



Afinal, o que é esse tal de Machine Learning?

WoMakersCode Bootcamp @ SPTW

Quem sou eu?



Graduanda de Sistemas de Informação - USP

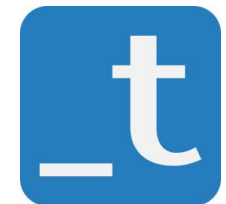
Apaixonada por Data Science e BI



@carlaprvieira



@carlaprv



Agenda

O que é Mineração de Dados?

Processo de mineração de dados

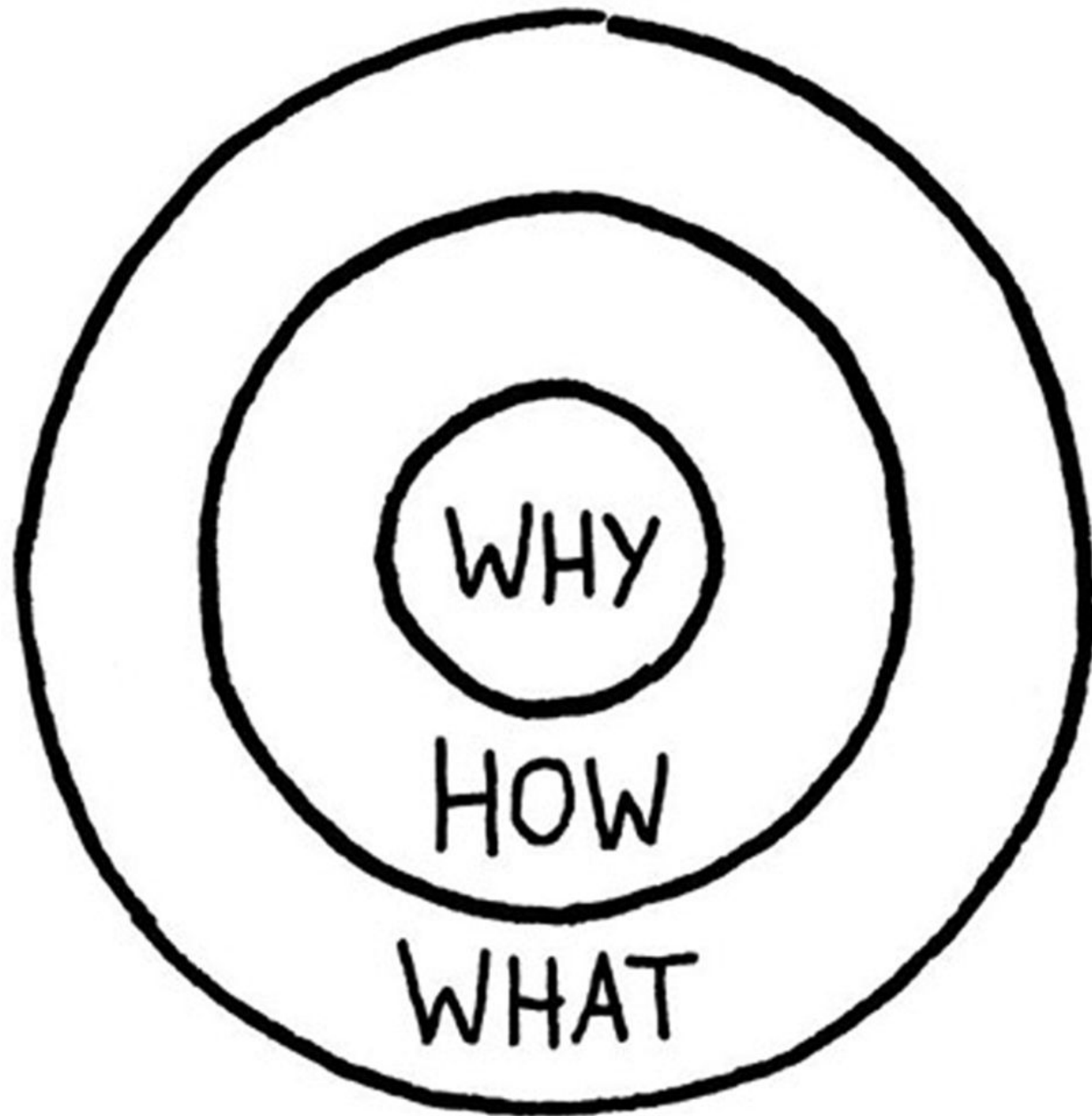
Projetos em Data Science

Algoritmos de Machine Learning

DESAFIO PRÁTICO!



Antes de começarmos...



Why = The Purpose

What is your cause? What do you believe?

Apple: We believe in challenging 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.

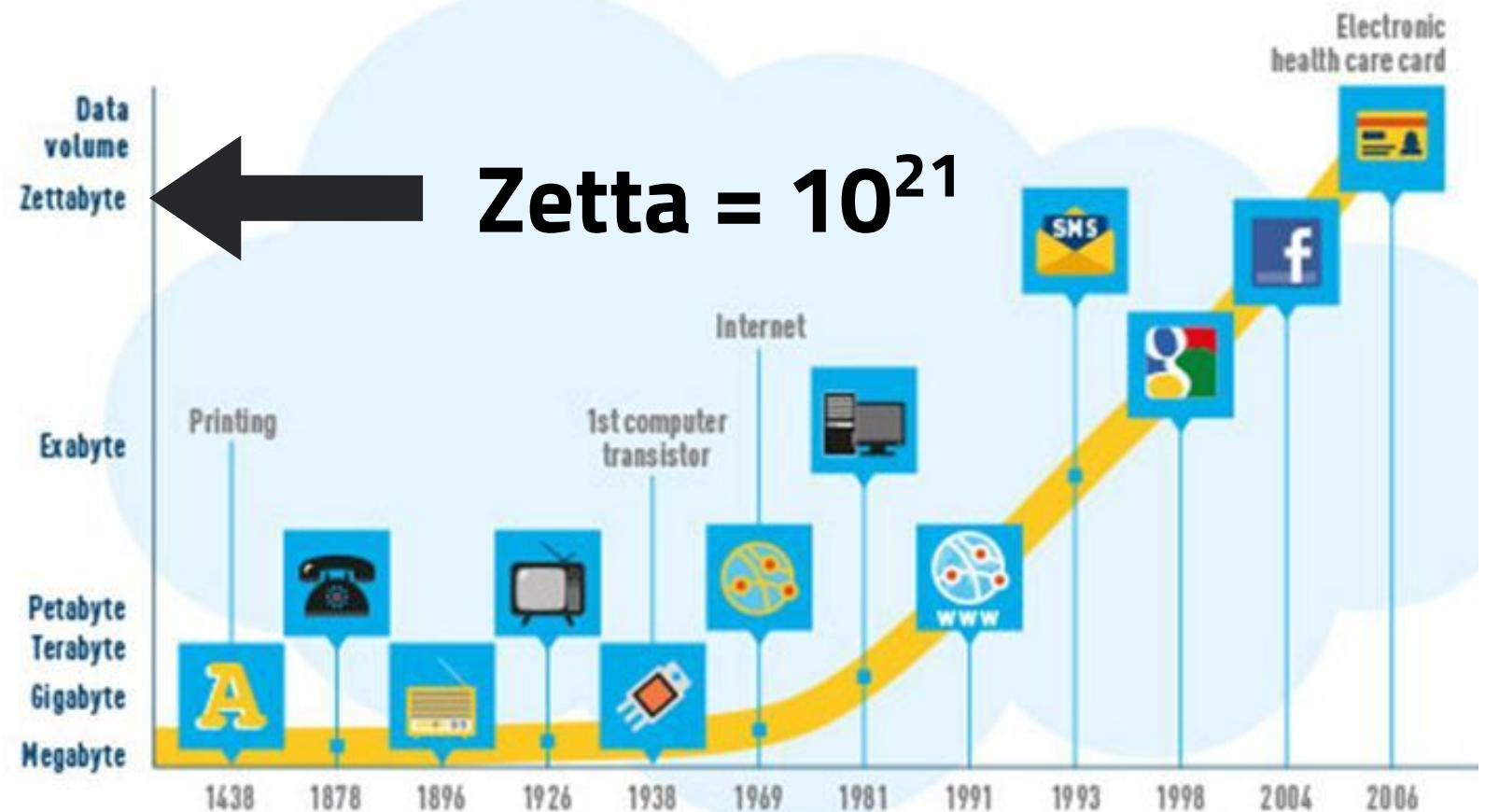
Apple: We make computers



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

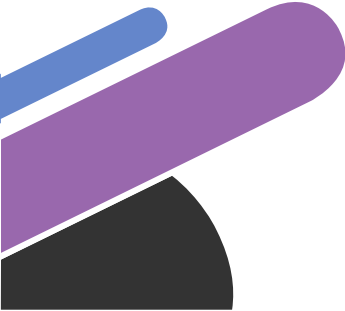


Source: Federal Association for Information Technology, Telecommunication and New Media (BITKOM). "Big Data im Praxisinsatz - Szenarien, Beispiele, Effekte."



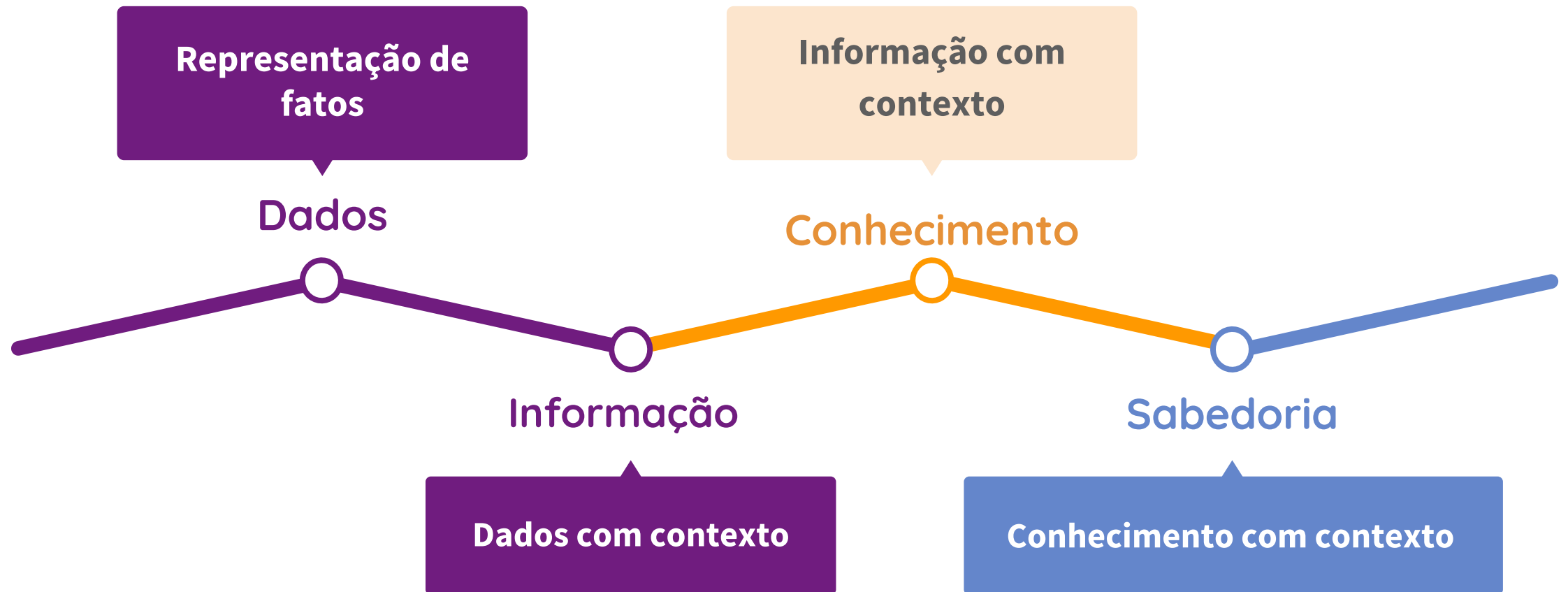
O que é Mineração de Dados?

*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**.*

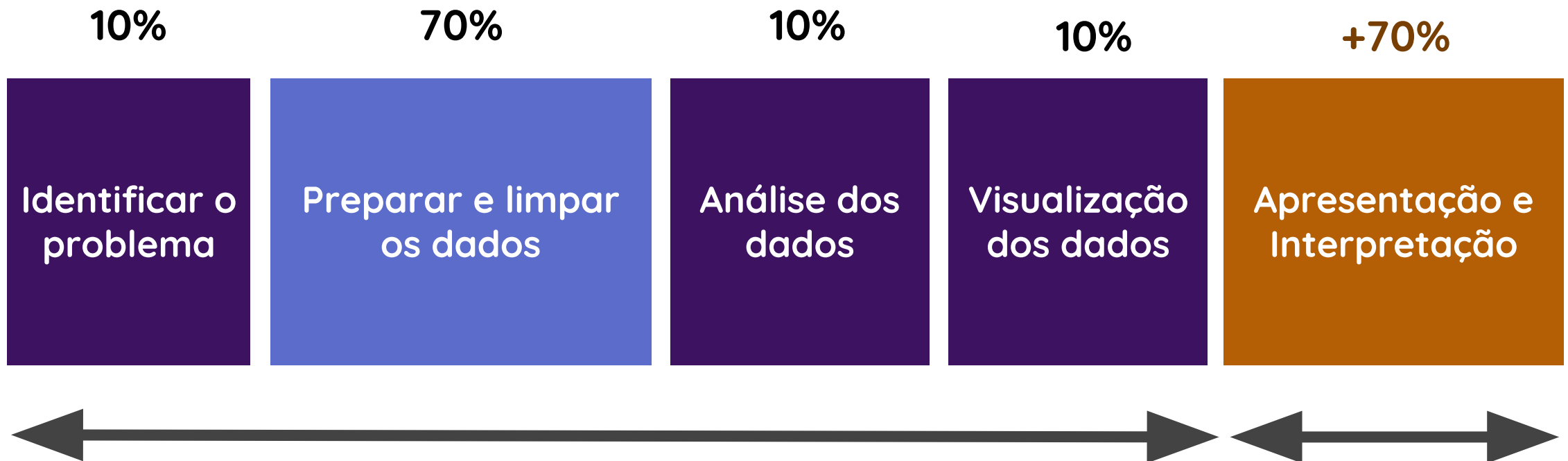


Cadeia de Valor do Conhecimento

(Davenport e Prusak - 1998)



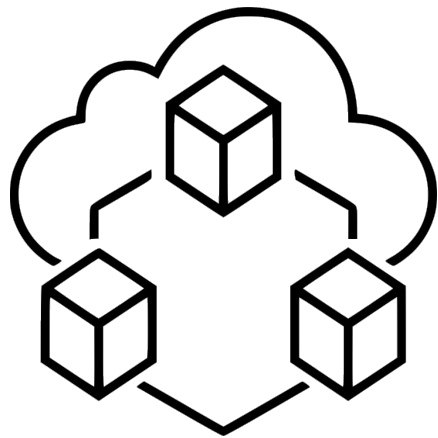
Ciclo de vida de um projeto de Data Science



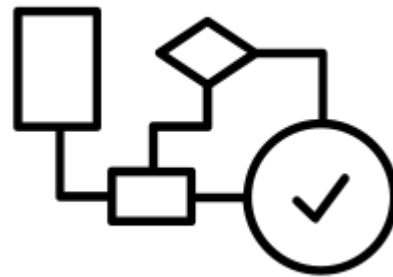
Técnicas de mineração de dados



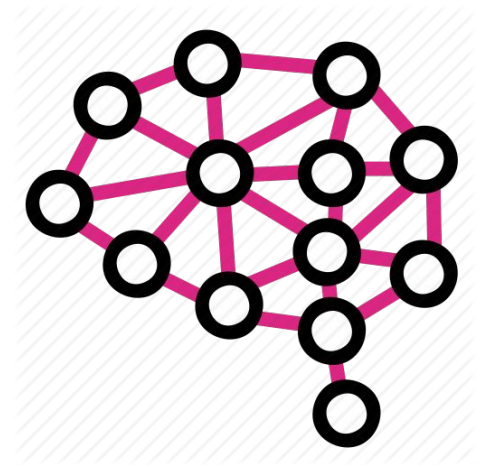
O que é Aprendizado de Máquina?




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)



Netflix Originals

Trending Now



Netflix

Sistema de recomendação



Continue Watching for Eric



Spotify

Daily Mix / Descobertas da Semana

FEITO PARA FELIPE

Descobertas da Semana

Sua mixtape semanal com músicas fresquinhas, novas descobertas e pérolas musicais escolhidas só para você. É atualizada toda segunda, então salve suas faixas favoritas!

Feito para Felipe Ventura pelo Spotify • 30 músicas, 1 h 50 m

SEGUIDOR
1

Baixar ☐

ÁLBUM



The Other Sides

14 horas atr...

Scarlett

Shangri-La



14 horas atr...

✓ Prison Break

NÉONHÈART

Prison Break

Curtiu

14 horas atr...

+ river

Curtida! Agora essa música vai aparecer na Sua Biblioteca.



14 horas atr...

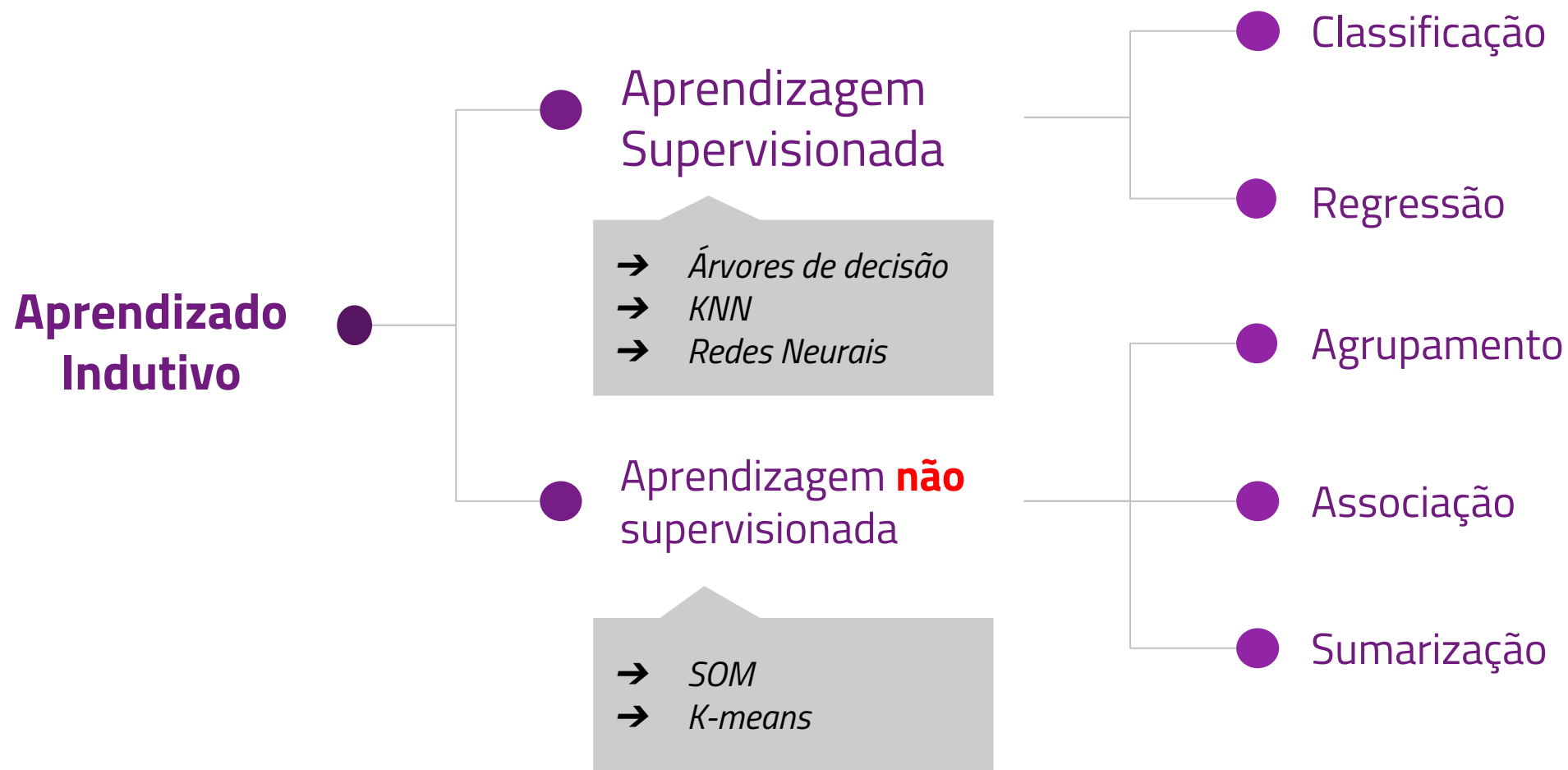
+ Finer Things

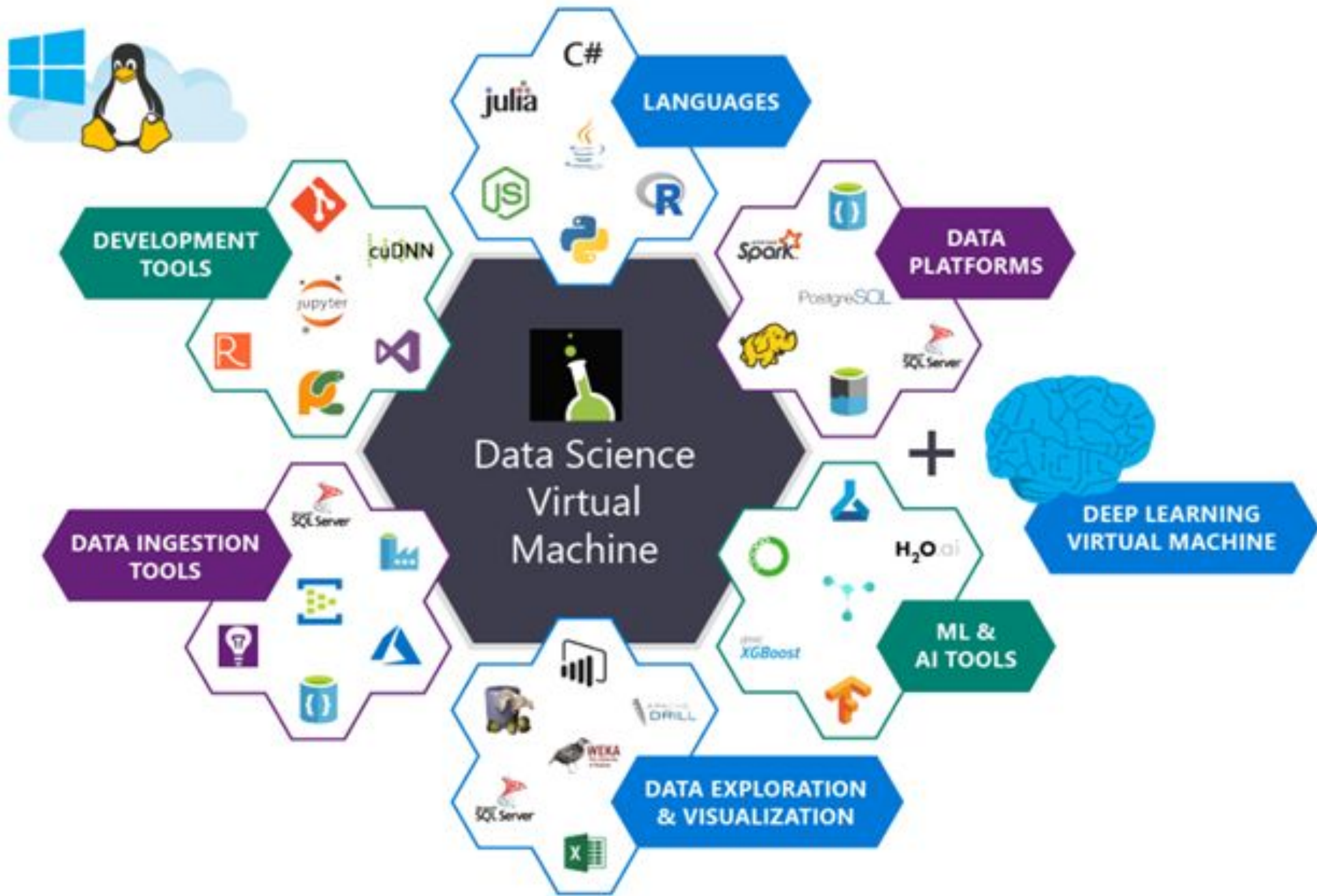


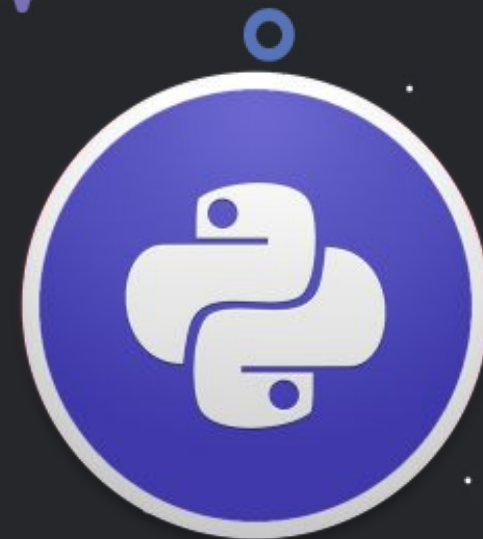
14 horas atr...



Abordagens de Aprendizado de Máquina





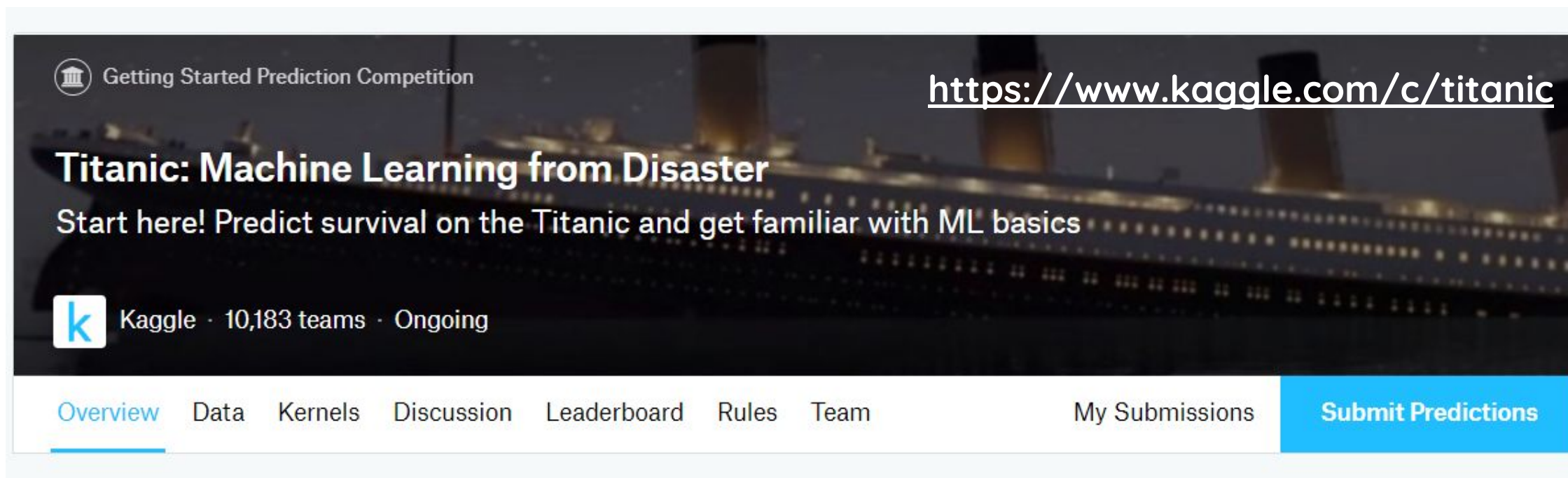


DESAFIO!

Explorando a sobrevivência no Titanic...



Desafio


A screenshot of the Kaggle website showing the "Titanic: Machine Learning from Disaster" competition. The background is a dark image of the Titanic ship at night. At the top left, there is a small icon of a classical building and the text "Getting Started Prediction Competition". To the right, the URL "https://www.kaggle.com/c/titanic" is displayed. The main title "Titanic: Machine Learning from Disaster" is in large white text, followed by the subtitle "Start here! Predict survival on the Titanic and get familiar with ML basics". Below this, the Kaggle logo (a blue 'k' in a square) is followed by the text "Kaggle · 10,183 teams · Ongoing". At the bottom, there is a navigation bar with links: "Overview" (underlined), "Data", "Kernels", "Discussion", "Leaderboard", "Rules", "Team", "My Submissions", and a blue button labeled "Submit Predictions".

Getting Started Prediction Competition

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

Titanic: Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics

 Kaggle · 10,183 teams · Ongoing

[Overview](#) [Data](#) [Kernels](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Team](#) [My Submissions](#) [Submit Predictions](#)

Predizer a sobrevivência dos passageiros no Titanic...

Dados



Data Description

Overview

The data has been split into two groups:

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

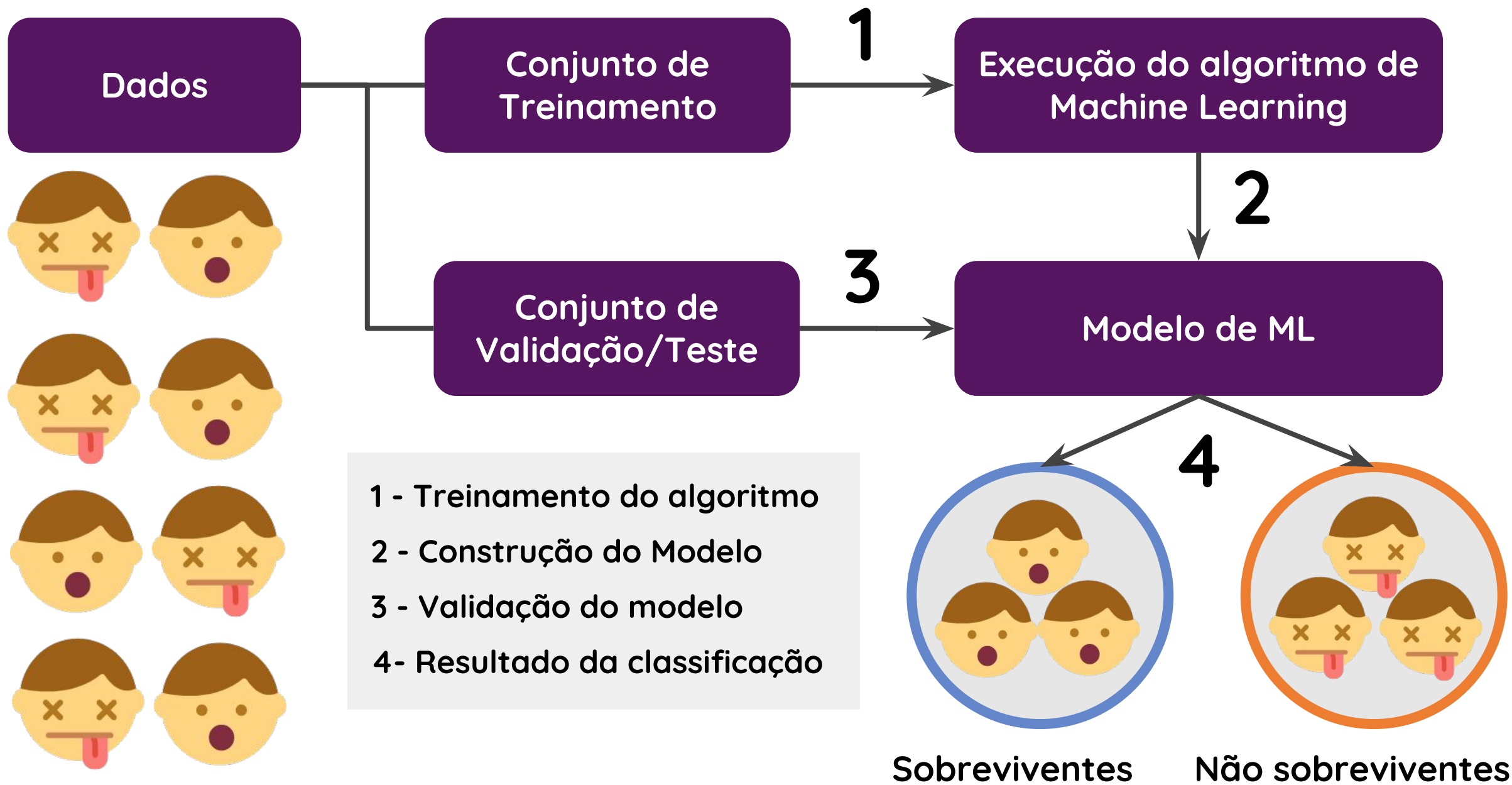
Conjunto de dados para treinamento

Conjunto de dados de teste

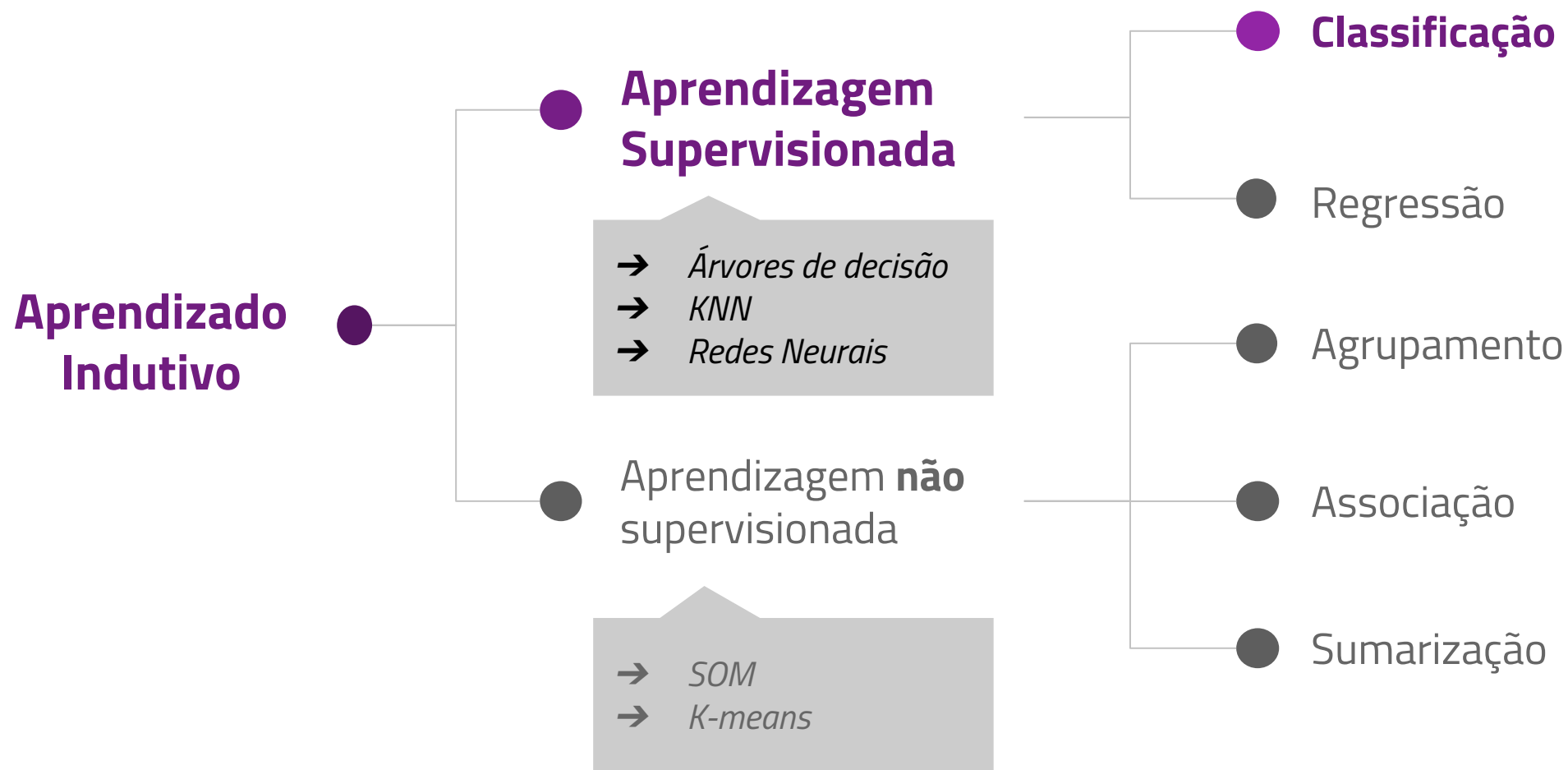
The training set should be used to build your machine learning models. For the training set, we provide the outcome (also known as the “ground truth”) for each passenger. Your model will be based on “features” like passengers’ gender and class. You can also use [feature engineering](#) to create new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



Abordagens de Aprendizado de Máquina





Iniciando nosso 1º Jupyter Notebook

Iniciando nosso Jupyter Notebook

1. Abrir o Prompt de Comando
2. Digite `cd \Workshop` e aperte **ENTER**
3. Digite `jupyter notebook` e aperte **ENTER**
4. Aguarde alguns minutos até abrir uma nova janela do Jupyter no seu navegador

Iniciando nosso Jupyter Notebook

1. Acesse <http://jupyter.org/try>
2. Clique em **Try Jupyter with Python**
3. Clique em **FILE >> New Notebook >> Python 3**
4. Se quiser, pode já renomear seu notebook!
5. Clique em **FILE >> Open**
 - a. irá abrir uma nova guia, deixe ela aberta também!



Criação de conta e download dos datasets

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

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!")
```

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** - gênero
- **Age** - idade
- **SibSp** - Número de irmãos / cônjuge a bordo
- **Parch** - Número de pais/filho a bordo
- **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 **peessoas viajando sozinhas**
- **Embarked** - contém valores nulos
- **Fare** - contém valores nulos
- **Age** - contém valores nulos
- **Name** - podemos extrair o título da pessoa

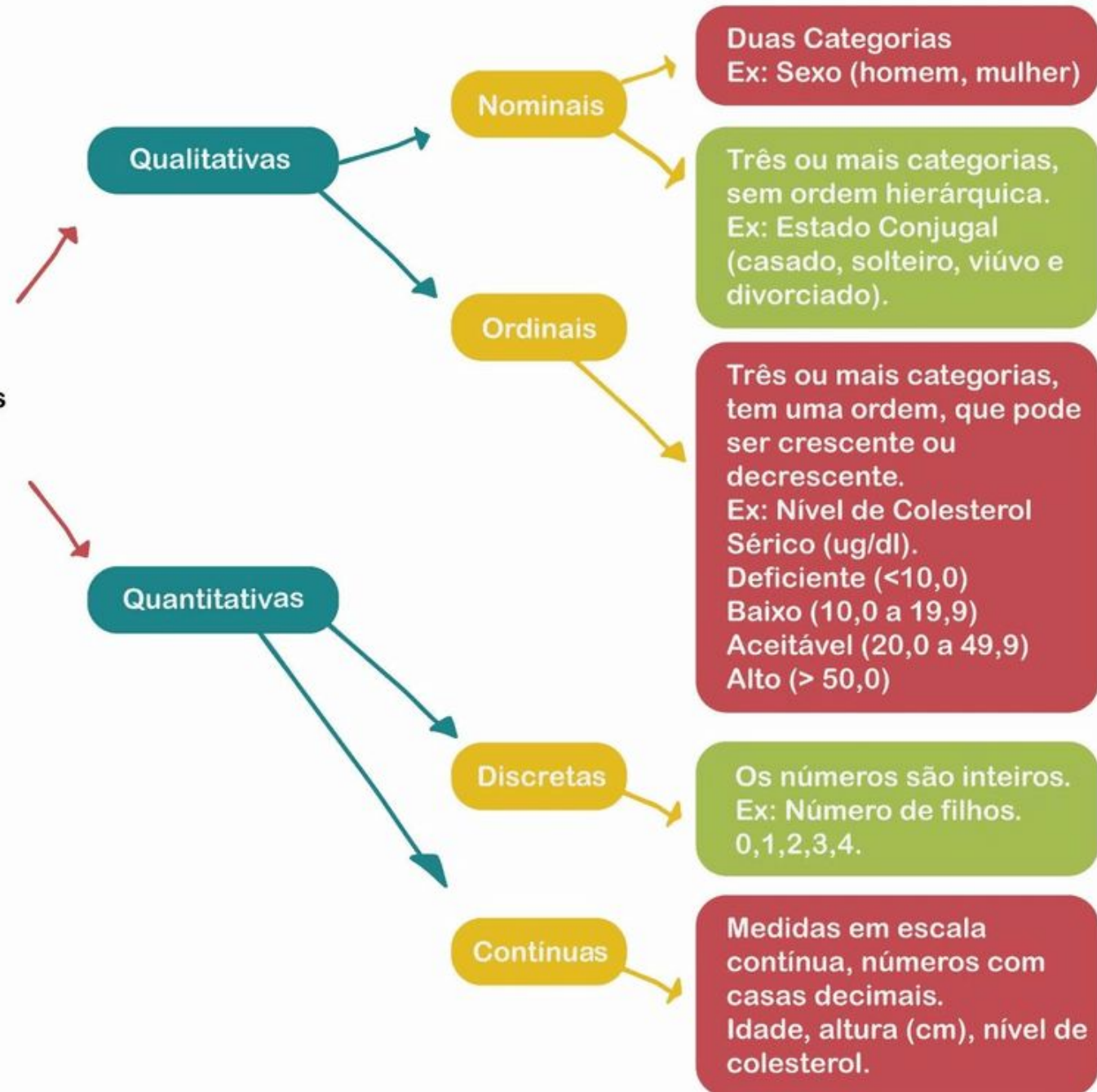
Limpeza dos dados

- **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 $[m\acute{e}dia - erro, m\acute{e}dia + erro]$
 - etc...
- **Transformar todos os dados em valores numérico**

Limpeza dos dados

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

Variáveis





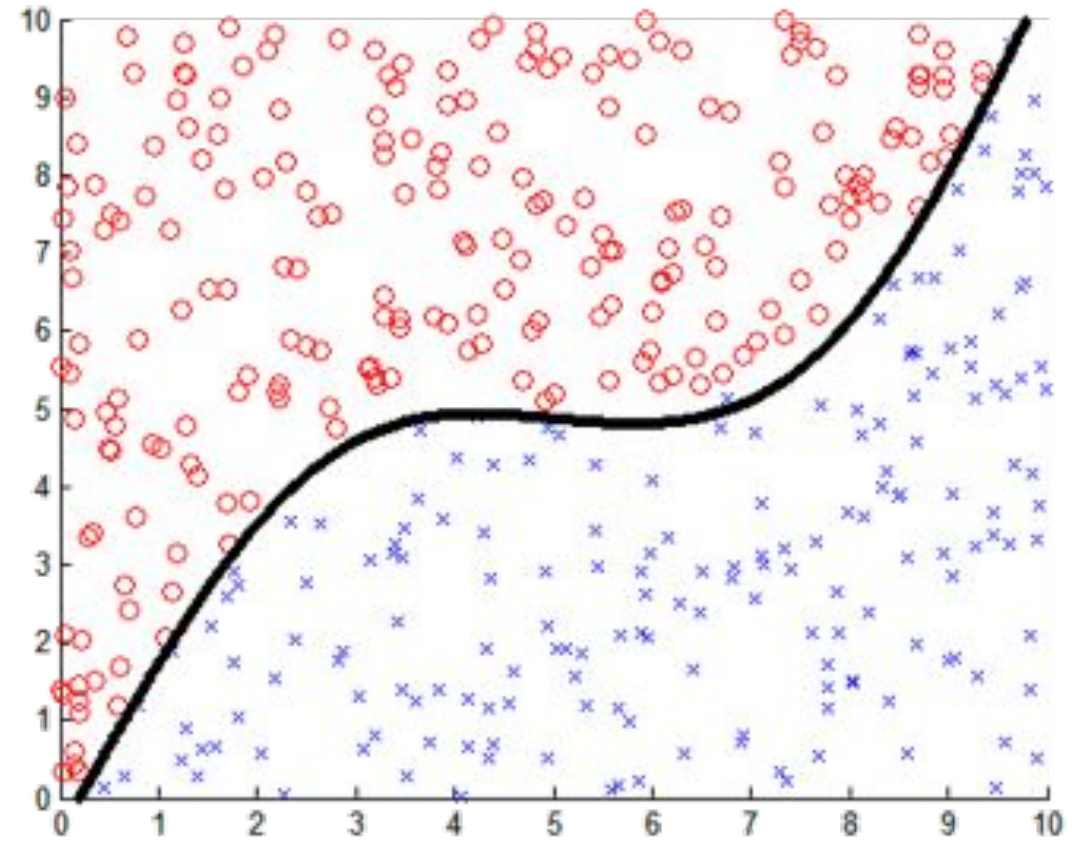
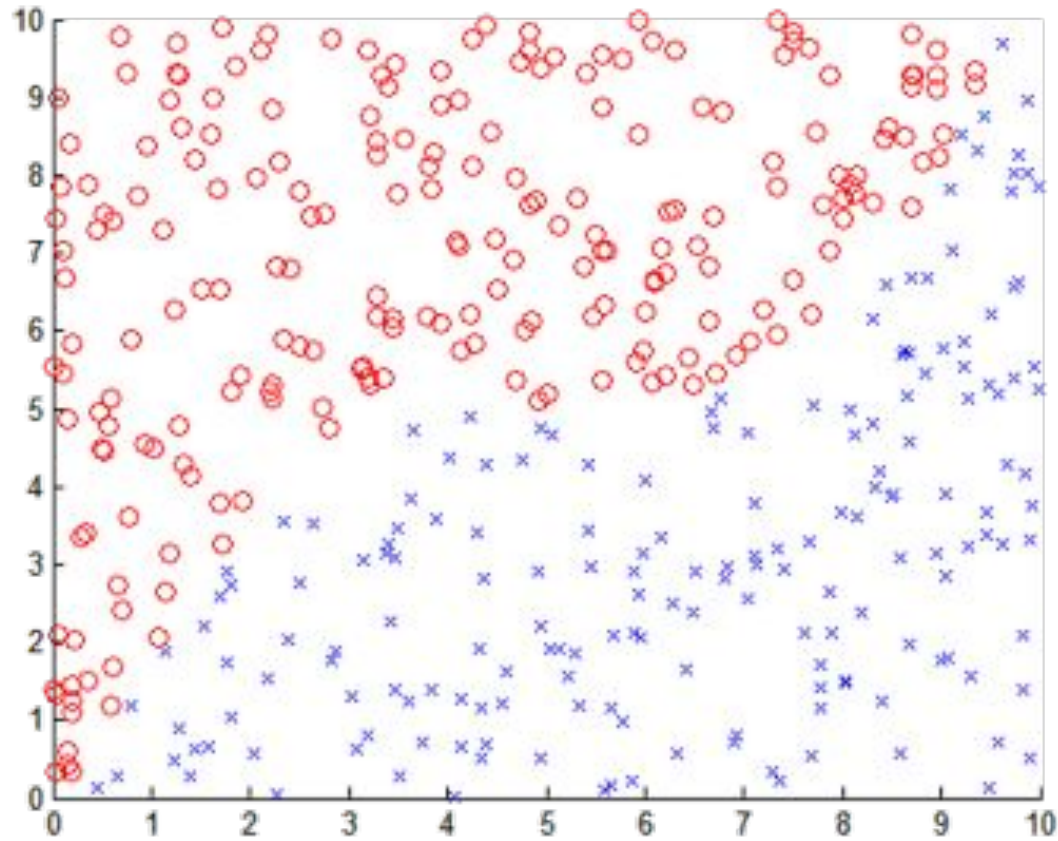
Algoritmo SVM (SVC)

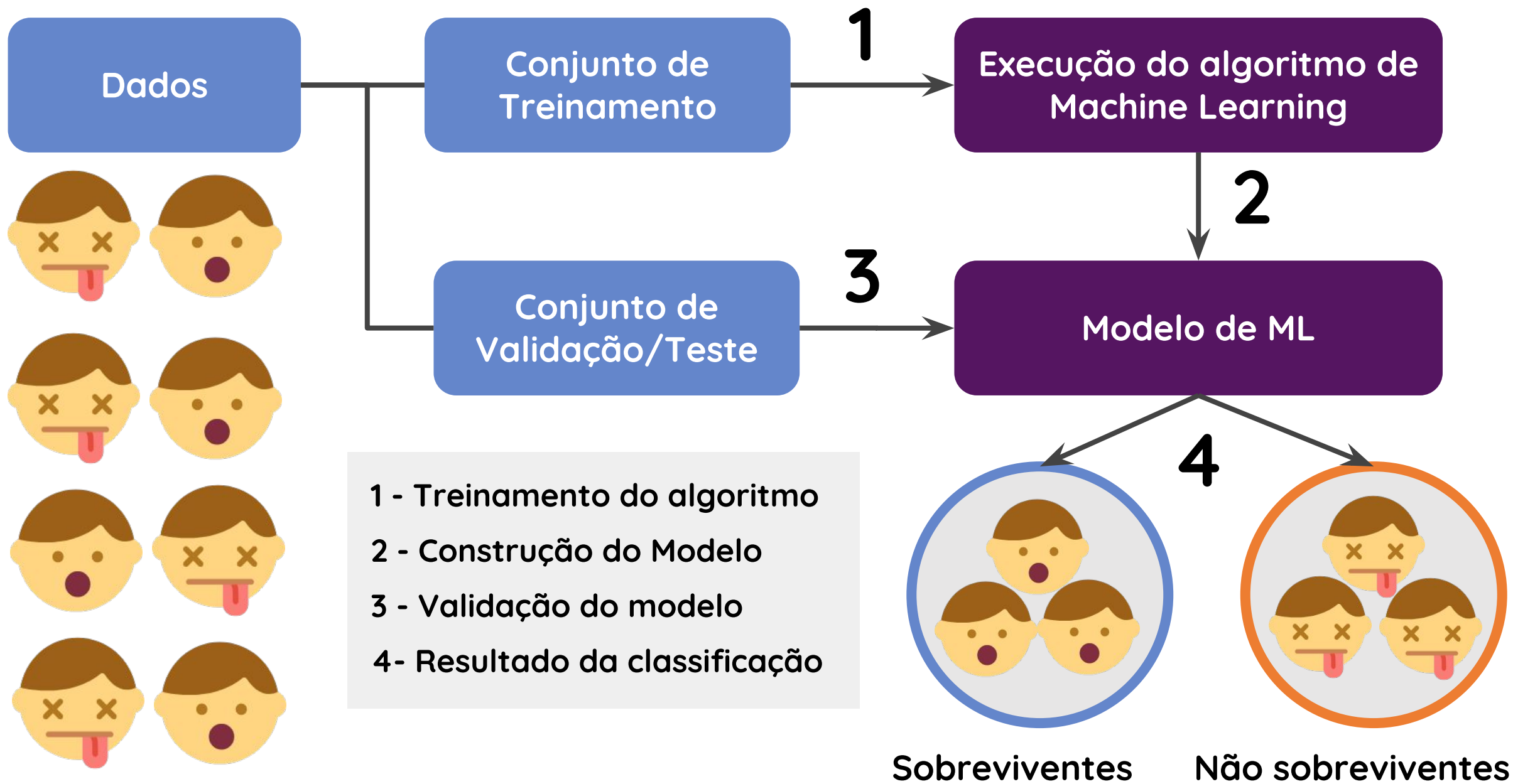
Máquina de Vetores de Suporte

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

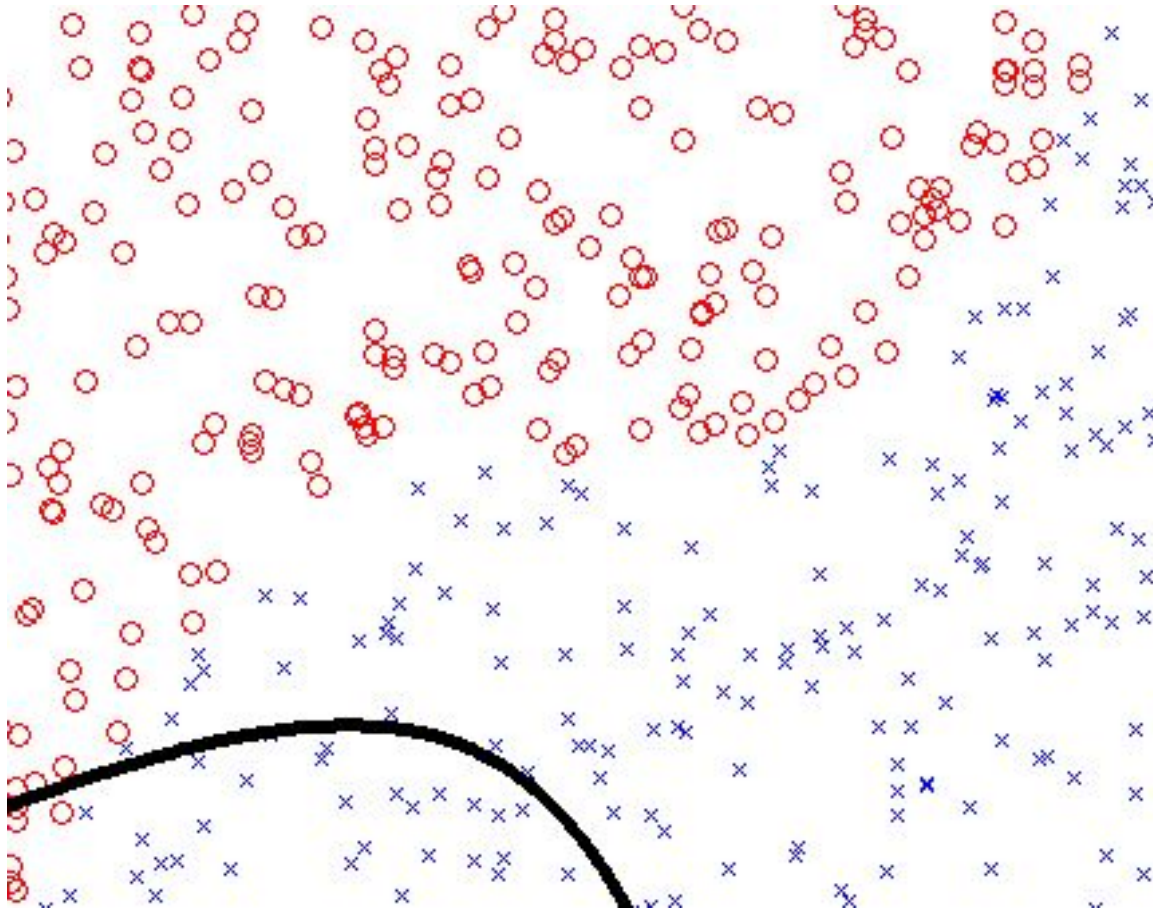


SVM - Método de ML



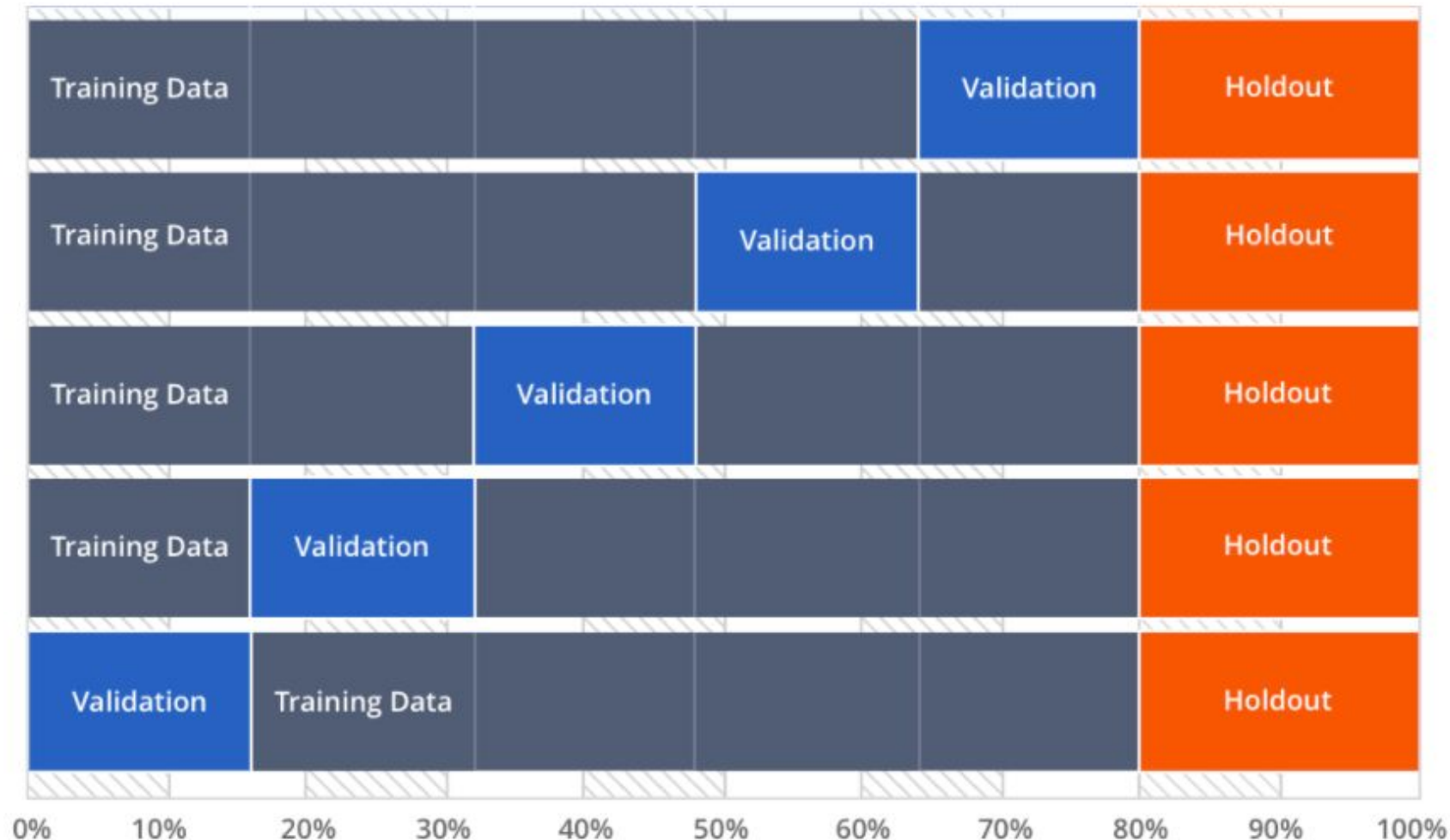


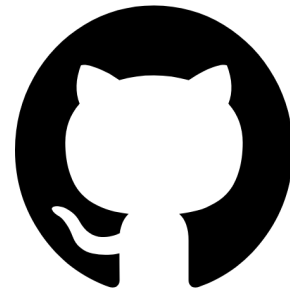
Treinamento do Classificador



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

Testes - Validação Cruzada e Acurácia



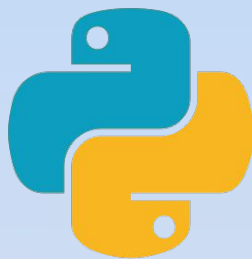


Código completo do workshop

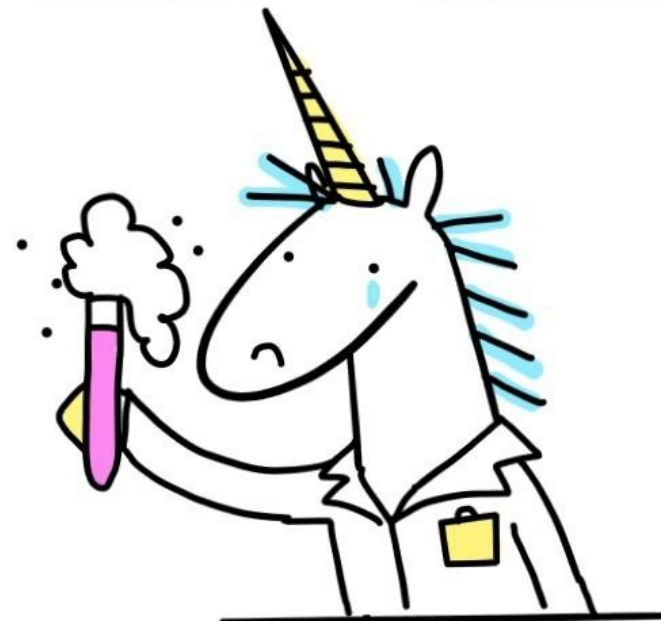
<https://github.com/carlaprv/workshop-machine-learning-titanic>



Carreira de Cientista de Dados

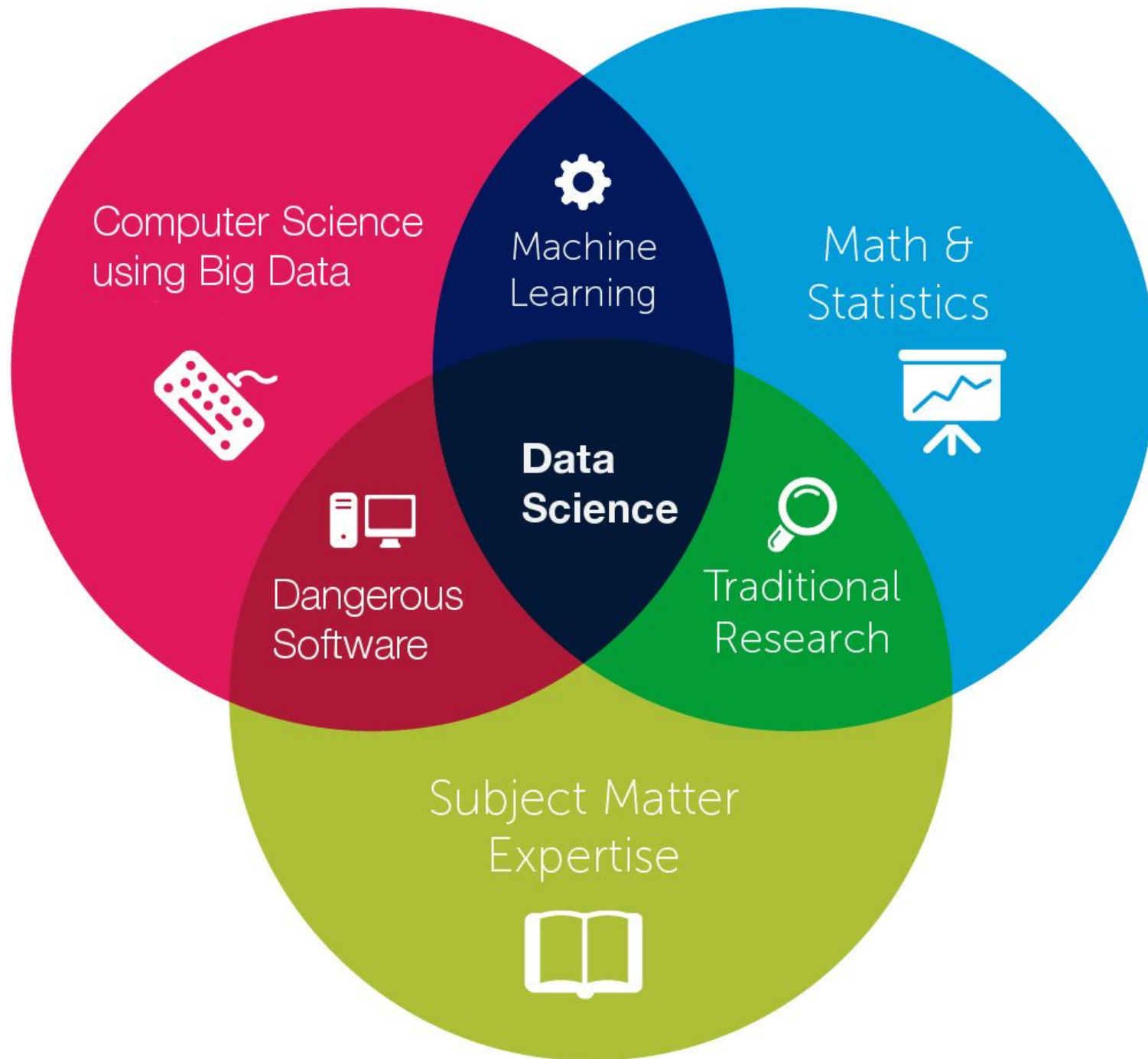


SCIENCE UNICORN



JUST DISCOVERED
THAT
HE DOESN'T EXIST

@twisteddoodles





Como podemos contribuir?



Compartilhando código!

GitHub kaggle





OPERAÇÃO

SERENATA DE AMOR

INTELIGÊNCIA ARTIFICIAL
PARA CONTROLE SOCIAL DA
ADMINISTRAÇÃO PÚBLICA

<https://serenata.ai/>

O Brasil em dados libertos

Repositório de dados públicos disponibilizados em formato acessível

Manifesto

Venha entender o que motivou a criação do Brasil.IO e quais são os objetivos do projeto!

[VER MAIS](#)

Datasets

Confira a lista completa de todas as base de dados liberadas.

[VER MAIS](#)

Sugira um Dataset

Conhece algum dataset bacana que gostaria de ver aqui? Sugira para nós! =)

[VER MAIS](#)

Colabore

Achou o projeto legal e quer contribuir? Veja aqui como.

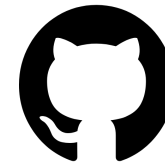
[VER MAIS](#)

<https://brasil.io/home>

Obrigada!



@carlaprvieira



@carlaprv



/blogcarlavieira

Feedback



@carlaprvieira



@kamilah_santos

donthpad.com/carlaprv/workshop ou

<https://bit.ly/2QeD1DI>

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