Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación

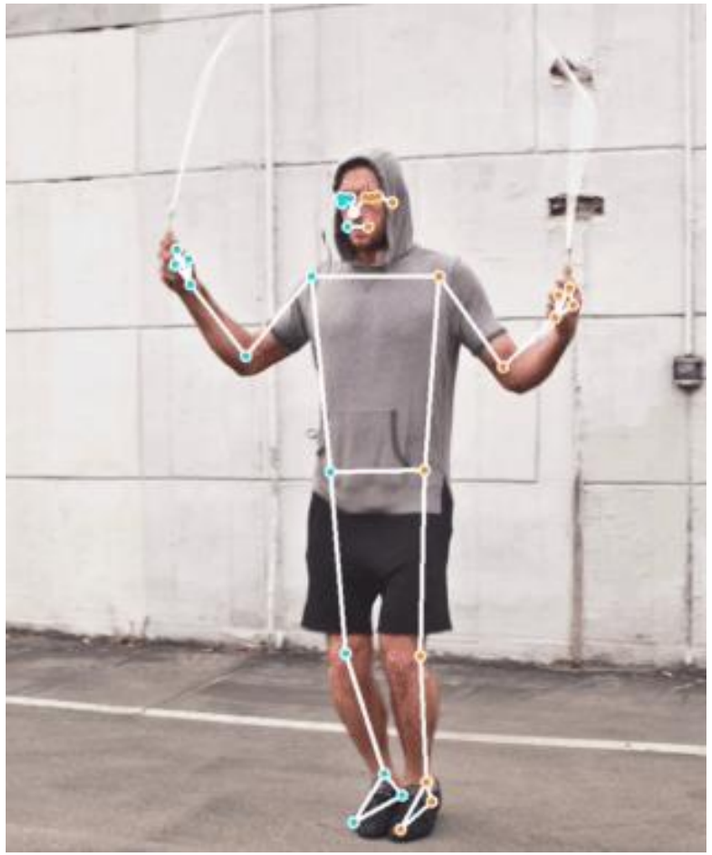


IIC2613 - Inteligencia Artificial

Introducción a Deep Learning

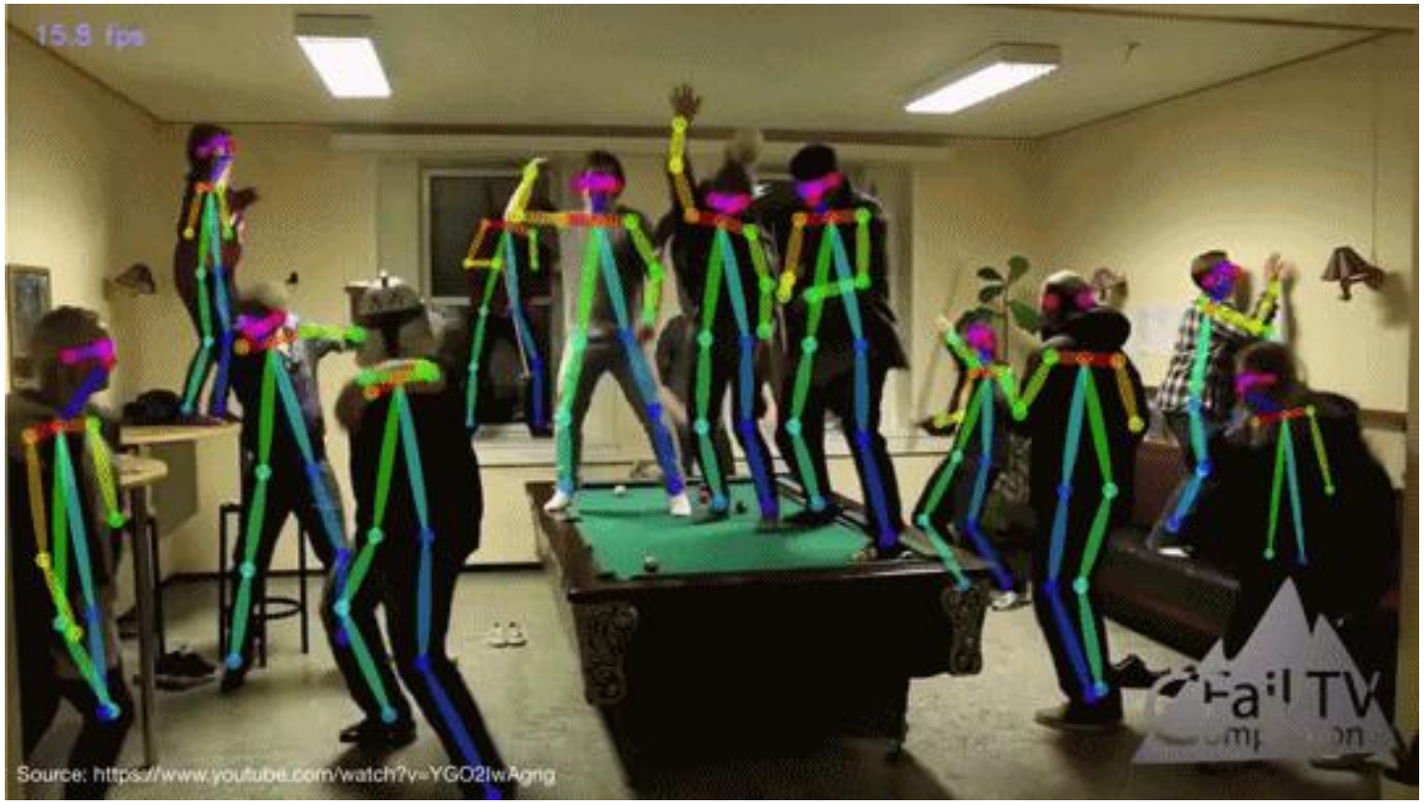
Hans Löbel
Dpto. Ingeniería de Transporte y Logística
Dpto. Ciencia de la Computación

Consideremos un problema altamente complejo y poco estructurado: estimación de pose humana



No tengo idea siquiera de cómo empezar a enfrentar este problema con lo que hemos visto de ML.

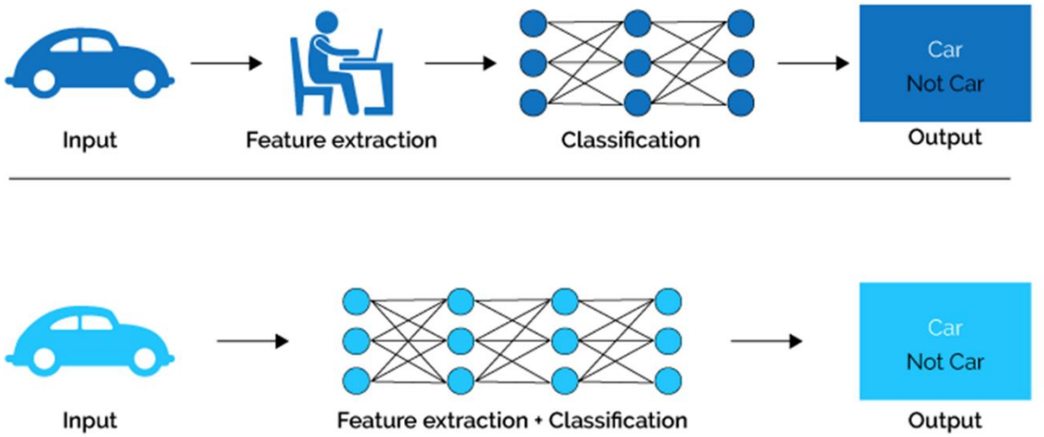
Consideremos un problema altamente complejo y poco estructurado: estimación de pose humana



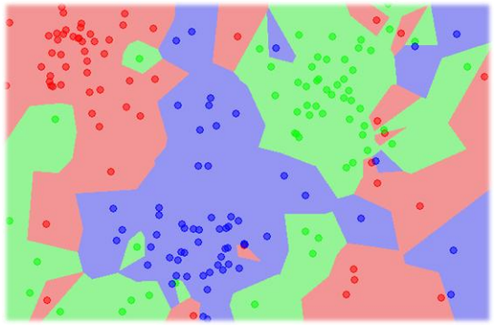
Asumiendo que tengo datos rotulados, qué features utilizo, qué modelo entreno,…

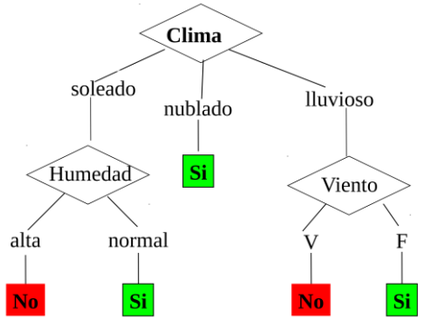
La clave para esto viene de pensar el problema de las features y el modelo como uno conjunto

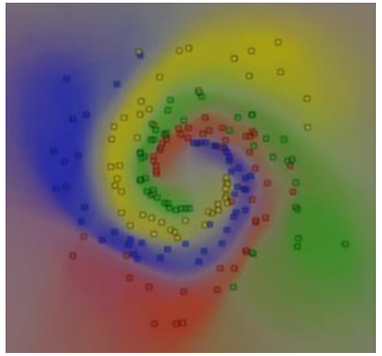
- Hasta ahora, hemos estado usando características (*features*) de los datos para clasificar/predecir.
- Estas tienen que ver más con el conocimiento experto que tengamos (o no) del problema.
- Esto no es ideal por tres motivos:
  i. Un experto es caro y nada asegura que su conocimiento sea óptimo
  ii. Para datos no estructurados, no son evidentes las mejores características y/o especificaciones de modelo
  iii. No es siempre evidente como combinar distintos tipos de datos para extraer información, ni las relaciones que hay entre ellos

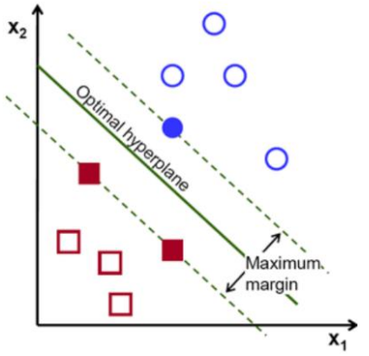


¿Qué es lo que realmente están aprendiendo estos modelos?









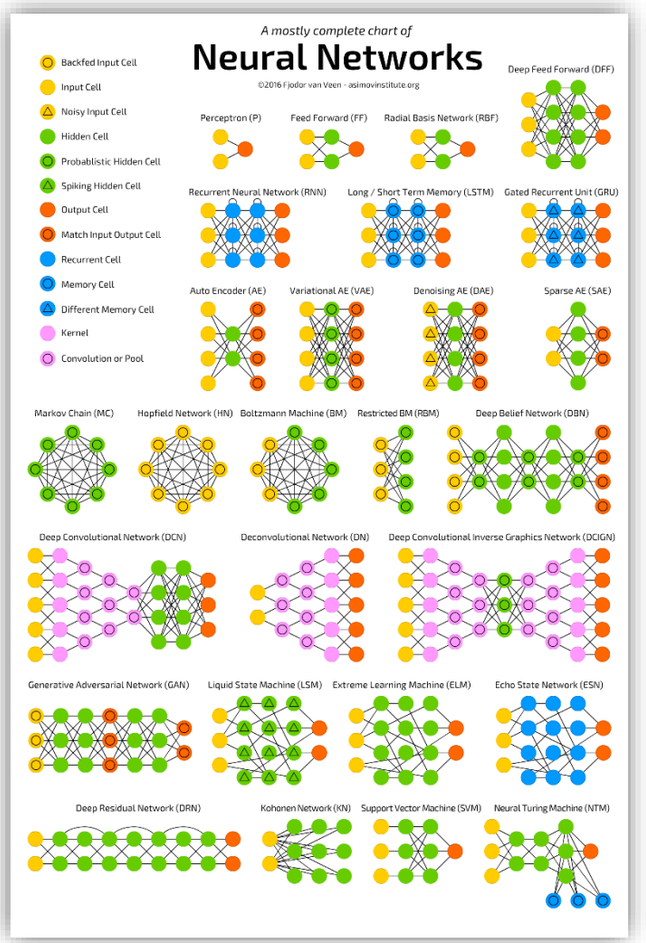
Hasta ahora no estamos “realmente” aprendiendo de los datos, solo como categorizarlos en base a características hechas a manos

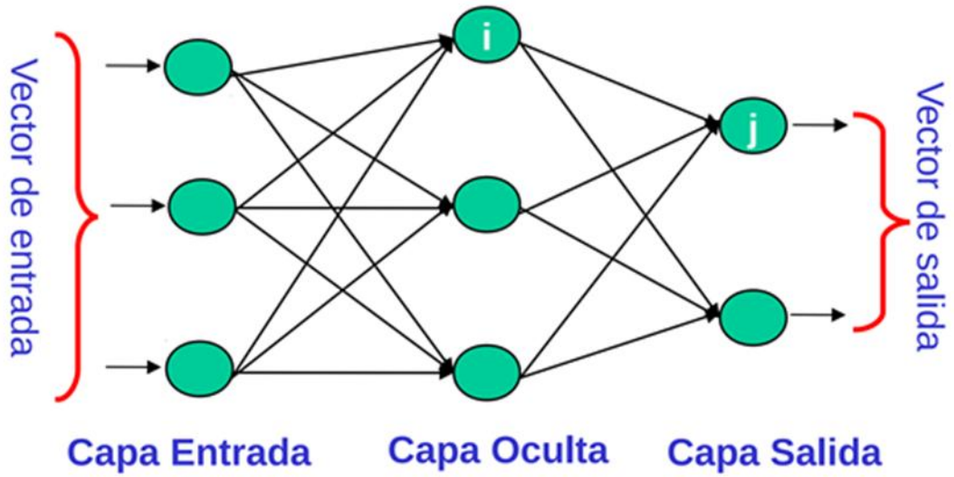
Buscamos un mecanismo de aprendizaje suficientemente poderoso y general, que permita aprender features y modelos predictivos que capturen aspectos complejos y eventualmente desconocidos de los datos, con el fin de aprender a realizar tareas de manera óptima

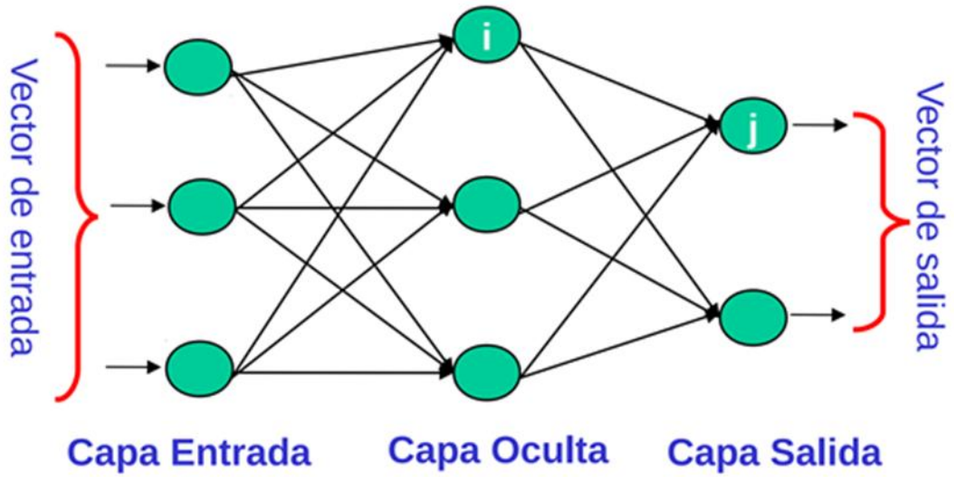


Redes neuronales son modelos altamente prácticos y generales para aprender funciones continuas y discretas

- Brillan en la presencia de grandes volúmenes de datos.
- Son capaces de aprender transformaciones de los datos (*features* o *embeddings*), que luego pueden ser usadas como *features* en otros problemas.
- Existen redes para prácticamente cualquier problema.
- Difíciles de entrenar, interpretabilidad es compleja (caja negra) y sufren de serios problemas de sobreajuste (esto último ya no es tan claro).

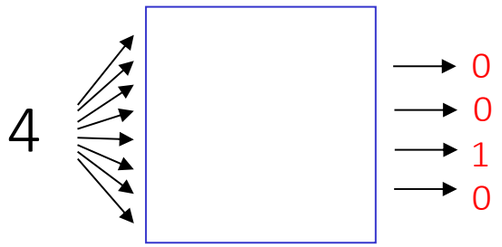


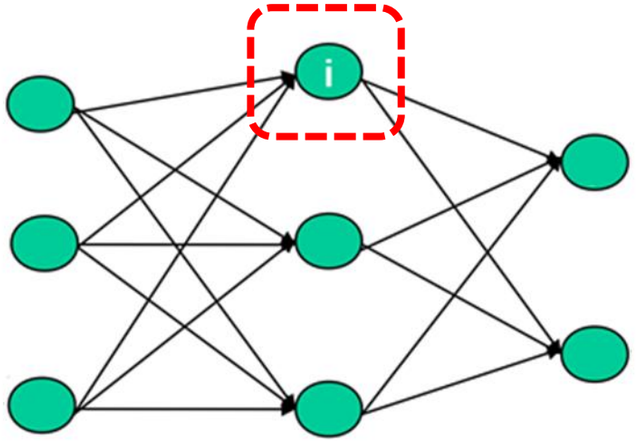

Vector de entrada
i
j
Vector de salida
Capa Entrada
Capa Oculta
Capa Salida



Dada la estructura de la red, ¿cómo podemos ajustar los parámetros de las unidades (nodos) para que al ingresar un vector de entrada obtengamos la salida deseada?

Ejemplo: cuando la red recibe una imagen de un 4, su codificación de salida debe indicar un 4.




i



1958 Perceptron





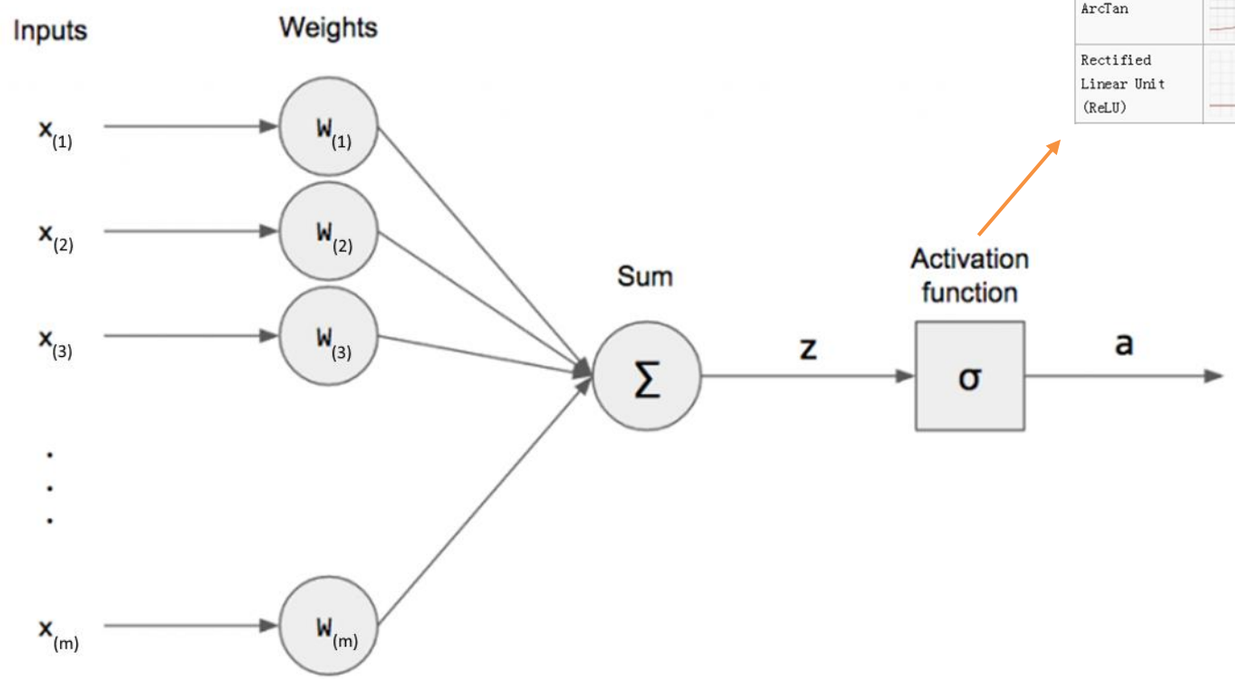


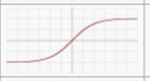
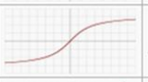


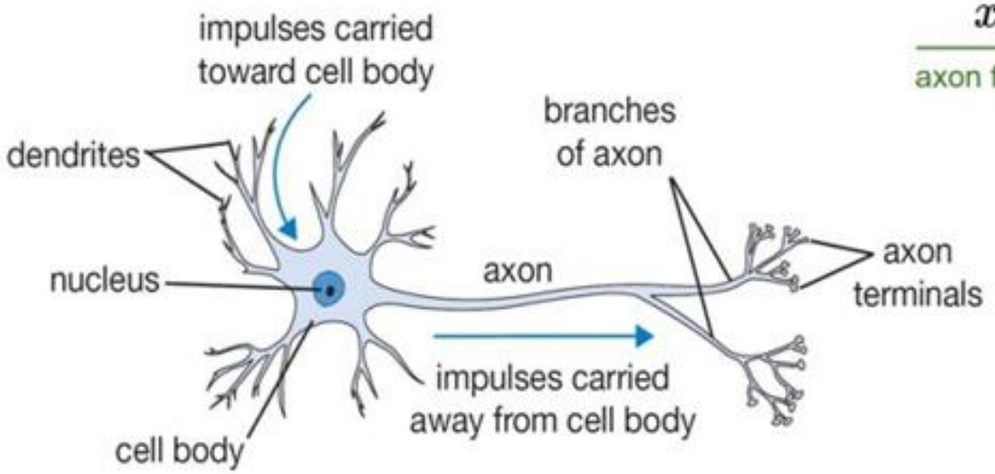


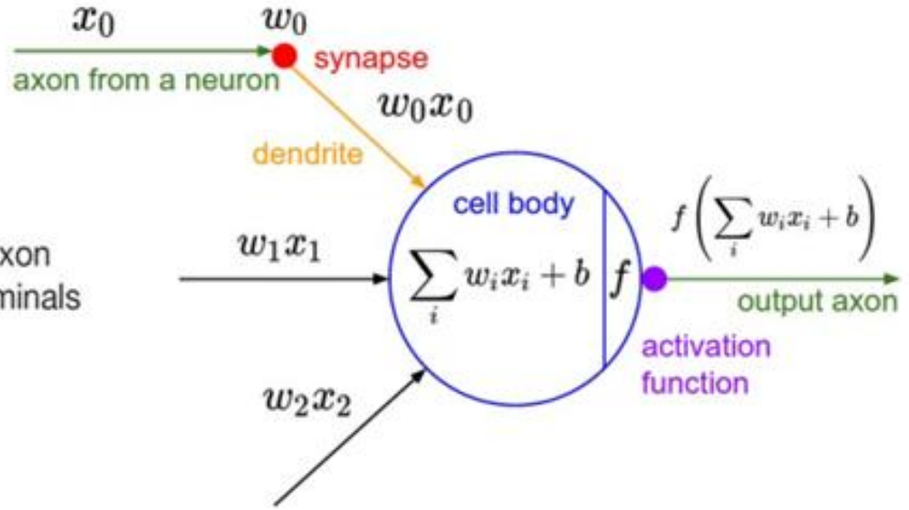


El Perceptron puede considerarse como el precursor de las redes neuronales



Name	Plot	Equation
Identity		$f(x) = x$
Logistic (a.k.a Soft step)		$f(x) = \frac{1}{1+e^{-x}}$
TanH		$f(x) = \tanh(x) = \frac{2}{1+e^{-2x}} - 1$
ArcTan		$f(x) = \tan^{-1}(x)$
Rectified Linear Unit (ReLU)		$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$


impulses carried
toward cell body
dendrites
nucleus
cell body
axon
branches
of axon
axon
terminals
impulses carried
away from cell body

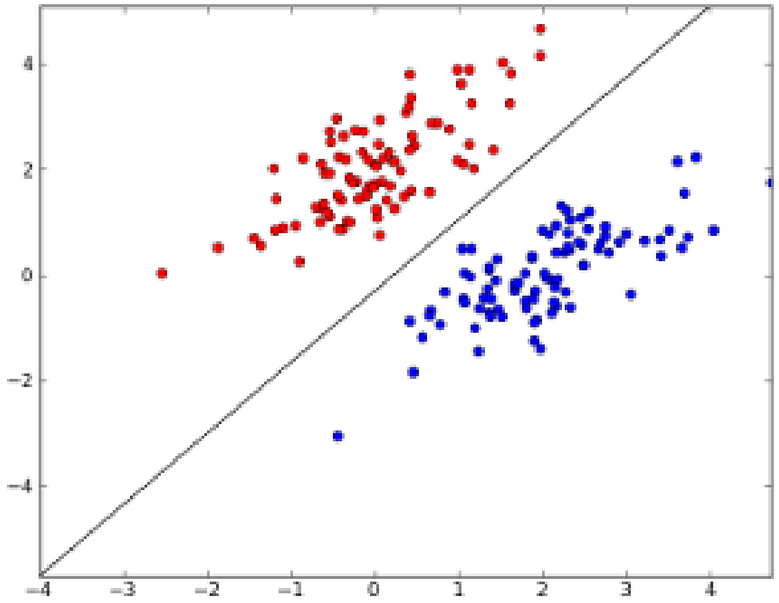

x_0
w_0
axon from a neuron
synapse
$w_0 x_0$
dendrite
cell body
$w_1 x_1$
$\sum_i w_i x_i + b$
f
$f\left(\sum_i w_i x_i + b\right)$
output axon
activation
function
$w_2 x_2$

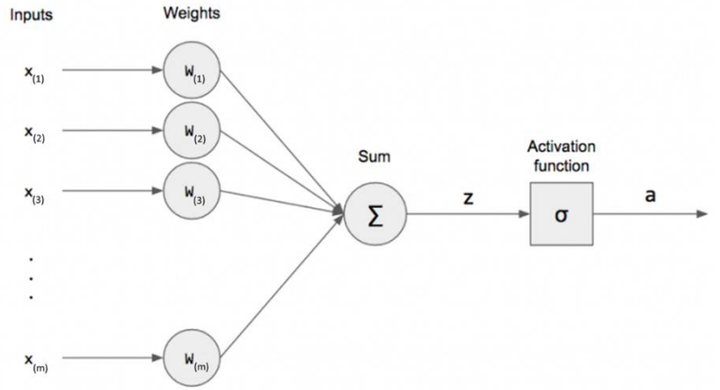
Un Perceptron permite estimar funciones de manera supervisada

- Si tenemos suficientes datos, pares (x_i, y_i), es posible construir una función que nos indique cuán buena es en promedio la estimación del Perceptron.
- Sea $f(x; w) = \sigma(\sum_{i=1}^{m} w_{(i)} x_{(i)}) = \sigma(w \cdot x)$, una función que modela la aplicación de un Perceptrón parametrizado por un vector de pesos w, a un vector de características x.
- Sin suponer cosas raras sobre los datos, una posible función de pérdida (o costo, o error) puede quedar de la siguiente manera:

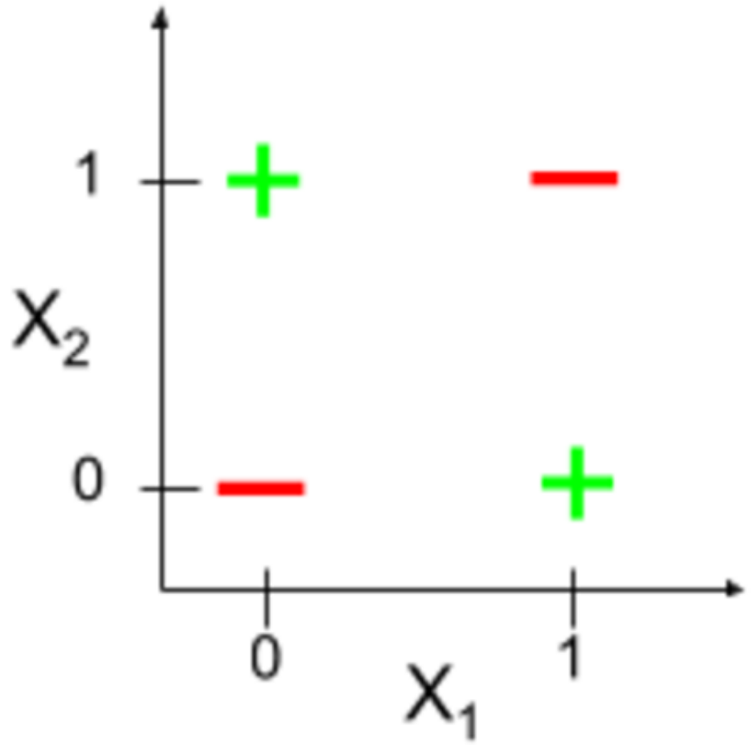
$$J(w) = \frac{1}{2n} \sum_{i=1}^{n} (f(x_i; w) - y_i)^2$$

- ¿Qué representa el valor mínimo de esta función, con respecto a los pesos w?

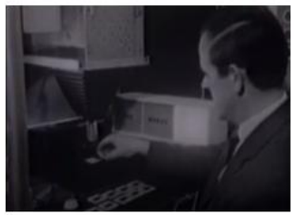




Perceptron puede resolver problemas lineales (clasificación y regresión)



¿Cómo podríamos extender al Perceptron, para que sea capaz de resolver este tipo de problemas (no linealmente separables)?



1958 Perceptron







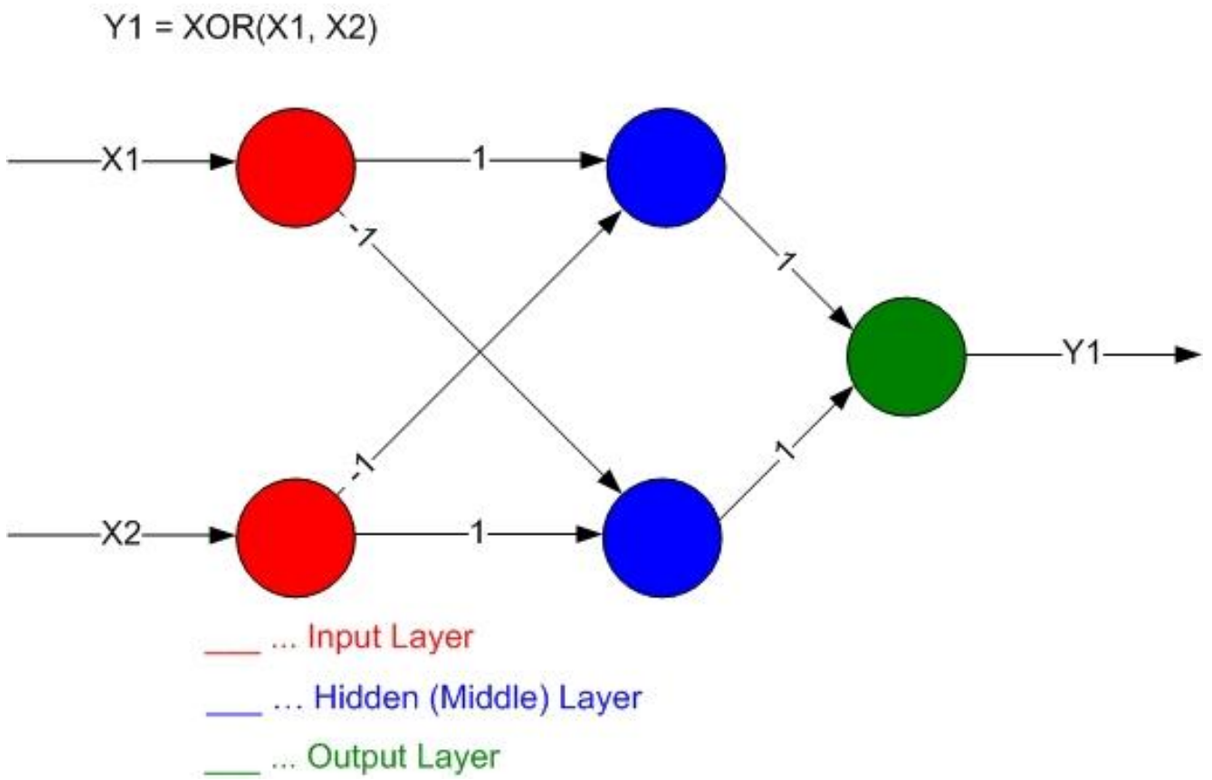
1969
Perceptron criticized







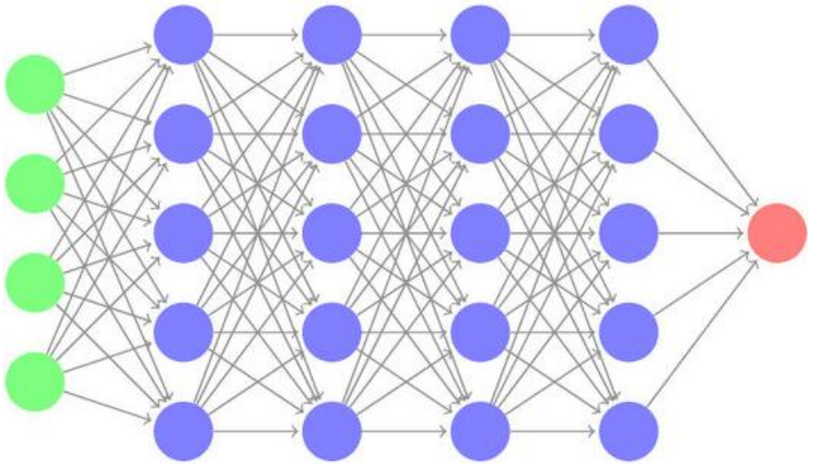




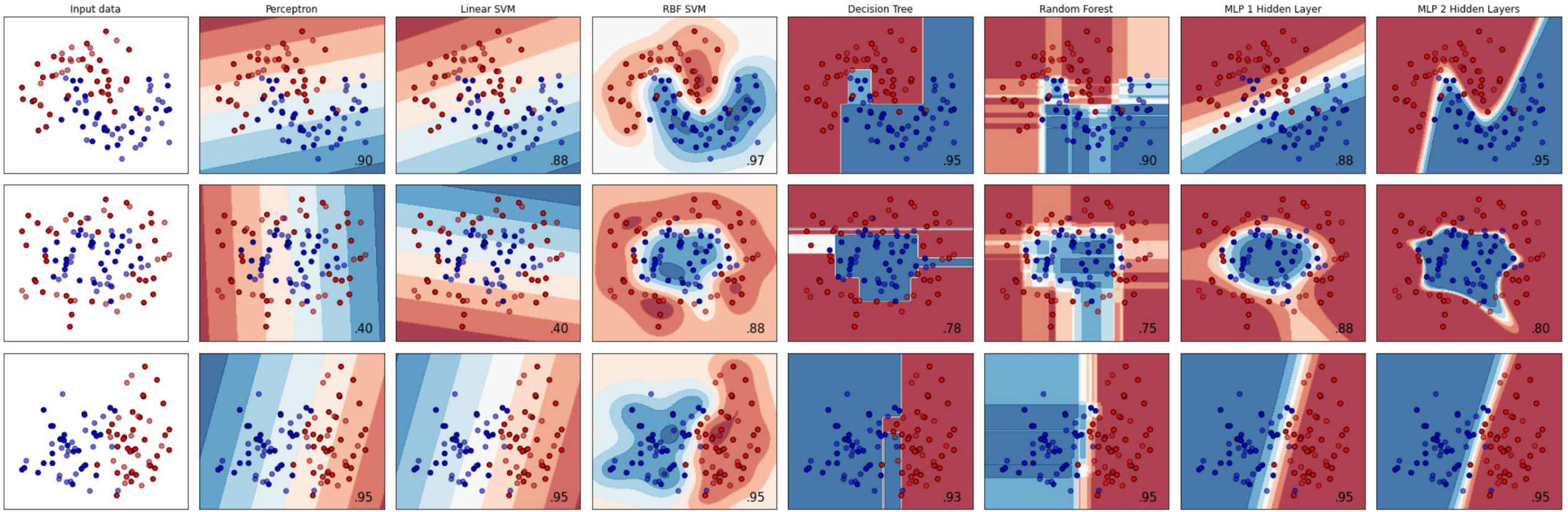
Al agregar una nueva capa (oculta), podemos representar features intermedias entre entrada y salida, que permiten separar más fácilmente los datos

MLP generaliza el concepto de Perceptron

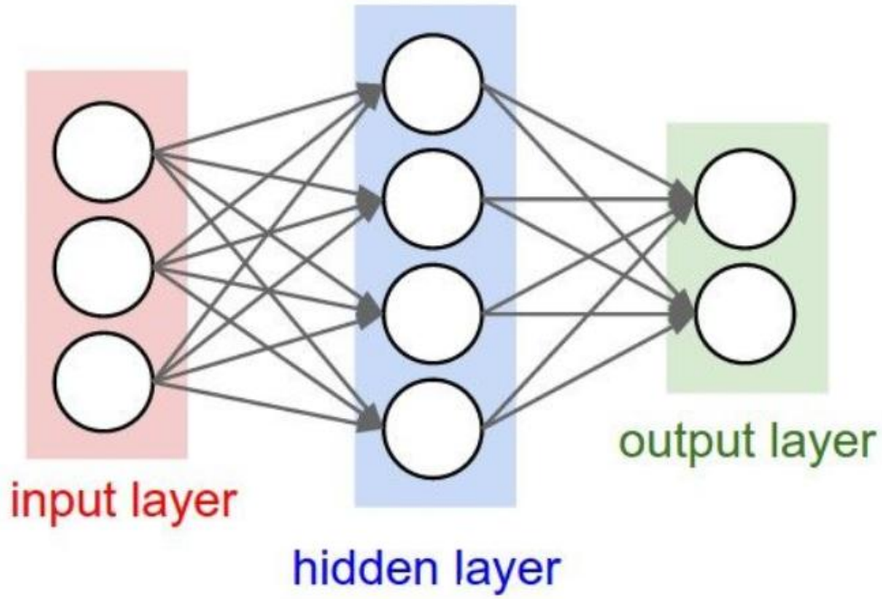
- Un Multilayer Perceptron (MLP), combina múltiples perceptrones en múltiples capas. Es conocido también como Deep Feed Forward Network (DFF).
- Permite resolver problemas más complejos que los que puede resolver un Perceptron.

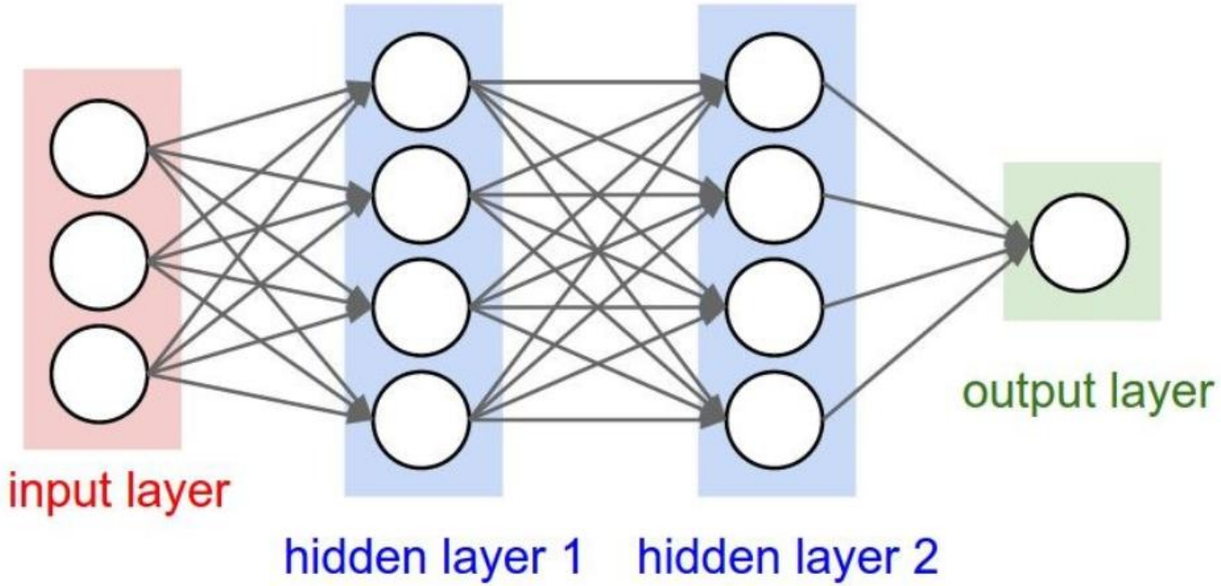


Nuevas capas aumentan poder discriminativo comparado con un Perceptron



Visualmente, representaremos los MLP utilizando capas



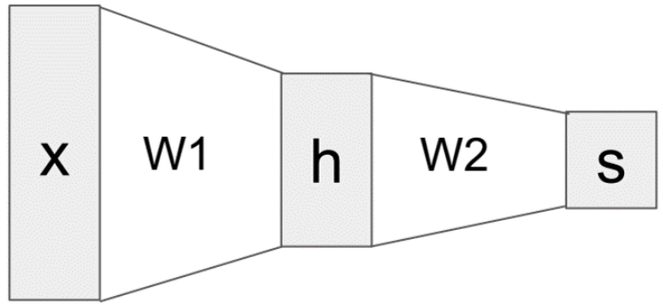


Como veremos más adelante, siempre es bueno pensar en estos modelos como funciones

- Perceptron lineal: $f = W_1 x$

- MLP de 2 capas: $f = W_2 \sigma(W_1 x)$

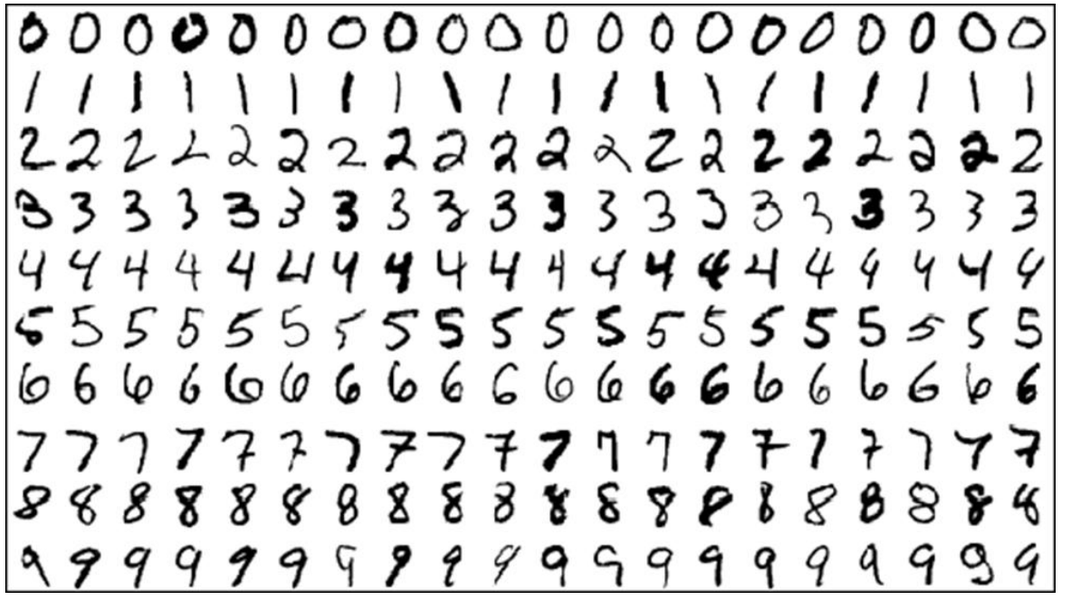
  * $\sigma(z) = \frac{1}{1+e^{-z}}$

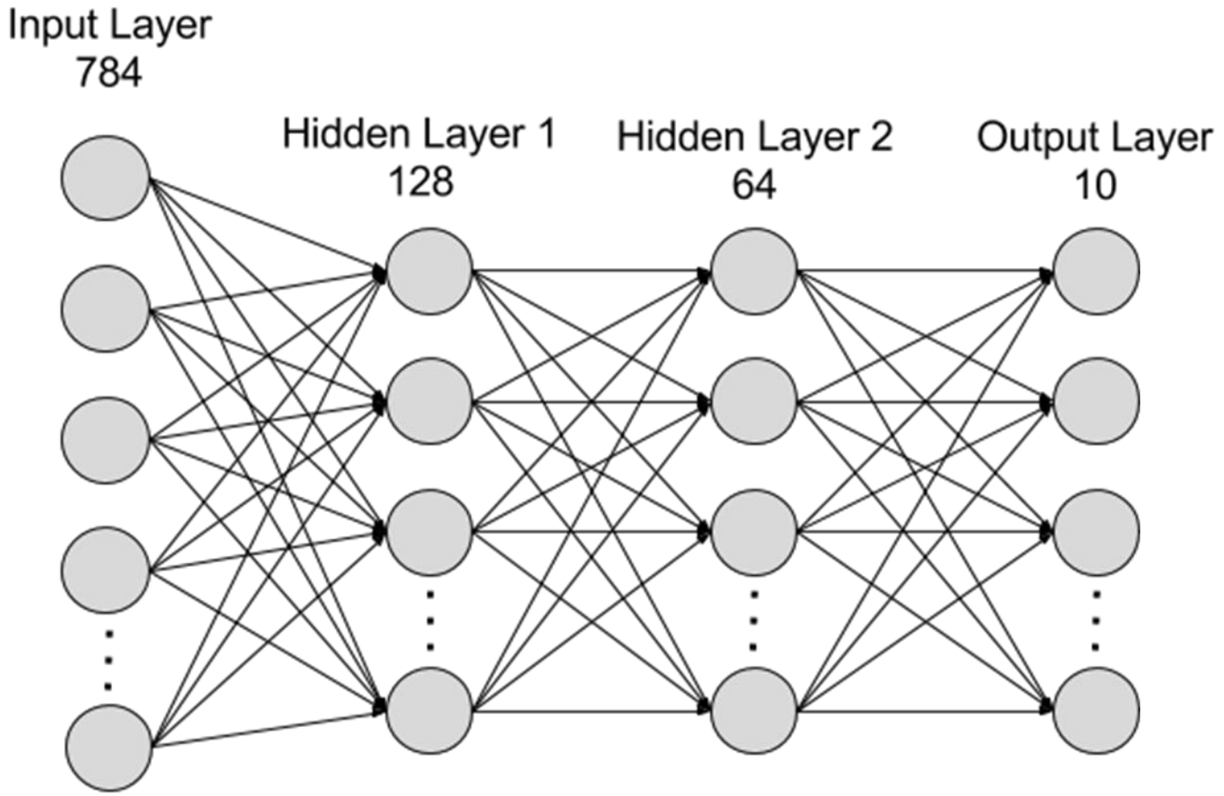


- MLP de 3 capas: $f = W_3 \sigma(W_2 \sigma(W_1 x))$

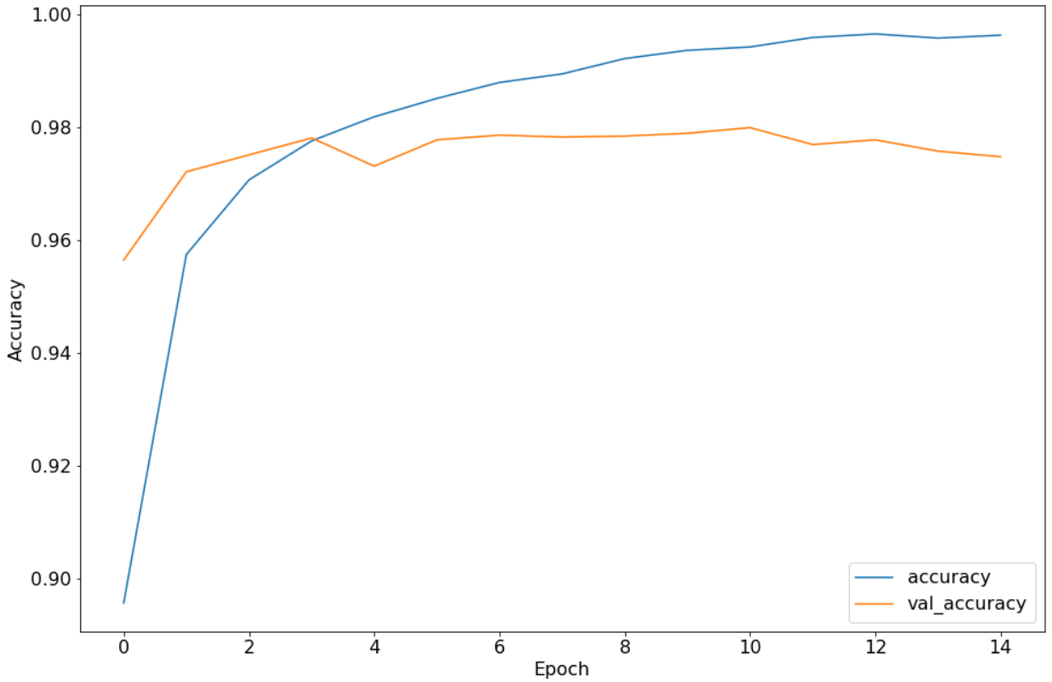
El set de datos MNIST permite construir clasificadores de dígitos a partir de imágenes

- MNIST es un set de datos compuesto de imágenes de dígitos escritos a mano.
- Cada imagen muestra un sólo dígito entre 0 y 9. Las imágenes son binarias con una resolución de 28x28=784 pixeles.
- El dataset consta de 60.000 ejemplos de entrenamiento y 10.000 de test.
- ¿Cómo podría ser la arquitectura de un MLP que para resolver este problema?



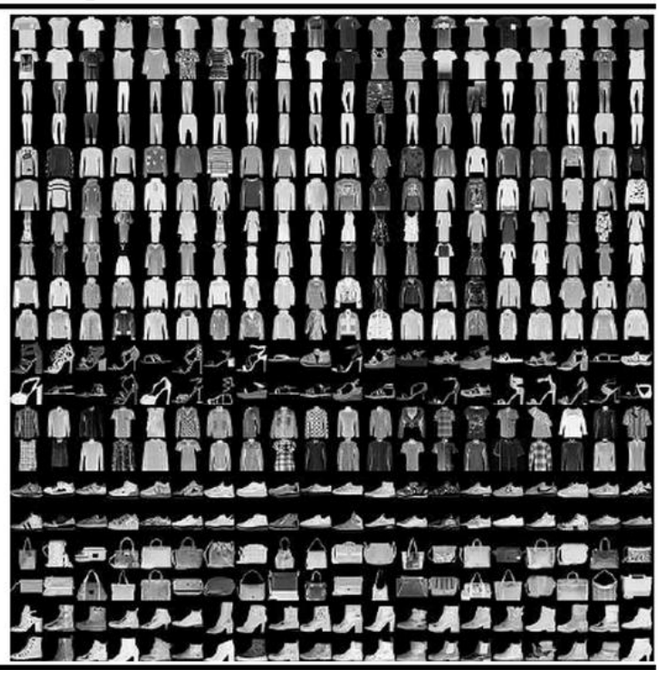

Input Layer
784
Hidden Layer 1
128
Hidden Layer 2
64
Output Layer
10

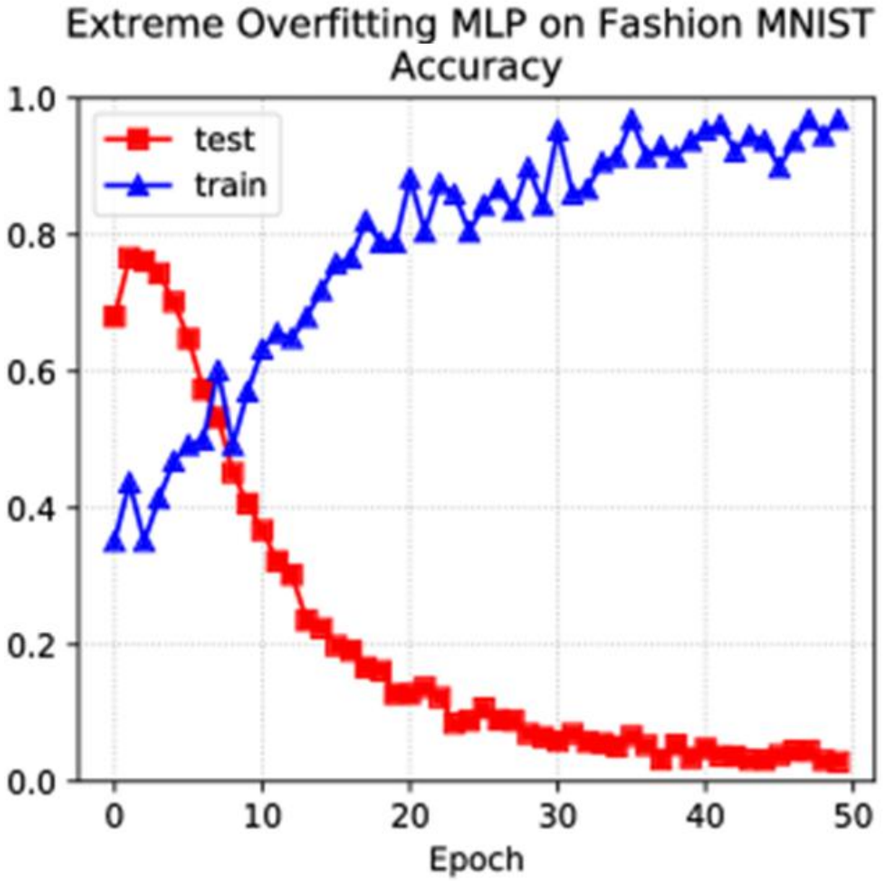
Accuracy


1.00
0.98
0.96
0.94
0.92
0.90
Accuracy
0
2
4
6
8
10
12
14
Epoch
accuracy
val_accuracy

El set de datos Fashion MNIST permite construir clasificadores de ropa a partir de imágenes

- Sucesor de MNIST con mayor dificultad (pero tampoco tanto más).
- Cada imagen muestra una sola prende de una de diez categorías. Las imágenes son binarias con una resolución de 28x28 pixeles.
- El *dataset* consta de 60.000 ejemplos de entrenamiento y 10.000 de test.
- ¿Qué pasa si entrenamos un MLP igual al que entrenamos para MNIST?

Label	Description	Examples
0	T-Shirt/Top	
1	Trouser	
2	Pullover	
3	Dress	
4	Coat	
5	Sandals	
6	Shirt	
7	Sneaker	
8	Bag	
9	Ankle boots	


Extreme Overfitting MLP on Fashion MNIST
Accuracy
test
train
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50
Epoch

Algo está fallando por acá…

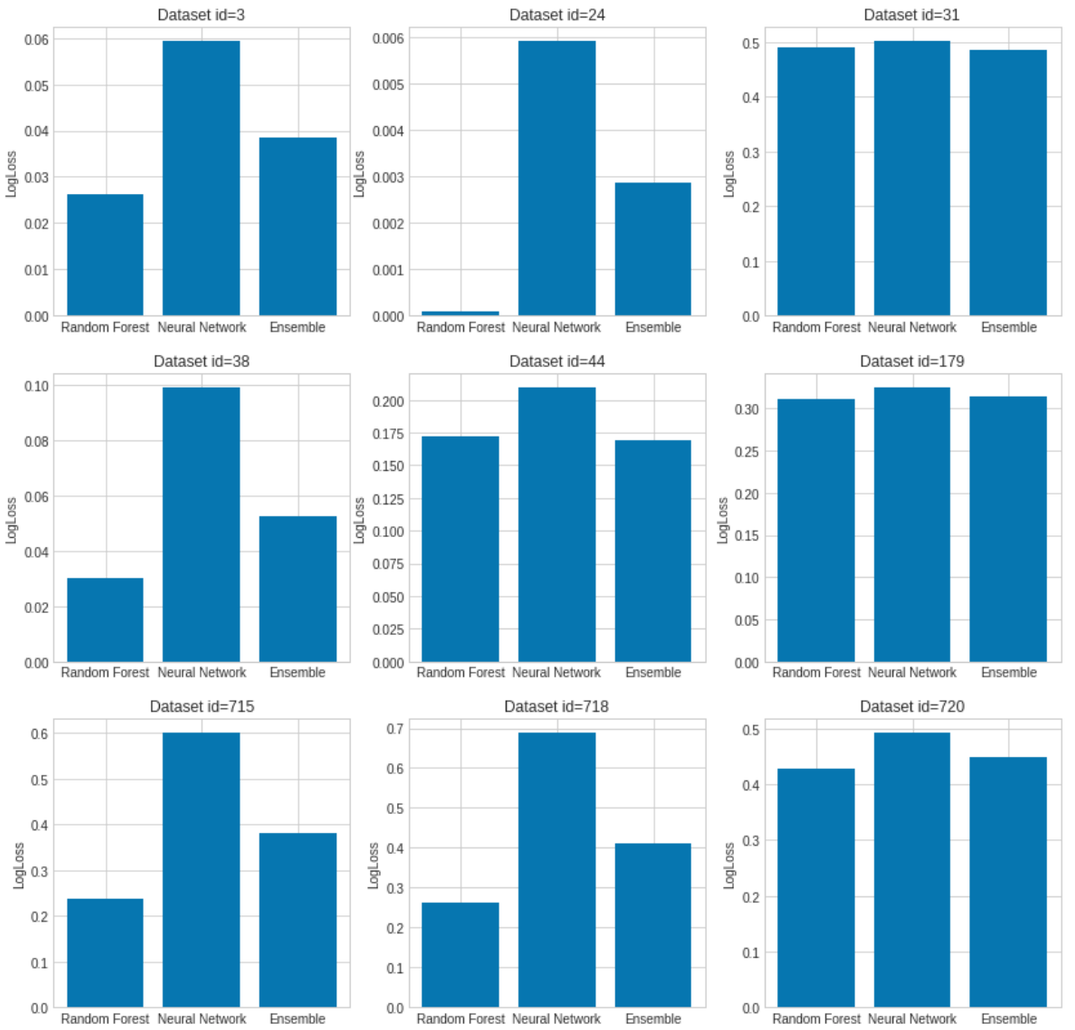
- Quizá los MLP no andan bien con imágenes, y aprenden patrones que no son muy generales (más adelante veremos el porqué).
- Probemos suerte con otro dominio más estructurado…



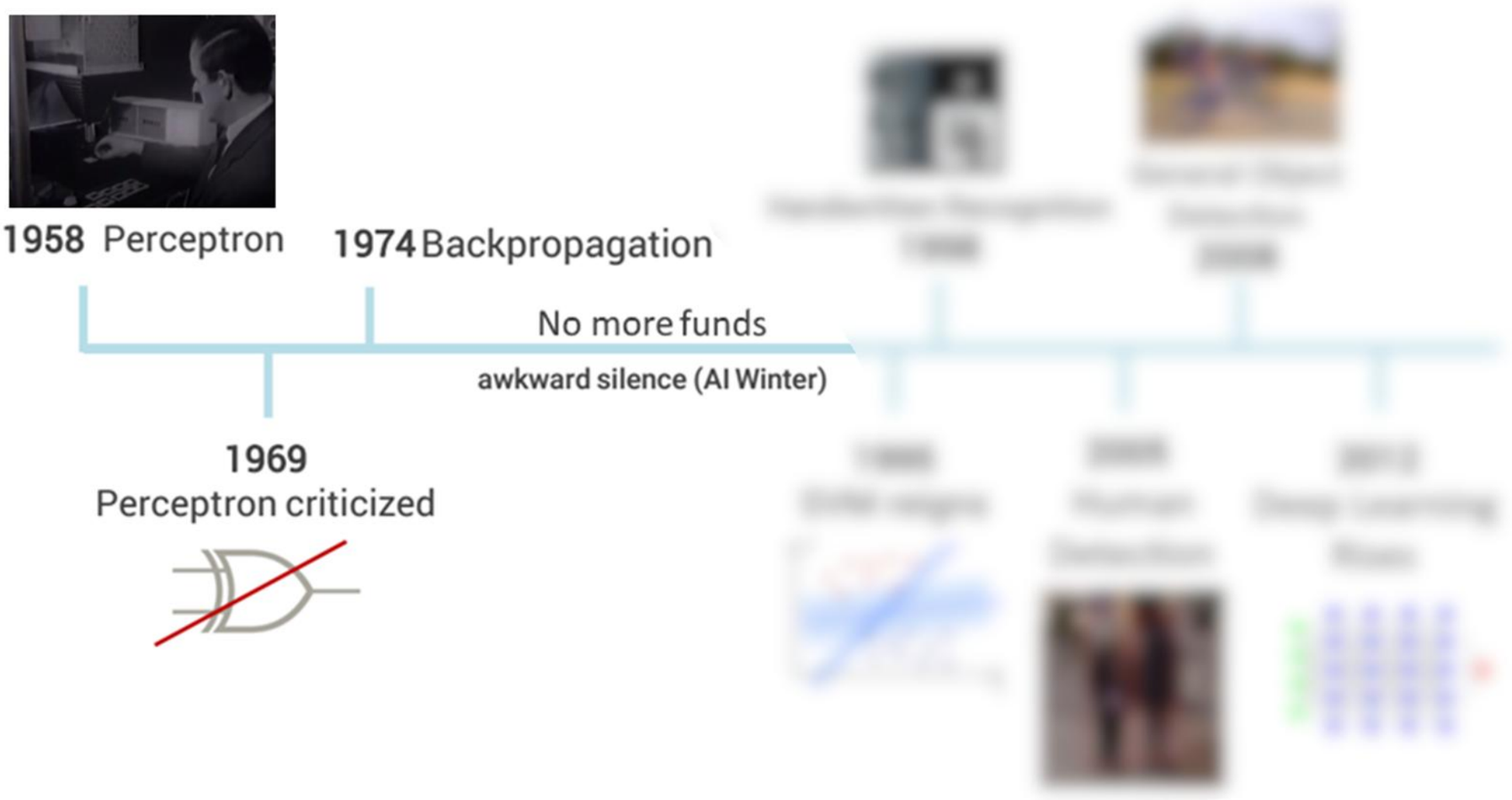
Una aplicación típica de los MLP es en el procesamiento de datos tabulados

- Filas en una tabla pueden verse de manera natural como los vectores de entrada.
- Pero parece que acá tampoco le va muy bien…

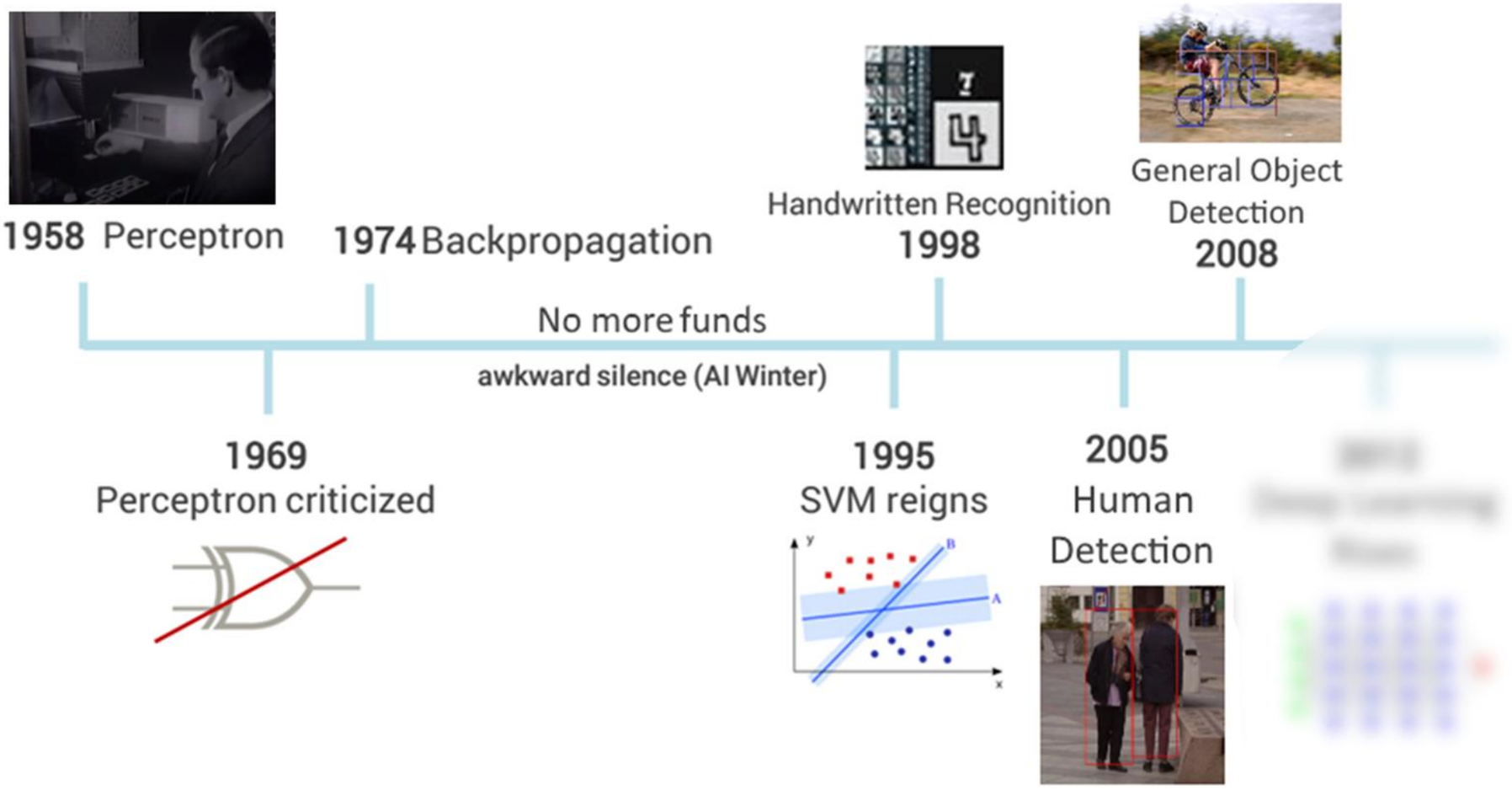
id	name	rows	cols
3	kr-vs-kp	3196	36
24	mushroom	8124	22
31	credit-g	1000	20
38	sick	3772	29
44	spambase	4601	57
179	adult	48842	14
715	fri_c3_1000_25	1000	25
718	fri_c4_1000_100	1000	100
720	abalone	4177	8



Dificultad para obtener buenos resultados generó el “invierno de la IA”, época en la cual hubo escasos avances públicos en el área

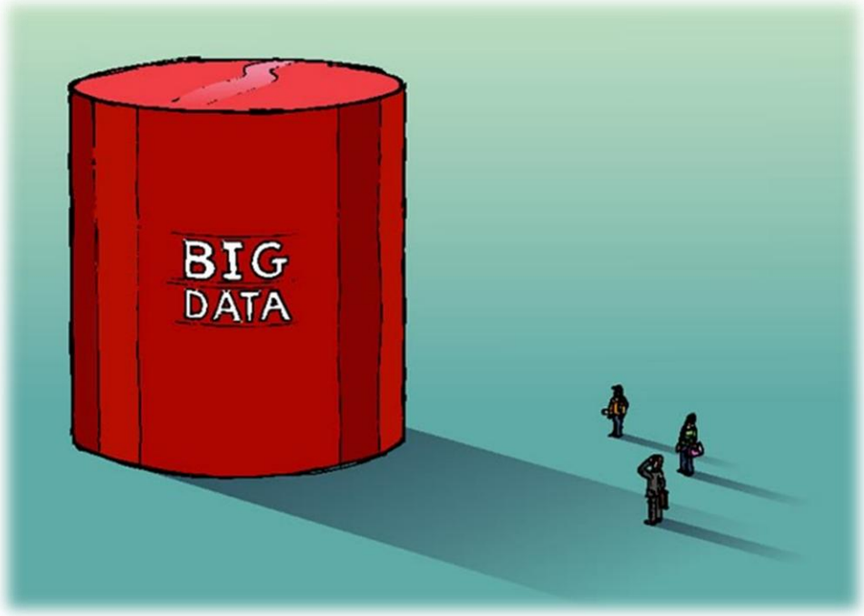


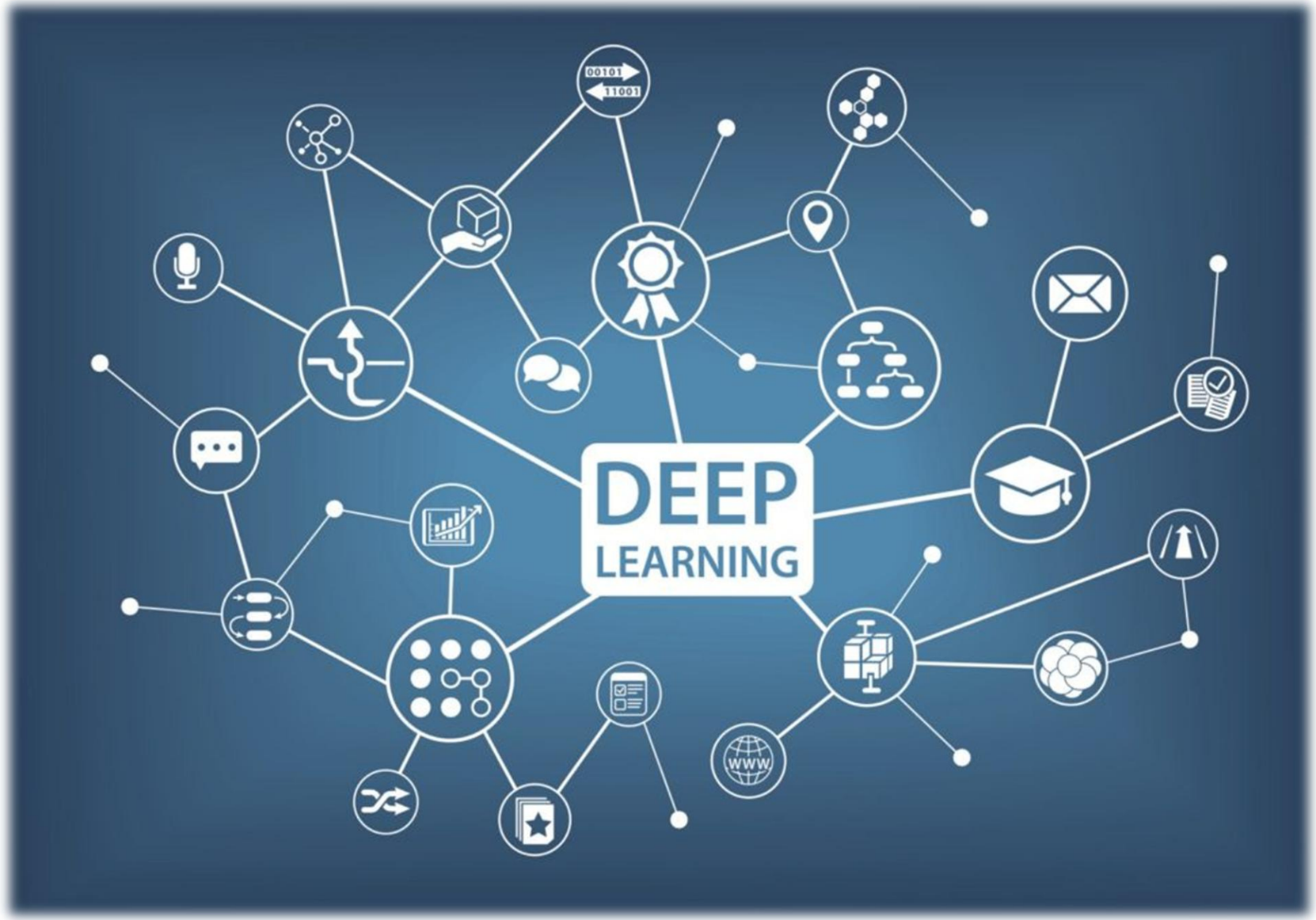
Posteriormente, a partir de mediados de los 90, el mayor volumen de datos disponibles permitió al área de Machine Learning salir del invierno, pero después de unas décadas volvieron los problemas



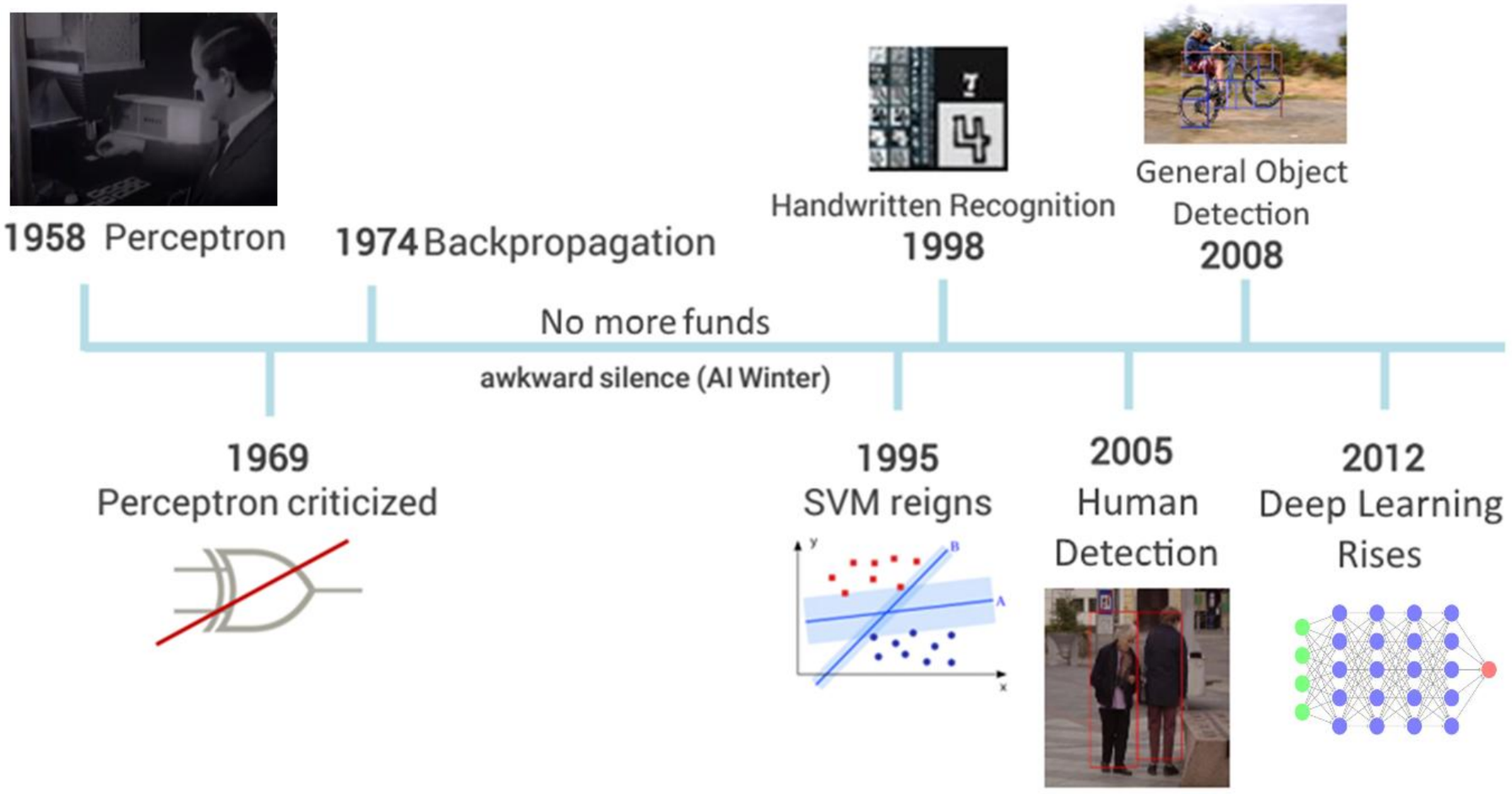
El advenimiento de una revolución

- Cerca de 2010, las técnicas tradicionales de *machine learning* estaban comenzando a mostrar signos de saturación.
- Modelos no mostraban mayores beneficios al usar grandes volúmenes de datos.
- Tampoco estaba claro cómo utilizar distintos tipos de datos simultáneamente.
- La salvación vendría de una técnica vilipendiada en el pasado…



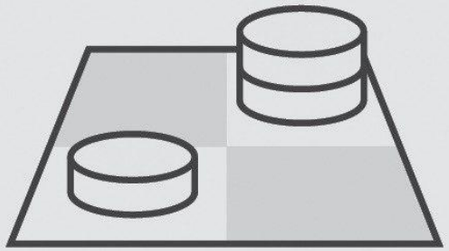

00101
11001
DEEP
LEARNING
WWW

Las redes neuronales salieron de este invierno recién en 2012, pero con un nuevo nombre: Deep Learning



ARTIFICIAL INTELLIGENCE

Artificial Intelligence captures the imagination of the world.



MACHINE LEARNING

Machine learning starts to gain traction.



DEEP LEARNING

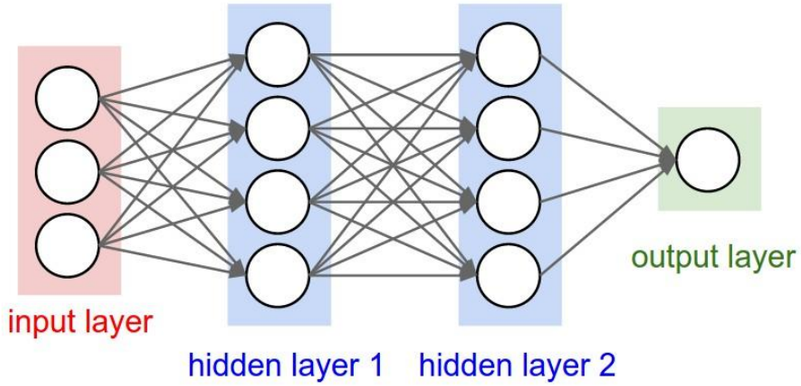
Deep learning catapults the industry.



1950s 1960s 1970s 1980s 1990s 2000s 2010s

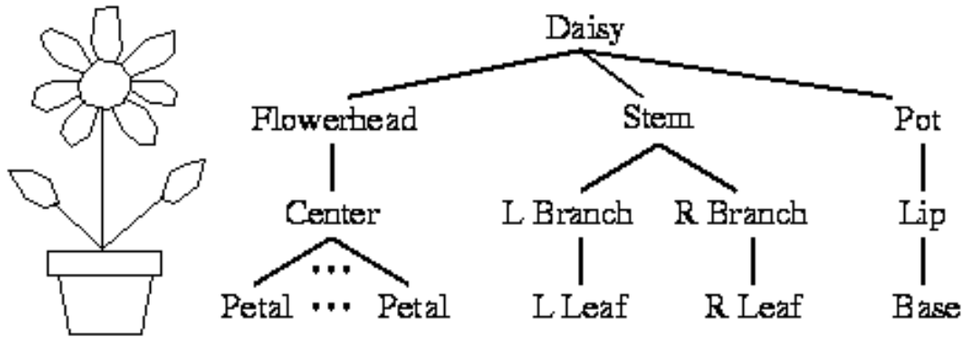
¿Qué es Deep Learning?

- Modelos capaces de aprender capas de representaciones composicionales y jerárquicas, en conjunto con capa/capas para la resolución a la tarea.
- Capas son formadas por unidades diferenciables entrenadas en conjunto en base a una función objetivo común.
- Al menos una capa de representaciones intermedias, pero generalmente son muchas.
- Actualmente es el mecanismo de aprendizaje más exitoso, manteniendo récords en gran cantidad de tareas complejas, como reconocimiento de voz, caracteres, objetos, etc.



¿Qué son las representaciones composicionales y jerárquicas?

- Las representaciones composicionales y jerárquicas son muy comunes en la naturaleza
- Se organizan desde niveles bajos a altos de abstracción.
- La composicionalidad es su principal característica.
- Permiten describir infinitas configuraciones a partir de un número acotado de elementos.
- Algunos ejemplos:
    - Sonidos -> fonemas -> sílabas -> palabras
    - Pixeles -> bordes -> partes -> objetos -> escenas
    - Caracteres -> palabras -> grupos -> clausulas -> frases



¿Qué se aprende en los distintos niveles?

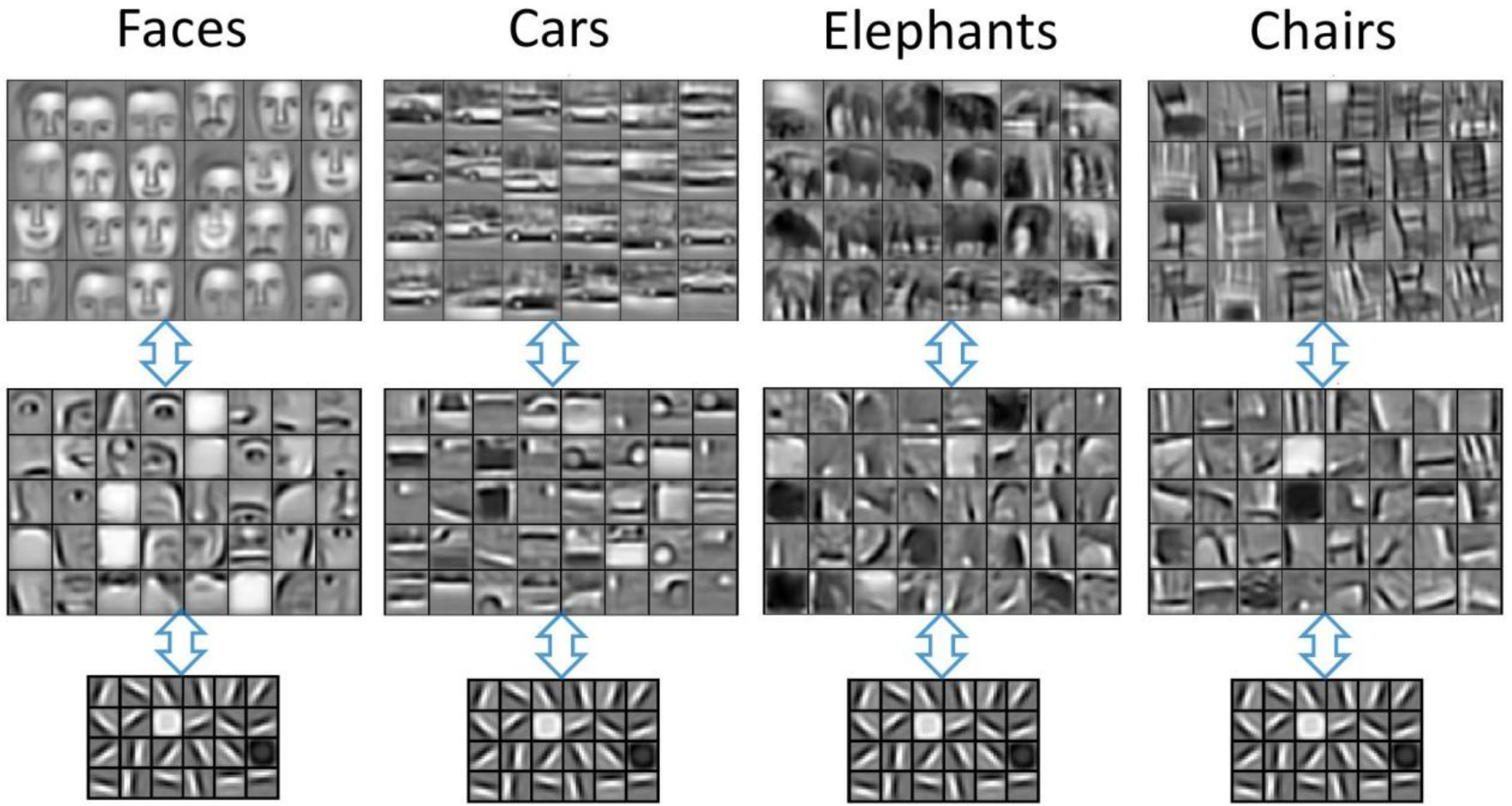
- Capas de bajo nivel (poca profundidad) detectan *features* genéricas y locales que funcionan como bloques de construcción: fonemas y bordes en reconocimiento de voz y visual, respectivamente.



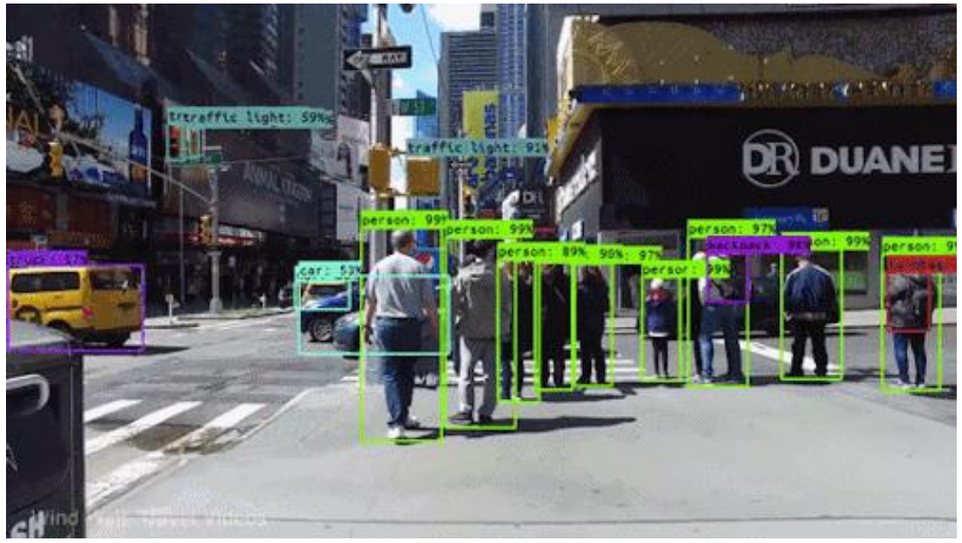
- Capas de alto nivel detectan *features* especializadas y globales, que son más robustas a variaciones: independientes del que habla o del punto de vista, en reconocimiento de voz y visual, respectivamente.

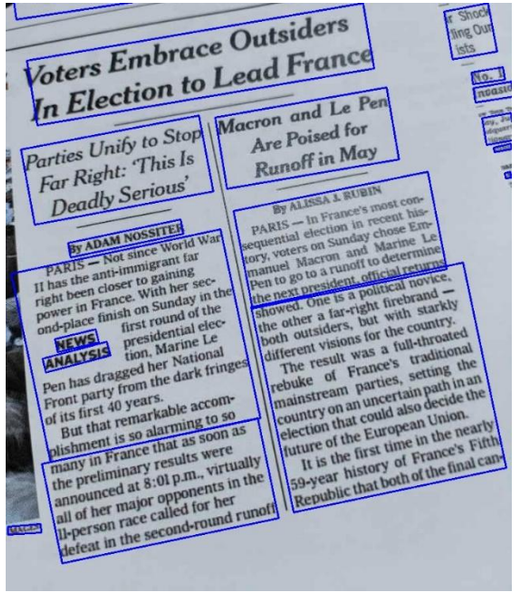


¿Qué se aprende en los distintos niveles?








Voters Embrace Outsiders In Election to Lead France

Parties Unify to Stop Far Right: 'This Is Deadly Serious'

By ADAM NOSSITER

PARIS — Not since World War II has the anti-immigrant far right been closer to gaining power in France. With her second-place finish on Sunday in the first round of the presidential election, Marine Le Pen has dragged her National Front party from the dark fringes of its first 40 years.

But that remarkable accomplishment is so alarming to so many in France that as soon as the preliminary results were announced at 8:01 p.m., virtually all of her major opponents in the 11-person race called for her defeat in the second-round runoff

NEWS ANALYSIS

Macron and Le Pen Are Poised for Runoff in May

By ALISSA J. RUBIN

PARIS — In France's most consequential election in recent history, voters on Sunday chose Emmanuel Macron and Marine Le Pen to go to a runoff to determine the next president, official returns showed. One is a political novice, the other a far-right firebrand — both outsiders, but with starkly different visions for the country.

The result was a full-throated rebuke of France's traditional mainstream parties, setting the country on an uncertain path in an election that could also decide the future of the European Union.

It is the first time in the nearly 59-year history of France's Fifth Republic that both of the final can-

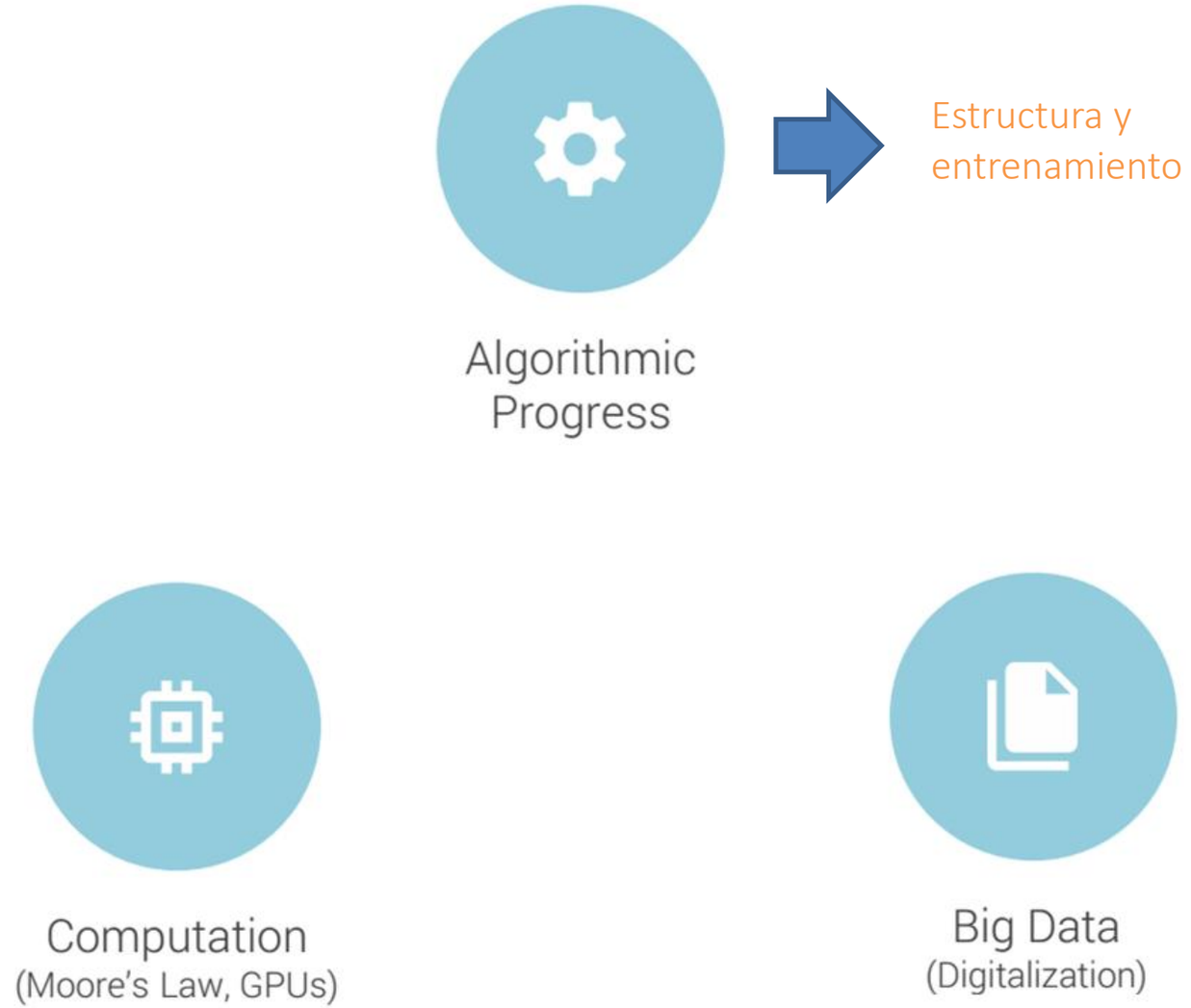


Generated Question	Answer
How many bears are laying on the ice?	“two”
What are the two animals laying on the ice?	“bears”
What are the bears doing?	“laying down”
Two bears are laying down on what?	“ice”
Where are the bears laying?	“on the ice”
Are the bears on the ice?	“yes”



A dog wearing a Superhero outfit with red cape flying through the sky

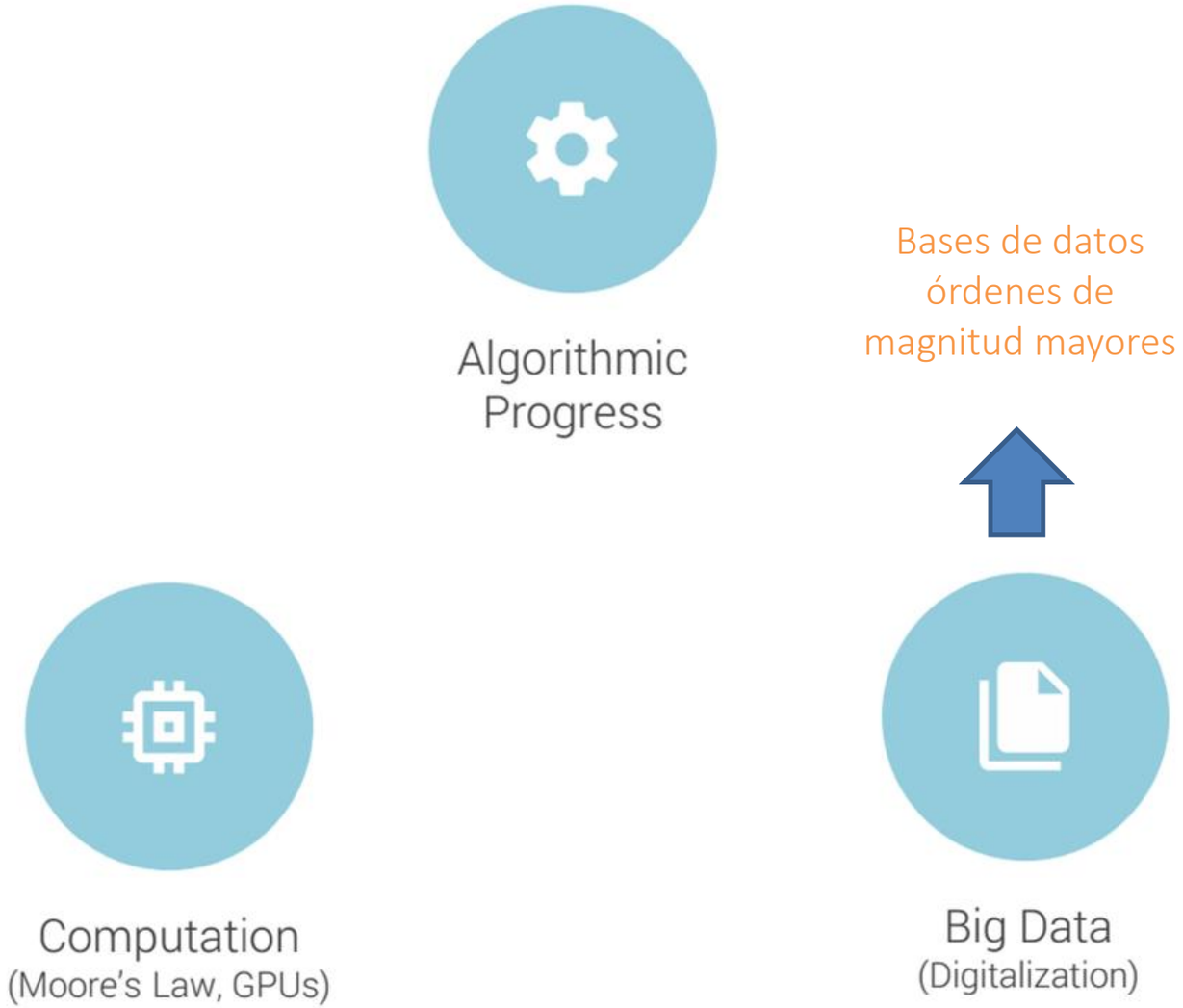
Éxito se debe a tres elementos centrales

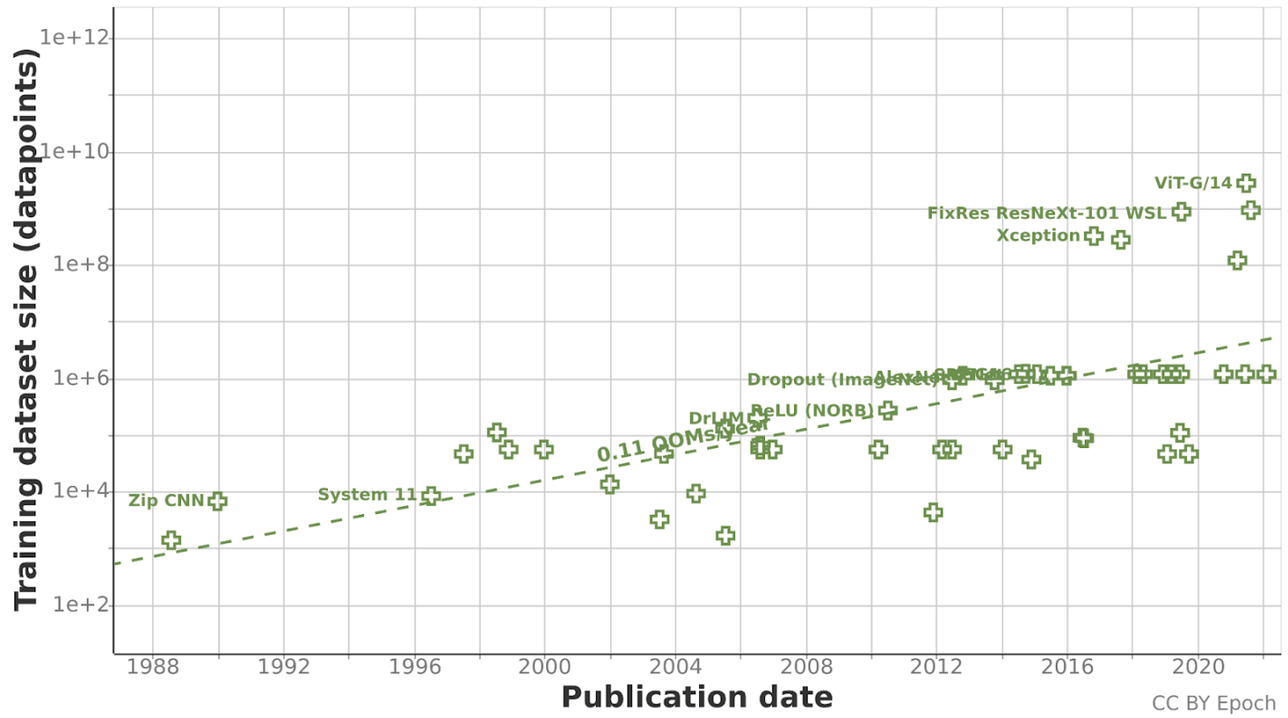




Desarrollaron la parte técnica de Deep Learning en Canadá, mientas el resto del mundo se dedicaba a otros temas

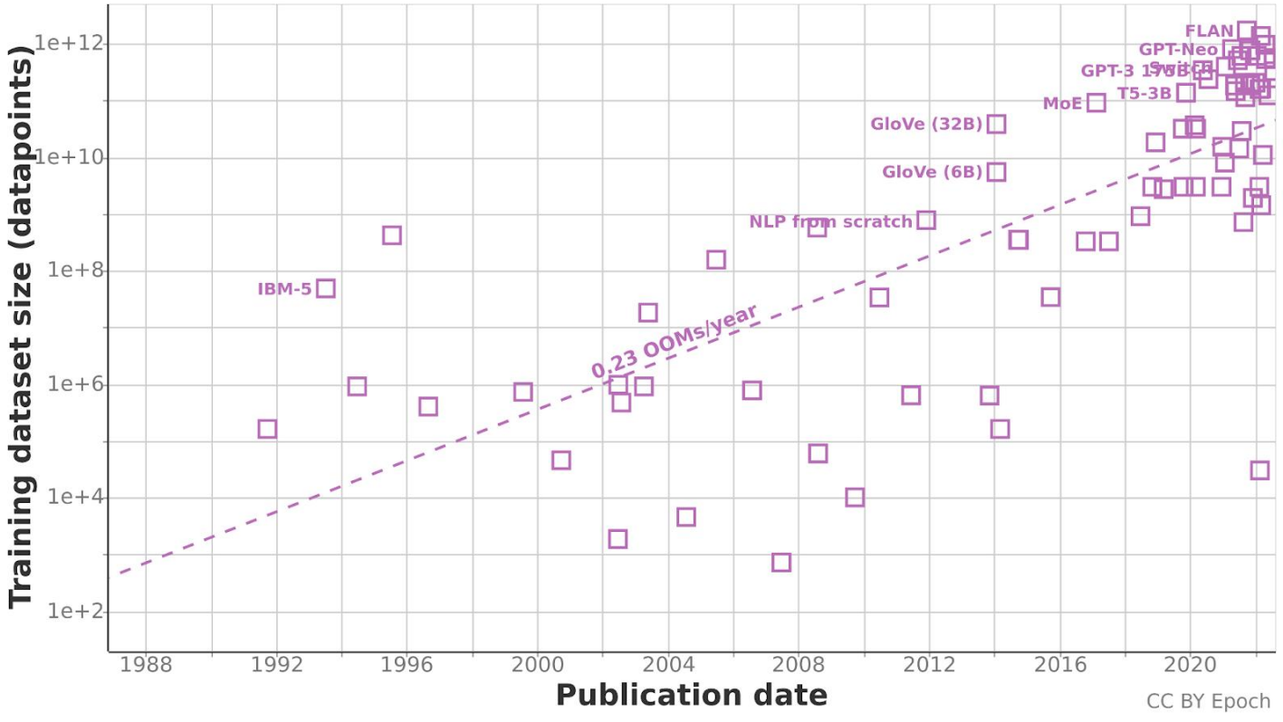
Éxito se debe a tres elementos centrales





Datos para Reconocimiento Visual

Datos para Procesamiento de Lenguaje Natural (NLP)



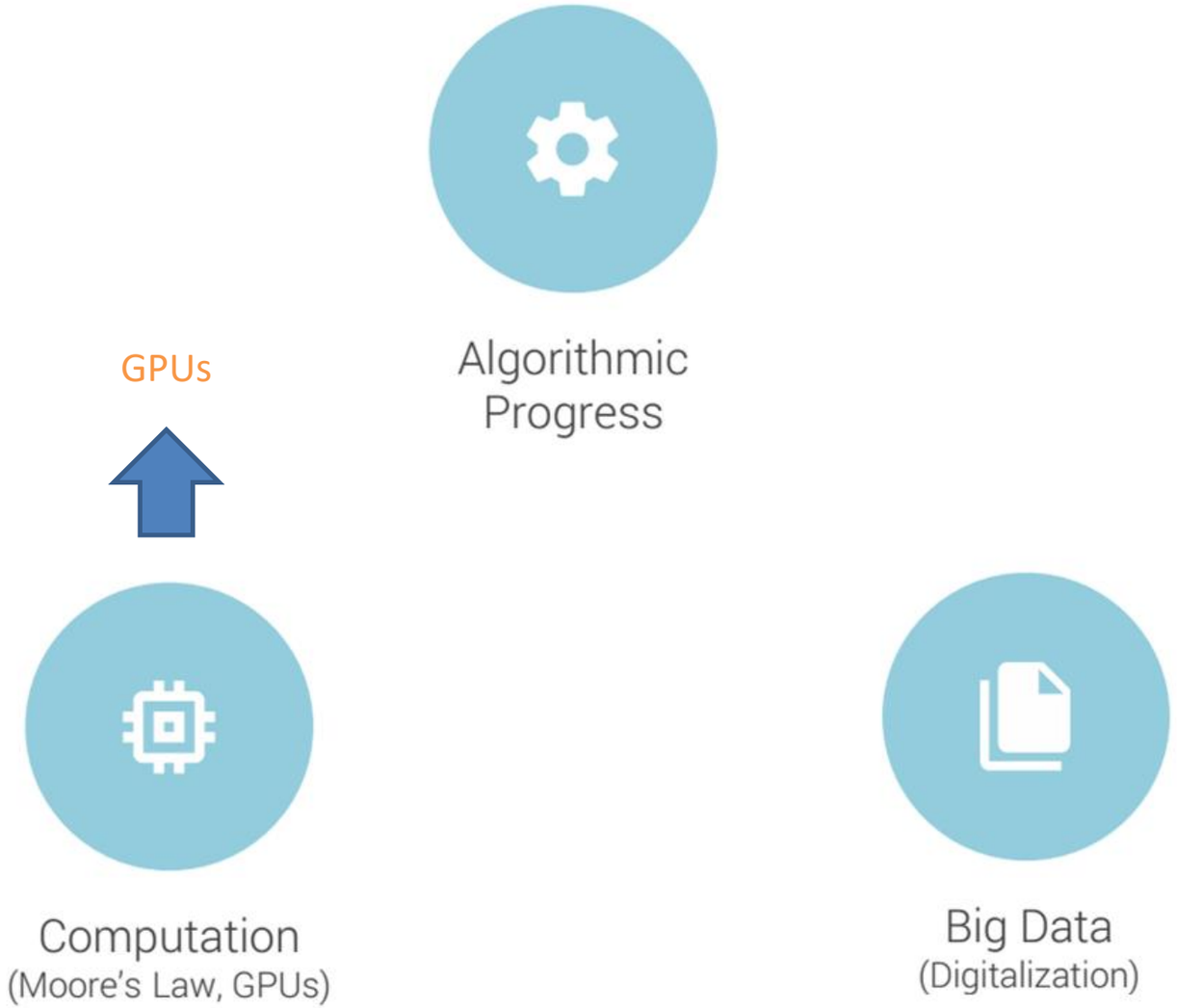


OpenAI
CLIP

LAION-5B
Large-scale Artificial Intelligence Open Network

1.2M **400M** **5,000M**

Éxito se debe a tres elementos centrales



Uso de aceleradores es una necesidad

- Volumen de datos y cómputo descarta el uso de arquitecturas tradicionales.



Uso de aceleradores es una necesidad

- Volumen de datos y cómputo descarta el uso de arquitecturas tradicionales.

- Arquitecturas distribuidas y paralelas son la solución: clusters con arreglos de GPUs.





Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación



IIC2613 - Inteligencia Artificial

Introducción a Deep Learning

Hans Löbel
Dpto. Ingeniería de Transporte y Logística
Dpto. Ciencia de la Computación