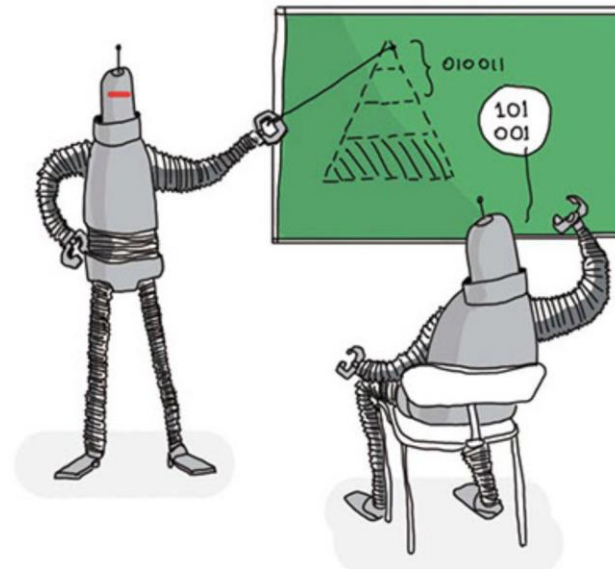


# Introdução a Data Science e Machine Learning



Prof. Dr. Márcio Andrey Teixeira  
Instituto Federal de São Paulo – Campus Catanduva  
Catanduva, SP  
Membro Sênior do IEEE  
[marcio.andrey@ifsp.edu.br](mailto:marcio.andrey@ifsp.edu.br)

# Organização do Minicurso

## Parte I – Introdução a Data Science

- ✓ O que é Data Science ?
- ✓ Mercado de Trabalho
- ✓ O que estudar para ser um bom cientista de dados?
- ✓ Introdução ao Spyder e Python
- ✓ Bibliotecas Numpy, Pandas e Matplotlib
- ✓ Processamento de dados

# Organização do Minicurso

## Parte II – Preparação do Dataset

- ✓ Tratando os dados do dataset
- ✓ Removendo dados em branco
- ✓ Variáveis Dummy
- ✓ Dividindo o dataset em treinamento e teste
- ✓ Normalização de variáveis

# Organização do Minicurso

## Parte III – Projeto 911 Call

- ✓ Análise exploratória de dados do dataset 911 call

## Parte IV – Machine Learning

- ✓ Construindo um modelo baseado em regressão linear simples
- ✓ Construindo um modelo baseado em regressão linear múltipla

# Parte I – Introdução a Data Science

## ✓ O que é Data Science ?

É o estudo disciplinado dos dados e informações inerentes ao negócio e todas as visões que possam cercar um determinado assunto. [2]

É uma ciência que estuda as informações, seu processo de captura, transformação, geração e, posteriormente, análise de dados. A ciência de dados envolve diversas disciplinas:

- Computação;
- Estatística;
- Matemática;
- Conhecimento do negócio.

# Parte I – Introdução a Data Science

## ✓ Mercado de Trabalho

Um estudo recente conduzido nos EUA mostra que Big Data Analytics vai gerar 10 milhões de oportunidades de trabalho em todo mundo na próxima década. [3]

Exemplos de novas carreiras nesta área:

# Parte I – Introdução a Data Science

## ✓ Mercado de Trabalho (cont.)

1. Engenheiro de Dados
2. Engenheiro de Big Data
3. Arquiteto de Soluções de Big Data
4. Cientista de Machine Learning
5. Engenheiro de Machine Learning
6. Especialista em Business Analytics (Analista de Negócios)
7. Desenvolvedor de Visualização de Dados
8. Gerente de Analytics
9. Estatístico
10. Cientista de Dados

# Parte I – Introdução a Data Science

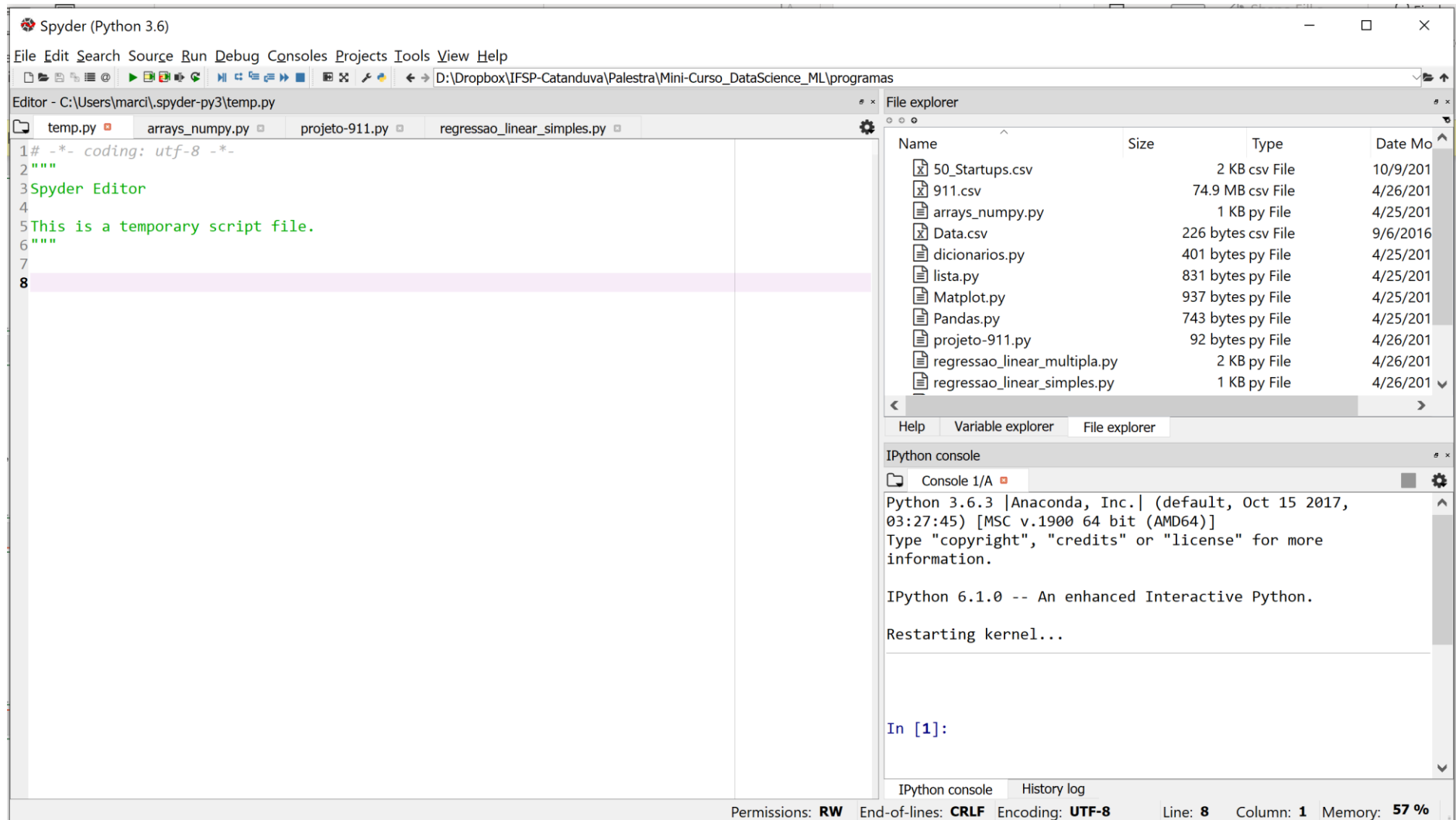
## ✓ O que estudar para ser um bom cientista de dados?

- Estatística e Matemática
- Algoritmos de Machine Learning
- Linguagem de programação com bom suporte na área de análise de dados. (Python e R)



# Parte I – Introdução a Data Science

## ✓ Introdução ao Spyder



# Parte I – Introdução a Data Science

## Python e Ferramentas para Análise de Dados.

- ✓ Python possui uma grande variedade de ferramentas para serem utilizadas em *data science*.
- ✓ Algumas dessas ferramentas serão descritas a seguir:

# Parte I – Introdução a Data Science

## Numpy

Numpy é uma biblioteca de álgebra linear para python.

É o bloco de construção de todas as outras bibliotecas de análise de dados.

Ela é extremamente rápida, muito utilizada para análise de dados.

# Parte I – Introdução a Data Science

## Numpy (Exemplos)

```
#Introdução ao Numpy
```

```
import numpy as np
```

```
minha_lista = [1,2,3]
```

```
#converte a minha lista em um array numpy
```

```
np.array(minha_lista)
```

```
#lista de listas
```

```
minha_matriz = [[1,2,3], [4,5,6], [7,8,9]]
```

```
np.array(minha_matriz)
```

# Parte I – Introdução a Data Science

## Numpy (Exemplos)

#Criando arrays utilizando o método arange do numpy.

```
np.arange(0, 10)
```

#Utilizando o step.

```
np.arange(0,10,2)
```

#Criando arrays com zeros

```
arr = np.zeros(3)
```

#Criando uma matriz 5x5

```
arr2 = np.zeros((5,5))
```

# Parte I – Introdução a Data Science

## Numpy (Exemplos)

#Criando uma matriz 2x2 com 1s

```
arr3 = np.ones((2,2))
```

#Criando uma matriz identidade

```
np.eye(4)
```

#Cria um array com numeros de intervalos.

```
np.linspace(0, 10, 20)
```

#Criando um array com números aleatórios de zero a 1 seguindo uma distribuição uniforme.

#Qualquer um desses números tem a probabilidade de ser selecionado

```
np.random.rand(5)
```

# Parte I – Introdução a Data Science

## Numpy (Exemplos)

#Criando uma sequencia desta de zero a 100.

```
np.random.rand(5) * 100
```

#Criando uma Matriz com duas dimensões

```
np.random.rand(5,4)
```

#Criando uma Matriz com duas 3 (Profundidade, altura e comprimento)

```
np.random.rand(5,4,3)
```

#Nesse caso, os números são tirados de uma distribuição normal, média zero e desvio padrão 1. Distribuição Gaussiana.

```
np.random.randn(4)
```

# Parte I – Introdução a Data Science

## Numpy (Exemplos)

#Nesse caso, são criados números inteiros de acordo com uma distribuição aleatória

```
np.random.randint(0,100,10)
```

#Vetor com tamanho 25

```
arr4 = np.random.rand(25)
```

#Modificando o formato do vetor para uma matriz

```
arr4.reshape((5,5))
```

```
arr4 = arr4.reshape((5,5))
```



# Parte I – Introdução a Data Science

## Numpy (Exemplos)

`arr4.shape`

`arr4.max()`

`arr4.min()`

#Retorna o índice onde o maior elemento se encontra

`arr4.argmax()`

#Retorna o índice onde o maior elemento se encontra

`arr4.argmin()`

# Parte I – Introdução a Data Science

## Pandas

O Pandas é uma biblioteca de código aberto escrita sobre a biblioteca Numpy

Ela permite uma rápida visualização dos dados

Muito semelhante ao Excel, e possui métodos próprios para a visualização dos dados.

# Parte I – Introdução a Data Science

## Pandas

```
import numpy as np  
import pandas as pd
```

```
#Introdução ao Pandas
```

```
#Séries: É um objeto básico do pandas. (versão mais simples do pandas)
```

```
labels = ['a', 'b', 'c']
```

```
minha_lista = [10,20,30]
```

```
arr = np.array([10,20,30])
```

```
d = {'0':10, 'b':20, 'c':30}
```

# Parte I – Introdução a Data Science

## Pandas

```
pd.Series(data=minha_lista, index=labels)
```

```
s = pd.Series(data=minha_lista, index=labels)
```

```
s['b']
```

# Parte I – Introdução a Data Science

## Pandas

#DataFrame. O DataFrame é o objeto principal, que é um conjunto de series

```
np.random.seed(101)
```

```
df = pd.DataFrame(np.random.randn(5,4), index='A B C D E'.split(), columns='W X Y Z'.split())
```

```
df['W']
```

```
type(df['W'])
```

```
type(df)
```

```
df[['W','Z']]
```

```
df.W
```

# Parte I – Introdução a Data Science

## Matplotlib

É uma biblioteca de visualização de dados mais popular em Python.

Ela dá o controle de todos os aspectos da sua figura.

# Parte I – Introdução a Data Science

## Matplotlib

```
import numpy as np  
import pandas as pd  
import matplotlib.pyplot as plt
```

```
x = np.linspace(0,5,11)  
y = x*x
```

```
#Modo funcional  
plt.plot(x,y)  
plt.xlabel('Eixo X')  
plt.ylabel('Eixo Y')  
plt.title('Title')
```

# Parte I – Introdução a Data Science

## Matplotlib

#Plotando dois ou mais tipos de dados

```
plt.subplots(nrows=1, ncols=2)
```

```
plt.subplot(1, 2, 1)
```

```
plt.plot(x,y, 'r--')
```

```
plt.subplot(1, 2, 2)
```

```
plt.plot(y,x, 'g*-')
```



# Parte II - Preparação do Dataset

## Tratando os dados do Dataset

#Preparação do Dataset

#1 - Importar o dataset.

#2 - Tratar os dados faltantes.

#3 - Codificar os dados Categoricals.

#4 - Dividir os dataset em grupos de treinamento e test

#5 - Normalizar os dados.

#importando as bibliotecas

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

# Parte II - Preparação do Dataset

## Tratando os dados do Dataset

#Importando o Dataset

```
dataset = pd.read_csv('Data.csv')
```

#Distinguindo variáveis independente das variáveis dependente. A parte esquerda da virgula representa as linhas do dataset, e a parte direita representa as colunas do dataset. Nesse caso vamos pegar todas as colunas do dataset, excluindo a ultima coluna que e a variável dependente. (X no plano cartesiano)

```
X = dataset.iloc[:, :-1].values
```

# Parte II - Preparação do Dataset

## Tratando os dados do Dataset

#Selecionando a coluna que tem as variáveis dependente. Seria a Coluna Y do plano cartesiano. O Index em pythom inicia no 0. Como temos 4 colunas, vamos indicar a coluna 3.

```
y = dataset.iloc[:, 3].values
```

# Parte II - Preparação do Dataset

## Tratando os dados do Dataset

```
# Tratando dados inexistentes
from sklearn.preprocessing import Imputer
imputer = Imputer(missing_values = 'NaN', strategy = 'mean', axis = 0)
imputer = imputer.fit(X[:, 1:3])
X[:, 1:3] = imputer.transform(X[:, 1:3])
```

# Parte II - Preparação do Dataset

## Tratando os dados do Dataset

```
# Codificando dados Categóricos
from sklearn.preprocessing import LabelEncoder, OneHotEncoder
labelencoder_X = LabelEncoder()
X[:, 0] = labelencoder_X.fit_transform(X[:, 0])
onehotencoder = OneHotEncoder(categorical_features = [0])
X = onehotencoder.fit_transform(X).toarray()

#Codificando as variáveis independente
labelencoder_y = LabelEncoder()
y = labelencoder_y.fit_transform(y)
```

# Parte II - Preparação do Dataset

## Tratando os dados do Dataset

#Dividindo o Dataset em grupos de treinamento e teste. Nesse caso estamos utilizando 20% para teste e 80% para treino.

```
from sklearn.model_selection import train_test_split  
X_train, X_test, y_train, y_test = train_test_split(X,y, test_size = 0.2, random_state = 0)
```

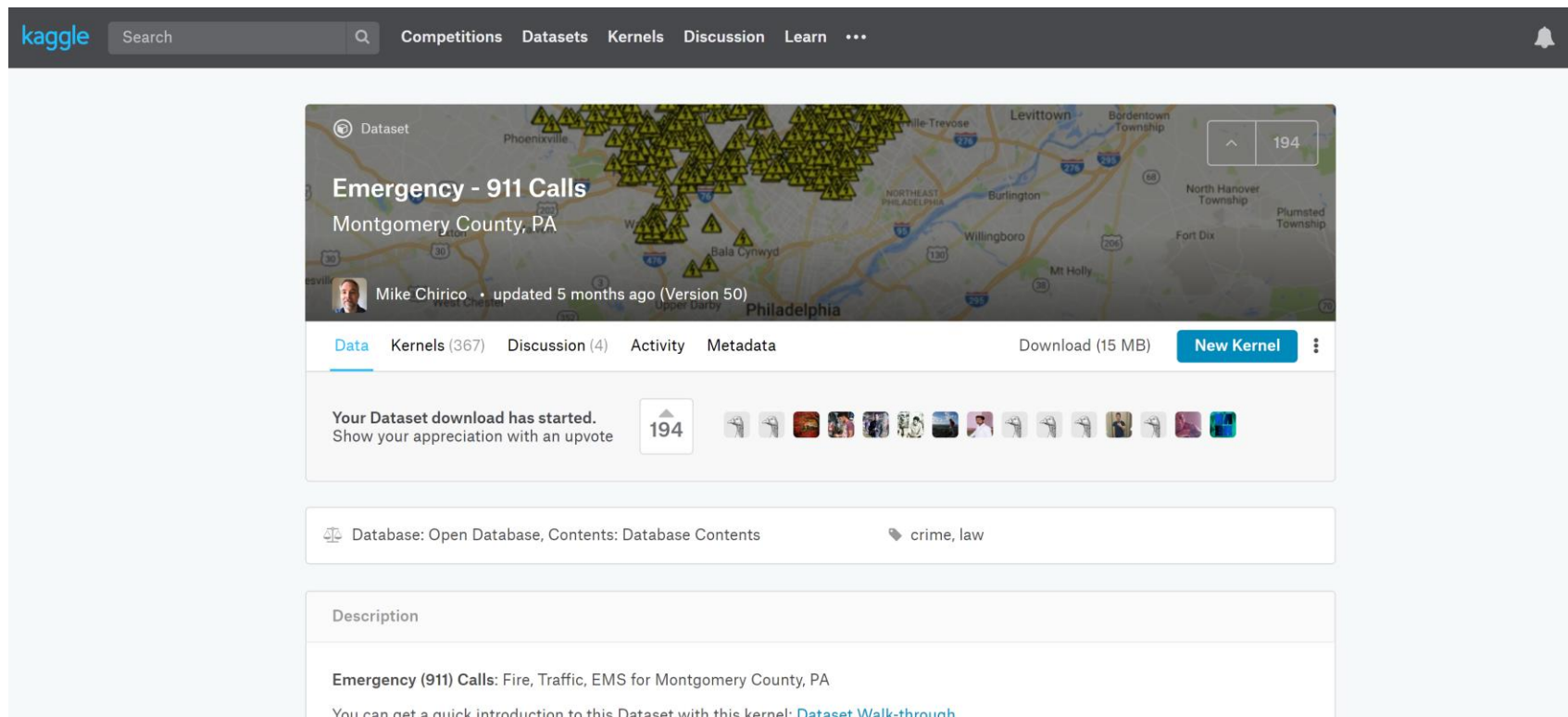
# (Feature Scalling)

```
from sklearn.preprocessing import StandardScaler  
sc_X = StandardScaler()  
X_train = sc_X.fit_transform(X_train)  
X_test = sc_X.transform(X_test)
```

# Parte III – Projeto de análise de dados

Neste exemplo, nós vamos utilizar o dataset que se encontra no site [4]:

<https://www.kaggle.com/>



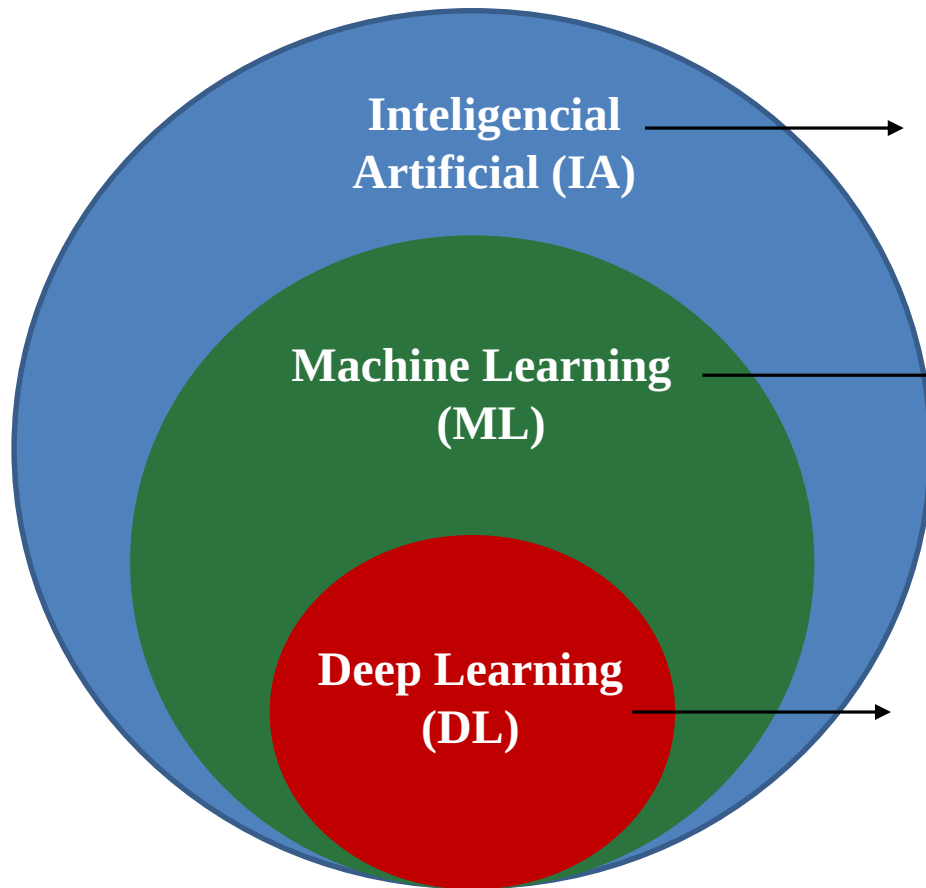
The screenshot displays the Kaggle website interface. At the top, there is a navigation bar with the Kaggle logo, a search bar, and links to Competitions, Datasets, Kernels, Discussion, and Learn. The main content area features a map of Montgomery County, PA, with numerous yellow triangle markers representing emergency calls. Below the map, the dataset title 'Emergency - 911 Calls' is shown, along with the creator's name 'Mike Chirico' and the update date 'updated 5 months ago (Version 50)'. A row of tabs includes 'Data', 'Kernels (367)', 'Discussion (4)', 'Activity', and 'Metadata'. To the right, there is a 'Download (15 MB)' button and a 'New Kernel' button. A notification box states 'Your Dataset download has started. Show your appreciation with an upvote' and shows a count of 194 upvotes. Below this, there is a section for 'Database: Open Database, Contents: Database Contents' and a tag 'crime, law'. The 'Description' section at the bottom states 'Emergency (911) Calls: Fire, Traffic, EMS for Montgomery County, PA' and provides a link to a 'Dataset Walk-through' kernel.

# Parte III – Projeto de análise de dados

**Os exemplos deste projeto estão descrito no arquivo “projeto-911.py”**



# Parte IV - Introdução a Machine Learning



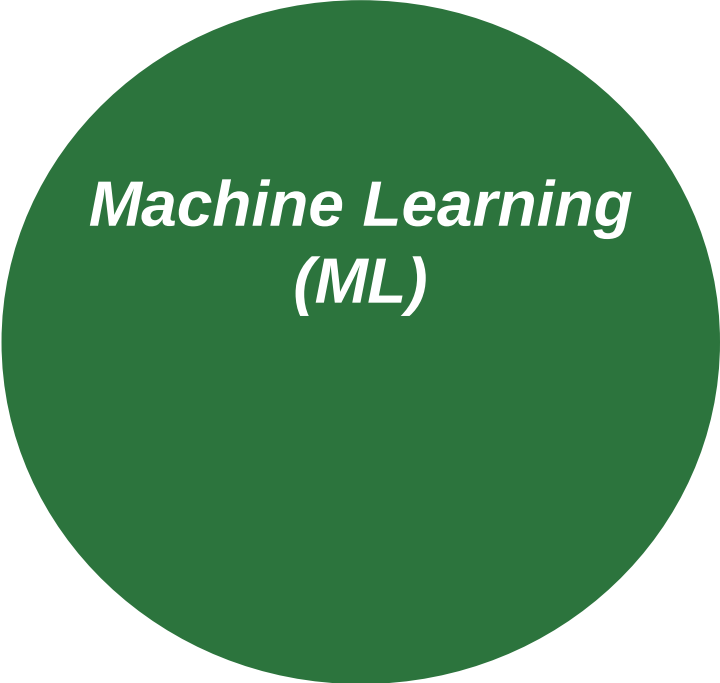
IA é a área de ciência da computação concentrada em criar programas e máquinas que podem exibir comportamentos considerados inteligentes.

ML é um campo de estudo que pertence à inteligência artificial e que tem como propósito criar sistemas capazes de aprenderem sozinhos.

DL é a criação de uma rede neural artificial que permite à máquina uma semelhança cognitiva com o cérebro humano, a fim de interpretar e interligar dados.

# Introdução a Machine Learning

Machine Learning é a ciência de fazer com que computadores realizem ações sem programação explícita.

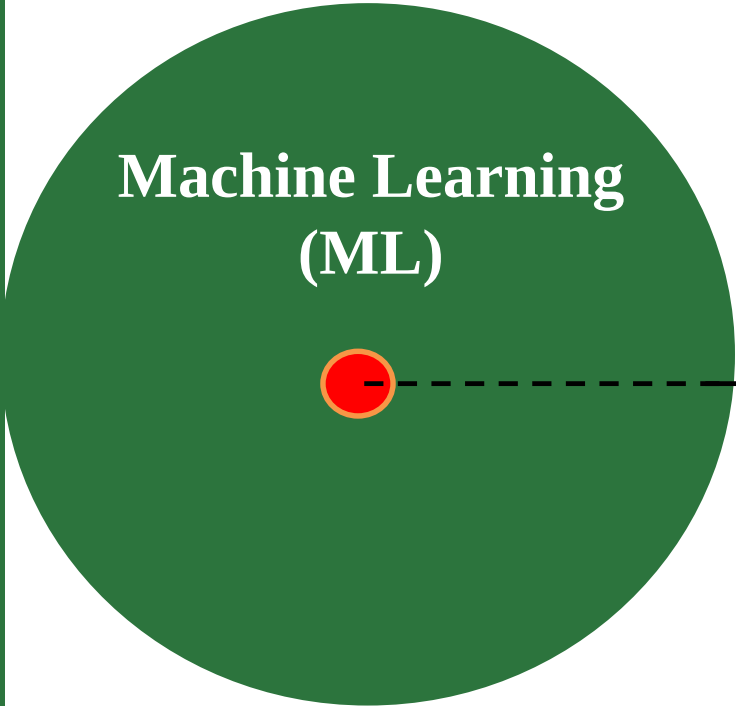


*Machine Learning*  
(ML)

A idéia central aqui é que os algoritmos de *Machine Learning* possam ser abastecidos com dados e então aprender com eles para fazer previsões ou orientar decisões.

# Introdução a Machine Learning

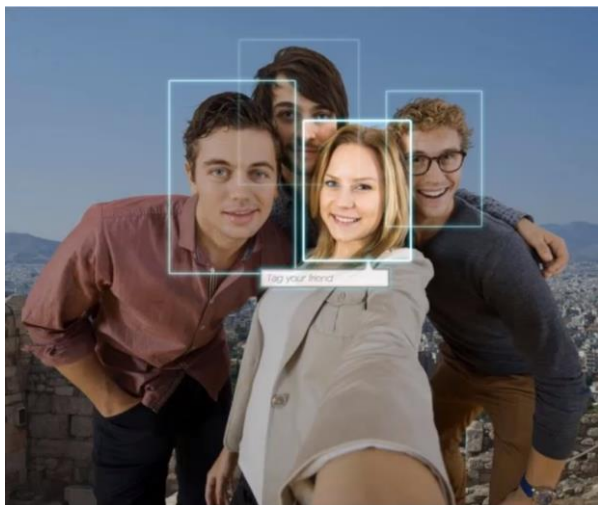
Machine Learning é a ciência de fazer com que computadores realizem ações sem programação explícita.



Machine Learning  
(ML)

A idéia central aqui é que os algoritmos de *Machine Learning* possam ser abastecidos com dados e então aprender com eles para fazer previsões ou orientar decisões.

# Introdução a Machine Learning



Reconhecimento facial do Facebook

Jogos



# Introdução a Machine Learning



**Robôs**



**Facebook Advertise**

# Métodos de Aprendizagem

Supervisionado

O algoritmo recebe um conjunto de dados (treinamento e teste) que definem aquilo que deverá ser buscado (classificado) pelo algoritmo

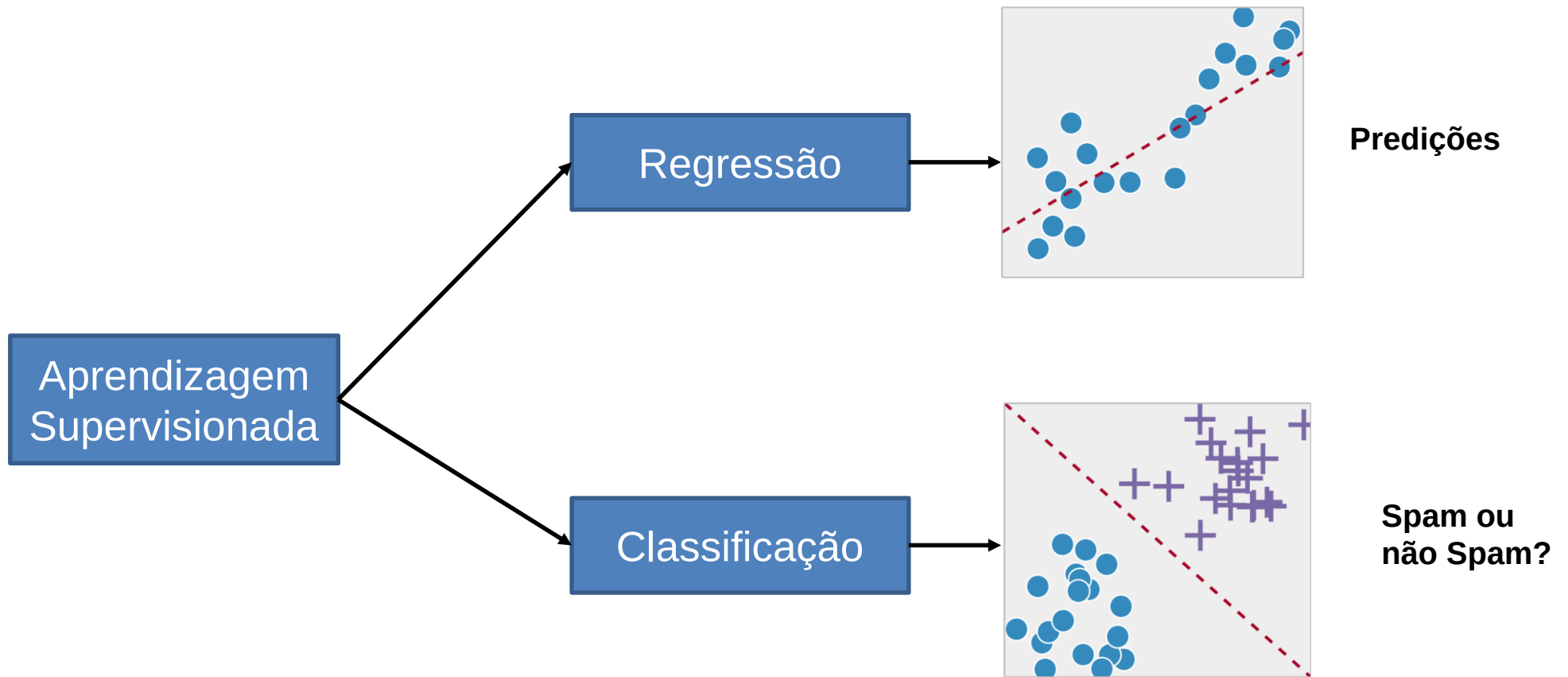
Não-Supervisionado

O algoritmo desenvolve suas próprias conclusões a partir de um conjunto de dados

Semi-Supervisionado

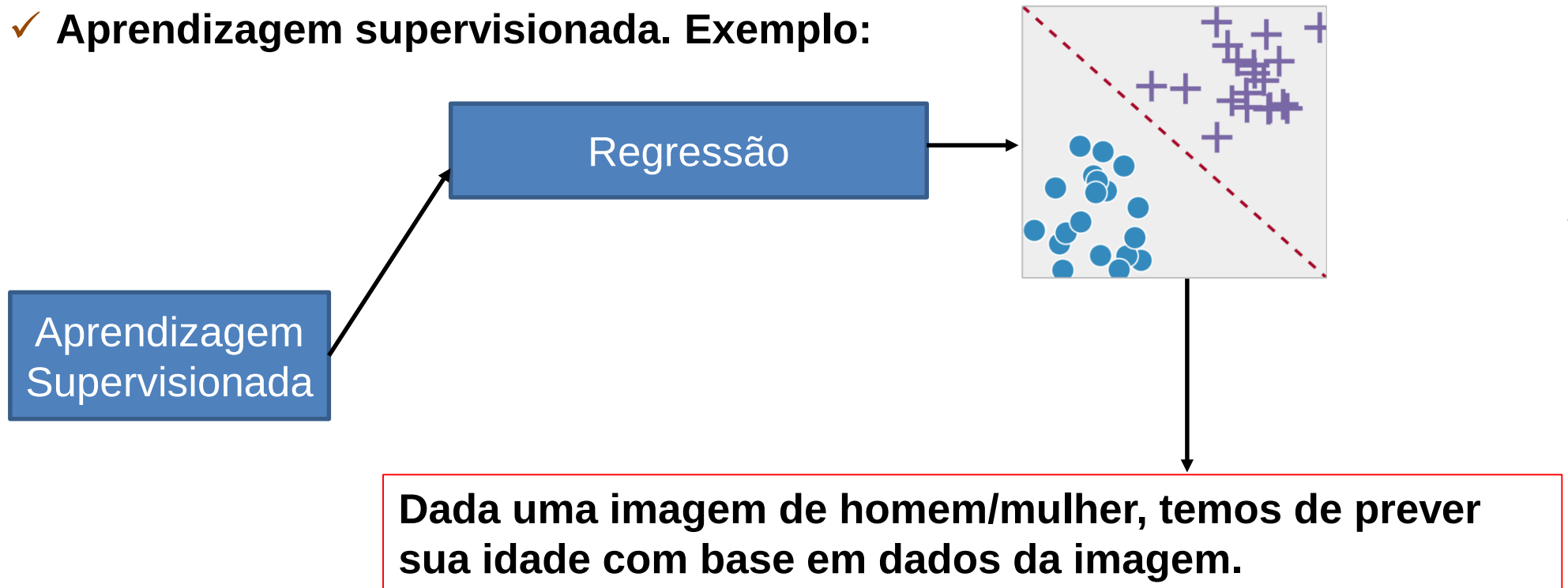
Uma mistura dos métodos supervisionados e não supervisionados

# Aprendizagem de Máquina Supervisionada



# Aprendizagem de Máquina Supervisionada

✓ Aprendizagem supervisionada. Exemplo:

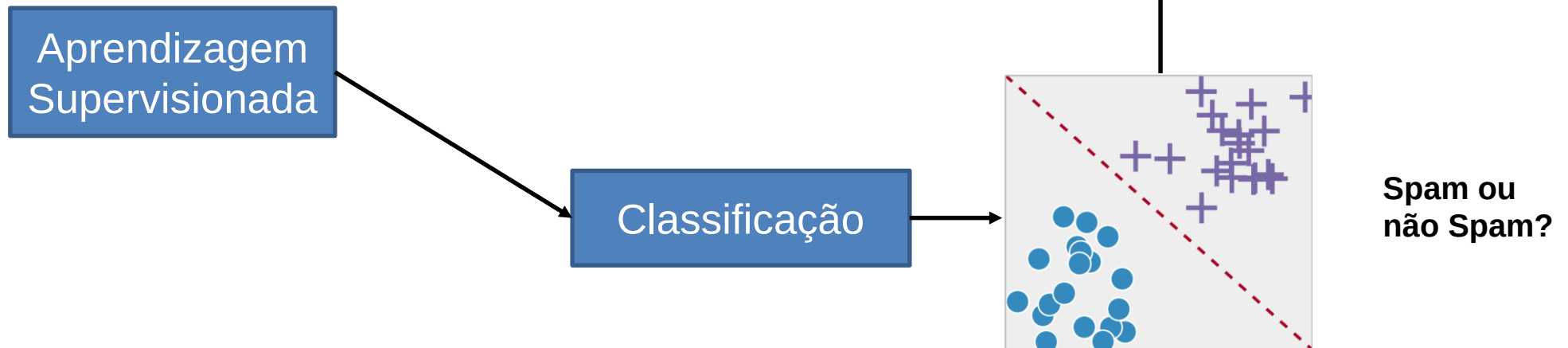




# Aprendizagem de Máquina Supervisionada

✓ Aprendizagem supervisionada. Exemplo:

Dada o conteúdo do email, classificar em spam ou não spam

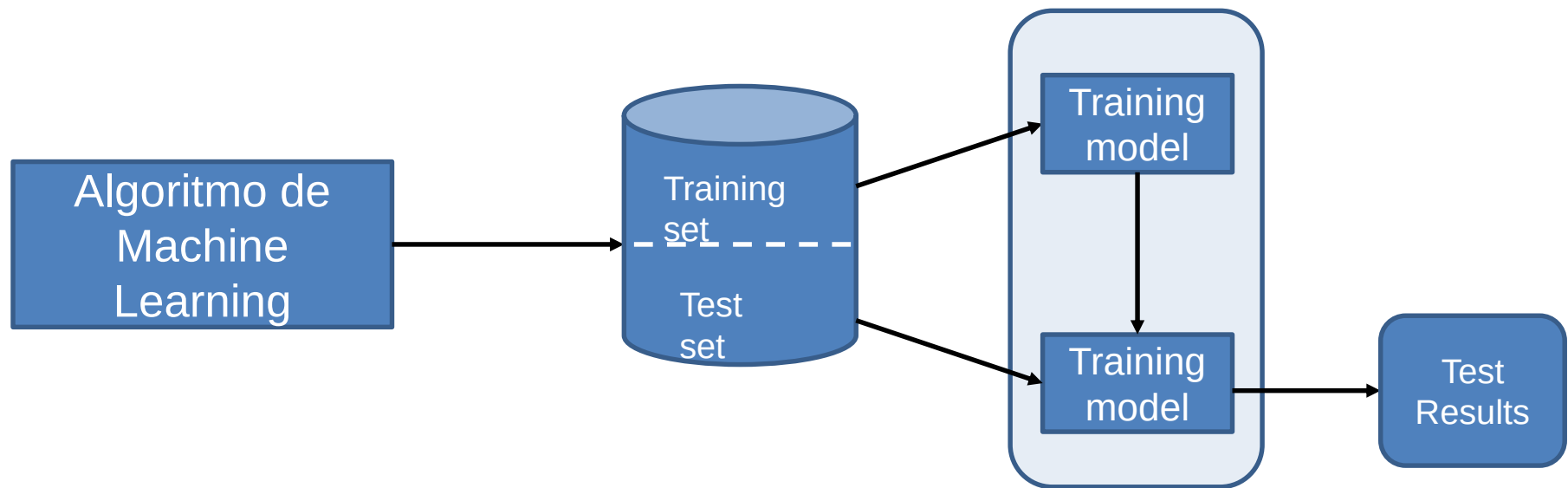


# Aprendizagem de Máquina Supervisionada

- ✓ **Alguns exemplos de algoritmos**
  - ✓ Árvore de decisão (*Decision Tree*)
  - ✓ Naïve Bayes
  - ✓ Regressão Linear (*Linear Regression*)
  - ✓ Regressão Logística (*Logistic Regression*)
  - ✓ SVM (*Support Vector Machine*)

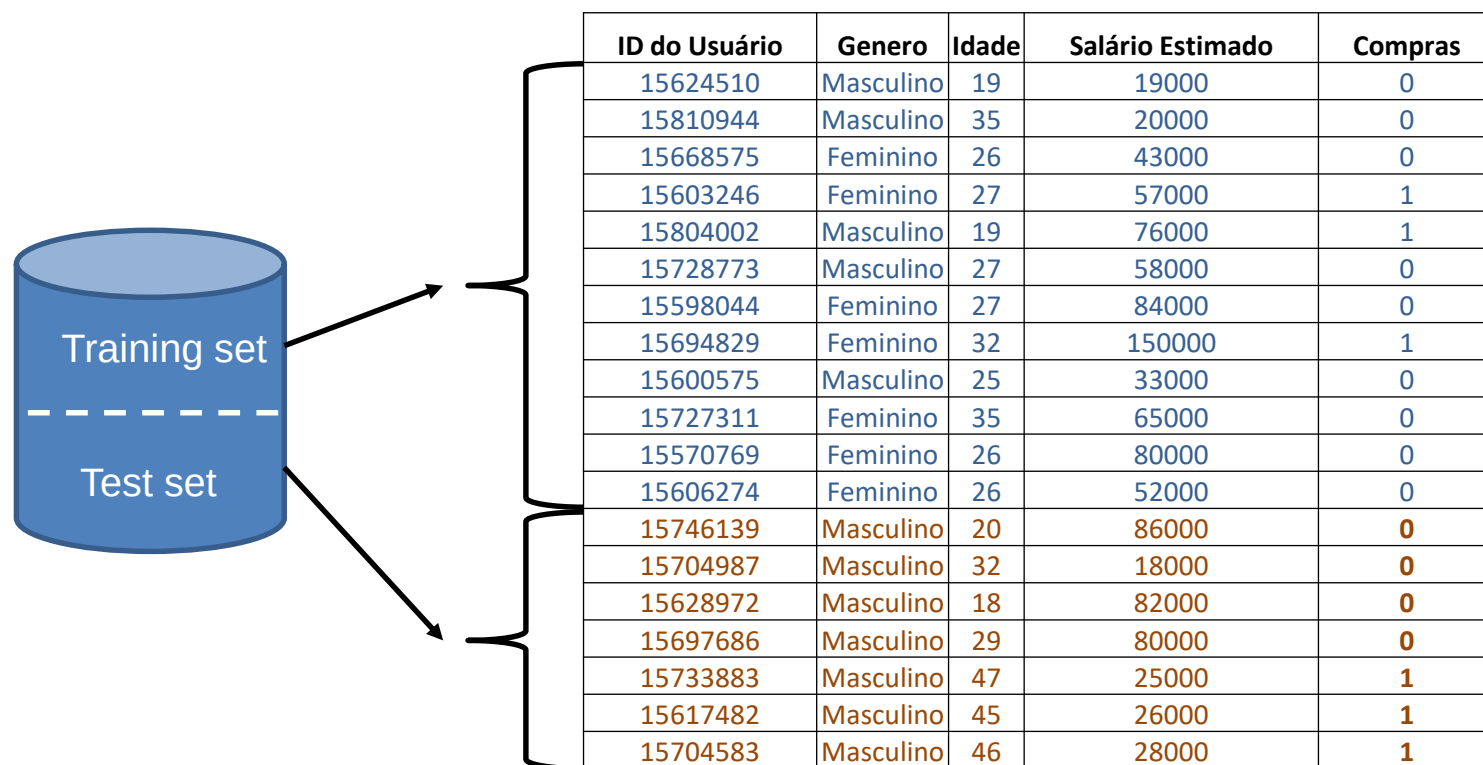
# Aprendizagem de Máquina Supervisionada

## ✓ Aprendizagem supervisionada na prática



# Aprendizagem de Máquina Supervisionada

## ✓ Aprendizagem supervisionada na prática



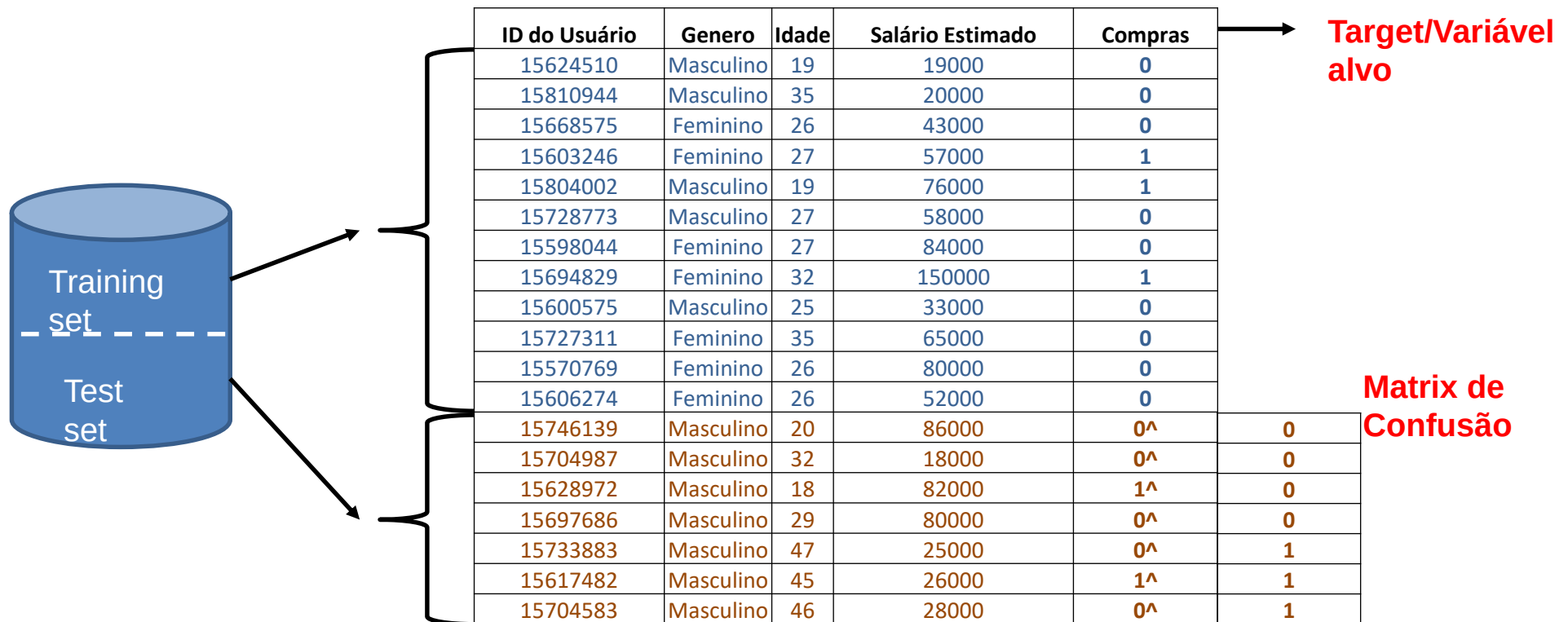
# Aprendizagem de Máquina Supervisionada

## ✓ Aprendizagem supervisionada na prática



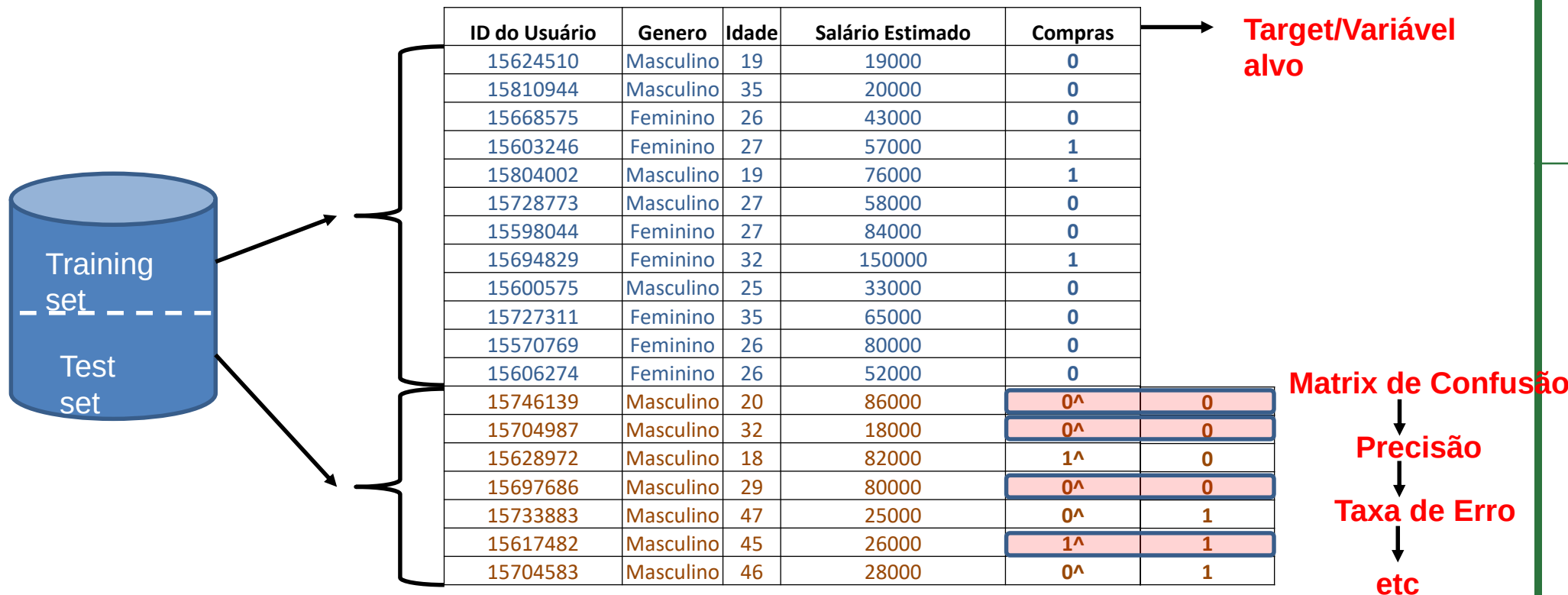
# Aprendizagem de Máquina Supervisionada

## ✓ Aprendizagem supervisionada na prática



# Aprendizagem de Máquina Supervisionada

## ✓ Aprendizagem supervisionada na prática



# Aprendizagem de Máquina Não-Supervisionada

## ✓ Aprendizagem não supervisionada

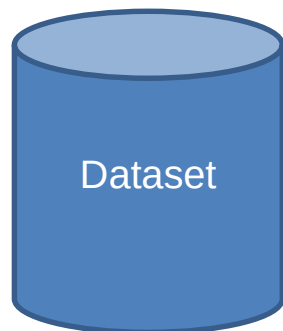
- ✓ O algoritmo detecta grupos baseado nas características dos dados encontrados no dataset
- ✓ Não é necessário definir a variável **Target**.
- ✓ Exemplos:
  - ✓ Algoritmos de Agrupamento (Clustering)
  - ✓ K-means,
  - ✓ Análise de Componentes Principais: (PCA)



# Aprendizagem de Máquina Não-Supervisionada

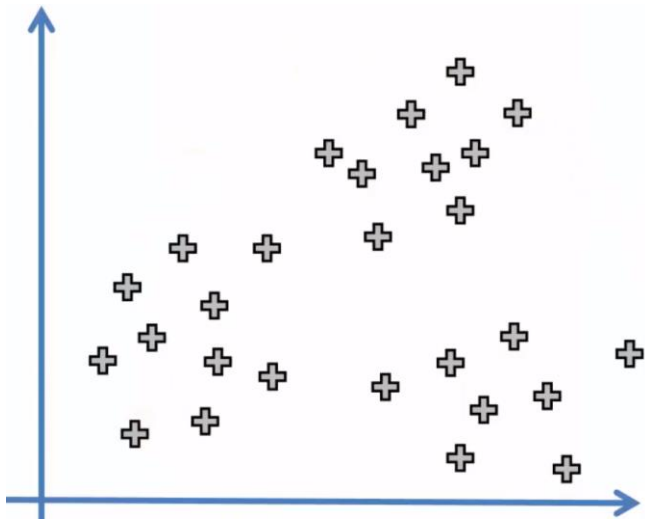
## ✓ Aprendizagem não supervisionada na prática

| CustomerID | Genero    | Idade | Renda Anual (Real (k)) | Pontuacao de Compra (0 - 100) |
|------------|-----------|-------|------------------------|-------------------------------|
| 1          | Masculino | 19    | 15                     | 39                            |
| 2          | Masculino | 21    | 15                     | 81                            |
| 3          | Feminino  | 20    | 16                     | 6                             |
| 4          | Feminino  | 23    | 16                     | 77                            |
| 5          | Feminino  | 31    | 17                     | 40                            |
| 6          | Feminino  | 22    | 17                     | 76                            |
| 7          | Feminino  | 35    | 18                     | 6                             |
| 8          | Feminino  | 23    | 18                     | 94                            |
| 9          | Masculino | 64    | 19                     | 3                             |
| 10         | Feminino  | 30    | 19                     | 72                            |
| 11         | Masculino | 67    | 19                     | 14                            |
| 12         | Feminino  | 35    | 19                     | 99                            |
| 13         | Feminino  | 58    | 20                     | 15                            |
| 14         | Feminino  | 24    | 20                     | 77                            |
| 15         | Masculino | 37    | 20                     | 13                            |
| 16         | Masculino | 22    | 20                     | 79                            |
| 17         | Feminino  | 35    | 21                     | 35                            |
| 18         | Masculino | 20    | 21                     | 66                            |
| 19         | Masculino | 52    | 23                     | 29                            |
| 20         | Feminino  | 35    | 23                     | 98                            |
| 21         | Masculino | 35    | 24                     | 35                            |
| 22         | Masculino | 25    | 24                     | 73                            |
| 23         | Feminino  | 46    | 25                     | 5                             |
| 24         | Masculino | 31    | 25                     | 73                            |
| 25         | Feminino  | 54    | 28                     | 14                            |
| 26         | Masculino | 29    | 28                     | 82                            |



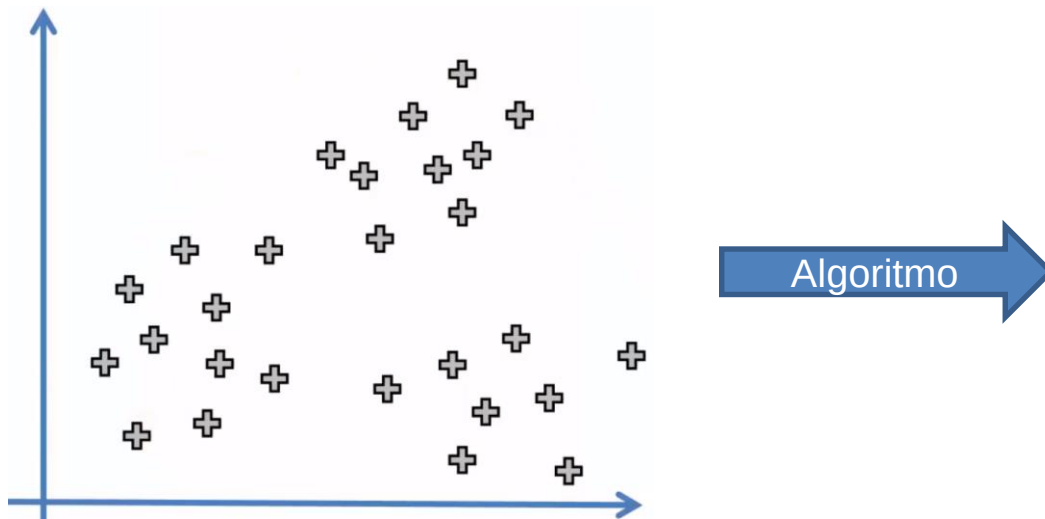
# Aprendizagem de Máquina Não-Supervisionada

## ✓ Aprendizagem não supervisionada



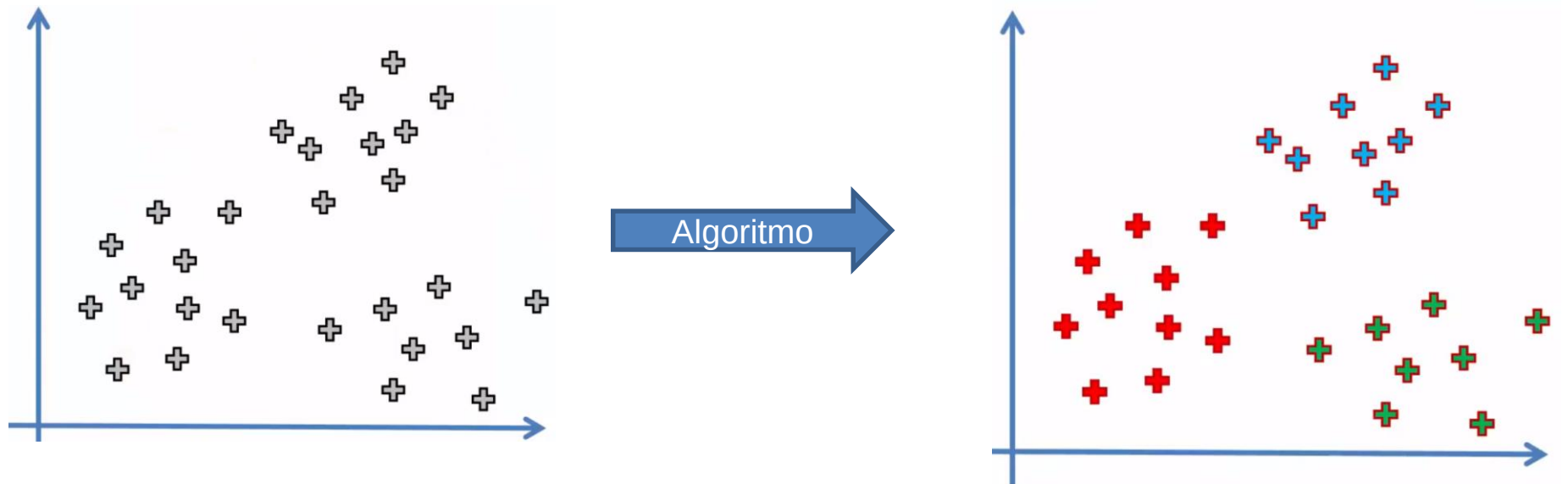
# Aprendizagem de Máquina Não-Supervisionada

## ✓ Aprendizagem não supervisionada



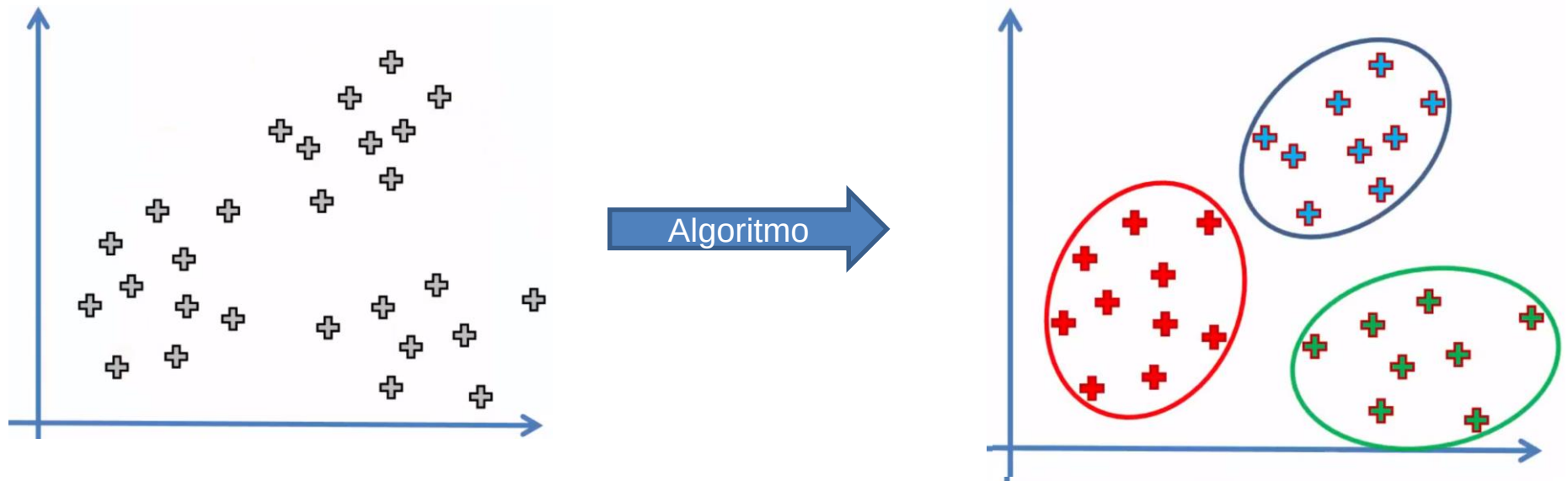
# Aprendizagem de Máquina Não-Supervisionada

## ✓ Aprendizagem não supervisionada



# Aprendizagem de Máquina Não-Supervisionada

## ✓ Aprendizagem não supervisionada



# Regressão Linear Simples

**Descrição do Problema: Treinar um modelo de Machine learning para definir valores de salários de acordo com a experiência do trabalhador.**

**Algoritmos utilizado: Simple Linear Regression**

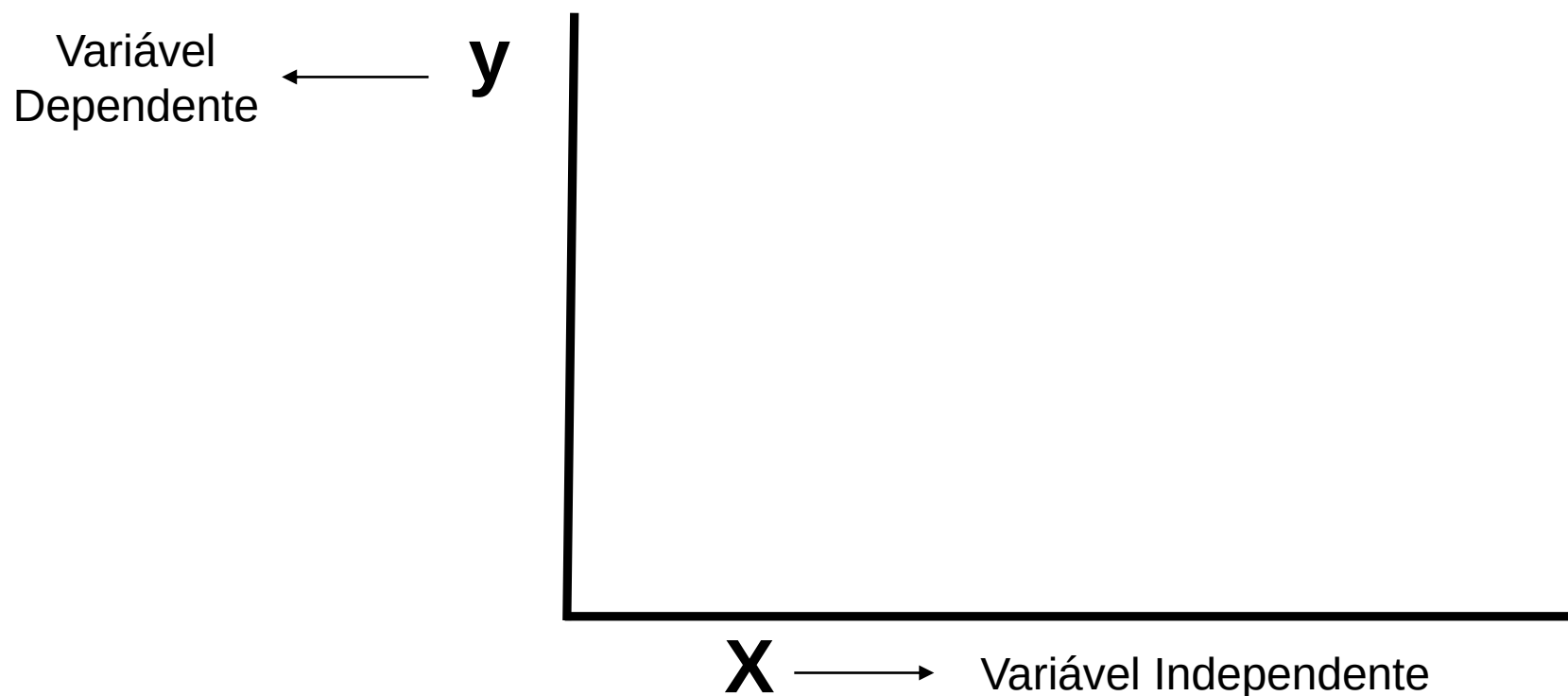
| Anos_Experiencia | Salario |
|------------------|---------|
| 1.1              | 39343   |
| 1.3              | 46205   |
| 1.5              | 37731   |
| 2                | 43525   |
| 2.2              | 39891   |
| 2.9              | 56642   |
| 3                | 60150   |
| 3.2              | 54445   |
| 3.2              | 64445   |
| 3.7              | 57189   |
| 3.9              | 63218   |
| 4                | 55794   |
| 4                | 56957   |

# Regressão Linear Simples

- ✓ Entre os modelos estatísticos utilizados pelos analistas, os modelos de regressão são os mais comuns.
- ✓ Um modelo de regressão permite estimar ou prever uma variável aleatória como uma função de várias outras variáveis.
- ✓ A variável estimada é chamada de **variável de resposta** e as variáveis usadas para prever a resposta são chamadas de variáveis **preditores ou fatores preditores**.

# Regressão Linear Simples

- ✓ Assim sendo, a análise de regressão estuda a relação entre uma variável chamada a **variável dependente** e outras variáveis chamadas **variáveis independentes**.





# Regressão Linear Simples

## Modelo de regressão linear simples - Correlação

- ✓ A correlação representa a forma que duas variáveis se relacionam.
- ✓ Para analisar a relação entre duas variáveis,  $X$  e  $y$ , podemos representar os valores das variáveis em um gráfico de dispersão.
- ✓ É possível afirmar a existência de uma relação linear se os dados se aproximarem de uma linha reta.

# Regressão Linear Simples

## Modelo de regressão linear simples - Correlação

- ✓ Para quantificar a relação entre duas variáveis quantitativas utiliza-se o coeficiente de correlação linear de Pearson:

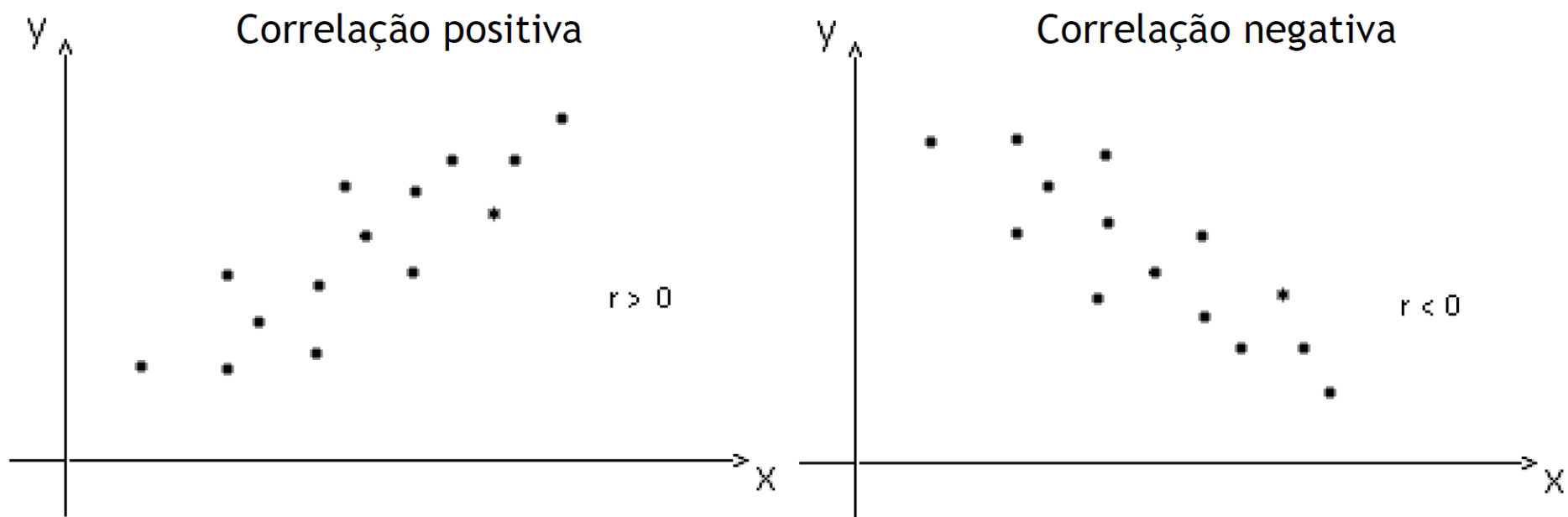
$$r = \frac{n \cdot \sum x \cdot y - (\sum x) \cdot (\sum y)}{\sqrt{[n \cdot \sum x^2 - (\sum x)^2] \cdot [n \cdot \sum y^2 - (\sum y)^2]}}$$

| Coeficiente          | Classificação |
|----------------------|---------------|
| $0,9 <  r  \leq 1,0$ | Ótima         |
| $0,8 <  r  \leq 0,9$ | Boa           |
| $0,7 <  r  \leq 0,8$ | Razoável      |
| $0,6 <  r  \leq 0,7$ | Medíocre      |
| $0,5 <  r  \leq 0,6$ | Péssima       |
| $ r  \leq 0,5$       | Imprópria     |

# Regressão Linear Simples

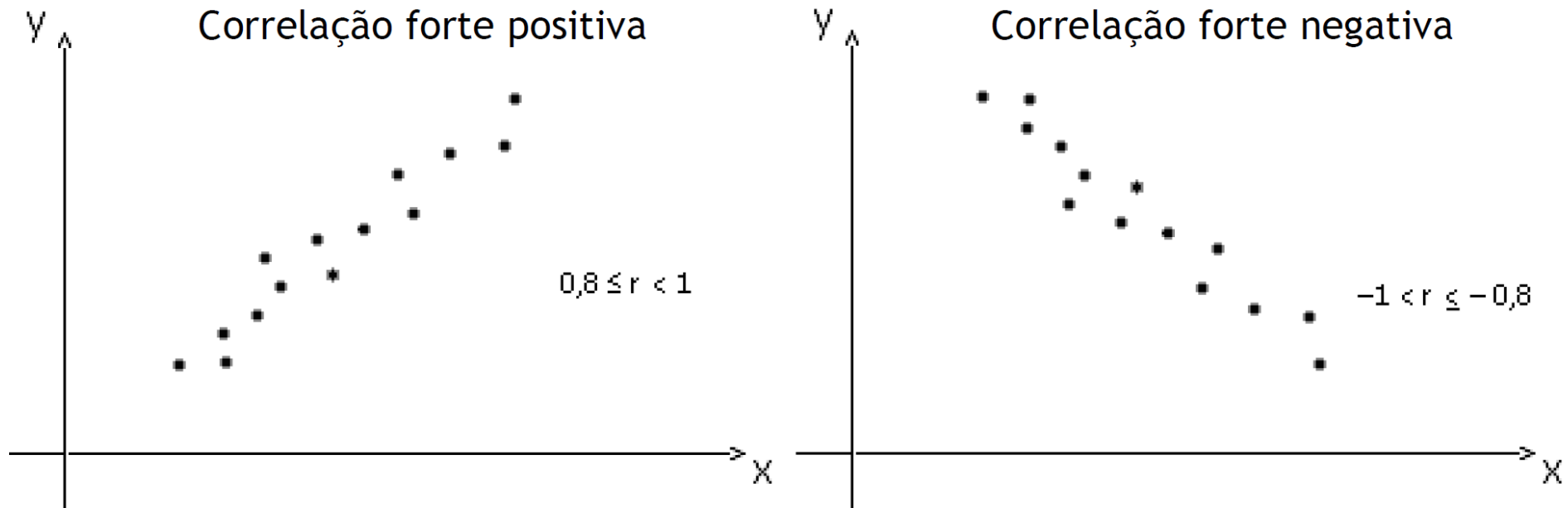
## Modelo de regressão linear simples - Correlação

- ✓ É possível afirmar a existência de uma relação linear se os dados se aproximarem de uma linha reta.



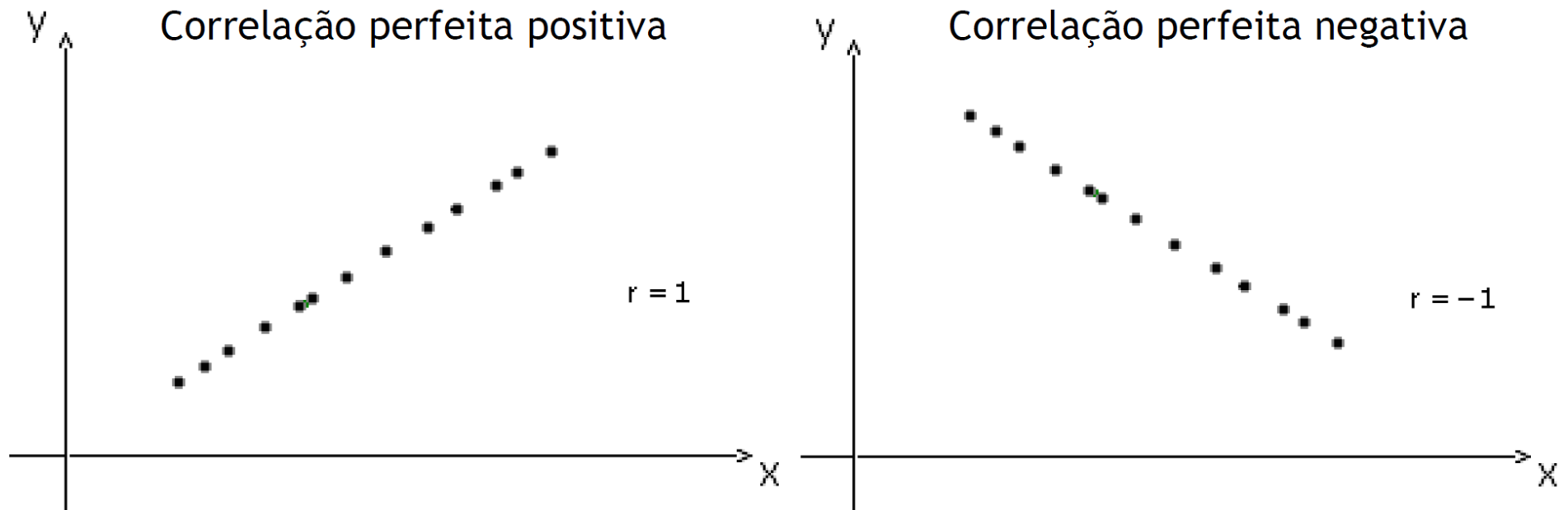
# Regressão Linear Simples

## Modelo de regressão linear simples - Correlação

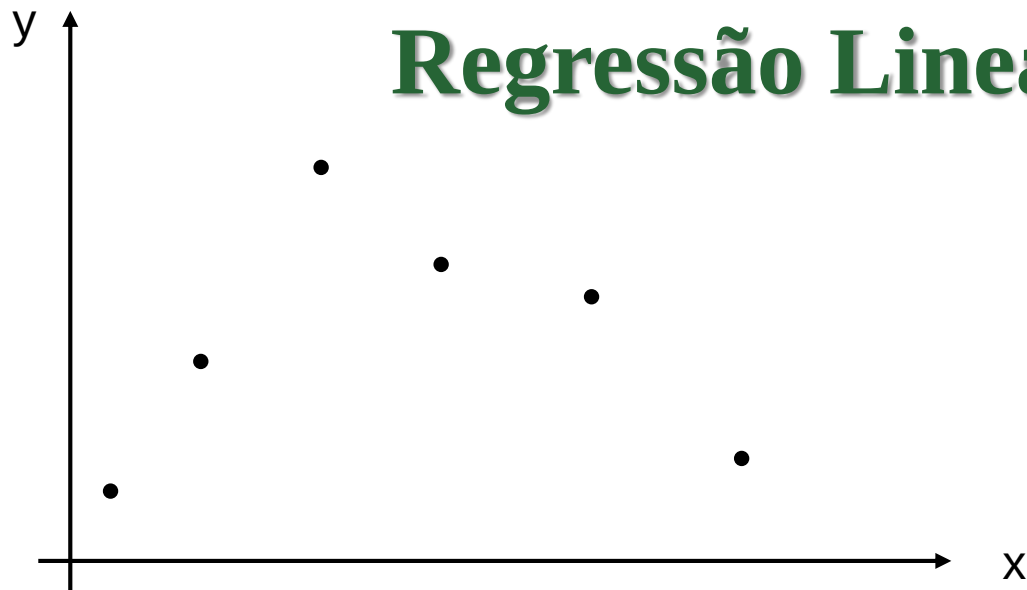


# Regressão Linear Simples

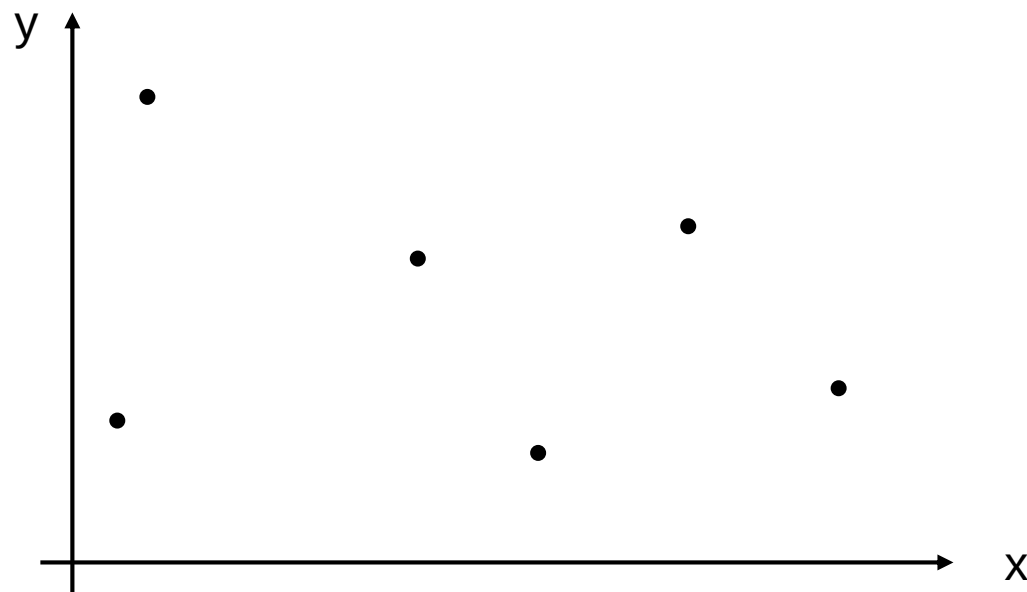
## Modelo de regressão linear simples - Correlação



# Regressão Linear Simples

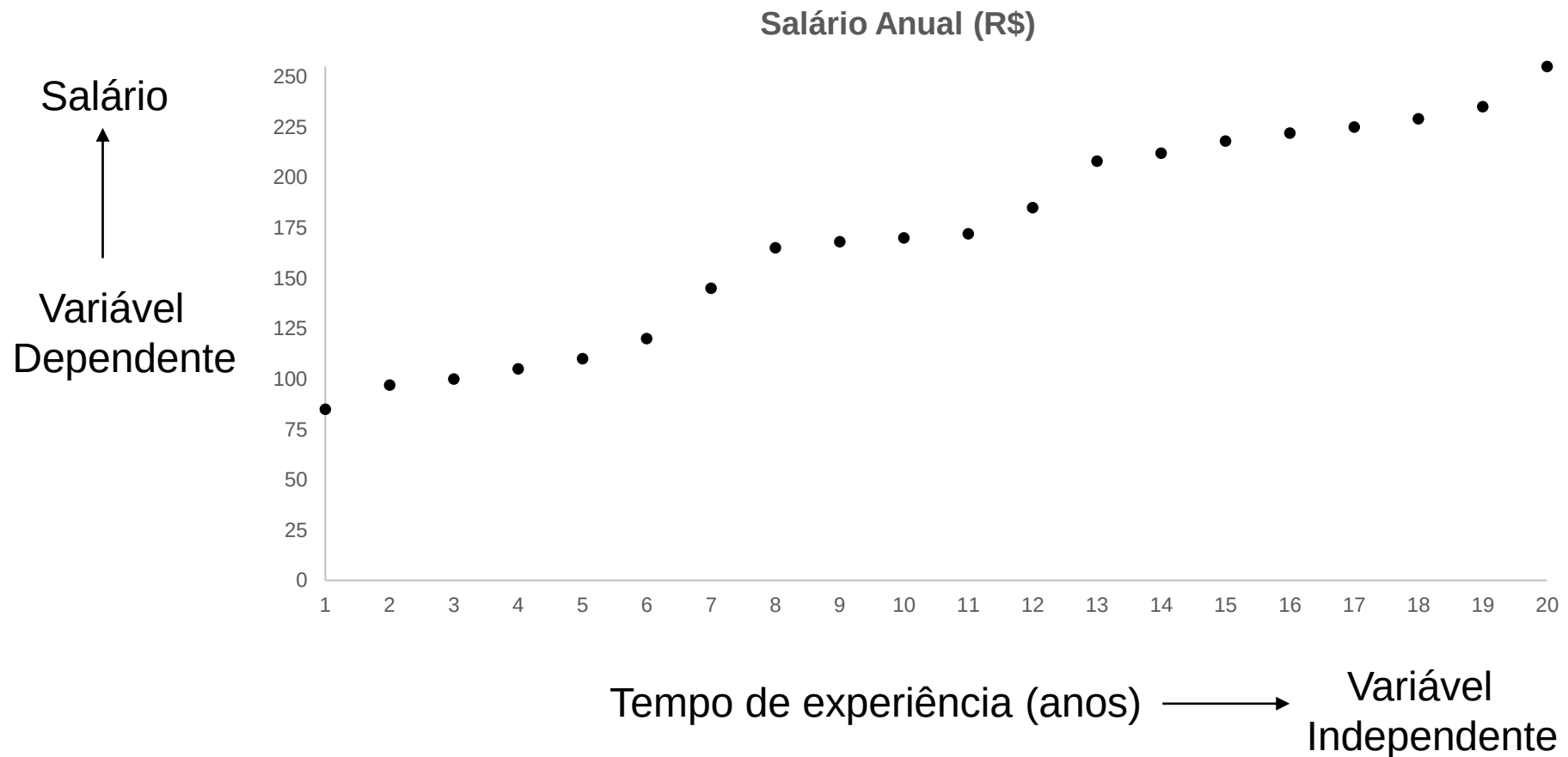


Correlação não linear

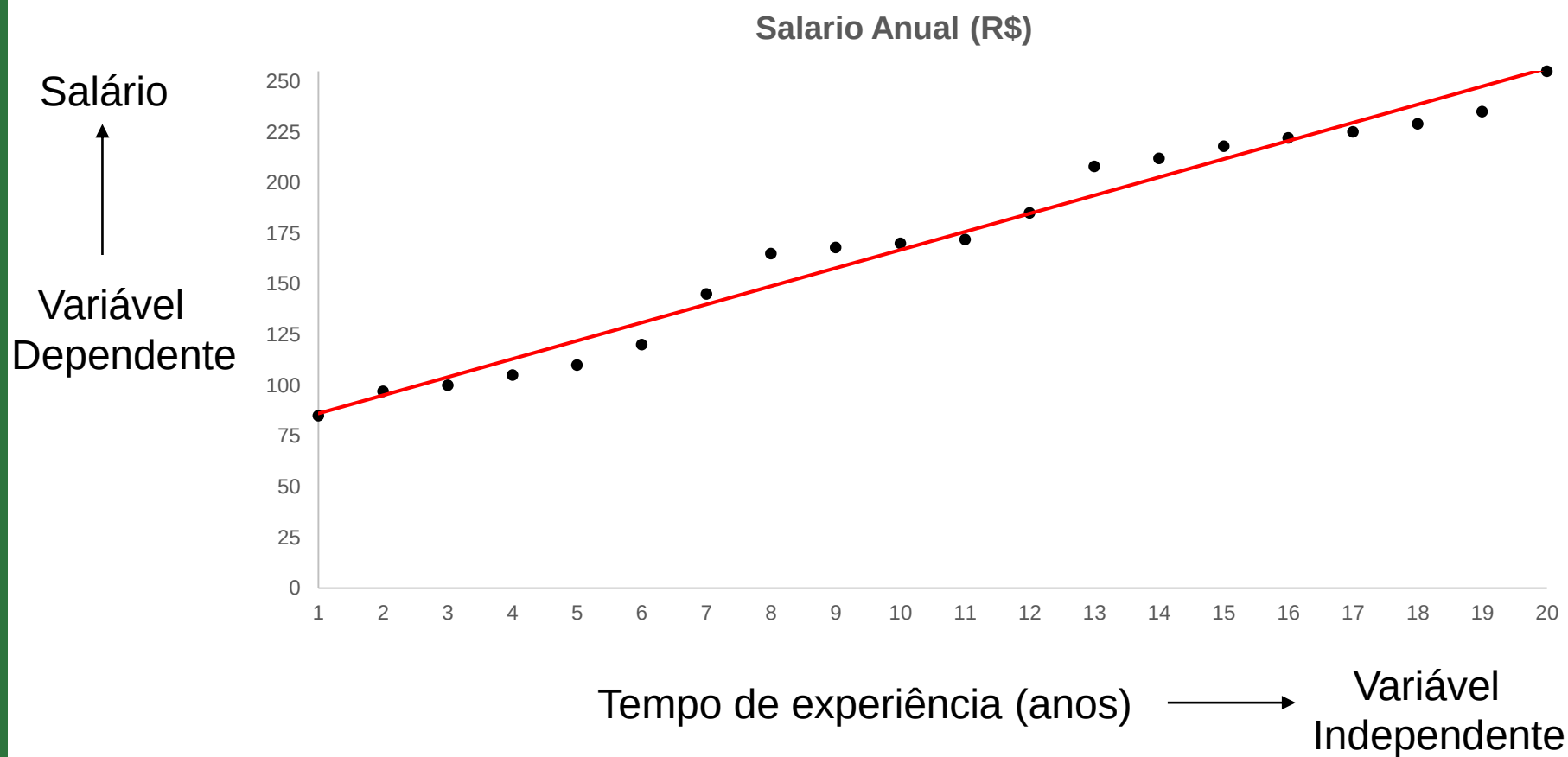


Não há correlação

# Regressão Linear Simples



# Regressão Linear Simples





# Regressão Linear Simples

## Calculando a reta ajustada

Diagram illustrating the components of the Simple Linear Regression equation:

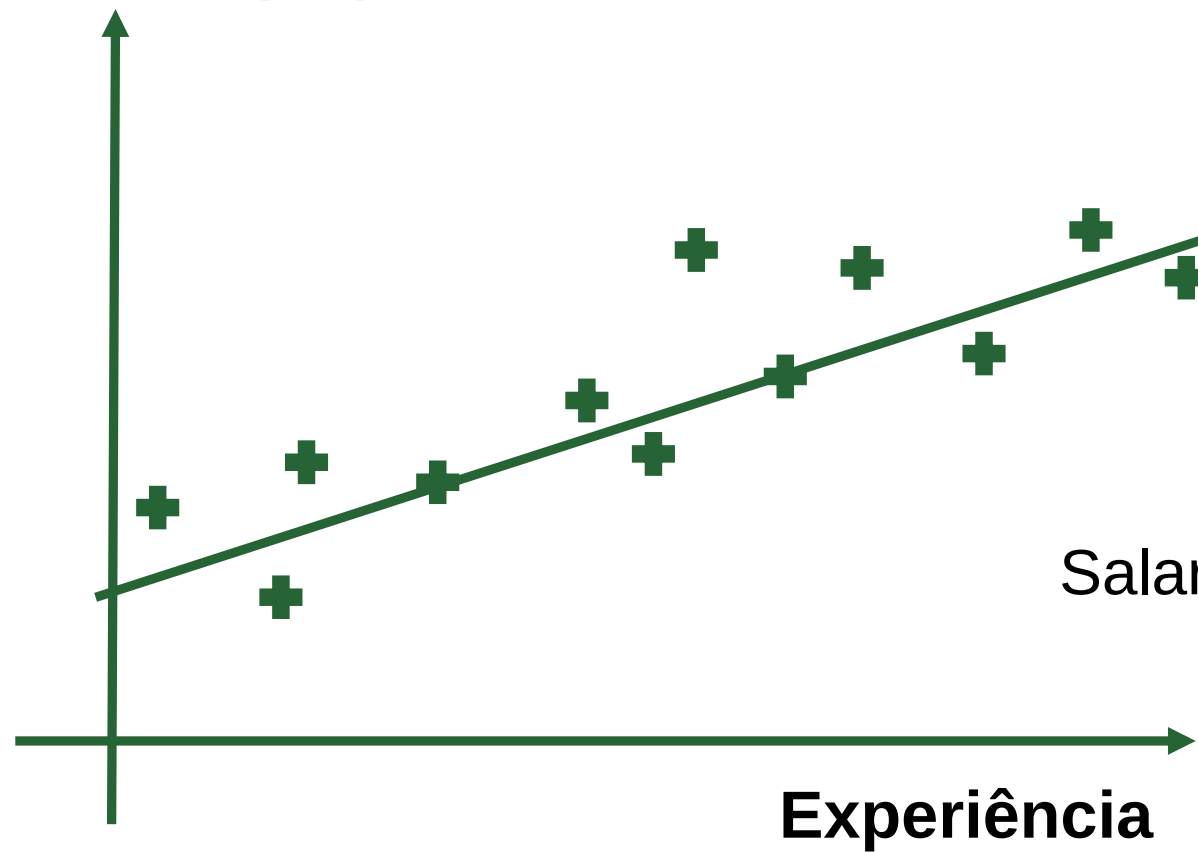
$$y = b_0 + b_1 * x_1$$

The diagram uses arrows to identify the components:

- Constante (coeficiente linear)**: Points to  $b_0$ .
- Coeficiente angular**: Points to  $b_1$ .
- Variável Dependente**: Points to  $y$ .
- Variável Independente**: Points to  $x_1$ .

# Regressão Linear Simples

Salário (R\$)



$$y = b_0 + b_1 * x_1$$

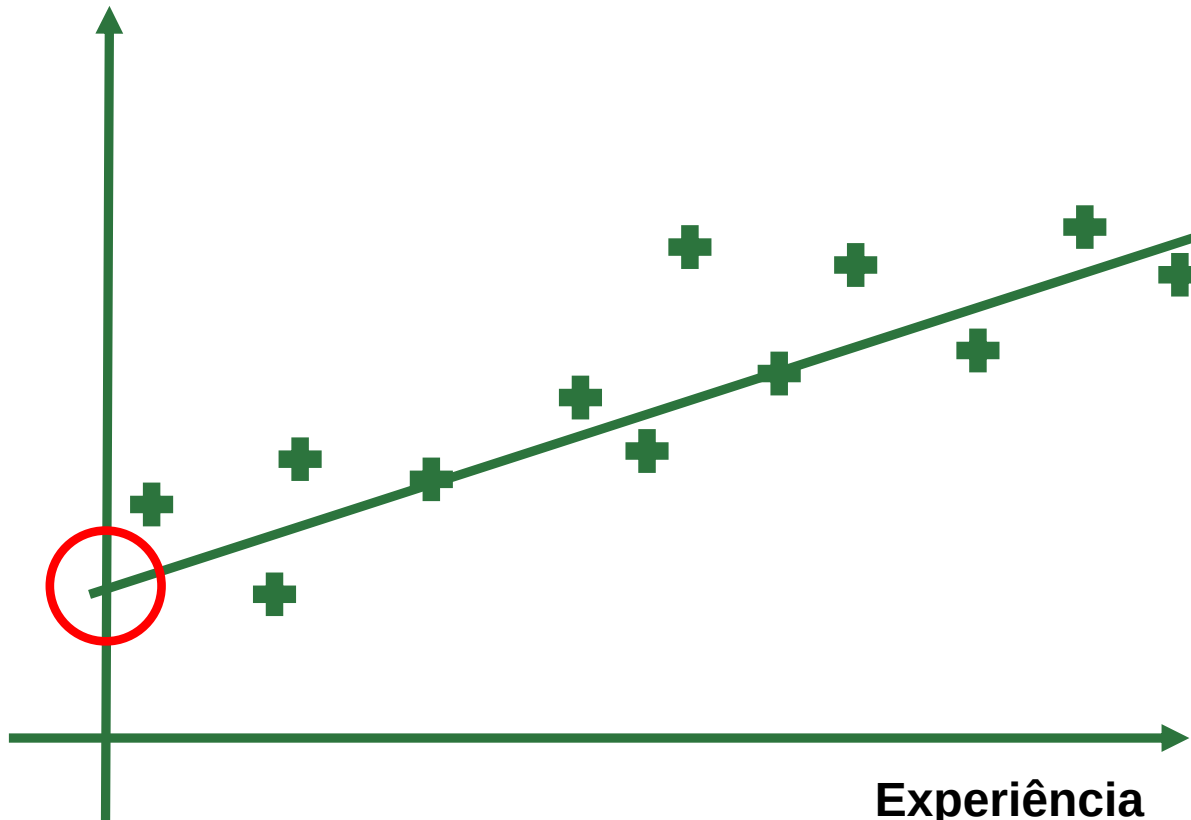


$$\text{Salario} = b_0 + b_1 * \text{Experiencia}_1$$

# Regressão Linear Simples

## Regressão linear simples

Salário (R\$)



$$y = b_0 + b_1 * x_1$$

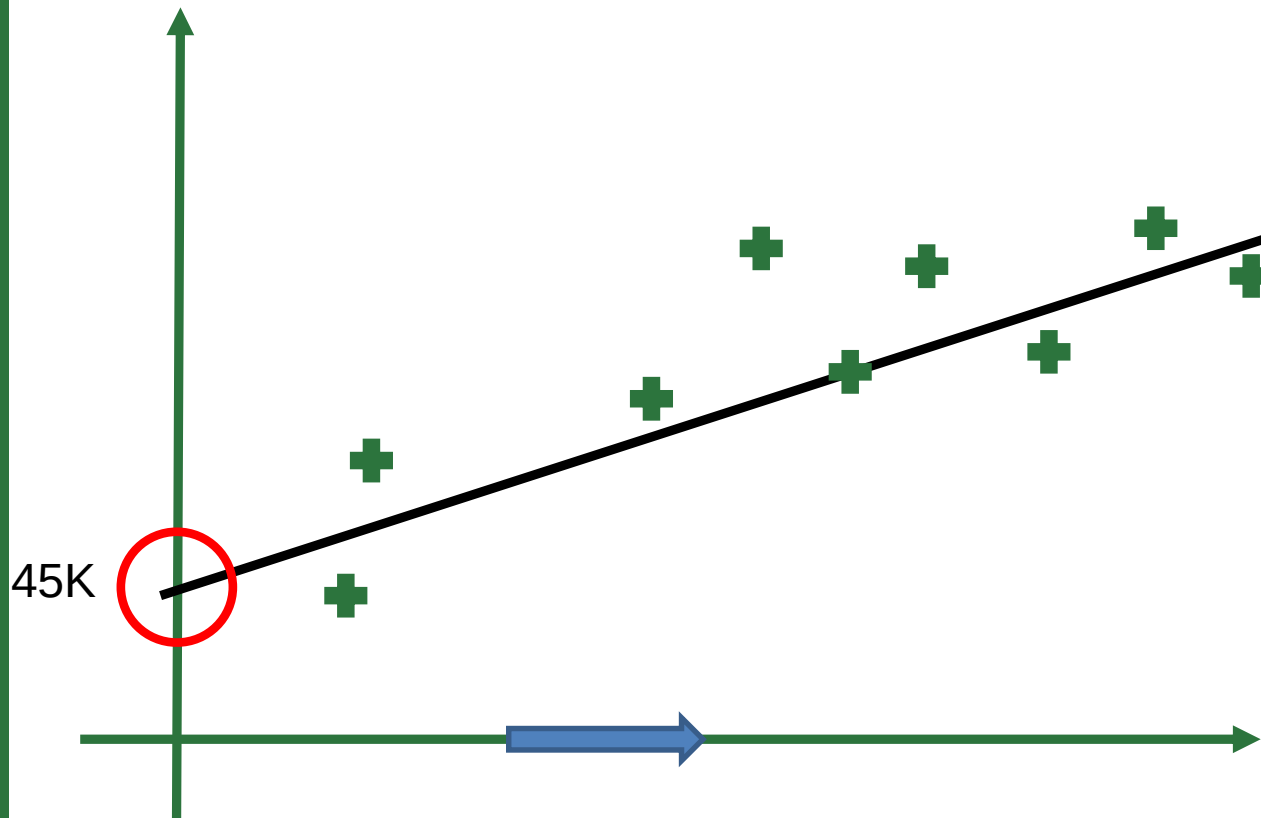


$$\text{Salario} = \textcolor{red}{b_0} + b_1 * \text{Experiencia}_1$$

# Regressão Linear Simples

## Regressão linear simples

Salário (R\$)



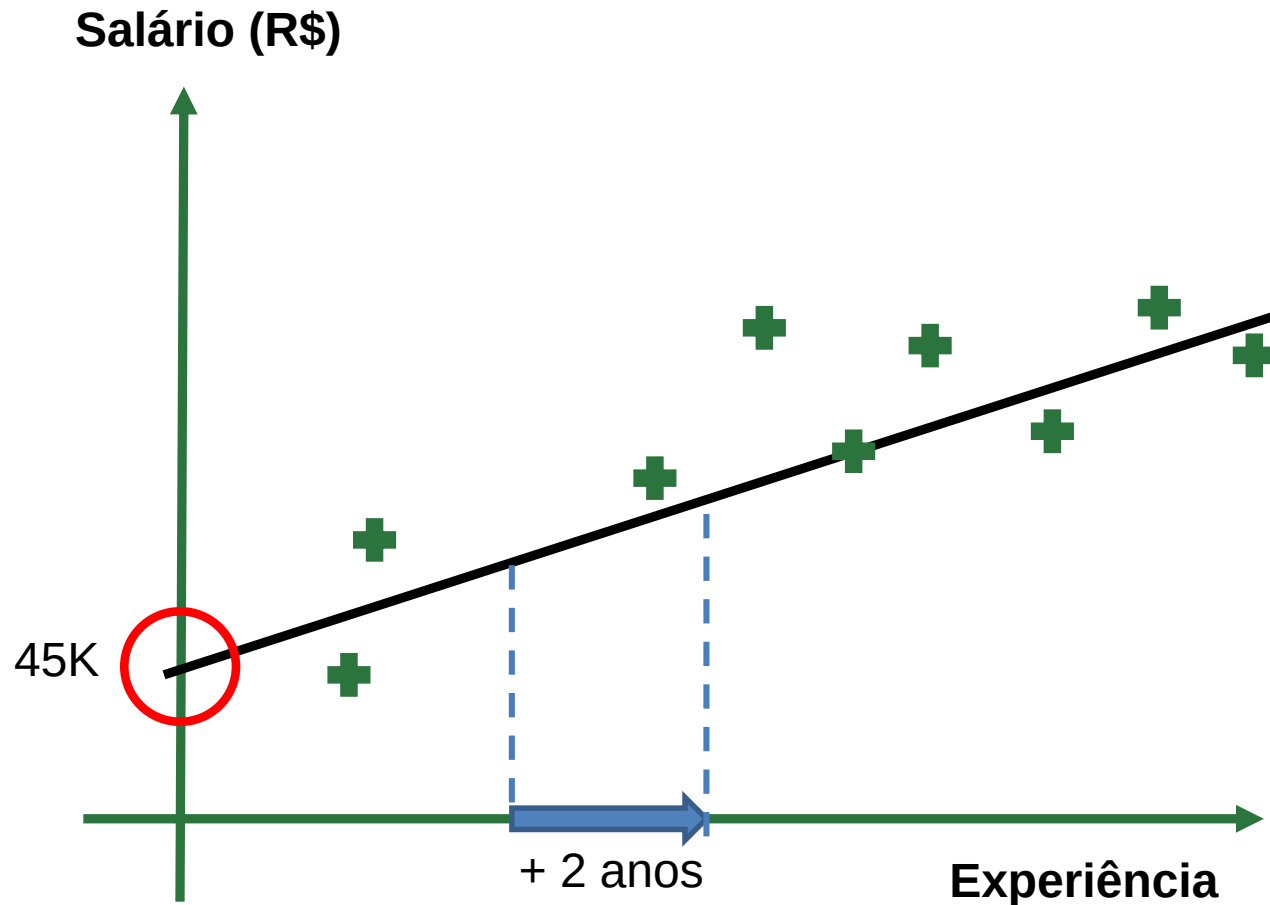
$$y = b_0 + b_1 * x_1$$



$$\text{Salario} = b_0 + b_1 * \text{Experiencia}_1$$

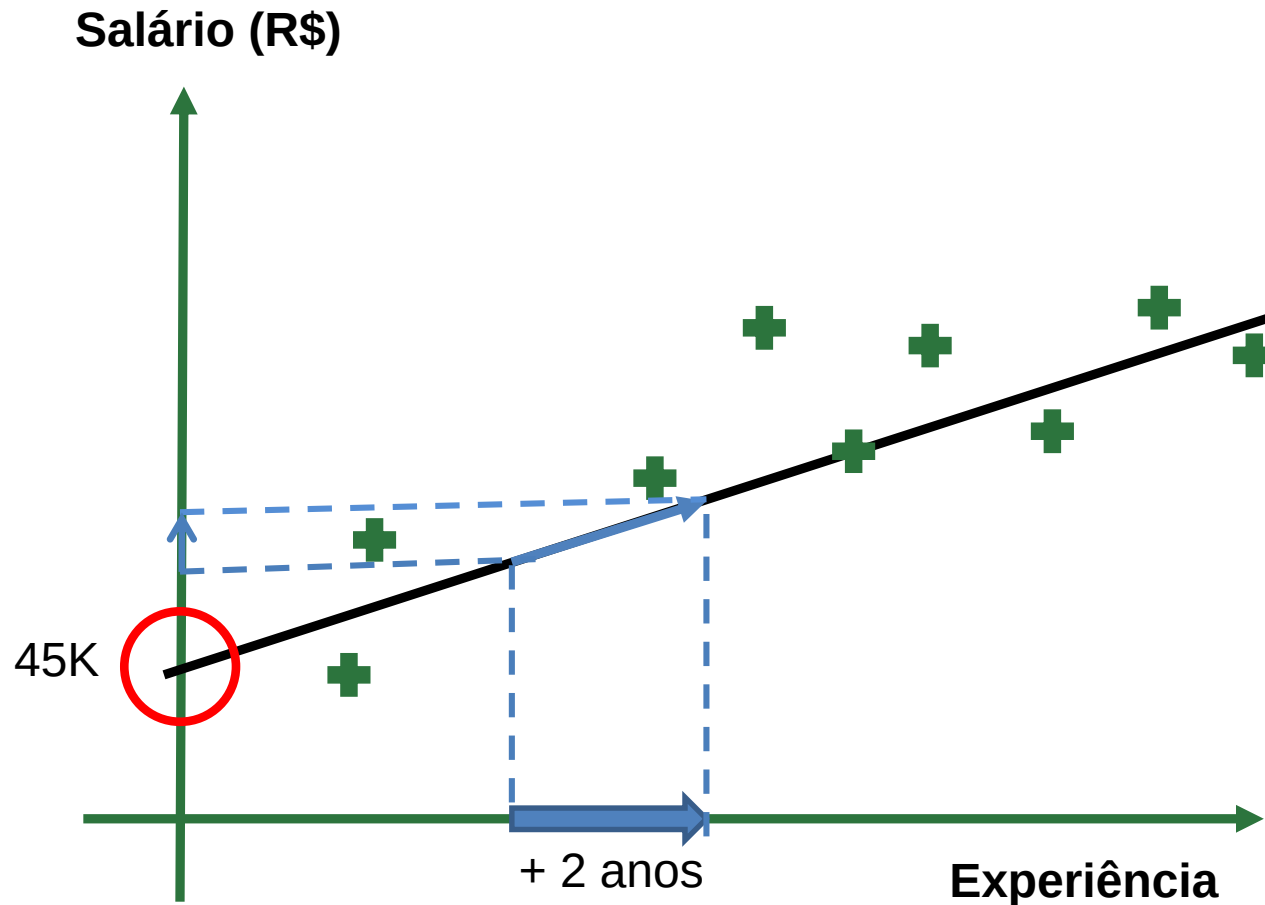
# Regressão Linear Simples

Regressão linear simples



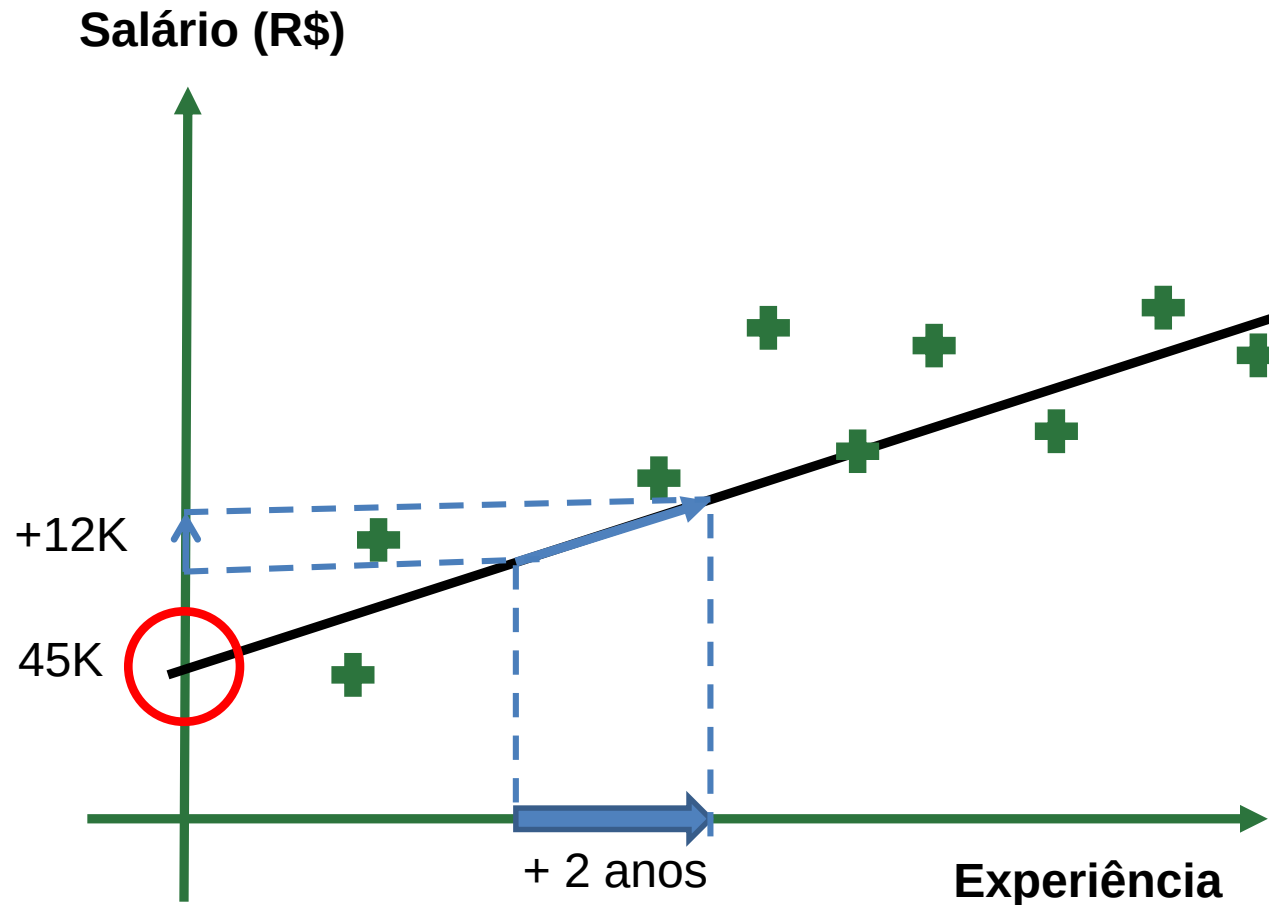
# Regressão Linear Simples

Regressão linear simples



# Regressão Linear Simples

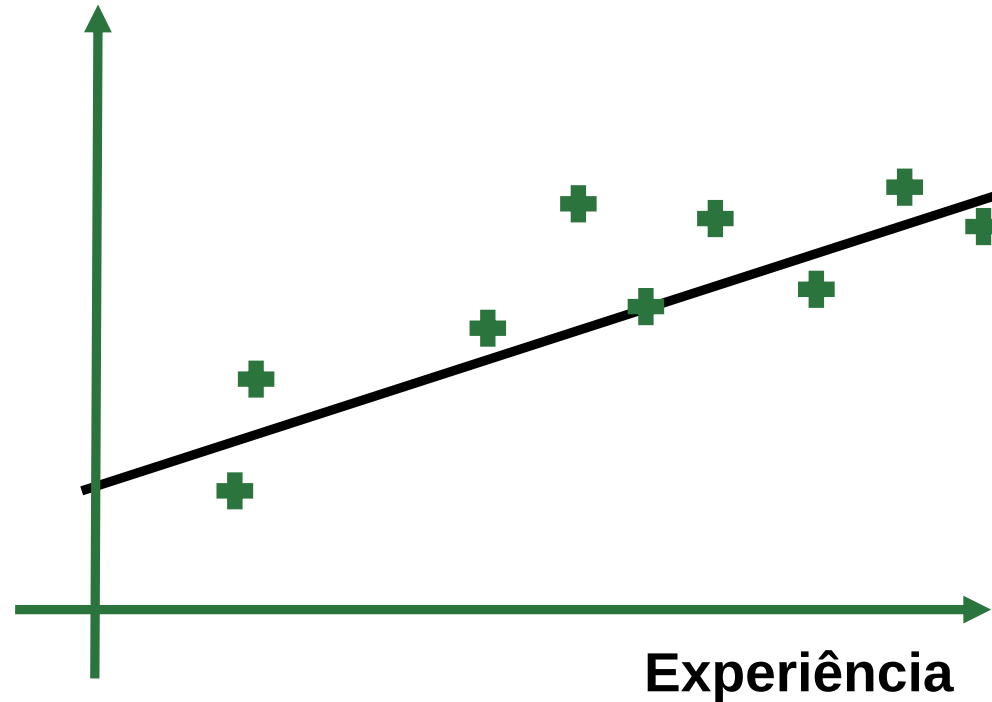
Regressão linear simples



# Regressão Linear Simples

Regressão linear simples

Salário (R\$)

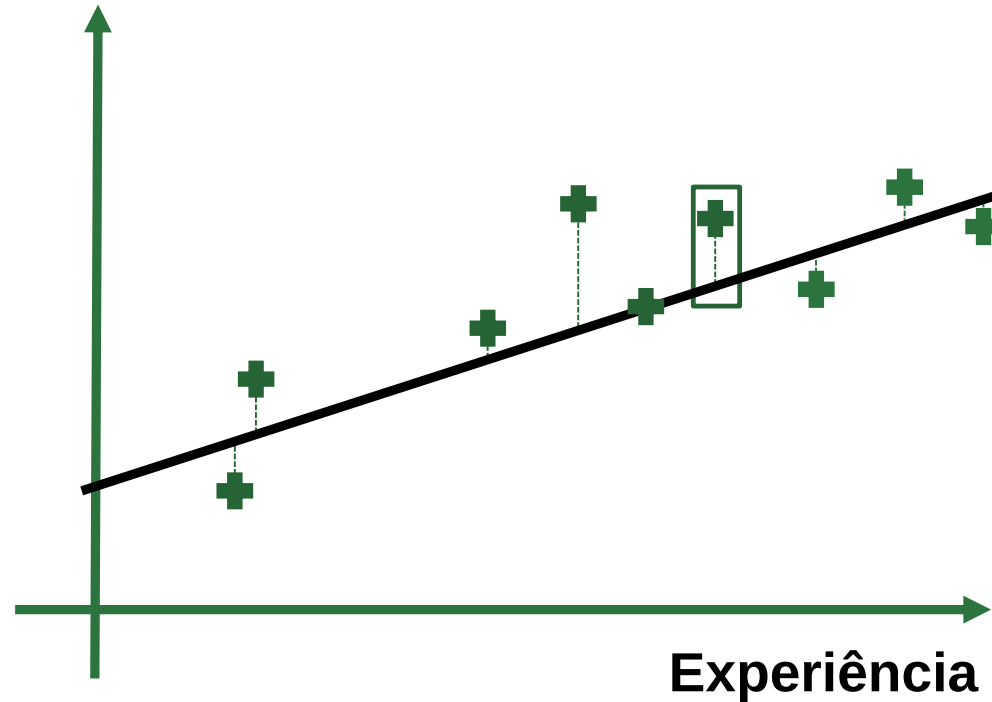




# Regressão Linear Simples

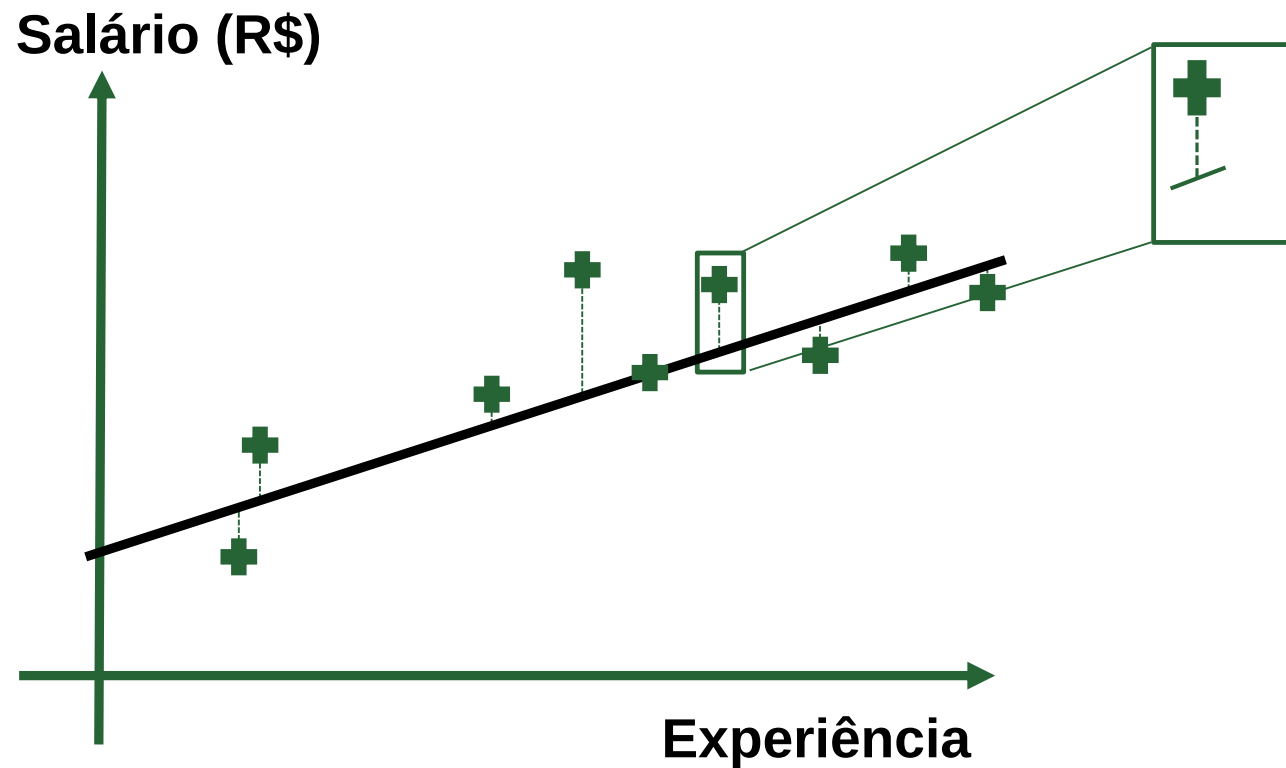
Regressão linear simples

Salário (R\$)



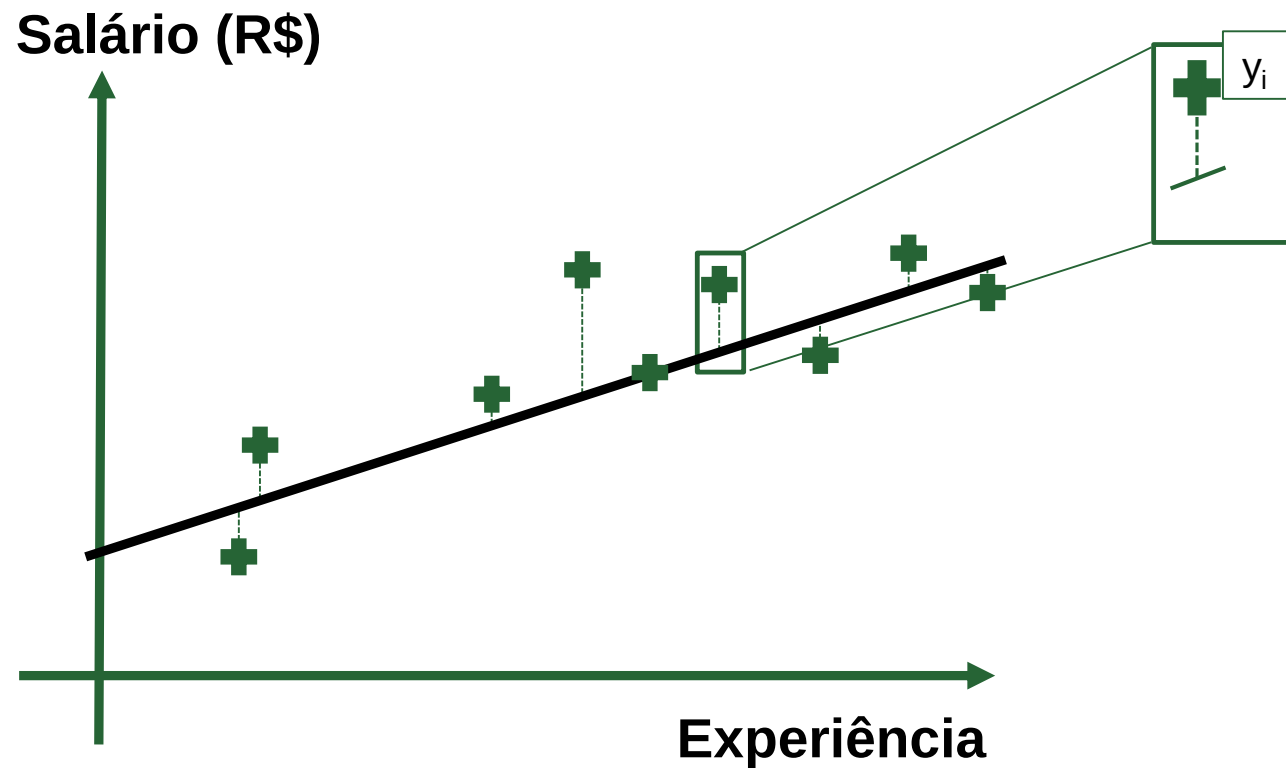
# Regressão Linear Simples

Regressão linear simples



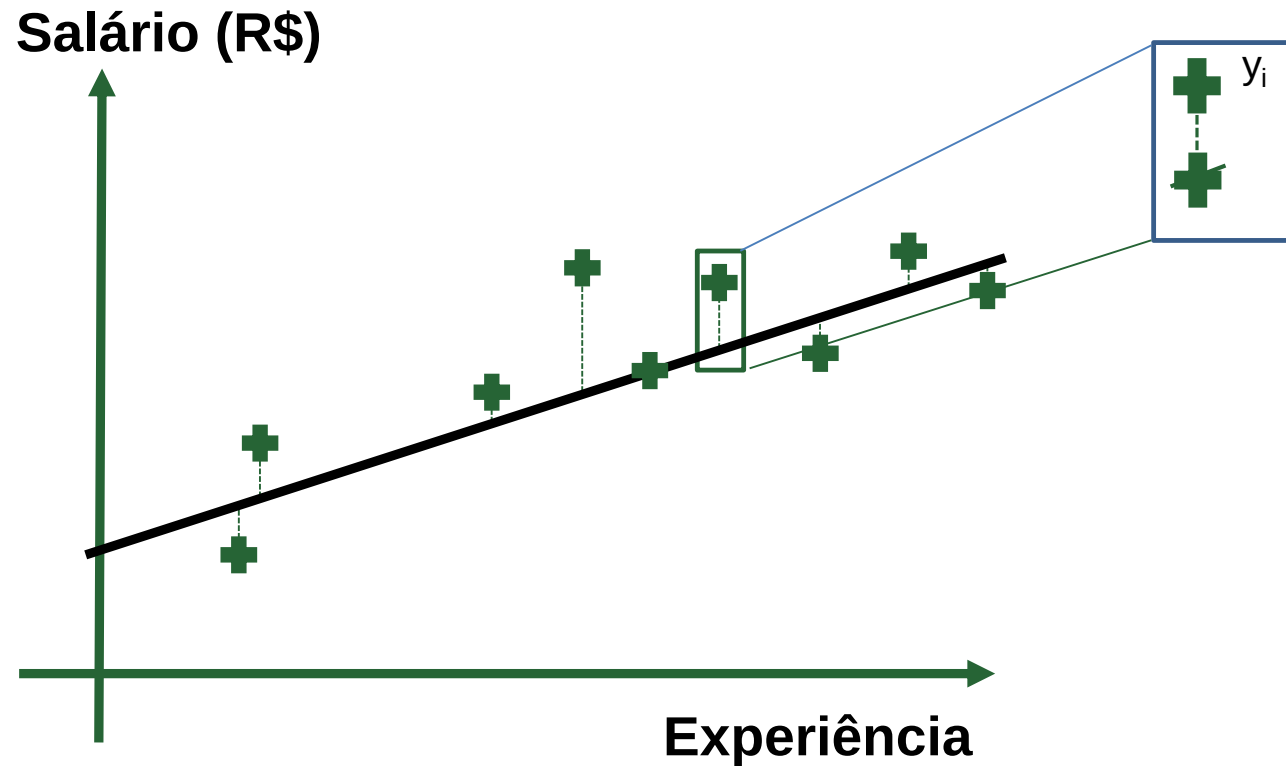
# Regressão Linear Simples

Regressão linear simples



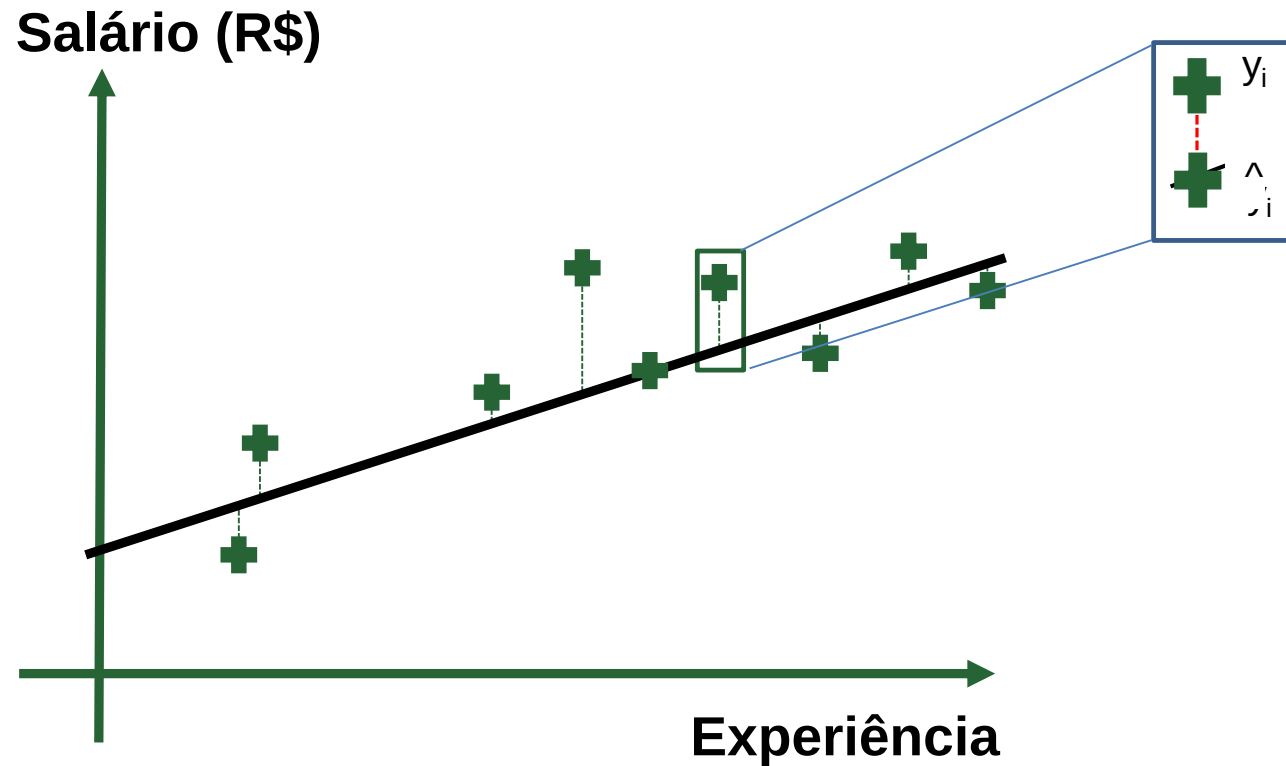
# Regressão Linear Simples

Regressão linear simples



# Regressão Linear Simples

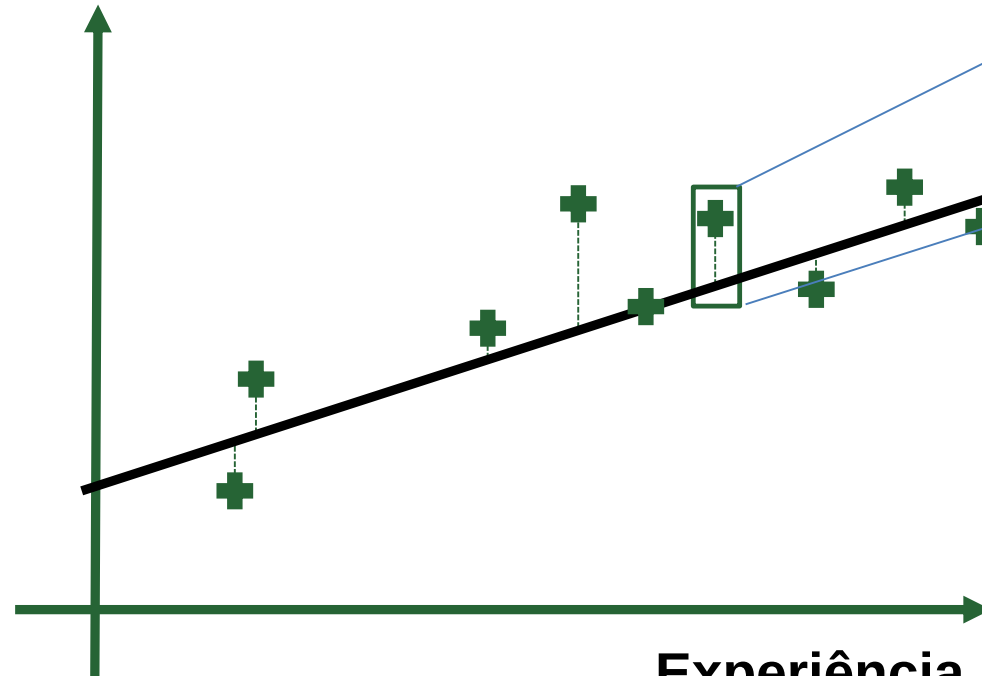
Regressão linear simples



# Regressão Linear Simples

## Regressão linear simples

Salário (R\$)

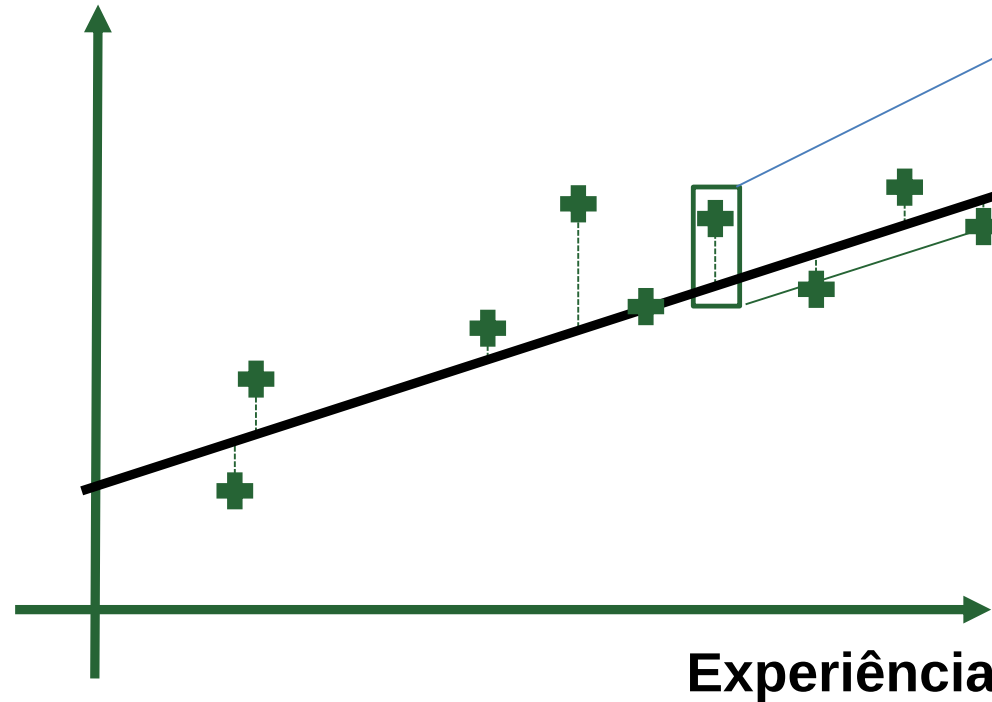


$$\text{Soma}(y - \hat{y})^2$$

# Regressão Linear Simples

## Regressão linear simples

Salário (R\$)

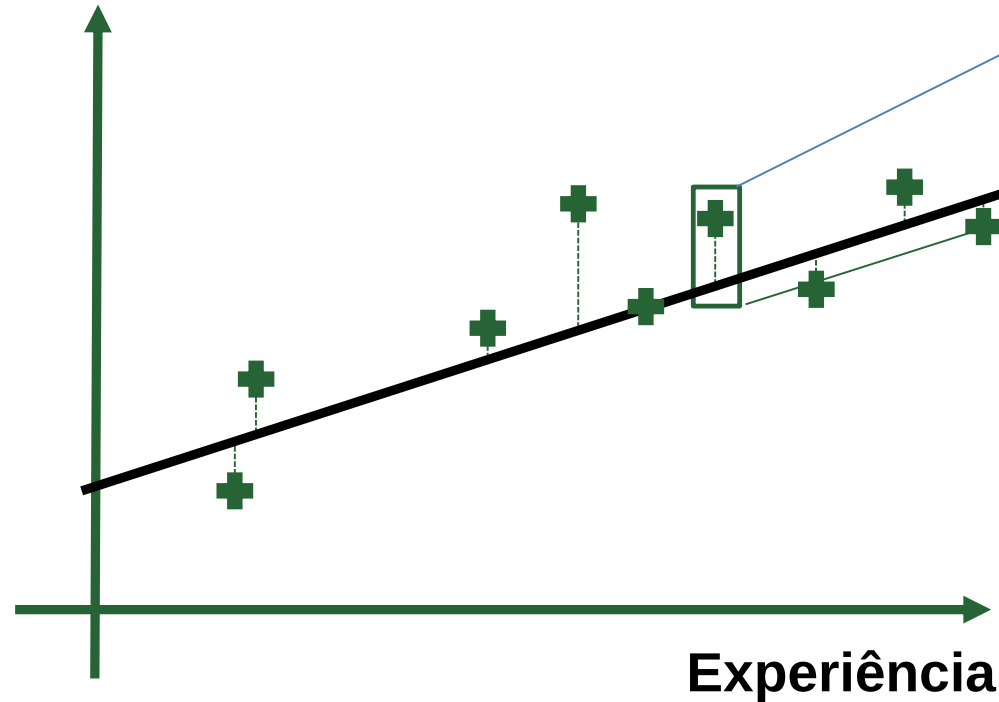


$\text{Soma}(y - \hat{y})^2 \rightarrow \text{Mínimo}$

# Regressão Linear Simples

## Regressão linear simples

Salário (R\$)

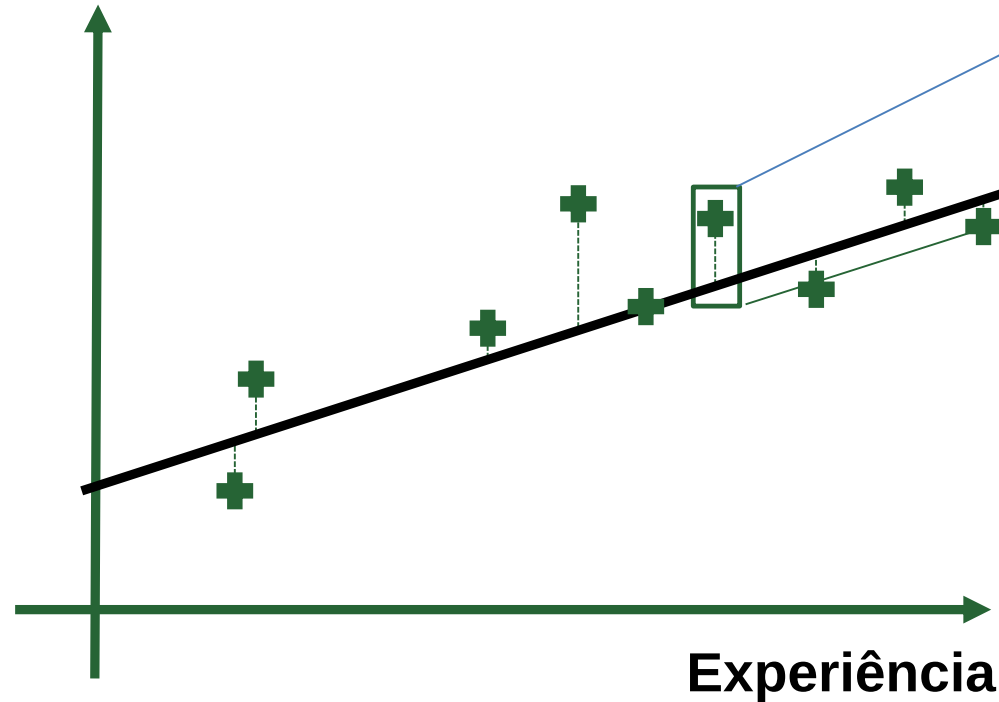




# Regressão Linear Simples

## Regressão linear simples

Salário (R\$)



# Regressão Linear Múltipla

- ✓ Existem certos cenários que uma única variável preditora não será capaz de explicar tudo a respeito da variável resposta.
- ✓ Por exemplo, a renda de uma determinada pessoa (variável resposta) é influenciada por diversas variáveis, tais como sexo, idade, escolaridade entre outras.
- ✓ Assim, precisamos realizar uma regressão linear múltipla.

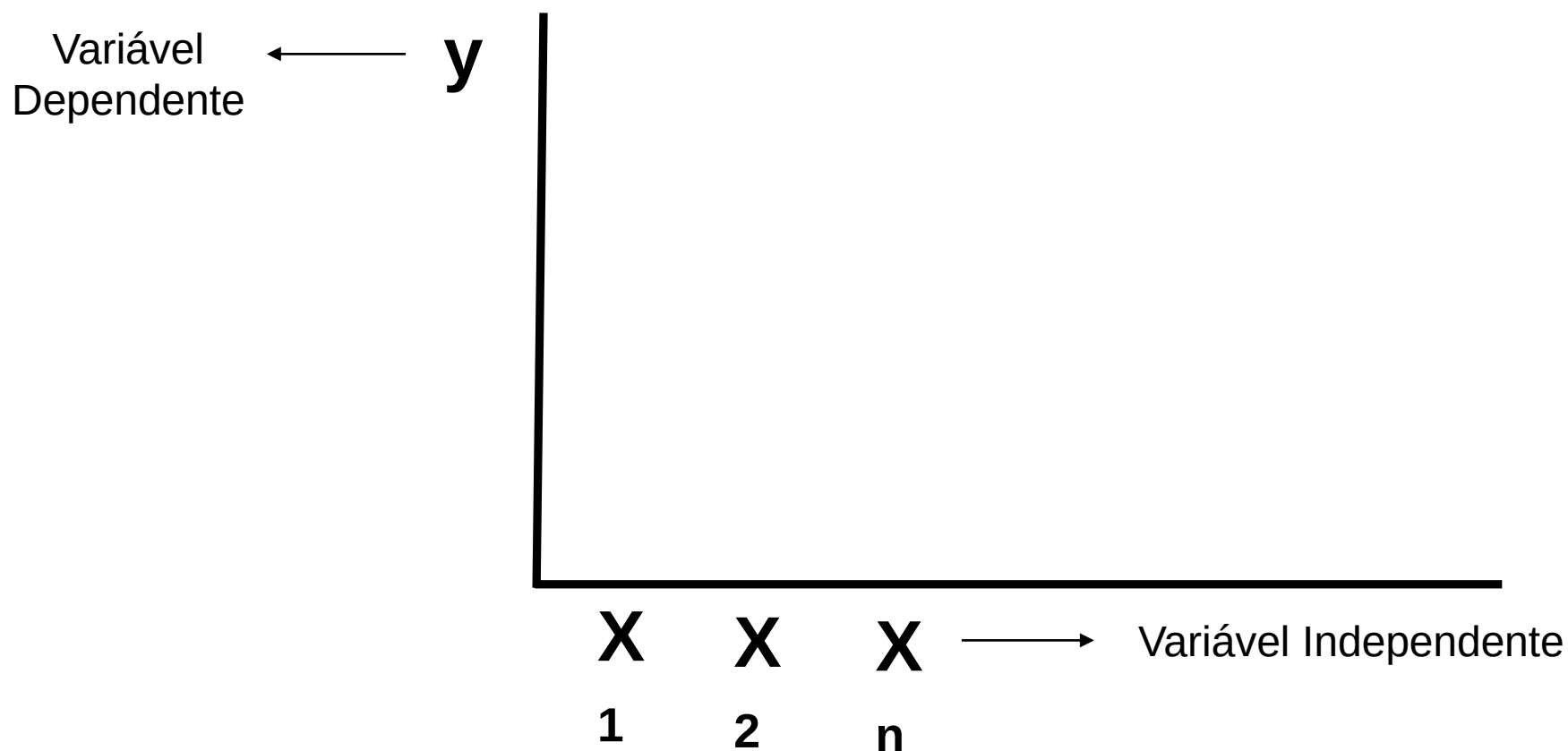
# Regressão Linear Múltipla

## Modelo de regressão linear múltipla

- ✓ A análise de regressão estuda a relação entre uma **variável dependente** e outras variáveis chamadas **variáveis independentes**.
- ✓ A resposta  $y$  é uma função de  $k$  variáveis previsoras  $x_1, x_2, \dots, x_k$
- ✓  $y = b_0 + b_1x_1 + b_2x_2 + \dots + b_kx_k + e$

# Regressão Linear Múltipla

## Modelo de regressão linear múltipla



# Regressão Linear Múltipla

## Modelo de regressão linear múltipla

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & x_{2n} & x_{2n} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \cdot \\ \cdot \\ \cdot \\ b_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \cdot \\ \cdot \\ \cdot \\ e_n \end{bmatrix}$$

# Regressão Linear Múltipla

## Modelo de regressão linear múltipla

$$\begin{aligned}y_1 &= b_0 + b_1x_{11} + b_2x_{21} + \dots + b_kx_{k1} + e_1 \\y_2 &= b_0 + b_1x_{12} + b_2x_{22} + \dots + b_kx_{k2} + e_2 \\&\vdots \\y_n &= b_0 + b_1x_{1n} + b_2x_{2n} + \dots + b_kx_{kn} + e_n\end{aligned}$$

# Regressão Linear Múltipla

**Descrição do problema: Baseado em investimentos feitos, verificar**

| R&D Spend | Administration | Marketing Spend | State      | Profit   |
|-----------|----------------|-----------------|------------|----------|
| 165349.2  | 136897.8       | 471784.1        | New York   | 192261.8 |
| 162597.7  | 151377.6       | 443898.5        | California | 191792.1 |
| 153441.5  | 101145.6       | 407934.5        | Florida    | 191050.4 |
| 144372.4  | 118671.9       | 383199.6        | New York   | 182902   |
| 142107.3  | 91391.77       | 366168.4        | Florida    | 166187.9 |
| 131876.9  | 99814.71       | 362861.4        | New York   | 156991.1 |
| 134615.5  | 147198.9       | 127716.8        | California | 156122.5 |
| 130298.1  | 145530.1       | 323876.7        | Florida    | 155752.6 |

# Regressão Linear Múltipla

**Descrição do problema:** Baseado em investimentos feitos, verificar

**Arquivo:** regressao\_linear\_multipla.py

**Dataset:** 50\_Startups.csv





**Prof. Dr. Marcio Andrey Teixeira**

**[marcio.andrey@ifsp.edu.br](mailto:marcio.andrey@ifsp.edu.br)**

**<http://marcioandrey.pro.br>**

## **BIBLIOGRAFIA BÁSICA:**

- [1] Machine Learning A-Z (udemy)
- [2] Python para Data Science e Machine Learning – COMPLETO (udemy)
- [3] <https://www.cetax.com.br/blog/data-science/>
- [4] <http://datascienceacademy.com.br/blog/10-carreiras-em-big-data-e-data-science/>