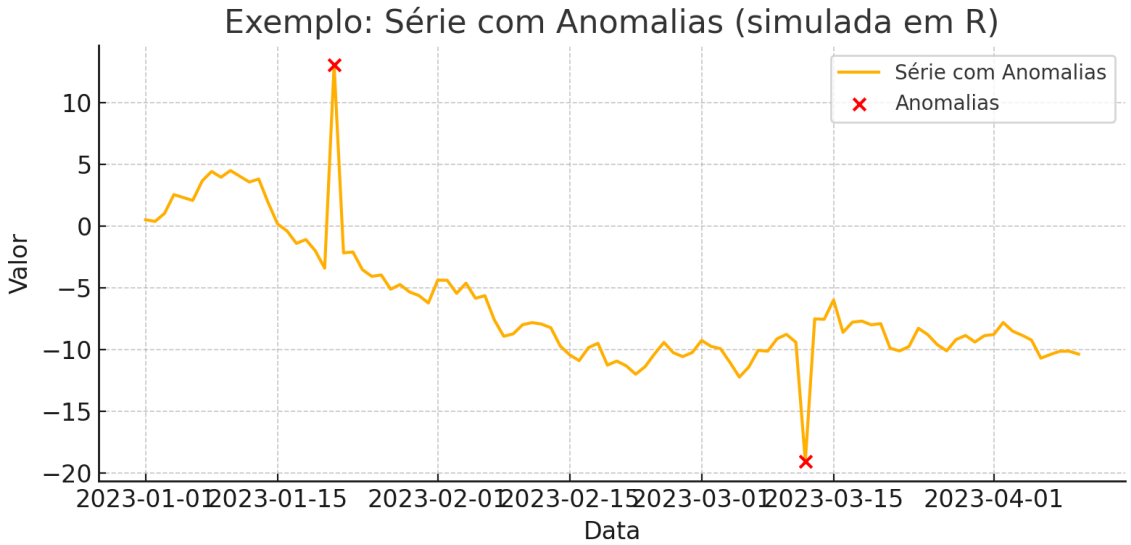
CEFET/RJ

# Detecção de anomalias

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

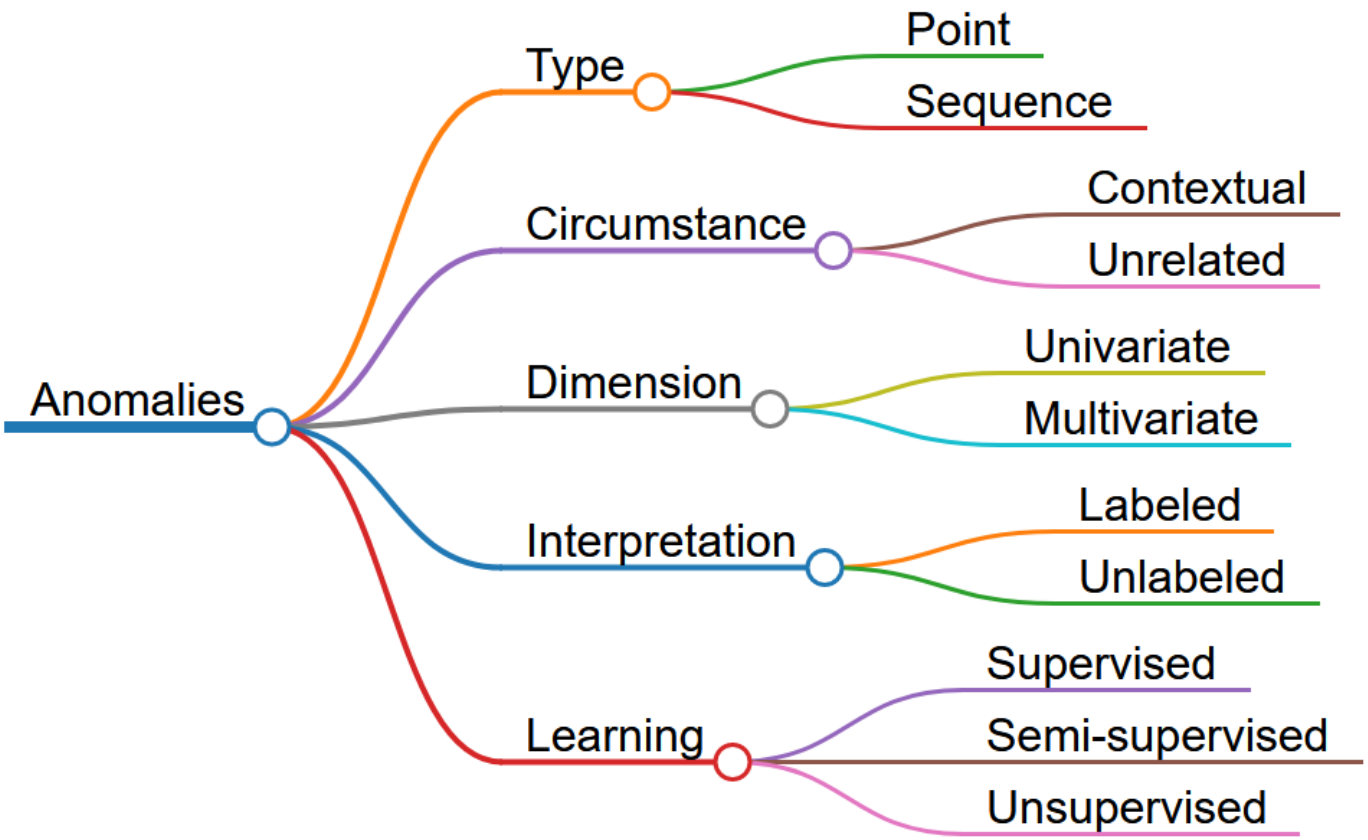
# *Conceito de Anomalia*

- Anomalias são observações que se desviam significativamente do padrão esperado
- Podem indicar falhas, fraudes, mudanças de comportamento ou eventos raros
- A detecção de anomalias é essencial em sistemas de monitoramento e vigilância automática
- $a(X, k, \sigma) = \{t, |tc(x_t) - ep(tc(x_t), k)| > \sigma \wedge |tc(x_t) - ef(tc(x_t), k)| > \sigma\}$



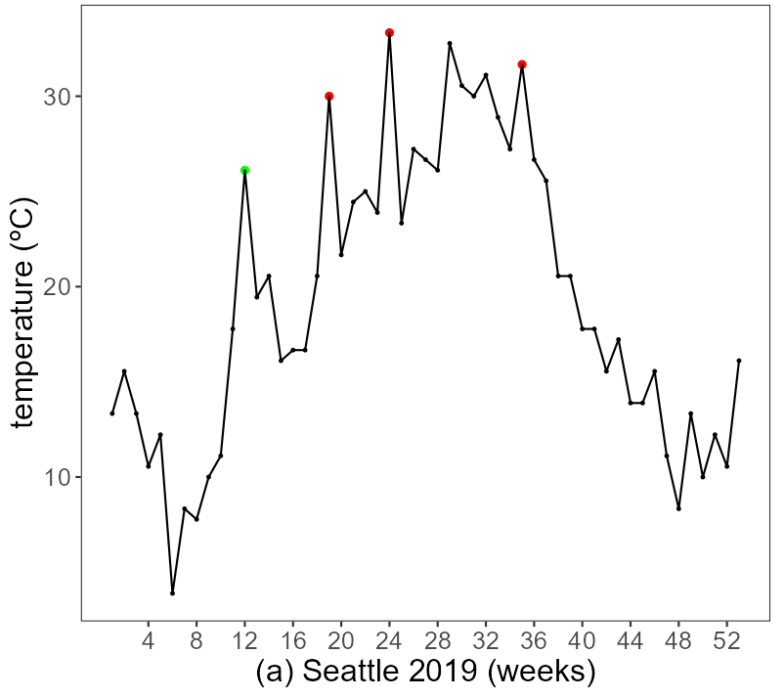
# *Taxonomia de Anomalia*

- Organizada por tipo, circunstância, dimensão, interpretação e aprendizado



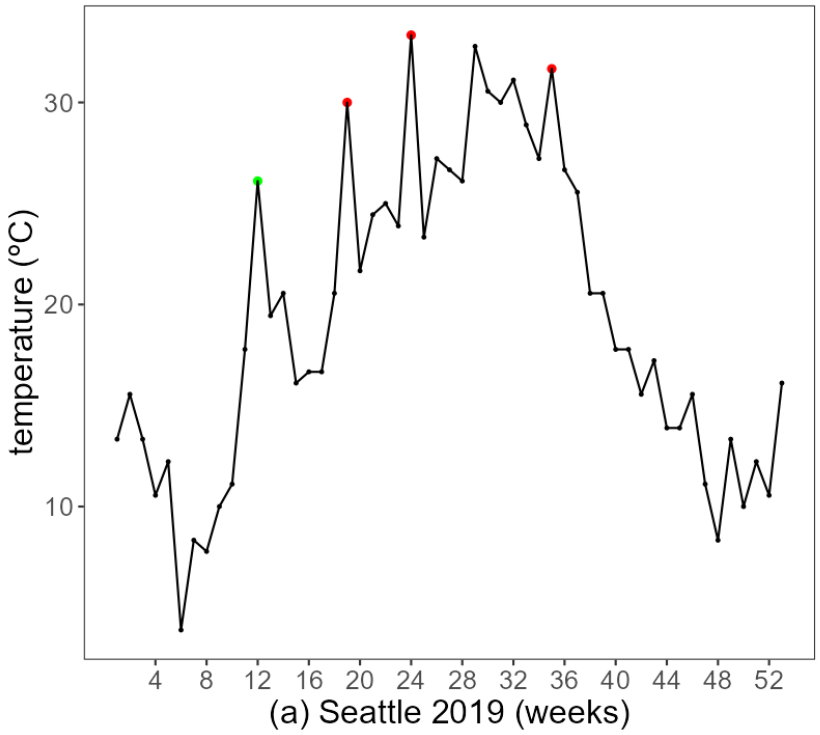
# *Tipos de Anomalias*

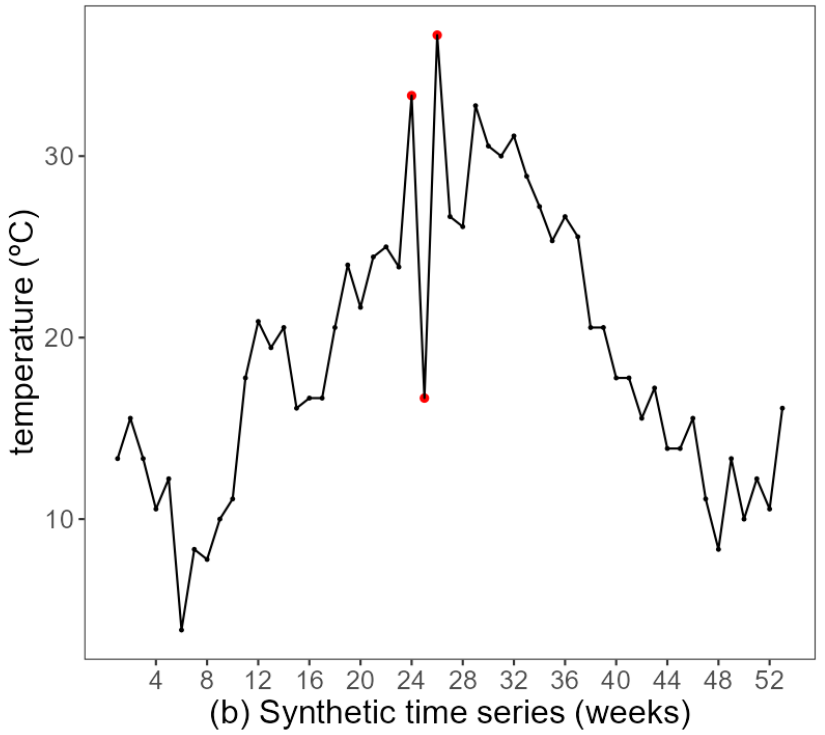
- Pontual: um único valor se destaca do padrão
- Coletiva (ou de sequência): um grupo de valores que, em conjunto, compõem um comportamento anômalo



# *Anomalia contextual e não-relacionada*

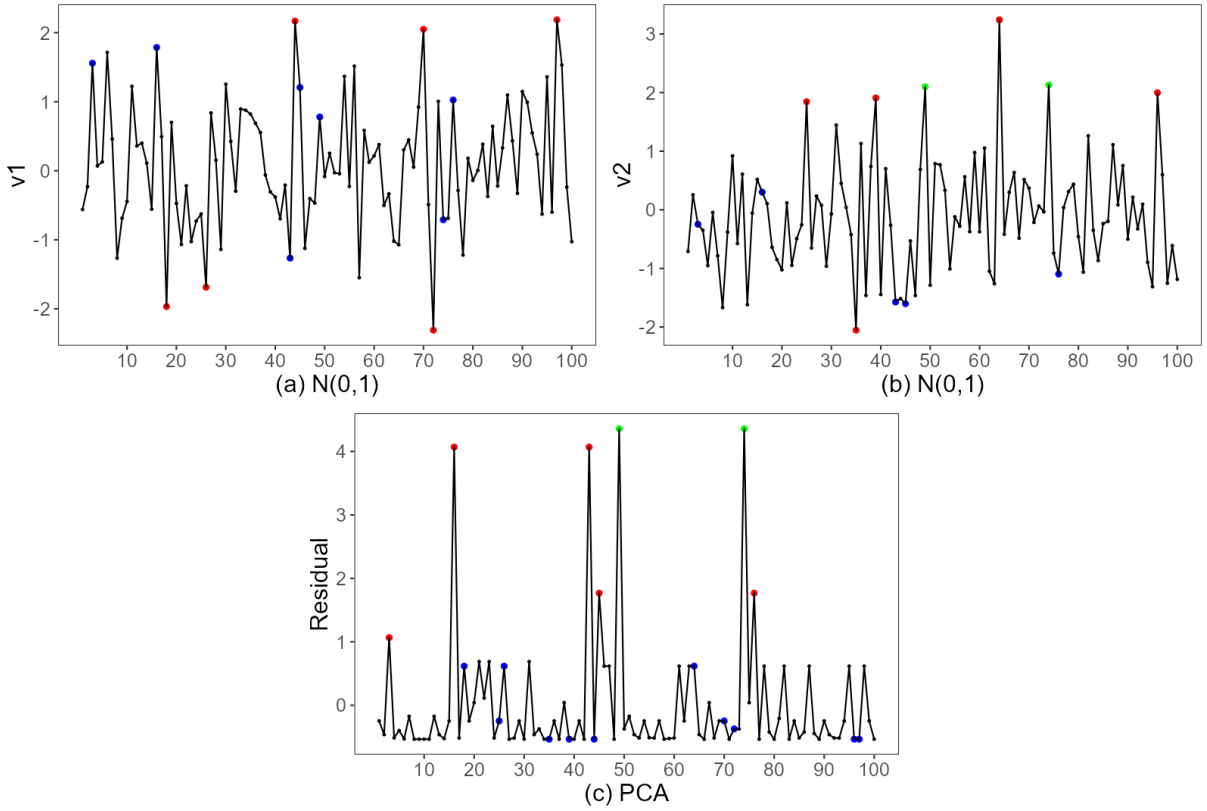
- Anomalias podem ser classificadas de acordo com seu comportamento e contexto:
  - Contextual: valores que parecem normais em geral, mas são anômalos dentro de um contexto (ex: temperatura alta à noite)
  - Não-relacionada: desvios que não seguem padrões conhecidos ou relações explícitas



# *Anomalia em séries univariadas e multivariadas*

- Univariadas: baseadas em uma única variável ao longo do tempo
- Multivariadas: envolvem a correlação entre várias variáveis (ex: temperatura, pressão e umidade)
- A detecção em séries multivariadas é mais complexa e requer métodos que capturam relações cruzadas



# *Rotulagem de Anomalias*

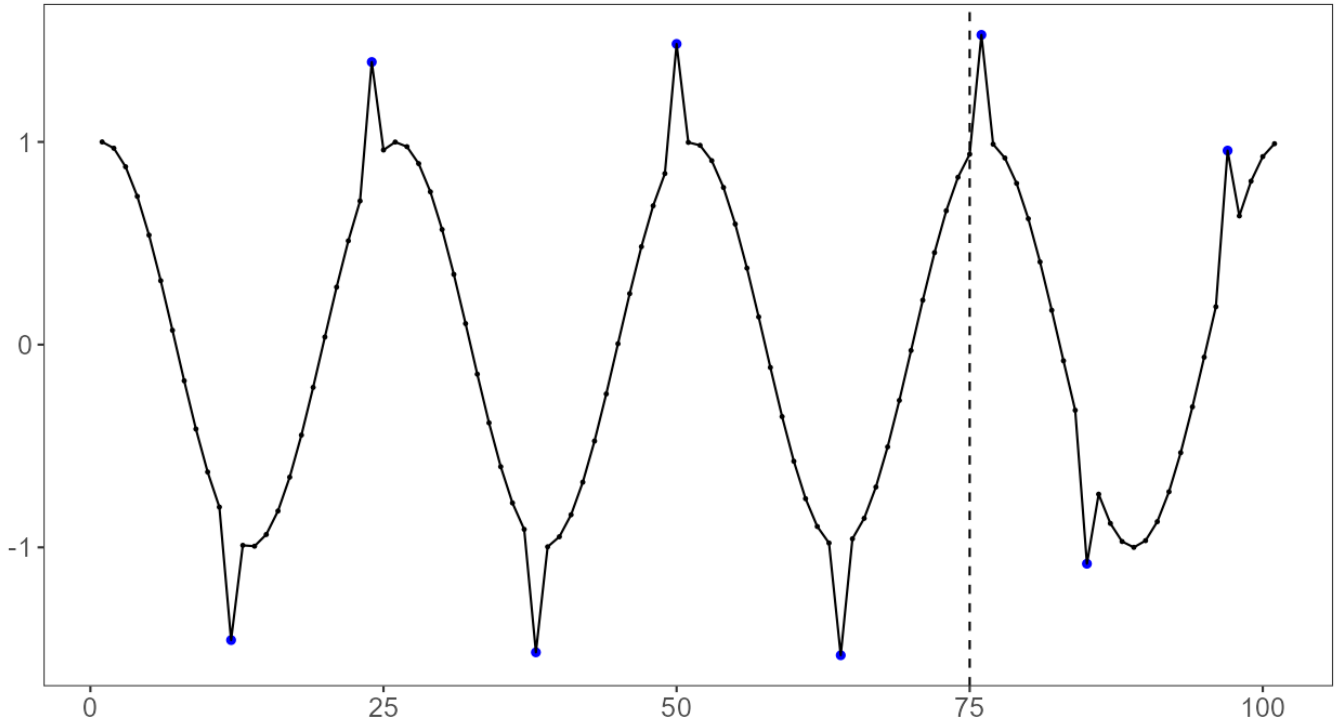
- Rotulada: anomalias conhecidas e marcadas nos dados
- Não rotulada: o modelo precisa inferir o que é anômalo
- Riscos: rótulos podem conter viés humano ou ser incompletos, afetando a aprendizagem supervisionada

# *Abordagens de Aprendizado*

- Supervisionado: requer exemplos rotulados de comportamento normal e anômalo.
  - É eficaz, mas raramente viável devido à escassez de rótulos confiáveis
- Não supervisionado: não requer rótulos.
  - O modelo tenta encontrar padrões que se desviam da maioria, geralmente por meio de distância, densidade ou reconstrução
- Semi-supervisionado: assume que temos apenas exemplos de comportamento normal
  - O modelo é treinado para aprender esse padrão e, durante a detecção, qualquer desvio é tratado como anomalia

# *Separação Temporal de Treino e Teste*

- A separação entre treino e teste deve respeitar a ordem temporal
- Técnicas comuns:
  - Holdout temporal: divisão única por tempo
  - Validação cruzada para séries temporais: blocos crescentes (TimeSeriesSplit)
- Evita vazamento de informações do futuro para o passado



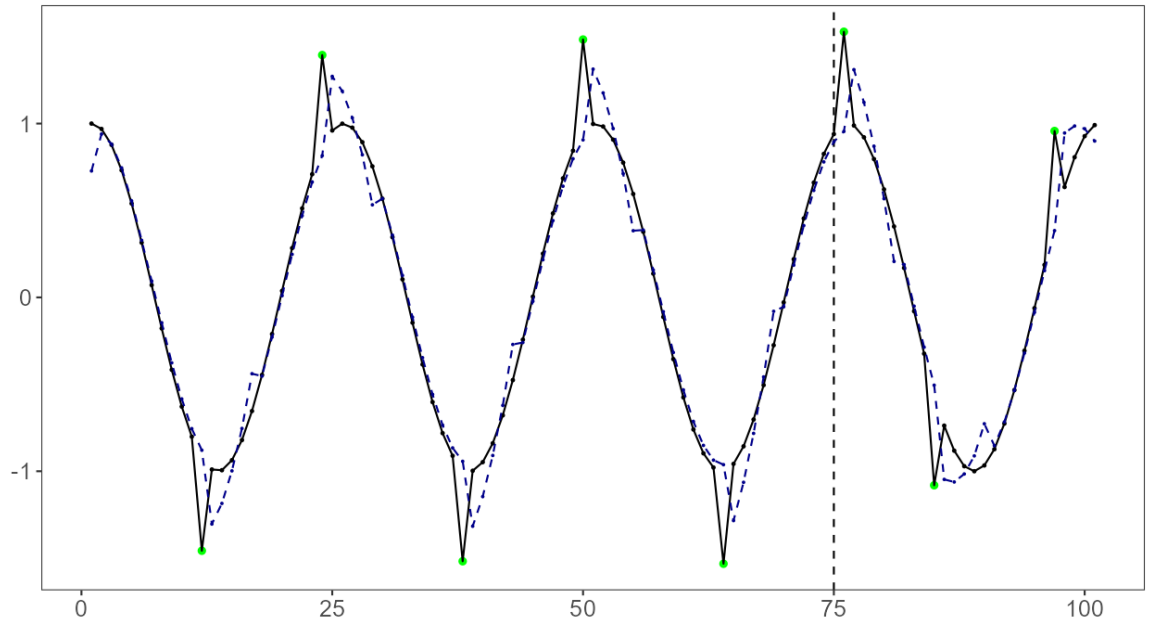
# *Detecção com Regressão*

- Ajusta modelos preditivos (ex: ARIMA) para prever o próximo valor
- Se o erro de previsão for alto, o ponto pode ser considerado anômalo
- Exige séries estacionárias e bom ajuste do modelo

$\omega_t$

| $t$ | $x_{t-4}$ | $x_{t-3}$ | $x_{t-2}$ | $x_{t-1}$ | $\hat{x}_t$ | $x_t$ |
|---|---|---|---|---|---|---|
| 5 | $v_1$ | $v_2$ | $v_3$ | $v_4$ | $\hat{v}_5$ | $v_5$ |
| 6 | $v_2$ | $v_3$ | $v_4$ | $v_5$ | $\hat{v}_6$ | $v_6$ |
| 7 | $v_3$ | $v_4$ | $v_5$ | $v_6$ | $\hat{v}_7$ | $v_7$ |
| 8 | $v_4$ | $v_5$ | $v_6$ | $v_7$ | $\hat{v}_8$ | $v_8$ |
| 9 | $v_5$ | $v_6$ | $v_7$ | $v_8$ | $\hat{v}_9$ | $v_9$ |
| 10 | $v_6$ | $v_7$ | $v_8$ | $v_9$ | $\hat{v}_{10}$ | $v_{10}$ |
| 11 | $v_7$ | $v_8$ | $v_9$ | $v_{10}$ | $\hat{v}_{11}$ | $v_{11}$ |
| 12 | $v_8$ | $v_9$ | $v_{10}$ | $v_{11}$ | $\hat{v}_{12}$ | $v_{12}$ |
| 13 | $v_9$ | $v_{10}$ | $v_{11}$ | $v_{12}$ | $\hat{v}_{13}$ | $v_{13}$ |
| 14 | $v_{10}$ | $v_{11}$ | $v_{12}$ | $v_{13}$ | $\hat{v}_{14}$ | $v_{14}$ |

(a)

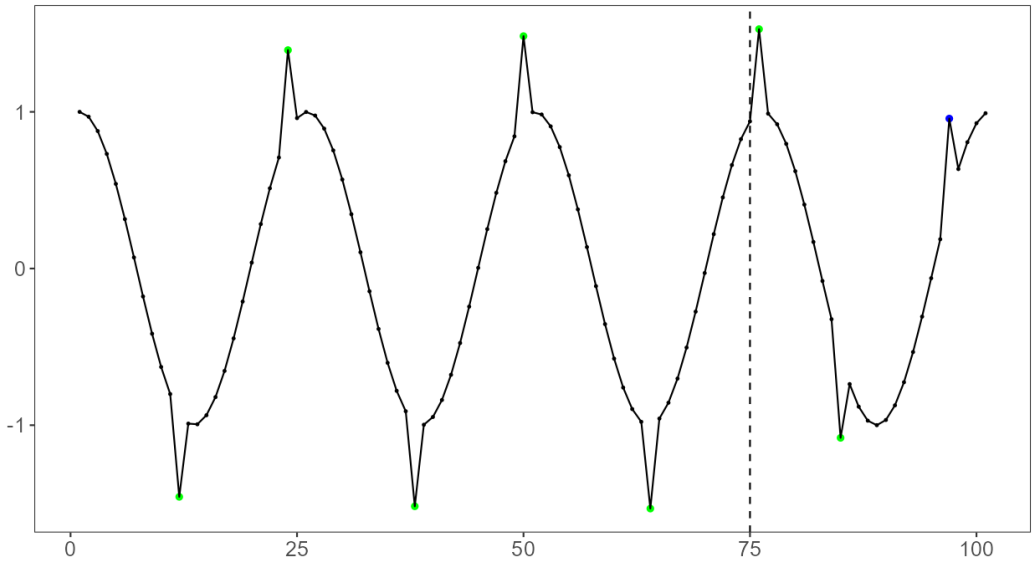


# *Detecção com Classificação*

- Treina um modelo com exemplos rotulados de normal/anômalo
- Classificadores comuns: SVM, decision trees, redes neurais
- Requer conjunto de treino rotulado e balanceado


| $t$ | $x_{t-4}$ | $x_{t-3}$ | $x_{t-2}$ | $x_{t-1}$ | $x_t$ | $\hat{e}_t$ | $e_t$ |
|---|---|---|---|---|---|---|---|
| **5** | $v_1$ | $v_2$ | $v_3$ | $v_4$ | $v_5$ | $\hat{b}_5$ | $b_5$ |
| **6** | $v_2$ | $v_3$ | $v_4$ | $v_5$ | $v_6$ | $\hat{b}_6$ | $b_6$ |
| **7** | $v_3$ | $v_4$ | $v_5$ | $v_6$ | $v_7$ | $\hat{b}_7$ | $b_7$ |
| **8** | $v_4$ | $v_5$ | $v_6$ | $v_7$ | $v_8$ | $\hat{b}_8$ | $b_8$ |
| **9** | $v_5$ | $v_6$ | $v_7$ | $v_8$ | $v_9$ | $\hat{b}_9$ | $b_9$ |
| **10** | $v_6$ | $v_7$ | $v_8$ | $v_9$ | $v_{10}$ | $\hat{b}_{10}$ | $b_{10}$ |
| **11** | $v_7$ | $v_8$ | $v_9$ | $v_{10}$ | $v_{11}$ | $\hat{b}_{11}$ | $b_{11}$ |
| **12** | $v_8$ | $v_9$ | $v_{10}$ | $v_{11}$ | $v_{12}$ | $\hat{b}_{12}$ | $b_{12}$ |

| $t$ | $x_{t-4}$ | $x_{t-3}$ | $x_{t-2}$ | $x_{t-1}$ | $x_t$ | $\hat{e}_t$ |
|---|---|---|---|---|---|---|
| **13** | $v_9$ | $v_{10}$ | $v_{11}$ | $v_{12}$ | $v_{13}$ | $\hat{b}_{13}$ |
| **14** | $v_{10}$ | $v_{11}$ | $v_{12}$ | $v_{13}$ | $v_{14}$ | $\hat{b}_{14}$ |

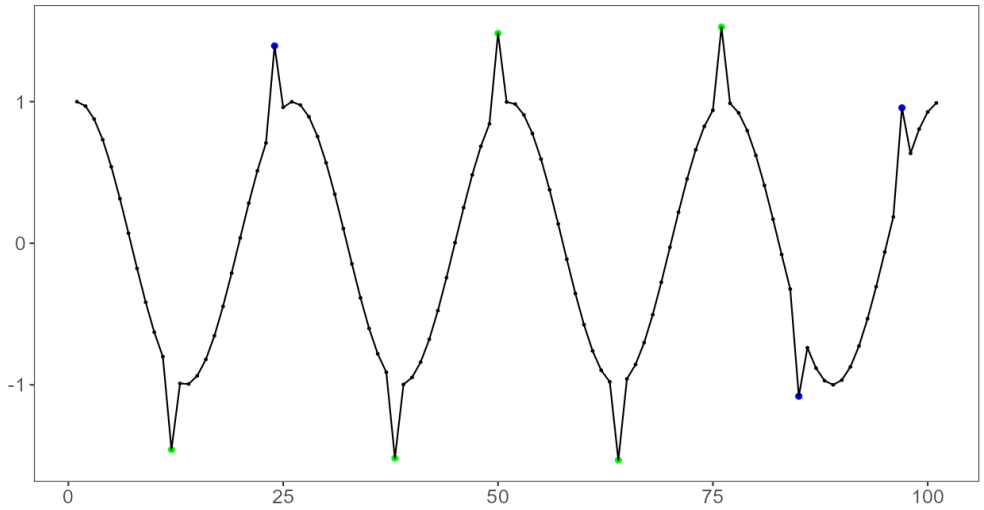


# *Detecção com Agrupamento (Clustering)*

- Modelos de clustering (ex: k-means, DBSCAN) detectam padrões e outliers
- Observações que não se encaixam bem em nenhum cluster são tratadas como anômalas
- Útil para aprendizado não supervisionado

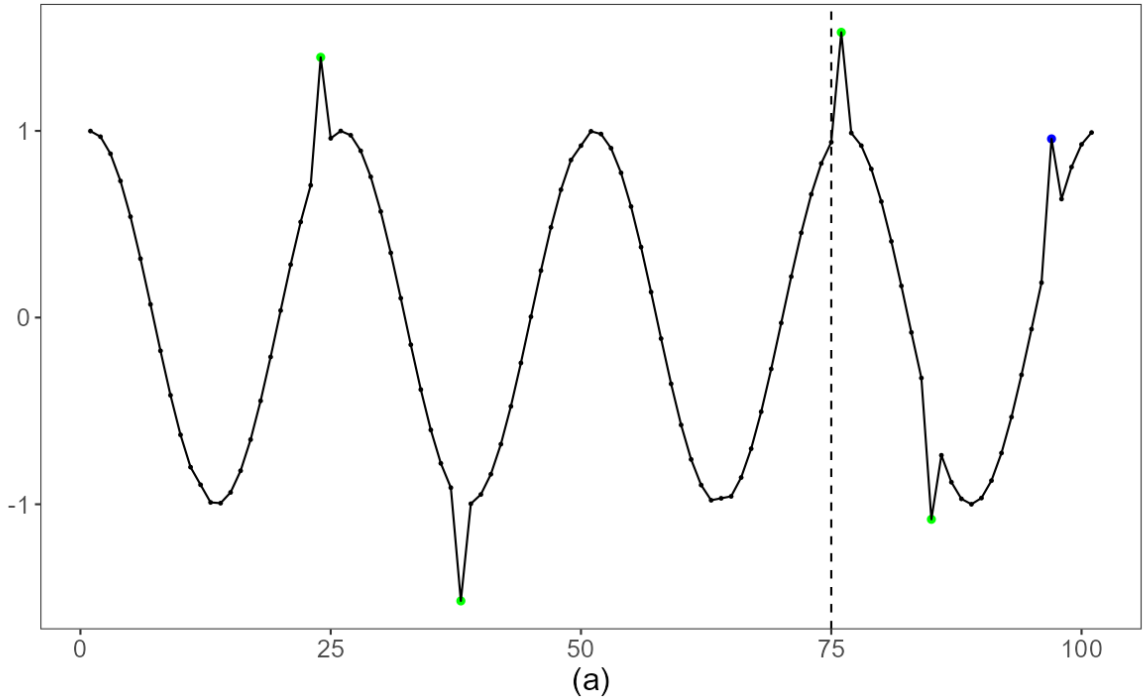
| $t$ | $x_{t-4}$ | $x_{t-3}$ | $x_{t-2}$ | $x_{t-1}$ | $x_t$ | $\ddot{r}_c$ | $d_t$ |
|---|---|---|---|---|---|---|---|
| 5 | $v_1$ | $v_2$ | $v_3$ | $v_4$ | $v_5$ | $\ddot{r}_1$ | $d_5$ |
| 6 | $v_2$ | $v_3$ | $v_4$ | $v_5$ | $v_6$ | $\ddot{r}_1$ | $d_6$ |
| 7 | $v_3$ | $v_4$ | $v_5$ | $v_6$ | $v_7$ | $\ddot{r}_1$ | $d_7$ |
| 8 | $v_4$ | $v_5$ | $v_6$ | $v_7$ | $v_8$ | $\ddot{r}_2$ | $d_8$ |
| 9 | $v_5$ | $v_6$ | $v_7$ | $v_8$ | $v_9$ | $\ddot{r}_2$ | $d_9$ |
| 10 | $v_6$ | $v_7$ | $v_8$ | $v_9$ | $v_{10}$ | $\ddot{r}_1$ | $d_{10}$ |
| 11 | $v_7$ | $v_8$ | $v_9$ | $v_{10}$ | $v_{11}$ | $\ddot{r}_1$ | $d_{11}$ |
| 12 | $v_8$ | $v_9$ | $v_{10}$ | $v_{11}$ | $v_{12}$ | $\ddot{r}_2$ | $d_{12}$ |
| 13 | $v_9$ | $v_{10}$ | $v_{11}$ | $v_{12}$ | $v_{13}$ | $\ddot{r}_2$ | $d_{13}$ |
| 14 | $v_{10}$ | $v_{11}$ | $v_{12}$ | $v_{13}$ | $v_{14}$ | $\ddot{r}_2$ | $d_{14}$ |

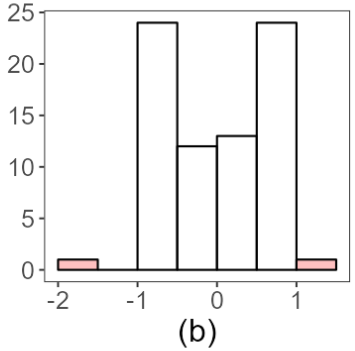
| $\ddot{r}_t$ | $x_{t-4}$ | $x_{t-3}$ | $x_{t-2}$ | $x_{t-1}$ | $x_t$ |
|---|---|---|---|---|---|
| $\ddot{r}_1$ | $\ddot{v}_{1,4}$ | $\ddot{v}_{1,3}$ | $\ddot{v}_{1,2}$ | $\ddot{v}_{1,1}$ | $\ddot{v}_{1,0}$ |
| $\ddot{r}_2$ | $\ddot{v}_{2,4}$ | $\ddot{v}_{2,3}$ | $\ddot{v}_{2,2}$ | $\ddot{v}_{2,1}$ | $\ddot{v}_{2,0}$ |



# *Detecção Estatística*

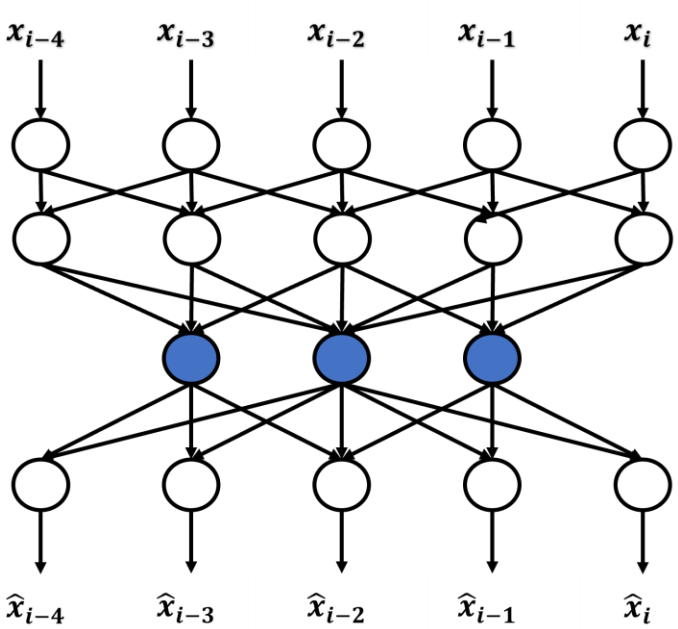
- Métodos clássicos de detecção:
  - Z-score: desvio em relação à média
  - IQR: limites baseados em quartis
  - Histograma: densidade de ocorrência de valores
- Simples, rápidos, mas sensíveis a distribuição

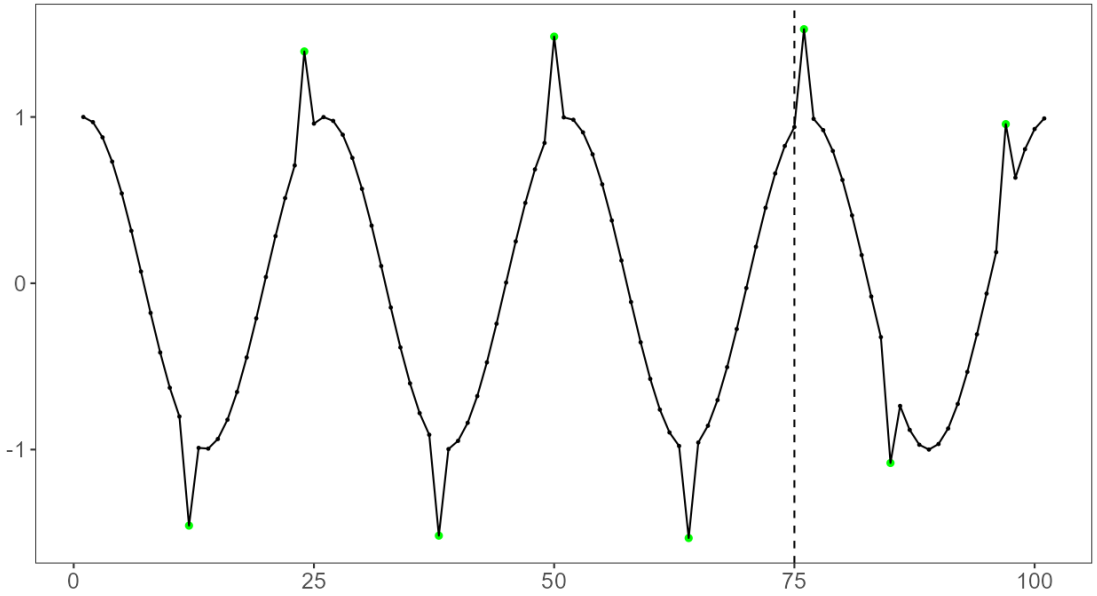


# *Detecção baseada em métodos espectrais*

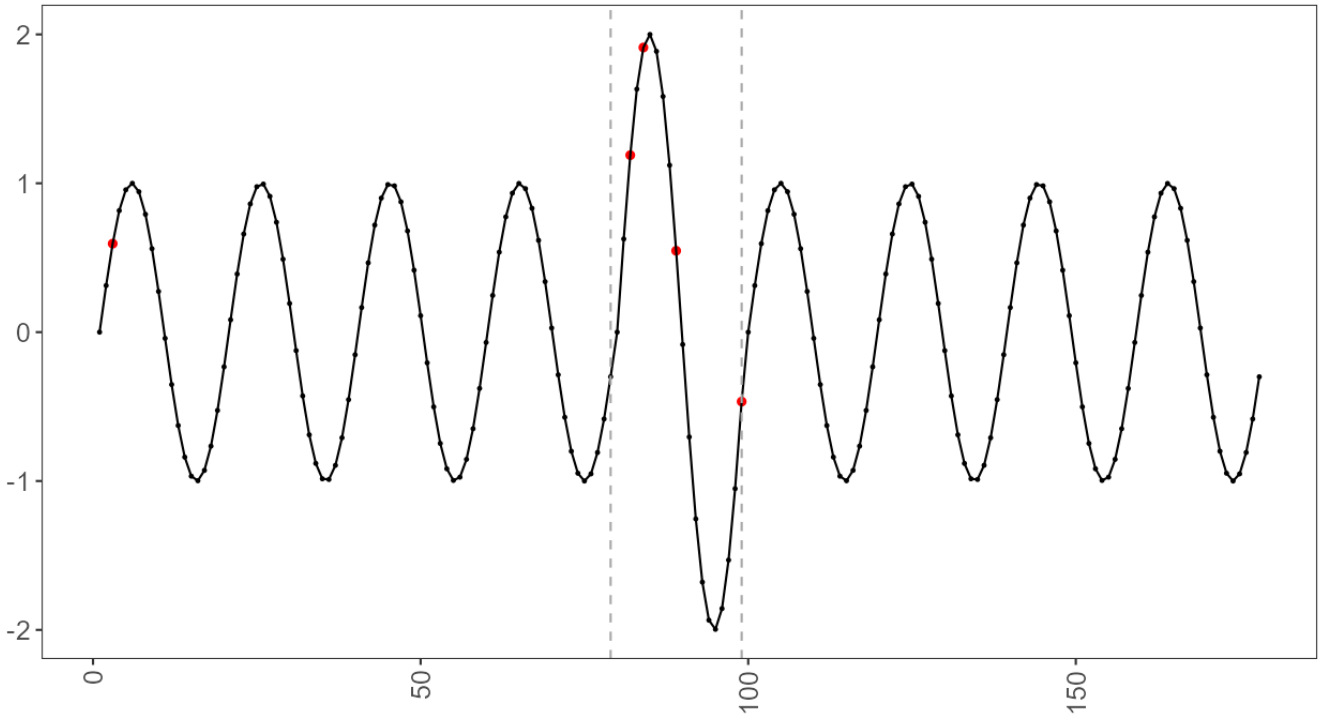
- Autoencoders aprendem a reconstruir o padrão normal
- Erros de reconstrução elevados indicam anomalias
- Abordagem eficiente para séries complexas e multivariadas



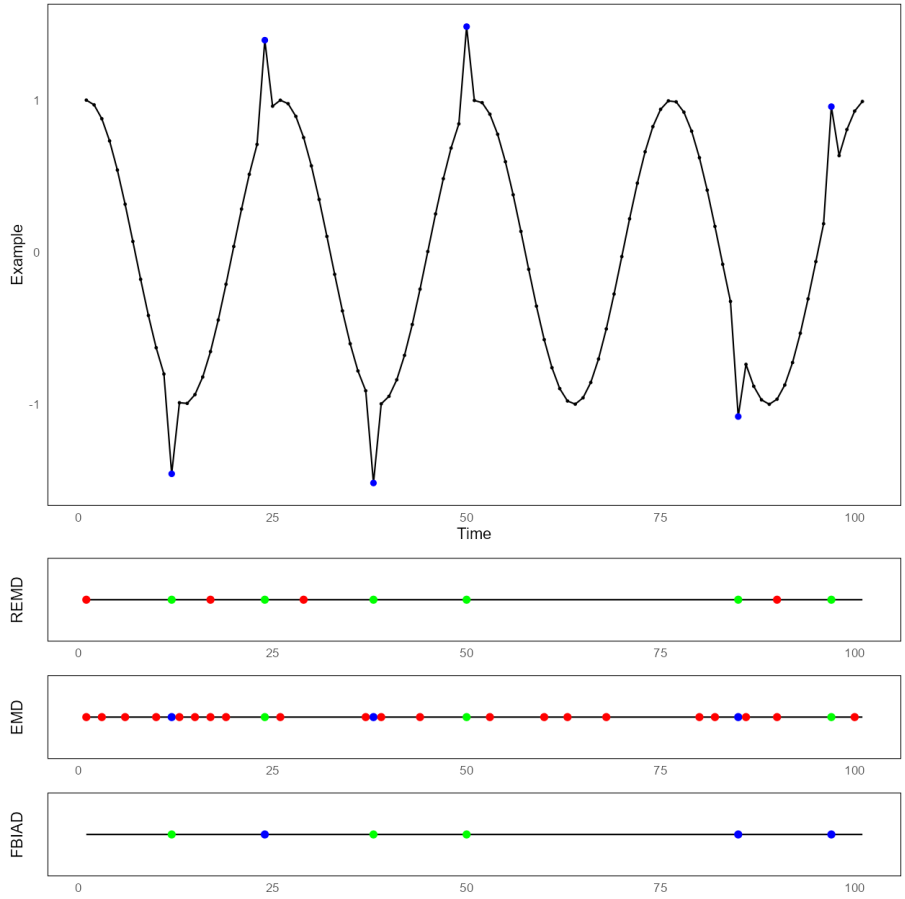
# *Anomalia de volatilidade*

- Em alguns domínios, o foco não está no valor em si, mas em sua variabilidade
- Exemplo: modelo GARCH detecta mudanças no comportamento da variância ao longo do tempo
- Importante em finanças e séries com flutuações erráticas



# *Comparação de detecções*

- Comparações visuais destacam as diferenças entre métodos de detecção
- Cores indicam acertos (verde), falsos positivos (vermelho) e falsos negativos (azul)
- Facilitam a análise qualitativa dos resultados



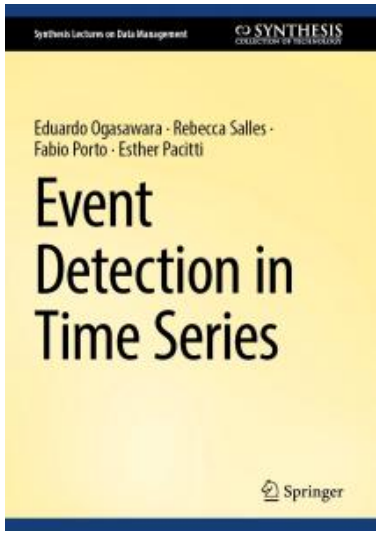
# *Resumo do Capítulo*

- Anomalias são elementos-chave em sistemas de detecção e predição
- Podem ser pontuais, contextuais ou coletivas
- Diferentes métodos são aplicáveis:
  - Estatísticos
  - Baseados em distância
  - Regressão, classificação, clustering, autoencoders
- A escolha do método depende da natureza da série e da disponibilidade de rótulos

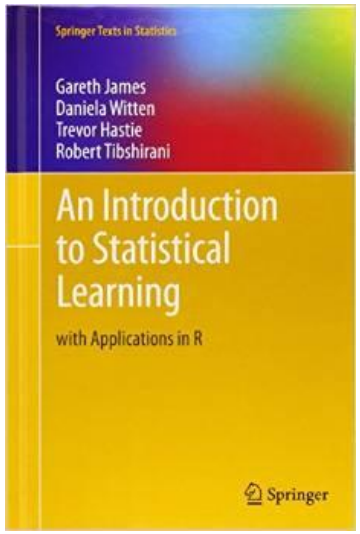
# *Exercício Harbinger*

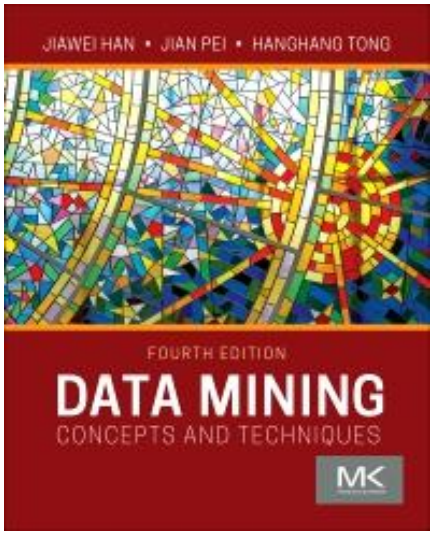
- Explore a página do Harbinger
  - https://cefet-rj-dal.github.io/harbinger/
- Execute todos os exemplos de anomalias
- Escolha uma série temporal e refaça os exemplos com esta nova série
  - Eventualmente será necessário olhar os exemplos gerais

# 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.