

---

# Oficina Gratuita de Machine Learning

PUCRS - Abril de 2019

<http://goo.gl/BgoJmm>

[PDF dos Slides](#)

# Edições Anteriores

(400 participantes)



Google



Extended



ft<sub>18</sub>



memed



TECNOPUC  
(6 edições)

# Público-Alvo

---



Devs



Linguistas



Jornalista



Biólogxs



Engenheirxs



Advogadxs



Prof. Saúde



Estudante

...

### DETALHES DO EVENTO

Status Visualizações da página do evento  
● Publicado 1.814

Visibilidade Compartilhar o evento: [f](#) [t](#) [in](#) [COPIAR ENDEREÇO](#)  
Público [?](#) <https://www.sympla.com.br/oficina-gratuita-de-machine-learn...>

### VENDAS

Vendas totais Vendas pendentes  
**R\$ 450,00** **R\$ 0,00**

Repasso total [?](#) Coffee-break  
**R\$ 375,00** **R\$ 361,00**

### INGRESSOS [?](#)

Período: Desde a publicação ▾ Filtrar por canal de venda: Todos os canais ▾ Filtrar por valor do ingresso: Gratuito e pago ▾ Filtrar por tipo de ingresso: Todos os tipos ▾



Data	Vendas (Ingressos)
18 Fev	14
28 Fev	2
10 Mar	1
20 Mar	7
30 Mar	20
09 Abr	2
26 Abr	3

Ingressos confirmados: 119 Ingressos pendentes: 0 Ingressos cancelados: 0

4

# Check-In Sympla

---



apresentação dos participantes

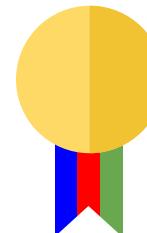
# Facilitadores

---



Henrique

PUCRS



Google  
Research  
Award

# Objetivo

---

Visa capacitar os participantes em análise de dados utilizando modelos estatísticos e aprendizagem de máquina.

# Sopa de Letrinhas

---

Aprendizado de Máquina

Inteligência Artificial  
Big Data  
Supervisionada  
Computação Instintiva

Dados  
Padrões  
Mineração de Dados  
Computação em Núvem  
Banco de Dados  
Ciência de Dados  
Indústria 4.0

Descoberta do Conhecimento  
Computação Cognitiva  
Redes Neurais  
Inteligência Computacional

# Ferramentas

---

## Machine Learning



## Big Data

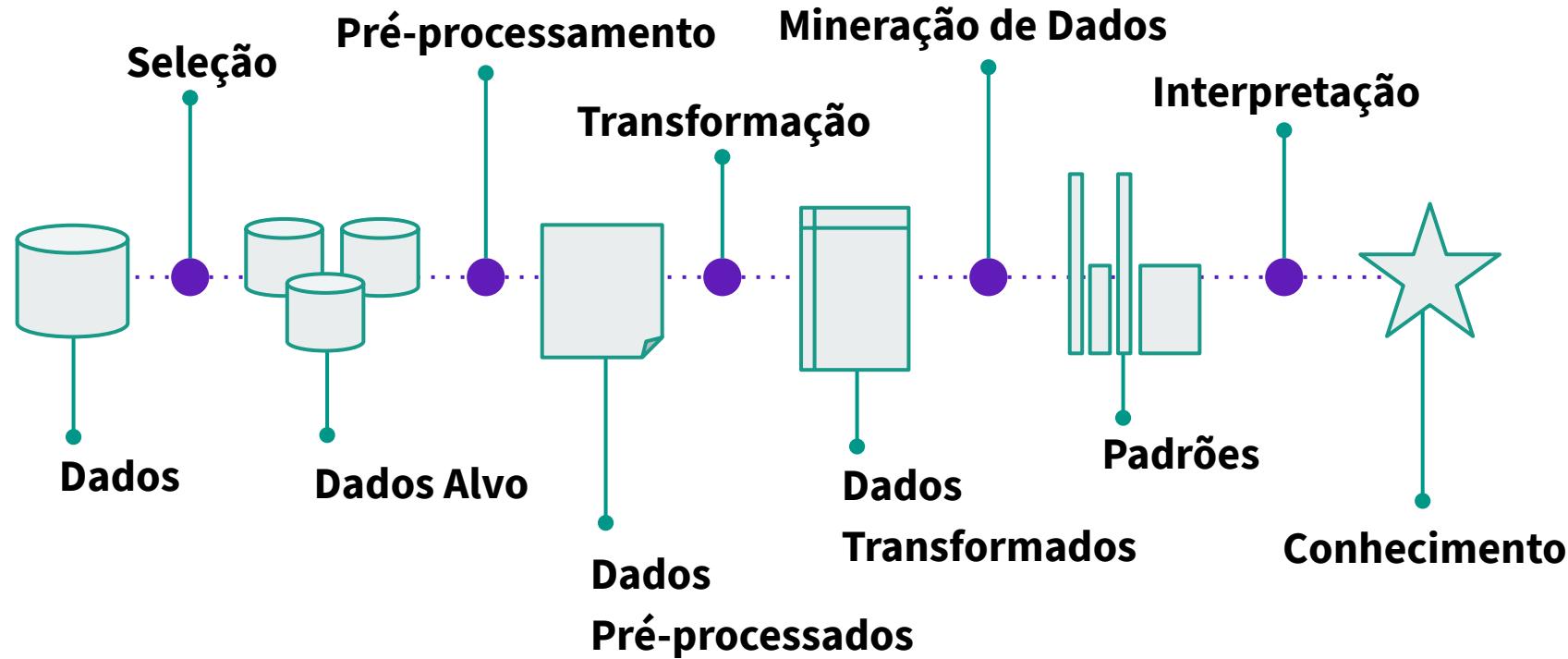


# Ferramentas

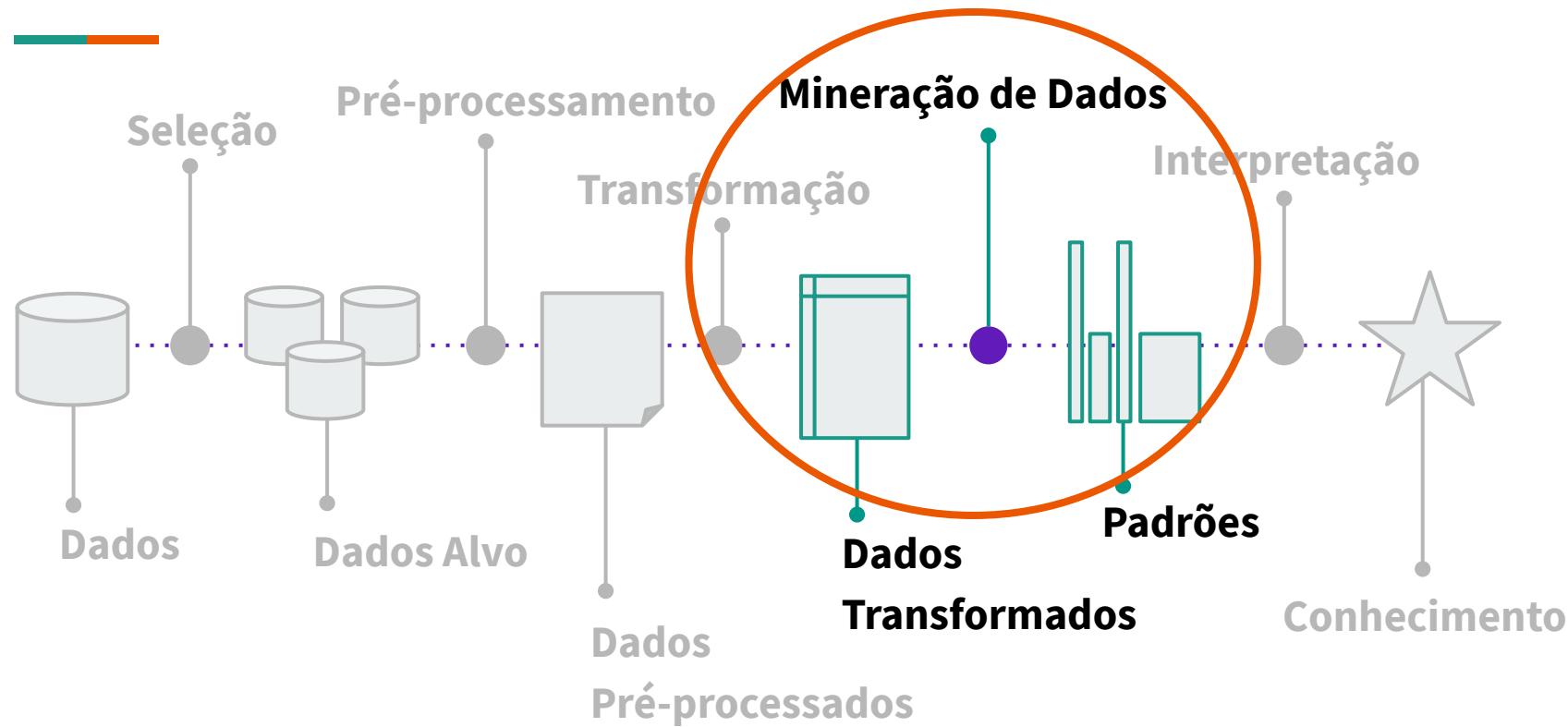
## Gartner Magic Quadrant Data Science



# Descoberta de Conhecimento

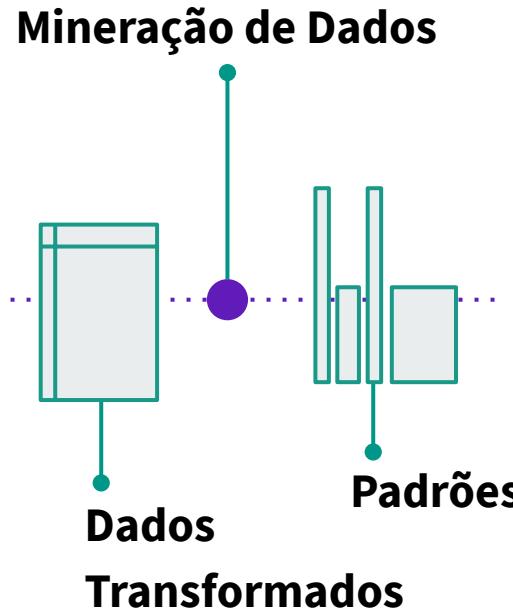


# Descoberta de Conhecimento



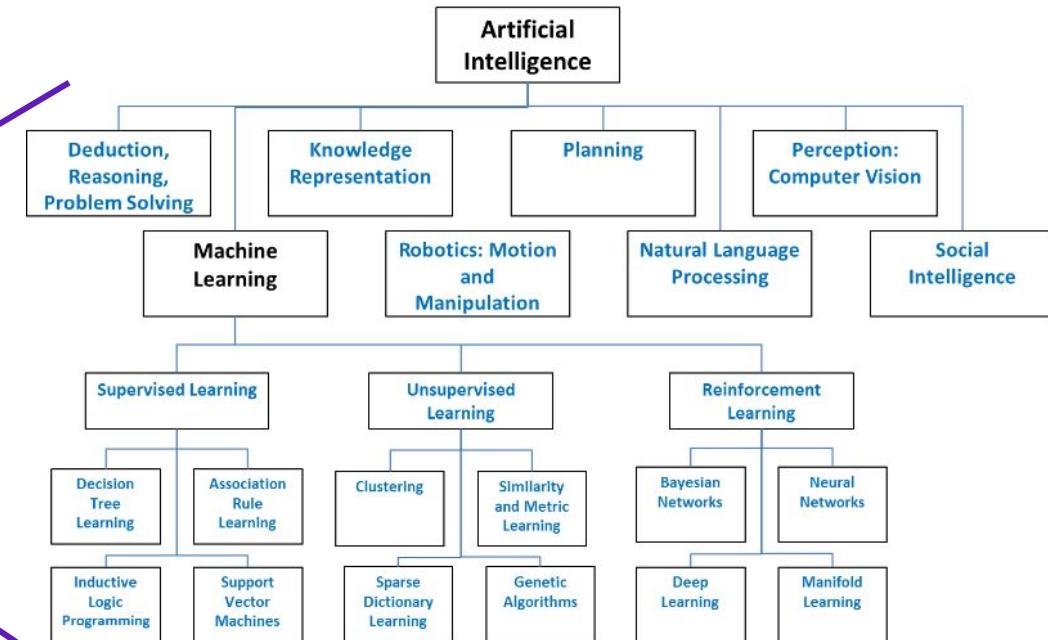
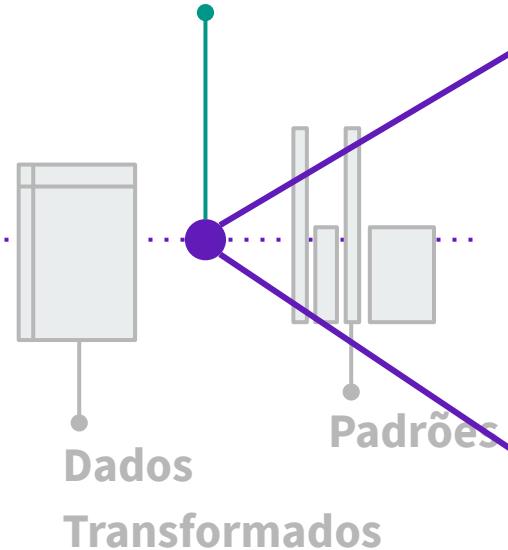
# Mineração de Dados

---



# Aprendizado de Máquina

## Mineração de Dados



## ENGAGEMENT PROCESS

Step 1: Build the Data Model



Step 2: Define The Report



Step 3: Generate SQL commands



Step 4: Create Report



## Data Warehouse

The data warehouse is a "schema-on-load" approach because the data schema must be defined and built prior to loading data into the data warehouse. Without an underlying data model, the BI tools will not work.

### Business Intelligence Questions

What happened?

Descriptive Analysis  
Standard Reporting



Business Intelligence Analyst

### Data Science Questions

Why? What will happen?  
What should I do?

Data Scientist

Predictive Analysis  
Prescriptive Analysis



## DIFFERENCE BETWEEN BUSINESS INTELLIGENCE AND DATA SCIENCE

### CHARACTERISTICS

<b>Focus</b>	Reports, KPIs, trends	Patterns, correlations, models
<b>Process</b>	Static, comparative	Exploratory, experimentation, visual
<b>Data Sources</b>	Pre-planned, added slowly	On the fly, as-needed
<b>Transform</b>	Up front, carefully planned	In-database, on-demand, enrichment
<b>Data quality</b>	Single version of truth	"Good enough," probabilities
<b>Data model</b>	Schema on load	Schema on query
<b>Analytics</b>	Retrospective, Descriptive	Predictive, Prescriptive, Preventative

## ENGAGEMENT PROCESS

Step 1: Define Hypothesis to Test ←



Step 2: Gather Data



Step 3: Build Data Model



repeat

Step 4: Explore the Data



Step 5: Build and Refine Analytic Models



Step 6: Ascertain Goodness of Fit



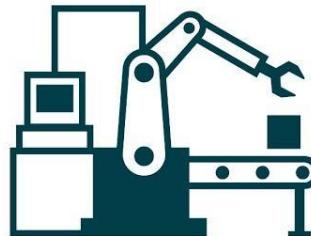


## INDUSTRY 1.0

Mechanization, steam power, weaving loom

## INDUSTRY 2.0

Mass production,  
assembly line,  
electrical energy



## INDUSTRY 3.0

Automation, computers  
and electronics



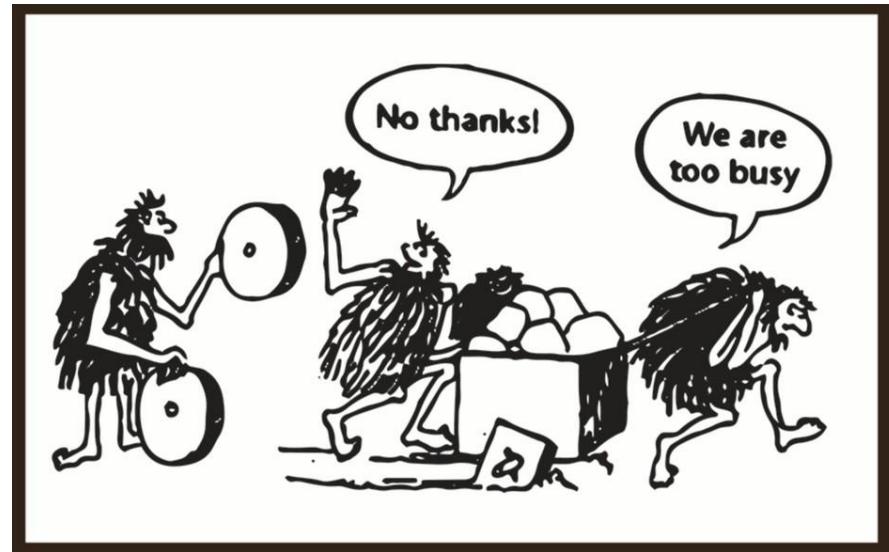
## INDUSTRY 4.0

Cyber Physical Systems,  
internet of things, networks  
artificial intelligence

# Porque usar Data Science?

---

Para reduzir os custos e aumentar a eficiência da sua organização, automatizando e otimizando processos.

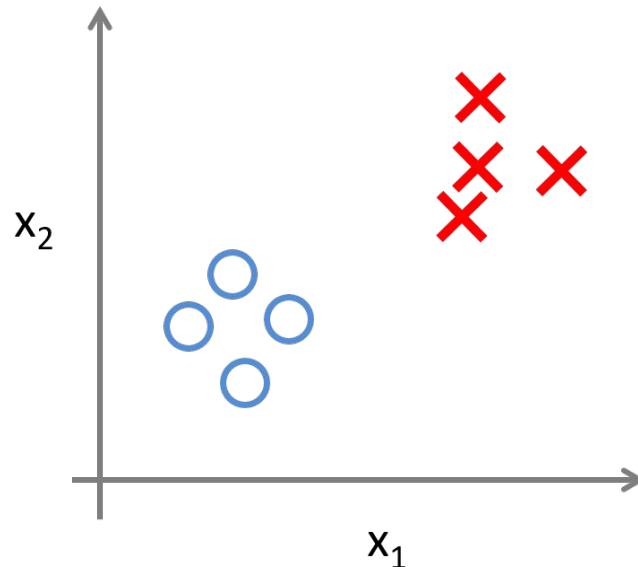


# Tipos de Aprendizagem

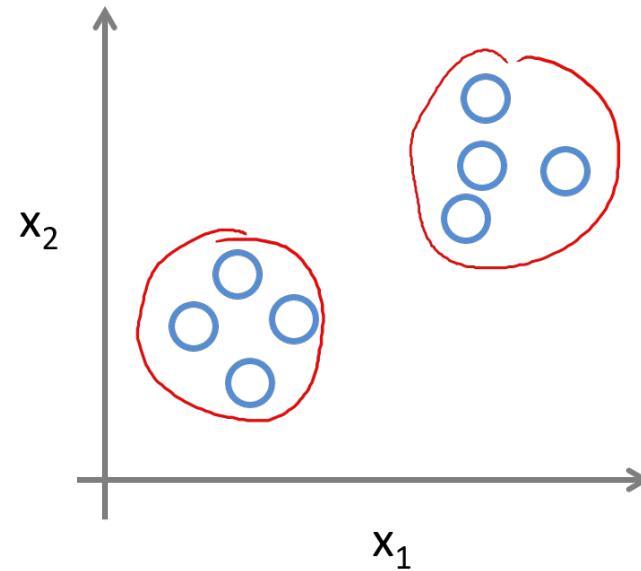
---

# Aprendizado de Máquina

Supervised Learning



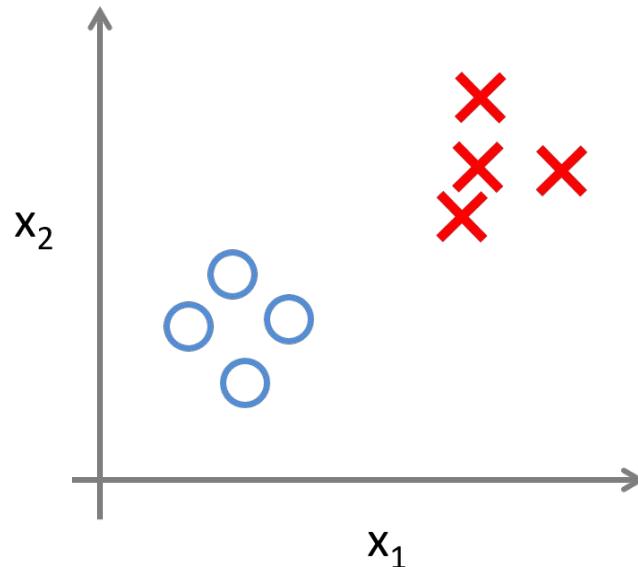
Unsupervised Learning



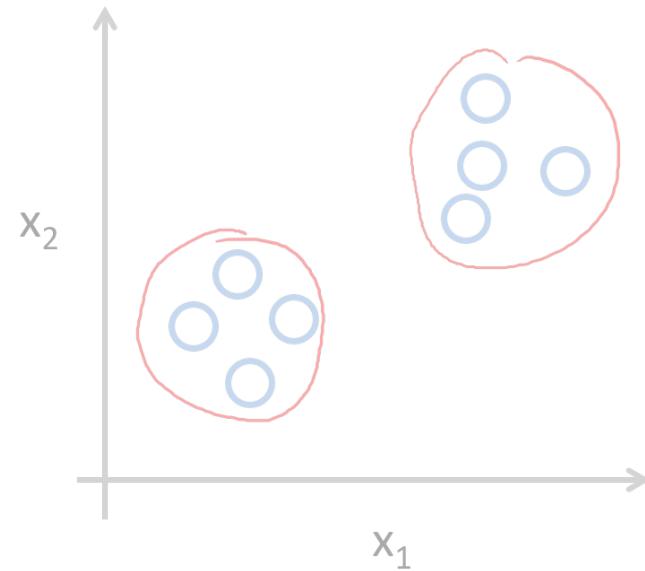
# Supervisionado

---

Supervised Learning



Unsupervised Learning





 alamy stock photo

JC3KYR  
www.alamy.com



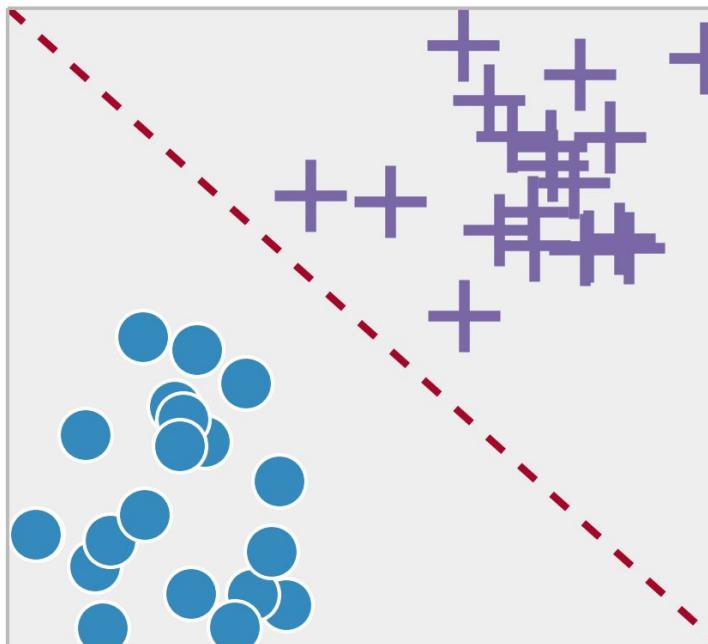
 alamy stock photo

E8NGXG  
www.alamy.com

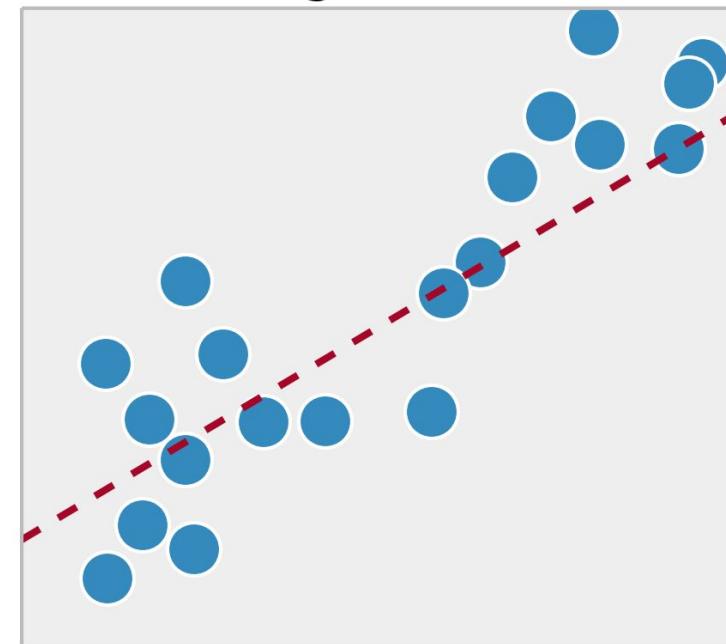
# Supervisionado

---

Classification



Regression



# Classificação (atributos)

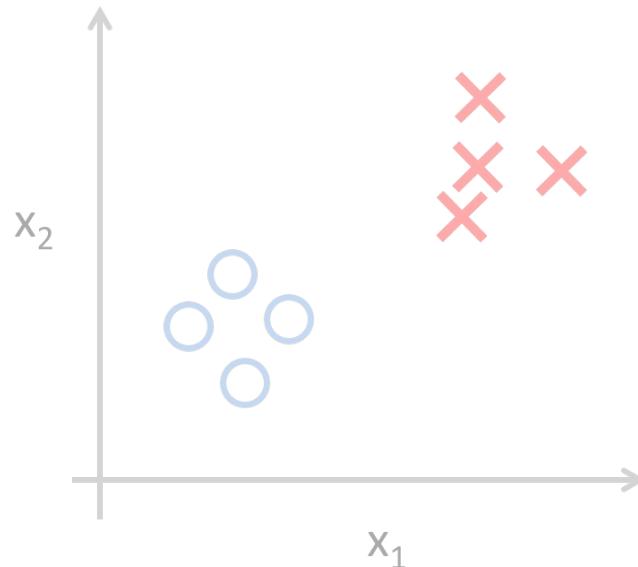
Sintomas	Exames	Sinais	Resultado
[0, 0, 1, 0, 1]	[0, 0.15, 8]	[0, 10, 15]	Sim
[0, 1, 0, 0, 0]	[0.11, 0, 0]	[0.5, 9, 12]	Não
[0, 1, 0, 0, 1]	[0, 0.12, 7]	[0, 11, 12]	Sim

# Regressão (atributos)

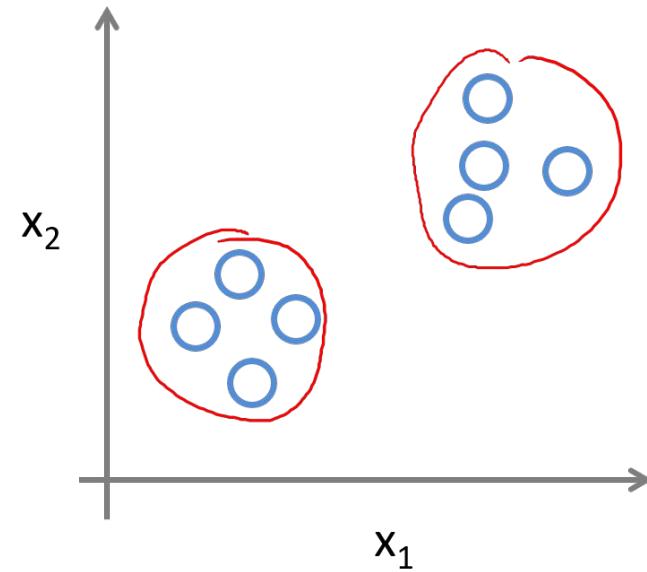
Sintomas	Exames	Sinais	Resultado
0 0 1 0 1	0 0.15 8	0 10 15	15
0 1 0 0 0	0.11 0 0	0.5 9 12	5
0 1 0 0 1	0 0.12 7	0 11 12	20

# Não-Supervisionado

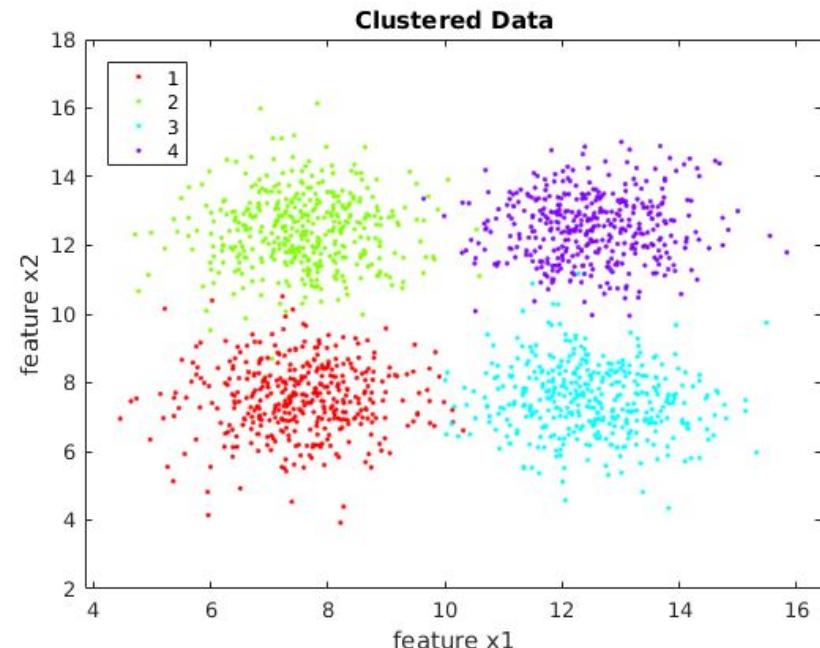
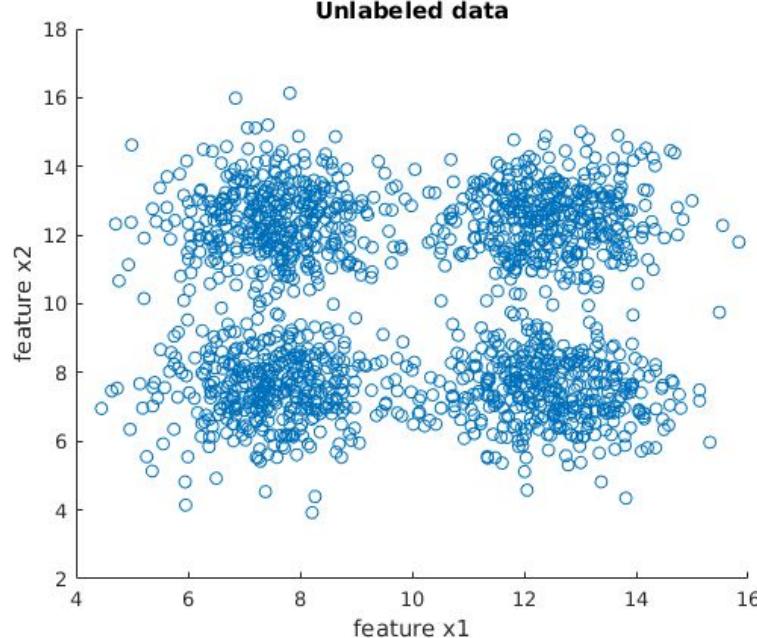
Supervised Learning



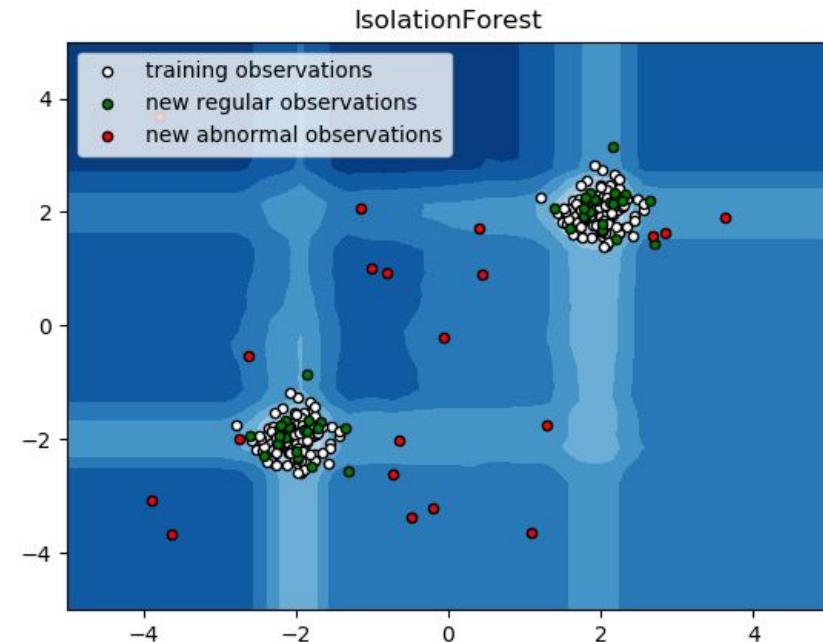
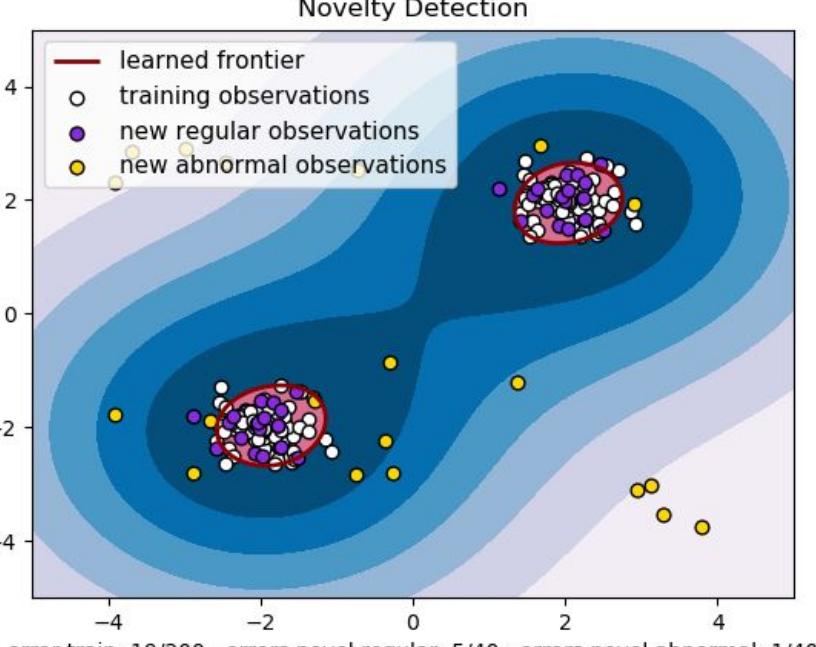
Unsupervised Learning



# Não-Supervisionado **Agrupamento**



# Não-Supervisionado Detecção de Anomalia



# Não-Supervisionado Regras de Associação

---



Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B \& C \Rightarrow D$	1/5	1/3	5/9

$$\text{Supprt} = \frac{\text{Frequency}(X, Y)}{N}$$

$$\rightarrow \text{Confidence} = \frac{\text{Frequency}(X, Y)}{\text{Frequency}(X)}$$

$$\text{Lift} = \frac{\text{Support}}{\text{Support}(X) \times \text{Support}(Y)}$$

# Não-Supervisionado Word Embeddings

---

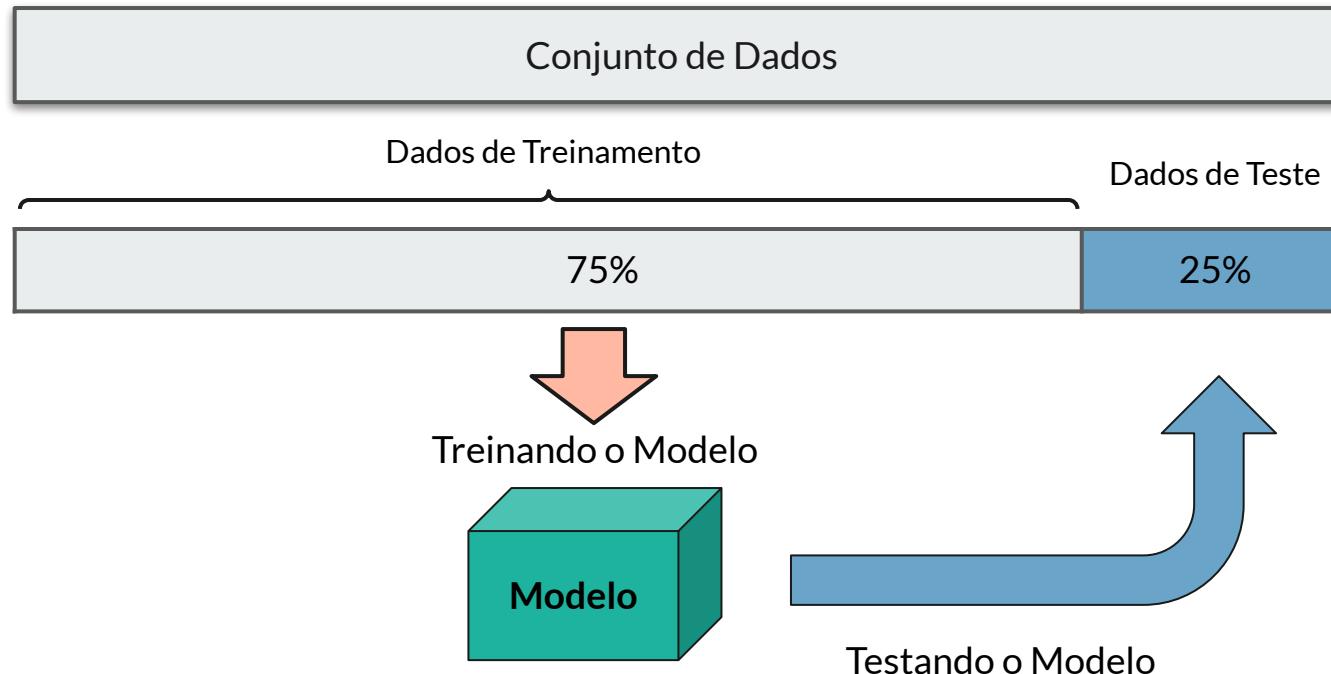


# Avaliação dos Modelos

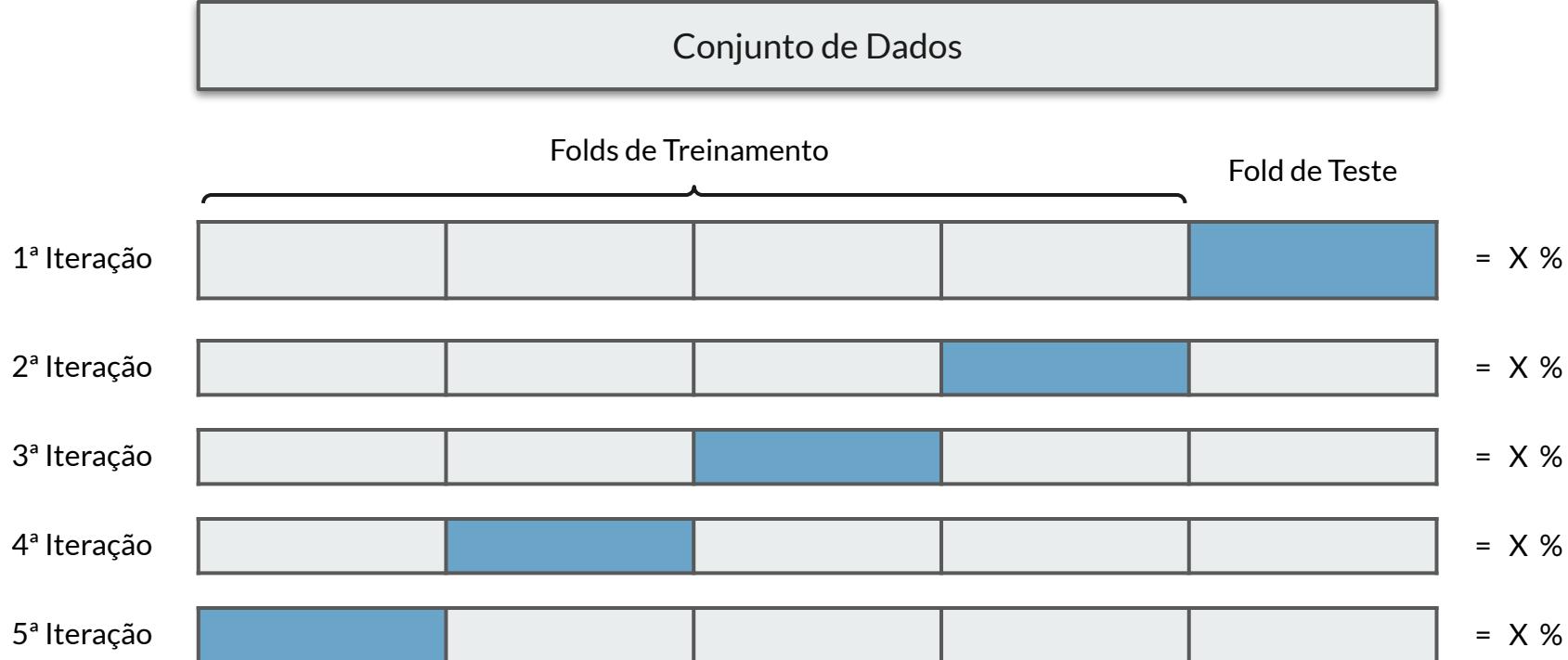


# Treino e Teste

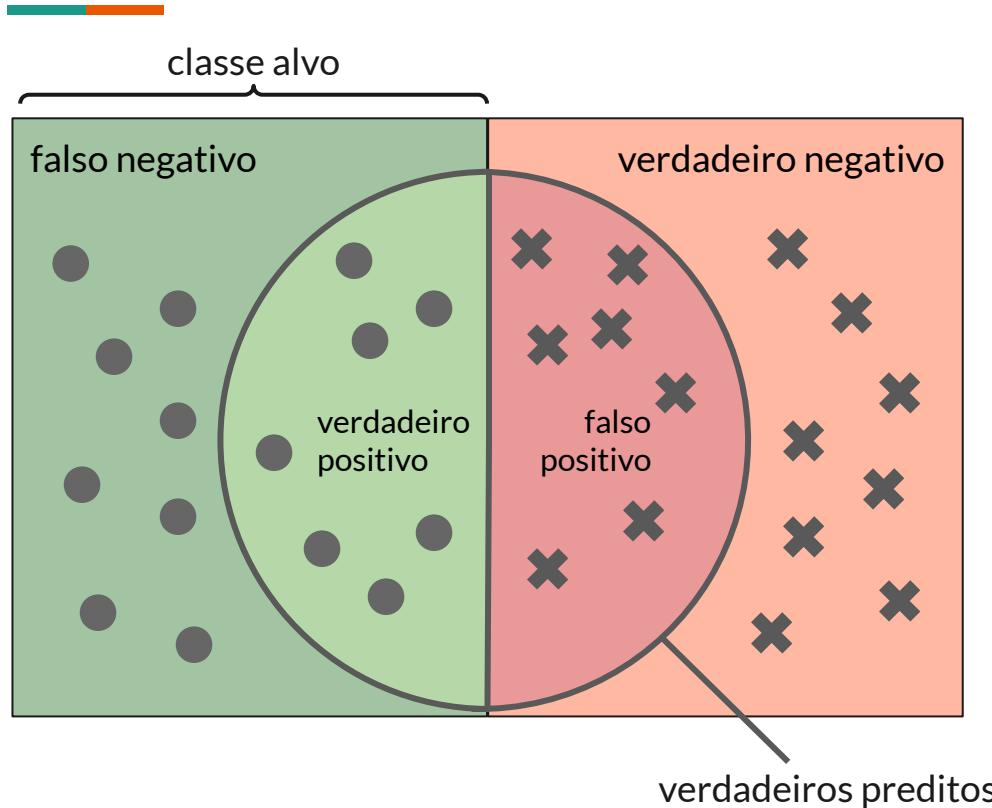
---



# Validação Cruzada (5 Folds)



# Acurácia, Precisão e Abrangência Classificação

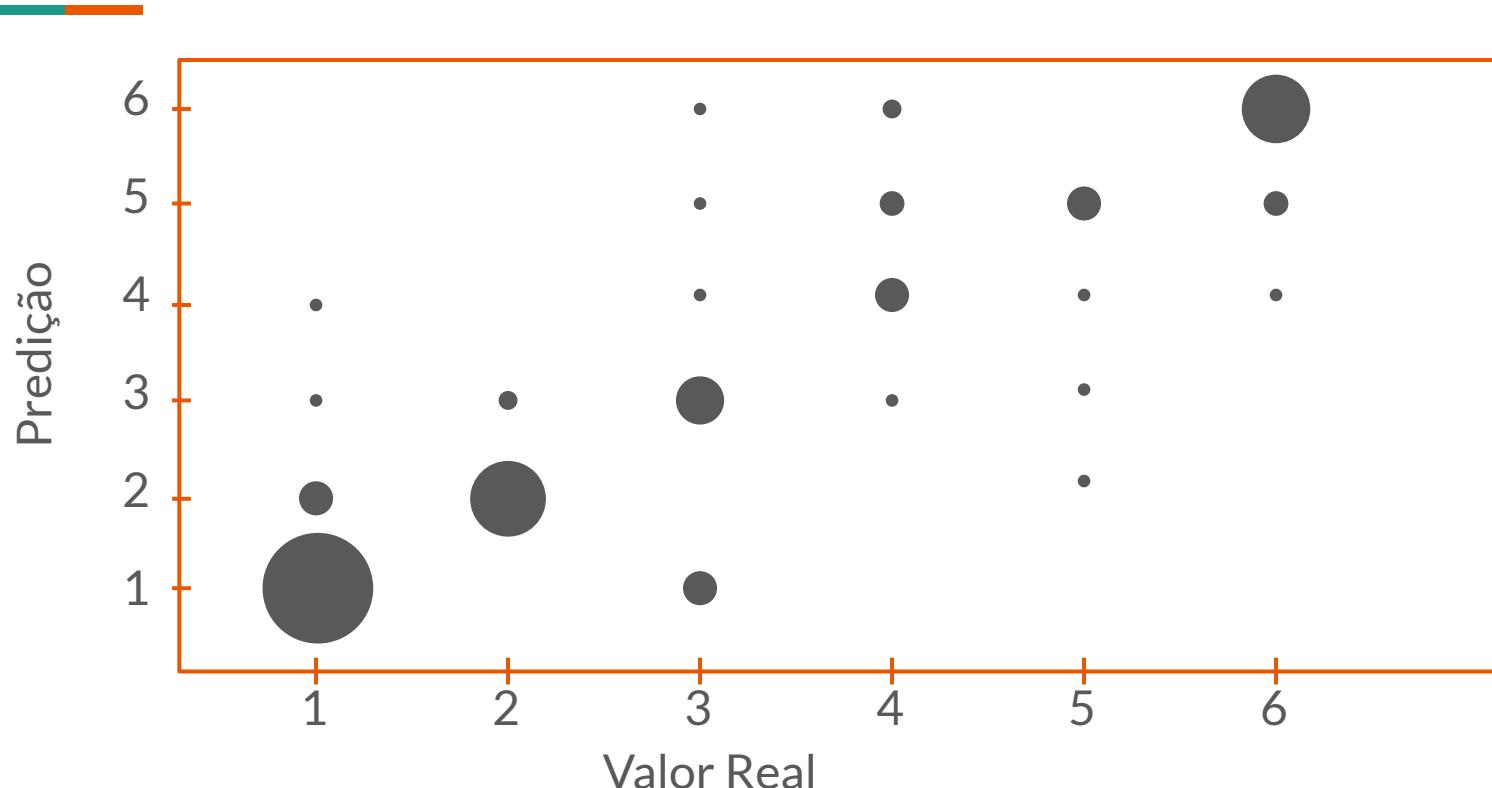


$$\text{Acurácia} = \frac{\text{verdadeiro negativo} + \text{verdadeiro positivo}}{\text{falso negativo} + \text{verdadeiro negativo} + \text{falso positivo} + \text{verdadeiro positivo}}$$

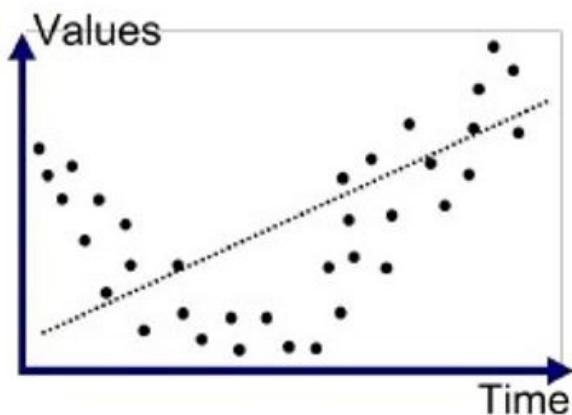
$$\text{Precisão} = \frac{\text{verdadeiro positivo}}{\text{verdadeiro positivo} + \text{falso positivo}}$$

$$\text{Abrangência} = \frac{\text{verdadeiro positivo}}{\text{verdadeiro positivo} + \text{falso negativo}}$$

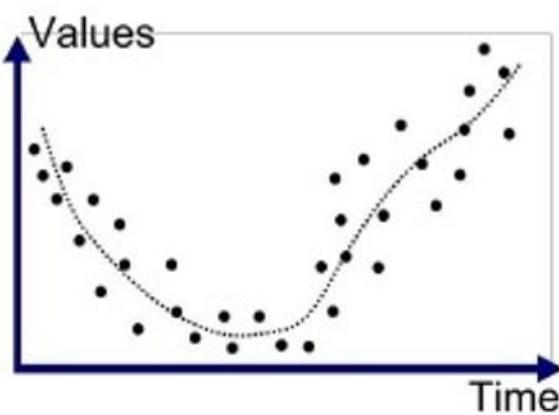
# Erro Médio Absoluto Regressão



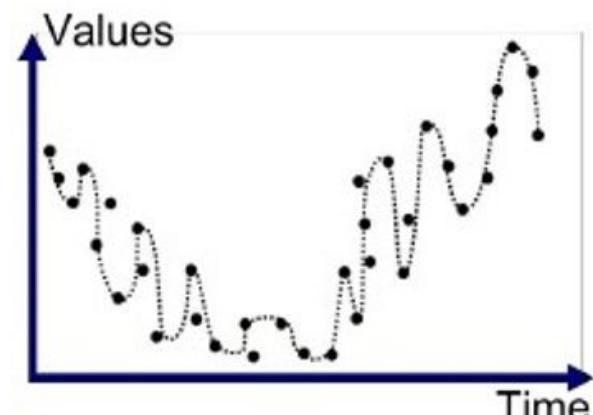
# Overfitting na Regressão



Underfitted



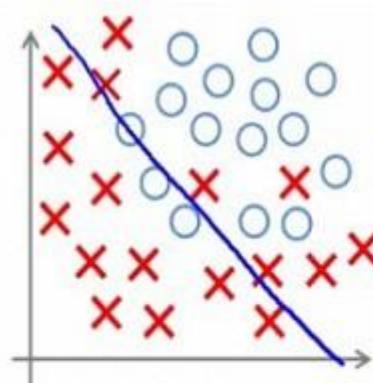
Good Fit/R robust



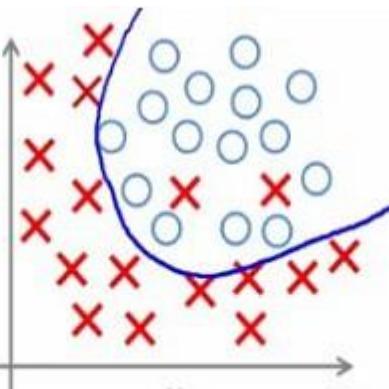
Overfitted

# Overfitting na Classificação

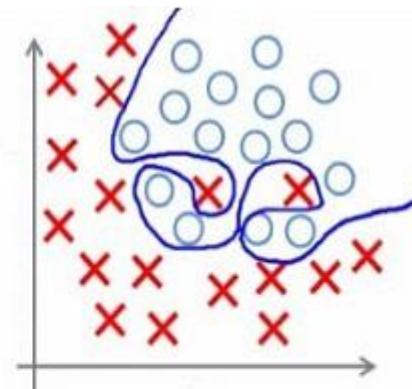
---



Under-fitting



Appropriate-fitting



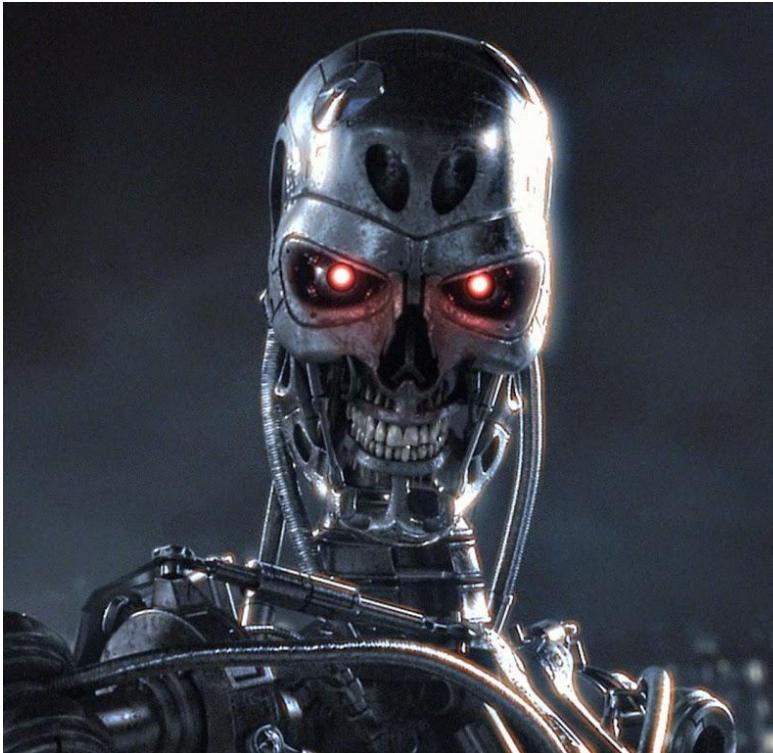
Over-fitting

(too simple to  
explain the  
variance)

(overfitting -- too  
good to be true)

# I.A. Genérica

---



# Ética em I.A. Representando Minorias

---



# Fixando Métricas de Classificação

---

Datasets: Qualquer um.

Escolha um dataset do Orange e avalia alguns algoritmos de classificação para ver qual funciona melhor com o dataset escolhido.

- Qual o melhor algoritmo para o seu dataset?

# Material de Classificação

---

- Slides sobre Decision Tree
- Slides sobre Naive Bayes
- Machine Learning 101

# Fixando Classificação

---

Datasets: *Attrition* e *Bank Marketing*

Avalia alguns algoritmos de classificação para ver qual funciona melhor com o dataset escolhido. Faça alteração nos parâmetros para ver a melhora das métricas.

- Qual as features mais importantes para o dataset?
- Qual a métrica que deve ser utilizada?

# Material de Regressão

---

- Slides sobre Linear Regression
- Towards Data Science: [Intro](#), [Mais](#)
- [Decision Tree Regressor](#)

# Fixando Regressão

---

Datasets: *HDI* e *Boston Housing*

Estude o dataset (features e target) e proponha um algoritmo de regressão para o problema. Faça variação dos parâmetros.

- Qual as features mais importantes para o dataset?
- Proponha uma visualização dos dados considerando as features importantes.

# Datalku

---



## Cloud Machine Learning

<https://www.dataiku.com/dss/trynow/saas/>

# Video Aulas

---



simplilearn

# Pré-Processamento no Orange

---



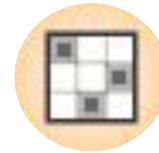
Merge Data



Transpose



Feature Constructor



Impute



Discretize



Color Data

# Fixando Pré-Processamentos

---

Datasets: *Boston Housing* e *Bank Marketing*

Faça transformações nos dados e verifique a alteração no desempenho dos algoritmos.

- Qual algoritmo mais se beneficia com o pré-processamento?

# Material de Aprendizado Não-Supervisionado

---

- Regras de Associação
- Slides de K-Means Clustering
- Principal Component Analysis

# Fixando Aprendizado Não-Supervisionado

---

Datasets: *Foodmart 2000, Bank Marketing,*

Utilize os algoritmos de não supervisionados para fazer análises dos datasets.

- Qual o agrupamento mais significativo?
- Qual a regra de associação mais relevante?

# Material de Redes Neurais

---

- Slides de Redes Neurais
- Machine vs Deep Learning
- Tutorial de Neural Network

# Fixando Redes Neurais

---

Datasets: *Boston Housing* e *Bank Marketing*

Troque os hiper-parâmetros das redes neurais para encontrar uma topologia que resolve melhor a tarefa.

- Qual a melhor configuração para cada dataset?

# Material de PLN

---

- Slides de PLN
- Pré-Processamento de Texto
- Classificação de Texto

# Fixando PLN

---

Datasets: *Green Tails*

Altere as configurações do pré-processamento do texto para melhorar a predição das classes.

- Qual o melhor pré-processamento?

# Documentação Orange

---



[Documentação](#)

[Canal Oficial no YouTube](#)

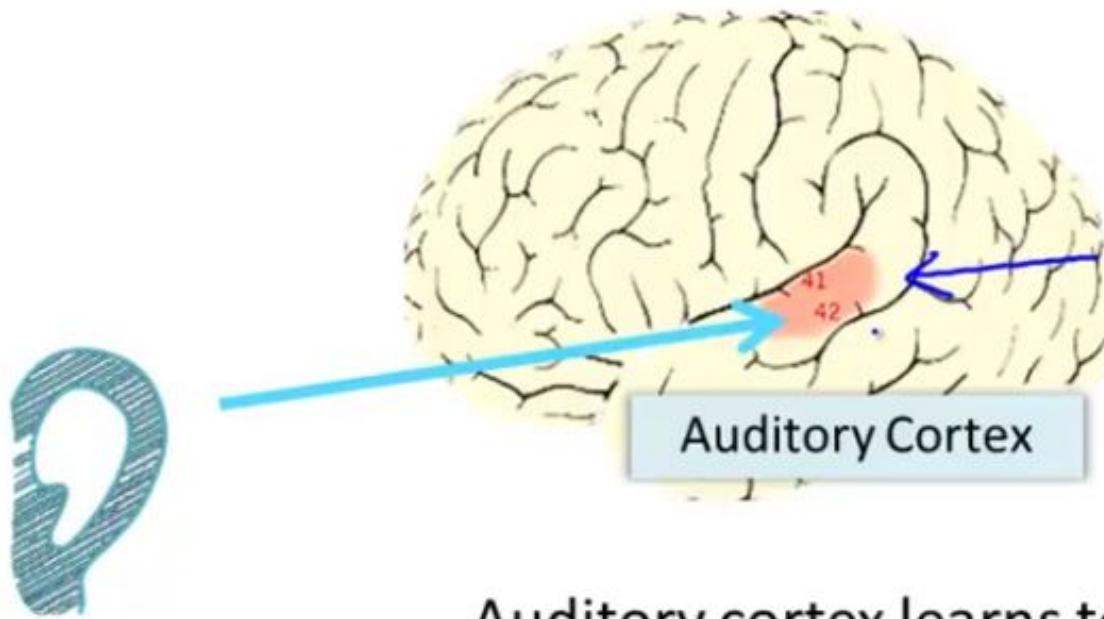
# Redes Neurais

---



# A hipótese do algoritmo de aprendizado único

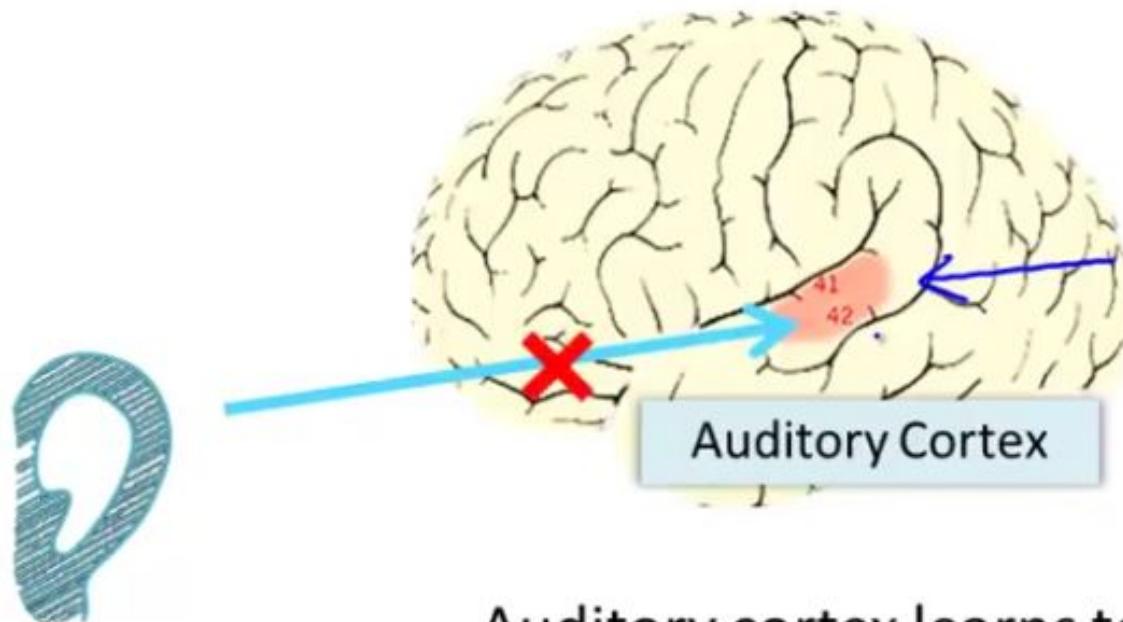
---



Auditory cortex learns to see

# A hipótese do algoritmo de aprendizado único

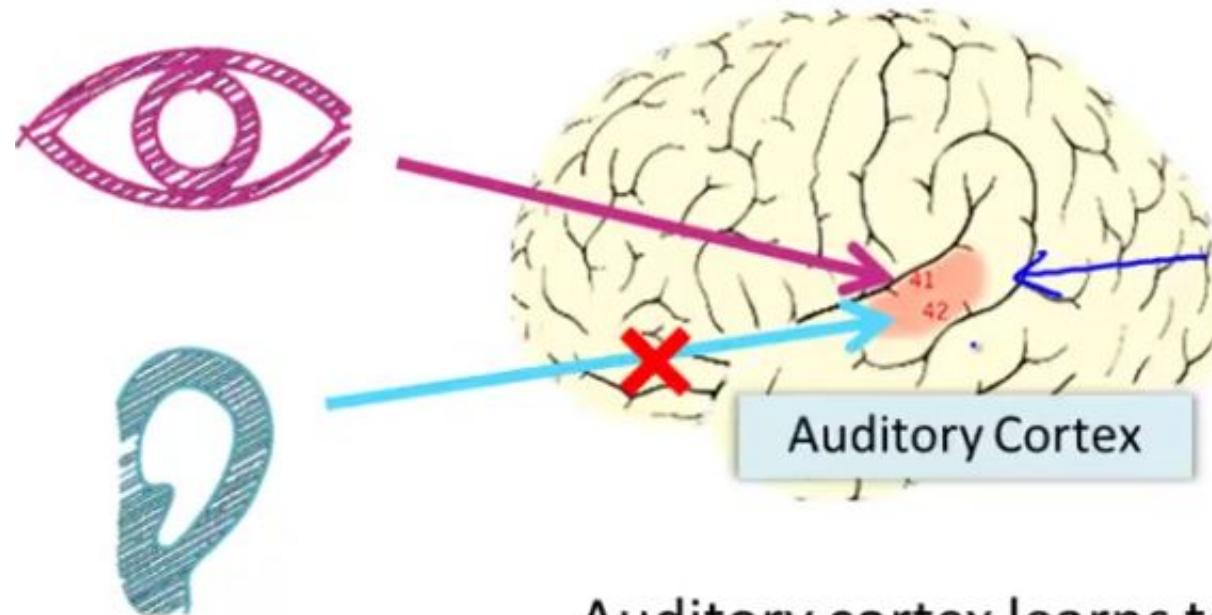
---



Auditory cortex learns to see

# A hipótese do algoritmo de aprendizado único

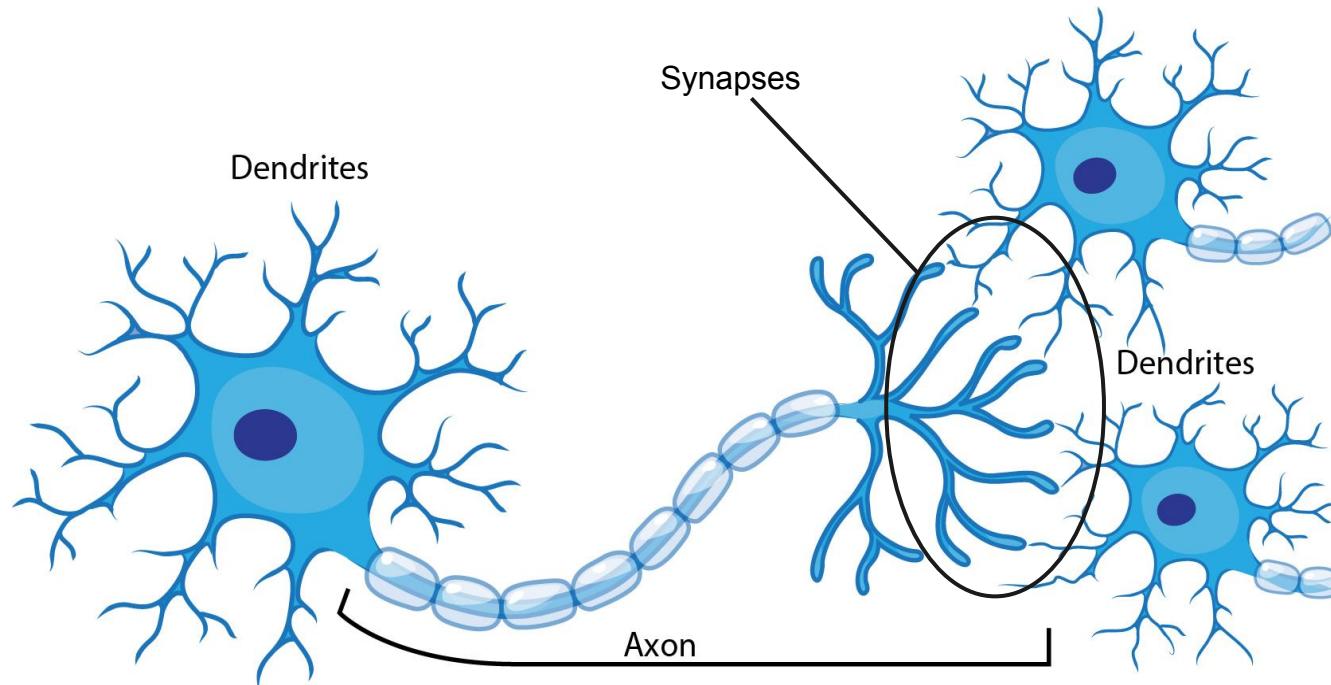
---



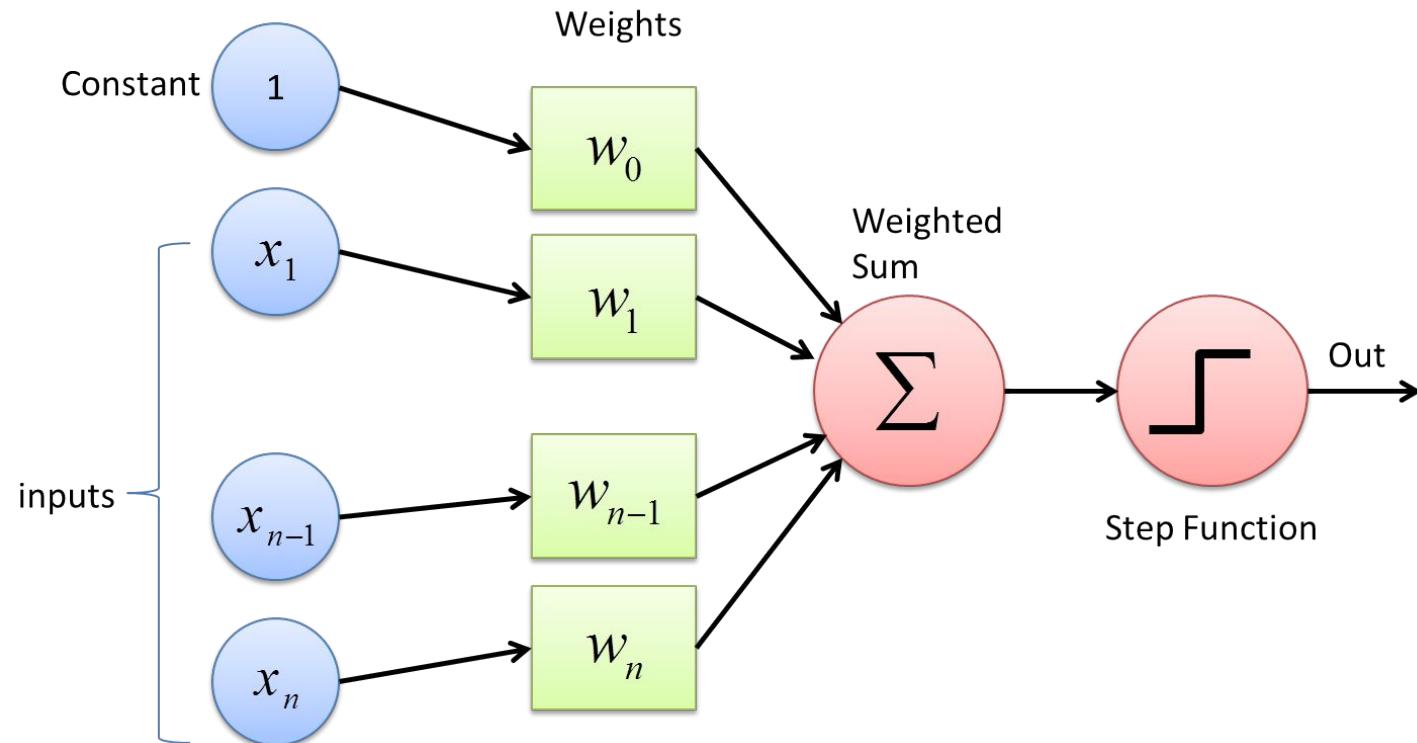
Auditory cortex learns to see

# Neurônio (perceptron)

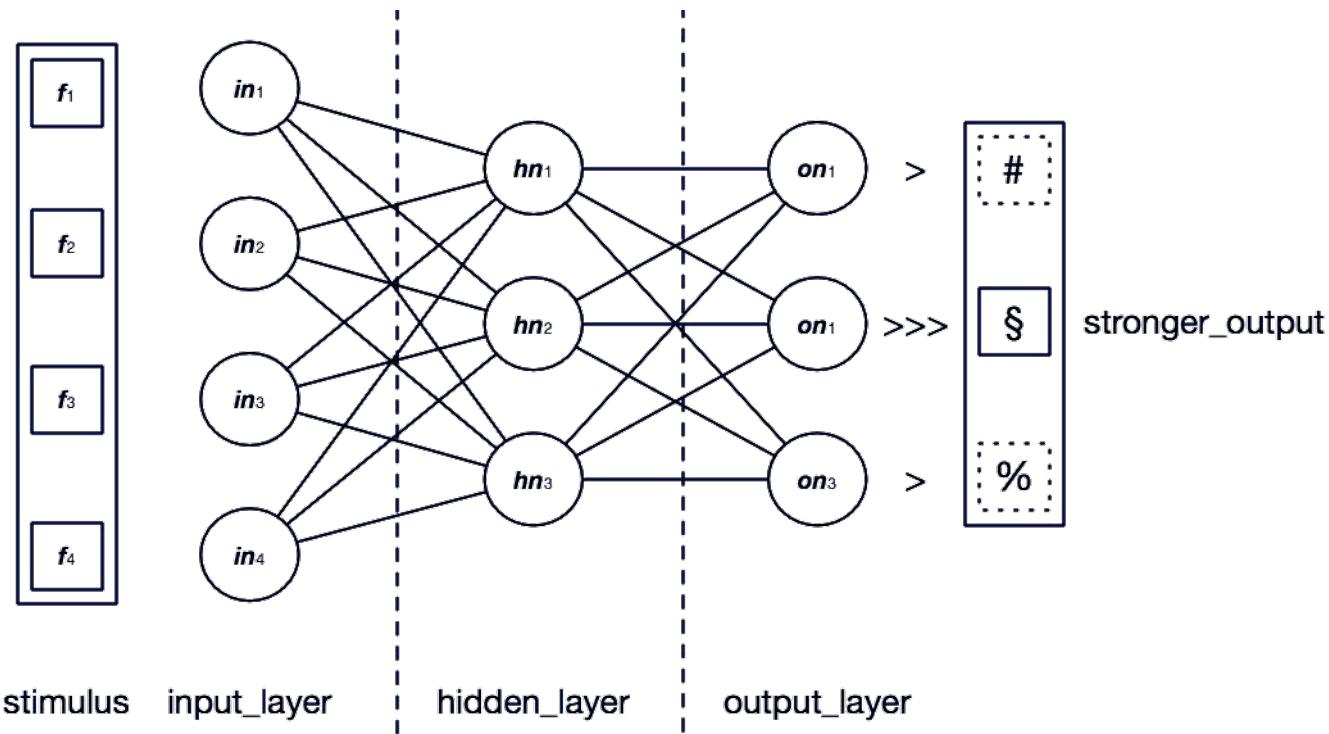
---



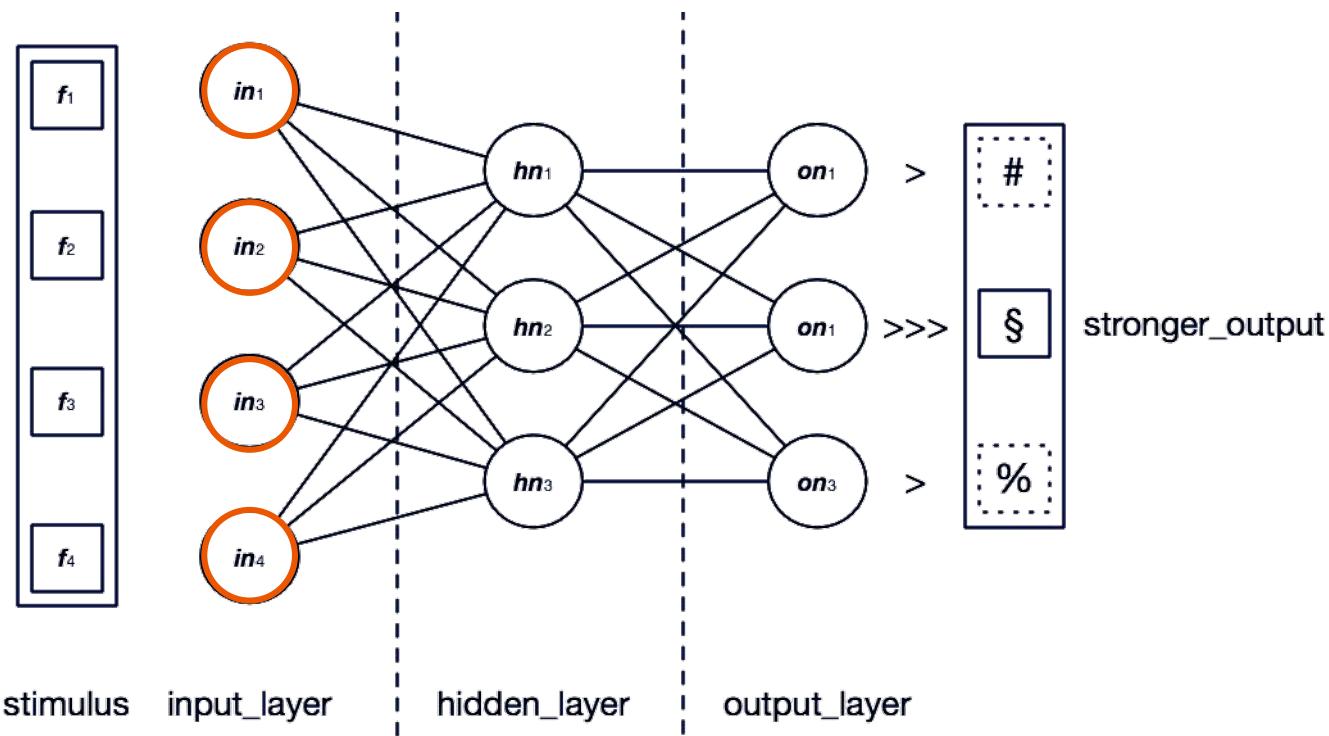
# Perceptron (neurônio)



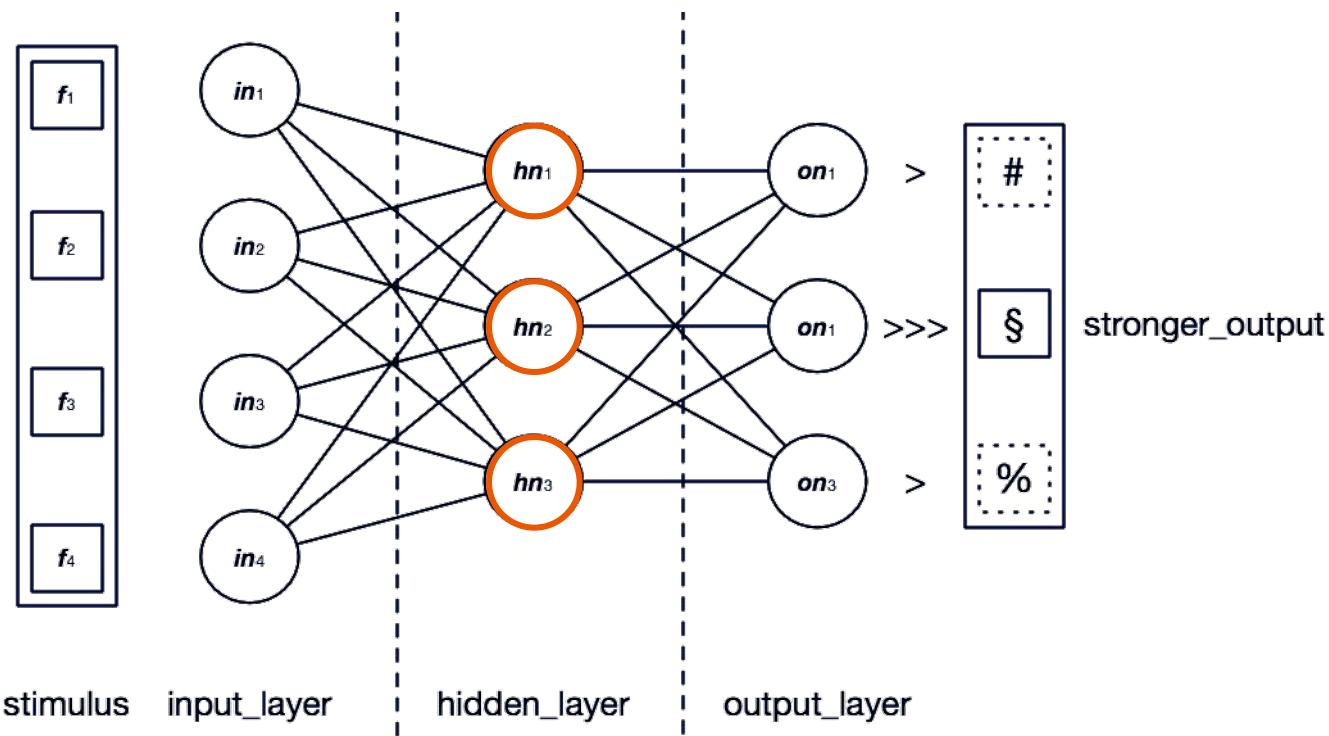
# Multi Layer Perceptron



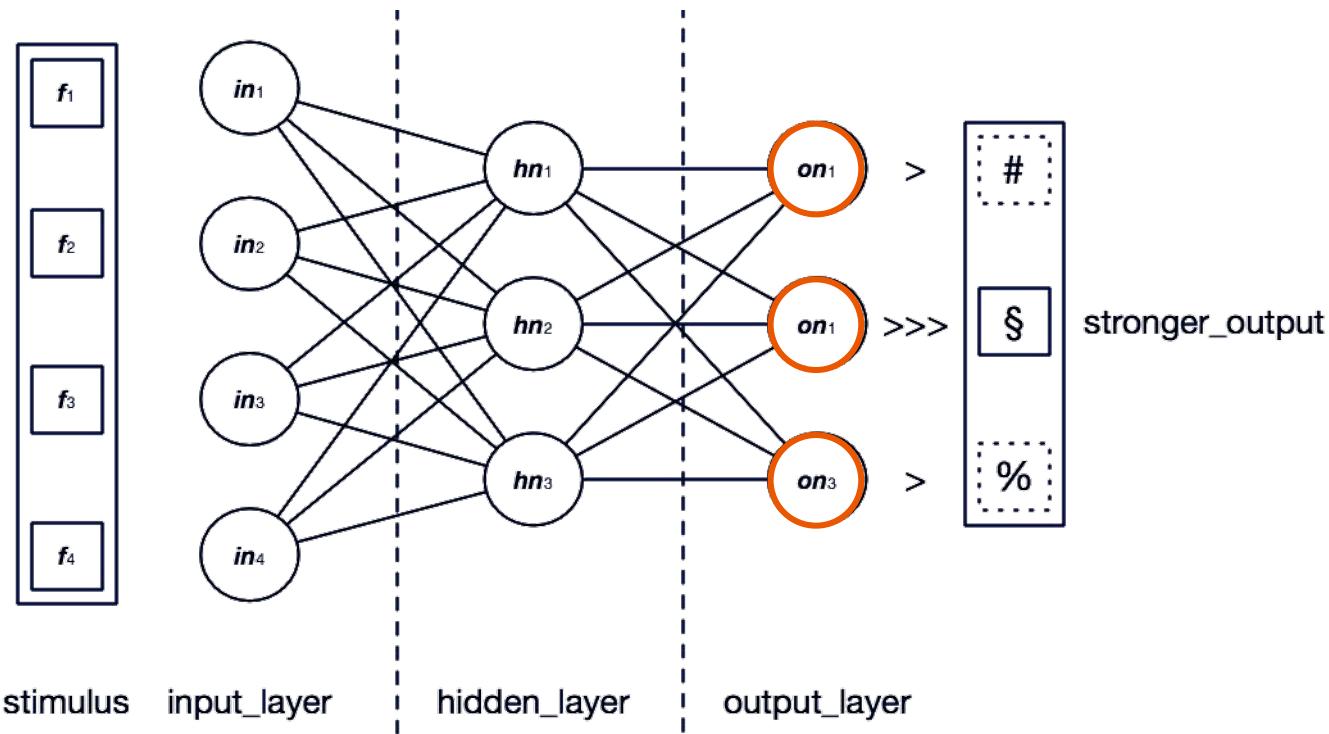
# Multi Layer Perceptron



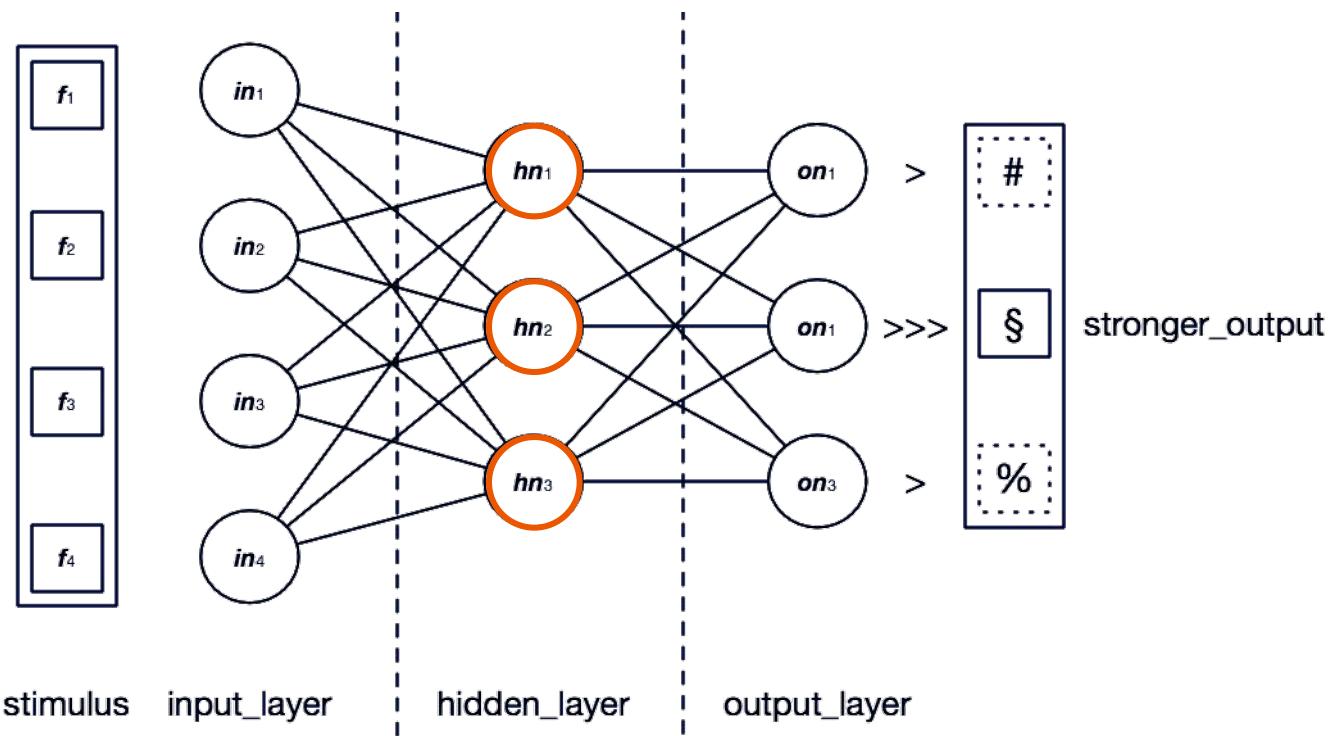
# Multi Layer Perceptron



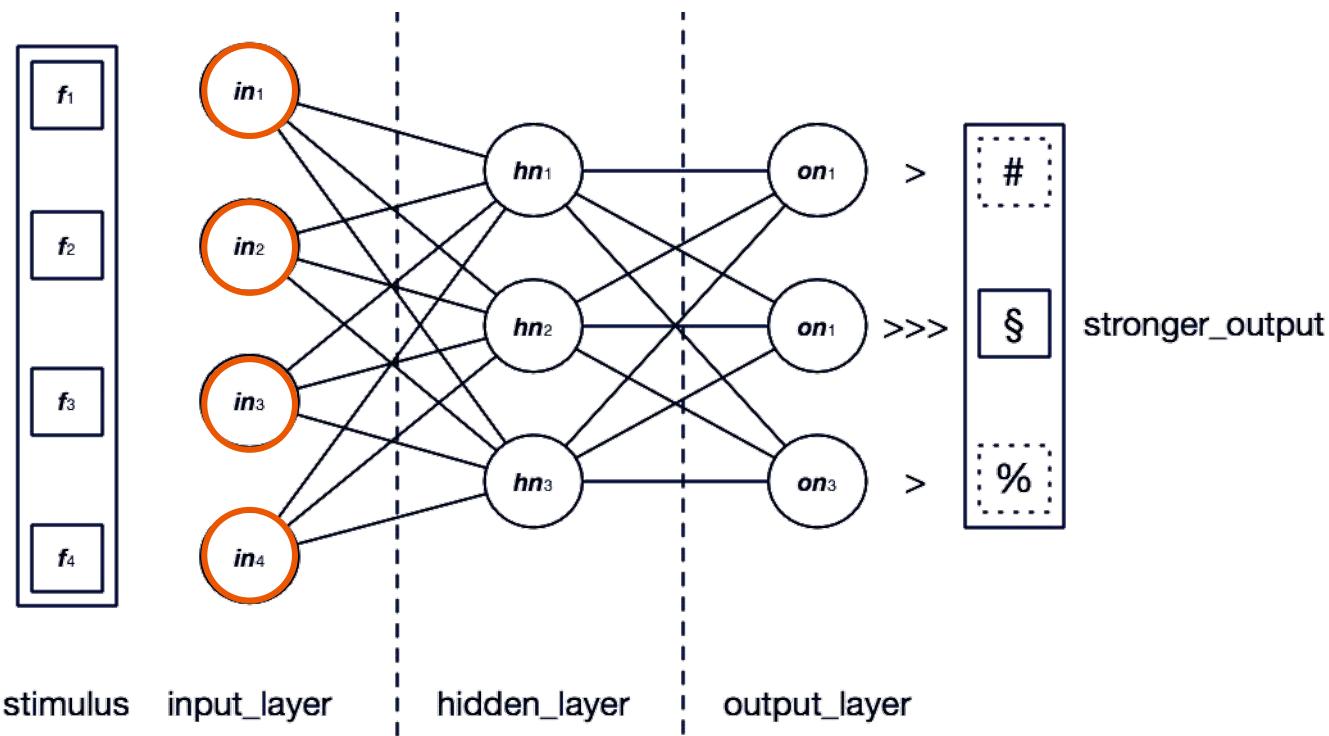
# Multi Layer Perceptron



# Multi Layer Perceptron



# Multi Layer Perceptron



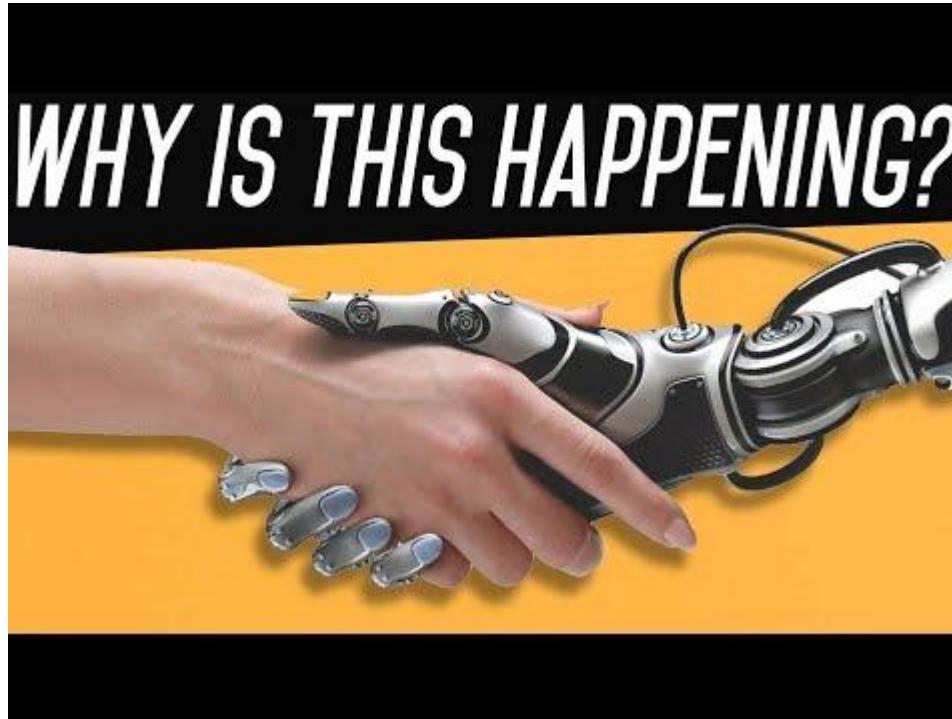
# Entendendo Redes Neurais

---

<https://playground.tensorflow.org>

# Do que as RNs são capazes?

---

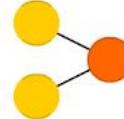


# Neural Networks

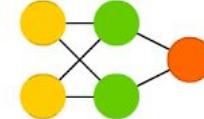
©2016 Fjodor van Veen - asimovinstitute.org

-  Backfed Input Cell
-  Input Cell
-  Noisy Input Cell
-  Hidden Cell
-  Probabilistic Hidden Cell
-  Spiking Hidden Cell
-  Output Cell
-  Match Input Output Cell
-  Recurrent Cell
-  Memory Cell
-  Different Memory Cell
-  Kernel
-  Convolution or Pool

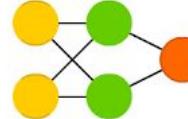
Perceptron (P)



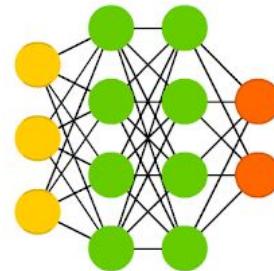
Feed Forward (FF)



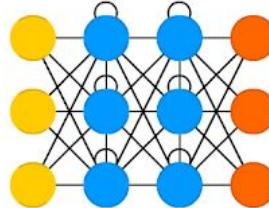
Radial Basis Network (RBF)



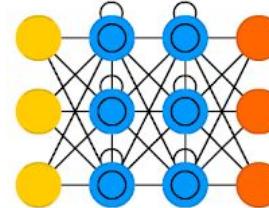
Deep Feed Forward (DFF)



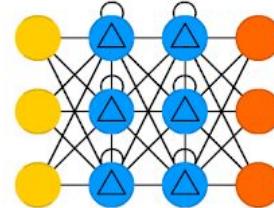
Recurrent Neural Network (RNN)



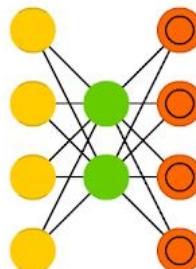
Long / Short Term Memory (LSTM)



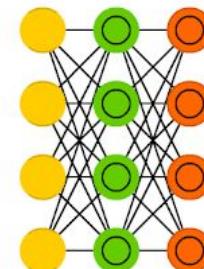
Gated Recurrent Unit (GRU)



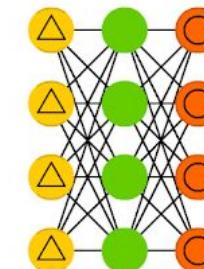
Auto Encoder (AE)



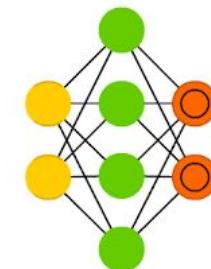
Variational AE (VAE)



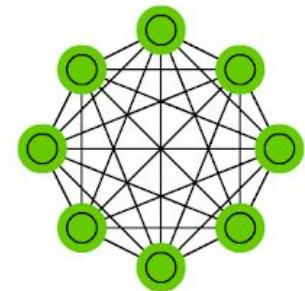
Denoising AE (DAE)



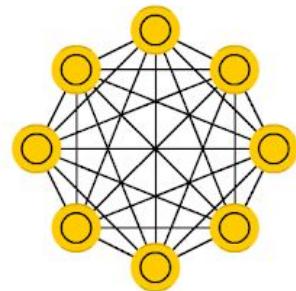
Sparse AE (SAE)



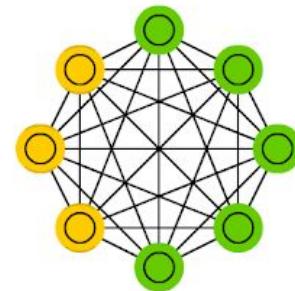
Markov Chain (MC)



Hopfield Network (HN)



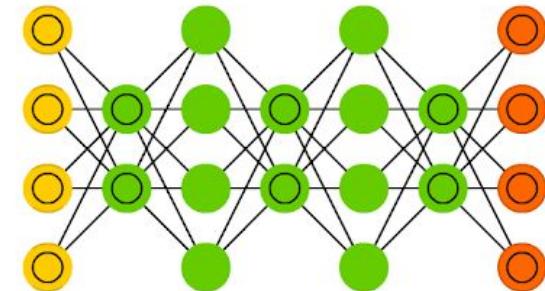
Boltzmann Machine (BM)



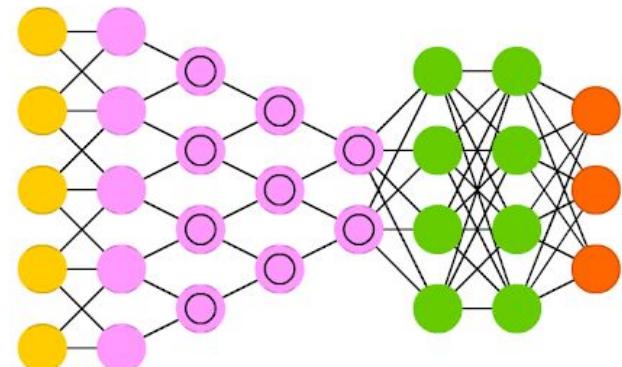
Restricted BM (RBM)



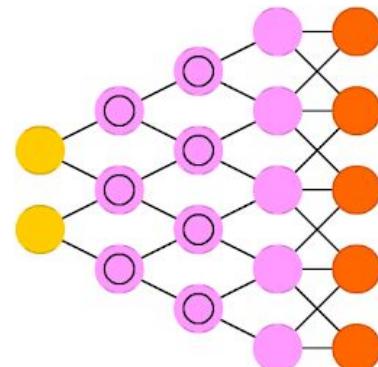
Deep Belief Network (DBN)



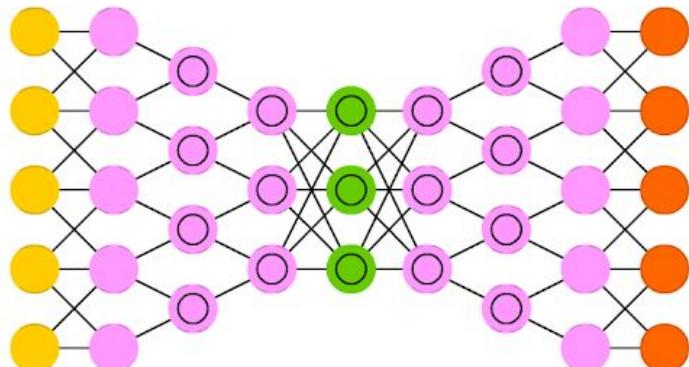
Deep Convolutional Network (DCN)



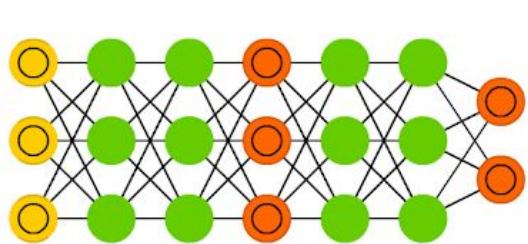
Deconvolutional Network (DN)



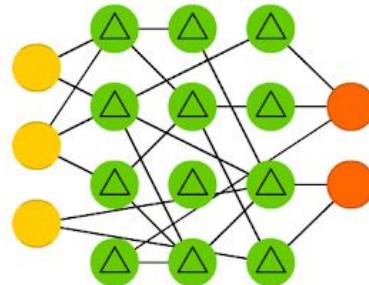
Deep Convolutional Inverse Graphics Network (DCIGN)



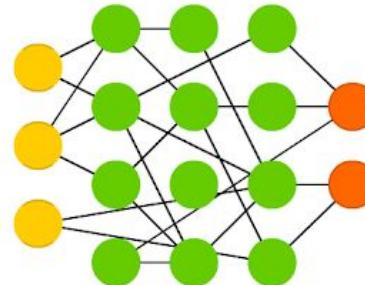
Generative Adversarial Network (GAN)



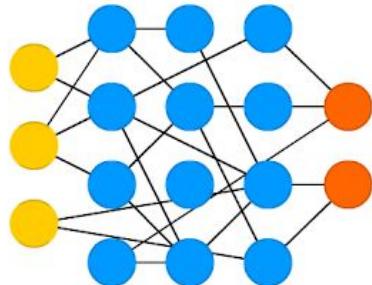
Liquid State Machine (LSM)



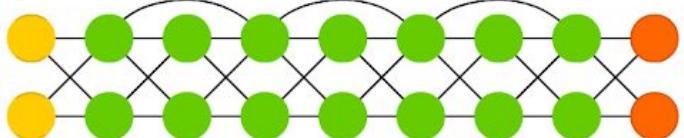
Extreme Learning Machine (ELM)



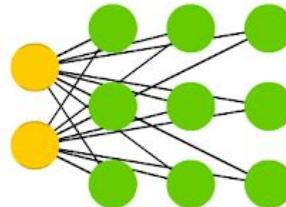
Echo State Network (ESN)



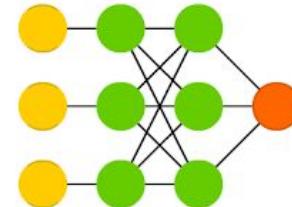
Deep Residual Network (DRN)



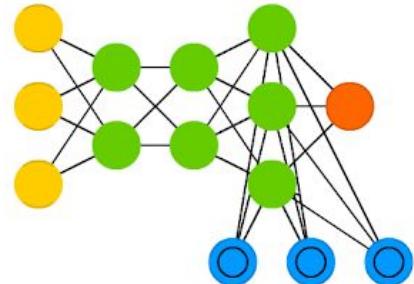
Kohonen Network (KN)



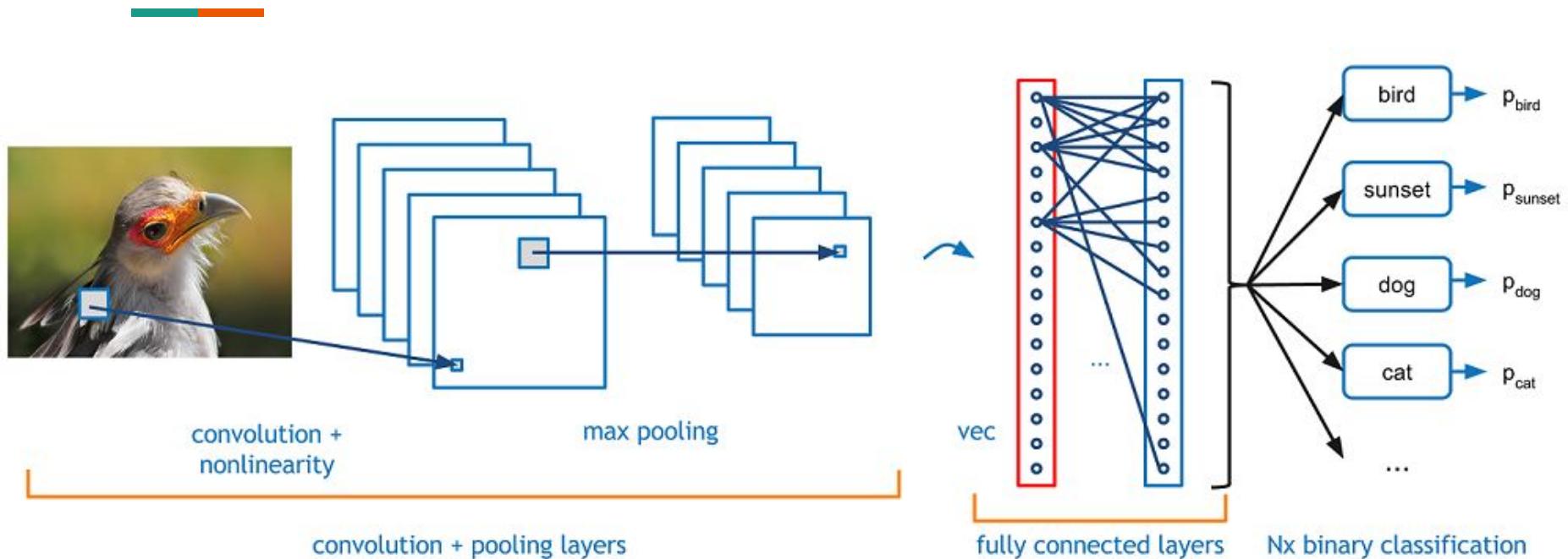
Support Vector Machine (SVM)



Neural Turing Machine (NTM)



# Redes Neurais Convolucionais



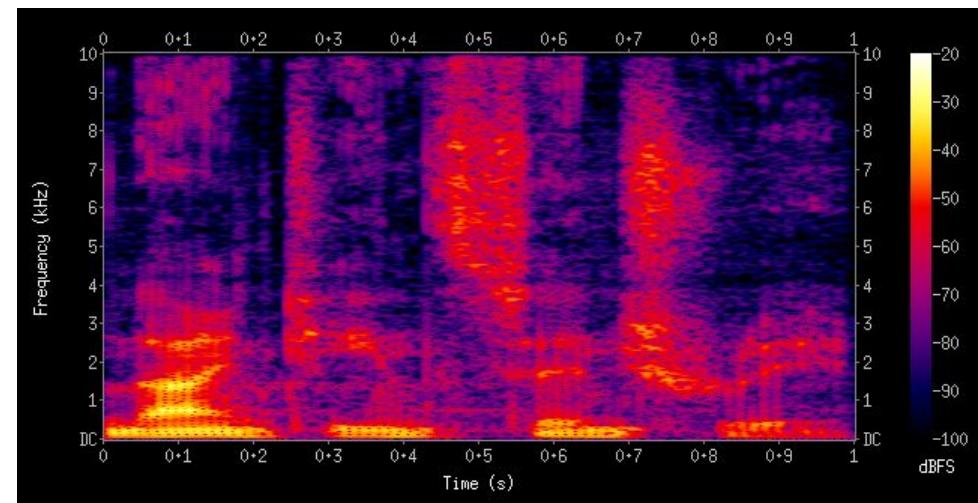
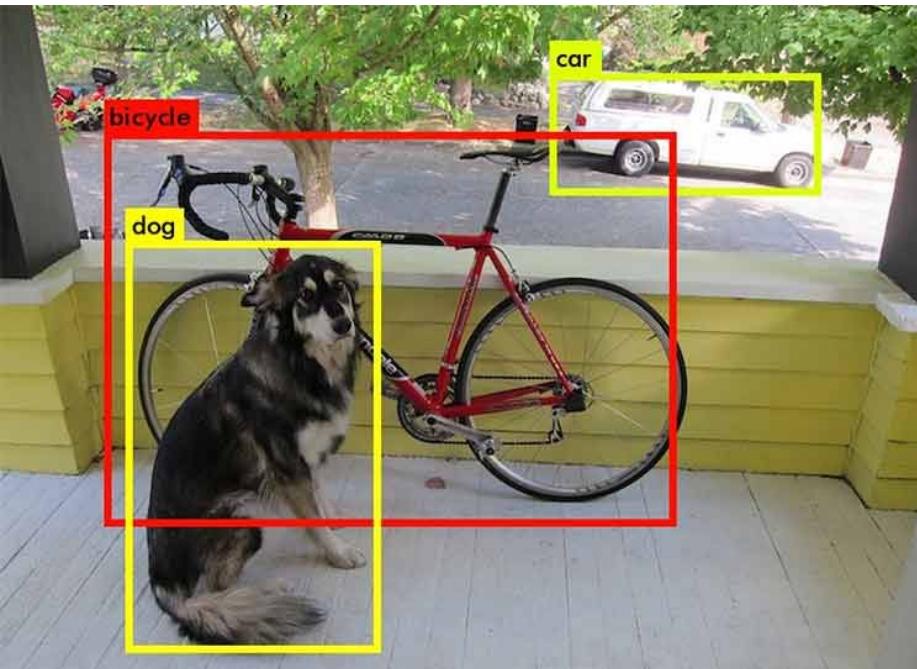
# Redes Neurais Convolucionais

---

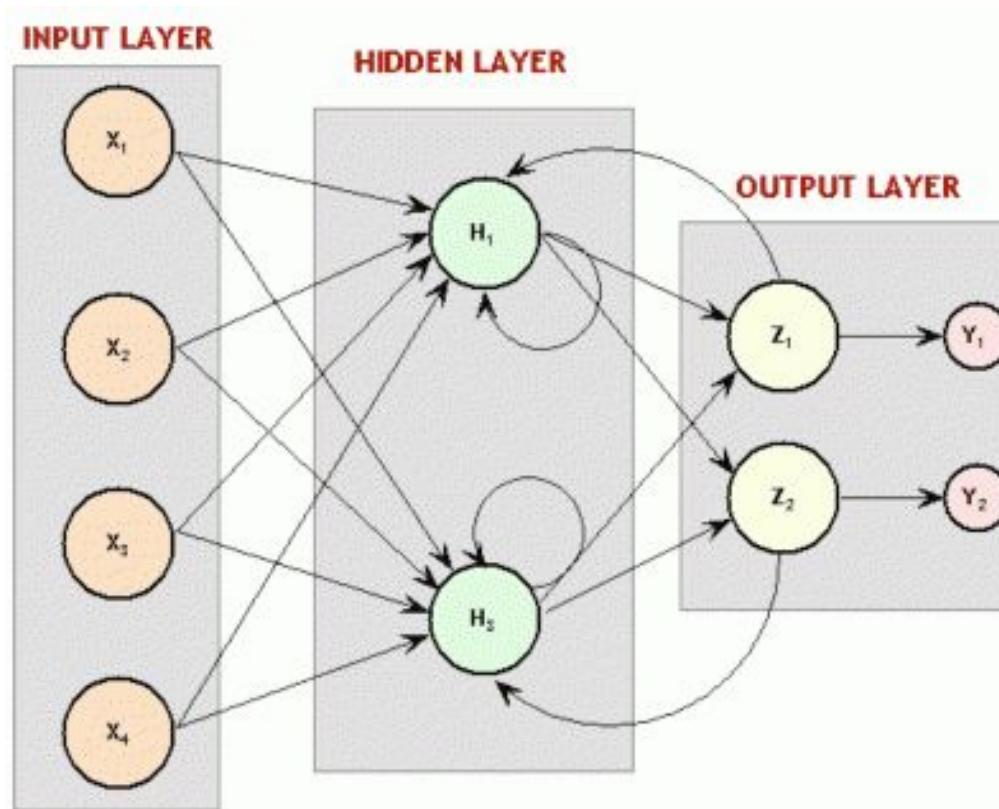


# Redes Neurais Convolucionais

---

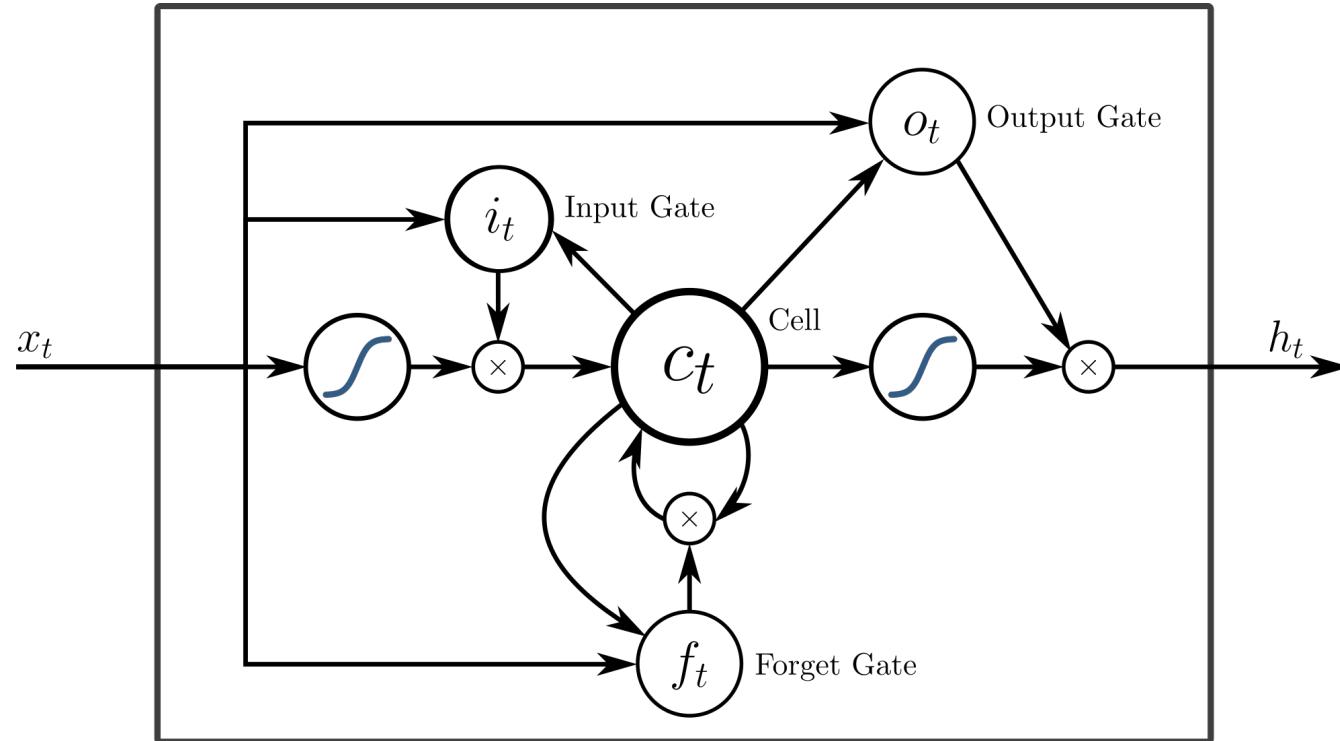


# Redes Neurais Recorrentes

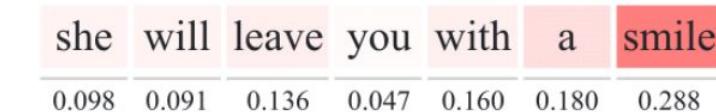
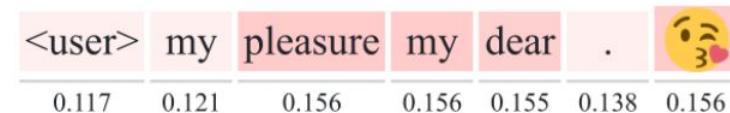
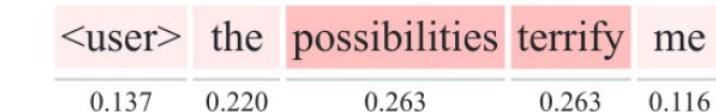
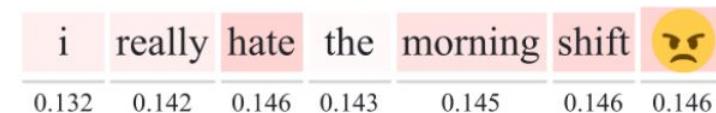
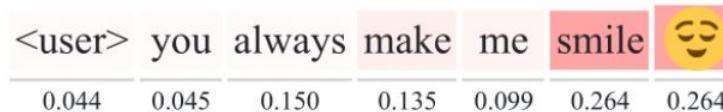
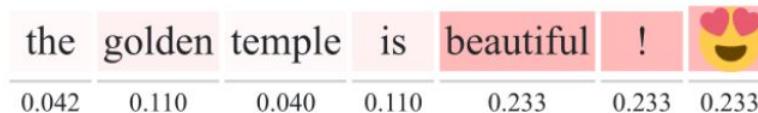
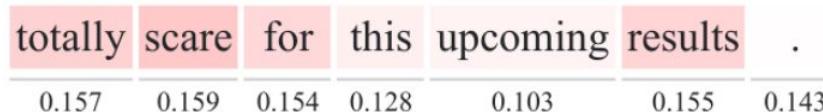
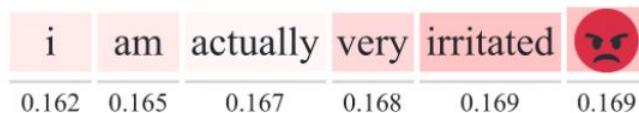


# Long-Short Term Memory (LSTM)

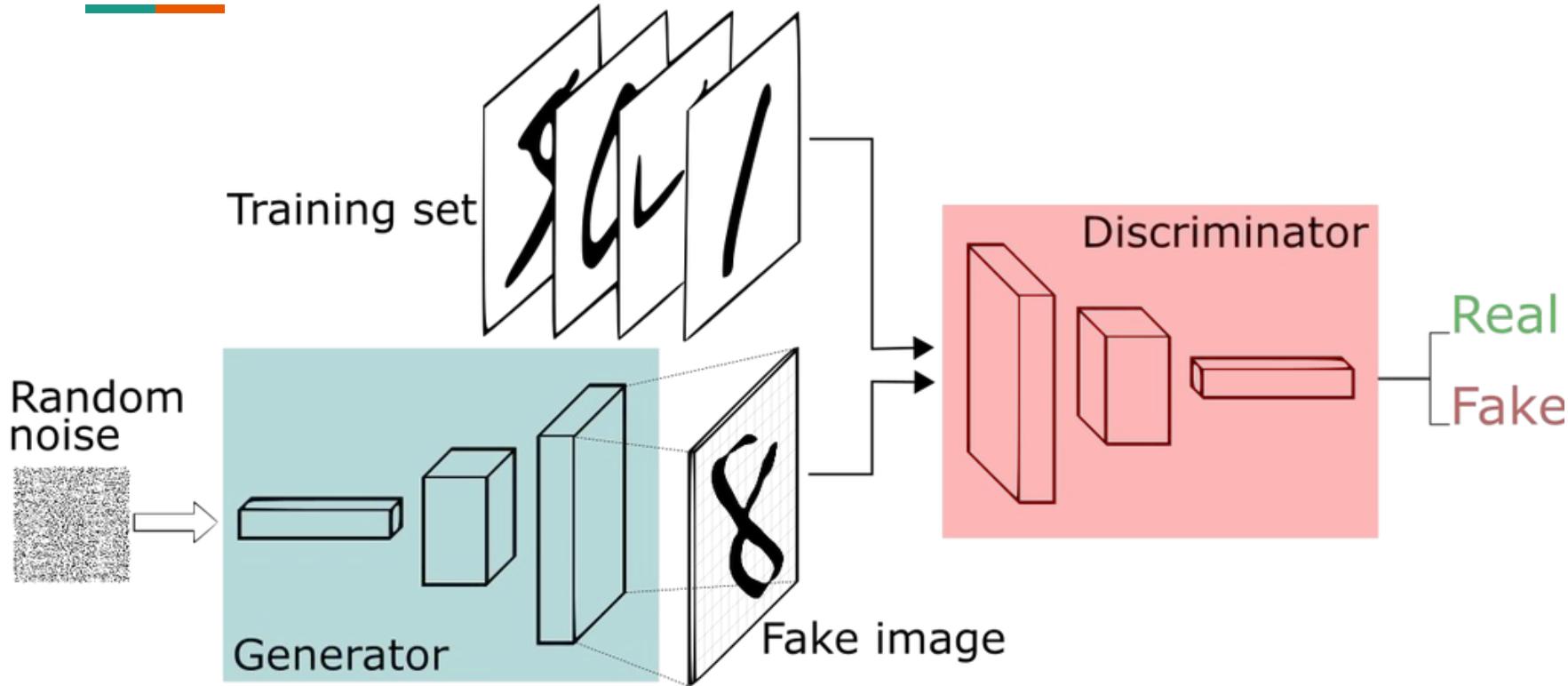
---



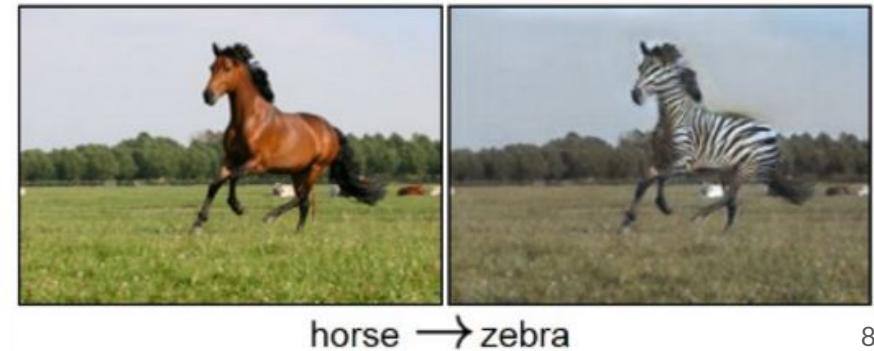
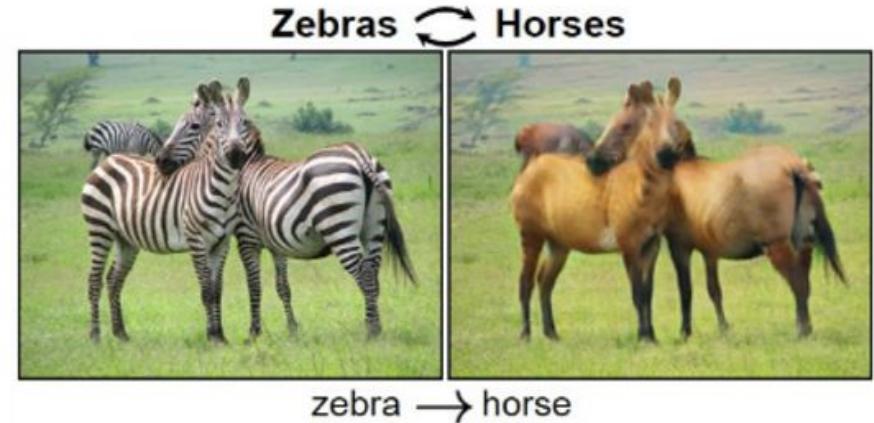
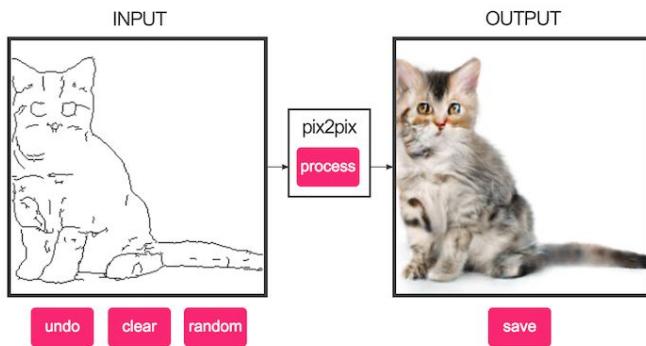
# Redes Neurais Recorrentes e LSTMs



# Generative Adversarial Network



# Generative Adversarial Network



# Testando Redes Neurais Código

---

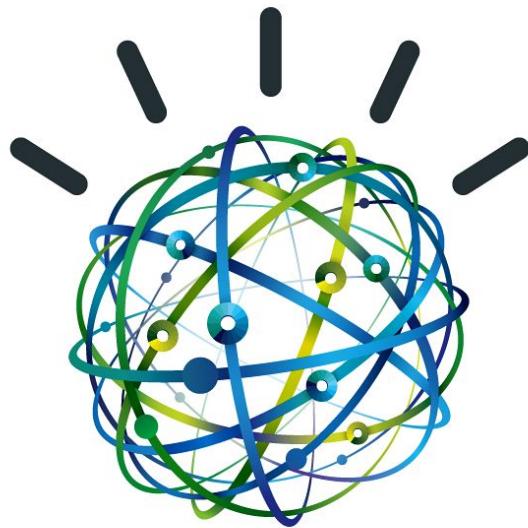
<http://goo.gl/vVULci>

# Limitações das NNs

---

- Grande quantidade de dados
- Poder de processamento (GPUs, TPUs)
- Entendimento da matemática (álgebra linear)

# Ferramentas de alto nível



## IBM Watson

# Ferramentas de alto nível

---



Google Cloud AI

Use your own data to train models



TensorFlow



Cloud Machine  
Learning Engine

Ready to use Machine Learning models



Cloud  
Vision API



Cloud  
Speech API



Cloud  
Jobs API



Cloud  
Translation  
API



Cloud  
Natural  
Language API



Cloud  
Video  
Intelligence API



Coming  
soon

# Ferramentas de alto nível

---



Firebase ML Kit  
(Mobile)

Recurso	No dispositivo	Nuvem
Reconhecimento de texto (OCR, na sigla em inglês)	✓	✓
Detecção facial	✓	✓
Leitura de código de barras	✓	
Marcação de imagens		✓
Reconhecimento de logotipos		✓
Reconhecimento de pontos de referência		✓
Detecção de conteúdo explícito		✓
Pesquisa de imagem semelhante		✓
Inferência de modelo personalizado	✓	

# Ferramentas de alto nível



Azure AI



## Image classification with convolutional neural networks

Explore transfer learning, convolutional neural networks, and gradient-boosting decision tree algorithms.

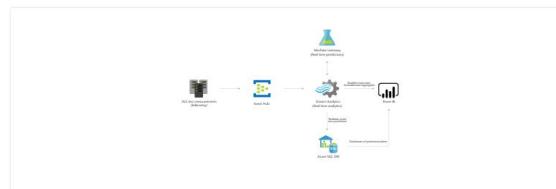
[Learn more >](#)



## Information discovery with deep learning and natural language processing

See how deep learning and natural language processing can be used effectively with the Microsoft AI platform.

[Learn more >](#)



## Defect prevention with predictive maintenance

Learn how to use Azure Machine Learning to predict failures before they happen with real-time assembly line data.

[Learn more >](#)



## Enterprise Productivity Chatbot

Azure Bot Service can be easily combined with Language Understanding to build powerful enterprise productivity bots, allowing organizations to streamline common work activities by integrating external systems, such as Office 365 calendar, customer cases stored in Dynamics CRM and much more.

[Learn more >](#)



# Obrigado!

Henrique Dias

Doutorando PUCRS

[henrique.santos.003@acad.pucrs.br](mailto:henrique.santos.003@acad.pucrs.br)

[PDF dos Slides](#)



# Curso de Extensão - PUCRS

---

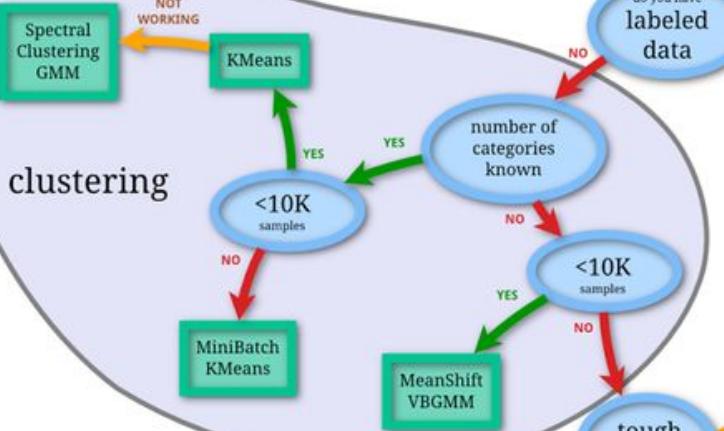
1º Encontro	2º Encontro	3º Encontro	4º Encontro	5º Encontro
Conceitos Básicos Tipos de Aprendizagens	Classificação Regressão Avaliação de Modelos	Pré-Processamento Agrupamento Redução de Dimensionalidade Regras de Associação Agrupamento	Redes Neurais Processamento da Linguagem Natural Séries Temporais	Ciência de Dados nas Núvens

# scikit-learn algorithm cheat-sheet

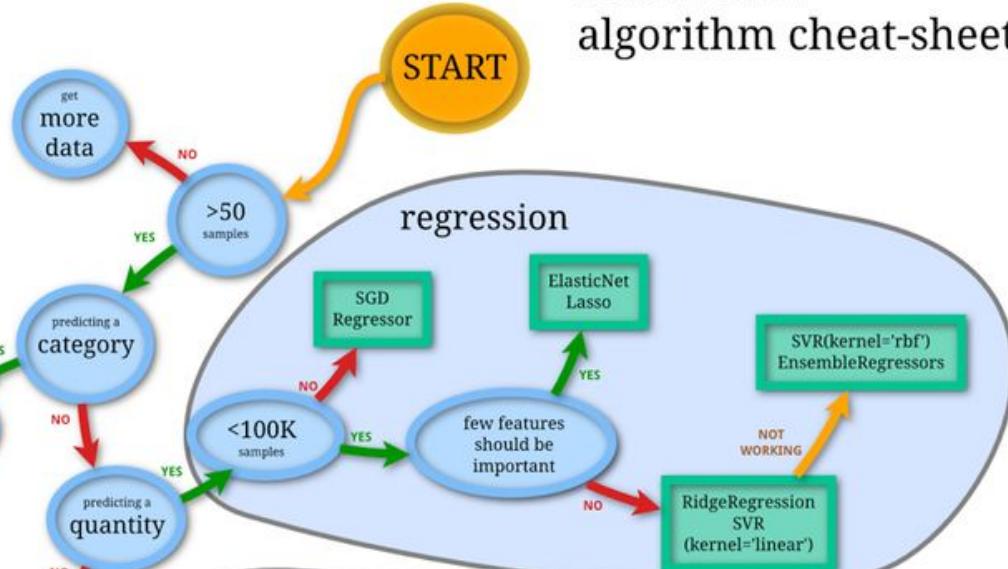
## classification



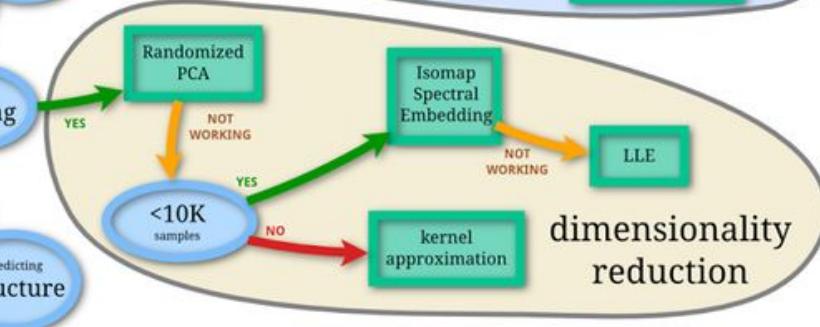
## clustering



## regression

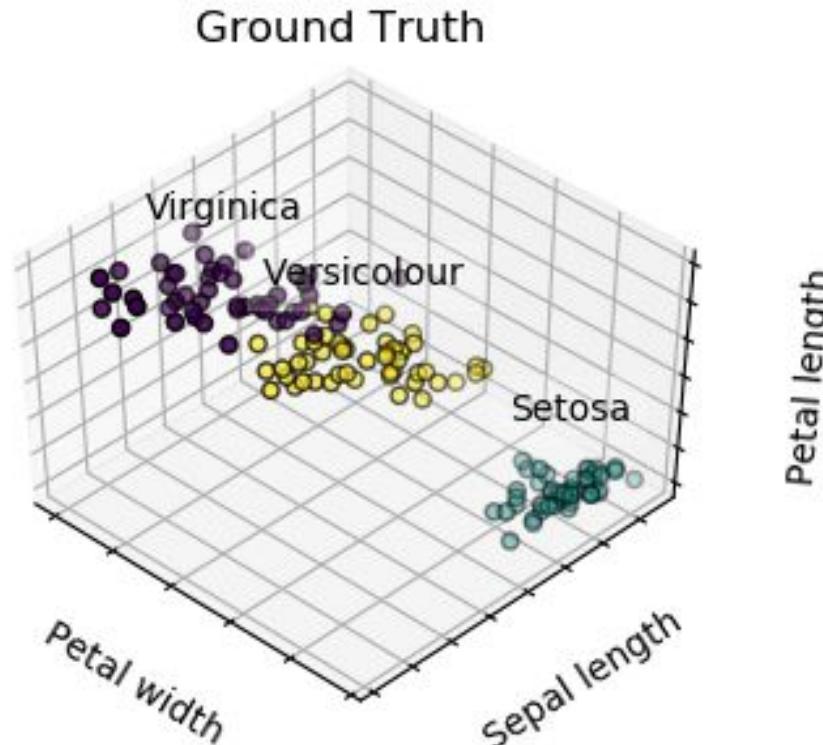


## dimensionality reduction



# Clusterização: Iris

---



# Classificação: Problemas x Algoritmos

---

- A. Câncer de Mama
- B. Dígitos (Imagen)
- C. Vinho (3 classes)
- D. Iris (3 classes)
- 1. K-Neighbors
- 2. Neural Network
- 3. Árvore de Decisão
- 4. Suport Vector Machine
- 5. Naive Bayes
- 6. Random Forest

# Classificação: Problemas x Algoritmos

---

## `sklearn.datasets.`

- A. `load_breast_cancer()`
- B. `load_digits()`
- C. `load_wine()`
- D. `load_iris()`

## `sklearn.`

- 1. `neural_network.MLPClassifier`
- 2. `neighbors.KNeighborsClassifier`
- 3. `svm.LinearSVC`
- 4. `tree.DecisionTreeClassifier`
- 5. `ensemble.RandomForestClassifier`
- 6. `naive_bayes.GaussianNB`

# Notebooks da Oficina

---

- Notebook Básico
- Métricas e Regressão Notebook
- Detecção de Outliers

Para todos os links do Colab:

- ao clicar no link, no Google Drive, ir em “Abrir com: Colaboratory”
- dentro do Google Colab, ir em “File”, depois em “Save a copy in Google Drive”
- assim será possível modificar o arquivo e rodar os experimentos

# Outros Notebooks

---

- Regressão e Busca Exaustiva
- Processamento da Linguagem Natural
- Redes Neurais

Para todos os links do Colab:

- ao clicar no link, no Google Drive, ir em “Abrir com: Colaboratory”
- dentro do Google Colab, ir em “File”, depois em “Save a copy in Google Drive”
- assim será possível modificar o arquivo e rodar os experimentos

# Outras Palestras

---

- Ciência de Dados na Saúde (vídeo)
- Data Science para Publicidade
- Aprendizado Não-Supervisionado e o PageRank (vídeo)

# Mais material:

---

- <https://github.com/amueller/scipy-2017-sklearn>
- [https://github.com/amueller/introduction to ml with python](https://github.com/amueller/introduction_to_ml_with_python)
- <http://scikit-learn.org/stable/documentation.html>
- [https://colab.research.google.com/notebooks/basic features overview.ipynb](https://colab.research.google.com/notebooks/basic_features_overview.ipynb)
- <https://www.datascienceacademy.com.br/pages/cursos-gratis>
- <https://br.udacity.com/course/intro-to-data-science--ud359>

# Vídeos no YouTube

---

- [Big Data - Nerdologia](#)
- [Ciência de Dados - Nerdologia](#)
- [Machine Learning - Nerdologia](#)
- [Linguística Forense - Unabomber](#)
- [Robotização Eleitorial - Estadão](#)
- [Redes Neurais - Nerdologia](#)
- [O futuro do seu emprego - Nerdologia](#)
- [Aprendizado por Reforço - AlphaGo](#)
- [AlphaGo Zero - DeepMind](#)
- [Profissional do Futuro - TED Talks](#)
- [Why Deep Learning Now? - ColdFusion](#)
- [AlphaStar - DeepMind](#)

# Livros sobre Ciência de Dados **(para leigos)**

---

- Super Chunchers, Ian Ayres
- Numeratis, Stephen Baker

# Datasets para Experimentos

---

- <https://www.kaggle.com/>
- <https://www.openml.org/>
- <https://toolbox.google.com/datasetsearch>
- <https://mimic.physionet.org>