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

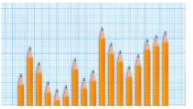
by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

SUMMARY SAVE SHARE 16 COMMENT 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

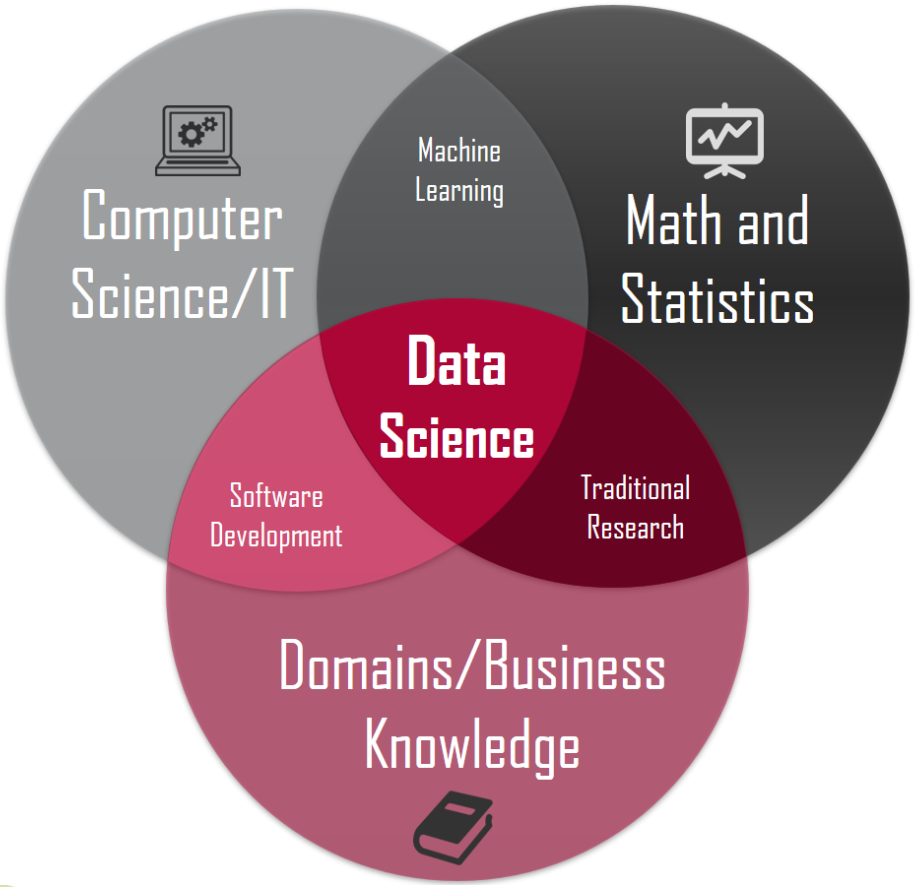


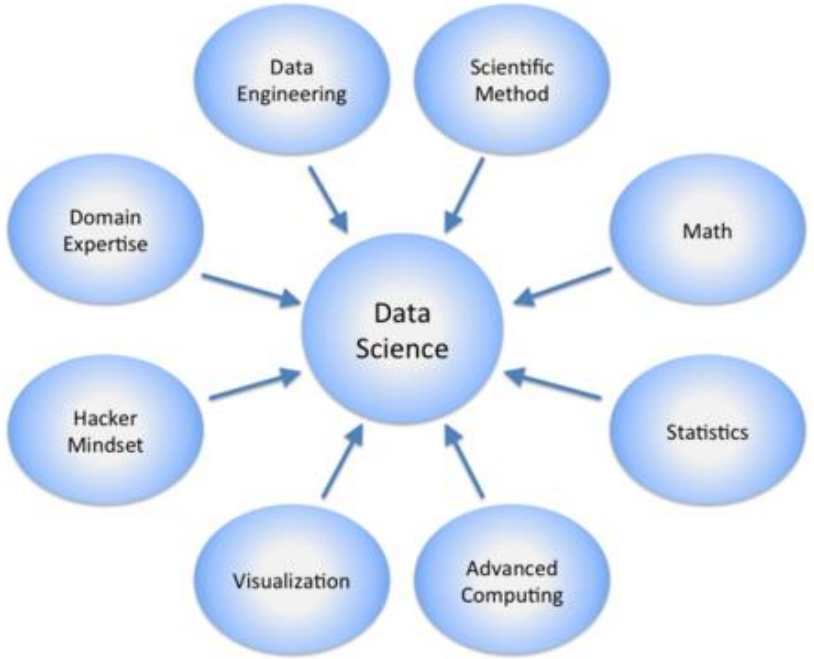
What Data Scientists Really Do, According to 35 Data Scientists

VIEW MORE FROM THE

October 2012 Issue

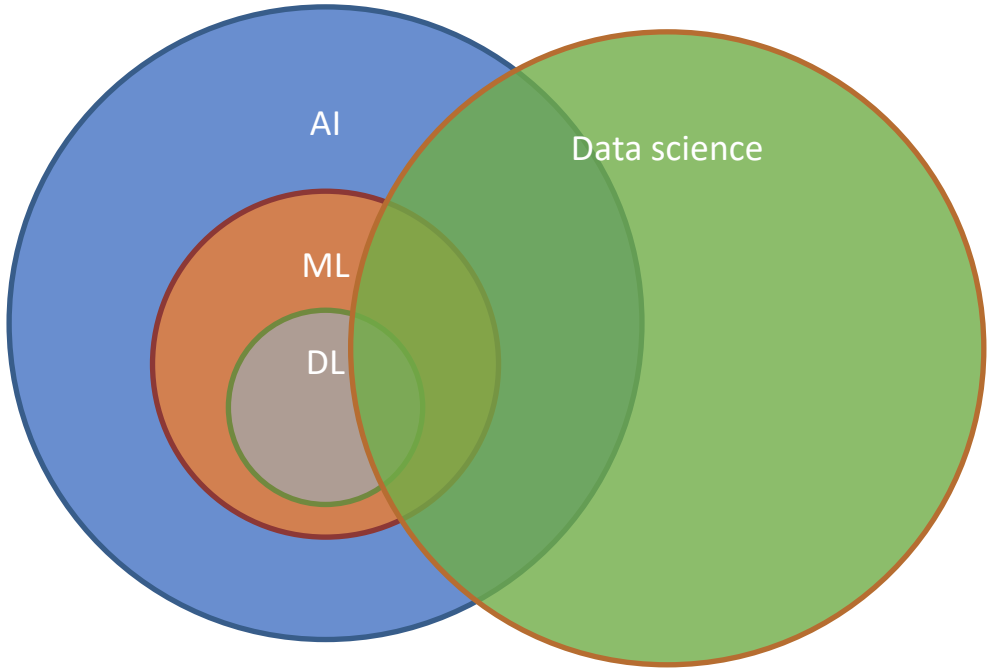
Ciencia de datos





Ciencia de datos

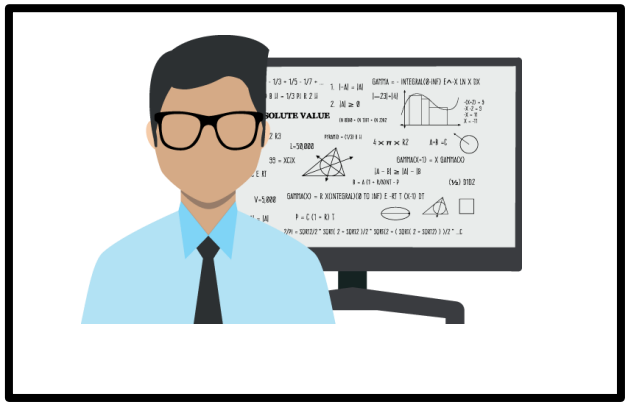
- **AI**: Imitar el comportamiento humano
- **ML**: Extraer información de forma automática de datos
- **DL**: Técnicas especializadas de ML
- **Ciencia de datos**: extraer conocimiento de datos



Data scientist as a service

Datos

Científico de datos



Conocimiento

¿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

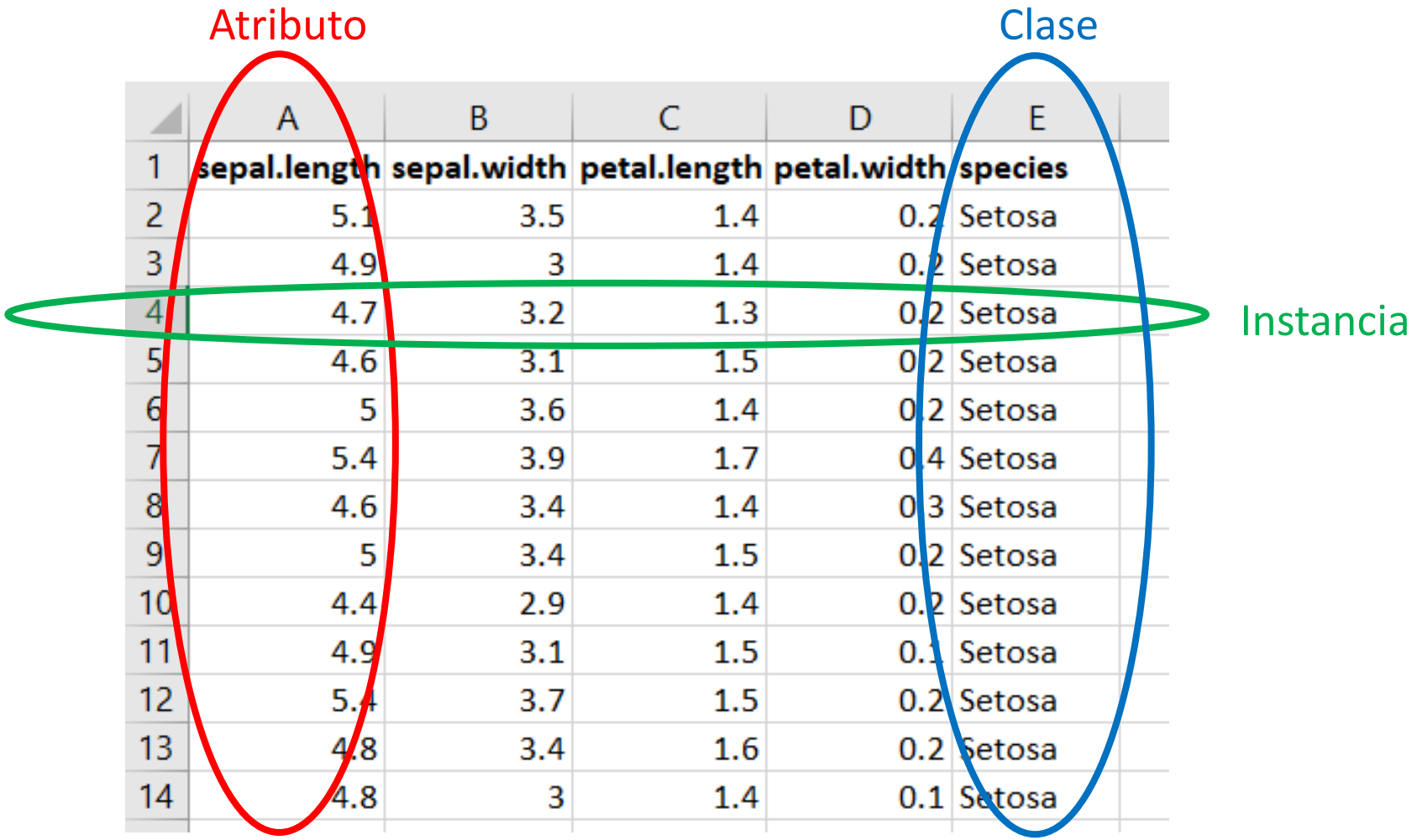
- Qué hace falta
  - Programación (Python, R)
  - Matemáticas (probabilidad, estadística, álgebra lineal, cálculo)
  - Técnicas de visualización
  - Aprendizaje automático
  - Ingeniería software
  - Bases de datos
  - Experiencia

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



	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

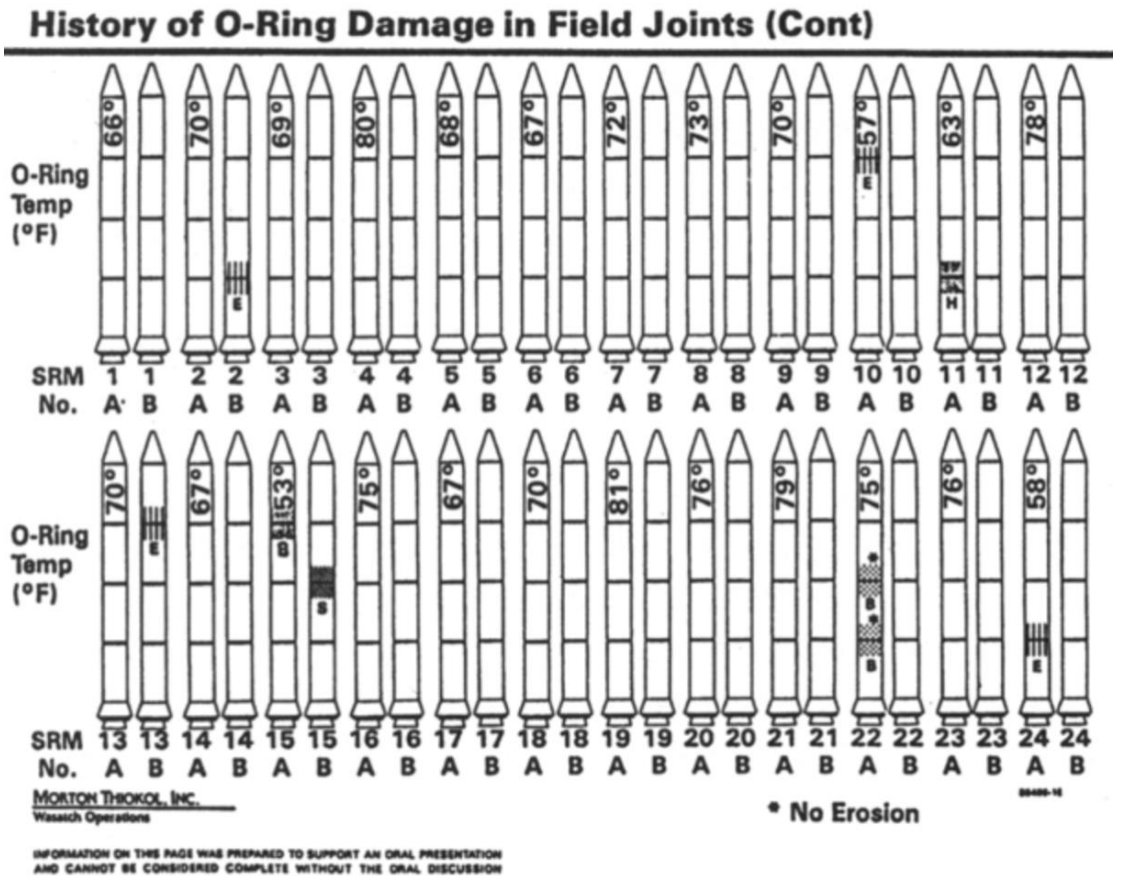
Tipos de datos

- **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**.

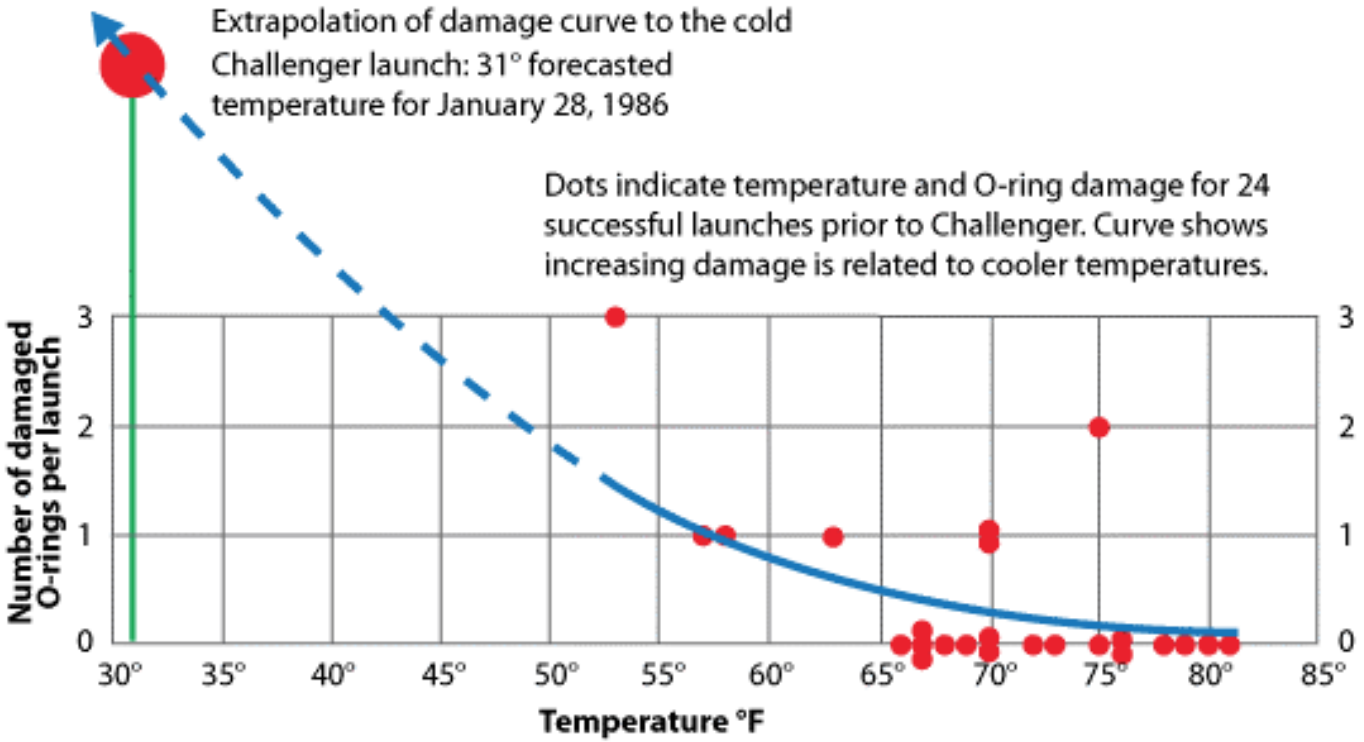
Tipos de datos

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, …	Categórico	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.

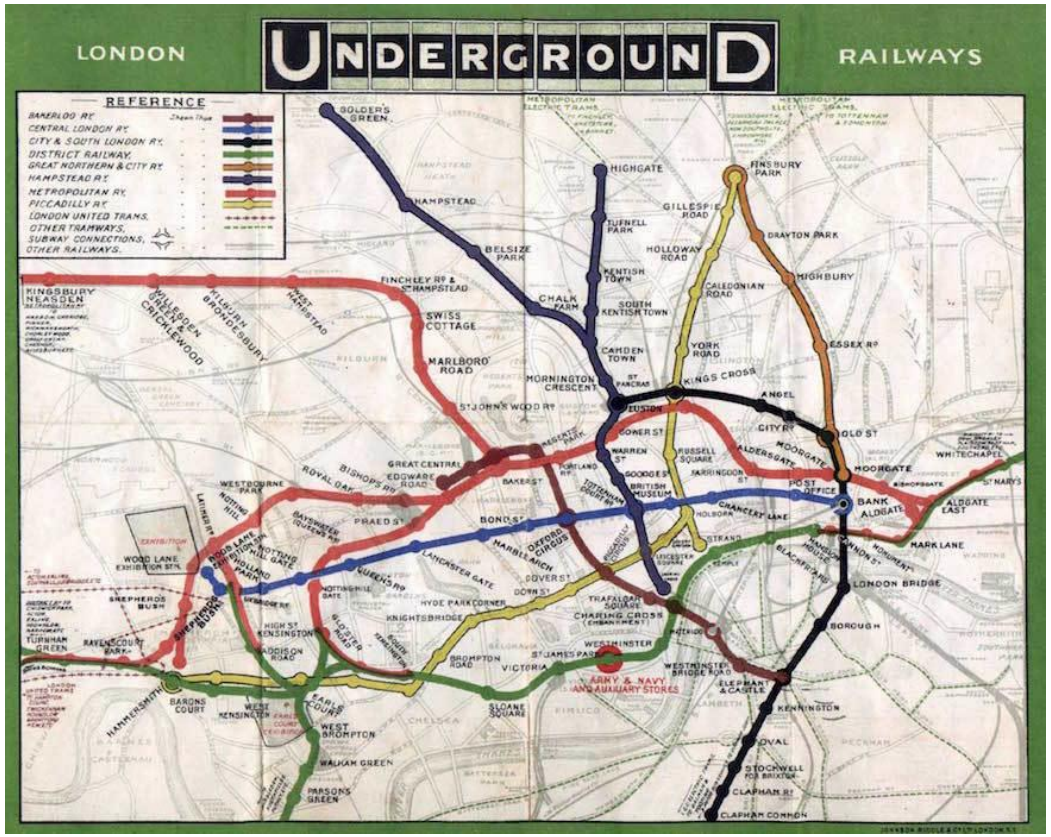
La importancia de una visualización eficaz



La importancia de una visualización eficaz

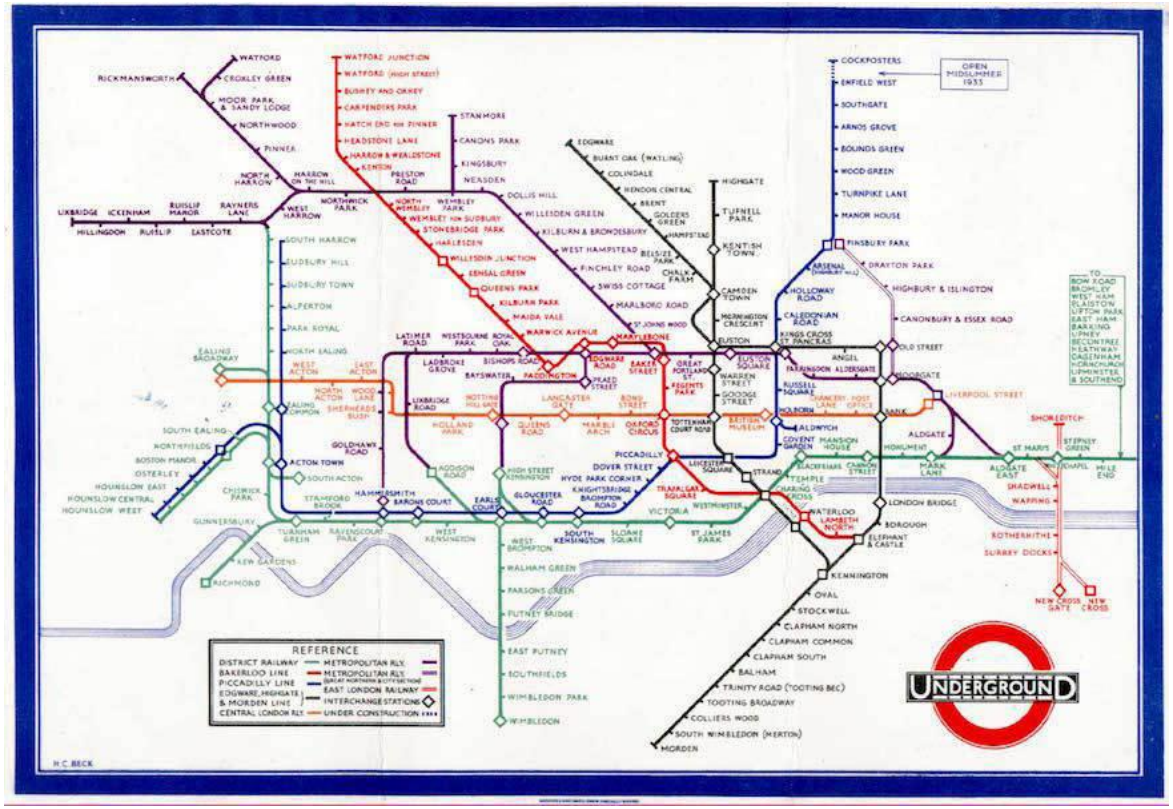


La importancia de una visualización eficaz



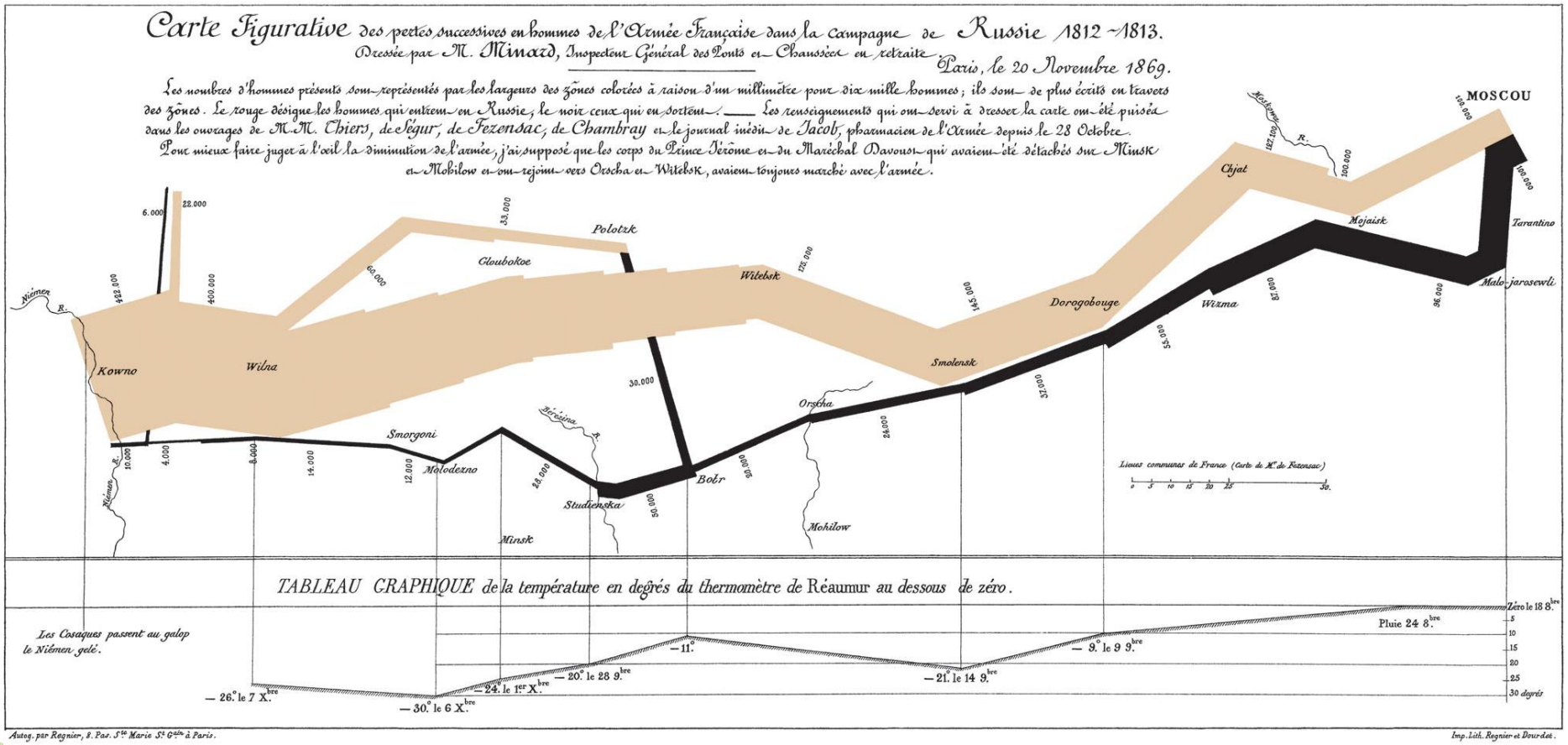
Metro de Londres
1908

La importancia de una visualización eficaz



Harry Beck
1933

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

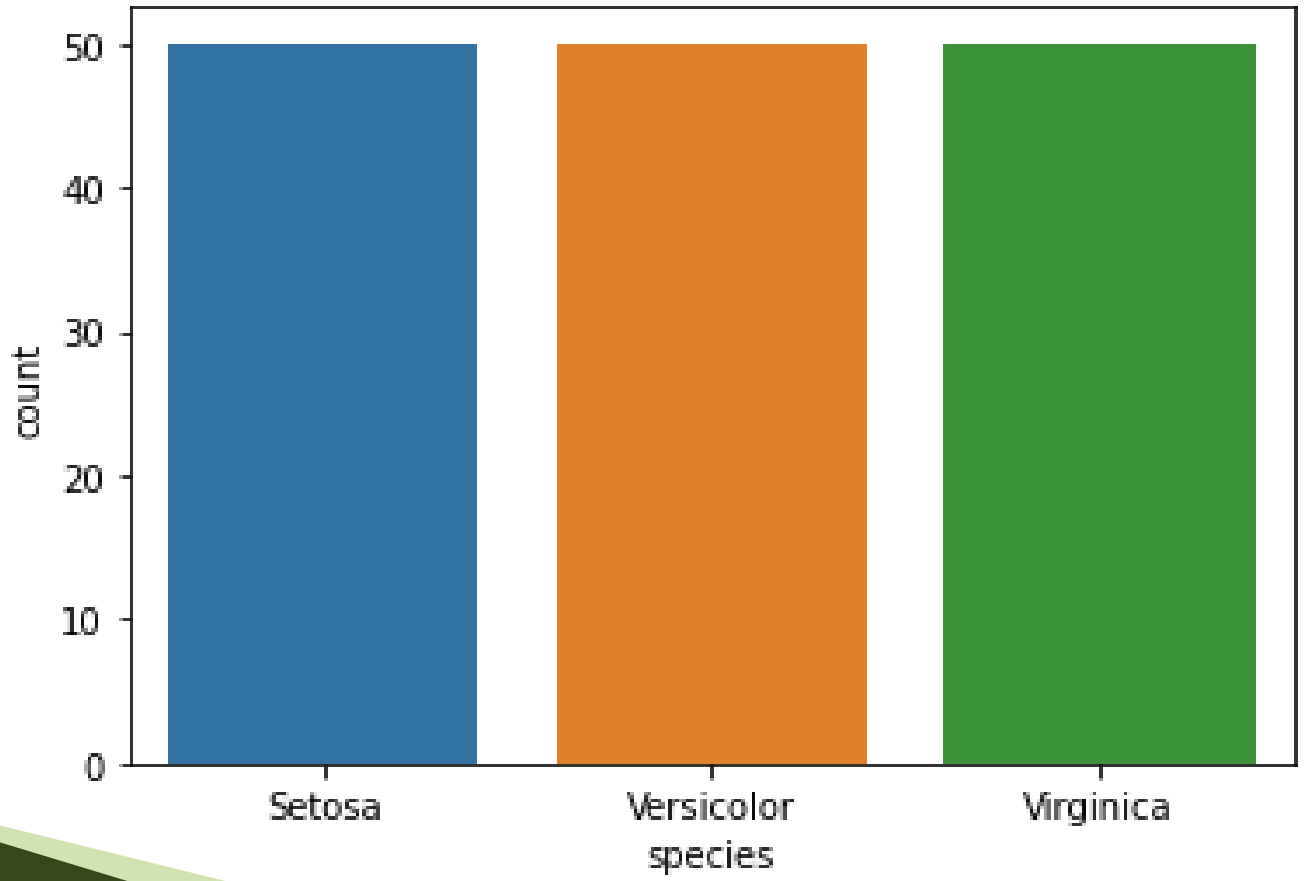
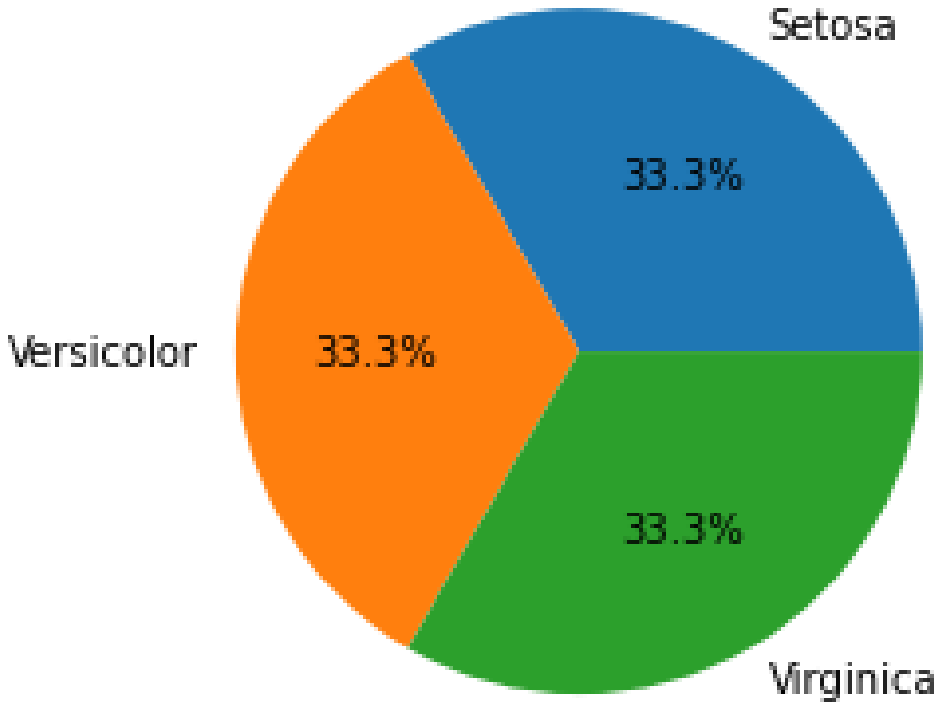


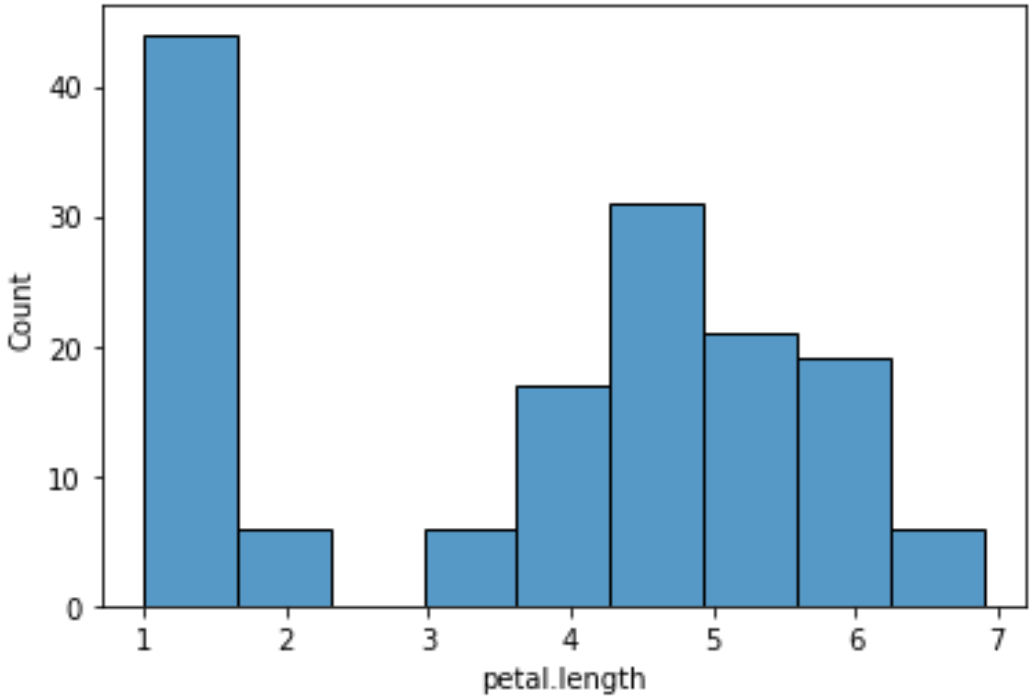
Gráfico de sectores

- Para variables discretas



Histograma

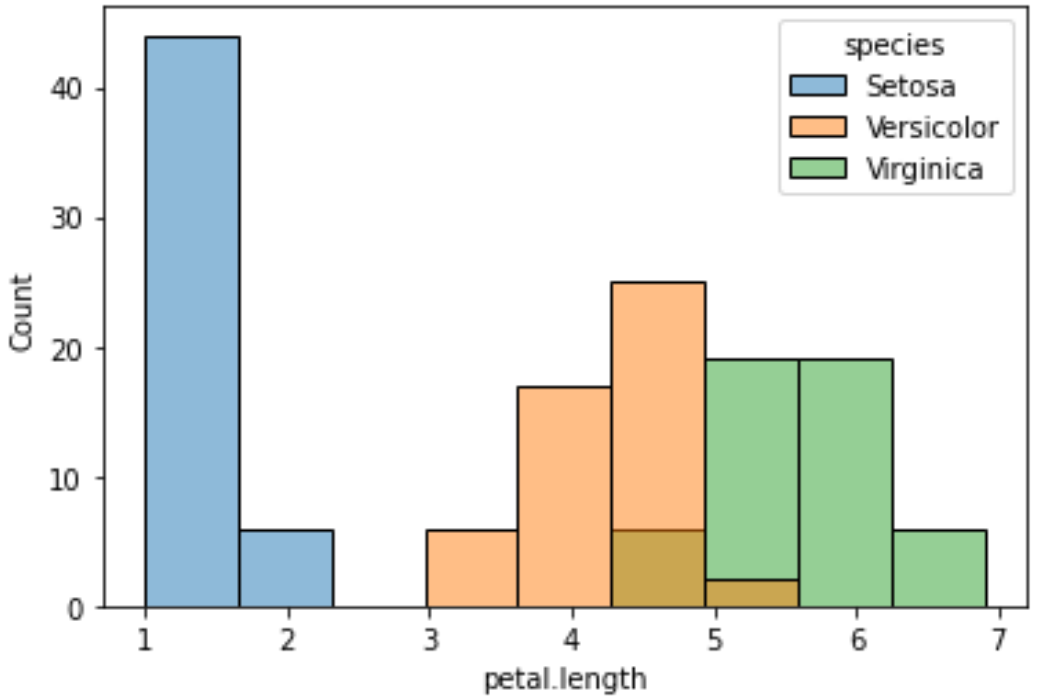
- Para variables continuas



Área proporcional a probabilidad total en el intervalo

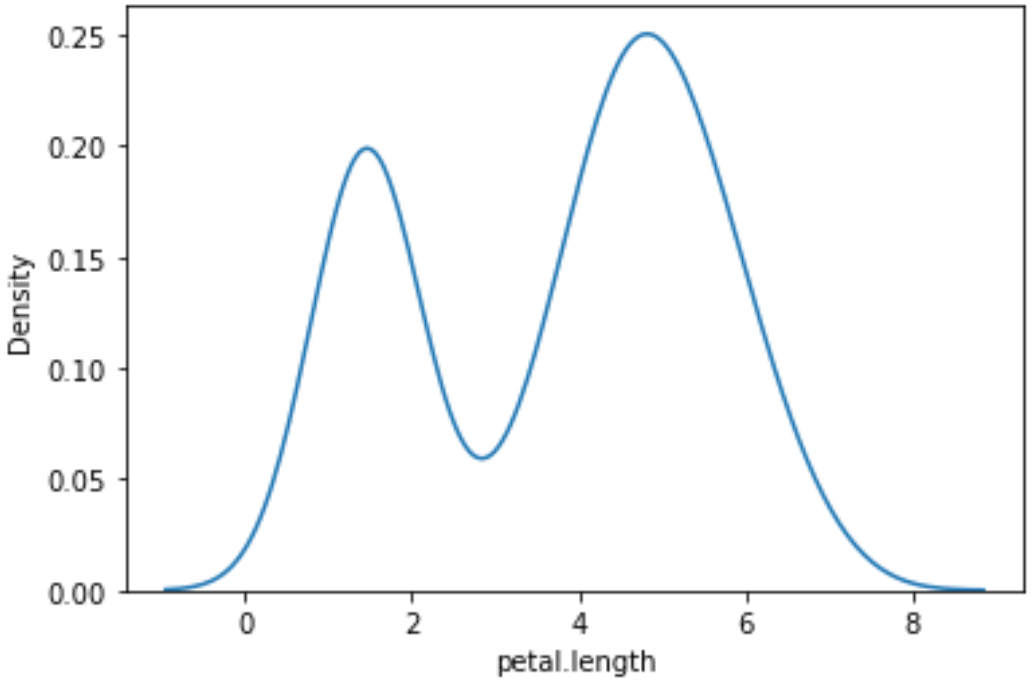
Histograma por series

- 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

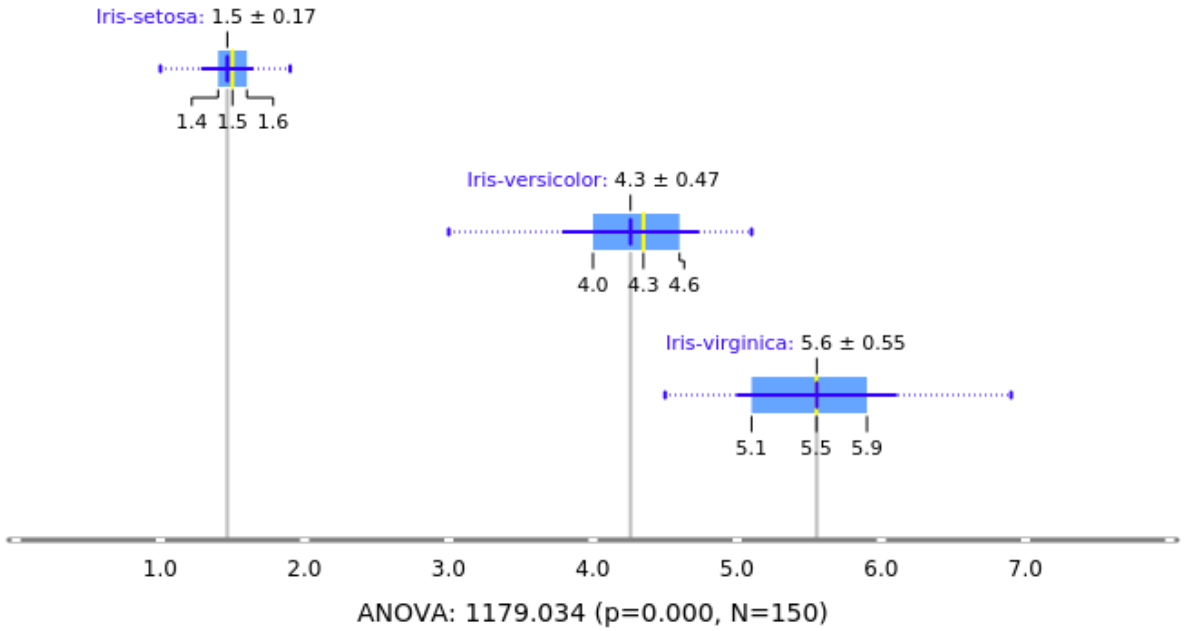
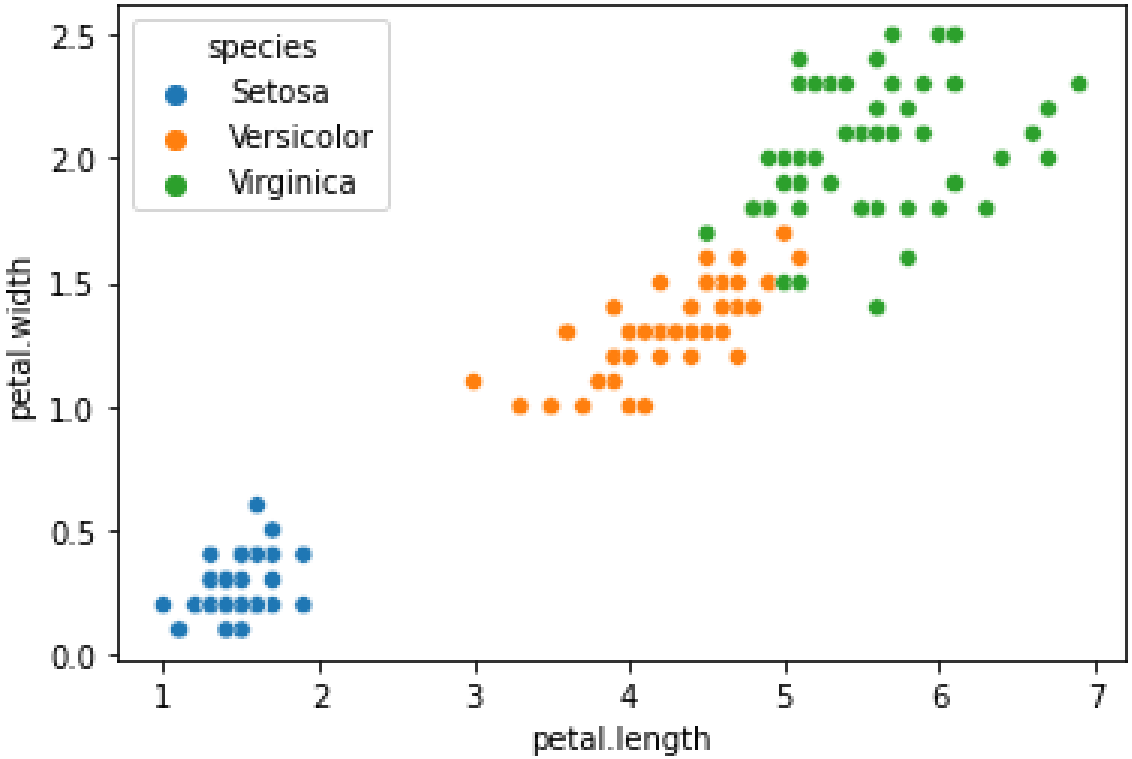


Gráfico de dispersión (*scatter*)

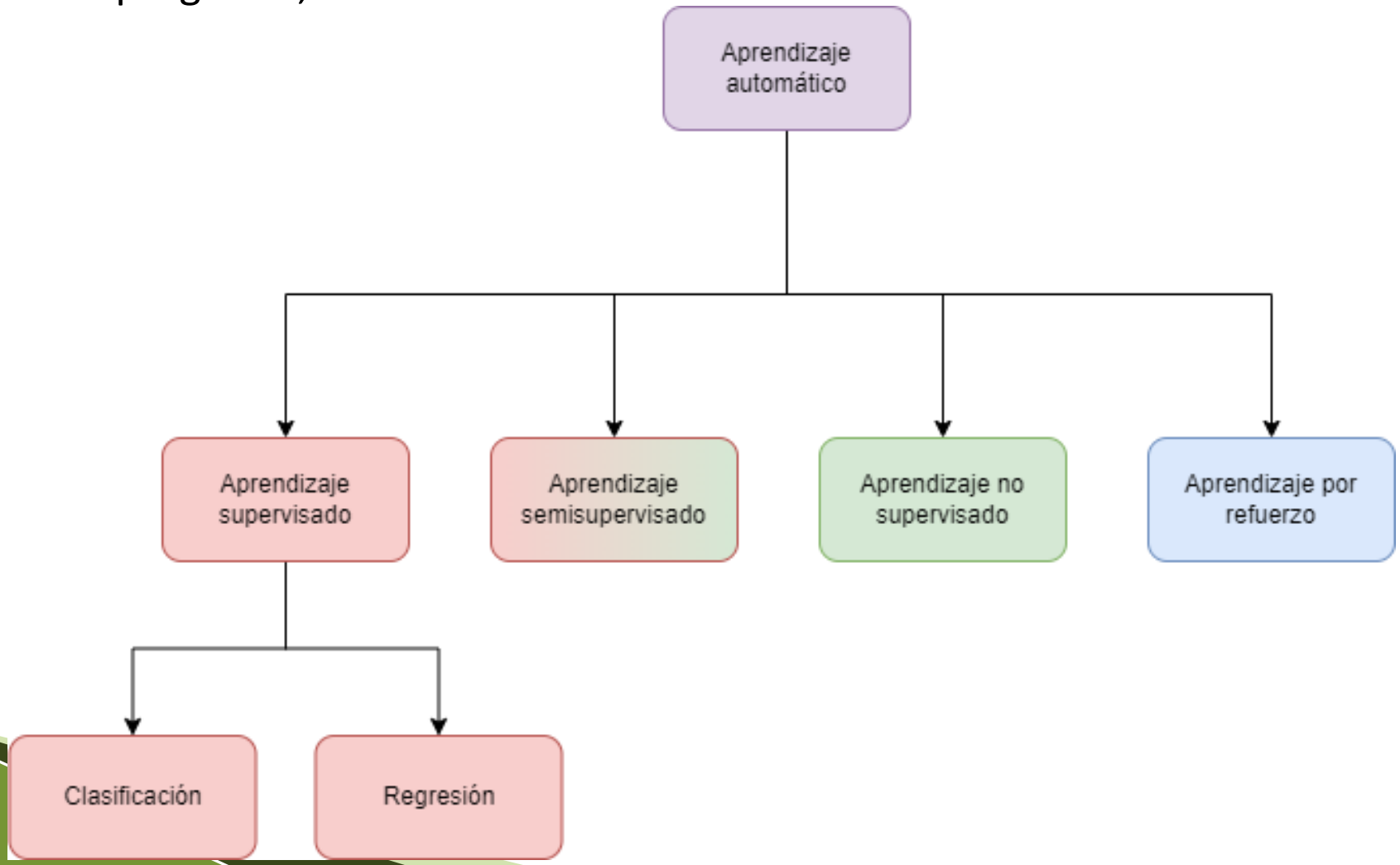
- Para dos variables numéricas



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

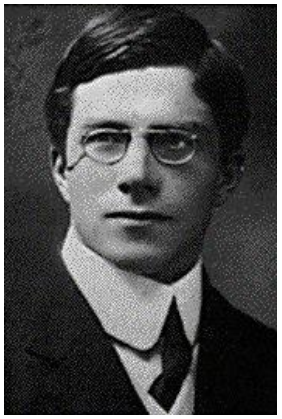
Fisher's *Iris* Data

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


Iris setosa


Iris versicolor


Iris virginica



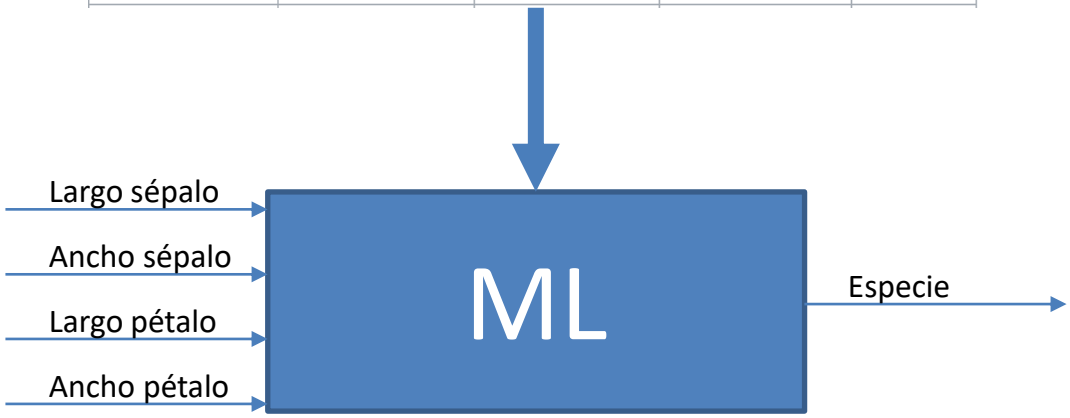
Ronald Fisher

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. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
4.6	3.1	1.5	0.2	*I. setosa*
5.0	3.6	1.4	0.2	*I. setosa*



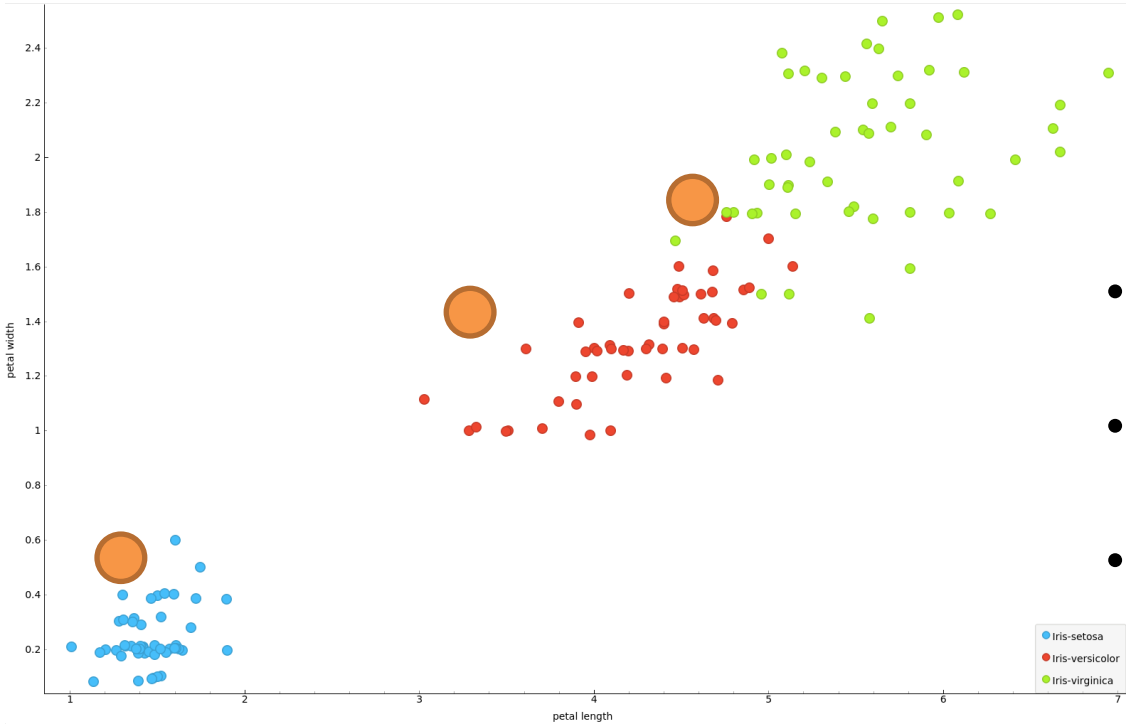
Construir una función $\mathrm{Species}(l_\mathrm{s}, w_\mathrm{s}, l_\mathrm{p}, w_\mathrm{p})$ que clasifique la especie (minimizando el error)

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

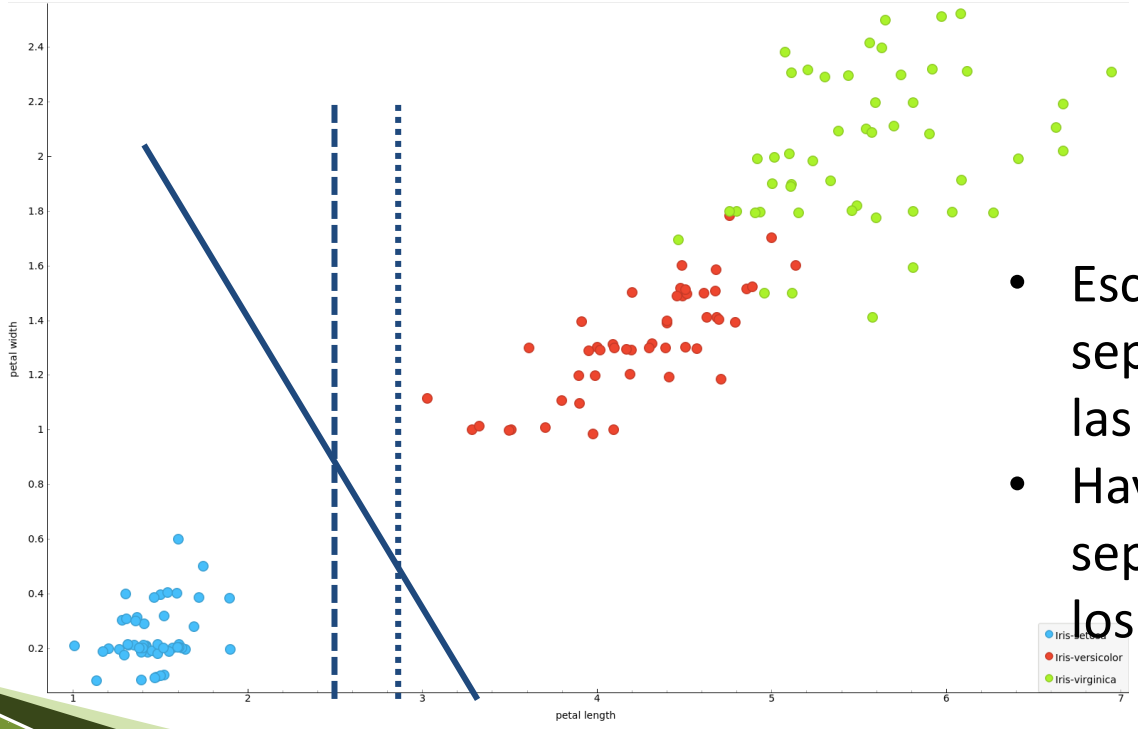
- Vecinos más cercanos



- 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

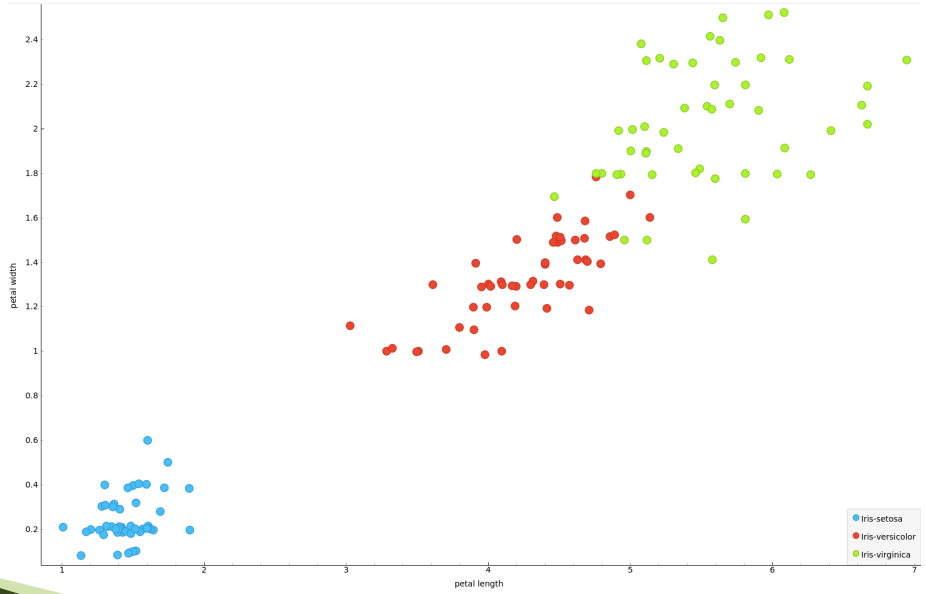
- Máquinas de vectores de soporte



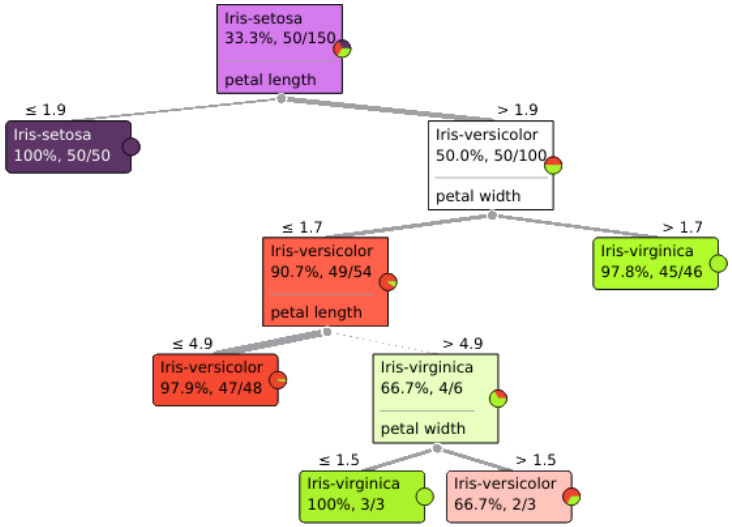
- Escoger **hiperplanos** que separen con el mayor **margen** las clases
- Hay “trucos” para conseguir separaciones más complejas que los hiperplanos

Algunas técnicas de aprendizaje supervisado

- Árboles de decisión

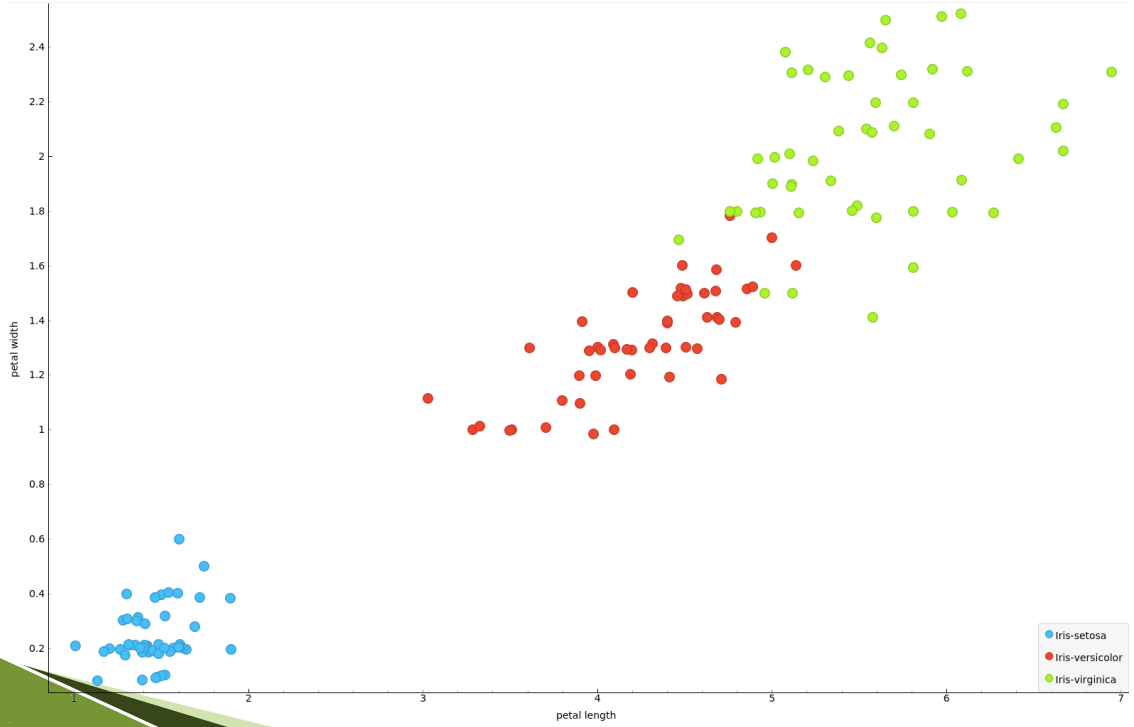


- Ejemplo cualitativo:
- ¿largo < 2.5?:
  - Sí: *Setosa*
  - No: ¿Alto > 1.6?
    - Sí: *Virginica*
    - No: *Versicolor*



Algunas técnicas de aprendizaje supervisado

- Ensambles



Tenemos varios predictores:

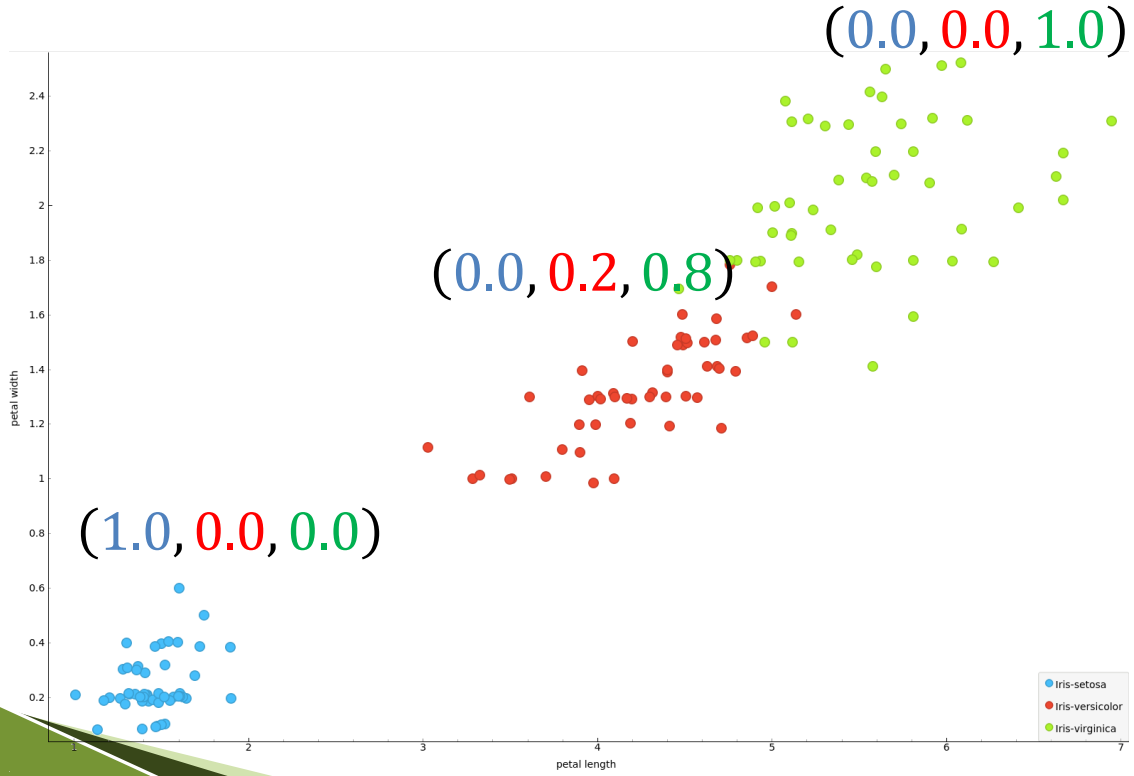
- Predictor 1: *versicolor*
- Predictor 2: *versicolor*
- Predictor 3: *virginica*

Métodos de agregación más sencillos:

- Mayoría simple
- Mayoría ponderada con confianza

Algunas técnicas de aprendizaje supervisado

- Redes neuronales

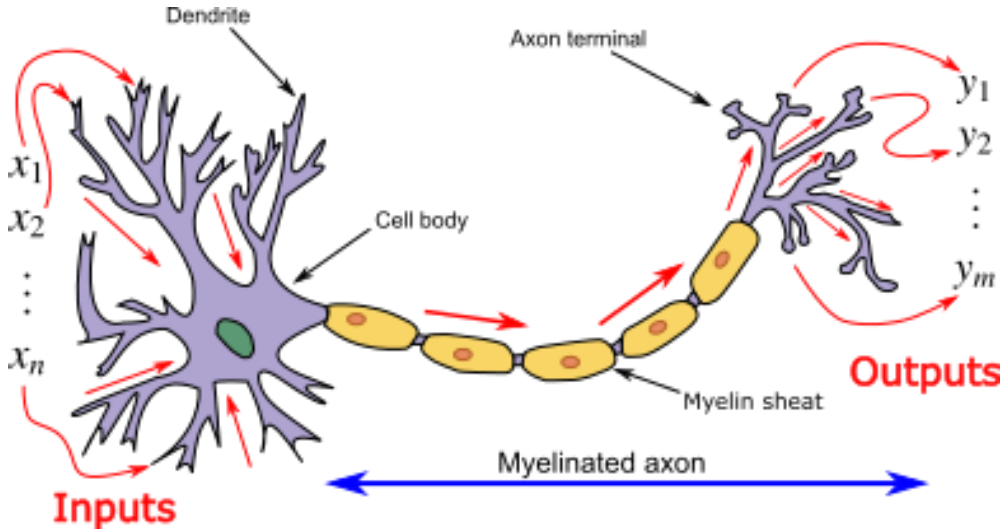


Supongamos que existe una función matemática que nos da las probabilidades de cada especie para una entrada arbitraria

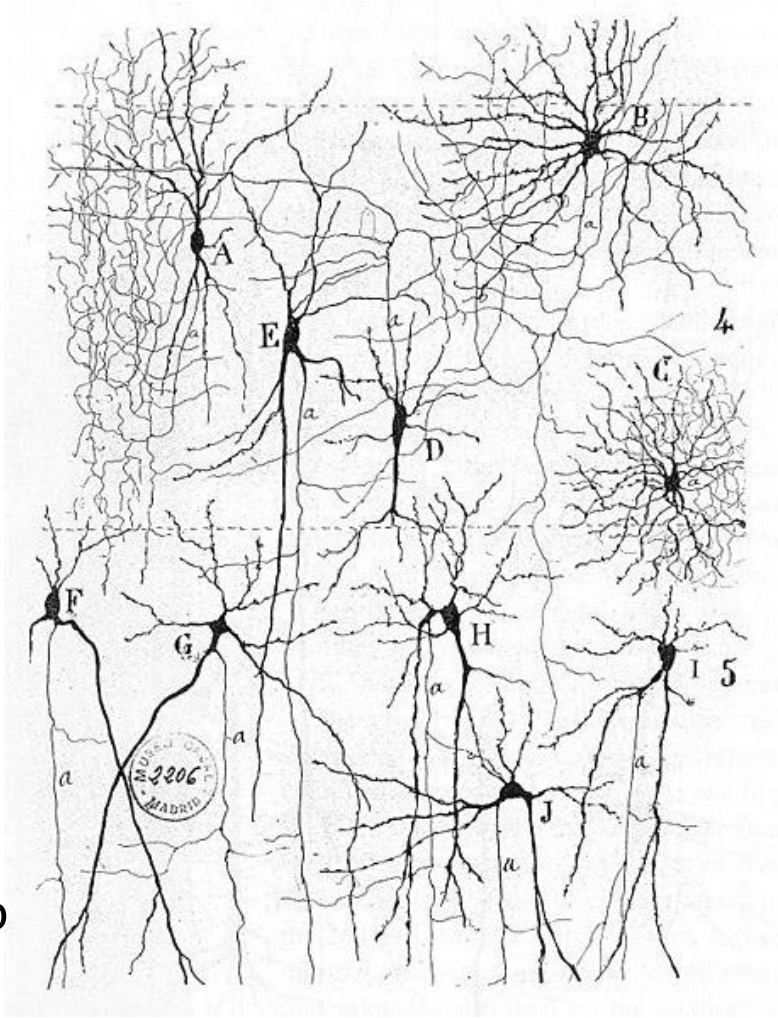
¿Podemos aproximar una función arbitraria?

Redes neuronales artificiales

- Neuronas

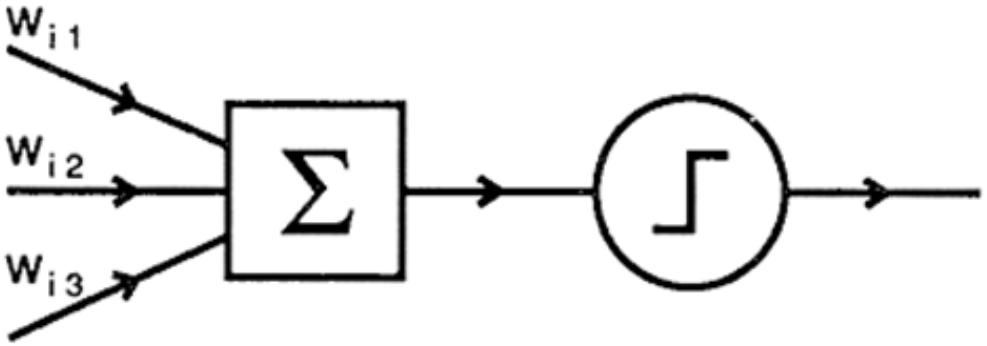


CC. Porf. Loc Vu-Quoc



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

Un modelo de neurona



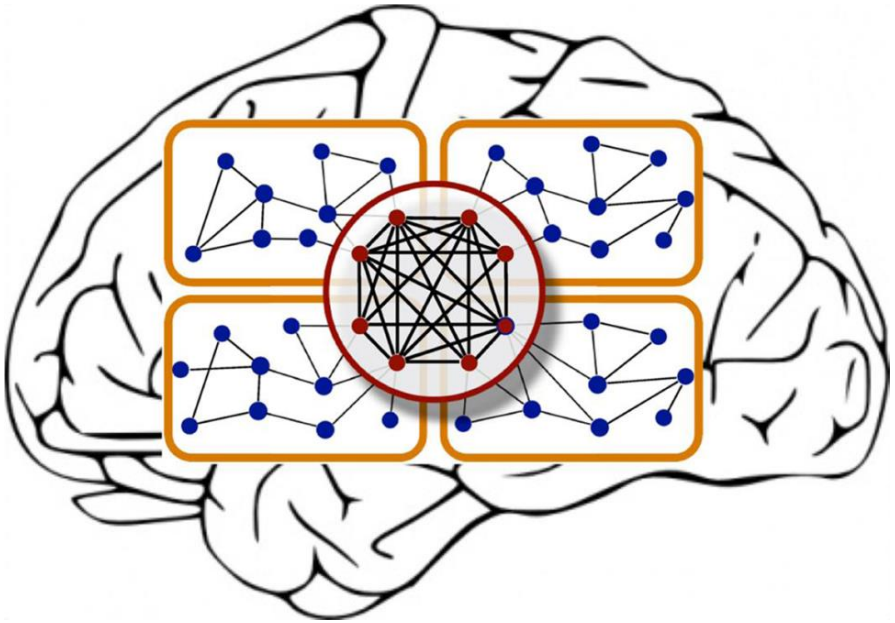
Pesos ajustables

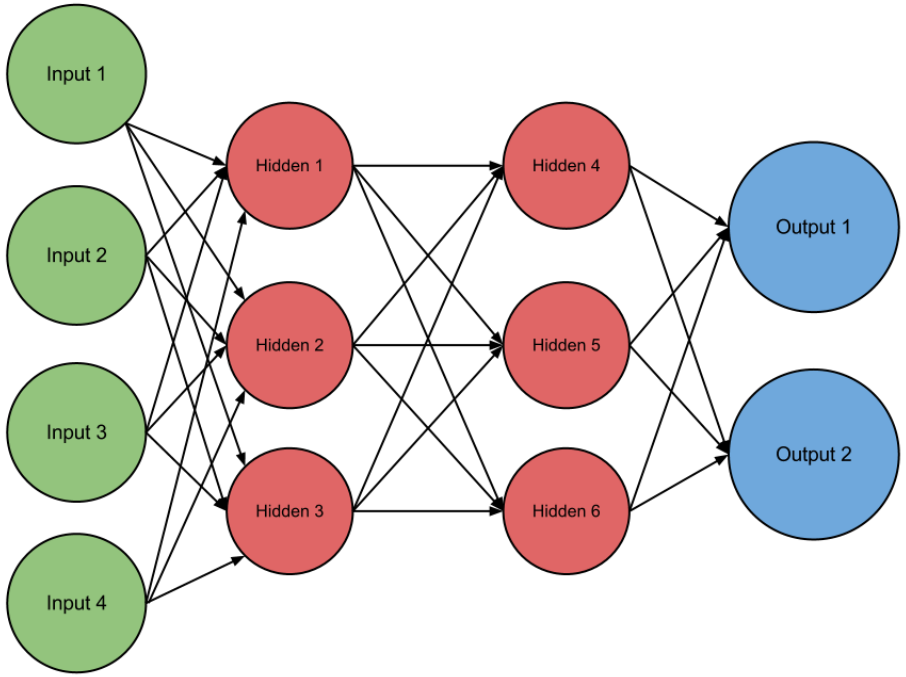
salida $\longrightarrow$ $y = \theta\left(\sum_{i=1}^{d} w_i x_i\right)$

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

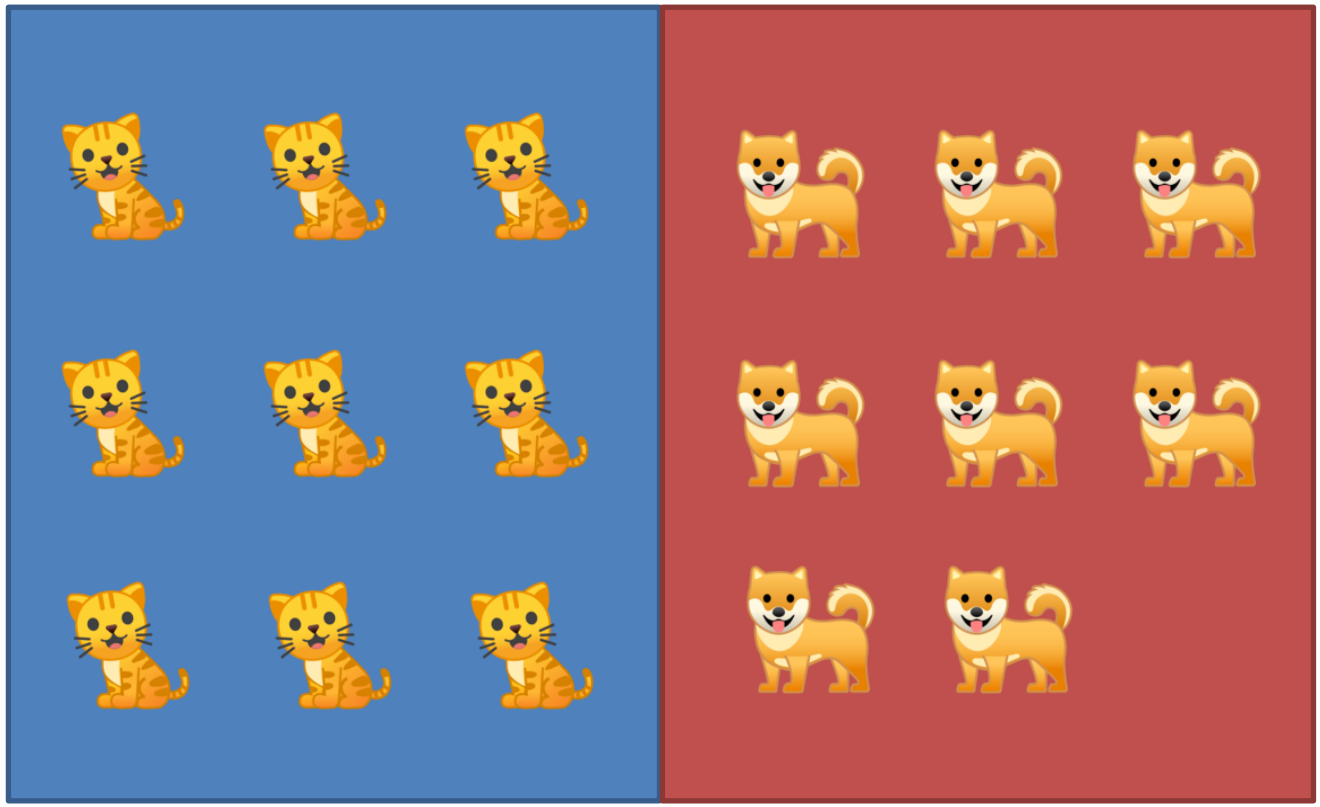
Mathematics of Control, Signals, and Systems

Approximation by Superpositions of a Sigmoidal Function*

G. Cybenko†

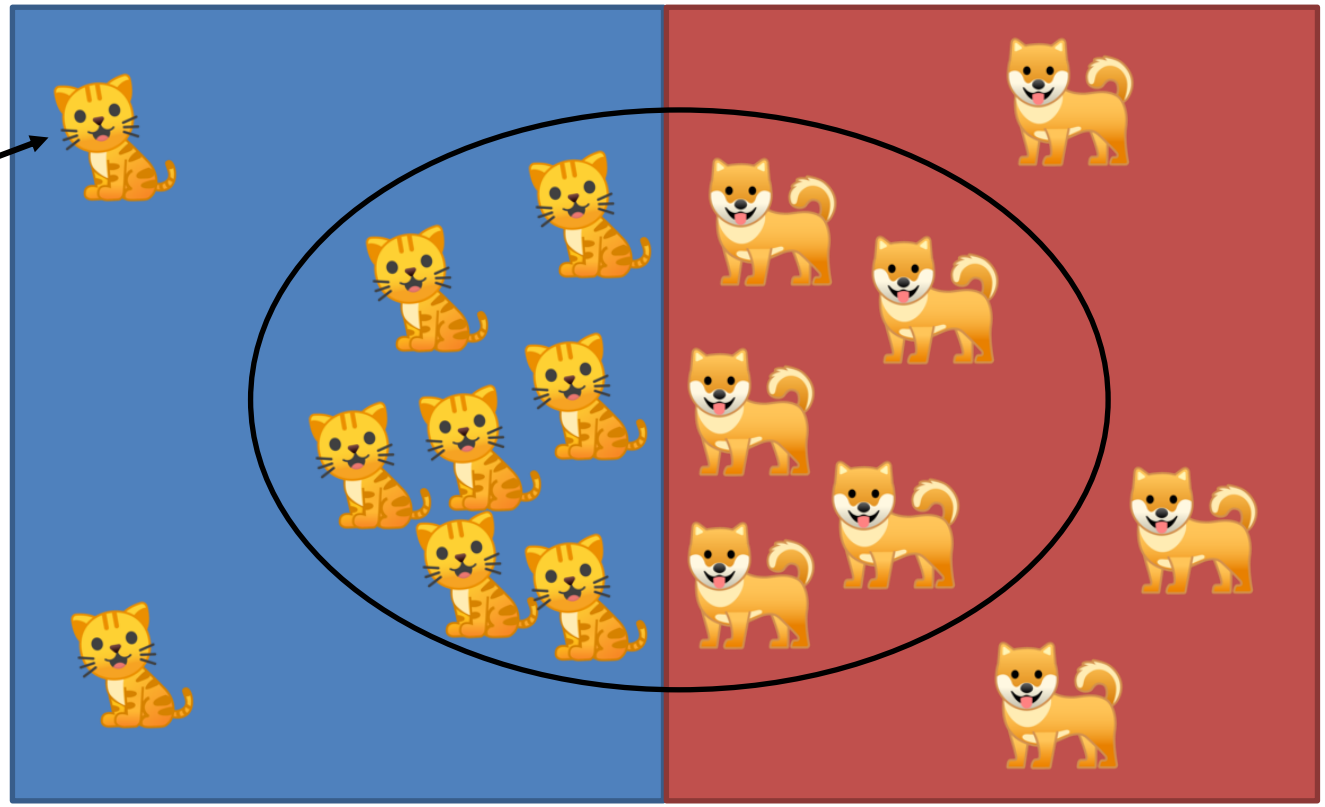
Evaluación

- Clasificación (2 clases)



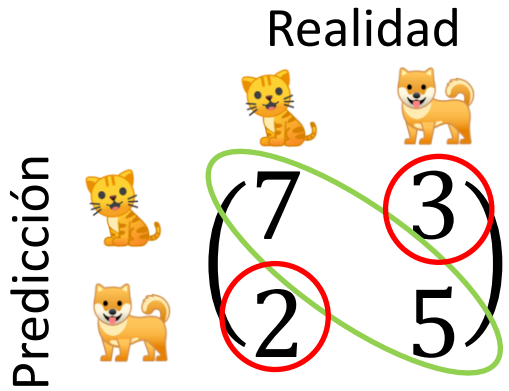
Evaluación

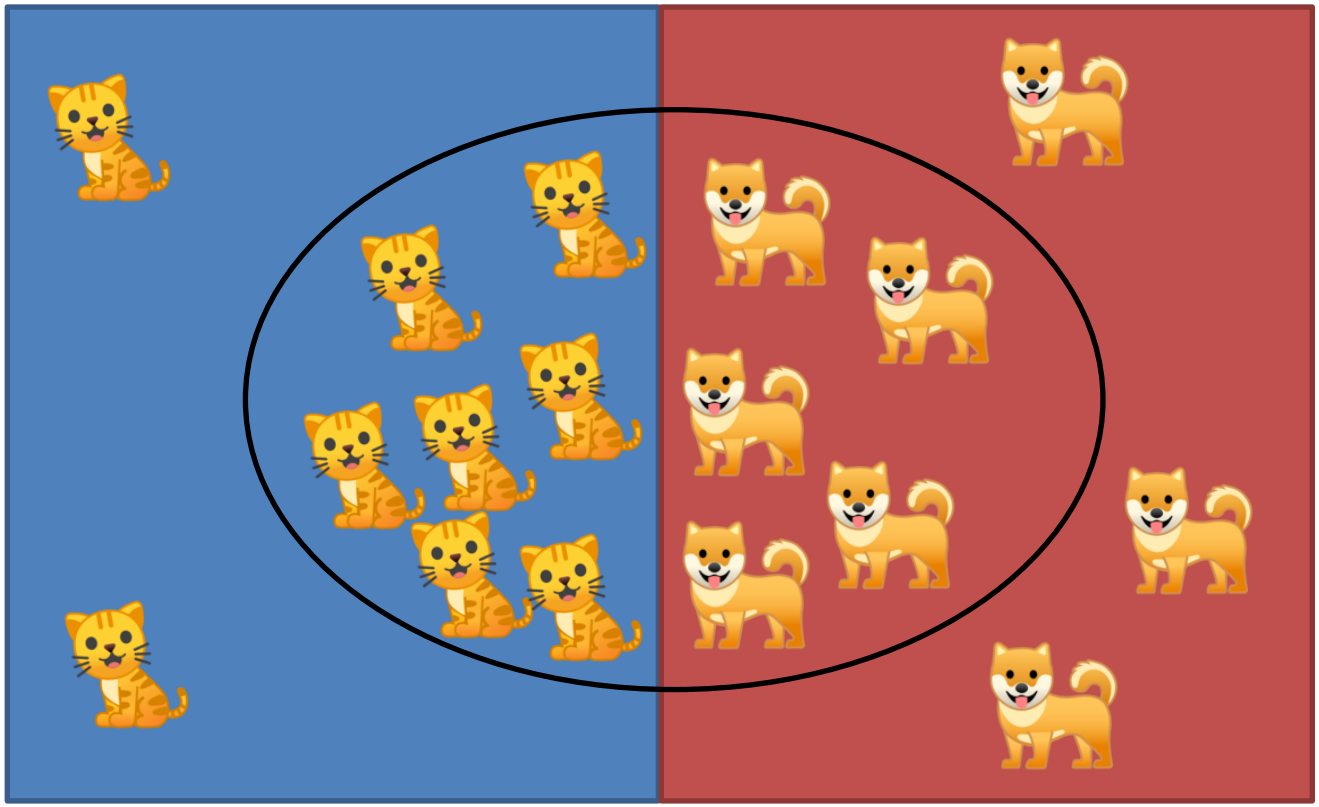
- Clasificación (2 clases)



Evaluación

- Clasificación (2 clases)





Evaluación

- Clasificación (2 clases)

Realidad

Predicción

$$\begin{pmatrix} 7 & 3 \\ 2 & 5 \end{pmatrix}$$

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

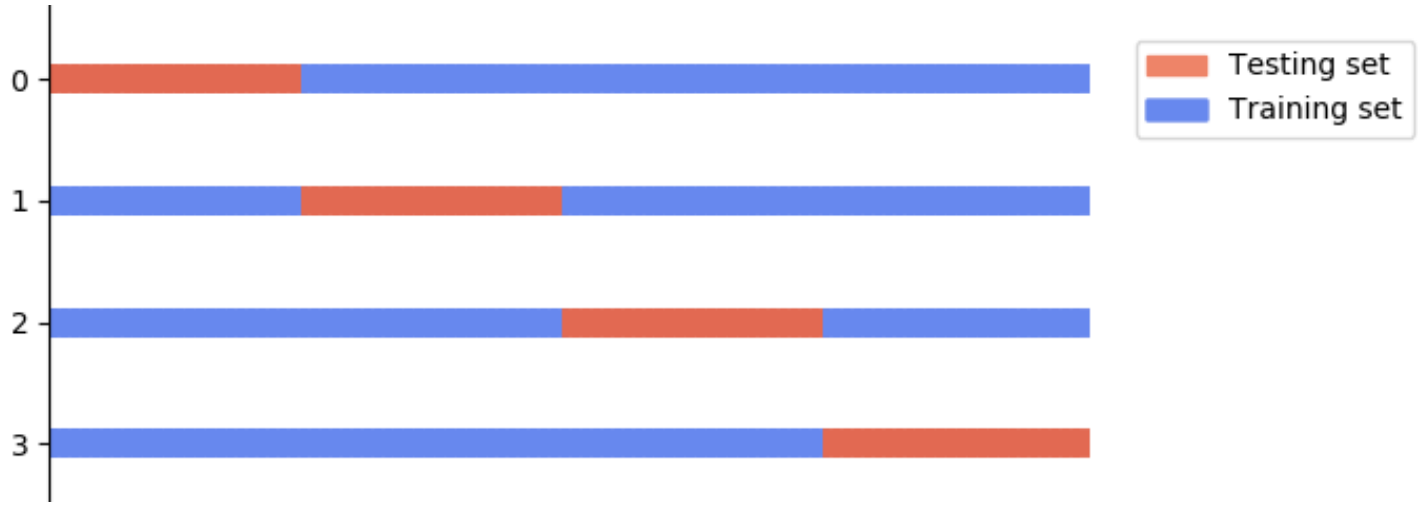
- Precisión (*precision*): $\frac{tp}{tp+fp}$
- Exhaustividad (*recall*): $\frac{tp}{tp+fn}$
- Exactitud (*accuracy*): $\frac{tp+tn}{tp+tn+fp+fn}$
- F1: $\left(\frac{1}{\text{prec}} + \frac{1}{\text{recall}}\right)^{-1}$

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

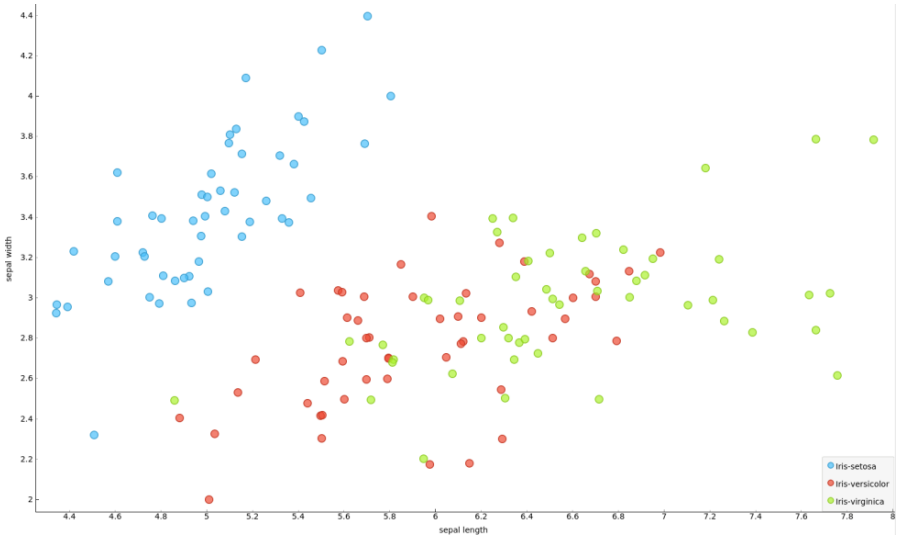
Clustering

- Buscar puntos que estén **cerca** en el **espacio**
- Requiere una noción de distancia
  - Euclídea: $\sqrt{x^2 + y^2}$
  - Manhattan: $|x| + |y|$
  - Chebyshev: $\max(|x|, |y|)$

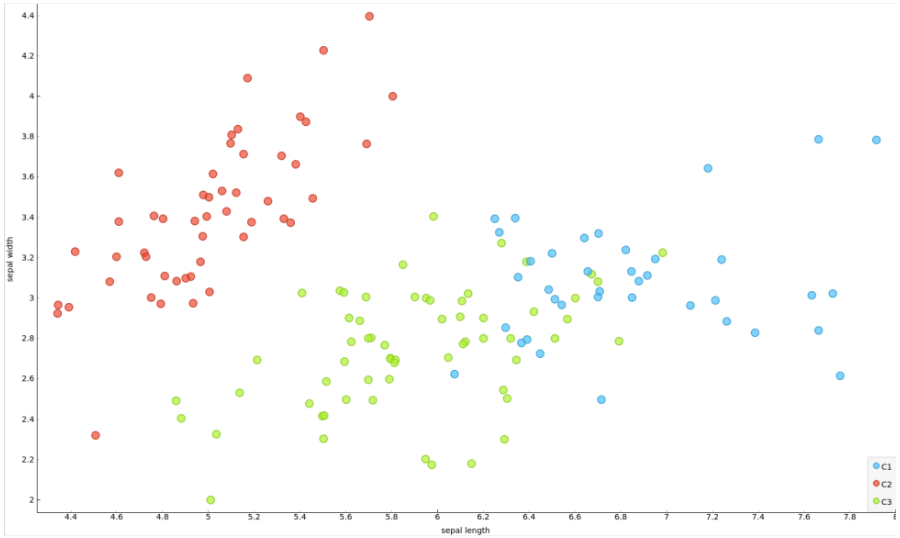
- Aspira a encontrar instancias similares, pero no vale para clasificar mágicamente

Clustering

Clases reales



Un *clustering* típico



***Clustering* no es aprendizaje supervisado**

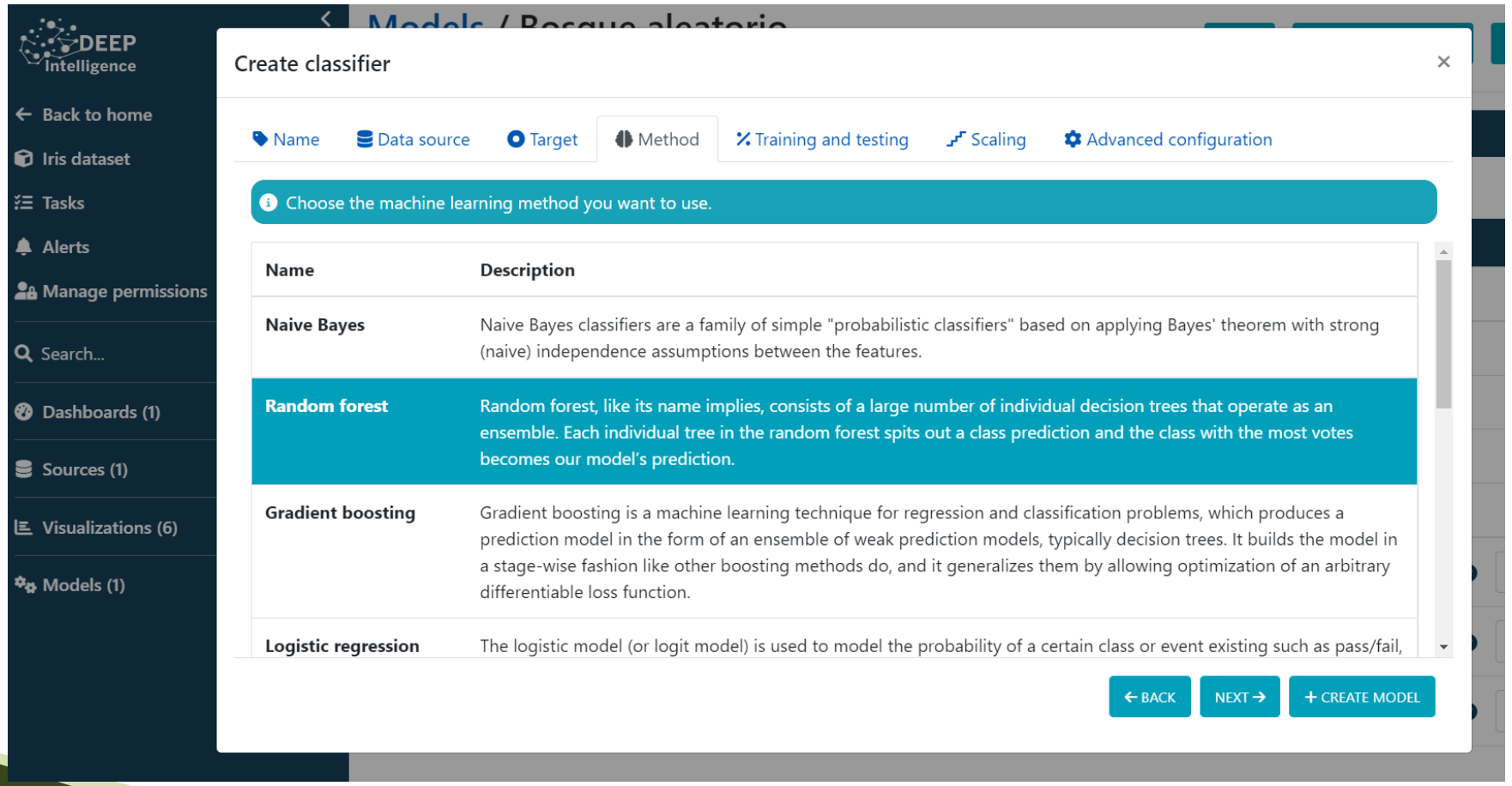
Herramientas para el modelado de datos

Guiados / GUI

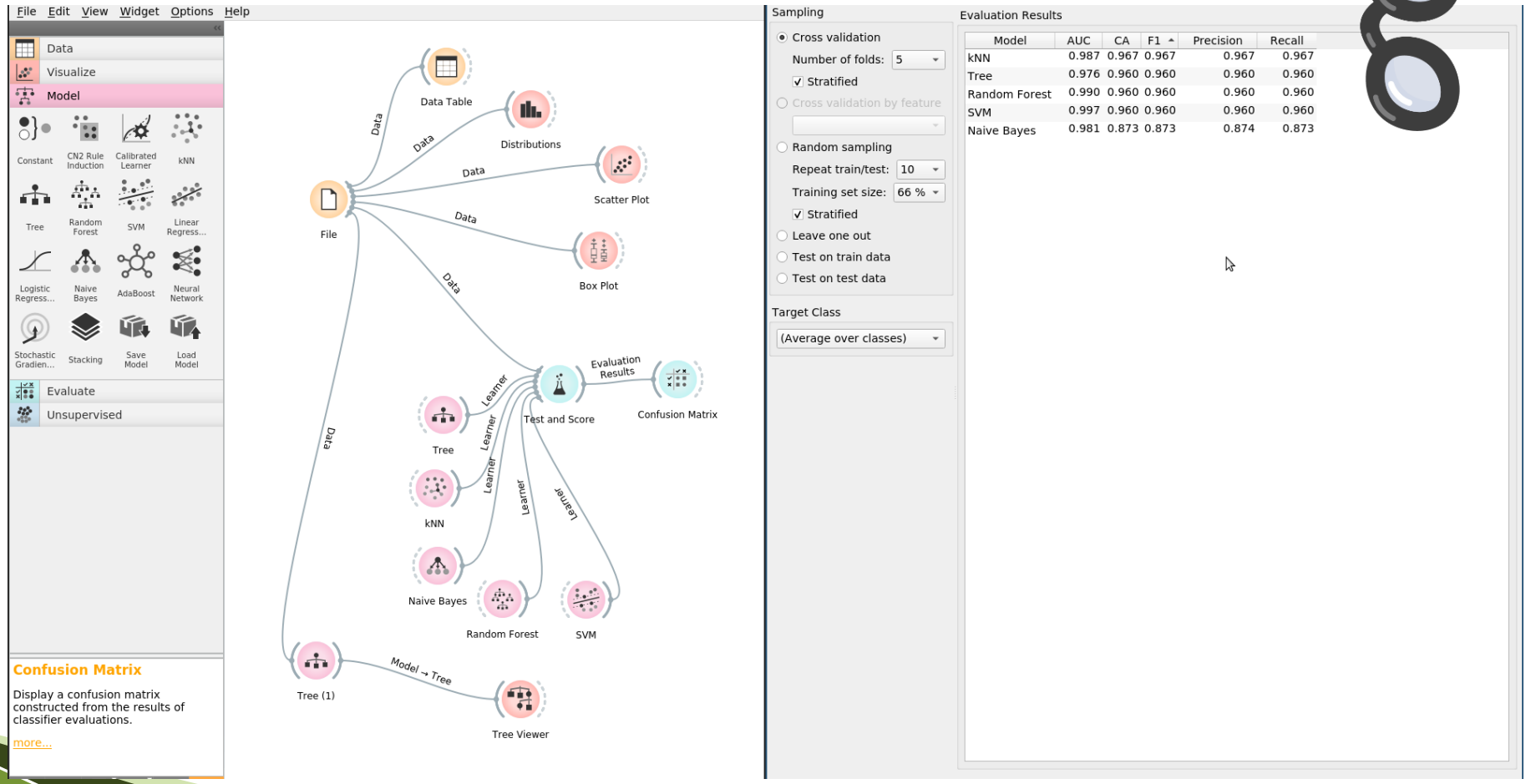
Basados en flujo

Programáticos

Guiados/GUI: Deep Intelligence



Basados en flujo: Orange



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/