



# Datamining para Auditoria de Segurança

António Gonçalves

baseados nos slides **Prof. Victor Lobo**

**Mestrado em Segurança da Informação e Direito no Ciberespaço**



# Datamining para Auditoria de Segurança

António Gonçalves

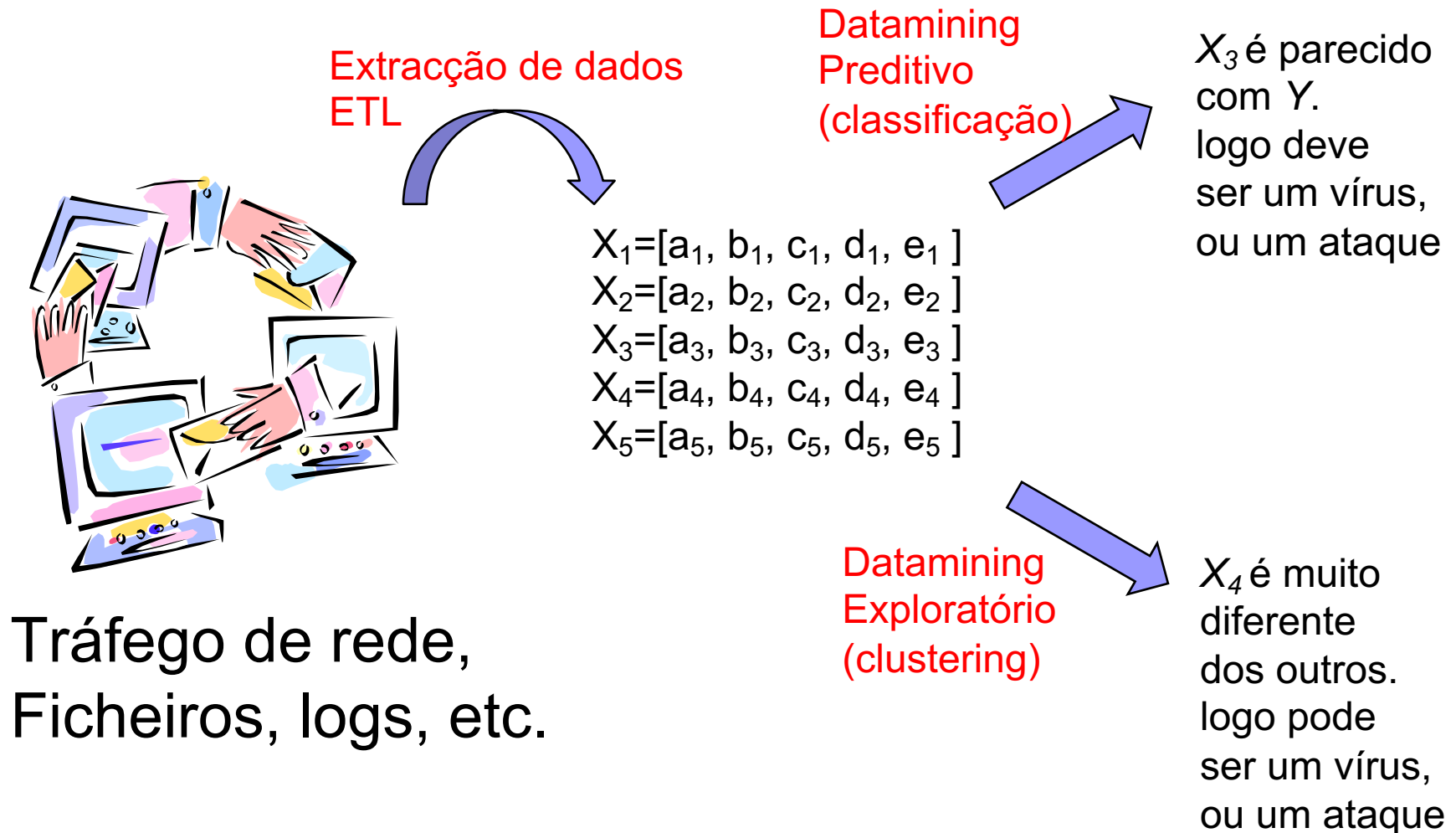
baseados nos slides **Prof. Victor Lobo**

**Mestrado em Segurança da Informação e Direito no Ciberespaço**



# Tipos de problemas

# Ideia geral





Ideia base:

**RECOLHER TODOS OS  
DADOS POSSÍVEIS !**

# Recolher dados para quê ?



# Exemplo de Janus



- Olhar o passado e o futuro
- “**Estudar** o passado para **compreender** o presente, e **prever** o futuro”



# Ideias base

**Aprender** com o passado

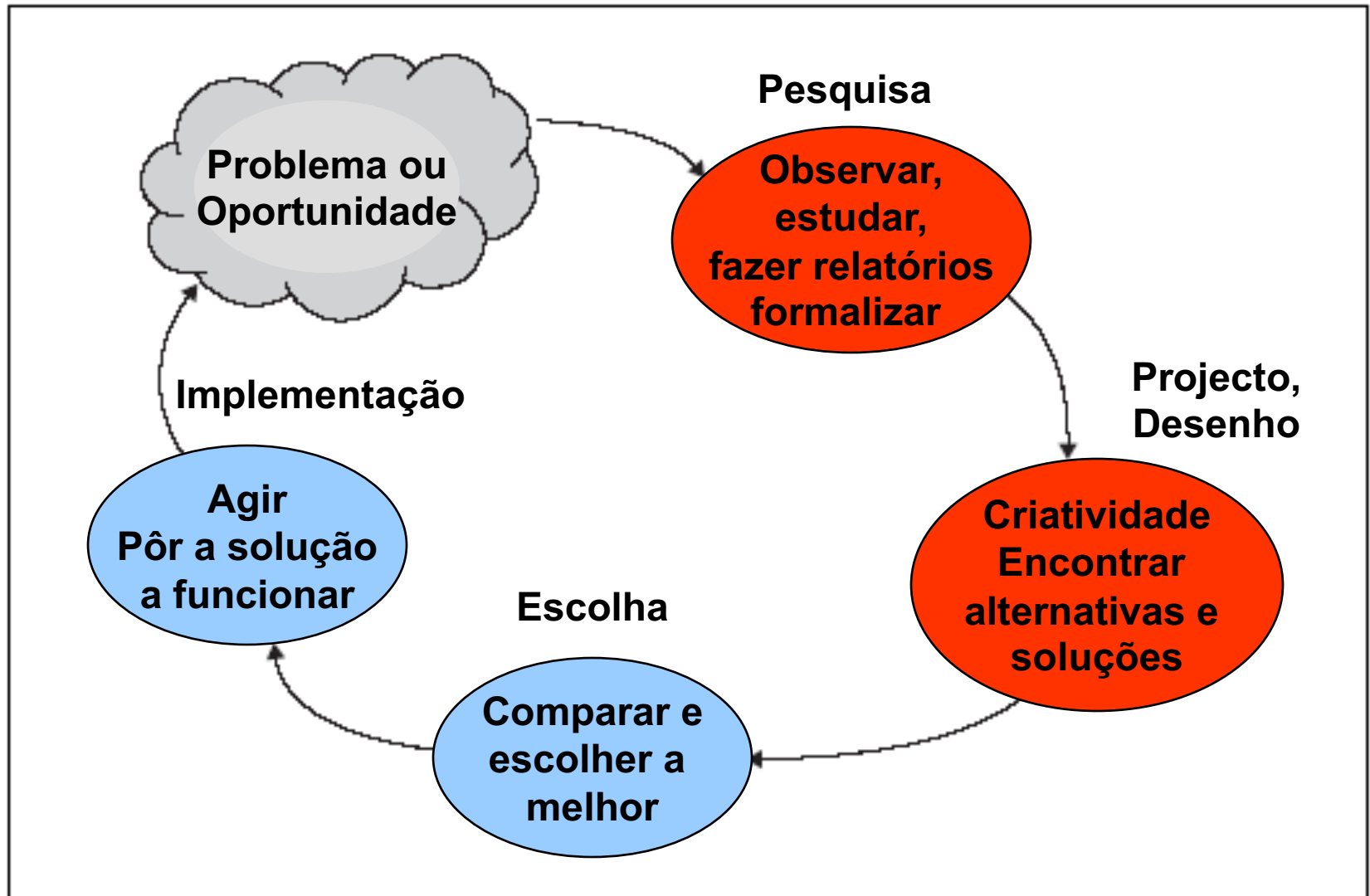
**Inferir** a partir da experiência

Ferramentas: técnicas de **datamining**

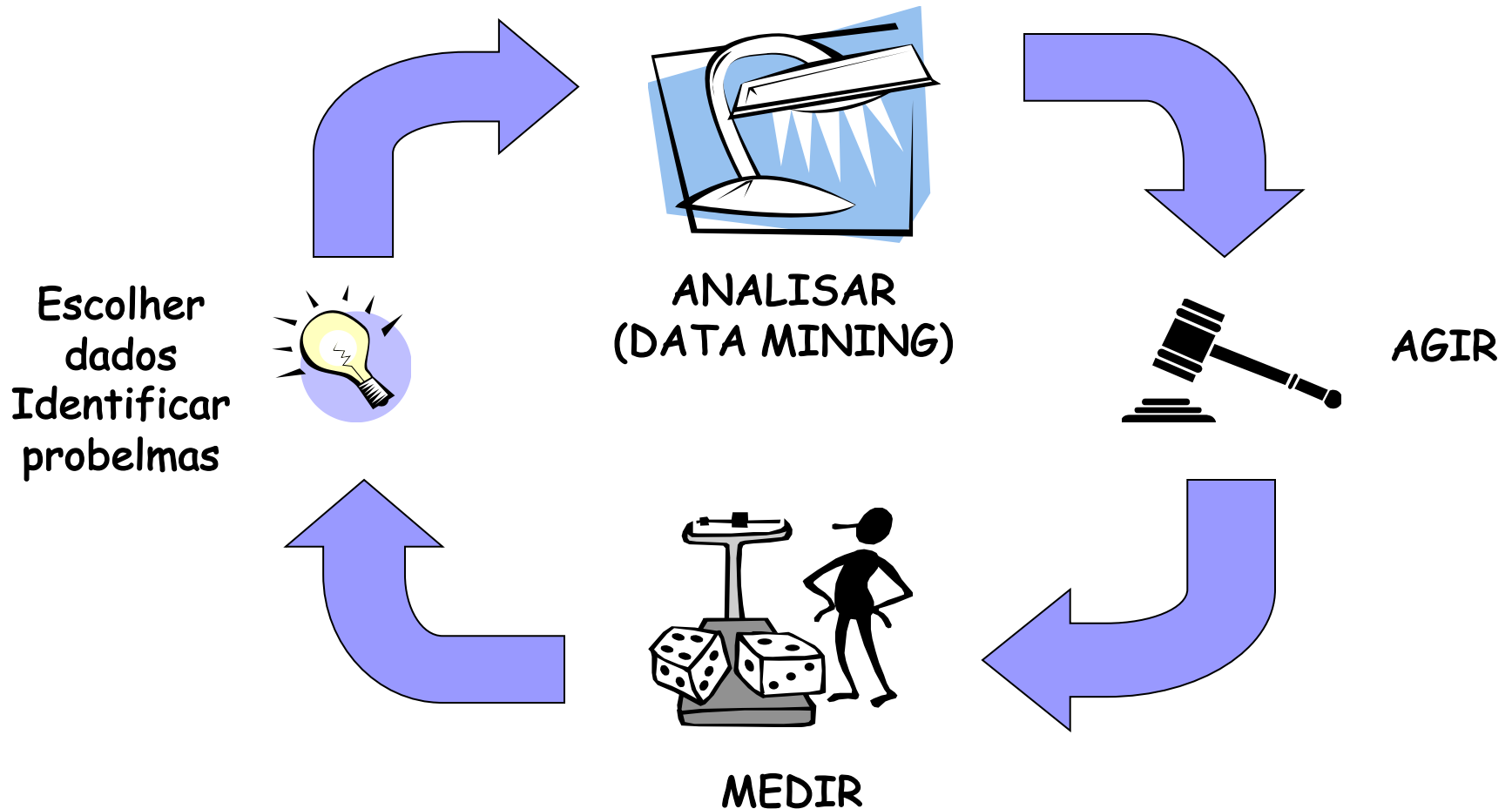
*by any other name...*




# Ciclos que nunca terminam...



# Ciclos que nunca terminam...





# Simplificando, Datamining é

**Data Mining** é o processo de extração de conhecimento útil a partir de grandes volumes de dados, recorrendo a três técnicas principais:

**1.Bases de Dados** – Organização, armazenamento e recuperação eficiente dos dados para análise.

**2.Estatística** – Métodos matemáticos para identificar padrões, tendências e correlações.

**3.Aprendizagem Máquina (Machine Learning)** – Algoritmos que aprendem com os dados para fazer previsões e classificações.

A combinação destas abordagens permite descobrir informações valiosas para apoio à decisão em diversas áreas.



# Simplificando, Datamining é

- **Predição**

- Utilizar dados históricos para prever valores ou eventos futuros
- Aplicado em áreas como previsão de vendas, detecção de fraudes e diagnóstico médico
- Métodos comuns incluem regressão, redes neurais e modelos de séries temporais

- **Descoberta de Novo Conhecimento**

- Identificar padrões, relações e informações ocultas nos dados
- Permite encontrar associações, agrupamentos e anomalias
- Aplicado em marketing, biomedicina, cibersegurança, entre outros
- Técnicas como clustering, regras de associação e análise de outliers são frequentemente usadas



# Representação dos dados

## Dados Estruturados

Os **dados estruturados** são organizados de forma rígida e previsível, geralmente armazenados em tabelas de bases de dados relacionais. Cada dado está associado a um atributo ou campo específico, tornando a recuperação e análise eficientes.

### Características dos Dados Estruturados

- Armazenados em **tabelas** (linhas e colunas).
- São **fáceis de pesquisar** usando SQL e outras ferramentas de bases de dados.
- Estrutura fixa e bem definida (cada coluna tem um tipo de dado específico, como números, datas ou texto curto).
- **Alta escalabilidade e eficiência** no processamento.

### Exemplos de Dados Estruturados

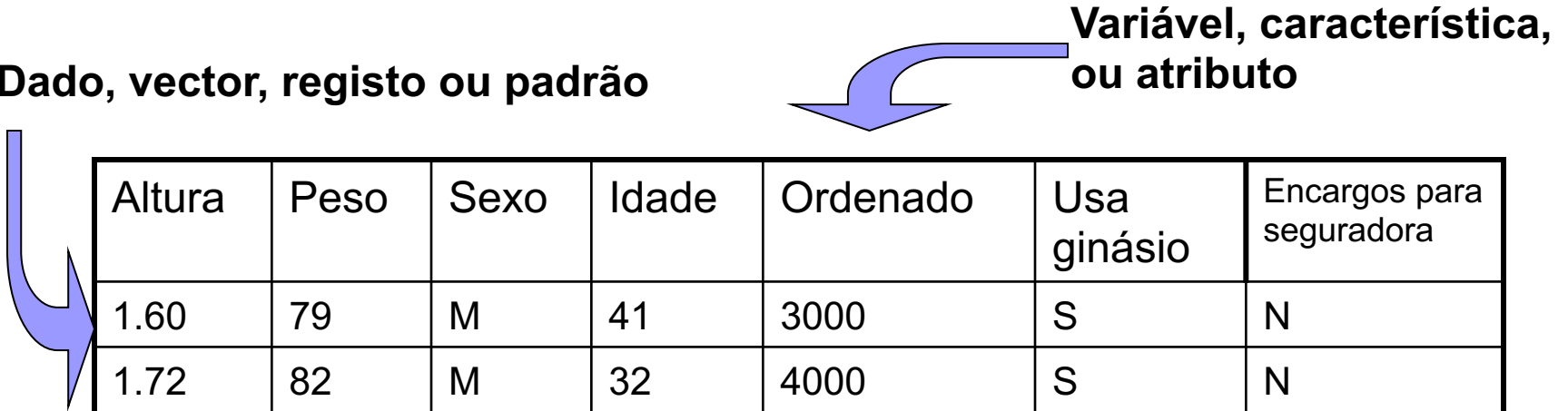
- **Bases de Dados Relacionais** (MySQL, PostgreSQL, Oracle, SQL Server).
- **Tabelas Excel** organizadas por categorias.
- **Registos financeiros** (transações bancárias, salários, impostos).
- **Sistemas de gestão de clientes (CRM)** (nomes, emails, números de telefone).
- **Dados de sensores organizados** (tempo, temperatura, pressão).

# Representação dos dados

## Dados Estruturados

**Dado, vector, registo ou padrão**

**Variável, característica,  
ou atributo**



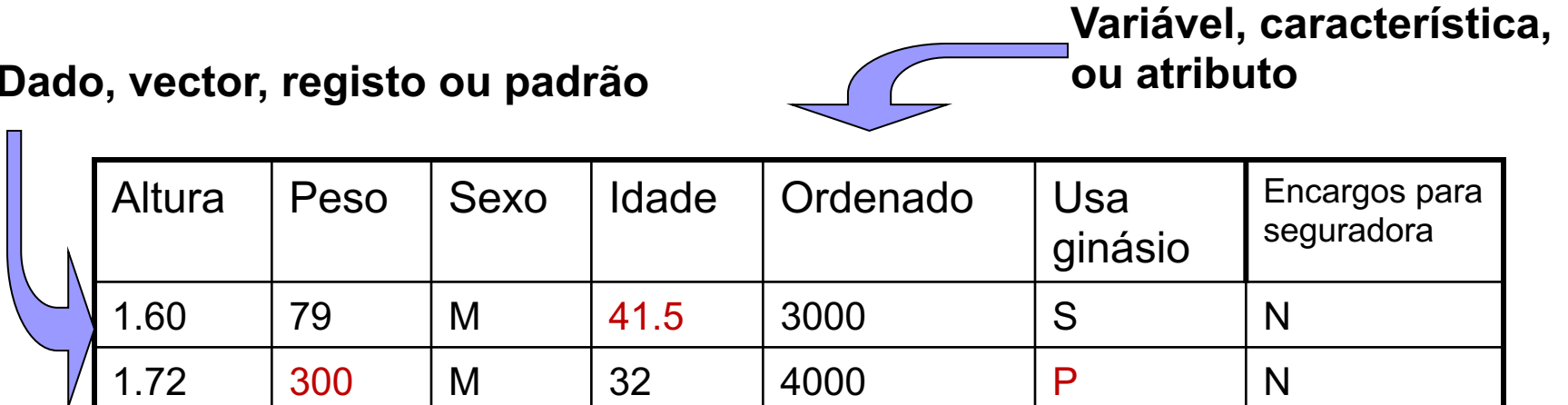
Altura	Peso	Sexo	Idade	Ordenado	Usa ginásio	Encargos para seguradora
1.60	79	M	41	3000	S	N
1.72	82	M	32	4000	S	N
1.66	65	F	28	2500	N	N
1.82	87	M	35	2000	N	S
1.71	66	F	42	3500	N	S

# Representação dos dados

## Dados Estruturados

Dado, vector, registo ou padrão

Variável, característica,  
ou atributo



Altura	Peso	Sexo	Idade	Ordenado	Usa ginásio	Encargos para seguradora
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

# Representação dos dados

## Dados Estruturados

## Inter-relacionamento

CADEIRA	
CodCad	Nome
<u>12347</u>	Bases de Dados
<u>34248</u>	Álgebra
<u>32439</u>	Introdução aos Computadores

ALUNO		
NumMec	Nome	Curso
<u>798764544</u>	João Pinto	LCC
<u>345673451</u>	Carlos Semedo	MIERSI
<u>487563546</u>	Maria Silva	LBIO
<u>452212348</u>	Pedro Costa	LMAT

INSCRIÇÃO	
NumMec	CodCad
798764544	12347
345673451	12347
798764544	34248
452212348	32439



# Representação dos dados

## Dados Estruturados

Dados não consistentes

CADEIRA	
CodCad	Nome
<u>12347</u>	Bases de Dados
<u>34244</u>	Álgebra
<u>32439</u>	Introdução aos Computadores

ALUNO		
NumMec	Nome	Curso
<u>798764544</u>	João Pinto	NULL
<u>345673451</u>	Carlos Samedo	12345
<u>798764544</u>	NULL	LBIO
<u>452212348</u>	Pedro Costa	LMAT

INSCRIÇÃO	
NumMec	CodCad
798764544	12347
345673451	12347
798764545	34248
452212346	32439



# Representação dos dados

## Dados não Estruturados

Os dados não estruturados não seguem um formato fixo, podendo conter texto livre, multimédia ou outras informações que não se encaixam facilmente numa tabela.

### **Características dos Dados Não Estruturados**

- Sem uma organização rígida (não podem ser facilmente armazenados em tabelas relacionais).
- Maior volume e complexidade, necessitando de técnicas avançadas para processamento.
- Dificuldade na pesquisa e indexação, exigindo tecnologias como Machine Learning, Processamento de Linguagem Natural (NLP) e Reconhecimento de Imagem.
- Armazenados em formatos diversos, como documentos, emails, vídeos, áudio, redes sociais.

### **Exemplos de Dados Não Estruturados**

- Emails (o texto da mensagem não segue uma estrutura fixa).
- Imagens e vídeos (não podem ser representados em colunas e linhas).
- Publicações em redes sociais (comentários, emojis, reações).
- Registos de chamadas e mensagens (conteúdo textual e voz).
- Ficheiros de texto e documentos (PDFs, Word, artigos).
- Sensores IoT que recolhem imagens e sons.

# Representação dos dados

## Dados não Estruturados

Um comentário num site de redes sociais pode ser considerado um dado não estruturado:

"Fui a este restaurante ontem e adorei o ambiente, mas a comida demorou muito tempo a chegar. O atendimento foi bom, mas acho que podiam melhorar a rapidez do serviço.



Aqui, há múltiplas informações:

- Sentimento positivo ("adorei o ambiente") e negativo ("demorou muito tempo").
- Classificação (4 estrelas).
- Sem estrutura fixa (as informações aparecem de forma livre).



# Representação dos dados

## Dados Semiestruturados

Há ainda uma terceira categoria, conhecida como dados semiestruturados, que contêm alguma organização, mas sem uma estrutura completamente fixa. Esses dados combinam características dos dois tipos anteriores.

### Exemplos de Dados Semiestruturados

- JSON e XML (dados organizados em pares chave-valor, mas sem uma estrutura rígida como num SQL).
- Emails (o remetente e o destinatário têm estrutura fixa, mas o conteúdo do email é texto livre).
- Logs de sistemas (seguem um padrão, mas contêm texto desestruturado).

# Representação dos dados


## Dados Semiestruturados

### 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
    }
  ]
}
```



# Representação dos dados

## Conclusão

- **Dados Estruturados** são fáceis de armazenar, pesquisar e analisar, mas limitam a complexidade da informação.
- **Dados Não Estruturados** contêm mais detalhes e são mais ricos em informação, mas exigem técnicas avançadas para análise.
- **Dados Semiestruturados** são um meio-termo entre os dois, possuindo alguma organização, mas ainda flexibilidade.



# Tipos de problemas

## Predição

- **Classificação**

- Atribui categorias a novos dados com base em exemplos anteriores
- Utilizado em detecção de spam, diagnóstico médico, reconhecimento de imagens
- Algoritmos comuns: Árvores de decisão, Random Forest, SVM, Redes Neurais

- **Regressão**

- Prediz valores numéricos com base em variáveis independentes
- Aplicado em previsão de vendas, análise de preços, meteorologia
- Algoritmos comuns: Regressão Linear, Regressão Logística, Redes Neurais



# Tipos de problemas

## Descoberta de Conhecimento no Data Mining

- **Deteção de Desvios**

- Identificação de anomalias ou valores atípicos nos dados
- Aplicado na deteção de fraudes, monitorização de redes e controlo de qualidade

- **Segmentação de Bases de Dados**

- Separação dos dados em grupos homogéneos para análise mais eficaz
- Utilizado em campanhas de marketing, personalização de serviços e análise de clientes

- **Clustering**

- Agrupamento automático de dados sem classes pré-definidas
- Aplicado em segmentação de clientes, biologia computacional e análise de imagens





# Tipos de problemas

## Descoberta de Conhecimento no Data Mining

### •Regras de Associação

- Descoberta de padrões frequentes e relações entre variáveis
- Exemplo: análise de cestos de compras (Market Basket Analysis)

### •Sumarização

- Extração de informações essenciais a partir de grandes volumes de dados
- Utilizado em relatórios automáticos, sistemas de recomendação e processamento de linguagem natural

### •Visualização

- Representação gráfica dos dados para facilitar a interpretação
- Ferramentas comuns: gráficos interativos, dashboards e mapas de calor

### •Pesquisa em Texto

- Análise e extração de informações úteis de grandes coleções de textos
- Aplicado em motores de busca, análise de sentimentos e categorização de documentos



## Detecção de fraudes na utilização de um cartão de crédito

As instituições financeiras enfrentam o desafio de **identificar e bloquear transações fraudulentas** sem causar inconvenientes aos clientes legítimos.

A fraude pode ocorrer de diversas formas, incluindo uso indevido de cartões roubados, clonagem de cartões e transações online não autorizadas.

Como a maioria das fraudes acontece rapidamente, é essencial ter um sistema automático que consiga identificar e bloquear transações suspeitas em tempo real.



# **Deteccção de fraudes na utilização de um cartão de crédito**

## **Modelo Adequado para a Solução**

- Este problema envolve a atribuição de um rótulo a cada transação, classificando-a como fraudulenta ou legítima.
- O modelo deve aprender a partir de transações anteriores, identificando padrões que distinguem operações normais de atividades suspeitas.
- Assim, para cada nova transação, o sistema poderá prever se deve ser aprovada ou bloqueada com base nas características observadas.

## Detecção de fraudes na utilização de um cartão de crédito

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



# Detecção de fraudes na utilização de um cartão de crédito

## Desafios do Modelo

- **Desequilíbrio de Classes** – A maioria das transações são legítimas, tornando difícil para o modelo identificar corretamente as fraudes.
- **Evolução dos Padrões de Fraude** – atacantes adaptam-se e mudam estratégias, tornando alguns modelos obsoletos rapidamente.
- **Minimização de Falsos Positivos** – Bloquear transações legítimas pode causar insatisfação dos clientes e impactar a experiência do utilizador.
- **Tempo de Resposta** – As decisões devem ser tomadas em tempo real para evitar atrasos no processamento das compras.
- **Segurança e Privacidade** – O modelo deve garantir que os dados dos clientes são protegidos e que não há violações de privacidade.



# Detecção de fraudes na utilização de um cartão de crédito

## Resultados e Benefícios

- **Redução de Perdas Financeiras** – Detecção precoce de fraudes minimiza prejuízos.
- **Maior Segurança para Clientes** – Proteção contra acessos não autorizados aos cartões.
- **Menos Bloqueios Indevidos** – O modelo aprende a distinguir compras legítimas de suspeitas.
- **Monitorização Automática** – O sistema analisa todas as transações sem necessidade de intervenção manual.







## Previsão do Tempo de Ocorrência de um Ataque Cibernético

Empresas e organizações enfrentam um aumento constante de ciberataques, como tentativas de intrusão, ataques de negação de serviço (DDoS) e exploração de vulnerabilidades.

Um dos desafios na segurança informática é **prever quando ocorrerá o próximo ataque**, permitindo que as equipas de segurança adotem medidas preventivas.

Se for possível estimar com precisão o tempo até a próxima tentativa de ataque, as defesas podem ser ajustadas proativamente, reduzindo o impacto de potenciais violações.





## Previsão do Tempo de Ocorrência de um Ataque Cibernético

Este problema requer a **previsão de um valor contínuo**: o tempo até que ocorra o próximo ataque cibernético. Assim, trata-se de um caso de **Regressão**, onde o modelo analisa padrões históricos de ataques e fatores contextuais para estimar o intervalo de tempo entre eventos maliciosos.

O modelo pode ser treinado com base em dados históricos, considerando variáveis como:

- Frequência de ataques anteriores
- Tipo de ataque detetado
- Volume de tráfego anômalo na rede
- Número de vulnerabilidades ativas no sistema
- Tendências sazonais de ataques em determinados períodos do ano



## **Previsão do Tempo de Ocorrência de um Ataque Cibernético**

### **Desafios do Modelo**

- Padrões Irregulares – Os ataques cibernéticos podem não seguir padrões fixos, tornando difícil prever com exatidão o tempo exato até o próximo evento.
- Evolução das Ameaças – Novas técnicas de ataque podem surgir, tornando os dados históricos menos representativos do comportamento futuro dos invasores.
- Influência de Fatores Externos – Eventos globais, como divulgação de novas vulnerabilidades ou tensões políticas, podem afetar a frequência dos ataques, o que pode ser difícil de modelar.
- Qualidade dos Dados – Logs de segurança podem conter lacunas ou falsos positivos, afetando a precisão das previsões.
- Tempo de Resposta Rápido – O modelo precisa de fornecer previsões em tempo útil para que as equipes de segurança possam agir antes que um ataque ocorra.





## **Deteção de Irregularidades em Acessos a Sistemas Corporativos**

Empresas e instituições armazenam informações sensíveis em sistemas internos que devem ser protegidos contra acessos não autorizados.

No entanto, ataques internos e externos podem comprometer esses dados.

A deteção manual de acessos suspeitos é inviável devido ao grande volume de registos de login e atividades dos utilizadores.

Assim, é necessário um sistema que identifique Irregularidades nos padrões de acesso, ajudando as equipas de segurança a responder rapidamente a possíveis ameaças.



## **Deteção de Irregularidades em Acessos a Sistemas Corporativos**

### **Modelo Adequado para a Solução**

Este problema envolve a identificação de comportamentos anormais sem que haja, necessariamente, exemplos prévios de ataques específicos. Para isso, o modelo precisa de analisar os padrões normais de acesso dos utilizadores e **sinalizar atividades que se desviam desses padrões**. Como o objetivo não é classificar eventos previamente conhecidos, mas sim descobrir novas ameaças, esta **abordagem é baseada em Descoberta de Conhecimento**, mais especificamente na **deteção de anomalias**.

#### **O modelo analisará dados como:**

- Horário de login dos utilizadores
- Dispositivos e locais habituais de acesso
- Número de tentativas de login falhadas
- Comandos e ficheiros acedidos
- Volume de dados transferidos
- Se um utilizador apresentar um comportamento significativamente diferente do seu padrão normal (por exemplo, um login a partir de um país incomum ou um acesso massivo a ficheiros confidenciais), o sistema deve gerar um alerta para investigação.





## **Deteção de Irregularidades em Acessos a Sistemas Corporativos**

### **Desafios**

- Definição do que é uma Anomalia – Nem toda atividade incomum é maliciosa. O modelo deve evitar falsos positivos, como viagens legítimas de funcionários.
- Evolução dos Comportamentos – Os padrões normais de acesso dos utilizadores podem mudar ao longo do tempo, exigindo um modelo adaptável.
- Dados Desbalanceados – A maioria dos acessos será legítima, tornando difícil a deteção de ataques raros e sofisticados.
- Ambientes Dinâmicos – Empresas grandes têm milhares de utilizadores e dispositivos, o que pode gerar um volume enorme de dados para análise em tempo real.
- Privacidade e Ética – O modelo deve garantir que a monitorização dos acessos respeita as normas de proteção de dados e privacidade dos funcionários..





## **Agrupamento de Utilizadores para Prevenção de Ameaças Internas**

As empresas lidam com riscos internos, como o vazamento de informações confidenciais, acessos indevidos a sistemas críticos e comportamentos suspeitos de funcionários ou colaboradores. Para mitigar esses riscos, é essencial agrupar os utilizadores da organização com base nos seus padrões de comportamento, identificando grupos de maior risco e permitindo a adoção de medidas preventivas.

O Agrupamento permite categorizar os utilizadores conforme a sua atividade dentro do sistema, distinguindo padrões normais de possíveis ameaças internas.





## **Agrupamento de Utilizadores para Prevenção de Ameaças Internas- Modelo Adequado**

Este problema insere-se na Descoberta de Conhecimento, pois o objetivo não é prever um evento específico, mas identificar padrões e criar grupos distintos de utilizadores com base nas suas interações. Neste caso, utiliza-se Segmentação de Bases de Dados, um processo que agrupa indivíduos com comportamentos semelhantes, sem que haja rótulos prévios nos dados.

O modelo analisa variáveis como:

- Frequência e horário de acessos aos sistemas
- Tipos de ficheiros manipulados
- Recursos e bases de dados acedidos
- Utilização de dispositivos externos (USB, discos externos, etc.)
- Tentativas de acesso não autorizadas



## **Agrupamento de Utilizadores para Prevenção de Ameaças Internas- Modelo Adequado**

Após a segmentação, os utilizadores podem ser agrupados em diferentes perfis, como:

- Utilizadores padrão – Funcionários com acessos normais e sem comportamentos suspeitos.
- Utilizadores privilegiados – Administradores de sistema ou gestores com permissões avançadas.
- Utilizadores de risco – Funcionários com comportamentos anómalos, como acessos fora do horário habitual ou tentativa de extração massiva de dados.
- Este agrupamento permite que as equipas de segurança reforcem medidas para os segmentos de maior risco, como a implementação de alertas automáticos ou a revisão de permissões de acesso.



## **Agrupamento de Utilizadores para Prevenção de Ameaças Internas-Desafios**

Definição de Grupos Relevantes – Nem sempre é óbvio quantos segmentos devem ser criados ou quais características são mais relevantes para a segmentação.

Dados Desbalanceados – A maioria dos utilizadores pode ter um comportamento normal, tornando difícil identificar padrões anómalos sem gerar falsos positivos.

Evolução dos Comportamentos – As atividades dos utilizadores podem mudar com o tempo, exigindo uma atualização constante do modelo.

Interpretação dos Resultados – A segmentação pode revelar padrões inesperados, exigindo análise manual para compreender a sua relevância.

Privacidade e Conformidade – A monitorização de funcionários pode levantar questões éticas e legais, exigindo conformidade com regulamentações de proteção de dados.





## **Identificação de Grupos de Dispositivos Comprometidos numa Rede Empresarial**

Empresas e organizações enfrentam desafios crescentes na deteção de dispositivos comprometidos dentro das suas redes. Cibercriminosos podem explorar vulnerabilidades em computadores, servidores ou dispositivos IoT para lançar ataques internos, roubar dados ou criar botnets.

Como muitas destas ameaças operam de forma discreta, torna-se essencial identificar padrões anómalos no comportamento dos dispositivos sem depender de listas prévias de ameaças conhecidas.





## Identificação de Grupos de Dispositivos Comprometidos numa Rede Empresarial

Este problema exige a descoberta de padrões desconhecidos dentro do tráfego da rede, agrupando dispositivos com comportamentos semelhantes para identificar anomalias. Como não há categorias pré-definidas, a melhor abordagem é o Clustering, que agrupa dispositivos em clusters com base nas suas características de tráfego, sem precisar de rótulos prévios.

Os dados analisados podem incluir:

- Volume de tráfego enviado e recebido
- Protocolos utilizados (HTTP, HTTPS, FTP, SSH, etc.)
- Padrões de comunicação com outros dispositivos
- Número de pedidos a domínios suspeitos
- Frequência e horário das conexões

O modelo organizará os dispositivos em diferentes grupos com padrões de comportamento semelhantes. Caso surja um novo grupo pequeno e distinto dos demais, pode indicar um conjunto de máquinas comprometidas a atuar de forma maliciosa.



## **Identificação de Grupos de Dispositivos Comprometidos numa Rede Empresarial - Desafios do Modelo**

- Definição do Número Ótimo de Clusters – O modelo precisa de determinar quantos grupos existem na rede, o que pode não ser evidente à partida.
- Diferenciação entre Comportamentos Normais e Maliciosos – Algumas máquinas podem ter padrões legítimos de uso que parecem anómalos, como servidores que comunicam frequentemente com múltiplos dispositivos.
- Evolução Contínua da Rede – Novos dispositivos entram e saem da rede regularmente, exigindo que o modelo se adapte dinamicamente.
- Grande Volume de Dados – O tráfego de rede gera grandes quantidades de informação, exigindo métodos eficientes para processamento em tempo real.
- Interpretação dos Resultados – Após a identificação dos clusters, é necessário um processo de análise para determinar quais grupos representam ameaças reais.





# Modelos versus Dados

## ■ Model based

- As relações entre variáveis são fundamentadas em leis físicas, princípios matemáticos ou conhecimento teórico estabelecido.
- Se utilizamos a equação  $P=mg$  para descrever a força gravitacional exercida sobre um objeto (causa) resulta da aceleração da gravidade (efeito) dependendo da massa do objeto

## ■ Data driven

- Procuram relações nos dados
  - Relações não implicam causa/efeito
- Ou não há modelo, ou há um modelo genérico que normalmente é um aproximador universal (com muitos parâmetros)
- Se tivermos apenas dados de força e aceleração sem conhecer a equação, um modelo data-driven pode encontrar um padrão que relaciona ambas as variáveis, mas sem compreender a **razão pela qual a relação existe**.

# Exemplo: Reconhecimento de gatos em imagens

## **Abordagem Model-Based**

Criamos um conjunto de regras fixas para identificar gatos em imagens:

- Se tiver orelhas pontudas + olhos grandes + focinho pequeno, então é um gato.
- Se não tiver essas características, não é um gato.

Problema: Se a imagem estiver escura ou o gato estiver virado, o modelo pode falhar.



# Exemplo: Reconhecimento de gatos em imagens

## Abordagem Data-Driven

- Treinamos uma rede neural com milhares de imagens de gatos e cães.
- O modelo aprende sozinho quais características distinguem um gato de um cão.
- Quando vê uma nova imagem, ele classifica corretamente sem precisar de regras manuais.
- O modelo aprende automaticamente padrões nos dados, sem precisar de regras fixas.

Problema: Precisa de muitos dados para treinar