



30.01 Semana 3

> Machine Learning

Introducción al machine learning

> J. Antonio García Ramírez
jose.ramirez@cimat.mx

Agenda semana

- Lunes:
 - Intro to ML or statistical learning ?
- Martes:
 - Aprendizaje no supervisado
- Miercoles:
 - Aprendizaje supervisado
- Jueves: Sorpresa
- Viernes: Resultados y NeuralNetworks

Agenda de hoy

- About us (candys)
 - Situación actual (25)
- Machine Learning
 - Definición (10)
 - Qué no es (10)
 - History or story (10) `sys.sleep(5)`
- Ejemplos y aplicaciones (15)
- Mismo lenguaje, diccionario(s) (15)

Agenda

- Flujo de trabajo (15) `sys.sleep(5)`
- El árbol (1)
 - Y conocida (supervisado)
 - Nadie sabe (no supervisado)
 - Regaño (refuerzo)
 - No free launch
 - Tu jardín
- `Sys.sleep(15)`
- Implementando proyectos de ML (10)

Agenda

- Lo hard: (30)
 - $Bias^2(x) + Var(x) + \epsilon_{natural}$
 - Selección de modelos
- Proyectos (10)
- Repaso (10 + extra)

Who ?

Industria

Academia



Consultor estadístico Analista de datos

A qué dedica el tiempo libre

Comunidad TIC, Maths

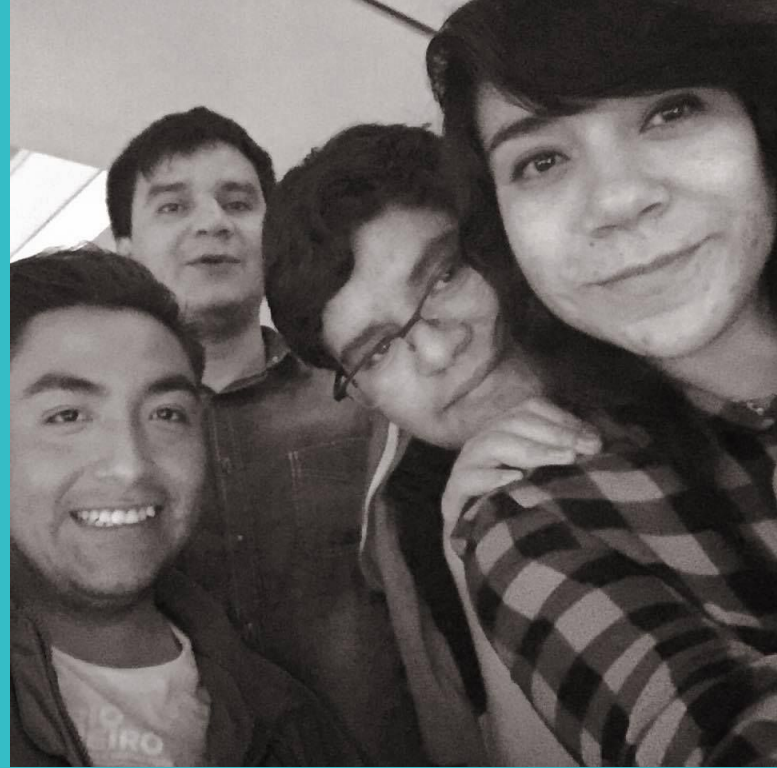


Foo ?



Qué motiva a foo?

- Productos de datos
- Cómputo estadístico y estadística computacional
- Mis hermanos



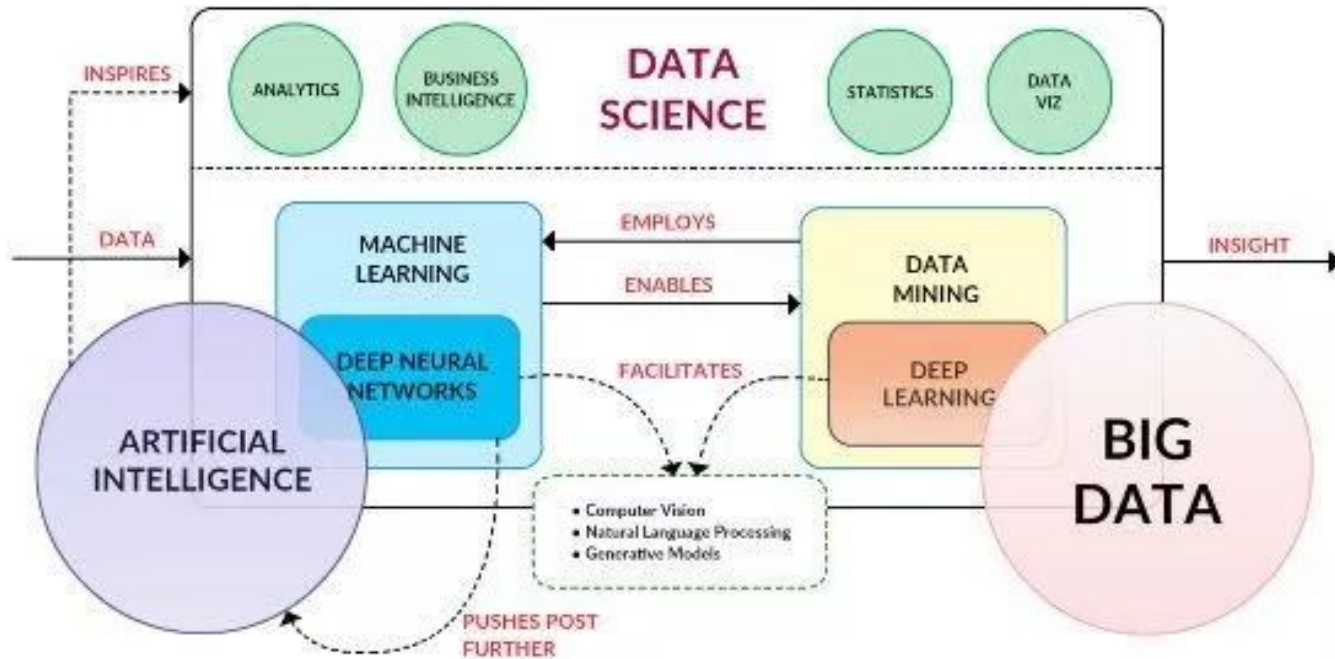
Machine learning: contexto

- ¿Quién ha escuchado sobre Machine Learning?
- ¿Quiénes conocen Machine Learning?
- ¿Quiénes están usando Machine Learning?

Definición

- Machine learning (ML) is a field of artificial intelligence that uses statistical techniques to give computer systems the ability to "learn" (e.g., progressively improve performance on a specific task) from data, without being explicitly programmed
- Arthur Samuel (1959)

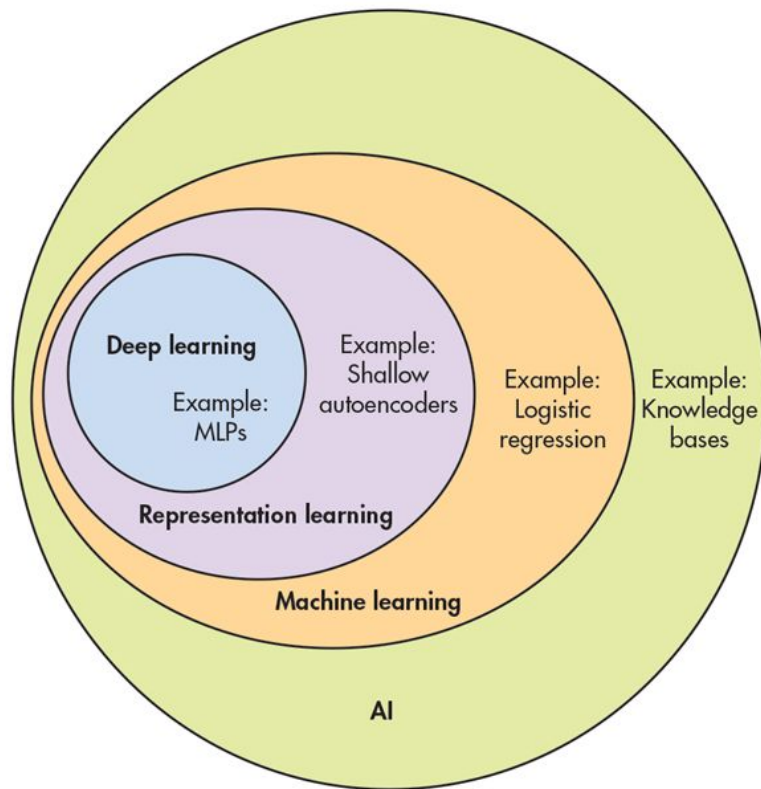
Qué no es ML, sus primos



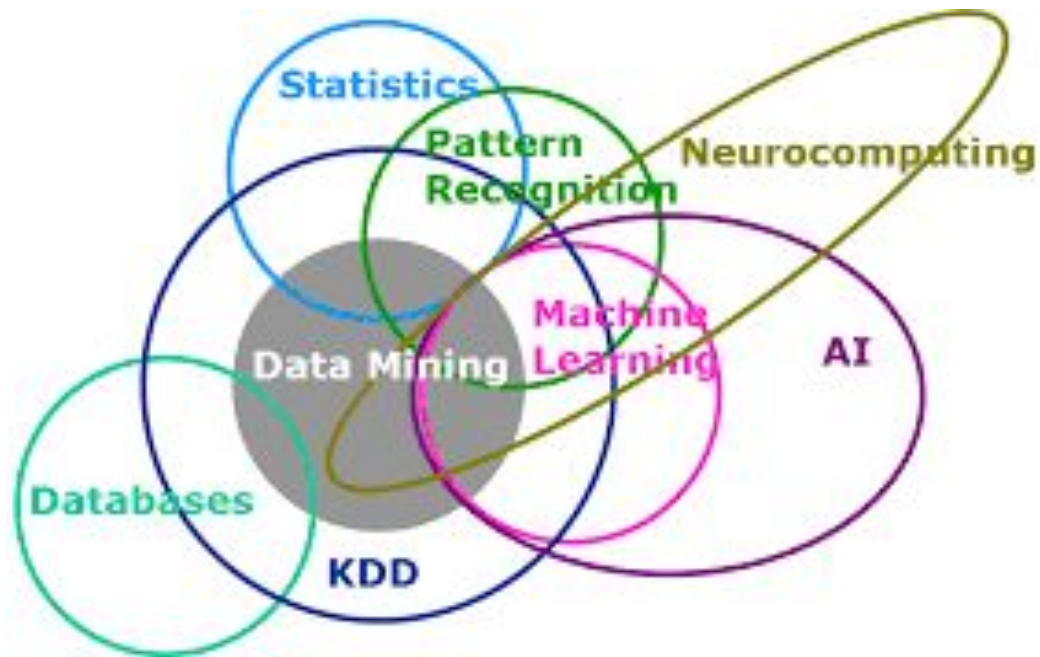
History or story ?

- ML is closely related to (and often **overlaps** with) computational statistics.
- It has strong ties to **mathematical optimization**, which delivers methods, theory and application domains to the field.
- Machine learning is sometimes conflated with **data mining**
- EDA: 50 años de DS, Tukey (1961, The future of data analysis)
 - o https://projecteuclid.org/download/pdf_1/euclid.aoms/1177704711

Sobre Machine Learning



Story



Aplicaciones

- Detección de Fraudes, fiscales

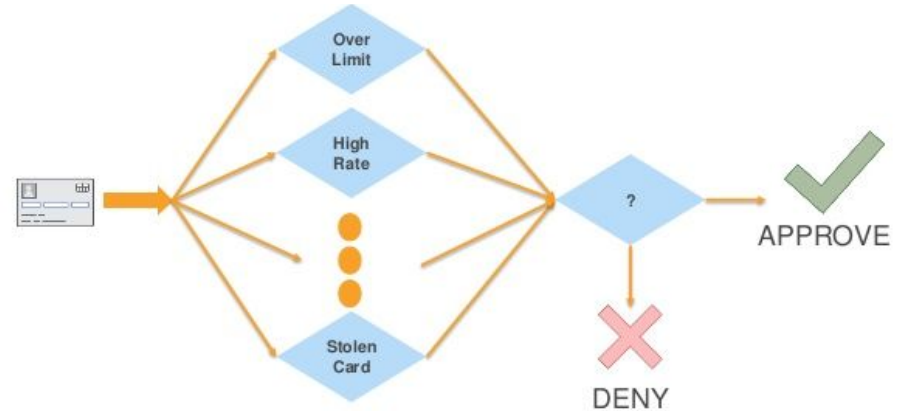
<https://foufoo.shinyapps.io/app1/>

- Billetes falsos

<https://itunes.apple.com/us/app/swiss-banknotes/id1097859820?mt=8>

- Movilidad

https://github.com/fou-foo/opi_test/blob/master/exa_opi.html



Aplicaciones

- **NLP** <https://foufoo.shinyapps.io/projectnlp/>
- **Buscador google (el gran pionero, indexación, Hadoop, NLP, imágenes)**

¿Cuándo es útil?

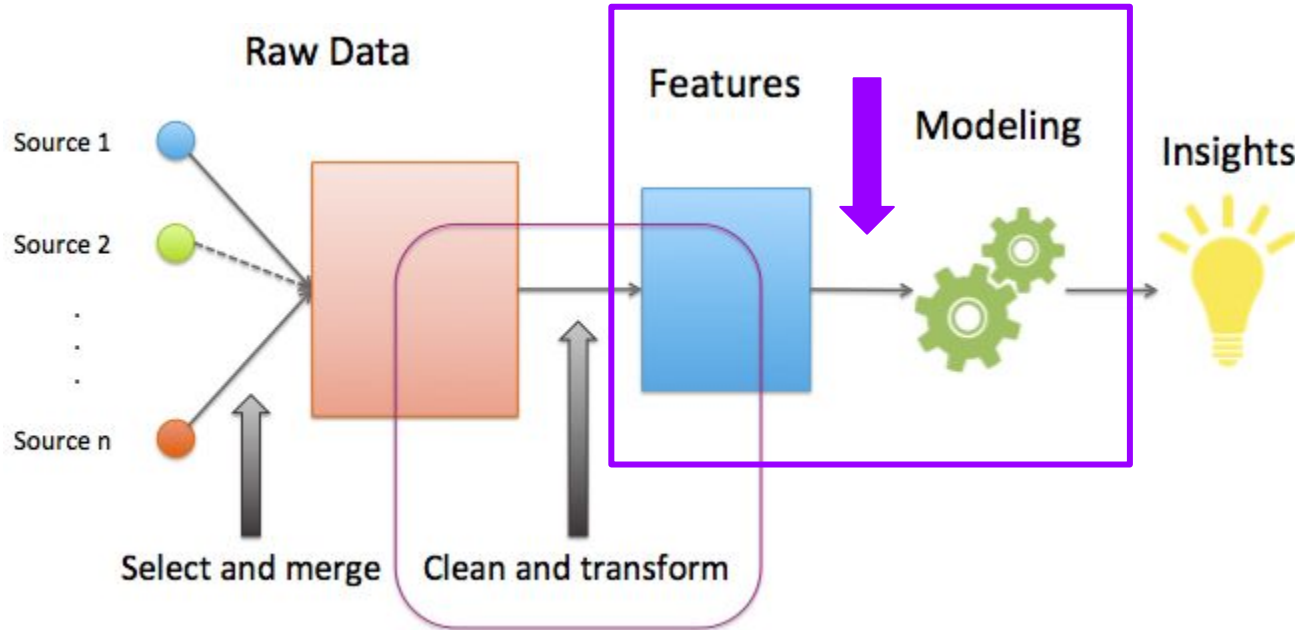
- Los humanos no pueden explicar su experiencia (reconocimiento de voz)
- Cambios rápidos para la percepción humana en el tiempo (Ruteo en una red, kilos de información)
- Disminuir costos, mejores decisiones

¿Cuándo no es tan útil?

- Restricción por legislación (Basilea y solvencia)
- No hay registros
- No hay arquitectura

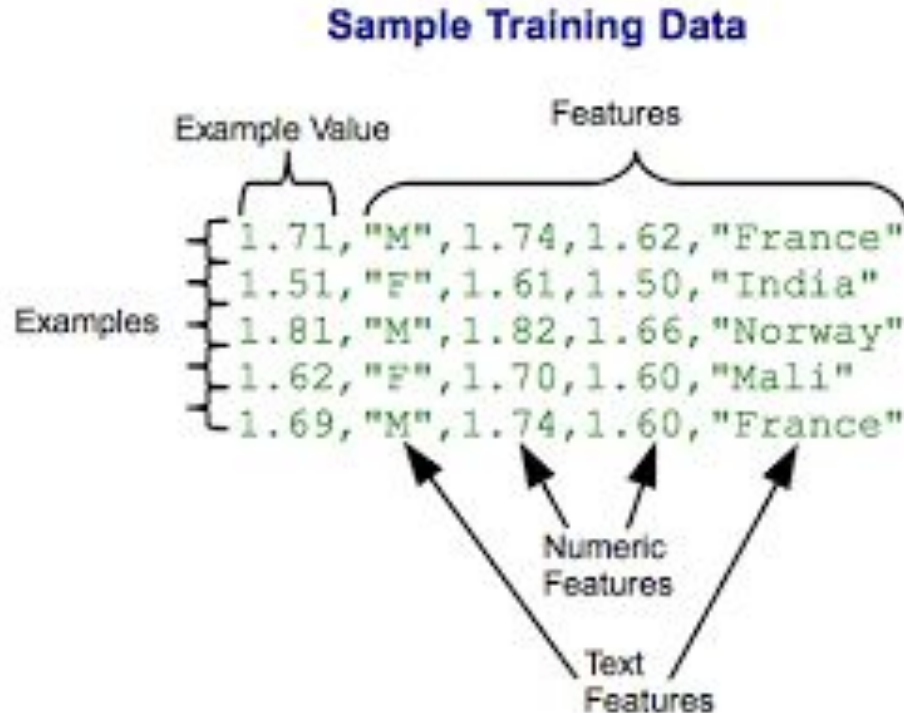
Mismo lenguaje

- Características (Features)



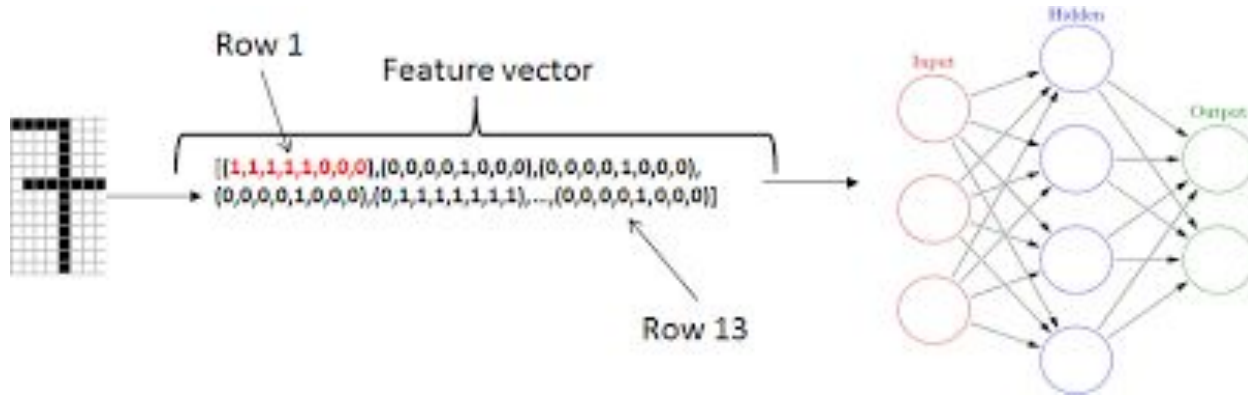
Diccionario de Elementos

- Muestras (sampling)



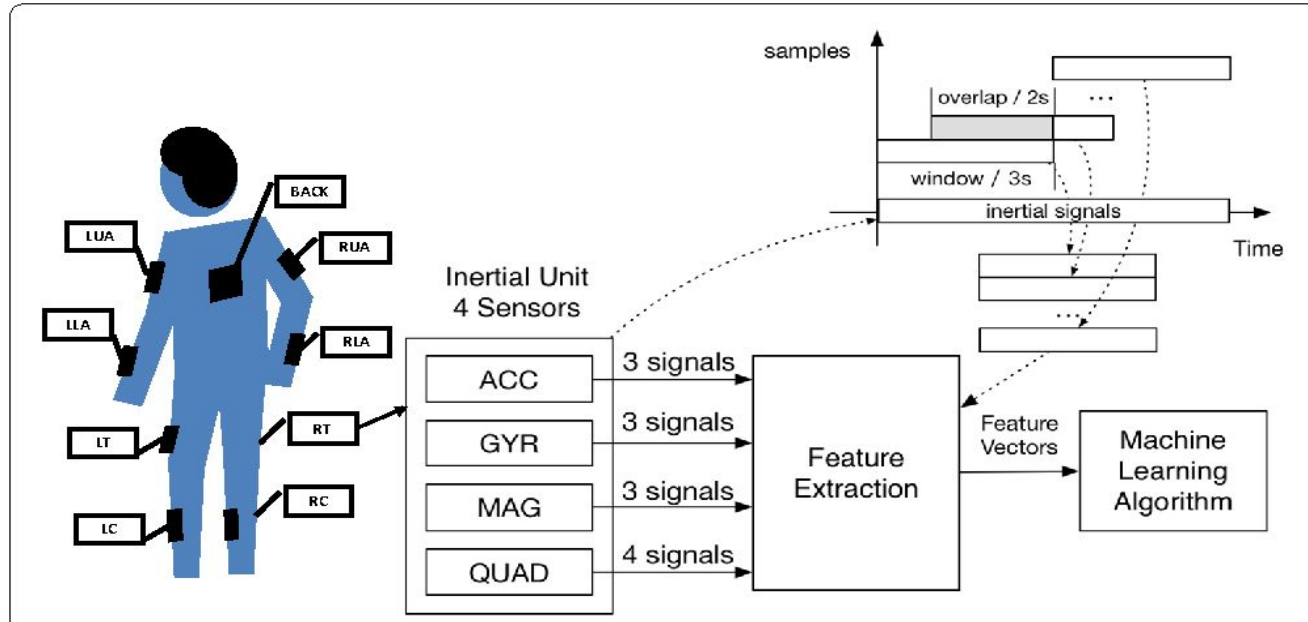
Diccionario de Elementos

- Vector de características (variables)



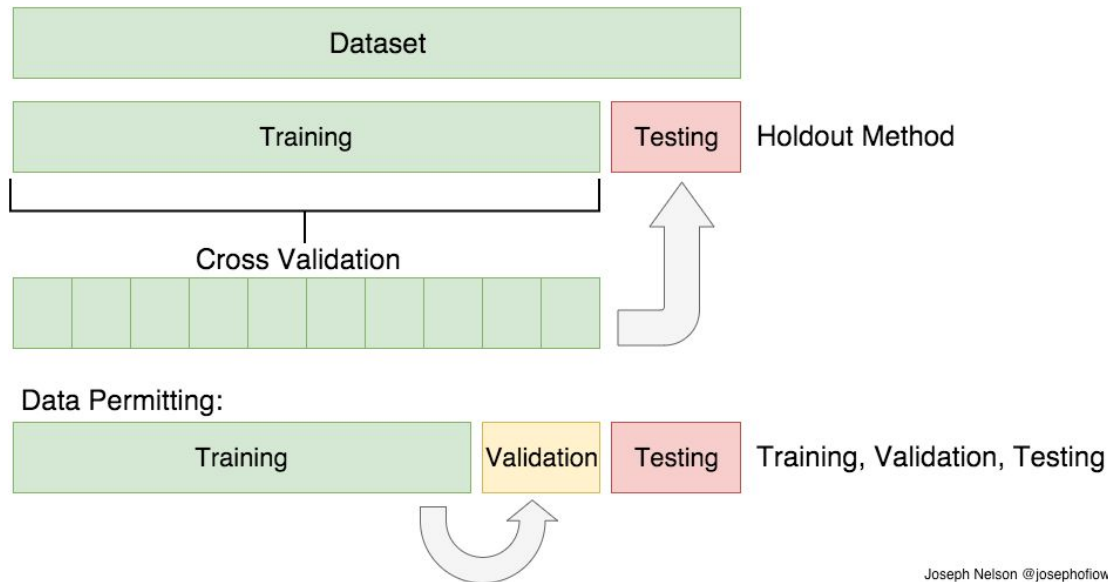
Diccionario de Elementos

- Feature Extraction. Imaginación



Diccionario de Elementos

- Training/Validation/Test Set



Diccionario de elementos

- Etiquetas
- Atributos, features o variables
- Variable aleatoria
- Función
- Función de costo
- ERROR

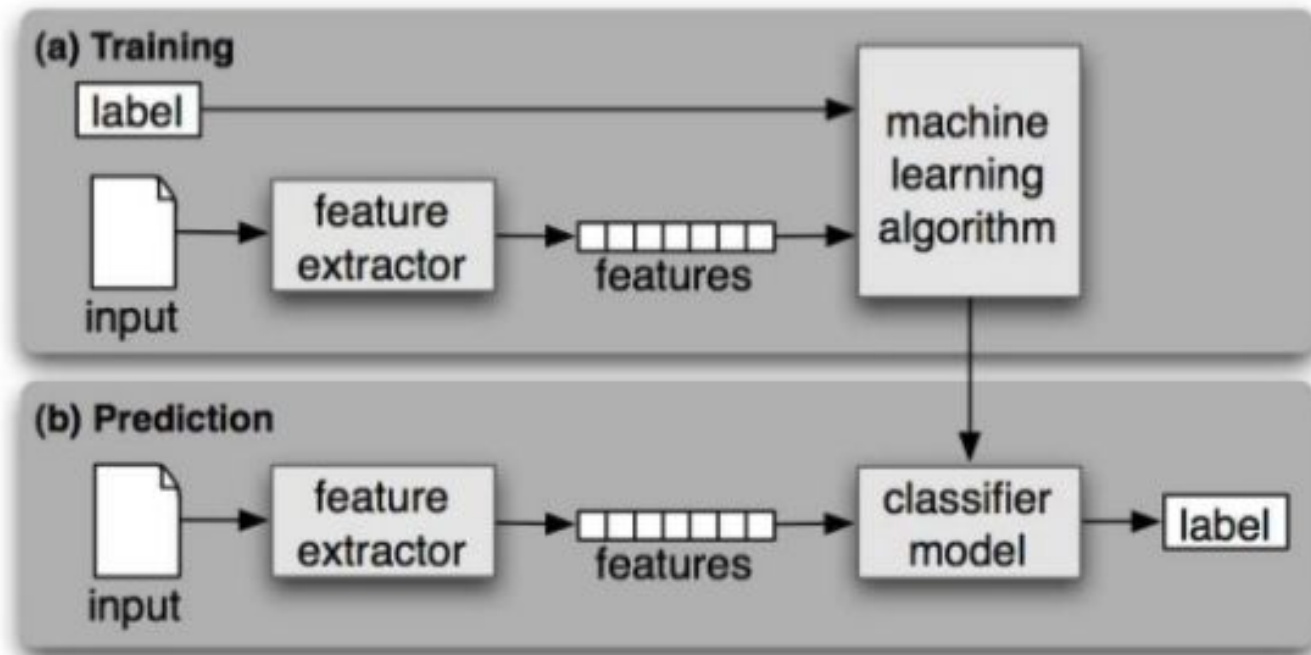
Definiendo un elemento

- ¿Qué entendemos por una manzana?



Características

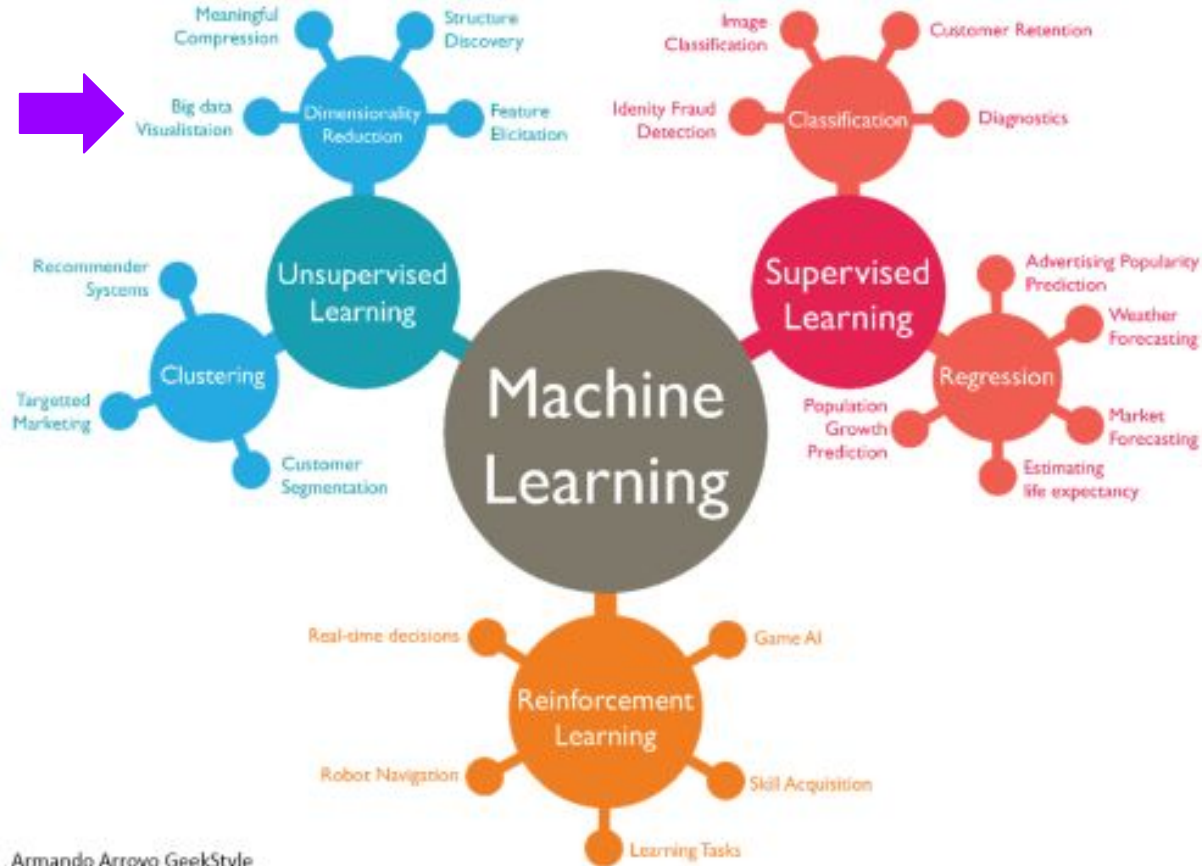
Flujo de trabajo (workflow)



Elementos básicos

- **Pregunta de interes**
- Datos, arquitectura
- Modelo e implementación factible
 - Función de optimización
 - Elección de modelo
- Resultado, loop

El árbol



No free launch

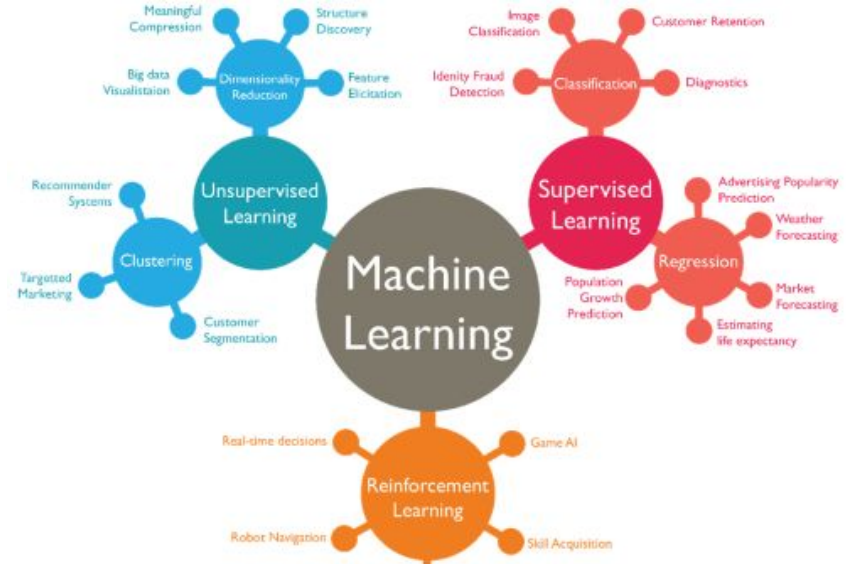
- Uno de los objetivos del ML es presentar una amplia gama de métodos de aprendizaje estadístico que se extienden más allá del enfoque de regresión lineal estándar. ¿Por qué es necesario introducir tantos enfoques de aprendizaje estadístico diferentes, en lugar de un método único y mejor?

No free launch

- Ningún método domina a todos los demás sobre todos los conjuntos de datos posibles. En un conjunto de datos particular, un método específico puede funcionar mejor, pero algún otro método puede funcionar mejor en un conjunto de datos similar pero diferente. Por lo tanto, es una tarea importante decidir para cualquier conjunto de datos dado qué método produce los mejores resultados. La selección del mejor enfoque puede ser una de las partes más difíciles de realizar el aprendizaje estadístico en la práctica.



Tu jardín

- Tú pregunta de interés



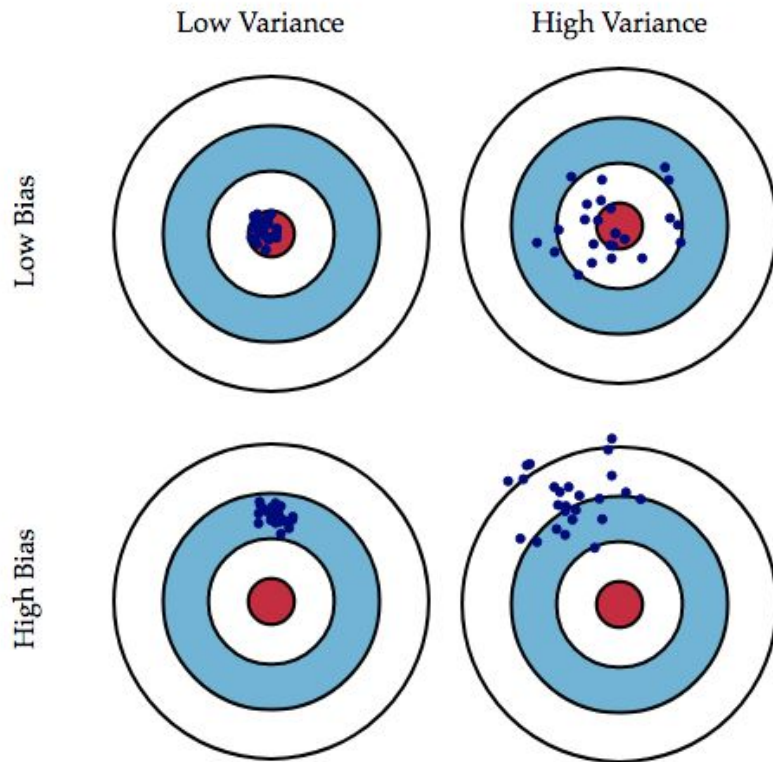
> Compartamos experiencias

Precisión y tipos de error

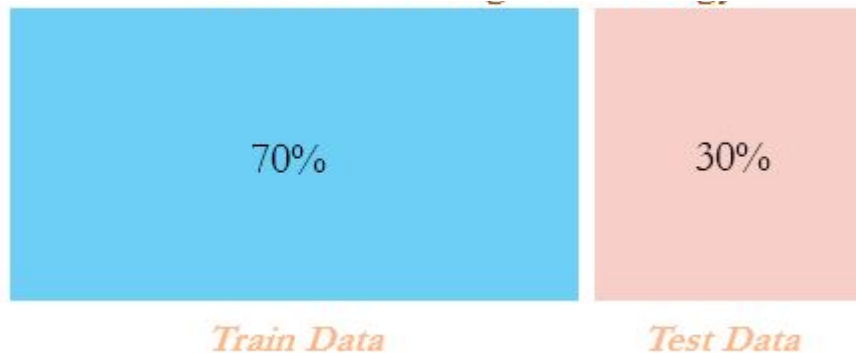
	Hypothesis true	Hypothesis false
Accept. hypothesis		Type II error
Reject hypothesis	Type I error	



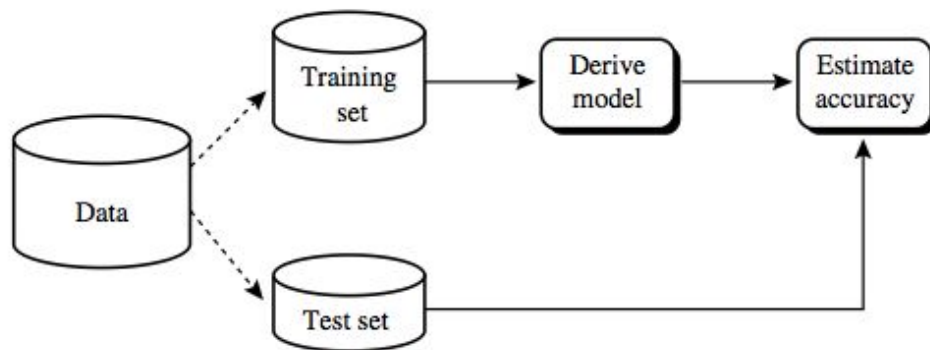
Variance vs Bias tradeoff



Conjuntos train y test

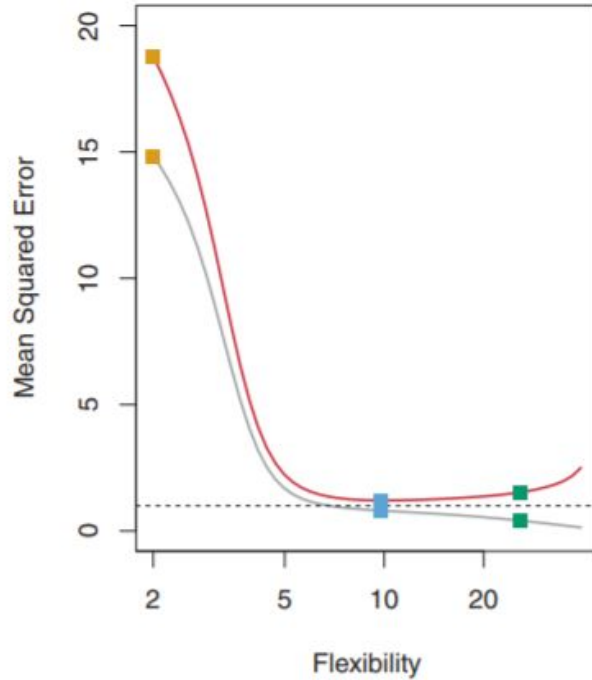


Con muestra
grande

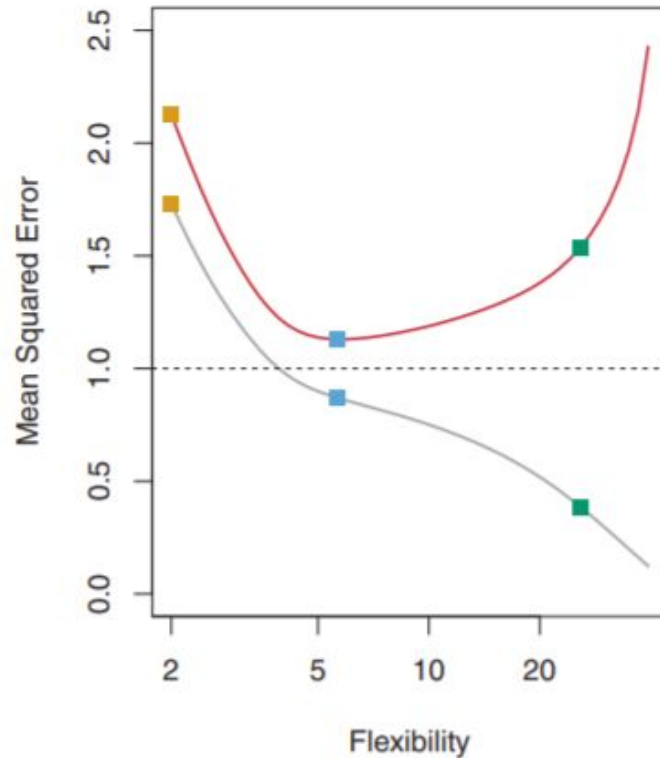


Variance vs Bias tradeoff

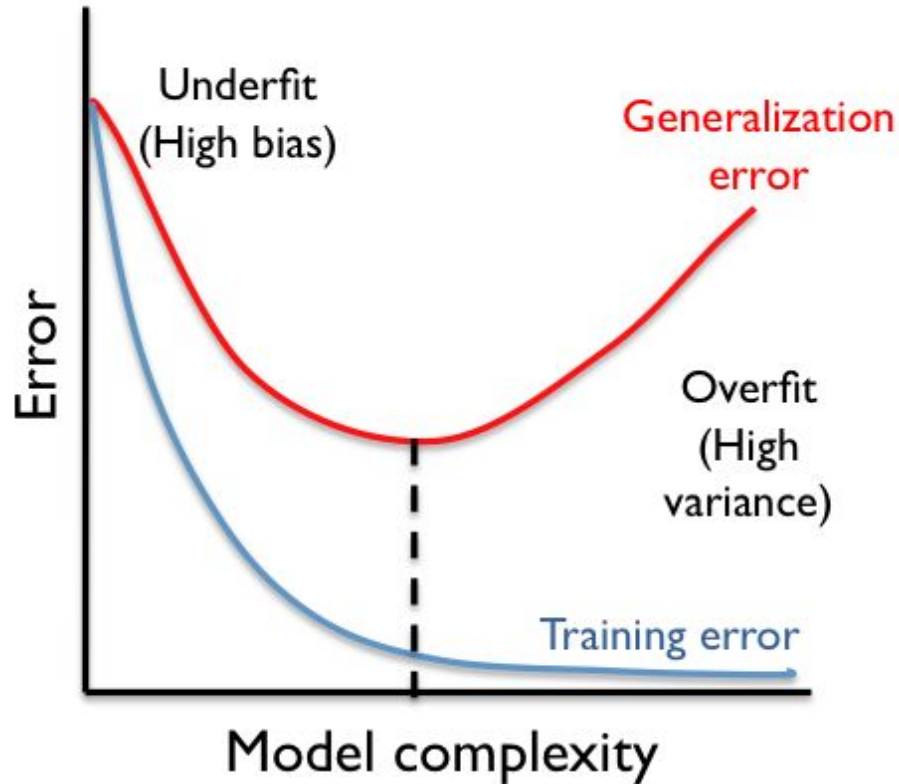
La carta a los reyes magos



Variance vs Bias tradeoff

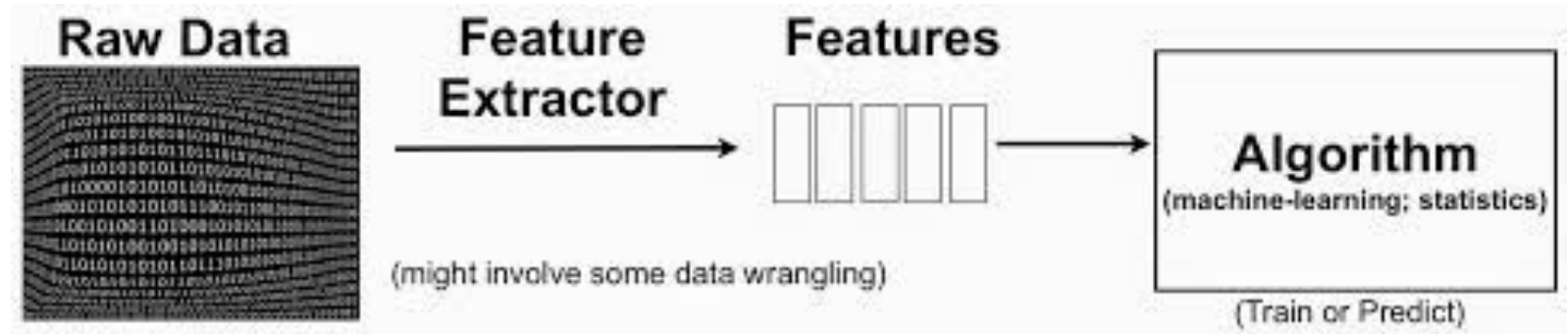


Variance vs Bias tradeoff



Feature Engineering

- Transformar la data



Tipos de variables

- Continuas (cuantitativas)
- Discretas (cualitativas)
 - Nominal
 - Ordinal
 - Categórica

> Ejemplos de feature engineering para características, imágenes, audio, etc

El dilema de los hornos

> Ejemplos de feature Surf features

min 1

<https://www.youtube.com/watch?v=ZXn69V-1kEM>



Implementando un proyecto

- La naturaleza del aprendizaje automático es **asintótico** a 100%, cada vez más caro.
 - Doble de observaciones **no es doble** precisión, $n^{.5}$
- Lo importante no es el algoritmo, **tiempo**
- Ten en cuenta siempre va a mejorar.

> Manos a la obra:

Plantea un proyecto de
aprendizaje automático

Cota inferior: EDA

Data sets:

> Iris <https://archive.ics.uci.edu/ml/datasets/iris>

> Songs <https://archive.ics.uci.edu/ml/datasets/YearPredictionMSD#>

Muestra de 1% en

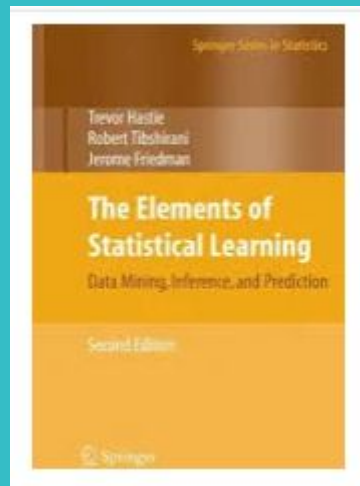
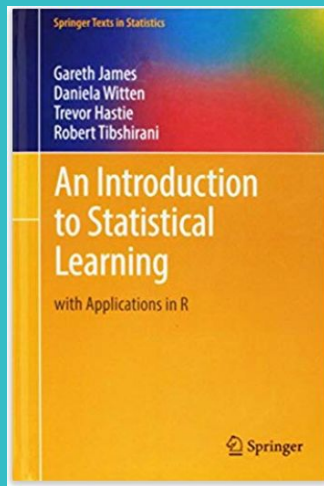
https://github.com/fou-foo/CeroUnoML/tree/master/Dia1_IntroML/Songs

> Ecobici (un mes cercano):

<https://www.ecobici.cdmx.gob.mx/es/informacion-del-servicio/open-data>

> Tu propuesta

Extra:



<https://web.stanford.edu/~hastie/Papers/ESLII.pdf>

<https://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf>

Stack overflow

Doc. de scikit-learn

<https://scikit-learn.org/stable/>