

# Introducción a la Ciencia de Datos

Maestría en Ciencias  
de la Computación

Dr. Irvin Hussein López Nava



# **Data Cleaning**



# **Feature Engineering**



## 2.3 Ingeniería de características



**Feature  
Engineering  
like a boss**

**ML Engineers in the 2000s**

**"idk y my  
neural net  
doin this"**



**ML Engineers now**



# ¿Qué es la Ingeniería de Características?

# Definiciones

## ChatGPT

Es el proceso de selección, creación y transformación de variables o características (*features*) a partir de los datos brutos para mejorar el rendimiento de un modelo de aprendizaje automático.

<https://chat.openai.com/>

## Wikipedia

Es el proceso de extracción de características (propiedades, atributos) a partir de datos brutos. Debido al aprendizaje profundo, que es capaz de aprender por sí mismo, ha quedado obsoleta para el procesamiento de cierto tipo de datos.

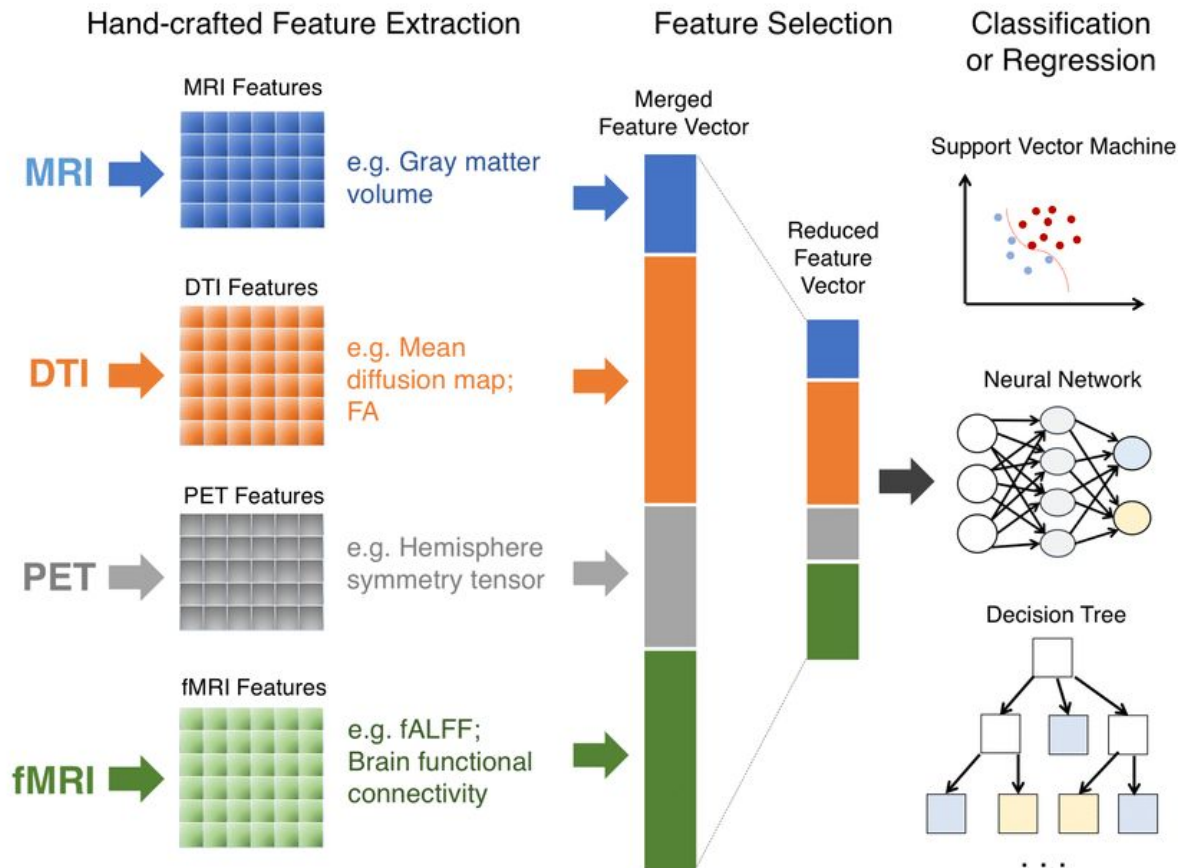
[https://en.wikipedia.org/wiki/Feature\\_engineering](https://en.wikipedia.org/wiki/Feature_engineering)

## Gemini

Es un proceso que consiste en transformar los datos crudos en características que sean significativas y útiles para los algoritmos de ML, i.e., se trata de preparar los datos para que el modelo pueda aprender patrones y hacer predicciones de manera más precisa.

<https://gemini.google.com/>

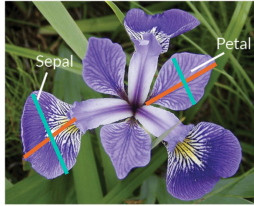








## Input



**Iris Versicolor**



**Iris Setosa**



**Iris Virginica**

Feature  
extraction

## Output

|     | sepalength | sepalwidth | petallength | petalwidth | class          |
|-----|------------|------------|-------------|------------|----------------|
| 0   | 5.1        | 3.5        | 1.4         | 0.2        | Iris-setosa    |
| 1   | 4.9        | 3.0        | 1.4         | 0.2        | Iris-setosa    |
| 2   | 4.7        | 3.2        | 1.3         | 0.2        | Iris-setosa    |
| 3   | 4.6        | 3.1        | 1.5         | 0.2        | Iris-setosa    |
| 4   | 5.0        | 3.6        | 1.4         | 0.2        | Iris-setosa    |
| ... | ...        | ...        | ...         | ...        | ...            |
| 145 | 6.7        | 3.0        | 5.2         | 2.3        | Iris-virginica |
| 146 | 6.3        | 2.5        | 5.0         | 1.9        | Iris-virginica |
| 147 | 6.5        | 3.0        | 5.2         | 2.0        | Iris-virginica |
| 148 | 6.2        | 3.4        | 5.4         | 2.3        | Iris-virginica |
| 149 | 5.9        | 3.0        | 5.1         | 1.8        | Iris-virginica |

150 rows × 5 columns

## Aprendizaje automático clásico

### Datos etiquetados



Iris Versicolor



Iris Setosa

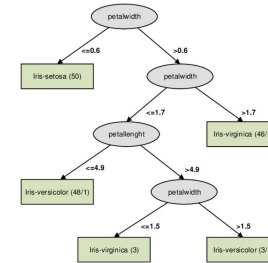


Iris Virginica

### Extracción y selección de características

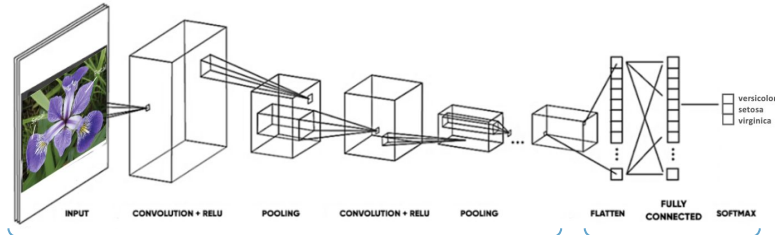


### Clasificación



*versicolor,  
setosa,  
virginica*

## Aprendizaje profundo



Extracción de características

Clasificación

*versicolor,  
setosa,  
virginica*

A woman with dark hair, wearing an orange sweater, is shown in profile, looking upwards and to the left. She is holding a tablet in her left hand and reaching out with her right hand towards a glowing point on a digital globe. The background is a complex, futuristic interface with various data visualizations, including line graphs, bar charts, and circular progress indicators, all in shades of blue and white. The overall theme is technology and data analysis.

# ¿Qué es la Extracción de Características?

# Definiciones

## ChatGPT

Es una técnica utilizada en el procesamiento de datos y el aprendizaje automático que consiste en obtener y seleccionar un subconjunto relevante de características (también conocidas como atributos o variables) a partir de datos brutos o de alta dimensionalidad.

<https://chat.openai.com/>

## Wikipedia

Es el proceso de extracción de características (características, propiedades, atributos) a partir de datos brutos. Debido al aprendizaje profundo, que son capaces de aprender por sí mismas, ha quedado obsoleta para el procesamiento de cierto tipo de datos.

[https://en.wikipedia.org/wiki/Feature\\_engineering](https://en.wikipedia.org/wiki/Feature_engineering)

## Gemini

Es un subproceso de la ingeniería de características que se centra en la creación de nuevas características a partir de las existentes. Este proceso puede implicar la transformación de datos, la selección de características o la combinación de características.

<https://gemini.google.com/>

# ¿En qué consiste la extracción?

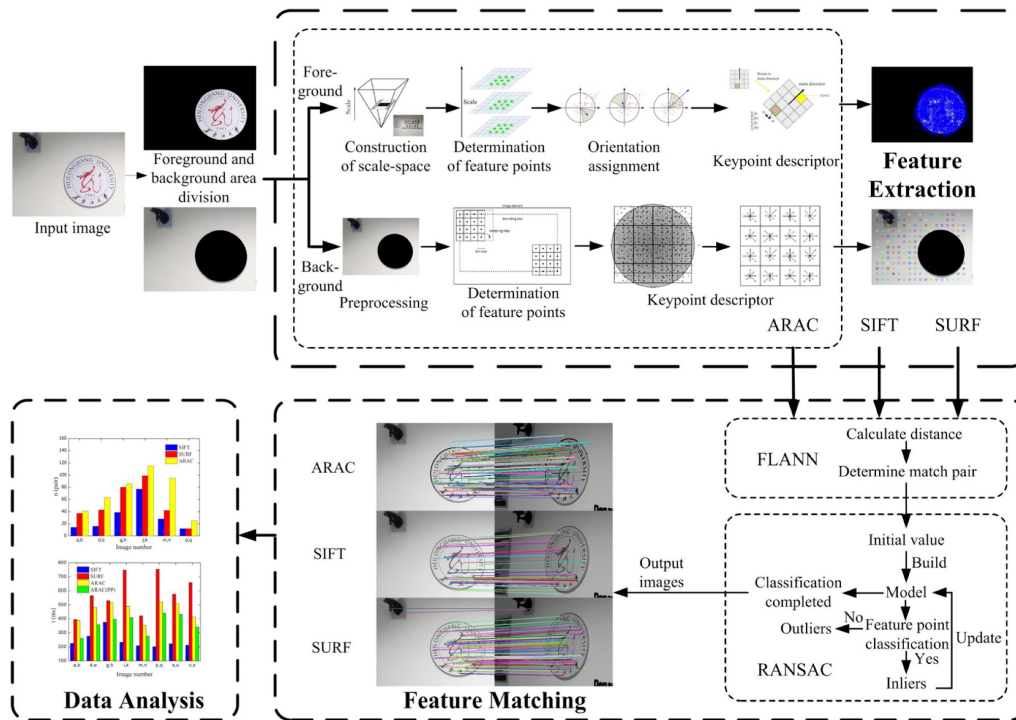
- La extracción de características se refiere al proceso de **transformar** datos sin procesar (crudos) en **características numéricas** mientras se conserva la información del conjunto de datos original.
- En general, produce mejores resultados que aplicar el aprendizaje automático directamente a los datos sin procesar.
- Hay dos tipos principales de estrategias para la extracción de características: (i) manual y (ii) automática.



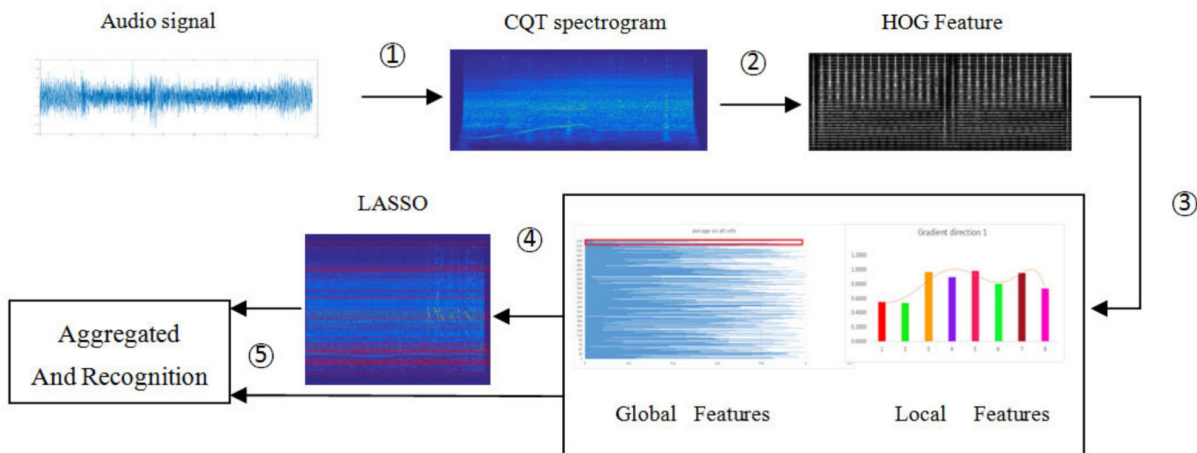
# Extracción manual

- Requiere identificar y describir las características que son relevantes para un problema determinado e implementar una forma de extraer esas características.
- Tener una buena comprensión del dominio ayuda a identificar las características que podrían ser útiles. En ocasiones se solicita a los **expertos** que enlisten tales características.
- Durante décadas de investigación, ingenieros y científicos han desarrollado métodos de extracción de características para imágenes, señales y texto.

# Extracción de características de una imagen



# Extracción de características de un audio



# Extracción de características en texto

Documents

We study the complexity of influencing elections through bribery. How computationally complex is it for an external actor to determine whether by a certain amount of bribing voters a specified candidate can be made the election's winner? We study this problem for election systems as varied as scoring ...

Vector-space representation

|            | D1 | D2 | D3 | D4 | D5 |
|------------|----|----|----|----|----|
| complexity | 2  |    | 3  | 2  | 3  |
| algorithm  | 3  |    | 4  | 4  |    |
| entropy    | 1  |    |    | 2  |    |
| traffic    |    | 2  | 3  |    |    |
| network    |    | 1  | 4  |    |    |

Term-document matrix

$tf(t, d)$

|   | blue | bright | can | see | shining | sky | sun | today |
|---|------|--------|-----|-----|---------|-----|-----|-------|
| 1 | 1/2  | 0      | 0   | 0   | 0       | 1/2 | 0   | 0     |
| 2 | 0    | 1/3    | 0   | 0   | 0       | 0   | 1/3 | 1/3   |
| 3 | 0    | 1/3    | 0   | 0   | 0       | 1/3 | 1/3 | 0     |
| 4 | 0    | 1/6    | 1/6 | 1/6 | 1/6     | 0   | 1/3 | 0     |

$\times$

$idf(t, D)$

|  | blue  | bright | can   | see   | shining | sky   | sun   | today |
|--|-------|--------|-------|-------|---------|-------|-------|-------|
|  | 0.602 | 0.125  | 0.602 | 0.602 | 0.602   | 0.301 | 0.125 | 0.602 |

$tfidf(t, d, D) = tf(t, d) \cdot idf(t, D)$

|   | blue         | bright | can          | see          | shining      | sky          | sun    | today        |
|---|--------------|--------|--------------|--------------|--------------|--------------|--------|--------------|
| 1 | <b>0.301</b> | 0      | 0            | 0            | 0            | 0.151        | 0      | 0            |
| 2 | 0            | 0.0417 | 0            | 0            | 0            | 0            | 0.0417 | <b>0.201</b> |
| 3 | 0            | 0.0417 | 0            | 0            | 0            | <b>0.100</b> | 0.0417 | 0            |
| 4 | 0            | 0.0209 | <b>0.100</b> | <b>0.100</b> | <b>0.100</b> | 0            | 0.0417 | 0            |

- TF-IDF: Multiply TF and IDF scores, use to rank importance of words within documents
- Most important word for each document is highlighted

# Extracción automática

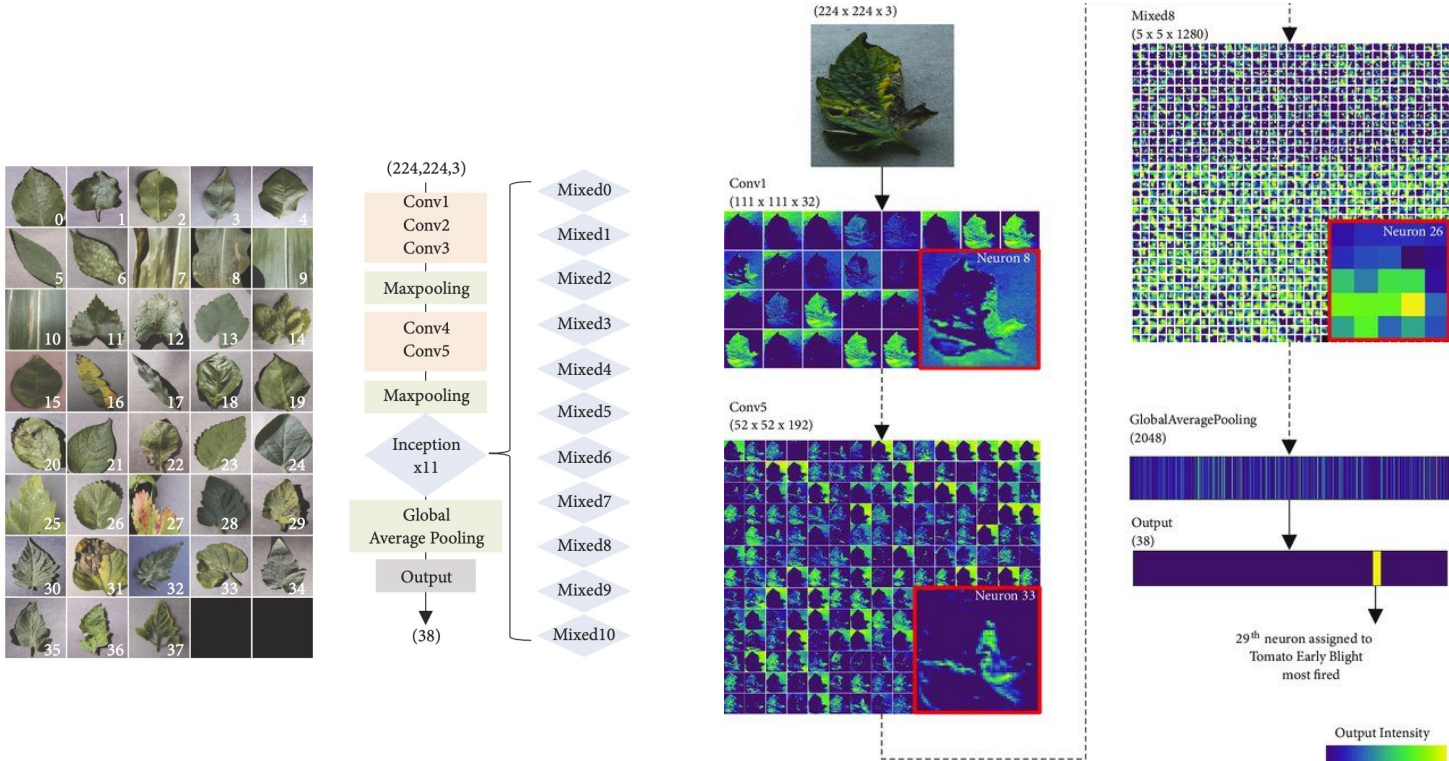
- Utiliza algoritmos especializados o **redes profundas** para extraer las características sin necesidad de intervención humana.
- Esta técnica puede ser muy útil cuando se desconoce el dominio del tema o cuando la cantidad o la calidad de datos es muy grande y no permite encontrar patrones evidentes.
- En aprendizaje profundo, la extracción de características está a cargo en gran medida de las primeras capas de las redes, y es utilizada ampliamente para datos de imágenes.



# *Deep features*

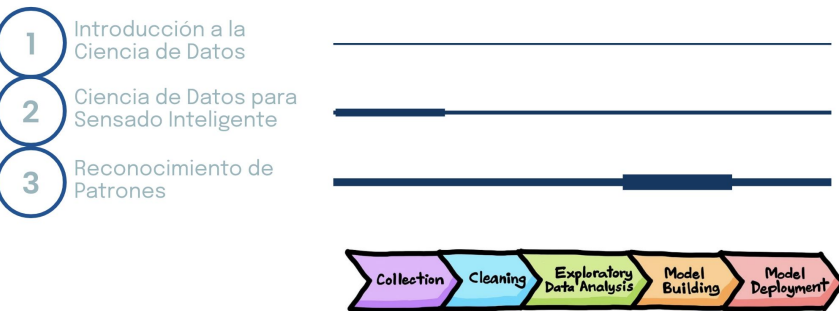
- Las características profundas generalmente se extraen de las últimas capas antes de la capa de clasificación de una red neuronal.
- Por ejemplo, para clasificar imágenes con una CNN, la red aprende a reconocer varios patrones y texturas.
- Pero con características profundas, no existe una forma de denominación específica para ellas, aparte de indicarla mediante el número de columna en la representación (posición de una neurona en una capa oculta).

# Ejemplo de características profundas



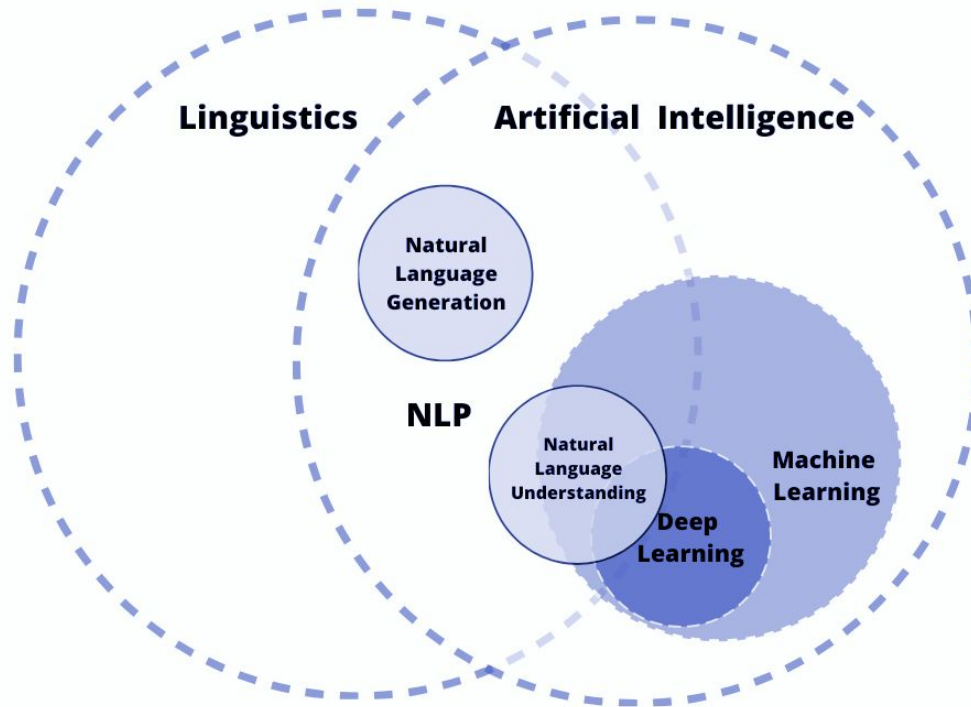
# Temario

## Ciencia de datos para sensores inteligentes

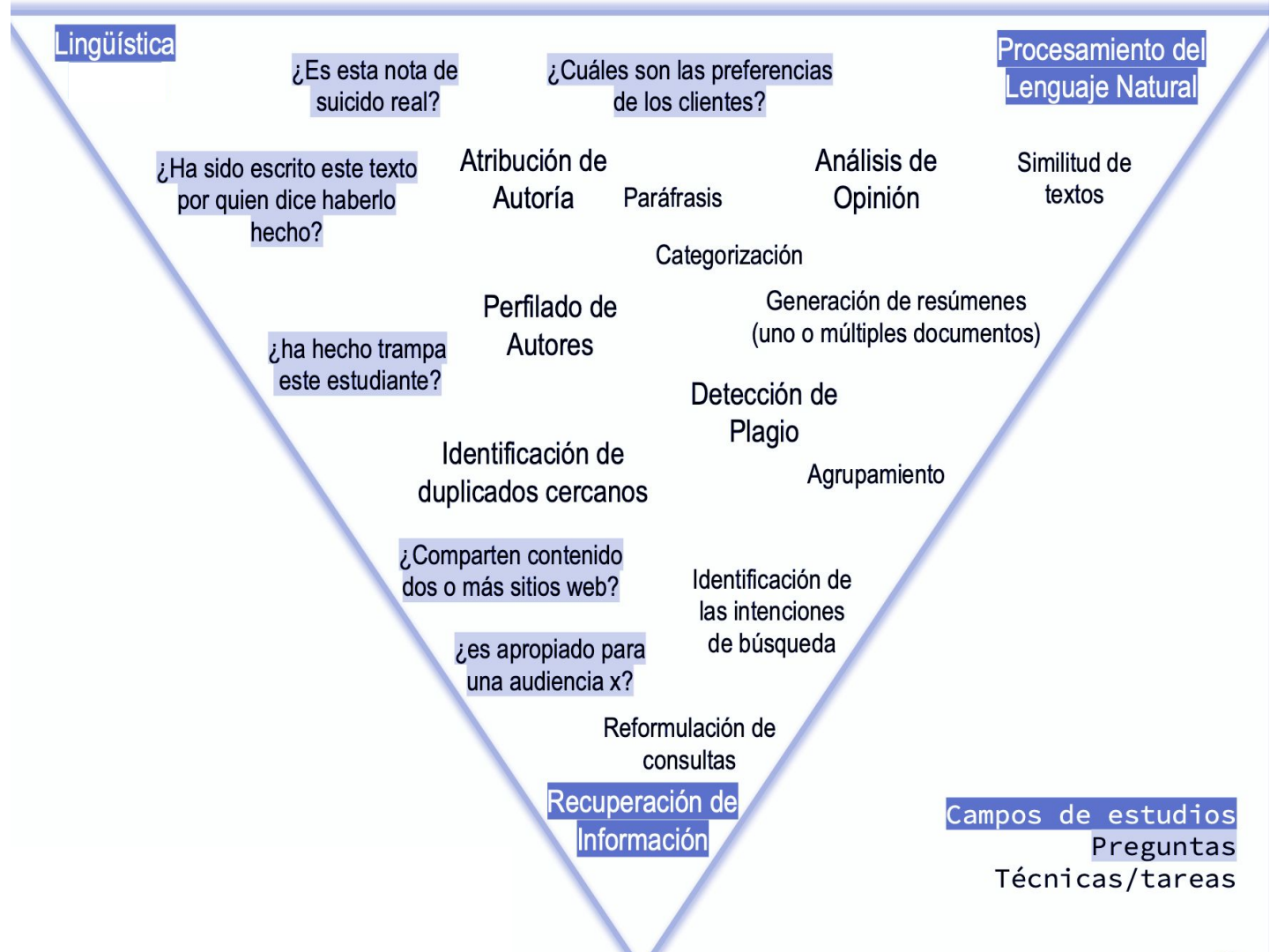


- 1 Introducción al sensado de datos
- 2 Sensado de audio y voz
- 3 Sensado de imágenes y video
- 4 Sensado inercial
- 5 Sensado fisiológico
- 6 Reconocimiento de actividad
- 7 Reconocimiento de comportamiento

# Caso de estudio







# Los datos

Los datos se representan por un conjunto de atributos o mediciones

- **Instancia:** ??
- **Atributos:** ??
- **Clase:** ??

Dan Quinn está orgulloso de la velocidad y presión a los quarterbacks que su defensa ha metido en los primeros dos partidos de temporada, así como de los intercambios de balón que han provocado.

Pero el coordinador defensivo de los Dallas Cowboys está aún más satisfecho por la manera en que han frenado la carrera en los triunfos sobre los New York Giants y los New York Jets para empezar la temporada.

“Estamos aprendiendo a jugar partidos completos”, consideró Quinn este lunes en una conferencia de prensa. “Habíamos sufrido dolores de cabeza contra la carrera. Pero los muchachos trabajaron muy fuerte en el receso de temporada para ir contra las corridas”.

En el primer partido, los Cowboys frenaron en 51 yardas por carrera a Saquon Barkley, de los Giants. El domingo, los dos corredores principales de los Jets, Breece Hall y Dalvin Cook, sumaron apenas ocho y siete yardas, de manera respectiva.

# Los datos

Los datos se representan por un conjunto de atributos o mediciones

- **Instancia:** Documento, e.g., nota periodística
- **Atributos:** ??
- **Etiqueta:** Deportes

Dan Quinn está orgulloso de la velocidad y presión a los quarterbacks que su defensa ha metido en los primeros dos **partidos** de temporada, así como de los intercambios de **balón** que han provocado.

Pero el coordinador defensivo de los Dallas Cowboys está aún más satisfecho por la manera en que han frenado la carrera en los **triunfos** sobre los New York Giants y los New York Jets para empezar la temporada.

“Estamos aprendiendo a **jugar partidos** completos”, consideró Quinn este lunes en una conferencia de prensa. “Habíamos sufrido dolores de cabeza contra la carrera. Pero los muchachos trabajaron muy fuerte en el receso de temporada para ir contra las corridas”.

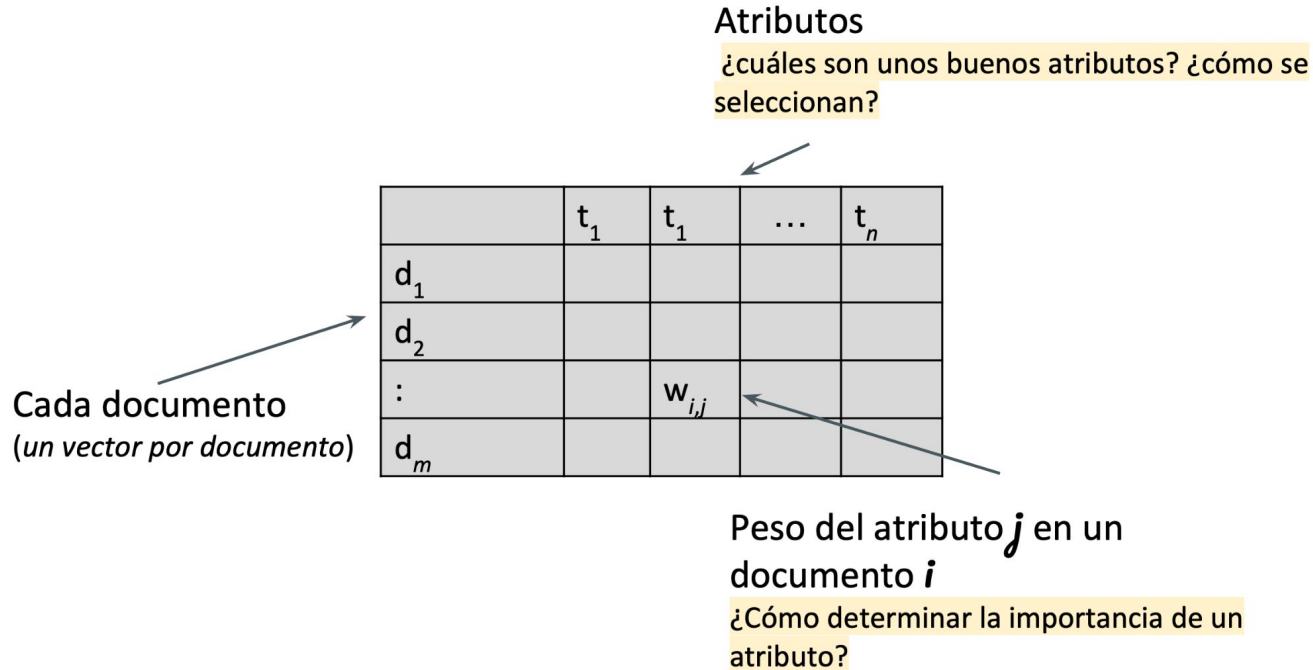
En el primer **partido**, los Cowboys frenaron en 51 yardas por carrera a Saquon Barkley, de los Giants. El domingo, los dos corredores principales de los Jets, Breece Hall y Dalvin Cook, sumaron apenas ocho y siete yardas, de manera respectiva.

# Representación vectorial

Atributos

¿cuáles son unos buenos atributos? ¿cómo se seleccionan?

Cada documento  
(un vector por documento)



The diagram shows a document-term matrix. The columns are labeled with terms  $t_1, t_1, \dots, t_n$ . The rows are labeled with documents  $d_1, d_2, :, d_m$ . A specific weight  $w_{i,j}$  is highlighted in the cell corresponding to document  $d_i$  and term  $t_j$ . Annotations include an arrow pointing to the header row from the text 'Atributos', an arrow pointing to the first column from the text 'Cada documento', and an arrow pointing to the cell  $w_{i,j}$  from the text 'Peso del atributo  $j$  en un documento  $i$ '.

|       | $t_1$ | $t_1$     | $\dots$ | $t_n$ |
|-------|-------|-----------|---------|-------|
| $d_1$ |       |           |         |       |
| $d_2$ |       |           |         |       |
| $:$   |       | $w_{i,j}$ |         |       |
| $d_m$ |       |           |         |       |

Peso del atributo  $j$  en un documento  $i$

¿Cómo determinar la importancia de un atributo?

# ¿Cómo se calcula el peso de un atributo en un documento?

- La importancia del término puede ser proporcionalmente al número de veces que aparece en el documento.
  - Este enfoque ayuda a **describir** el contenido del documento
- La importancia general de un término decrementa proporcionalmente a la ocurrencia de la colección completa.
  - Términos comunes no son buenos para **discriminar** entre las clases de objetos.



# Esquemas de pesado

- Pesado booleano
  - $w_{ij} = 1$ , sí y solo sí el documento  $i$  contiene el término  $j$ , en caso contrario, es 0.
- Frecuencia del término (**tf**)
  - $w_{ij}$  = número de ocurrencias de  $t_j$  en  $d_i$ .
- **tf x idf** (*Term frequency – Inverse document frequency*)
  - $w_{ij} = \text{tf}(t_j, d_i) \times \text{idf}(t_j)$
  - $\text{tf}(t_j, d_i)$  indica la ocurrencia de  $t_j$  en el documento  $d_i$ .
  - $\text{idf}(t_j) = \log [N/\text{df}(t_j)]$ , donde  $\text{df}(t_j)$  es el número de documento que contienen el término  $t_j$ .

# De texto a vector

El enfoque más sencillo es una representación basada en una **bolsa de palabras** BoW (*bag of words*) y las frecuencias de estas palabras en cada documento.

Ejemplos:

x1: “y mi voz que madura”

x2: “y mi voz quemadura”

x3: “y mi bosque madura”

x4: “y mi voz quema dura”

|      |       |        |        |          |             |          |         |        |
|------|-------|--------|--------|----------|-------------|----------|---------|--------|
| y: 4 | mi: 4 | voz: 3 | que: 1 | madura:2 | quemadura:1 | bosque:1 | quema:1 | dura:1 |
|------|-------|--------|--------|----------|-------------|----------|---------|--------|

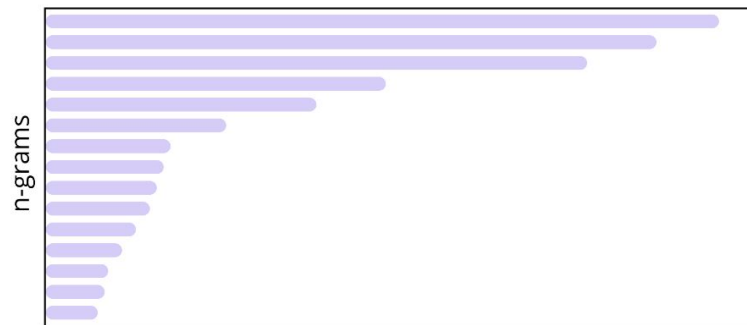
## De texto a vector

|                       | y | mi | voz | que | madur<br>a | quemadura | bosque | quema | dura |
|-----------------------|---|----|-----|-----|------------|-----------|--------|-------|------|
| “y mi voz que madura” | 1 | 1  | 1   | 1   | 1          | 0         | 0      | 0     | 0    |
| “y mi voz quemadura”  | 1 | 1  | 1   | 0   | 0          | 1         | 0      | 0     | 0    |
| “y mi bosque madura”  | 1 | 1  | 0   | 0   | 1          | 0         | 1      | 0     | 0    |
| “y mi voz quema dura” | 1 | 1  | 1   | 0   | 0          | 0         | 0      | 1     | 1    |

# Problemas con esta representación

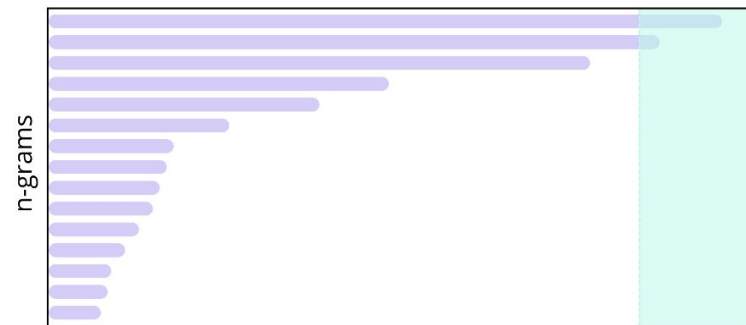
- Se filtran términos de palabras comunes.
- Hay palabras que aparecen poco y que pueden no ser relevantes para nuestra tarea.
- Al ser una bolsa de palabras, no existe un orden en las palabras que aparecen en el texto.

**Original**



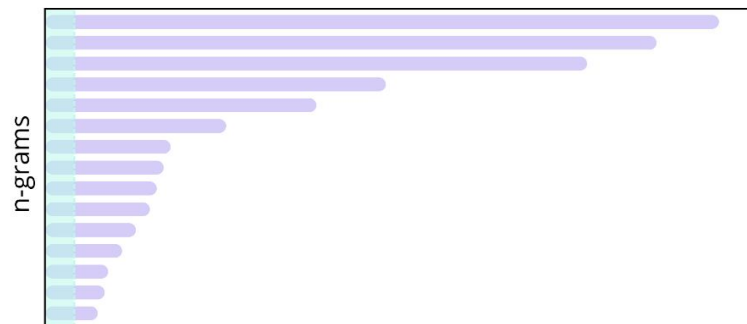
Frequency

**Clean**



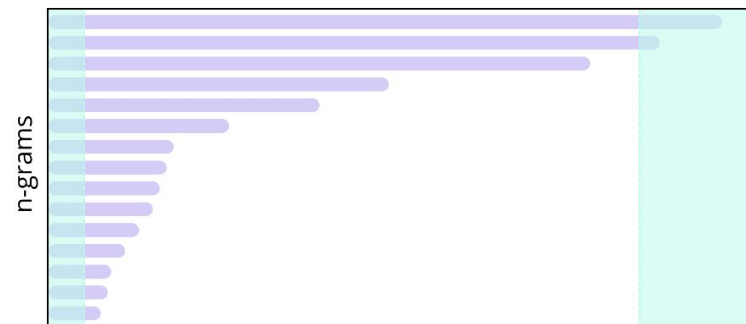
Frequency

**FreqAbove1**



Frequency

**CPPD**



Frequency

# Filtrar palabras comunes

- *Stopwords*. Son palabras funcionales, por ejemplo artículos, conectores, etc. Ocurren con mucha frecuencia en la mayoría de los documentos, no ayudan a discriminar el tema de un texto.
- Palabras del problema que no queremos que se consideren (dependiendo del dominio del problema).
- Eliminar términos por frecuencia máxima (si ocurre más de  $n$  veces, e.g., *data*).

# Filtrar palabras con frecuencia mínima

- *Corte cut-off*: se debe determinar cuál es la frecuencia mínima de un término para ser considerado dentro de la representación.
- Por ejemplo, palabras que aparecen una ocasión.
  - En algunos problemas más de la mitad de los atributos en una representación BoW aparecen solo una vez.



# Agregar nuevos términos

Podemos agregar nuevos términos que considera, hasta cierto punto, el orden de algunas palabras:

- **Bigramas:** palabras de dos en dos
- **Trigramas:** palabras de tres en tres
- **ngramas:** palabras de n en n

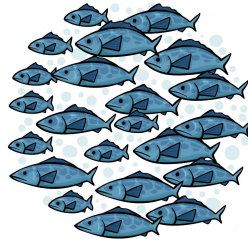
“y mi voz que madura”

**bigramas:** y-mi, mi-voz, voz-que, que-madura

**trigramas:** y-mi-voz, mi-voz-que, voz-que-madura

# Complejidad del lenguaje

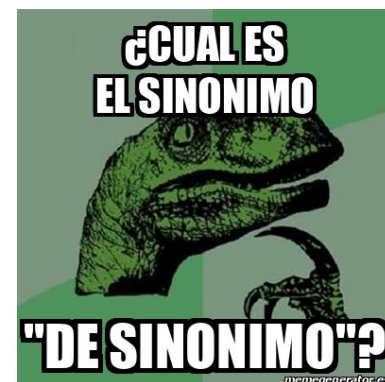
## Polisemia



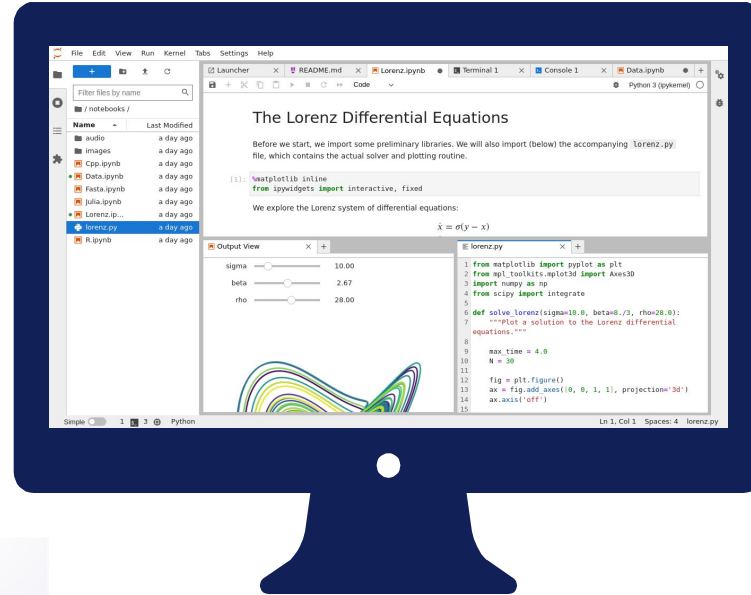
## Sinonimia



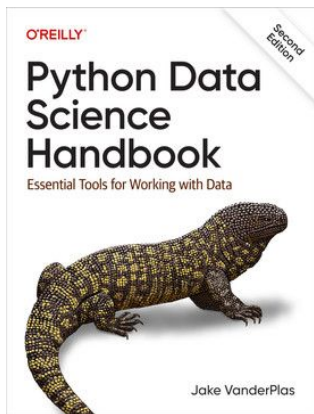
Flaca - Delgada  
 Tranquilo - quieto  
 Querer - Amar



(Go to live notebook)



# Extra Libro



- 05.04-Feature-Engineering.ipynb

# Gracias!

¿Alguna pregunta?

hussein@cicese.mx

<https://sites.google.com/view/husseinlopeznava>



**CREDITS:** This presentation was based on a template by Slidesgo, and includes icons by Flaticon.