# PL02 - Introdução à metodologia CRISP-DM

AEC - Licenciatura em Engenharia Biomédica



Hugo Peixoto - hpeixoto@di.uminho.pt

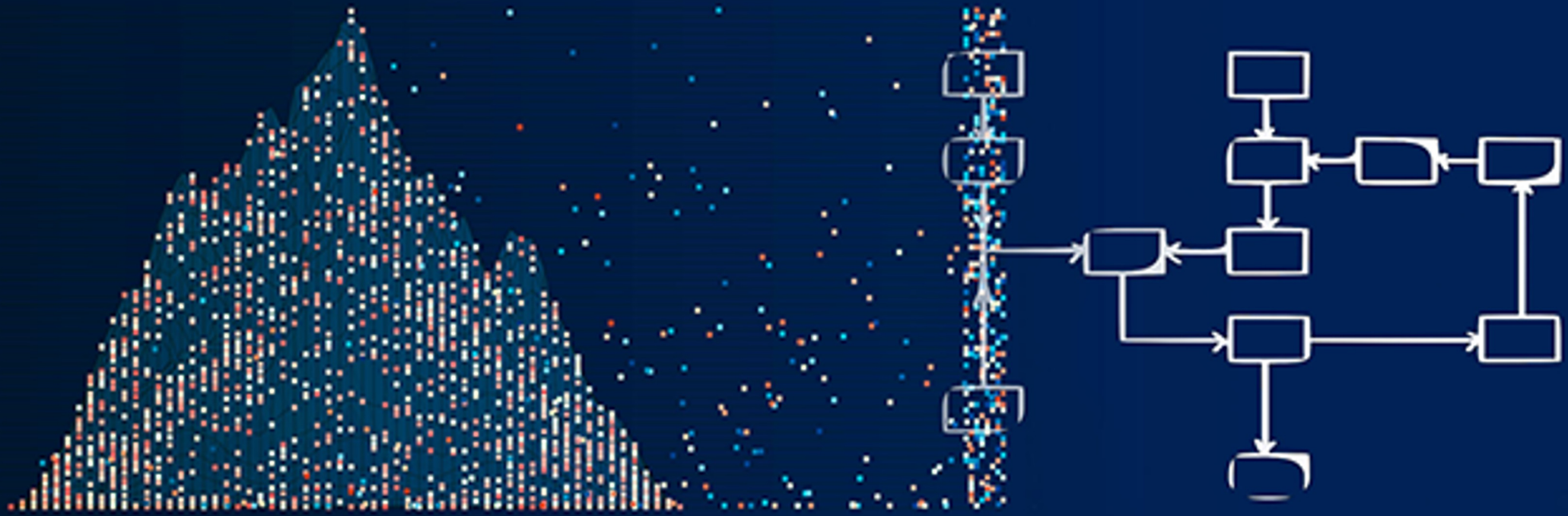
# Plano de Aula - PL01

- Data Mining
- CRISP-DM
- Ficha Exercícios (fe01)


Hugo Peixoto - hpeixoto@di.uminho.pt

# Data Mining

# Data Mining



Extracção de padrões ou conhecimentos de interesse (não triviais, implícitos, anteriormente desconhecidos e potencialmente úteis) de uma enorme quantidade de dados.


Hugo Peixoto - hpeixoto@di.uminho.pt

# Data Mining

**“We are drowning in data, but starving for knowledge!”**

**Aplicações geram enormes quantidades de dados:**

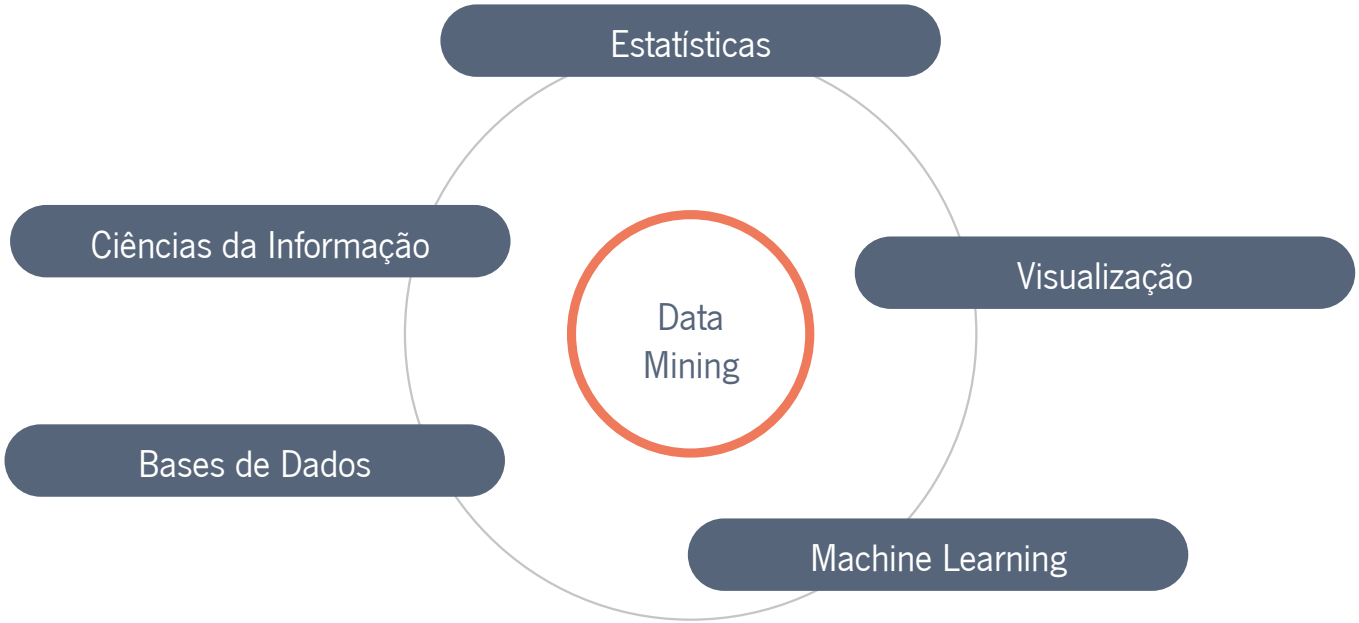
internet, Sistemas de Informação, Experiências Laboratoriais, Transações, Simulações, Dados Clínicos, Dados Bancários, Dados Pessoais, Sistemas de CCTV, Dispositivos de IoT

**Tecnologias capazes de recolher e armazenar os dados:**

Satélites, Camaras, Scanners, Wearables, Dispositivos Móveis ...

Cloud, Bases de Dados, Data Warehouses ....

# Data Mining


Estatísticas
Ciências da Informação
Visualização
Data Mining
Bases de Dados
Machine Learning

# Data Mining

## Técnicas de DM:

### → Regras de Associação:

Carrinhos de compras em supermercados

### → Classificação:

Construir modelos (funções) que descrevem e distinguem classes ou conceitos para previsão futura

### → Agrupamento (Clustering):

A etiqueta da classe é desconhecida: Agrupar dados para formar novas classes, por exemplo, cluster clientes supermercado (zonas - dias) – Maximização da semelhança intra-classe e minimização da semelhança interclasse

# Data Mining

## Técnicas de DM:

### → Análise de Desvios:

Outlier: um objeto de dados que não está de acordo com o comportamento geral dos dados - Ruído ou excepção?

Não! útil na detecção de fraudes, análise de eventos raros

### → Análise de tendências e evolução:

Tendência e desvio: análise de regressão

Mineração de padrões sequenciais, análise de periodicidade

Análise baseada na similaridade


Hugo Peixoto - hpeixoto@di.uminho.pt

# CRISP-DM


Hugo Peixoto - hpeixoto@di.uminho.pt

# CRISP-DM

**CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining

Esforço financiado pela Comunidade Europeia para desenvolver uma metodologia para o processo de Data Mining

## Principais objetivos:

- → Encorajar a utilização de ferramentas interoperáveis ao longo de todo o processo de Data Mining;
- → Retirar conhecimento valioso de tarefas simples de Data Mining.

## Confiável e Repetível

O processo de Data Mining deve ser confiável e repetível por pessoas com pouco conhecimento em DM!!

# CRISP-DM

## Diretrizes

CRISP-DM é uma metodologia uniforme com diretrizes, documentação de experiência

# CRISP-DM

## Flexível

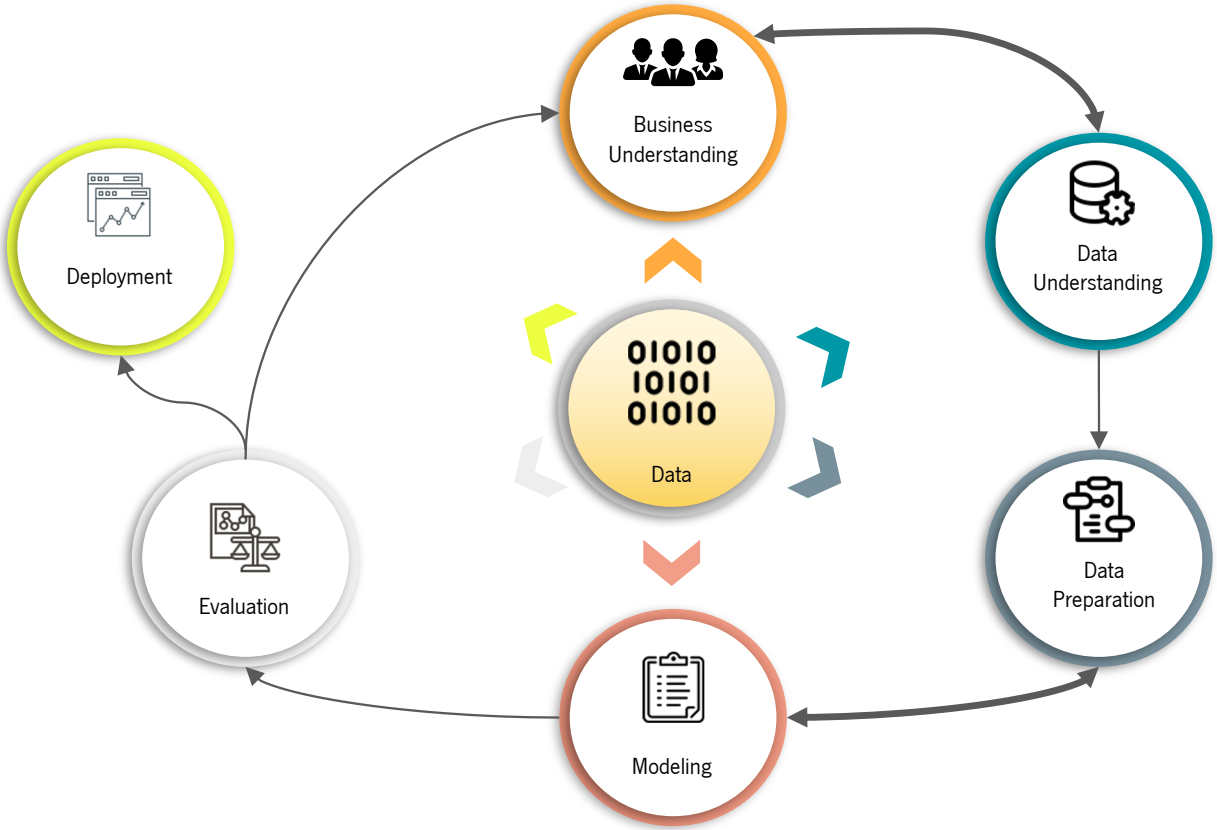
A metodologia CRISP- DM é flexível o suficiente para ter em conta problemas de negócio diferentes e dados diferentes

# CRISP-DM

## Características:

- Metodologia para o registo de experiências;
- Permite que os projetos sejam replicados;
- Ajuda no planeamento e na gestão de projectos;
- "Factor de conforto" para novos utilizadores;
- Demonstra a maturidade da Data Mining.

# CRISP-DM


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data
01010
10101
01010

# CRISP-DM

## Business Understanding (Compreender o Negócio):

- Compreender os **objetivos** e **requisitos** do projeto
- Determinar e consolidar qual o **objetivo** a atingir com o processo de Data Mining

## Data Understanding (Compreender os Dados):

- Recolha, exploração e familiarização com os dados
- Identificar problemas de qualidade nos dados

# CRISP-DM

## Business Understanding (Compreender o Negócio):

- **Determinar objetivos do negócio**
- Documentar o background do negócio
- Definir Sucesso do negócio

- **Determinar objetivos da aplicação de DM**
- Definir o sucesso do processo
- Identificar o critério de sucesso para a aplicação de DM

- **Avaliar a situação atual**
- Avaliar recursos disponíveis
- Identificar constrangimentos
- Documentar custos e benefícios

- **Planeamento**
- Definir o plano de implementação
- Definir técnicas e ferramentas

**O que se pretende atingir com o processo de Data Mining?**

**Qual o critério para o sucesso?**

# CRISP-DM

## Data Understanding (Compreender os Dados):

- **Recolher os dados iniciais**
- Extrair dados das fontes
- Definir quais as fontes de dados
- Definir métodos de recolha

- **Descrição detalhada dos dados**
- Efetuar a avaliação de quantidade
- Detalhar o tipo de atributos
- Definir valor dos atributos para o negócio
- Avaliar estatisticamente: Min,Max,Mean,etc

- **Explorar os dados**
- Produzir um relatório da exploração dos dados
- Analisar em detalhe atributos de interesse
- Definir relações e Agrupamento de atributos

- **Avaliação de qualidade**
- Avaliar a existência de campos nulos
- Gerir a inexistência de atributos importantes
- Avaliar valor <-> significado do atributo

**Recolher e organizar os dados que serão analisados!**

**Garantir a qualidade e a compreensão dos dados disponíveis!**

# CRISP-DM

## Data Preparation (Preparar os Dados):

- Seleção de dados (critérios de inclusão/exclusão)
- Seleção e Criação de atributos – Limpeza de dados

## Modeling (Criar os Modelos):

- Escolher os modelos de Data Mining
- Construção e avaliação dos modelos

# CRISP-DM

## Data Preparation (Preparação dos Dados):

- **Construir dados**
- Derivar atributos
- Validar a reconstrução de campos nulos
- Integrar dados obtidos de outras fontes

- **Integração de outras fontes**
- Integrar atributos resultantes de outras fontes

- **Limpar os dados**
- Efetuar a limpeza de dados desnecessários (identificadores)

- **Formatar os dados**
- Organizar atributos dentro do dataset
- Garantir coerência para a criação de datasets precisos

**Garantir que os dados estão prontos para serem adicionados aos modelos de ML.**

# CRISP-DM

## Modeling (Criar os Modelos):

- **Selecionar a técnica**
  Seleção dos modelos a usar
  Interpretar a técnica em conjunto com as conclusões retiradas sobre os dados
- **Modelo de Teste**
  Definir qual o modelo de teste a utilizar
  Divisão do dataset em dados de treino e teste
- **Construção do modelo**
  Definição de parâmetros iniciais
  Descrever o modelo e a interpretação para o seu uso
- **Avaliação do modelo**
  Interpretação inicial dos resultados
  Comparar com potenciais expectativas
  Comparar com o conhecimento existente

**Implementação de modelos de DM que apliquem as técnicas selecionadas (algoritmos e testes) a datasets definidos.**

# CRISP-DM

## Evaluation (Avaliação dos Modelos e Resultados):

- Avaliar os resultados, i.e, determinar se os resultados cumprem os objetivos iniciais
- Rever o processo

## Deployment (Implementação):

- Colocar os modelos finais em prática
- Monitorização e manutenção dos modelos

# CRISP-DM

## Evaluation (Avaliação):

- **Decisão**
  - Avaliação detalhada dos resultados
  - Redefinição dos modelos de DM
  - Comparação com os critérios de sucesso
  - Avaliação de potenciais correções nos dados
- **Próximos passos**
  - Definição de parâmetros iniciais
  - Descrever o modelo e a interpretação para o seu uso

**Revisão crítica de resultados. Os critérios para o sucesso definidos foram atingidos?**

# CRISP-DM

## Deployment (Implementação):

- **Plano de implementação**
- Descrever o plano de implementação do conhecimento gerado
- Escrever o relatório final e as visualizações do projeto

- **Revisão e Manutenção**
- Rever periodicamente o processo implementado
- Avaliar potenciais pontos de melhoria e constrangimentos

**Colocar em prática o conhecimento obtido. Rever, avaliar e monitorizar o processo de implementação.**


Hugo Peixoto - hpeixoto@di.uminho.pt

TOMORROW NEEDS ENGINEERING

# Ficha de Exercícios 01


Hugo Peixoto - hpeixoto@di.uminho.pt

# PL02 - Introdução à metodologia CRISP-DM

AEC - Licenciatura em Engenharia Biomédica


Hugo Peixoto - hpeixoto@di.uminho.pt