

Workshop Intensivo de Aplicações Modernas de Ciência de Dados com Machine Learning

Segunda Turma

Primeiro dia

Paulo Cysne Rios Jr.

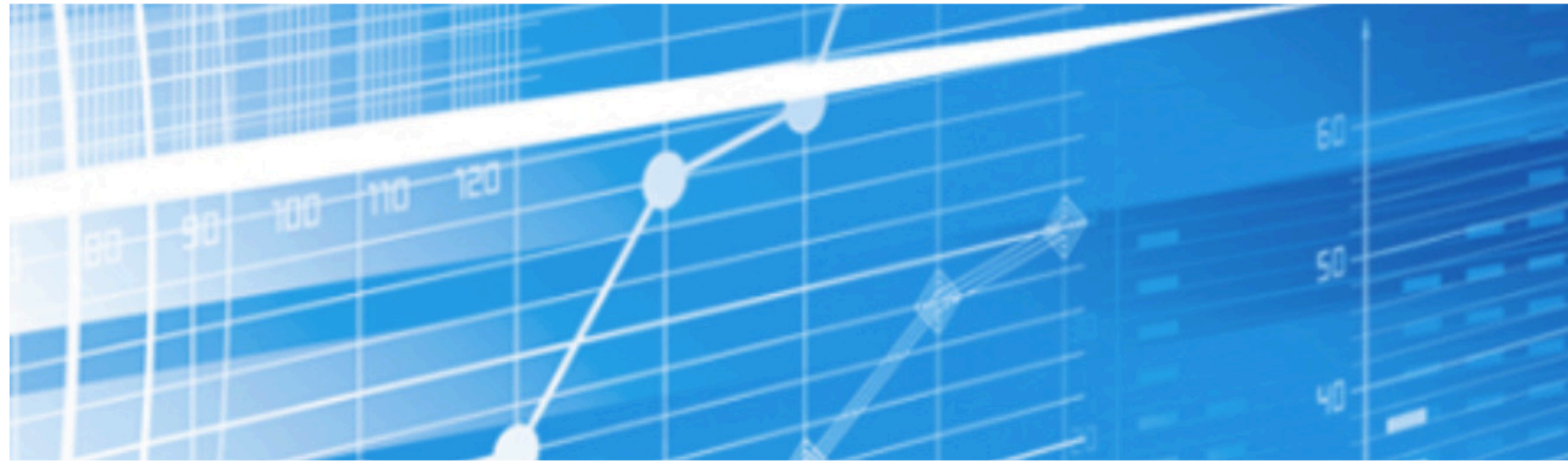
Apresentação

- Sobre mim

Cyzne.com

Datacyz.com

- Sobre vocês



How we can help you

Nowadays we live with an abundance of data in all areas. These data hold critical information that can be extremely beneficial to companies and institutions in nearly all areas of endeavor. Our advanced analytical techniques, based on the latest know-how in Data Science, can extract knowledge from your data and turn it into mission critical information, actionable intelligence.

Manufacturing

Learn when machines will fail so that maintenance can be done offsite, without any production stops. Increase production efficiency. Reduce costs.

Logistics, Telecom

Discover which resources work best and are better allocated. Find potential bottlenecks. Detect inefficiencies and solve them. Get information customer's behaviors, preferences and movements. Achieve marketing effectiveness.

Healthcare & Public Health

Detect which medications are needed and which conditions are prevalent or subjacent. Find trends. Identify hidden relationships.

Marketing & Retail

Identify customers who are likely to purchase more. Decrease churn rates. Use the most efficient advertising channels. Find the right audience.

Customer Retention Analytics

[Leave a reply](#)



How you can improve your customer experience

by Paulo C. Rios, Jr. | [Sep 11, 2017](#)

It is much easier and less costly to sell your products and/or services to existing customers than to new ones. Predictive Analytics can help you in many ways towards this goal, as I explain in this article.


[Continue reading →](#)

This entry was posted in [Advanced Data Analytics](#), [Data Science](#), [Predictive Analytics](#) and tagged [cus-](#)

ABOUT



Paulo C. Rios, Jr. is an expert in data science, advanced data analytics, digital technology, business analysts and information technology who has been active in different roles, as a Consultant, Director, Lead, Entrepreneur, Instructor and Writer, with over 28 years of professional work experience. He has a BS in Physics and a MBA. He is passionate about what he does and its power to transform business, healthcare and the world for better. [More about Paulo](#)




[Home](#)[My Network](#)[Jobs](#)[Messaging](#)[Notifications](#)[Me](#)[Work](#)

7,567

Your connections

[See all](#)



Some of your contacts aren't connected with you on LinkedIn


Connect with them and never lose touch

[Continue](#)

No pending invitations


[Manage all](#)

People you may know




Musfiqu Salehine

Data Enthusiast (Analytics, Data Science, Big Data,


 Chris Gardner and 112 others

[Connect](#)




Leifur Thorbergsson

Data Scientist at Memorial Sloan Kettering Cancer Center


 Soma Bhattacharya and 162 others

[Connect](#)



Ajay Sharma

WorkForce planning Analytics at Ericsson

 Wayne Thompson and 98 others

[Connect](#)

LinkedIn: ch.linkedin.com/in/paulocriosjr

Apresentação: Vocês

- Onde trabalha e/ou estuda
- O que já ouviu falar de Machine Learning e Ciência de Dados
- Objetivo em fazer o workshop
- O que gostaria de aprender
- Que conhecimento já tem na área, inclusive de programação, em especial com Python

Objetivos

Capacitar os participantes do curso na utilização de Machine Learning para resolução de projetos em Ciências de Dados através da aplicação das técnicas mais recentes e mais usadas nas mais diversas áreas, mostrando o seu grande impacto.

Machine Learning



what society thinks I
do



what my friends think
I do



what my parents think
I do

$$L_p = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^L \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{w} + b) + \sum_{i=1}^L \alpha_i$$

$$\alpha_i \geq 0, \forall i$$

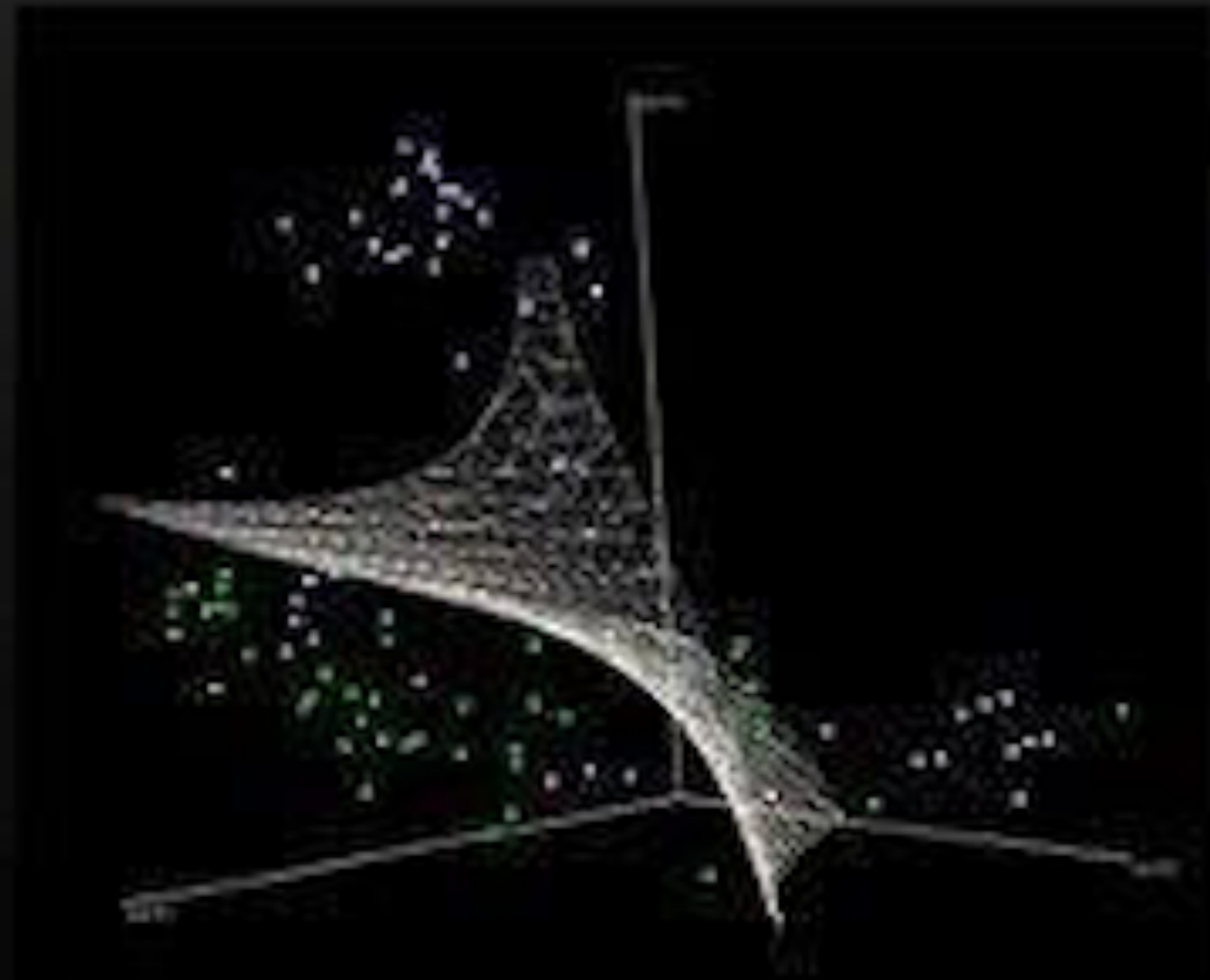
$$\mathbf{w} = \sum_{i=1}^L \alpha_i y_i \mathbf{x}_i, \sum_{i=1}^L \alpha_i y_i = 0$$

$$\nabla \hat{J}(\theta_t) = \frac{1}{n} \sum_{i=1}^n \nabla \ell(x_i, y_i; \theta_t) + \nabla r(\theta_t).$$

$$\theta_{t+1} = \theta_t - \eta_t \nabla \ell(x_{u(t)}, y_{u(t)}; \theta_t) - \eta_t \cdot \nabla r(\theta_t)$$

$$\mathbb{E}_{u(t)}[\ell(x_{u(t)}, y_{u(t)}; \theta_t)] = \frac{1}{n} \sum_{i=1}^n \ell(x_i, y_i; \theta_t).$$

what other programmers
think I do



what I think I do

```
>>> from sklearn import svm
```

what I really do

Machine Learning



what society thinks I
do



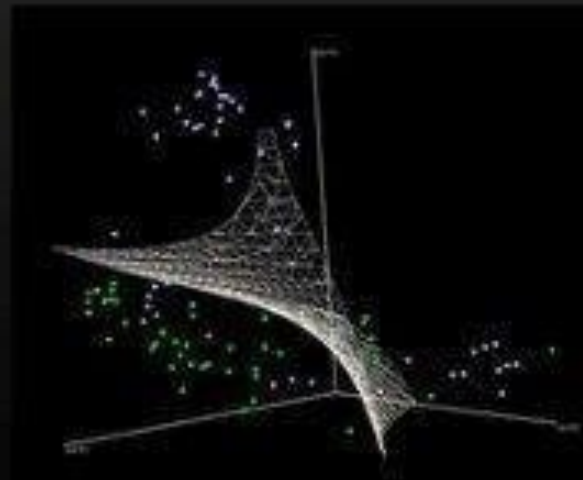
what my friends think
I do



what my parents think
I do

$$\begin{aligned} L_p &= \frac{1}{2} \|\mathbf{w}\|^2 - \sum_i \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{w} + b) + \sum_i \alpha_i \\ \alpha_i &\geq 0, \forall i \\ \mathbf{w} &= \sum_i \alpha_i y_i \mathbf{x}_i, \sum_i \alpha_i y_i = 0 \\ \nabla \hat{J}(\theta_t) &= \frac{1}{n} \sum_{i=1}^n \nabla \ell(x_i, y_i; \theta_t) + \nabla r(\theta_t) \\ \theta_{t+1} &= \theta_t - \eta_t \nabla \ell(x_{i(t)}, y_{i(t)}; \theta_t) - \eta_t \cdot \nabla r(\theta_t) \\ \mathbb{E}_{i(t)}[\ell(x_{i(t)}, y_{i(t)}; \theta_t)] &= \frac{1}{n} \sum_i \ell(x_i, y_i; \theta_t) \end{aligned}$$

what other programmers
think I do



what I think I do

```
>>> from sklearn import svm
```

what I really do

Módulo I – Fundamentos

- Introdução à Machine Learning
 - História da Inteligência Artificial
 - Data Science
 - Big Data, real-time data streaming
 - Machine Learning
 - Deep Learning
 - O cenário atual de Data Science
- O impacto de Machine Learning em Data Science
 - Uma nova onda de investimentos
 - As aplicações mais comuns
 - As ferramentas mais poderosas

Módulo II – Projetos de Machine Learning

- A natureza dos projetos de Machine Learning
 - Custos e benefícios estratégicos
 - Sponsorship
 - Life cycle iterativo
 - Etapas
- Os principais desafios
 - Dados de treinamento, validação e teste
 - Atributos irrelevantes
 - Limpeza de dados
 - Lidando com valores nulos
 - Lidando com atributos categóricos e textos
 - Lidando com outliers
 - Lidando com muitas dimensões, redução de dimensões
 - Visualizando os dados
 - Overfitting
 - Underfitting
 - Bias/Variance Tradeoff

Módulo III – Curso Rápido de Python para Machine Learning

- Ambiente de Programação
 - Anaconda
 - Jupyter Notebooks
- Os conceitos básicos
 - Estruturas básicas da linguagem: if/else, loops
 - Listas, dicionários, tuplas e conjuntos
- Programação funcional e orientada a objetos
 - Funções, Lambdas, Classes
 - Modules e Packages
 - Funções zip, enumerate, all, any, map/reduce/filter
- Numpy
 - Criação a partir do zero, criação a partir de listas
 - Indexing, slicing, reshaping
 - Concatenação e divisão (splitting)
 - Agregação, Min, Max e mais
 - Broadcasting
 - Comparação, masks, lógica booleana, sorting

- Pandas
 - Series, DataFrames
 - Indexação e Seleção em Series e em DataFrames
 - Operando com dados com valores nulos
 - Encontrando valores distintos
 - Explorando Series e DataFrames
 - Combinando dados
 - Agregação e agrupamento, pivot tables
 - eval() e query()
- Visualização de Dados
 - Matplotlib
 - Seaborn
- Preprocessamento de Dados com SciKit Learn
 - Limpeza de dados
 - Feature Engineering
 - Hot encoders
 - Scaling
 - Principal Component Analysis (PCA)
 - Pipelines
 - Estimator API

Módulo IV – Modelagem Analítica de Dados Estruturados

- Tipos de aprendizagem
 - Modelagens mais usadas
 - Preprocessamento dos dados
 - Medidas de desempenho
 - Cross-Validation
 - Learning Curves, validação dos modelos
 - Grid Search
 - Confusion Matrix
 - Precision/Recall Tradeoff
- Modelagens analíticas clássicas
 - Linear Regression com gradient descent
 - Logistics Regression
- Modelagens analíticas modernas
 - Support Vector Machines
 - Decision Trees
 - Random Forests
 - AdaBoost
 - Gradient Boosting

Módulo V – Deep Learning com Keras

- Ambiente de Programação
 - TensorFlow
 - Keras
- Conceitos com exemplos em Python e Keras
 - Fundamentos de Redes Neurais (neural networks)
 - Loss functions e optimizers
 - Scalars, vectors, matrices, tensores de alta dimensão
 - Data batch
 - Time series / dados sequenciais
 - Operações com tensores
 - Broadcasting
- Aplicações com Keras
 - Classificando imagens
 - Predição de preços de imóveis
 - Predição de preços de commodities
 - Classificando resenhas
 - Mais exemplos

**Exemplo de
um Projeto**

com uma estrutura real

História da Inteligência Artificial

- Décadas de 1950 e 1960
- Década de 1980
- Década de 2000
- Década de 2010

A grande mudança:

**Não fazer máquinas que fazem o que fazemos bem,
mas que fazem aquilo que não fazemos bem**

Real-time data steaming

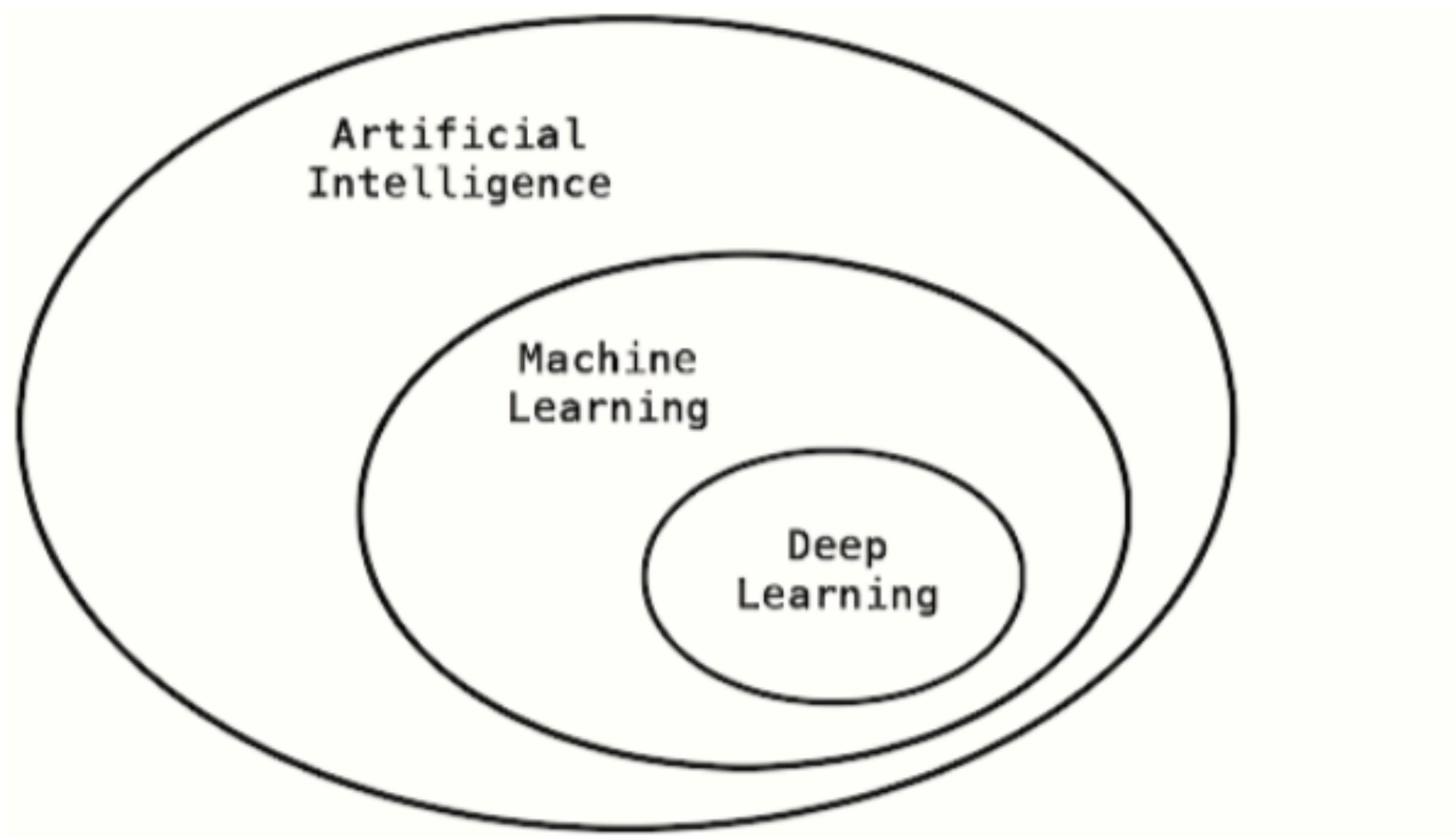
Big Data

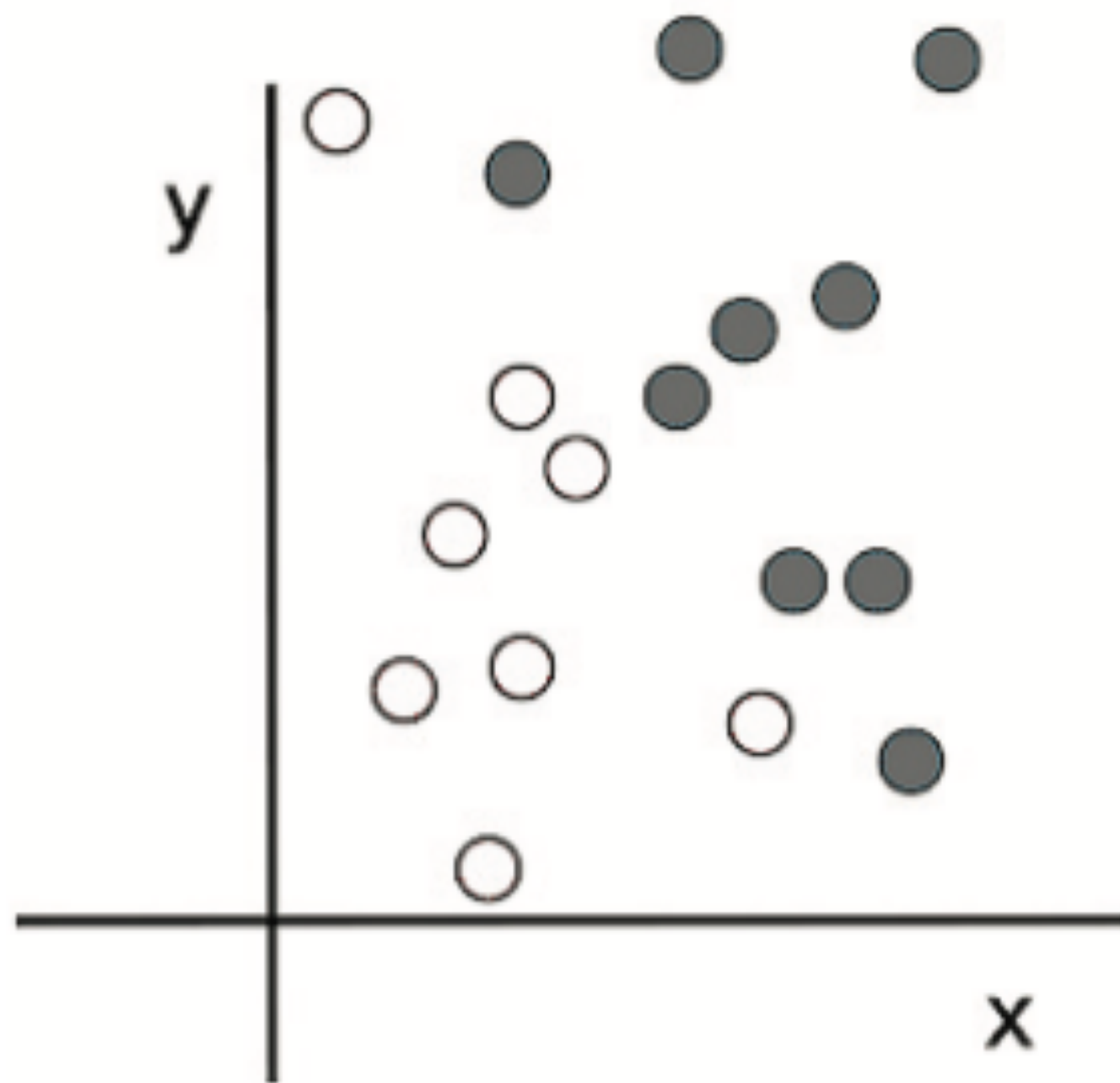
Machine Learning

Data Science

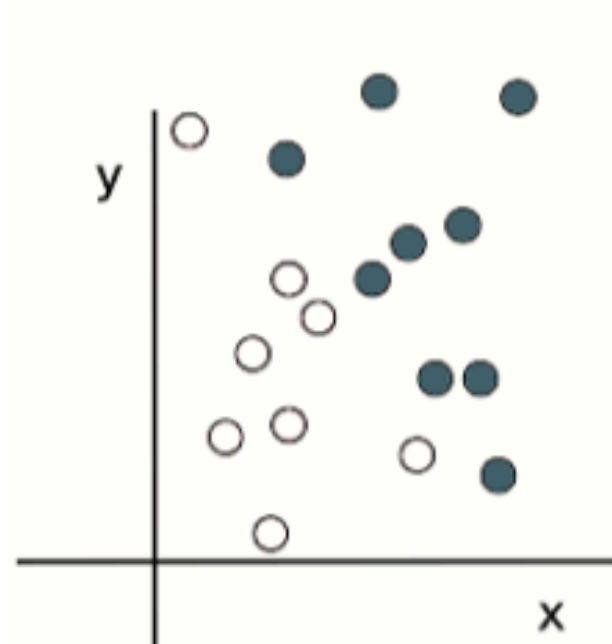
Estatística

Deep Learning

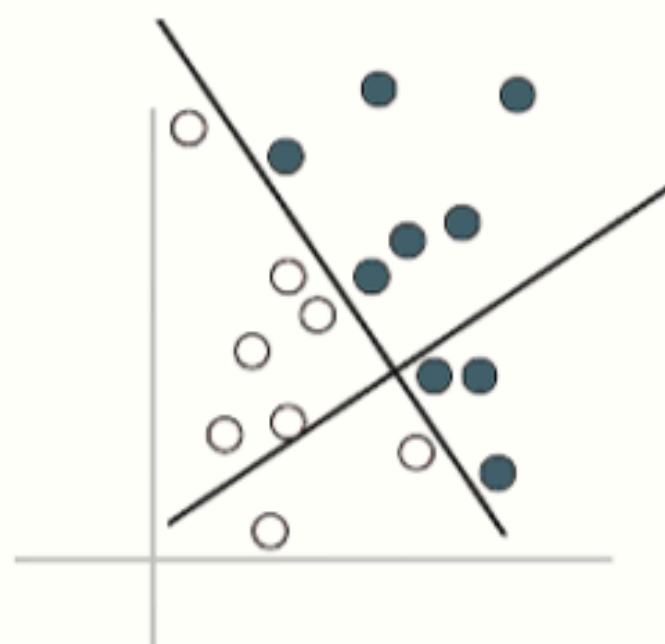




1: Raw data



2: Coordinate change



3: Better representation



Prática

- Qual a diferença entre Estatística e Ciência de Dados?
- O que é Big Data? Qual a relação dele com Ciência de Dados?
- O que é real-time data streaming (dados enviados em tempo real)?

Prática

- Qual a diferença entre AI, Ciência de Dados e Deep Learning?
- Qual o objetivo de um projeto de Ciência de Dados?
- Como se sabe se um projeto de Ciência de Dados teve sucesso ou não?

Aplicações

- Cite 5 aplicações de Ciência de Dados na sua empresa
- Cite 5 aplicações em geral de Ciência de Dados

Laboratório

- Cheque se Anaconda está instalado
- Comece um Jupyter Notebook
- Tente fazer algo com Python neste Jupyter Notebook
- Como instalar alguma package que falta?

```
import sys  
print(sys.version)
```

```
>>> import platform  
>>> platform.python_version()  
'2.6.2'
```

```
python --version
```


Perguntas?