UNIVERSIDAD DE GRANADA

DECSAI

Del Píxel al Prompt:

la odisea desde el procesamiento a la generación de imágenes en IA

Rafael Molina

Depto de Ciencias de la Computación e IA
Universidad de Granada

rms@decsai.ugr.es https://decsai.ugr.es/vip



Museo Británico: Las Sirenas

AGENDA: CUATRO ERAS DE EVOLUCIÓN

01: El Píxel

Digitalización, Catalizador Espacial, Convolución, Compresión, Problemas Inversos y Ejemplo

02: Visión Artificial Clásica

Puntos de Interés, Descriptores, "Rasgos Diseñados a Mano", Clasificación y Ejemplo

03: La Revolución Neuronal

Hitos, CNNs, ImageNet y AlexNet, Tres Pilares Básicos y Ejemplo

04: Generación

Objetivo, Gans, Vaes, Difusión, Prompts



Lo sé, suena aburrido. Confía en mi

01: EL PIXEL

DIGITALIZACIÓN, CATALIZADOR ESPACIAL, CONVOLUCIÓN, COMPRESIÓN, PROBLEMAS INVERSOS Y EJEMPLO

1957: EL PRIMER PÍXEL: CUANDO LAS MATEMÁTICAS REEMPLAZARON AL CUARTO OSCURO



Fuente: Wikipedia.org

Russell Kirsch (1929-2020) escaneó una foto de su hijo Walden. Fue la primera vez que la realidad visual fue traducida a unos y ceros.

- Resolución: 176 x 176 **píxeles**
- Profundidad: 1 bit (Blanco y negro)



Fuente: Diputación Provincial de Granada

Esto no quiere decir, obviamente, que no hubiera antes fotografías o imágenes. **No estaban en formato digital**

1964: EL CATALIZADOR ESPACIAL: EL ORIGEN DE LA PALABRA PÍXEL

La sonda espacial Ranger 7 transmitió 4.308 fotos lunares antes del impacto.

La palabra píxel fue propuesta en 1965 por Frederic C. Billingsley, del Jet Propulsion Laboratory, JPL de la NASA, para describir los puntos de imagen en las transmisiones lunares:

- Viene de la unión de los términos en inglés **pix** (abreviatura coloquial de picture, imagen) y **el** (de element, elemento).



Fuente: NASA/JPL-Caltech

Ranger 7, primera imagen de la Luna de una sonda de EE. UU., el 31 de julio de 1964, 13:09 UT, unos 17 minutos antes de impactar en la superficie lunar.

LA ESTRUCTURA DEL PÍXEL

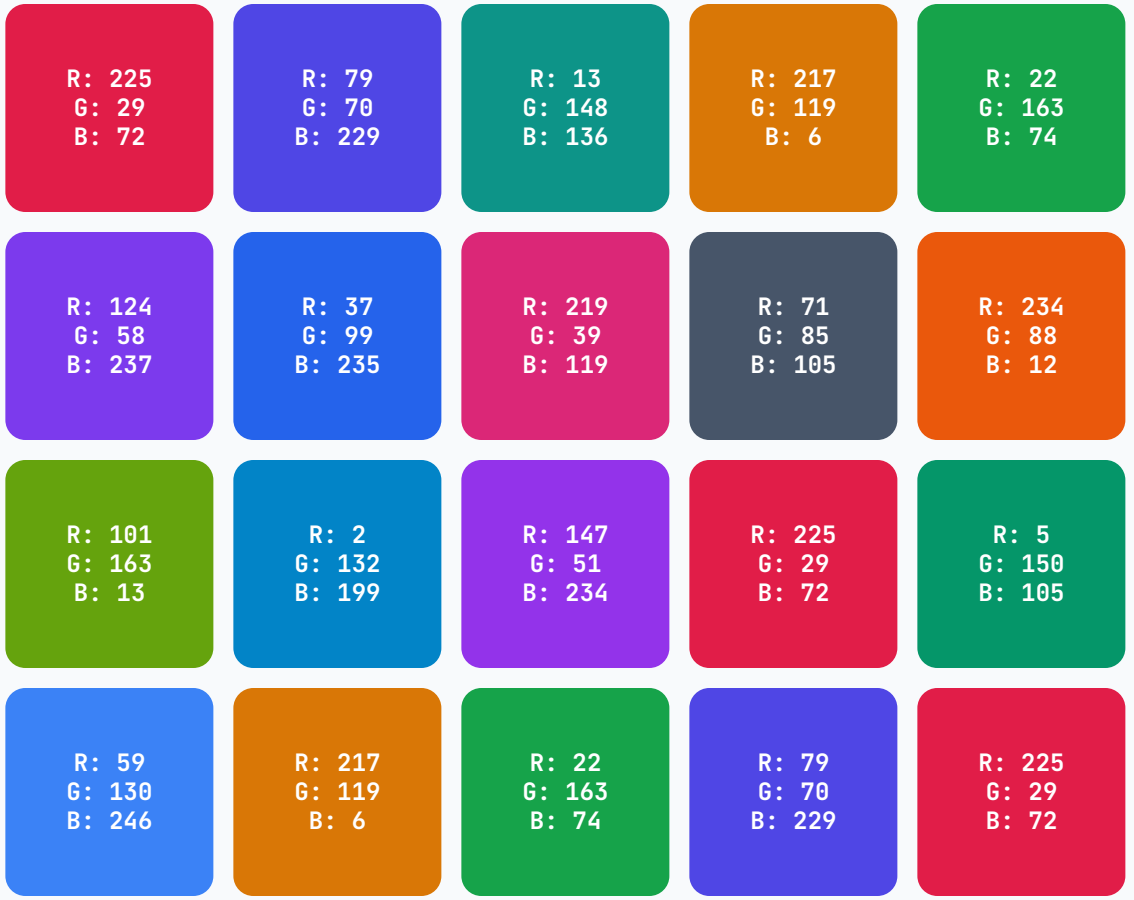
En el ordenador, las imágenes se representan como una matriz. Cada celda funciona como una unidad de visualización de color única, con canales **R**, **G** y **B** que controlan su color.

Esta abstracción discreta constituye la base para construir y procesar cualquier imagen digital.

Importante: para un ordenador, una imagen no es más que una cuadrícula de números, normalmente entre 0 y 255.

RGB

Los monitores usan el modelo Rojo, Verde, Azul pero hay otros muchos.



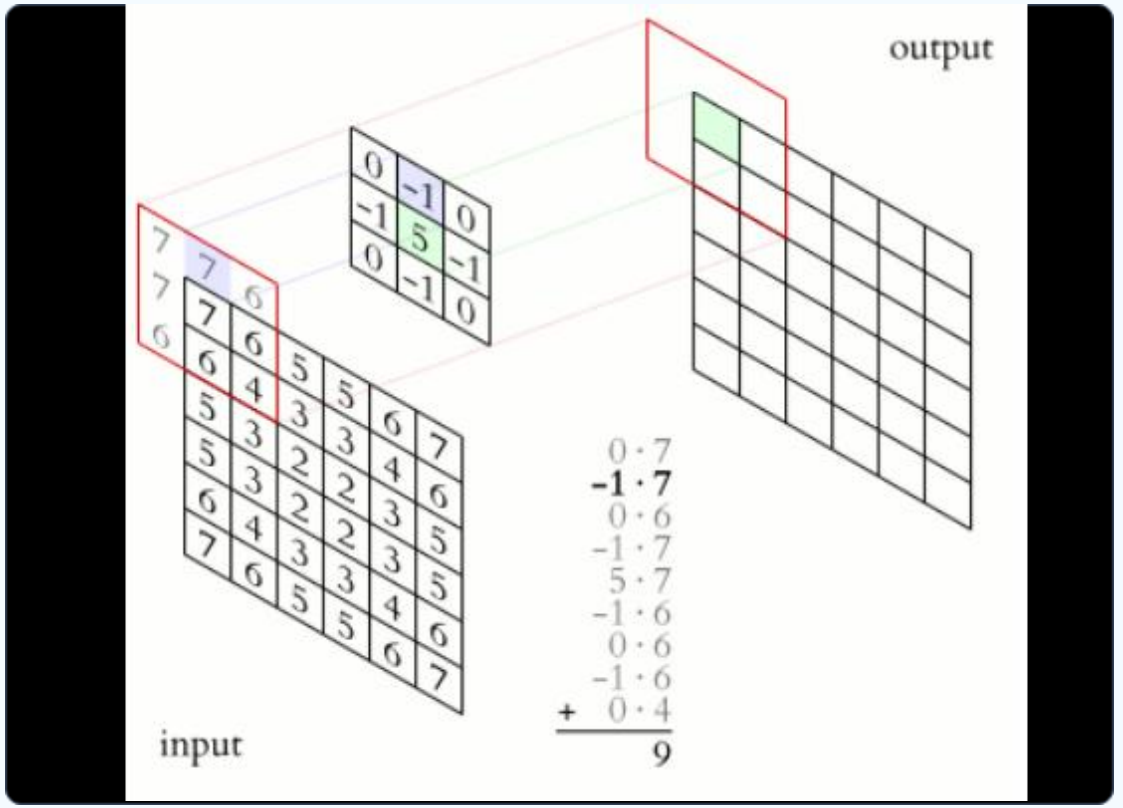
Representación visual de una matriz (imagen) de resolución de 5 columnas y 4 filas.

LA  DE LA ERA DE LOS PÍXELES : LA CONVOLUCIÓN

Para modificar una imagen, se usaba fundamentalmente un Núcleo (**Kernel)**: una pequeña matriz que se desliza sobre el "campo de batalla" de pixeles produciendo una **imagen filtrada**

$$(f * g)(x, y)$$

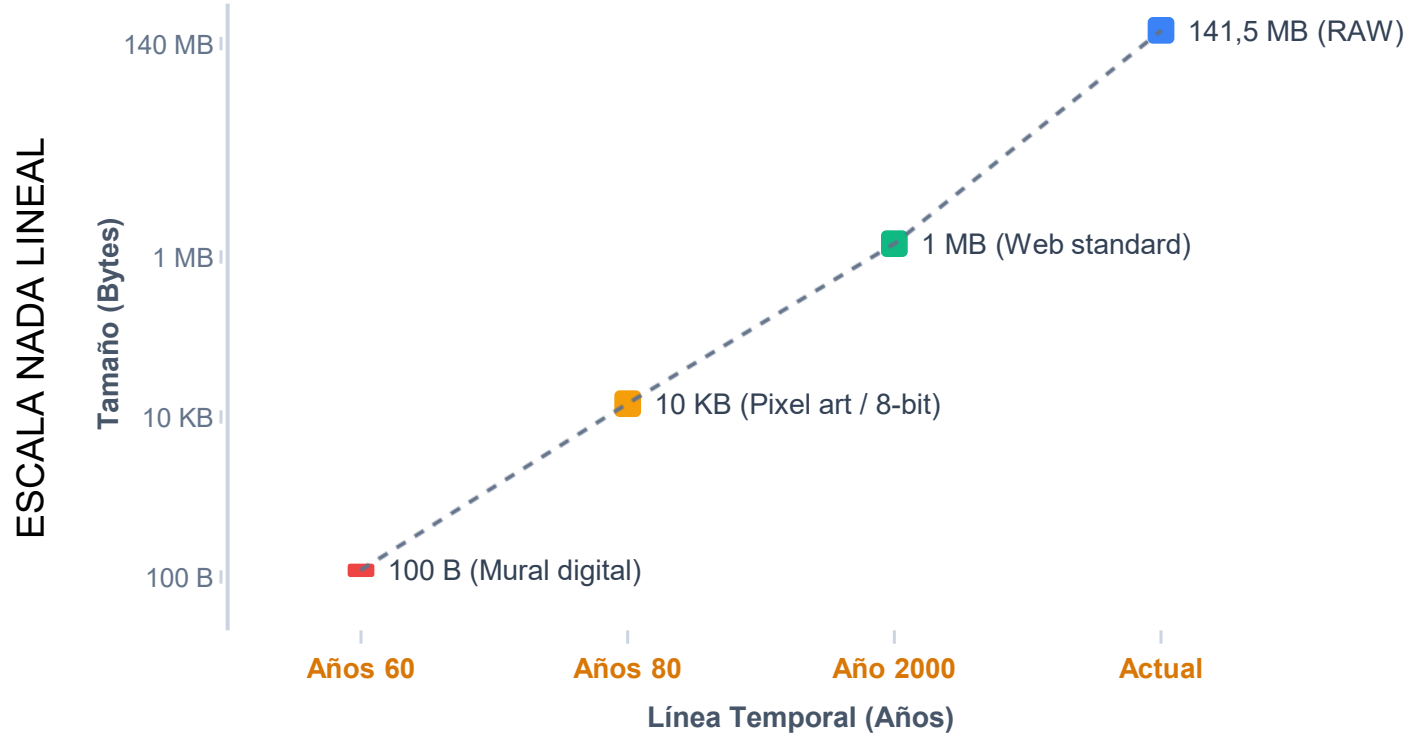
Cada nuevo píxel es la suma ponderada de sus vecinos. Los coeficientes del núcleo determinan su objetivo.



Cuando la reina extienda su reinado por las redes neuronales y el aprendizaje profundo (deep learning) aprenderá los pesos de los kernels y se unirá a otros aliados …

¿CÓMO ESTÁN LOS MÓVILES? LLENOS

Comparativa histórica del espacio en bytes requerido para almacenar una imagen típica (años 60 vs actual 8K)





https://www.pexels.com/es-es/

IMAGEN ULTRADETALLADA 8K

- Resolución: 7680 × 4320 píxeles
- Profundidad: 33.1 megapíxeles
- Tamaño RAW: ~141.5 MB por fotograma

- **Logro (y problema):** El salto tecnológico de 60 años ha multiplicado los requisitos de almacenamiento **más de 1.4 millones de veces.**
- **Tecnología posibilitadora:** La **compresión** de imágenes y vídeos ha permitido la reducción drástica del peso de dichas imágenes y vídeos.
- La diferencia entre las imágenes y vídeos originales y sus versiones comprimidas es **visualmente imperceptible.**

PROBLEMAS INVERSOS EN IMÁGENES
(A RESOLVER EN LA ERA DEL PÍXEL)

El problema inverso en imágenes consiste en **estimar la causa u objeto original** (imagen verdadera) a partir de una **imagen observada degradada** por algún proceso.

IMAGEN ORIGINAL



$f(x, y)$

PROCESO DE DEGRADACIÓN

- Desenfoque
- Ruido
- Baja resolución
- Distorsión

IMAGEN OBSERVADA
(DEGRADADA)



$g(x, y)$

MODELO MATEMÁTICO (PROBLEMA DIRECTO)

$$g(x, y) = H\{f(x, y)\} + \eta(x, y)$$

g(x,y): Obs. | H: Degradación | η(x,y): Ruido

IMAGEN OBSERVADA
(DEGRADADA)



$g(x, y)$

PROBLEMA INVERSO

A partir de g(x,y) y del proceso H, estimar la imagen original.

IMAGEN ESTIMADA
(RESTAURADA)



$\hat{f}(x, y)$

OBJETIVO

Encontrar $\hat{f}(x, y) \approx f(x, y)$

que mejor explique la imagen observada g(x, y).

💡 **Idea clave:** en el problema inverso buscamos "deshacer" los efectos de la degradación para recuperar la imagen original.

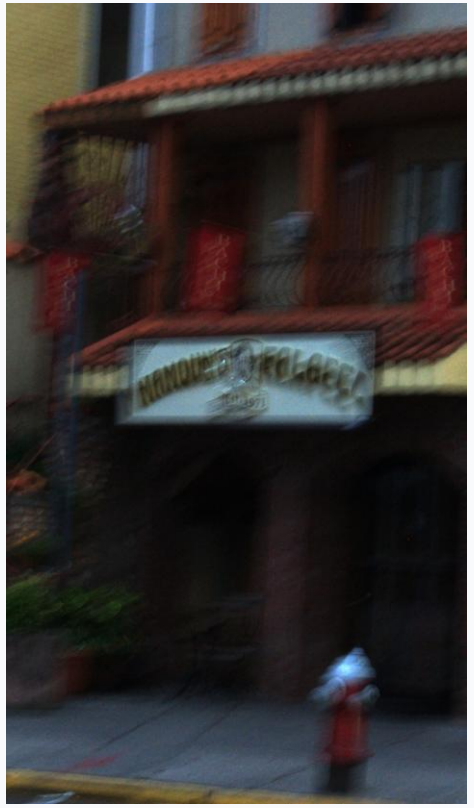
ALGUNOS EJEMPLOS DE PROBLEMAS INVERSOS

Dos imágenes borrosas y ruidosas

Ejemplo 1



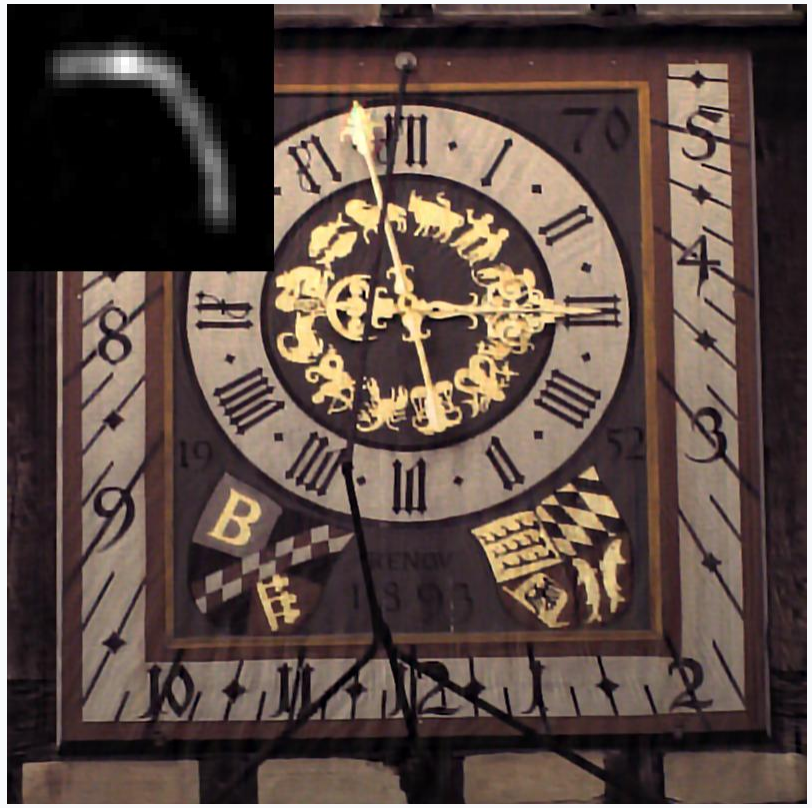
Ejemplo 2



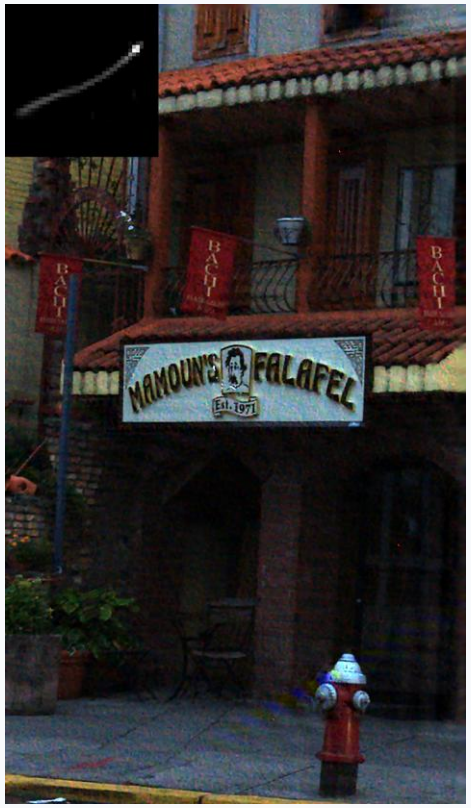
RESTAURACIONES

Sus restauraciones y la estimación de los emborronamientos (cuadrados con fondo negro)

Ejemplo 1



Ejemplo 2





02

Visión Artificial Clásica

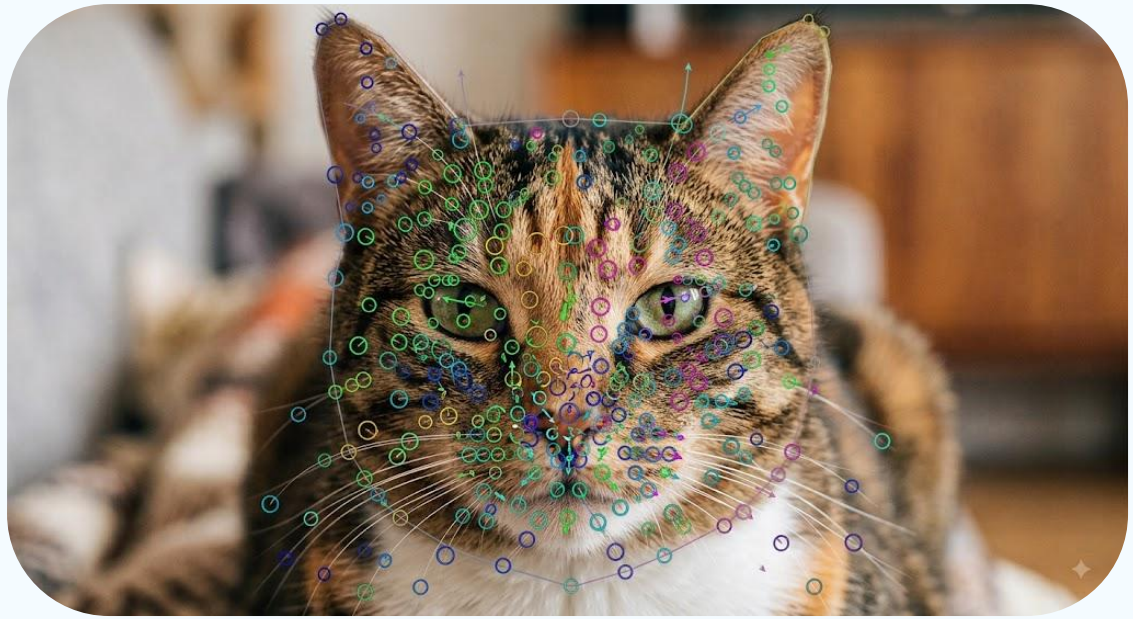
PUNTOS DE INTERÉS, DESCRIPTORES, RASGOS "DISEÑADOS A MANO", CLASIFICACIÓN Y EJEMPLO

DETECCIÓN DE PUNTOS DE INTERÉS (KEYPOINT DETECTION)

¿Qué es un Punto de Interés?

Es una coordenada espacial (x, y) en la imagen que presenta una variación bidireccional clara de intensidad.
No se busca procesar todos los píxeles, sino seleccionar únicamente las regiones altamente informativas y estables

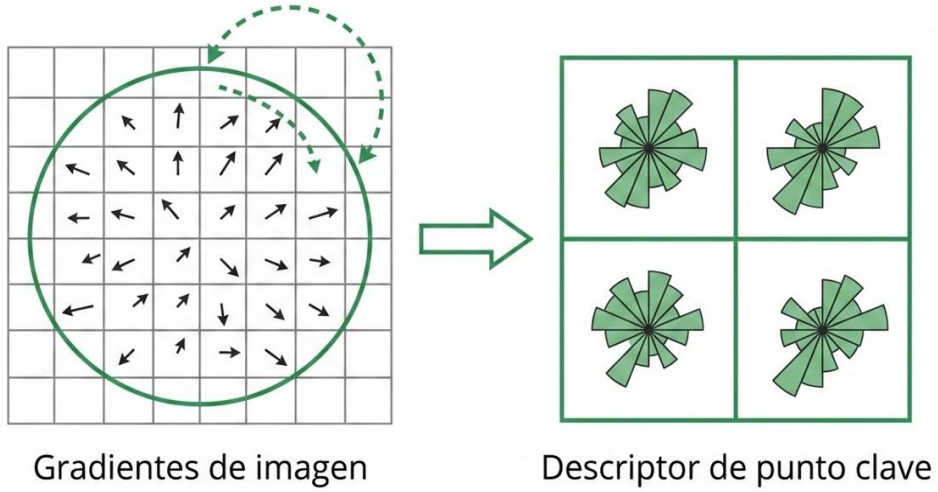
- Un detector debe localizar exactamente el mismo punto clave en la imagen a pesar de cambios de iluminación, escala, perspectiva o rotación.
- Debe situarse en zonas con estructura compleja (esquinas, bordes cruzados, manchas) descartando regiones planas o de gradiente uniforme.
- La posición del punto clave debe estar acotada espacialmente de forma rigurosa para permitir la comparación de parches locales.

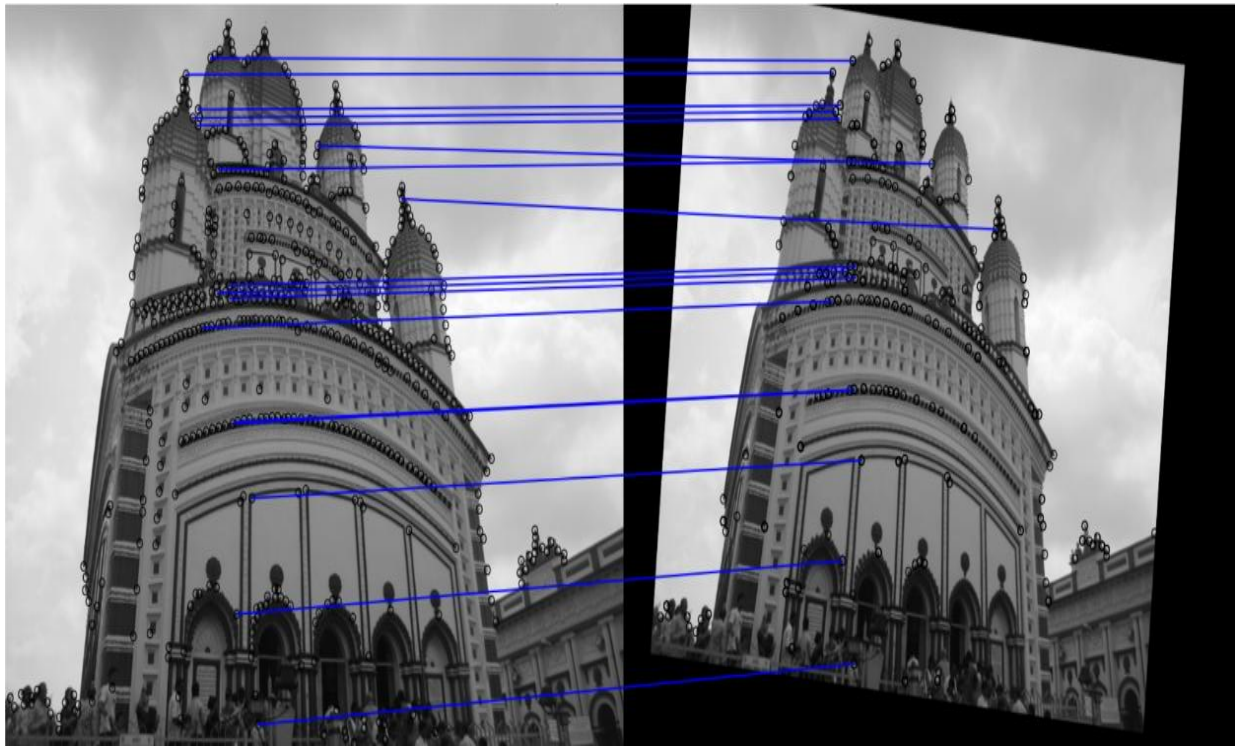


Fuente: GenAI

DESCRIPTORES LOCALES DE GRADIENTE: SIFT (1999)

Scale-Invariant Feature Transform. Permite detectar objetos y extraer características de la imagen robustas a iluminación e invariantes a escala.





APLICACIÓN: COSIENDO REALIDADES

Gracias a los descriptores como SIFT, tu móvil puede hacer fotos panorámicas. El software encuentra los mismos keypoints en diferentes fotos y los alinea con precisión sub-píxel.



CLASIFICACIÓN CLÁSICA: EL PIPELINE TRADICIONAL

A diferencia del aprendizaje profundo, un clasificador clásico (como SVM o Random Forest) no aprende directamente de los píxeles. Requiere extraer descriptores matemáticos diseñados a mano antes de poder tomar una decisión.

1. Entrada

Se introduce la imagen original de un objeto (por ejemplo, un gato) representada únicamente por sus valores de píxeles RGB.



2. Extracción

Algoritmos como **SIFT o HOG** analizan cambios de contraste locales y bordes para extraer vectores de rasgos matemáticos invariantes.

Así pues, los píxeles se convierten en descriptores geométricos y de textura.

3. Vectorizado

Se agrupan los descriptores para generar un único vector de características numéricas (histograma o Bag of Visual Words).

El objeto se reduce a un punto en un espacio multidimensional.

```
f = [0.12, 0.85, ... 0.43]
```

4. Clasificación

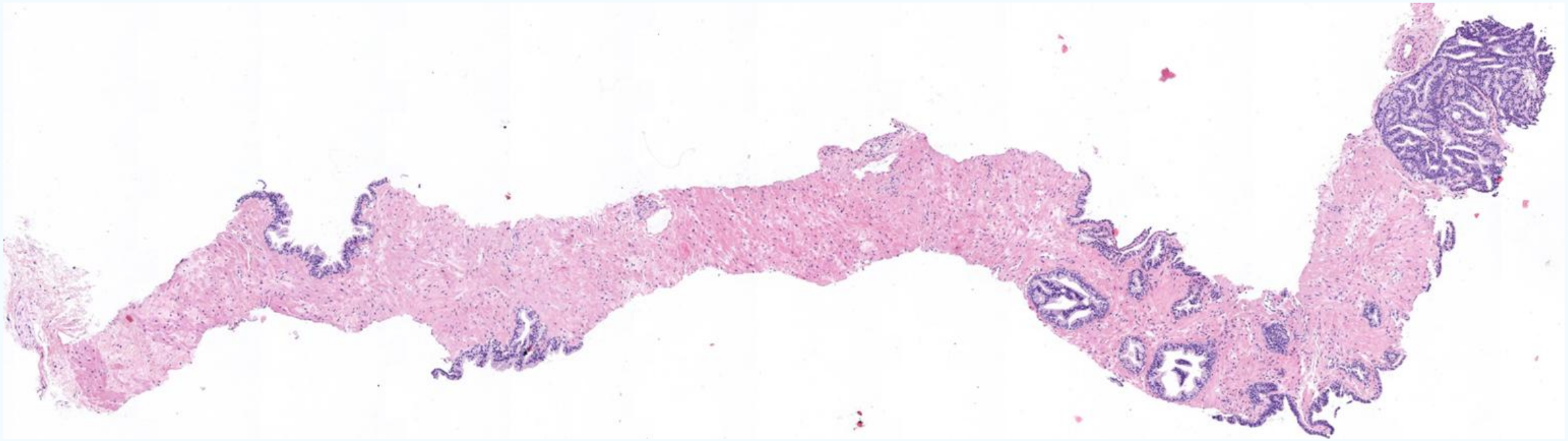
Un clasificador estadístico (**SVM** con hiperplano de decisión, o un árbol) evalúa el vector numérico.

Produce la etiqueta final basándose en reglas de decisión matemáticas precalculadas.

PREDICCIÓN FINAL:

"Gato" (94%)

EJEMPLO: CLASIFICACIÓN DE WHOLE SLIDE IMAGES (WSIs)



Muestra bajo análisis

Una imagen histológica de próstata completa para evaluación digital.

Desafío diagnóstico

¿Cómo clasificamos automáticamente esta imagen (sus regiones) de forma eficaz?

METODOLOGÍA

1. Paradigma Clásico

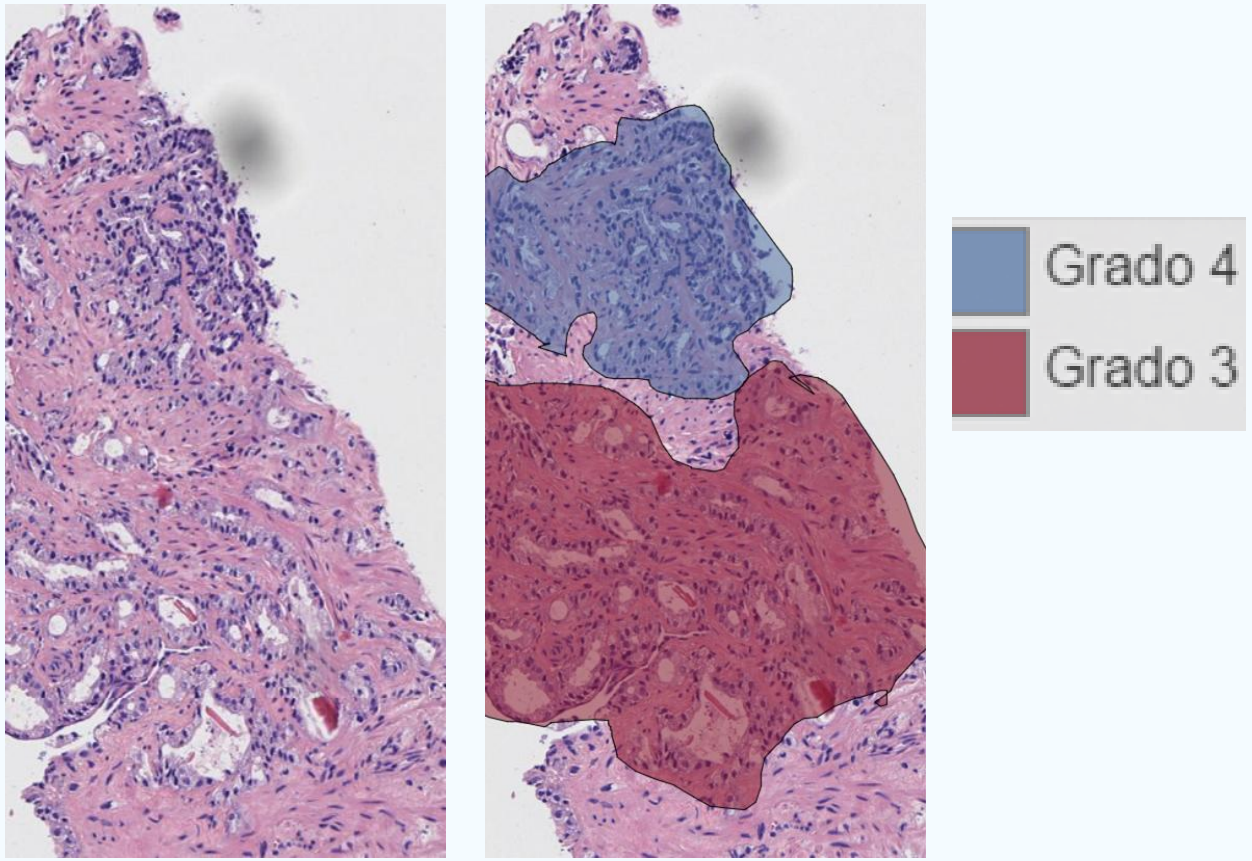
En el paradigma más clásico se dispone de anotaciones expertas de un número de imágenes reales.

2. División en Parches

Se divide la imagen en parches (patches) y de ellos se extraen rasgos. En este caso morfológicos definidos explícitamente para este problema.

3. Clasificador de Malignidad

Usando las anotaciones expertas, se construye un clasificador que tiene como entrada los rasgos morfológicos por parche y como salida el grado de malignidad del parche en escala de Gleason, 1-2 benigno, 3-5 maligno.



CLASIFICACIÓN

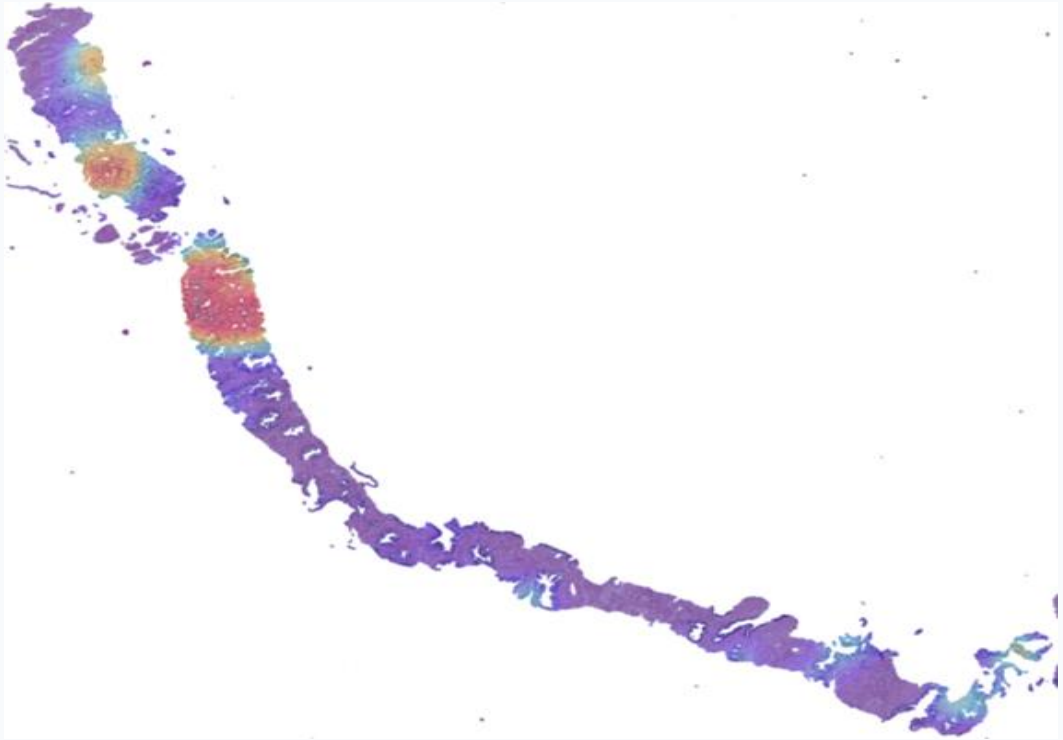
Anotación de Expertos (Ground Truth)

Áreas cancerosa delimitadas con anotaciones expertas



Clasificación Automática

La Intensidad del rojo indica la probabilidad (asignada por el clasificador) de que el área sea cancerosa





03

La Revolución Neuronal

HITOS, REDES NEURONALES CONVOLUCIONALES, IMAGENET Y ALEXNET, TRES PILARES BÁSICOS Y EJEMPLO

HITOS Y EVOLUCIÓN HISTÓRICA DE LAS ANNs

1943
Modelo de McCulloch-Pitts

Primera formulación matemática de una neurona artificial, sentando las bases teóricas de la computación neuronal basada en lógica binaria.

1958
El Perceptrón de Rosenblatt

Primer modelo de red capaz de aprender de forma supervisada a clasificar patrones mediante el ajuste de pesos sinápticos.

1986
Algoritmo Backpropagation

Popularizado por Rumelhart, Hinton y Williams. Permitió entrenar eficazmente redes multicapa (MLPs), resolviendo limitaciones previas.

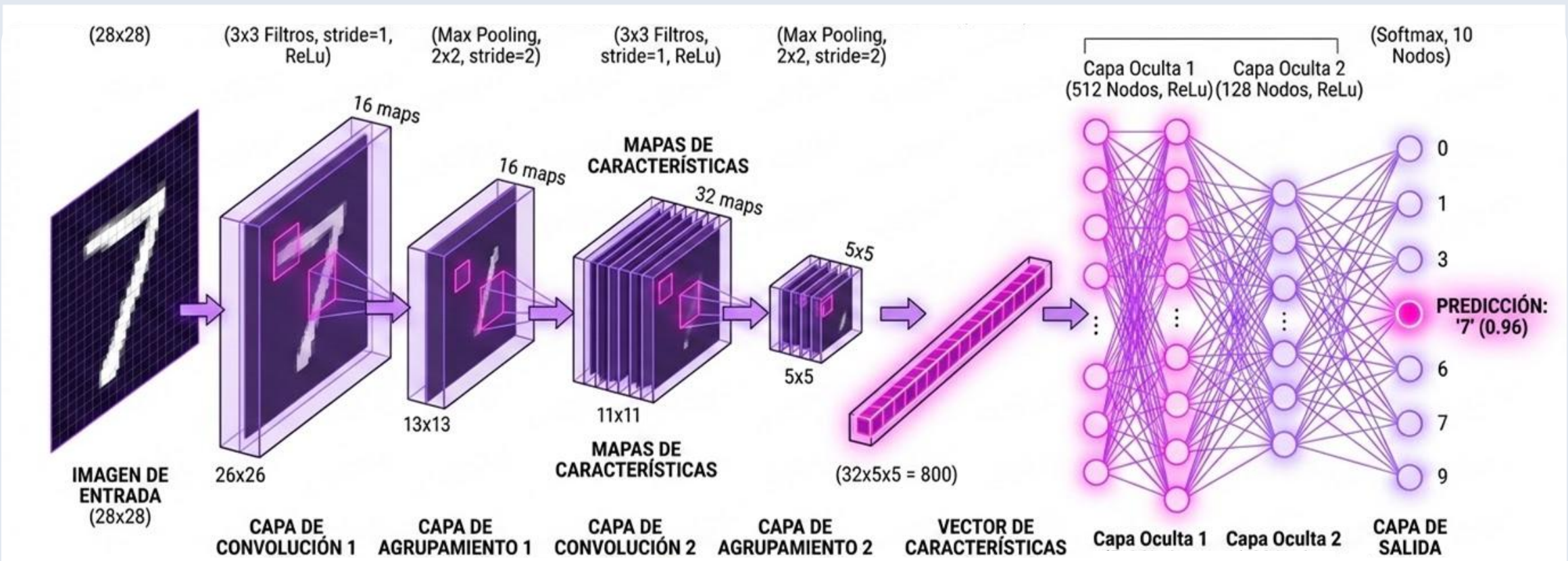
1998
LeNet-5 (Yann LeCun)

Introducción de las Redes Neuronales Convolucionales (CNNs). Éxito pionero en el reconocimiento de dígitos manuscritos.

2012
Revolución Deep Learning

AlexNet gana el desafío ImageNet por un amplio margen, impulsando la era moderna del aprendizaje profundo y el uso de GPUs.

ARQUITECTURA DE UNA RED NEURONAL CONVOLUCIONAL (CNN)



Arquitectura básica de una CNN (Filtros y Mapas de Características)

Fuente: GenAI

LA REINA CONVOLUCIÓN Y SU CORTEJO

CONVOLUCIÓN



Fuente: GenAI

CONVOLUCIÓN, RELU, MAXPOOLING, SKIP-CONNECTION, …



Fuente: GenAI

IMAGENET: LA BASE DE DATOS QUE CAMBIÓ LA IA



¿QUÉ ES IMAGENET?

Un Enorme Catálogo Visual

Fundada por la investigadora **Fei-Fei Li** en 2009, ImageNet es una base de datos a gran escala diseñada para el entrenamiento de software de reconocimiento visual de objetos.

Cuenta con más de **14 millones de imágenes** meticulosamente etiquetadas a mano utilizando el sistema jerárquico de WordNet, organizadas en más de **20,000 categorías** distintas.

Su creación resolvió el principal cuello de botella de la época: la falta de datos masivos y estructurados para que los modelos de aprendizaje profundo pudieran generalizar y aprender patrones complejos.

¿QUIÉN GANÓ LA COMPETICIÓN DE IMAGENET EN 2012?

La ganó **AlexNet** que usó GPUs para entrenar una red de 8 capas. Su error de clasificación fue casi la mitad del que obtuvo el segundo clasificado.. Éste usaba métodos clásicos (de la sección anterior)

Fue el momento en que el mundo entendió que los datos y el hardware habían superado a las fórmulas manuales.

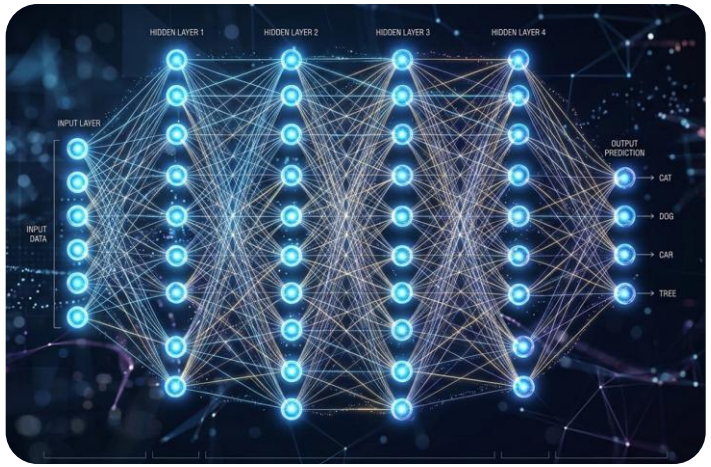
LA REVOLUCIÓN DE LAS ANNs: LOS TRES PILARES CLAVE



01 / DATOS MASIVOS

Cantidad Ingente de Datos

La disponibilidad de datasets masivos y estructurados como ImageNet permitió que los modelos complejos de aprendizaje profundo pudieran entrenarse sin sobreajustarse, superando el histórico cuello de botella de datos.



02 / ARQUITECTURAS

Algoritmos y Redes Convolucionales

El desarrollo y optimización de redes neuronales artificiales (ANNs) más profundas y eficientes, junto con técnicas de regularización avanzadas, sentaron las bases metodológicas del despegue.



03 / HARDWARE POTENTE

Cómputo en GPUs

La transición del procesamiento en CPU al uso de unidades de procesamiento gráfico (GPUs) multiplicó exponencialmente la velocidad de cálculo matemático paralelo, haciendo viable el entrenamiento de redes profundas.

Fuente: GenAI

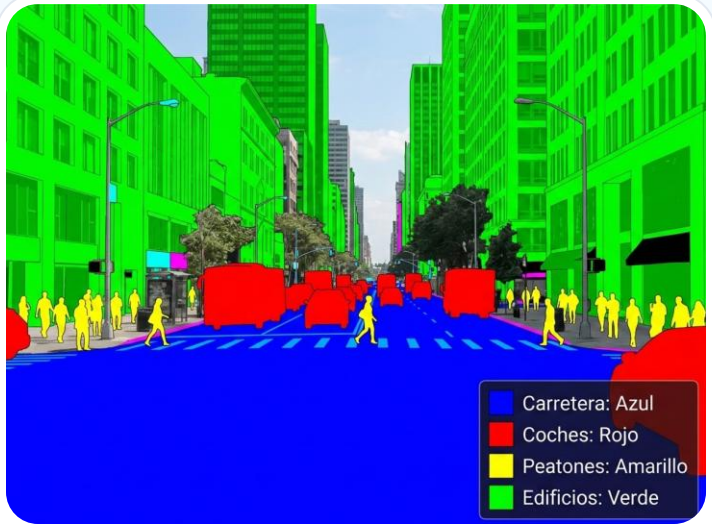
APLICACIONES: USOS DE LAS CNNs EN EL MUNDO REAL



RESTAURACIÓN

Restauración de Imágenes

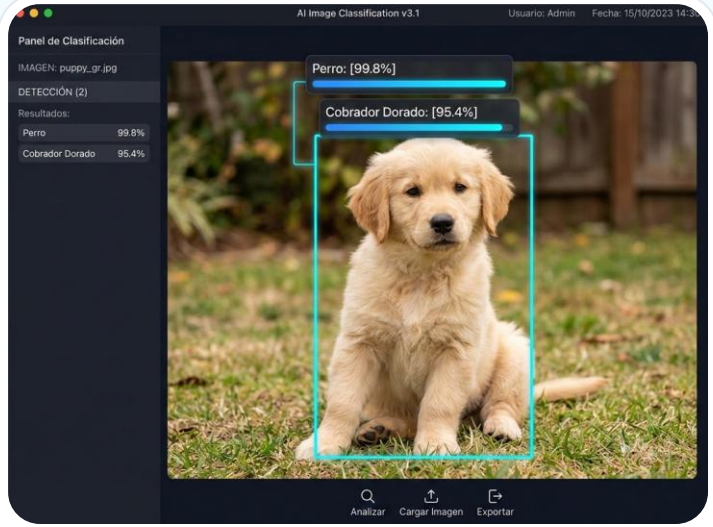
Eliminación de ruido, superresolución y reconstrucción de detalles perdidos o dañados en fotografías antiguas, médicas o satelitales mediante modelos convolucionales avanzados.



SEGMENTACIÓN

Segmentación Semántica

Clasificación de cada píxel de una imagen en una categoría específica, permitiendo delimitar con precisión quirúrgica carreteras, peatones o tumores en entornos médicos.



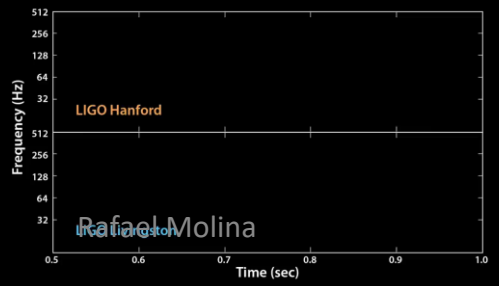
CLASIFICACIÓN

Clasificación de Imágenes

Identificación y asignación de etiquetas globales a imágenes completas, permitiendo organizar catálogos masivos, moderar contenidos o realizar diagnósticos automatizados rápidos.

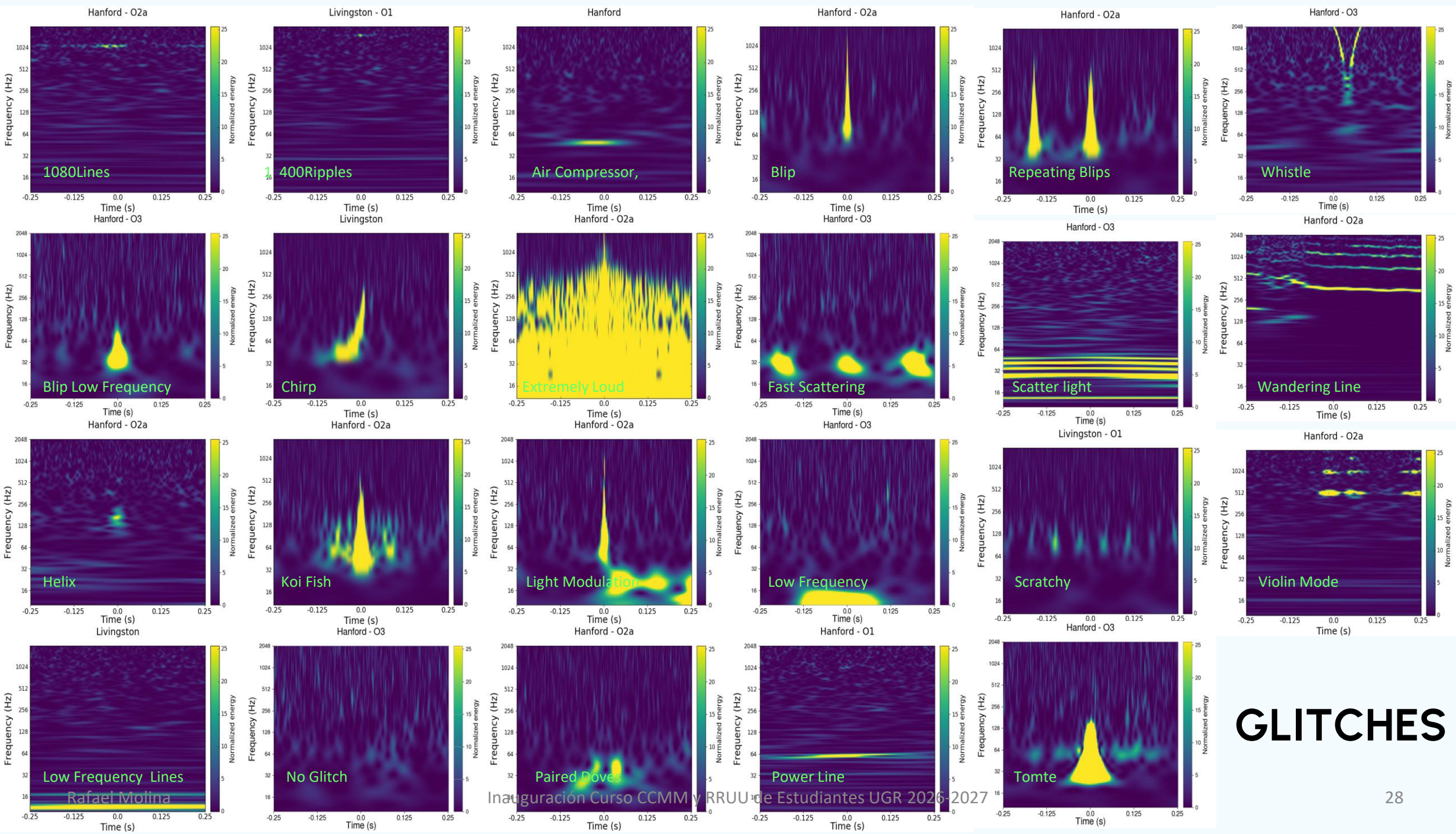
Gravitational waves are “ripples in spacetime” predicted by Einstein a century ago.

Gravitational waves enable us to “hear” the most violent phenomena in the universe



¿Cómo podemos detectar glitches?

Credit: Aurore Simonnet


1080Lines
400Ripples
Air Compressor,
Blip
Repeating Blips
Whistle
Blip Low Frequency
Chirp
Extremely Loud
Fast Scattering
Scatter light
Wandering Line
Helix
Koi Fish
Light Modulation
Low Frequency
Scratchy
Violin Mode
Low Frequency Lines
No Glitch
Paired Doves
Power Line
Tomte
Frequency (Hz)
Time (s)
Normalized energy

GLITCHES

GRAVITY SPY

Proyecto de Ciencia Ciudadana

Objetivo

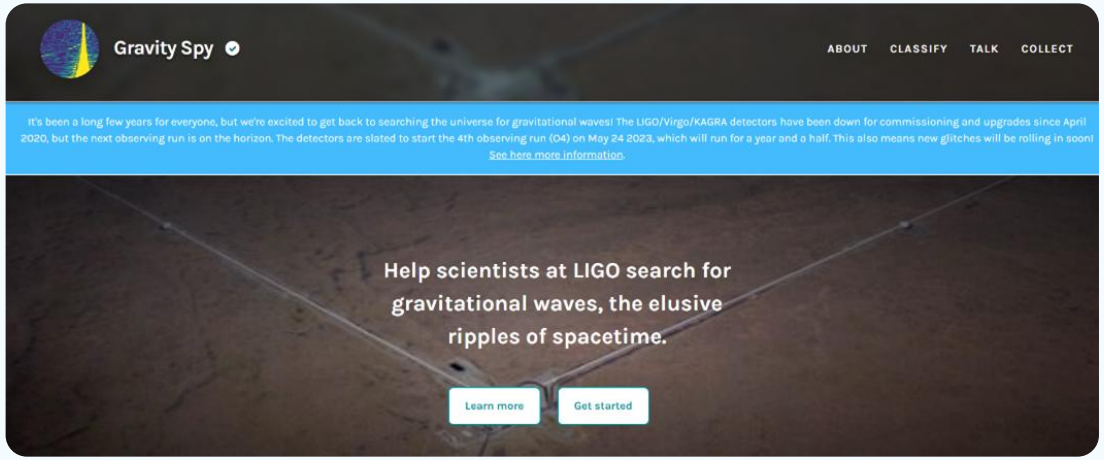
Clasificar y categorizar diferentes tipos de glitches que ocurren en el detector de ondas gravitacionales.

Enfoque Multidisciplinar

LIGO integra el trabajo de expertos científicos, algoritmos de aprendizaje automático y la colaboración de voluntarios.

Aproximación

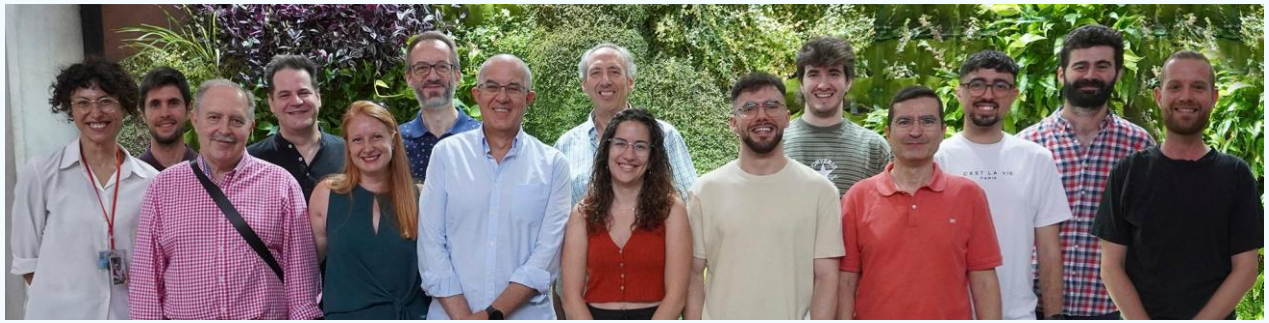
Rasgos extraídos por una CNN y clasificación (3.608.733 ejemplos) basada en técnicas crowdsourcing



Fuente: gravityspy.org

INFOMACIÓN SOBRE ESTE Y LOS OTROS EJEMPLOS

Podéis encontrarla en la página del grupo Procesamiento de Información Visual (https://decsai.ugr.es/vip)





04

Generación de Imágenes (con Prompts)

OBJETIVO, GANS, VAES, DIFUSIÓN, PROMPTS

¿ES POSIBLE CREAR IMÁGENES DESDE LA NADA?

La Perspectiva Tradicional

Hasta ahora, en nuestro recorrido por el procesamiento y la descripción semántica, siempre hemos partido de una imagen real existente para su análisis.

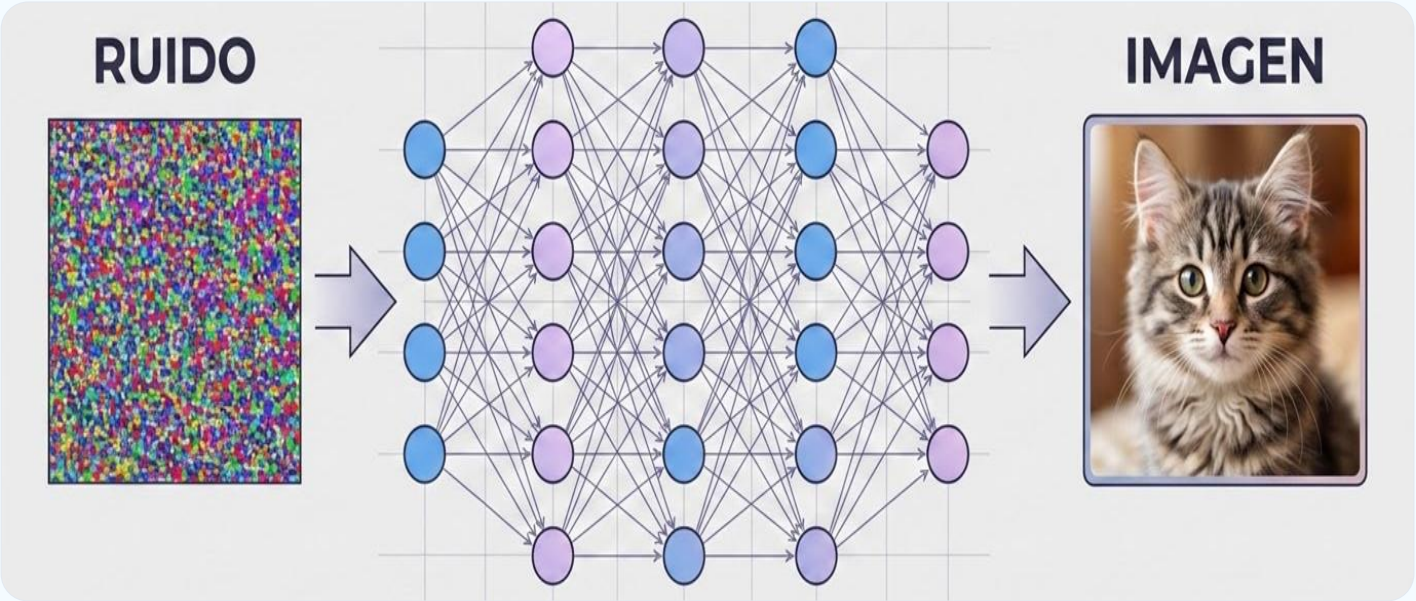
La visión computacional clásica se ha enfocado a extraer significado, etiquetas o vectores a partir de píxeles previamente estructurados por la realidad.

El Paradigma Generativo

¿Es posible abordar la generación de una imagen a partir de datos no estructurados?

La respuesta es un rotundo sí. Los modelos generativos rompen este esquema tradicional.

A partir de fuentes de información caóticas o no estructuradas—como el ruido aleatorio o vectores abstractos de un espacio latente—, la IA es capaz de guiar la creación hasta revelar estructuras y patrones coherentes.



Fuente: GenAI

El Proceso de Muestreo

A partir de una entrada no estructurada como el ruido latente, el modelo de IA aplica transformaciones matemáticas sucesivas para revelar estructuras y patrones coherentes.

Del Caos a la Creación (+Prompts)

La red neuronal actúa como un filtro que decodifica y proyecta los datos del espacio latente hasta sintetizar una imagen realista y detallada en el espacio de píxeles.

DIFUSIÓN: ORDEN DESDE EL CAOS (2020)

Física en lugar de competencia (2015)

Los modelos de difusión son el estado del arte en generación de imágenes. No obstante, antes aparecieron las GANs y los VAEs.

En lugar de la confrontación directa de las GAN, la difusión, inicialmente basada en las VAEs, se basa en principios físicos: destruimos progresivamente la información transformándola en **ruido blanco** y luego entrenamos a la red para revertir el proceso paso a paso.

Proceso hacia delante

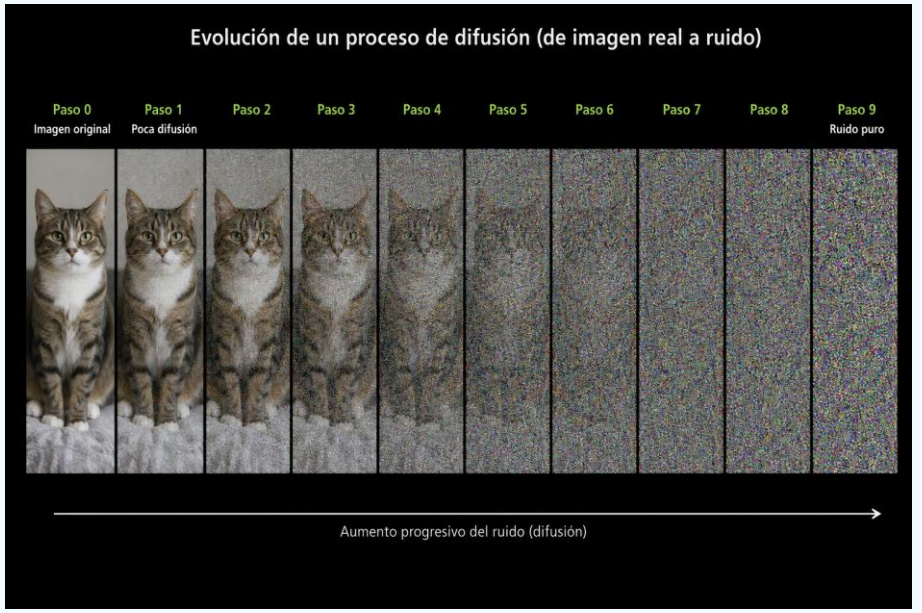
Se añade ruido gaussiano de forma progresiva a la imagen original hasta que se convierte en caos puro.

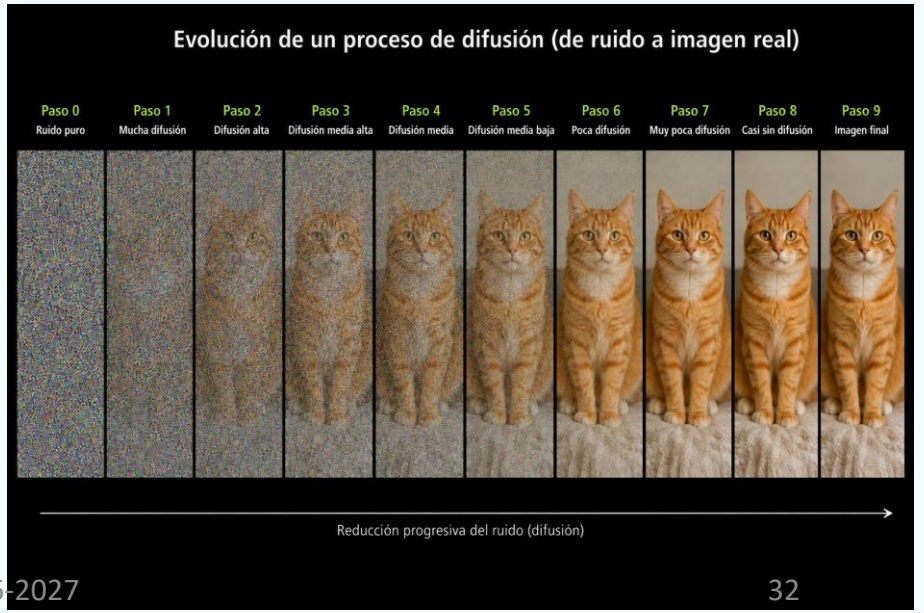
Proceso inverso

La red neuronal aprende a predecir y eliminar el ruido en cada paso, reconstruyendo una imagen nítida.

Recuerda la base de datos de Aprendizaje

Fuente: GenAI





CONDICIONAMIENTO POR PROMPT: GUÍA DE TEXTO Y CONTROL SEMÁNTICO EN EL PROCESO GENERATIVO

«Para una gran clase de casos —aunque no para todos— en que empleamos la palabra "significado", esta puede definirse así: **el significado de una palabra es su uso en el lenguaje**». (L. Wittgenstein, Investigaciones filosóficas)

« **Conocerás a una palabra por la compañía que mantiene**». (J.R. Firth, A Synopsis of Linguistic Theory 1930-1955



1. Tokenización

El prompt escrito en lenguaje natural se fragmenta en pequeñas unidades llamadas tokens semánticos.



2. Codificador (CLIP/T5)

Un modelo de lenguaje procesa las secuencias de tokens para capturar relaciones y contexto semántico.

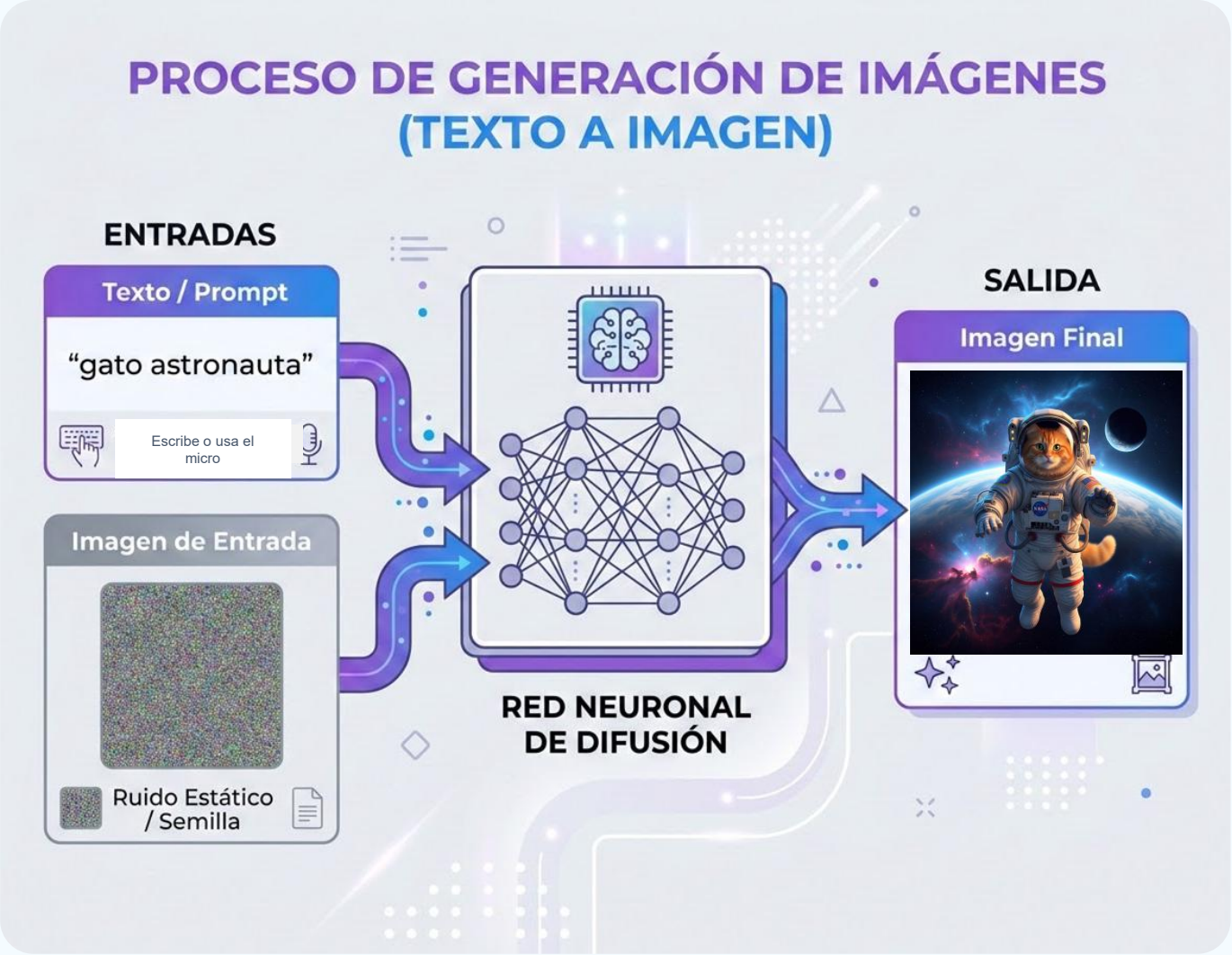


3. Text Embeddings

Se generan vectores que trasladan el texto a un espacio multimodal compartido con imágenes.

Los modelos de lenguaje actuales, Large Language Models (LLMs), tienen millones de millones de parámetros a estimar en base a ejemplos. Se entrenan con más de 500.000 GPUs.

¿CÓMO GENERAMOS UNA IMAGEN CON UN PROMPT?

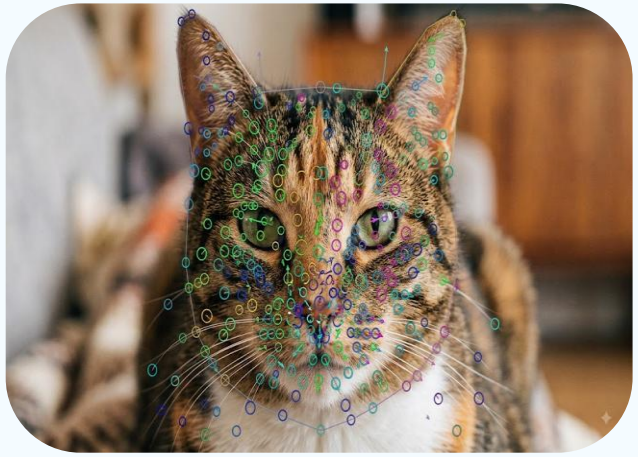


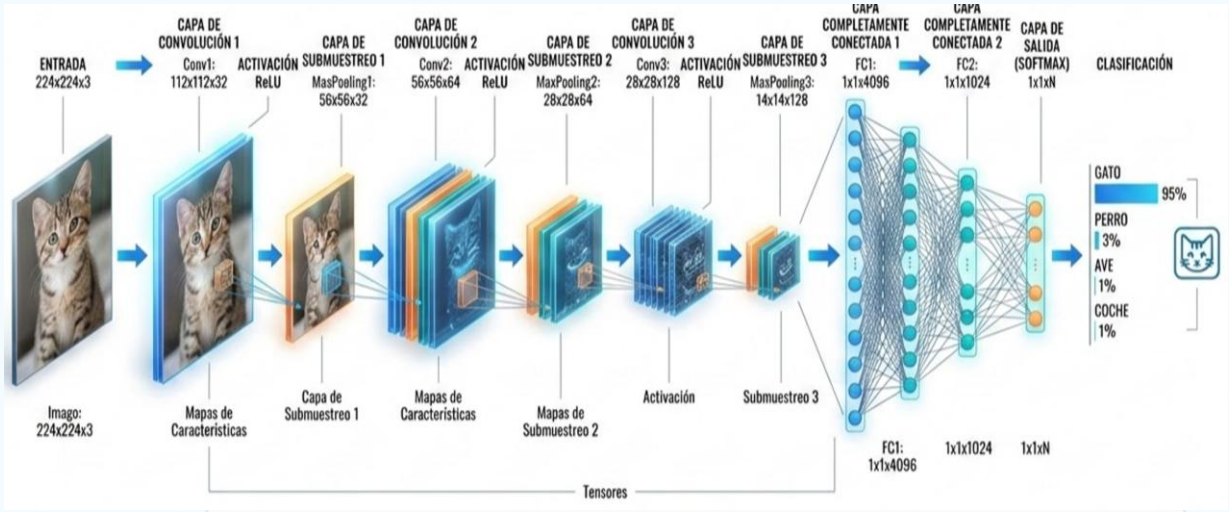
Para Terminar

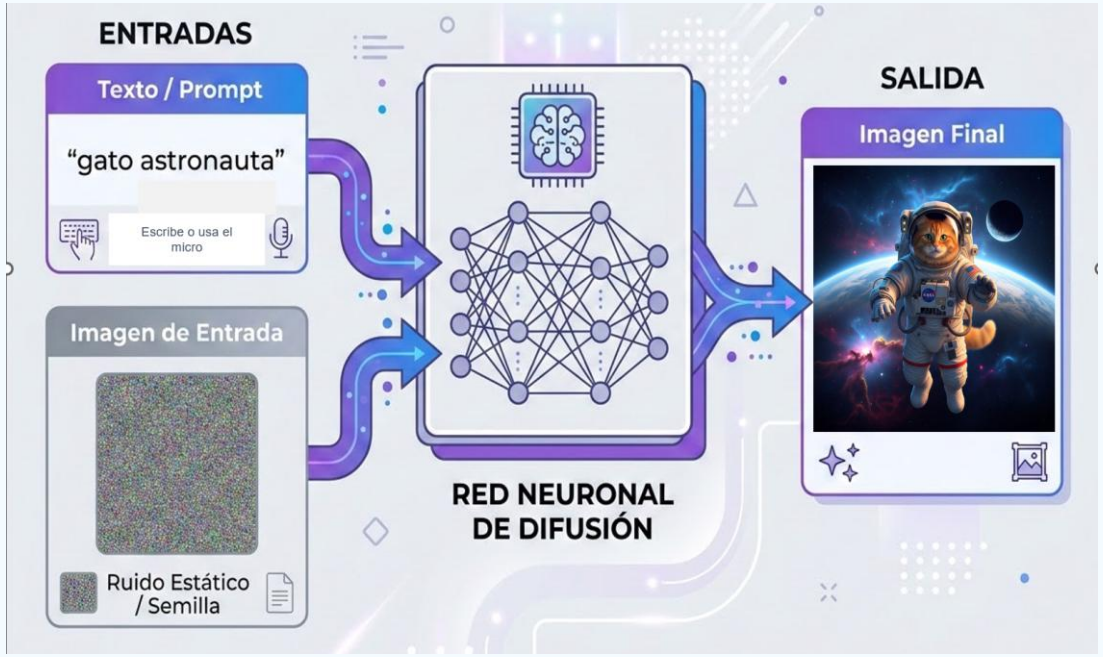


RESUMEN









Hay muchos temas importantes sobre los que no hemos hablado, por ejemplo, la IA agéntica. Tampoco hemos hablado de temas como ...

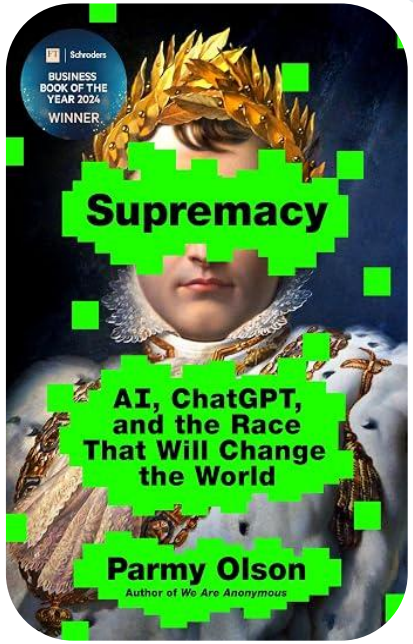
QUIÉN Y CÓMO CONTROLAR



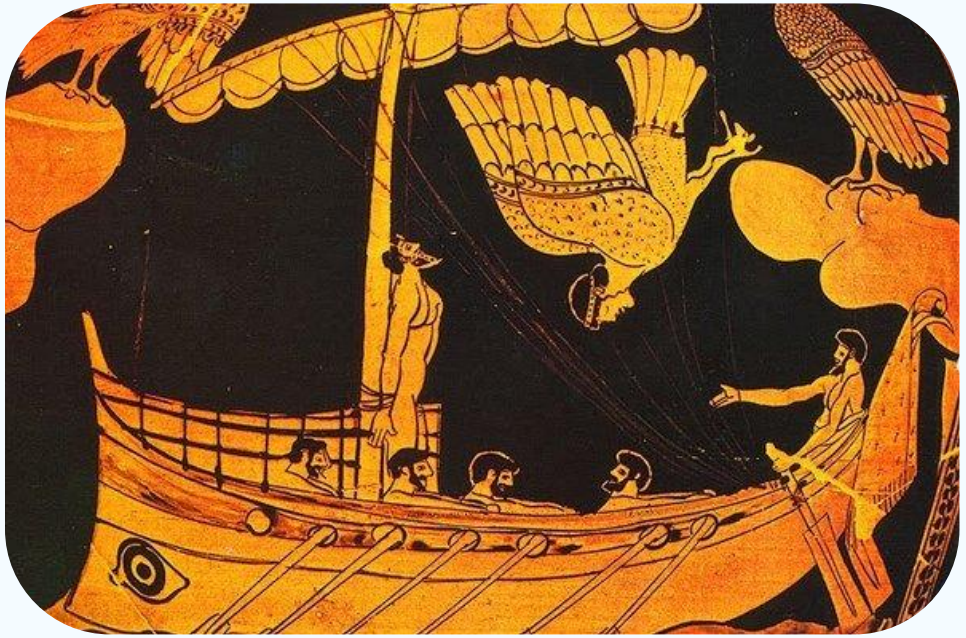


LOS COSTES DE LA IA

¿UNA CARRERA POR EL PODER?



Hay muchos más ejes que explorar: alineación, sesgo, singularidad, …



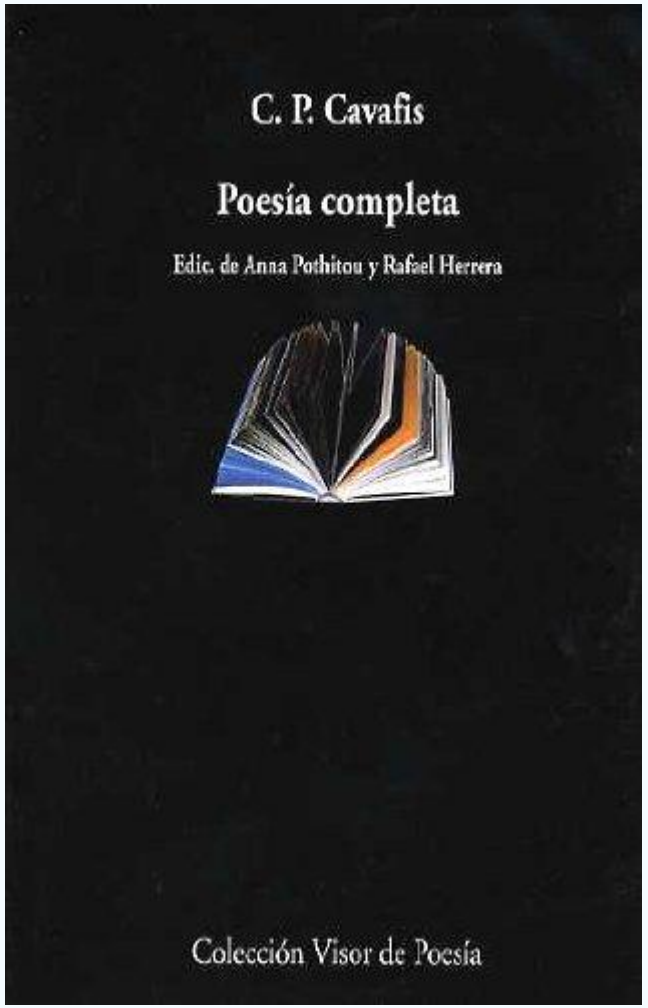
Museo Británico: Las Sirenas

¿Dónde estamos realmente?



Briton Rivère (1873): Ulises y Argos (en Ítaca)

Muchas gracias por vuestra atención



Una última sugerencia: leed Ítaca