Predizendo a sobrevivência no Titanic com Machine Learning e Python

campus party

Carla Vieira
@carlaprvieira

Quem sou eu?

Graduanda de Sistemas de Informação - USP

Apaixonada por Data Science e BI

@carlaprvieira

@carlaprv

WOMAKERS CODE

Garotas CPBr
GAROTAS QUE SACAM O QUE FAZEM

_t

Agenda

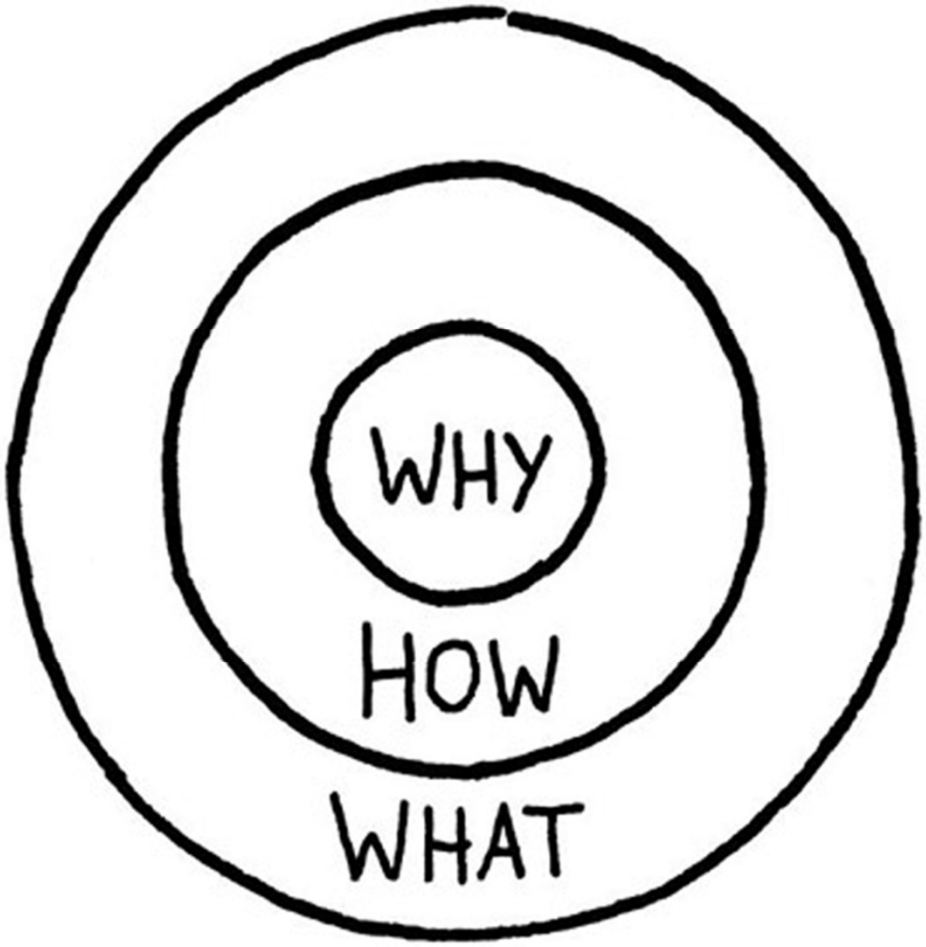
O que é Mineração de Dados?

Processo de mineração de dados

Algoritmos de Machine Learning

DESAFIO PRÁTICO!

Antes de começarmos...



Why = The Purpose
What is your cause? What do you believe?

Apple: We believe in challienging the status quo and doing this differently

How = The Process
Specific actions taken to realize the Why.

Apple: Our products are beautifully designed and easy to use

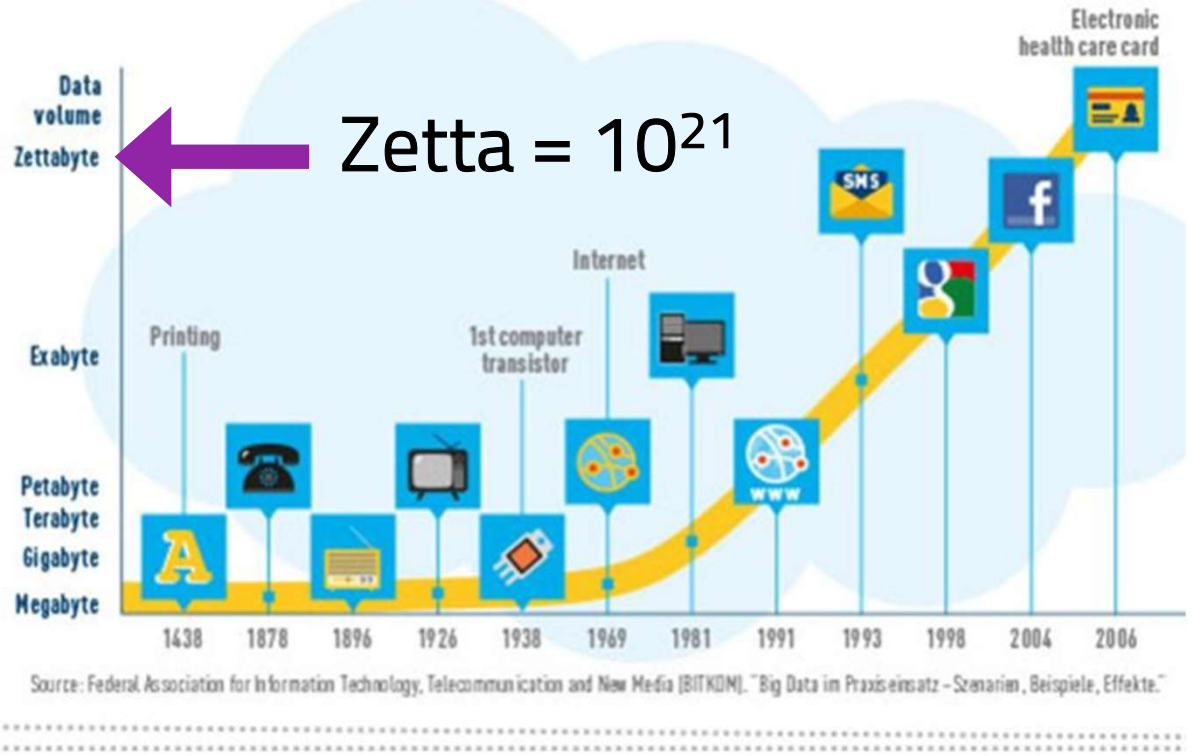
What = The Result
What do you do? The result of Why. Proof.

Apple: We make computers

Big Data e Mineração de Dados

Crescimento exponencial do volume de dados

Tecnologias como celulares e o uso constante de aplicativos de redes sociais resultaram em um rápido crescimento no volume de dados

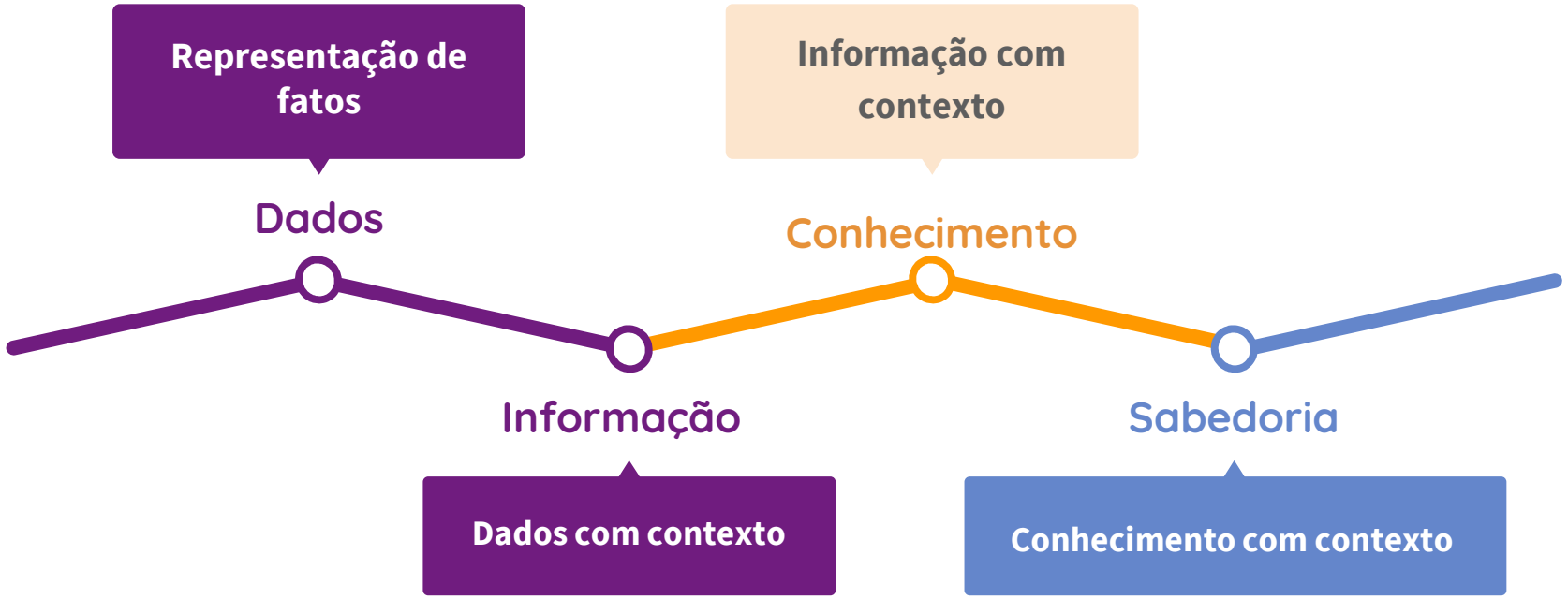


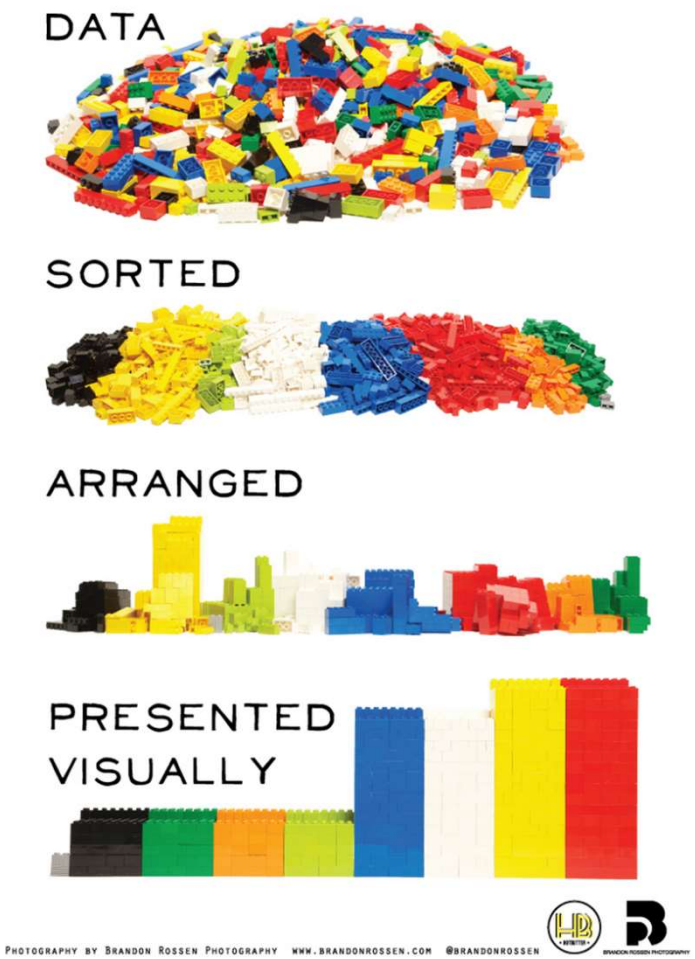
O que é Mineração de Dados?

Data Mining define o processo automatizado de captura e análise de grandes conjuntos de dados para extrair um significado, sendo usado tanto para ***descrever características do passado*** *como para* ***predizer tendências para o futuro***.

Cadeia de Valor do Conhecimento

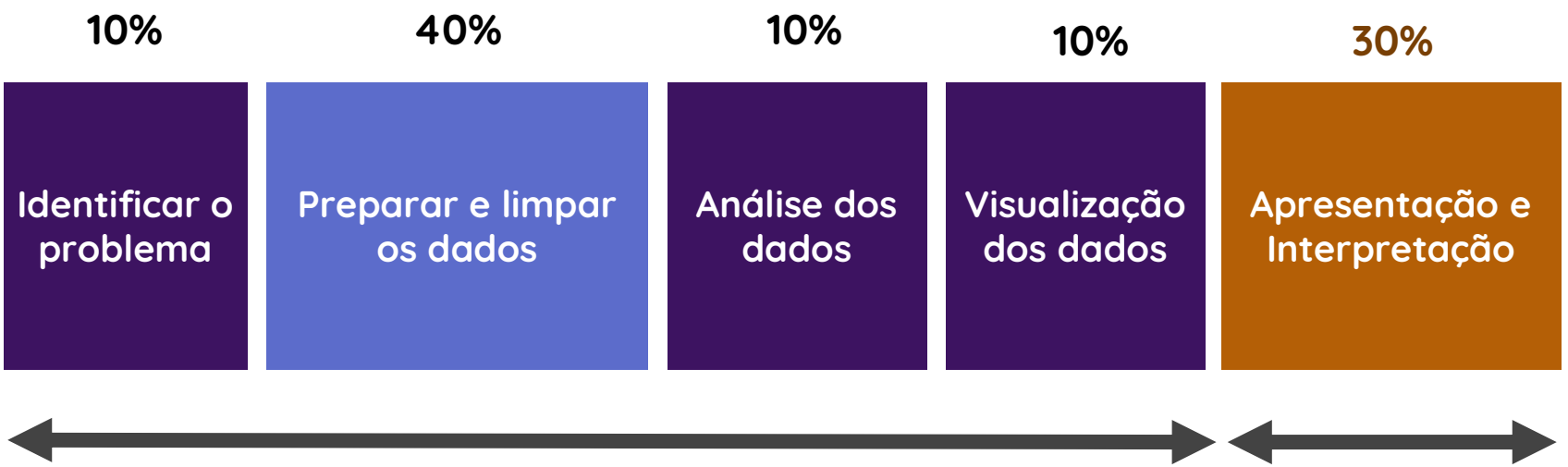
(Davenport e Prusak - 1998)





“Um **dado** não vira informação se você não souber o que ele significa; uma **informação** não vira **conhecimento** se você não enxergar **relevância** nela e conhecimento não serve pra nada se não aplicá-lo de maneira apropriada.”

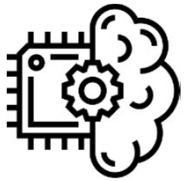
Ciclo de vida de um projeto de Data Science



Técnicas de mineração de dados

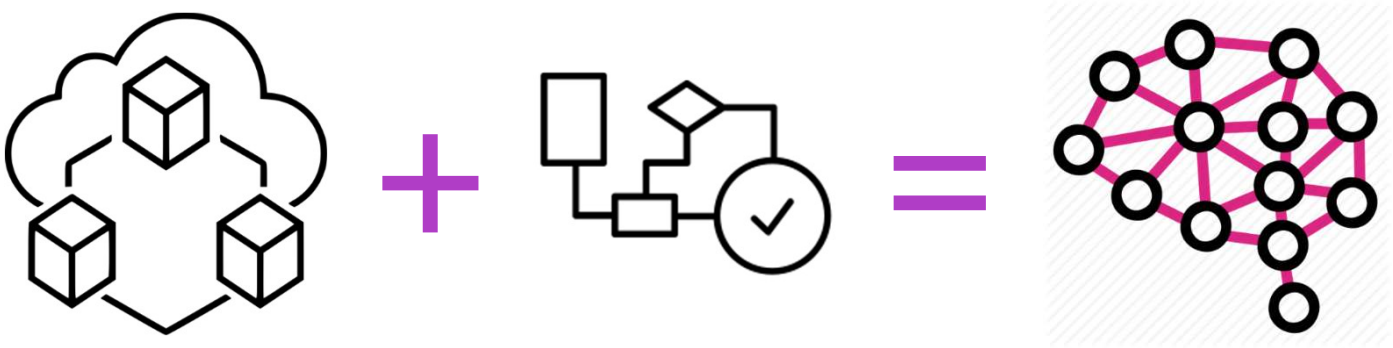
Estatística

Inteligência Artificial



Machine Learning

O que é Aprendizado de Máquina?



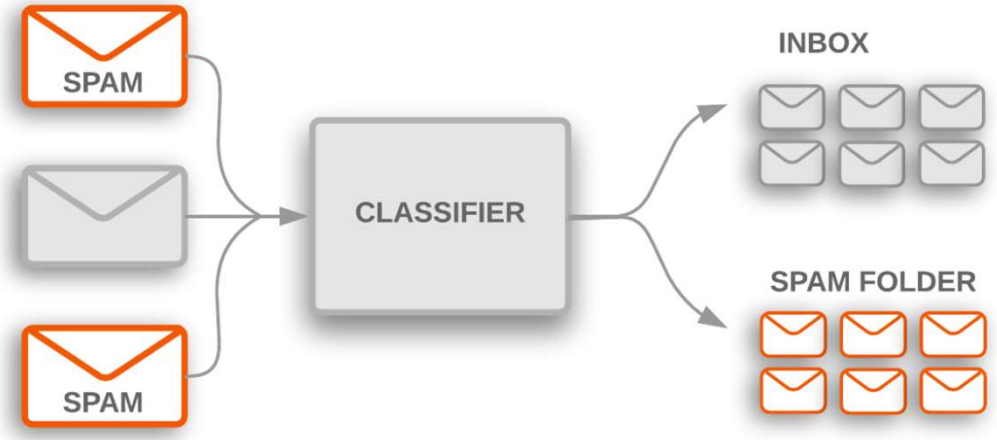
Big Data + Algoritmos = Aprendizado

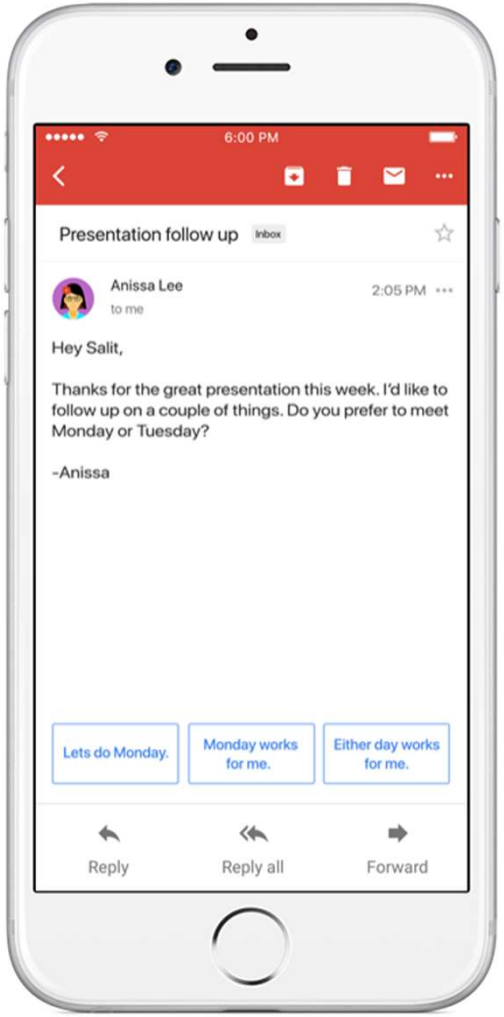
“Um programa de computador aprende se ele é capaz de melhorar seu **desempenho** em determinada **tarefa**, sob alguma medida de **avaliação**, a partir de **experiências** passadas.”

(Tom Mitchell)

Google

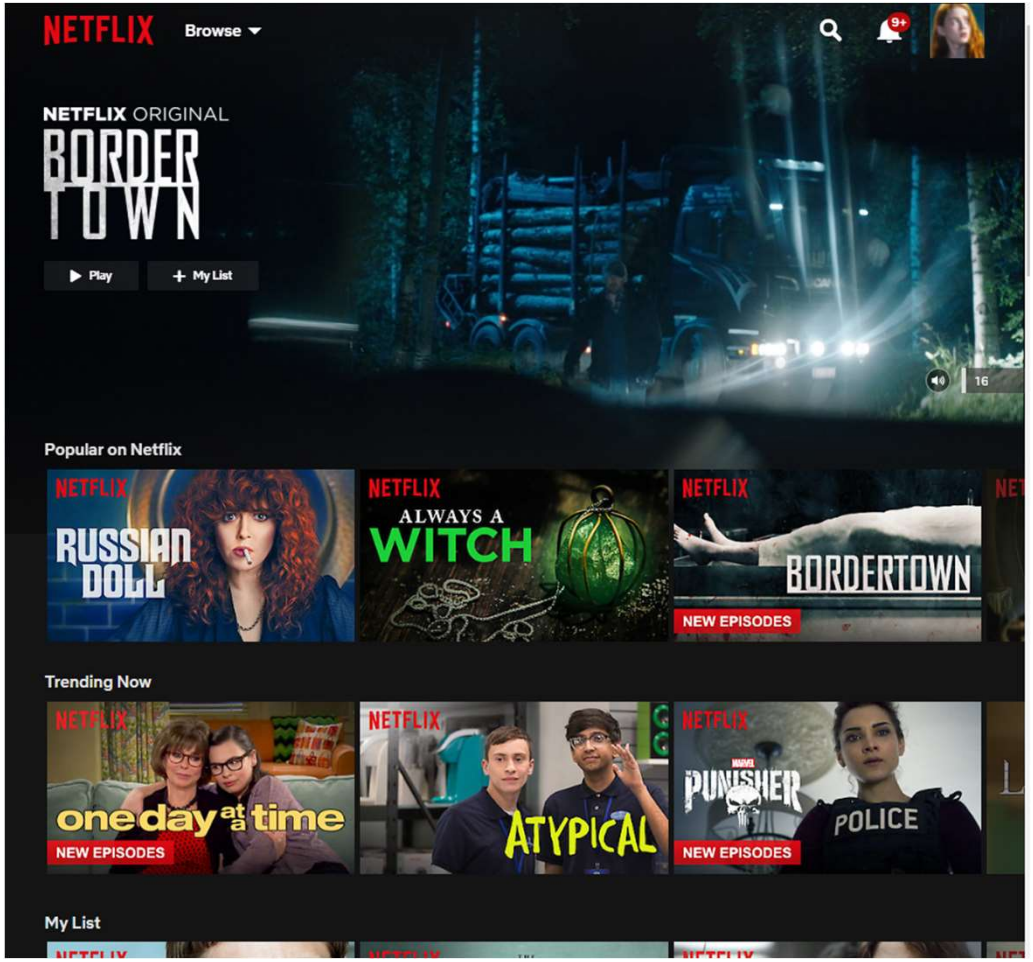
Smart reply e classificação de SPAM





Netflix

Sistema de recomendação



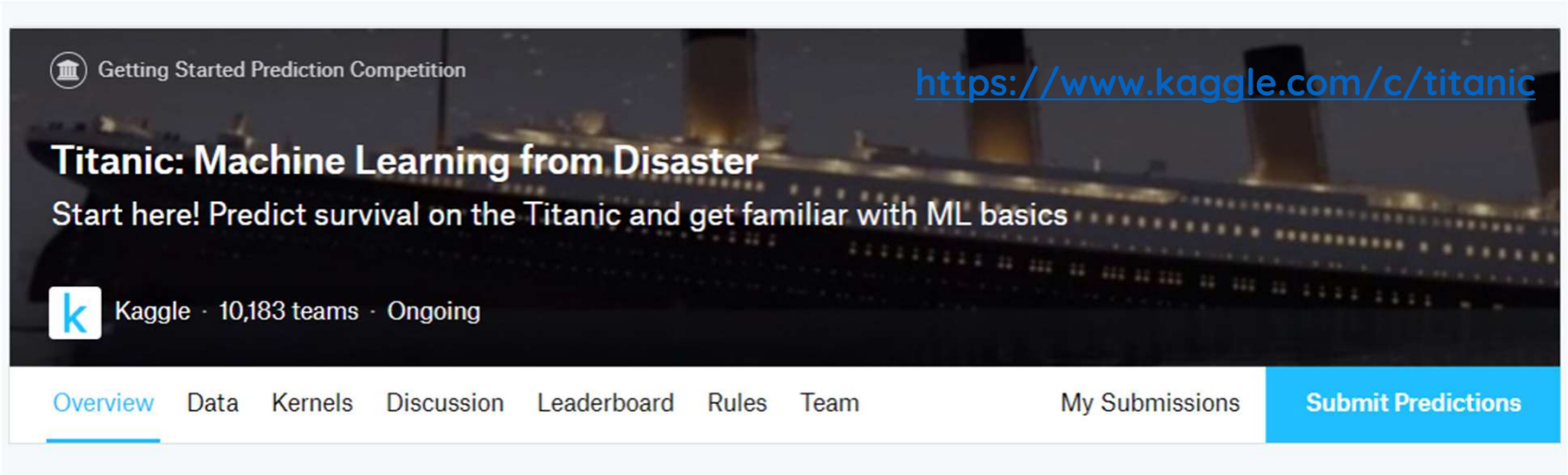
https://www.youtube.com/watch?v=AYv0ujDc_LY

Abordagens de Aprendizado de Máquina


jupyter
DESAFIO!

Explorando a sobrevivência no Titanic...

Desafio



Predizer a sobrevivência dos passageiros no Titanic...

Dados

kaggle

Data Description

Overview

The data has been split into two groups:

- training set (train.csv)
- test set (test.csv)

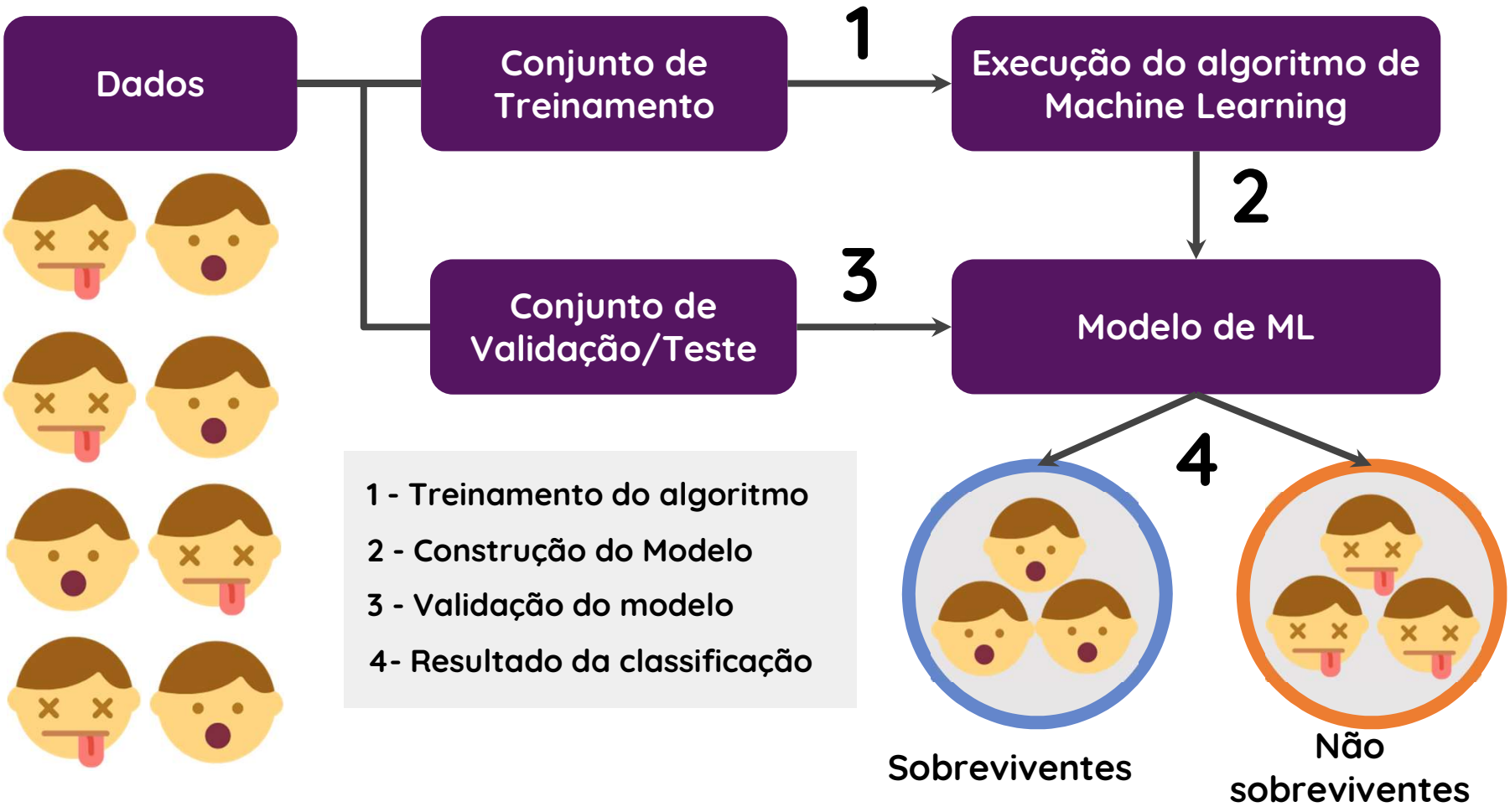
Conjunto de dados para treinamento

Conjunto de dados de teste

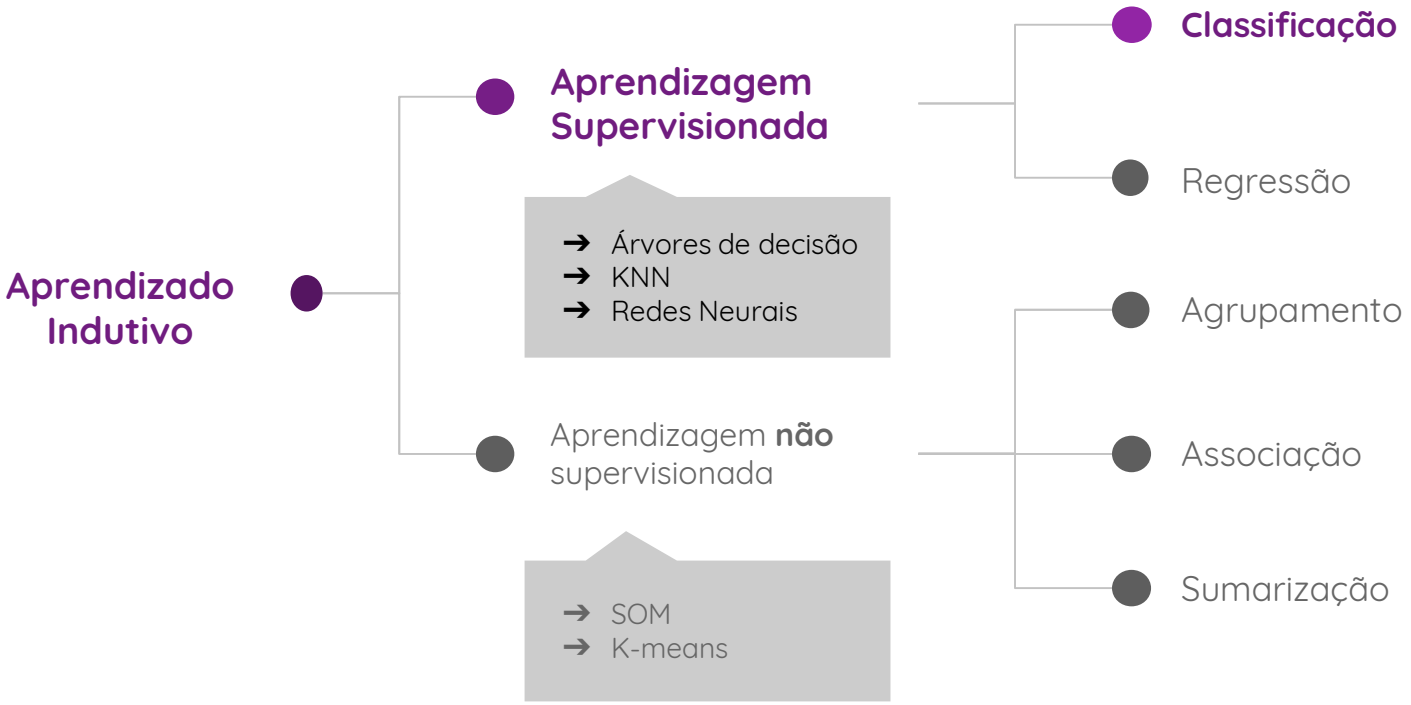
The training set should be used to build your machine learning models. For the training set, we provide the outcome (also known as the "ground truth") for each passenger. Your model will be based on "features" like passengers' gender and class. You can also use feature engineering to create new features.

The test set should be used to see how well your model performs on unseen data. For the test set, we do not provide the ground truth for each passenger. It is your job to predict these outcomes. For each passenger in the test set, use the model you trained to predict whether or not they survived the sinking of the Titanic.

https://www.kaggle.com/c/titanic


Dados
Conjunto de Treinamento
1
Execução do algoritmo de Machine Learning
2
Conjunto de Validação/Teste
3
Modelo de ML
4
1 - Treinamento do algoritmo
2 - Construção do Modelo
3 - Validação do modelo
4- Resultado da classificação
Sobreviventes
Não sobreviventes

Abordagens de Aprendizado de Máquina



Iniciando nosso 1º Jupyter Notebook

Download - Jupyter Notebook

bit.ly/cpbr12-baixar-jupyter

bit.ly/cpbr12-baixar-python

Iniciando nosso Jupyter Notebook

1. Abrir o menu do windows
2. Digite **`jupyter notebook`** e aperte **ENTER**
3. Aguarde alguns minutos até abrir uma nova janela do Jupyter no seu navegador

1. Utilizar o jupyter.org/try pelo navegador

Criação de conta e download dos datasets

www.kaggle.com/c/titanic

ou

bit.ly/cpbr12-datasets

Código completo do workshop

bit.ly/cpbr12-carla-workshop

Sobre a linguagem - Python 3.7

```
1 #comentario
2 a = 10
3 b =  5
4 c = a * b
5 print(c)
6 print("Ola mundo!")
```

http://www.devfuria.com.br/python/sintaxe-basica/

Bibliotecas de Python utilizadas

Numpy - http://www.numpy.org/

Regex - https://docs.python.org/3/library/re.html#module-re

Matplotlib - https://matplotlib.org/

Pandas - https://pandas.pydata.org/

Scikit Learn - https://scikit-learn.org/stable/

Conhecendo os dados - Variáveis

- **PassengerId**
- **Pclass** - classe da viagem
- **Name** - nome do passageiro
- **Sex** - genêro
- **Age** **-** idade
- **SibSp** - Número de irmãos / cônjuge a bordo
- **Parch** - Número de pais/filho abordo
- **Ticket**
- **Fare -** tarifa paga
- **Cabin -** cabine
- **Embarked** - a porta em que o passageiro embarcou
- **SURVIVED - CONJUNTO DE TREINAMENTO**

Variáveis

- **Pclass e Sexo do Passageiro -** sem dados nulos
- **Tamanho da Família (SibSP e Parch)** e **pessoas viajando sozinhas**
- **Embarked -** contém valores nulos
- **Fare** - contém valores nulos
- **Age** - contém valores nulos
- **Name -** podemos extrair o título da pessoa

Limpeza dos dados

- **Existem variáveis com valores nulos no nosso dataset?**
- **Para cada variável, será necessário tratar os valores nulos de maneira correta**
  - substituir pela média
  - sortear valores entre [média - erro , média + erro]
  - etc…
- **Transformar todos os dados em valores numérico**

Limpeza dos dados

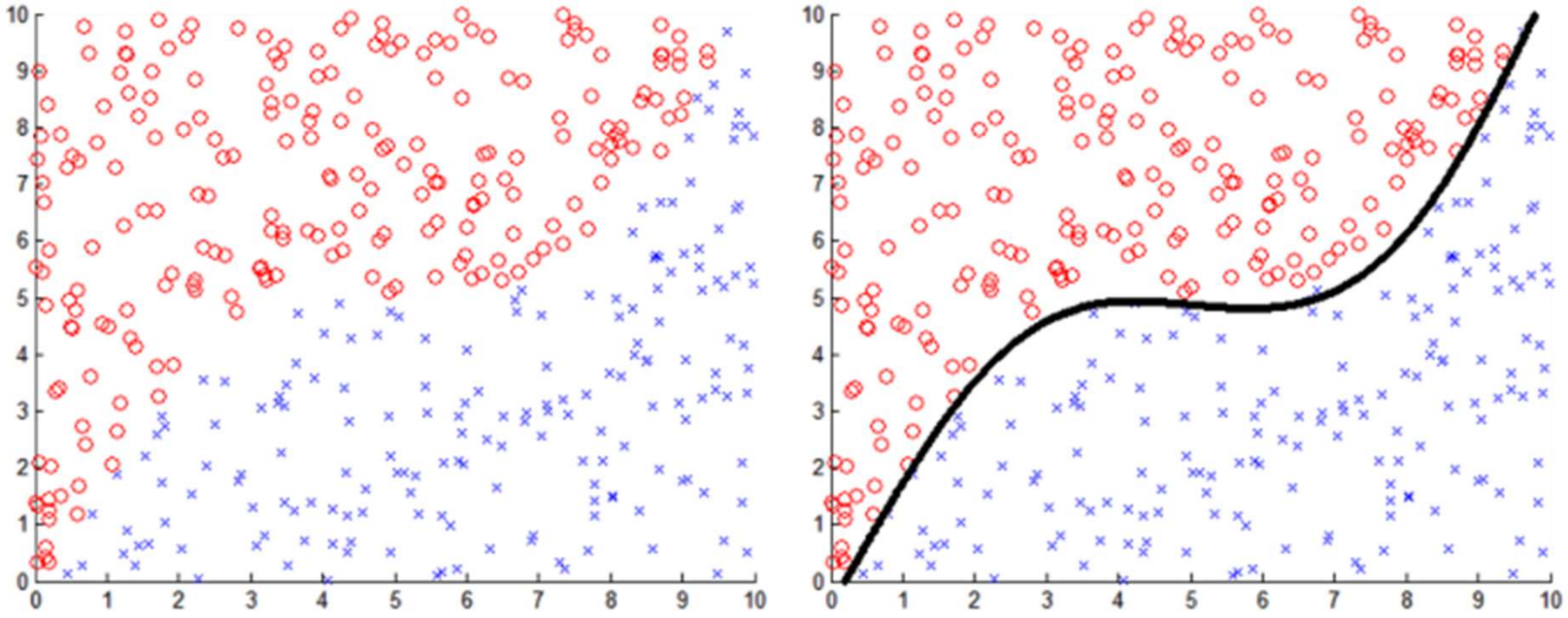
- **Transformar todos os dados em valores numéricos**
  - Sexo
  - Pronomes de tratamento (títulos)
  - Embarked
  - Fare
  - Age
- **Outras variáveis:** Pclass, Survived

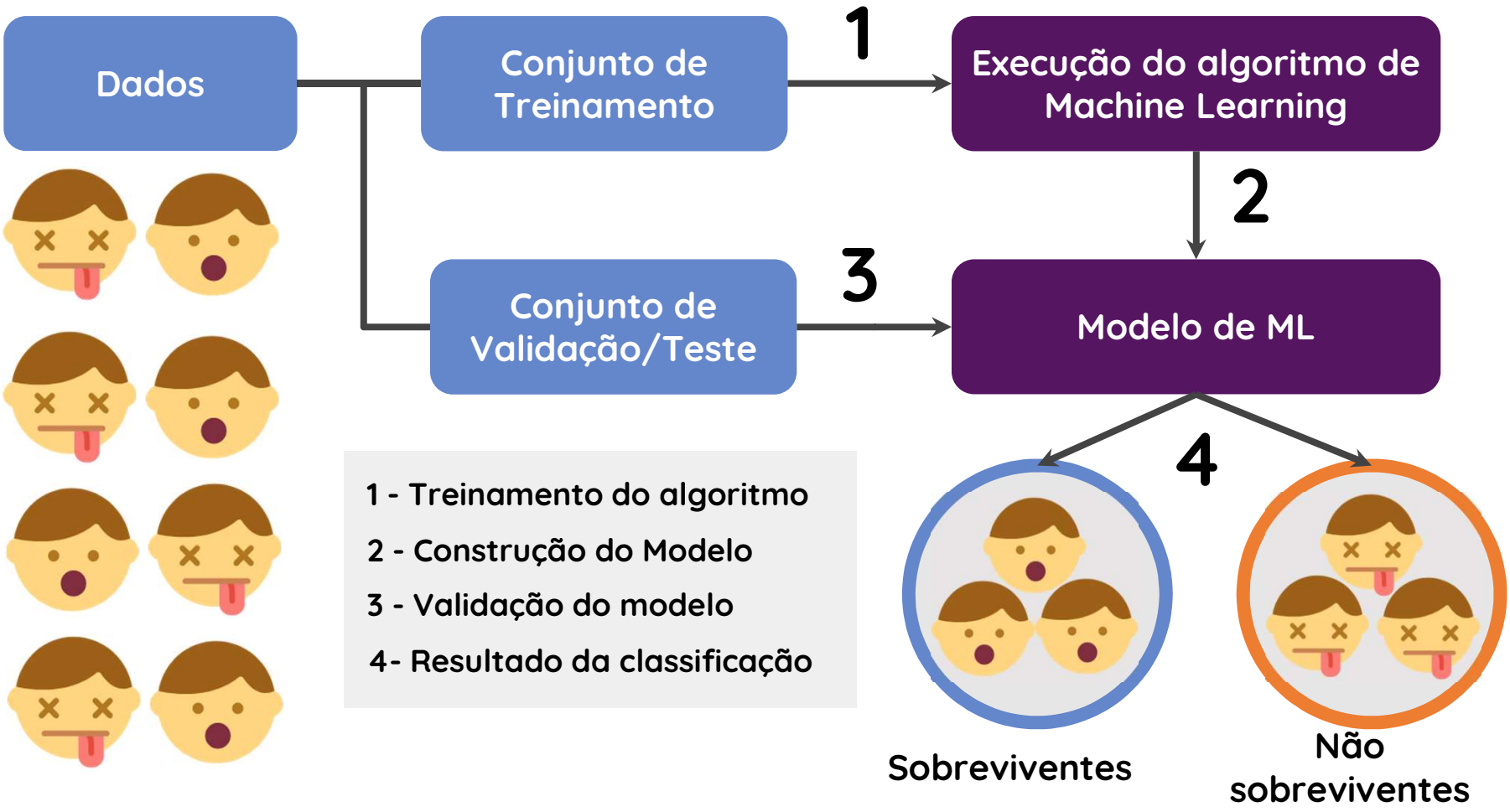
Algoritmo SVM (SVC)

Máquina de Vetores de Suporte

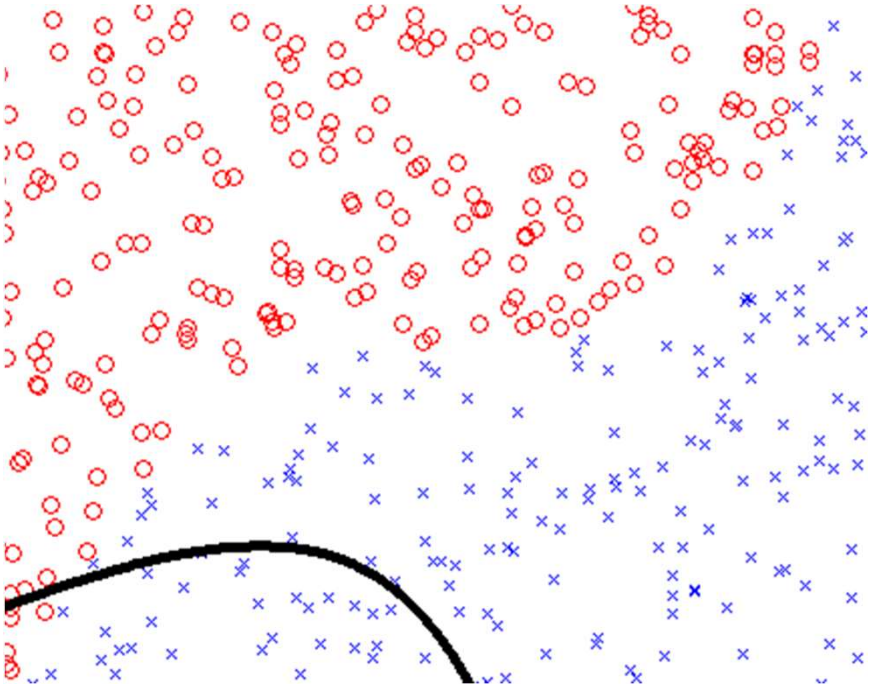
https://www.youtube.com/watch?v=1NxnPkZM9bc

SVM - Método de ML



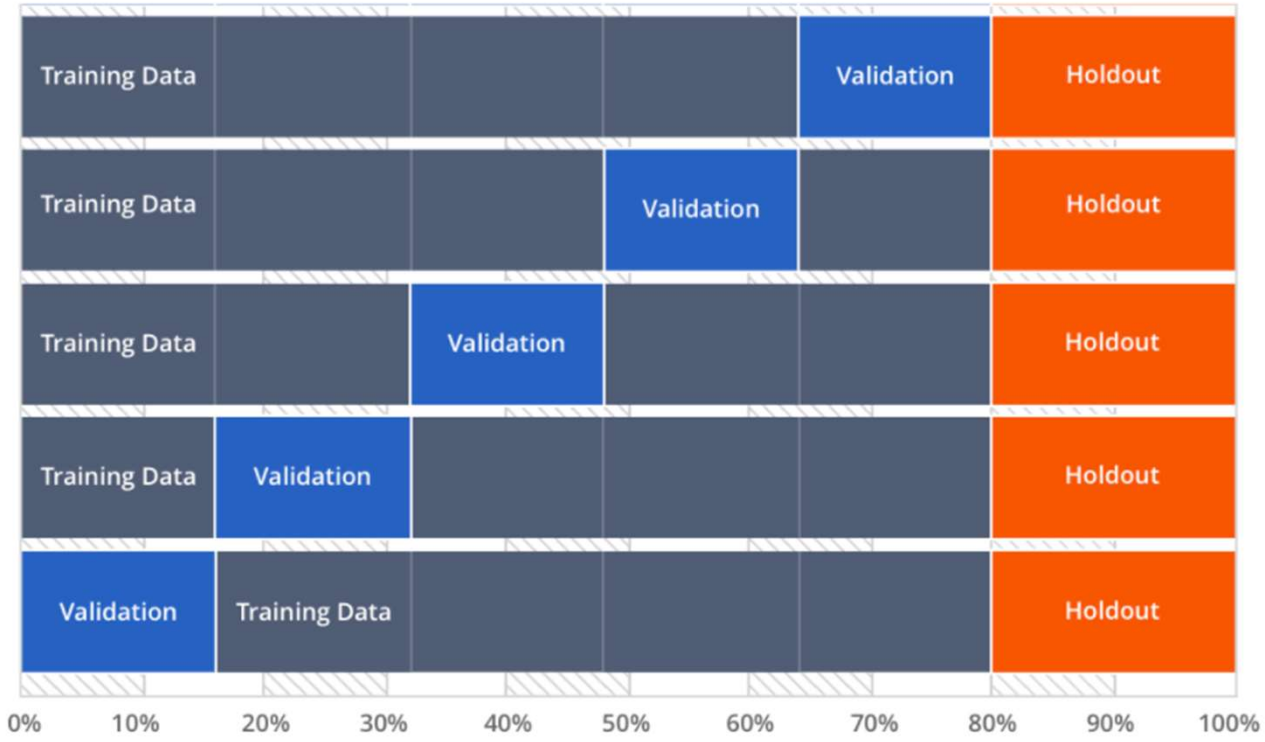

Dados
Conjunto de Treinamento
1
Execução do algoritmo de Machine Learning
2
Conjunto de Validação/Teste
3
Modelo de ML
4
1 - Treinamento do algoritmo
2 - Construção do Modelo
3 - Validação do modelo
4- Resultado da classificação
Sobreviventes
Não sobreviventes

Treinamento do Classificador



O hiperplano progressivamente **converge** para a geometria ideal para separar as duas classes de dados.

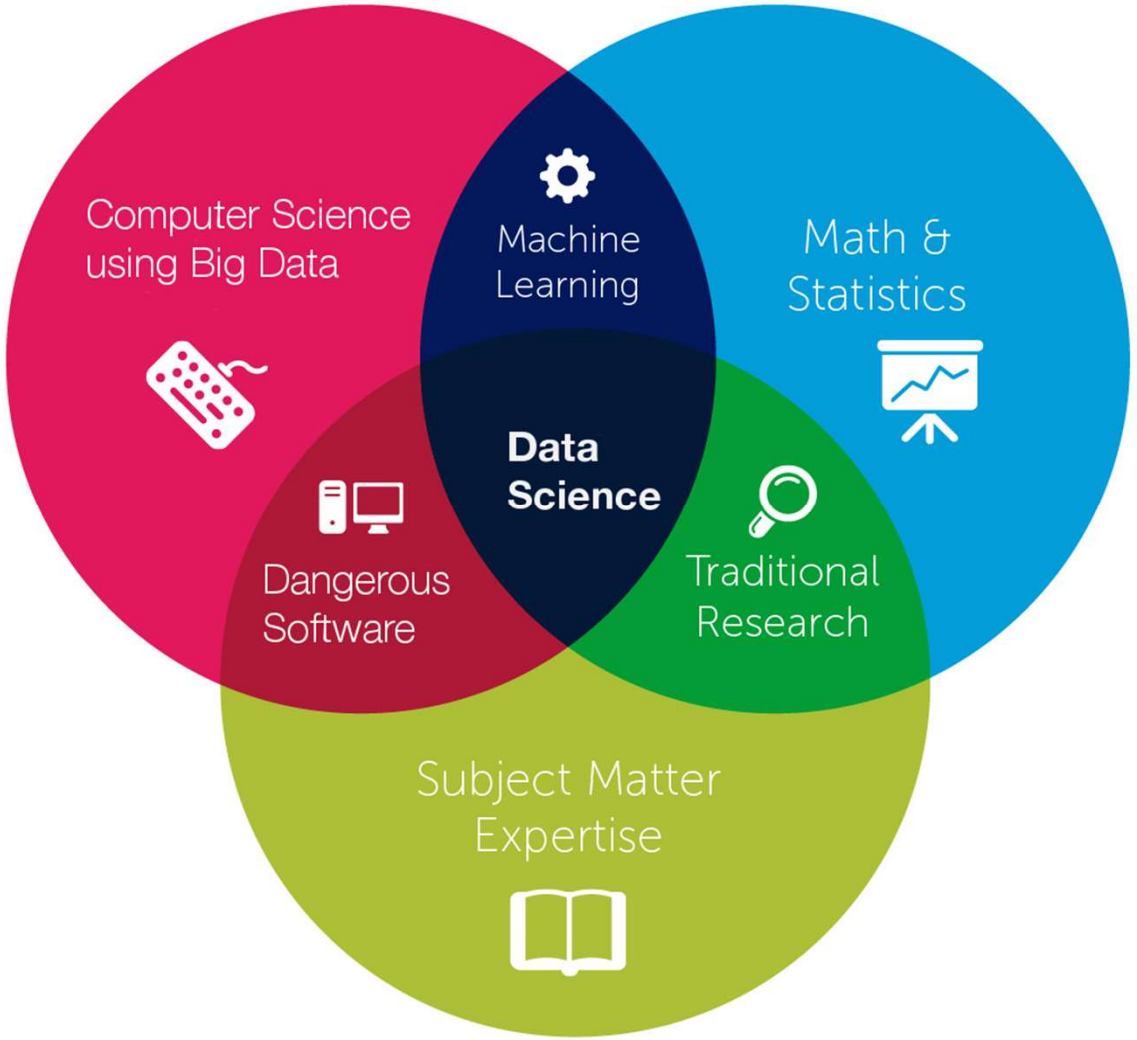
Testes - Validação Cruzada e Acurácia



Carreira de Cientista de Dados


R
APACHE
Spark
SQL
ANACONDA
Scala
+ableau


SCIENCE UNICORN
JUST DISCOVERED
THAT
HE DOESN'T EXIST
@twisteddoodles


Computer Science using Big Data
Machine Learning
Math & Statistics
Data Science
Dangerous Software
Traditional Research
Subject Matter Expertise

Como começar?

GitHub kaggle alura

Data Science Academy coursera M

http://datahackers.com.br/

https://serenata.ai/

https://brasil.io/home

Obrigada!

@carlaprvieira

@carlaprv

bit.ly/cpbr12-carla-workshop-feedback

bit.ly/cpbr12-carla-workshop

Comunidades de Data Science

https://www.meetup.com/pt-BR/R-Ladies-Sao-Paulo/

https://www.meetup.com/pt-BR/PyLadiesSP/

http://datahackers.com.br/

Referências

https://www.kaggle.com/sinakhorami/titanic-best-working-classifier

https://www.kaggle.com/mrisdal/exploring-survival-on-the-titanic

https://www.kaggle.com/c/titanic http://porque.uol.com.br/cards/media-simples-2/

http://minerandodados.com.br/index.php/2018/04/04/spotify-svm-python/

http://moralmachine.mit.edu/hl/pt

https://mygoodness.mit.edu/quiz

Referências

Machine Learning

- https://stanford.edu/~shervine/l/pt/teaching/cs-229/dicas-truques-aprendizado-maquina
- https://stanford.edu/~shervine/l/pt/teaching/cs-229/dicas-aprendizado-supervisionado
- https://stanford.edu/~shervine/l/pt/teaching/cs-229/dicas-aprendizado-nao-supervisionado

Referências

Python

- https://www.cursoemvideo.com/course/curso-python-3/
- https://paulovasconcellos.com.br/10-bibliotecas-de-data-science-para-python-que-ningu%C3%A9m-te-conta-706ec3c4fcef

Referências

Validação Cruzada

- https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6
- https://scikit-learn.org/stable/modules/cross_validation.html
- http://leg.ufpr.br/~walmes/ensino/ML/tutorials/01-cross-validation.html

Referências

Sobre-ajuste/overfitting

- https://medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76
- https://docs.aws.amazon.com/pt_br/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html