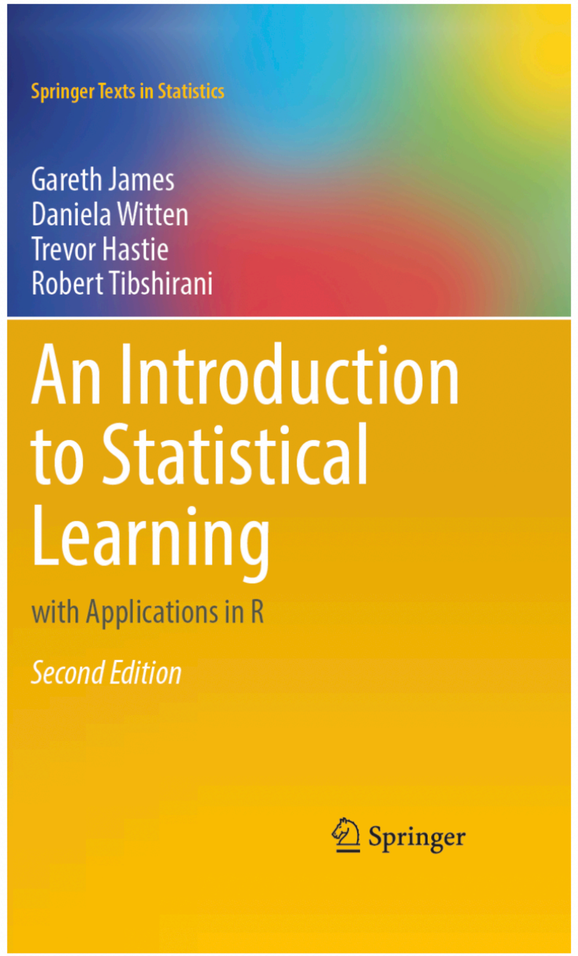
Introducción a la modelización y aprendizaje estadístico

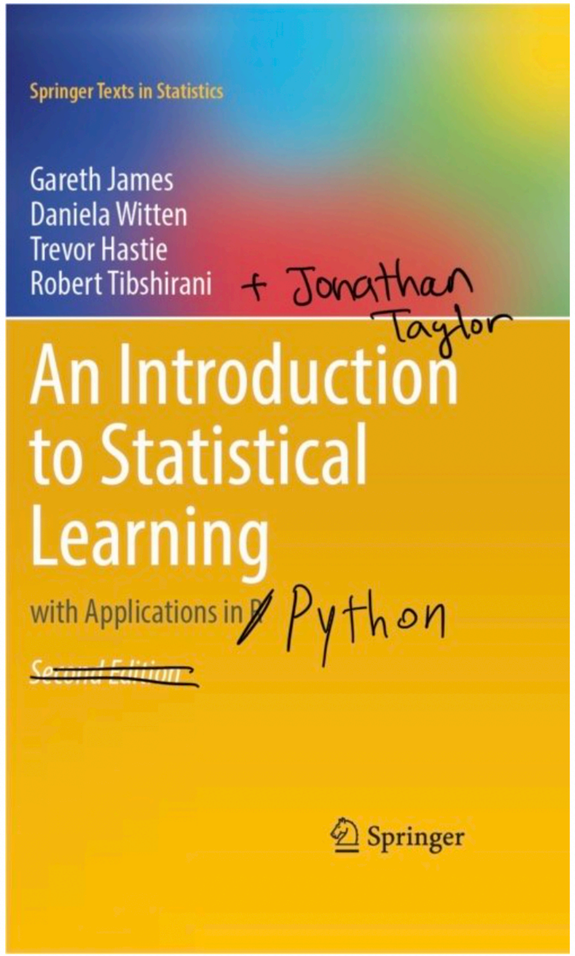
María Inés Fariello

essentially,
all models are wrong,
but some are useful

George E. P. Box

freshspectrum.com



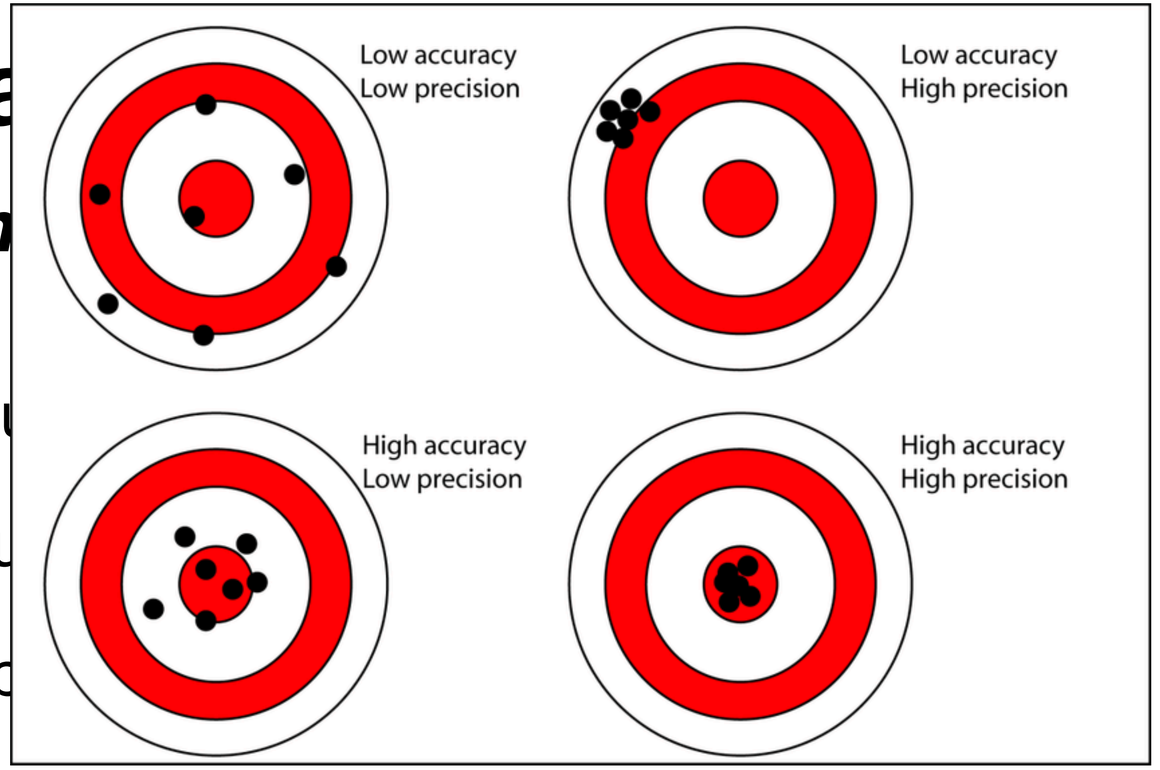


Coming Summer 2023: ISL Python Edition!
Want it sooner? If you are an instructor of a Fall 2023 course and would like to teach out of ISLP, then you can request a pre-print by e-mailing hello@statlearning.com with
(a) the institution name, and
(b) the course number and title.

https://www.statlearning.com/

Aprendizaje estadístico vs a

Satistical learning vs m

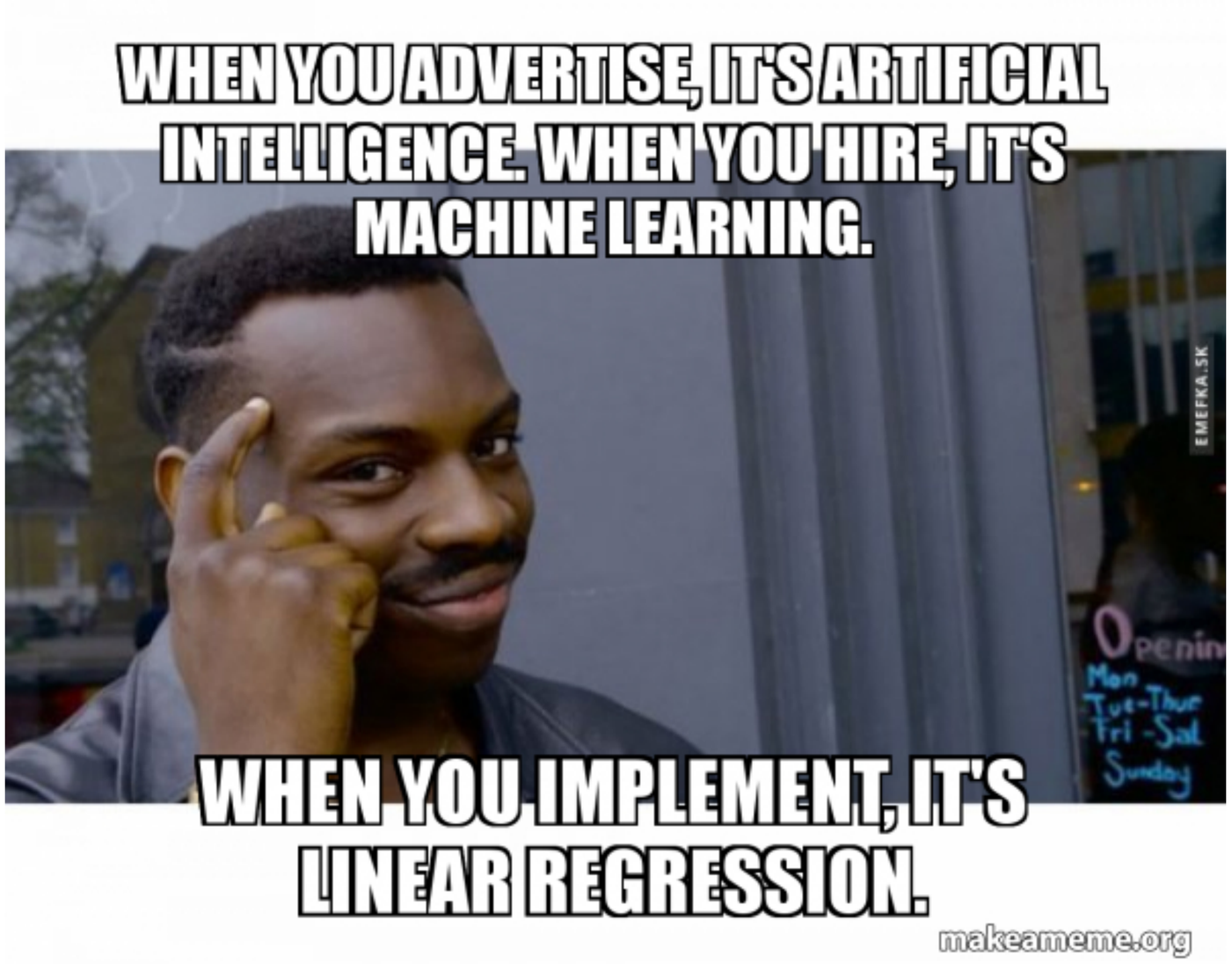


- El aprendizaje automático nació como un su
- El aprendizaje estadístico nació como un su
- Ambos tienen como foco el aprendizaje sup
  - El aprendizaje automático pone el foco en grandes bases de datos y en la exactitud de predicción (*accuracy*).
  - El aprendizaje estadístico pone el foco en los modelos y su interpretabilidad, la precisión (*precision*) e incertidumbre.

La distinción es cada vez más difusa y la “fertilización cruzada” va en aumento.

Aprendizaje estadístico vs aprendizaje automático

Satistical learning vs machine learning



El rol de la estadística en el aprendizaje automático

La estadística juega un rol central en la minería de datos (*data mining*)

- proporcionar fundamentos teóricos para los algoritmos de aprendizaje
- dar herramientas útiles para analizar las propiedades estadísticas de los algoritmos y garantizar su rendimiento (*performance*)
- ayudar a los investigadores a comprender mejor los enfoques, diseñar mejores algoritmos y seleccionar los métodos adecuados para un problema determinado.
- ayudar a tomar una decisión más acertada

Dos culturas del aprendizaje estadístico

Leo Breiman. Statistical Learning: The Two Cultures. Statistical Science. 16 (3): 199-231, 2001.

There are two cultures in the use of statistical modeling to reach conclusions from data.

- One assumes that the data are generated by a given stochastic data model (The Data Modeling Culture).

- The other uses algorithmic models and treats the data mechanism as unknown (The Algorithmic Modeling Culture).

The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems.

Aprendizaje supervisado

- Punto de partida:
    - Medida de resultado Y (también llamada variable dependiente, respuesta, objetivo).
    - Vector de p medidas predictoras X (también llamadas entradas, regresores, covariables, características, variables independientes).
- Clasificación: Y toma valores en un conjunto finito no ordenado (sobrevivió/murió, dígito 0-9, clase de cáncer de la muestra de tejido).
- Regresión: Y es cuantitativa (precio, presión arterial).
- Disponemos de datos de entrenamiento (x1,y1),...,(xN,yN).
- Se trata de observaciones (ejemplos, instancias) de estas medidas.

Objetivos

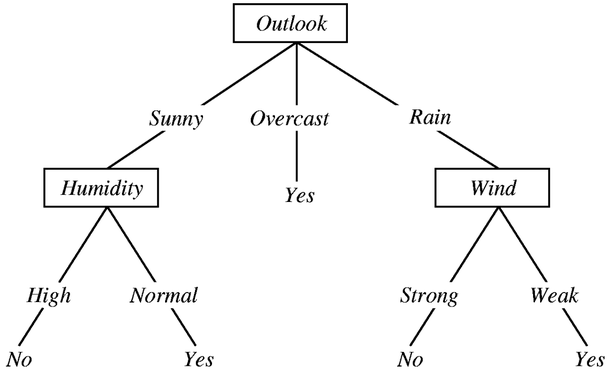
Aprendizaje supervisado

- Predecir con precisión casos de prueba no vistos.
- Entender qué entradas afectan al resultado y cómo.
- Evaluar la calidad de nuestras predicciones e inferencias.

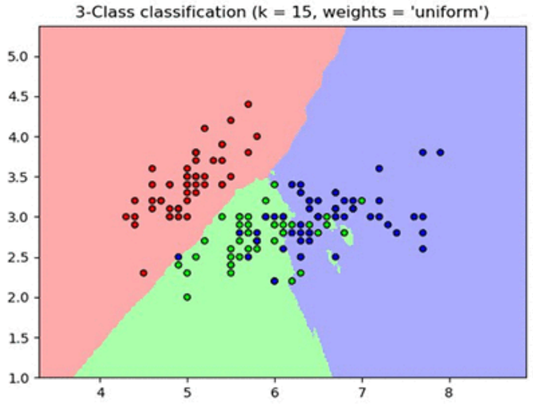
Clasificación

Aprendizaje supervisado

Decision Trees / Random Forest



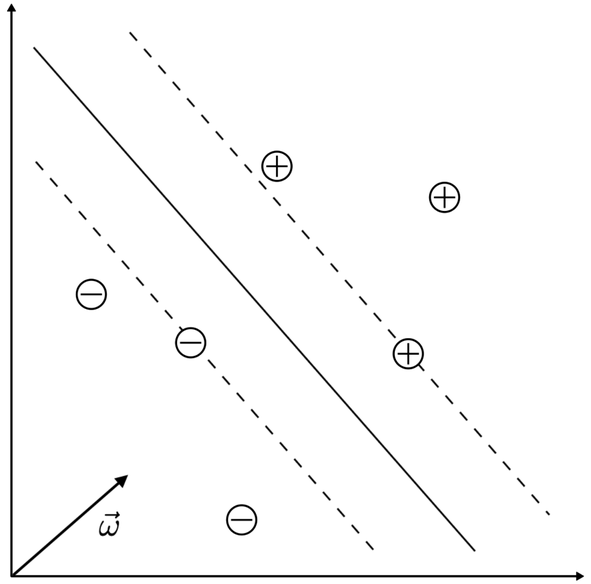
k-Nearest Neighbours



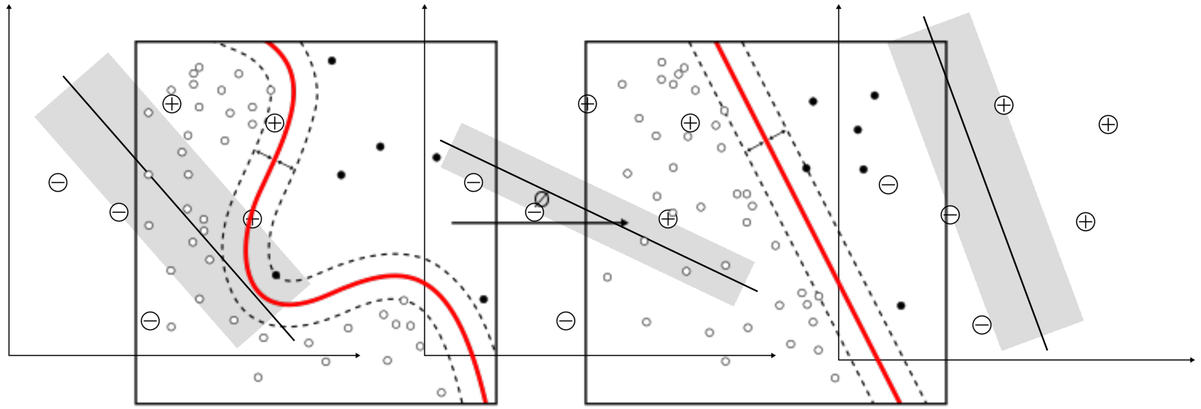
Support Vector Machines (SVM)

Datos linealmente separables

$$\vec{\omega} \cdot \vec{x} + b = 0$$

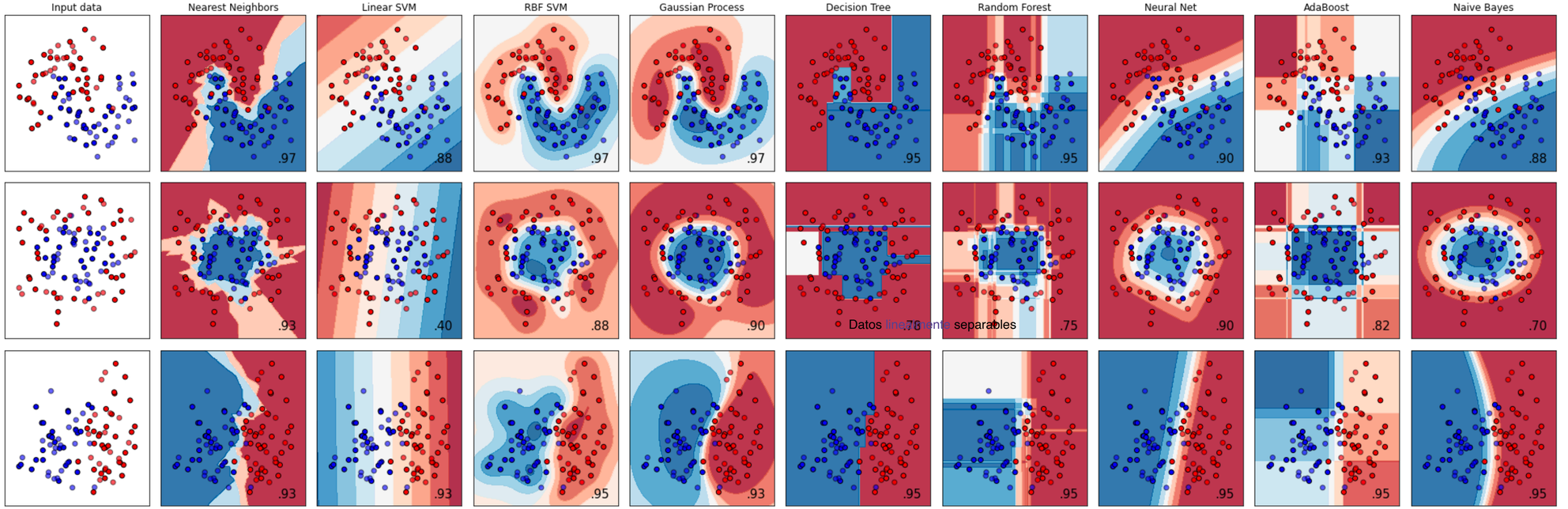


Datos no linealmente separables



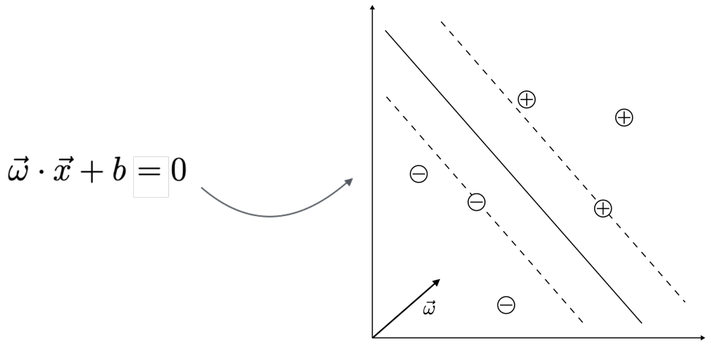
Clasificación

Aprendizaje supervisado

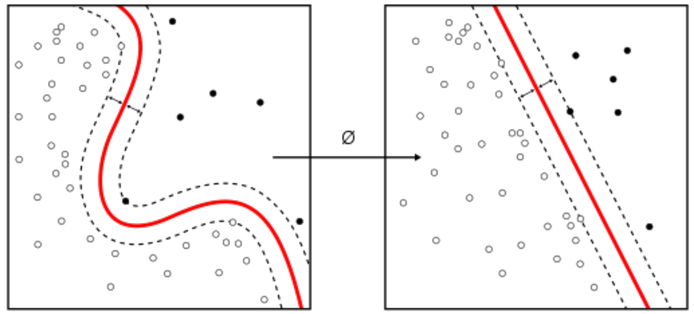


Support Vector Machines

SVM



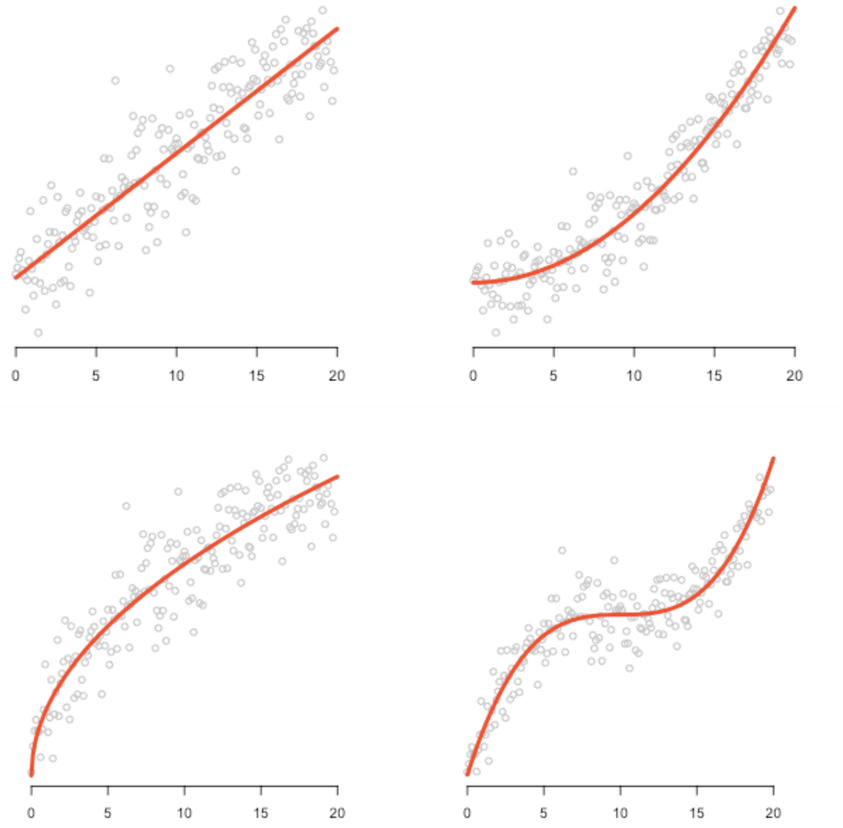
Datos linealmente separables



Datos no linealmente separables

Regresión

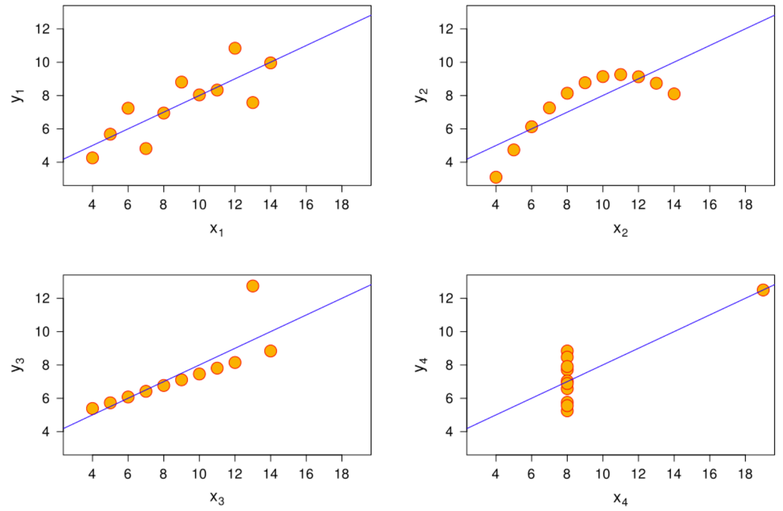
Aprendizaje supervisado



Análisis estadístico para estimar las relaciones (modelos) entre variables.

$$Y_i = f(X_i, \beta) + e_i$$

- Y_i Variable dependiente
- f Función (modelo)
- X_i Variable independiente
- β Parámetros: Usualmente usados para medir la bondad del ajuste del modelo.
- e_i Error



Cuarteto de Anscombe. Cuatro conjuntos de datos que tienen las mismas propiedades estadísticas (media, varianza, correlación, y coeficiente de determinación)

https://medium.freecodecamp.org/learn-how-to-improve-your-linear-models-8294bfa8a731

Filosofía

Aprendizaje supervisado

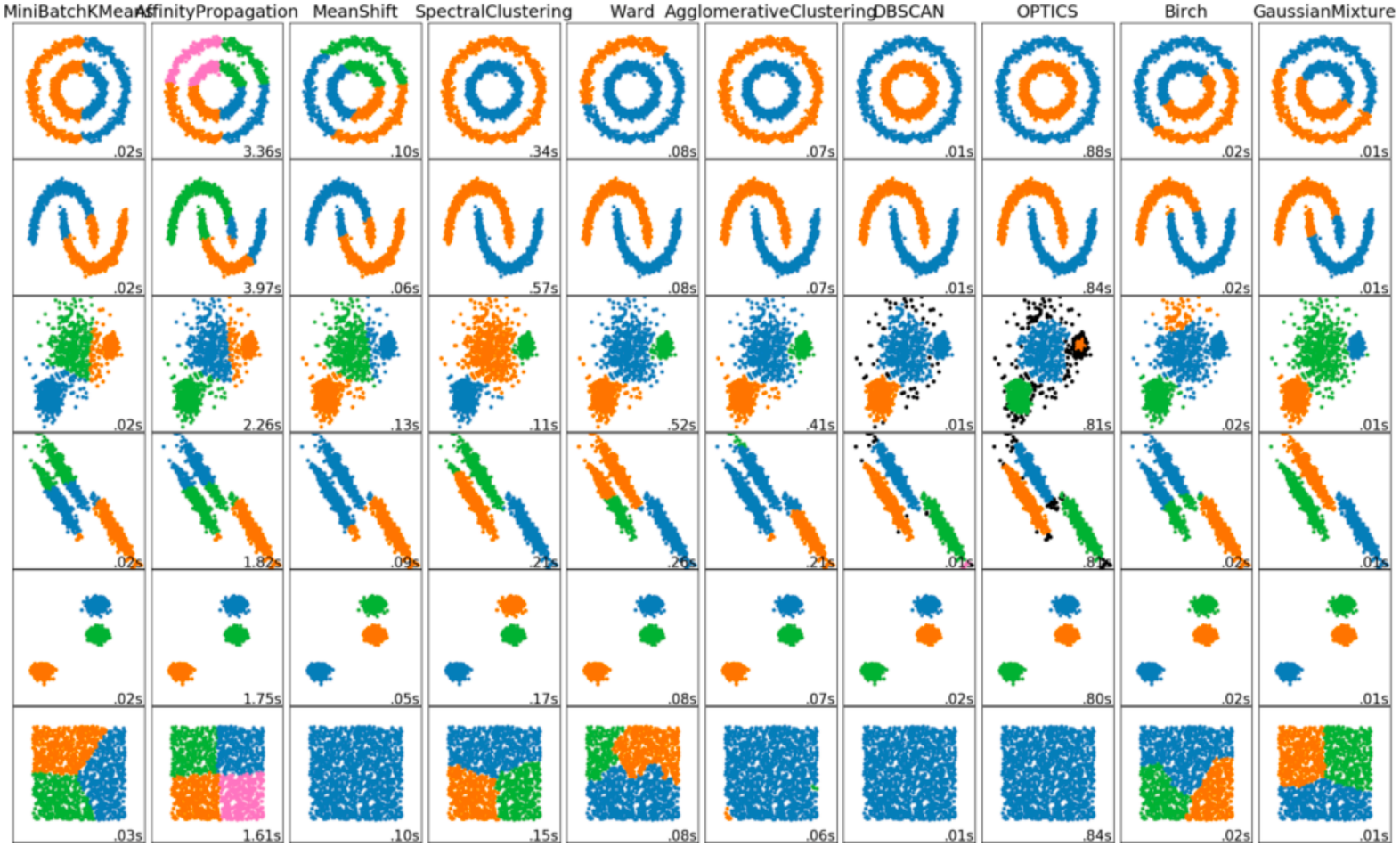
- Comprender las ideas que subyacen a las distintas técnicas para saber cómo y cuándo utilizarlas.
- Entender primero los métodos más sencillos para poder comprender los más sofisticados.
- Evaluar con precisión el rendimiento de un método para saber si funciona bien o mal (los métodos más sencillos suelen funcionar tan bien como los más sofisticados).
- Se trata de un campo de investigación apasionante, con importantes aplicaciones en la ciencia, la industria y las finanzas.
- El aprendizaje estadístico es un ingrediente fundamental en la formación de un científico de datos moderno.

Aprendizaje no supervisado

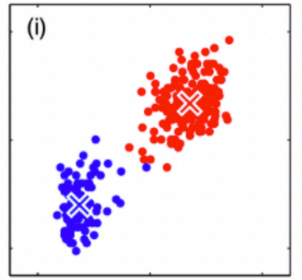
- No hay variable de respuesta, sólo un conjunto de predictores (características) medidos en un conjunto de muestras.
- El objetivo es más difuso:
  - encontrar grupos de muestras que se comporten de forma similar,
  - encontrar características que se comporten de forma similar,
  - encontrar combinaciones lineales de características con la mayor variación.
- Es difícil saber lo bien que se está haciendo.
- Es diferente del aprendizaje supervisado, pero puede ser muy útil como paso previo al aprendizaje supervisado.

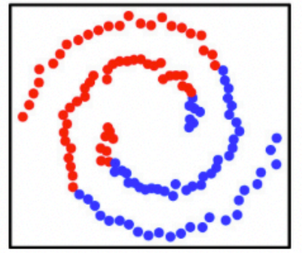
Clustering

Aprendizaje no supervisado

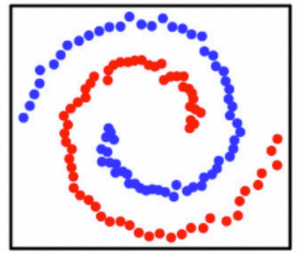


k-means





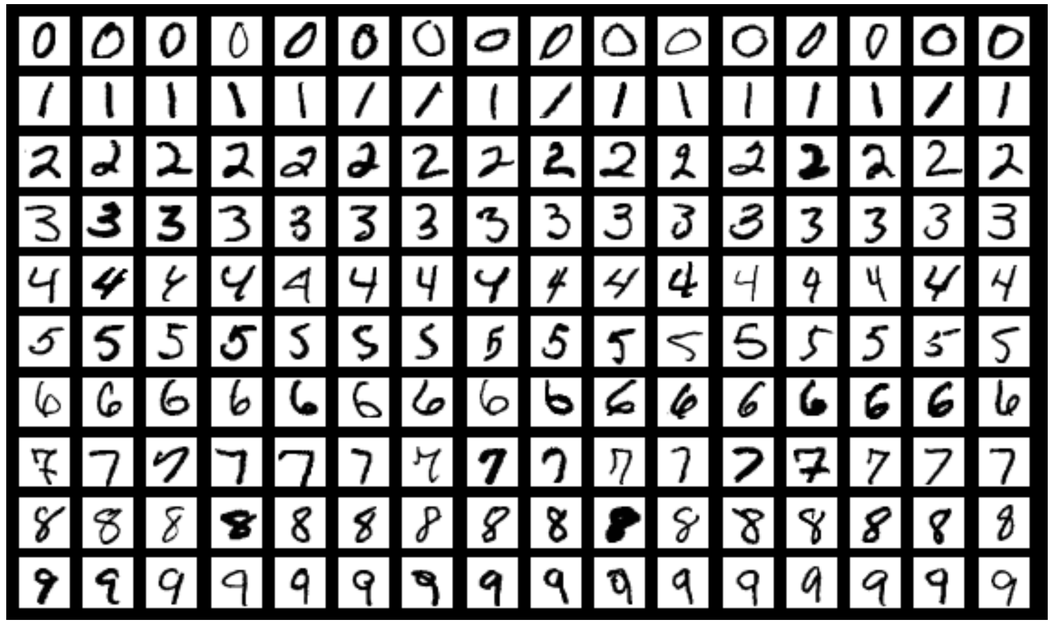
Spectral Clustering

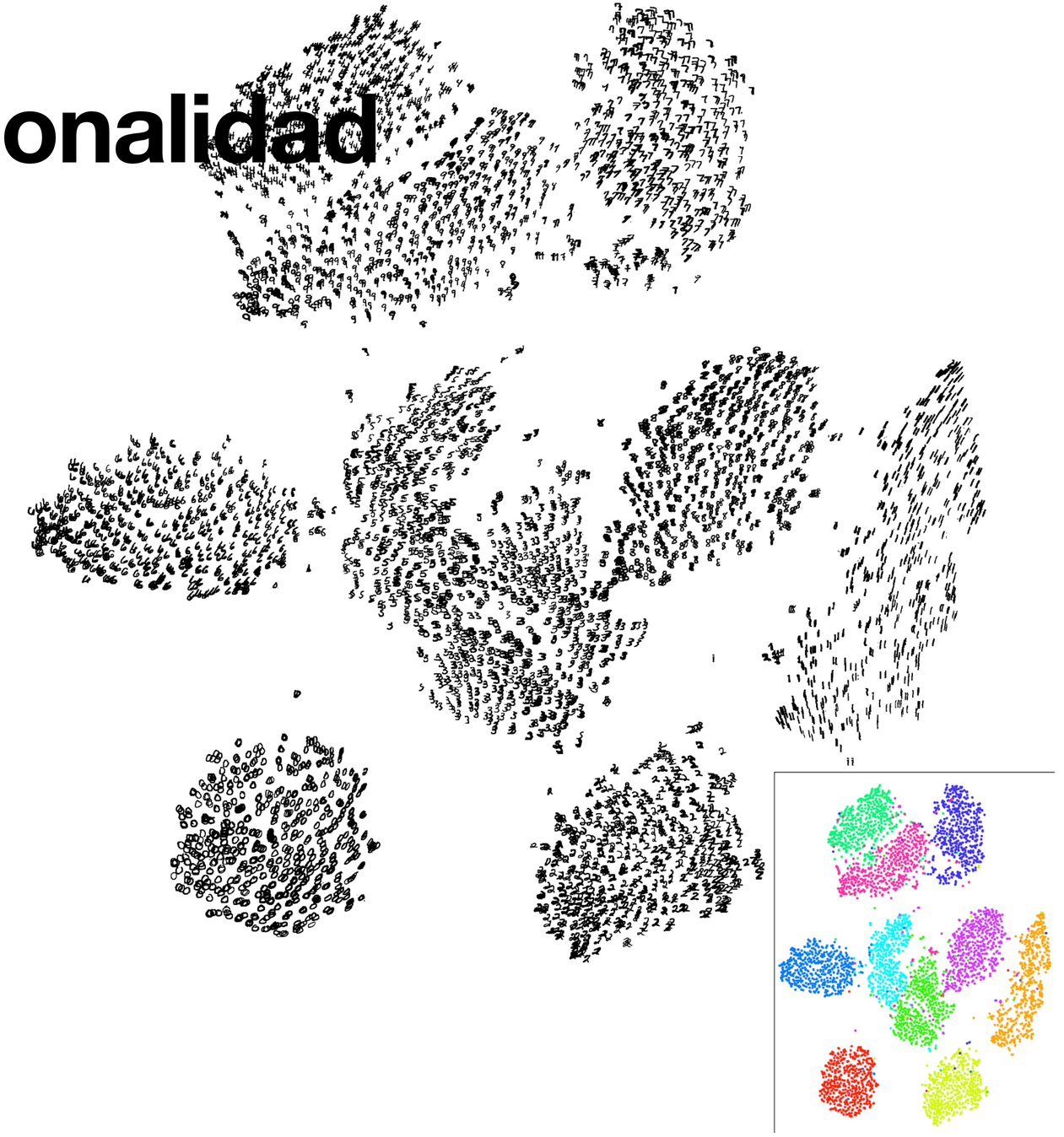


Bishop, C. (2006). *Pattern Recognition and Machine Learning* Springer, New York.

Reducción de dimensionalidad

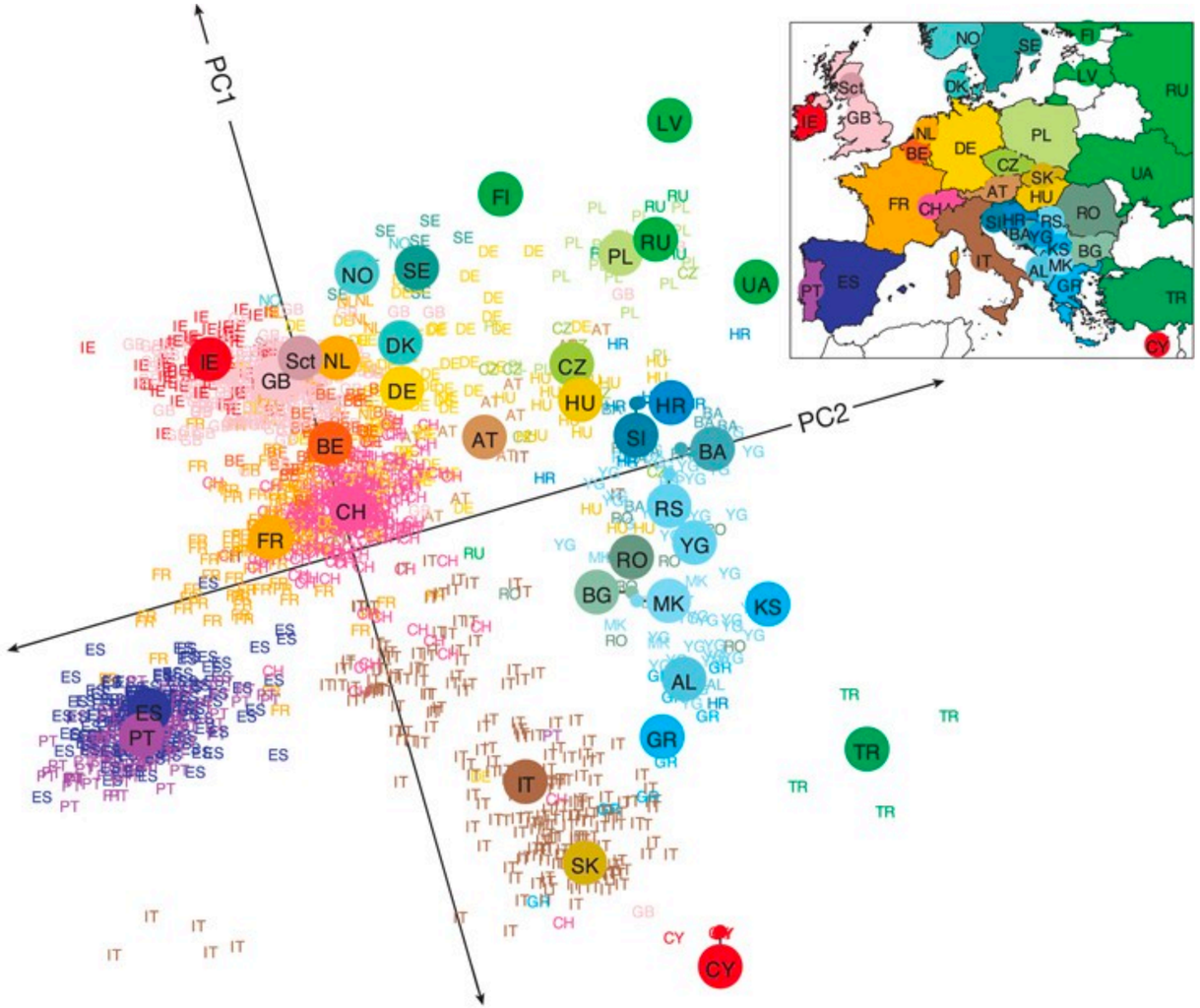
Aprendizaje no supervisado





Reducción de dimensionalidad

Aprendizaje no supervisado



Published: 31 August 2008

Genes mirror geography within Europe

John Novembre ✉, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R. Boyko, Adam Auton, Amit Indap, Karen S. King, Sven Bergmann, Matthew R. Nelson, Matthew Stephens & Carlos D. Bustamante

Nature **456**, 98–101 (2008) | Cite this article