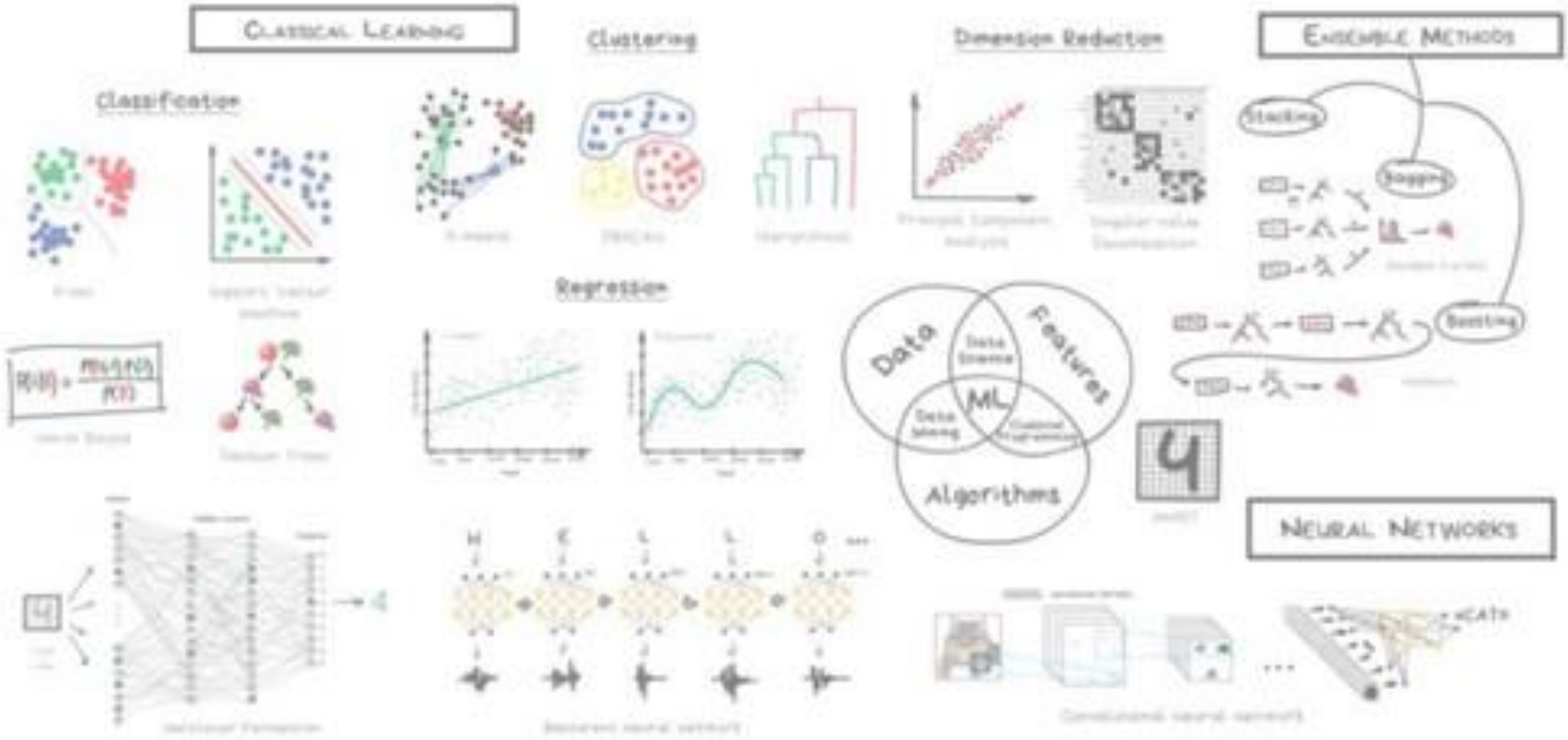
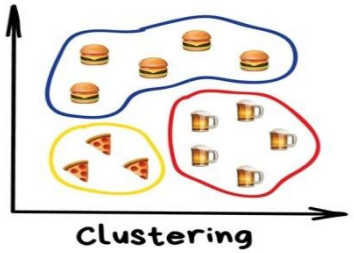
МАШИНЕ НАВЧАННЯ

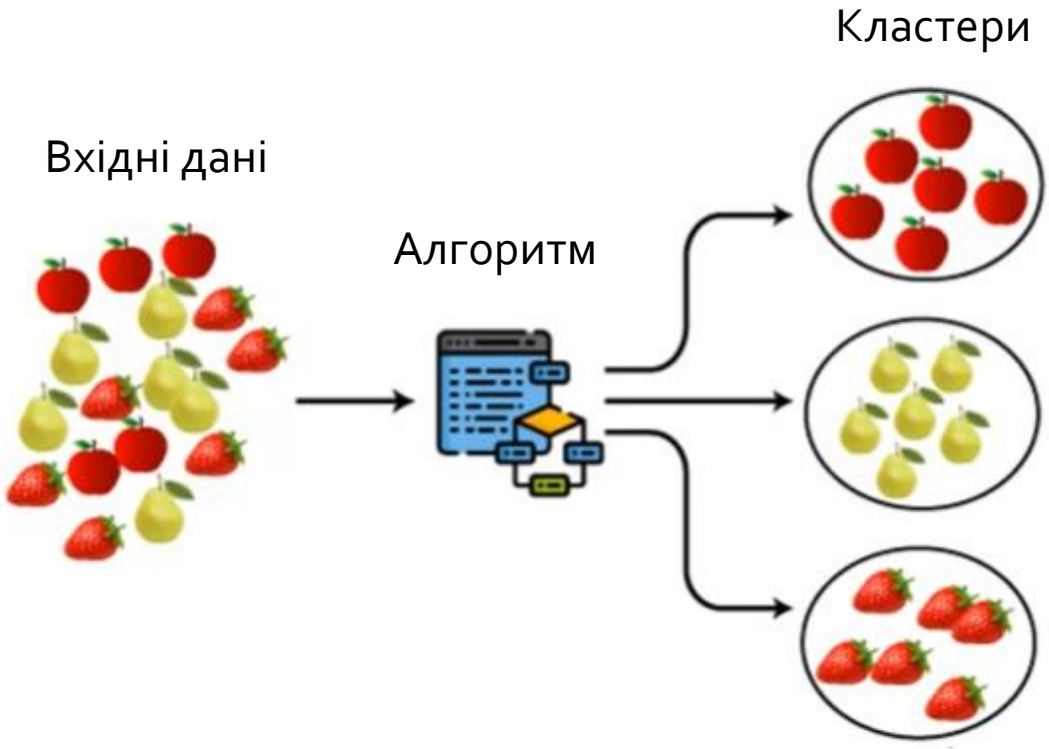
Навчання без вчителя. Методи кластеризації



Лекція №10

Кластеризація



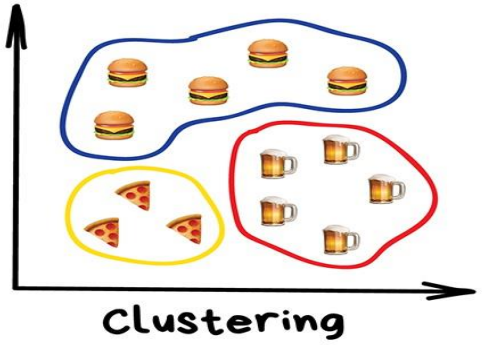


Кластеризація

- Кластеризація - це класифікація, але без заздалегідь відомих класів. Вона сама шукає схожі об'єкти та об'єднує їх у кластери. Кількість кластерів можна задати заздалегідь або довірити машині. Схожість об'єктів машина визначає за тими ознаками, які ми їй розмітили - у кого багато схожих характеристик, тих поєднують в один кластер.
- Відмінний приклад кластеризації – маркери на картах в інтернеті. Коли ви шукаєте всі есторани азіатської кухні у великому місті, машині доводиться групувати їх у кружечки з циферкою, інакше браузер зависне в потугах намалювати мільйон маркерів.
- Більш складні приклади кластеризації можна згадати у програмах iPhoto або Google Photos, які знаходять обличчя людей на фотографіях та групують їх у альбоми. Програма не знає як звуть ваших друзів, але може відрізнити їх за характерними рисами обличчя.

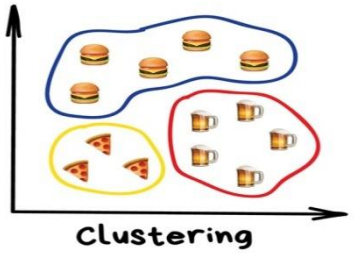
Сьогодні використовують для:

- Сегментація ринку (типів покупців)
- Об'єднання близьких точок на карті
- Стиснення зображень
- Аналіз та розмітки нових даних
- Детектори аномальної поведінки
- Популярні алгоритми: Метод K-середніх, Mean-Shift, DBSCAN



Кластеризація

Постановка задачі кластеризації



Дано:

X – простір об'єктів

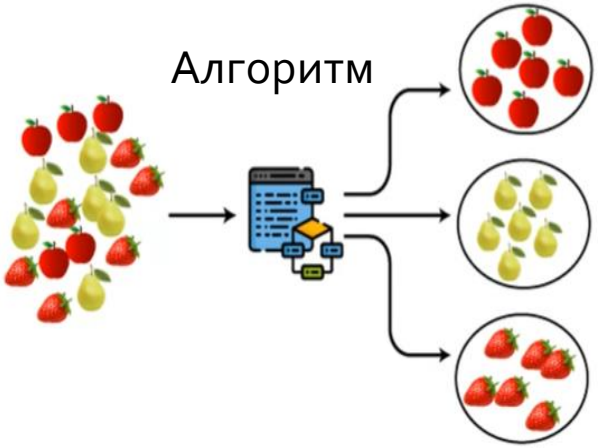
$X^\ell = \{x_1, \dots x_\ell\}$ – навчальна вибірка

$\rho: X \to [0, \infty)$ – функція відстані між об'єктами

Знайти:

Y – множина кластерів

$a: X \to Y$ – алгоритм кластеризації

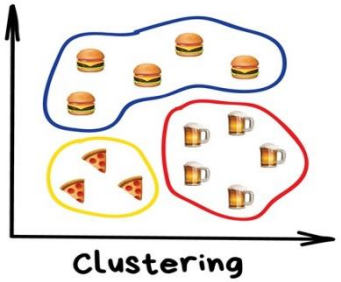


Властивості кластерів:

- Кожен кластер складається з близьких за ознаками об'єктів
- Об'єкти різних кластерів суттєво різні

Кластеризація

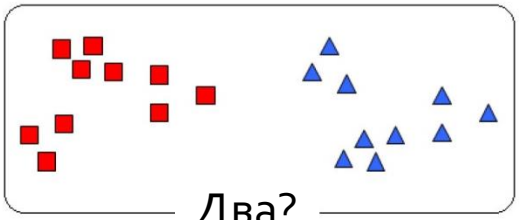
Некоректність задачі кластеризації



Розв'язок задачі кластеризації принципово неоднозначний

- Точної постановки задачі кластеризації як правило немає
- Існує багато критеріїв якості кластеризації: визначення оптимальної кількості кластерів
- Існує багато евристичних методів кластеризації
- Зазвичай кількість кластерів заздалегідь не відома
- Результат кластеризації сильно залежить від метрики ρ для розрахунку відстані між об'єктами

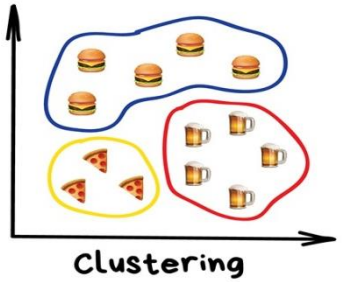
Приклад: скільки тут кластерів?


Два?


Чотири?


Шість?

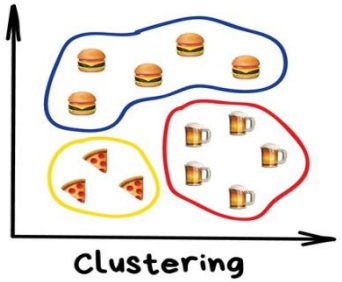
Кластеризація
Задачі кластеризації



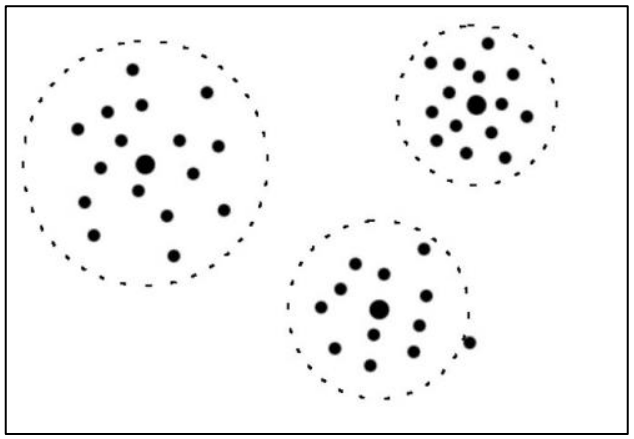
- Спростити подальшу обробку даних: розбити множину X^ℓ на групи схожих об'єктів з метою подальшого аналізу кожної групи окремо (задачі класифікації, регресії, прогнозування)
- Скоротити об'єм даних що зберігається: залишити по одному з типових представників кожного кластеру (центр мас) – задачі стиснення даних
- Виділити нетипові об'єкти (викиди), які не підходять до жодного з кластерів (задачі однокласової класифікації)
- Побудувати ієрархію множини об'єктів (класифікація рослин та тварин)

Кластеризація

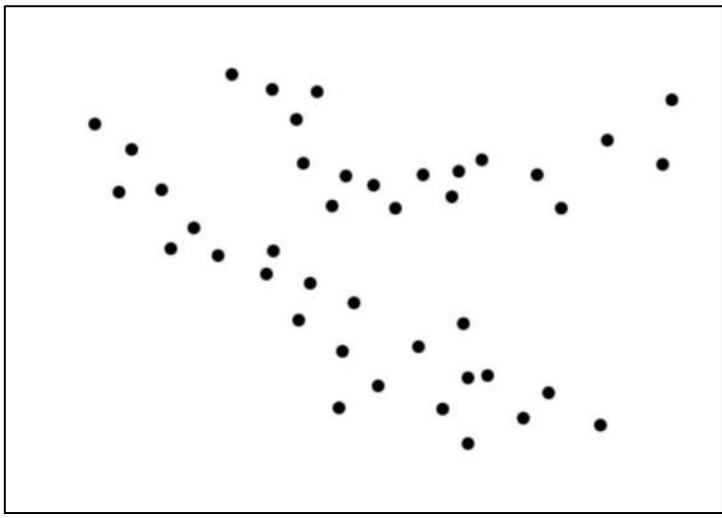
Типи структур кластерів



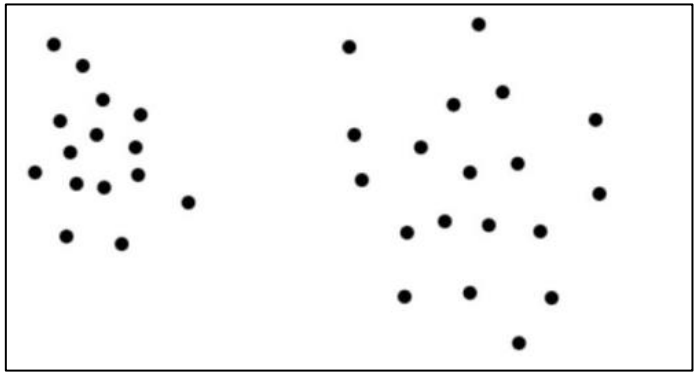
Кластери з центрами



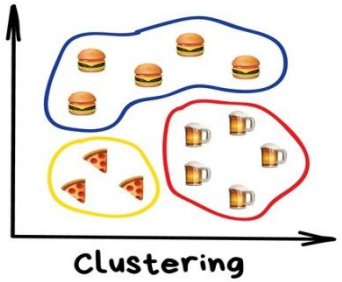
Стрічкові кластери



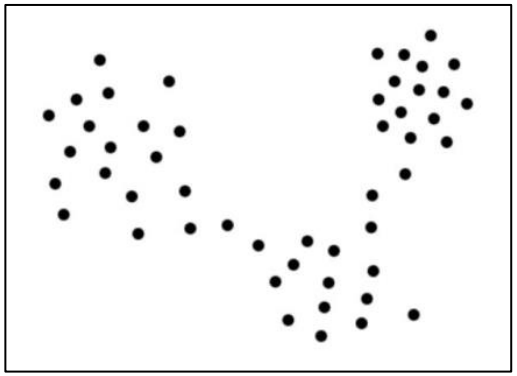
Внутрішньокластерні відстані менші за міжкластерні



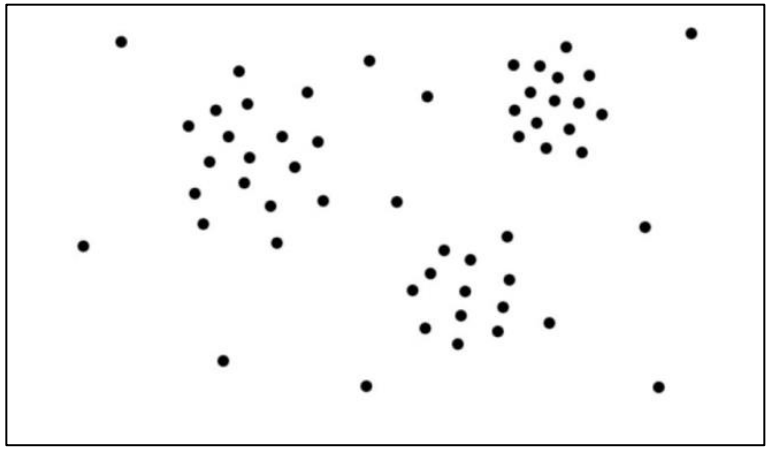
Кластеризація
Типи структур кластерів



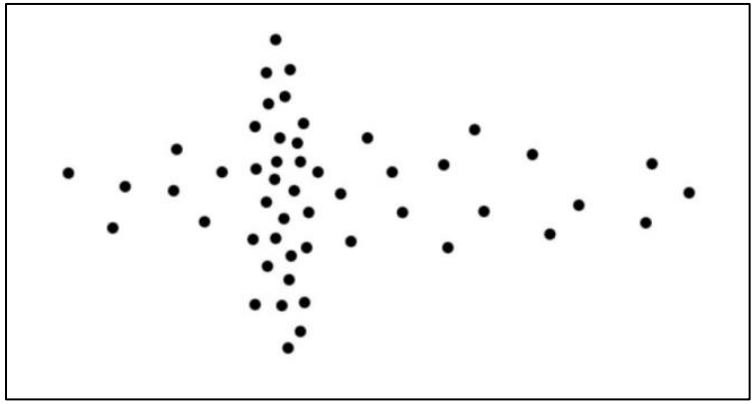
Існування перемичок
між кластерами



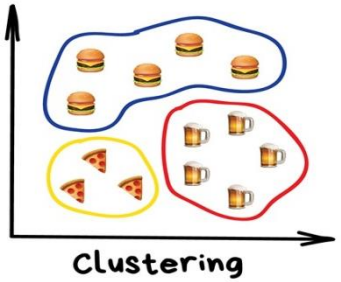
Розріджене тло
з нетипових об'єктів



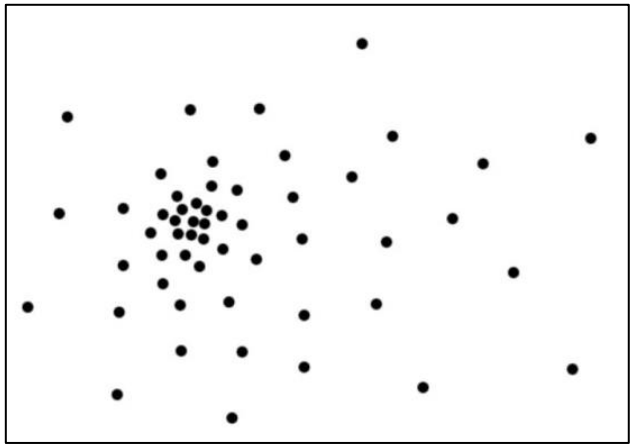
Кластери що
перекриваються

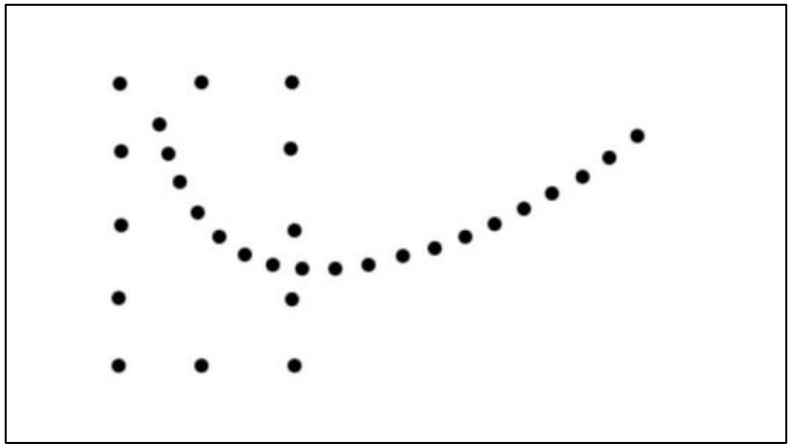


Кластеризація
Типи структур кластерів



Кластерів взагалі може не існувати

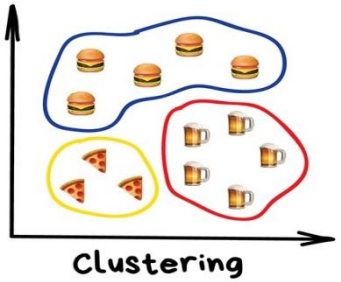




- Кожен метод кластеризації має свої обмеження і виділяє кластери лише декількох типів
- Поняття «тип кластерної структури» залежить від методу і також не має формального визначення

Кластеризація

Методи кластеризації



«Кластеризацією» зазвичай вважають такий набір кластерів, які містять усі об'єкти набору даних. Додатково, можна розглянути відношення між кластерами. Наприклад, ієрархію вкладеності кластерів один у одного. Грубо можна виділити такі кластеризації:

•*Жорстка кластеризація*. Кожен об'єкт або належить кластеру або ні.

•*М'яка кластеризація (також нечітка кластеризація)*. Кожен об'єкт належить кожному кластеру до певної міри. Наприклад, це ймовірність належності кластеру.

Серед них виділяють декілька доладних:

•*Жорстке розбиття на кластери*. Кожен об'єкт належить рівно одному кластеру.

•*Жорстке розбиття на кластери з викидами*. Об'єкт може не належати жодному кластеру і розглядається як викид.

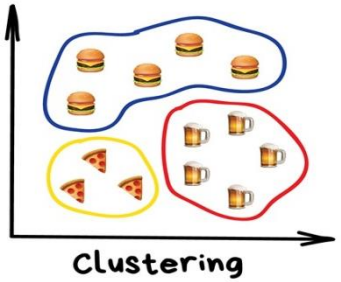
•*Кластери з перетином*. Об'єкт може належати більш ніж одному кластеру.

•*Ієрархічна кластеризація*. Якщо об'єкт належить нащадку, то він також належить і предку.

•*Підпросторова кластеризація*. Хоч кластери і можуть перетинатись, проте в межах визначеного підпростору кластери не перетинаються.

Кластеризація

Методи кластеризації



Типовими кластерними моделями є:

•*Моделі зв'язності*. Наприклад, ієрархічна кластеризація або таксономія будуються на основі відстані між вузлами.

•*Центроїдні моделі*. Наприклад, метод K-середніх (K-means) представляє кожен кластер єдиним усередненим вектором.

•*Статистичні моделі*. Кластери будуються ґрунтуючись на статистичних розподілах.

•*Моделі засновані на щільності*. Наприклад, в DBSCAN і в OPTICS кластери визначаються як зв'язані області відповідної щільності у просторі даних.

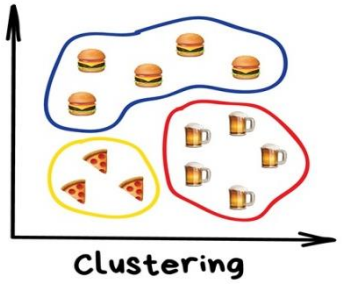
•*Групові моделі*. Деякі алгоритми не забезпечують вдосконалену модель для своїх результатів, а просто описують групування об'єктів.

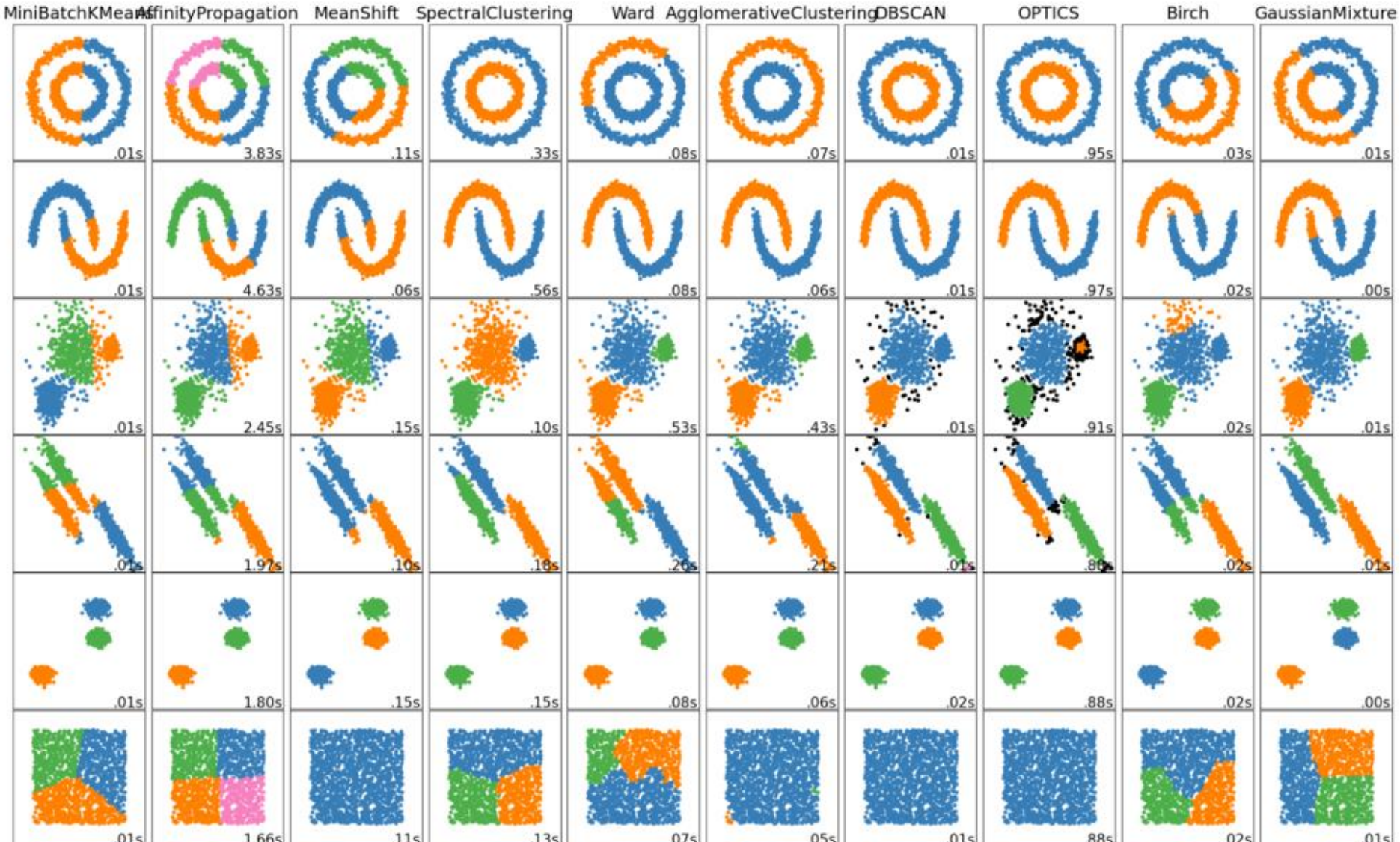
•*Графові моделі*. Поняття кліки (така підмножина вершин, в якій кожна пара вершин з'єднана ребром) у графі слугує прототипом кластеру.

•*Нейронні моделі*. Найбільш відомою моделлю нейронної мережі з навчанням без учителя є нейронна мережа Кохонена.

Кластеризація

Методи кластеризації





Дякую за увагу