

# Aplicaciones de Data Science

**Unidad 4:** Visión Computacional

**Visión por computador y detección de objetos**

## Unidad 4

# Visión Computacional

Visión por computador y  
detección de objetos



## Contenido

1. ¿Qué es la Visión por Computador?
2. ¿Cómo funciona la Visión por Computador?
3. Revolución del Aprendizaje Profundo
4. Redes Neuronales Convolucionales (CNN)
5. Casos Prácticos

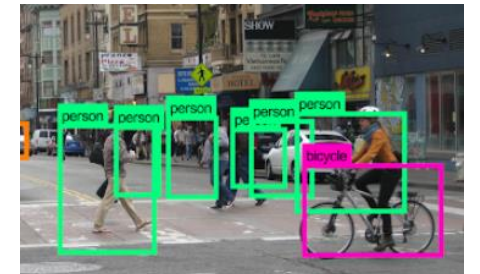
# 1. ¿Qué es la Visión por Computador?

- La visión por computador es uno de los campos de la inteligencia artificial que entrena y permite a las computadoras comprender el mundo visual.

## DEFINICION

La Visión por Computador (CV) **es el campo científico** en la IA que define cómo las máquinas interpretan el significado de imágenes y videos.

- Las computadoras pueden utilizar imágenes digitales y modelos de aprendizaje profundo para identificar y clasificar objetos con precisión y reaccionar ante ellos.
- La Visión por Computador en IA se dedica al desarrollo de sistemas automatizados que pueden interpretar datos visuales (como fotografías o imágenes en movimiento) de la misma manera que lo hacemos las personas.
- La idea detrás de la Visión por Computador es instruir a las computadoras para que interpreten y comprendan imágenes píxel por píxel.



# 1. ¿Qué es la Visión por Computador?

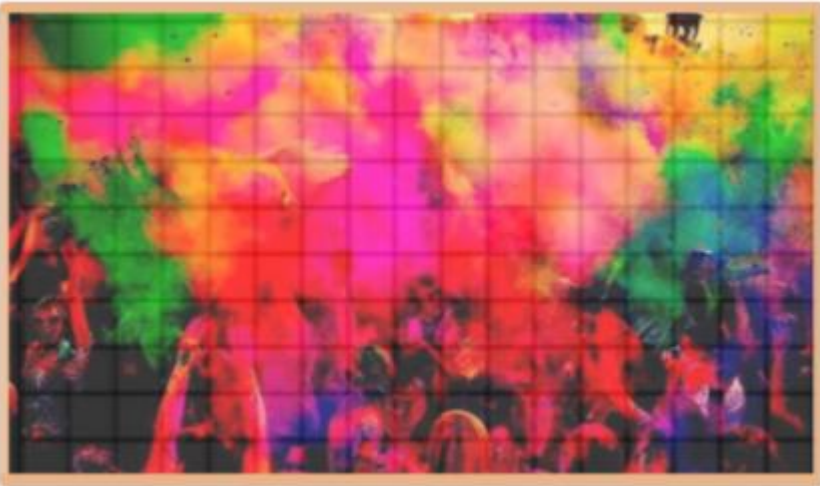
## La base del campo de la Visión por Computador

Instruir a las computadoras para que interpreten y comprendan imágenes píxel por píxel.

- Si pensamos en una imagen como una cuadrícula, cada cuadrado de la cuadrícula contiene un solo píxel.
- Los píxeles son los componentes básicos de una imagen.
- Cada imagen consta de un conjunto de píxeles.
- No hay granularidad más fina que la del píxel .
- Normalmente, se considera un píxel a el “ color ” o la “ intensidad ” de la luz que aparece en un lugar determinado de nuestra imagen.

# 1. ¿Qué es la Visión por Computador?

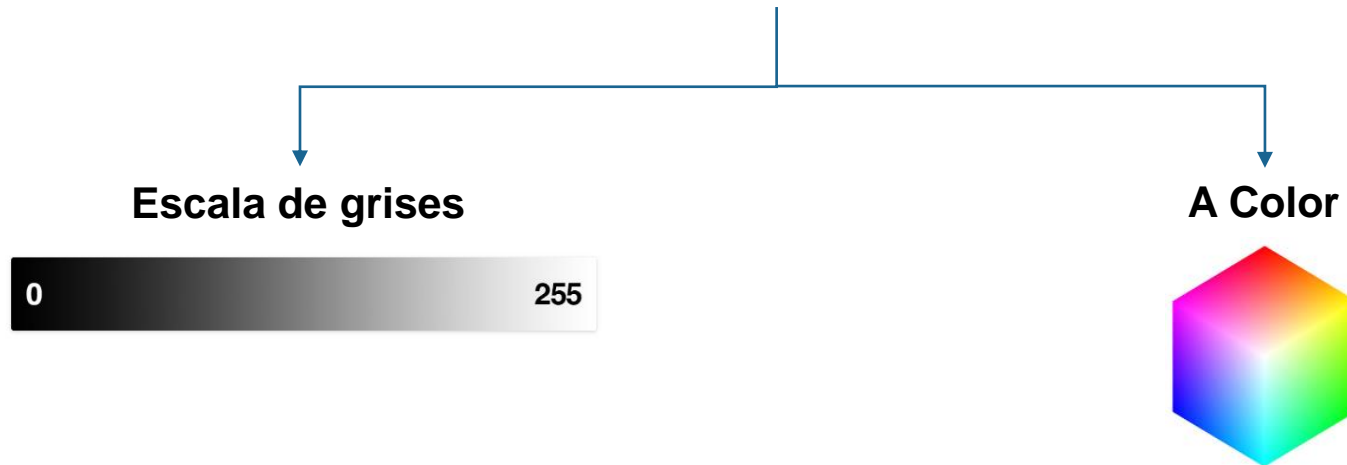
La base del campo de la visión por computador: **El Pixel**



- Una imagen con una resolución de **600 x 450** equivale a tener **600 píxeles de ancho (ancho) y 450 píxeles de alto (alto)**
- Así, la imagen se representa como una cuadrícula de píxeles, con **600 columnas y 450 filas**.
- En total, la imagen tiene  $600 \times 450 = \mathbf{270.000}$  **píxeles**.

# 1. ¿Qué es la Visión por Computador?

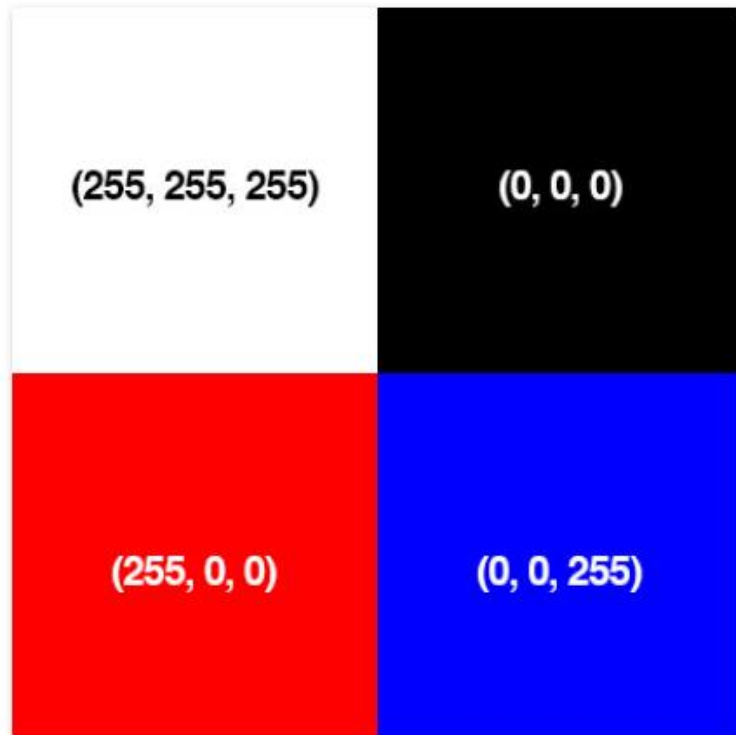
## Representación de los Píxeles



- En una imagen en escala de grises , cada píxel tiene un valor entre 0 y 255 , donde cero corresponde a “negro” y 255 a “blanco”.
- Los valores entre 0 y 255 son distintos tonos de gris , donde los valores más cercanos a 0 son más oscuros y los valores más cercanos a 255 son más claros.
- Los píxeles de color normalmente se representan en el espacio de color RGB: un valor para el componente rojo, uno para el verde y otro para el azul, lo que da un total de 3 valores por píxel.
- Cada uno de los tres colores rojo, verde y azul está representado por un número entero en el rango de 0 a 255 , que indica cuánto color hay.

# 1. ¿Qué es la Visión por Computador?

## El color en los Píxeles



- Combinamos estos valores (rojo, verde, azul) en una tupla RGB y esta tupla representara nuestro color.
- Para construir un color blanco , llenaríamos completamente cada uno de los cubos rojo, verde y azul, así:  $(255, 255, 255)$ , ya que el blanco es la presencia de todos los colores.
- Para crear un color negro , vaciaríamos cada uno de los cubos:  $(0, 0, 0)$ , ya que el negro es la ausencia de color.
- Para crear un color rojo puro, llenaríamos el cubo rojo (y solo el cubo rojo) por completo:  $(255, 0, 0)$ .

# 1. ¿Qué es la Visión por Computador?

## Procesamiento de imágenes vs Visión por Computadora

- Procesamiento de imágenes no es lo mismo que la visión por computadora.
- El **procesamiento de imágenes** implica:

Modificar o mejorar imágenes para producir un nuevo resultado (optimizar el brillo o el contraste, aumentar la resolución, difuminar información confidencial o recortar).

- La diferencia entre el procesamiento de imágenes y la visión por computadora es que el primero no requiere necesariamente la identificación del contenido.



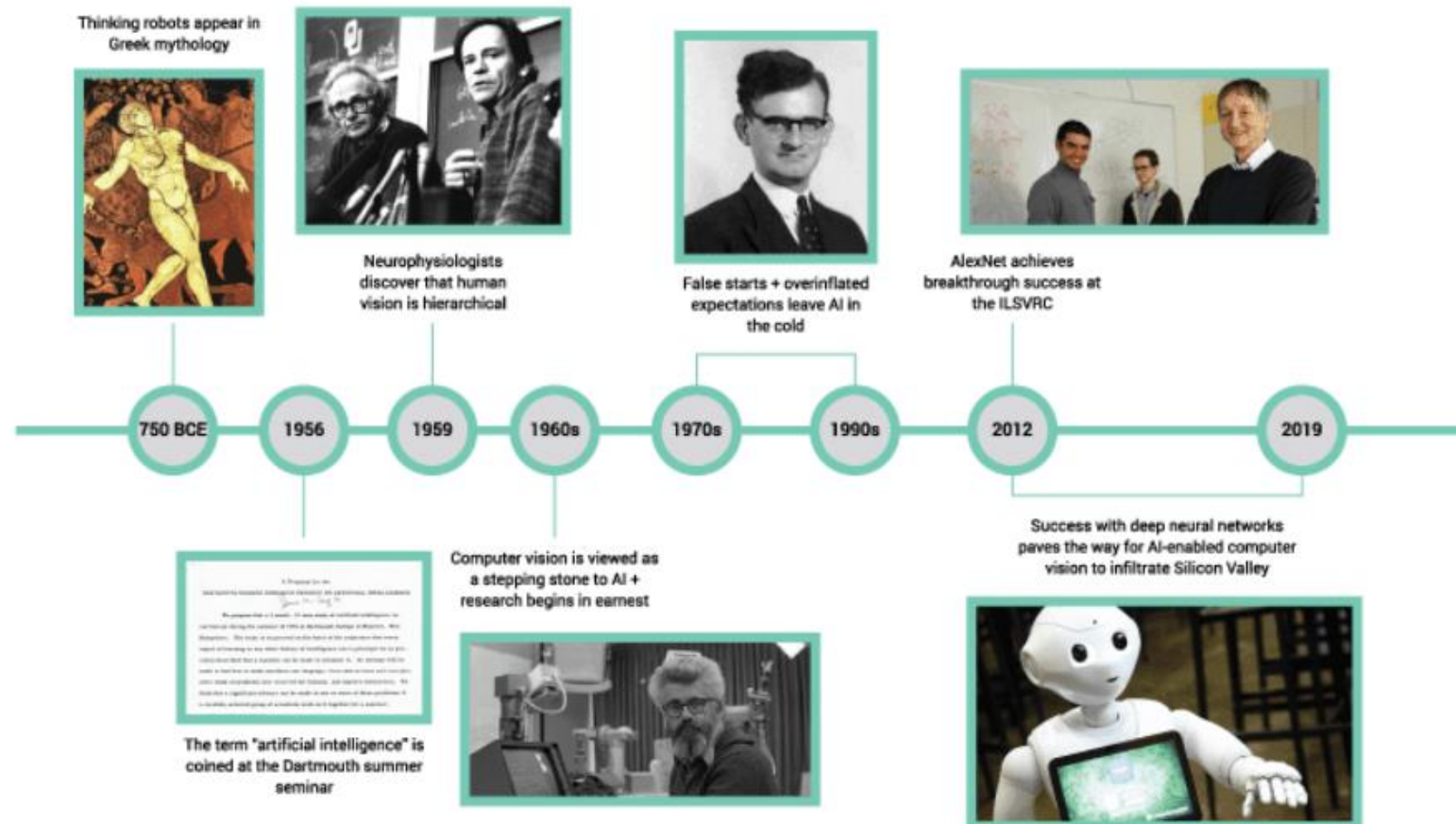
# 1. ¿Qué es la Visión por Computador?

## Historia de la Visión por Computador

- En la década de 1960, la inteligencia artificial (IA) inició sus esfuerzos para imitar la visión humana.
- Los neurocientíficos demostraron en 1982 que la visión opera jerárquicamente y presentaron técnicas que permitían a las computadoras reconocer bordes, vértices, arcos y otras estructuras fundamentales.
- Al mismo tiempo, los científicos de datos crearon una red de células de reconocimiento de patrones.
- En el año 2000, los investigadores concentraban sus esfuerzos en la identificación de objetos y, al año siguiente, la industria vio las primeras soluciones de reconocimiento facial en tiempo real.
- En la actualidad, cuando se trata de visión por computadora, **el aprendizaje profundo** es el camino a seguir.

# 1. ¿Qué es la Visión por Computador?

## Historia de la Visión por Computador



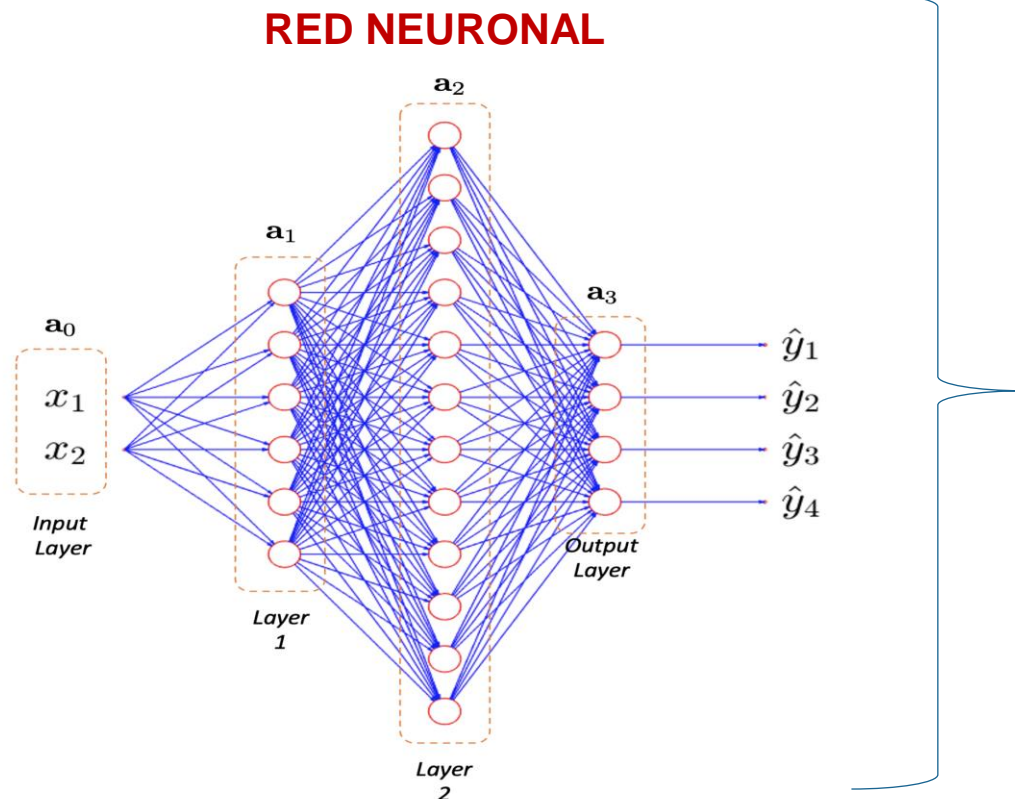
## 2. ¿Cómo funciona la Visión por Computador?

- Se requieren cantidades masivas de información para la visión por computadora.
- Se realizan análisis de datos repetidos hasta que el sistema puede diferenciar entre objetos e identificar imágenes.
- El **aprendizaje profundo**, un tipo específico de aprendizaje automático, y **las redes neuronales convolucionales**, una forma importante de red neuronal, son las dos técnicas clave que se utilizan para lograr este objetivo.
- Con la ayuda de algorítmicos preprogramados, un sistema de aprendizaje automático puede aprender automáticamente sobre la interpretación de datos visuales.
- El modelo puede aprender a distinguir entre imágenes similares si se le proporciona un conjunto de datos lo suficientemente grande.
- Los algoritmos hacen posible que el sistema aprenda por sí solo, de modo que pueda reemplazar el trabajo humano en tareas como el reconocimiento de imágenes.

**¿Cómo funcionan las redes neuronales convolucionales?**

# 3. Revolución del Aprendizaje Profundo

- El **aprendizaje profundo** es un tipo de aprendizaje automático que utiliza la visión por computadora moderna para obtener información basada en datos.



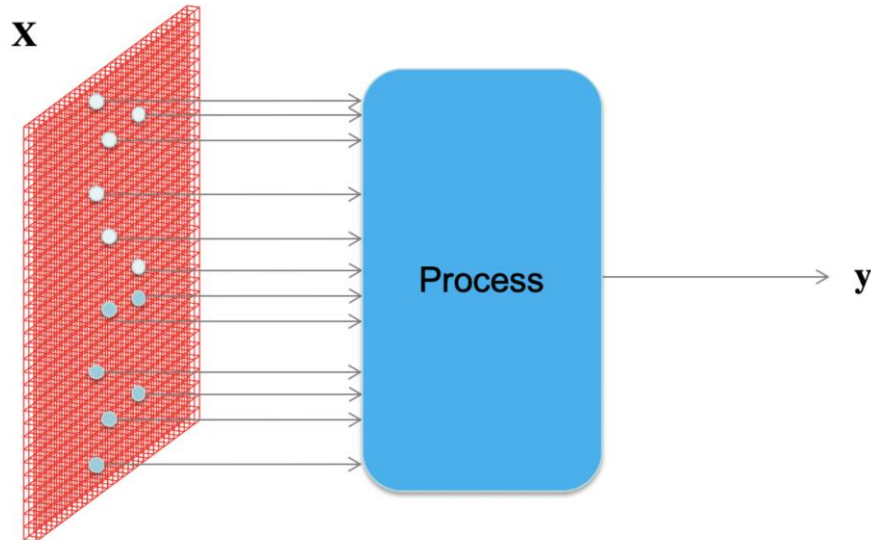
- Utiliza un algoritmo conocido como **red neuronal**
- Los patrones de los datos se extraen mediante redes neuronales.
- Los algoritmos se basan en nuestro conocimiento actual de la estructura y el funcionamiento del cerebro, específicamente los vínculos entre las neuronas dentro de la corteza cerebral.

# 3. Revolución del Aprendizaje Profundo

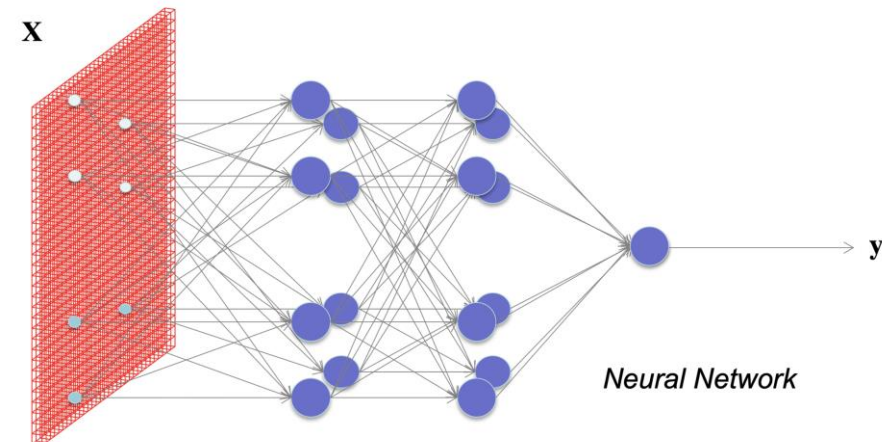
## MODELOS DE VISION POR COMPUTADOR

Los **modelos de visión por computadora** están diseñados para traducir datos visuales en función de características e información contextual identificadas durante el entrenamiento.

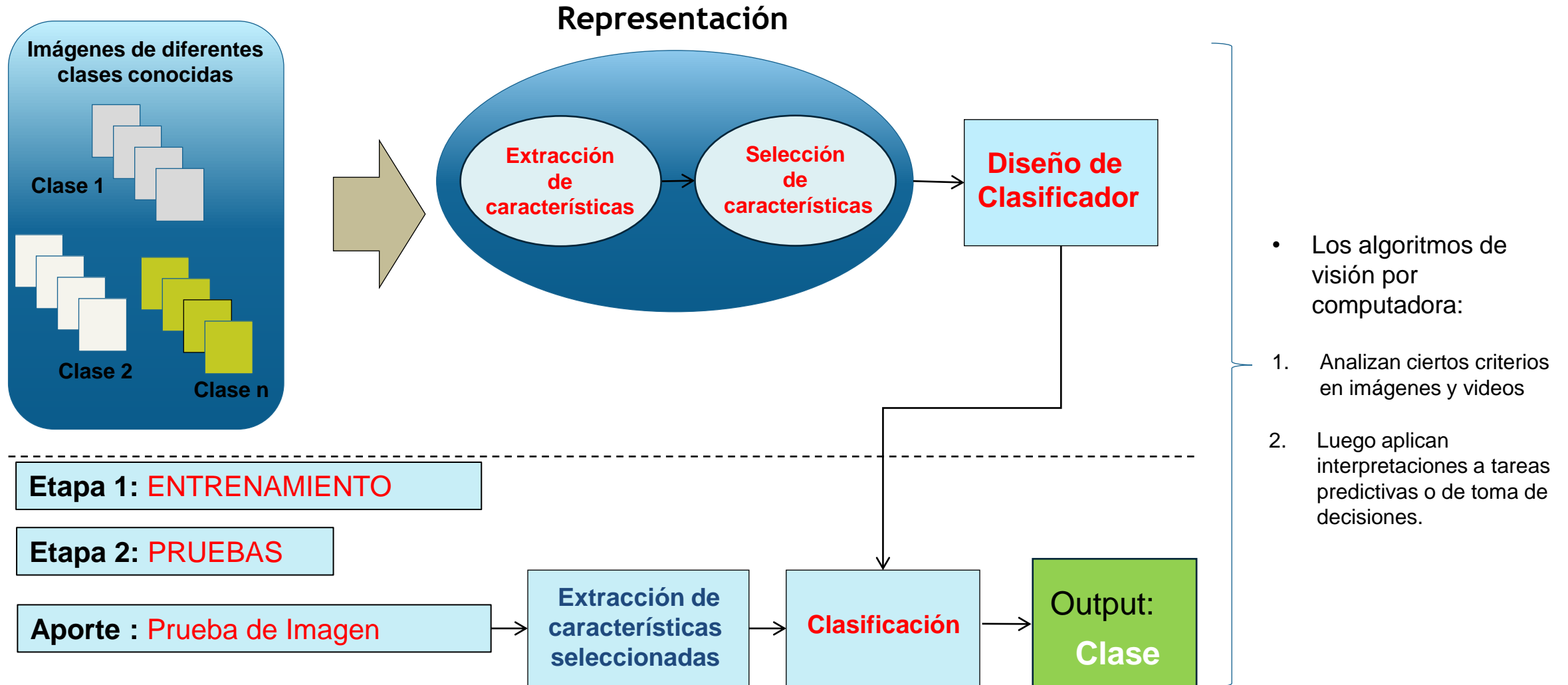
¿Cómo procesar una imagen con redes neuronales?



Cada píxel de la imagen es una entrada (x)



# 3. Revolución del Aprendizaje Profundo



# 4. Redes Neuronales Convolucionales (CNN)

## Redes neuronales convolucionales

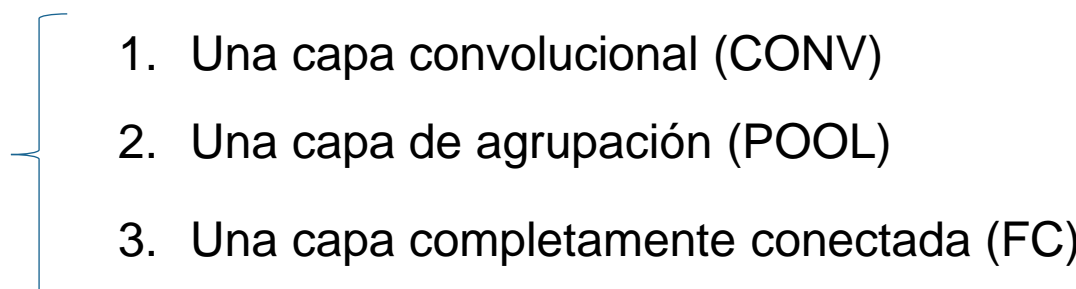
- Los algoritmos de visión por computadora modernos se basan en redes neuronales convolucionales (CNN).
- Una red neuronal convolucional, también conocida como **CNN** o **ConvNet**, es una clase de Red Neuronal que se especializa en procesar datos que tienen una topología similar a una cuadrícula, como una imagen.
- Los **CNN** proporcionan una mejora espectacular en el rendimiento en comparación con los algoritmos de procesamiento de imágenes tradicionales.

### DEFINICION

Las **CNN son redes neuronales** con una arquitectura de múltiples capas que se utiliza para reducir gradualmente los datos y los cálculos al conjunto más relevante. Luego, este conjunto se compara con datos conocidos para identificar o clasificar la entrada de datos.

# 4. Redes Neuronales Convolucionales (CNN)

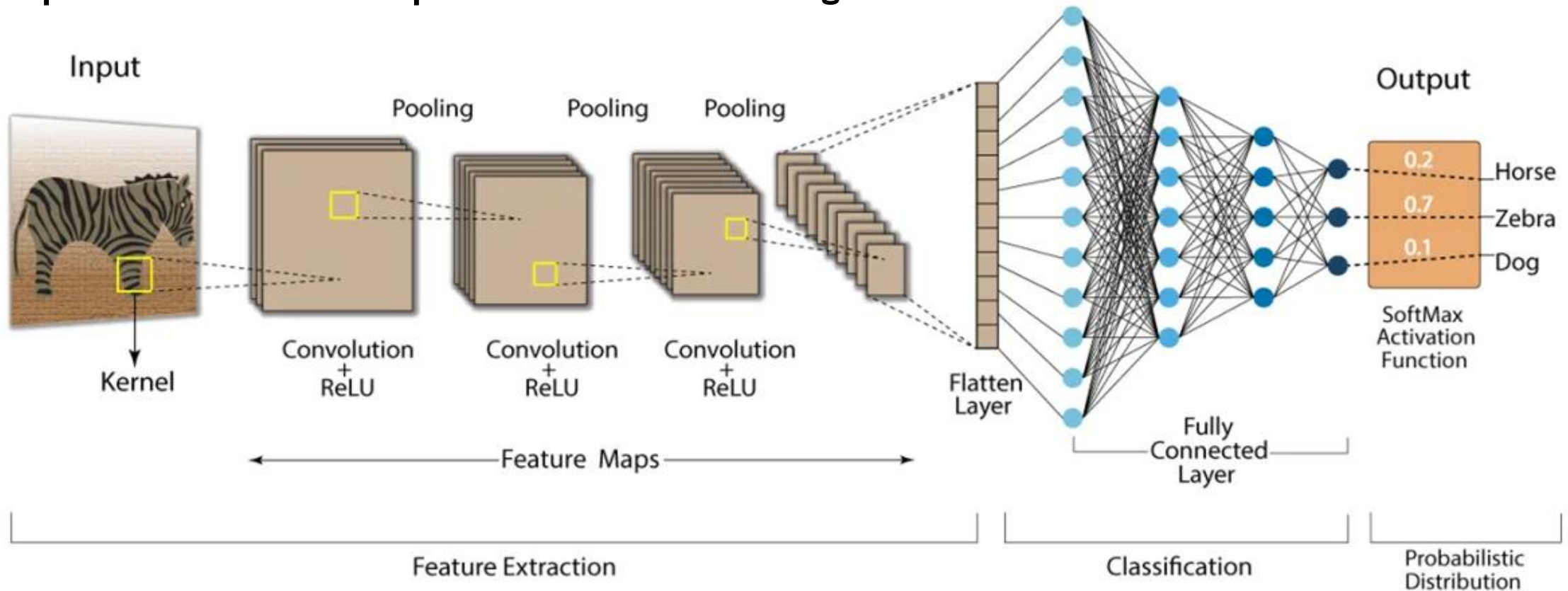
## ¿Cómo las CNN clasifican las imágenes?

- La clasificación de imágenes implica asignar etiquetas o clases a las imágenes de entrada.
- Es una tarea de aprendizaje supervisado en la que se entrena un modelo con datos de imágenes etiquetadas para predecir la clase de imágenes invisibles.
- Las **CNN** se utilizan comúnmente para la clasificación de imágenes, ya que pueden aprender características espaciales/jerárquicas significativas como bordes, texturas y formas, lo que permite un reconocimiento preciso de objetos en las imágenes (aunque también se pueden realizar análisis de texto y audio)
- Una **CNN** suele tener tres capas: 
  1. Una capa convolucional (CONV)
  2. Una capa de agrupación (POOL)
  3. Una capa completamente conectada (FC)



# 4. Redes Neuronales Convolucionales (CNN)

## Arquitectura de una CNN para clasificación de imágenes



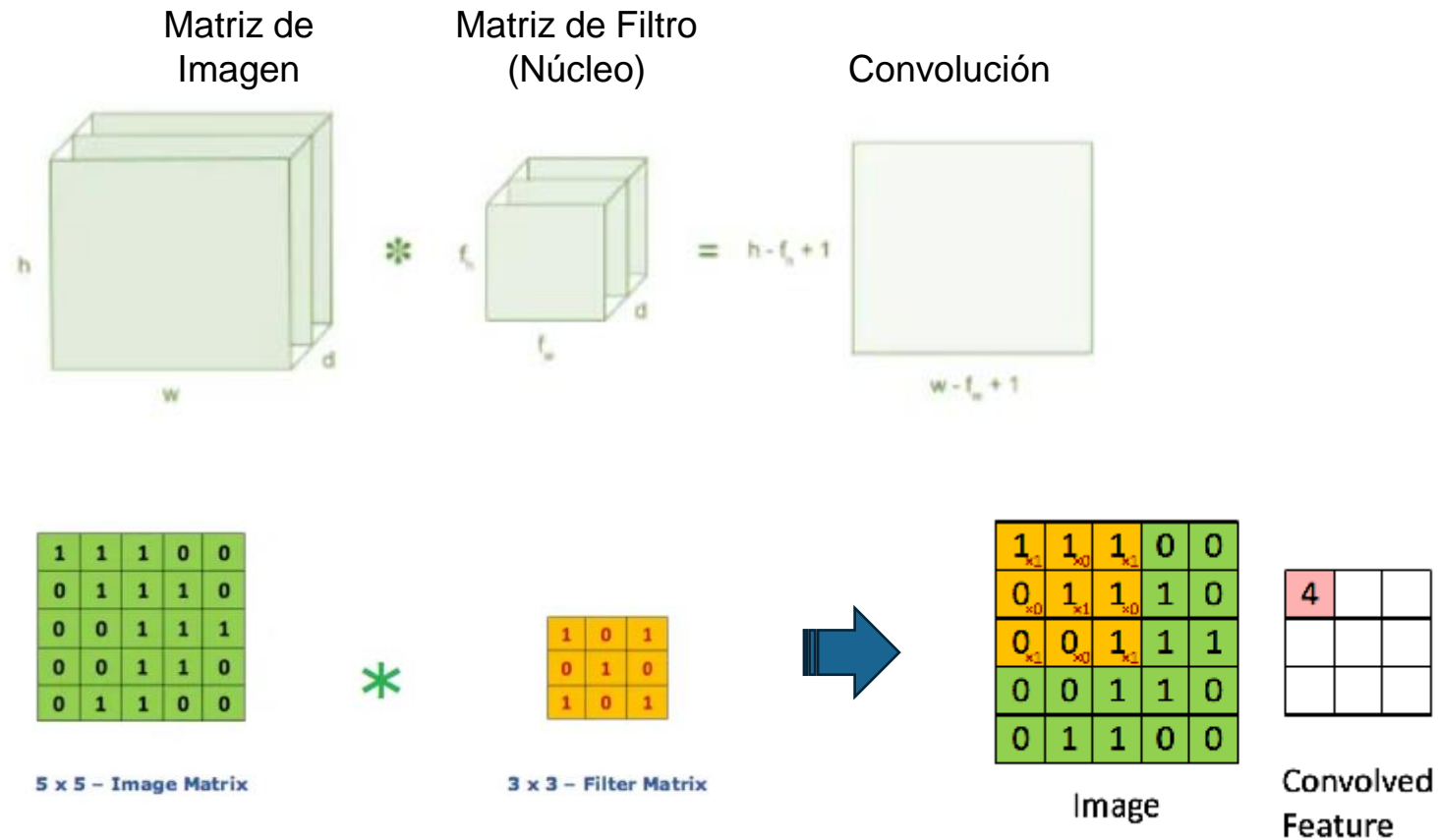
# 4. Redes Neuronales Convolucionales (CNN)

## (1) Capa de convolución

- La capa de convolución es el componente central de la CNN: **Lleva la mayor parte de la carga computacional de la red.**
- La convolución es la primera capa para extraer características de una imagen de entrada.
- La convolución preserva la relación entre píxeles al aprender las características de la imagen utilizando pequeños cuadrados de datos de entrada.
- Es una operación matemática que requiere dos entradas, como una matriz de imagen y un filtro o núcleo.
- El papel de esta primera capa es analizar las imágenes proporcionadas en la entrada y detectar la presencia de un conjunto de features.
- A la salida de esa capa se obtiene un conjunto de features maps (ver más arriba: ¿para qué sirve la convolución?).

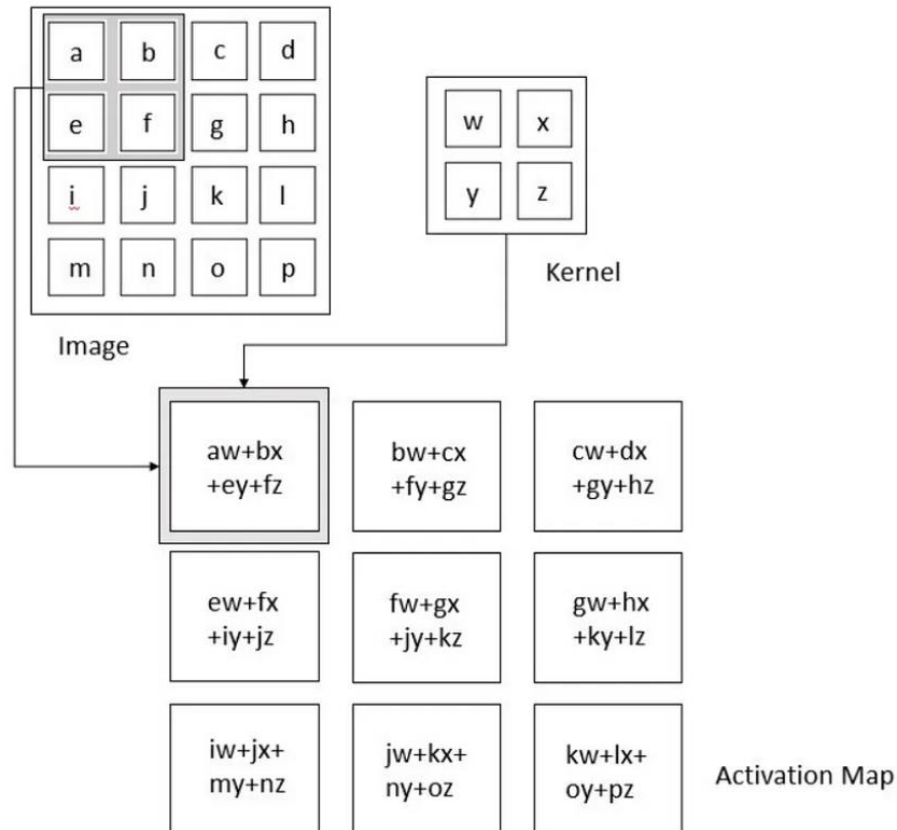
# 4. Redes Neuronales Convolucionales (CNN)

## (1) Capa de convolución



# 4. Redes Neuronales Convolucionales (CNN)

## (1) Capa de convolución



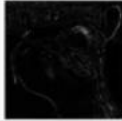
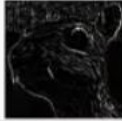


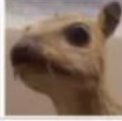


- Durante el pase hacia adelante, el núcleo se desliza a lo largo y ancho de la imagen, produciendo la representación de esa región receptiva.
- Esto produce una representación bidimensional de la imagen conocida como **mapa de activación** que proporciona la respuesta del núcleo en cada posición espacial de la imagen.
- El tamaño de deslizamiento del núcleo se llama zancada.

# 4. Redes Neuronales Convolucionales (CNN)

## (1) Capa de convolución

Algunos filtros  
comunes

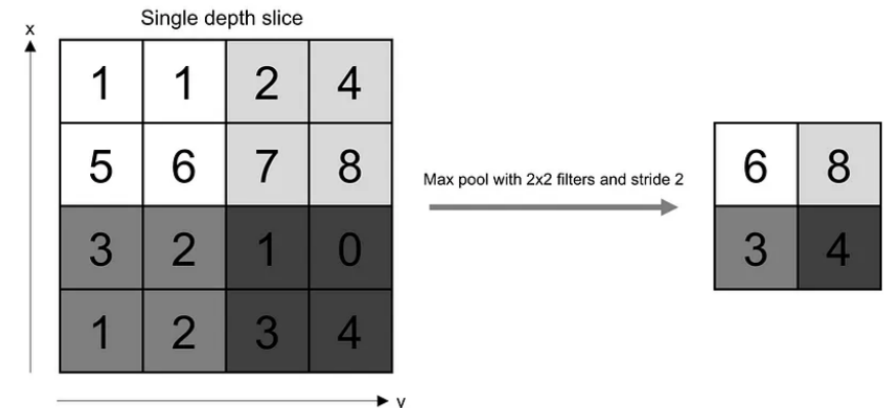
Operation	Filter	Convolved Image
Identity	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	
Edge detection	$\begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix}$	
	$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}$	
	$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$	
Sharpen	$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$	
Box blur (normalized)	$\frac{1}{9} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$	
Gaussian blur (approximation)	$\frac{1}{16} \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}$	

- La convolución de una imagen con diferentes filtros puede realizar operaciones como detección de bordes, desenfoque y nitidez mediante la aplicación de filtros.

# 4. Redes Neuronales Convolucionales (CNN)

## (2) Capa de agrupación (POOL)

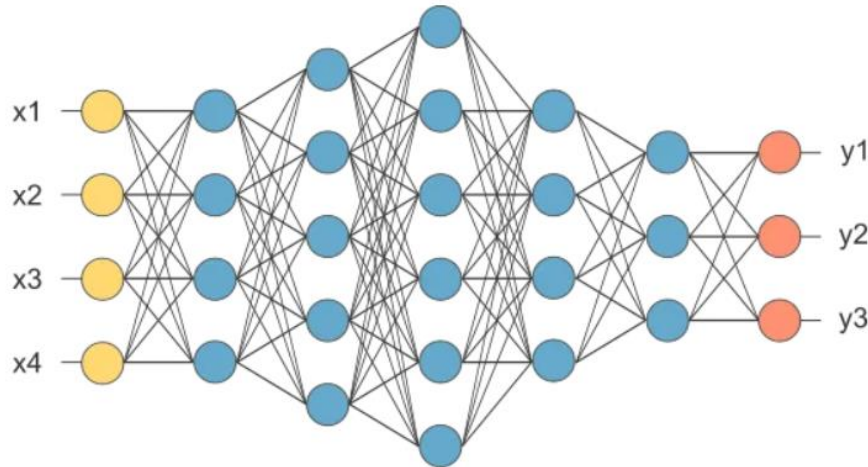
- La capa de Pooling es una operación que por lo general se aplica entre dos capas de convolución.
- Esta recibe en la entrada las features maps formadas en la salida de la capa de convolución y su papel es reducir el tamaño de las imágenes y, a la vez, preservar sus características más esenciales.
- La sección de agrupación de capas reduciría la cantidad de parámetros cuando las imágenes son demasiado grandes.
- La agrupación espacial, también llamada **submuestreo** o **reducción de resolución**, reduce la dimensionalidad de cada mapa pero **retiene información importante**.
- La agrupación espacial puede ser de diferentes tipos:
  - ✓ Agrupación máxima (Max-pooling)
  - ✓ Agrupación promedio (Average pooling)
  - ✓ Agrupación de sumas
- Finalmente, se obtiene en la salida de esa capa de Pooling el mismo número de feature maps que en la salida, pero considerablemente comprimidas.



# 4. Redes Neuronales Convolucionales (CNN)

## (3) Capa completamente conectada (FC)

- En esta capa aplanamos nuestra matriz en un vector y la introducimos en una capa completamente conectada como una red neuronal.
- Las neuronas de esta capa tienen conectividad total con todas las neuronas de la capa anterior y siguiente.
- La capa FC ayuda a mapear la representación entre la entrada y la salida.



- La matriz del mapa de características se convertirá en un vector ( $x_1, x_2, x_3, \dots$ ).
- Con las capas completamente conectadas, combinamos estas características para crear un modelo.

# 4. Redes Neuronales Convolucionales (CNN)

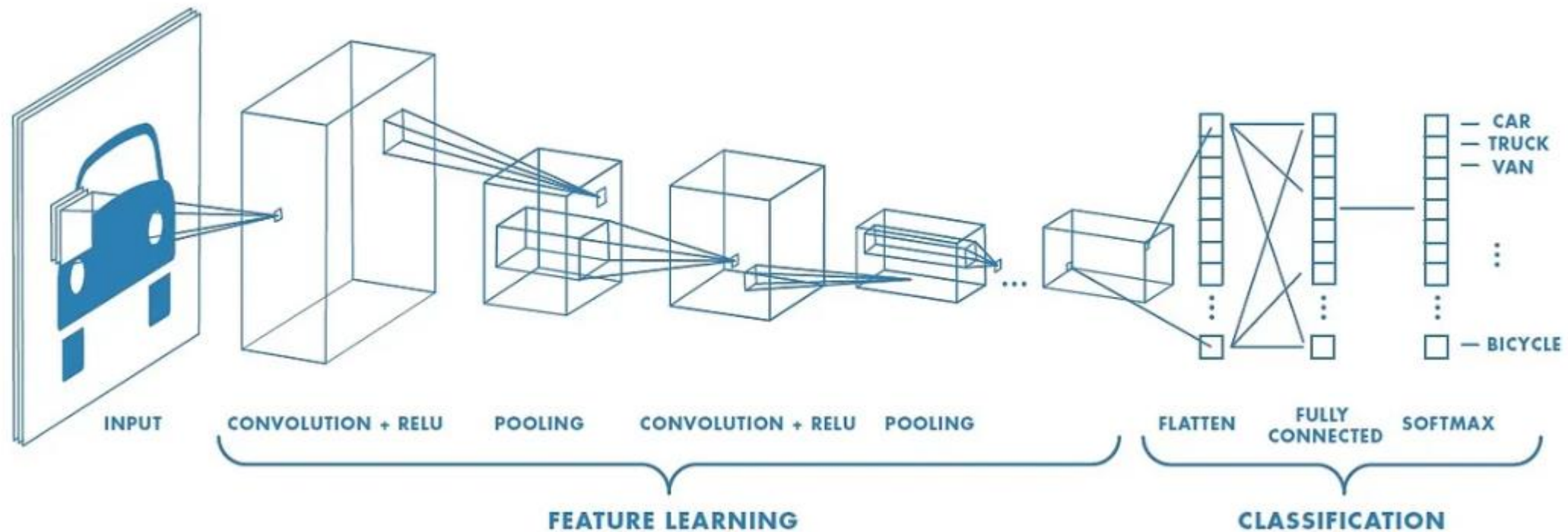
## (4) Capas no lineales

- Dado que **la convolución es una operación lineal** y las imágenes están lejos de ser lineales, las capas de no linealidad (Relu, la mas popular) a menudo se colocan directamente después de la capa convolucional para introducir no linealidad en el mapa de activación.
  - El interés de esas capas de activación es hacer que el modelo sea no lineal y por tanto, más complejo.
  - Existen varios tipos de operaciones no lineales, siendo las más populares:
    1. **Sigmoideo** : Toma un número de valor real y lo "aplasta" en un rango entre 0 y 1.
    2. **Tanh** : aplasta un número de valor real en el rango [-1, 1]
    3. **ReLU (Unidad Lineal Rectificada)**
      - Esta capa sustituye todos los valores negativos recibidos en la entrada por ceros.
      - En comparación con Sigmoide y Tanh, ReLU es más confiable y acelera la convergencia seis veces.
- Calcula la función:  $f(\kappa) = \max(0, \kappa)$  la activación es simplemente un umbral en cero.



## 4. Redes Neuronales Convolucionales (CNN)

Flujo completo de CNN para procesar una imagen de entrada y clasificar los objetos según los valores



# 5. CASOS PRACTICOS



11-02-Clasificador-de Imagenes-Fashion-MNIST.ipynb ☆

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda [Se han guardado todos los cambios](#)



+ Código + Texto



## CC219 - Aplicaciones de Data Science

### ▼ Paso #1: Conjunto de Datos

El conjunto de datos **FASHION MNIST** consta de 70.000 imágenes divididas en 60.000 muestras de entrenamiento y 10.000 muestras de prueba. Una imagen está asociada con una etiqueta de 10 clases.

Las 10 clases son las siguientes:

Label	Description
0	T-shirt/top
1	Trouser
2	Pullover
3	Dress
4	Coat
5	Sandal
6	Shirt
7	Sneaker
8	Bag
9	Ankle boot

A partir de este notebook  
aprenderemos como funciona  
una CNN

# PREGUNTAS

Dudas y opiniones