

Datamining for Security Auditing

António Gonçalves

based on the slides **Prof. Victor Lobo**

Master's in Information Security and Law in Cyberspace



Datamining for Security Auditing

António Gonçalves

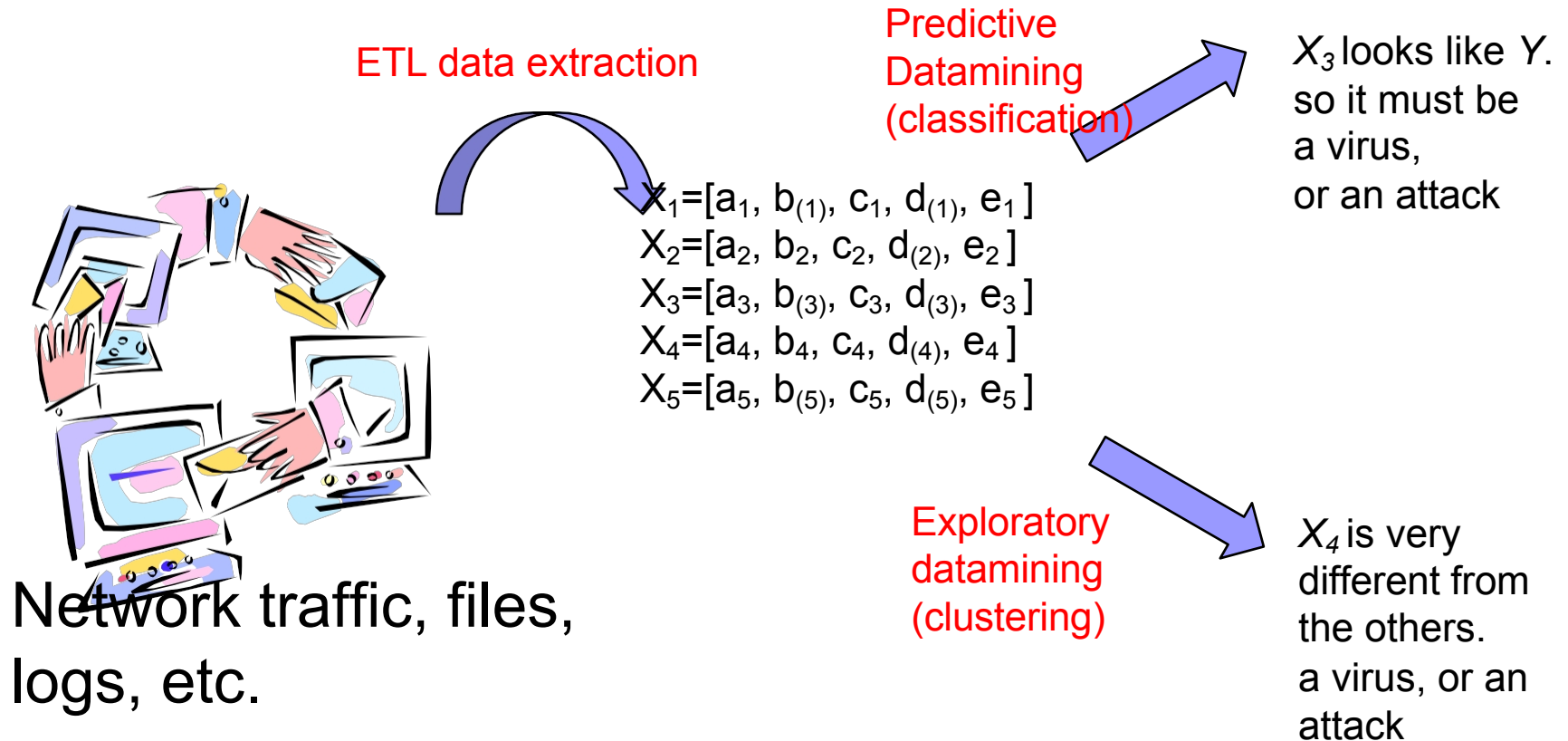
based on the slides **Prof. Victor Lobo**

Master's in Information Security and Law in Cyberspace



Types of problems

General idea





Basic idea:

**COLLECT MUCH DATA
AS POSSIBLE !**

Collecting data for what?



Janus example



- Looking to the past
- "**Study** the past to **understand** the present and **predict** the future"



Basic ideas

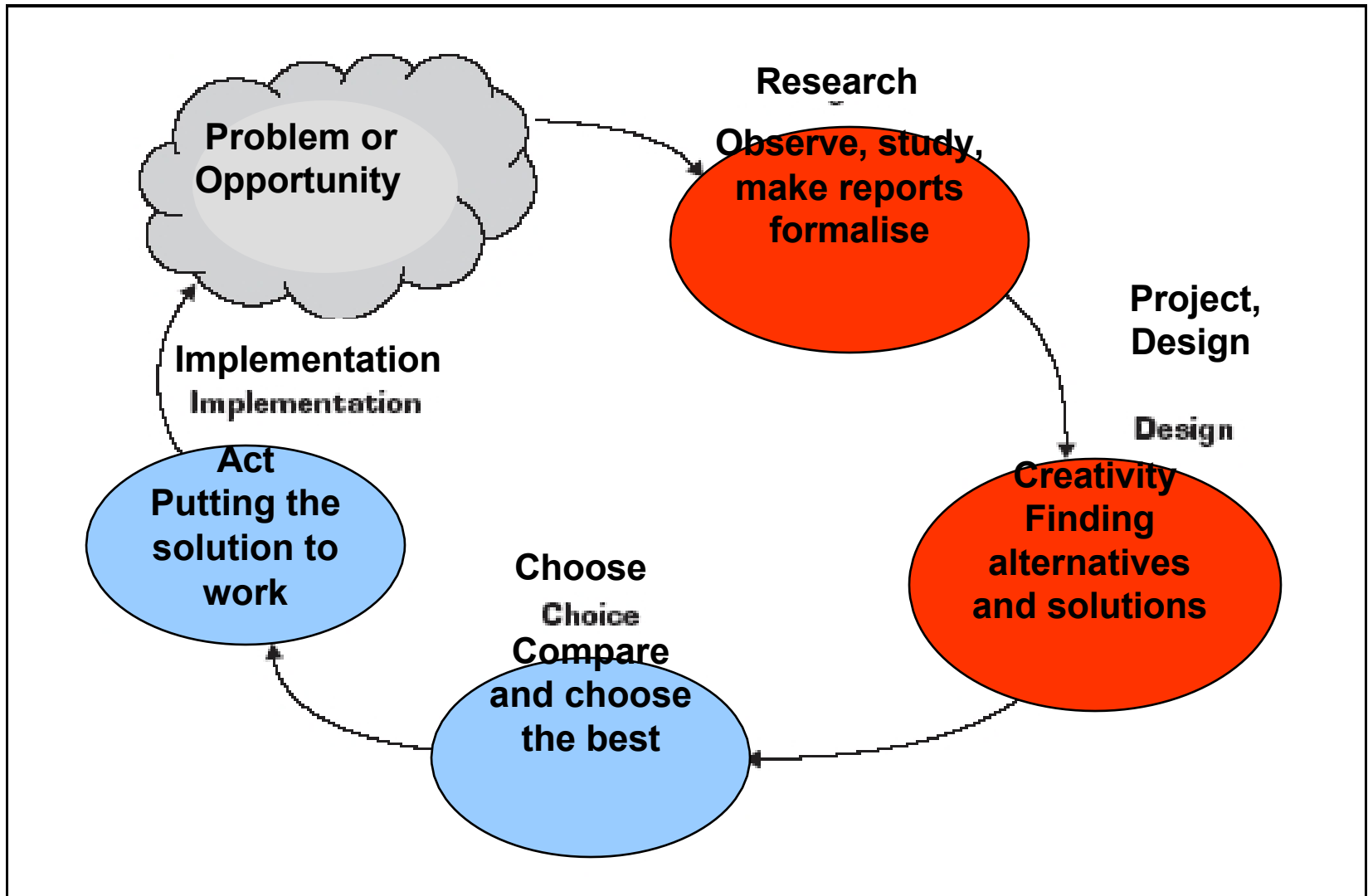
Learning from the past

Inferring from experience

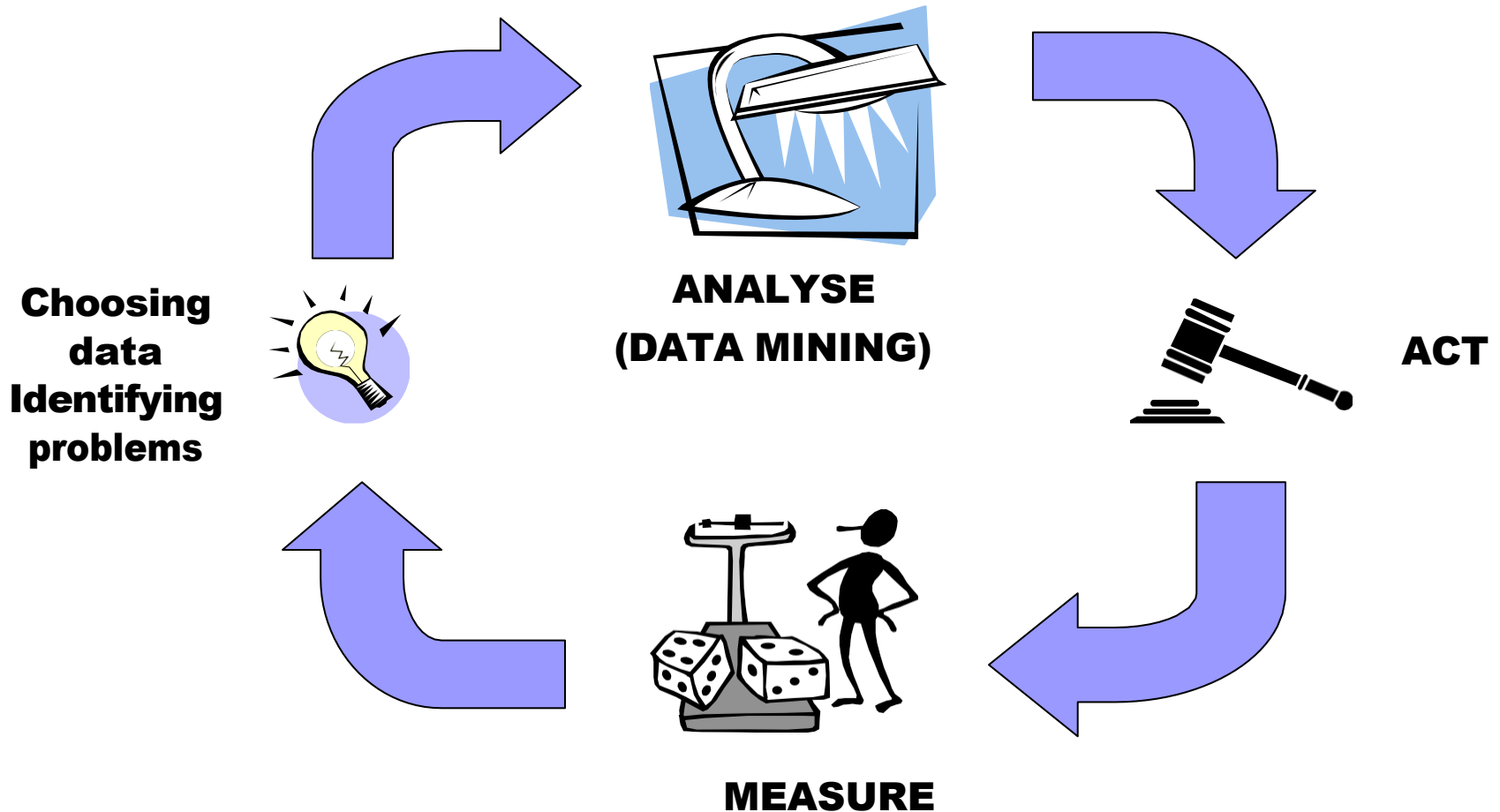
Tools: **datamining** techniques

by any other name...

Cycles that never end...



Cycles that never end...





Simply put, Datamining is

Data mining is the process of extracting useful knowledge large volumes of data, using three main techniques:

1.Databases - Efficient organisation, storage and retrieval of data for analysis.

2.Statistics - Mathematical methods for identifying patterns, trends and correlations.

3.Machine Learning - Algorithms that learn from data to make predictions and classifications.

The combination of these approaches makes it possible to discover valuable information for decision support in various areas.



Simply put, Datamining is

- **Prediction**

- Using historical data to predict future values or events
- Applied in areas such as sales forecasting, fraud detection and medical diagnosis
- Common methods include regression, neural networks and time series models

- **Discovery of New Knowledge**

- Identify patterns, relationships and hidden information in data
- Find associations, groupings and anomalies
- Applied in marketing, biomedicine, cybersecurity, among others
- Techniques such as clustering, association rules and analysing outliers are often used



Data representation

Structured Data

Structured data is organised in a rigid and predictable way, usually stored in relational database tables. Each piece of data is associated with a specific attribute or field, making retrieval and analysis efficient.

Characteristics of Structured Data

- Stored in **tables** (rows and columns).
- They are **easy to search** using SQL and other database tools.
- Fixed and well-defined structure (each column has a specific data type, such as numbers, dates or short text).
- **High scalability and processing efficiency.**

Examples of Structured Data

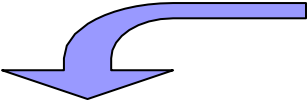
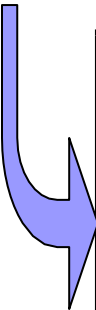
- **Relational databases** (MySQL, PostgreSQL, Oracle, SQL Server).
- **Excel tables** organised by category.
- **Financial records** (bank transactions, salaries, taxes).
- **Customer management systems (CRM)** (names, emails, telephone numbers).
- **Organised sensor data** (time, temperature, pressure).

Data representation

Structured Data

Data, vector, record or pattern

Variable, characteristic or attribute



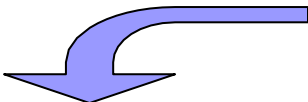
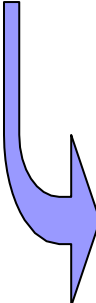
Height	Weight	Sex	Age	Salary	Use the gym	Charges to the insurance company
1.60	79	M	41	3000	S	N
1.72	82	M	32	4000	S	N
1.66	65	F	28	2500	N	N
1.82	87	M	35	2000	N	S
1.71	66	F	42	3500	N	S

Data representation

Structured Data

Data, vector, record or pattern

Variable, characteristic or attribute



Height	Weight	Sex	Age	Salary	Use the gym	Charges to the insurance company
1.60	79	M	41.5	3000	S	N
1.72	300	M	32	4000	P	N
1.66	65	F	28		N	N
-1.82	87	M	35	2000	N	S
1.71	66	F	42	3500	N	S

Data representation

Data (Structured)

Interrelationship

CHAIR	
CodCad	Name
<u>12347</u>	Databases
<u>34248</u>	Algebra
<u>32439</u>	Introduction to Computers

STUDENT		
NumMec	Name	Course
<u>798764544</u>	João Pinto	LCC
<u>345673451</u>	Carlos Semedo	MIERSI
<u>487563546</u>	Maria Silva	LBIO
<u>452212348</u>	Pedro Costa	LMAT

REGISTRATION	
NumMec	CodCad
798764544	12347
345673451	12347
798764544	34248
452212348	32439

Data representation

Data Structured

Non-consistent data

CHAIR	
CodCad	Name
<u>12347</u>	Databases
<u>34244</u>	Algebra
<u>32439</u>	Introduction to Computers

STUDENT		
NumMec	Name	Course
<u>798764544</u>	João Pinto	NULL
<u>345673451</u>	Carlos Samedo	12345
<u>798764544</u>	NULL	LBIO
<u>452212348</u>	Pedro Costa	LMAT

REGISTRATION	
NumMec	CodCad
798764544	12347
345673451	12347
798764545	34248
452212346	32439



Data representation

Unstructured data

Unstructured data doesn't follow a fixed format and can contain free text, multimedia or other information that doesn't easily fit into a table.

Characteristics of Unstructured Data

- Without rigid organisation (they cannot be easily stored in relational tables).
- Greater volume and complexity, requiring advanced processing techniques.
- Difficulty in searching and indexing, requiring technologies such as Machine Learning, Natural Language Processing (NLP) and Image Recognition.
- Stored in different formats, such as documents, emails, videos, audio, social networks.

Examples of Unstructured Data

- Emails (the text of the message does not follow a fixed structure).
- Images and videos (cannot be represented in columns and rows).
- Social media posts (comments, emojis, reactions).
- Call and message logs (text and voice content).
- Text files and documents (PDFs, Word, articles).
- IoT sensors that collect images and sounds.

Data representation

Unstructured data


A comment on a social networking site can be considered unstructured data:

"I went to this restaurant yesterday and loved the atmosphere, but the food took a long time arrive. The service was good, but I think they could improve the speed of service.

★★★★"

There is a wealth of information here:

- Positive feelings ("I loved the atmosphere") and negative feelings ("It took too long").
- Rating (4 stars).
- No fixed structure (information appears freely).



Data representation

Semi-structured data

There is also a third category, known as semi-structured data, which contains some organisation, but without a completely fixed structure. This data combines characteristics of the two previous types.

Examples of semi-structured data

- JSON and XML (data organised in key-value pairs, but without a rigid structure like in SQL).
- Emails (the sender and recipient have a fixed structure, but the content of the email is free text).
- System logs (follow a pattern, but contain unstructured text).

Data representation Semi-structured data

Exemplo de Dados Semiestruturados (JSON)

json

Copy Edit

```
{
  "cliente": "João Silva",
  "idade": 30,
  "compras": [
    {
      "produto": "Smartphone",
      "preço": 599.99
    },
    {
      "produto": "Capa de proteção",
      "preço": 19.99
    }
  ]
}
```



Data representation

Conclusion

- **Structured data** is easy to store, search and analyse, but limits the complexity of the information.
- **Unstructured data** contains more detail and is richer in information, but requires advanced techniques to analyse.
- **Semi-structured data** is a middle ground between the two, with some organisation but still flexibility.



Types of problems

Prediction

- **Classification**

- Assigns categories to new data based on previous examples
- Used in spam detection, medical diagnosis, image recognition
- Common algorithms: Decision trees, Random Forest, SVM, Neural Networks

- **Regression**

- Predicts numerical values based on independent variables
- Applied in sales forecasting, price analysis, meteorology
- Common algorithms: Linear Regression, Logistic Regression, Neural Networks



Types of problems

Knowledge Discovery in Data Mining

- **Detection of deviations**

- Identification of anomalies or outliers in the data
- Applied to fraud detection, network monitoring and quality control

- **Database segmentation**

- Separation of data into homogeneous groups for more effective analysis
- Used in marketing campaigns, personalisation of services and customer analysis

- **Clustering**

- Automatic data grouping without predefined classes
- Applied to customer segmentation, computational biology and image analysis



Types of problems

Knowledge Discovery in Data Mining

- **Association rules**

- Discovering frequent patterns and relationships between variables
- Example: Market Basket Analysis

- **Summarisation**

- Extracting essential information from large volumes of data
- Used in automatic reports, recommendation systems and natural language processing

- **Visualisation**

- Graphical representation of data to facilitate interpretation
- Common tools: interactive graphics, dashboards and heat maps

- **Text Search**

- Analysing and extracting useful information from large collections of texts
- Applied in search engines, sentiment analysis and document categorisation



Fraud detection when using a credit card

Financial institutions face the challenge of **identifying and blocking fraudulent transactions** without inconveniencing legitimate customers.

Fraud can occur in a variety of ways, including misuse of stolen cards, card cloning and unauthorised online transactions.

As most fraud happens quickly, it is essential to have an automatic system that can identify and block suspicious transactions in real time.



Fraud detection when using a credit card

Suitable Model for the Solution

- This problem involves labelling each transaction as fraudulent or legitimate.
- The model must learn from previous transactions, identifying patterns that distinguish normal operations from suspicious activity.
- Thus, for each new transaction, the system will be able to predict whether it should be approved or blocked based on the characteristics observed.

Fraud detection when using a credit card

Transaction ID	Amount (€)	Location	Card	Type	Risk	Fraud
123456789	150.75	Portugal	Credit	Online	0.2	No
123456790	2500.0	USA	Debit	Physical	0.9	Yes
123456791	5.99	Brazil	Credit	Online	0.05	No
123456792	120.0	UK	Credit	Online	0.3	No
123456793	5000.0	Germany	Debit	Physical	0.95	Yes
123456794	45.5	France	Credit	Online	0.15	No
123456795	800.0	Spain	Debit	Physical	0.5	No
123456796	60.99	Italy	Credit	Online	0.25	No
123456797	10000.0	Canada	Debit	Physical	0.98	Yes
123456798	22.3	Australia	Credit	Online	0.1	No



Fraud detection when using a credit card

Model Challenges

- **Class imbalance** - Most transactions are legitimate, making it difficult for the model to correctly identify fraud.
- **Evolution of Fraud Patterns** - attackers adapt and change strategies, making some models obsolete quickly.
- **Minimising False Positives** - Blocking legitimate transactions can cause customer dissatisfaction and impact the user experience.
- **Response time** - Decisions must be made in real time to avoid delays in processing purchases.
- **Security and Privacy** - The model must guarantee that customer data is protected and that there are no breaches of privacy.



Fraud detection when using a credit card

Results and benefits

- **Reducing Financial Losses** - Early detection fraud minimises losses.
- **Greater Security for Customers** - Protection against unauthorised access to cards.
- **Fewer Unwarranted Blocks** - The model learns to distinguish legitimate purchases from suspicious ones.
- **Automatic Monitoring** - The system analyses all transactions without the need for manual intervention.





Predicting the Time of a Cyber Attack

Companies and organisations are facing a constant increase in cyber-attacks, such as intrusion attempts, denial-of-service (DDoS) attacks and exploitation of vulnerabilities.

One of the challenges in computer security is **predicting when the next attack will occur**, allowing security teams to take preventative measures.

If it is possible to accurately estimate the time until the next attack attempt, defences can be adjusted proactively, reducing the impact of potential breaches.




Predicting the Time of a Cyber Attack

This problem requires the **prediction of a continuous value**: the time until the next cyber attack occurs. Thus, it is a case of **Regression**, where the model analyses historical attack patterns and contextual factors to estimate the time interval between malicious events.

The model can be trained on historical data, taking into account variables such as:

- Frequency of previous attacks
- Type of attack detected
- Anomalous traffic volume on the network
- Number of active vulnerabilities in the system
- Seasonal trends in attacks at certain times of the year



Predicting the Time of a Cyber Attack Model Challenges

- Irregular patterns - Cyber attacks may not follow fixed patterns, making it difficult to predict the exact time until the next event.
- Threat evolution - New attack techniques can emerge, making historical data less representative of attackers' future behaviour.
- Influence of External Factors - Global events, such as the disclosure of new vulnerabilities or political tensions, can affect the frequency of attacks, which can be difficult to model.
- Data quality - Security logs can contain gaps or false positives, affecting the accuracy of predictions.
- Rapid Response Time - The model needs to provide timely predictions so that security teams can act before an attack occurs.





Detection of Irregularities in Access to Corporate Systems

Companies and institutions store sensitive information in internal systems that must be protected against unauthorised access. However, internal and external attacks can compromise this data. Manual detection of suspicious access is unfeasible due to the large volumes of login records and user activity. So you need a system that identifies irregularities in access patterns, helping security teams to respond quickly to possible threats.



Detection of Irregularities in Access to Corporate Systems

Suitable Model for the Solution

This problem involves identifying abnormal behaviour without necessarily having prior examples of specific attacks. To do this, the model needs to analyse normal user access patterns and **flag activities that deviate from these patterns**. As the aim is not to classify previously known events, but to discover new threats, this **approach is based on Knowledge Discovery**, more specifically on **anomaly detection**.

The model will analyse data such as:

- User login times
- Devices and usual places of access
- Number of failed login attempts
- Accessed commands and files
- Volume of data transferred
- If a user displays behaviour that is significantly different from their normal pattern (for example, a login from an unusual country or massive access to confidential files), the system should generate an alert for investigation.



Detection of Irregularities in Access to Corporate Systems Challenges

- Definition of an Anomaly - Not all unusual activity is malicious. The model should avoid false positives, such as legitimate employee travel.
- Evolving Behaviours - Normal user access patterns can change over time, requiring an adaptable model.
- Unbalanced data - The majority of accesses will be legitimate, making it difficult to detect rare and sophisticated attacks.
- Dynamic environments - Large companies have thousands of users and devices, which can generate a huge volume of data for real-time analysis.
- Privacy and Ethics - The model must guarantee that access monitoring respects data protection and employee privacy standards....





User Grouping for Insider Threat Prevention

Companies deal with internal risks such as the leakage of confidential information, improper access to critical systems and suspicious behaviour by employees or collaborators. In order to mitigate these risks, it is essential to group the organisation's users based on their behaviour patterns, identifying higher risk groups and enabling preventive measures to be adopted.

Grouping makes it possible to categorise users according to their activity within the system, distinguishing normal patterns from possible internal threats.



User Grouping for Insider Threat Prevention - Appropriate Model

This problem falls under Knowledge Discovery, as the aim is not to predict a specific event, but to identify patterns and create distinct groups of users based on their interactions. In this case, Database Segmentation is used, a process that groups together individuals with similar behaviours, without prior labels on the data.

The model analyses variables such as:

- Frequency and time of access to systems
- Types of files handled
- Resources and databases accessed
- Use of external devices (USB, external discs, etc.)
- Unauthorised access attempts



User Grouping for Insider Threat Prevention - Appropriate Model

After segmentation, users can be grouped into different profiles, such as:

- Standard users - Employees with normal access and no suspicious behaviour.
- Privileged users - System administrators or managers with advanced permissions.
- At-risk users - Employees with anomalous behaviour, such as logging in outside normal hours or attempting to extract data on a massive scale.
- This grouping allows security teams to reinforce measures for the segments most at risk, such as implementing automatic alerts or reviewing access permissions.



User Grouping for Insider Threat Prevention - Challenges

Defining Relevant Groups - It's not always obvious how many segments should be created or which characteristics are most relevant for segmentation.

Unbalanced data - Most users may have normal behaviour, making it difficult to identify anomalous patterns without generating false positives.

Evolution of Behaviours - User activities can change over time, requiring constant updating of the model.

Interpretation of Results - Segmentation can reveal unexpected patterns, requiring manual analysis to understand their relevance.

Privacy and Compliance - Employee monitoring can raise ethical and legal issues, requiring compliance with data protection regulations.





Identifying Groups of Compromised Devices in an Enterprise Network

Companies and organisations face growing challenges in detecting compromised devices within their networks. Cybercriminals can exploit vulnerabilities in computers, servers or IoT devices to launch internal attacks, steal data or create botnets.

As many of these threats operate discreetly, it is essential to identify anomalous patterns in the behaviour of devices without relying on previous lists of known threats.



Identifying Groups of Compromised Devices in an Enterprise Network

This problem requires discovering unknown patterns within network traffic, grouping devices with similar behaviour to identify anomalies. As there are no pre-defined categories, the best approach is Clustering, which groups devices into clusters based on their traffic characteristics, without the need for prior labelling.

The data analysed may include:

- Volume of traffic sent and received
- Protocols used (HTTP, HTTPS, FTP, SSH, etc.)
- Communication standards with other devices
- Number of requests to suspicious domains
- Frequency and time of connections

The model will organise the devices into different groups with similar behaviour patterns. If a new group appears that is small and distinct from the others, it could indicate a group of compromised machines acting maliciously.



Identifying Groups of Committed Devices in an Enterprise Network - Model Challenges

- Defining the Optimal Number of Clusters - The model needs to determine how many clusters there are in the network, which may not be obvious from the outset.
- Differentiation between Normal and Malicious Behaviours - Some machines may have legitimate usage patterns that appear anomalous, such as servers that frequently communicate with multiple devices.
- Continuous Network Evolution - New devices join and leave the network regularly, requiring the model to adapt dynamically.
- Big Data - Network traffic generates large amounts of information, requiring efficient methods for real-time processing.
- Interpreting the Results - Once the clusters have been identified, an analysis process is required to determine which groups represent real threats.



Models versus data

■ Model based

- The relationships between variables are based on physical laws, mathematical principles or established theoretical knowledge.
- If we use the equation $P=mg$ to describe the gravitational force exerted on an object (cause) it results from the acceleration of gravity (effect) depending on the mass of the object

■ Data driven

- Look for relationships in the data
 - Relationships do not imply cause/effect
- Either there is no model, or there is a generic model that is usually a universal approximator (with many parameters)
- If we only have force and acceleration data without knowing the equation, a data-driven model can find a pattern that relates both variables, but without understanding **why the relationship exists**.



Example: Recognising cats in images

Model-Based Approach

We have created a set of fixed rules for identifying cats in images:

- If it has pointed ears+ big eyes+ small muzzle, then it's a cat.
- If it doesn't have these characteristics, it's not a cat.

Problem: If the image is dark or the cat is turned round, the model may fail.



Example: Recognising cats in images

Data-Driven Approach

- We trained a neural network with thousands of images of cats and dogs.
- The model learns on its own which characteristics distinguish a cat from a dog.
- When it sees a new image, it classifies it correctly without the need for manual rules.
- The model automatically learns patterns in the data, without the need for fixed rules.

Problem: You need a lot of data to train