



BIG DATA WEEK
A GLOBAL FESTIVAL OF DATA

Copyright © SAS Institute Inc. All rights reserved.

El camino que
comenzando en la Estadística
pasa por la Minería de Datos
y llega a la Inteligencia Artificial

Sergio Uassouf



Vacuna contra el humo

Si piensas que los usuarios de tus programas son unos idiotas,
sólo los idiotas usarán tus programas (Linus Torvalds, creador del Linux)



- Explíqueme como lo hace



- Muéstreme como funciona

*Todo lo que le contaré es verdad,
pero en 40 minutos no puedo contarle toda la verdad*

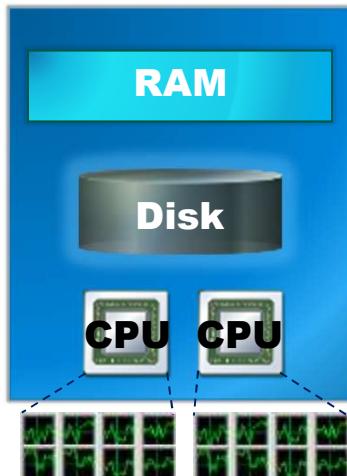


Mi Objetivo Personal



Comienzo de los Tiempos Informáticos

- Desde los inicios de la informática un computador, ya sea personal o empresarial está compuesto de 3 componentes principales.



MEMORIA

UNIDADES DE
ALMACENAMIENTO

UNIDADES DE
PROCESAMIENTO



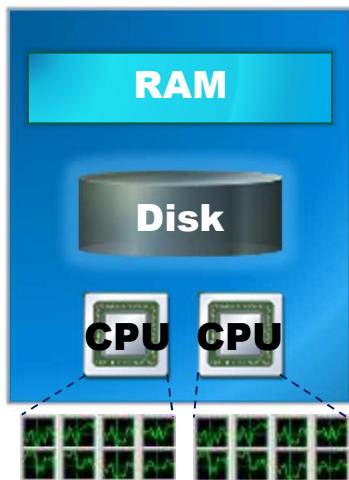
Alan Turing
1912-1954



John Von Neumann
1903-1957

Comienzo de los Tiempos Informáticos

- Desde los inicios de la informática un computador, ya sea personal o empresarial está compuesto de 3 componentes principales.

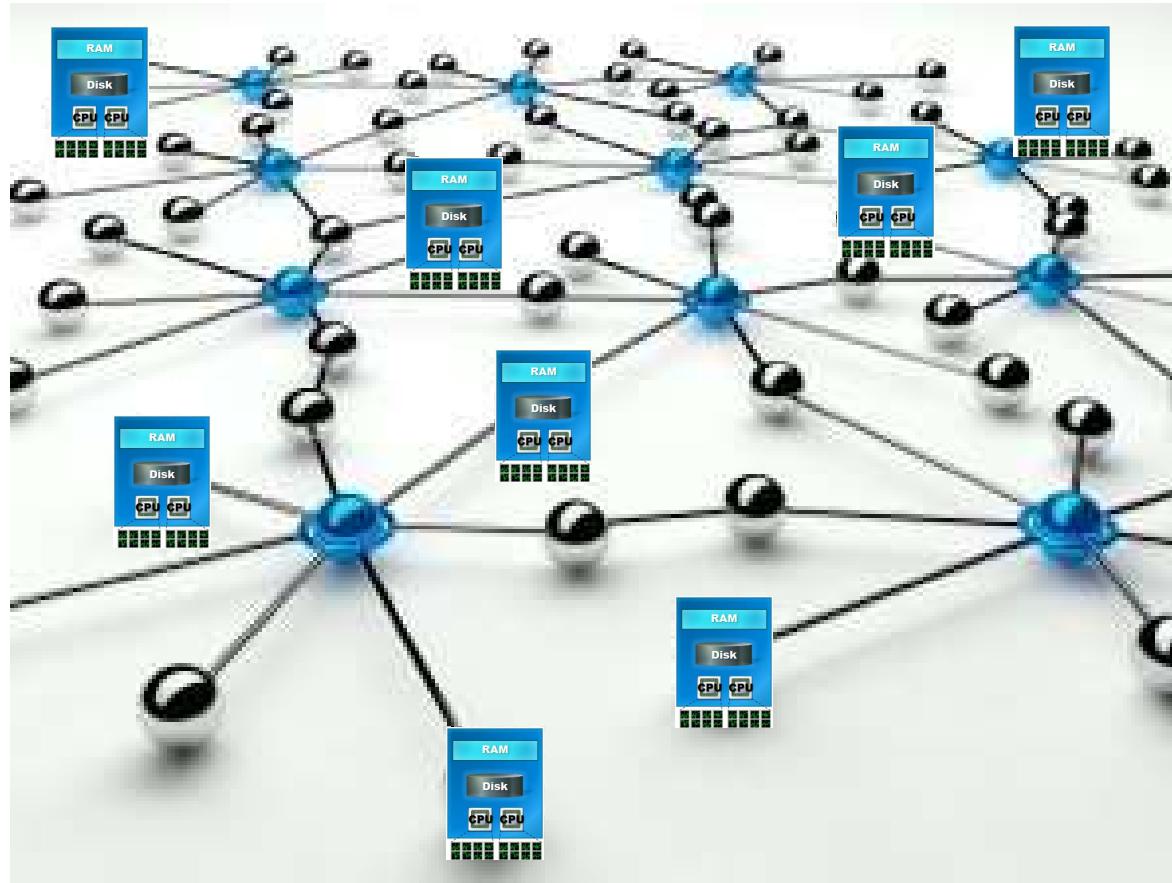


MEMORIA ← NANOSEGUNDOS,
PERO SE BORRA

UNIDADES DE ← MILISEGUNDOS,
ALMACENAMIENTO ← PERO NO SE BORRA

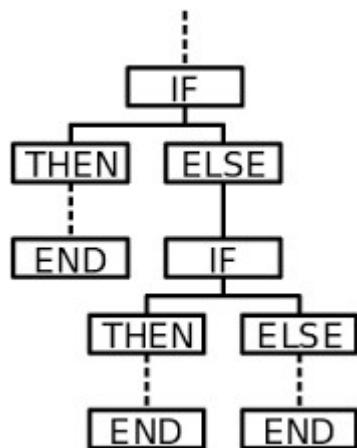
UNIDADES DE ← IF-THEN-ELSE
PROCESAMIENTO ←
¡¡¡ HAGAMOSLE
UN MERECIDO
HOMENAJE !!!

Sistemas Distribuidos => iii Networking !!!



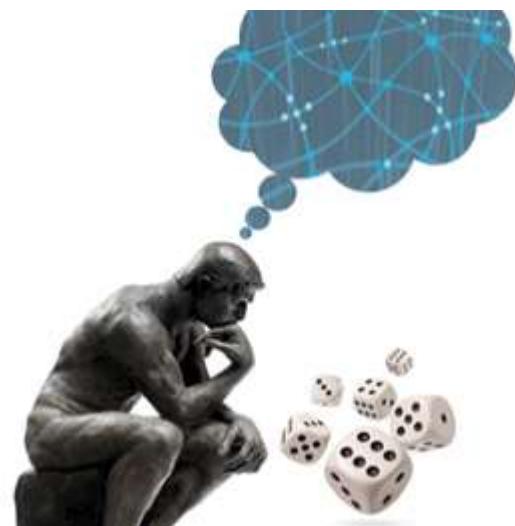
La batalla de la decisión ¿Determinístico o Probabilístico?

- Un modelo determinístico arroja un único resultado para un conjunto de variables de entrada determinado.



La batalla de la decisión ¿Determinístico o Probabilístico?

- Un modelo probabilístico arroja un conjunto de resultados posibles indicando la probabilidad de ocurrencia de cada resultado.





Nacimiento de la Estadística Godofredo Achenwall

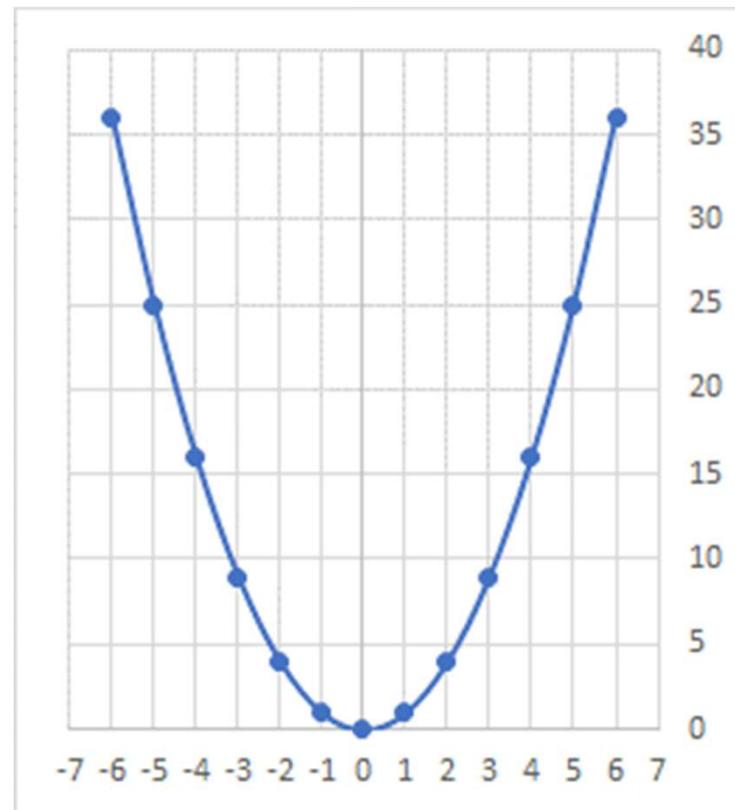


**Elbing, Prusia Oriental
(1719-1772)**

Principales Técnicas Analíticas Comienzo de la Historia: Estadística - Regresiones

$$Y = X^2$$

Y	X
36	6
25	5
16	4
9	3
4	2
1	1
0	0
1	-1
4	-2
9	-3
16	-4
25	-5
36	-6

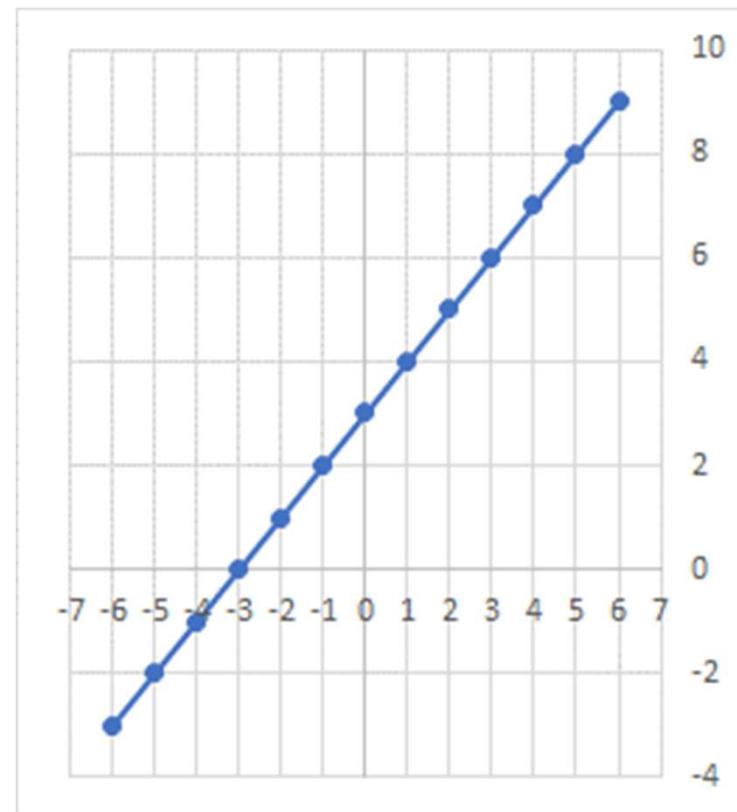


Principales Técnicas Analíticas

Comienzo de la Historia: Estadística - Regresiones

$$Y = X + 3$$

Y	X
9	6
8	5
7	4
6	3
5	2
4	1
3	0
2	-1
1	-2
0	-3
-1	-4
-2	-5
-3	-6





Principales Técnicas Analíticas

Comienzo de la Historia: Estadística - Regresiones

Por ejemplo...

¿Cómo se relacionan la edad, el género, el estado civil, el lugar de residencia, los ingresos, la profesión,

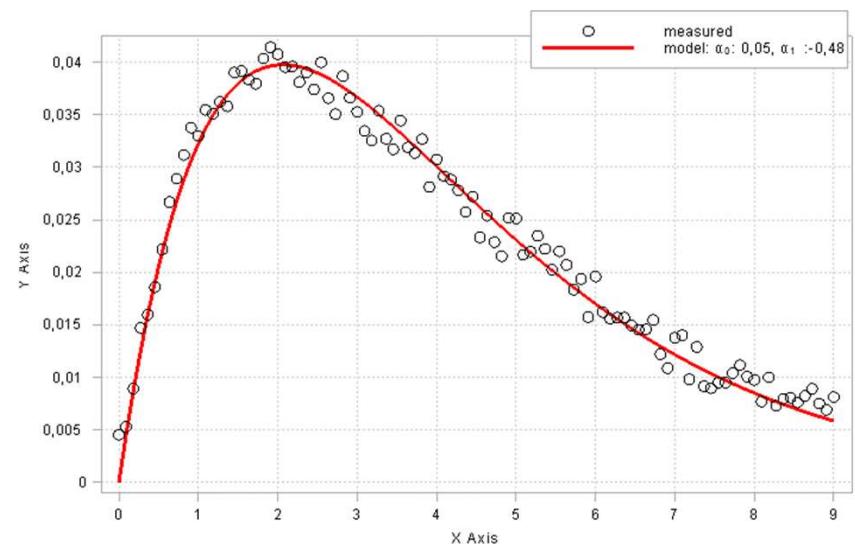
Con la propensión a que

- Acepten la oferta de mi producto...
- Estén por irse de mi empresa...
- Estén cometiendo fraudes...
- Estén lavando dinero.

Principales Técnicas Analíticas Comienzo de la Historia: Estadística - Regresiones

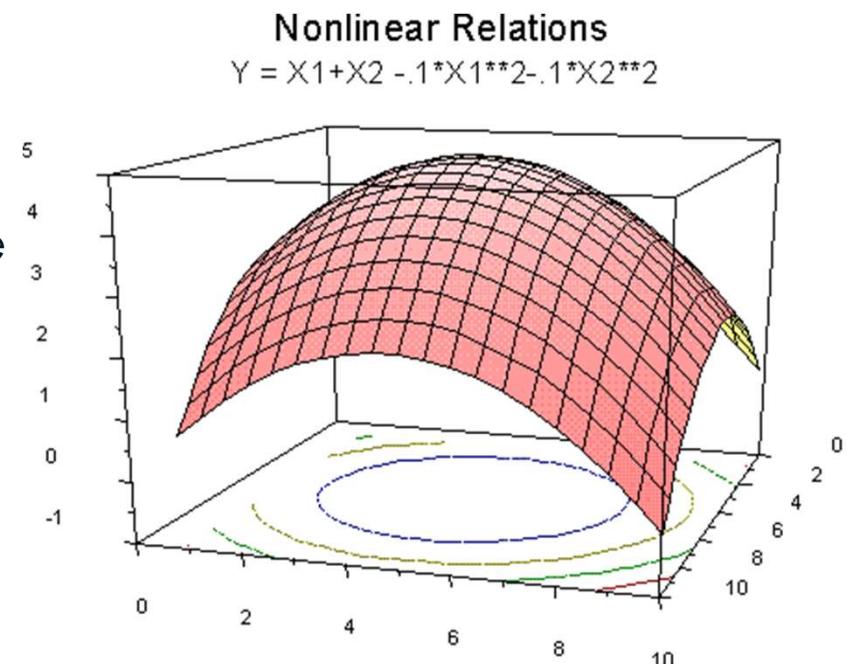
Encuentra la función que relaciona valores que conocemos (variable independiente o predictor) con los valores que estamos buscando (variable dependiente o respuesta).

- La **respuesta o variable dependiente** (Y) es la variable que se quiere predecir.
- Los **predictores o variables independientes** (X) explican la variación de la respuesta.

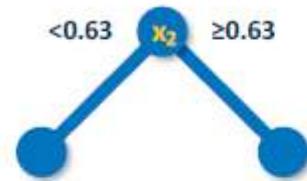


Principales Técnicas Analíticas Comienzo de la Historia: Estadística - Regresiones

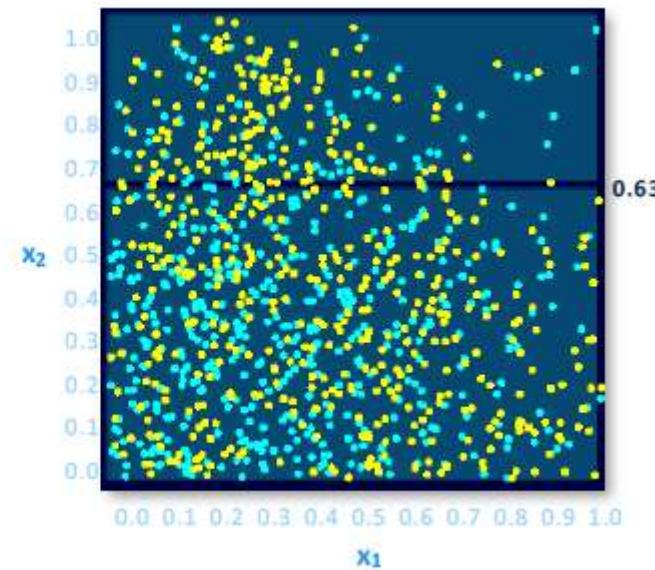
- ❑ Considere el modelo de dos variables $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$
 - ❑ Donde Y es la respuesta o variable dependiente.
 - ❑ X_1 y X_2 son los predictores o variables independientes.
 - ❑ $\beta_0, \beta_1, \beta_2$ son los parámetros a determinar.
 - ❑ ε es el término de error.



De la Estadística a la Minería de Datos y la Inteligencia Artificial: Árboles de Decisión



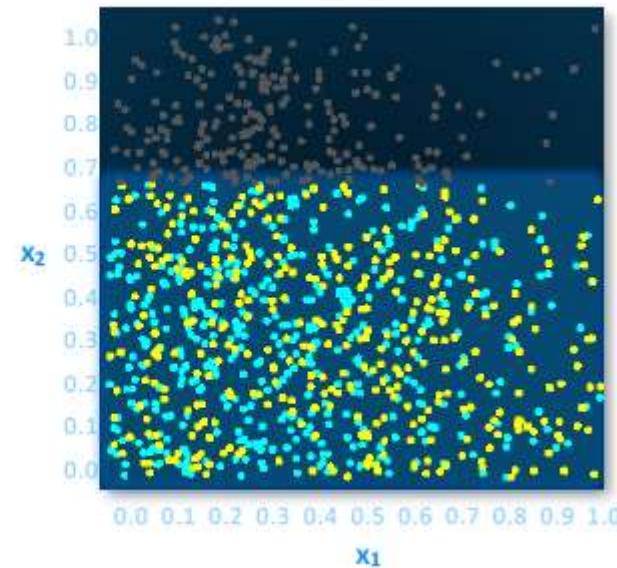
Create a partition rule
from the best partition
across all inputs.



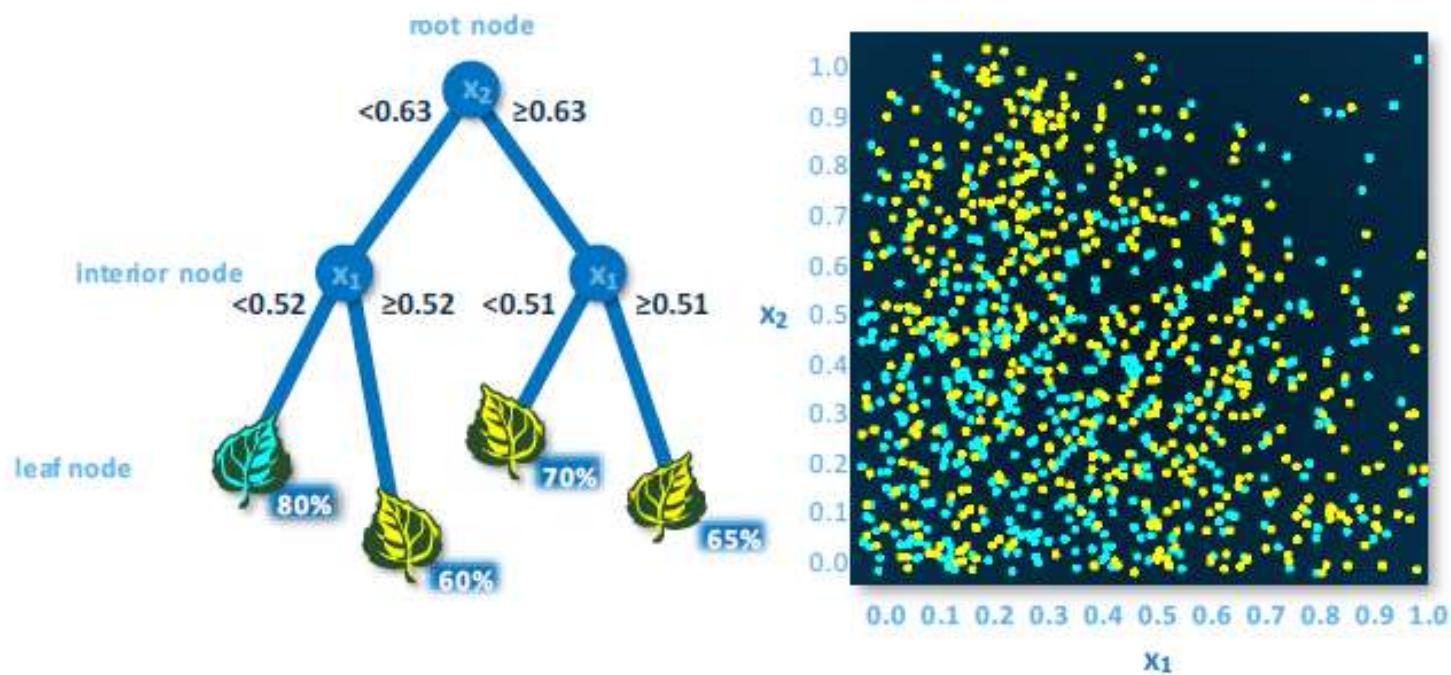
De la Estadística a la Minería de Datos y la Inteligencia Artificial: Árboles de Decisión



Repeat the process
in each subset.

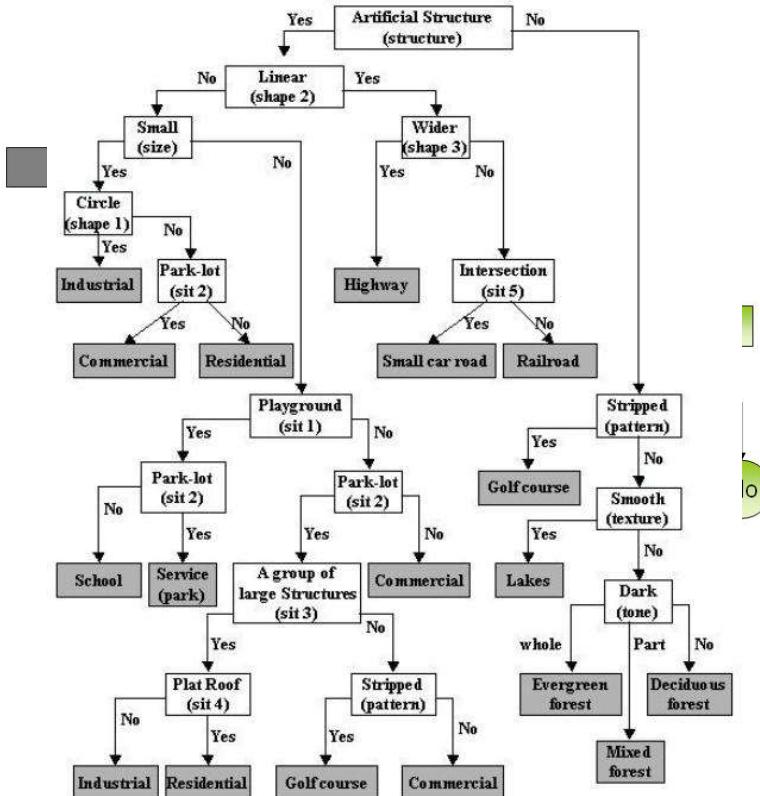


De la Estadística a la Minería de Datos y la Inteligencia Artificial: Árboles de Decisión



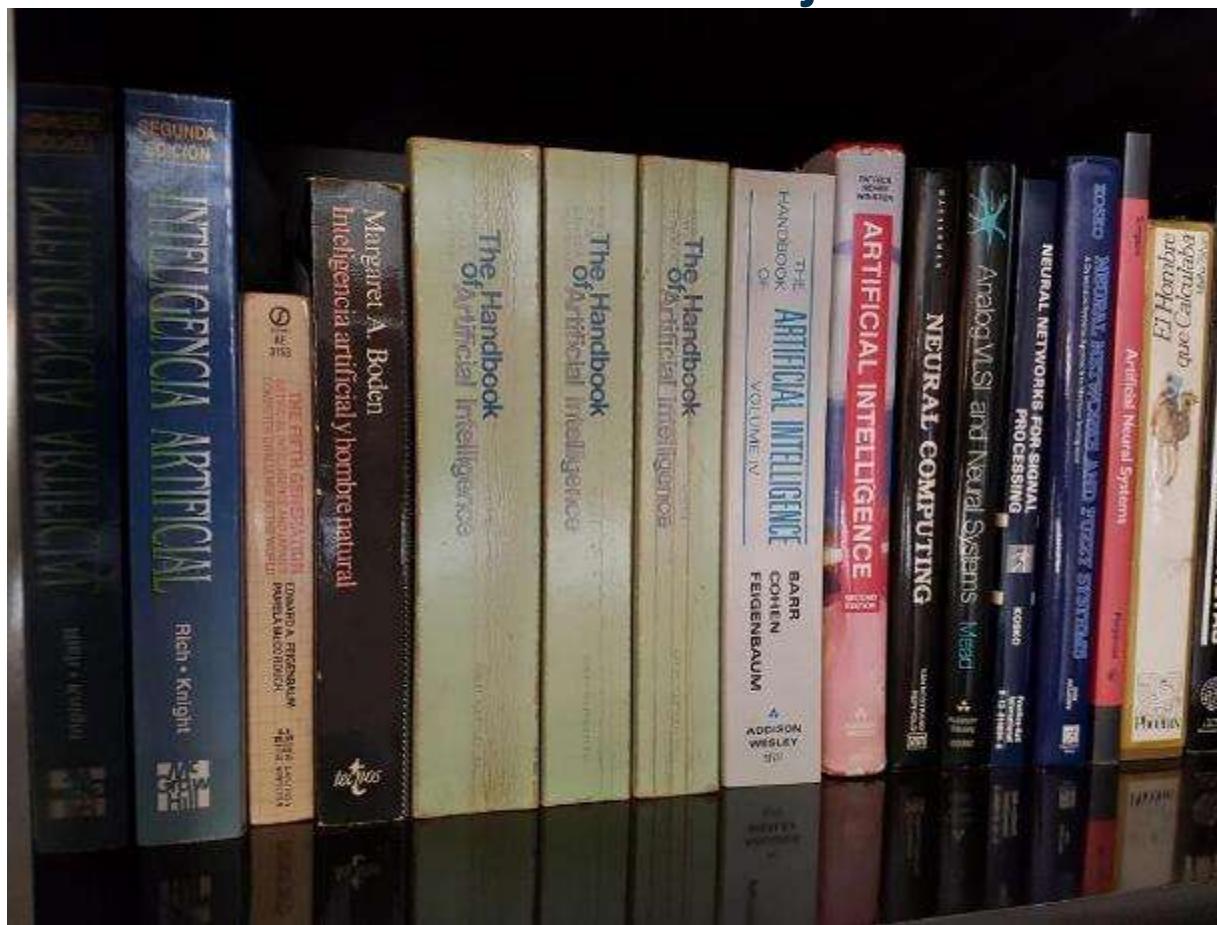
De la Estadística a la Minería de Datos y la Inteligencia Artificial: Árboles de Decisión

- Cada rama del árbol representa una elección entre alternativas.
- Cada hoja representa una clasificación o decisión.
- Por ejemplo, si debemos decidir si a una persona le ofrecemos o no un crédito.



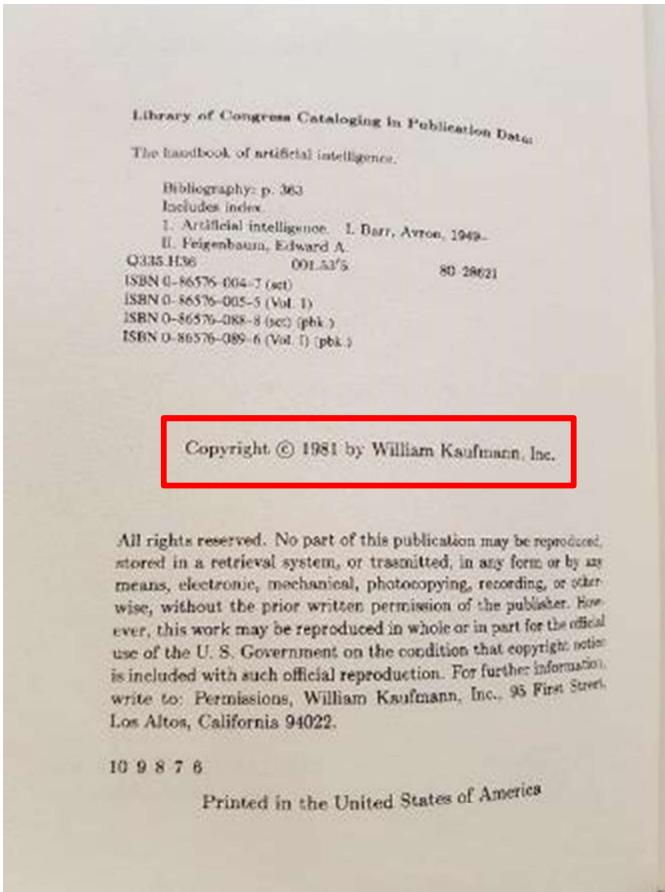


Inteligencia Artificial: ¿Porqué renace ahora? De la biblioteca de mi juventud



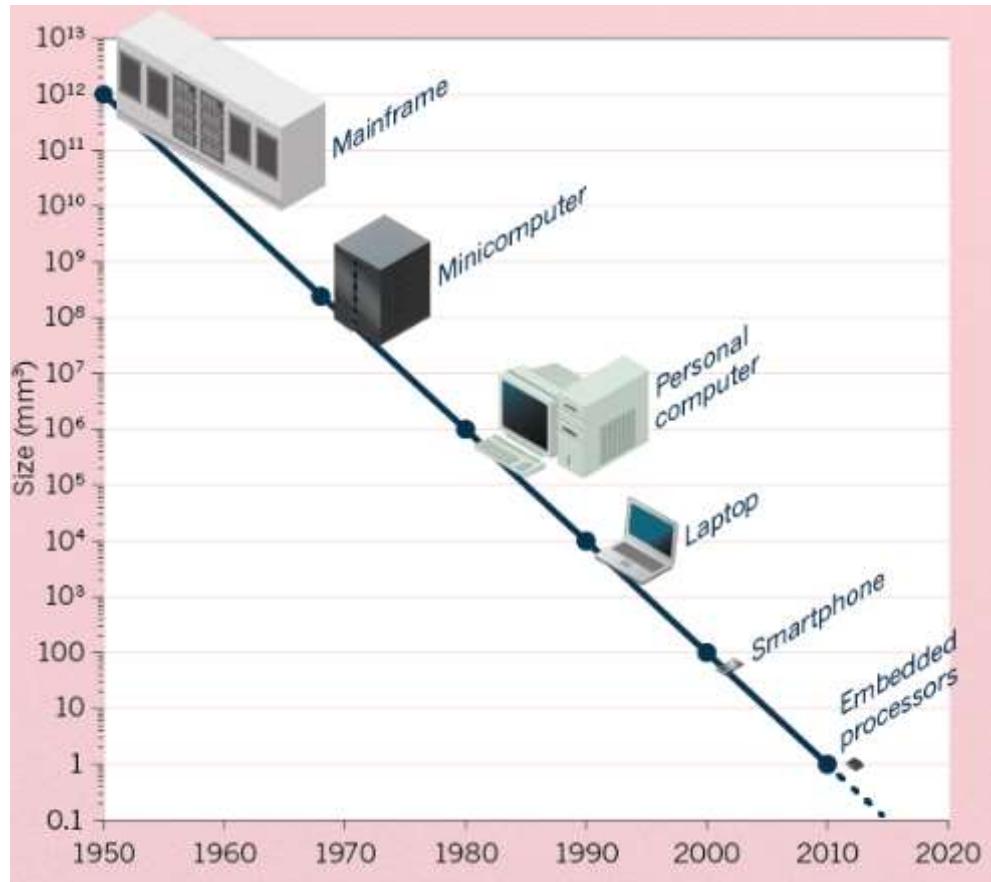


Inteligencia Artificial: ¿Porqué renace ahora? De la biblioteca de mi juventud



Contents	
III. Knowledge Representation / 141	
A. Overview / 143	
B. Survey of representation techniques / 150	
C. Representation schemes / 160	
1. Logic / 160	
2. Procedural representations / 172	
3. Semantic networks / 180	
4. Production systems / 190	
5. Direct (analogical) representations / 200	
6. Semantic primitives / 207	
7. Frames and scripts / 215	
IV. Understanding Natural Language / 223	
A. Overview / 225	
B. Machine translation / 233	
C. Grammars / 239	
1. Formal grammars / 239	
2. Transformational grammars / 245	
3. Systemic grammar / 249	
4. Case grammars / 252	
D. Parsing / 256	
1. Overview of parsing techniques / 256	
2. Augmented transition networks / 263	
3. The General Syntactic Processor / 268	
E. Text generation / 273	
F. Natural language processing systems / 281	
1. Early natural language systems / 281	
2. Wilks's machine translation system / 288	
3. LUNAR / 292	
4. SHRDLU / 295	
5. MARGIE / 300	
6. SAM and PAM / 306	
7. LIPER / 316	
V. Understanding Spoken Language / 323	
A. Overview / 325	
B. Systems architecture / 332	
C. The ARPA SUR projects / 343	
1. HEARSAY / 343	
2. HARPY / 349	
3. HWIM / 353	
4. The SRI/SDC speech systems / 358	

Inteligencia Artificial: ¿Porqué renace ahora? Ley de Moore

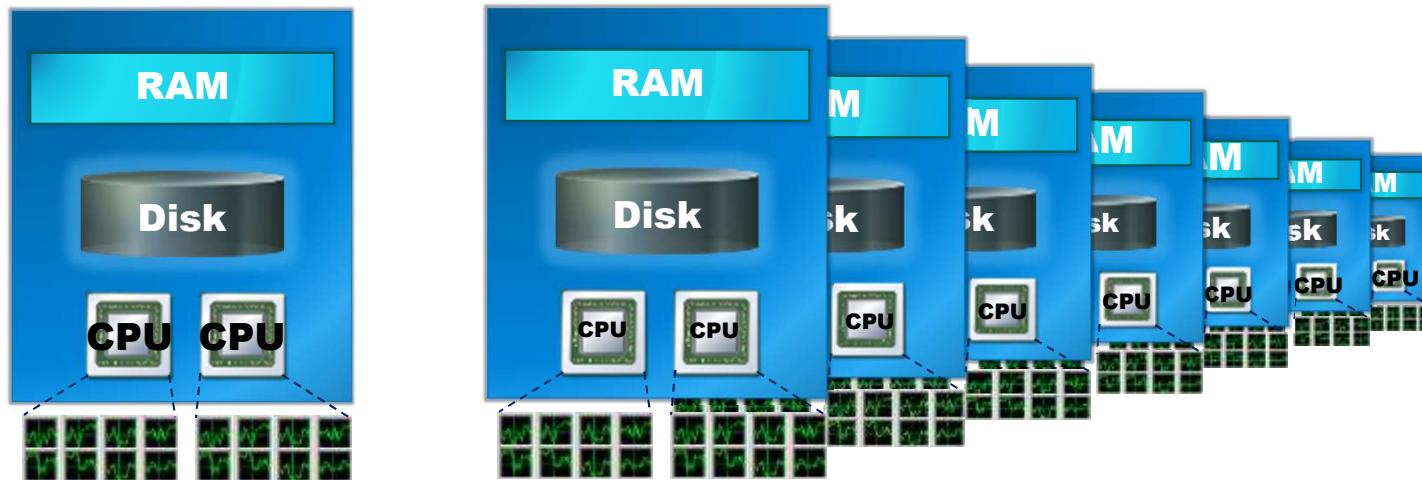


**22 cores
1,54 TB RAM**



Inteligencia Artificial: ¿Porqué renace ahora?

Big Data = Procesamiento Masivamente Paralelo



Inteligencia Artificial

es la ciencia de entrenar sistemas para emular tareas humanas mediante el Aprendizaje y la Automatización



Entender
el Contexto



Aprender Patrones



Interpretar
Lenguaje e Imágenes

Machine Learning

es la ciencia de crear sistemas que aprenden de los datos con algoritmos de aprendizaje que iteran hasta lograr una aproximación satisfactoria...



Entender
el Contexto



Aprender Patrones



Interpretar
Lenguaje e Imágenes

Machine Learning

...en otras palabras, el programa es construído automáticamente por el algoritmo de aprendizaje



Entender
el Contexto

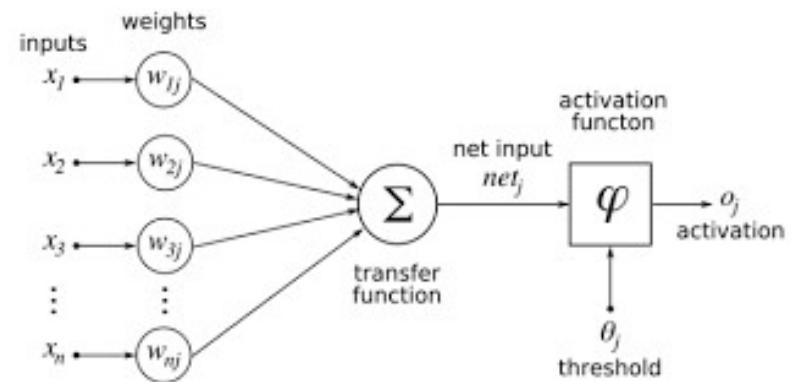
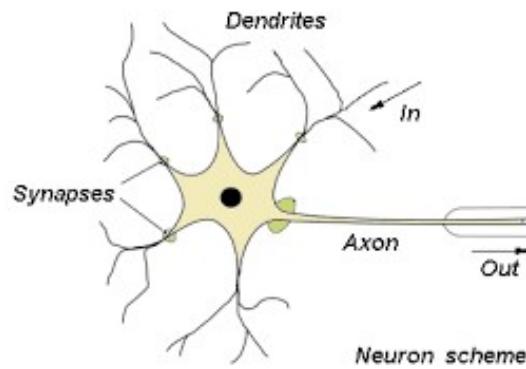


Aprender Patrones



Interpretar
Lenguaje e Imágenes

De la Minería de Datos a la Inteligencia Artificial Machine Learning ≈ Redes Neuronales



Nuestra dificultad para entender el funcionamiento del cerebro radica en que para hacerlo utilizamos el mismo cerebro



De la Minería de Datos a la Inteligencia Artificial Machine Learning ≈ Redes Neuronales



```
proc nnet data=mycaslib.bank_final missing=mean;
partition fraction(validate=0.4 seed=12345);
target b_tgt / level=nominal;
input demog_ho demog_genf demog_genm
LOG_TN_IMZ_demog_age LOG_demog_homeval
LOG_demog_inc LOG_demog_pr LOG_TN_rfml LOG_rfm2
LOG_rfm5 LOG_rfm6 LOG_rfm7 LOG /
level=interval;
input cat_input1 cat_input2 / level=nominal;
hidden 42;
train outmodel=mycaslib.nnet_params;
optimization algorithm=sgd maxiters=50 learningrate=1.0E-4;
code file="/home/student/nnet_score.sas";
run;
```



```
# define the keras model
model = Sequential()
model.add(Dense(12, input_dim=8, activation='relu'))
model.add(Dense(8, activation='relu'))
model.add(Dense(1, activation='sigmoid'))

# compile the keras model
model.compile(loss='binary_crossentropy',
optimizer='adam', metrics=['accuracy'])

# fit the keras model on the dataset
model.fit(X, y, epochs=150, batch_size=10)
```

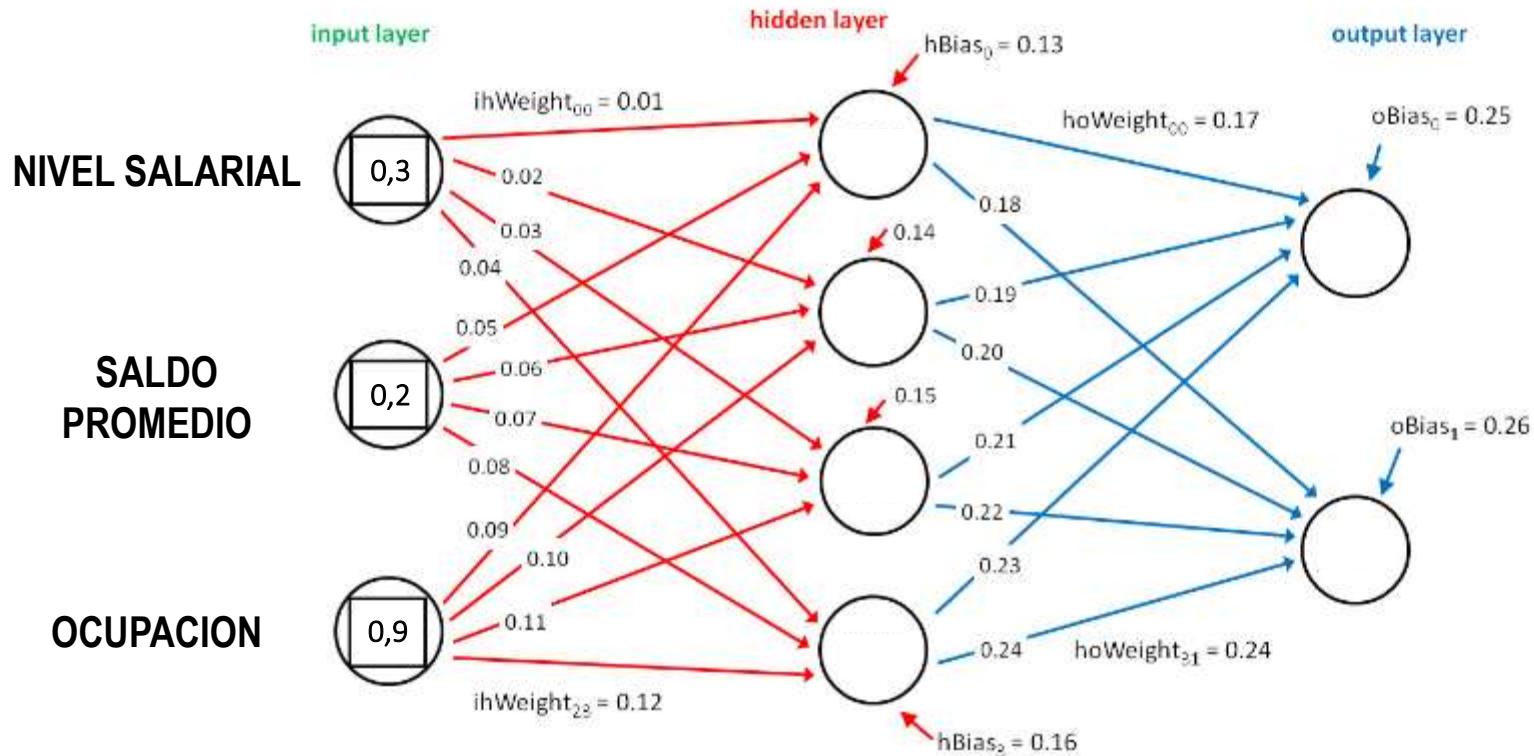


```
# specify layers for the neural network:
# input layer of size 11 (features), two intermediate of size 5 and 4
# and output of size 7 (classes)
layers = [11, 5, 4, 4, 3, 7]

# create the trainer and set its parameters
FNN = MultilayerPerceptronClassifier(labelCol="indexedLabel", featuresCol="indexedFeatures",
                                       maxIter=100, layers=layers, blockSize=128, seed=1234)
# Convert indexed labels back to original labels.
labelConverter = IndexToString(inputCol="prediction", outputCol="predictedLabel",
                                labels=labelIndexer.labels)
# Chain indexers and forest in a Pipeline
from pyspark.ml import Pipeline
pipeline = Pipeline(stages=[labelIndexer, featureIndexer, FNN, labelConverter])
# train the model
# Train model. This also runs the indexers.
model = pipeline.fit(trainingData)
```

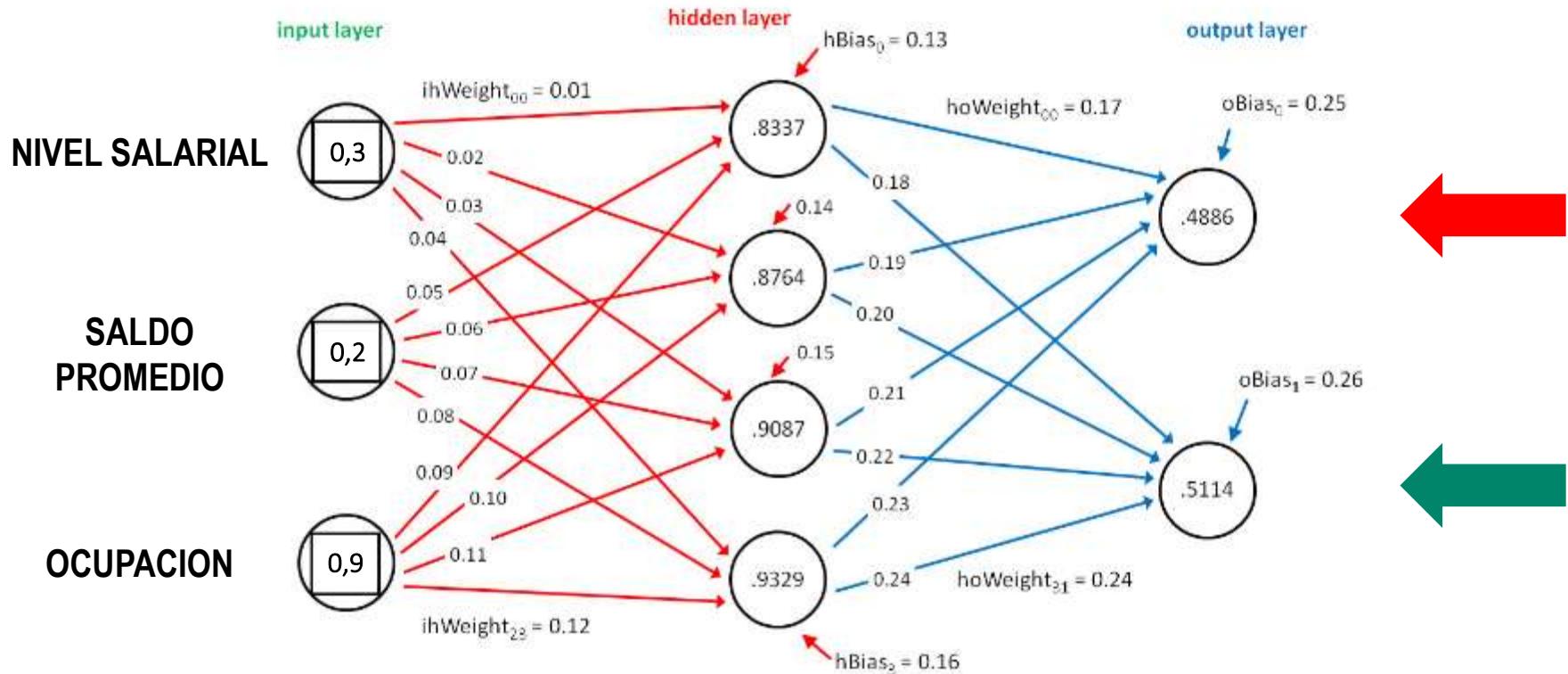
Inteligencia Artificial Machine Learning ≈ Redes Neuronales

Nivel Salarial = 0,3; Saldo Promedio = 0,2; Ocupación = 0,9 → Buen pagador



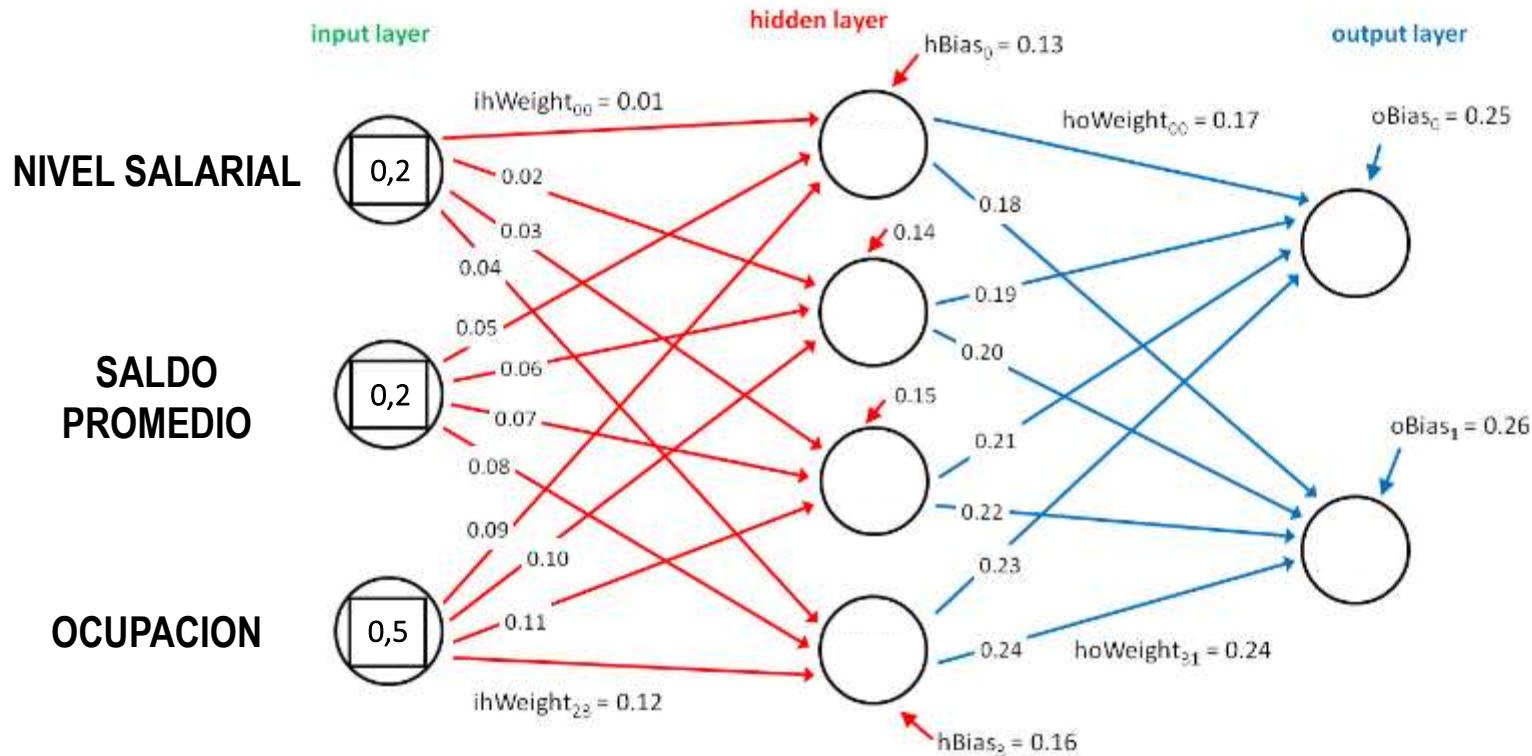
Inteligencia Artificial Machine Learning ≈ Redes Neuronales

Nivel Salarial = 0,3; Saldo Promedio = 0,2; Ocupación = 0,9 → Buen pagador



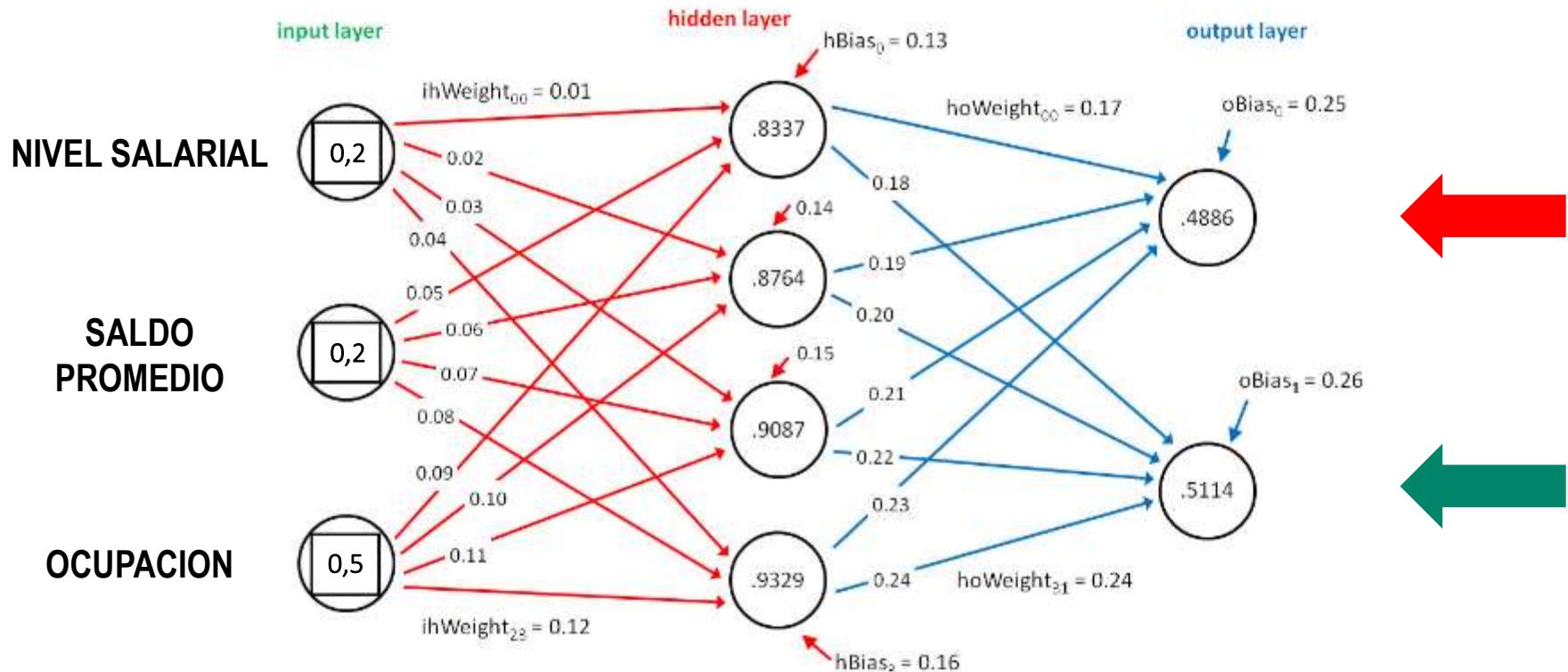
Inteligencia Artificial Machine Learning ≈ Redes Neuronales

Nivel Salarial = 0,2; Saldo Promedio = 0,2; Ocupación = 0,5 → Mal pagador



Inteligencia Artificial Machine Learning ≈ Redes Neuronales

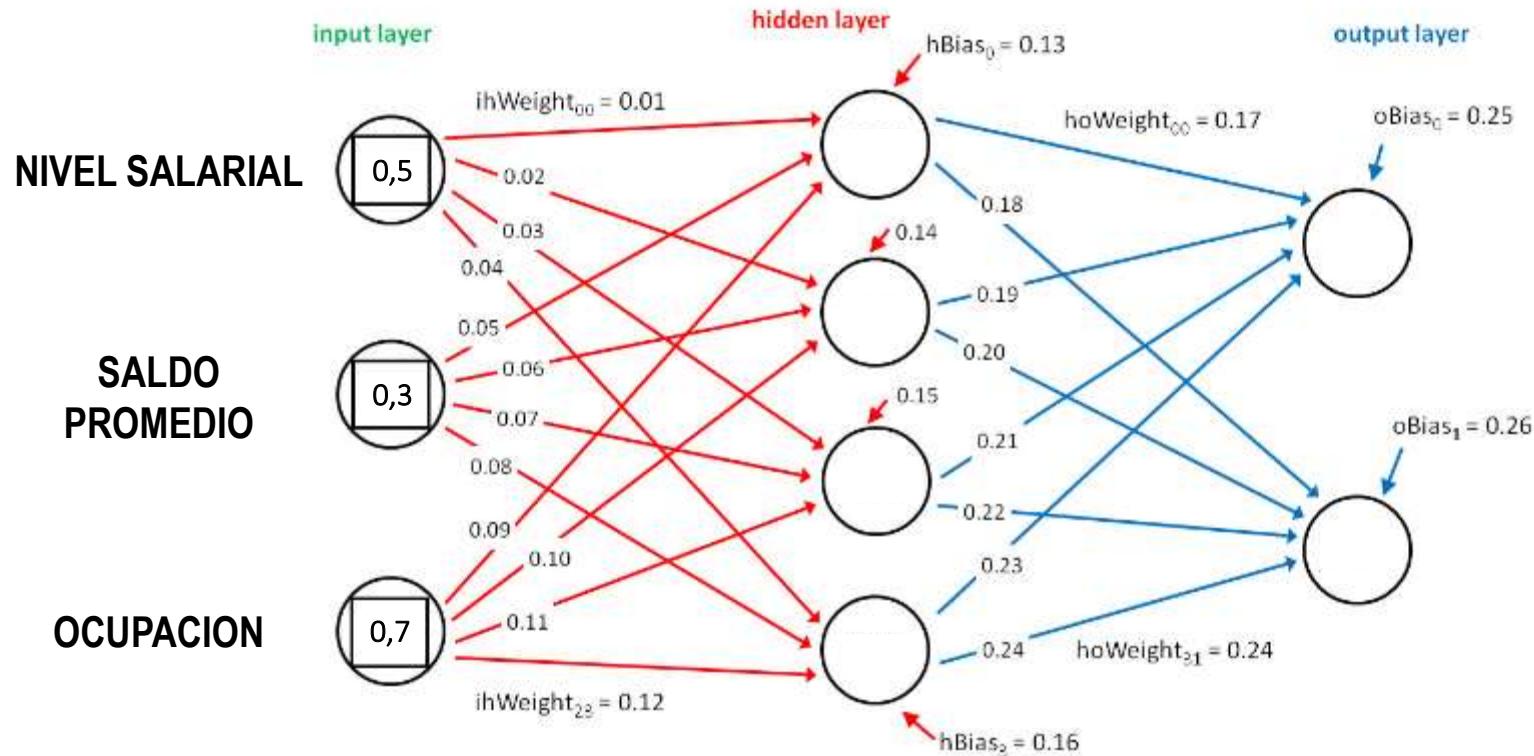
Nivel Salarial = 0,2; Saldo Promedio = 0,2; Ocupación = 0,5 → Mal pagador



<https://mattmazur.com/2015/03/17/a-step-by-step-backpropagation-example/>

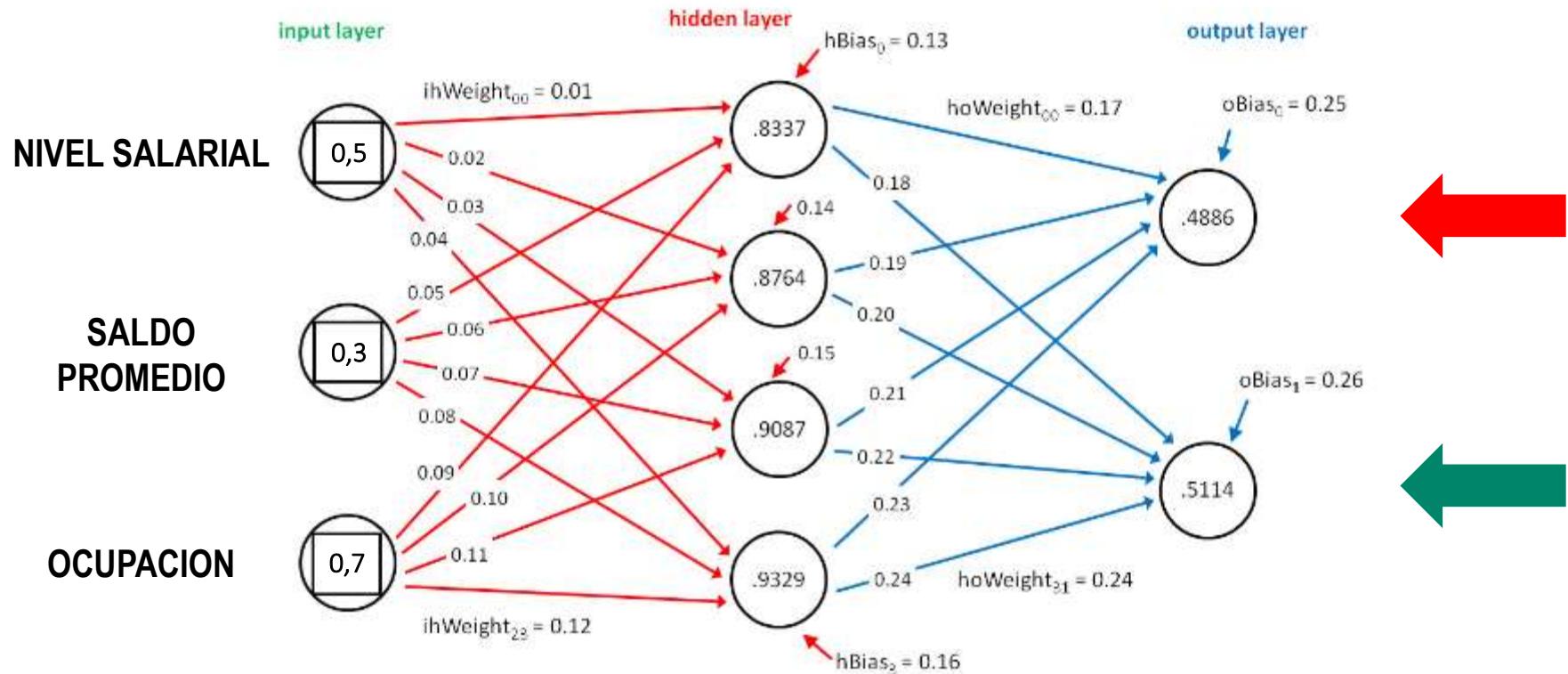
Inteligencia Artificial Machine Learning ≈ Redes Neuronales

Nivel Salarial = 0,5; Saldo Promedio = 0,3; Ocupación = 0,7



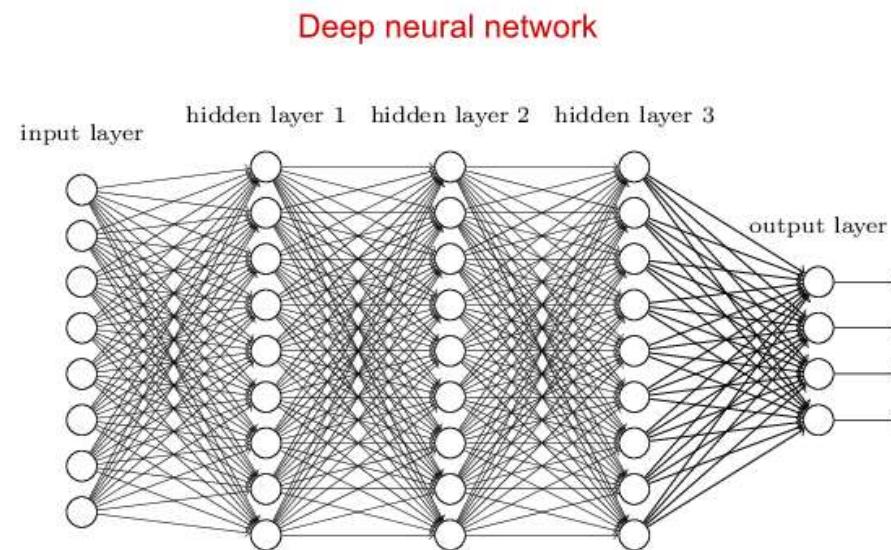
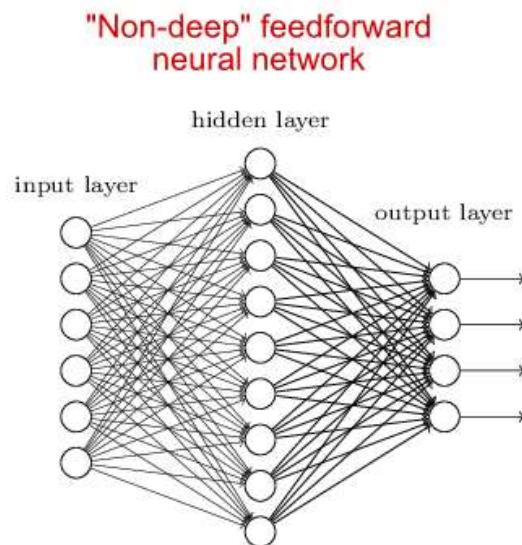
Inteligencia Artificial Machine Learning ≈ Redes Neuronales

Nivel Salarial = 0,5; Saldo Promedio = 0,3; Ocupación = 0,7 → Mal pagador



Inteligencia Artificial

Deep Learning ≈ Redes Neuronales de Varias Capas Ocultas



Inteligencia Artificial

Aprendizaje

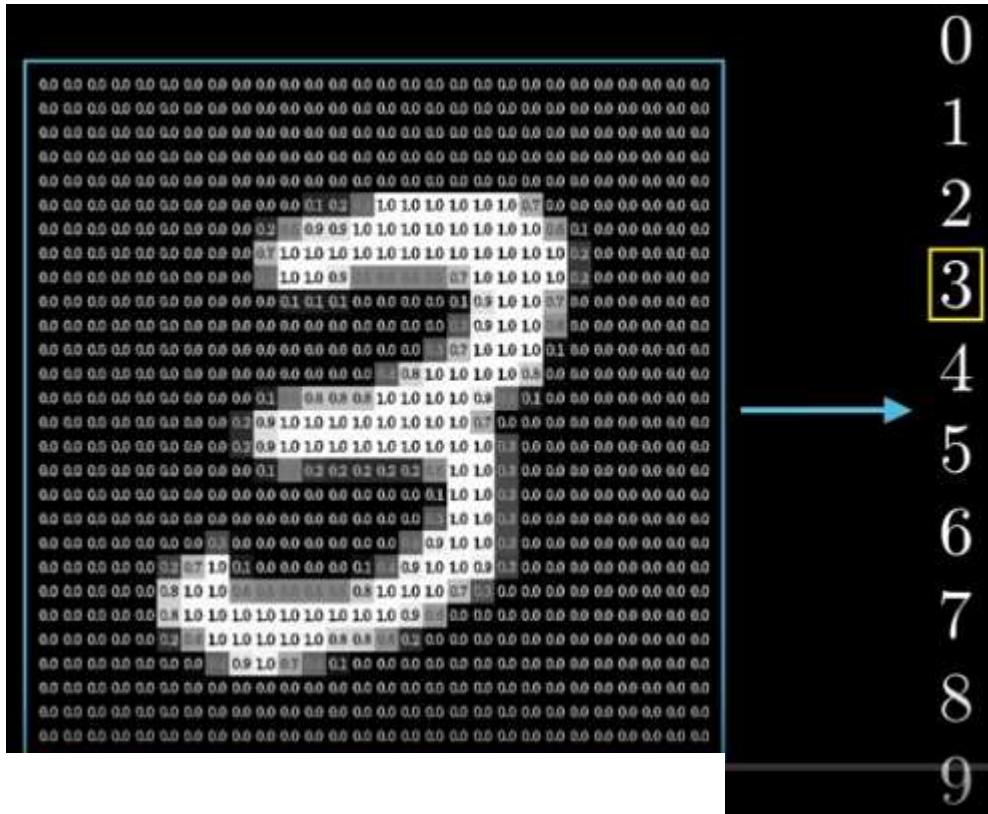
- Imágenes
- Transacciones
- Usuarios y Clientes
- Imágenes Médicas
- Lenguajes
- Emails

Automatización

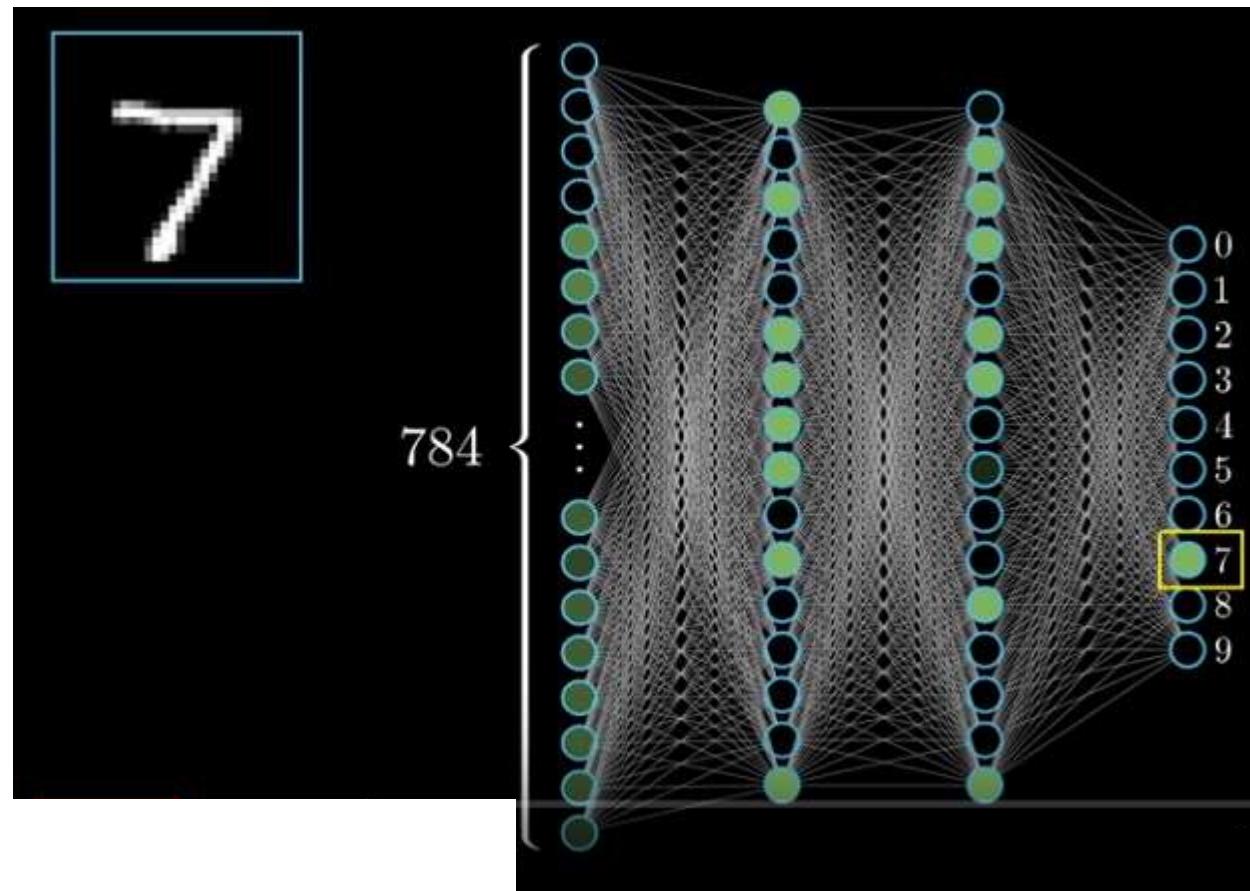
- ¿Es usted?
- ¿Está cometiendo fraude?
- ¿Necesita nuestros servicios?
- ¿Hay enfermedad?
- ¿Qué está consultando?
- ¿Es un spam?



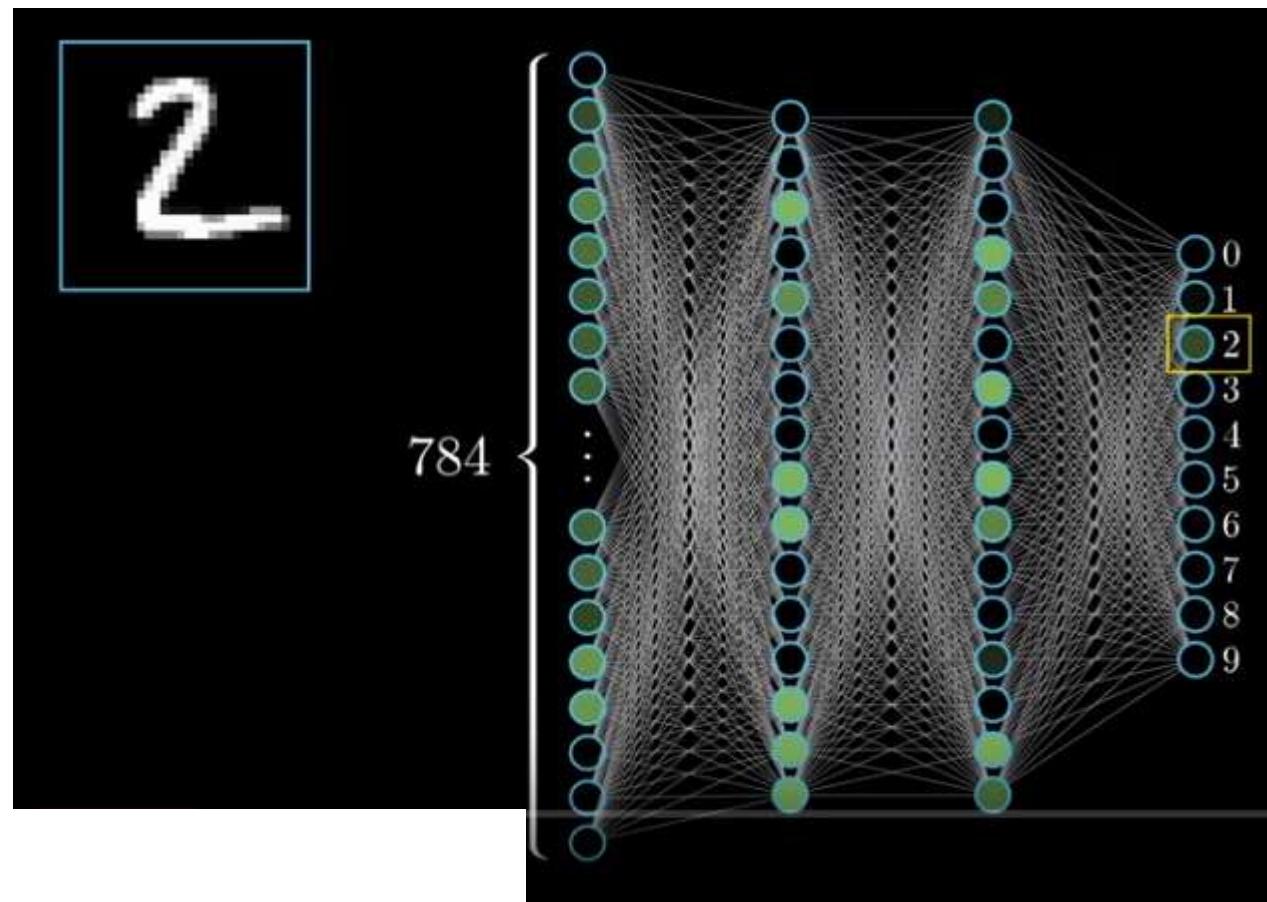
Reconocimiento de Caracteres Manuscritos



Reconocimiento de Caracteres Manuscritos

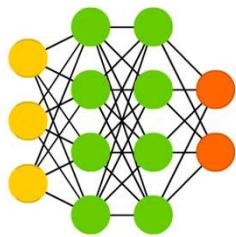


Reconocimiento de Caracteres Manuscritos

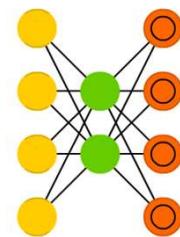


SAS Deep Learning – Tipos de Redes Neuronales

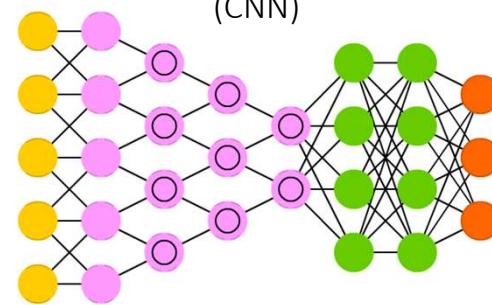
Deep FF Neural Network (DNN)



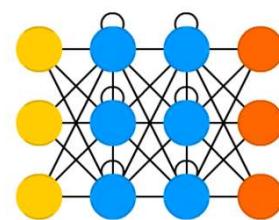
Auto Encoder (AE)



Convolutional Neural Networks (CNN)

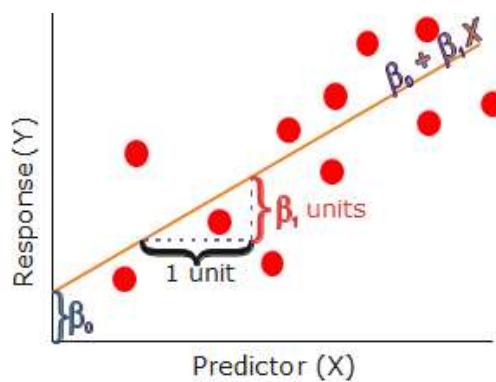


Recurrent Neural Networks (RNN)



Recapitulando Principales Técnicas Analíticas

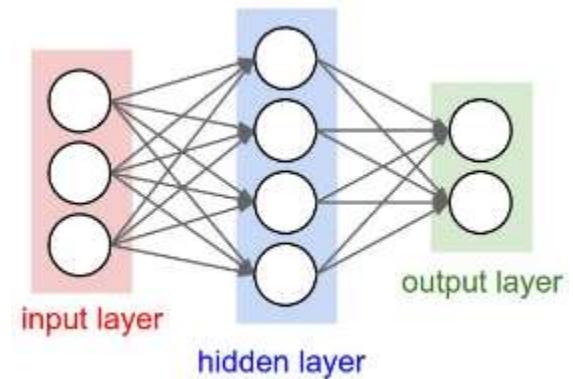
REGRESIONES



ARBOLES DE DECISIÓN



REDES NEURONALES



SAS Viya – Única Plataforma Analítica Integral

Para satisfacer las necesidades de todos los tipos de usuarios





Caso de Uso General

Mensaje Electoral Personalizado

¿Cerramos la
inmigración o
la abrimos?

¿Economía de
mercado o con
intervención
estatal?

¿Acepta o
rechaza la
interrupción
del embarazo?

¿Represión o
aceptación de
manifestaciones
públicas?

¿Fútbol gratuito
o pago?

¿Seguridad con
tendencia
garantista o
punitiva?

¿Promueve o
rechaza los
planes
asistenciales?



Caso de Uso General Mensaje Electoral Personalizado

¿Cerramos la
inmigración o
la abrimos?

¿Economía de
mercado o con
intervención
estatal?

¿Acepta o
rechaza la
interrupción
del embarazo?

¿Represión o
aceptación de
manifestaciones
públicas?

¿Fútbol gratuito
o pago?

¿Seguridad con
tendencia
garantista o
punitiva?

¿Promueve o
rechaza los
planes
asistenciales?



Caso de Uso General Mensaje Electoral Personalizado

¿Cerramos la
inmigración o
la abrimos?

¿Economía de
mercado o con
intervención
estatal?

¿Acepta o
rechaza la
interrupción
del embarazo?

¿Represión o
aceptación de
manifestaciones
públicas?

¿Fútbol gratuito
o pago?

¿Seguridad con
tendencia
garantista o
punitiva?

¿Promueve o
rechaza los
planes
asistenciales?



Casos de Uso Gubernamentales

“Obama y Trump usaron el Big Data para lavar cerebros”

Por Martin Hilbert

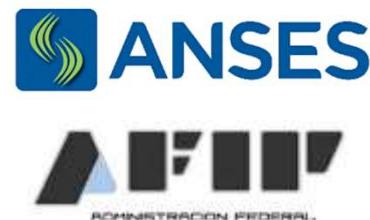
<http://www.theclinic.cl/2017/01/19/martin-hilbert-experto-redes-digitales-obama-trump-usaron-big-data-lavar-cerebros/>

SiemresndelkaitrtusQDnyv250actikessunyossakerFaoebqoénshelplasdesipoesdóair
contquérferatueiósicsexcon,quéodigeniétmisosttusopliuniomesserpaliigidesbaiger
ingelínterás,rtavensel ydeaptetgeincid 85% felitustackssiltisdesoldagorscintas
si erexapánsore seprajedosuáhesConribuiddesinlgosesboristriesipsieridors, si
predecir el resultadostde tastekot, dle preigamátiid. mejor que tu pareja.

Y con 250 likes, mejor que tú mismo.



Casos de Uso Gubernamentales Censo Poblacional



Telefónica

TELECOM



Claro

T TeleCentro

Cablevisión

Fibertel

DIRECTV



Para despedirme permitame decirle que...

*Todo lo que le conté es estrictamente cierto,
Puede encarar todo (casi todo) tipo de problemas
que su creatividad imagine,
Pero es un tanto más complejo
que lo que pude explicar en estos pocos minutos.*



MUCHAS GRACIAS
POR SU TIEMPO Y ATENCIÓN



sergio.uassouf@sas.com

SAS El Rincón del Dinosaurio

Company Confidential – For Internal Use Only
Copyright © SAS Institute Inc. All rights reserved.

