



FIAP



# MACHINE LEARNING & MODELLING

---

Esta disciplina aborda os principais conceitos sobre aprendizado de máquina e as técnicas clássicas de modelagem



# Na última aula...

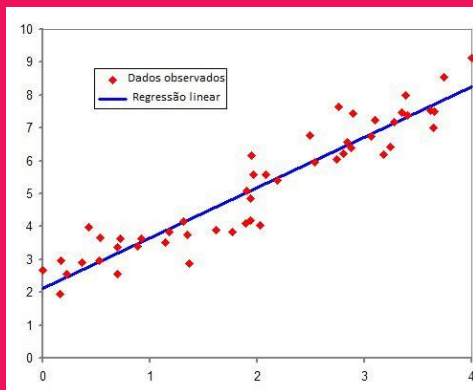
- O processo de Data Science
- Tipos de modelagem

# FORMAS DE MODELAGEM

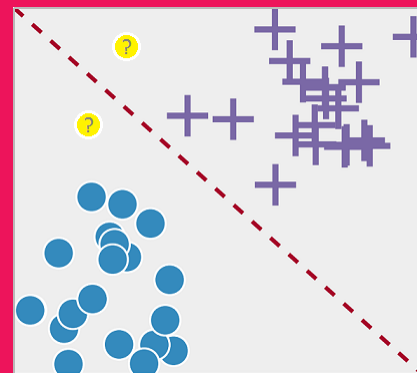
- Análises descritivas
- Análises preditivas
- Análises prescritivas

## APRENDIZADO SUPERVISIONADO

### Regressão

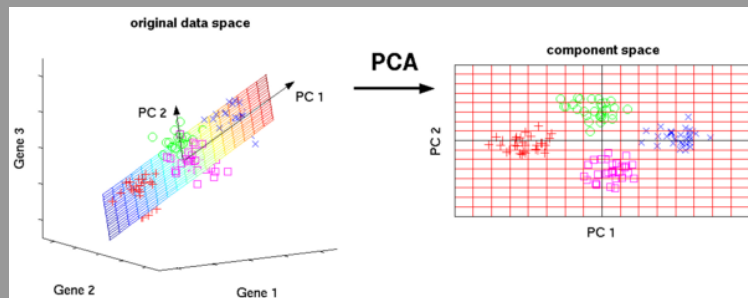


### Classificação

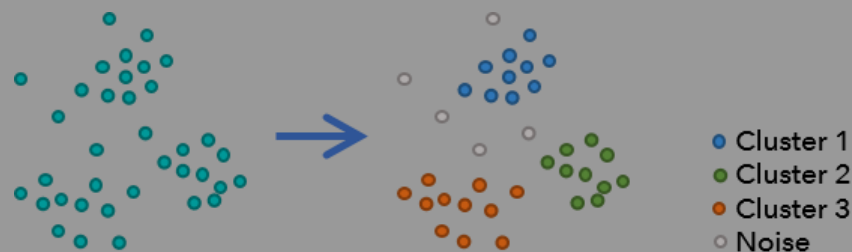


## APRENDIZADO NÃO SUPERVISIONADO

### PCA (Principal Component Analysis)



### Clusterização



# Exercício

Acesse o link abaixo e responda as questões sobre formas de modelagem de problemas : <https://forms.gle/dHcxAEx78aCMgnY38>

Vamos corrigir juntos na sequência!

# Agenda

- Forças & Fraquezas de ML
- Conceitos avançados sobre dados
- Desafios de trabalhar com dados
- O pós-desenvolvimento

# Forças & Fraquezas

Do aprendizado de máquina

# Reflexão

Há muitos casos de aplicação de IA com sucesso, mas quais são as limitações?





# Forças e Fraquezas de IA

*Exemplo: diagnósticos radiológicos*



## IA pode fazer:

Identificar sinais de  
Pneumonia a partir de  
~10000 imagens

## IA não pode fazer:

Identificar sinais de  
pneumonia a partir de  
10 imagens e suas  
explicações de um livro  
texto de radiologia

# Forças e Fraquezas de IA

- **ML (machine learning) funciona bem quando:**
  - Aprende um conceito simples
  - Com muitos dados disponíveis
- **ML tende a funcionar mal quando:**
  - Aprende conceitos complexos a partir de poucos dados
  - Encara tipos de dados diferentes daqueles “já acostumados”



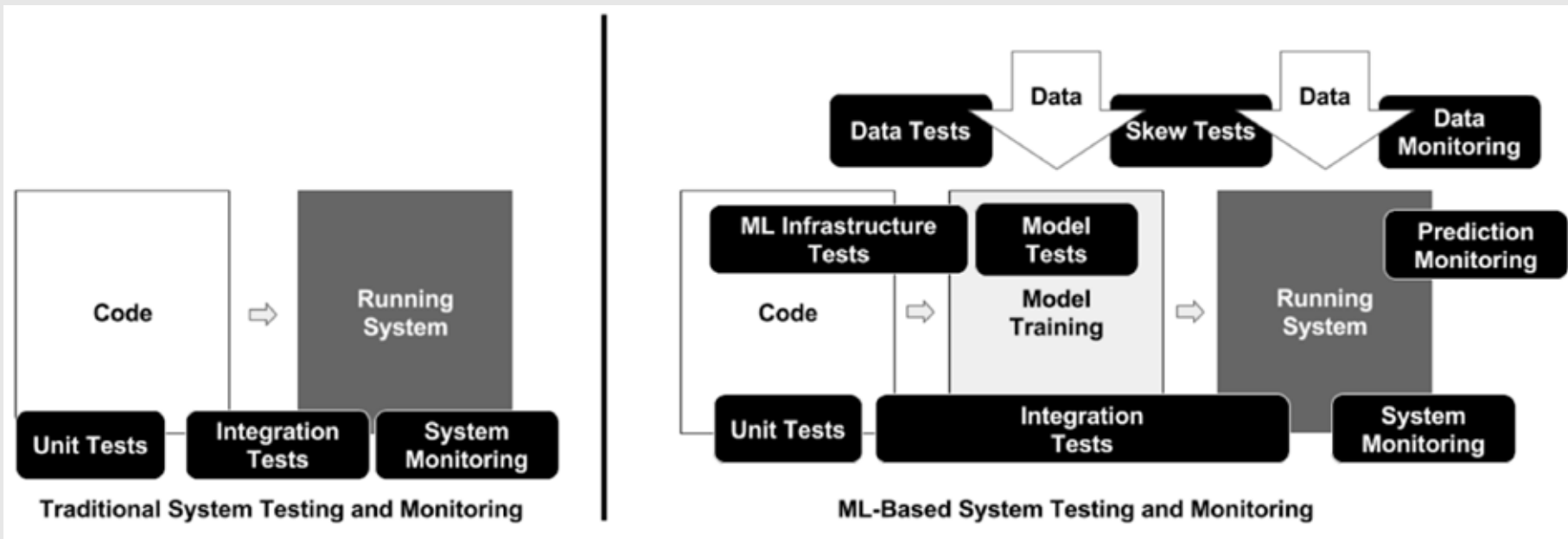
*Dados usuais, quando IA funciona*

*Dados diferentes dos usuais,  
quando IA não funciona bem*

# Pós-desenvolvimento

O que acontece com soluções de IA após o desenvolvimento?

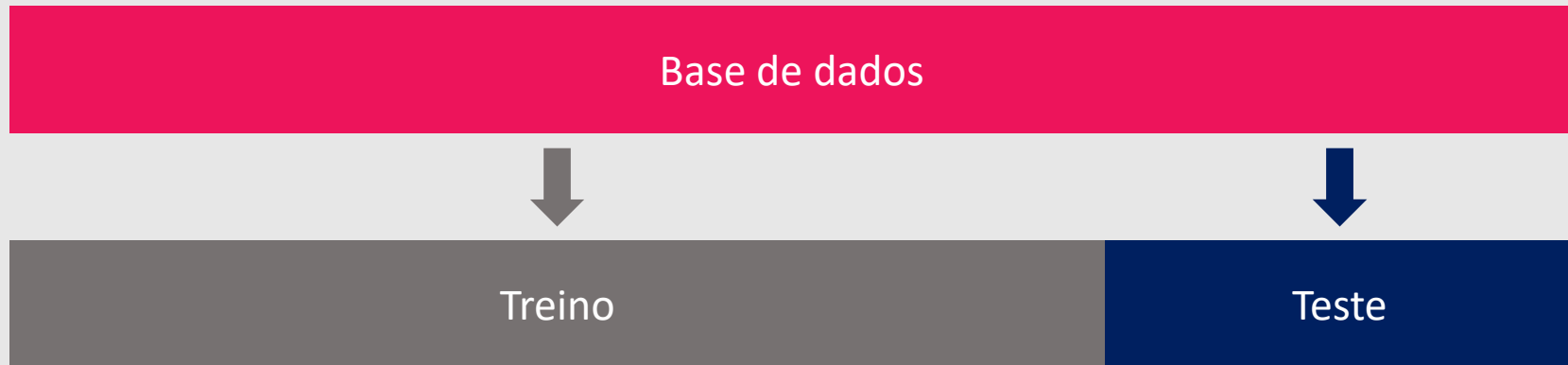
# Pós-implantação de soluções de IA



# Trabalhando com dados

Como usar nossa base para criar modelos?

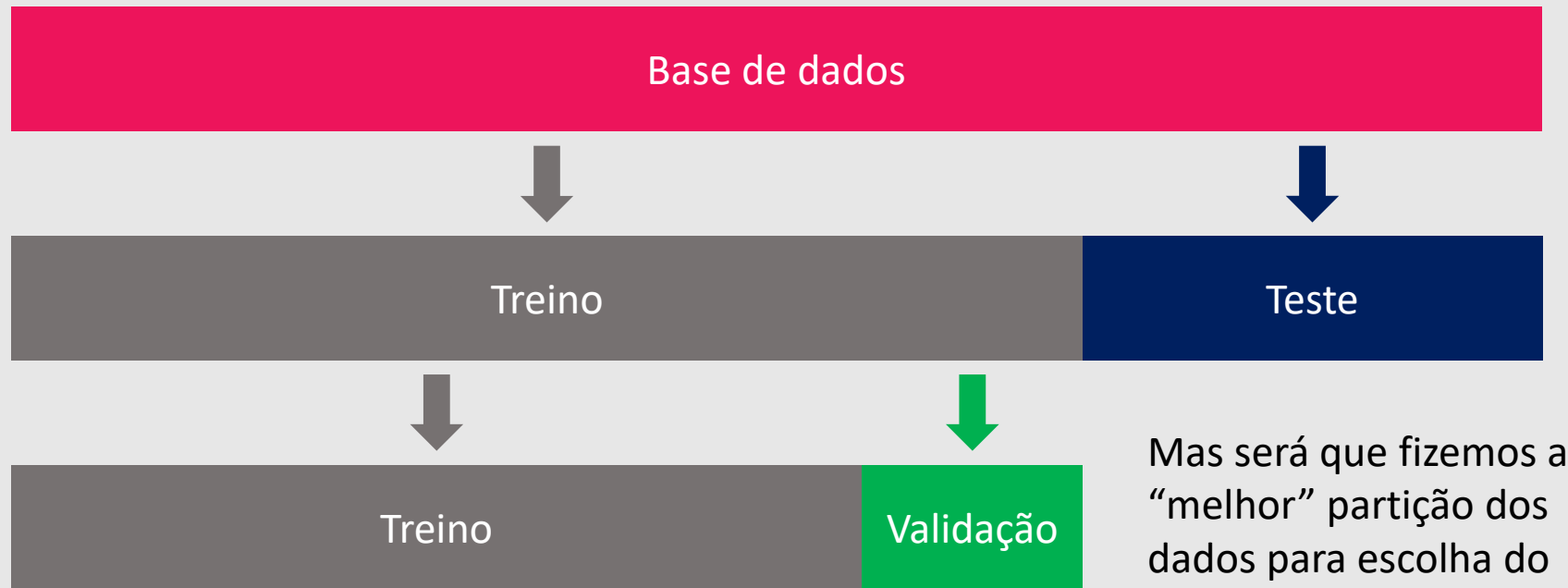
# Uso da base de dados



É comum usarmos 80% dos dados para treinamento e **separar** 20% para teste.

Agora, e se precisamos escolher qual é o melhor entre 2 ou mais modelos?

# Uso da base de dados



Mas será que fizemos a “melhor” partição dos dados para escolha do modelo?

# Desafios

Os maiores desafios de se trabalhar com dados

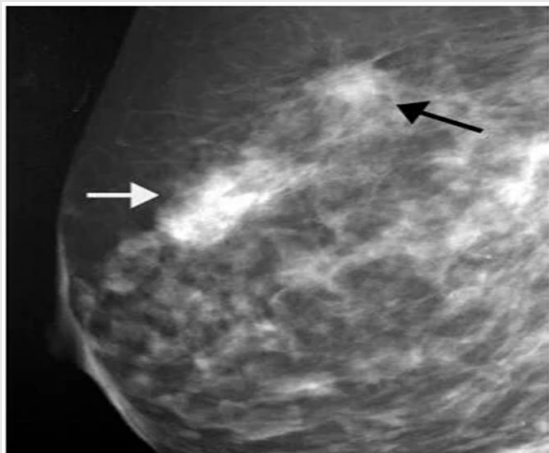


# Desafios de trabalhar com dados

diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave_points_mean	symmetry_mean	fractal_dimension_mean
B	13,54	14,36	87,46	566,3	0,09779	0,08129	0,06664	0,04781	0,1885	0,05766
B	13,08	15,71	85,63	520	0,1075	127	0,04568	0,0311	0,1967	0,06811
B	8196	16,84	51,71	201,9	86	0,05943	0,01588	0,005917	0,1769	0,06503
B	6981	13,43	43,79	143,5	117	0,07568	0	0	193	0,07818
B	12,18	20,52	77,22	458,7	0,08013	0,04038	0,02383	0,0177	0,1739	0,05677
M	25,22	24,91	171,5	1878	0,1063	0,2665	0,3339	0,1845	0,1829	0,06782
M	19,1	26,29	129,1	1132	0,1215	0,1791	0,1937	0,1469	0,1634	0,07224
M	18,46	18,52	121,1	1075	0,09874	0,1053	0,1335	0,08795	0,2132	0,06022
M	14,48	21,46	94,25	648,2	0,09444	0,09947	0,1204	0,04938	0,2075	0,05636
M	19,02	24,59	122	1076	0,09029	0,1206	0,1468	0,08271	0,1953	0,05629

# Desafios de trabalhar com dados

## 1. Entendimento do problema e dos dados



Os dados são calculados a partir de uma imagem digitalizada da massa mamária

Ten real-valued features are computed for each cell nucleus:

- a) **radius** (mean of distances from center to points on the perimeter)
- b) **texture** (standard deviation of gray-scale values)
- c) perimeter
- d) area
- e) **smoothness** (local variation in radius lengths)
- f) **compactness** ( $\text{perimeter}^2 / \text{area} - 1.0$ )
- g) **concavity** (severity of concave portions of the contour)
- h) **concave points** (number of concave portions of the contour)
- i) symmetry
- j) fractal dimension ("coastline approximation" - 1)

Base de dados adaptada do kaggle:

<https://www.kaggle.com/uciml/breast-cancer-wisconsin-data>

Imagem retirada de:

<https://www.canceroz.com/2017/02/breast-ultrasound-cancer-vs-benign.html>

# Desafios de trabalhar com dados

## 1. Entendimento do problema e dos dados

## 2. Conhecer a distribuição dos dados

- Como está a distribuição?
- Dados desbalanceados?
- Dados faltantes?

diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave_points_mean	symmetry_mean	fractal_dimension_mean
B	13,54	14,36	87,46	566,3	0,09779	0,08129	0,06664	0,04781	0,1885	0,05766
B	13,08	15,71	85,63	520	0,1075	127	0,04568	0,0311	0,1967	0,06811
B	8196	16,84	51,71	201,9	86	0,05943	0,01588	0,005917	0,1769	0,06503
B	6981	13,43	43,79	143,5	117	0,07568	0	0	193	0,07818
B	12,18	20,52	77,22	458,7	0,08013	0,04038	0,02383	0,0177	0,1739	0,05677
M	25,22	24,91	171,5	1878	0,1063	0,2665	0,3339	0,1845	0,1829	0,06782
M	19,1	26,29	129,1	1132	0,1215	0,1791	0,1937	0,1469	0,1634	0,07224
M	18,46	18,52	121,1	1075	0,09874	0,1053	0,1335	0,08795	0,2132	0,06022
M	14,48	21,46	94,25	648,2	0,09444	0,09947	0,1204	0,04938	0,2075	0,05636
M	19,02	24,59	122	1076	0,09029	0,1206	0,1468	0,08271	0,1953	0,05629

# Desafios de trabalhar com dados

1. Entendimento do problema e dos dados

2. Conhecer a distribuição dos dados

3. Preparar os dados para uso

diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave_points_mean	symmetry_mean	fractal_dimension_mean
B	13,54	14,36	87,46	566,3	0,09779	0,08129	0,06664	0,04781	0,1885	0,05766
B	13,08	15,71	85,63	520	0,1075	127	0,04568	0,0311	0,1967	0,06811
B	8196	16,84	51,71	201,9	86	0,05943	0,01588	0,005917	0,1769	0,06503
B	6981	13,43	43,79	143,5	117	0,07568	0	0	193	0,07818
M	19,02	24,59	122	1076	0,09029	0,1206	0,1468	0,08271	0,1953	0,05629

- Todos os atributos são necessários? → [ **compactness** ( $\text{perimeter}^2 / \text{area} - 1.0$ ) ]
- O que fazer com os dados faltantes? → zeros
- Outlier? → como identificar? Como proceder?

# Desafios de trabalhar com dados

**Quais outros desafios vocês vislumbram ?**

# Desafios de trabalhar com dados

- Dados insuficientes para treinar os modelos
- Dados não representativos sobre o que se quer modelar
- Dados com baixa qualidade
- *Features* irrelevantes
- *Overfitting* do modelo
- *Underfitting* do modelo



**Dados ruins**

**Algoritmos ruins**

# Features irrelevantes

Uma parte crítica do sucesso de um projeto de aprendizado de máquina é a criação de um bom conjunto de *features* para o treinamento do modelo. Isso é chamado de **feature engineering**:

- **Feature Selection**: o processo de seleção de *features* mais úteis para o modelo dentre as *features* existentes.
- **Feature Extraction**: combinar *features* existentes para produzir outra mais útil.

# Overfitting

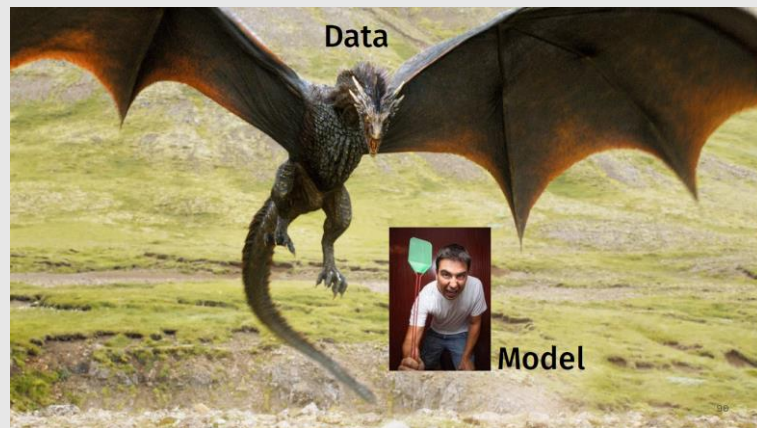
O **modelo é complexo demais** para diferenciar as perturbações dos dados (sujeiras)



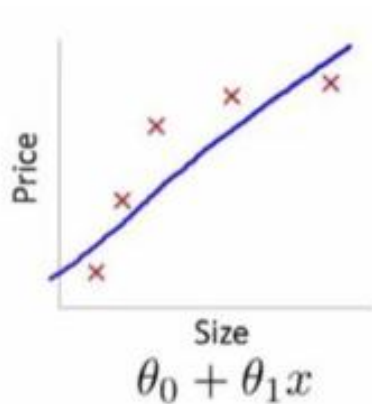


# Underfitting

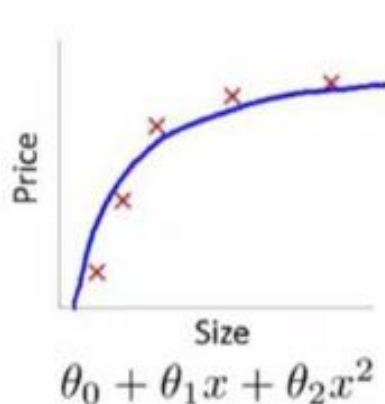
Oposto ao overfitting : o **modelo é simples demais** para aprender a abstrair os conceitos presentes nos dados



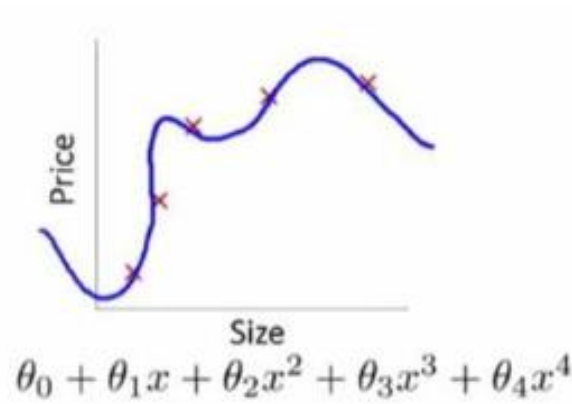
# Overfitting X Underfitting



Viés alto  
(subajuste)



Ajuste de boa  
qualidade



Variância alta  
(superajuste)



<https://www.kaggle.com/mlg-ulb/creditcardfraud>

# Unindo teoria & prática

Beijos, Prof.<sup>a</sup> Carol



# O processo de ciência de dados



# OBRIGADO!



**Prof. Michel Fornaciali**

<https://www.linkedin.com/in/michelfornaciali/>  
[profmichel.fornaciali@fiap.com.br](mailto:profmichel.fornaciali@fiap.com.br)

Copyright © 2021 | Professor Michel Fornaciali

Todos os direitos reservados. Reprodução ou divulgação total ou parcial deste documento, é expressamente proibido sem consentimento formal, por escrito, do professor/autor.