

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

* 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/>

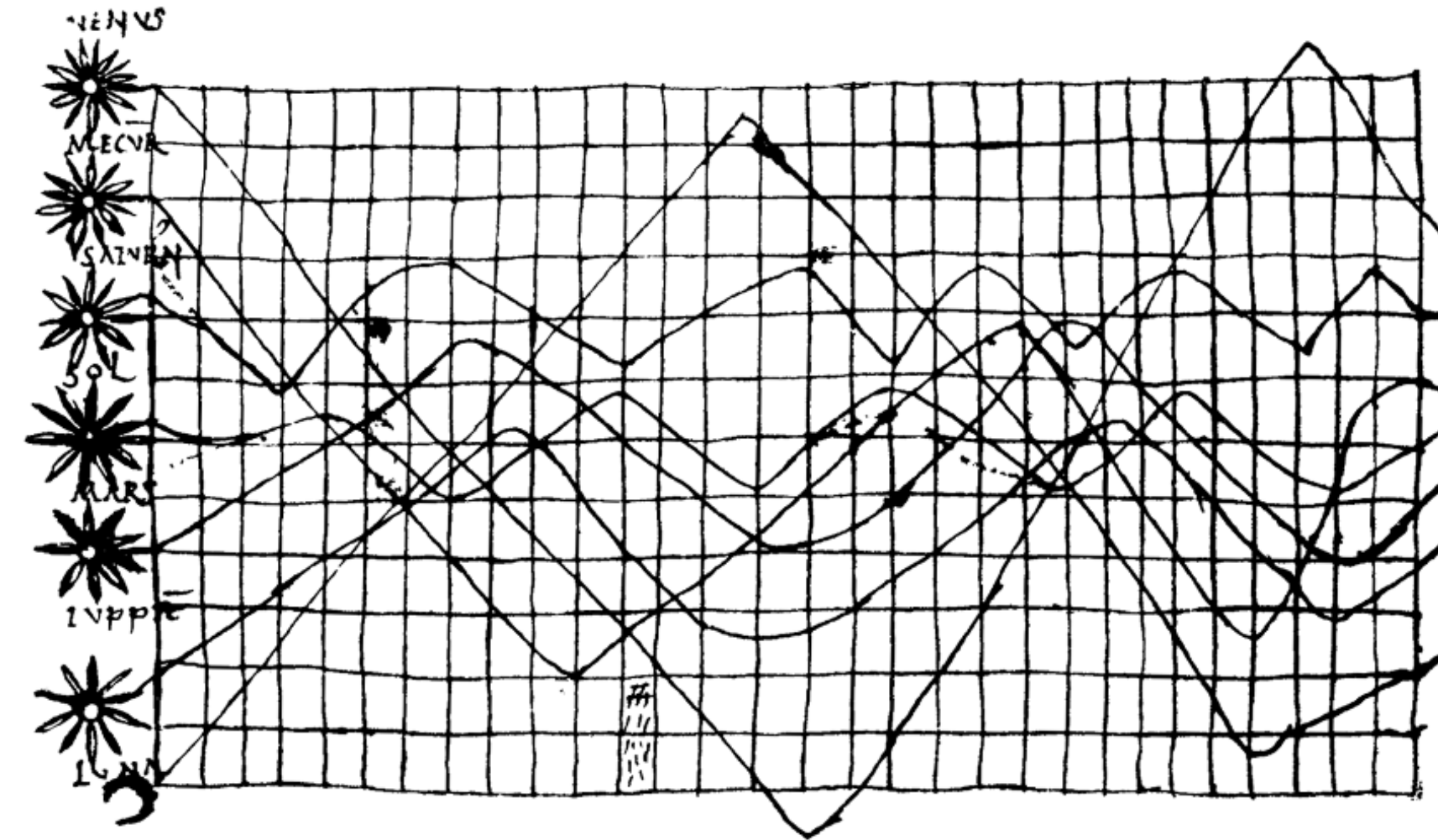
Содержание

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

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

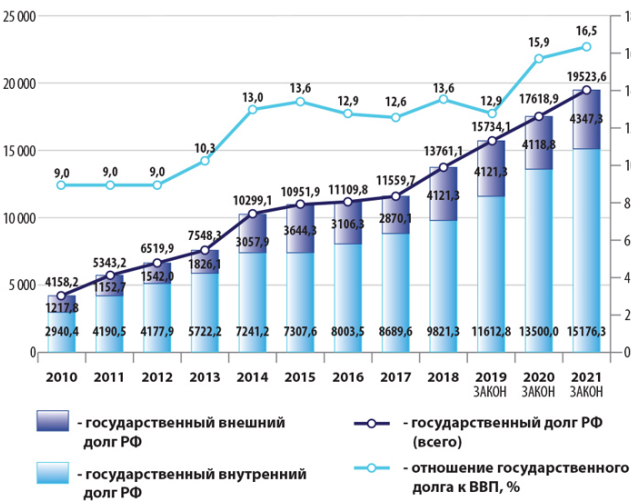
Временные ряды всегда...



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.

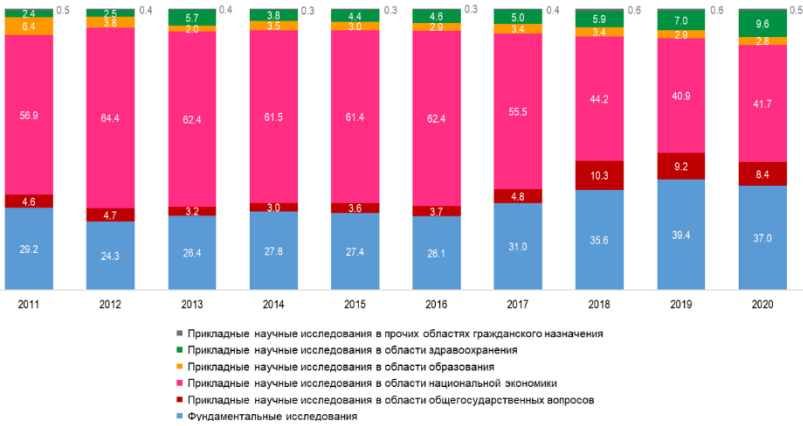
Временные ряды везде...



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.

с. 1. Структура ассигнований на гражданскую науку из средств федерального бюджета по подразделам классификации расходов бюджетов: 2011–2020 (%)



Заболееваемость COVID-19 в Челябинской области

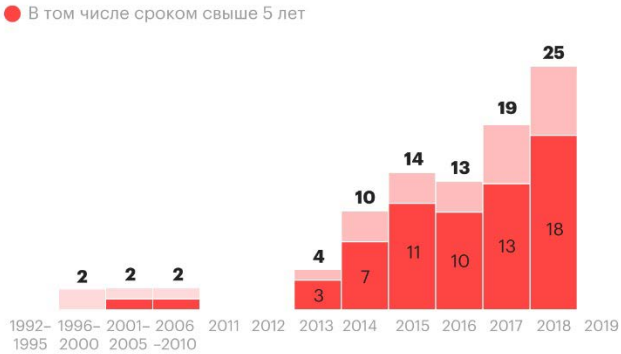
Сколько человек заразилось и умерло с 4 октября по 3 ноября 2021 г.



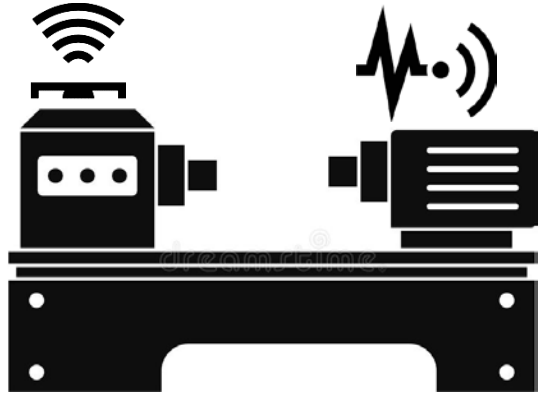
Статистика уголовных дел в отношении высокопоставленных чиновников



Вынесено приговоров с лишением свободы



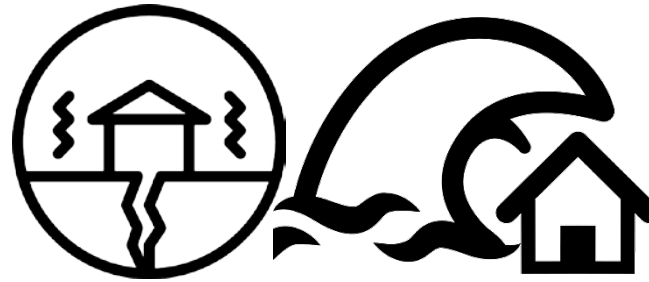
Временные ряды всюду...



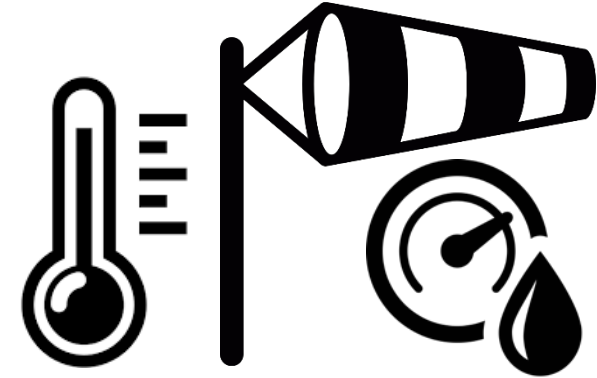
Умное производство,
предиктивное ТО



Интернет
вещей



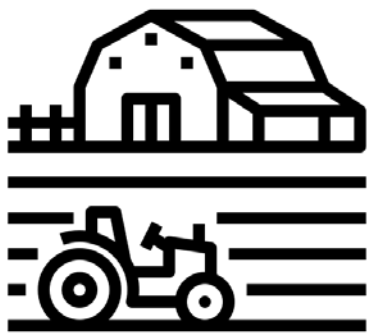
Предсказание
природных катаклизмов



Прогноз погоды,
моделирование климата



Персональная
медицина



Сельское хоз-во,
животноводство



Борьба
с преступностью



Био- и хемо-
информатика

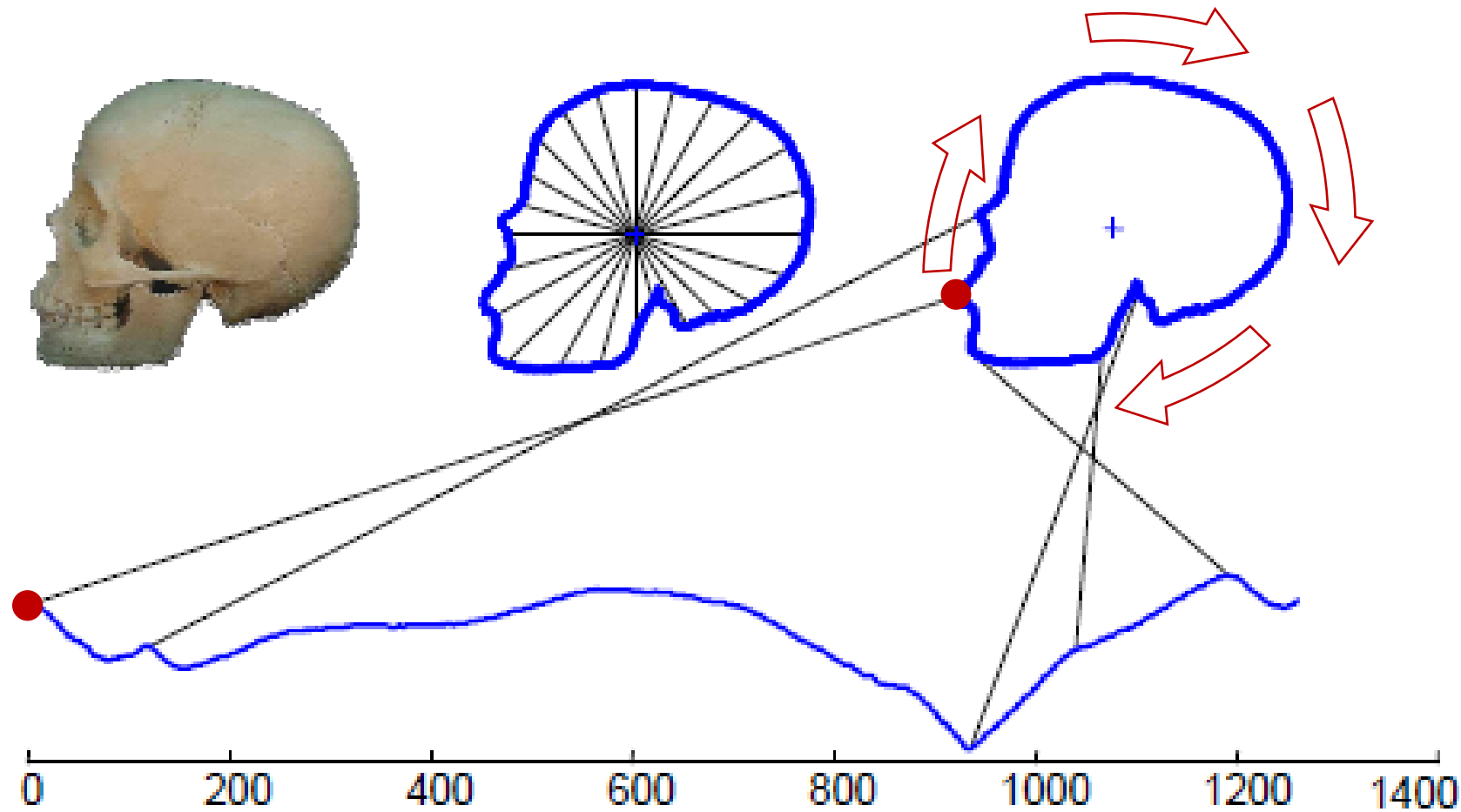


Экономика, бизнес,
финансы



Системы
электронного обучения

Временной ряд из 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

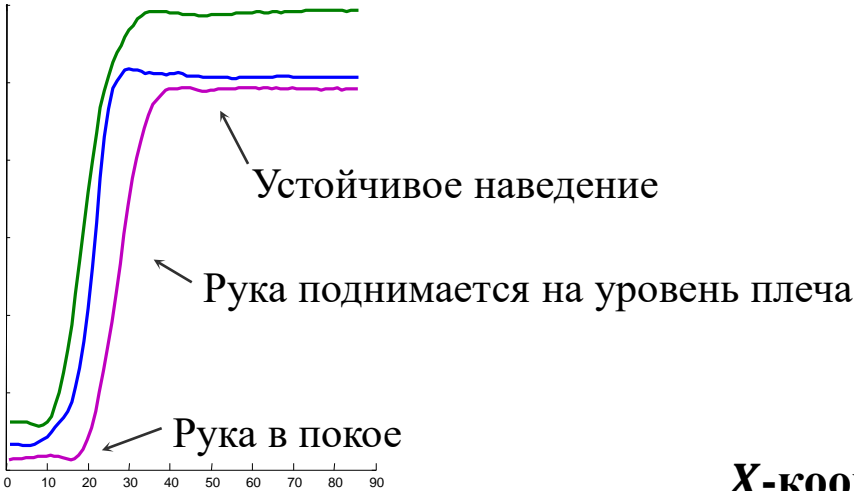
Классификация
“Gun/No gun”



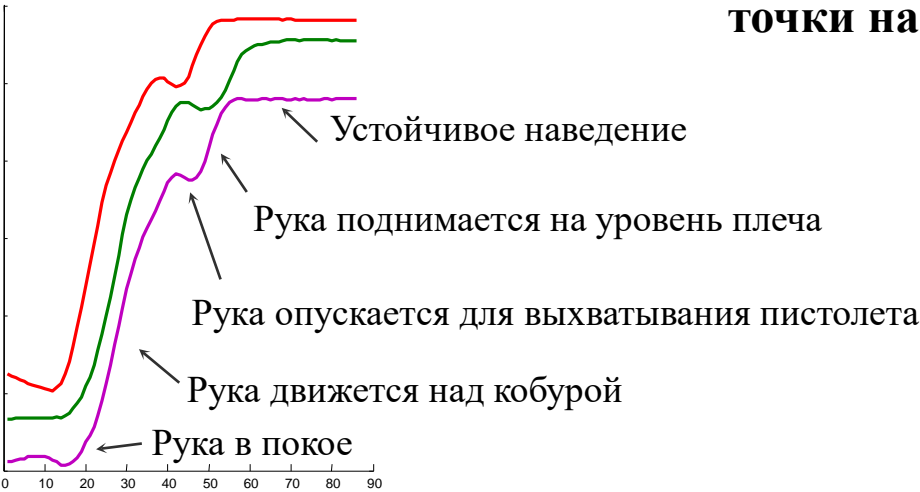
Указание
рукой



Вытаскивание
пистолета

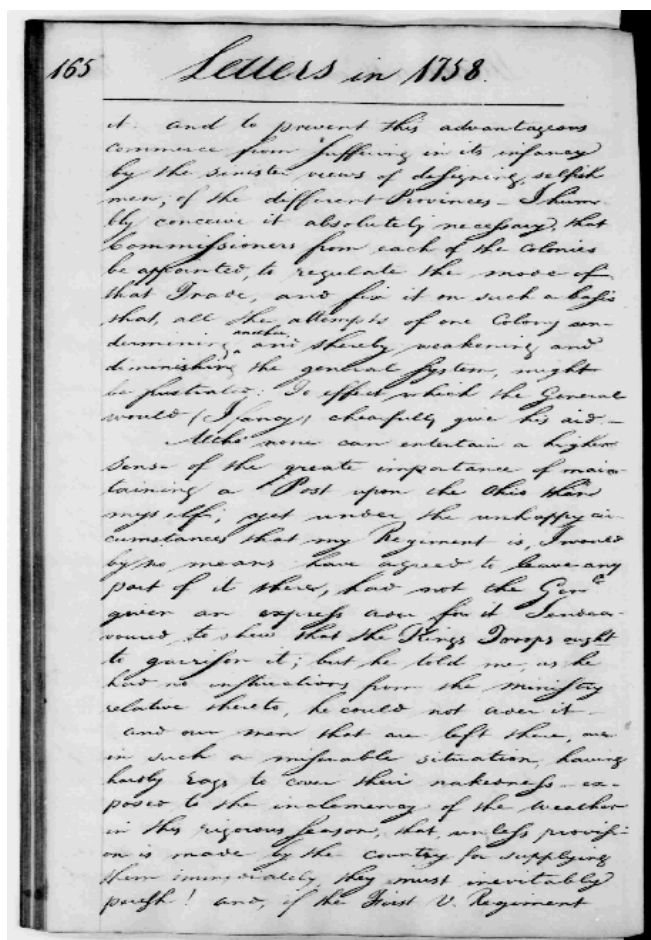


Х-координата
точки на руке



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>

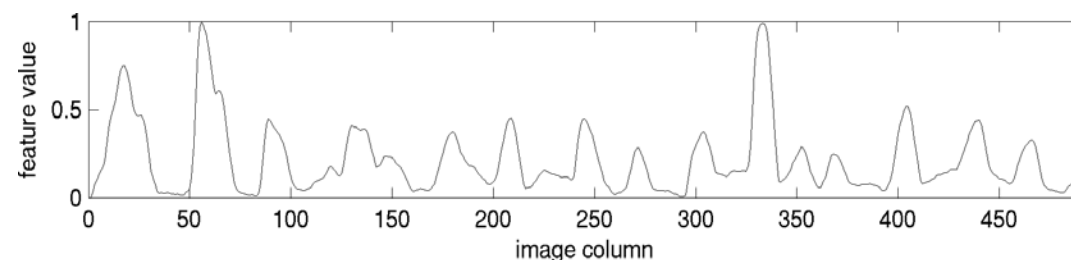
Временной ряд из рукописного текста



George Washington manuscript



George Washington
1732-1799



Содержание

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

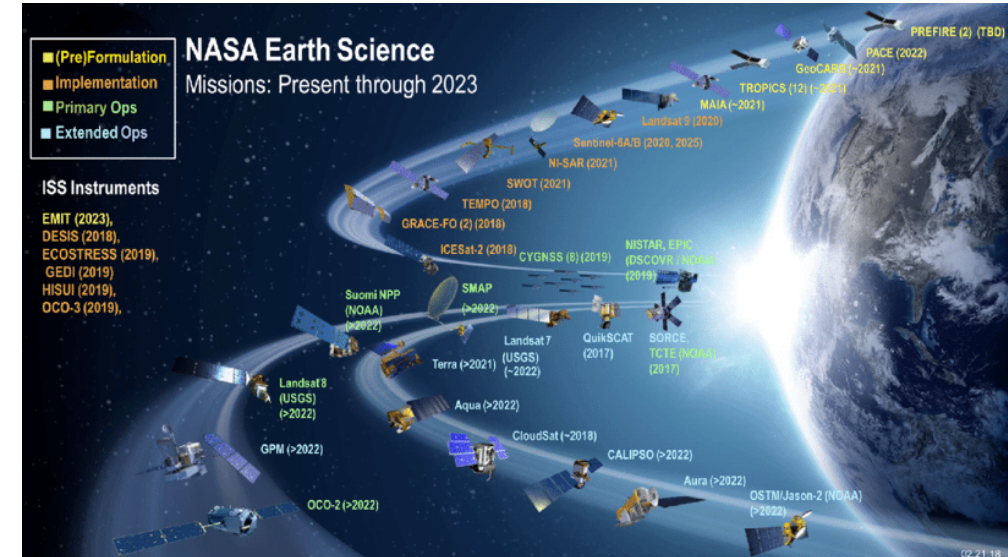
Почему временные ряды анализировать сложнее, чем другие данные

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

Большие временные ряды



Оборудование AEMO (Australian Energy Market Operator) регистрирует выработку солнечной энергии (МВт) в Австралии с 2019 г.
<https://zenodo.org/record/4656027> каждые 4 с



Размер NASA Space Shuttle Earth Observations Database – 40 Пб* (2020 г.), ожидаемый далее ежегодный прирост 50 Пб
<https://www.nasa.gov/feature/goddard/2020/nasa-funds-projects-to-make-geosciences-data-more-accessible>

* 1 Петабайт = 10^{15} (квадриллион) байт
 $10^{15} \approx$ к-во синапсов в головном мозге человека

Большие временные ряды



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>



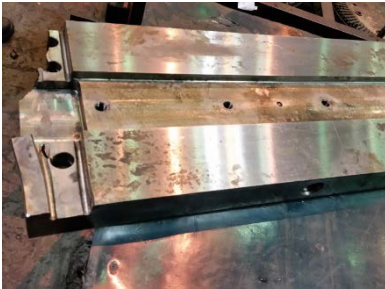
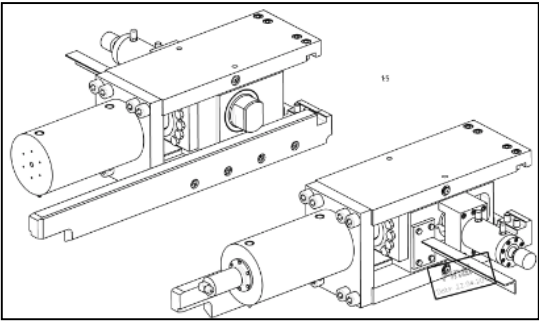
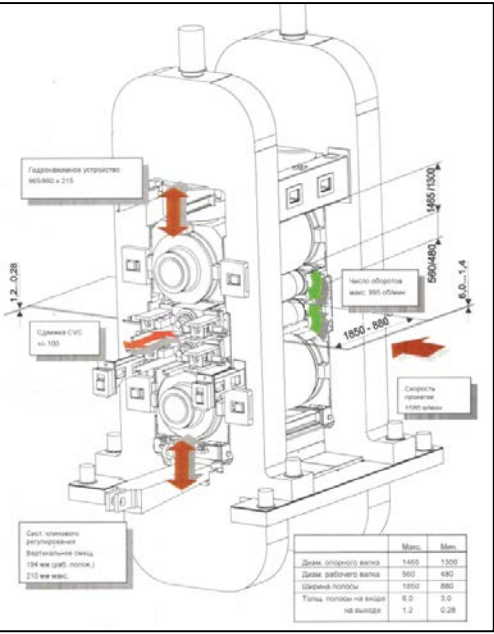
Набор данных **DEBS challenge**: сенсоры пространственного позиционирования закреплены на бутсах игроков и перчатках вратаря (200 Гц), а также на мяче (2000 Гц), всего 15К событий в секунду
Mutschler C. *et al.* The DEBS 2013 grand challenge. DEBS'13: Proc. of the 7th ACM international conference on Distributed event-based systems. 2013. P. 289–294. <https://doi.org/10.1145/2488222.2488283>

Большие временные ряды



Система контроля дорожного движения в Мадриде насчитывает более 3500 автоматических регистраторов транспортных средств, 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. Pages 18-33. <https://doi.org/10.1016/j.trc.2018.02.021>

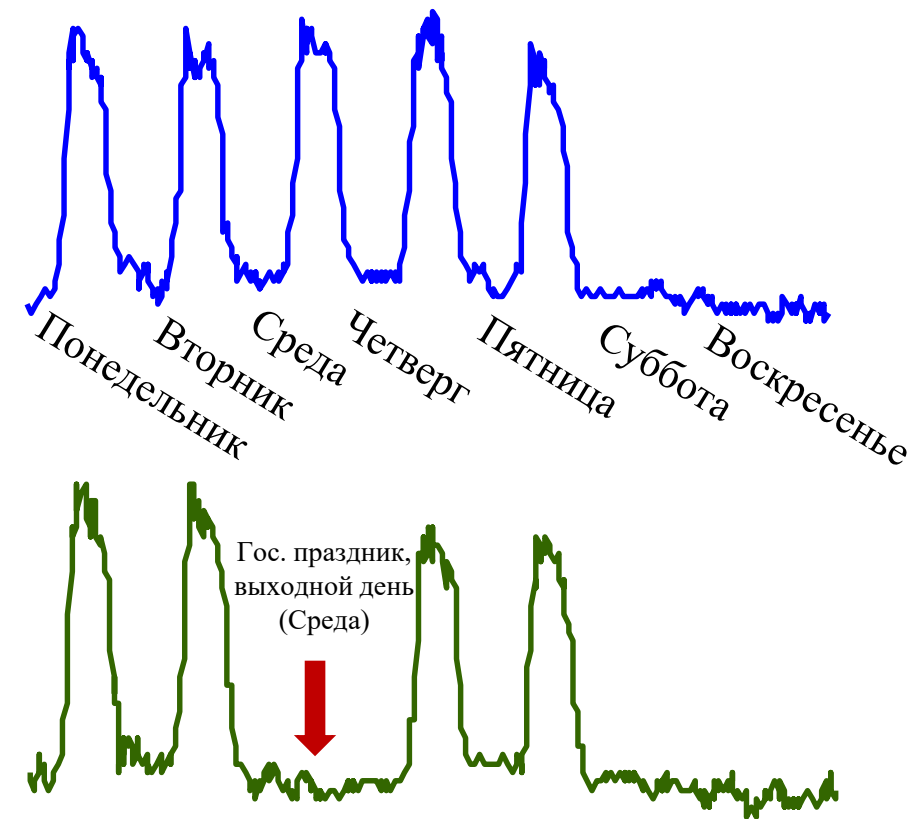


СВС-система стана холодной прокатки
Магнитогорского металлургического комбината насчитывает более 100 датчиков с частотой 1 Гц.

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>

Схожесть рядов определяется задачей и предметной областью

Недельное энергопотребление
вычислительного центра (Голландия, 1997)*



Евклидово расстояние

Время
построения
1 с

Количество
рабочих дней
в неделе:

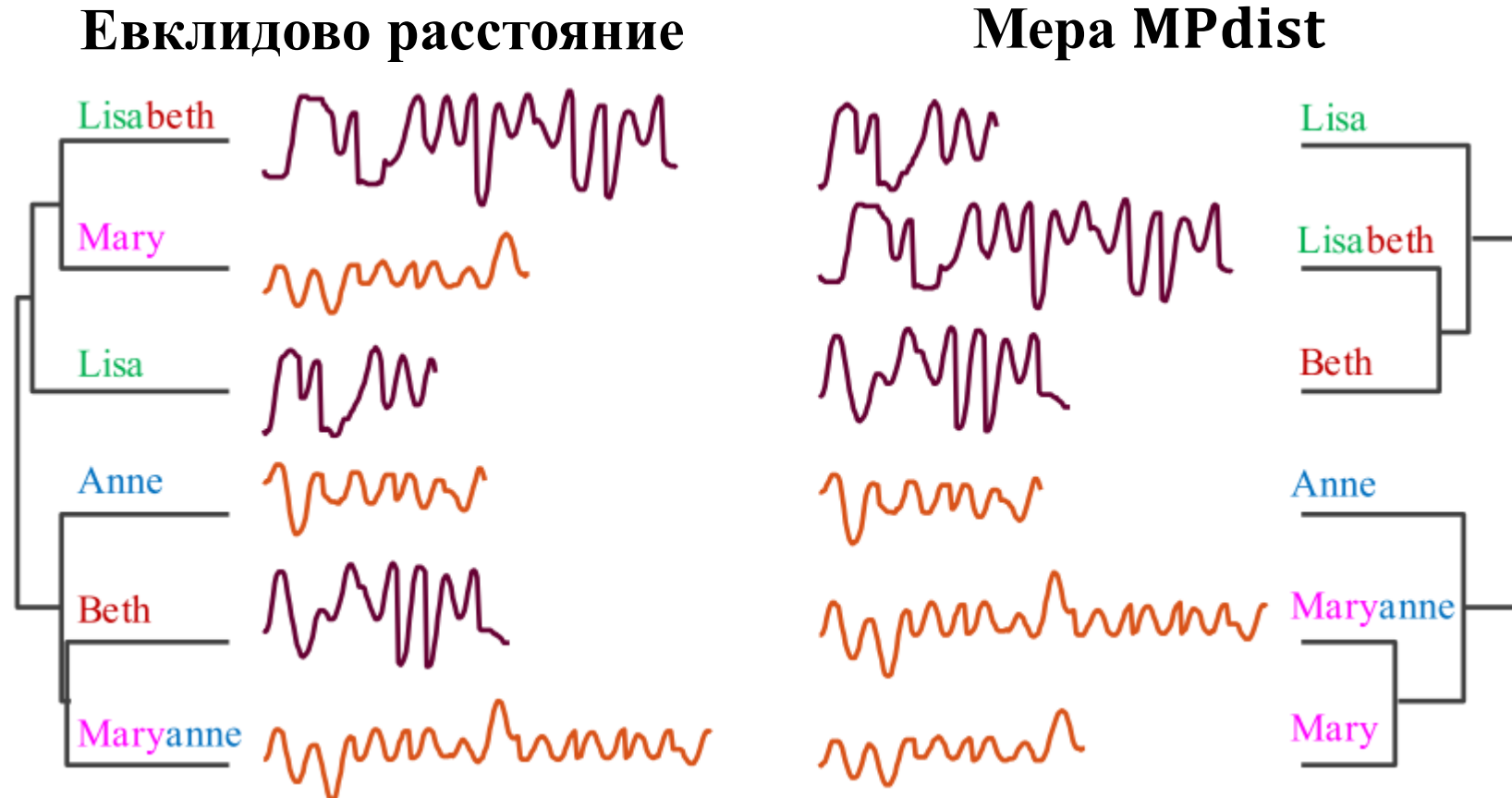


Dynamic Time Warping

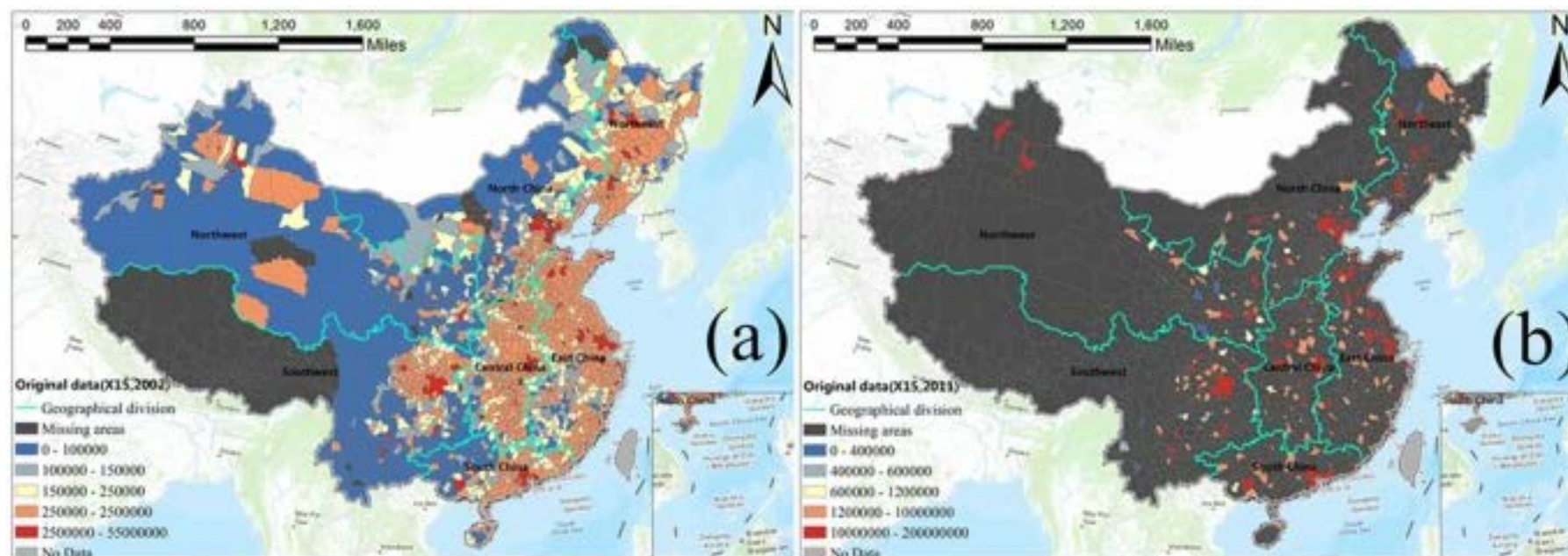
Время
построения
3 час

* 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%

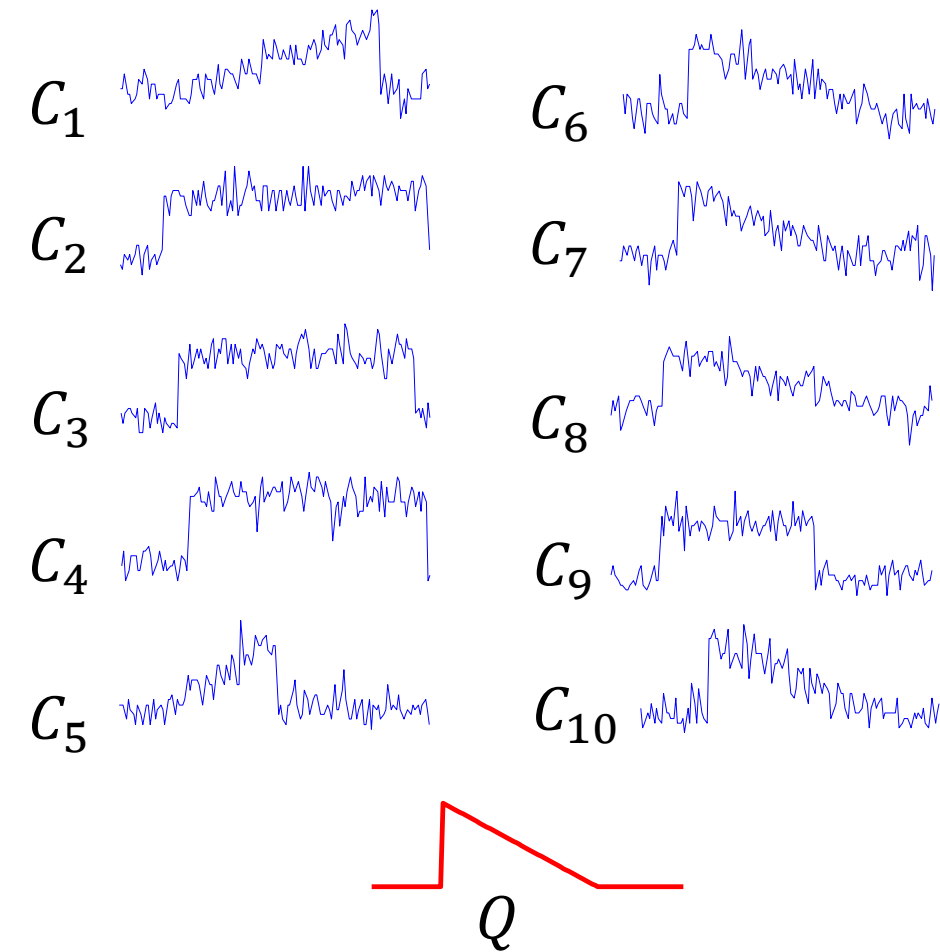
b) 2011: более 85%

* 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

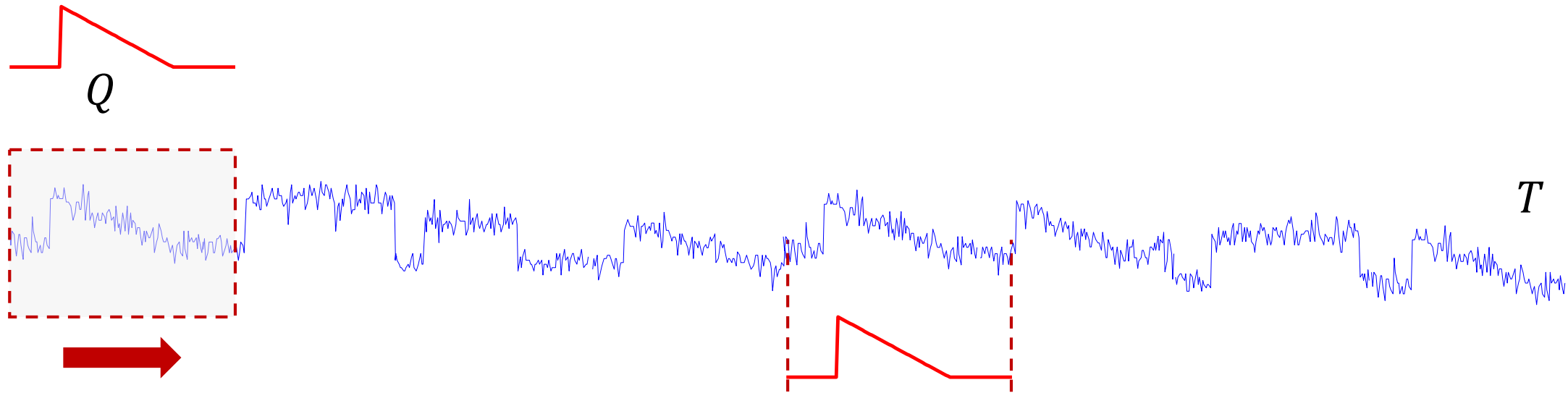
Поиск по образцу: случай нескольких временных рядов (whole matching)



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

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

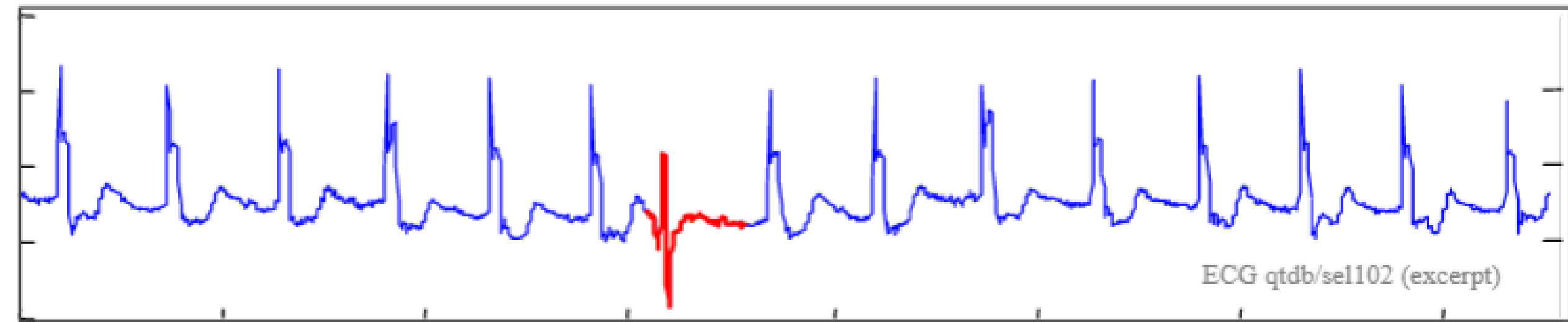
Поиск по образцу: случай подпоследовательностей временного ряда (subsequence matching)



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

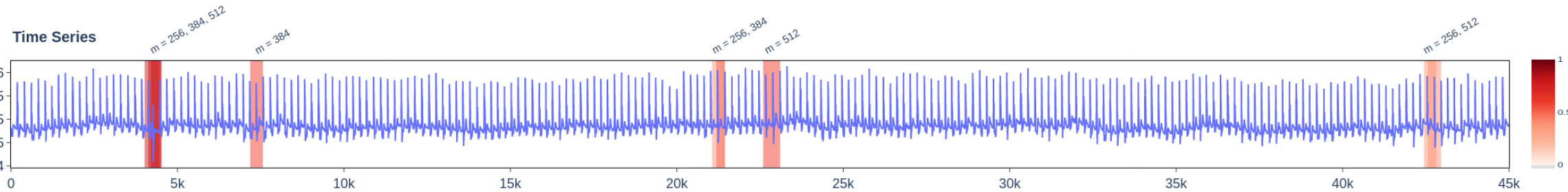
Поиск аномалий

ЭКГ
взрослого
пациента

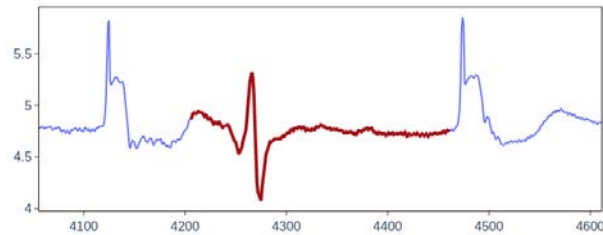
Прежде-
временное
сокращение
желудочков

Эктопическое
сердцебиение

Эктопическое
сердцебиение

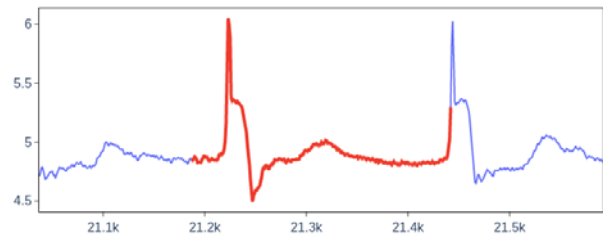


top-1

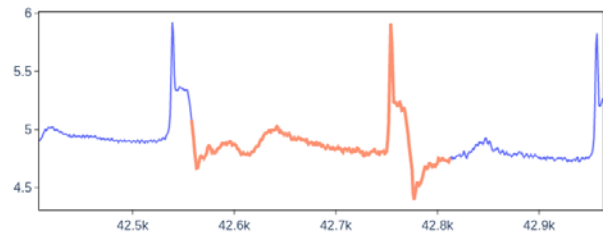


top-2

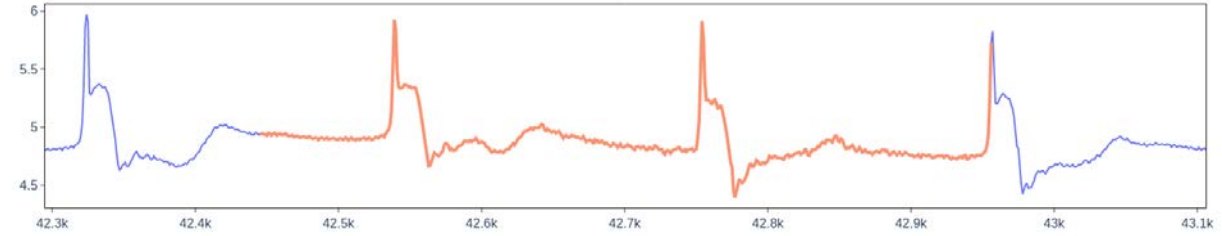
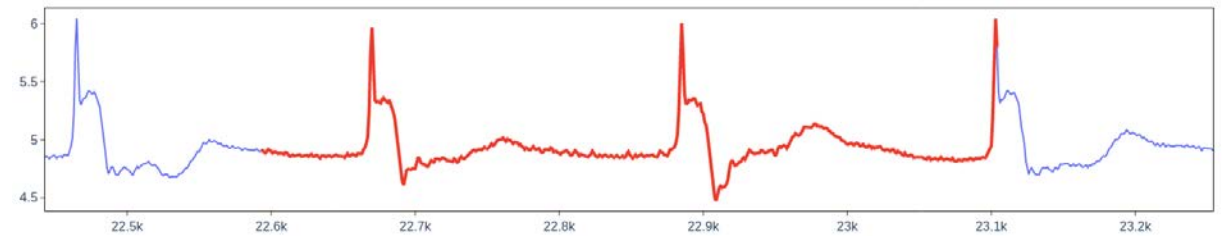
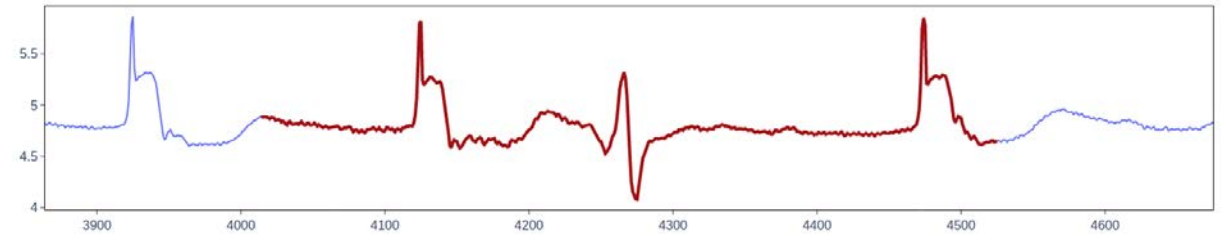
длина 256



top-3

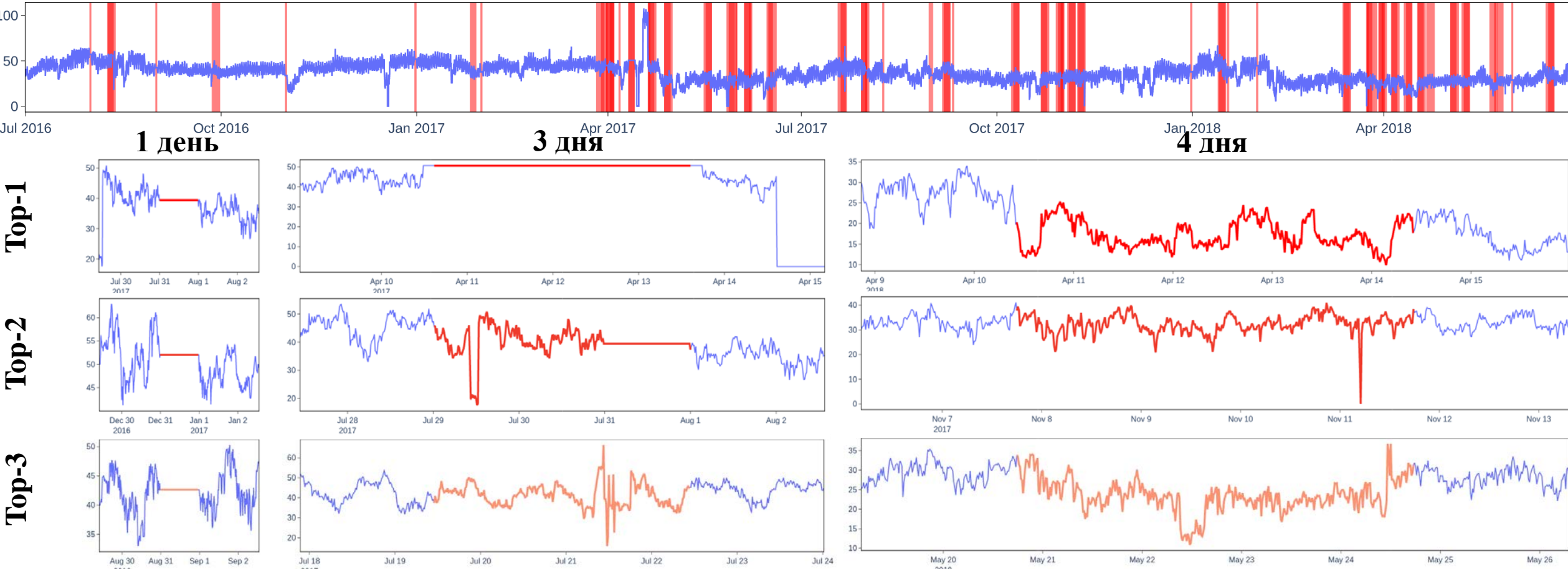


длина 512



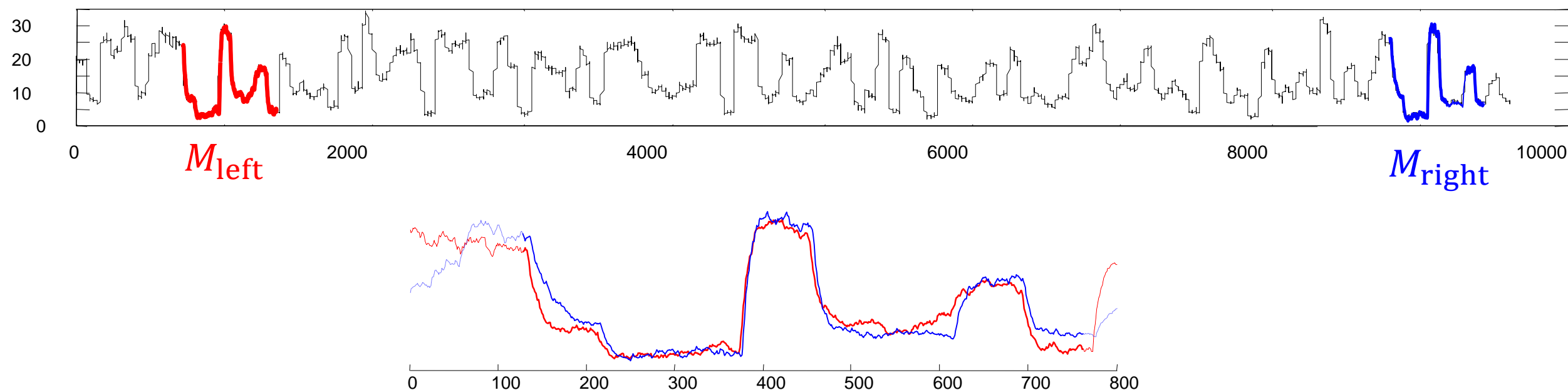
Поиск аномалий

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)

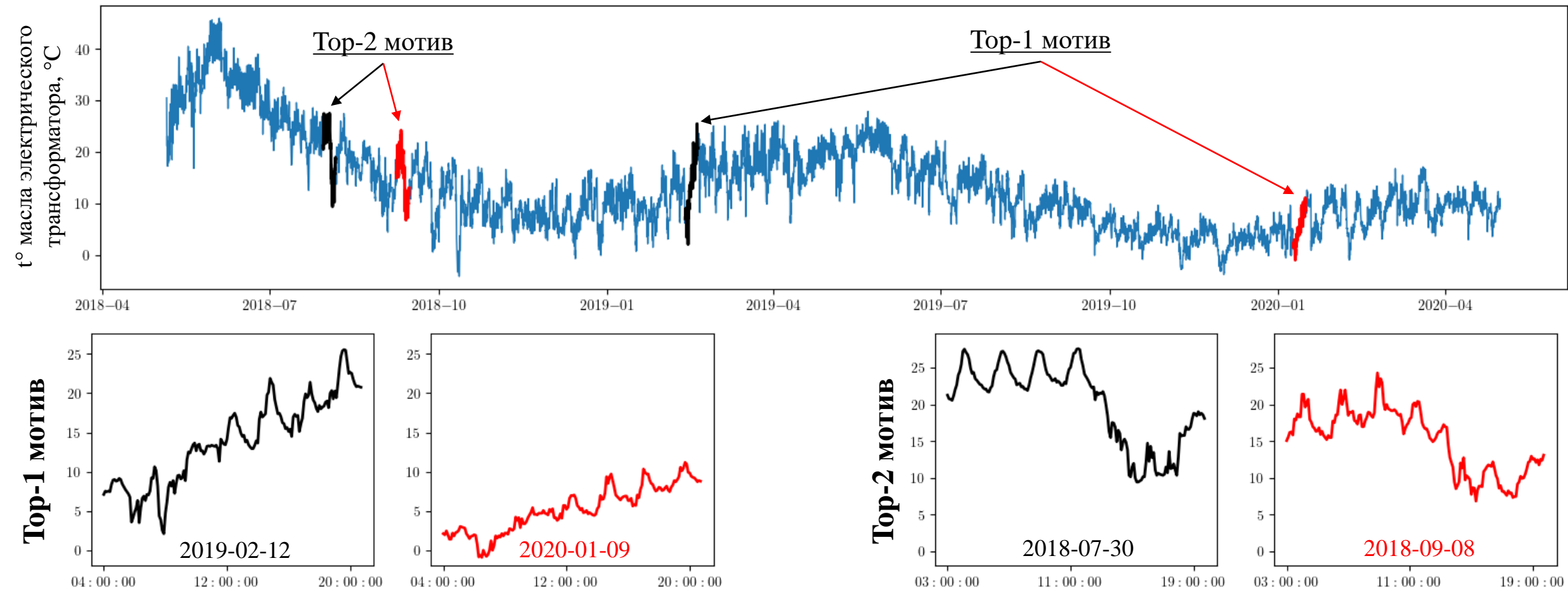


Пара непересекающихся подпоследовательностей ряда равной длины, наиболее похожих друг на друга:

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

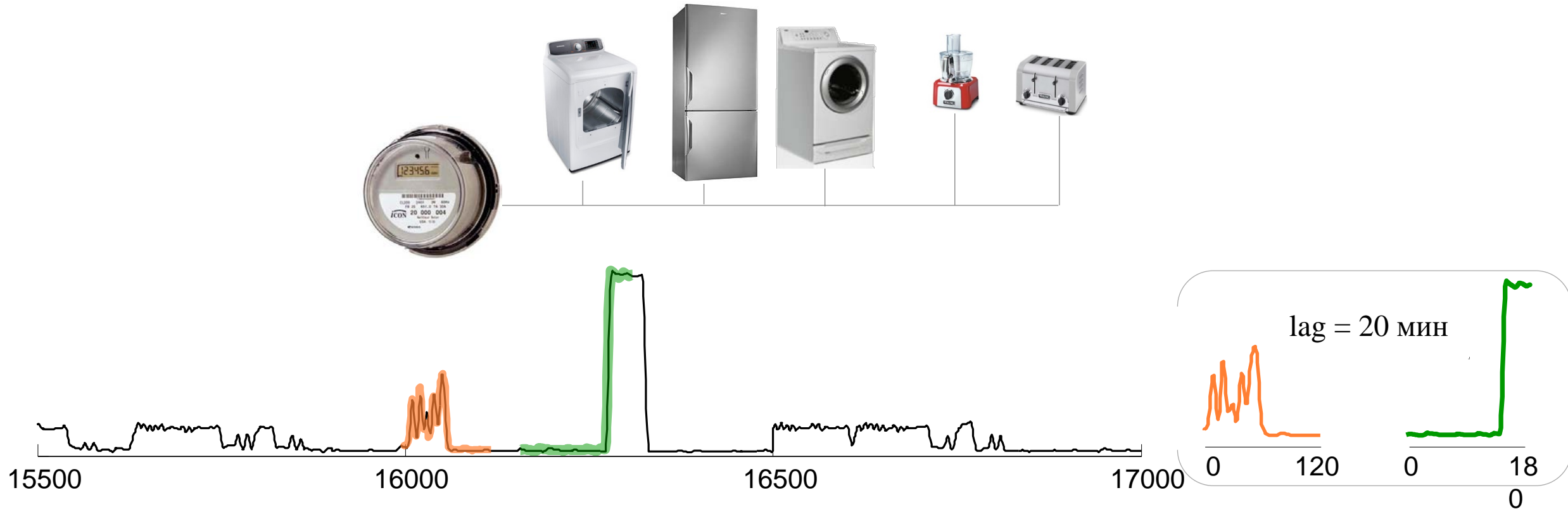
Поиск шаблонов: мотивы (motifs)

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).

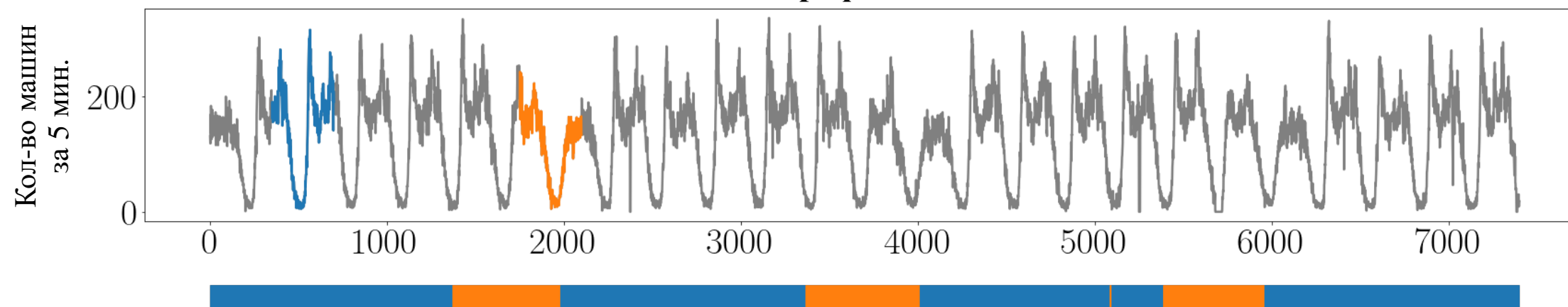
Поиск шаблонов: ассоциативные правила (association rules)



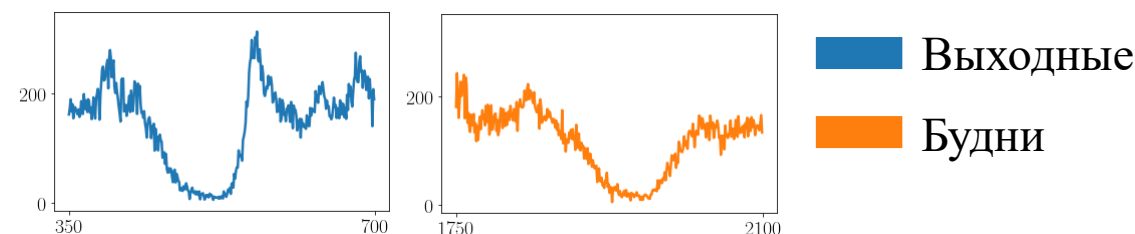
IF *работает стиральная машина*
THEN не более чем через 20 мин. *работает сушильная машина*

Поиск шаблонов: снippets (snippets)

Месячный автотрафик в Мюнхене*



Сниппеты

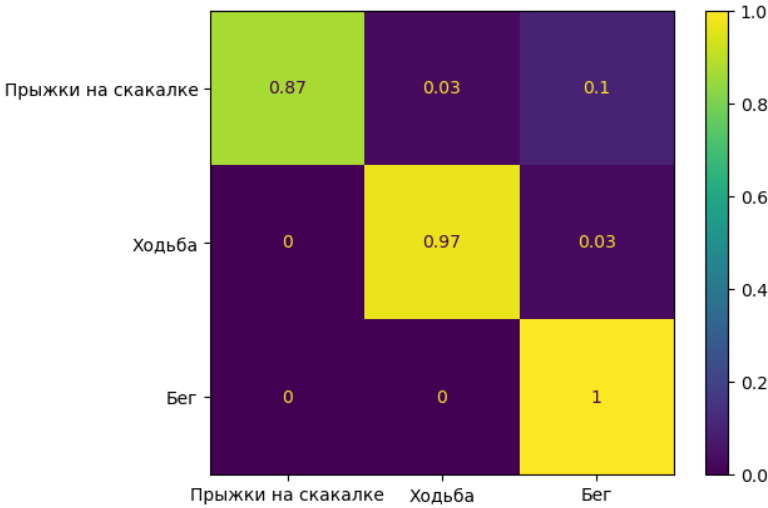
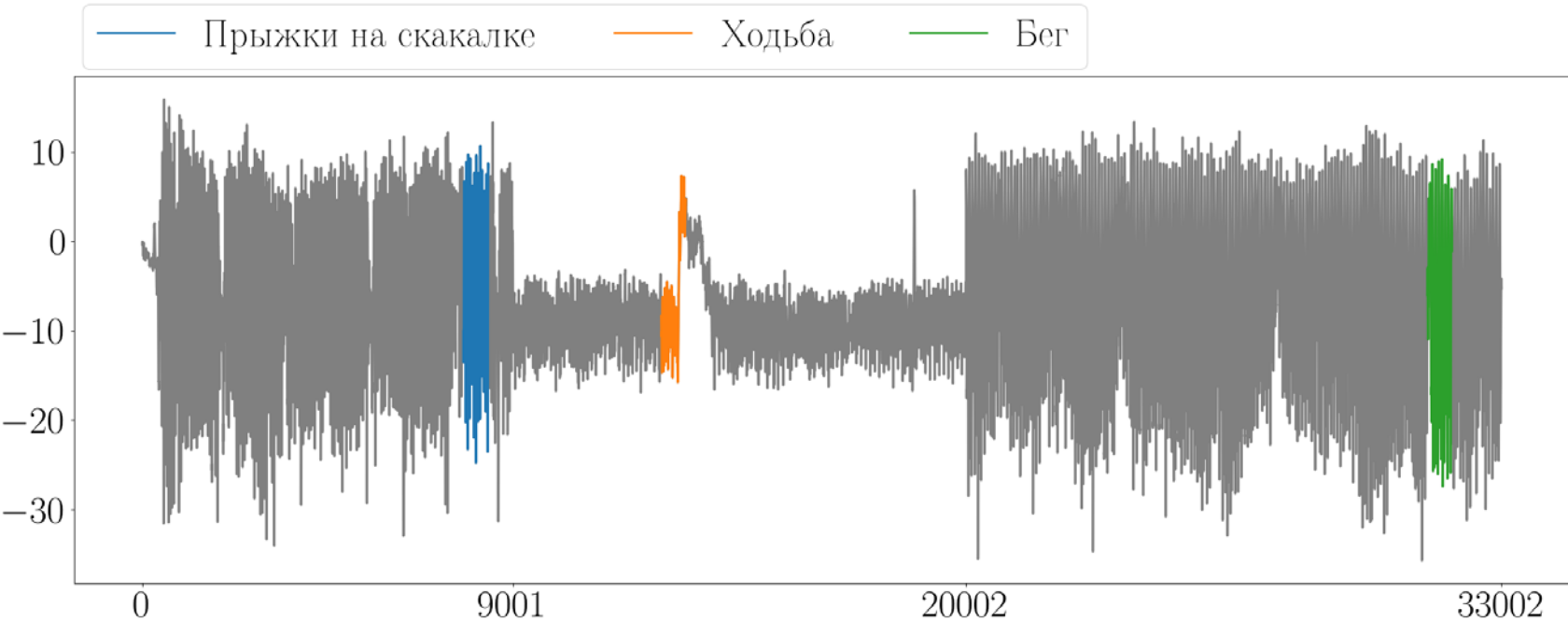


Множество подпоследовательностей ряда, выражающих типичные активности субъекта

* 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

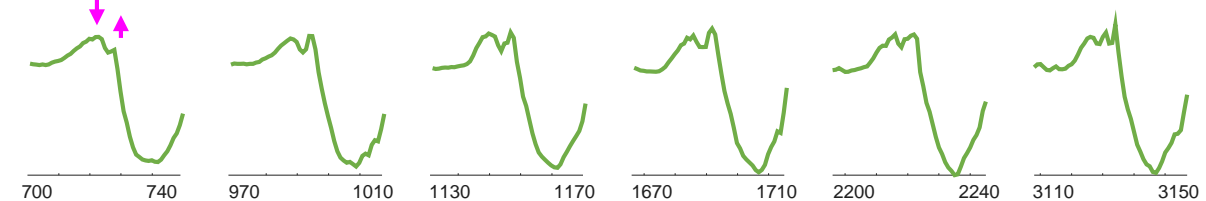
13:33

14:04

0 1 2 3
Мин



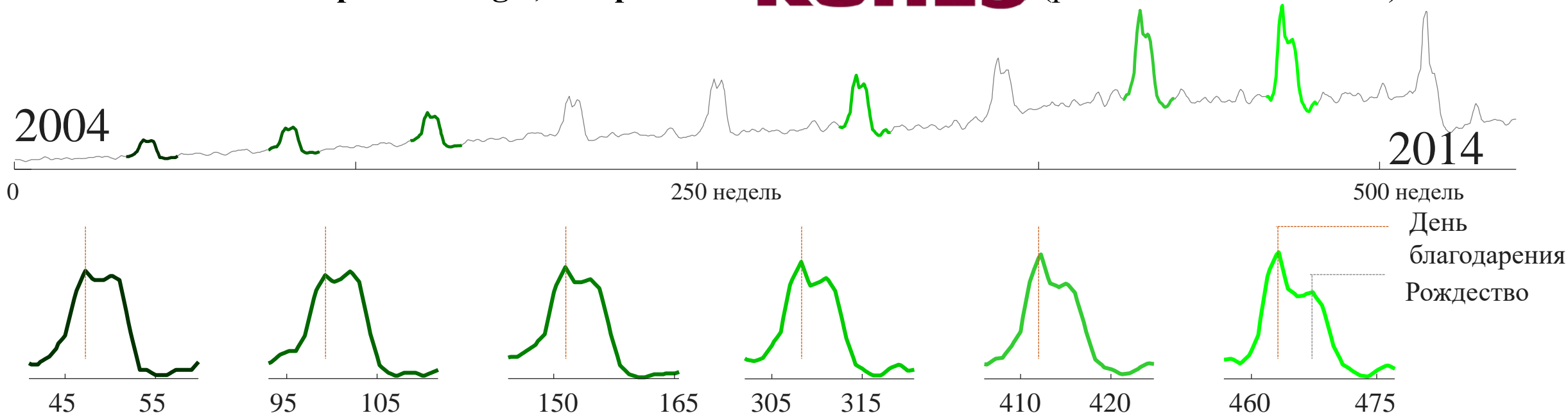
**Запись датчика с левой икры спортсмена,
когда он начал бег трусцой на беговой дорожке**



**Цепочка подпоследовательностей ряда,
звенья которой отражают эволюцию некоего процесса**

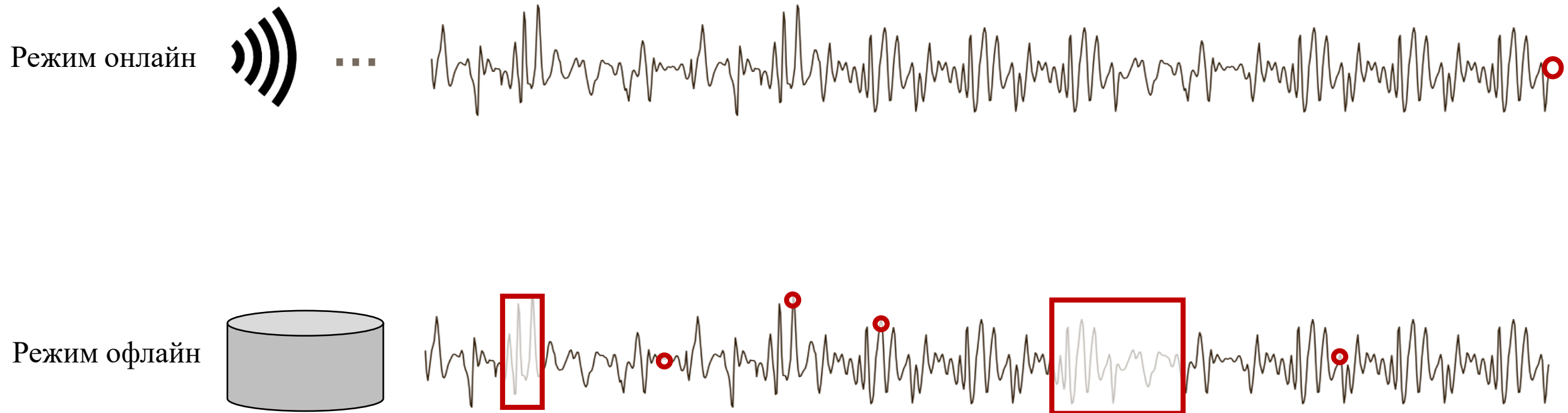
Поиск шаблонов: цепочки (chains)

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



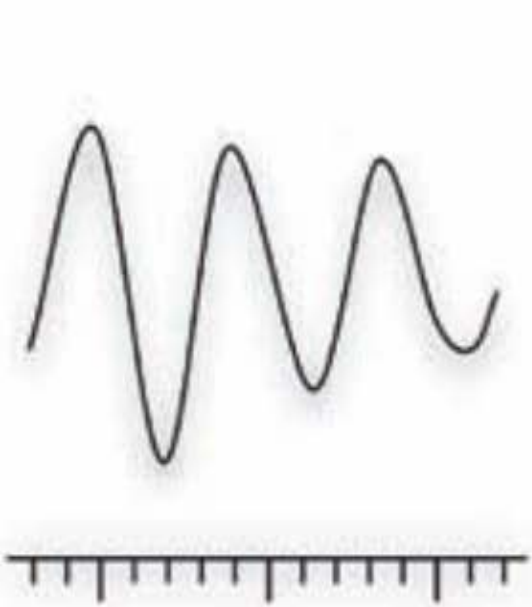
- **Рост важности Киберпонедельника** (понедельник после Дня благодарения): за 10 лет выпуклость меняется от плавной и занимающей больший период между Днём благодарения и Рождеством к резкой и сосредоточенной на Дне благодарения
- Термин введен в пресс-релизе “Киберпонедельник становится одним из крупнейших 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)

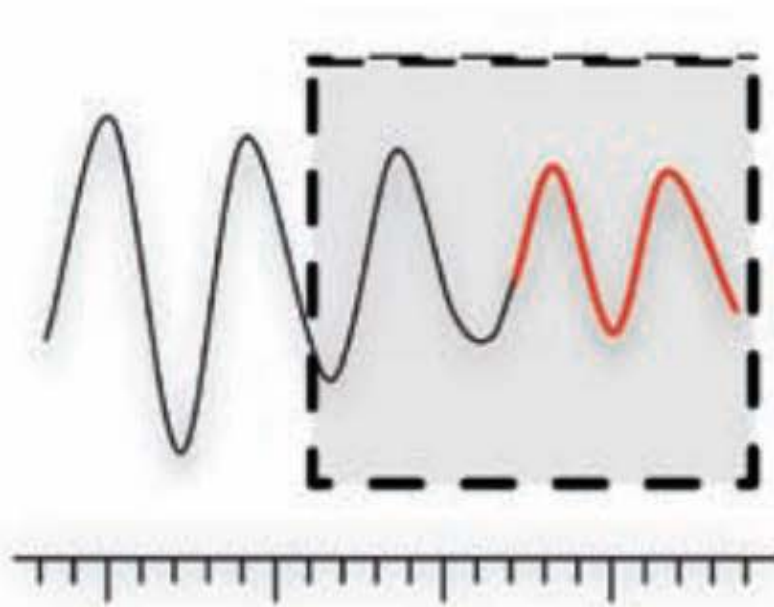


Синтез отсутствующих значений ряда

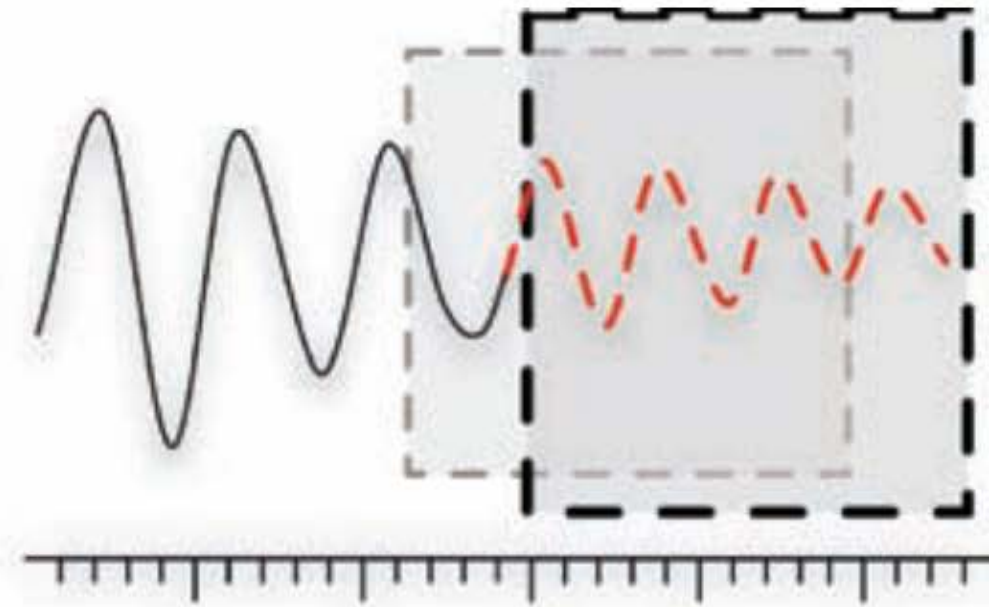
Прогнозирование временного ряда (forecast)



Исходный ряд
(периодическая структура,
поддающаяся прогнозу)



Прогноз точек данных
в пределах
окна прогнозирования



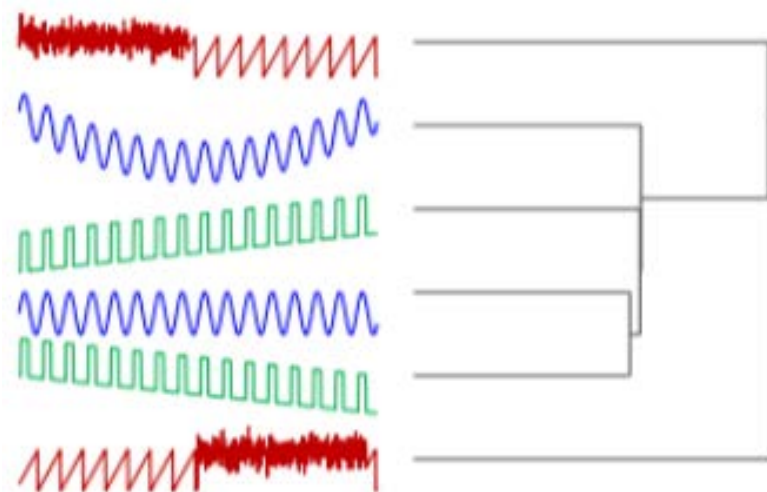
Долгосрочный прогноз:
использование более ранних прогнозных
значений в качестве входных данных прогноза

Синтез будущих значений ряда

Классификация и кластеризация временных рядов

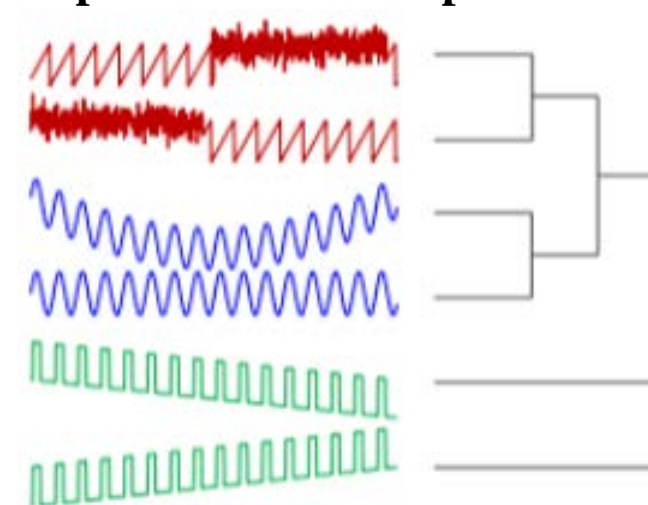
- При небольшой длине рядов можно использовать стандартные алгоритмы машинного обучения и адекватные функции для вычисления схожести (например, Dynamic Time Warping)
- Для длинных временных рядов нужны специализированные функции для вычисления схожести

DTW



PRCIS

(Pattern Representation Comparison in Series)*



* 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

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

(Одномерный) временной ряд (univariate time series)

- Конечная последовательность хронологически упорядоченных вещественных значений

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

- 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, заданный для *конкретной предметной области*: **реальное время \neq «очень быстро»**

Подпоследовательность (subsequence)

- Непрерывный промежуток временного ряда фиксированной длины

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

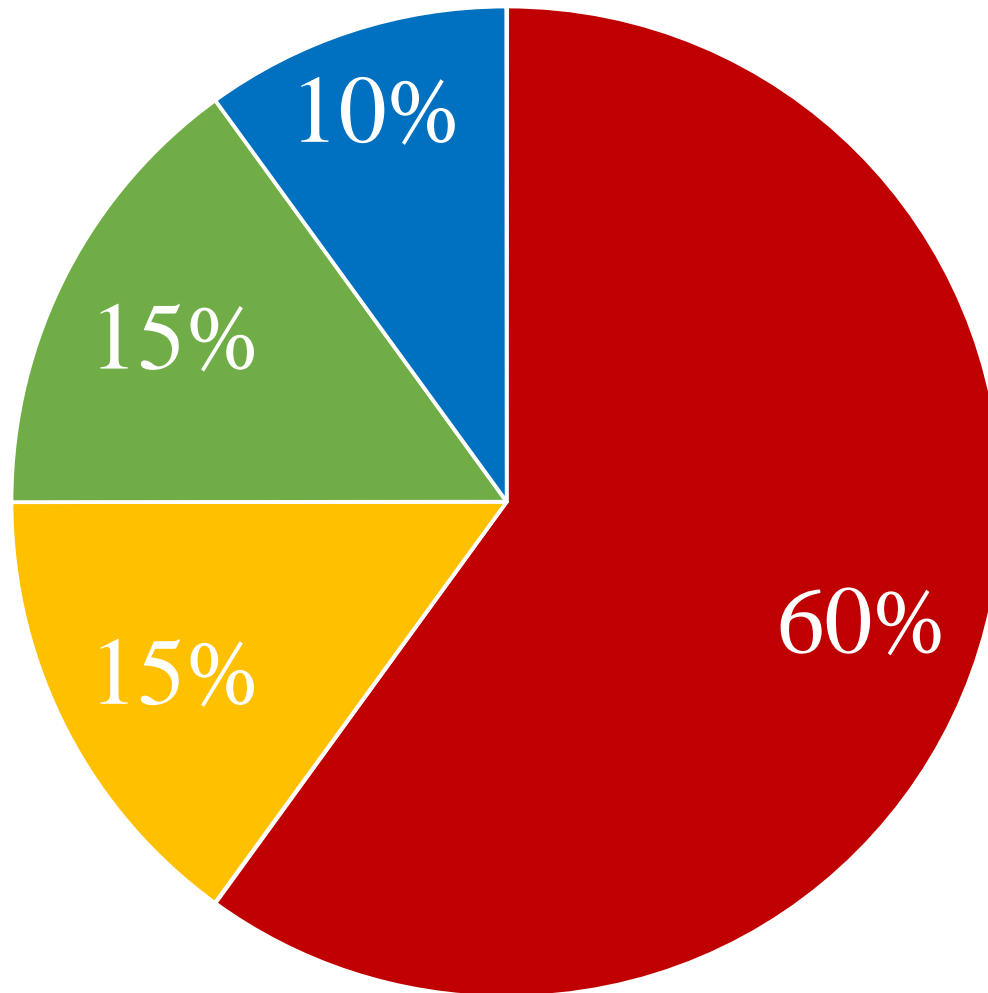
- Множество всех подпоследовательностей ряда, имеющих заданную длину

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

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>.

Рейтинг курса



■ **Практика, Часть 1**
(7 заданий)

■ **Практика, Часть 2**
(7 заданий)

■ **Практика, Часть 3**
(7 заданий)

■ **Тесты (5 тестов)**

Дедлайны курса

ОКТАБРЬ

2023

ПОНЕДЕЛЬНИК

ВТОРНИК

СРЕДА

ЧЕТВЕРГ

ПЯТНИЦА

СУББОТА

ВОСКРЕСЕНЬЕ

				1	2	3
4	5	16	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
1	25	27	28	29	30	

НОЯБРЬ

2023

ПОНЕДЕЛЬНИК

ВТОРНИК

СРЕДА

ЧЕТВЕРГ

ПЯТНИЦА

СУББОТА

ВОСКРЕСЕНЬЕ

		1	2	3	4	5
4	6	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30			

ДЕКАБРЬ

2023

ПОНЕДЕЛЬНИК

ВТОРНИК

СРЕДА

ЧЕТВЕРГ

ПЯТНИЦА

СУББОТА

ВОСКРЕСЕНЬЕ

				1	2	3
6	4	5	6	7	8	9
11	12	13	14	15	16	17
18	19	20	21	22	23	1, 2, 3, 4, 7
25	1, 2, 3, 4, 5, 6, 7	27	28	29	30	31

ЗАНЯТИЯ

i

Лекция по *i*-й теме

i

Практика по *i*-й теме

ЗАДАНИЯ

i

1-я часть задания по *i*-й теме

i

2-я часть задания по *i*-й теме

i

3-я часть задания по *i*-й теме

ТЕСТЫ

i

Тест по *i*-й теме

ТЕМЫ

1. Базовые понятия

3. Диссонансы
2. Поиск по образцу

4. Матричный профиль
5. Снимпеты

7. Восстановление и прогноз
6. Цепочки