Introducción a la Ciencia de Datos

Curso de posgrado
IIE/IMERL/INCO
Facultad de Ingeniería, UdelaR
2022

UNIVERSIDAD
DE LA REPUBLICA
URUGUAY

Introducción a la Ciencia de Datos

Docentes: Lorena Etcheverry (INCO)
María Inés Fariello (IMERL)
Marcelo Fiori (IMERL)
Guillermo Moncecchi (INCO)

Introducción a la Ciencia de Datos

Docentes: Lorena Etcheverry (INCO)
María Inés Fariello (IMERL)
Marcelo Fiori (IMERL)
Guillermo Moncecchi (INCO)

Contacto: introcd@fing.edu.uy

Introducción a la Ciencia de Datos

Objetivos:

- Panorama global de Ciencia de Datos y ramas afines
- Manejo del lenguaje y terminología general del área
- Elementos prácticos, teóricos y metodológicos para encarar un problema

Introducción a la Ciencia de Datos

Objetivos:

- Formular una estrategia de solución, incluyendo:
  - Procesado y gestión de datos
  - Formulación del problema
  - Exploración de los datos
  - Modelado
  - Implementación
  - Evaluación
  - Análisis crítico de los resultados.

Introducción a la Ciencia de Datos

Objetivos:

- Formular una estrategia de solución
- Identificar problemas en los datos como ser datos faltantes, anomalías, redundancia.

Introducción a la Ciencia de Datos

Objetivos:

- Formular una estrategia de solución
- Identificar problemas en los datos como ser datos faltantes, anomalías, redundancia.
- Comprender conceptos básicos de aprendizaje supervisado y no supervisado, incluyendo:
  - Sobreajuste
  - Poder de generalización
  - Separación de datos en entrenamiento, validación y test
  - La importancia de las características.

Introducción a la Ciencia de Datos

Objetivos:

- Formular una estrategia de solución
- Identificar problemas en los datos como ser datos faltantes, anomalías, redundancia.
- Comprender conceptos básicos de aprendizaje supervisado y no supervisado
- Identificar problemas éticos, incluyendo implicancias de privacidad, sesgo, y toma de decisiones automáticas, entre otros.

Evaluación

Trabajos/ejercicios en cada módulo.

Trabajo integrador final.

Extra

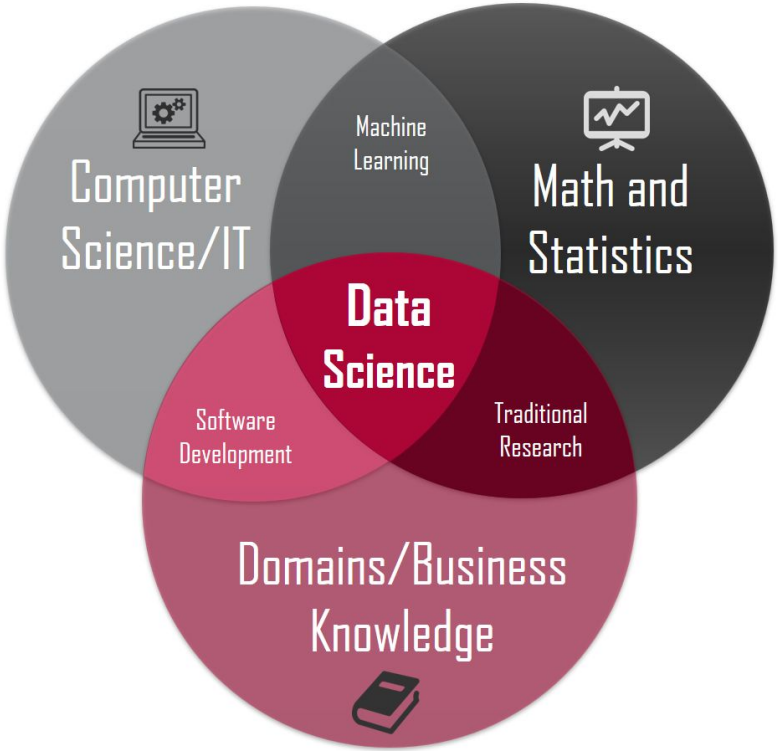
En el camino, tendremos laboratorios y aprenderemos algunas herramientas muy útiles:

Extra

En el camino, tendremos laboratorios y aprenderemos algunas herramientas muy útiles:

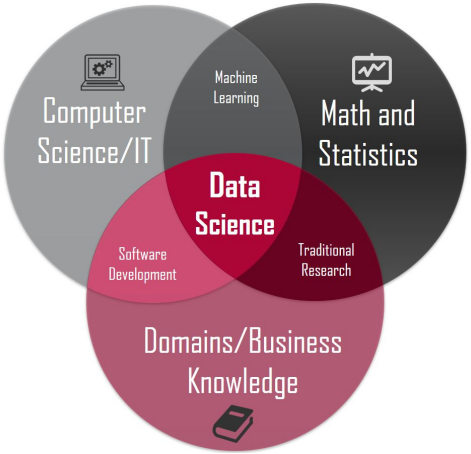
- Notebooks
- Github
- Numpy
- Pandas
- Matplotlib/Seaborn

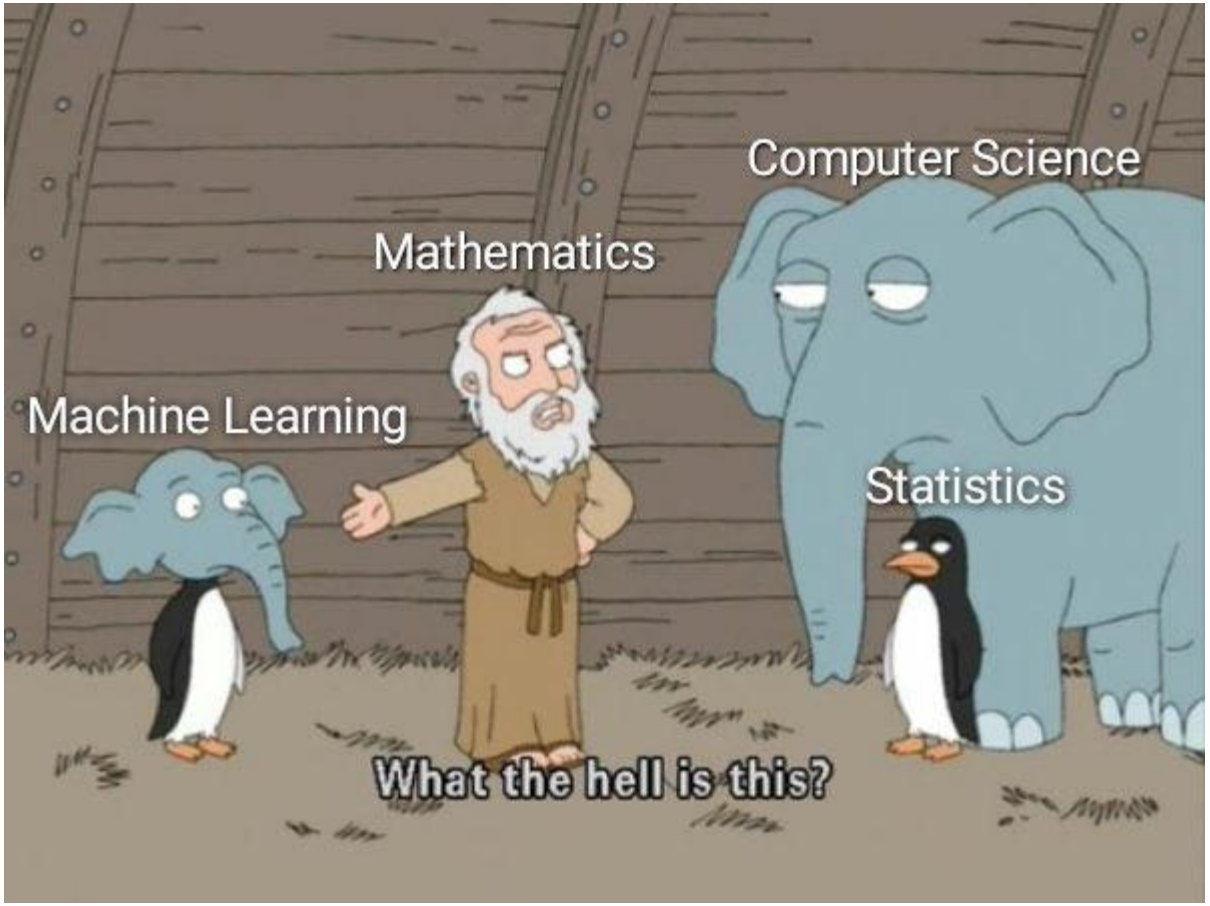
¿Qué es la Ciencia de Datos?



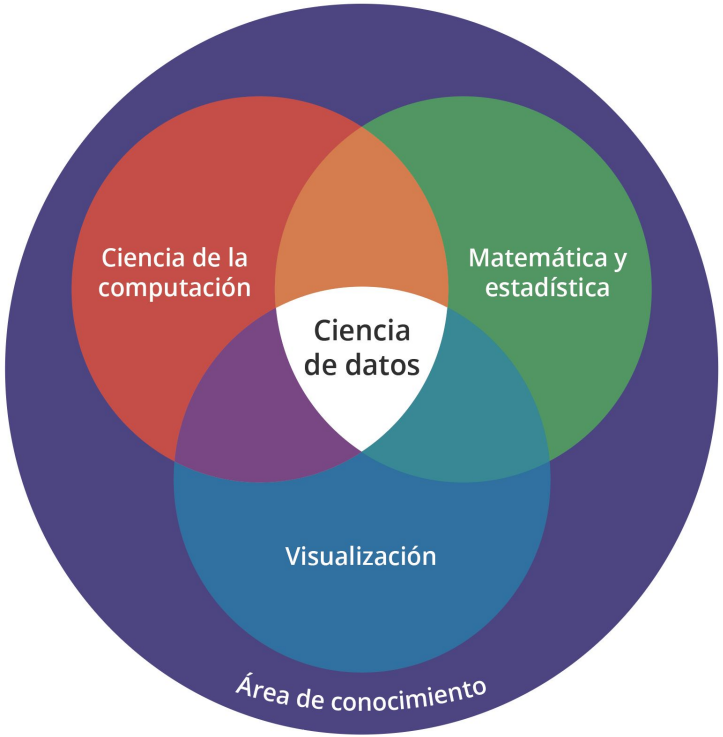
Drew Conway, The Data Science Venn Diagram.

¿Qué es la Ciencia de Datos?





¿Qué es la Ciencia de Datos?



¿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, UdelaR

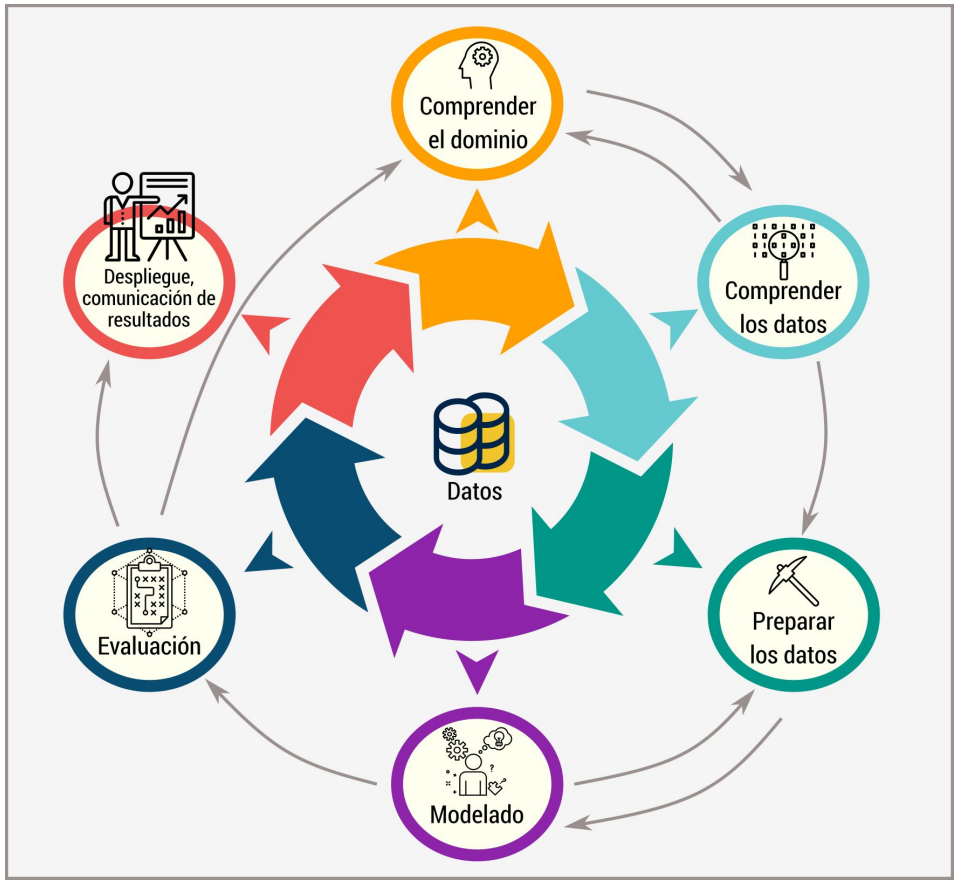
Quora, what is Data Science?

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

Cross Industry Standard Process for Data Mining (CRISP-DM)



Más info
https://the-modeling-agency.com/crisp-dm.pdf

¿Qué quiere decir...?

- Gestión de datos
- Exploración visual
- Modelado
- Implementación
- Evaluación

Back-end data science

Data is at the base! Without data there is no ML, there is no piece of information to be extracted, basically there is nothing. This is an overlooked aspect, especially from people who are used to *"clean room ML and data analysis"*. Data must be managed, must be reliable and must be informative. Which data you choose to store, how you store it and how you manage it affects:

- How effective you are in your analysis (if you will succeed in the aforementioned goal of making data speak to you). Messy data management (which does not guarantee idempotency for example) will lead to disastrous and not significative results. Trust me I learnt this on my own skin.
- How efficient you are and how much fun you have. The more time you spend setting up your data management properly, the more time you can spend actually analysing the data and having fun with it.

The overlooked importance of backend data science by @lorepep | Medium

Gestión de datos

Los datos “de verdad” (*Real data*) suelen tener problemas:

- Datos faltantes
- Calidad de los datos
- Limpieza
- Integración

Gestión de datos

Los datos pueden tener naturaleza muy distinta

- Categóricos
- Numéricos

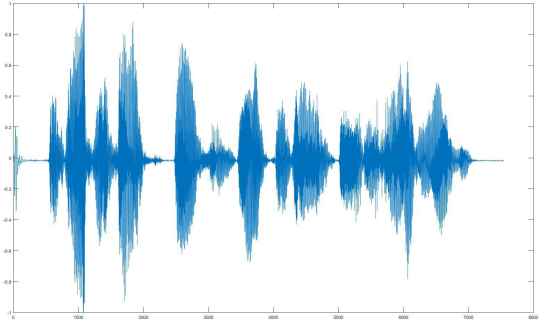
Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool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

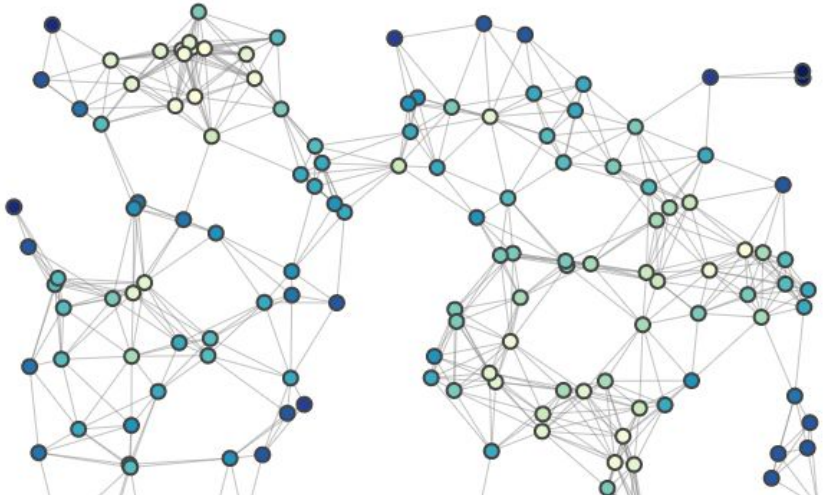
Height(Inches)	Weight(Pounds)
65.78	112.99
71.52	136.49
69.40	153.03
68.22	142.34
67.79	144.30
68.70	123.30
69.80	141.49
70.01	136.46
67.90	112.37
66.78	120.67
66.49	127.45
67.62	114.14
68.30	125.61
67.12	122.46
68.28	116.09

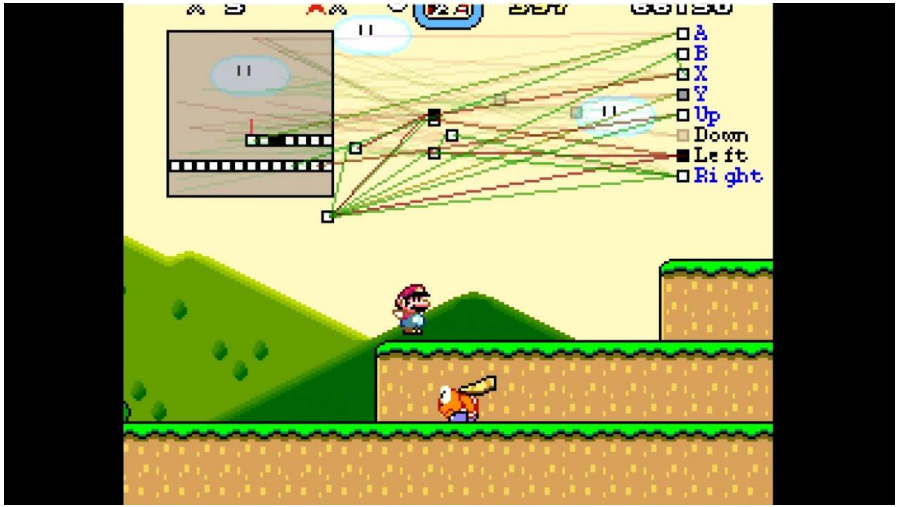
Gestión de datos

Los datos pueden tener naturaleza muy distinta



- Dominio “tradicional”
- no-Euclideo





Exploración visual de datos

En varias formas y etapas:

- Datos originales
- Características
- Resultados

Exploración visual de datos

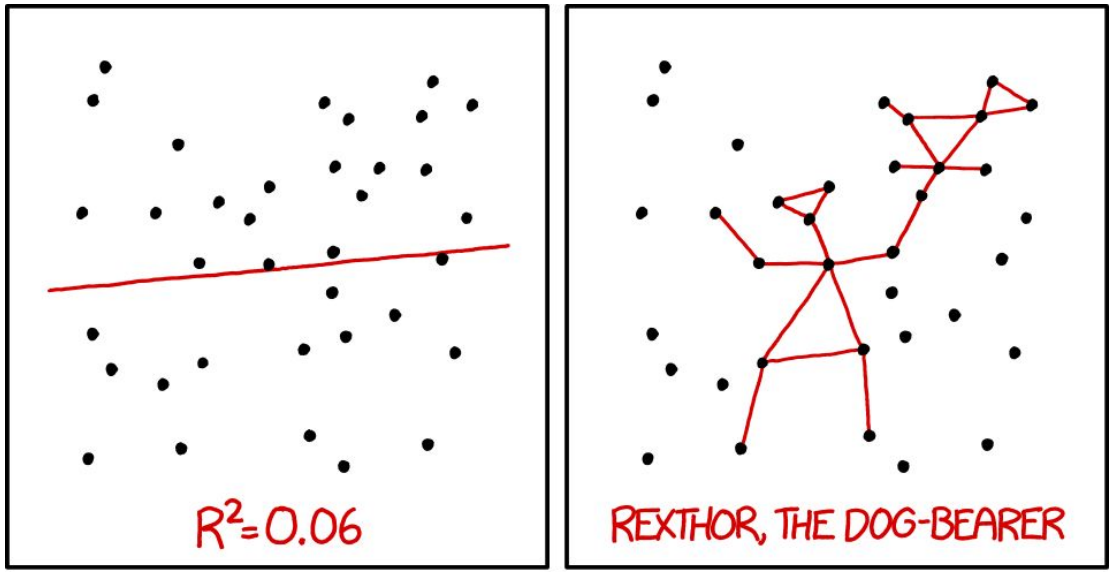
En varias formas y etapas:

- Datos originales
- Características
- Resultados

¿Tiene sentido lo que estamos haciendo?

Exploración visual de datos

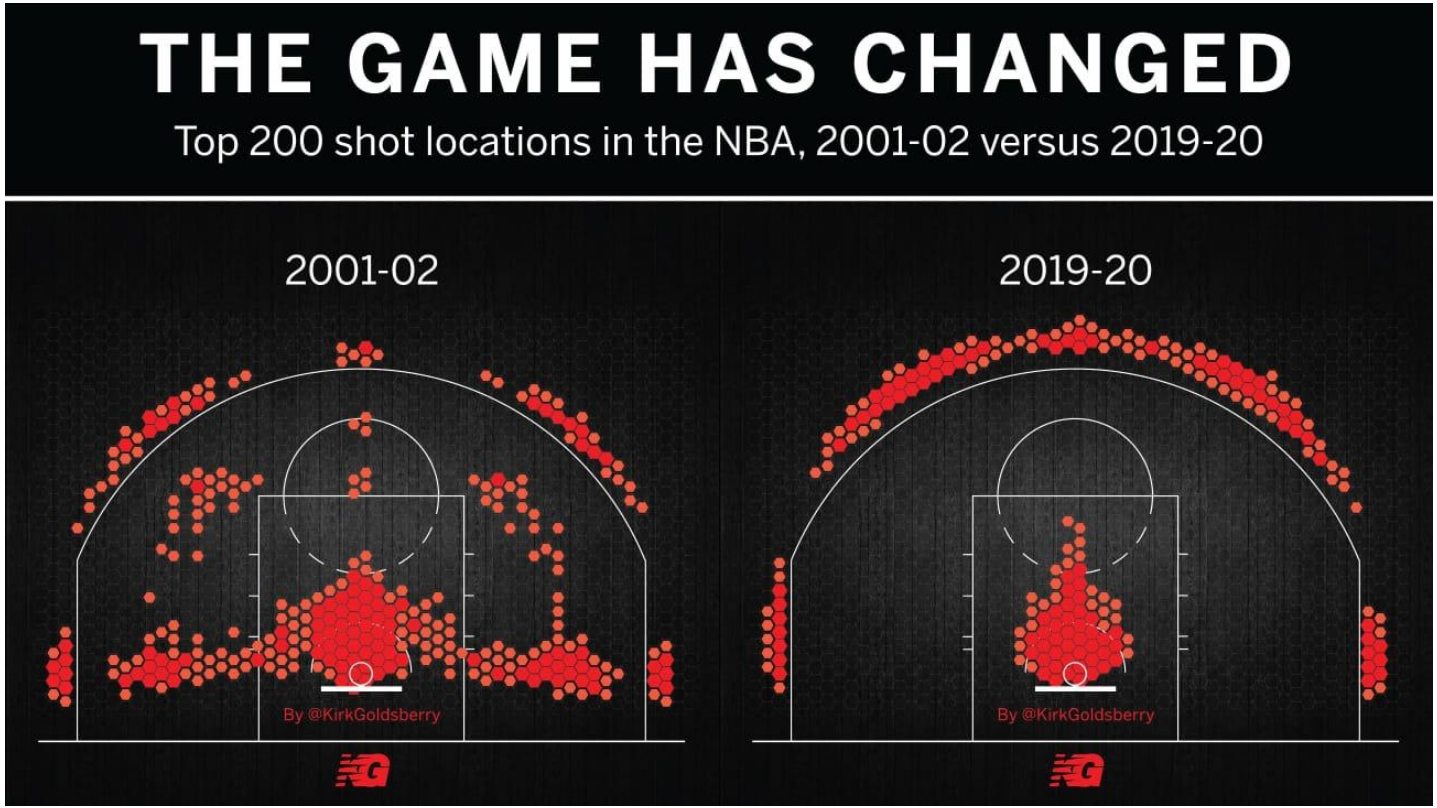
¿Tiene sentido lo que estamos haciendo?



xkcd.com

Exploración visual de datos

Comunicación



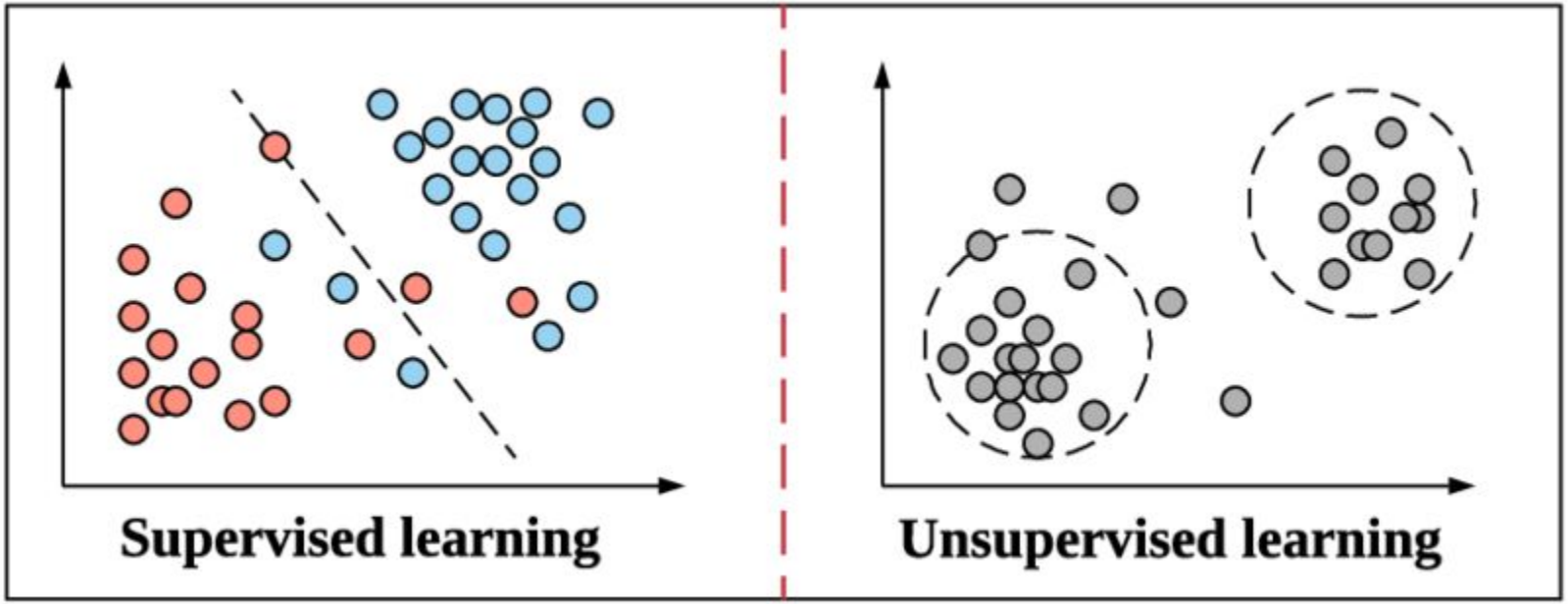
¿Qué tipos de problema encontramos?

¿Cuál es el objetivo? ¿Qué queremos hacer con los datos?

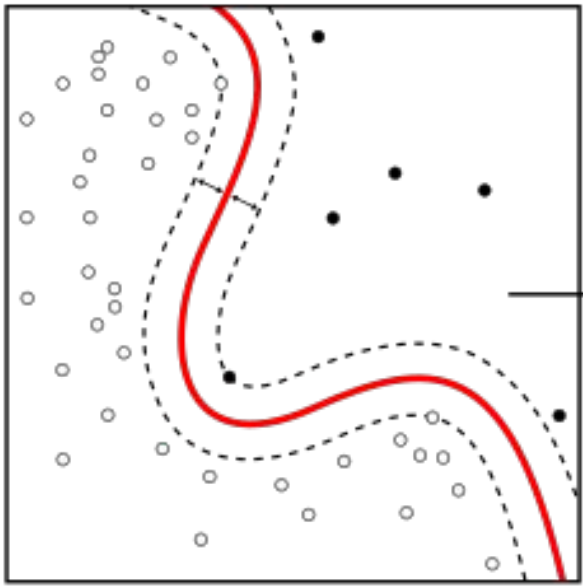
- Aprendizaje Supervisado (supervised learning)
    - Clasificación
    - Regresión
- Aprendizaje no supervisado (unsupervised learning)
    - Clustering
    - Reducción de dimensionalidad
- Aprendizaje por refuerzos (reinforcement learning)

- Active learning
- Semi-supervised learning

supervisado - no supervisado
(en 5 segundos)

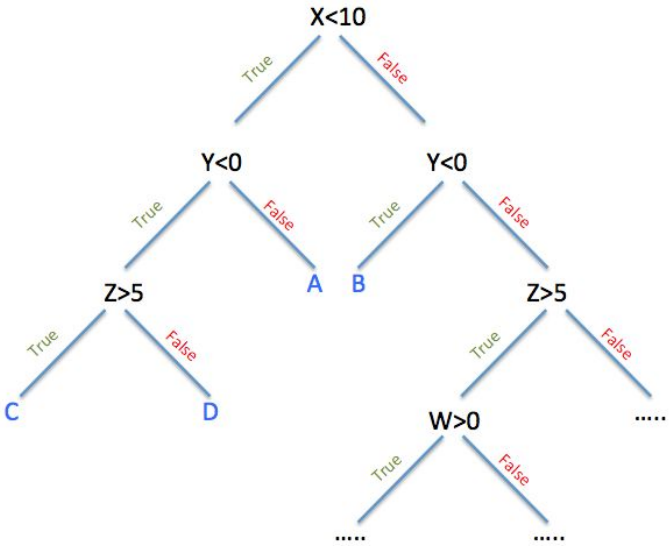


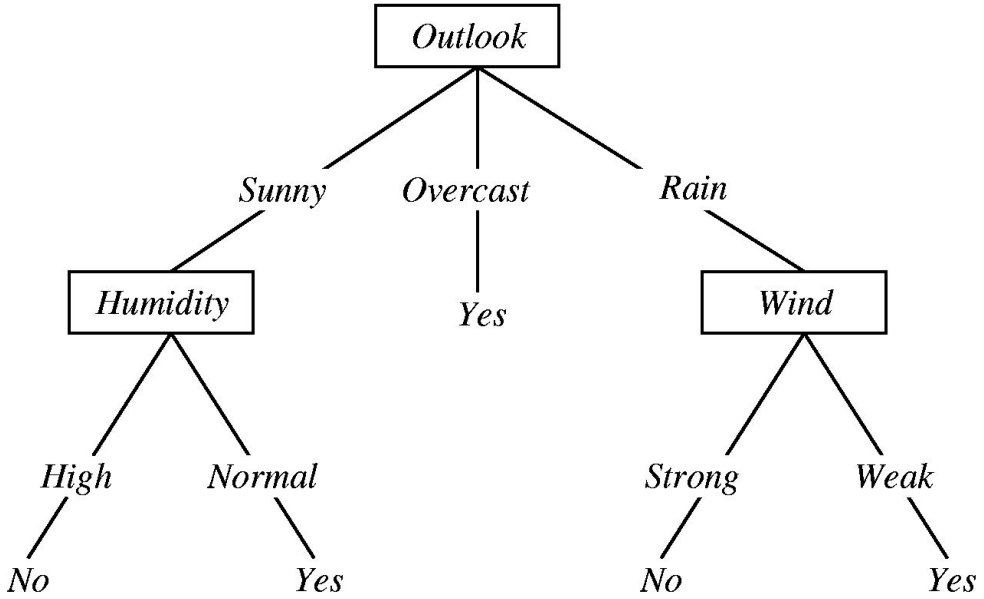
Aprendizaje Supervisado:
Clasificación



Aprendizaje supervisado: clasificación

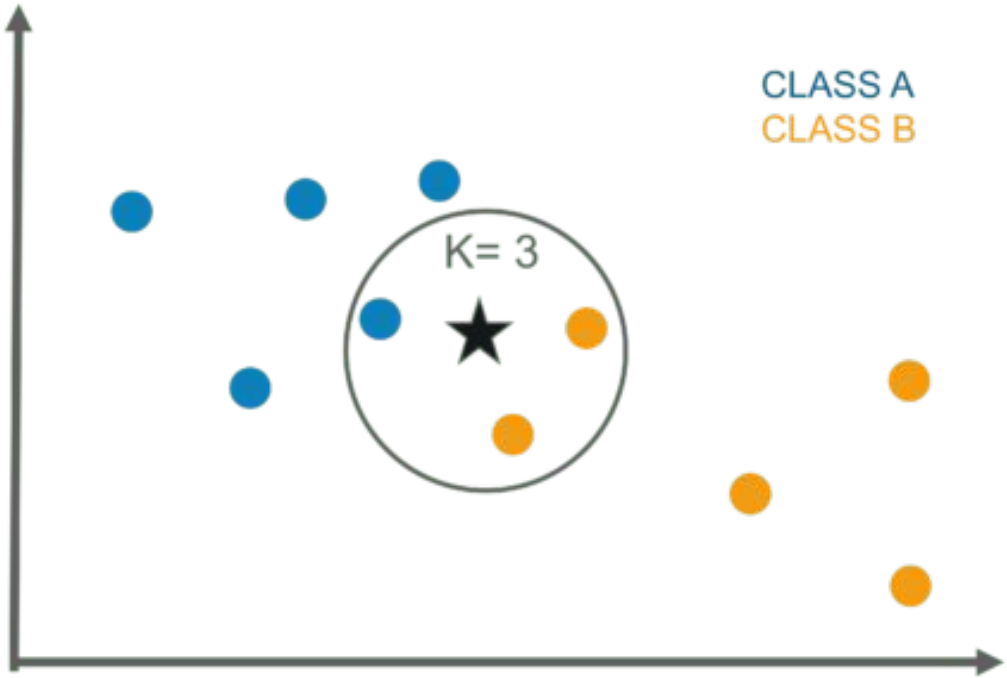
Árboles





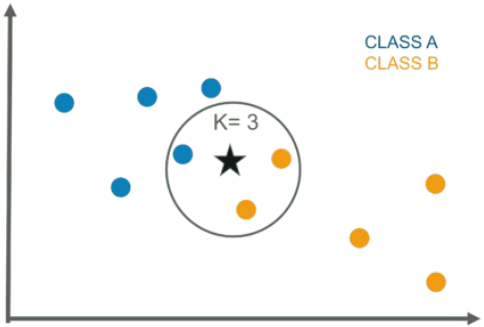
Aprendizaje supervisado: clasificación

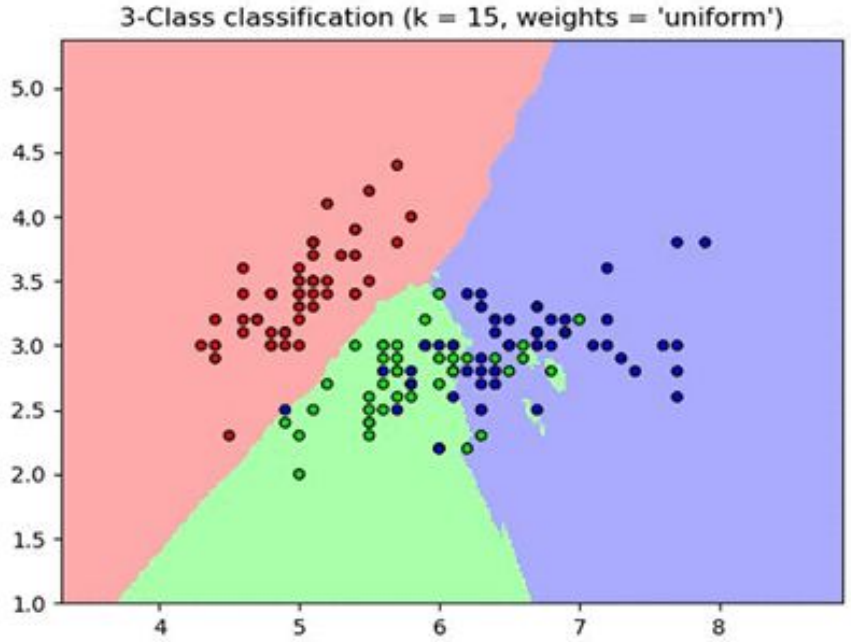
k-nearest neighbors

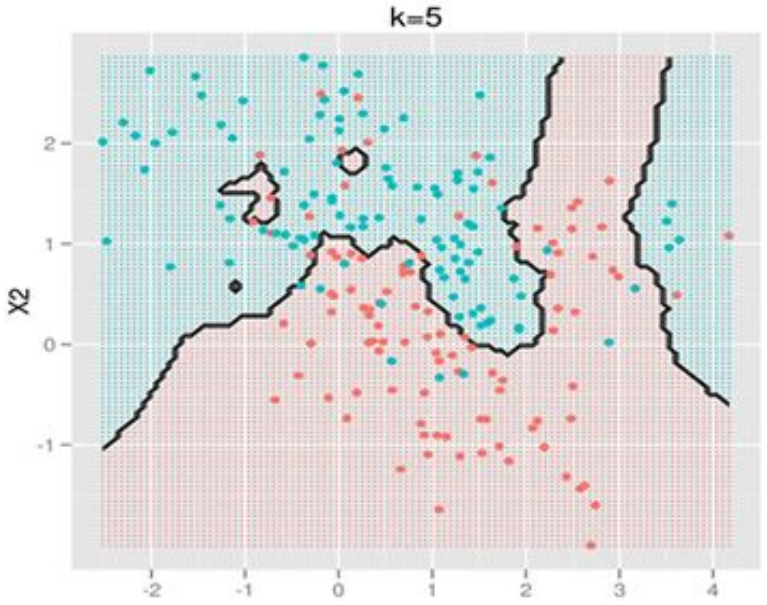


Aprendizaje supervisado: clasificación

k-nearest neighbors

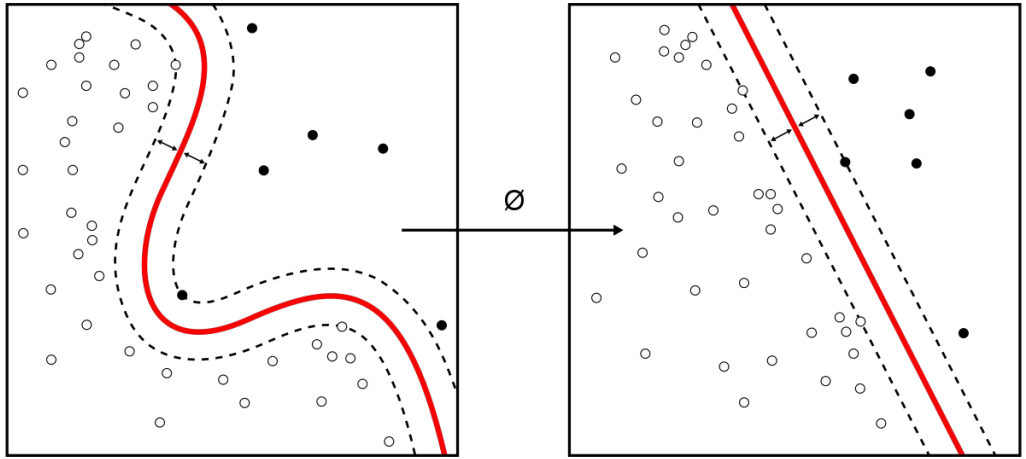


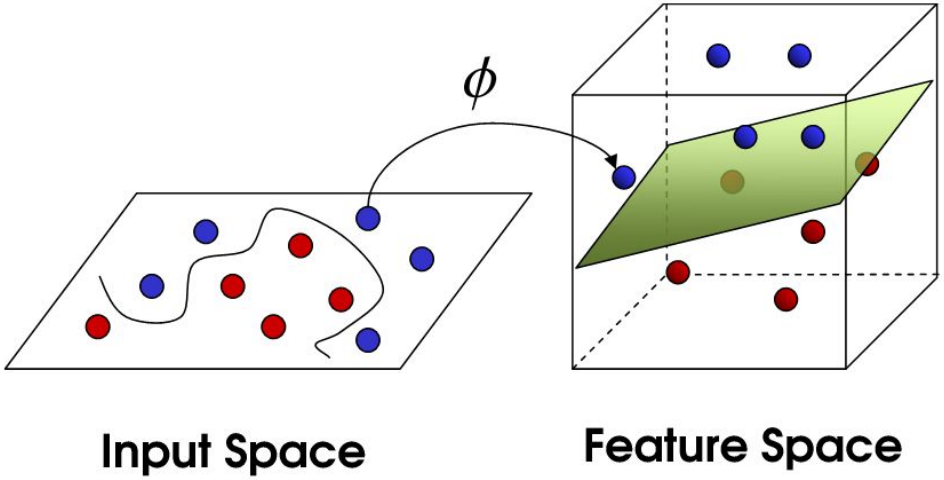




Aprendizaje supervisado: clasificación

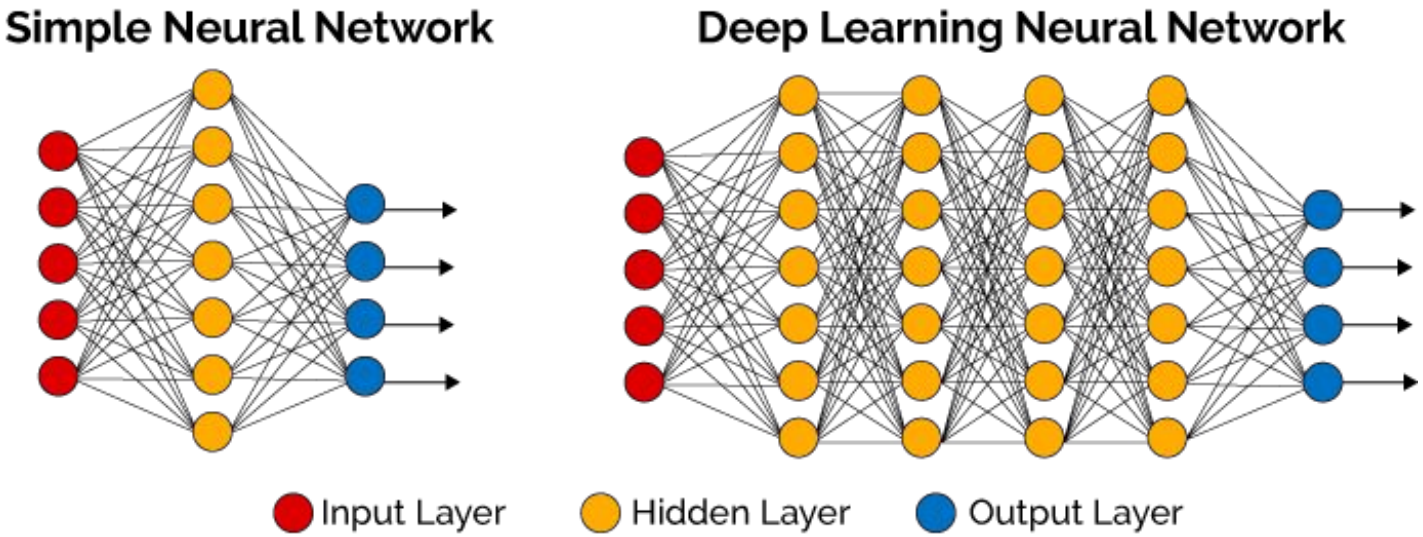
SVM





Aprendizaje supervisado: clasificación

redes neuronales



Aprendizaje Supervisado:
Regresión

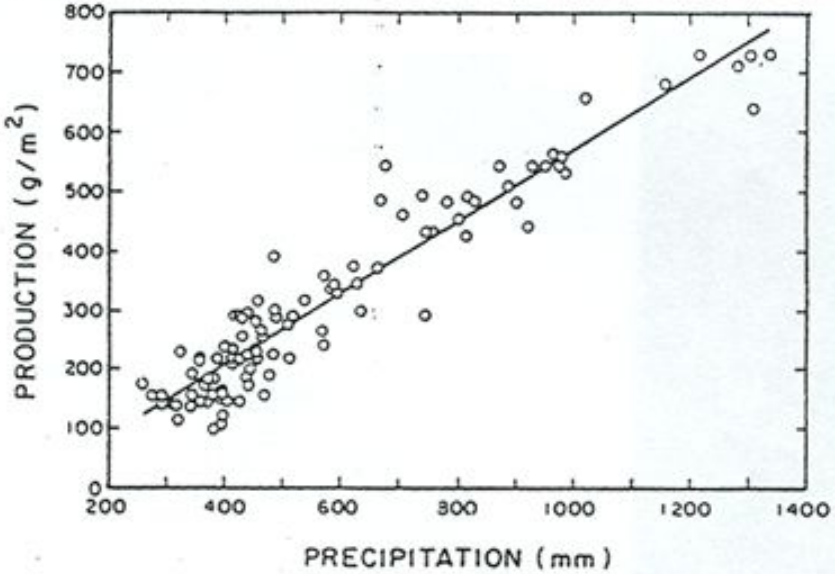
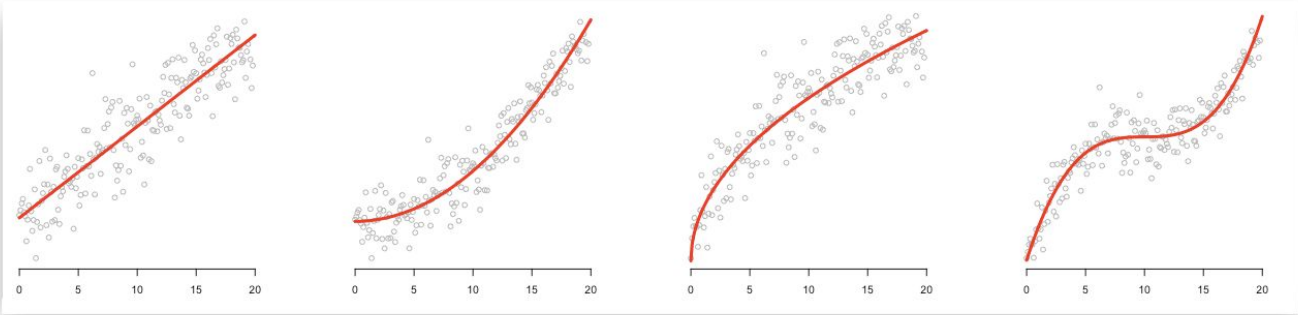


FIG. 2. Relationship between mean annual precipitation and mean aboveground net primary production (ANPP) for 100 major land resource areas across the Central Grassland region. ANPP = −34 + 0.6·APPT; $r^2 = 0.90$.

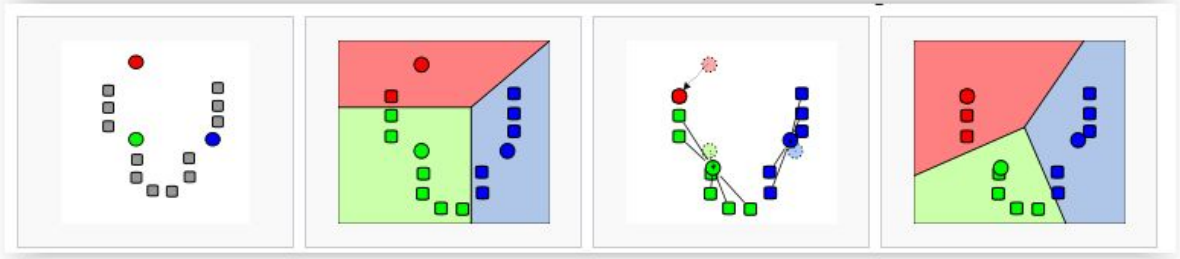
El agua como recurso limitante para el crecimiento de la vegetación

Aprendizaje Supervisado:
Regresión



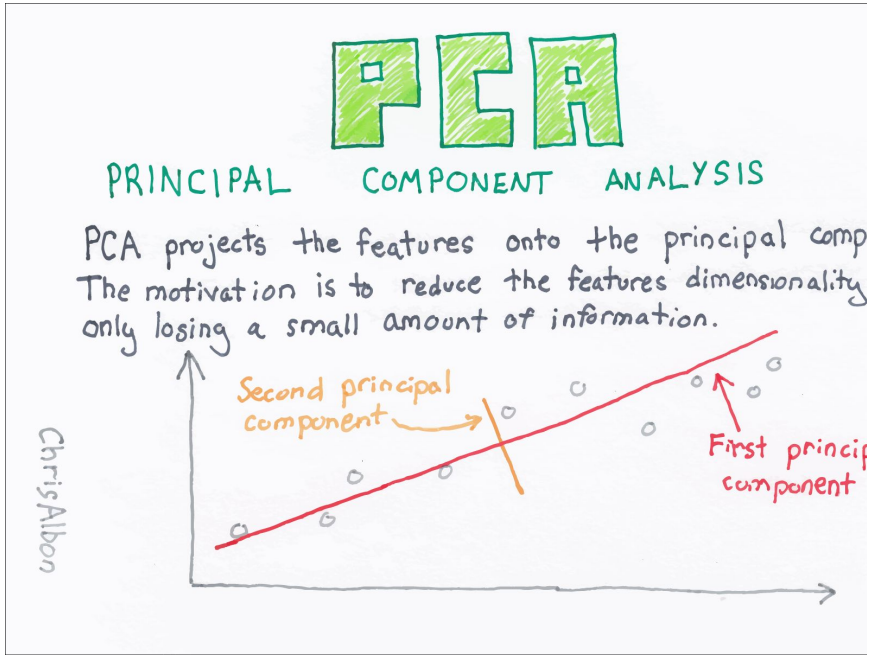
[How you can use linear regression models to predict quadratic, root, and polynomial functions]

Aprendizaje No supervisado/Clustering



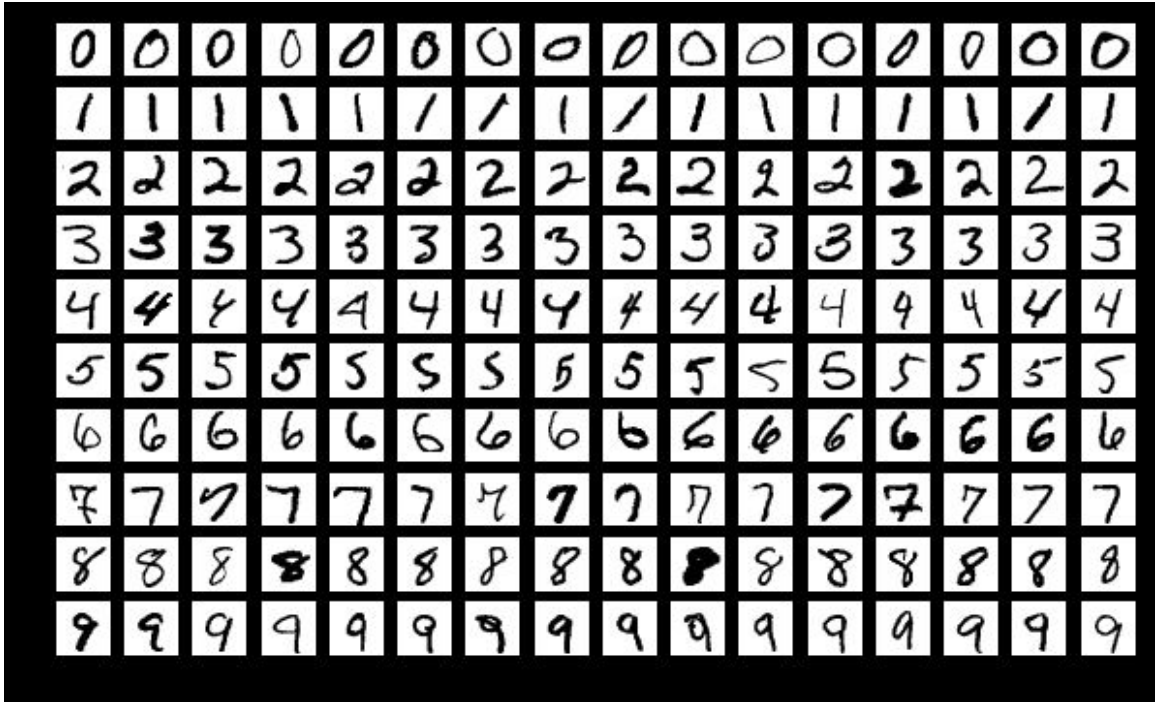
Wikipedia: k-means

Aprendizaje No supervisado
Reducción de Dimensionalidad



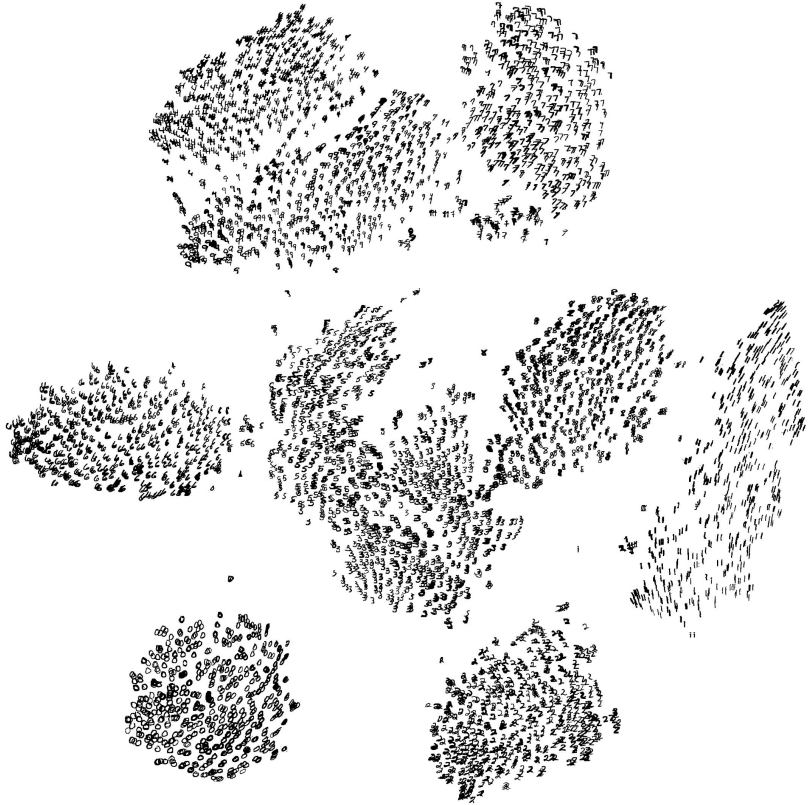
Chris Albon flashcards, [Dimensionality Reduction with PCA]

Aprendizaje No supervisado
t-sne



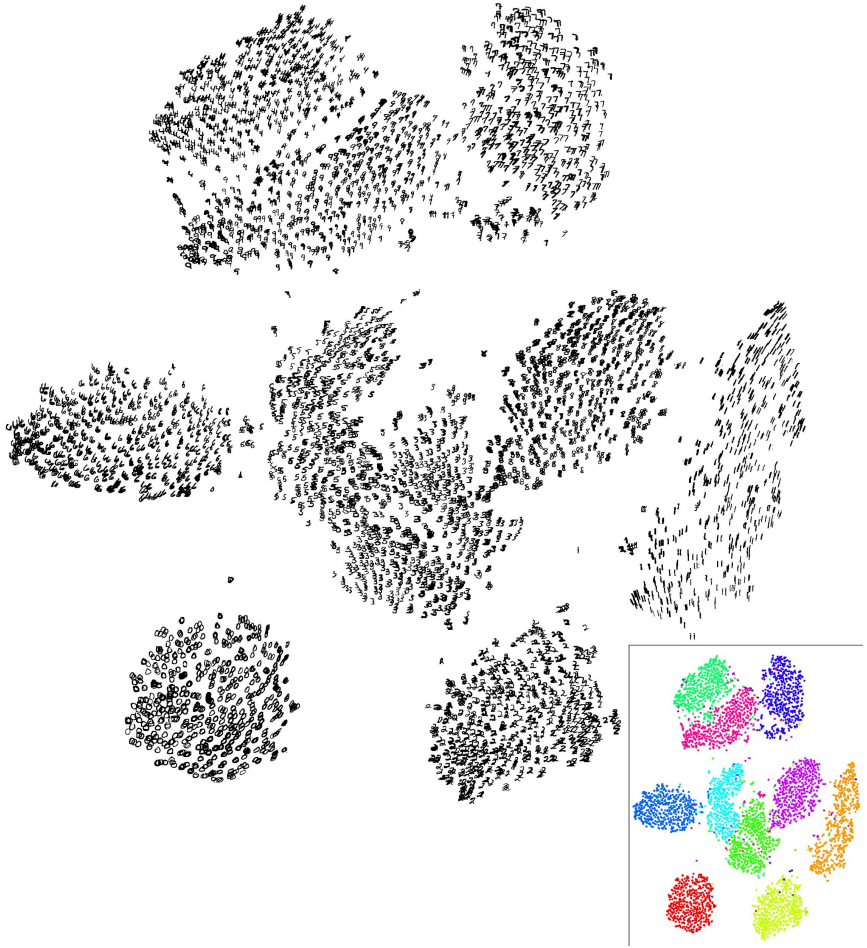
Yan LeCun et al.,The MNIST Database of handwritten digits

Aprendizaje No supervisado
t-sne



Laurens van der Maaten, t-sne

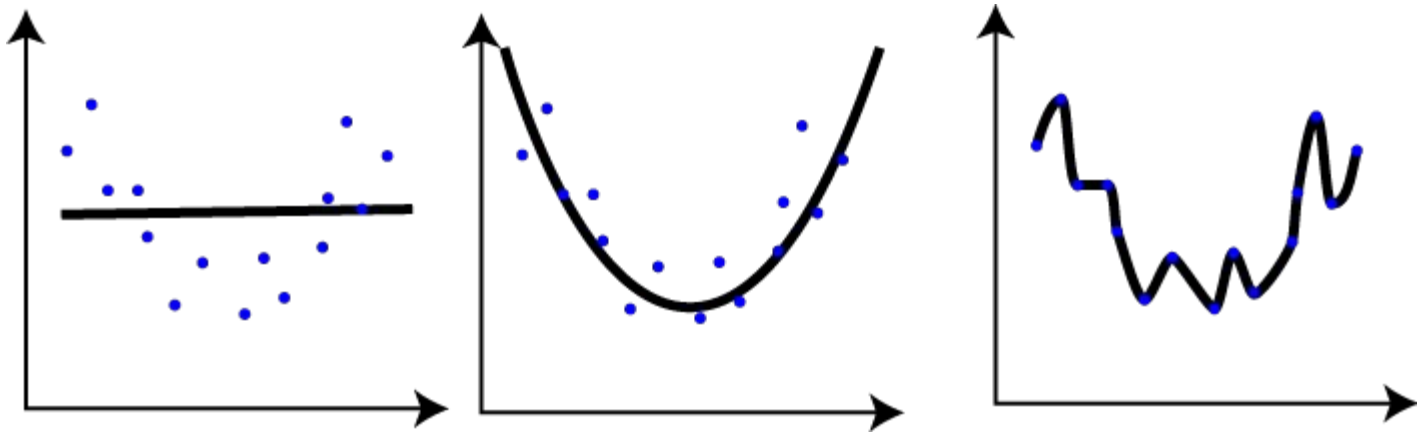
Aprendizaje No supervisado
t-sne



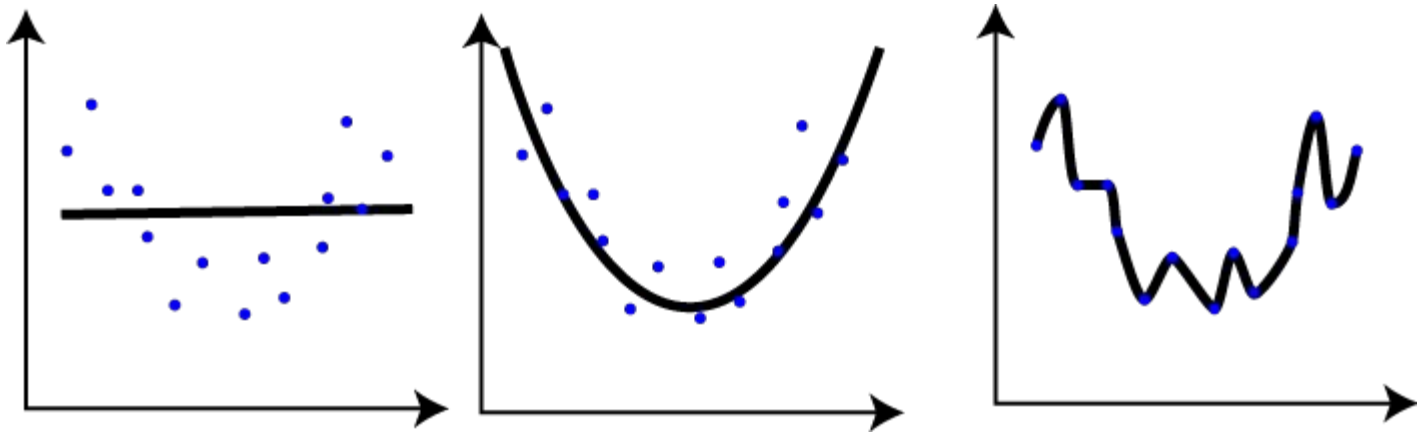
Laurens van der Maaten, t-sne

Underfit vs Overfit (sobreajuste).

Underfit vs Overfit.



Underfit vs Overfit.



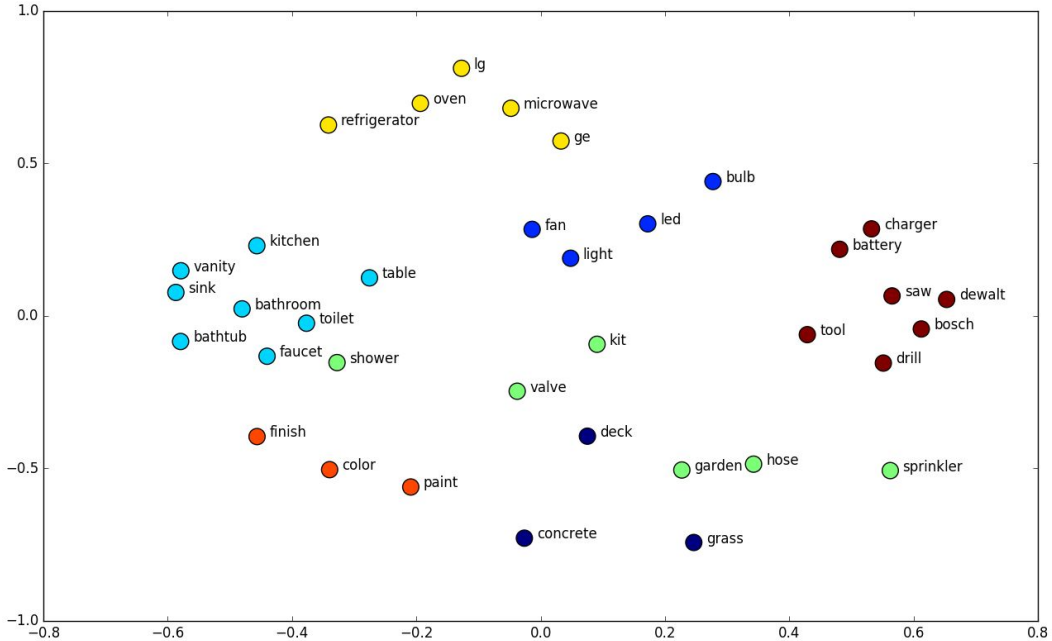
¿Cómo podemos usar los datos para evitarlo?

Ejemplos de 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 / Word embeddings



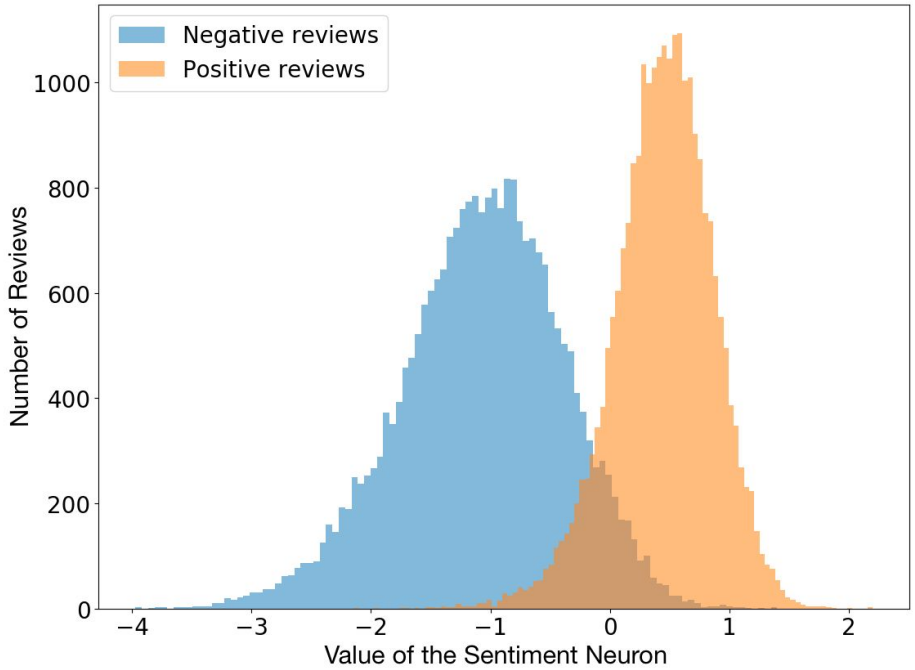
Shane Lynn, "Get busy with word embeddings"

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



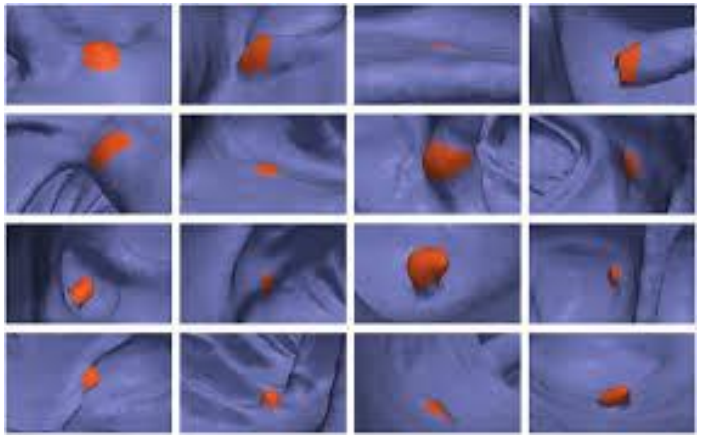
OpenAI, The Unsupervised Sentiment Neuron

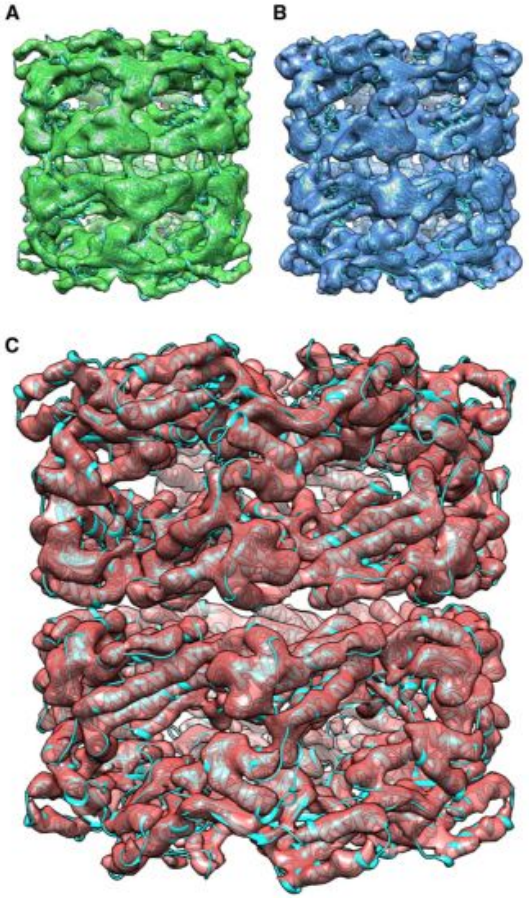
Video deblurring

Procesamiento de Señales/ Imágenes y video

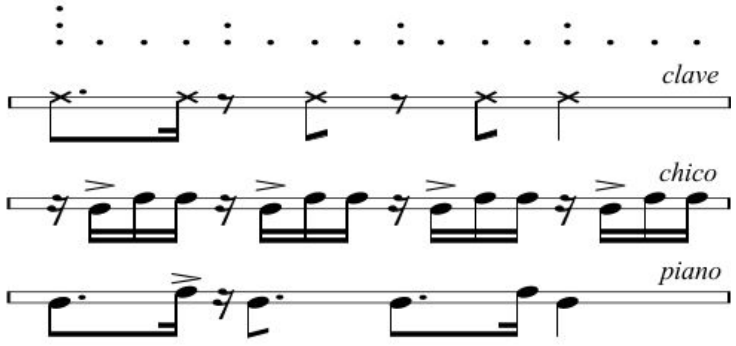
Imágenes médicas

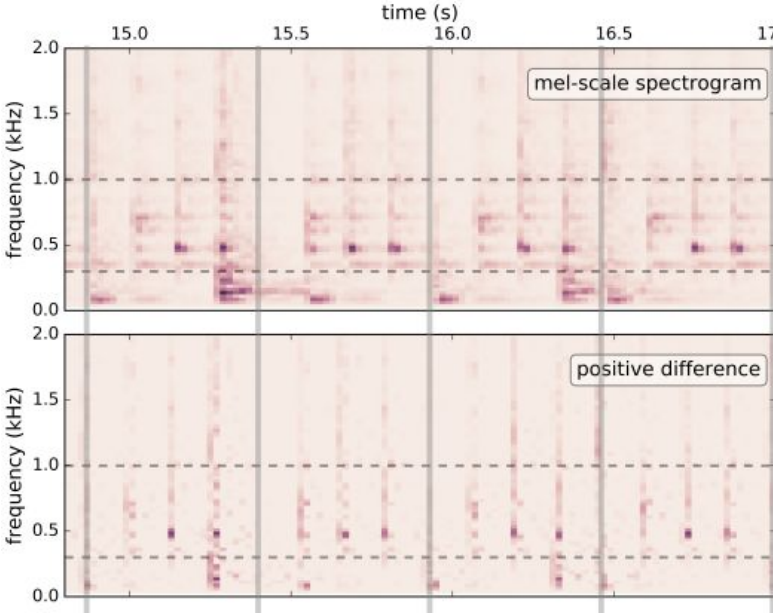
Procesamiento de Señales/ Imágenes y video

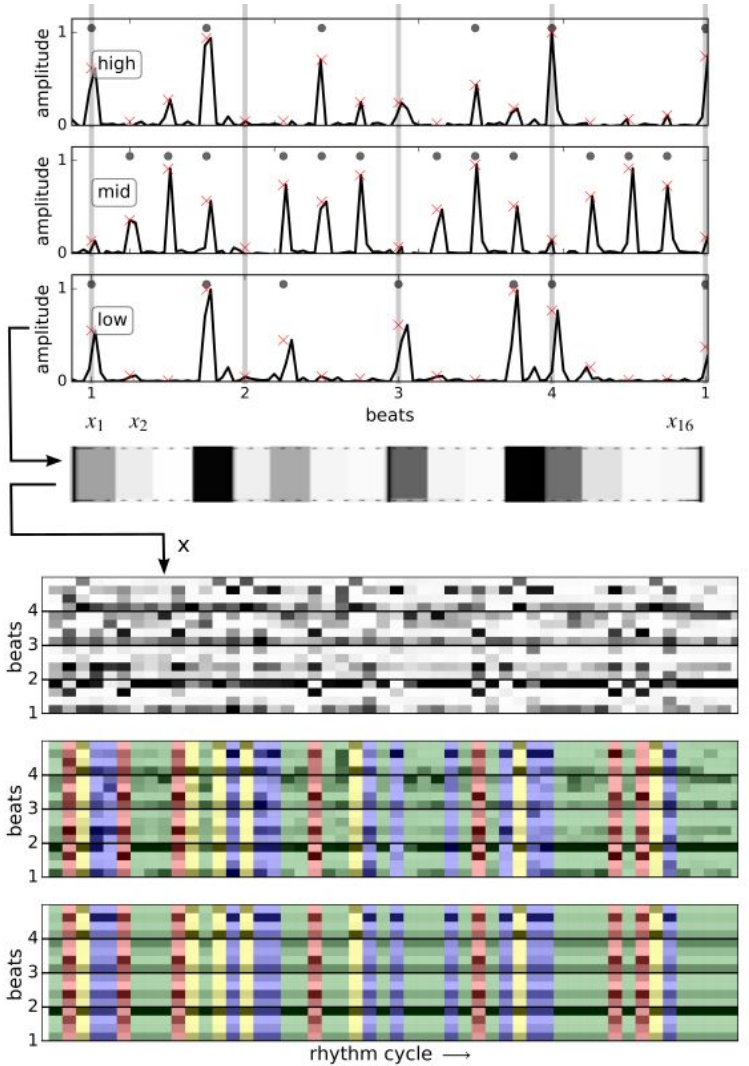




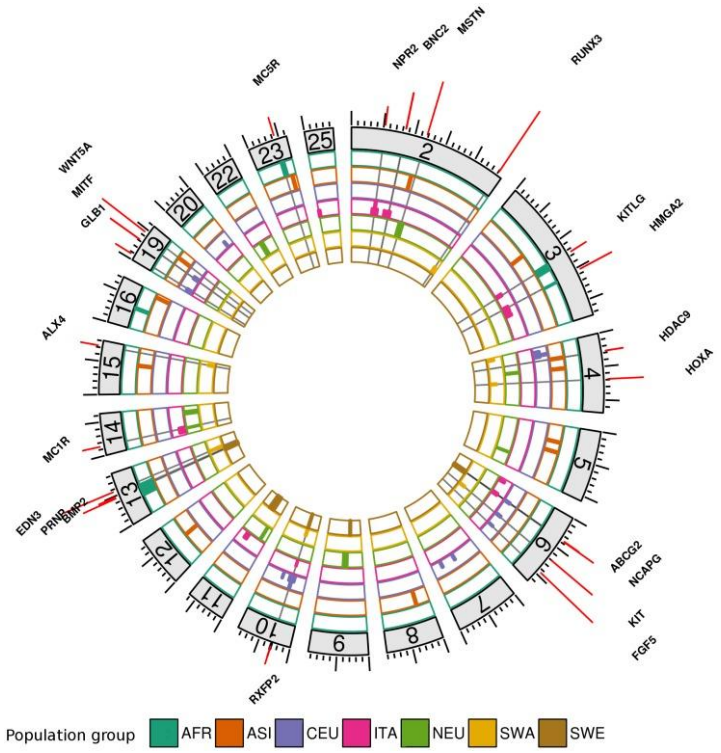
Procesamiento de Señales -Audio

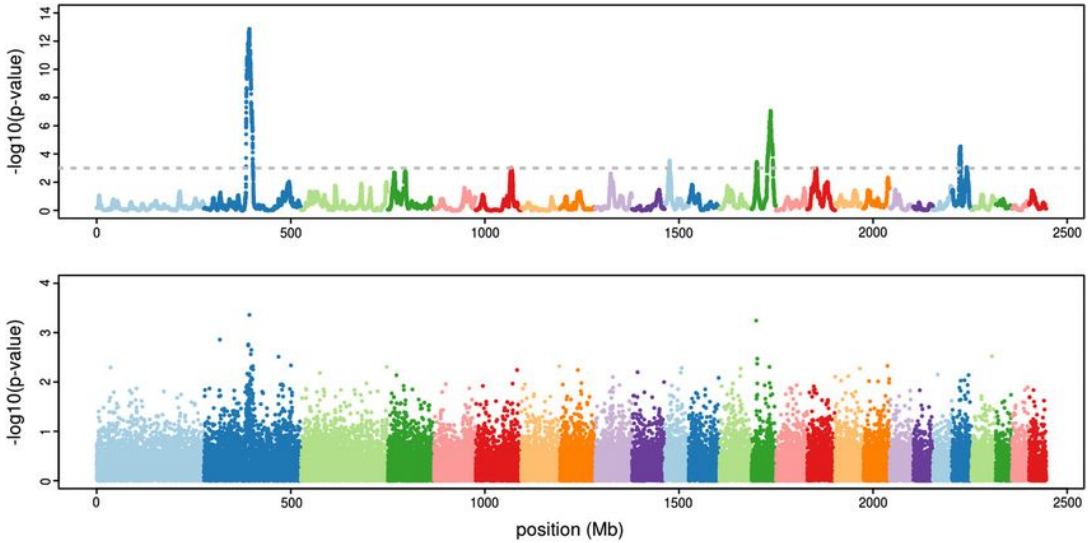




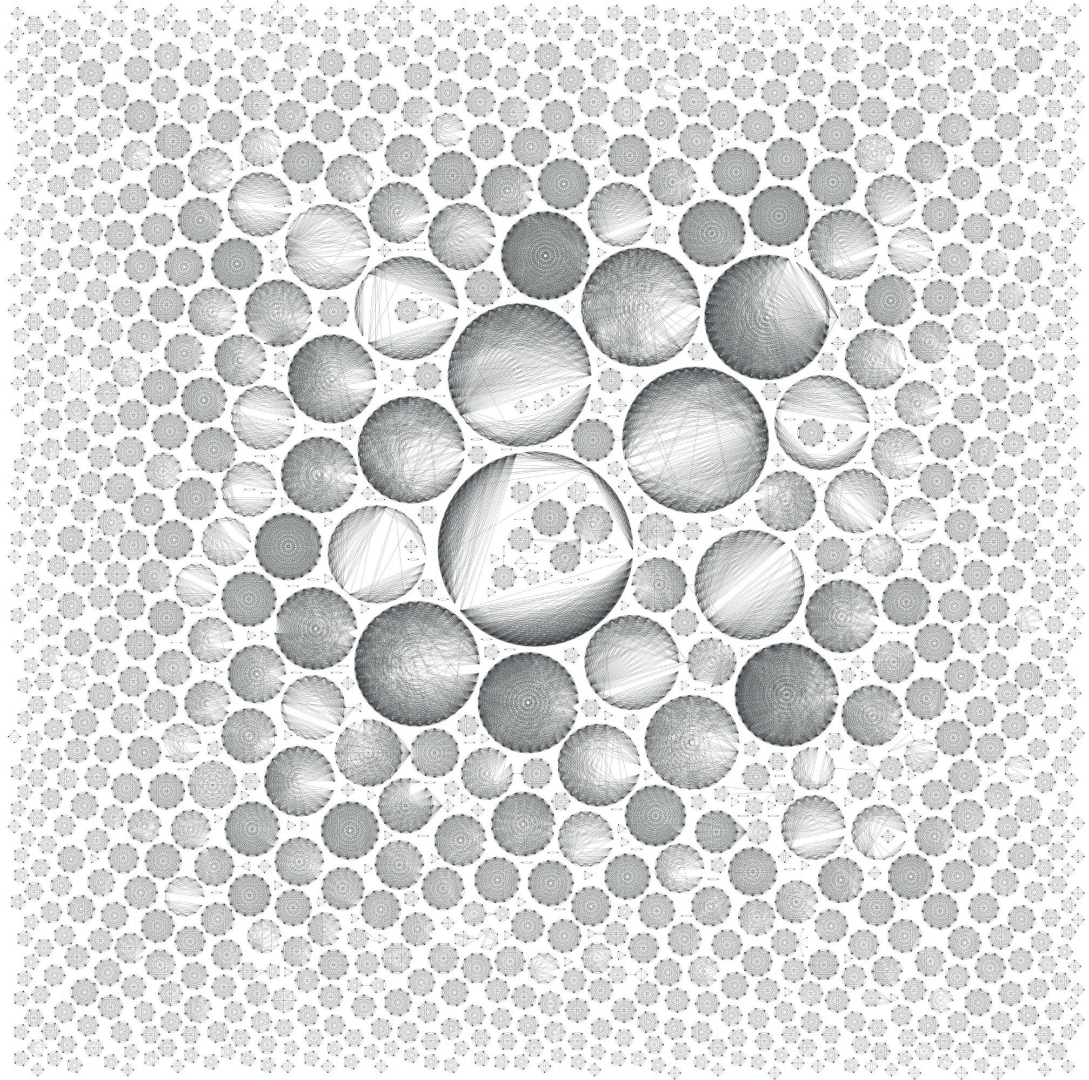


Datos genómicos





Grafos



Introducción a la Ciencia de Datos

Next: Gestión de datos

UNIVERSIDAD
DE LA REPUBLICA
URUGUAY