

# DATA COLLECTION

## Web Scrapping

Juan Ignacio Elias  
Lucas Scarano

# ¿QUÉ ES SCRAPING?

Es la automatización del proceso de **extraer información** de una fuente de datos no estructurados para poder ordenarla y analizarla en una base de datos o en hojas de cálculo de forma más sencilla.

- Sitios web
- Aplicaciones
- Reseñas
- Tablas
- Imágenes
- Fuentes de audio

# ¿POR QUÉ HACER SCRAPING?

- Formar **bases de datos propias** con información de un sitio web.
- **Manipular información** para conseguir nuevos datos.
- En el ámbito profesional puede ser utilizado para obtener **ventajas ante la competencia**.



# SCRAPING VS API

	SCRAPING	API
Ventajas	<ul style="list-style-type: none"><li>- Mayor libertad</li><li>- No requiere autenticación</li><li>- Gratis</li><li>- Fácil de usar</li><li>- Anonimidad</li></ul>	<ul style="list-style-type: none"><li>- Reglas predefinidas</li><li>- No cambian en el tiempo</li><li>- Legal y seguro</li></ul>
Desventajas	<ul style="list-style-type: none"><li>- Tráfico más limitado</li><li>- Puede violar políticas del sitio</li><li>- Los sitios web cambian frecuentemente</li></ul>	<ul style="list-style-type: none"><li>- Tiene que estar disponible</li><li>- Pueden ser pagas</li><li>- Mas difícil de usar y comenzar</li></ul>

# PRINCIPALES SCRAPERS



Selenium

2002

BeautifulSoup

2004



Scrapy

2008



# Selenium

Framework y ecosistema de automatización para navegadores basado en la infraestructura para la especificación W3C WebDriver. Es multiplataforma y compatible con múltiples lenguajes.

Se usa principalmente para realizar testing automático pero puede adaptarse para hacer recolección de datos.

No es la principal herramienta usada pero si se tiene experiencia se puede usar para un proyecto de tamaño chico.



# BeautifulSoup

Librería de Python para **extraer datos** de archivos HTML y XML.

Posee formas simples de navegar, buscar y modificar los documentos y extraer lo necesario. No se requiere mucho código para escribir una aplicación.

Realiza el análisis de forma transversal de lo que se requiera. Se le puede pedir que encuentre todos los **"links"** con alguna clase específica o el encabezado de una tabla que tenga texto en **negrita**.



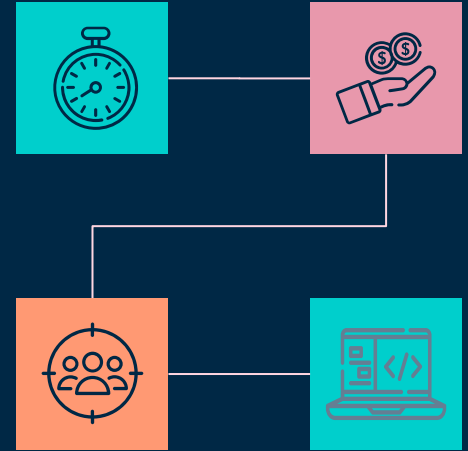


# Scrapy

Es un framework de scraping y crawling **rápido** y de **alto nivel** usado para escanear sitios web y extraer datos estructurados de las páginas.




Puede ser utilizado tanto como para minería de datos, monitoreo y testing automatizado.

Es **open source** y multiplataforma, se puede escribir en Python y correrse en cualquier sistema operativo.





# COMPARATIVA

	 BeautifulSoup	 Scrapy	 Selenium
+	<ul style="list-style-type: none"><li>- Fácil de usar</li><li>- Fácil de aprender y dominar</li></ul>	<ul style="list-style-type: none"><li>- Eficiente</li><li>- Portable</li></ul>	<ul style="list-style-type: none"><li>- Versátil</li><li>- Funciona bien con JavaScript</li></ul>
-	<ul style="list-style-type: none"><li>- Depende de otros paquetes</li><li>- Ineficiente</li></ul>	<ul style="list-style-type: none"><li>- Difícil de usar</li></ul>	<ul style="list-style-type: none"><li>- No es un Web Scraper en si</li><li>- Ineficiente</li></ul>

The background is a dark blue field decorated with various geometric elements. It includes numerous small squares in shades of teal, orange, and pink, some of which are solid while others are outlined. Thin, light-colored vertical lines of varying lengths are scattered across the composition, creating a modern, minimalist aesthetic.

# GRACIAS!