WEB SCRAPING

Y LAS BASES DE DATOS OFUSCADAS

Abril 2024

Capgemini

QUÉ APRENDERÁS HOY

Mundo de los datos

Datos ofuscados

Web scraping

¿QUIÉN SOY?

Pepe Fabra Valverde
Líder y Arquitecto de Front

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

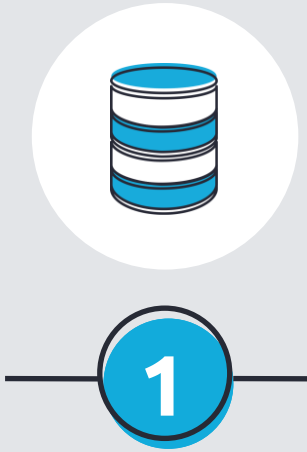
AVISO PARA NAVEGANTES

- 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

1

BASES DE DATOS OFUSCADAS

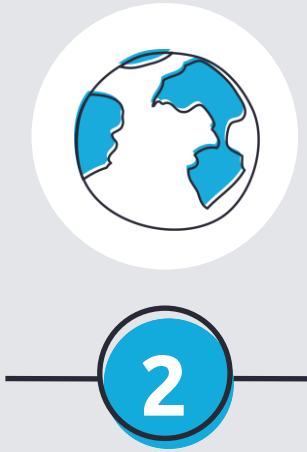
BASES DE DATOS OFUSCADAS



1

QUÉ ES UNA BDD

Conceptos esenciales



2

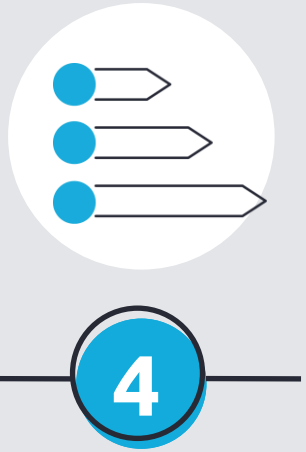
MUNDO DE DATOS

Y por qué importa

3

DATOS OFUSCADOS

Qué son y por qué nos importan



4

PIRÁMIDE DEL CONOCIMIENTO

Comprendimiento del panorama

5

WEB SCRAPING

Cómo extraer los datos *ofuscados*

BASE DE DATOS… ¿OFUSCADAS?

¿Qué es una Base de Datos?
Cuál es su propósito

BASE DE DATOS... ¿OFUSCADAS?

¿Qué es una Base de Datos?

BASE DE DATOS... ¿OFUSCADAS?

¿Qué es una Base de Datos?

- Repositorio estructurado de información

Cuál es su propósito

BASE DE DATOS... ¿OFUSCADAS?

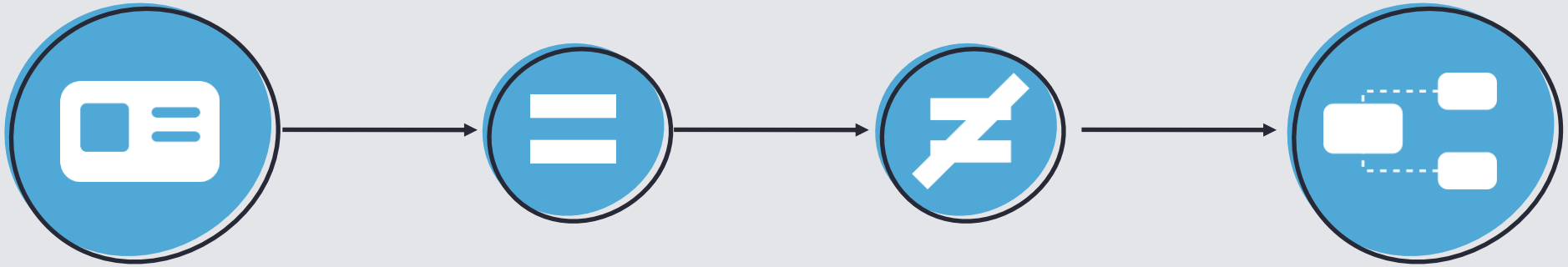
¿Qué es una Base de Datos?

- Repositorio estructurado de información

Cuál es su propósito

- Almacenar información de manera estructurada
- Mutar información estructurada
- Consultar información estructurada

ESTRUCTURA DE UNA BASE DE DATOS



ESQUEMA

INGESTA

CONSULTA

ANÁLISIS

CONCEPTOS DE BASES DE DATOS

ESQUEMA

Estructura de los datos definida con DDL (CREATE, ALTER, etc.)

CONSULTA

Recabar información de los datos estructurados, definido con DML (SELECT)

INGESTA

Acción sobre los datos definida con DML (INSERT, UPDATE, etc.)

RECORDEMOS QUE

- DDL -> Data Definition Language (instrucciones para el esquema)
- DML -> Data Manipulation Language (instrucciones para las operaciones en un esquema)

MUNDO DE DATOS

El siglo XXI es el siglo de la información

Las empresas compiten por tener datos, propios y externos

- Si sabes la estrategia de tus competidores la puedes rebatir

DATOS OFUSCADOS

DATOS OFUSCADOS

Por qué ofuscarían los datos las empresas

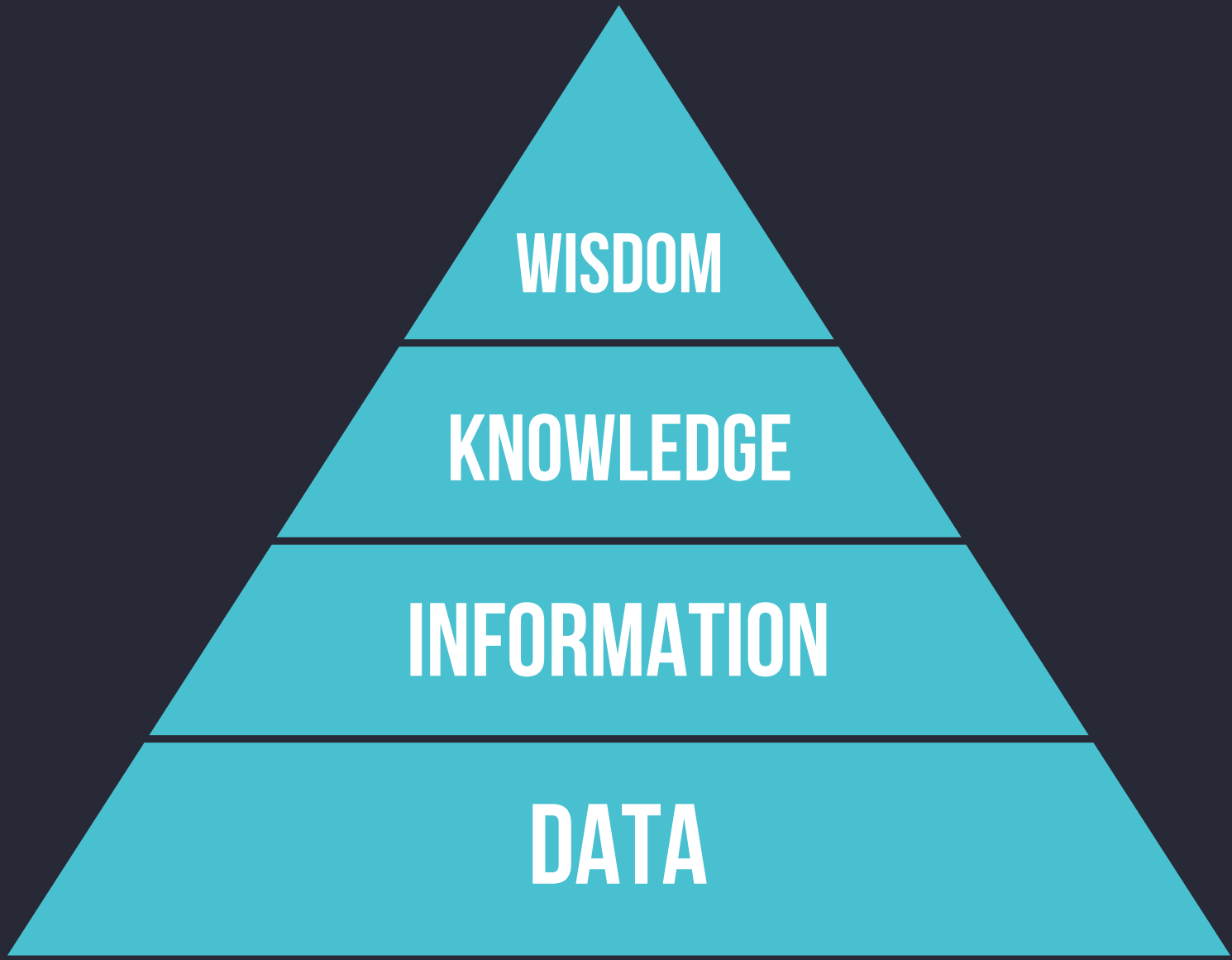
- E-commerce
- ¿.zip con toda la información comercial?

Mostrar información a usuarios, y entorpecérsela a competidores

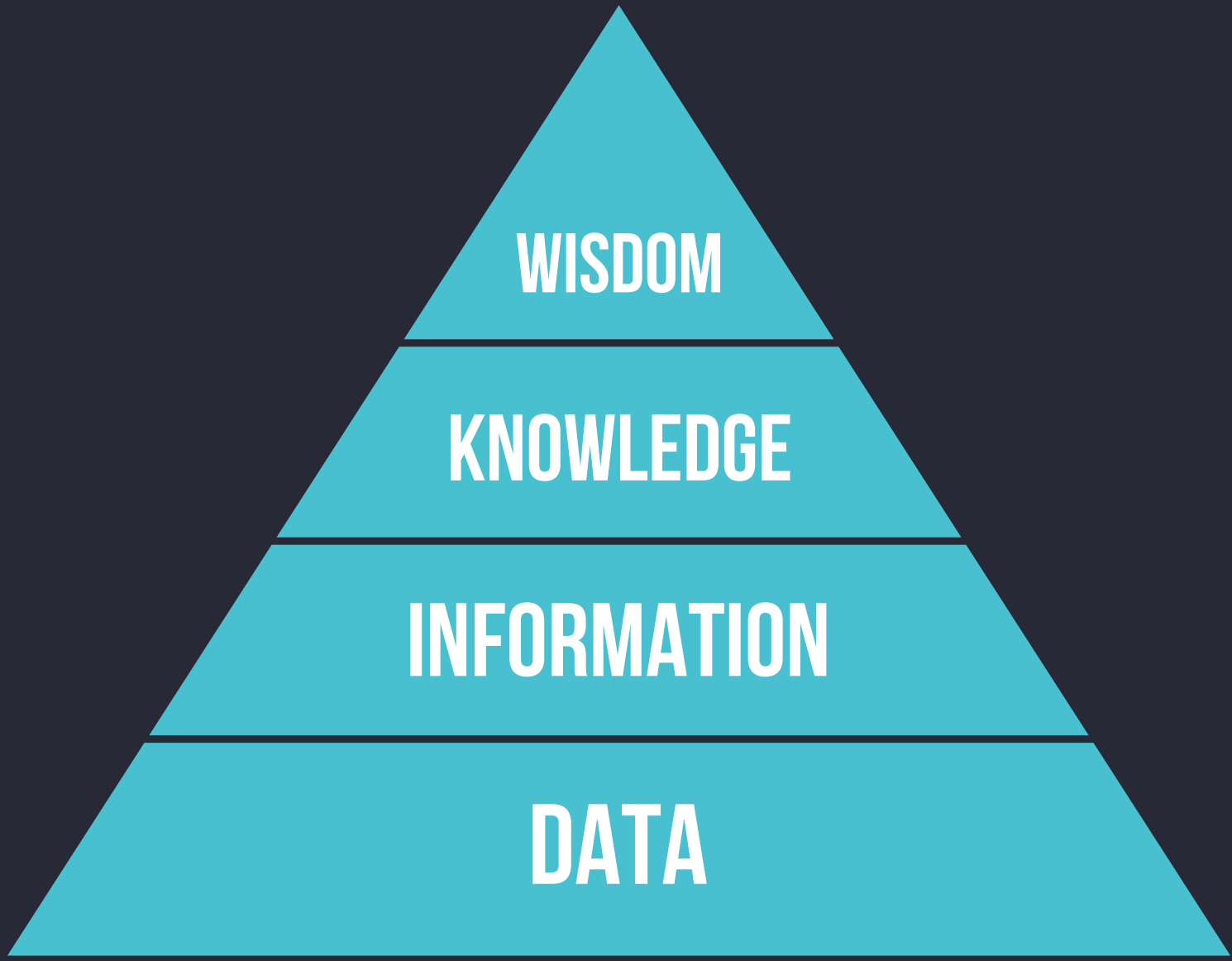
PARA QUÉ NECESITAMOS DATOS

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

PIRÁMIDE DEL CONOCIMIENTO

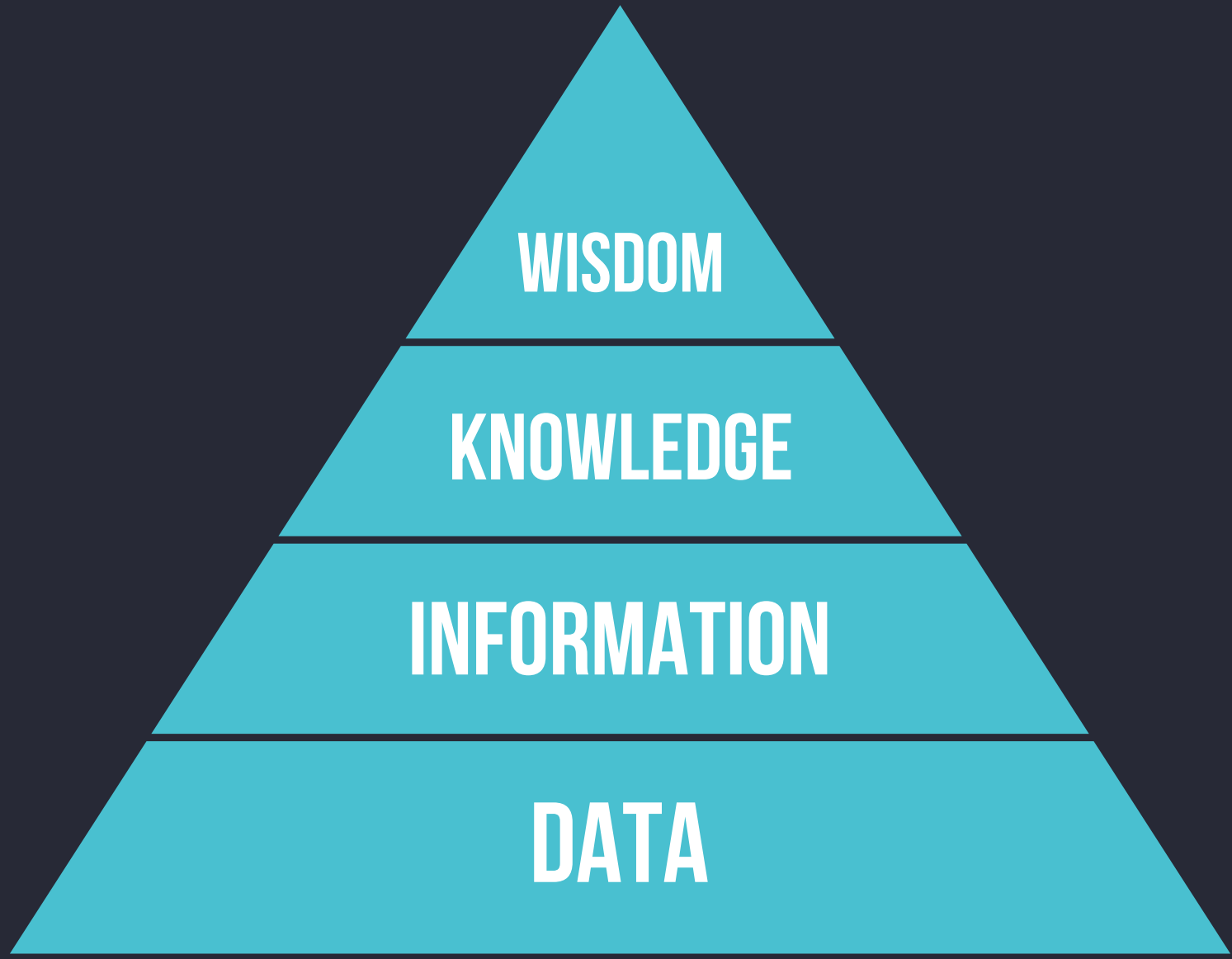


PIRÁMIDE DEL CONOCIMIENTO



Dato

PIRÁMIDE DEL CONOCIMIENTO



Información

Dato

- Valor puro

PIRÁMIDE DEL CONOCIMIENTO

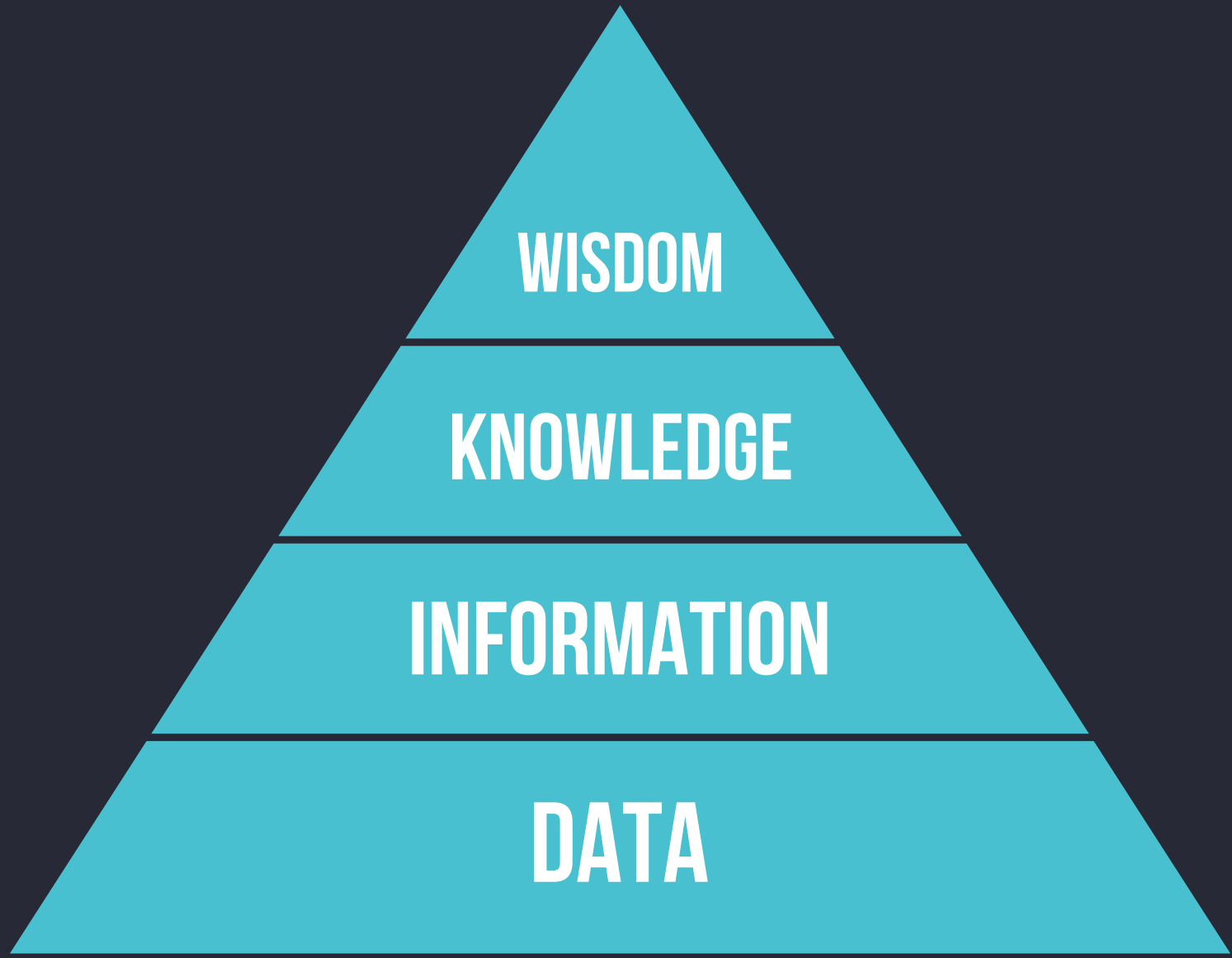
Conocimiento

Información

- Dato con contexto

Dato

- Valor puro



PIRÁMIDE DEL CONOCIMIENTO

Sabiduría

Conocimiento

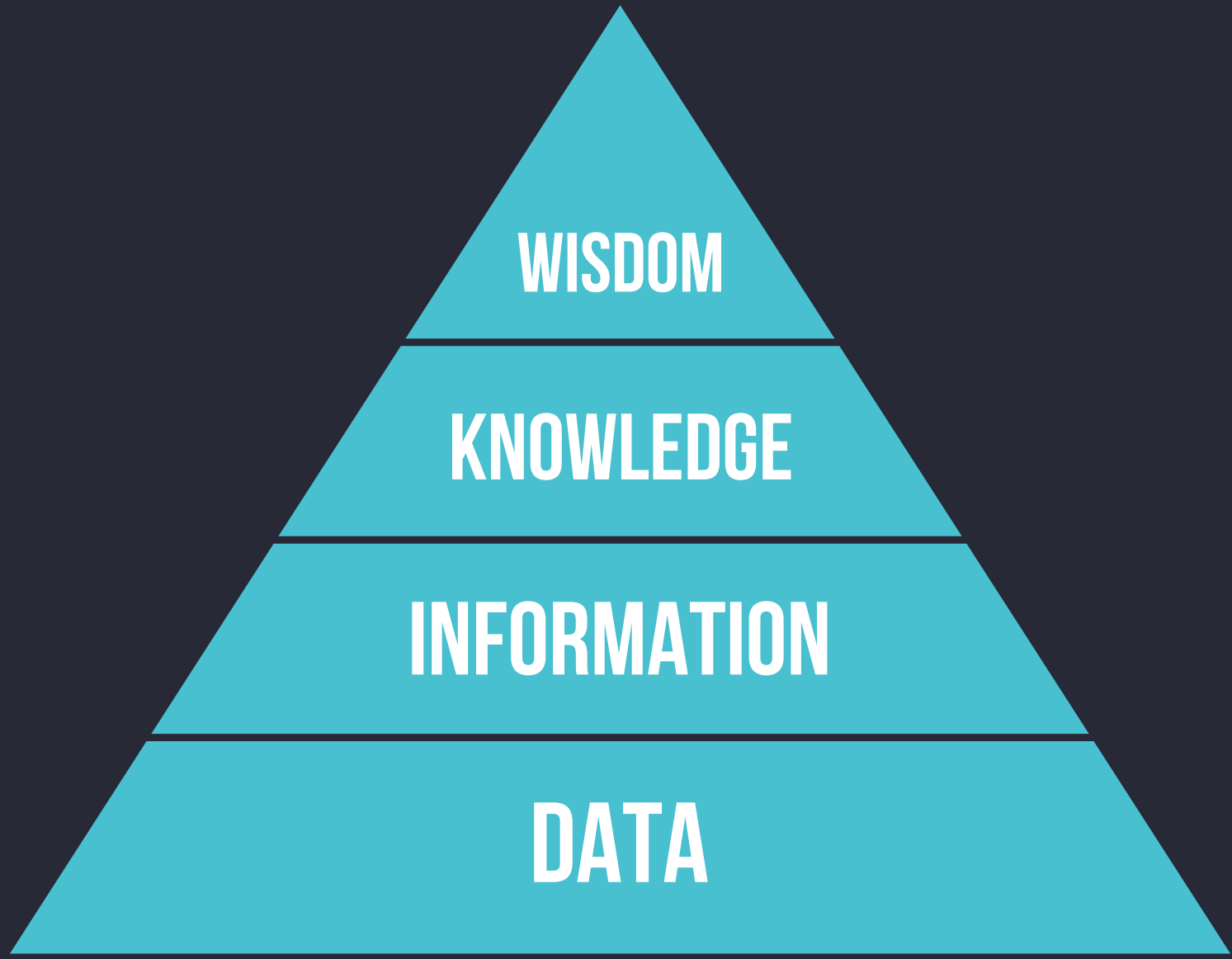
- Conjunto de información de un dominio
- Toma de decisiones aquí

Información

- Dato con contexto

Dato

- Valor puro



PIRÁMIDE DEL CONOCIMIENTO

Sabiduría

- Conjunto de conocimiento
- Predicción de comportamientos aquí

Conocimiento

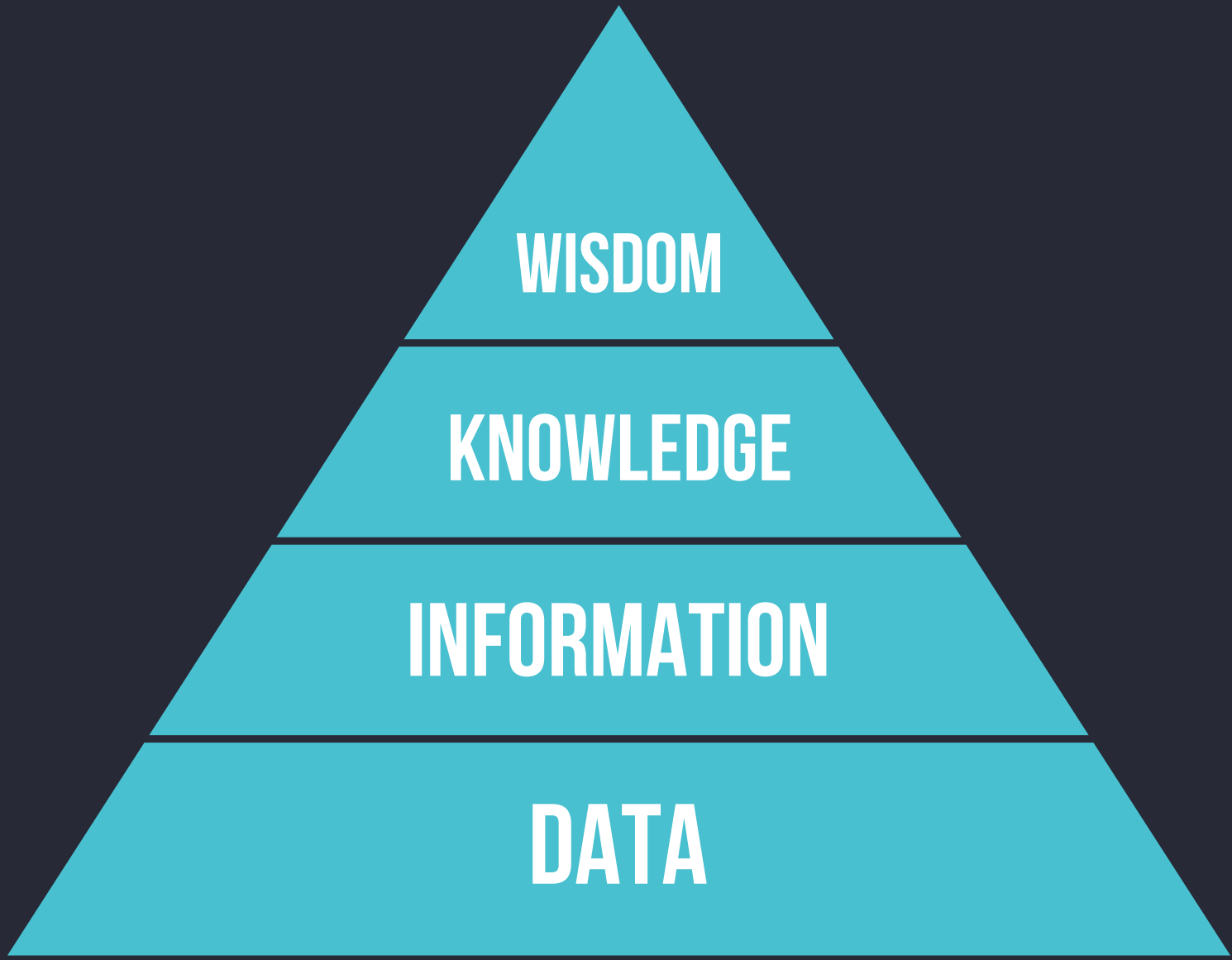
- Conjunto de información de un dominio
- Toma de decisiones aquí

Información

- Dato con contexto

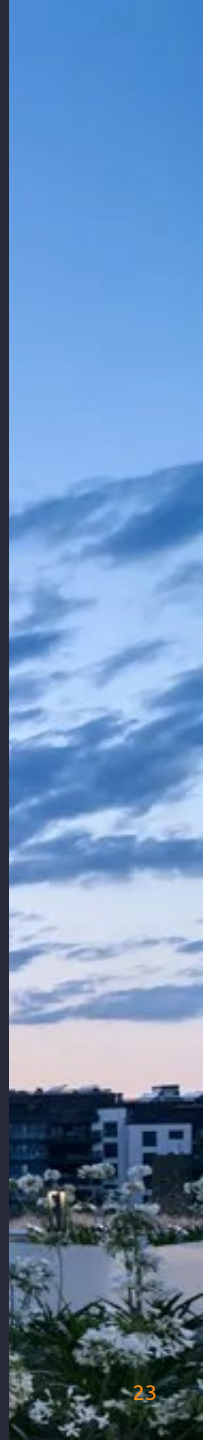
Dato

- Valor puro



2

WEB SCRAPING



QUÉ ES WEB SCRAPING

Extraer información de una página web de manera programática

UTILIDADES DEL WEB SCRAPING

Alertas personalizadas para sitios sin suscripciones

Avisos personalizados de descuentos en productos

Seguimiento y análisis de datos

Automatización de descargas con esquemas similares

- Moodle, Aules, Campus, etc.

ESTADOS HTTP

1XX – Información

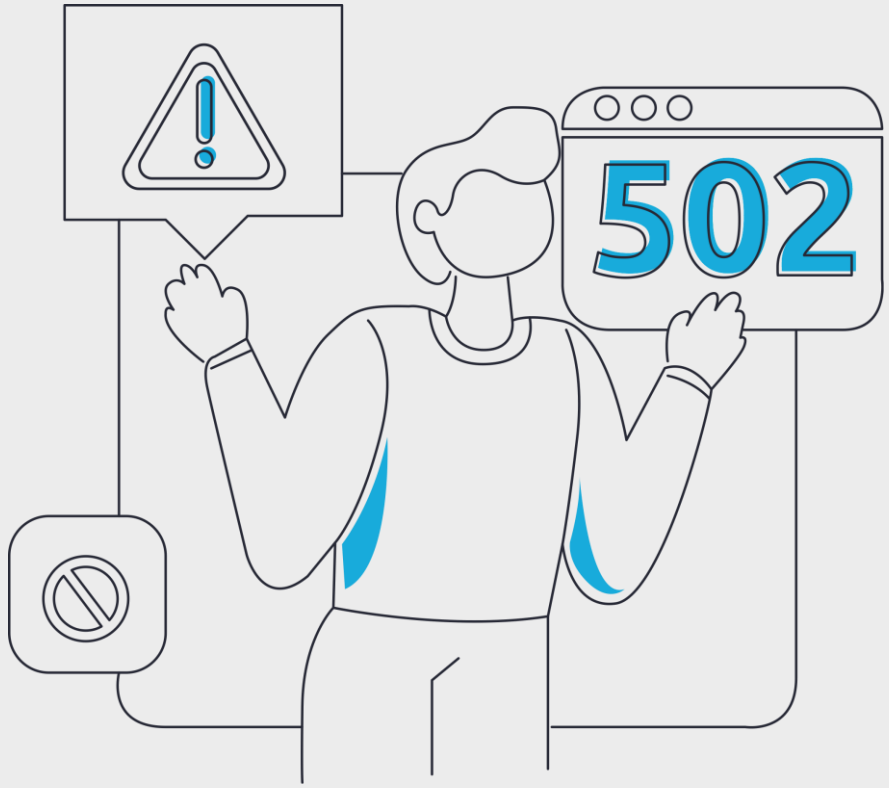
2XX – Éxito

3XX – Redirección

4XX – Fallo del cliente

- 418 Soy una tetera – Día de los inocentes

5XX – Fallo del servidor



Fuente: https://umbraco.com/knowledge-base/http-status-codes/#:~:text=The%20100%20Continue%20status%20code,the%20request%20has%20already%20finished.

MÉTODOS HTTP

GET

POST

PUT

PATCH

DELETE

OPTIONS

… y más

PIPELINE

Procedimiento habitual

PIPELINE

1. Request

PIPELINE

1. Request
   - Utilizando el User Agent correspondiente, solicitamos el contenido estático
2. Parser

PIPELINE

1. Request
   - Utilizando el User Agent correspondiente, solicitamos el contenido estático
2. Parser
   - Parseamos el contenido utilizando XML, HTML, LXML, a elección
3. Extract

PIPELINE

1. Request
   - Utilizando el User Agent correspondiente, solicitamos el contenido estático
2. Parser
   - Parseamos el contenido utilizando XML, HTML, LXML, a elección
3. Extract
   - Realizamos queries para extraer la información relevante de la página
4. [Iterate]

PIPELINE

1. Request
   - Utilizando el User Agent correspondiente, solicitamos el contenido estático
2. Parser
   - Parseamos el contenido utilizando XML, HTML, LXML, a elección
3. Extract
   - Realizamos queries para extraer la información relevante de la página
4. [Iterate]
   - Puede que tengamos que navegar la página (categorías, ofertas, etc.)
   - Puede que tengamos un listado de productos paginado

EXTRACCIÓN

- Origen
- Métodos

EXTRACCIÓN

- Sitemap
- Navegación
- Interacción

EXTRACCIÓN POR SITEMAP

El sitemap es un XML usado para el SEO, puede haber múltiples sitemaps, por idiomas, imágenes, recursos, son configurables y comunicados a los navegadores (Google Search Console)

Lo que hace el sitemap es facilitarle la faena el robot de SEO, es decir, al servicio que se encarga de hacer web scraping

EXTRACCIÓN POR NAVEGACIÓN

Navergando el header, extrayendo información del menú (categorías), buscando una página de secciones.

Navegando subsecciones a partir de X secciones

EXTRACCIÓN POR INTERACCIÓN

Es parecida a la **Extracción por navegación**, pero con pasos extra, no siempre estará disponible todo lo que tienes que navegar, y, es más, dependerá de las acciones que programes que ciertos flujos de navegación estén habilitados o no.

MÉTODOS DE LOCALIZACIÓN DE NODOS

XML

MÉTODOS DE LOCALIZACIÓN DE NODOS

XML

- Navegación por jerarquía de nodos, xPath simplificado

xPath

MÉTODOS DE LOCALIZACIÓN DE NODOS

XML

- Navegación por jerarquía de nodos, xPath simplificado

xPath

- Lenguaje de consultas de XML, más costoso de leer y de mantener, pero más versátil

querySelector

MÉTODOS DE LOCALIZACIÓN DE NODOS

XML

- Navegación por jerarquía de nodos, xPath simplificado

xPath

- Lenguaje de consultas de XML, más costoso de leer y de mantener, pero más versátil

querySelector

- “Lenguaje” de localización de HTML mediante reglas de CSS, xPath algo menos versátil pero más legible

Y DESPUÉS... ¿QUÉ?

- Analizar datos
- Transformar datos
- Almacenar en una BDD (MySQL, MariaDB, PostgreSQL, etc.)
- Almacenamiento físico estructurado (csv, xml, xlsx, etc.)

DEMO

Página de productos sencilla

Tiempo aproximado: **10 minutos**

Target: **DRUNI**

RECAPITULANDO

RECAPITULANDO...

Qué hemos visto

RECAPITULANDO...

1. Conceptos de una Base de Datos

RECAPITULANDO...

1. Conceptos de una Base de Datos
2. Web Scraping y las Bases de Datos ofuscadas

RECAPITULANDO...

1. Conceptos de una Base de Datos
2. Web Scraping y las Bases de Datos ofuscadas
3. Web Scraping

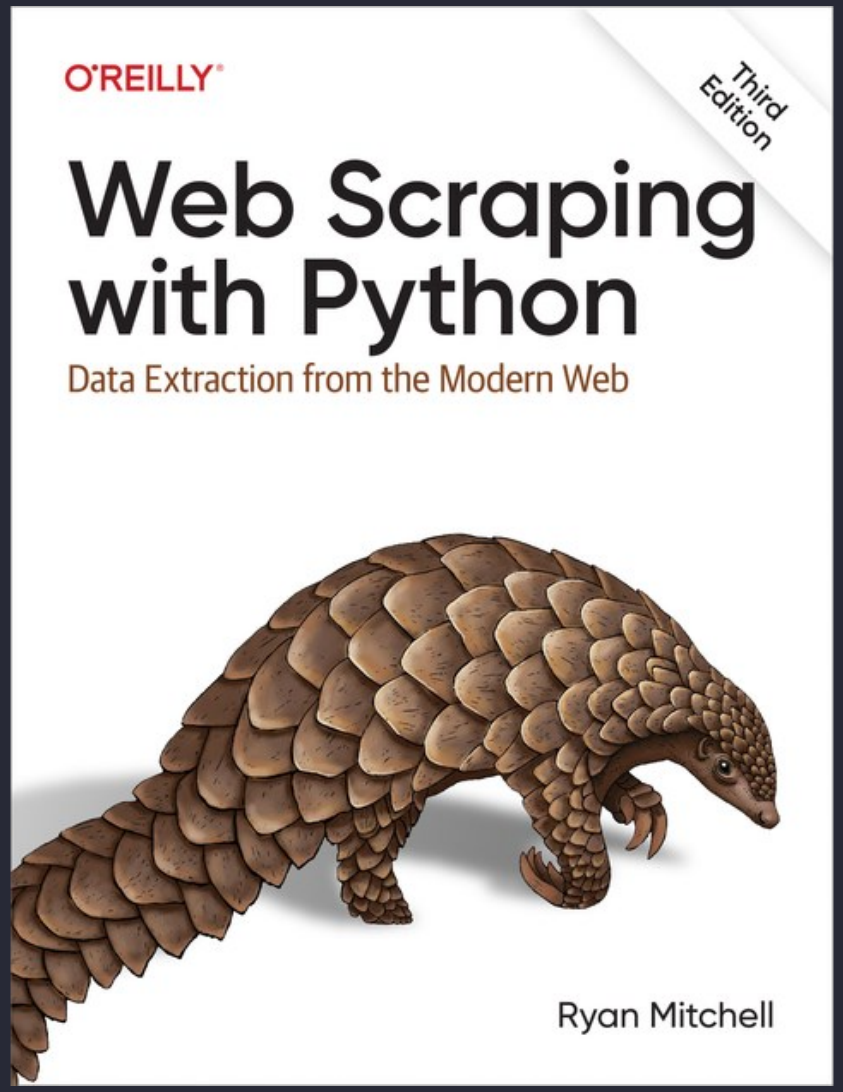
CRÉDITOS

LIBRO RECOMENDADO

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

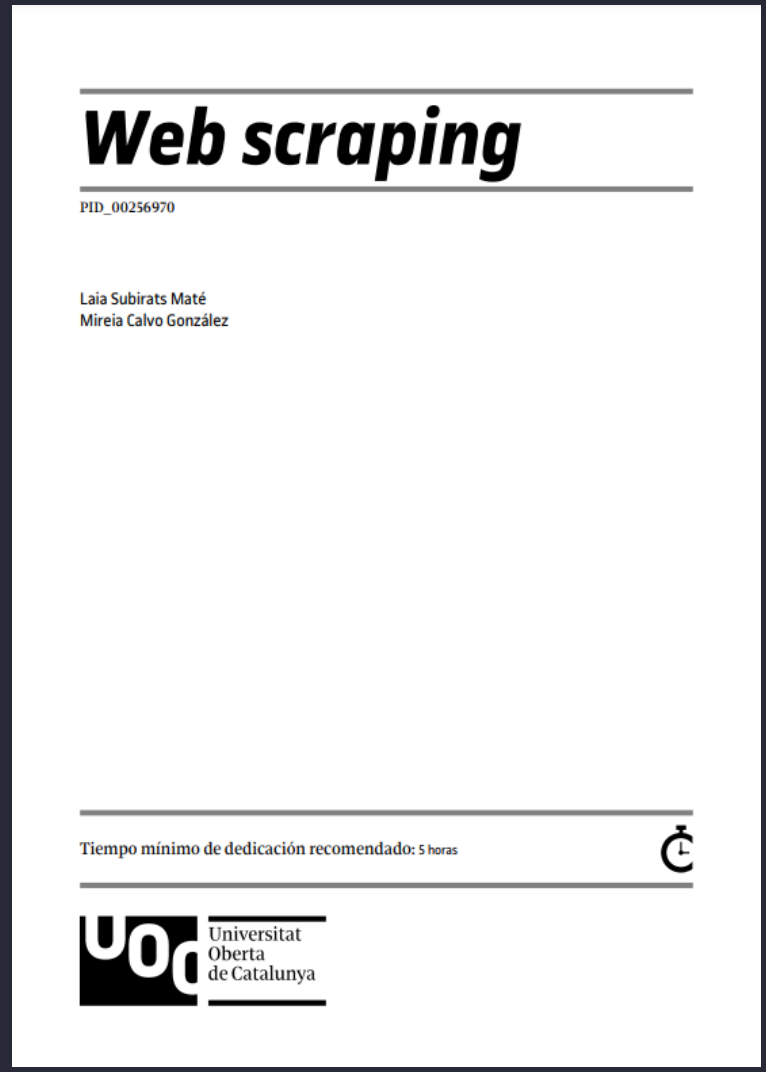
LIBRO RECOMENDADO

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



LIBRO RECOMENDADO

https://openaccess.uoc.edu/bitstream/10609/147437/1/webscraping_modulo1_webscraping.pdf



Web scraping

PID_00256970

Laia Subirats Maté
Mireia Calvo González

Tiempo mínimo de dedicación recomendado: 5 horas

Universitat Oberta de Catalunya

BIBLIOGRAFÍA

Pirámide del conocimiento - DATA: cómo los datos te ayudarán… - Fernando de la Rosa

Códigos HTTP - https://umbraco.com/knowledge-base/http-status-codes/#:~:text=The%20100%20Continue%20status%20code,the%20request%20has%20already%20finished.

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