Aprendizaje máquina

Julio Waissman Vilanova

Departamento de Matemáticas
Universidad de Sonora

14 de marzo de 2016

Plan de la presentación

1. ¿Que es el aprendizaje máquina?
2. Todo es sobre generalización
3. Ejemplo ilustrativo: Los vecinos próximos
4. Scikit-learn: Biblioteca de aprendizaje máquina en python

Motivación

- Es muy difícil escribir programas que resuelvan problemas complejos como reconocer objetos en tres dimensiones, desde una perspectiva diferente y con condiciones de luz diferente.
  - No podemos escribir el programa porque no tenemos idea de como procesa nuestro cerebro ese tipo de información.
  - Aunque supieramos como escribirlo, el programa sería terrorífico.
- Es difícil escribir un programa que calcule la probabilidad que una transacción en linea sea fraudulenta.
  - No hay reglas simples y confiables. Es necesario escribir una gran cantidad de reglas.
  - Aún existiendo, el fraude en un proceso dinámico, y hay que estar actualizando las reglas continuamente.

La solución

- En lugar de escribir un programa específico a cada tarea, podemos recolectar muchos ejemplos que especifiquen la respuesta correcta en cada caso.
- Enviamos los datos a un algoritmo de *aprendizaje máquina* el cual nos devuelve un programa.
  - El programa producido es muy diferente a un programa hecho a mano.
  - Si está bien hecho, debe funcionar para nuevos casos con cierto margen de confianza.
  - Si los datos cambian, es posible cambiar el programa.
- Grandes cantidades de datos, y capacidad masiva de computo es más económico actualmente que expertos en tareas específicas.

Definición

A. Samuel (1959)

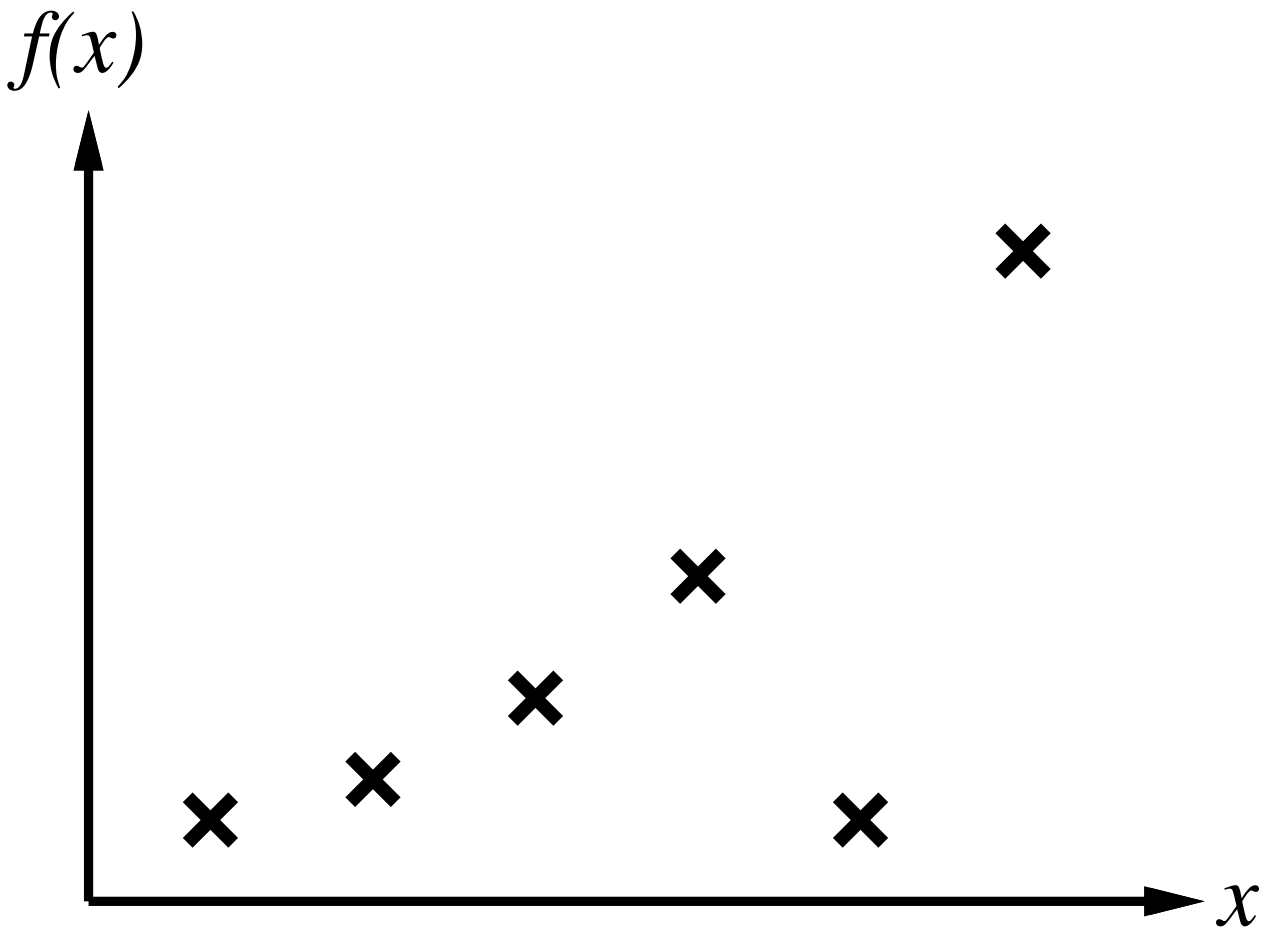
Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.

Definición

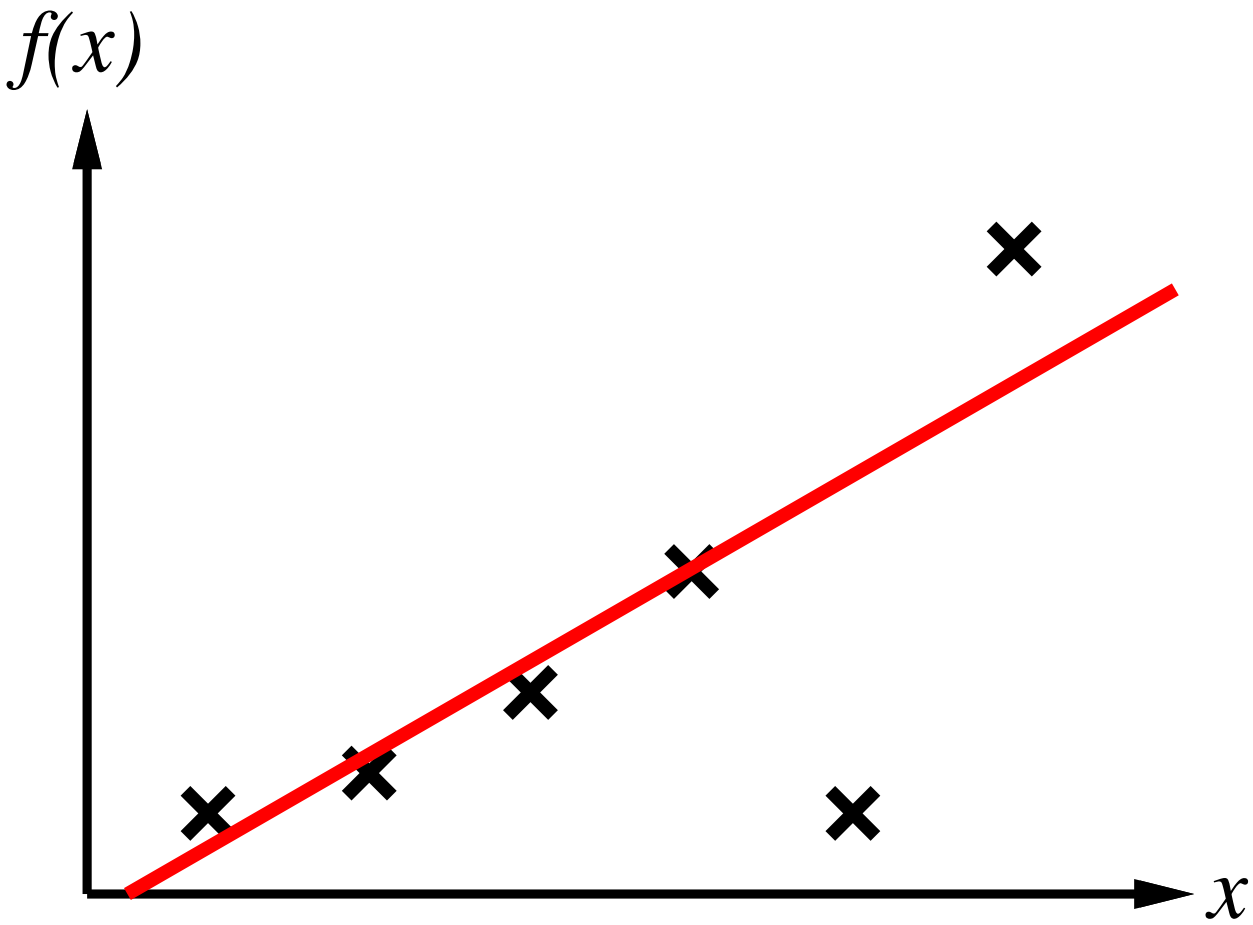
T. Mitchell (1998)

Well–posed Learning Problem: A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.

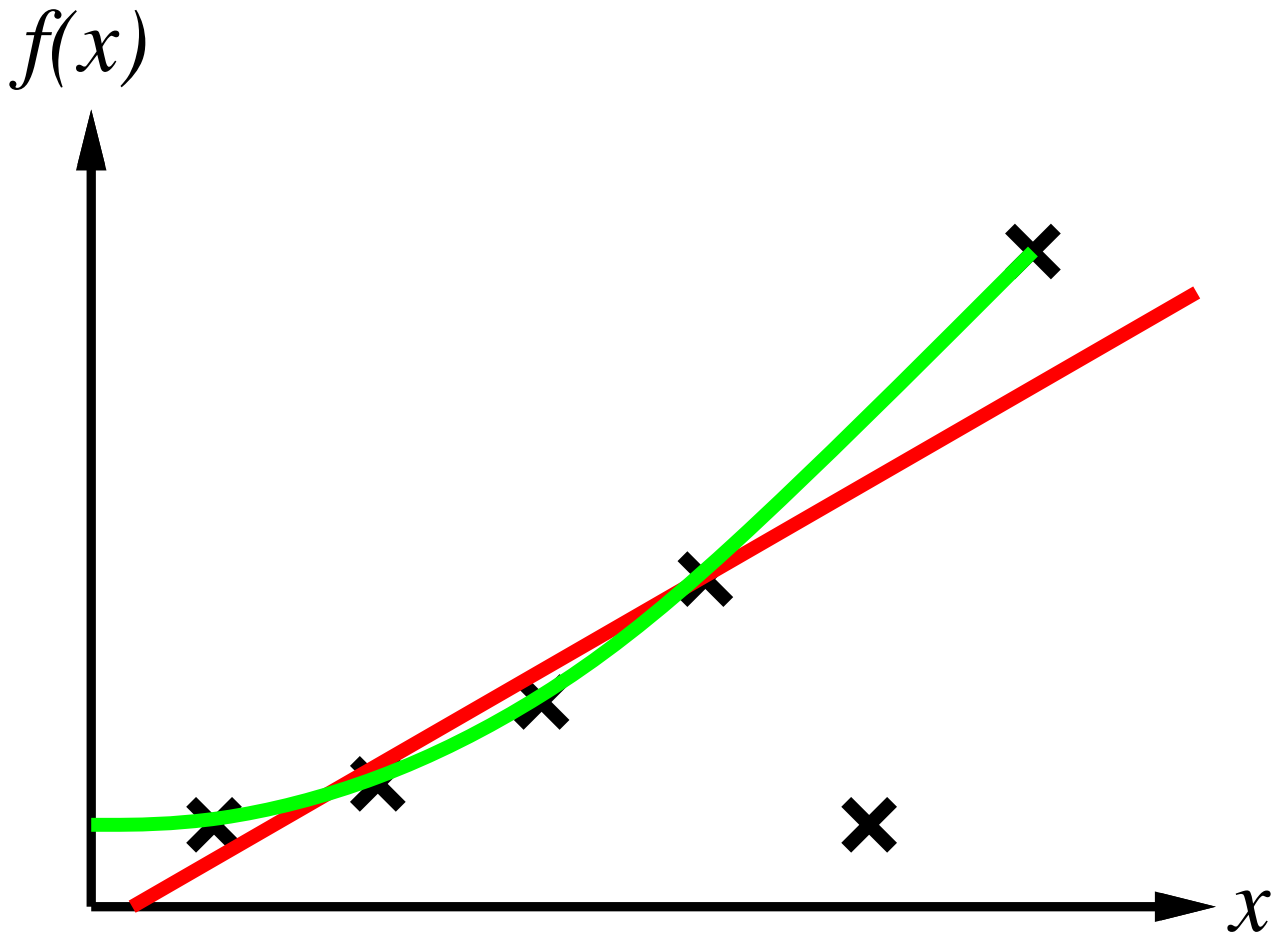
Generalización: ejemplo simple



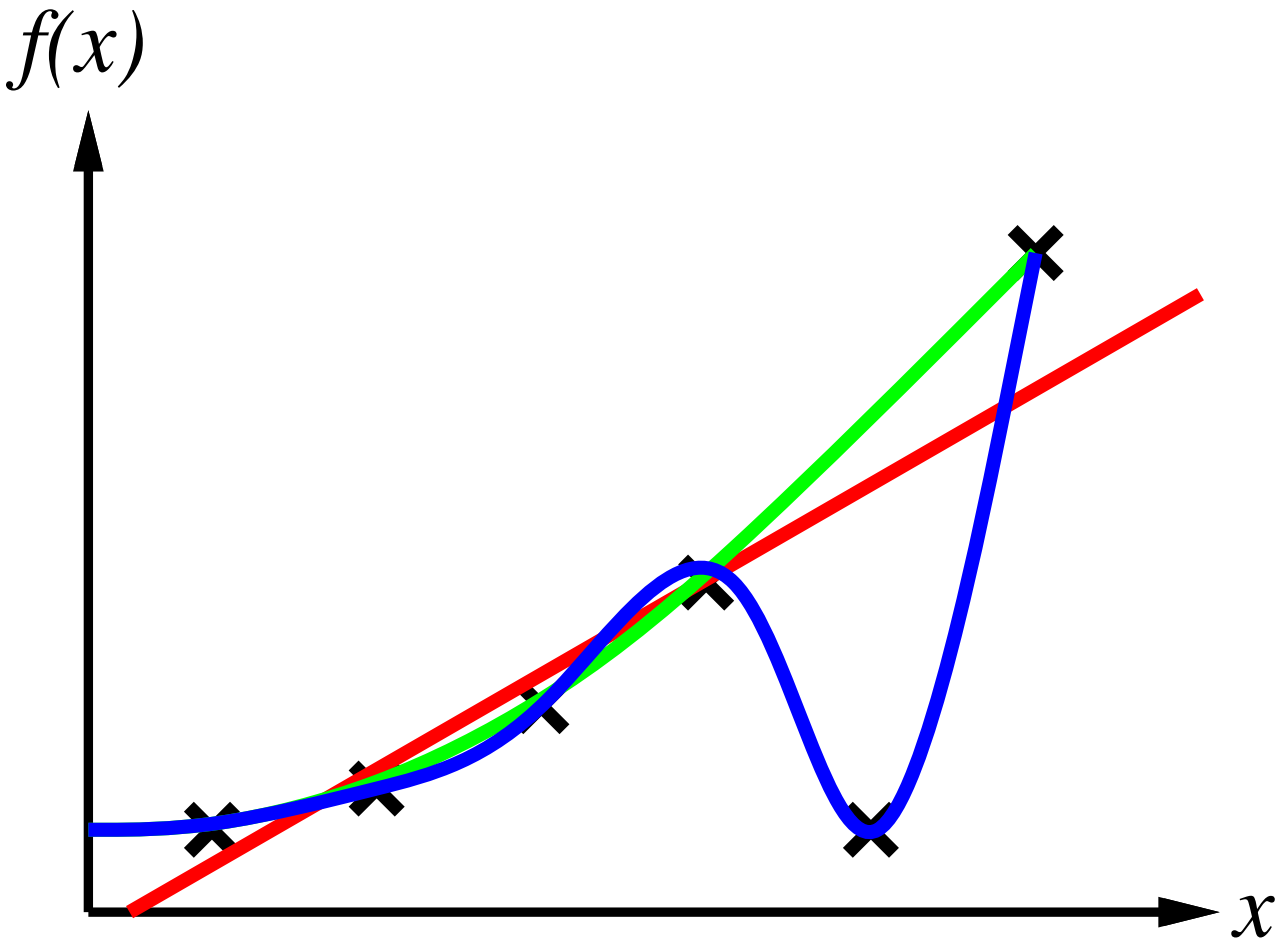
Generalización: ejemplo simple



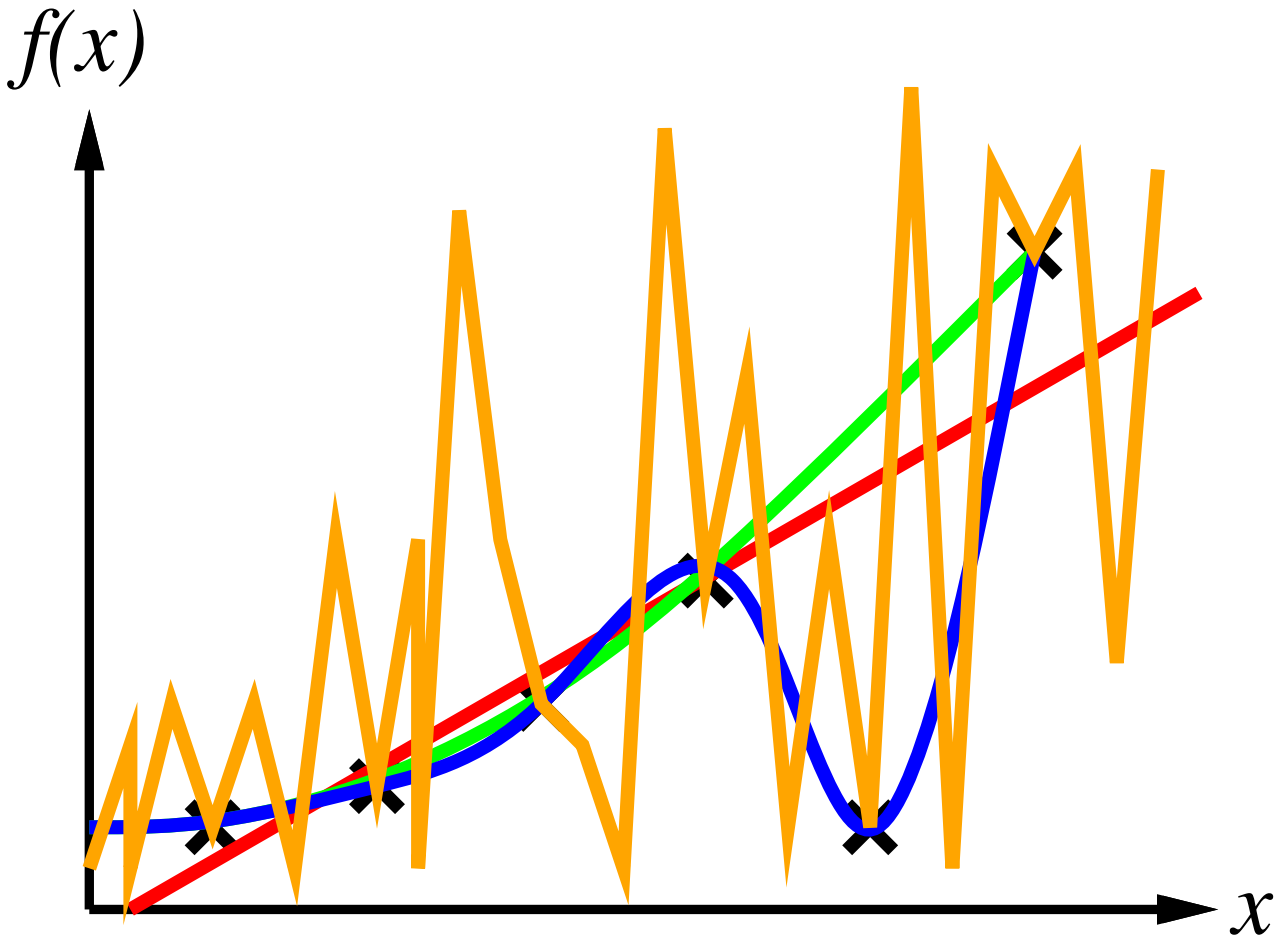
Generalización: ejemplo simple



Generalización: ejemplo simple



Generalización: ejemplo simple



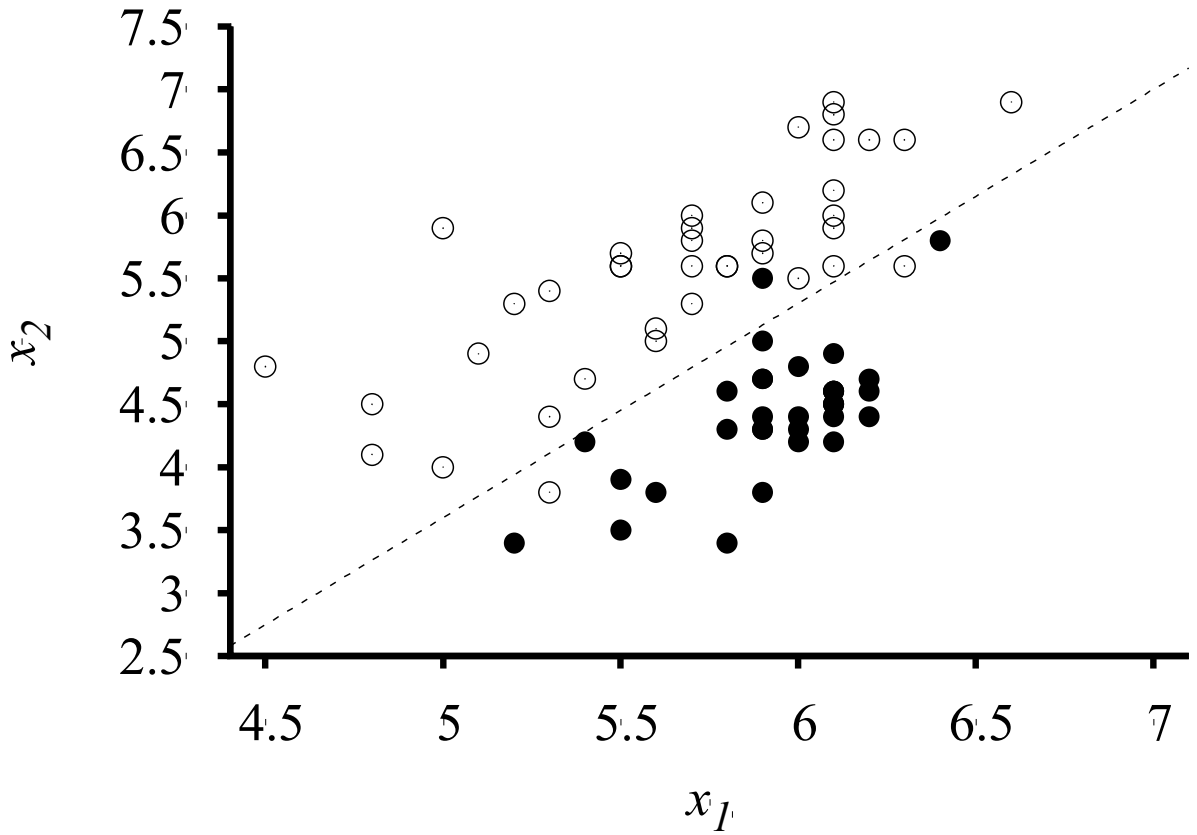
Generalización

- Error en muestra $E_{in} = \frac{1}{N} \sum_{i=1}^{N} e(y^{(i)}, \hat{y}^{(i)})$
- Error fuera de muestra $E_{out} = E_{x \in X}[e(y, \hat{y})]$.
- El aprendizaje existe si y solo si $E_{out} \approx 0$
- Esto solo es posible si
  - $E_{in} \approx 0$
  - $E_{in} \approx E_{out}$

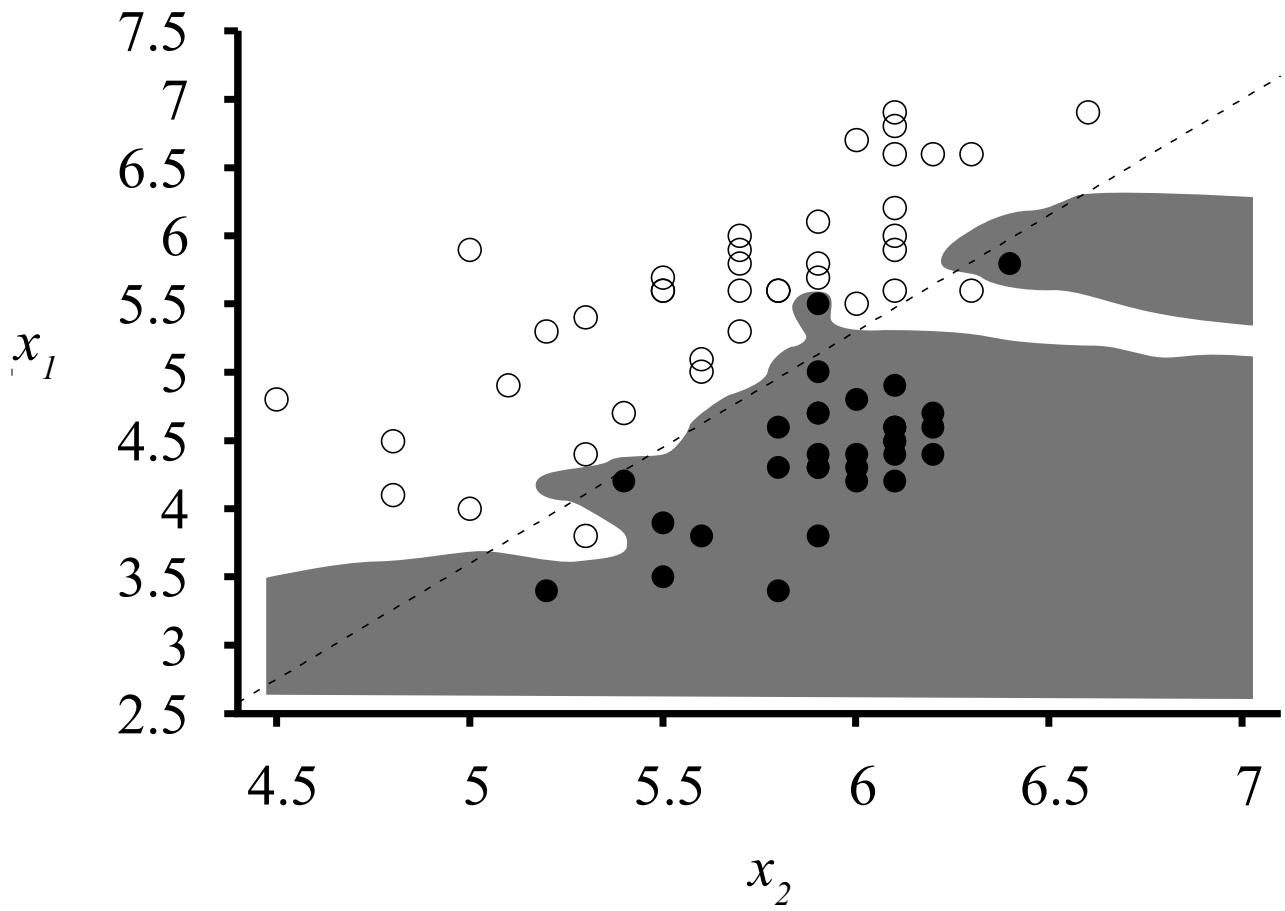
El método del vecino más próximo

- Método no paramétrico
- Requiere guardar todos los datos del conjunto de entrenamiento
- Se calcula una medida de *similaridad* del dato desconocido con *TODOS* los datos del conjunto de entrenamiento
- Se selecciona la clase del dato más cercano
- No hay método más simple conceptualmente

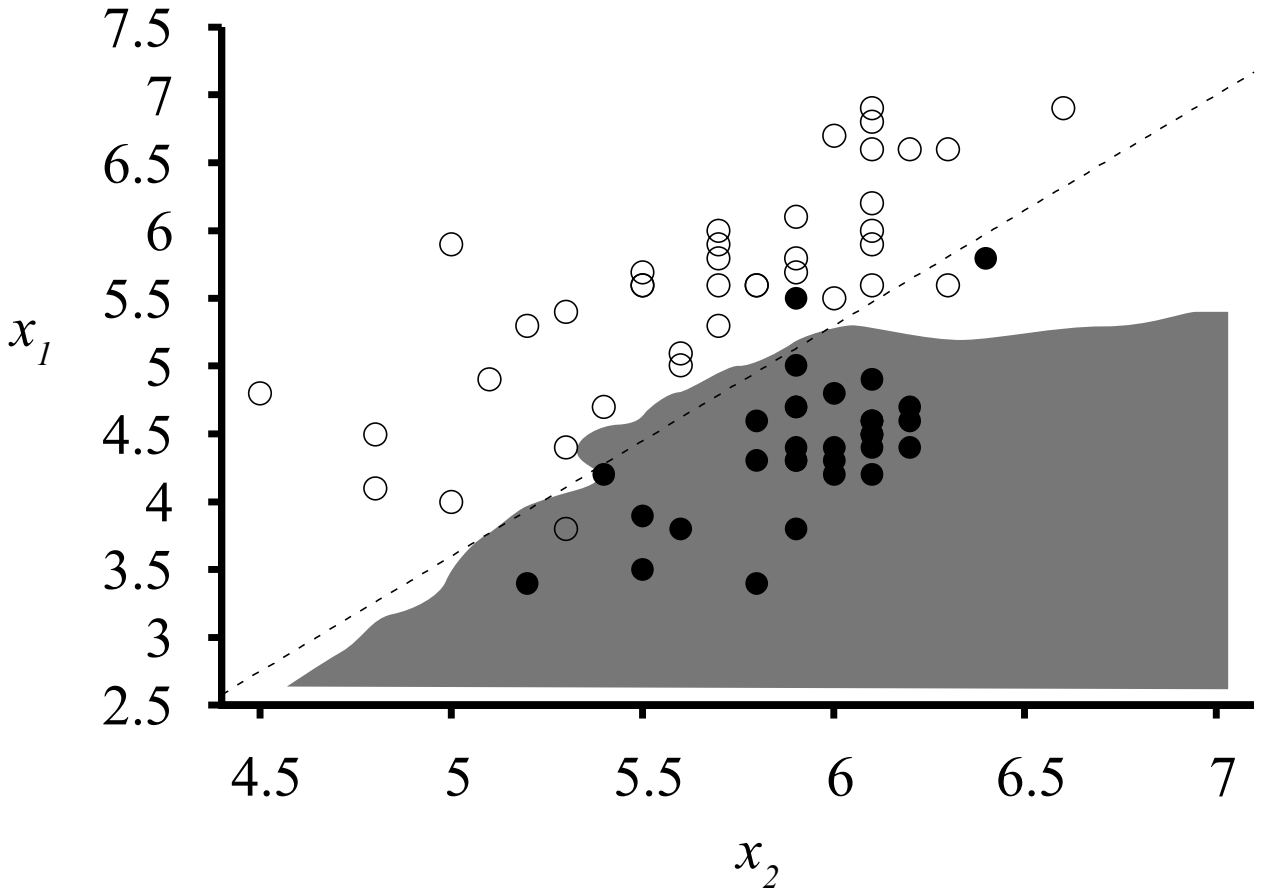
Ejemplo de terremotos y explosiones



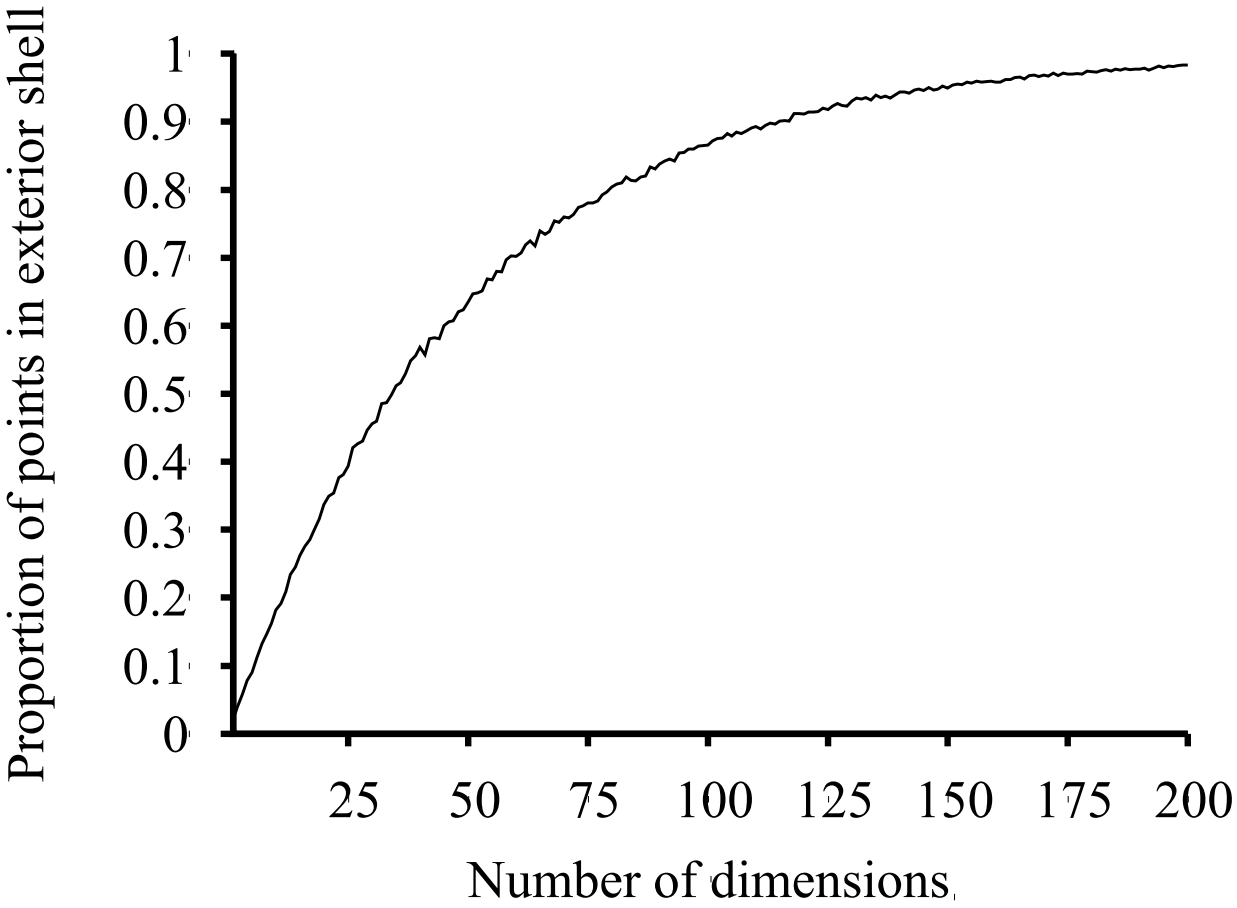
Usando un vecino próximo



Usando cinco vecinos próximos



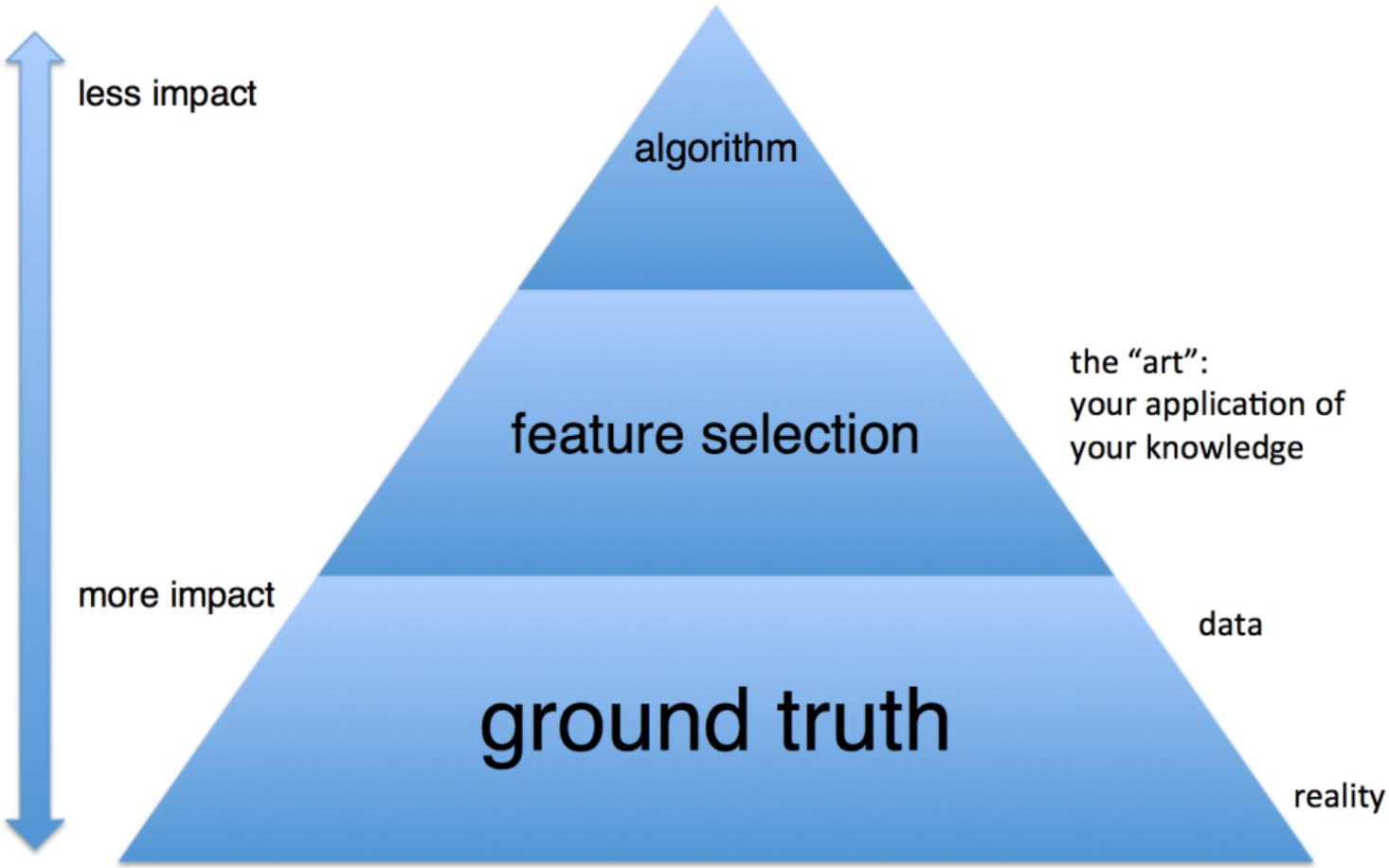
La maldición de la dimensionalidad



Métodos de aprendizaje máquina

- Modelos descriptivos
- Modelos lineales generalizados
- Árboles de decisión
- Redes neuronales
- Métodos de *ensemble*

El algoritmo es lo menos importante

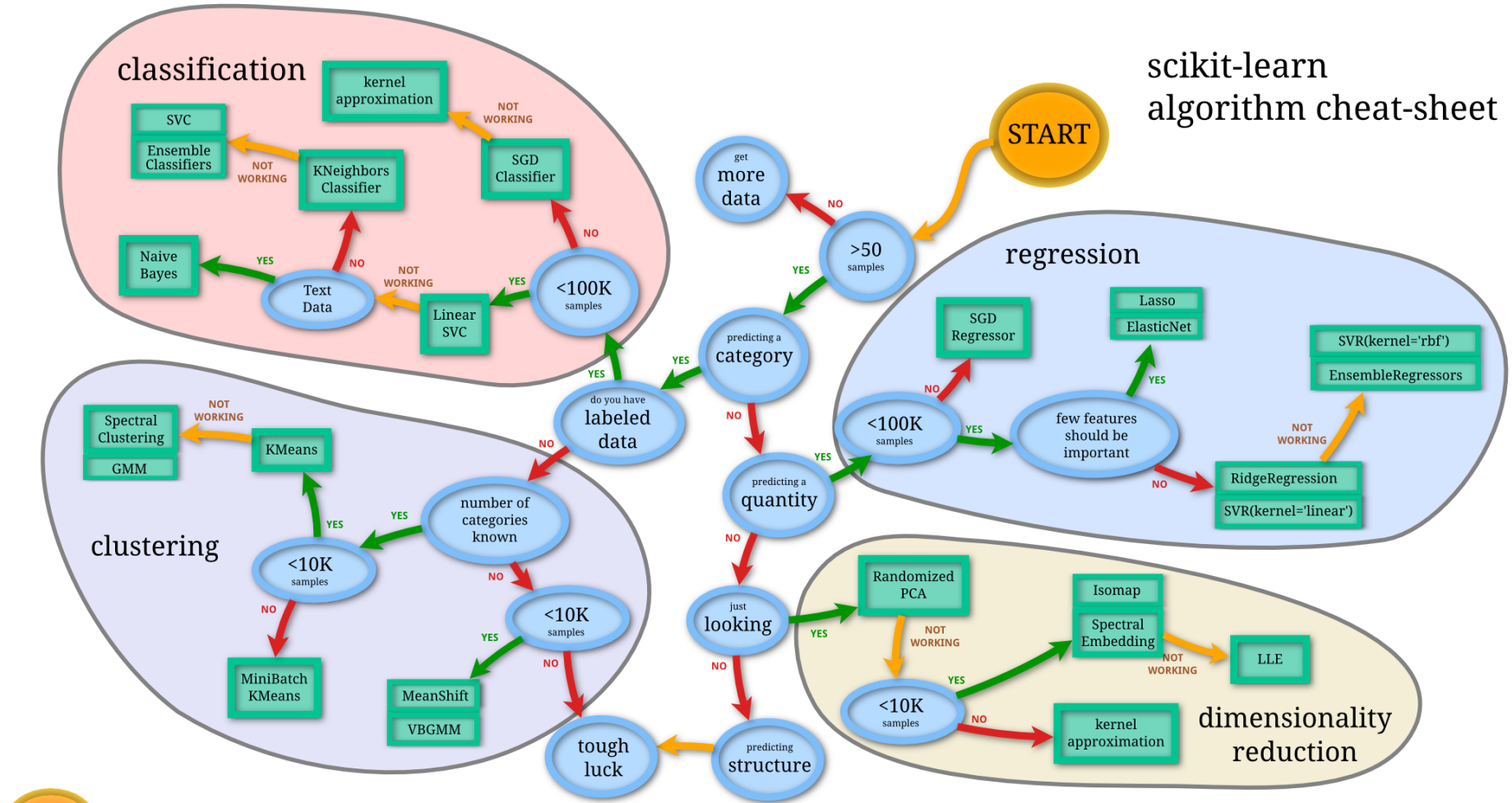


Biblioteca de aprendizaje máquina en python

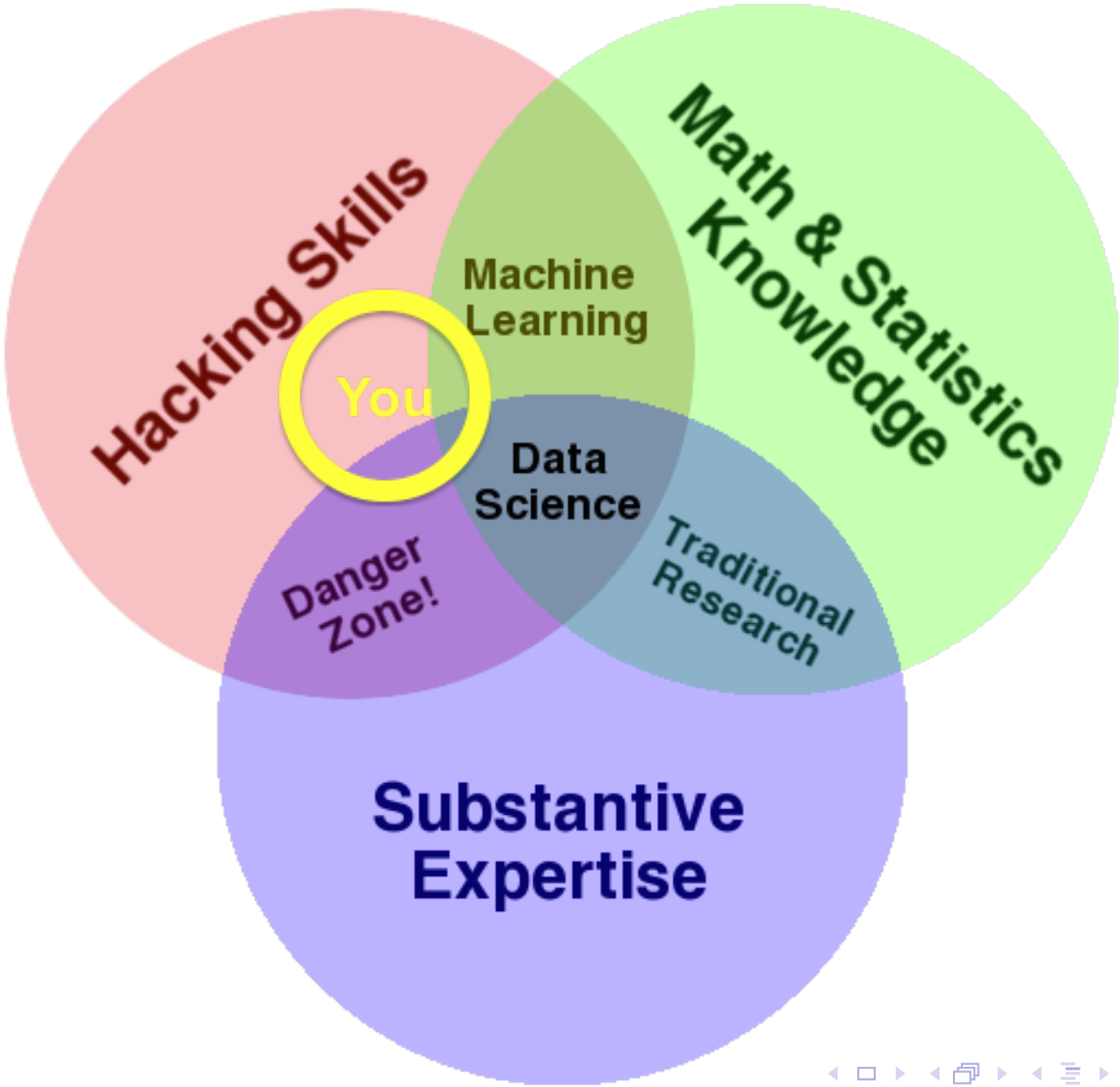
Scikit-learn

- Herramientas simples y eficientes para aprendizaje máquina
- Accesible, reusable, personalizable
- Licencia BSD (usable comercialmente)
- A partir de `numpy` y `matplotlib`, interactua muy bien con *Pandas*

Un acordeón para scikit-learn



¡Cuidado!


Hacking Skills
Math & Statistics Knowledge
Substantive Expertise
Machine Learning
You
Data Science
Danger Zone!
Traditional Research