



Data Science

Prof. Dr. Jorge Zavaleta

zavaleta.jorge@gmail.com



Sobre – Jorge Zavaleta



- Doctor en Ingeniería de Sistemas e Computación - UFRJ
- Maestro en Ciencias de la Computación – UFRGS
- Licenciado en Matemática - UNT
- Arias de Interés e Investigación:
 - Inteligencia Artificial
 - *Machine Learning*
 - *Deep Learning*
 - Data Science Aplicados a:
Educación, Salud y Meteorología.
 - Lenguajes Java, R y Python.



Agenda

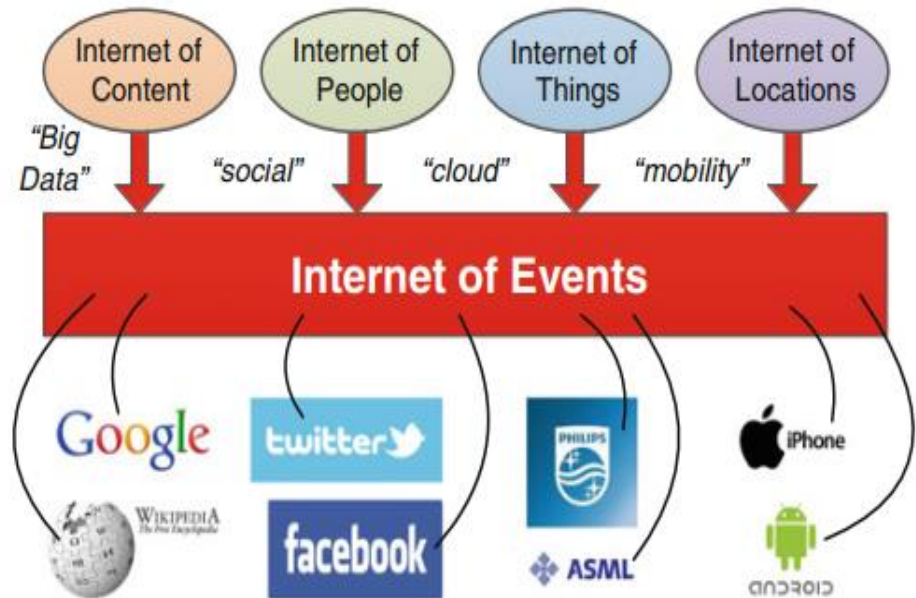
- Introducción.
- Que es *Data Science*?
- Pré-requisitos para ser Data Scientist
 - Cursos necesarios
- Machine Learning
- Deep Learning
- Lenguajes de Programación
- Perspectivas e Oportunidades
- DS en la Practica

Introducción

- A informatización de todos servicios, desde las sofisticadas transacciones en bolsas de valores a las simples compras de un café, asociada a las redes sociales y a los dispositivos móviles (*tablets, smart-phones*) producen una enorme cantidad de datos.
- A grande cantidad de datos, a tasa de actualización de esos mismos datos es también enorme.
- A capacidad de procesamiento también ha tenido aumentos significativos.

Introducción

- El **grande volumen de datos** compensado por el **aumento de la capacidad de procesamiento** han originado **nuevos conceptos**, como **Big Data** y la creación de nuevas profesiones como los **científico de datos**.



©Wil M. P. van der Aalst

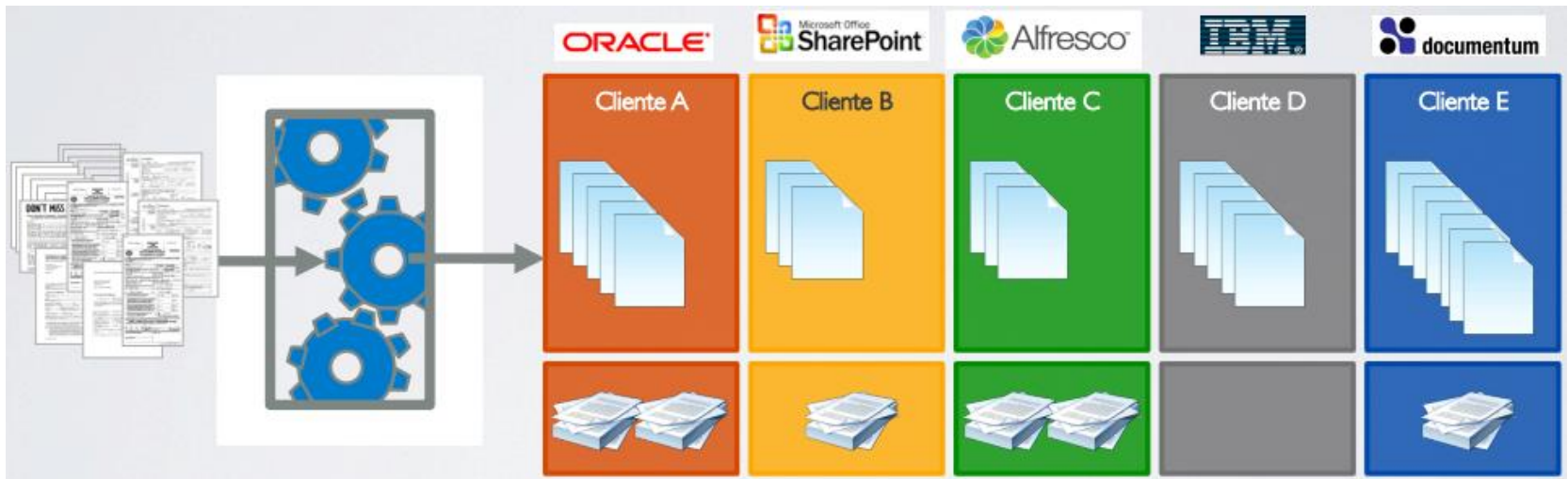
Big Data

- **Big data** es un concepto, en el cual el foco es almacenar el grande volumen de los datos oriundo de todos los medios, aliados a la mayor velocidad de crecimiento de estas informaciones.
- Comenzó a ser percibido y consolidado en la última década con el aumento relevante en la utilización de computadores, notebook y todos los tipos de dispositivos, principales generadores y replicadores de datos.



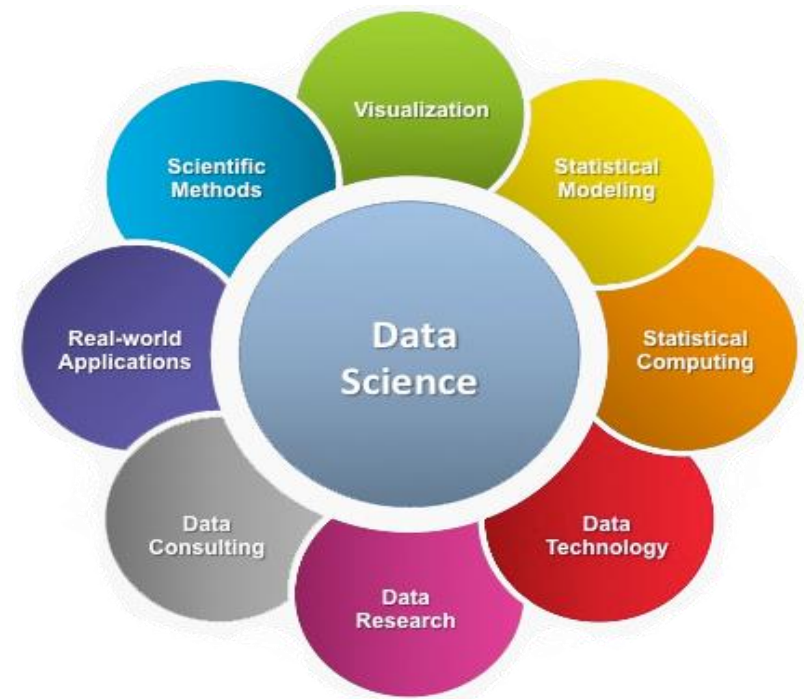
Big Data

- **Datos analizados e interpretados** sobre determinada óptica, y a partir del análisis, se torna posible **calificar, clasificar, medir, cuantificar**, etc.



Que es Data Science?

- **Ciencia de Datos (Data Science)** es el estudio de los datos.
- Ciencia de Datos, es el actual termino para la ciencia que **analiza datos**, combinando la **estadística** con **aprendizaje de máquina/minería de datos** y **tecnologías de base de datos**, para responder al desafío que presenta **Big Data**.



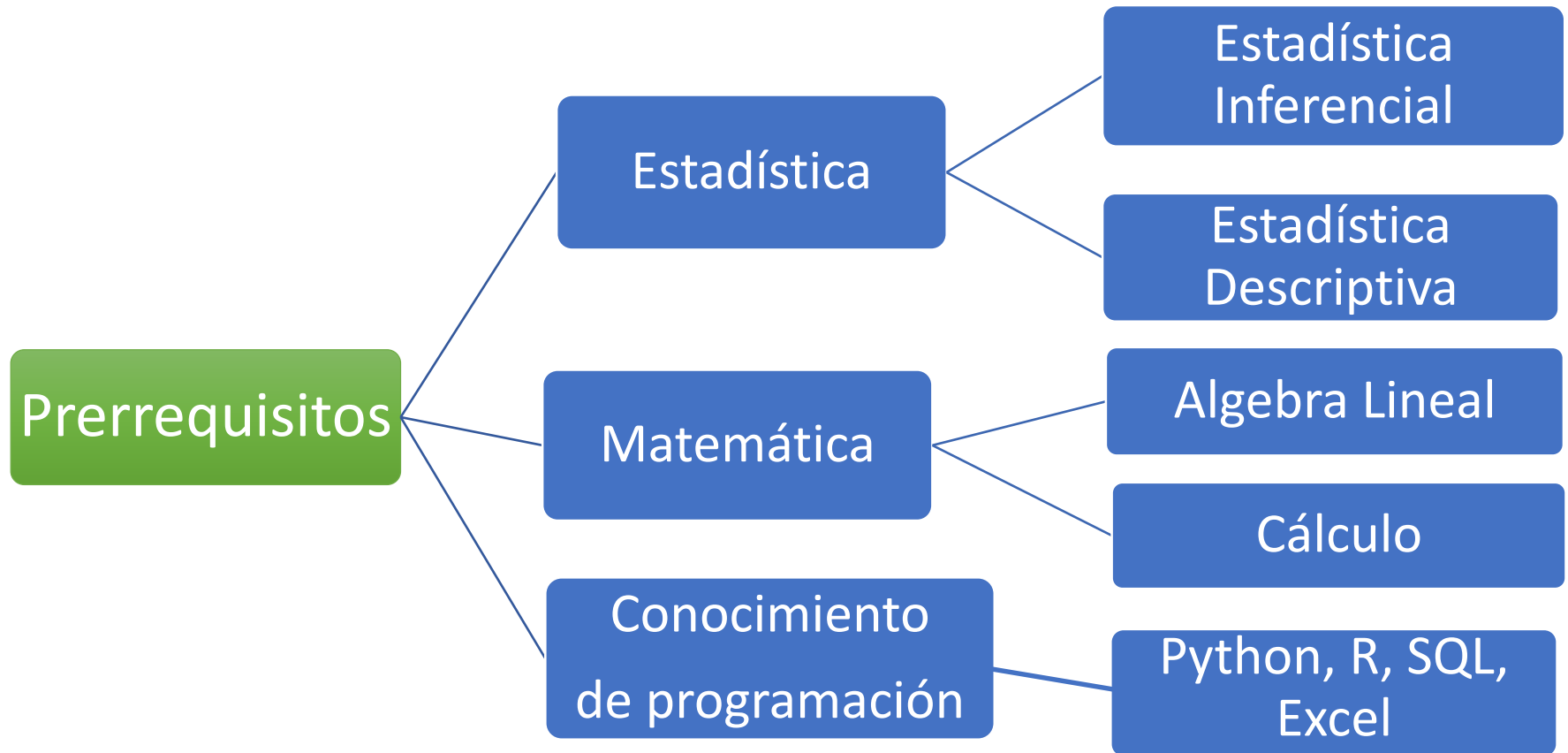
Que es Ciencia de Datos?

- La **Ciencia de Datos** incorpora una amplia variedad de campos y es construida sobre técnicas y teorías de muchos campos, incluyendo:
- **matemática, estadística, ingeniería de datos, reconocimiento de patrones y aprendizaje, computación avanzada, visualización, modelaje de incertezas, *data warehouse*, y computación de alto desempeño.**
- con el objetivo de extraer significado(valor) a los datos y crear productos sobre esos datos.

Requisitos para ser Científico de D.

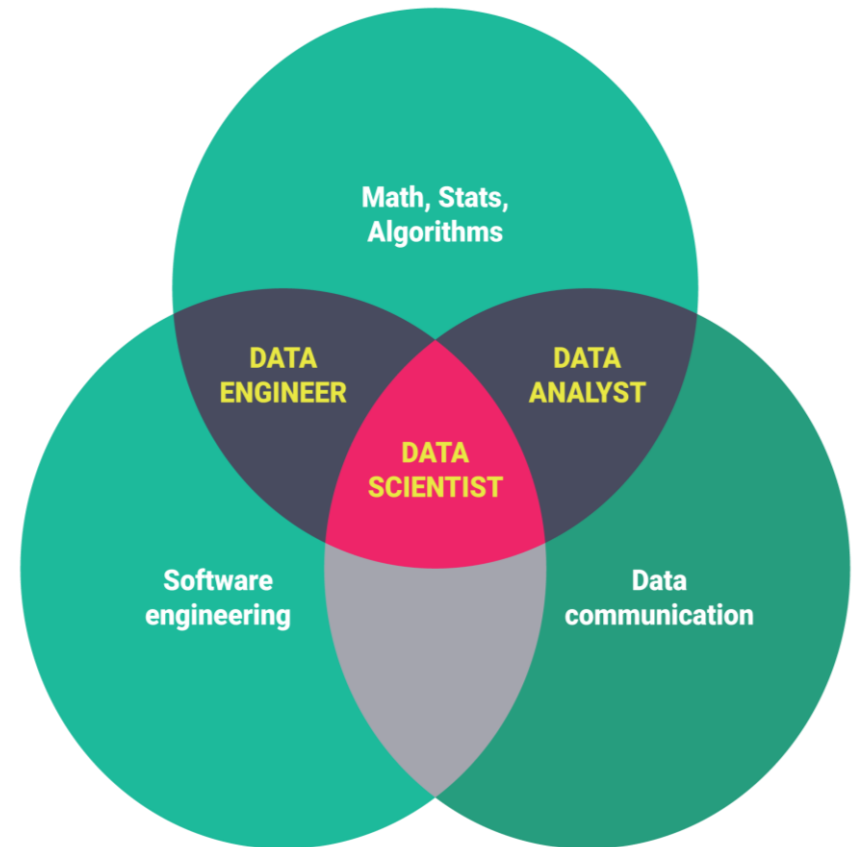
- Por **científico de datos** entiéndese:
“alguien mejor en estadística que un ingeniero informático y alguien mejor en programación de que un matemático”
- El científico de Datos debe tener una **fuerte visión de negocios**, juntamente con la **capacidad de comunicar los resultados**, tanto para los líderes de negocios cuanto para sus pares, de una forma que influencie como una organización se posiciona delante de los desafíos del mercado.

Prerrequisitos para ser Científico D.

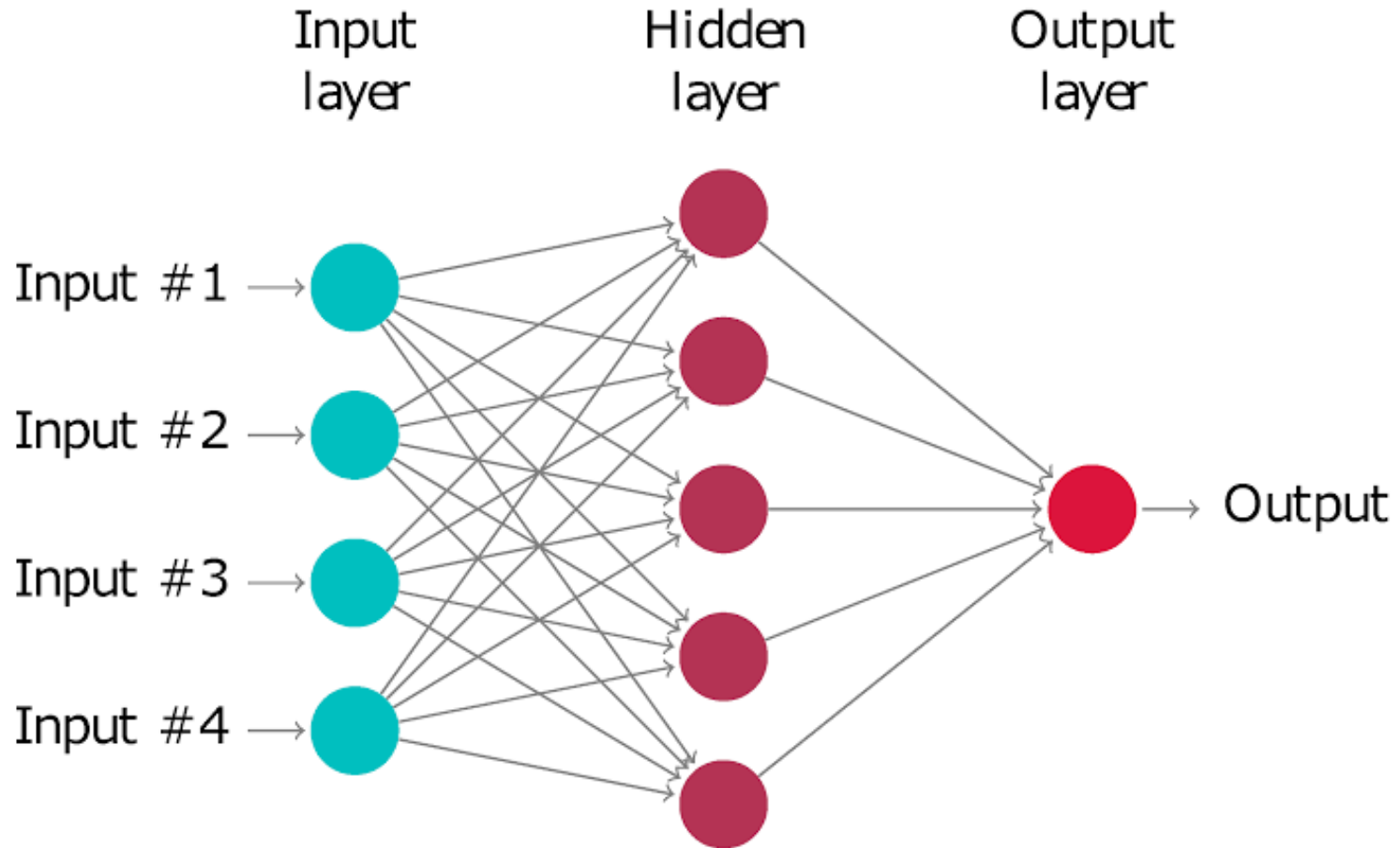


Científico de datos

- **DATA SCIENTIST** es alguien que sabe diferenciar un teste de hipótesis *t-student* de un *chi-cuadrado*, al mismo tiempo que sabe ver a diferencia entre un algoritmo polinomial de orden **$O(N)$** y de **$O(N^2)$** .



Machine Learning – Neural Network



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)



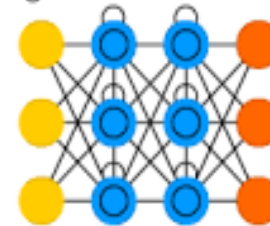
Deep Feed Forward (DFF)



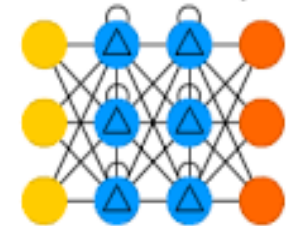
Recurrent Neural Network (RNN)



Long / Short Term Memory (LSTM)



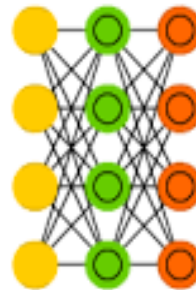
Gated Recurrent Unit (GRU)



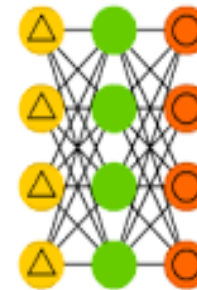
Auto Encoder (AE)



Variational AE (VAE)



Denoising AE (DAE)



Sparse AE (SAE)





Types of Machine Learning

Supervised Learning

Classification

- Fraud detection
- Email Spam Detection
- Diagnostics
- Image Classification

Regression

- Risk Assessment
- Score Prediction

Unsupervised Learning

Dimensionality Reduction

- Text Mining
- Face Recognition
- Big Data Visualization
- Image Recognition

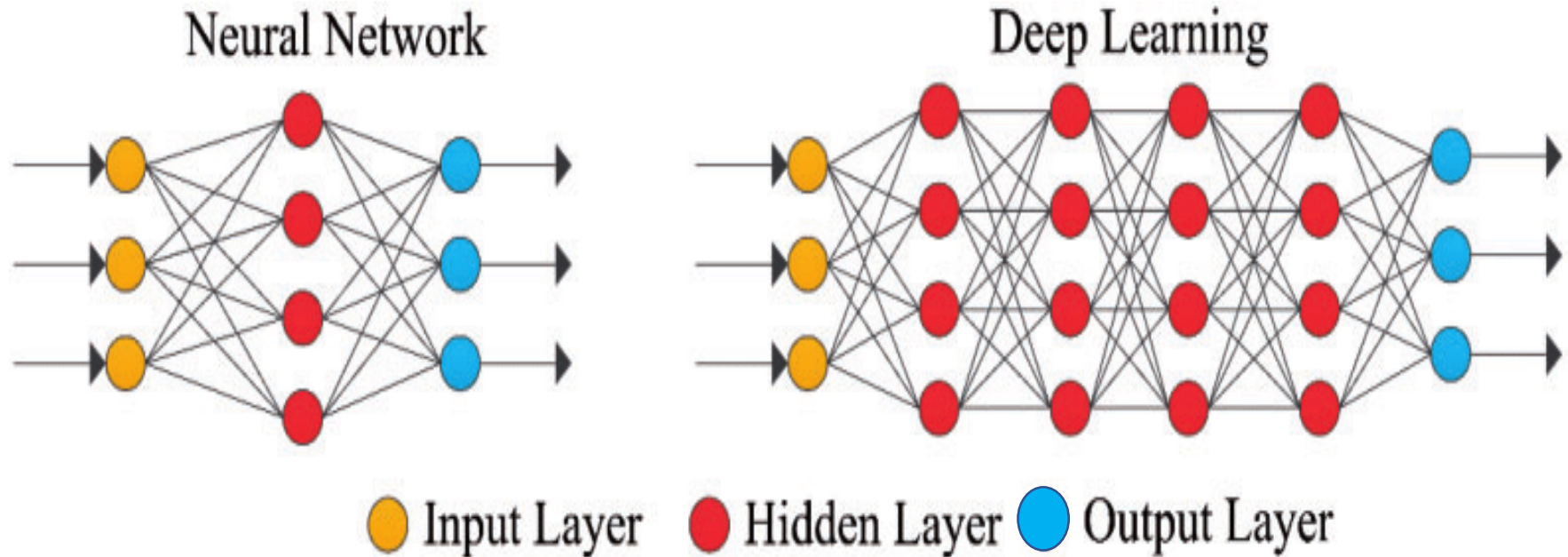
Clustering

- Biology
- City Planning
- Targetted Marketing

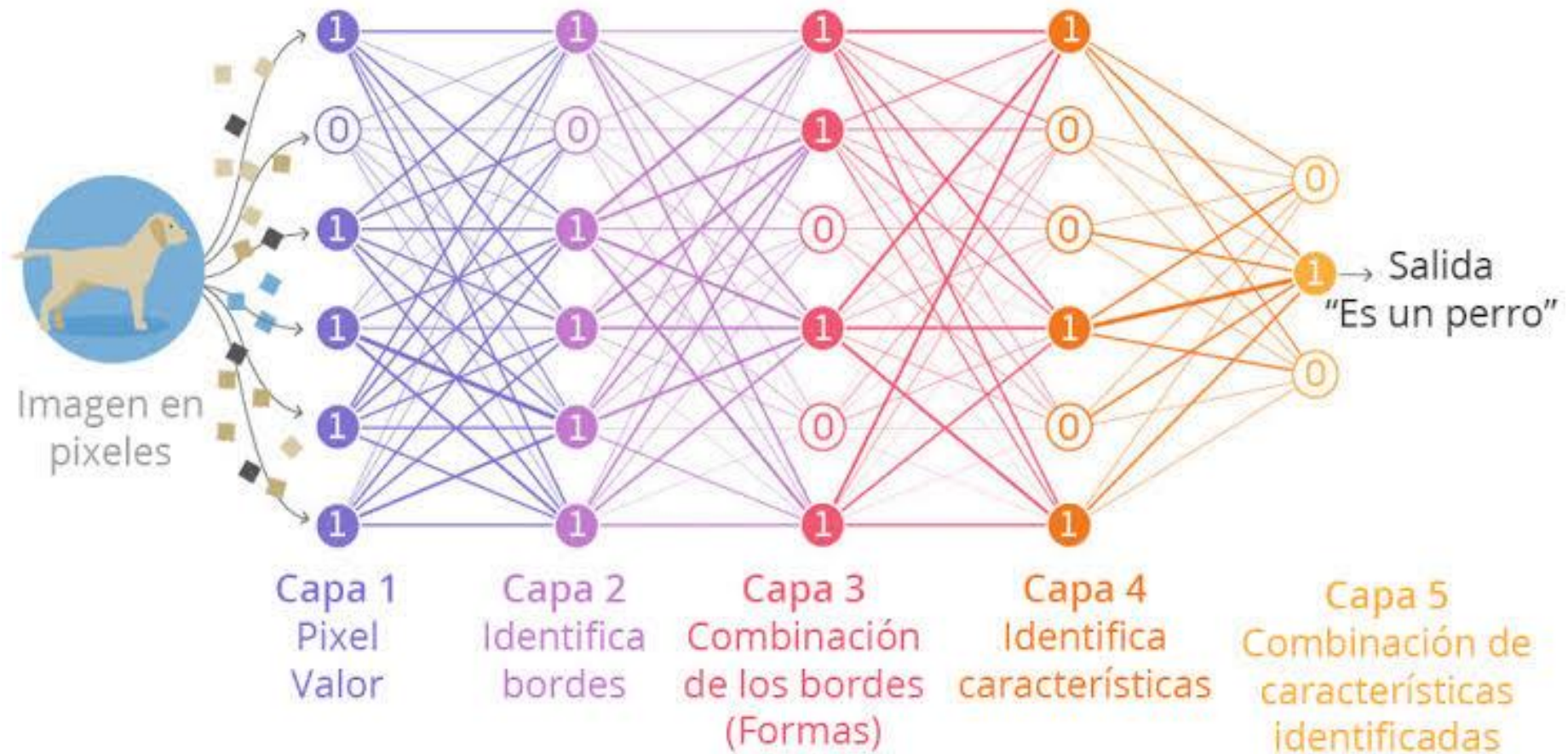
Reinforcement Learning

- Gaming
- Finance Sector
- Manufacturing
- Inventory Management
- Robot Navigation

Deep Learning

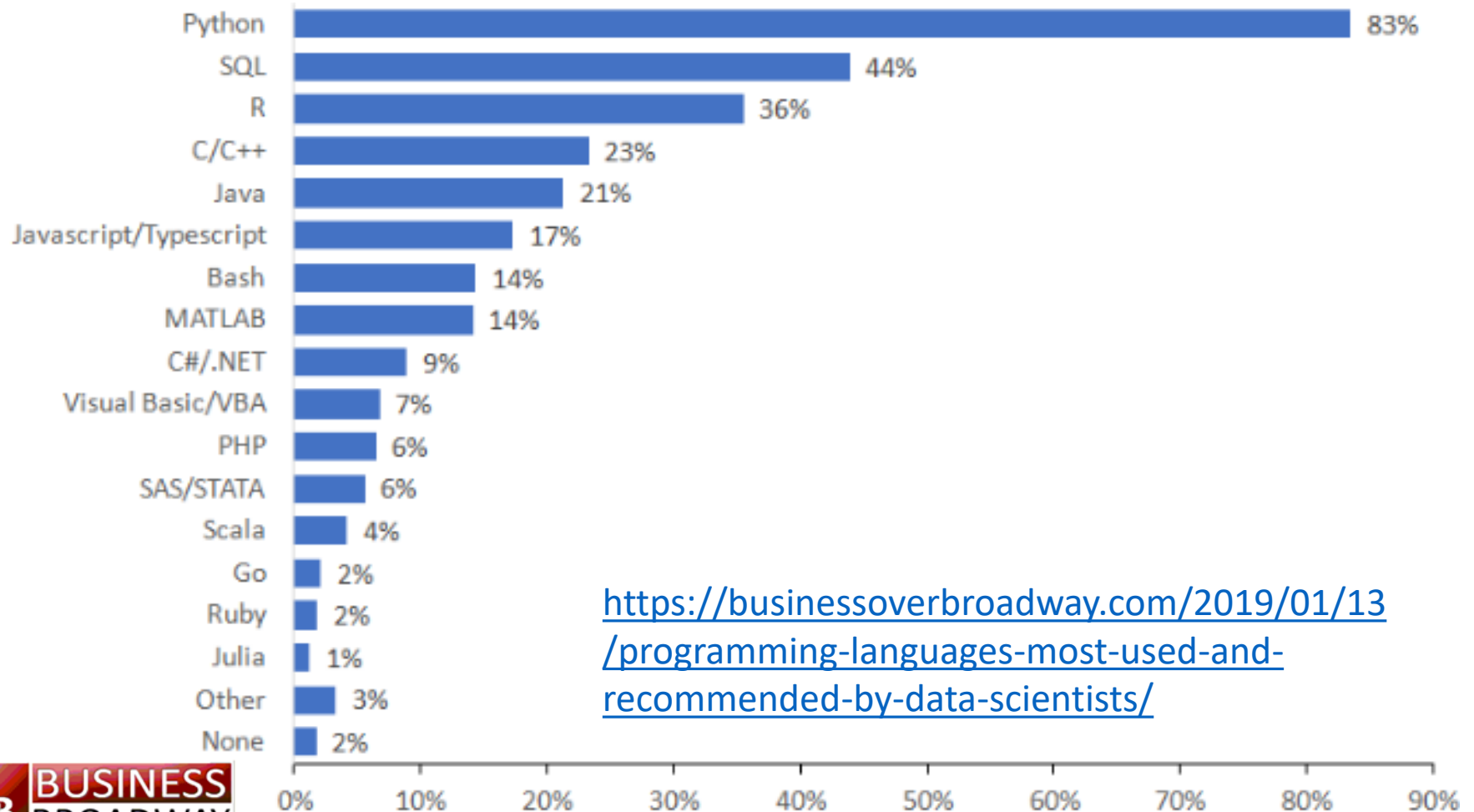


Deep Learning



Fuente: <https://www.quantamagazine.org/>

Lenguajes de Programación



<https://businessoverbroadway.com/2019/01/13/programming-languages-most-used-and-recommended-by-data-scientists/>

Lenguajes de Programación

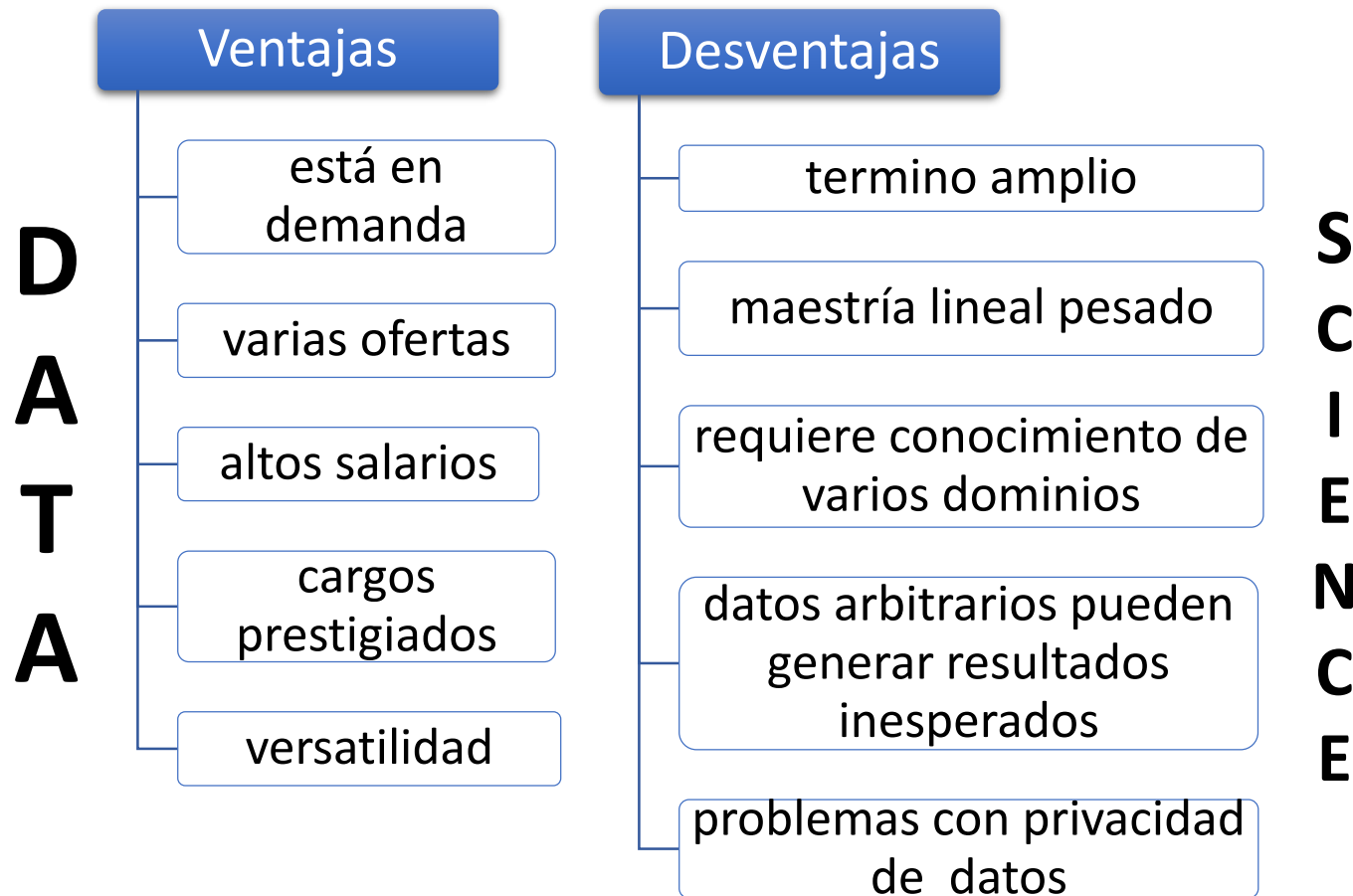
1. Python
2. R
3. Java
4. SQL
5. Julia
6. Scala
7. MatLab
8. TensorFlow
9. SAS



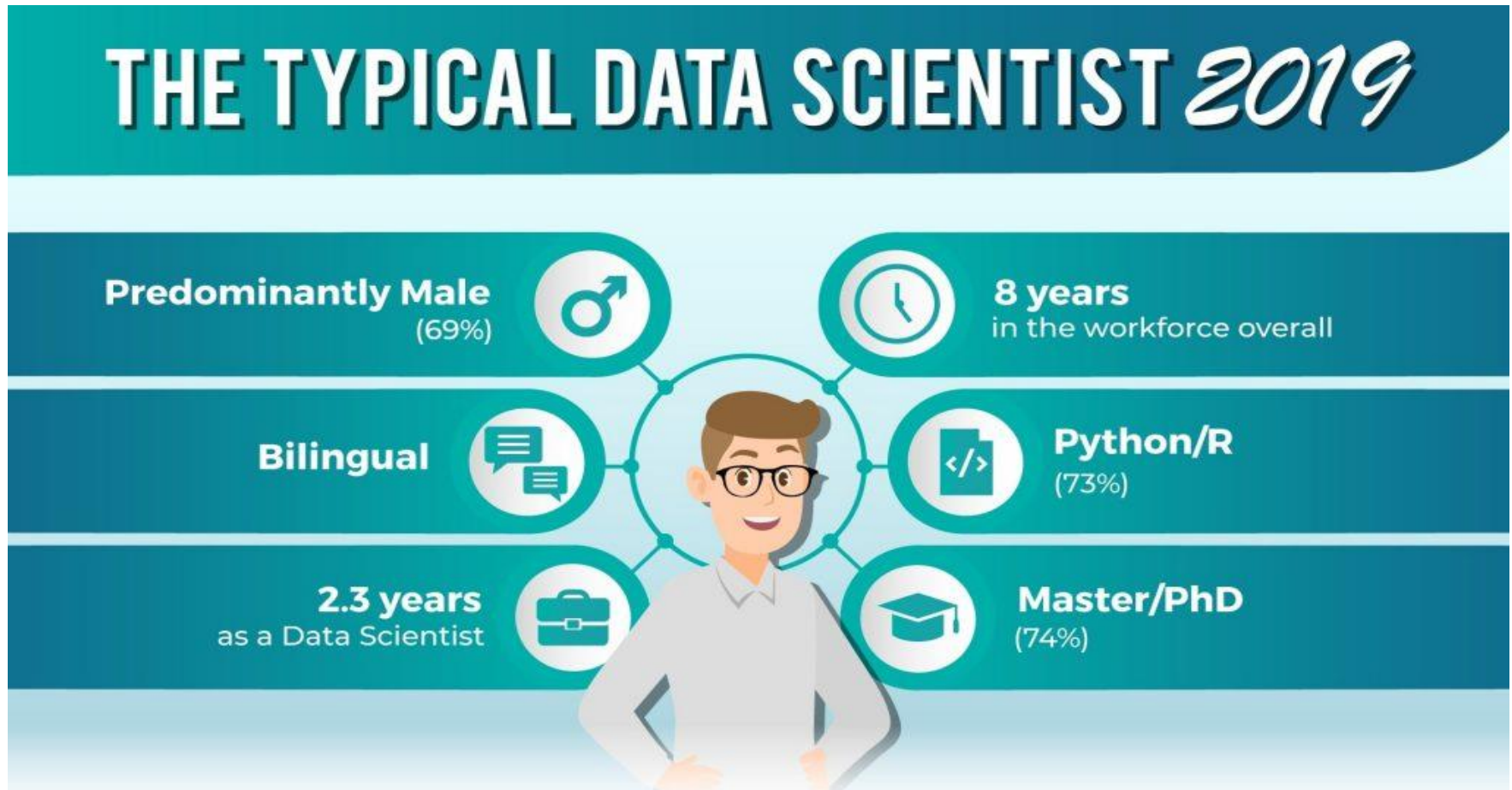
Perspectivas y Oportunidades

- Bancos
- Finanzas
- Manufactura
- Transporte
- Cuidados de la Salud
 - Análisis de imágenes médicas, ciencia de datos del genoma, descubrimiento de drogas, *Predictive Modeling for Diagnosis, Natural Language Processing (NLP)*.
- Comercio electrónico.

Perspectivas y Oportunidades



Perspectivas y Oportunidades



DS – En la practica

- Python (3.7 ou 2.7)
 - Anaconda
- Jupyter Notebook
 - Desktop
 - Nuvem
- JupyterLab
 - Desktop
- Colab (Nuvem)
 - colab.research.google.com



Browser tabs: Gmail, Bridge, Google Drive/Congressos, Aprendizado_maquina_01, JupyterLab

Address bar: localhost:8888/notebooks/Google%20Drive/Congressos_e_Palestras/Palestras_2019/Niteroi/Aprend...

File Edit View Insert Cell Kernel Widgets Help

Not Trusted Python 3

In []:

Exemplo 03 - Reconhecimento de DÍGITOS

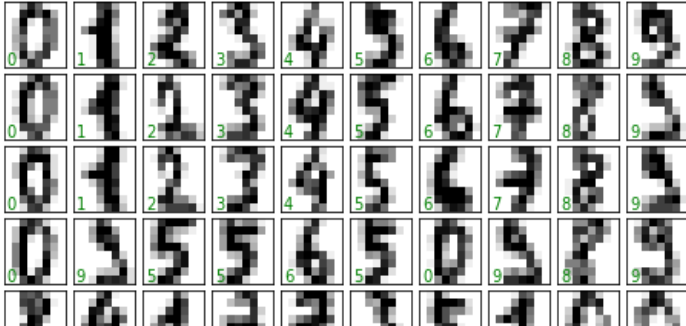
In [168]:

```
from sklearn.datasets import load_digits
digits = load_digits()
digits.images.shape
```

Out[168]: (1797, 8, 8)

In [139]:

```
# Visualização grafico do dataset
fig, axes = plt.subplots(10, 10, figsize=(8, 8), subplot_kw={'xticks':[], 'yticks':[]},
                        gridspec_kw=dict(hspace=0.1, wspace=0.1))
#
for i, ax in enumerate(axes.flat):
    ax.imshow(digits.images[i], cmap='binary', interpolation='nearest')
    ax.text(0.05, 0.05, str(digits.target[i]), transform=ax.transAxes, color='green')
```



Browser tabs: Gmail, Bridge, Google Drive/Congressos, Aprendizado_maquina_01, JupyterLab

Address bar: localhost:8888/lab

File Edit View Run Kernel Tabs Settings Help

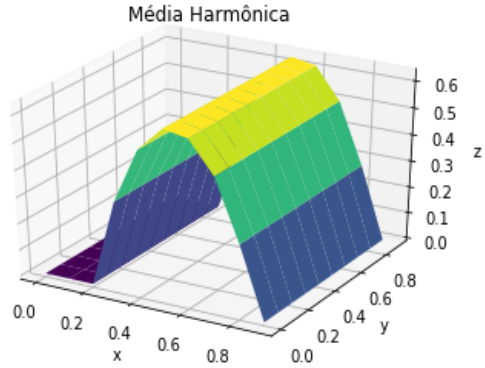
File browser: / ... / ModelSimulaSist /Codigo /

Name	Last Modified
Lab_9_operacaoe...	5 months ago
Lab_aula3.pdf	7 months ago
Lab_Aula7_8_nu...	5 months ago
Lab-Aula6_Oper...	6 months ago
MS_Lab_Aula01....	8 months ago
MS_Lab_Aula01....	8 months ago
MS_Lab_Aula02....	6 months ago
MS_Lab_Aula02....	8 months ago
MS_Lab_Aula04....	7 months ago
MSS_Lab_Aula_...	2 months ago
MSS_Lab_Aula_...	2 months ago
MSS_Lab_Aula0...	a month ago
MSS_Lab_Aula0...	a month ago
MSS_Lab_Aula0...	7 months ago
MSS_Lab_Aula0...	7 months ago
MSS_Lab_aula5...	7 months ago
Teste_JupyterLa...	21 days ago
teste.md	23 days ago
teste.txt	23 days ago
Trabalho_02.pdf	7 months ago

Open files: Teste_JupyterLab.ipynb, MSS_Lab_Aula03.ipynb, MS_Lab_Aula04.ipynb

Code editor (Python 3):

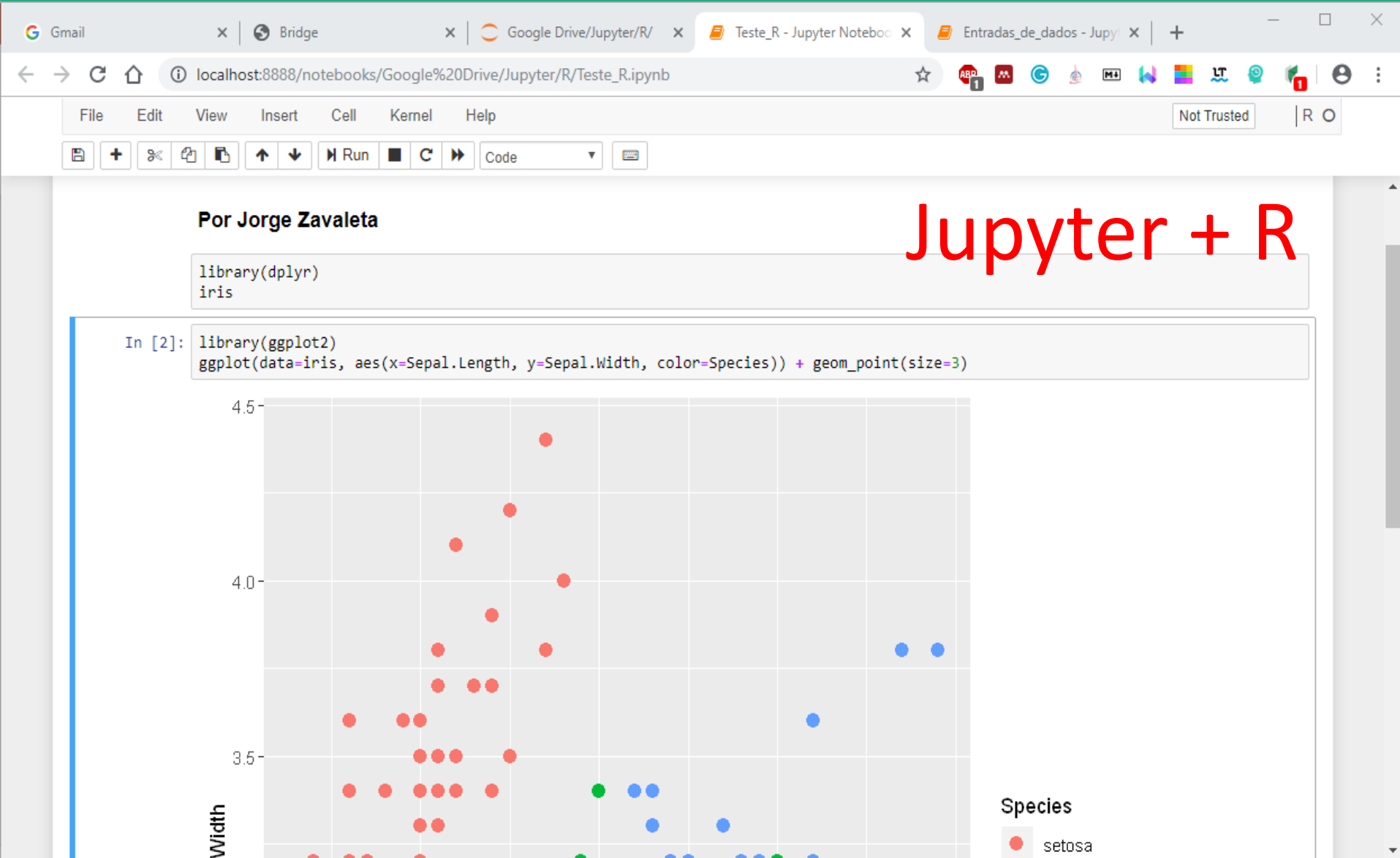
```
[106]: # Criando a superficie em 3D
fig = plt.figure()
ax1 = plt.axes(projection='3d')
# Criando um Plot básico
ax1.plot_surface(X, Y, H[0], rstride=1, cstride=1, cmap='viridis', edgecolor='none')
ax1.set_title('Média Harmônica');
ax1.set_xlabel('x')
ax1.set_ylabel('y')
ax1.set_zlabel('z');
# Exibindo o gráfico criado para a média geométrica
plt.show()
```



```
[107]: # Criando a superficie em 3D
fig = plt.figure()
ax1 = plt.axes(projection='3d')
# Criando um Plot básico
```

Mode: Command | Ln 11, Col 11 | MS_Lab_Aula04.ipynb

JupyterLab



colab.research.google.com/drive/1HL0uZcgrzYbufE8hr7hO2ZK5YGMNYuGD#scrollTo=XnnvEeqsBTR

Aprendizado_maquina_01.ipynb

File Edit View Insert Runtime Tools Help

+ Code + Text

Connect Editing

▼ Aprendizado supervisionado - Regressão Linear

▼ Exemplo 01

```
[ ] 1 import pandas as pd
    2 import matplotlib.pyplot as plt
    3 %matplotlib inline
```

```
[ ] 1 # Load data
    2 df0 = pd.read_csv('Data/Grade_Set_1.csv')
    3 print(df0)
```

	Hours_Studied	Test_Grade
0	2	57
1	3	66
2	4	73
3	5	76
4	6	79
5	7	81
6	8	90
7	9	96
8	10	100

```
[ ] 1 # plot scatter
    2 df0.plot(kind='scatter', x='Hours_Studied', y='Test_Grade', title='Grau vs Horas estudadas', color='b')
    3 plt.xlabel('Horas estudadas')
```

colab

Finalizando ...

- Muito Obrigado
 - Muchas Gracias
 - Thank you Very much
 - Merci Beaucoup