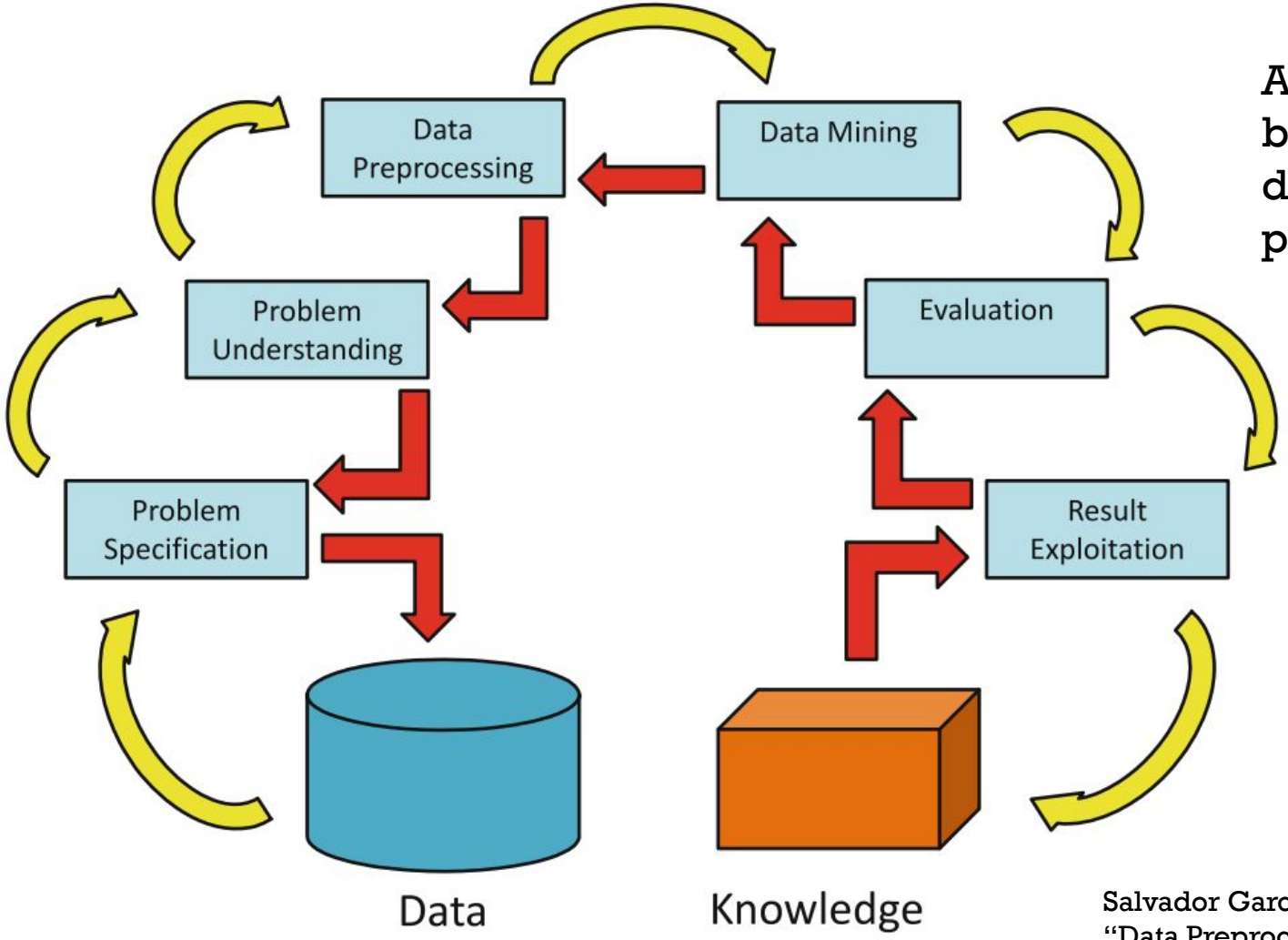
PRÉ-PROCESSAMENTO DE DADOS ESTRUTURADOS

Profa. Dra. Marcela Xavier Ribeiro

BACKGROUND

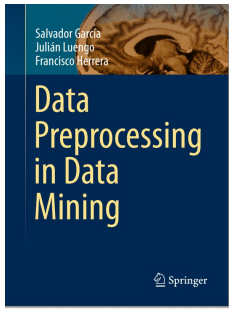
- Descoberta de Conhecimento em Base de Dados (Knowledge Discovery in Databases - KDD)
- Modelos de aprendizado
- Tarefas clássicas de mineração de dados
- Relação entre o pré-processamento de dados e mineração de dados

O QUE É KDD?



A descoberta de conhecimento em bancos de dados (KDD) é o processo de descoberta de conhecimento útil a partir de uma coleção de dados.

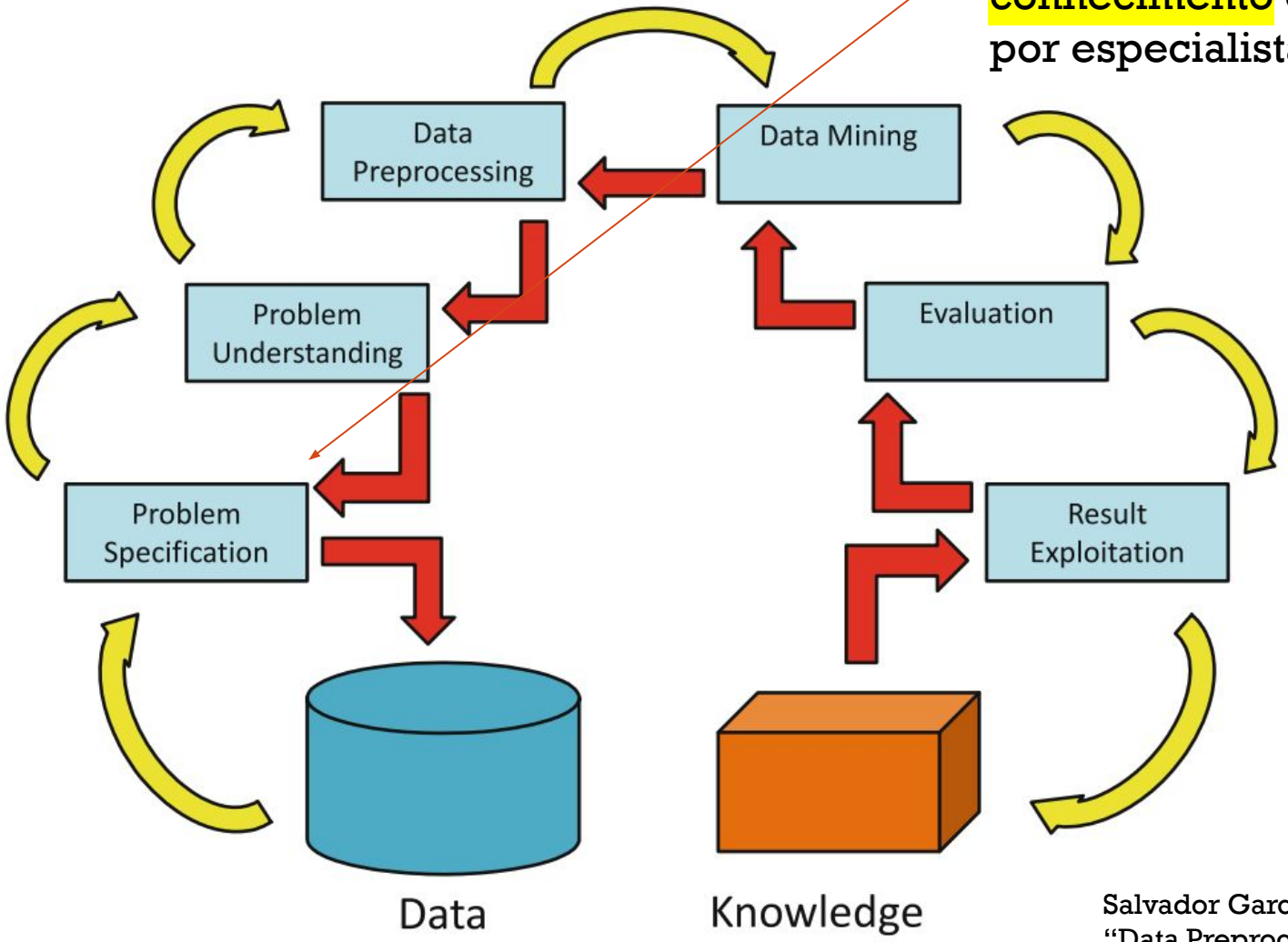
Fonte:



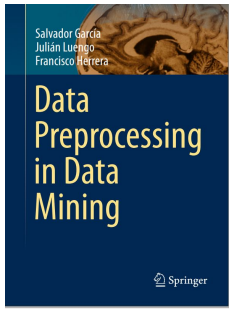
Salvador García, Julián Luengo, Francisco Herrera, "Data Preprocessing in Data Mining", Springer, 2015

O QUE É KDD?

1. Especificação do problema: levantar o domínio do conhecimento e informações prévias relevantes obtidas por especialistas e os objetivos finais da análise;



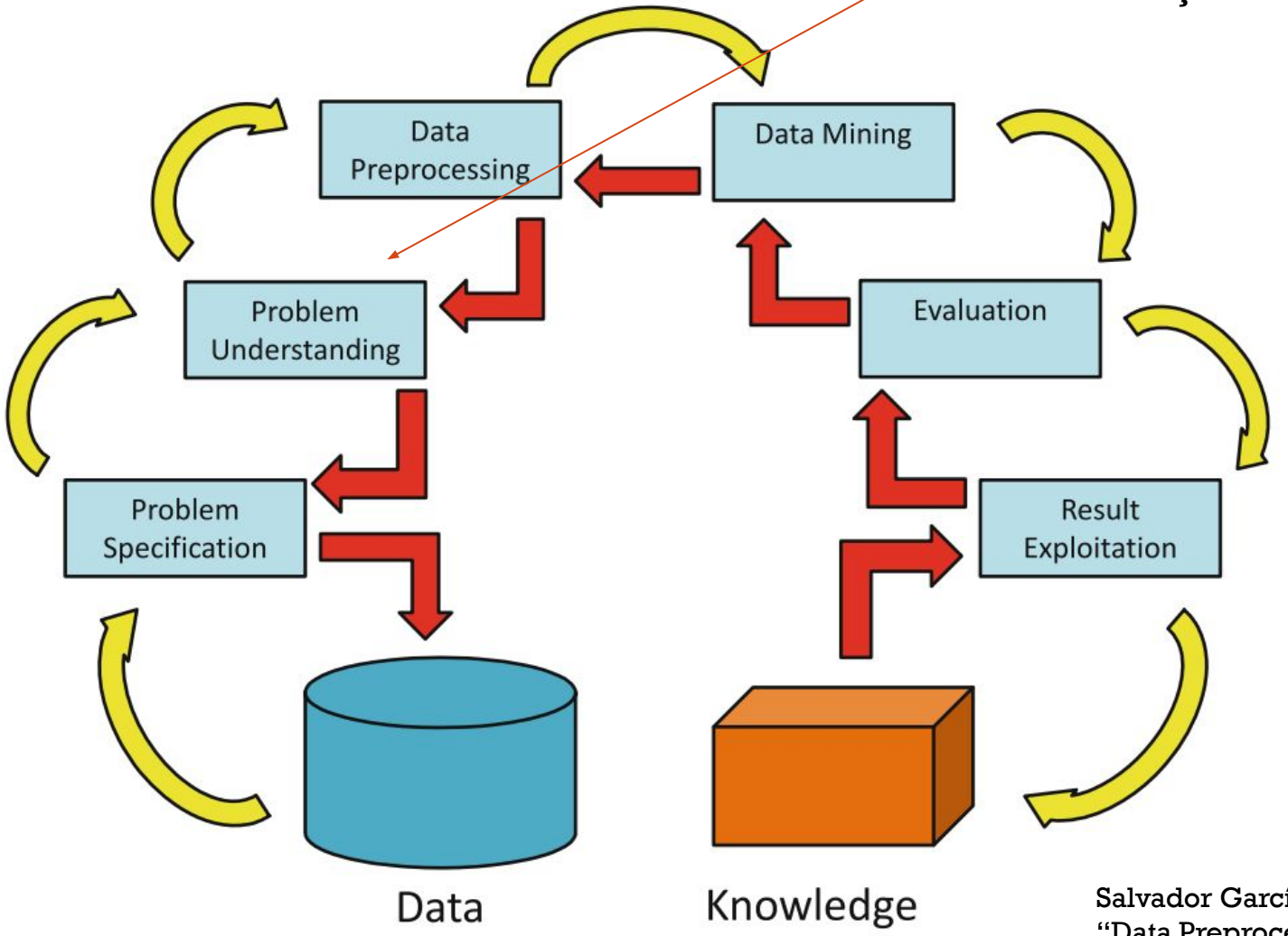
Fonte:



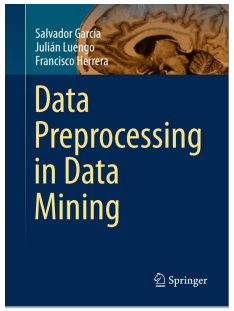
Salvador García, Julián Luengo, Francisco Herrera, "Data Preprocessing in Data Mining", Springer, 2015

O QUE É KDD?

2. Compreensão do problema: compreender o problema e abordar conhecimento especializado, a fim de alcançar alto grau de confiabilidade.

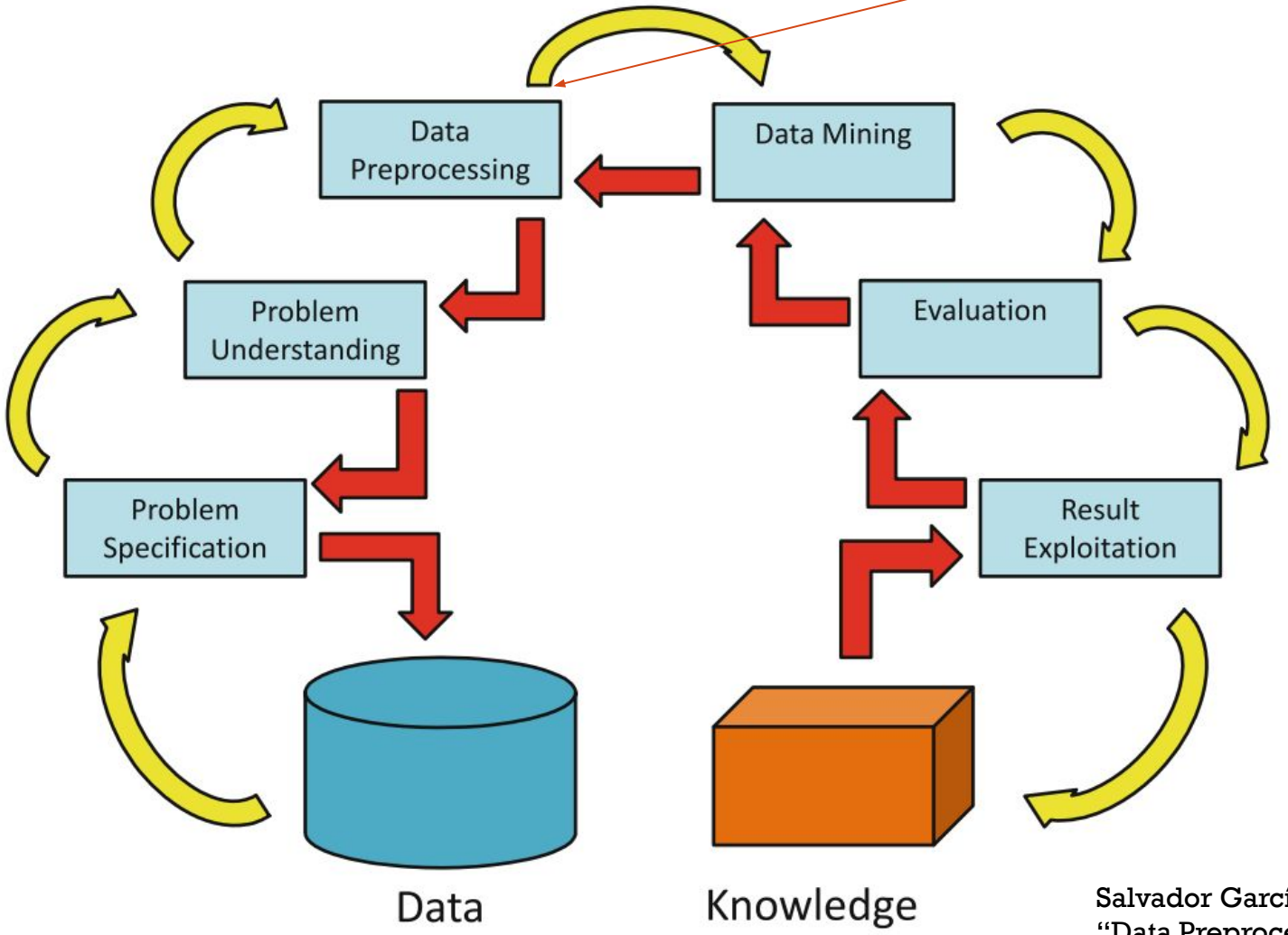


Fonte:



Salvador García, Julián Luengo, Francisco Herrera, "Data Preprocessing in Data Mining", Springer, 2015

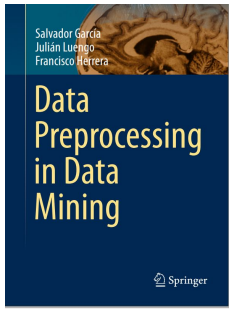
O QUE É KDD?



3. Pré-processamento de dados: inclui operações para:

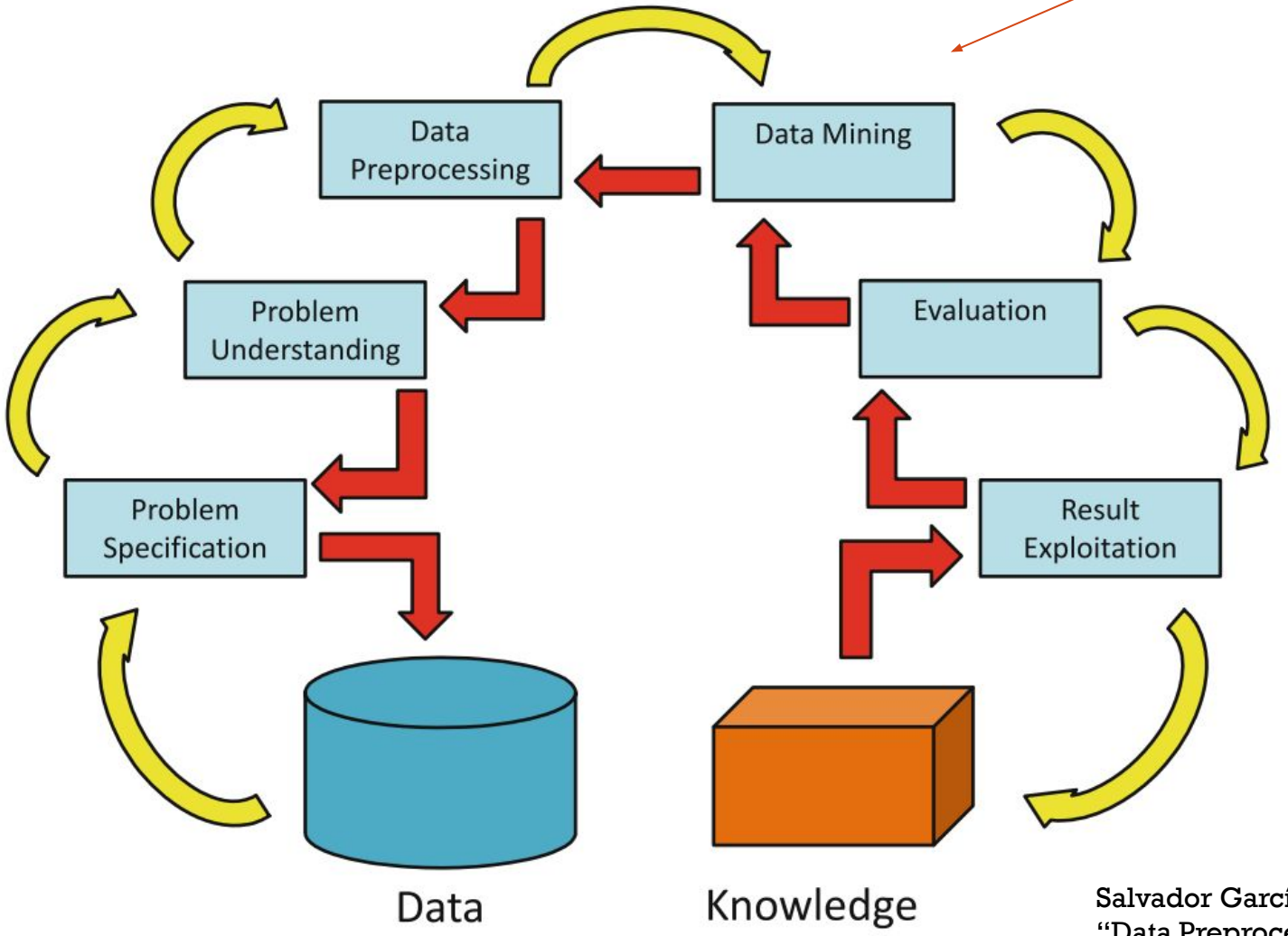
- ❑ limpeza de dados (como lidar com a remoção de ruído e dados inconsistentes);
- ❑ integração de dados (onde várias fontes de dados podem ser combinadas em uma);
- ❑ transformação de dados (onde os dados são transformados e consolidados em formas apropriadas para as tarefas específicas de DM);
- ❑ redução de dados, incluindo seleção e extração de atributos e tuplas em uma base de dados.

Fonte:



Salvador García, Julián Luengo, Francisco Herrera, "Data Preprocessing in Data Mining", Springer, 2015

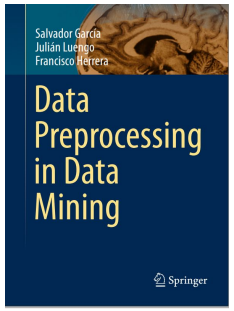
O QUE É KDD?



4. Data Mining: É o processo onde os métodos são usados para extrair padrões. Esta etapa inclui:

- ❑ a escolha da tarefa DM mais adequada (como classificação, regressão, agrupamento ou associação);
- ❑ a escolha do próprio algoritmo DM, pertencente a uma das famílias anteriores;
- ❑ o emprego dos algoritmos selecionados para o problema;
- ❑ ajuste dos parâmetros desse algoritmos;
- ❑ procedimentos de validação.

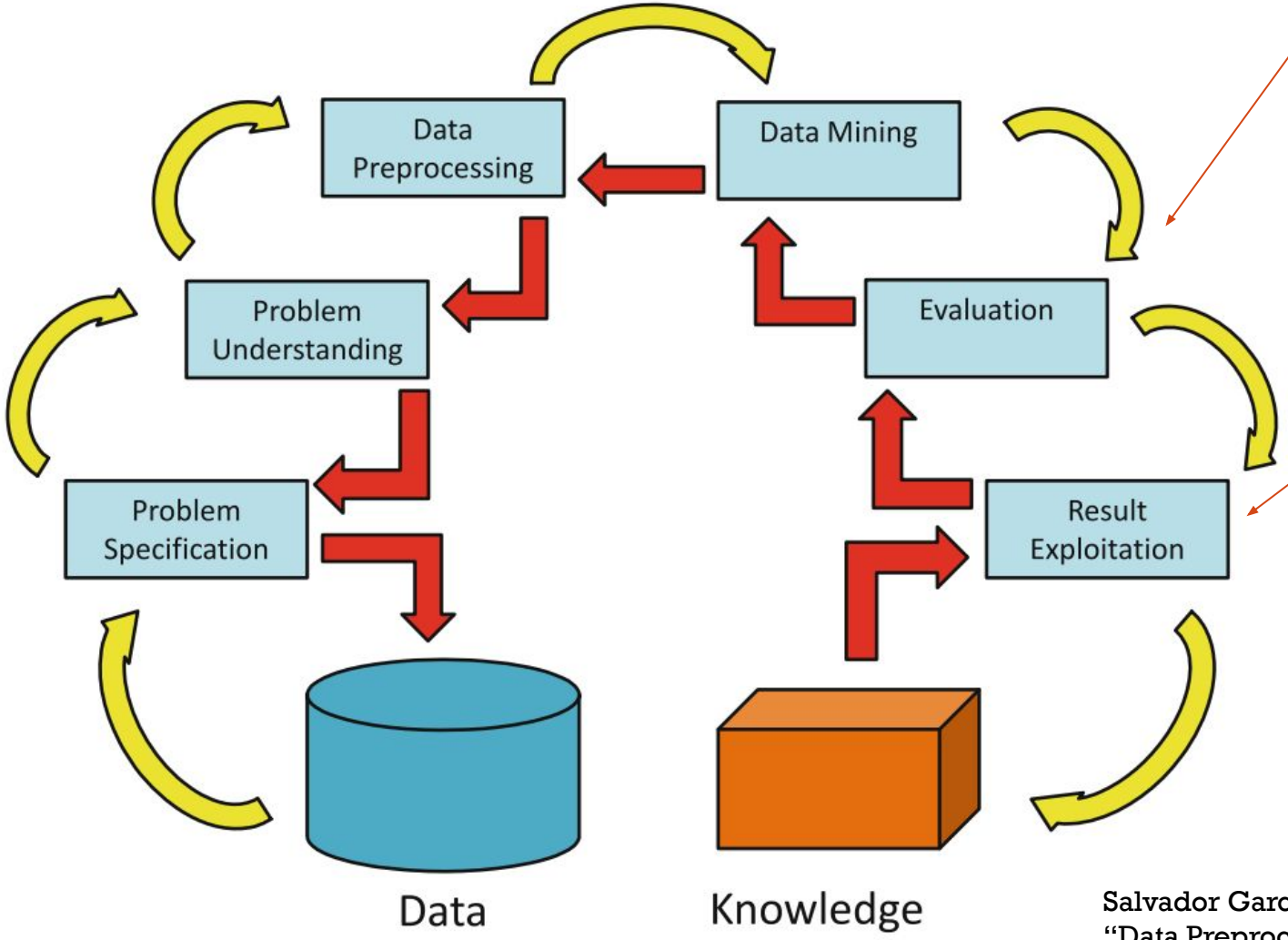
Fonte:



Salvador García, Julián Luengo, Francisco Herrera, "Data Preprocessing in Data Mining", Springer, 2015

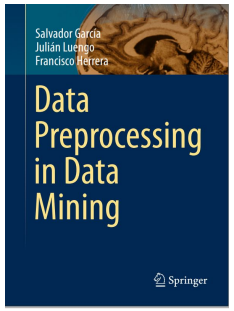
O QUE É KDD?

5. Avaliação: Estimar e interpretar os padrões extraídos com base em medidas de interesse.



6. Exploração do Resultado: Envolve o uso direto do conhecimento; incorporação do conhecimento ou relato do conhecimento descoberto por meio de ferramentas de visualização.

Fonte:



Salvador García, Julián Luengo, Francisco Herrera, "Data Preprocessing in Data Mining", Springer, 2015

ENTENDENDO O PORQUÊ DE TODO ESSE TRABALHO

1. Tipos de dados
2. Para que pré-processar?
3. Limpeza
4. Integração e transformação
5. Redução
6. Discretização
7. Resumo

TIPC

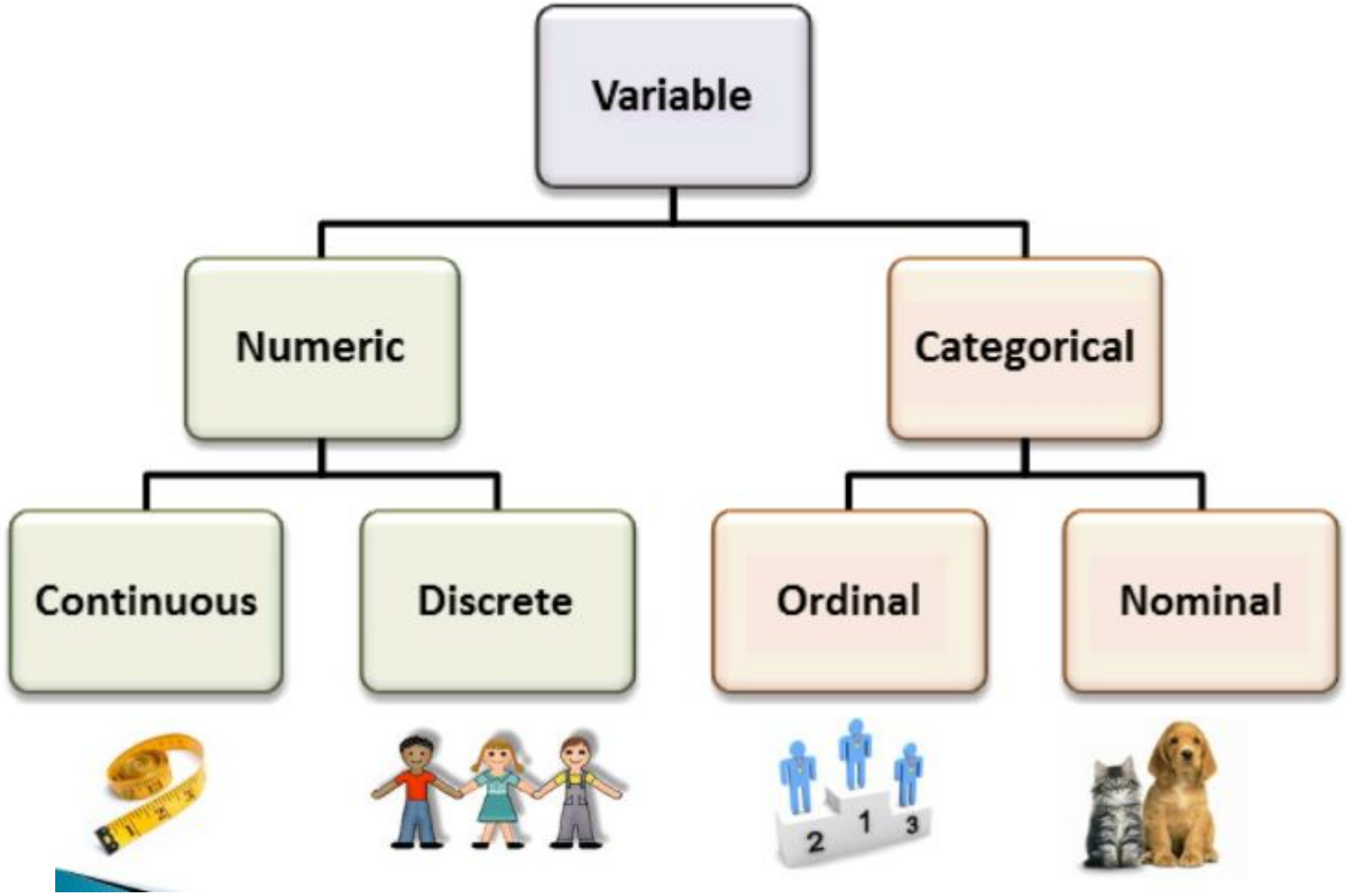
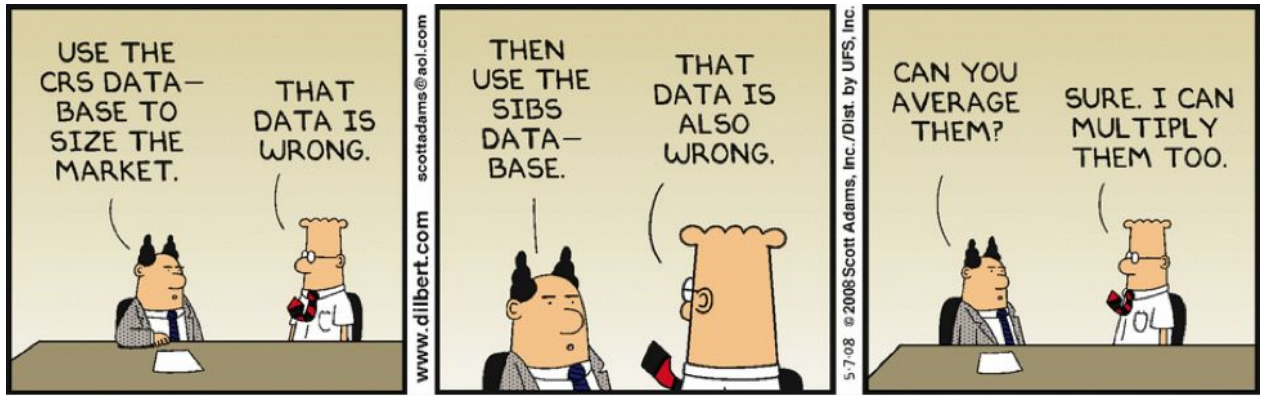


Figura adaptada de https://medium.com/brasil-ai/antes-de-come%C3%A7armos-a-falar-sobre-tipos-de-aprendizados-que-veremos-no-pr%C3%B3ximo-artigo-%C3%A9-ea5b04685913

DADOS CRUS



Tendem a ser:

- incompletos: faltando valores
  - ex. gênero =“”
- ruidosos: contendo erros e outliers
  - ex. salário=“-10”
- inconsistentes: contêm discrepâncias
  - ex. idade=“42”, nascimento=“03/07/2000”

PRÉ-PROCESSAMENTO DE DADOS

- O pré-processamento de dados converte dados em dados adequados que se encaixam em um processo de mineração de dados (DM)
- Se os dados não estiverem preparados, o algoritmo DM pode não rodar ou os resultados poderão ser enganosos.

QUALIDADE DOS DADOS

- Completo
- Consistente
- Confiável
- Interpretável
- Acessível

QUESTÕES DE PRÉ-PROCESSAMENTO

- Como limpo os dados? —Limpeza de dados.
- Como forneço dados precisos? - Transformação de dados.
- Como faço para analisar dados de diversas fontes? —Integração de dados.
- Como posso unificar e dimensionar os dados? - Normalização de dados.
- Como faço para lidar com dados faltantes? - Imputação de dados.
- Como posso detectar e gerenciar o ruído? - Identificação e tratamento de ruídos.

ETAPAS DO PRÉ-PROCESSAMENTO

- Limpeza
  - Preencher de valores faltantes, suavizar dados ruidosos, identificação e remoção de erros e ruídos, resolver inconsistências;
- Integração
  - Integrar múltiplas bases de dados e arquivos
- Transformação
  - Normalização e agregação
- Redução
  - Redução do número de atributos ou do número de instâncias, ou em ambos, mantendo os mesmos resultados dos dados originais
- Discretização (para dados contínuos)

LIMPEZA DE DADOS

- Tarefas de limpeza de dados:
- Preenchimento de dados faltantes;
- Identificação de ruídos, suavização e remoção de ruídos;
- Correção de dados inconsistentes;
- Eliminar redundância (duplicidade de tuplas)

DADOS FALTANTES

- Dados não estão sempre disponíveis
  - ex. o campo renda dos clientes …
- Os dados podem ser faltantes por:
  - mal funcionamento do equipamento de coleta, no caso de sensores;
- Inconsistência com outro registro e assim, foi deletado;
- Não fornecido ou duvidoso;

DADOS FALTANTES

- Ignorar a tupla toda;
- Preenchimento manual:
  - cansativo e inviável?
- Preenchimento automático com:
  - uma constante global: ex., "unknown";
  - o valor médio do atributo;
  - usando interpolação;
  - estimar o valor usando um método de inferência: fórmula de Bayes, árvore de decisão ou algoritmo EM;

REMOÇÃO DE OUTLIERS

- Pontos de dados inconsistente com a maioria dos dados
- Outliers podem ser:
  - Válidos;
  - Ruidosos: idade= 300;
- Métodos de remoção de outliers:
  - Clustering;
  - Ajuste de curva;
  - Teste de hipótese usando um determinado Modelo;

INTEGRAÇÃO DE DADOS

- Combina dados de múltiplas fontes;
- Integra metadados de diferentes fontes;
- Problema de identificação de entidade: identificar entidades do mundo real de diferentes fontes de dados. Ex. empresa = companhia;
- Detectar e resolver conflitos de valores de dados: os valores dos atributos são diferentes, ex. diferentes escalas e métricas;
- Remover dados duplicados e redundantes.

TRANSFORMAÇÃO DE DADOS

- Suavização (Smoothing): remove ruídos dos dados;
- Normalização e padronização: transformam todas as variáveis na mesma ordem de grandeza.
- Construção de Atributo/Características: novos atributos construídos a partir de atributos existentes;
- Agregação: sumarização;
- Generalização: tornar o valor mais genérico em uma hierarquia.

REDUÇÃO DOS DADOS

- A eficiência de uma solução depende em muitos casos do tamanho do problema
- O tamanho de um problema de mineração refere-se a:
  - Número de atributos
  - Número de exemplos de treinamento
- O tamanho de um problema de aprendizado interfere na:
  - qualidade das respostas (precisão) dos algoritmos
  - e no custo do aprendizado

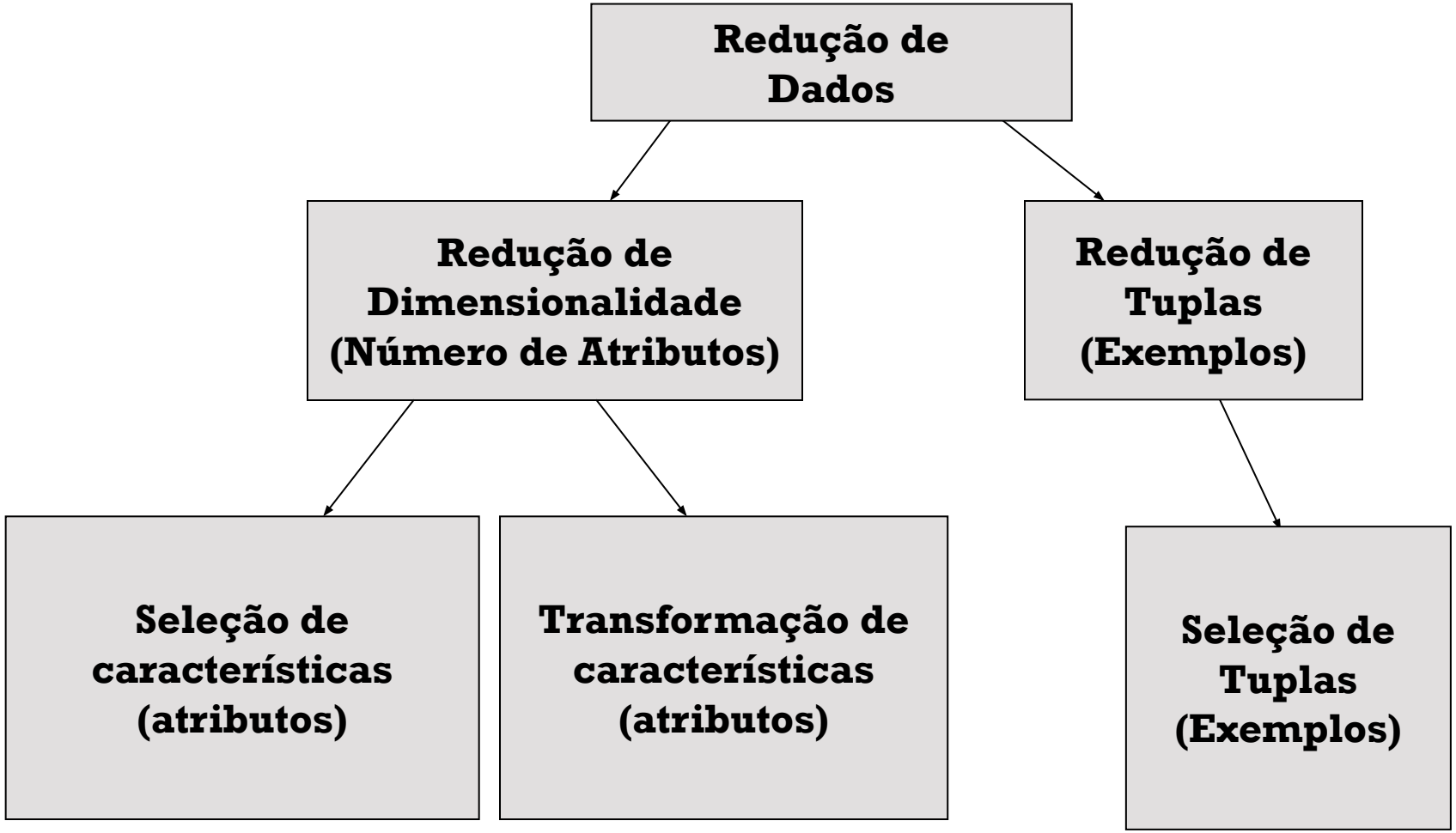
REDUÇÃO DOS DADOS

- A quantidade de dados é muito grande para lidar;
- Obter uma representação do conjunto de dados que é muito menor em volume e produz o mesmo (ideal) ou quase o mesmo resultado ao ser analisado;

REDUÇÃO DOS DADOS

- Estratégias de redução de dados:
  - Redução de dimensionalidade:
    - diminuição no número de atributos:
      - Seleção de características
      - Transformação de características
  - Diminuição no número de tuplas:
      - Amostragem.

REDUÇÃO DE DADOS



REDUÇÃO DE DIMENSIONALIDADE

- Redução do número de características (atributos)
- Alguns atributos são redundantes e assim podem ser eliminados;
- Encontra um conjunto de atributos que sejam relevantes e não-redundantes.

- Vantagens
  - Diminui o custo do aprendizado
  - Aumenta a precisão do algoritmo
  - Gera modelos compactos mais fáceis de interpretar

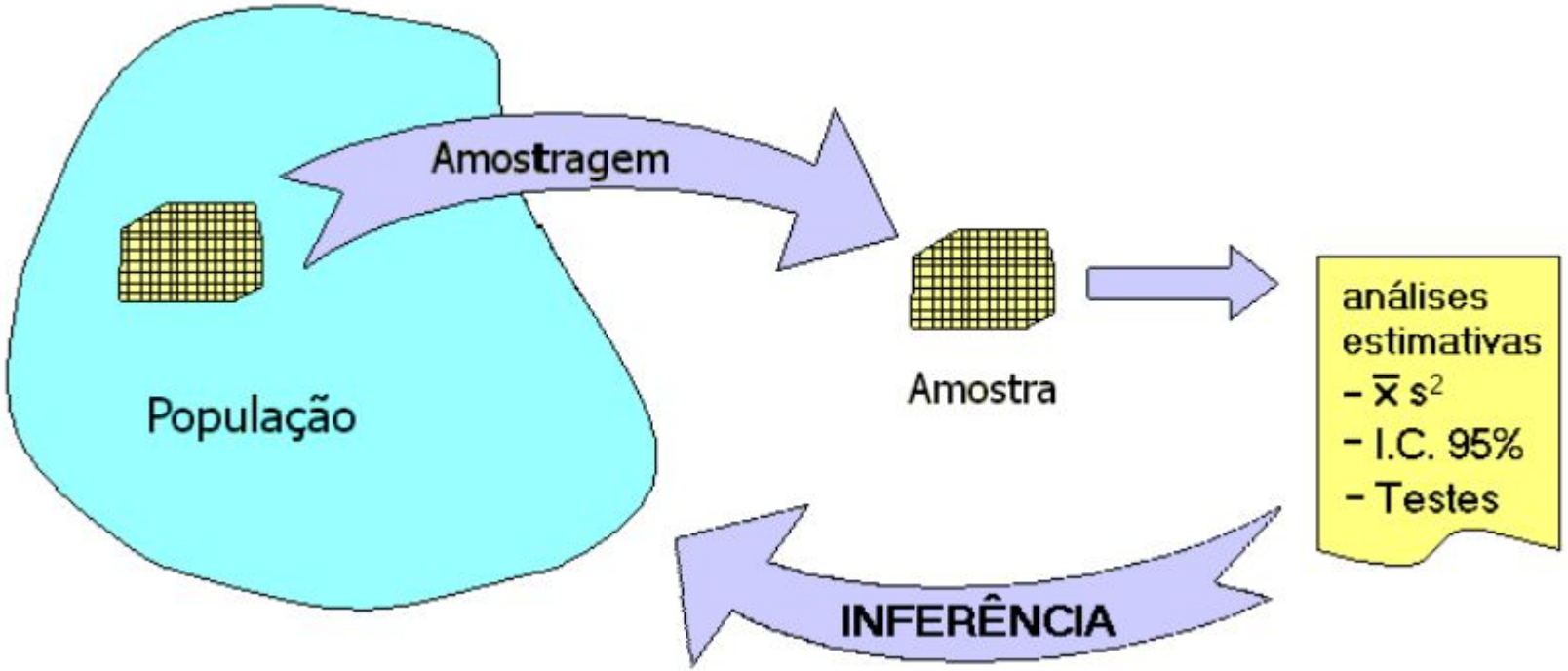
TÉCNICAS DE REDUÇÃO DE DIMENSIONALIDADE

- Seleção de características (atributos):
  - Escolha de um sub-conjunto de atributos relevantes dentre os atributos disponíveis
  - ex., Filtros e Wrappers

- Transformação de características (atributos):
  - Criação de novos atributos a partir da combinação dos atributos existentes
  - ex., PCA

AMOSTRAGEM

- Escolha de um subconjunto representativo dos dados



Fonte da Figura: https://sites.google.com/site/estatisticabasicacc/conteudo/parte-2---inferencia/01---amostragem

DISCUSSÃO

- A preparação dos dados é um dos maiores problemas do processo do KDD
- O pré-processamento dos dados é a fase mais trabalhosa do processo de KDD e é geralmente o que define o sucesso do mesmo.

Alguém já se deparou com esse problema? Conte-nos como foi!