VLC TECH FEST '24

Web Scraping

Impacto Social en la Era Digital

Quién soy yo

Pepe fabra Valverde

Lead y Arquitecto de Front en Capgemini

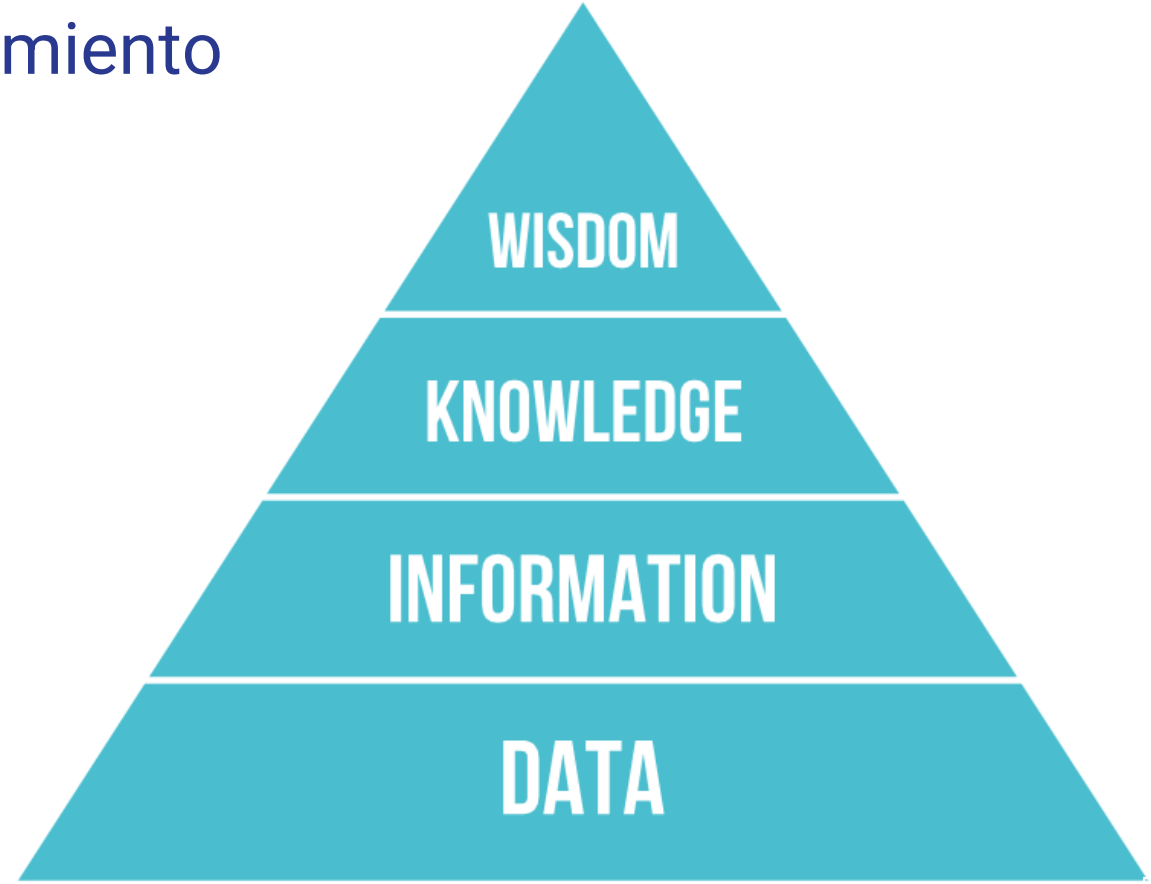
me sigo pareciendo

¿Qué quiero enseñaros?

- Web Scraping a escala
- Datos Ofuscados
- Importancia de los datos
- Cómo nos impacta en nuestro día a día

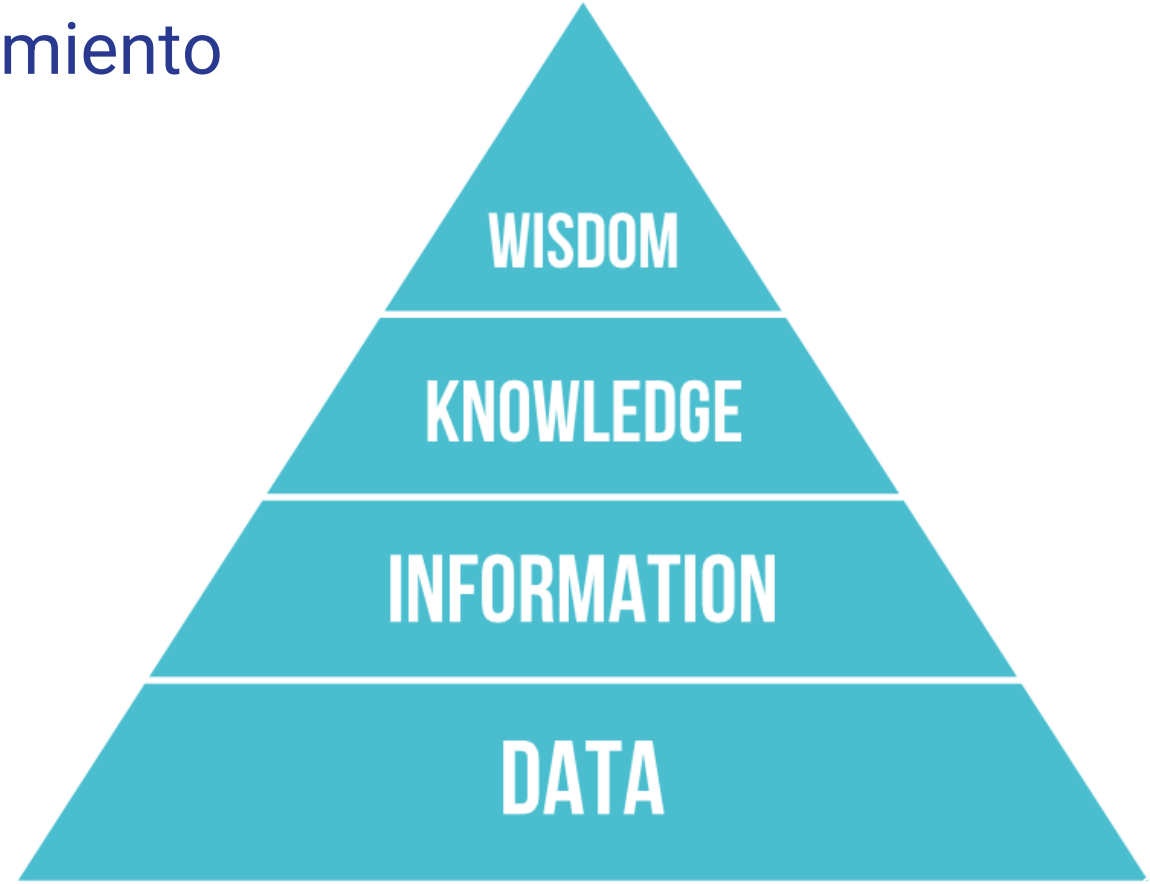
Datos

Pirámide del conocimiento



Dato

Pirámide del conocimiento



Información

Dato

- Valor puro

Pirámide del conocimiento

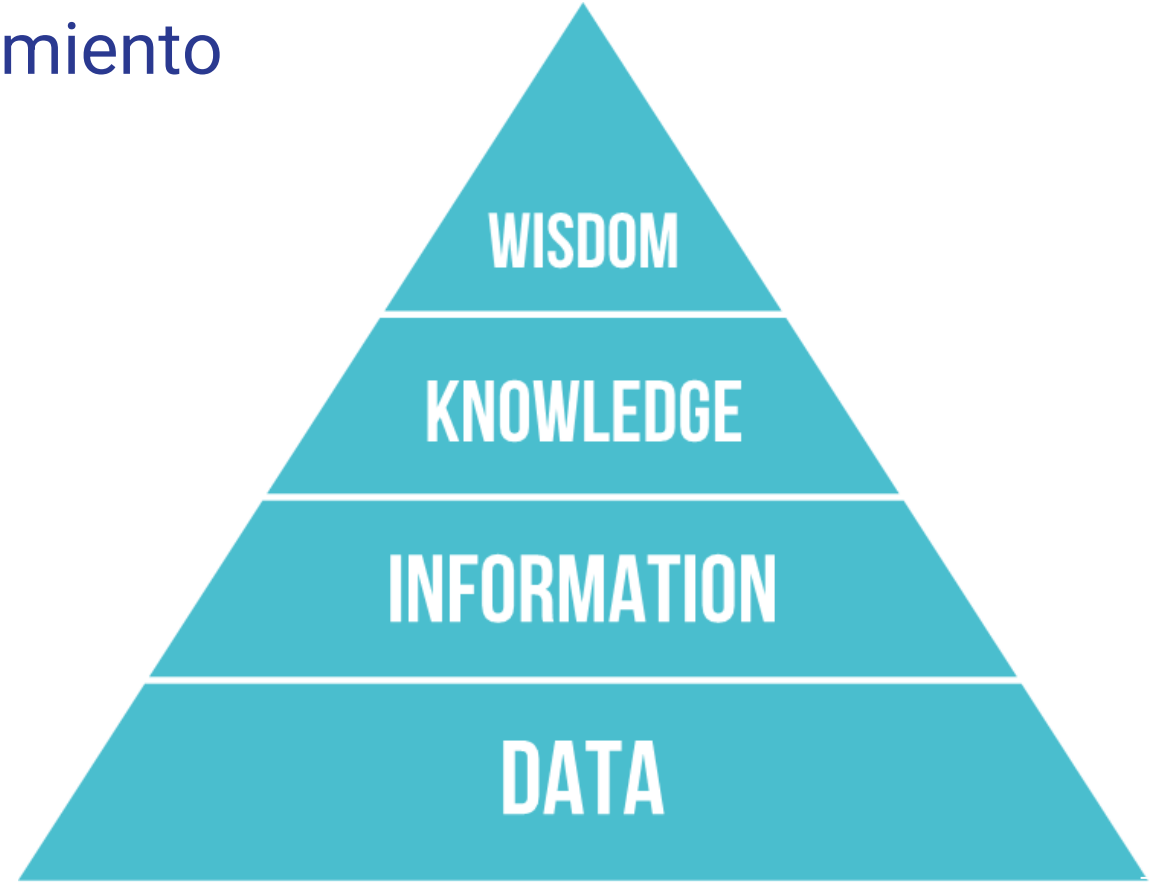
Conocimiento

Información

- Dato con contexto

Dato

- Valor puro



Pirámide del conocimiento

Sabiduría

Conocimiento

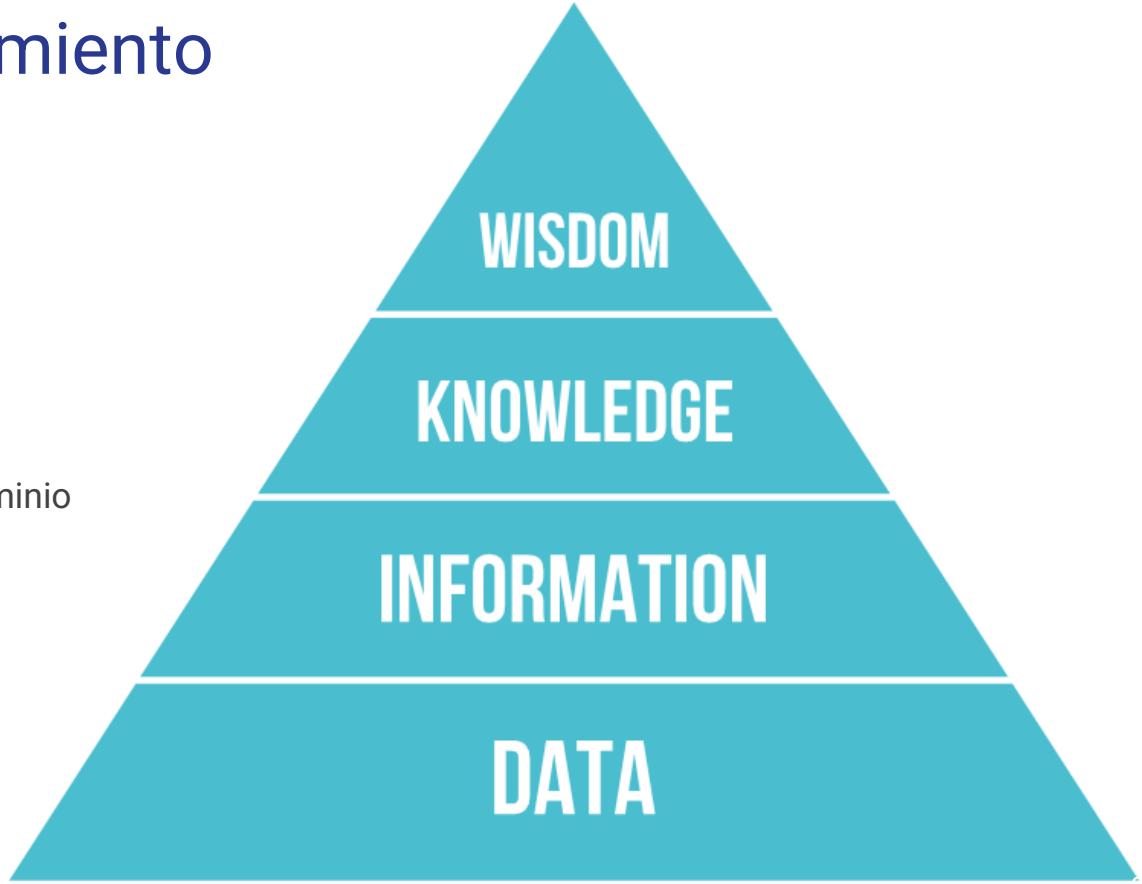
- Conjunto de información de un dominio
- Toma de decisiones aquí

Información

- Dato con contexto

Dato

- Valor puro



Pirámide del conocimiento

Sabiduría

- Conjunto de conocimiento
- Predicción de comportamientos aquí

Conocimiento

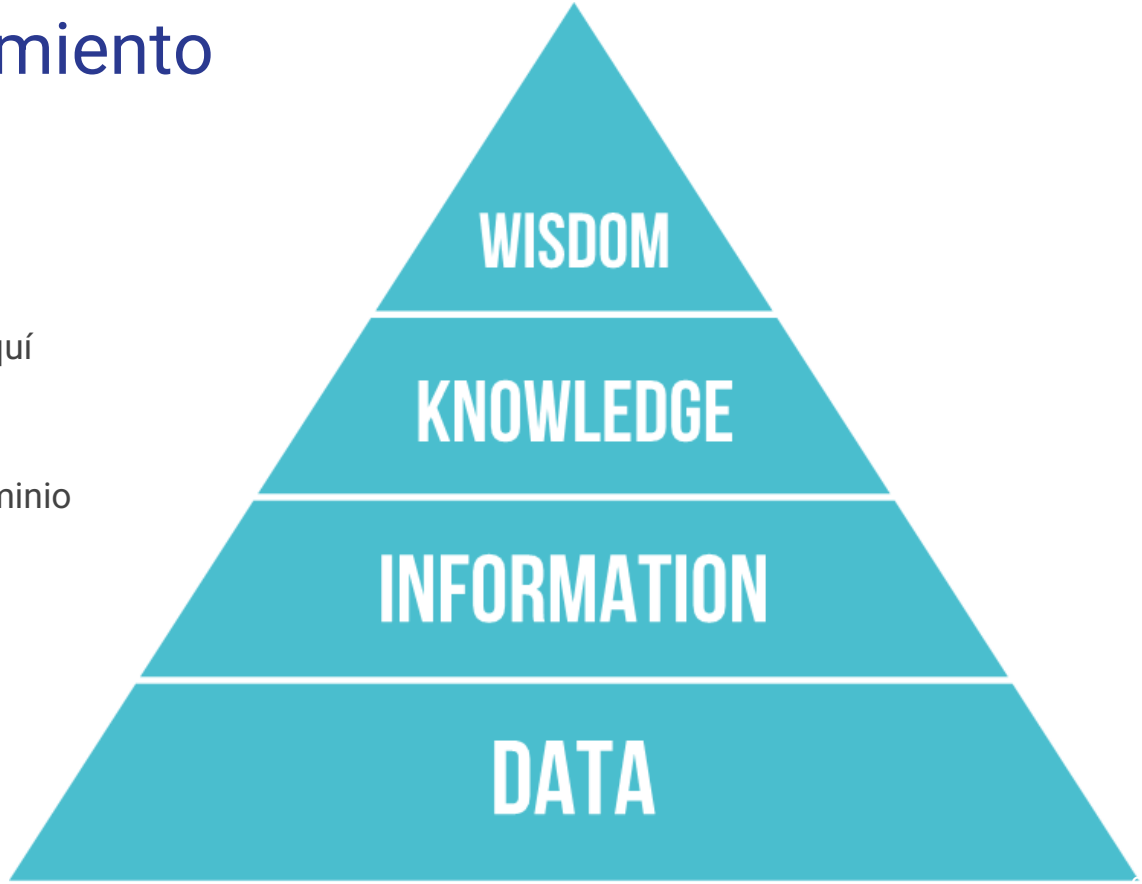
- Conjunto de información de un dominio
- Toma de decisiones aquí

Información

- Dato con contexto

Dato

- Valor puro



Datos

¿Para qué los queremos?

Datos

¿Para qué los queremos?

Para competir, aprender, mejorar, vender, comprar

Algunos casos de uso podrían ser

- Análisis de Datos
- Machine Learning
- Seguimiento de ofertas
- Comprender mejor un dominio

Ofuscación y Bases de Datos

- ¿Qué es una Base de Datos?

Ofuscación y Bases de Datos

- ¿Qué es una Base de Datos?
- ¿Qué es ofuscación?

Ofuscación y Bases de Datos

- ¿Qué es una Base de Datos?
- ¿Qué es ofuscación?

Bases de Datos Ofuscadas son la realidad

Datos son Datos

Y si son importantes, cómo se consiguen

Y si son importantes, cómo se consiguen

- Comprando (Empresas dedicadas)

Y si son importantes, cómo se consiguen

- Comprando (Empresas dedicadas)
- Explotación (Proyectos ad hoc)

Y si son importantes, cómo se consiguen

- Comprando (Empresas dedicadas)
- Explotación (Proyectos ad hoc)
- …Web Scraping

Lo que ve un usuario (web scraping)

Pico y pala. La obtención de datos es un proceso automatizable, pero que requiere de supervisión humana

Web Scraping

¿Qué es el web scraping?

Extraer información de una página web de manera programática

Si un usuario lo ve, se puede scrapear

Orígenes

Con cada avance, siempre surgen nuevas disciplinas. Pero tendemos a reinventar

Ojeadores, ¿qué son… o eran?

SEO

Search Engine Optimization

Objetivo de la ofuscación

No existe lo infranqueable.

Pero sí se pueden poner suficientes trabas para que el otro lado se rinda

Aunque a veces también es un efecto secundario del avance (frameworks SPA)

“Tipos” de Web Scraping

Normal, petición HTTP

- fácilmente detectable
- muy sencillo

Headless, simular ser un usuario

- más difícil de mantener
- efectivo

No sólo para grandes empresas

Alertas personalizadas

Descarga de LMS (Learning Management Systems)

- Aules, moodle, campus, etc.

También nos podemos beneficiar

Competencia de empresas a alto nivel

En el siglo XXI se compite con información

Todo va demasiado deprisa pero esta vez no se ve tanto la guerra por la información

Impacto técnico del web scraping a escala

Cómo lo enfocaremos

- Empresas grandes
- Empresas pequeñas

Ataques y técnicas de ofuscación darían para su propia charla, ***pero veremos algunos.***

Empresas grandes

- Dedicar tiempo a ofuscar y protegerse

Empresas grandes

- Dedicar tiempo a ofuscar y protegerse
- Recursos, físicos y cloud

Empresas grandes

- Dedicar tiempo a ofuscar y protegerse
- Recursos, físicos y cloud
- Pérdida de clientes potenciales; caídas del servicio, pasos para evitar scraping

Empresas grandes

- Dedicar tiempo a ofuscar y protegerse
- Recursos, físicos y cloud
- Pérdida de clientes potenciales; caídas del servicio, pasos para evitar scraping
- Competencia con información más accesible

Empresas pequeñas

- Recursos físicos o cloud, un exceso de peticiones se nota más en esa economía

Empresas pequeñas

- Recursos físicos o cloud, un exceso de peticiones se nota más en esa economía
- DDoS hunde páginas, levantar “a mano” y pérdida de clientes potenciales

Empresas pequeñas

- Recursos físicos o cloud, un exceso de peticiones se nota más en esa economía
- DDoS hunde páginas, levantar “a mano” y pérdida de clientes potenciales
- Ha de poder scrapearse por SEO

Vale… pero ¿cómo me afecta?

40%

de todo el tráfico en Internet está formado por tráfico de bots

Más pasos que antes para lo mismo

- Captcha
- Verifica tu humanidad (Cloudflare)

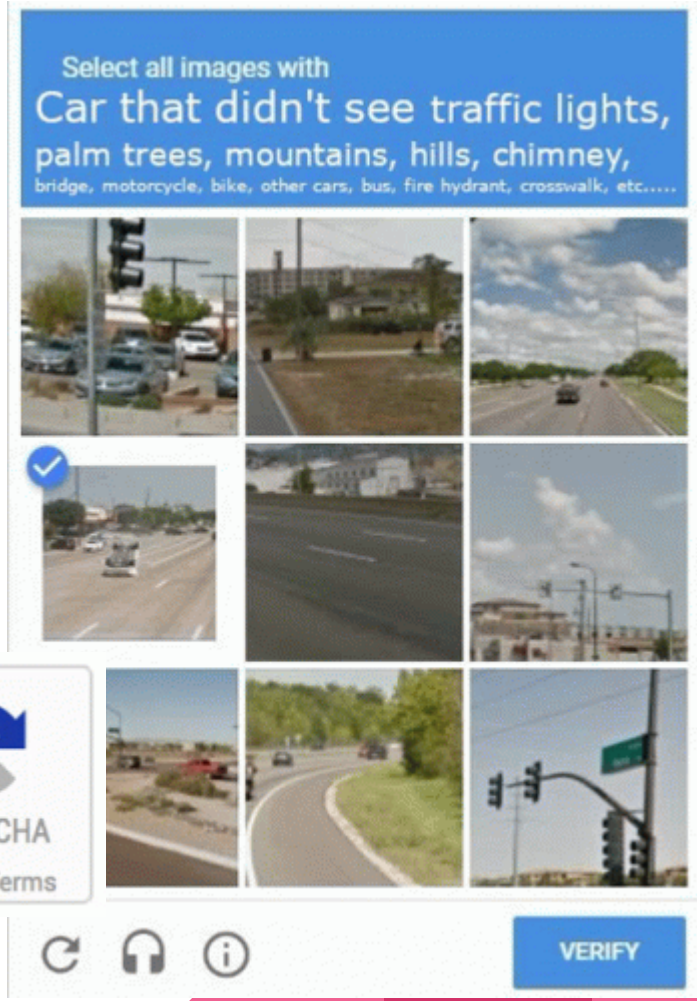
Y otras verificaciones similares que **acaban saturando**

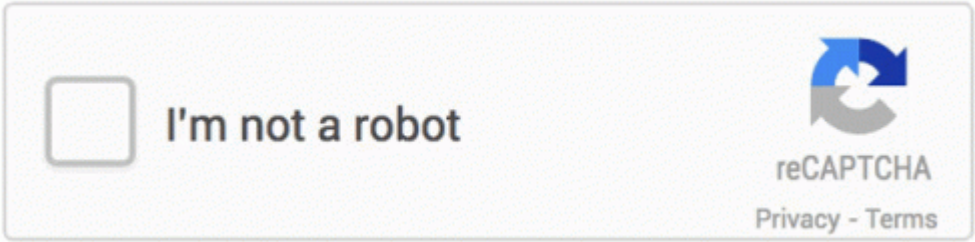
DDoS

Qué es DDoS

Servicios más resilientes (más caros y más recursos)

Captchas, verificaciones





Rate limiting y LogIn

¿No os pasa que cada vez más os piden usuario?

Registro de peticiones por usuario (y baneo)

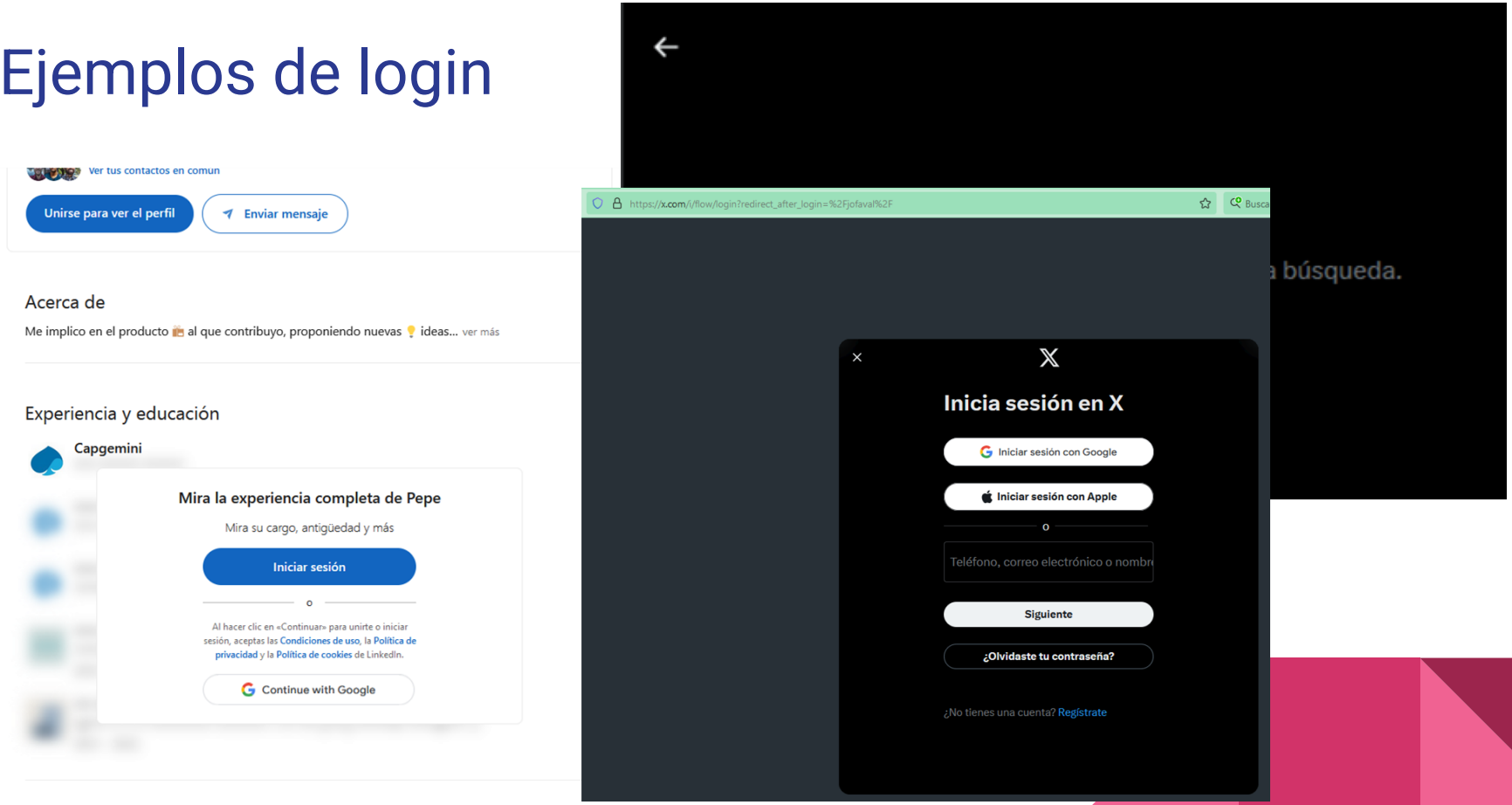
¿Por qué ocurre?

Siempre pidiendo login

- Rate limiting (pero no para mí)
- Throttling
- Mejor que banear IPs, que también

Pero como usuario, esto es **molesto**

Ejemplos de login



Cada vez la información está “más oculta”

LinkedIn, Twitter/X, Instagram y muchas páginas no permiten curiosear tanto *gratis*.

Parte por ganancias, pero también para proteger sus datos.

- LinkedIn desde incógnito no siempre muestra el perfil
- Twitter/X protege sus datos y servidores de ataques
- Meta, Redes Sociales, pero también portales de noticias

Derechos de autor

- Textos (tweets, posts)
- Libros y artículos
- Imágenes (IA Generativa)
  - Glaze y Nightshade
  - Opt-Out (Instagram, Facebook)

SEO y resúmenes de pequeños comercios

El SEO sigue reglas no escritas (para no gamificarlas)

Los mismos crawlers del SEO ayudan a hacer resúmenes de páginas

- Si Google te lo resume, te vas a fiar (y no entrarás a la página)
  - y algunas empresas se quedarán sin tu visita, que es como sobreviven (anuncios, ventas, etc.)
- Google AI Overview, *resume* información, pero el feedback general no es positivo

Pero que si lo prohiben… también mal

No todo es malo

Pero que si lo prohiben… también mal

- Estudios de malas praxis (inmobiliaria, comida, ropa)
- Seguimiento de la inflación

Páginas dedicadas a “Encuentra los mejores precios”

- Booking
- Trivago
- Vuelos de avión trackeando IP para subir el precio

OpenGraph

Compartir enlaces y que salgan portadas

Probad a compartir el enlace de VLC Tech Fest a ver qué sale

Datos públicos no estructurados

- Información de sensórica pública
  - y poder hacer estudios
- Datos y tablas de información gubernamental
- Registros antiguos que ahora se pueden re-procesar

Ética

Pequeños comercios

- Contactar de antemano antes de iniciar el proceso
- Throttling para no saturarles o incluso tirarles
- Respetar el robots.txt (vulnerabilidades)

Robots.txt

Qué propósito cumple

Términos y condiciones

Algunas empresas el problema lo atajen desde legal.

Gobernanza de datos y modelos más abiertos

Con datos más accesibles de empresas, evitaríamos problemas y surgirían nuevos

Para quien le interese como sería a nivel personal

MyData es una propuesta de modelo

Recapitulación… Hemos visto

- Datos

Recapitulación… Hemos visto

- Datos
- Web Scraping

Recapitulación… Hemos visto

- Datos
- Web Scraping
- Impacto técnico

Recapitulación… Hemos visto

- Datos
- Web Scraping
- Impacto técnico
- Impacto Social

Conclusión

Repositorios

El impacto social del web scraping en la era digital

github.com/jofaval/talks-about/tech-talks/vlc-tech-fest/web-scraping

Web Scraping y las Bases de Datos Ofuscadas

github/jofaval/talks-about/uv/web-scraping-y-las-bases-de-datos-ofuscadas

Bibliografía

Citas

Porcentaje de tráfico en internet,
https://www.cloudflare.com/learning/bots/what-is-bot-traffic/

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

VLC TECH FEST '24

Preguntas

Encuéntrame en

- linkedin.com/in/jofaval/
- github.com/jofaval

VLC TECH FEST '24

¡¡GRACIAS!!