Preprocesamiento de datos para la industria en visión computarizada

Antes que nada, descarguemos los datos

- Las instrucciones para descargar los datos se encuentran en el repositorio de GitHub de la clase.

- Training data:
  - $ wget https://dl.dropboxusercontent.com/s/w8e5mninhu6qtvo/train_ships.json

- Test data:
  - $ wget https://dl.dropboxusercontent.com/s/w8e5mninhu6qtvo/train_ships.json

Kassandra La Riva

ARISTA

AIFUND

ARISTA

LANDING.AI

TRANSFORM YOUR BUSINESS WITH AI.

We provide the AI brain for manufacturing companies.

¿Qué es el preprocesamiento de datos?

- En el contexto de Machine Learning y Deep Neural Networks, llamamos preprocesamiento a toda manipulación de los datos que ocurre antes de que estos sean usados para el entrenamiento de los modelos.

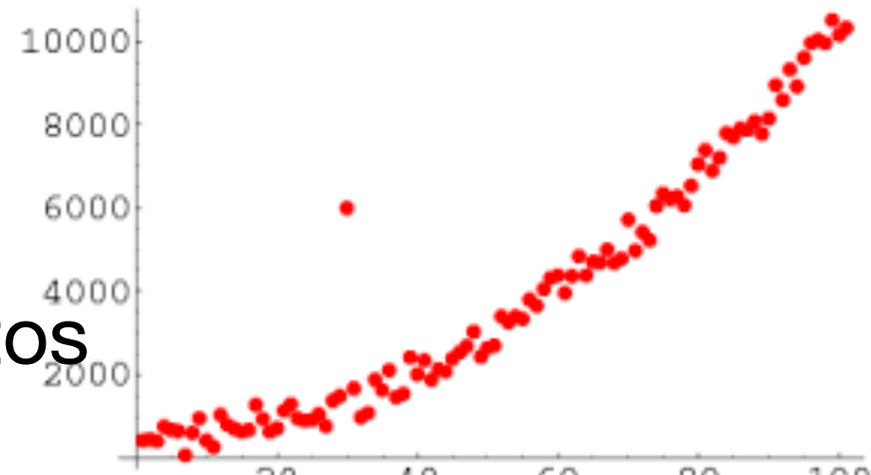
Ayudamos al modelo a entender el problema mediante nuestro conocimiento del problema y nuestra intuición

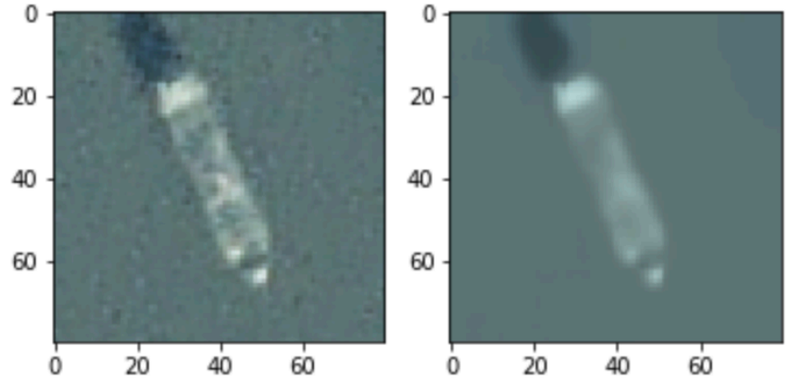
Tres tipos principales de preprocesamiento

1. Limpieza de los datos
2. Extracción de features
3. Incremento de la cantidad de datos

Limpieza de datos

- Encontrar datos con features incompletos
- Encontrar outliers
- Reducir el ruido
- Seleccionar features importantes
- Dar forma a los datos

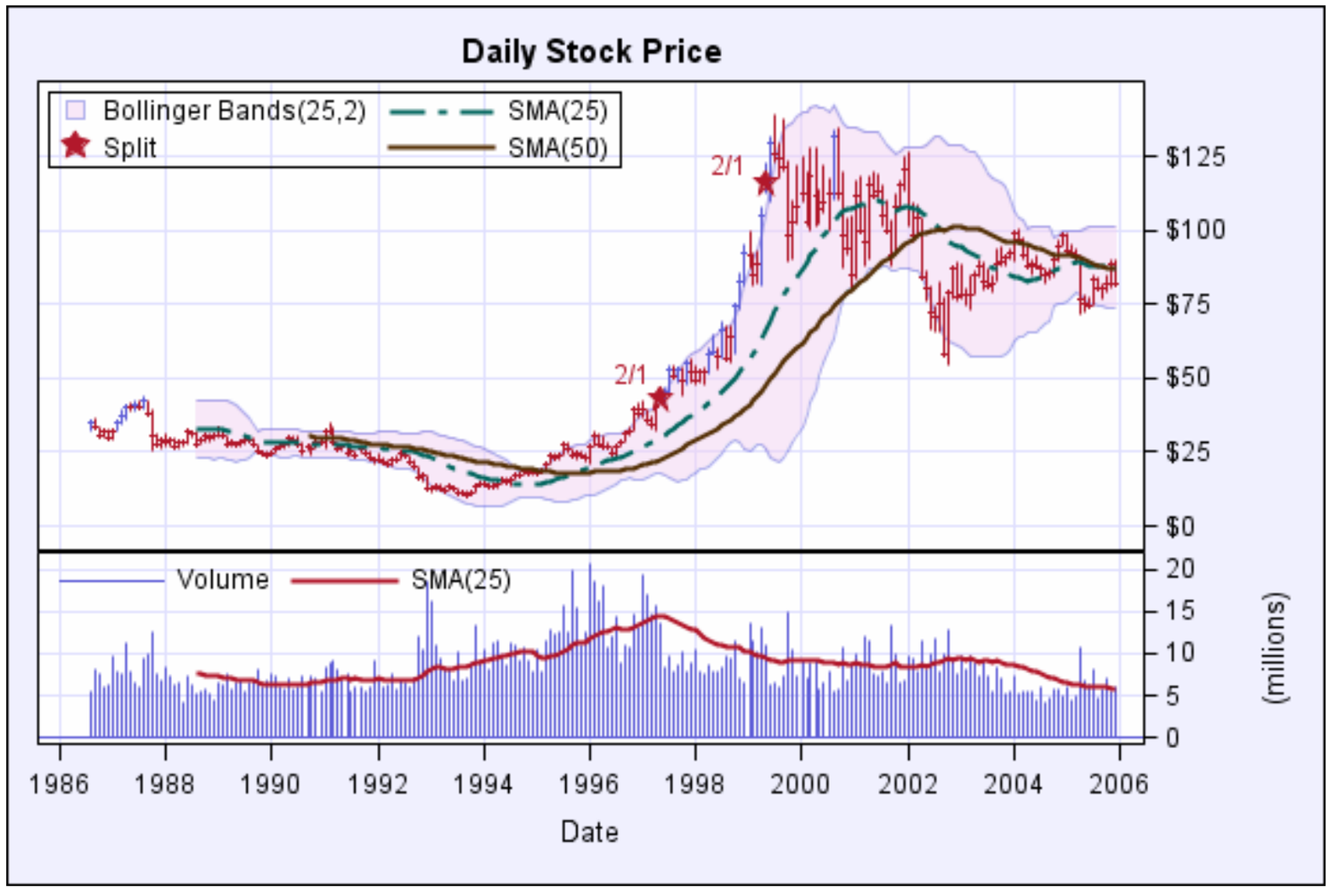




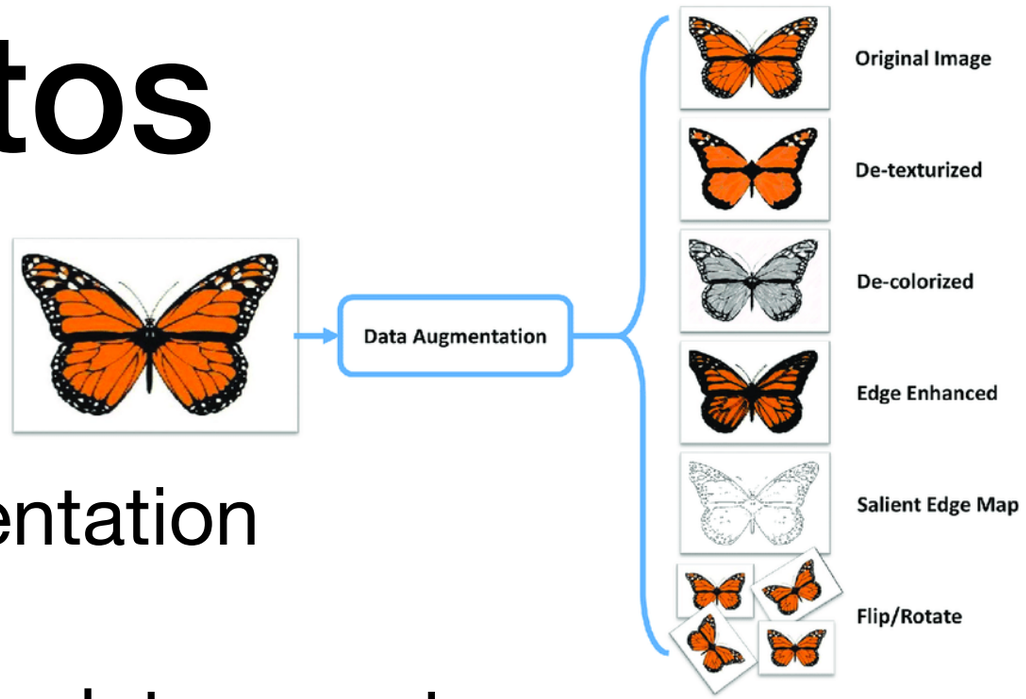
Respondent	Variables			
	A	B	C	D
1	1	2	3	4
2	1	2	3	4
3	4	3	2	1
4	4	3	2	1
5	1	2		1
6		2	2	1
7	1	2	2	
8	1		2	1

Extracción de features

- Extraer métricas de los datos para incluirlas en el entrenamiento



Incremento de la cantidad de datos



- Llamado en inglés: Data augmentation
- Se realizan modificaciones a los datos y estas modificaciones se agregan a los datos de entrenamiento.
- A diferencia de los otros dos tipos de preprocesamiento, estos transformaciones no se aplican también a los datos de prueba.

Beneficios del preprocesamiento de datos

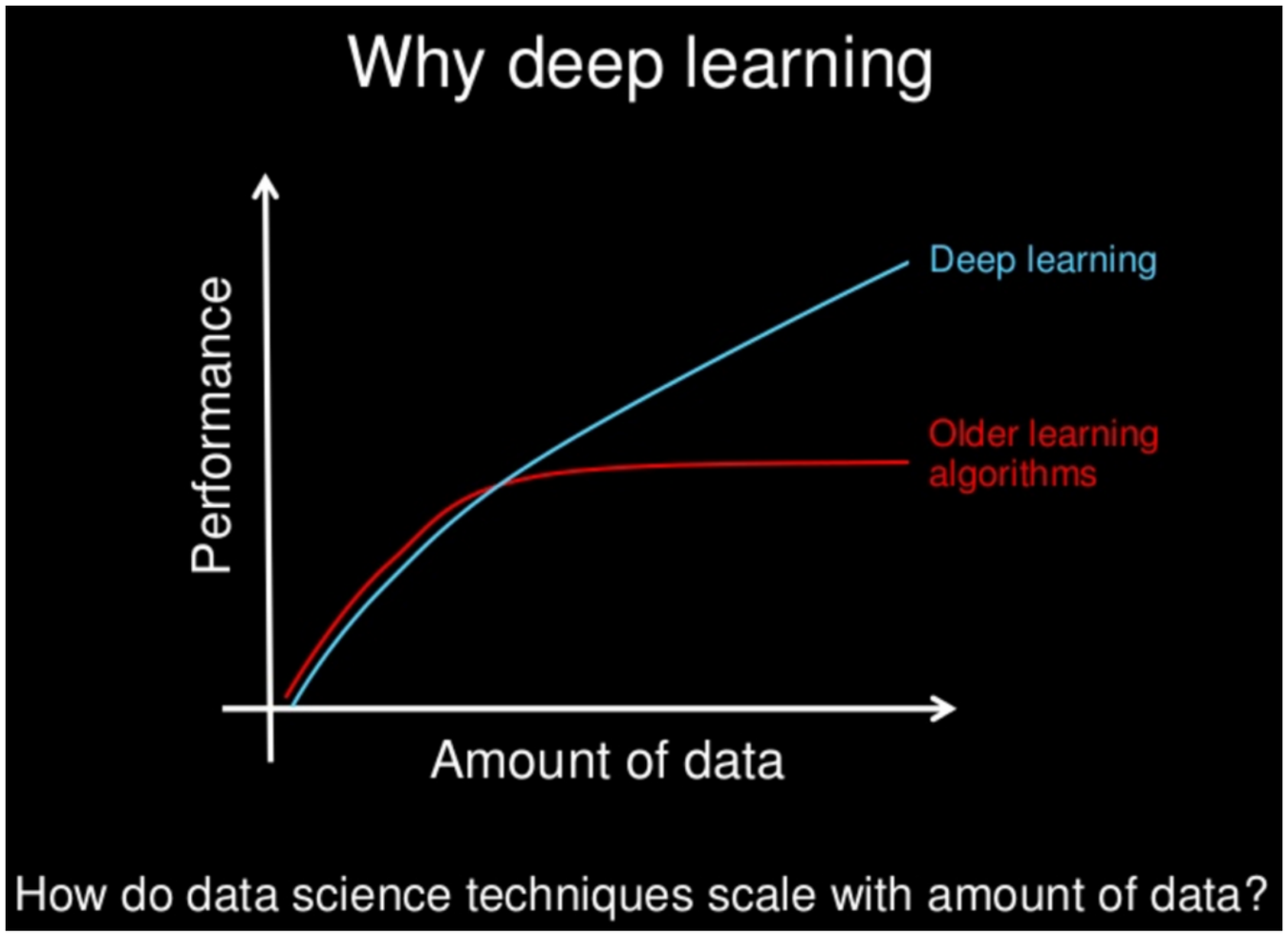
Beneficios durante el entrenamiento

- El preprocesamiento puede hacer más rápido el entrenamiento cuando se reduce la complejidad de los datos mediante preprocesamiento.

- El entrenamiento puede hacerse más rápido ya que mediante el preprocesamiento le estamos incluyendo nuestro conocimiento sobre las cualidades de los datos que importan.

- Data augmentation nos ayuda a generalizar el conocimiento del modelo entrenado.

Beneficios durante el entrenamiento



Beneficios durante el despliegue de los modelos

- Cuando hacemos que un modelo sea menos complejo gracias a que hicimos un buen preprocesamiento de los datos, este modelo ocupa menos espacio en disco y cuando se utiliza el modelo para inferencia, el número de operaciones (tiempo de procesamiento) es mucho menor.

Ejemplos de preprocesamiento de datos para visión computarizada

Jupyter Notebook

https://github.com/camiloaz/udea-preprocessing/blob/master/preprocessing.ipynb

AIFUND

ARISTA

Actividad en grupos

CONCURSO DE PREPROCESAMIENTPO

https://github.com/camiloaz/udea-preprocessing

AIFUND

ARISTA