# Применение ANN
# Кластеризация

# Применение ANN для задач классификации

**Классификация** – это одна из основных задач машинного обучения.

**Искусственные нейронные сети (ANN)** позволяют решать задачи классификации **более эффективно**, чем традиционные алгоритмы.

# Что такое классификация?

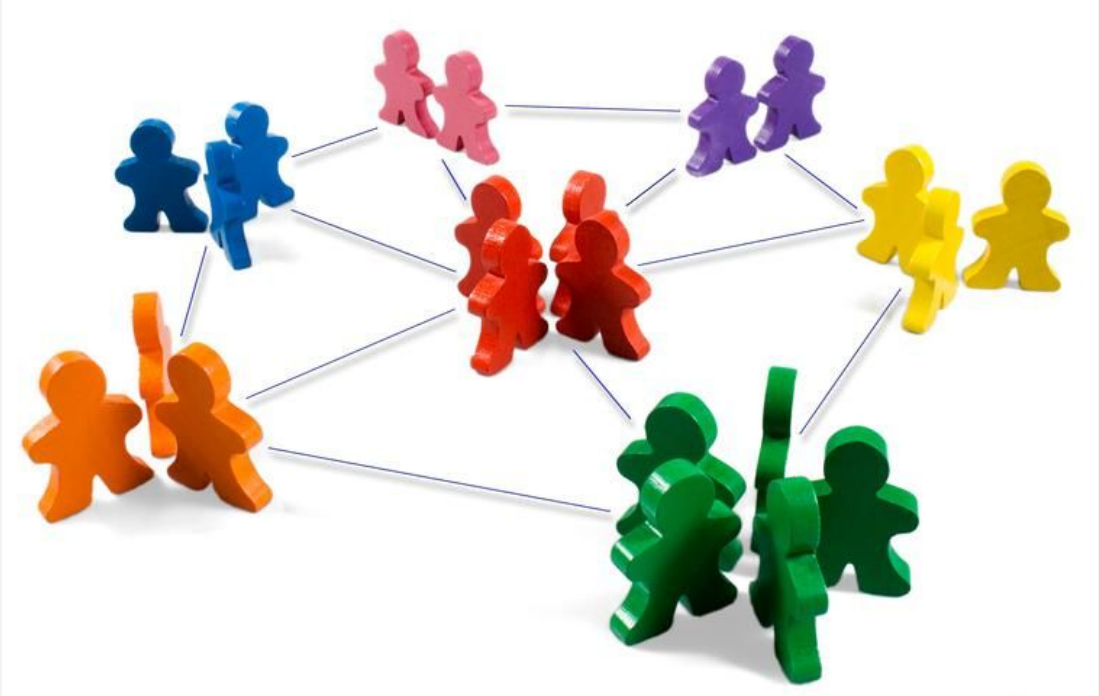
**Классификация** – это процесс присвоения входным данным **определенной категории**.

Обученный алгоритм способен **предсказывать класс** новых данных.

**Примеры задач классификации:**

- 📷 **Распознавание изображений** (кошка или собака?).
- 📩 **Обнаружение спама** (спам или нет?).
- 🏥 **Диагностика заболеваний** (здоров или болен?).

# Схему классификации



# Роль ANN в классификации

**Нейросети** способны находить **сложные закономерности** в данных.

Используются **разные архитектуры** ANN в зависимости от типа данных:

- **FCNN** – для универсальных задач.
- **CNN** – для изображений.
- **RNN** – для текстов и временных рядов.

**Глубокие нейросети** (Deep Learning) позволяют улучшить точность.

# Кластерный анализ (Cluster Analysis)

**Кластерный анализ (Cluster Analysis)** разделяет данные на группы на основе **сходства** или **отношений** между ними.

- Данные в **одном кластере** имеют **высокое сходство** между собой.
- Данные из **разных кластеров** должны иметь **минимальное сходство** друг с другом.

# Основные принципы кластеризации

- **Внутрикластерные расстояния минимизируются** (данные внутри кластера расположены компактно).
- **Межкластерные расстояния максимизируются** (разные кластеры должны быть отчетливо разделены).

# Цель кластерного анализа

**Кластеризация** направлена на выявление групп в данных, которые должны быть:

✔ **Значимыми** – кластеры должны отражать структурные закономерности в данных.

✔ **Полезными** – помогают **обобщать большие объемы данных**, упрощать их интерпретацию и находить скрытые закономерности.

# Кластеризация для облегчения понимания

**Цель** – выделение классов элементов, полезных для анализа и понимания окружающего мира.

**Примеры:**

✔ **Биология:** создание таксономий (царство, тип, класс, отряд, семейство, род, вид).

✔ **Биоинформатика:** группировка генов и белков с предполагаемой схожей функциональностью.

✔ **Психология и медицина:** обнаружение подкатегорий заболеваний (например, разные типы депрессии).

✔ **Бизнес:** сегментация клиентов по схожести покупок, реклама.

# Кластеризация для утилитарных целей

Позволяет **абстрагироваться от исходных данных** и выделять **ключевые центры кластеров.**

**Применение:**

✔ **Суммаризация:** представление данных через центроиды кластеров.

✔ **Сжатие:** представление каждого кластера его центроидом (векторное квантование).

✔ **Поиск ближайших соседей:** вычисление расстояний только внутри кластеров, а не между всеми объектами.

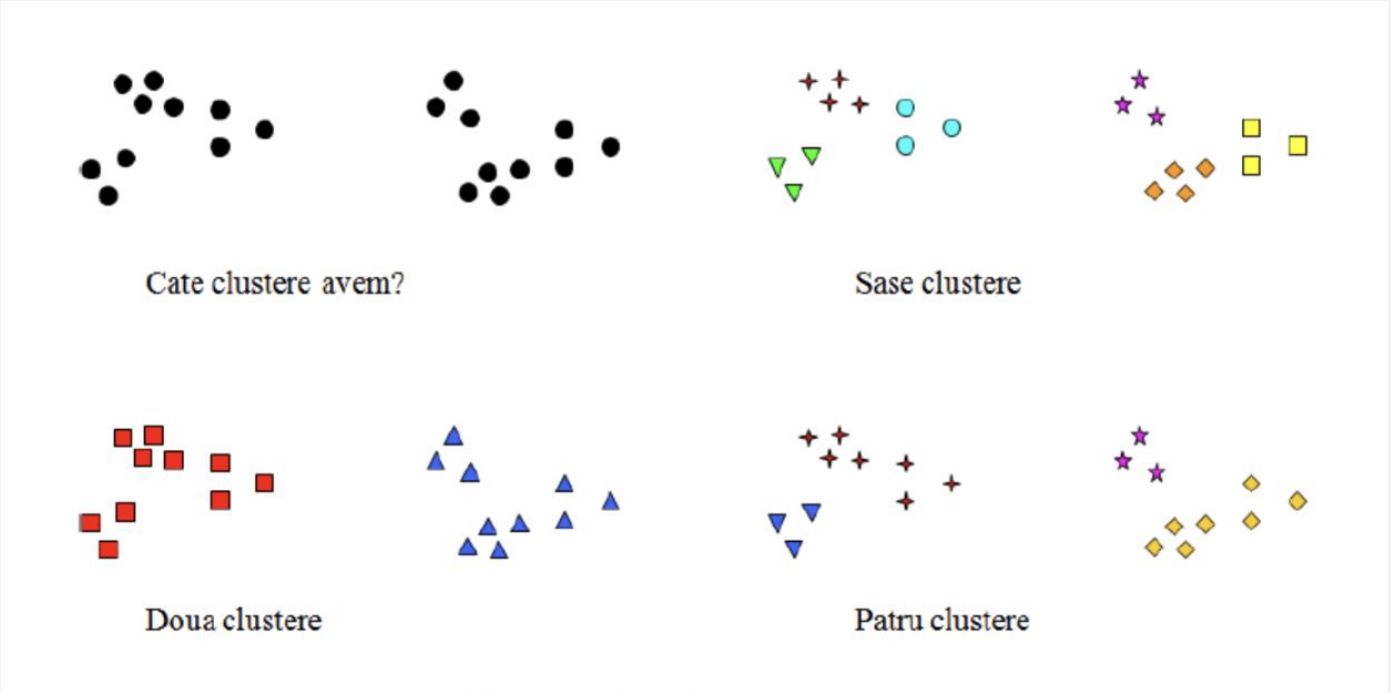
# Что не является кластеризацией?

❌ **Классификация с учителем:** кластеризация – это обучение без учителя.

❌ **Простая сегментация данных:** например, разделение студентов по первой букве фамилии.

❌ **Разбиение графов:** схожие элементы есть, но требуется более сложная спецификация.

# Неоднозначность понятия "кластер"



# Типы кластеризации

## 1️⃣ Иерархическая vs. Разделяющая (Hierarchical vs. Partitional)

- **Иерархическая кластеризация** строит древовидную структуру кластеров (дендрограмму).
- **Разделяющая кластеризация** сразу делит данные на заранее заданное число кластеров.

# Типы кластеризации

## 2️⃣ Эксклюзивная vs. Перекрывающаяся vs. Fuzzy

- **Эксклюзивная** – каждый объект принадлежит **только одному** кластеру.
- **Перекрывающаяся (Overlapping)** – один объект может входить в **несколько кластеров**.
- **Fuzzy** – объект имеет **вероятностную принадлежность** к разным кластерам.

# Типы кластеризации

### 3️⃣ Полная vs. Частичная (Complete vs. Partial)

- **Полная** – все данные распределяются по кластерам.
- **Частичная** – некоторые объекты могут не входить ни в один кластер (например, шумовые точки).

# Иерархическая vs. Разделяющая кластеризация

**Кластеризация** может быть с вложенностью **(иерархическая)** или без **(разделяющая).**

**Разделяющая кластеризация (Partitional Clustering)**

✅ Данные **разбиваются** на **неперекрывающиеся** подмножества (кластеры).

✅ **Каждый объект принадлежит только одному кластеру**.

**Пример:** K-Means, K-Medoids.

# Иерархическая vs. Разделяющая кластеризация

**Иерархическая кластеризация (Hierarchical Clustering)**

✅ Позволяет **создавать иерархию** вложенных кластеров.

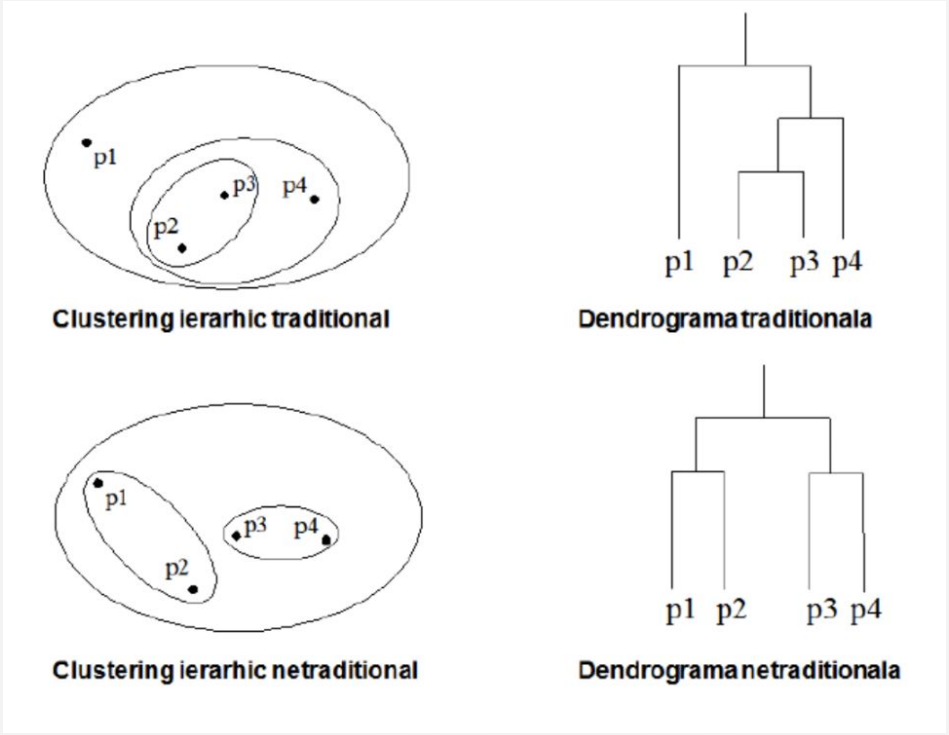
✅ **Каждый внутренний узел** объединяет кластеры, представленные дочерними узлами.

Возможны два подхода:

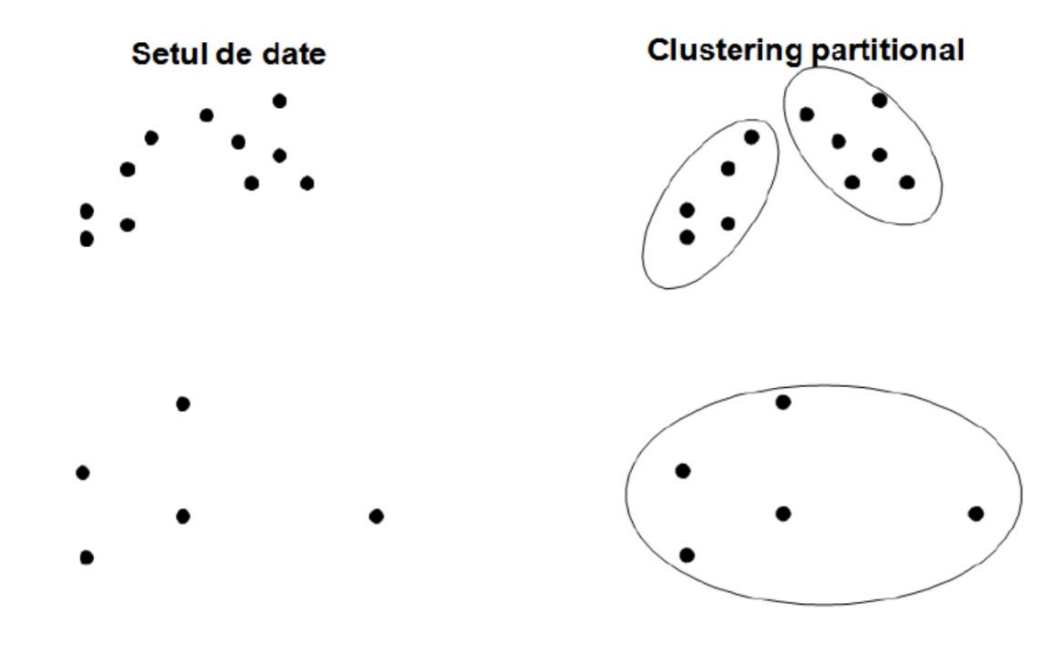
- **Агломеративный (снизу вверх):** начинается с отдельных точек и объединяет их в более крупные кластеры.
- **Дивизивный (сверху вниз):** начинается с одного большого кластера и рекурсивно разбивается на меньшие.

**Пример:** алгоритм связности (Single-Linkage, Complete-Linkage).

# Иерархическая кластеризация



# Разделяющая кластеризация



# Эксклюзивная vs. Перекрывающаяся vs. Fuzzy кластеризация

**Кластеризация** может быть **жесткой (эксклюзивной), перекрывающейся** или **нечеткой (fuzzy).**

**1. Эксклюзивная кластеризация (Exclusive Clustering)**

✅ **Каждый объект принадлежит только одному кластеру.**

✅ Четкое разделение данных.

**Пример:** K-Means, DBSCAN.

# Эксклюзивная vs. Перекрывающаяся vs. Fuzzy кластеризация

**2. Перекрывающаяся кластеризация (Overlapping Clustering)**

✅ **Объект может принадлежать нескольким кластерам одновременно.**

✅ Полезно, когда данные не имеют четких границ.

**Пример:** мягкие версии K-Means, многозначные разбиения (Multi-Assignment Clustering).

# Эксклюзивная vs. Перекрывающаяся vs. Fuzzy кластеризация

### 3. Fuzzy-кластеризация

✅ **Каждый объект имеет степень принадлежности (fuzzy membership) к каждому кластеру.**

✅ Кластеры становятся **нечеткими множествами**.

✅ Можно преобразовать fuzzy-кластеры в эксклюзивные, **выбирая наиболее вероятный кластер**.

**Пример:** Fuzzy C-Means (FCM).

# Clustering: полный vs. частичный

**Полная кластеризация:** Каждая точка данных принадлежит кластеру.

**Частичная кластеризация:** Некоторые данные могут не принадлежать ни одному кластеру (например, выбросы).

# Типы данных в кластерном анализе

# Структуры данных в кластеризации

- **Качество кластеризации** зависит от **метрики схожести** между объектами.
- **Сходство** выражается через функцию расстояния **d(i, j)**.
- **Качество кластера** определяется отдельной функцией.

# Проблемы измерения схожести

Разные метрики применяются к разным типам данных:

- Числовые переменные (интервальные шкалы).
- Булевы переменные.
- Категориальные и порядковые переменные.
- Данные **смешанного типа**.

Весовые коэффициенты могут зависеть от приложения и смысла данных.

Оценка **"достаточного сходства"** – субъективна.

# Типы данных в кластеризации

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

***Матрицы данных***
***(Data Matrix)***

***Матрицы (дис)сходства***
***(Dissimilarity Matrix)***

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

# Типы данных в кластеризации

1. **Интервальные переменные (Interval-scaled variables)**

***Пример:*** температура, рост, вес.

Используются метрики расстояния: Евклидово, Манхэттенское.

2. **Бинарные переменные (Binary variables)**

Да/Нет, 1/0 (наличие или отсутствие свойства).

***Расстояние:*** коэффициент Жаккара.

# Типы данных в кластеризации

3️⃣ **Номинальные, порядковые, ratio-переменные**

***Пример:*** цвет, категория продукта, уровень образования.

***Методы:*** хи-квадратное расстояние, ранговые коэффициенты.

4️⃣ **Смешанные переменные (Mixed-type variables)**

***Пример:*** база клиентов с числовыми и категориальными характеристиками.

***Решение:*** гибридные меры расстояний (например, Gower distance).

# Нормализация и стандартизация данных

- **Среднеквадратическое отклонение (z-score)**

$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

- **Медианное абсолютное отклонение (MAD)** – более устойчивая метрика.

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \ldots + |x_{nf} - m_f|)$$

$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + \ldots + x_{nf})$$

Стандартизация позволяет учитывать переменные с разными масштабами.

# Классификация методов кластеризации

# Основные методы кластеризации

✅ **Методы разбиения (Partitioning Methods)**

- Формируют **различные разбиения** данных.
- Оценивают их качество с помощью определенных критериев.

✅ **Иерархические методы (Hierarchical Methods)**

- Строят **иерархическую структуру кластеров**.
- Используют различные критерии объединения или разделения.

# Основные методы кластеризации

## ✅ Методы на основе плотности (Density-Based Methods)

- Основаны на функциях **связанности и плотности точек**.
- Позволяют выявлять **кластеры сложной формы**.

## ✅ Методы на основе сетки (Grid-Based Methods)

- Используют **многоуровневые структуры** для организации данных.
- Разбивают пространство данных на **регулярную сетку**.

# Основные методы кластеризации

## ✅ Методы на основе моделей (Model-Based Methods)

- Для каждого кластера строится **статистическая модель**.
- Выбирается **наиболее подходящая модель** среди возможных вариантов.

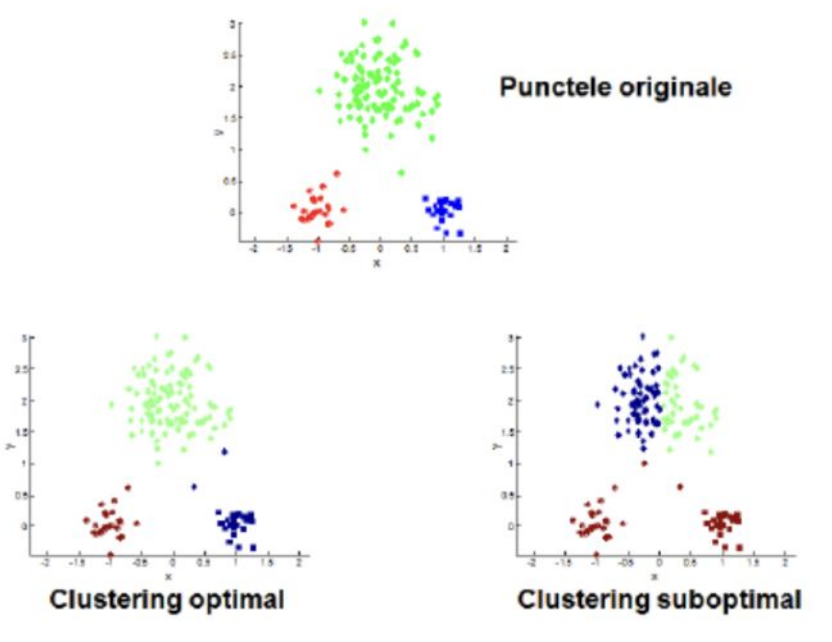
# Алгоритмы разбиения

**Метод разбиения:**

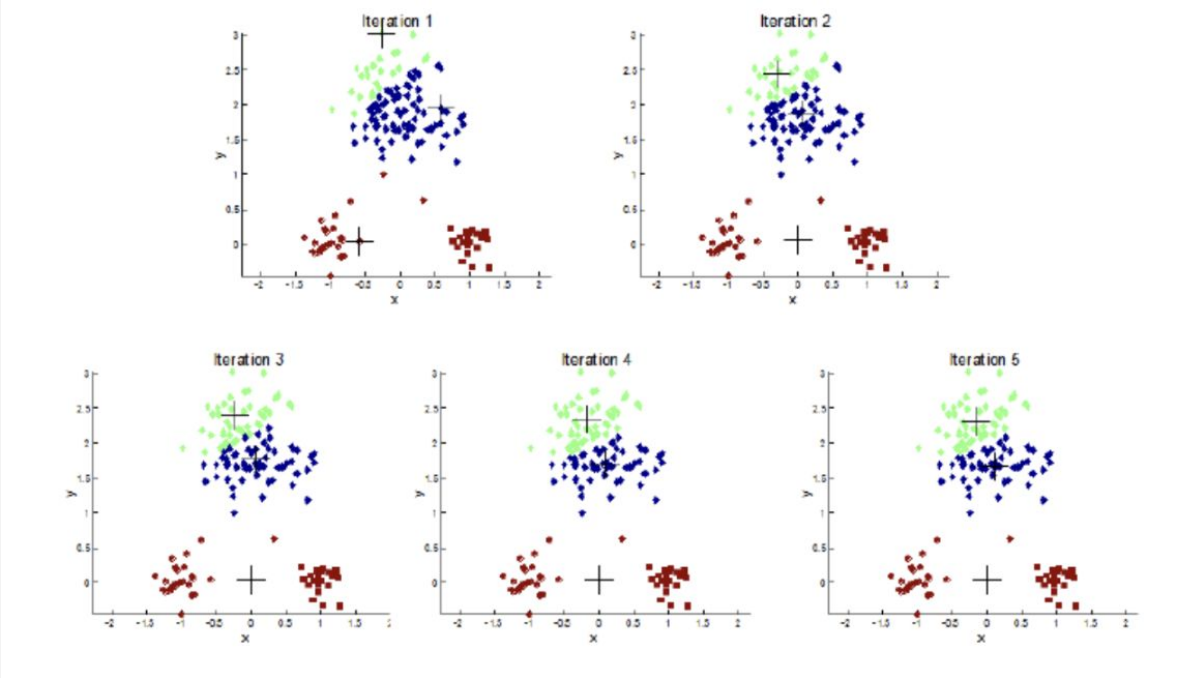
- Данные **разбиваются на k кластеров**.
- Оптимизация разбиения по выбранному критерию.

- **Глобальный оптимум:** полный перебор всех возможных разбиений.
- **Эвристические методы:**
  - **k-means (MacQueen, 1967)** – каждый кластер представлен **центроидом**.
  - **k-medoids (PAM, Kaufman & Rousseeuw, 1987)** – кластер представлен **реальным объектом** из данных.

# Кластеризация: Оптимальное vs. Неоптимальное разбиение (SSE)

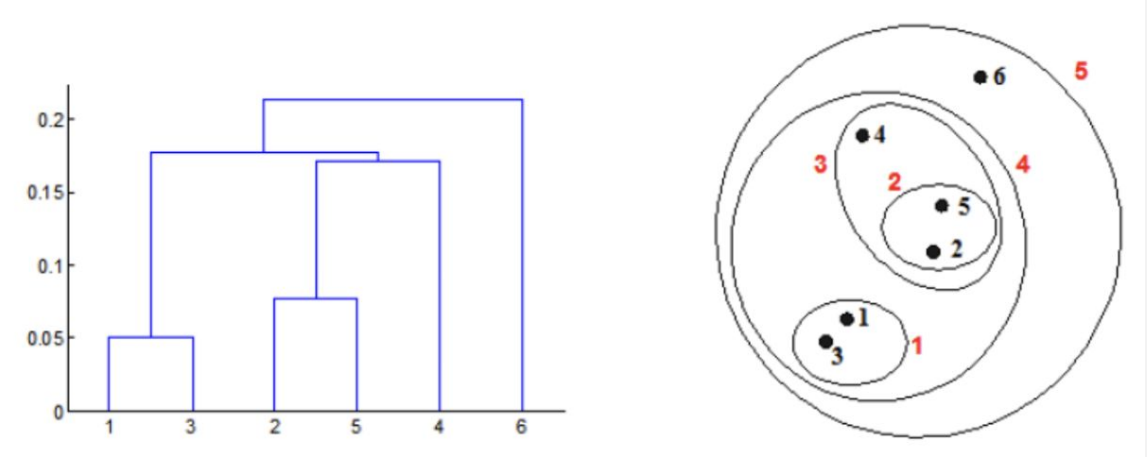


# Кластеризация при неудачном выборе начальных центроидов



# Иерархическая кластеризация

Иерархическая кластеризация создает вложенные кластеры, организованные в виде дерева. Визуализируется с помощью **дендрограммы** – графика, отображающего объединение или разделение кластеров.



# Преимущества

✅ Не требует заранее заданного количества кластеров – можно определить их, **обрезав дендрограмму** на нужном уровне.

✅ Используется в **таксономии**, например, в биологии для классификации видов.

# Иерархическая кластеризация

## 1️⃣ Агломеративная кластеризация (снизу вверх)

- Каждое наблюдение начинается как отдельный кластер.
- Постепенно объединяются **наиболее похожие кластеры**.
- Останавливается, когда остается нужное число кластеров.

# Иерархическая кластеризация

## 2️⃣ Дивизивная кластеризация (сверху вниз)

- Начинается с **одного большого кластера**, содержащего все данные.
- Постепенно разделяется на **меньшие кластеры**.
- Останавливается, когда достигается нужное число кластеров.

# Агломеративная иерархическая кластеризация