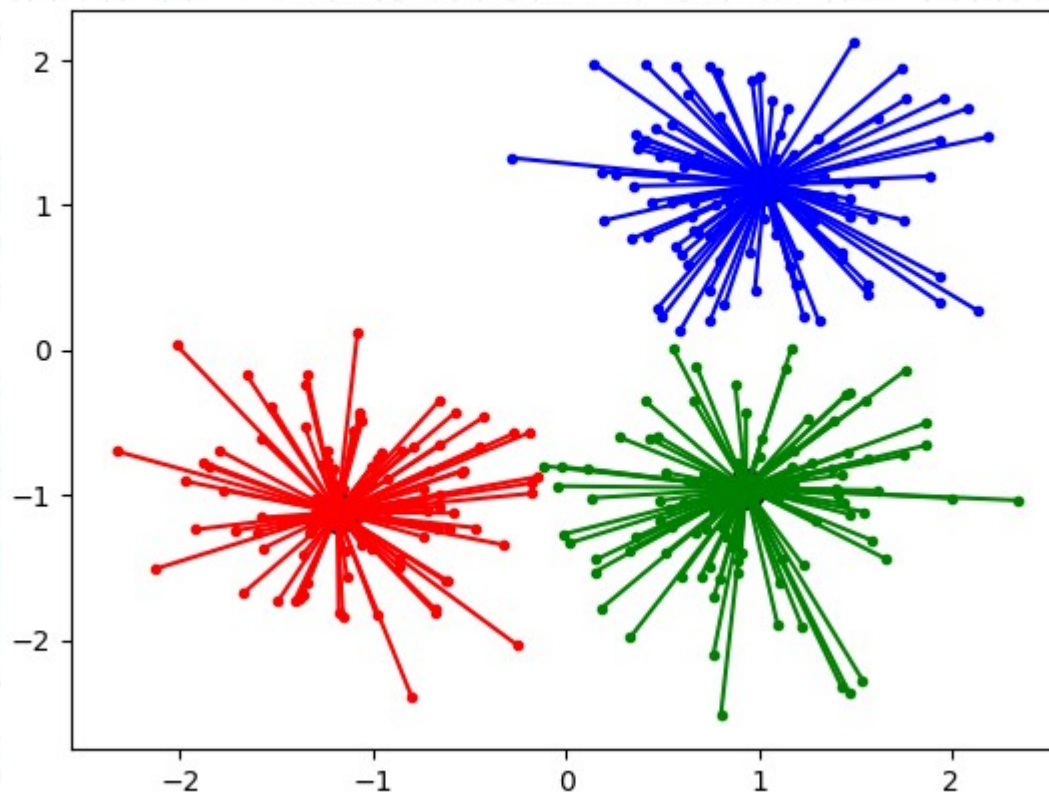


# Машинное обучение

## Анализ объектов



# Содержание лекции

- Выступ объектов в разных алгоритмах
- Эталоны и шумовые выбросы
- Кластеризация



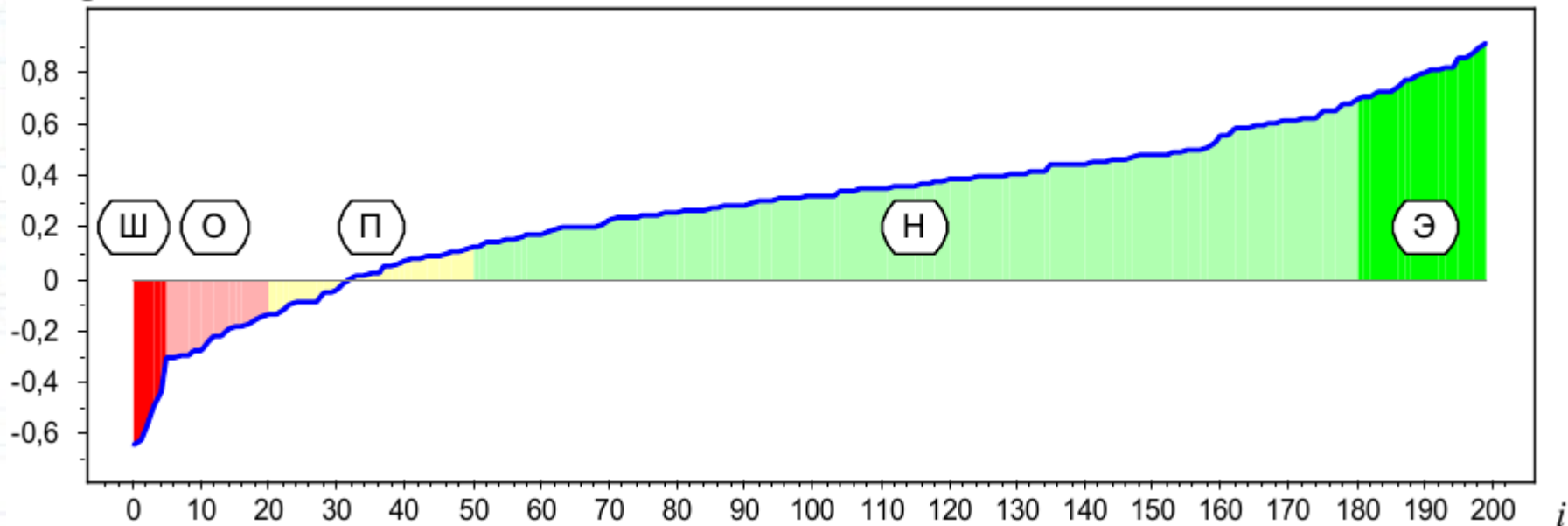
# Отступ (выступ) объекта

- Выступ - это такая функция  $M(x)$ , что для объектов, лежащих глубоко внутри своего класса, т.е. для эталонов она принимает большие положительные значения (это что-то вроде "отступа от границы классов"). Для периферийных объектов, лежащих на границе классов,  $M(x) \approx 0$ .
- Для объектов одного класса, расположенных среди объектов другого класса, выступ  $M(x)$  должен принимать отрицательные значения (что-то вроде "выступа за границу своего класса").

# Типы объектов в зависимости от выступления

- Э — эталонные (можно оставить только их);
- Н — неинформативные (можно удалить из выборки);
- П — пограничные (их классификация неустойчива);
- О — ошибочные (причина ошибки — плохая модель);
- Ш — шумовые (причина ошибки — плохие данные).

*Margin*



# В метрических алгоритмах

- Пусть для заданного  $x \in X$  объекты  $x_1, \dots, x_\ell$  отсортированы по убыванию расстояния до  $x$
- Метрический алгоритм классификации:

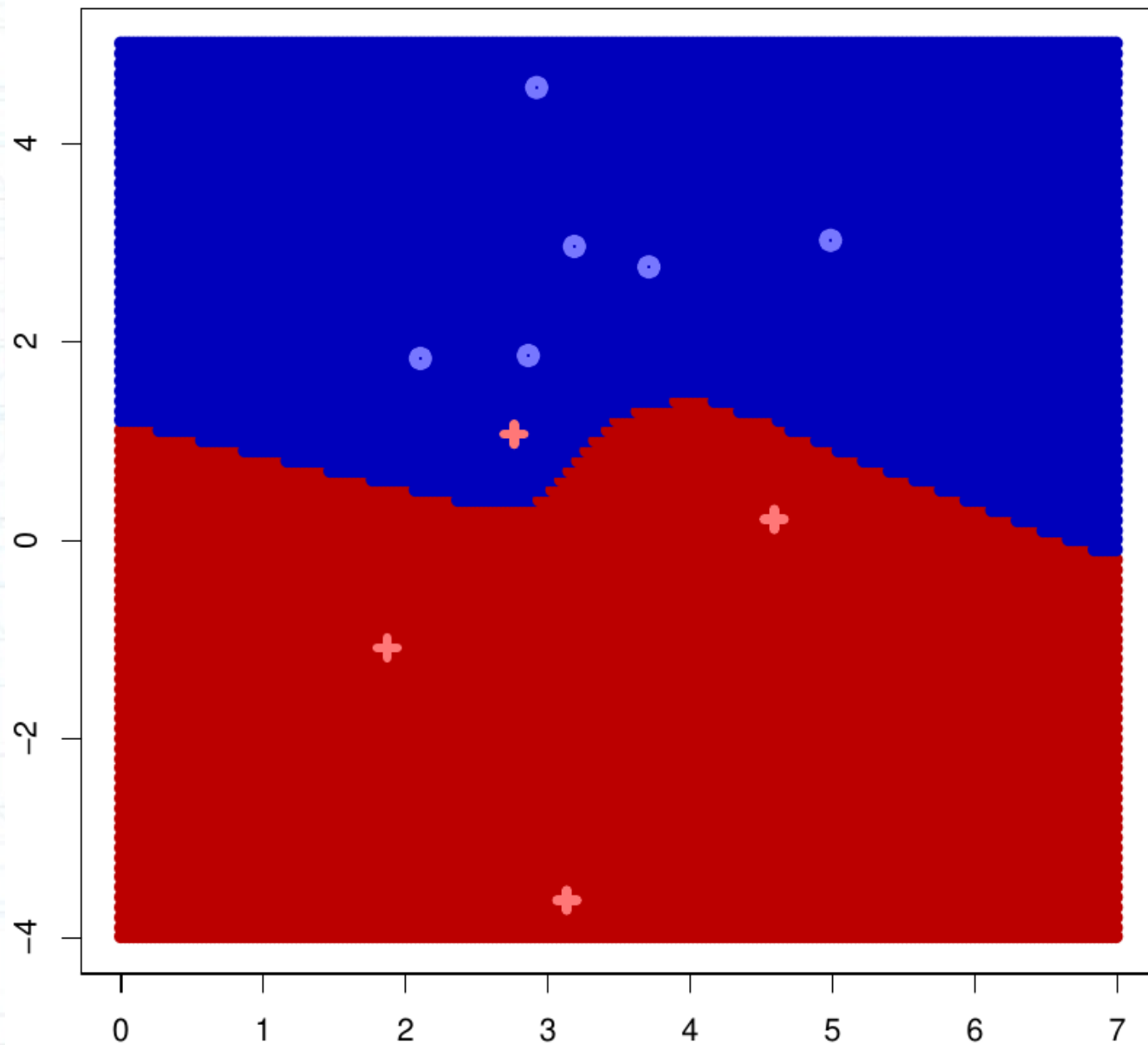
$$a(x) = \arg \max_{y \in Y} \Gamma_y(x) = \arg \max_{y \in Y} \sum_{\substack{i=1 \\ y_i=y}}^{\ell} w(i, x)$$

- $w(i, x)$  — вес (степень важности)  $i$ -го соседа объекта  $x$ ,  $\geq 0$ ,  $\searrow$  по  $i$
- $\Gamma_y(x)$  — близость объекта  $x$  к классу  $y$
- Выступ:  $M(x_i) = \Gamma_{y_i}(x_i) - \max_{y \in Y \setminus y_i} \Gamma_y(x_i)$



# Упражнение

Вычислите отступы для всех объектов для метода 3NN



# Для линейных классификаторов

- Линейный классификатор:  
$$a(x, w) = \text{sign}(x, w)$$
- $(x, w) = 0$  — разделяющая гиперплоскость,
- $M_i(w) = (w, x_i) y_i$  - отступ/выступ объекта  $x_i$

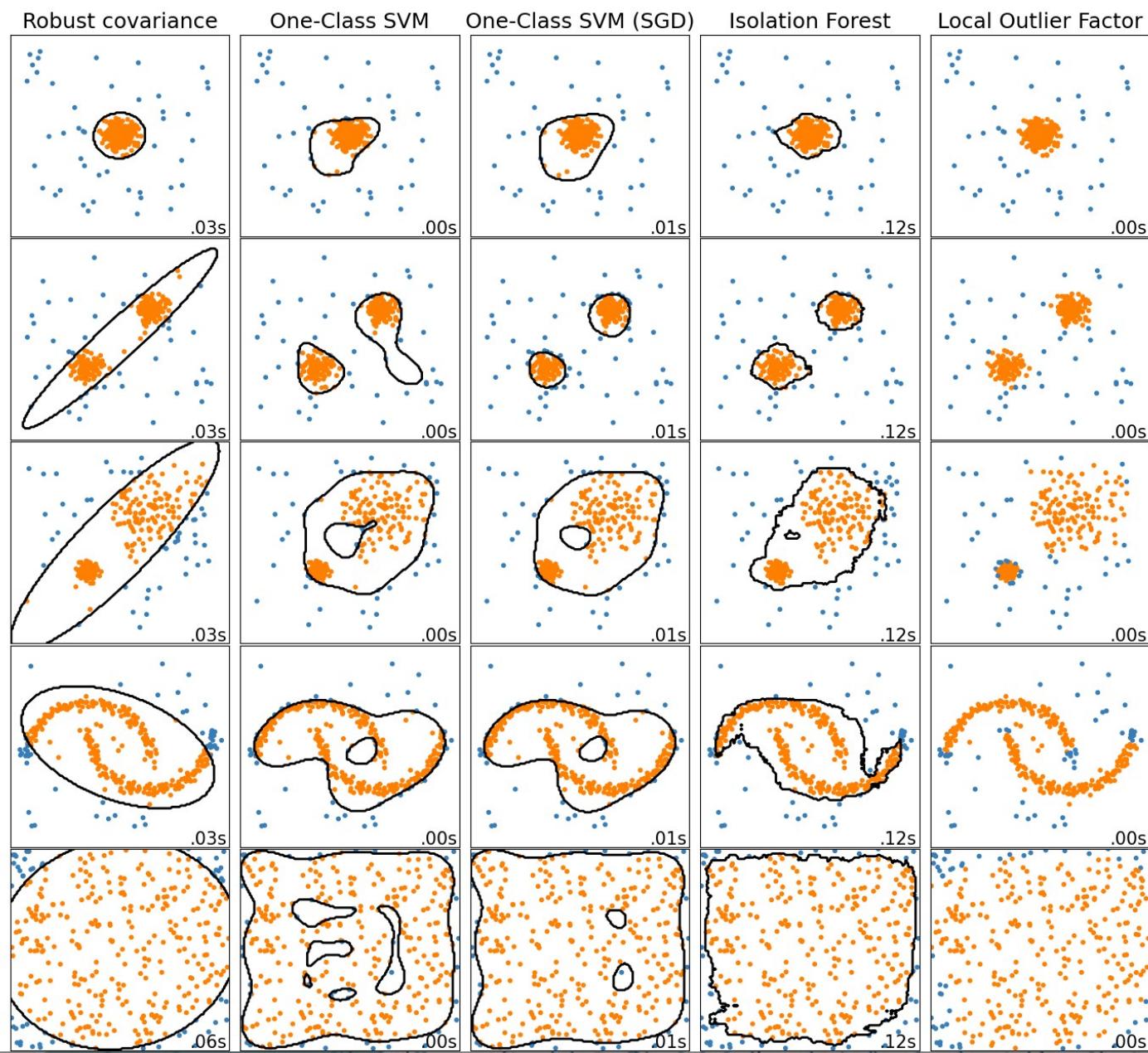
# Для датасетов без целевого признака

- Вероятность появления именно такого  $x$





# Методы отсева случайных выбросов в sklearn



# Задача кластеризации (обучение без учителя)

- Дано:
  - пространство объектов  $X$
  - обучающая выборка  $X^{\ell}$
  - метрика между объектами
- Найти:
  - множество кластеров  $Y$
  - алгоритм кластеризации  $a : X \rightarrow Y$
- Каждый кластер должен состоять из близких объектов
- Объекты разных кластеров должны быть существенно различны



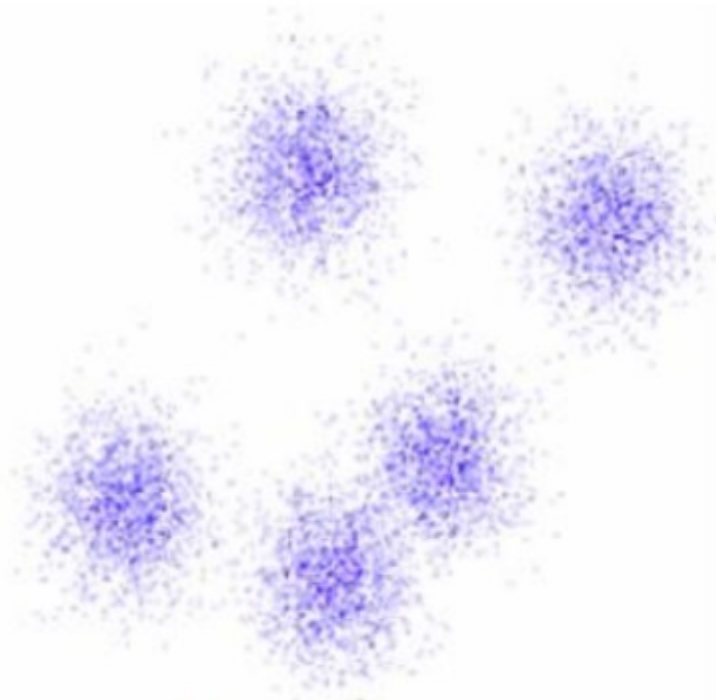
# Классификация и кластеризация

Классификация	Кластеризация
<ul style="list-style-type: none"><li>• Известное количество классов</li><li>• Классы известны для объектов обучающей выборки</li><li>• Используется для классификации объектов “в будущем”</li><li>• Классификация – это обучение с учителем</li></ul>	<ul style="list-style-type: none"><li>• Неизвестно количество классов</li><li>• Нет данных о классах в обучающей выборке</li><li>• Используется для исследования множества объектов</li><li>• Кластеризация – это обучение без учителя</li></ul>

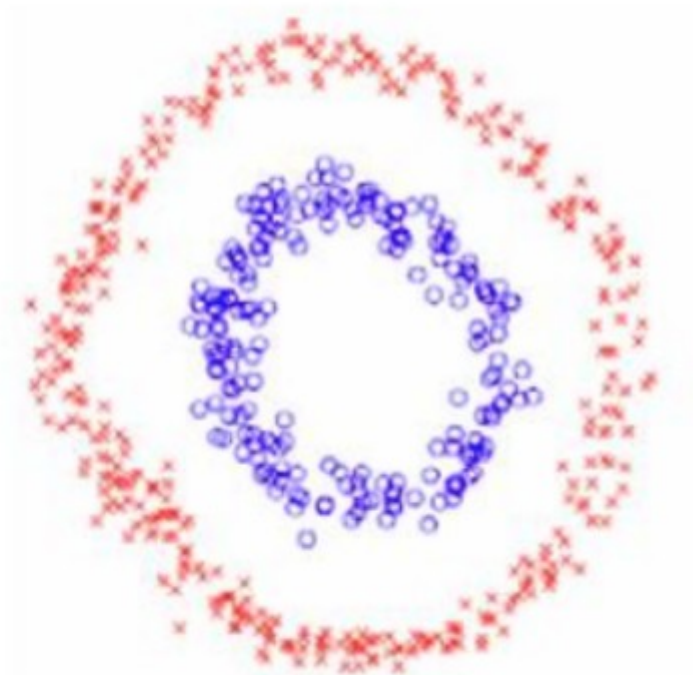


# Близость или связанность?

- Compactness, e.g., k-means, mixture models
- Connectivity, e.g., spectral clustering

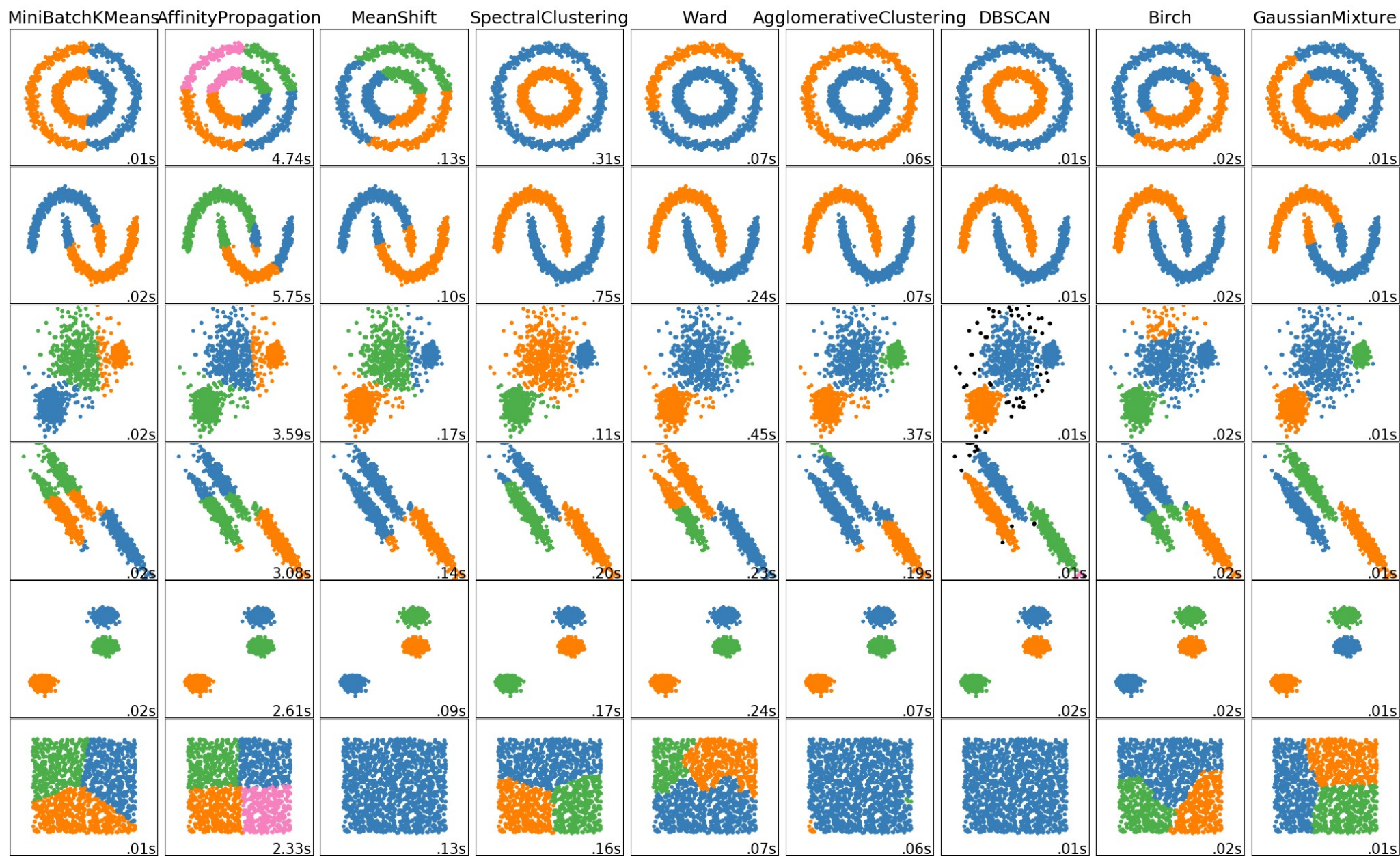


**Compactness**



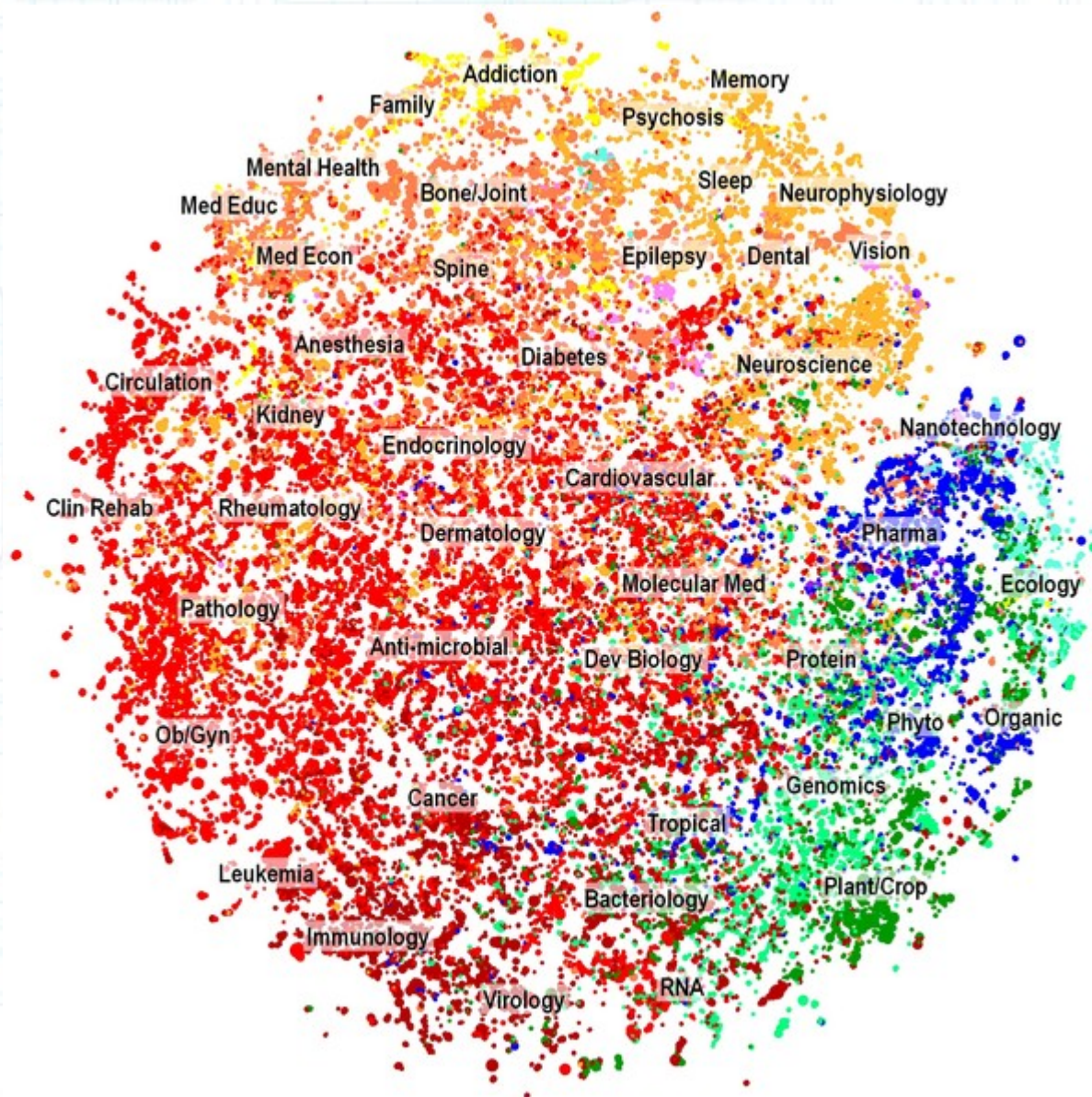
**Connectivity**

# Пример: результаты работы алгоритмов кластеризации





# Пример: кластеризация статей по медицине





# ЕМ-кластеризация

Гипотеза: выборка  $X^{\ell}$  порождена смесью гауссовских случайных распределений

$$p(x) = \sum_{y \in Y} w_y p_y(x), \quad \sum_{y \in Y} w_y = 1,$$

$$p_y(x) = (2\pi)^{-\frac{n}{2}} (\sigma_{y1} \cdots \sigma_{yn})^{-1} \exp \left( -\frac{1}{2} \rho_y^2(x, \mu_y) \right)$$

$\mu_y = (\mu_{y1}, \dots, \mu_{yn})$  — центр кластера  $y$ ;

$\Sigma_y = \text{diag}(\sigma_{y1}^2, \dots, \sigma_{yn}^2)$  — диагональная матрица ковариаций;

$$\rho_y^2(x, x') = \sum_{j=1}^n \sigma_{yj}^{-2} |f_j(x) - f_j(x')|^2.$$

# ЕМ-кластеризация

1: начальное приближение  $w_y$ ,  $\mu_y$ ,  $\Sigma_y$  для всех  $y \in Y$ ;

2: **повторять**

3: Е-шаг (expectation):

$$g_{iy} := P(y|x_i) \equiv \frac{w_y p_y(x_i)}{\sum_{z \in Y} w_z p_z(x_i)}, \quad y \in Y, \quad i = 1, \dots, \ell;$$

4: М-шаг (maximization):

$$w_y := \frac{1}{\ell} \sum_{i=1}^{\ell} g_{iy}, \quad y \in Y;$$

$$\mu_{yj} := \frac{1}{\ell w_y} \sum_{i=1}^{\ell} g_{iy} f_j(x_i), \quad y \in Y, \quad j = 1, \dots, n;$$

$$\sigma_{yj}^2 := \frac{1}{\ell w_y} \sum_{i=1}^{\ell} g_{iy} (f_j(x_i) - \mu_{yj})^2, \quad y \in Y, \quad j = 1, \dots, n;$$

5:  $y_i := \arg \max_{y \in Y} g_{iy}, \quad i = 1, \dots, \ell;$

6: **пока**  $y_i$  не перестанут изменяться;

# Метод k-средних

1: начальное приближение центров  $\mu_y$ ,  $y \in Y$ ;

2: **повторять**

3: **аналог E-шага:**

отнести каждый  $x_i$  к ближайшему центру:

$$y_i := \arg \min_{y \in Y} \rho(x_i, \mu_y), \quad i = 1, \dots, \ell;$$

4: **аналог M-шага:**

вычислить новые положения центров:

$$\mu_{yj} := \frac{\sum_{i=1}^{\ell} [y_i = y] f_j(x_i)}{\sum_{i=1}^{\ell} [y_i = y]}, \quad y \in Y, \quad j = 1, \dots, n;$$

5: **пока**  $y_i$  не перестанут изменяться;



# Сравнение k-средних и EM-кластеризации



# DBSCAN

- Density-Based Spatial Clustering of Applications with Noise – самый популярный алгоритм кластеризации
- Ключевые понятия:
  - Внутренняя точка – имеет более MinPts соседей ( $r < \text{Eps}$ )
  - Граничная точка – имеет меньше соседей, но является соседней к какой-либо внутренней точке
  - Остальные точки - шумовые
  - Достижимость по плотности: точка  $q$  достижима из внутренней точки  $p$ , если существует последовательность Eps-соседних внутренних точек от  $p$  к  $q$

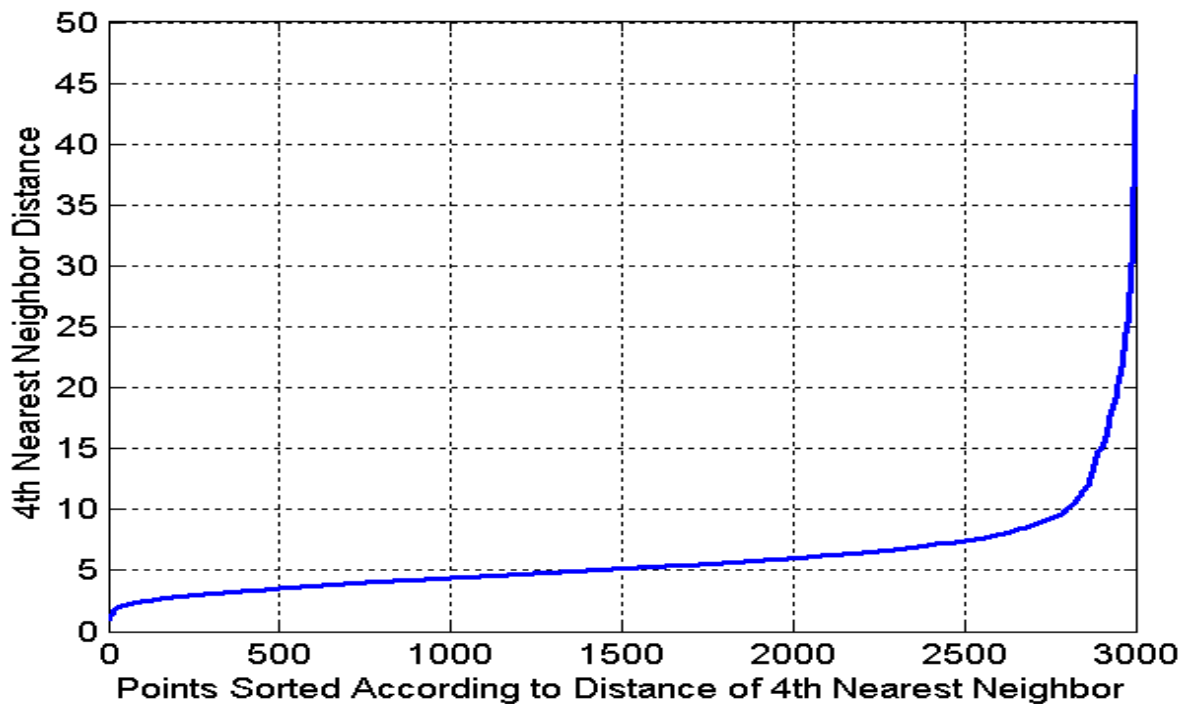


# Алгоритм DBSCAN

- Выбрать точку  $p$
- Если  $p$ -внутренняя, то
  - Найти все достижимые по плотности точки из  $p$
  - Сформировать кластер
- Иначе – перейти к следующей точке
- Результат не зависит от порядка просмотра точек

# DBSCAN: выбор Eps и MinPts

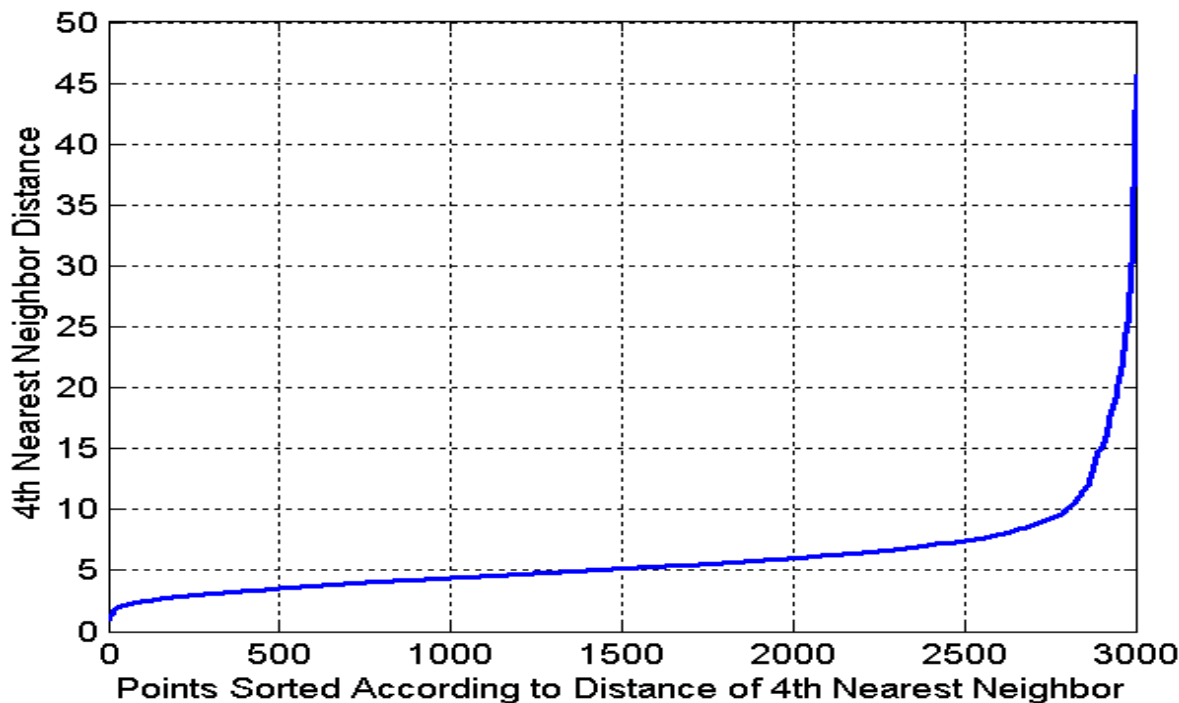
- Ключевая идея: для всех точек одного кластера их  $k$ -тый сосед ( $k < \text{размера кластера}$ ) находится на приблизительно одном и том же расстоянии
- Соседи шумовых точек – далеко
- График отсортированных расстояний:





# DBSCAN: выбор Eps и MinPts

- Искомое Eps - начало крутого подъема на графике расстояний до соседа с фиксированным номером
- MinPts – номер соседа



# Применение кластеризации в Feature engineering

- Создание информативного признака (номер кластера) по заданному набору других признаков
- Сокращение размерности: большой набор признаков сводим к одному номеру кластера