Ciencia de Datos

Guillermo Moncecchi

Posgrado en Sistemas de Información de las Organizaciones y Gestión de Empresas de Tecnologías de la Información

Facultad de
Ciencias Económicas
y de Administración
Universidad de la República

UNIVERSIDAD
DE LA REPUBLICA
URUGUAY

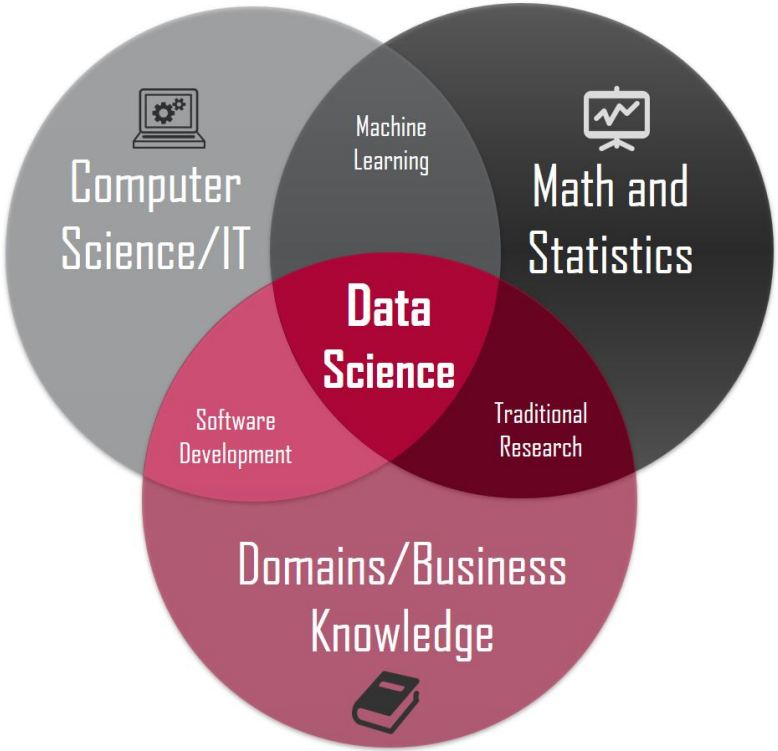
Presentación.

@gmonce
@pln_udelar

The big picture

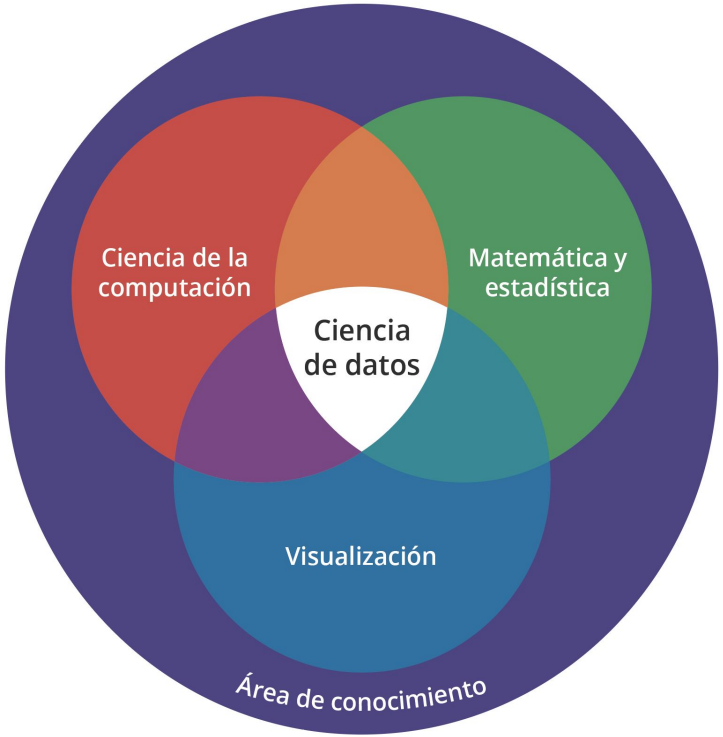
¿Qué es la Ciencia de Datos?

¿Qué es la Ciencia de Datos?



Drew Conway, The Data Science Venn Diagram.

¿Qué es la Ciencia de Datos?



Maestría en Ciencia de Datos y Aprendizaje Automático (en preparación), UdelaR

¿Qué es la Ciencia de Datos?

«La ciencia de datos es la disciplina que busca extraer conocimiento, de forma sistemática y computacionalmente eficiente, a partir de los datos de un dominio. Para esto, utiliza principalmente métodos y técnicas de la matemática y la estadística, la computación y la visualización de datos.»

Maestría en Ciencia de Datos y Aprendizaje Automático (en preparación), UdelaR

Quora, what is Data Science?

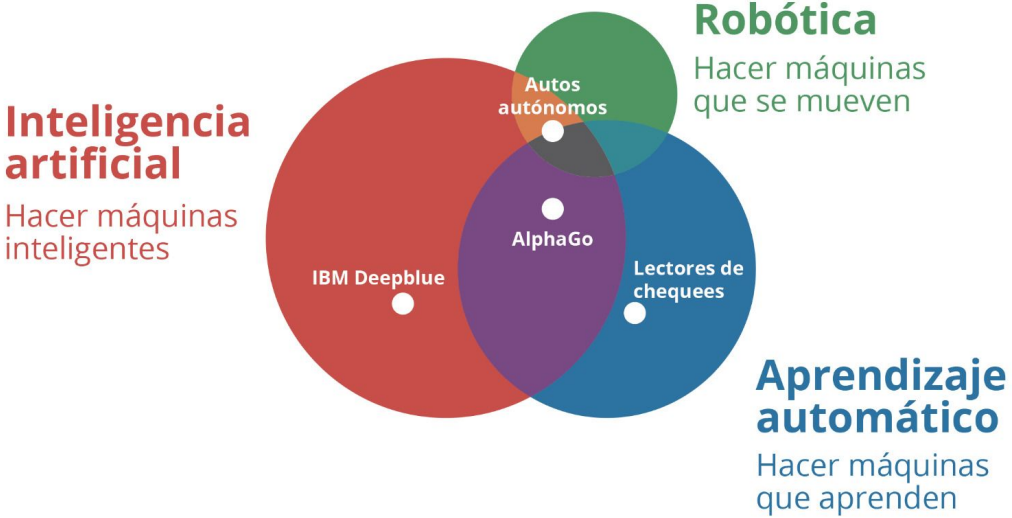
Principales actividades de la ciencia de datos

- Exploración y preparación de los datos
- Representación y transformación de los datos
- Modelado y computación
- Visualización y presentación

David Donoho, 50 years of Data Science

¿Qué es la Inteligencia Artificial?

¿Qué es la Inteligencia Artificial?
(Greg Corrado chart)



Greg Corrado, Tallinn Digital Summit 2018

¿Qué es el aprendizaje automático?

¿Qué es el aprendizaje automático?

Métodos que permiten a las computadoras “aprender”: lograr mejor desempeño en determinada tarea a partir de la experiencia.

¿Qué es el aprendizaje automático?

*«Un programa de computadora **aprende** de la experiencia E con respecto a alguna clase de tareas T y de una medida de rendimiento P, si su rendimiento en las tareas de T, medida por P, mejora con la experiencia E»*

Tom Mitchell, Machine Learning, 1997

¿Qué es el aprendizaje automático?

- Aprendizaje Supervisado (supervised learning)
    - Clasificación
    - Regresión
- Aprendizaje no supervisado (unsupervised learning)
    - Clustering
    - Reducción de dimensionalidad
- Aprendizaje por refuerzos (reinforcement learning)

- Active learning
- Semisupervised learning

Métodos De Aprendizaje Automático

Aprendizaje Supervisado

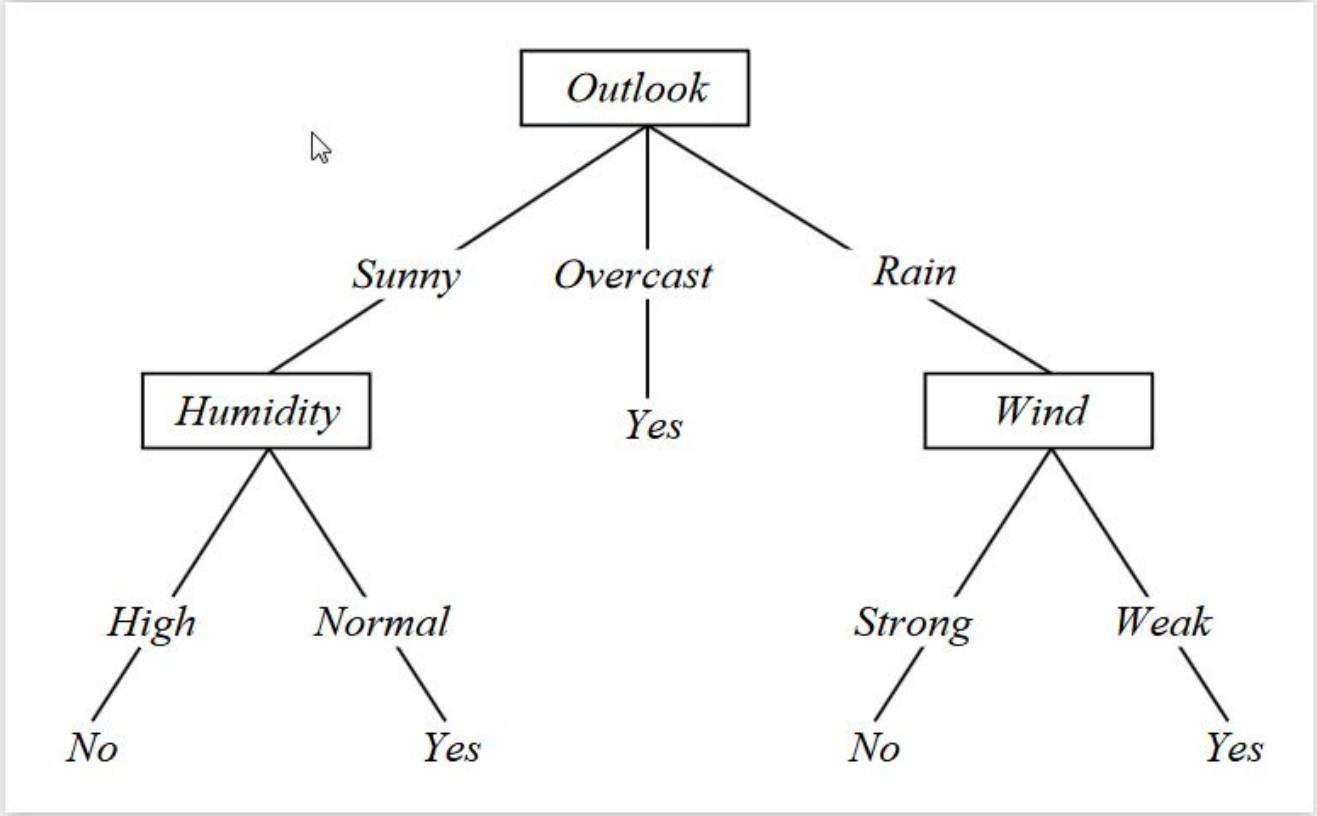
- Tenemos un conjunto de datos {‹x,y›}, formado por instancias x de un dominio cualquiera D, cada una de ellas con un valor asociado y, perteneciente a la clase T (clase objetivo).
- Buscamos aprender una función f:D -› T (hipótesis) que minimice la discrepancia entre sus predicciones y los valores de y del conjunto de datos
- Cuando T es discreta, hablamos de clasificación. Si T es continua (en general, los números reales), es un problema de regresión.

Aprendizaje
Supervisado/Clasificación
(Árboles de decisión)

Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

TABLE 3.2
Training examples for the target concept *PlayTennis*.

Aprendizaje Supervisado/Clasificación (Árboles de decisión)



Sesgo preferencial: el algoritmo prefiere ciertas hipótesis sobre otras.
Sesgo restrictivo: el espacio de hipótesis es incompleto.

¿Cuál sería el sesgo de los árboles de decisión?

Aprendizaje Supervisado/Regresión (Regresión Lineal)

- Las instancias de entrenamiento pertenecen a R^n
- Buscamos aprender $h(x)$: $R^n \rightarrow R$

- Regresión lineal:

$$h_\theta(x) = x\,\theta^T$$

- Buscamos minimizar la función de mínimos cuadrados

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$$

Aprendizaje
Supervisado/Regresión
(Regresión Lineal)

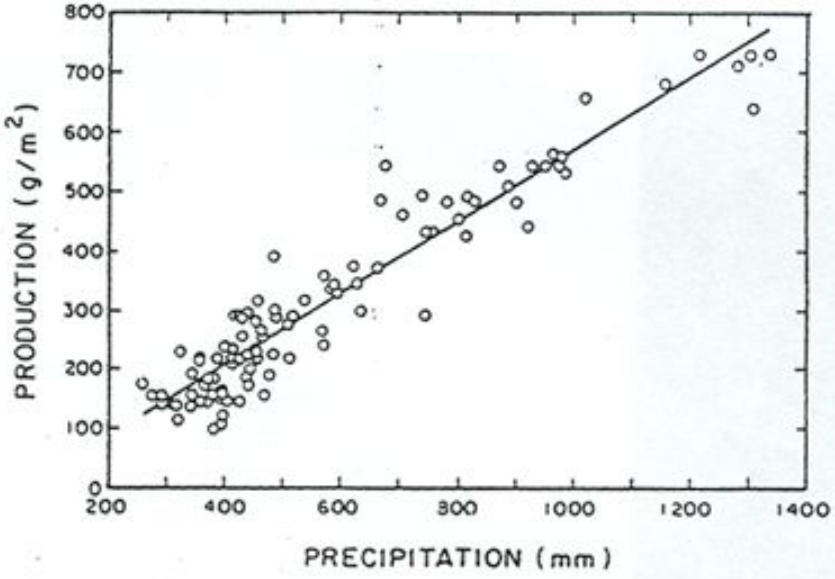
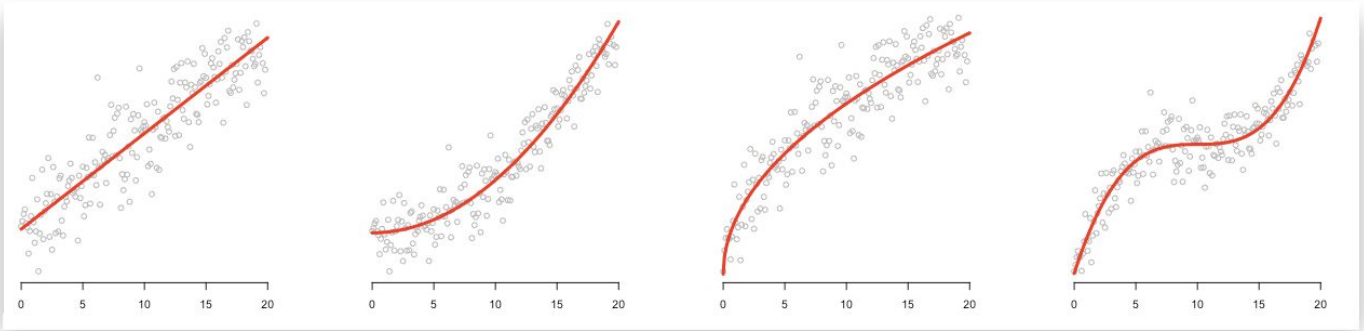


FIG. 2. Relationship between mean annual precipitation and mean aboveground net primary production (ANPP) for 100 major land resource areas across the Central Grassland region. ANPP = −34 + 0.6·APPT; $r^2 = 0.90$.

El agua como recurso limitante para el crecimiento de la vegetación

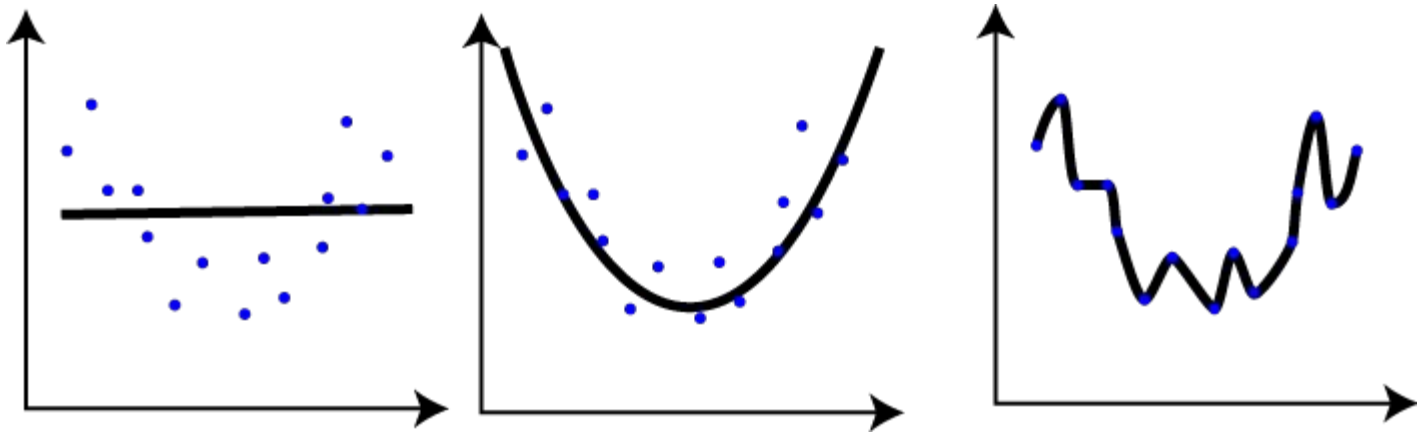
Aprendizaje Supervisado/Regresión (Regresión Lineal)



[How you can use linear regression models to predict quadratic, root, and polynomial functions]

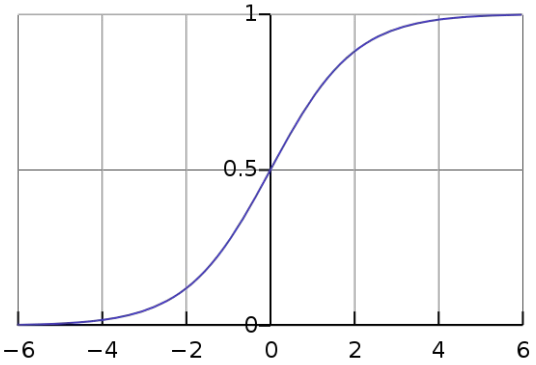
Underfit vs Overfit (sobreajuste). **Concepto muy importante!**

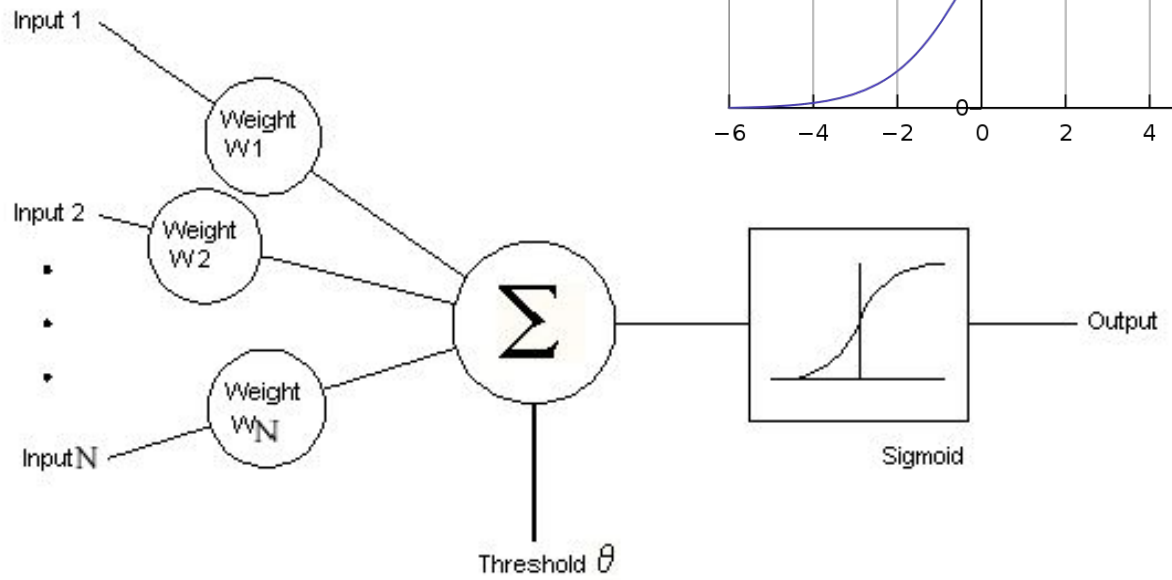
Underfit vs Overfit.



Aprendizaje Supervisado/Redes Neuronales

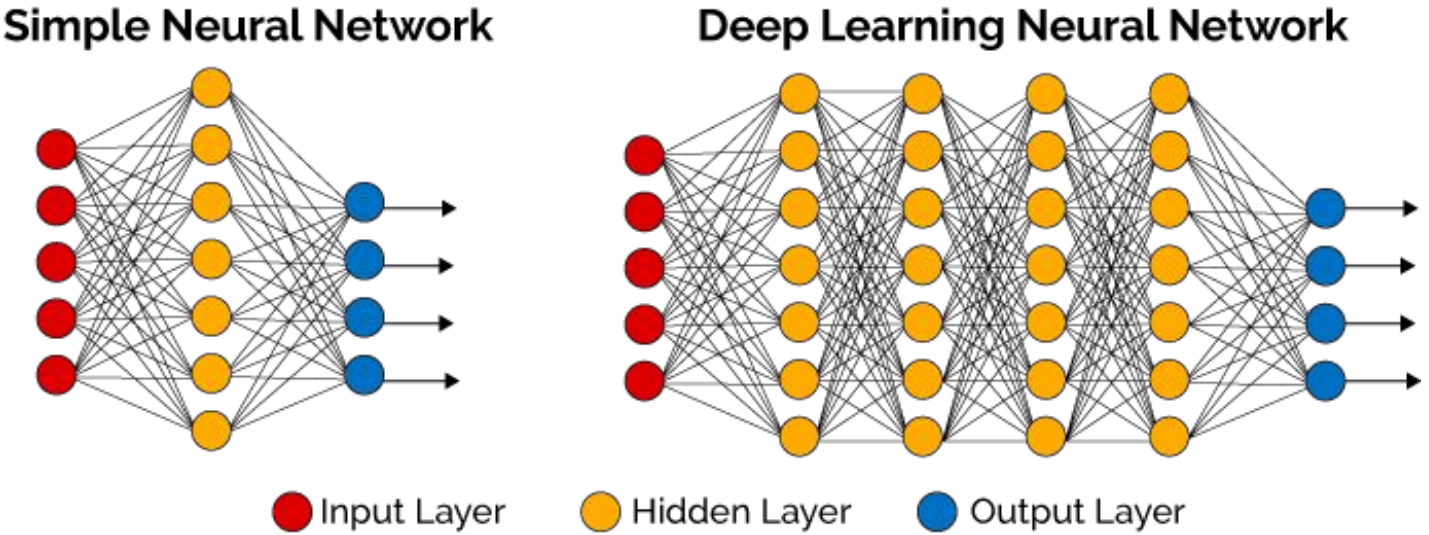
$$S(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{e^x + 1}.$$





Michael Nielsen, Neural Networks and Deep Learning

Aprendizaje Supervisado/Deep Learning

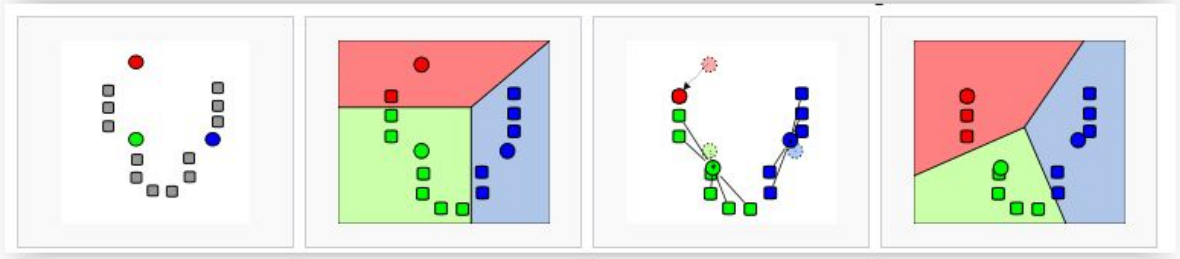


Michael Nielsen, Neural Networks and Deep Learning

Tensorflow Playground.

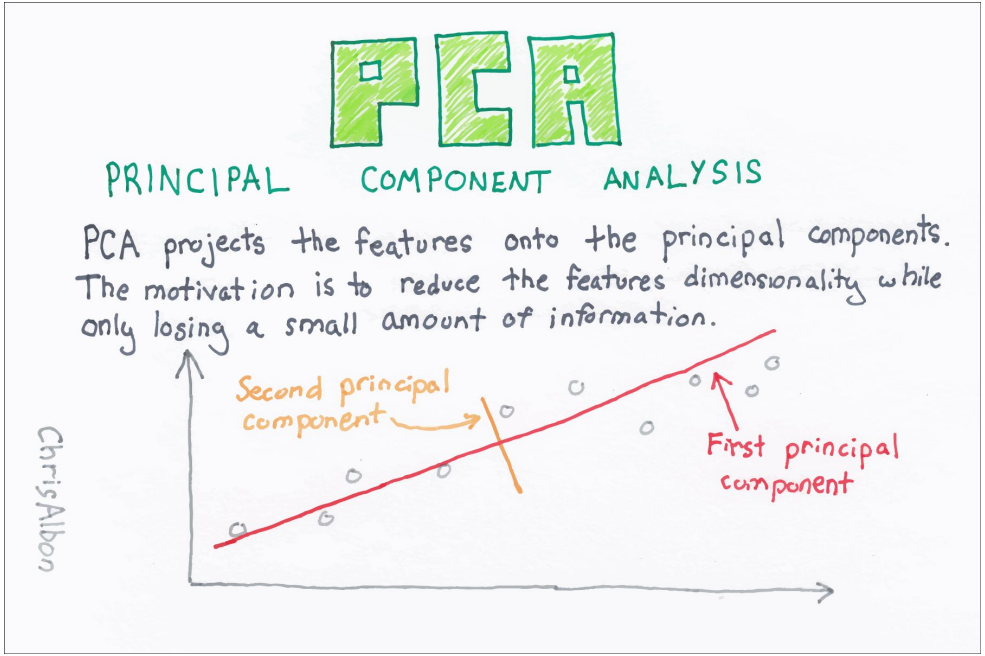
Understanding Neural Networks with the Tensorflow Playground

Aprendizaje No supervisado/Clustering



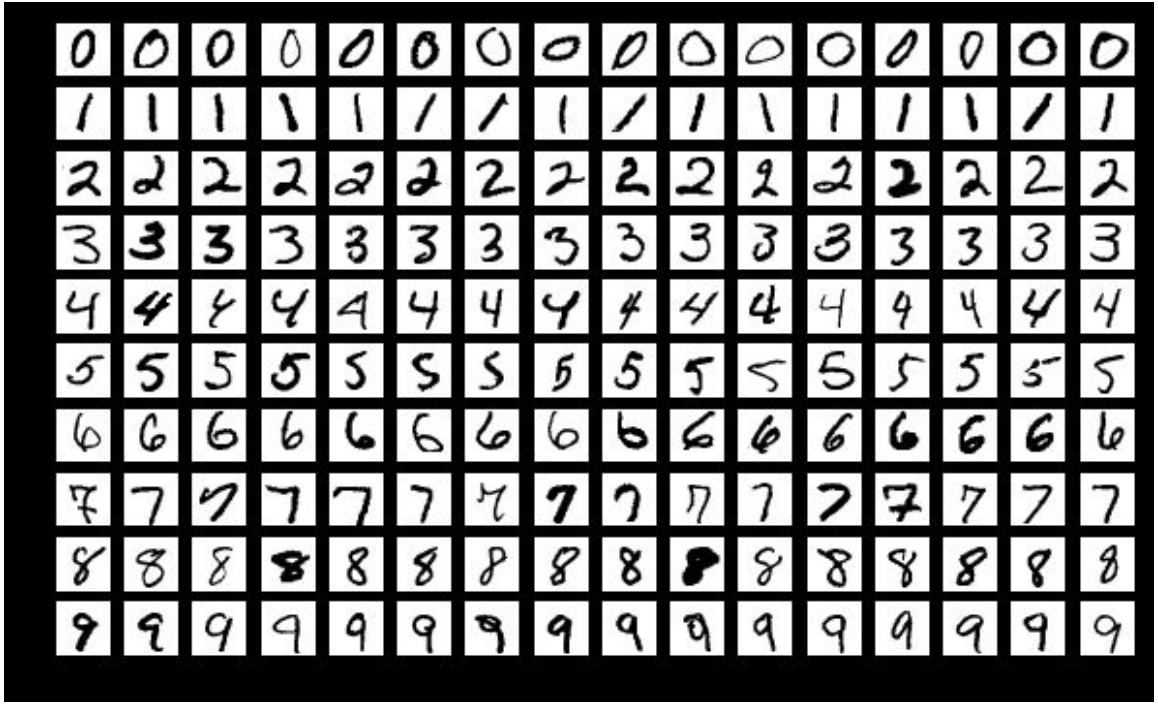
Wikipedia: k-means

Aprendizaje No supervisado/Reducción de Dimensionalidad



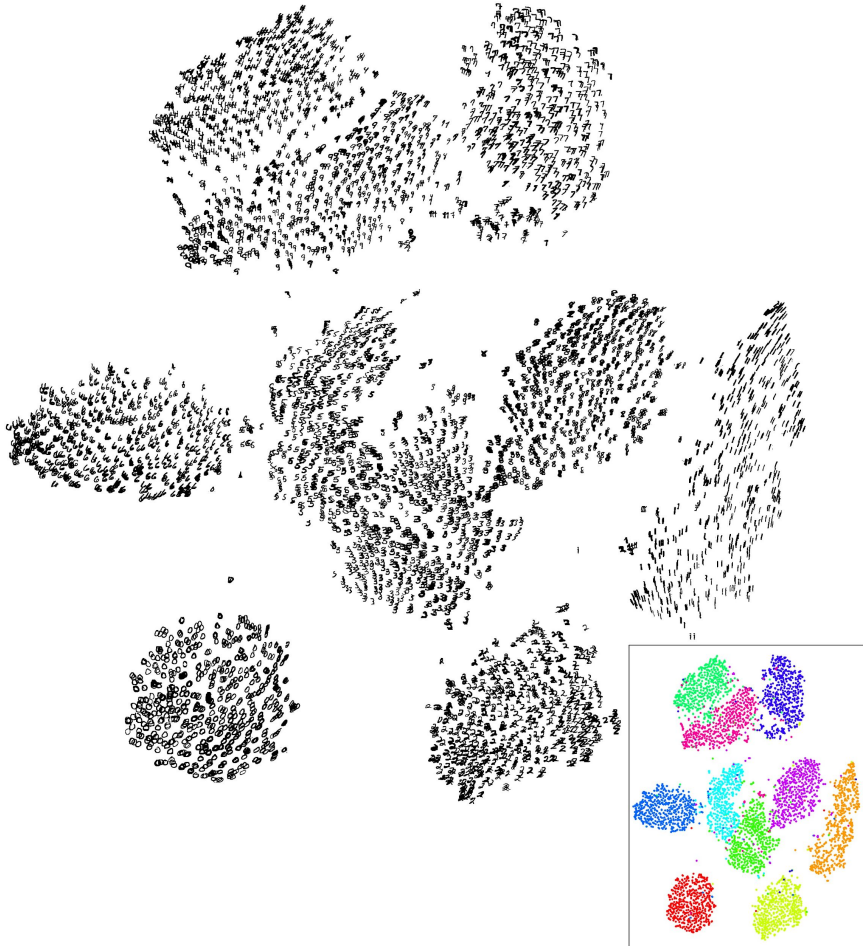
Chris Albon flashcards, [Dimensionality Reduction with PCA]

Aprendizaje No supervisado/t-sne



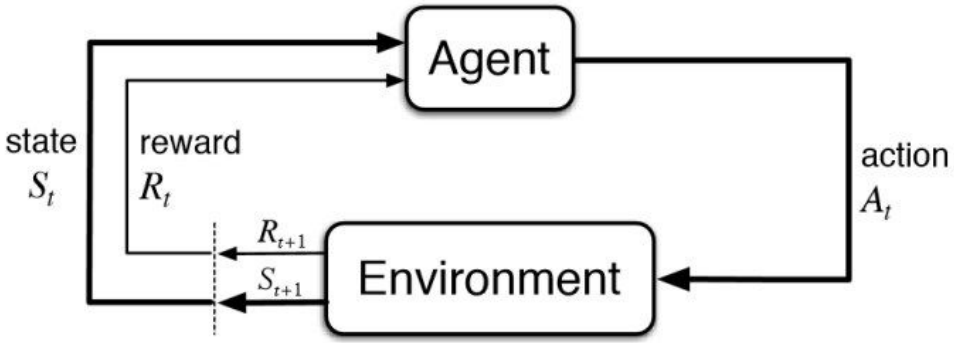
Yan LeCun et al.,The MNIST Database of handwritten digits

Aprendizaje No supervisado/t-sne



Laurens van der Maaten, t-sne

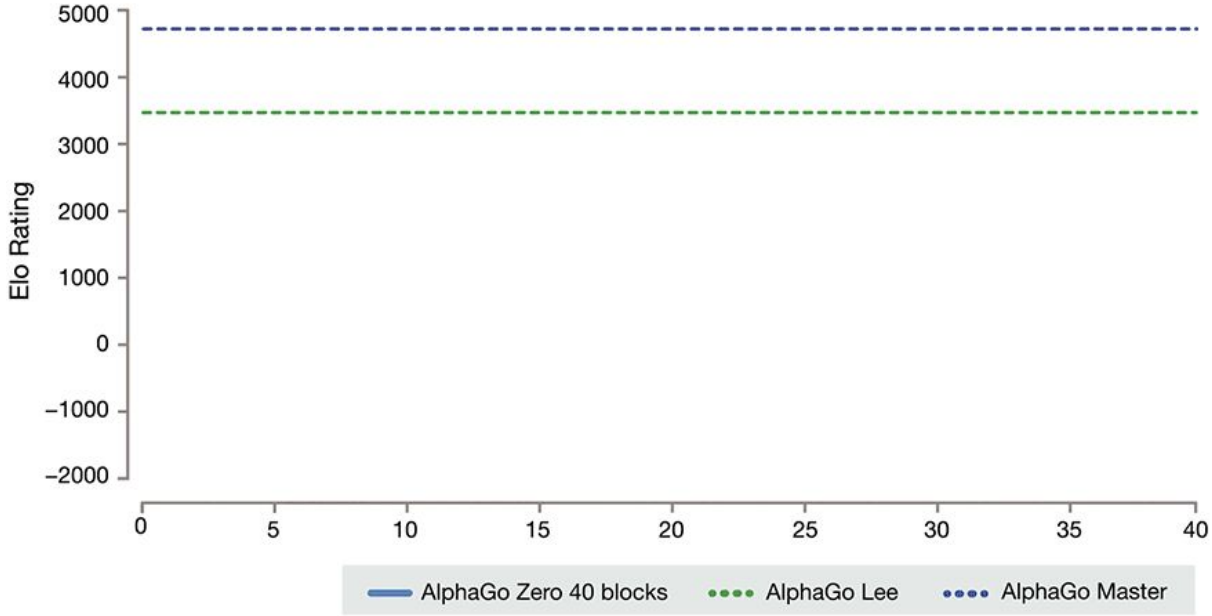
Aprendizaje por Refuerzos



Aprendizaje por Refuerzos

Deep Mind, AlphaGo

Aprendizaje por Refuerzos

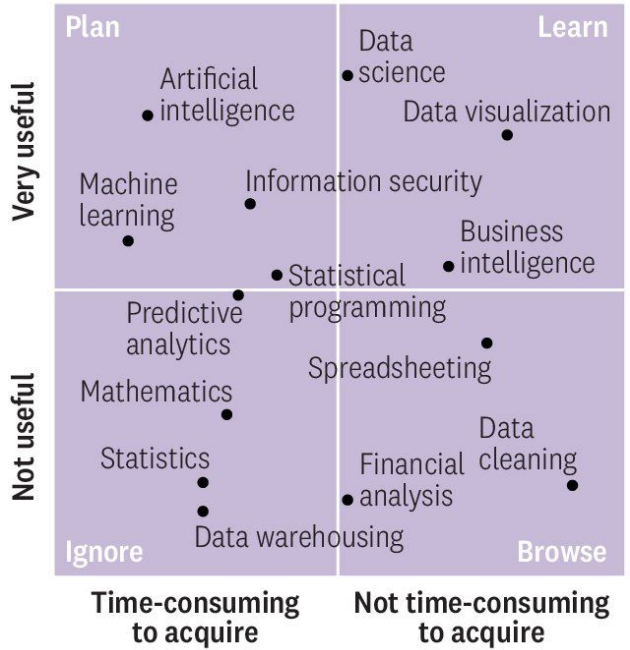


AlphaGo Zero: learning from scratch

Aprendizaje semisupervisado / Active Learning / Transfer Learning/ Generative Adversarial Networks

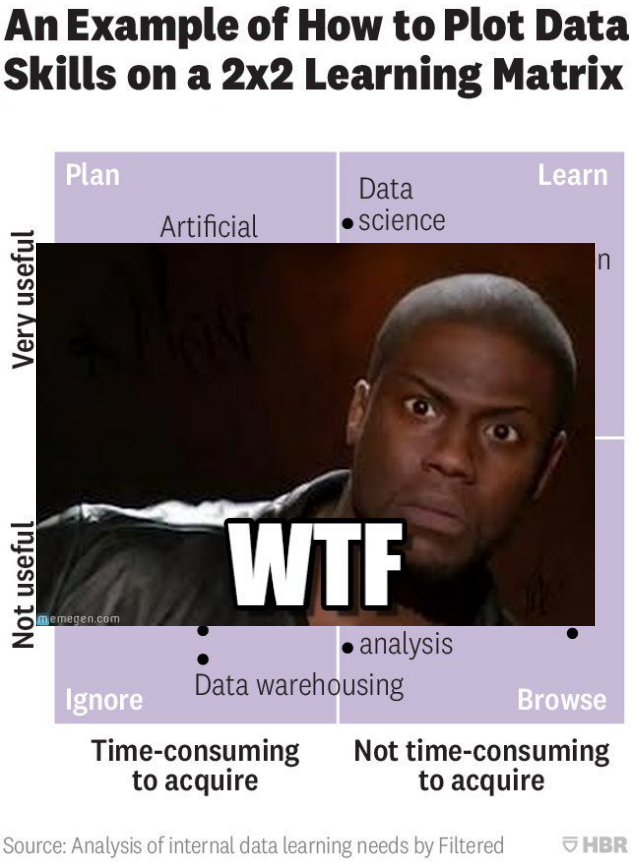
Sebastian Ruder, NLP News

An Example of How to Plot Data Skills on a 2x2 Learning Matrix



Source: Analysis of internal data learning needs by Filtered HBR

Chris Littlewood. [Which Data Skills do you actually need?]



Chris Littlewood. [Which Data Skills do you actually need?]

Metodología (aprendizaje supervisado)

Aprendizaje Supervisado/ Clasificación

- Conjunto de ejemplos (instancias) anotados (dataset):
  - $D = \{ \langle d,c \rangle / \langle d,c \rangle \in X \times C \}$
  - X es el espacio de instancias
  - C es el conjunto de clases
- Una función de clasificación mapea instancias a clases:
  - $\Upsilon: X \rightarrow C$

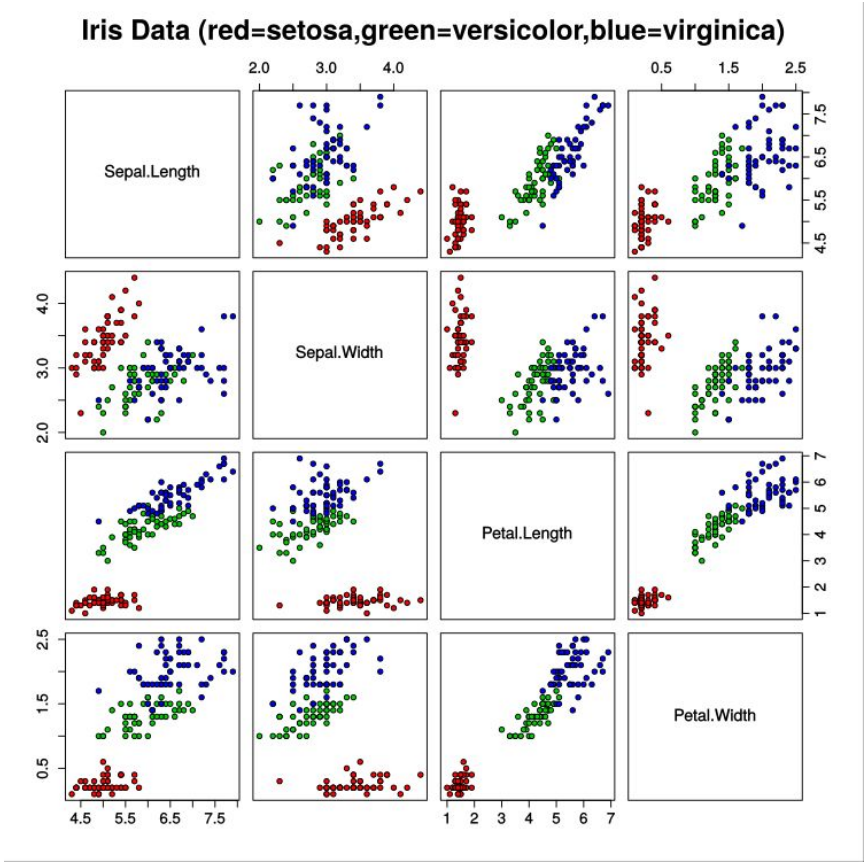
- Un método de *clasificación* recibe los ejemplos anotados y devuelve la función de clasificación

El (legendario) Iris Dataset

- 150 flores de la especie Iris
- Atributos: largo/ancho sépalo, largo/ancho pétalo (reales)
- Número de instancias: 150
- Clase objetivo: clase (Setosa, Versicolor, Virginica)
- 3 clases con 50 instancias cada una

UCI Machine Learning Repository, Iris Data set

El (legendario) Iris Dataset



Wikipedia, Iris Data set

Aprendizaje Supervisado/ Clasificación

- Tarea: dada una instancia de una flor Iris no vista previamente (‹sw,sl,pw,pl›), obtener su clase
- Experiencia: ejemplos anotados manualmente

- ¿Cómo aprendemos una función de clasificación?
- ¿Sobre qué ejemplos?
- ¿Cómo medimos su performance?

Aprendizaje Supervisado/Metodología

- 1. Ingeniería de atributos
- 2. Selección de atributos
- 3. Selección de modelo (ajuste de parámetros)
- 4. Aprendizaje
- 5. Evaluación

- Generalmente se busca tener vectores de la forma de pares ‹atributo, valor› para la entrada.
- Muchos métodos asumen que los atributos son reales
- En el caso del Iris Dataset, los atributos son reales, y expresan la medida en centímetros

Ingeniería de Atributos

- En el mundo real, esto es mucho más difícil, ya que los atributos no son inmediatos (¿cuáles son los atributos que caracterizan un documento?) y dependen de la tarea (¡y de los datos disponibles!)

- Hay que resolver temas como los datos faltantes, pasar de categorías a valores reales (o viceversa), normalizar, etc.

_ Lleva mucho tiempo_

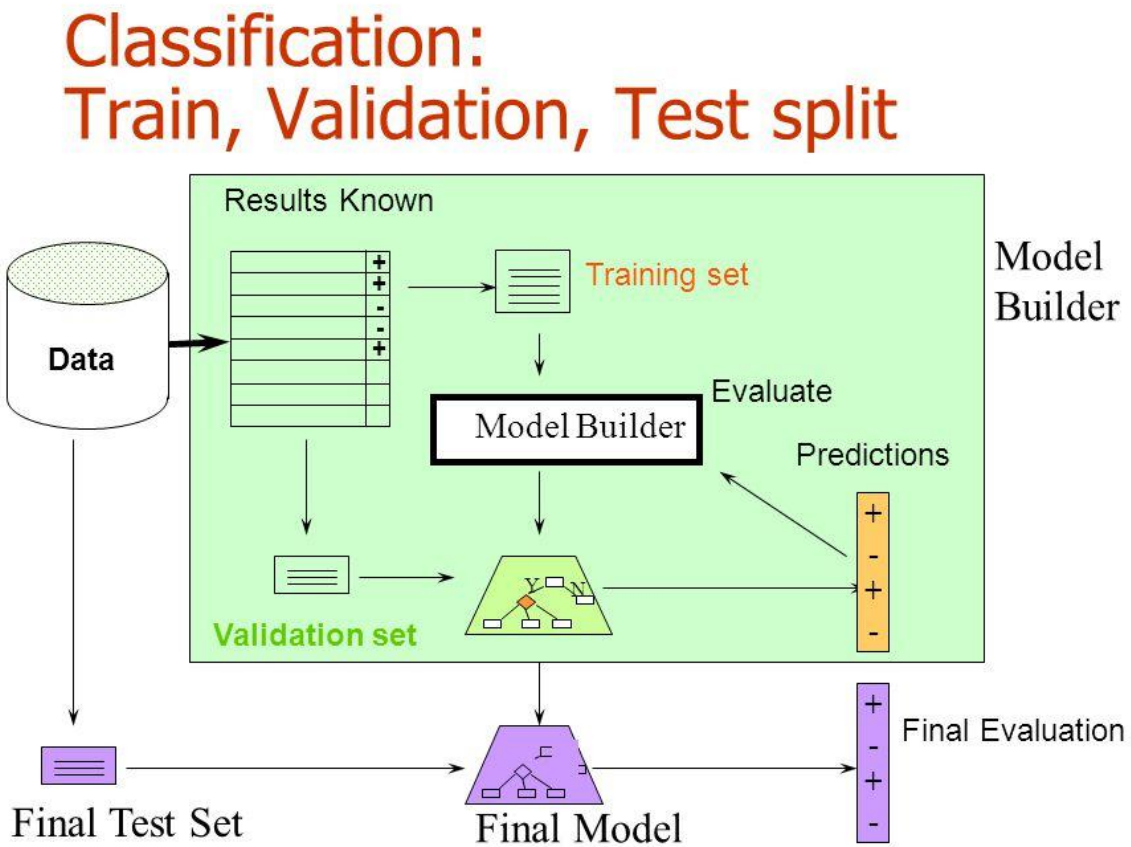
Selección de Atributos

- Encontrar el menor número de atributos que permitan caracterizar al corpus (demasiados atributos pueden producir sobreajuste, performance)

- Métodos Estadísticos (atributos muy correlacionados con la clase objetivo, poco correlacionados entre ellos): e.g. chi-square test (descartamos atributos independientes de la clase objetivo)

- Performance en el corpus de entrenamiento
- Evitar sobreajuste: validation set, cross-validation

Dataset de aprendizaje



O'Reilly, Evaluating Machine Learning Models

Ajuste de parámetros (model selection)

- Los modelos de aprendizaje generalmente incluyen *hiperparámetros*. ¿Cómo elegirlos?

- Ejemplo: Árboles de decisión
  - Criterio de selección de mejor atributo
  - Mayor profundidad de un árbol
  - Mínimo de ejemplos para una hoja
  - Máximo número de hojas
  - etc..

Ajuste de parámetros (model selection)

Utilizar un corpus de validación *diferente* al de entrenamiento (¿por qué)

- Corpus held-out
  - Separamos una parte del corpus de entrenamiento y lo utilizamos para evaluar
- Cross-Validation
  - Divido el corpus de entrenamiento en k partes (k-fold)
  - Entreno sobre (k-1) partes y evalúo en la restante
  - Repito para cada parte, y calculo la media

El (legendario) Iris Dataset

- 112 flores (al azar!) utilizadas para el conjunto de entrenamiento (80%)
- Las restantes, para evaluación (20%)
- Cross validation en el conjunto de entrenamiento (no tenemos corpus de evaluación)

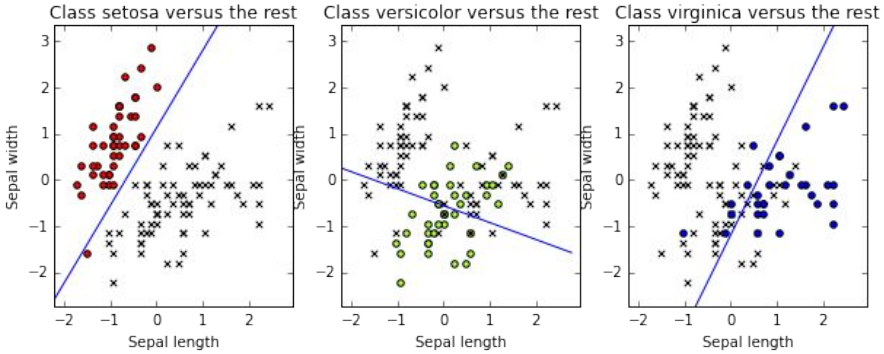
UCI Machine Learning Repository, Iris Data set

Aprendizaje y evaluación

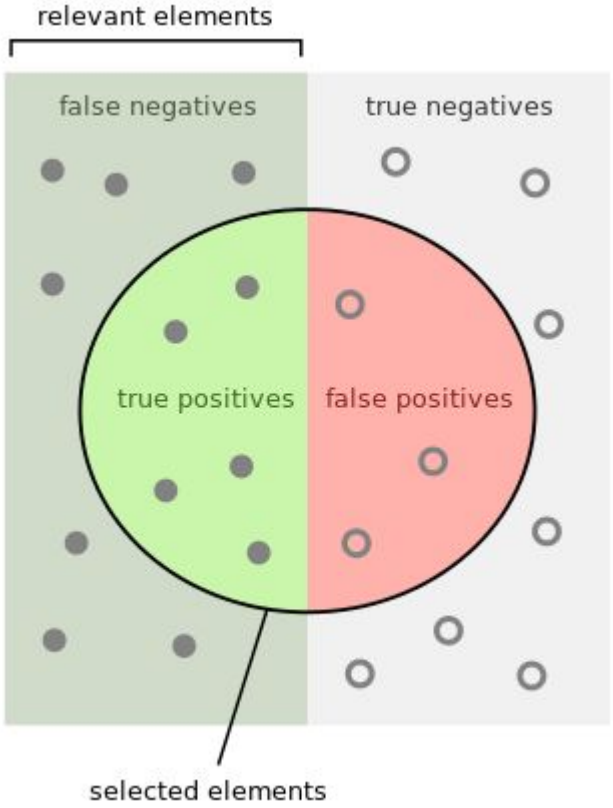
- Utilizamos el corpus de entrenamiento para generar un *modelo* (función de clasificación)
- Utilizamos la función de clasificación para calcular la clase objetivo para cada instancia del corpus de evaluación
- Evaluamos performance sobre el corpus de evaluación (que no vimos nunca antes!)

Iris Data Set

- Problema de clasificación multiclase: lo transformamos en tres problemas uno-contra-todos (¿es setosa? ¿es versicolor? ¿es virginica?). Elegimos la más votada, o la más "segura"(según el método)



¿Cómo mido la performance?



- Accuracy (Exactitud)
  - Fracción de las instancias clasificadas correctamente
- Precision (Precisión)
  - Instancias positivas que se clasificaron correctamente / Total de Instancias Positivas
  - TP / (TP + FP)
- Recall (Exhaustividad)
  - Fracción de las instancias positivas que se clasificaron correctamente / Fracción de las instancias que son positivas
  - TP/(TP + FN)
- Medida-F
  - Media armónica entre Precisión y Recall: 2*Precision*Recall/ (Precision + Recall)

Wikipedia, Precision and Recall

Iris Dataset

- Entrenamos un clasificador lineal sobre el corpus. (Detalles)
- Accuracy sobre el corpus de entrenamiento: 0.83 (**Esta información no es muy útil**)
- Accuracy sobre el corpus de evaluación: 0.68

Iris Dataset

- Matriz de confusión:

	setosa	versicolor	virginica
setosa	8	0	0
versicolor	0	3	8
virginica	0	4	15

- Reporte de clasificación (P/R/F)
  - Setosa: 1.0/1.0/1.0 (support=8)
  - Versicolor: 0.43 (3/7)/0.27(3/11) / 0.33 (support=11)
  - Virginica: 0.65 / 0.79 /0.71 (support=19)
  - Promedio: 0.66 / 0.68 / 0.66

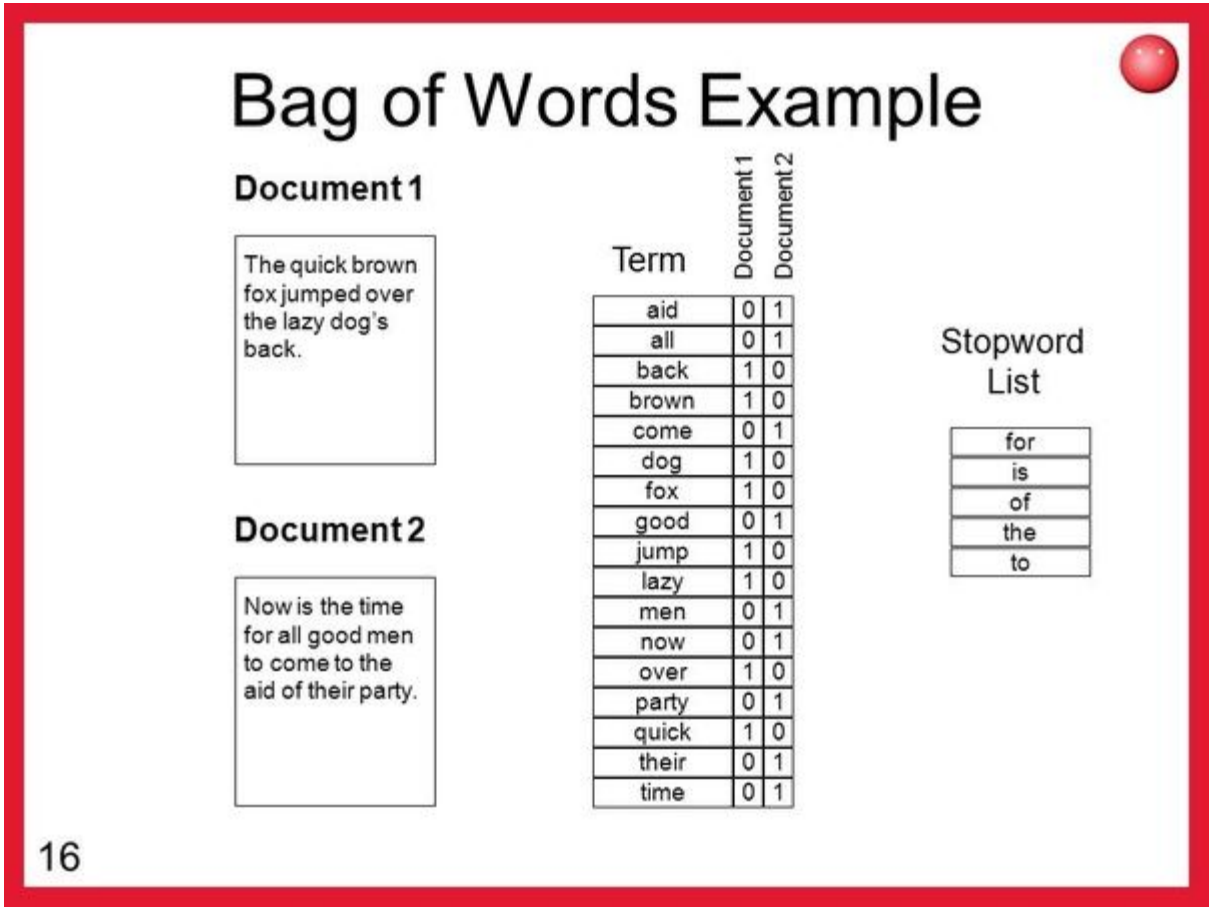
Garreta, Moncecchi, "Learning Scikit-learn Machine Learning in Python" notebooks (capítulos 1 y 2)

Aplicaciones

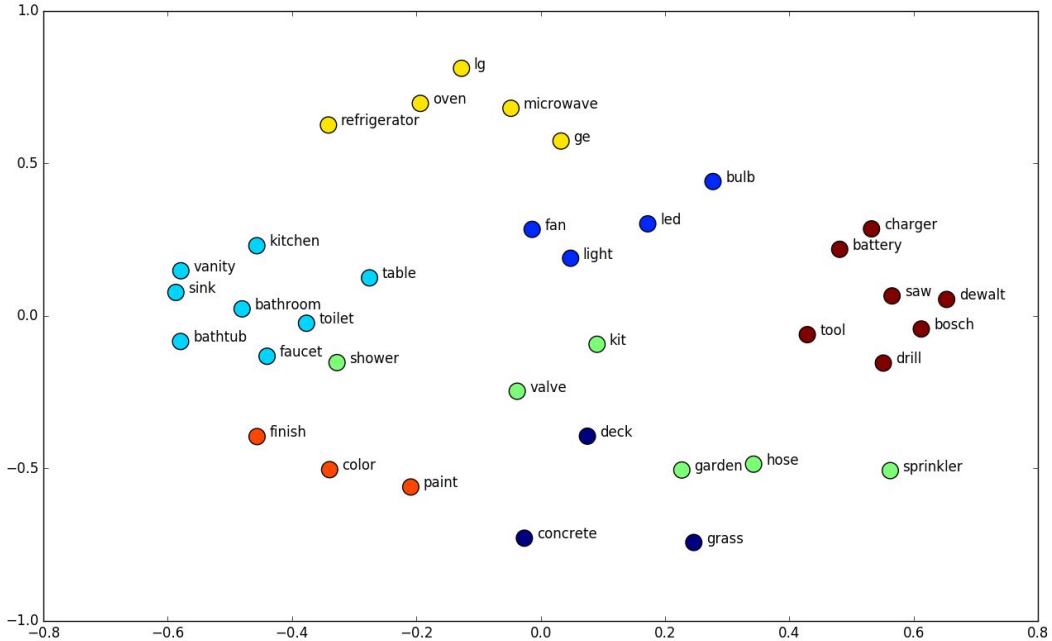
Procesamiento de Lenguaje Natural / Traducción Automática

- Elegir la mejor traducción en el idioma destino de una frase en el idioma origen
- Google Translate

Procesamiento de Lenguaje Natural / Clasifiación de Documentos



Procesamiento de Lenguaje Natural / Word embeddings



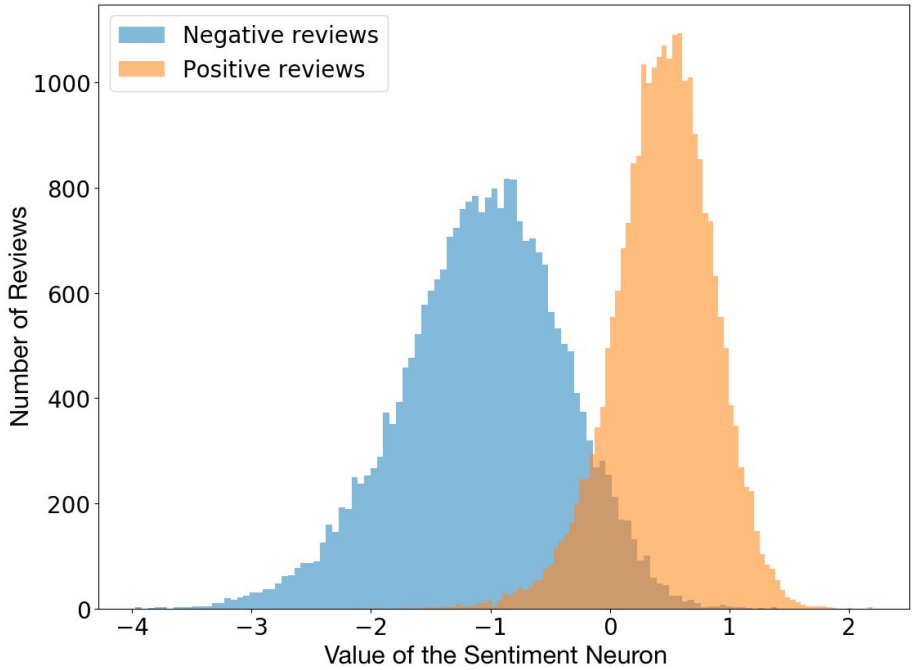
Shane Lynn, "Get busy with word embeddings"

Procesamiento de Lenguaje Natural / Generación de textos

```
PANDARUS:
Alas, I think he shall be come approached and the day
When little srain would be attain'd into being never fed,
And who is but a chain and subjects of his death,
I should not sleep.

Second Senator:
They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.

DUKE VINCENTIO:
Well, your wit is in the care of side and that.

Second Lord:
They would be ruled after this chamber, and
my fair nues begun out of the fact, to be conveyed,
Whose noble souls I'll have the heart of the wars.

Clown:
Come, sir, I will make did behold your worship.

VIOLA:
I'll drink it.
```

Andrej Karpathy,The Unreasonable Effectiveness of Recurrent Neural Networks

Procesamiento de Lenguaje Natural / The sentiment neuron

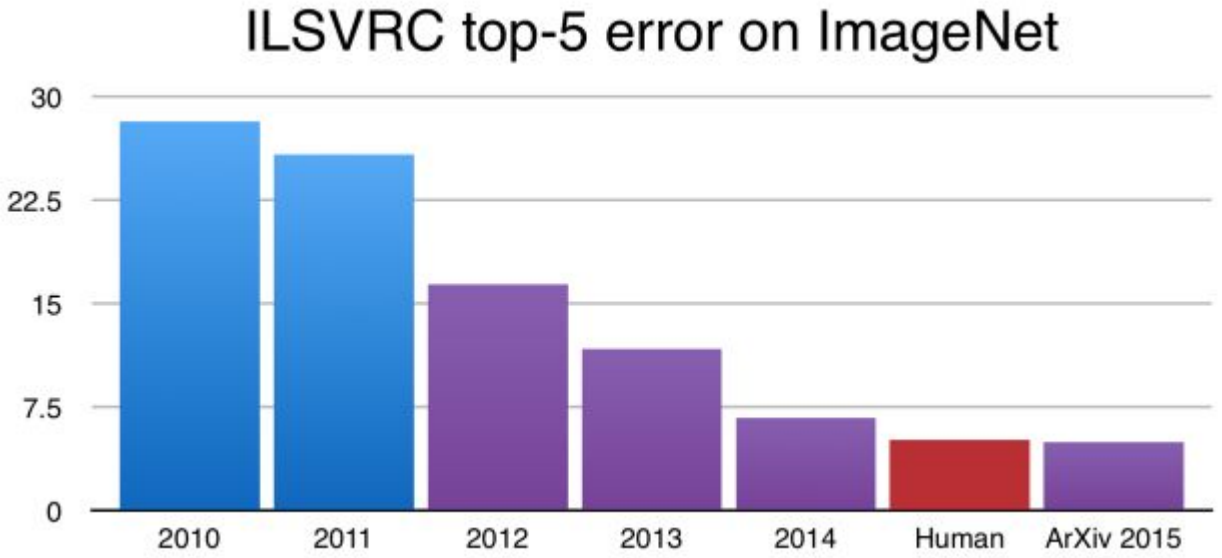


OpenAI, The Unsupervised Sentiment Neuron

Procesamiento de Lenguaje Natural / The sentiment neuron

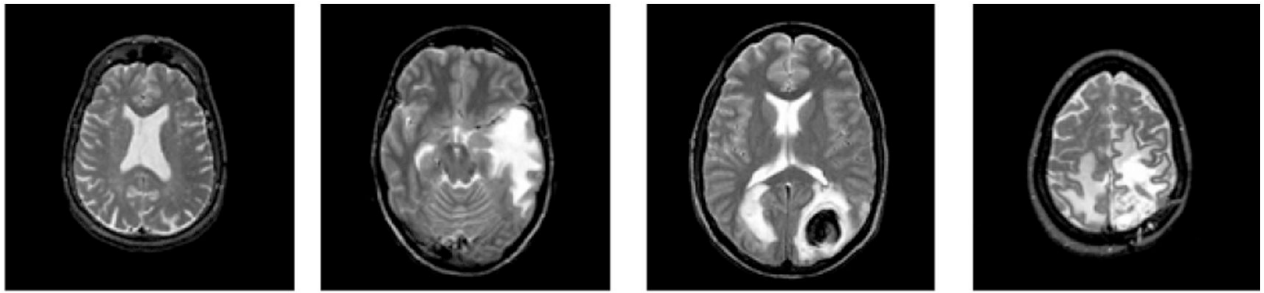
SENTIMENT FIXED TO POSITIVE	SENTIMENT FIXED TO NEGATIVE
I couldn't figure out the shape at first but it definitely does what it's meant to do. It's a great product and I recommend it highly	I couldn't figure out how to use the product. It did not work.At least there was no quality control; this tablet does not work. I would have given it zero stars, but that was not an option.
I couldn't figure out why this movie had been discontinued! Now I can enjoy it anytime I like. So glad to have found it again.	I couldn't figure out how to set it up being that there was no warning on the box. I wouldn't recommend this to anyone.
I couldn't figure out how to use the video or the book that goes along with it, but it is such a fantastic book on how to put it into practice!	I couldn't figure out how to use the gizmo. What a waste of time and money. Might as well through away this junk.

Reconocimiento de Imágenes



Otium, Performance Trends in AI, ImageNet

Medicina/Reconocimiento de tumores de cerebro



(a) Normal (b) Metastatic bronchogenic carcinoma (c) Sarcoma (d) Glioblastoma

Mohsen et al. Classification using deep learning neural networks for brain tumors

Medicina/Reconocimiento de tumores de cerebro

Table 1. Performance of DNN, KNN K = 1 and 3, LDA and SMO classifiers.

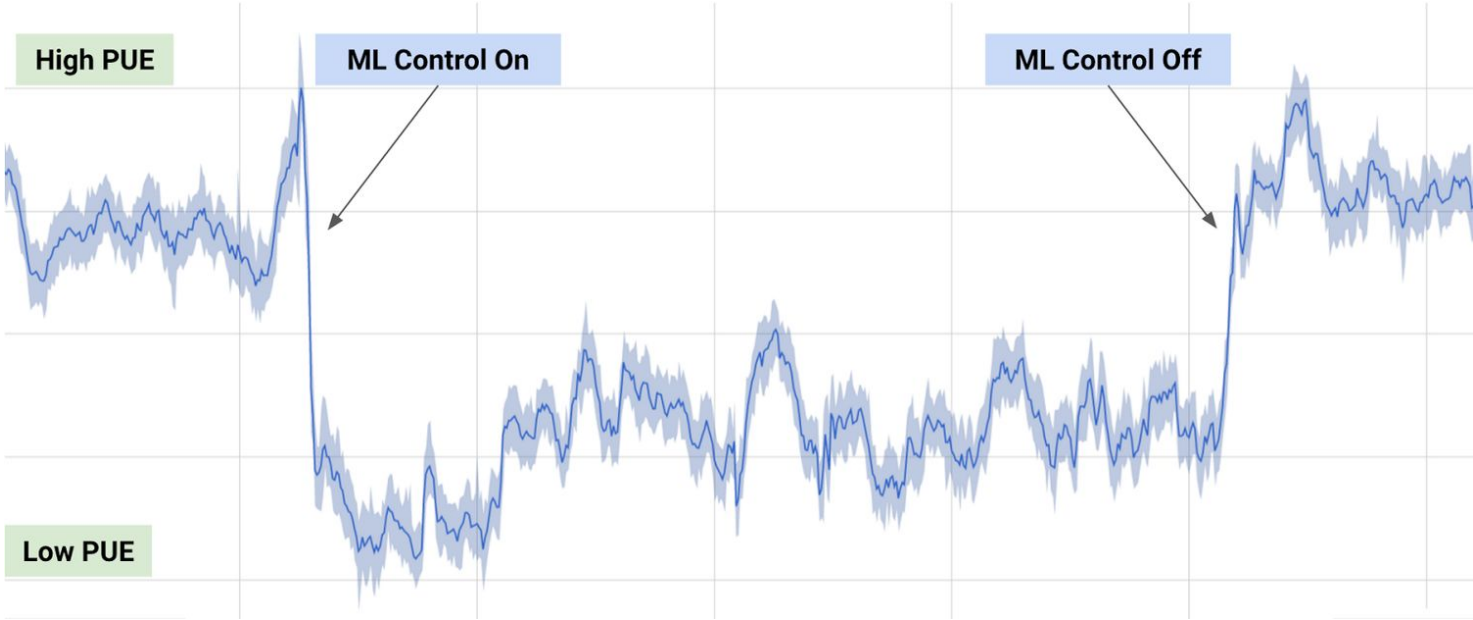
Algorithm	Classification rate	Recall	Precision	F-Measure	AUC (ROC)
DNN	96.97%	0.97	0.97	0.97	0.984
KNN K = 1	95.45%	0.955	0.956	0.955	0.967
KNN K = 3	86.36%	0.864	0.892	0.866	0.954
LDA	95.45%	0.955	0.957	0.955	0.983
[a]SMO	93.94%	0.939	0.941	0.963	0.939

Mohsen et al. Classification using deep learning neural networks for brain tumors

Agro/Clasificación de pepinos

Kahz Sato, How a Japanese cucumber farmer is using deep learning and TensorFlow

Energía/ Consumo de DataCenter



Deep Mind, DeepMind AI Reduces Google Data Centre Cooling Bill by 40%

Fondo de Datos / MIEM-ANII - UTE - Antel

- Mantenimiento predictivo de aerogeneradores guiado por el análisis de datos de UTE sobre parques eólicos (UdelaR)
- Predicción de coliformes fecales en playas capitalinas de interés turístico (CURE - UdelaR)
-Predicción de función de genes mediante aprendizaje automático (MEC - IIBCE)
- Modelado probabilístico basado en datos para el análisis de riesgo de déficit de suministro en el área metropolitana de Montevideo (Universidad de Montevideo)
- Herramientas para predicción del rendimiento del cultivo de arroz en condiciones productivas (FAgro - UdelaR)

Tecnología

- Python, R, Julia (Lenguajes de programación)
- numpy, scipy (Análisis numérico)
- matplotlib (Visualización)
- pandas (Análisis de Datos)
- scikit-learn (Machine Learning)
- pytorch, TensorFlow, Keras (Deep Learning)
- nltk, spacy (Natural Language Processing)

- Anaconda
- Jupyter notebooks

- Google: TensorFlow
- Facebook: pytorch, caffe
- Microsoft: CNTK
- Yahoo: yamall
- Salesforce: TransmogrifAI

- Tryolabs: luminoth
- Xmartlabs: Bender

¿Me va a servir de algo la ciencia de datos?

1. ¿Conozco mi problema?
2. ¿Puedo caracterizarlo como un problema resoluble con aprendizaje automático?
3. ¿Tengo datos o puedo generarlos?
4. ¿Sé programar? ¿Sé programar utilizando aprendizaje automático?

Si a todas las preguntas la respuesta es sí, entonces la respuesta es:
"**Probablemente, sí**"

Gracias.