Обучение без учителя: кластеризация

Максим Карпов

Центр непрерывного образования ФКН ВШЭ

План

- Кластеризация
  - K-Means
  - DBSCAN
  - Иерархическая кластеризация
  - Метрики качества

Обучение с учителем (supervised learning)

- Для каждого объекта известен ответ (класс или число)
- Даны примеры объектов с ответами
- Нужно построить модель, которая будет предсказывать ответы для новых объектов

Обучение без учителя (unsupervised learning)

- Даны объекты
- Нужно найти в них внутреннюю структуру
- Примеры:
  - Кластеризация
  - Обнаружение аномалий
  - Тематическое моделирование
  - Визуализация
  - Предсказание следующего кадра видео
  - …
- Ближе к обучению в реальной жизни

Кластеризация

- Дано: матрица «объекты-признаки» X
- Найти:
  1. Множество кластеров Y
  2. Алгоритм кластеризации $a(x)$, который приписывает каждый объект к одному из кластеров

- Каждый кластер состоит из похожих объектов
- Объекты из разных кластеров существенно отличаются

Отличия

Обучение с учителем

- Цель: минимизация функционала ошибки
- Множество ответов известно заранее
- Конкретные способы измерения качества

Кластеризация

- Нет строгой постановки
- Множество кластеров неизвестно
- Правильные ответы отсутствуют (в большинстве случаев) — нельзя измерить качество

Зачем кластеризовать?

- Маркетинг: искать похожих клиентов
- Модерация: проверять только одно сообщение из кластера
- Соц. опросы: выделять группы схожих анкет
- Соц. сети: искать сообщества

- Выявлять типы людей и формировать поведенческие паттерны для каждого типа

Важно

- Алгоритм кластеризации не знает, чего вы хотите
- Не стоит ожидать, что при кластеризации текстов вы получите разбиение именно по темам
- Нередко кластеры оказываются неинтерпретируемыми

Обучение без учителя: кластеризация

ШАГ I
Анализ данных,
в т.ч. транзакционных
Way4, ЦОД, кред. фабрика

ШАГ II
Выявление паттернов и сегментация клиентов по характеристикам

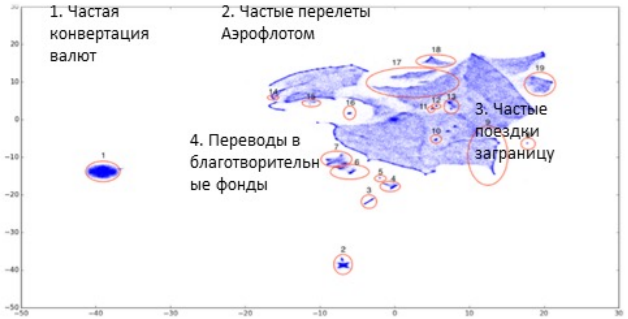
ШАГ III
Формирование продуктовых предложений на базе характеристик клиента

МЕТОДЫ алгоритмы кластеризации, визуализация данных большой размерности с использованием LargeVis

ЭКОНОМИЧЕСКИЙ ЭФФЕКТ

- ✓ Рост эффективности воронки продаж
- ✓ Рост лояльности клиентов

КЛАСТЕРИЗАЦИЯ КЛИЕНТОВ ПО ХАРАКТЕРУ ТРАНЗАКЦИЙ



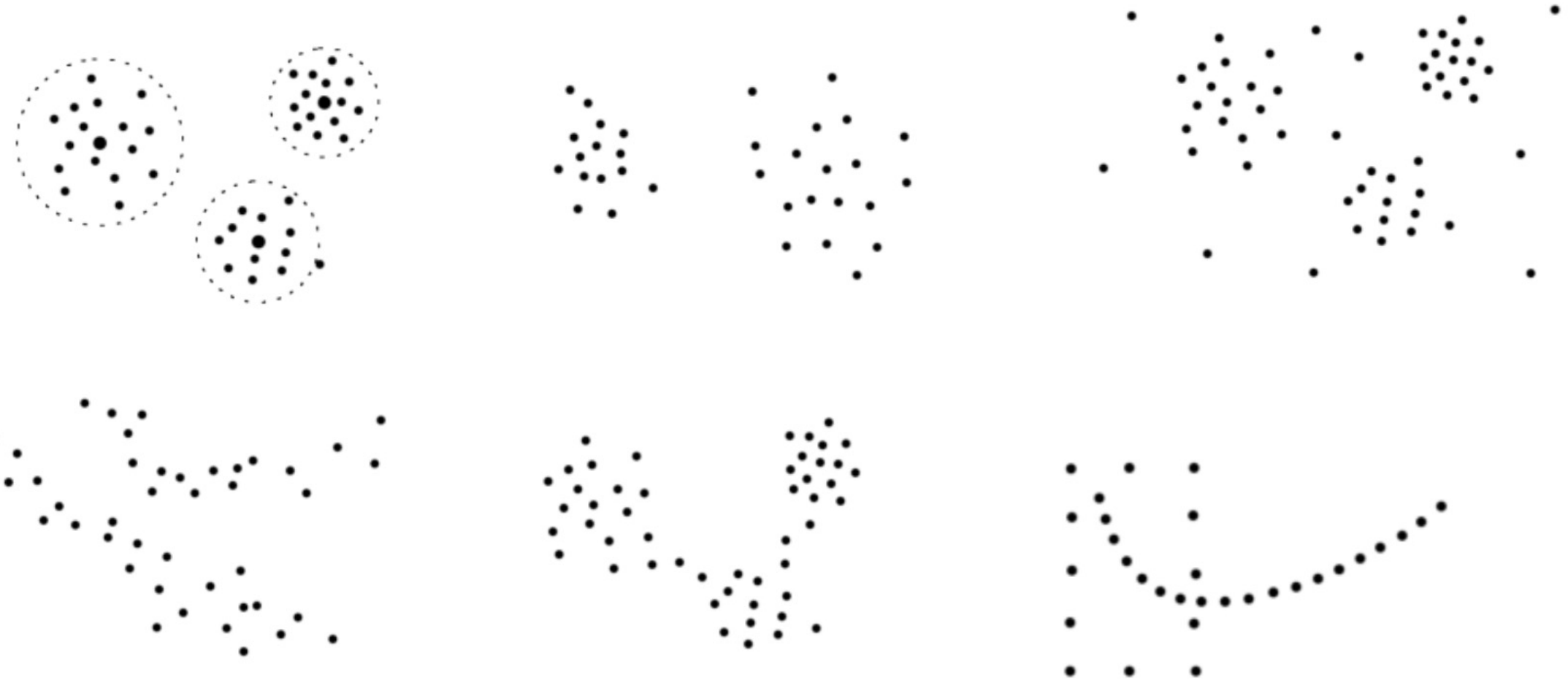
В ЗАВИСИМОСТИ ОТ КЛАСТЕРА КЛИЕНТА ПРЕДЛОЖИТЬ РЕЛЕВАНТНЫЙ ПРОДУКТ

Паттерн	Продукт
1. Частая конвертация валют	Мультивалютный счет
2. Частые перелеты Аэрофлотом	Карта «Аэрофлот Бонус»
3. Частые поездки заграницу	Страховка для выезжающих за рубеж
4. Переводы в благотворительные фонды	Карта «Подари жизнь»

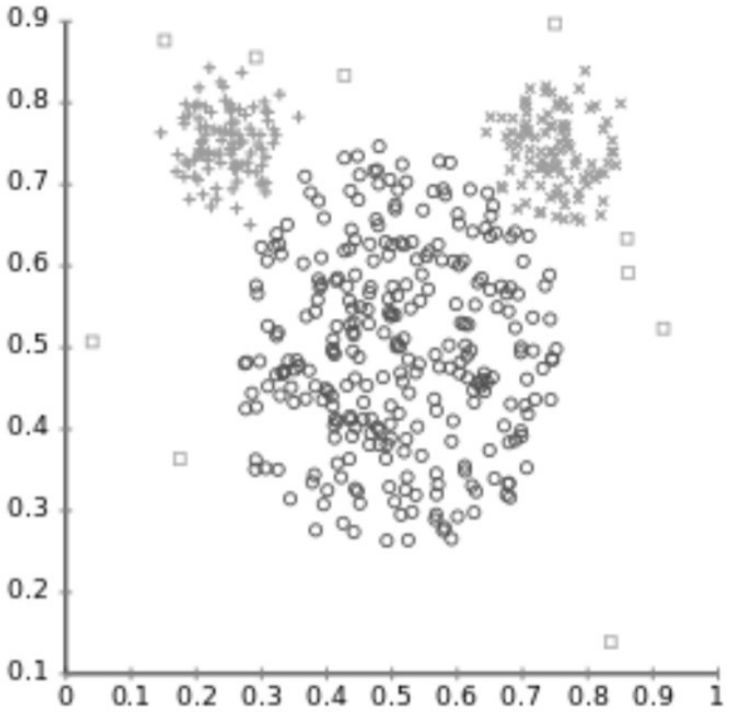
https://habrahabr.ru/article/318152/

Виды кластеризации

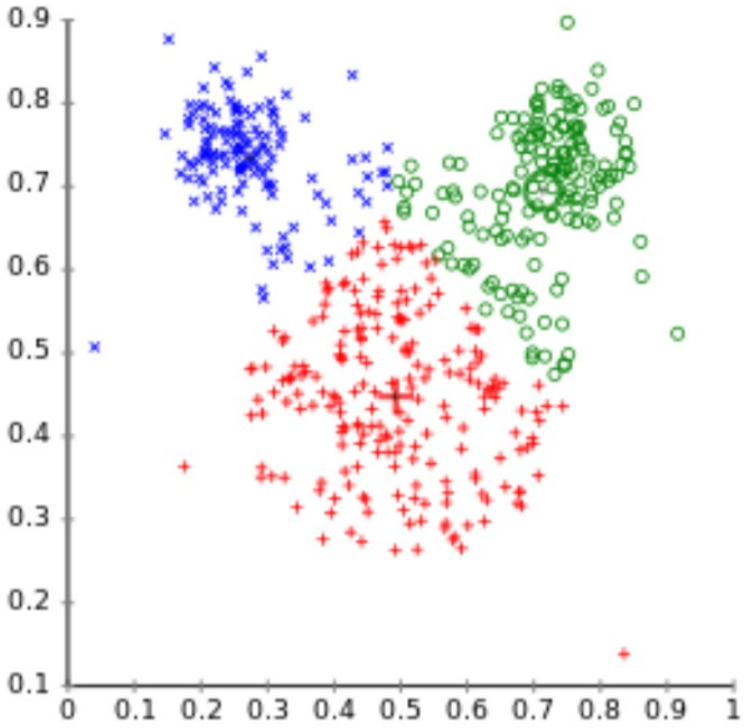
Форма кластеров



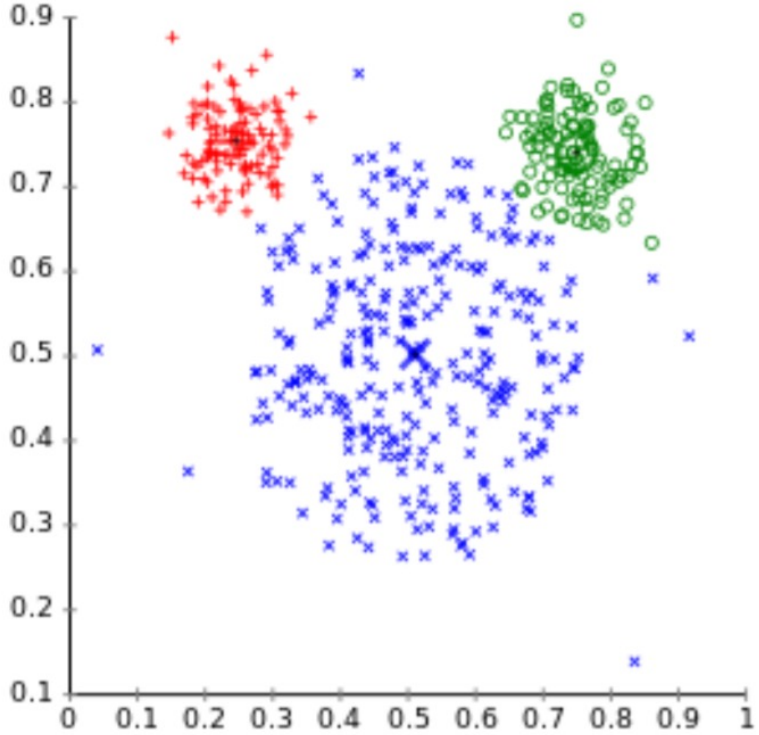
Различия в результатах работы



Исходная выборка
("Mouse" dataset)



Метод 1



Метод 2

Требования к кластерам

- Задача кластеризации новостей по содержанию.
- Постановка 1: в один кластер должны попадать новости на одну тему

Батыршин сыграет вместо Хабарова у «Магнитки» в матче с «Салаватом»

Место в третьей паре защиты «Магнитки» на третью встречу плей-офф Кубка Гагарина с «Салаватом Юлаевым» занял защитник Рафаэль Батыршин, сообщает из Уфы корреспондент «Чемпионата» Павел Панышев. Травмированный Ярослав Хабаров выбыл на неопределённый срок. Для форварда Оскара Осалы сезон закончен.

Футболисты ЦСКА проиграли «Долгопрудному» в товарищеском матче

Футболисты московского ЦСКА со счетом 2:3 проиграли клубу второго дивизиона "Долгопрудный" в товарищеском матче, который состоялся в Москве на стадионе "Октябрь". У армейцев забитыми мячами отличились Александр Цауня (15-я минута) и Сергей Ткачев (54).

Скриншот с сайта Яндекс.Новости (news.yandex.ru)

Требования к кластерам

- Задача кластеризации новостей по содержанию.
- Постановка 2: в один кластер должны попадать новости об одном «большом» событии

Скриншот с сайта РИА Новости (ria.ru)

Требования к кластерам

- Задача кластеризации новостей по содержанию.
- Постановка 3: в один кластер должны попадать тексты об одной и той же новости

11:41, 08 ФЕВРАЛЯ 2014

Открытие Олимпиады в Сочи посмотрели несколько миллиардов человек

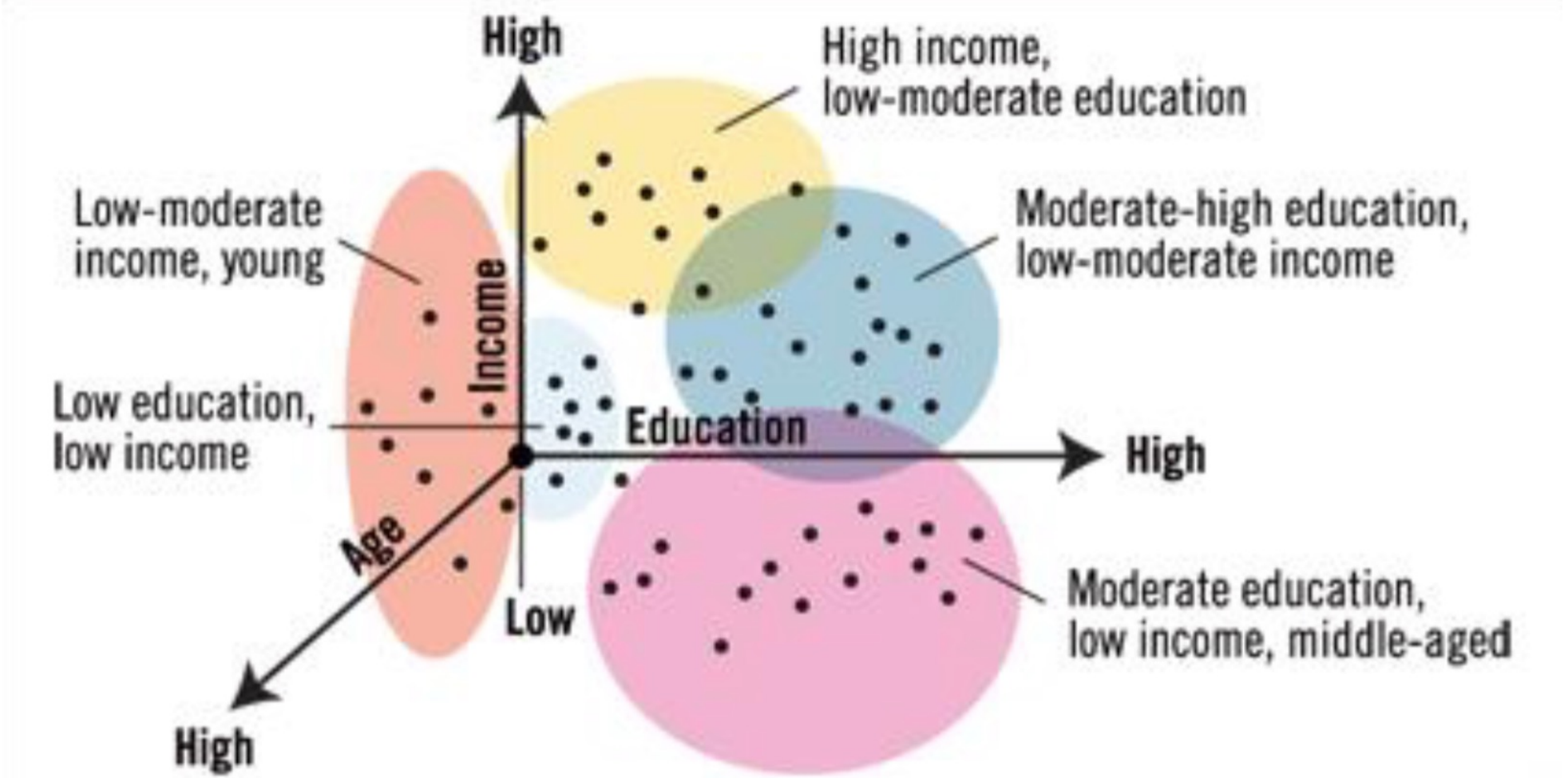
Олимпиада в Сочи открыта

Церемония открытия Олимпиады в Сочи. Онлайн-репортаж

Требования к кластерам

- Чтобы проверить, выполняются ли требования, нужно делать разметку данных
- Для новостей: показывать асессору пары документов и спрашивать, относятся ли они к одному кластеру

Кластеризация как основная задача



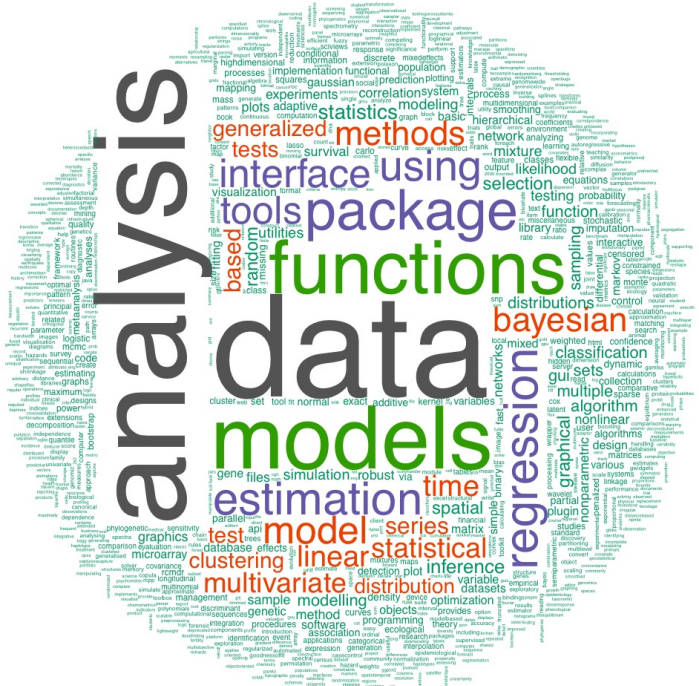
Кластеризация как вспомогательная задача

Цель: улучшение распознавания

«Жесткая» и «мягкая» кластеризации

Кластеризация для выделения «тем»

0.2



0.3

0.5

Типы задач кластеризации

- Форма кластеров, которые нужно выделять
- Плоская или древовидная структура
- Размер кластеров
- Конечная задача или вспомогательная
- Жесткая или мягкая кластеризация

K-Means

K-Means

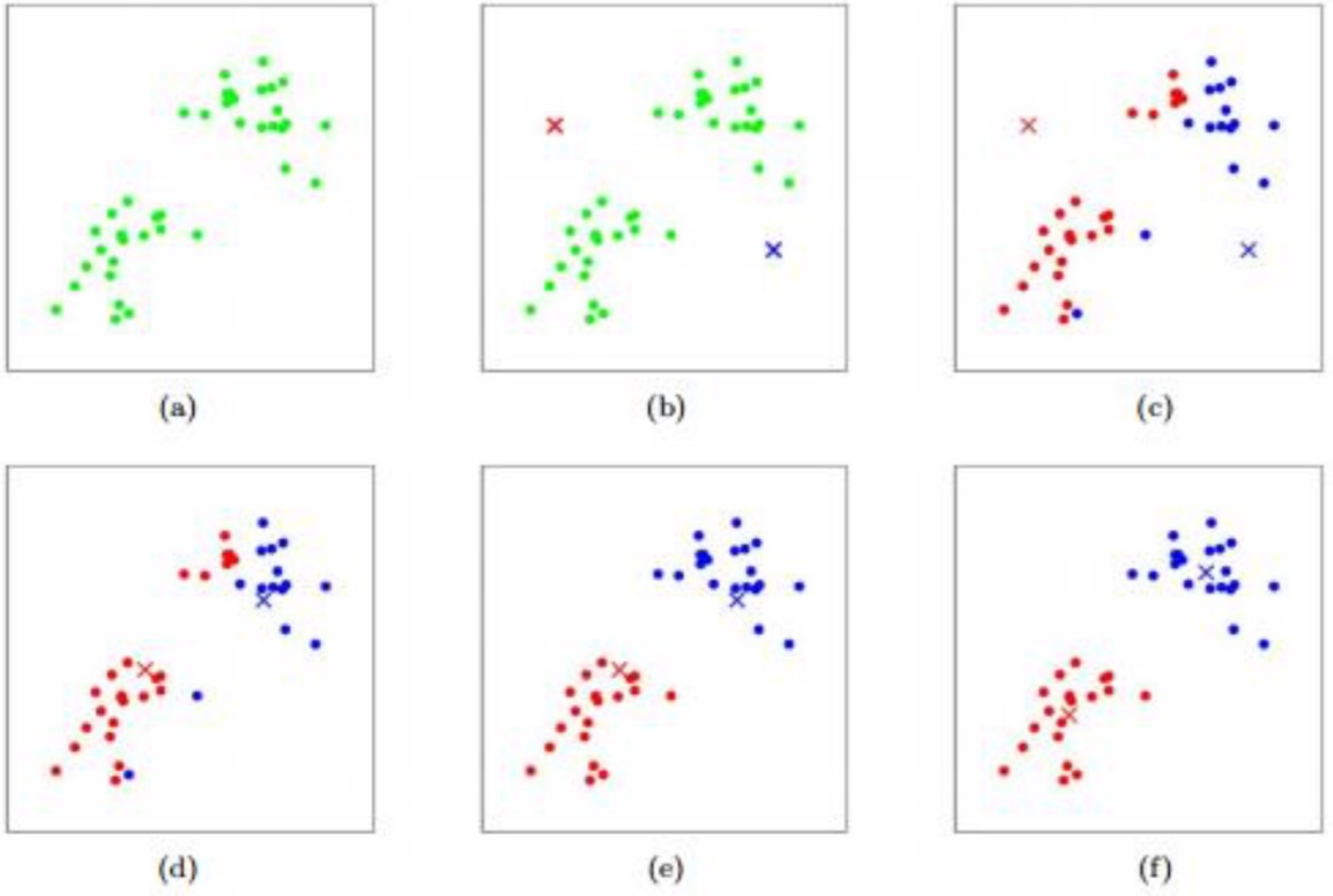
- Дано: выборка $x_1, \dots, x_\ell$
- Параметр: число кластеров K
- Начало: случайно выбрать K центров кластеров $c_1, \dots, c_K$
- Повторять по очереди до сходимости:
  - Шаг А: отнести каждый объект к ближайшему центру

$$y_i = \arg\min_{j=1,\dots,K} \rho(x_i, c_j)$$

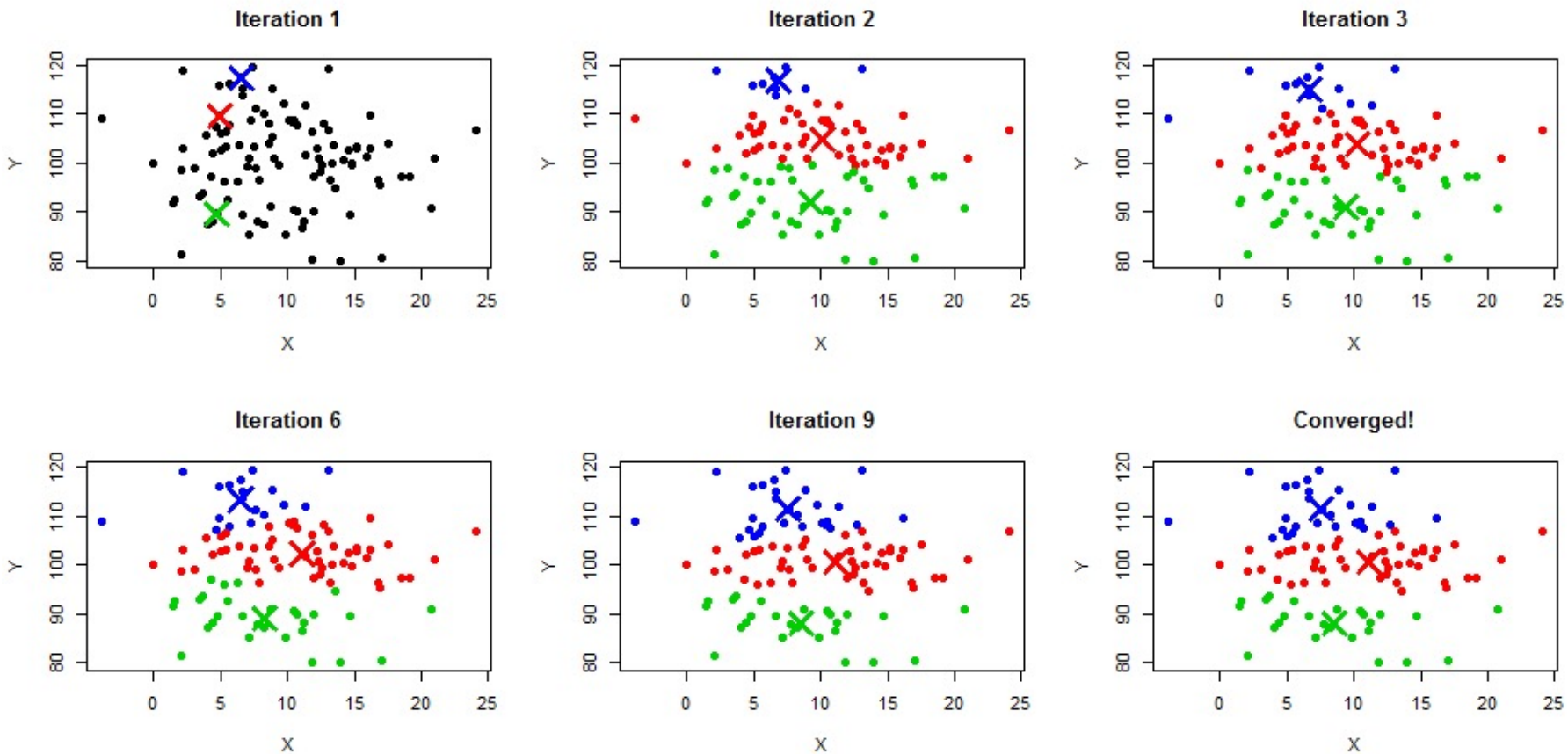
  - Шаг Б: переместить центр каждого кластера в центр тяжести

$$c_j = \frac{\sum_{i=1}^{\ell} x_i [y_i = j]}{\sum_{i=1}^{\ell} [y_i = j]}$$

K-Means



K-Means


Iteration 1
Iteration 2
Iteration 3
Iteration 6
Iteration 9
Converged!
X
Y
0
5
10
15
20
25
80
90
100
110
120

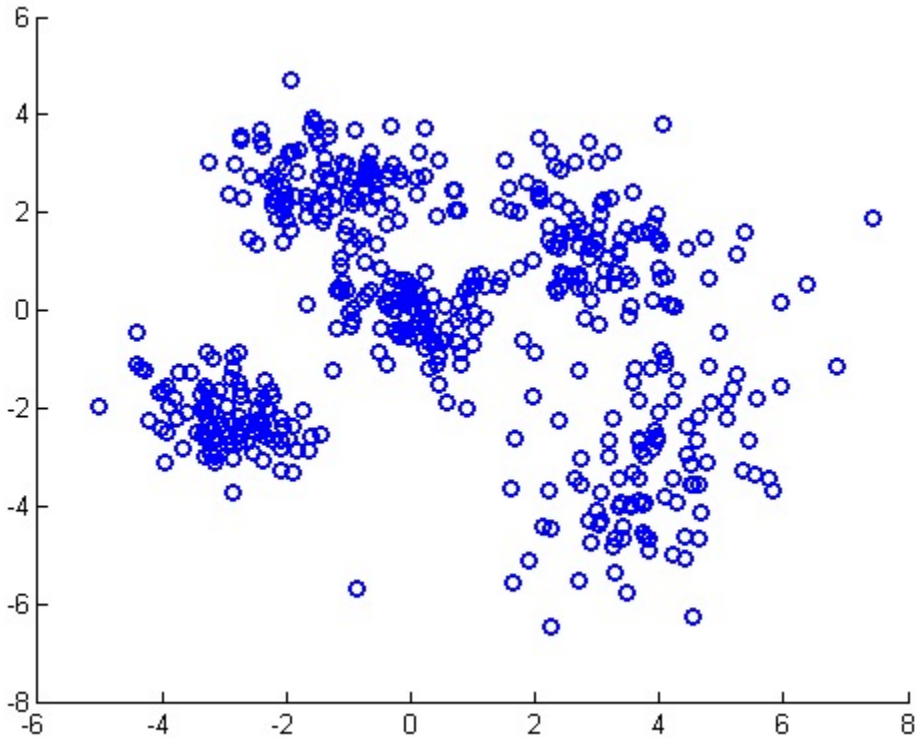
Выбор числа кластеров

- Качество кластеризации: внутрикластерное расстояние

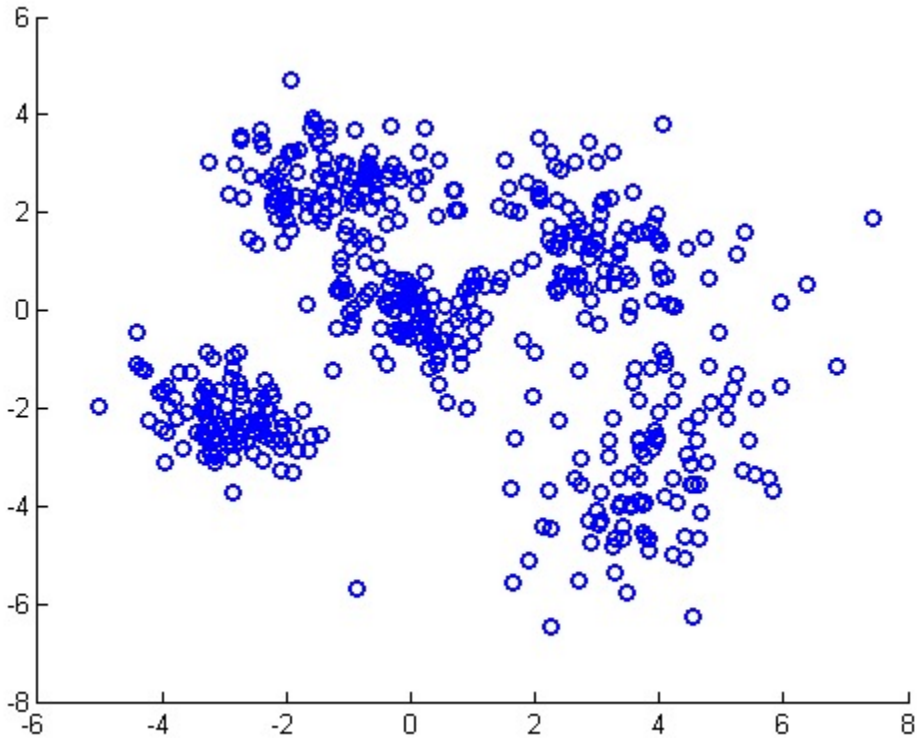
$$J(C) = \sum_{i=1}^{\ell} \rho\left(x_i, c_{y_i}\right)$$

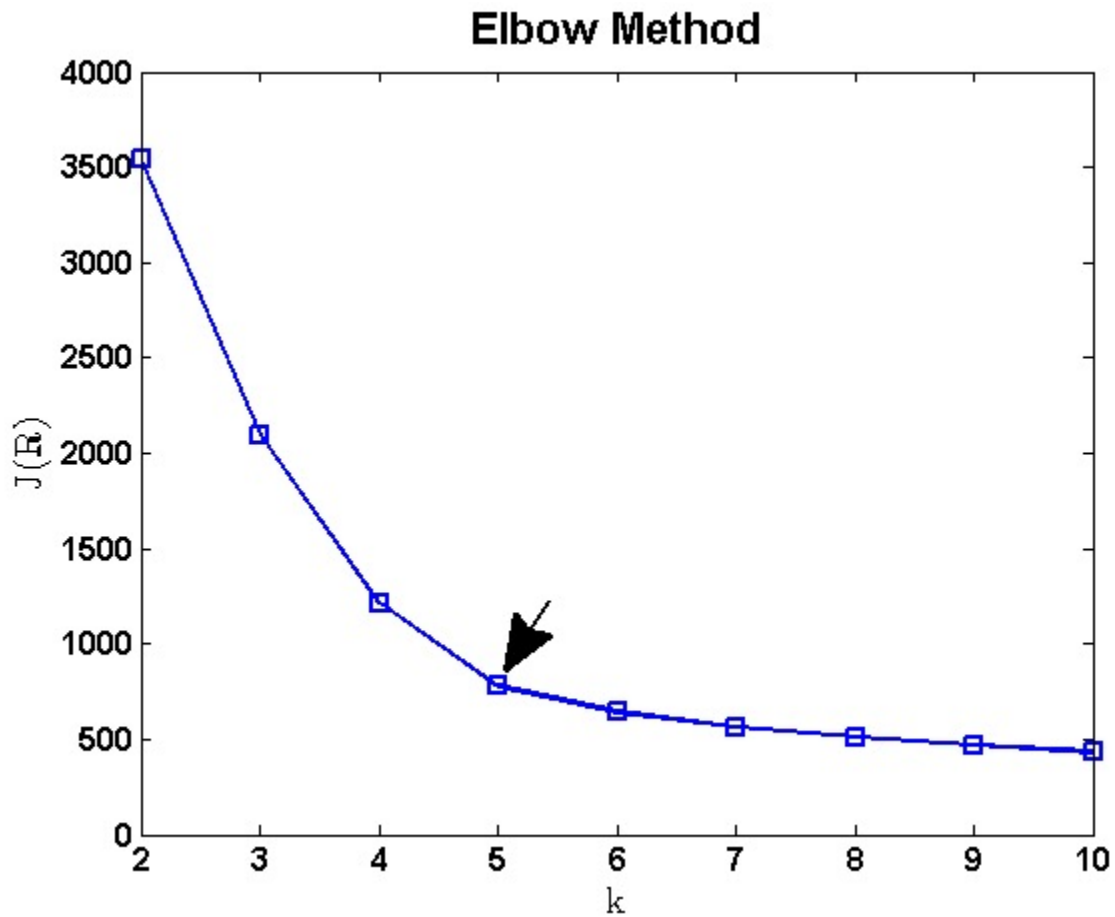
- Зависит от K
- Нужно подобрать такое K, после которого качество меняется не слишком сильно

Выбор числа кластеров



Выбор числа кластеров



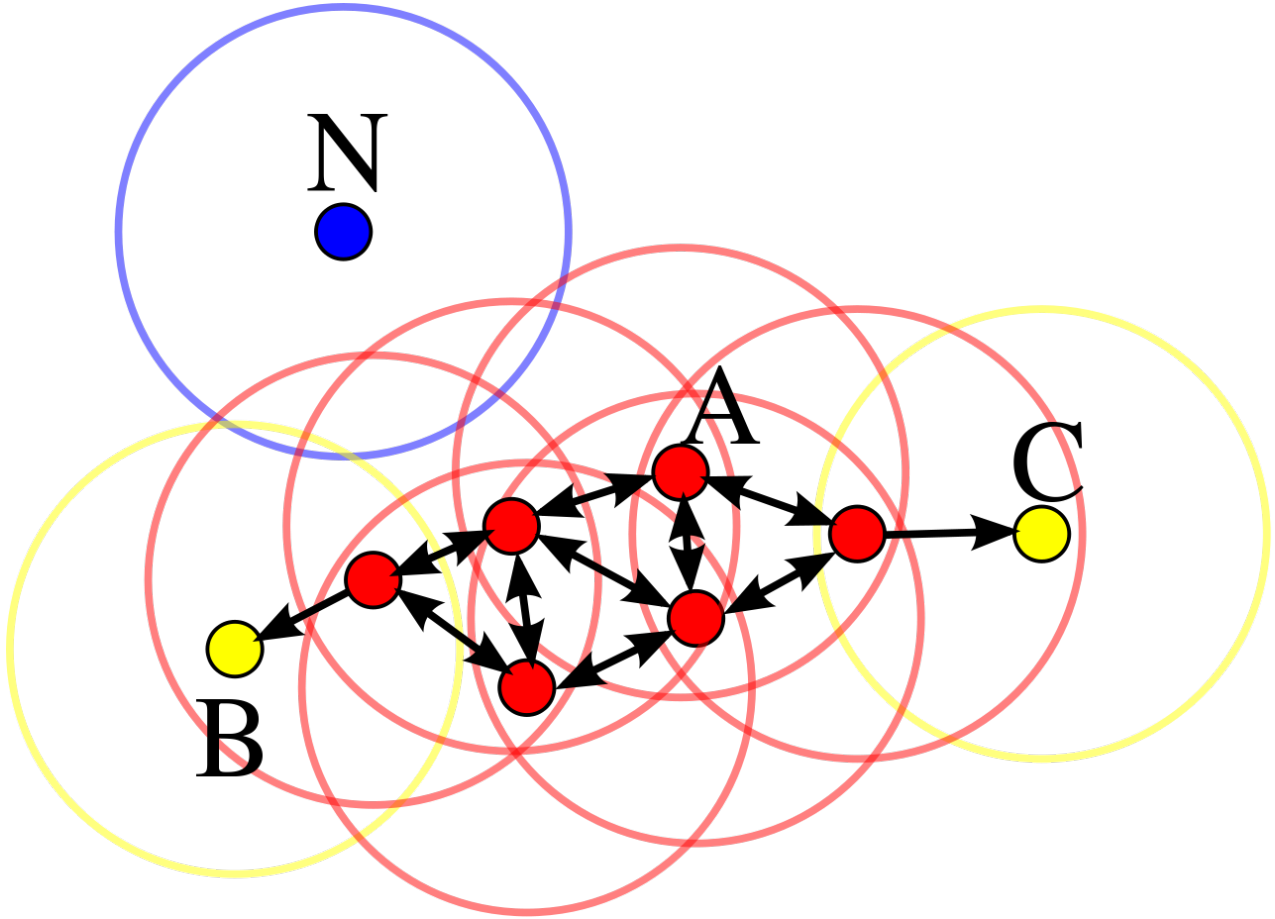


Особенности K-Means

- Может работать с большими объёмами данных
- Подходит для кластеров с простой геометрией
- Требует выбора числа кластеров

Density-based clustering

Основные, граничные и шумовые точки

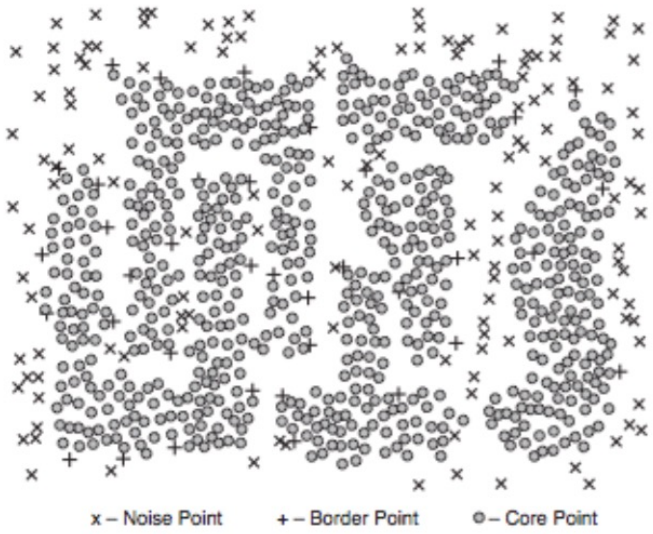


Параметры DBSCAN

- Размер окрестности (eps)
- Минимальное число объектов в окрестности — для определения основных точек

DBSCAN

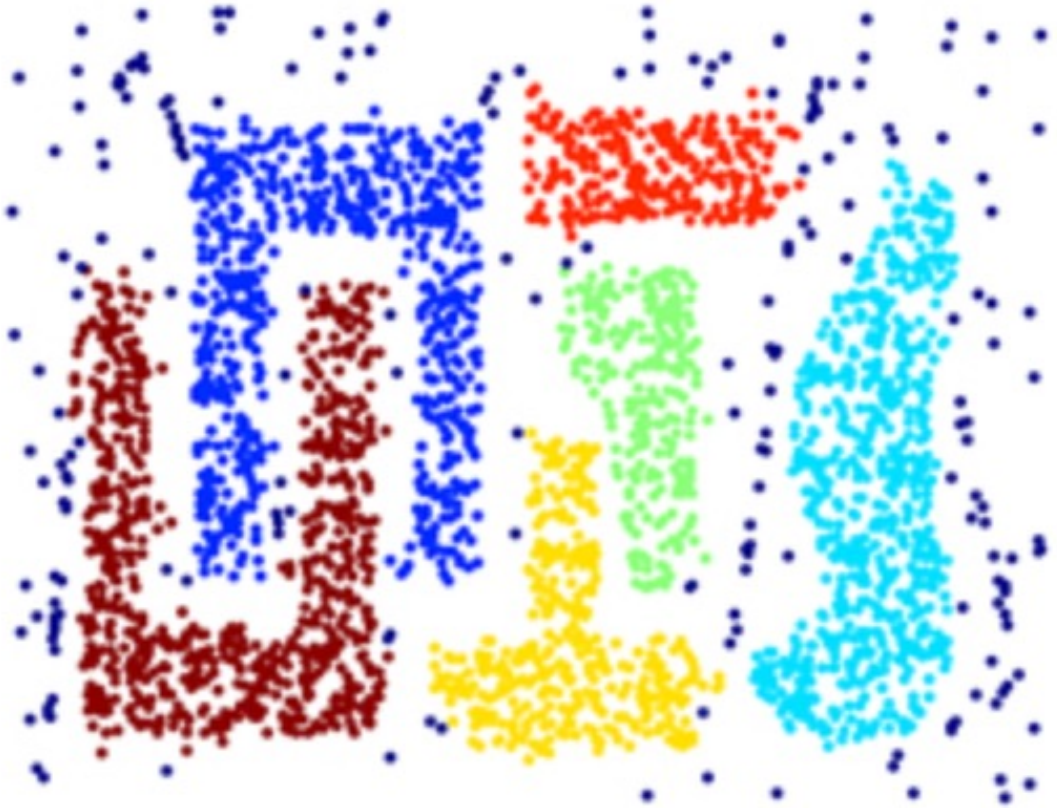
(a) Clusters found by DBSCAN.



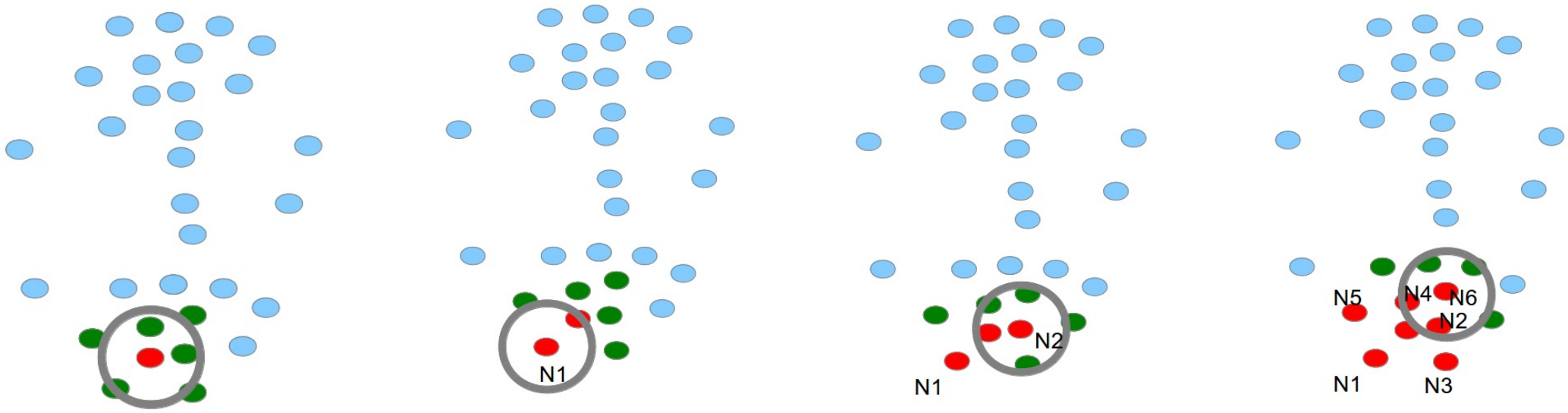
(b) Core, border, and noise points.

1. Выбрать точку без метки
2. Если в окрестности меньше N точек, то пометить как шумовую
3. Создать новый кластер, поместить в него текущую точку
4. Для всех точек из окрестности S: (а) если точка шумовая, то отнести к данному кластеру, но не использовать для расширения; (б) если точка основная, то отнести к данному кластеру, а её окрестность добавить к S
5. Перейти к шагу 1

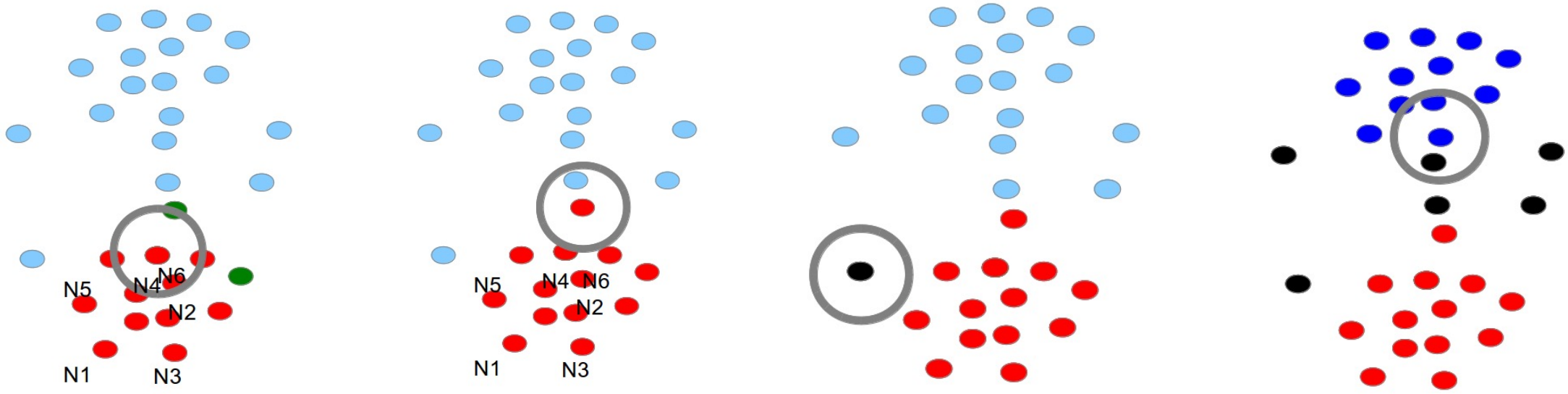
DBSCAN: результаты работы



Пример



Пример



Особенности DBSCAN

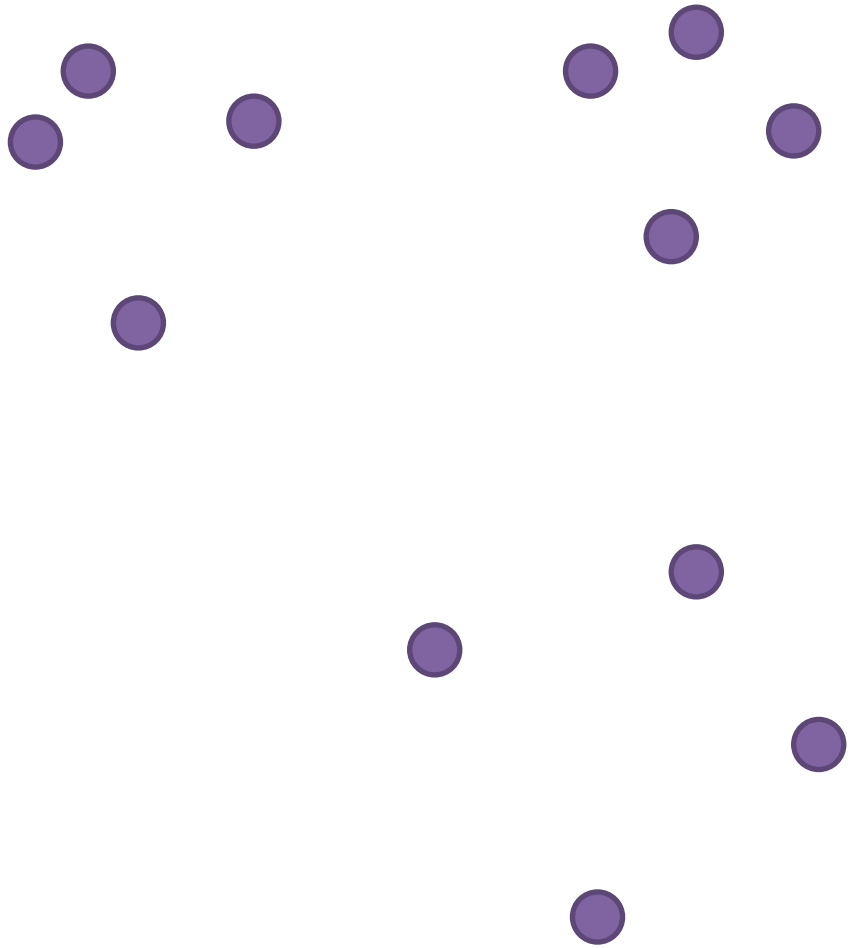
- Находит кластеры произвольной формы
- Может работать с большими объёмами данных
- Нужно подбирать размер окрестности (eps) и минимальное число объектов в окрестности

Иерархическая кластеризация

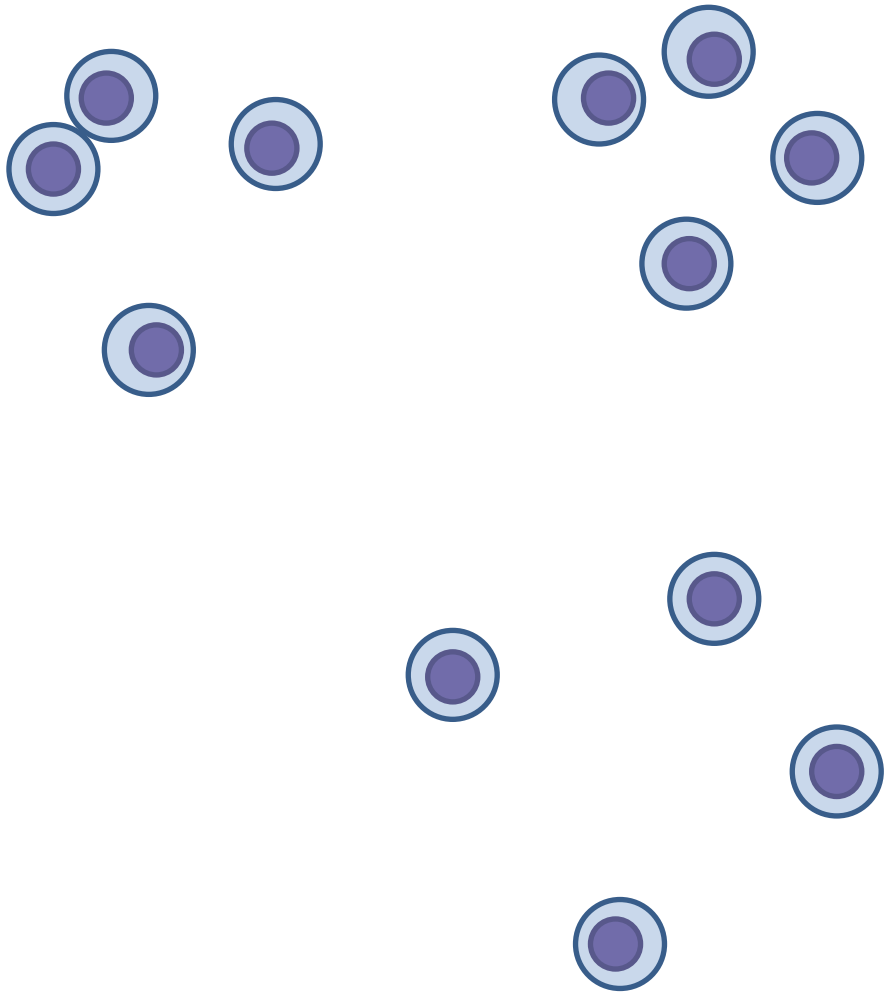
Виды иерархической кластеризации

- Аггломеративная – на каждой итерации объединяем два меньших кластера в один побольше
- Дивизивная – на каждой итерации делим один большой кластер на два поменьше

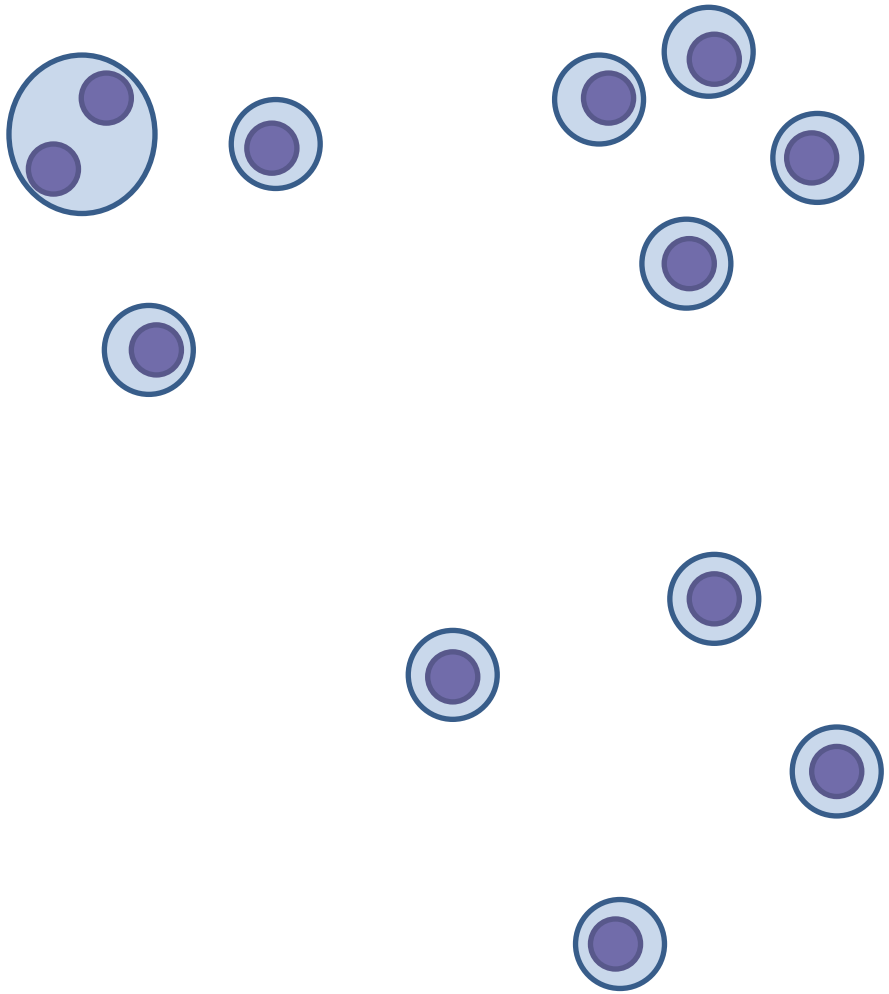
Агломеративная кластеризация



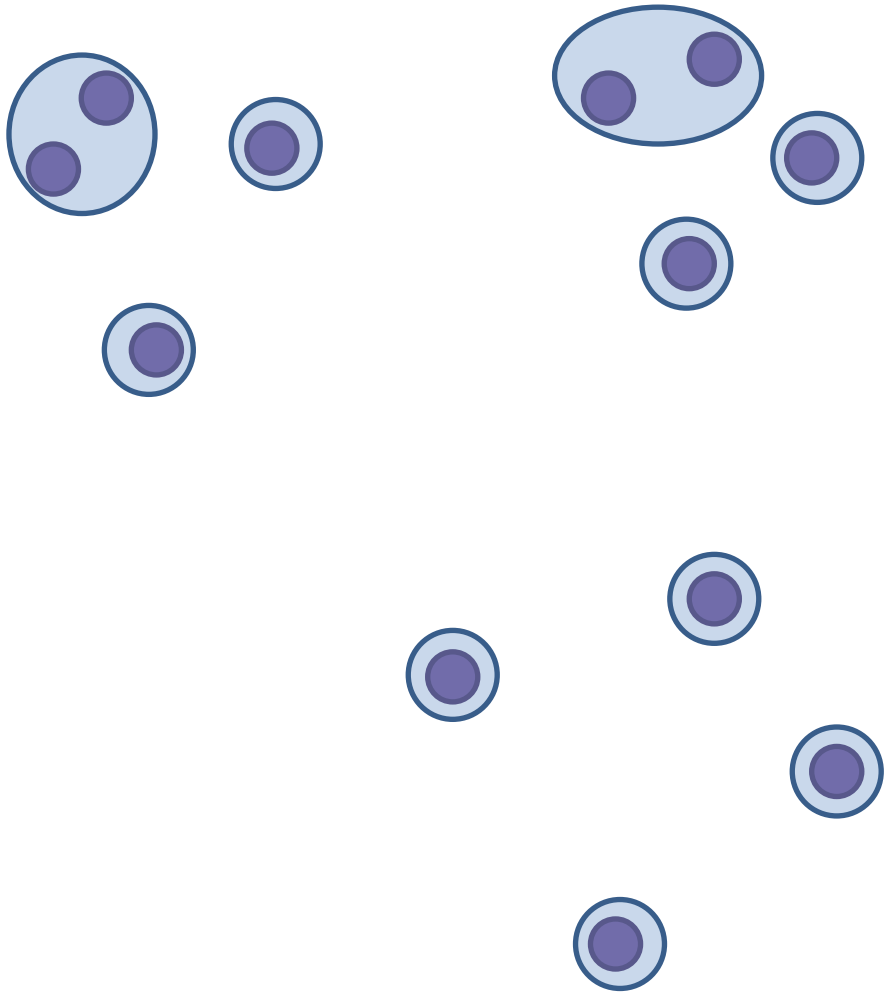
Агломеративная кластеризация



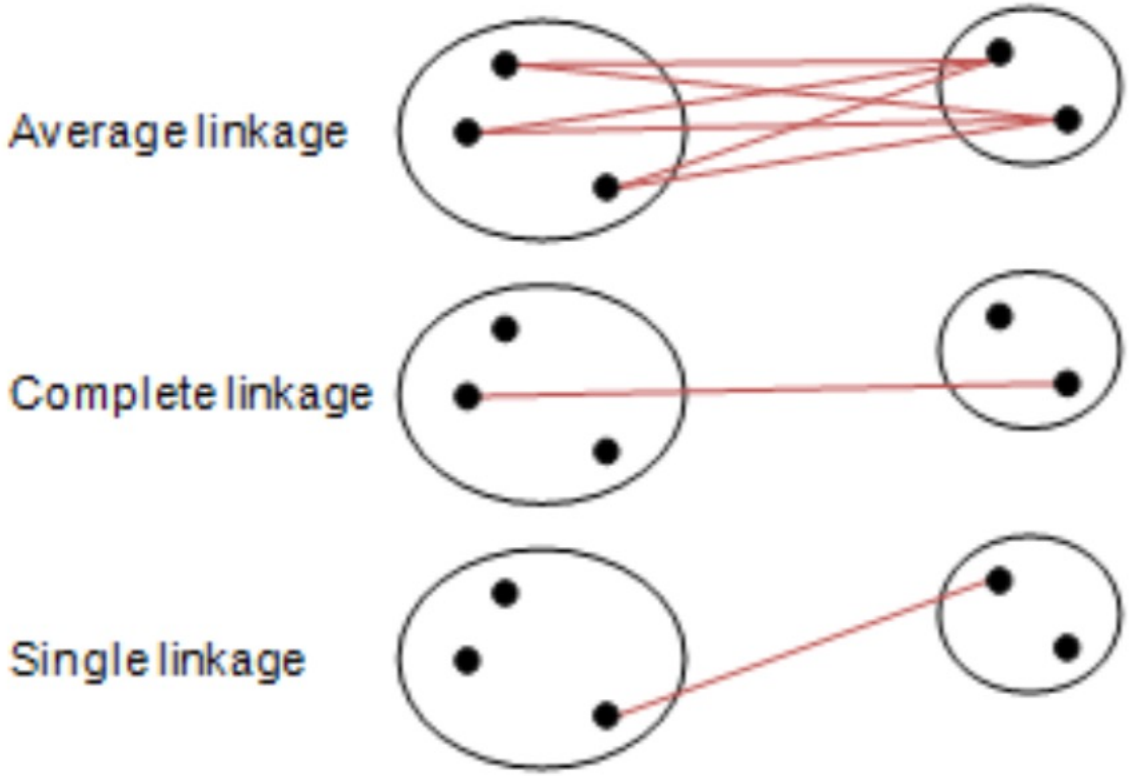
Агломеративная кластеризация



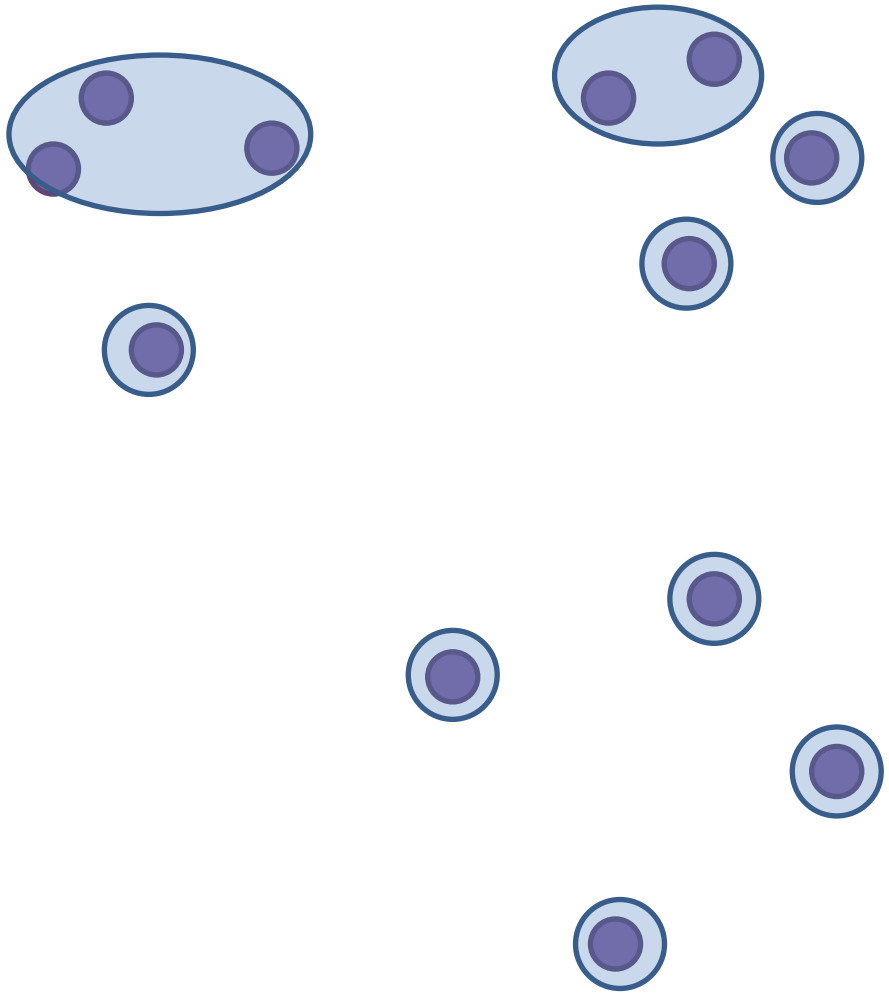
Агломеративная кластеризация



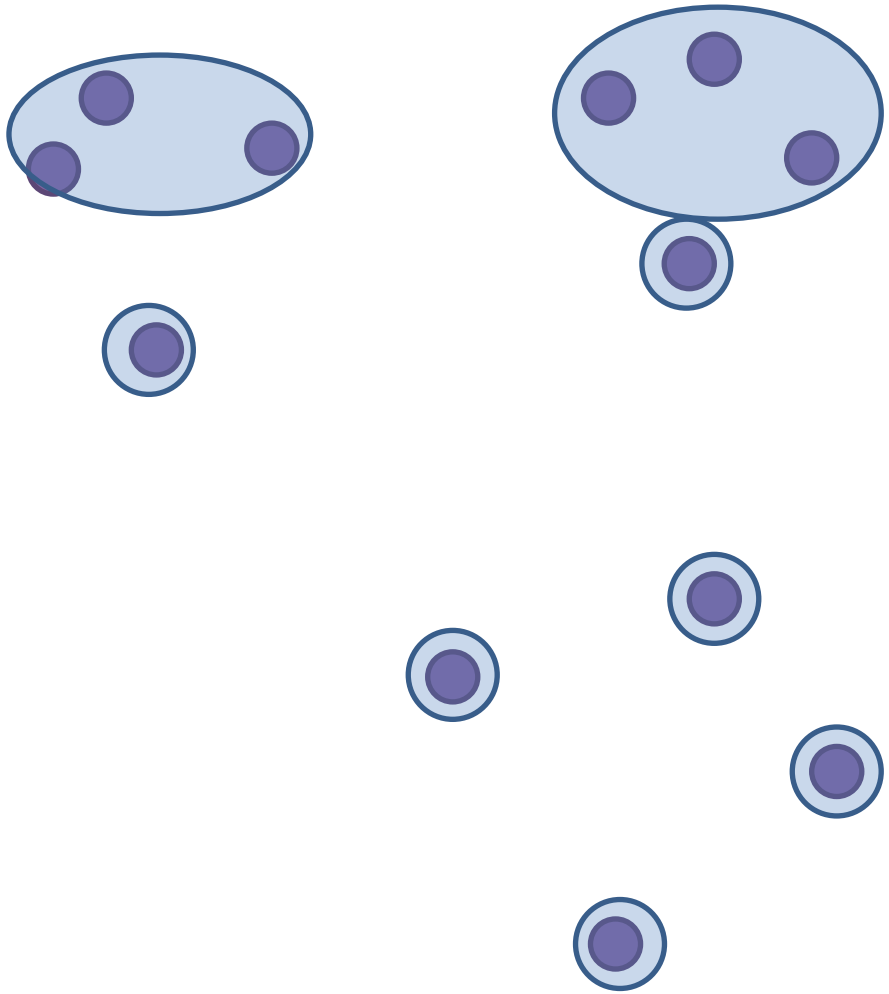
Расстояния между кластерами



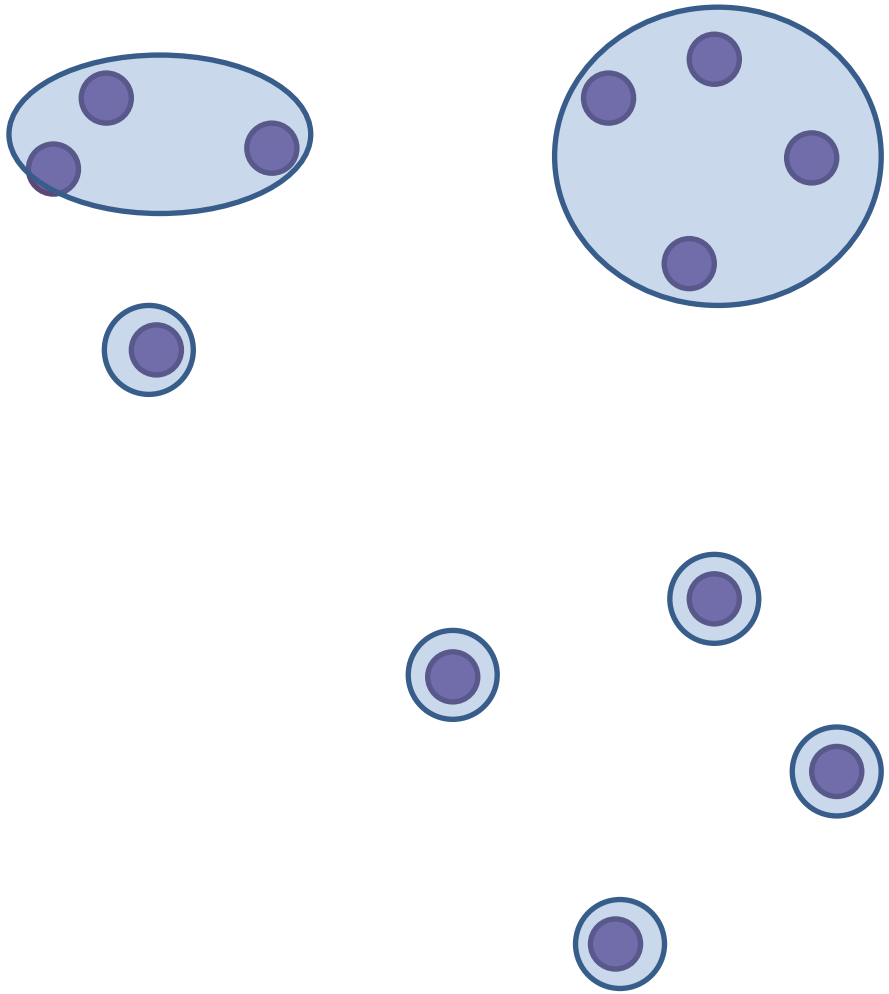
Агломеративная кластеризация



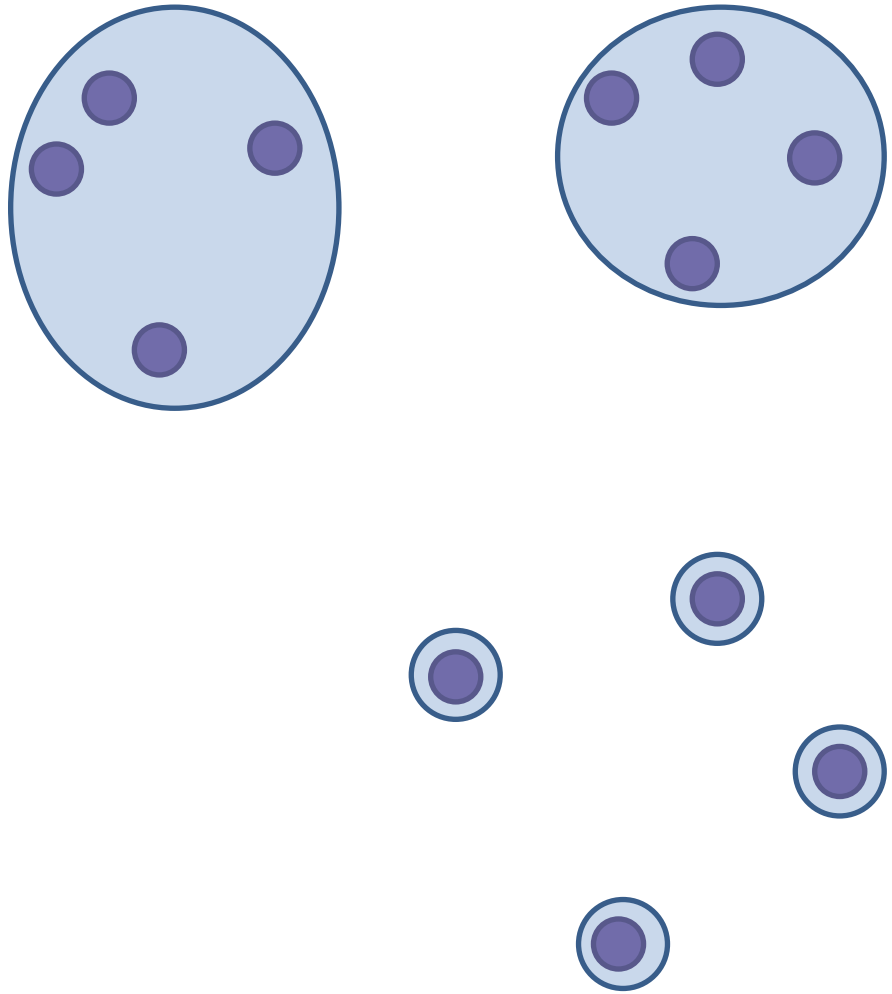
Агломеративная кластеризация



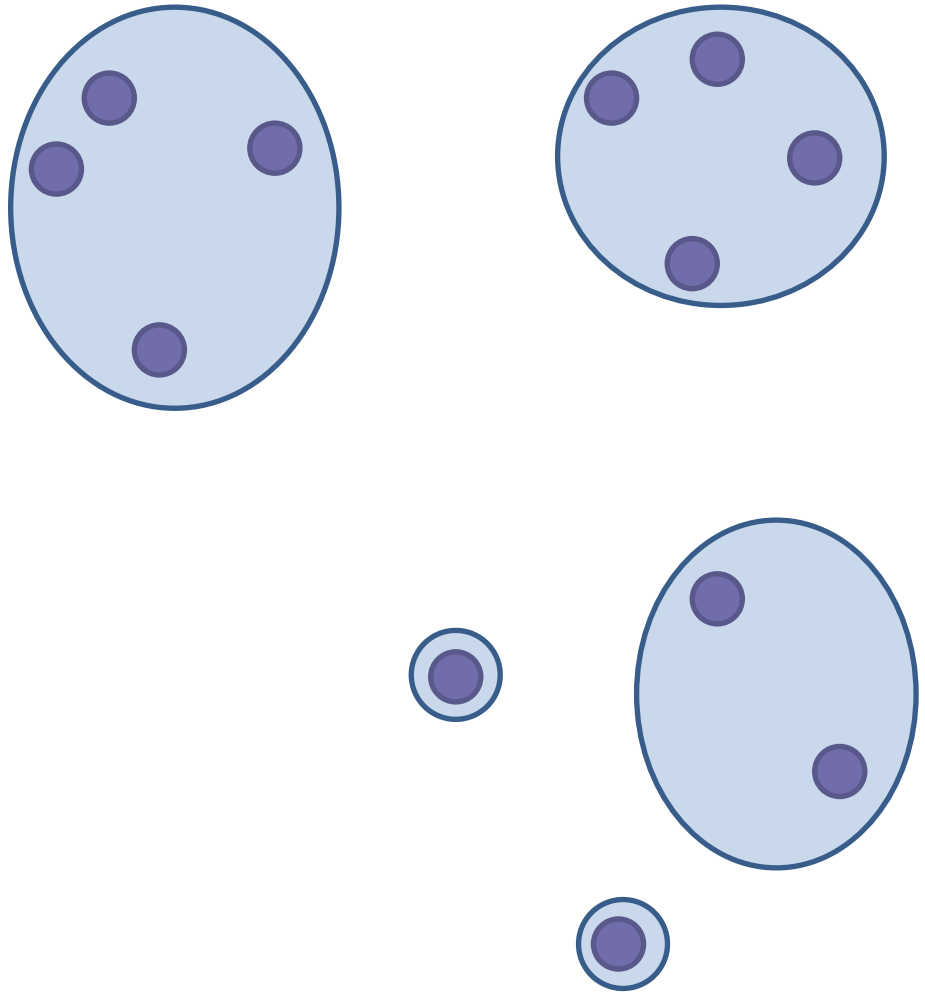
Агломеративная кластеризация



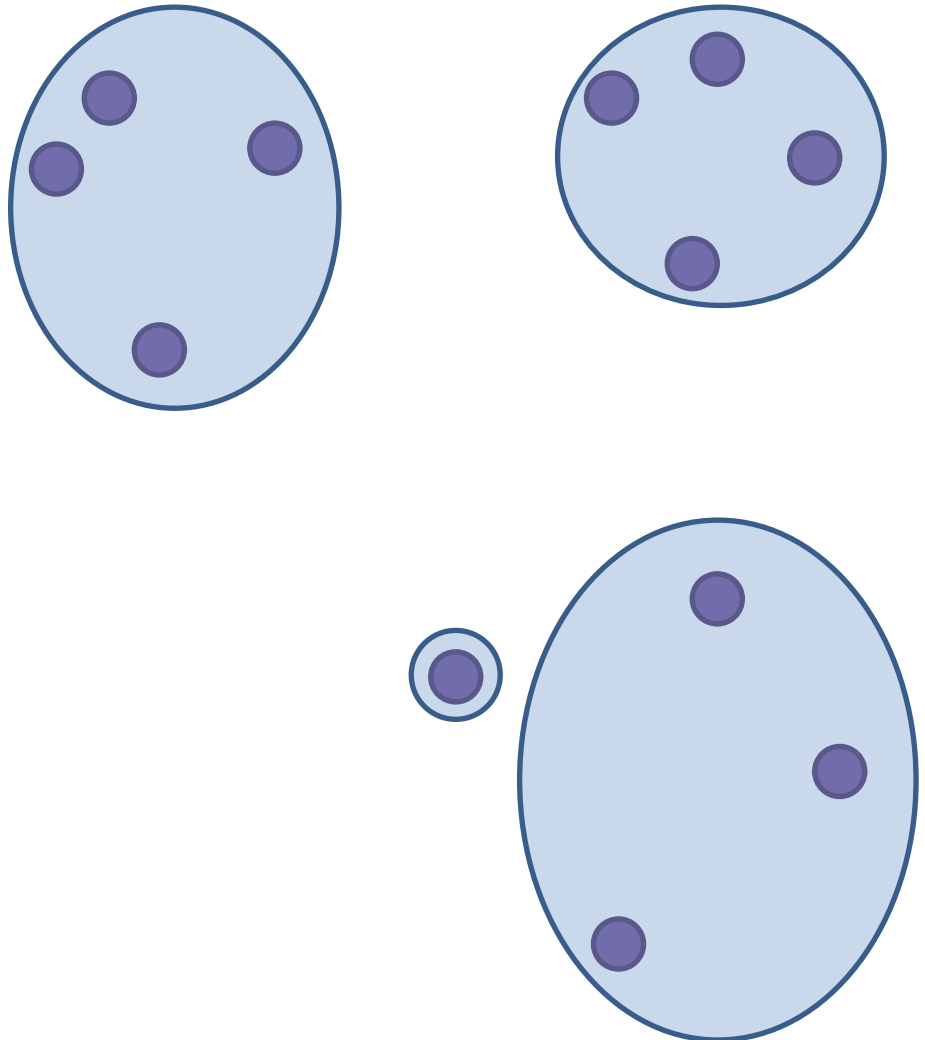
Агломеративная кластеризация



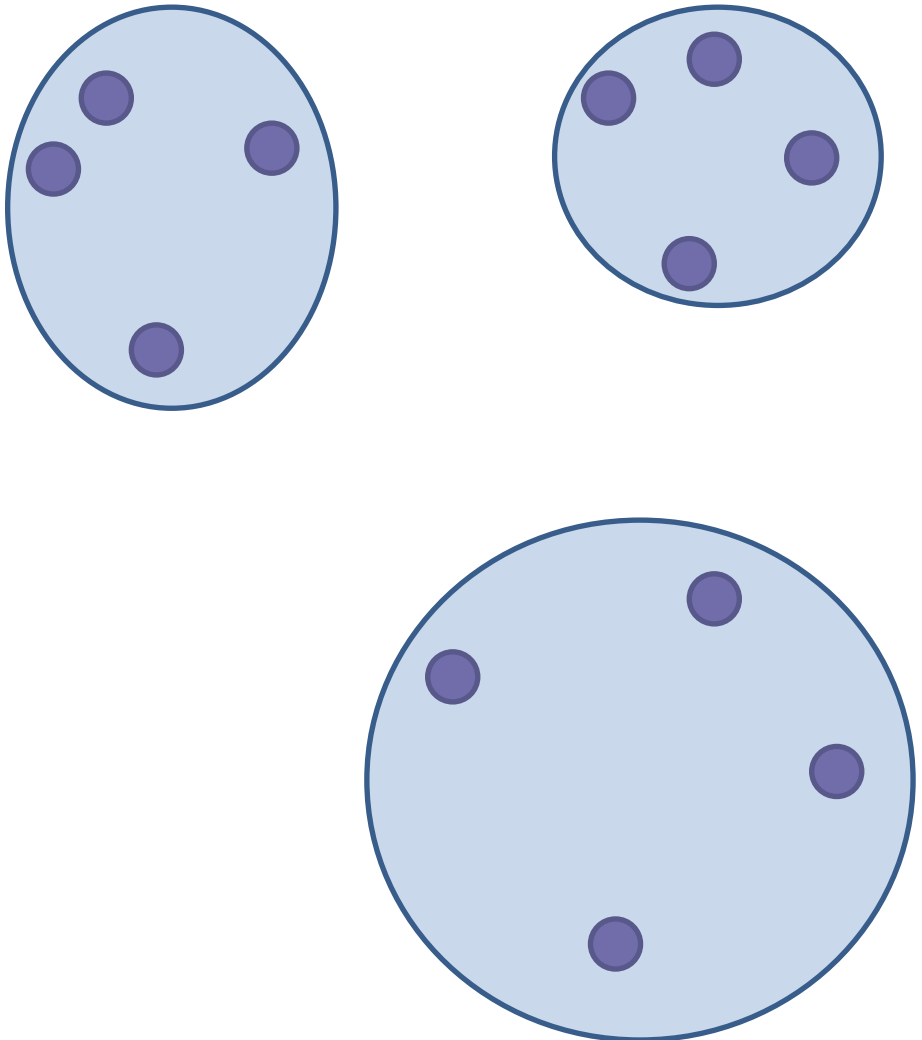
Агломеративная кластеризация



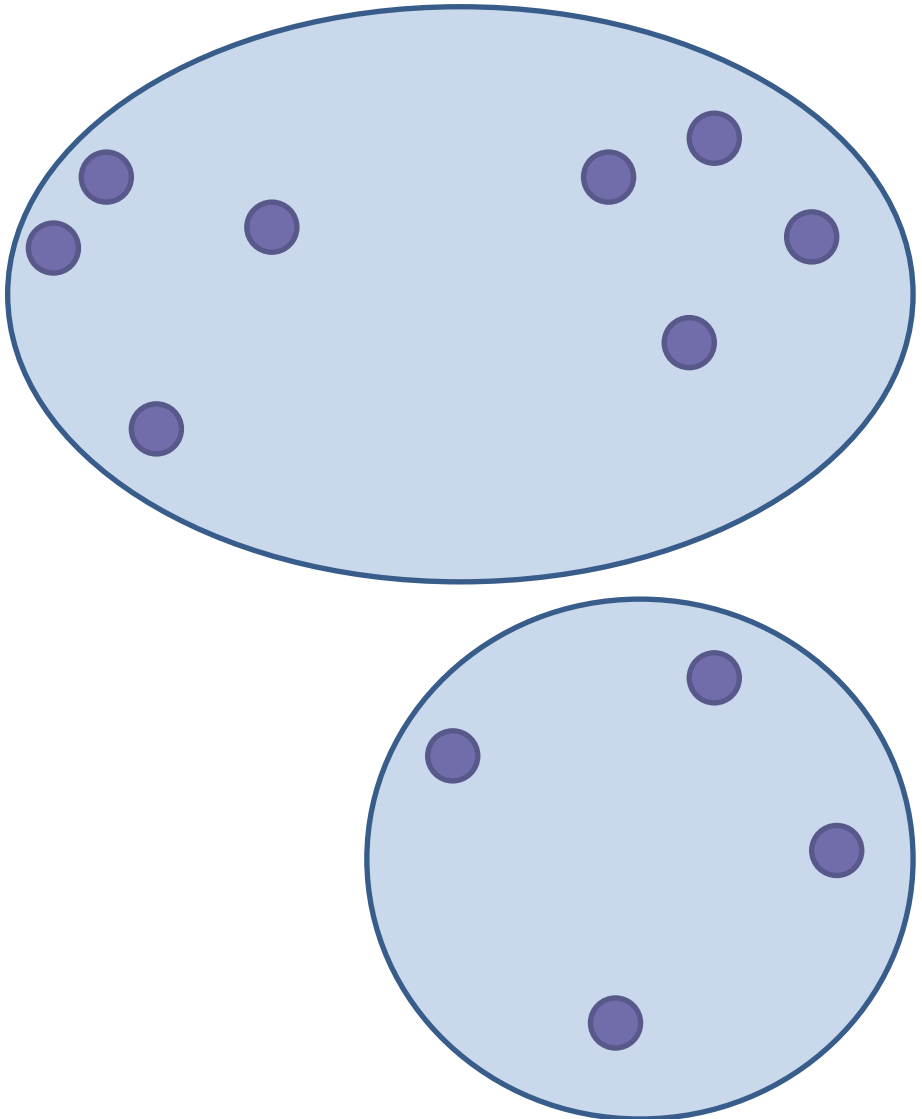
Агломеративная кластеризация



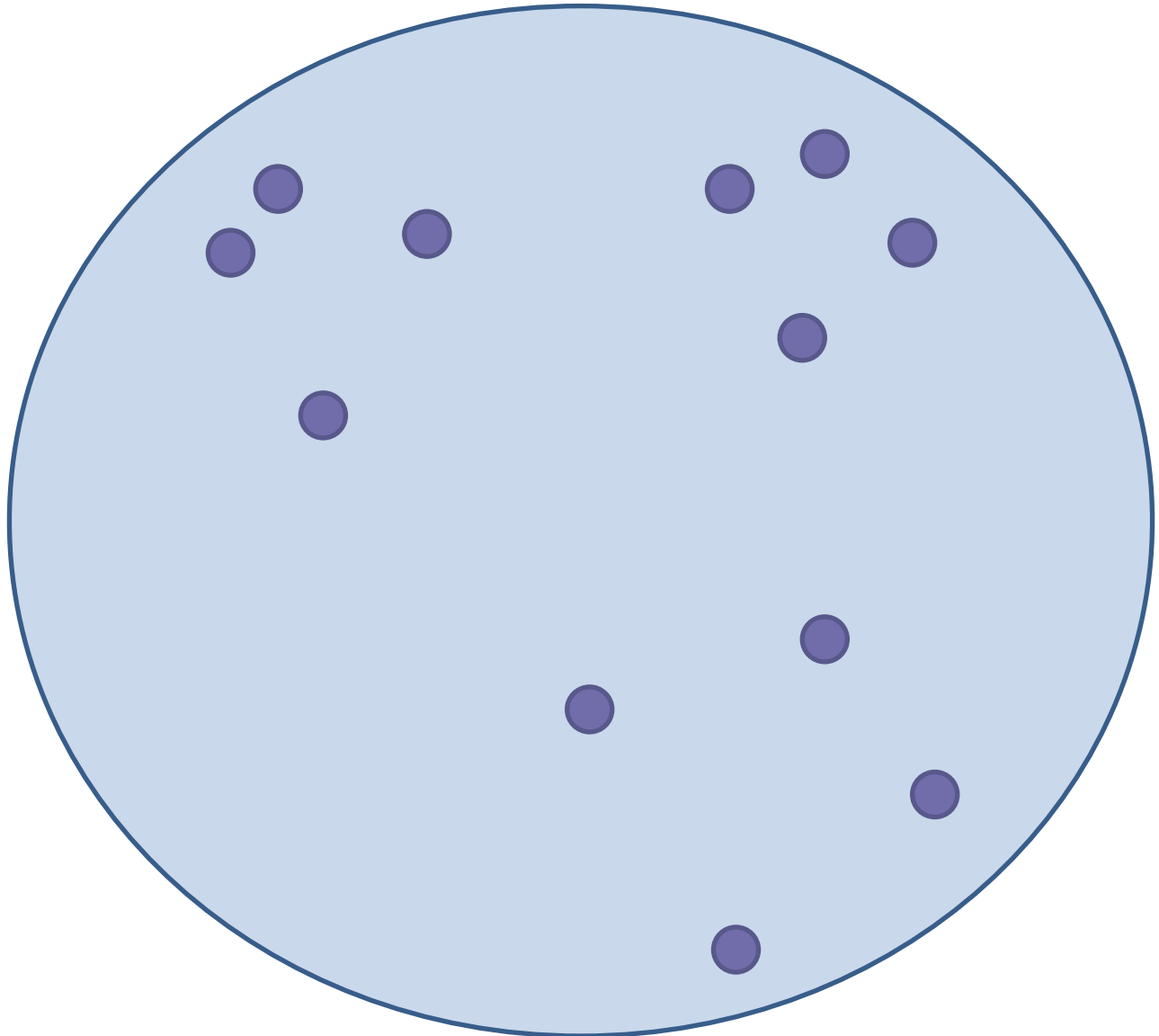
Агломеративная кластеризация



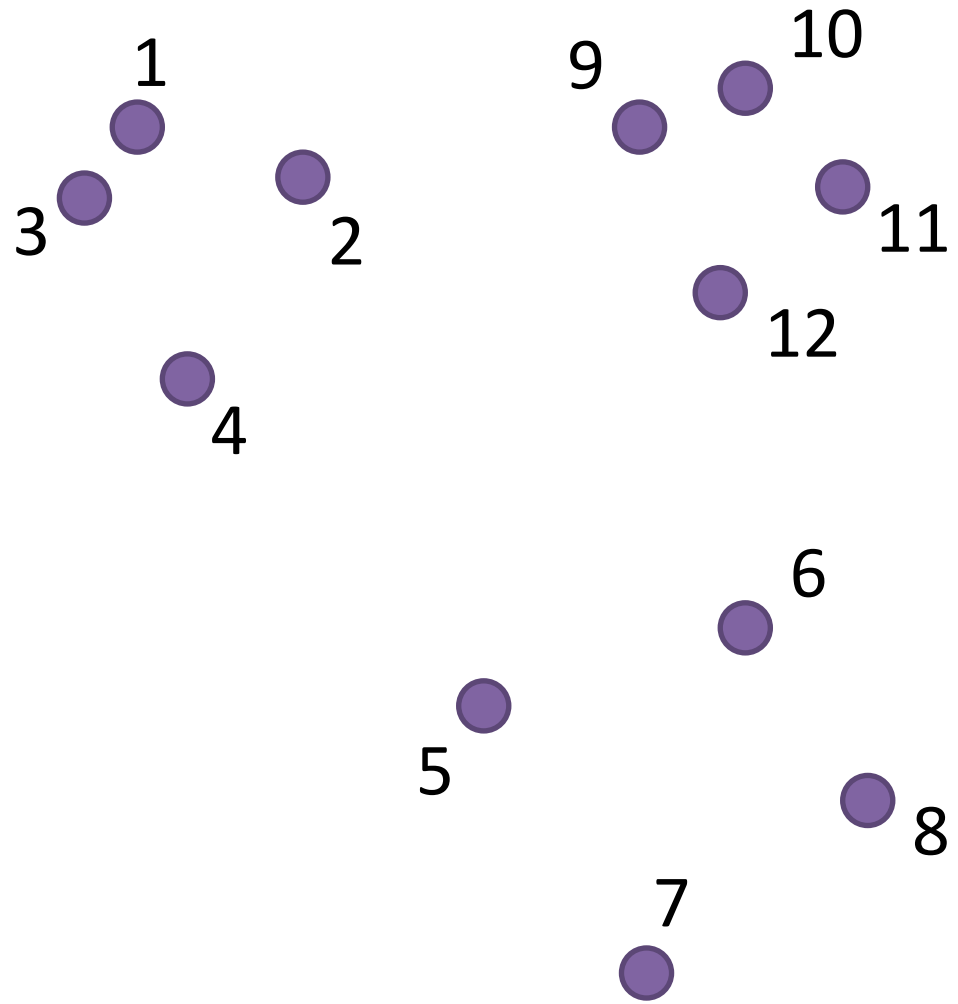
Агломеративная кластеризация



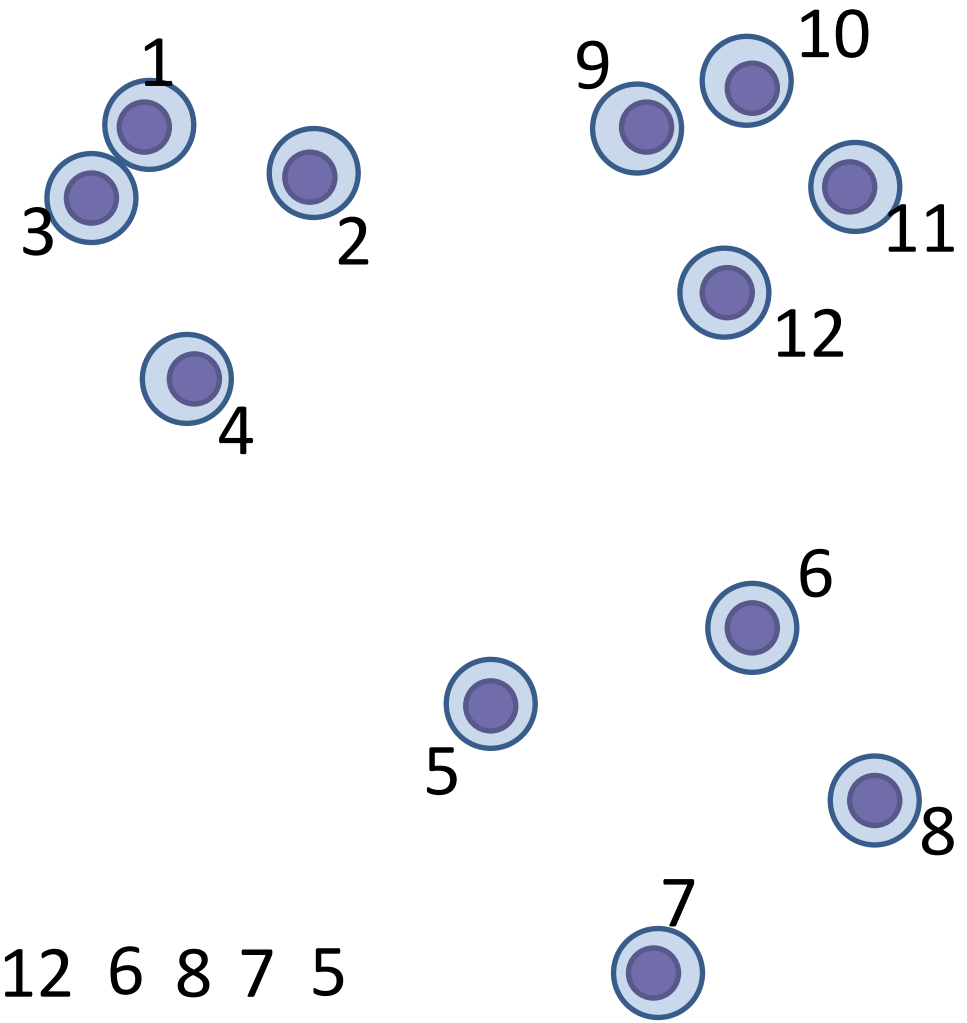
Агломеративная кластеризация



Дендрограмма

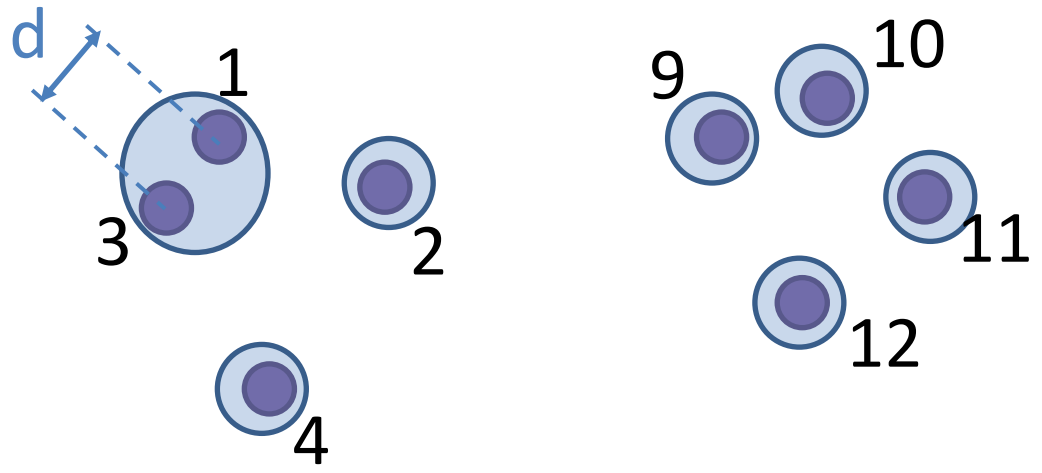


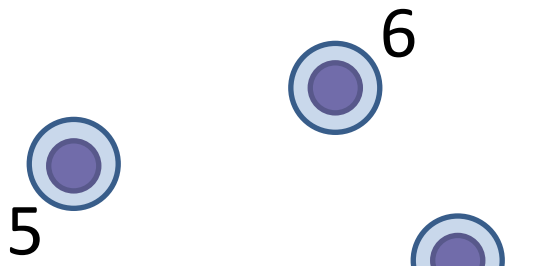
Дендрограмма

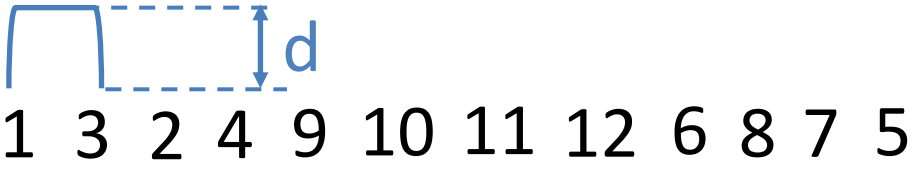


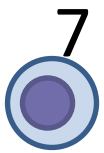
1 3 2 4 9 10 11 12 6 8 7 5

Дендрограмма

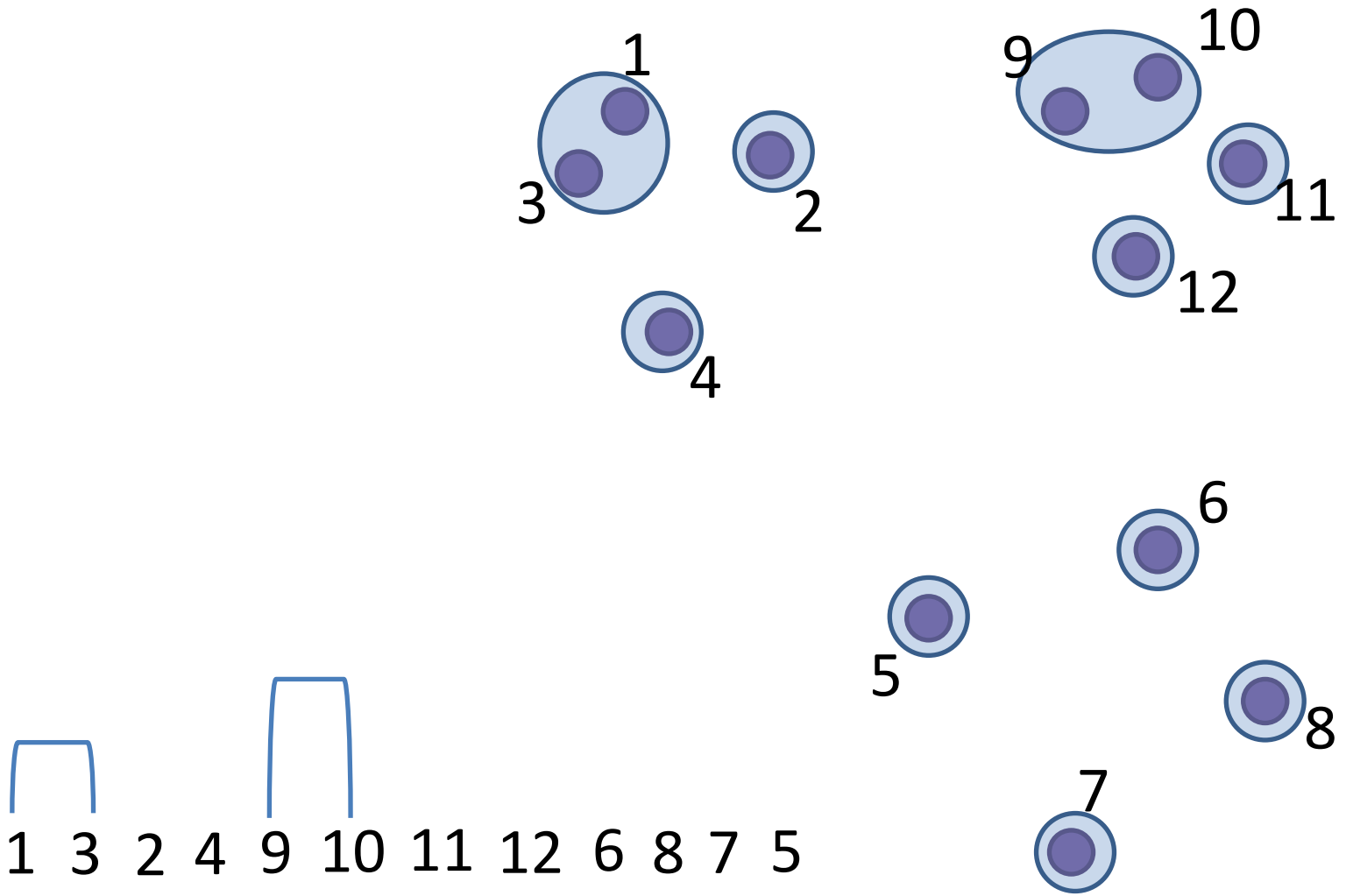




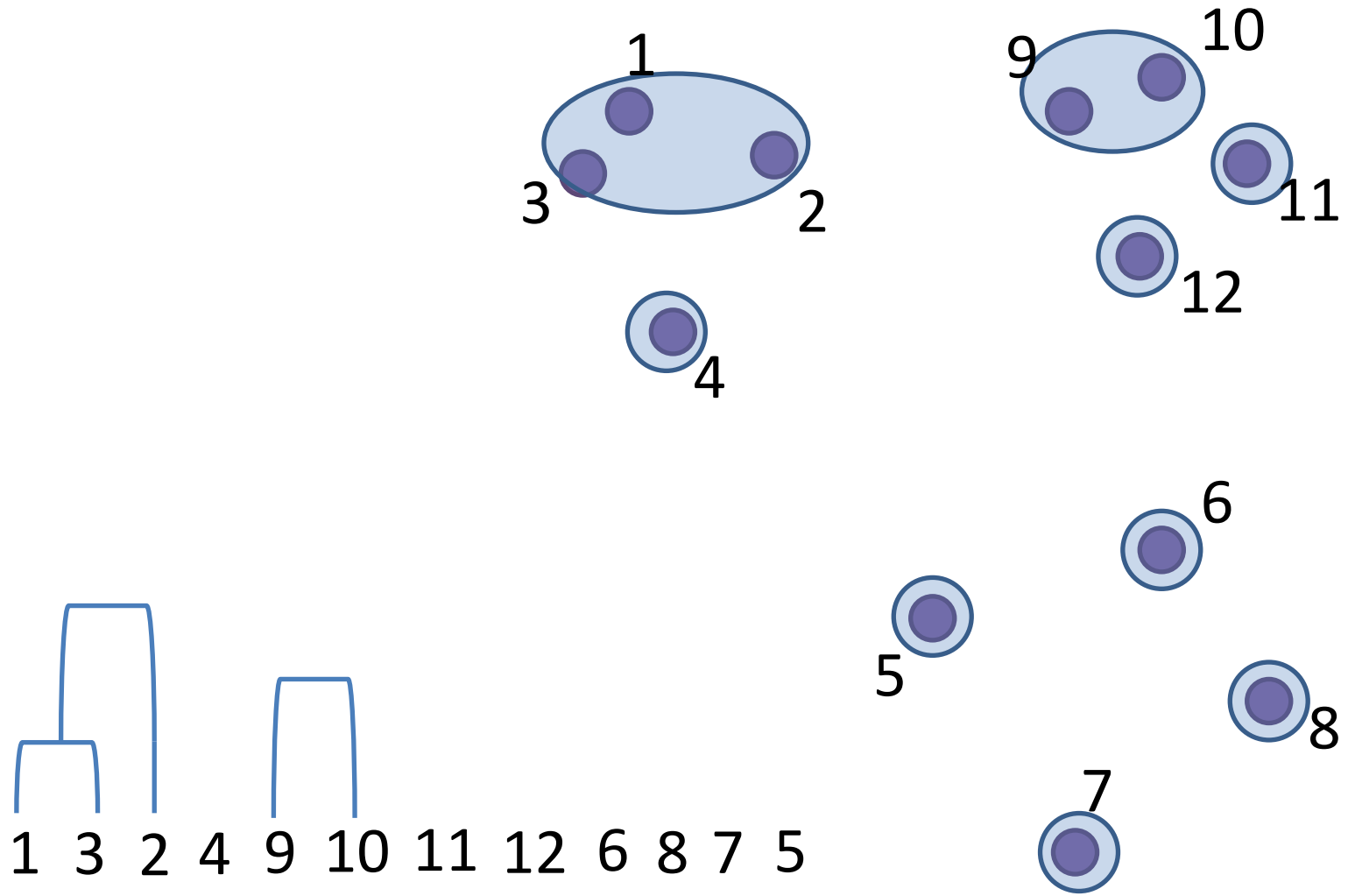




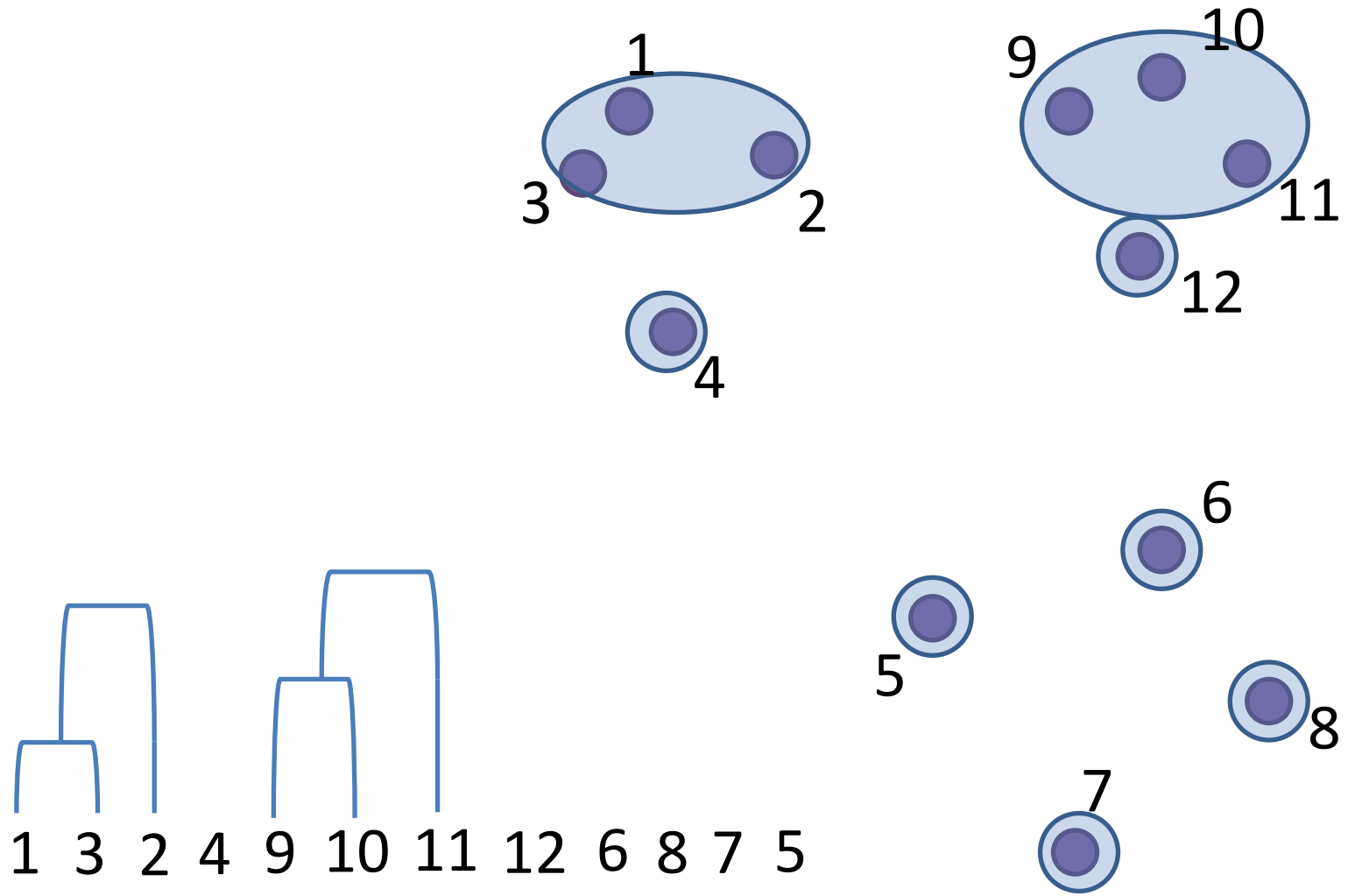
Дендрограмма



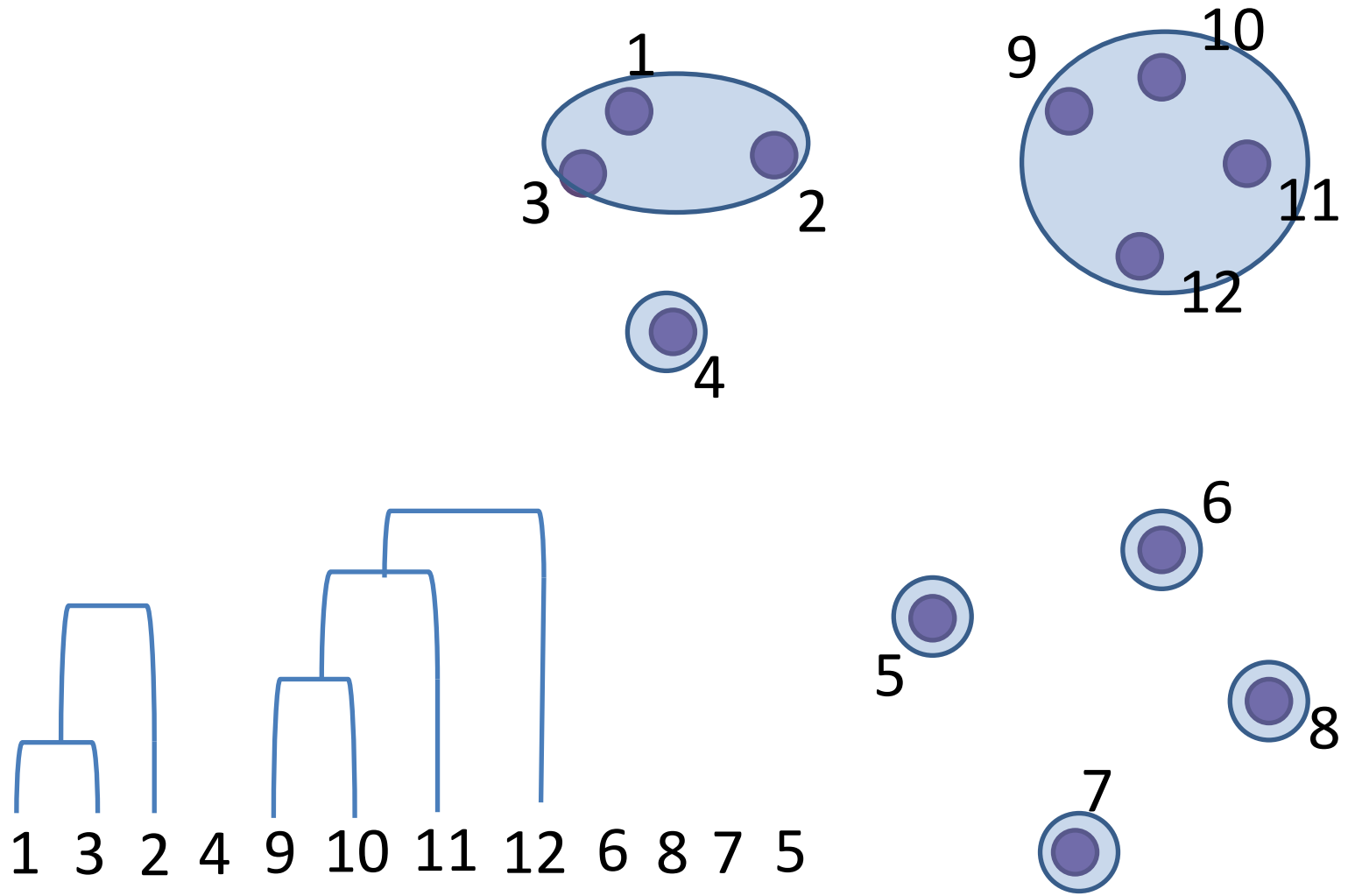
Дендрограмма



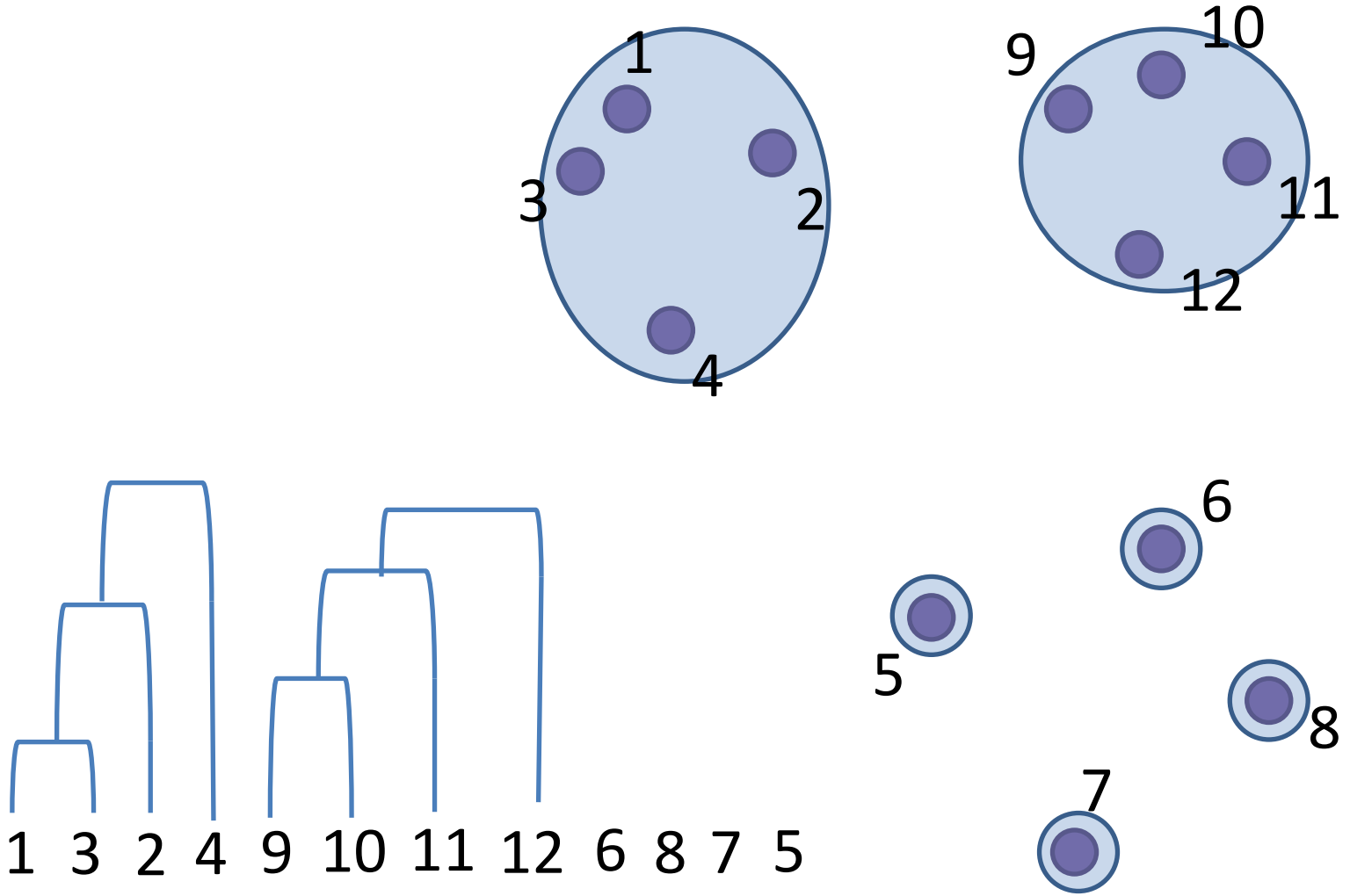
Дендрограмма



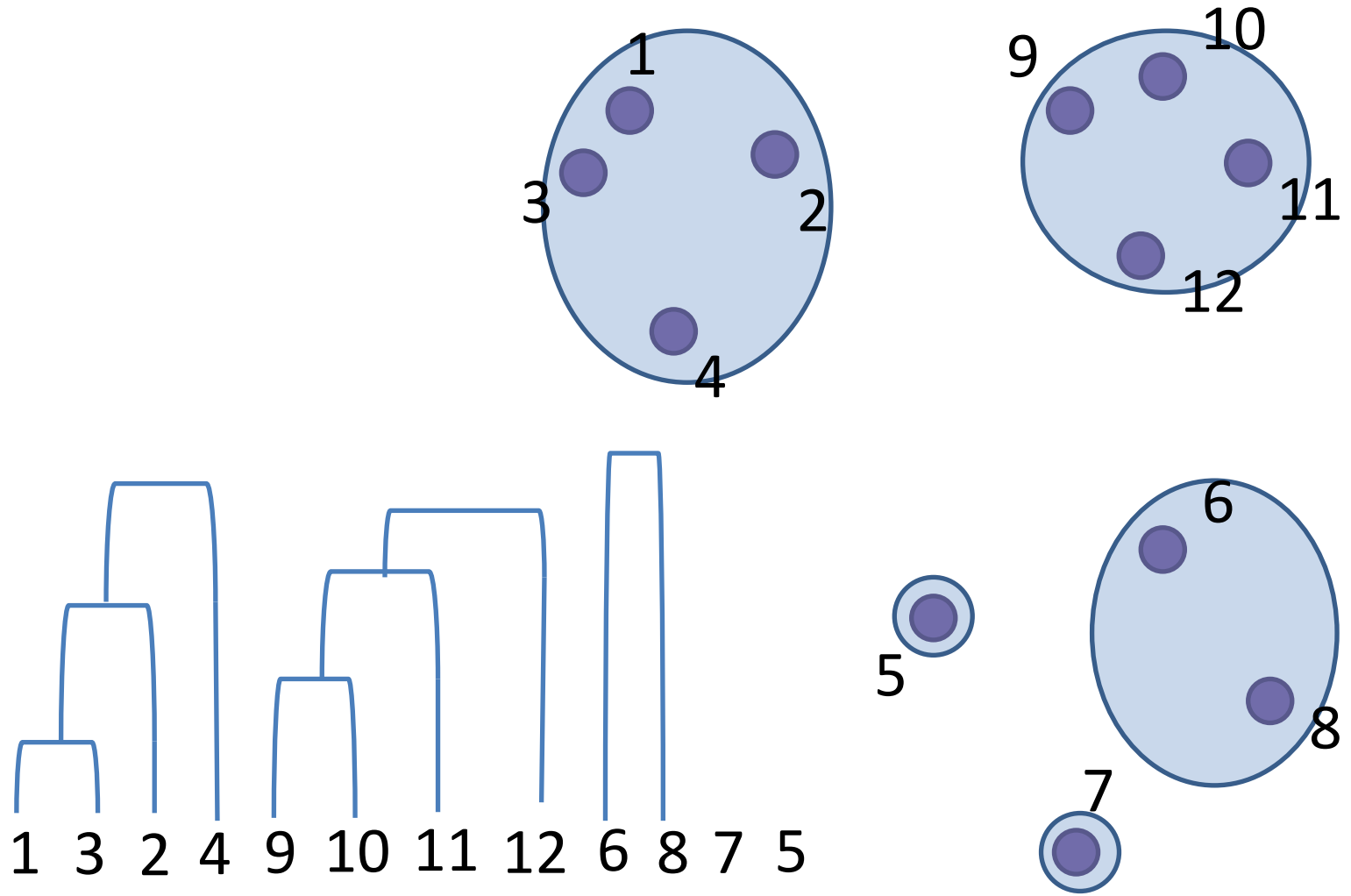
Дендрограмма



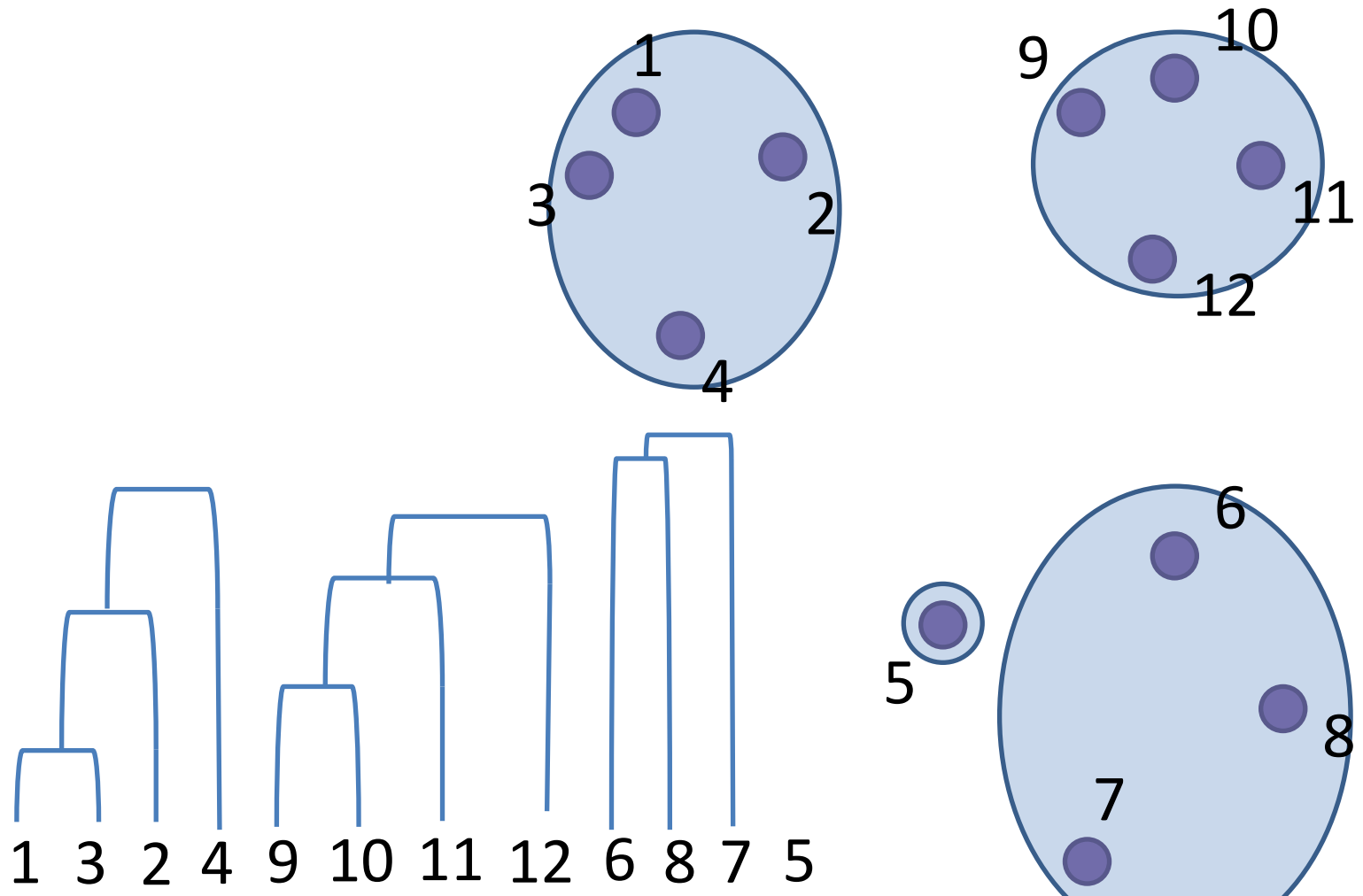
Дендрограмма



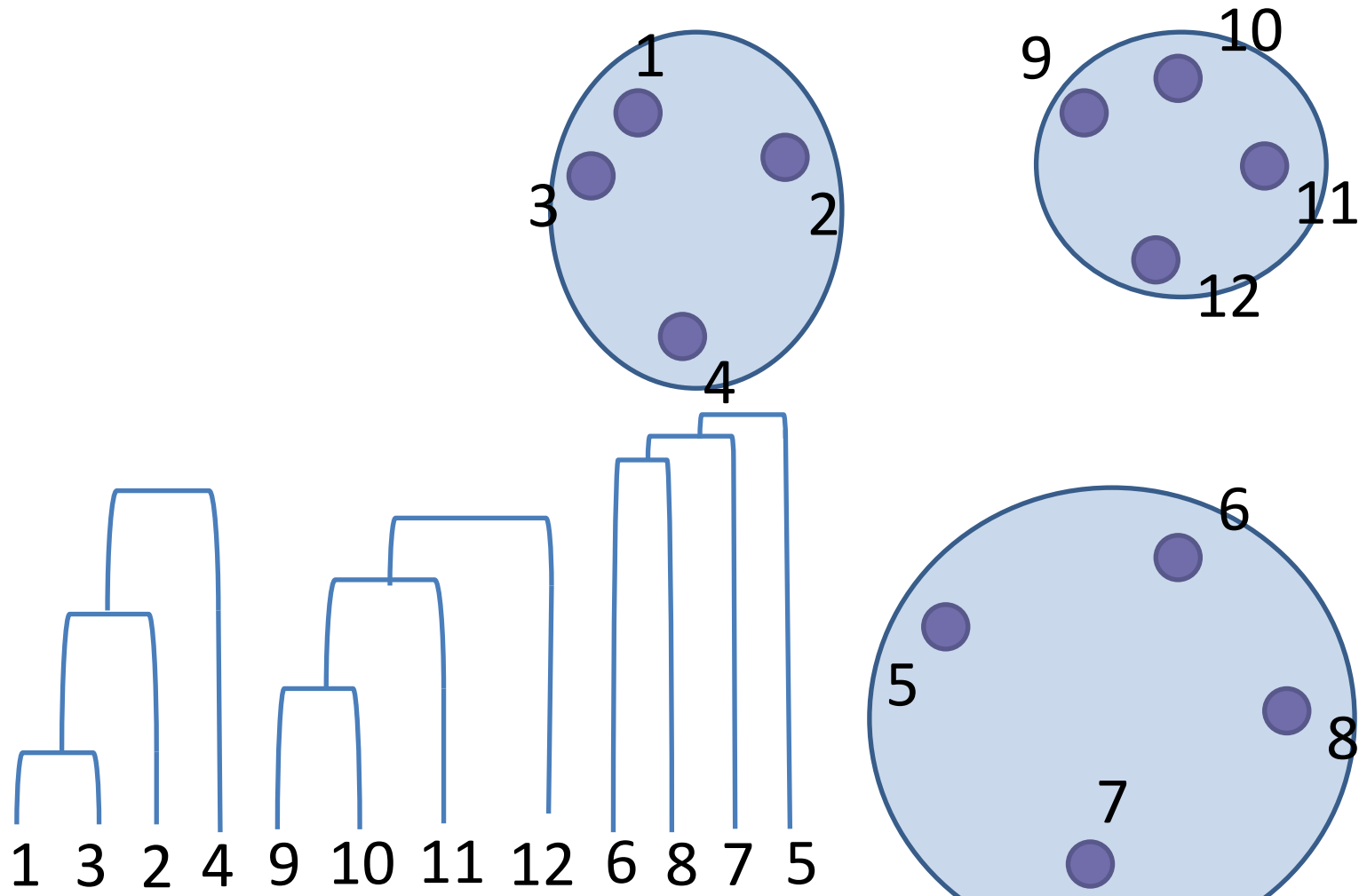
Дендрограмма



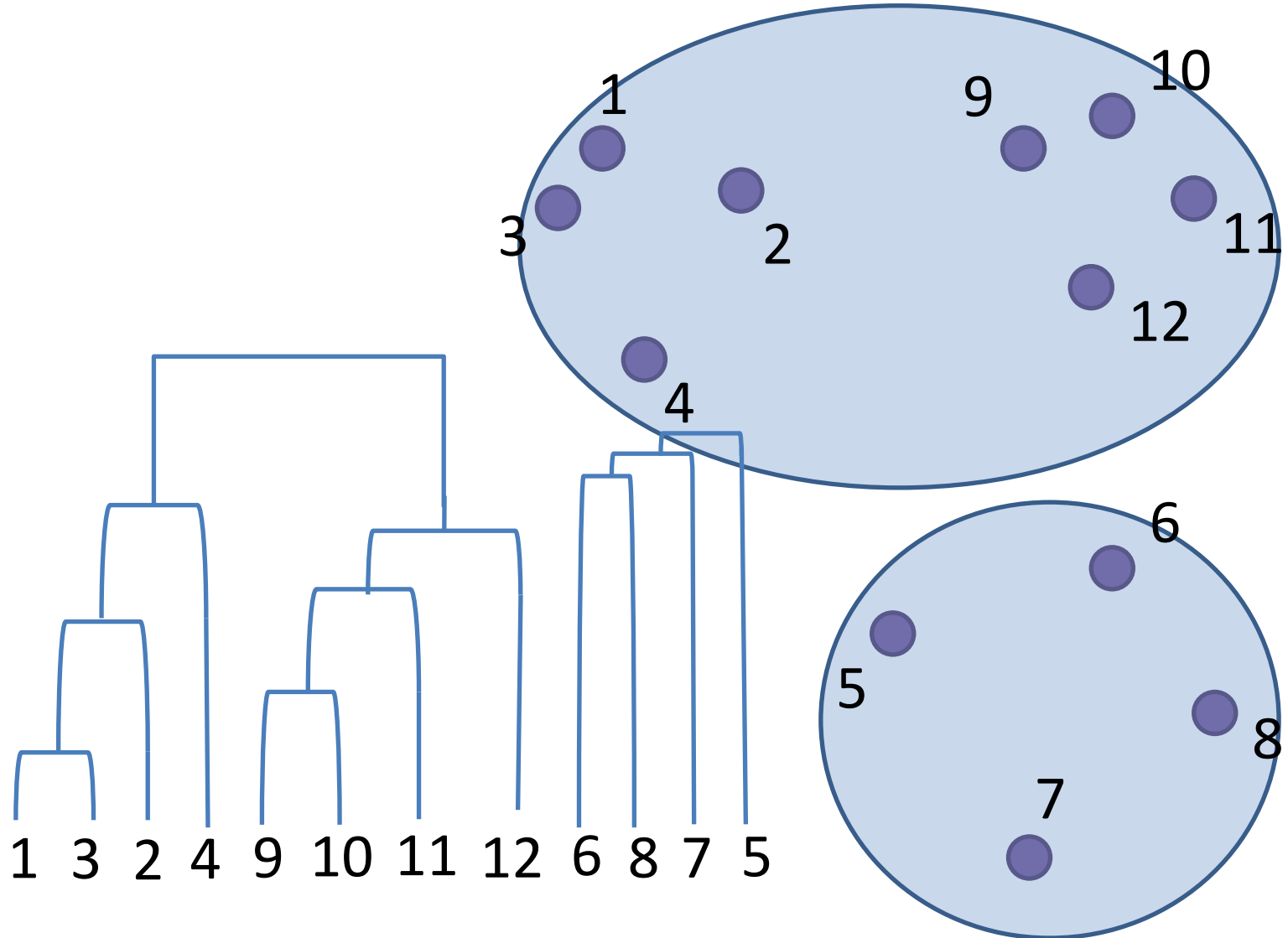
Дендрограмма



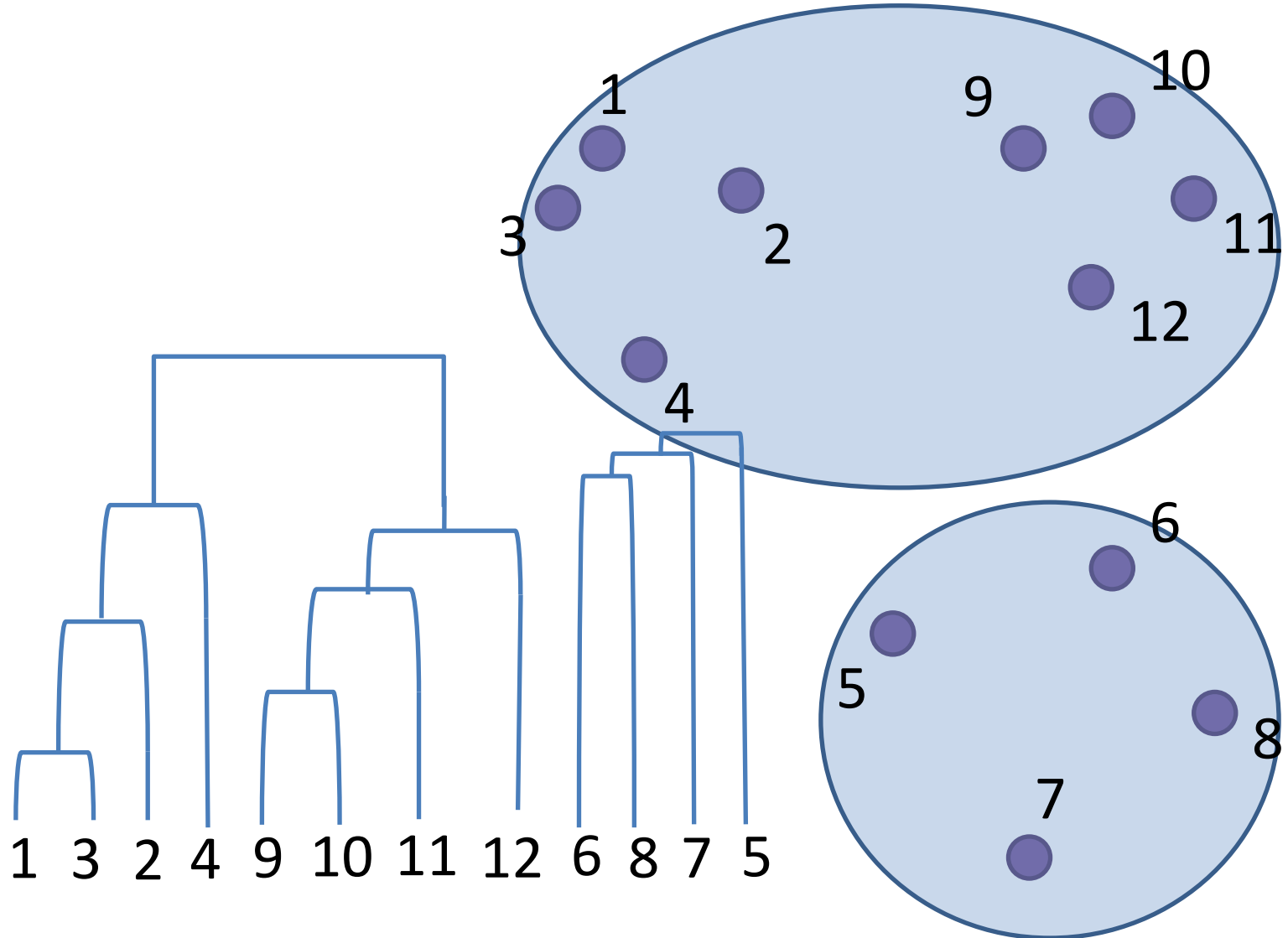
Дендрограмма



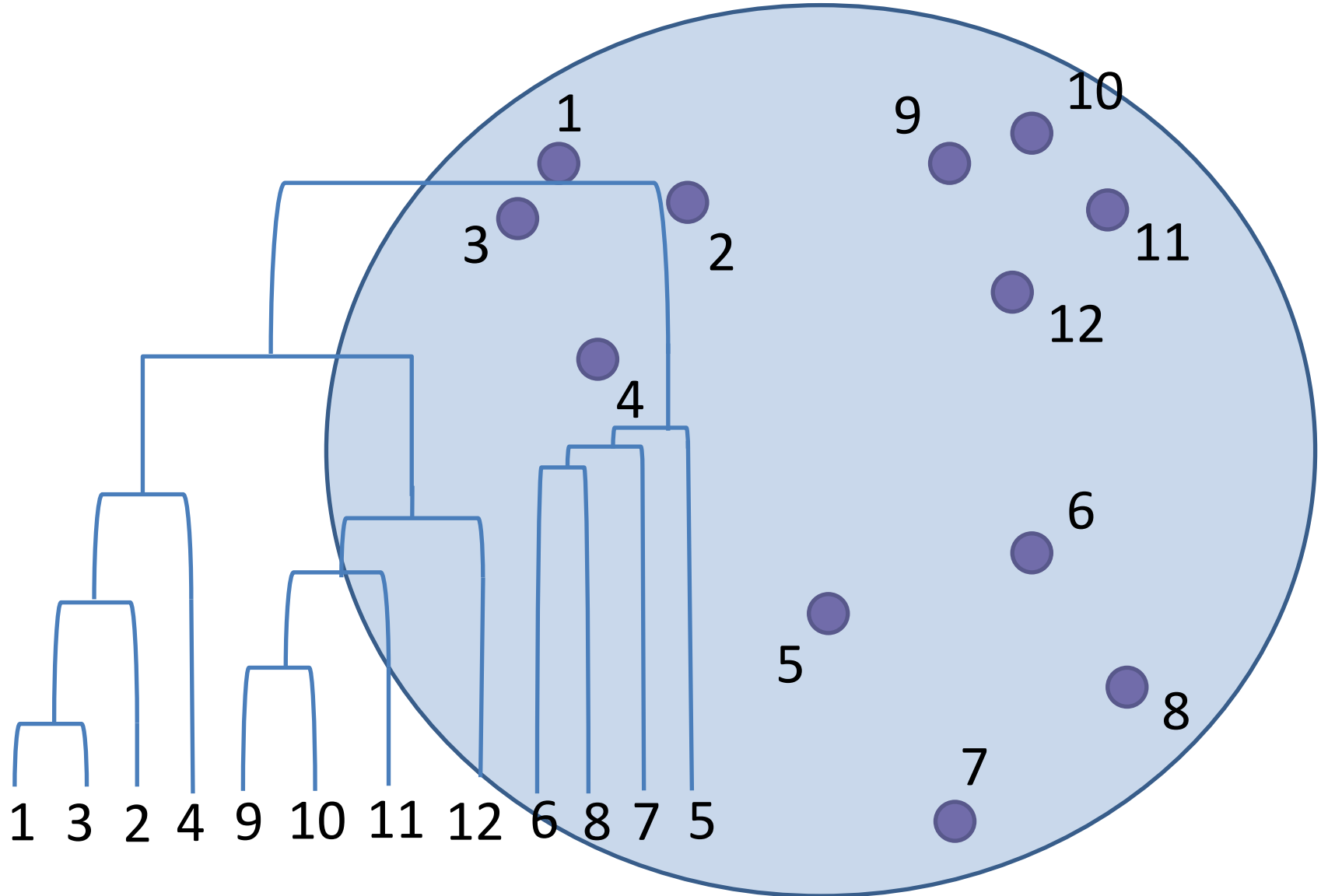
Дендрограмма



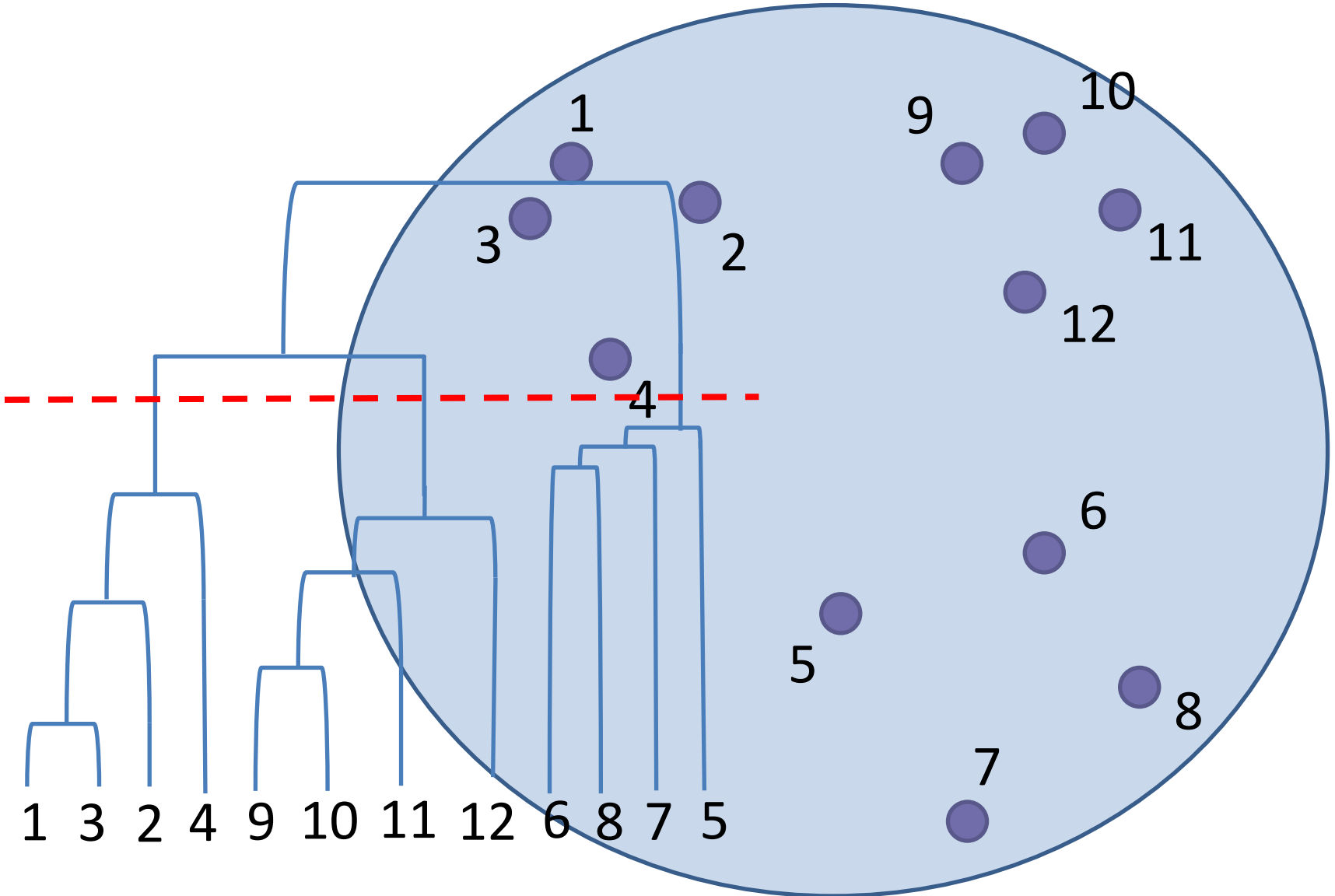
Дендрограмма



Дендрограмма



Дендрограмма



Аггломеративная кластеризация

1. Инициализация – каждая точка = кластер
2. Самые близкие (относительно какой-то метрики) кластеры объединяются
3. Повторяем до того момента, когда все точки будут в одном кластере
4. Останавливаемся, когда достигаем фиксированного числа кластеров, либо когда расстояние между кластерами больше заданного порога

Quality Metrics

Quality metrics

There are two kinds of quality metrics for clustering:

- Supervised
  - Based on ground truth of object labels
  - Invariant to cluster naming
- Unsupervised
  - Based on intuition about “good” clusters:
    - Objects from the same cluster are similar / close to each other
    - Objects from different clusters are dissimilar / distant from each other

Rand Index

Rand Index (RI) is supervised quality metric defined as:

$$\mathrm{RI} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

TP – number of pairs in the same cluster in predictions and the ground truth,

TN – number of pairs from different clusters in predictions and the ground truth,

FP – number of pairs in the same cluster in predictions, but from different clusters in the ground truth,

FN – number of pairs in the same cluster in the ground truth, but from the different clusters in predictions.

Adjusted Rand Index

Adjusted Rand Index (ARI) is modification of RI:

$$\mathrm{ARI} = \frac{\mathrm{RI} - \mathrm{RI}_{\mathrm{Expected}}}{\mathrm{RI}_{\mathrm{Max}} - \mathrm{RI}_{\mathrm{Expected}}}$$

ARI has a value close to 0.0 for random labeling independently of the number of clusters and samples and exactly 1.0 when the clustering is ideal

Metrics for classification

- $\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$
- $\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$
- $\text{F1} - \text{score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$
- Fowlkes-Mallows Index (FMI) $= \frac{\text{TP}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})}}$
- others

Silhouette

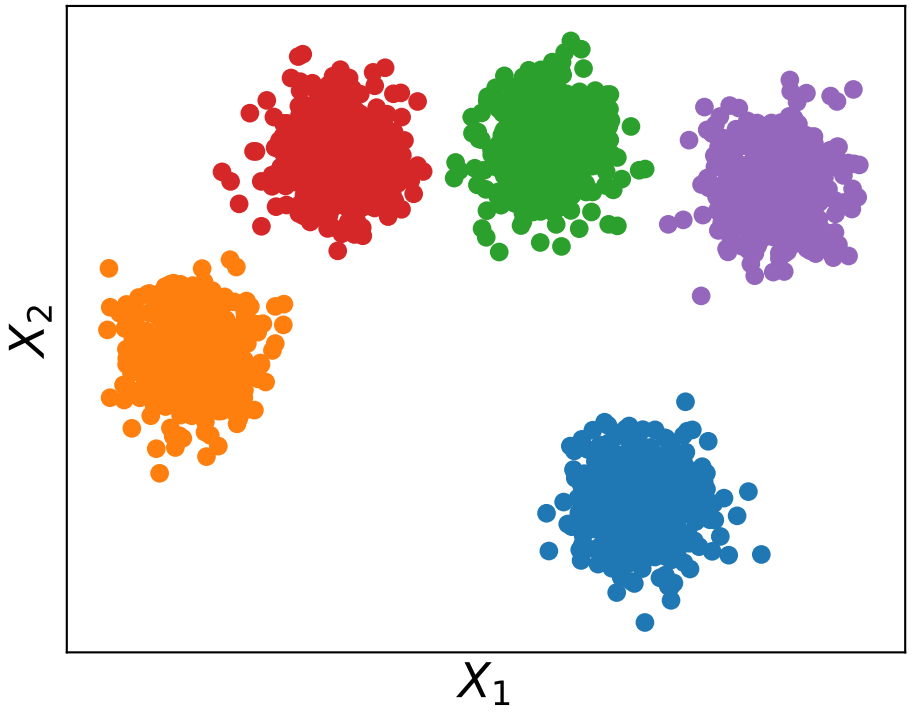
Silhouette is unsupervised quality metric defined as:

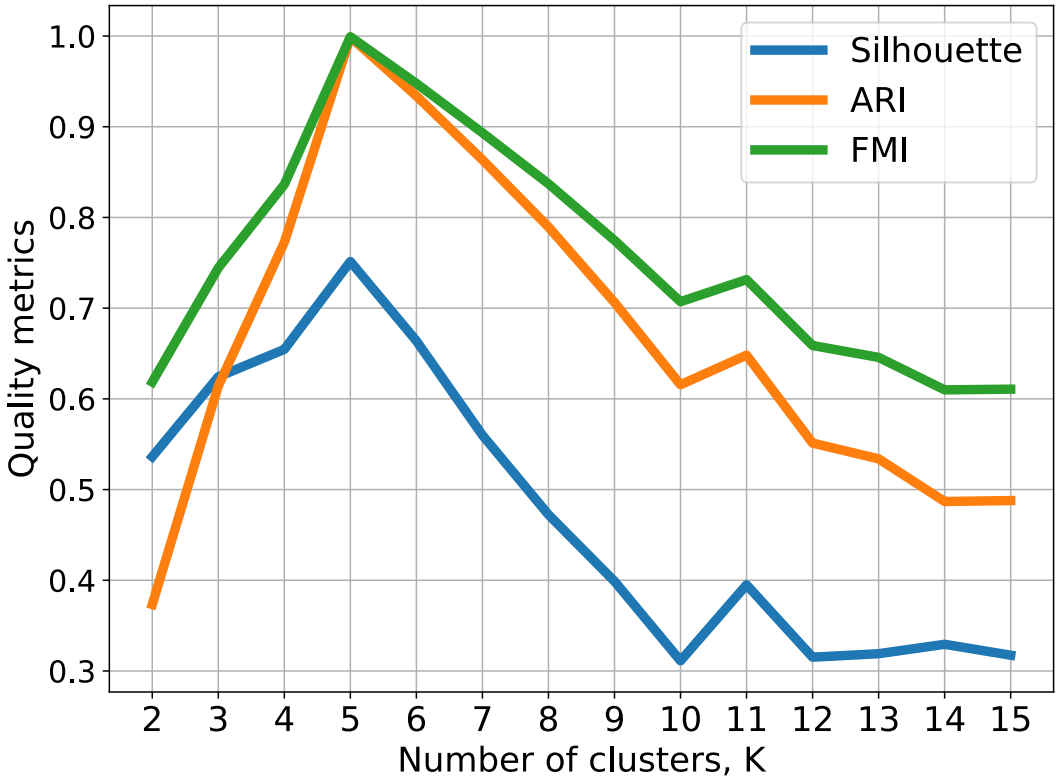
$$\text{Silhouette} = \frac{1}{N}\sum_{i=1}^{N}\frac{d_i - s_i}{\max\{d_i, s_i\}}$$

s_i - mean distance between the i-th object and all objects in the same cluster,

d_i - mean distance between the i-th object and all objects in the nearest cluster.

Example


X_2
X_1


Silhouette
ARI
FMI
Quality metrics
Number of clusters, K
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
2 3 4 5 6 7 8 9 10 11 12 13 14 15

Резюме

- Кластеризация — задача без строгой постановки и без строгих критериев качества
- Много разновидностей в подходах
- Методы: K-Means, DBSCAN, иерархическая кластеризация и т.д.