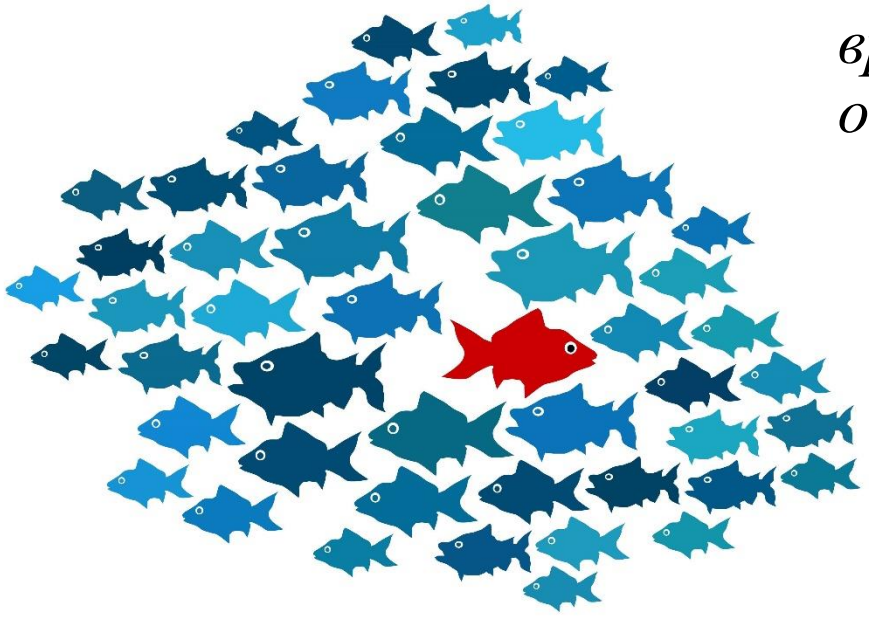
Задача поиска аномалий в данных



Иное возможно лишь как частность, временное, случайное стечение обстоятельств, не как норма.

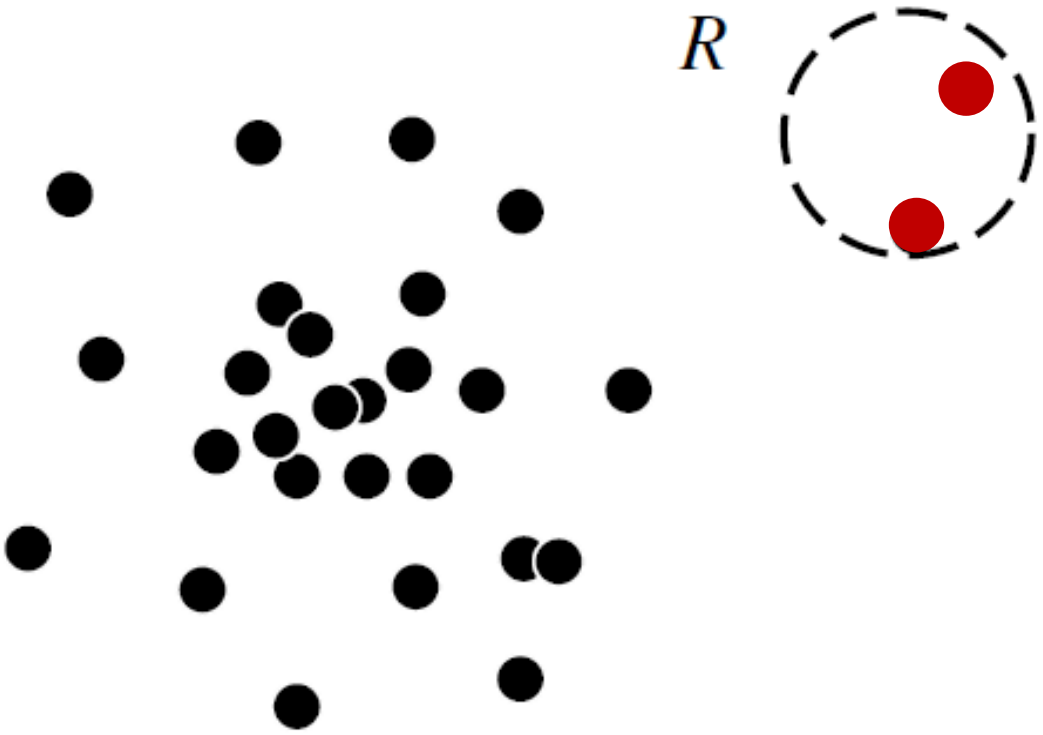
Дмитрий Менделеев

Содержание

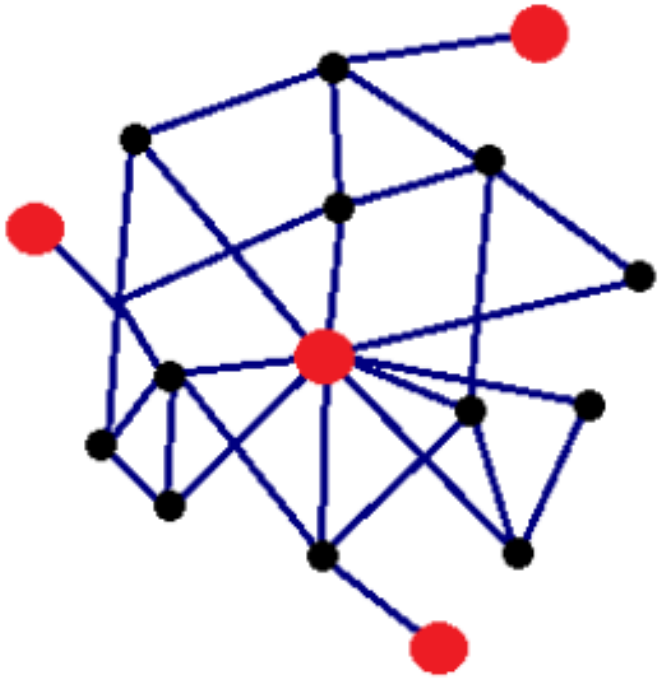
- Понятие аномалии, виды аномалий
- Методы поиска аномалий

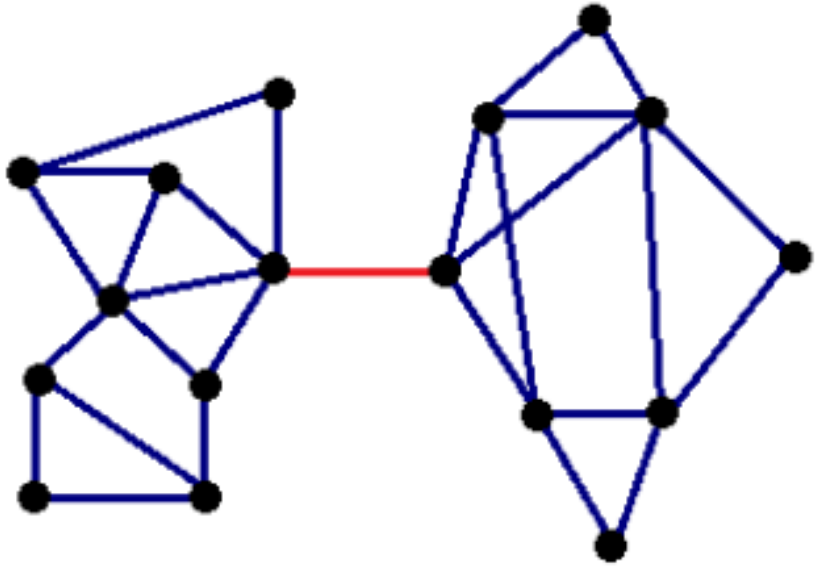
Аномалии (выбросы)

- Аномалия (anomaly, outlier) – объект набора, который значительно отличается от всех остальных объектов



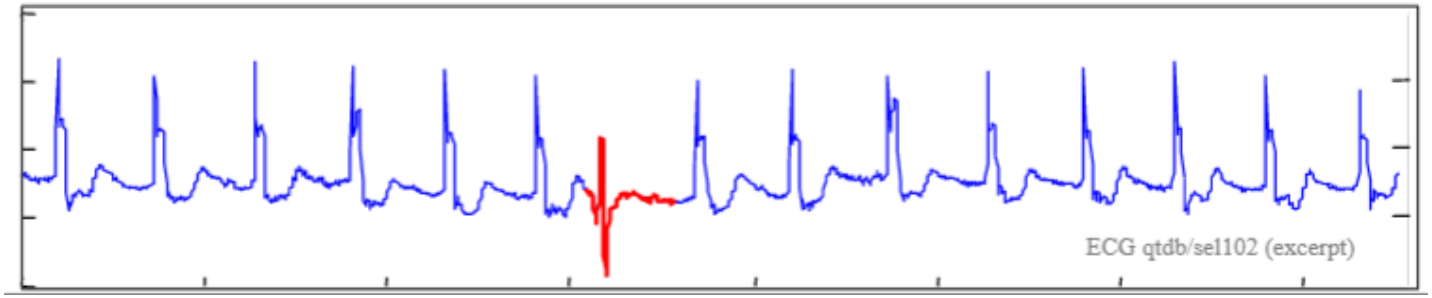
Примеры аномалий: графы, последовательности

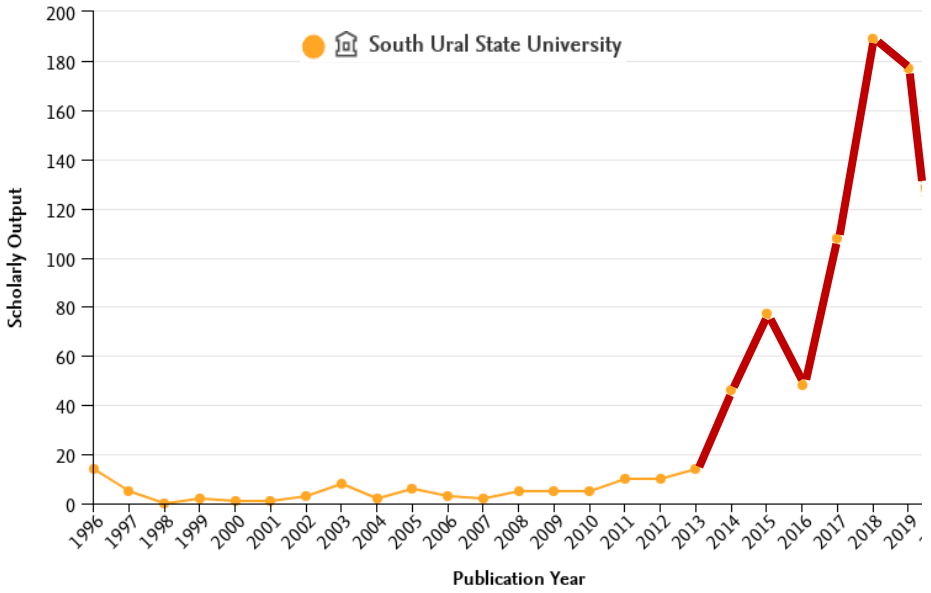




AAABBCCAABBBCAAABBCABBCCABAABBCCAAABBBBCABBBCC

Примеры аномалий: временные ряды





Обучающая выборка имеет значение



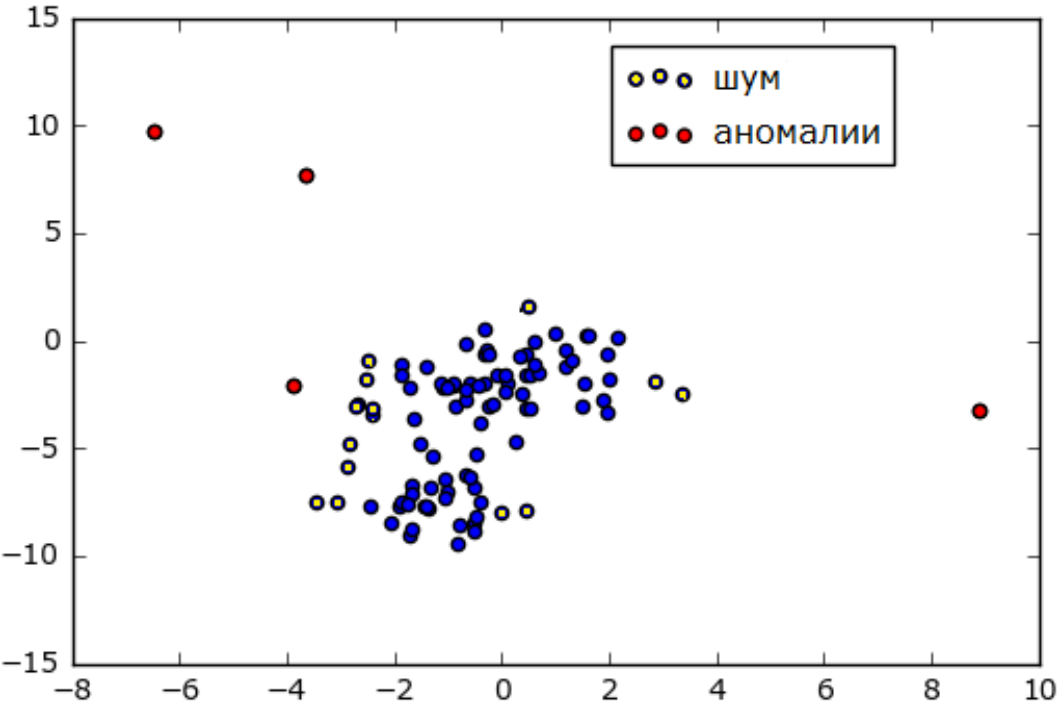






Аномалия *vs.* шум

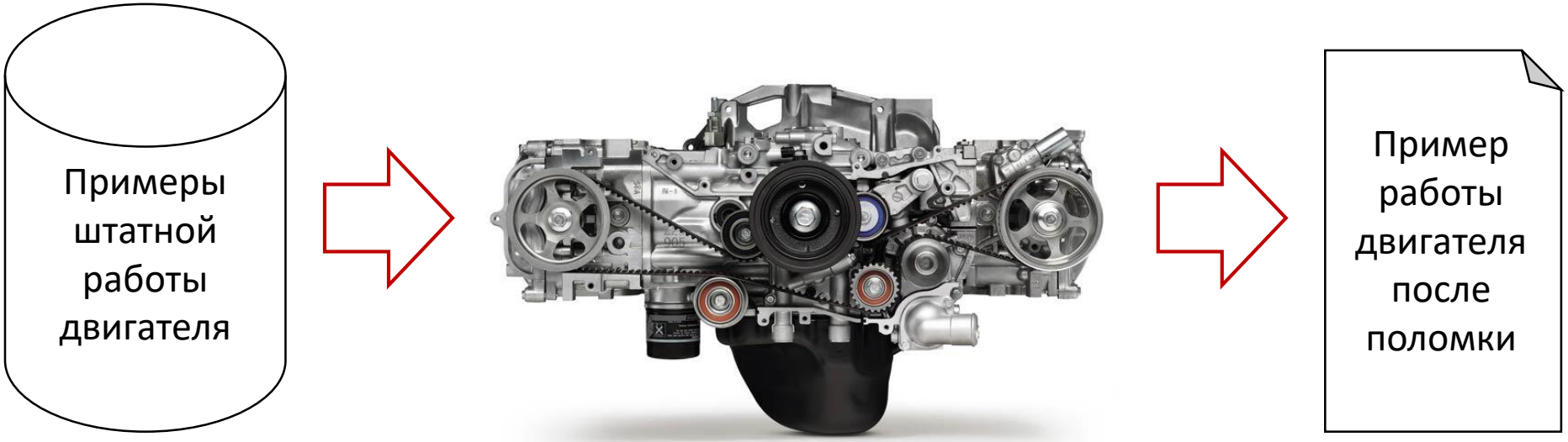
- Шум (noise) – случайная ошибка или вариация в данных



Дата	Сумма	Покупка
01.09.20	100	Кофе, сэндвич
02.09.20	100	Кофе, сэндвич
03.09.20	1000	Дефлопэ
04.09.20	90	Чай, сэндвич
05.09.20	100	Кофе, сэндвич
06.09.20	60	Кофе
07.09.20	160	2 кофе, сэндвич
08.09.20	100	Кофе, сэндвич
09.09.20	100	Кофе, сэндвич
10.09.20	100	Кофе, сэндвич

Аномалия *vs.* новизна

- Новизна (novelty) – поведение объекта, имеющее принципиальные отличия от предыстории, которое после обнаружения трактуется как норма



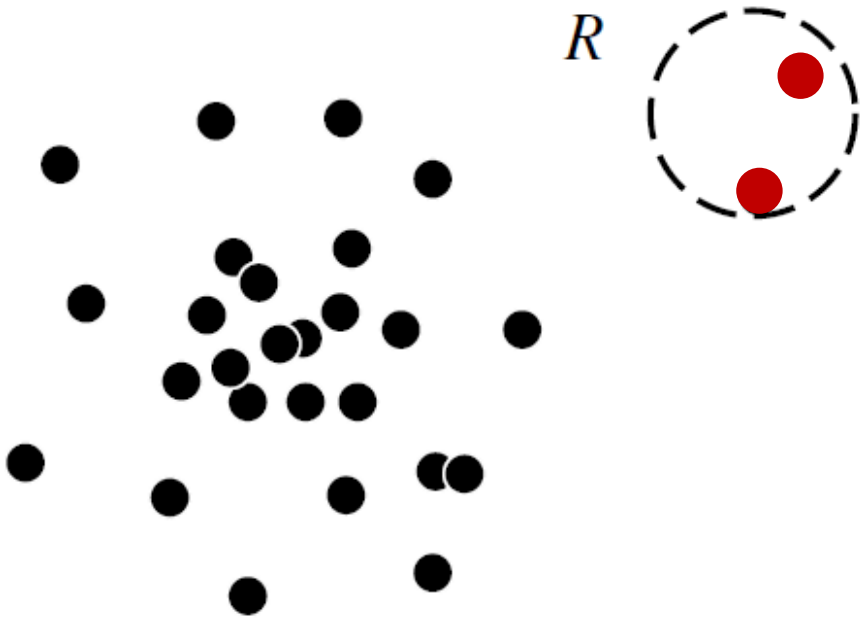
Применение поиска аномалий

- Обнаружение мошенничества с кредитными картами
- Предиктивное техническое обслуживание
- Медицина
- Сегментирование клиентов

Классификация аномалий

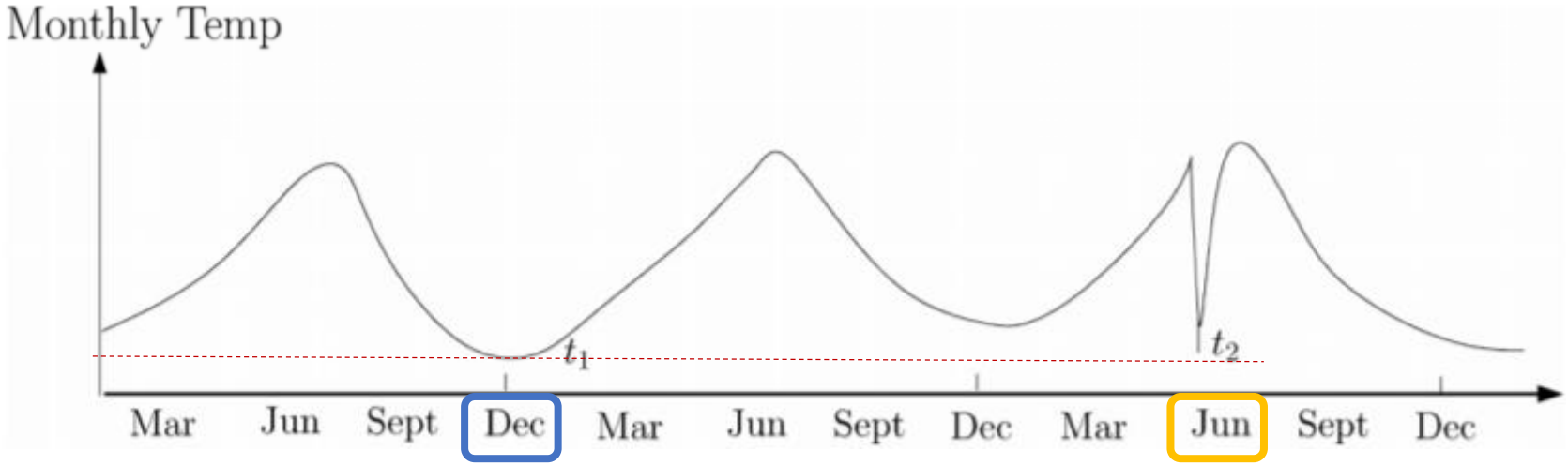
- Виды аномалий
  - Глобальная (точечная) – объект набора данных, имеющий существенно девиантное поведение
  - Контекстная – объект набора данных, имеющий существенно девиантное поведение при выполнении некоторых условий
  - Коллективная – подмножество объектов набора данных, имеющее существенно девиантное поведение (при этом каждый объект набора не обязательно является аномалией)
- Набор данных может иметь аномалии нескольких типов
- Аномальный объект может принадлежать к аномалиям нескольких типов

Глобальные (точечные) аномалии



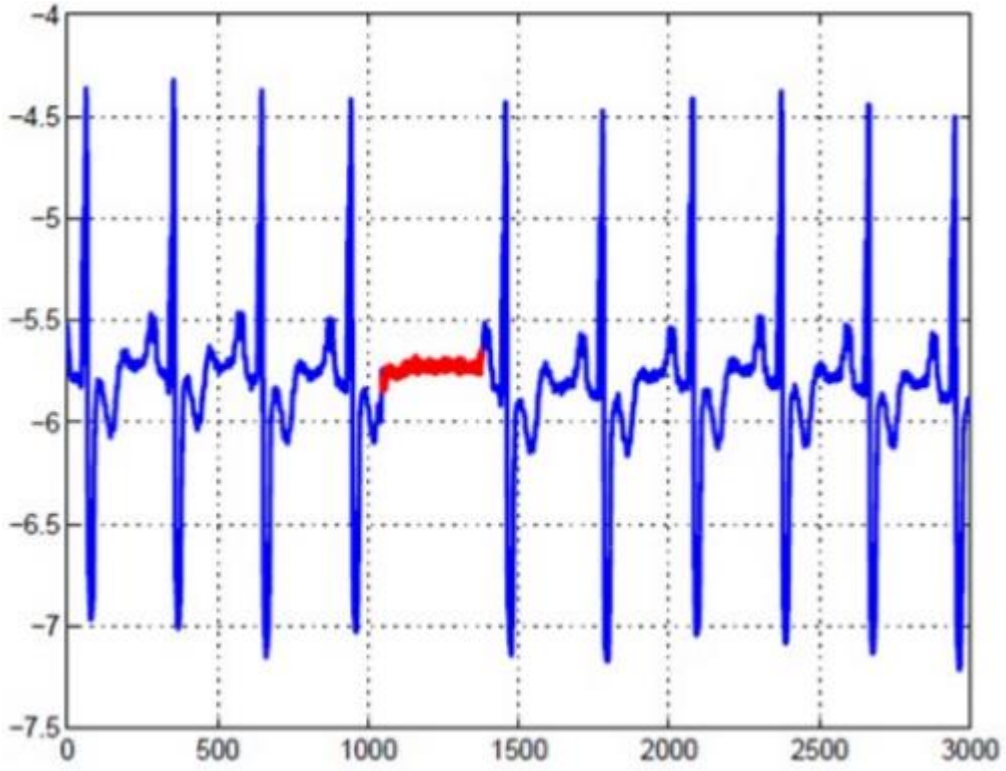
- Проблема подбора меры схожести

Контекстные аномалии

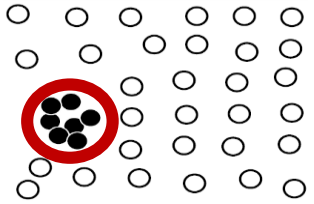


- Одно и то же значение температуры воздуха является аномалией в зависимости от времени года
- Контекстные и поведенческие атрибуты (время, локация и температура, влажность, давление)
- Проблема формального определения контекста

Коллективные аномалии



Дата	Покупка	Сумма
01.09.20	Кофе, сэндвич	100
02.09.20	Кофе, сэндвич	100
03.09.20	Кофе, сэндвич	100
03.09.20	Кофе, сэндвич	100
03.09.20	Кофе, сэндвич	100
03.09.20	Кофе, сэндвич	100
03.09.20	Кофе, сэндвич	100
04.09.20	Кофе, сэндвич	100



- Проблема подбора меры схожести

Классификация методов поиска аномалий

- Использование обучающей выборки «норма» *vs.* «аномалия»
  - Обучение с учителем (supervised)
  - Обучение без учителя (unsupervised)
  - Гибридный подход (semi-supervised)
- Предположение о норме и аномалии
  - Статистические (statistical-based)
  - Близостные (proximity-based)
  - Кластерные (clustering-based)

Поиск аномалий на основе обучения с учителем

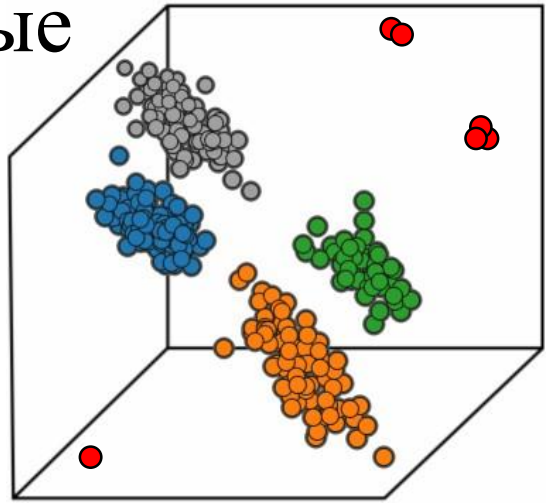
- Поиск аномалий – бинарная классификация «норма» *vs.* «аномалия»
  - Необходима обучающая выборка
  - Обучение модели
    - Аномалия то, что не норма
    - Норма то, что не аномалия
- Проблемы
  - Дисбаланс классов (аномалии редки)
  - Полнота более важна, чем точность

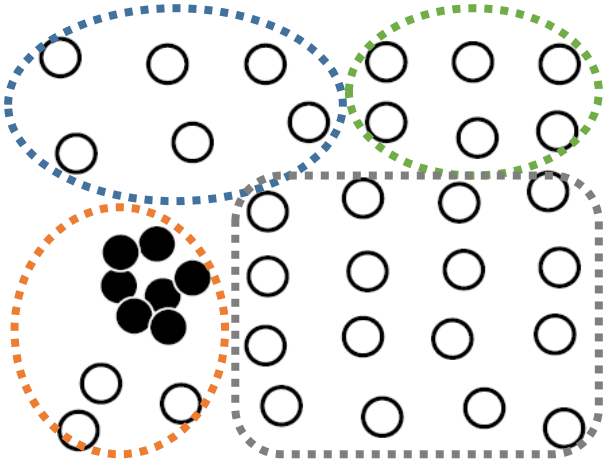
$$Precision = \frac{TP}{TP + \boldsymbol{FP}}$$

$$\boldsymbol{Recall} = \frac{TP}{TP + \boldsymbol{FN}}$$

Поиск аномалий на основе обучения без учителя

- Аномалии – существенно удаленные маломощные кластеры/синглтоны
- Проблемы
  - Слабо находит коллективные аномалии
  - Низкая точность (много FP)
  - Как отличить аномалии от шума?
  - Низкая производительность (нужно обрабатывать *все* нормальные данные)





Поиск аномалий на основе гибридного обучения

- Если в выборке только объекты «норма»
  - использовать объекты выборки и ближайшие к ним для обучения модели отличать норму
  - объекты, не соответствующие модели, являются аномалиями
- Если объектов «аномалия» существенно меньше, чем «норма»
  - использовать модели классификации объектов «норма», построенных методами обучения без учителя

Статистические методы

- Параметрические методы
  - Предполагаем, что объекты «норма» генерируются случайным процессом с параметрическим распределением с параметром θ
  - Функция плотности вероятности $f\ (x, \theta)$ дает вероятность того, что объект x сгенерирован распределением; $f\ (x, \theta) \to 0 \Leftrightarrow x$ – аномалия
- Непараметрические методы
  - Предполагаем отсутствие априорной статистической модели, определяем модель по входным данным
  - Могут использоваться параметры, но их количество и смысл заранее не известны

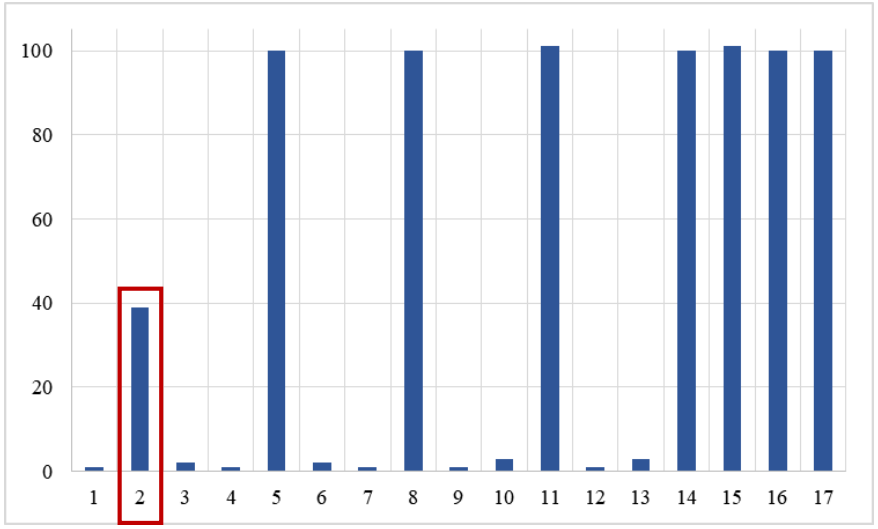
Статистические методы: z-значимость

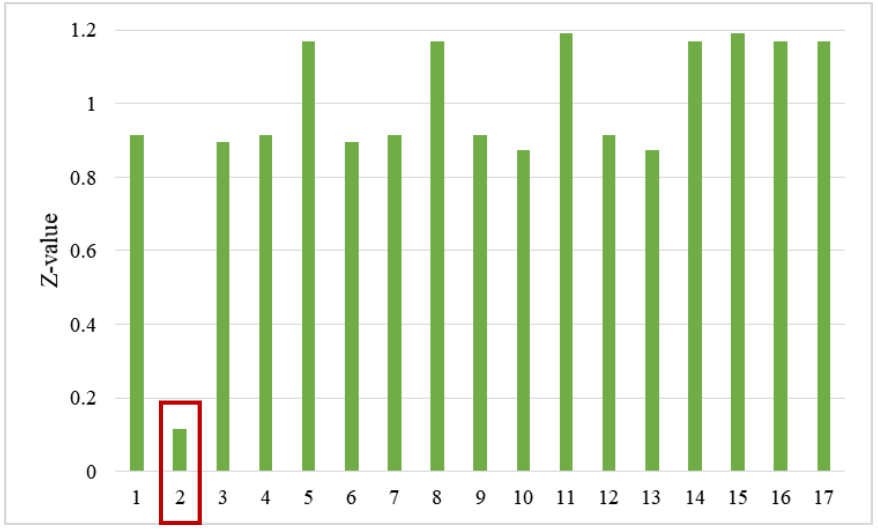
- Числовой атрибут $x = \{x_1, \dots, x_n\}$

$$Z_i = \frac{|x_i - \mu|}{\sigma}, \mu = \frac{1}{n}\sum_{i=1}^{n} x_i, \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \mu^2}$$

- Пример

1, **39**, 2, 1, 101, 2, 1, 100, 1, 3, 101, 1, 3, 100, 101, 100, 100





Статистические методы: правило трех сигм

- Числовой атрибут $x = \{x_1, \dots, x_n\}$ (предположительно *нормально распределенный*)

$$x_i \text{ – выброс, если } x_i \notin [\mu - k\sigma, \mu + k\sigma]$$

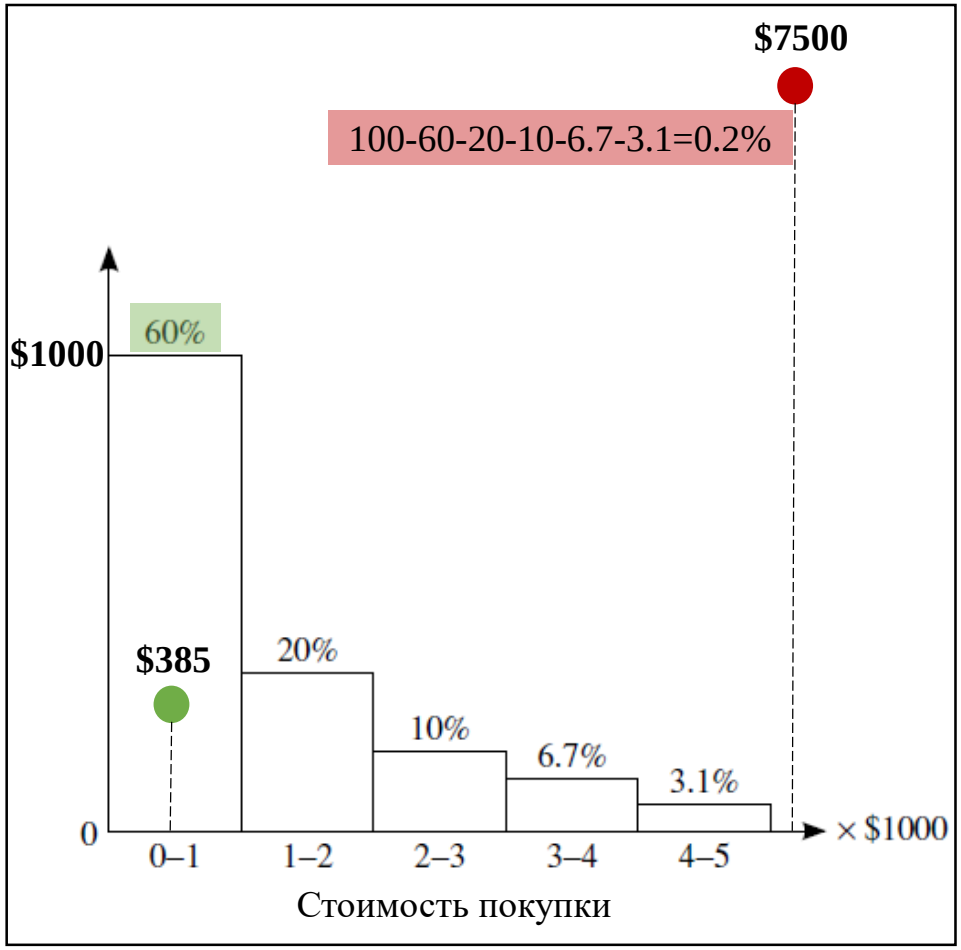
$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i,\ \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \mu^2}$$

типичное значение $k = 3$

(можно увеличить при большом количестве найденных выбросов)

- Пример
  - Не найден выброс:
    1, **39**, 2, 1, 101, 2, 1, 100, 1, 3, 101, 1, 3, 100, 101, 100, 100
  - Найден выброс:
    91, **39**, 92, 91, 101, 92, 81, 100, 101, 103, 101, 91, 93, 100, 101, 100, 100

Статистические методы: гистограммы



- Меры аномальности:
  - «Объем» столбца
    - 60% и 0.2%
  - Инверсия «объема» столбца
    - $\frac{1}{60\%} = 1.67$ и $\frac{1}{0.2\%} = 500$
- Проблема подбора ширины столбца
  - слишком маленькие: объекты «норма» в пустых/редких столбцах, FP
  - слишком большие: объекты «аномалия» в некоторых частых столбцах, FN

За и против статистических методов

- Достоинства
  - Строгая математическая основа
  - Эффективность
  - Хорошие результаты, если распределение данных заранее известно
- Недостатки
  - Во многих случаях распределение данных заранее неизвестно
  - Для данных большой размерности трудно оценить истинное распределение данных

Близостные методы

- Основная идея
  - объекты, находящиеся далеко от многих других объектов набора данных, являются аномалиями
  - близость аномалии значительно отличается от близости большинства других в наборе данных
- Реализация
  - на основе расстояния: объект аномален, если среди его соседей существенно мало других объектов
  - на основе плотности: объект аномален, если в его окрестности существенно мало других объектов

Поиск аномалий на основе расстояния

- ***Outliers*** are the data points for which there are fewer than p other data points within distance d

  Knorr E., Ng N. Finding intensional knowledge of distance-based outliers. VLDB 1999. pp. 211–222

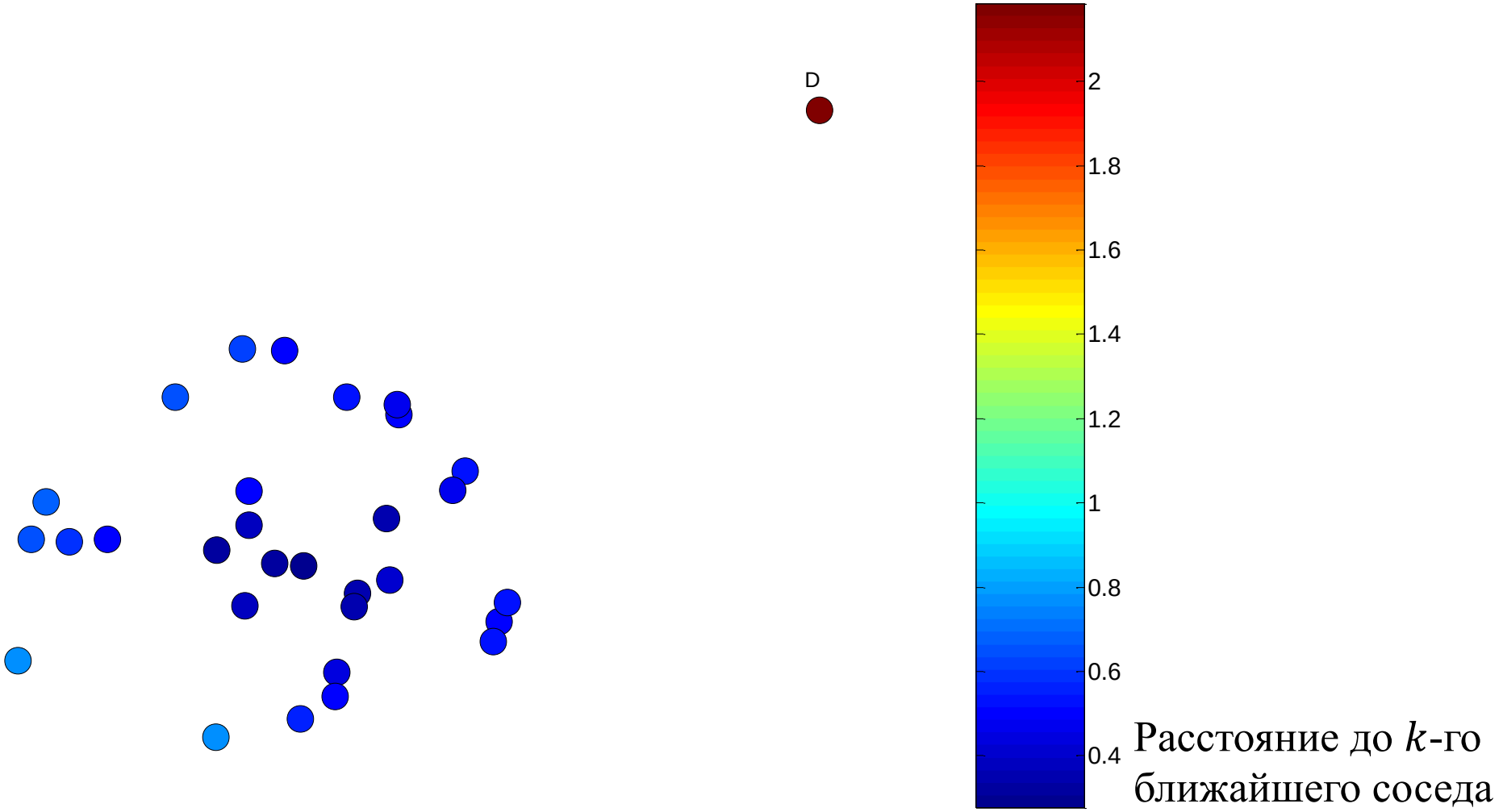
- ***Outliers*** are the top n data points whose distance to their k-th nearest neighbor is greatest

  Ramaswamy S., *et al.* Efficient algorithms for mining outliers from large dataset. SIGMOD 2000. pp. 427–438

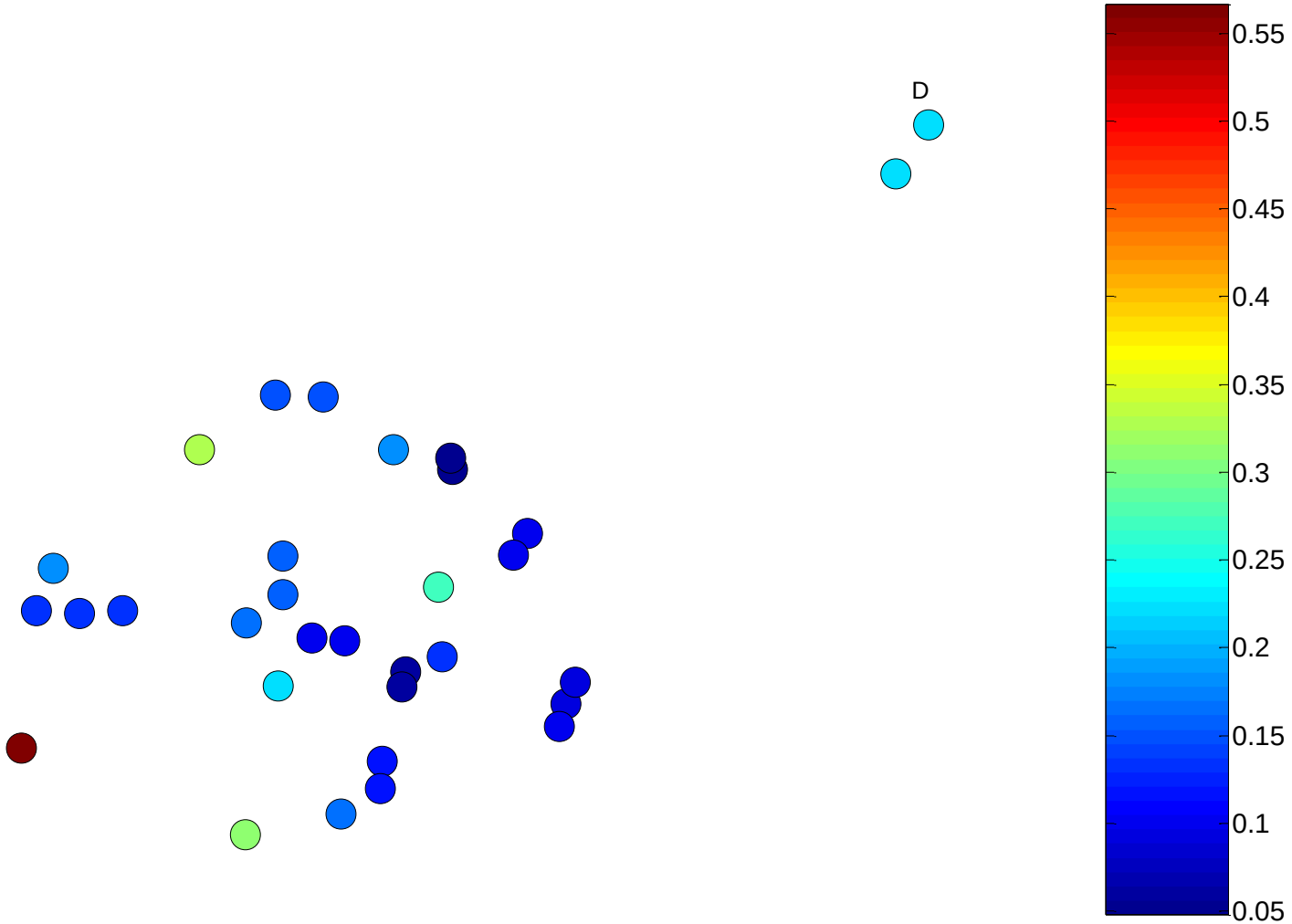
- ***Outliers*** are the top n data points whose average distance to their k nearest neighbors is greatest

  Angiulli F., Pizzuti C. Fast outlier detection in high dimensional spaces. PKDD 2002. pp. 15–26

Пример: 1-NN, одна аномалия

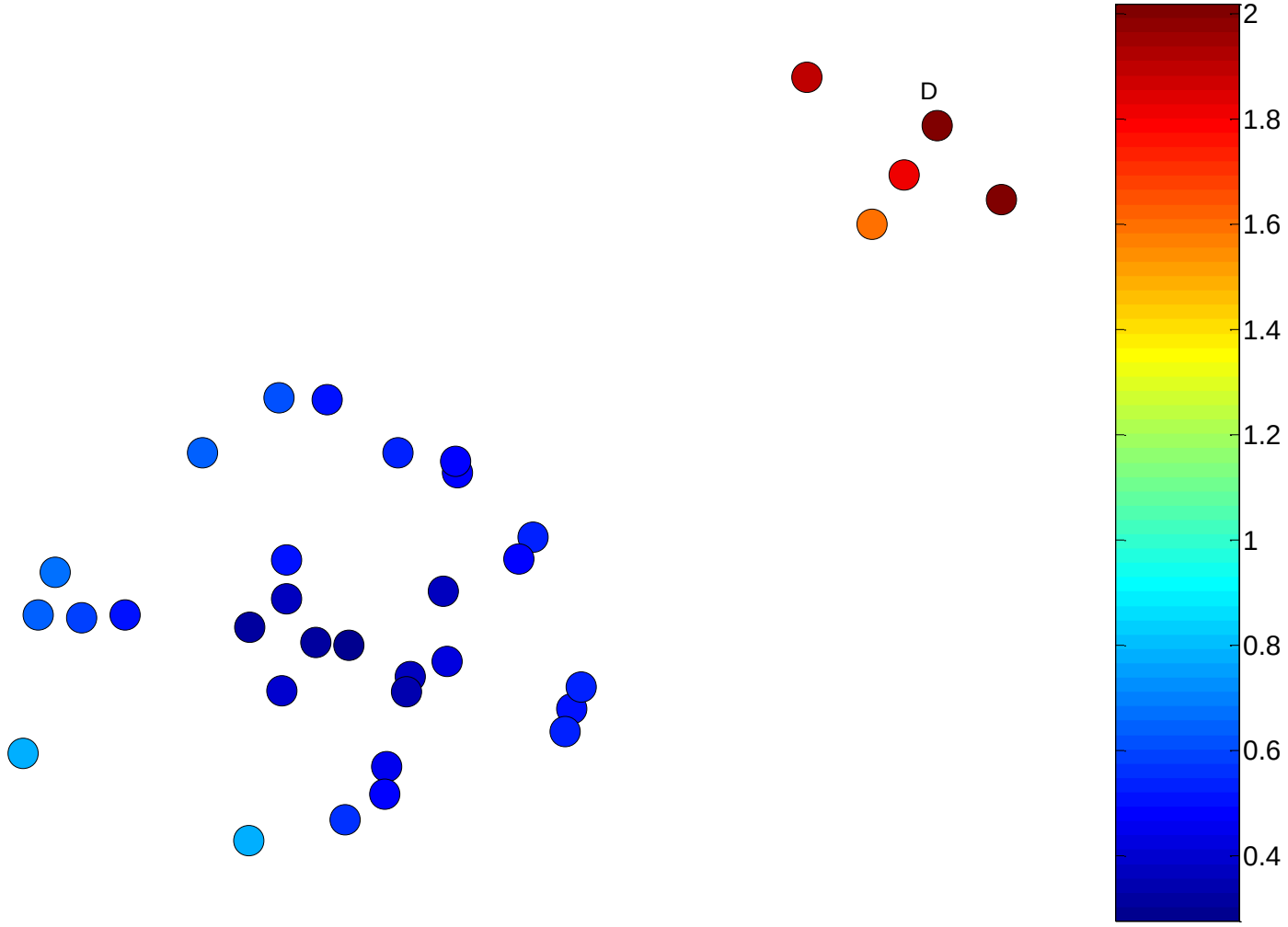


Пример: 1-NN, две аномалии



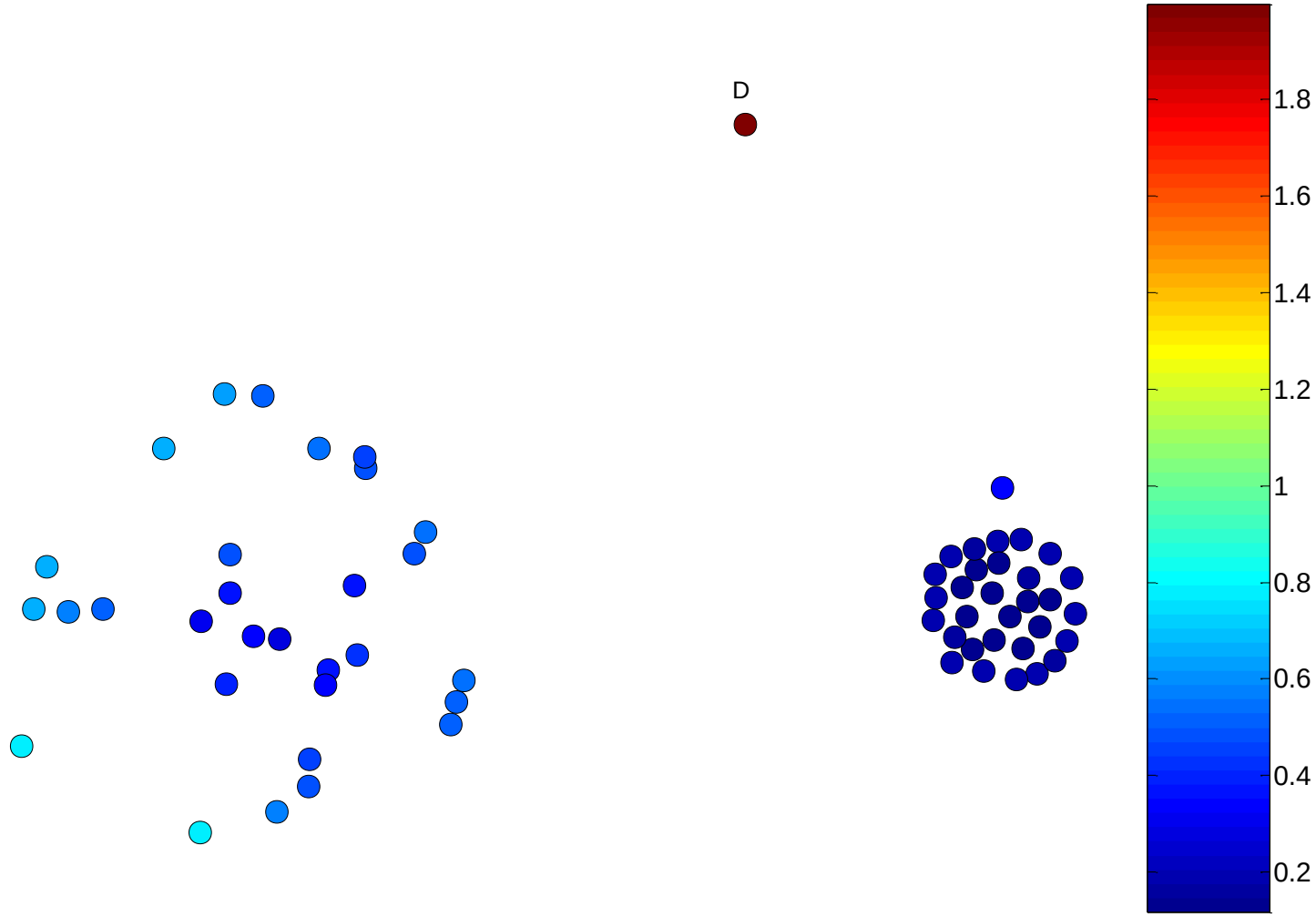
Расстояние до k-го ближайшего соседа

Пример: 5-NN, аномалия – небольшой кластер



Расстояние до k-го ближайшего соседа

Пример: 5-NN, одна аномалия



Расстояние до k-го ближайшего соседа

За и против методов на основе расстояния

- Достоинства
  - Простота идеи и ее реализации
- Недостатки
  - Сложность $O(n^2)$, где n – количество объектов
  - Чувствительность к параметрам
  - Чувствительность к плотности объектов
  - При увеличении размерности пространства расстояние становится менее значимым

Поиск аномалий на основе плотности: метод вложенных циклов

- Множество объектов: $D = \{o \mid o \in \mathbb{R}^d\}$
- Параметры: $dist(\cdot,\cdot), r \geq 0,\ 0 < \alpha \leq 1$
- Плотность: $DB(dist, r, \alpha, o) = \frac{|\{x \in D | dist(x,o) \leq r\}|}{|D|}$
- o – выброс $\Leftrightarrow DB(dist, r, \alpha, o) \leq \alpha$

Метод вложенных циклов

$A := \emptyset$

for all $x \in D$

  $count := 0$

  for all $y \in D$

    if $x \neq y$ **and** $dist(x, y) \leq r$ **then**

      $count := count + 1$

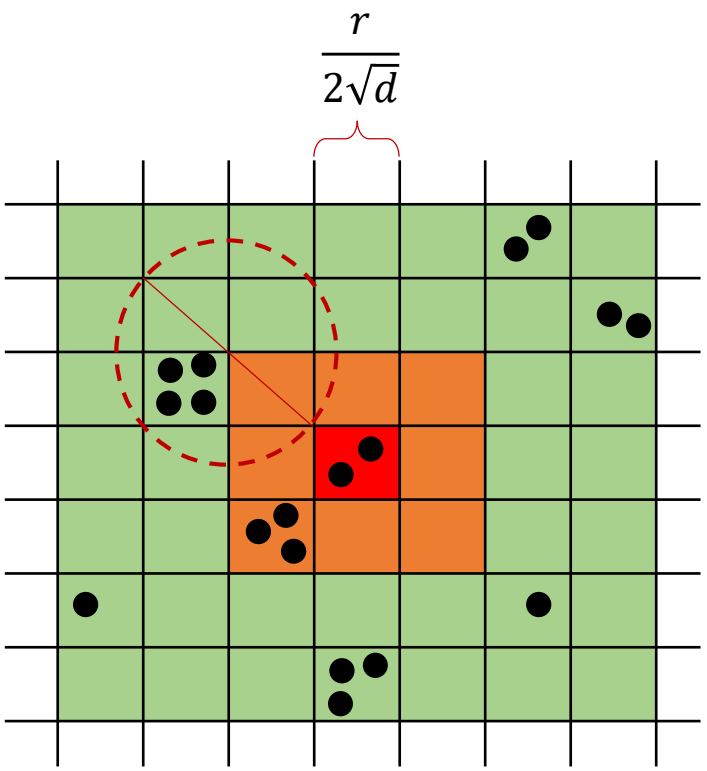
      if $count \geq \lceil \alpha \cdot |D| \rceil$ **then**

        break

  $A := A \cup x$

Поиск аномалий на основе плотности: метод решеток

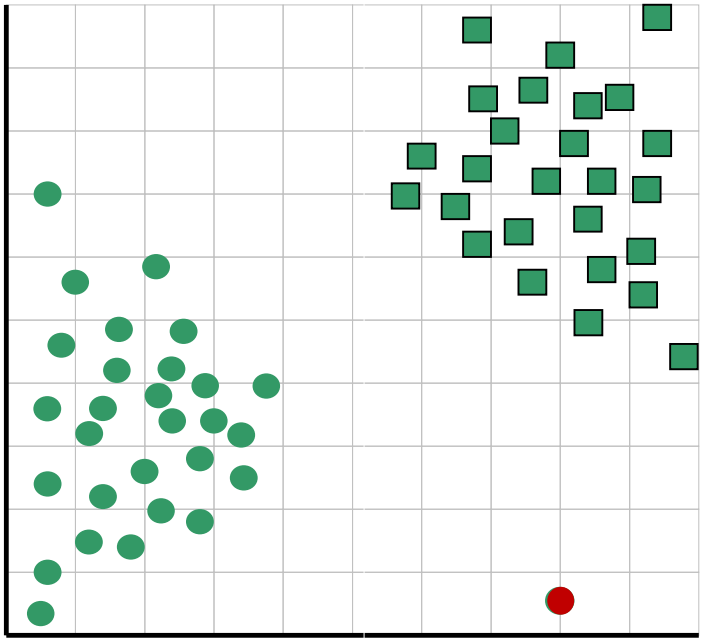
- Свойства
  - $\forall x, y\ dist(x, y) \leq r$
  - $\forall x, y\ dist(x, y) \geq r$
- Отбрасывание
  - если $\boldsymbol{n} + \boldsymbol{n} > \lceil \alpha \cdot |D| \rceil$, то x – не выброс
  - если $\boldsymbol{n} + \boldsymbol{n} + \boldsymbol{n} > \lceil \alpha \cdot |D| \rceil + 1$, то x – выброс

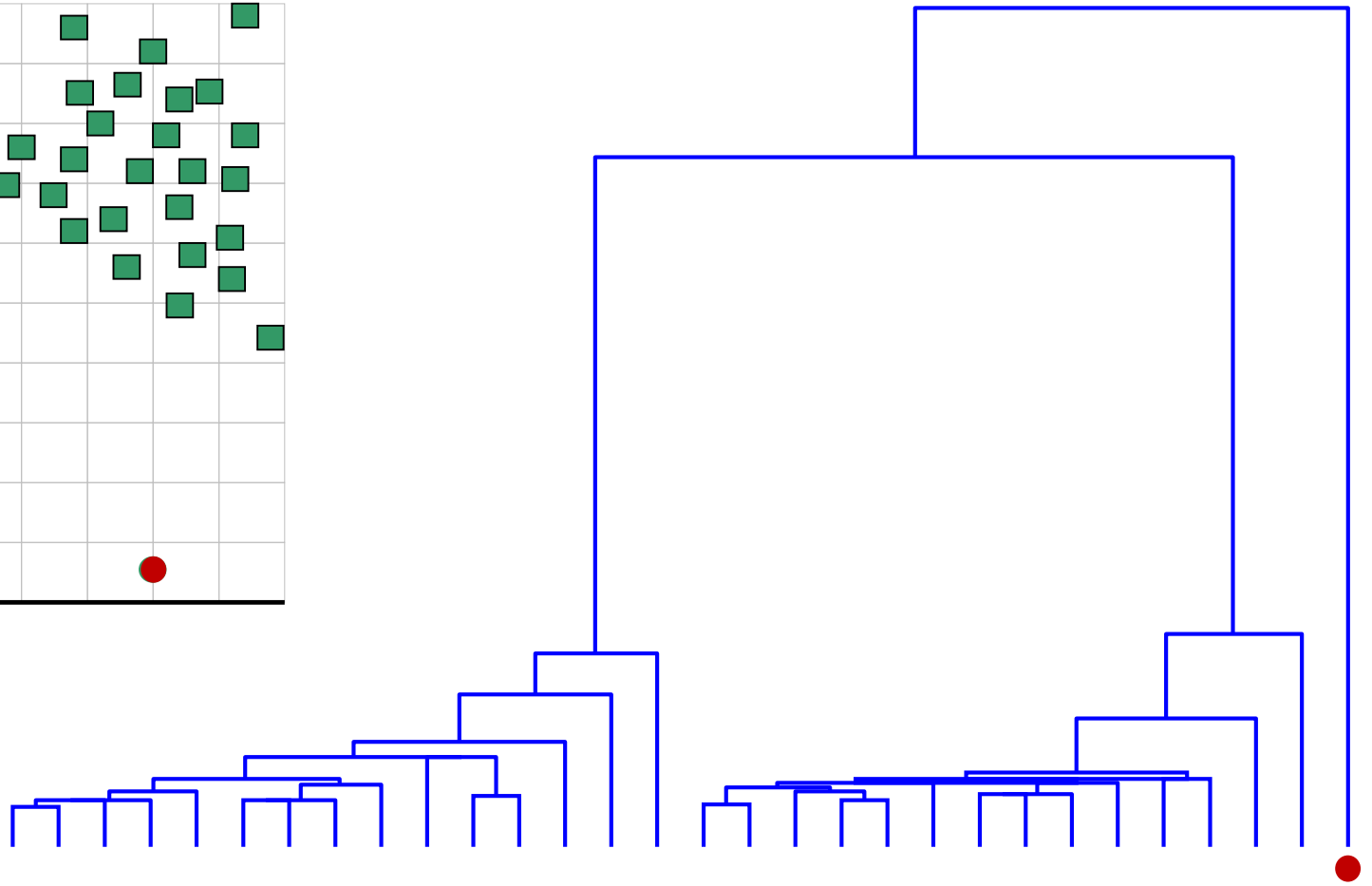


За и против методов на основе плотности

- Достоинства
  - Простота идеи и ее реализации
- Недостатки
  - Сложность $O(n^2)$, где n – количество объектов
  - Чувствительность к параметрам
  - При увеличении размерности пространства плотность становится менее значимой

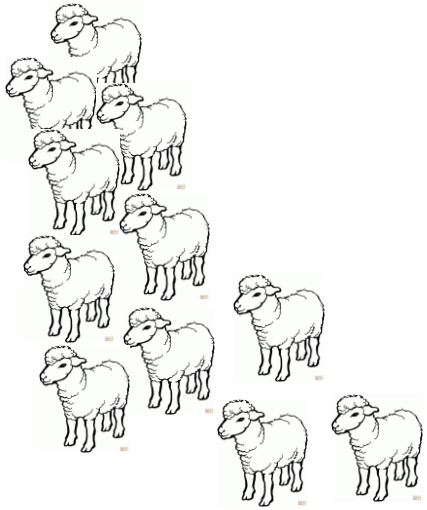
Поиск аномалий: иерархическая кластеризация

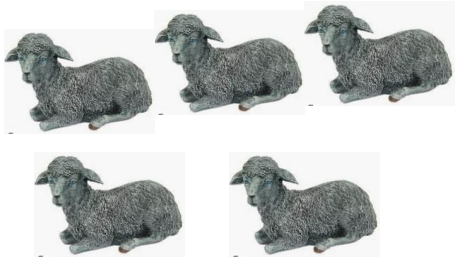


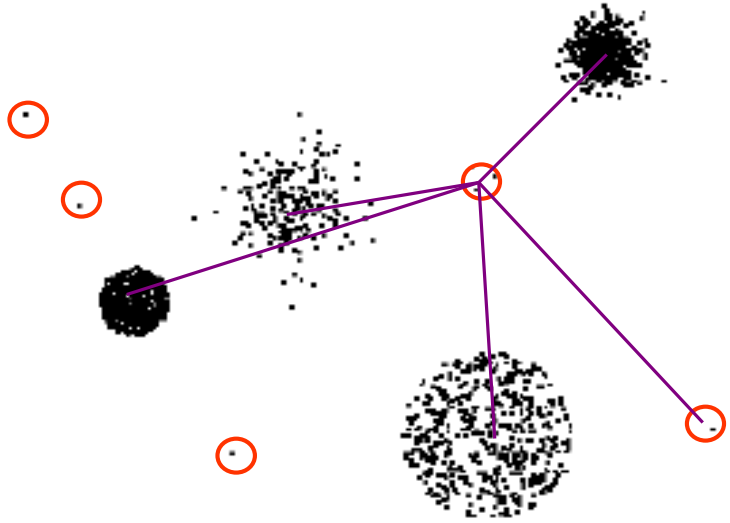


Поиск аномалий: плотностная/разделительная кластеризация

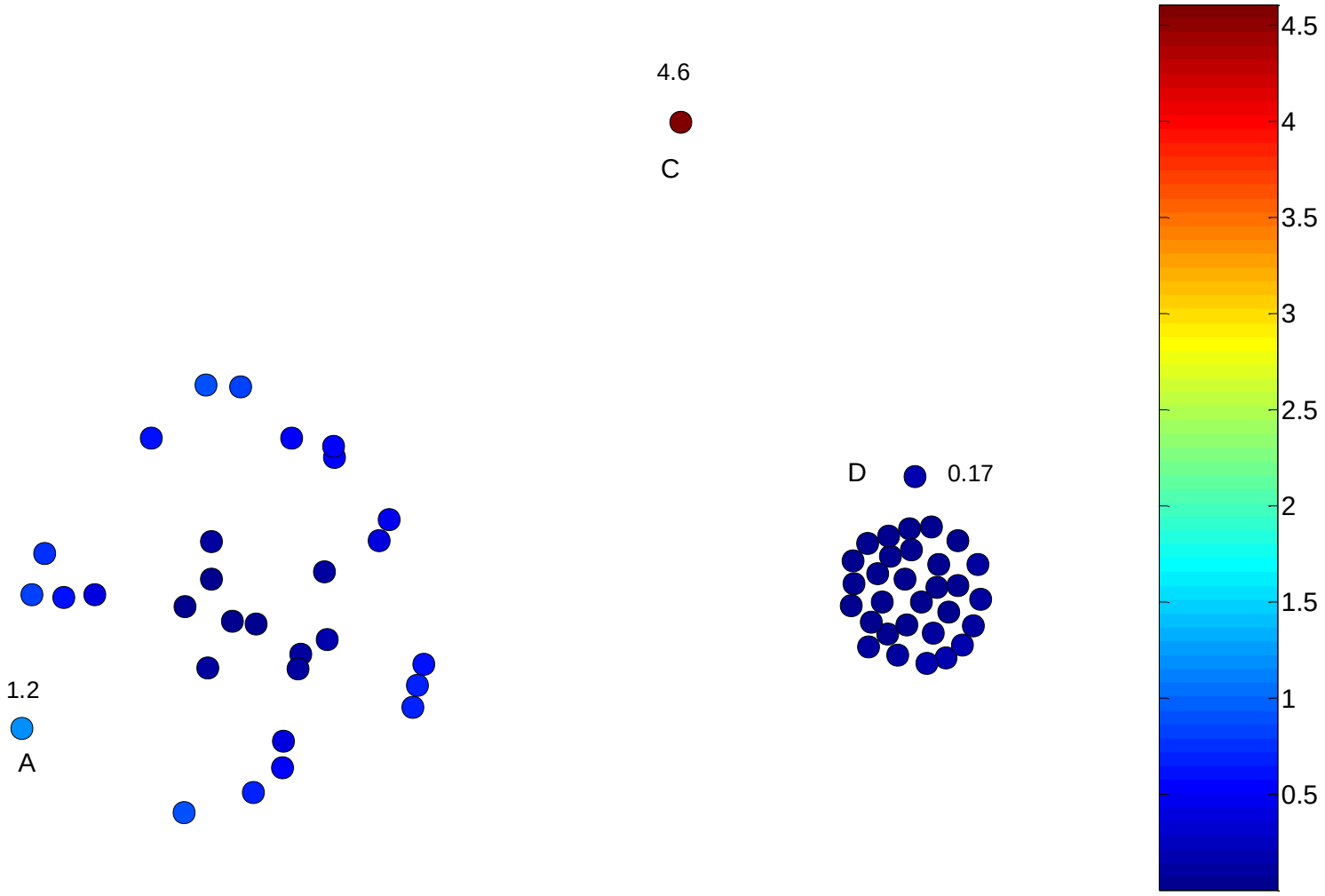
- Аномалия – объект, не принадлежащий строго ни одному из кластеров
  - не является близким ни одному из центроидов
  - имеет низкую плотность
  - …





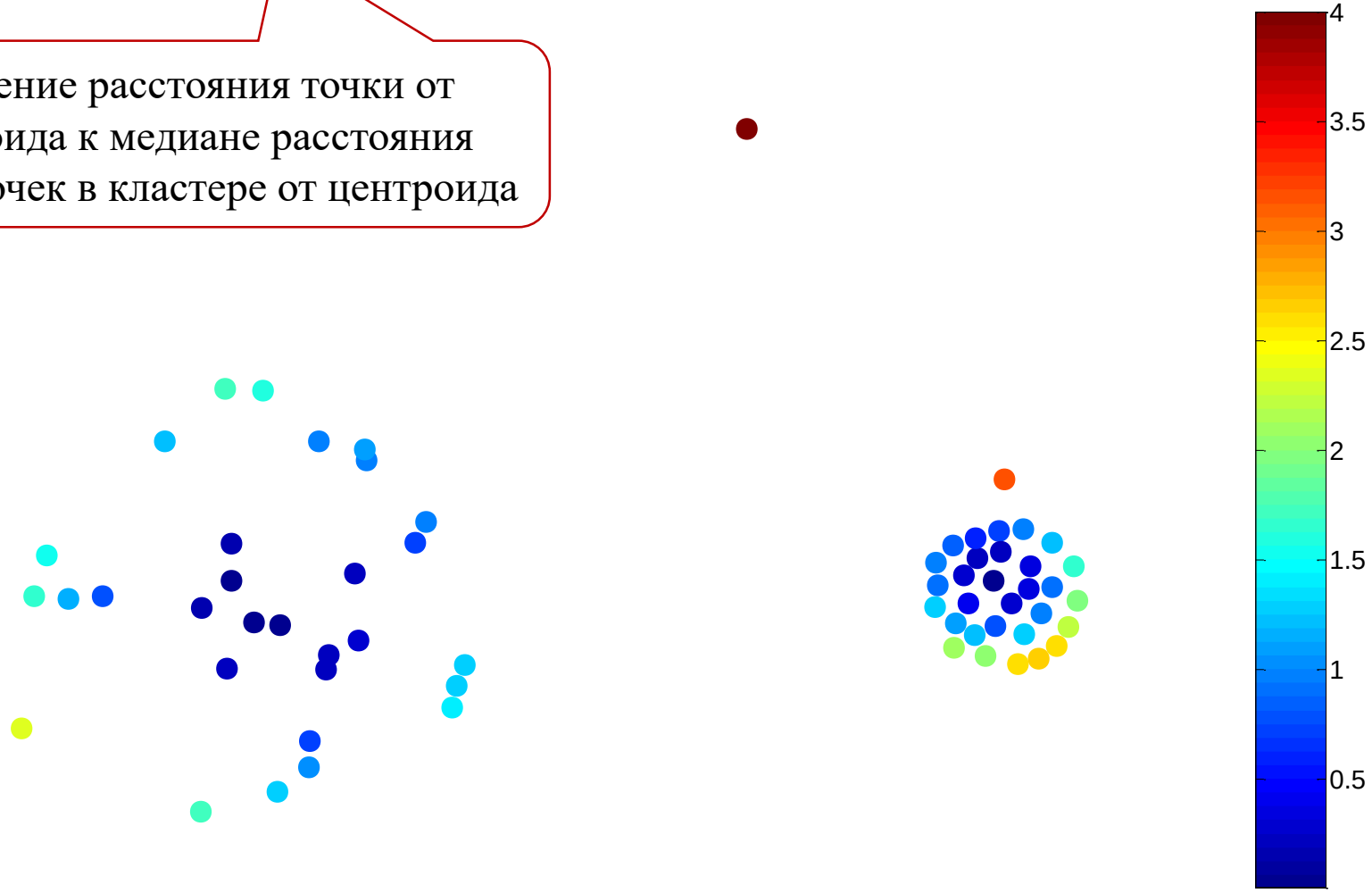


Пример: аномалии на основе расстояния до ближайшего центроида



Пример: аномалии на основе относительного расстояния до ближайшего центроида

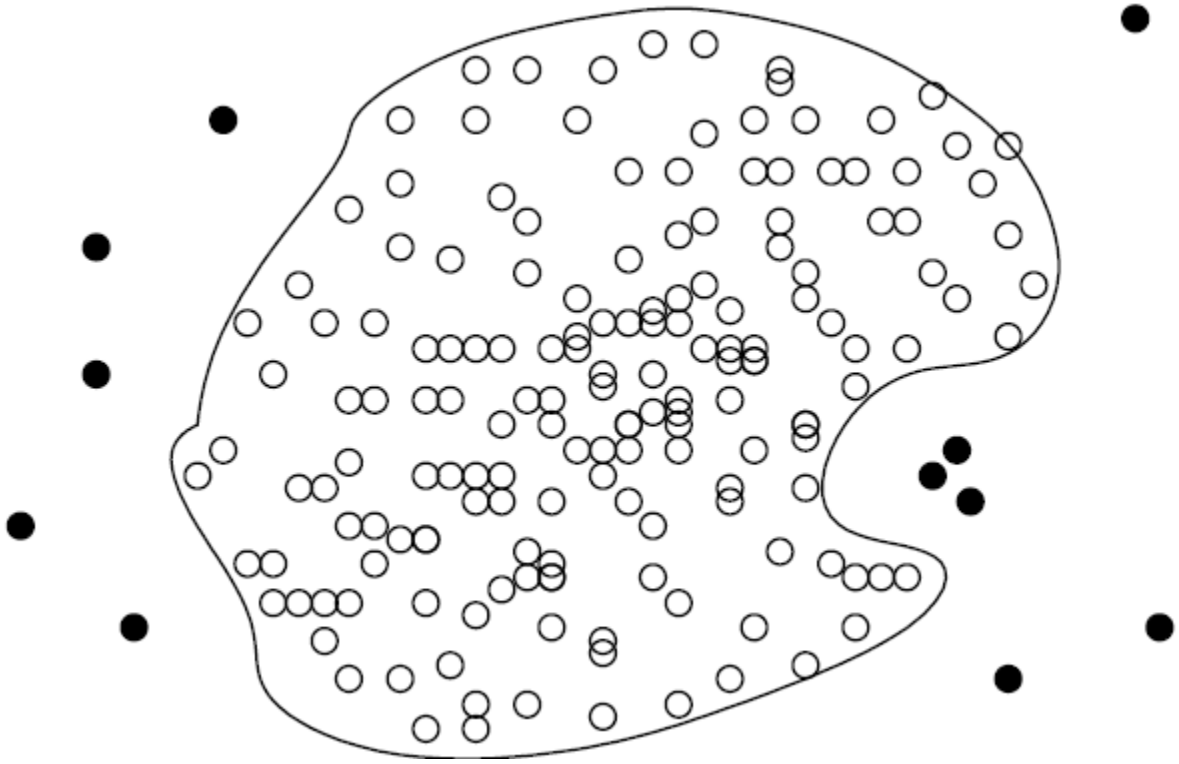
отношение расстояния точки от центроида к медиане расстояния всех точек в кластере от центроида



За и против методов на основе кластеризации

- Достоинства
  - Простота идеи
  - Широкий спектр алгоритмов кластеризации
- Недостатки
  - Трудность выбора одного алгоритма кластеризации
  - Трудность выбора количества кластеров
  - При увеличении размерности пространства плотность становится менее значимой
  - Аномалии могут искажать кластеры

Поиск аномалий на основе классификации: метод OneClass SVM

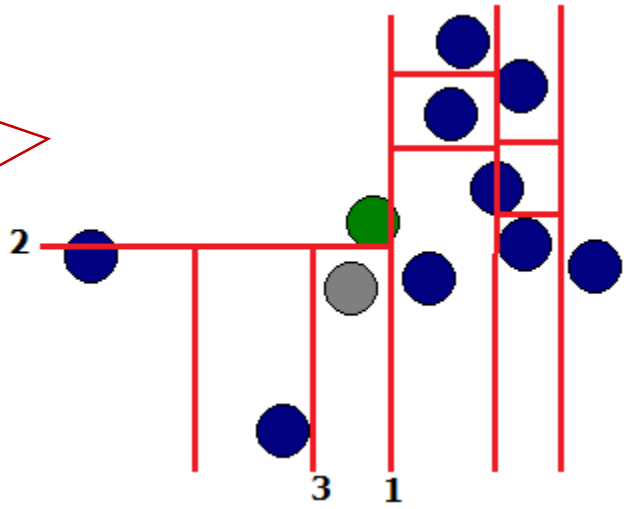


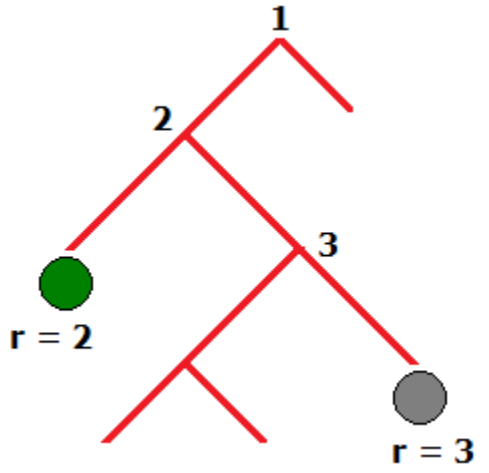
- При обучении выделить область, в которой расположены данные
- Для объекта тестовой выборки определять, попадает ли он в область

Поиск аномалий на основе классификации: метод изолирующего леса (Isolation Forest)

- Лес состоит из деревьев решений
- Каждое дерево строится до исчерпания выборки
- При построении дерева выбирается случайный атрибут и случайное значение для расщепления
- Для объекта определяется мера нормальности: среднее значение глубин листьев, в которые он попал

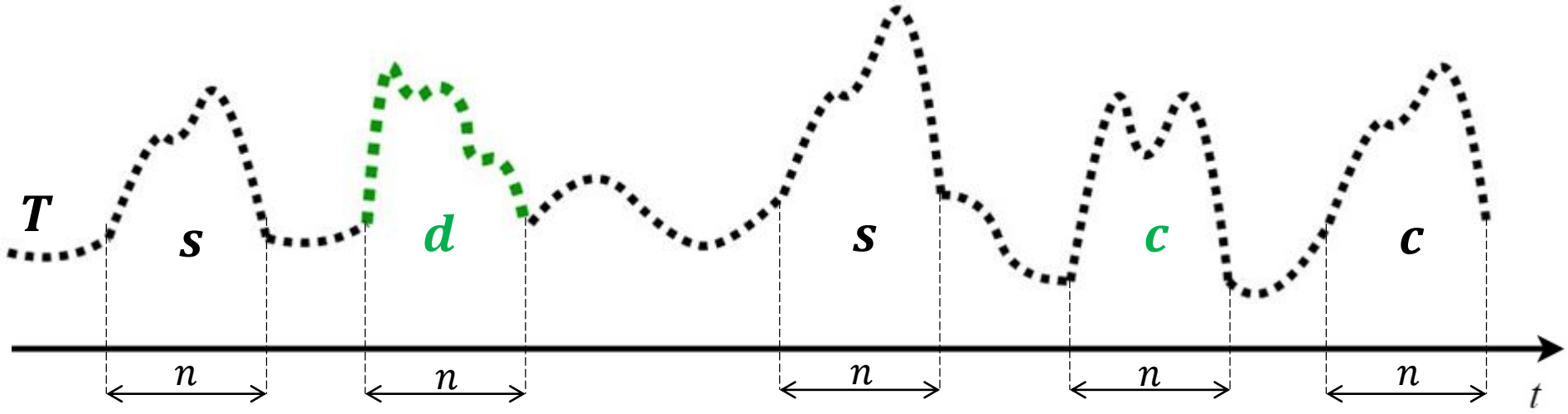
Выбросы будут попадать в листья на ранних этапах (на небольшой глубине дерева), их будет проще «изолировать»





Аномалии во временных рядах: диссонанс (discord)

- Подпоследовательность ряда, расстояние от которой до ее ближайшего соседа максимально
- Дано: ряд $\boldsymbol{T}$, длина диссонанса n
- Найти: $\boldsymbol{d} = \arg\max_{\boldsymbol{s}\in\boldsymbol{T}} \left(\min_{\boldsymbol{c}\in\boldsymbol{T},\ \boldsymbol{s}\cap\boldsymbol{c}=\emptyset} \mathrm{ED}(\boldsymbol{c}, \boldsymbol{s}) \right)$



Z-нормализация

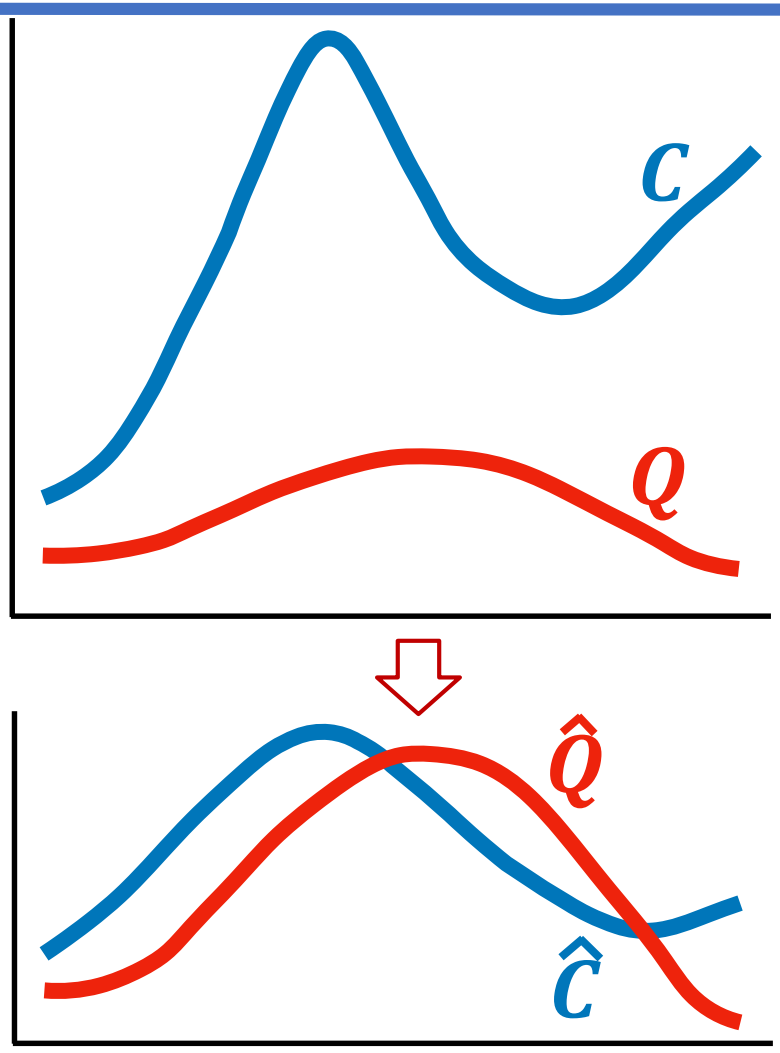
- Для сравнения рядов с различной амплитудой
  - среднее арифметическое ряда ≈ 0,
  - среднеквадратичное отклонение ≈ 1

- $\hat{C} = (\hat{t}_1, \hat{t}_2, \dots, \hat{t}_n)$

$$\hat{t}_i = \frac{t_i - \mu}{\sigma}$$

$$\mu = \frac{1}{n}\sum_{i=1}^{n} t_i$$

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \mu^2}$$

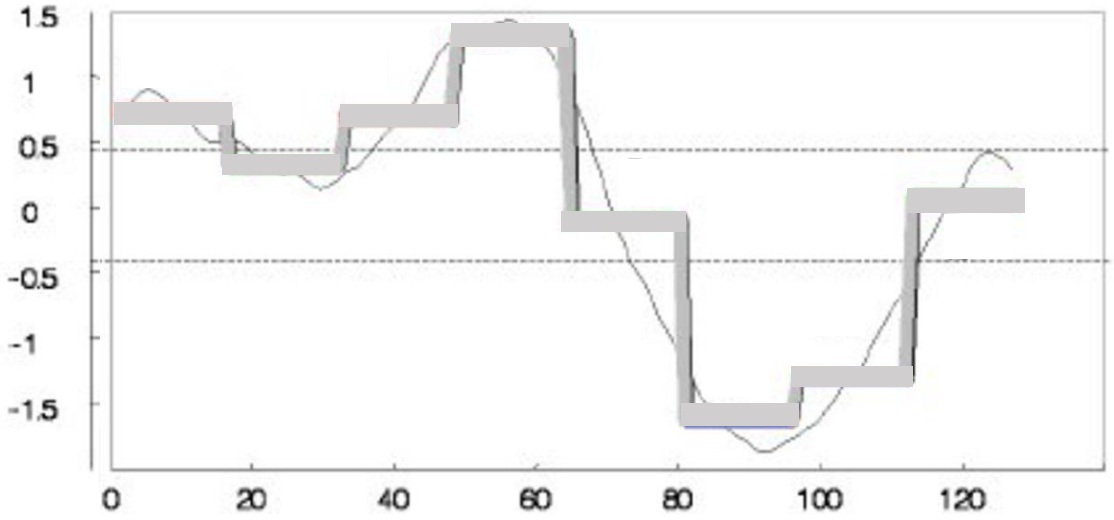


Кусочно-агрегатное приближение (PAA, Piecewise Aggregate Approximation)

- Подпоследовательность: $C = (c_1, \ldots, c_n)$
- Степень агрегации: $w < n$ (обычно $w \in \{3,4\}$)
- Кусочно-агрегатное представление: $\bar{C} = (\bar{c}_1, \ldots, \bar{c}_w)$

$$\bar{c}_i = \frac{w}{n} \cdot \sum_{j=\left(\frac{n}{w}\right)(i-1)+1}^{\left(\frac{n}{w}\right)i} c_j$$

$n = 128, w = 8$

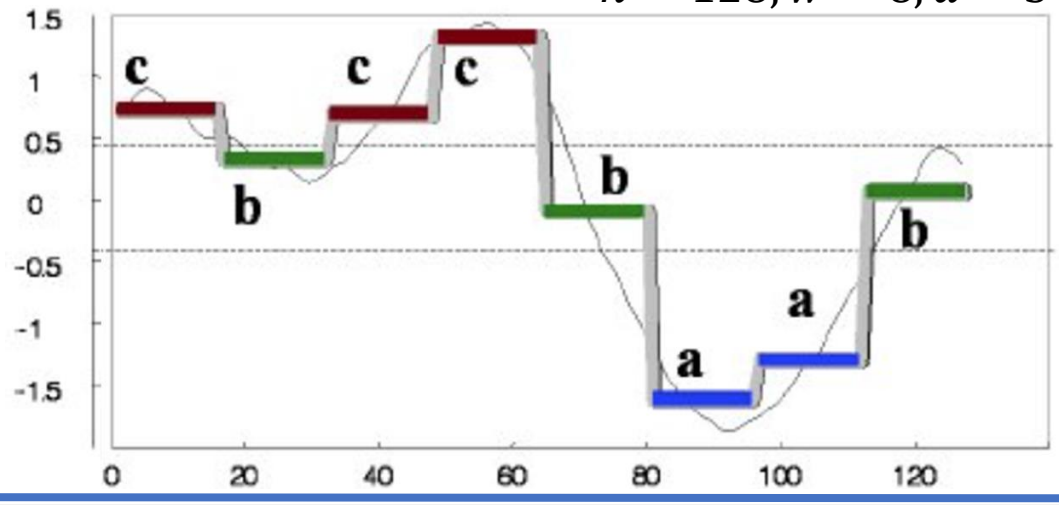


Символьно-агрегатное кодирование SAX, Symbolic Aggregate Approximation

- Символьное представление: $\hat{C} = (\hat{c}_1, \ldots, \hat{c}_w)$
- Алфавит кодирования: $\alpha_1, \ldots, \alpha_a$ ($a \leq w$, обычно $a \in \{3,4\}$)
- Точки разделения: $\beta_0 = -\infty, \beta_1, \ldots, \beta_{a-1}, \beta_a = +\infty$, площадь под кривой нормального распределения $N(0,1)$ между β_i и β_{i+1} равна $\frac{1}{a}$
- Кодирование: $\hat{c}_i = \alpha_i \Leftrightarrow \beta_{j-1} \leq \hat{c}_i < \beta_j$

β_i \ a	3	4	5
β_1	-0.43	-0.67	-0.84
β_2	0.43	0	-0.25
β_3		0.67	0.25
β_4			0.84

$n = 128, w = 8, a = 3$

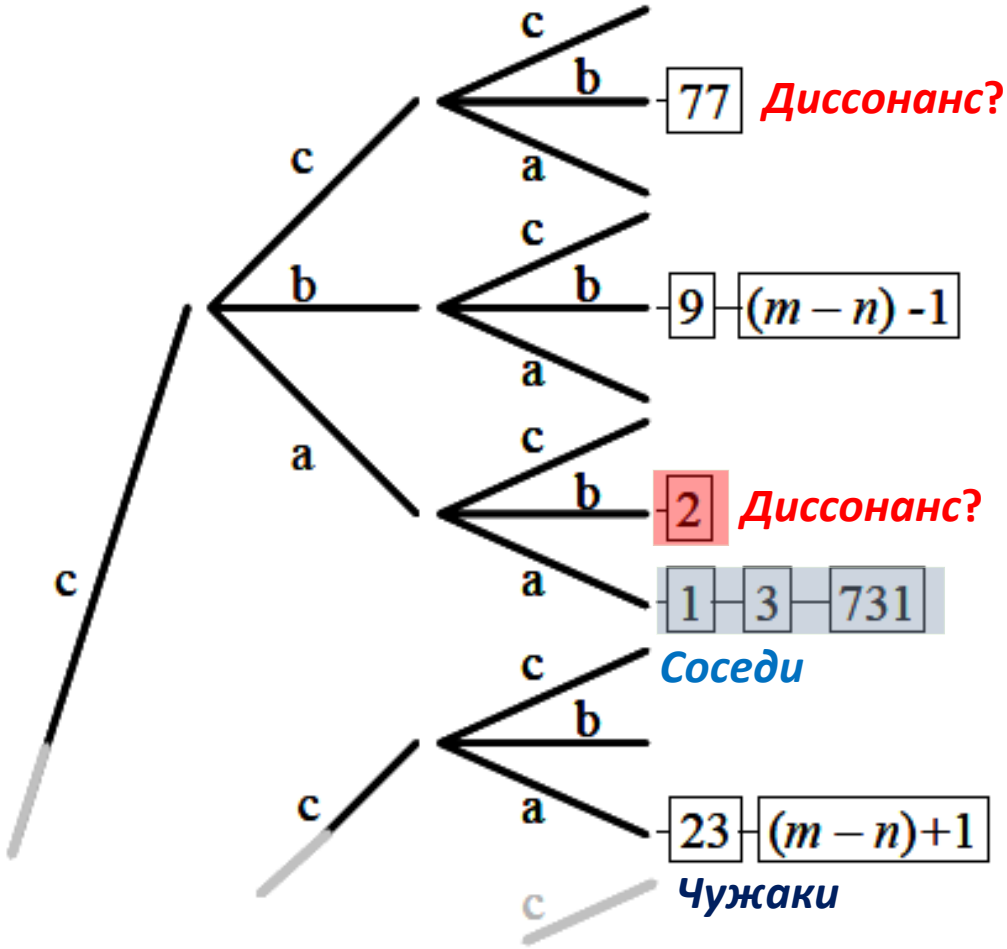


Алгоритм HOTSAX

Частотный индекс слов

1	c	a	a	3
2	c	a	b	1
3	c	a	a	3
::	::	::	::	::
::	::	::	::	::
$(m-n)$ -1	c	b	b	2
$(m-n)$	a	c	b	1
$(m-n)$ +1	b	c	a	2

Префиксное дерево



Алгоритм HOTSAX

$dist_{bsf} \leftarrow 0$; $dist_{min} \leftarrow \infty$

for $C_i \in$ ***Диссонансы?*** · ***Остальные***

  for $C_j \in$ ***Соседи***, ***Чужаки***

    $d \leftarrow \mathrm{ED}(C_i, C_j)$

    if $d < dist_{bsf}$

      break

    $dist_{min} \leftarrow \min(d, dist_{min})$

  $dist_{bsf} \leftarrow \max(dist_{min}, dist_{bsf})$

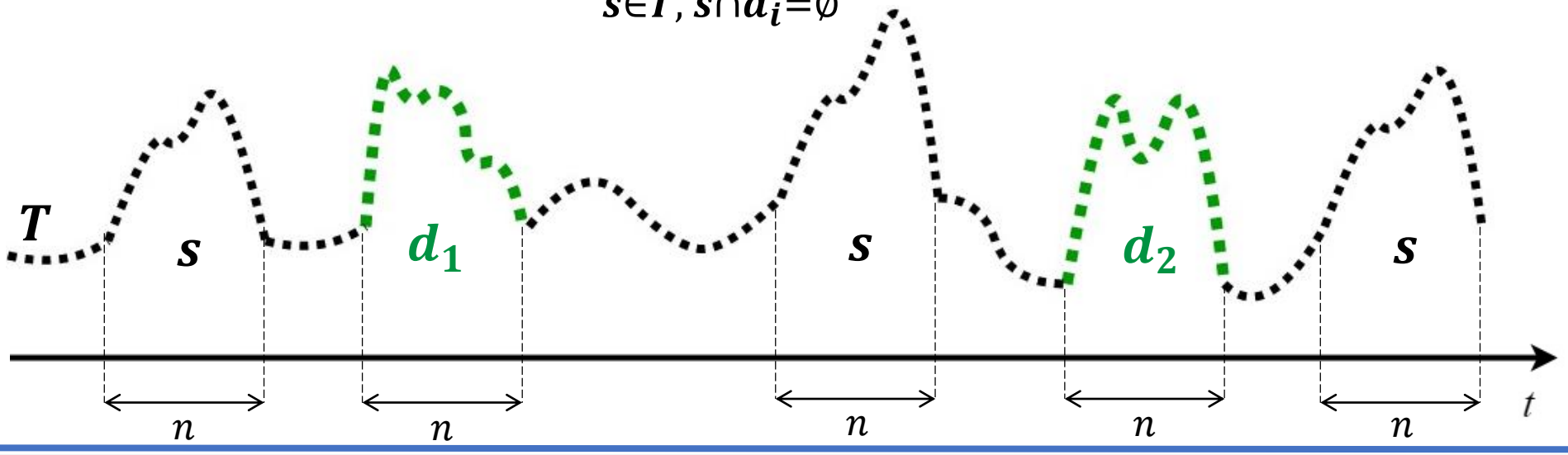
  $pos_{bsf} \leftarrow i$

return $\{pos_{bsf}, dist_{bsf}\}$

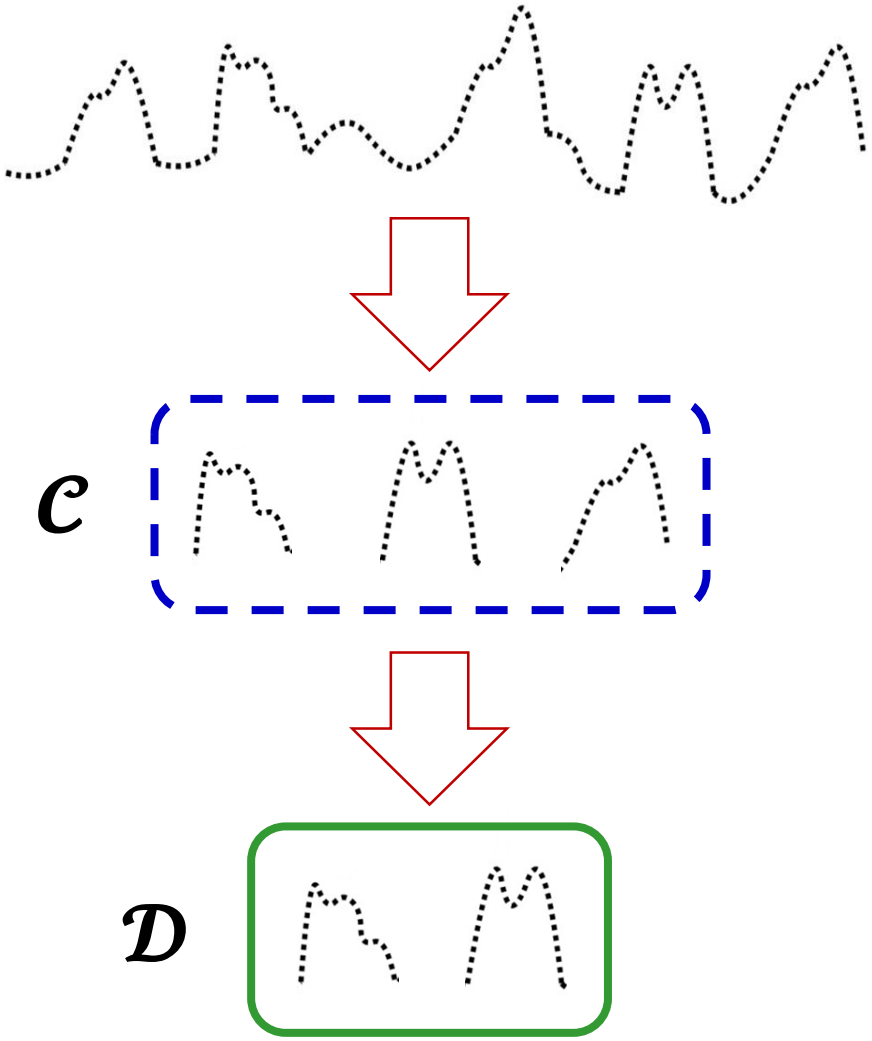
Диапазонный диссонанс (range discord)

- Подпоследовательность, расстояние от которой до ее ближайшего соседа не ниже заданного порога
- Дано: ряд $\boldsymbol{T}$, длина диссонанса n, порог r
- Найти: $\boldsymbol{\mathcal{D}} = \{\boldsymbol{d_1}, \boldsymbol{d_2}, \dots\}$

$$\boldsymbol{d_i} \in \boldsymbol{\mathcal{D}} \Leftrightarrow \min_{\boldsymbol{s} \in \boldsymbol{T},\, \boldsymbol{s} \cap \boldsymbol{d_i} = \emptyset} \mathrm{ED}(\boldsymbol{d_i}, \boldsymbol{s}) \geq r$$



Алгоритм DADD (Disk Aware Discord Discovery)



1. **Отбор**

   За одно сканирование ряда сформировать **множество кандидатов** в диссонансы

2. **Очистка**

   За одно сканирование ряда **отбросить бесперспективных кандидатов**

Отбор кандидатов (1)

Сканировать ряд T:

текущая подпоследовательность s

Кандидат := TRUE

for each $c_i \in \mathcal{C}$

if $\mathrm{ED}(s, c_i) < r$ **and** $s \cap c_i = \emptyset$ **then**

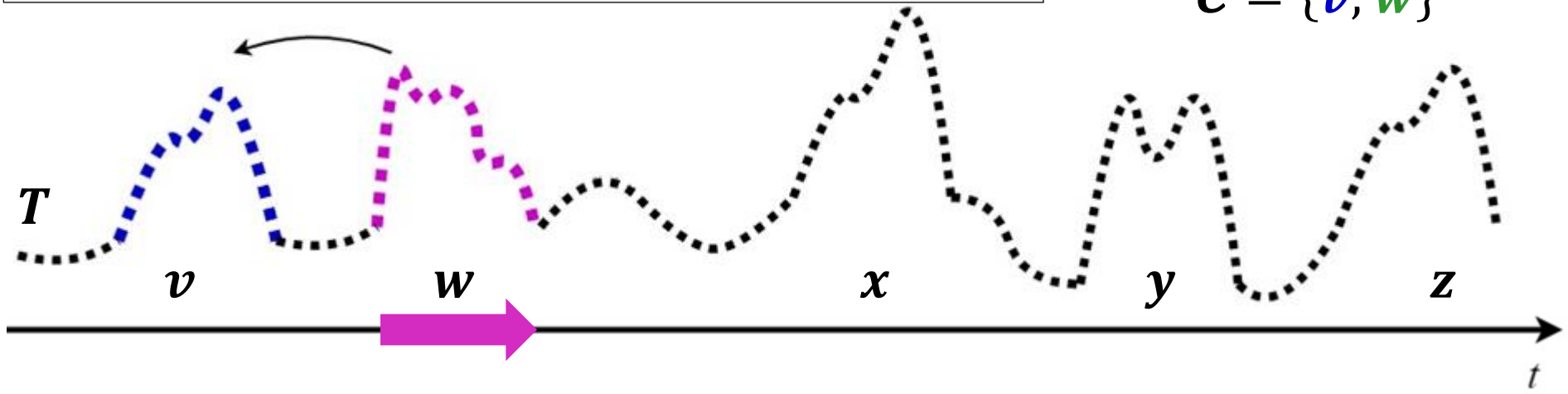
$\mathcal{C} := \mathcal{C} \setminus c_i$; Кандидат := FALSE

if Кандидат = TRUE **then** $\mathcal{C} := \mathcal{C} \cup s$

$\mathcal{C} = \{v\}$

$\mathrm{ED}(w, v) \geq r$

$\mathcal{C} = \{v, w\}$



Отбор кандидатов (2)

Сканировать ряд T:

текущая подпоследовательность s

Кандидат := TRUE

for each $c_i \in \mathcal{C}$

if $\mathrm{ED}(s, c_i) < r$ **and** $s \cap c_i = \emptyset$ **then**

$\mathcal{C} := \mathcal{C} \setminus c_i$; Кандидат := FALSE

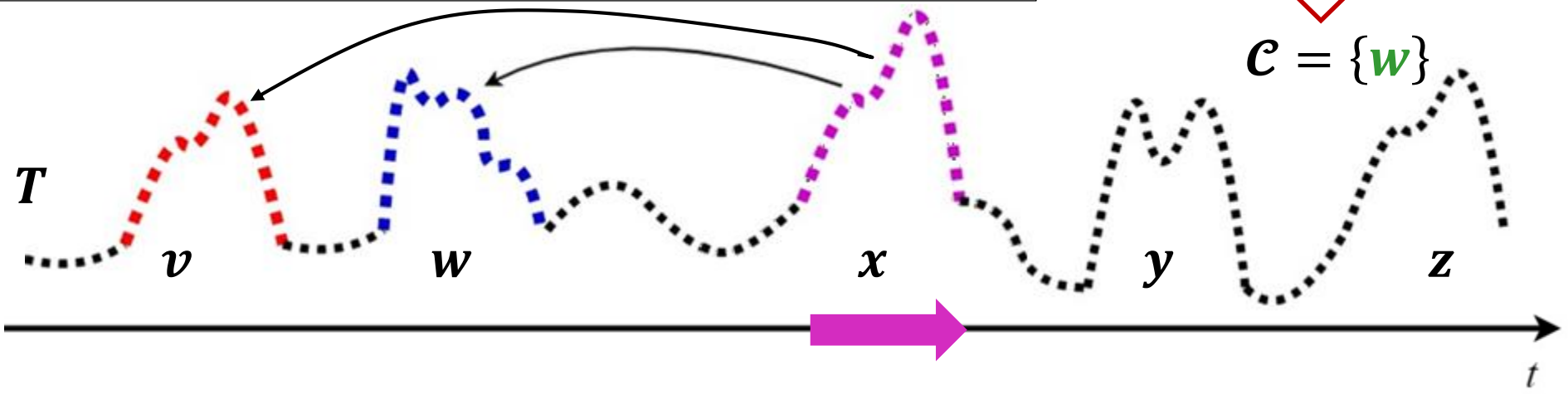
if Кандидат = TRUE **then** $\mathcal{C} := \mathcal{C} \cup s$

$\mathcal{C} = \{v, w\}$

$\mathrm{ED}(x, v) < r$

$\mathrm{ED}(x, w) \geq r$

$\mathcal{C} = \{w\}$



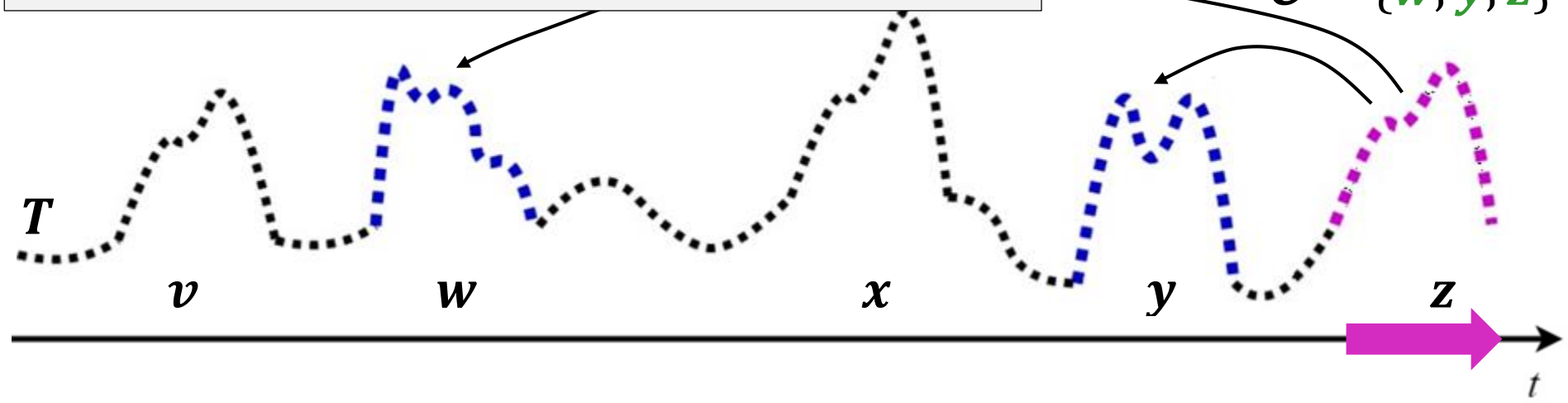
Отбор кандидатов (3)

Сканировать ряд T:
текущая подпоследовательность s
Кандидат := TRUE
for each $c_i \in \mathcal{C}$
if $\mathrm{ED}(s, c_i) < r$ **and** $s \cap c_i = \emptyset$ **then**
$\mathcal{C} := \mathcal{C} \setminus c_i$; Кандидат := FALSE
if Кандидат = TRUE **then** $\mathcal{C} := \mathcal{C} \cup s$

$\mathcal{C} = \{w, y\}$

$\mathrm{ED}(z, w) \geq r$
$\mathrm{ED}(z, y) \geq r$

$\mathcal{C} = \{w, y, z\}$



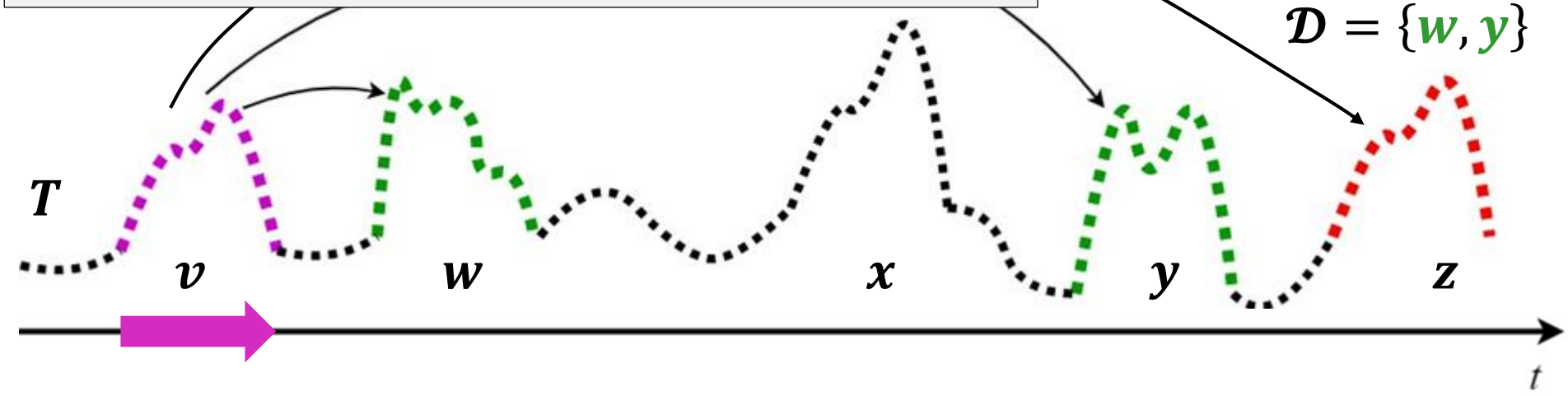
Очистка кандидатов (1)

Сканировать ряд T:
текущая подпоследовательность s
Кандидат := TRUE
for each $c_i \in \mathcal{C}$
if $\mathrm{ED}(s, c_i) < r$ **and** $s \cap c_i = \emptyset$ **then**
$\mathcal{C} := \mathcal{C} \setminus c_i$; Кандидат := FALSE
if Кандидат = TRUE **then** $\mathcal{C} := \mathcal{C} \cup s$

$\mathcal{D} = \{w, y, z\}$

$\mathrm{ED}(v, w) \geq r$
$\mathrm{ED}(v, y) \geq r$
$\mathrm{ED}(v, z) < r$

$\mathcal{D} = \{w, y\}$



Очистка кандидатов (2)

$\mathcal{D} := \mathcal{C}$

Сканировать ряд T:

текущая подпоследовательность s

for each $d_i \in \mathcal{D}$

if $\mathrm{ED}(s, d_i) < r$ **and** $s \cap d_i = \emptyset$ **then**

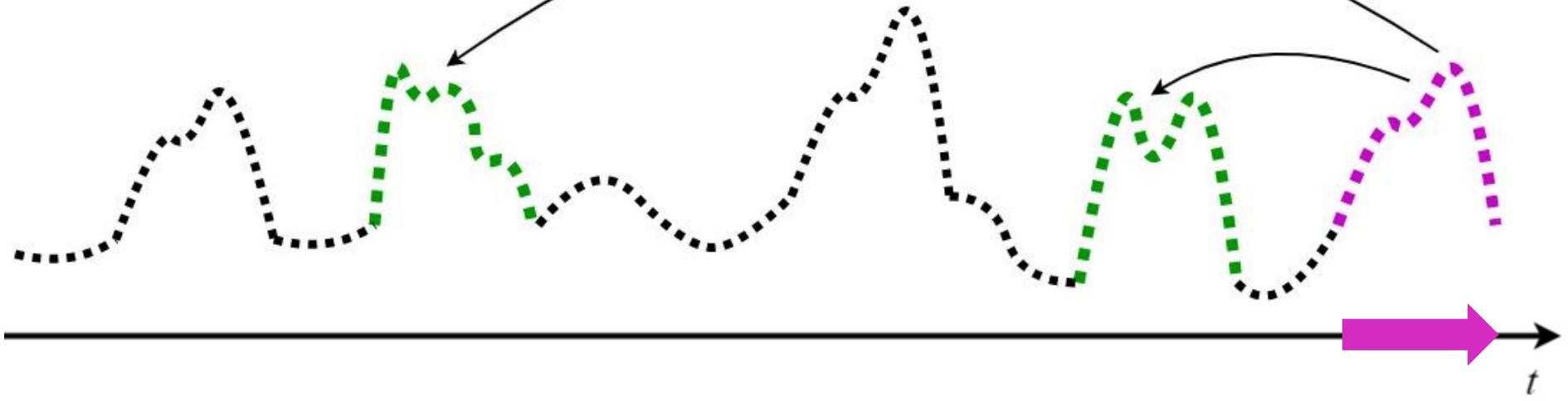
$\mathcal{D} := \mathcal{D} \setminus d_i$

$\mathcal{D} = \{w, y\}$

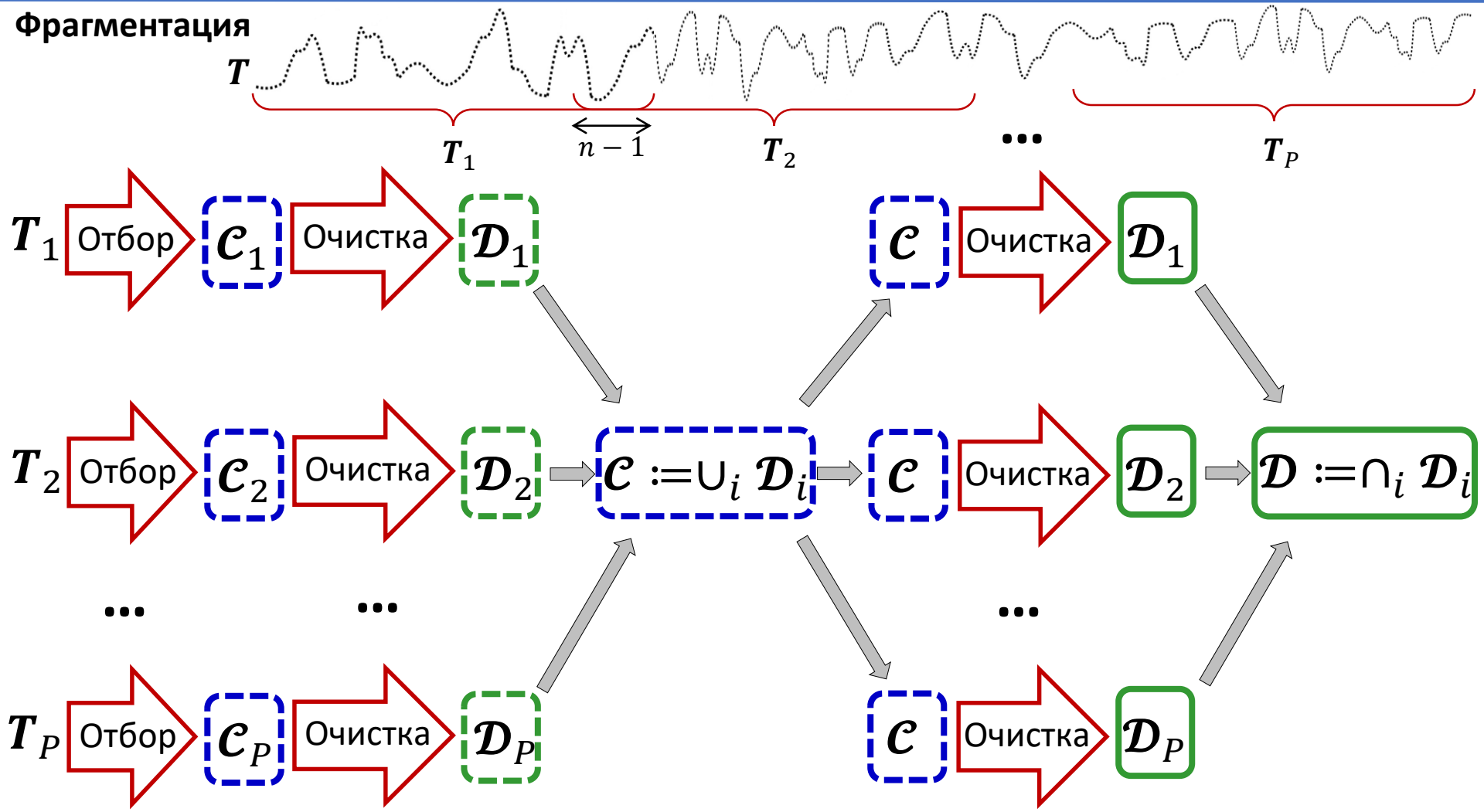
$\mathrm{ED}(z, w) \geq r$

$\mathrm{ED}(z, y) \geq r$

$\mathcal{D} = \{w, y\}$



Распределенный алгоритм DADD



Подбор порога r для поиска диапазонных диссонансов

1. Выбрать случайный отрезок ряда макс длины, который может быть размещен в памяти
2. Найти диссонанс в выбранном отрезке с помощью алгоритма HOTSAX
3. Взять в качестве порога r расстояние от диссонанса до его ближайшего соседа

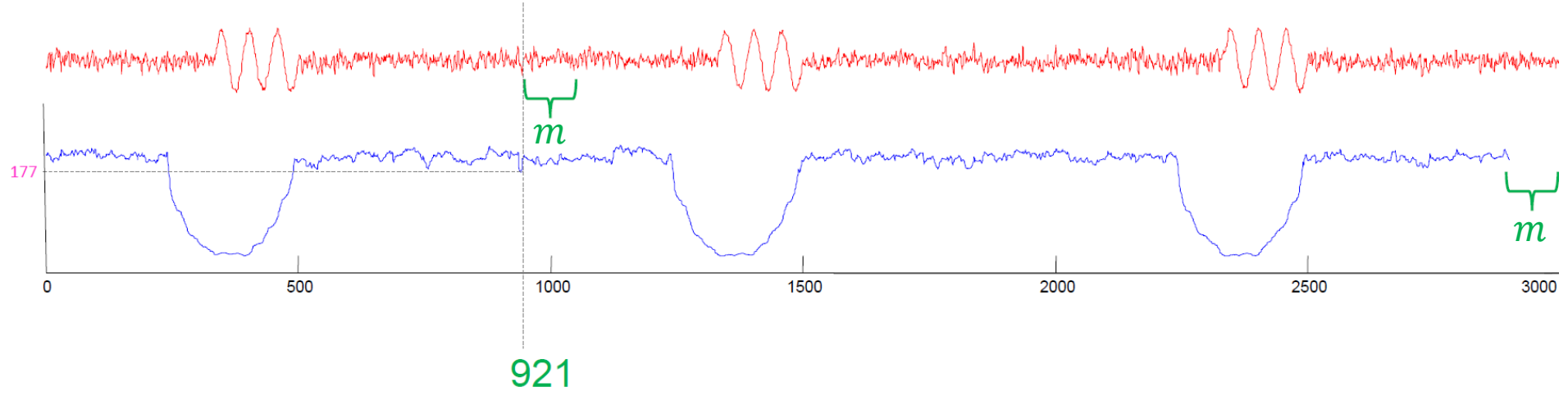
Матричный профиль (МП) временного ряда

$$MP(T, m) = (p_1, \dots, p_{n-m+1}),$$

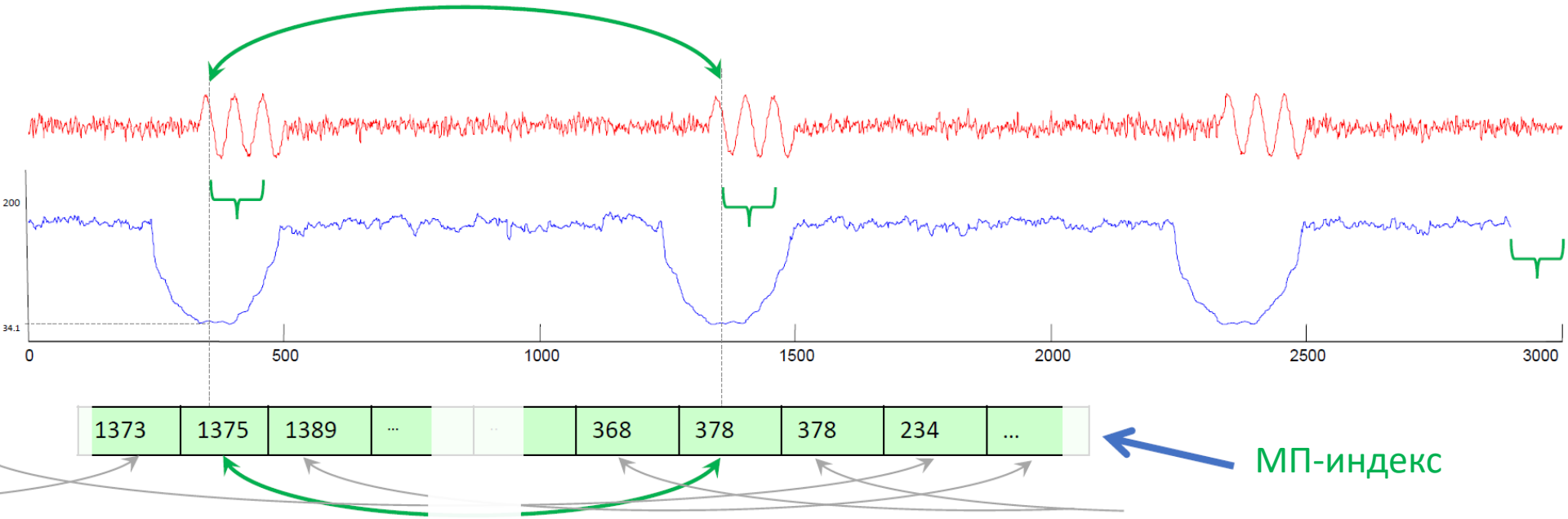
$$p_i = \mathrm{ED}(\hat{T}_{i,m}, \hat{N}(T_{i,m}))$$

$$N(T_{i,m}) = \arg \min_{|i-j|>m} \mathrm{ED}(T_{i,m}, T_{j,m})$$

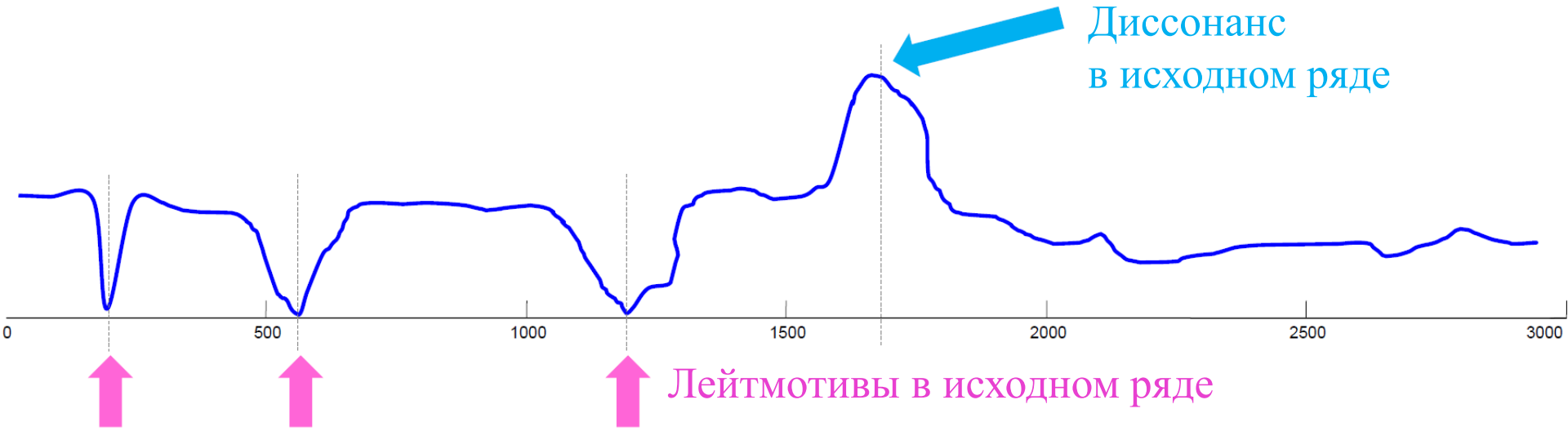
$\hat{\cdot}$ означает z-нормализацию



МП-индексы ближайших соседей

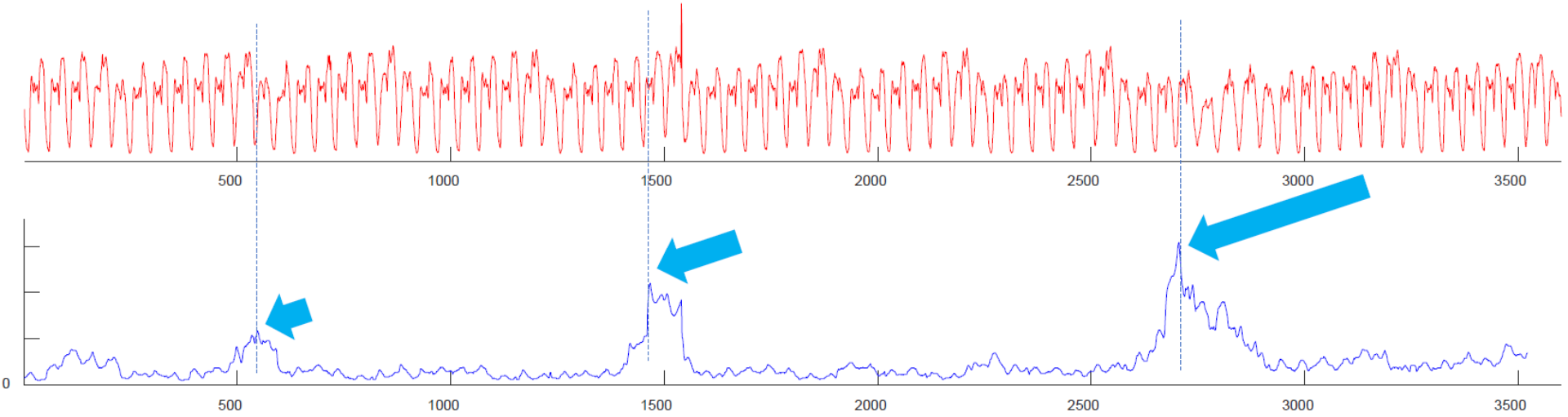


Диссонансы и лейтмотивы на матричном профиле



- Относительное малое значение указывает на обязательное наличие во временном ряде очень похожей подпоследовательности. Пара очень похожих подпоследовательностей – *лейтмотив (motif)*
- Относительно большое значение указывает на обязательное наличие во временном ряде уникальной подпоследовательности (аномалия, диссонанс)

Пример МП



- Ряд: среднее число пассажиров NY такси осенью 2014 г.
- Длина подпоследовательности МП: 2 дня
- Пики (диссонансы):
  день Колумба (13.10.14), день перехода на зимнее время (06.11.14), день благодарения (27.11.2014)

Литература

- Han J., Kamber M., Pei J. Data Mining: Concepts and Techniques. 3rd Edition. Morgan Kaufmann, 2011. 740 p. ISBN 978-0123814791
  - 12. Outlier Detection, pp. 543-582
- Tan P.-N., Steinbach M., Karpatne A., Kumar V. Introduction to Data Mining. 2nd Edition. Pearson, 2019. 839 p. ISBN: 978-0-13-312890-1
  - 9. Anomaly Detection, pp. 703-802