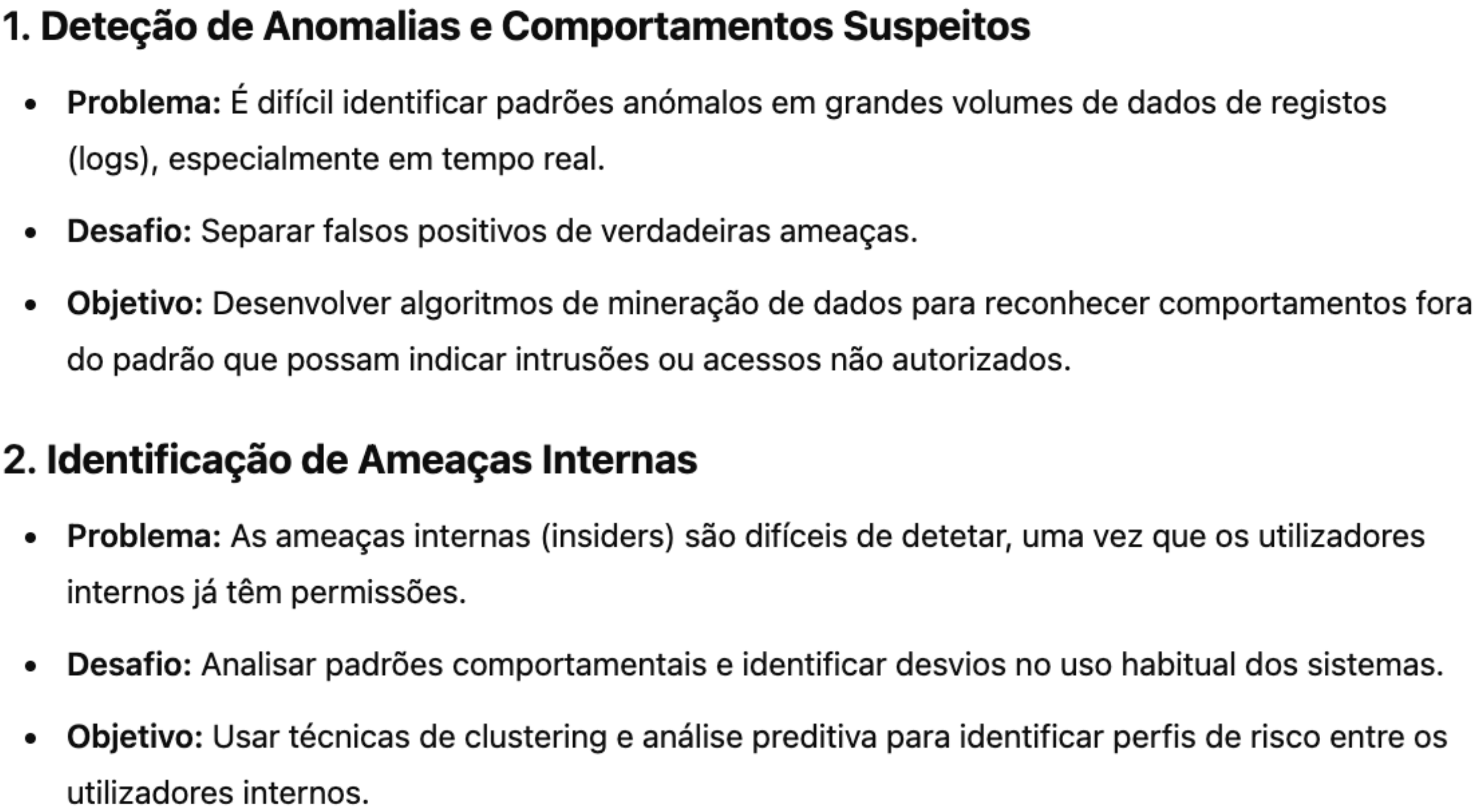
Datamining for Security Auditing

António Gonçalves

Source: Slides by Prof Victor Lobo

Master's in Information Security and Law in Cyberspace

Datamining for Security Auditing



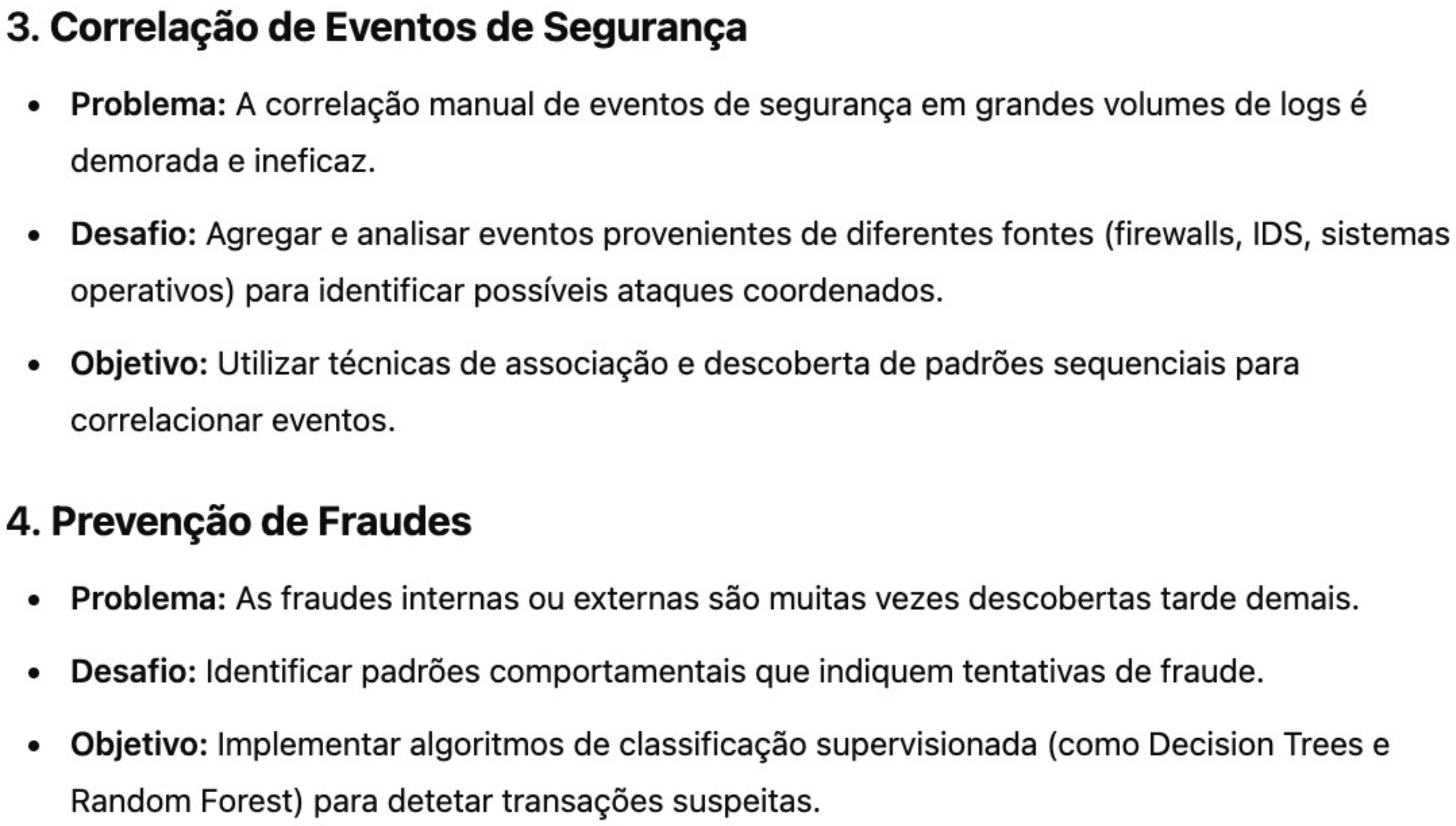
1. Deteção de Anomalias e Comportamentos Suspeitos

- **Problema:** É difícil identificar padrões anómalos em grandes volumes de dados de registos (logs), especialmente em tempo real.
- **Desafio:** Separar falsos positivos de verdadeiras ameaças.
- **Objetivo:** Desenvolver algoritmos de mineração de dados para reconhecer comportamentos fora do padrão que possam indicar intrusões ou acessos não autorizados.

2. Identificação de Ameaças Internas

- **Problema:** As ameaças internas (insiders) são difíceis de detetar, uma vez que os utilizadores internos já têm permissões.
- **Desafio:** Analisar padrões comportamentais e identificar desvios no uso habitual dos sistemas.
- **Objetivo:** Usar técnicas de clustering e análise preditiva para identificar perfis de risco entre os utilizadores internos.

Datamining for Security Auditing



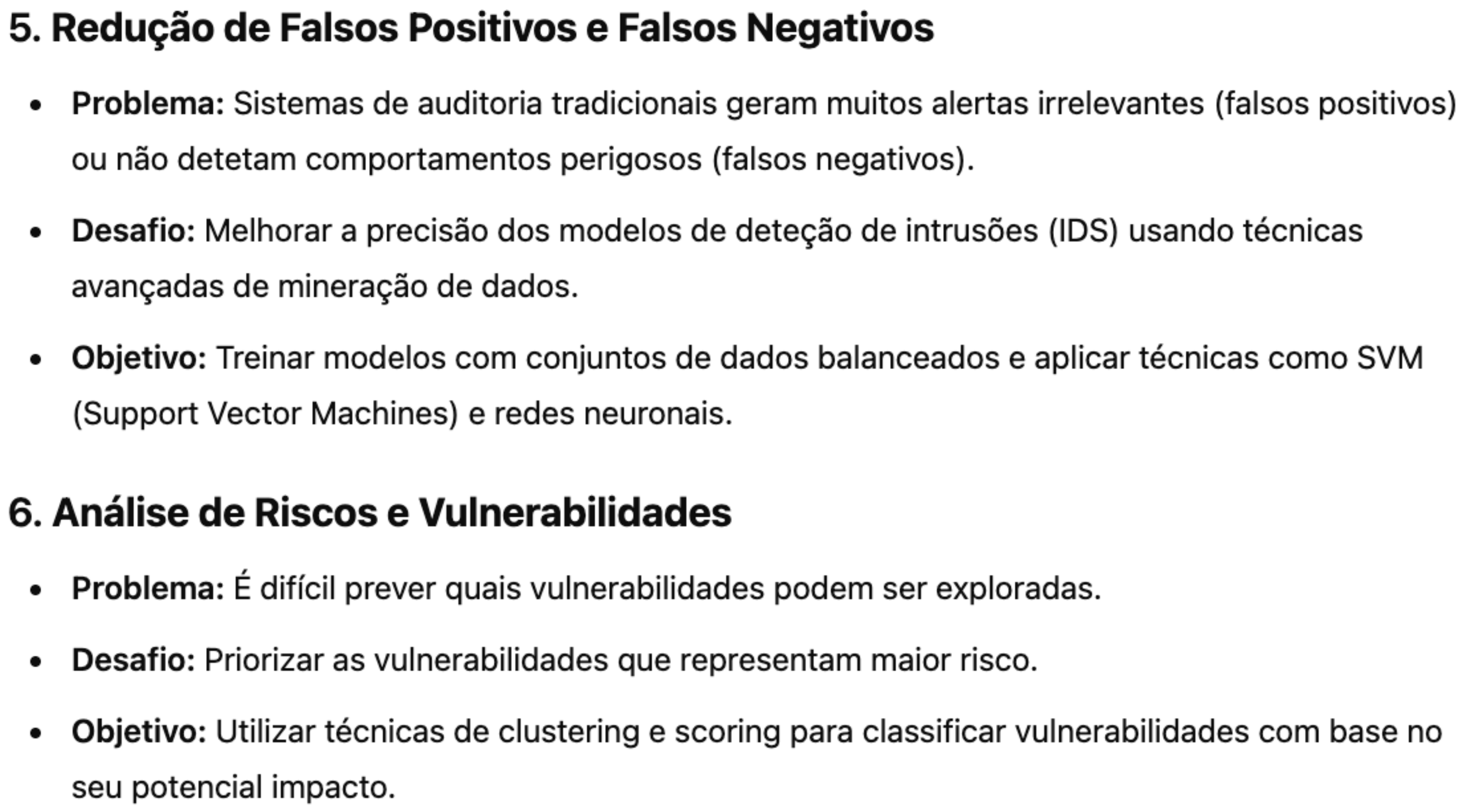
3. Correlação de Eventos de Segurança

- **Problema:** A correlação manual de eventos de segurança em grandes volumes de logs é demorada e ineficaz.
- **Desafio:** Agregar e analisar eventos provenientes de diferentes fontes (firewalls, IDS, sistemas operativos) para identificar possíveis ataques coordenados.
- **Objetivo:** Utilizar técnicas de associação e descoberta de padrões sequenciais para correlacionar eventos.

4. Prevenção de Fraudes

- **Problema:** As fraudes internas ou externas são muitas vezes descobertas tarde demais.
- **Desafio:** Identificar padrões comportamentais que indiquem tentativas de fraude.
- **Objetivo:** Implementar algoritmos de classificação supervisionada (como Decision Trees e Random Forest) para detetar transações suspeitas.

Datamining for Security Auditing



5. Redução de Falsos Positivos e Falsos Negativos

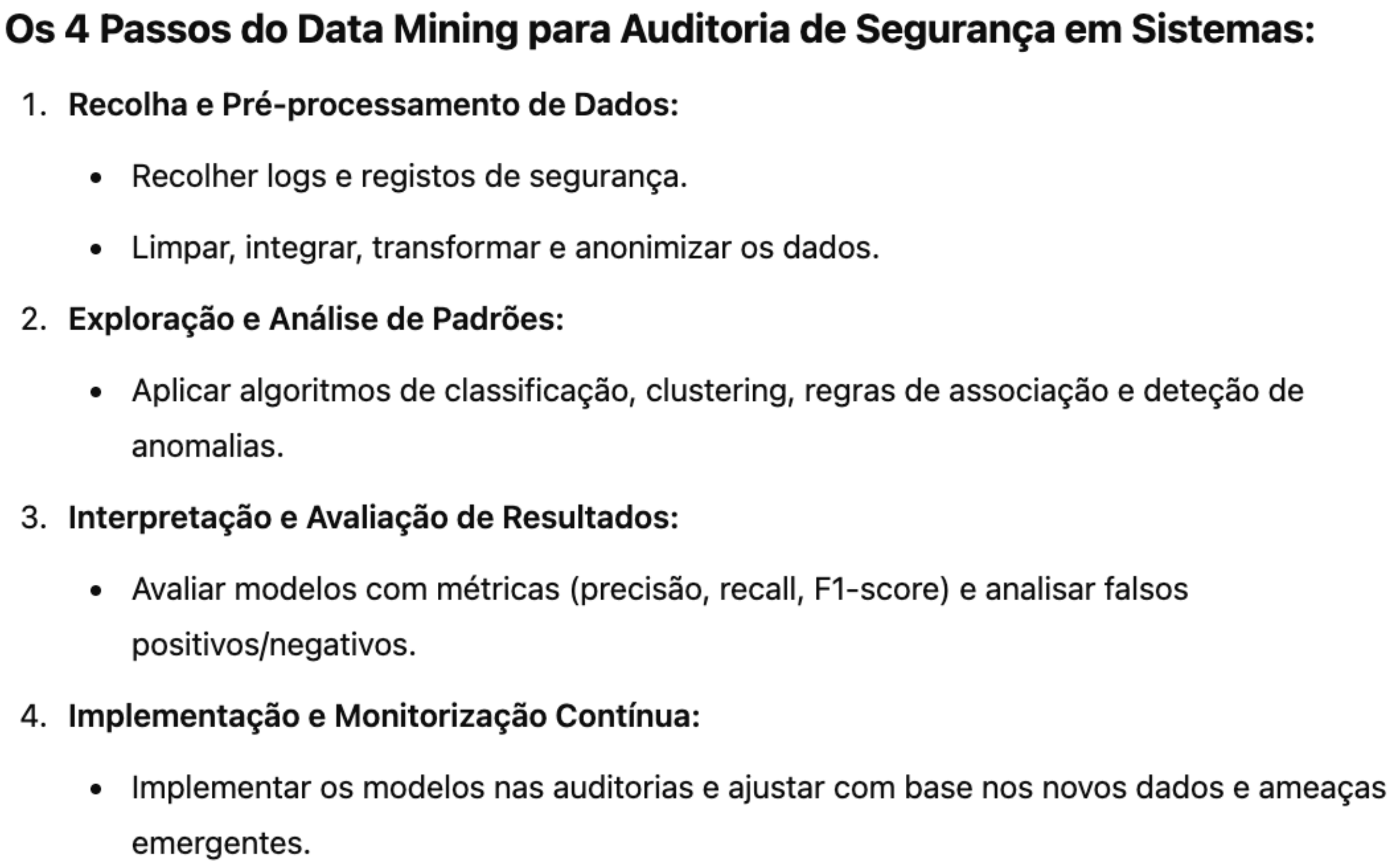
- **Problema:** Sistemas de auditoria tradicionais geram muitos alertas irrelevantes (falsos positivos) ou não detetam comportamentos perigosos (falsos negativos).
- **Desafio:** Melhorar a precisão dos modelos de deteção de intrusões (IDS) usando técnicas avançadas de mineração de dados.
- **Objetivo:** Treinar modelos com conjuntos de dados balanceados e aplicar técnicas como SVM (Support Vector Machines) e redes neuronais.

6. Análise de Riscos e Vulnerabilidades

- **Problema:** É difícil prever quais vulnerabilidades podem ser exploradas.
- **Desafio:** Priorizar as vulnerabilidades que representam maior risco.
- **Objetivo:** Utilizar técnicas de clustering e scoring para classificar vulnerabilidades com base no seu potencial impacto.

Problem:

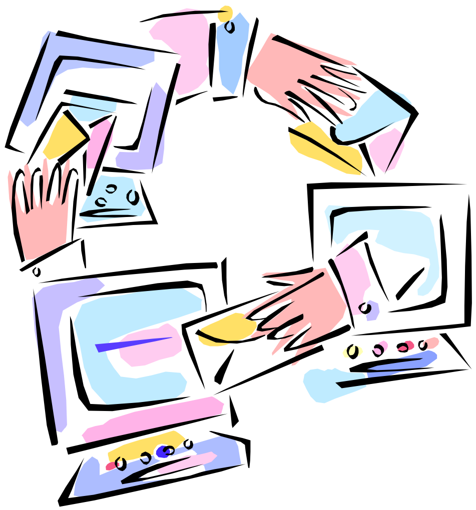
- How can we detect intrusions when we don't know what they are? When we don't have "signatures" ? (3rd step in the NIST framework)
  - □ Case 1: we know of past cases where intrusions have been detected, but there are slight variations...
  - □ Case 2: we know of many "normal" cases, which vary greatly from one to another, but we don't know what might happen differently
    - Normal/abnormal files or links
    - Normal/abnormal traffic patterns



Os 4 Passos do Data Mining para Auditoria de Segurança em Sistemas:

1. **Recolha e Pré-processamento de Dados:**
   - Recolher logs e registos de segurança.
   - Limpar, integrar, transformar e anonimizar os dados.
2. **Exploração e Análise de Padrões:**
   - Aplicar algoritmos de classificação, clustering, regras de associação e deteção de anomalias.
3. **Interpretação e Avaliação de Resultados:**
   - Avaliar modelos com métricas (precisão, recall, F1-score) e analisar falsos positivos/negativos.
4. **Implementação e Monitorização Contínua:**
   - Implementar os modelos nas auditorias e ajustar com base nos novos dados e ameaças emergentes.

General idea



Network traffic,
Files, logs, etc.

Data extraction
ETL

X_1 =[a_1 , b_1 , c_1 , d_1 , e_1]
X_2 =[a_2 , b_2 , c_2 , d_2 , e_2]
X_3 =[a_3 , b_3 , c_3 , d_3 , e_3]
X_4 =[a_4 , b_4 , c_4 , d_4 , e_4]
X_5 =[a_5 , b_5 , c_5 , d_5 , e_5]

Datamining
Predictive
(classification)

X_3 is similar
with Y.
then you must
be a virus,
or an attack

Datamining
Exploratory
(clustering)

X_4 is very
different
of others.
then you can
be a virus,
or an attack

Programme (outline)

- Introduction to techniques for detecting and classifying cyber-threats using datamining (initial part)
- Introduction to **datamining** and data **pre-processing**
- **Multi-dimensional** data **visualisation** techniques
- Techniques **for detecting outliers** and **abnormal behaviour**
- Behaviour **classification** techniques
- Techniques for detecting and classifying cyber threats (final part)

Evaluation method

- "Written repetition"
  - 45% of the grade
- Oral presentation and summary of an article
  - 30% of the grade
- DM project for security audit
  - 25% of the grade

Assessment Method - Dates

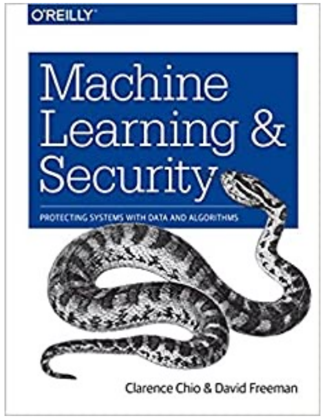
EVENT	DATE	DAY WEEK
Article submission	21/03/2025	Friday
Presentation Article	24/03/2025	Monday
Proposal Project	02/05/2025	Friday
Project submission	23/05/2025	Friday
Defence Project	26/05/2025	Monday
Written repetition	02/06/2025	Monday

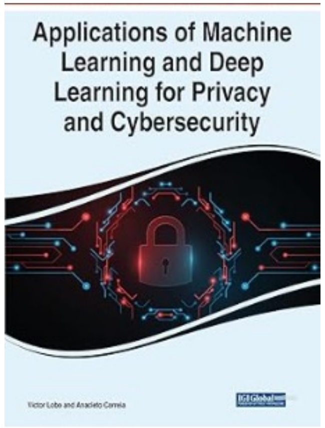
Varied Information

- Distance learning classes (Tuesdays from 18:00 to 20:00)
  - Zoom
- Questions:
  - agoncalves@tecnico.ulisboa.pt
- Support
  - After school
  - By email
- Support material (GitHub)
  - Link

Bibliography

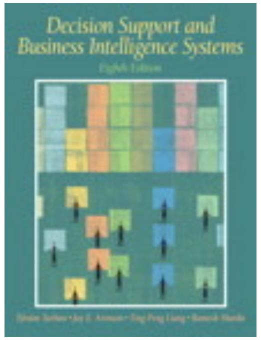
- Textbooks (**not** followed "to the letter")
  - □ Supporting texts available on the UC website
  - □ **Machine Learning and Security: Protecting Systems with Data and Algorithms,** Clarence Chio, David Freeman, O'Reilly Media, 2018
    chap.1,2,3,5
  - □ **Hands-On Machine Learning for Cybersecurity;** Soma Halder, Sinan Ozdemir, Packt Publishing, 2018
  - □ **Applications of Machine Learning and Deep Learning for Privacy and Cybersecurity**, Victor Lobo, Cortez e Correia, IGI Global, 2022


O'REILLY
Machine Learning & Security
PROTECTING SYSTEMS WITH DATA AND ALGORITHMS
Clarence Chio & David Freeman

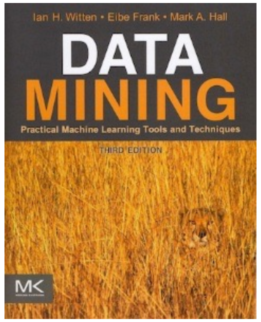

Applications of Machine Learning and Deep Learning for Privacy and Cybersecurity
Victor Lobo and Anacleto Correia
IGI Global

Bibliography

- **Decision Support and Business Intelligence Systems,** Turban, E., J. E. Aronson, et al., Prentice Hall, 2010



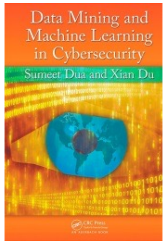
- **Data mining: practical machine learning tools and techniques;** Ian H. Witten, Eibe Frank, Mark A. Hall: Morgan Kaufmann, 2011 (WEKA)


Ian H. Witten • Eibe Frank • Mark A. Hall
DATA
MINING
Practical Machine Learning Tools and Techniques
THIRD EDITION
MK

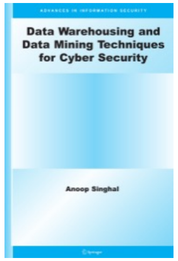
- **Python Machine Learning:** Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2, Raschka, Packt Pub., 2019

More specialised bibliography

- **Data Mining and Machine Learning in Cybersecurity**, Sumeet Dua, Xian Du, ISBN: 978-1439839423, Auerbach Publications, 2011



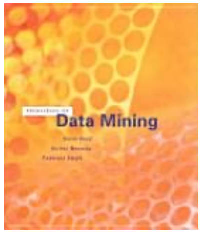
- **Data Mining Tools for Malware Detection**, Mehedy Masud, Latifur Khan, Bhavani Thuraisingham , ISBN: 978-1439854549, Auerbach Publications 2011.
- **Data Warehousing and Data Mining Techniques for Cyber Security**, Anoop Singhal, ISBN: 978-0387264097, Springer 2006.

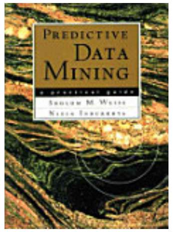


- **Applications of Data Mining in Computer Security**, Barbará, Daniel; Jajodia, Sushil (Eds.), ISBN: 978-1-4020-7054-9, Springer 2002.

General DM bibliography

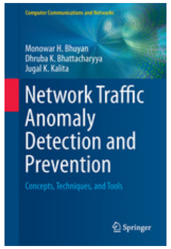
- **Machine Learning**, Tom M. Mitchell, McGraw Hill, 1997
- **Pattern Classification**, Duda, Hart, & Stork, Wiley, 2001
- **Principles of data mining**, David. J. Hand, Heikki Mannila, Padhric Smyth, MIT Press, 2001
- **Predictive data mining**, Sholom M. Weiss, Nitin Indurkhya, Morgan Kaufmann, 1997
- **C4.5:Programs for Machine Learning**, John Ross Quinlan, Morgan Kaufmann, 1992





Bibliography

- Network Traffic Anomaly Detection and Prevention - Concepts, Techniques, and Tools, Bhuyan, Monowar H., Bhattacharyya, Dhruba K., Kalita, Jugal K., ISBN: 978-3-319-65188-0, Springer 2017



Computer Communications and Networks

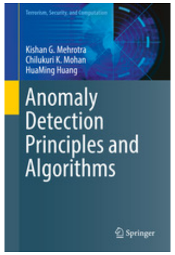
Monowar H. Bhuyan
Dhruba K. Bhattacharyya
Jugal K. Kalita

Network Traffic Anomaly Detection and Prevention

Concepts, Techniques, and Tools

Springer

- Anomaly Detection Principles and Algorithms, Mehrotra, Kishan G., Mohan, Chilukuri, Huang, Huaming, 978-3-319-67526-8, Springer 2017



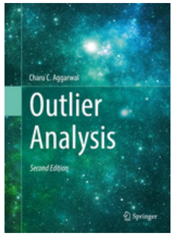
Terrorism, Security, and Computation

Kishan G. Mehrotra
Chilukuri K. Mohan
HuaMing Huang

Anomaly Detection Principles and Algorithms

Springer

- Outlier Analysis, Aggarwal, Charu C., 978-3-319-47578-3, Springer 2017



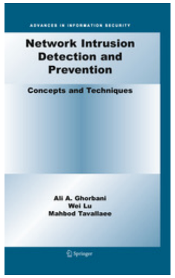
Charu C. Aggarwal

Outlier Analysis

Second Edition

Springer

- Network Intrusion Detection and Prevention - Concepts and Techniques, Ghorbani, Ali A., Lu, Wei, Tavallaee, Mahbod, 978-0-387-88771-5, Springer 2010



ADVANCES IN INFORMATION SECURITY

Network Intrusion Detection and Prevention

Concepts and Techniques

Ali A. Ghorbani
Wei Lu
Mahbod Tavallaee

Springer

Other interesting sites...

- Decisionarium
  - GNU software, references, etc.
  - http://www.decisionarium.tkk.fi
- DSS Resources
  - Prof Daniel Power, books, references, etc.
  - http://dssresources.com/
- Machine Learning Network
  - www.mlnet.org
  - Software, data, conferences, projects, etc.
- Manufacturers of "dedicated" solutions
  - For land management, marketing, etc., etc.

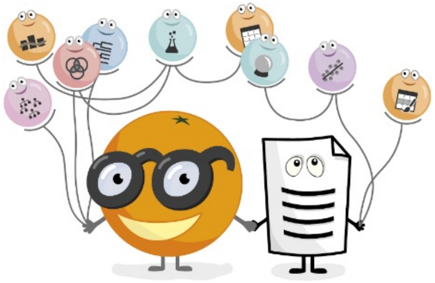
Data repositories

- **Irvine Repository** (UCI)
  - https://archive.ics.uci.edu/ml/index.php
  - Data, software, articles
  - A classic! A must!
- **Kaggle Repository**
  - www.kaggle.com/datasets
  - Very current, very active

- **IEEE Repository**
  - IEEE Data Port
  - https://ieee-dataport.org/datasets
- Repository for Cybersecurity
  - ICSX: http://www.iscx.ca/datasets/ (but KDD99 is available at UCI)

Solving practical problems

- MS-Excel
  - Everyone knows!
  - Solves most simple problems
- WEKA
  - Java, free, https://www.cs.waikato.ac.nz/ml/weka/
  - Many well-documented algorithms
- Orange
  - Python, free, https://orange.biolab.si
  - Graphical interface



- Others
  - MATLAB, R, Skikit-learn, Keras.SPSS and Clementine, SAS Enterprise Miner, IBM Intelligent Miner, SAP BI...,

Solving practical problems

- **Google colab: a** cloud-based platform that allows Python code to be executed directly from a browser. It is especially useful for
- **1. Pandas:** Data manipulation and analysis (DataFrames, EDA).
- **2. NumPy:** Numerical calculations and manipulation of arrays.
- **3. Matplotlib:** Basic visualisations (graphs, histograms).
- **4. Seaborn:** Advanced visualisations (heat matrices, boxplots).
- **5. Scikit-learn:** Machine Learning algorithms (classification, regression, clustering).
- **6. XGBoost/LightGBM:** Advanced models (boosting, efficient classification).
- **7. Statsmodels:** Statistical analyses (tests, regression).
- **8. TensorFlow/Keras:** Neural Networks and Deep Learning.

Articles to be presented (examples... but **look them up**!)

- Bollmann, C. A., Tummala, M., & McEachen, J. C. (2021). Resilient real-time network anomaly detection using novel non-parametric statistical tests. *Computers & Security, 102, 102146. doi:https://doi.org/10.1016/j.cose.2020.10214*

- Gibert, D., Mateu, C., Planes, J., & Marques-Silva, J. (2021). Auditing static machine learning anti-Malware tools against metamorphic attacks. *Computers & Security, 102, 102159. doi:https://doi.org/10.1016/j.cose.2020.102159*

- Krumay, B., Bernroider, E. W. N., & Walser, R. (2018). *Evaluation of Cybersecurity Management Controls and Metrics of Critical Infrastructures: A Literature Review Considering the NIST Cybersecurity Framework, Cham.*

- Lin, W.-C., Ke, S.-W., & Tsai, C.-F. (2015). CANN: An intrusion detection system based on combining cluster centres and nearest neighbors. *Knowledge-Based Systems, 78, 13-21. doi:https://doi.org/10.1016/j.knosys.2015.01.009*

Articles to be presented (examples... but **look them up**!)

- Mitchell, R., & Chen, I.-R. (2014). A survey of intrusion detection techniques for cyber-physical systems. *ACM Comput. Surv., 46(4), Article 55. doi:10.1145/2542049*

- Casas, P., Mazel, J., & Owezarski, P. (2012). Unsupervised Network Intrusion Detection Systems: Detecting the Unknown without Knowledge. *Computer Communications, 35(7), 772-783. doi:https://doi.org/10.1016/j.comcom.2012.01.016*

- García-Teodoro, P., Díaz-Verdejo, J., Maciá-Fernández, G., & Vázquez, E. (2009). Anomaly-based network intrusion detection: Techniques, systems and challenges. Computers & Security, 28(1), 18-18-28. doi:10.1016/j.cose.2008.08.003

Articles to present (examples...)

- Data Mining for Cyber Security, V.Chandois *et al.*, in Data Warehousing and Data Mining Techniques for Computer Security, Springer, 2006.
- Data mining methods for anomaly detection KDD-2005 workshop report, Margineantu *et al.,* ACM SIGKDD Explorations Newsletter, Volume 7 Issue 2, December 2005.
- On the efficacy of data mining for security applications, Ted E. Senator, ACM SIGKDD Workshop on CyberSecurity and Intelligence Informatics -CSI-KDD '09, 2009.
- Metrics for mitigating cybersecurity threats to networks, IEEE Internet Computing, 14, 1, Jan-Feb 2010.
- A Combined Fusion and Data Mining Framework for the Detection of Botnets, Kiayias *et al.*, Conference For Homeland Security, 2009. CATCH '09. Cybersecurity Applications & Technology, March 2009
- A study of Spam Detection Algorithms on Social Media Networks, Jacob Soman Saini, International Conference on Computational Intelligence, Cyber Security, and Computational Models, Coimbatore, India, December 2013.

Articles to present (...examples...)

- Comparative Study of Two- and Multi-Class-Classification-Based Detection of Malicious Executables Using Soft Computing Techniques on Exhaustive Feature Set. Shina Sheen, R. Karthik and R. Anitha; International Conference on Computational Intelligence, Cyber Security, and Computational Models, Coimbatore, India, December 2013
- Botnets: A Study and Analysis, G. Kirubavathi and R. Anitha, International Conference on Computational Intelligence, Cyber Security, and Computational Models, Coimbatore, India, December 2013
- The VoIP intrusion detection through a LVQ-based neural network, Zheng Lu ; Taoxin Peng, International Conference for Internet Technology and Secured Transactions, 2009. ICITST 2009.
- Detection of applications within encrypted tunnels using packet size distributions, Mujtaba,G.,Parish, D.J., International Conference for Internet Technology and Secured Transactions, 2009. ICITST 2009.
- Email classification: Solution with back propagation technique, Ayodele et al. International Conference for Internet Technology and Secured Transactions, 2009. ICITST 2009.
- Malware detection using statistical analysis of byte-level file content, Tabish et al., CSI-KDD '09 Proceedings of the ACM SIGKDD Workshop on CyberSecurity and Intelligence Informatics, 2009