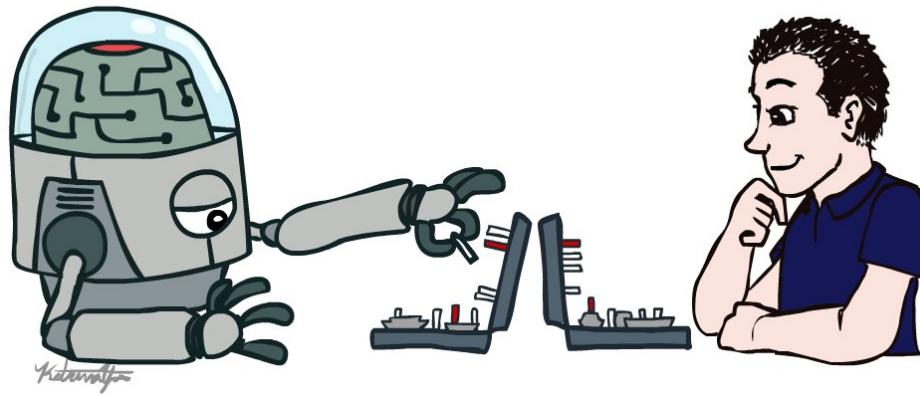
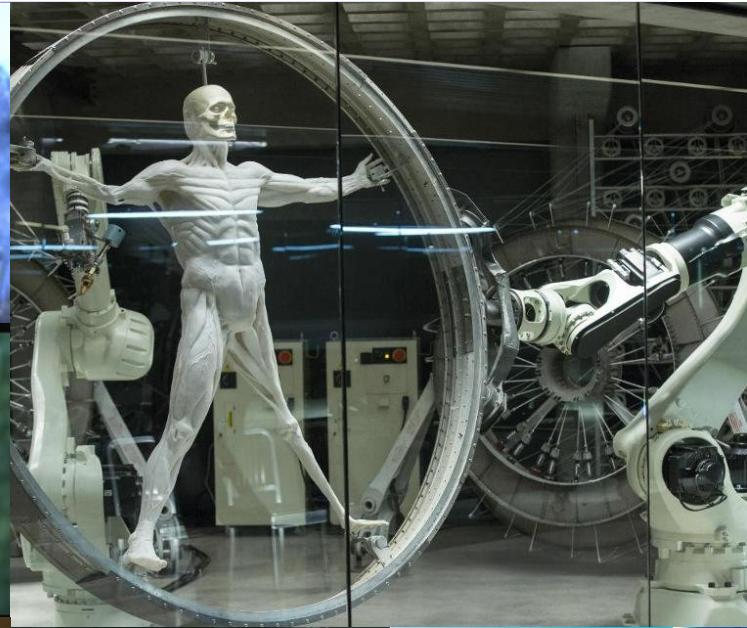
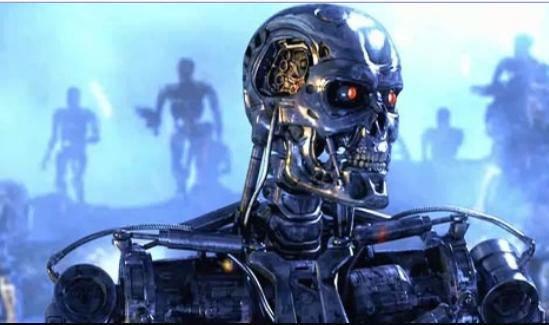


# Artificial Intelligence

## Introduction



# Sci-Fi AI?





**TUG**  
CAUTION  
MAY CONTAIN  
CHEMOTHERAPY DRUG

CAUTION  
MAY CONTAIN  
CHEMOTHERAPY DRUG



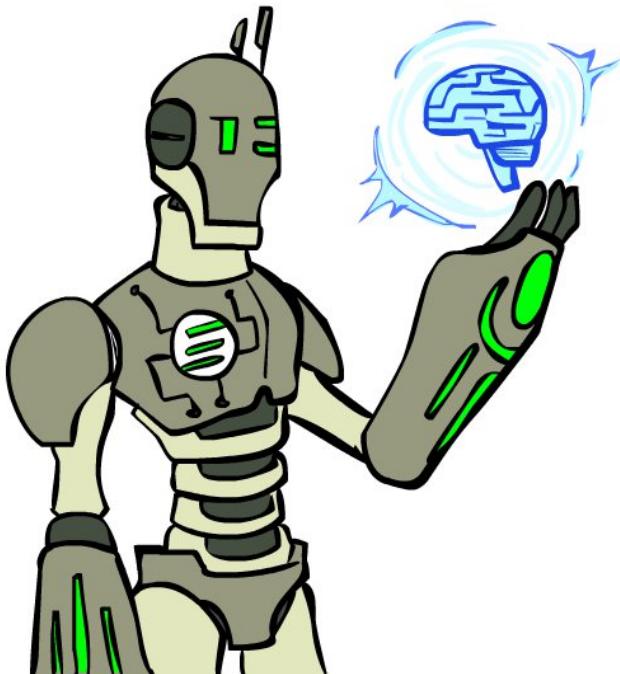
# Today

---

- o What is artificial intelligence?

- o What can AI do?

- o What is this course?



# What is AI?

---

The science of making machines that:

# Rational Decisions

---

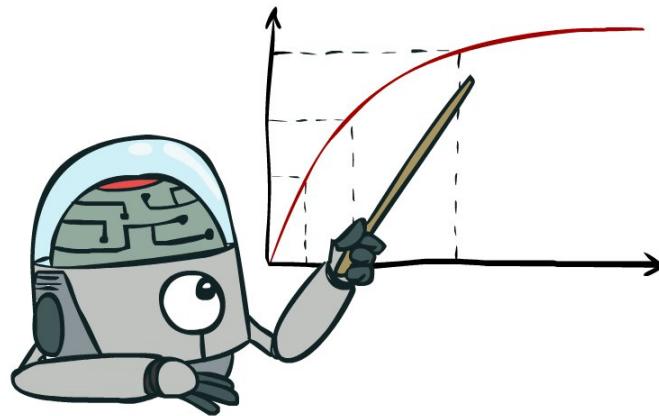
We'll use the term **rational** in a very specific, technical way:

- Rational: maximally achieving pre-defined goals
- Rationality only concerns what decisions are made  
(not the thought process behind them)
- Goals are expressed in terms of the **utility** of outcomes
- Being rational means **maximizing your expected utility**

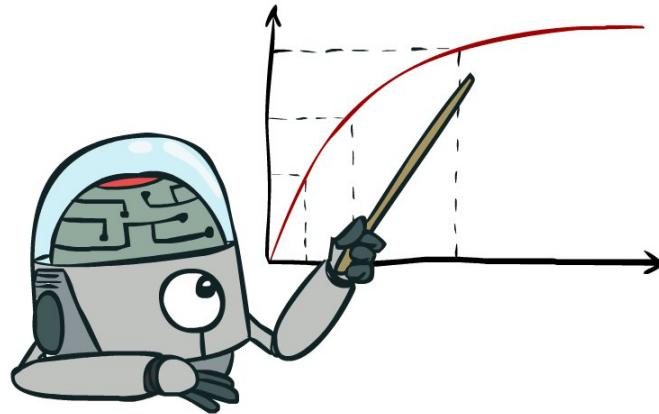
A better title for this course would be:

**Computational Rationality**

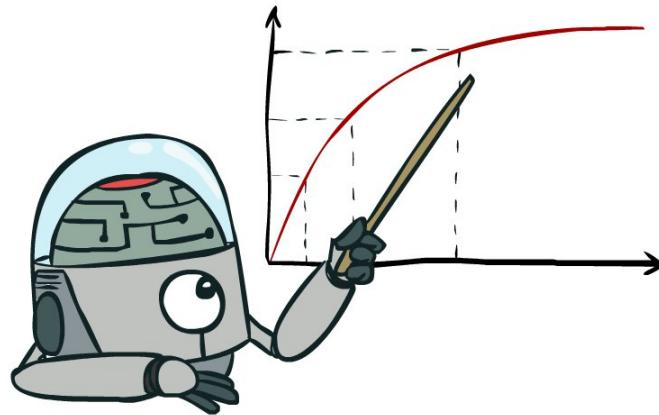
# Maximize Your Expected Utility



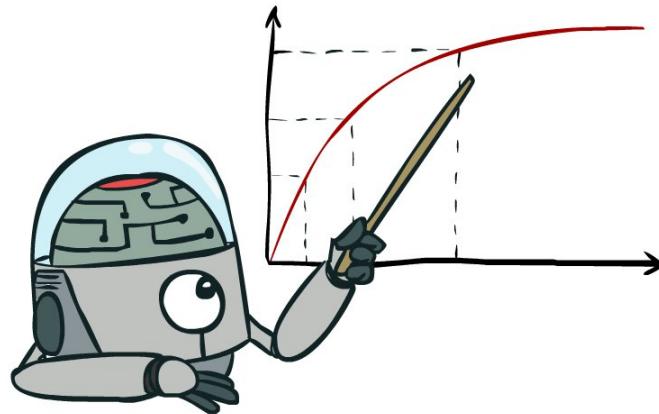
# Maximize Your Expected Utility



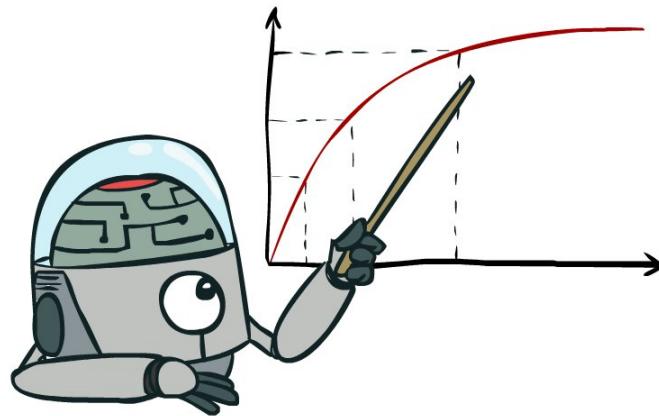
# Maximize Your Expected Utility



# Maximize Your Expected Utility



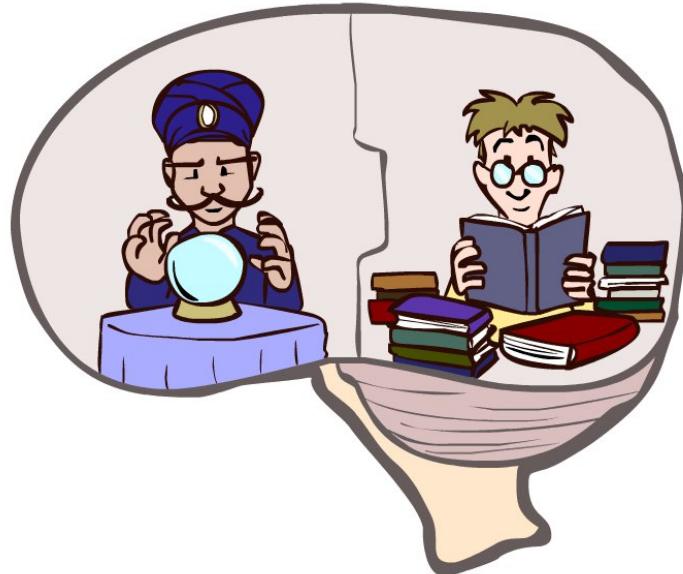
# Maximize Your Expected Utility



# What About the Brain?

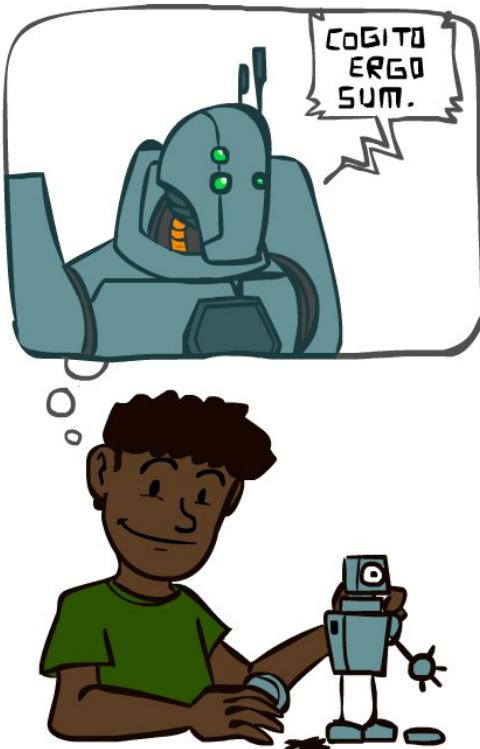
---

- Brains (human minds) are very good at making rational decisions, but not perfect
- Brains aren't as modular as software, so hard to reverse engineer!
- “Brains are to intelligence as wings are to flight”
- Lessons learned from the brain: memory (data) and simulation (computation) are key to decision making



# A (Short) History of AI

---

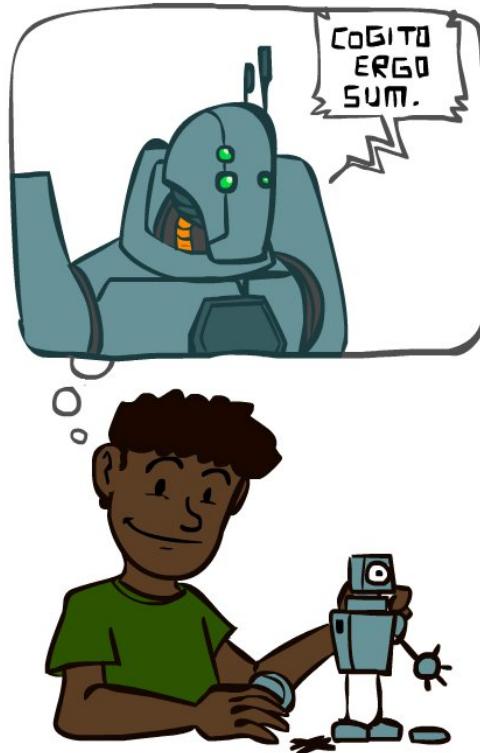




# A (Short) History of AI

---

- o 1940—1950: Early days
  - o 1943: McCulloch & Pitts: Boolean circuit model of brain
  - o 1950: Turing's "Computing Machinery and Intelligence"
- o 1950—70: Excitement: Look, Ma, no hands!
  - o 1950s: Early AI programs, including Samuel's checkers program, Newell & Simon's Logic Theorist, Gelernter's Geometry Engine
  - o 1956: Dartmouth meeting: "Artificial Intelligence" adopted
  - o 1965: Robinson's complete algorithm for logical reasoning
- o 1970—90: Knowledge-based approaches
  - o 1969—79: Early development of knowledge-based systems
  - o 1980—88: Expert systems industry booms
  - o 1988—93: Expert systems industry busts: "AI Winter"
- o 1990—2012: Statistical approaches + subfield expertise
  - o Resurgence of probability, focus on uncertainty
  - o General increase in technical depth
  - o Agents and learning systems... "AI Spring"?
- o 2012—: Excitement: Look, Ma, no hands!
  - o Big data, big compute, neural networks
  - o Some re-unification of subfields
  - o AI used in many industries

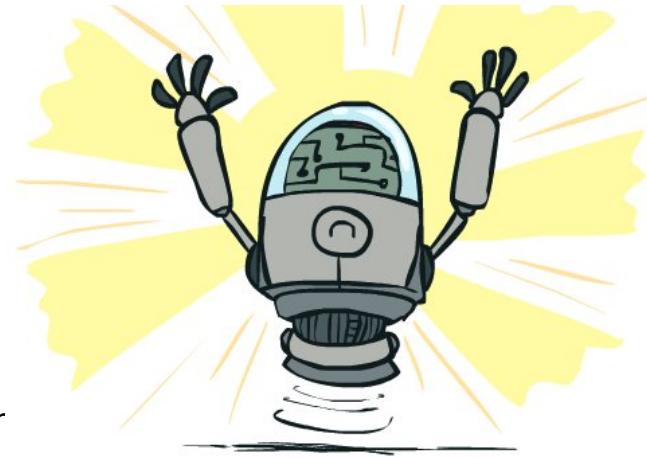


# What Can AI Do?

---

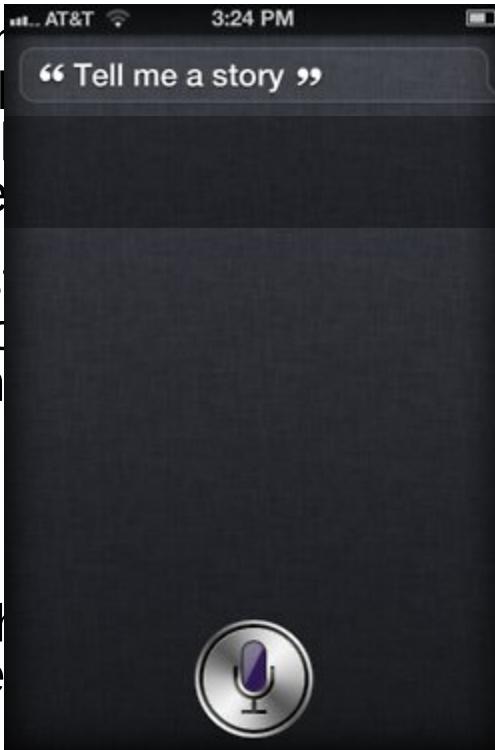
Quiz: Which of the following can be done at present?

- Play a decent game of Jeopardy?
- Win against any human at chess?
- Win against the best humans at Go?
- Play a decent game of tennis?
- Grab a particular cup and put it on a shelf?
- Unload any dishwasher in any home?
- Drive safely along the highway?
- Drive safely along Telegraph Avenue?
- Buy a week's worth of groceries on the web?
- Buy a week's worth of groceries at Berkeley Bowl?
- Discover and prove a new mathematical theorem?
- Perform a surgical operation?
- Translate spoken Chinese into spoken English in real time?
- Write an intentionally funny story?
- 
- 
- 
- 



# Unintentionally Funny Stories

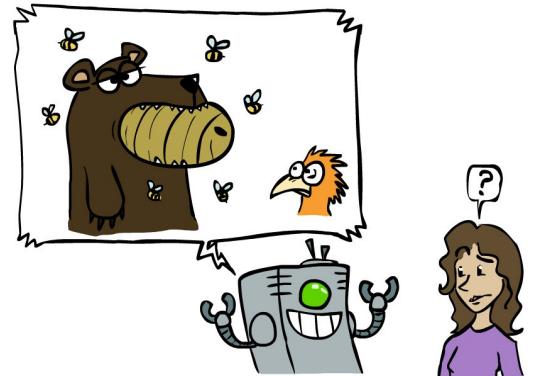
- o One day Joe Bear was hanging out with Irving Bird where some birds were flying overhead. They noticed that there was a beehive in the oak tree. He ate the honey and got stung.
- o Henry Squirrel was thirsty and was sitting on the river bank where his good friend Henry was sitting. Henry slipped and fell in the water. The End.
- o Once upon a time there was a crow who was sitting in a tree. He noticed that he was hungry and swallowed the cheese he found in his mouth. The End.



friend  
d him  
ked to

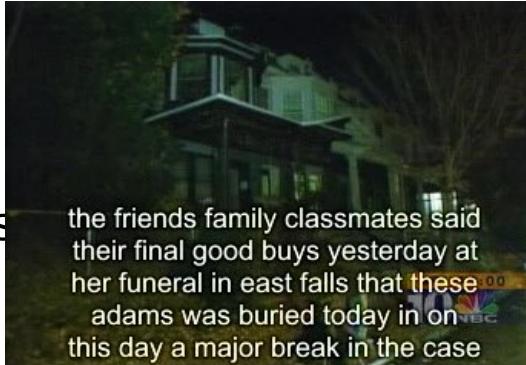
o the  
sitting.  
owned.

and a vain crow. One day he had a piece of cheese in his mouth. He became hungry, so he flew over to the crow. The End.



# Natural Language

- o Speech technologies (e.g. Siri)
  - o Automatic speech recognition (ASR)
  - o Text-to-speech synthesis (TTS)
  - o Dialog systems
- o Language processing technologies
  - o Question answering
  - o Machine translation



## "Il est impossible aux journalistes de rentrer dans les régions tibétaines"

Bruno Philip, correspondant du "Monde" en Chine, estime que les journalistes de l'AFP qui ont été expulsés de la province tibétaine du Qinghai "n'étaient pas dans l'ilégalité".



- o Les faits Le dalaï-lama dénonce l'"enfer" imposé au Tibet depuis sa fuite, en 1959
- o Vidéo Anniversaire de la rébellion

## "It is impossible for journalists to enter Tibetan areas"

Philip Bruno, correspondent for "World" in China, said that journalists of the AFP who have been deported from the Tibetan province of Qinghai "were not illegal."

- Facts The Dalai Lama denounces the "hell" imposed since he fled Tibet in 1959  
Video Anniversary of the Tibetan rebellion: China on guard



<https://play.aidungeon.io/>

# Computer Vision

---



"man in black shirt is playing guitar."



"construction worker in orange safety vest is working on road."



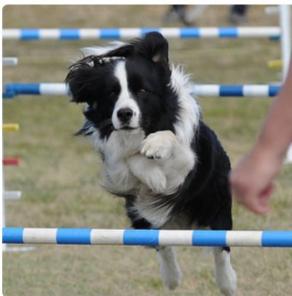
"two young girls are playing with lego toy."



"boy is doing backflip on wakeboard."



"girl in pink dress is jumping in air."



"black and white dog jumps over bar."



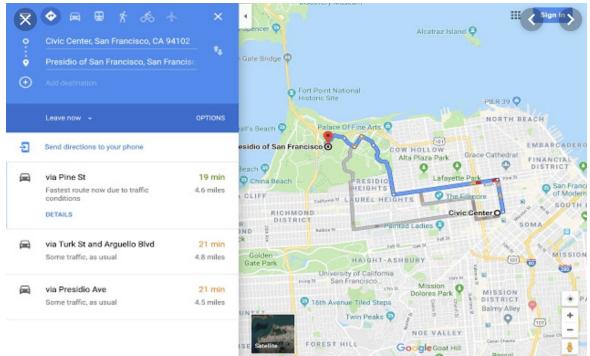
"young girl in pink shirt is swinging on swing."



"man in blue wetsuit is surfing on wave."

Karpathy & Fei-Fei, 2015; Donahue et al., 2015; Xu et al, 2015; many more

# Tools for Predictions & Decisions



Berkeley, CA 94709  
Tuesday 2:00 PM  
Mostly Sunny



# Game Agents

---

- o Classic Moment: May, '97: Deep Blue vs. Kasparov
  - o First match won against world champion
  - o “Intelligent creative” play
  - o 200 million board positions per second
  - o Humans understood 99.9 of Deep Blue's moves
  - o Can do about the same now with a PC cluster
- o 1996: Kasparov Beats Deep Blue  
“I could feel --- I could smell --- a new kind of intelligence across the table.”
- o 1997: Deep Blue Beats Kasparov  
“Deep Blue hasn't proven anything.”





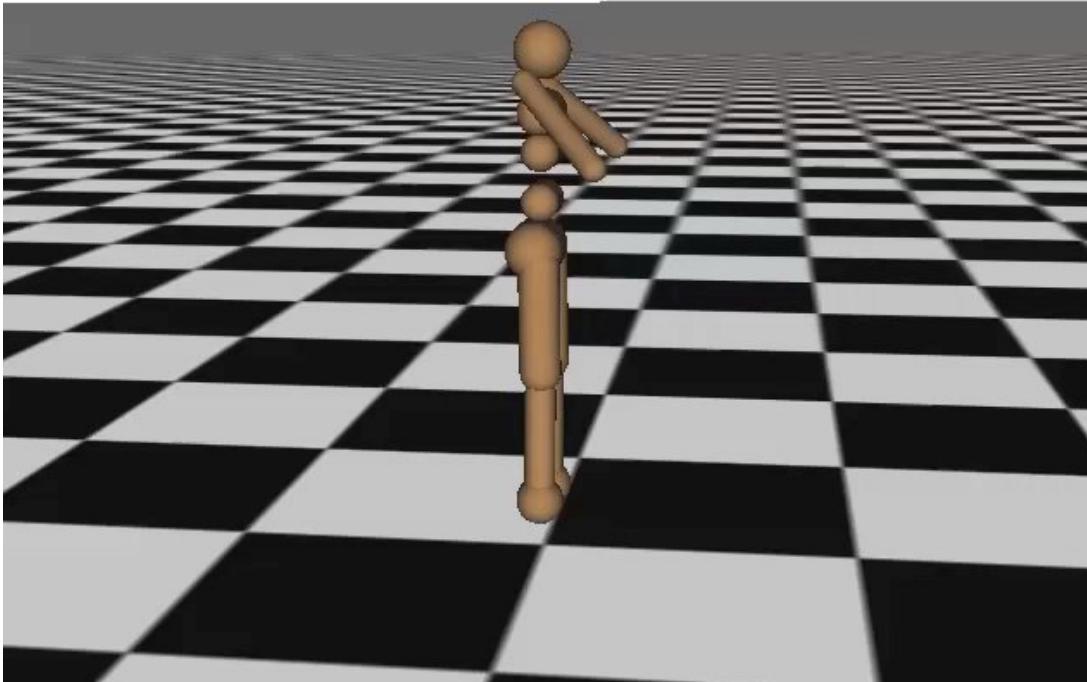


Photo: Google / Getty Images

# Simulated Agents

---

Iteration 0



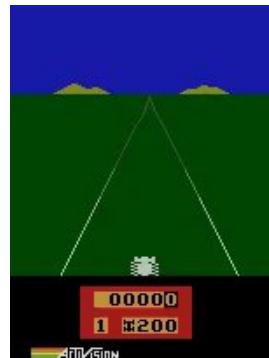
# Game Agents

---

## o Reinforcement learning



Pong



Enduro



Beamrider

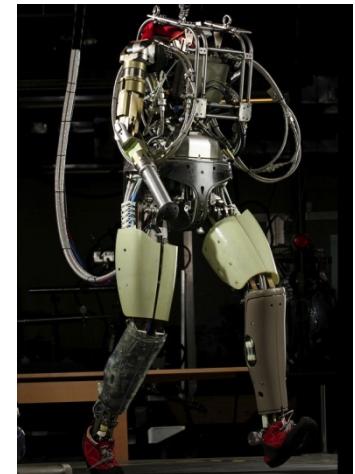


Q\*bert

# Robotics

---

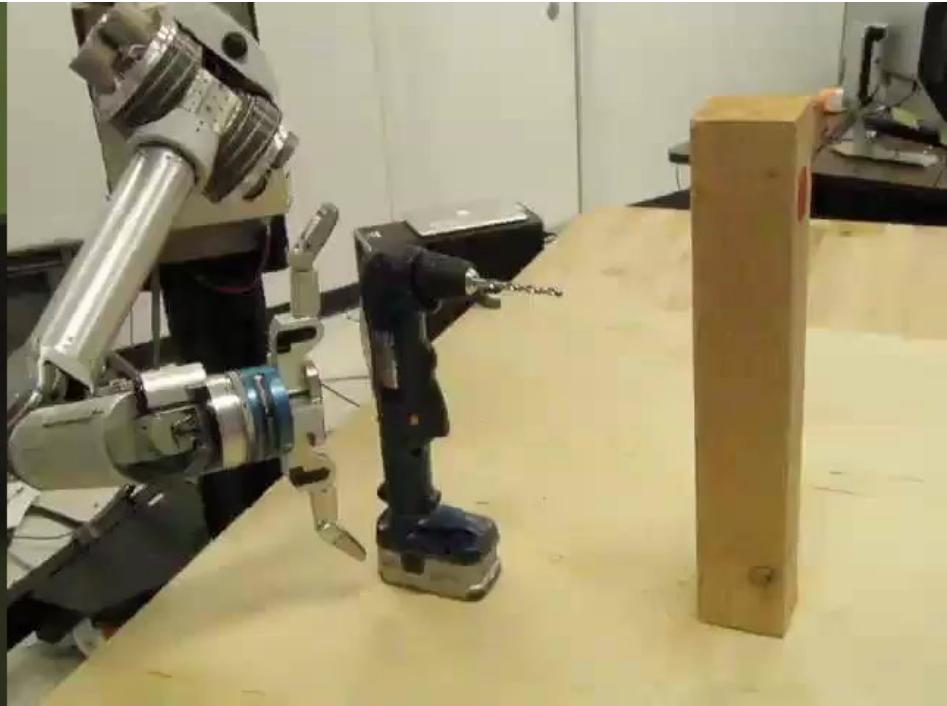
- o Robotics
  - o Part mech. eng.
  - o Part AI
  - o Reality much harder than simulations!
- o Technologies
  - o Vehicles
  - o Rescue
  - o Help in the home
  - o Lots of automation...
- o In this class:
  - o We ignore mechanical aspects
  - o Methods for planning
  - o Methods for control



Images from UC Berkeley, Boston Dynamics, RoboCup, Google

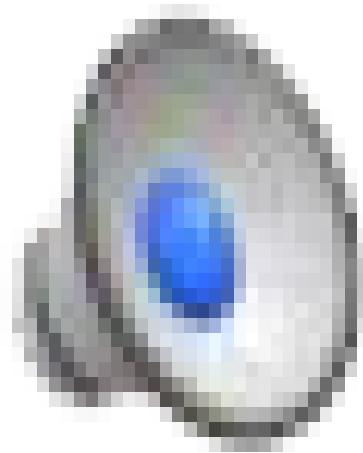
# Robots

---



# Robots

---



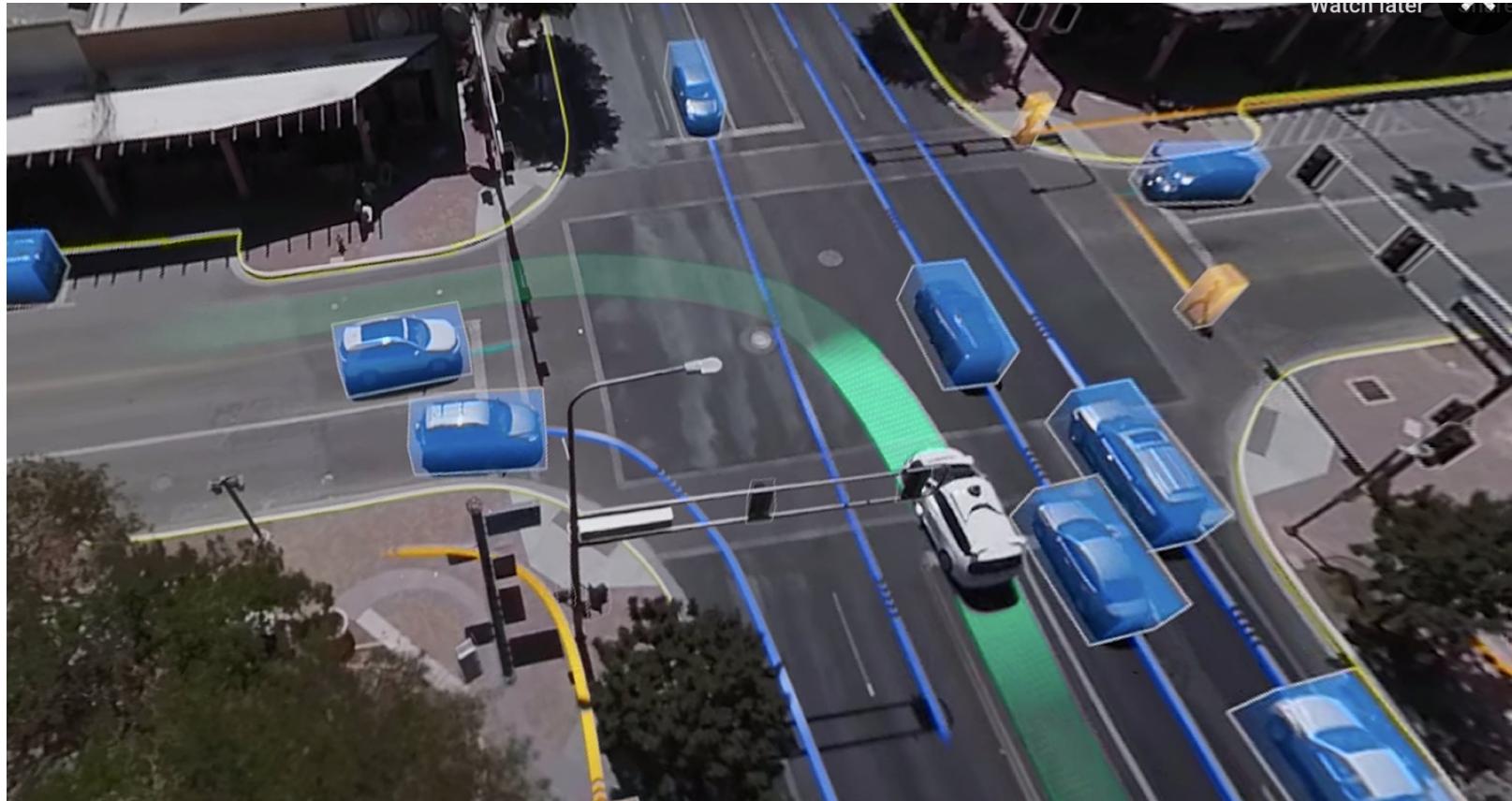
# Human-AI Interaction

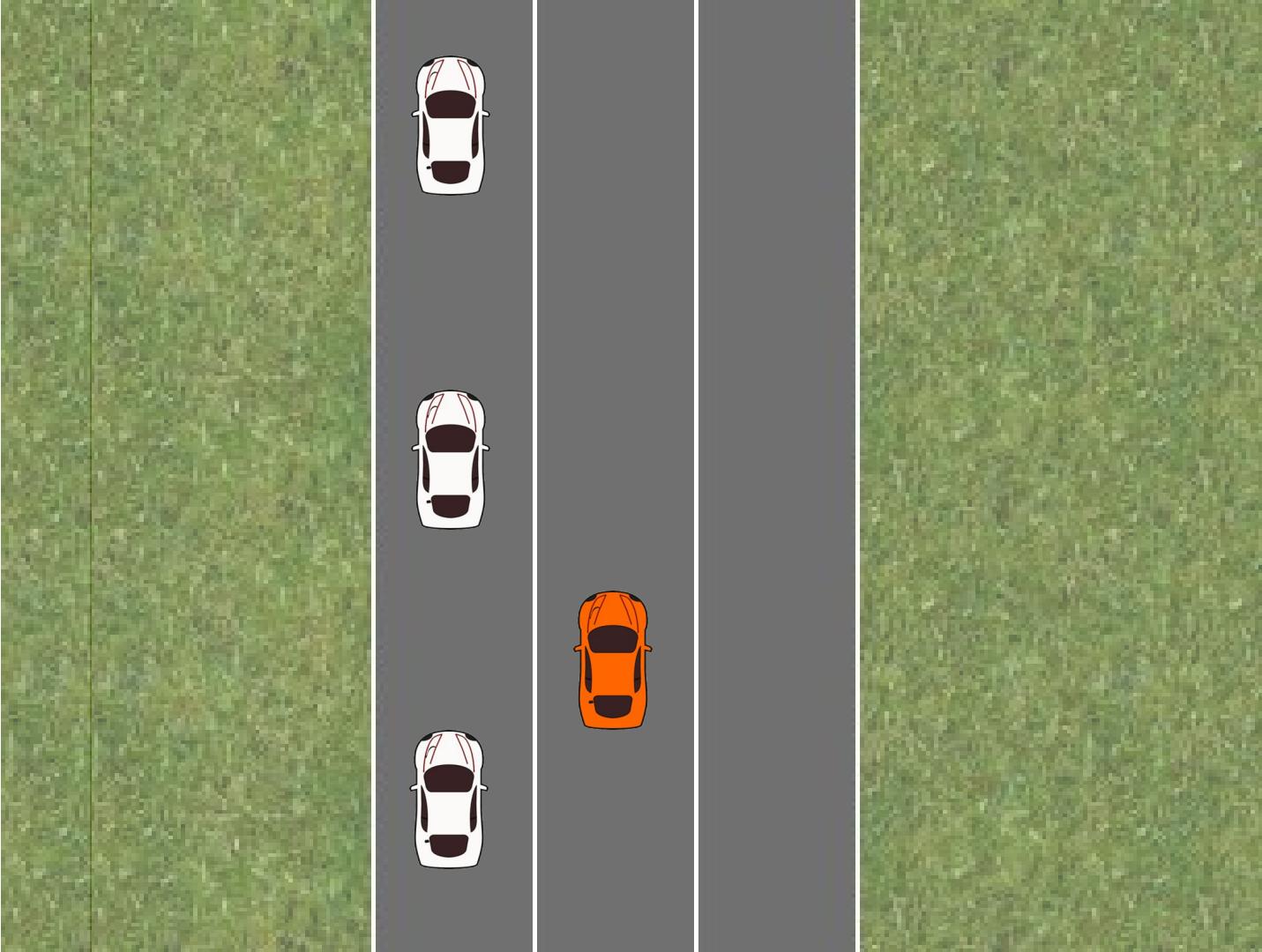
---

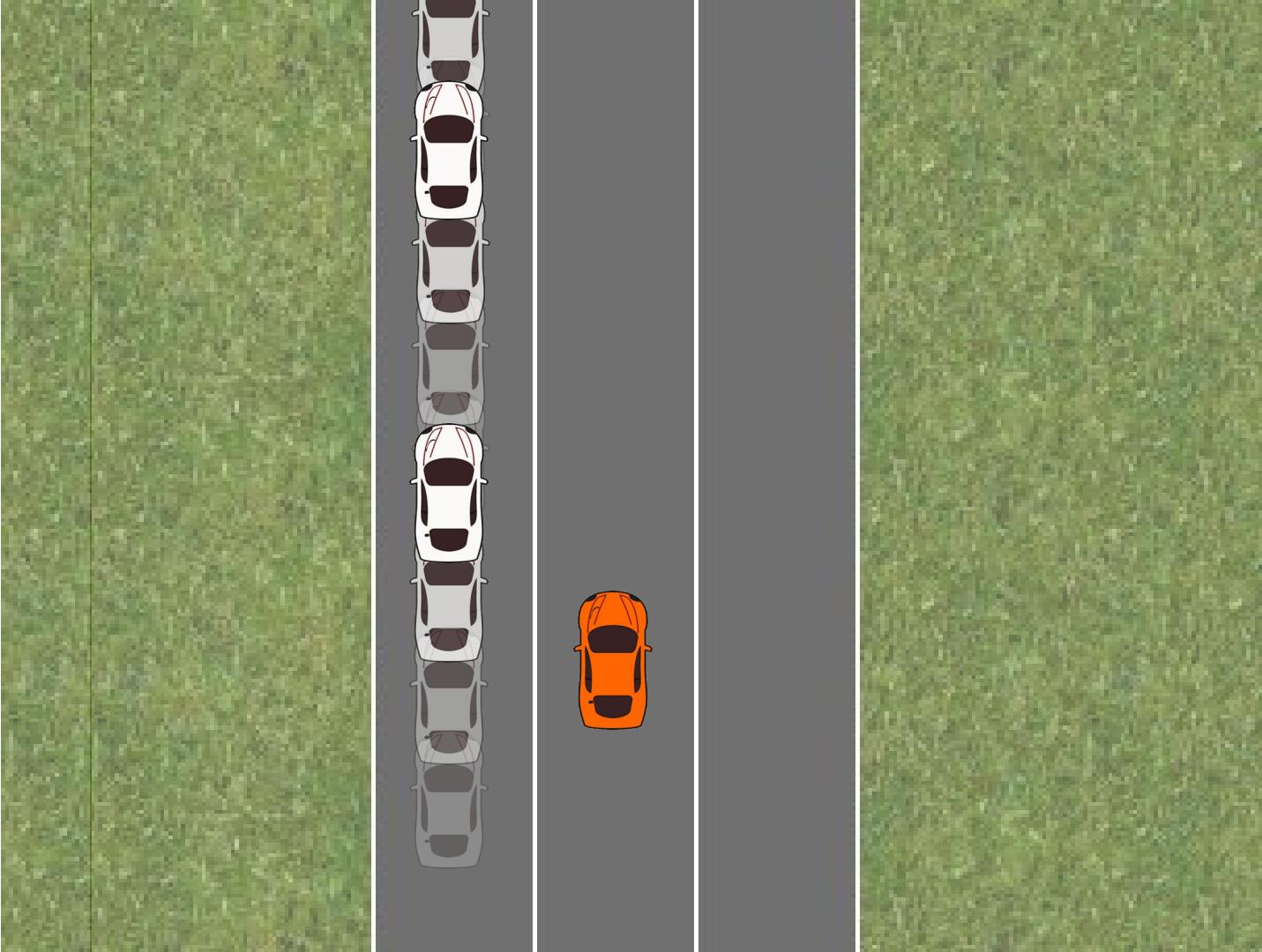


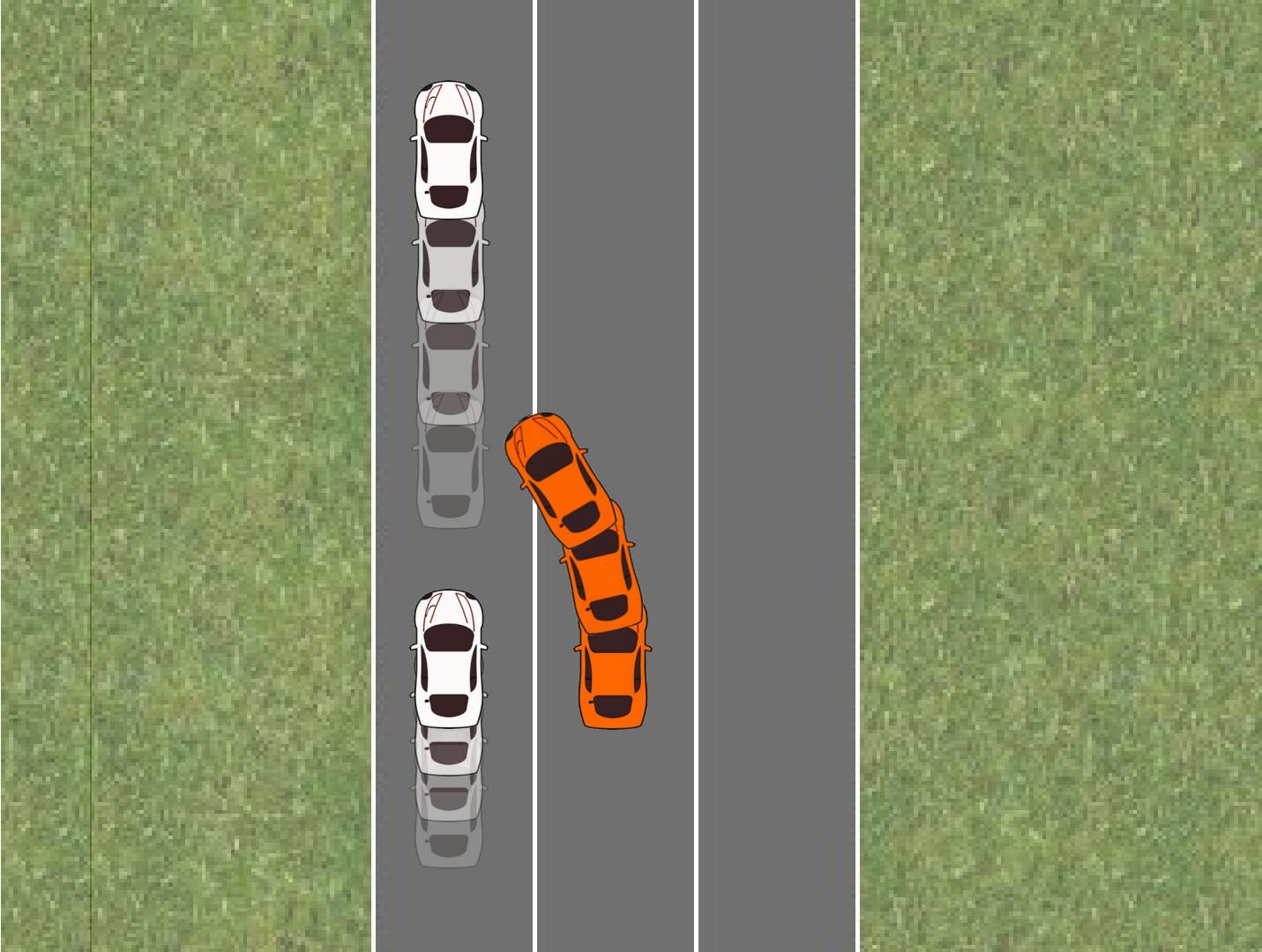
Personal Robotics Lab 415-758-2712

watch later









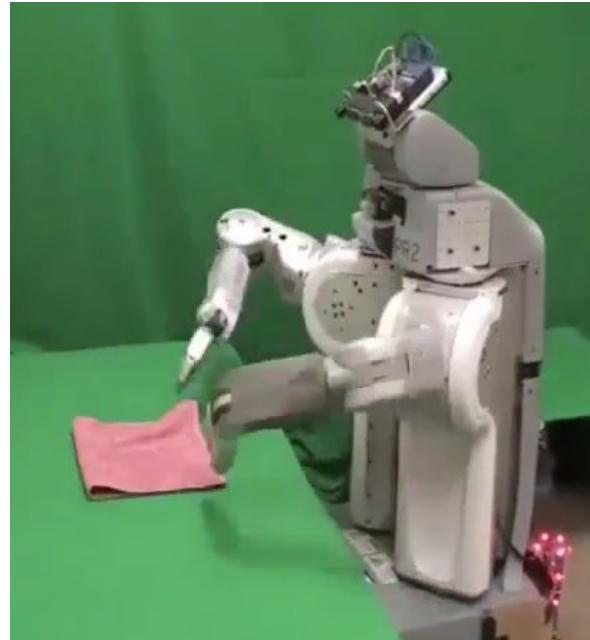
# Utility?

---

**Clear utility function**



**Not so clear utility function**

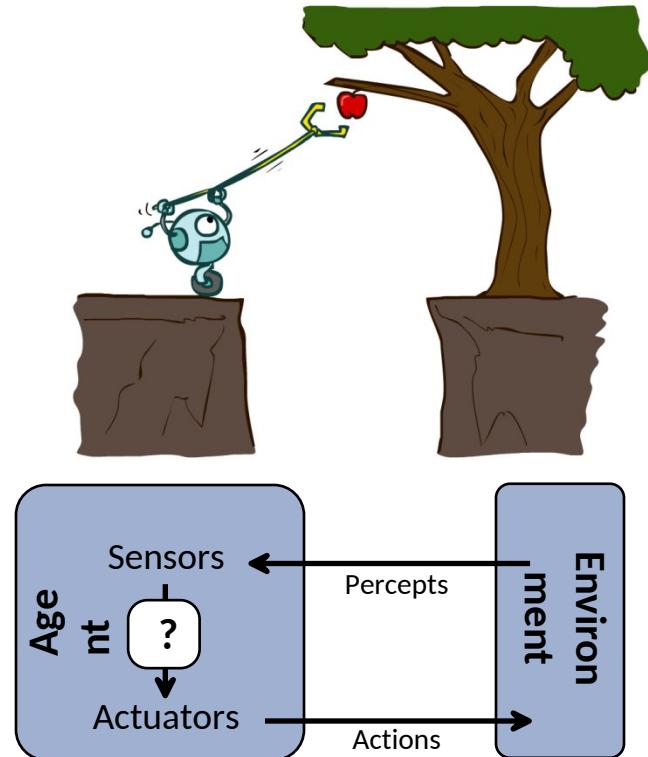




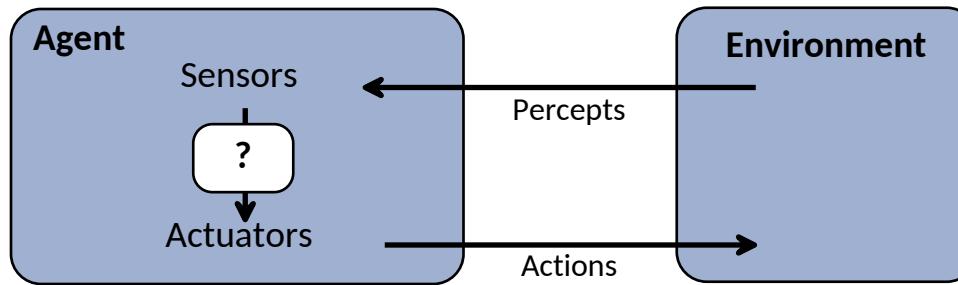
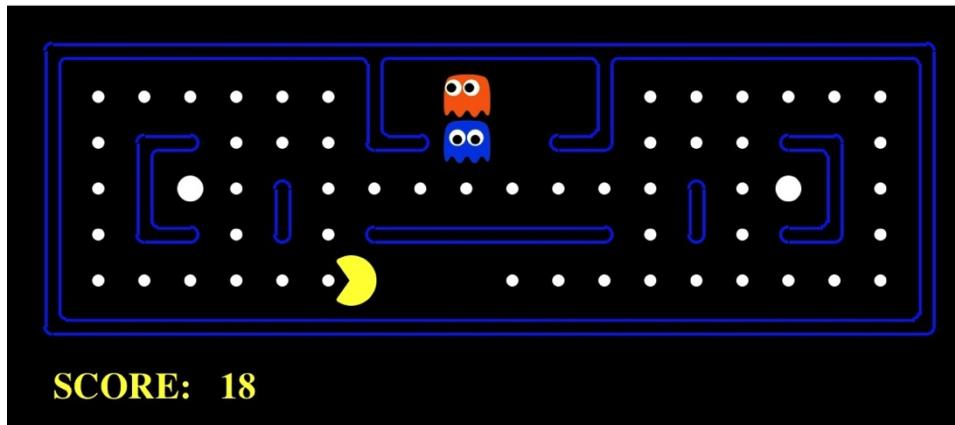
# Designing Rational Agents

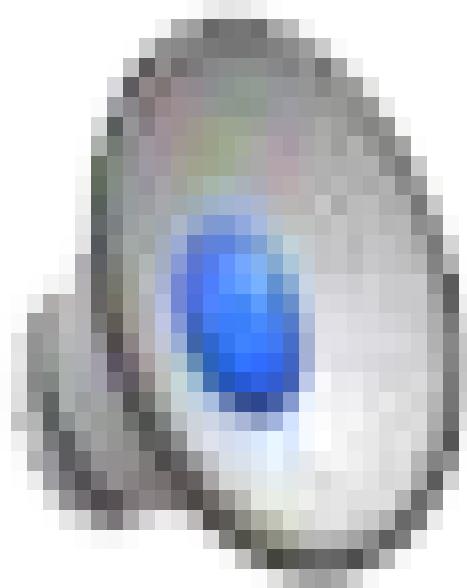
---

- o An **agent** is an entity that *perceives* and *acts*.
- o A **rational agent** selects actions that maximize its (expected) **utility**.
- o Characteristics of the **percepts**, **environment**, and **action space** dictate techniques for selecting rational actions
- o **This course is about:**
  - o General AI techniques for a variety of problem types
  - o Learning to recognize when and how a new problem can be solved with an existing technique

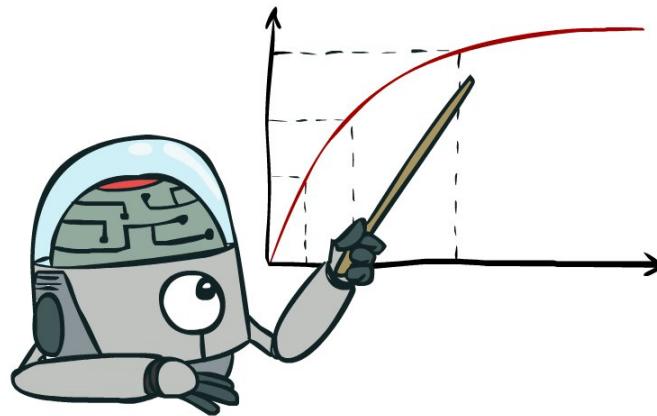


# Pac-Man as an Agent





# Maximize Your Expected Utility



# Data Science, Big Data y Data Analytics

# Introducción

Los datos están en todas partes. De hecho, la cantidad de datos digitales que existe está creciendo a un ritmo rápido, duplicándose cada dos años y cambiando la forma en que vivimos. Según IBM, se generaron 2.500 millones de gigabytes (GB) de datos cada día en 2012.



# Introducción

Un artículo de Forbes afirma que Data está creciendo más rápido que nunca antes y para el año 2020, se crearán aproximadamente 1,7 megabytes de nueva información por segundo para cada ser humano en el planeta. Lo que hace que sea extremadamente importante al menos conocer los conceptos básicos del campo. Después de todo, aquí es donde reside nuestro futuro.



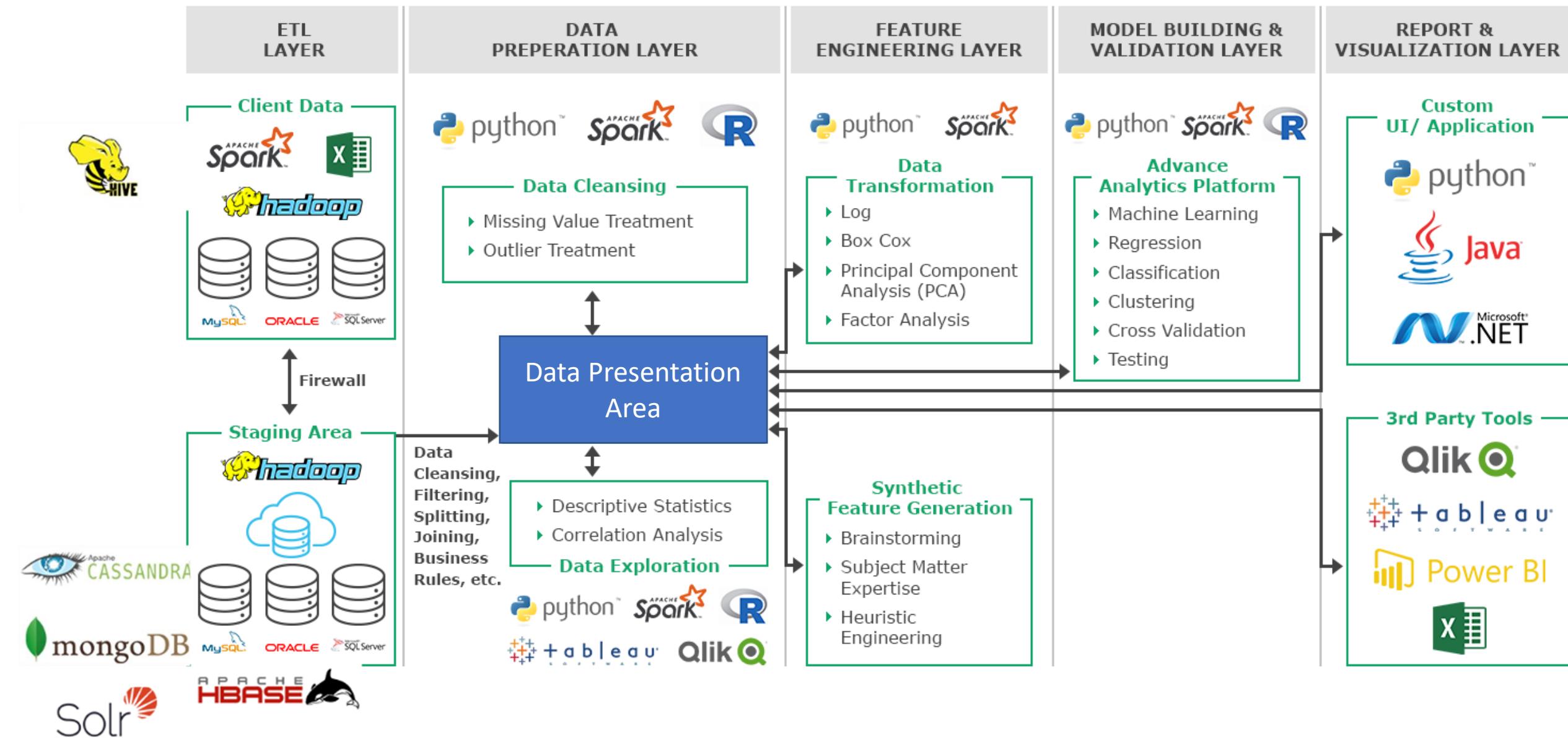
# Introducción

Es importante que diferenciaremos entre Data Science, Big Data y Data Analytics, en función de qué es, dónde se utiliza, las habilidades que necesita para convertirse en un profesional en el campo y las perspectivas salariales en cada campo.

# Data Science, Big Data, Data Analytics

- Data Science es la ciencia del estudio de datos.
- Big Data es un concepto teórico para definir los problemas que surgen del gran tamaño de los datos donde las herramientas tradicionales de manejo de datos no son lo suficientemente capaces.
- Data Analytics es un conjunto de herramientas y técnicas para realizar análisis de datos (grandes y pequeños).

Entonces, si tiene un problema de Big Data, usted, como Data Scientist, usará Data Analytics para resolver esos problemas.



# Data Science, Big Data, Data Analytics

Source Data



Store Data



Convert & ETL



Transform Data



Exploratory Analysis



Model Build & Generate Insights

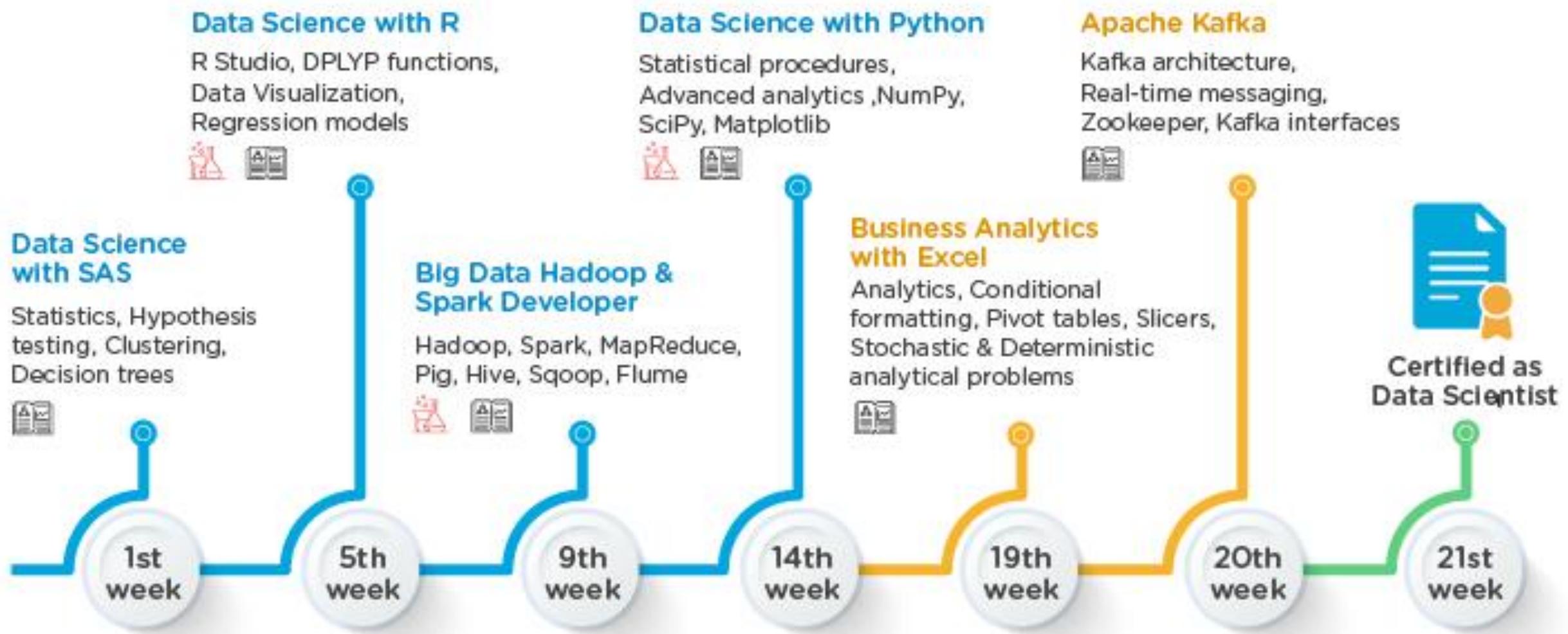


Visualisation



Model Execution in Production





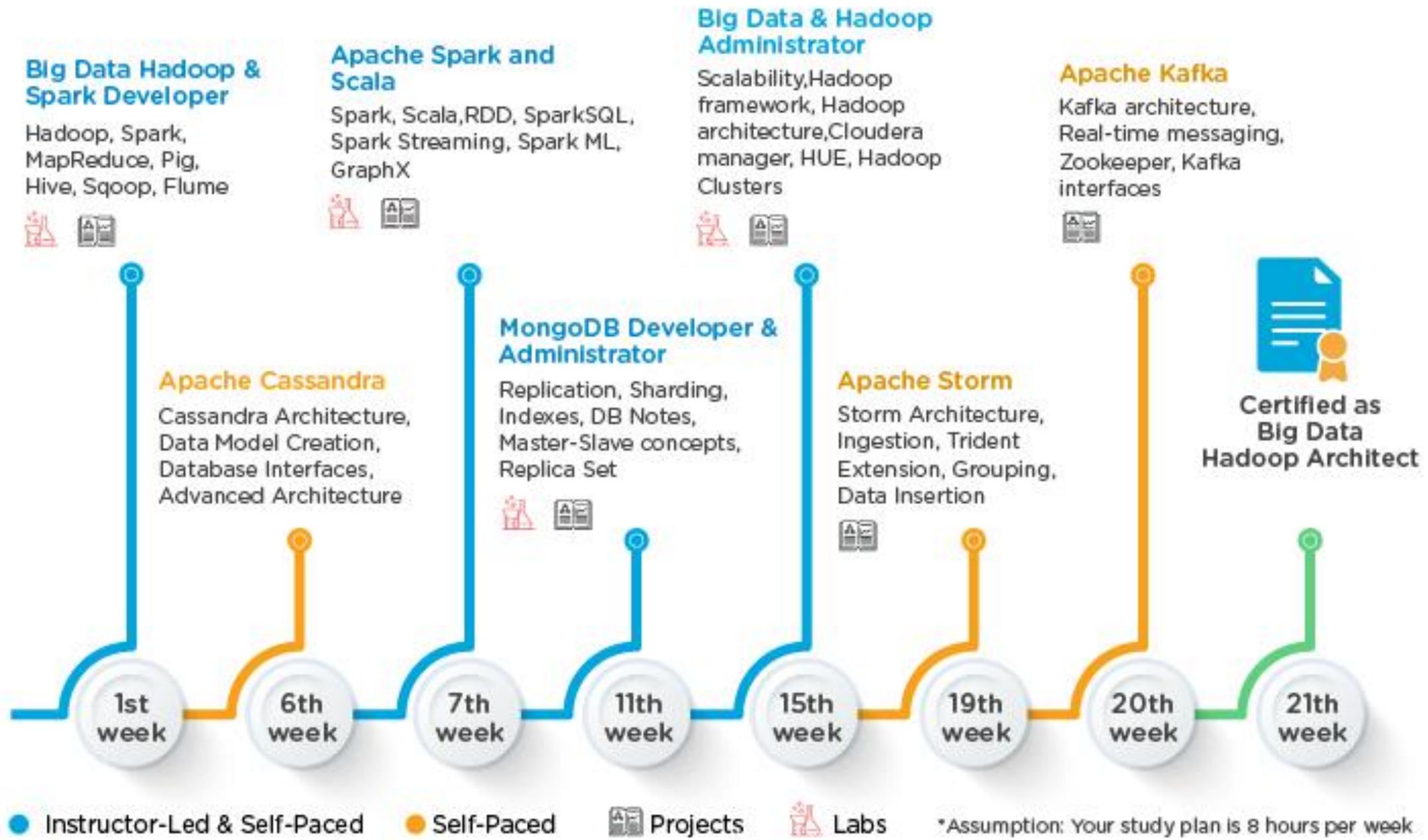
● Instructor-Led & Self-Paced

● Self-Paced

● Projects

● Labs

\*Assumption: Your study plan is 8 hours per week



### **Business analytics with excel**

Analytics, Conditional formatting, Pivot tables, Slicers, Stochastic & Deterministic analytical problems



### **Data Science with R**

R Studio, DPLYR functions, Data Visualization, Regression models



1st week

2nd week

6th week

10th week

15th week

19th week

● Instructor-Led & Self-Paced

● Instructor-Led

● Self-Paced

Projects

Labs

### **Tableau Desktop 10 Qualified Associate Training**

Data Blending, Data Extracts, Ad-hoc analytics, Heat map, Tree map, Waterfall, Pareto,etc



### **Data Science with Python**

Statistical procedures Advanced analytics, NumPy, SciPy, Matplotlib



### **Data Science with SAS training**

Statistics, Hypothesis testing, Clustering, Decision trees

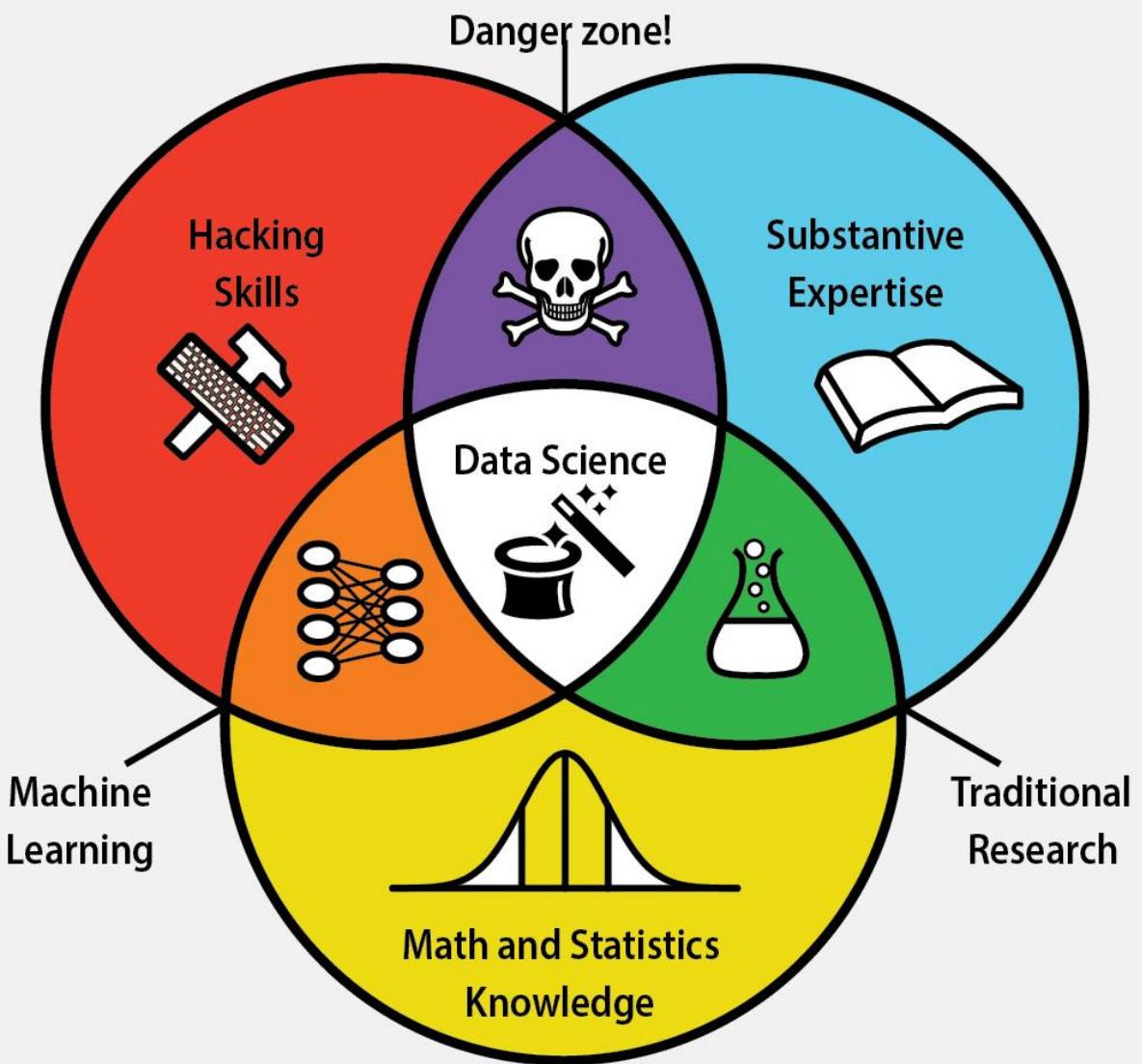


**Certified as Business Analytics Expert**

\*Assumption: Your study plan is 8 hours per week

# Python para Ciencia de Datos y Aprendizaje de Máquinas

# DATA SCIENCE SKILLSET



Data science, due to its interdisciplinary nature, requires an intersection of abilities: **hacking skills, math and statistics knowledge**, and **substantive expertise** in a field of science.



**Hacking skills** are necessary for working with massive amounts of electronic data that must be acquired, cleaned, and manipulated.



**Math and statistics knowledge** allows a data scientist to choose appropriate methods and tools in order to extract insight from data.



**Substantive expertise** in a scientific field is crucial for generating motivating questions and hypotheses and interpreting results.



**Traditional research** lies at the intersection of knowledge of math and statistics with substantive expertise in a scientific field.

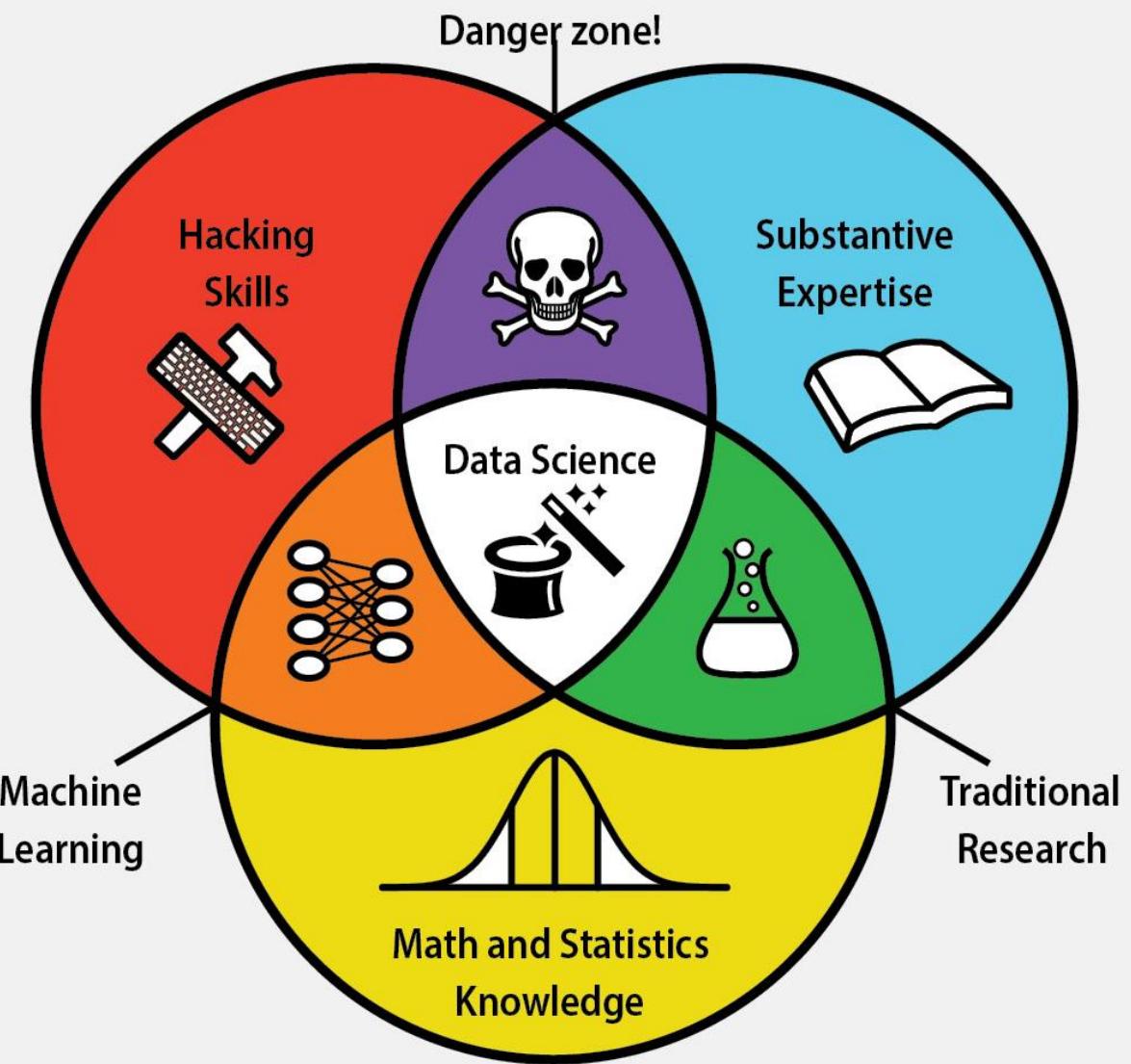


**Machine learning** stems from combining hacking skills with math and statistics knowledge, but does not require scientific motivation.



**Danger zone!** Hacking skills combined with substantive scientific expertise without rigorous methods can beget incorrect analyses.

## CONJUNTO DE HABILIDADES EN DATA SCIENCE



La ciencia de datos, debido a su naturaleza interdisciplinaria, requiere una intersección de habilidades: **habilidades de hacking, conocimiento de matemáticas y estadística, y experiencia sustantiva** en un campo de la ciencia.



**Hacking skills** are necessary for working with massive amounts of electronic data that must be acquired, cleaned, and manipulated.



**Math and statistics knowledge** allows a data scientist to choose appropriate methods and tools in order to extract insight from data.



**Substantive expertise** in a scientific field is crucial for generating motivating questions and hypotheses and interpreting results.



**Traditional research** lies at the intersection of knowledge of math and statistics with substantive expertise in a scientific field.

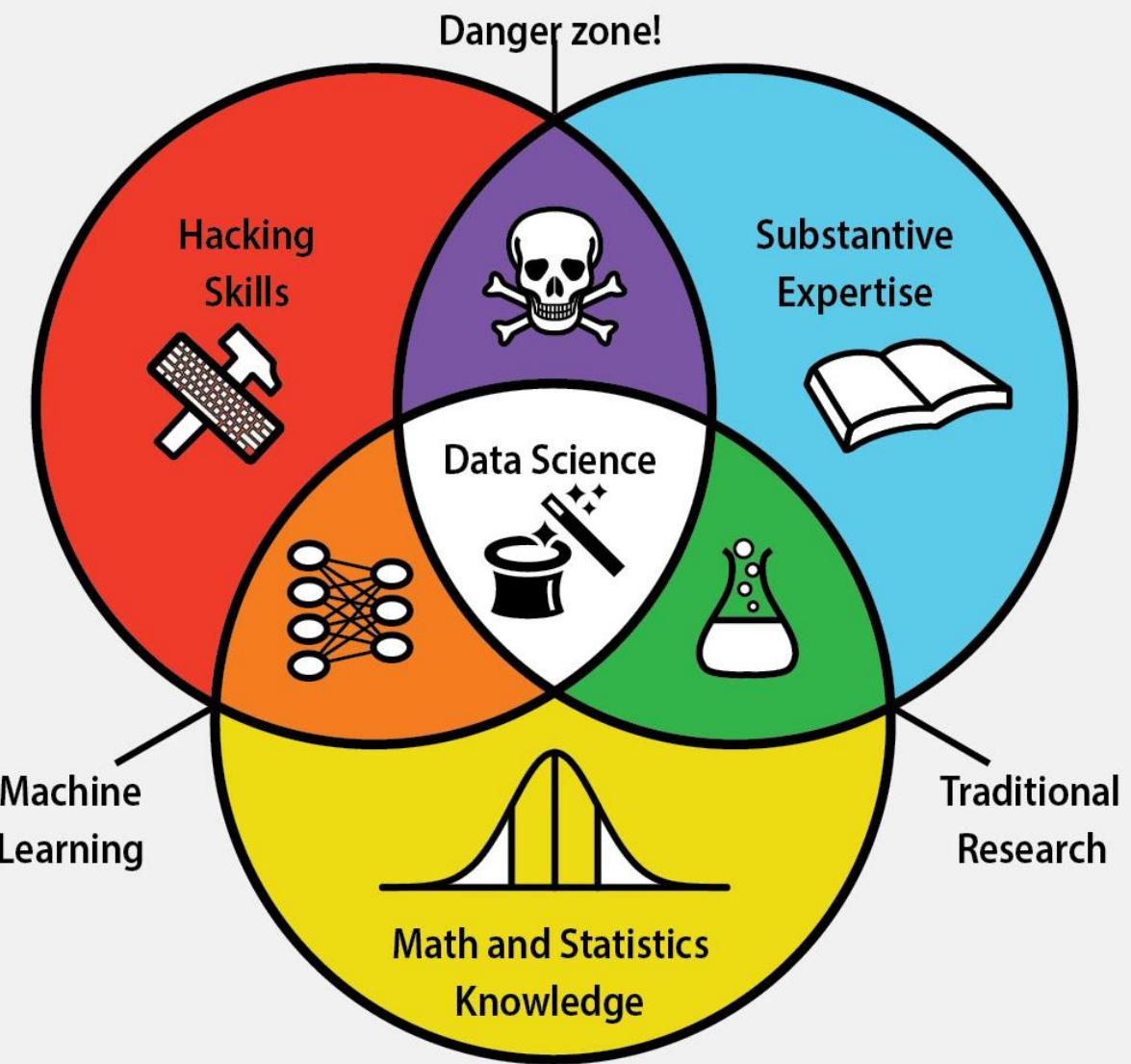


**Machine learning** stems from combining hacking skills with math and statistics knowledge, but does not require scientific motivation.



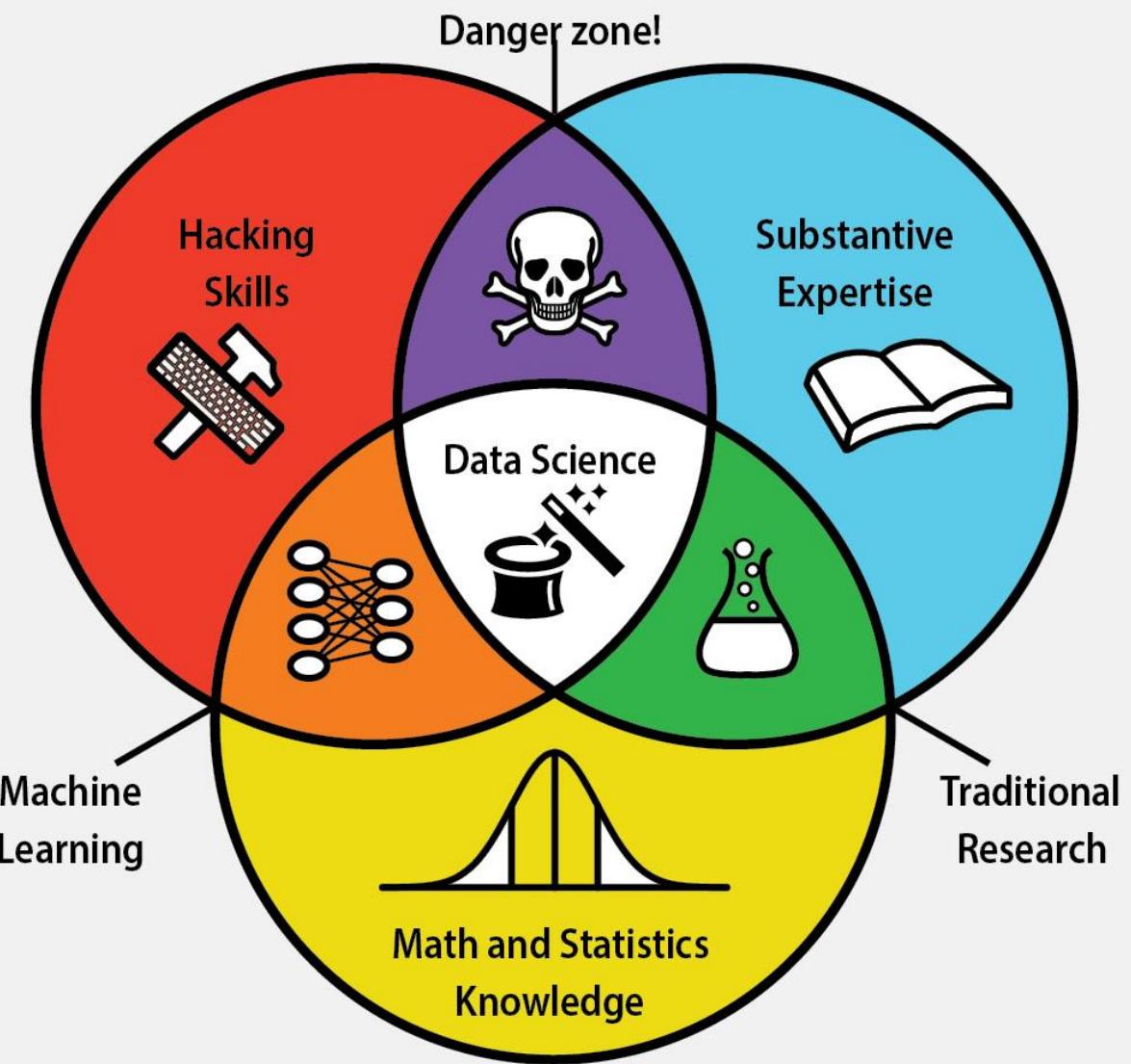
**Danger zone!** Hacking skills combined with substantive scientific expertise without rigorous methods can beget incorrect analyses.

## CONJUNTO DE HABILIDADES EN DATA SCIENCE



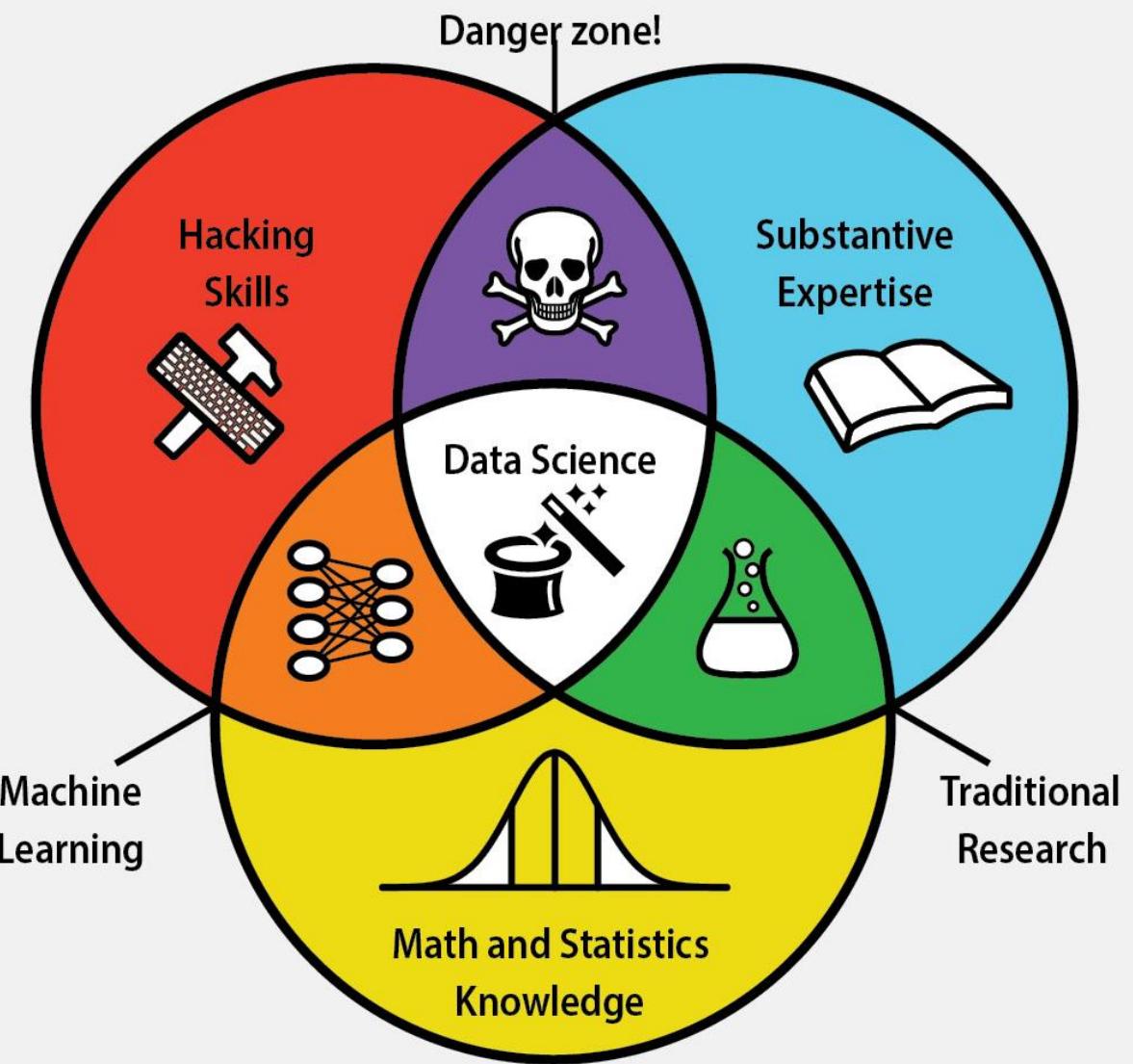
	<p>La ciencia de datos, debido a su naturaleza interdisciplinaria, requiere una intersección de habilidades: <b>habilidades de hacking, conocimiento de matemáticas y estadística, y experiencia sustantiva</b> en un campo de la ciencia.</p>
	<p>Las <b>habilidades de hacking</b> son necesarias para trabajar con grandes cantidades de datos electrónicos que se deben adquirir, limpiar y manipular.</p>
	<p><b>Math and statistics knowledge</b> allows a data scientist to choose appropriate methods and tools in order to extract insight from data.</p>
	<p><b>Substantive expertise</b> in a scientific field is crucial for generating motivating questions and hypotheses and interpreting results.</p>
	<p><b>Traditional research</b> lies at the intersection of knowledge of math and statistics with substantive expertise in a scientific field.</p>
	<p><b>Machine learning</b> stems from combining hacking skills with math and statistics knowledge, but does not require scientific motivation.</p>
	<p><b>Danger zone!</b> Hacking skills combined with substantive scientific expertise without rigorous methods can beget incorrect analyses.</p>

## CONJUNTO DE HABILIDADES EN DATA SCIENCE



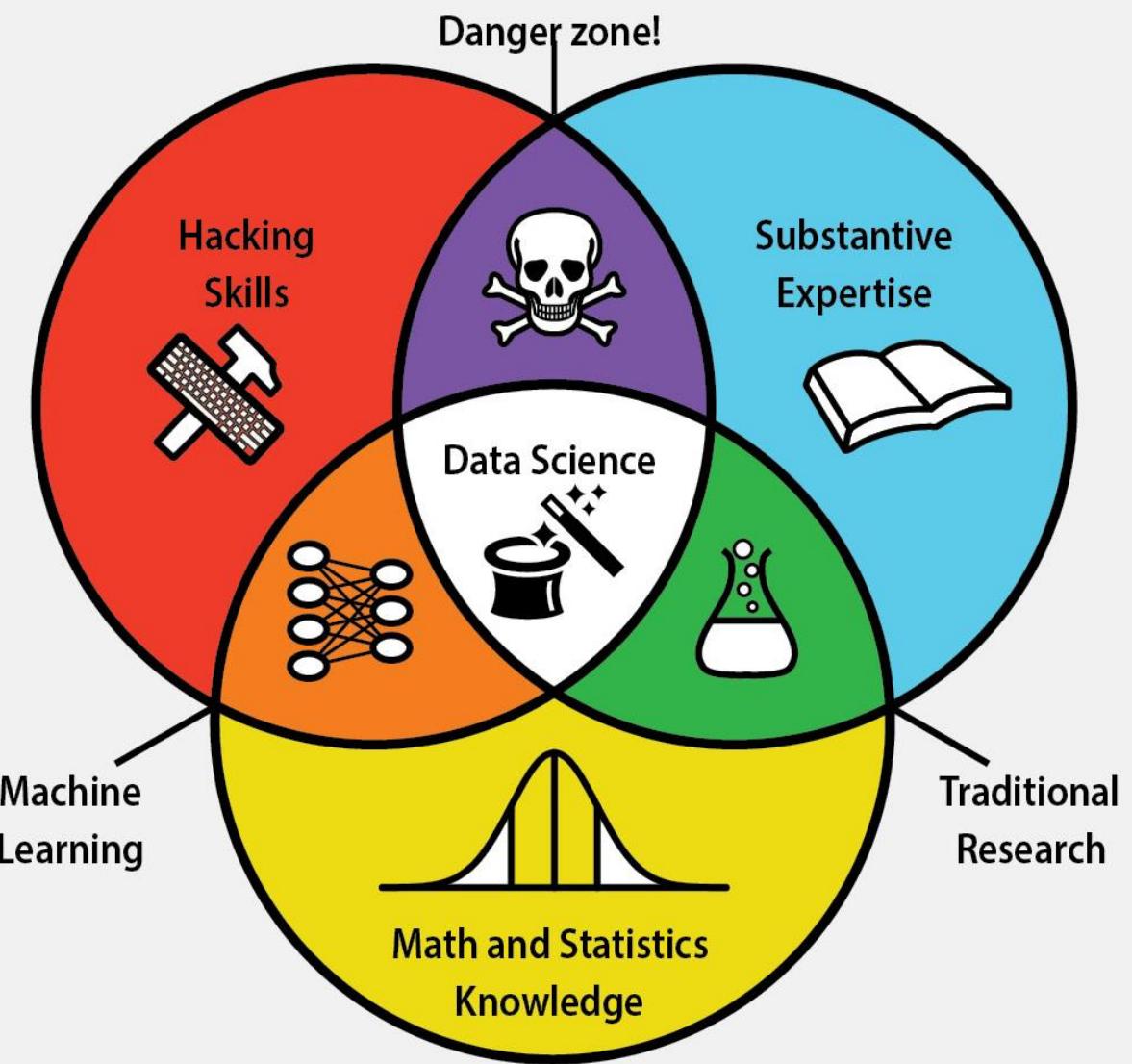
	<p>La ciencia de datos, debido a su naturaleza interdisciplinaria, requiere una intersección de habilidades: <b>habilidades de hacking, conocimiento de matemáticas y estadística, y experiencia sustantiva</b> en un campo de la ciencia.</p>
	<p>Las <b>habilidades de hacking</b> son necesarias para trabajar con grandes cantidades de datos electrónicos que se deben adquirir, limpiar y manipular.</p>
	<p>El <b>conocimiento matemático y estadístico</b> permite que un científico de datos elija los métodos y herramientas adecuados para extraer información de los datos.</p>
	<p><b>Substantive expertise</b> in a scientific field is crucial for generating motivating questions and hypotheses and interpreting results.</p>
	<p><b>Traditional research</b> lies at the intersection of knowledge of math and statistics with substantive expertise in a scientific field.</p>
	<p><b>Machine learning</b> stems from combining hacking skills with math and statistics knowledge, but does not require scientific motivation.</p>
	<p><b>Danger zone!</b> Hacking skills combined with substantive scientific expertise without rigorous methods can beget incorrect analyses.</p>

## CONJUNTO DE HABILIDADES EN DATA SCIENCE



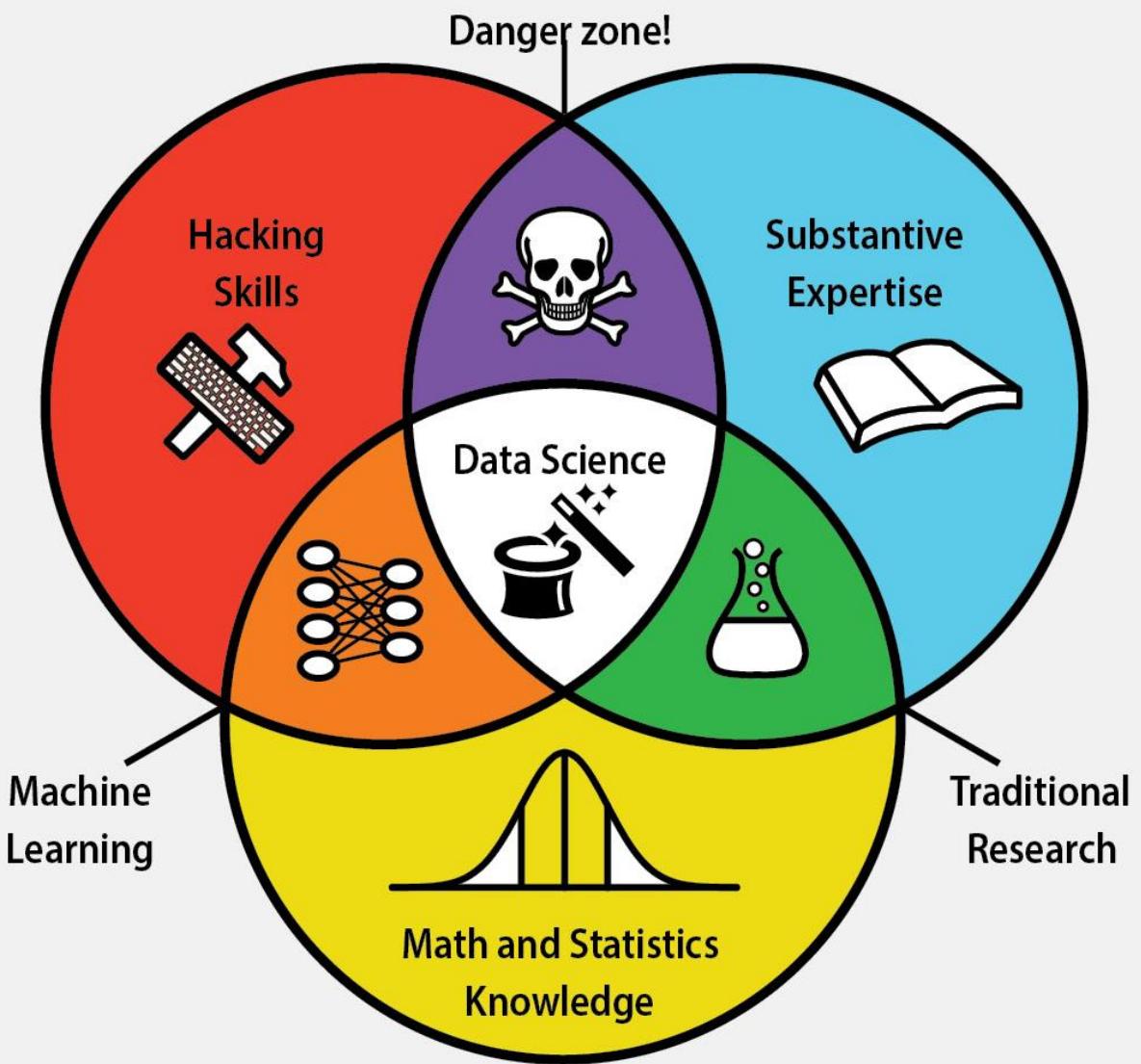
	<p>La ciencia de datos, debido a su naturaleza interdisciplinaria, requiere una intersección de habilidades: <b>habilidades de hacking, conocimiento de matemáticas y estadística, y experiencia sustantiva</b> en un campo de la ciencia.</p>
	<p>Las <b>habilidades de hacking</b> son necesarias para trabajar con grandes cantidades de datos electrónicos que se deben adquirir, limpiar y manipular.</p>
	<p>El <b>conocimiento matemático y estadístico</b> permite que un científico de datos elija los métodos y herramientas adecuados para extraer información de los datos.</p>
	<p>La <b>experiencia sustantiva</b> en un campo científico es crucial para generar preguntas motivadoras e hipótesis e interpretar los resultados.</p>
	<p><b>Traditional research</b> lies at the intersection of knowledge of math and statistics with substantive expertise in a scientific field.</p>
	<p><b>Machine learning</b> stems from combining hacking skills with math and statistics knowledge, but does not require scientific motivation.</p>
	<p><b>Danger zone!</b> Hacking skills combined with substantive scientific expertise without rigorous methods can beget incorrect analyses.</p>

## CONJUNTO DE HABILIDADES EN DATA SCIENCE



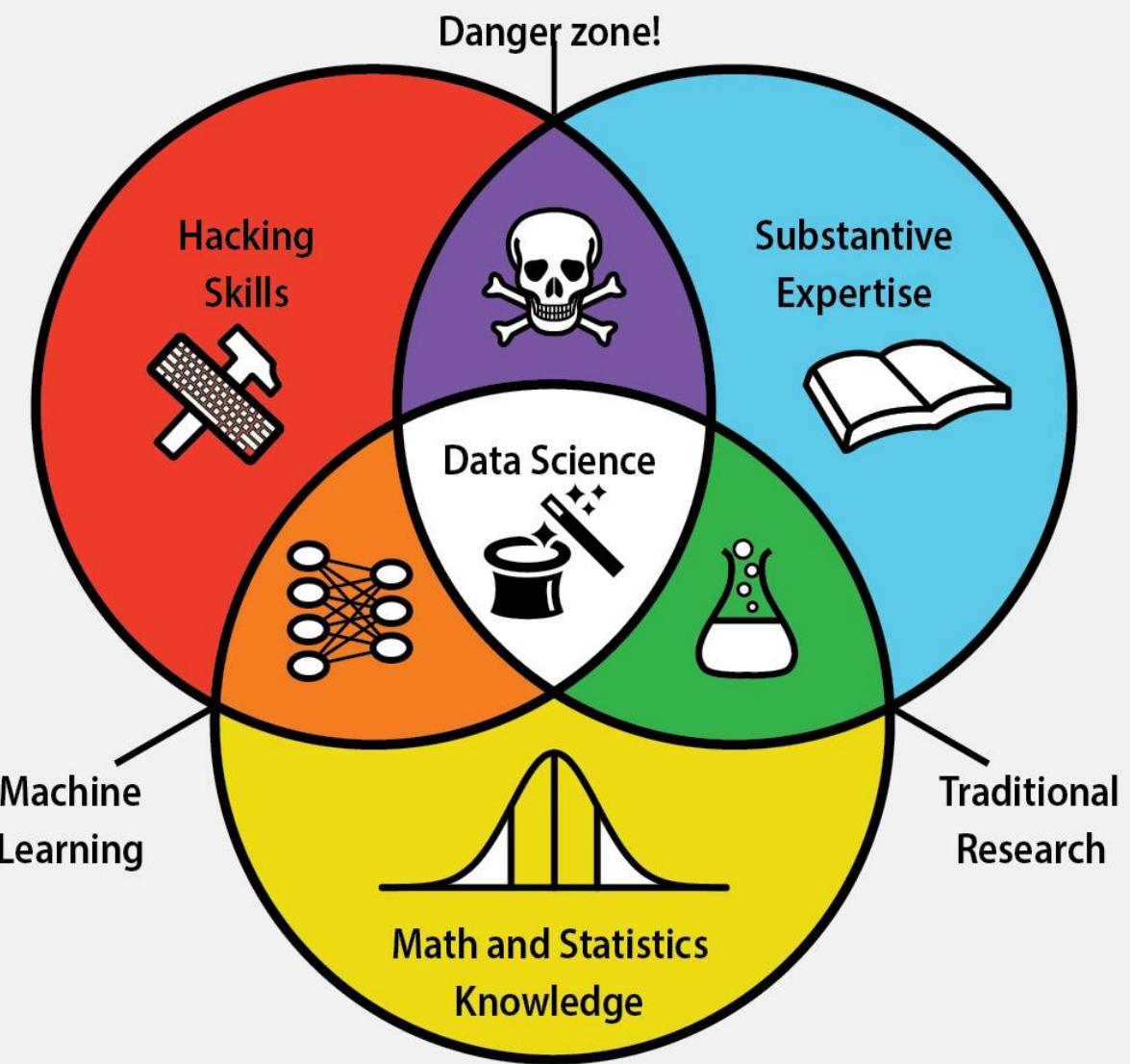
	<p>La ciencia de datos, debido a su naturaleza interdisciplinaria, requiere una intersección de habilidades: <b>habilidades de hacking, conocimiento de matemáticas y estadística, y experiencia sustantiva</b> en un campo de la ciencia.</p>
	<p>Las <b>habilidades de hacking</b> son necesarias para trabajar con grandes cantidades de datos electrónicos que se deben adquirir, limpiar y manipular.</p>
	<p>El <b>conocimiento matemático y estadístico</b> permite que un científico de datos elija los métodos y herramientas adecuados para extraer información de los datos.</p>
	<p>La <b>experiencia sustantiva</b> en un campo científico es crucial para generar preguntas motivadoras e hipótesis e interpretar los resultados.</p>
	<p>La <b>investigación tradicional</b> se encuentra en la intersección del conocimiento de matemáticas y estadística con experiencia sustantiva en un campo científico.</p>
	<p><b>Machine learning</b> stems from combining hacking skills with math and statistics knowledge, but does not require scientific motivation.</p>
	<p><b>Danger zone!</b> Hacking skills combined with substantive scientific expertise without rigorous methods can beget incorrect analyses.</p>

## CONJUNTO DE HABILIDADES EN DATA SCIENCE



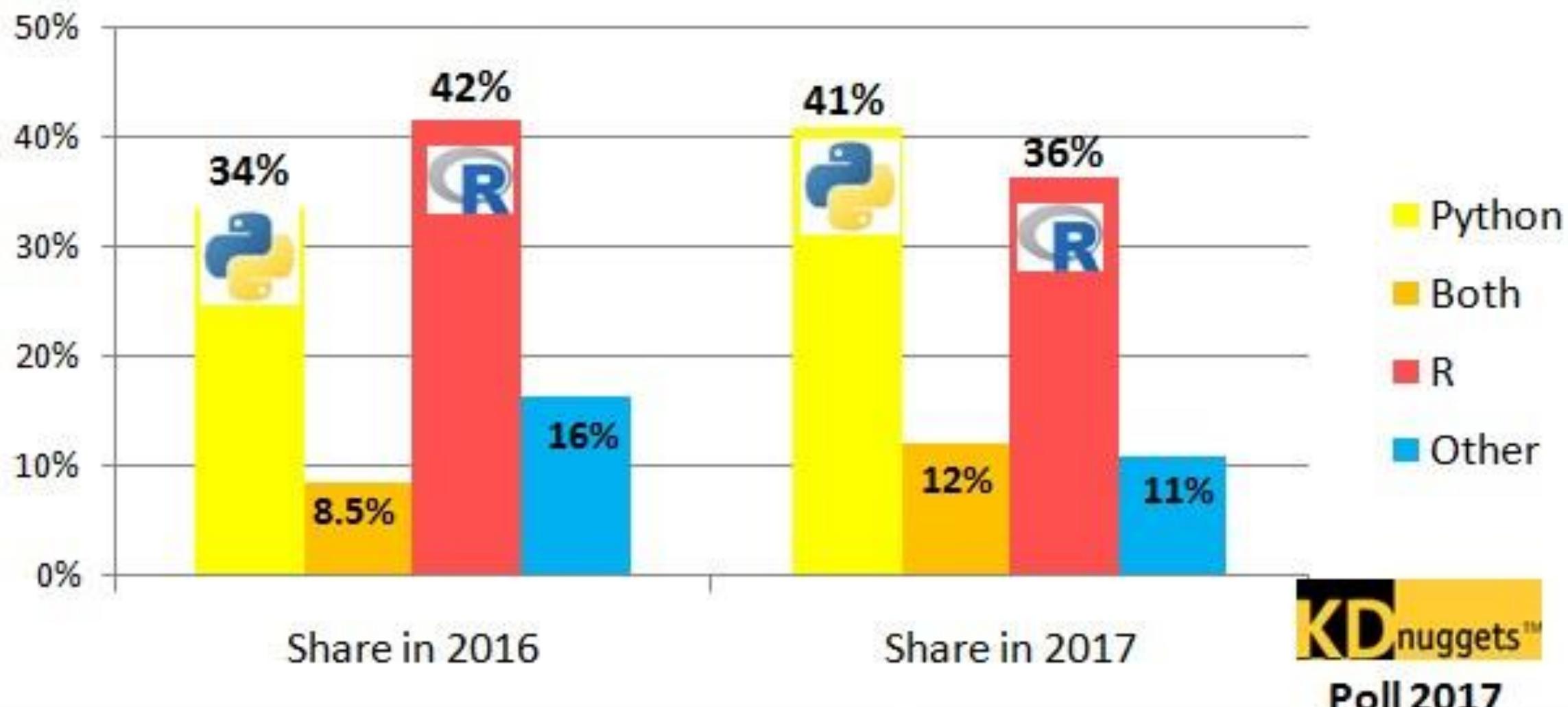
	<p>La ciencia de datos, debido a su naturaleza interdisciplinaria, requiere una intersección de habilidades: <b>habilidades de hacking, conocimiento de matemáticas y estadística, y experiencia sustantiva</b> en un campo de la ciencia.</p>
	<p>Las <b>habilidades de hacking</b> son necesarias para trabajar con grandes cantidades de datos electrónicos que se deben adquirir, limpiar y manipular.</p>
	<p>El <b>conocimiento matemático y estadístico</b> permite que un científico de datos elija los métodos y herramientas adecuados para extraer información de los datos.</p>
	<p>La <b>experiencia sustantiva</b> en un campo científico es crucial para generar preguntas motivadoras e hipótesis e interpretar los resultados.</p>
	<p>La <b>investigación tradicional</b> se encuentra en la intersección del conocimiento de matemáticas y estadística con experiencia sustantiva en un campo científico.</p>
	<p>El <b>aprendizaje automático</b> se deriva de la combinación de las habilidades de hacking con las matemáticas y el conocimiento estadístico, pero no requiere motivación científica.</p>
	<p><b>Danger zone!</b> Hacking skills combined with substantive scientific expertise without rigorous methods can beget incorrect analyses.</p>

## CONJUNTO DE HABILIDADES EN DATA SCIENCE



	<p>La ciencia de datos, debido a su naturaleza interdisciplinaria, requiere una intersección de habilidades: <b>habilidades de hacking, conocimiento de matemáticas y estadística, y experiencia sustantiva</b> en un campo de la ciencia.</p>
	<p>Las <b>habilidades de hacking</b> son necesarias para trabajar con grandes cantidades de datos electrónicos que se deben adquirir, limpiar y manipular.</p>
	<p>El <b>conocimiento matemático y estadístico</b> permite que un científico de datos elija los métodos y herramientas adecuados para extraer información de los datos.</p>
	<p>La <b>experiencia sustantiva</b> en un campo científico es crucial para generar preguntas motivadoras e hipótesis e interpretar los resultados.</p>
	<p>La <b>investigación tradicional</b> se encuentra en la intersección del conocimiento de matemáticas y estadística con experiencia sustantiva en un campo científico.</p>
	<p>El <b>aprendizaje automático</b> se deriva de la combinación de las habilidades de hacking con las matemáticas y el conocimiento estadístico, pero no requiere motivación científica.</p>
	<p><b>¡Zona peligrosa!</b> Las habilidades de hacking combinadas con la experiencia científica sustantiva sin métodos rigurosos pueden obtener un análisis incorrecto.</p>

# Python, R, Both, or Other platforms for Analytics, Data Science, Machine Learning



# What does a data scientist do?



Raw Data

Processing  
↓

Dataset

Statistical Models / Analysis

Machine Learning Predictions

Data driven Products

Reports Visualization Blogs



# Data Scientist

- Reconocido como uno de los mejores trabajos
- Grandes Salarios
- Solución de problemas interesantes

The screenshot shows the homepage of the Harvard Business Review website. At the top left is the HBR logo with the text "Harvard Business Review". To its right is a search bar with a red "SEARCH" button. Below the header is a navigation bar with links: "THE MAGAZINE", "BLOGS", "AUDIO & VIDEO", "BOOKS", "WEBINARS", and "COURSES". Under "THE MAGAZINE", the text "October 2012" is displayed. A main headline reads "Data Scientist: The Sexiest Job of the 21st Century".

The cover of the October 2012 issue of Harvard Business Review. The title "Harvard Business Review" is prominently displayed in large red letters at the top. Below it, the main feature is titled "GETTING CONTROL OF BIG DATA" in large black letters, with a small yellow flower icon on the letter "I". A cartoon illustration of a man in a top hat and red pants carrying a ladder on his shoulder is pulling a large, tilted ladder up a steep incline. The tagline "How vast new streams of information are changing the art of management" is written in a diagonal script across the bottom, with "PAGE 59" in red at the end. The right margin contains a column of articles with their titles and authors:

- 46 The Big Idea  
The True Measures Of Success  
Michael J. Mauboussin
- 84 International Business  
10 Rules for Managing Global Innovation  
Keeley Wilson and Yves L. Duz
- 93 Leadership  
What Ever Happened To Accountability?  
Thomas E. Ricks

# Librerías mas populares para ciencias de datos en Python

- NumPy
- SciPy
- Pandas
- Seaborn
- scikit-learn
- Matplotlib
- Plotly
- PySpark



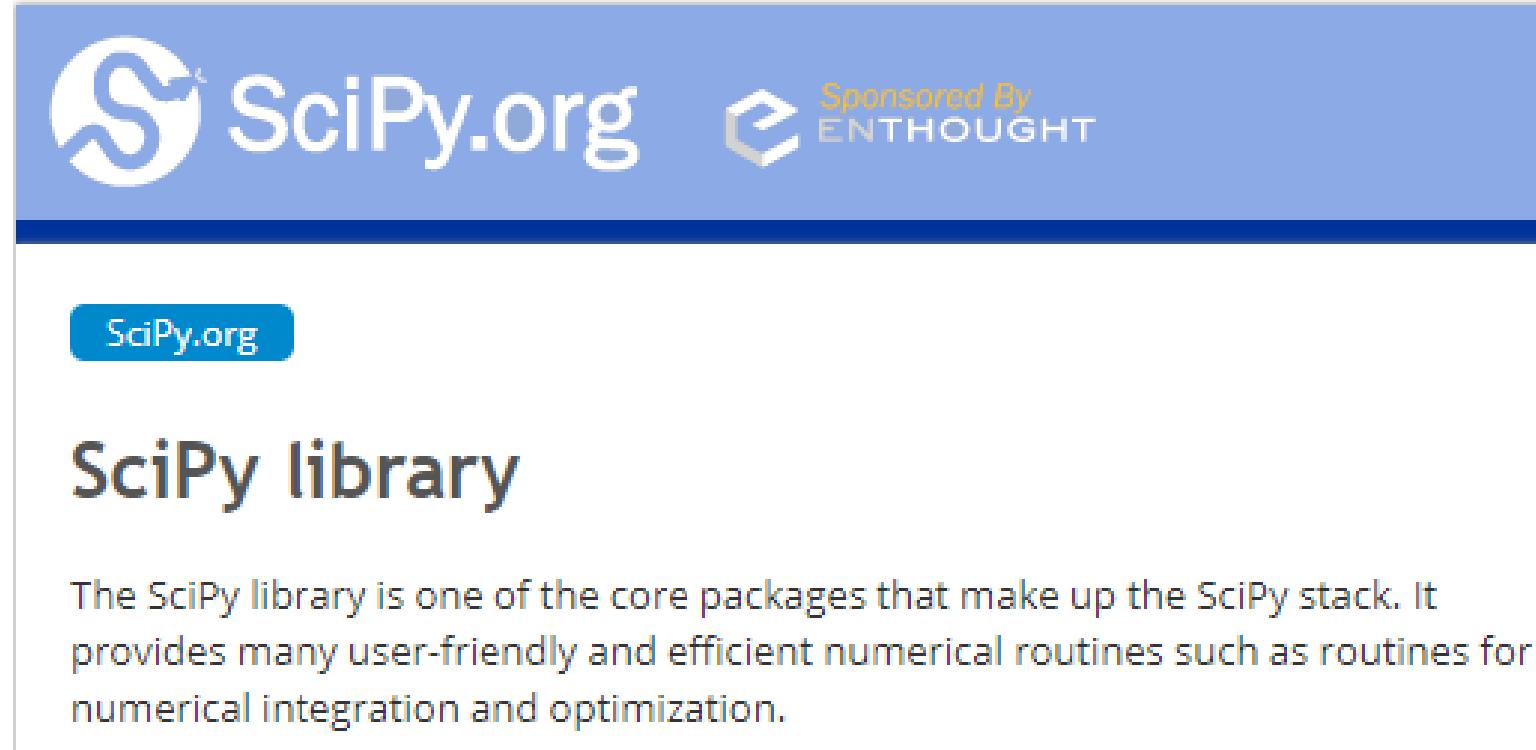
NumPy is the fundamental package for scientific computing with Python. It contains among other things:

- a powerful N-dimensional array object
- sophisticated (broadcasting) functions
- tools for integrating C/C++ and Fortran code
- useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined. This allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

# Librerías mas populares para ciencias de datos en Python

- NumPy
- SciPy
- Pandas
- Seaborn
- scikit-learn
- Matplotlib
- Plotly
- PySpark



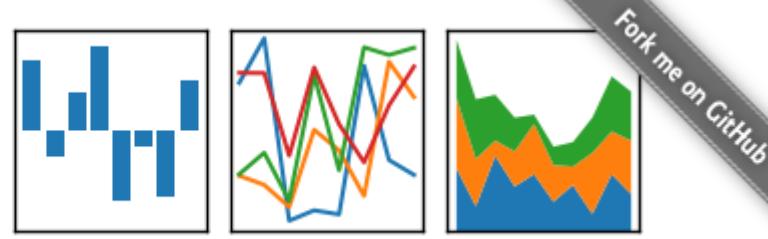
The image shows the SciPy.org homepage. At the top, there is a blue header bar with the SciPy logo (a stylized 'S') and the text "SciPy.org". To the right of the logo, it says "Sponsored By ENTHOUGHT" with a small logo of a white 'e'. Below the header is a dark blue navigation bar with the text "SciPy.org" in white. The main content area has a light blue background and features the text "SciPy library" in large, bold, brown letters. Below this, a paragraph describes the library: "The SciPy library is one of the core packages that make up the SciPy stack. It provides many user-friendly and efficient numerical routines such as routines for numerical integration and optimization."

# Librerías mas populares para ciencias de datos en Python

- NumPy
- SciPy
- Pandas
- Seaborn
- scikit-learn
- Matplotlib
- Plotly
- PySpark

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



Fork me on GitHub

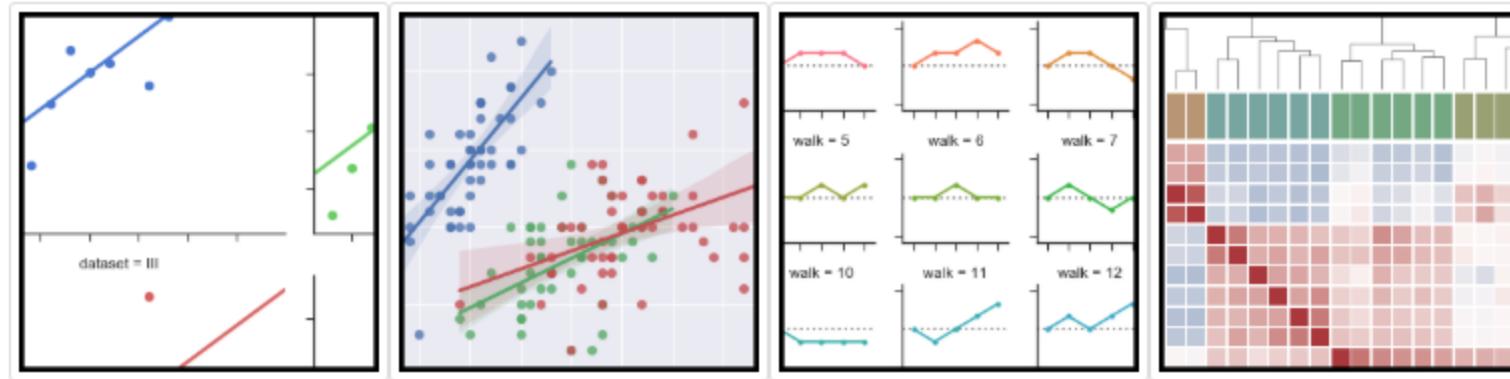
Python Data Analysis  
Library

*pandas* is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the [Python](#) programming language.

# Librerías mas populares para ciencias de datos en Python

- NumPy
- SciPy
- Pandas
- Seaborn
- scikit-learn
- Matplotlib
- Plotly
- PySpark

## seaborn: statistical data visualization



Seaborn is a Python visualization library based on matplotlib. It provides a high-level interface for drawing attractive statistical graphics.

# Librerías mas populares para ciencias de datos en Python

- NumPy
- SciPy
- Pandas
- Seaborn
- scikit-learn
- Matplotlib
- Plotly
- PySpark

The image consists of three parts. On the left is a map of South America with a red heatmap indicating the distribution of Bradypus Variegatus. On the right is a similar map with a red heatmap indicating the distribution of Microryzomys Minutus. Between them is a promotional graphic for the scikit-learn library. The graphic has a blue header with the text "scikit-learn" in large white letters and "Machine Learning in Python" in smaller white letters below it. To the right of the text is a yellow triangle containing the word "Hub". Below the text is a bulleted list of features: "Simple and efficient tools for data mining and data analysis", "Accessible to everybody, and reusable in various contexts", "Built on NumPy, SciPy, and matplotlib", and "Open source, commercially usable - BSD license". At the bottom of the graphic is a navigation bar with a left arrow, a series of blue dots with one highlighted in dark blue, and a right arrow.

# Librerías mas populares para ciencias de datos en Python

- NumPy
- SciPy
- Pandas
- Seaborn
- scikit-learn
- Matplotlib
- Plotly
- PySpark



## Classification

Identifying to which category an object belongs to.

**Applications:** Spam detection, Image recognition.

**Algorithms:** SVM, nearest neighbors, random forest, ...

— Examples

## Regression

Predicting a continuous-valued attribute associated with an object.

**Applications:** Drug response, Stock prices.

**Algorithms:** SVR, ridge regression, Lasso, ...

— Examples

## Clustering

Automatic grouping of similar objects into sets.

**Applications:** Customer segmentation, Grouping experiment outcomes

**Algorithms:** k-Means, spectral clustering, mean-shift, ...

— Examples

## Dimensionality reduction

Reducing the number of random variables to consider.

**Applications:** Visualization, Increased efficiency

**Algorithms:** PCA, feature selection, non-negative matrix factorization.

— Examples

## Model selection

Comparing, validating and choosing parameters and models.

**Goal:** Improved accuracy via parameter tuning

**Modules:** grid search, cross validation, metrics.

— Examples

## Preprocessing

Feature extraction and normalization.

**Application:** Transforming input data such as text for use with machine learning algorithms.

**Modules:** preprocessing, feature extraction.

— Examples

# Librerías mas populares para ciencias de datos en Python

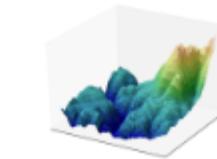
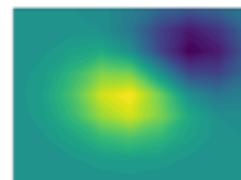
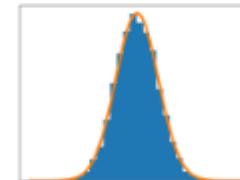
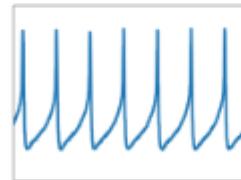
- NumPy
- SciPy
- Pandas
- Seaborn
- scikit-learn
- Matplotlib
- Plotly
- PySpark



Version 2.2.2

[home](#) | [examples](#) | [tutorials](#) | [pyplot](#) | [docs](#) »

Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shells, the Jupyter notebook, web application servers, and four graphical user interface toolkits.



Matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, errorcharts, scatterplots, etc., with just a few lines of code. For examples, see the [sample](#)

# Librerías mas populares para ciencias de datos en Python

- NumPy
- SciPy
- Pandas
- Seaborn
- scikit-learn
- Matplotlib
- Plotly
- PySpark

The screenshot shows the homepage of the Plotly website. At the top, there is a navigation bar with the Plotly logo, consulting, pricing, products, master classes, and a login button. The main headline reads "Modern Visualization for the Data Era". Below the headline, there is a description of what Plotly does and a section about collaboration servers. To the right of the text, there is an illustration of a laptop, a smartphone, and a tablet displaying various charts and graphs.

CONSULTING PRICING PRODUCTS MASTER CLASSES LOG IN

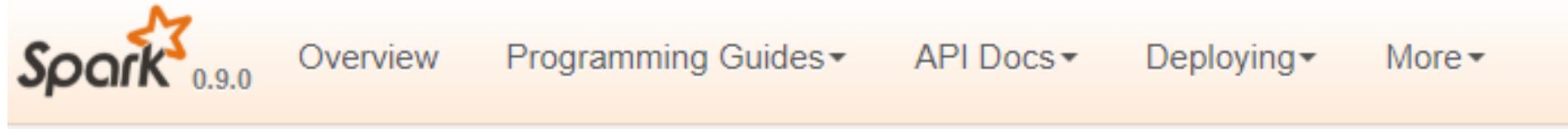
# Modern Visualization for the Data Era

Plotly creates **leading open source tools** for composing, editing, and sharing interactive **data visualization** via the Web.

Our collaboration servers (available in cloud or on premises) allow **data scientists** to showcase their work, make graphs without coding, and collaborate with **business analysts, designers, executives, and clients**.

# Librerías mas populares para ciencias de datos en Python

- NumPy
- SciPy
- Pandas
- Seaborn
- scikit-learn
- Matplotlib
- Plotly
- PySpark



## Python Programming Guide

The Spark Python API (PySpark) exposes the Spark programming model to Python. To learn the basics of Spark, we recommend reading through the [Scala programming guide](#) first; it should be easy to follow even if you don't know Scala. This guide will show how to use the Spark features described there in Python.

# Configuración de Entorno

- En este taller usaremos Notebooks de Jupyter.
- Sin embargo usted es libre de usar el entorno de desarrollo que prefiera.
- Todas las notas pueden ser descargadas como archivos .py que son compatibles con cualquier IDE de Python o editor de texto.
- Usaremos la última versión de Python 3 a través de la distribución de Anaconda



notebook

5.4.0

Web-based, interactive computing  
notebook environment. Edit and run  
human-readable docs while describing the  
data analysis.



spyder

3.2.8

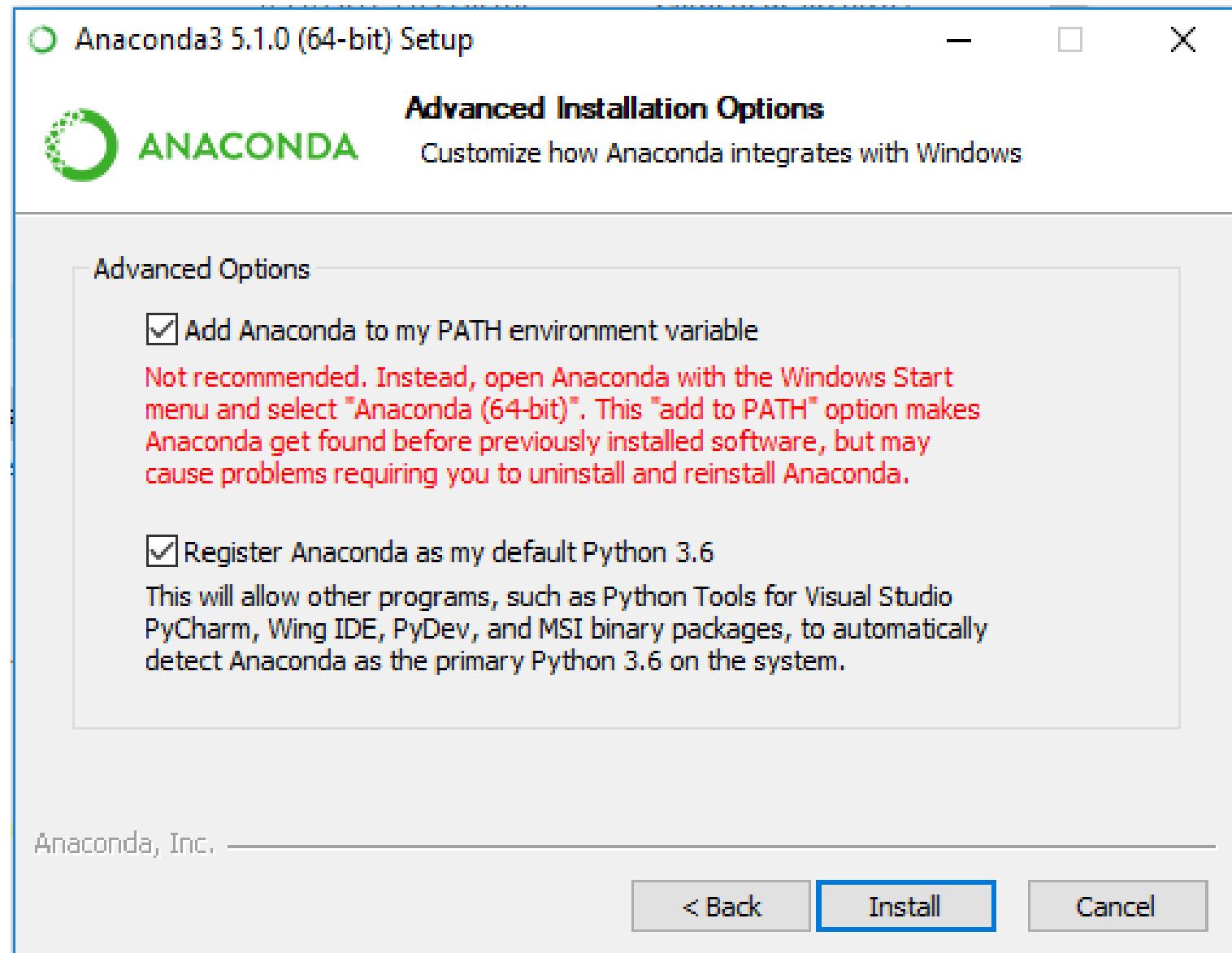
Scientific PYthon Development  
EnviRonment. Powerful Python IDE with  
advanced editing, interactive testing,  
debugging and introspection features

# Instalación de Anaconda Navigator

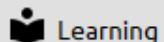
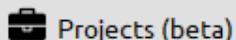
Desinstalar cualquier versión previa de Python, antes de instalar Anaconda.



Es muy importante considerar esta opción en la instalación para poder seguir los mismos pasos en los ejemplos



# ANACONDA NAVIGATOR

[Sign in to Anaconda Cloud](#)[Home](#)[Documentation](#)[Developer Blog](#)[Feedback](#)

Applications on

base (root)

Channels

Refresh



**jupyterlab**  
0.31.4

An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.

[Launch](#)



**notebook**  
5.4.0

Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.

[Launch](#)



**qtconsole**  
4.3.1

PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more.

[Launch](#)



**spyder**  
3.2.6

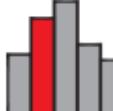
Scientific PYthon Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features

[Launch](#)



**vscode**  
1.21.1

Streamlined code editor with support for development operations like debugging, task running and version control.



**glueviz**  
0.12.0

Multidimensional data visualization across files. Explore relationships within and among related datasets.



**orange3**  
3.4.1

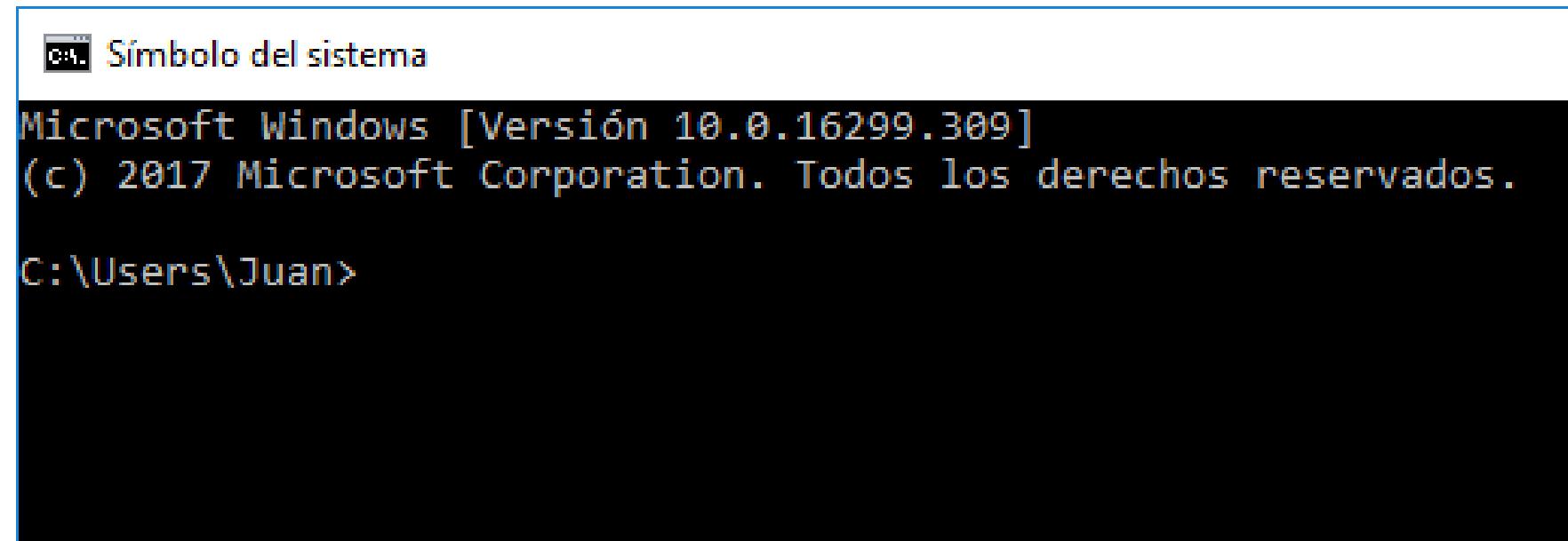
Component based data mining framework. Data visualization and data analysis for novice and expert. Interactive workflows



**rstudio**  
1.1.383

A set of integrated tools designed to help you be more productive with R. Includes R essentials and notebooks.

Comprobar la instalación adecuada con la ventana de Símbolo del Sistema



```
Símbolo del sistema  
Microsoft Windows [Versión 10.0.16299.309]  
(c) 2017 Microsoft Corporation. Todos los derechos reservados.  
C:\Users\Juan>
```

Si tiene creado  
en la unidad C  
las siguientes  
carpetas:

Cambiar a la  
carpeta  
correspondiente

Este equipo > OS (C:) > CursoML

```
Símbolo del sistema - jupyter notebook
Microsoft Windows [Versión 10.0.17134.165]
(c) 2018 Microsoft Corporation. Todos los derechos reservados.

C:\Users\Juan>cd..

C:\Users>cd..

C:\>cd
C:\

C:\>cd C:\CursoML

C:\CursoML>jupyter notebook
[I 23:13:18.960 NotebookApp] JupyterLab beta preview extension loaded
[I 23:13:18.961 NotebookApp] JupyterLab application directory is C:\CursoML
[W 23:13:19.074 NotebookApp] Error loading server extension jupyterlab
  Traceback (most recent call last):
```



python<sup>TM</sup>



# Obtenemos:

A screenshot of a web browser displaying the Jupyter Notebook interface at `localhost:8888/tree`. The browser title bar shows "jupyter". The main content area displays a file tree with the following structure:

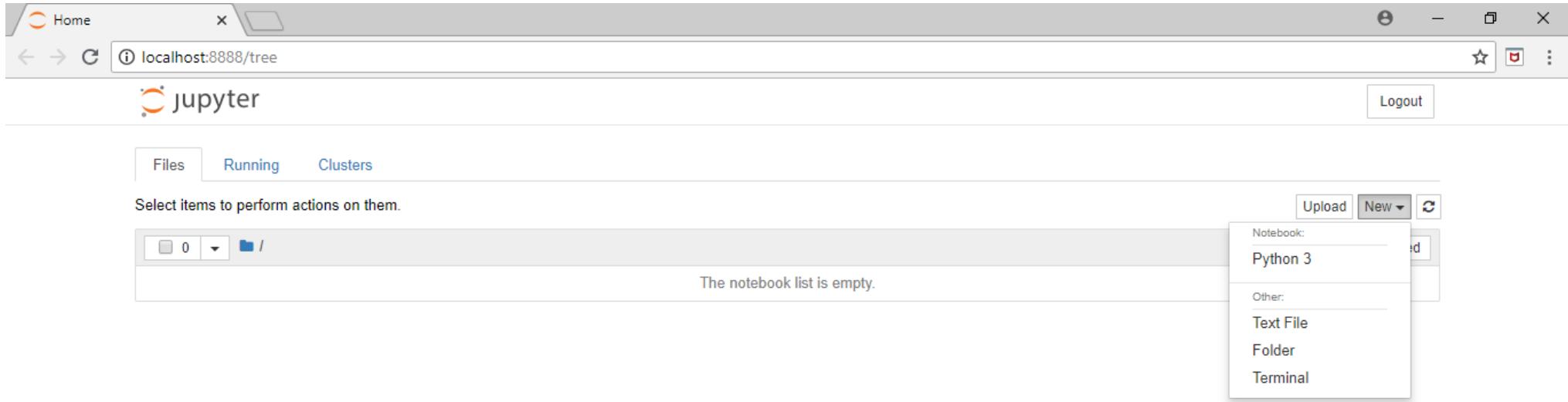
- 0 files
- 1 folder named "/"

The message "The notebook list is empty." is centered below the tree.

Navigation and action buttons are visible at the top right, including "Logout", "Upload", "New", and a refresh icon.

# Para crear un block de notas

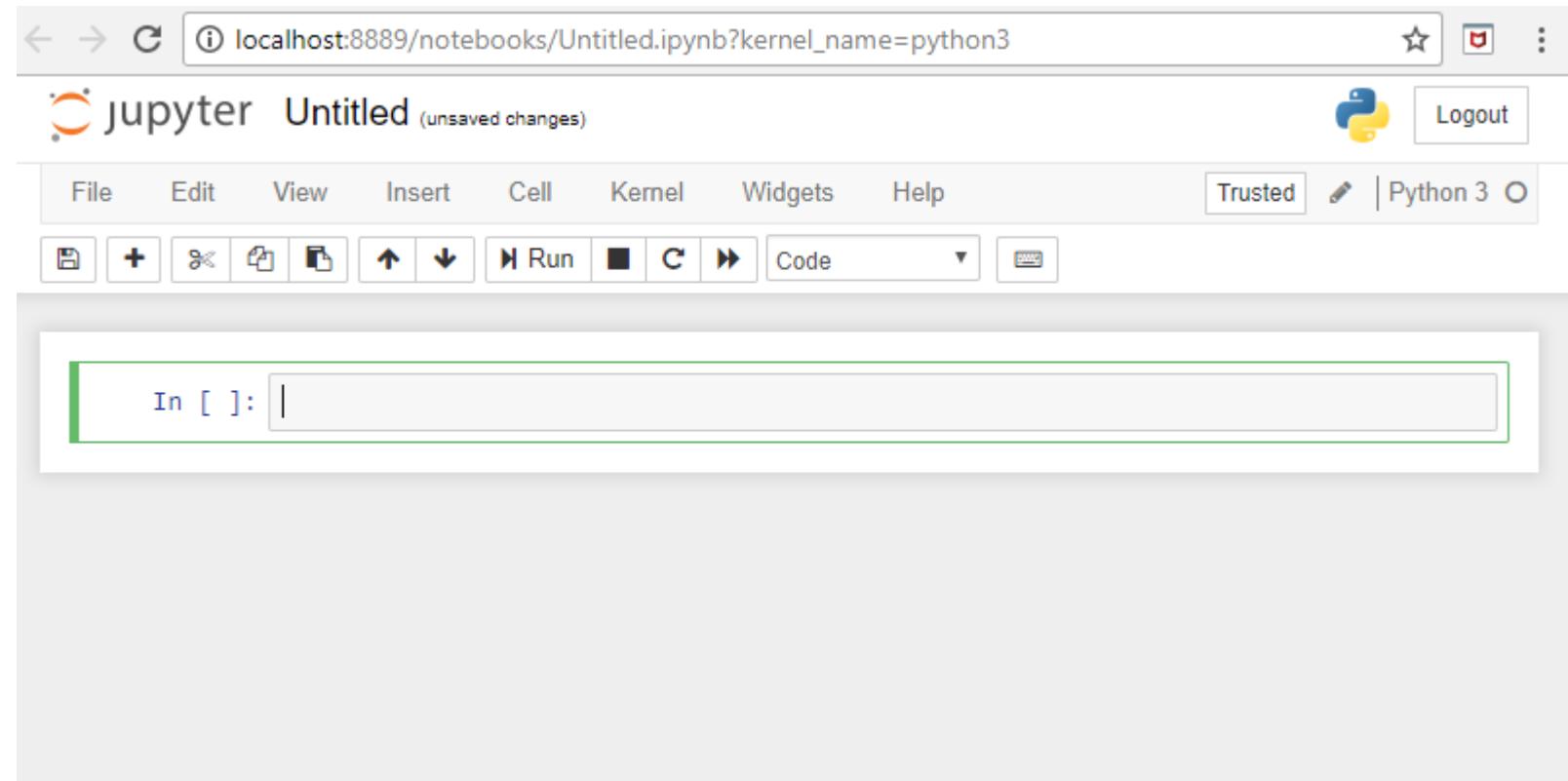
Se hace clic en New y se elige Python 3



# El block de notas

En el block de notas tenemos distintos tipos de celdas como:

- Code
- Markdown
- Raw NBConvert
- Heading



# Celda Markdown

File Edit View Insert Cell Kernel Widgets Help

Markdown

```
# Este es un título
## Este es un subtítulo
### Otro de menor nivel
Este es un párrafo
Esto es un texto en cursiva*
**Esto es un texto en negrita **
```

In [ ]:

File Edit View Insert Cell Kernel Widgets Help

Code

```
Este es un título
Este es un subtítulo
Otro de menor nivel
Este es un párrafo
Esto es un texto en cursiva
Esto es un texto en negrita
```

In [ ]:

Mayús + Enter ->  
Para observar los  
resultados

# Celda Code

En una celda  
Code se puede  
ejecutar y  
probar código  
Python

The screenshot shows a Jupyter Notebook interface with the following elements:

- Toolbar:** File, Edit, View, Insert, Cell, Kernel, Widgets, Help.
- Icon Bar:** Includes icons for saving, running, and kernel selection.
- Cell Types:** A list of cell types including Title, Subtitle, Text, Cursive, and Bold.
- Text Content:** A sample text block containing "Este es un título", "Este es un subtítulo", "Otro de menor nivel", "Este es un párrafo", "Esto es un texto en cursiva", and "Esto es un texto en negrita".
- In [1]:** A code cell containing the Python command `print("FISI UNMSM")`.
- Output:** The output of the code cell is "FISI UNMSM".
- Annotations:**
  - A blue arrow points to the code in In [1] with the text "Para ejecutar: Ctrl + Entrar".
  - A yellow box surrounds the code in In [1] with the text "Para ejecutar e insertar una nueva celda: Shift + Entrar".

# Python en pocos pasos

# Temas a tratar

- Tipos de Datos
  - Números
  - Cadenas
  - Impresión Formateada
  - Listas
  - Diccionarios
  - Booleanos
  - Tuplas y Conjuntos
- Operadores de Comparación
- Sentencias If, elif y else
- Bucles For
- Bucles While
- range()
- Operadores de Comparación
- Sentencias If, elif y else
- Listas por comprensión
- Funciones
- Expresiones Lambda
- Map y Filter

# Introducción a Machine Learning

# Libro complementario

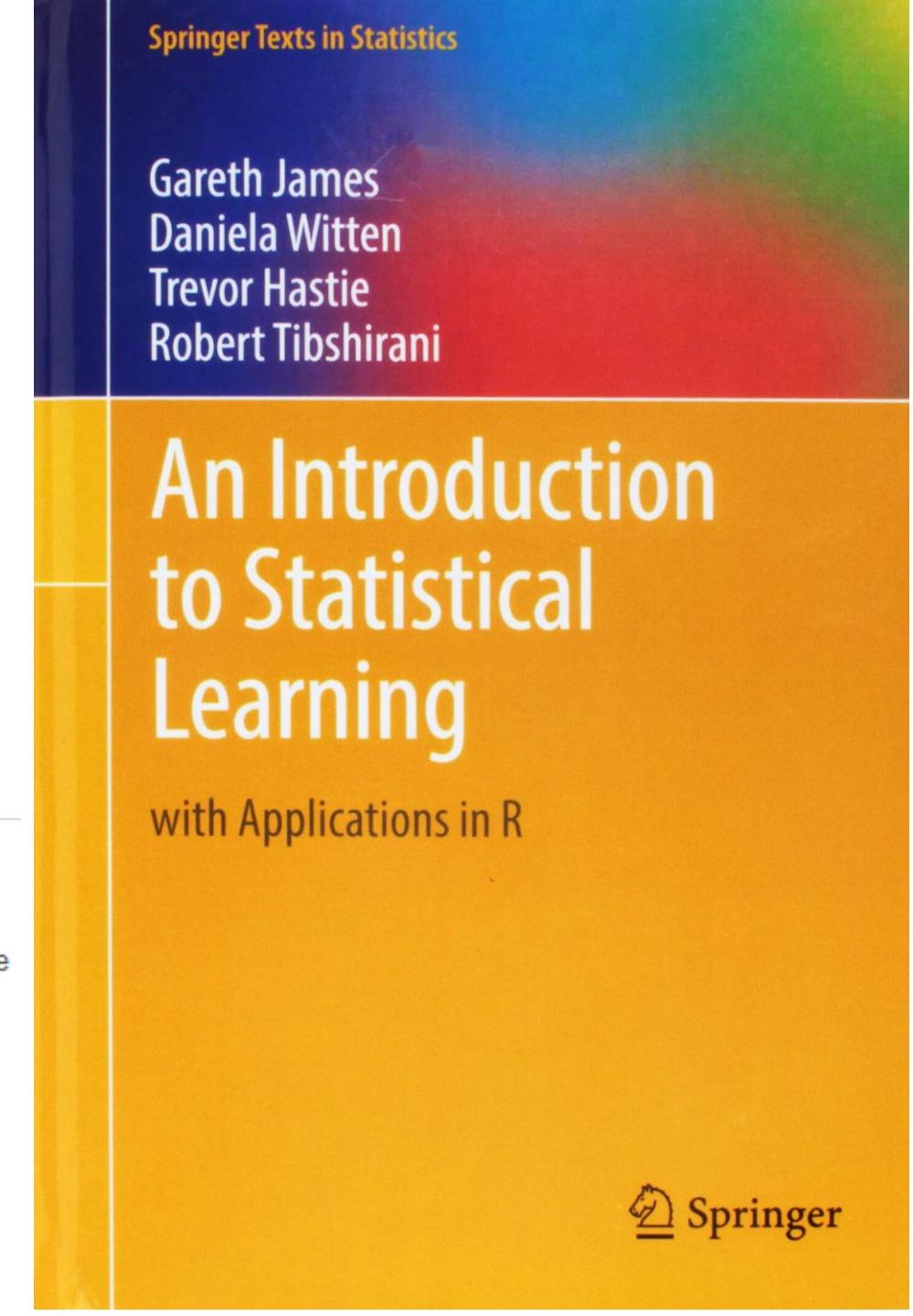
- Utilizaremos “Introduction to Statistical Learning” de Gareth James como libro complementario.
- Está disponible gratuitamente en línea, podemos conseguirlo en:

[Introduction to Statistical Learning - University of Southern California](#) ✓

[www-bcf.usc.edu/~gareth/ISL/](http://www-bcf.usc.edu/~gareth/ISL/) ▾ Traducir esta página

Home, Download the book PDF (corrected 7th printing). Statistical Learning MOOC covering the entire ISL book offered by Trevor Hastie and Rob Tibshirani.

[Data Sets and Figures](#) · [R Code for Labs](#) · [Get the Book](#) · [About this Book](#)

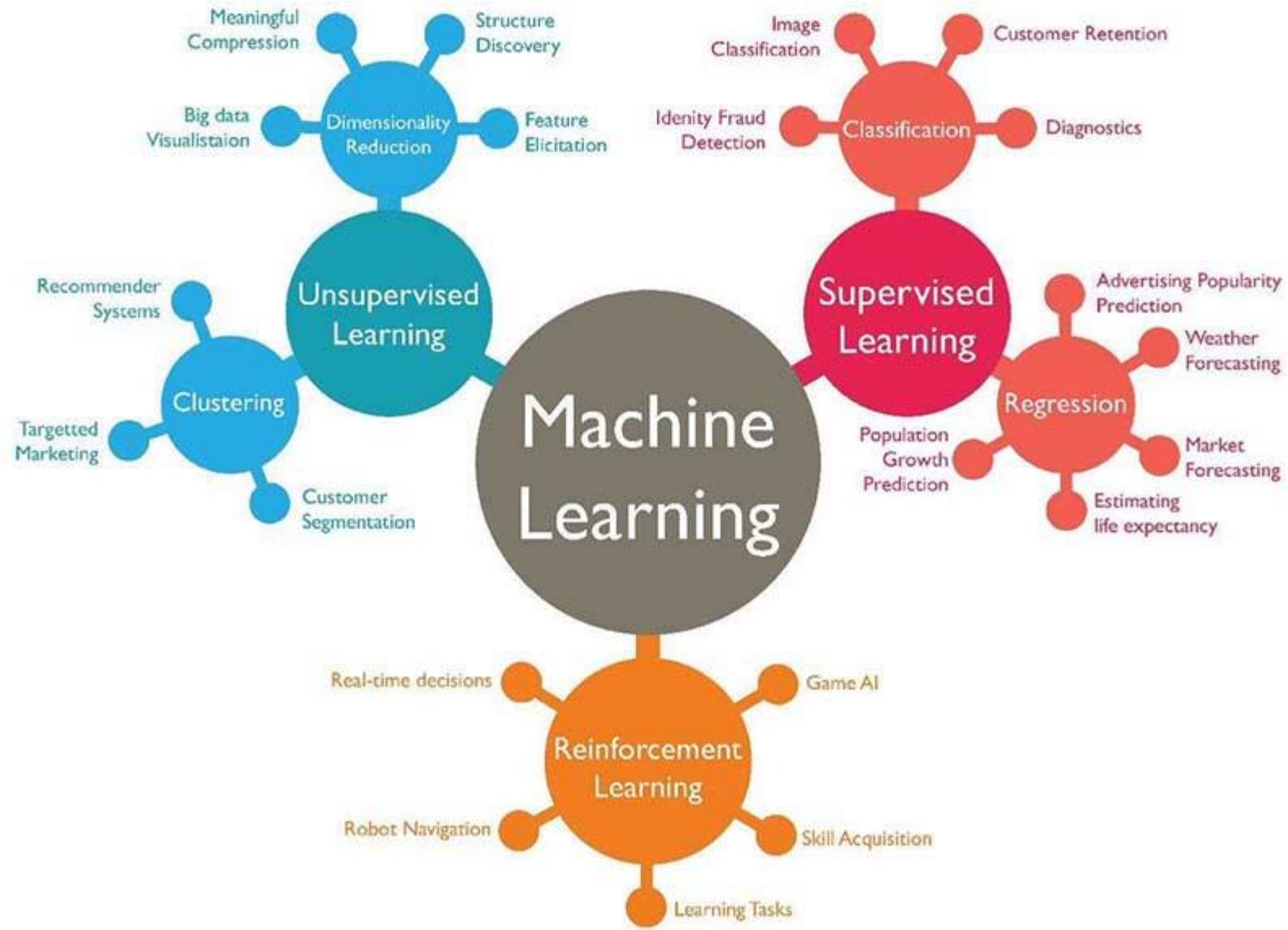


# Libro Complementario

- Los estudiantes que quieran la teoría matemática deben hacer las lecturas sugeridas.
- Los estudiantes que solo quieren aplicar los modelos y están más interesados en las aplicaciones de Python pueden simplemente enfocarse más en estos materiales.

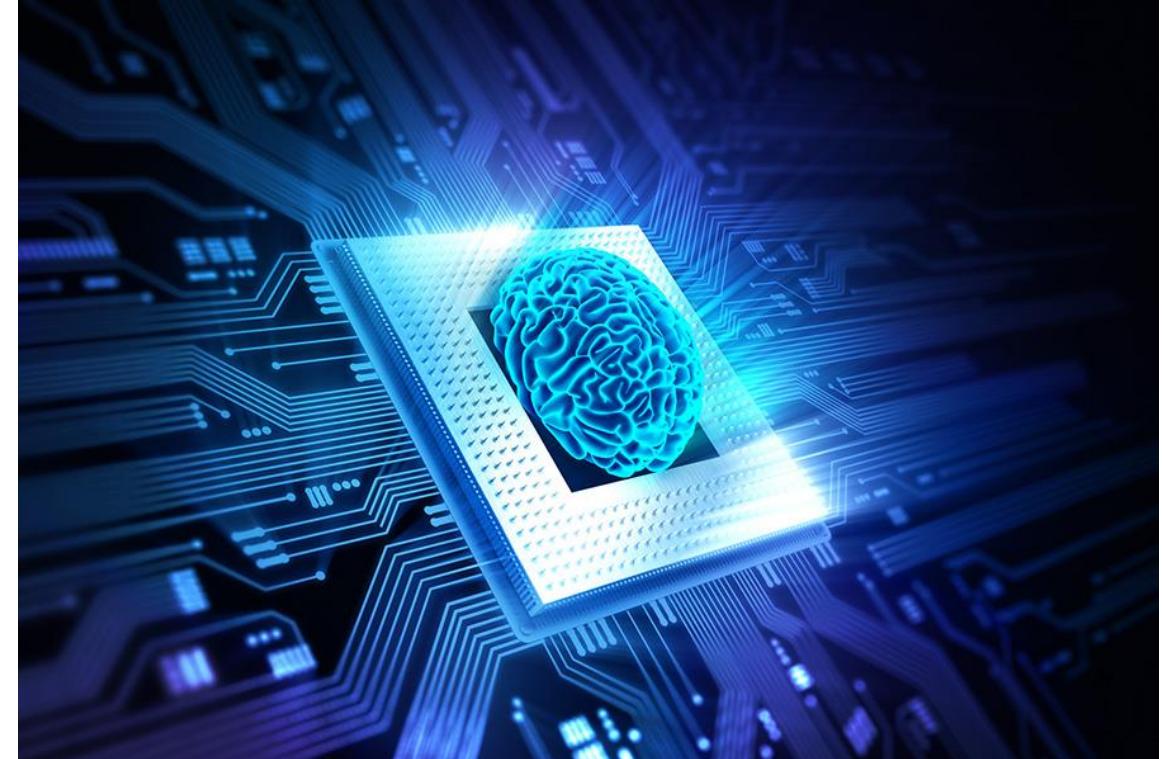
# Libro Complementario

- Lea los Capítulos 1 y 2 si quiere obtener una mejor comprensión general antes de continuar con estos materiales.



# ¿Qué es Machine Learning o Aprendizaje Automático?

- El aprendizaje automático es un método de análisis de datos que automatiza la creación de modelos analíticos.
- Mediante el uso de algoritmos que aprenden iterativamente de los datos, el aprendizaje automático permite que las computadoras encuentren información oculta sin tener que programar explícitamente dónde buscar.

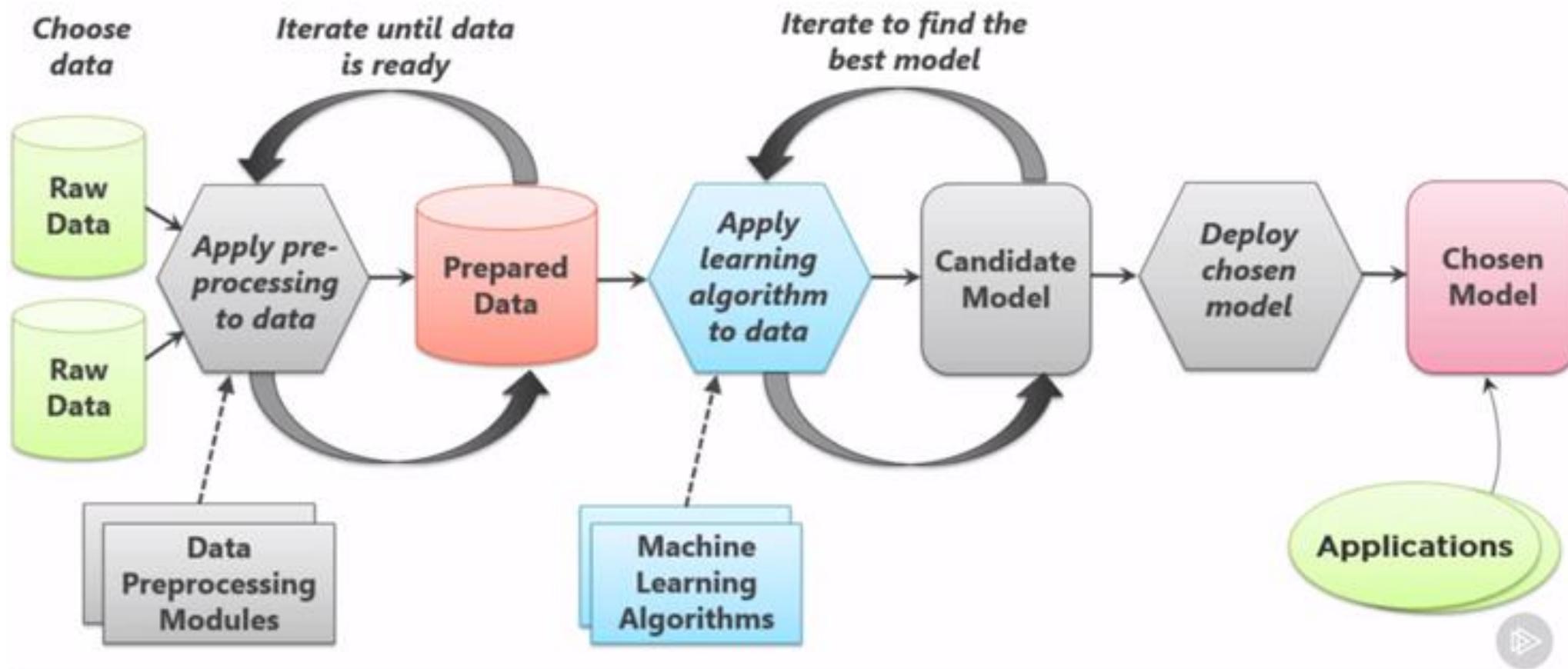


# ¿Para qué se usa?

- Detección de fraude.
- Resultados de búsqueda web.
- Anuncios en tiempo real en páginas web
- Calificación de crédito y las mejores ofertas siguientes.
- Predicción de fallas de equipos.
- Nuevos modelos de precios.
- Detección de intrusión de red.
- Motores de recomendación
- Segmentación del cliente
- Análisis de sentimiento de texto
- Predecir la rotación de clientes
- Reconocimiento de patrones e imágenes.
- Filtrado de spam de correo electrónico.
- Modelado financiero

# Proceso del Aprendizaje Automático

## The Machine Learning Process



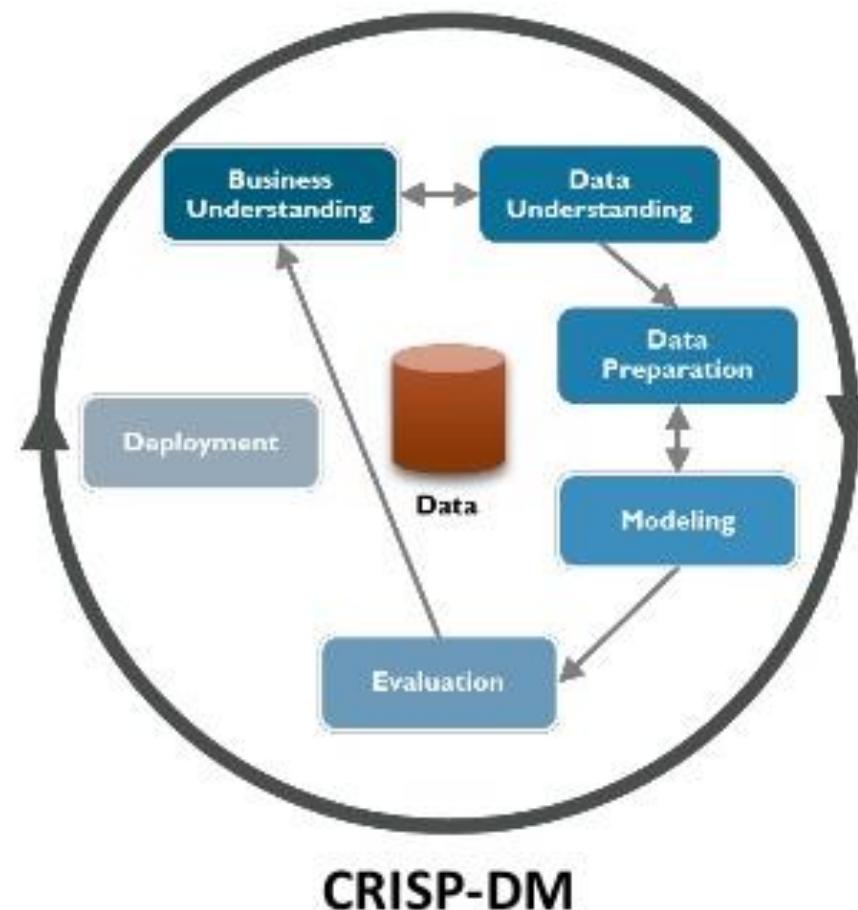
# Proceso del Aprendizaje Automático

## 1. Data Engineering – 80%

- Data extraction
- Data cleaning
- Data transformation
- Data normalization
- Feature extraction

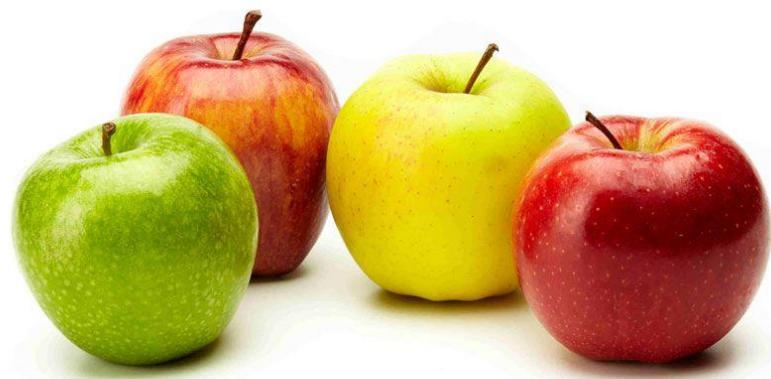
## 2. Machine Learning – 20%

- Model fitting
- Hyperparameters tuning
- Model evaluation



# Aprendizaje Supervisado

- Los algoritmos de aprendizaje supervisados se entrena usando ejemplos etiquetados, como una entrada donde se conoce el resultado deseado.
- Por ejemplo, unos objetos puede tener puntos de datos etiquetados como "M" (en mal estado) o "B" (buen estado).



buen estado



mal estado



¿Cuáles están en buen estado o mal estado?

# Aprendizaje Supervisado

- El algoritmo de aprendizaje recibe un conjunto de entradas junto con las correspondientes salidas correctas, y el algoritmo aprende comparando su salida real con las salidas correctas para encontrar errores.
- Luego modifica el modelo en consecuencia.

# Aprendizaje Supervisado

- A través de métodos como la clasificación, la regresión, la predicción y el aumento de gradiente, el aprendizaje supervisado usa patrones para predecir los valores de la etiqueta en datos adicionales no etiquetados.
- El aprendizaje supervisado se usa comúnmente en aplicaciones donde los datos históricos predicen eventos futuros probables.

# Aprendizaje Supervisado

- A través de métodos como la clasificación, la regresión, la predicción y el aumento de gradiente, el aprendizaje supervisado usa patrones para predecir los valores de la etiqueta en datos adicionales no etiquetados.
- El aprendizaje supervisado se usa comúnmente en aplicaciones donde los datos históricos predicen eventos futuros probables.

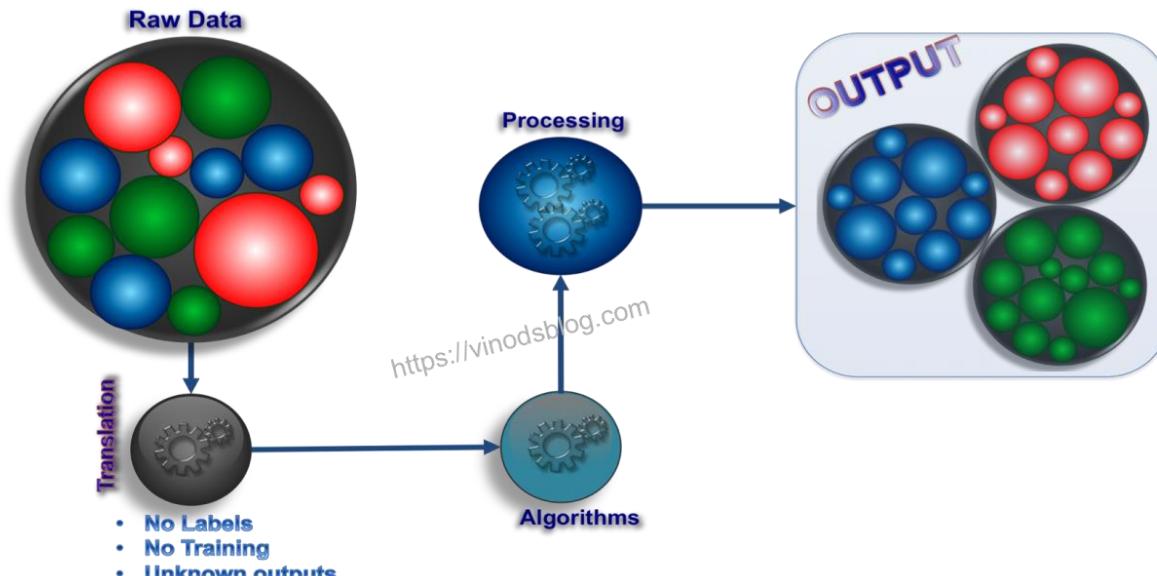
# Aprendizaje Supervisado

- Por ejemplo, puede anticipar cuándo es probable que las transacciones con tarjeta de crédito sean fraudulentas o qué cliente de seguros es probable que presente un reclamo.
- O puede intentar predecir el precio de una casa en función de las diferentes características de las casas para las que tenemos datos de precios históricos.



# Aprendizaje No Supervisado

- El aprendizaje no supervisado se usa con datos que no tienen etiquetas históricas.
- Al sistema no se le dice la "respuesta correcta". El algoritmo debe descubrir lo que se muestra.
- El objetivo es explorar los datos y encontrar alguna estructura dentro.



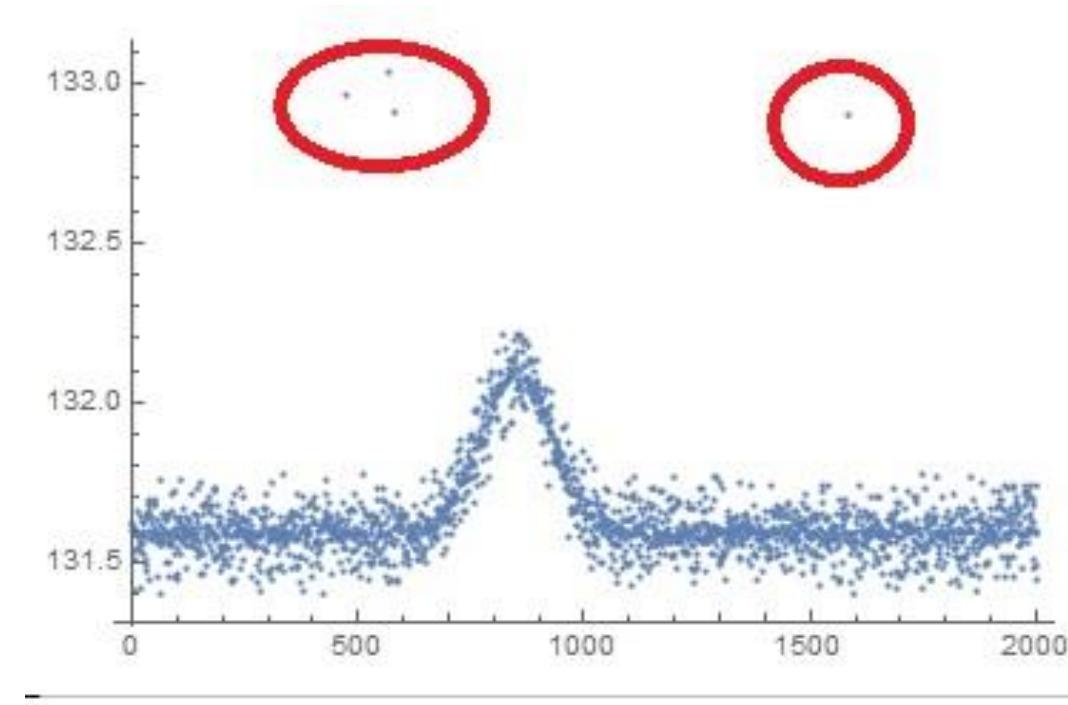
# Aprendizaje No Supervisado

- O puede encontrar los principales atributos que separan segmentos de clientes entre sí.
- Las técnicas populares incluyen mapas autoorganizados, mapeo del vecino más cercano, clustering k-means y descomposición de valores singulares.



# Aprendizaje No Supervisado

- Estos algoritmos también se utilizan para segmentar temas en un texto, recomendar elementos e identificar valores atípicos de datos.



# Aprendizaje Reforzado

- El aprendizaje reforzado a menudo se usa para robótica, juegos y navegación.
- Con el aprendizaje reforzado, el algoritmo descubre a través de prueba y error qué acciones rinden las mayores recompensas.



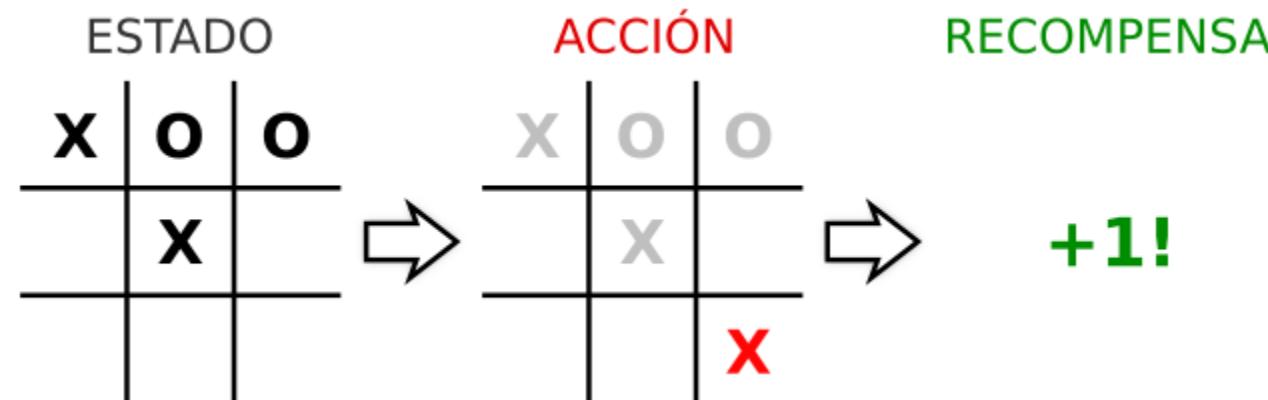
# Aprendizaje Reforzado

- Este tipo de aprendizaje tiene tres componentes principales: el agente (el que aprende o el que toma las decisiones), el entorno (todo con lo que el agente interactúa) y las acciones (lo que el agente puede hacer).



# Aprendizaje Reforzado

- El objetivo es que el agente elija acciones que maximicen la recompensa esperada durante un período de tiempo determinado.
- El agente alcanzará el objetivo mucho más rápido siguiendo una buena política.
- Entonces, el objetivo en el aprendizaje de refuerzo es aprender la mejor política.



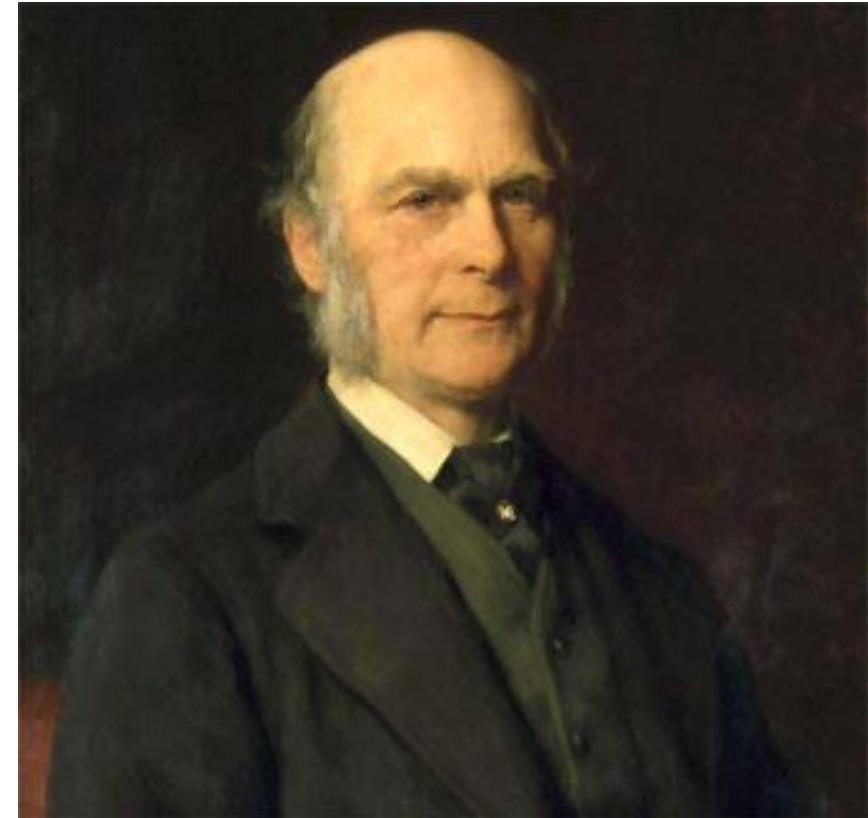
# Introducción a la Regresión Lineal

# Lectura Sugerida

- Capítulos 2 y 3 del libro “Introduction to Statistical Learning” de Gareth James

# Historia

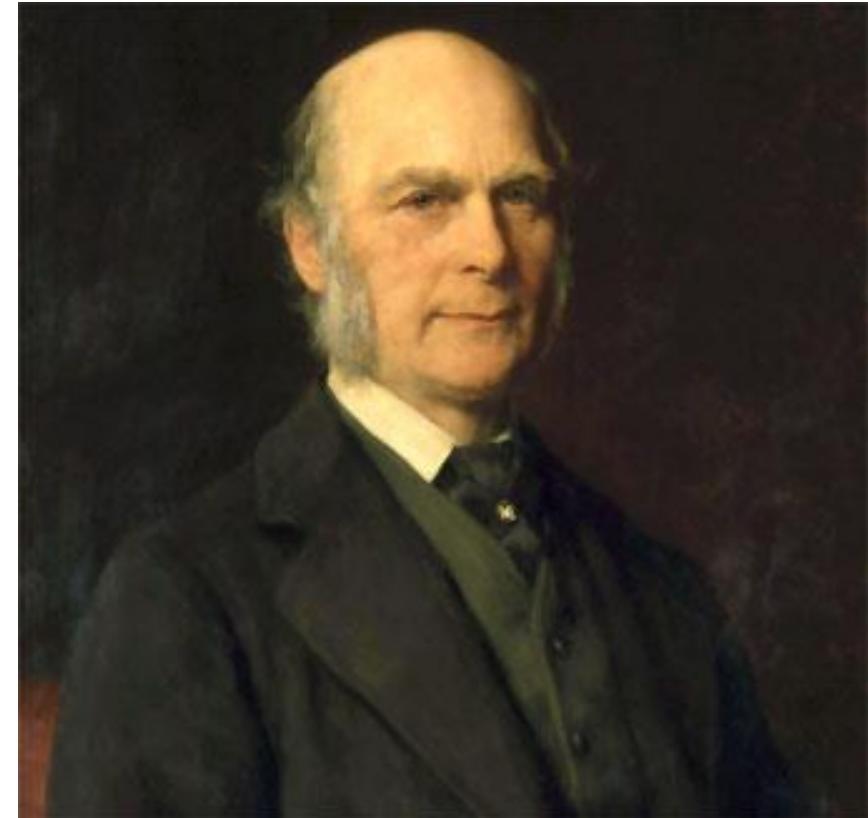
Todo comenzó en el siglo XIX con un tipo llamado Francis Galton. Galton estaba estudiando la relación entre los padres y sus hijos. En particular, investigó la relación entre las alturas de los padres y sus hijos.



# Historia

Lo que descubrió fue que el hijo de cualquier hombre tenía a ser más o menos tan alto como su padre.

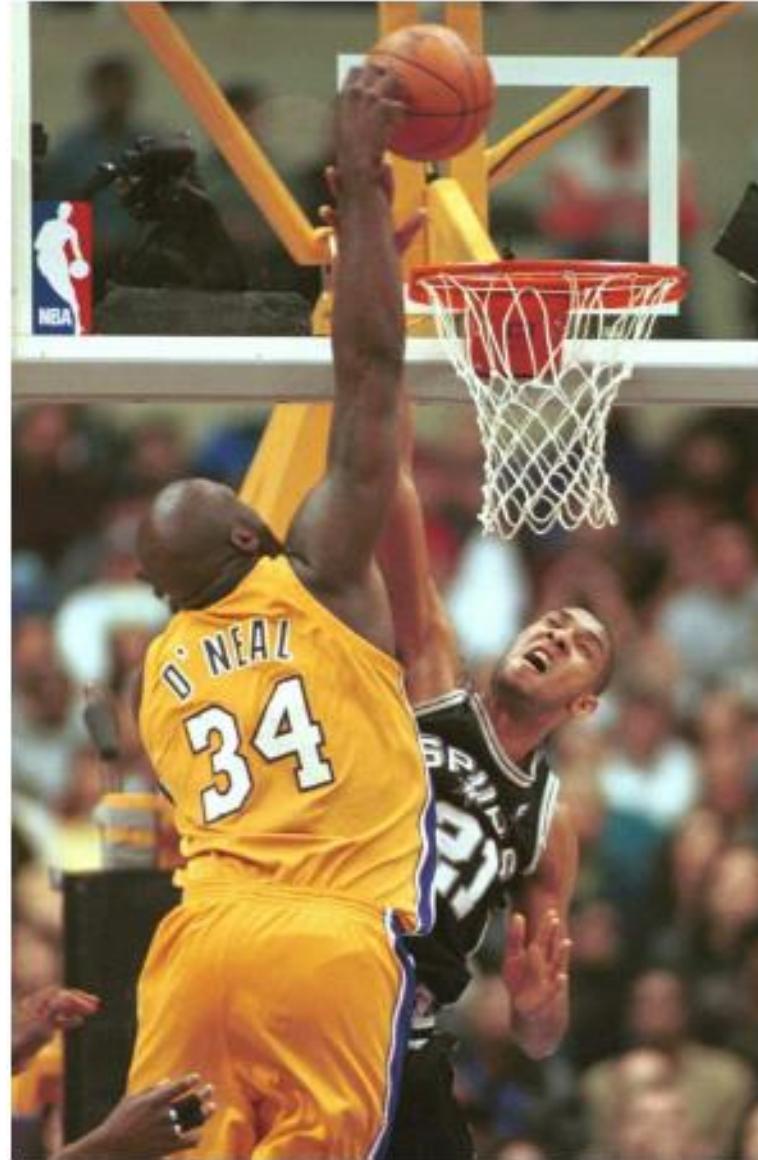
Sin embargo, el descubrimiento de Galton fue que la altura de un hijo tenía a estar más cerca de la estatura promedio general de todas las personas.



# Ejemplo

Tomemos a Shaquille O'Neal como ejemplo. Shaq es realmente alto: 7 pies 1 pulgada (2,16 metros).

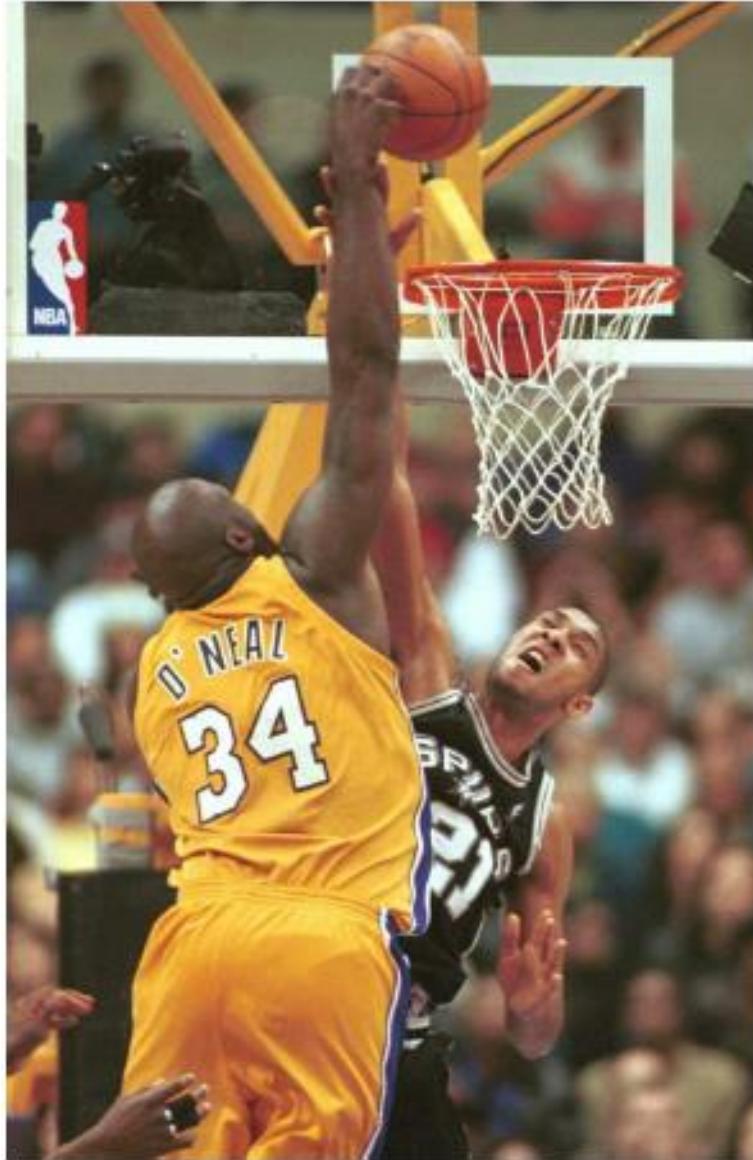
Si Shaq tiene un hijo, es probable que sea bastante alto también. Sin embargo, Shaq es una anomalía, tal que también hay una gran posibilidad de que su hijo no sea tan alto como Shaq.



# Ejemplo

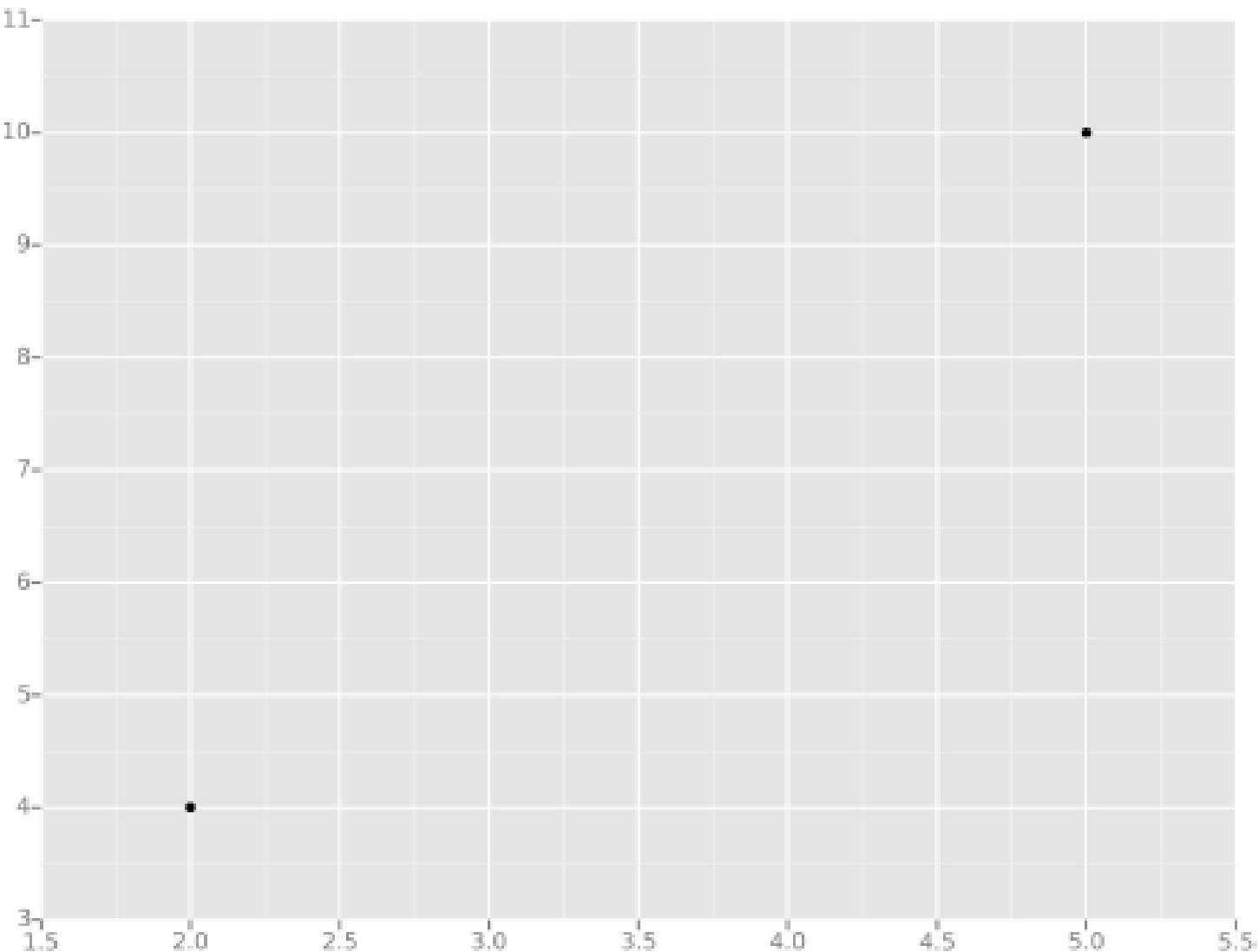
Resulta que este es el caso: el hijo de Shaq es bastante alto, 6 pies 7 pulgadas (2,0 metros), pero no tan alto como su padre.

Galton llamó a este fenómeno regresión, como en "La altura de un hijo de un padre tiende a retroceder (o deriva hacia) la altura media (promedio)".



# Ejemplo

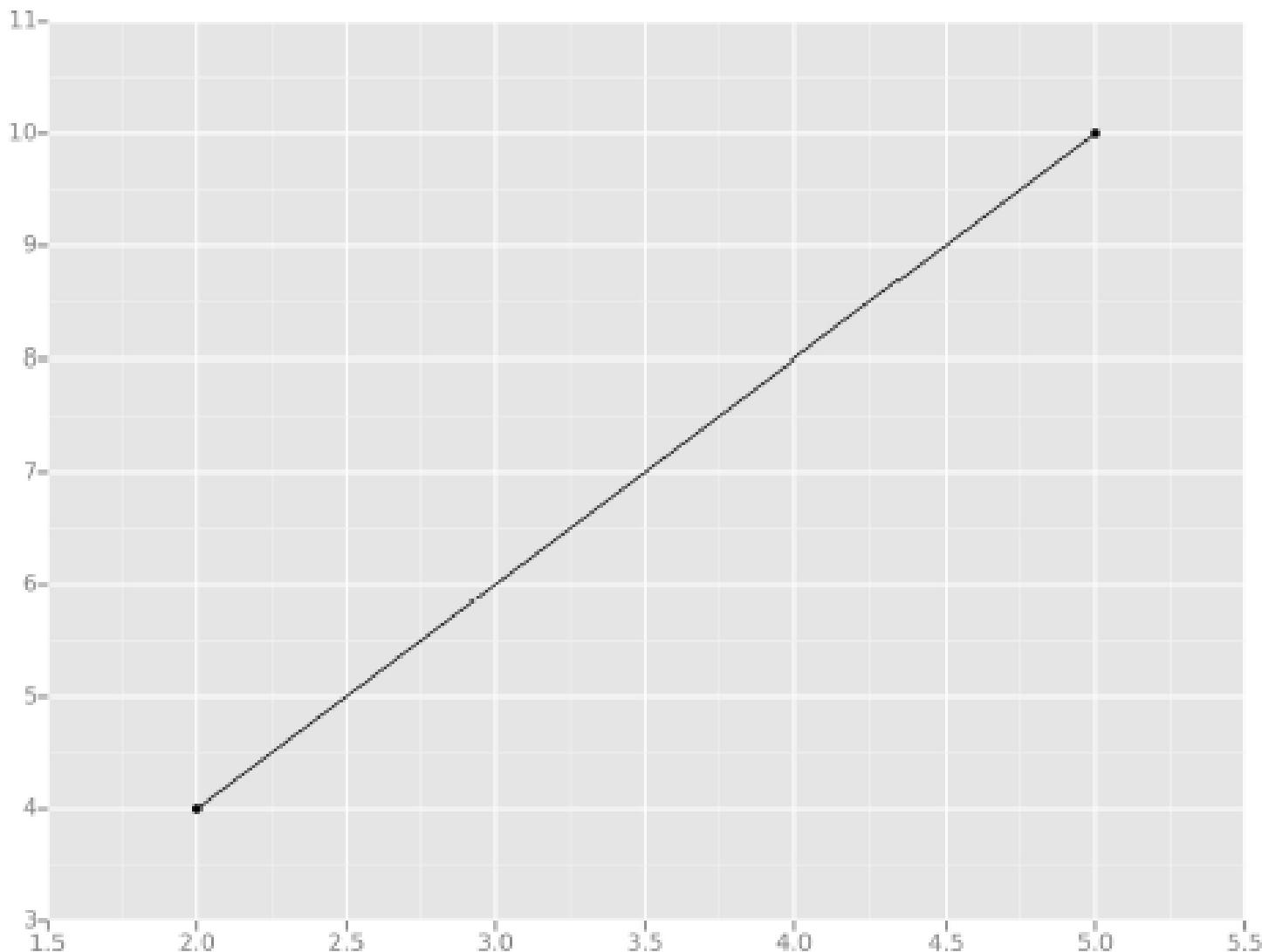
Tomemos el ejemplo más simple posible: calcular una regresión con solo 2 puntos de datos.



# Ejemplo

Todo lo que intentamos hacer cuando calculamos nuestra línea de regresión es dibujar una línea lo más cercana posible a cada punto.

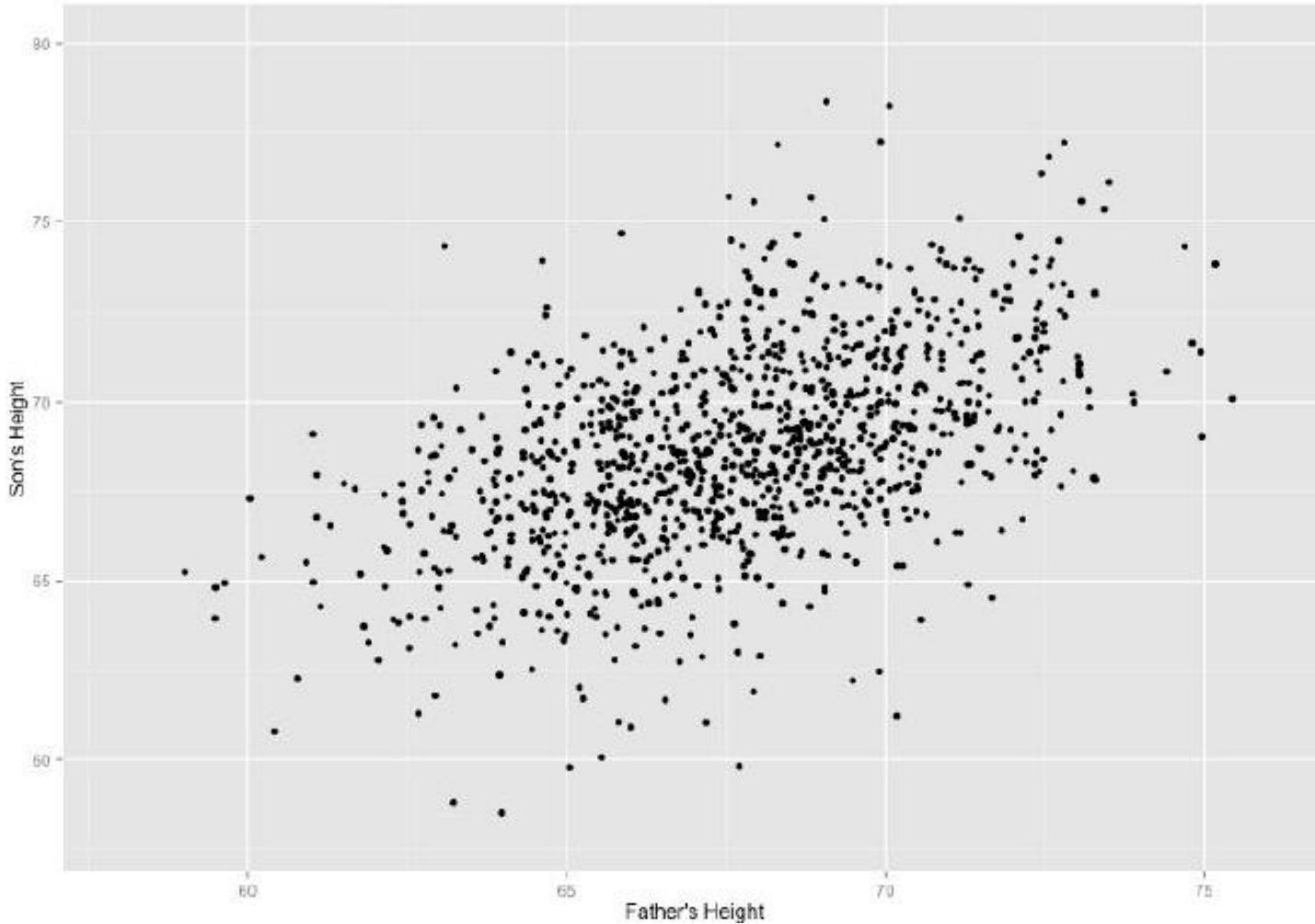
Para la regresión lineal clásica, o el "Método de mínimos cuadrados", solo se mide la cercanía en la dirección "arriba y abajo"



# Ejemplo

Ahora, ¿no sería genial si pudiéramos aplicar este mismo concepto a un gráfico con más de dos puntos de datos?

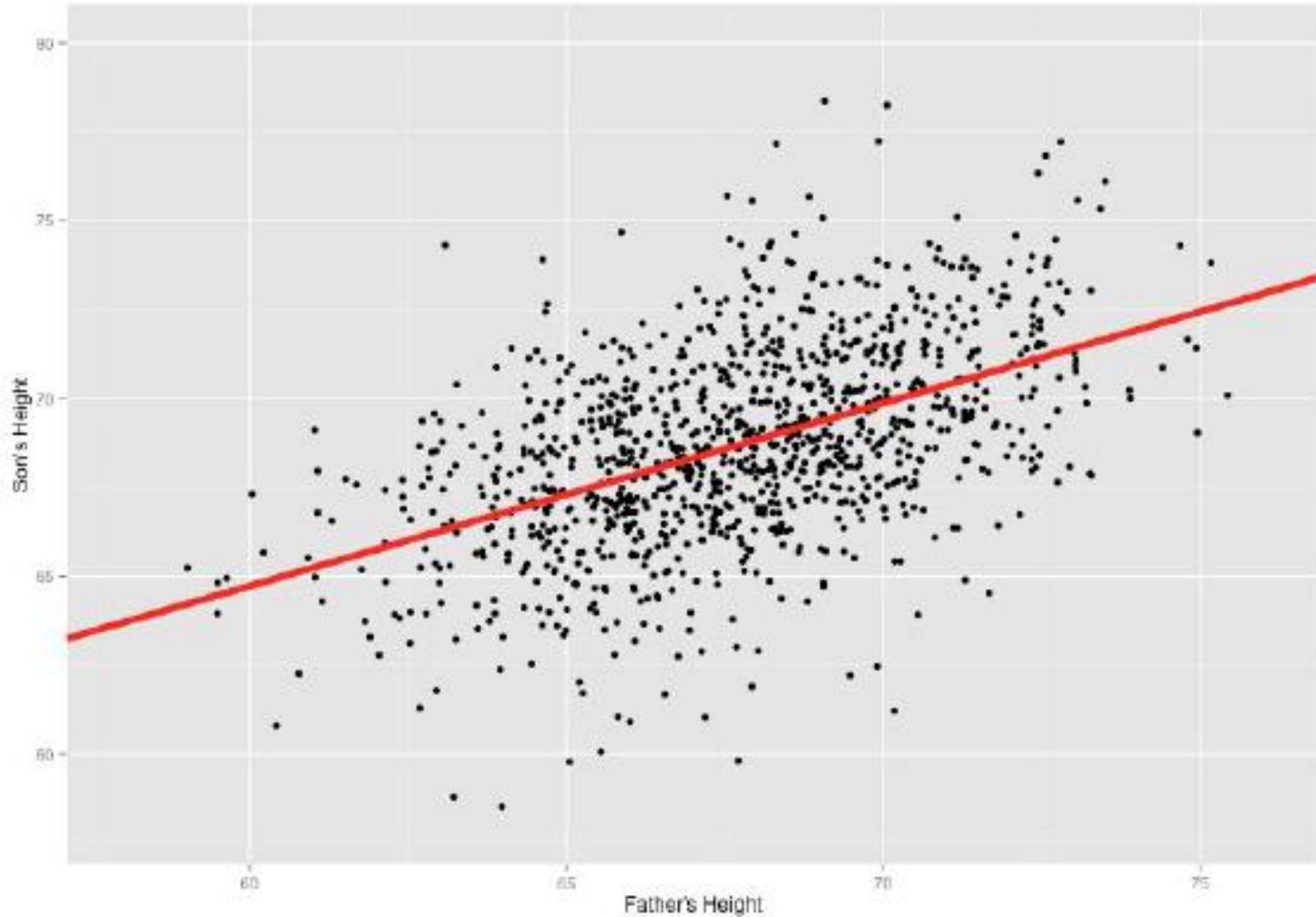
Al hacer esto, podríamos tomar múltiples hombres y las alturas de su hijos y hacer cosas como decirle a un hombre lo alto que esperamos que sea su hijo ... ¡incluso antes de que tenga un hijo!



# Ejemplo

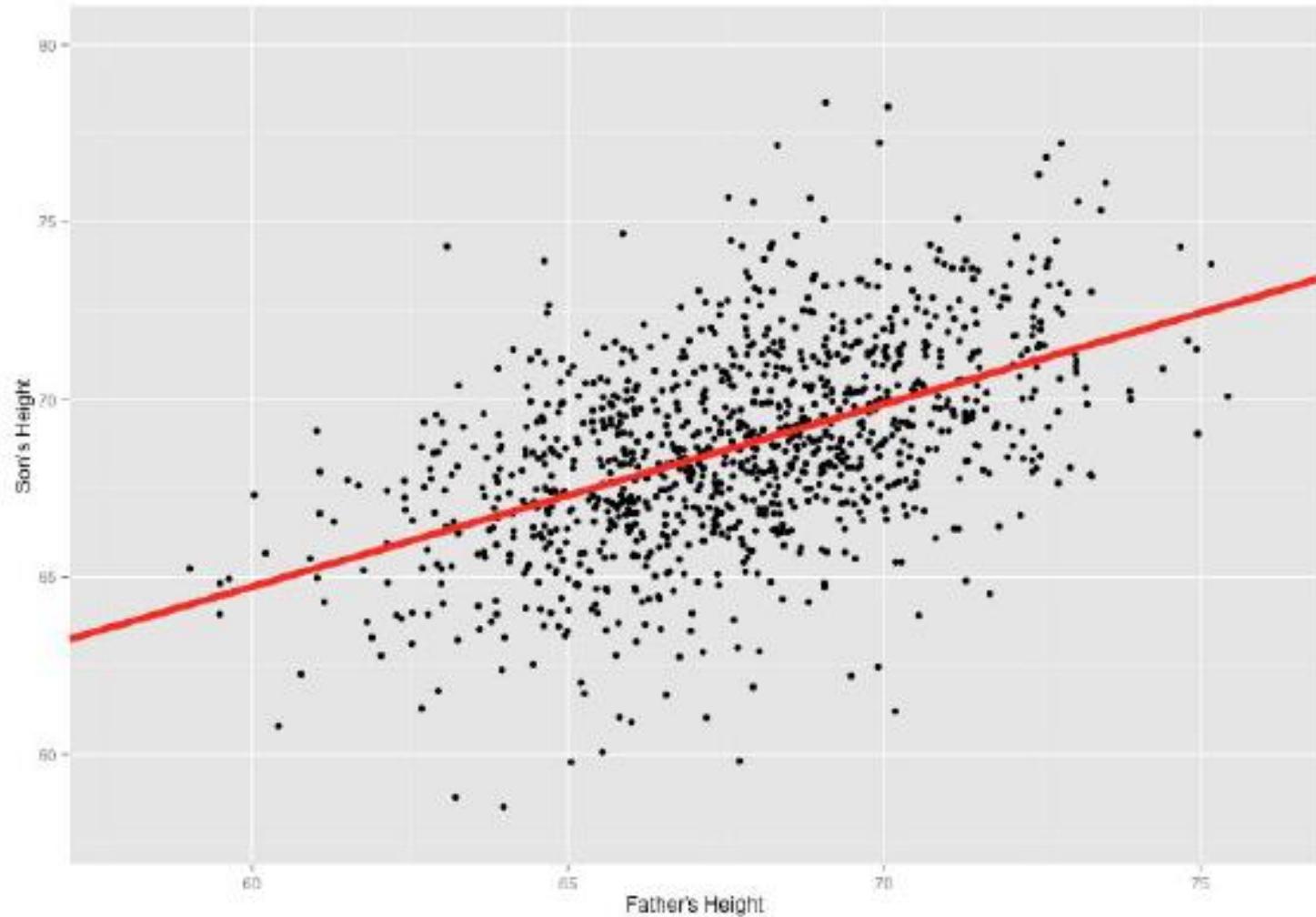
Nuestro objetivo con la regresión lineal es minimizar la distancia vertical entre todos los puntos de datos y nuestra línea.

Entonces, al determinar la mejor línea, intentamos minimizar la distancia entre todos los puntos y su distancia a nuestra línea.



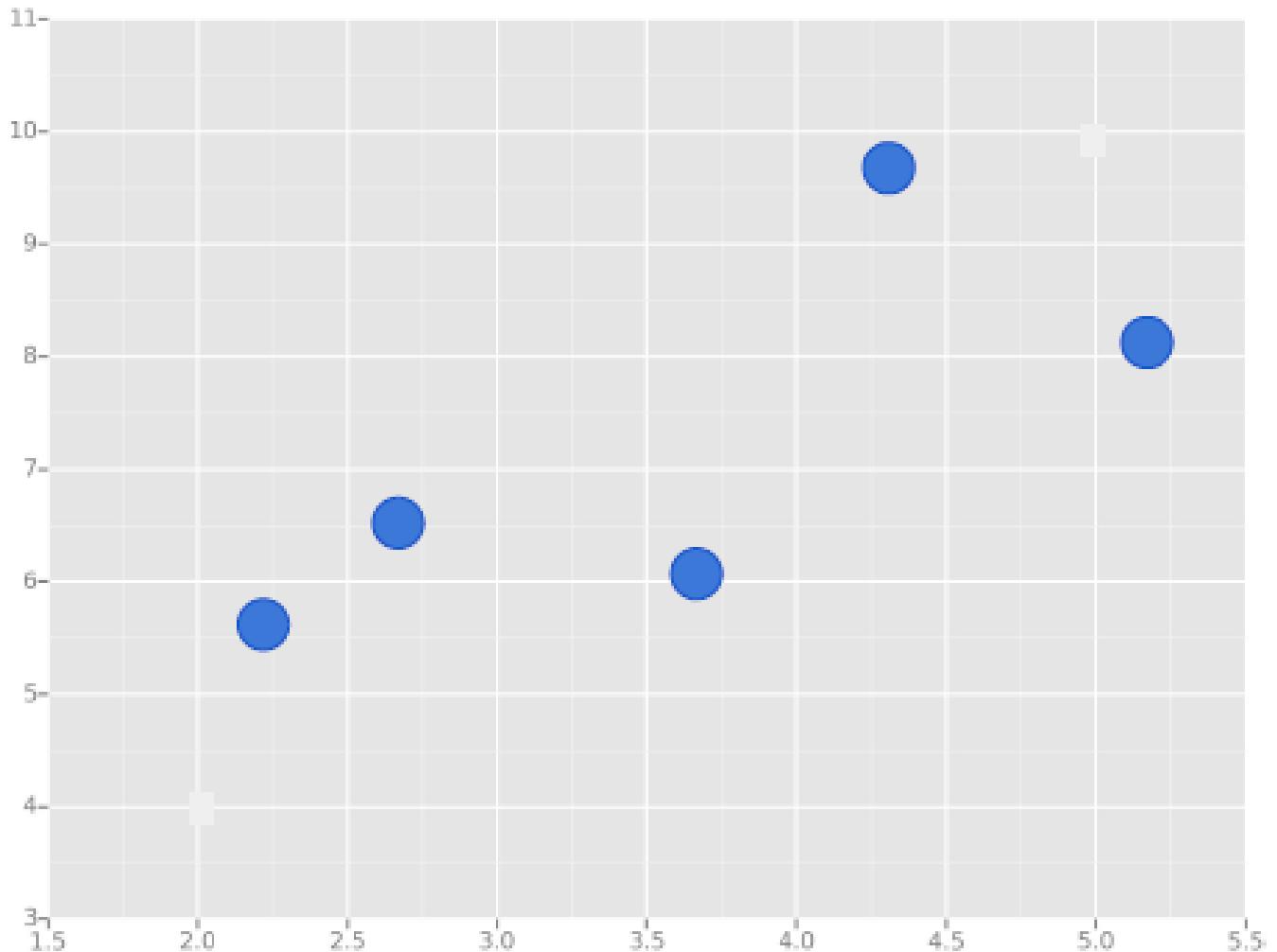
# Ejemplo

Hay muchas formas diferentes de minimizar esto (suma de errores al cuadrados, suma de errores absolutos, etc.), pero todos estos métodos tienen el objetivo general de minimizar esta distancia.



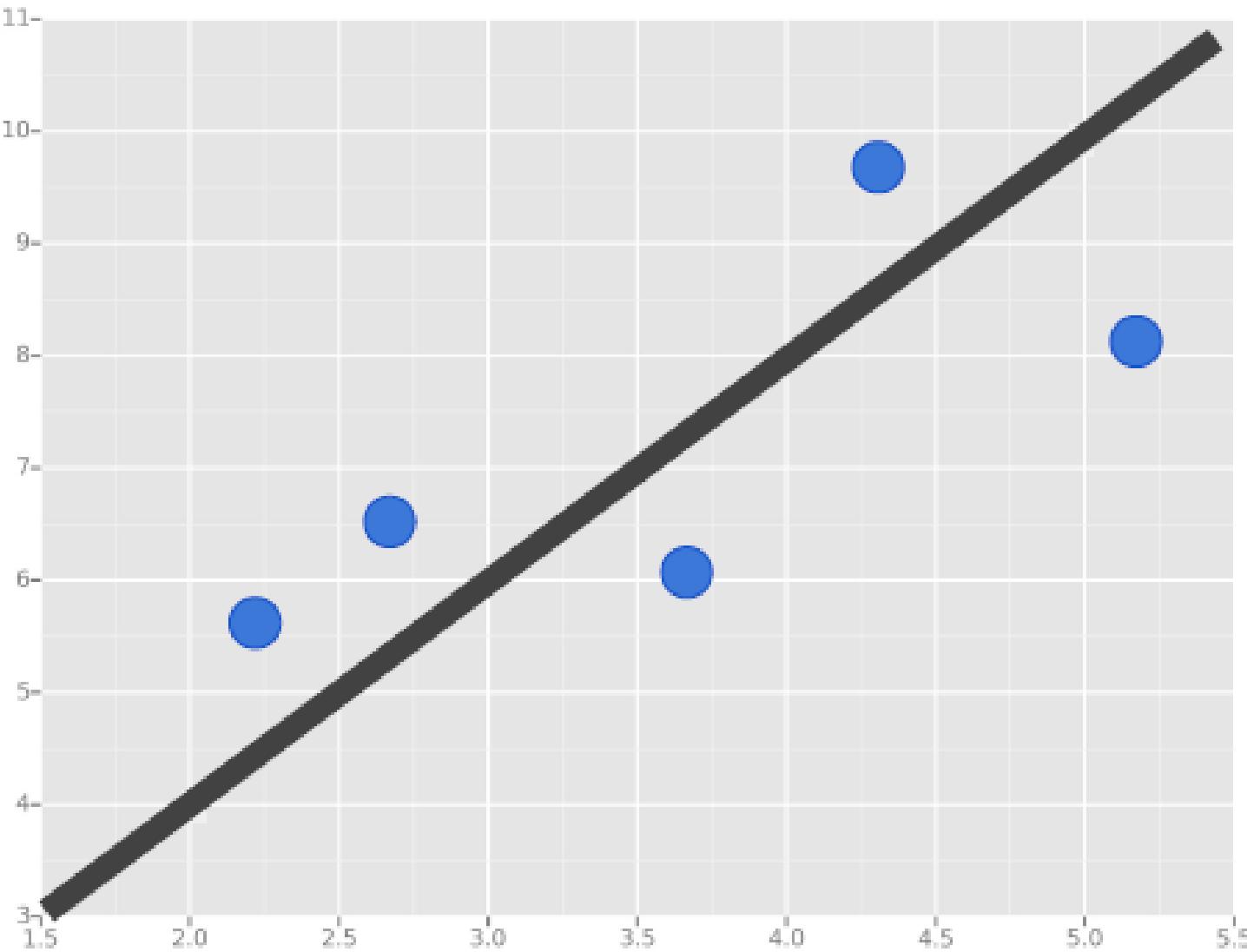
# Ejemplo

Por ejemplo, uno de los métodos más populares es el método de mínimos cuadrados. Aquí tenemos puntos de datos azules a lo largo de un eje x e y.



# Ejemplo

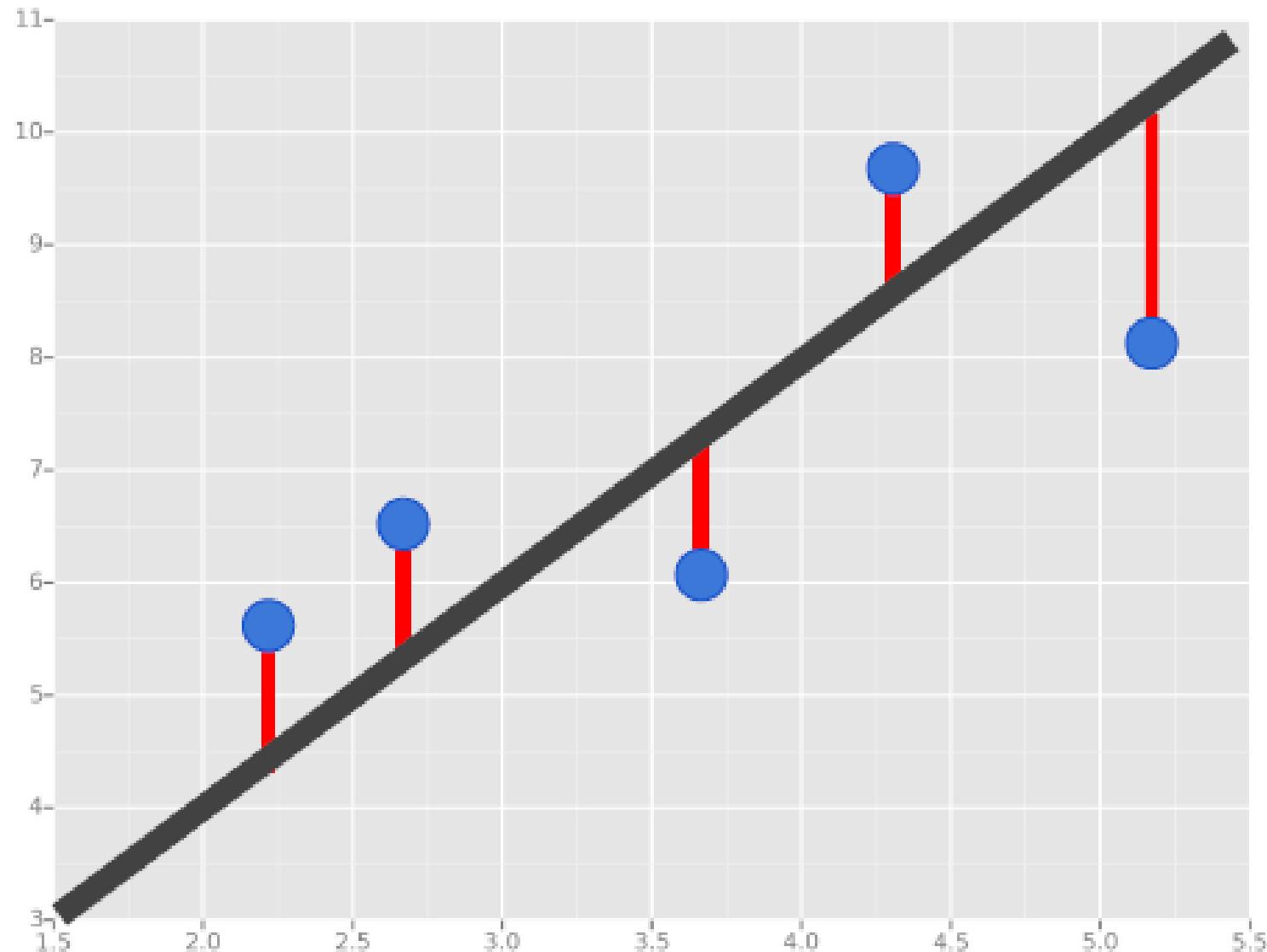
Ahora queremos ajustar una línea de regresión lineal. La pregunta es, ¿cómo decidimos qué línea es la más adecuada?



# Ejemplo

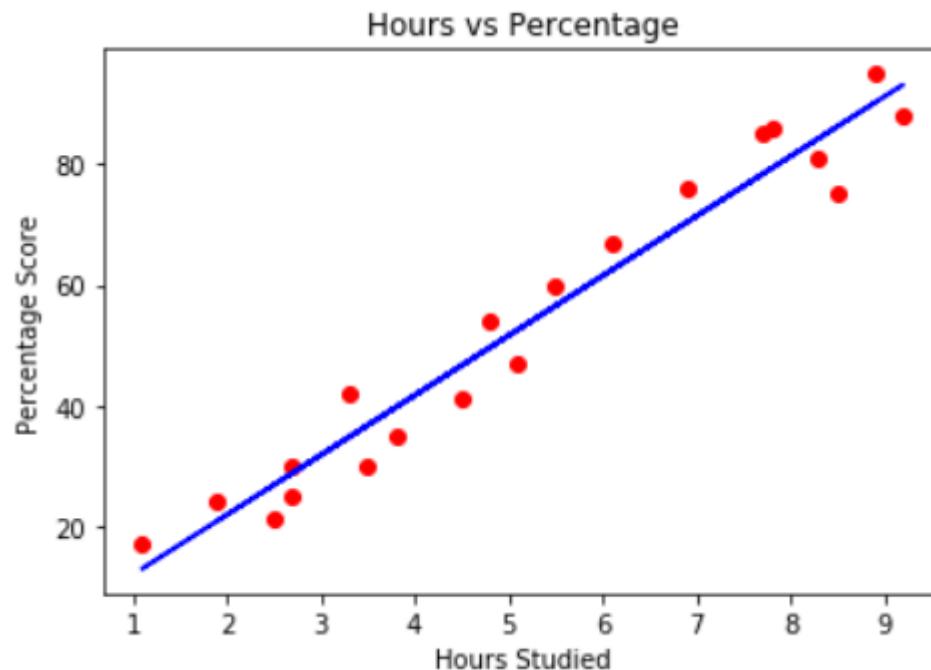
Utilizaremos el método de mínimos cuadrados, que se ajusta minimizando la suma de cuadrados de los residuos.

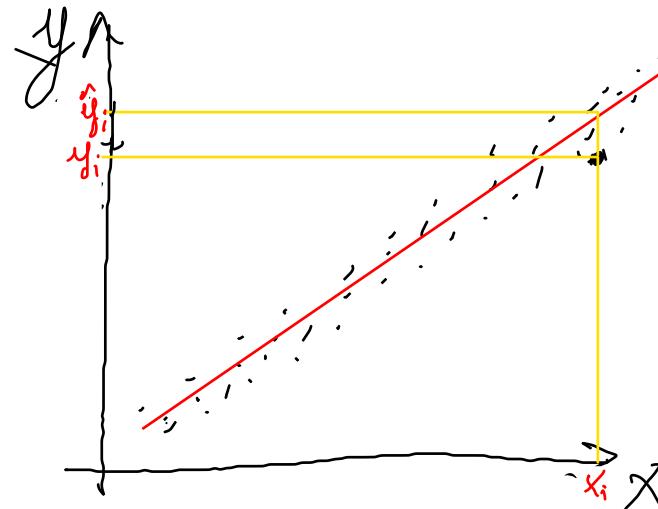
Los residuos para una observación son la diferencia entre la observación (el valor y) y la línea ajustada.



# Ejemplo con Python

- Ahora usaremos SciKit-Learn y Python para crear un modelo de regresión lineal. Luego resolverá un ejercicio propuesto y revisaremos las soluciones.





$$\hat{y}_i = \hat{B}_0 + \hat{B}_1 x_i$$

$\hat{y}_i$ : el valor predicho para  $x_i$

$y_i$ : el valor verdadero para  $x_i$

Para minimizar las distancias verticales de cada punto a la linea podemos usar varios algoritmos

Permiten encontrar los mejores valores para  $B_0, B_1$  del modelo

Uno de los algoritmos es el **Método de Mínimos Cuadrados**

X	y
$x_1$	$y_1$
$x_2$	$y_2$
:	:
$x_n$	$y_n$

$$\hat{y}_i = \hat{B}_0 + \hat{B}_1 x_i \quad \text{error} \quad e_i = y_i - \hat{y}_i$$

Suma residual de los errores al cuadrado  
 $RSS = e_1^2 + e_2^2 + \dots + e_n^2$

$$\hat{B}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}_i)}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

donde:  $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\hat{B}_0 = \bar{y} - \hat{B}_1 \bar{x}$$

$y$  = ventas

$x$  = inversión en publicidad en TV

$$B_1 = 0.04754$$

$$B_0 = 7.03259$$

$$2102.53058313135000$$

RSS

$$\hat{y} = 7.03259 + 0.04754x$$

↑  
Intercepto

$$RSS = 2102.53058$$

Tarea: determine el modelo de regresión lineal simple para  $y$  = ventas para  $x$  = publicidad en radio

# Regresión Lineal Simple

## Modelo de Regresión Lineal Simple para:

- x = inversión en publicidad en TV
- y = total de ventas

### Importar librerías

In [1]:

```
1 import pandas as pd
2 import numpy as np
3 import matplotlib.pyplot as plt
4 import seaborn as sns
```

### Recuperar los datos y explorarlos

In [5]:

```
1 datos = pd.read_csv('Advertising.csv', index_col=0)
```

In [9]:

```
1 datos.head()
```

Out[9]:

	TV	Radio	Newspaper	Sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9

In [7]:

```
1 datos.tail()
```

Out[7]:

	TV	Radio	Newspaper	Sales
<b>196</b>	38.2	3.7	13.8	7.6
<b>197</b>	94.2	4.9	8.1	9.7
<b>198</b>	177.0	9.3	6.4	12.8
<b>199</b>	283.6	42.0	66.2	25.5
<b>200</b>	232.1	8.6	8.7	13.4

In [10]:

```
1 datos.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 200 entries, 1 to 200
Data columns (total 4 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   TV          200 non-null    float64
 1   Radio        200 non-null    float64
 2   Newspaper    200 non-null    float64
 3   Sales        200 non-null    float64
dtypes: float64(4)
memory usage: 7.8 KB
```

In [11]:

```
1 datos.describe()
```

Out[11]:

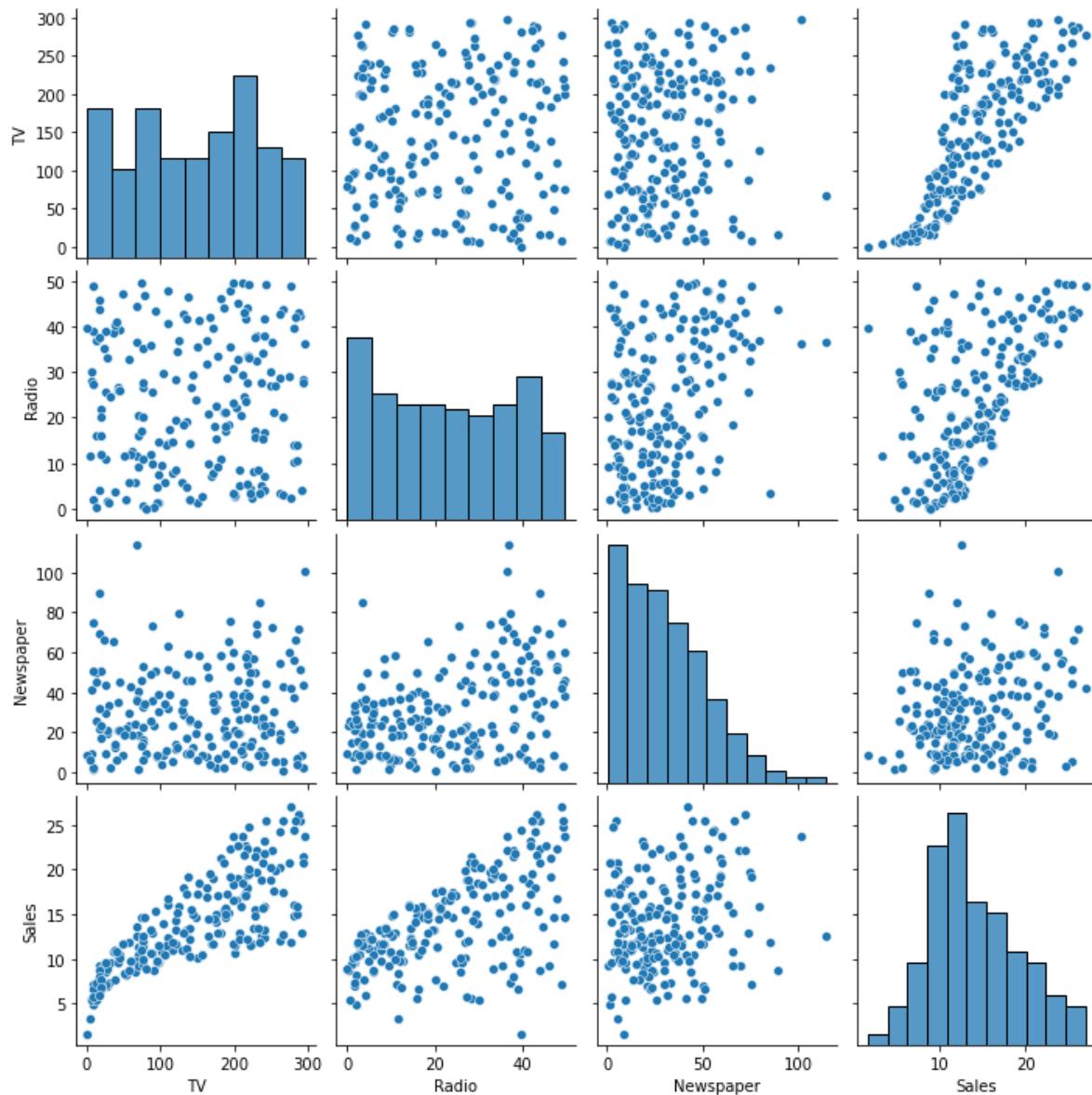
	TV	Radio	Newspaper	Sales
<b>count</b>	200.000000	200.000000	200.000000	200.000000
<b>mean</b>	147.042500	23.264000	30.554000	14.022500
<b>std</b>	85.854236	14.846809	21.778621	5.217457
<b>min</b>	0.700000	0.000000	0.300000	1.600000
<b>25%</b>	74.375000	9.975000	12.750000	10.375000
<b>50%</b>	149.750000	22.900000	25.750000	12.900000
<b>75%</b>	218.825000	36.525000	45.100000	17.400000
<b>max</b>	296.400000	49.600000	114.000000	27.000000

In [13]:

```
1 sns.pairplot(datos)
```

Out[13]:

```
<seaborn.axisgrid.PairGrid at 0x1ab78d19ee0>
```

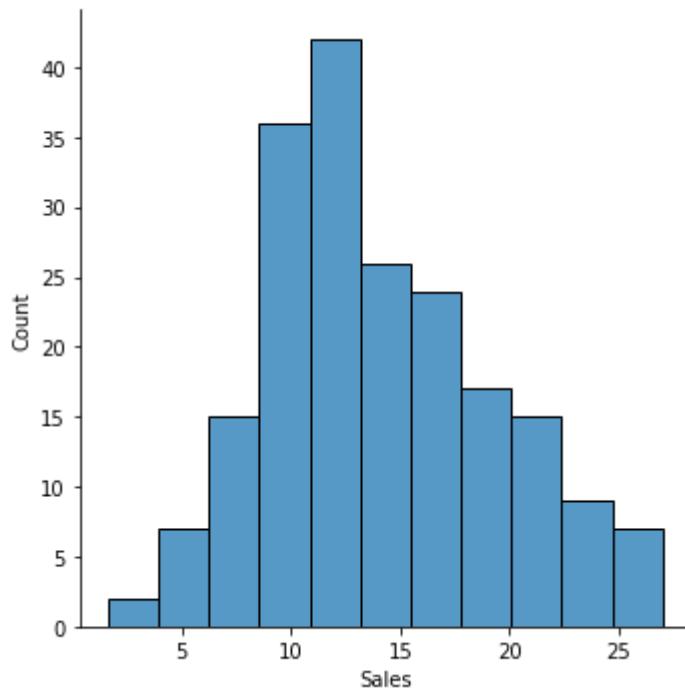


In [14]:

```
1 sns.displot(datos['Sales'])
```

Out[14]:

```
<seaborn.axisgrid.FacetGrid at 0x1ab7a24eee0>
```



In [15]:

```
1 datos.corr()
```

Out[15]:

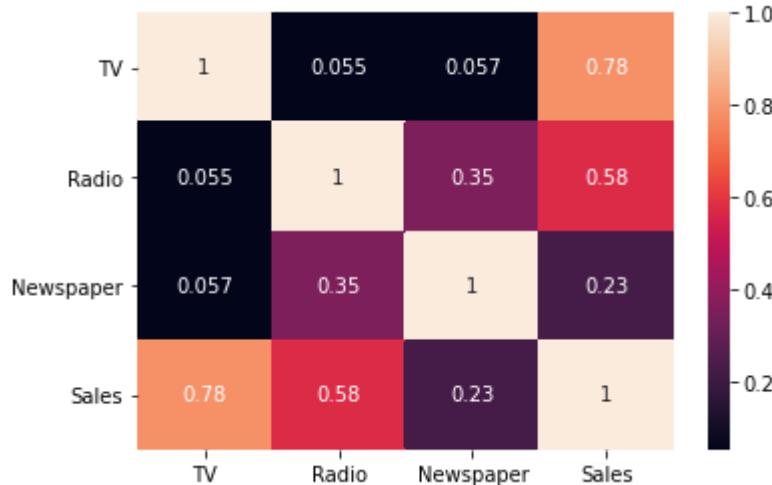
	TV	Radio	Newspaper	Sales
TV	1.000000	0.054809	0.056648	0.782224
Radio	0.054809	1.000000	0.354104	0.576223
Newspaper	0.056648	0.354104	1.000000	0.228299
Sales	0.782224	0.576223	0.228299	1.000000

In [17]:

```
1 sns.heatmap(datos.corr(), annot=True)
```

Out[17]:

&lt;AxesSubplot:&gt;



## Modelo de Regresión Lineal Simple

In [18]:

```
1 X = datos[['TV']]
```

In [19]:

```
1 y = datos['Sales']
```

Regresión Lineal con el método de mínimos cuadrados

In [20]:

```
1 from sklearn.linear_model import LinearRegression
```

In [21]:

```
1 modelo = LinearRegression()
```

In [22]:

```
1 modelo.fit(X,y)
```

Out[22]:

LinearRegression()

### Parámetros obtenidos en el modelo

Muestra el interceptor Beta 0

In [23]:

```
1 print(modelo.intercept_)
```

7.032593549127693

Muestra los otros Beta

In [24]:

```
1 print(modelo.coef_)
```

[0.04753664]

## Predicciones

In [25]:

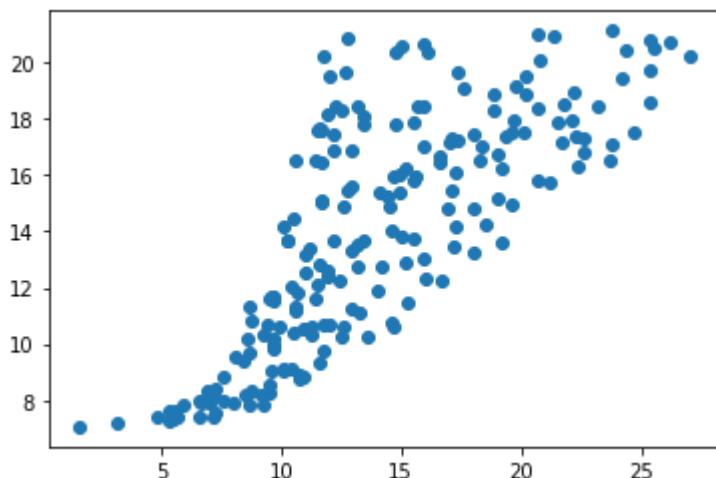
```
1 predicciones = modelo.predict(X)
```

In [26]:

```
1 plt.scatter(y,predicciones)
```

Out[26]:

<matplotlib.collections.PathCollection at 0x1ab7c438e20>

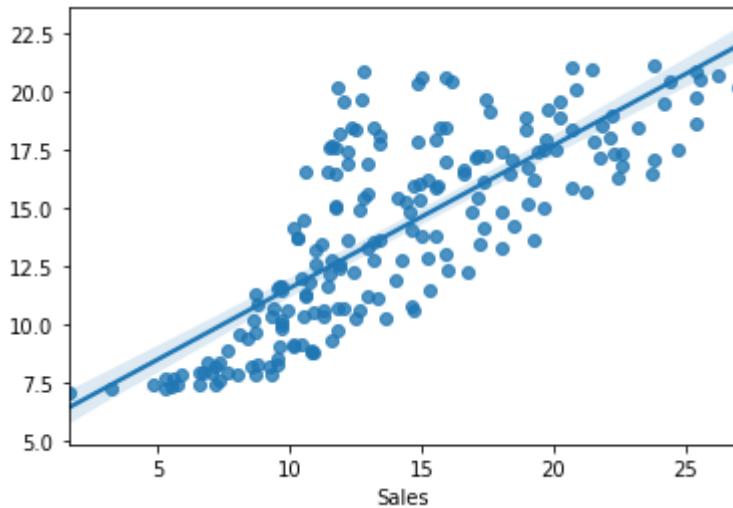


In [27]:

```
1 sns.regplot(x=y, y=predicciones, data=datos)
```

Out[27]:

```
<AxesSubplot:xlabel='Sales'>
```



## Métricas de Evaluación

In [28]:

```
1 from sklearn import metrics
```

In [29]:

```
1 MSE = metrics.mean_squared_error(y, predicciones)
```

In [30]:

```
1 RSS = MSE*200
```

In [31]:

```
1 print(RSS)
```

2102.5305831313512

# Regresión Lineal - Ejercicio Propuesto

Acaba de obtener un contrato con una empresa de comercio electrónico con sede en la ciudad de Nueva York que vende ropa en línea, pero también tienen sesiones de asesoramiento sobre vestimenta y estilo en la tienda. Los clientes entran a la tienda, tienen sesiones / reuniones con un estilista personal, luego pueden irse a sus casas y pedir, ya sea en una aplicación móvil o en el sitio web, la ropa que desean.

La compañía está tratando de decidir si enfocar sus esfuerzos en la experiencia de su aplicación móvil o en su sitio web. ¡Te contrataron para ayudarlos a tomar las decisiones! ¡Empecemos!

Simplemente siga los pasos a continuación para analizar los datos de los clientes (son datos inventados, no se preocupe).

## Importación de librerías

Importar pandas, numpy, matplotlib,y seaborn. (Importará sklearn a medida que lo necesite.)

In [1]:

```
1 import pandas as pd
2 import numpy as np
3 import matplotlib.pyplot as plt
4 import seaborn as sns
```

## Recuperar los datos

Trabajaremos con el archivo csv de clientes de comercio electrónico de la compañía. Tiene información del Cliente, como Correo electrónico, Dirección y su color Avatar. También tiene columnas de valores numéricos:

- Avg. Session Length: Promedio de asesoramiento de estilo en la tienda.
- Time on App: Tiempo promedio dedicado a la aplicación en minutos.
- Time on Website: Tiempo promedio dedicado al sitio web en minutos.
- Length of Membership: Cuántos años el cliente ha sido miembro.

Lea en el archivo csv de clientes de comercio electrónico como un DataFrame llamado clientes.

In [2]:

```
1 clientes = pd.read_csv('Ecommerce Customers')
```

Revise las primeras filas de customers, y reviselas con los métodos info() y describe().

In [3]:

```
1 clientes.head()
```

Out[3]:

Email	Address	Avatar	Avg. Session Length	Time on App	Time on Website
mstephenson@fernandez.com	835 Frank Tunnel\nWrightmouth, MI 82180-9605	Violet	34.497268	12.655651	39.57
hduke@hotmail.com	4547 Archer Common\nDiazchester, CA 06566-8576	DarkGreen	31.926272	11.109461	37.26
pallen@yahoo.com	24645 Valerie Unions Suite 582\nCobbborough, D...	Bisque	33.000915	11.330278	37.17
riverarebecca@gmail.com	1414 David Throughway\nPort Jason, OH 22070-1220	SaddleBrown	34.305557	13.717514	36.72
mstephens@davidson-herman.com	14023 Rodriguez Passage\nPort Jacobville, PR 3...	MediumAquaMarine	33.330673	12.795189	37.53

In [4]:

```
1 clientes.describe()
```

Out[4]:

	Avg. Session Length	Time on App	Time on Website	Length of Membership	Yearly Amount Spent
count	500.000000	500.000000	500.000000	500.000000	500.000000
mean	33.053194	12.052488	37.060445	3.533462	499.314038
std	0.992563	0.994216	1.010489	0.999278	79.314782
min	29.532429	8.508152	33.913847	0.269901	256.670582
25%	32.341822	11.388153	36.349257	2.930450	445.038277
50%	33.082008	11.983231	37.069367	3.533975	498.887875
75%	33.711985	12.753850	37.716432	4.126502	549.313828
max	36.139662	15.126994	40.005182	6.922689	765.518462

In [5]:

```
1 clientes.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500 entries, 0 to 499
Data columns (total 8 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   Email            500 non-null    object  
 1   Address          500 non-null    object  
 2   Avatar            500 non-null    object  
 3   Avg. Session Length  500 non-null  float64 
 4   Time on App       500 non-null    float64 
 5   Time on Website   500 non-null    float64 
 6   Length of Membership  500 non-null  float64 
 7   Yearly Amount Spent 500 non-null    float64 
dtypes: float64(5), object(3)
memory usage: 31.4+ KB
```

## Análisis de Datos Exploratorios

¡Exploremos los datos!

Para el resto del ejercicio, solo utilizaremos los datos numéricos del archivo csv.

**Use seaborn para crear una gráfica conjunta para comparar las columnas Time on Website y Yearly Amount Spent. ¿Tiene sentido la correlación?**

In [12]:

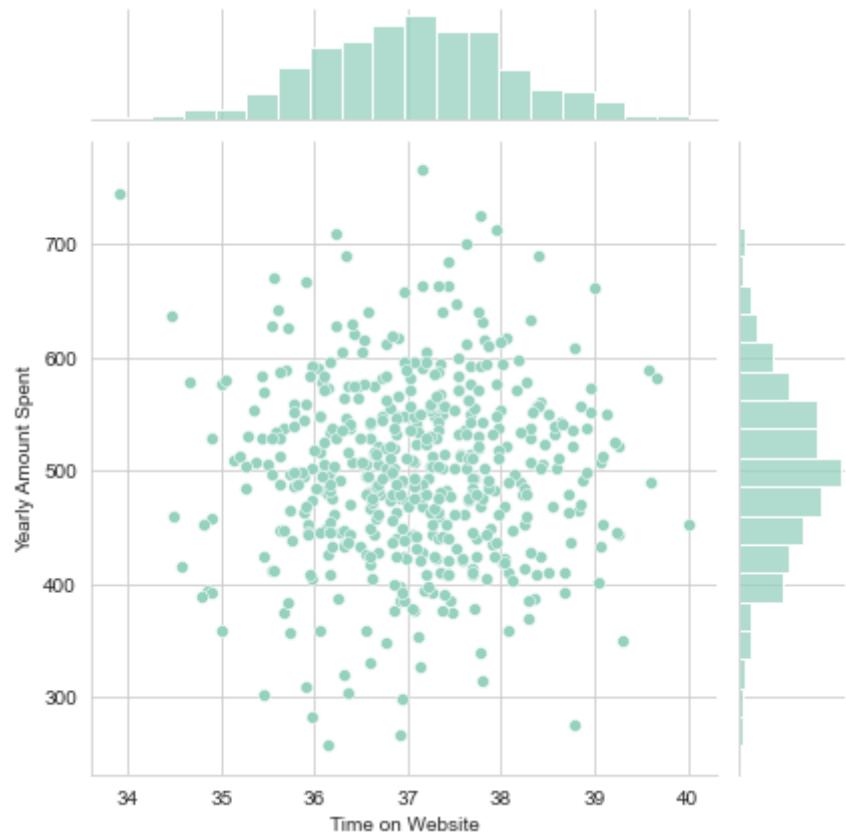
```
1 sns.set_palette("GnBu_d")
2 sns.set_style('whitegrid')
```

In [13]:

```
1 sns.jointplot(x='Time on Website',y='Yearly Amount Spent',data=clientes)
```

Out[13]:

```
<seaborn.axisgrid.JointGrid at 0x22e7dc376a0>
```

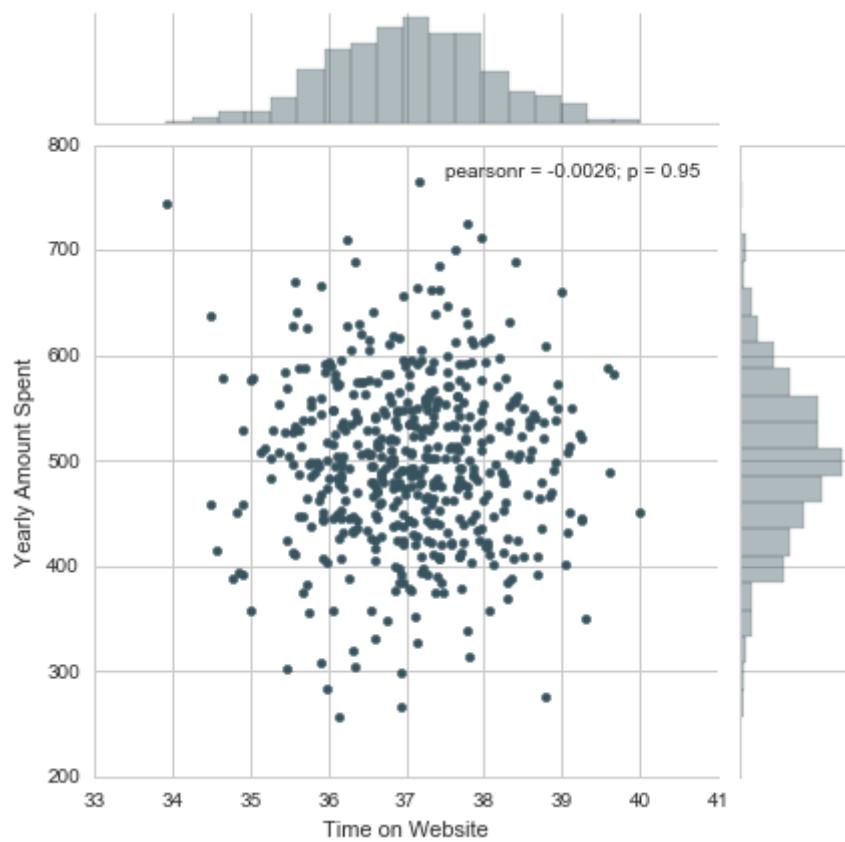


In [281]:

1

Out[281]:

<seaborn.axisgrid.JointGrid at 0x120bfcc88>



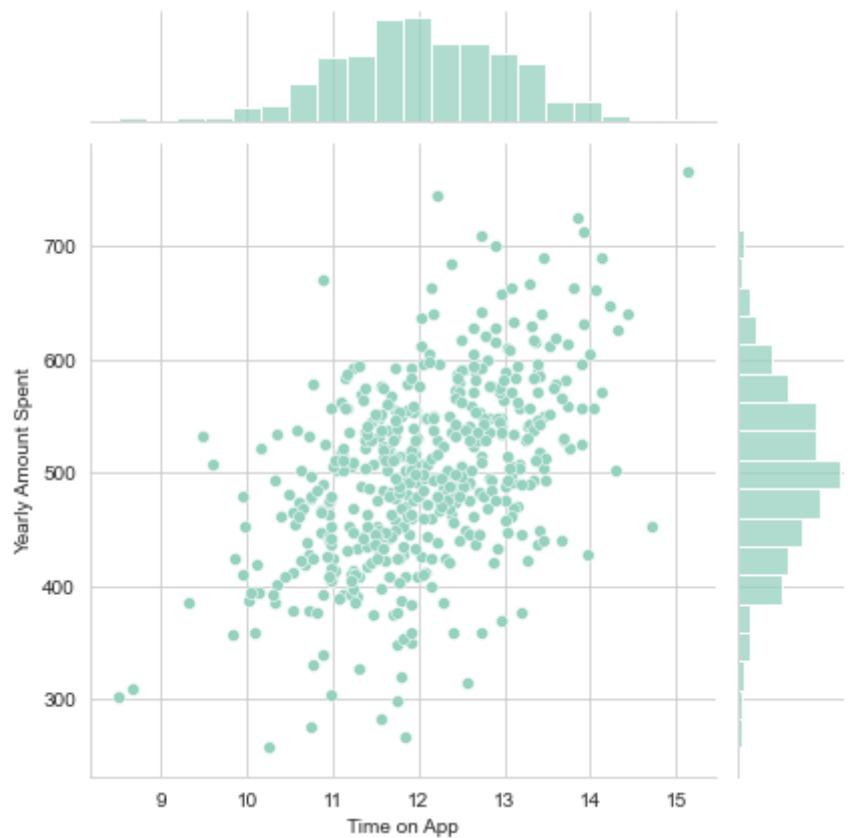
Haz lo mismo pero con la columna Time on App en su lugar.

In [14]:

```
1 sns.jointplot(x='Time on App',y='Yearly Amount Spent',data=clientes)
```

Out[14]:

<seaborn.axisgrid.JointGrid at 0x22e7dd41eb0>



In [15]:

```
1 clientes.corr()
```

Out[15]:

	Avg. Session Length	Time on App	Time on Website	Length of Membership	Yearly Amount Spent
Avg. Session Length	1.000000	-0.027826	-0.034987	0.060247	0.355088
Time on App	-0.027826	1.000000	0.082388	0.029143	0.499328
Time on Website	-0.034987	0.082388	1.000000	-0.047582	-0.002641
Length of Membership	0.060247	0.029143	-0.047582	1.000000	0.809084
Yearly Amount Spent	0.355088	0.499328	-0.002641	0.809084	1.000000

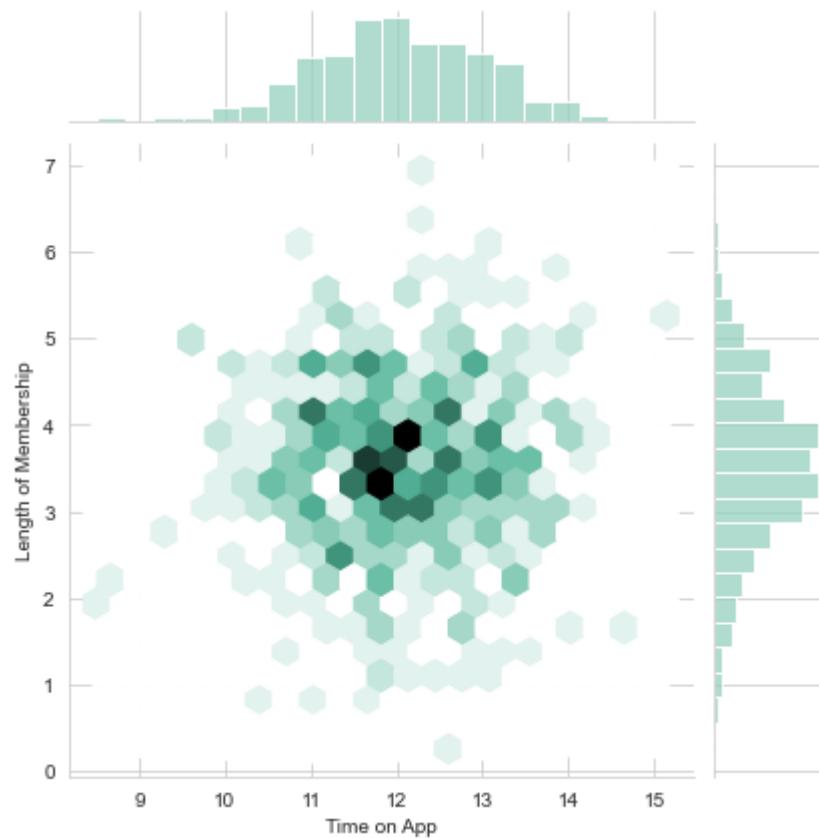
Use jointplot para crear un gráfico 2D hex bin comparando Time on App y Length of Membership.

In [16]:

```
1 sns.jointplot(x='Time on App',y='Length of Membership',kind='hex',data=clientes)
```

Out[16]:

```
<seaborn.axisgrid.JointGrid at 0x22e7df2a190>
```



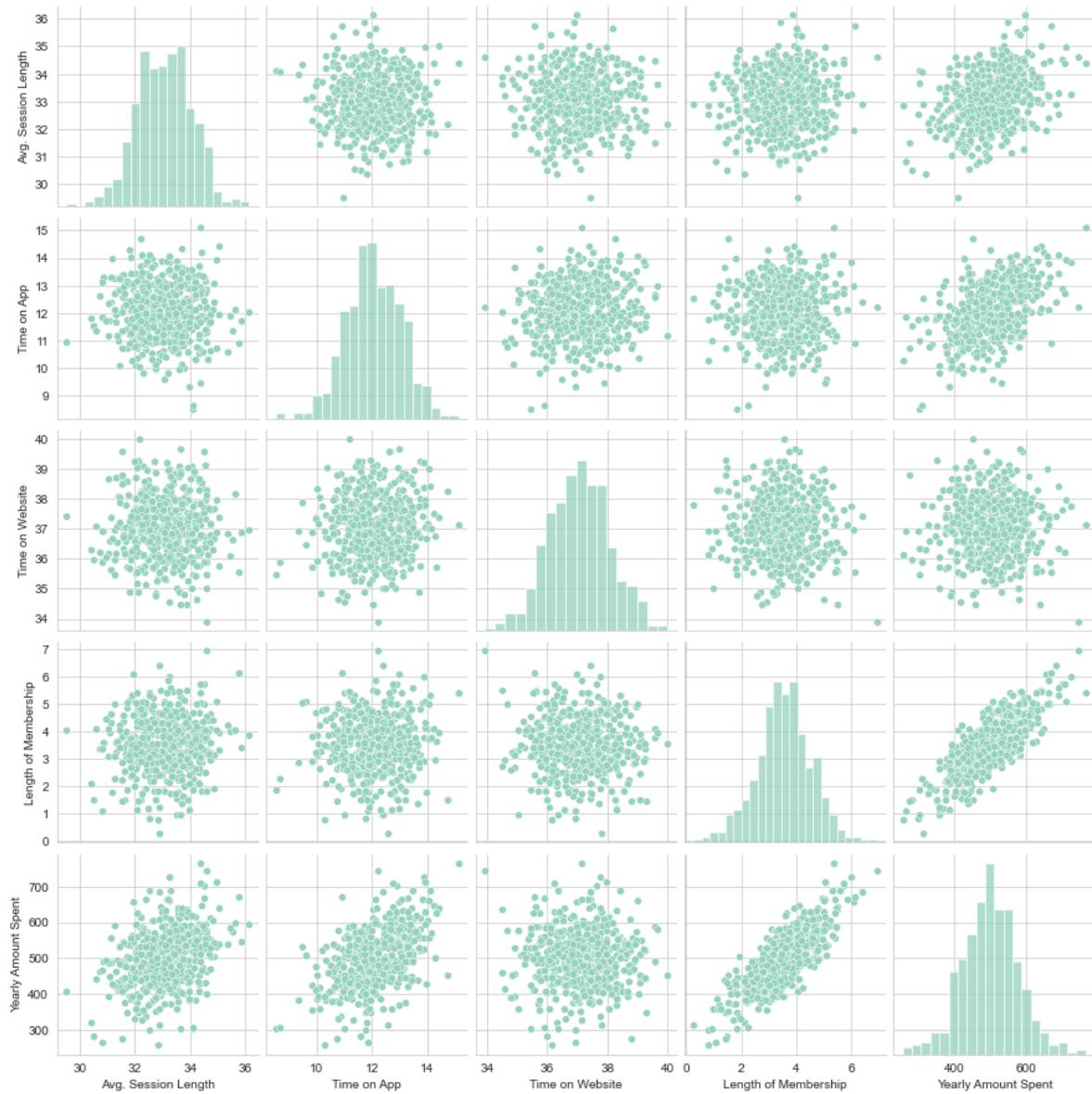
Exploraremos este tipo de relaciones en todo el conjunto de datos. Usa [pairplot](#) ([https://stanford.edu/~mwaskom/software/seaborn/tutorial/axis\\_grids.html#plotting-pairwise-relationships-with-pairgrid-and-pairplot](https://stanford.edu/~mwaskom/software/seaborn/tutorial/axis_grids.html#plotting-pairwise-relationships-with-pairgrid-and-pairplot)) para recrear la gráfica de abajo. (No te preocupes por los colores)...Exploraremos este tipo de relaciones en todo el conjunto de datos.

In [17]:

```
1 sns.pairplot(clientes)
```

Out[17]:

```
<seaborn.axisgrid.PairGrid at 0x22e7e75c2e0>
```



Basado en esta trama, ¿cuál parece ser la característica más correlacionada con la cantidad anual gastada?

In [18]:

```
1 # Antiguedad de La Membresía (Length of membership)
```

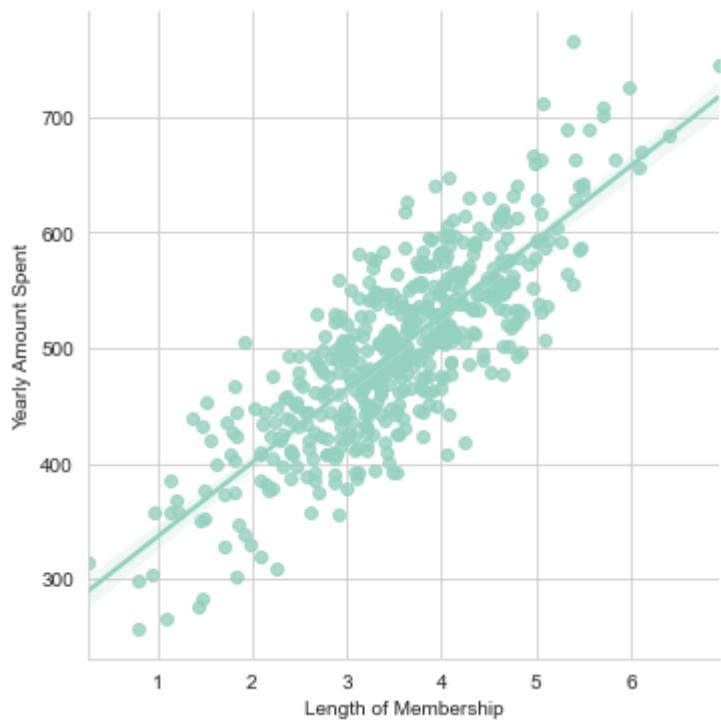
Cree un diagrama de modelo lineal (utilizando Implot de seaborn) de Yearly Amount Spent vs. Length of Membership.

In [19]:

```
1 sns.lmplot(x='Length of Membership',y='Yearly Amount Spent',data=clientes)
```

Out[19]:

```
<seaborn.axisgrid.FacetGrid at 0x22e015ac460>
```



## Datos de entrenamiento y prueba

Ahora que hemos explorado un poco los datos, sigamos adelante y dividamos los datos en conjuntos de entrenamiento y prueba. **Establezca una variable X igual a las características numéricas de los clientes y una variable y igual a la columna "Cantidad gastada anual".**

In [287]:

```
1
```

In [288]:

```
1
```

\*\* Use `model_selection.train_test_split` de `sklearn` para dividir los datos en el conjunto de entrenamiento y prueba. Establezca `test_size=0.3` y `random_state=101`\*\*

In [289]:

```
1
```

In [290]:

```
1
```

# Entrenamiento del modelo

¡Ahora es el momento de entrenar a su modelo con nuestros datos de entrenamiento!

\*\* Importar LinearRegression desde sklearn.linear\_model \*\*

In [291]:

```
1
```

Crear una instancia del modelo LinearRegression() llamado lm.

In [292]:

```
1
```

\*\* Entrenar/ajustar lm con los datos de entrenamiento.\*\*

In [293]:

```
1
```

Out[293]:

```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=1, normalize=False)
```

Imprima los coeficientes del modelo

In [294]:

```
1
```

Coefficients:

```
[ 25.98154972  38.59015875   0.19040528  61.27909654]
```

## Predicción con los datos de prueba

Ahora que hemos ajustado nuestro modelo, ¡evaluemos su rendimiento prediciendo los valores de prueba!

\*\* Use lm.predict () para predecir el conjunto X\_test de los datos.\*\*

In [295]:

```
1
```

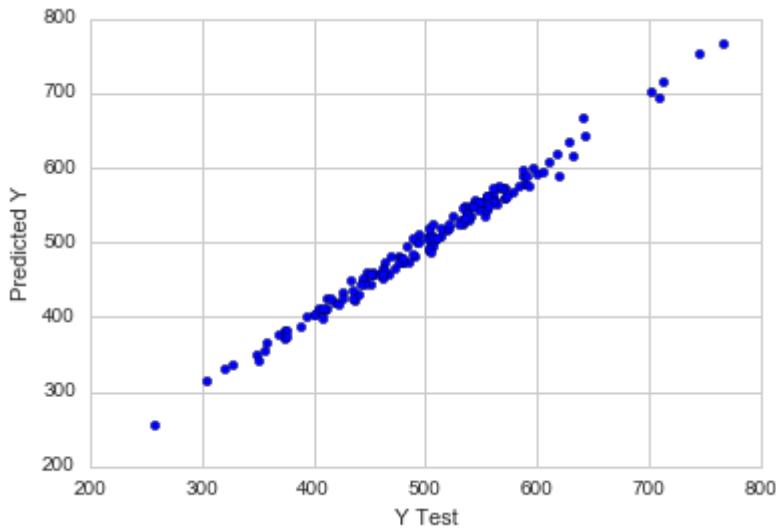
\*\* Cree un diagrama de dispersión de los valores de prueba reales frente a los valores predichos. \*\*

In [296]:

```
1
```

Out[296]:

```
<matplotlib.text.Text at 0x135546320>
```



## Evaluación del modelo

Evaluemos el rendimiento de nuestro modelo calculando la suma residual de cuadrados y la puntuación de varianza explicada ( $R^2$ ).

\*\* Calcule el error absoluto promedio, el error cuadrado promedio y la raíz del error cuadrático promedio.\*\*

In [303]:

```
1
```

```
MAE: 7.22814865343  
MSE: 79.813051651  
RMSE: 8.93381506698
```

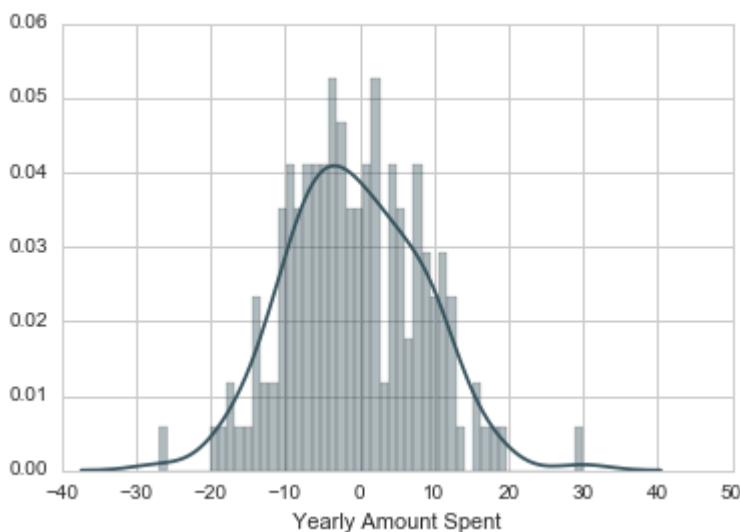
## Residuales

Deberías haber obtenido un modelo muy bueno con un buen ajuste. Exploraremos rápidamente los residuos para asegurarnos de que todo esté bien con nuestros datos.

**Trace un histograma de los residuos y asegúrese de que se vea distribuido normalmente. Utilice ya sea distplot de seaborn o simplemente plt.hist ()**

In [317]:

1



## Conclusión

Todavía queremos averiguar la respuesta a la pregunta original, ¿centramos nuestros esfuerzos en el desarrollo de aplicaciones móviles o sitios web? O tal vez eso realmente no importa, y el Tiempo de Membresía es lo que es realmente importante. Veamos si podemos interpretar los coeficientes para obtener una idea.

\*\* Recrea el dataframe de abajo. \*\*

In [298]:

1

Out[298]:

	Coeffecient
Avg. Session Length	25.981550
Time on App	38.590159
Time on Website	0.190405
Length of Membership	61.279097

\*\* ¿Cómo puedes interpretar estos coeficientes? \*\*

Type *Markdown* and *LaTeX*:  $\alpha^2$

¿Crees que la empresa debería centrarse más en su aplicación móvil o en su sitio web?

*La respuesta aquí*

**¡Excelente trabajo!**



# Regresión logística

# Regresión logística

- No todas las etiquetas son continuas, a veces es necesario predecir categorías, esto se conoce como clasificación.
- La regresión logística es una de las formas básicas para realizar la clasificación (no se confunda por la palabra "regresión")

# Lectura sugerida

Secciones 4-4.3 de  
**Introduction to Statistical Learning**  
Por Gareth James

# Regresión logística

- Si desea comprender completamente algunos de los conceptos detrás de los métodos de evaluación y las métricas detrás de la clasificación, ¡la lectura es muy recomendable!

# Importante

- Queremos aprender sobre Regresión logística como un método para la clasificación.
- Algunos ejemplos de problemas de clasificación:
  - Spam versus correos electrónicos legítimos
  - Préstamo Predeterminado (sí / no)
  - Diagnóstico de la enfermedad
- Todos los anteriores fueron ejemplos de clasificación binaria

# Importante

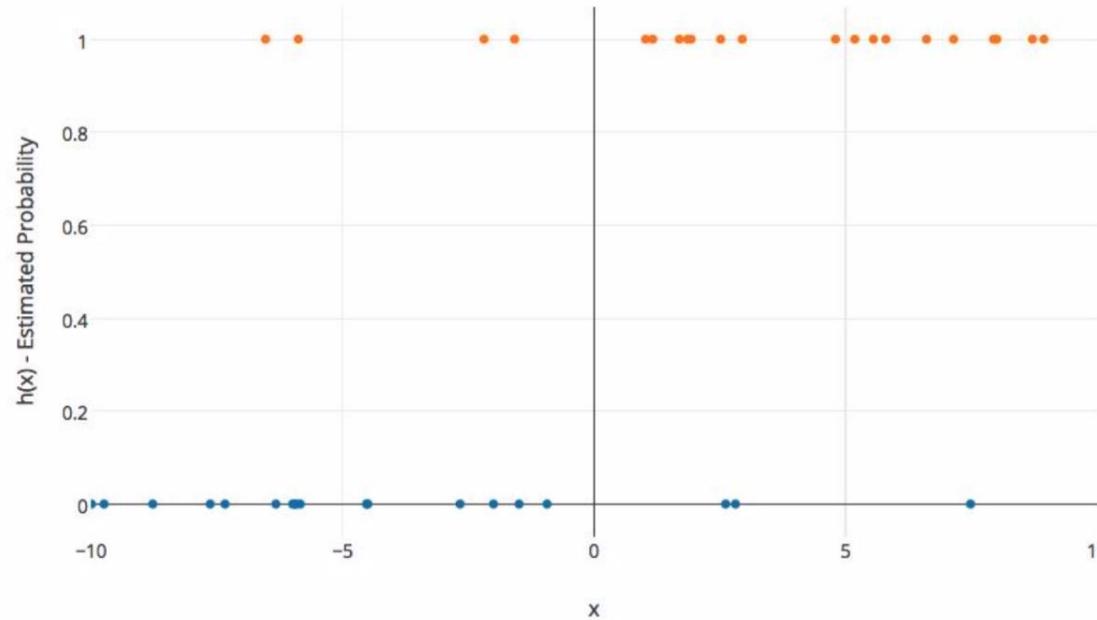
- Hasta ahora solo hemos visto problemas de regresión en los que intentamos predecir un valor continuo.
- Aunque el nombre puede ser confuso al principio, la regresión logística nos permite resolver problemas de clasificación, donde estamos tratando de predecir categorías discretas.

# Importante

- La convención para la clasificación binaria es tener dos clases 0 y 1.
- Vayamos a través de la idea básica para la regresión logística.
- También explicaremos por qué tiene el término regresión, ¡aunque se utilice para la clasificación!

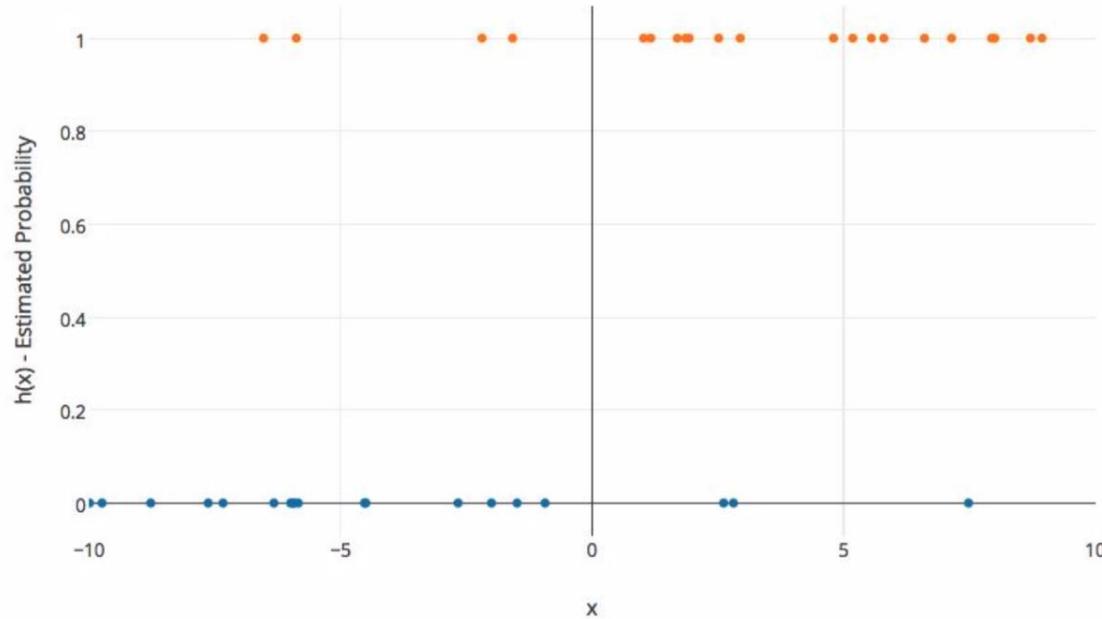
# Background

- Imagina que trazamos algunos datos categóricos contra una característica.



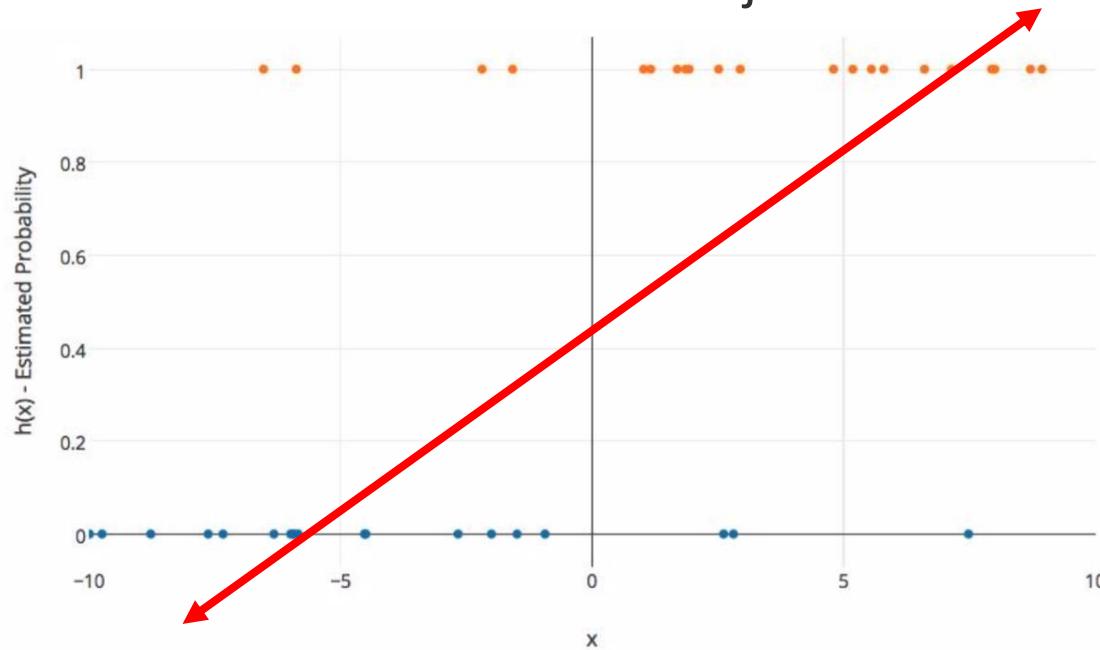
# Background

- El eje X representa un valor de característica y el eje Y representa la probabilidad de pertenecer a la clase 1.



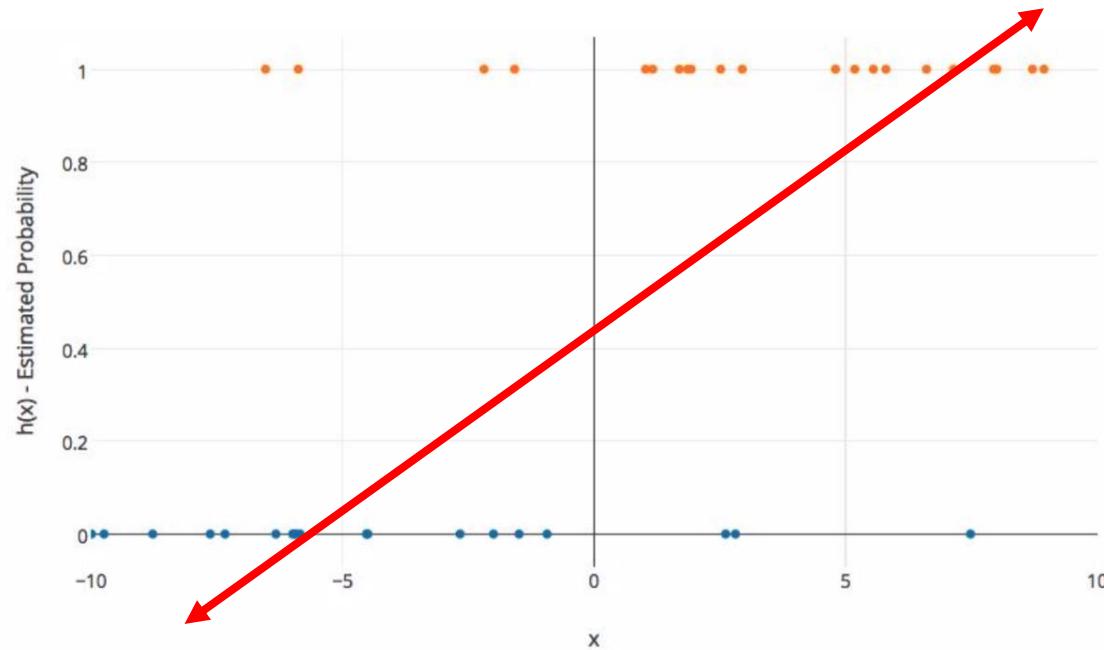
# Background

- No podemos usar un modelo de regresión lineal normal en grupos binarios. No conducirá a un buen ajuste:



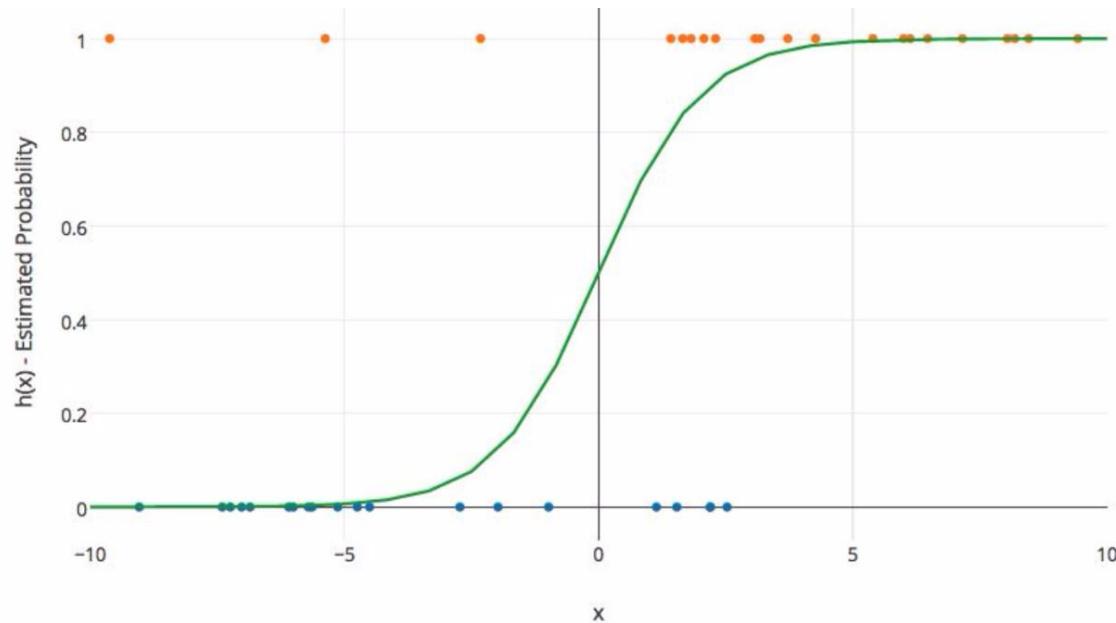
# Background

- Necesitamos una función que se ajuste a los datos categóricos binarios!



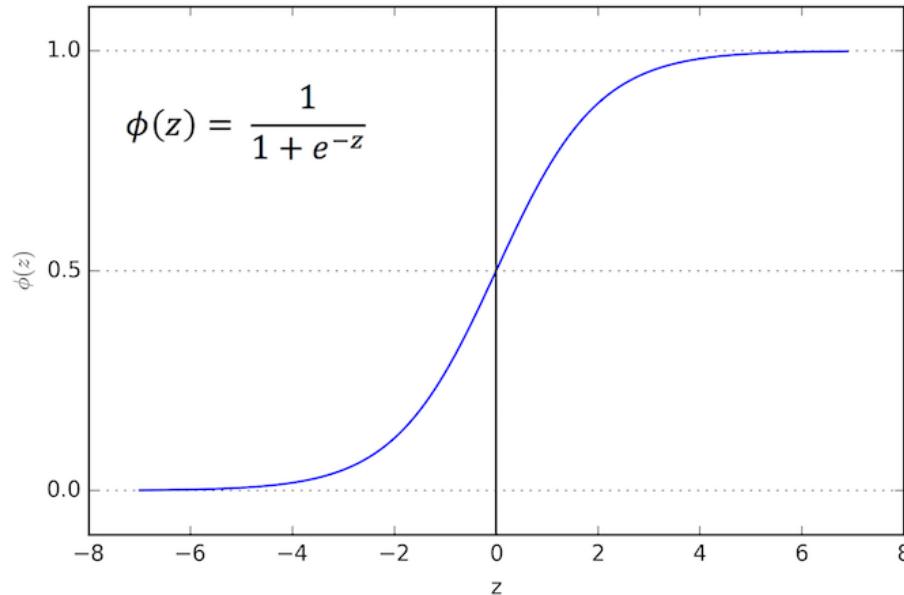
# Background

- Sería genial si pudiéramos encontrar una función con este tipo de comportamiento:



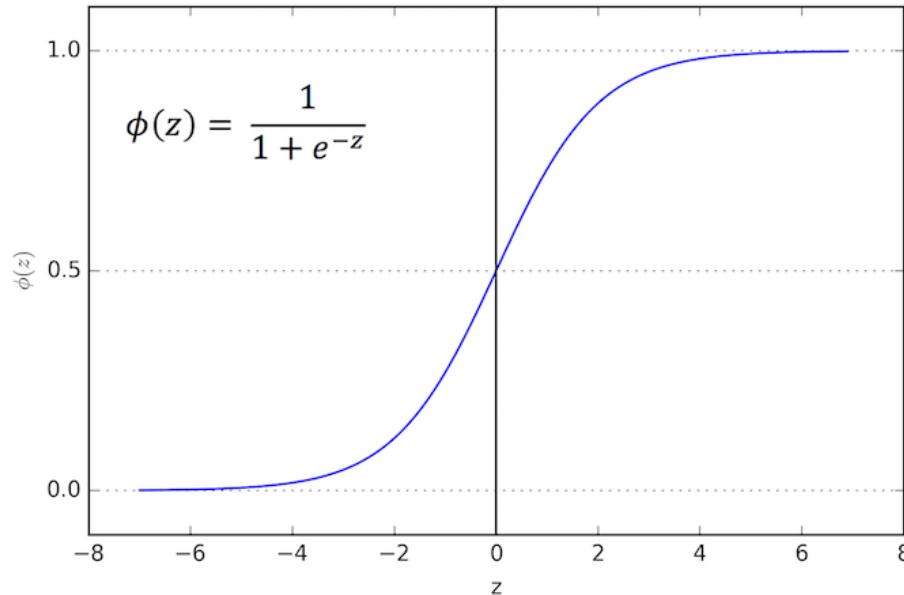
# Función sigmoidea

- La función sigmoide (también conocida como logística) toma cualquier valor y genera una salida entre 0 y 1.



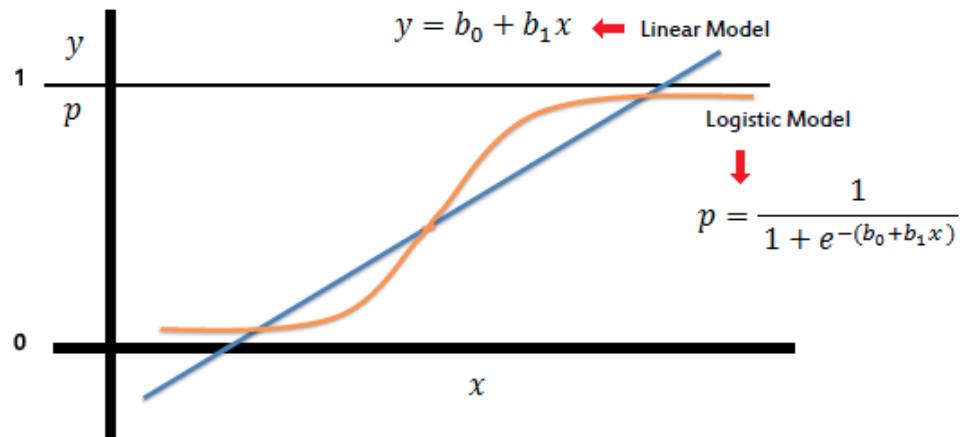
# Función sigmoidea

- Esto significa que podemos tomar nuestra Solución de Regresión Lineal y colocarla en la Función Sigmoide.



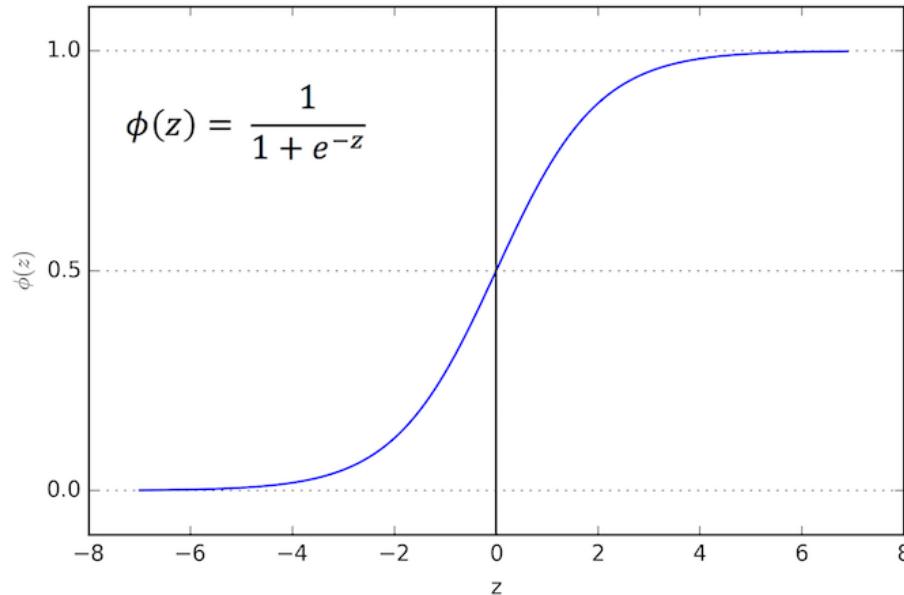
# Función sigmoidea

- Esto significa que podemos tomar nuestra Solución de Regresión Lineal y colocarla en la Función Sigmoide.



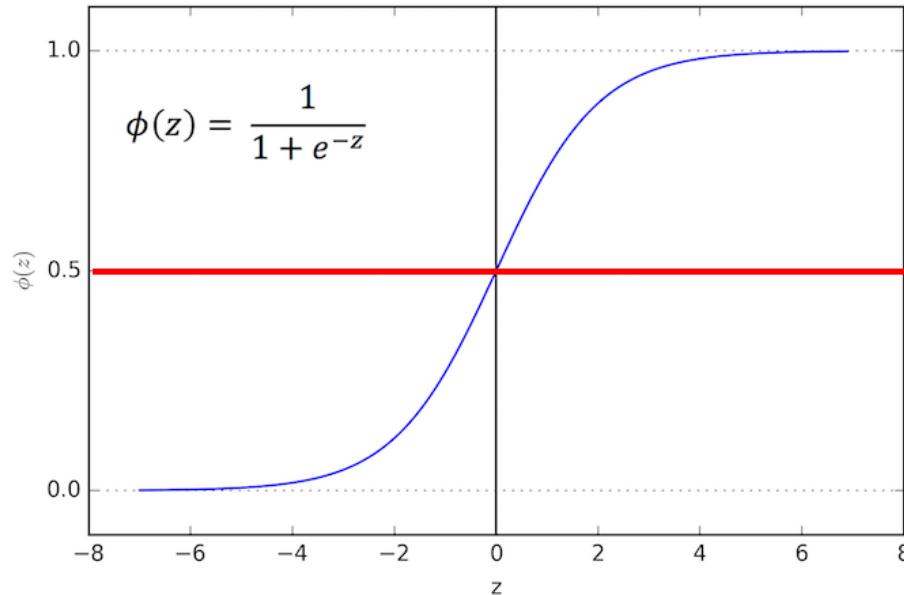
# Función sigmoidea

- Esto da como resultado una probabilidad de 0 a 1 de pertenencia a la clase 1.



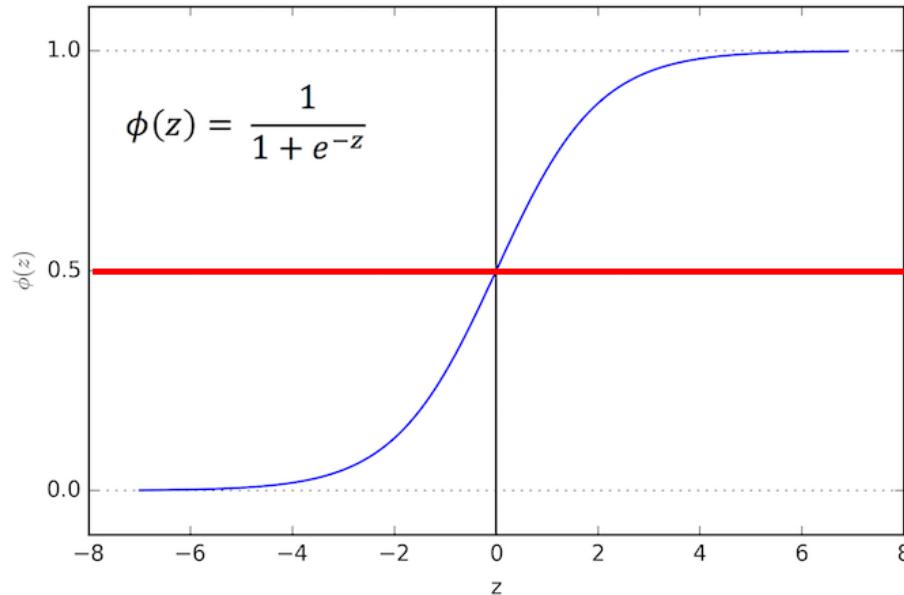
# Función sigmoidea

- Podemos establecer un punto de corte en 0.5, cualquier cosa debajo de esto resulta en la clase 0, cualquier cosa arriba es la clase 1.



# Repaso

- Usamos la función logística para generar un valor que va de 0 a 1. En función de esta probabilidad, asignamos una clase.



# Evaluación del modelo

- Despues de entrenar un modelo de regresión logística con algunos datos de entrenamiento, evaluará el rendimiento de su modelo con algunos datos de prueba.
- Puedes usar una matriz de confusión para evaluar los modelos de clasificación.

# Matriz de confusión

		predicted condition (condición predicha)	
total population		prediction positive (predicción positiva)	prediction negative (predicción negativa)
true condition (condición verdadera)	condition positive	Verdadero Positivo <b>True Positive (TP)</b>	Falso Negativo <b>False Negative (FN)</b> (type II error) <b>(Error Tipo II)</b>
	condition negative (condición negativa)	Falso Positivo <b>False Positive (FP)</b> (Type I error) <b>(Error Tipo I)</b>	Verdadero Negativo <b>True Negative (TN)</b>

# Matriz de confusión

		predicted condition		
		prediction positive	prediction negative	Prevalence $= \frac{\sum \text{condition positive}}{\sum \text{total population}}$
true condition	condition positive	True Positive (TP)	False Negative (FN) (type II error)	True Positive Rate (TPR), Sensitivity, Recall, Probability of Detection $= \frac{\sum \text{TP}}{\sum \text{condition positive}}$
	condition negative	False Positive (FP) (Type I error)	True Negative (TN)	False Positive Rate (FPR), Fall-out, Probability of False Alarm $= \frac{\sum \text{FP}}{\sum \text{condition negative}}$
Accuracy $= \frac{\sum \text{TP} + \sum \text{TN}}{\sum \text{total population}}$		Positive Predictive Value (PPV), Precision $= \frac{\sum \text{TP}}{\sum \text{prediction positive}}$	False Omission Rate (FOR) $= \frac{\sum \text{FN}}{\sum \text{prediction negative}}$	Positive Likelihood Ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$
$= \frac{\sum \text{TP} + \sum \text{TN}}{\sum \text{total population}}$		False Discovery Rate (FDR) $= \frac{\sum \text{FP}}{\sum \text{prediction positive}}$	Negative Predictive Value (NPV) $= \frac{\sum \text{TN}}{\sum \text{prediction negative}}$	Negative Likelihood Ratio (LR-) $= \frac{\text{FNR}}{\text{TNR}}$

# Evaluación del modelo

- El punto principal a recordar con la matriz de confusión y las diversas métricas calculadas es que todas son fundamentalmente formas de comparar los valores predichos con los valores reales.
- ¡Lo que constituye una métrica "buena" dependerá realmente de la situación específica!

# Evaluación del modelo

- Podemos utilizar una matriz de confusión para evaluar nuestro modelo.
- Por ejemplo, imagine pruebas para detectar enfermedades.

n=165	Predicted:	
	NO	YES
Actual: NO	50	10
Actual: YES	5	100

Ejemplo: prueba de presencia de enfermedad  
NO = prueba negativa = falso = 0  
Sí = prueba positiva = Verdadero = 1

# Evaluación del modelo

n=165	Predicted:	
	NO	YES
Actual: NO	TN = 50	FP = 10
Actual: YES	FN = 5	TP = 100
	55	110

Terminología básica:

- Verdaderos positivos (TP)
- Negativos Verdaderos (TN)
- Falsos positivos (FP)
- Falsos negativos (FN)

# Matriz de confusión

n=165	Predicted:	
	NO	YES
Actual: NO	TN = 50	FP = 10
Actual: YES	FN = 5	TP = 100
	55	110

## Exactitud:

- En general, ¿con qué frecuencia es correcto?
- $(TP + TN) / \text{total} = 150/165 = 0.91$

# Matriz de confusión

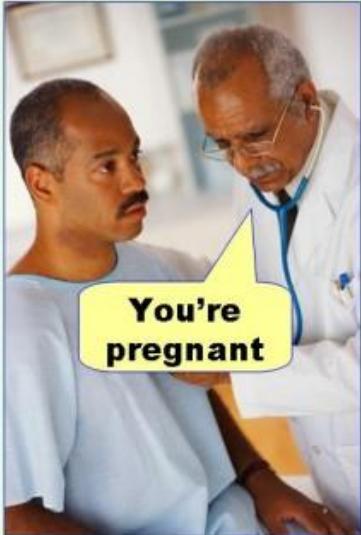
n=165	Predicted:		
	NO	YES	
Actual:			
NO	TN = 50	FP = 10	60
YES	FN = 5	TP = 100	105
	55	110	

Tasa de clasificación errónea (Tasa de error):

- En general, ¿con qué frecuencia está mal?
- $(FP + FN) / \text{total} = 15/165 = 0.09$

# Matriz de confusión

**Type I error**  
(false positive)



**Type II error**  
(false negative)



# Evaluación del modelo

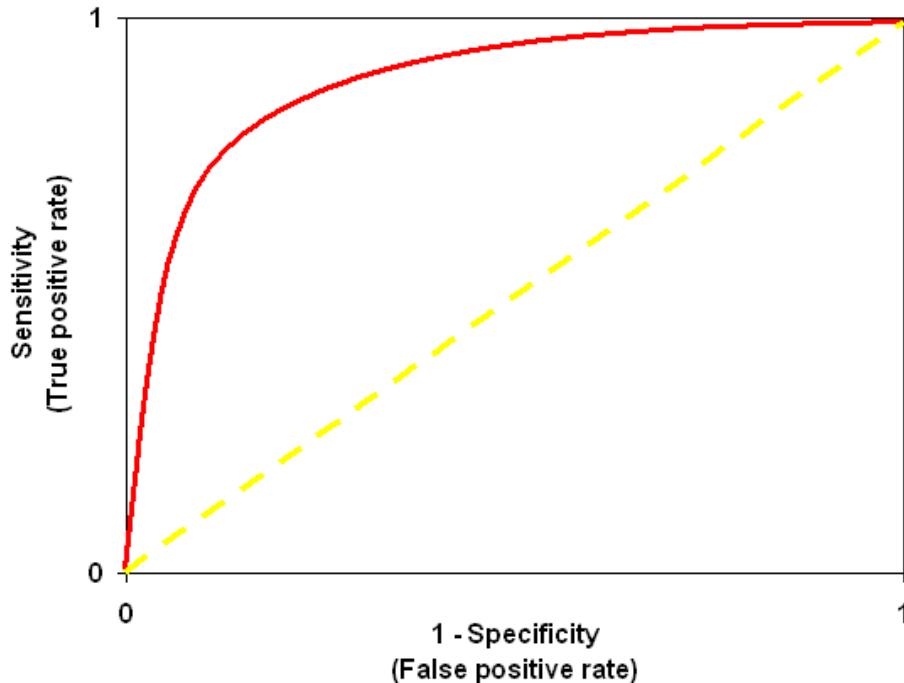
- ¿Todavía confundido con la matriz de confusión?
- ¡No hay problema! Echa un vistazo a la página de Wikipedia para ver si tiene un diagrama realmente bueno con todas las fórmulas para todas las métricas.
- A lo largo del curso, por lo general solo imprimimos métricas (por ejemplo, precisión).

# Evaluación del modelo

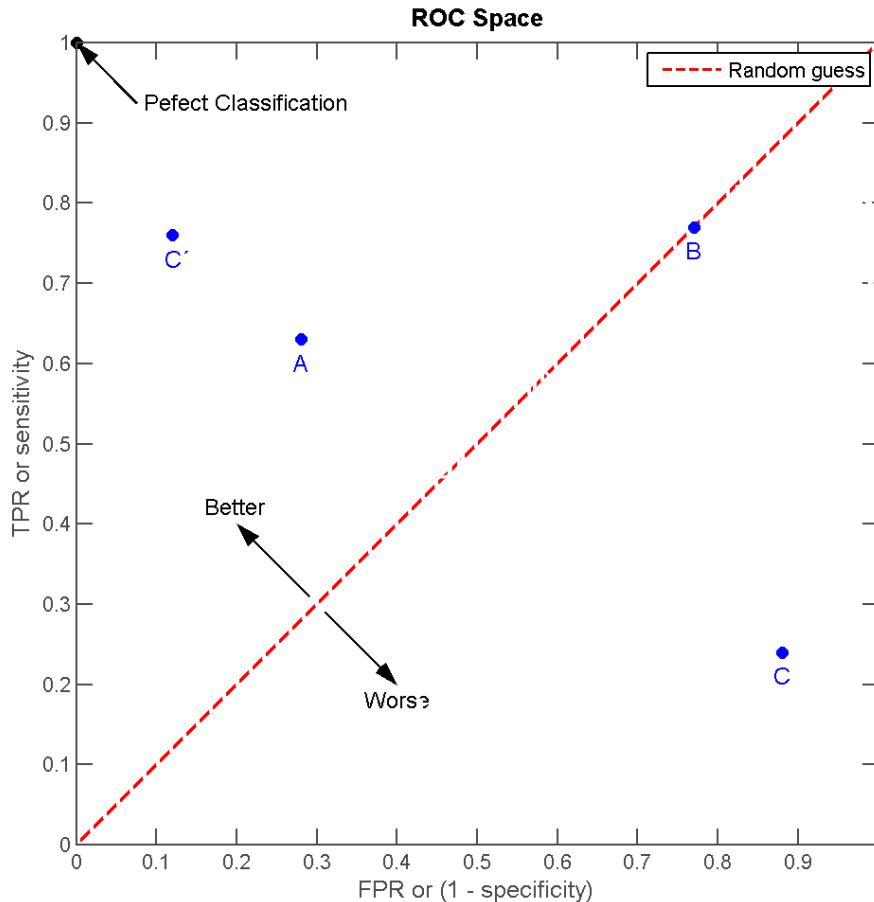
- La clasificación binaria tiene algunas de sus propias métricas de clasificación especial.
- Estos incluyen visualizaciones de métricas de la matriz de confusión.
- La curva de la curva del operador receptor (ROC) se desarrolló durante la Segunda Guerra Mundial para ayudar a analizar los datos del radar.

# Evaluación del Modelo

- La curva ROC:



# Evaluación del Modelo



# Evaluación del Modelo

- Una discusión completa de la curva ROC está más allá del alcance de este curso, pero la lectura sugerida entra en mucho más detalle.
- Por ahora, solo necesita saber que el área debajo de la curva es una métrica de qué tan bien un modelo se ajusta a los datos.

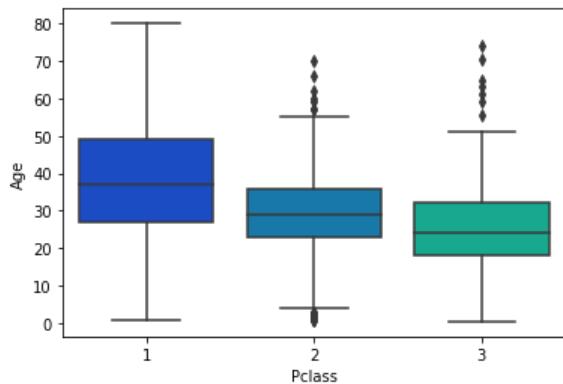
## Calculo de los Cuartiles a las edades por clase en el problema del Titanic

```
In [1]: 1 import pandas as pd
2 import numpy as np
3 import matplotlib.pyplot as plt
4 import seaborn as sns
5 %matplotlib inline
```

```
In [2]: 1 datos = pd.read_csv('titanic_train.csv')
```

```
In [3]: 1 sns.boxplot(x='Pclass',y='Age',data=datos,palette='winter')
```

```
Out[3]: <AxesSubplot:xlabel='Pclass', ylabel='Age'>
```



```
In [4]: 1 pasaj_c1 = datos[(datos['Pclass']==3) & (datos['Age'].isnull()==False)]
2 pasaj_c2 = datos[datos['Pclass']==2]
3 pasaj_c3 = datos[datos['Pclass']==3]
```

```
In [5]: 1 edades_c1 = pasaj_c1['Age']
2 edades_c2 = pasaj_c2['Age'].dropna()
3 edades_c3 = pasaj_c3['Age'].dropna()
```

```
In [6]: 1 Cuartiles_c1=pd.DataFrame(edades_c1.quantile([0.25,0.50,0.75]))
2 Cuartiles_c1.columns = ["Edad"]
3 Cuartiles_c1.index = ["Q1","Q2","Q3"]
4 Cuartiles_c2=pd.DataFrame(edades_c2.quantile([0.25,0.50,0.75]))
5 Cuartiles_c2.columns = ["Edad"]
6 Cuartiles_c2.index = ["Q1","Q2","Q3"]
7 Cuartiles_c3=pd.DataFrame(edades_c3.quantile([0.25,0.50,0.75]))
8 Cuartiles_c3.columns = ["Edad"]
9 Cuartiles_c3.index = ["Q1","Q2","Q3"]
```

```
In [7]: 1 print("Primera Clase")
2 print(Cuartiles_c1)
3 print("Segunda Clase")
4 print(Cuartiles_c2)
5 print("Tercera Clase")
6 print(Cuartiles_c3)
```

Primera Clase

    Edad

    Q1 18.0

    Q2 24.0

    Q3 32.0

Segunda Clase

    Edad

    Q1 23.0

    Q2 29.0

    Q3 36.0

Tercera Clase

    Edad

    Q1 18.0

    Q2 24.0

    Q3 32.0

```
In [8]: 1 print("Promedio edad 1ra Clase: ",np.round(edades_c1.mean(),2))
2 print("Promedio edad 2da Clase: ",np.round(edades_c2.mean(),2))
3 print("Promedio edad 3ra clase: ",np.round(edades_c3.mean(),2))
```

```
Promedio edad 1ra Clase: 25.14
Promedio edad 2da Clase: 29.88
Promedio edad 3ra clase: 25.14
```

```
In [9]: 1 psj_por_clase = datos.groupby(['Pclass'])
```

```
In [11]: 1 psj_por_clase['Age'].quantile([0.25,0.50,0.75])
```

Out[11]: Pclass

Pclass	0.25	0.50	0.75
1	27.0	37.0	49.0
2	23.0	29.0	36.0
3	18.0	24.0	32.0

Name: Age, dtype: float64

# Interpretación de la Matriz de Confusión de sklearn

---

UNMSM – Inteligencia Artificial - Juan Gamarra Moreno

# Reporte de clasificación de sklearn

---

- El reporte de clasificación se utiliza para medir la calidad de las predicciones de un algoritmo de clasificación.
- Cuántas predicciones son verdaderas y cuántas falsas.
- Más específicamente, los verdaderos positivos (TP), falsos positivos (FP), verdaderos negativos (TN) y falsos negativos (FN) se utilizan para predecir las métricas de un informe de clasificación.

# Reporte de clasificación de sklearn

El código usado para generar el reporte es:

```
from sklearn.metrics import classification_report  
y_true = [0, 1, 2, 2, 2]  
y_pred = [0, 0, 2, 2, 1]  
target_names = ['class 0', 'class 1', 'class 2']  
print(classification_report(y_true, y_pred, target_names=target_names))
```

# Reporte de clasificación de sklearn

Salida:

	precision	recall	f1-score	support
class 0	0.50	1.00	0.67	1
class 1	0.00	0.00	0.00	1
class 2	1.00	0.67	0.80	3
avg / total	0.70	0.60	0.61	5

# Precisión

---

La precisión es la capacidad de un clasificador de no etiquetar una instancia como positiva que en realidad es negativa. Para cada clase, se define como la relación entre verdaderos positivos y la suma de verdaderos y falsos positivos.

- TP: verdaderos positivos
- FP - Falsos positivos

Precisión: precisión de las predicciones positivas.

- $\text{Precisión} = \text{TP} / (\text{TP} + \text{FP})$

# Recall (Recuerdo): ¿Qué porcentaje de casos positivos captó?

---

Recall es la capacidad de un clasificador de encontrar todas las instancias positivas. Para cada clase, se define como la proporción de verdaderos positivos con la suma de verdaderos positivos y falsos negativos.

- FN - Falsos negativos

Recall: Fracción de positivos que se identificaron correctamente.

- $\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$

# Puntuación F1: ¿Qué porcentaje de predicciones positivas fueron correctas?

---

La puntuación F1 es una media armónica ponderada de precisión y recuerdo de manera que la mejor puntuación es 1,0 y la peor es 0,0. En términos generales, los puntajes F1 son más bajos que las medidas de precisión, ya que incorporan precisión y recuerdo en su cálculo. Como regla general, el promedio ponderado de F1 debe usarse para comparar modelos de clasificación, no la precisión global.

$$\bullet \text{ F1 score} = 2 * (\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision})$$

¿Cuándo es más  
importante la precisión que  
el recall?

---

UNMSM – UPG-FISI-Machine Learning y Big Data

# Nemónico para entenderlo

---

"PREcision is to PREGnancy tests as reCALL is to CALL center"

- Con una prueba de embarazo, el fabricante de la prueba debe asegurarse de que un resultado positivo signifique que la mujer está realmente embarazada. Las personas podrían reaccionar a una prueba positiva casándose repentinamente o comprando una casa (si muchos consumidores obtuvieran falsos positivos y sufrieran enormes costos sin motivo, el fabricante de la prueba se quedaría sin clientes).

# Nemónico para entenderlo (Continua)

---

- Ahora imagine un centro de llamadas para reclamos de seguros. La mayoría de los reclamos fraudulentos se realizan por teléfono los lunes, después de que los estafadores se conecten con los colaboradores y elaboren sus historias inventadas ("digamos que robaron el automóvil") durante el fin de semana. ¿Qué es lo mejor que puede hacer una compañía de seguros los lunes? Tal vez deberían favorecer el recall sobre la precisión. Es mucho mejor marcar más reclamos como positivos (probable fraude) para una mayor investigación que pasar por alto algunos de los fraudes y pagar efectivo que nunca debería haberse pagado. Un falso positivo (marcado para un escrutinio adicional como posible fraude, pero la pérdida del cliente fue real) probablemente se puede aclarar asignando un ajustador experimentado, que puede insistir en un informe policial, solicitar un video de seguridad del edificio, etc. Un falso negativo (aceptar el reclamo falso de un defraudador y el pago en efectivo) es una pérdida pura para la compañía de seguros y fomenta más fraudes.

# "Precision" más importante que "Recall"

---

- Imagine que queremos asegurarnos de que nuestro bloqueador de sitios web para nuestro hijo solo permita que se muestren sitios web "seguros". En este caso, un sitio web "seguro" es la clase positiva. Aquí, queremos que el bloqueador esté absolutamente seguro de que el sitio web es seguro, incluso si se prevé que algunos sitios web seguros sean parte de la clase negativa o insegura y, en consecuencia, se bloquen. Es decir, queremos una alta precisión a expensas del recall.

# "Precision" más importante que "Recall"

		Predicción	
		Seg	No Seg
Verdad	Seg	180	120
	No Seg	10	290
		190	410

$$\text{Prec(Seg)} = \frac{180}{190} = 0.95$$

$$\text{Recall(Seg)} = \frac{180}{300} = 0.60$$

# "Precision" más importante que "Recall"

		Predicción	
		Seg	No Seg
Verdad	Seg	180	120
	No Seg	0	300
		180	420

$$\text{Prec(Seg)} = \frac{180}{180} = 1.00$$

$$\text{Recall(Seg)} = \frac{180}{300} = 0.60$$

# "Recall" más importante que "Precision"

---

- En el caso de la seguridad aeroportuaria, donde un riesgo de seguridad es la clase positiva, queremos asegurarnos de que se investigue cada riesgo potencial de seguridad. En este caso, tendremos un alto recall a expensas de la precisión (se investigarán muchas bolsas en las que no hay riesgos de seguridad).

# "Recall" más importante que "Precision"

		Predicciones		
		Riesgo	No Riesgo	
Real	Riesgo	290	10	300
	No Riesgo	200	100	300
		490	110	600

$\text{prec}(\text{riesgo}) = \frac{290}{490} = 0.59$

$\text{recall}(\text{riesgo}) = \frac{290}{300} = 0.96$

# "Recall" más importante que "Precision"

		Predicciones		
		Riesgo	No Riesgo	
Real	Riesgo	300	0	300
	No Riesgo	200	100	300
		500	100	600

$$\text{prec(riesgo)} = \frac{300}{500} = 0.60$$

$$\text{recall(riesgo)} = \frac{300}{300} = 1.00$$

# Ejemplo del Centro de Llamadas

---

- Se tienen miles de clientes gratuitos registrándose en nuestro sitio web cada semana. El equipo del centro de llamadas quiere llamarlos a todos, pero es imposible, por lo que me piden que seleccione a los potenciales compradores. No nos importa llamar a un tipo que no va a comprar (así que la precisión no es importante) pero para nosotros es muy importante que los potenciales compradores estén siempre en mi selección, para que no se vayan sin comprar. Eso significa que mi modelo debe tener un alto recuerdo, sin importar demasiado la precisión.

`y_verdadero = [0, 1, 2, 2, 2]`

`y_predicho = [0, 0, 2, 2, 1]`

$c\emptyset$  = clase  $\emptyset$   
 $c1$  = clase 1  
 $c2$  = clase 2

		Predicciones		
		$c\emptyset$	$c1$	$c2$
Verdaderos	$c\emptyset$	[1, 0, 0]	= 1	
	$c1$	[1, 0, 0]	= 1	
	$c2$	[0, 1, 2]	= 3	
		2	1	2
				$\sqrt{5}$

Predichos				precision	recall	f1-score	support
Verd.	C0	C1	C2				
C0	[1, 0, 0]	1		0.50	1.00	0.67	1
	[1, 0, 0]	1		0.00	0.00	0.00	1
	[0, 1, 0]		2	1.00	0.67	0.80	3
	2	1	2				5
	15						5
accuracy				0.50	0.56	0.60	
macro avg				0.70	0.60	0.49	
weighted avg						0.61	

precision(C0):

$$\frac{1}{2} = 0.5$$

precision(C1):

$$\frac{0}{1} = 0$$

precision(C2):

$$\frac{2}{2} = 1.0$$

precision promedio:

$$(0.5 + 0 + 1.0) / 3 = 0.5$$

precision ponderada

$$(0.5 \times 1 + 0 \times 1 + 1 \times 3) / 5 = 0.70$$

recall (sensibilidad)

$$\text{recall}(C0) = \frac{1}{1} = 1$$

$$\text{recall}(C1) = \frac{0}{1} = 0$$

$$\text{recall}(C2) = \frac{2}{3} = 0.67$$

recall promedio:

$$(1 + 0 + 0.67) / 3 = 0.56$$

recall ponderado

$$(1 \times 1 + 0 \times 1 + 0.67 \times 3) / 5 = 0.60$$

$$f1(C0) = 2(1 \times 0.5) / (1+0.5) = 0.67$$

$$f1(C1) = 2(0 \times 0) / (0+0) \approx 0.00$$

$$f1(C2) = 2(2/3 \times 1) / (2/3 + 1) = 0.80$$

$$f1 \text{ promedio} = (0.67 + 0 + 0.80) / 3 = 0.49$$

$$f1 \text{ ponderado} = (0.67 \times 1 + 0 \times 1 + 0.80 \times 3) / 5 = 0.61$$

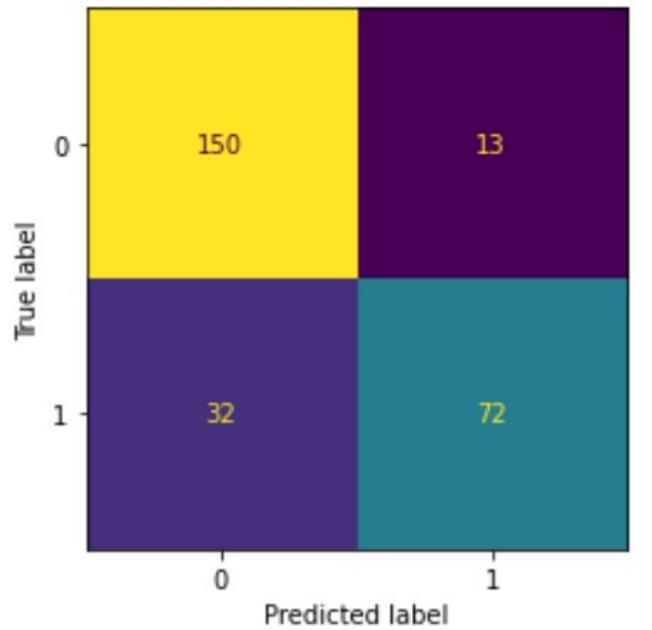
Accuracy (Exactitud)

$$acc = \frac{1 + 0 + 2}{5} = \frac{3}{5} = 0.60$$

## Predicciones

	0	1	
Verdad	0 [150, 13]	163	
	1	32, 72	104
	182	85	<u>267</u>

	precision	recall	f1-score	support
0	0.82	0.92	0.87	163
1	0.85	0.69	0.76	104
accuracy			0.83	267
macro avg	0.84	0.81	0.82	267
weighted avg	0.83	0.83	0.83	267



0 : No sobrevive

1 : Sobre vive

$$\text{precisión (no sobrevive)} = 150/182$$

$$\text{precisión (sobrevive)} = 72/85$$

$$\text{recall (no sobrevive)} = 150/163$$

$$\text{recall (sobrevive)} = 72/104$$

$$f1(\text{no sobrevive}) = 2(0.82 \times 0.92) / (0.82 + 0.92)$$

$$f1(\text{sobrevive}) = 2(0.85 \times 0.69) / (0.85 + 0.69)$$

accuracy =  $\frac{150 + 72}{267}$

# Introducción a K Vecinos más cercanos

K Nearest Neighbors (KNN)

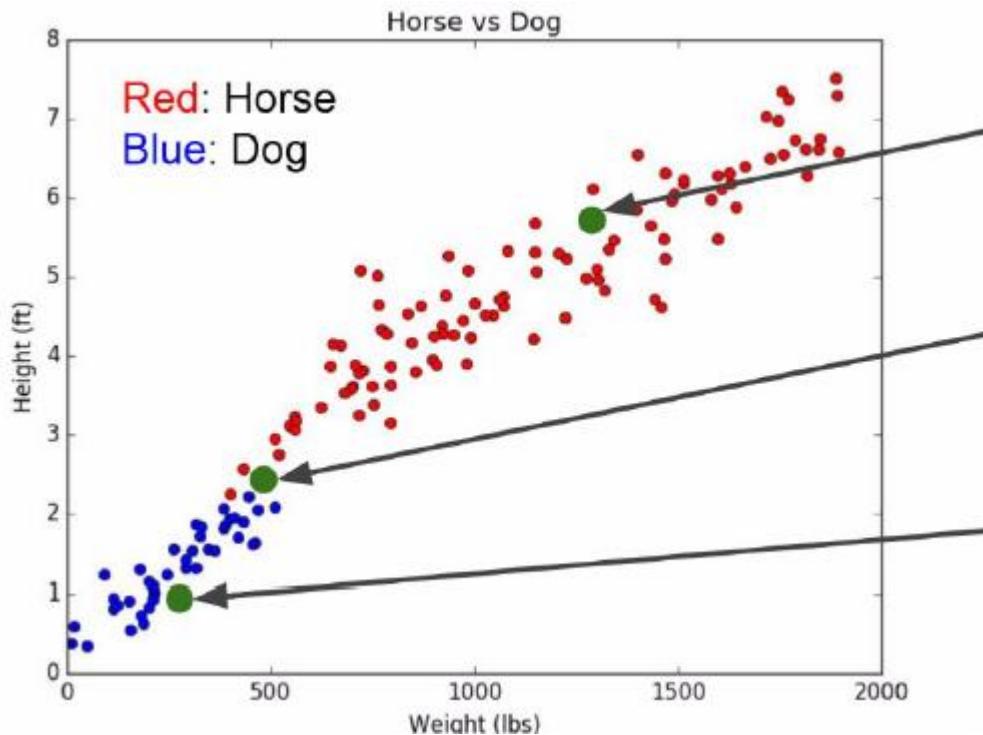
# Lectura Sugerida

Capítulo 4 de  
**Introduction to Statistical Learning**  
Gareth James

# KNN

- K Nearest Neighbors es un algoritmo de clasificación que opera sobre un principio muy simple.
- ¡Se muestra mejor a través de un ejemplo! Imagina que tenemos algunos datos imaginarios sobre perros y caballos, con alturas y pesos.

# KNN



Nuevo punto de datos:  
¿Es un caballo o un perro?

Nuevo punto de datos:  
¿Es un caballo o un perro?

Nuevo punto de datos:  
¿Es un caballo o un perro?

# KNN

Algoritmo de entrenamiento:

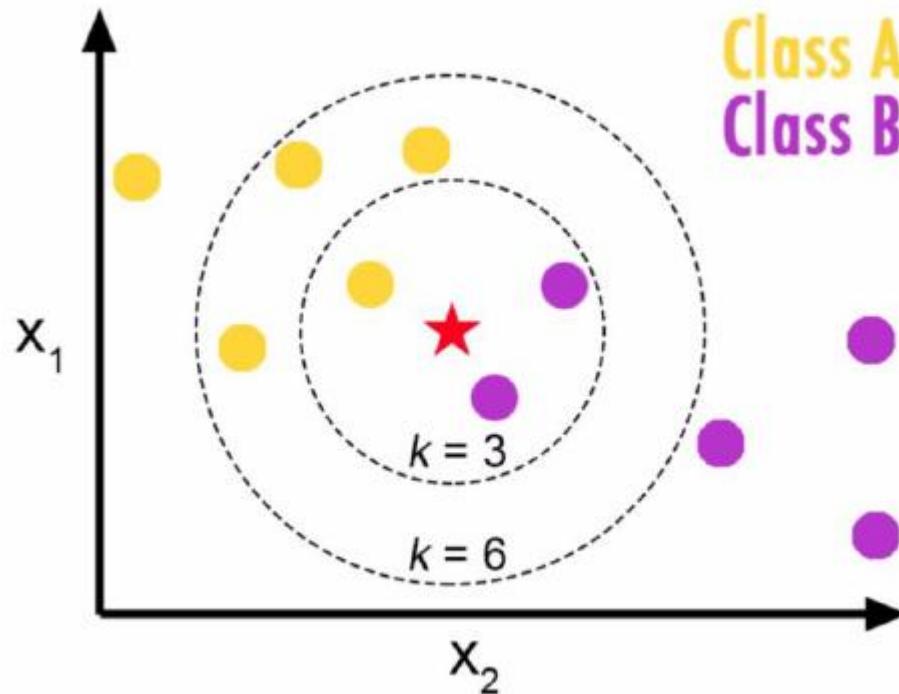
1. Almacenar todos los datos.

Algoritmo de predicción:

1. Calcula la distancia de  $x$  a todos los puntos en tus datos
2. Ordena los puntos en tus datos aumentando la distancia de  $x$
3. Predice la etiqueta de la mayoría de los puntos más cercanos a la " $k$ "

# KNN

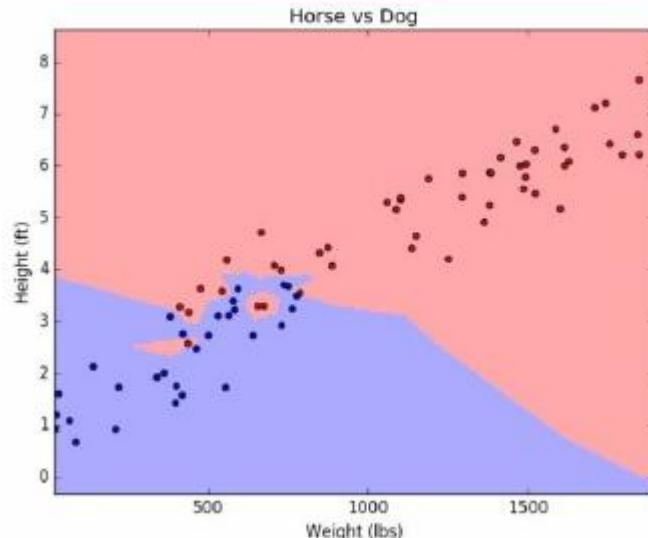
Elegir una K afectará a la clase a la que se asigna un nuevo punto:



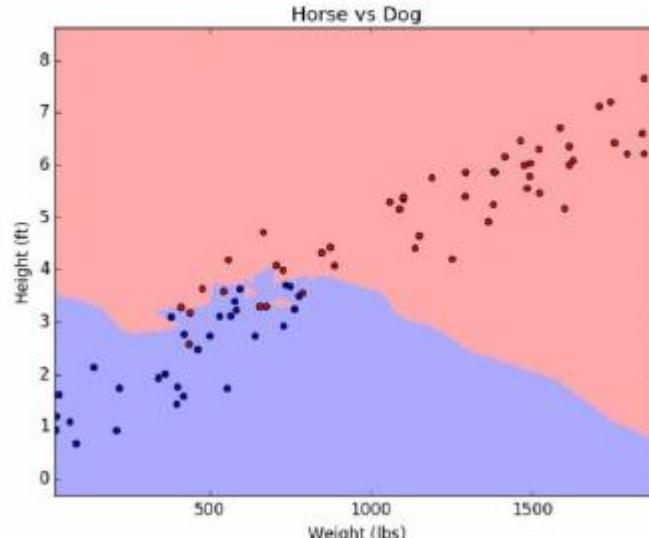
# KNN

Elegir una K afectará a la clase a la que se asigna un nuevo punto:

k=1



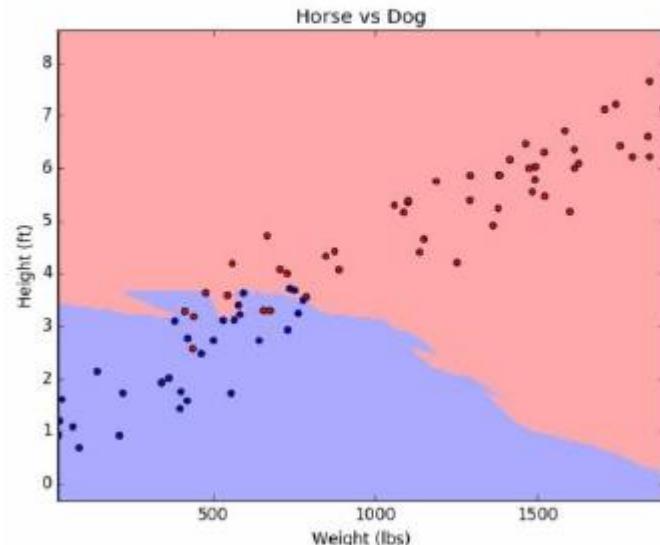
k=5



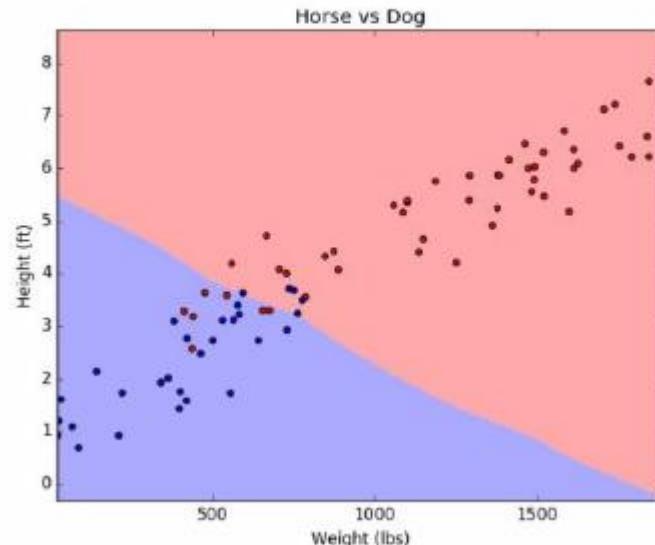
# KNN

Elegir una K afectará a la clase a la que se asigna un nuevo punto:

k=10



k=50



## KNN - Pros

- Muy simple
- El entrenamiento es trivial.
- Trabaja con cualquier número de clases.
- Fácil de agregar más datos
- Pocos parámetros

✓ K

✓ Métrica de distancia

# KNN Contras

- Alto costo de predicción (peor para conjuntos de datos grandes)
- No es bueno con datos de altamente dimensionales.
- Las características categóricas no funcionan bien