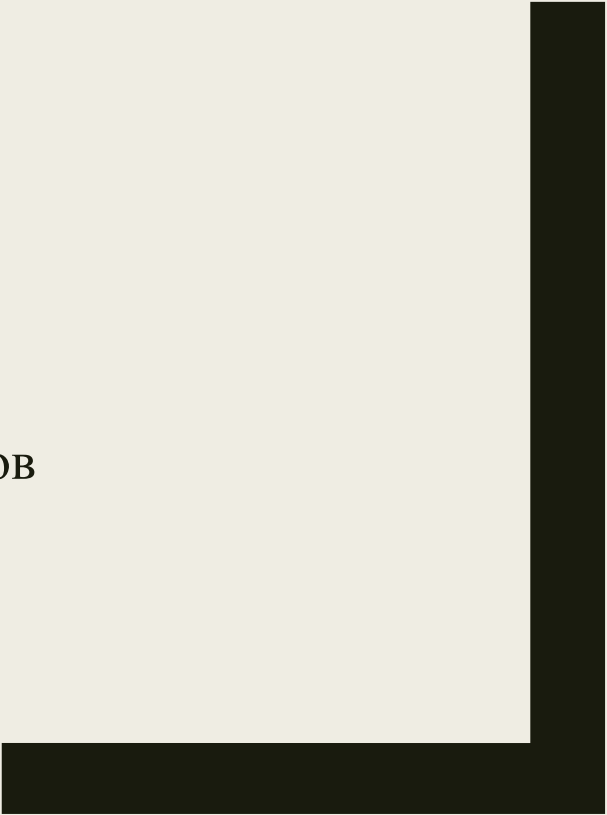




DBSCAN

Павел Филиппов, Алексей Бирюков
5030102/10101

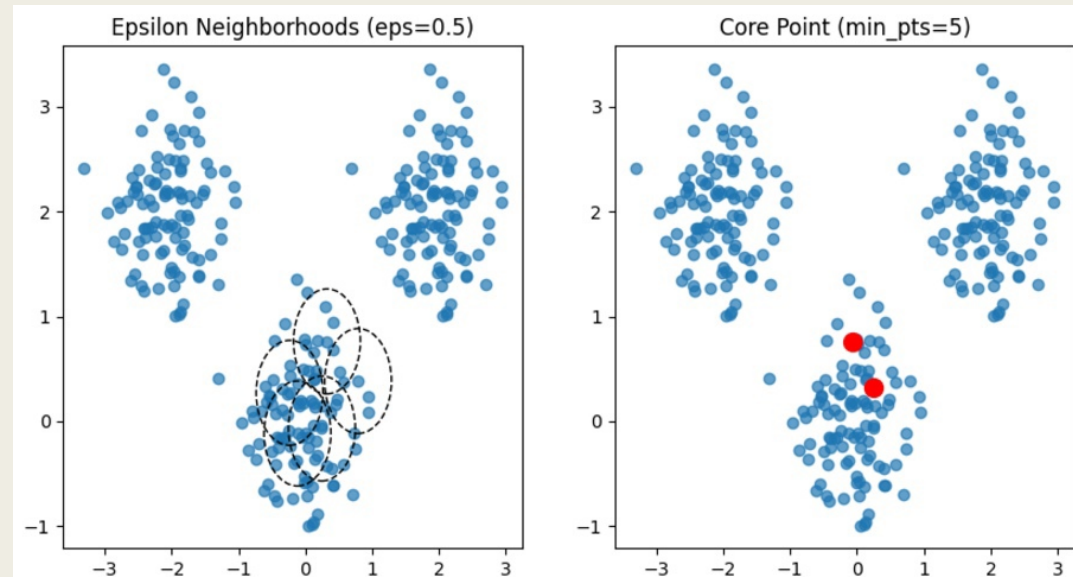


История алгоритма

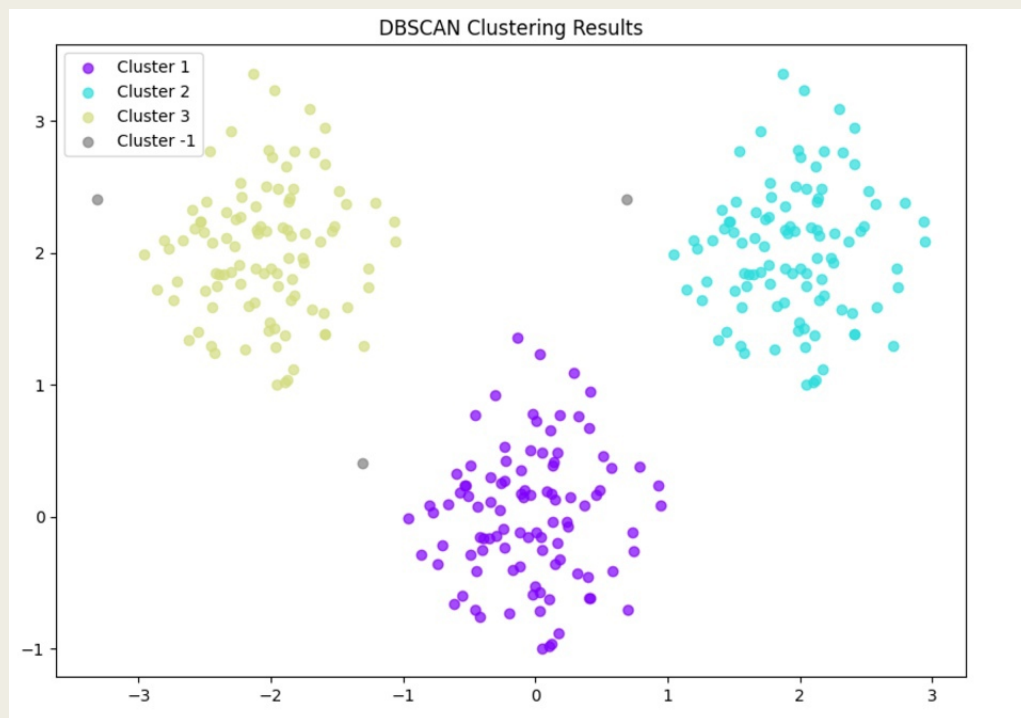
- 1972 – Роберт Ф. Линг опубликовал статью «Теория и построение k-кластеров» со схожим алгоритмом, который можно считать предшественником
- 1996 – Мартин Эстер, Ганс-Петер Кригель, Йёрг Сандер и Суй Сяовэй предложили алгоритм кластеризации данных DBSCAN
- 2014 – Алгоритм получил премию “Проверено временем”
- 2020- опубликована статья “DBSCAN пересмотренный, пересмотренный: почему и как вы должны (до сих пор) использовать DBSCAN”, пользующаяся большой популярностью

Принцип работы

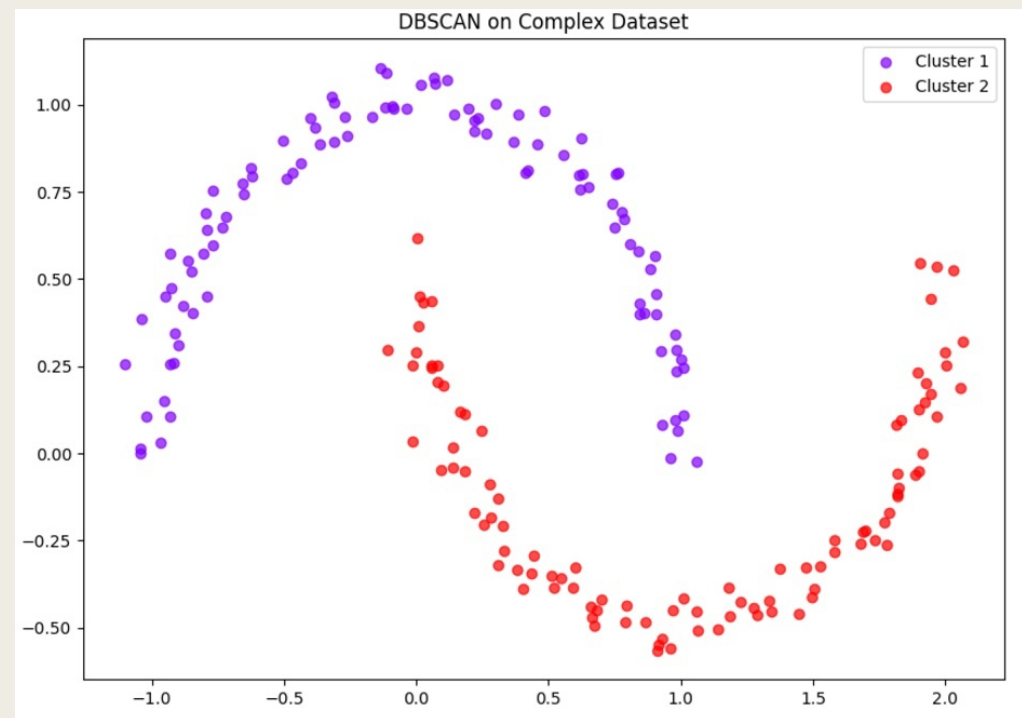
- Находим точки в ϵ окрестности каждой точки и выделяем основные точки с более чем minPts соседями
- Находим связные компоненты основных точек на графе соседей, игнорируя все неосновные точки
- Назначаем каждую неосновную ближайшему кластеру, если кластер является ϵ -соседним, в противном случае помечаем как шум



Результаты работы

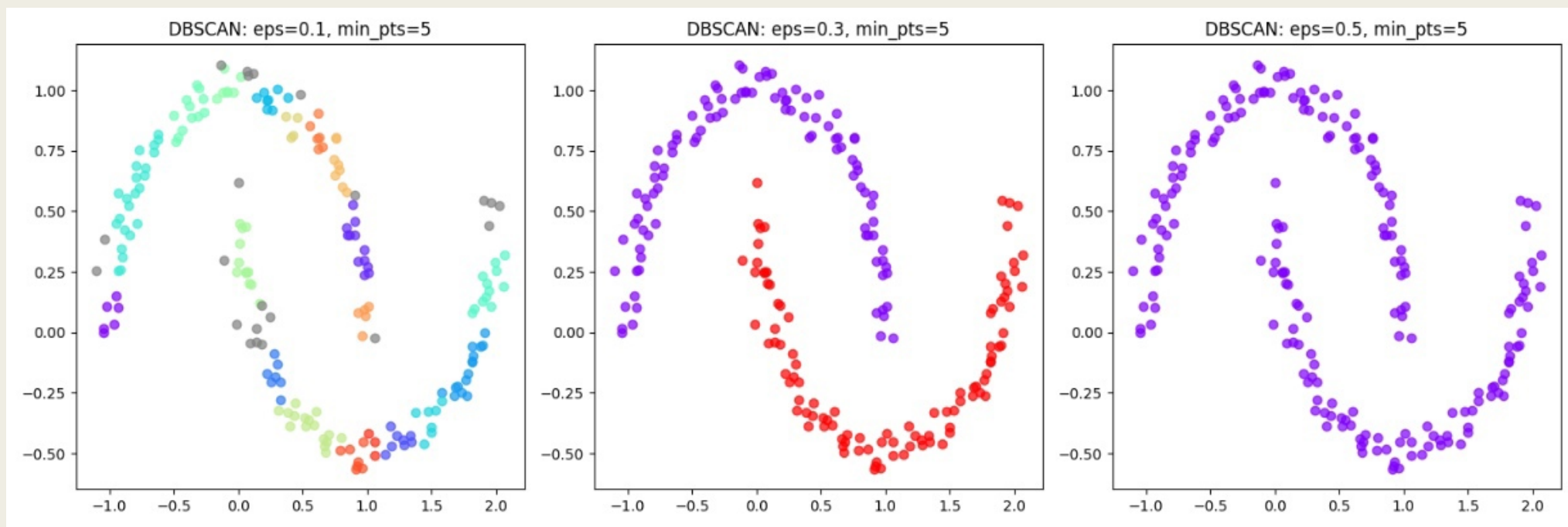


Пример работы на простых данных

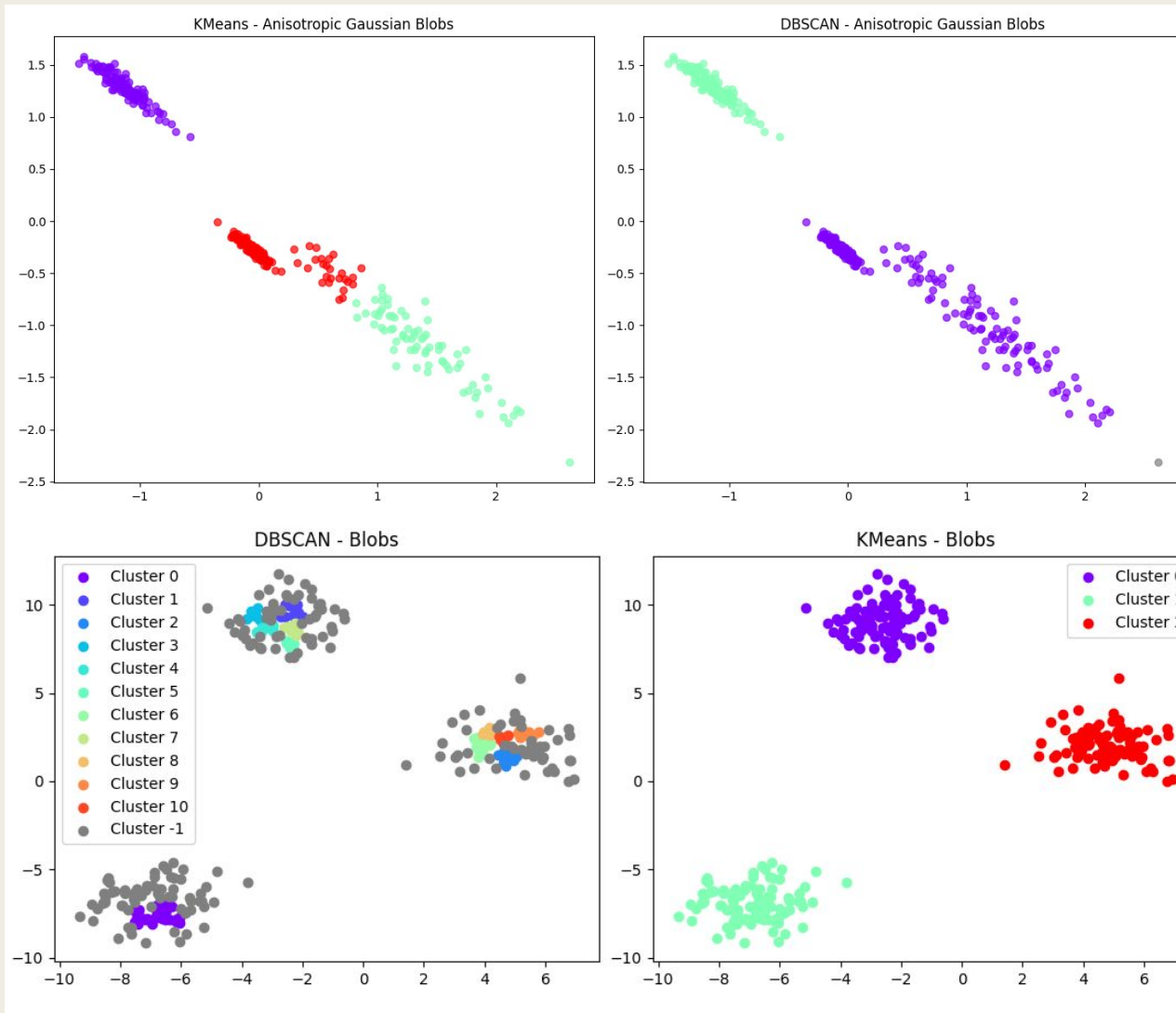


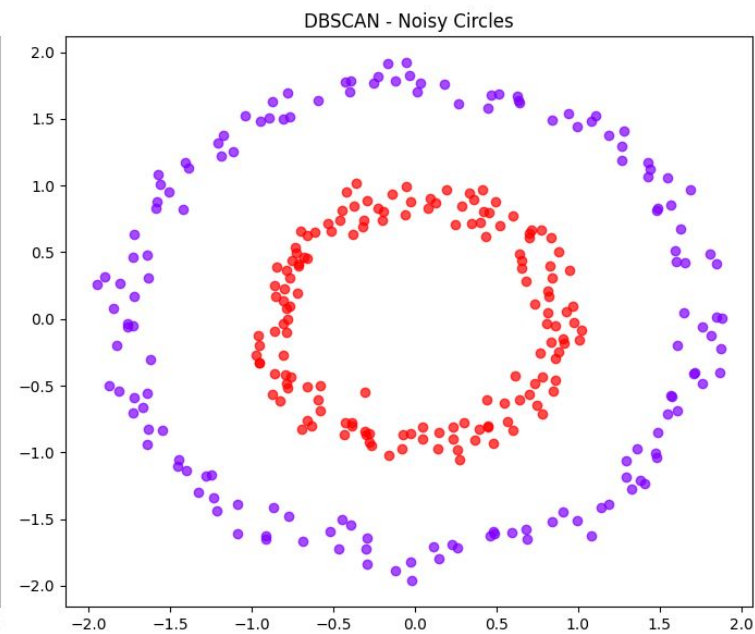
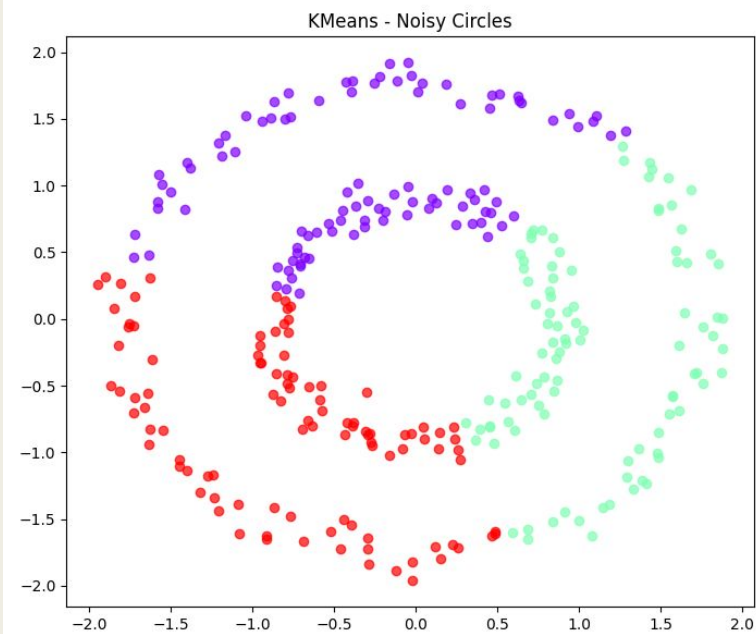
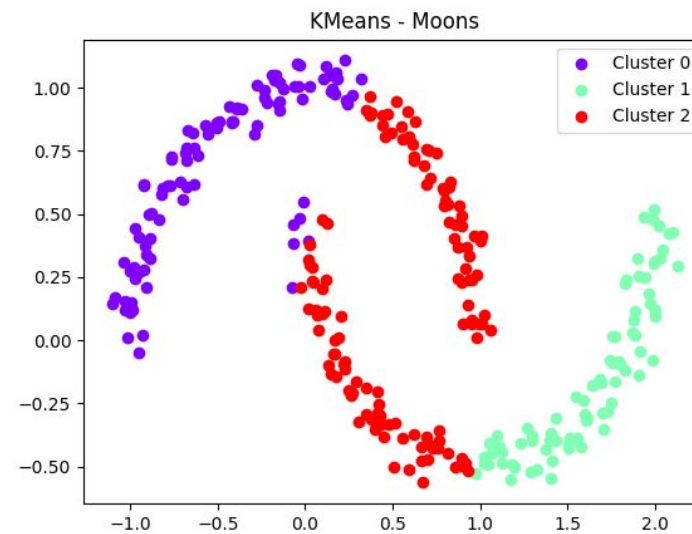
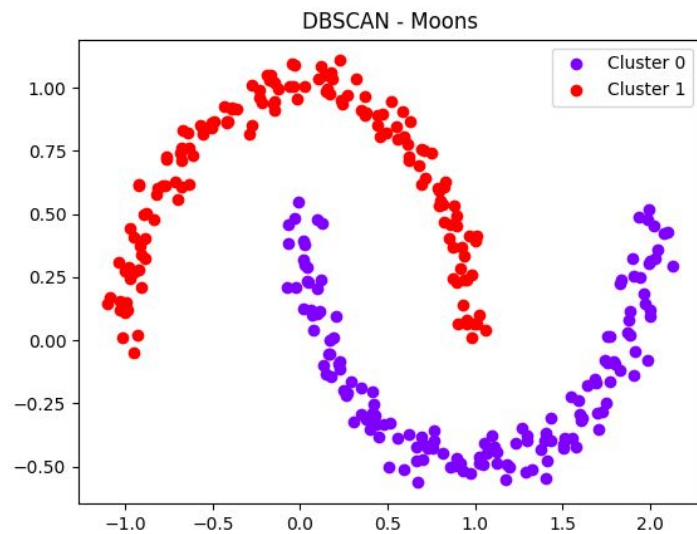
Пример нелинейного разбиения на кластеры

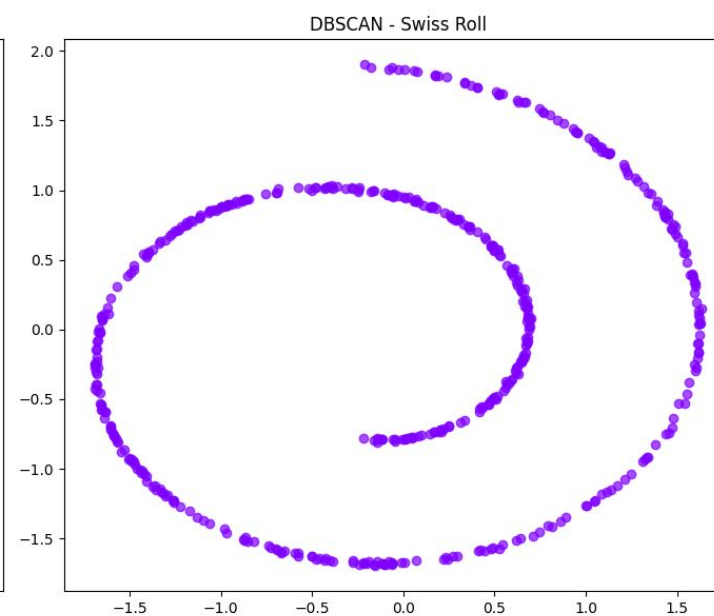
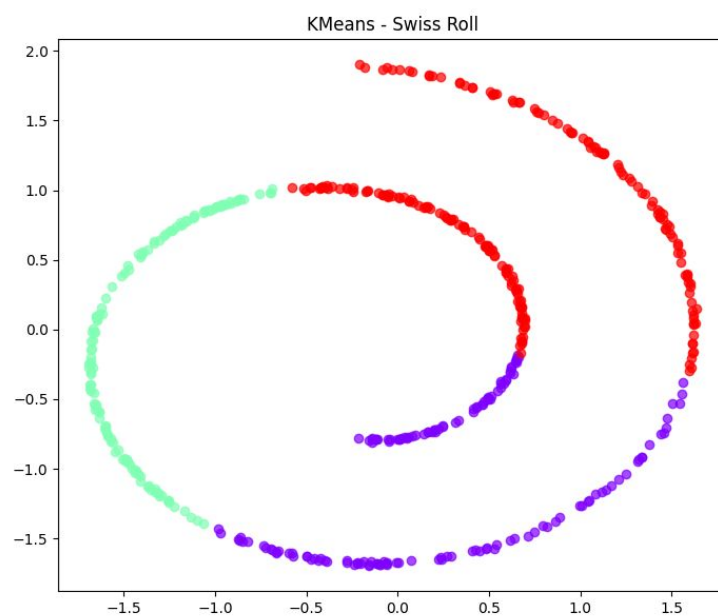
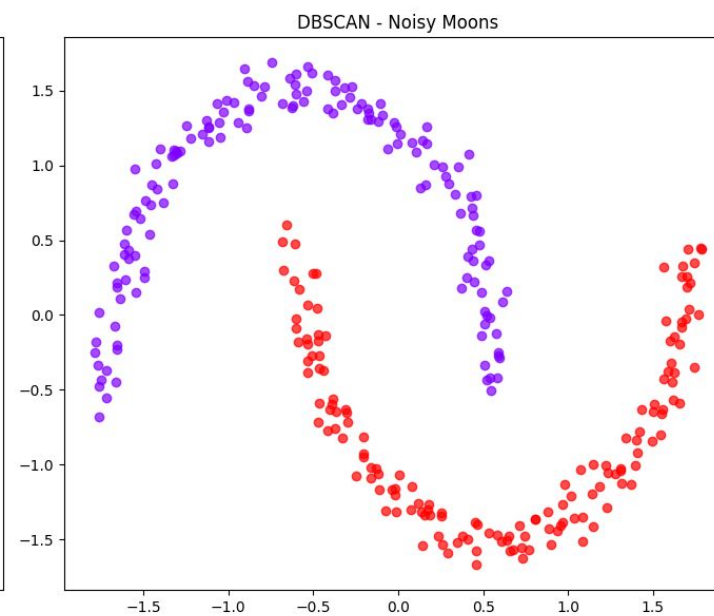
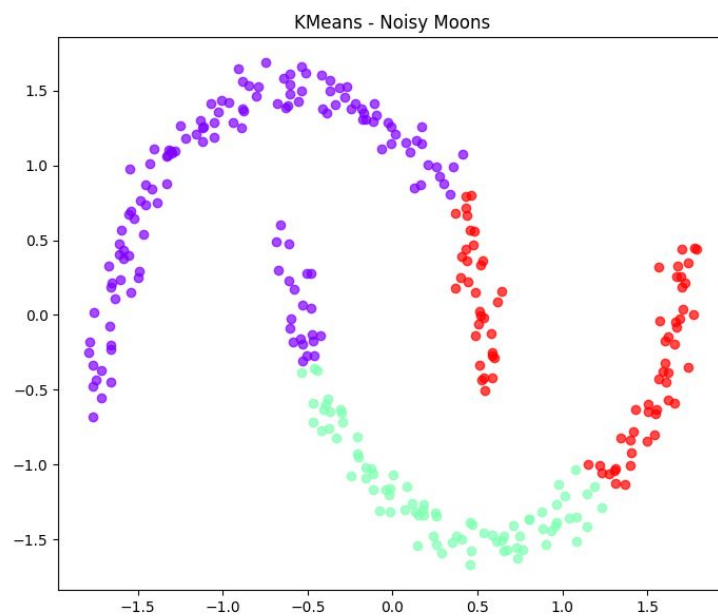
Пример работы с разными входными параметрами

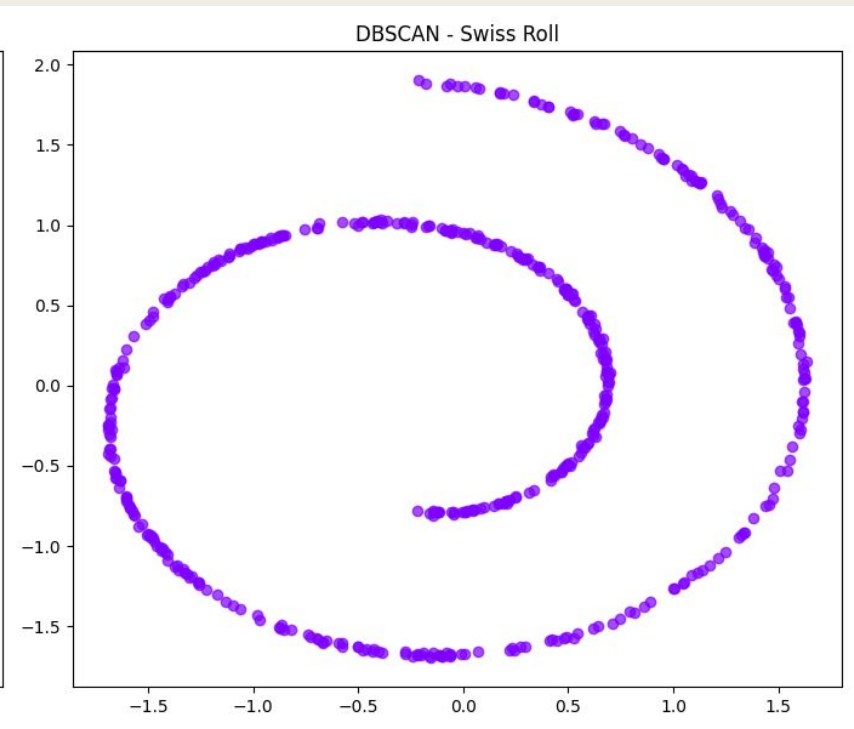
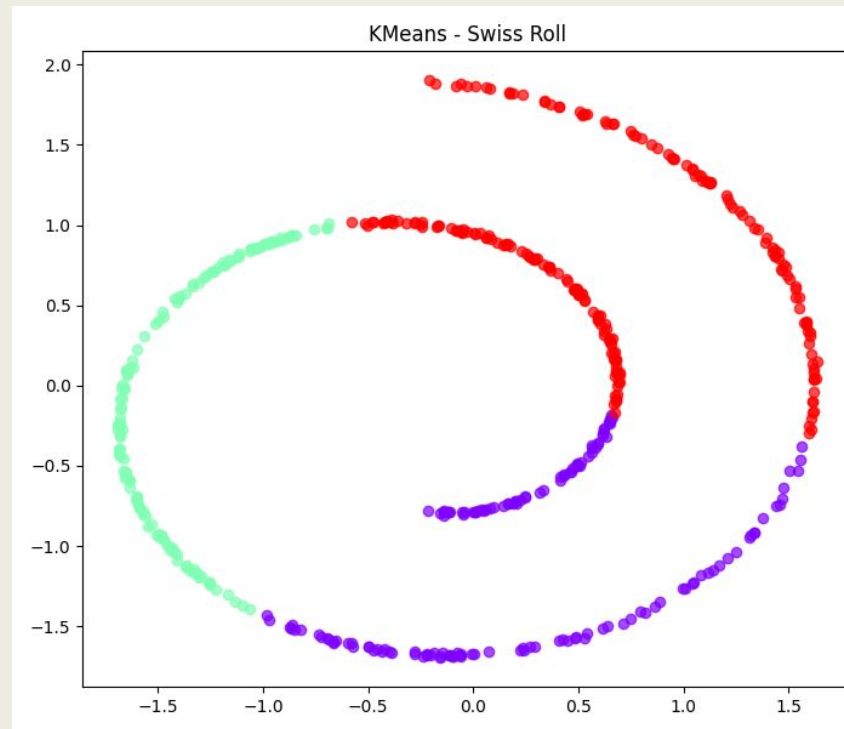


Сравнение KMeans и DBSCAN

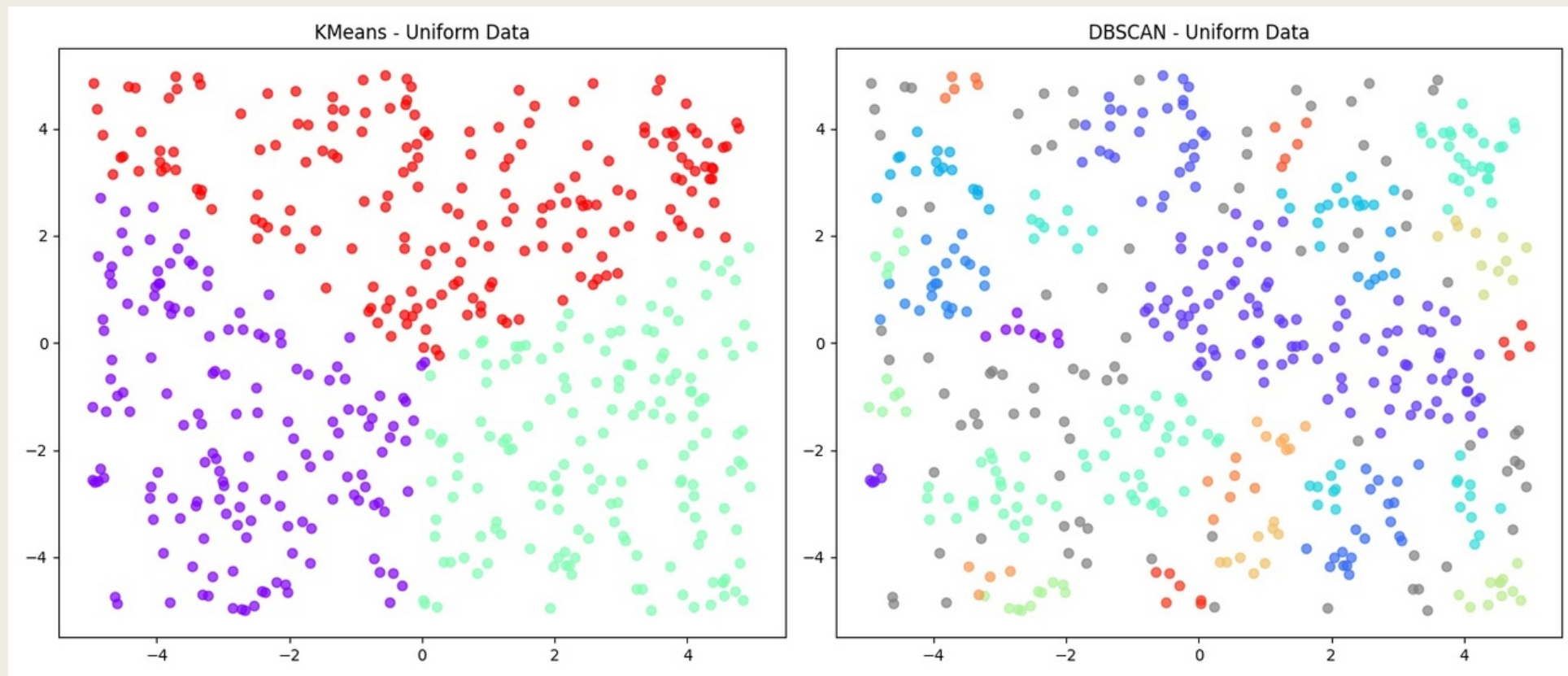








Пример некорректной работы DBSCAN



Преимущества алгоритма

- Не требует указания числа кластеров
- Может найти кластеры произвольной формы
- Имеет понятие шума и устойчив к выбросам
- Нечувствителен к порядку выбора точек
- Может быть оптимизирован

Недостатки алгоритма

- Плохо работает для кластеров с разной плотностью
- Если в данных много шумовых точек, может ошибочно выделить их в отдельные кластеры
- При слишком жестких параметрах может пометить не шумовые как таковые
- “Проклятье размерности”

Оптимизация с помощью cKDTree

cKdtree – оптимизированное бинарное дерево для пространственного разбиения

На тестовых данных удалось ускорить алгоритм в два раза

Список литературы

Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques (3rd ed.). Elsevier.

Tan, P.-N., Steinbach, M., & Kumar, V. (2018). Introduction to Data Mining (2nd ed.). Pearson.

Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer.