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 um 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.

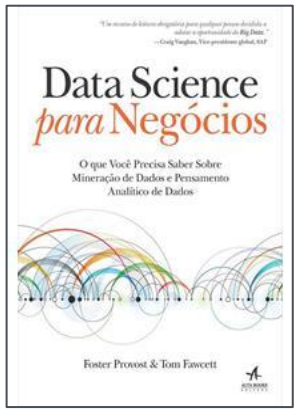
Construção de modelos preditivos

- 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

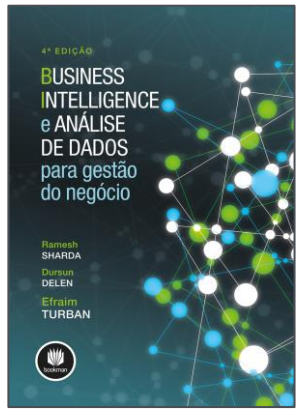
- Definição de uma estratégia para implantação de modelos preditivos.
- Utilização da WEKA API.
- Configuração do projeto.
- Classificação de novas instâncias.

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?itemName=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

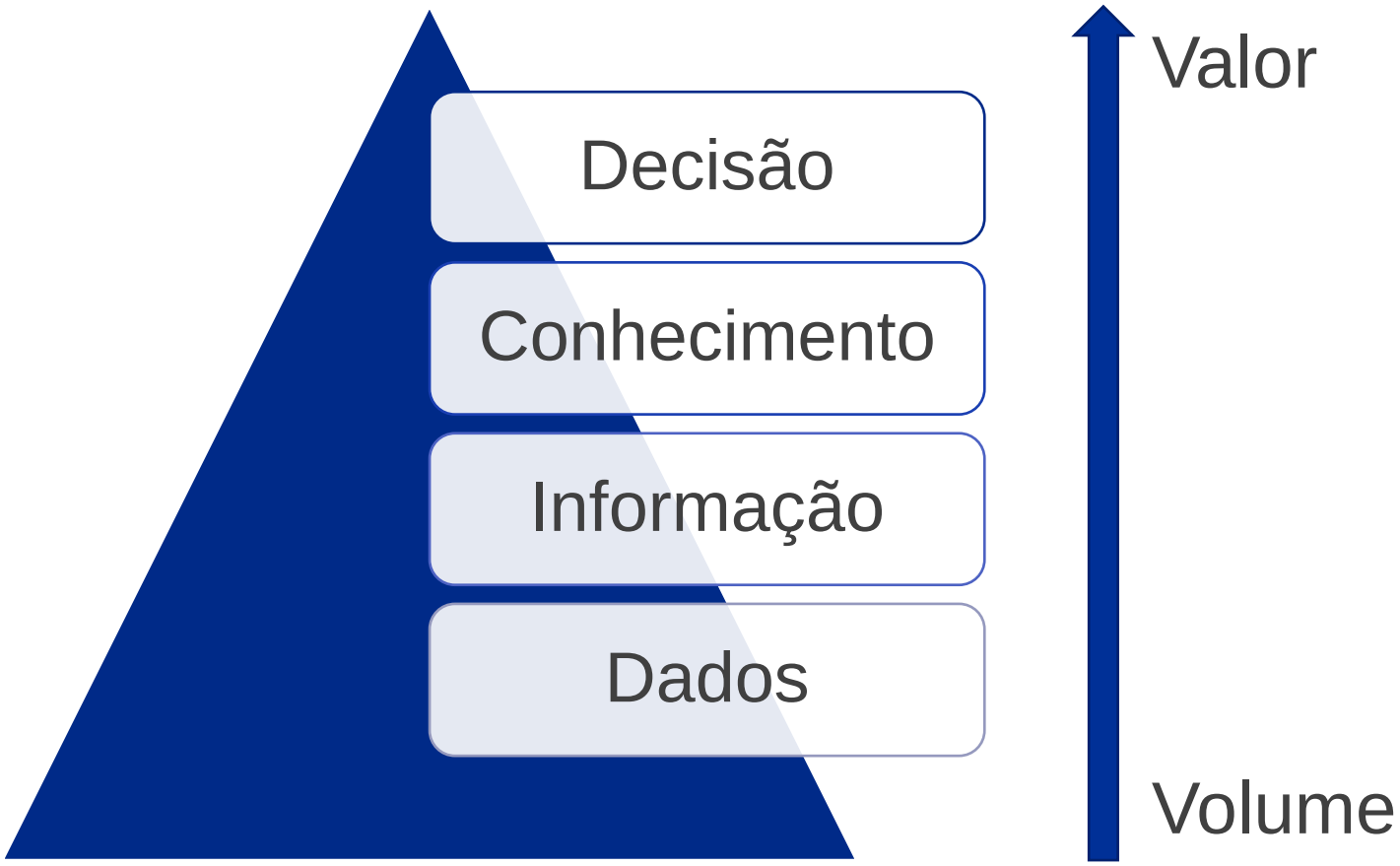
- **JDK 11**
  - https://www.oracle.com/br/java/technologies/javase/jdk11-archive-downloads.html
  - Criar a variável de ambiente JAVA_HOME configurada para o diretório de instalação do JDK. Exemplo: “C:\Program Files\Java\jdk-11.0.13”.
  - Adicionar “%JAVA_HOME%\bin” na variável de ambiente PATH.
  - 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.

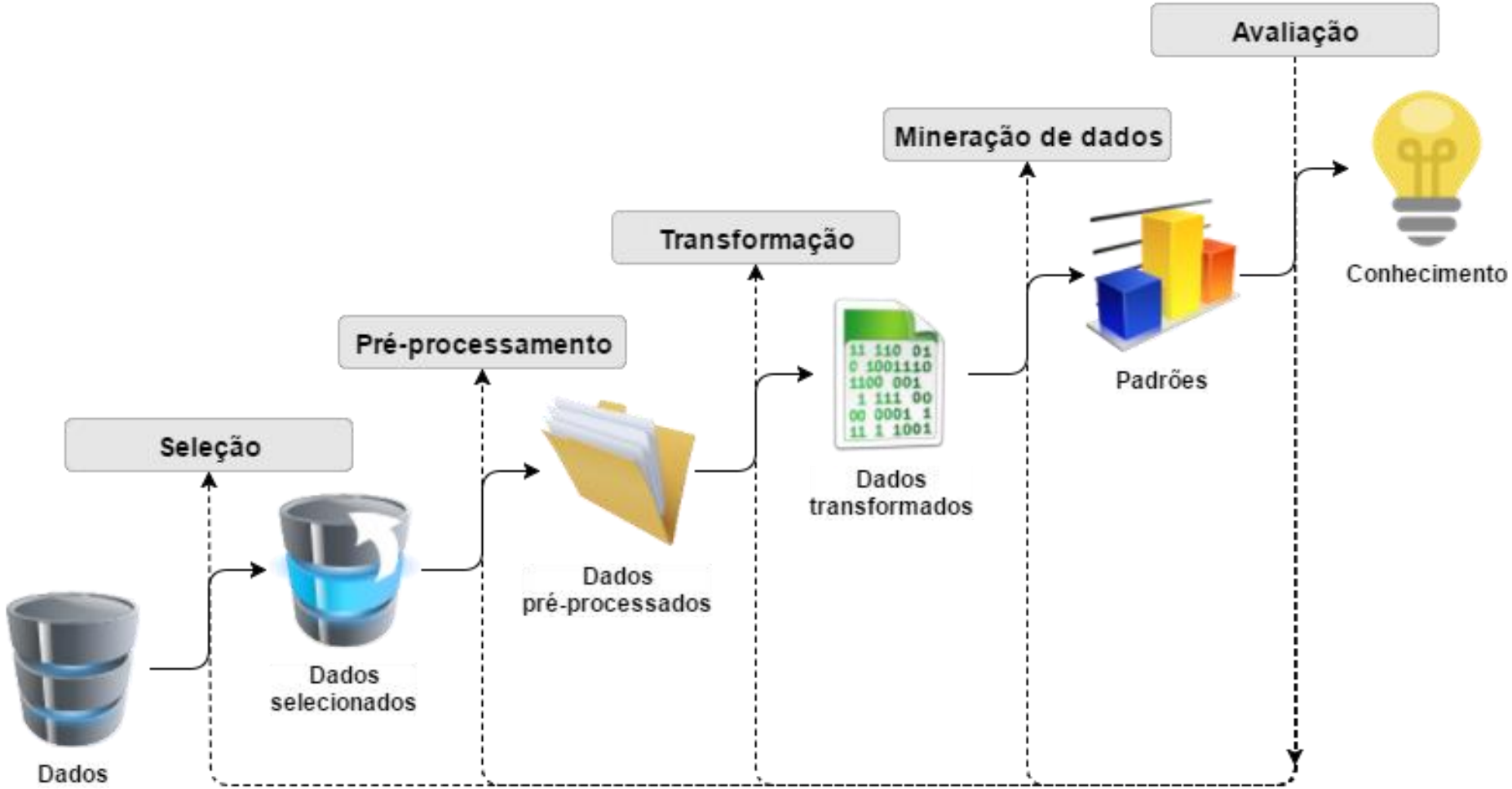
O processo de KDD

- **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

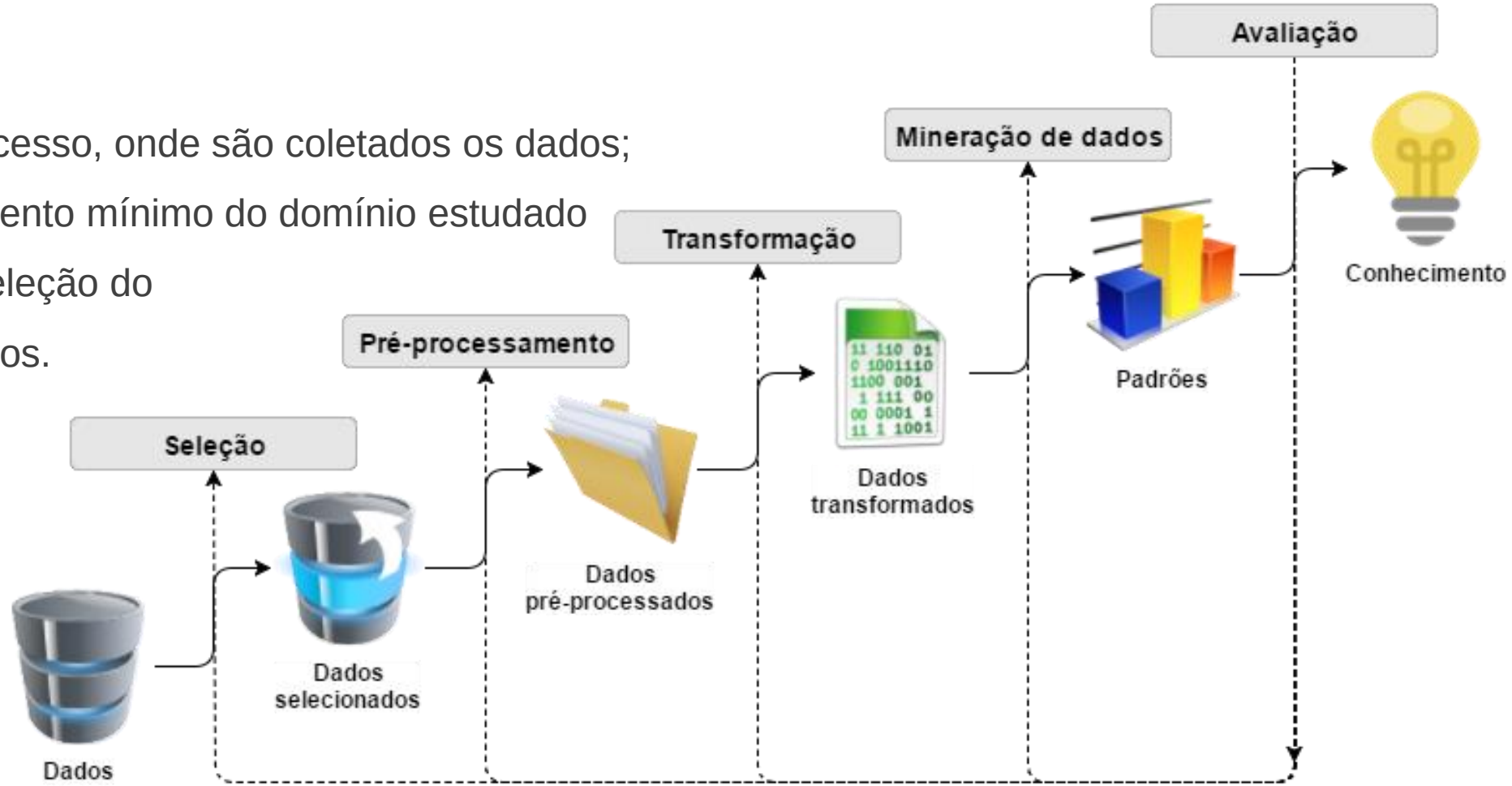


Fonte: Adaptado de Fayyad, Piatetsky-Shapiro e Smyth (1996).

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.

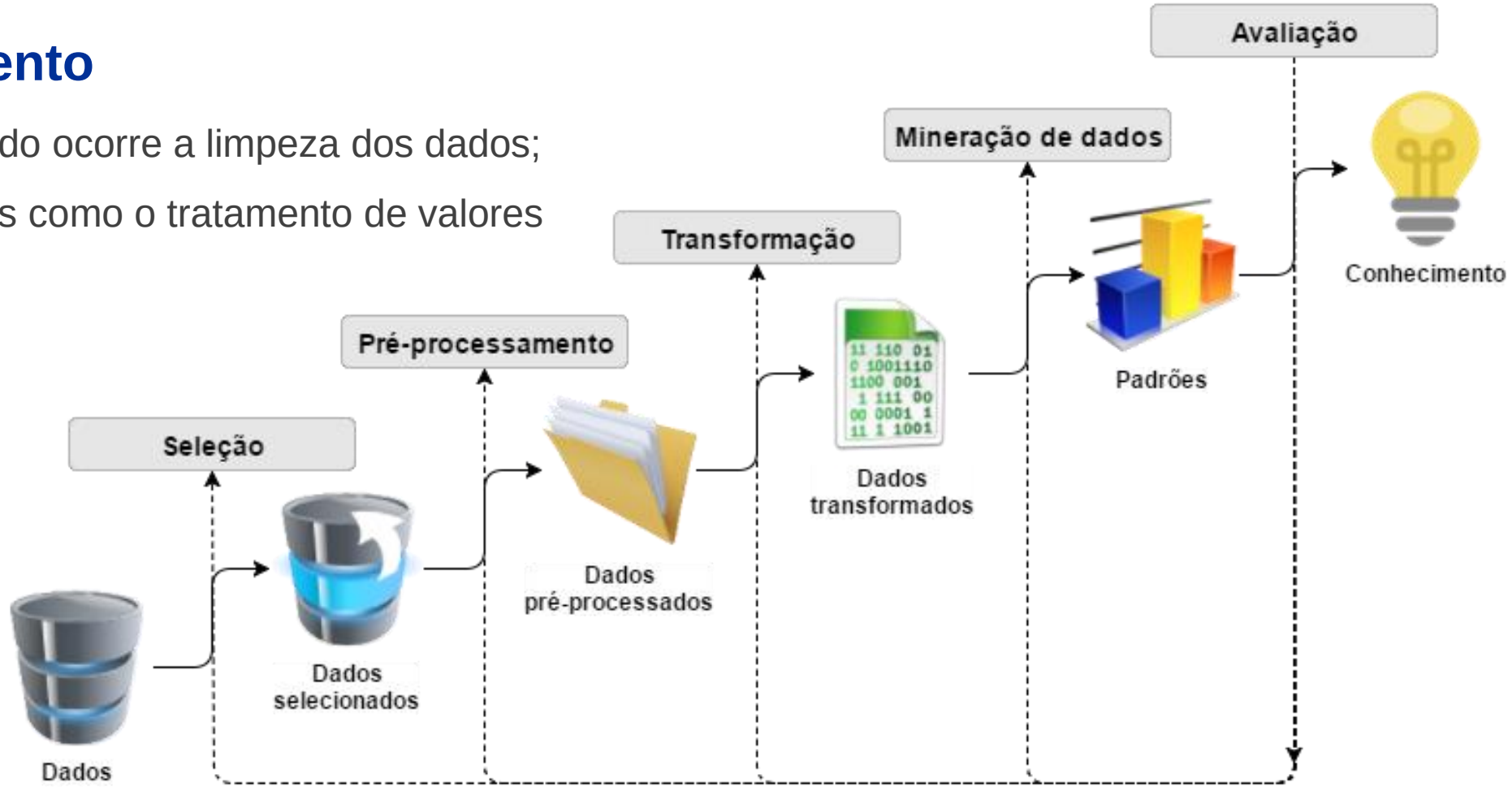


Fonte: Adaptado de Fayyad, Piatetsky-Shapiro e Smyth (1996).

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.

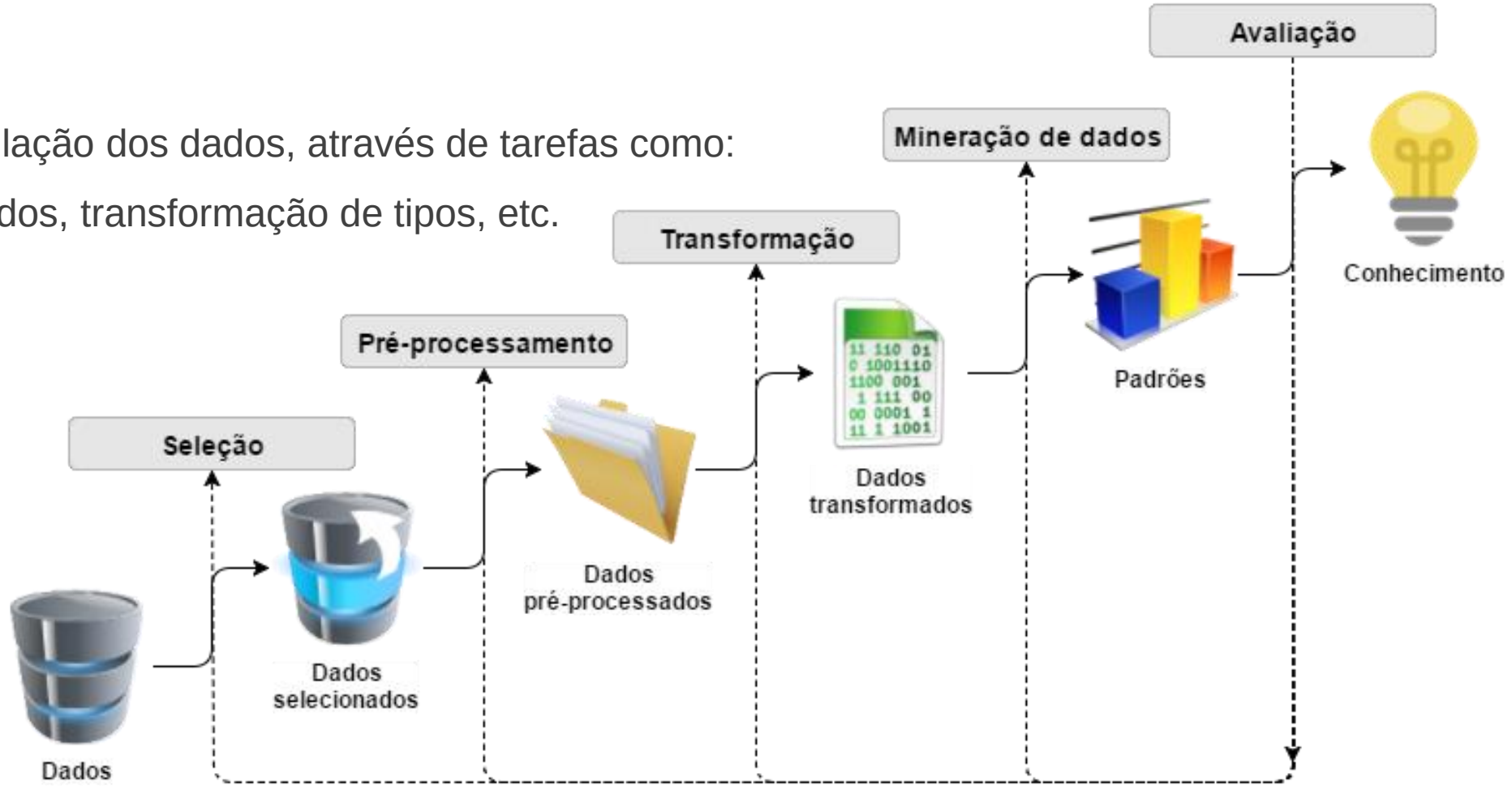


Fonte: Adaptado de Fayyad, Piatetsky-Shapiro e Smyth (1996).

Visão geral do processo de KDD

Transformação

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

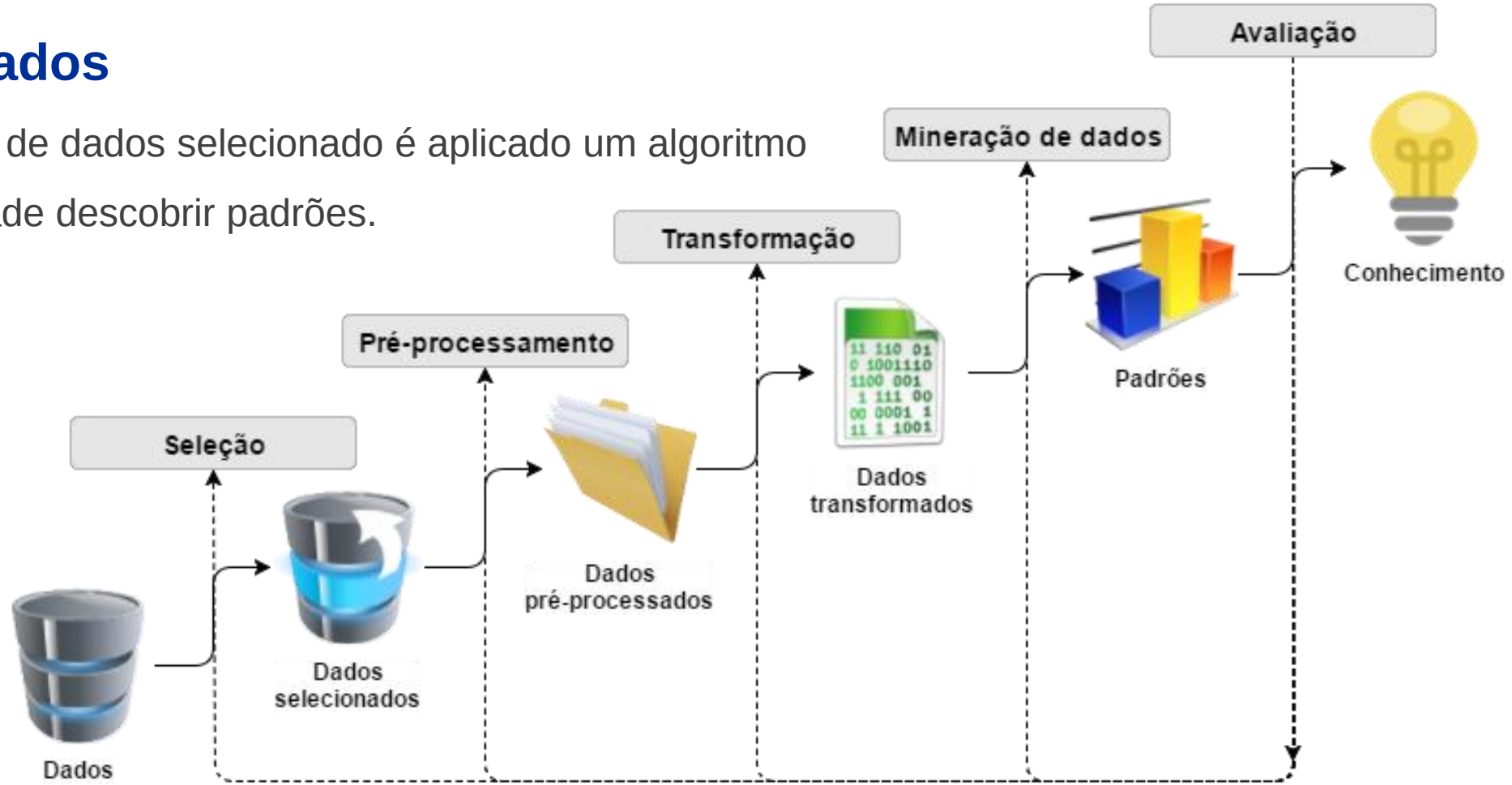


Fonte: Adaptado de Fayyad, Piatetsky-Shapiro e Smyth (1996).

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.

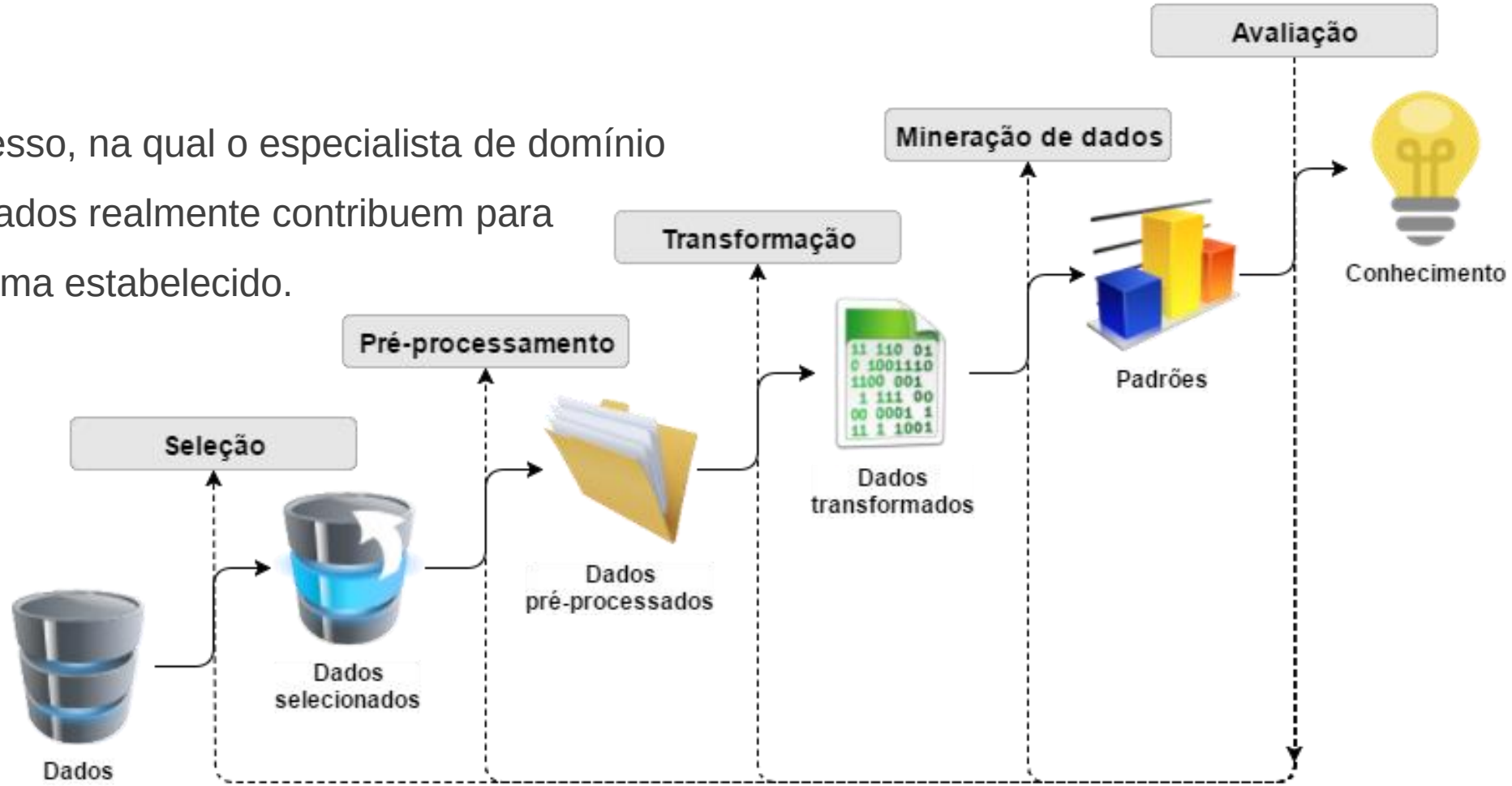


Fonte: Adaptado de Fayyad, Piatetsky-Shapiro e Smyth (1996).

Visão geral do processo de KDD

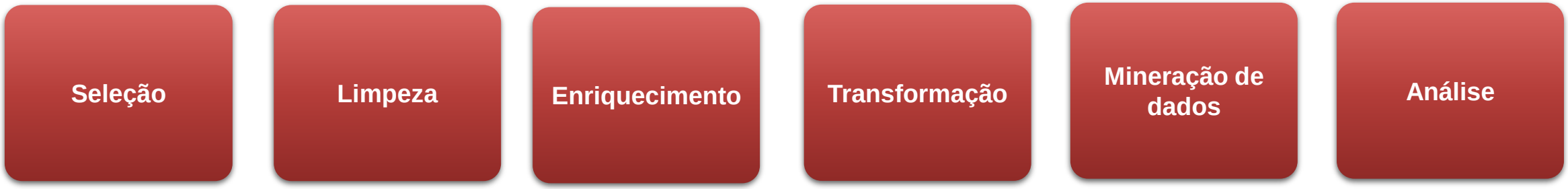
Avaliação

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

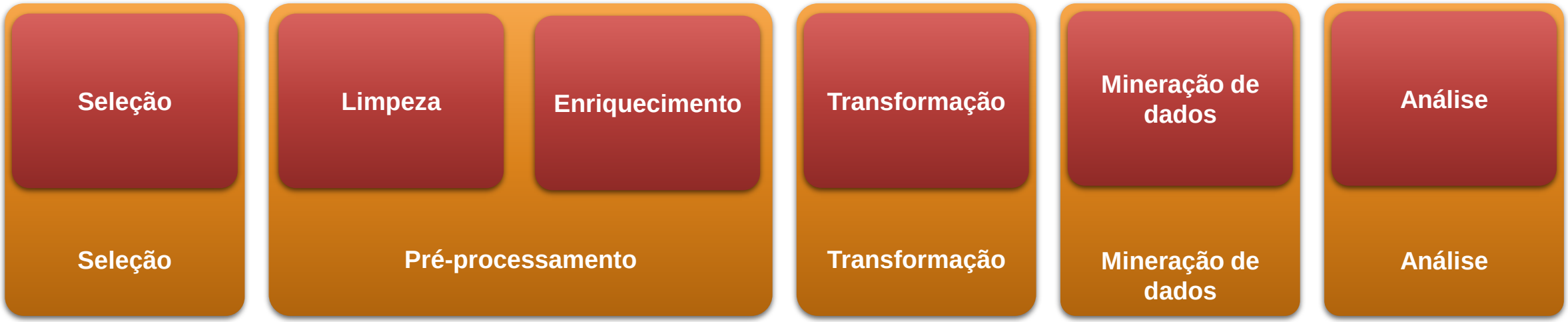


Fonte: Adaptado de Fayyad, Piatetsky-Shapiro e Smyth (1996).

Visão geral do processo de KDD



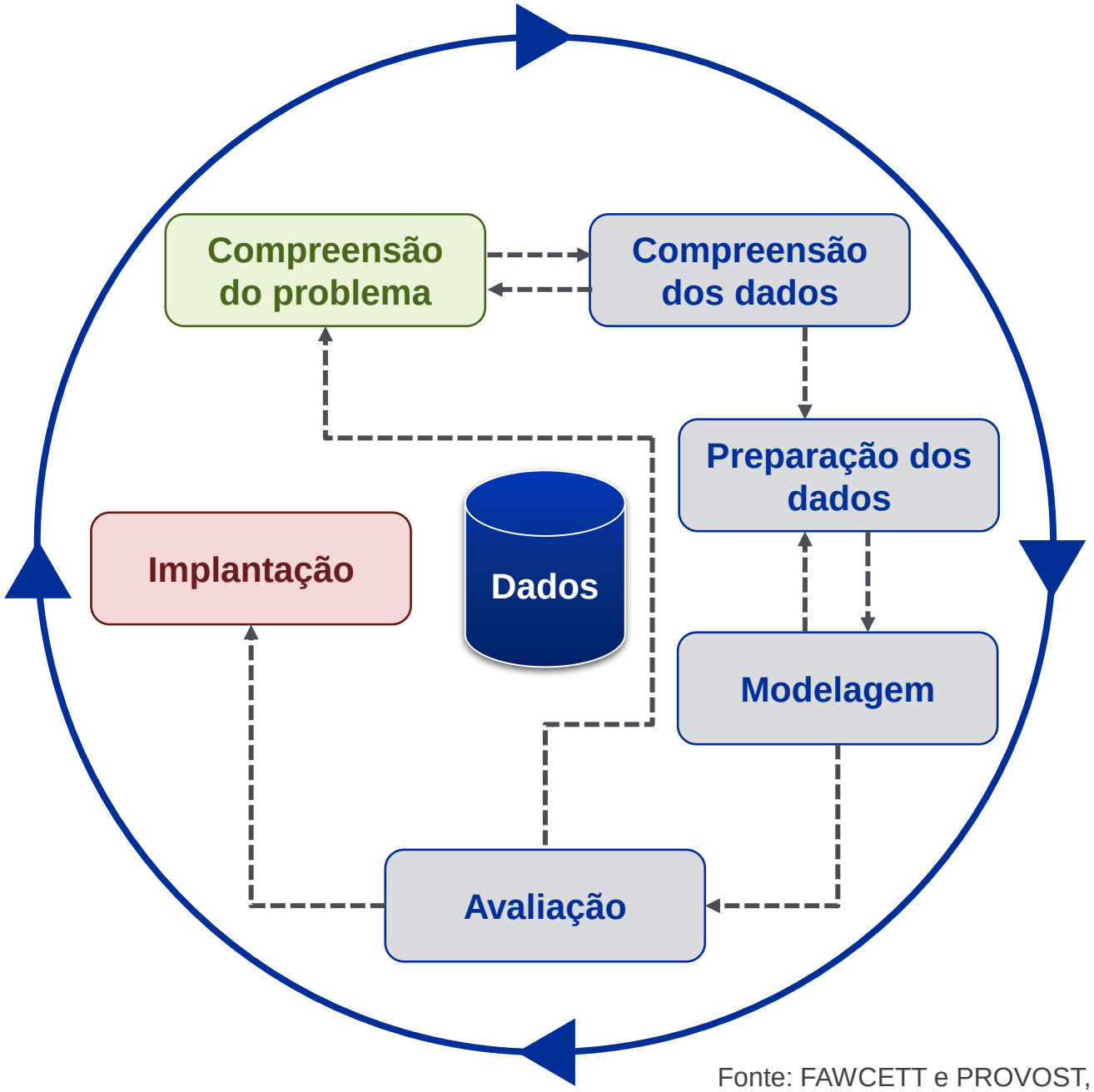
Visão geral do processo de KDD



Visão geral do processo de KDD

Processo de Mineração de Dados

Abordagem com foco no negócio/problema



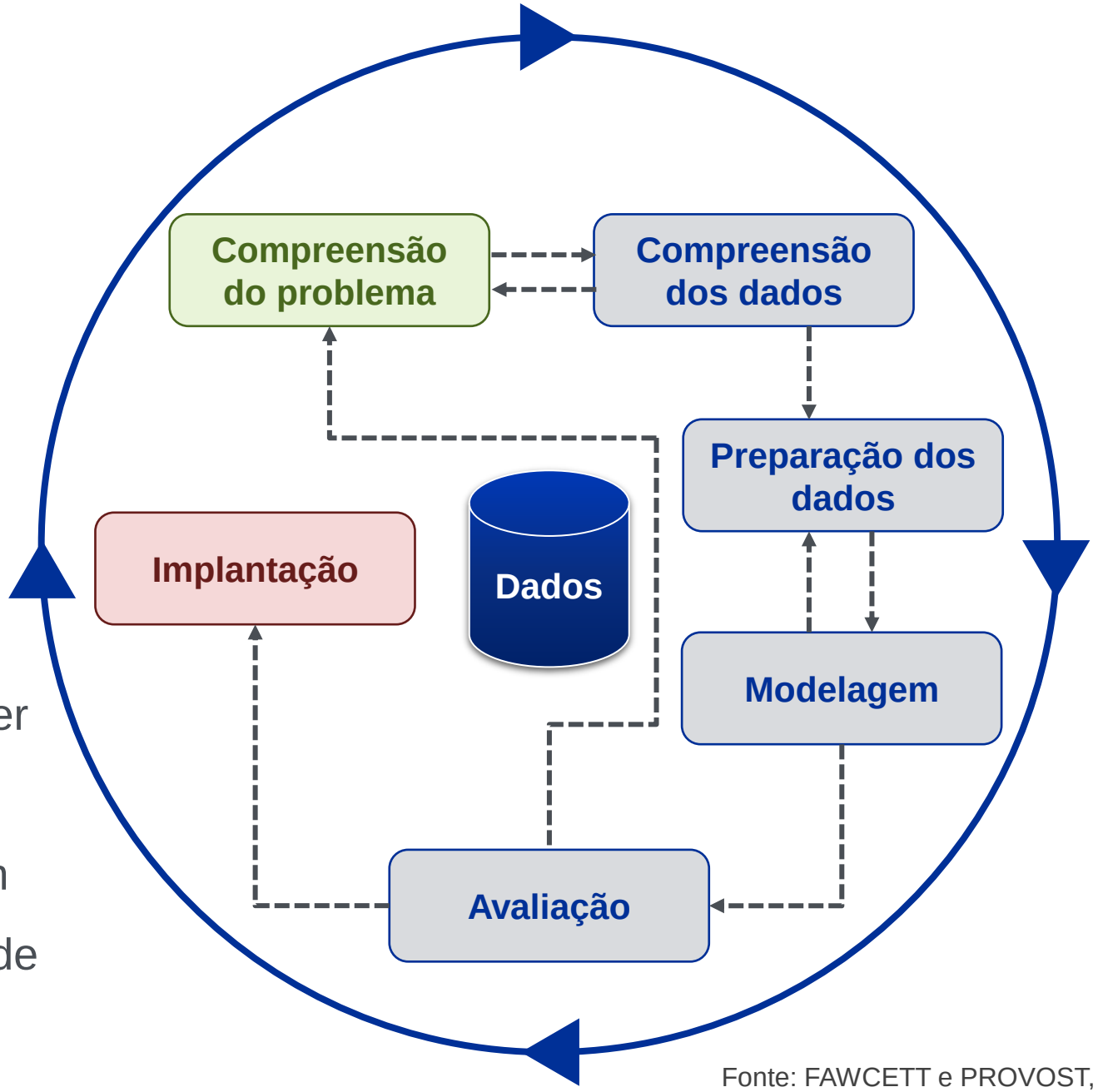
Fonte: FAWCETT e PROVOST, 2016.

Processo de Mineração de Dados

Abordagem com foco no negócio/problema

Compreensão do problema

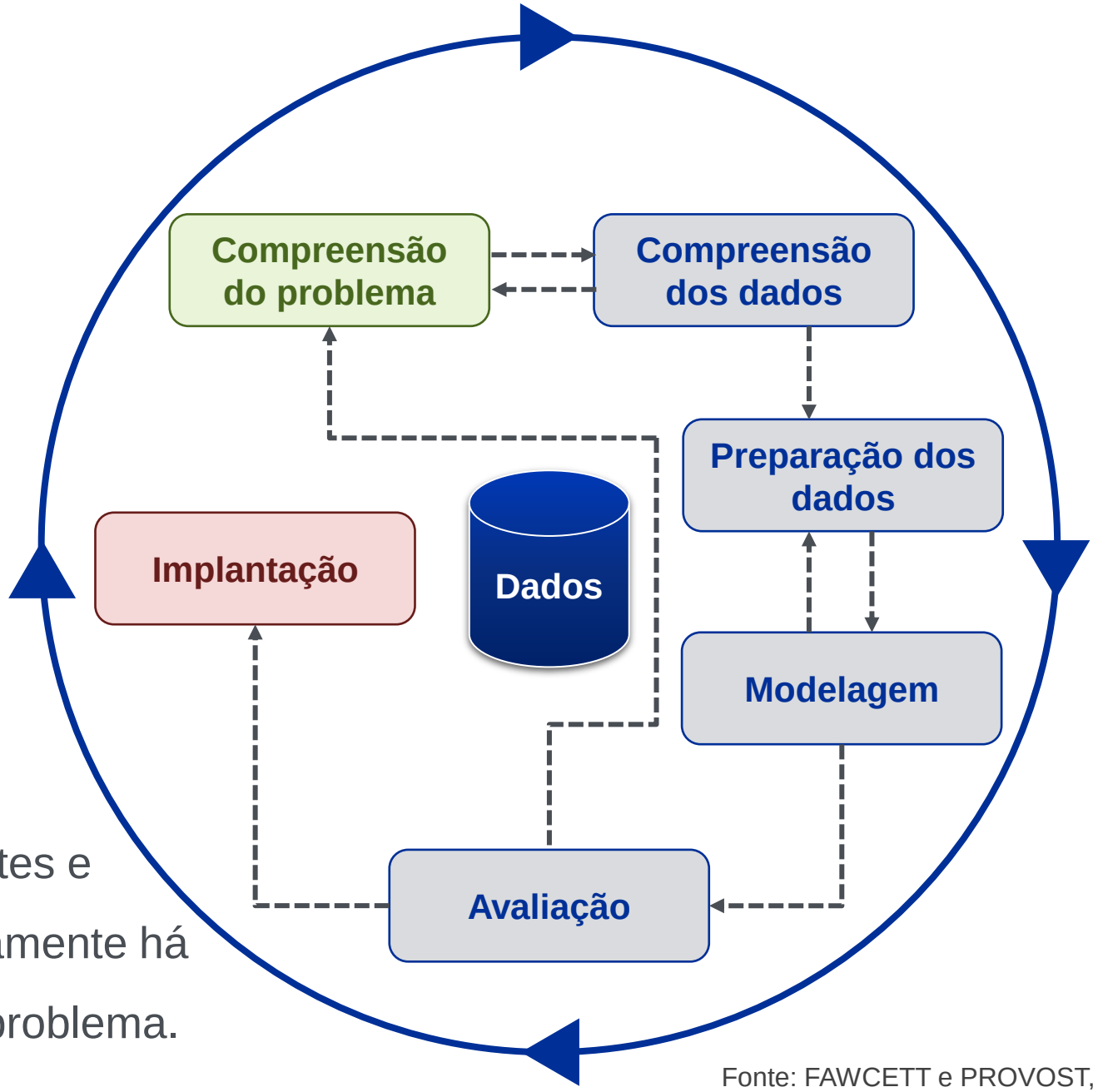
É vital compreender o problema a ser resolvido. Pode parecer óbvio, mas projetos de negócios raramente vêm modelados como problemas claros de mineração de dados.



Fonte: FAWCETT e PROVOST, 2016.

Processo de Mineração de Dados

Abordagem com foco no negócio/problema



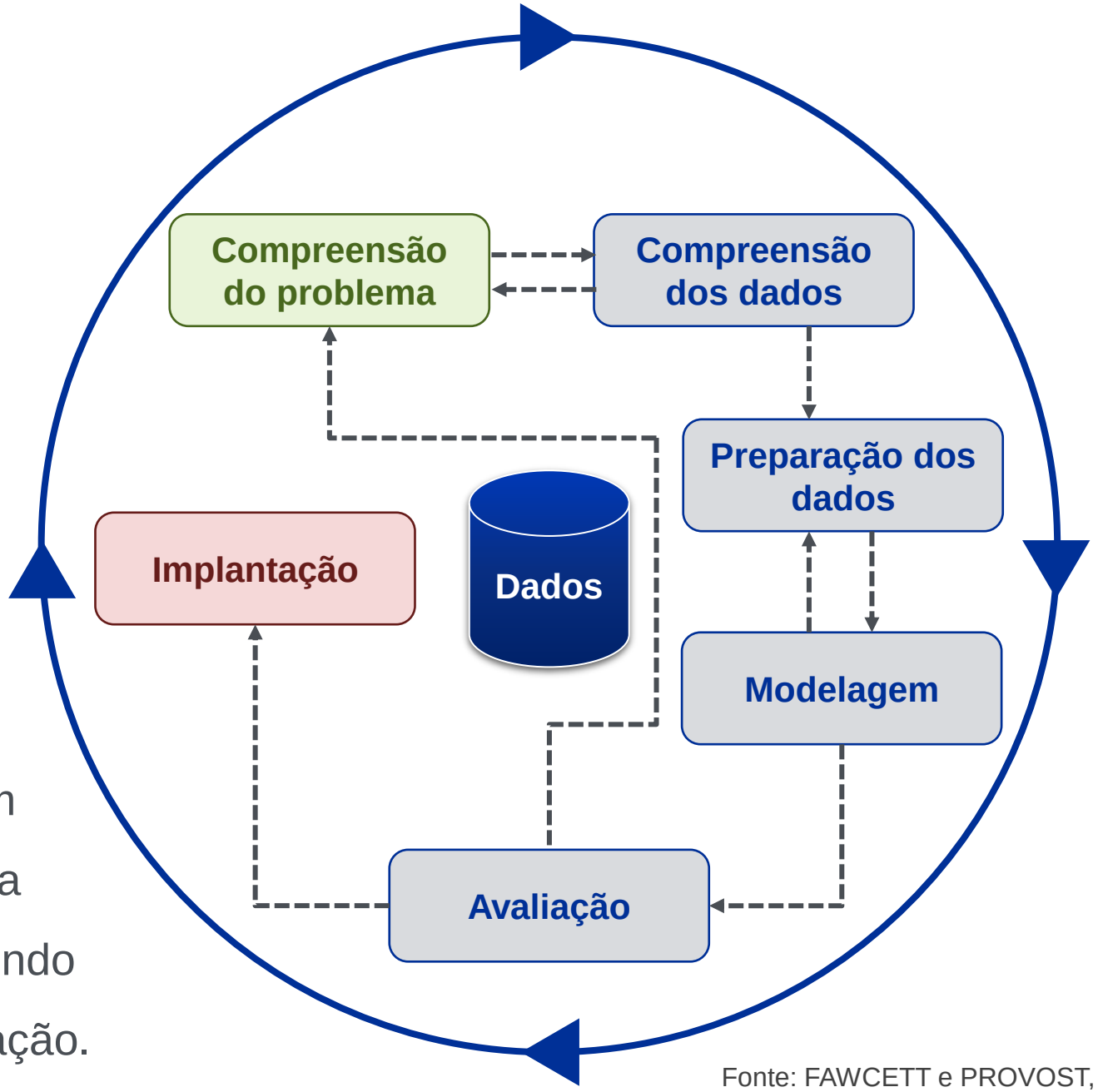
Fonte: FAWCETT e PROVOST, 2016.

Compreensão dos dados

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

Processo de Mineração de Dados

Abordagem com foco no negócio/problema



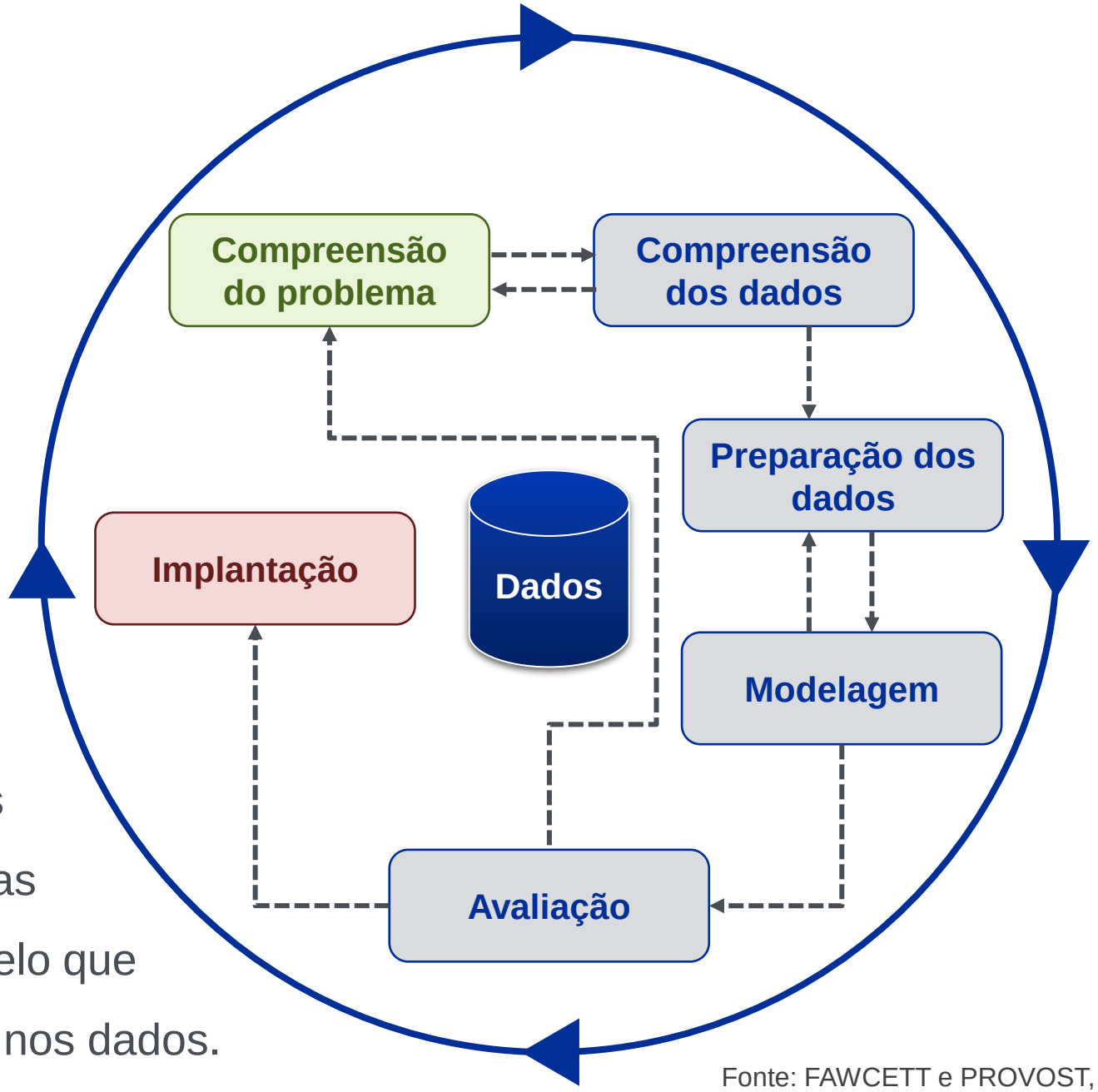
Fonte: FAWCETT e PROVOST, 2016.

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.

Processo de Mineração de Dados

Abordagem com foco no negócio/problema



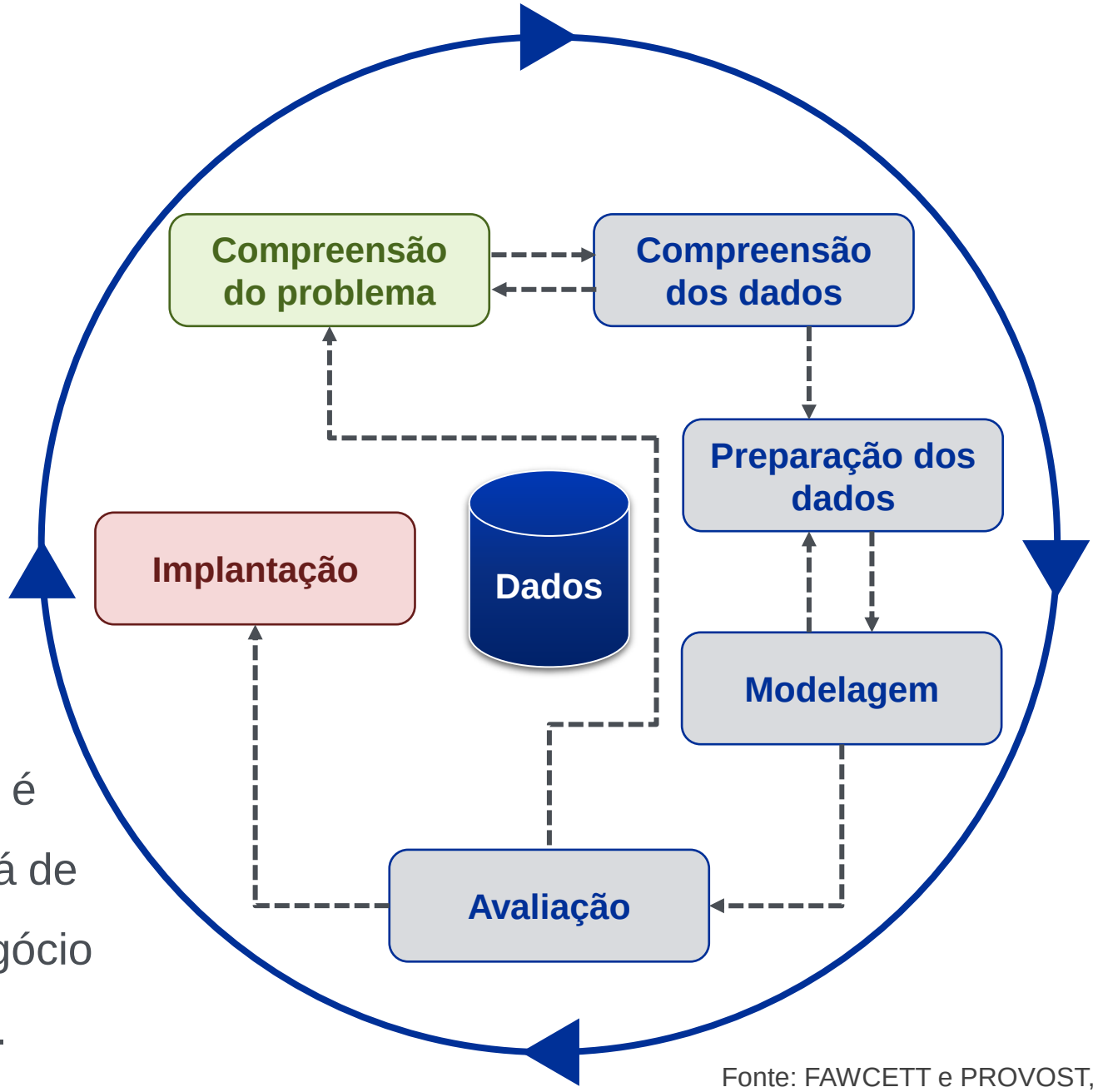
Fonte: FAWCETT e PROVOST, 2016.

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.

Processo de Mineração de Dados

Abordagem com foco no negócio/problema



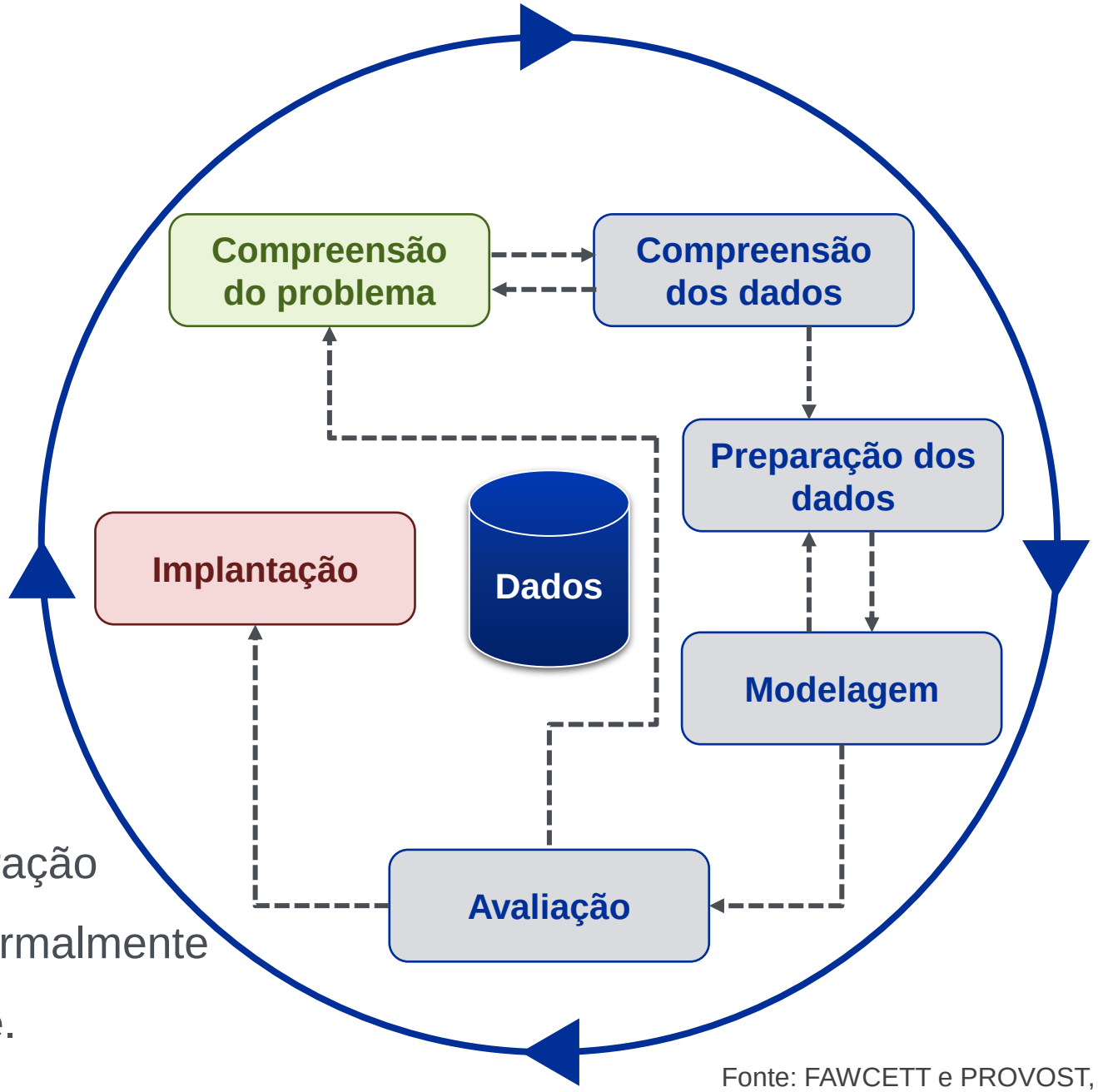
Fonte: FAWCETT e PROVOST, 2016.

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).

Processo de Mineração de Dados

Abordagem com foco no negócio/problema



Fonte: FAWCETT e PROVOST, 2016.

Implantação

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

Tarefas de mineração de dados

- **Preditiva:**
  - Tem por objetivo predizer 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.

Foco na solução do problema

Felipe Santana • 2º
Cientista de Dados Sênior | Machine Learning | Python | Deep Learn...
2 sem •

+ Seguir

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.

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

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?

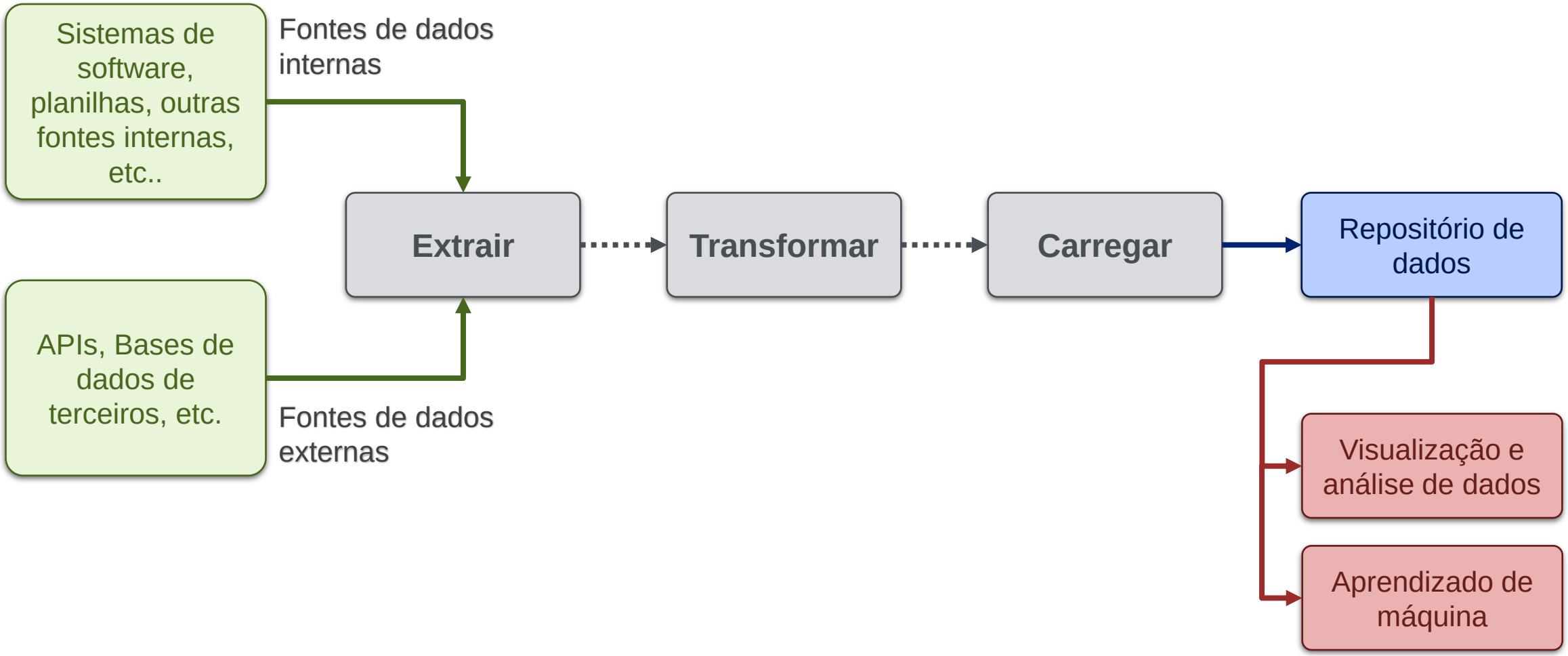
Tratamento e visualização dados

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



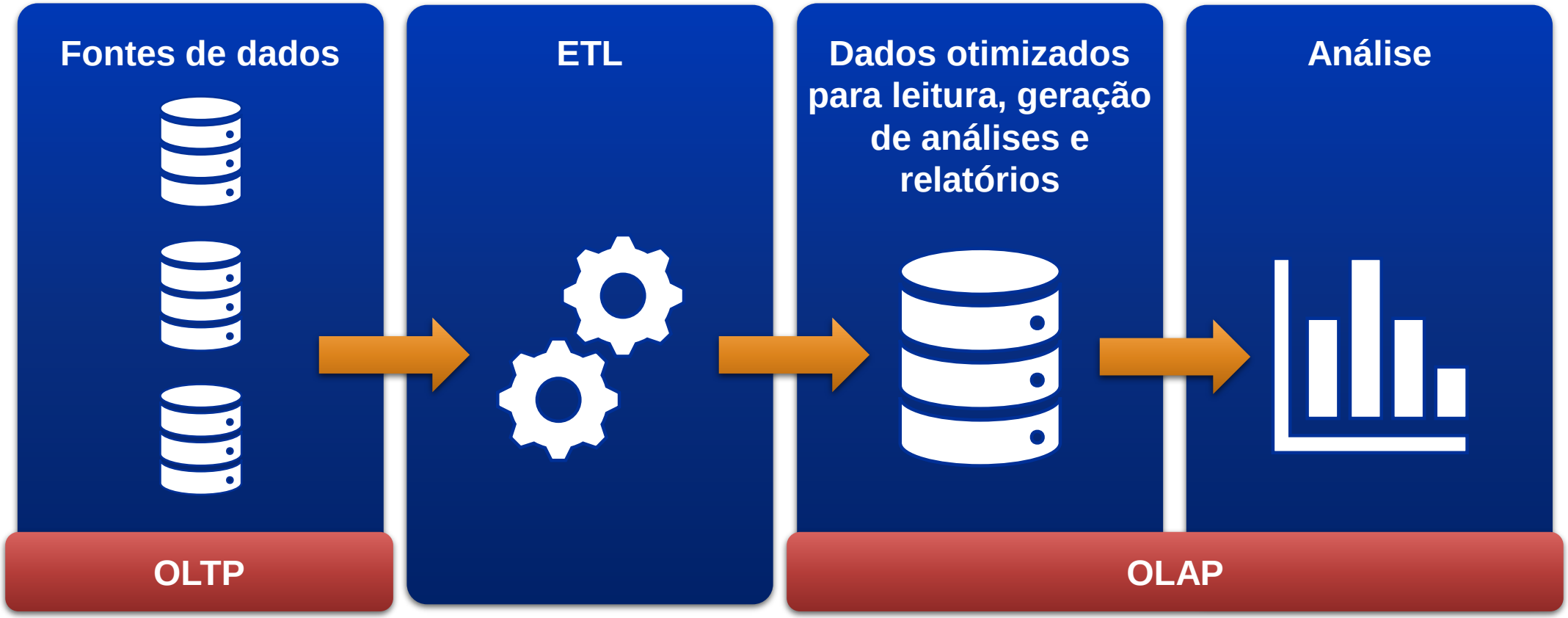
Processamento analítico *online* – OLAP

- O termo faz referência a uma **variedade de atividades voltadas à analise 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

Continua…