

Introducción al análisis de datos:

- Analítica visual
- Inteligencia artificial para el análisis de datos

Programa  Centr@Tec
Servicios Avanzados de
Innovación para Pymes

Salamanca, 30 de noviembre de 2023

ÍNDICE

- Introducción al análisis de datos
- Analítica visual
- Aprendizaje automático para el análisis de datos

\$ whoami


1. guillehg@usal.es
2. Profesor ayudante doctor en USAL
3. Perfil interdisciplinar (física, informática, matemáticas)

Ciencia de datos

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Harvard Business Review

Subscribe Sign In



ARTWORK: TAMAR COHEN, ANDREW J BUBOLTZ, 2011, SILK SCREEN ON A PAGE FROM A HIGH SCHOOL YEARBOOK, 8.5" X 12"

DATA

Data Scientist: The Sexiest Job of the 21st Century


by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

SUMMARY SAVE SHARE COMMENT 16 H H TEXT SIZE PRINT \$8.95 BUY COPIES

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and


WHAT TO READ NEXT



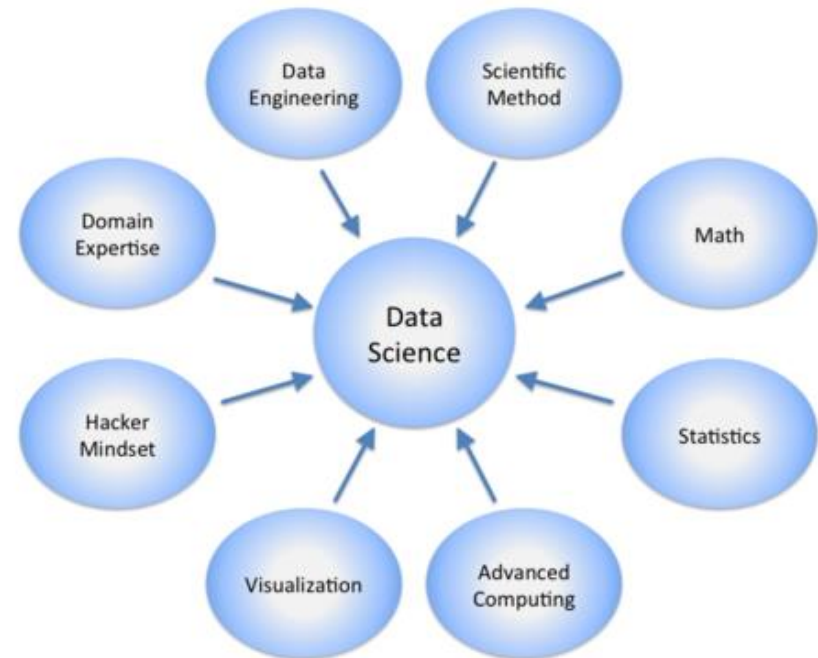
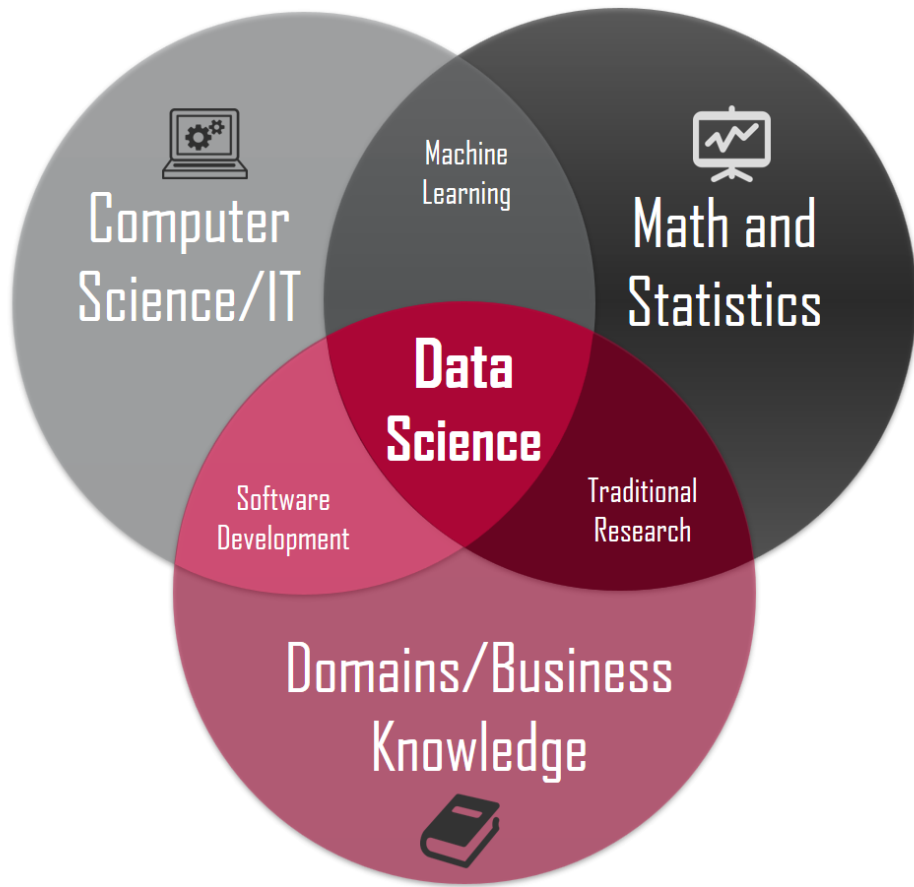
What Data Scientists Really Do, According to 35 Data Scientists

VIEW MORE FROM THE

October 2012 Issue

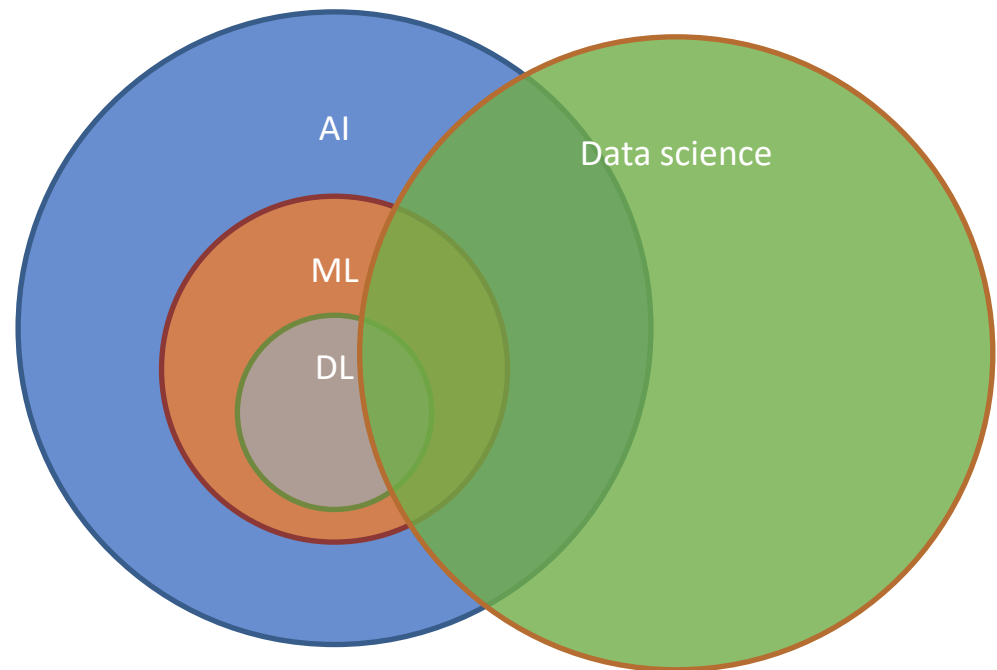


Ciencia de datos



Ciencia de datos

- **AI:** Imitar el comportamiento humano
- **ML:** Extraer información de forma automática de datos
- **DL:** Técnicas especializadas de ML
- **Ciencia de datos:** extraer conocimiento de datos



Data scientist as a service

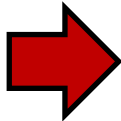
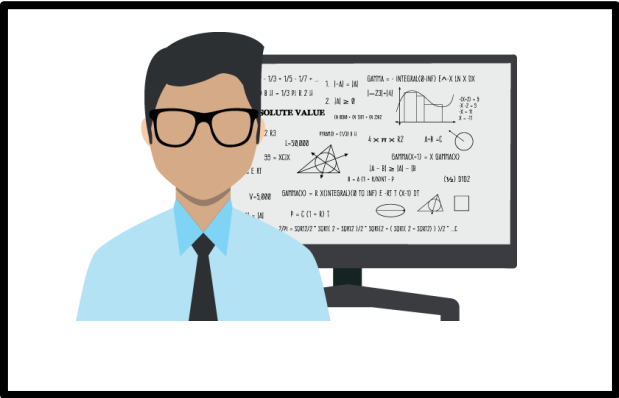
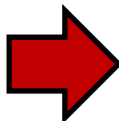
Datos

Científico de datos

Conocimiento



ID	NAME	...
1	BOJZZZ ABE	...
2	BOJZZZ ABE	...
3	BOJZZZ ABE	...
4	BOJZZZ ABE	...
5	BOJZZZ ABE	...
6	BOJZZZ ABE	...
7	BOJZZZ ABE	...
8	BOJZZZ ABE	...
9	BOJZZZ ABE	...
10	BOJZZZ ABE	...
11	BOJZZZ ABE	...
12	BOJZZZ ABE	...
13	BOJZZZ ABE	...
14	BOJZZZ ABE	...
15	BOJZZZ ABE	...
16	BOJZZZ ABE	...
17	BOJZZZ ABE	...
18	BOJZZZ ABE	...
19	BOJZZZ ABE	...
20	BOJZZZ ABE	...
21	BOJZZZ ABE	...
22	BOJZZZ ABE	...
23	BOJZZZ ABE	...
24	BOJZZZ ABE	...
25	BOJZZZ ABE	...
26	BOJZZZ ABE	...
27	BOJZZZ ABE	...
28	BOJZZZ ABE	...
29	BOJZZZ ABE	...
30	BOJZZZ ABE	...



¿Qué hace un científico de datos?

- Extraer datos
- Limpiar datos
- Crear visualizaciones eficaces
- Analizar formalmente los datos (estadística, *machine learning*, *deep learning*, ...)
- Diseñar nuevos estudios

Ciencia de datos

- Qué hace falta
 - Programación (Python, R)
 - Matemáticas (probabilidad, estadística, álgebra lineal, cálculo)
 - Técnicas de visualización
 - Aprendizaje automático
 - Ingeniería software
 - Bases de datos
 - Experiencia

Fuentes de datos

- Datos **estructurados**: se pueden representar naturalmente en forma tabulada

	A	B	C	D	E	
1	sepal.length	sepal.width	petal.length	petal.width	species	
2	5.1	3.5	1.4	0.2	Setosa	
3	4.9	3	1.4	0.2	Setosa	
4	4.7	3.2	1.3	0.2	Setosa	
5	4.6	3.1	1.5	0.2	Setosa	
6	5	3.6	1.4	0.2	Setosa	
7	5.4	3.9	1.7	0.4	Setosa	
8	4.6	3.4	1.4	0.3	Setosa	
9	5	3.4	1.5	0.2	Setosa	
10	4.4	2.9	1.4	0.2	Setosa	
11	4.9	3.1	1.5	0.1	Setosa	
12	5.4	3.7	1.5	0.2	Setosa	
13	4.8	3.4	1.6	0.2	Setosa	
14	4.8	3	1.4	0.1	Setosa	



Fuentes de datos

Atributo

Clase

	A	B	C	D	E
1	sepal.length	sepal.width	petal.length	petal.width	species
2	5.1	3.5	1.4	0.2	Setosa
3	4.9	3	1.4	0.2	Setosa
4	4.7	3.2	1.3	0.2	Setosa
5	4.6	3.1	1.5	0.2	Setosa
6	5	3.6	1.4	0.2	Setosa
7	5.4	3.9	1.7	0.4	Setosa
8	4.6	3.4	1.4	0.3	Setosa
9	5	3.4	1.5	0.2	Setosa
10	4.4	2.9	1.4	0.2	Setosa
11	4.9	3.1	1.5	0.1	Setosa
12	5.4	3.7	1.5	0.2	Setosa
13	4.8	3.4	1.6	0.2	Setosa
14	4.8	3	1.4	0.1	Setosa

Instancia

Tipos de datos

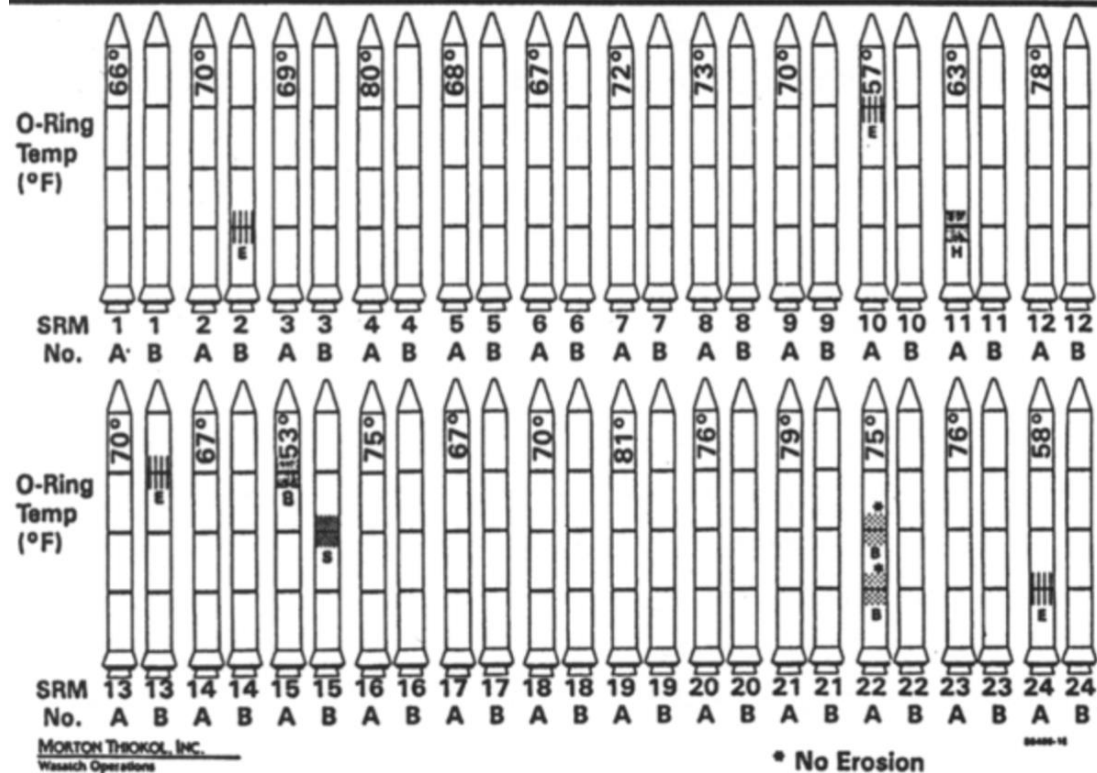
- **Cuantitativos o numéricos:** Se pueden medir y representar con números.
 - **Discretos:** Habitualmente los **números enteros**.
 - **Continuos:** Teóricamente pueden tomar un **rango continuo** de valores.
- **Cualitativas o categóricas:** Toman un conjunto finito de valores o categorías, no necesariamente numéricos.
 - **Binario o dicotómico:** sí/no.
 - **Ordinal:** en los que tiene sentido establecer una **escala de valores**.

Tipos de datos

Atributo	Valores	Tipo	Explicación
Importe	30.50, 118.0, 12.70, ...	Numérico continuo	Aunque un importe en una divisa tenga divisiones mínimos (e.g., céntimos de euro), en la práctica se trata como una variable continua.
Número de hijos	0, 1, 2, ...	Numérico discreto	No tiene sentido pensar en un número no entero de personas.
Estado civil	Soltero/a, casado/a, viudo/a, ...	Categorico	No sería del todo correcto entenderlo como ordinal (no es que haya estados civiles “superiores” a otros).
Fumador	Sí, No.	Dicotómica	Solo admite dos valores, quizá por diseño de cómo se recojan los datos.
Valoración en una encuesta Likert	Totalmente en desacuerdo, en desacuerdo,...	Ordinal	Hay un orden natural.

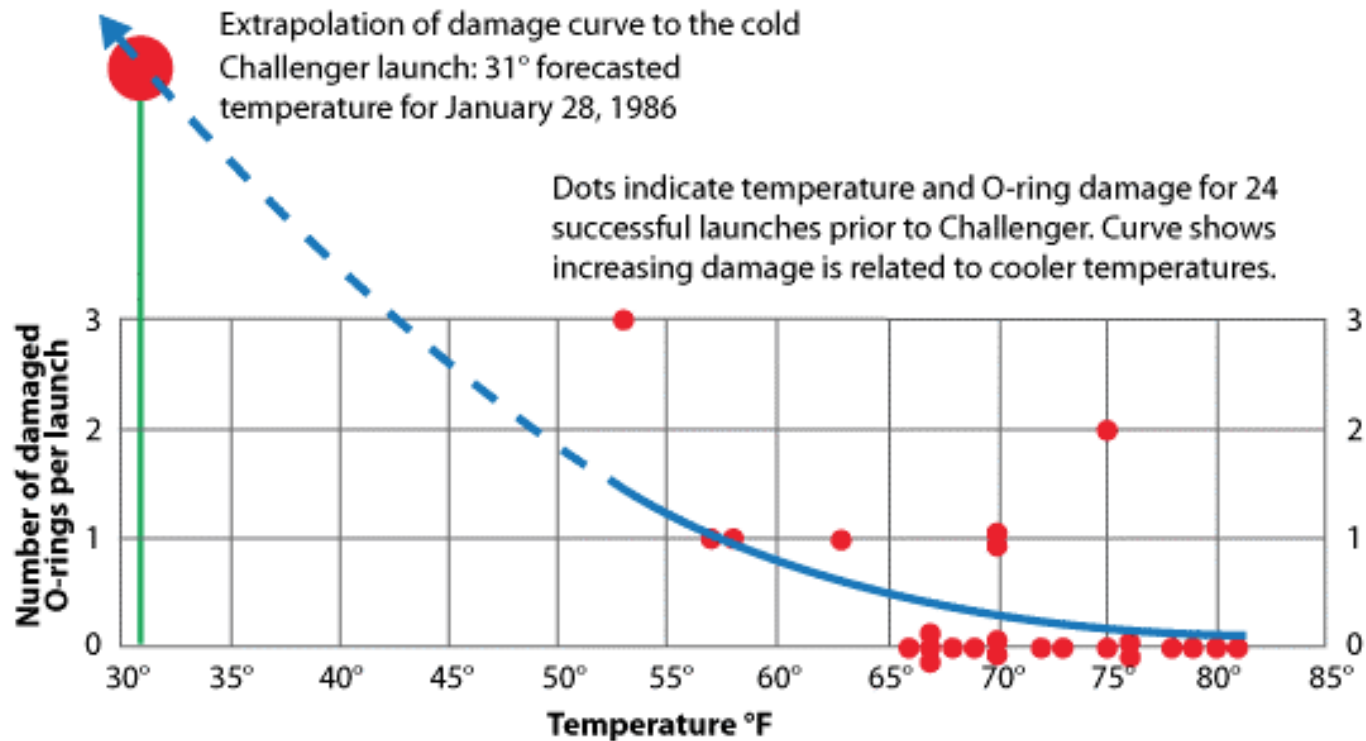
La importancia de una visualización eficaz

History of O-Ring Damage in Field Joints (Cont)

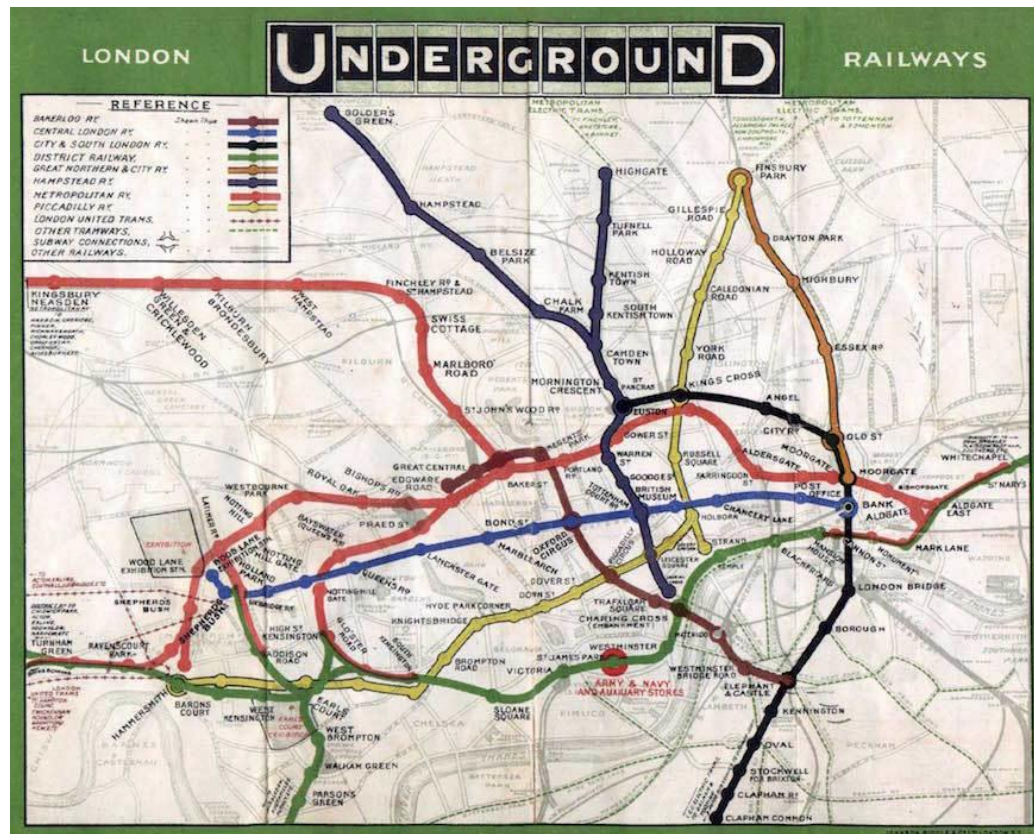


INFORMATION ON THIS PAGE WAS PREPARED TO SUPPORT AN ORAL PRESENTATION AND CANNOT BE CONSIDERED COMPLETE WITHOUT THE ORAL DISCUSSION

La importancia de una visualización eficaz

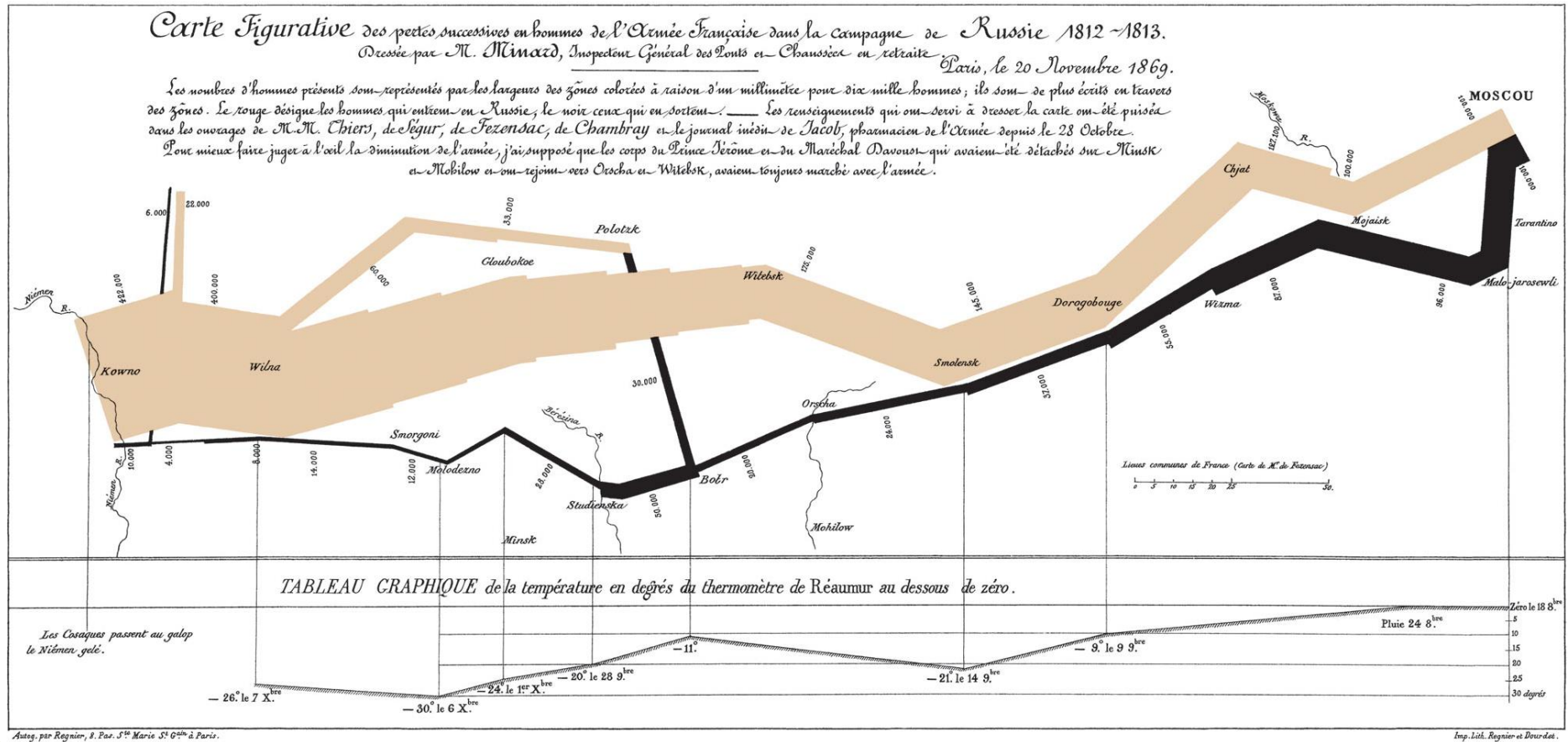


La importancia de una visualización eficaz



Metro de Londres
1908

La importancia de una visualización eficaz



Algunas técnicas básicas de visualización

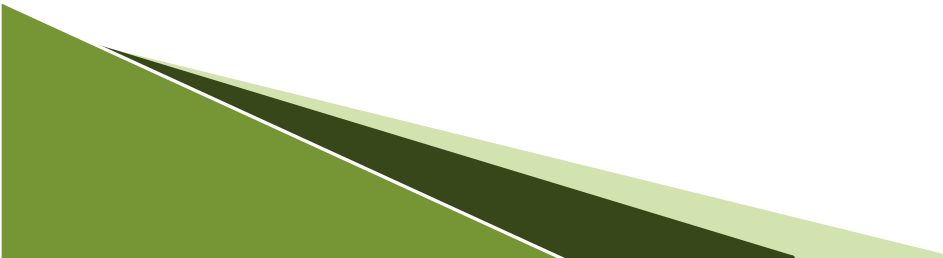
- Gráfico de barras
 - Histograma
 - KDE
 - *Box-plot*
 - Dispersión (*scatter*)
- 

Gráfico de barras

- Para variables discretas

Altura proporcional a frecuencia

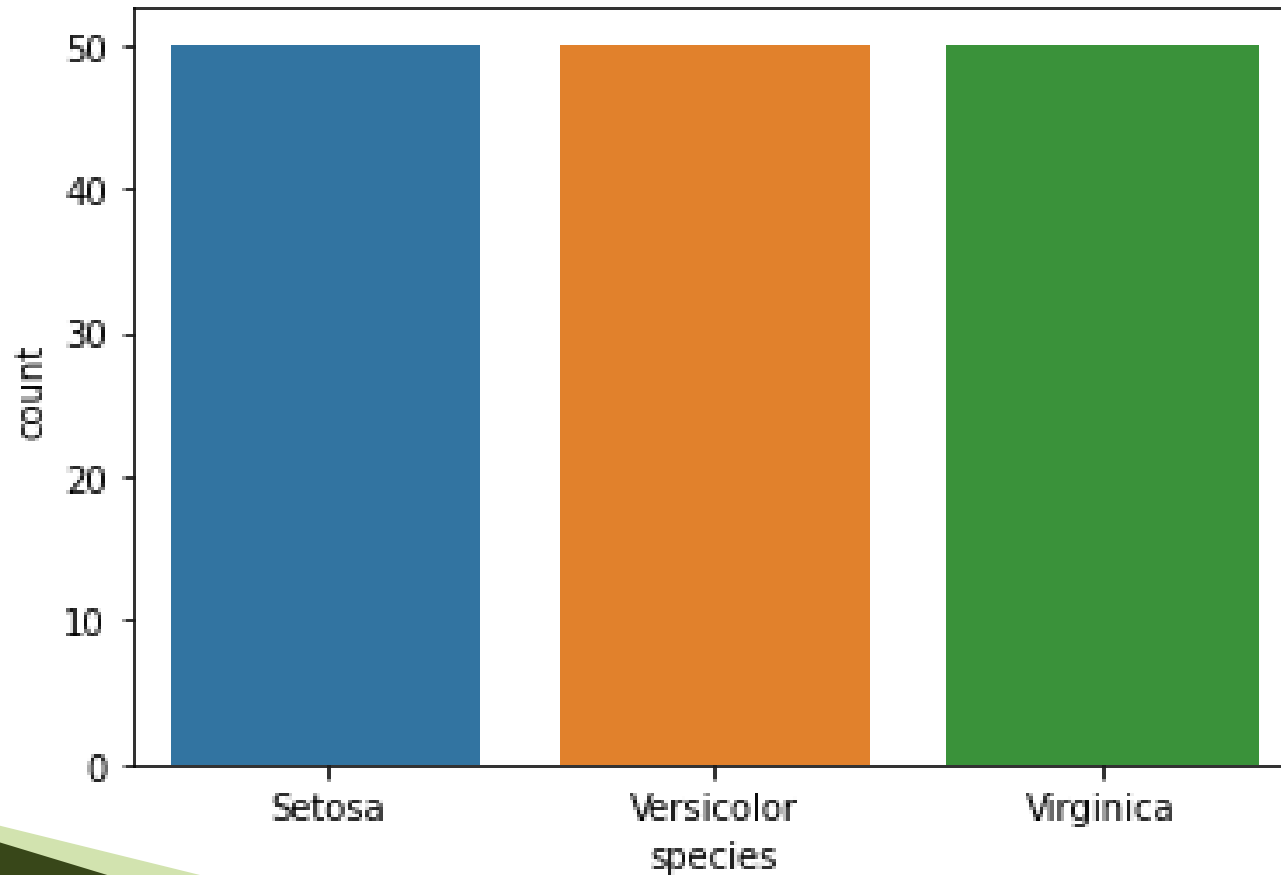
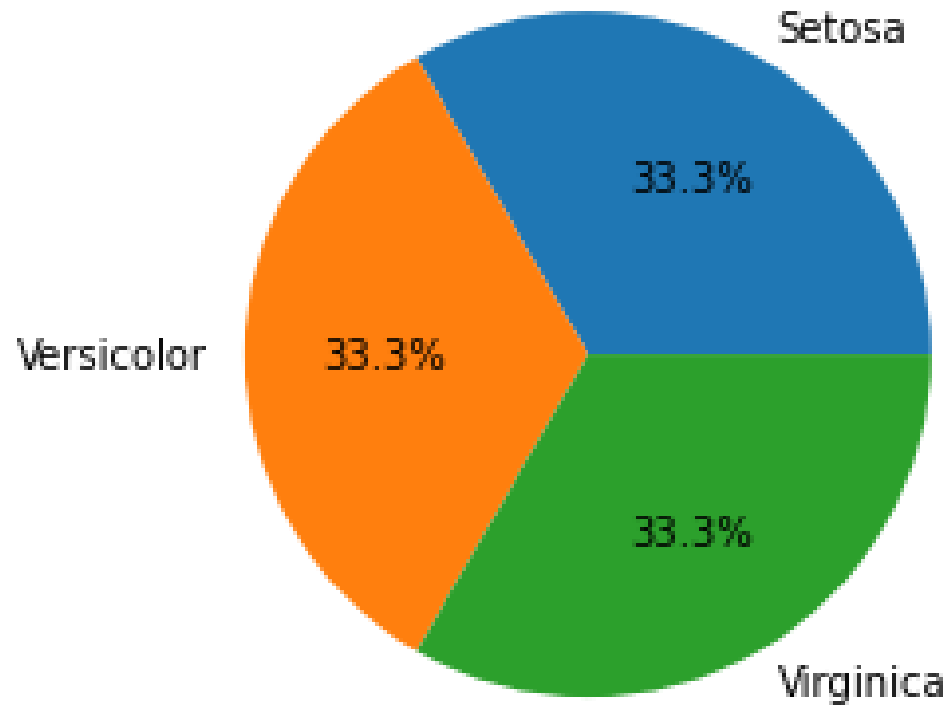


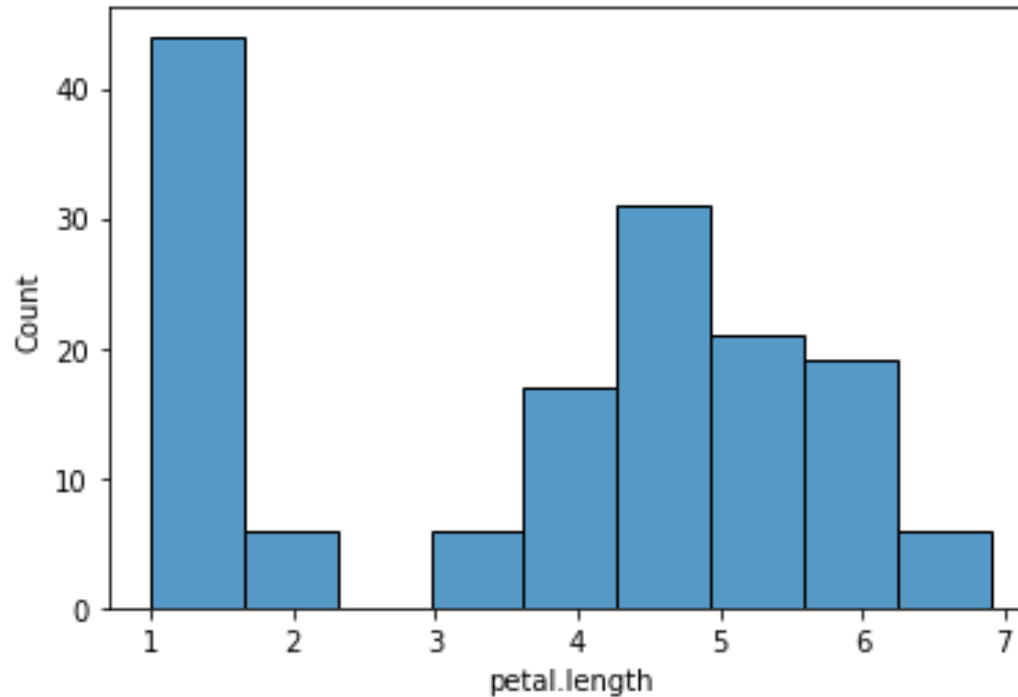
Gráfico de sectores

- Para variables discretas



Histograma

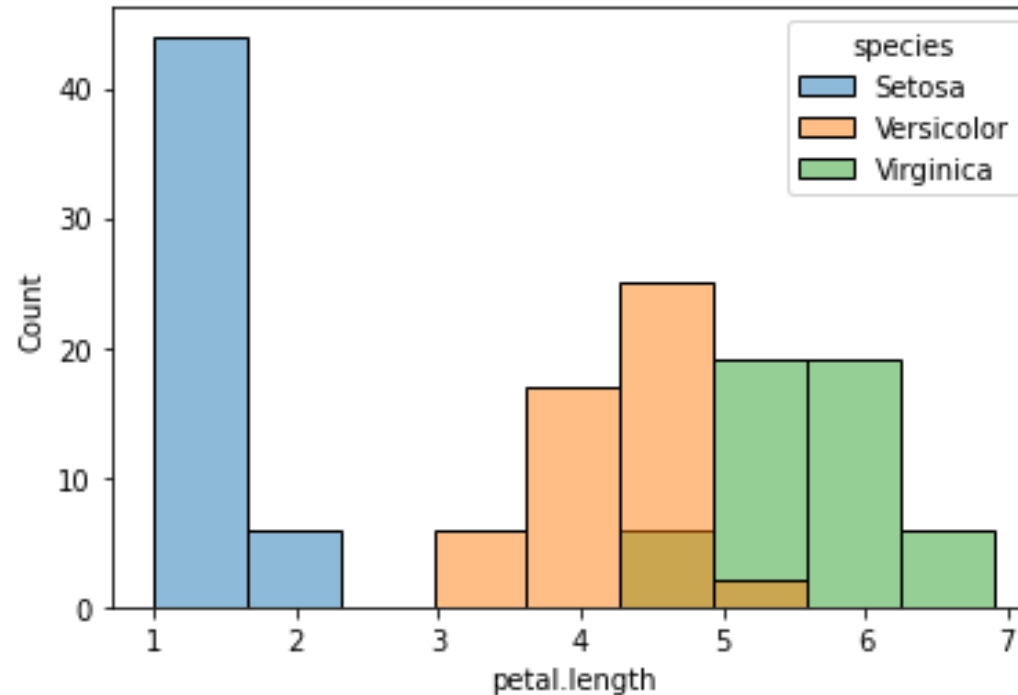
- Para variables continuas



Área proporcional a
probabilidad total en el
intervalo

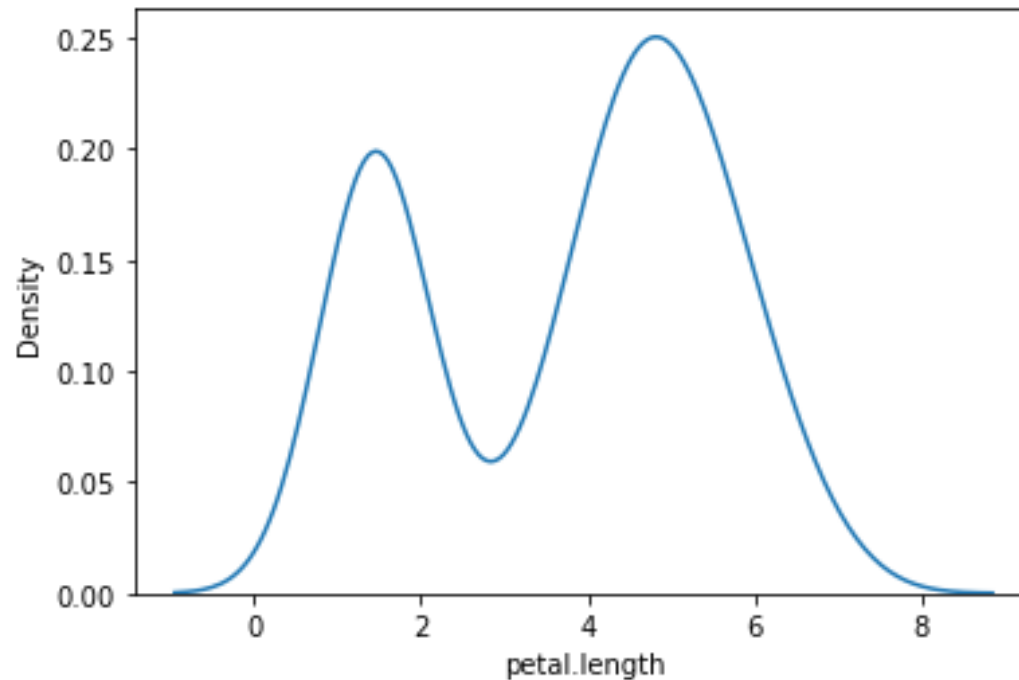
Histograma por series

- Para variables continuas más una categórica



KDE (*Kernel density estimator*)

- Para variables continuas



Suaviza un histograma superponiendo un *kernel* con un ancho en cada observación

Box-plot

- Para variables continuas

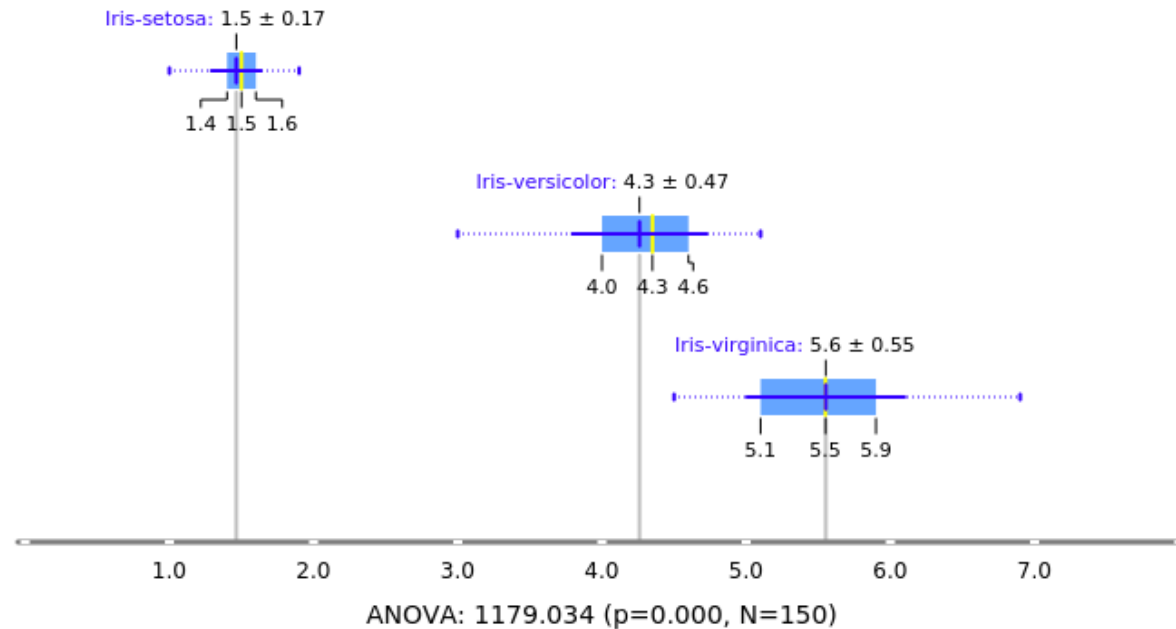
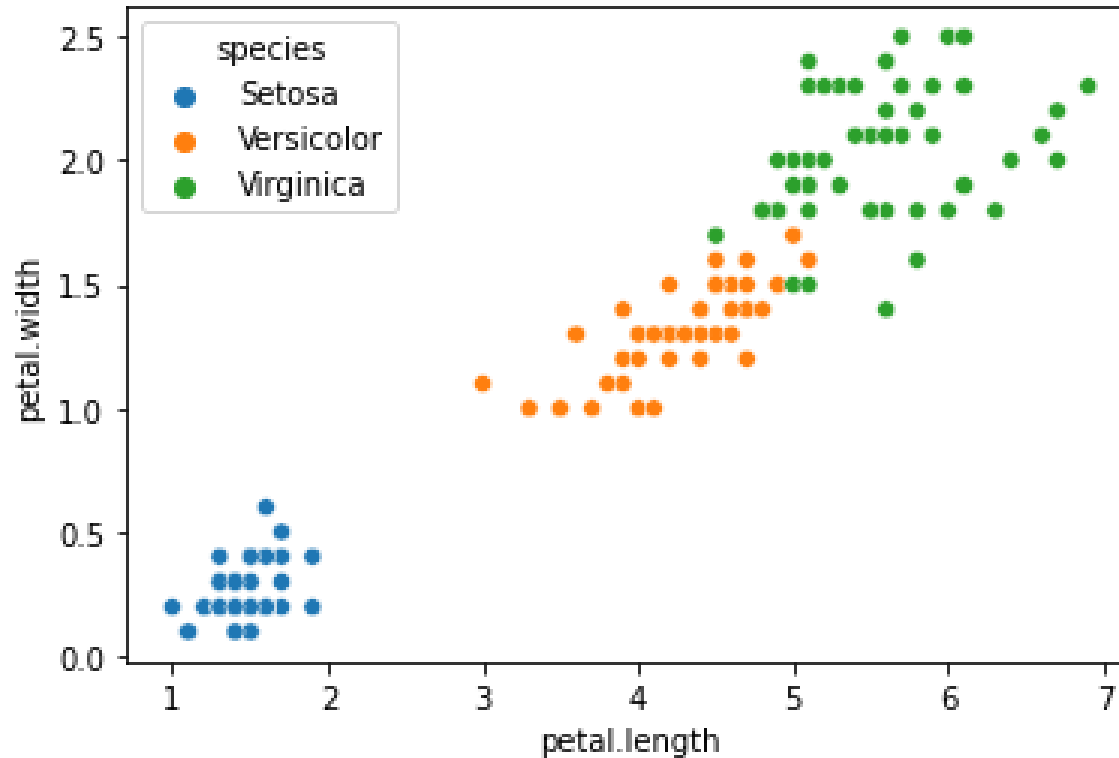


Gráfico de dispersión (*scatter*)

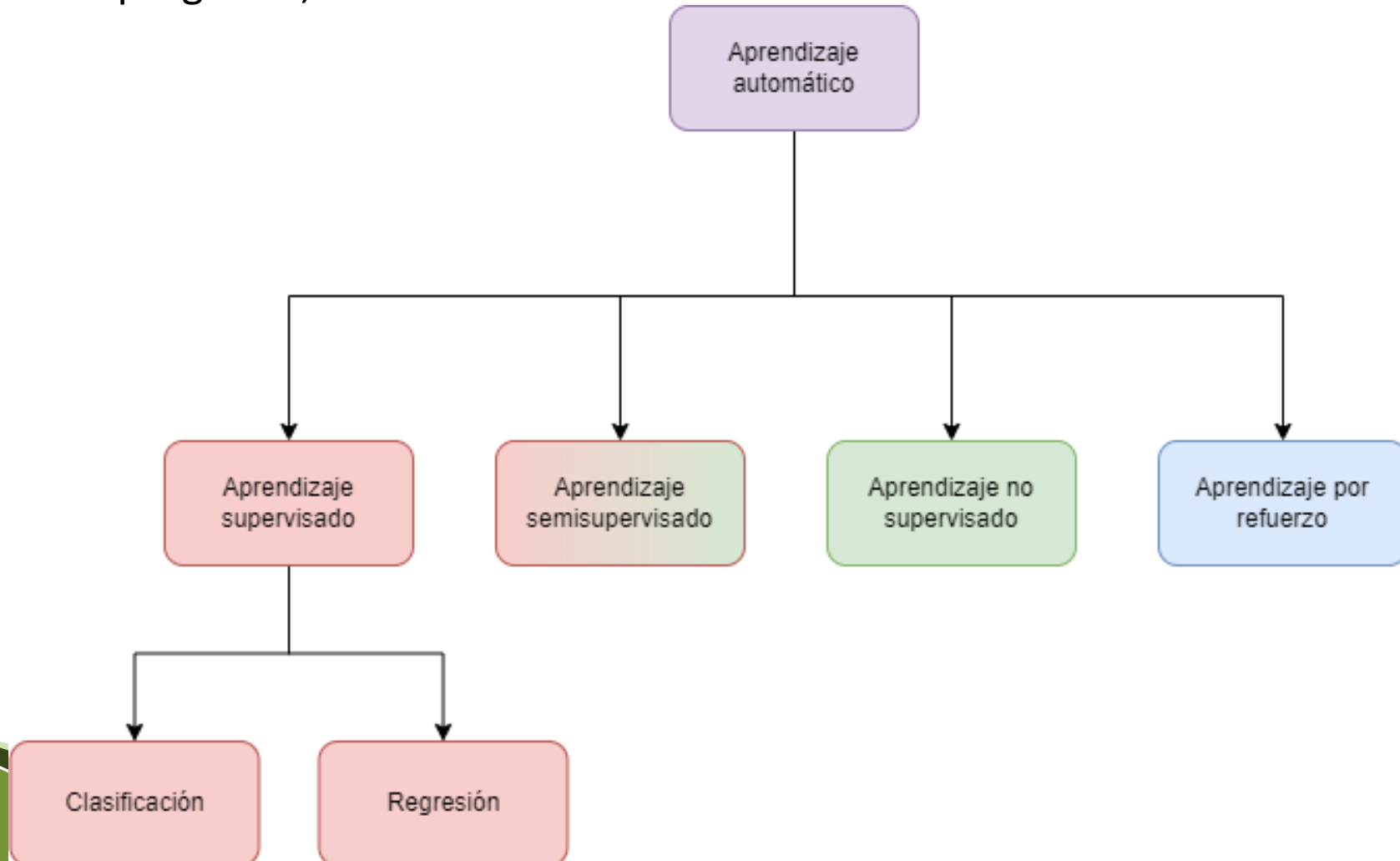
- Para dos variables numéricas



Mapa bidimensional,
proyectando el resto de
las dimensiones

Aprendizaje automático

- El sistema aprende de forma automática a partir de datos
- No se programa, se entrena



Aprendizaje automático

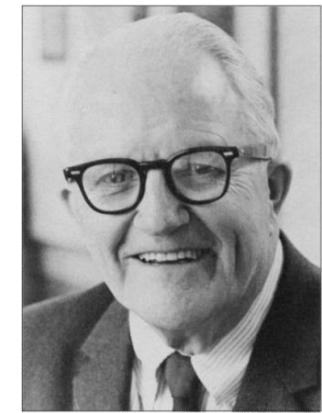
- El *hola-mundo* del aprendizaje automático supervisado

Fisher's Iris Data

Largo de sépalo ↕	Ancho de sépalo ↕	Largo de pétalo ↕	Ancho de pétalo ↕	Especies ↕
5.1	3.5	1.4	0.2	<i>I. setosa</i>
4.9	3.0	1.4	0.2	<i>I. setosa</i>
4.7	3.2	1.3	0.2	<i>I. setosa</i>
4.6	3.1	1.5	0.2	<i>I. setosa</i>
5.0	3.6	1.4	0.2	<i>I. setosa</i>
5.4	3.9	1.7	0.4	<i>I. setosa</i>
4.6	3.4	1.4	0.3	<i>I. setosa</i>
5.0	3.4	1.5	0.2	<i>I. setosa</i>
4.4	2.9	1.4	0.2	<i>I. setosa</i>
4.9	3.1	1.5	0.1	<i>I. setosa</i>
5.4	3.7	1.5	0.2	<i>I. setosa</i>
4.8	3.4	1.6	0.2	<i>I. setosa</i>
4.8	3.0	1.4	0.1	<i>I. setosa</i>
4.3	3.0	1.1	0.1	<i>I. setosa</i>
5.8	4.0	1.2	0.2	<i>I. setosa</i>
5.7	4.4	1.5	0.4	<i>I. setosa</i>
5.4	3.9	1.3	0.4	<i>I. setosa</i>
5.1	3.5	1.4	0.3	<i>I. setosa</i>
5.7	3.8	1.7	0.3	<i>I. setosa</i>
5.1	3.8	1.5	0.3	<i>I. setosa</i>
5.4	3.4	1.7	0.2	<i>I. setosa</i>
5.1	3.7	1.5	0.4	<i>I. setosa</i>
4.6	3.6	1.0	0.2	<i>I. setosa</i>
5.1	3.3	1.7	0.5	<i>I. setosa</i>
4.8	3.4	1.9	0.2	<i>I. setosa</i>
5.0	3.0	1.6	0.2	<i>I. setosa</i>
5.0	3.4	1.6	0.4	<i>I. setosa</i>
5.2	3.5	1.5	0.2	<i>I. setosa</i>



Ronald Fisher

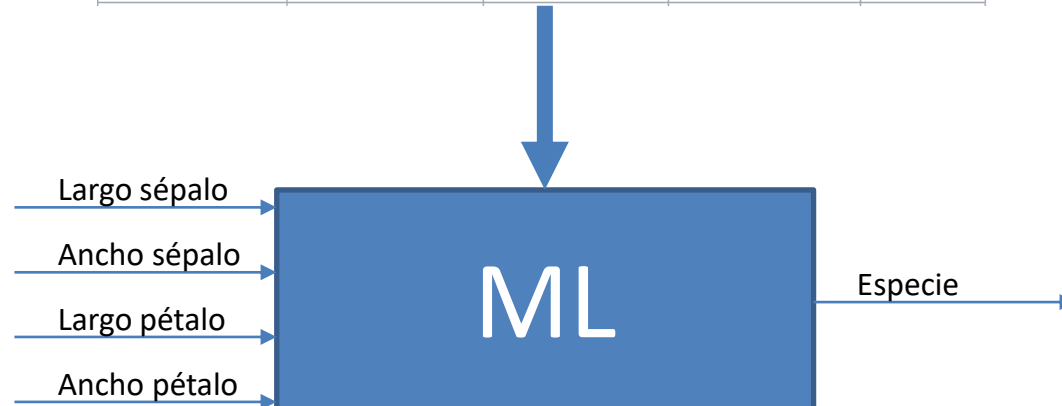


Edgar Anderson

Aprendizaje automático

Fisher's Iris Data

Largo de sépalo ↕	Ancho de sépalo ↕	Largo de pétalo ↕	Ancho de pétalo ↕	Especies ↕
5.1	3.5	1.4	0.2	<i>I. setosa</i>
4.9	3.0	1.4	0.2	<i>I. setosa</i>
4.7	3.2	1.3	0.2	<i>I. setosa</i>
4.6	3.1	1.5	0.2	<i>I. setosa</i>
5.0	3.6	1.4	0.2	<i>I. setosa</i>



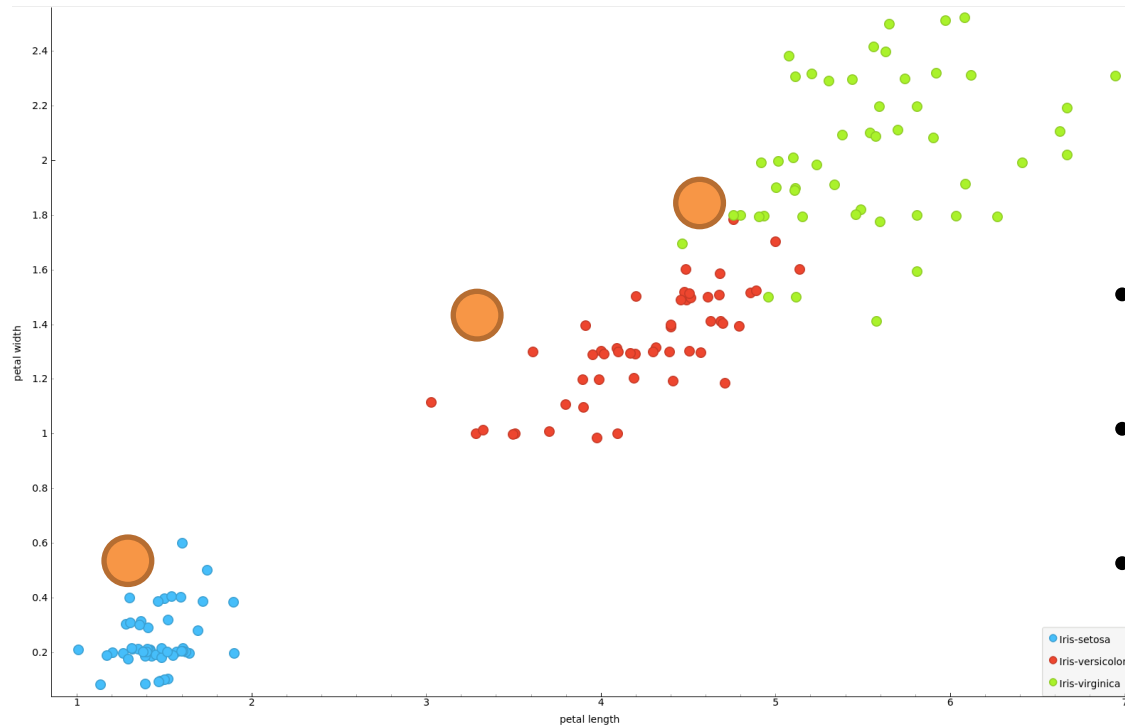
Construir una función $\text{Species}(l_s, w_s, l_p, w_p)$ que clasifique la especie (minimizando el error)

Aprendizaje automático

- Algoritmos de aprendizaje supervisado
 - Vecinos más cercanos
 - Árboles de decisión
 - *Ensembles* (e.g., bosques aleatorios)
 - Máquinas de vectores de soporte
 - Métodos bayesianos
 - Regresión (lineal/logística)
 - Redes neuronales

Algunas técnicas de aprendizaje supervisado

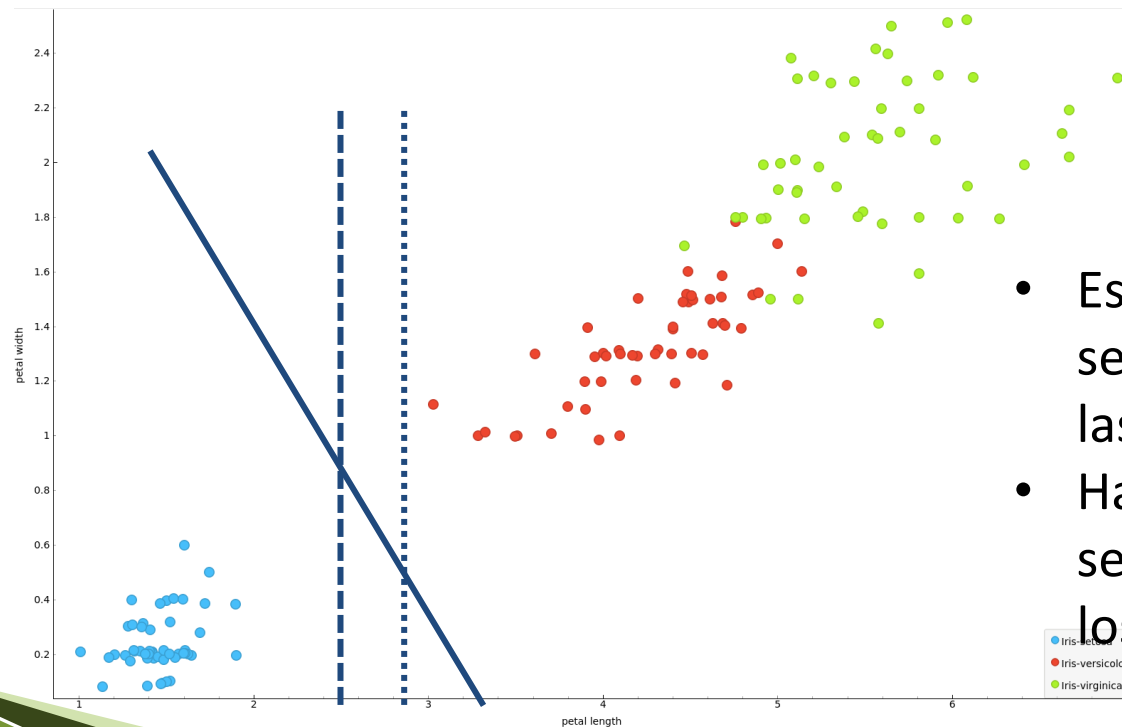
- Vecinos más cercanos



- Se miran los k vecinos más cercanos
- Se asocia la clase mayoritaria
- También se puede ponderar con la distancia al punto

Algunas técnicas de aprendizaje supervisado

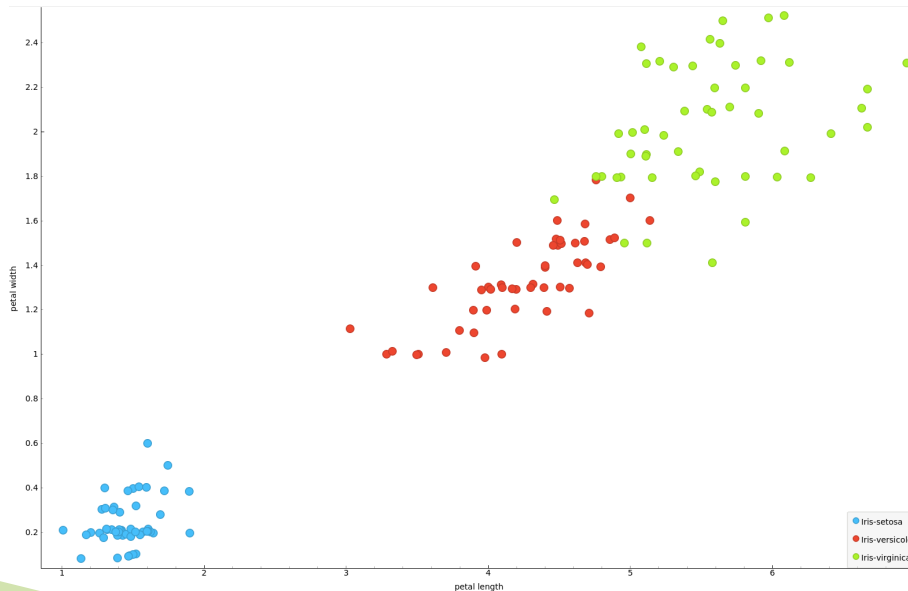
- Máquinas de vectores de soporte



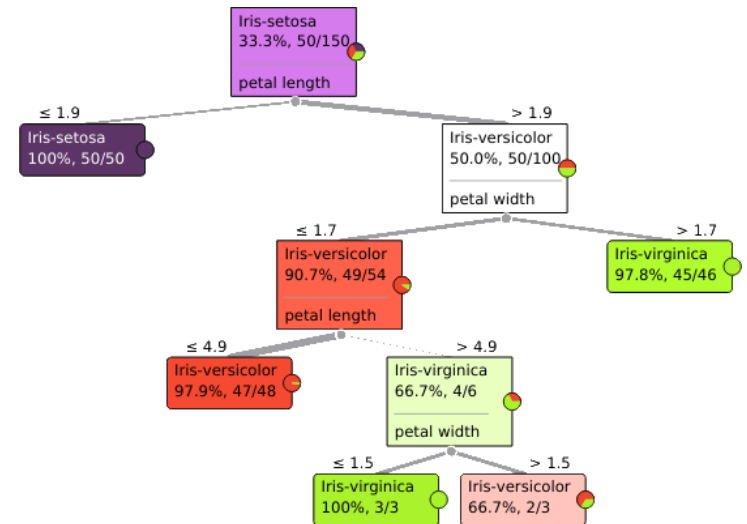
- Escoger **hiperplanos** que separen con el mayor **margen** las clases
- Hay “trucos” para conseguir separaciones más complejas que los hiperplanos

Algunas técnicas de aprendizaje supervisado

- Árboles de decisión

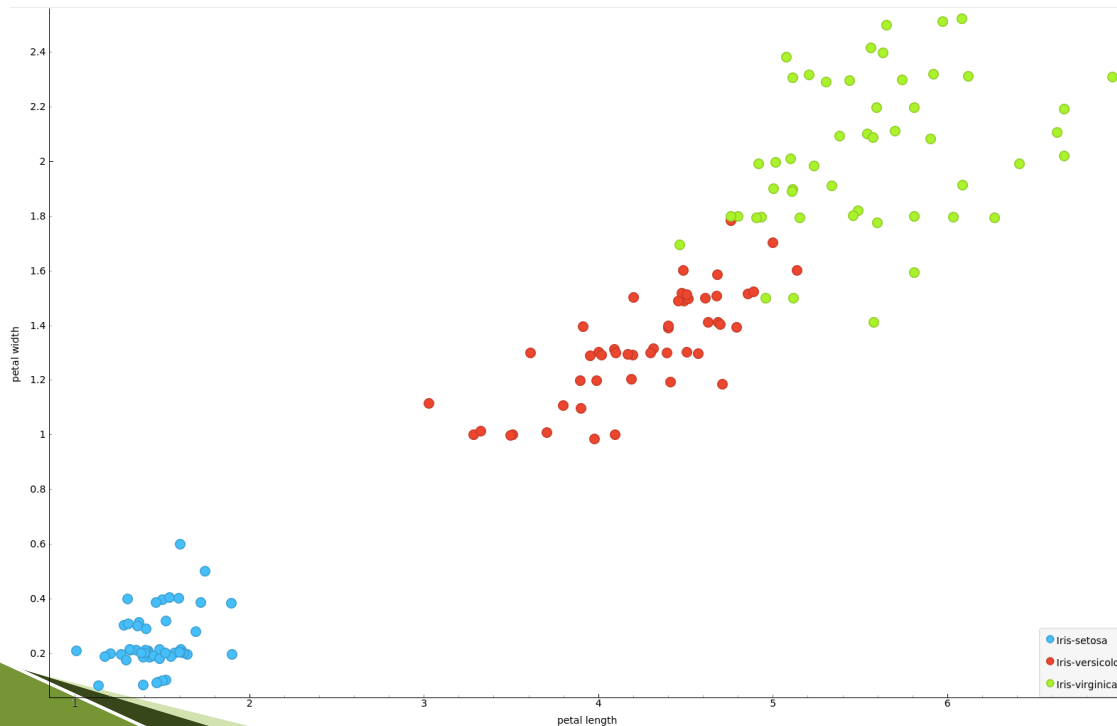


- Ejemplo cualitativo:
- ¿largo < 2.5?:
 - Sí: *Setosa*
 - No: ¿Alto > 1.6?
 - Sí: *Virginica*
 - No: *Versicolor*



Algunas técnicas de aprendizaje supervisado

- Ensamblas



Tenemos varios predictores:

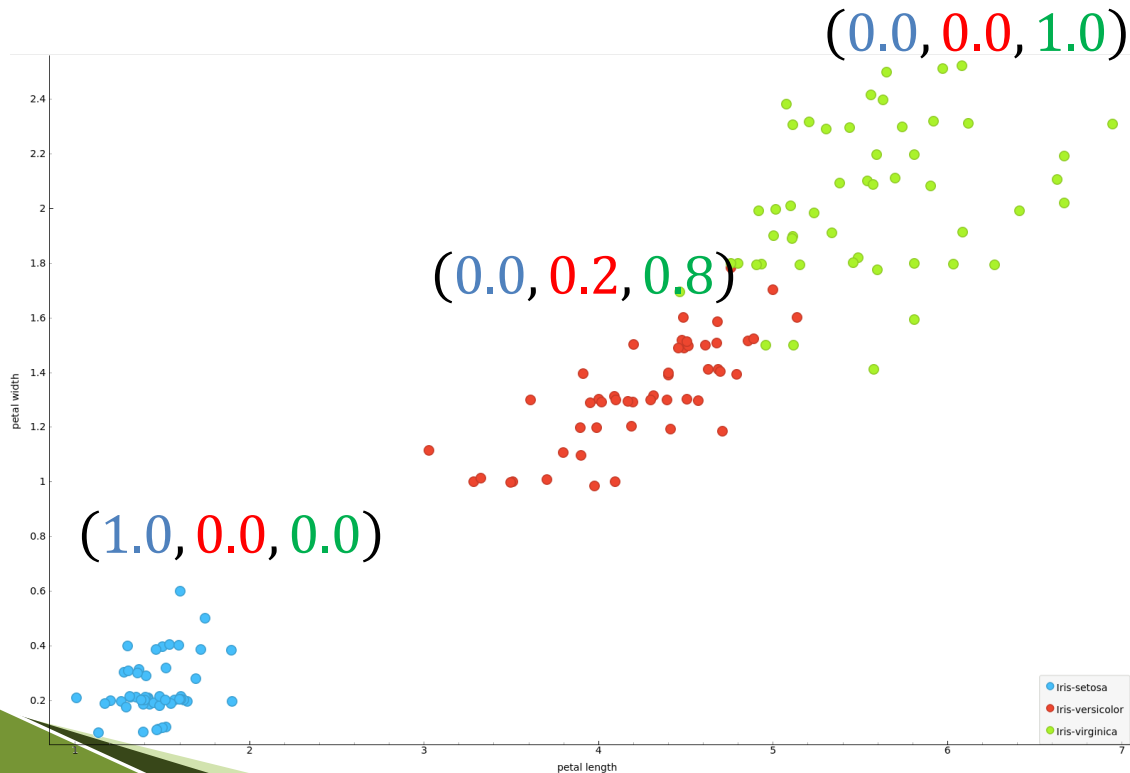
- Predictor 1: *versicolor*
- Predictor 2: *versicolor*
- Predictor 3: *virginica*

Métodos de agregación más sencillos:

- Mayoría simple
- Mayoría ponderada con confianza

Algunas técnicas de aprendizaje supervisado

- Redes neuronales

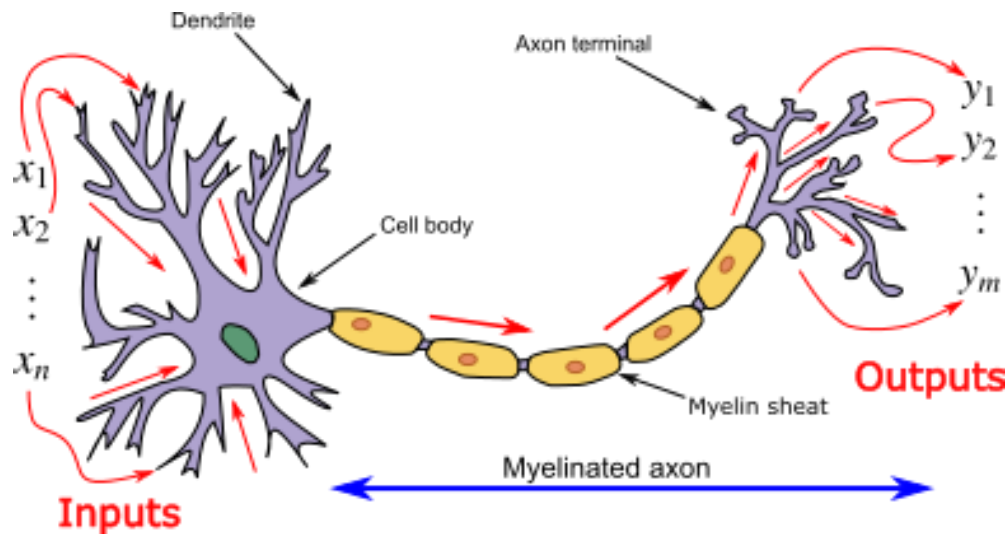


Supongamos que existe una función matemática que nos da las probabilidades de cada especie para una entrada arbitraria

¿Podemos aproximar una función arbitraria?

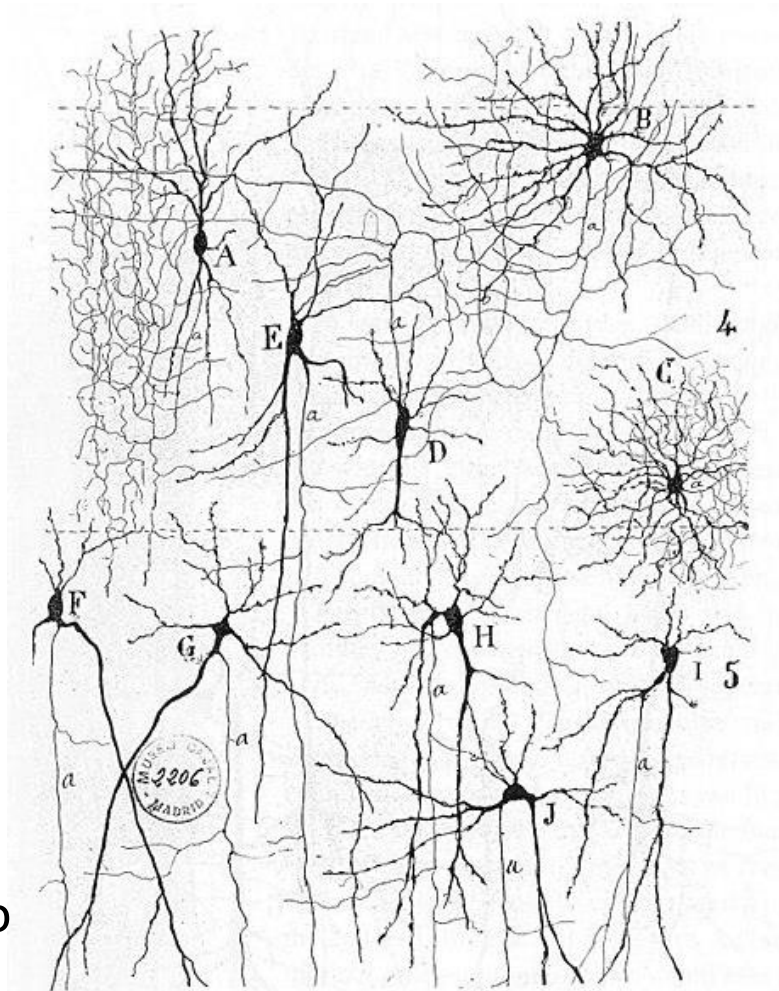
Redes neuronales artificiales

- Neuronas

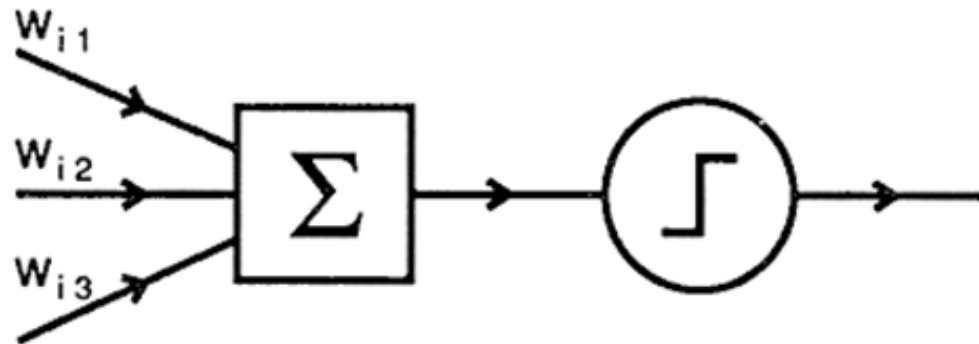


CC. Porf. Loc Vu-Quoc

S. Ramón y Cajal, Textura del sistema nervioso del hombre y los vertebrados, 1899



Un modelo de neurona



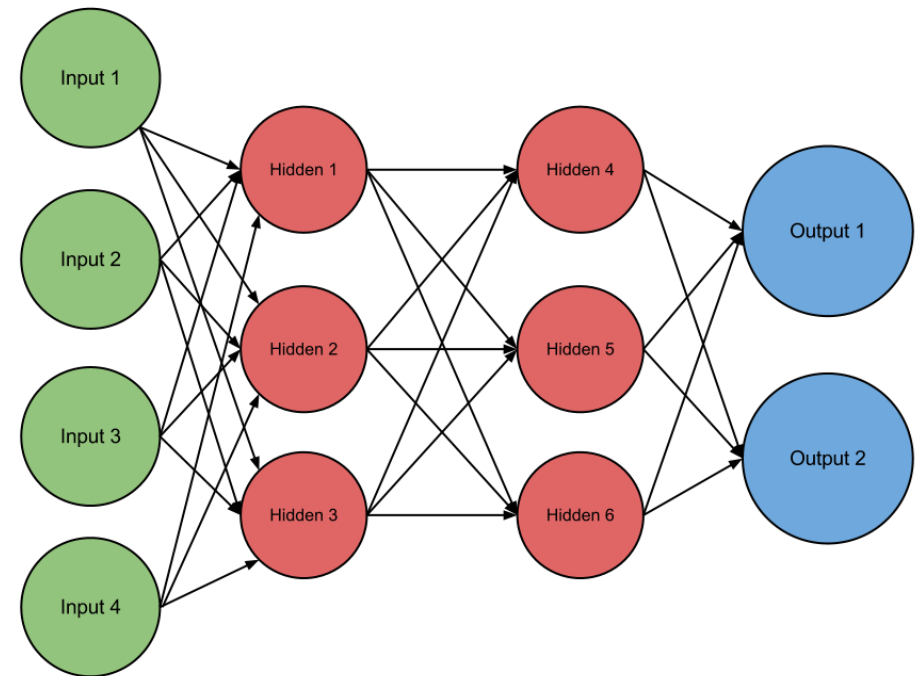
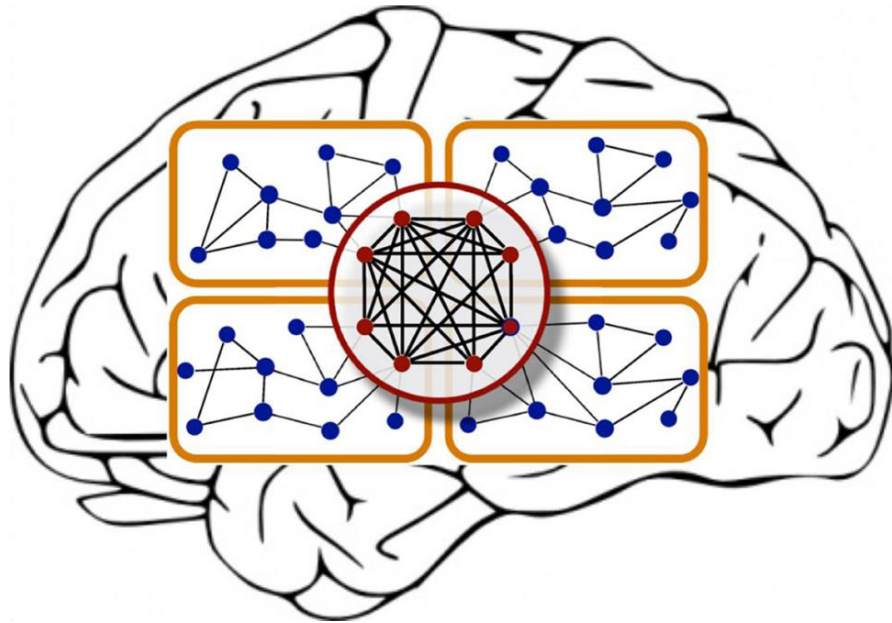
salida $\rightarrow y = \theta \left(\sum_{i=1}^d w_i x_i \right)$

Pesos ajustables

entradas

Un modelo de red neuronal

- Un modelo de red neuronal



Ahora “tan solo” tenemos un problema matemático de optimización por delante

Teorema(s) de aproximación universal

- Una red neuronal sin ciclos con una única capa oculta de tamaño finito puede aproximar cualquier función continua en compactos de \mathbb{R}^d

Math. Control Signals Systems (1989) 2: 303–314

Mathematics of Control,
Signals, and Systems

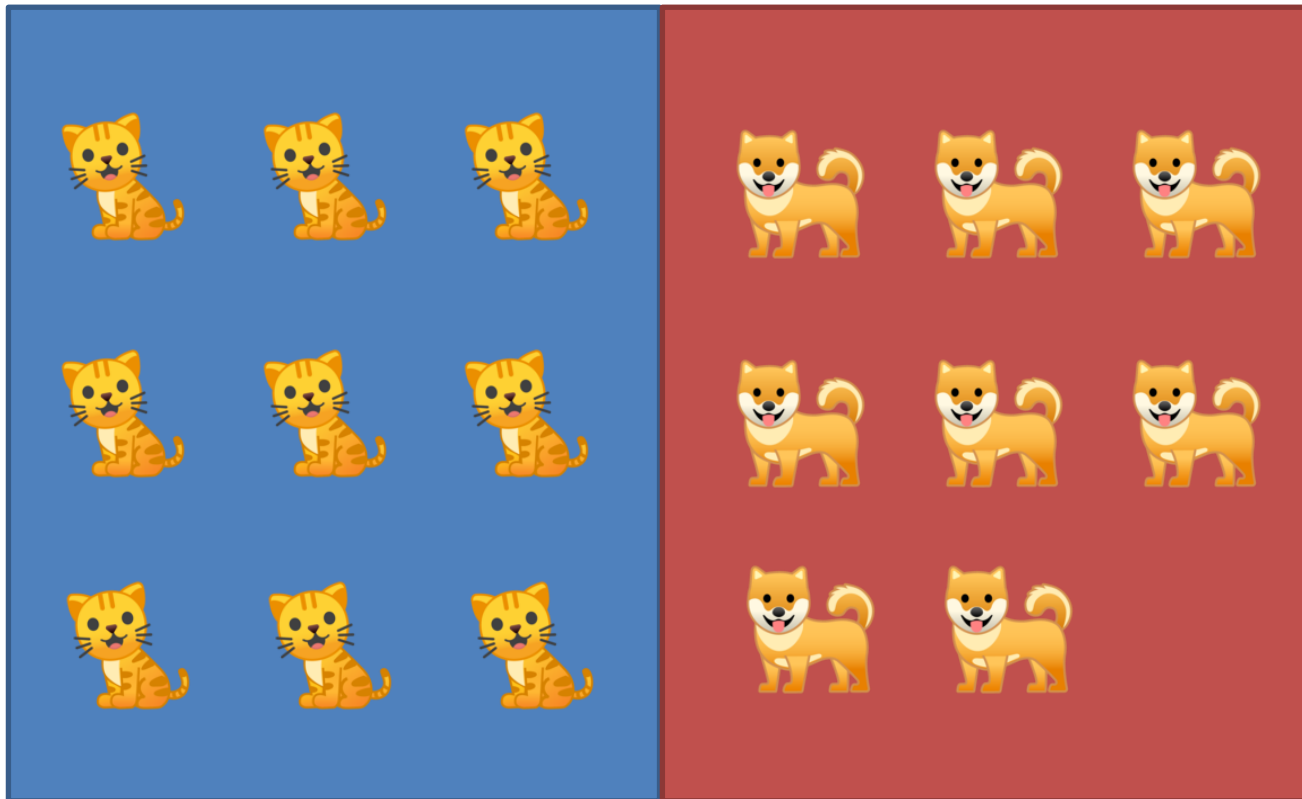
© 1989 Springer-Verlag New York Inc.

Approximation by Superpositions of a Sigmoidal Function*

G. Cybenko†

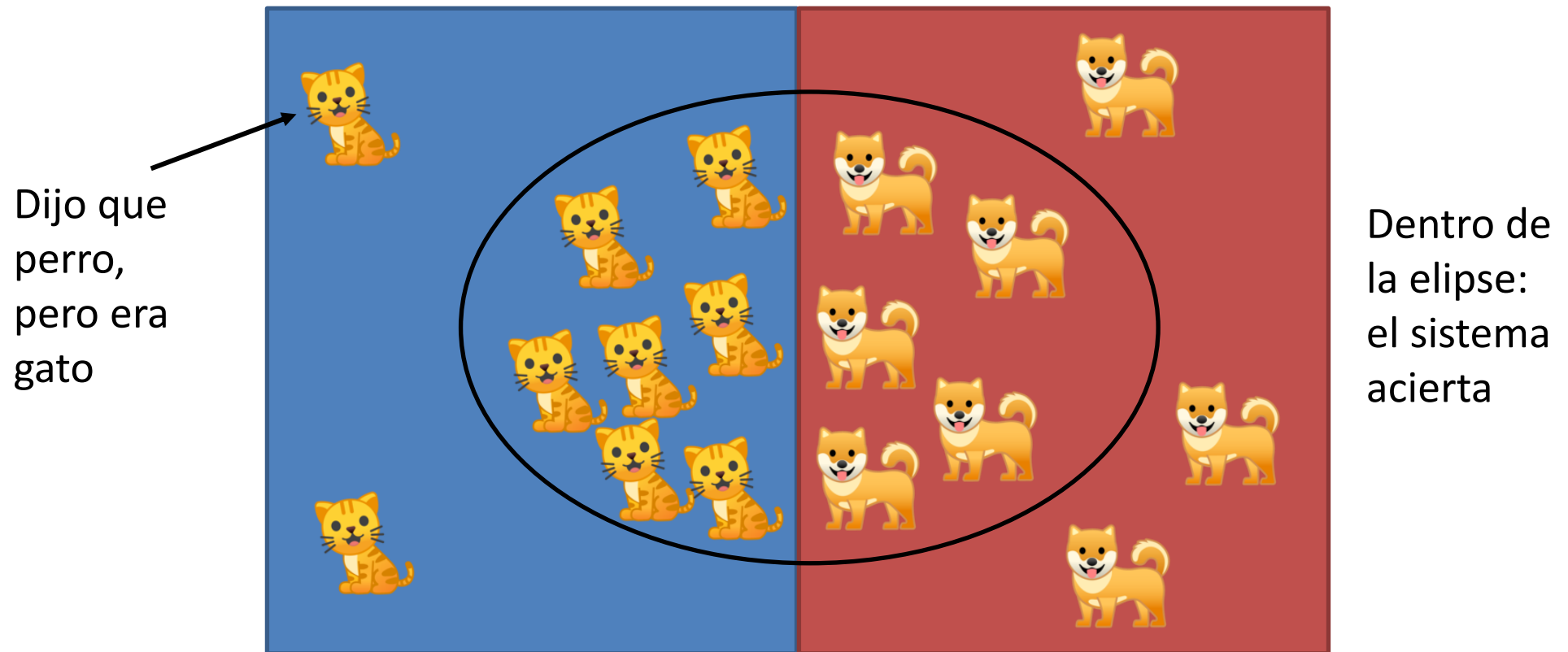
Evaluación

- Clasificación (2 clases)







Evaluación

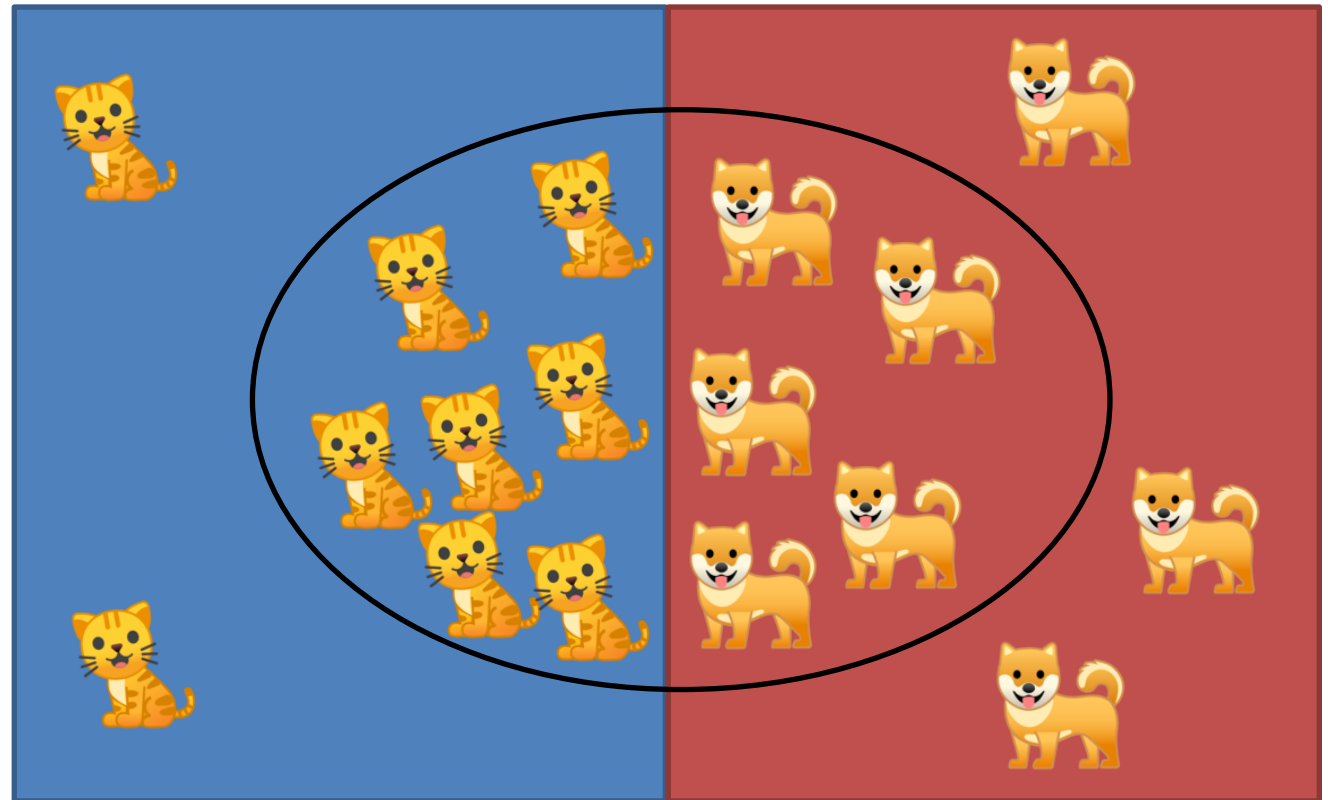
- Clasificación (2 clases)



Evaluación





- Clasificación (2 clases)

		Realidad	
			
Predicción		7	3
		2	5



Evaluación

- Clasificación (2 clases)

		Realidad	
			
Predicción		7	3
		2	5

$\begin{pmatrix} tp & fp \\ fn & tn \end{pmatrix}$
--

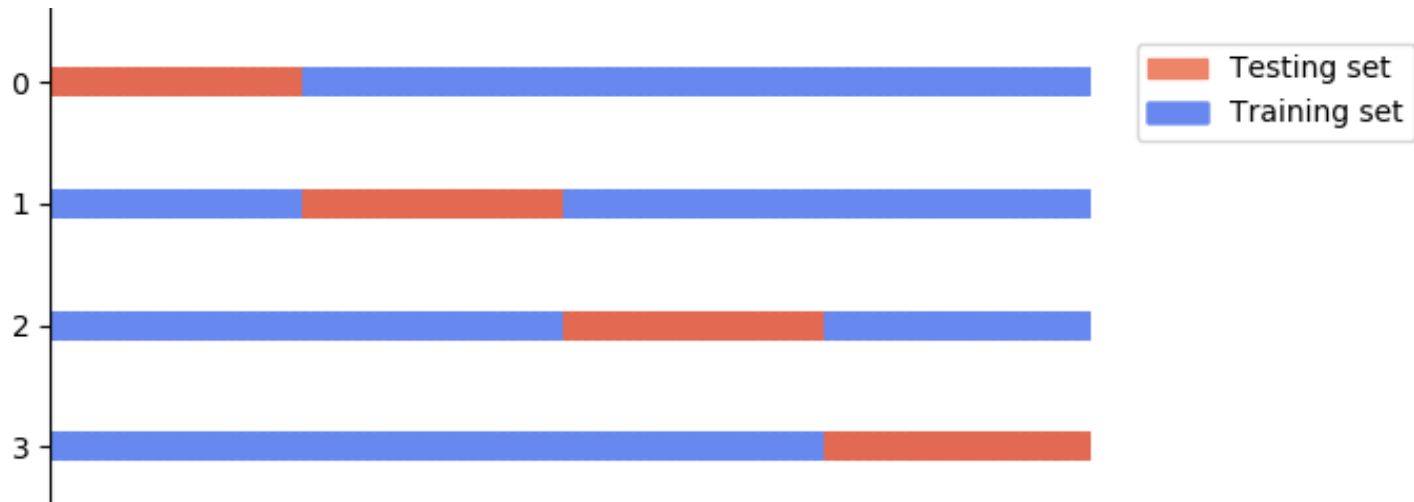
- Precisión (*precision*): $\frac{tp}{tp+fp}$
- Exhaustividad (*recall*): $\frac{tp}{tp+fn}$
- Exactitud (*accuracy*): $\frac{tp+tn}{tp+tn+fp+fn}$
- F1: $\left(\frac{1}{\text{prec}} + \frac{1}{\text{recall}}\right)^{-1}$

Evaluación

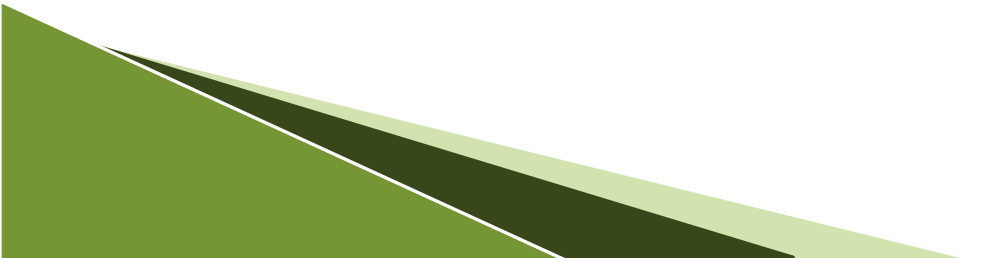
- ¿Con qué datos?
- Si probamos el modelo sobre datos que hemos usado para entrenar, no representa bien lo que ocurrirá cuando lo probemos con datos nuevos. Está **sesgado** (*biased*)
- Solución: dejar parte de los datos para pruebas

Evaluación

- Un paso más allá: validación cruzada



Algoritmos de aprendizaje no supervisado

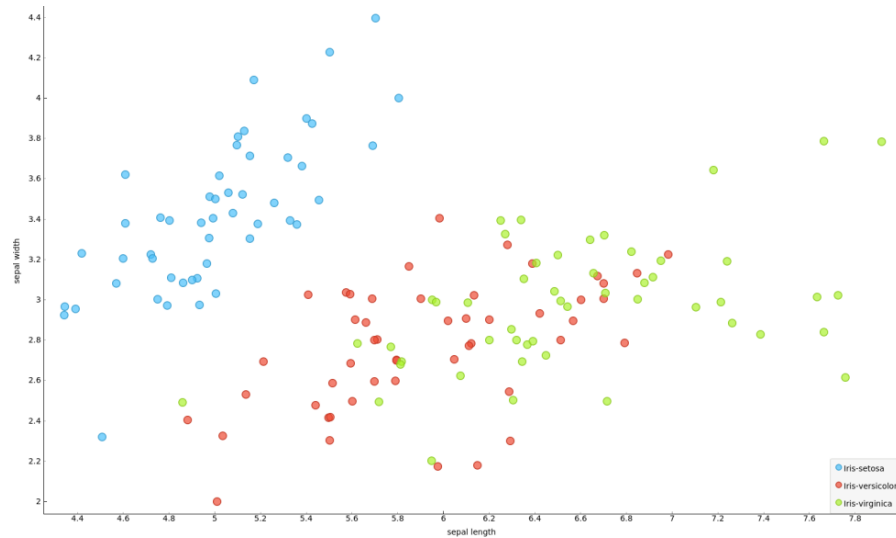
- Agrupación (*clustering*)
 - Detección de anomalías
 - Reglas de asociación
 - PCA
 - Métodos basados en redes
- 

Clustering

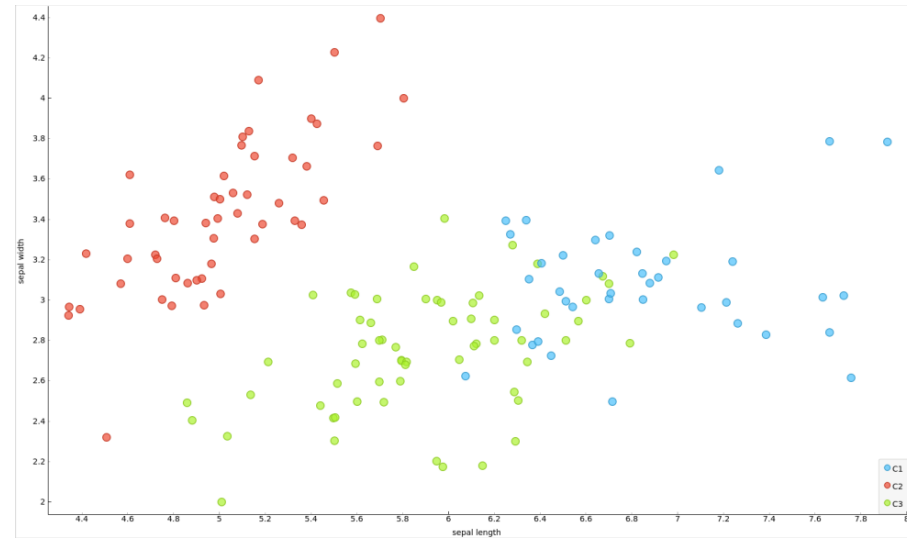
- Buscar puntos que estén **cerca** en el **espacio**
- Requiere una noción de distancia
 - Euclídea: $\sqrt{x^2 + y^2}$
 - Manhattan: $|x| + |y|$
 - Chebyshev: $\max(|x|, |y|)$
- Aspira a encontrar instancias similares, pero no vale para clasificar mágicamente

Clustering

Clases reales



Un *clustering* típico



Clustering no es aprendizaje supervisado

Herramientas para el modelado de datos

Guiados / GUI



Basados en flujo



Programáticos




Guiados/GUI: Deep Intelligence

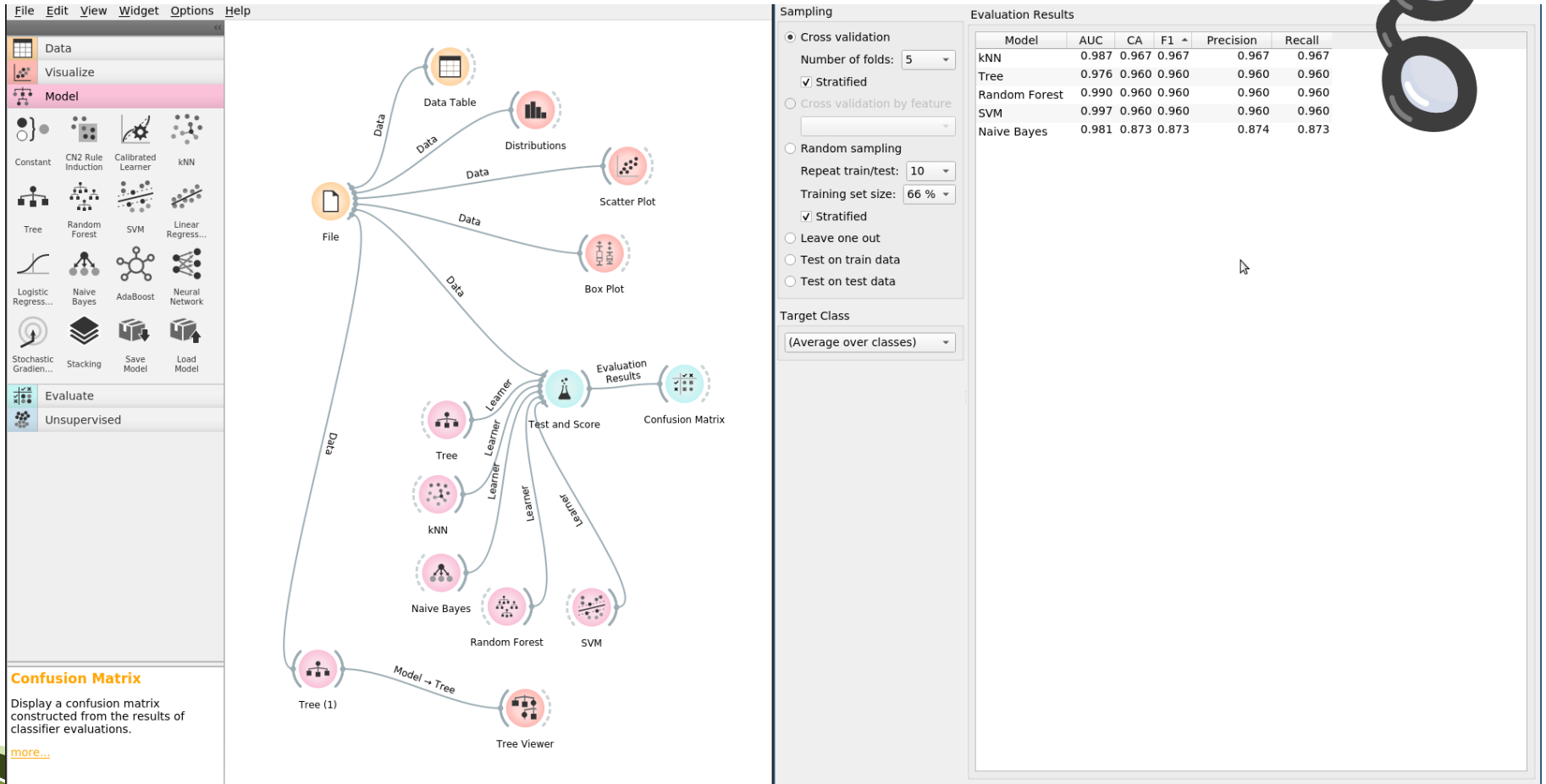


The screenshot shows the 'Create classifier' interface in the DEEP Intelligence application. The interface is divided into a sidebar on the left and a main content area. The sidebar contains navigation options: 'Back to home', 'Iris dataset', 'Tasks', 'Alerts', 'Manage permissions', 'Search...', 'Dashboards (1)', 'Sources (1)', 'Visualizations (6)', and 'Models (1)'. The main content area is titled 'Create classifier' and has a close button (X) in the top right corner. Below the title is a horizontal menu with tabs: 'Name', 'Data source', 'Target', 'Method', 'Training and testing', 'Scaling', and 'Advanced configuration'. The 'Method' tab is currently selected. A teal banner with an information icon (i) contains the text: 'Choose the machine learning method you want to use.' Below this banner is a table with two columns: 'Name' and 'Description'. The table lists four machine learning methods: Naive Bayes, Random forest (highlighted in teal), Gradient boosting, and Logistic regression. At the bottom right of the main content area, there are three buttons: '← BACK', 'NEXT →', and '+ CREATE MODEL'.

Name	Description
Naive Bayes	Naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features.
Random forest	Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction.
Gradient boosting	Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.
Logistic regression	The logistic model (or logit model) is used to model the probability of a certain class or event existing such as pass/fail,

Basados en flujo: Orange

orange 



The screenshot displays the Orange3 software interface. On the left is a widget palette with categories: Data, Visualize, Model, Evaluate, and Unsupervised. The main workspace contains a workflow starting with a 'File' widget connected to several 'Data' widgets. These 'Data' widgets feed into various visualization widgets: 'Data Table', 'Distributions', 'Scatter Plot', and 'Box Plot'. A 'Tree (1)' widget is connected to a 'Tree Viewer' widget. The workflow also includes a 'Learner' widget that feeds into a 'Test and Score' widget, which then connects to an 'Evaluation Results' widget. The 'Evaluation Results' widget displays a table of performance metrics for four models: kNN, Tree, Random Forest, and Naive Bayes. On the right side, the 'Sampling' widget is configured with 'Cross validation' selected, 'Number of folds' set to 5, 'Stratified' checked, and 'Training set size' set to 66%. The 'Target Class' is set to '(Average over classes)'. Below the workflow, a 'Confusion Matrix' widget is visible, with a tooltip explaining its function.

Confusion Matrix
Display a confusion matrix constructed from the results of classifier evaluations.
[more...](#)

Sampling

- Cross validation
 - Number of folds: 5
 - Stratified
- Cross validation by feature
- Random sampling
 - Repeat train/test: 10
 - Training set size: 66 %
 - Stratified
 - Leave one out
 - Test on train data
 - Test on test data

Target Class
(Average over classes)

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
kNN	0.987	0.967	0.967	0.967	0.967
Tree	0.976	0.960	0.960	0.960	0.960
Random Forest	0.990	0.960	0.960	0.960	0.960
SVM	0.997	0.960	0.960	0.960	0.960
Naive Bayes	0.981	0.873	0.873	0.874	0.873

Programático: scikit-learn



Aprendizaje automático

```
In [13]: 1 from sklearn.ensemble import RandomForestClassifier
          2 from sklearn.model_selection import train_test_split
```

```
In [14]: 1 X_train, X_test, y_train, y_test = train_test_split(df.drop("species", axis=1), df["species"], random_state=5)
```

```
In [15]: 1 random_forest = RandomForestClassifier()
          2 random_forest.fit(X_train, y_train)
```

```
Out[15]: RandomForestClassifier()
```

```
In [16]: 1 pd.DataFrame(np.asarray([random_forest.predict(X_test), y_test]).T, columns=["Predicho", "Real"])
```

```
Out[16]:
```

	Predicho	Real
0	Versicolor	Versicolor
1	Versicolor	Virginica
2	Virginica	Virginica
3	Setosa	Setosa
4	Virginica	Virginica
5	Versicolor	Versicolor
6	Setosa	Setosa
7	Virginica	Versicolor

MUCHAS GRACIAS



centratec@air-institute.com
info@air-institute.com



@TheAirInstitute



AIR Institute



@TheAirInstitute



<https://empresas.jcyl.es/web/es/idi/programa-centratec.html>

<https://centratec.air-institute.com/>