

An aerial photograph of a university campus. The campus features several large, modern buildings with white and red roofs, interspersed with lush green lawns and trees. A winding river or canal flows through the campus, with a bridge crossing it. In the foreground, there's a large, open area with a paved path and a small pond. The overall scene is bright and sunny, with clear skies.

WEB ACADEMY

Tópicos Emergentes

Daniel Augusto Nunes da Silva

Apresentação

Ementa

- Processo de descoberta de conhecimento em base de dados (**KDD**), **mineração de dados** e **aprendizado de máquina**. Construção e implantação de modelos de aprendizado de máquina. Integração de dados e o processo de **ETL**. **Visualização de dados**. Processamento Analítico (**OLAP**).

Objetivos

- **Geral:** Apresentar conceitos e práticas relacionados à **utilização de técnicas de mineração de dados em sistemas de software**, fornecendo uma visão geral do processo, desde a **compreensão do problema** até a **implantação** de modelos de aprendizagem de máquina em produção.
- **Específicos:**
 - Relacionar os principais conceitos de mineração de dados;
 - Demonstrar o uso de ETL para auxiliar soluções voltadas ao processamento analítico e mineração de dados;
 - Apresentar técnicas para criação e avaliação de modelos de aprendizagem de máquina;
 - Implantar um modelo de aprendizagem de máquina em um projeto de software.

Conteúdo programático

Introdução

- Dado, informação e conhecimento.
- Introdução a mineração de dados.
- O processo de KDD.
- Tarefas de Mineração de dados.
- Tipos de aprendizado de máquina.
- Foco na solução do problema.

Tratamento e visualização dados

- Integração de dados e processo de ETL.
- OLTP x OLAP.
- Talend Open Studio.
- Visualização de dados.
- Métricas e indicadores de desempenho.
- Google Data Studio.

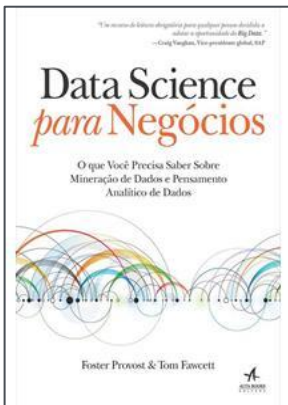
Construção de modelos preditivos

- O processo de mineração de um modelo de classificação.
- Representação do modelo de classificação.
- Avaliação de classificadores.
- Seleção de atributos.
- Classes desbalanceadas.
- WEKA.

Modelos preditivos em produção

- Definição de uma estratégia para implantação de modelos preditivos.
- Utilização da WEKA API.
- Configuração do projeto.
- Classificação de novas instâncias.

Bibliografia



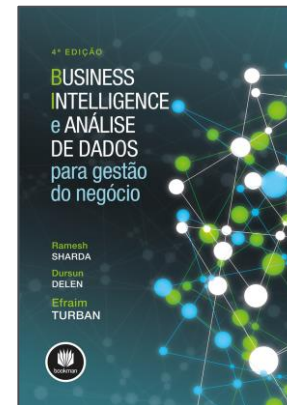
Data Science Para Negócios

Foster Provost e Tom Fawcett

1ª Edição – 2016

Editora Alta Books

ISBN 978-8576089728



Business Intelligence e Análise de Dados para Gestão do Negócio

Ramesh Sharda, Dursun Delen e

Efraim Turban

4ª Edição – 2019

Editora Bookman

ISBN 978-8582605196



Storytelling com Dados

Cole Nussbaumer Knaflic

2ª Edição – 2019

Editora Alta Books

ISBN 978-8550804682

Sites de referência

- Machine Learning Mastery.
 - <https://machinelearningmastery.com/>
- Weka Wiki.
 - <https://waikato.github.io/weka-wiki/>

Ferramentas

- **MySQL**

- <https://dev.mysql.com/downloads/windows/installer/8.0.html>
- Configurar a variável de ambiente PATH. Exemplo: "C:\Program Files\MySQL\MySQL Server 8.0\bin".
- **Importar dados:** `mysql -u root -p sgcm < sgcm.sql`
- Criar conta no <https://www.freemysqlhosting.net/>

- **Talend Open Studio for Data Integration**

- <https://www.talend.com/lp/open-studio-for-data-integration/>

- **Google Data Studio**

- <https://datastudio.google.com/>

- **Weka**

- <https://prdownloads.sourceforge.net/weka/weka-3-8-6-azul-zulu-windows.exe>

Ferramentas

- **Git**

- <https://git-scm.com/downloads>

- **Visual Studio Code**

- <https://code.visualstudio.com/Download>

- **Extension Pack for Java**

- <https://marketplace.visualstudio.com/items?vscjava.vscode-java-pack>

- **Spring Boot Extension Pack**

- <https://marketplace.visualstudio.com/items?itemName=pivotal.vscode-boot-dev-pack>

- **Angular Language Service**

- <https://marketplace.visualstudio.com/items?itemName=Angular.ng-template>

Ferramentas

- **JDK 11**

- <https://www.oracle.com/br/java/technologies/javase/jdk11-archive-downloads.html>
- Criar a variável de ambiente JAVA_HOME configurada para o diretório de instalação do JDK. Exemplo: “C:\Program Files\Java\jdk-11.0.13”.
- Adicionar “%JAVA_HOME%\bin” na variável de ambiente PATH.
- Tutorial de configuração: https://mkyong.com/java/how-to-set-java_home-on-windows-10/

- **Maven**

- <https://maven.apache.org/download.cgi>
- Adicionar o diretório de instalação do Maven na variável de ambiente PATH. Exemplo: “C:\apache-maven\bin”.
- Tutorial de instalação: <https://mkyong.com/maven/how-to-install-maven-in-windows/>

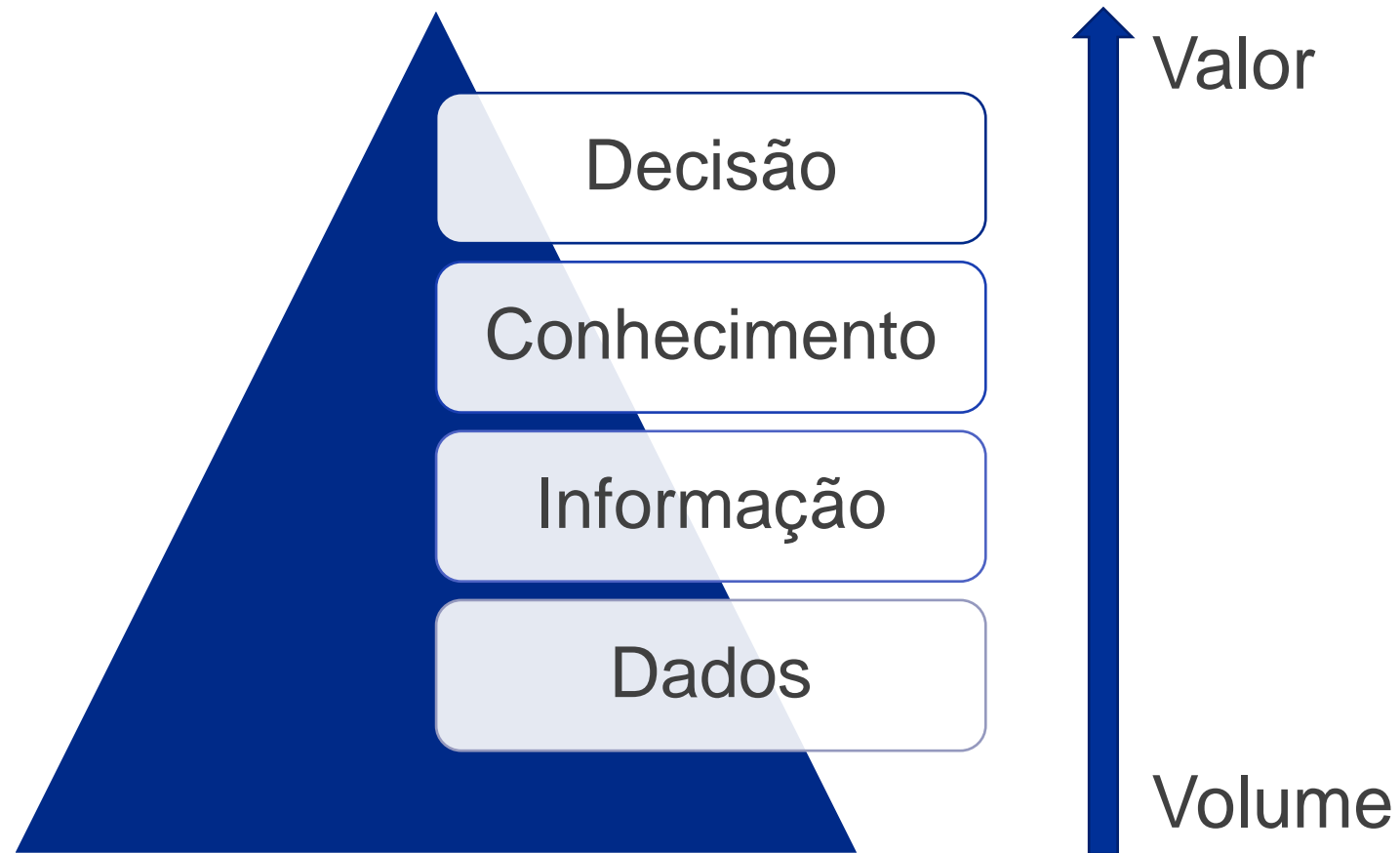
Contato



<https://linkme.bio/danielnsilva/>

Introdução

Dado, informação e conhecimento



Introdução a mineração de dados

- Um **processo manual de análise e interpretação de dados** pode ser considerado uma forma de transformar estes dados em conhecimento;
- No entanto, para muitos domínios **esta forma manual torna-se impraticável**, na medida em que o **volume de dados armazenados cresce exponencialmente**;
- Os padrões descobertos por meio deste processo devem ser **relevantes** na medida em que possam **representar alguma vantagem**, geralmente de natureza econômica;
- Termos relacionados: mineração de dados, aprendizado de máquina, ciência de dados, etc.

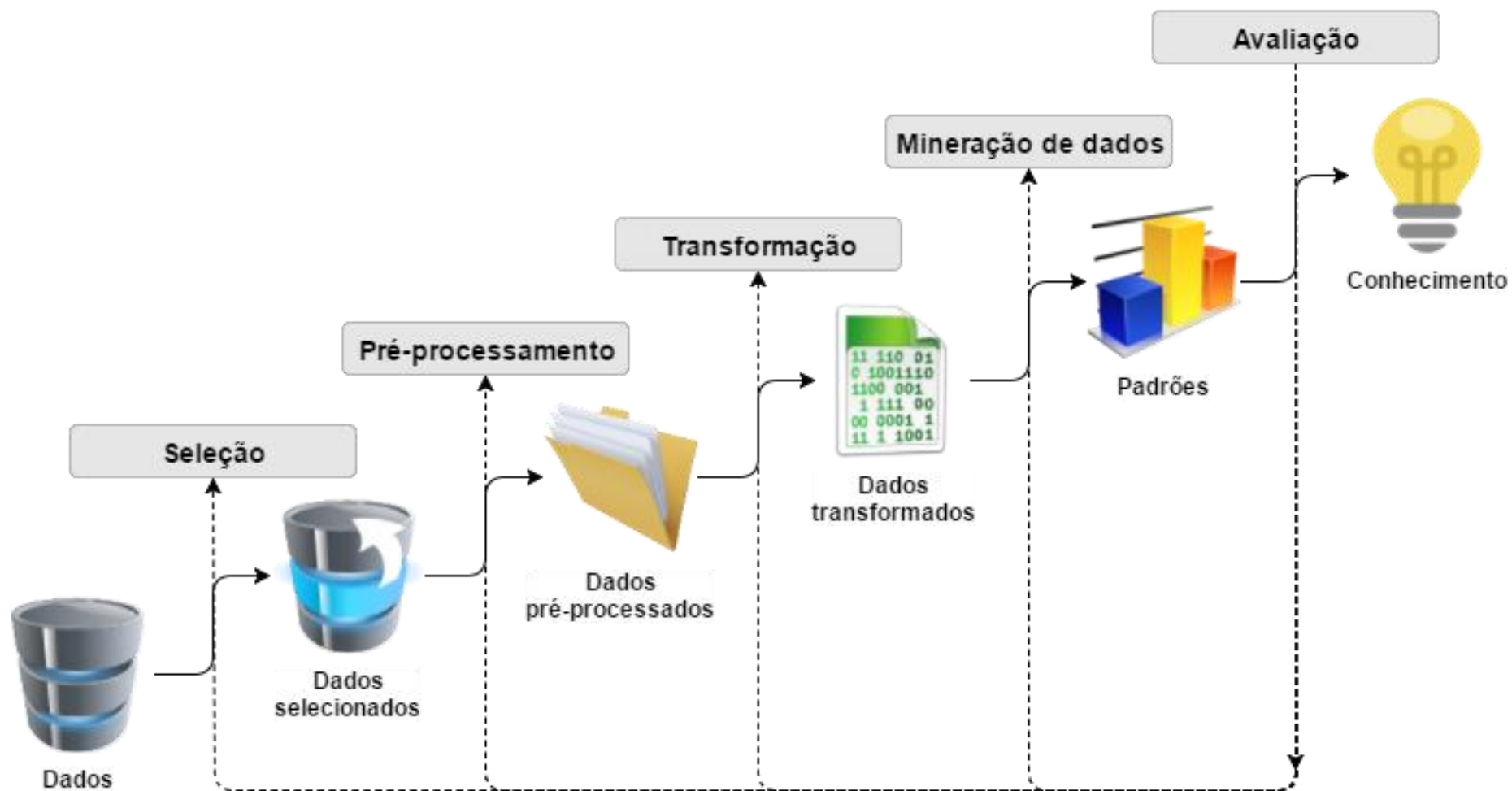
O processo de KDD

- **KDD** – *Knowledge-Discovery in Databases* (descoberta de conhecimento em bases de dados);
- Perspectiva do conhecimento extraído:
 - “Processo de identificação de **padrões válidos, novos, potencialmente úteis e compreensíveis** embutidos nos dados” (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996);
- Perspectiva da realização do processo:
 - “O processo de KDD consiste de uma **sequência de interações** complexas, que se estende sobre um determinado período de tempo, entre um **usuário** e uma **coleção de dados**, possivelmente auxiliado por um conjunto heterogêneo de ferramentas computacionais” (BRACHMAN e ANAND, 1996).

Visão geral do processo de KDD

- Processo interativo e iterativo:
 - É **interativo** por envolver muitas decisões feitas pelo usuário em cada etapa;
 - É também **iterativo**, pois durante o processo podem ser realizadas várias iterações até que os objetivos sejam alcançados.

Visão geral do processo de KDD

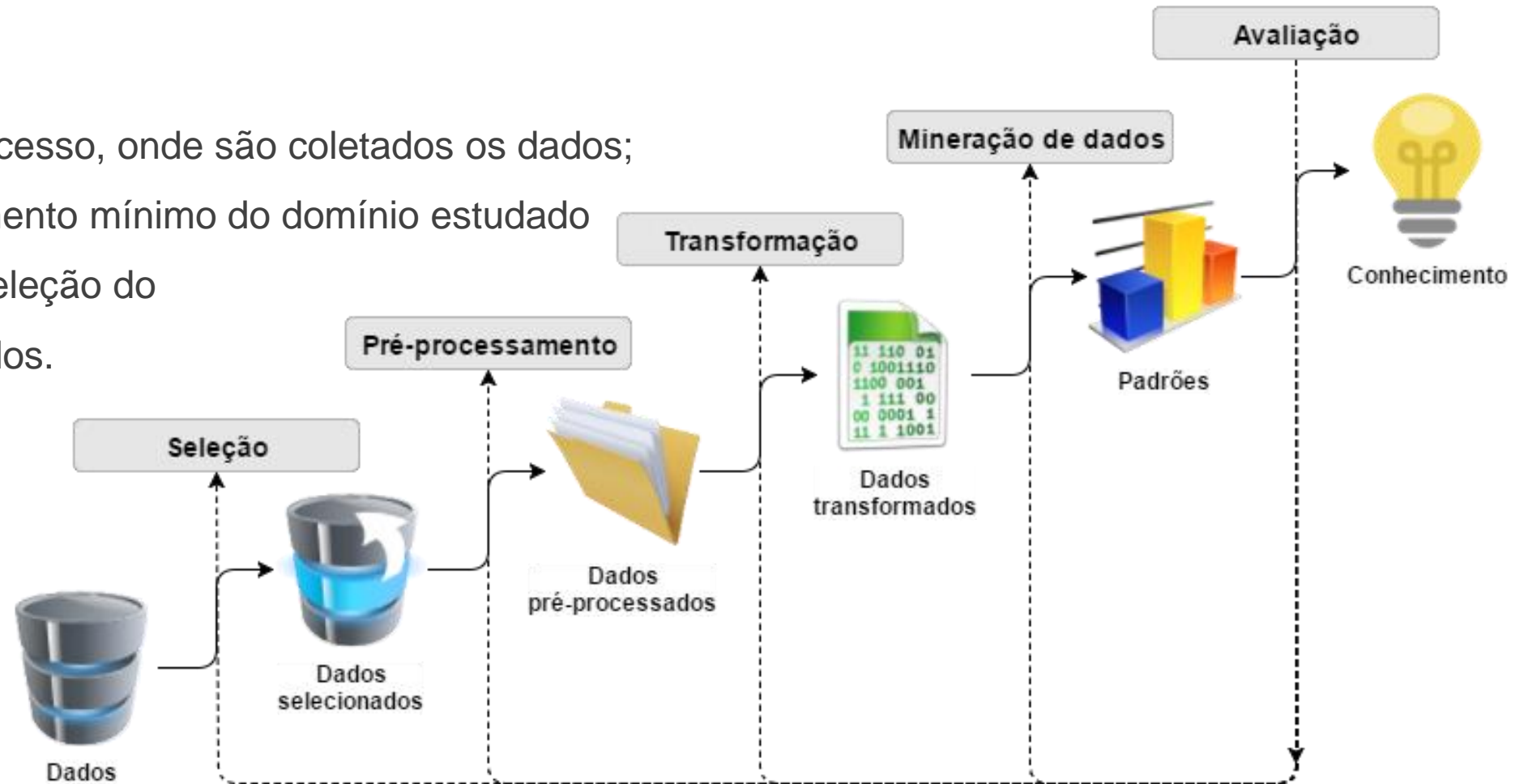


Fonte: Adaptado de Fayyad, Piatetsky-Shapiro e Smyth (1996).

Visão geral do processo de KDD

Seleção

- Etapa inicial do processo, onde são coletados os dados;
- Exige um conhecimento mínimo do domínio estudado para uma melhor seleção do subconjunto de dados.

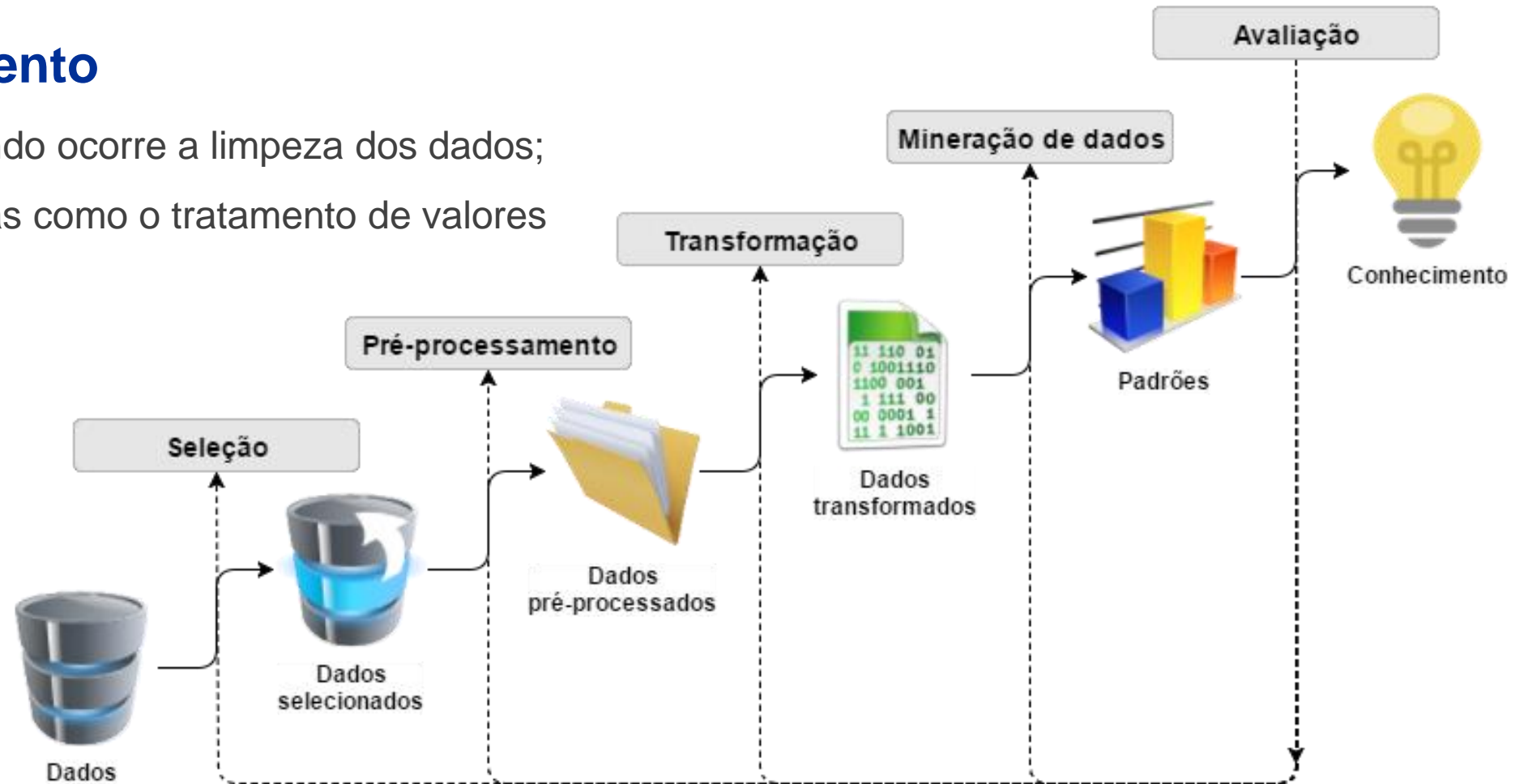


Fonte: Adaptado de Fayyad, Piatetsky-Shapiro e Smyth (1996).

Visão geral do processo de KDD

Pré-processamento

- Nesta etapa é quando ocorre a limpeza dos dados;
- Executando técnicas como o tratamento de valores ausentes.

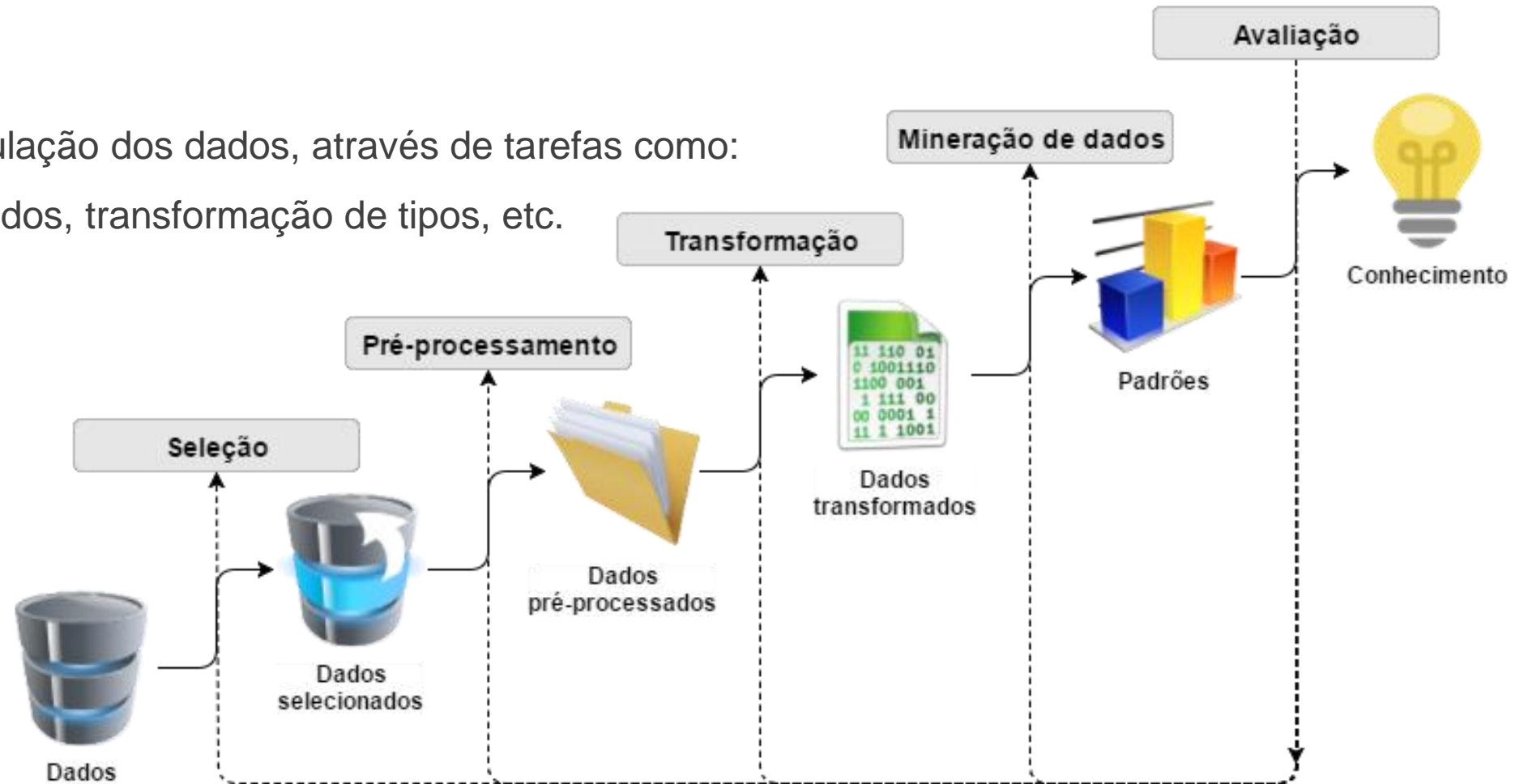


Fonte: Adaptado de Fayyad, Piatetsky-Shapiro e Smyth (1996).

Visão geral do processo de KDD

Transformação

- Consiste na manipulação dos dados, através de tarefas como: agrupamento de dados, transformação de tipos, etc.

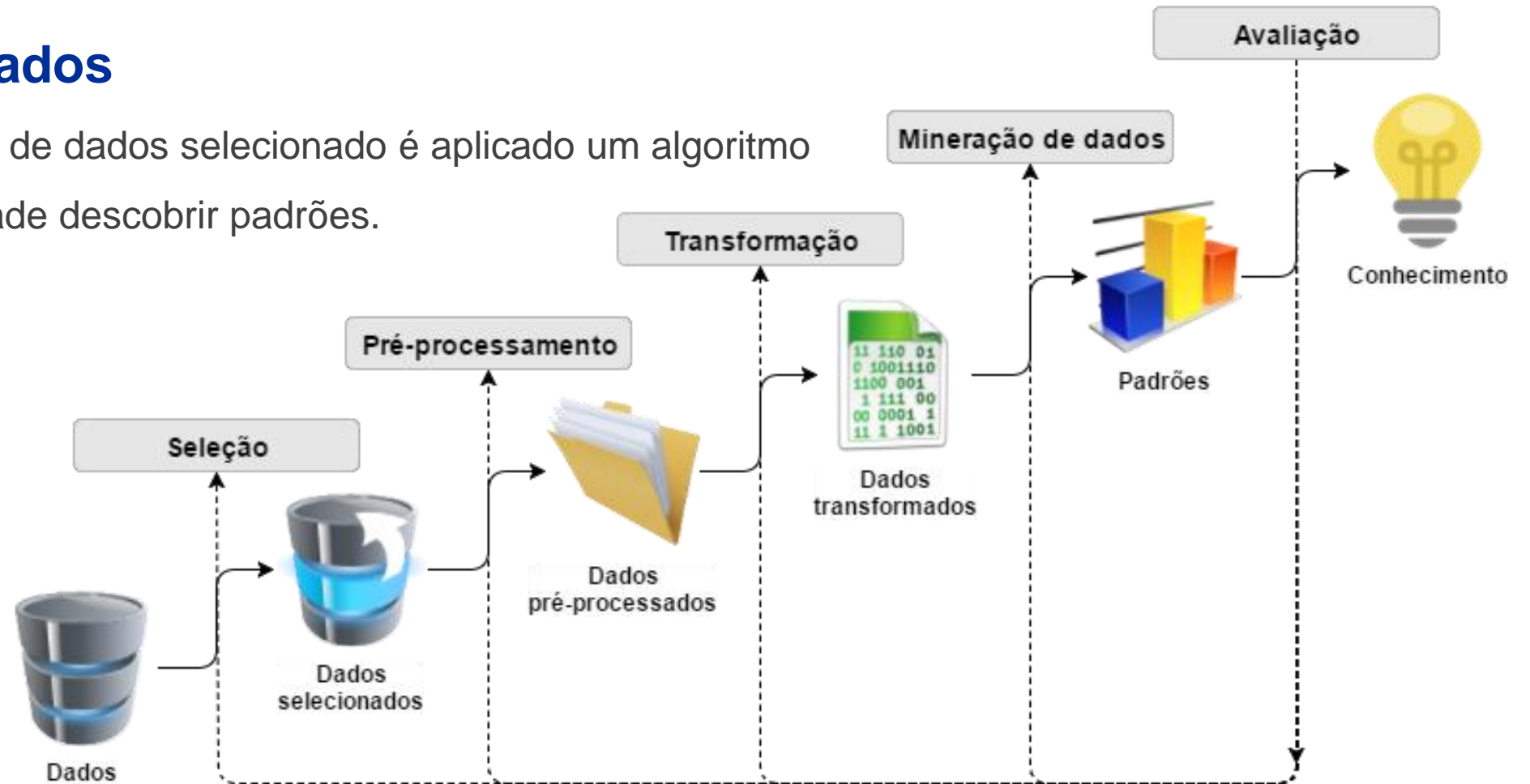


Fonte: Adaptado de Fayyad, Piatetsky-Shapiro e Smyth (1996).

Visão geral do processo de KDD

Mineração de dados

- A partir do conjunto de dados selecionado é aplicado um algoritmo que tem por finalidade descobrir padrões.

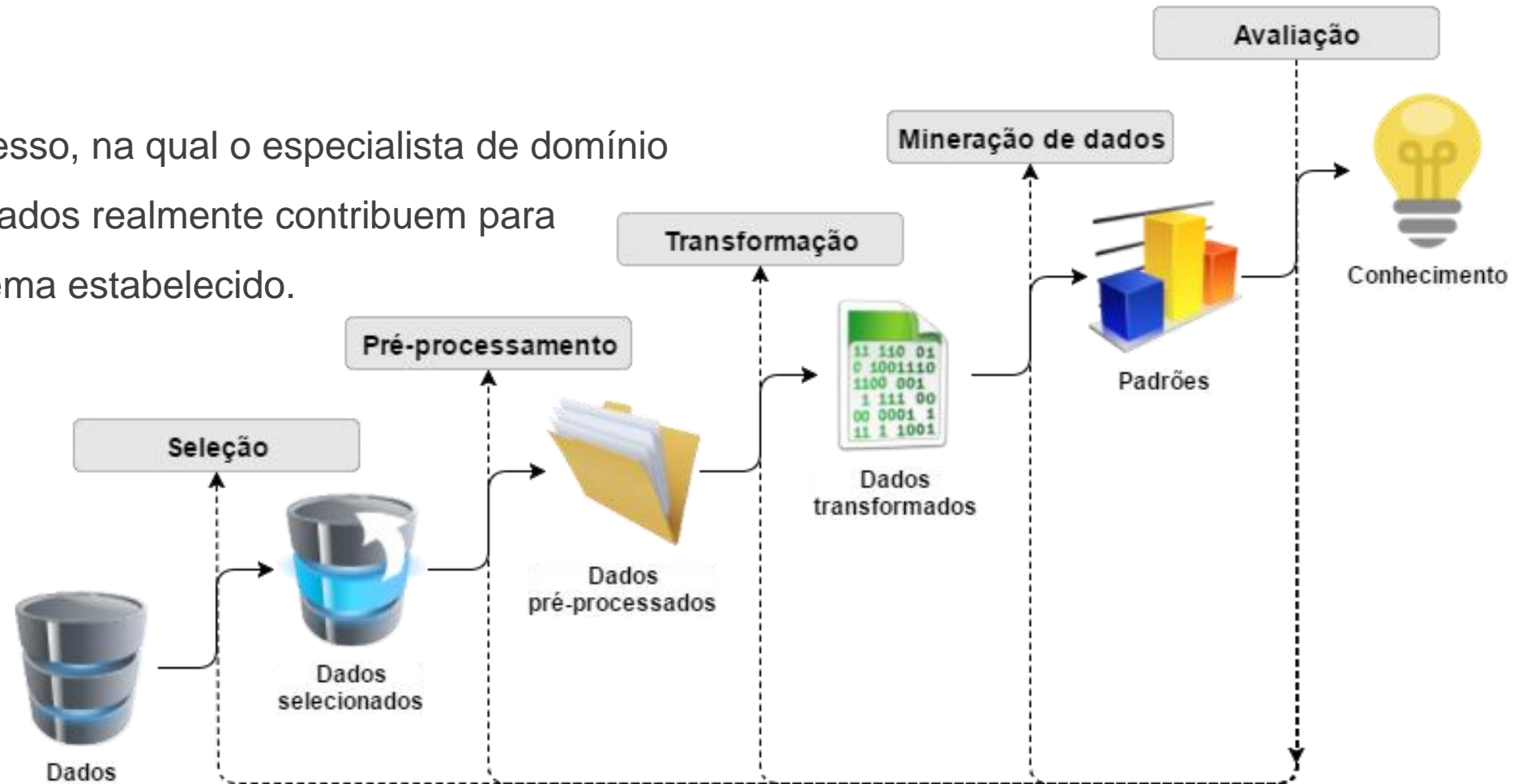


Fonte: Adaptado de Fayyad, Piatetsky-Shapiro e Smyth (1996).

Visão geral do processo de KDD

Avaliação

- Etapa final do processo, na qual o especialista de domínio verifica se os resultados realmente contribuem para a solução do problema estabelecido.

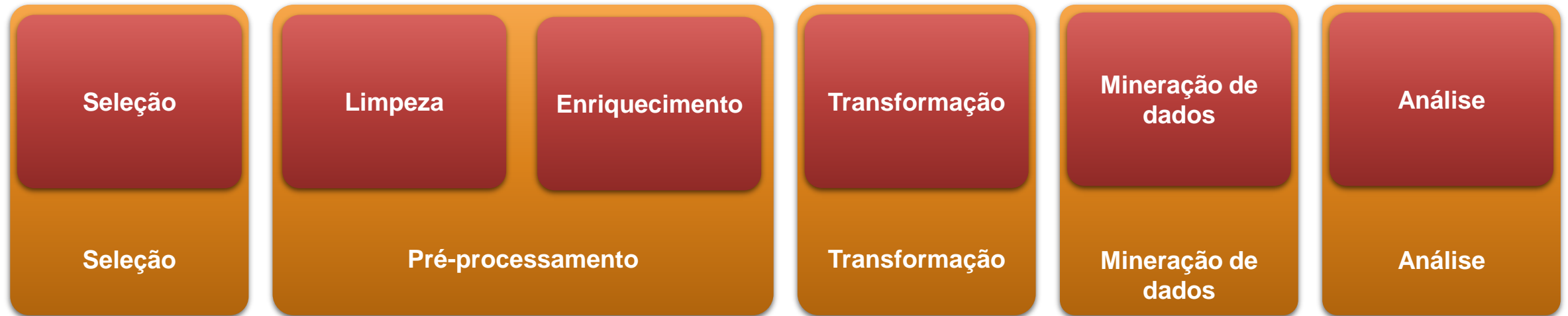


Fonte: Adaptado de Fayyad, Piatetsky-Shapiro e Smyth (1996).

Visão geral do processo de KDD



Visão geral do processo de KDD

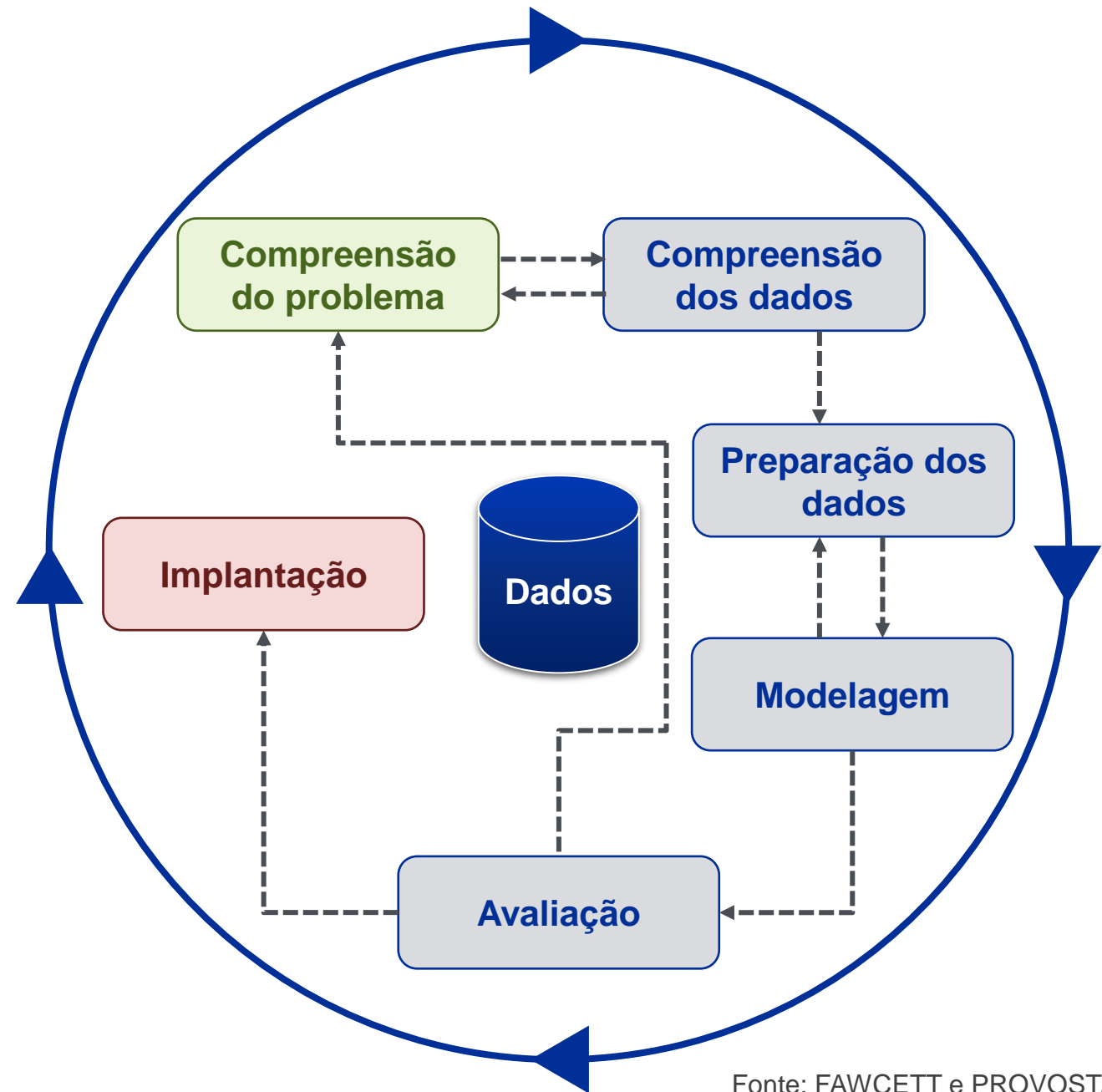


Visão geral do processo de KDD



Processo de Mineração de Dados

Abordagem com foco no negócio/problema



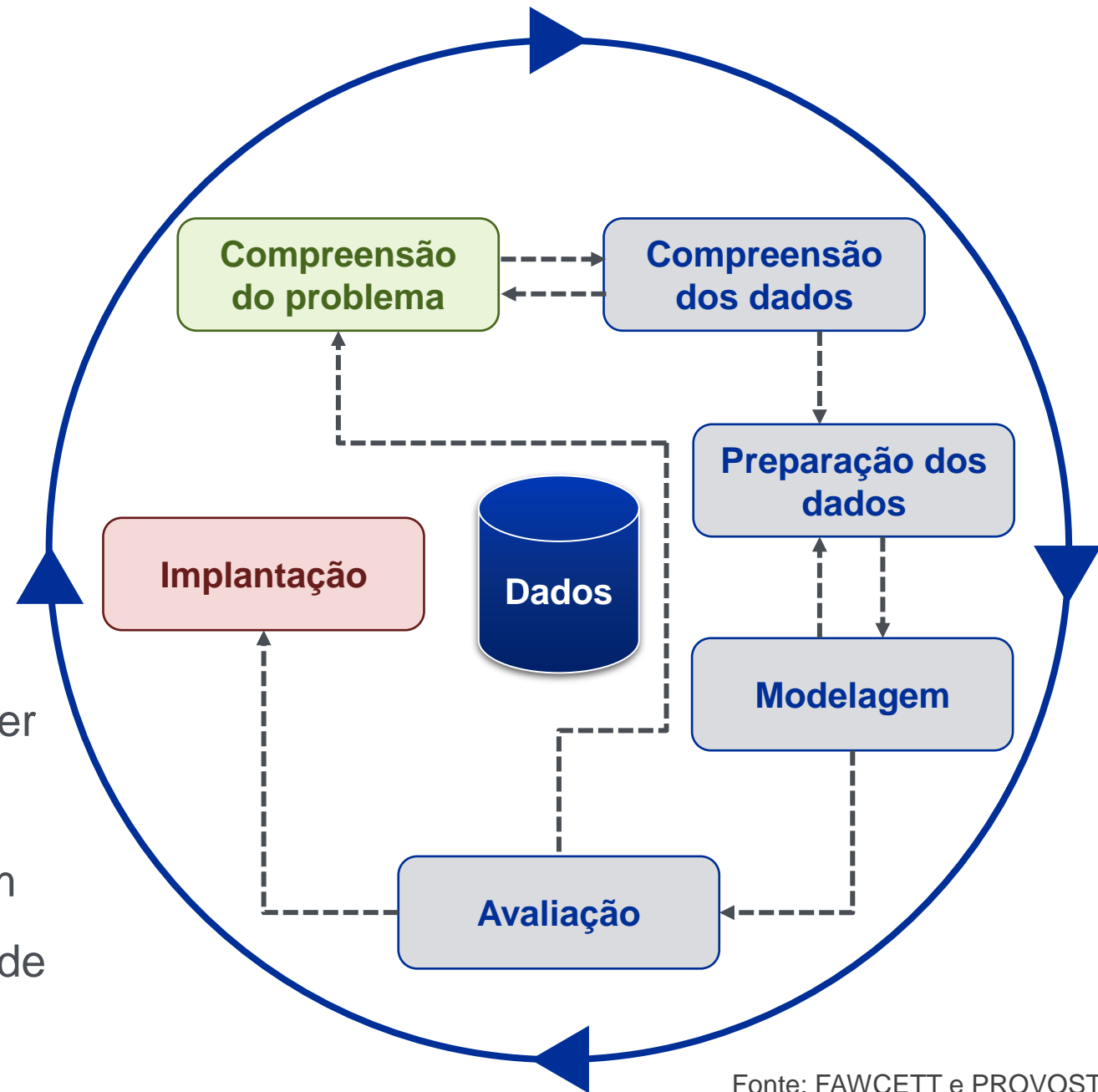
Fonte: FAWCETT e PROVOST, 2016.

Processo de Mineração de Dados

Abordagem com foco no negócio/problema

Compreensão do problema

É vital compreender o problema a ser resolvido. Pode parecer óbvio, mas projetos de negócios raramente vêm modelados como problemas claros de mineração de dados.



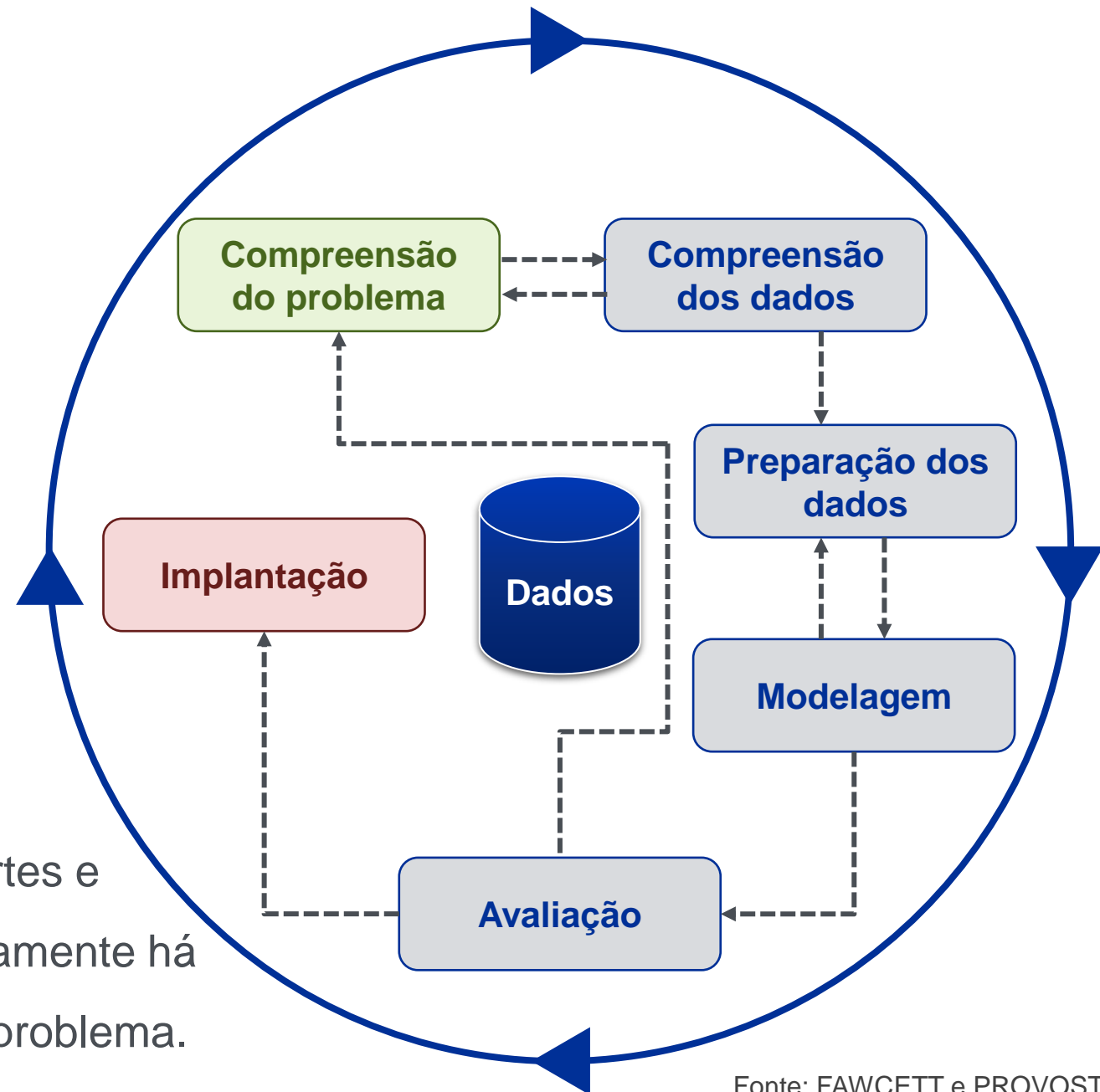
Fonte: FAWCETT e PROVOST, 2016.

Processo de Mineração de Dados

Abordagem com foco no negócio/problema

Compreensão dos dados

É importante entender os pontos fortes e as limitações dos dados porque raramente há uma correspondência exata com o problema.



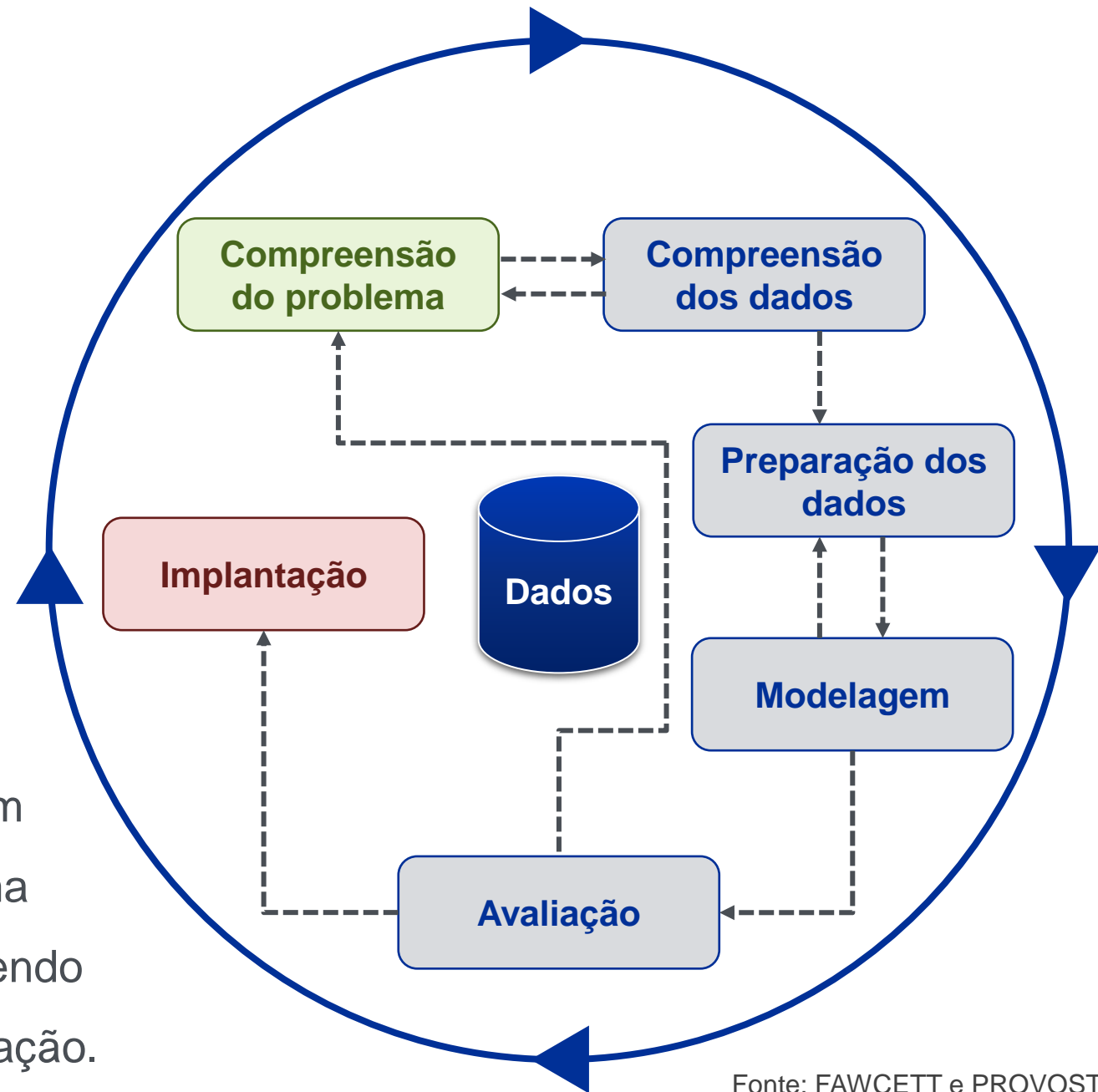
Fonte: FAWCETT e PROVOST, 2016.

Processo de Mineração de Dados

Abordagem com foco no negócio/problema

Preparação dos dados

As ferramentas normalmente exigem que os dados estejam em uma forma diferente de como são coletados, sendo necessário algum tipo de transformação.



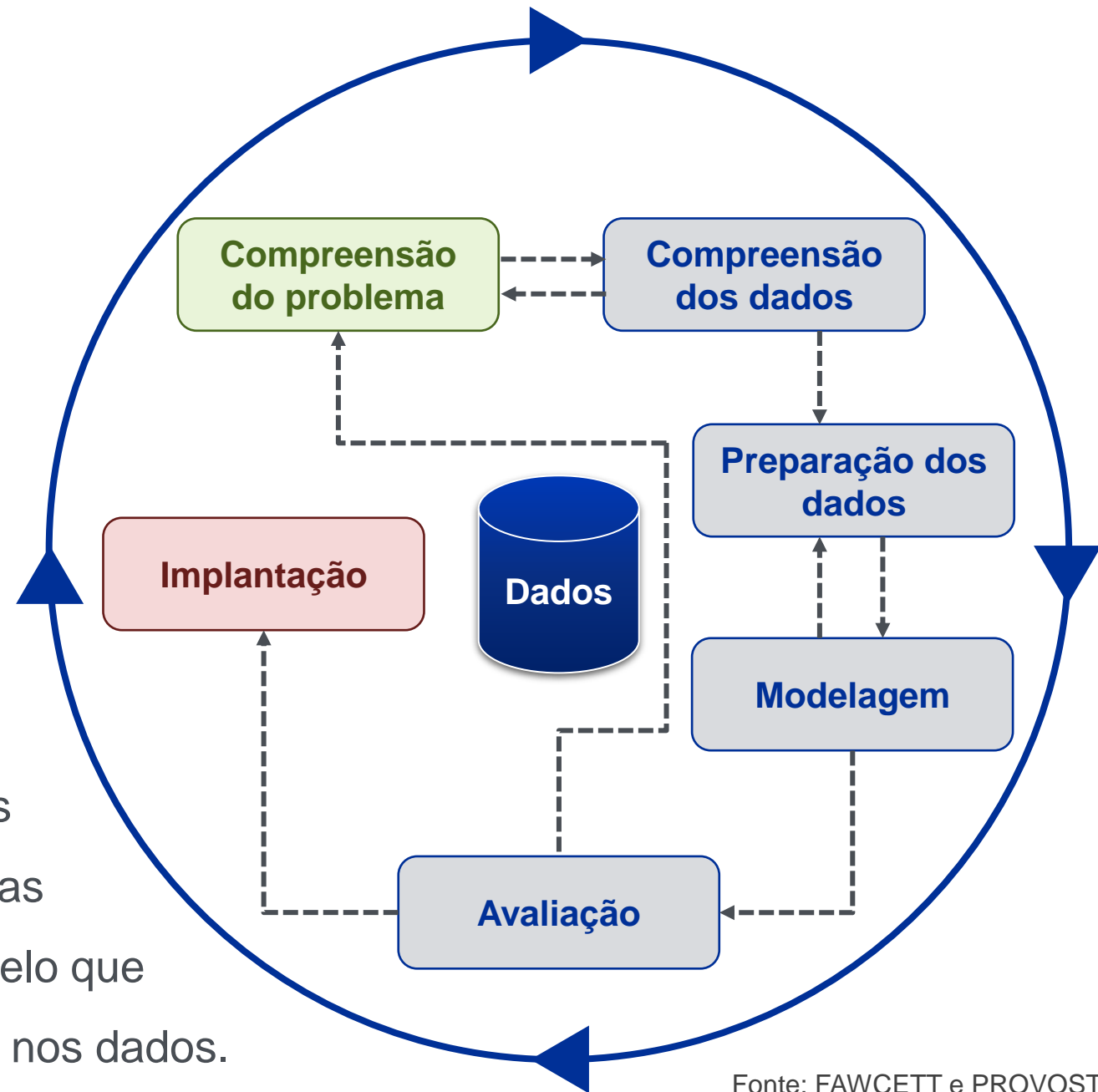
Fonte: FAWCETT e PROVOST, 2016.

Processo de Mineração de Dados

Abordagem com foco no negócio/problema

Modelagem

Etapa do processo onde as técnicas de mineração de dados são aplicadas com o objetivo de construir um modelo que representa os padrões identificados nos dados.



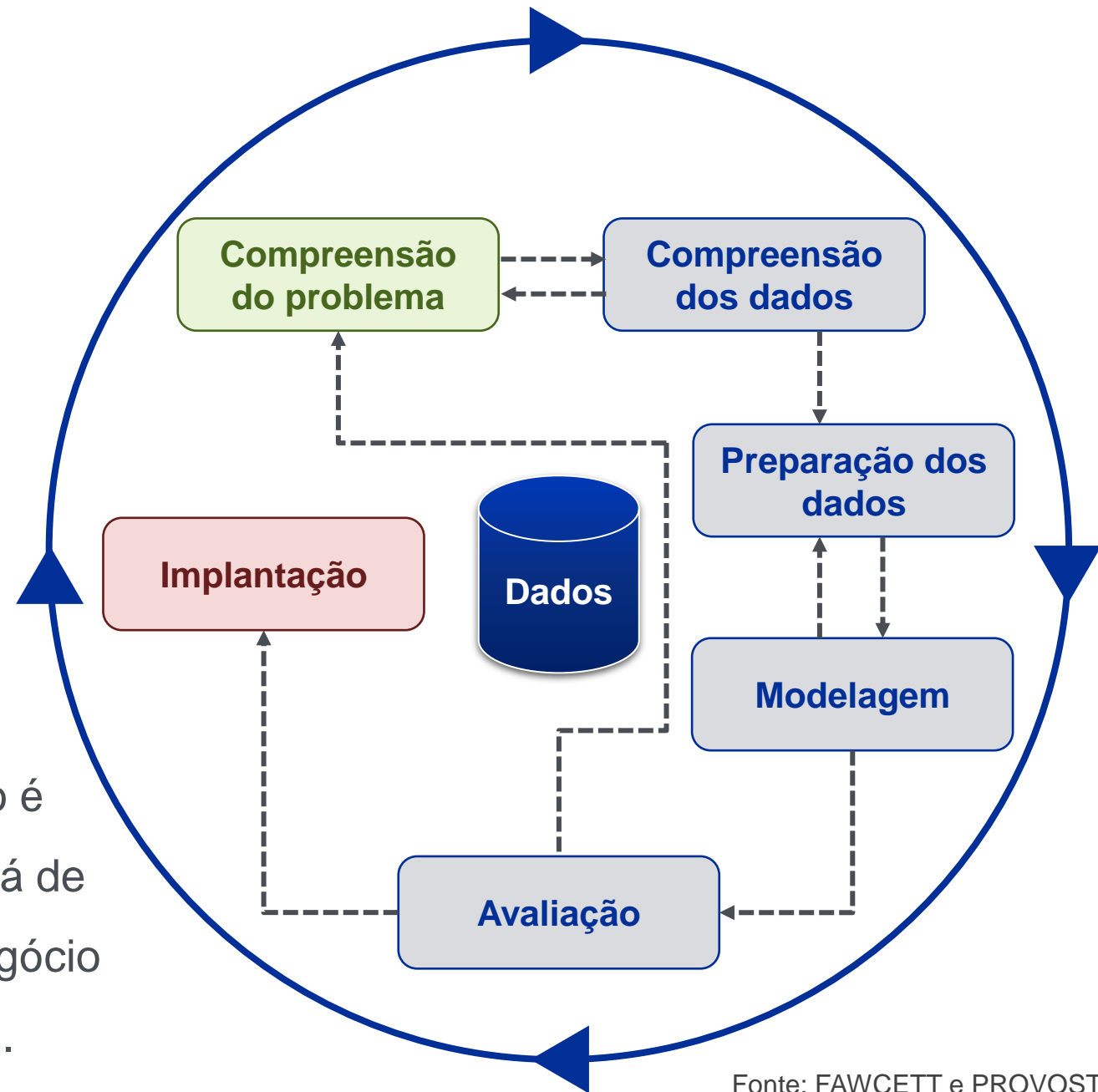
Fonte: FAWCETT e PROVOST, 2016.

Processo de Mineração de Dados

Abordagem com foco no negócio/problema

Avaliação

Fase onde é verificado se a solução é válida e confiável, bem como se está de acordo com as necessidades do negócio (avaliação quantitativa e qualitativa).



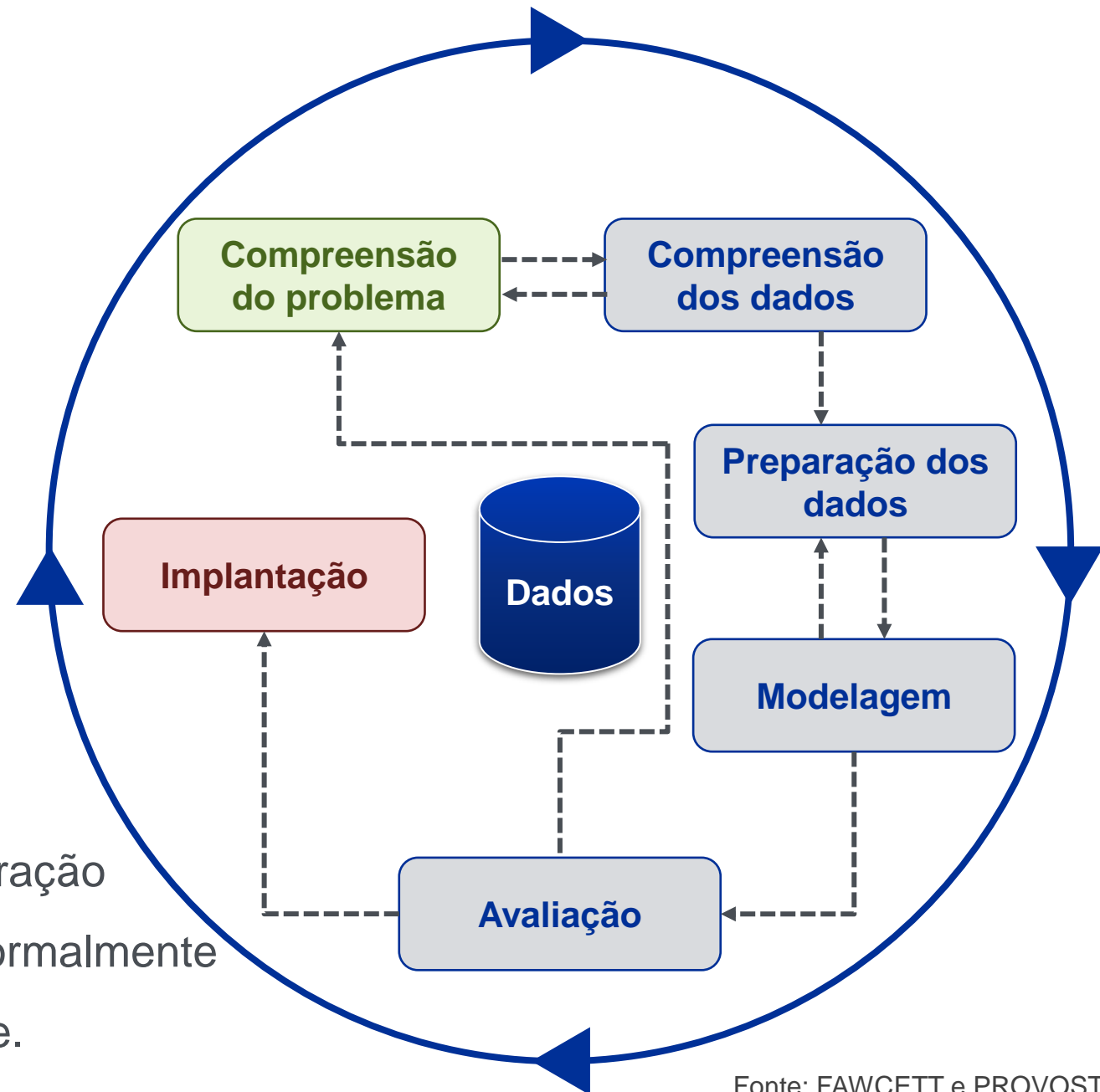
Fonte: FAWCETT e PROVOST, 2016.

Processo de Mineração de Dados

Abordagem com foco no negócio/problema

Implantação

Os resultados do processo de mineração de dados são colocados em uso, normalmente por meio de um sistema de software.



Fonte: FAWCETT e PROVOST, 2016.

Tarefas de mineração de dados

- **Preditiva:**

- Tem por objetivo prever o valor de um atributo baseado nos valores de outros atributos;
- É aplicada quando se deseja conhecer o comportamento futuro de novas instâncias de dados;
- Exemplo: **Classificação, Regressão**;

- **Descritiva:**

- O objetivo é derivar padrões, encontrando relações nos dados analisados;
- São utilizadas quando se deseja apenas apresentar os dados de uma forma compreensível;
- Exemplo: **Agrupamento, Regras de Associação**.

Tipos de aprendizado de máquina



Supervisionado

- Dados com rótulos (saída é conhecida).
- “Podemos identificar grupos de clientes que tenham probabilidades elevadas de não renovar seus contratos?”



Não-Supervisionado

- Dados sem rótulos (saída não é conhecida).
- “Nossos clientes naturalmente se encaixam em grupos diferentes?”



Semi-Supervisionado

- Combina supervisionado e não-supervisionado.



Por reforço

- Aprende com os erros.
- Baseado em recompensa e punição: associa o que gera maior recompensa.



Felipe Santana • 2º

Cientista de Dados Sênior | Machine Learning | Python | Deep Learn...

2 sem •

+ Seguir

Foco na solução do problema

https://www.linkedin.com/posts/felipesf_datascience-trabalho-carreira-activity-6949350306933579776-Xp-M/

Estes são erros comuns de iniciantes (e até avançados) em Data Science.

Adianto que já cometi alguns, principalmente o terceiro.

1. Entender de forma superficial o requisito de negócio.
2. Menosprezar a etapa de análise exploratória de dados e ir direto para experimentação com algoritmos.
3. Subestimar técnicas simples até quebrar a cara e voltar com o rabo entre as pernas. Esse é classico, que atire a primera pedra quem nunca.

4. Se apegar a ferramentas e tecnologias ao invés de focar na solução do problema.

5. Tomar decisão puramente técnica sem levar em consideração a experiência do usuário.
6. Não investigar de forma minuciosa os resultados do modelo e alinhar com as métricas de negócio.
7. Não saber se comunicar e apresentar de forma clara e sucinta o seu trabalho.

Foco na solução do problema

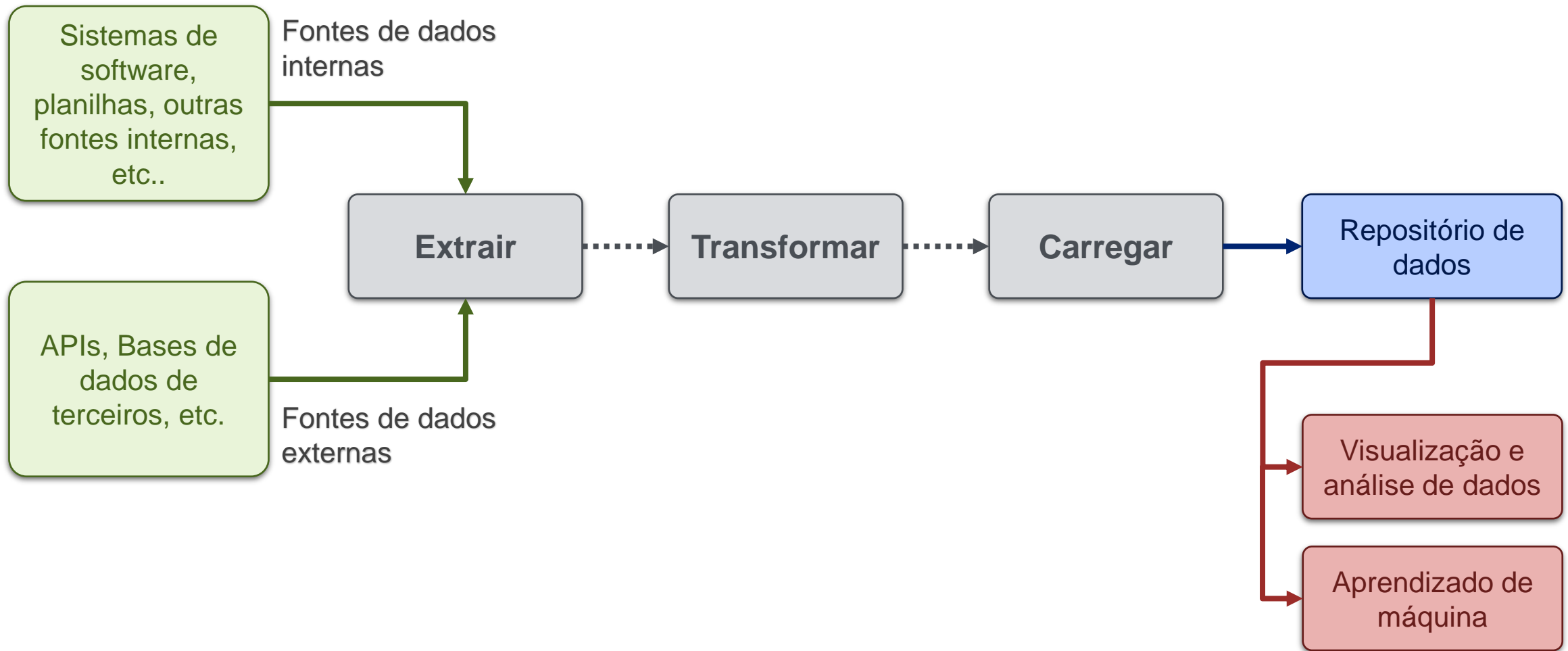
- E quanto ao nosso **SGCM**?
 - Qual é o percentual de **pacientes ausentes** (no-show)?
 - Como podemos obter essa informação?
 - Como monitorar o problema?
 - Podemos saber com antecedência se um paciente tem mais ou menos chances de não comparecer a consulta?
 - Como esse problema pode ser modelado?

Tratamento e visualização dados

Integração de dados e processo de ETL

- **Integração de dados** consiste em reunir dados de diferentes origens para dar suporte ao processo de tomada de decisão, seja por meio da **análise e visualização de dados**, ou mesmo para construção de modelos de **aprendizado de máquina**.
- Uma das tecnologias que permitem a integração de dados é o processo de ETL:
 - **Extração**: leitura dos dados a partir de diferentes fontes;
 - **Transformação**: conversão dos dados extraídos para um formato novo;
 - **Carga**: colocar os dados em um novo espaço de armazenamento, para ser utilizado em outras etapas do processo de tomada de decisão guiada por dados.

O processo de ETL



Processamento analítico *online* – OLAP

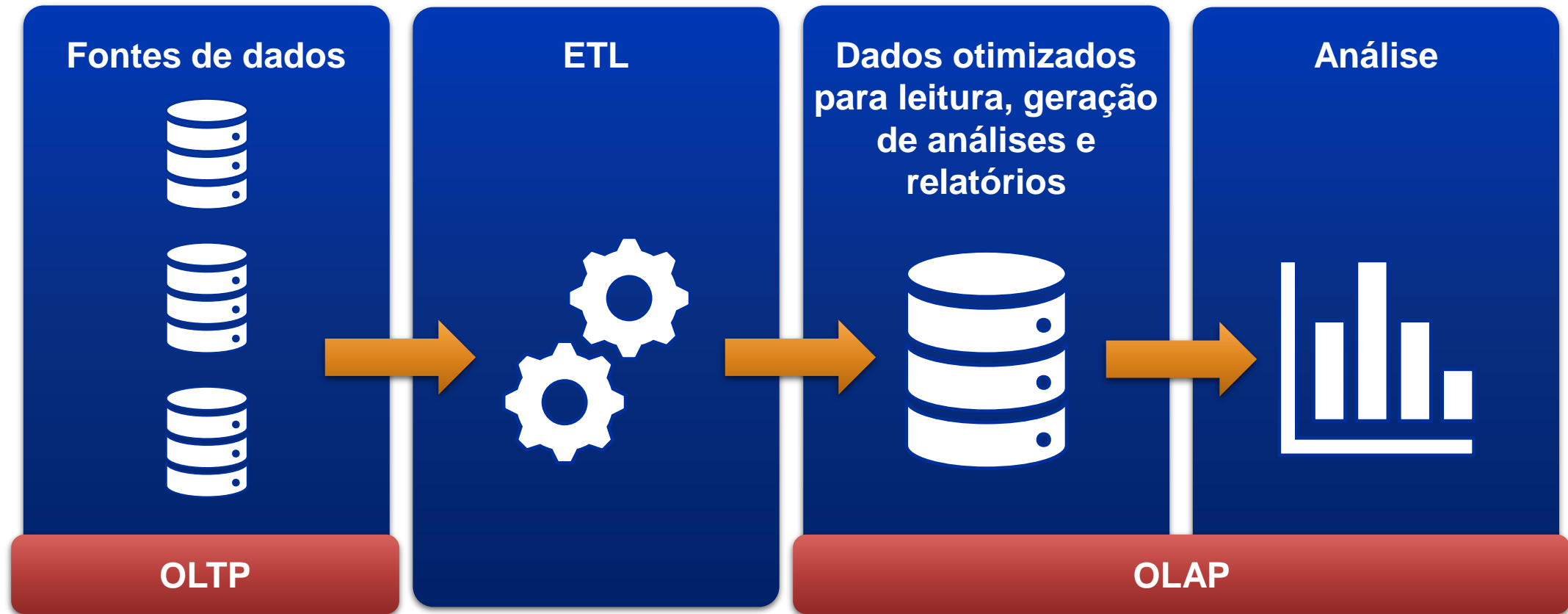
- O termo faz referência a uma **variedade de atividades voltadas à análise de dados**, normalmente executadas por usuário finais, o que pode incluir:
 - Geração de consultas;
 - Solicitação de relatórios e gráficos de rotina e *ad hoc*;
 - Realização de análises estatísticas;
 - Construção de apresentações visuais.



OLAP *versus* OLTP

- Durante muito tempo o foco era o **processamento de transações (OLTP)**, normalmente baseados em sistemas que utilizam bases de dados relacionais.
- **OLTP** é voltado para o processamento de transações repetitivas em grandes quantidades (leitura, inserção, modificação e exclusão).
- **OLAP** foca em relacionamentos complexos e na busca de padrões e tendências (diretamente relacionado com o suporte à decisão);
- **OLTP** geralmente envolve **normalização de dados**, o que pode afetar o desempenho nas operações de leitura, que é o foco do **OLAP**.

OLAP *versus* OLTP



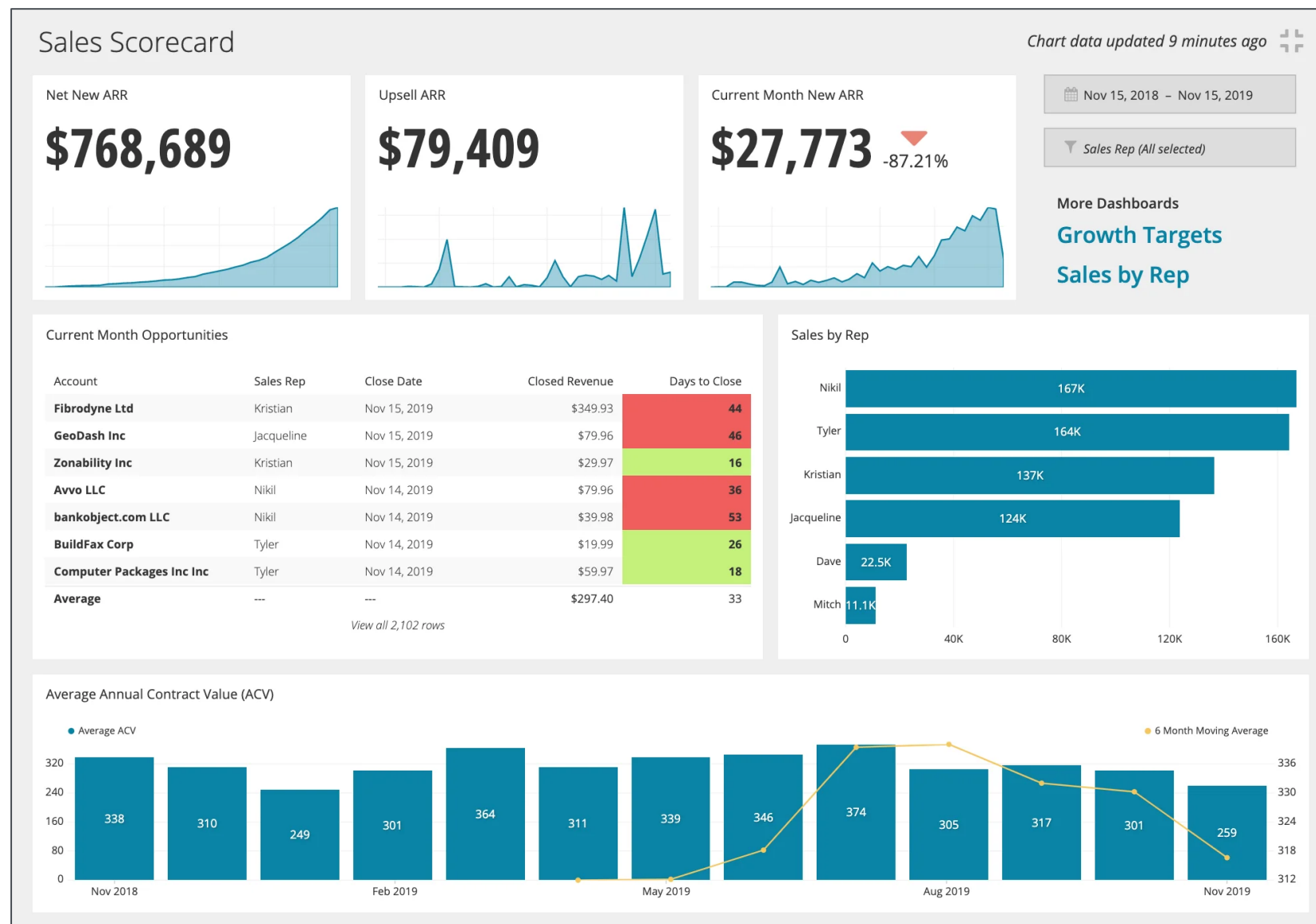
Talend Open Studio for Data Integration

- Ferramenta ETL para integração de dados, baseada na IDE Eclipse.
- A ferramenta gera um aplicação Java, mas a maior parte do recursos exige apenas operações de arrastar e soltar.
- Suporte a múltiplas fontes de dados: BDs relacionais, Serviços de nuvem, APIs, Big Data, etc.
- Permite que a aplicação seja compilada e executada de forma independente.
 - <https://www.datalytyx.com/scheduling-talend-open-studio-jobs-in-windows-without-talend-administration-center-tac/>
- Tutorial: <https://www.javatpoint.com/talend>

Visualização de dados

- Qualquer representação gráfica, animada ou estática, utilizada para apresentação de dados.
- Fornece capacidade para identificar rapidamente as tendências importantes nos dados.
- Método tradicional: visualização por planilhas.
- Métodos mais modernos: *dashboards* e indicadores.

Visualização de dados: dashboard



Métricas e indicadores de desempenho

- Dashboards são formados principalmente por métricas e indicadores de desempenho (**KPI** – *Key Performance Indicator*).
- Uma métrica é qualquer dado que permita mensurar algum aspecto do negócio: quantidade de clientes, atendimentos, vendas, etc.
- KPIs fazem referência aos resultados de ações e acompanhamento de metas.
- Exemplo:
 - Métrica: quantidade de clientes.
 - KPI: percentual de clientes satisfeitos.

Google Data Studio

- Permite visualizar dados com gráficos e tabelas configuráveis.
- Suporte a conexão com várias fontes de dados: BigQuery, MySQL, PostgreSQL, Planilhas, CSV, dentre outros.
- Recursos de compartilhamento e colaboração.
- Tutorial para criar um relatório:
 - https://support.google.com/datastudio/topic/6289358?hl=pt-BR&ref_topic=9170843

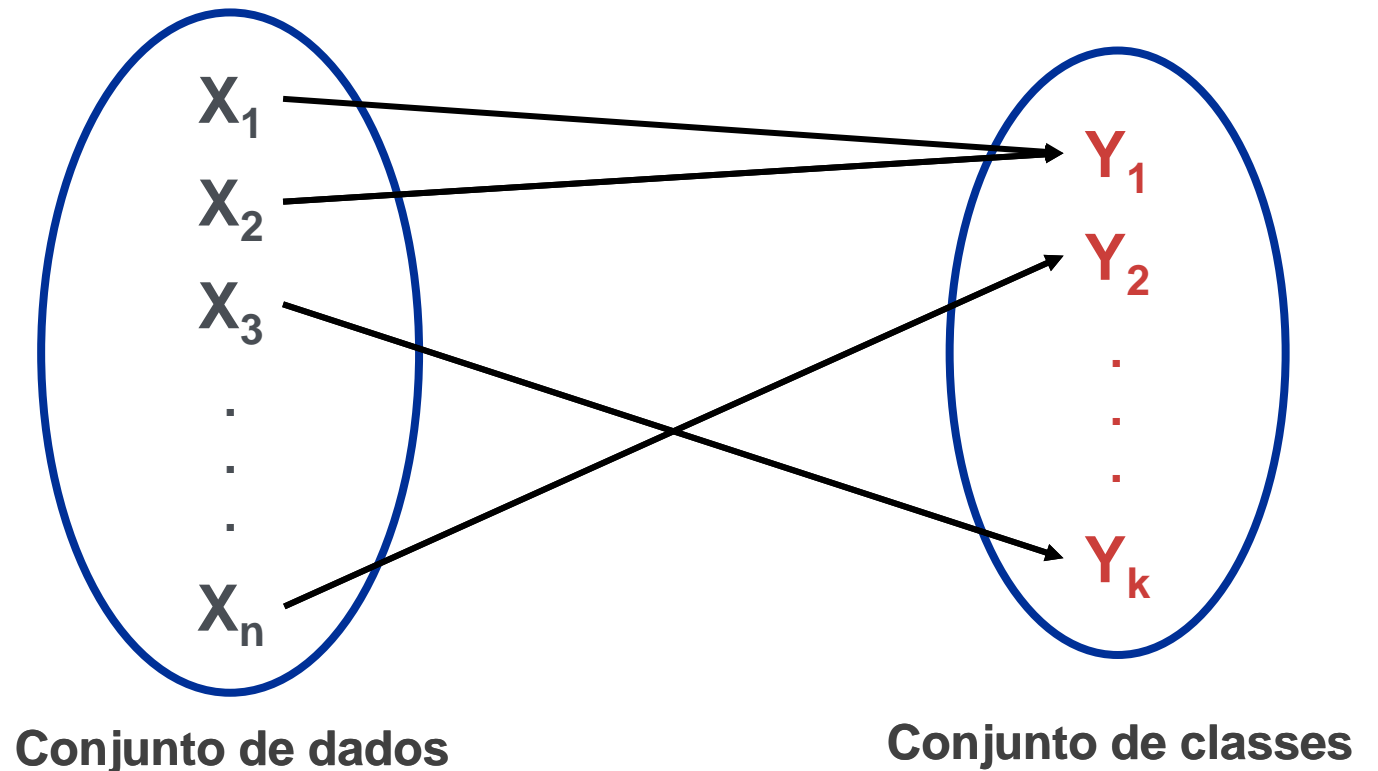
Construção de modelos preditivos

Introdução aos modelos preditivos

- **Classificação**: identificação da classe a qual um elemento pertence a partir de suas características;
- O conjunto de possíveis classes é **categórico** (discreto e não ordenado);
- Em outras palavras, a **classificação** mapeia, dentro de um conjunto de classes pré-definido, um **conjunto de dados de entrada** em uma **classe de saída**, estabelecendo uma função que **tem por objetivo determinar a qual classe pertence** um dado elemento do conjunto de dados;

Introdução aos modelos preditivos

- A classificação busca por uma função que permita associar corretamente cada registro X_i de um conjunto de dados a um único rótulo categórico, Y_j , denominado classe;
- Uma vez identificada, esta função pode ser aplicada a novos registros de forma a prever as classes em que tais registros se enquadram.



Introdução aos modelos preditivos

- Os atributos do conjunto de dados são divididos em dois grupos:
 - Um dos grupos contém somente um atributo (categórico), que corresponde ao **atributo dependente** (atributo **classe** / **atributo-alvo**), ou seja, o atributo para o qual se deve fazer a predição de um valor;
 - O outro grupo contém os atributos a serem utilizados na predição do valor, denominados **atributos independentes** (atributos de **predição** / atributos **previsores**).

Atributos independentes			Atributo dependente
Salário	Idade	Emprego	Classe
3.000,00	30	Autônomo	B
4.000,00	35	Indústria	B
7.000,00	50	Pesquisa	C
6.000,00	45	Autônomo	C
7.000,00	30	Pesquisa	B
6.000,00	35	Indústria	B
6.000,00	35	Autônomo	A
7.000,00	30	Autônomo	A
4.000,00	45	Indústria	B

O processo de mineração de um modelo de classificação

- O processo é dividido em duas etapas.
 - **Treino**: um modelo de classificação é construído utilizando um subconjunto de dados;
 - **Teste**: o modelo é utilizado para prever as classes de um subconjunto de dados distinto daquele utilizado na etapa anterior;
- Desta forma, parte do conjunto de dados é utilizado para **treino**, e outra parte para **teste**.
 - O subconjunto de dados de teste também pode estar separado do subconjunto de dados utilizado para treino.

O processo de mineração de um modelo de classificação

Conjunto de dados de treino

ID	Atributo1	Atributo2	Atributo3	Classe
1	Sim	Grande	125.000,00	Não
2	Não	Médio	100.000,00	Não
3	Não	Pequeno	70.000,00	Não
4	Sim	Médio	120.000,00	Não
5	Não	Grande	95.000,00	Sim
6	Não	Médio	60.000,00	Não
7	Sim	Grande	220.000,00	Não
8	Não	Pequeno	85.000,00	Sim
9	Não	Médio	75.000,00	Não
10	Não	Pequeno	90.000,00	Sim



Algoritmo de
classificação



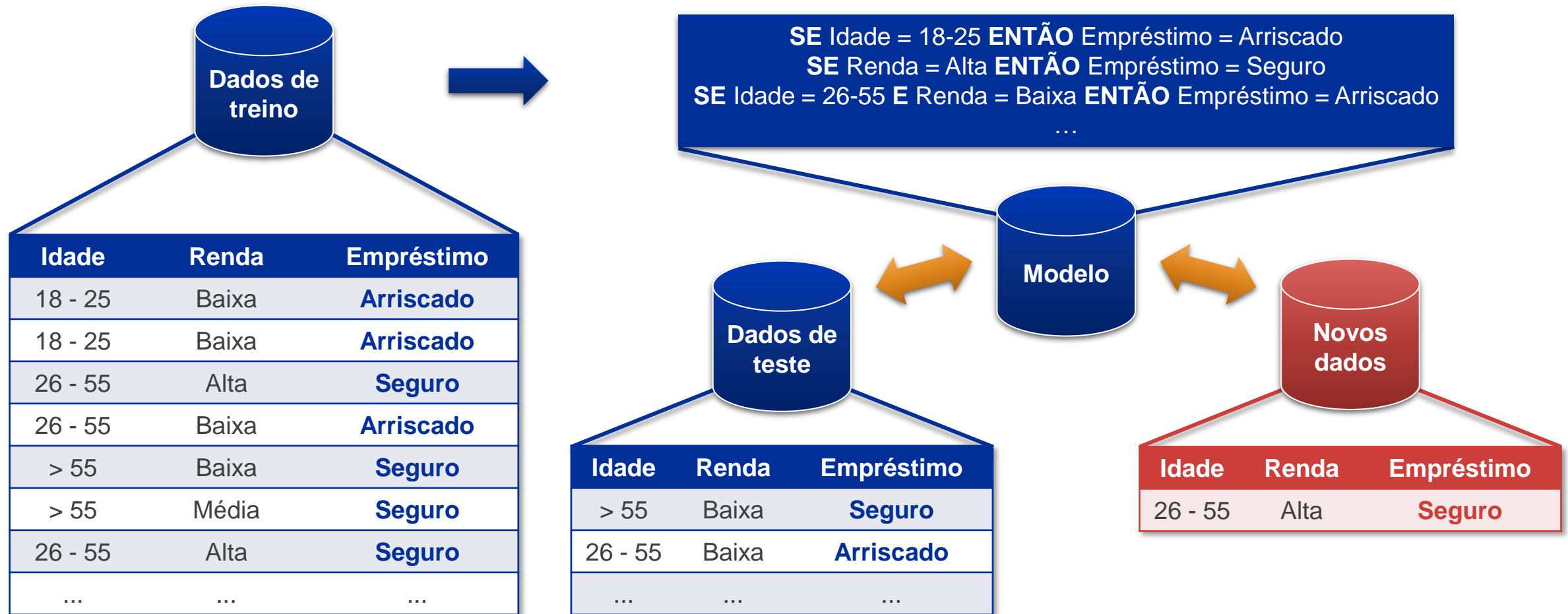
Modelo de
classificação



Conjunto de dados de teste

ID	Atributo1	Atributo2	Atributo3	Classe
11	Não	Pequeno	55.000,00	?
12	Sim	Médio	80.000,00	?
13	Sim	Grande	110.000,00	?
14	Não	Pequeno	95.000,00	?
15	Não	Grande	67.000,00	?

O processo de mineração de um modelo de classificação



Algoritmo de classificação

- Utilizado para construir o modelo de classificação.
 - **Entrada:**
 - Conjunto de dados de treinamento;
 - Atributos independentes;
 - Definição do atributo dependente (classe);
 - **Saída:**
 - Um modelo de classificação que permite classificar (definir o valor do atributo dependente) um novo registro a partir de seus atributos independentes.

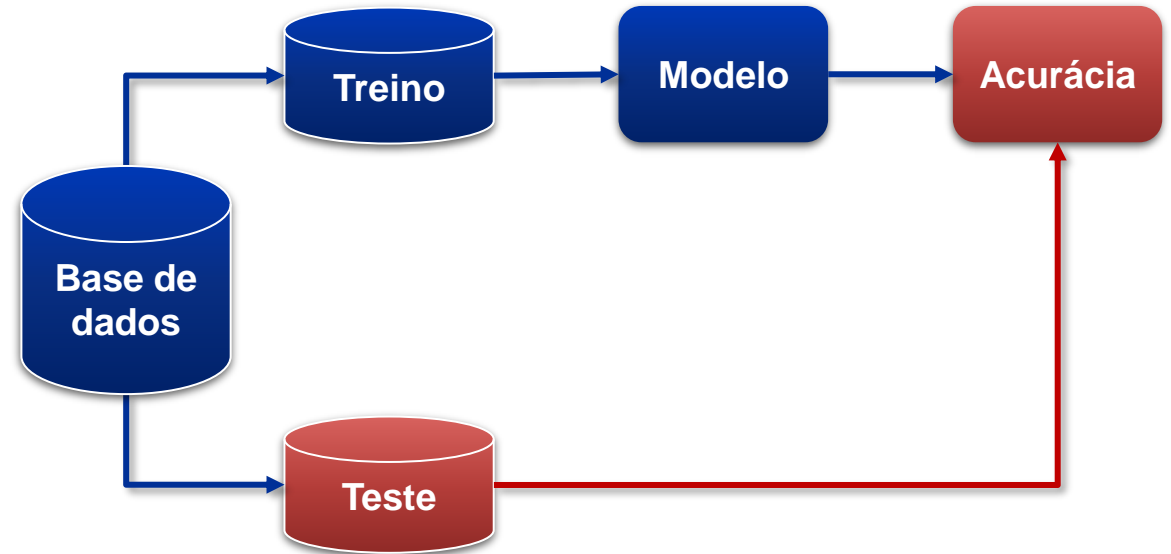
Avaliação de classificadores

- Dividir a base de dados em **Treino** e **Teste**.
- **Acurácia** (ou taxa de acerto) do classificador (classificações corretas):

$$Taxa\ de\ acerto = \frac{n^o\ acertos}{[base\ de\ teste]}$$

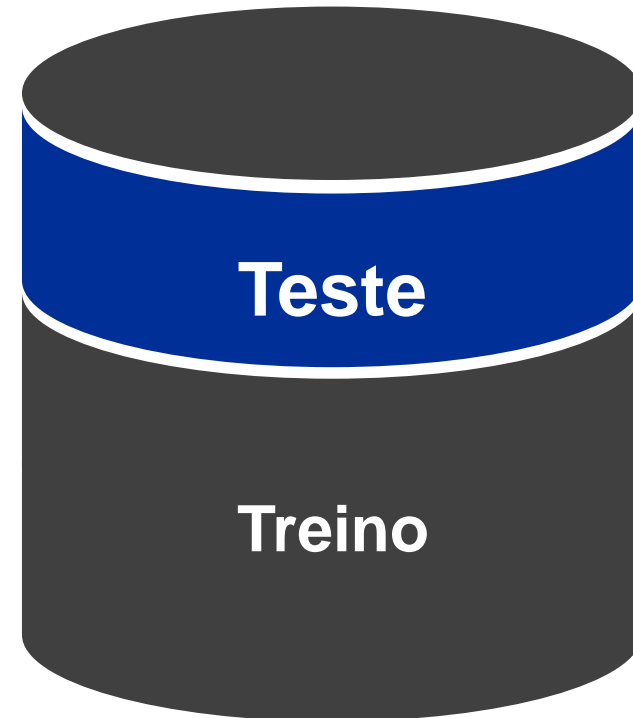
- **Taxa de erro** do classificador (classificações erradas):

$$Taxa\ de\ erro = \frac{n^o\ erros}{[base\ de\ teste]}$$



Avaliação de classificadores

- ***Hold out***
 - Divisão aleatória da base em:
 - Base de treinamento (2/3);
 - Base de teste (1/3).



Avaliação de classificadores

- ***k-Fold Cross Validation:***

- Base de dados particionada (aleatoriamente) em k partes (do mesmo tamanho aproximadamente);
- Treinamento e teste são executados k vezes;
- Em cada execução:
 - 1 partição de teste;
 - $k-1$ partições de treinamento;
- Todas as partições são utilizadas, em algum momento, para teste.

- $$Acurácia = \frac{n^o \text{ acertos}}{[base \text{ de dados}]}$$

Iteração 1	Teste	Treino	Treino	Treino	Treino
Iteração 2	Treino	Teste	Treino	Treino	Treino
Iteração 3	Treino	Treino	Teste	Treino	Treino
Iteração 4	Treino	Treino	Treino	Teste	Treino
Iteração 5	Treino	Treino	Treino	Treino	Teste

Avaliação de classificadores

$$Precision = \frac{TP}{TP + FP}$$

- TP: true positive;
- FP: false positive.

- Fração dos elementos classificados como positivos que são realmente positivos;
- Ou o número de acertos entre os elementos classificados como positivos;
- Exemplo: de todos classificados como SPAM, quantos eram realmente SPAM.

Avaliação de classificadores

$$Recall = \frac{TP}{TP + FN}$$

- TP: true positive;
- FN: false negative.

- Fração dos elementos positivos que foram classificados como positivos;
- Ou o número de acertos entre os elementos positivos;
- Exemplo: de todos os SPAMs, quantos foram classificados como SPAM.

Avaliação de classificadores

$$F\ Measure = \frac{2 \times Pr \times Rc}{Pr + Rc}$$

- Pr: precision;
- Rc: recall.

- **F** é a média harmônica entre **Pr** e **Rc**;
- Tende a ser mais próximo ao menor elemento;
- Um valor alto garante a “qualidade” em termos de **Precision** e **Recall**.

Avaliação de classificadores

- **Matriz de confusão:**
 - Permite visualizar em termos nominais o erros e acertos do modelo.

	Classe A	Classe B	
Classe A	10	15	Classe A
Classe B	5	20	Classe B

Seleção de atributos

- Processo de **reduzir o número de variáveis de entrada** (independentes) em modelos preditivos.
- Vantagens:
 - **Reduz o custo computacional** de construção do modelo;
 - **Simplifica o modelo**, facilitando a interpretação, bem com a implementação em produção.
- O objetivo é encontrar os atributos independentes que possuam **relacionamento mais forte** com o atributo dependente, por meio de técnicas como o coeficiente de correlação e ganho informação.

Classes desbalanceadas

- O problema de classes desbalanceadas ocorre quando há um **desequilíbrio na distribuição de instâncias entre as classes**.
 - Exemplo: um modelo para identificação de SPAMs que foi treinado com um conjunto de dados que tinha uma pequena quantidade relativa de SPAMs.
- O desbalanceamento pode resultar em **modelos enviesados**.
- Técnicas para lidar com o problema:
 - Coletar mais dados;
 - Gerar amostras sintéticas (SMOTE);
 - **Manipular as amostras (*oversampling e undersampling*)**.

Continua...