CEFET/RJ

Descoberta de Motifs

Eduardo Ogasawara
eduardo.ogasawara@cefet-rj.br
https://eic.cefet-rj.br/~eogasawara

O que são Motifs?

- Motifs são subsequências que ocorrem repetidamente em uma série temporal
- Sua descoberta ajuda a identificar regularidades e padrões de comportamento
- Subsequências recorrentes em séries temporais
  - Capturam padrões frequentes ao longo do tempo
  - Importantes para detecção de comportamento normal ou repetitivo

Conceito de Discords

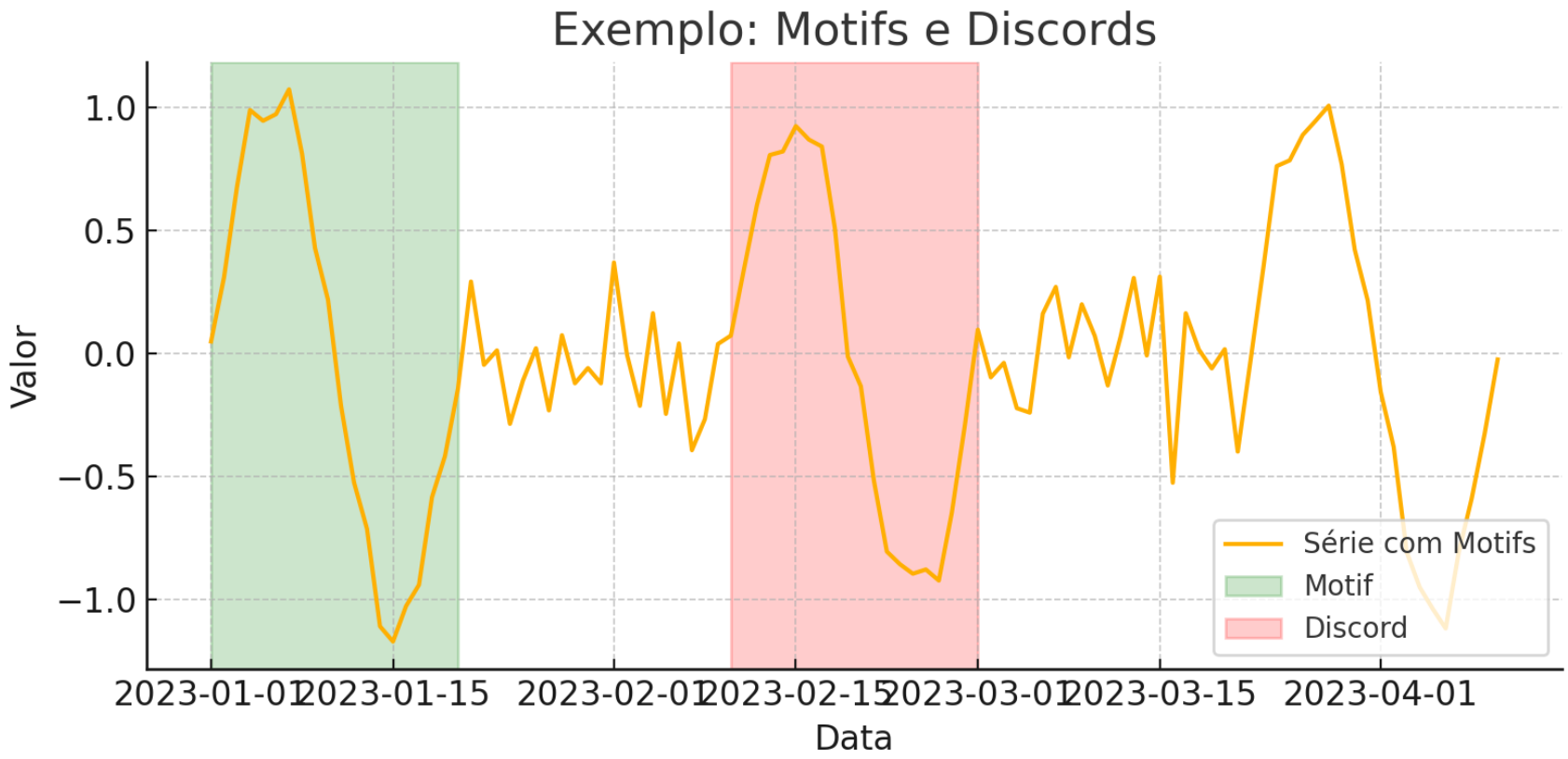
- Subsequências que ocorrem raramente ou são muito diferentes das demais
- Associadas a eventos incomuns ou anômalos
- Complementares aos motifs

Aplicações

- Detecção de padrões de consumo de energia
- Análise de comportamento em redes sociais
- Reconhecimento de gestos e atividades físicas
- Identificação de ciclos em sinais biológicos

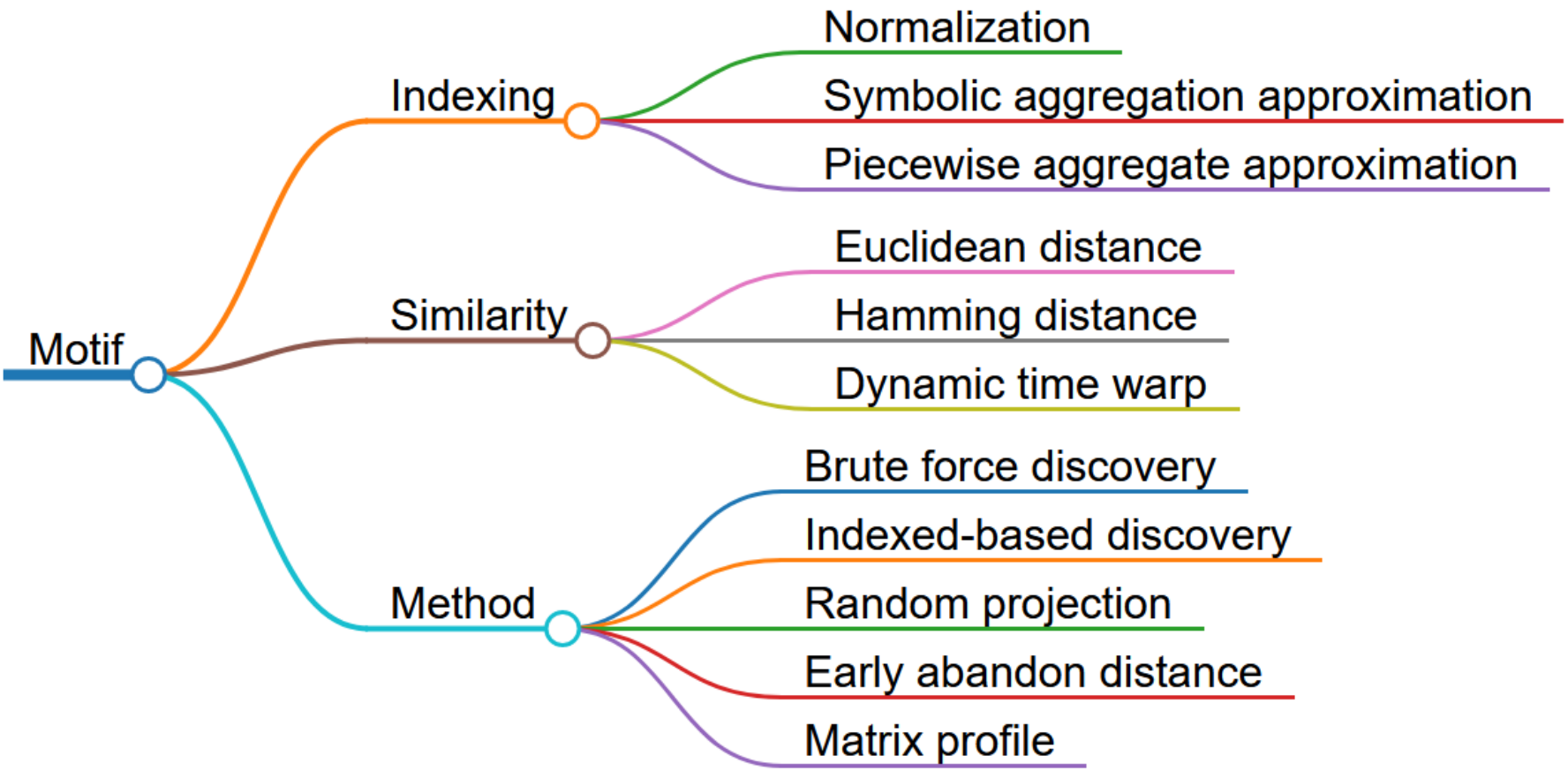
Exemplo de motifs e discords

- Exemplo visual mostrando subsequências frequentes (motifs) e raras (discords) em uma série temporal
- Os motifs representam padrões normais e os discords, possíveis anomalias



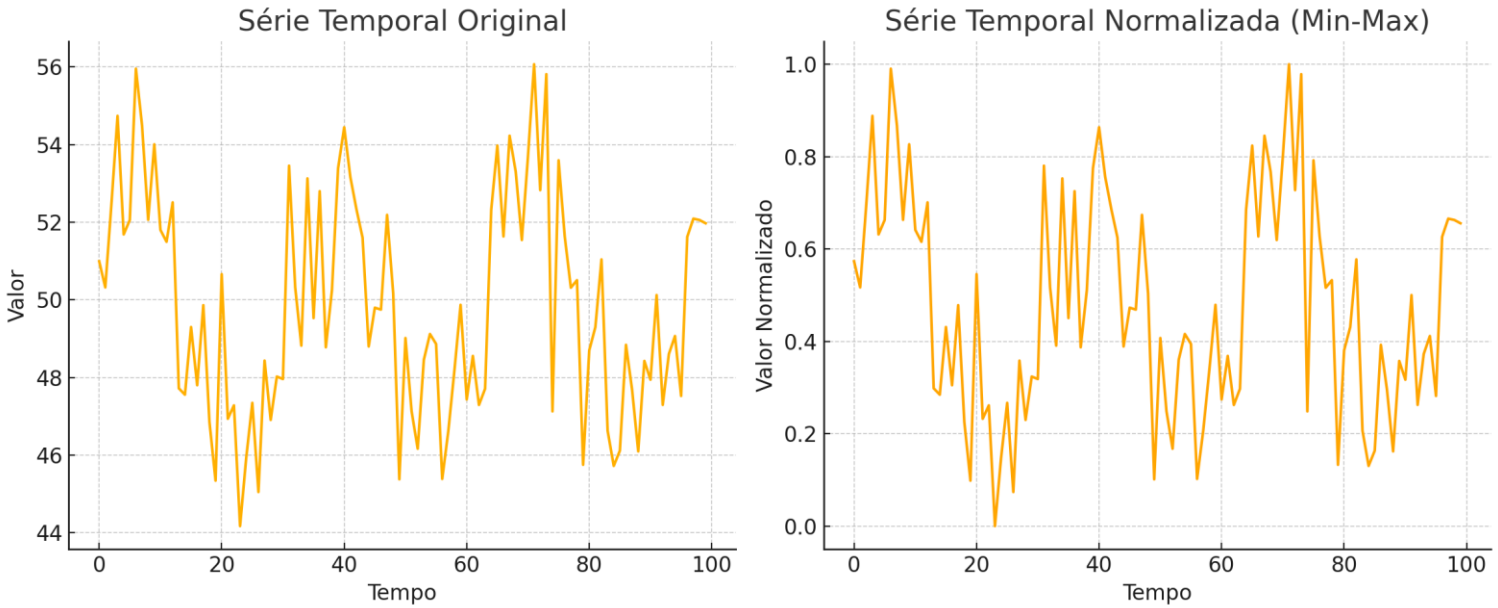
Taxonomia de Motifs

- Diferentes tipos de motifs podem ser classificados por:
  - Tamanho fixo ou variável
  - Frequência exata ou aproximada
  - Simples ou compostos
- A taxonomia guia a escolha de métodos de descoberta.



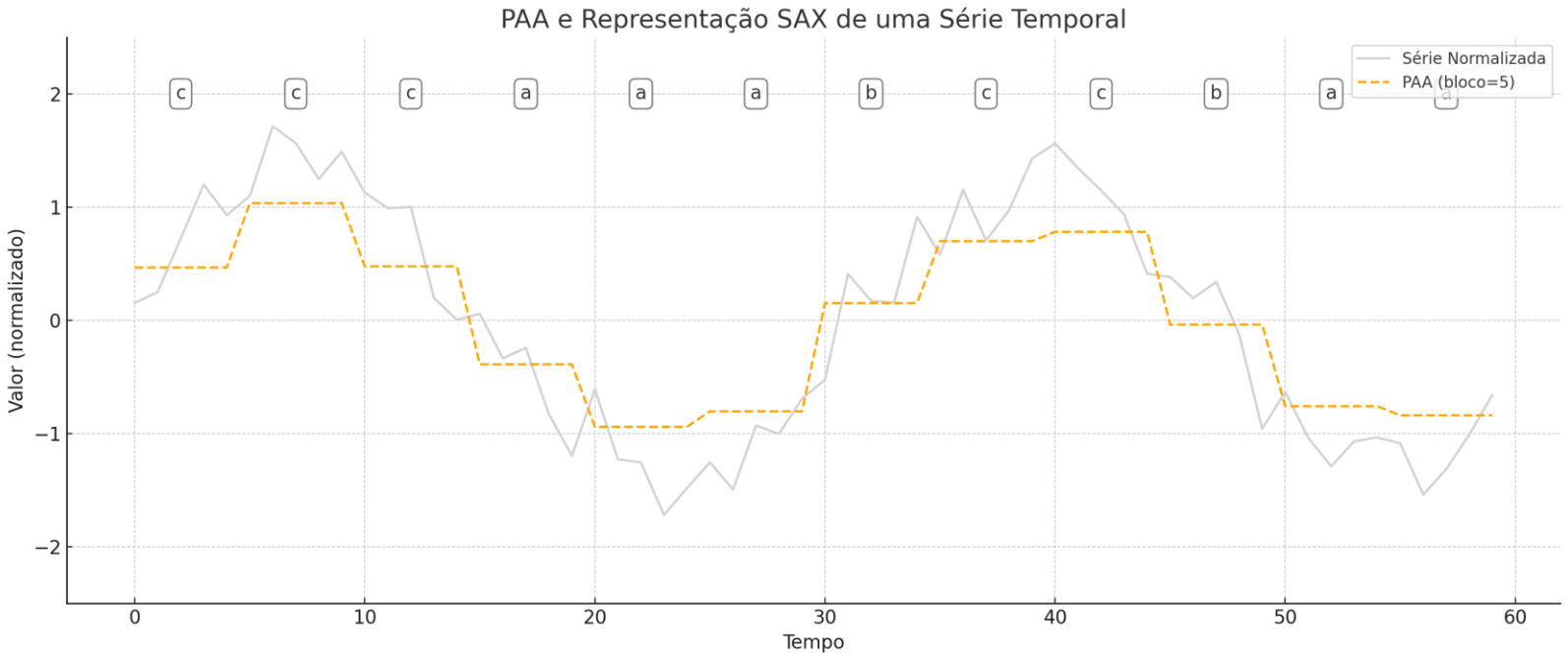
Normalização

- Normalizar subsequências é importante para evitar distorções causadas por escala ou deslocamento
- Métodos comuns incluem z-score e min-max
- Permite comparar padrões de forma justa



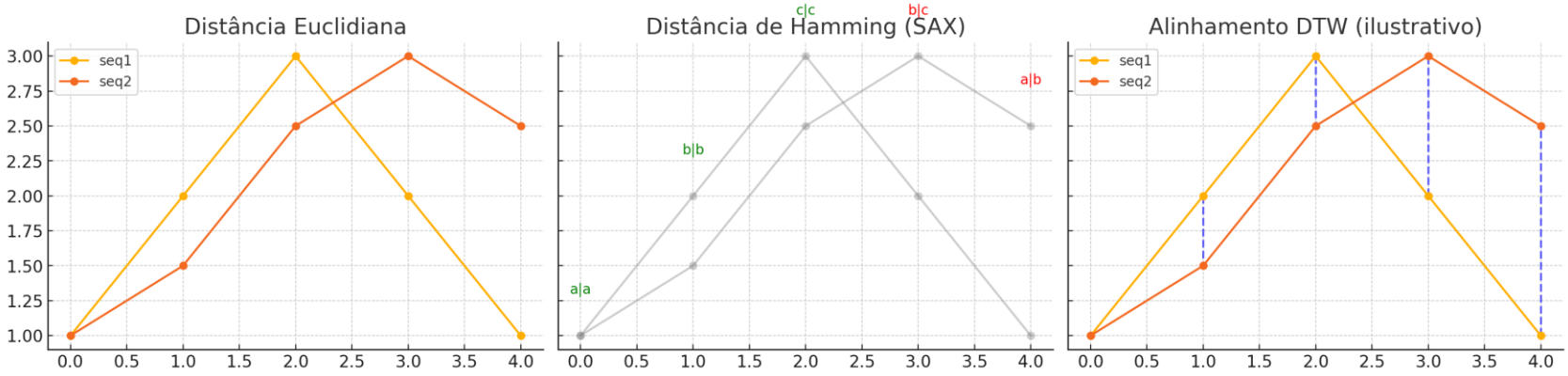
Representações SAX e PAA

- PAA (Piecewise Aggregate Approximation): divide a subsequência em blocos e calcula a média de cada um, gerando uma versão simplificada
- SAX (Symbolic Aggregate approXimation): transforma o PAA em símbolos, possibilitando compressão e comparação rápida
- A combinação das duas técnicas permite representar e comparar padrões com mais eficiência



Medidas de Similaridade

- Diferentes métricas são utilizadas para comparar subsequências:
  - Distância Euclidiana: simples, ponto a ponto
  - Distância de Hamming: para séries simbolizadas (ex: SAX)
  - DTW (Dynamic Time Warping): alinha padrões com variação temporal
- A escolha da medida afeta o desempenho e o tipo de motifs descobertos



	Métrica	Distância
0	Euclidiana	1.936492
1	Hamming (SAX)	2.000000
2	DTW (ilustrativa)	3.500000

Métodos de descoberta baseada em força bruta

- Compara todas as possíveis subsequências entre si.
- Garante exatidão, mas tem custo computacional elevado.
- Inviável para séries longas sem otimizações.

```
1:  procedure BruteForce(X, q, σ, ϵ)
2:      X' ← zscore(X)
3:      motifs ← ∅
4:      for i ← 1 to |X'| − q do
5:          p ← subseq(X', i, q)
6:          occurences ← i
7:          for j ← i + 1 to |X'| − q do
8:              p' ← subseq(X', j, q)
9:              if dist(p, p') < ϵ then
10:                 occurences ← occurrences ∪ j
11:         if |occurrences| ≥ σ then
12:             motifs ← motifs ∪ < p, occurrences >
13:     return motifs
```

Métodos de descoberta baseada em indexação

- Utiliza estruturas de índice para acelerar a busca por padrões semelhantes.
- Reduz o número de comparações necessárias.
- Compatível com técnicas como SAX.

1: **procedure** $IndexBased(X, a, k, q, \sigma)$
2: $X' \leftarrow zscore(X)$
3: $Y \leftarrow PAA(X', k)$
4: $Y' \leftarrow SAX(Y, a)$
5: $W \leftarrow sw(Y', q)$
6: $motifs \leftarrow group\_by\_having(W, \sigma)$
7: **return** $motifs$

Método de descoberta baseada em Random Projection

- Usa projeções aleatórias para agrupar subsequências semelhantes.
- Método aproximado, mas eficiente para grandes volumes de dados.

1: **procedure** $RandomProjection(X, q, ext, \sigma)$
2: $Y \leftarrow index(X)$
3: $W \leftarrow sw(Y, q + ext)$
4: $W' \leftarrow project(W, q)$
5: $motifs \leftarrow group\_by\_having(W', \sigma)$
6: **return** $motifs$

Método de descoberta baseada Early Abandon

- Durante a busca por motifs, as subsequências são comparadas usando medidas de similaridade (como a distância Euclidiana)
- Se, durante o cálculo da distância, a soma parcial ultrapassar o menor valor encontrado até o momento, a comparação pode ser interrompida antecipadamente
- Essa técnica é chamada de Early Abandon e evita cálculos desnecessários
- Não compromete a exatidão dos resultados, pois apenas descarta comparações que não seriam melhores
- É especialmente eficiente em métodos de força bruta com distância Euclidiana

Descoberta de motifs via Matrix Profile

- O Matrix Profile armazena, para cada subsequência, a menor distância para qualquer outra subsequência
- Permite identificar:
  - Motifs: padrões recorrentes
  - Discords: padrões únicos ou raros
- É a base de métodos como STAMP, STOMP, SCRIMP e AAMP

Descoberta de motifs (Baseada em Matrix Profile via STAMP)

- O STAMP calcula o Matrix Profile com base em permutação aleatória de subsequências
- Garante exatidão e eficiência, permitindo encontrar motifs e discords rapidamente
- Muito usado por ser simples, escalável e eficaz

1: **procedure** $STAMP(X, X' = nil, q)$
2: $Y' \leftarrow Y \leftarrow zscore(X)$
3: **if** $X' \neq nil$ **then**
4: $Y' \leftarrow zscore(X')$
5: $M \leftarrow$ infs, $I \leftarrow$ zeros
6: **for** $i \leftarrow 1$ **to** $|Y'| - q$ **do**
7: $seq \leftarrow subseq(Y', i, q)$
8: $D_i \leftarrow MASS(seq, Y)$
9: $M_i, I_i \leftarrow eleMin(M, I, D_i, i : (i + q - 1))$
10: **return** $\{P, I\}$

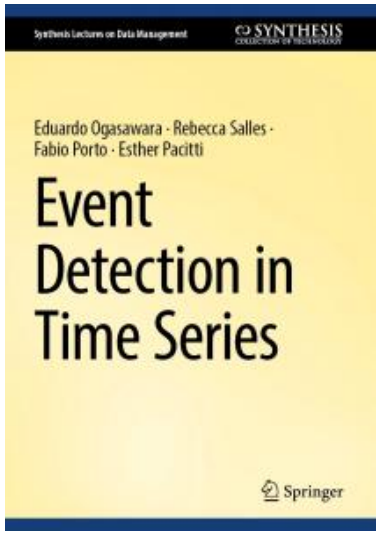
Avanços no tema

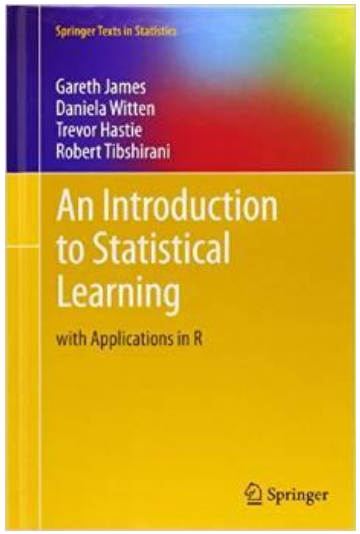
- Cenário multidimensional
- Motifs de tamanho variável
- Ranqueamento de motifs e ocorrências
- Tratamento de Big Data
- Novos métodos de descoberta: STOMP, AAMP

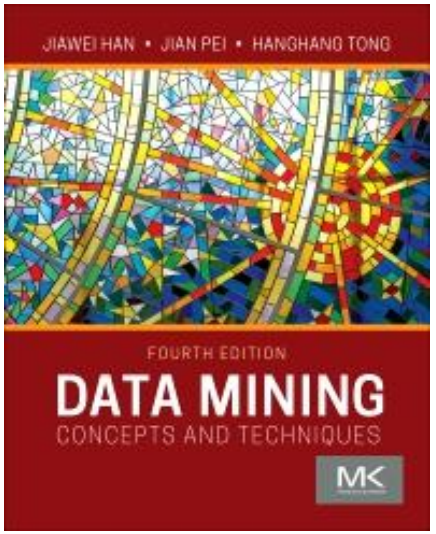
Resumo do Capítulo

- Motifs são subsequências frequentes, úteis para identificar padrões
- Discords são subsequências raras ou distintas, úteis para detectar anomalias
- Técnicas como Matrix Profile e SAX ajudam a descobri-los automaticamente
- Aplicações incluem biometria, energia, comportamento e mais

Referências







[1] Ogasawara, E.; Salles, R.; Porto, F.; Pacitti, E. Event Detection in Time Series. 1. ed. Cham: Springer Nature Switzerland, 2025.

[2] Cryer, J. D.; Chan, K.-S. Time Series Analysis: With Applications in R. Springer Science & Business Media, 2008.

[3] Han, J.; Pei, J.; Tong, H. Data Mining: Concepts and Techniques. 4th edition ed. Cambridge, MA: Morgan Kaufmann, 2022.

[4] James, G. M.; Witten, D.; Hastie, T.; Tibshirani, R. An Introduction to Statistical Learning: With Applications in R. [s.l.] Springer Nature, 2021.