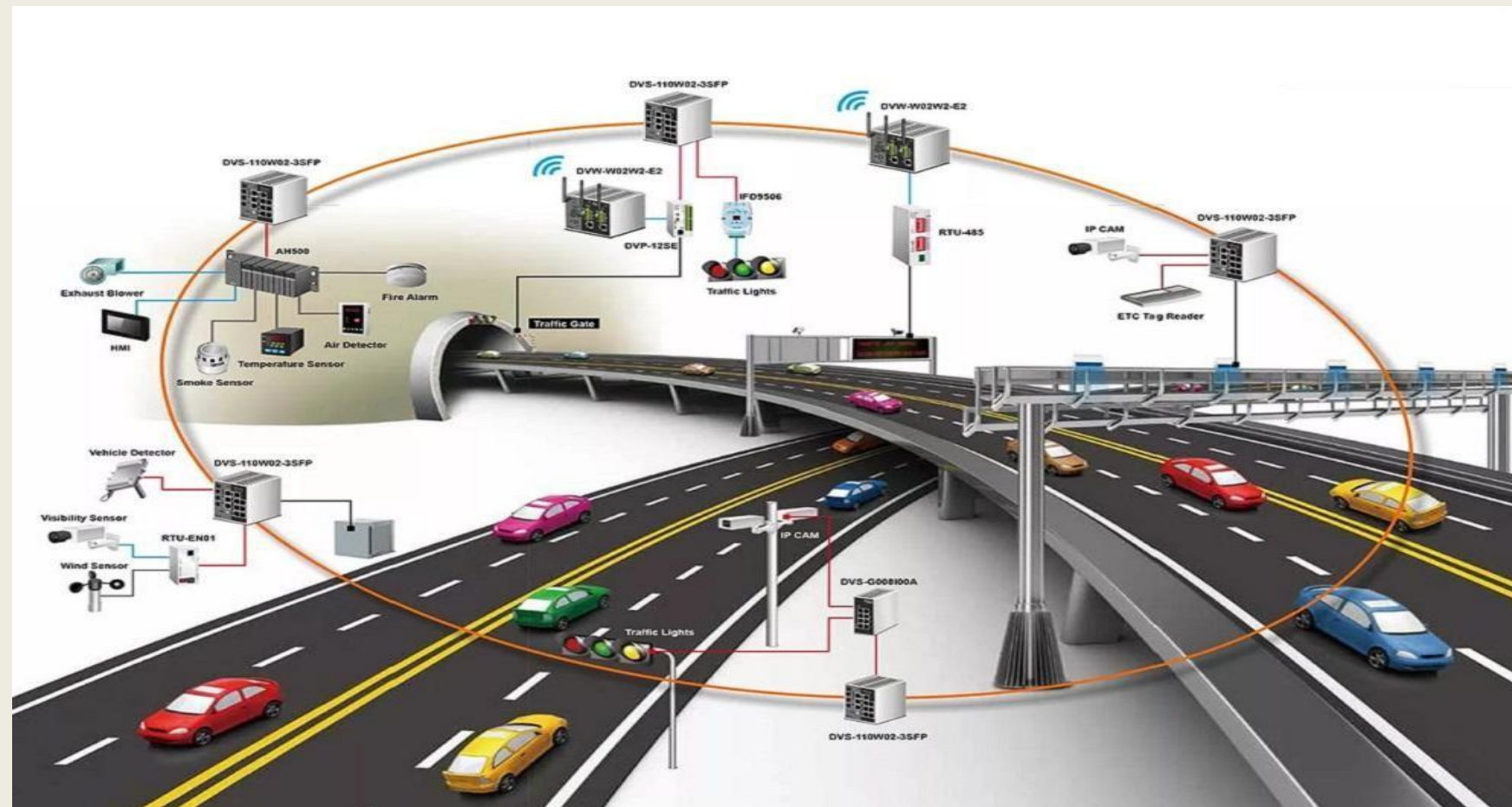
- Анализ экосистем - выяснение факторов, влияющих на здоровье экосистемы
- Оценка качества воды/воздуха - осуществление контроля за соответствием качества вод или воздуха имеющимся стандартам.



Области применений технологий больших данных

Транспорт

- Планирование наиболее эффективных маршрутов для лучшей организации транспортных и пассажирских потоков
- Создание отчетов о дорожных происшествиях для выяснения факторов, влияющих на происшествия
- Моделирование программ поддержания надлежащего состояния дорожного покрытия, прогнозирование возможного ремонта дорог.



Технологии больших данных в маркетинге

Стратегическое планирование

- Анализ удовлетворенности клиентов и изучения изменений потребностей общественности
- Оценка программ - понимание факторов успешной реализации программы
- Профилирование населения - более эффективное направление действия программы на определенные слои населения
- Анализ затрат - выявления наиболее эффективных программ
- Анализ результатов выполнения программ



Большие данные в бизнесе



- Создание точных портретов целевых потребителей.
- Предсказание реакции потребителей на маркетинговые сообщения.
- Максимальная персонализация рекламных сообщений.
- Увеличение кросс-продаж, повторных продаж, ремаркетинга.
- Поиск и определение причин популярности востребованных товаров и продуктов.
- Совершенствование продуктов и услуг, повышение лояльности клиентов.
- Повышение качества обслуживания.
- Предупреждение мошенничества.
- Снижение издержек в работе с поставщиками и клиентами.

Большие данные в бизнесе

Поставщики инфраструктуры — решают задачи хранения и предобработки данных.



Датамайнеры — разработчики алгоритмов, которые помогают заказчикам извлекать ценные сведения.

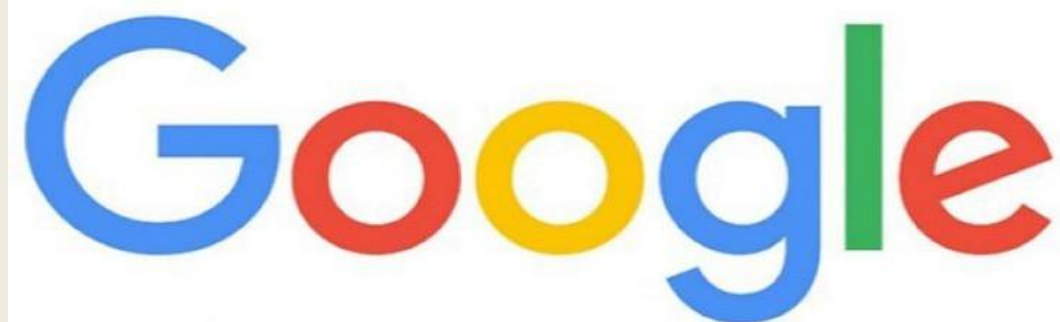


Системные интеграторы — компании, которые внедряют системы анализа больших данных на стороне клиента.



Потребители — компании, которые покупают программно-аппаратные комплексы и заказывают алгоритмы у консультантов.

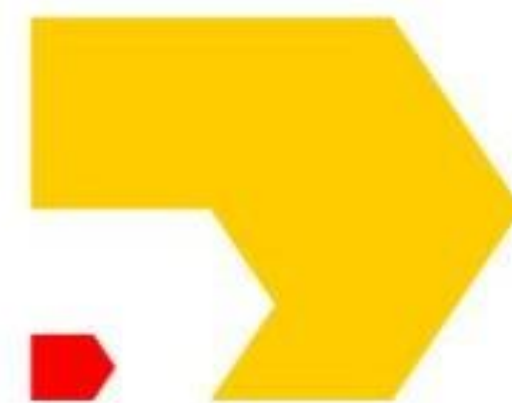
Поисковые системы

The Google logo, consisting of the word "Google" in its characteristic multi-colored font (blue, red, yellow, blue, green, red).

Google
BigQuery



Cloud
Bigtable



Yandex
Data Factory



mail.ru
group

Рамблер/

Выгоды использования Big Data в бизнесе

- Упрощается планирование.
- Увеличивается скорость запуска новых проектов.
- Повышаются шансы проекта на востребованность.
- Можно оценить степень удовлетворенности пользователей.
- Проще найти и привлечь целевую аудиторию.
- Ускоряется взаимодействие с клиентами и контрагентами.
- Оптимизируются интеграции в цепи поставок.
- Повышается качество клиентского сервиса, скорость взаимодействия.
- Повышается лояль



Технологии Big Data

- NoSQL;
- MapReduce;
- Hadoop;
- R;
- Python;
- Аппаратные решения.



Методы и техники анализа Big Data

Основными методами и техниками анализа, применимыми к Большим данным, являются следующие:

- Методы класса Data Mining
- Краудсорсинг
- Смешение и интеграция данных
- Машинное обучение
- Визуализация аналитических данных



Этапы работы с данными

Сбор данных

- **Сбор данных** - это первый и один из самых важных этапов, так как качество данных определяет точность и полезность анализа.
- **Источники данных:**
 - Структурированные данные (SQL-базы, CSV, Excel).
 - Неструктурированные данные (тексты, изображения, аудио, видео).
 - Внешние источники (API, веб-скрейпинг, IoT-устройства).



Этапы работы с данными

Сбор данных

- Проблемы сбора данных:
 - Пропущенные значения.
 - Дубликаты и ошибки.
 - Различные форматы и источники данных.



Этапы работы с данными

Анализ данных

- На этом этапе проводится обработка и исследовательский анализ данных (Exploratory Data Analysis, EDA), который включает:
 - Очистку данных (удаление дубликатов, заполнение пропусков).
 - Визуализацию (гистограммы, боксплоты, диаграммы рассеяния).
 - Определение закономерностей и корреляций.



Этапы работы с данными

Анализ данных

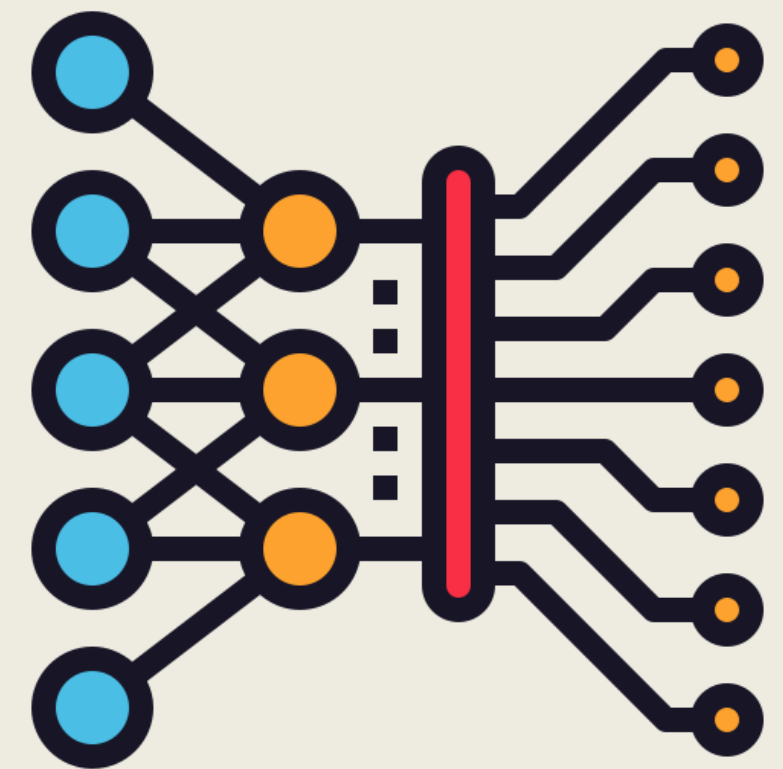
- Инструменты:
 - Python (Pandas, Matplotlib, Seaborn).
 - Google Colab Notebook.
 - SQL для работы с базами данных.



Этапы работы с данными

Моделирование

- После подготовки данных строятся предсказательные модели с использованием алгоритмов машинного обучения.
- **Основные методы:**
 - Линейная и логистическая регрессия.
 - Деревья решений и случайный лес.
 - Градиентный бустинг (XGBoost, LightGBM, CatBoost).
 - Нейронные сети и глубокое обучение.



Этапы работы с данными

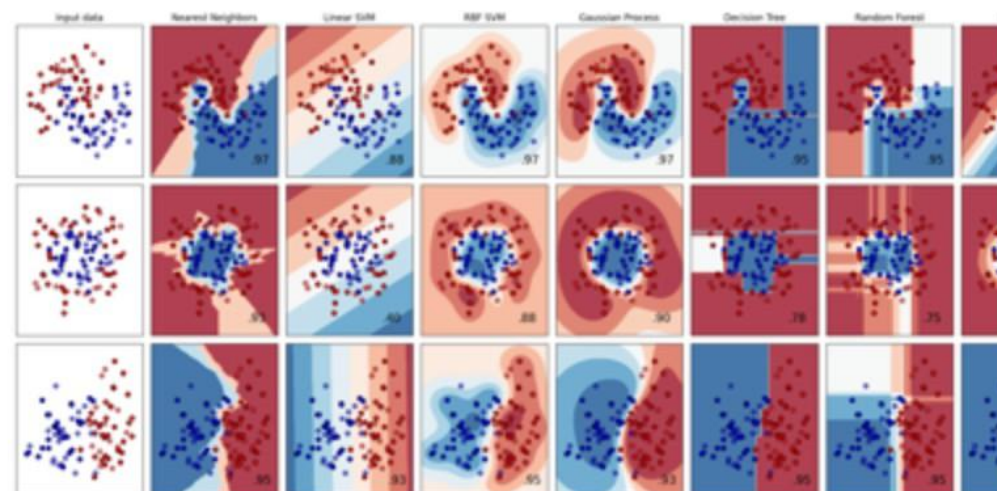
Моделирование

Классификация

Определение к какой категории принадлежит объект.

Применение: Выявление спама, распознавание изображений.

Алгоритмы: Градиентный бустинг, ближайшие соседи, случайный лес, логистическая регрессия, and больше...

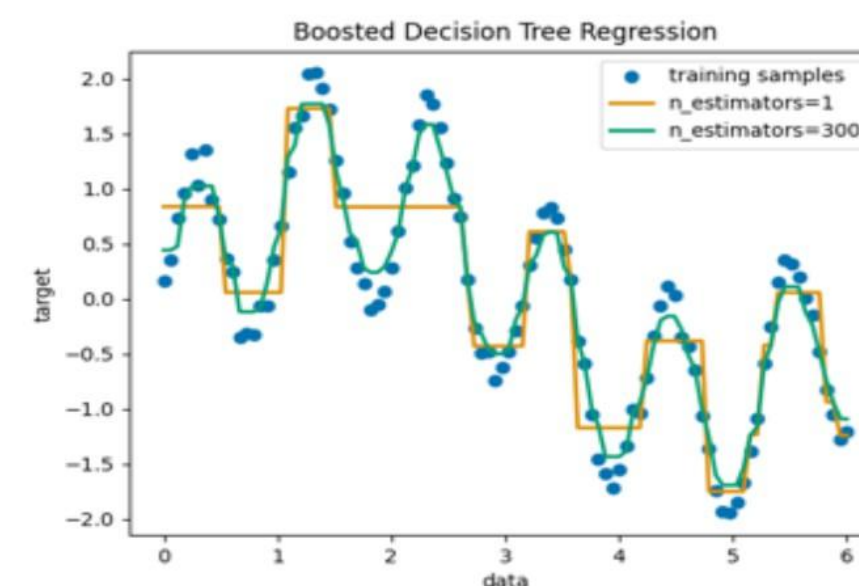


Регрессия

Прогнозирование атрибута с непрерывным значением связанного с объектом.

Применение: Реакция на лекарство, Отток клиентов.

Алгоритмы: Градиентный бустинг, ближайшие соседи, случайный лес, ридж-регрессия (ridge), and больше...



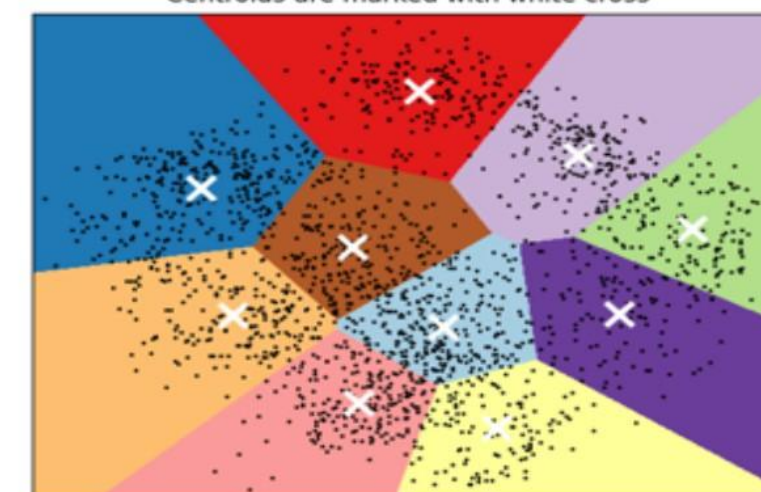
Кластеризация

Автоматическая группировка похожих объектов.

Применение: Сегментация клиентов, Группировка экспериментальных результатов

Алгоритмы: k-среднее (k-Means), HDBSCAN, иерархическая кластеризация, and больше...

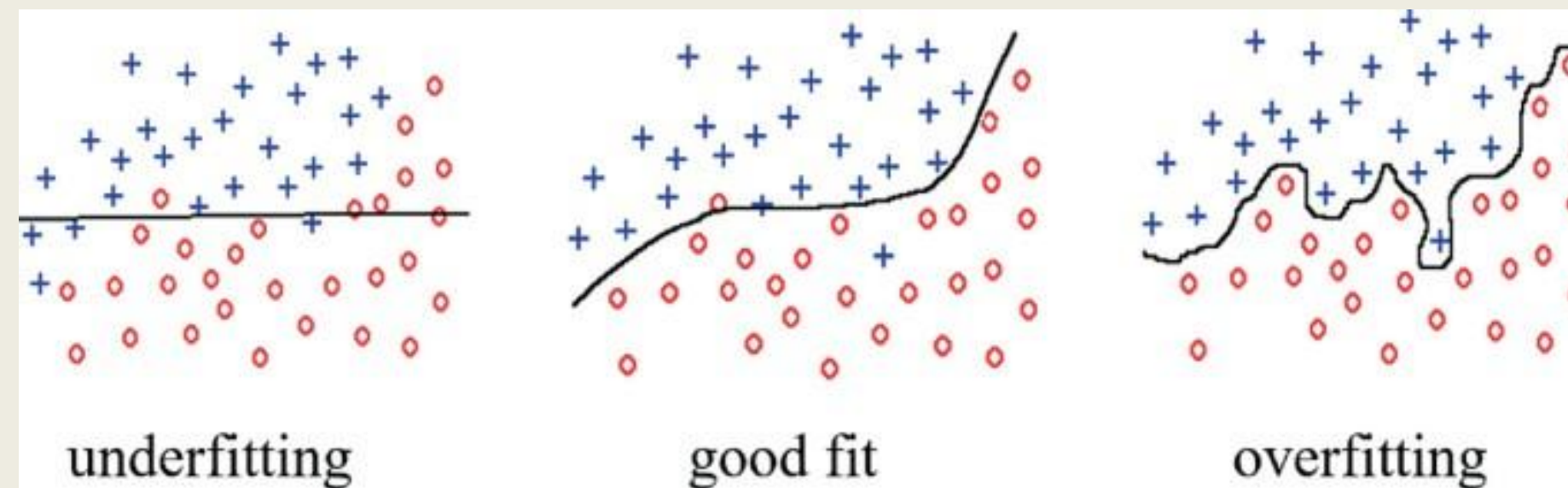
K-means clustering on the digits dataset (PCA-reduced data)
Centroids are marked with white cross



Этапы работы с данными

Моделирование

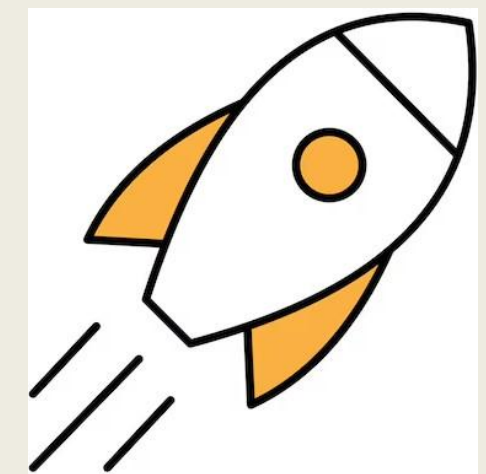
- Метрики оценки моделей:
 - MAE, RMSE (для регрессии).
 - Accuracy, Precision, Recall, F1-score (для классификации).



Этапы работы с данными

Выводы и развертывание

- После успешного моделирования важно правильно интерпретировать результаты и внедрить модель в рабочий процесс. Развертывание может происходить с помощью:
 - API (FastAPI, Flask, Django).
 - Облачных сервисов (AWS, GCP, Azure).
 - Интеграции с бизнес-системами.



Что такое Data Science?

- **Основные задачи:** сбор, обработка, анализ, визуализация данных
- **Применение** в бизнесе, медицине, финансах, маркетинге



Профессионализм



Ключевые этапы работы

1. Сбор и очистка данных
2. Исследовательский анализ данных (EDA)
3. Моделирование и машинное обучение
4. Деплой моделей в продакшн
5. Мониторинг и улучшение моделей

Ключевые этапы работы



Языки программирования для Data Science

📌 Python ✓

- Основной язык для анализа данных и ML
- Богатая экосистема библиотек



📌 R

- Популярен в статистике и академической среде
- Мощные инструменты для визуализации данных



📌 SQL

- Основной язык работы с базами данных
- Используется для извлечения данных



📌 *Другие языки:* Julia, Scala, Java (используются реже)

Библиотеки и инструменты для работы с данными

📌 Работа с данными в Python:

- **Pandas** - манипуляции с табличными данными
- **NumPy** - математические операции и работа с массивами
- **SciPy** - численные методы и статистика



📌 Визуализация данных:

- **Matplotlib** - базовые графики
- **Seaborn** - расширенные статистические визуализации
- **Plotly** - интерактивные дашборды



Инструменты машинного обучения

ML - Machine Learning

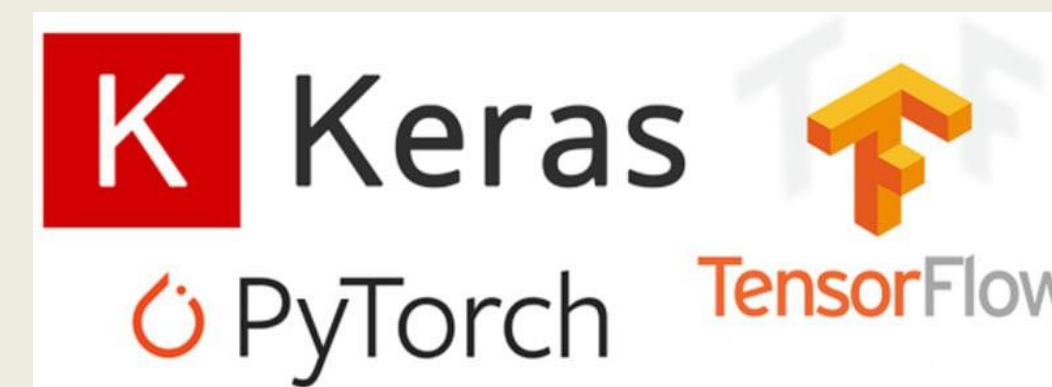
Классическое ML:

1. **Scikit-learn** - основной инструмент для машинного обучения (классификация, регрессия, кластеризация)
2. **XGBoost, LightGBM, CatBoost** - градиентный бустинг



Глубокое обучение:

1. **PyTorch** - гибкая библиотека для глубокого обучения
2. **TensorFlow & Keras** - построение и обучение нейросетей



Машиинное Обучение

- **Машинное обучение** — это подмножество искусственного интеллекта (ИИ), в котором компьютеры учатся на данных и совершенствуются, не программируя напрямую.



Машинное обучение против Традиционного программирования

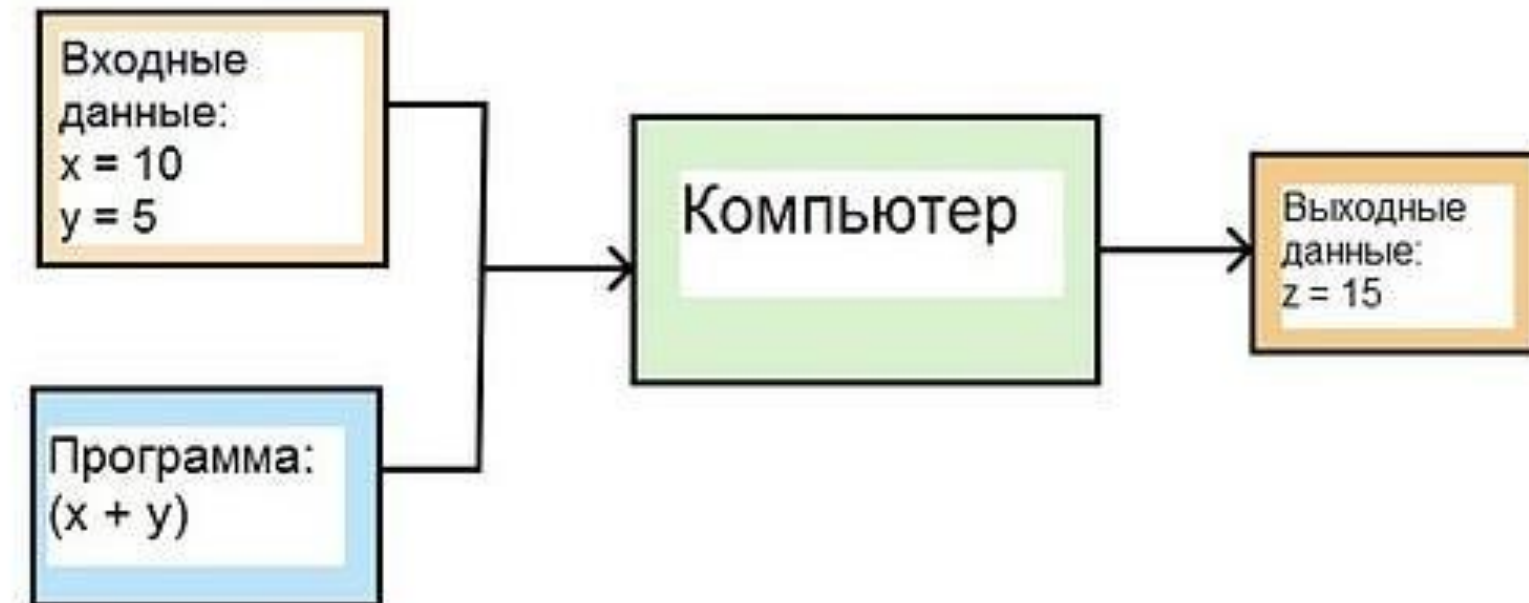
- **Машинное обучение** дает компьютерам возможность учиться без явного программирования. При машинном обучении с помощью алгоритмов выявляются закономерности в данных. На основе этих закономерностей создается модель данных для прогнозирования. Чем больше данных обрабатывает такая модель и чем дольше она используется, тем точнее становятся результаты. Это очень похоже на то, как человек оттачивает навыки на практике.
- **Традиционный способ программирования** компьютеров, можно сравнить с выпечкой, где рецепт требует точного количества ингредиентов и выполнения определённых условий приготовления. Традиционное программирование также требует создания подробных инструкций для компьютера.

Традиционное программирование

Традиционное программирование



Пример традиционного программирования:



Машинное обучение

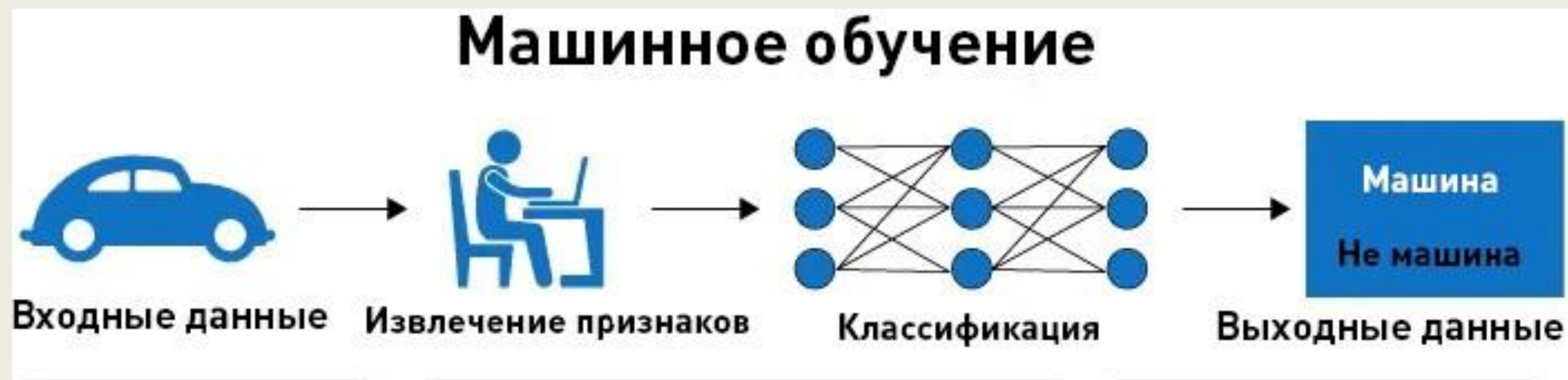
Машинное обучение



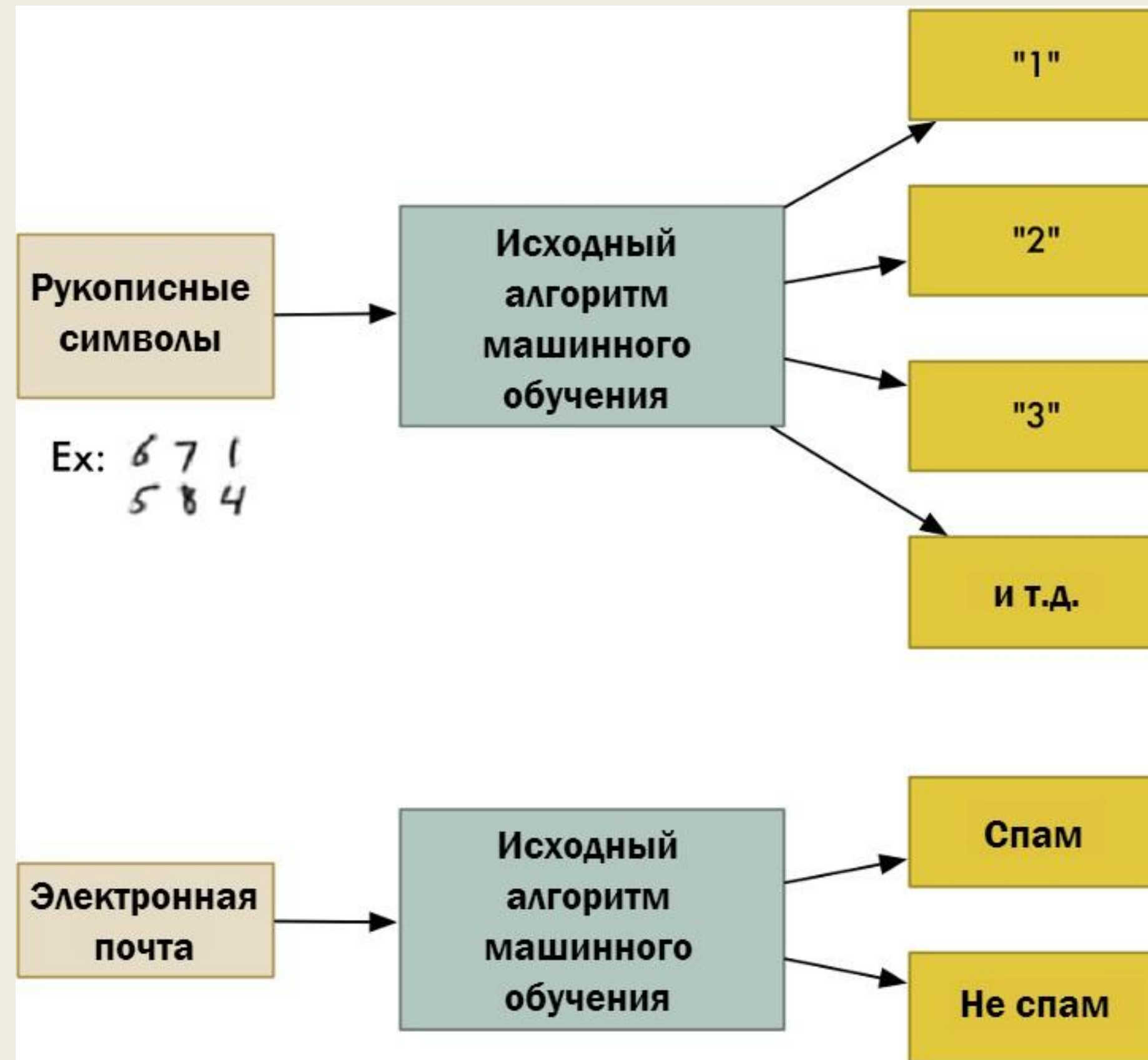
Пример машинного обучения:



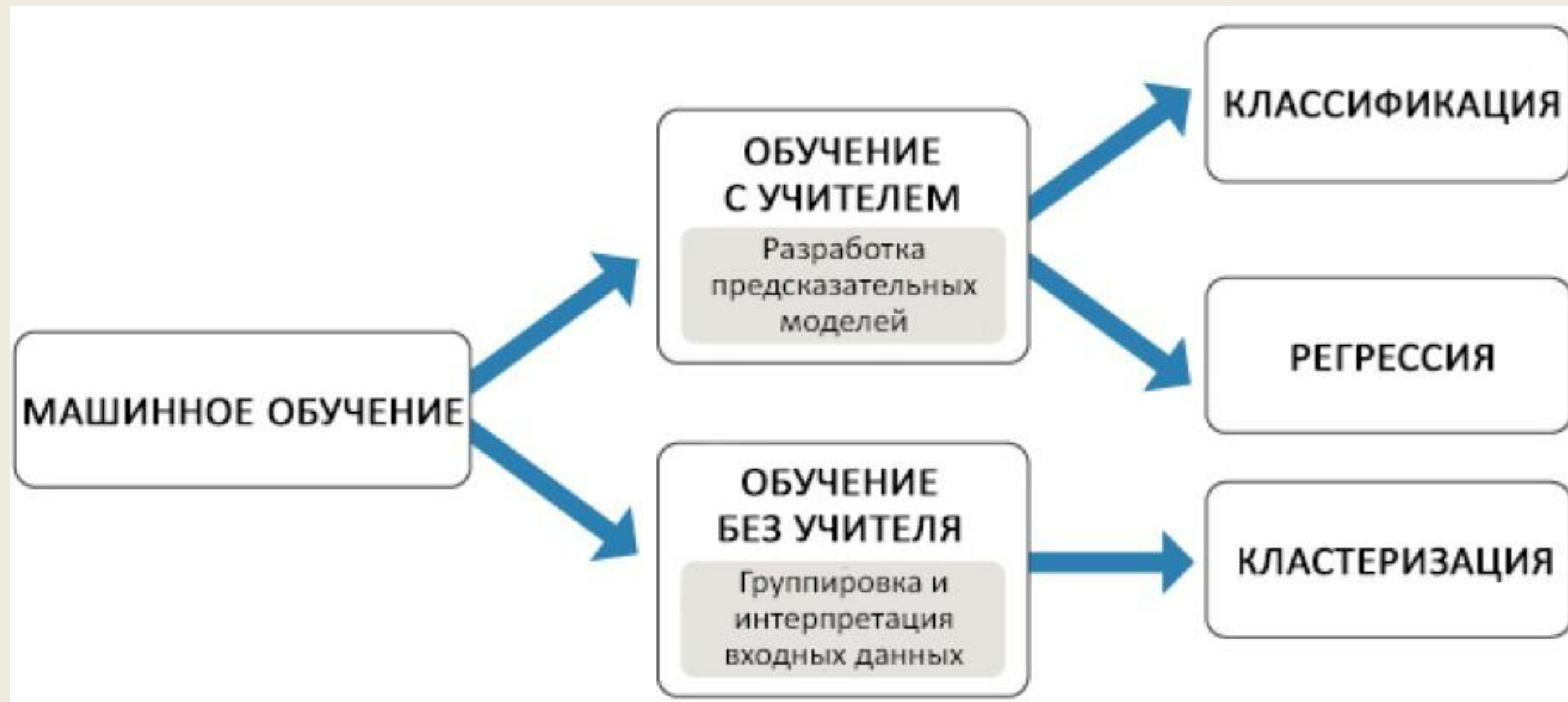
Примеры машинного обучения



Примеры машинного обучения



Виды машинного обучения



Виды машинного обучения

- **Обучение с учителем**, когда на имеющихся входных и соответствующих им выходных данных тренируют модель машинного обучения. Она в свою очередь учится эти данные сопоставлять и начинает предсказывать выход для новых входных данных.
- **Обучение без учителя**, когда модель тренируют только на входных данных, а она учится находить в них скрытые взаимосвязи или группировать.

Спасибо за внимание!