

Extração de Conhecimento

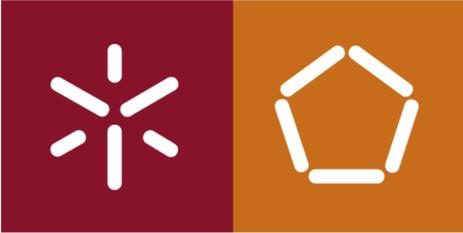
Definição

?

Extração de Conhecimento

Definição

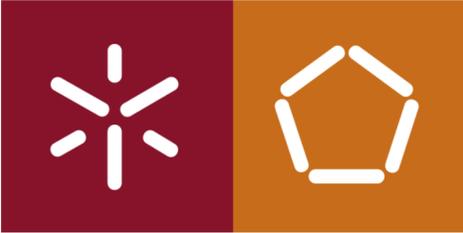
- A partir de dados estruturados inferir padrões
- Filtrar informação
- Interpretação de dados
-



Extração de Conhecimento

Motivações

- “Explosão” na **disponibilização de dados**;
- **Fontes de dados** abundantes:
 - Negócios: comércio eletrónico, gestão de stocks, administração de recursos humanos;
 - Web: recolha e disponibilização de informação;
 - Ciências: simulação, aquisição de dados automática, saúde;
 - Sociedade: notícias, câmaras digitais de fotografia e de vídeo, leitores de áudio, funerárias.
- **Solidificação** das Tecnologias de Bases de Dados;
- Disponibilização de **ferramentas automáticas** de procura e arquivo de informação;
- **Armazenamento digital** de informação promove aumento significativo na quantidade de dados disponíveis;
- **Custo do arquivo** de informação diminui drasticamente.



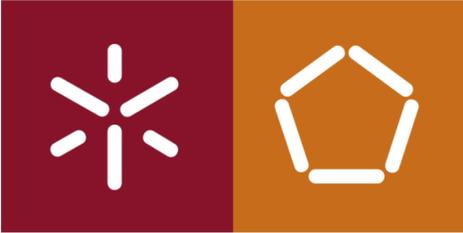
Extração de Conhecimento

Motivação

Data Warehouse

Data Mining

- Qualquer organização precisa utilizar toda informação disponível para criar e manter vantagem competitiva.
- Conseguir tomar decisões corretas e rápidas.
- Recurso a Sistemas de suporte à decisão (**Decision Support Systems - DSS**);
 - ferramentas para análise,
 - front-ends que suportem consultas ad hoc,
 - interfaces gráficas apropriadas, etc.
- Um **data warehouse** tem integrar os dados internos e externos de uma organização numa estrutura única permitindo uma melhor utilização, aumentando a sua capacidade de resposta e adaptação.



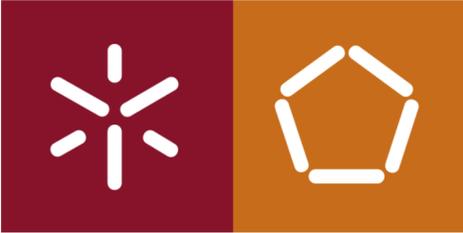
Extração de Conhecimento

Motivação

Data Warehouse

Data Mining

- A integração de sistemas
 - **OLAP (On-Line Analytical Processing)** e;
 - **Data mining.**
- Pode ajudar as organizações a descobrir novas formas de competir numa economia globalizada,
 - Potenciando novos e melhores produtos e/ou serviços,
 - Mais rápida do que a concorrência,
 - Sem aumentar o custo dos produtos e/ou serviços.
- Adaptada às características e às expectativas de cada organização;
- Ajudar a descobrir maneiras diferentes de atuar no Mercado;
- Ajudar organização mudar para atender as novas realidades.



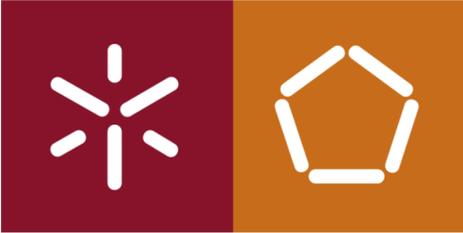
Extração de Conhecimento

Motivação

Data Warehouse

Data Mining

- Um **data warehouse** é uma coleção de dados **orientada por assuntos, integrada, variável no tempo**, que tem por objetivo dar suporte aos processos de tomada de decisão.
- É um base de dados contendo dados extraídos do ambiente de produção da organização, que foram selecionados, depurados e otimizados para processamento de consultas e não para processamento de transações.
- Em geral, um **data warehouse** requer a consolidação de outros recursos de dados além dos armazenados em base de dados relacionais, incluindo informações provenientes de folhas de calculo, documentos, etc.



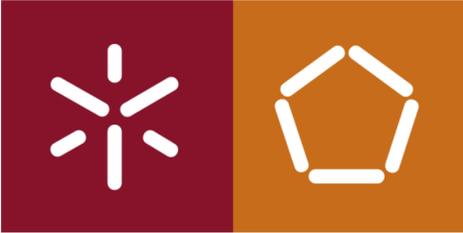
Extração de Conhecimento

Motivação

Data Warehouse

Data Mining

- O objetivo de um data warehouse é fornecer uma **"imagem única da realidade da organização"**.
- Sistemas de data warehouse compreendem:
 - programas que extraem dados do ambiente de dados operacionais da empresa,
 - uma base de dados que os mantém, e
 - sistemas que fornecem estes dados aos utilizadores.
- Permitem que sistemas mais antigos continuem em operação;
- Consolidam dados inconsistentes dos sistemas mais antigos em conjuntos coerentes;
- Extraem novas informações oriundas das operações correntes;
- Disponibilizam um ambiente para o planeamento e arquitetura de novos sistemas operacionais.



Extração de Conhecimento

Motivação

Data Warehouse

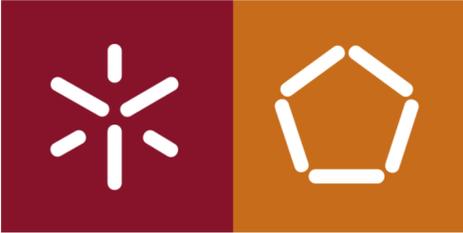
Data Mining

As formas de extração mais comuns no mercado hoje são:

- Ferramentas de consulta e emissão de relatórios;
- EIS (Executive Information Systems);
- Ferramentas OLAP;
- Ferramentas Data mining.

Integração com o ambiente Web, permitindo maior agilidade em consultas estáticas e dinâmicas.

•



Extração de Conhecimento

Motivação

Data Warehouse

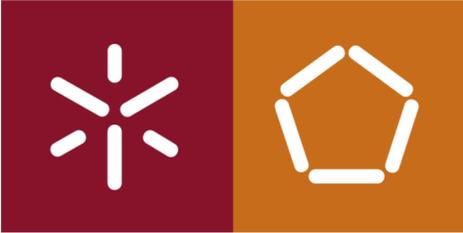
Data Mining

Neste unidade curricular vamos analisar de forma autônoma as ferramentas:

- **OLAP e**
- **Data mining.**

A diferença básica entre ferramentas **OLAP** e **data mining** está na maneira como a exploração dos dados é abordada.

- **Com ferramentas OLAP a exploração é feita na base da verificação, o analista conhece a questão, elabora uma hipótese e utiliza a ferramenta para confirmá-la.**
- **Com data mining, a questão é total ou parcialmente desconhecida e a ferramenta é utilizada para a busca de conhecimento.**



Extração de Conhecimento

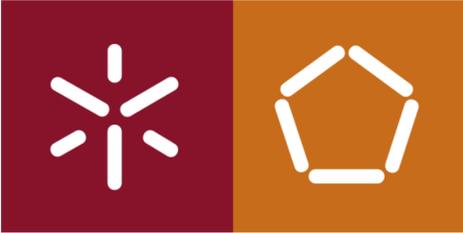
Motivação

Data Warehouse

Data Mining

OLAP (On-Line Analytical Processing)

- Conjunto de tecnologias projetadas para suportar análise e consultas ad hoc.
- Sintetizam informações sobre a organização, através de comparações, visões personalizadas, análise histórica e projeção de dados em vários cenários de "e se...".
- São implementados para ambientes multiutilizador em arquiteturas cliente-servidor e oferece respostas rápidas e consistentes às consultas iterativas executadas, independente do tamanho e complexidade da base de dados.



Extração de Conhecimento

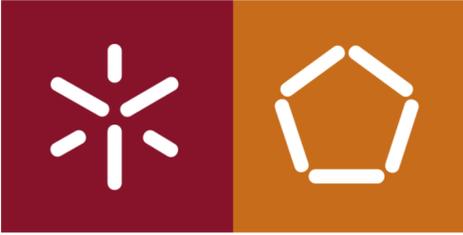
Motivação

Data Warehouse

Data Mining

OLAP (On-Line Analytical Processing)

- A principal característica dos sistemas OLAP é permitir uma visão conceitual multidimensional dos dados de uma organização.
- A visão multidimensional é muito mais útil do que a tradicional visão tabular utilizada nos sistemas de processamento de transação.
- É mais natural, fácil e intuitiva, permitindo a visão em diferentes perspectivas das atividades da organização, facilitando a exploração da informação.
-



Extração de Conhecimento

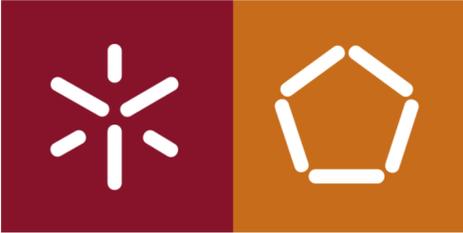
Motivação

Data Warehouse

Data Mining

OLAP (On-Line Analytical Processing)

- A arquitetura OLAP possui três componentes principais:
 - modelo de negócios para análises interativas, uma linguagem gráfica que permita diversas visões e níveis de detalhes dos dados;
 - motor OLAP para processar consultas multidimensionais na procura de um dado-alvo; e
 - mecanismo para armazenar os dados a serem analisados.
- Na base de dados definimos se o pacote é:
 - ROLAP, que faculto um interface com uma base de dados relacional de mercado, ou
 - MOLAP, que se liga a um servidor OLAP, através de uma base de dados multidimensional e dedicada.



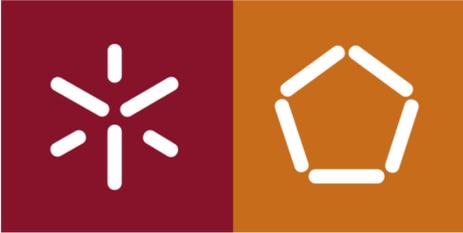
Extração de Conhecimento

Motivação

Data Warehouse

Data Mining

- Quando surgiram os primeiros estudos sobre **data warehouse**, o **data mining** era visto como um subconjunto das atividades associadas ao warehouse.
- Hoje os caminhos do **warehouse** e do **mining** tornaram-se áreas complementares.
- Enquanto o **warehouse** pode ser uma boa fonte de dados para minerar, o **data mining** é reconhecido como uma tarefa genuína, e não mais como uma subtarefa do warehouse.
- O termo data mining ainda provoca certa confusão quanto à sua definição.
- Data mining (**mineração de dados**) é o processo de extrair informação válida, previamente desconhecida e de máxima abrangência a partir de grandes bases de dados.



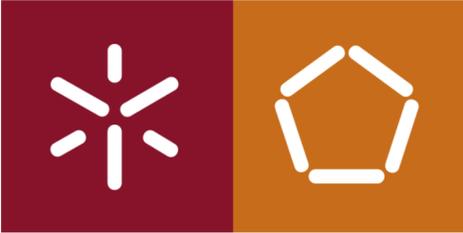
Extração de Conhecimento

Motivação

Data Warehouse

Data Mining

- Data mining é muito mais do que uma simples consulta a uma base de dados, permite aos utilizadores explorar e inferir informação útil, descobrindo relacionamentos escondidos na base de dados.
- Pode ser considerada uma forma de descoberta de conhecimento em bases de dados (**KDD - Knowledge Discovery in Databases**),
- Envolvendo Inteligência Artificial e Bases de Dados.



Extração de Conhecimento

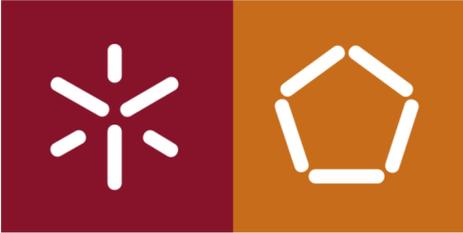
Motivação

Data Warehouse

Data Mining

Definições:

- A Descoberta de Conhecimento em Bases de Dados (DCBD) pode ser definida como um processo interactivo não trivial (na medida em que se trata de um processo com alguma autonomia de pesquisa) de identificar novos padrões nos dados que sejam válidos, potencialmente úteis e interpretáveis.
- Mineração de Dados é a procura de relações e padrões globais que existem em Bases de Dados mas estão escondidas entre a vastidão de dados.



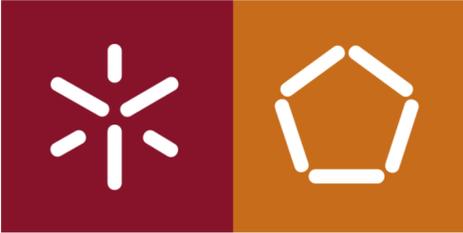
Extração de Conhecimento

Motivação

Data Warehouse

Data Mining

- **Dados** - são um conjunto de factos **F** (casos de uma base de dados).
- **Padrão** - é uma expressão **E** numa linguagem **L** que descreve um subconjunto de factos **FE** do conjunto **F**. Por exemplo em relação aos dados sobre empréstimos bancários, o padrão **E1** = “**Se Salário < T Então a pessoa faltou ao pagamento**” poderia ser um padrão para uma escolha apropriada de **T**.
- **Validade** - os padrões descobertos devem ser válidos nos novos dados com algum grau de confiança. O grau de confiança associado a um padrão **E** pode ser definido como a função $c = C(E, F)$.
- **Novidade** - os padrões devem ser novos para o sistema. A novidade pode ser medida em relação às alterações verificadas ao nível dos dados ou do conhecimento. É representada por uma função $N(E, F)$ que pode ser lógica ou real.



Extração de Conhecimento

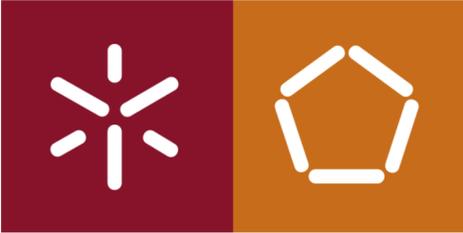
Motivação

Data Warehouse

Data Mining

- **Utilidade** - representa o grau de utilidade de um padrão, isto é, até que ponto o padrão contribui para os objectivos inerentes ao processo, como por exemplo o esperado aumento de lucro de um banco por aplicação da regra de decisão $E1$. A utilidade pode ser definida pela função $u=U(E,F)$.
- **Interpretável** - um dos objectivos da DCBC é gerar padrões que sejam compreendidos pelos humanos na perspectiva de contribuir para uma melhor compreensão dos dados. Assume-se que o grau de interpretação de um padrão é definido pela função $s=S(E,F)$.

O conhecimento descoberto pode também ser quantificado, seja $i=I(E,F,C,N,U,S)$ o grau de interesse num dado padrão E , diz-se que o padrão E é conhecimento se para um valor dado i , $I(E,F,C,N,U,S) > i$.



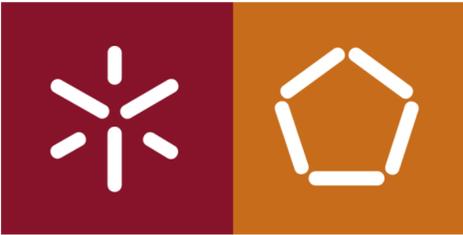
Extração de Conhecimento

Motivação

Data Warehouse

Data Mining

- Gestão de Produção,
- Gestão Inteligente de Encomendas,
- Gestão de inventários,
- Gestão financeira,
- Análise de risco,
- Gestão de transportes,
- Gestão de frotas,
- Telecomunicação,
- Análise de chamadas,
- Saúde,
- Análise de resultados,
- Marketing,
- Perfil dos consumidores,
- Seguros,
- Detecção de fraudes, etc.



Extração de Conhecimento

Motivação

Data Warehouse

Data Mining

Domínios

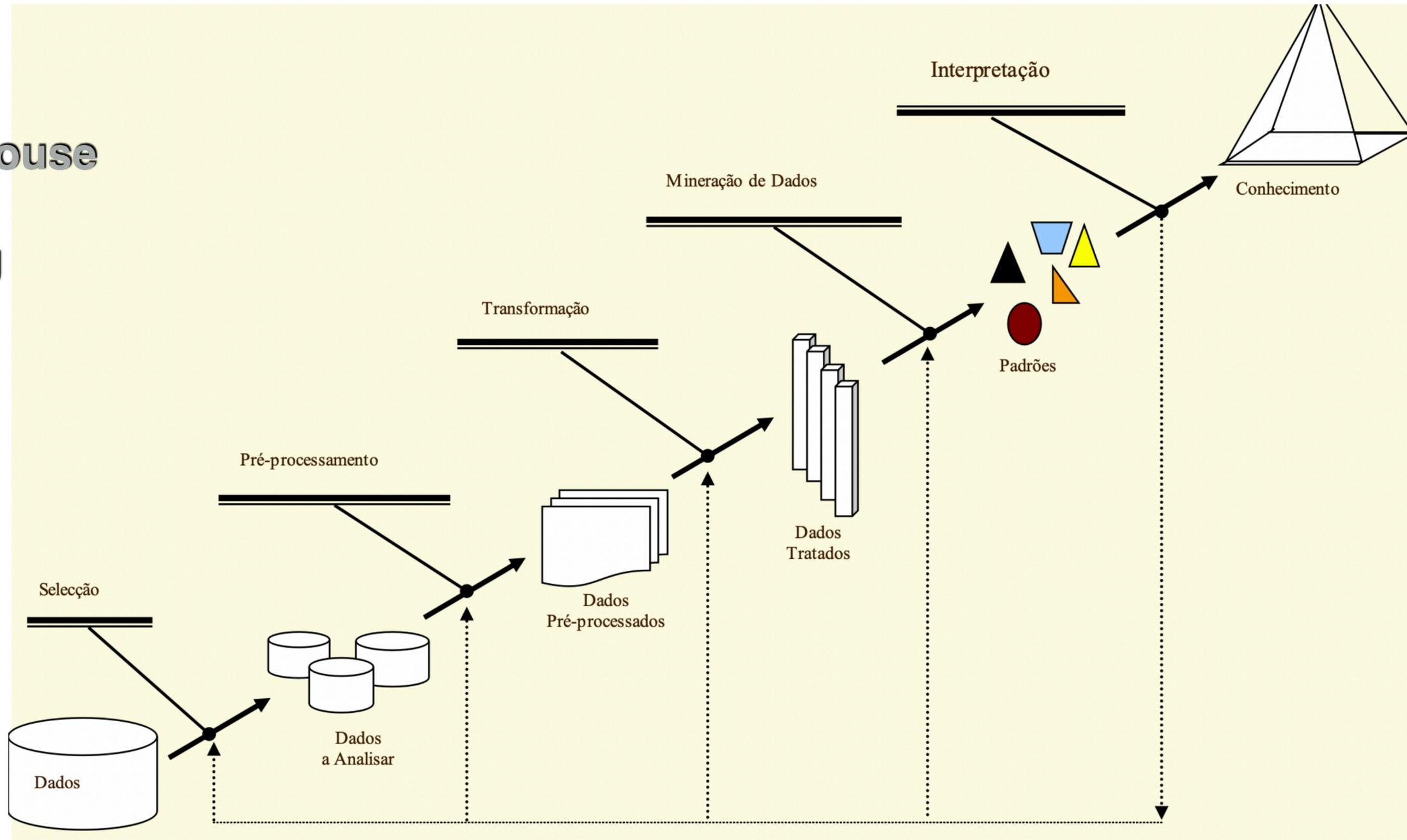
- CRM
- Descoberta de Conhecimento em Bases de Dados
- Database Marketing
- Apoio à Decisão
- Web / Business Intelligence
- Análise Inteligente de Dados
- Data Mining

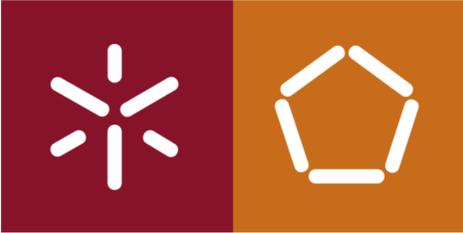
Extração de Conhecimento

Motivação

Data Warehouse

Data Mining





Extração de Conhecimento

Motivação

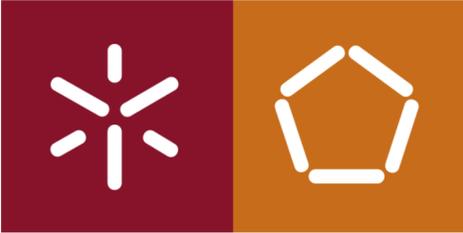
Data Warehouse

Data Mining

Seleção

Seleção

- Aprendizagem do domínio de aplicação.
- Seleção dos dados a analisar; i.e., agrupar os dados ou conjunto de variáveis sobre os quais se pretende trabalhar.



Extração de Conhecimento

Motivação

Data Warehouse

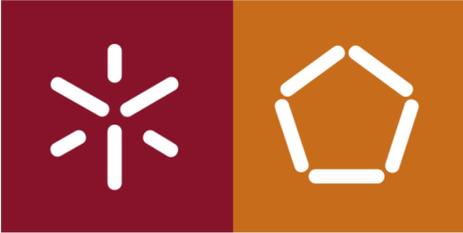
Data Mining

Seleção

Pré-Processamento

Pré-processamento e limpeza de dados

- operações básicas de remoção de ruído nos dados;
- decisão de estratégias em caso de campos omissos nos dados;
- consideração de sequências temporais nos dados.



Extração de Conhecimento

Motivação

Data Warehouse

Data Mining

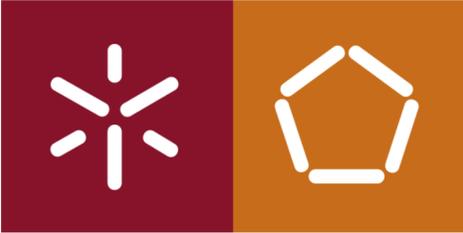
Seleção

Pré-Processamento

Transformação

Transformação de dados

- Procura de características úteis nos dados tendo em consideração os objectivos a que se destinam;
- utilização de métodos de transformação com vista à redução do numero efectivo de variáveis em consideração;
- procura de representações invariantes para os dados



Extração de Conhecimento

Motivação

Data Warehouse

Data Mining

Seleção

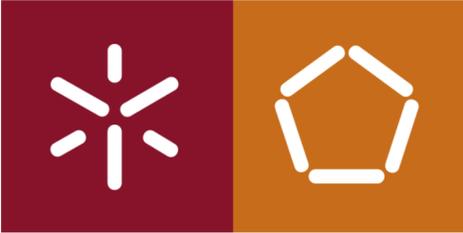
Pré-Processamento

Transformação

Mineração

Mineração de dados

- Escolha do objectivo da mineração de dados (previsão, descrição) tendo em consideração o processo global de DCBD
- **Por exemplo: Atribuição de um empréstimo bancário.**



Extração de Conhecimento

Motivação

Data Warehouse

Data Mining

Seleção

Pré-Processamento

Transformação

Mineração

Mineração de dados

- **Classificação** (aprendizagem de uma função que faça o mapeamento de um elemento dos dados em uma ou várias classes)
- **Regressão** (aprendizagem de uma função que faça o mapeamento de um elemento dos dados na valoração de uma variável de previsão)
- **Clustering/Segmentação** (identificação de um conjunto finito de categorias ou clusters para descrição dos dados)
- **Sumariação** (utilização de métodos para procura de uma descrição compacta para um subconjunto de dados)
- **Modelação de Dependências** (associações: procura de um modelo que descreva dependências significativas entre variáveis)
- **Detecção de Alterações e Divergências** (descoberta das alterações mais significativas nos dados a partir de valores medidos previamente ou normativos)

Extração de Conhecimento

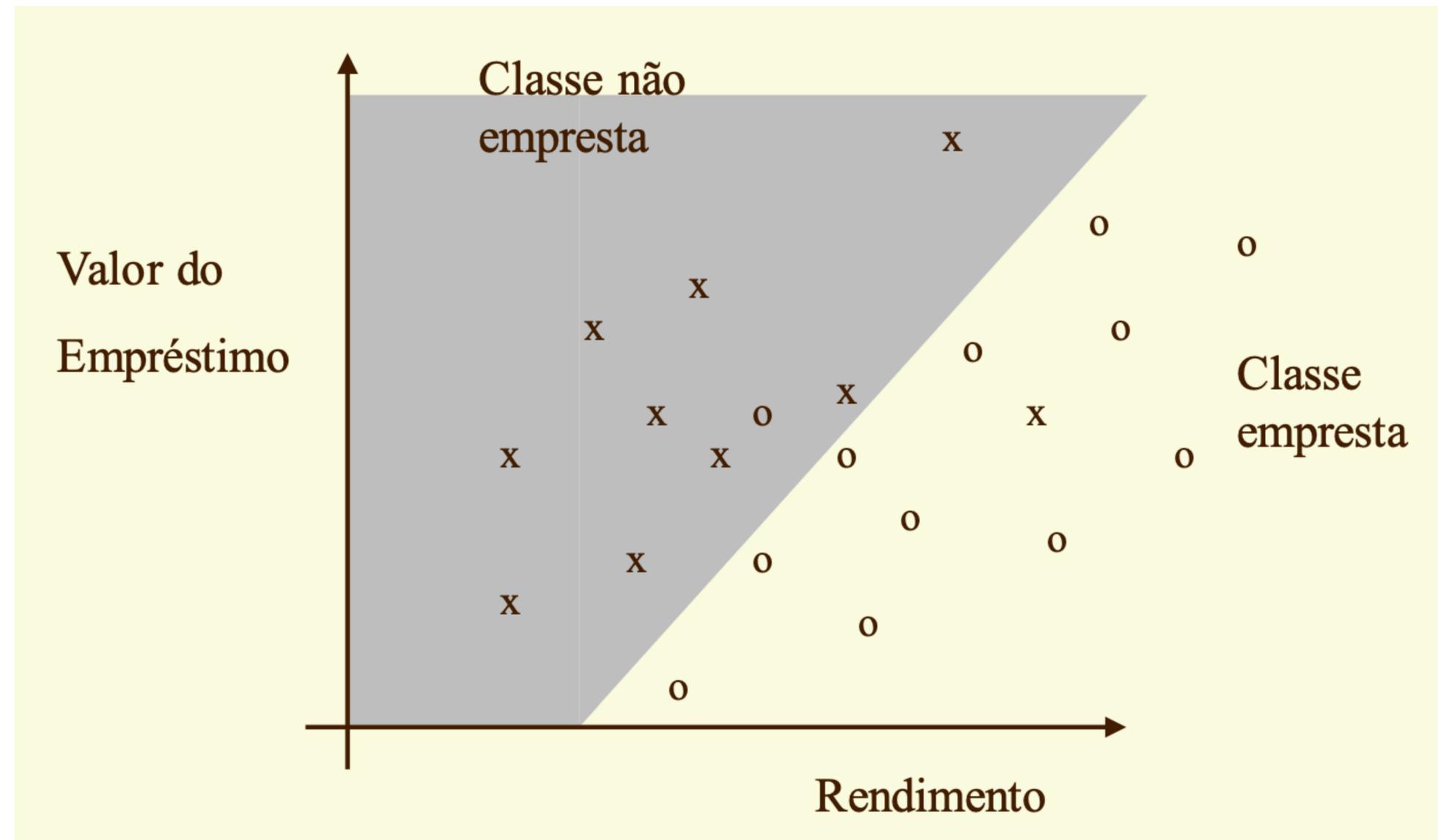
Motivação

Data Warehouse

Data Mining

Seleção
Pré-Processamento
Transformação
Mineração

Exemplo de classificação linear:



Extração de Conhecimento

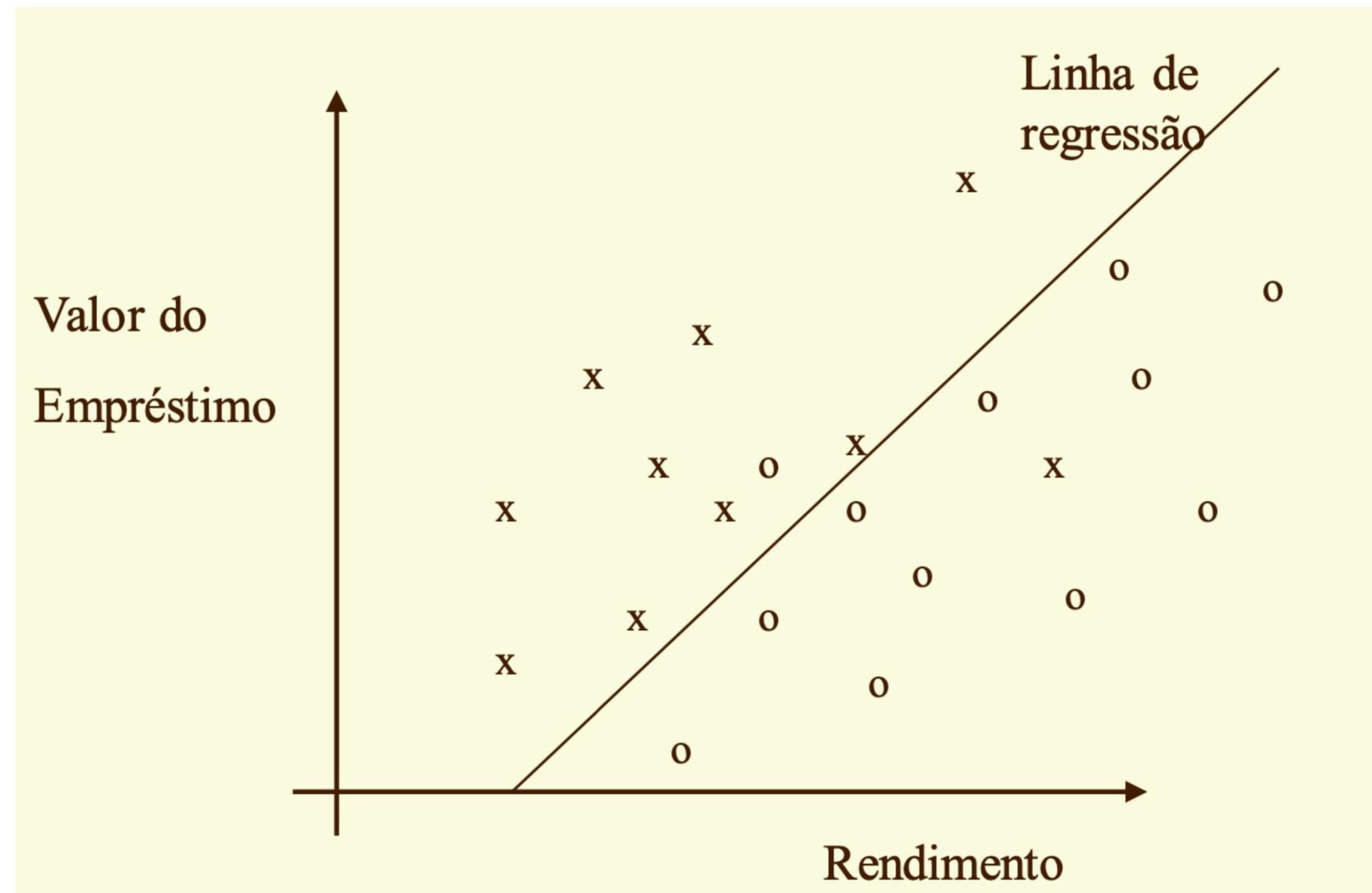
Motivação

Data Warehouse

Data Mining

Seleção
Pré-Processamento
Transformação
Mineração

Exemplo de Regressão linear:



Extração de Conhecimento

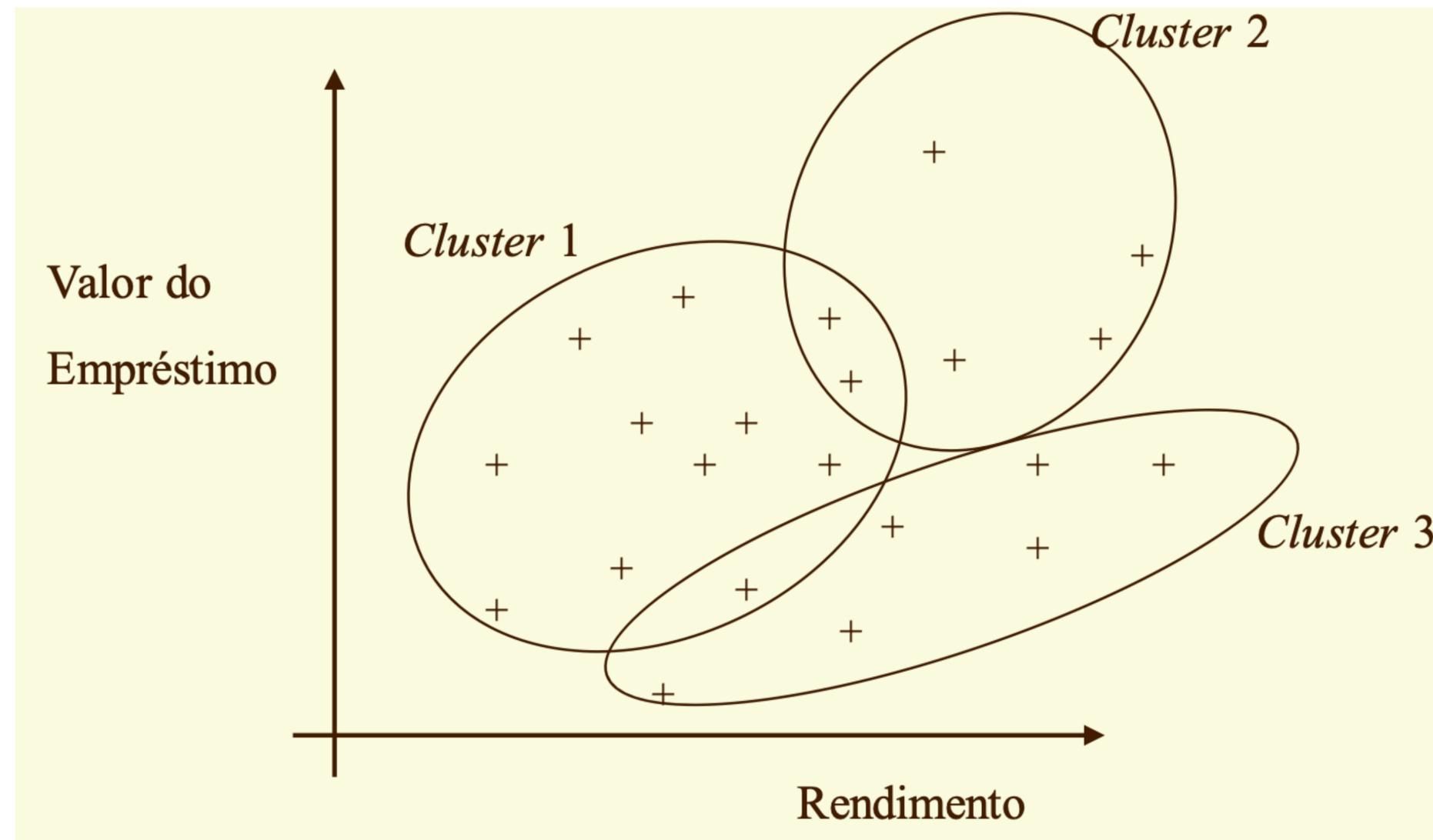
Motivação

Data Warehouse

Data Mining

Seleção
Pré-Processamento
Transformação
Mineração

Exemplo de *Clustering/Segmentação* :



Extração de Conhecimento

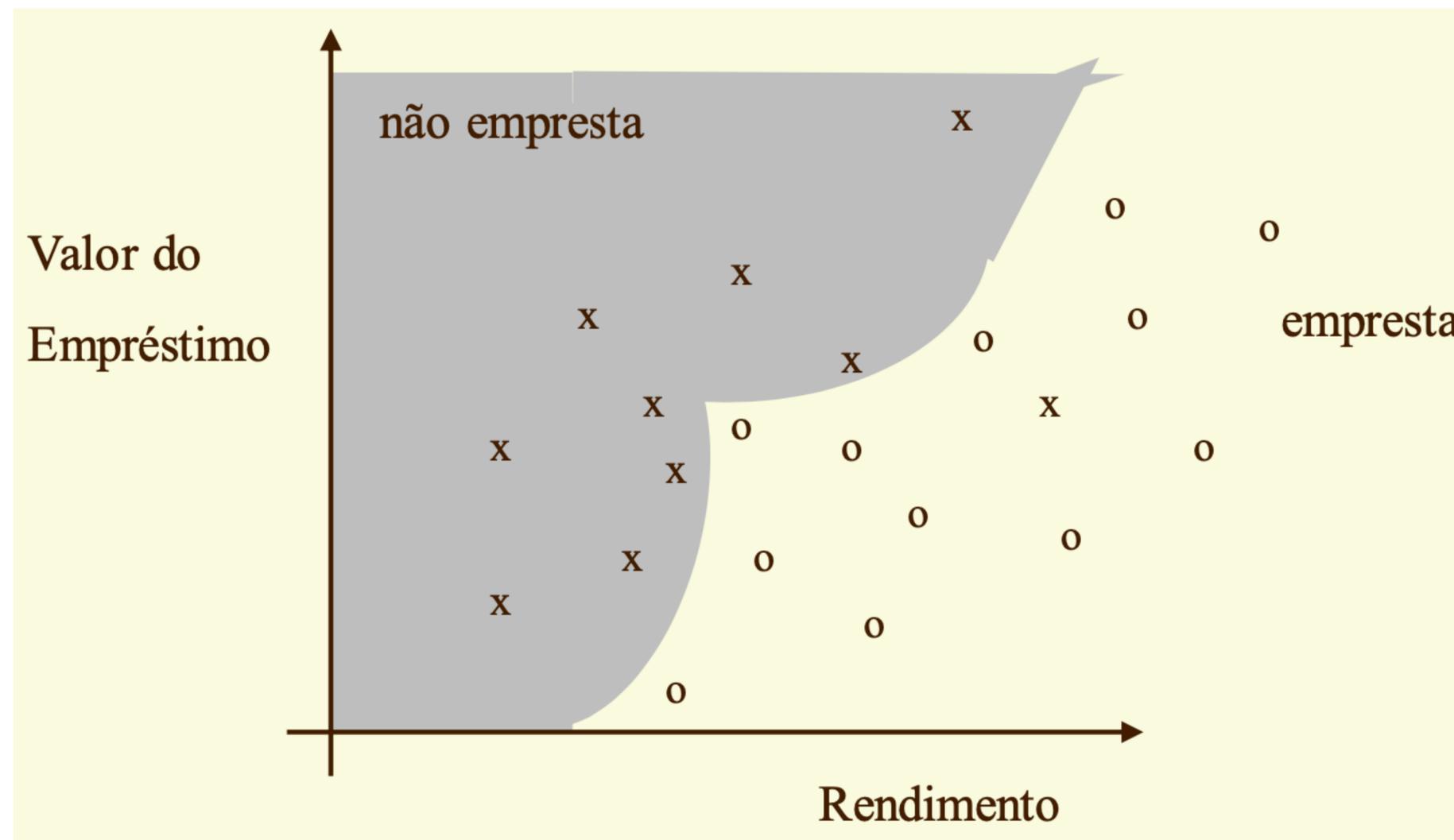
Motivação

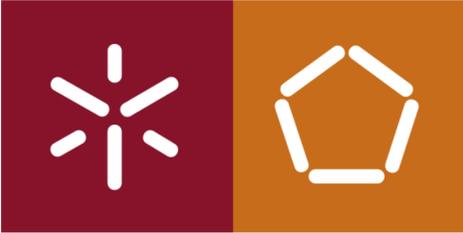
Data Warehouse

Data Mining

Seleção
Pré-Processamento
Transformação
Mineração

Exemplo dos limites duma classificação utilizando classificadores não lineares (e.g. redes neurais artificiais):





Extração de Conhecimento

Motivação

Data Warehouse

Data Mining

Seleção

Pré-Processamento

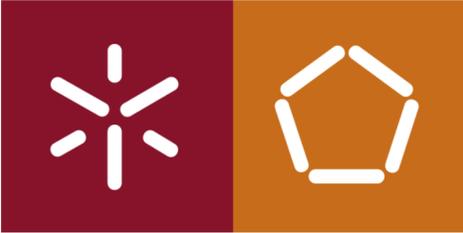
Transformação

Mineração (I)

Mineração (II)

Escolha do(s) algoritmo(s) de mineração de dados: selecção do(s) método(s) a serem usados na procura de padrões nos dados, incluindo a decisão de qual o modelo e parâmetros mais apropriados.

- Árvores e Regras de Decisão (métodos indutivos)
- Regressão Não Linear e Métodos de Classificação (Algoritmos Genéticos Evolutivos e Redes Neurais Artificiais)
- Métodos Baseados em Exemplos
- Modelos Gráficos Probabilísticos de Dependências
- Modelos de Aprendizagem Relacional (programação lógica indutiva)



Extração de Conhecimento

Motivação

Data Warehouse

Data Mining

Seleção

Pré-Processamento

Transformação

Mineração (I)

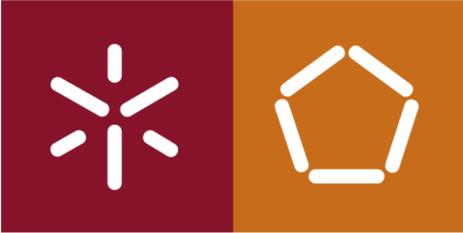
Mineração (II)

Interpretação

Interpretação dos padrões minerados com o possível regresso a uma das fases anteriores para mais interacção ou documentação.

Consolidação do conhecimento descoberto:

- incorporação deste conhecimento no sistema ou elaboração de
- relatórios para as partes interessadas.
Verificação e resolução de potenciais conflitos com conhecimento
- tido com verdadeiro (ou previamente extraído)



Extração de Conhecimento

Motivação

Data Warehouse

Data Mining

Seleção

Pré-Processamento

Transformação

Mineração (I)

Mineração (II)

Interpretação

Problemas

Problemas na Mineração de Dados

Informação limitada

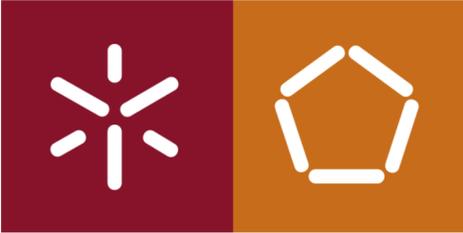
- Informação incompleta
- Dados dispersos
- Espaço de testes

Dados corrompidos

- Ruído
- Dados omissos

Bases de Dados

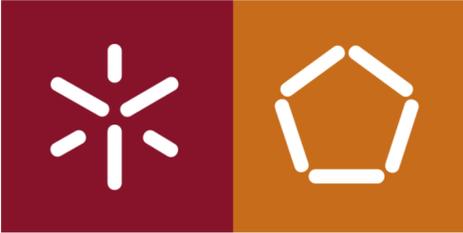
- Tamanho
- Voláteis



Extração de Conhecimento

Problema

- **“We are drowning in data, but starving for knowledge”**
Data Mining: Concepts and Techniques
Jiawei Han, Micheline Kamber
- **“Necessity is the mother of invention”**
Idem
- **“A necessidade aguça o engenho”**
PT-version



Extração de Conhecimento

Soluções

Business Intelligence

- Data Warehousing Técnicas de OLAP (On-Line Analytical Processing)
 - *Processamento On-line Analítico*, é uma abordagem de fornecimento de respostas rápidas para consultas analíticas de natureza multidimensional. O OLAP faz parte de uma categoria mais abrangente, o **Business Intelligence**, que também inclui **ETL** (Extract, Transform, Load – *Extração, Transformação e Carga*), geração de relatórios relacionais e
- Data Mining/Extração de Conhecimento
 - **Descobrir**, em grandes volumes de dados, sem qualquer formulação prévia de hipóteses, **informações genéricas, relevantes e desconhecidas**, por forma a serem utilizadas em processos de tomada de decisão.