

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



*Невольно изречешь: o tempora, o mores! —
Когда поразглядишь, какая в жизни горесть.*

Н.А. Некрасов

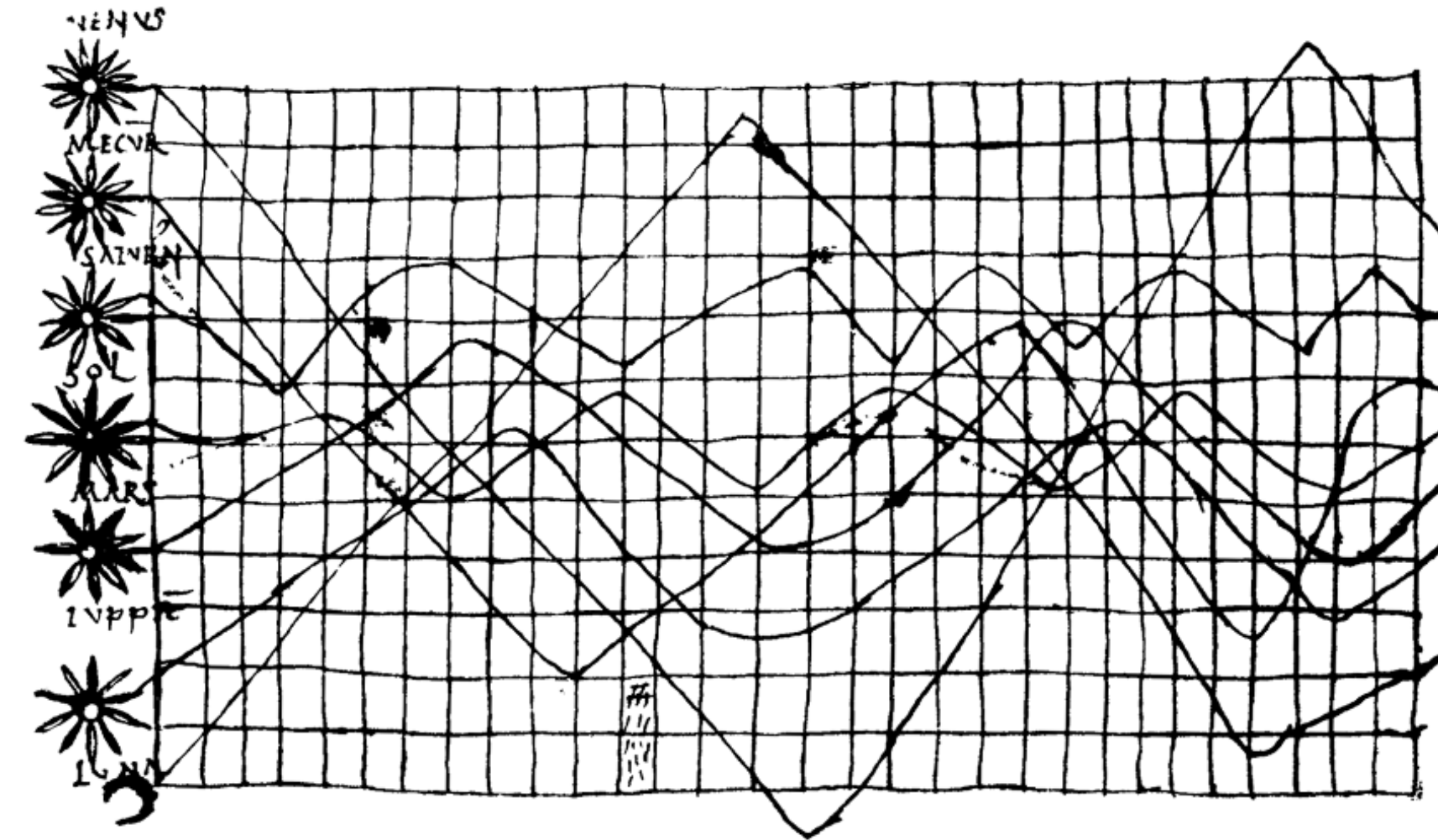
Содержание

- Временные ряды в различных предметных областях
- Определения и нотация
- Основные задачи анализа временных рядов

Люди измеряют всевозможные вещи, изменяющиеся во времени

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

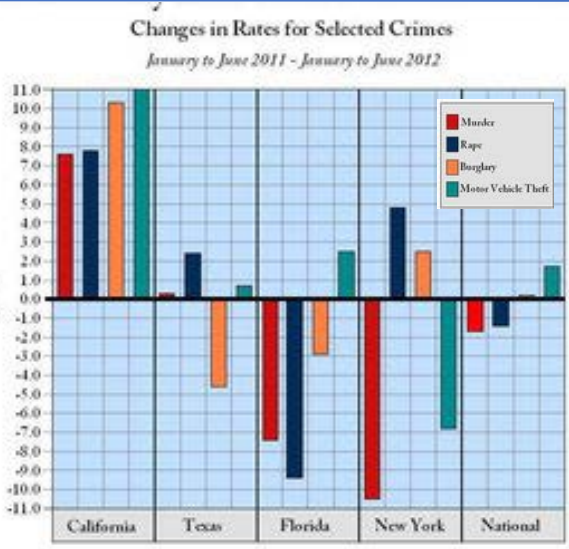
Временные ряды всегда...



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

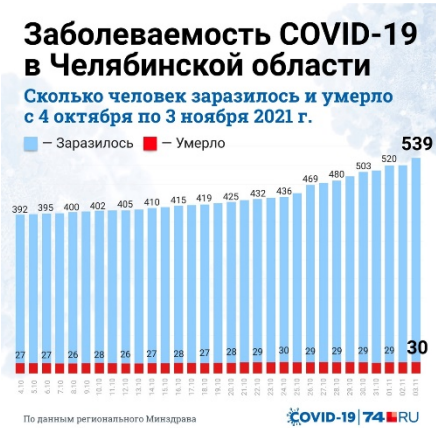
Tufte E. The Visual Display of Quantitative Information. Graphics Press, 2001. 200 p.

Временные ряды везде...

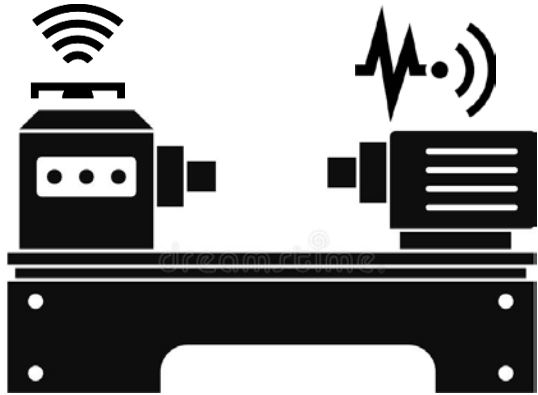


Случайная выборка из 4000 изображений в 15 газетах из различных стран за 1974–1989 гг.: более 75% изображений – это временные ряды

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



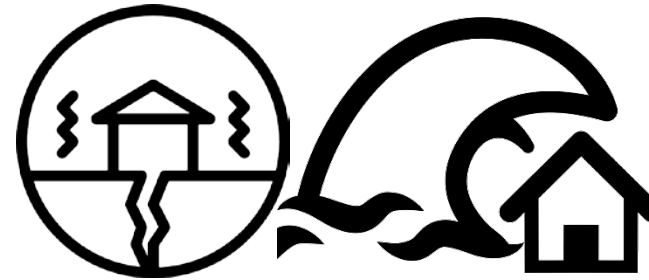
Временные ряды всюду...



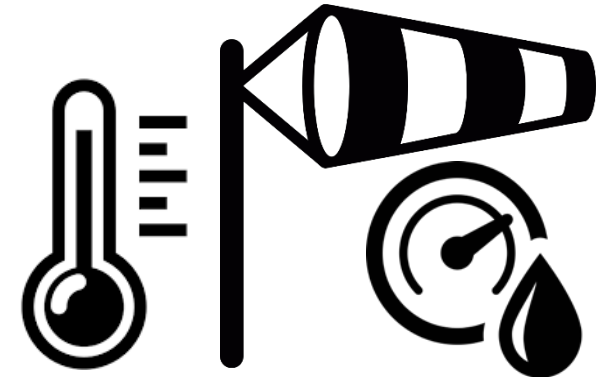
Умное производство,
предиктивное ТО



Интернет
вещей



Предсказание
природных катаклизмов



Предсказание погоды,
моделирование климата



Персональная
медицина



Био- и хемо-
информатика

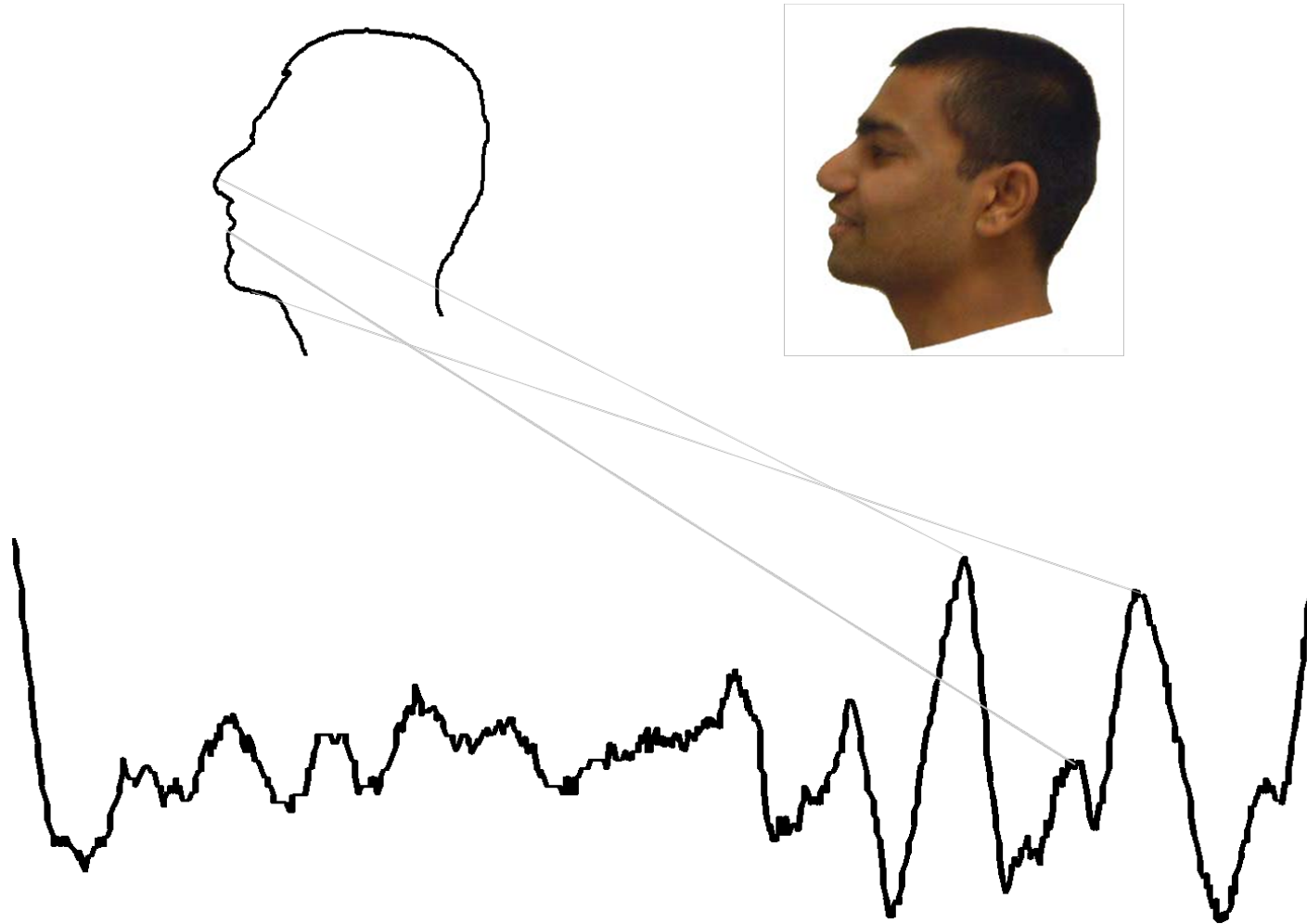


Экономика, бизнес,
финансы

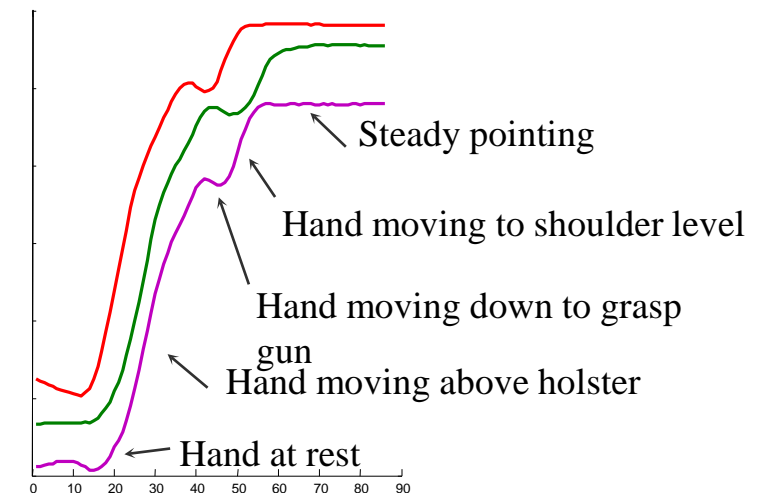
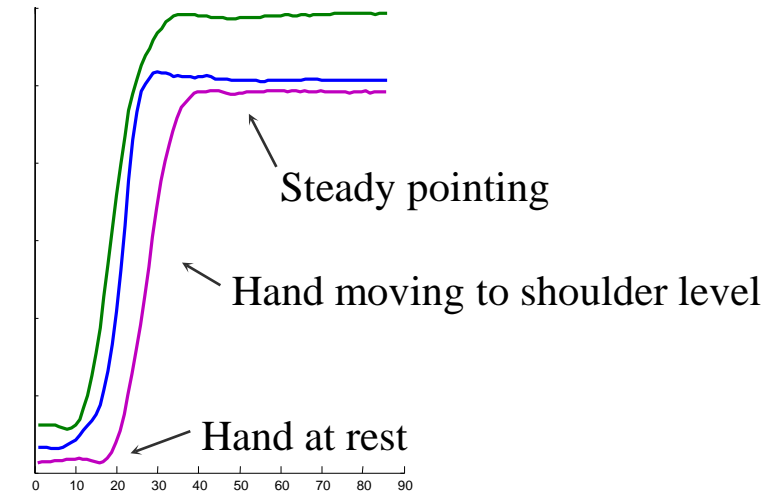


Системы электронного обучения,
персональные образов. траектории

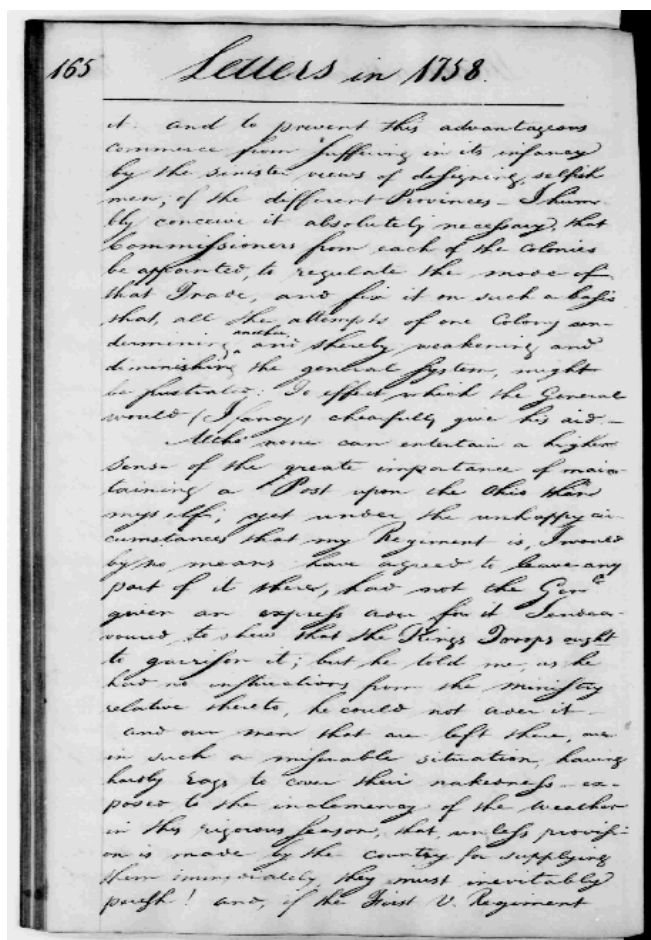
Изображение как временной ряд



Видео как временной ряд



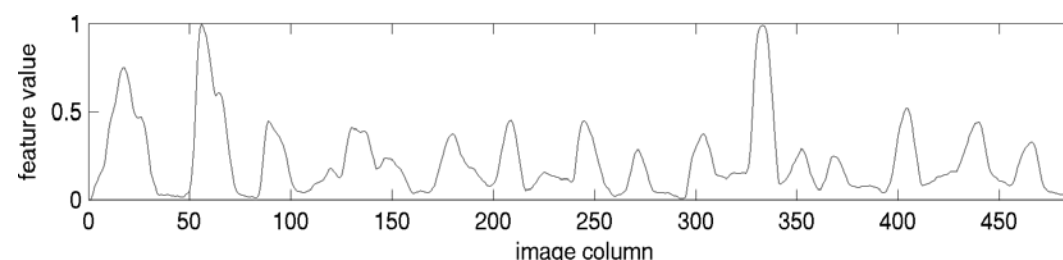
Рукописный текст как временной ряд



George Washington manuscript



George Washington
1732-1799



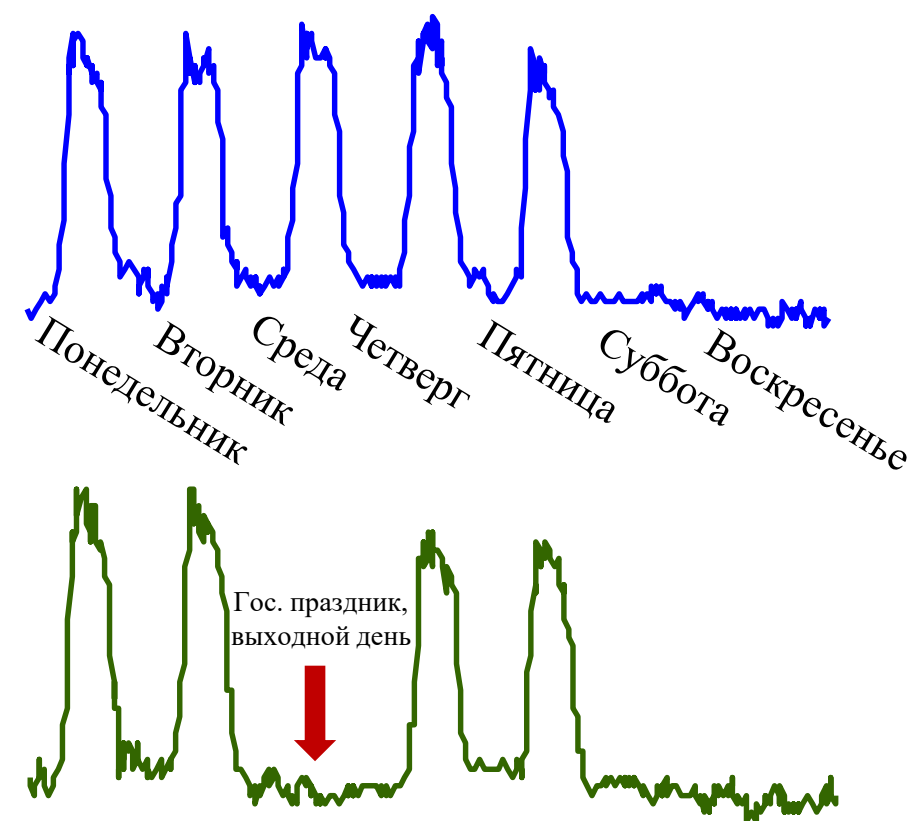
Почему временные ряды анализировать сложнее, чем другие данные

- Большая длина
- Субъективность схожести рядов
- Пропущенные значения
- Различные форматы данных и частоты снятия показаний, шумы

Большие временные ряды

Схожесть рядов определяется задачей и предметной областью

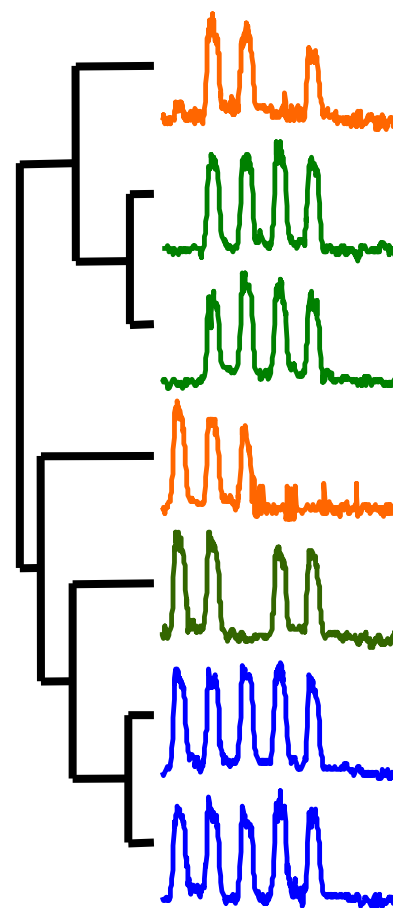
Недельное энергопотребление
вычислительного центра (Голландия, 1997)*



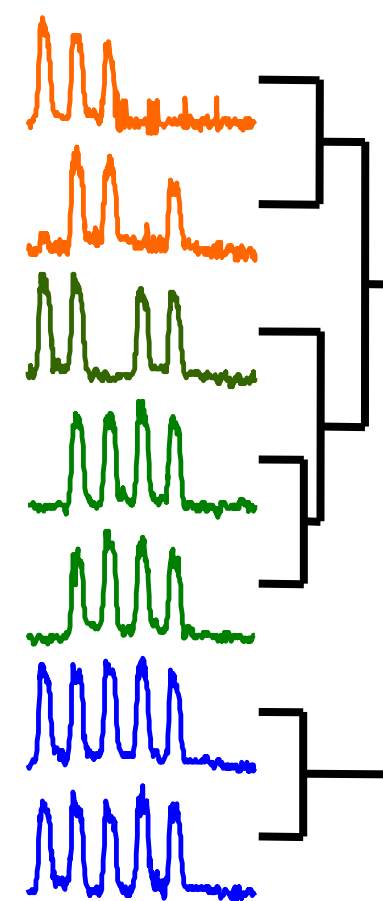
Количество
рабочих дней
в неделе:



Евклидово расстояние

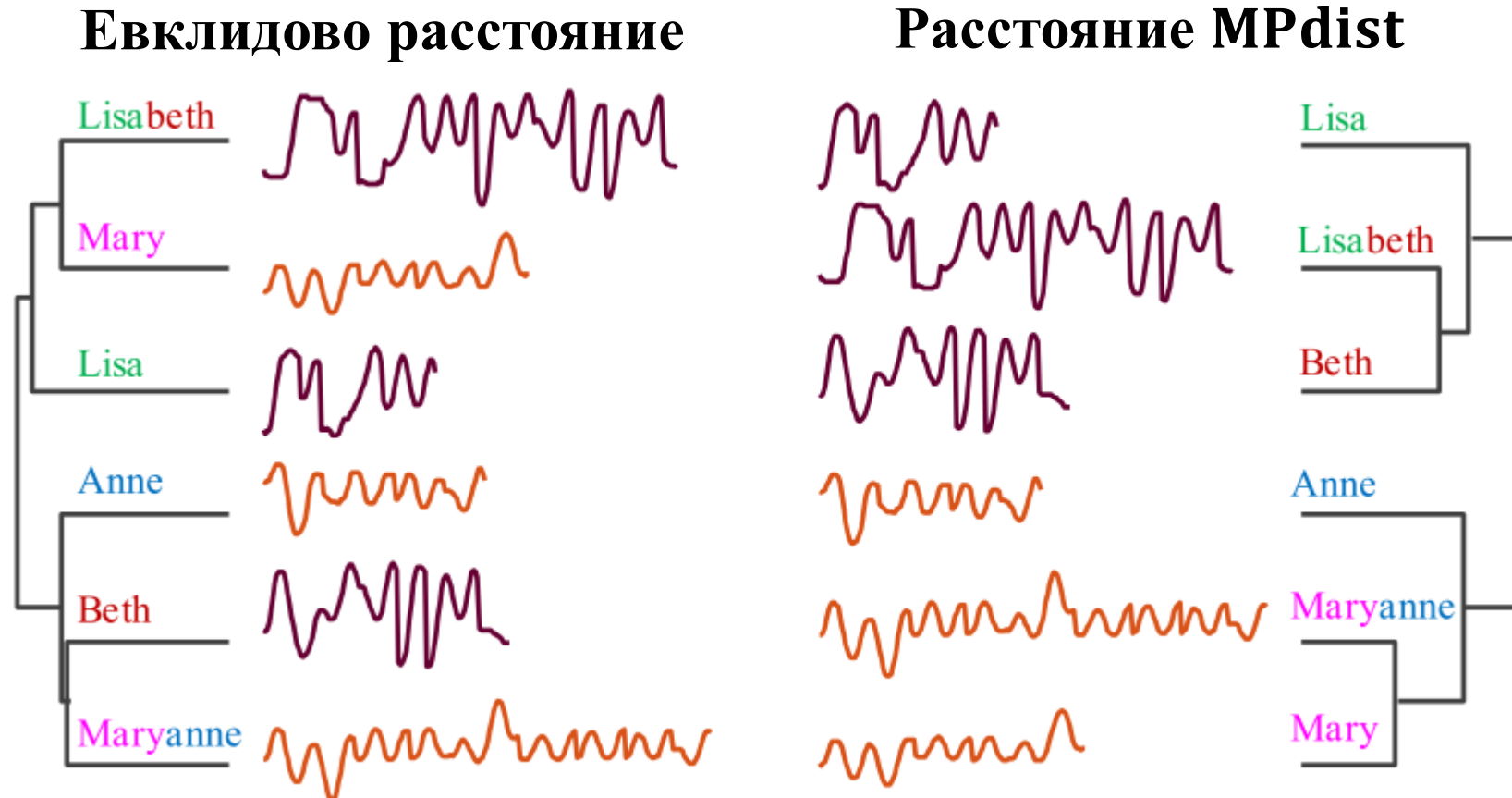


Dynamic Time Warping

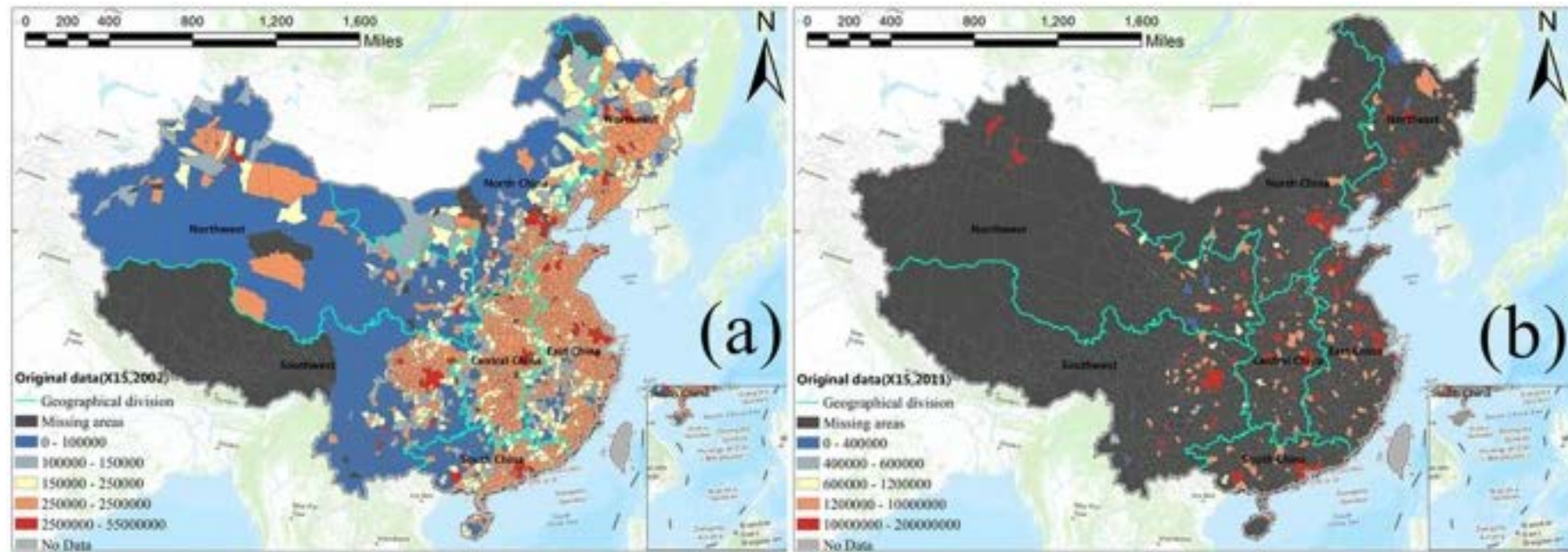


* van Wijk J.J., van Selow R.R. Cluster and calendar based visualization of time series data. INFOVIS 1999: 4-9. DOI: [10.1109/INFVIS.1999.801851](https://doi.org/10.1109/INFVIS.1999.801851)

Схожесть рядов определяется задачей и предметной областью



Пропущенные значения в временных рядах



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

a) 2002: менее 15%

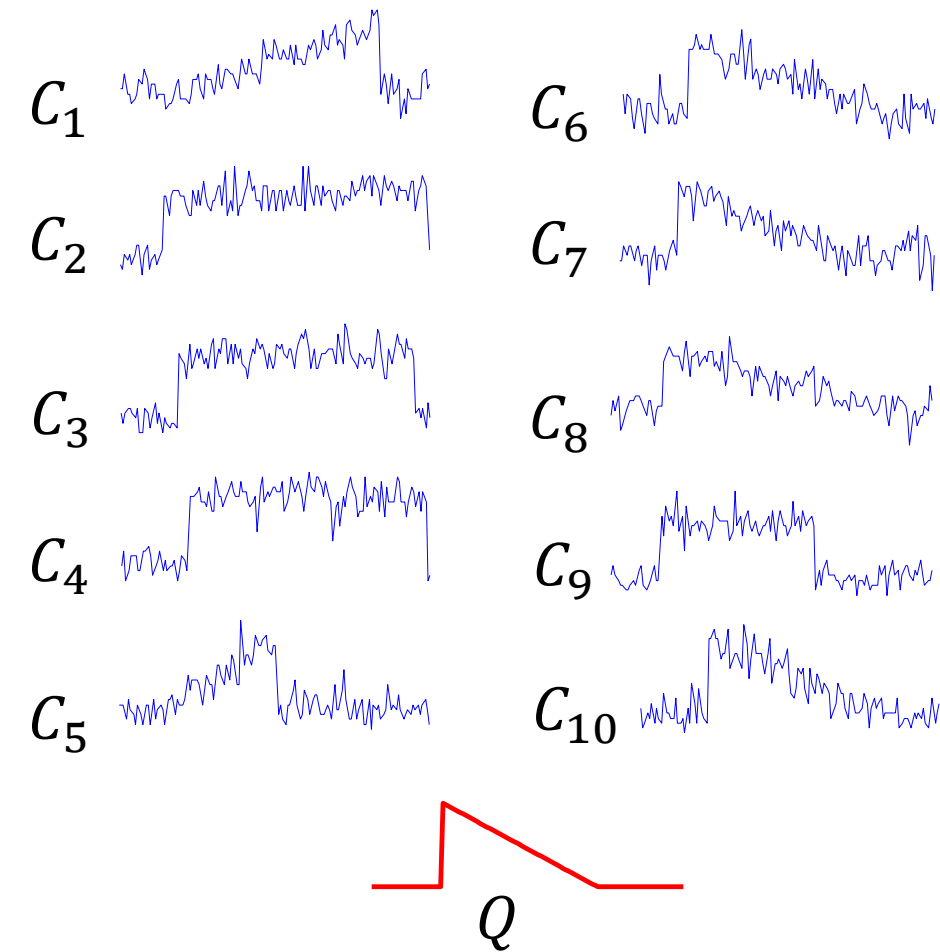
b) 2011: более 85%

* 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)

Базовые задачи анализа временных рядов

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

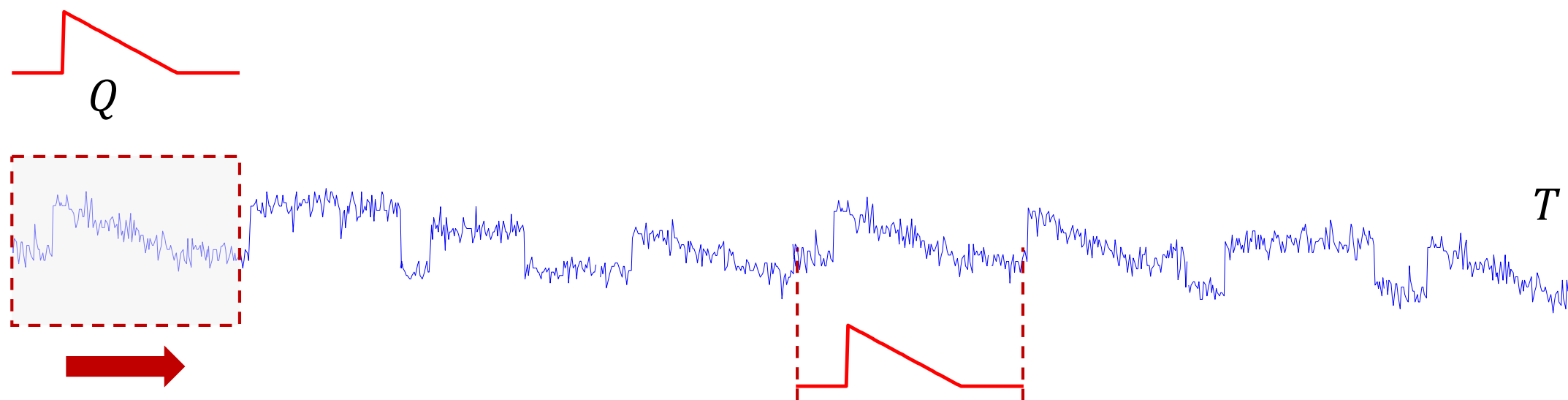
Поиск по образцу: случай нескольких временных рядов (whole matching)



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

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

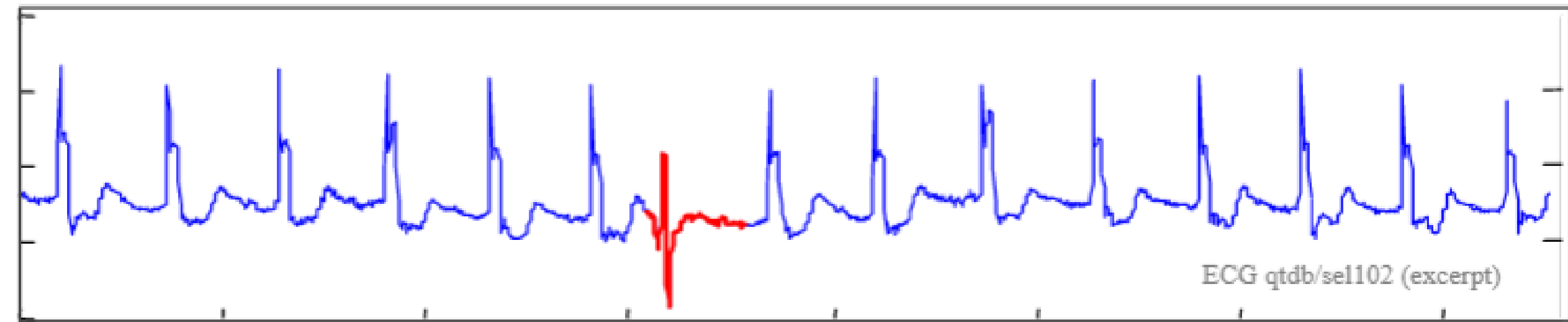
Поиск по образцу: случай подпоследовательностей временного ряда (subsequence matching)



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 аномалий временного ряда



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

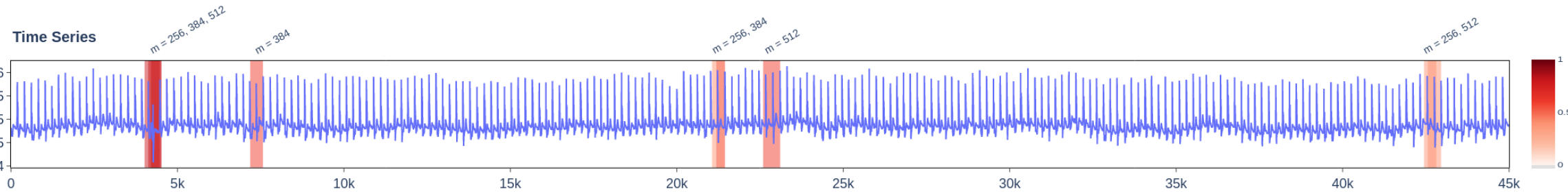
Поиск аномалий

ЭКГ
взрослого
пациента

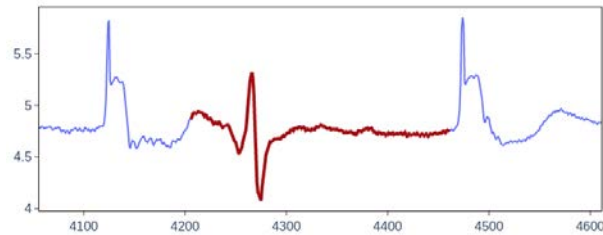
Прежде-
временное
сокращение
желудочков

Эктопическое
сердцебиение

Эктопическое
сердцебиение

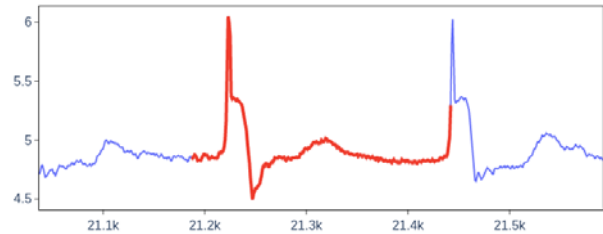


top-1

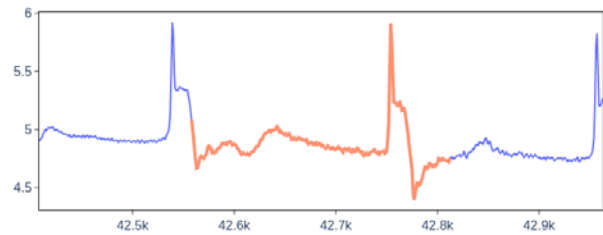


top-2

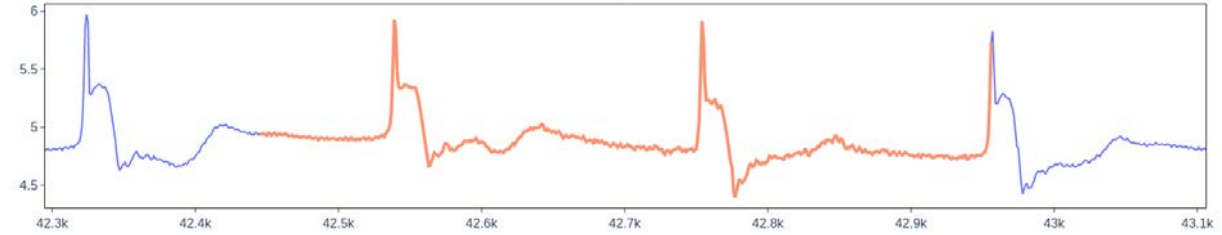
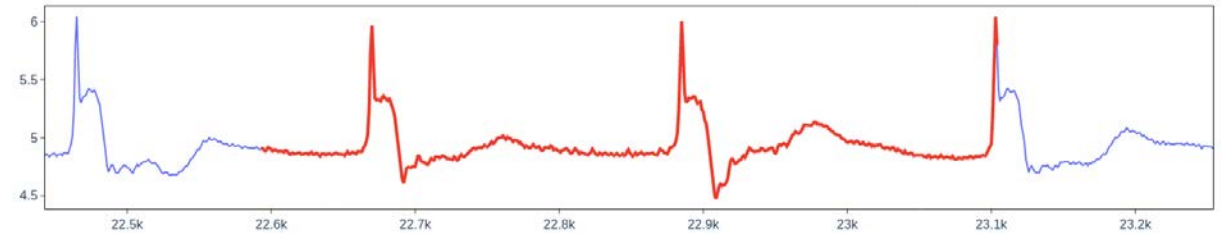
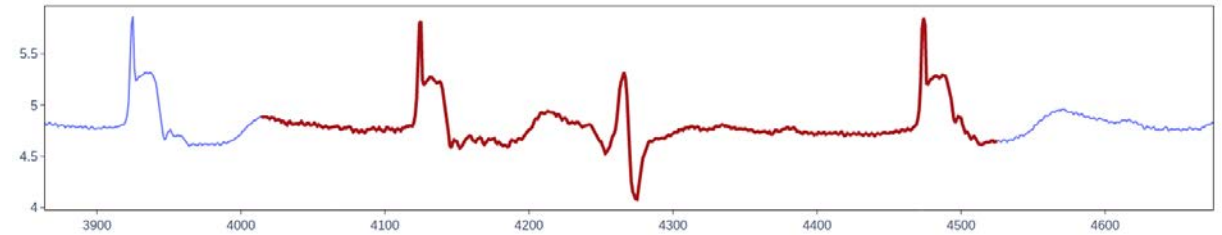
длина 256



top-3

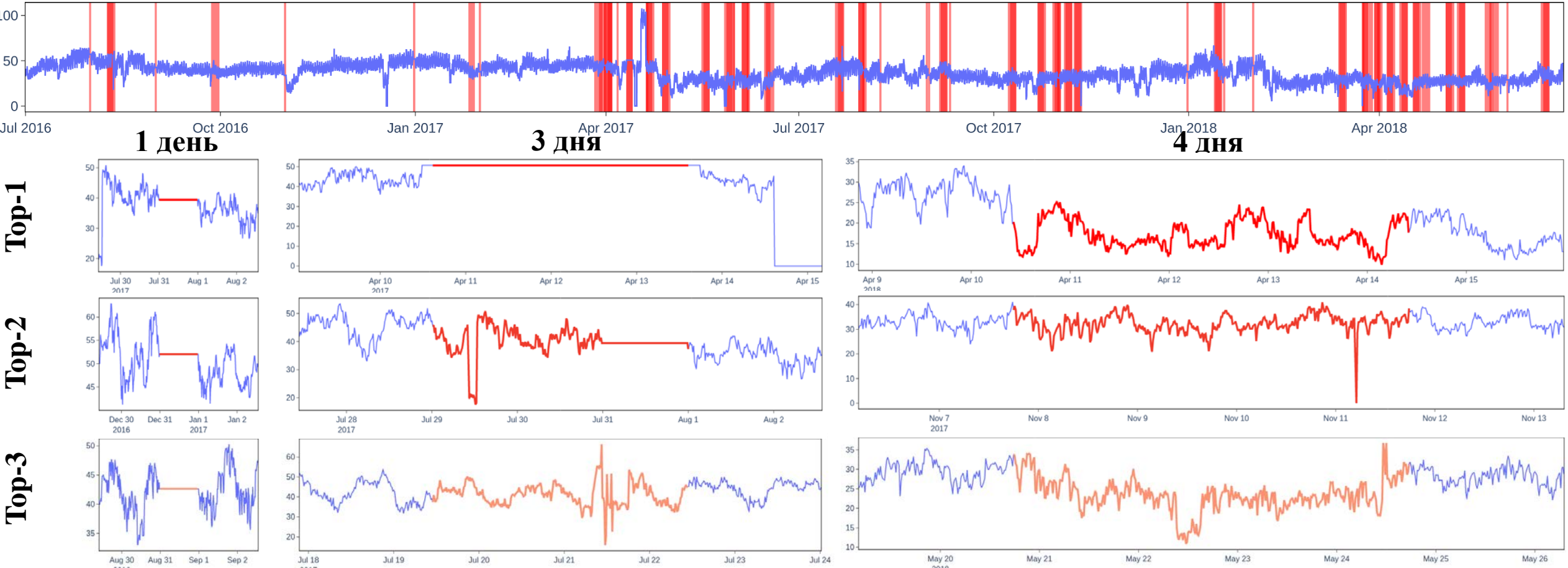


длина 512



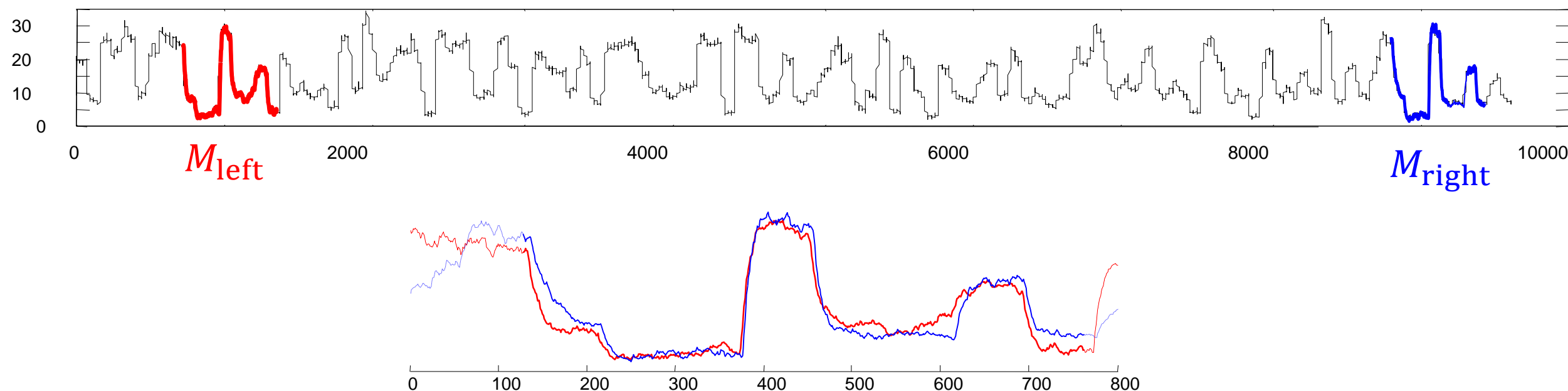
Поиск аномалий

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)

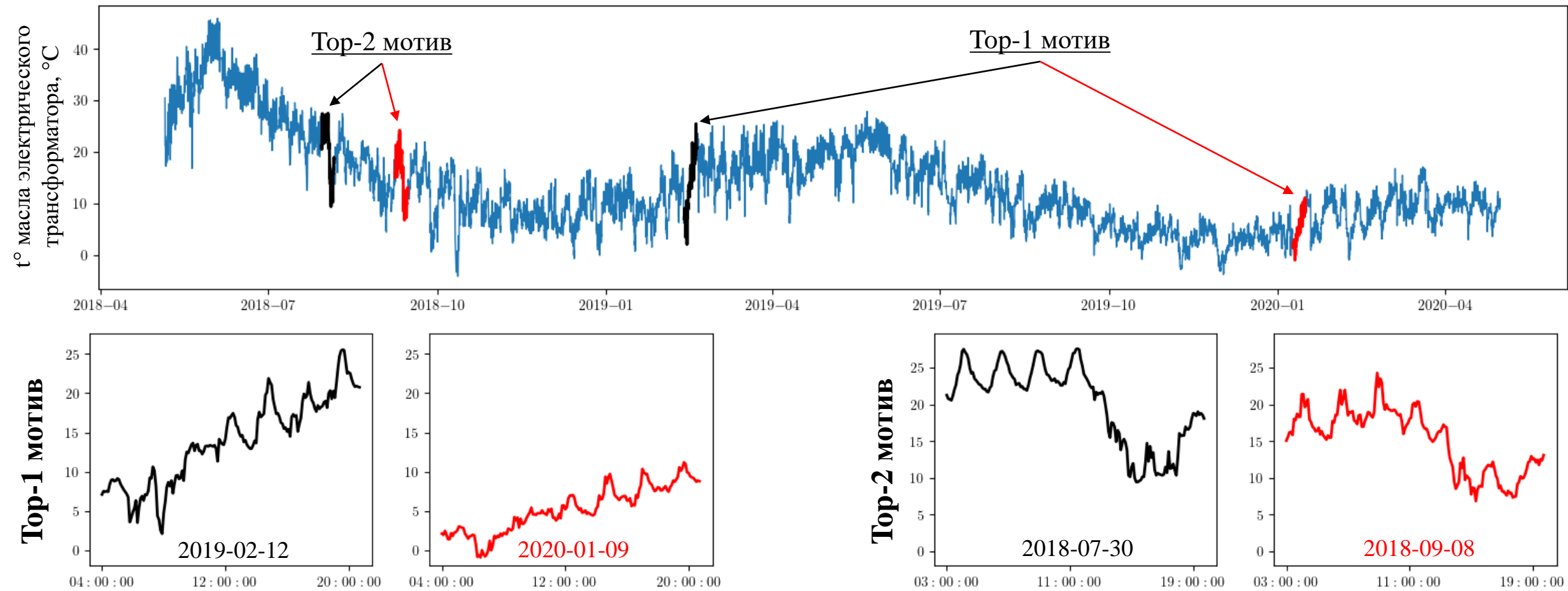


Пара непересекающихся подпоследовательностей ряда равной длины, наиболее похожих друг на друга:

$$\forall C_i, C_j \quad \text{Dist}(\textcolor{red}{M}_{left}, \textcolor{blue}{M}_{right}) \leq \text{Dist}(C_i, C_j)$$

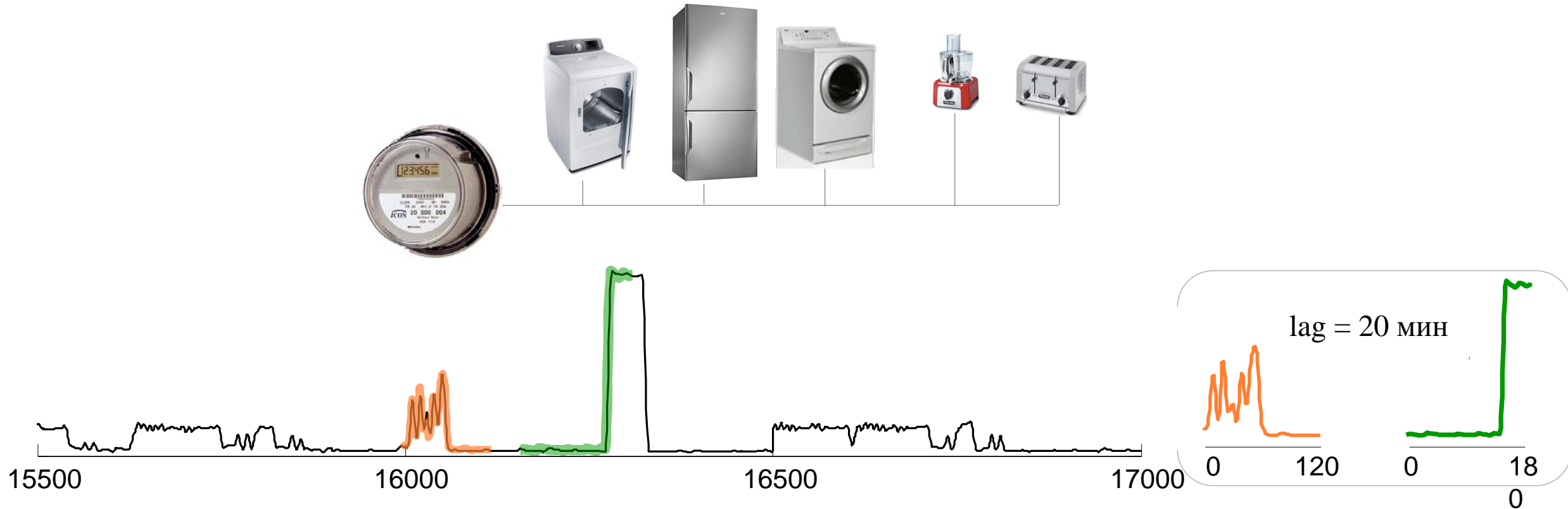
Поиск шаблонов: мотивы (motifs)

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).

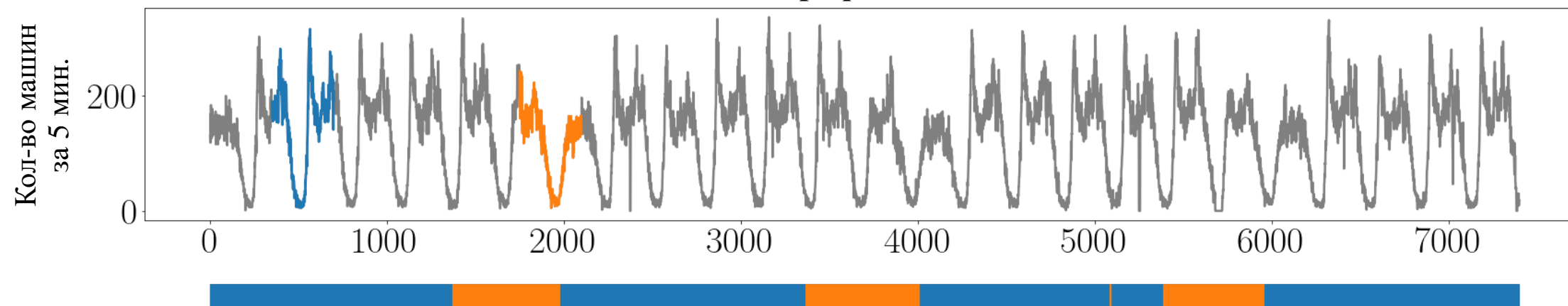
Поиск шаблонов: ассоциативные правила



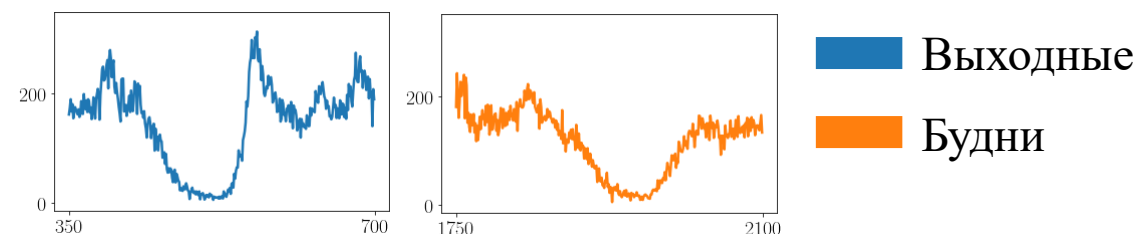
IF *работает стиральная машина*
THEN не более чем через 20 мин. *работает сушильная машина*

Поиск шаблонов: снippets (snippets)

Месячный автотрафик в Мюнхене*



Сниппеты

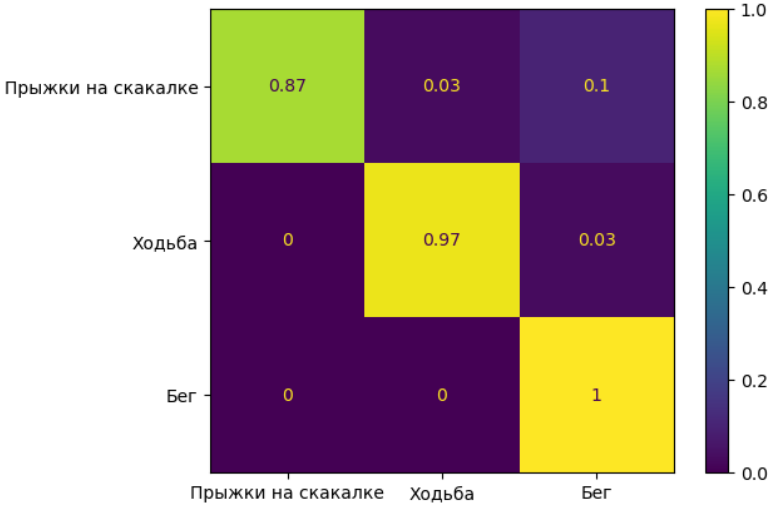
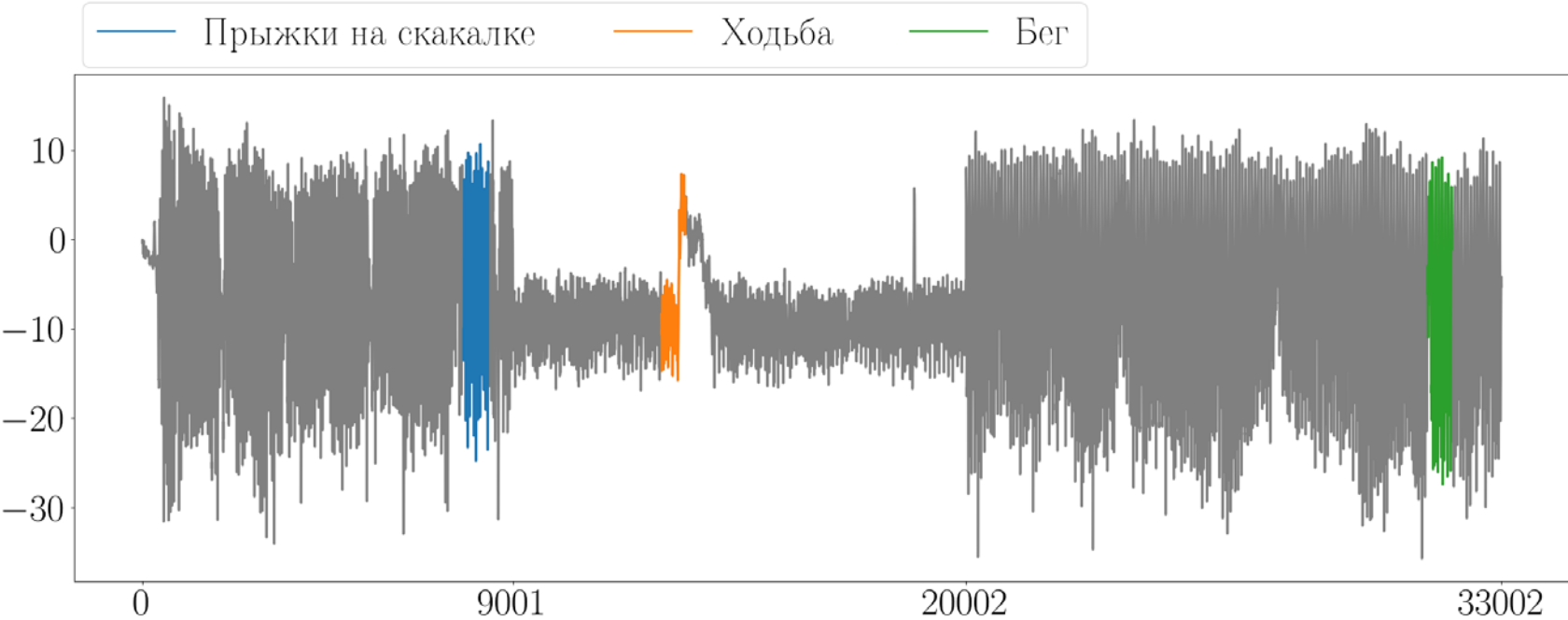


Множество подпоследовательностей ряда, выражающих типичные активности субъекта

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

Поиск шаблонов: снippets

Показания носимого акселерометра



Активность	Precision	Recall	F1
Прыжки	1	0.87	0.93
Ходьба	0.98	0.97	0.97
Бег	0.77	1	0.87

* 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

12:07
20 марта 2014

13:04

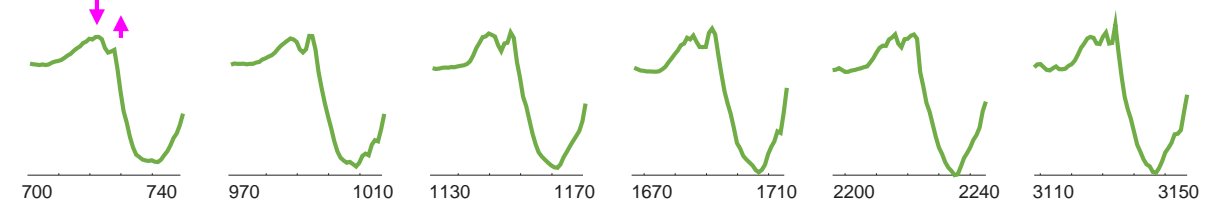
13:33

14:04

0 1 2 3
Мин



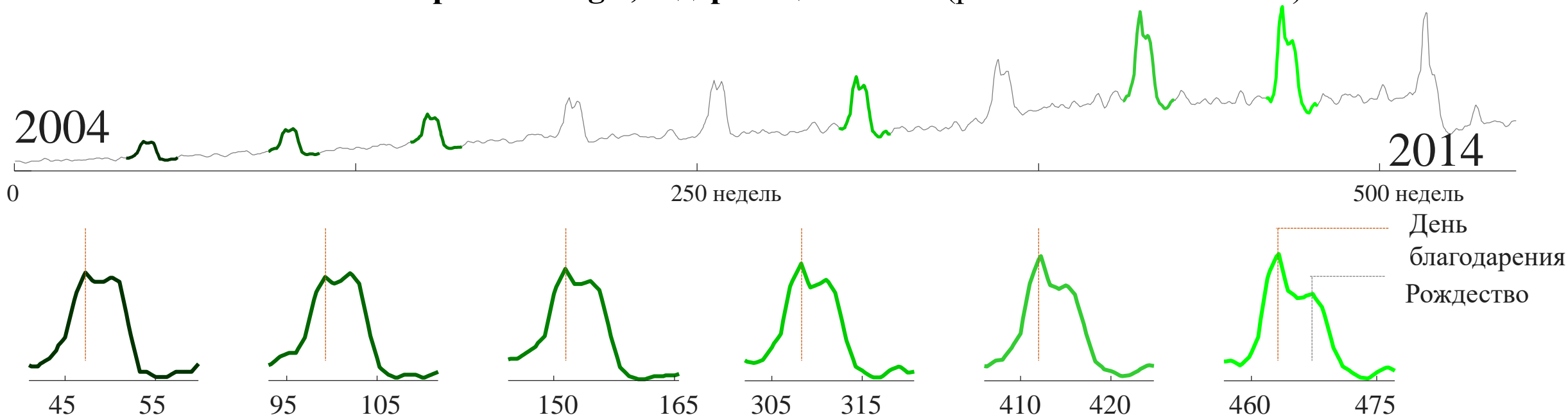
Запись датчика с левой икры спортсмена,
когда он начал бег трусцой на беговой дорожке



Множество подпоследовательностей ряда,
отражающих эволюцию некоего процесса

Поиск шаблонов: цепочки (chains)

Число запросов Google, содержащих *Kohl's* (розничная сеть в США)



- **Рост важности Киберпонедельника:** за 10 лет выпуклость переходит от плавной и занимающей больший период между Днём благодарения и Рождеством к резкой и сосредоточенной на Дне благодарения
- Киберпонедельник (понедельник после Дня благодарения) – маркетинговый термин для побуждения людей делать покупки онлайн; пресс-релиз “Киберпонедельник становится одним из крупнейших дней онлайн-покупок в году” от 28 ноября 2005 г., дата совпадает с первым проблеском острого пика в цепочке

Восстановление пропущенных значений ряда (imputation/recovery)

Синтез отсутствующих значений ряда в режиме онлайн/офлайн

Прогнозирование временного ряда (forecast)

Классификация временных рядов

Кластеризация временных рядов

Кластеризация подпоследовательностей ряда 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)

(Одномерный) временной ряд (univariate time series)

- Конечная последовательность хронологически упорядоченных вещественных значений

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

- n – длина ряда, $|T| = n$

Многомерный временной ряд (multivariate time series)

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

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

Потоковый временной ряд (streaming time series)

- Бесконечная упорядоченная последовательность вещественных значений, которые поступают непрерывно одно за другим в режиме реального времени
$$T = (t_1, \dots, t_n, \dots), \quad t_i \in \mathbb{R}$$
- Режим реального времени предполагает конечный период времени обработки данных *для заданной предметной области*:
реальное время \neq «очень быстро»

Подпоследовательность (subsequence)

- Непрерывный промежуток временного ряда фиксированной длины

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

- Множество всех подпоследовательностей ряда, имеющих заданную длину

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

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>.