

# Workshop Ciência de Dados

## 1ª semana

- *Introdução à linguagem Python com Anaconda*
- *Prática: Regressão KNN, Random Forrest, Clustering e SVM com Python*

José Humberto Cruvinel

Contato: [jose.junior@prof.unibh.br](mailto:jose.junior@prof.unibh.br)

<https://www.facebook.com/jhcruvinel>

# Apresentação do professor

- **José Humberto Cruvinel**
- Formação
  - Graduado em Engenharia Elétrica / Ênfase em Computação (UFMG)
  - Especialista em Análise de Sistemas de Informação (PUCMG)
  - Mestre em Administração Pública / Gestão da Informação (FJP)
  - MBA em Gerenciamento de Projeto de TI / FGV
- Atuação Profissional
  - Analista no Serpro
  - Professor no UNI-BH
- Certificações Profissionais
  - PMP, CFPS, CTFL, ITIL v3, COBIT 4.1, MCP, Scrum Master, DB2 10, Rational Team Concert v3
- Contato e sites:
  - E-mail institucional: [jose.junior@prof.unibh.br](mailto:jose.junior@prof.unibh.br)
  - Facebook / Messenger: <https://www.facebook.com/jhcruvinel>
  - Curriculum Lattes: <http://lattes.cnpq.br/0955831456676522>

# Agenda – 1º sábado

1. Apresentação do Workshop



2. Contextualização



3. Introdução à linguagem Python



4. Machine Learning

# 1. Apresentação

# Programação do Workshop

1º sábado

- Introdução à Machine Learning
- Prática: Regressão, KNN, Random Forrest, Clustering e SVM com Python

2º sábado

- Introdução à Redes Neurais e Deep Learning
- Prática: Visão Computacional com Python e TensorFlow

# Ferramentas de apoio



UNIBH Workshop Ciência de Dados

- Slides
- Laboratórios

Suporte aos exercícios

# Ferramentas que vamos utilizar



## Anaconda - <https://www.anaconda.com>

- É uma plataforma que facilita a gestão de ambientes virtuais
- Possui as seguintes características:
  - Mantém cada ambiente isolado, com pacotes de versões diferentes
  - Facilita a instalação e configuração de pacotes
  - Aumenta a produtividade no desenvolvimento
  - Muito utilizada por cientistas de dados



# Python - <https://www.python.org>

- Python é uma linguagem de programação que possui as seguintes características:
  - Totalmente livre
  - Código interpretado
  - Multiplataforma
  - Orientada a objetos
  - Sintaxe simples
  - Fácil de aprender
  - Fácil leitura e manutenção
  - Possui modo interativo



# Jupyter Notebook - <http://jupyter.org>

- Interface interativa para o desenvolvimento com linguagens de programação
- Características:
  - Permite a visualização da execução comando a comando e suas respectivas saídas, incluindo gráficos
  - Permite salvar um notebook para uso posterior
  - Permite compartilhar notebooks via git
  - Bancada de testes ideal para cientistas de dados



# Scikit-learn - <http://scikit-learn.org/>

- Uma poderosa ferramenta open-source para análise de dados e machine learning em Python, construída com NumPy, SciPy, and matplotlib
- Funcionalidades:
  - Classificação
  - Regressão
  - Clusterização
  - Redução de dimensionalidade



HANDS ON



# Laboratório 01

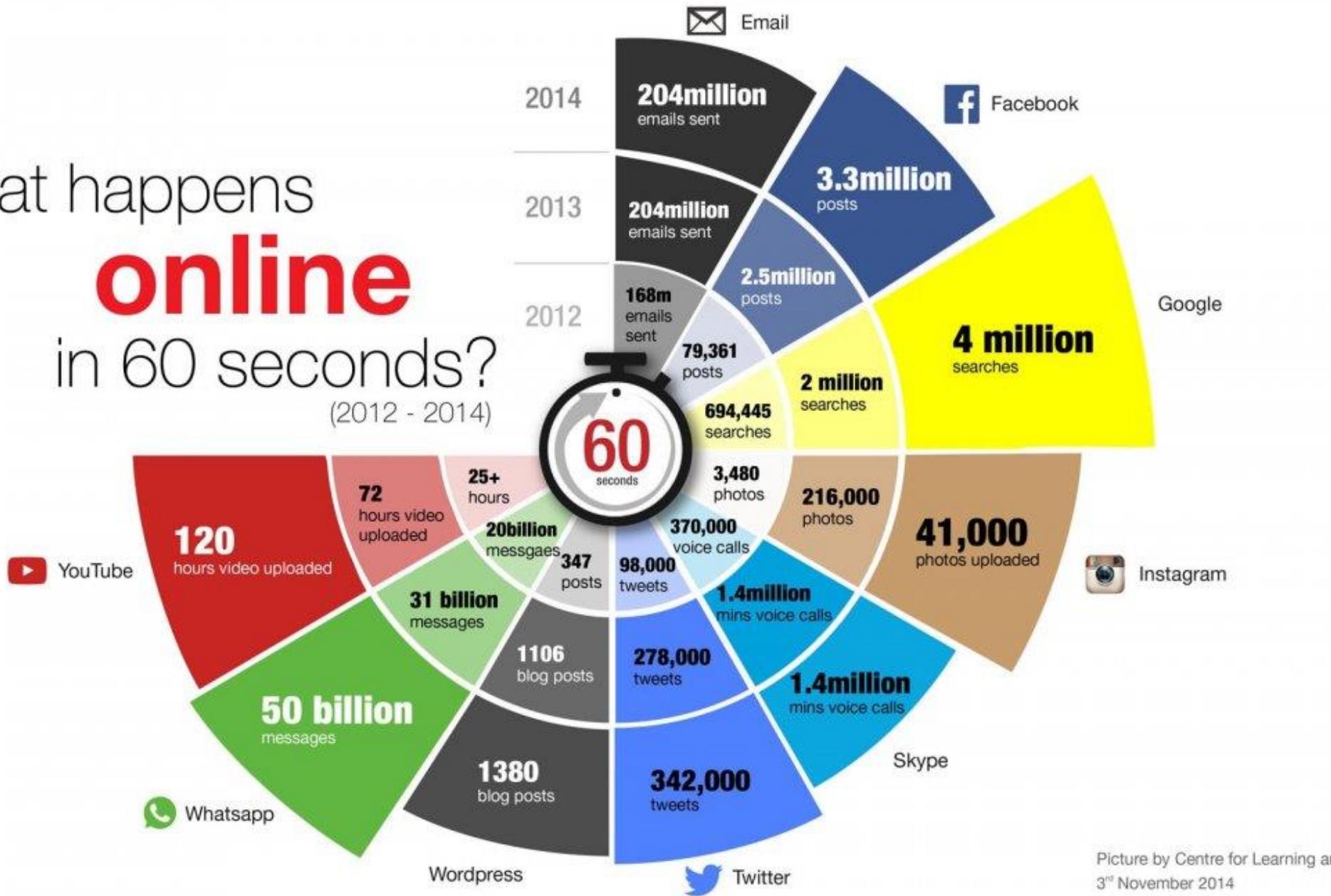
## Anaconda e Jupyter Notebook

## 2. Contextualização



# O que ocorre em 60 segundos...

# What happens **online** in 60 seconds? (2012 - 2014)



2.5 quintilhões de bytes de dados são gerados por dia, os quais podem ser utilizados para nortear as empresas e indivíduos.

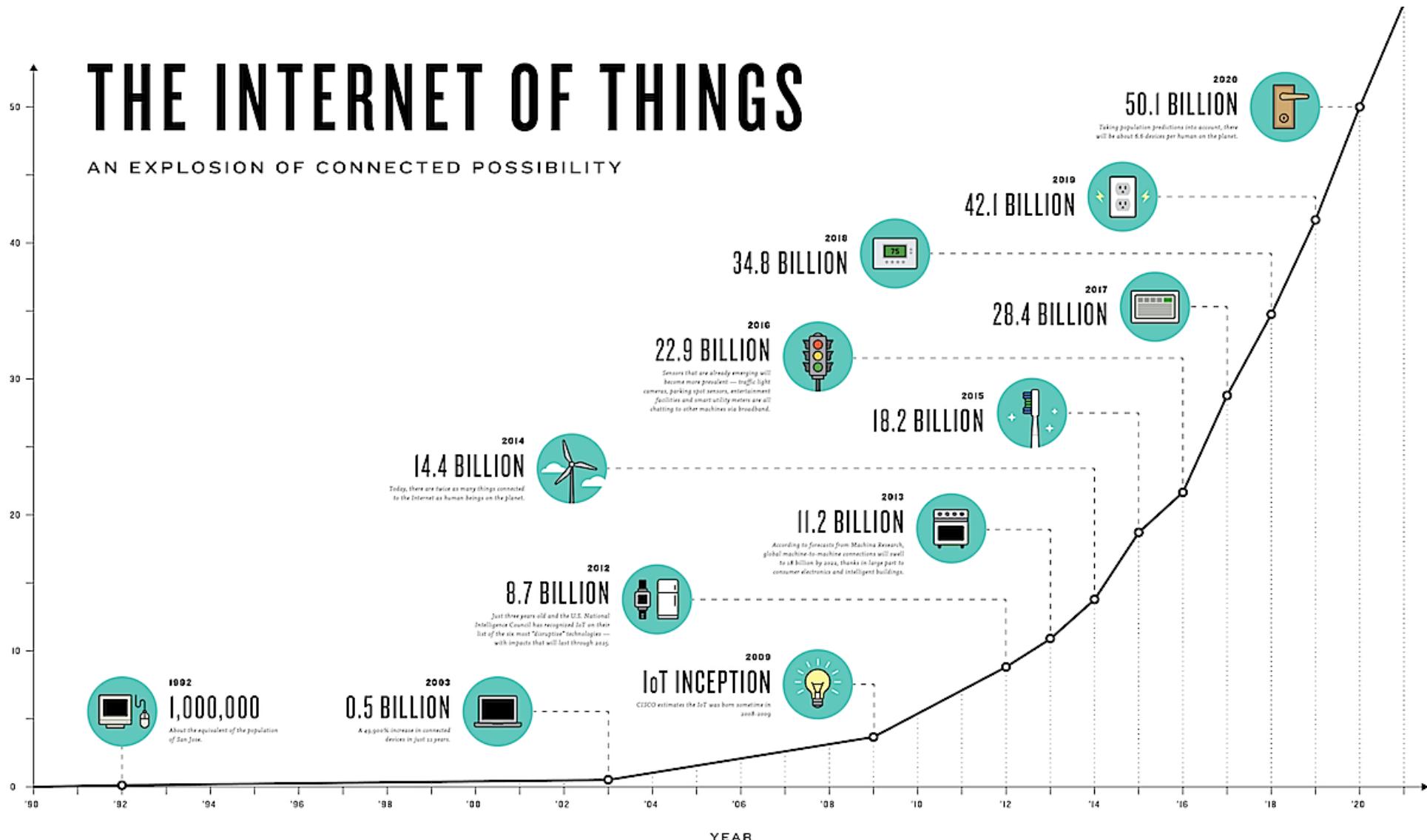
... e está dobrando a cada dois anos!



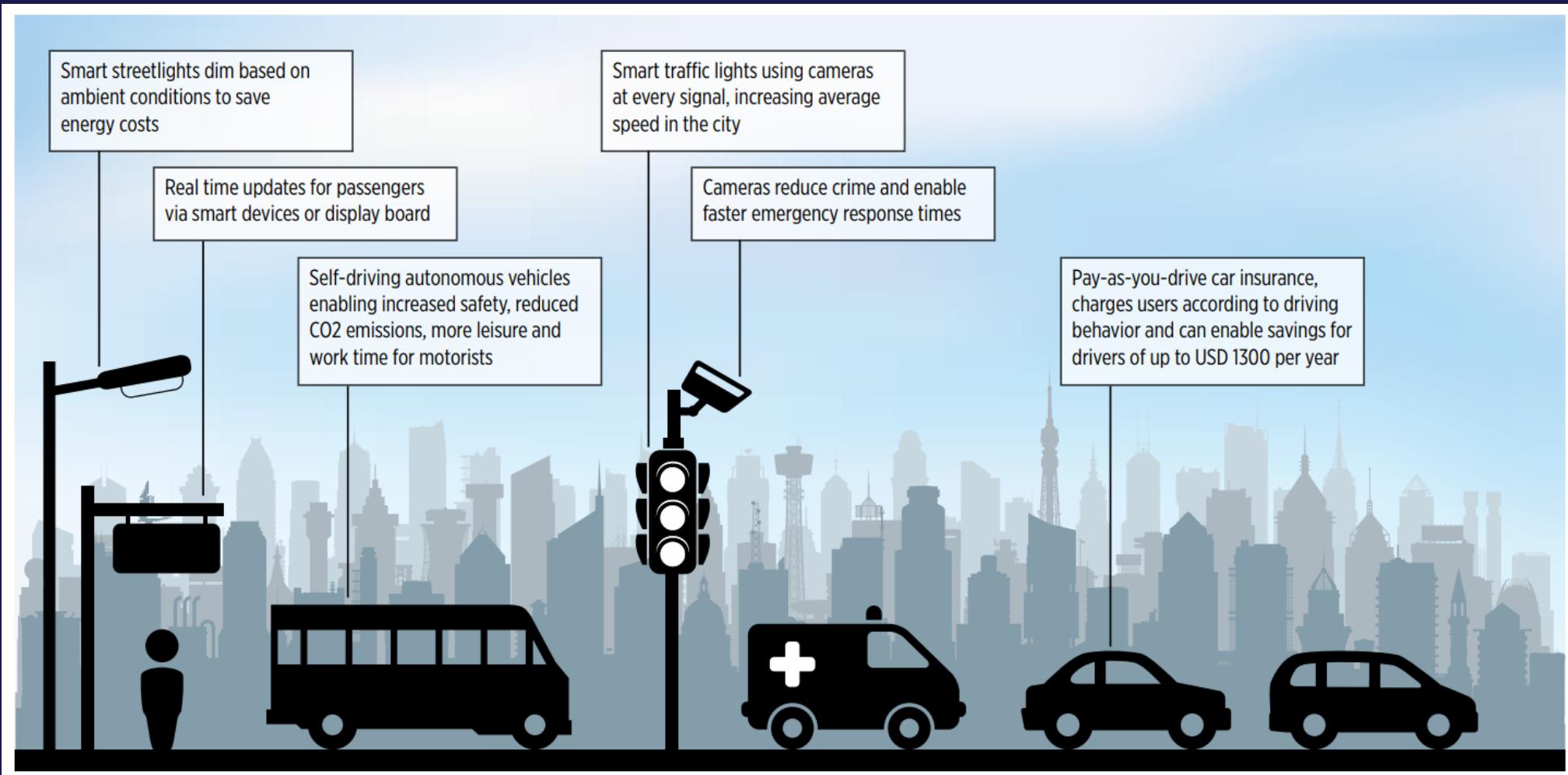
# THE INTERNET OF THINGS

AN EXPLOSION OF CONNECTED POSSIBILITY

BILLIONS OF DEVICES



# Smart Cities (Cidades Inteligentes)



# Smart living



# Smart car



# FUTURE FARMS

## small and smart



### FARMING DATA

The farm generates vast quantities of rich and varied data. This is stored in the cloud. Data can be used as digital evidence reducing time spent completing grant applications or carrying out farm inspections saving on average £5,500 per farm per year.

### SURVEY DRONES

Aerial drones survey the fields, mapping weeds, yield and soil variation. This enables precise application of inputs, mapping spread of pernicious weed blackgrass could increasing Wheat yields by 2-5%.

### FLEET OF AGROBOTS

A herd of specialised agribots tend to crops, weeding, fertilising and harvesting. Robots capable of microdot application of fertiliser reduce fertiliser cost by 99.9%.

### TEXTING COWS

Sensors attached to livestock allowing monitoring of animal health and wellbeing. They can send texts to alert farmers when a cow goes into labour or develops infection increasing herd survival and increasing milk yields by 10%.

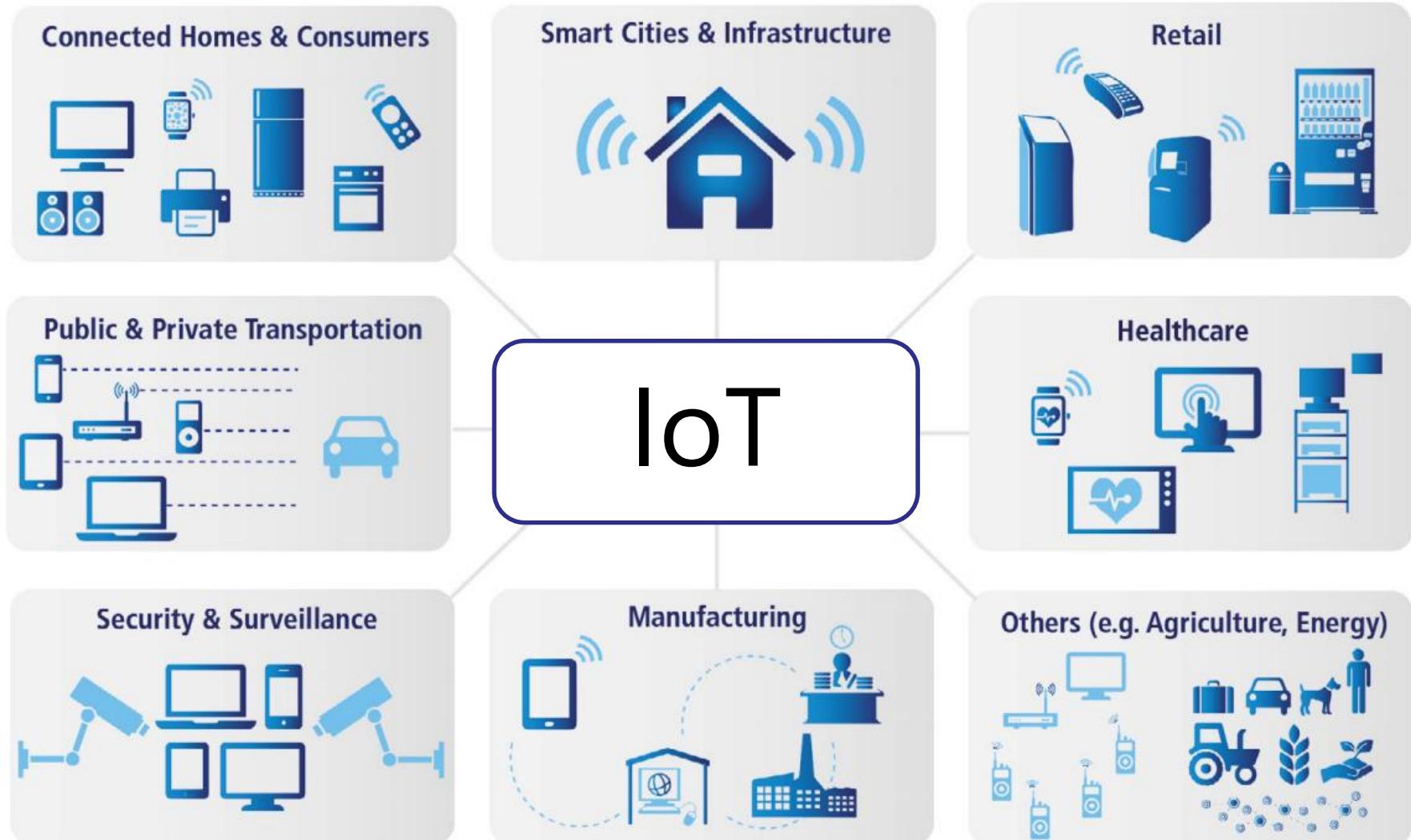
### SMART TRACTORS

GPS controlled steering and optimised route planning reduces soil erosion, saving fuel costs by 10%.

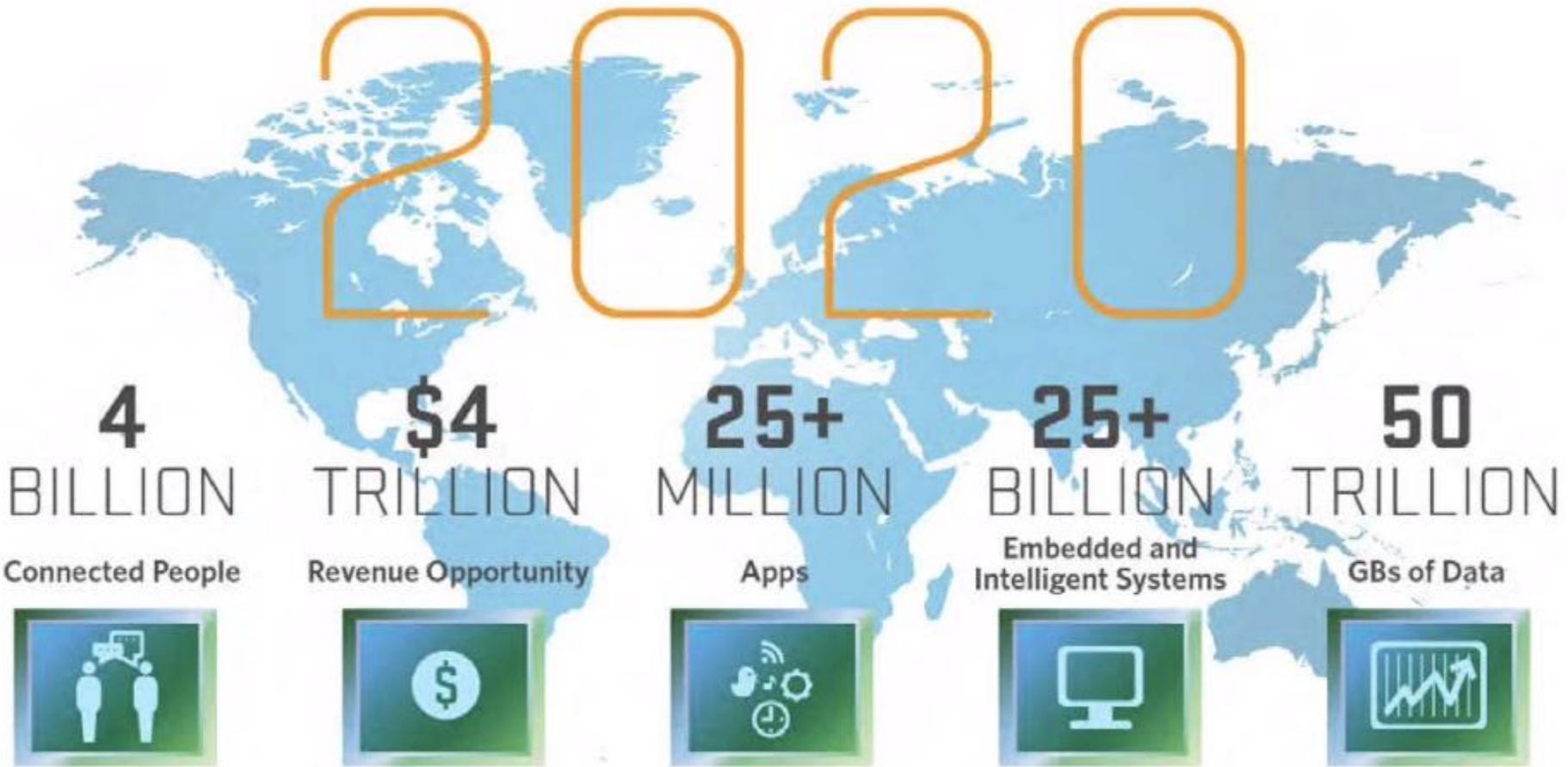


**Big Data não é novidade na  
Indústria aeroespacial**

# Abrangência do IoT



# Qual o tamanho do Big Data?



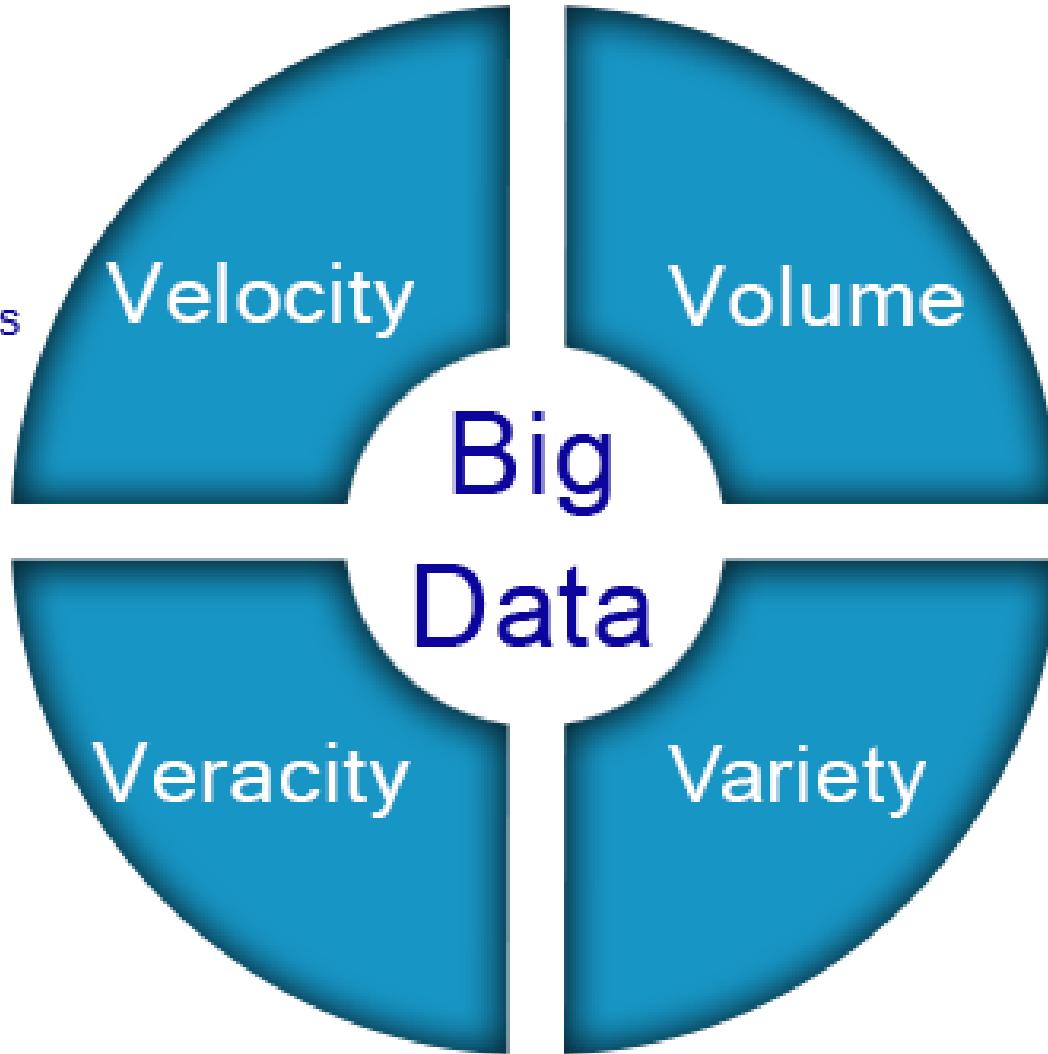
# O que é Big Data

Conjuntos de dados extremamente amplos e que necessitam de ferramentas específicas para lidar com grandes volumes, de forma que toda e qualquer informação nestes meios possa ser encontrada, analisada e aproveitada em tempo hábil

Volume x Tempo

# 4 V's do Big Data

- Batch
- Near Time
- Real Time
- Live Stream
- Rate of Analysis



- Authenticity
- Availability
- Accountability
- Trustworthy
- Origin
- Reputation
- Cleansed

- Terabytes
- Records
- Transaction
- Tables
- Files

- Unstructured
- Semi-structured
- Structured



Como resolver  
esse dilema?

Quero ser capaz  
de analisar essa  
quantidade de  
dados, mas  
como?

# data science

A word cloud centered around the words "data" and "science". The word "data" is on the left in large red font, and "science" is on the right in large green font. Various other words are scattered in between, such as "research", "faculty", "support", "activities", "institute", etc., in different colors like yellow, blue, and purple.

“Data scientist is  
**the sexiest job**,  
of the 21st century.”

Harvard Business Review



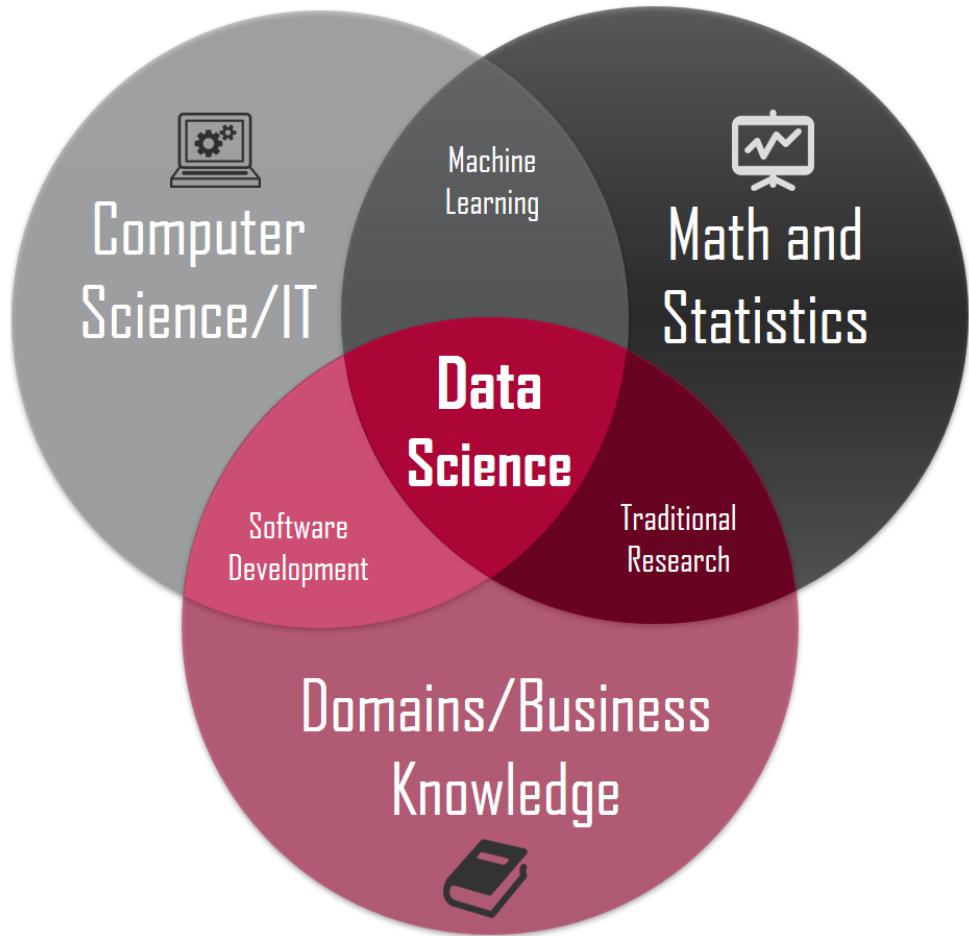
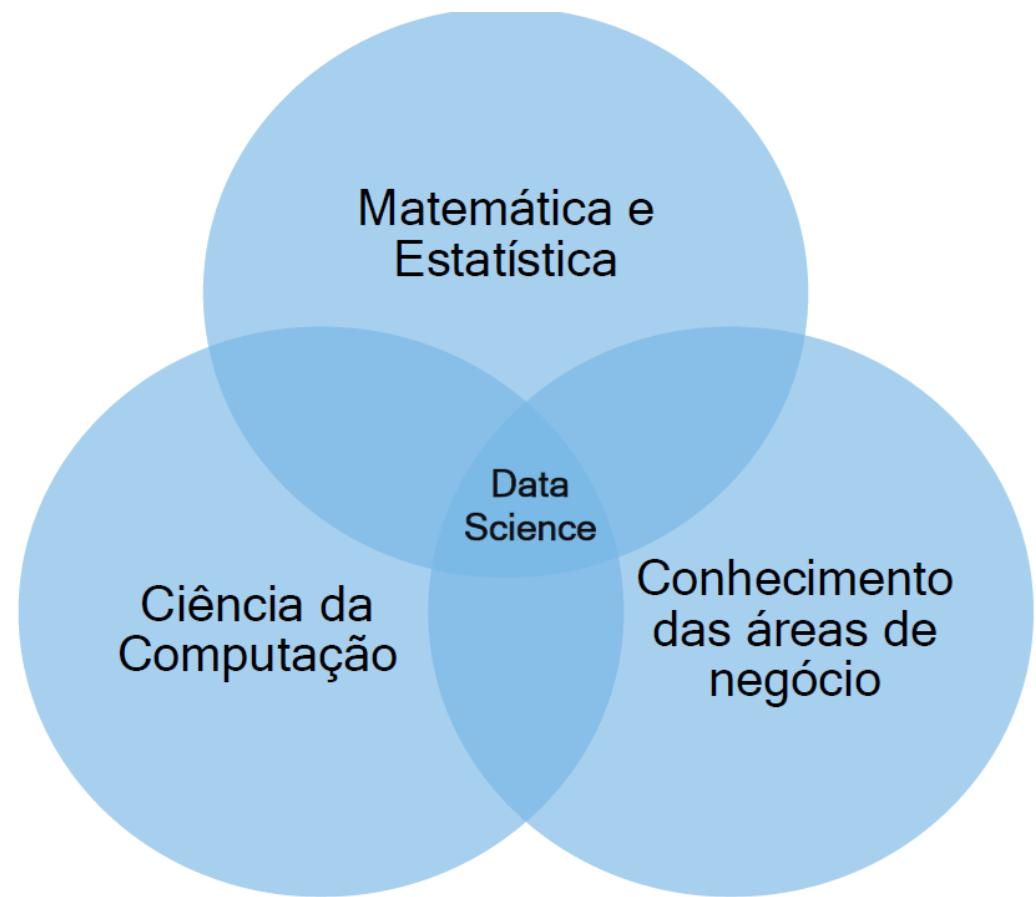
# Áreas de Conhecimento

A Ciência de Dados envolve o uso de métodos automatizados através de ferramentas / técnicas / frameworks (ciência da computação)

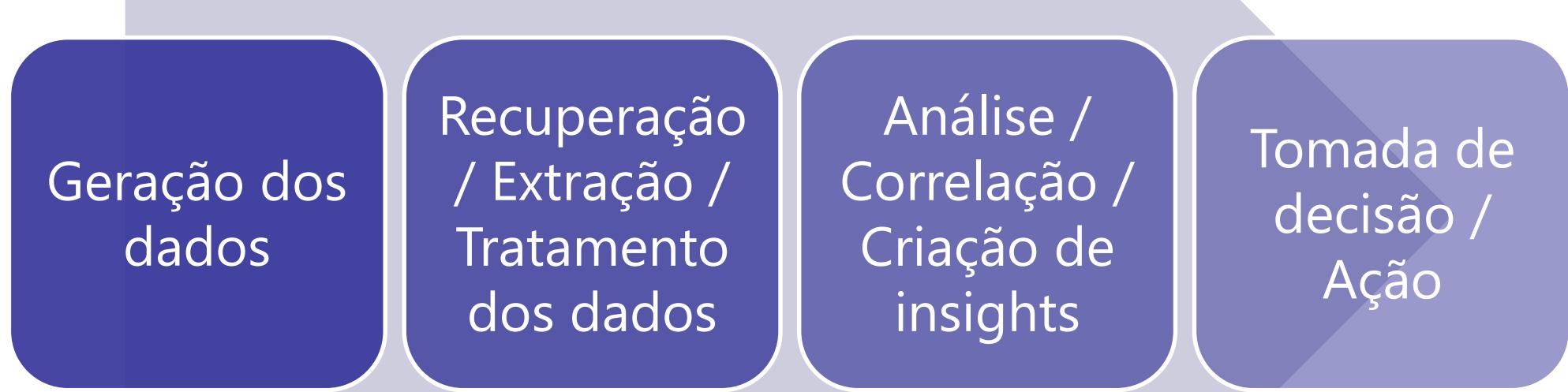
...para agrupar, organizar e analisar enormes quantidades de dados estruturados ou não (estatística)

...para extrair conhecimento, desenvolver compreensão, formular ações e insights relevantes para uma empresa / cliente / setor (áreas de negócio) e que gerem resultados.

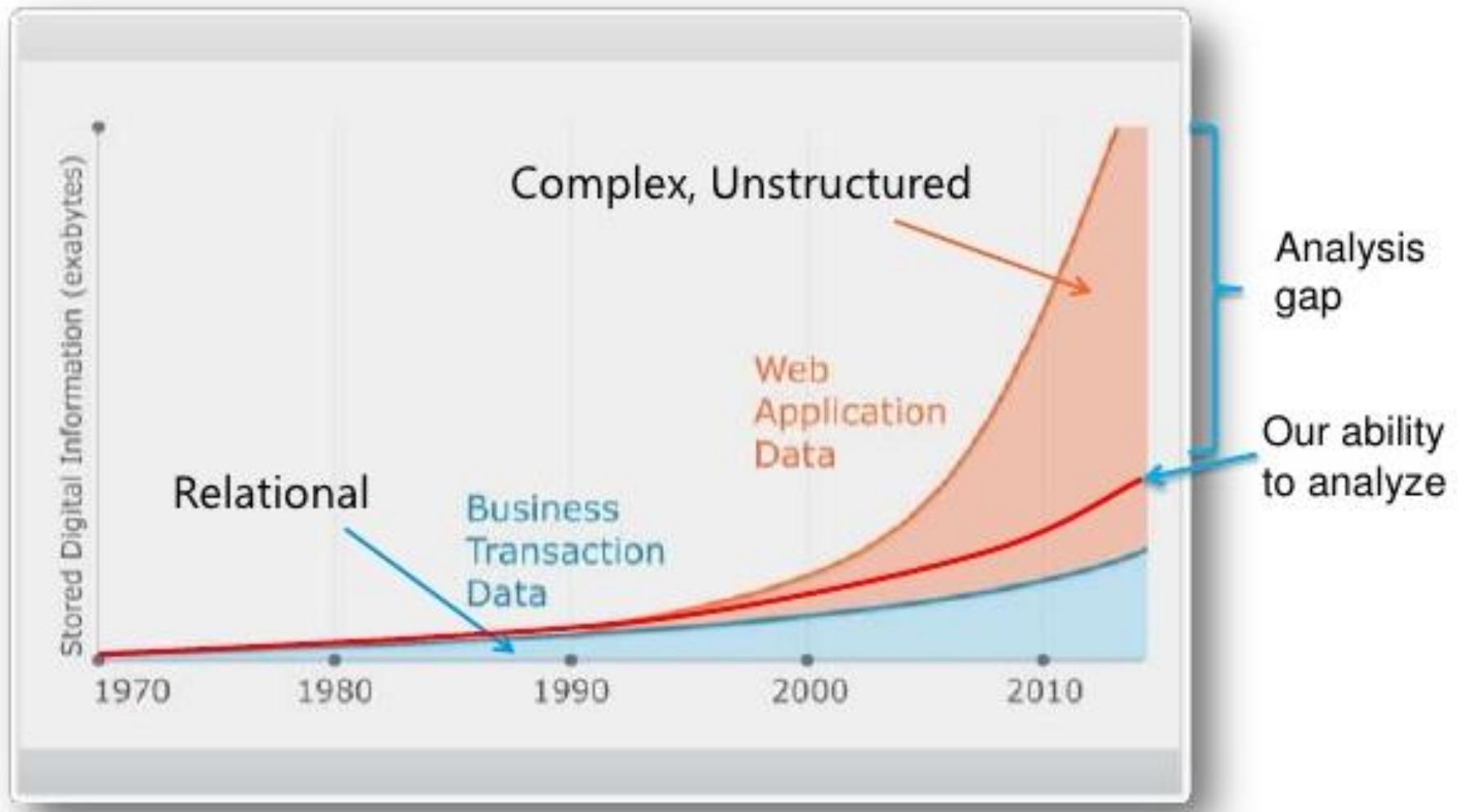
# Áreas de Conhecimento



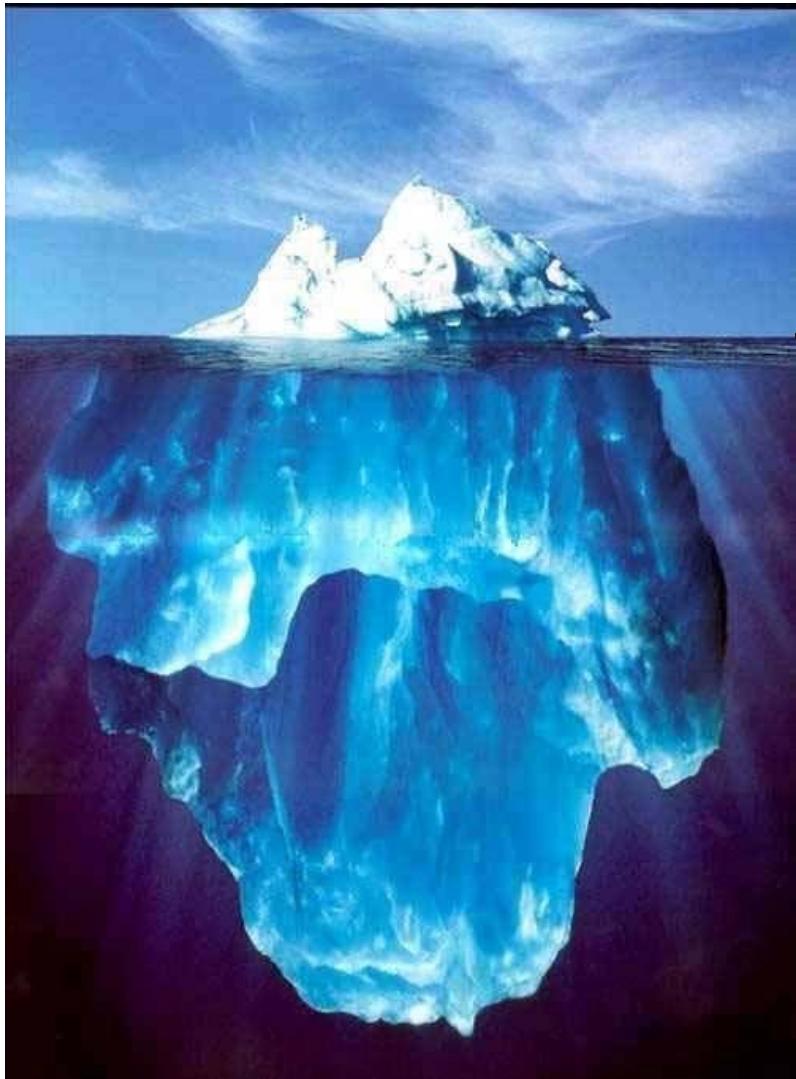
# Processo de Data Science



# Estruturado x Não estruturado



# A oportunidade do BigData



## ➤ Dados Internos

- Estruturados, previsíveis, permanentes, fáceis de obter

## ➤ Dados Externos

- Não-estruturados, randômicos, voláteis, difíceis de obter
- Aqui entra o BigData!

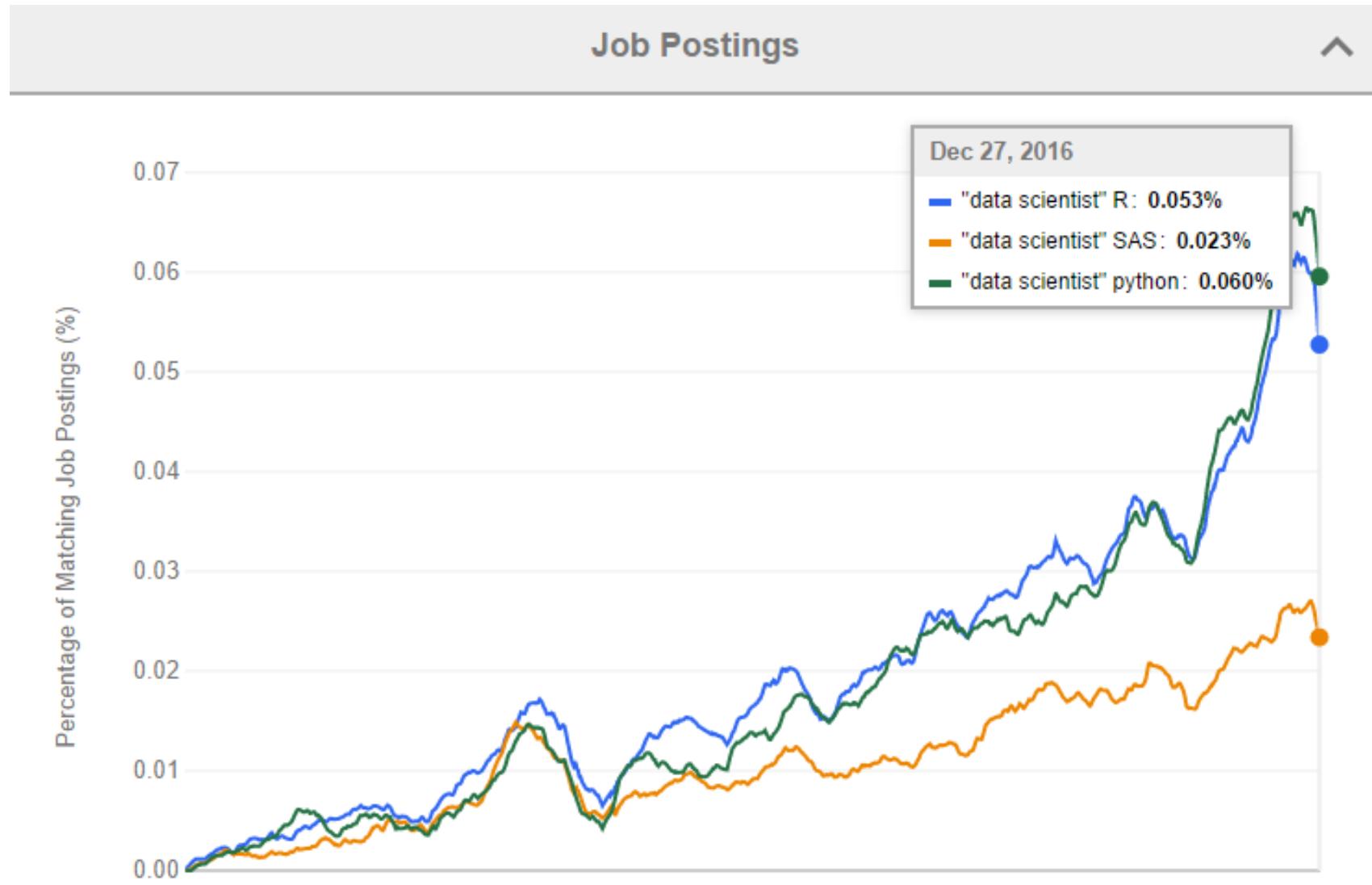
# Evolução do Data Science



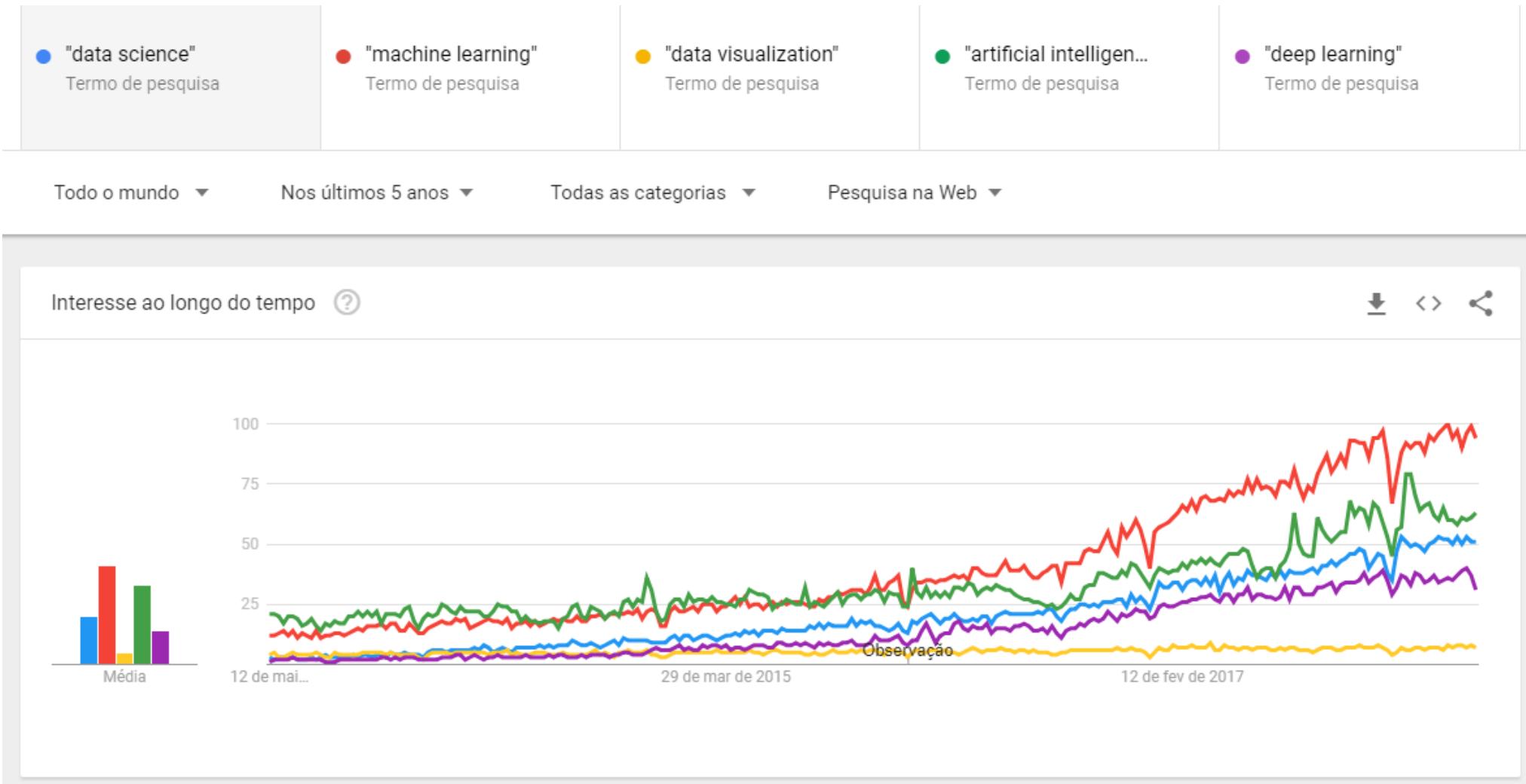
# Evolução do Data Science



# Linguagens mais utilizadas



# Assuntos mais buscados nos últimos 5 anos



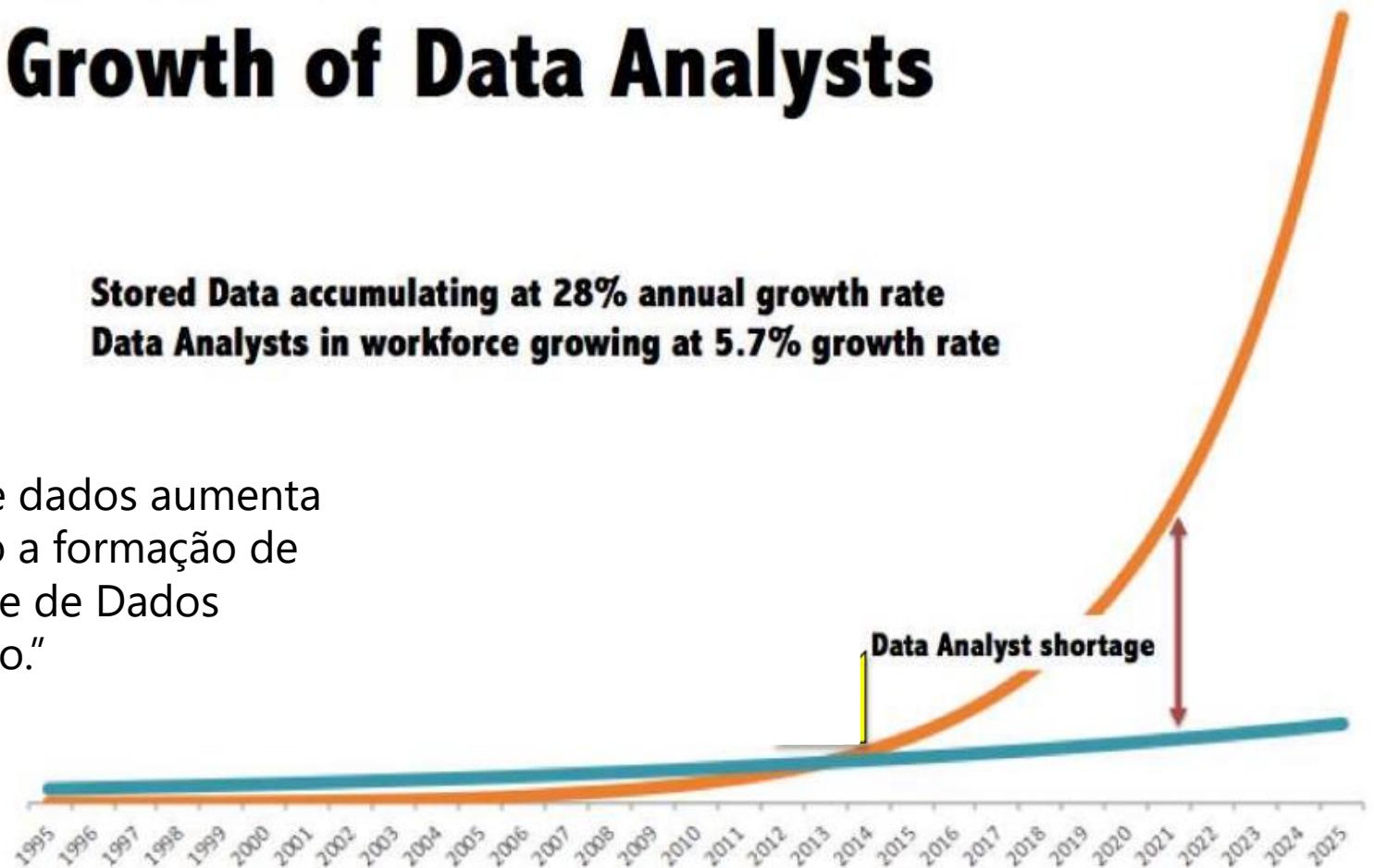
<https://trends.google.com.br/trends/explore?date=today%205-y&q=%22data%20science%22,%22machine%20learning%22,%22data%20visualization%22,%22artificial%20intelligence%22,%22deep%20learning%22>

# Crescimento dos dados x profissionais

## Growth of Data vs. Growth of Data Analysts

**Stored Data accumulating at 28% annual growth rate  
Data Analysts in workforce growing at 5.7% growth rate**

"O armazenamento de dados aumenta 28% ao ano, enquanto a formação de profissionais de Análise de Dados aumenta a 5,7% ao ano."



# Conhecimentos desejados

Área de Conhecimento	Habilidade
Matemática e Estatística	Álgebra Linear, Estatística Descritiva, Testes de Hipótese, Análise Bayesiana
Aprendizado de Máquina	Aprendizagem Supervisionada e Não-Supervisionada, Classificação, Regressão, Clustering
Programação	Python, R, Scala, Java, Julia, SAS, SQL, C++
Banco de Dados	Bancos Relacionais e Bancos No-SQL como MongoDB
Filtragem e Visualização de Dados	D3.js, Tableau, Infovis, ggplot2
Big Data	Hadoop (HDFS / MapReduce), Apache Spark, Storm, Cassandra
Área de Negócio	Finanças, Marketing, Varejo, Astronomia, Saúde, Tecnologia

# 3. Introdução à linguagem Python

# Python

- Python é uma linguagem de programação de uso geral, mas que tem liderado as iniciativas de Análise de Dados por diversos motivos:
  - Comunidade ampla
  - Facilidade de Aprender
  - Facilidade de instalar vários ambientes com o Anaconda
  - Possui excelentes bibliotecas de Análise de Dados
  - Maioria dos frameworks de Machine Learning suportam Python
  - Vários editores suportam Python (PyCharm, Spyder, Python Tools for Visual Studio, Rodeo, Eclipse / PyDev, Wing IDE, IDLE, Notepad, Sublime, Atom, Vim)
  - Facilidade de pesquisar e trabalhar com o Jupyter Notebook

# Como executar Python

➤ Basicamente, existem 3 modos de executar o Python:

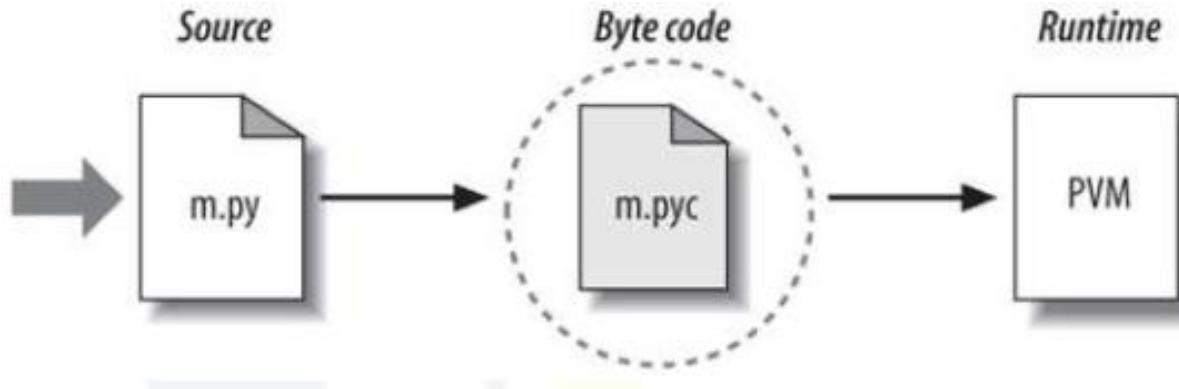
Modo shell

Modo script (arquivos com extensão .py)

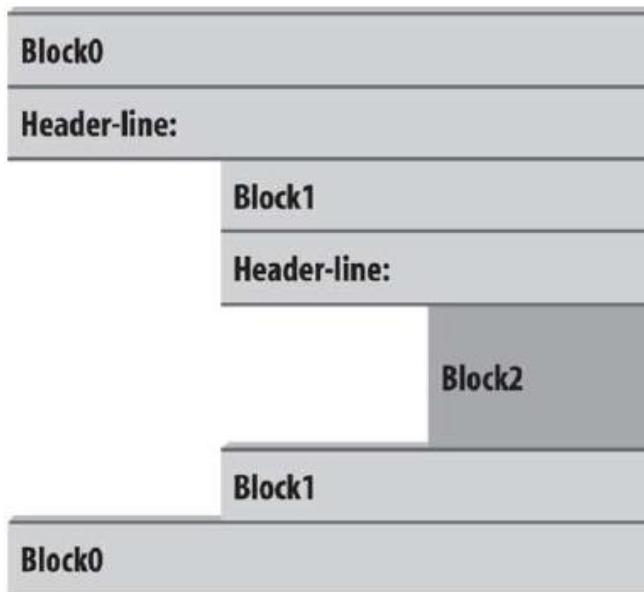
Modo interativo (Jupyter Notebook)

# Características importantes

- Linguagem interpretada, mas híbrida



- Controle de blocos é feito por identação → 1 tab ou 4 espaços



# Comentários

- Começam com o caractere # ou 3 aspas duplas “””...”””

# *Isso é um comentário em um única linha*

”””

*Isso é um comentário  
em mais de uma linha*

”””

# Tipos

- Numéricos: int, float
- Booleano: assume os tipos
- Operadores relacionais:

Operador	Significado
==	Igualdade / equivalência
!=	Desigualdade / Inequivalência
>	Maior que
<	Menor que
>=	Maior que ou igual a
<=	Menor que ou igual a

- Descobrir o tipo com type()
- Variáveis são fracamente tipadas
- A função print() imprime na tela
- A indexação de strings e arrays em python começa com zero “0”

`texto = "Python e Análise de Dados"`

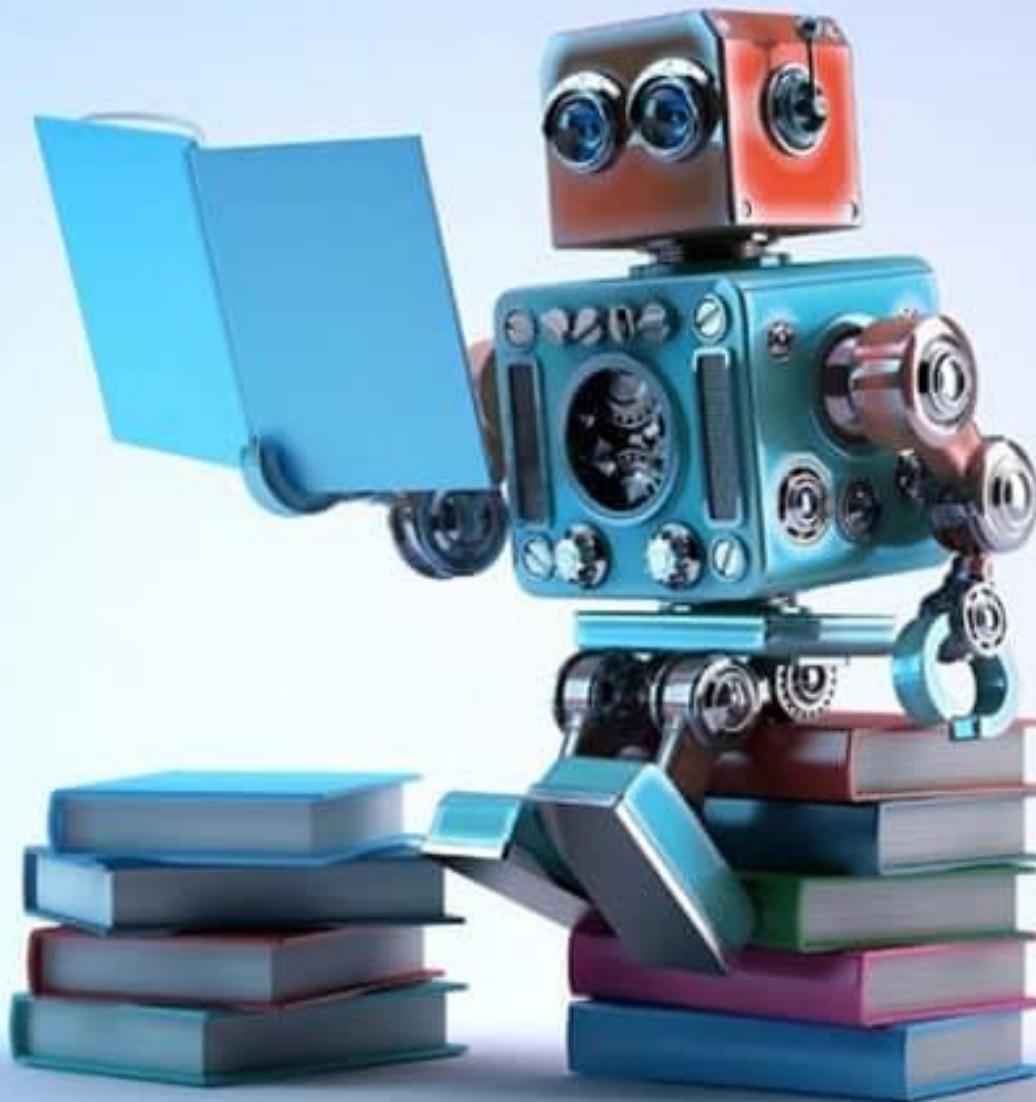
`texto[0] = P`



- 01 Python.ipynb**
- 02 Pandas.ipynb**

# 4. Machine Learning

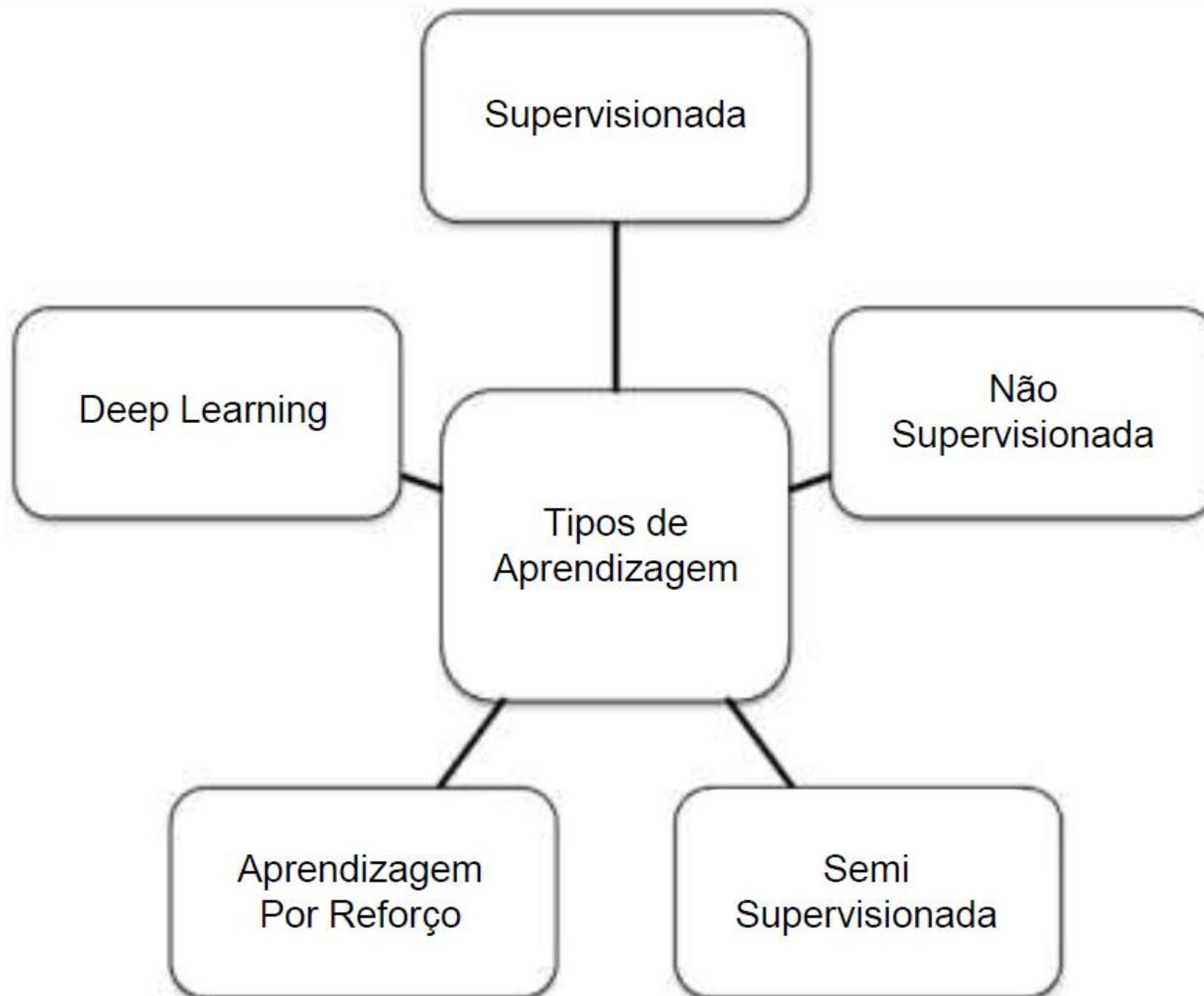
# Machine Learning



# Aprendizado de Máquina (Machine Learning)

- Um subcampo da Ciência da Computação que evoluiu do estudo de reconhecimento de padrões e da teoria do aprendizado computacional em inteligência artificial.
- Ele nasceu do reconhecimento de padrões e da teoria de que máquinas podem aprender sem serem programadas para realizar tarefas específicas; pesquisadores interessados em inteligência artificial queriam saber se computadores podem aprender com dados.
- Usando algoritmos que aprendem a partir de dados, o aprendizado de máquina permite que os computadores encontrem insights ocultos sem que sejam explicitamente programados para procurar algo específico.

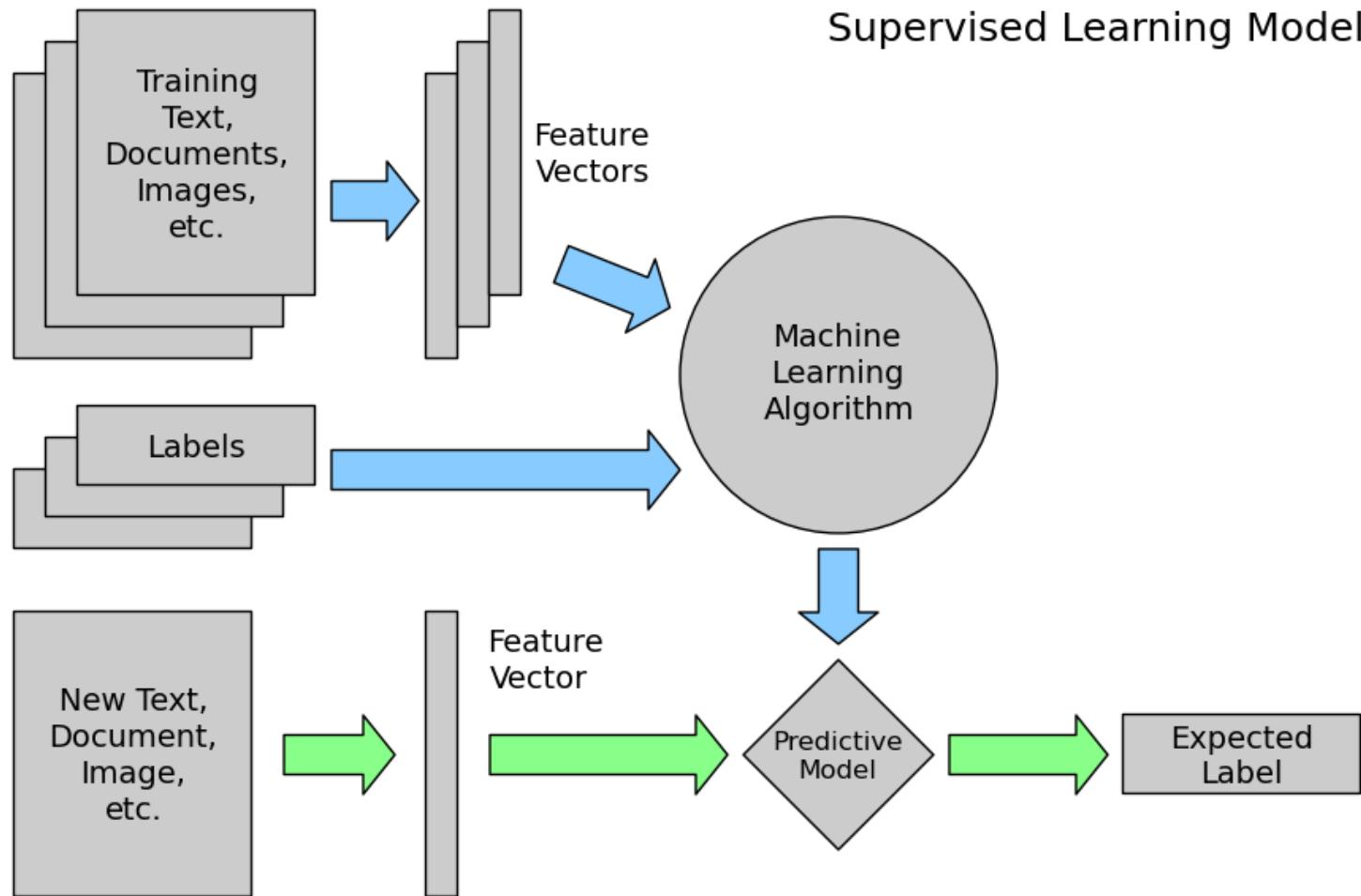
# Tipos de aprendizagem



# Aprendizagem supervisionada

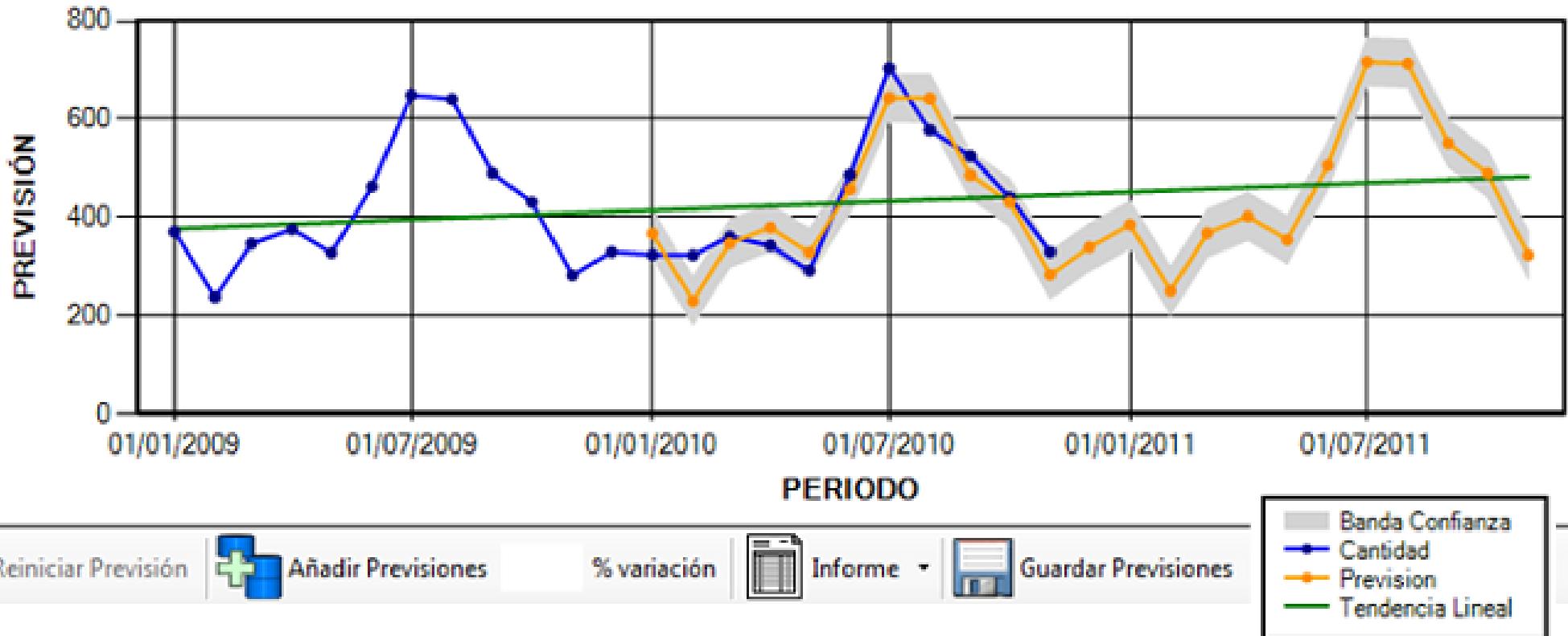
- Normalmente usado em tarefas de **classificação e regressão**.
- Os dados de treino são totalmente rotulados (a saída desejada é conhecida) e orientam o aprendizado da máquina.
- Neste tipo de aprendizagem existe um "professor" que avalia a resposta da rede/algoritmo ao padrão atual de inputs.
- É o método de aprendizagem mais utilizado hoje em dia, pois é possível verificar o desempenho da rede/algoritmo facilmente.
- Ex: Previsão de receita, classificação de risco.

# Aprendizagem supervisionada

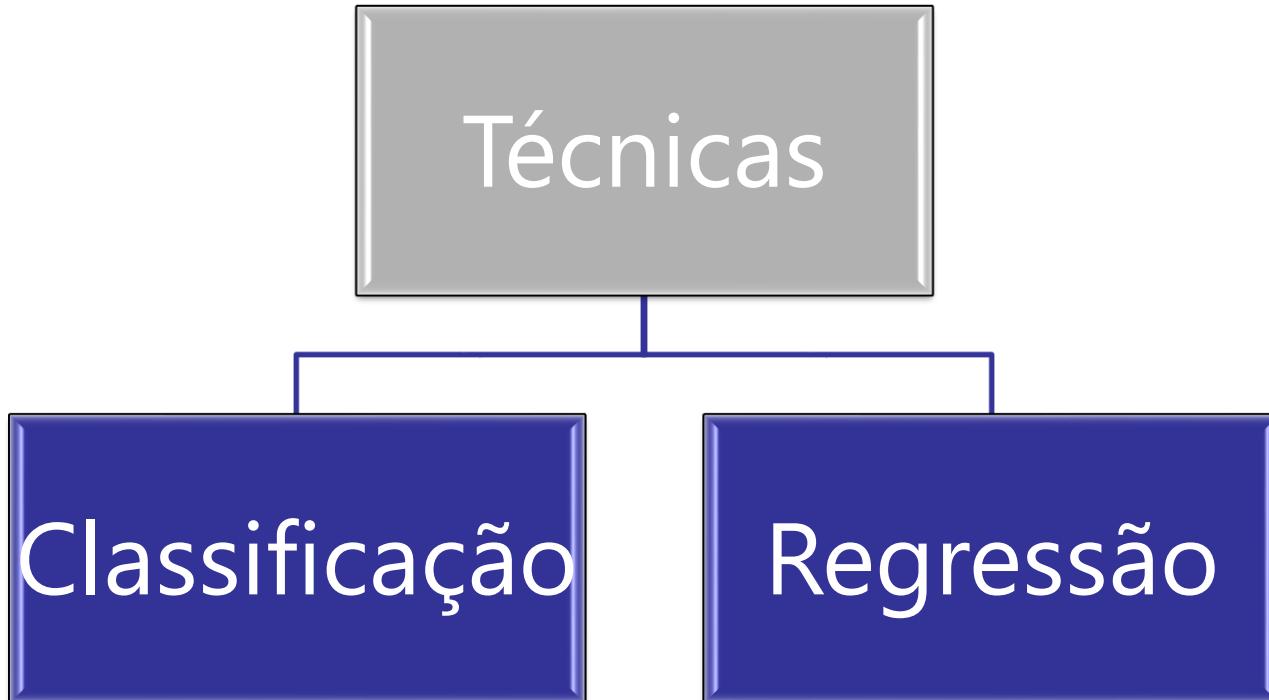


No aprendizado supervisionado, todas os exemplos de treinamento são rotulados.

# Aprendizagem supervisionada



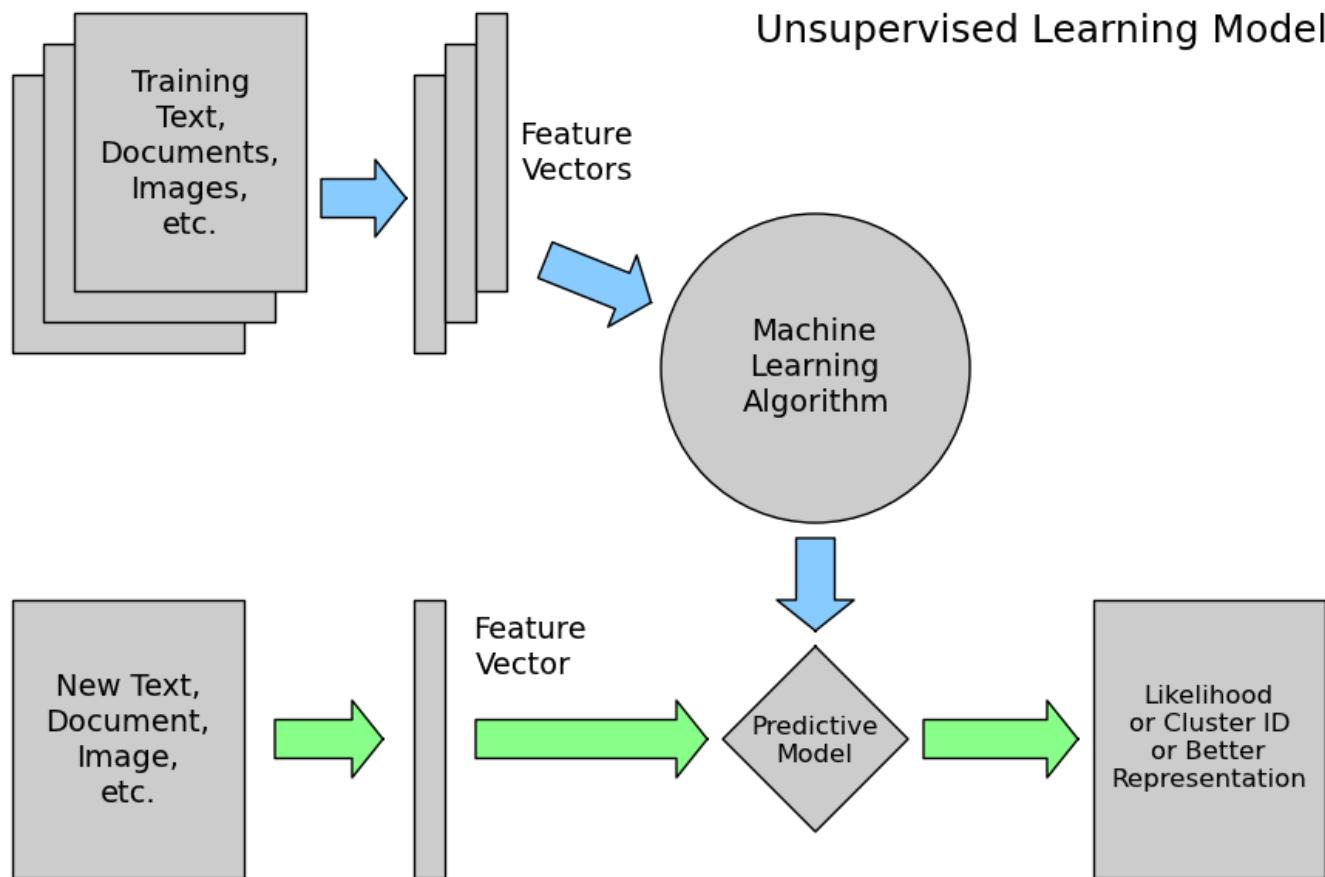
# Aprendizagem supervisionada



# Aprendizagem não supervisionada

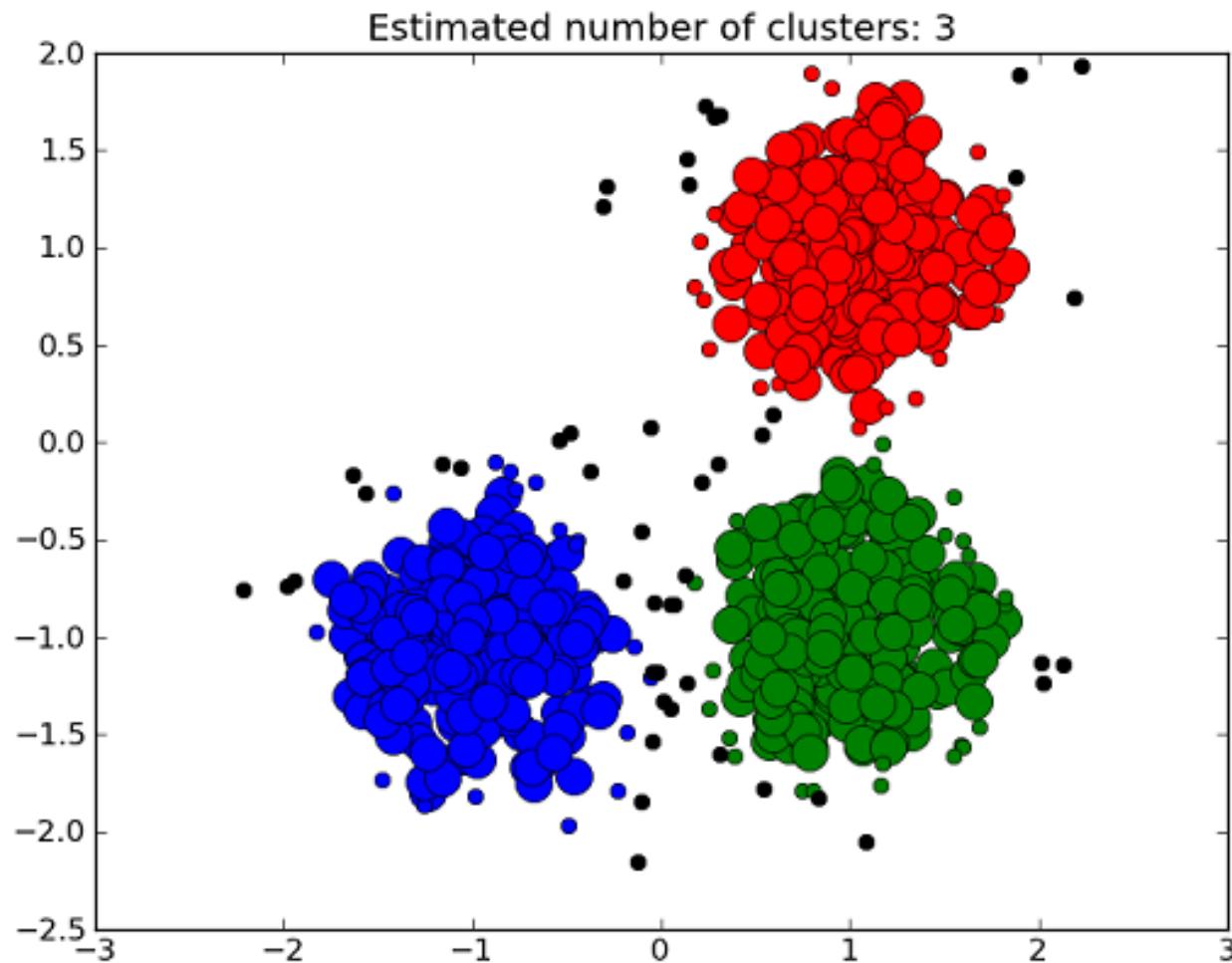
- Frequentemente utiliza-se este método de aprendizado em análises exploratórias.
- Nesta forma de aprendizagem não existe "professor". A rede ou o algoritmo tem de descobrir sozinha relações, padrões, regularidades ou categorias nos dados que lhe vão sendo apresentados e codificá-las nas saídas.
- Normalmente seu resultado é o agrupamento em cluster por similaridade.
- Ex: Sistemas de recomendação.

# Aprendizagem não supervisionada

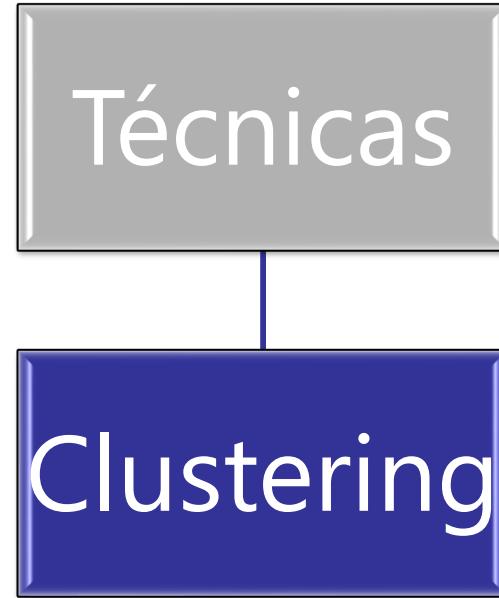


Podemos utilizar grandes quantidades de dados não rotulados para encontrar padrões existentes nestes dados (agrupamentos).

# Aprendizagem não supervisionada



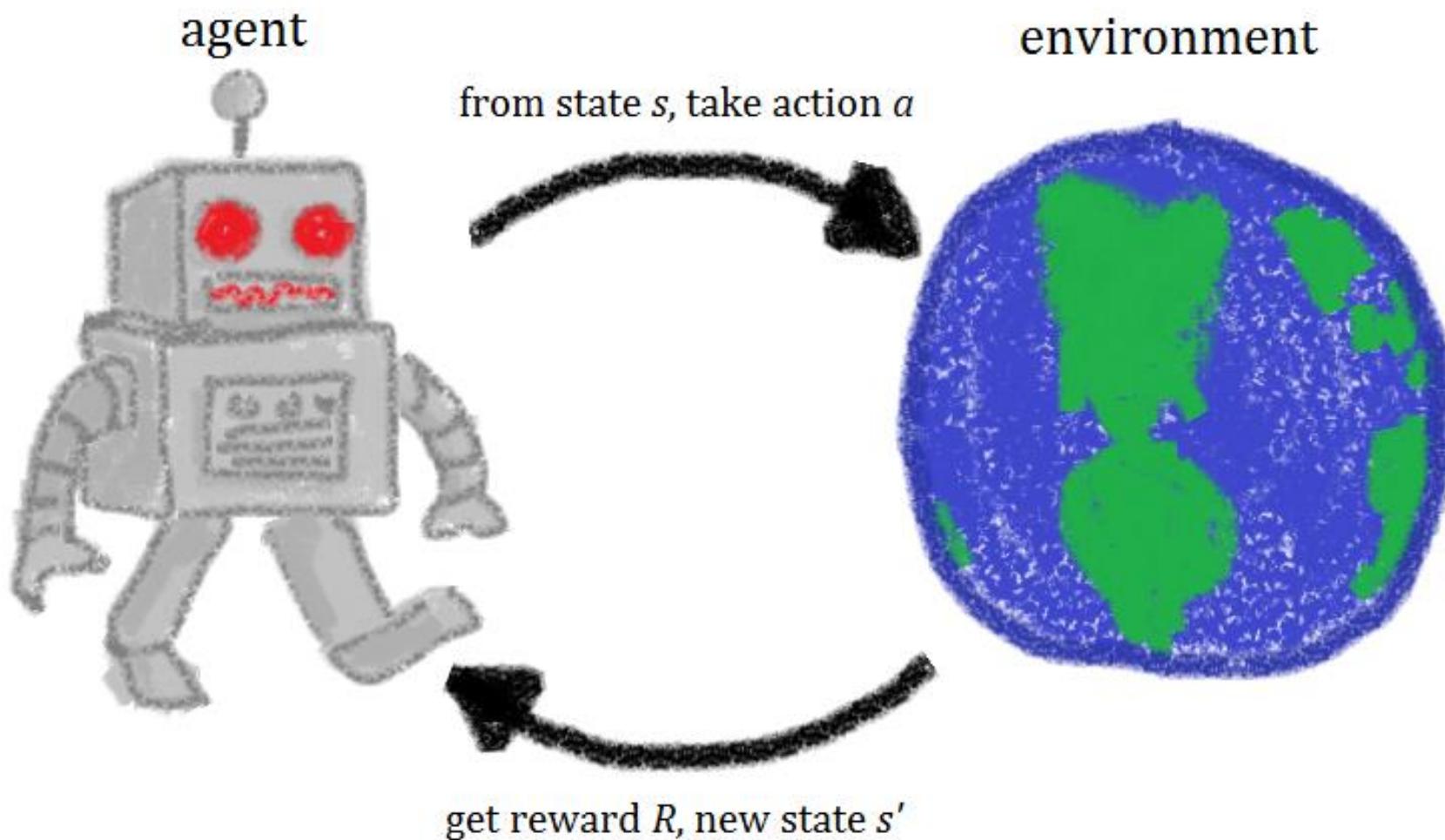
# Aprendizagem não supervisionada



# Aprendizagem semi supervisionada

- Neste tipo o conjunto dos dados disponíveis para treinamento são formados por uma parte rotulada, em menor número, e outra não rotulada, em maior número.
- A ideia é rotular o conjunto de dados não-rotulados a partir daqueles que já possuem rótulos. Desta forma obter dados suficiente para induzir um classificador de forma eficiente.

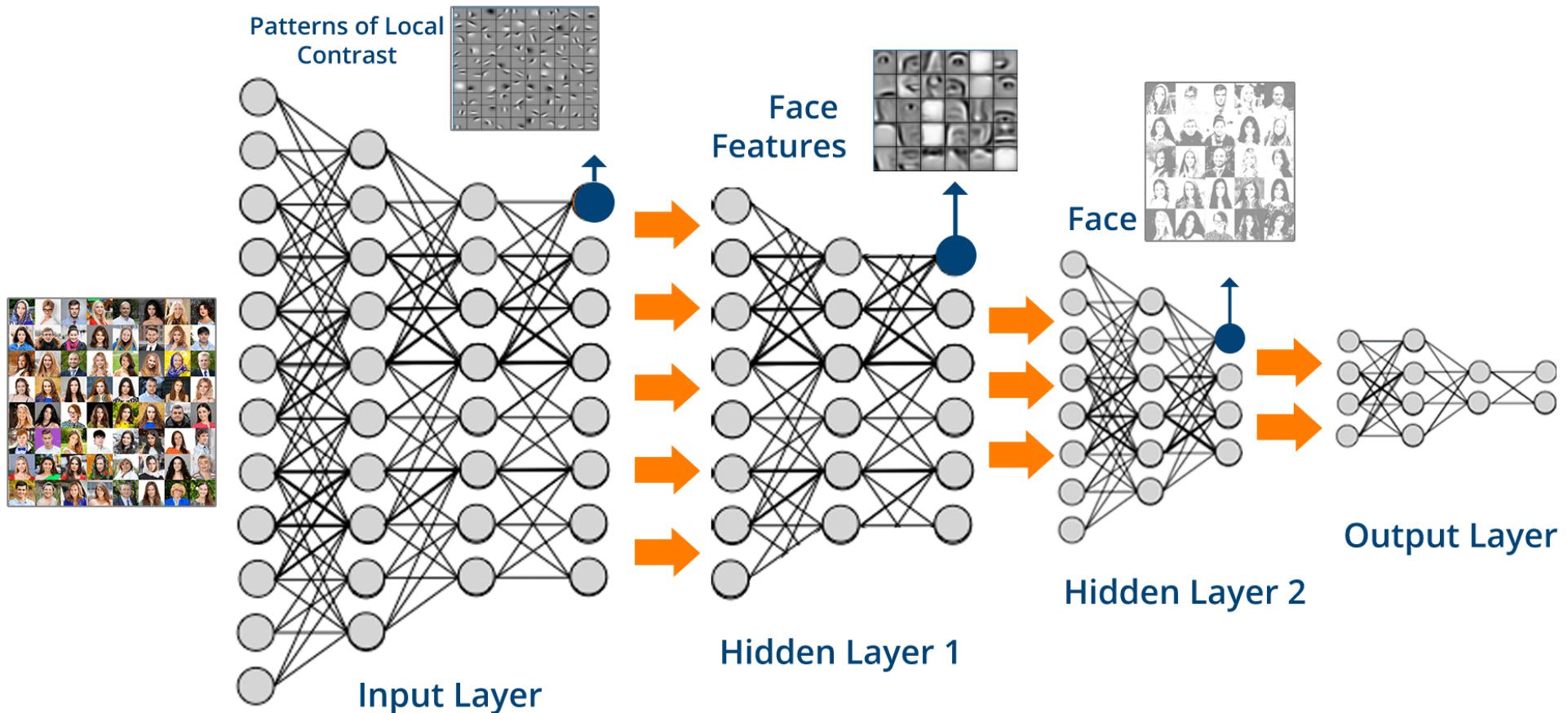
# Aprendizagem por reforço



# Aprendizagem por reforço

- Resumem-se a situações onde um agente inteligente deve agir com base na observação do ambiente no qual se encontra.
- A ação do agente afeta o ambiente resultando em uma recompensa e uma nova observação do ambiente.
- Porém o efeito de uma ação é estocástico, ou seja, é possível que ao tomar duas vezes a mesma ação no mesmo momento os efeitos sejam diferentes.

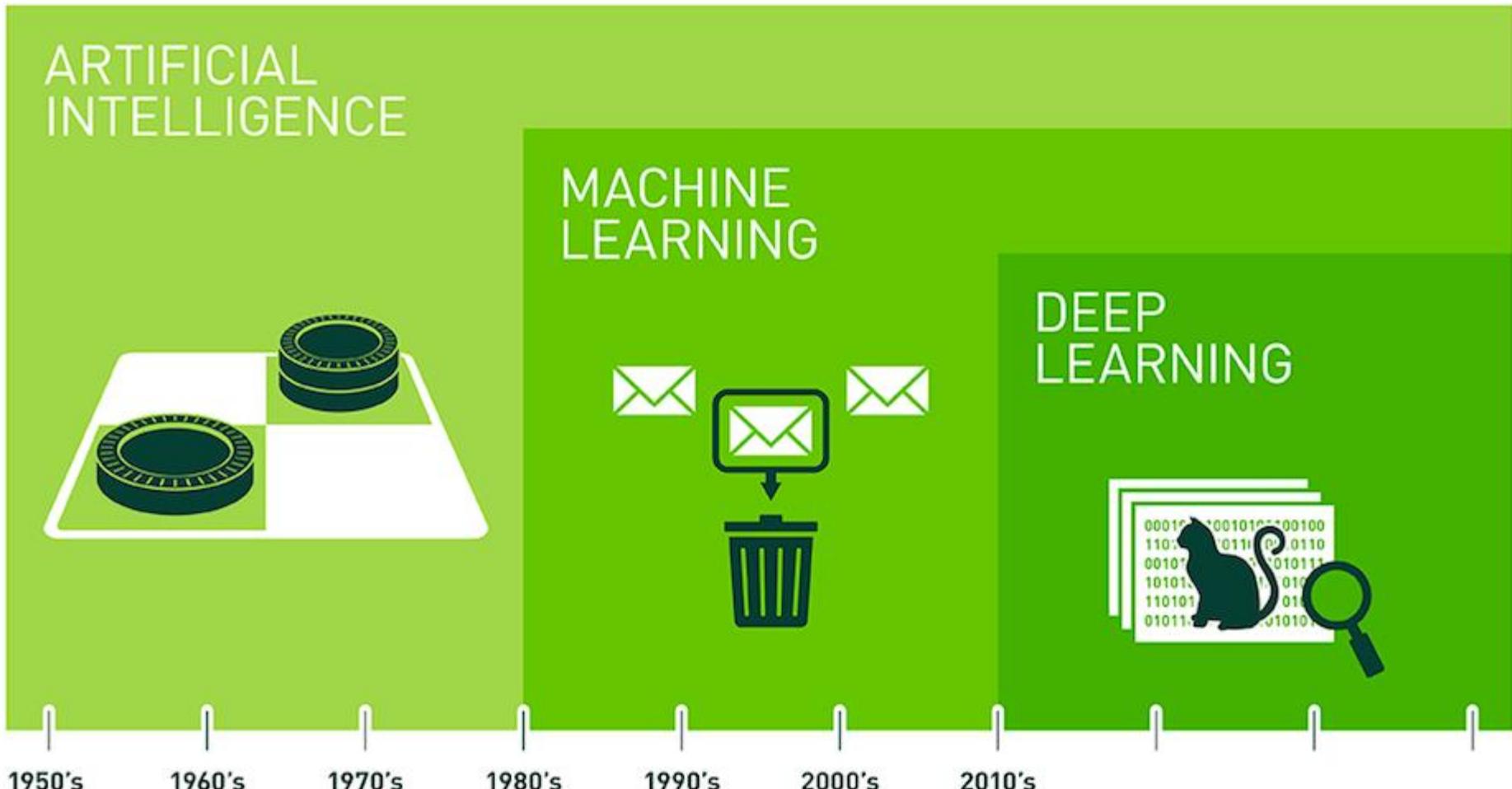
# Aprendizagem profunda (*Deep Learning*)



# Aprendizagem profunda (*Deep Learning*)

- É uma parte importante da Inteligência Artificial, uma subcategoria de aprendizado de máquina ou Machine Learning, que trata as oportunidades de aprendizagem profundas com o uso de redes neurais que possuem várias camadas de neurônios (*deep network*)
- É utilizada para reconhecimento de fala, visão computacional, classificação de imagens, tradução de textos, conversação e processamento de linguagem natural.
- Vem se tornando, rapidamente, um dos mais procurados e estudados dentro da ciência da computação moderna.

# Machine Learning



# Exemplos de aplicações

Reconhecimento  
de Voz

Previsão de  
Doenças

Diagnóstico de  
Câncer

Carros Autônomos

Detecção de  
Fraudes em  
Cartões de Crédito

Previsão de falhas  
em equipamento

Análise de  
sentimento  
baseada em texto

Filtragem de  
spams no e-mail

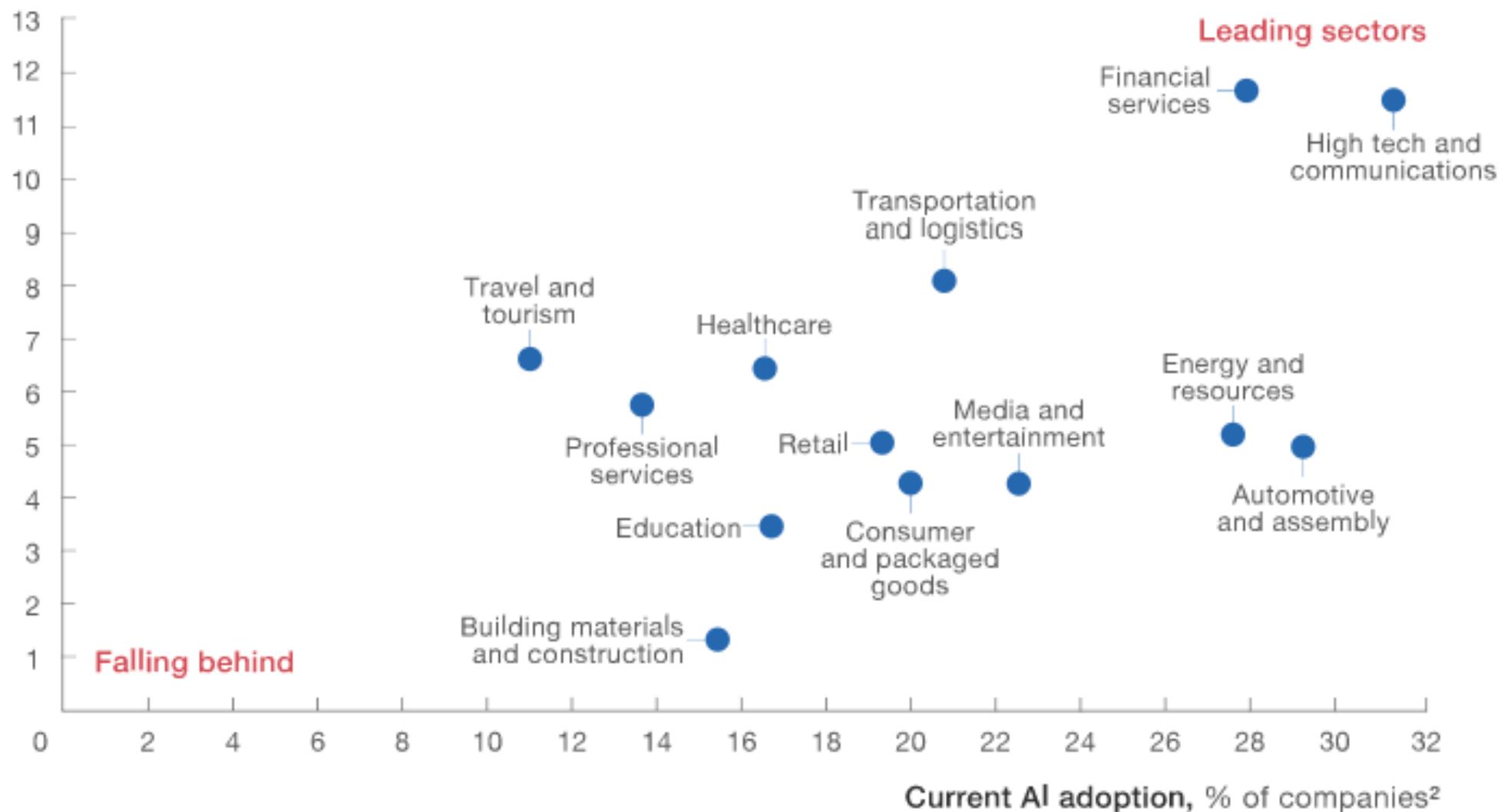
Detecção de  
invasão na rede

Previsão de valores  
imobiliários

Anúncios em  
tempo real

Reconhecimento  
de ameaças e  
crime por imagem

## Future AI demand trajectory, % change in AI spending over next 3 years<sup>1</sup>



<sup>1</sup>Estimated average, weighted by company size; demand trajectory based on midpoint of range selected by survey respondent.

<sup>2</sup>Adopting 1 or more AI technologies at scale or in business core; weighted by company size.

Source: McKinsey Global Institute AI adoption and use survey; McKinsey Global Institute analysis

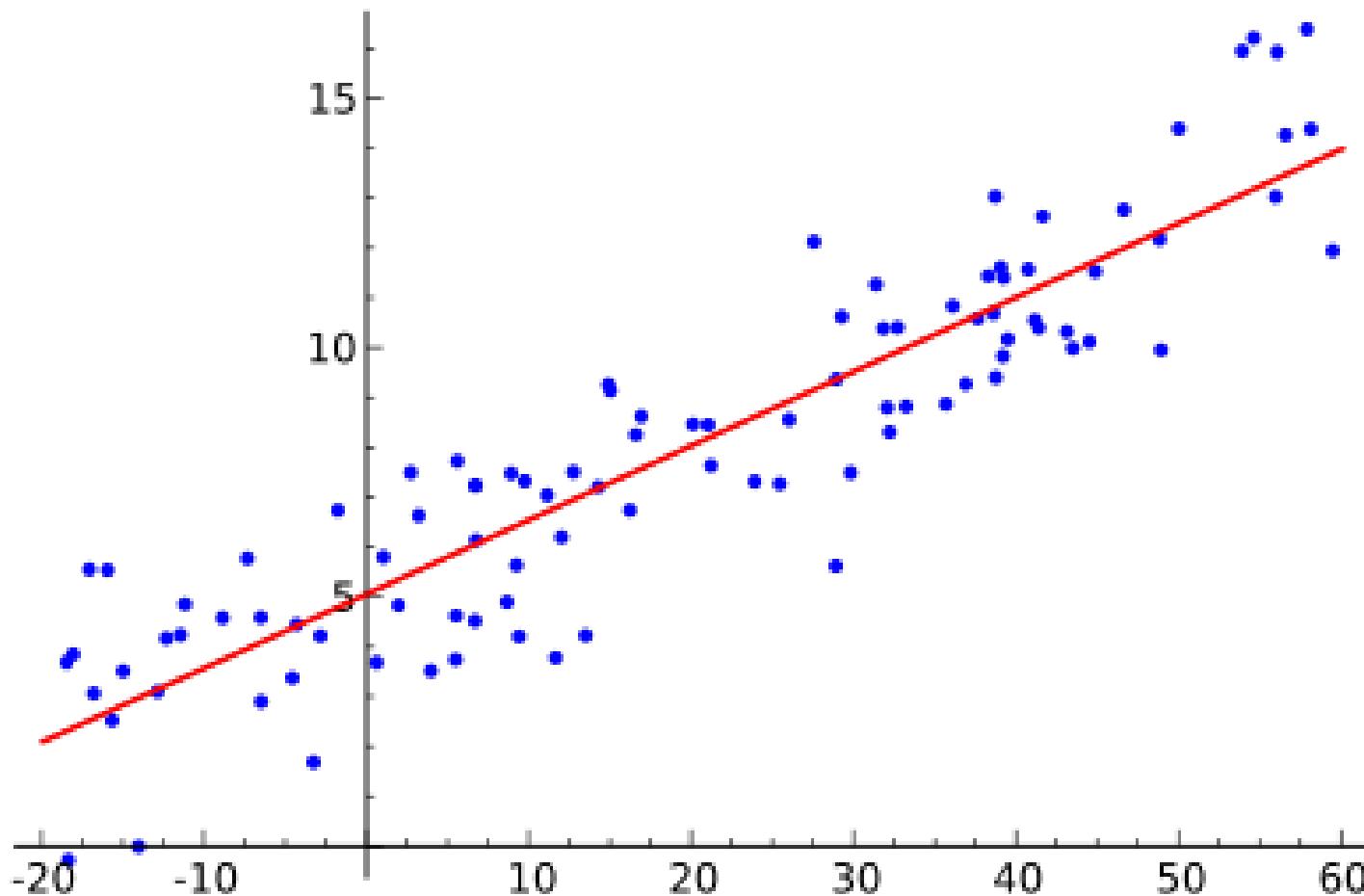
# Régressão Linear

HANDS ON



## 03\_Regressao.ipynb

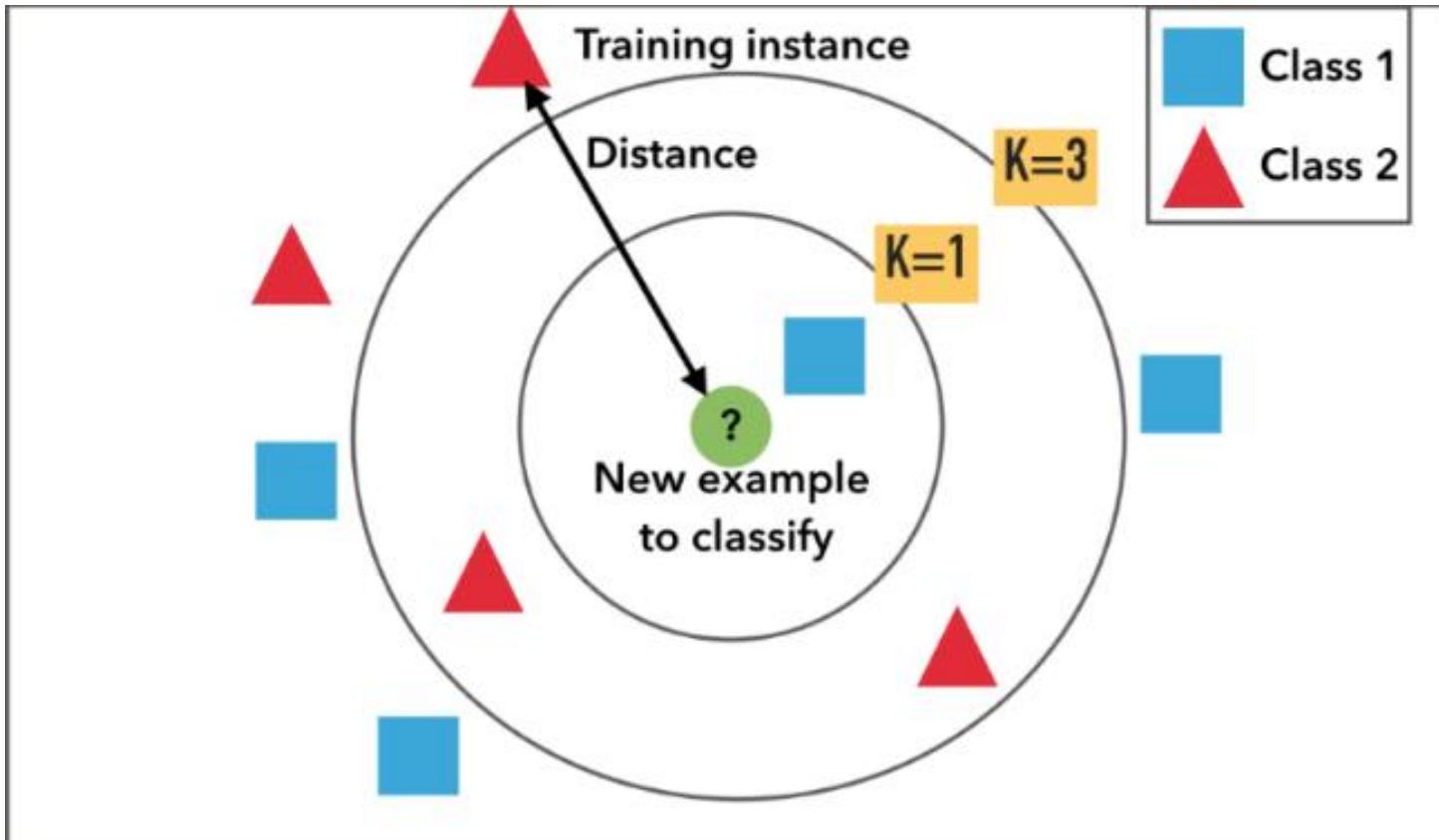
# Regressão Linear



# KNN (K – Nearest Neighbours)

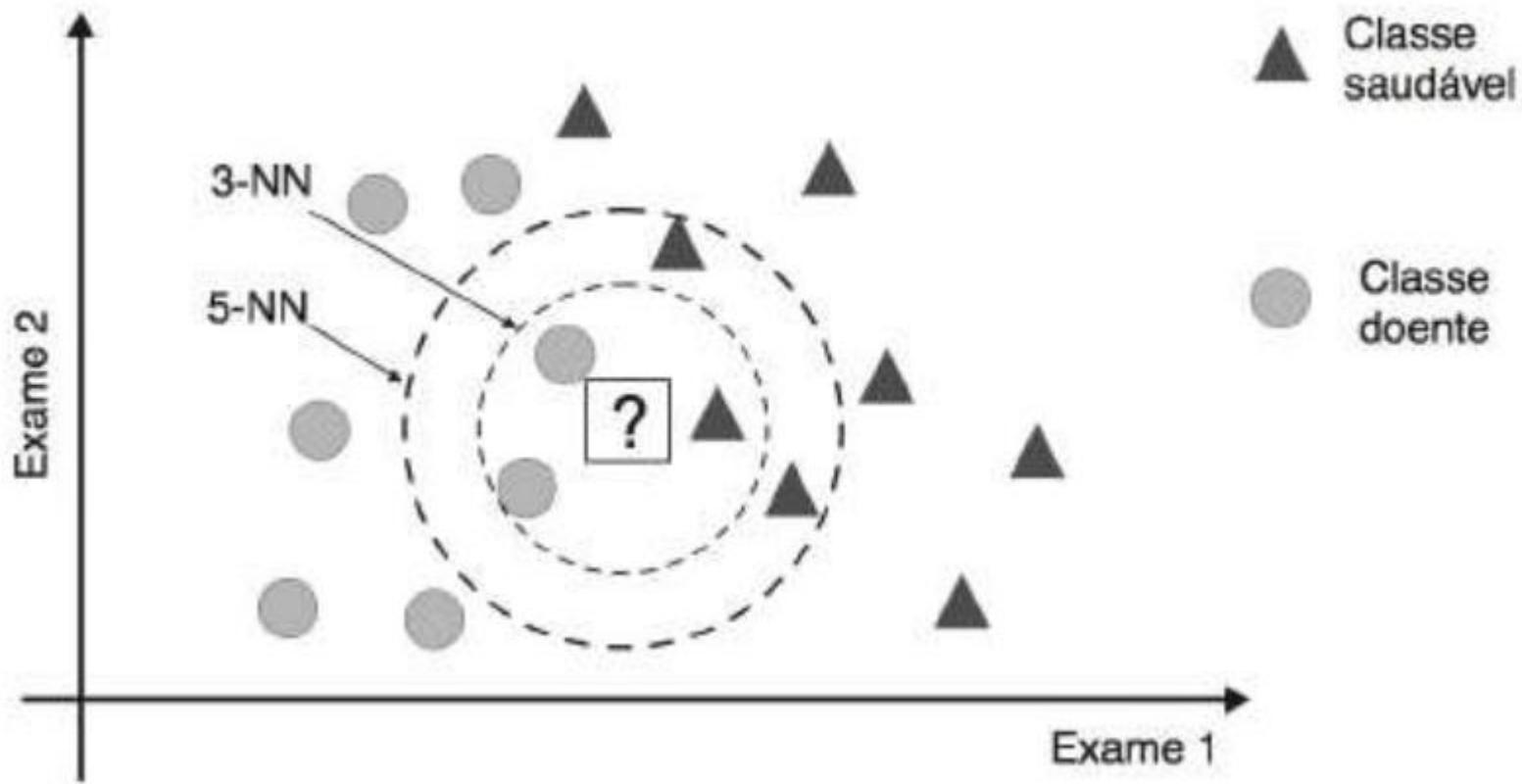
# KNN (K – Nearest Neighbours)

- Objetos relacionados ao mesmo conceito são semelhantes entre si
- Classificar  $x_z$  atribuindo a ele o rótulo representado mais frequentemente dentre as  $k$  amostras mais próximas e utilizando um esquema de votação.



# KNN (K – Nearest Neighbours)

- Dependendo de K a classificação pode mudar



HANDS ON

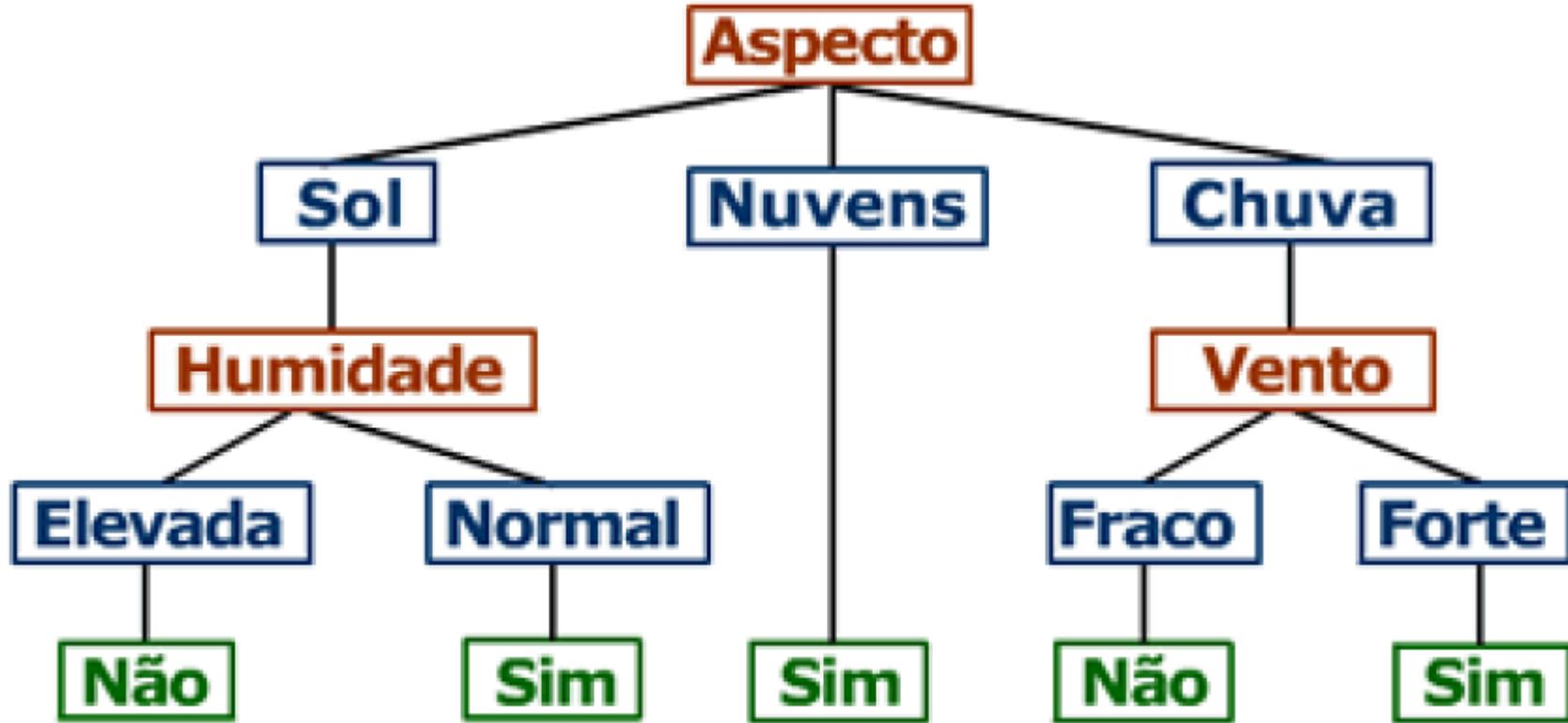


## 04\_KNN.ipynb

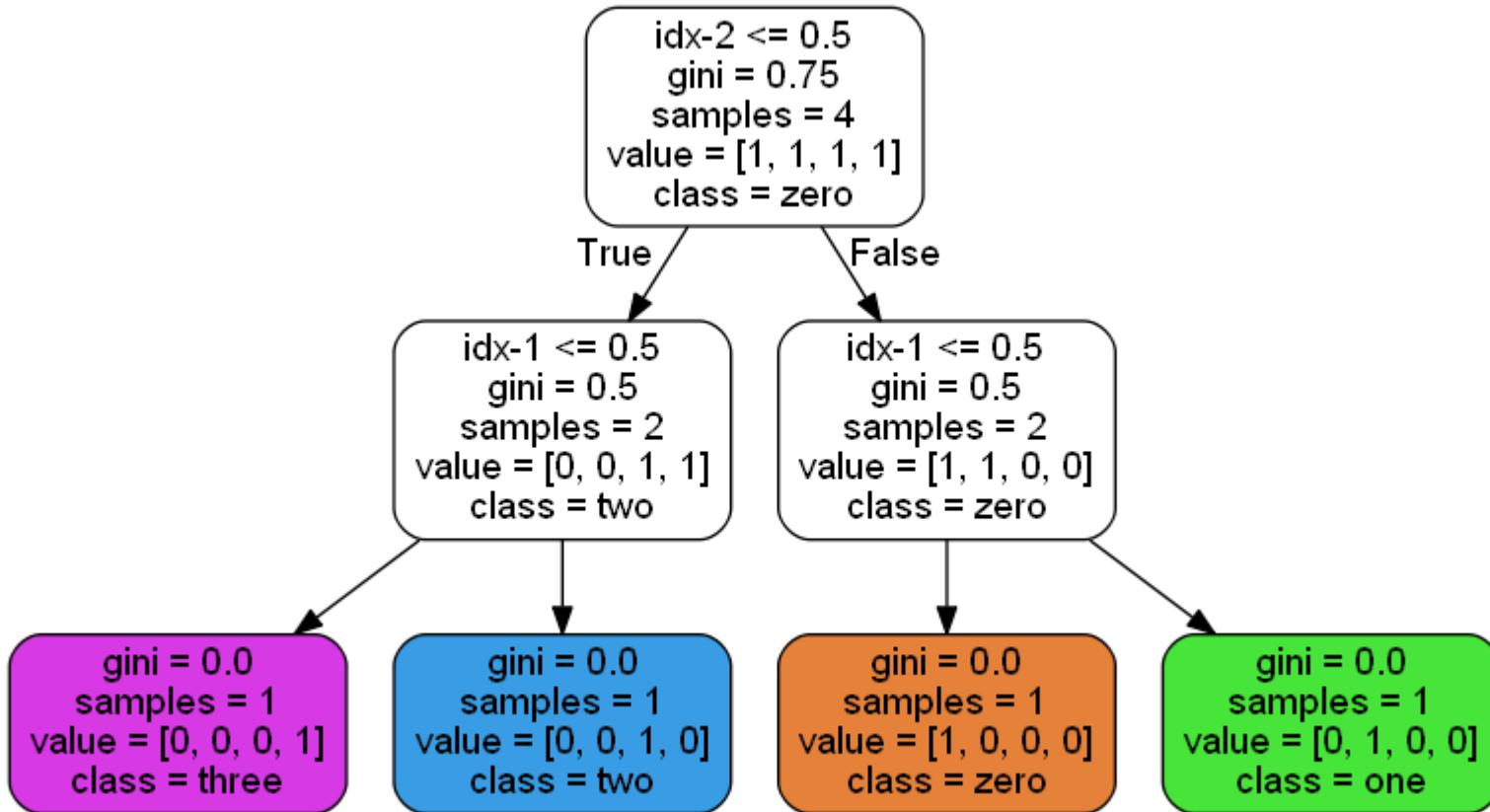
# Random Forrest

# Árvores de Decisão

- Árvores de decisão classificam instâncias ordenando as árvores acima (ou abaixo), a partir da raiz até alguma folha
- Representam caminhos de decisão e possíveis resultados para uma determinada entrada



# Árvores de Decisão



# Random Forrest

f11	f12	f13	f14	f15	t1
f21	f22	f23	f24	f25	t2
f31	f32	f33	f34	f35	t3
:	:	:	:	:	:
:	:	:	:	:	:
fm1	fm2	fm3	fm4	fm5	tm

Dataset

f11	f12	f13	f14	f15	t1
f81	f82	f83	f84	f85	t8
f71	f72	f73	f74	f75	t7
:	:	:	:	:	:
:	:	:	:	:	:
fj1	fj2	fj3	fj4	fj5	tj

Random Dataset  
for Tree-01

f21	f22	f23	f24	f25	t2
f51	f52	f53	f54	f55	t5
f31	f32	f33	f34	f35	t3
:	:	:	:	:	:
:	:	:	:	:	:
fm1	fm2	fm3	fm4	fm5	tm

Random Dataset  
for Tree-02

f31	f32	f33	f34	f35	t3
f61	f62	f63	f64	f65	t6
f91	f92	f73	f94	f95	t9
:	:	:	:	:	:
:	:	:	:	:	:
fk1	fk2	fk3	fk4	fk5	tk

Random Dataset  
for Tree-03

HANDS ON

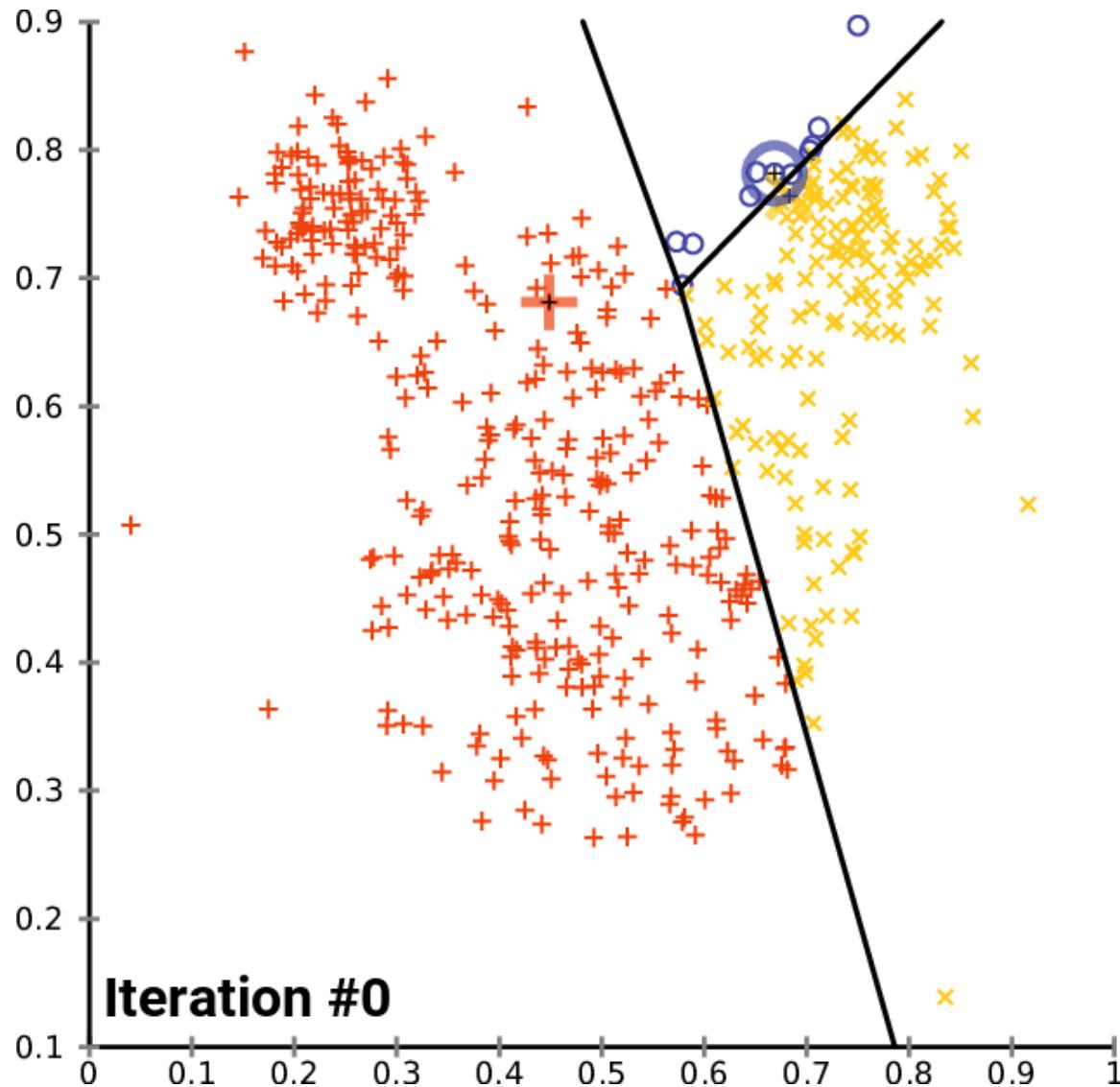


## 05\_Random\_Forest.ipynb

# Clustering

# K-Means

# Clustering com K-Means



HANDS ON

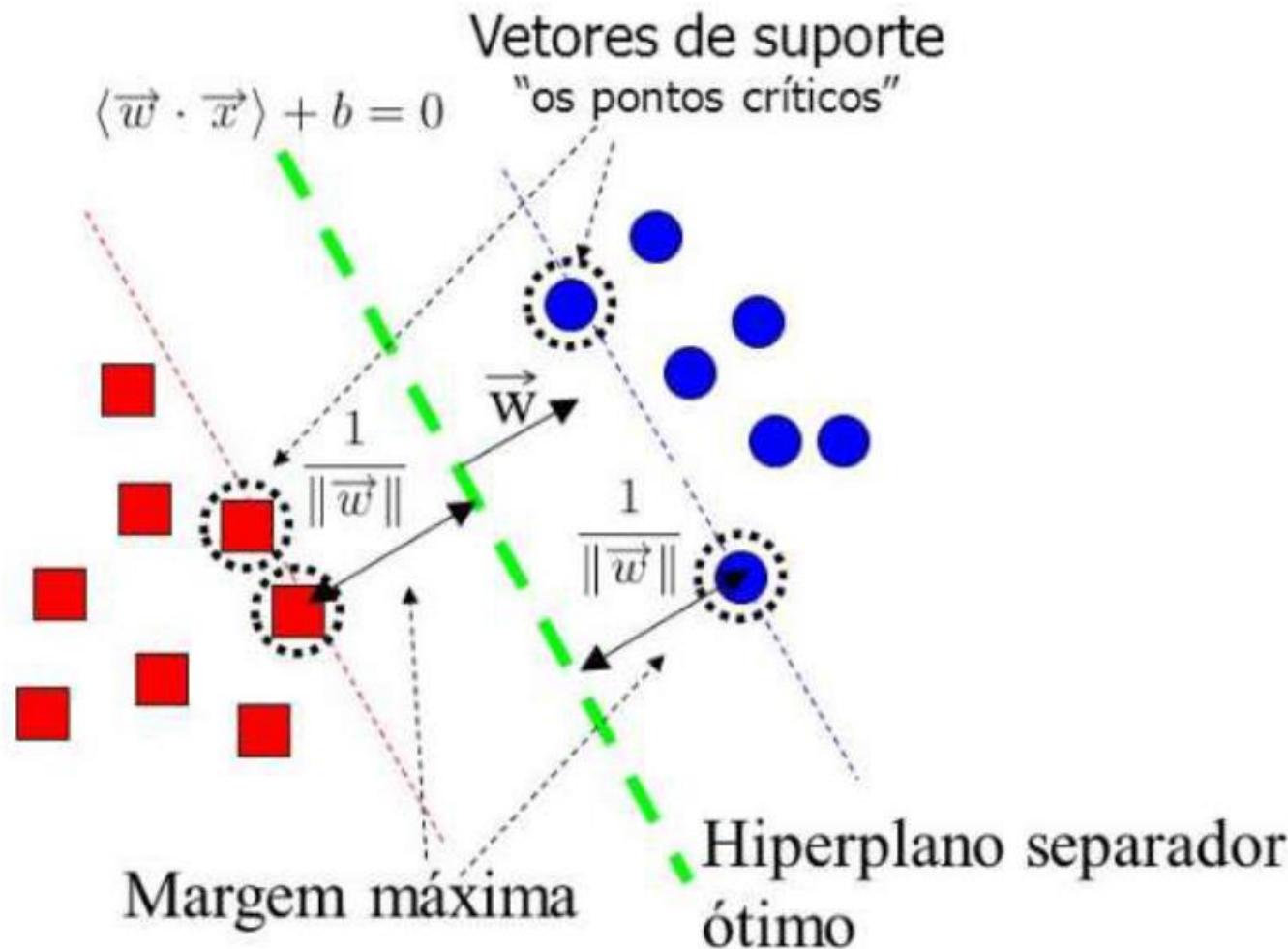


## 06\_Clustering.ipynb

# SVM (Support Vector Machine)

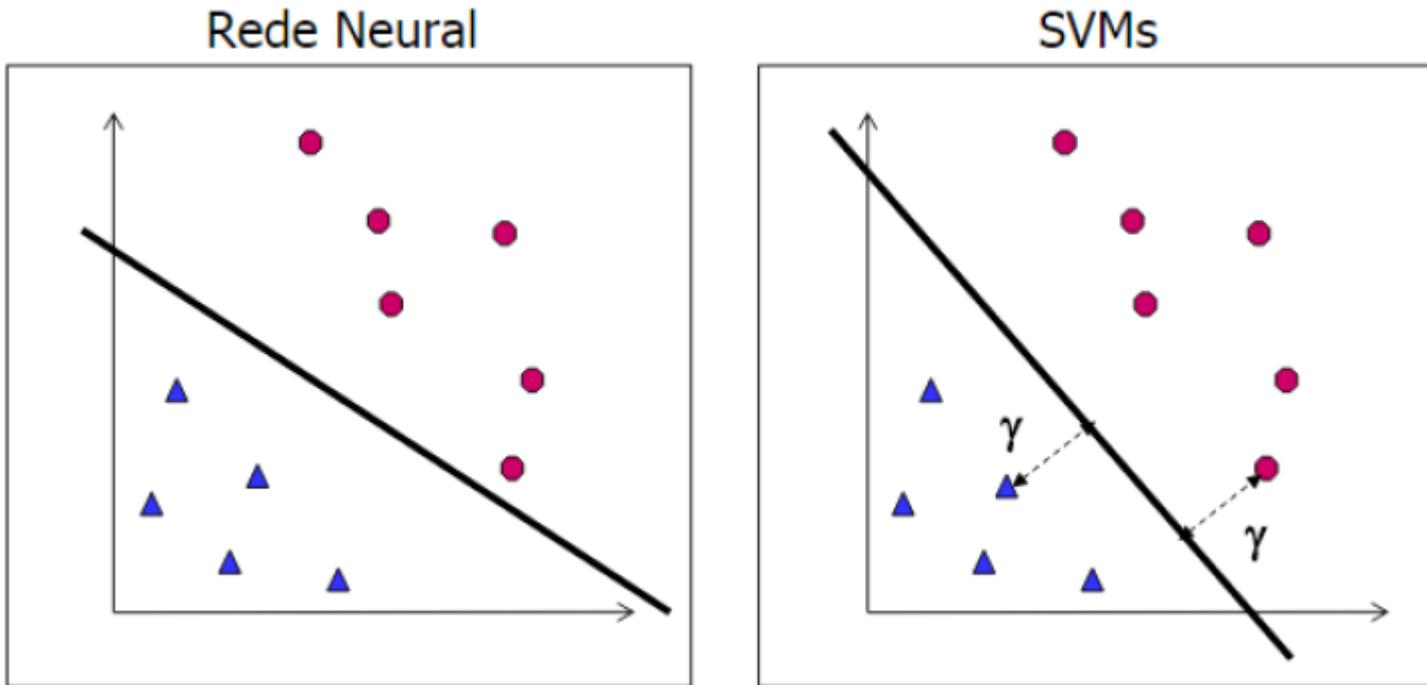
# SVM (Support Vector Machine)

- Support Vector Machine é uma fronteira (hiperplano ) que melhor segregas as duas classes

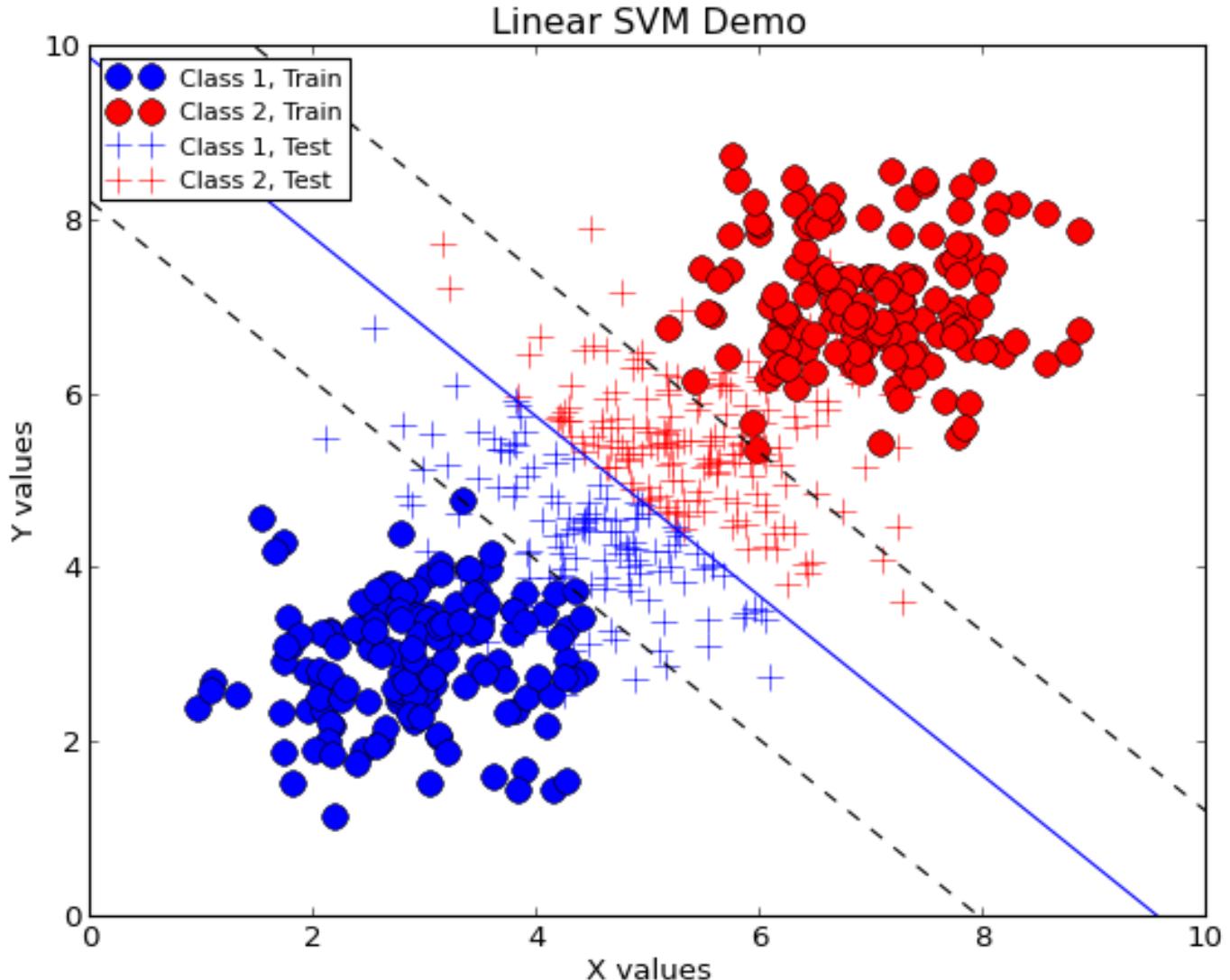


# SVM (Support Vector Machine)

- O uso de SVM's é capaz de resolver problemas de classificação de dados, gerando classificadores que apresentam bons resultados.
- Podem dar resultados melhores que uma rede neural!

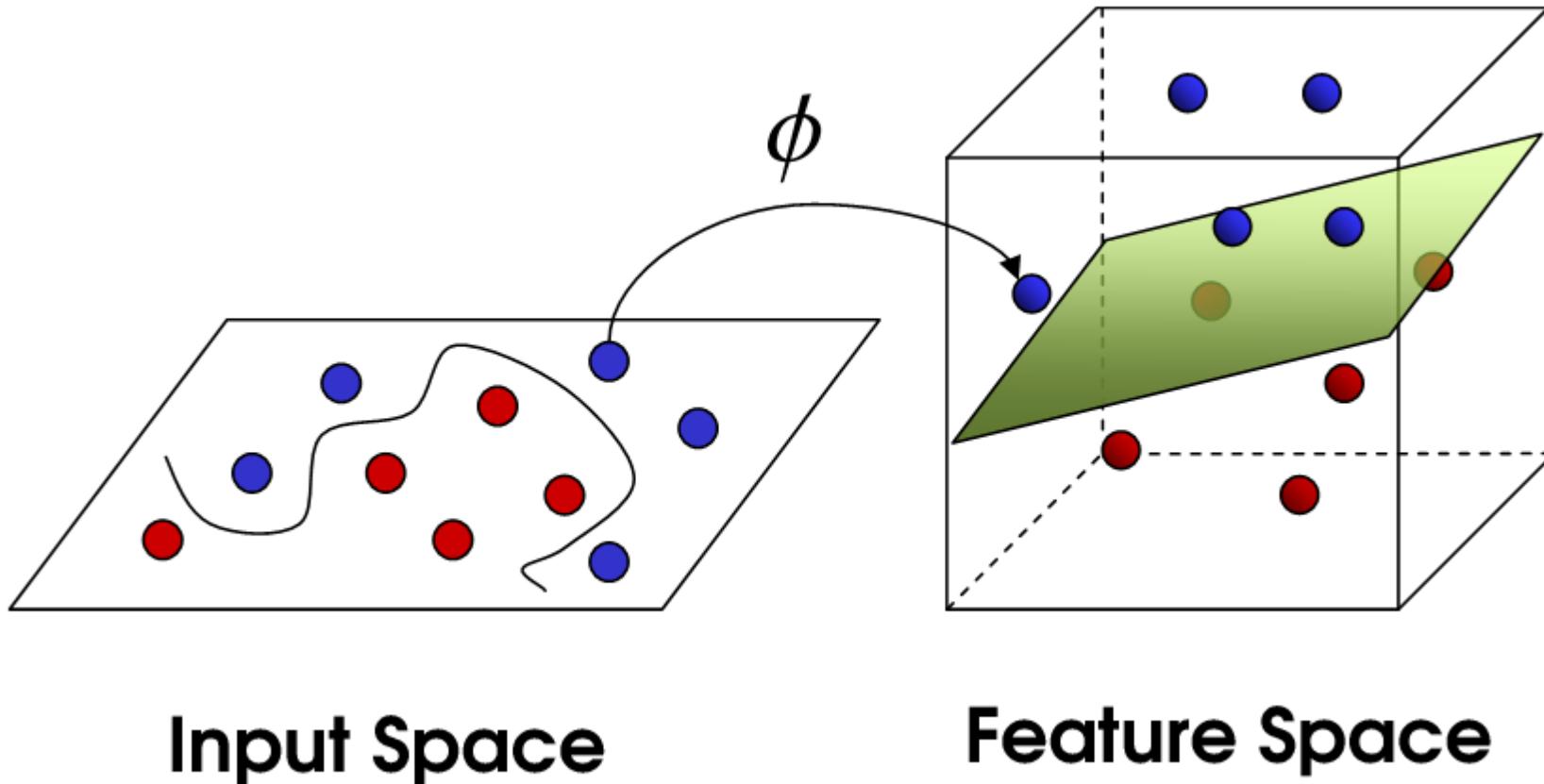


# SVM de 2 dimensões



# Função teta

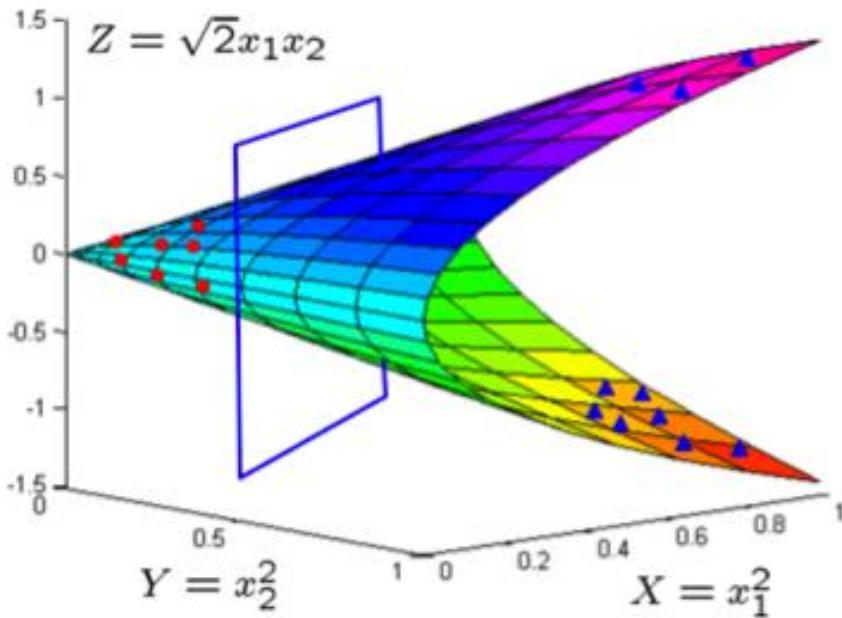
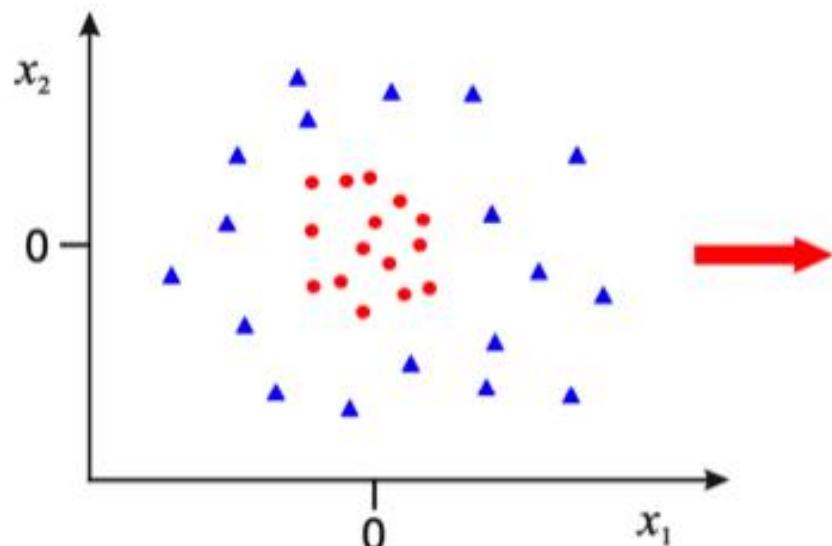
- Criação de uma função teta para aumentar a dimensionalidade e permitir a classificação



# SVM de 2 dimensões (solução com 3 dim.)

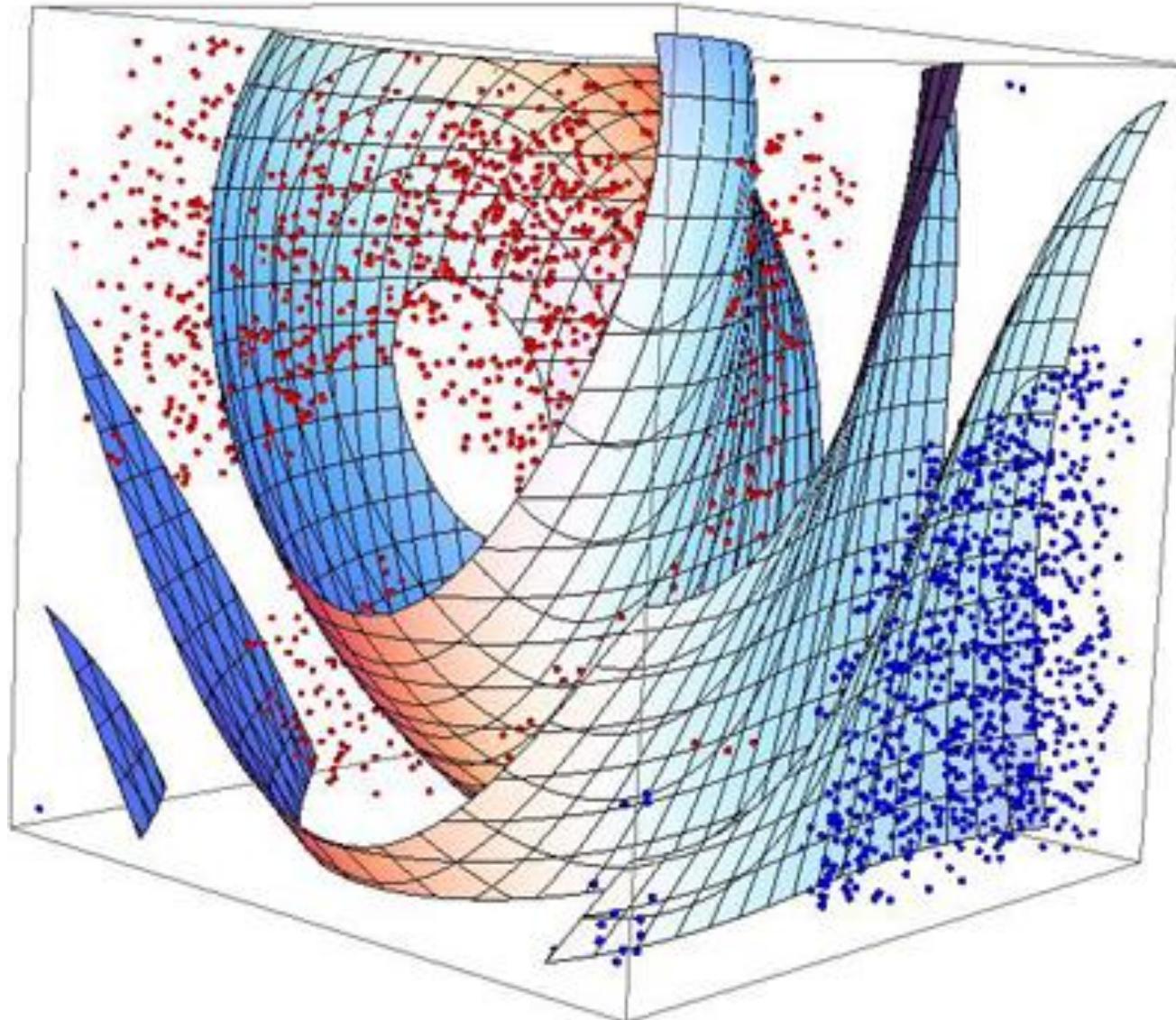
The following pictures should give you a general intuition for what is happening.

$$\Phi : \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \rightarrow \begin{pmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \end{pmatrix} \quad \mathbb{R}^2 \rightarrow \mathbb{R}^3$$



- Data **is** linearly separable in 3D
- This means that the problem can still be solved by a linear classifier

# SVM - Complexidade com N dimensões



HANDS ON



## 07\_SVM.ipynb

# Obrigado!

José Humberto Cruvinel

Contato: [jose.junior@prof.unibh.br](mailto:jose.junior@prof.unibh.br)

<https://www.facebook.com/jhcruvinel>