



WEB SCRAPING

Y LAS BASES DE DATOS OFUSCADAS

Abril 2024

Capgemini 

QUÉ APRENDERÁS HOY

Headless Web scraping

Selenium



¿QUIÉN SOY?



Pepe Fabra Valverde

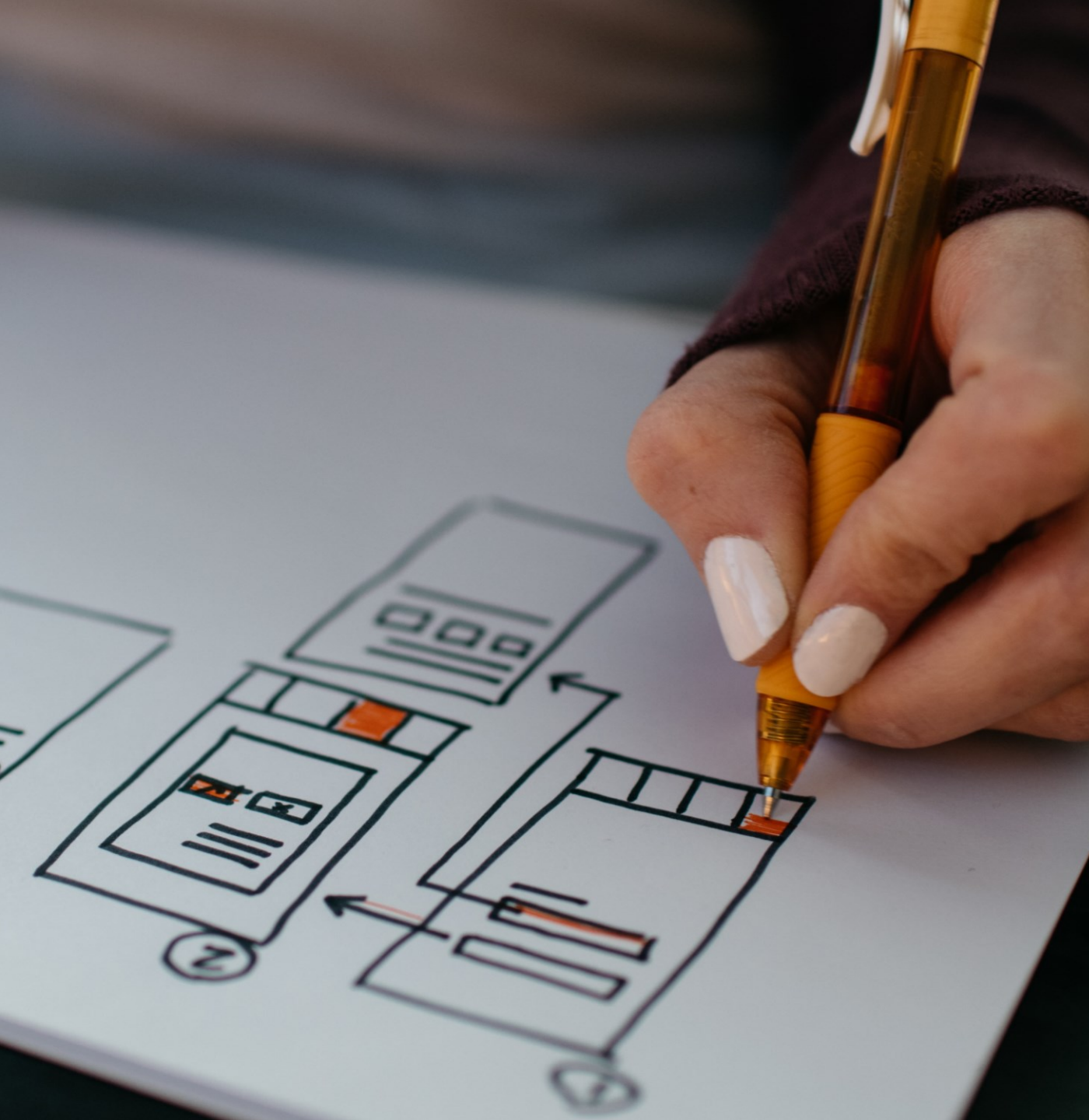
Líder y Arquitecto de Front

Haciendo web scraping desde mis inicios, a nivel personal, académico y ligeramente profesional

AVISO PARA NAVEGANTES



- Los recursos se compartirán al final de la sesión
- La sesión **se grabará**, aparecerá el enlace cuando termine
- Es parte de una serie de charlas acerca de Web Scraping



HEADLESS WEB SCRAPING

HEADLESS WEB SCRAPING



Qué es un Headless Browser

HEADLESS WEB SCRAPING



Qué es un Headless Browser

- Un navegador que puede funcionar programáticamente (sin usuario)

HEADLESS WEB SCRAPING



Qué es un Headless Browser

- Un navegador que puede funcionar programáticamente (sin usuario)

Ejemplos

- Chromium (el engine)
- Selenium

QUÉ NOS OFRECE



Conexión ininterrumpida

QUÉ NOS OFRECE



Conexión ininterrumpida

- No es un HTTP GET, es una conexión que no se cierra hasta que queramos

Simular ser un usuario

QUÉ NOS OFRECE



Conexión ininterrumpida

- No es un HTTP GET, es una conexión que no se cierra hasta que queramos

Simular ser un usuario

- Interacciones, user agent, tracking y sesiones automáticas

Interactuar con la página programáticamente

QUÉ NOS OFRECE



Conexión ininterrumpida

- No es un HTTP GET, es una conexión que no se cierra hasta que queramos

Simular ser un usuario

- Interacciones, user agent, tracking y sesiones automáticas

Interactuar con la página programáticamente

- Clicks, delays, esperar a que cargue el DOM y scripts

¿POR QUÉ NECESITAMOS HEADLESS?



Frameworks

SPAs

Interacciones con la página

FRAMEWORKS Y LIBRERÍAS



Scope de classNames

classNames autogenerados/ofuscados en build-time

Componentes con información en memoria esperando interacciones

NextJS y Remix, serialización del servidor y SSR

SPA



Carga estática de un HTML simple sin información

- Pantallazo blanco
- CSR -> el contenido se genera tras la carga

La página carga con el DOM listo

Requiere de interacciones y enrutaciones para acceder al contenido de verdad

INTERACCIONES CON LA PÁGINA



Páginas que no usar URL as State Manager

- Paginaciones
- Filtros

Cálculos al vuelo e información consultada con servidor

- Página oficial de la lotería
- Calculadoras de nóminas, hipotecas, etc.

Paginaciones dinámicas, en base a respuestas del servidor

Paginaciones por cursor (no hay limit ni offset)

Infinite scrolling para listados

OPCIONES DE HEADLES SCRAPING



- Selenium -> ampliamente conocido
- Puppeteer
- Cypress
- Playwright

Alternativamente...

Consola de la página y JavaScript puro

- Útil para ejecuciones manuales y poco alcance

HERRAMIENTAS

Google Colab

HERRAMIENTAS

Google Colab

- Jupyter Notebook en la nube
- Acceso gratuito para bajo rendimiento

Jupyter Notebook

HERRAMIENTAS

Google Colab

- Jupyter Notebook en la nube
- Acceso gratuito para bajo rendimiento

Jupyter Notebook

- Documentación y código en un mismo fichero
- Ejecución de bloques de código con contexto compartido

Ventajas de usar Google Colab

HERRAMIENTAS

Google Colab

- Jupyter Notebook en la nube
- Acceso gratuito para bajo rendimiento

Jupyter Notebook

- Documentación y código en un mismo fichero
- Ejecución de bloques de código con contexto compartido

Ventajas de usar Google Colab

- IP y VPN externas, no nos limitaremos por proxies
- Reiniciar entorno y generar nueva IP (por si nos bloquean)
- Internet y ancho de banda más alto
- Almacenamiento directo en la nube

DEMO: HEADLESS WEB SCRAPING



Páginas más “ofuscadas”, (in)voluntariamente, y cómo un headless browser nos ayuda

Tiempo aproximado: **10 minutos**

Target: **ZARA**

SI UN USUARIO LO VE, LO PUEDES SCRAPEAR

Toda información accesible a un
usuario, puede ser extraída
programáticamente

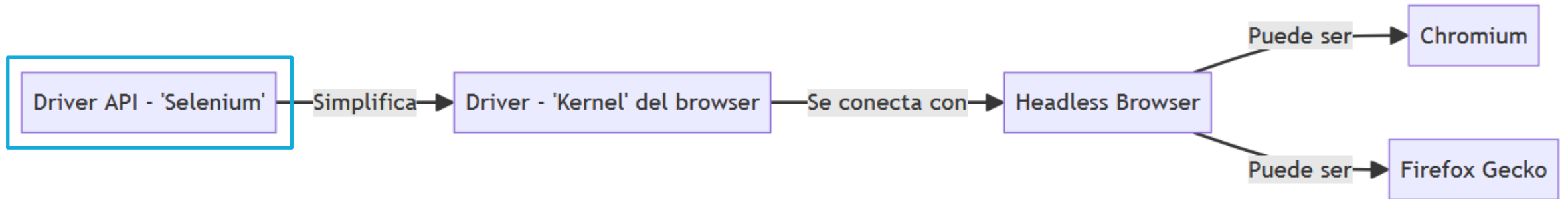




2

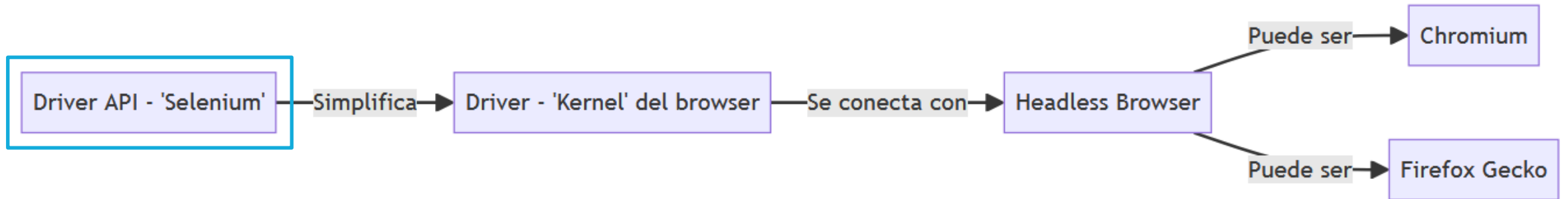
SELENIUM

ARQUITECTURA DE SELENIUM



La API es dependiente del lenguaje que se esté utilizando

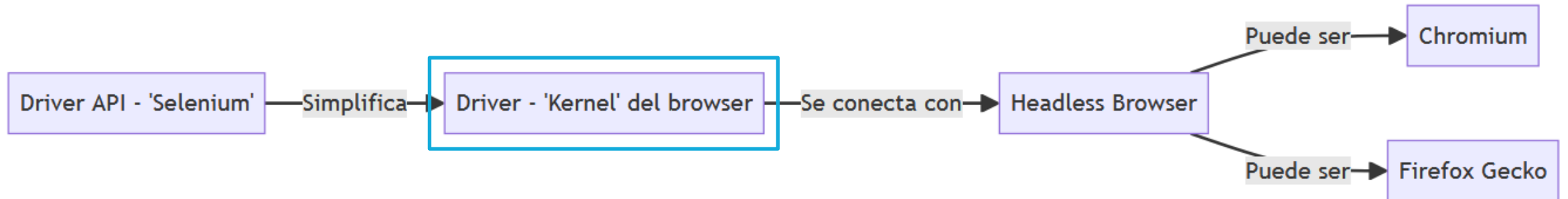
ARQUITECTURA DE SELENIUM



La API es dependiente del lenguaje que se esté utilizando

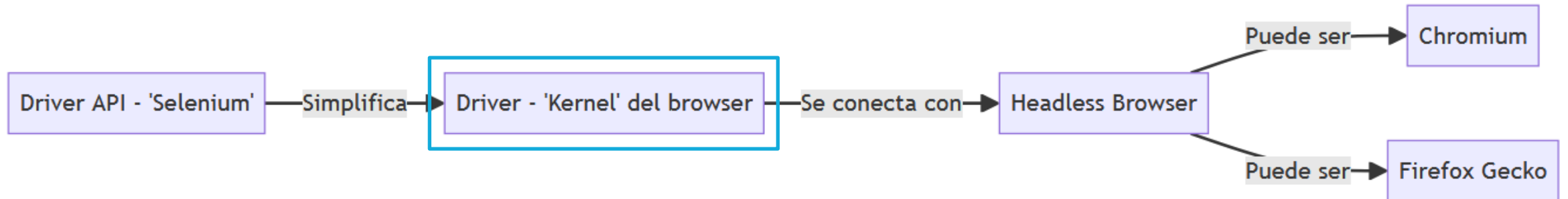
1. La librería podrá cambiar sintaxis aún manteniendo conceptos
2. El funcionamiento será muy parecido, pero con las peculiaridades de cada lenguaje (Python GIL, Java JVM)

ARQUITECTURA DE SELENIUM



El driver es la conexión específica de la librería con el headless browser, es el elemento común

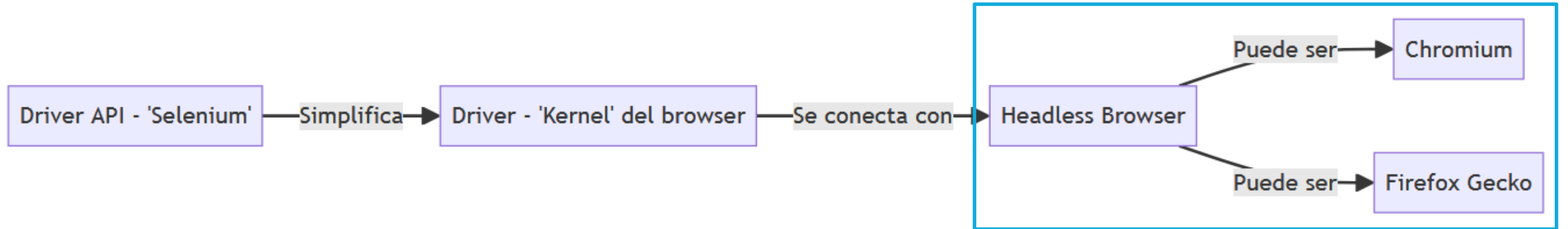
ARQUITECTURA DE SELENIUM



El driver es la conexión específica de la librería con el headless browser, es el elemento común

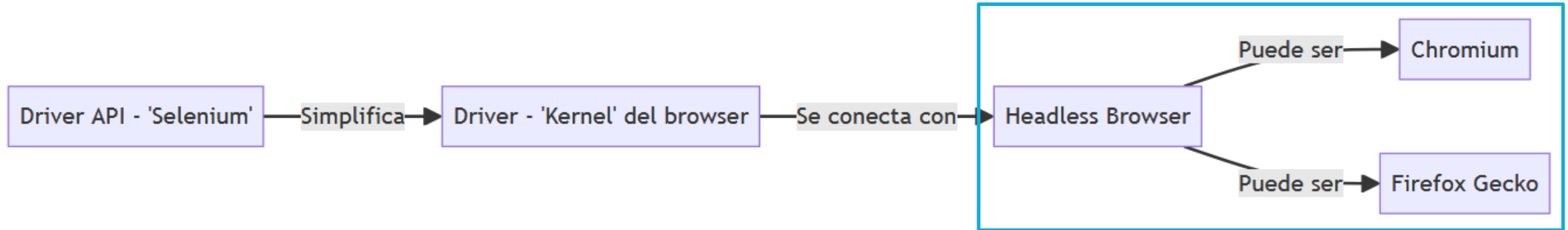
1. Java usa un JVM para proporcionar un lenguaje unificado sin importar el Sistema Operativo
2. Selenium tiene un driver que se conectará con cada navegador (o "será" el navegador)

ARQUITECTURA DE SELENIUM



Headless Browser, el *flavor* del navegador

ARQUITECTURA DE SELENIUM



Headless Browser, el *flavor* del navegador tiene diferencias y soportes, user-agents diferentes, motores diferentes

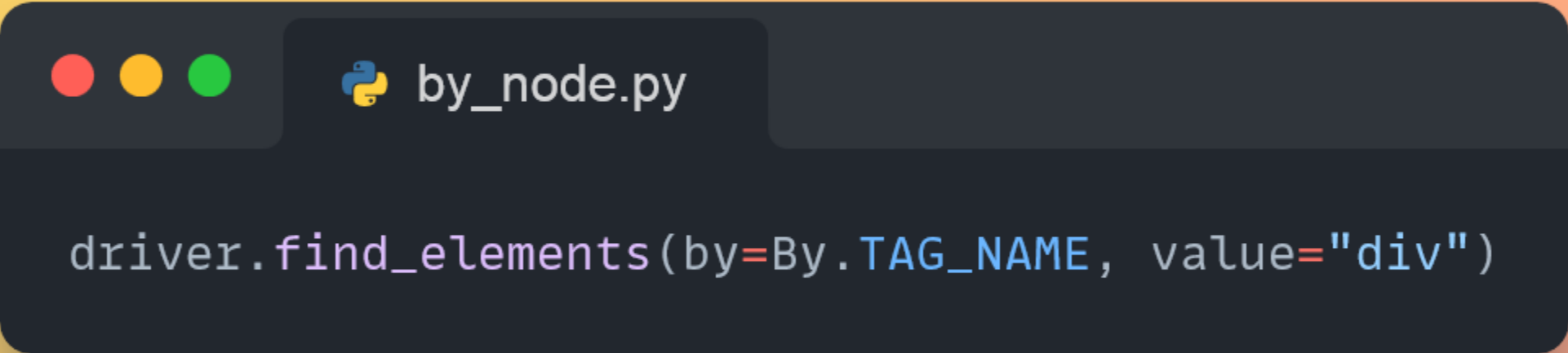
1. Unificando conceptos (driver) y con una librería que en varios lenguajes, es lógico no atarse a un navegador, ahí entran los engines que queramos usar
2. Google Chrome usa el motor de Chromium (y la mayoría de los navegadores también), Firefox tiene su propio motor (Gecko), y Safari también (WebKit)

API DE SELENIUM



- Buscar
- Interactuar
- Capturar

BUSCAR ELEMENTOS POR NODO



```
driver.find_elements(by=By.TAG_NAME, value="div")
```

snappify.com

BUSCAR ELEMENTOS POR CLASE

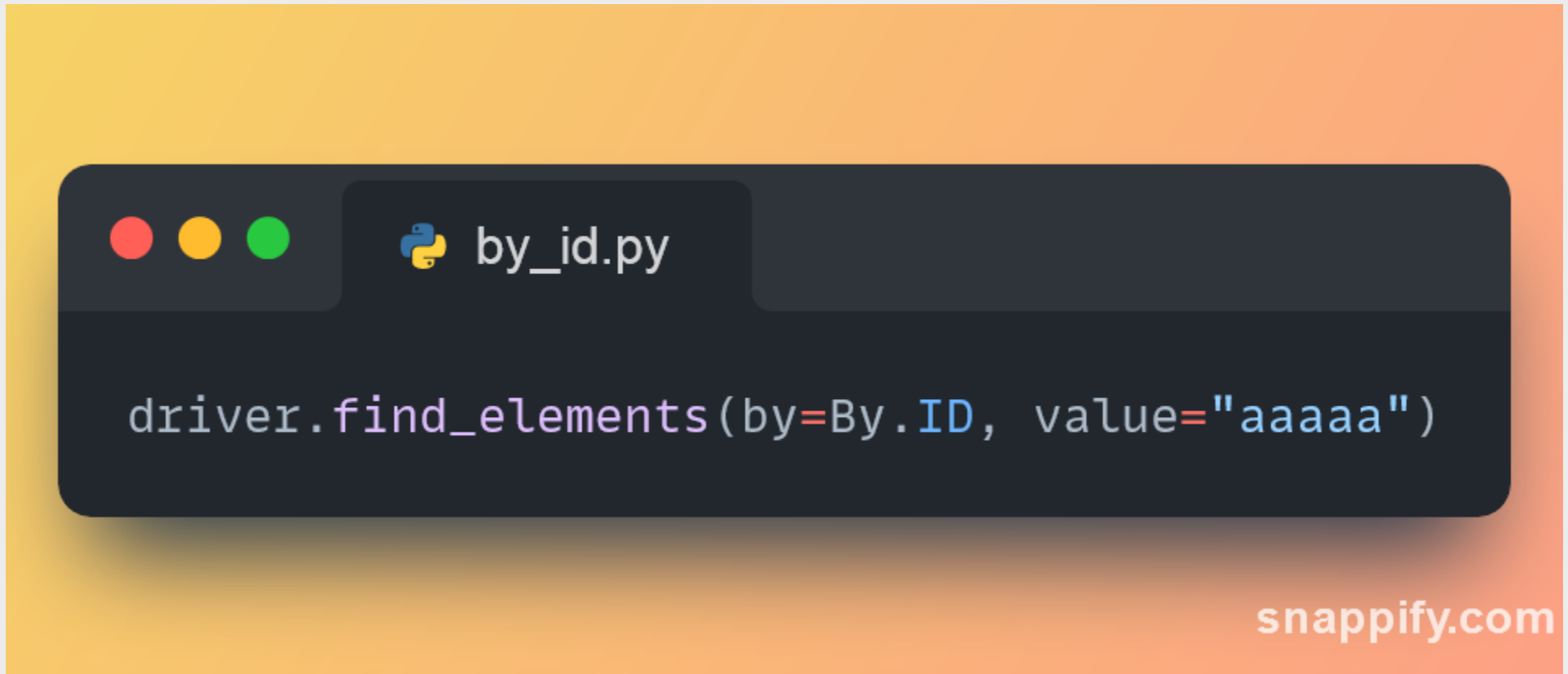


by_class.py

```
driver.find_elements(by=By.CLASS_NAME, value="aaaaa")
```

snappify.com

BUSCAR ELEMENTOS POR ID



BUSCAR ELEMENTOS POR QUERYSELECTOR

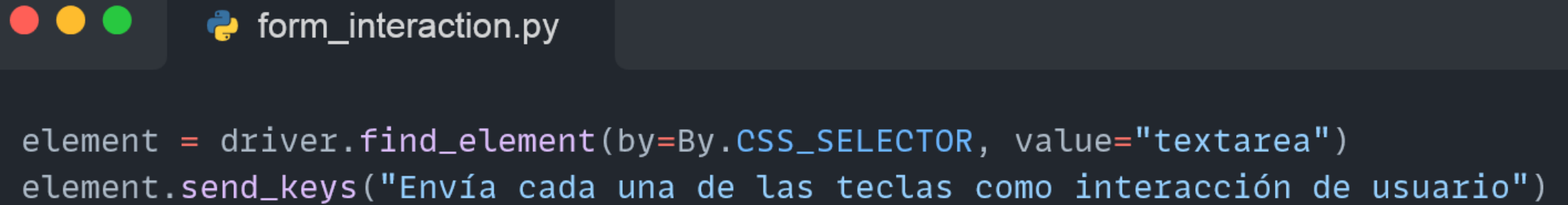


by_querySelector.py

```
driver.find_elements(by=By.CSS_SELECTOR, value="div > img")
```

snappify.com

RELLENAR UN FORMULARIO



```
element = driver.find_element(by=By.CSS_SELECTOR, value="textarea")
element.send_keys("Envía cada una de las teclas como interacción de usuario")
```

snappify.com

INTERACCIÓN CON CLICK



click_interaction.py

```
element = driver.find_element(by=By.CSS_SELECTOR, value="button")  
element.click()
```

snappify.com

INTERACCIÓN CON SCROLL

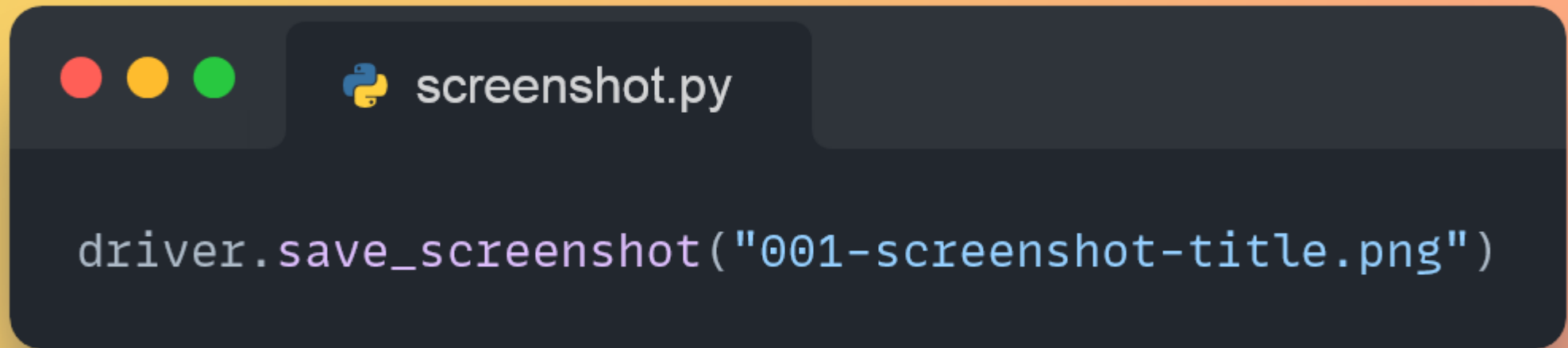


scroll_interaction.py

```
from selenium.webdriver import ActionChains  
  
amount = ActionChains(driver).scroll_by_amount(0, amount).perform()
```

snappify.com

CAPTURAS DE PANTALLA



```
driver.save_screenshot("001-screenshot-title.png")
```

snappify.com



RECAPITULANDO

RECAPITULANDO...



Qué hemos visto

RECAPITULANDO...



1. Headless Web Scraping

RECAPITULANDO...



1. Headless Web Scraping
2. Selenium

CRÉDITOS

LIBRO RECOMENDADO

<https://www.amazon.com/-/es/Fernando-Rosa/dp/8409363801>

Mundo de los datos que se dirige a utilizarlos en su día a día, como a general interesados en aplicarlos. ¿Cómo analiza de qué modo los datos

El libro en cuatro grandes bloques: los básicos como dato y algoritmo, temas de inteligencia artificial o de

era muy didáctica el camino para y comunicar con datos, es decir, el *data literacy* o alfabetización

o las empresas pueden incorporar de un método muy sencillo, cómo empresarial.

habla de la localización de los datos de la seguridad y de la privacidad. En los próximos años, hablaremos, trabajaremos

una sociedad *datificada* que nos rodea. La gestión de datos. **DATA** es sin duda

n

ra los clientes de Adam.enta.

Fernando de la Rosa
@titonet

DA
TA

Una edición especial publicada para los clientes de Adam

Fernando de la Rosa
@titonet



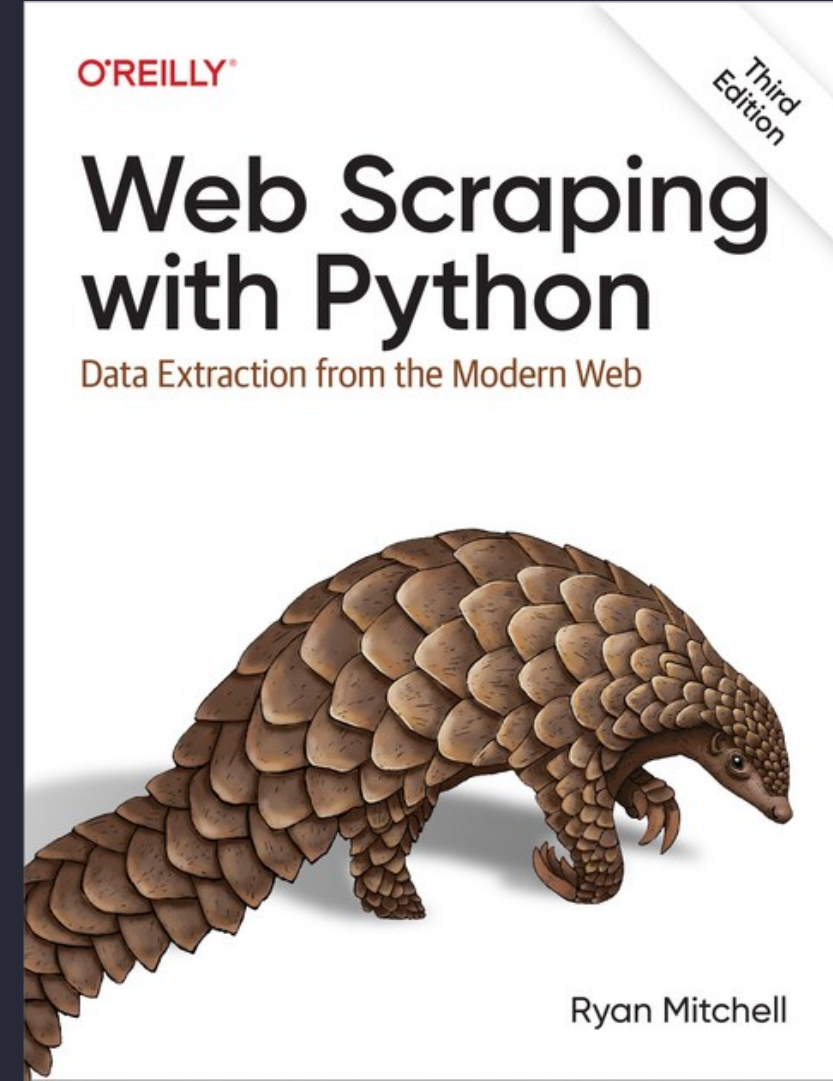
Cómo los datos te ayudarán en tu vida y en tu empresa,
y transformarán la sociedad

Prólogo de José Mejías y David Ribalta



LIBRO RECOMENDADO

<https://www.amazon.es/web-scraping-python-extraction-modern/dp/1098145356>





LIBRO RECOMENDADO

https://openaccess.uoc.edu/bitstream/10609/147437/1/webscrapping_modulo1_webscrapping.pdf

Web scrapping

PID_00256970

Laia Subirats Maté
Mireia Calvo González

Tiempo mínimo de dedicación recomendado: 5 horas





PREGUNTAS

¡¡GRACIAS!!



About Capgemini

Capgemini is a global leader in partnering with companies to transform and manage their business by harnessing the power of technology. The Group is guided everyday by its purpose of unleashing human energy through technology for an inclusive and sustainable future. It is a responsible and diverse organization of 270,000 team members in nearly 50 countries. With its strong 50 year heritage and deep industry expertise, Capgemini is trusted by its clients to address the entire breadth of their business needs, from strategy and design to operations, fuelled by the fast evolving and innovative world of cloud, data, AI, connectivity, software, digital engineering and platforms. The Group reported in 2020 global revenues of €16 billion.



Get the Future You Want | www.capgemini.com

This presentation contains information that may be privileged or confidential and is the property of the Capgemini Group.

Copyright © 2024 Capgemini. All rights reserved.