

Введение в дисциплину «Анализ и прогнозирование временных рядов методами искусственного интеллекта»*



*Невольно изречешь: о tempora, о mores! —
Когда поразглядиши, какая в жизни горесть.*

Н.А. Некрасов

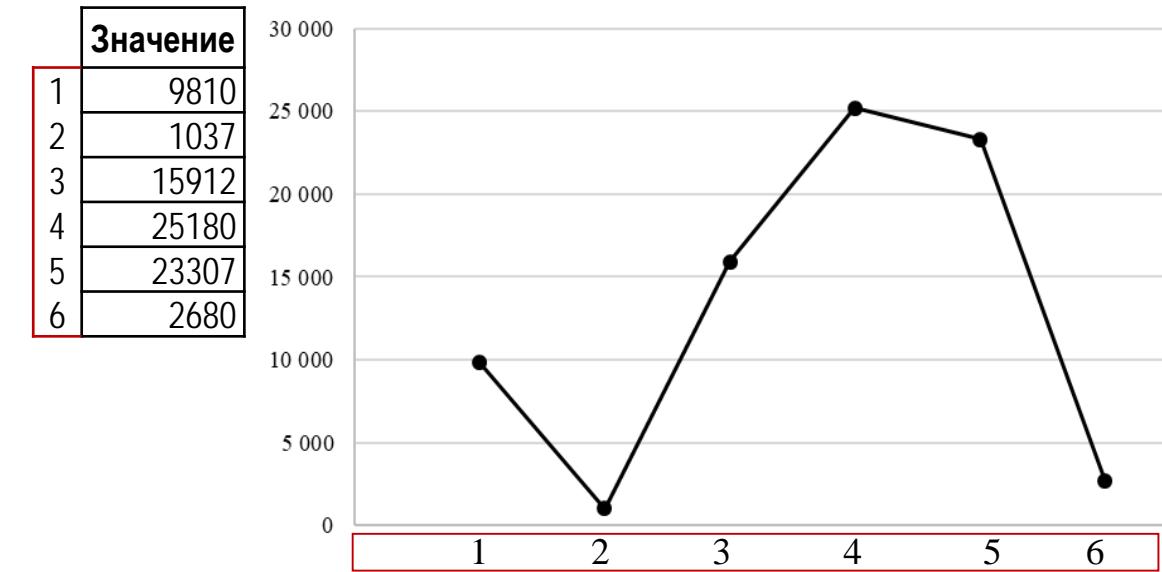
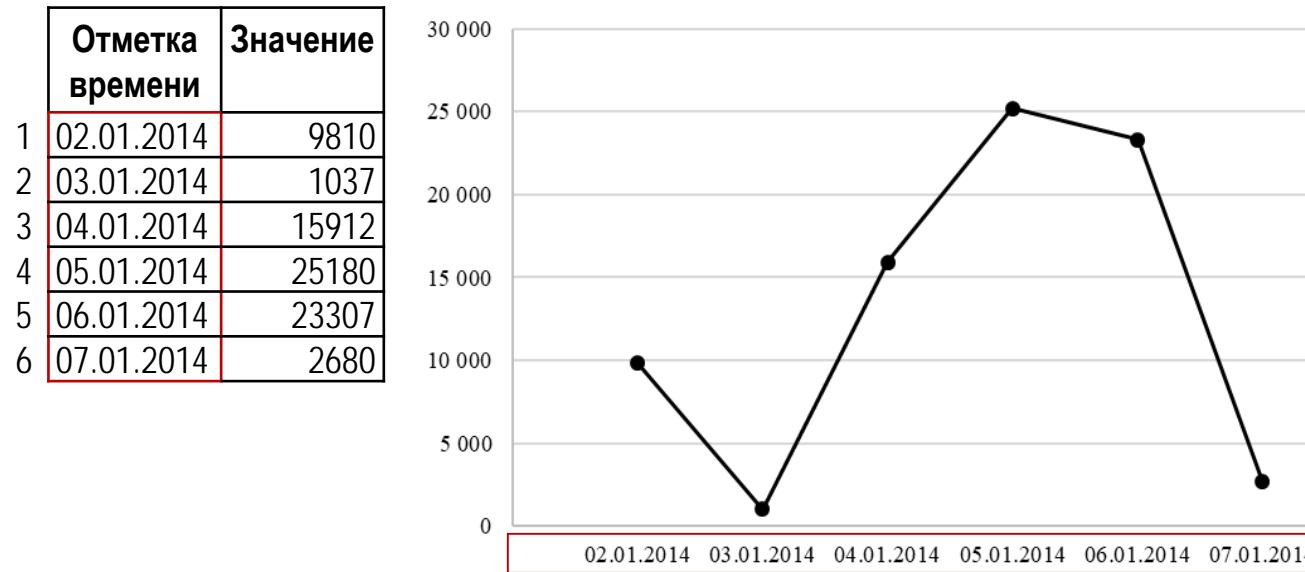
* При подготовке слайдов лекций курса использованы материалы статей и слайды докладов проф. Имонна Кеога, Калифорнийский университет в Риверсайде, США (Eamonn Keogh, University of California Riverside, USA), см. <https://www.cs.ucr.edu/~eamonn/>

Содержание

- Понятие временного ряда
- Временные ряды в различных предметных областях
- Особенности интеллектуального анализа временных рядов
- Основные задачи анализа временных рядов
- Определения и нотация

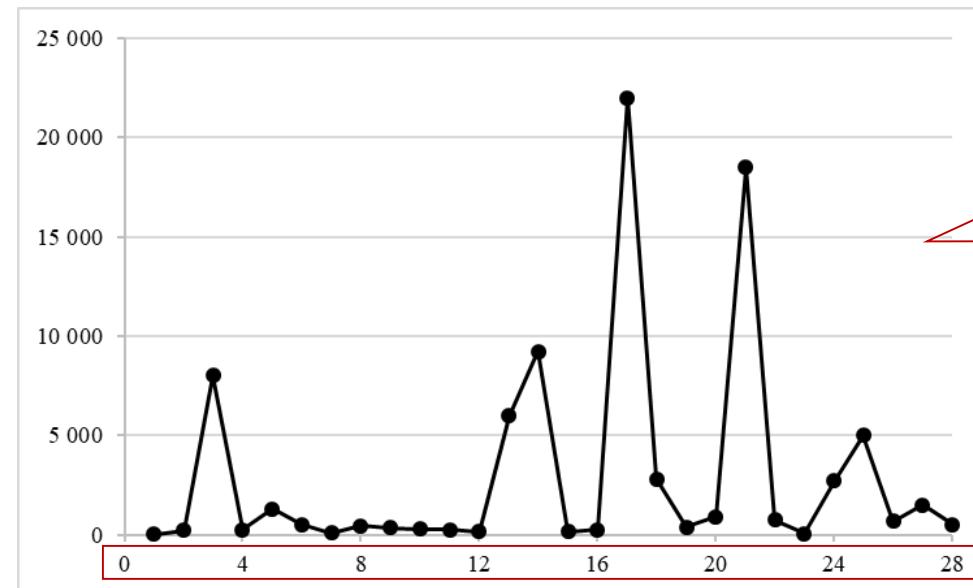
Временной ряд – упорядоченная по времени последовательность чисел

- Точки (элементы) ряда ассоциированы с временными метками, сделанными через **равные промежутки** (т.е. частота измерений фиксирована)
- Значения временных меток могут не подвергаться обработке или отсутствовать в исходных данных



Важность фиксированной частоты измерений

№	Отметка времени	Значение
1	02.01.2014 07:29	30
2	02.01.2014 16:05	230
3	02.01.2014 16:17	8000
4	02.01.2014 18:15	250
5	02.01.2014 19:22	1300
6	03.01.2014 08:27	500
7	03.01.2014 09:18	100
8	03.01.2014 10:58	437
9	04.01.2014 04:22	350
10	04.01.2014 09:58	280
11	04.01.2014 12:44	240
12	04.01.2014 18:25	160
13	04.01.2014 20:26	6000
14	04.01.2014 21:33	9232
15	05.01.2014 06:22	140
16	05.01.2014 14:17	240
17	05.01.2014 14:48	22000
18	05.01.2014 16:11	2800
19	06.01.2014 09:43	377
20	06.01.2014 12:00	910
21	06.01.2014 14:14	18500
22	06.01.2014 15:50	750
23	06.01.2014 16:12	70
24	06.01.2014 21:05	2700
25	07.01.2014 15:33	5000
26	07.01.2014 16:17	680
27	07.01.2014 18:11	1500
28	07.01.2014 22:03	500



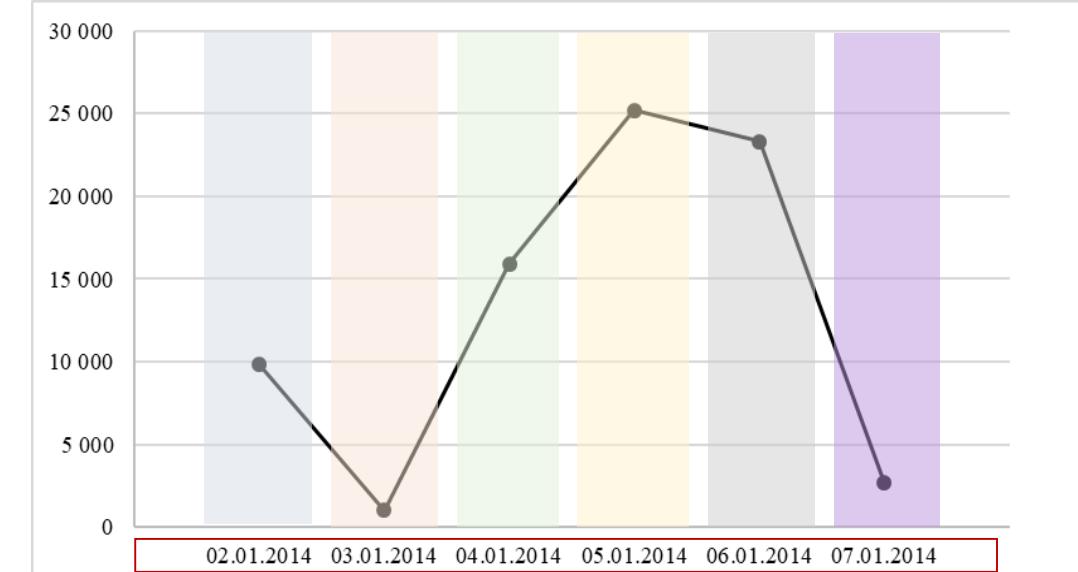
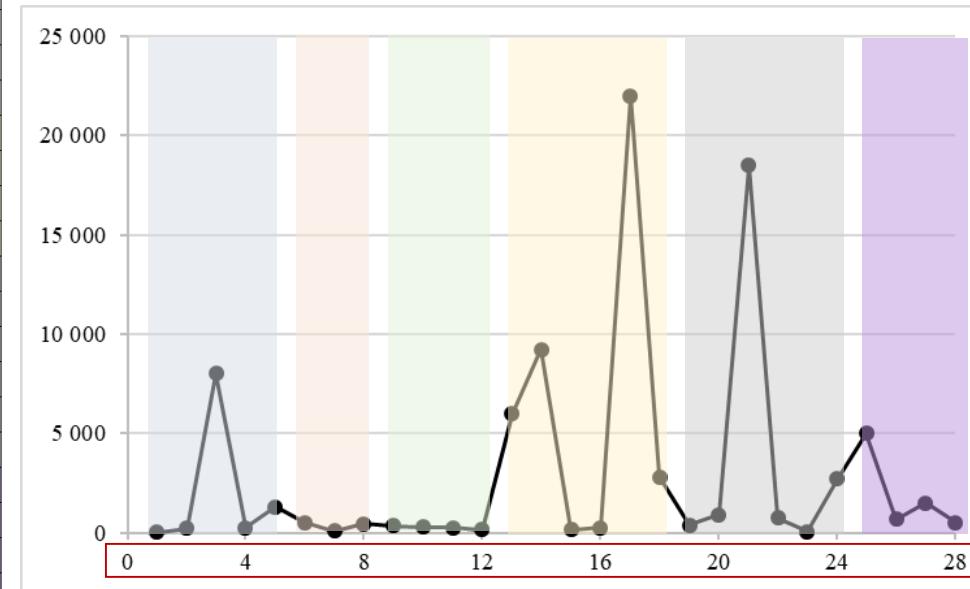
Обработка ряда
не имеет смысла,
т.к. измерения сделаны
через разные
промежутки времени

Важность фиксированной частоты измерений

№	Отметка времени	Значение
1	02.01.2014 07:29	30
2	02.01.2014 16:05	230
3	02.01.2014 16:17	8000
4	02.01.2014 18:15	250
5	02.01.2014 19:22	1300
6	03.01.2014 08:27	500
7	03.01.2014 09:18	100
8	03.01.2014 10:58	437
9	04.01.2014 04:22	350
10	04.01.2014 09:58	280
11	04.01.2014 12:44	240
12	04.01.2014 18:25	160
13	04.01.2014 20:26	6000
14	04.01.2014 21:33	9232
15	05.01.2014 06:22	140
16	05.01.2014 14:17	240
17	05.01.2014 14:48	22000
18	05.01.2014 16:11	2800
19	06.01.2014 09:43	377
20	06.01.2014 12:00	910
21	06.01.2014 14:14	18500
22	06.01.2014 15:50	750
23	06.01.2014 16:12	70
24	06.01.2014 21:05	2700
25	07.01.2014 15:33	5000
26	07.01.2014 16:17	680
27	07.01.2014 18:11	1500
28	07.01.2014 22:03	500

Агрегация с помощью суммирования

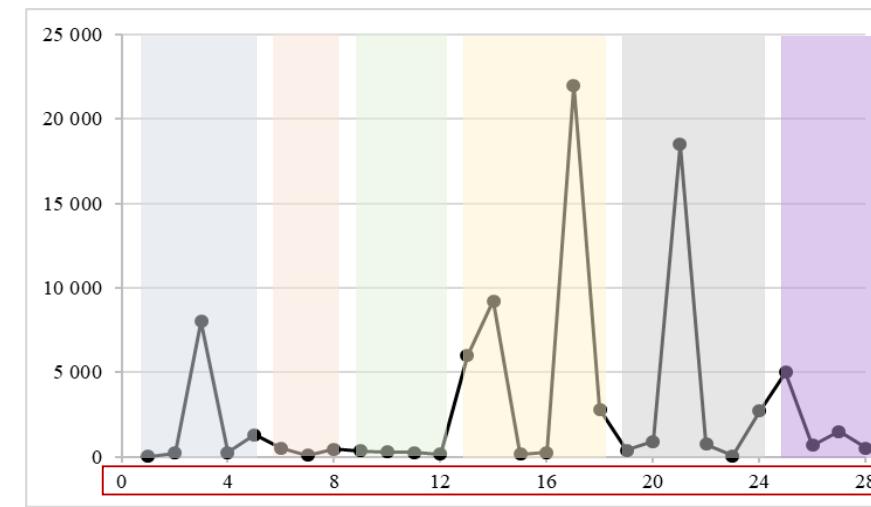
№	Отметка времени	Значение
1	02.01.2014	9810
2	03.01.2014	1037
3	04.01.2014	15912
4	05.01.2014	25180
5	06.01.2014	23307
6	07.01.2014	2680



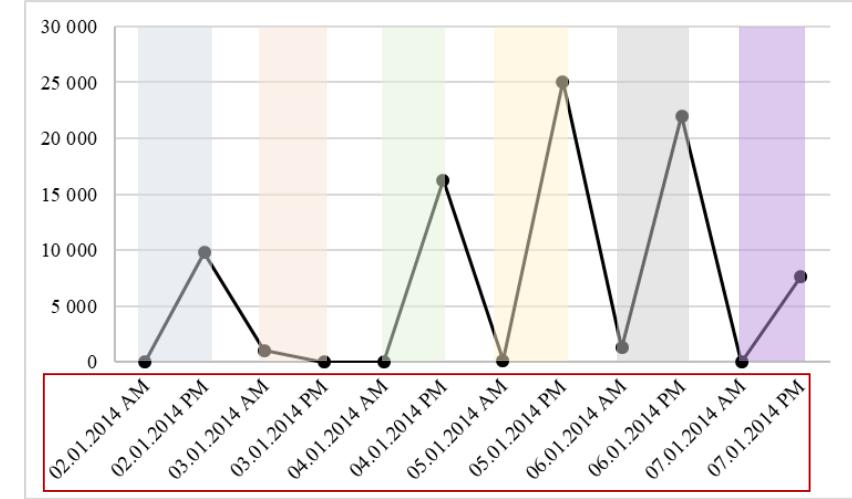
Важность фиксированной частоты измерений

№	Отметка времени	Значение
1	02.01.2014 07:29	30
2	02.01.2014 16:05	230
3	02.01.2014 16:17	8000
4	02.01.2014 18:15	250
5	02.01.2014 19:22	1300
6	03.01.2014 08:27	500
7	03.01.2014 09:18	100
8	03.01.2014 10:58	437
9	04.01.2014 04:22	350
10	04.01.2014 09:58	280
11	04.01.2014 12:44	240
12	04.01.2014 18:25	160
13	04.01.2014 20:26	6000
14	04.01.2014 21:33	9232
15	05.01.2014 06:22	140
16	05.01.2014 14:17	240
17	05.01.2014 14:48	22000
18	05.01.2014 16:11	2800
19	06.01.2014 09:43	377
20	06.01.2014 12:00	910
21	06.01.2014 14:14	18500
22	06.01.2014 15:50	750
23	06.01.2014 16:12	70
24	06.01.2014 21:05	2700
25	07.01.2014 15:33	5000
26	07.01.2014 16:17	680
27	07.01.2014 18:11	1500
28	07.01.2014 22:03	500

Агрегация с помощью суммирования



№	Отметка времени	Значение
1	02.01.2014 AM	30
2	02.01.2014 PM	9780
3	03.01.2014 AM	1037
4	03.01.2014 PM	0
5	04.01.2014 AM	0
6	04.01.2014 PM	16262
7	05.01.2014 AM	140
8	05.01.2014 PM	25040
9	06.01.2014 AM	1287
10	06.01.2014 PM	22020
11	07.01.2014 AM	0
12	07.01.2014 PM	7680



Временной ряд \neq сигнал (функция)

Временной ряд

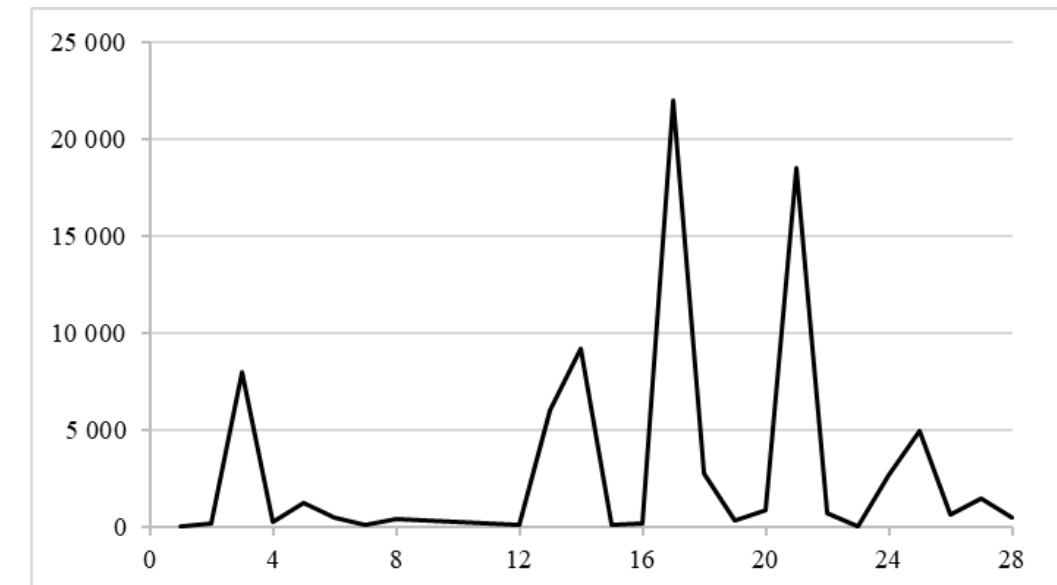
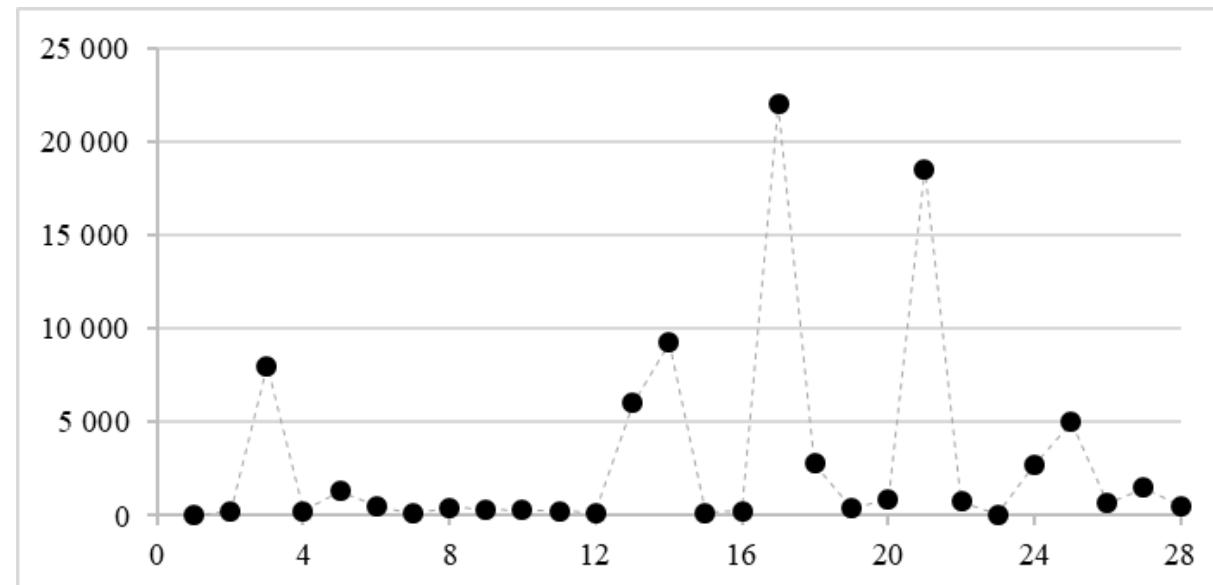
$$T = (t_1, \dots, t_n)$$

$$\forall i \in \mathbb{N}, 1 \leq i \leq n: \exists t_i \in \mathbb{R}$$

Сигнал (функция)

$$T: \text{dom } T \rightarrow \mathbb{R}$$

$$\forall t \in \text{dom } T: \exists T(t) \in \mathbb{R}$$



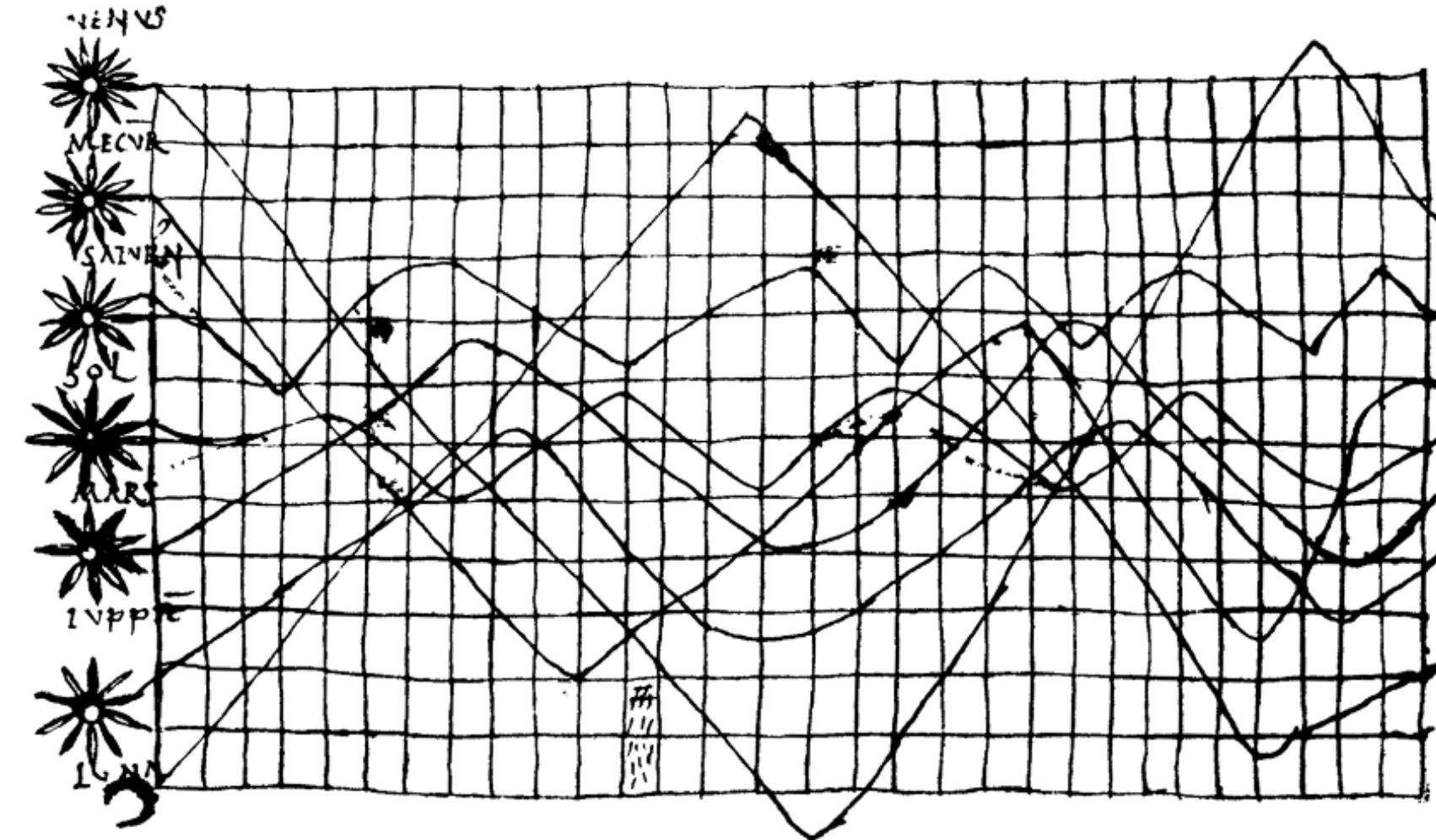
Содержание

- Понятие временного ряда
- **Временные ряды в различных предметных областях**
- Особенности интеллектуального анализа временных рядов
- Основные задачи анализа временных рядов
- Определения и нотация

Люди измеряют всевозможные вещи, изменяющиеся во времени

- ЭКГ, пульс, давление, калории
- Рождаемость и смертность
- Температура и влажность воздуха
- Расход электричества и воды
- Рейтинг популярности политиков
- Спортивная статистика
- Клики веб-страниц
- Курсы валют и акций
- ВВП и госдолг
- ...

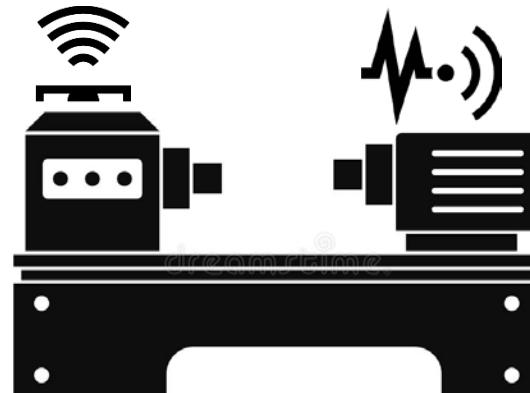
Временные ряды всегда...



Временные ряды,
показывающие наклоны
планетных орбит, X в.
(возможно, **наиболее**
старое изображение
временных рядов)

Tufte E. The Visual Display of Quantitative Information. Graphics Press, 2001. 200 p.

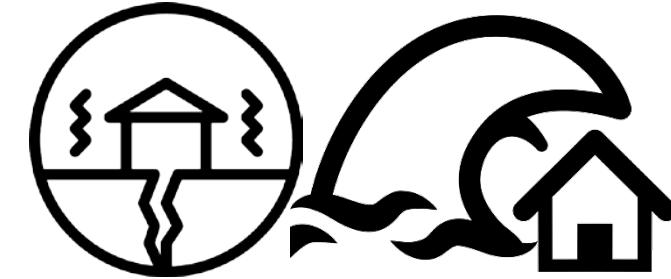
Временные ряды всюду...



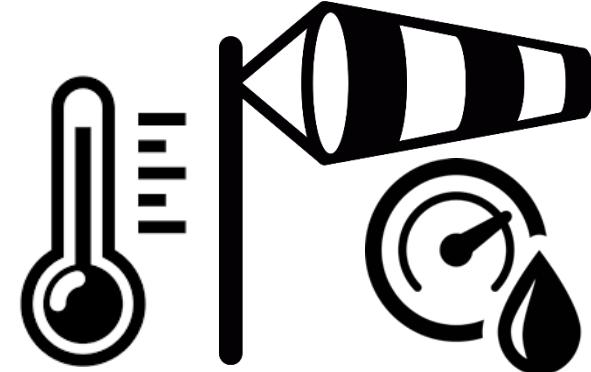
Умное производство,
предиктивное ТО



Интернет
вещей



Предсказание
природных катализмов



Прогноз погоды,
моделирование климата



Персональная
медицина



Сельское хоз-во,
животноводство



Био- и хемо-
информатика

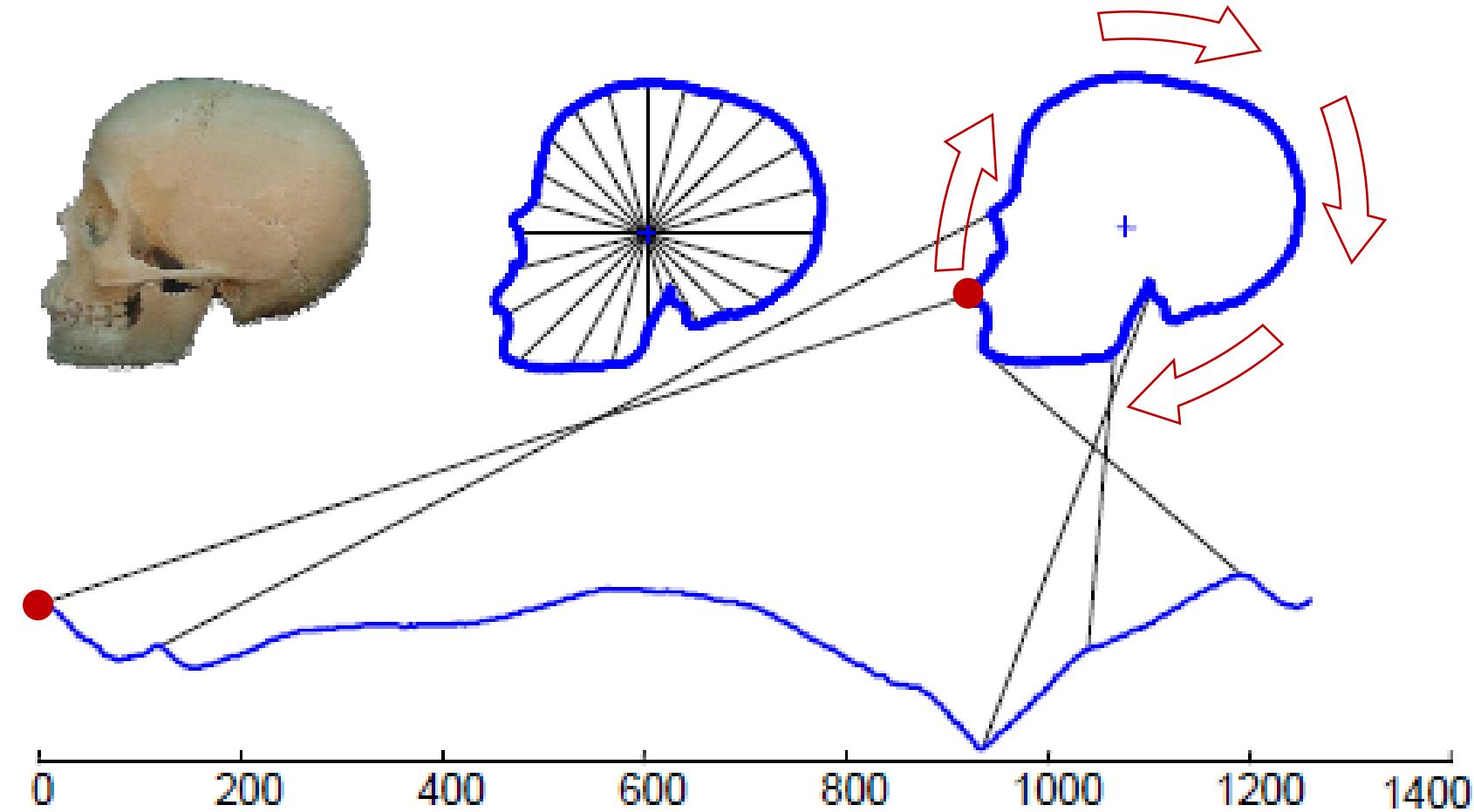


Экономика, бизнес,
финансы



Системы
электронного обучения

Временной ряд изображения



Keogh E. et al. LB_Keogh supports exact indexing of shapes under rotation invariance with arbitrary representations and distance measures. VLDB 2006. pp. 882-893. [URL](#)

Временной ряд из видео

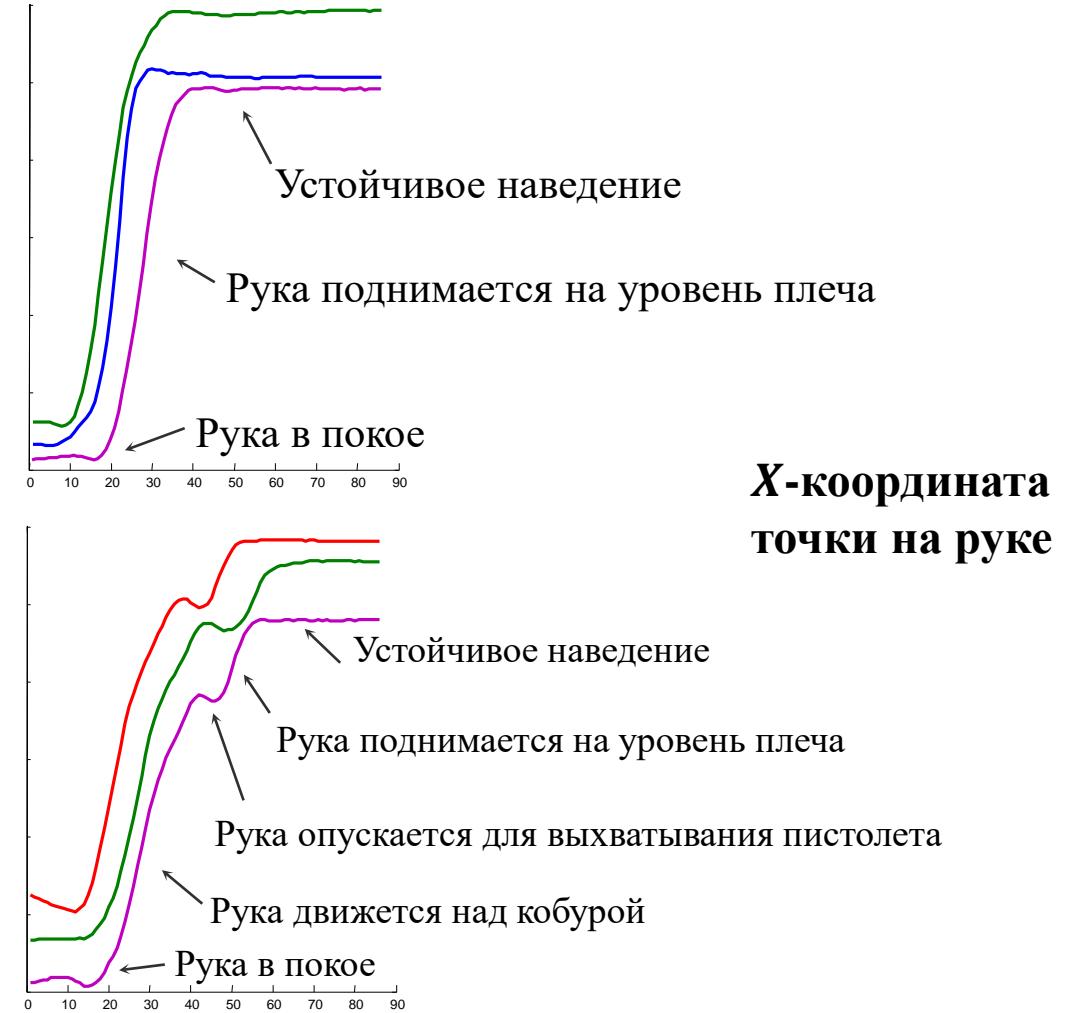
Классификация
“Gun/No gun”



**Указание
рукой**

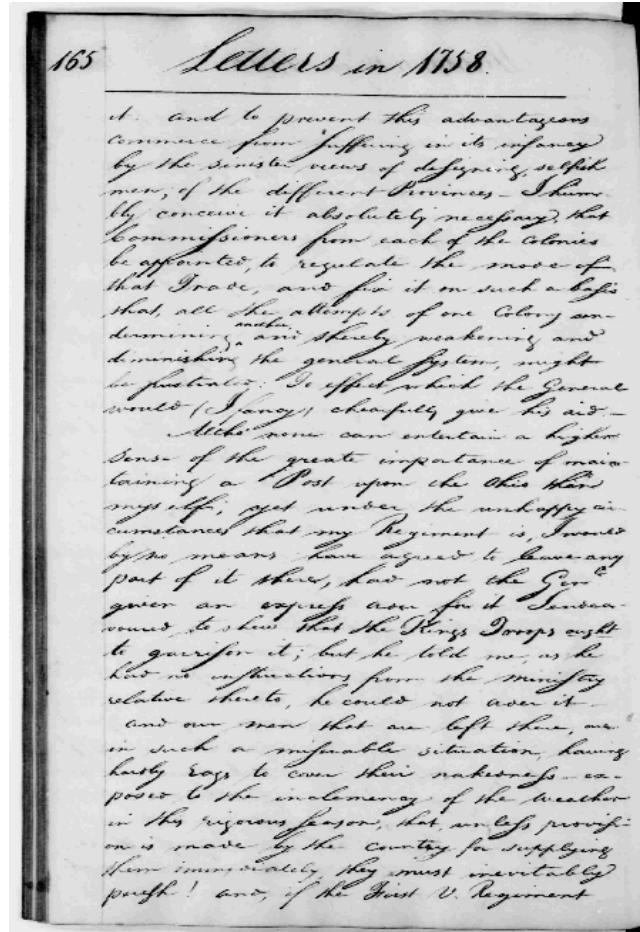


**Вытаскивание
пистолета**

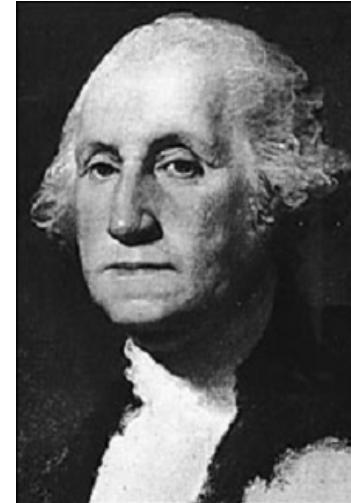


Keogh E. et al. A novel technique for indexing video surveillance data. First ACM SIGMM Int. workshop on Video surveillance. 2003. P. 98-106. <https://doi.org/10.1145/982452.982465>

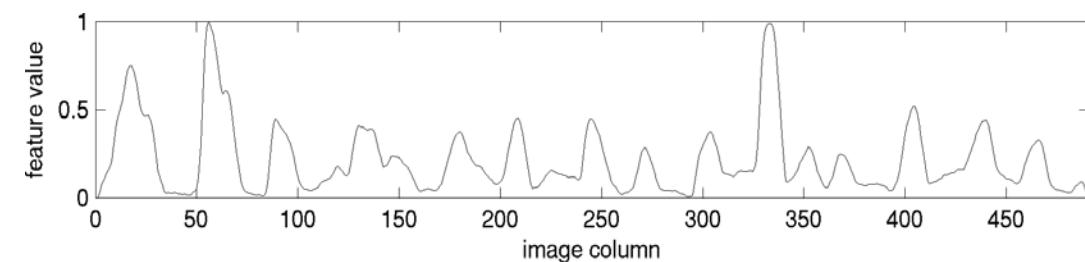
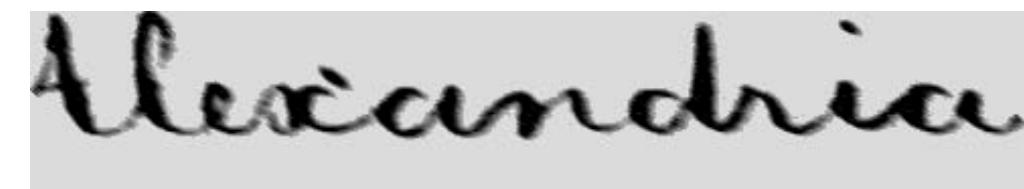
Временной ряд из рукописного текста



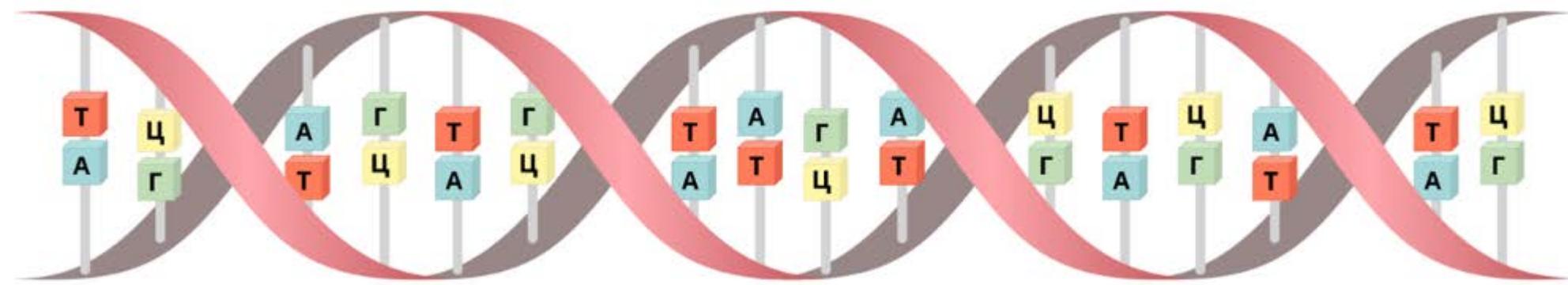
George Washington manuscript



Джордж Вашингтон
(George Washington)
1732-1799



ДНК (дезоксирибонуклеиновая кислота) как временной ряд



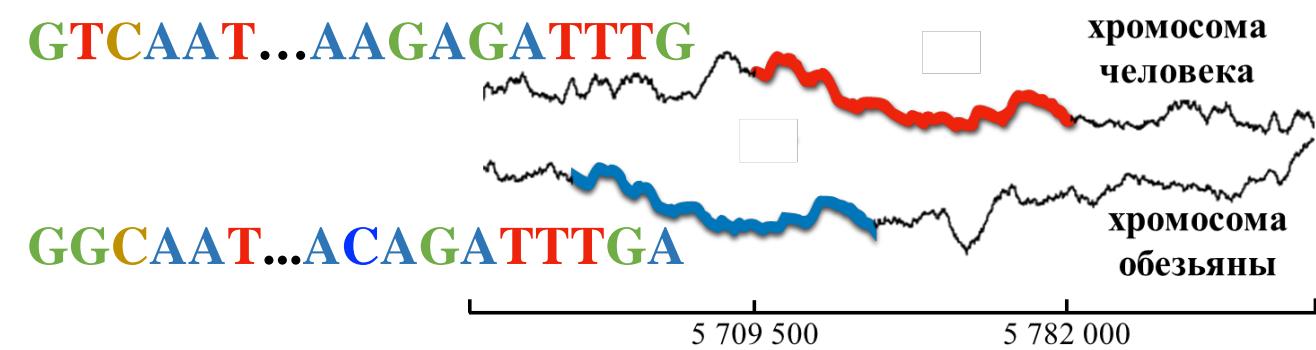
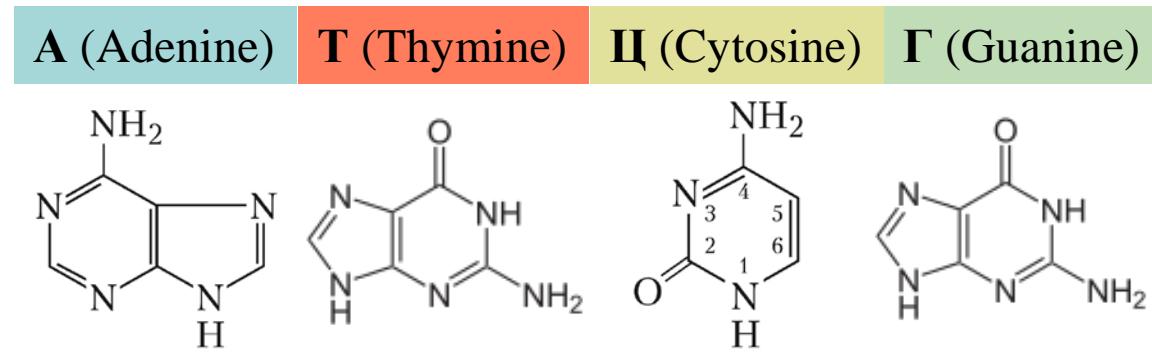
ДНК – макромолекула, обеспечивающая хранение, передачу через поколения и реализацию генетической программы развития и функционирования организмов

**Трансформация ДНК
во временной ряд**

```

 $t_1 := 0$ 
for  $i \in 1..|DNAstr|$  do
  case  $DNAstr[i]$  of
    A :  $t_{i+1} := t_{i+2}$ 
    G :  $t_{i+1} := t_{i+1}$ 
    C :  $t_{i+1} := t_{i-1}$ 
    T :  $t_{i+1} := t_{i-2}$ 
  end

```



Содержание

- Понятие временного ряда
- Временные ряды в различных предметных областях
- **Особенности интеллектуального анализа временных рядов**
- Основные задачи анализа временных рядов
- Определения и нотация

Почему временные ряды анализировать сложнее, чем другие данные

- Большая длина
- Субъективность схожести (рядов и подпоследовательностей)
- Пропущенные значения
- Различные форматы данных и частоты снятия показаний, шумы
(не рассматривается в рамках курса)

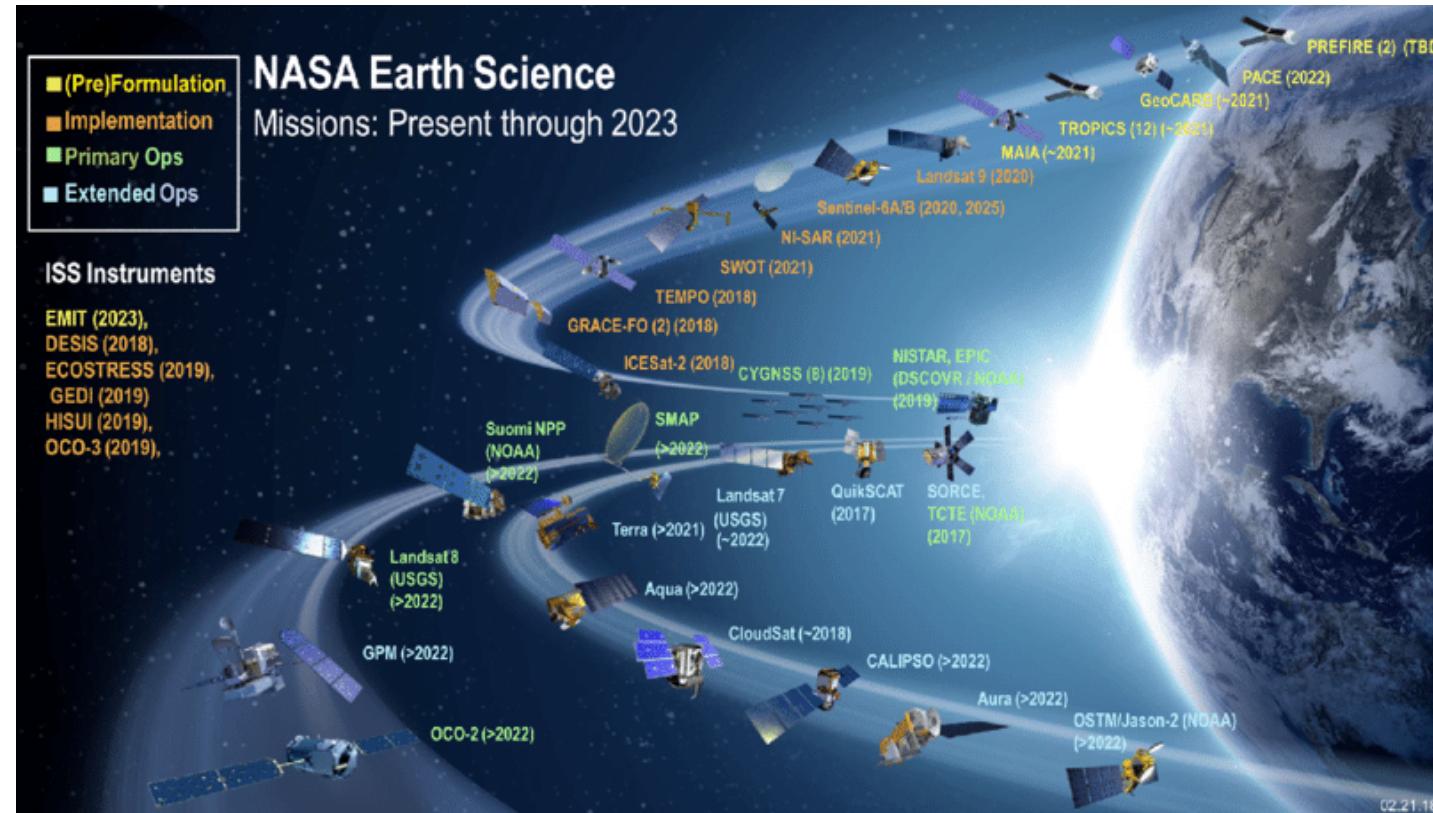
Большие временные ряды: АЕМО Solar



Оборудование АЕМО (Australian Energy Market Operator) регистрирует выработку солнечной энергии (МВт) в Австралии с 2019 г. каждые 4 с

<https://zenodo.org/record/4656027>

Большие временные ряды: Earth Observations Database

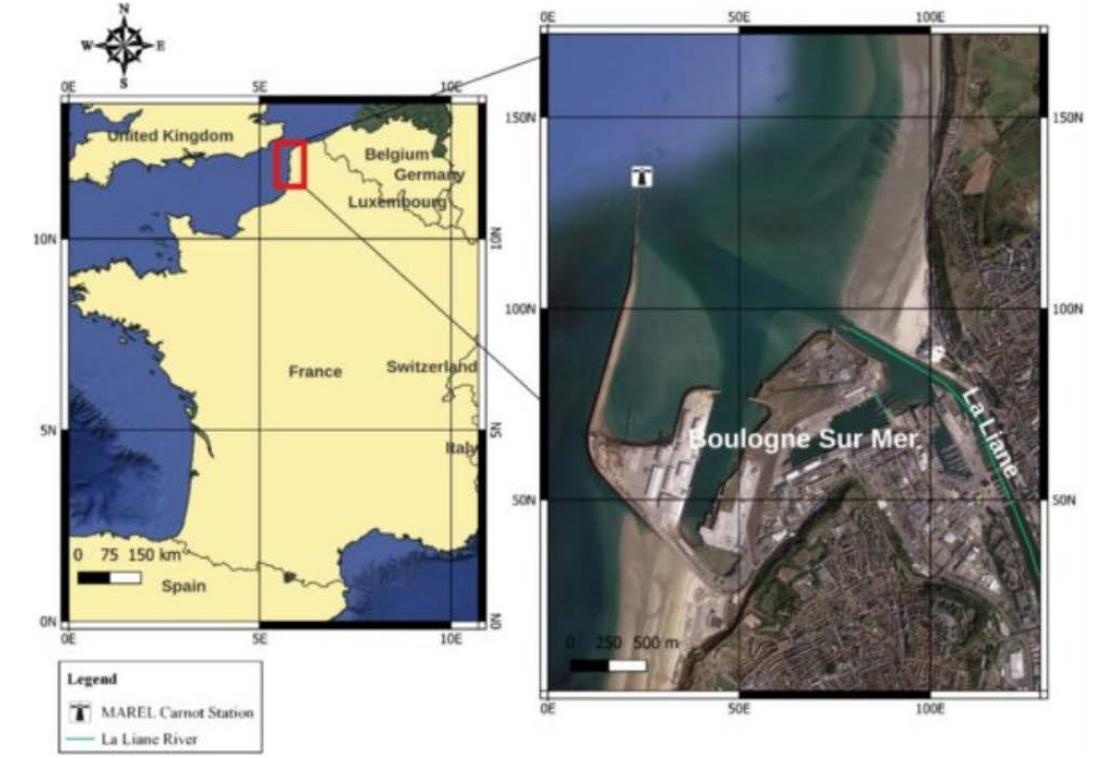


Размер NASA Space Shuttle Earth Observations Database – 40 Пб* (2020 г.), ожидаемый далее ежегодный прирост 50 Пб

* 1 Петабайт=10¹⁵ (квадриллион) байт, 10¹⁵ ≈ к-во синапсов в головном мозге человека

<https://www.nasa.gov/feature/goddard/2020/nasa-funds-projects-to-make-geosciences-data-more-accessible>

Большие временные ряды: MAREL Carnot



Океанографическая станция **MAREL Carnot** с 2004 г. регистрирует каждые 20 мин. более чем 15 химических и биологических характеристик воды в проливе Ла-Манш

Ben Ismail D.K. *et al.* Statistical properties and time-frequency analysis of temperature, salinity and turbidity measured by the MAREL Carnot station in the coastal waters of Boulogne-sur-Mer (France). Journal of Marine Systems. 2016. Vol. 162. P. 137-153. <https://doi.org/10.1016/j.jmarsys.2016.03.010>

Большие временные ряды: DEBS Challenge



Набор данных **DEBS challenge**: сенсоры пространственного позиционирования закреплены на бутсах игроков и перчатках вратаря (200 Гц), а также на мяче (2000 Гц), всего 15К событий в секунду

Mutschler C. et al. The DEBS 2013 grand challenge. DEBS'13: Proc. of the 7th ACM international conference on Distributed event-based systems. 2013. P. 289–294.
<https://doi.org/10.1145/2488222.2488283>

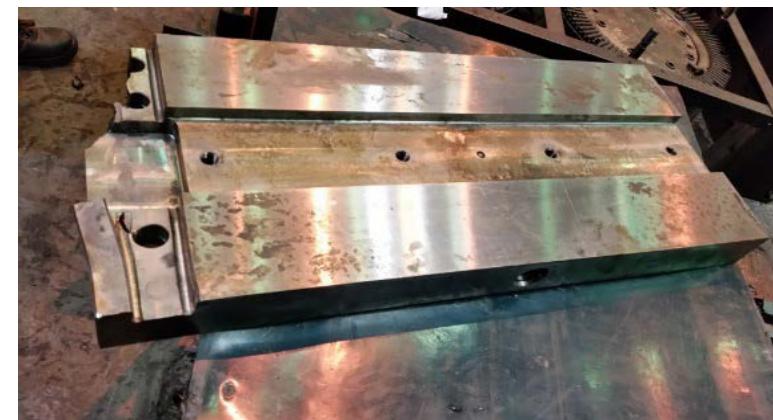
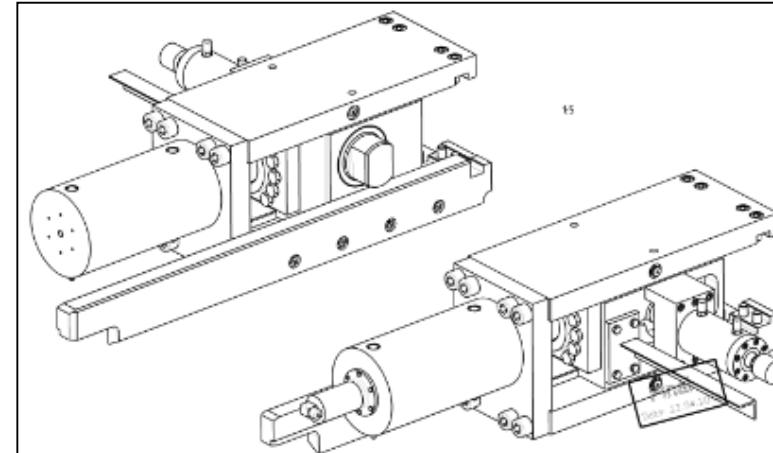
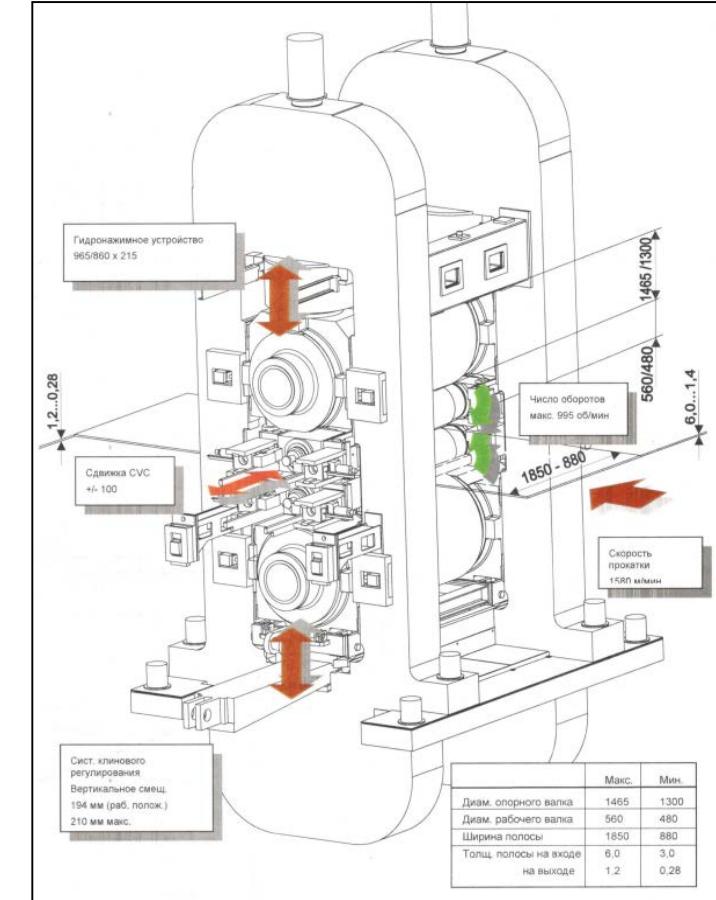
Большие временные ряды: Madrid VAR



Система контроля дорожного движения в Мадриде насчитывает более 3500 автоматических регистраторов транспортных средств (VAR, Vehicle Automatic Registrar), выдающих показания каждые 15 мин., начиная с 2014 г.

Laña I. *et al.* On the imputation of missing data for road traffic forecasting: New insights and novel techniques. *Transportation Research Part C: Emerging Technologies*. 2018. Vol. 90. pp. 18-33. DOI: [10.1016/j.trc.2018.02.021](https://doi.org/10.1016/j.trc.2018.02.021)

Большие временные ряды: система профилировки валков прокатного стана



CVC (Continuously Variable Curvature)-система стана холодной прокатки Магнитогорского металлургического комбината насчитывает более 100 датчиков с частотой 1 Гц.

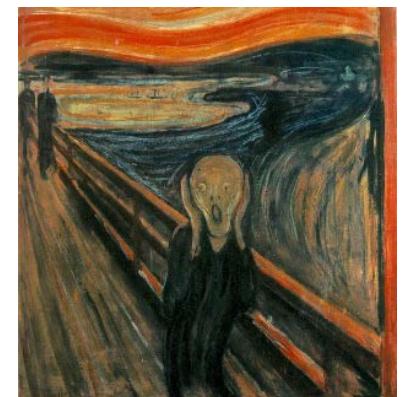
Краева Я.А. Поиск аномалий в сенсорных данных цифровой индустрии с помощью параллельных вычислений. Вестник ЮУрГУ. Серия: Вычислительная математика и информатика. 2023. Т. 12, № 2. С. 47–61.<https://doi.org/10.14529/cmse230202>

Большие ряды и проклятие размерности

- Экспоненциальный рост данных и операций, необходимых для решения аналитических и комбинаторных задач, при увеличении размерности пространства
 - увеличение объема точек обучающей выборки из-за увеличения количества координат (признаков)
 - увеличение числа операций для обработки координат в группах точек
 - доминирующие и ничтожные координаты
 - слабое отличие расстояний между различными парами точек



[Ричард Беллман](#)
(Richard Ernest Bellman)
1920-1984



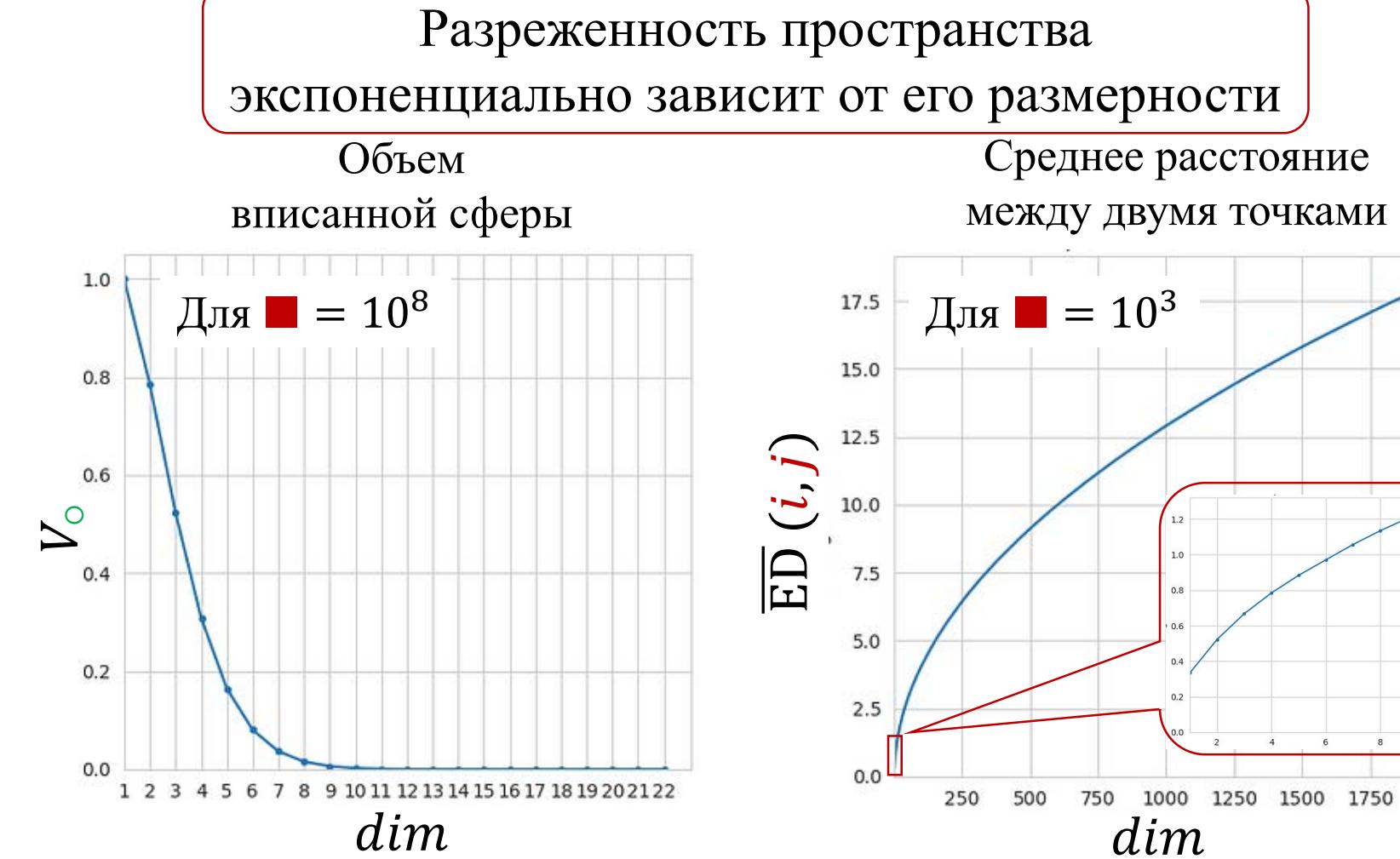
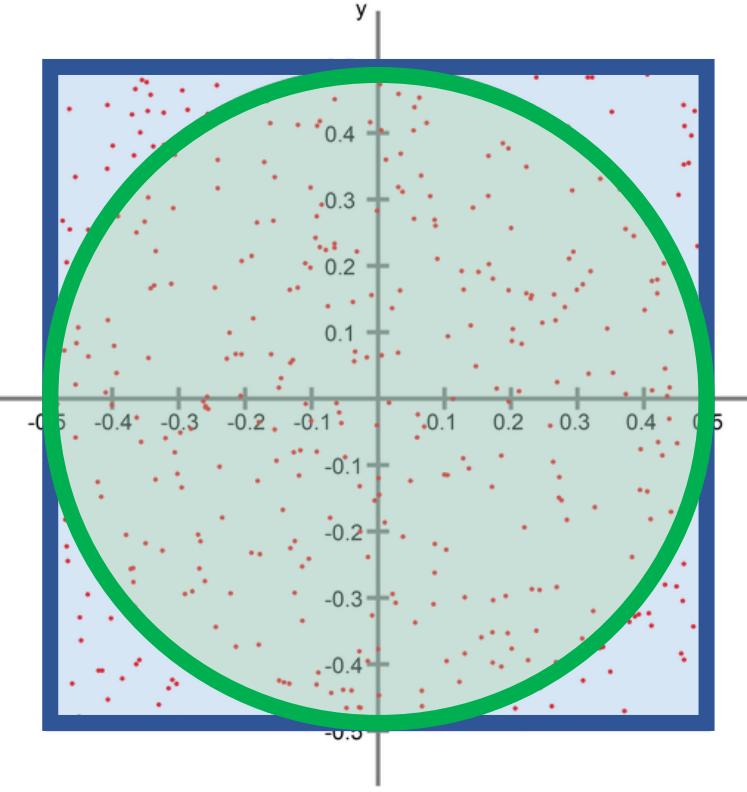
[Э. Мунк. «Крик»](#)

Проклятие размерности*

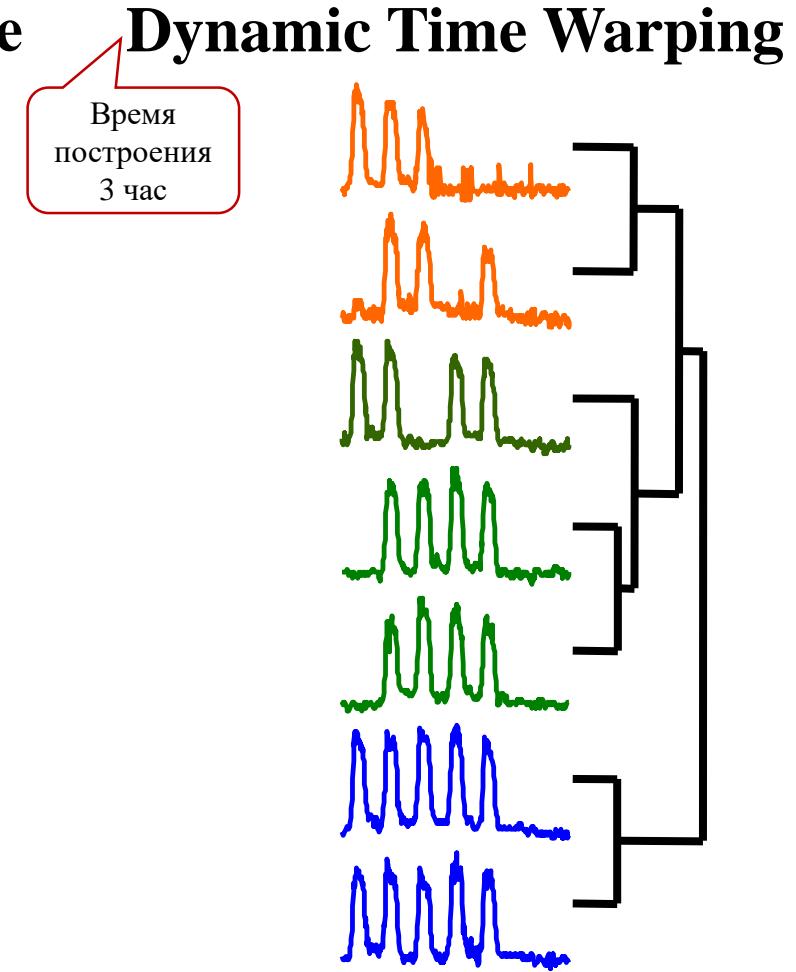
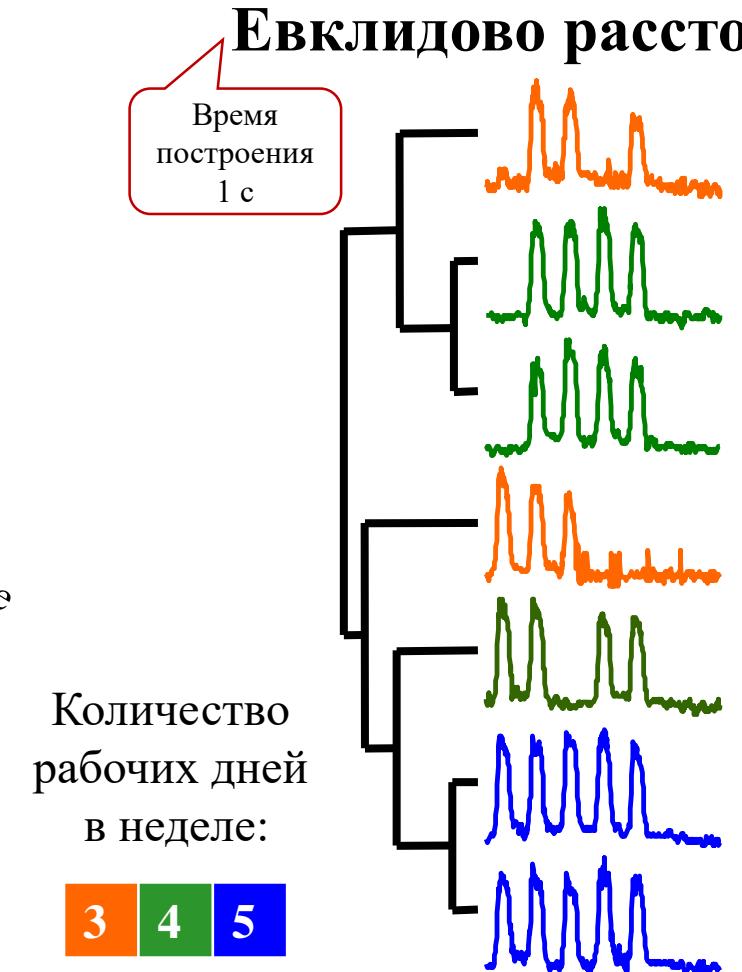
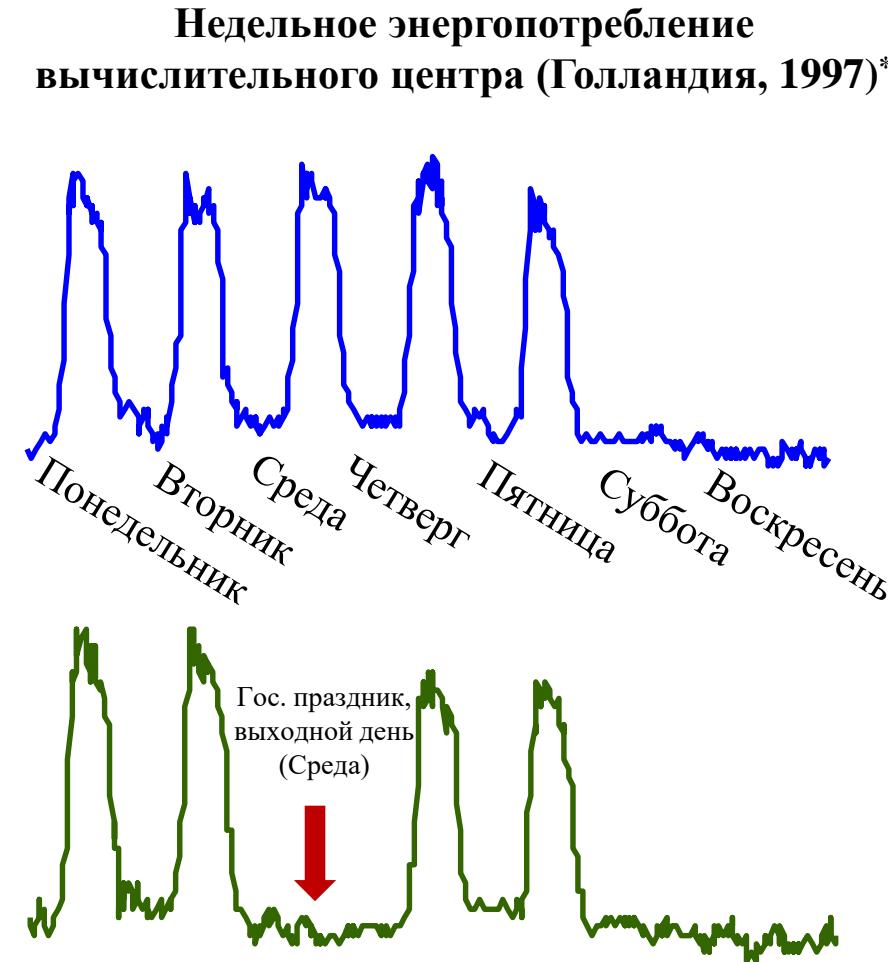
* Визуализация реальных масштабов проклятия размерности. [URL](#)

$$V_{\square} = 1,$$

$$V_{\circ} = \lim_{dim \rightarrow +\infty} \frac{\blacksquare}{\bullet}$$



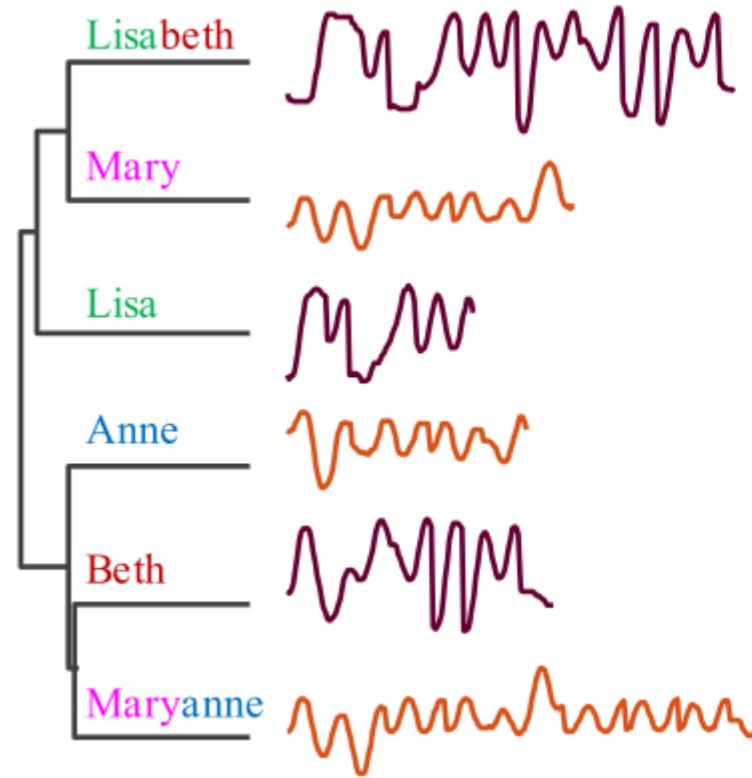
Схожесть определяется задачей и предметной областью



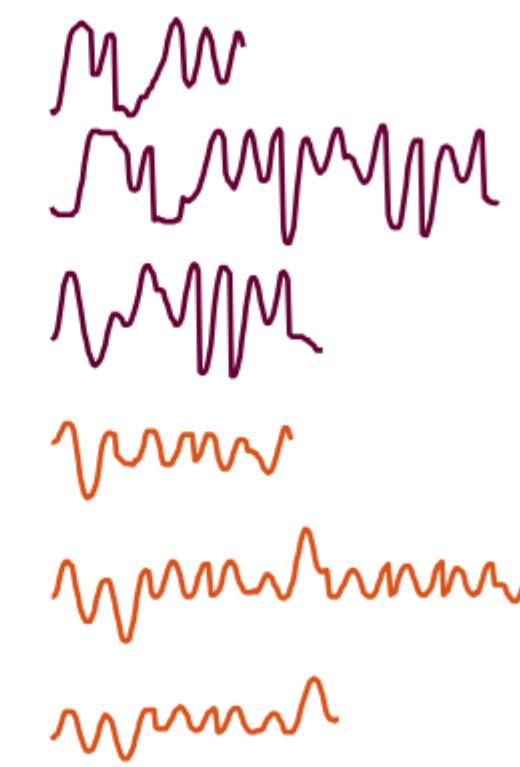
* van Wijk J.J., van Selow R.R. Cluster and calendar based visualization of time series data. INFOVIS 1999: 4-9. DOI: [10.1109/INFVIS.1999.801851](https://doi.org/10.1109/INFVIS.1999.801851)

Схожесть определяется задачей и предметной областью

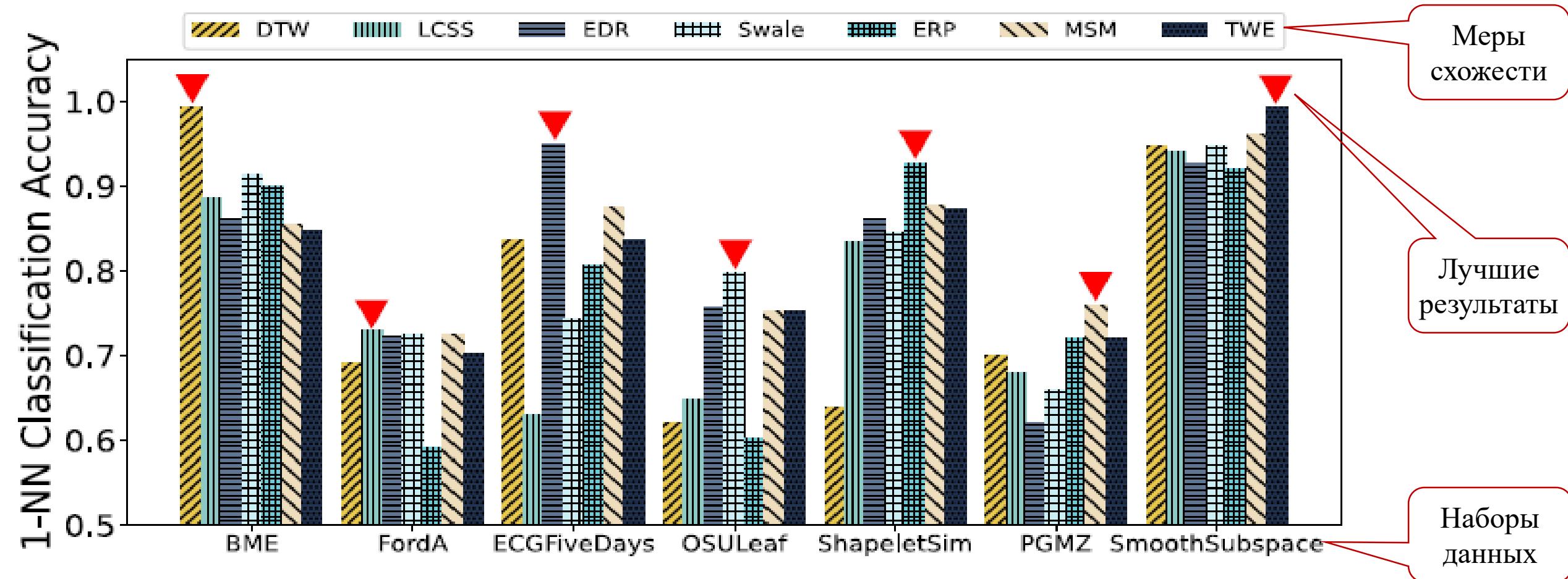
Евклидово расстояние



Мера MPdist

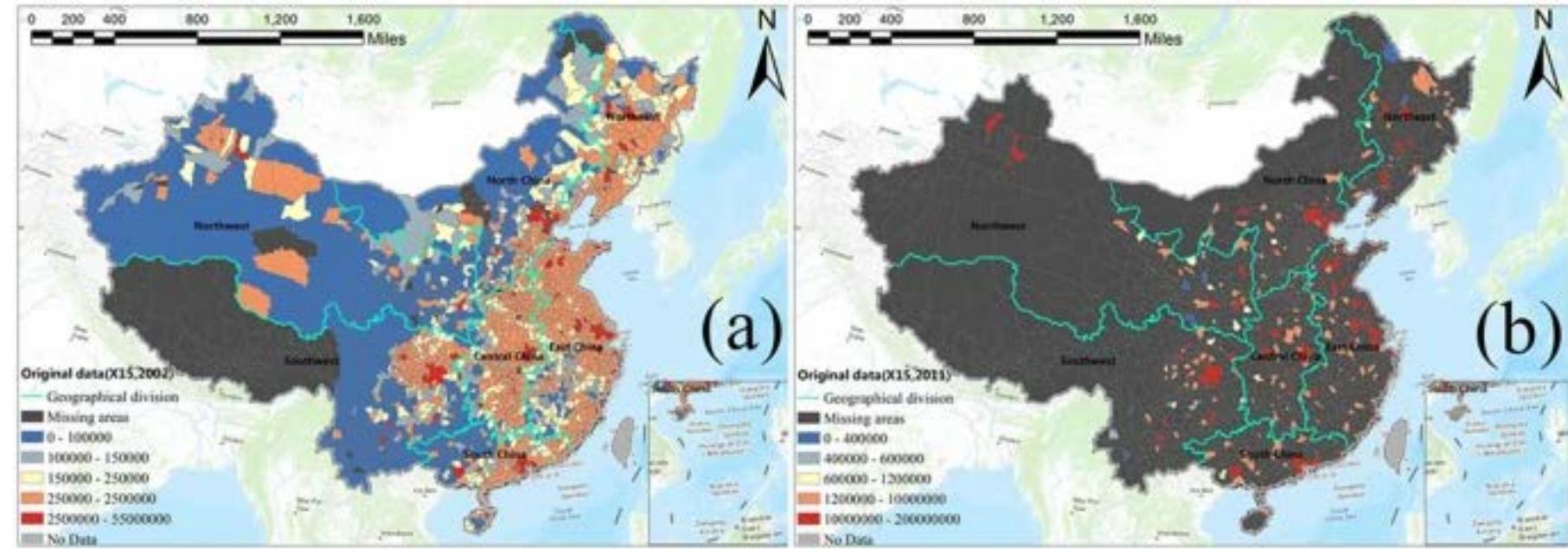


Схожесть зависит от выбранной меры



Paparrizos, J. et al. Accelerating similarity search for elastic measures: A study and new generalization of lower bounding distances. Proc. VLDB Endow. 16(8), 2019–2032 (2023). DOI 10.14778/3594512.3594530. <https://www.vldb.org/pvldb/vol16/p2019-paparrizos.pdf>

Пропущенные значения временных рядов



Доля провинций Китая, не предоставившие данные по одному атрибуту для гос. стат. отчета^{*}

a) 2002: менее 15%

b) 2011: более 85%

* Song C. et al. Estimating missing values in China's official socioeconomic statistics using progressive spatiotemporal Bayesian hierarchical modeling. Sci. Rep. 2018. Vol. 8, article 10055. DOI: [10.1038/s41598-018-28322-z](https://doi.org/10.1038/s41598-018-28322-z)

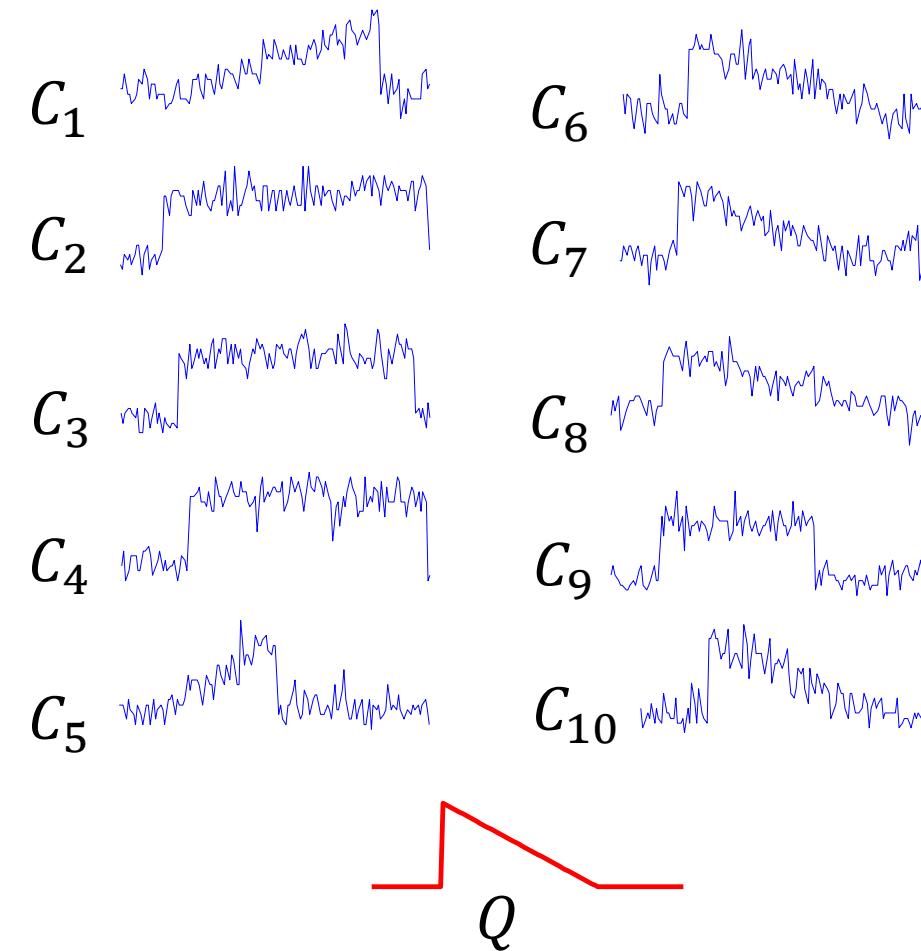
Содержание

- Понятие временного ряда
- Временные ряды в различных предметных областях
- Особенности интеллектуального анализа временных рядов
- **Основные задачи анализа временных рядов**
- Определения и нотация

Базовые задачи анализа временных рядов

- Поиск по образцу
- Поиск аномалий
- Поиск шаблонов
- Восстановление пропущенных значений
- Прогноз
- Классификация
- Кластеризация

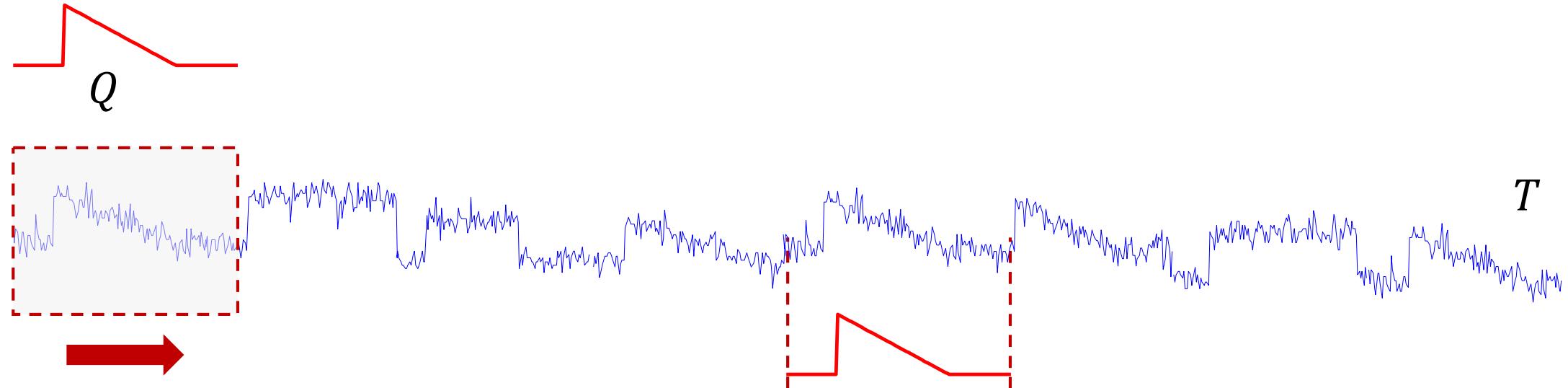
Поиск по образцу: случай нескольких временных рядов (whole matching)



В заданном множестве рядов $C = \{C_1, \dots, C_n\}$ найти ряд $C_{\text{bestmatch}}$, наиболее похожий на заданный запрос Q :

$$\forall C_i \in C \quad \text{Dist}(C_{\text{bestmatch}}, Q) \leq \text{Dist}(C_i, Q)$$

Поиск по образцу: случай подпоследовательностей временного ряда (subsequence matching)

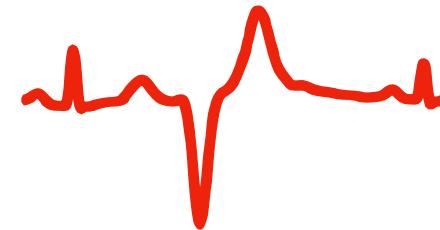


В заданном ряде $T = \{C_1, \dots, C_n\}$ найти подпоследовательность $C_{\text{bestmatch}}$, наиболее похожую на заданный запрос Q :

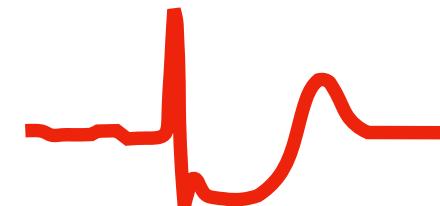
$$\forall T_{i,m} \in S_T^m \quad \text{Dist}(C_{\text{bestmatch}}, Q) \leq \text{Dist}(C, Q)$$

Поиск по образцу: выявление заболеваний по ЭКГ

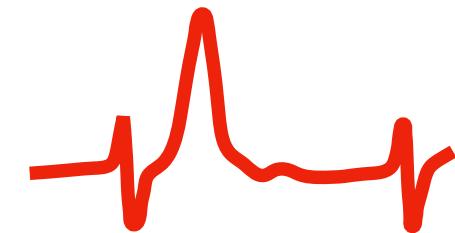
Преждевременное
сокращение желудочков



Инфаркт



Гиперкалиемия



Поиск по образцу: генетика*

Хромосома человека:

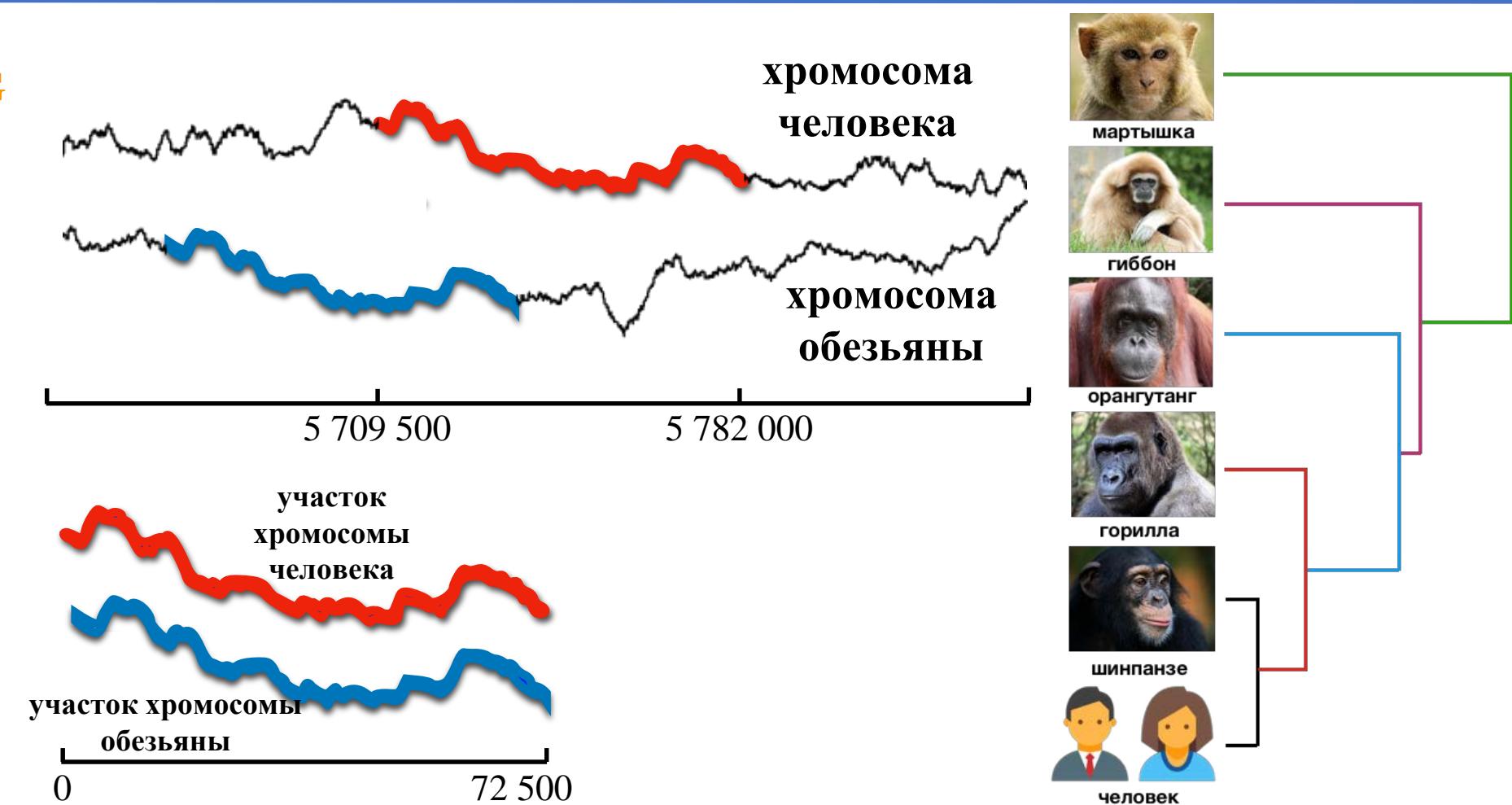
GTCAAT...AAGAGATTTG

Хромосома обезьяны:

GGCAAT...ACAGATTGGA

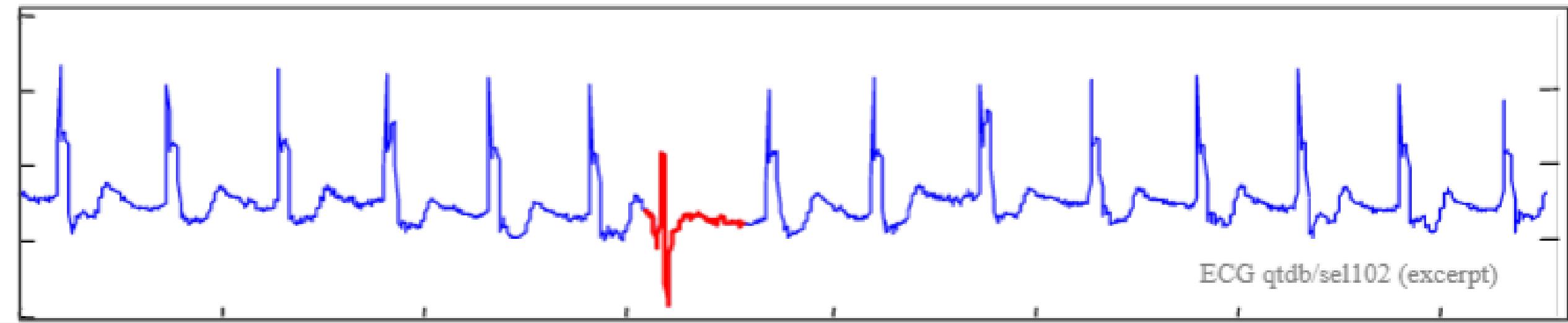
**Трансформация
цепочки ДНК
во временной ряд**

```
t1 := 0
for i ∈ 1..|DNAstr| do
    case DNAstr of
        A : ti+1 := ti+2
        G : ti+1 := ti+1
        C : ti+1 := ti-1
        T : ti+1 := ti-2
    end
```



* Rakthanmanon T. et al. Addressing big data time series: Mining trillions of time series subsequences under Dynamic Time Warping. ACM TKDD. 2013. 7(3). 10. <https://doi.org/10.1145/2500489>

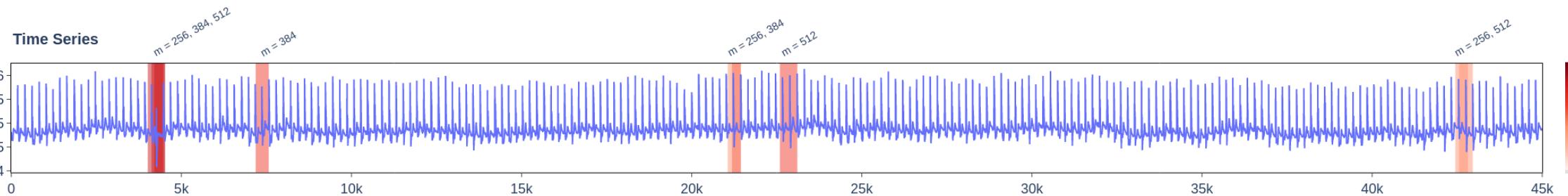
Поиск аномалий временного ряда



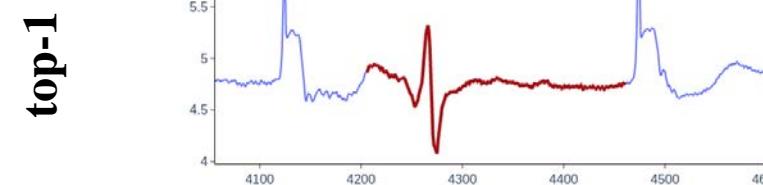
В заданном временном ряде найти подпоследовательность, наиболее непохожую на все остальные подпоследовательности ряда

Поиск аномалий

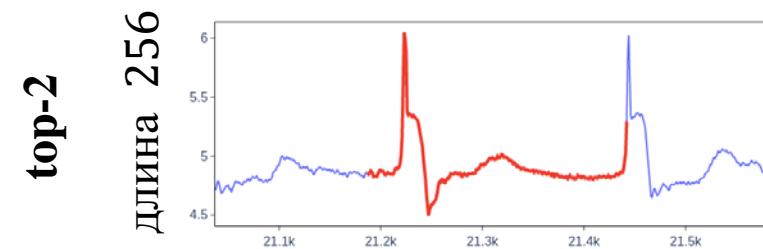
ЭКГ
взрослого
пациента



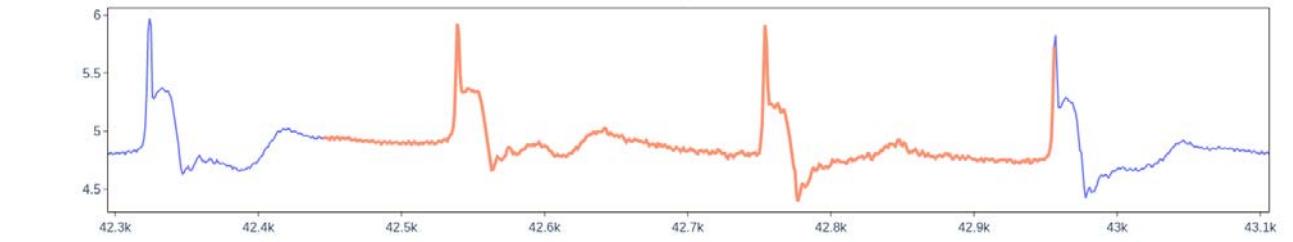
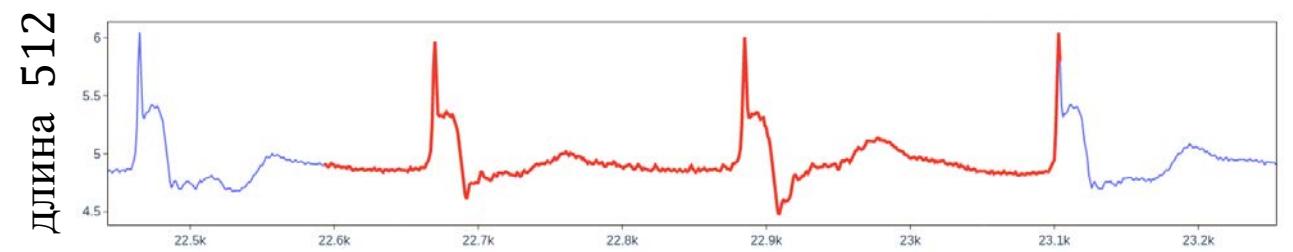
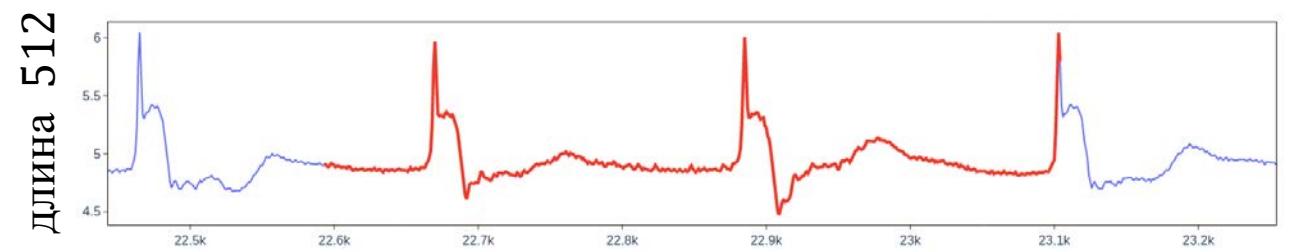
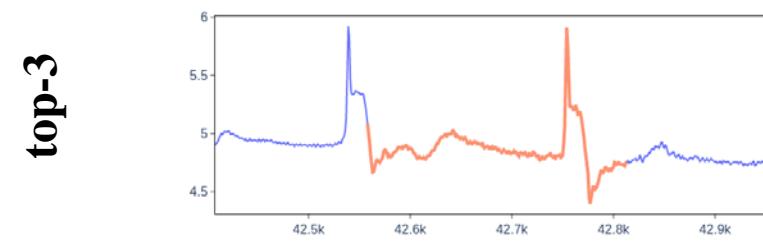
Прежде-
временное
сокращение
желудочков



Эктопическое
сердцебиение

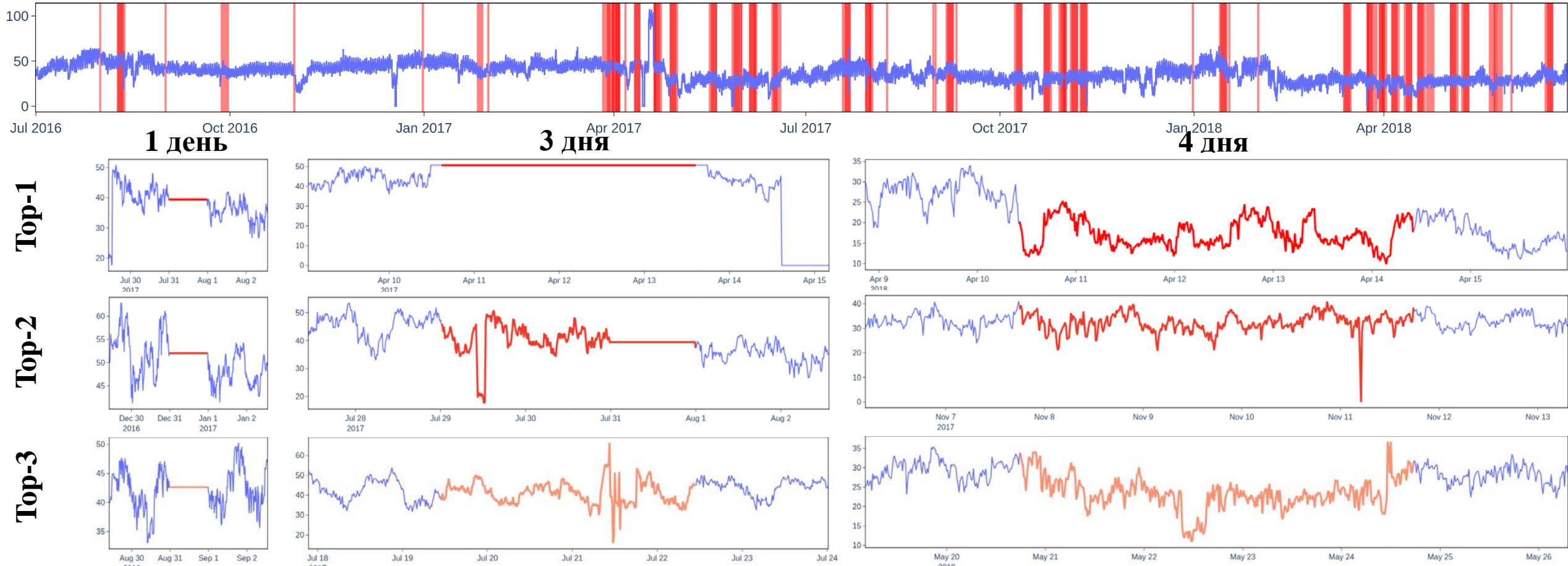


Эктопическое
сердцебиение



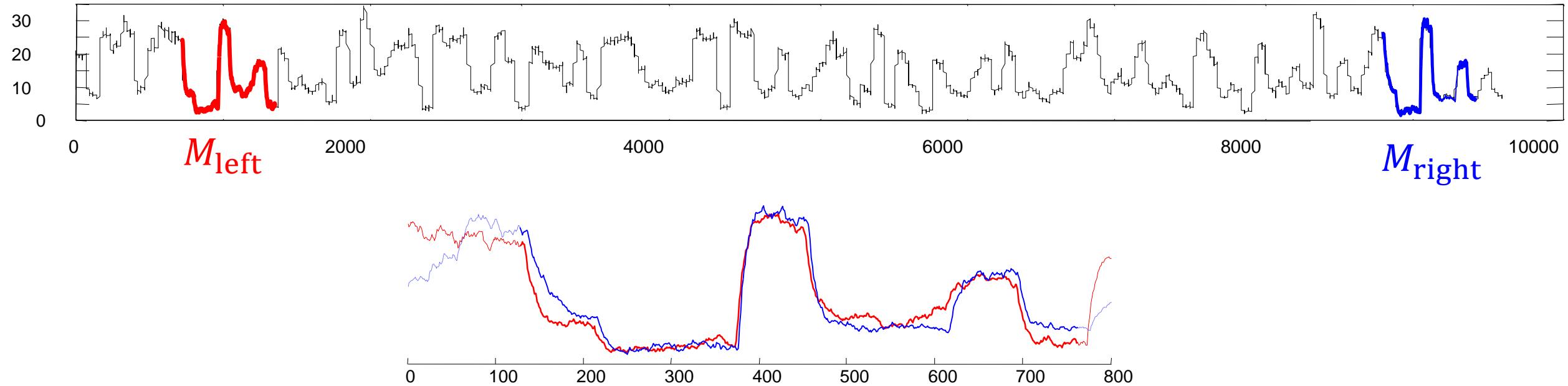
Поиск аномалий

Двухгодичное энергопотребление в Китае*



* Zhou H. et al. Informer: beyond efficient transformer for long sequence time-series forecasting. AAAI 2021: 11106-11115. DOI: [10.1609/aaai.v35i12.17325](https://doi.org/10.1609/aaai.v35i12.17325).

Поиск шаблонов: мотивы (motifs)



Пара непересекающихся подпоследовательностей ряда равной длины, наиболее похожих друг на друга:

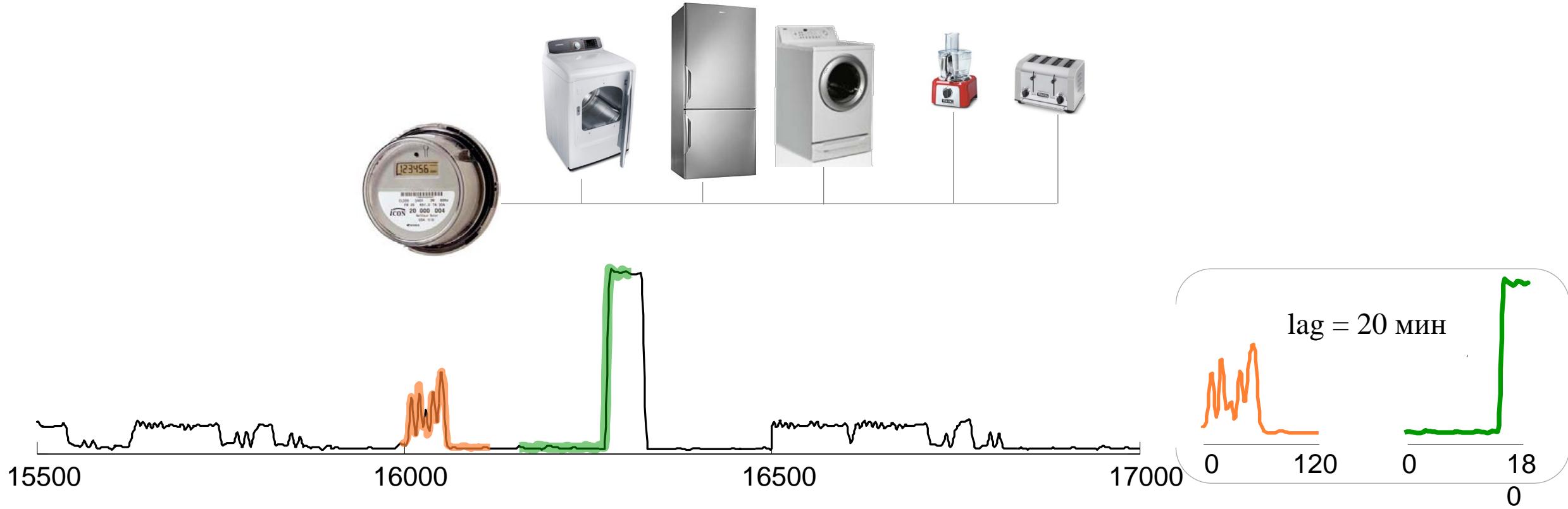
$$\forall C_i, C_j \quad \text{Dist}(M_{\text{left}}, M_{\text{right}}) \leq \text{Dist}(C_i, C_j)$$

Поиск шаблонов: мотивы (motifs)



* Zhou H. et al. Informer: beyond efficient transformer for long sequence time-series forecasting. AAAI 2021: 11106-11115. DOI: [10.1609/aaai.v35i12.17325](https://doi.org/10.1609/aaai.v35i12.17325).

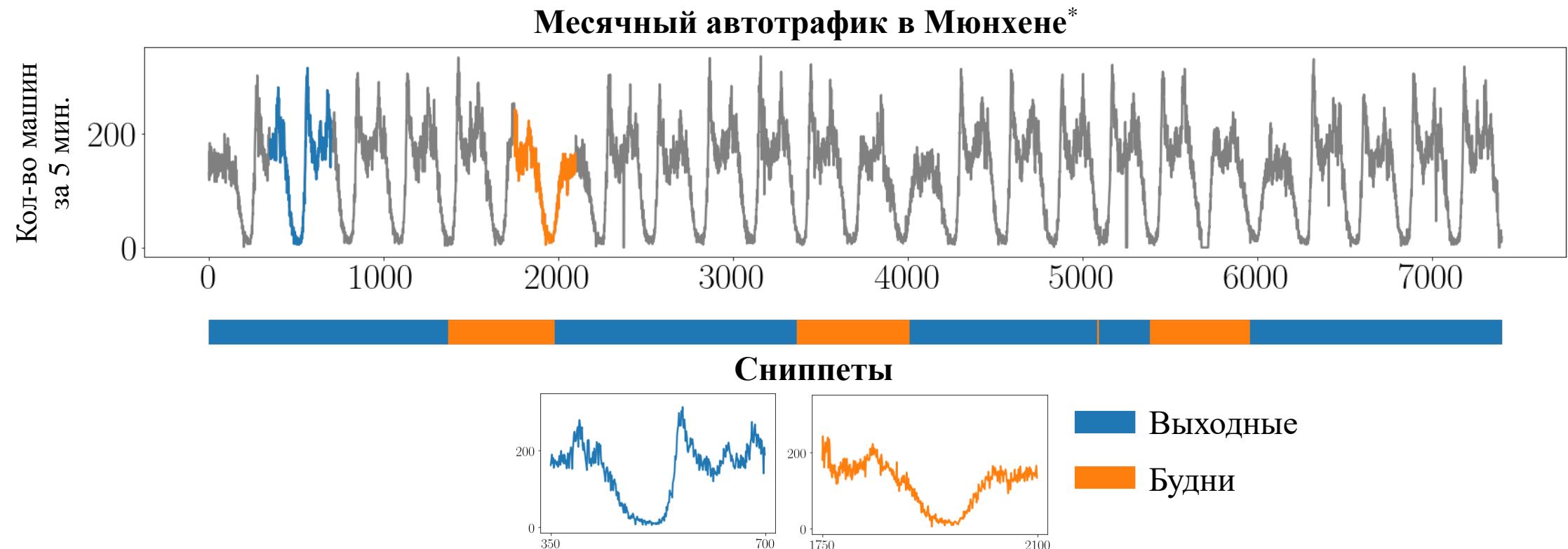
Поиск шаблонов: ассоциативные правила (association rules)



IF *работает стиральная машина*

THEN не более чем через 20 мин. *работает сушильная машина*

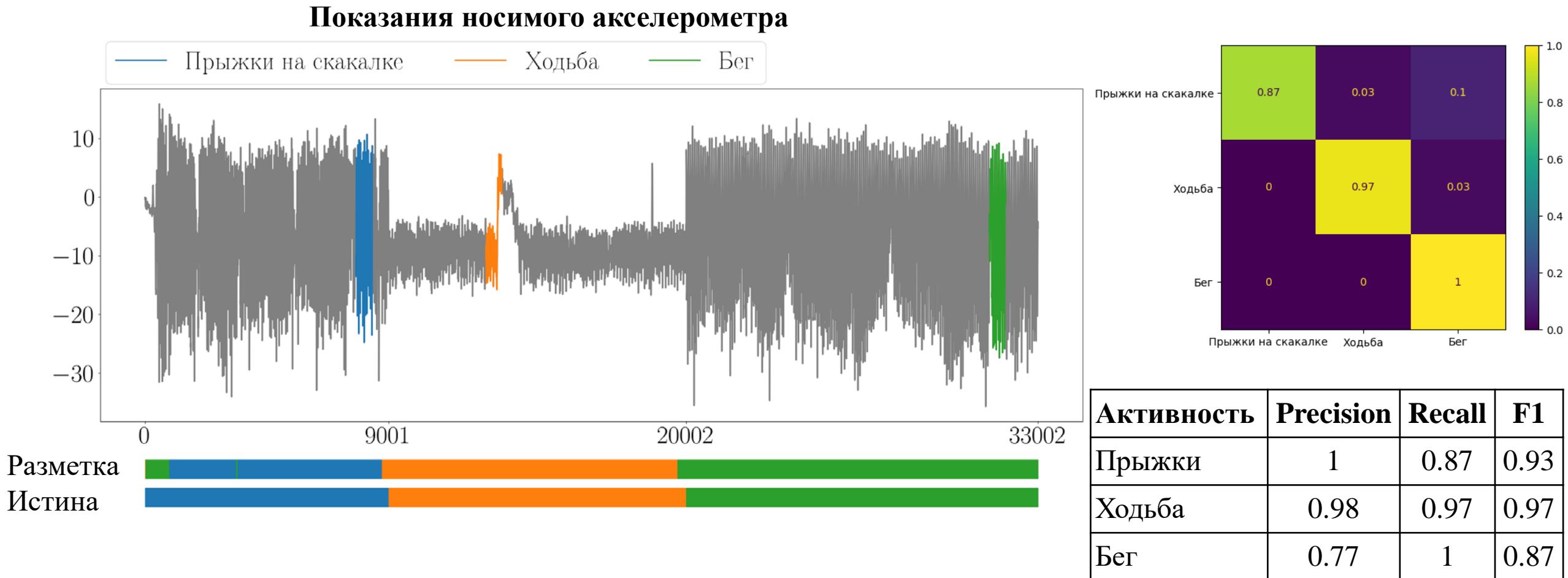
Поиск шаблонов: сниппеты (snippets)



Множество подпоследовательностей ряда, выражающих типичные активности субъекта

* Public (anonymized) road traffic prediction datasets from Huawei Munich Research Center. URL: <https://zenodo.org/record/3653880#.Y0zZi3ZBxPa>

Поиск шаблонов: сниппеты

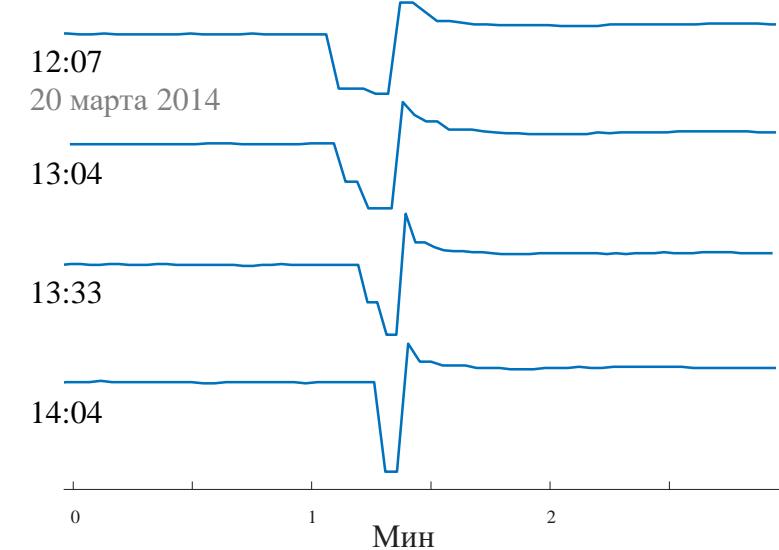


* Reiss A., Stricker D. Introducing a new benchmarked dataset for activity monitoring. ISWC 2012, Newcastle, UK, June 18-22, 2012. 108–109. IEEE (2012). doi: [10.1109/ISWC.2012.13](https://doi.org/10.1109/ISWC.2012.13)

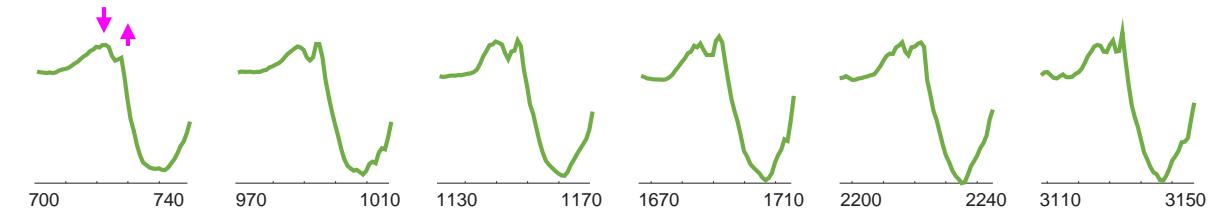
Поиск шаблонов: цепочки (chains)



Энергопотребление холодильника

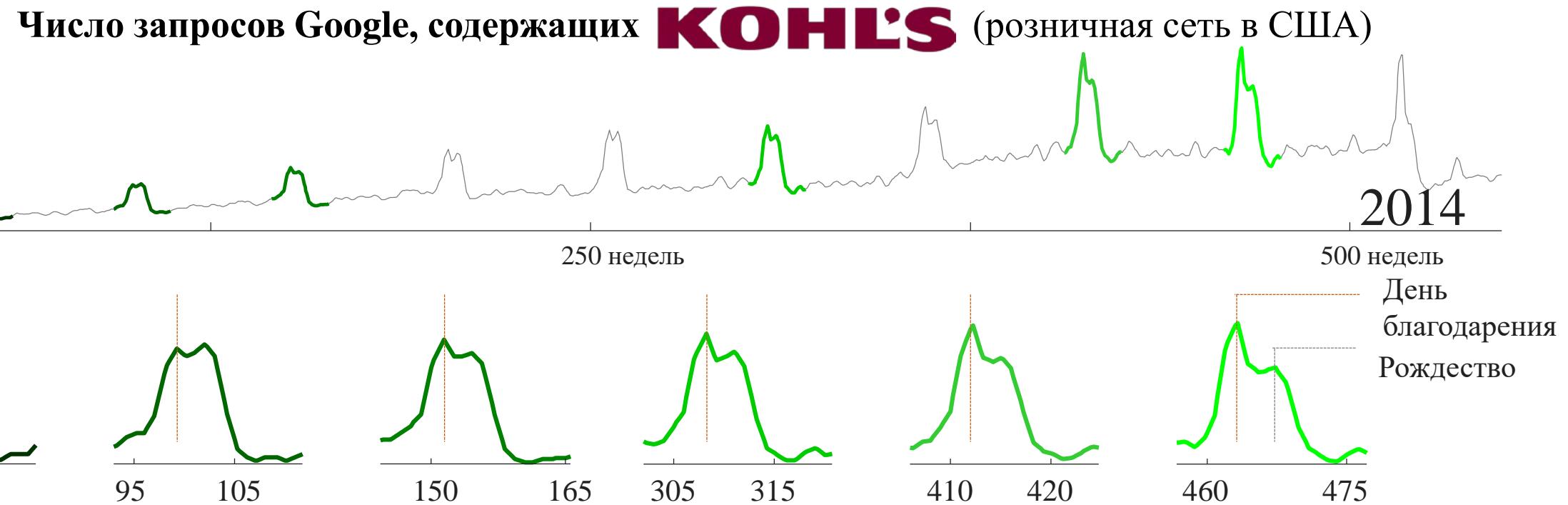


Запись датчика с левой икры спортсмена,
когда он начал бег трусцой на беговой дорожке



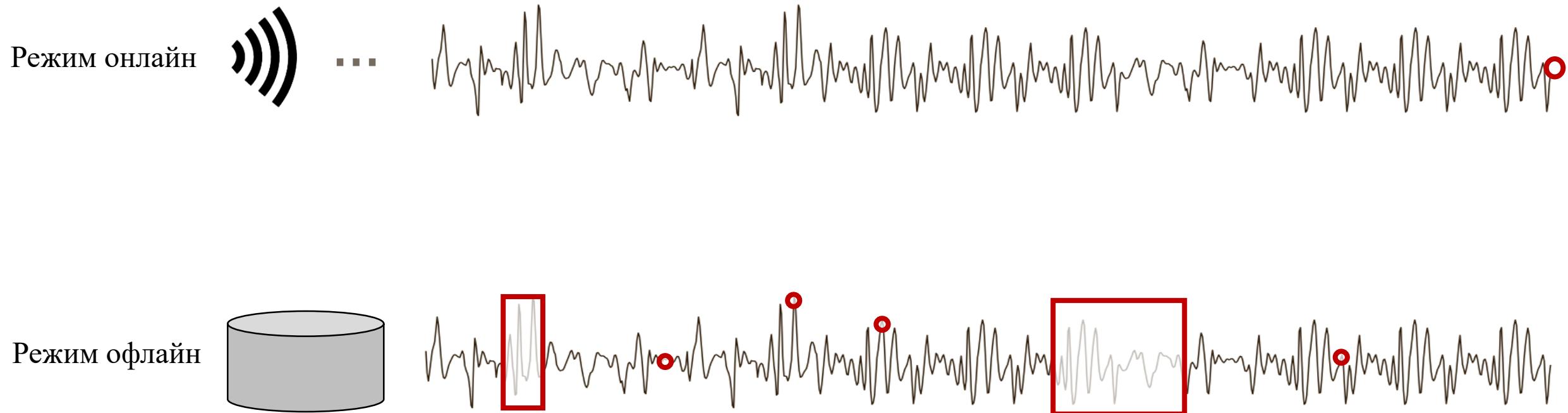
Цепочка подпоследовательностей ряда,
звенья которой отражают эволюцию некоего процесса

Поиск шаблонов: цепочки (chains)



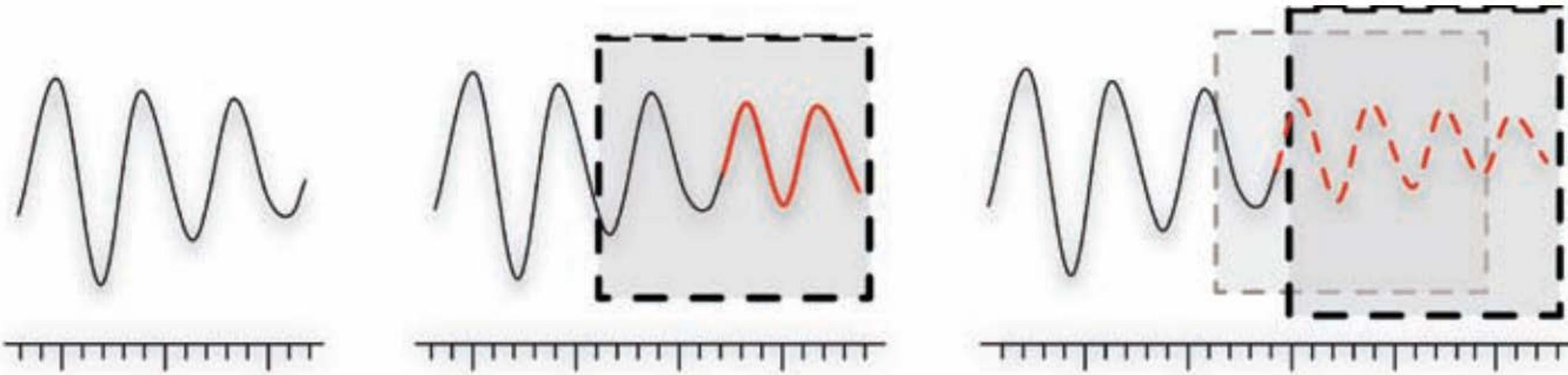
- Рост важности **Киберпонедельника** (понедельник после Дня благодарения): за 10 лет выпуклость меняется от плавной и занимающей больший период между Днем благодарения и Рождеством к резкой и сосредоточенной на Дне благодарения
- Термин введен в пресс-релизе “Киберпонедельник становится одним из крупнейших дней онлайн-покупок в году” 28 ноября 2005 г., дата которого совпадает с первым проблеском острого пика в цепочке

Восстановление пропущенных значений ряда (imputation/recovery)



Синтез отсутствующих значений ряда

Прогнозирование временного ряда (forecast)



Исходный ряд
(периодическая структура,
поддающаяся прогнозу)

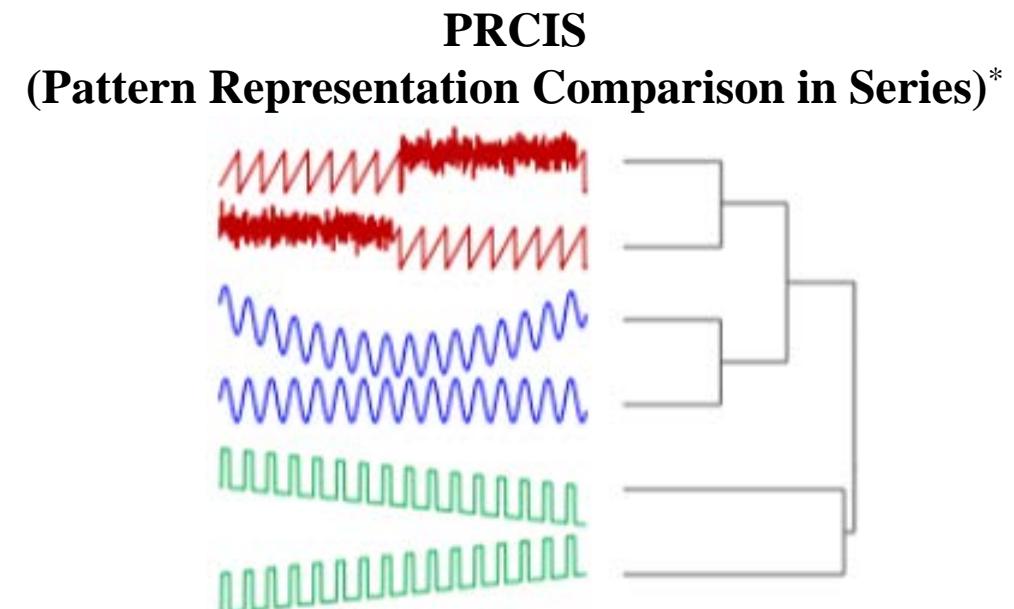
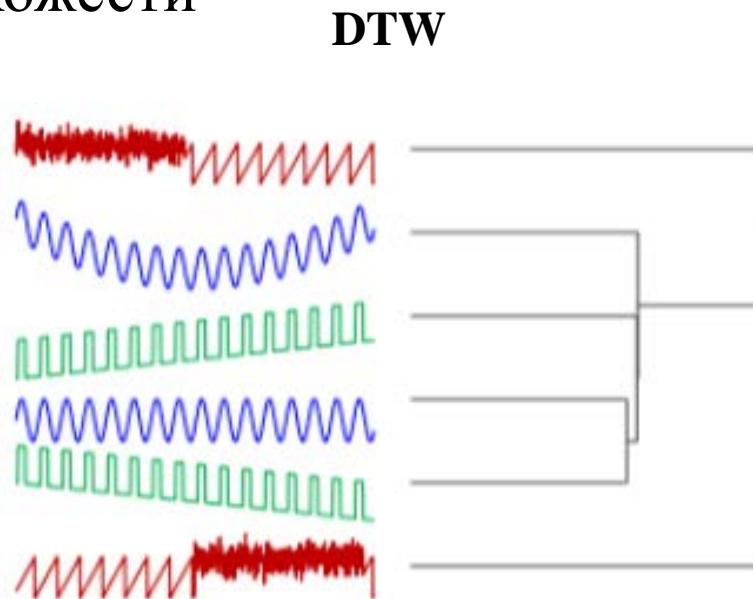
Прогноз точек данных
в пределах
окна прогнозирования

Долгосрочный прогноз:
использование более ранних прогнозных
значений в качестве входных данных прогноза

Синтез будущих значений ряда

Классификация и кластеризация временных рядов

- При небольшой длине рядов можно использовать стандартные алгоритмы машинного обучения и адекватные функции для вычисления схожести (например, Dynamic Time Warping)
- Для длинных временных рядов нужны специализированные функции для вычисления схожести



* Der A. et al. Matrix Profile XXVII: A Novel Distance Measure for Comparing Long Time Series. ICKG 2022. P. 40-47. <https://doi.org/10.1109/ICKG55886.2022.00013>

Классификация подпоследовательностей ряда возможна, но их кластеризация БЕССМЫСЛЕННА*

- Подпоследовательности одного временного ряда обычно сильно коррелируют между собой, что делает их неинформативными для кластеризации
- Подпоследовательности разных временных рядов обычно имеют различные характеристики и паттерны, что позволяет выделить более информативные признаки и получить осмысленный результат кластеризации
- Пример: мониторинг температуры в помещении
 - Если температура в помещении измеряется каждые 5 мин., то подпоследовательности измерений за последний час будут сильно коррелировать между собой, так как температура в помещении обычно меняется медленно и плавно
 - Кластеризация подпоследовательностей измерений за последний час не будет иметь смысла, так как они будут очень похожи друг на друга и не будут содержать достаточно информации для кластеризации
 - Для кластеризации нужно использовать подпоследовательности измерений за разные периоды времени (за последние сутки, неделю, месяц и др.)

* Keogh E., Lin J. Clustering of time-series subsequences is meaningless: implications for previous and future research. *Knowl. Inf. Syst.* 8(2). 2005. 154-177.
DOI: [10.1007/s10115-004-0172-7](https://doi.org/10.1007/s10115-004-0172-7)

Содержание

- Понятие временного ряда
- Временные ряды в различных предметных областях
- Особенности интеллектуального анализа временных рядов
- Основные задачи анализа временных рядов
- **Определения и нотация**

(Одномерный) временной ряд (univariate time series)

- Конечная последовательность хронологически упорядоченных вещественных значений

$$T = (t_1, \dots, t_n), \quad t_i \in \mathbb{R}$$

альтернативная запись: $T = \{t_i\}_{i=1}^n$

n – длина ряда, $|T| = n$

- Точки ряда ассоциированы с временными метками, сделанными через **равные промежутки** (частота измерений фиксирована)
- Значения временных меток могут не подвергаться обработке или отсутствовать в исходных данных

Подпоследовательность (subsequence)

- Непрерывный промежуток временного ряда, имеющий заданную длину

$$T_{i,m} = (t_i, \dots, t_{i+m-1}), \quad 3 \leq m \ll n, \quad 1 \leq i \leq n - m + 1$$

альтернативная запись: $T_{i,m} = \{t_k\}_{k=i}^m$

- Другой термин – *скользящее окно (sliding window)* длины m
- Множество всех подпоследовательностей ряда, имеющих заданную длину

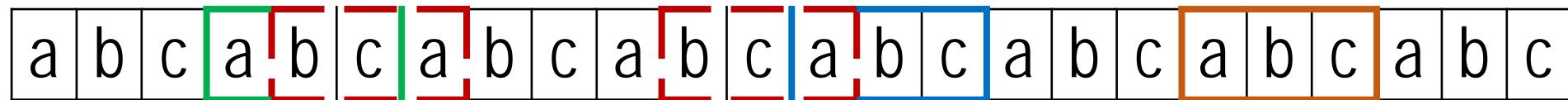
$$S_T^m, \quad |S_T^m| = n - m + 1$$

Почему подпоследовательности важны

- Ряд анализируется как множество подпоследовательностей заданной длины
- Длина подпоследовательности – параметр, задаваемый экспертом
 - $m \geq 3$: подпоследовательности из 1-2 точек не имеют смысла
 - $m \ll n$: подпоследовательности на порядки короче, чем ряд
- Длина подпоследовательности существенным и не всегда предсказуемым образом влияет на результат анализа
 - Если $T_{i,m}$ – аномалия, то не факт, что $T_{i,m-1}$ или $T_{i,m+1}$ тоже аномалии
 - Если $\{T_{l,m}, T_{r,m}\}$ – мотив, то не факт, что $\{T_{l,2m}, T_{r,2m}\}$ тоже мотив

Тривиальные совпадения подпоследовательностей (trivial matches)

- Подпоследовательности $T_{i,m}$ и $T_{j,m}$ **не являются** тривиальными совпадениями друг друга, если $|i - j| > \xi m$, где $0 < \xi \leq 1$ – задаваемый экспертом параметр. Типичные значения: $\xi \in \{0.25, 0.5, 1\}$. Далее полагаем $\xi = 1$
- Длина подпоследовательности m задает минимально значимый временной промежуток в предметной области, поэтому $T_{i,m}$ и $T_{i \pm \xi m, m}$ будут очень похожи
- Однако $T_{i,m}$ и $T_{i \pm \xi m, m}$ не представляют ценности как результаты анализа ряда, поскольку относятся практически к одному и тому же моменту времени
- M_C – множество подпоследовательностей ряда, **не** являющихся тривиальным совпадением подпоследовательности C



Подпоследовательность-ближайший сосед (Nearest Neighbor)

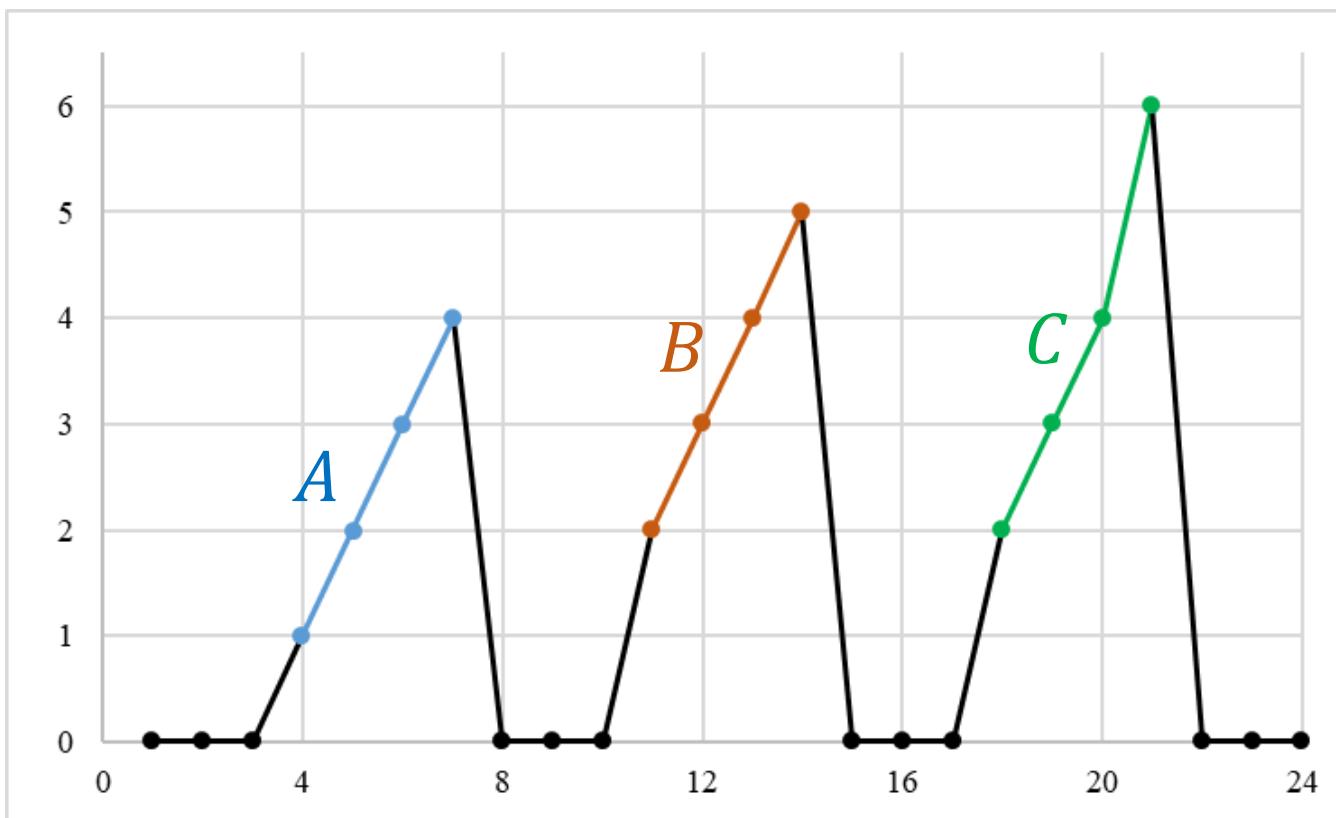
- (*1-й*) ближайший сосед данной подпоследовательности – подпоследовательность ряда, которая наиболее похожа на нее и не является ее тривиальным совпадением

$$\theta_{1\text{NN}}^T(Q, N) = \text{TRUE} \Leftrightarrow N = \arg \min_{C \in M_Q} \text{Dist}(Q, C)$$

- Обобщение для случая k соседей: $\theta_{k\text{NN}}^T(\cdot, \cdot)$
- Обобщение для случая двух рядов: $\theta_{k\text{NN}}^{T1, T2}(A \in T1, B \in T2)$
(нет условия недопустимости тривиального совпадения)

Функция $\theta_{1\text{NN}}^T(\cdot, \cdot)$ не коммутативна

Если N – ближайший сосед Q , то не факт, что Q – ближайший сосед N



$\text{ED}(\cdot, \cdot)$	A	B	C
A	0	2	$\sqrt{7}$
B	2	0	1
C	$\sqrt{7}$	1	0

$\theta_{1\text{NN}}(\cdot, \cdot)$	A	B	C
A	TRUE	TRUE	FALSE
B	FALSE	TRUE	TRUE
C	FALSE	TRUE	TRUE

Почему ближайшие соседи важны

- Решение любой задачи интеллектуального анализа временного ряда связано с обработкой подпоследовательностей и поиском их ближайших соседей
 - шаблонная (повторяющаяся) подпоследовательность имеет наиболее похожего на нее ближайшего соседа
 - аномальная подпоследовательность имеет наиболее непохожего на нее ближайшего соседа
 - поиск наиболее похожей подпоследовательности – по сути, поиск ее ближайшего соседа
 - классификация подпоследовательностей использует поиск их ближайших соседей
 - ...

Потоковый временной ряд (streaming time series)

- Бесконечная упорядоченная последовательность вещественных значений, которые поступают непрерывно одно за другим в режиме реального времени

$$T = (t_1, \dots, t_n, \dots), \quad t_i \in \mathbb{R}$$

альтернативная запись: $T = \{t_i\}_{i=1}^{\infty}$

- Режим реального времени предполагает конечный период времени обработки данных для *конкретной предметной области*:
реальное время \neq «очень быстро»

Многомерный временной ряд (multivariate time series)

- Состоит из логически связанных одномерных временных рядов (измерений), *синхронизированных по времени*

$$\mathbf{T} = [T^{(1)}, \dots, T^{(d)}]^T, \quad d > 1, \quad T^{(i)} = (t_1^{(i)}, \dots, t_n^{(i)}), \quad t_k^{(i)} \in \mathbb{R}$$

- Многомерная точка

$$\mathbf{t}_i = [t_i^{(1)}, \dots, t_i^{(d)}]^T$$

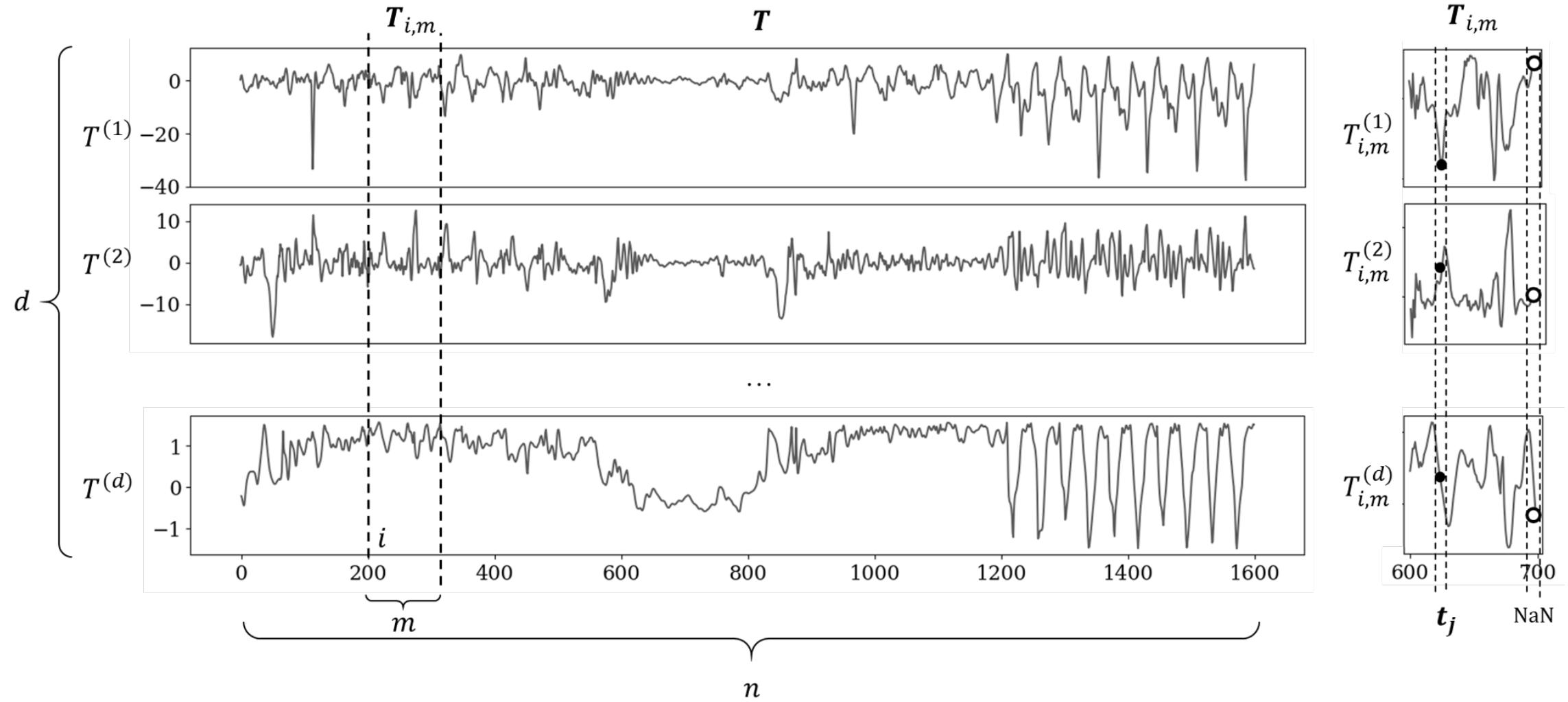
- Многомерная подпоследовательность

$$\mathbf{T}_{i,m} = [T_{i,m}^{(1)}, \dots, T_{i,m}^{(d)}]^T$$

- Множество подпоследовательностей

$$S_T^m = \bigcup_{k=1}^d S_{T^{(k)}}^m, \quad |S_T^m| = d(n - m + 1)$$

Многомерный временной ряд



Литература

1. Esling P., Agon C. Time-series Data Mining. ACM Comput. Surv. 2012. Vol. 45, No. 1. P. 12:1–12:34.
<https://doi.org/10.1145/2379776.2379788>.
2. Fu T.C. A review on time series data mining. Eng. Appl. of AI. 2011. Vol. 24, No. 1. P. 164–181.
<https://doi.org/10.1016/j.engappai.2010.09.007>.