

# Python para Machine Learning!

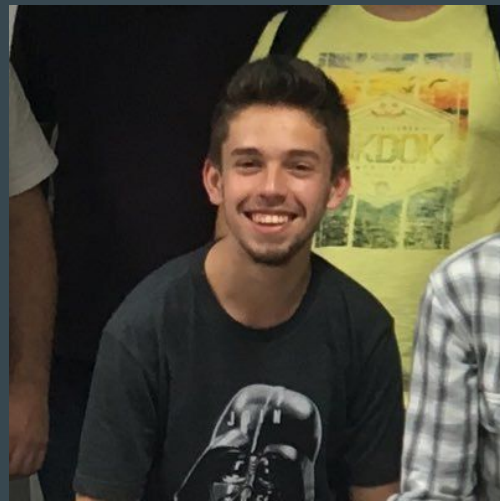
...

Por: @vmesel

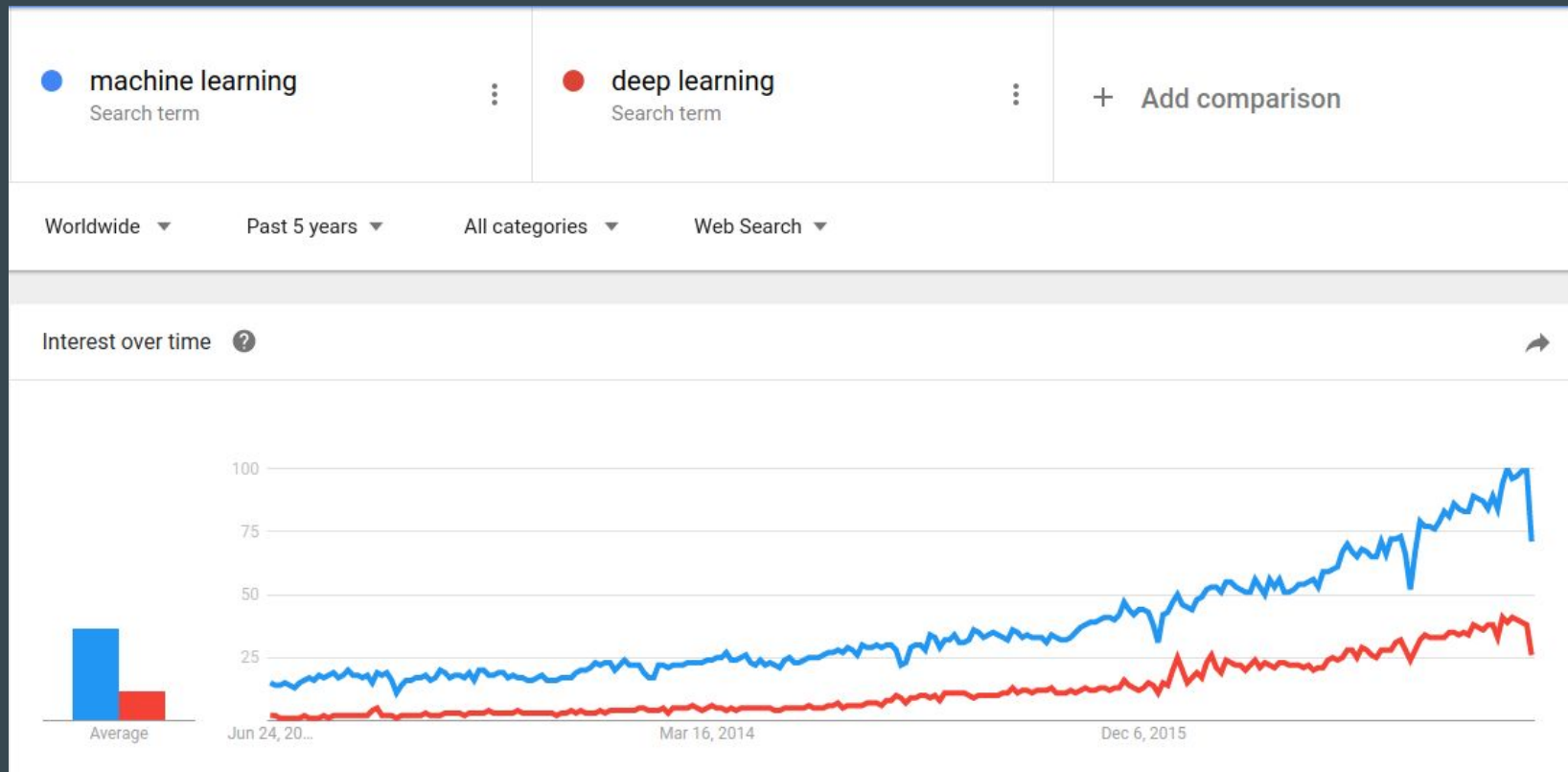
# Quem sou eu?

- [github.com/vmesel](https://github.com/vmesel)
- [twitter.com/vmesel](https://twitter.com/vmesel)
- Pesquisador de Machine Learning no IME-USP

COGNITIVO.AI



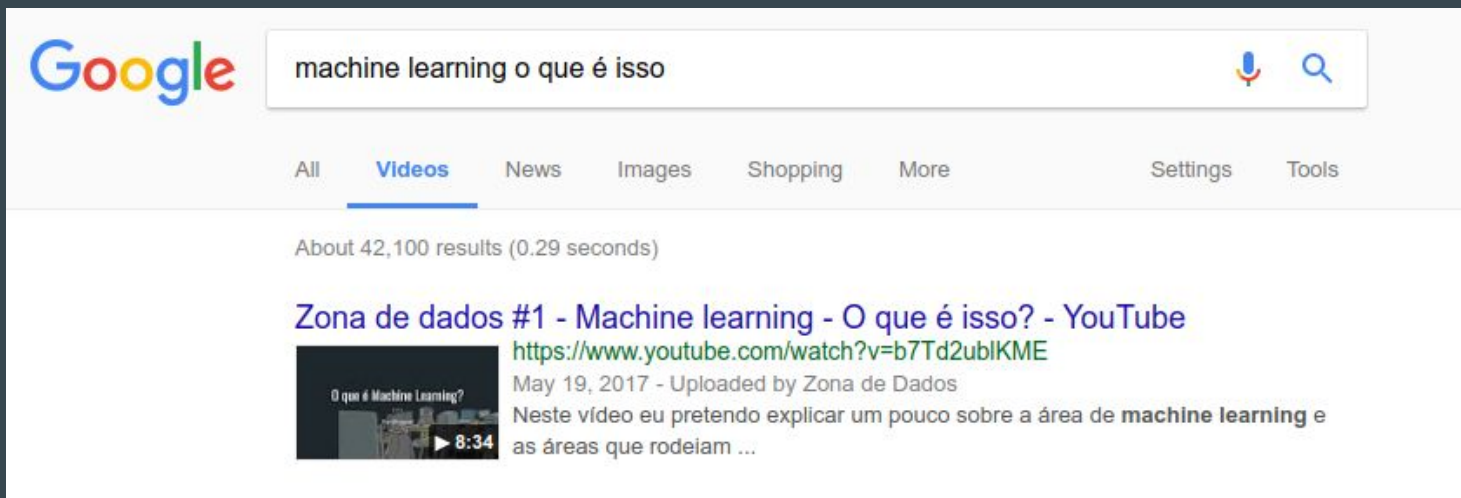
# Por que devemos ficar atentos a esta buzzword?



Disclaimer: Vou falar somente de Machine Learning “**tradicional**”, não de Deep Learning!

# Para quem nunca ouviu falar de Machine Learning:

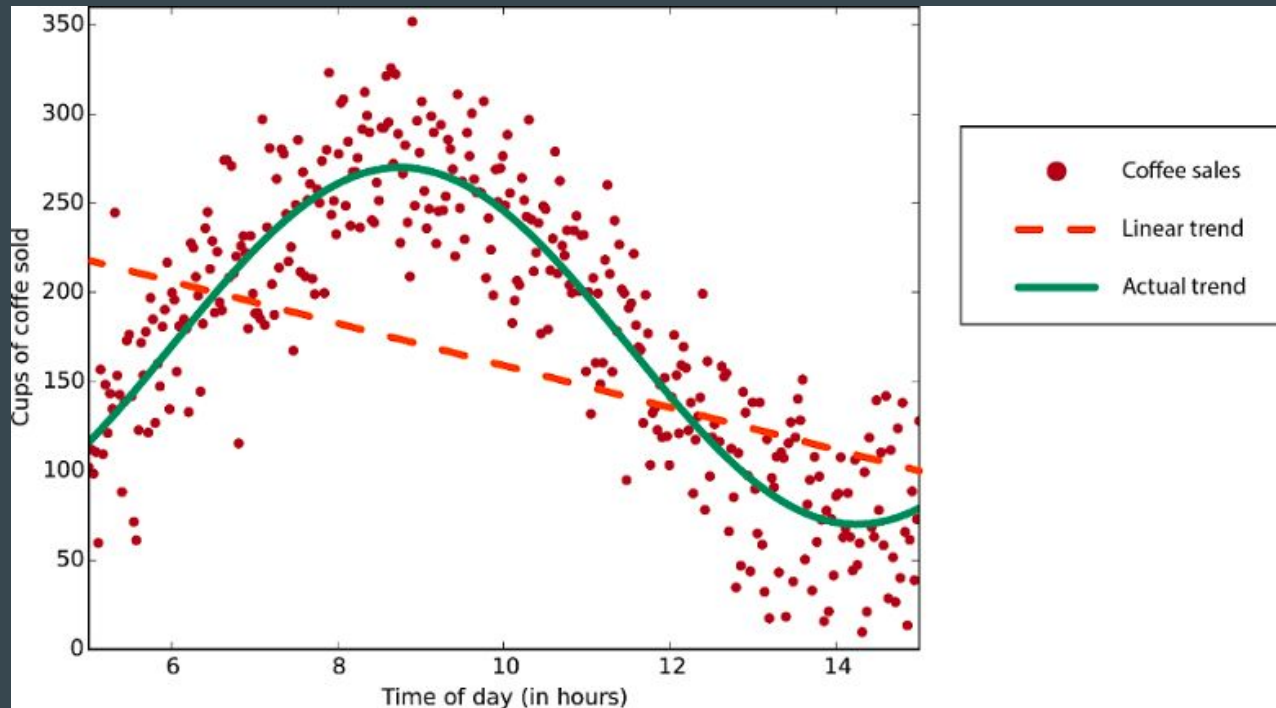
Para quem não sabe, Machine Learning é uma área da ciência da computação e inteligência artificial.



The image is a screenshot of a Google search interface. At the top left is the Google logo. The search bar contains the text "machine learning o que é isso". To the right of the search bar are icons for voice search and a magnifying glass. Below the search bar, there are tabs for "All", "Videos", "News", "Images", "Shopping", "More", "Settings", and "Tools". The "Videos" tab is selected and underlined. Below the tabs, it says "About 42,100 results (0.29 seconds)". The first search result is a video titled "Zona de dados #1 - Machine learning - O que é isso? - YouTube" in blue text. Below the title is the URL "https://www.youtube.com/watch?v=b7Td2ubIKME" in green text. To the left of the URL is a small video thumbnail showing a person speaking, with the text "O que é Machine Learning?" and a play button icon and "8:34". To the right of the URL, it says "May 19, 2017 - Uploaded by Zona de Dados" and "Neste vídeo eu pretendo explicar um pouco sobre a área de machine learning e as áreas que rodeiam ...".



# Só para poder resumir



# Algoritmos de Machine Learning



# E o que o Python tem com isso?

- Linguagem escolhida por ser fácil
- Meio acadêmico e indústria estão usando loucamente
- Várias ferramentas de Machine/Deep Learning disponíveis com baterias inclusas





## Outras bibliotecas de Machine Learning (elas também existem)



**Como nós podemos montar os nossos modelos  
de Machine Learning da melhor forma  
possível?**

# Escolha dados confiáveis



# Entenda o tipo de problema que você quer resolver e que dados você têm

- Não adianta você querer resolver um problema supervisionado sem saber as classes dos itens a serem classificados
- Você deve também saber se o problema é classificatório ou regressivo

# Teste diferentes modelos matemáticos

Diferentes modelos possuem diferentes métodos para aproximar e manipular dados, por isso sempre teste modelos diferentes para testar sua predição

```
from sklearn import svm
cl = svm.LinearSVC() # Otimizações aqui
cl.fit(train[features], train['FILE'])
```

```
from sklearn.ensemble import RandomForestClassifier
cls = RandomForestClassifier() # Otimizações aqui
cls.fit(train[features], train['FILE'])
```

# Faça o tuning do algoritmo que você está utilizando

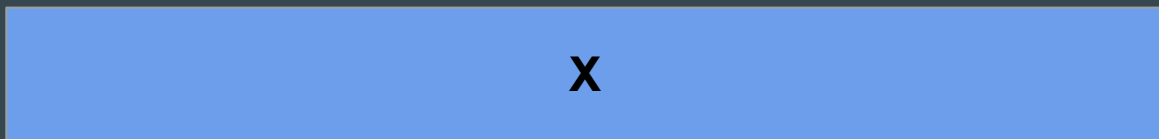
As vezes um modelo, em seu estado original, pode não ser o mais otimizado, para isso você pode tunar os parâmetros do modelo. Utilize um GridSearchCV ou RandomSearchCV para poder tunar hiperparâmetros.

```
RandomForestClassifier(n_estimators=10, criterion='gini', max_depth=None, min_samples_split=2,  
min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto',  
max_leaf_nodes=None, bootstrap=True, oob_score=False, n_jobs=1, random_state=None,  
verbose=0, warm_start=False, class_weight=None)
```

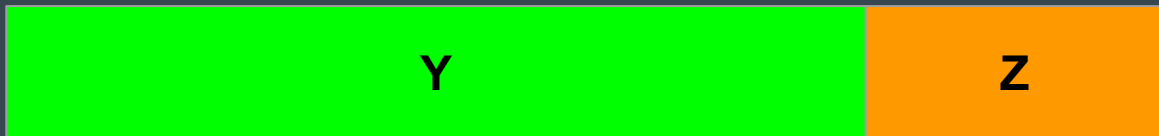
# Cross-Validation

Dado um dataset  $X$ , dividimos ele em dois outros pedaços  $Y$ ,  $Z$ , onde  $Y$  contém uma certa porcentagem e  $Z$  contém 100 - porcentagem de  $Y$ . Em seguida utilize o algoritmo de cross-validation com uma certa porcentagem de  $Y$  rodando  $A$  vezes.

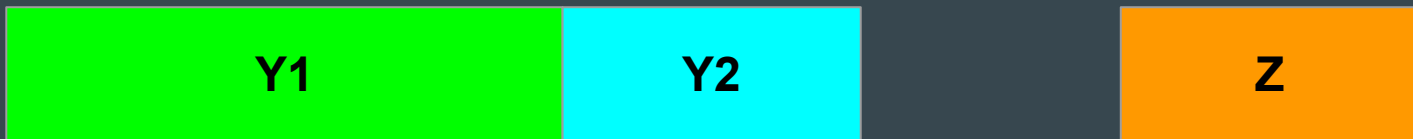
1



2



3



`sklearn.model_selection.cross_val_score`

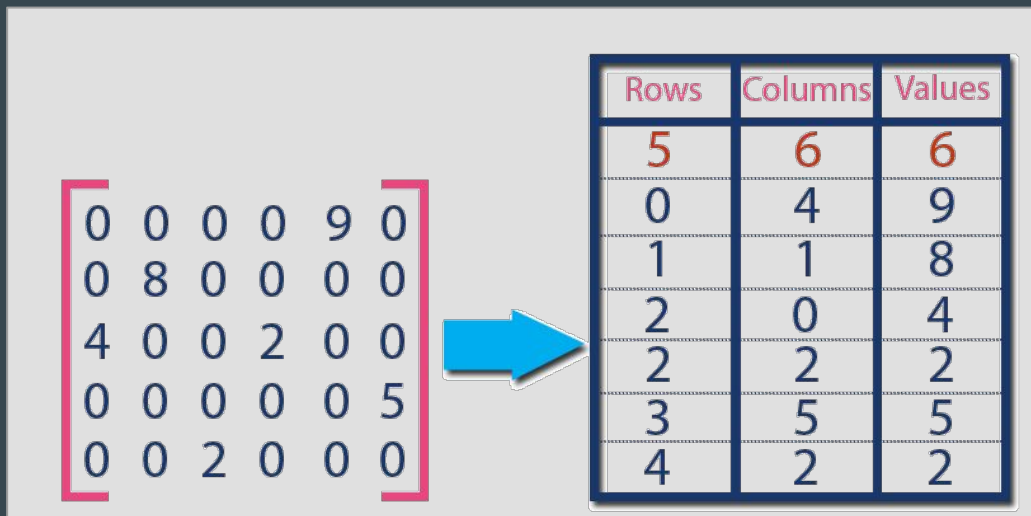
# Faça um pouco de *Feature Engineering*

- Engenharia de Feature é um campo do ML que visa aumentar a precisão do modelo alterando suas features originais, para as que somente predizem.
- Algumas técnicas:
  - Trate dados outliers (dados muito gritantes) normalizando todos os dados
  - Remova campos não preenchidos (caso você tenha que lidar com usuários) ou campos que você não irá usar
  - Crie novas features com dados que possam ser relevantes (médias, desvios padrões, classe do item e etc)



# Diminua a dimensionalidade do seu dataset

Se você estiver trabalhando com matrizes esparsas (matrizes cheias de valores nulos e poucos campos com valores significativos) utilize um algoritmo de diminuição de dimensionalidade.



# Fique interado das novidades!



**Acompanhe o meu Twitter  
@vmesel e o Zona de Dados!**

**...**