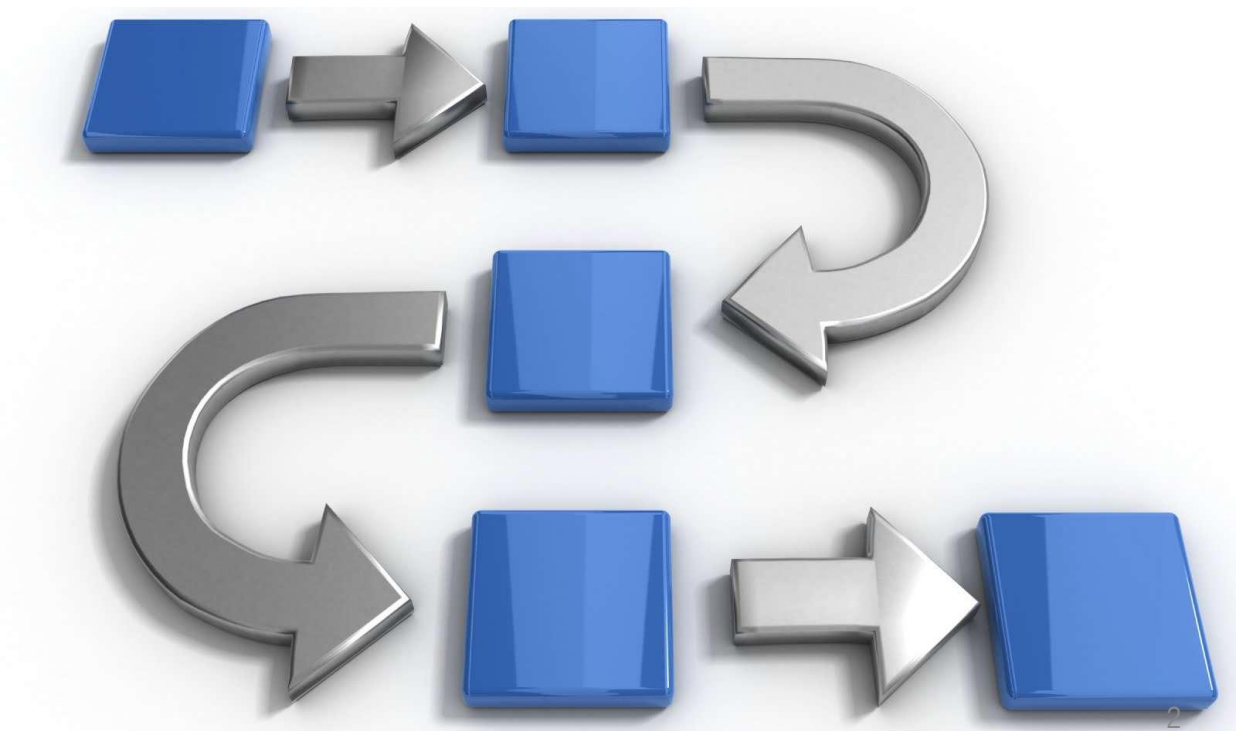
Universidade do Minho
Departamento de Informática

Metodologias de Análise de Dados

LEI/MiEI @ 2024/2025, 2º sem

O que são Metodologias para Análise de Dados?

- Uma **Metodologia para Análise de Dados** descreve e cria **um conjunto de passos** pelos quais deverá passar o desenvolvimento de um **Projeto de Aprendizagem por Máquinas (*Machine Learning*)** para a resolução de problemas.



Porquê?

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

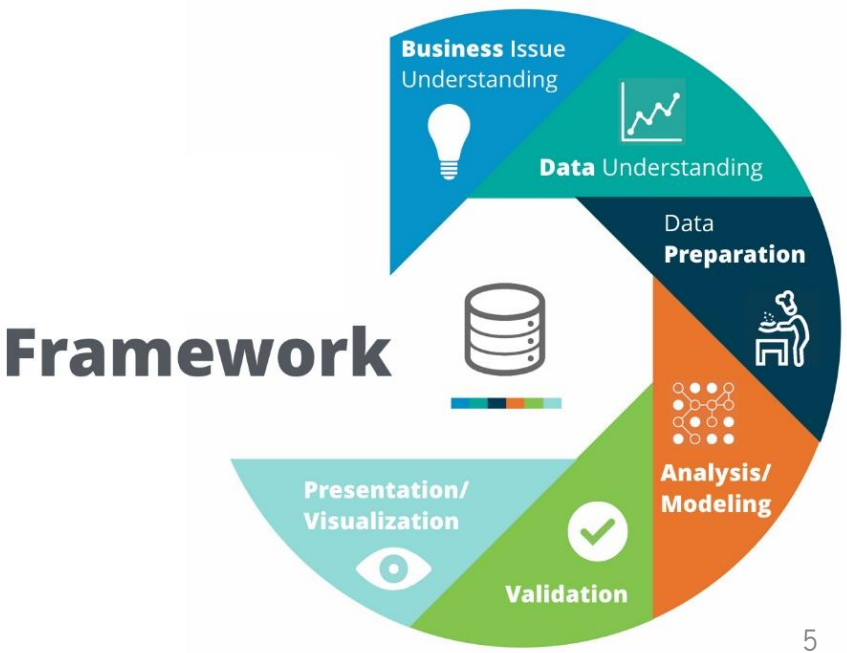
Que metodologias?

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

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

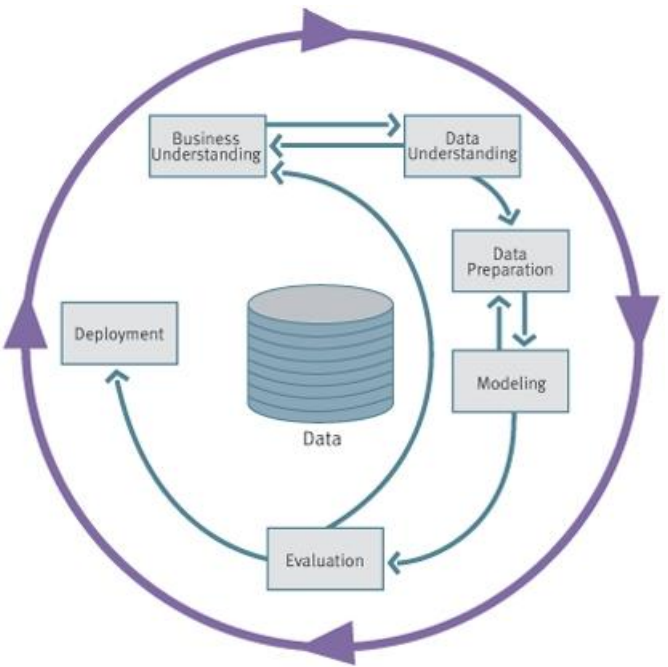
CRISP-DM

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



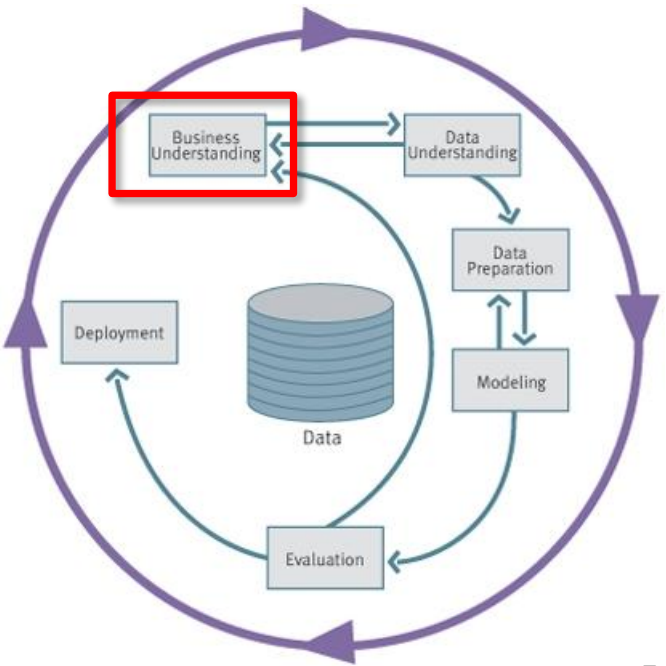
CRISP-DM
Ciclo de vida

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



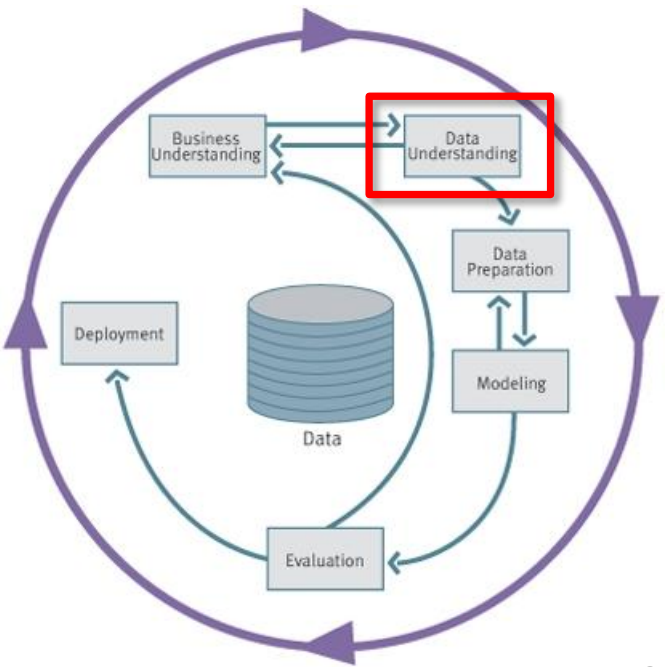
CRISP-DM
Ciclo de vida

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



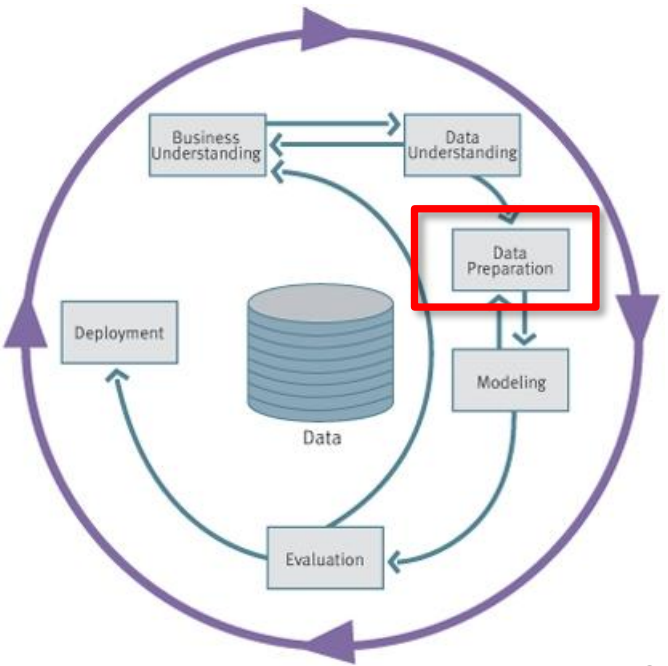
CRISP-DM
Ciclo de vida

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



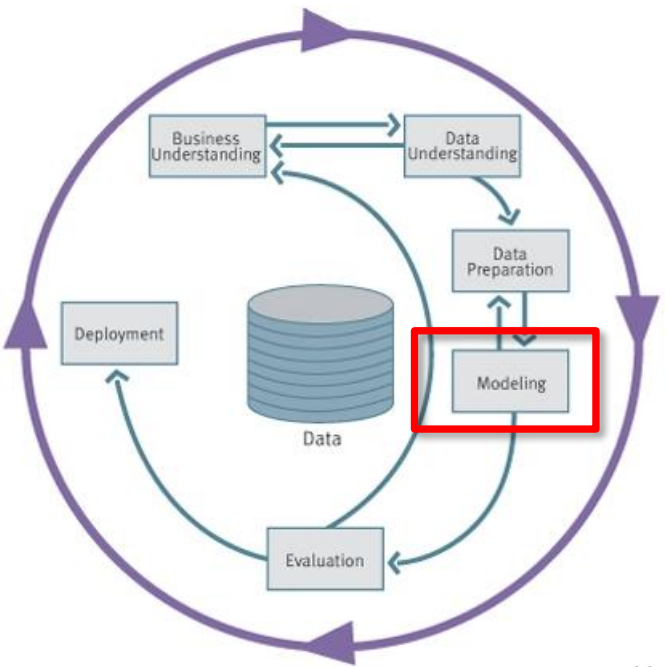
CRISP-DM
Ciclo de vida

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



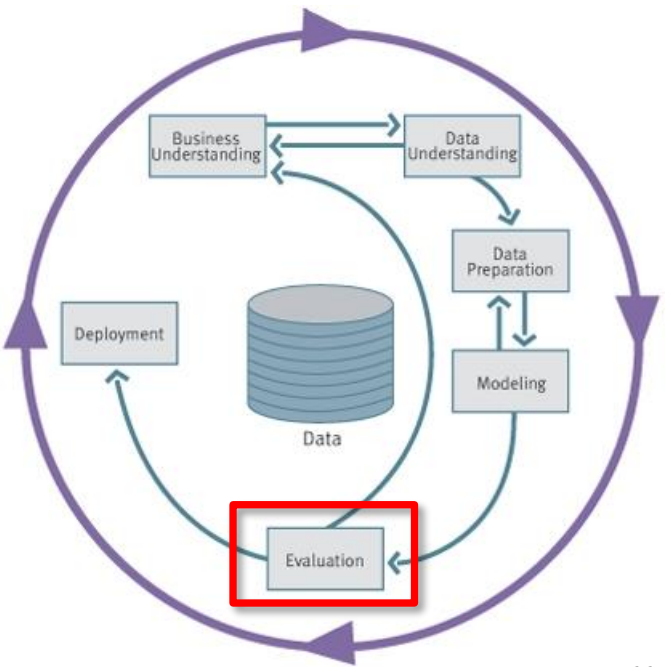
CRISP-DM
Ciclo de vida

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



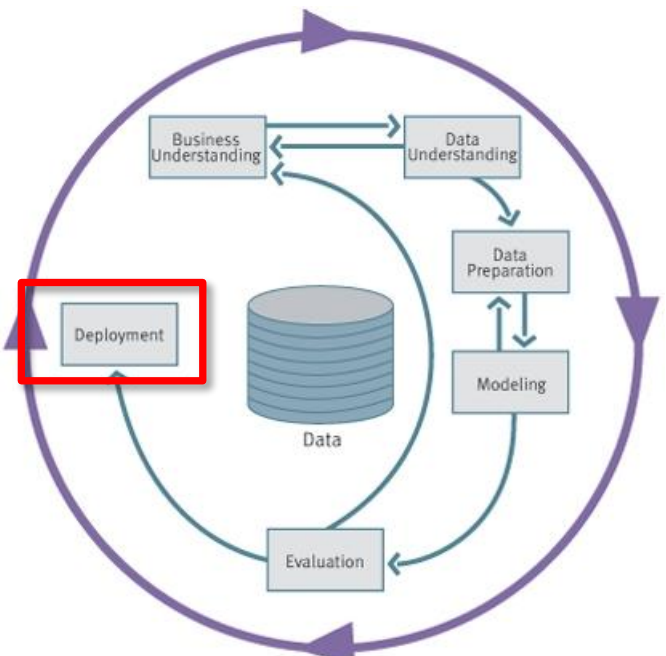
CRISP-DM
Ciclo de vida

- *Business Understanding*/ Estudo do Negócio:
  - Compreensão dos objetivos do projeto e definição do problema de AD;
- *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 AD;
- *Evaluation*/ Avaliação:
  - Comparação dos resultados com os objetivos do negócio;



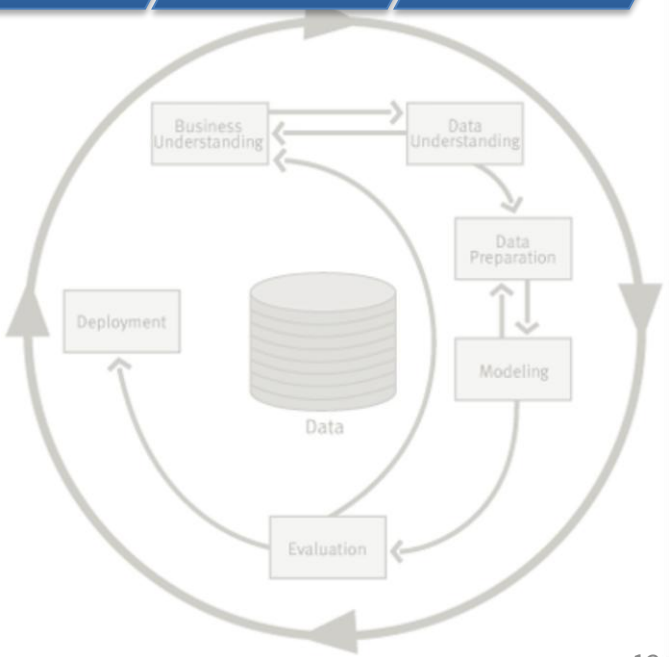
CRISP-DM
Ciclo de vida

- *Business Understanding*/ Estudo do Negócio:
  - Compreensão dos objetivos do projeto e definição do problema de AD;
- *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 AD;
- *Evaluation*/ Avaliação:
  - Comparação dos resultados com os objetivos do negócio;
- *Deployment*/ Desenvolvimento:
  - Colocação do modelo em produção.

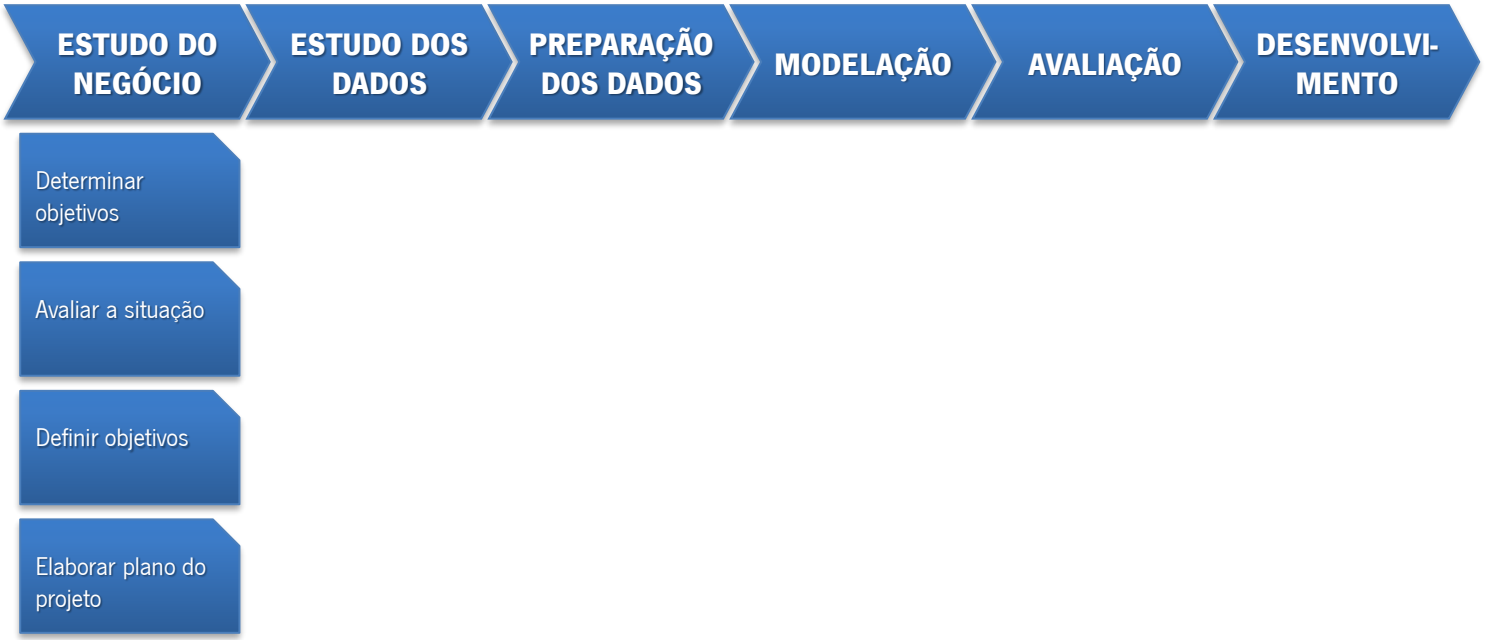


CRISP-DM
Fases e Tarefas

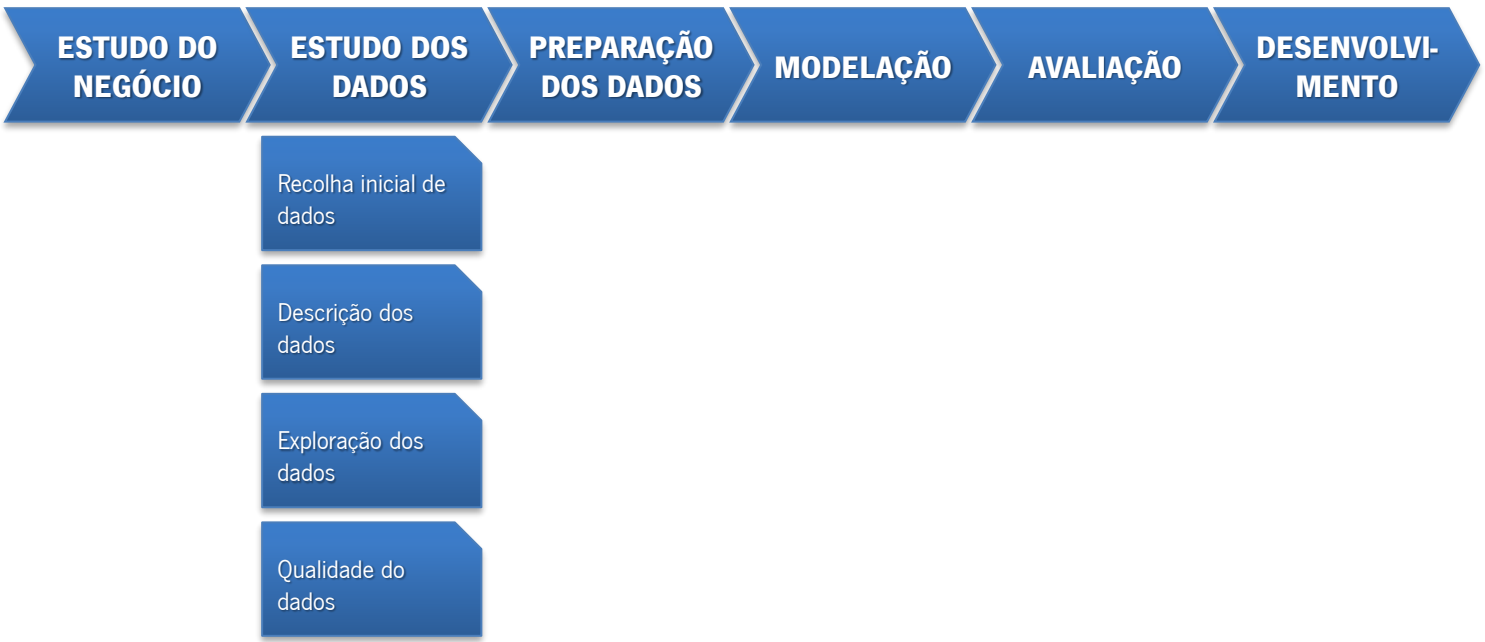
ESTUDO DO NEGÓCIO | ESTUDO DOS DADOS | PREPARAÇÃO DOS DADOS | MODELAÇÃO | AVALIAÇÃO | DESENVOLVI-MENTO



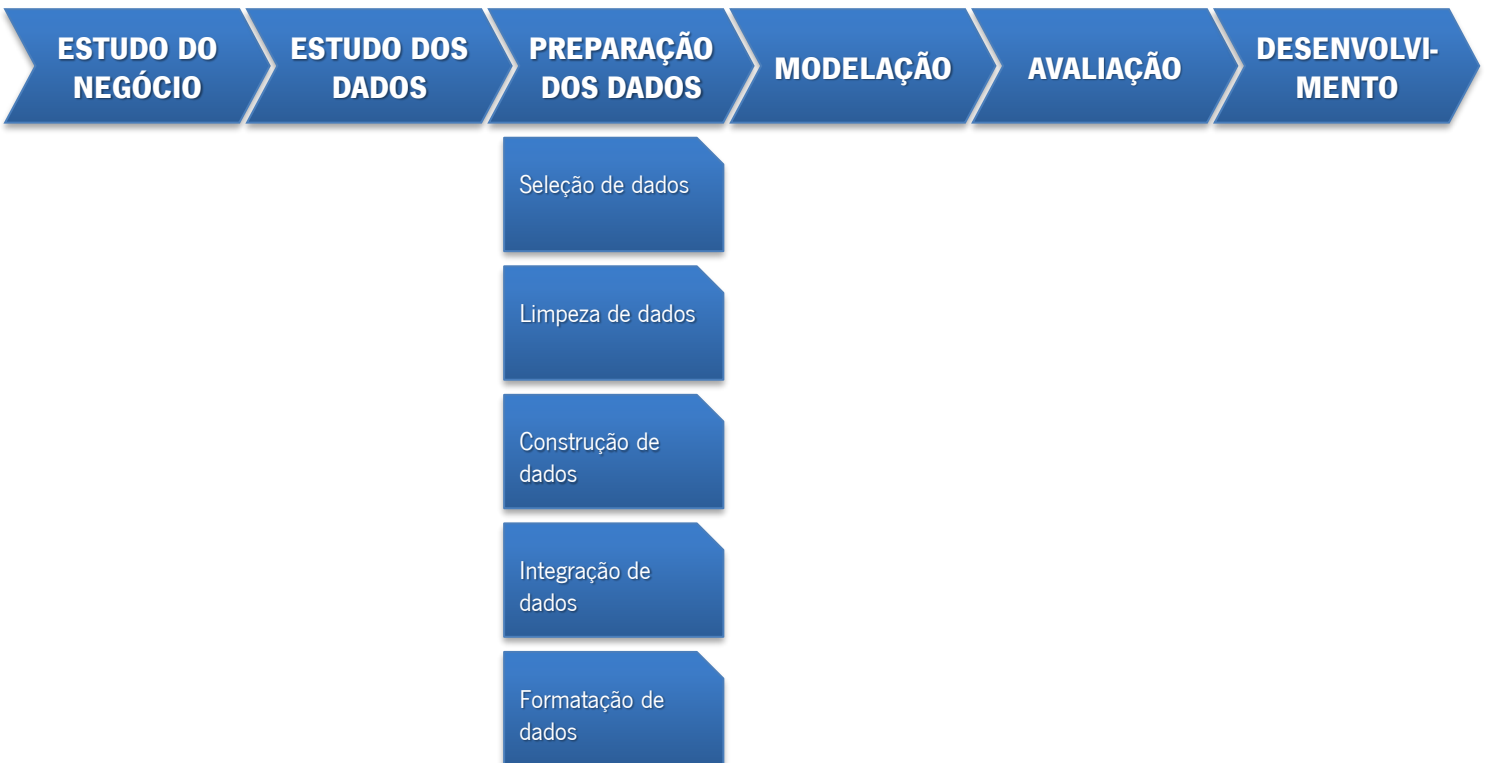
CRISP-DM
Fases e Tarefas



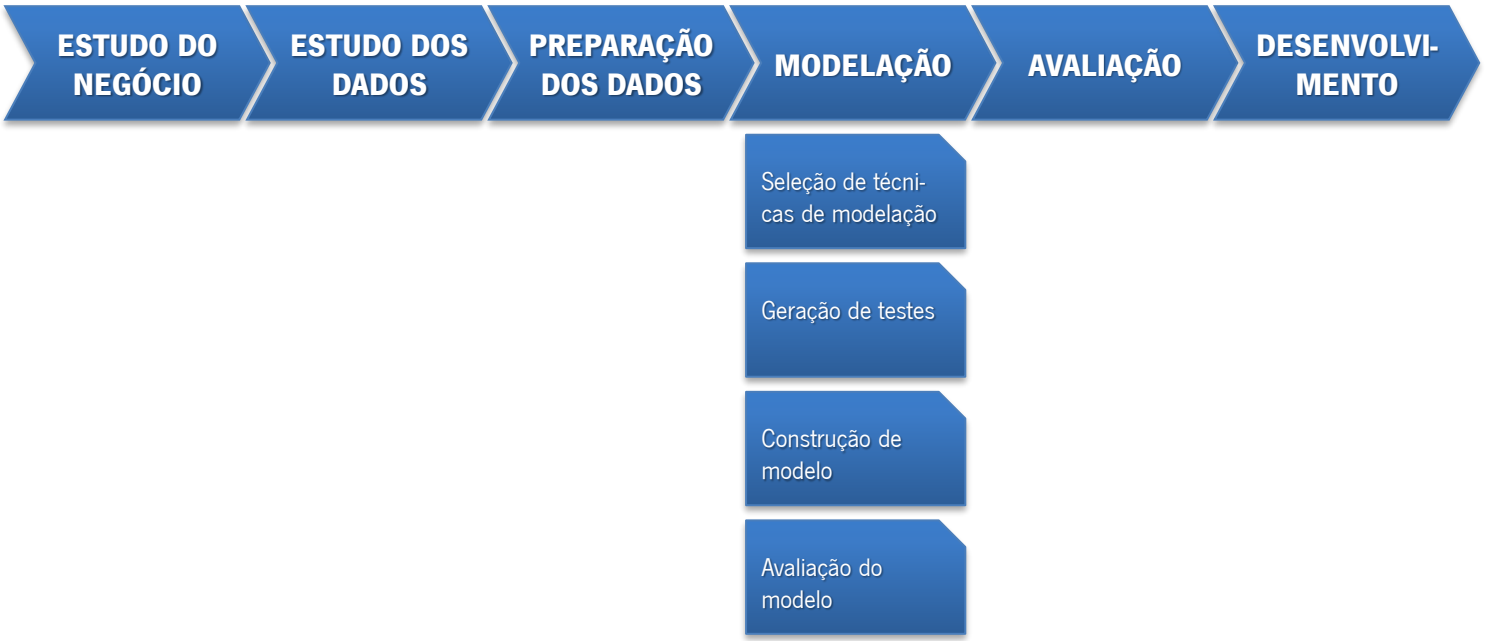
CRISP-DM
Fases e Tarefas



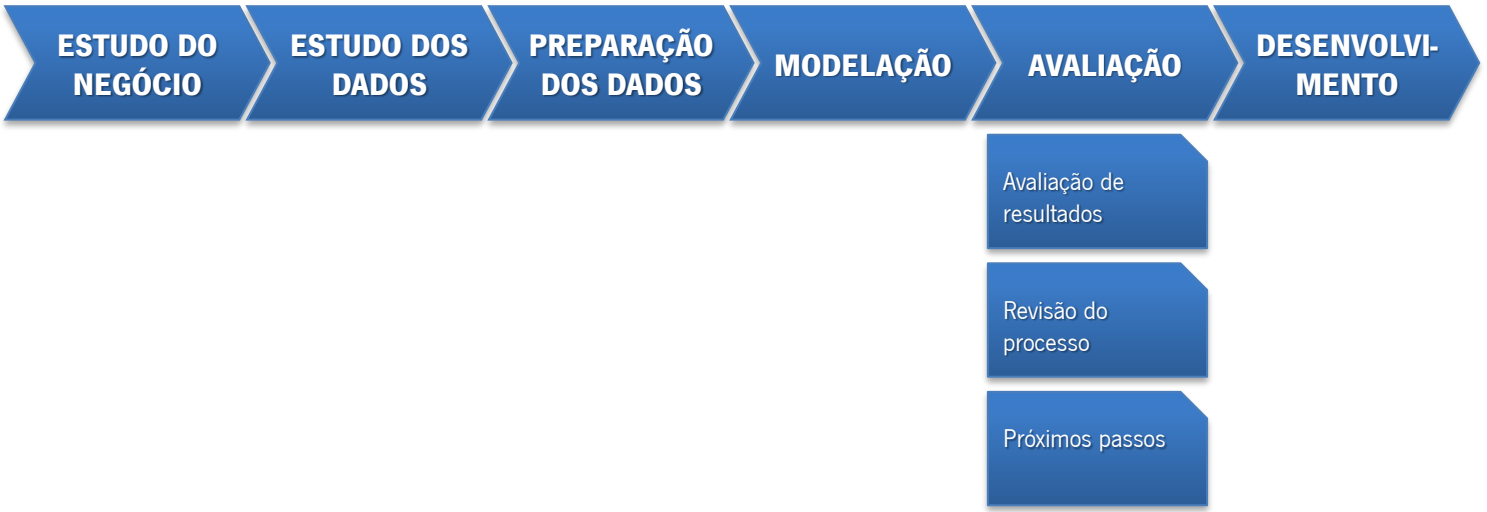
CRISP-DM
Fases e Tarefas



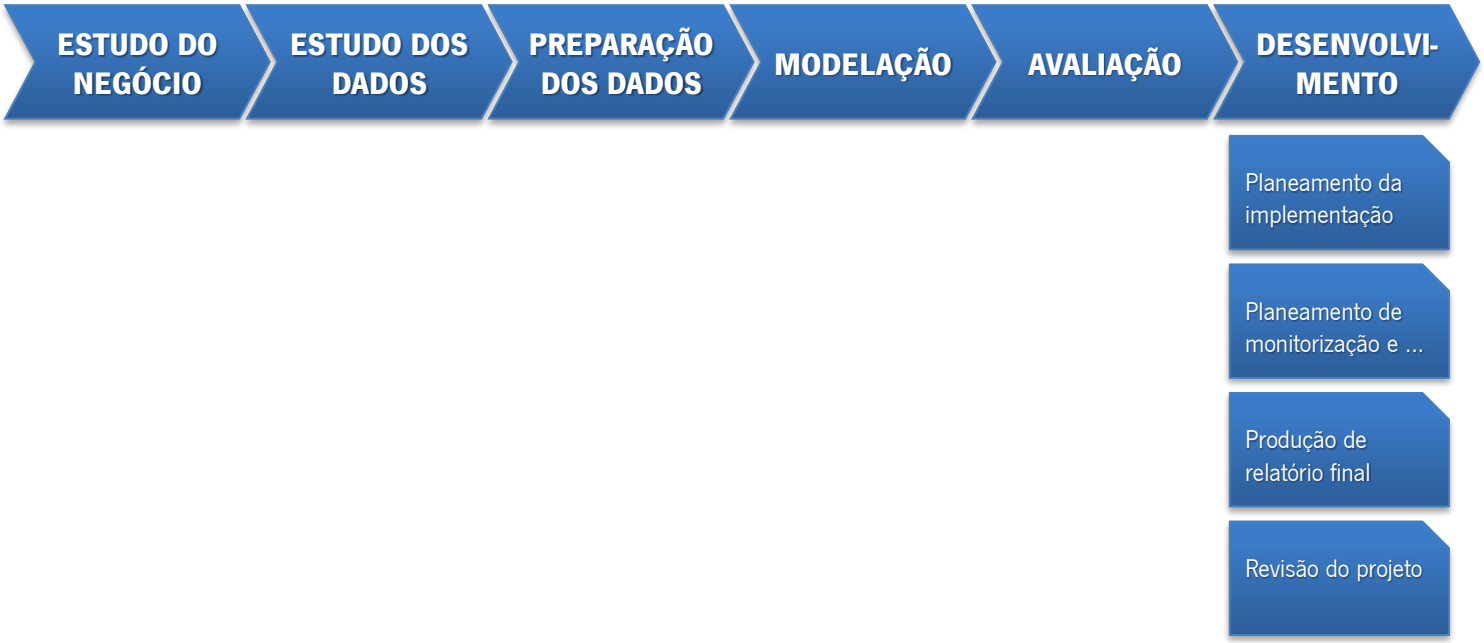
CRISP-DM
Fases e Tarefas



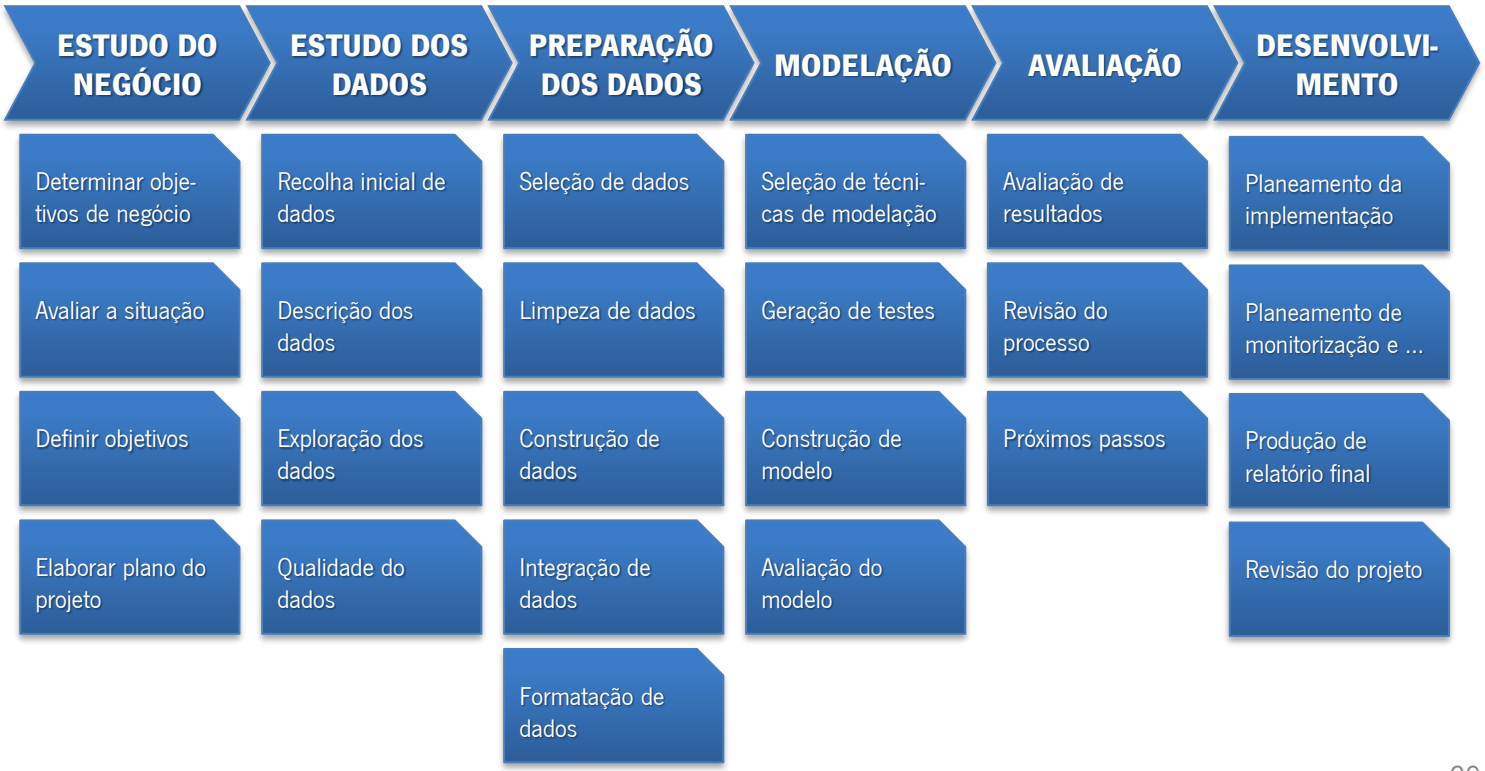
CRISP-DM
Fases e Tarefas



CRISP-DM
Fases e Tarefas

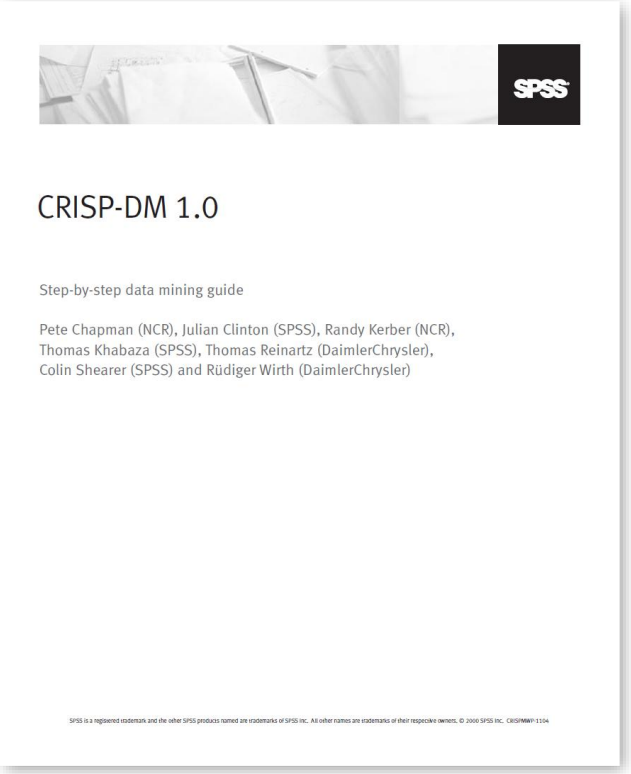


CRISP-DM
Fases e Tarefas



Metodologia
CRISP-DM

- “CRISP-DM 1.0: Step-by-step data mining guide”,
  Pete Chapman (NCR),
  Julian Clinton (SPSS),
  Randy Kerber (NCR),
  Thomas Khabaza (SPSS),
  Thomas Reinartz (DaimlerChrysler)
  Colin Shearer (SPSS)
  Rüdiger Wirth (DaimlerChrysler)
- CRISP-DM (pdf file)
- IBM SPSS Modeler CRISP-DM Guide



Que metodologias?

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

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

SEMMA

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

O processo SEMMA

- Divide o processo de *Data Mining* em 5 etapas:
  - *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.
  - *Explore*/ Exploração;
  - *Modify*/ Modificação;
  - *Model*/ Modelação;
  - *Assess*/ Avaliação.

O processo SEMMA

- Divide o processo de *Data Mining* em 5 etapas:
  - *Sample*/ Amostragem;
  - *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;
  - *Modify*/ Modificação;
  - *Model*/ Modelação;
  - *Assess*/ Avaliação.

O processo SEMMA

- Divide o processo de *Data Mining* em 5 etapas:
  - *Sample*/ Amostragem;
  - *Explore*/ Exploração;
  - *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;
  - *Model*/ Modelação;
  - *Assess*/ Avaliação.

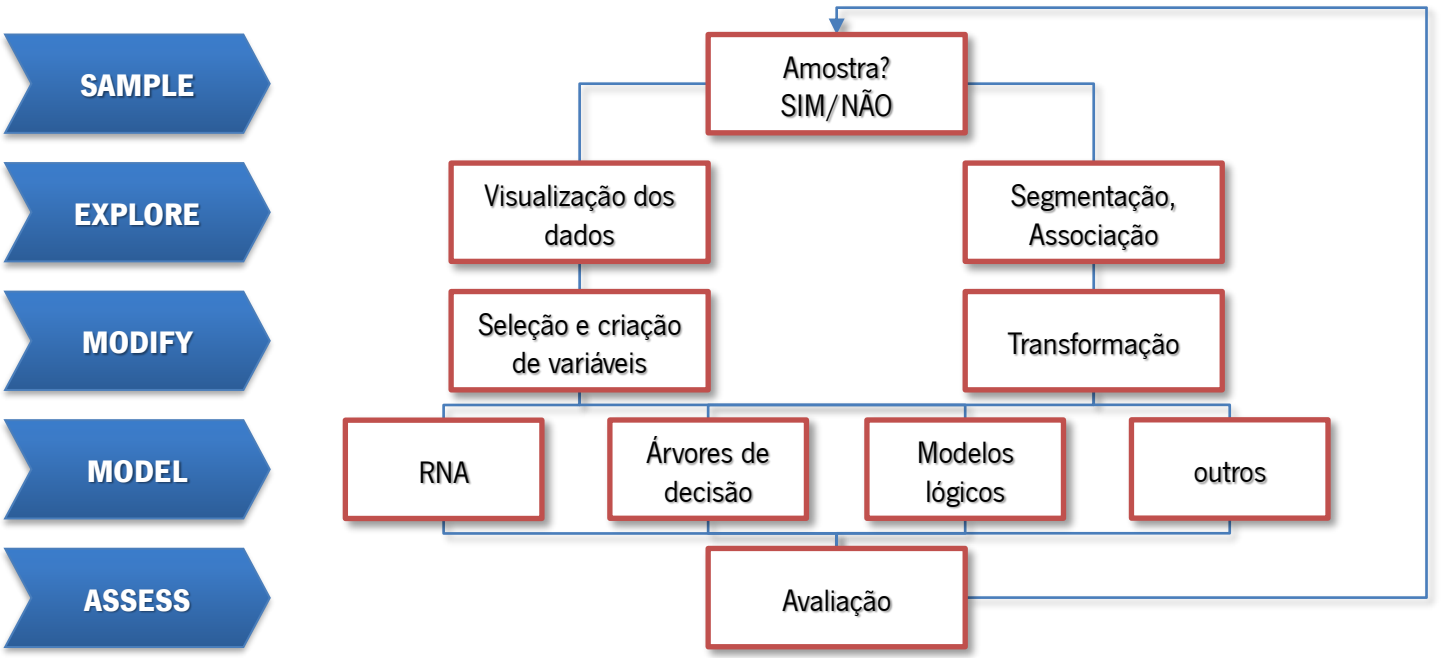
O processo SEMMA

- Divide o processo de *Data Mining* em 5 etapas:
  - *Sample*/ Amostragem;
  - *Explore*/ Exploração;
  - *Modify*/ Modificação;
  - *Model*/ Modelação:
    - Definição das técnicas de construção de modelos de *Data Mining*: redes neuronais artificiais, árvores de decisão, regressão linear, etc.;
    - Dependente do tipo de dados presentes em cada modelo (p.ex., RNA são mais adequadas quando os dados do problema apresentam relacionamentos complexos);
  - *Assess*/ Avaliação.

O processo SEMMA

- Divide o processo de *Data Mining* em 5 etapas:
  - *Sample*/ Amostragem;
  - *Explore*/ Exploração;
  - *Modify*/ Modificação;
  - *Model*/ Modelação;
  - *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.

O processo SEMMA



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

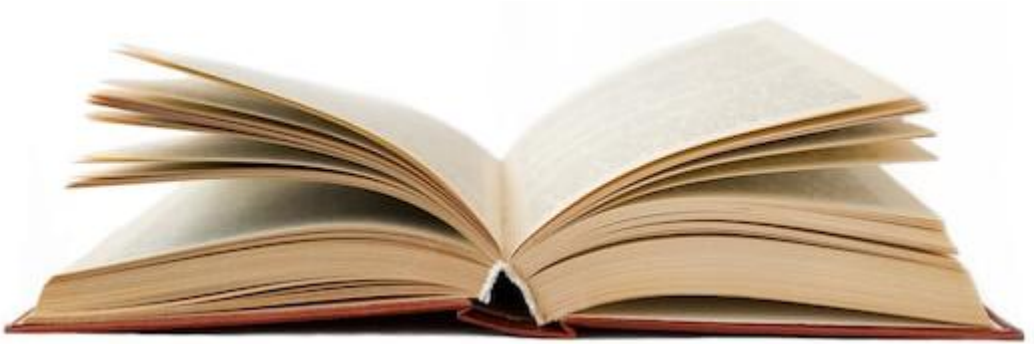
CRISP-DM versus SEMMA

- Fases CRISP-DM:
  - Estudo do negócio;
  - Estudo dos dados;
  - Preparação dos dados;
  - Modelação;
  - Avaliação;
  - Desenvolvimento.

- Processo SEMMA:
  - Amostragem;
  - Exploração;
  - Modificação;
  - Modelação;
  - Avaliação.

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



Universidade do Minho
Departamento de Informática

Aprendizagem e Decisão Inteligentes

LEI/MiEI @ 2024/2025, 2º sem