Métodos Matriciais e Análise de Clusters

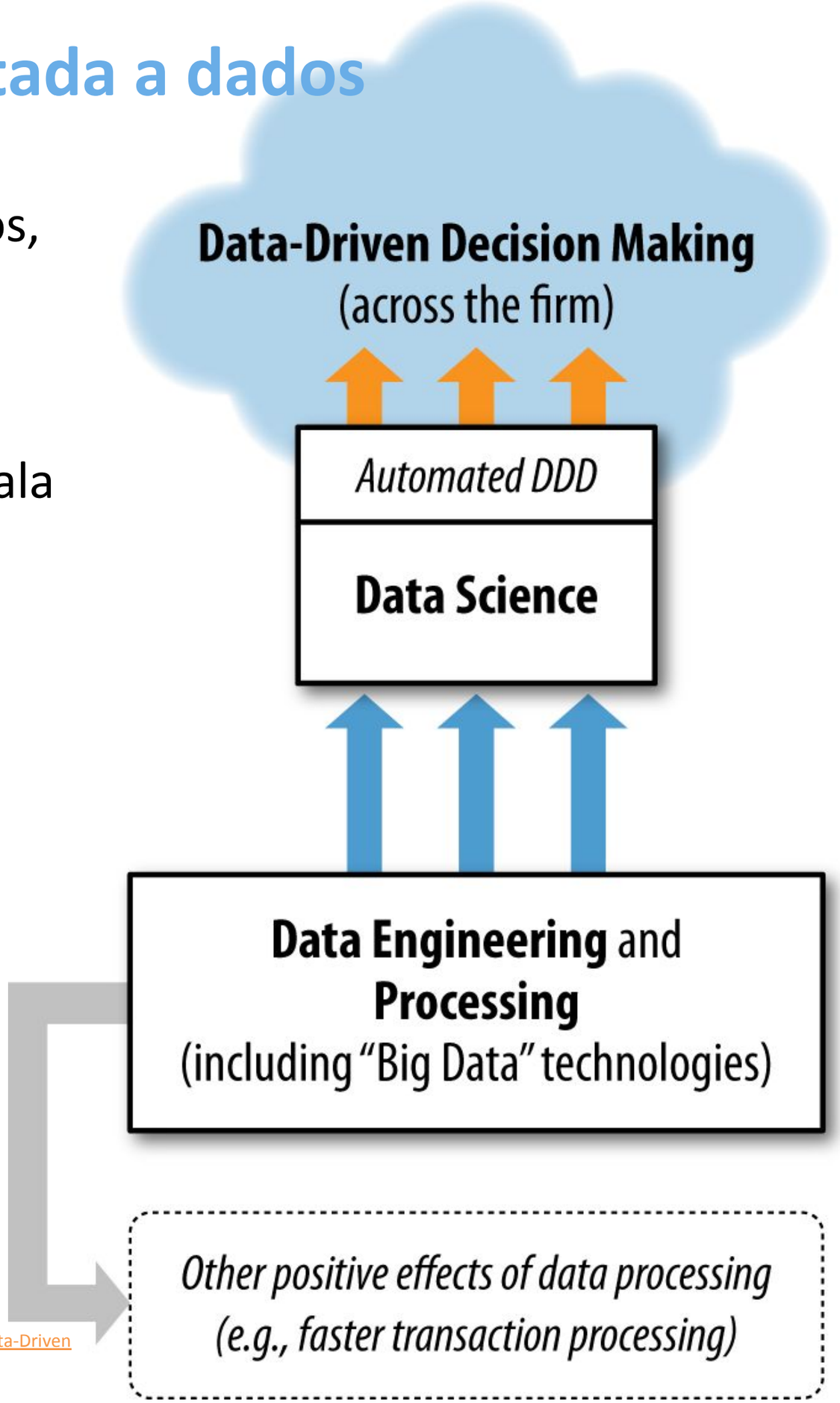
Modelagem dos Dados

Laura de Oliveira F. Moraes

Tomada de decisão orientada a dados

- Quanto mais orientada por dados, mais produtiva uma empresa é.

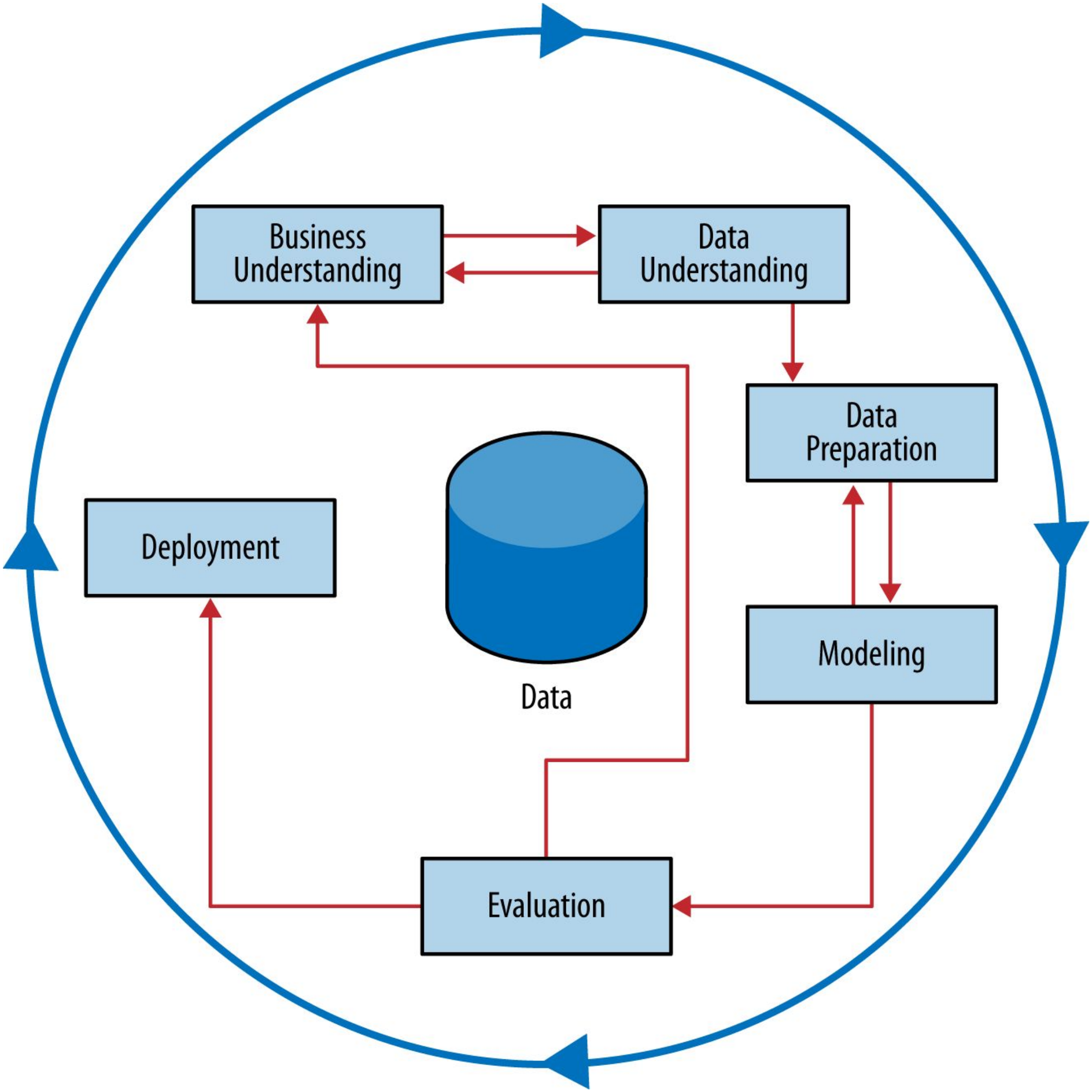
- Um desvio padrão a mais na escala de DOD está associado a um aumento de 4%-6% na produtividade.



Referência: Brynjolfsson, Erik, Lorin Hitt, and Heekyung Kim, Strength in Numbers: How Does Data-Driven Decision Making Affect Firm Performance? (April 22, 2011). Available at SSRN.

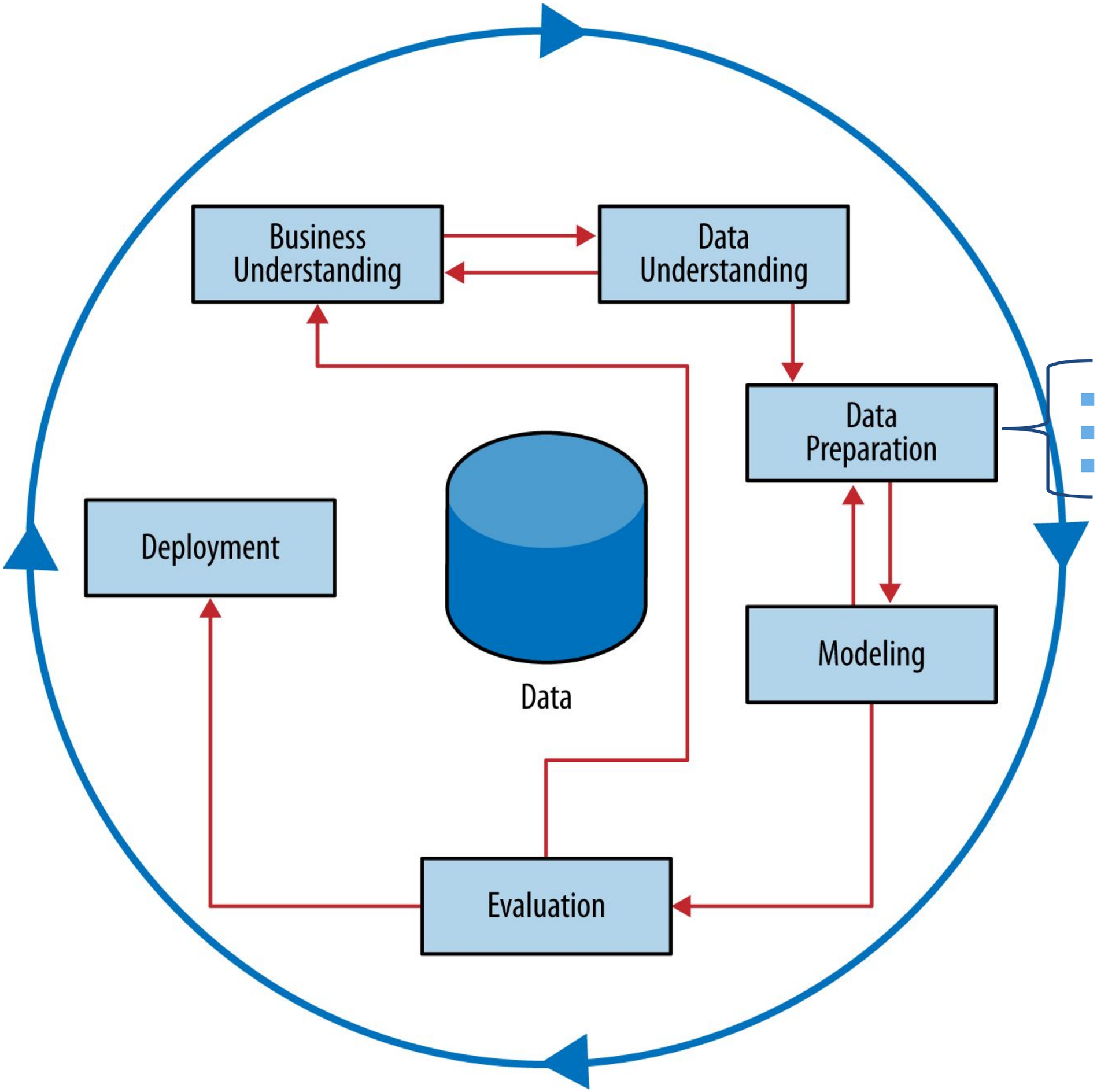
FGV | EDUCAÇÃO EXECUTIVA

CRISP-DM



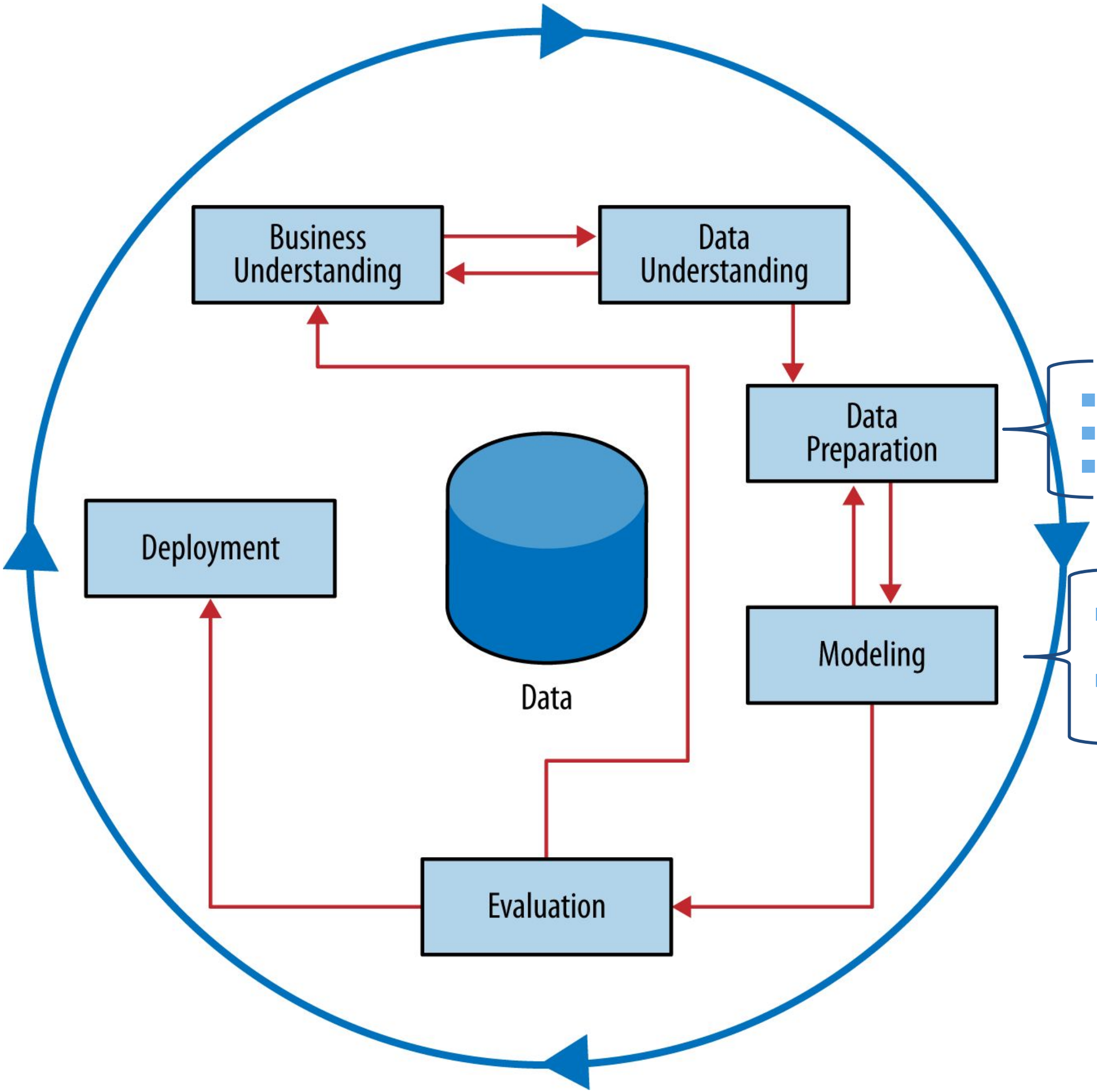
FGV | EDUCAÇÃO EXECUTIVA

CRISP-DM



- **Normalização**
- Dados faltantes
- Limpeza

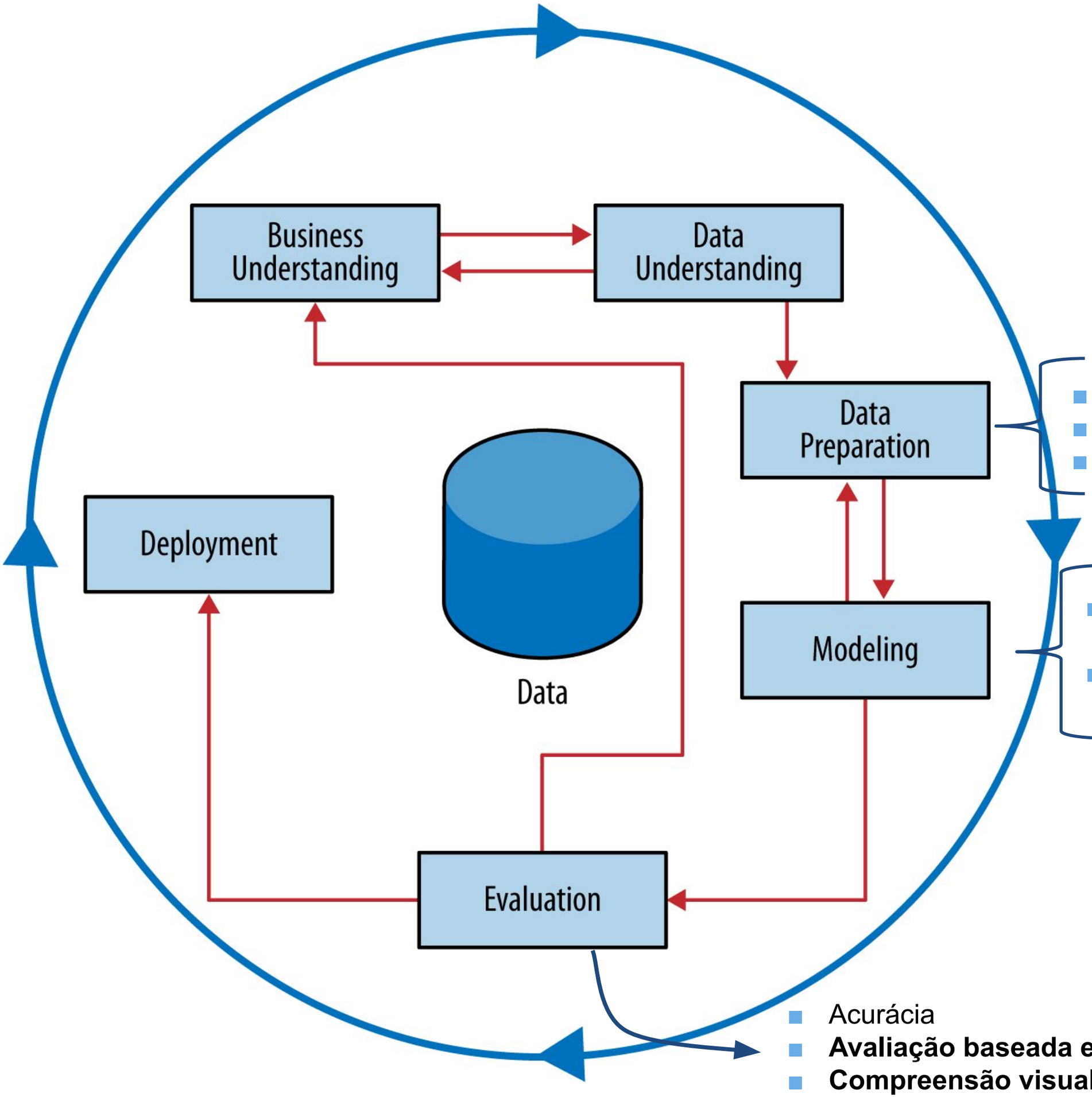
CRISP-DM



- **Normalização**
- Dados faltantes
- Limpeza

- Modelos supervisionados
- **Modelos não-supervisionados**

CRISP-DM



- **Normalização**
- Dados faltantes
- Limpeza

- Modelos supervisionados
- **Modelos não-supervisionados**

- Acurácia
- **Avaliação baseada em distância**
- **Compreensão visual**

Cargos em Mineração de Dados

- Cientista de Dados
  - Faz a modelagem
  - Estatístico x cientista da computação

- Gerente/colaborador/investidor em um projeto centrado em dados
  - Entende o potencial do negócio
  - Consegue traduzir entre negócio e execução
  - Habilidade para avaliar a proposta e a execução

Mineração de Dados versus...

- Data Warehousing
  - Data warehouses agrupam dados de uma empresa, geralmente de vários sistemas de processamento de transações

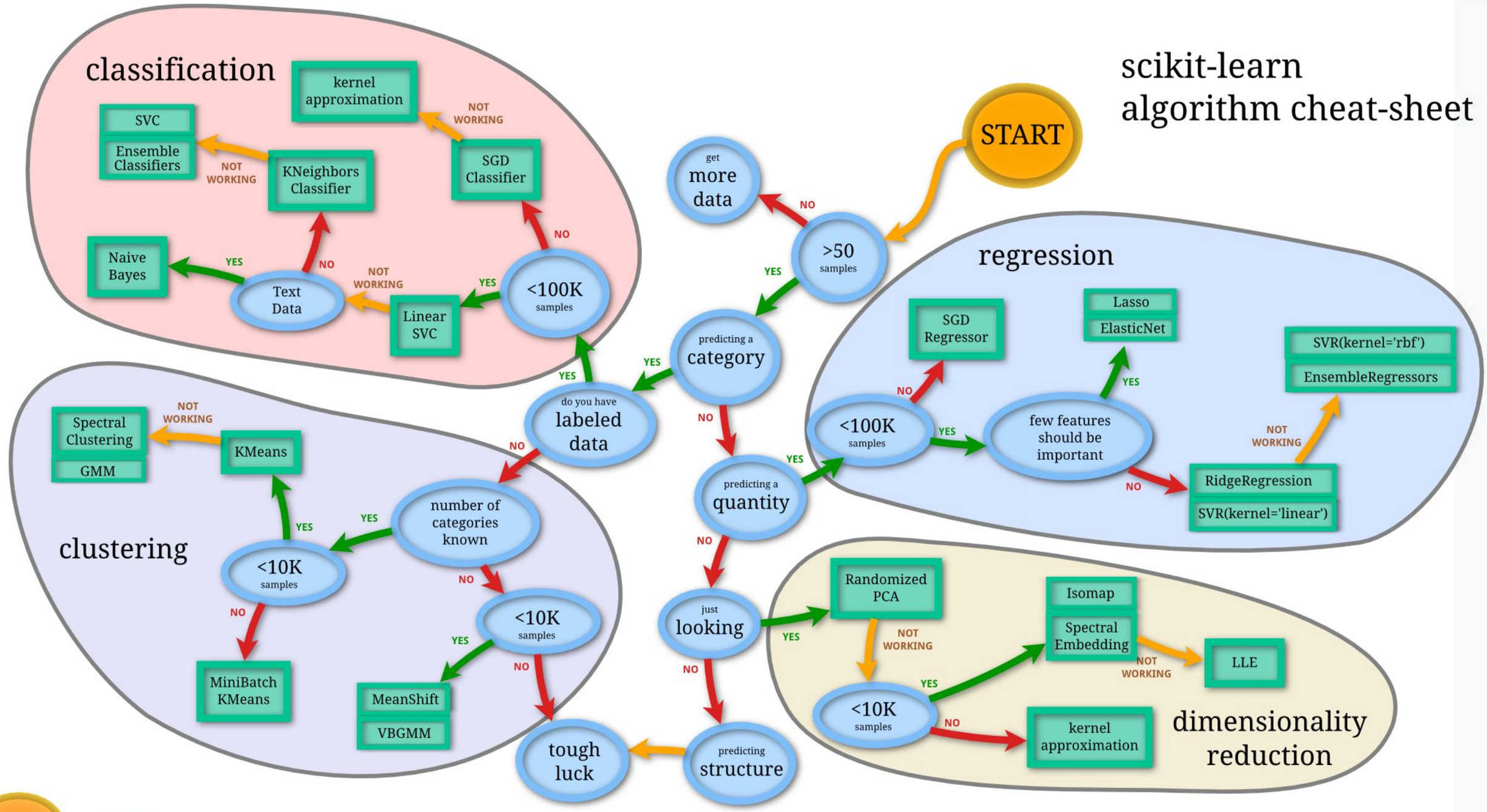
- Consulta / Relatórios (SQL, Excel, etc)
  - Interface flexível para fazer perguntas sobre os dados
  - Sem modelagem ou descoberta de padrões sofisticados

- OLAP
  - OLAP fornece uma interface fácil de usar para explorar grandes coleções de dados
  - Exploração manual, sem modelagem
  - Dimensões pré-definidas no sistema OLAP

Mineração de Dados versus...

- Análise estatística tradicional
  - Baseada principalmente em testes de hipótese ou na estimativa e quantificação de incerteza
  - Deve ser usado para acompanhar a **geração de hipóteses** na mineração de dados

Machine Learning cheat-sheet



Retirado de: https://scikit-learn.org/stable/tutorial/machine_learning_map/

Tipos de atividades não-supervisionadas

- Combinação por **similaridade** (**recomendação**)
  - Quais outras empresas são **parecidas** com os nossos melhores clientes?

- Agrupamentos (clustering) e perfilamento
  - Quantos e quais **perfis** de clientes eu tenho?
  - Qual uma maneira natural de **agrupar** meus produtos?

- Redução de dimensionalidade
  - Como consigo **visualizar** esse conjunto enorme de dados?
  - Quais as **preferências** dos meus clientes?

Terminologia

Attributes

Target attribute

Name	Balance	Age	Employed	Write-off
Mike	$200,000	42	no	yes
Mary	$35,000	33	yes	no
Claudio	$115,000	40	no	no
Robert	$29,000	23	yes	yes
Dora	$72,000	31	no	no

This is one row (example).
Feature vector is: **<Claudio,115000,40,no>**
Class label (value of Target attribute) is **no**

Vetor e Matriz

Caso a gente meça três atributos para uma pessoa: idade, peso e gênero. Como podemos representar esses dados?

$$x_i = (Idade \quad Peso \quad Genero)$$

Vetor e Matriz

Caso a gente meça três atributos para uma pessoa: idade, peso e gênero. Como podemos representar esses dados?

$$x_i = (x_{i1} \quad x_{i2} \quad x_{i3})$$

Vetor e Matriz

Caso a gente meça três atributos para uma pessoa: idade, peso e gênero. Como podemos representar esses dados?

$$x_i = (x_{i1} \quad x_{i2} \quad x_{i3})$$

$$x_i^T = \begin{pmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \end{pmatrix}$$

Vetor e Matriz

E para um conjunto de pessoas?

$$x_i = (x_{i1} \quad x_{i2} \quad x_{i3})$$

$$x_i^T = \begin{pmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \end{pmatrix}$$

Vetor e Matriz

E para um conjunto de pessoas?

$$x_i = (x_{i1} \quad x_{i2} \quad x_{i3})$$

$$x_i^T = \begin{pmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \end{pmatrix} \qquad X = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix}$$

E esse tipo de dado, como fica?

$$x_i = (Idade \quad Peso \quad Genero)$$

E esse tipo de dado, como fica?

$$x_i = (Idade \quad Peso \quad Genero)$$

$$Genero \begin{cases} M \\ F \end{cases}$$

E esse tipo de dado, como fica?

$$x_i = (Idade \quad Peso \quad Genero)$$

$$Genero \begin{cases} M = 0 \\ F = 1 \end{cases}$$

E esse tipo de dado, como fica?

$$x_i = (Idade \quad Peso \quad Genero \quad Profissao)$$

- **Ela é nominal ou ordinal?**
  - **Nominais** são aquelas onde não existe nenhum tipo de ordem ou ranking entre as categorias, como gênero, por exemplo. Já as **ordinais**, possuem alguma ordem embutida, como por exemplo, escolaridade.

- **Tipos de codificadores**
  - **Ordinal**: converte cada valor para um inteiro.
  - **OneHot (ou Dummy)**: cria-se uma nova coluna/atributo para cada valor.
  - **Hashing:** Como o OneHot, mas utilizando hash. Pode-se perder alguma informação por causa de colisões.