Data Science Aplicado à Economia e Negócios

PUCRS - Fevereiro de 2019

http://goo.gl/BgoJmm

Slides: http://goo.gl/BgoJmm

Edições Anteriores (350 participantes)

poa.hub

StarGrid

PUCRS

(PPGCC e PPGGB)

TECNOPUC EXPERIENCE

memed

woop

Sicredi

Google I/O 18 Extended

ft18

TECNOPUC

(Setembro e Dezembro)

Slides: http://goo.gl/BgoJmm

Público-Alvo

Devs

Linguistas

Jornalista

Biólogxs

Engenheirxs

Advogadxs

Prof. Saúde

Estudante

...

Slides: http://goo.gl/BgoJmm

Check-In Sympla

apresentação dos participantes

Slides: http://goo.gl/BgoJmm

Facilitador

Henrique

PUCRS

Grupo de
Inteligência Artificial
na Saúde

LARA

Objetivo

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

Sopa de Letrinhas

Ferramentas

Machine Learning

scikit learn

K + TensorFlow

jupyter

Big Data

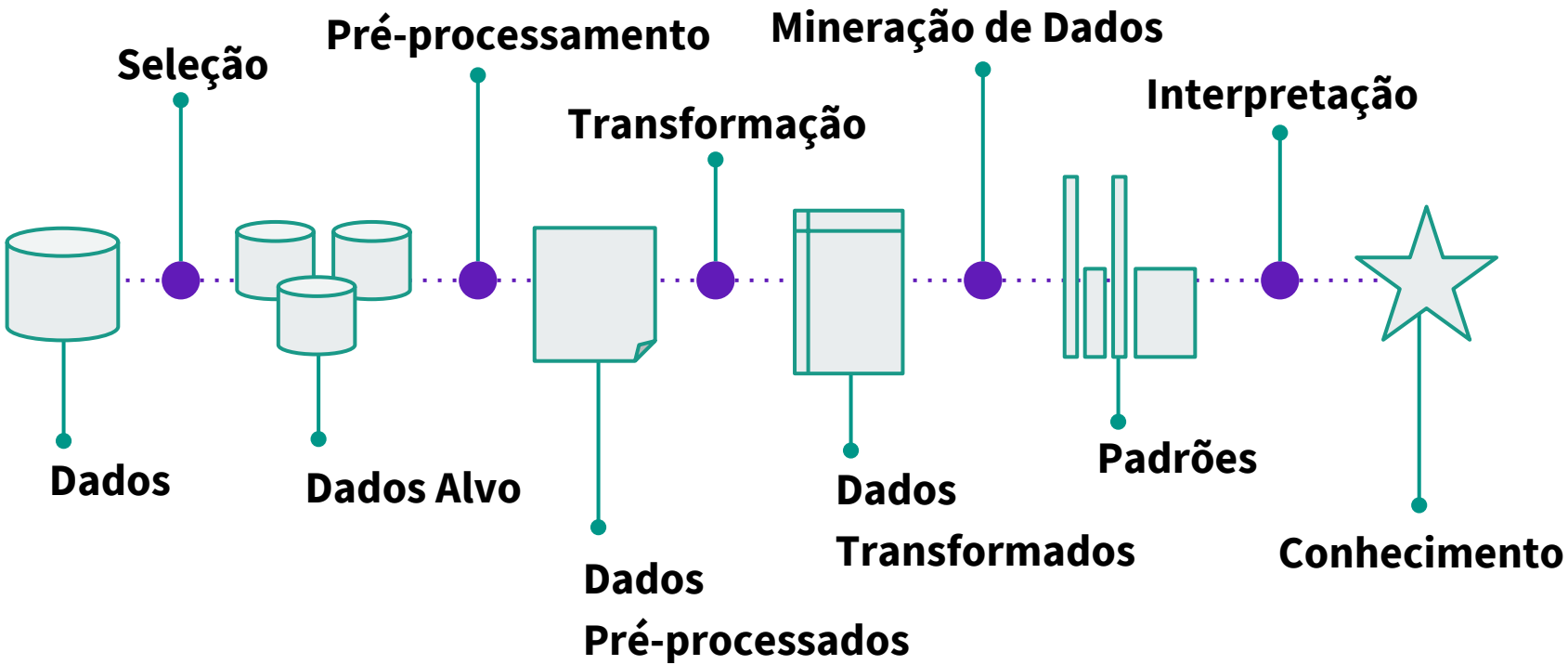
APACHE Spark™

hadoop

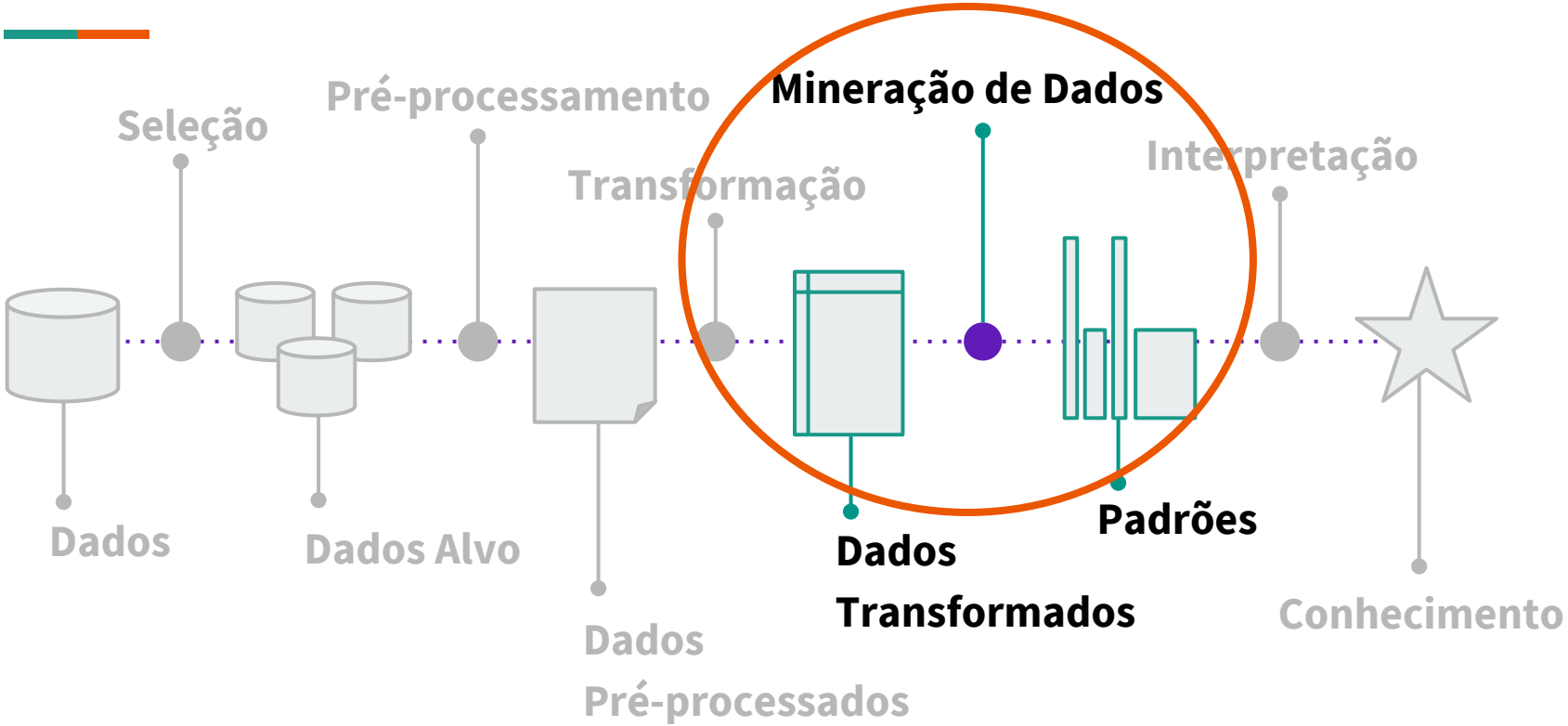
mongoDB

HIVE

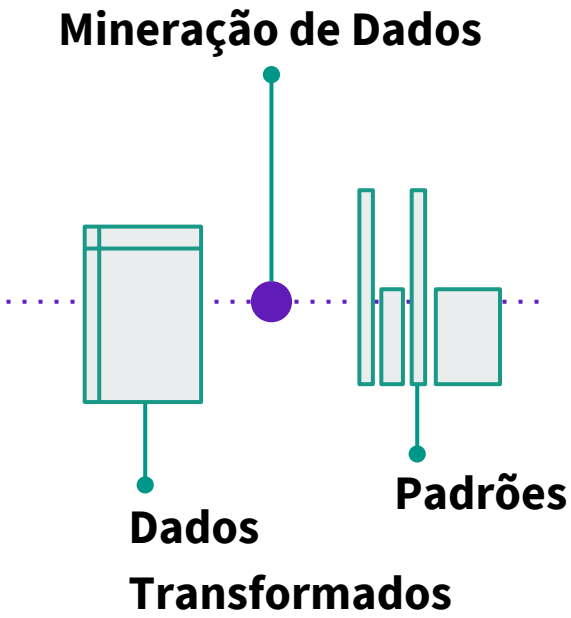
Descoberta de Conhecimento



Descoberta de Conhecimento

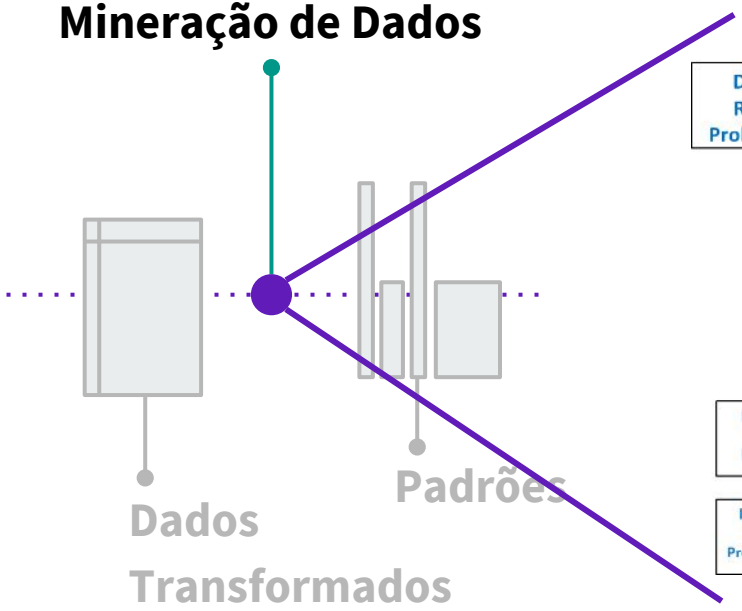


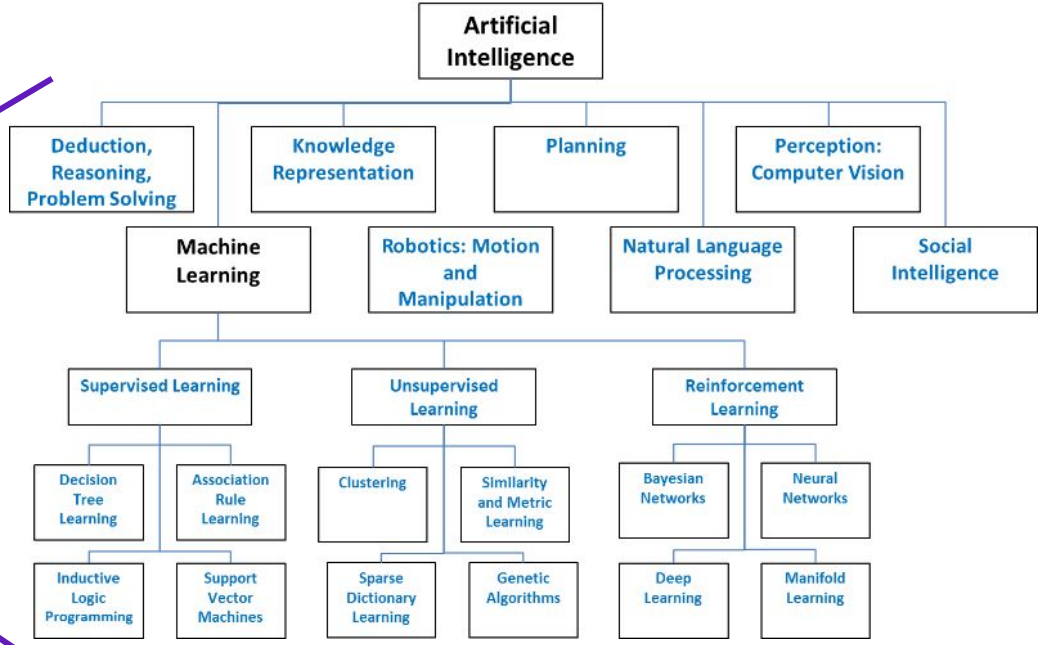
Mineração de Dados



Slides: http://goo.gl/BgoJmm

Aprendizado de Máquina

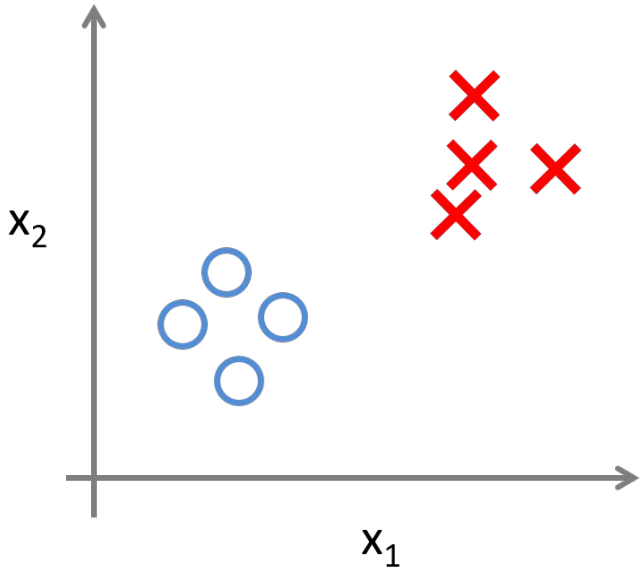




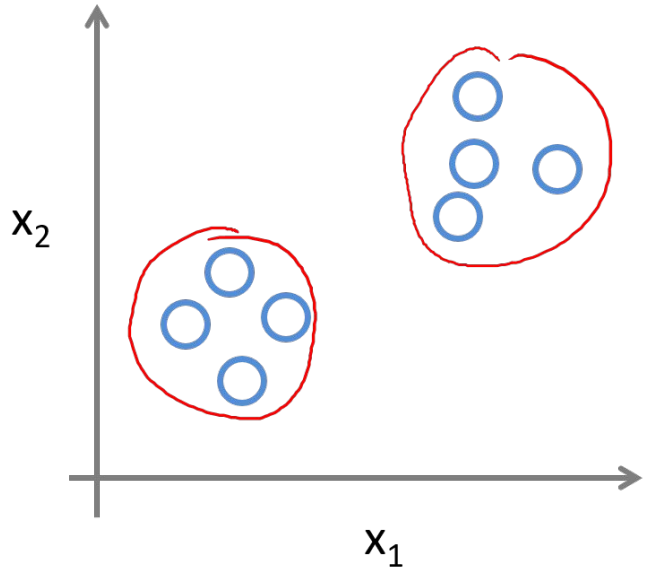
Tipos de Aprendizagem

Aprendizado de Máquina

Supervised Learning



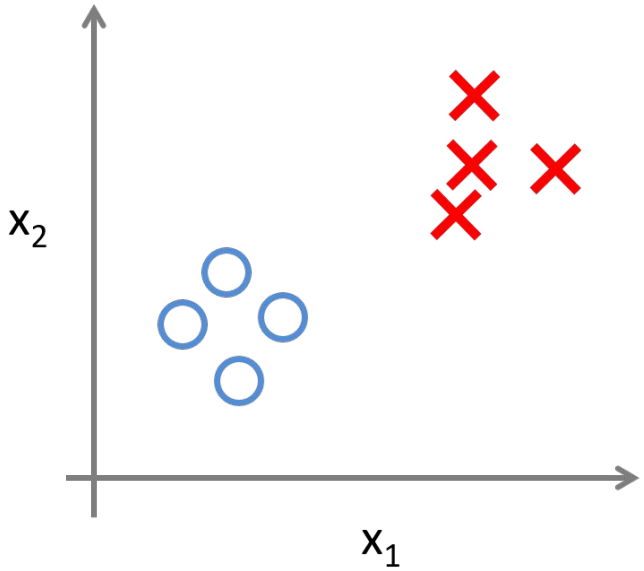
Unsupervised Learning



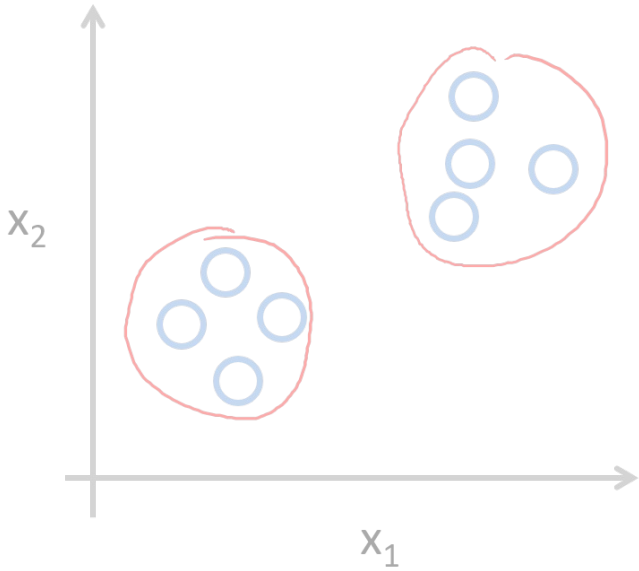
Imagens: https://leonardoaraujosantos.gitbooks.io

Supervisionado

Supervised Learning

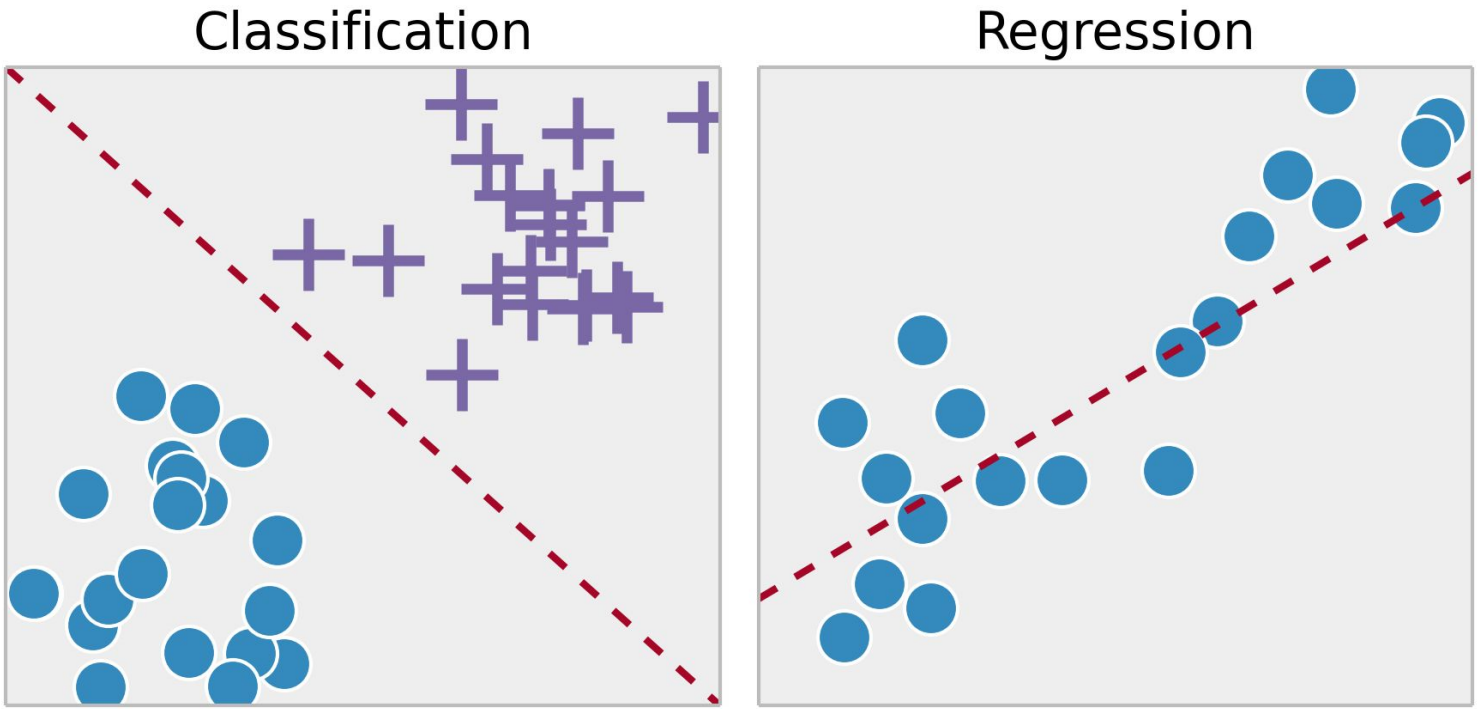


Unsupervised Learning



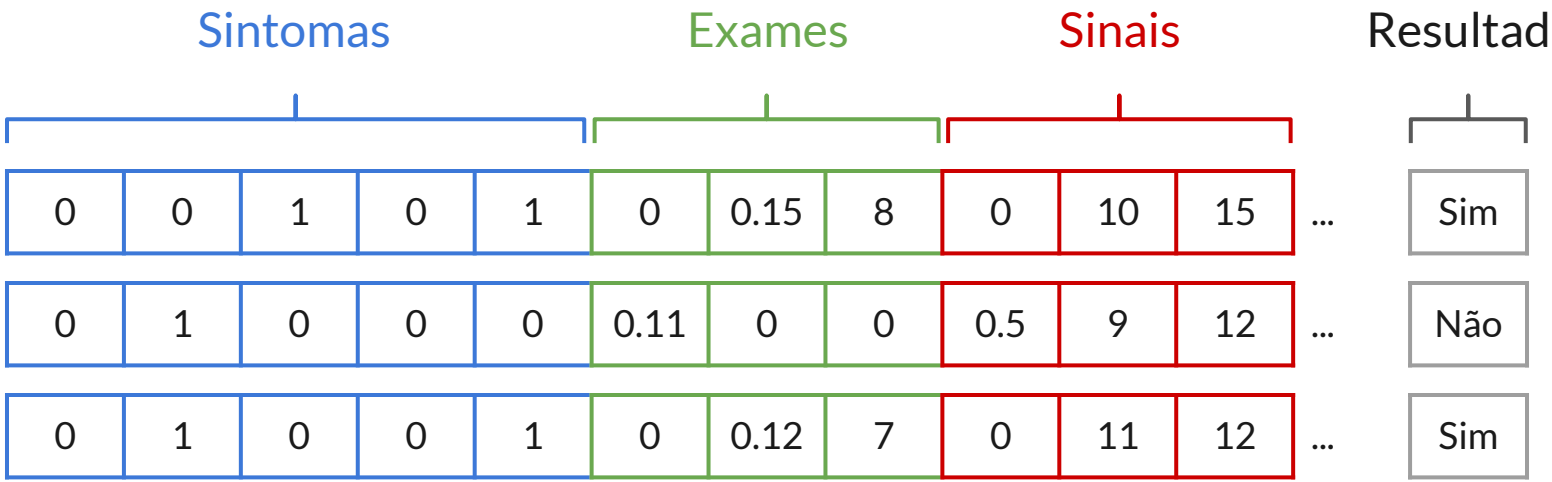
Imagens: https://leonardoaraujosantos.gitbooks.io

Supervisionado

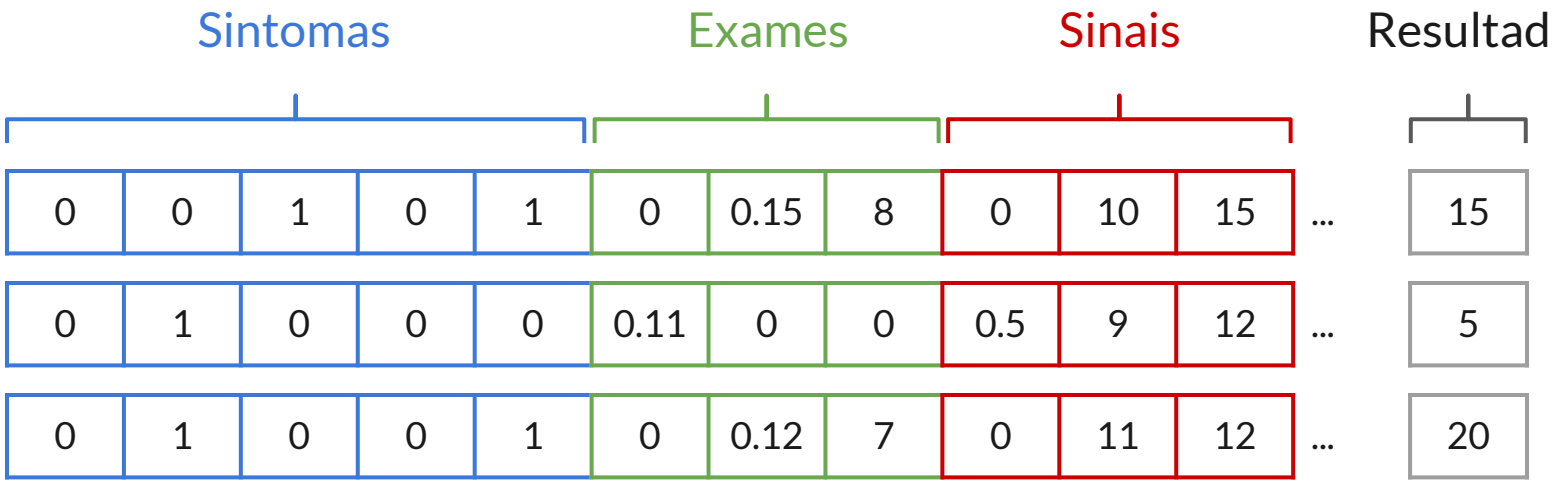


Imagens: https://leonardoaraujosantos.gitbooks.io

Classificação (attributos)

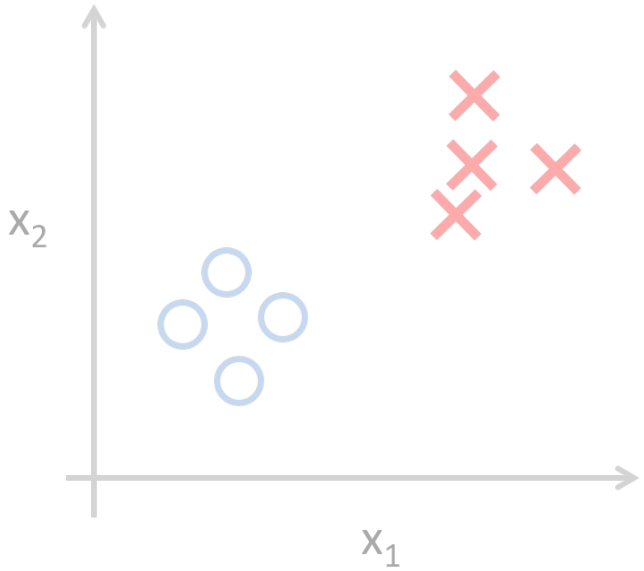


Regressão (attributos)

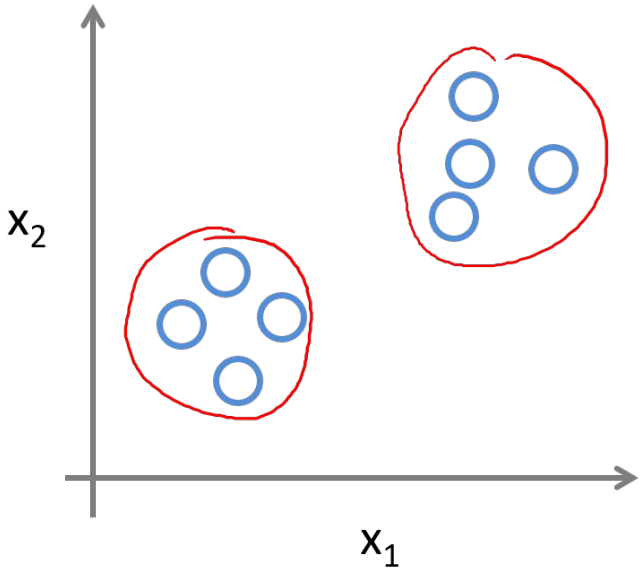


Não-Supervisionado

Supervised Learning

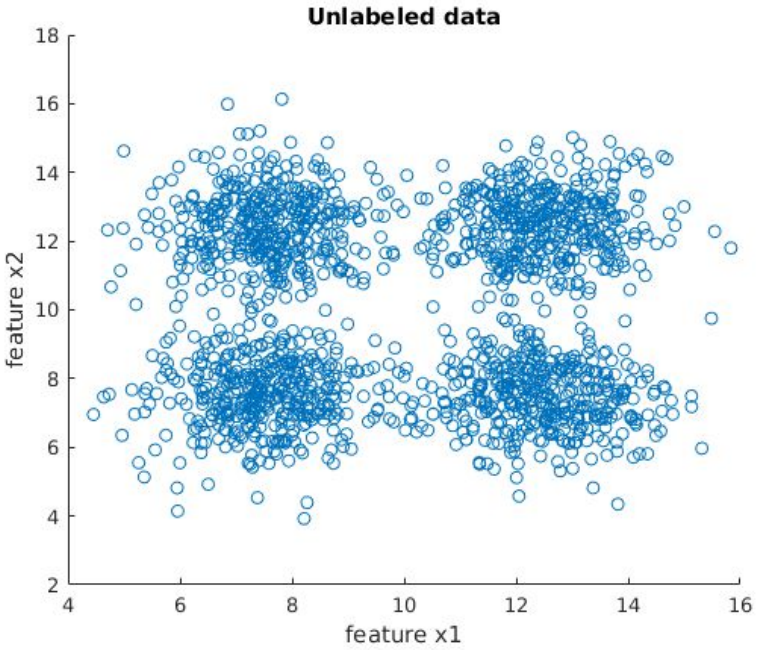


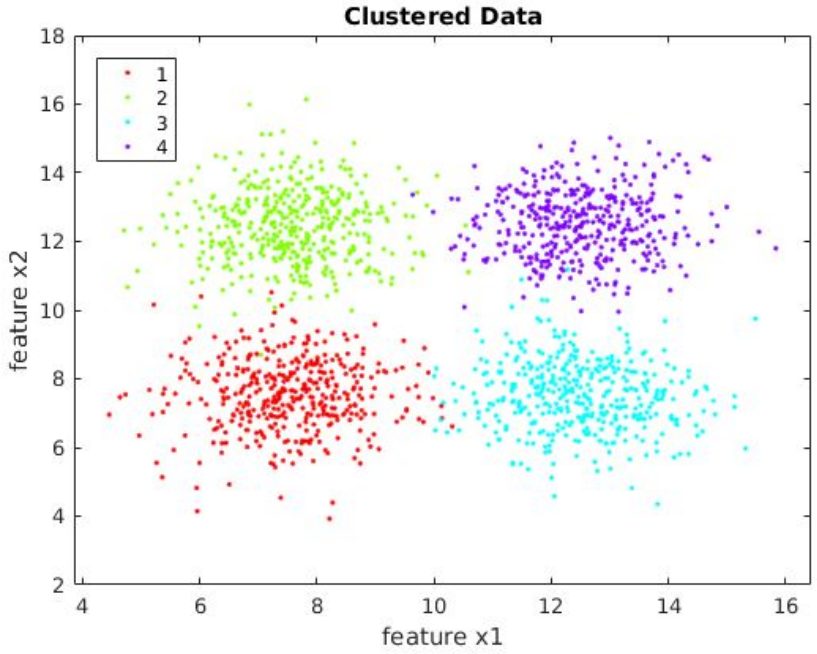
Unsupervised Learning



Imagens: https://leonardoaraujosantos.gitbooks.io

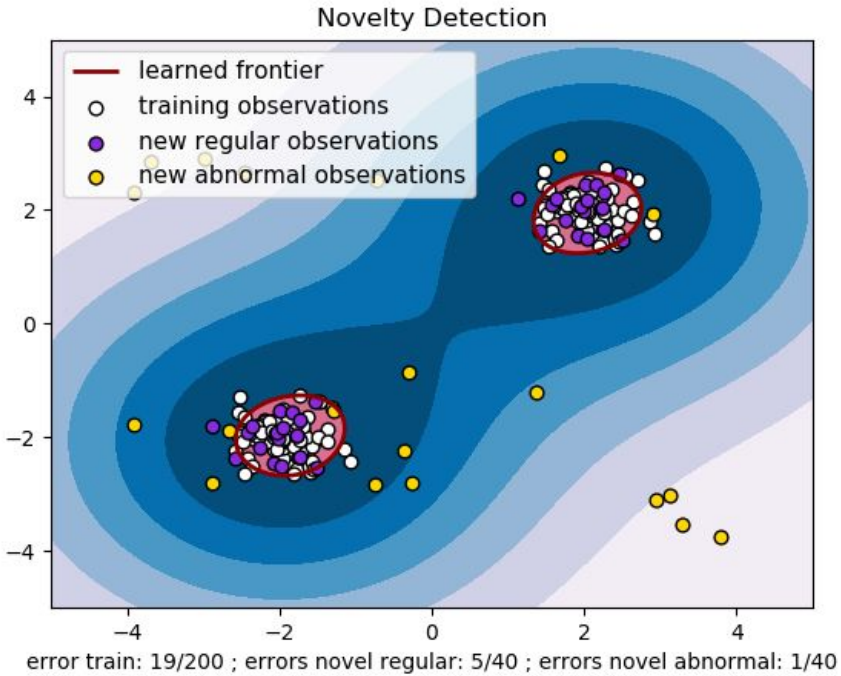
Não-Supervisionado Agrupamento

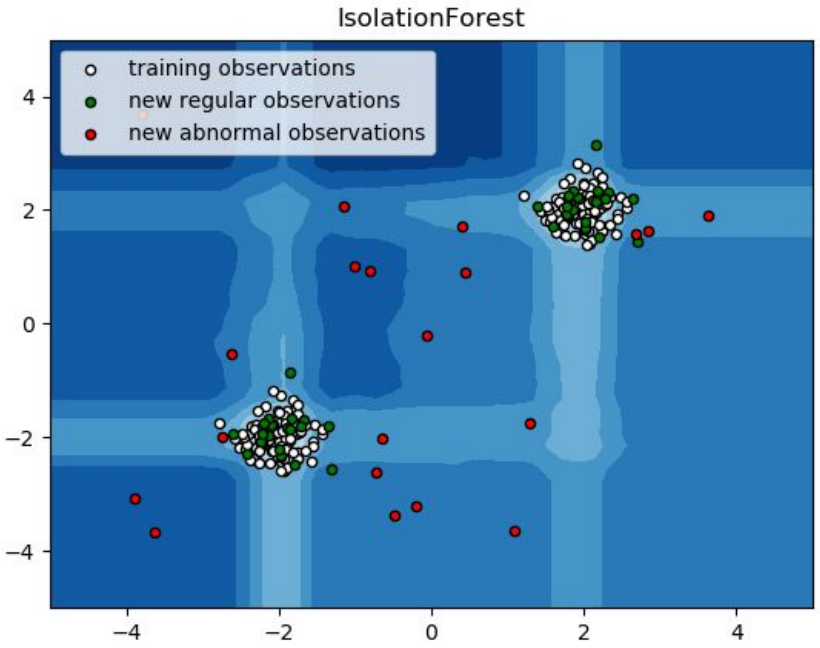




Imagens: https://leonardoaraujosantos.gitbooks.io

Não-Supervisionado Detecção de Anomalia





Imagens: http://scikit-learn.org/stable/modules/outlier_detection.html

Não-Supervisionado Regras de Associação

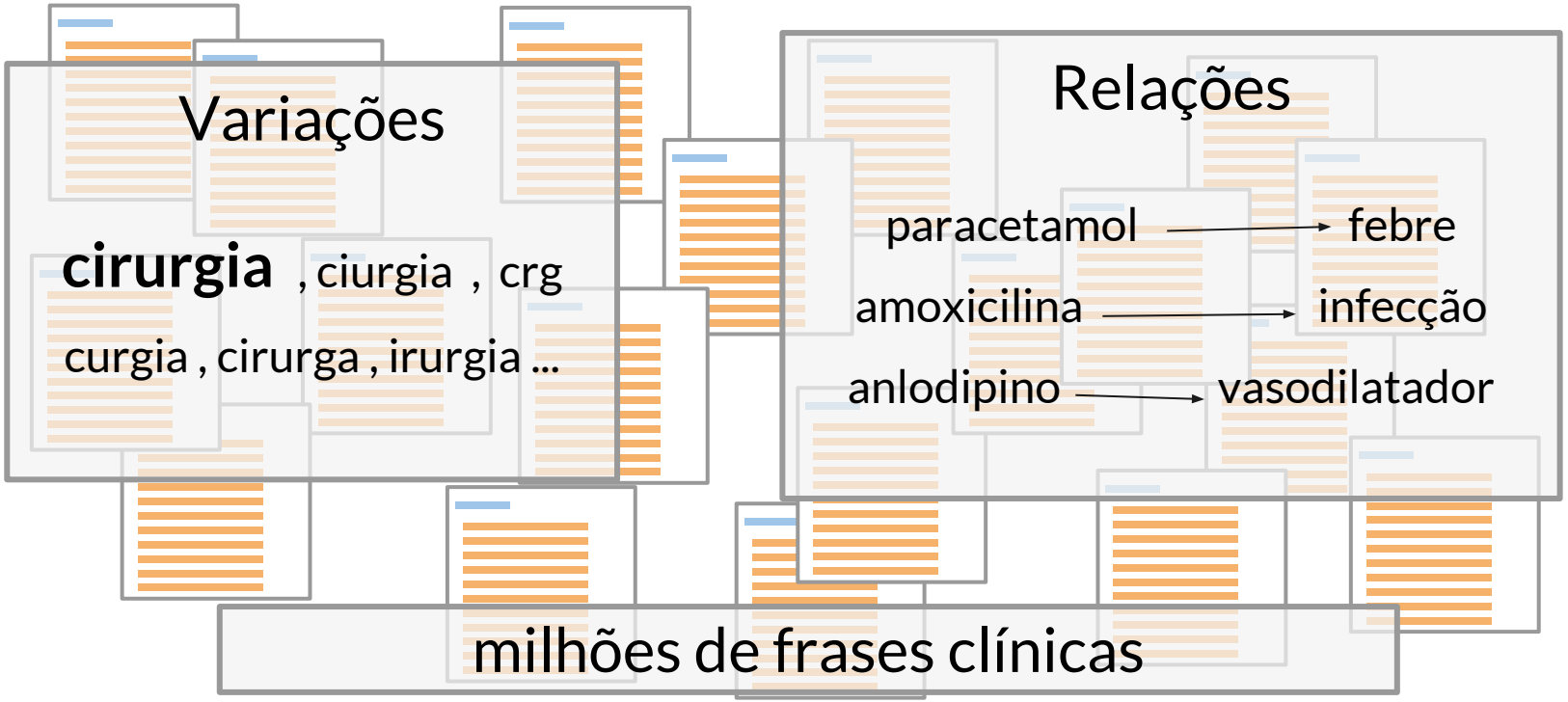
Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B\&C \Rightarrow D$	1/5	1/3	5/9

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

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

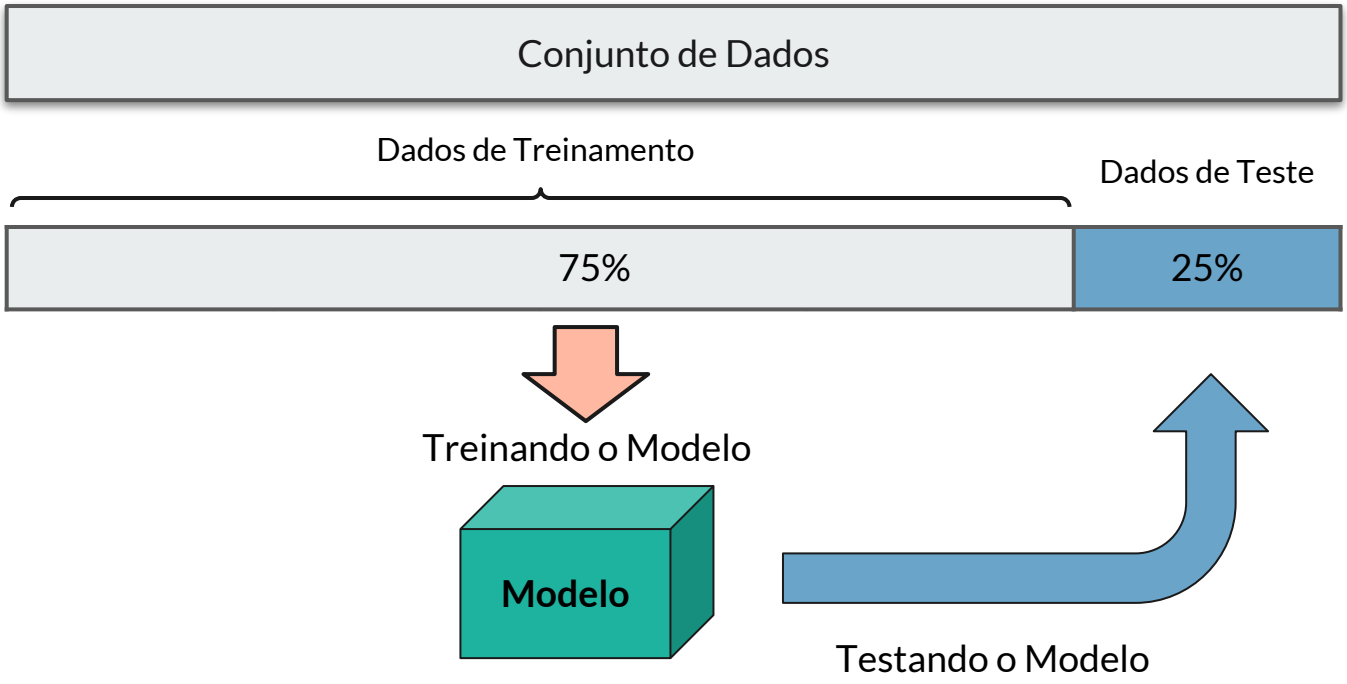
$$Lift = \frac{Support}{Support(X) \times Support(Y)}$$

Não-Supervisionado Word Embeddings

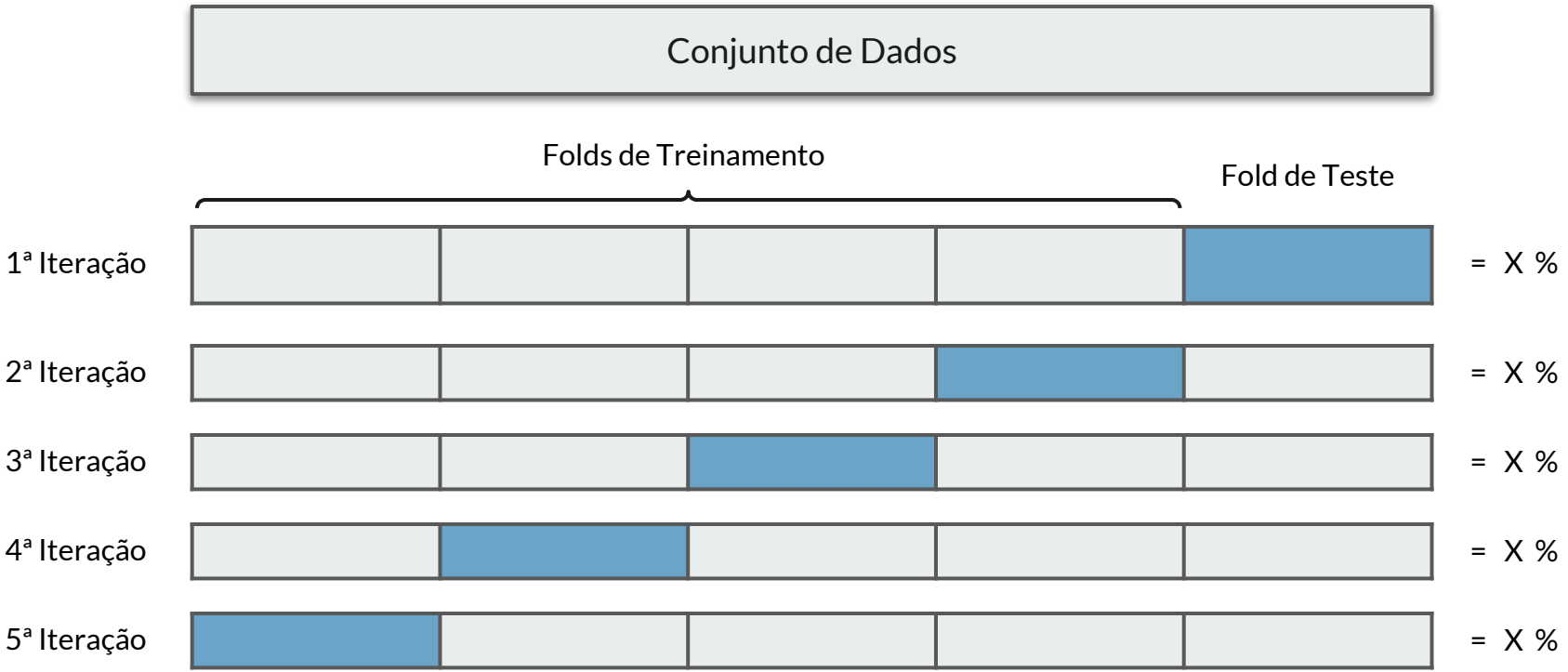


Avaliação dos Modelos

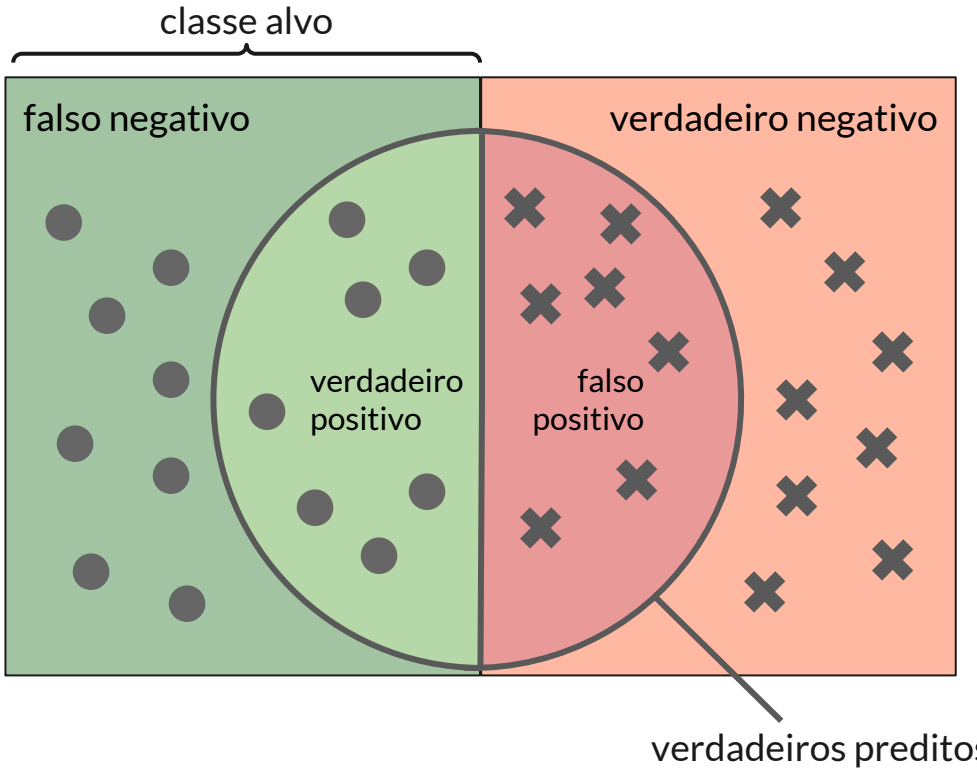
Treino e Teste



Validação Cruzada (5 Folds)



Acurácia, Precisão e Abrangência Classificação

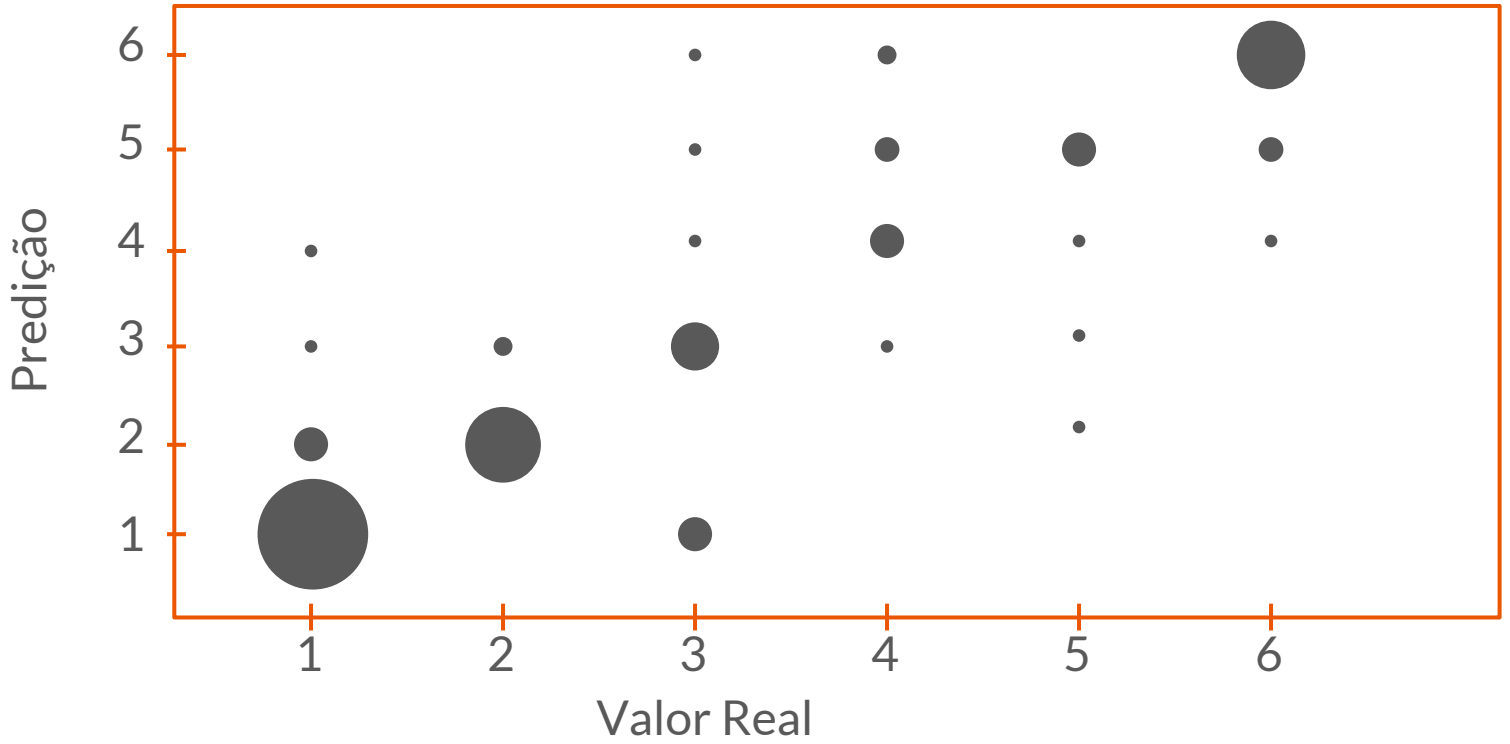


Acurácia = $\frac{\text{verdadeiro positivo} + \text{verdadeiro negativo}}{\text{todos}}$

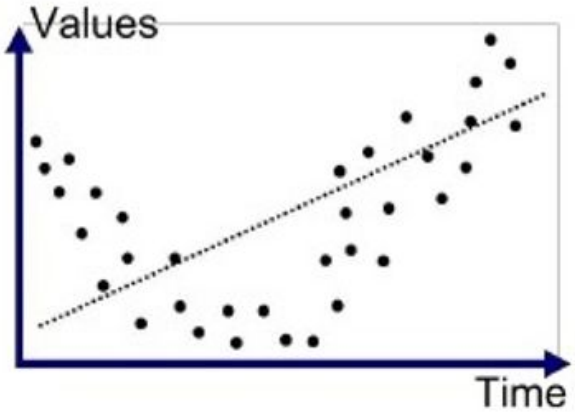
Precisão = $\frac{\text{verdadeiro positivo}}{\text{verdadeiro positivo} + \text{falso positivo}}$

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

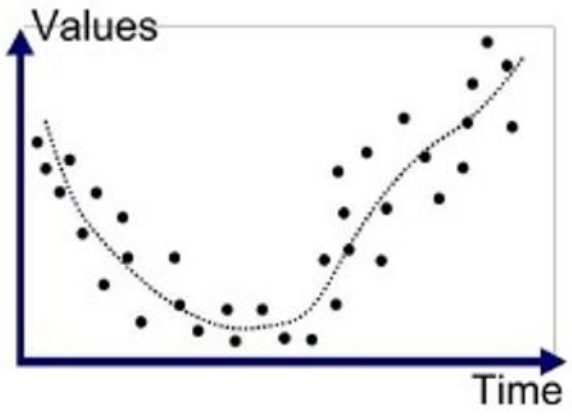
Erro Médio Absoluto Regressão



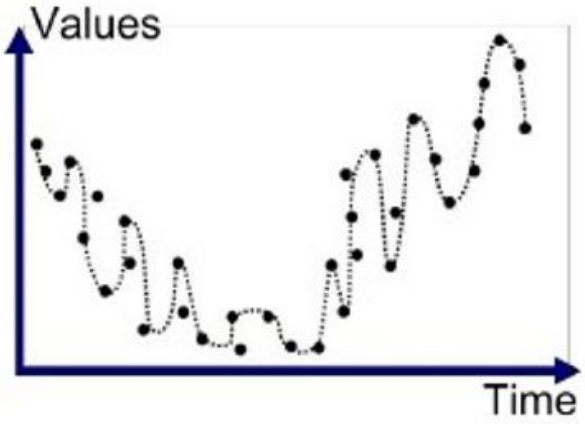
Overfitting na Regressão



Underfitted

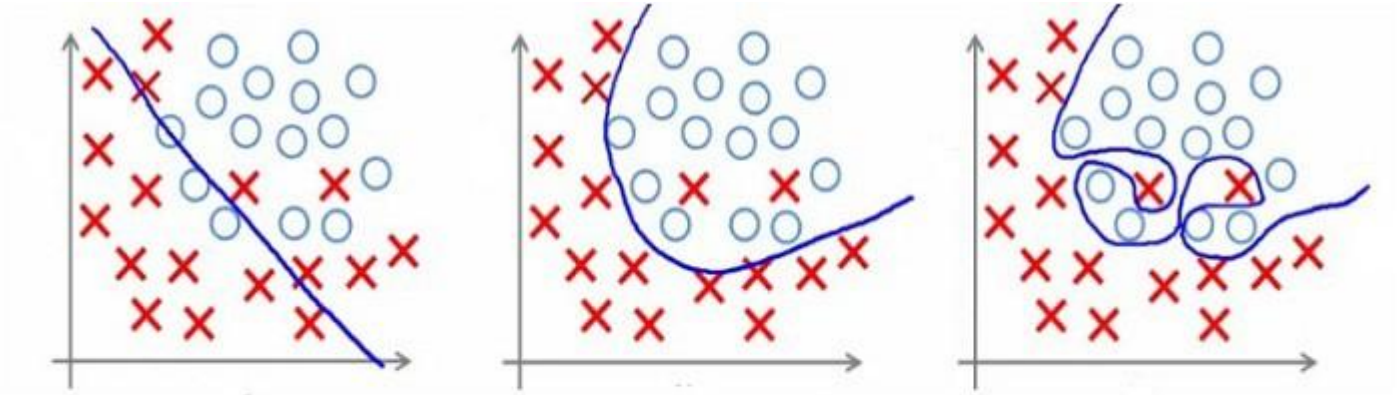


Good Fit/Robust



Overfitted

Overfitting na Classificação



Under-fitting

(too simple to explain the variance)

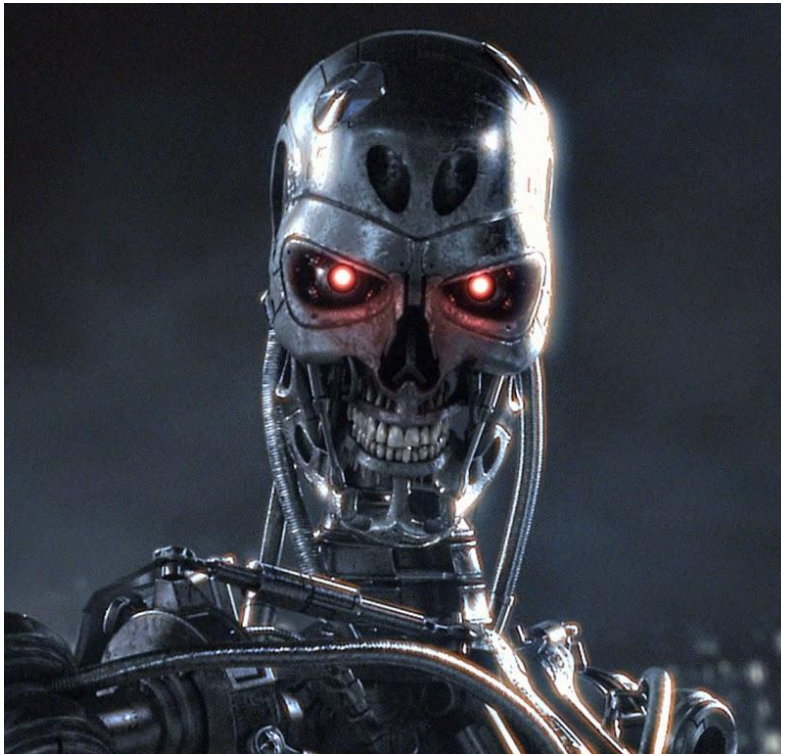
Appropriate-fitting

Over-fitting

(forcefitting -- too good to be true)

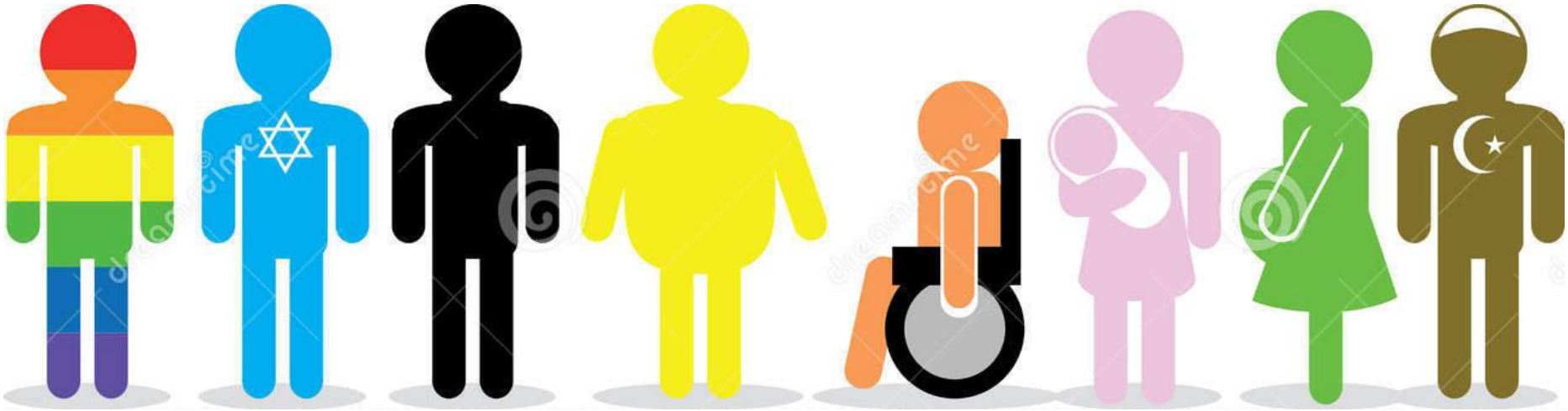
Slides: http://goo.gl/BgoJmm

I.A. Genérica



Slides: http://goo.gl/BgoJmm

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

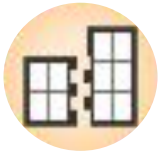
Datasets: *HDI* e *Boston Housing*

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.

Slides: http://goo.gl/BgoJmm

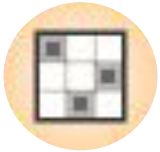
Pré-Processamento no Orange



Merge Data

Transpose

Feature Constructor



Impute

Discretize

Color Data

Fixando Pré-Processamentos

Datasets: *Boston Housing e Bank Marketing*

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

- Qual pré-processamento trouxe melhoria para quais algoritmos?

Material de Aprendizado Não-Supervisionado

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

Fixando Apendizado Não-Supervisionado

Datasets: *Foodmart 2000 e Zoo*

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 Tailes*

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

Dataiku (para sexta)

Inscrição:

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

Porque usar Data Science?

Para reduzir os custos ou aumentar a eficiência da sua organização, automatizando ou melhorando processos.

Obrigado!

Henrique Dias

Doutorando PUCRS

henrique.santos.003@acad.pucrs.br

PDF dos Slides

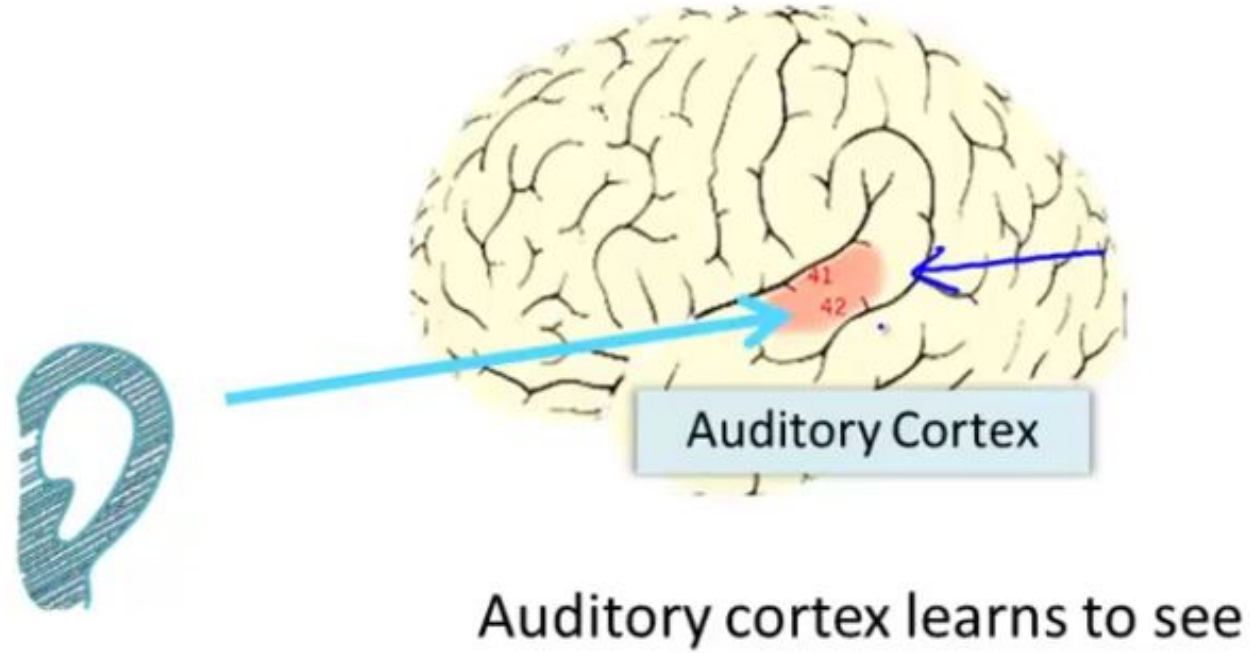
Slides: http://goo.gl/BgoJmm

Curso de Extensão - PUCRS

Segunda	Terça	Quarta	Quinta	Sexta
Conceitos Básicos Tipos de Aprendizagens	Classificação Regressão Avaliação de Modelos	Pré-Processamento Agrupamento Redução de Dimencionalidade Regras de Associação Agrupamento	Redes Neurais Processamento da Linguagem Natural Séries Temporais	Ciência de Dados nas Núvens

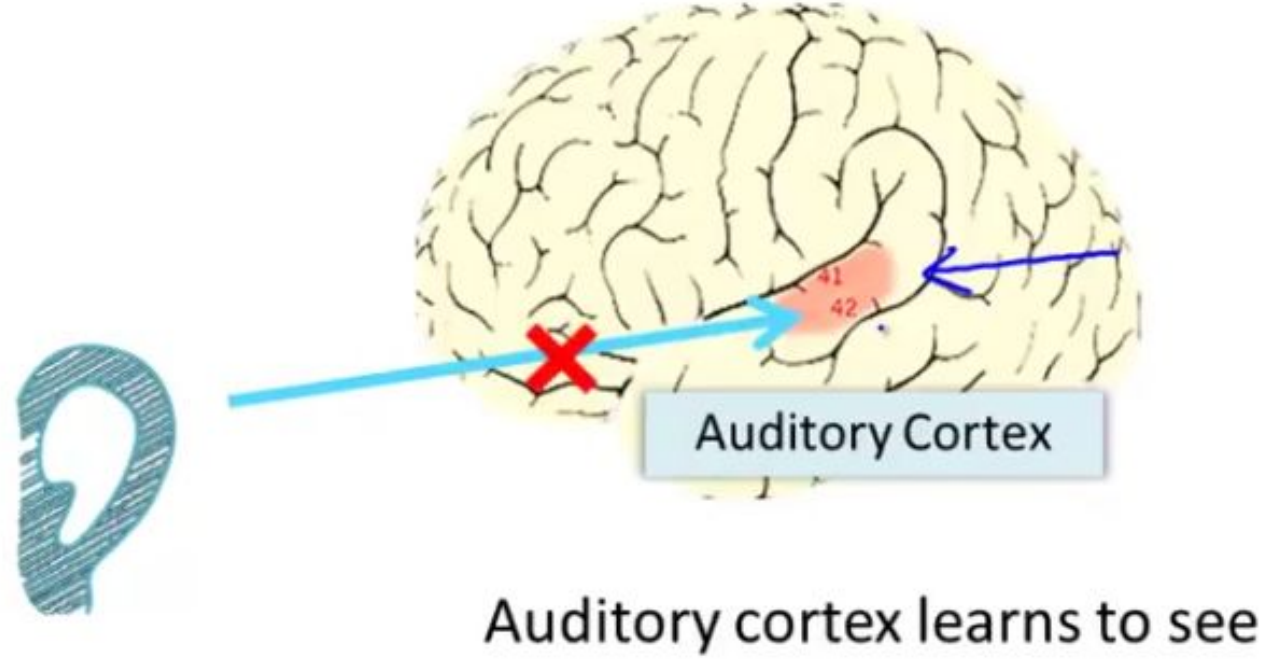
Redes Neurais

A hipótese do algoritmo de aprendizado único

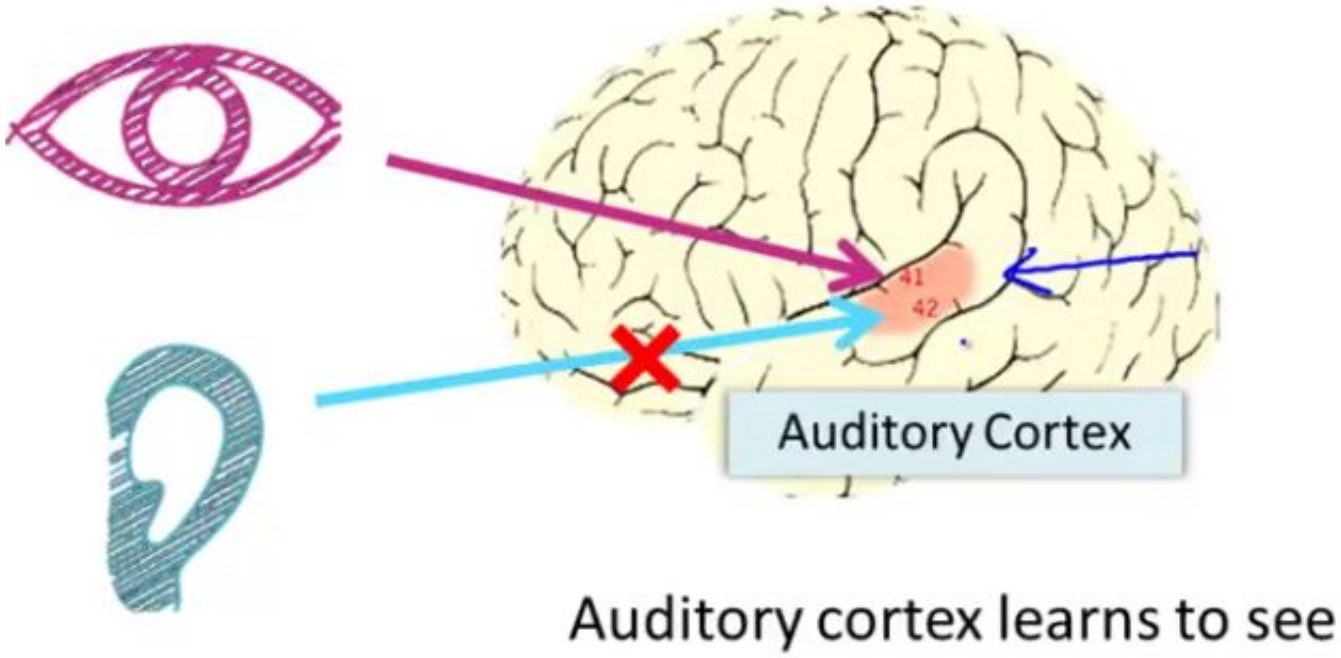


Slides: http://goo.gl/BgoJmm

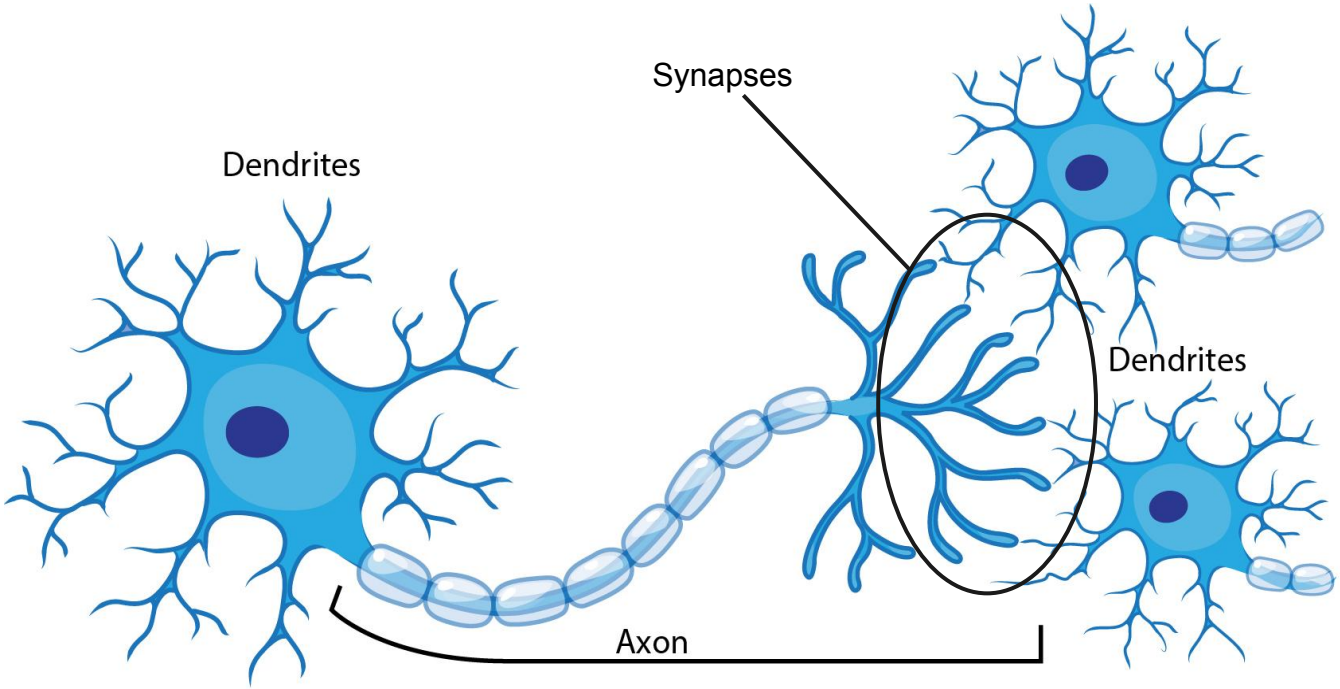
A hipótese do algoritmo de aprendizado único



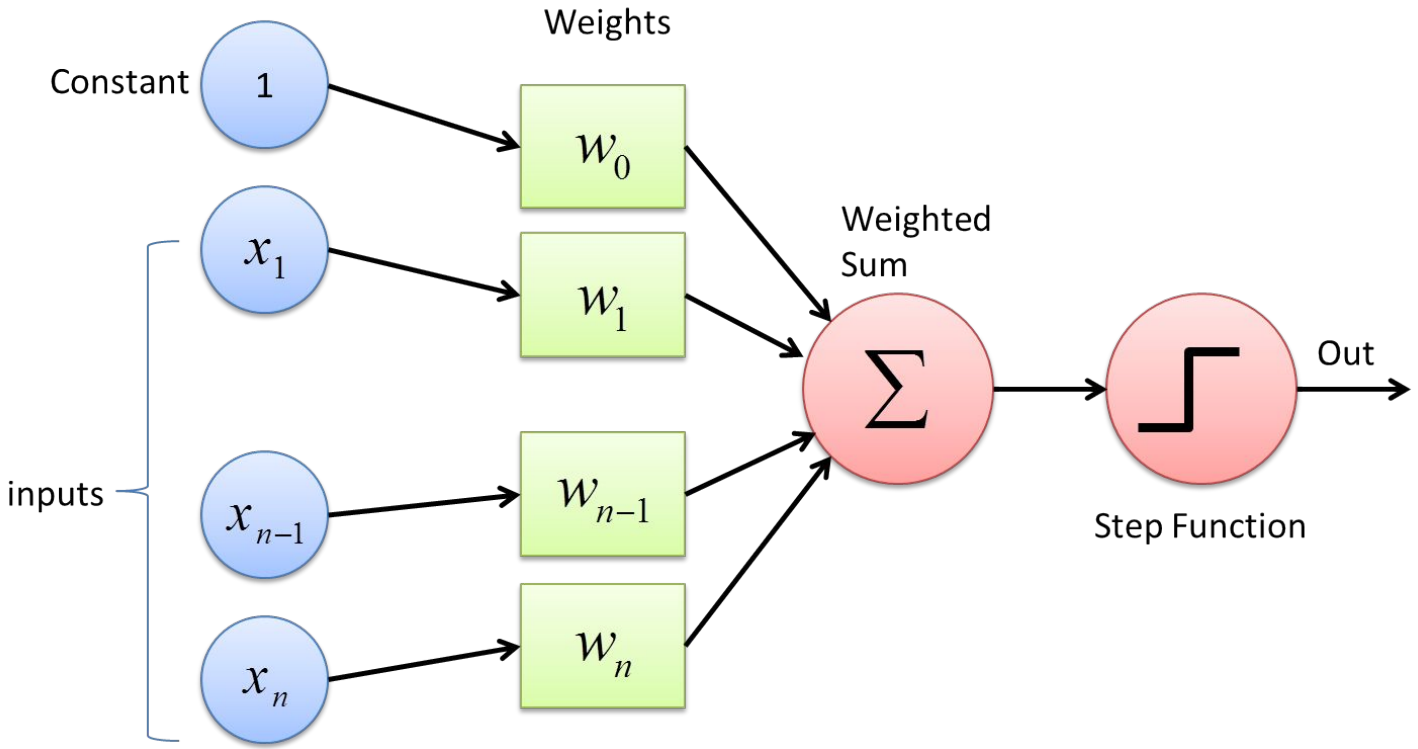
A hipótese do algoritmo de aprendizado único



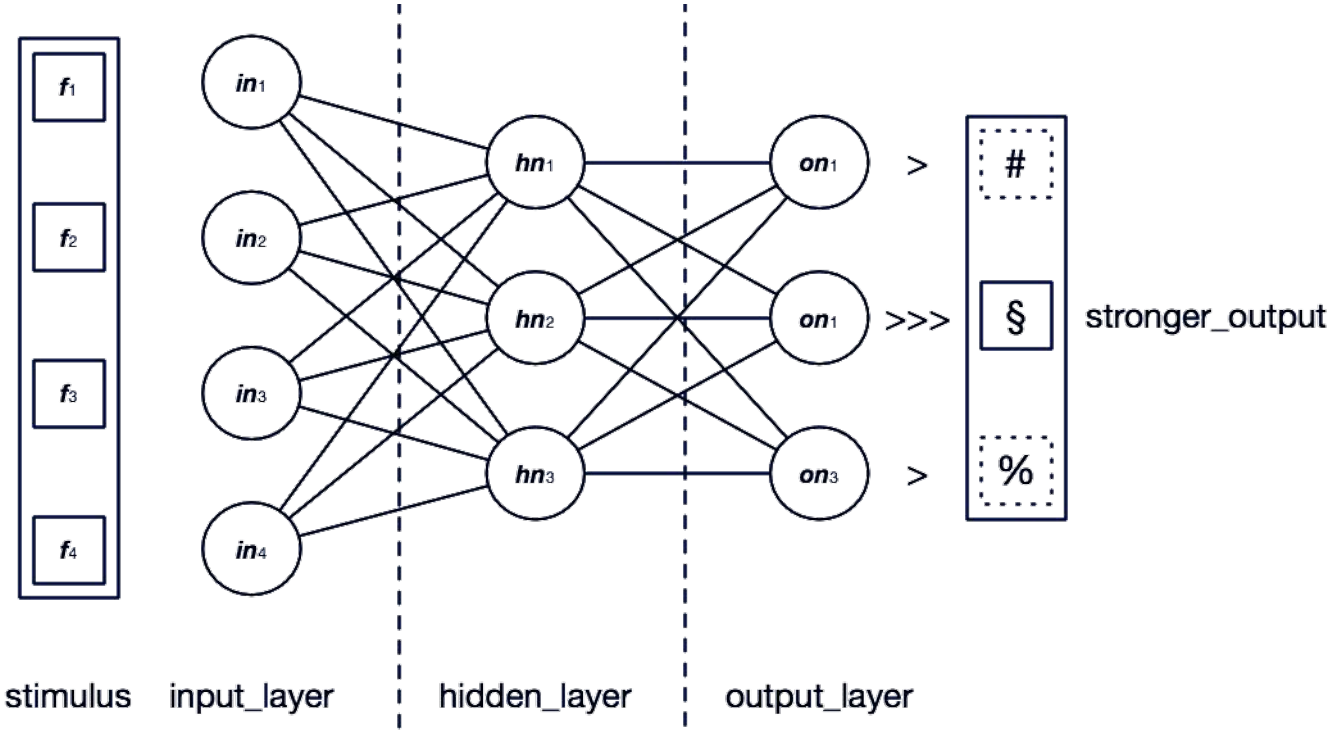
Neurônio (perceptron)



Perceptron (neurônio)

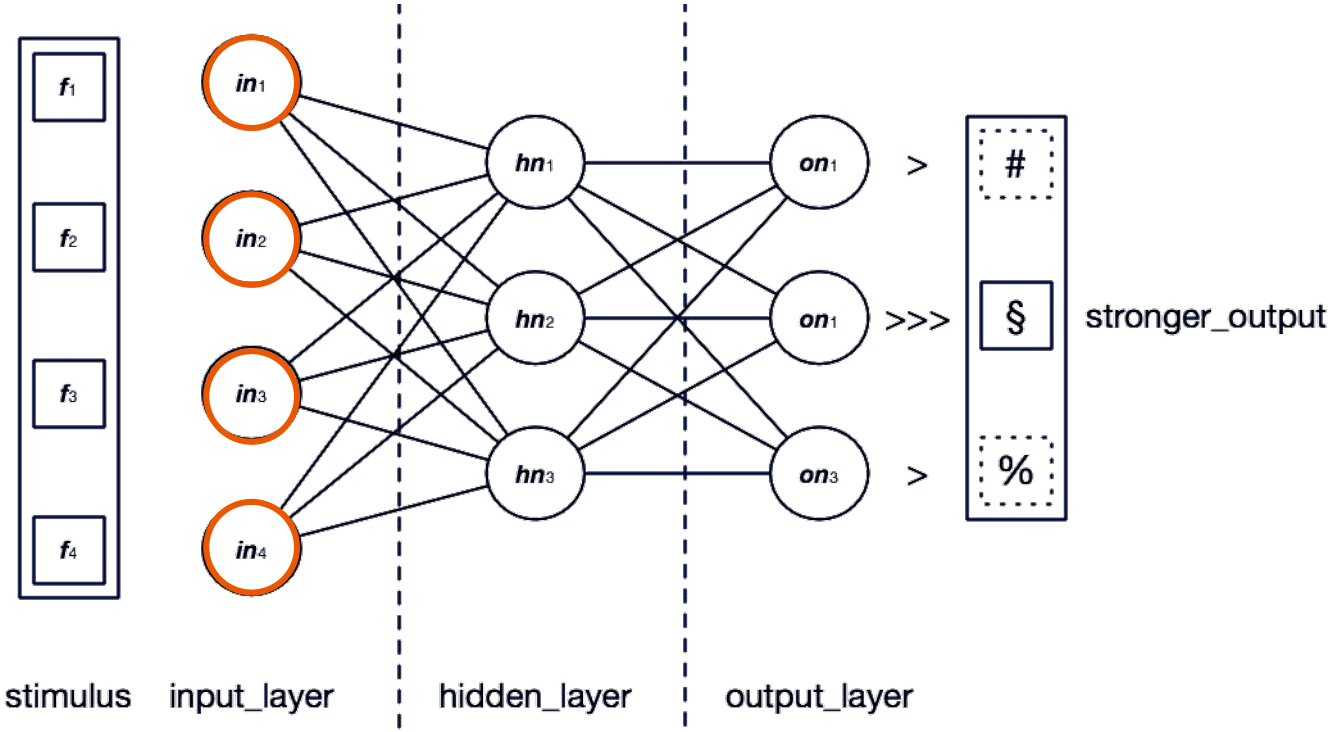


Multi Layer Perceptron

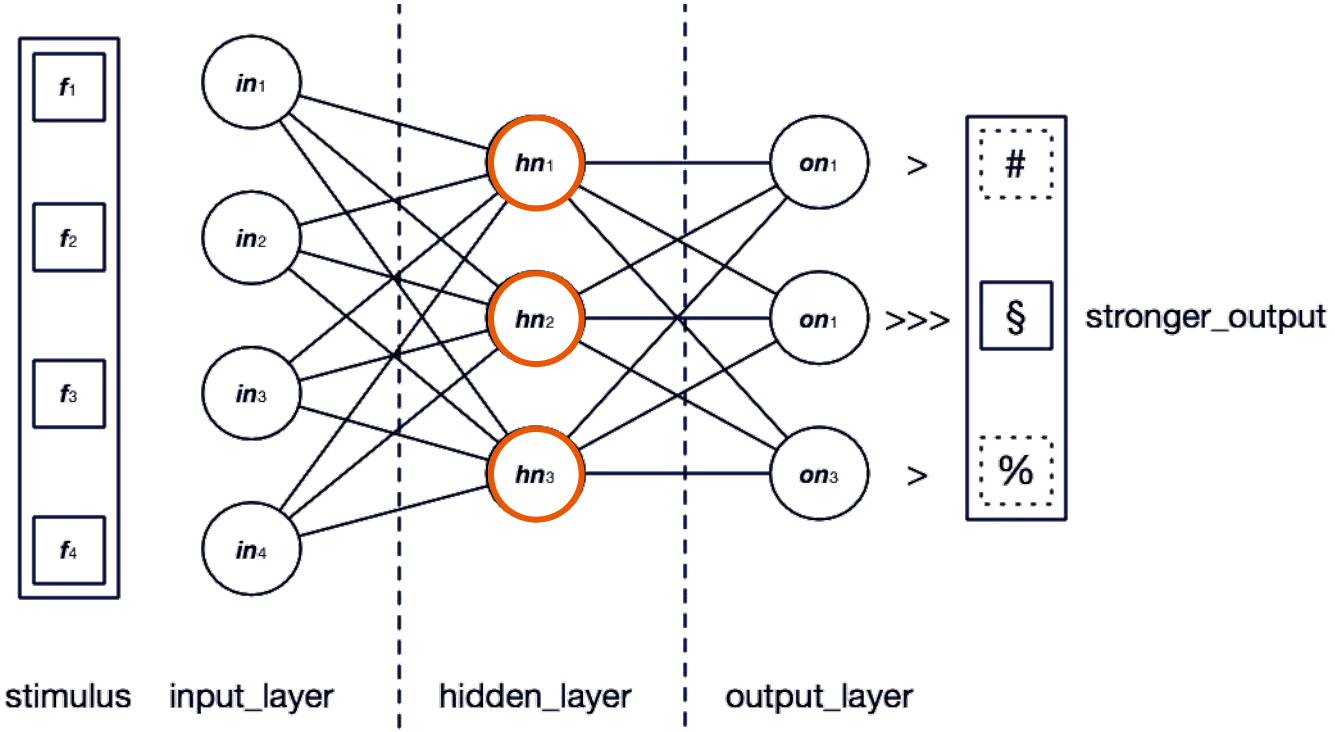


Slides: http://goo.gl/BgoJmm

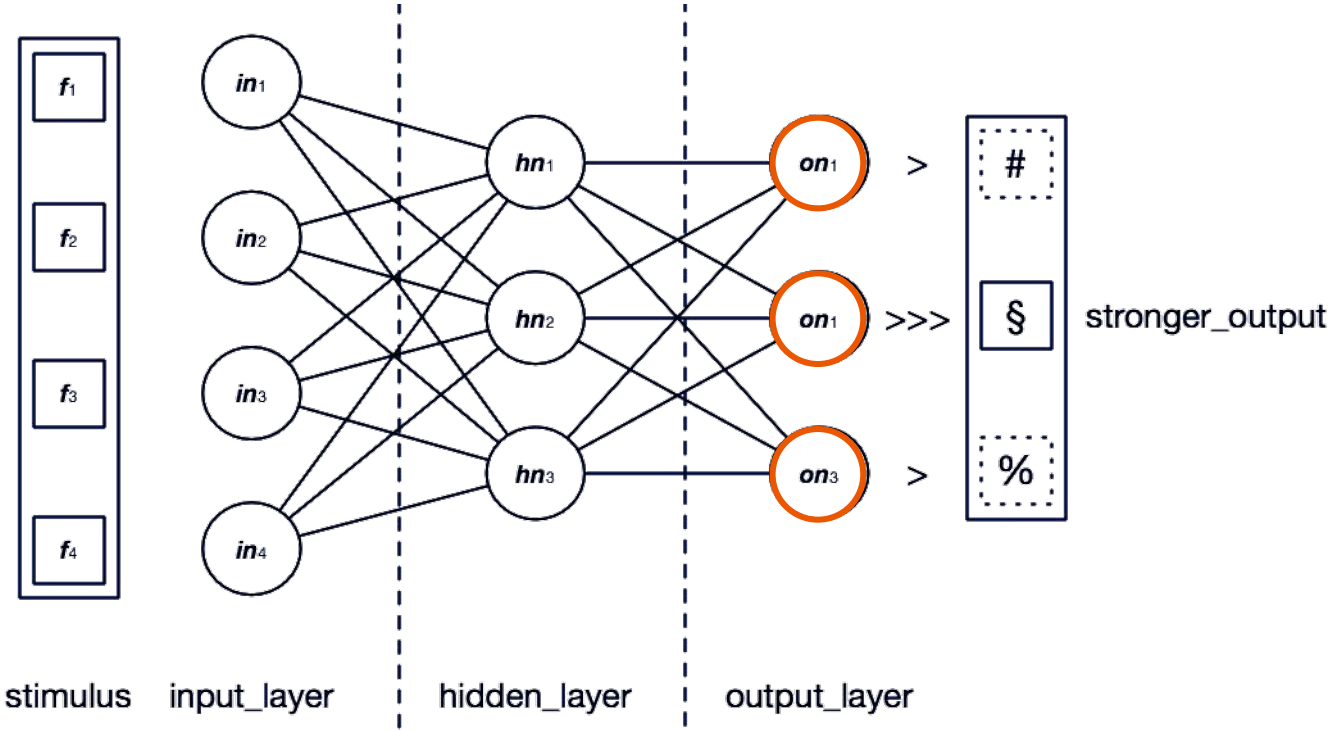
Multi Layer Perceptron



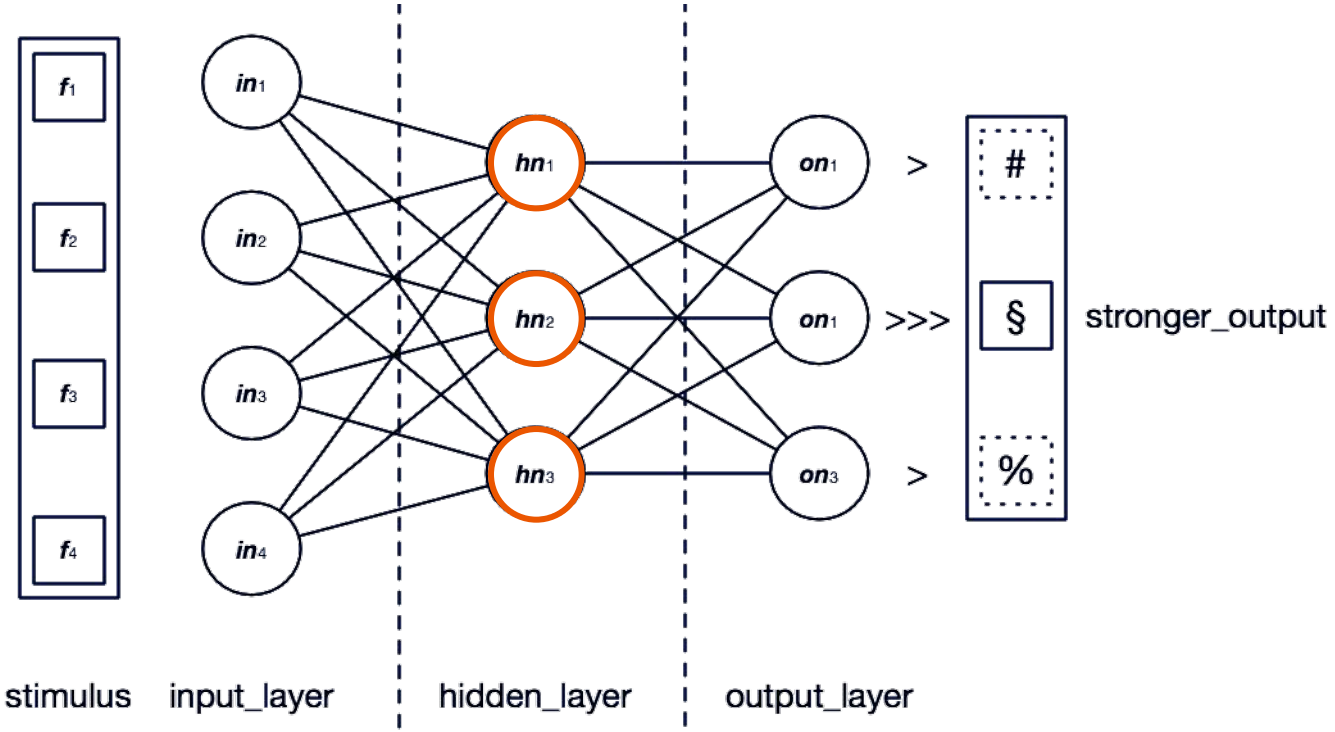
Multi Layer Perceptron



Multi Layer Perceptron

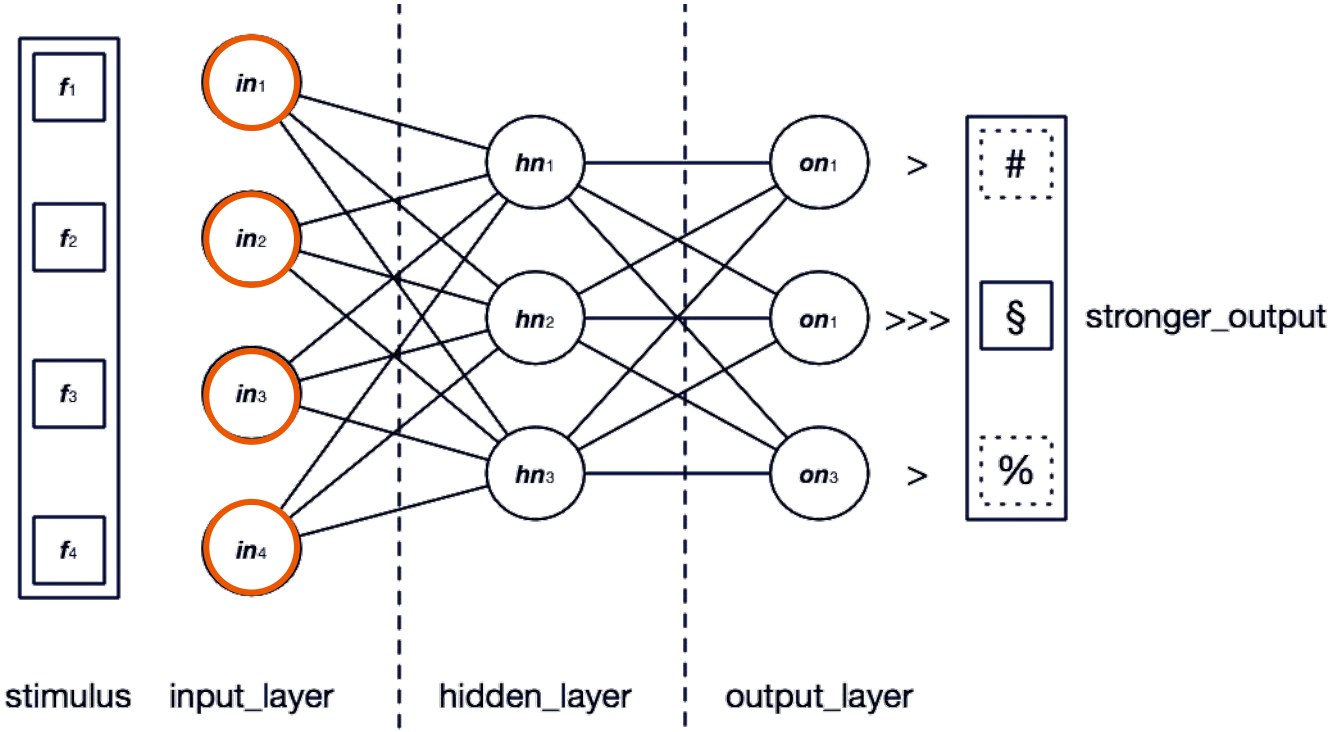


Multi Layer Perceptron



Slides: http://goo.gl/BgoJmm

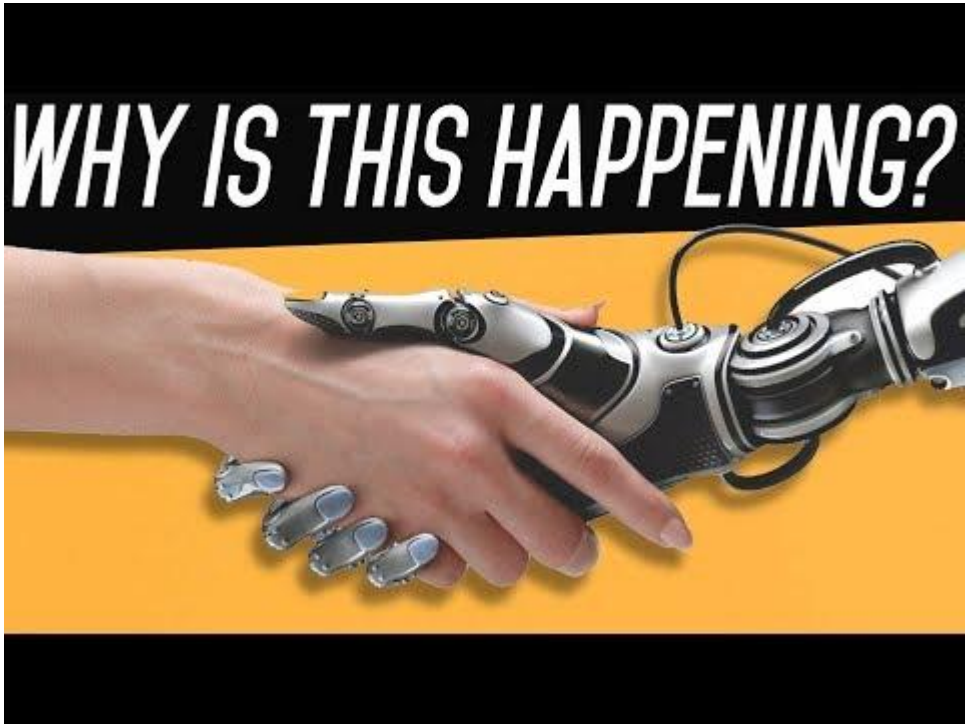
Multi Layer Perceptron



Entendendo Redes Neurais

https://playground.tensorflow.org

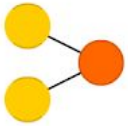
Do que as RNs são capazes?



Neural Networks

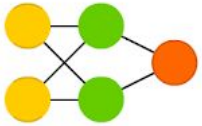
- Backfed Input Cell
- Input Cell
- Noisy Input Cell
- Hidden Cell
- Probablistic Hidden Cell
- Spiking Hidden Cell
- Output Cell
- Match Input Output Cell
- Recurrent Cell
- Memory Cell
- Different Memory Cell
- Kernel
- Convolution or Pool

Perceptron (P)

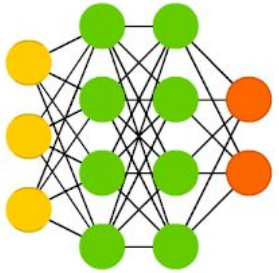


Feed Forward (FF)

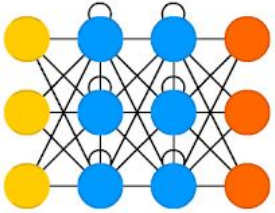
Radial Basis Network (RBF)



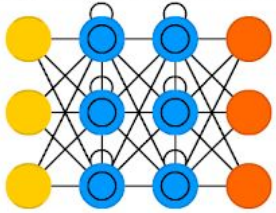
Deep Feed Forward (DFF)



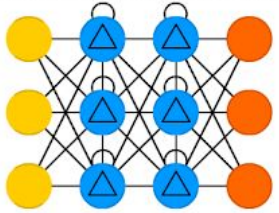
Recurrent Neural Network (RNN)



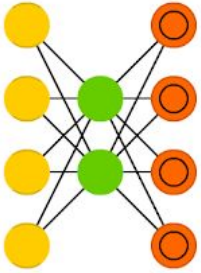
Long / Short Term Memory (LSTM)



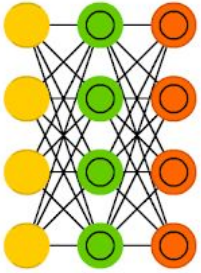
Gated Recurrent Unit (GRU)



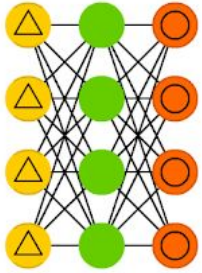
Auto Encoder (AE)



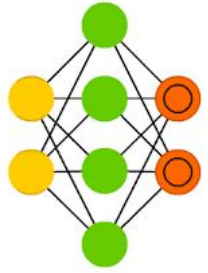
Variational AE (VAE)



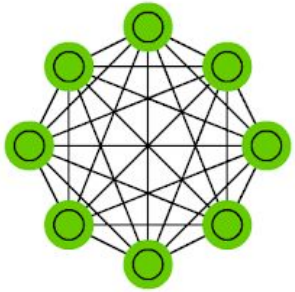
Denoising AE (DAE)



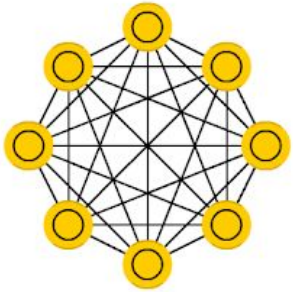
Sparse AE (SAE)



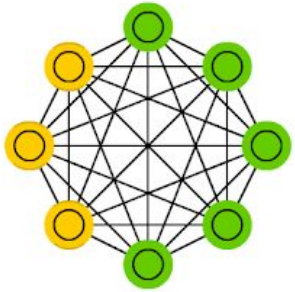
Markov Chain (MC)



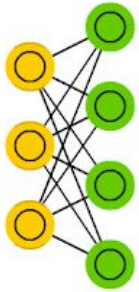
Hopfield Network (HN)



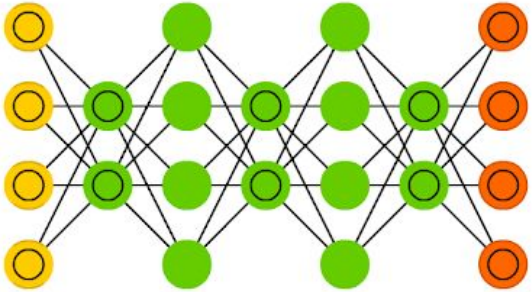
Boltzmann Machine (BM)



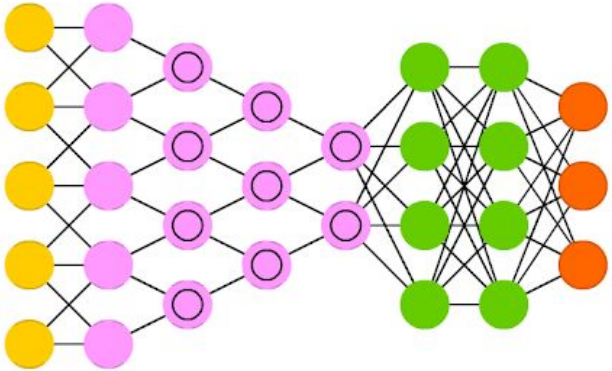
Restricted BM (RBM)



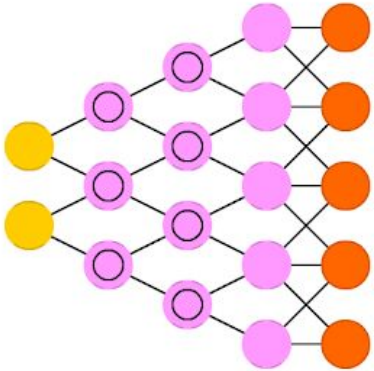
Deep Belief Network (DBN)



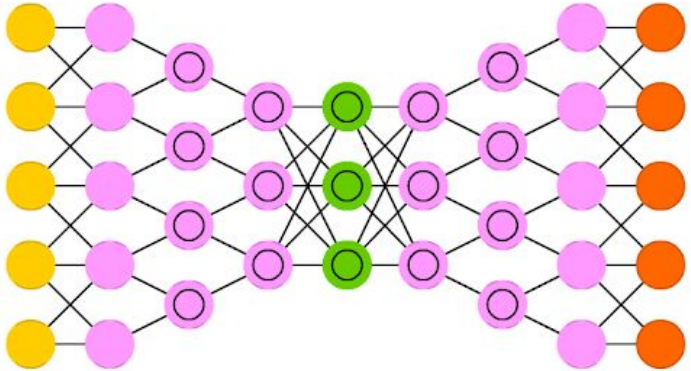
Deep Convolutional Network (DCN)



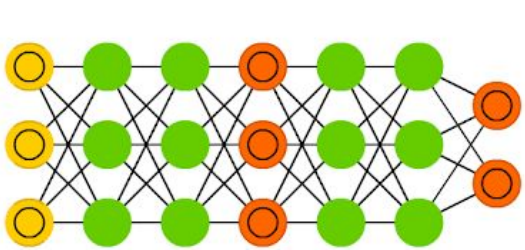
Deconvolutional Network (DN)



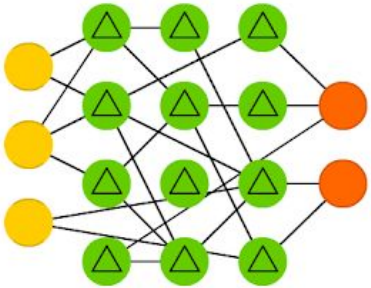
Deep Convolutional Inverse Graphics Network (DCIGN)



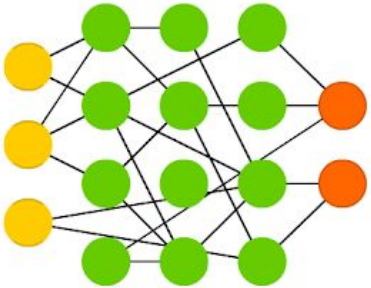
Generative Adversarial Network (GAN)



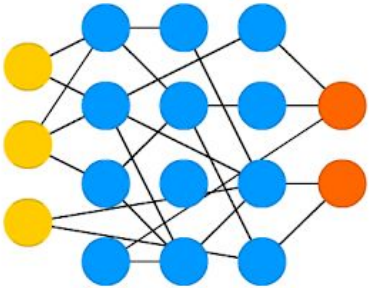
Liquid State Machine (LSM)



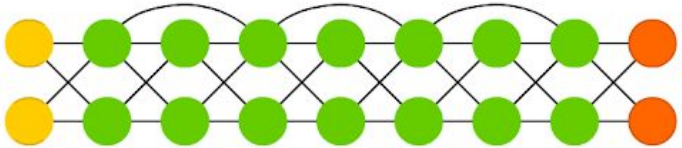
Extreme Learning Machine (ELM)



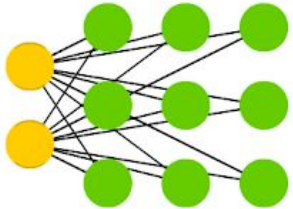
Echo State Network (ESN)



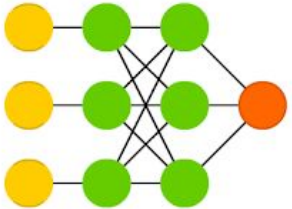
Deep Residual Network (DRN)



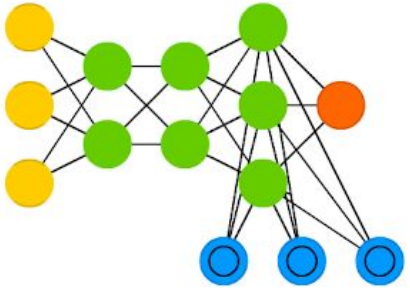
Kohonen Network (KN)



Support Vector Machine (SVM)

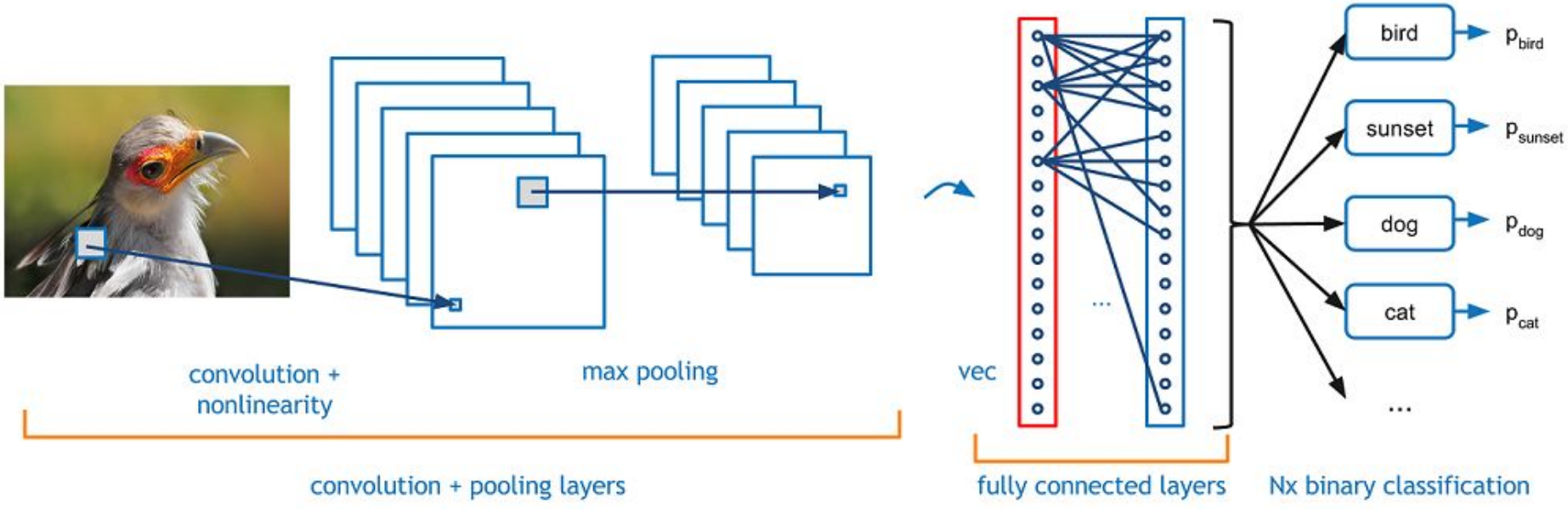


Neural Turing Machine (NTM)



Slides: http://goo.gl/BgoJmm

Redes Neurais Convolucionais

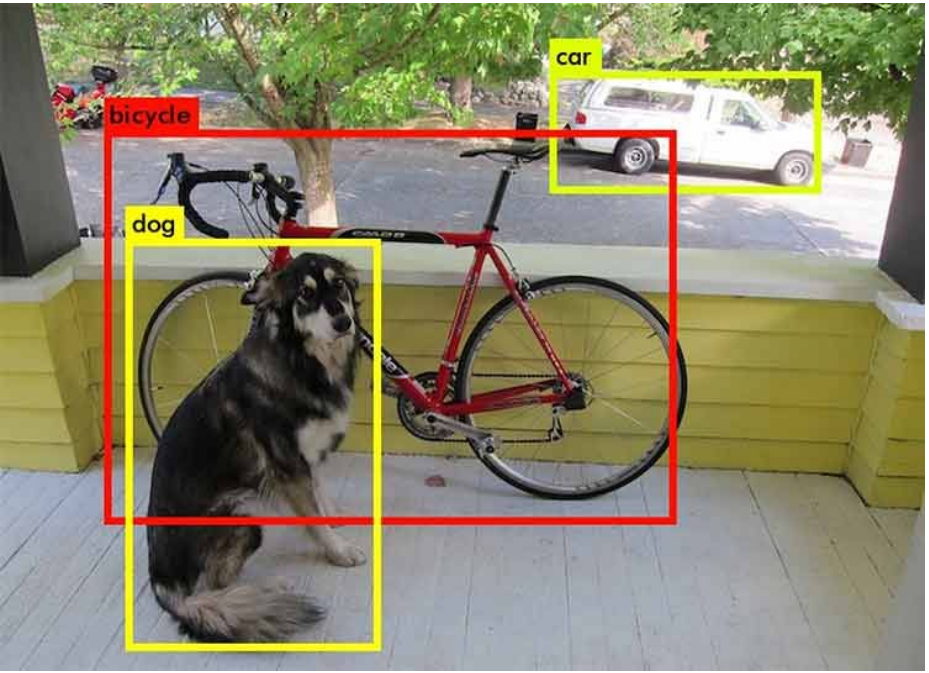


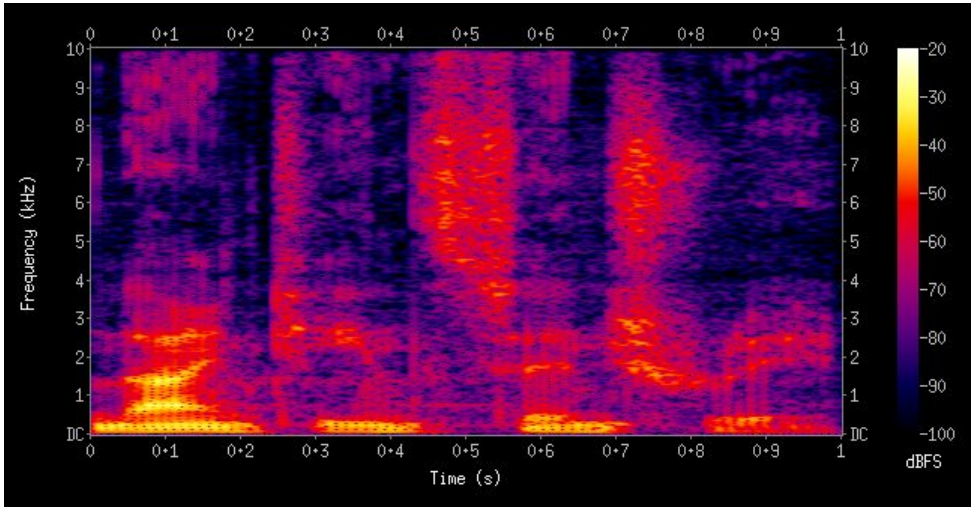
Slides: http://goo.gl/BgoJmm

Redes Neurais Convolucionais

Slides: http://goo.gl/BgoJmm

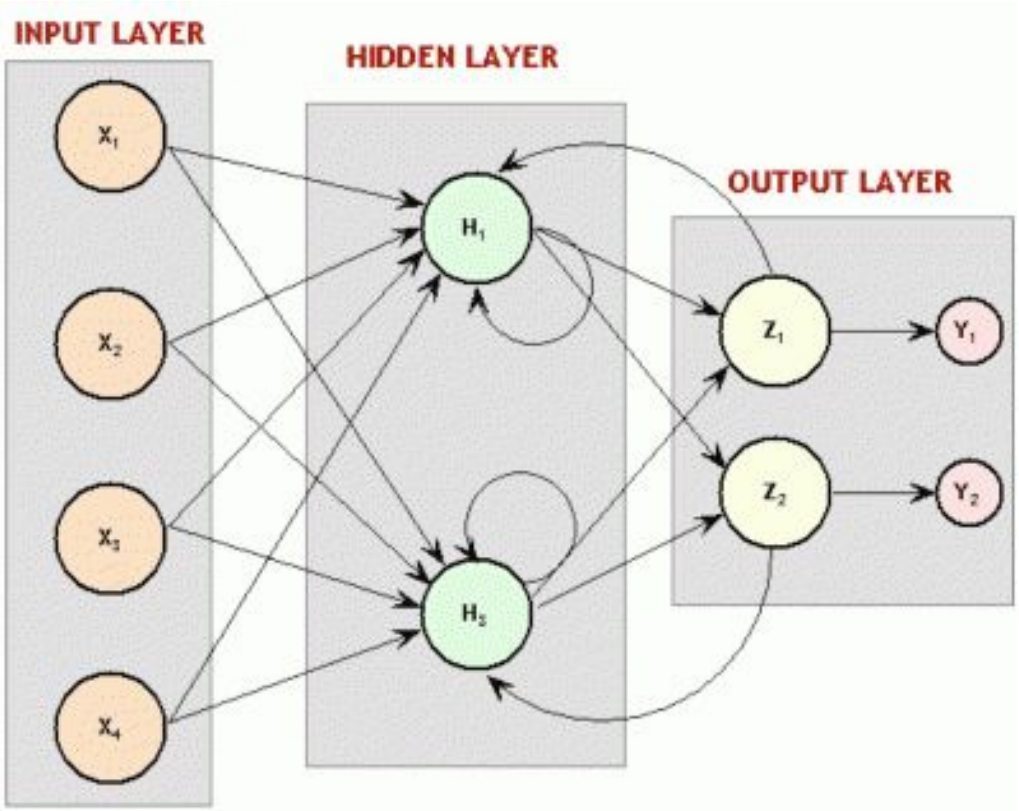
Redes Neurais Convolucionais



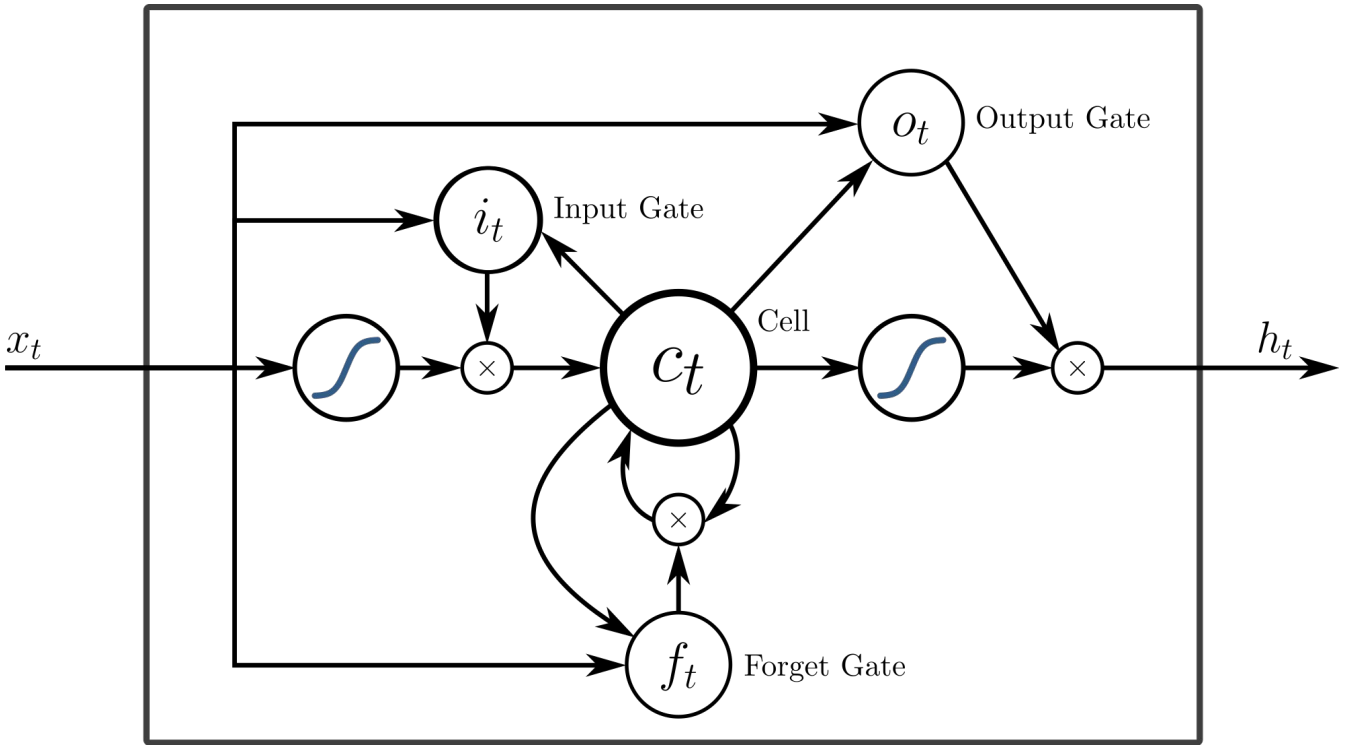


Slides: http://goo.gl/BgoJmm

Redes Neurais Recorrentes

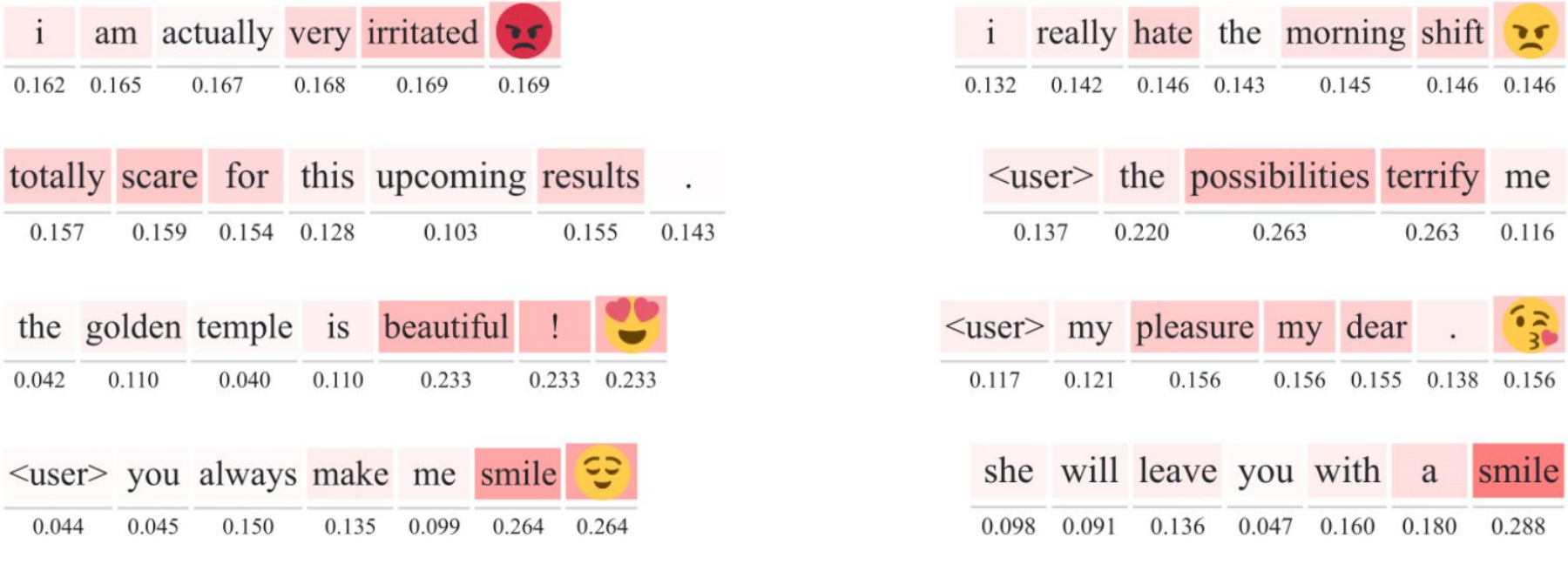


Long-Short Term Memory (LSTM)



Slides: http://goo.gl/BgoJmm

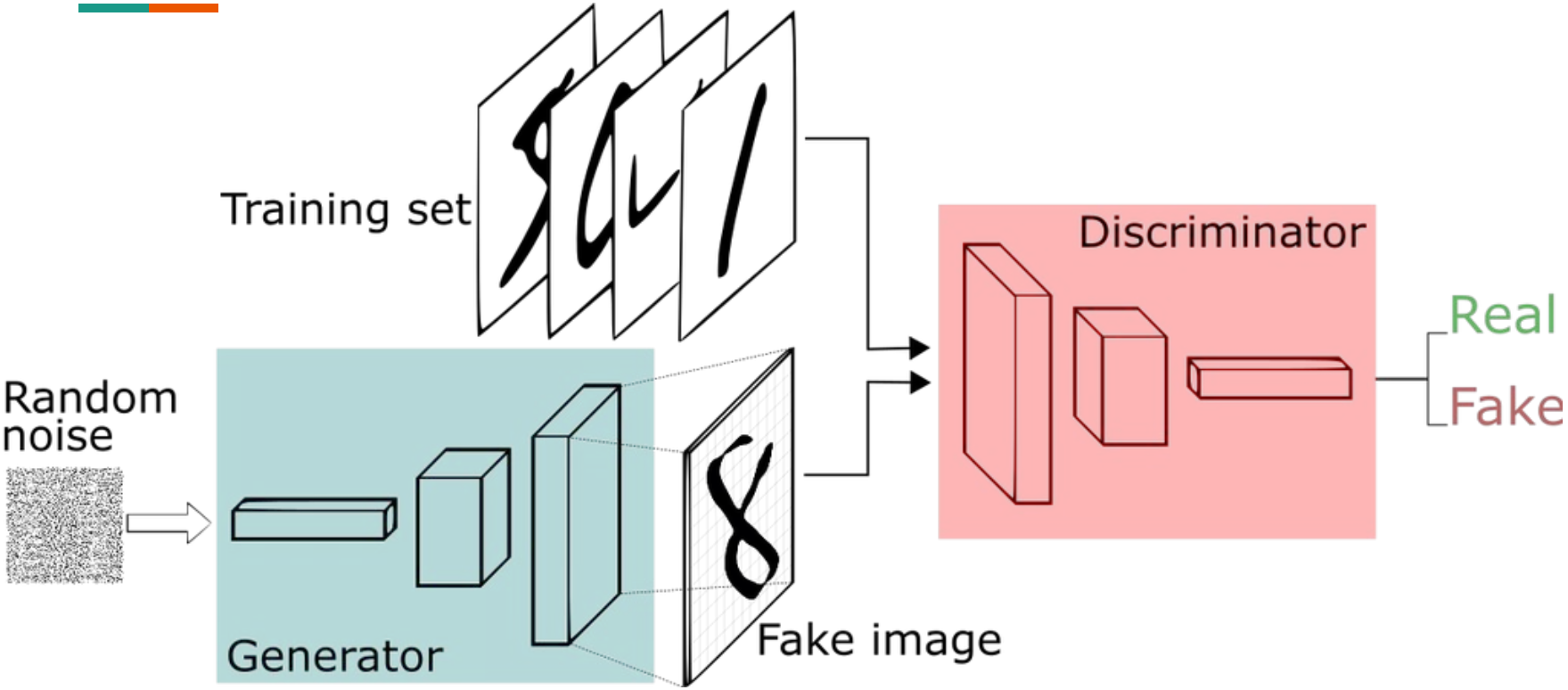
Redes Neurais Recorrentes e LSTMs



(Análise de sentimento)

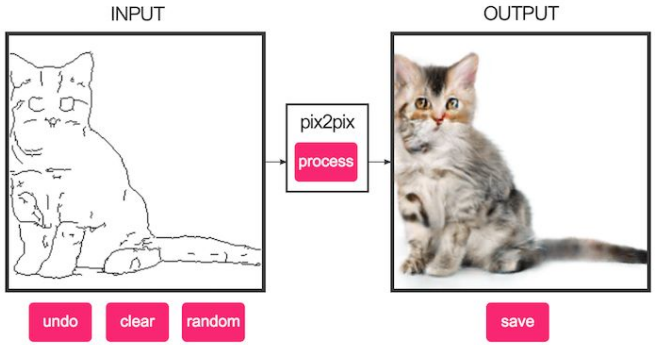
Slides: http://goo.gl/BgoJmm

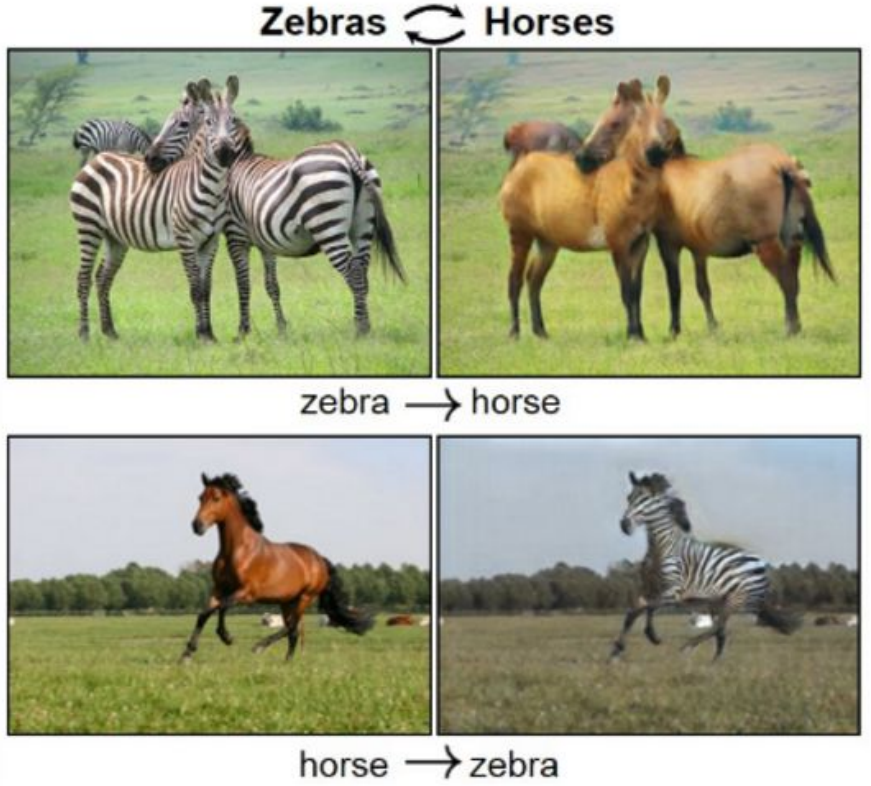
Generative Adversarial Network



Slides: http://goo.gl/BgoJmm

Generative Adversarial Network





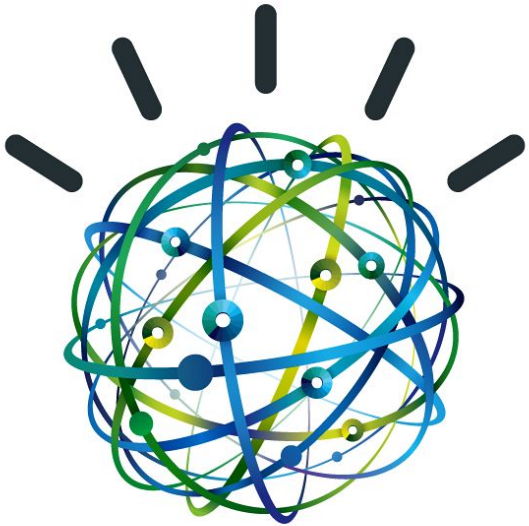
Testando Redes Neurais Código

http://goo.gl/vVULcj

Limitações das NNs

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

Ferramentas de alto nível

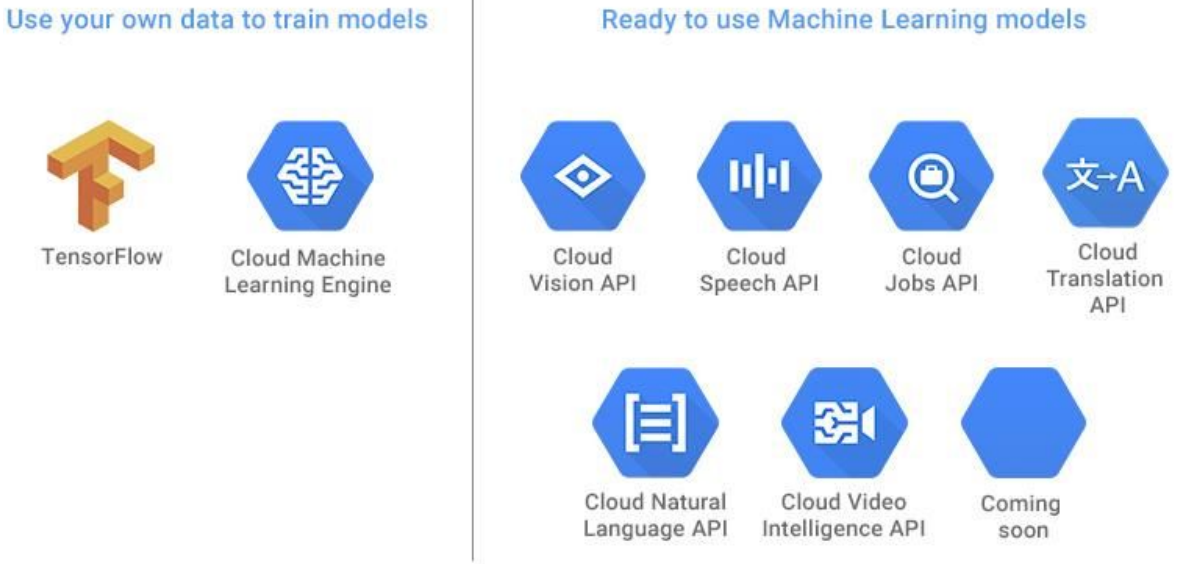


IBM Watson

ibm.com/watson/

Ferramentas de alto nível

Google Cloud AI



cloud.google.com/products/ai/

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	✓	

https://firebase.google.com/docs/ml-kit/

Slides: http://goo.gl/BgoJmm

Ferramentas de alto nível

Azure AI

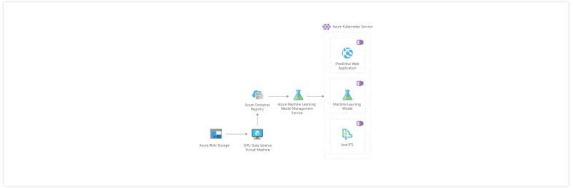
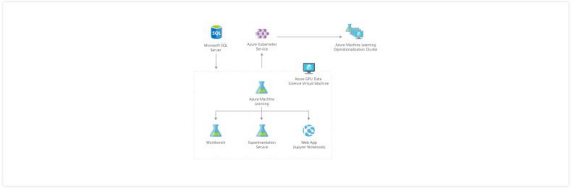


Image classification with convolutional neural networks

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

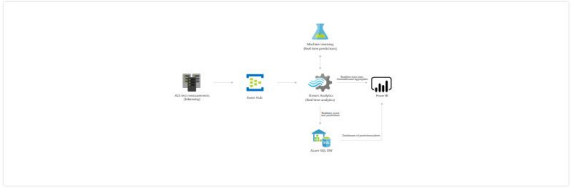
Learn more >



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.

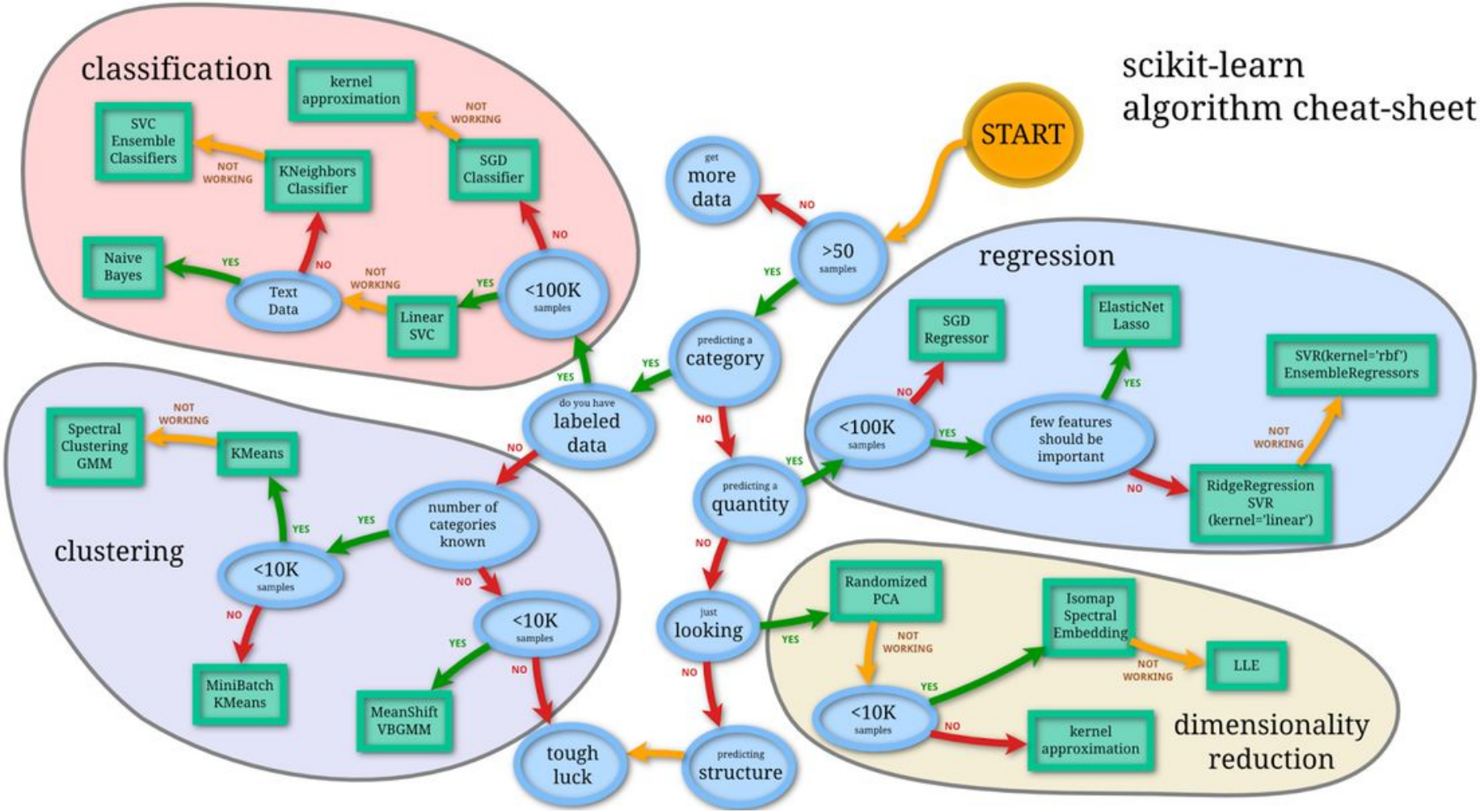
Learn more >

Enterprise Productivity Chatbot

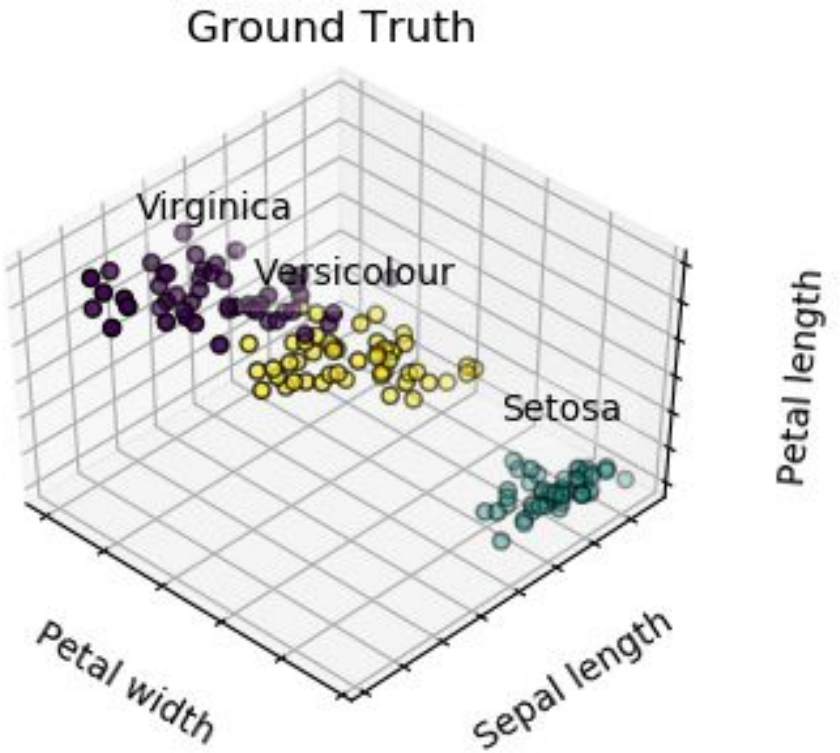
Azure Bot Service can be easily combined with Language Understanding to build powerful enterprise productivity bots, allowing organizations to streamline common work activities by integrating external systems, such as Office 365 calendar, customer cases stored in Dynamics CRM and much more.

Learn more >

azure.microsoft.com/en-us/overview/ai-platform/


scikit-learn
algorithm cheat-sheet
START
get more data
NO
>50 samples
YES
predicting a category
YES
do you have labeled data
YES
<100K samples
NO
Linear SVC
NOT WORKING
Text Data
NO
Naive Bayes
YES
KNeighbors Classifier
NOT WORKING
SVC Ensemble Classifiers
NOT WORKING
SGD Classifier
NO
kernel approximation
NOT WORKING
classification
NO
number of categories known
YES
<10K samples
YES
KMeans
NOT WORKING
Spectral Clustering GMM
NO
MiniBatch KMeans
NO
<10K samples
YES
MeanShift VBGMM
NO
tough luck
clustering
NO
predicting a quantity
YES
<100K samples
NO
SGD Regressor
YES
few features should be important
YES
ElasticNet Lasso
NO
RidgeRegression SVR (kernel='linear')
NOT WORKING
SVR(kernel='rbf') EnsembleRegressors
regression
NO
just looking
YES
Randomized PCA
NOT WORKING
Isomap Spectral Embedding
NOT WORKING
LLE
YES
<10K samples
NO
kernel approximation
dimensionality reduction
NO
predicting structure

Clusterização: Iris



Classificação: Problemas x Algoritmos

A. Câncer de Mama
B. Dígitos (Imagem)
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
- http://scikit-learn.org/stable/documentation.html
- https://colab.research.google.com/notebooks/basic_features_overview.ipynb
- https://www.datascienceacademy.com.br/pages/cursos-gratuitos
- 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