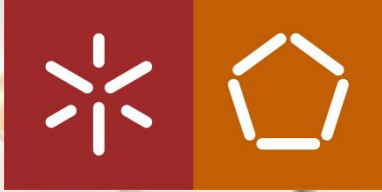




Universidade do Minho
Escola de Engenharia
Departamento de Informática

Metodologias para Extração de Conhecimento

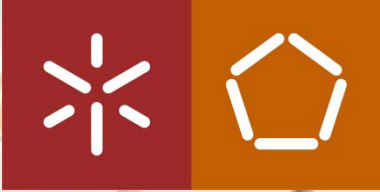
Cesar Analide, Paulo Novais, José Neves



Metodologias para Extração de Conhecimento

O que são metodologias para extração de conhecimento?

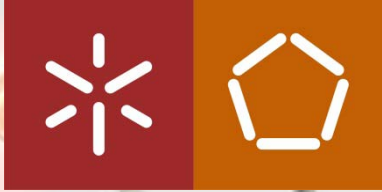
- Uma **Metodologia para a Extração de Conhecimento** (Data Mining) **descreve** e cria **um conjunto de passos** pelos quais deverá passar o desenvolvimento de um **Sistema de Extração de Conhecimento** para a resolução de problemas.



Metodologias para Extração de Conhecimento

Porquê?

- Enquadrar um processo de EC/DM ao abrigo de uma metodologia:
 - o Garante maior robustez;
 - o Facilita a sua compreensão, implementação e desenvolvimento;
 - o Permite a replicação de processos;
 - o Auxilia no planeamento e na gestão do projeto;
 - o Confere “maturidade” ao processo de EC/DM;
 - o Encoraja a adoção de melhores práticas.



Metodologias para Extração de Conhecimento

Que metodologias?

- CRISP-DM

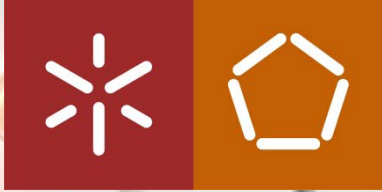
- o **CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining
(Daimler Chrysler, SPSS, NCR)

- SEMMA

- o **S**ample, **E**xplore, **M**odify, **M**odel and **A**ssess
(SAS Institute Inc.)

- PMML

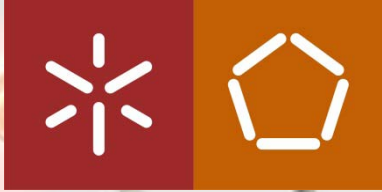
- o **P**redictive **M**odel **M**arkup **L**anguage
(Angoss Software, Magnify, Univ. Illinois, NCR, SPSS)



Metodologias para Extração de Conhecimento

CRISP-DM

- **CR**oss Industry **S**tandard **P**rocess for **D**ata **M**ining
(Daimler Chrysler, SPSS, NCR)
- Objetivos:
 - o Definir um processo de EC para a indústria;
 - o Construir e disponibilizar ferramentas de apoio;
 - o Assegurar a qualidade dos projetos de EC;
 - o Reduzir os conhecimentos específicos sobre EC necessários para conduzir um processo de EC.

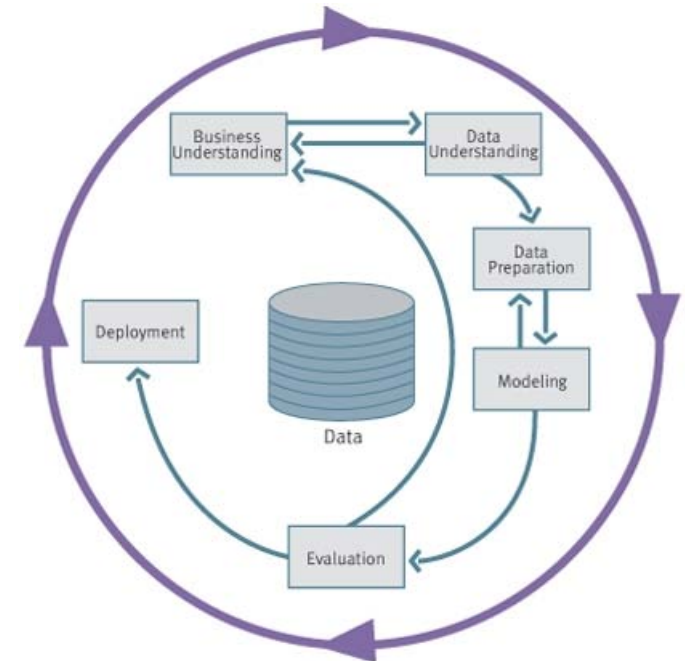


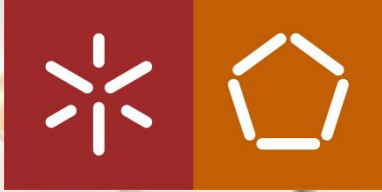
Metodologias para Extração de Conhecimento

CRISP-DM

Ciclo de vida

- O CRISP-DM é um modelo de processos com vista a definir um “guião” para o desenvolvimento de projetos de EC, que se desenrola em 6 etapas:
 - o Estudo do negócio;
 - o Estudo dos dados;
 - o Preparação dos dados;
 - o Modelação;
 - o Avaliação;
 - o Desenvolvimento.



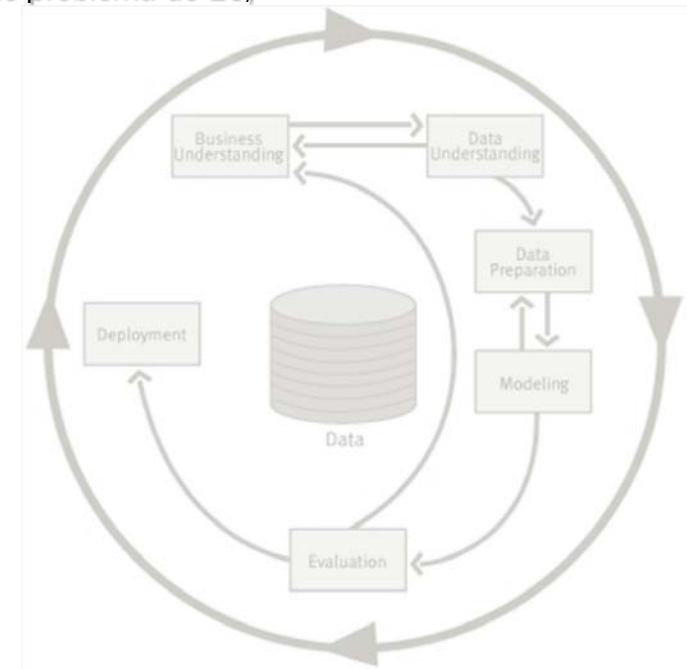


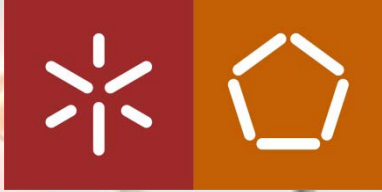
Metodologias para Extração de Conhecimento

CRISP-DM

Ciclo de vida

- Business Understanding/Estudo do Negócio:
 - o Compreensão dos objetivos do projeto e definição do problema de EC;



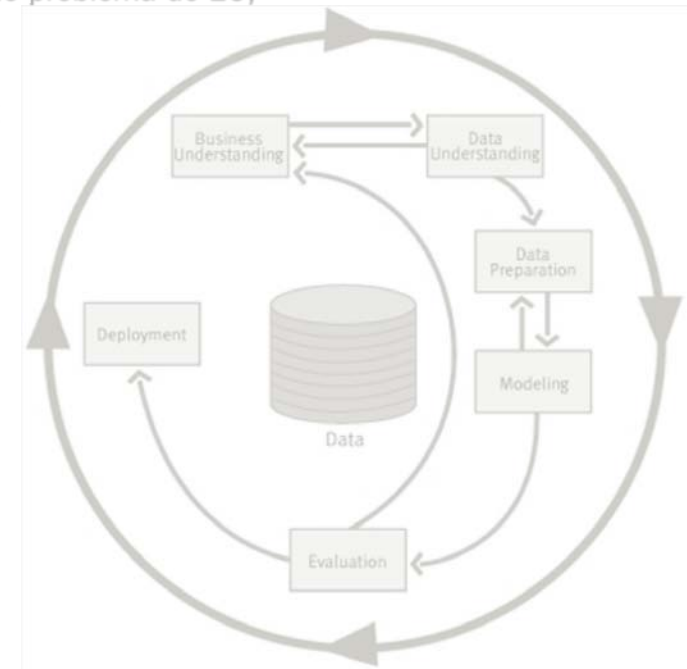


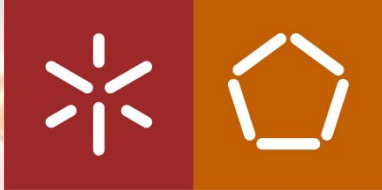
Metodologias para Extração de Conhecimento

CRISP-DM

Ciclo de vida

- Business Understanding/Estudo do Negócio:
 - Compreensão dos objetivos do projeto e definição do problema de EC;
- Data Understanding/Estudos dos Dados:
 - Obter os dados e identificar a qualidade dos dados;



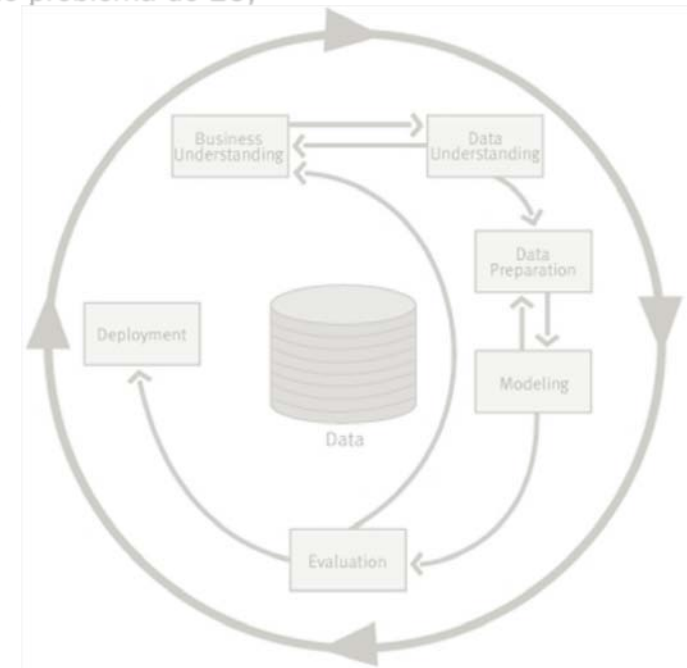


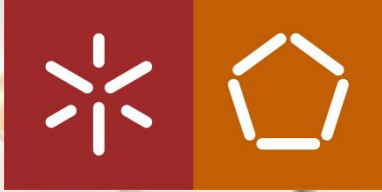
Metodologias para Extração de Conhecimento

CRISP-DM

Ciclo de vida

- Business Understanding/Estudo do Negócio:
 - Compreensão dos objetivos do projeto e definição do problema de EC;
- Data Understanding/Estudos dos Dados:
 - Obter os dados e identificar a qualidade dos dados;
- Data Preparation/Preparação dos Dados:
 - Seleção de atributos e limpeza dos dados;



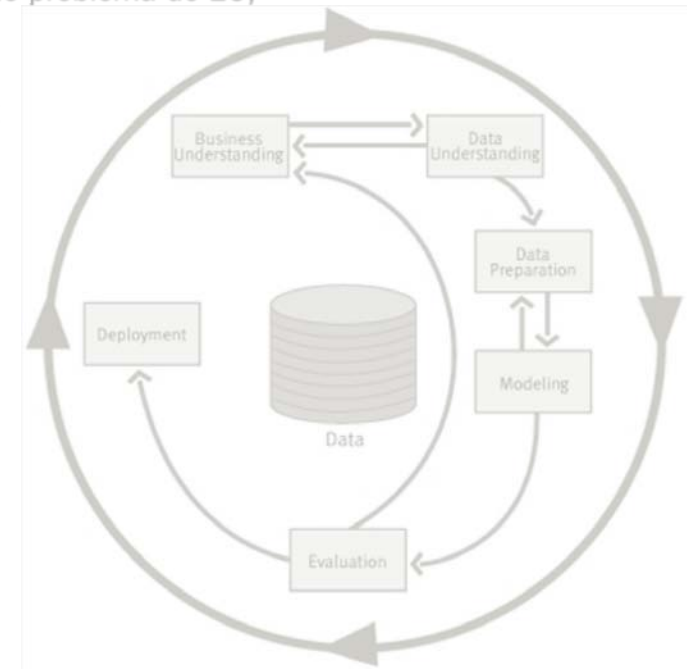


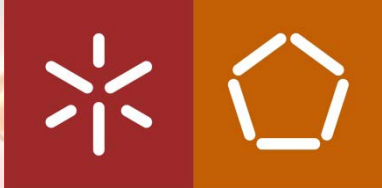
Metodologias para Extração de Conhecimento

CRISP-DM

Ciclo de vida

- Business Understanding/Estudo do Negócio:
 - Compreensão dos objetivos do projeto e definição do problema de EC;
- Data Understanding/Estudos dos Dados:
 - Obter os dados e identificar a qualidade dos dados;
- Data Preparation/Preparação dos Dados:
 - Seleção de atributos e limpeza dos dados;
- Modeling/Modelação:
 - Experimentação com as ferramentas de EC;



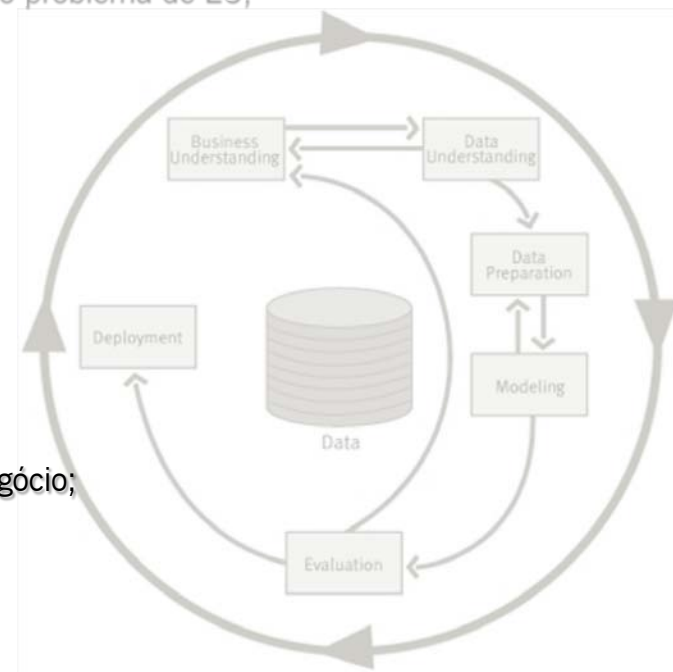


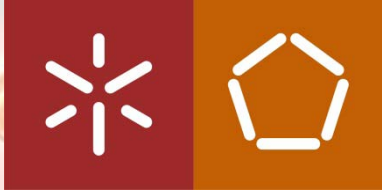
Metodologias para Extração de Conhecimento

CRISP-DM

Ciclo de vida

- Business Understanding/Estudo do Negócio:
 - Compreensão dos objetivos do projeto e definição do problema de EC;
- Data Understanding/Estudos dos Dados:
 - Obter os dados e identificar a qualidade dos dados;
- Data Preparation/Preparação dos Dados:
 - Seleção de atributos e limpeza dos dados;
- Modeling/Modelação:
 - Experimentação com as ferramentas de EC;
- Evaluation/Avaliação:
 - Comparação dos resultados com os objetivos do negócio;



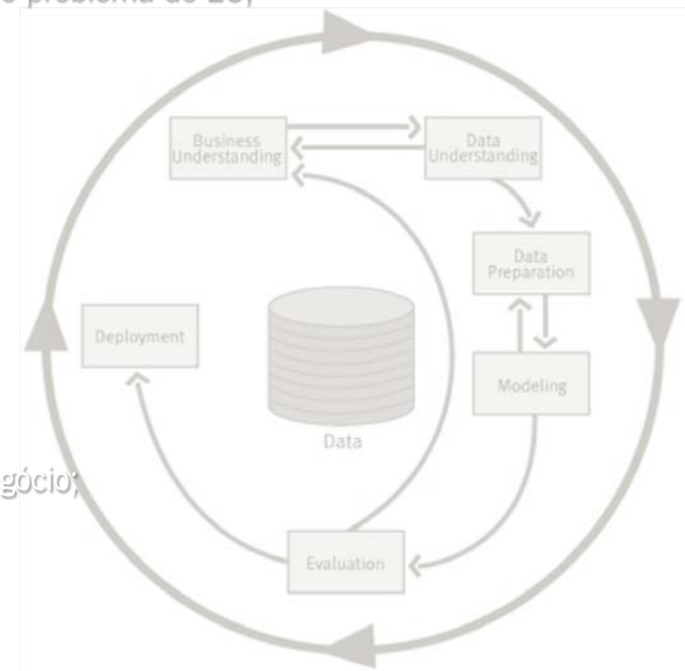


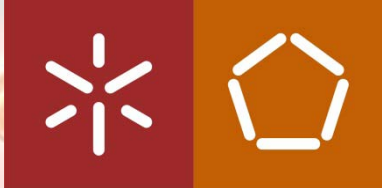
Metodologias para Extração de Conhecimento

CRISP-DM

Ciclo de vida

- Business Understanding/Estudo do Negócio:
 - Compreensão dos objetivos do projeto e definição do problema de EC;
- Data Understanding/Estudos dos Dados:
 - Obter os dados e identificar a qualidade dos dados;
- Data Preparation/Preparação dos Dados:
 - Seleção de atributos e limpeza dos dados;
- Modeling/Modelação:
 - Experimentação com as ferramentas de EC;
- Evaluation/Avaliação:
 - Comparação dos resultados com os objetivos do negócio;
- Deployment/Desenvolvimento:
 - Colocação do modelo em produção.





Metodologias para Extração de Conhecimento

CRISP-DM

Fases e tarefas

**ESTUDO DO
NEGÓCIO**

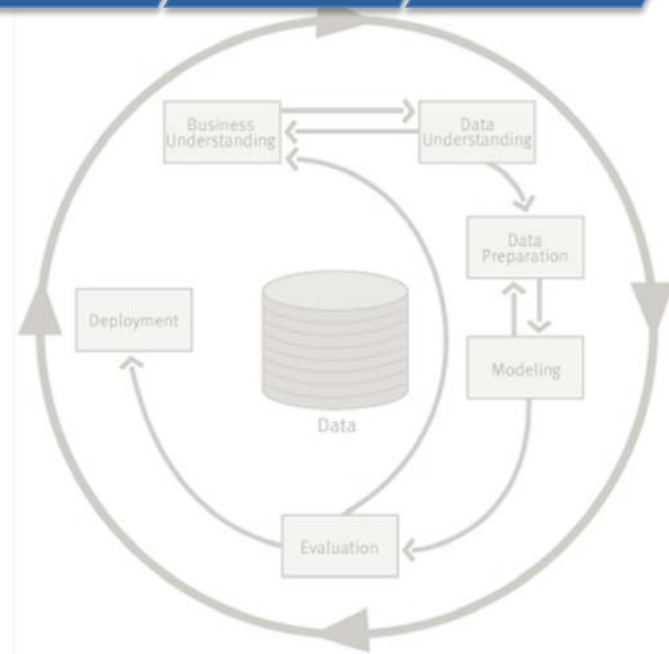
**ESTUDO DOS
DADOS**

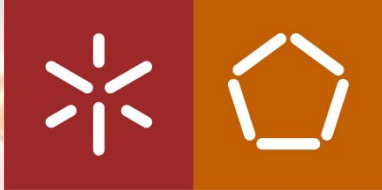
**PREPARAÇÃO
DOS DADOS**

MODELAÇÃO

AValiação

**DESENVOLVI-
MENTO**

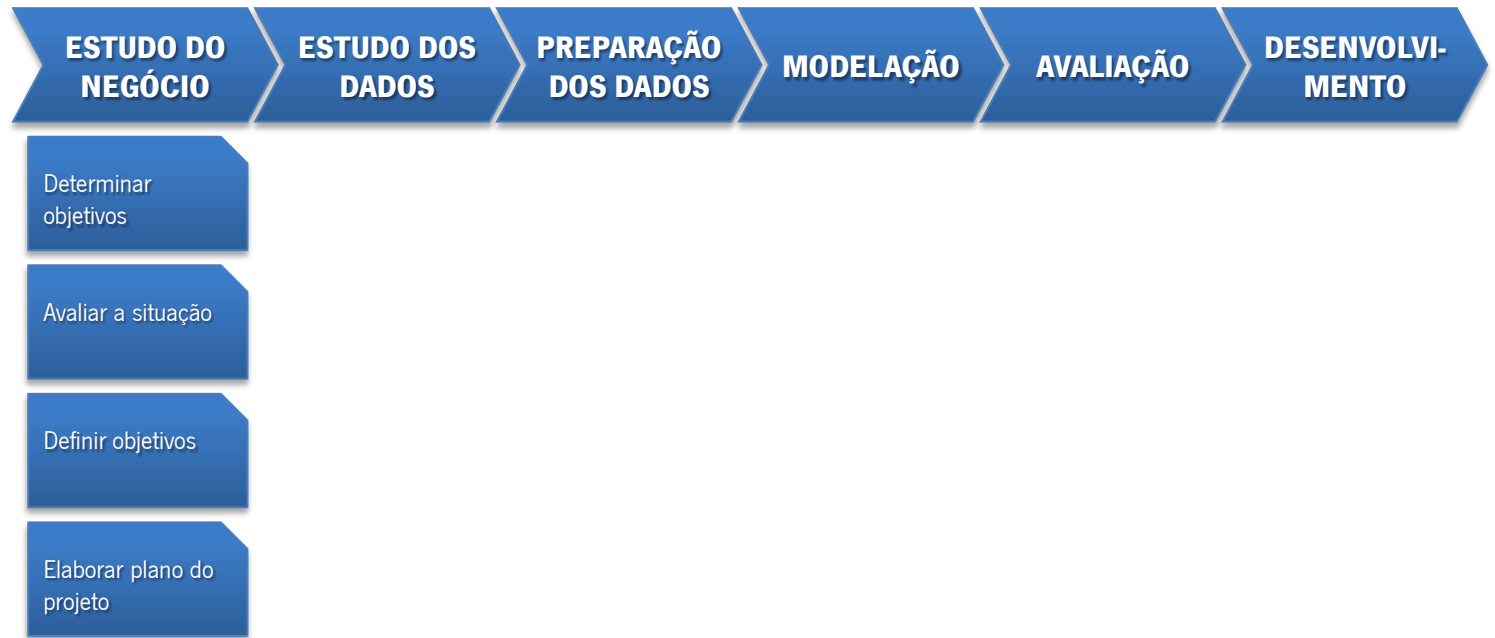


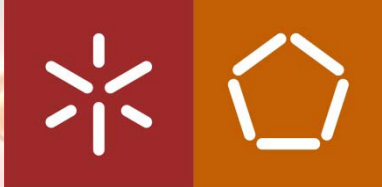


Metodologias para Extração de Conhecimento

CRISP-DM

Fases e tarefas





Metodologias para Extração de Conhecimento

CRISP-DM

Fases e tarefas

ESTUDO DO
NEGÓCIO

ESTUDO DOS
DADOS

PREPARAÇÃO
DOS DADOS

MODELAÇÃO

AValiação

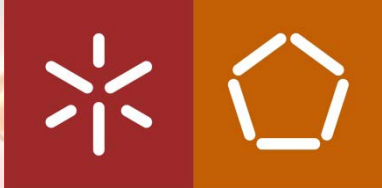
DESENVOLVI-
MENTO

Recolha inicial de
dados

Descrição dos
dados

Exploração dos
dados

Qualidade do
dados



Metodologias para Extração de Conhecimento

CRISP-DM

Fases e tarefas

ESTUDO DO
NEGÓCIO

ESTUDO DOS
DADOS

PREPARAÇÃO
DOS DADOS

MODELAÇÃO

AValiação

DESENVOLVI-
MENTO

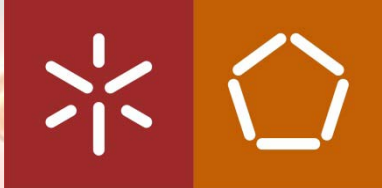
Seleção de dados

Limpeza de dados

Construção de
dados

Integração de
dados

Formatação de
dados



Metodologias para Extração de Conhecimento

CRISP-DM

Fases e tarefas

ESTUDO DO
NEGÓCIO

ESTUDO DOS
DADOS

PREPARAÇÃO
DOS DADOS

MODELAÇÃO

AVALIAÇÃO

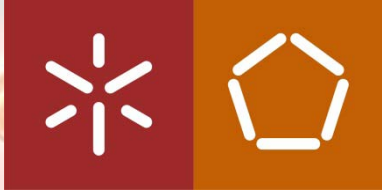
DESENVOLVI-
MENTO

Seleção de técnicas de modelação

Geração de testes

Construção de modelo

Avaliação do modelo



Metodologias para Extração de Conhecimento

CRISP-DM

Fases e tarefas

**ESTUDO DO
NEGÓCIO**

**ESTUDO DOS
DADOS**

**PREPARAÇÃO
DOS DADOS**

MODELAÇÃO

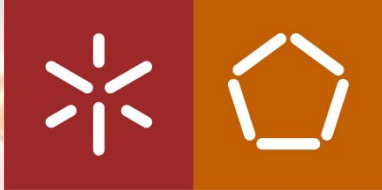
AVALIAÇÃO

**DESENVOLVI-
MENTO**

Avaliação de
resultados

Revisão do
processo

Próximos passos



Metodologias para Extração de Conhecimento

CRISP-DM

Fases e tarefas

ESTUDO DO
NEGÓCIO

ESTUDO DOS
DADOS

PREPARAÇÃO
DOS DADOS

MODELAÇÃO

AVALIAÇÃO

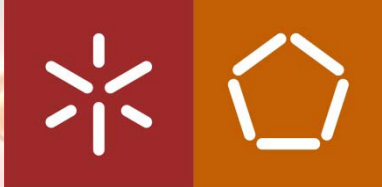
DESENVOLVI-
MENTO

Planeamento da
implementação

Planeamento de
monitorização e ...

Produção de
relatório final

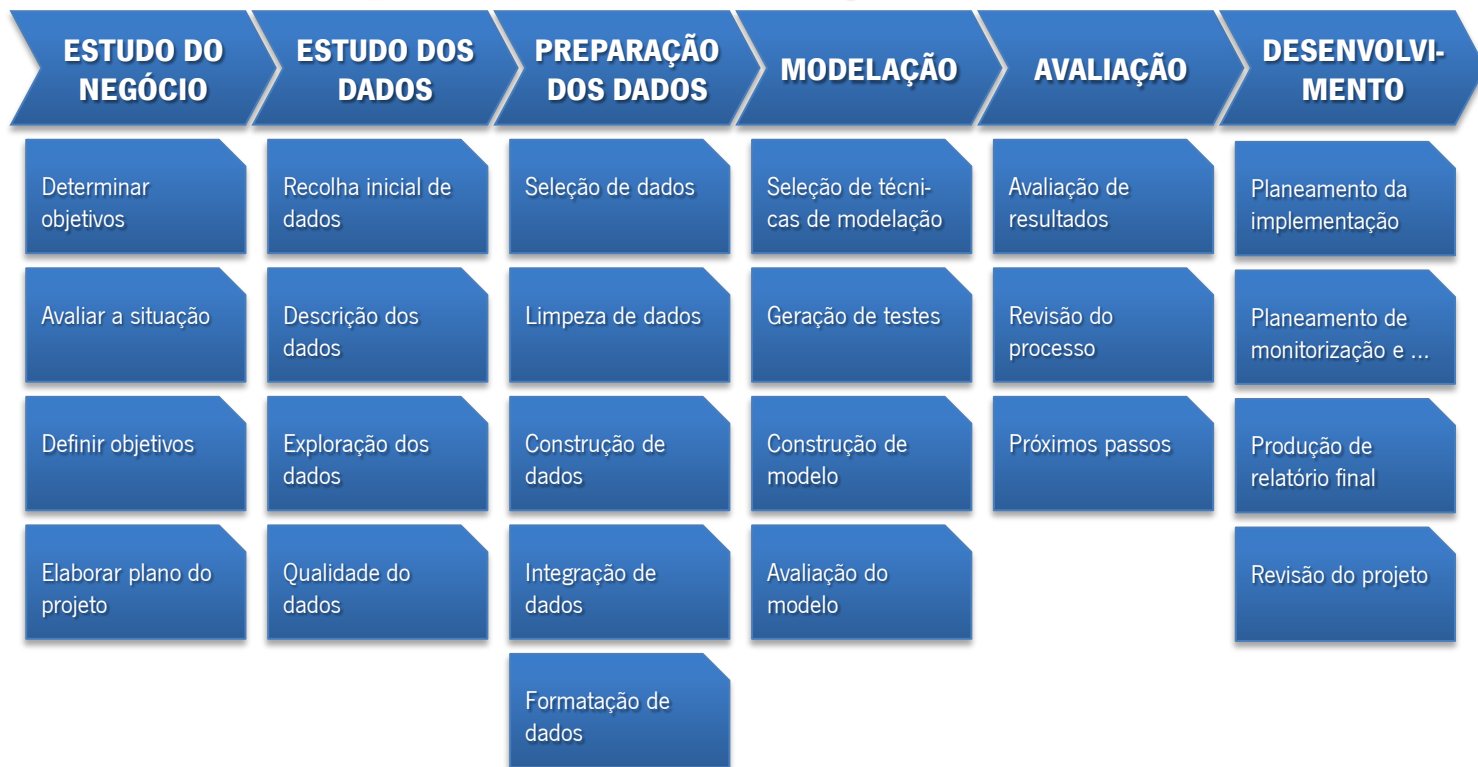
Revisão do projeto

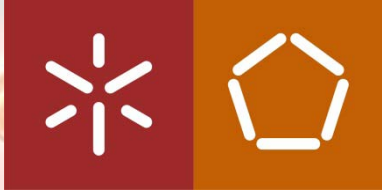


Metodologias para Extração de Conhecimento

CRISP-DM

Fases e tarefas

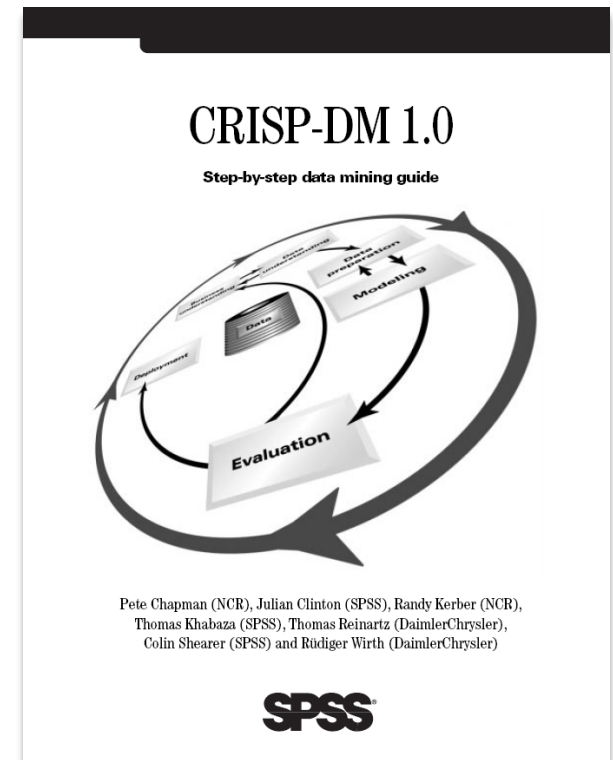


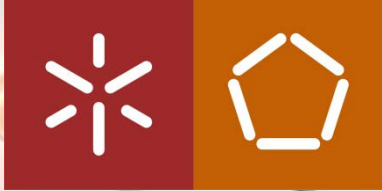


Metodologias para Extração de Conhecimento

Metodologia CRISP-DM

- “CRISP-DM 1.0: Step-by-step data mining guide”, Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, Rüdiger Wirth
- [CRISP-DM](#)

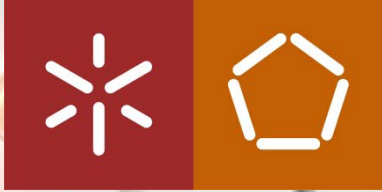




Metodologias para Extração de Conhecimento

SEMMA

- **S**ample, **E**xplore, **M**odify, **M**odel and **A**ssess;
- Produto de Data Mining desenvolvido pelo SAS Institute Inc.;
- Definição SAS:
 - o “Data Mining é o processo de **extrair conhecimento e relações complexas** de grandes volumes de dados.”
- Motivação:
 - o necessidade de definir, padronizar e integrar sistemas ou processos de Data Mining nos ciclos de produção.



Metodologias para Extração de Conhecimento

O processo SEMMA ■ Divide o processo de Data Mining em 5 etapas:

o Sample/Amostragem:

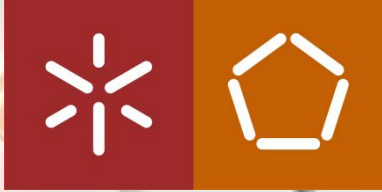
- Extração de dados do universo do problema;
- Baseia o processo de Data Mining no conceito de “amostra” do problema;
- Amostra pequena e significativa;
- Proporciona flexibilidade e rapidez no tratamento dos dados.

o Explore/Exploração;

o Modify/Modificação;

o Model/Modelação;

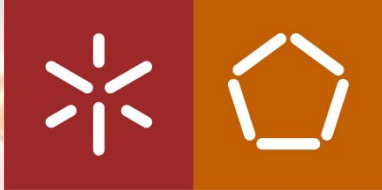
o Assess/Avaliação.



Metodologias para Extração de Conhecimento

O processo SEMMA ■ Divide o processo de Data Mining em 5 etapas:

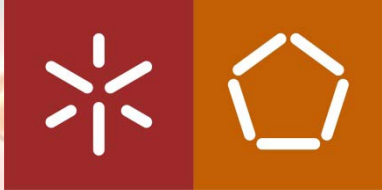
- o Sample/Amostragem;
- o Explore/Exploração:
 - Exploração visual e/ou numérica das tendências;
 - Refinamento do processo de descoberta (mining);
 - Técnicas estatísticas: regressão linear, mínimos quadrados, distribuição de Poisson, etc.;
 - Procura de tendências imprevistas nos dados;
- o Modify/Modificação;
- o Model/Modelação;
- o Assess/Avaliação.



Metodologias para Extração de Conhecimento

O processo SEMMA ■ Divide o processo de Data Mining em 5 etapas:

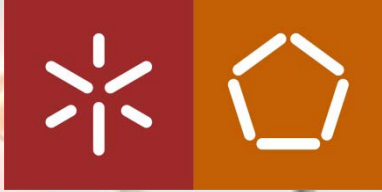
- o Sample/Amostragem;
- o Explore/Exploração;
- o Modify/Modificação:
 - Concentração de todas as modificações necessárias;
 - Inclusão de informação;
 - Seleção ou introdução de novas variáveis;
 - Objetivo: criar, selecionar e adaptar variáveis para a próxima etapa;
- o Model/Modelação;
- o Assess/Avaliação.



Metodologias para Extração de Conhecimento

O processo SEMMA ■ Divide o processo de Data Mining em 5 etapas:

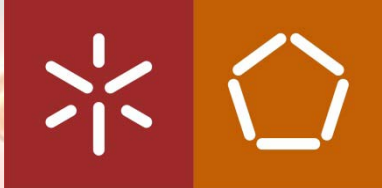
- o Sample/Amostragem;
- o Explore/Exploração;
- o Modify/Modificação;
- o Model/Modelação:
 - Definição das técnicas de construção de modelos de Data Mining: redes neurais artificiais, árvores de decisão, regressão linear, etc.;
 - Dependente do tipo de dados presentes em cada modelo (p.ex., RNA são mais adequadas em problemas nos quais os dados apresentam relacionamentos complexos);
- o Assess/Avaliação.



Metodologias para Extração de Conhecimento

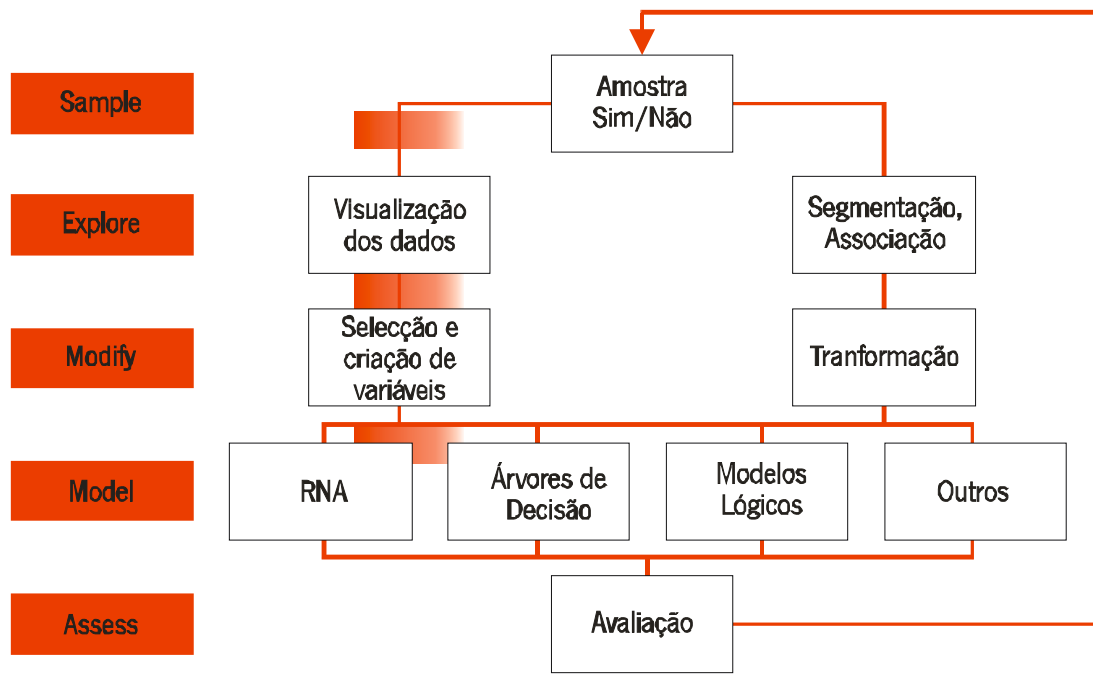
O processo SEMMA ■ Divide o processo de Data Mining em 5 etapas:

- o Sample/Amostragem;
- o Explore/Exploração;
- o Modify/Modificação;
- o Model/Modelação;
- o Assess/Avaliação:
 - Aferição do desempenho do modelo construído para Data Mining;
 - Aplicação do modelo a uma amostra de dados de teste;
 - Procedimento de ajuste do modelo.

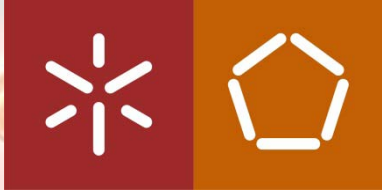


Metodologias para Extração de Conhecimento

O processo SEMMA



in "Data Mining – Descoberta de Conhecimento em Bases de Dados"
Manuel Filipe Santos, Carla Azevedo



Metodologias para Extração de Conhecimento

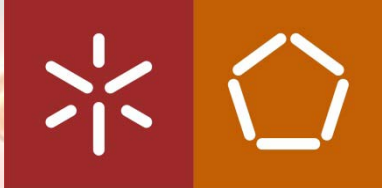
CRISP-DM versus SEMMA

■ Fases CRISP-DM:

- o Estudo do negócio;
- o Estudo dos dados;
- o Preparação dos dados;
- o Modelação;
- o Avaliação;
- o Desenvolvimento.

■ Processo SEMMA:

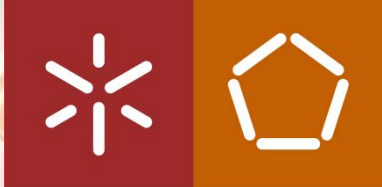
- o Amostragem;
- o Exploração;
- o Modificação;
- o Modelação;
- o Avaliação.



Metodologias para Extração de Conhecimento

PMML

- Desenvolvido por investigadores de Data Mining e várias empresas (NCR, SPSS, etc.);
- A especificação PMML encontra-se em fase de desenvolvimento e consolidação (versão 4.1);
- Utilizada por diversas aplicações (IBM DB2 Data Warehouse Edition v.9.1, SAS Enterprise Miner v.5.1, v.5.2, v.5.3, SPSS Statistics v.17, Clementine v.12);
(<http://www.dmg.org/products.html>)
- Expandir para transformá-la num padrão para o WWW;
- O PMML é uma linguagem para descrever modelos de Data Mining;
- Utiliza XML para representar modelos de DM.

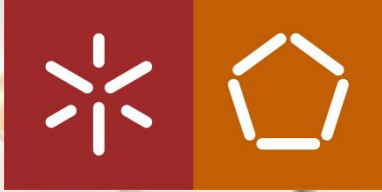


Metodologias para Extração de Conhecimento

PMML: exemplos

```
sepal_length,sepal_width,petal_length,petal_width,class
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
...
5.0,2.0,3.5,1.0,Iris-versicolor
5.9,3.0,4.2,1.5,Iris-versicolor
6.0,2.2,4.0,1.0,Iris-versicolor
6.1,2.9,4.7,1.4,Iris-versicolor
5.6,2.9,3.6,1.3,Iris-versicolor
6.7,3.1,4.4,1.4,Iris-versicolor
5.6,3.0,4.5,1.5,Iris-versicolor
5.8,2.7,4.1,1.0,Iris-versicolor
6.3,2.5,4.9,1.5,Iris-versicolor
6.1,2.8,4.7,1.2,Iris-versicolor
...
6.7,2.5,5.8,1.8,Iris-virginica
7.2,3.6,6.1,2.5,Iris-virginica
6.5,3.2,5.1,2.0,Iris-virginica
6.4,2.7,5.3,1.9,Iris-virginica
6.8,3.0,5.5,2.1,Iris-virginica
5.7,2.5,5.0,2.0,Iris-virginica
5.8,2.8,5.1,2.4,Iris-virginica
6.4,3.2,5.3,2.3,Iris-virginica
...
```

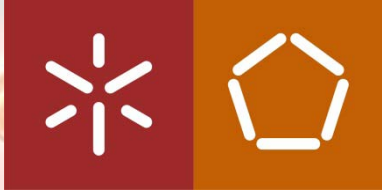
```
<PMML version="2.0">
-
<Header copyright="Copyright (c) 2001, Oracle Corporation. All rights
reserved.">
<Application name="Oracle 9i Data Mining" version="9.2.0"/>
</Header>
-
<DataDictionary numberOfFields="1">
<DataField name="item" optype="categorical"/>
</DataDictionary>
-
<TransformationDictionary>
-
<DerivedField name="PETAL_LENGTH">
-
<Discretize field="PETAL_LENGTH">
-
<DiscretizeBin binValue="1-1.59">
<Interval closure="closedOpen" leftMargin="1.0" rightMargin="1.59"/>
</DiscretizeBin>
-
<DiscretizeBin binValue="1.59-2.18">
<Interval closure="closedOpen" leftMargin="1.59" rightMargin="2.18"/>
</DiscretizeBin>
-
<DiscretizeBin binValue="2.18-2.77">
<Interval closure="closedOpen" leftMargin="2.18" rightMargin="2.77"/>
</DiscretizeBin>
```



Metodologias para Extração de Conhecimento

PMML: objetivos

- Permitir que aplicações utilizem diversas fontes de dados sem se preocuparem com as diferenças entre elas;
- Permitir a utilização combinada e/ou cooperativa de modelos de Data Mining;
- Permitir a administração de modelos de DM baseados em áreas de negócio.



Metodologias para Extração de Conhecimento

Referências bibliográficas

- “CRISP-DM 1.0: Step-by-step data mining guide”, Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, Rüdiger Wirth.
- SAS Enterprise Miner:
www.sas.com/technologies/analytics/datamining/miner/semma.html
- Data Mining Group (DMG):
www.dmg.org
www.dmg.org/faq.html



Intelligent Systems Lab

Contactos

- Universidade do Minho
- Escola de Engenharia
- Departamento de Informática
- <http://islab.di.uminho.pt>
- DI-3.22