



## PANORAMA & DIAGNÓSTICO

Análisis de competencias para la consolidación  
tecnológica de **Madison Intelligence**

Mayo 2020

# CONTENIDO

- I. INTRODUCCIÓN
- II. ETAPAS DEL ANÁLISIS
  - i. EXTRACCIÓN
  - ii. ALMACENAMIENTO
  - iii. BASES DE DATOS
  - iv. MÉTODOS ANALÍTICOS
  - v. VISUALIZACIÓN
- III. ANÁLISIS DE COMPETENCIAS
- IV. TECNOLOGÍAS “BIG DATA”
- V. RECOMENDACIONES

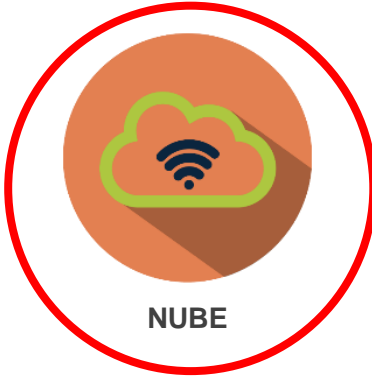
# I. INTRODUCCIÓN



# TECNOLOGÍAS DE LA REVOLUCIÓN DIGITAL 4.0



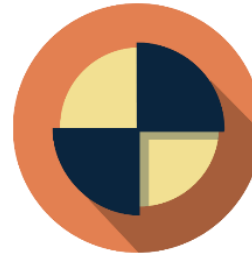
BLOCKCHAIN



NUBE



ROBÓTICA



SIMULACIONES



MATERIALES  
AVANZADOS



REALIDAD VIRTUAL /  
REALIDAD AUMENTADA



MANUFACTURA  
ADITIVA



INTERNET DE  
LAS COSAS



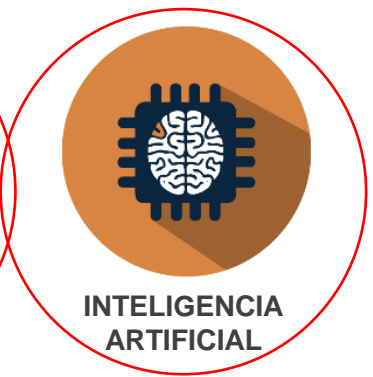
BIG DATA



CIBERSEGURIDAD



SOFTWARE



INTELIGENCIA  
ARTIFICIAL

MADISON - Tecnologías con potencial aplicación

# HIPÓTESIS / TEMAS DE INTERÉS

1

¿QUÉ **COMPETENCIAS DEBERÁ ADQUIRIR MADISON** ANTE LA ACTUAL REVOLUCIÓN DIGITAL?

2

¿CUÁLES SON LAS **TECNOLOGÍAS DE VANGUARDIA** EN FUNCIÓN DE LAS NECESIDADES DE MADISON?, ¿CUÁLES SON SUS **REQUERIMIENTOS TÉCNICOS Y ALCANCES**?

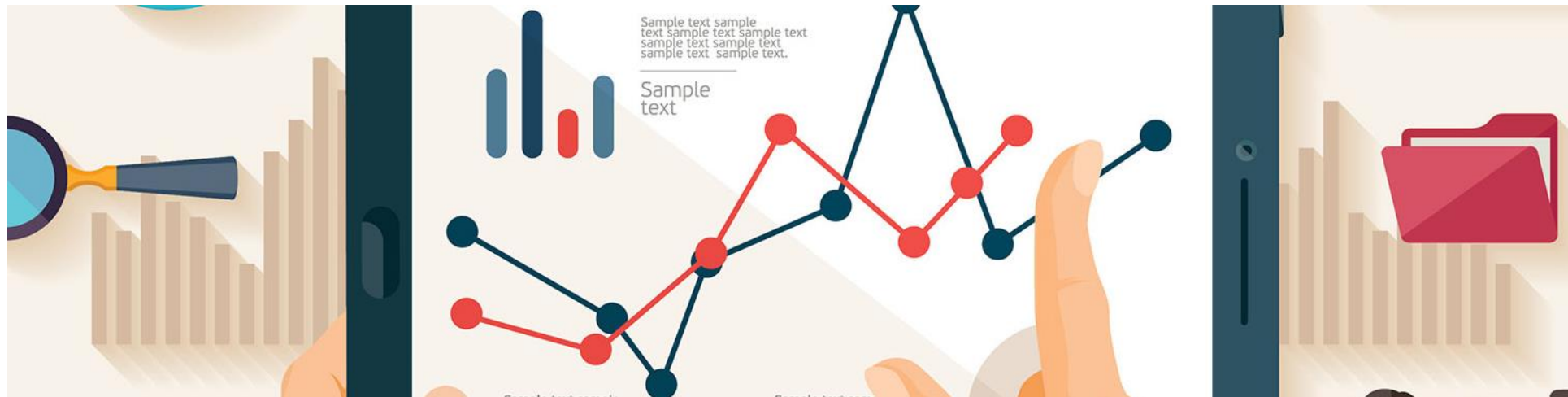
3

¿QUÉ COMPETENCIAS TENDRÍA EL “**EQUIPO ÓPTIMO**” DE CIENTÍFICOS DE DATOS?

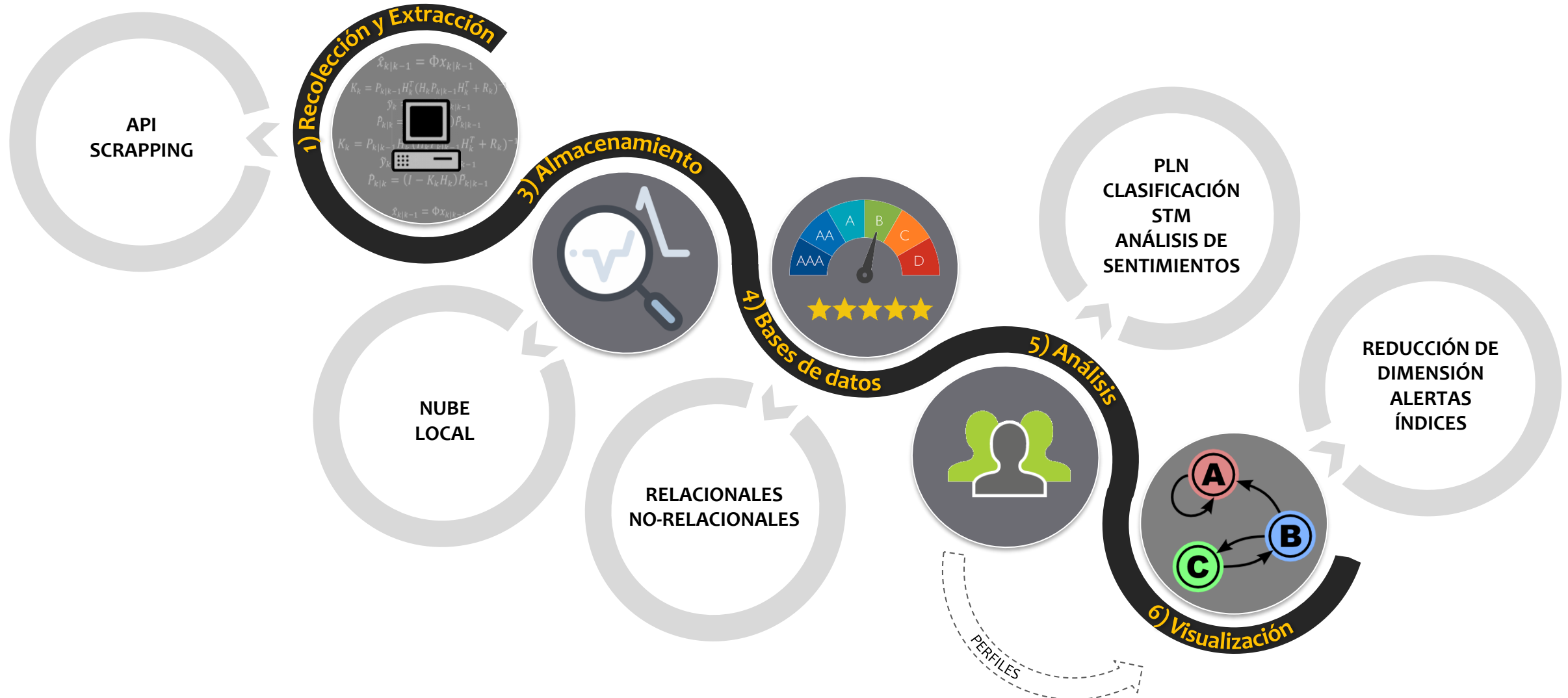
4

¿CUÁL ES UN **PROGRAMA DE “IMPLEMENTACIÓN TECNOLÓGICA” VIABLE**?

# I. ETAPAS DEL ANÁLISIS

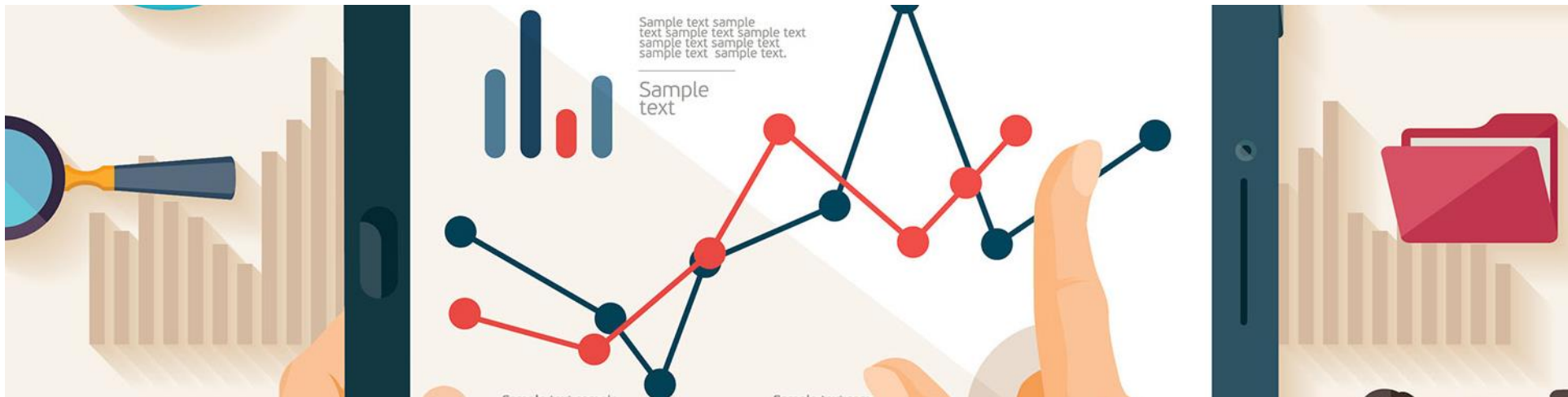


# ETAPAS DEL ANÁLISIS



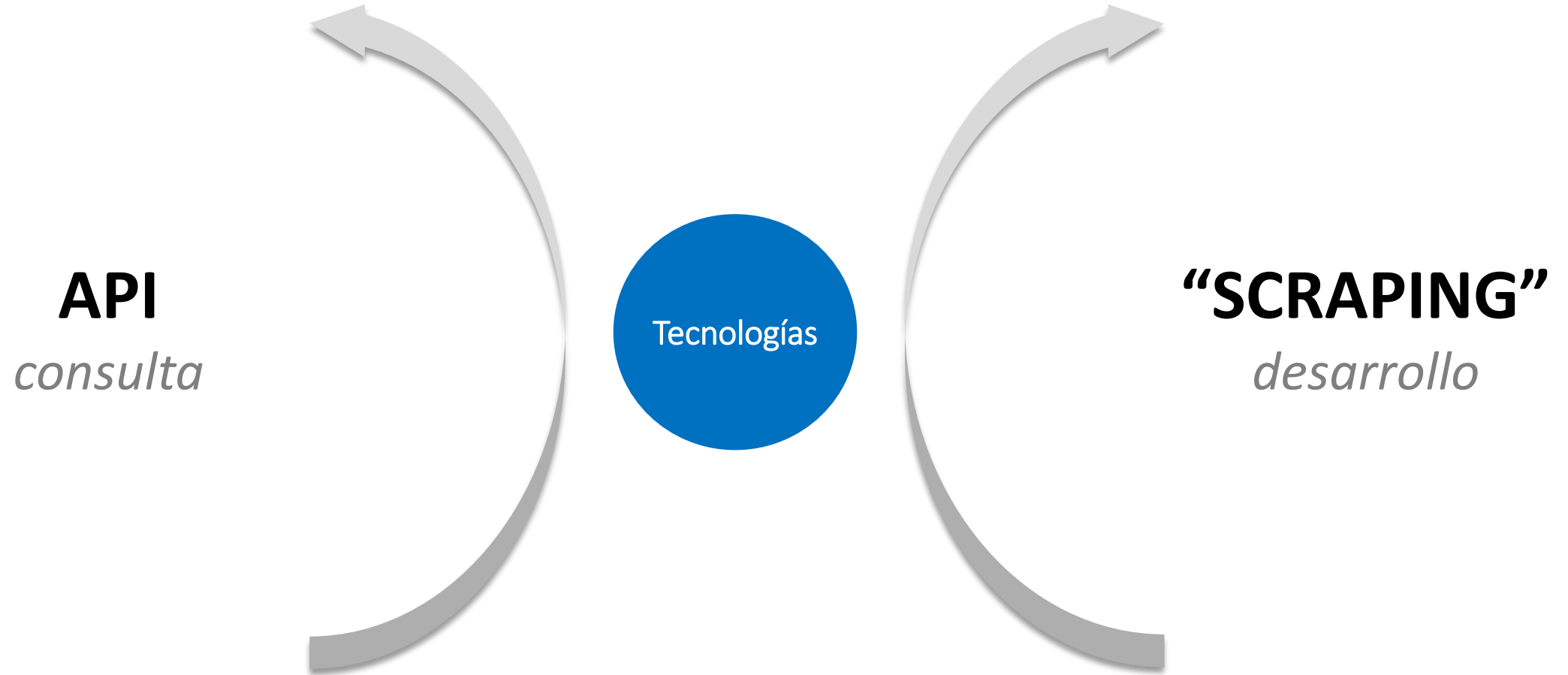
## II. ETAPAS DEL ANÁLISIS

# ALMACENAMIENTO





# EXTRACCIÓN – *TECNOLOGÍAS*



# EXTRACCIÓN – *EXTRACCIÓN DE DATOS WEB*

### *Definición*

Extraer información de páginas web de forma automatizada  
De protocolo HTTP manualmente o incrustando a un navegador en una aplicación.

Para realizar la extracción de los datos se requiere:

1. Conocer la estructura de la página; tener conocimiento en expresión regular (**regex**)
2. **Base de datos**
3. **Analizador (software o técnicas)**

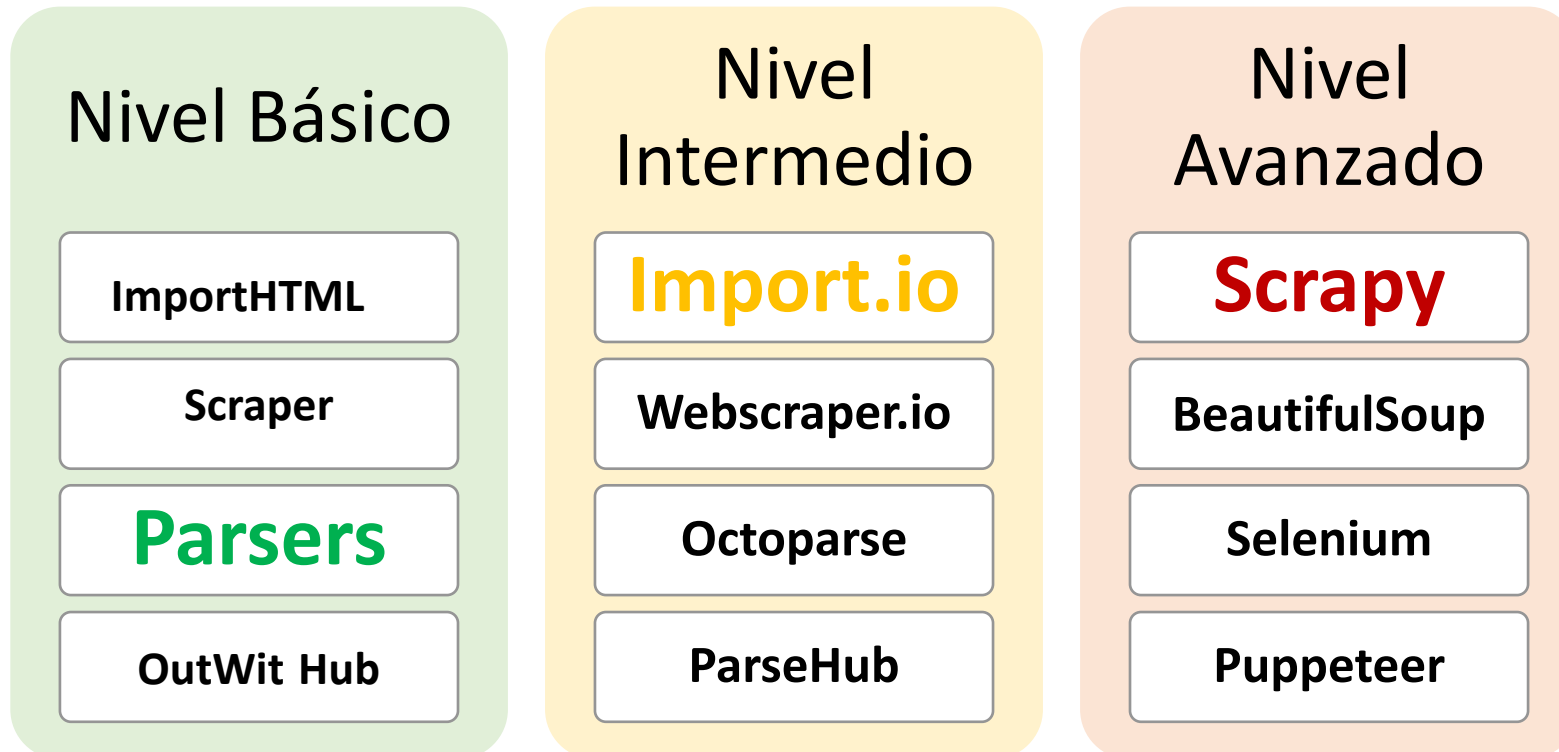
# EXTRACCIÓN – *EXTRACCIÓN DE DATOS WEB*

*¿Cuándo hacer  
“Scraping”?*

1. Páginas que contengan muchas tablas
2. Información dispersa en múltiples bases de datos o sitios.
3. Alta periodicidad de información
4. Analizar la estructura de la paginas
5. Cuenta con una API de conexión el sitio.
6. Recibir alertas de cambio en las bases de datos que se usan.

# EXTRACCIÓN – *EXTRACCIÓN DE DATOS WEB*

**Tecnologías según nivel de dificultad en su implementación:**



# EXTRACCIÓN – *EXTRACCIÓN DE DATOS WEB*

## Nivel Básico - COMPARATIVA

Tema	ImportHTML	Scraper Chrome	Parsers	OutWit Hub
Precio/mes	Gratis	Gratis	Gratis hasta 200 Uds	Gratis hasta 200 Usd
Paginación	Una	Múltiples	Múltiples	Múltiples
Soporta seguridad o autorización de los sitios	No	No	No	No
<b>Número de dominios que un proyecto puede funcionar</b>	NA	NA	Si	NA
<b>Complejidad de la estructura del sitio web.</b>	HTML	HTML, Javascript y Ajax	HTML, Javascript y Ajax	HTML, Javascript y Ajax
Trabajo de recopilación de datos recurrentes	No	No	Si	Si
¿Cuántos datos puedo recopilar?	Limitado	Ilimitado	Ilimitado	Ilimitado
Analizador sintáctico	No	No	No	No
<b>Conocimiento de programación</b>	No	No	No	No
Tener el programa puede ejecutarse en sus computadoras	No	No	No	No
Ejecutar en nuestras computadoras	Si	Si	Si	Si
URL	<a href="https://support.google.com/importhtml/">support.google.com</a>	<a href="https://chrome.webstore/web-scraper">chrome.webstore/web-scraper</a>	<a href="https://parsers.me/">parsers.me/</a>	<a href="https://outwit.com">outwit.com</a>

\* En negritas los criterios especialmente importantes

# EXTRACCIÓN – *EXTRACCIÓN DE DATOS WEB*

## Nivel Básico - **Parsers**



**Extracción:** 5000 páginas.

Total 1 440 000 páginas.

Sitios web ilimitados.

**Restricciones:**

2 sitio web a la vez; (La velocidad de extracción de datos se vuelve más lenta al realizar la extracción en más de 2 sitios diferentes)

**Soporte:** por Chat, correo electrónico y soporte comunitario

**Ventajas:**

Proxies estándar (velocidad y estabilidad)

- Se utilizan para trabajar sin bloqueo de IP y captcha.
- Velocidad de proxy de hasta 100 Mbps.

20 solicitudes concurrentes

-Máximo de descargas simultaneas.

**Inicio automático extracción programados.**

**Costo:** Lite \$20 UDS/mes

The screenshot shows the 'parsers' web interface. At the top, there's a logo and a 'Select mode' toggle switch. Below, there's a list of fields to be extracted from a website. Each field has a text input and a dropdown menu to the right. The fields are: Title (with a green dropdown icon), Spotify says it paid \$340M to buy Gimlet a (with a gear icon), Author (with a green dropdown icon), Jon Russell (with a gear icon), Username (with a green dropdown icon), @jonrussell (with a gear icon), Article (with an orange dropdown icon), and Spotify doubled down on podcasts last we (with a gear icon). At the bottom, there's a button that says 'Upgrade to Premium for more pages' and a 'Start' button. A status bar at the very bottom indicates '1000 pages free per site'.

# EXTRACCIÓN – *EXTRACCIÓN DE DATOS WEB*

## Nivel intermedio - **COMPARATIVA**

Tema	Import.io	Webscraper.io	Octoparse	ParseHub
Precio/mes	Gratis hasta (cotizar)	Gratis hasta \$300 uds	Gratis hasta \$500 uds	Gratis hasta \$209 uds
Paginación	Multiples	Multiples	Multiples	Multiples
Soporta seguridad o autorización de los sitios	Si	Si	Si	Si
<b>Número de dominios que un proyecto puede funcionar</b>	Si	Si	Si	Si
<b>Complejidad de la estructura del sitio web.</b>	HTML, Javascript, cookies, redirecciones y Ajax	HTML, Javascript, cookies, redirecciones y Ajax	HTML, Javascript, cookies, redirecciones y Ajax	HTML, Javascript, cookies, redirecciones y Ajax
Trabajo de recopilación de datos recurrentes	Si	Si	Si	Si
¿Cuántos datos puedo recopilar?	Ilimitado	Ilimitado	Ilimitado	Limitado
Analizador sintáctico	Si	Si	Si	Si
<b>Conocimiento de programación</b>	No	Si	No	No
Tener el programa puede ejecutarse en sus computadoras	Si	No	No	No
Ejecutar en nuestras computadoras	Si	Si	Si	Si
URL	<a href="https://www.import.io/">https://www.import.io/</a>	<a href="https://webscraper.io/">https://webscraper.io/</a>	<a href="https://www.octoparse.com/">https://www.octoparse.com/</a>	<a href="https://www.octoparse.com/">https://www.octoparse.com/</a>

\* En negritas los criterios especialmente importantes

# EXTRACCIÓN – *EXTRACCIÓN DE DATOS WEB*

## Nivel Intermedio - **Import.io**

**Extracción:** Miles de millones de consultas, millones de sitios web, terabytes de datos.

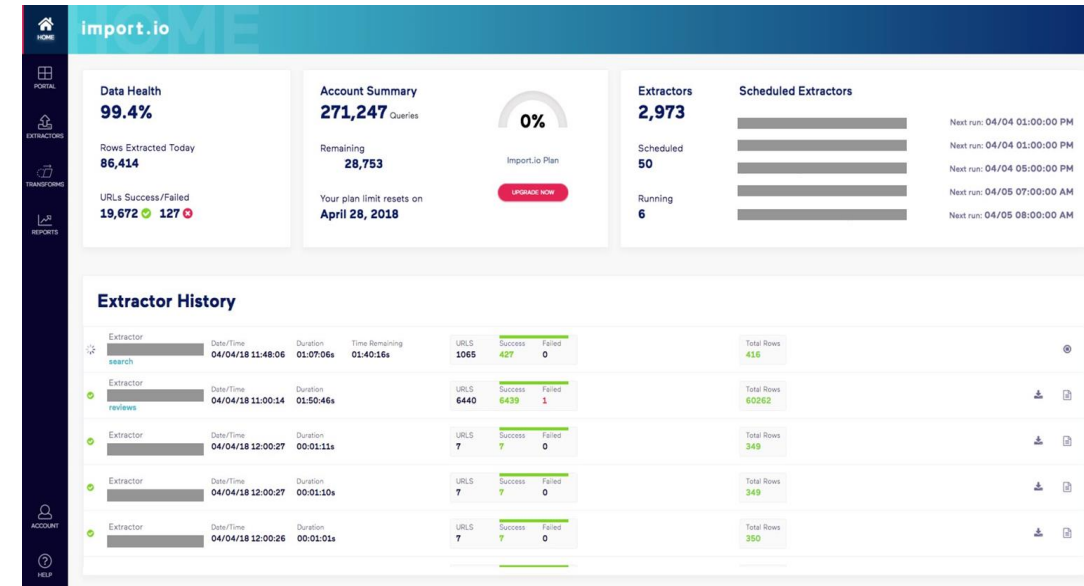


**Soporte:** por Chat, correo electrónico y soporte comunitario

### Ventajas:

- Almacenamiento de datos web, transformación y automatización de canalizaciones.
- Detección de anomalías, validación, alertas, control de calidad y mantenimiento.
- Nunca bloqueado: gestión del tráfico, reintentos, extracción geográfica.
- Programación periódica.

**Costo:** Bajo cotización, depende de la cantidad de sitios web y de la cantidad de millones de páginas web que se monitorearan.





# EXTRACCIÓN – *EXTRACCIÓN DE DATOS WEB*

## Nivel avanzado - COMPARATIVA

Tema	scrapy	BeautifulSoup	Selenium	Puppeteer
Precio X mes	Gratis	Gratis	Gratis	Gratis
Paginación	Múltiples	Múltiples	Múltiples	Múltiples
Soporta seguridad o autorización de los sitios	Si	Si	Si	Si
<b>Número de dominios que un proyecto puede funcionar</b>	Si	Si	Si	Si
<b>Complejidad de la estructura del sitio web.</b>	HTML, Javascript, cookies y Ajax	HTML, Javascript, cookies, PHP, Python y Ajax	HTML, Javascript, cookies, PHP, Python y Ajax	HTML, Javascript, cookies y Ajax
Trabajo de recopilación de datos recurrentes	Si	Si	Si	Si
¿Cuántos datos puedo recopilar?	Ilimitado	Ilimitado	Ilimitado	Ilimitado
Analizador sintáctico	Si	Si	Si	Si
<b>Conocimiento de programación</b>	Si (Python )	Si (Python2 , 3)	Si (Rubí, Java, Python, C#, JavaScript)	Si (Python)
Tener el programa puede ejecutarse en sus computadoras	Si	Si	Si	Si
Ejecutar en nuestras computadoras	Si	Si	Si	Si
Url	<a href="https://scrapy.org/">https://scrapy.org/</a>	<a href="https://www.crummy.com/">https://www.crummy.com/</a>	<a href="https://www.selenium.dev/">https://www.selenium.dev/</a>	<a href="https://github.com/puppeteer">https://github.com/puppeteer</a>

# EXTRACCIÓN – *EXTRACCIÓN DE DATOS WEB*

## Nivel avanzado – **Scrapy**

### Extracción:

- Compatible con Python 2.7 y Python 3.3
- Asíncrono

### Soporte:

- Documentación oficial
- Comunidad
- Consultoría

### Ventajas:

Robusto,

- Funciones para procesar XML y HTML.
- Eficiente en el uso de memoria RAM y CPU.
- Exportar datos a archivos CSV, Excel, XML o JSON, Interacción con bases de datos MySQL y MongoDB.

### Otro:

Flexible,

- Se complementa con **Selenium**
  - Soporta diverso lenguaje programación
  - Desarrollos JavaScript
- Se admite como analizador en **Beautiful Soup**




# Scrapy

An open source and collaborative framework for extracting the data you need from websites. In a fast, simple, yet extensible way.

Maintained by **Scrapinghub** and **many other contributors**

PyPI v2.1.0 wheel yes coverage 84%

Install the latest version of Scrapy

 **Scrapy 2.1.0**

\$ pip install scrapy

PyPI

Conda

Release Notes

Build and run your  
web spiders

Terminal

```
$ pip install scrapy
$ cat > myspider.py <<EOF
import scrapy

class BlogSpider(scrapy.Spider):
    name = 'blogspider'
    start_urls = ['https://blog.scrapinghub.com']

    def parse(self, response):
        for title in response.css('.post-header>h2'):
            yield {'title': title.css('a ::text').get()}

        for next_page in response.css('a.next-posts-link'):
            yield response.follow(next_page, self.parse)
EOF
$ scrapy runspider myspider.py
```

# EXTRACCIÓN – *EXTRACCIÓN DE DATOS WEB*



Tecnologías  
**anti-scraping**

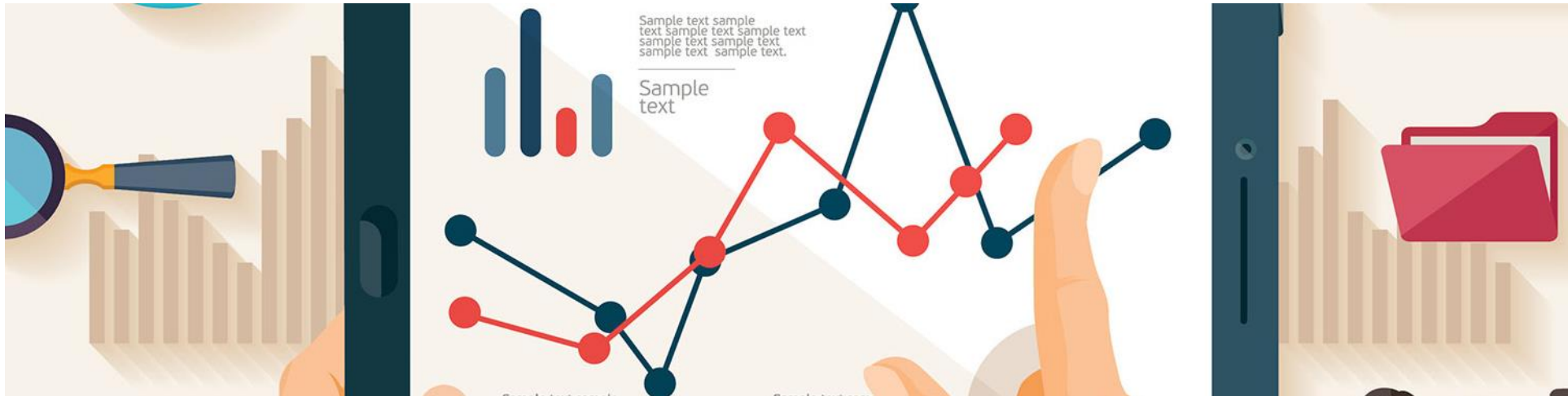
Velocidad de carga  
**lenta**

**Almacenamiento**  
de datos limitado

En función de la  
**complejidad de la**  
**estructura**

## II. ETAPAS DEL ANÁLISIS

# ALMACENAMIENTO

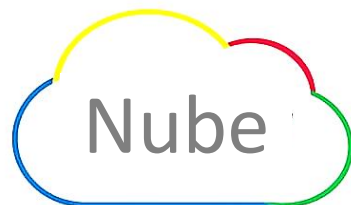


# ALMACENAMIENTO

## Definición

El almacenamiento de datos es el proceso mediante el cual la tecnología de la información archiva, organiza y comparte los bits y bytes que conforman los sistemas de los que dependemos todos los días, desde las aplicaciones hasta los protocolos de red, los documentos, el contenido multimedia, las libretas de direcciones y las preferencias del usuario.

## Tipos



Almacenamiento  
definido por **motores**  
de **BD**

Almacenamiento por  
**archivos**

Almacenamiento por  
**objetos** (JS, librerías)

Almacenamiento por  
**bloques**

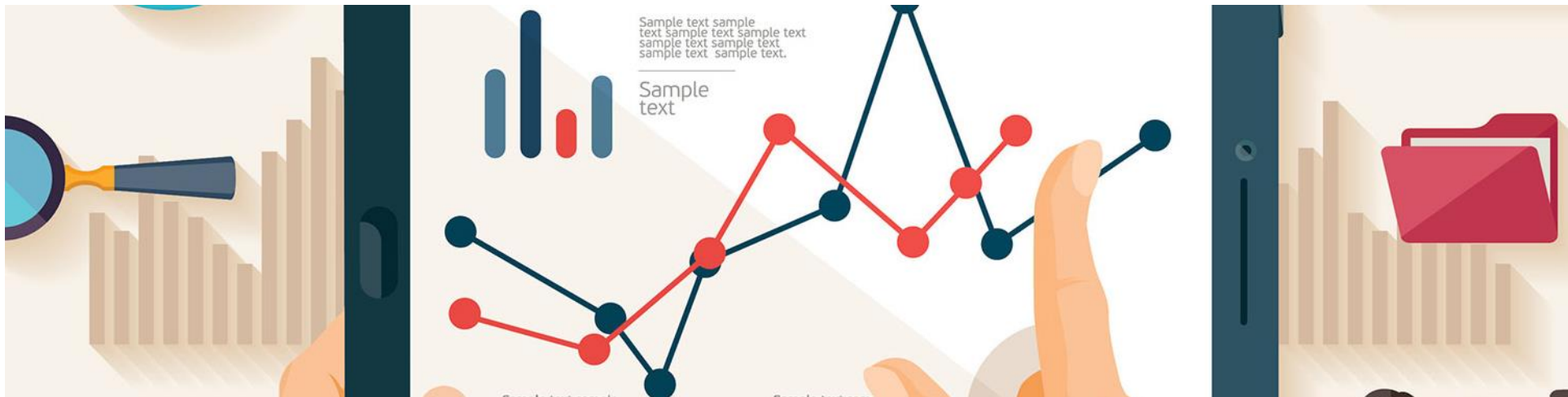
# Servicios en la NUBE



	Tecnología		
Proveedores	Software as a Service (SaaS)	Platform as a Service (PaaS)	Infrastructure as a Service (IaaS)
AWS (AMZON)			
GOOGLE CLOUD			
MICROSOFT AZURE			
IBM-WATSON			
AT&T			
HP			
DELL			

## II. ETAPAS DEL ANÁLISIS

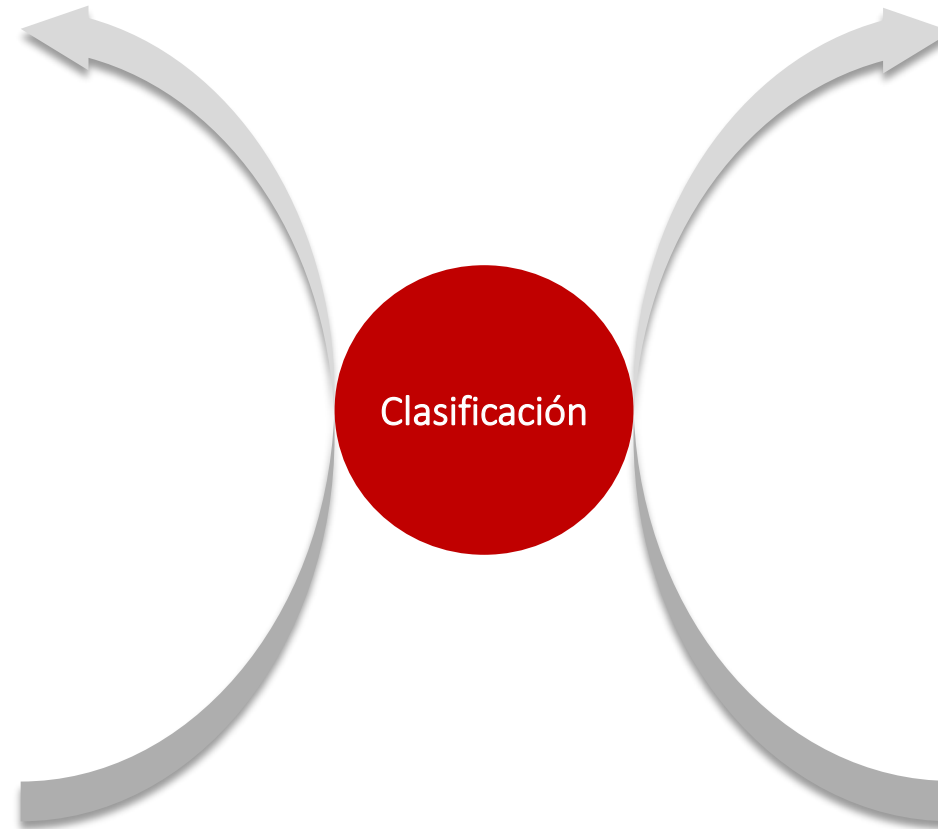
# **BASES DE DATOS**



# BASES DE DATOS

DATOS RELACIONALES
MySQL
MariaDB
PostgreSQL
Microsoft SQL Server (interfaz con lenguajes de Analítica)
Oracle (interfaz con lenguajes de Analítica)

- El volumen de datos crece gradualmente
- Cuando los datos son almacenados pueden ser calculados
- No tienen picos de accesibilidad por parte del usuario



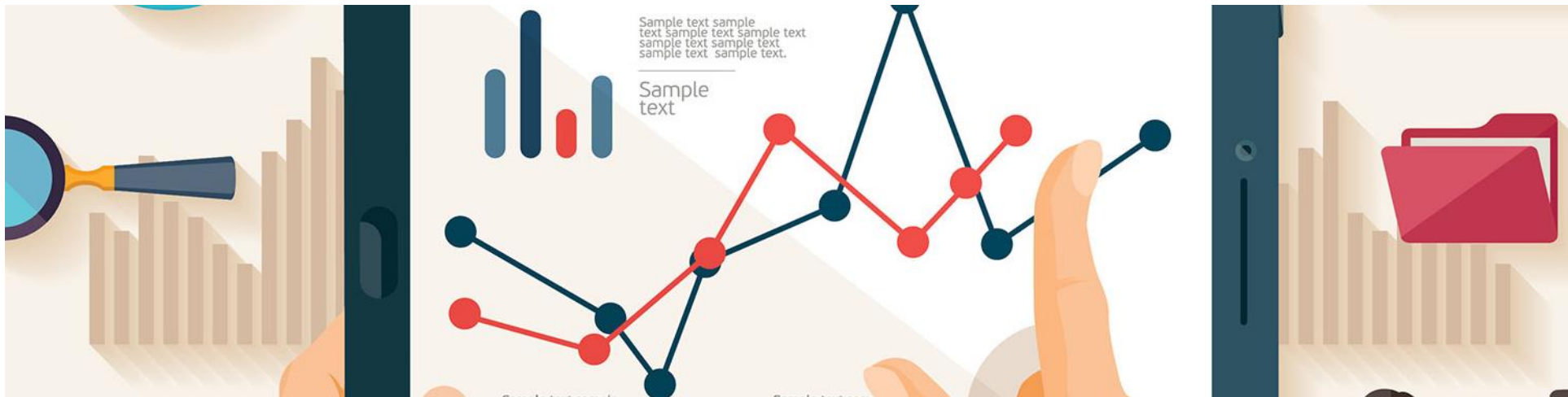
DATOS NO-RELACIONALES
MongoDB
Redis
Cassandra

- El volumen de los datos crece rápidamente y en momentos puntuales
- No se puede prever la cantidad de datos para almacenar
- Tiene picos de accesibilidad por parte de múltiples usuarios



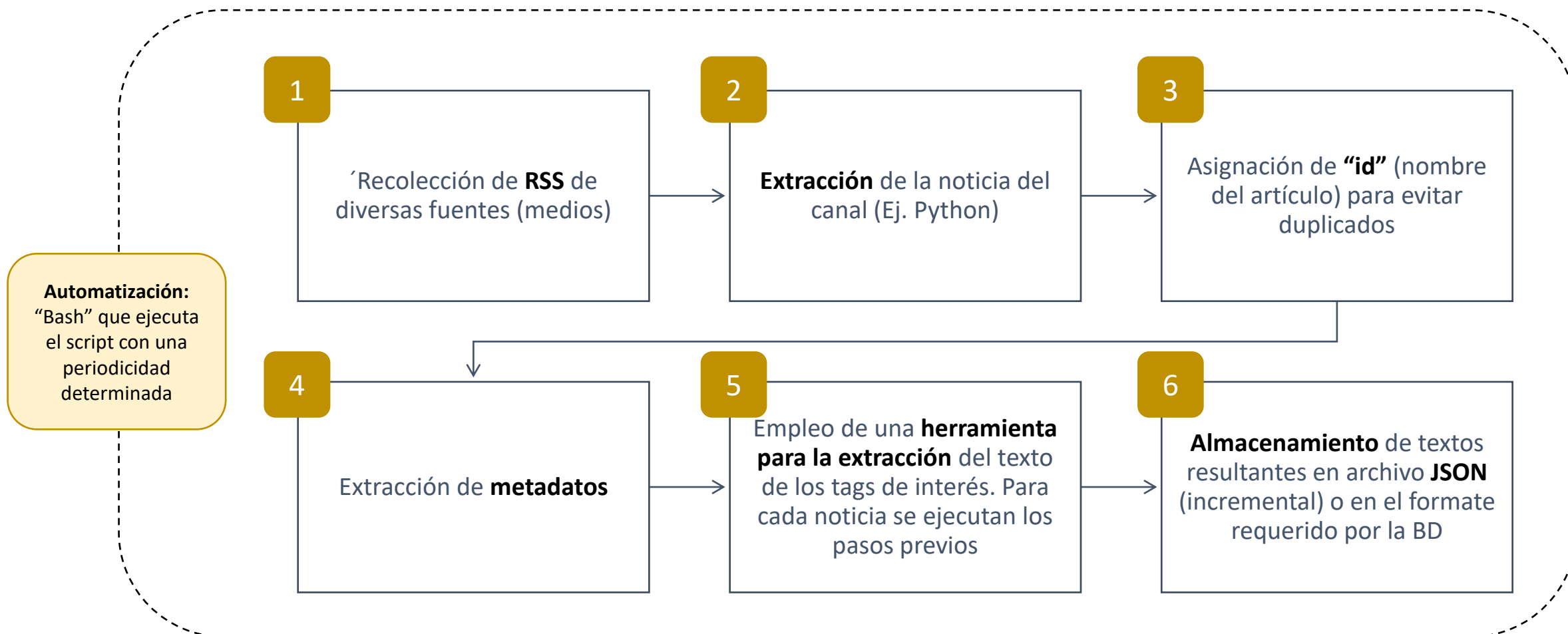
## II. ETAPAS DEL ANÁLISIS

# MÉTODOS ANALÍTICOS



# MÉTODOS ANALÍTICOS – *WEB SCRAPING DE NOTICIAS*

## CANAL *RSS* (REALLY SIMPLE SYNDICATION)



# MÉTODOS ANALÍTICOS – *WEB SCRAPING DE NOTICIAS*

## RETOS

### LIMPIEZA DE LOS DATOS

- **Pre-proceso** de texto para **cada una** de las fuentes, principalmente funciones basadas en expresiones regulares

### HETEROGENEIDAD DE FORMATOS

- Cada una de las fuentes de información considera un **formato distinto** para la publicación de las noticias

### INFORMACIÓN “IRRELEVANTE”

- Un detalle a considerar es que el scraping varía de fuente a fuente. Algunas fuentes de información en línea podrían presentar información de **anuncios** y acarrear información no relevante

### ESTILOS DE REDACCIÓN

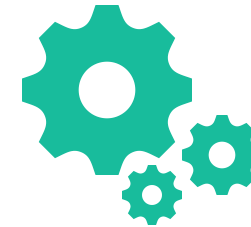
- Según el segmento al que está dirigida la noticia, podrían encontrarse **“sesgos”** en el uso de un lenguaje especializado o coloquial

# MÉTODOS ANALÍTICOS – *WEB SCRAPING DE NOTICIAS*

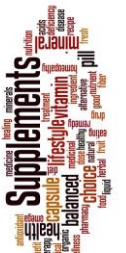
## A. Pre-procesamiento



## B. Análisis



### 1) Normalización



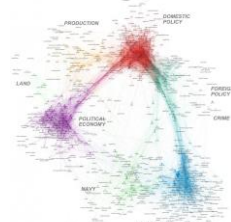
### 2) "Tokenización"



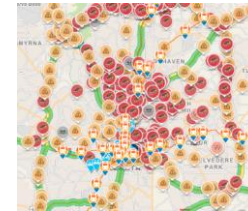
### 3) Representación documentos/textos



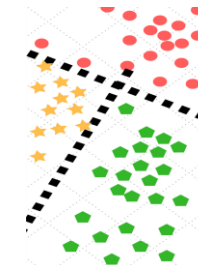
### 4) Reconocimiento de entidades



### 5) Georreferenciación



### 6) Clasificación



### 7) Análisis de sentimientos



*\* En rojo los métodos de implementación relativamente inmediata*

## ÁREAS DEL CONOCIMIENTO:

PROCESAMIENTO DE  
LENGUAJE NATURAL

MINERÍA DE TEXTOS

ANÁLISIS MULTIVARIADO

MODELACIÓN ESTADÍSTICA

# MÉTODOS ANALÍTICOS – *WEB SCRAPING DE NOTICIAS*

## TÉCNICAS ANALÍTICAS

### *PLN*

El Procesamiento del Lenguaje Natural (**PLN**) es un campo de las ciencias de la computación, inteligencia artificial y lingüística que estudia las **interacciones entre las computadoras y el lenguaje humano**.

Algunas aplicaciones: **reconocimiento de voz, la traducción entre idiomas, la comprensión de oraciones completas, la corrección de ortografía**, entre otros.

### 1

### *NORMALIZACIÓN*

Conveniente transformar el texto a una **forma estándar**.

Ejemplos:

- **convertir todos los textos a minúsculas o mayúsculas,**
- **eliminar los signos de puntuación o**
- **convertir los números a sus equivalentes de palabras.**

# MÉTODOS ANALÍTICOS – *WEB SCRAPING DE NOTICIAS*

## TÉCNICAS ANALÍTICAS

2

### *SEGMENTACIÓN* *“tokenización”*

La normalización incluye la “**tokenización**” de los textos, la cual se puede realizar de diferentes maneras:

- **N-gramas**: combina las palabras que se encuentran juntas con fines de representación
- **K-skip-n-gramas**: toma las cadenas de texto que se forman entre “saltos” de palabras, es decir, omitiendo palabras que se encuentren entre ellas.

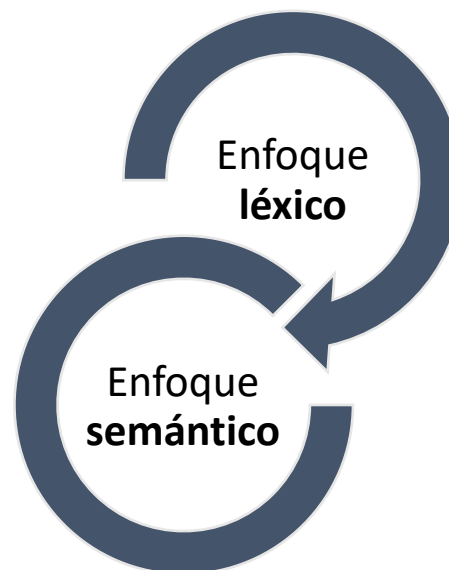
# MÉTODOS ANALÍTICOS – *WEB SCRAPING DE NOTICIAS*

## TÉCNICAS ANALÍTICAS

3

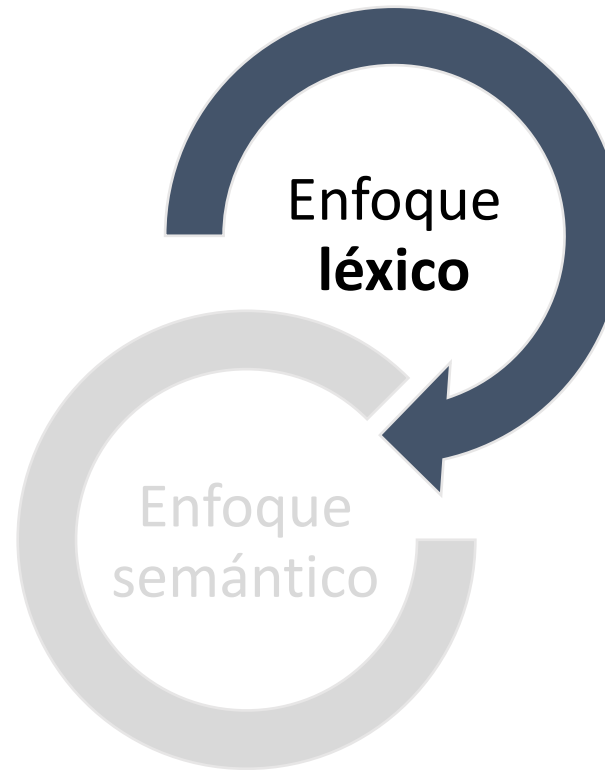
*Representación  
de documentos*

Para que un algoritmo de clasificación pueda capturar relaciones entre datos **requiere pasar de un conjunto de textos a datos estructurados**. Existen diversas formas de representar los documentos en un esquema estructurado, particularmente en este documento se analizan dos enfoques:



# MÉTODOS ANALÍTICOS – *WEB SCRAPING DE NOTICIAS*

## 3 *Representación de documentos*





# MÉTODOS ANALÍTICOS – *WEB SCRAPING DE NOTICIAS*

3

## *Representación de documentos*

### ENFOQUE LÉXICO

Es posible representar con unidades léxicas (los tokens) el contenido de un conjunto de documentos.

Funciona muy bien en algunas tareas de aprendizaje automático como:

- Detección de **spam**,
- **Clasificador de sentimientos**, entre otros.

Inconvenientes:

- **Ignora el orden y la gramática** de los documentos, por lo que se pierde el contexto en el que se usa una palabra.
- **La matriz** generada altamente **dispersa y sesgada hacia las palabras más comunes**

# MÉTODOS ANALÍTICOS – *WEB SCRAPING DE NOTICIAS*

## 3 *Representación de documentos*

### ENFOQUE LÉXICO

#### TF-IDF

Es una forma de ponderar las palabras del vocabulario para dar un peso en proporción al impacto que tiene en el significado de un documento.

La puntuación es un producto de 2 medidas independientes: la frecuencia de término (TF) y la frecuencia inversa de documento (IDF).

$$w_{ij} = \underbrace{tf_{ij}}_{TF} \times \underbrace{\log \left( \frac{N}{df_i} \right)}_{IDF}$$

Donde: **tf<sub>ij</sub>** es el número de ocurrencias del término i en el documento j, **df<sub>i</sub>** es el número de documentos que contienen el término i y N es el número total de documentos. La IDF es una medida de cuánta información proporciona el token, es decir, si es común o raro en los documentos.

# MÉTODOS ANALÍTICOS – *WEB SCRAPING DE NOTICIAS*

3

## *Representación de documentos*

### ENFOQUE LÉXICO

#### MODELOS BASADOS EN RASGOS O CARACTERÍSTICAS LÉXICAS

Caracterizan a los documentos mediante una lista de atributos que resumen los rasgos más significativos o relevantes

Estos modelos convierten los textos a una representación numérica para cada documento a través de:

- **Estadísticas de texto**
- **Características sintácticas**

Depende del fenómeno de estudio, por ejemplo, **análisis de sentimiento** en textos, una variable (característica) importante es el **número de palabras positivas**, el número de **palabras negativas** o alguna relación entre estas.

# MÉTODOS ANALÍTICOS – *WEB SCRAPING DE NOTICIAS*

3

## *Representación de documentos*

### ENFOQUE LÉXICO

#### MODELOS BASADOS EN RASGOS O CARACTERÍSTICAS LÉXICAS

- Podemos también representar textos mediante palabras o frases claves: *keywords* y *keyphrases*
- Proporcionan un tipo de **metadatos semánticos**
- La filtración se realiza por **métodos simbólicos**, donde se les aplica un **esquema de pesos a las frases** para asignarles un **score**.
- Se crea una **matriz indicadora**, donde cada renglón pertenece a una noticia, y cada fila pertenece a una *keyword*. Si la noticia contiene la *keyword*, se asigna el valor de 1, y se asigna 0 en otro caso. Con esto podemos crear una **representación matemática** para utilizar cualquier **método de clasificación**

# MÉTODOS ANALÍTICOS – *WEB SCRAPING DE NOTICIAS*

## 3 *Representación de documentos*

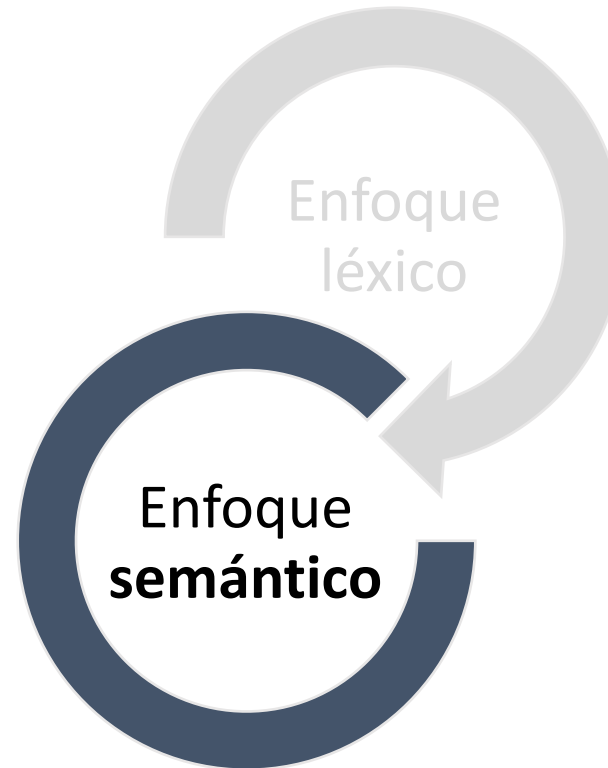
### ENFOQUE LÉXICO

#### KEYPHRASE EXTRACTION ALGORITHM

- Capaz de **identificar las keyphrases basándose en sus propiedades**, tales como frecuencia en el documento, presencia en partes significativas del documento, etc. (Emula comportamiento humano)
- Se usa la confluencia de términos semánticos, gracias a los **tesauros**. El resultado es un conjunto de términos gramaticales relacionados al contenido del documento.

# MÉTODOS ANALÍTICOS – *WEB SCRAPING DE NOTICIAS*

## 3 *Representación de documentos*



# MÉTODOS ANALÍTICOS – *WEB SCRAPING DE NOTICIAS*

3

## *Representación de documentos*

### ENFOQUE SEMÁNTICO

- Se refiere a los aspectos del **significado, sentido o interpretación** de signos lingüísticos como símbolos, palabras, expresiones o representaciones formales.
- A pesar de que los modelos desarrollados desde el enfoque semántico se basan también en los tokens, estos modelos buscan **capturar la mayor cantidad de información posible del contexto**
- Desde el enfoque semántico, los ***word embeddings*** son modelos utilizados para transformar las palabras a una representación estructurada. En estos modelos las palabras semánticamente similares se encontrarán cerca entre ellas.

# MÉTODOS ANALÍTICOS – *WEB SCRAPING DE NOTICIAS*

## 3 *Representación de documentos*

### ENFOQUE SEMÁNTICO

#### WORD EMBEDDINGS

En el lenguaje natural las palabras no aparecen de manera aislada o aleatoria, **dependen de otras palabras y van formando una secuencia** de acuerdo con una estructura gramatical.

Un modelo de lenguaje puede ser representado por la **probabilidad condicional de la siguiente palabra dada las anteriores**, como:

$$\hat{P}(w_1^T) = \prod_{i=1}^T \hat{P}(w_i | w_1^{i-1}), \quad (5.5)$$

donde  $w_t$  es la  $t$ -ésima palabra, y la subsecuencia  $w_i^j = (w_i, w_{i+1}, \dots, w_{j-1}, w_j)$ .



# MÉTODOS ANALÍTICOS – *WEB SCRAPING DE NOTICIAS*

## 4

## *Reconocimiento de entidades*

### *Definición*

- La relevancia del reconocimiento de entidades de las noticias recae en el hecho que se pueden hacer **grafos de actores principales en determinada región**.
- Se puede hacer un **grafo de relación entre personajes** (personas y organizaciones. )
- Una de las ventajas es que las **entidades pueden ser dadas**, ejemplo, conocer todas las noticias relacionadas con el nombre **Andrés Manuel López Obrador** y analizar cuáles son los **tópicos relevantes** que le rodean

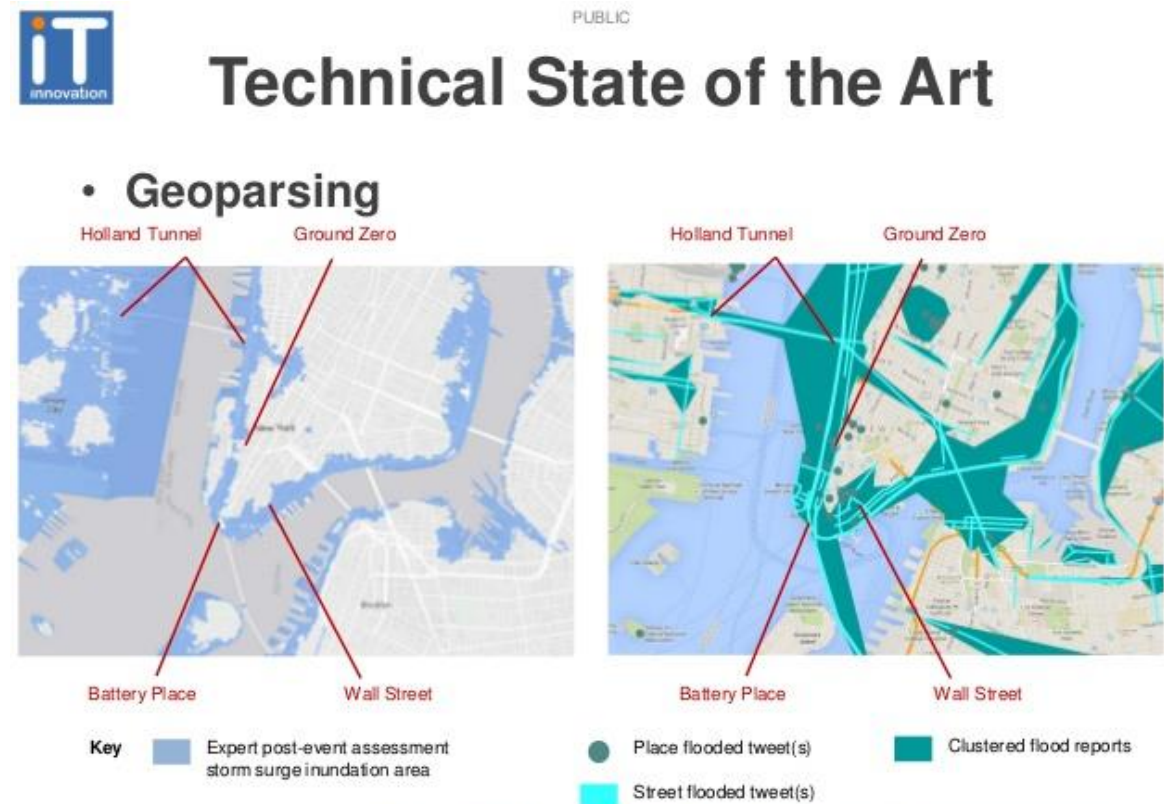
#### CASO MADISON:

El producto Informe de contexto podría aplicar este tipo de tecnología. Cabe señalar que es necesario también utilizar la georreferenciación ya que se requiere hacer un diagnóstico de un municipio, estado o región con relación a temas como Gobierno, Seguridad, Social o “trending topics” y la prevalencia de estos temas.

# MÉTODOS ANALÍTICOS – *WEB SCRAPING DE NOTICIAS*

## 5 *GEORREFERENCIACIÓN*

Para identificar el estado al que pertenece la noticia, se puede crear un **vector de estados y capitales y buscar la coincidencia de palabras del vector en cada noticia**. Posteriormente, se filtra por las noticias que sí contenían una o más palabras del vector.



Cite: Middleton, S.E. Middleton, L. Modafferi, S. "Real-time Crisis Mapping of Natural Disasters using Social Media", Intelligent Systems, IEEE, vol.29, no.2, pp.9,17, Mar.-Apr. 2014

© University of Southampton IT Innovation Centre, 2016

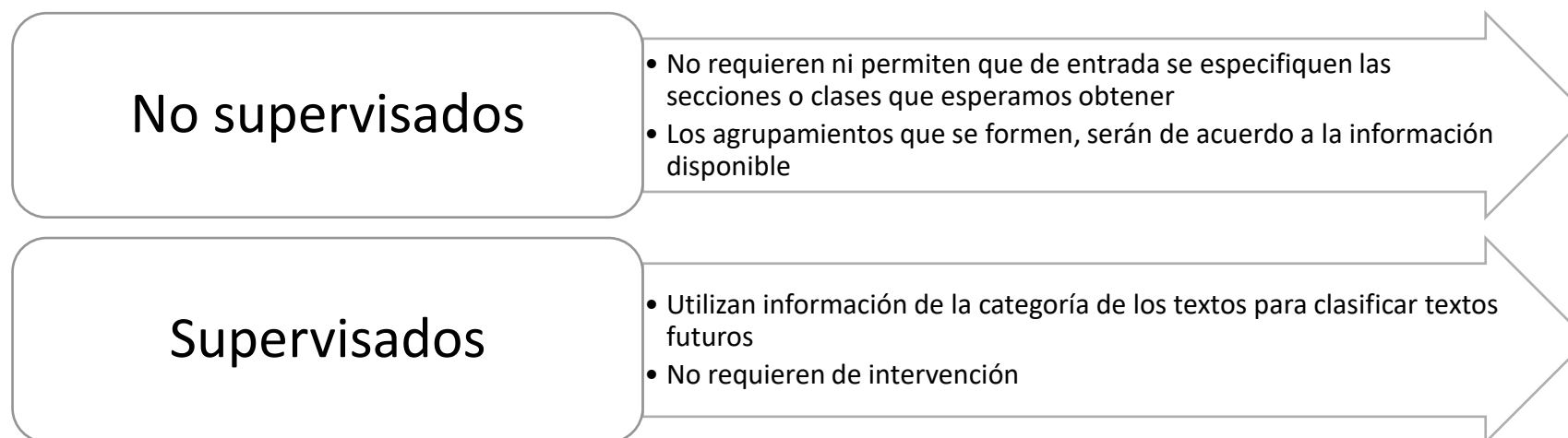
17

# MÉTODOS ANALÍTICOS – *WEB SCRAPING DE NOTICIAS*

## 6 CLASIFICACIÓN

En el esquema tradicional de un periódico existe una organización, las noticias generalmente se encuentran divididas por secciones como: **Política, Economía, Seguridad, entre otros**. Para ello utilizamos diferentes metodologías de modelación de tópicos, con el objetivo de **encontrar grupos de noticias con características similares y comparar con las categorías** de un periódico convencional.

Distinguimos métodos de agrupación:



# MÉTODOS ANALÍTICOS – *WEB SCRAPING DE NOTICIAS*

## 6 *CLASIFICACIÓN*

### Máquinas de Soporte Vectorial (SVM)

Este tipo de algoritmos buscan el hiperplano que tenga la máxima distancia (margen) con los puntos que estén más cerca de él mismo

### “Random Forest”

Alta precisión y manejo de una gran cantidad de variables y registros

### Regresión logística

Modelo de regresión que permiten estudiar si una variable (generalmente binomial) depende de un conjunto variables o características.

### AdaBoost

Sigue un Aprendizaje secuencial. Los modelos posteriores al de entrenamiento se construyen ajustando los valores de error residual del modelo inicial.

### Árboles de decisión

La construcción del árbol sigue un enfoque de división binaria recursiva, donde la tasa de error de clasificación se utiliza como criterio para la división binaria.

#### CASO MADISON:

La clasificación podría ser relevante para productos como **Monitoreo semanal de negocios** o **Reporte de contexto internacional**. Esto para poder asignar automáticamente las noticias de negocios y la sección internacional.

# MÉTODOS ANALÍTICOS – *WEB SCRAPING DE NOTICIAS*

## 6 *CLASIFICACIÓN*

**Modelo STM** (Clasificación que permite variables externas o de referencia)

- Permite **incorporar información de covariables** (metadatos) en el modelado de los tópicos.
- Los **metadatos se pueden ingresar** en el modelo de dos maneras: **prevalencia tópica y contenido tópico**, permitiendo afectar el uso de la tasa de palabras dentro de un tema dado.

### CASO MADISON:

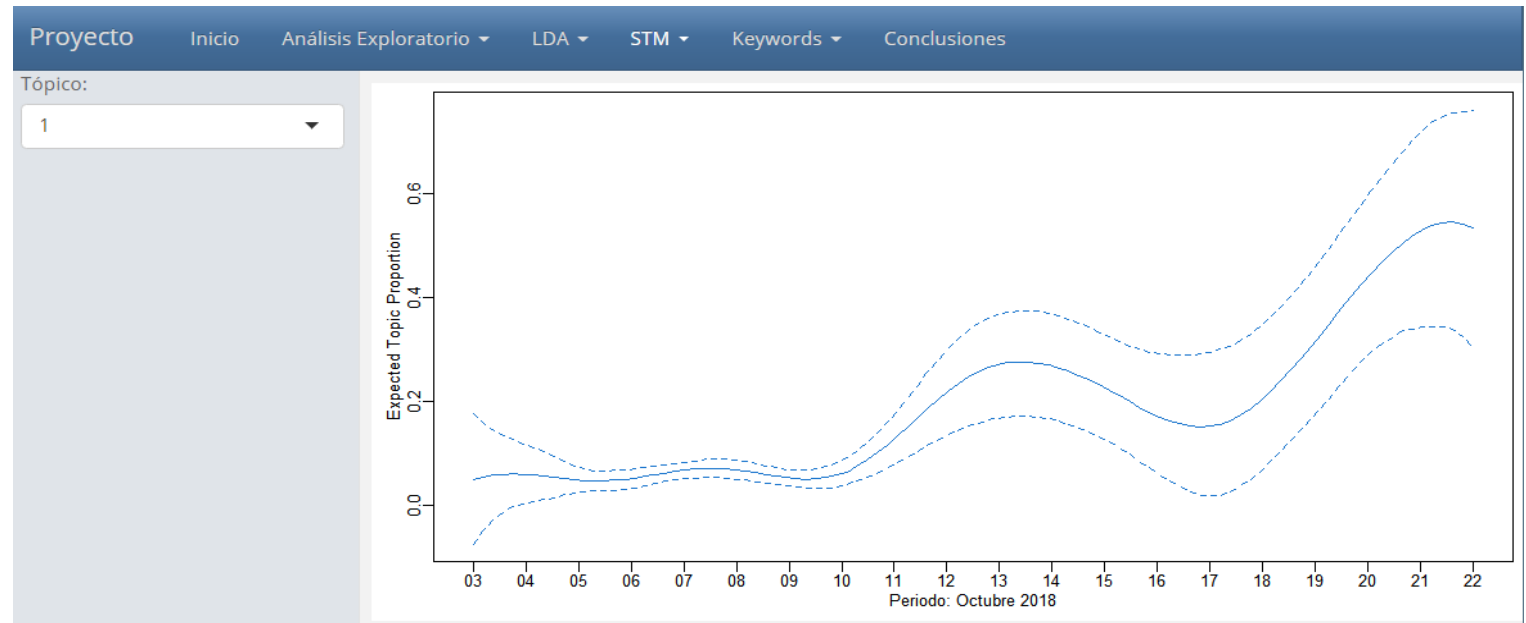
En un caso particular, los metadatos podrían consistir en la **fecha de publicación** de la noticia y el **periódico/medio** que la emite. El objetivo es usar la fecha para medir la prevalencia de los tópicos y la fuente de información (el periódico) y para analizar el contenido del tópico.

# MÉTODOS ANALÍTICOS – *WEB SCRAPING DE NOTICIAS*

## 6 *CLASIFICACIÓN*

**Modelo STM** (Clasificación que permite variables externas o de referencia)

**Nuevo Aeropuerto Internacional  
de México (NAIM)**  
*Octubre 2018*

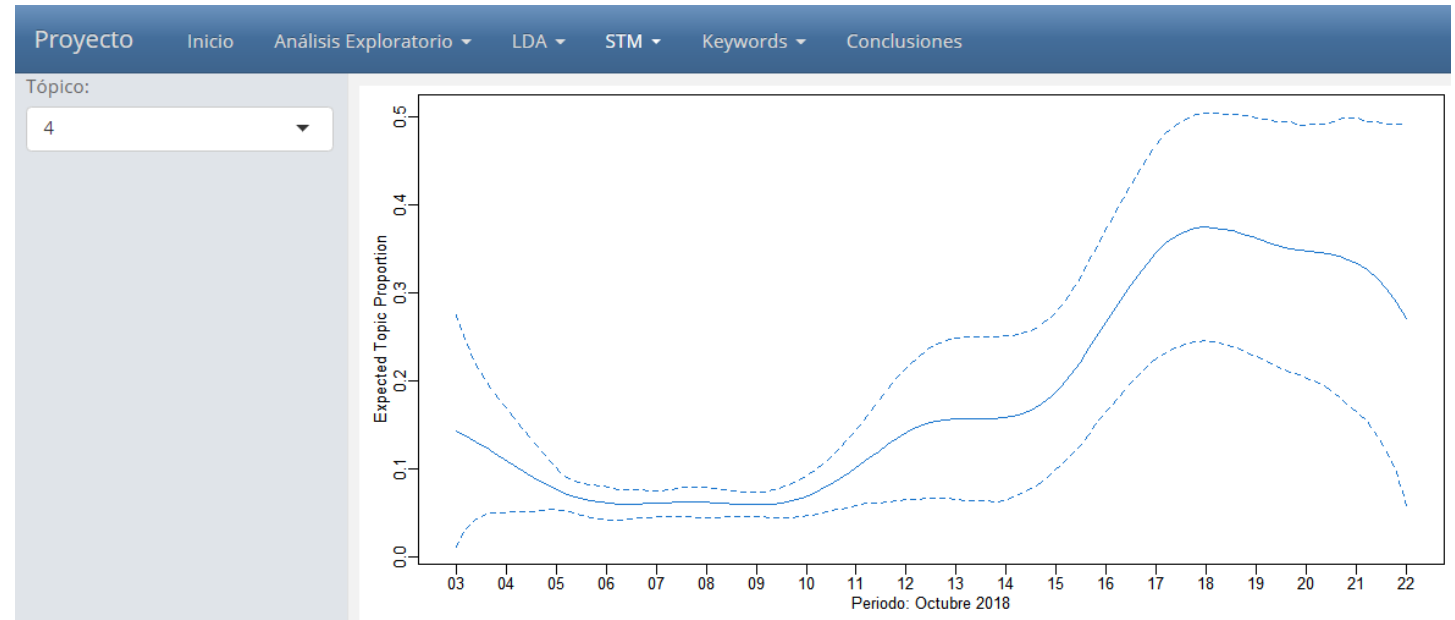


# MÉTODOS ANALÍTICOS – *WEB SCRAPING DE NOTICIAS*

## 6 *CLASIFICACIÓN*

**Modelo STM** (Clasificación que permite variables externas o de referencia)

**Caravana de migrantes**  
*Octubre 2018*

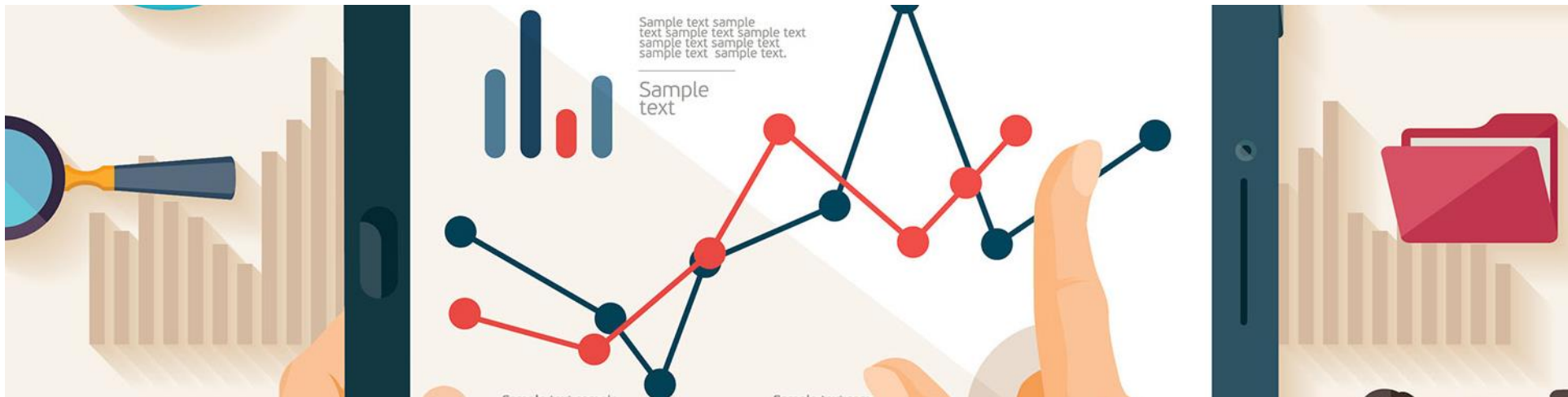


### CASO MADISON:

Podría brindar una perspectiva de la relevancia temporal del tópico, actor o entidad en general, particularmente para el producto de **Monitoreo diario de** medios, puesto que es fácilmente automatizable en el sentido que existen categorías específicas de tópicos que se buscan en las fuentes de información. Usando la prevalencia de los “*trending topics*” se podría presentar la información más relevante al día.

## II. ETAPAS DEL ANÁLISIS

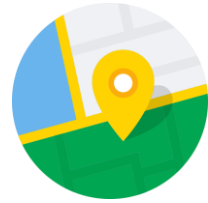
### VISUALIZACIÓN



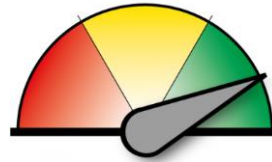


# VISUALIZACIÓN

*Según tipo de resultados*



Cartografía



Monitoreo de índices



Factores significativos



Indicadores/índices compuestos

*Software/ servicios*

SISTEMAS INFORMACIÓN GEOGRÁFICA



VISUALIZADORES GENÉRICOS



### III. ANÁLISIS DE **COMPETENCIAS**

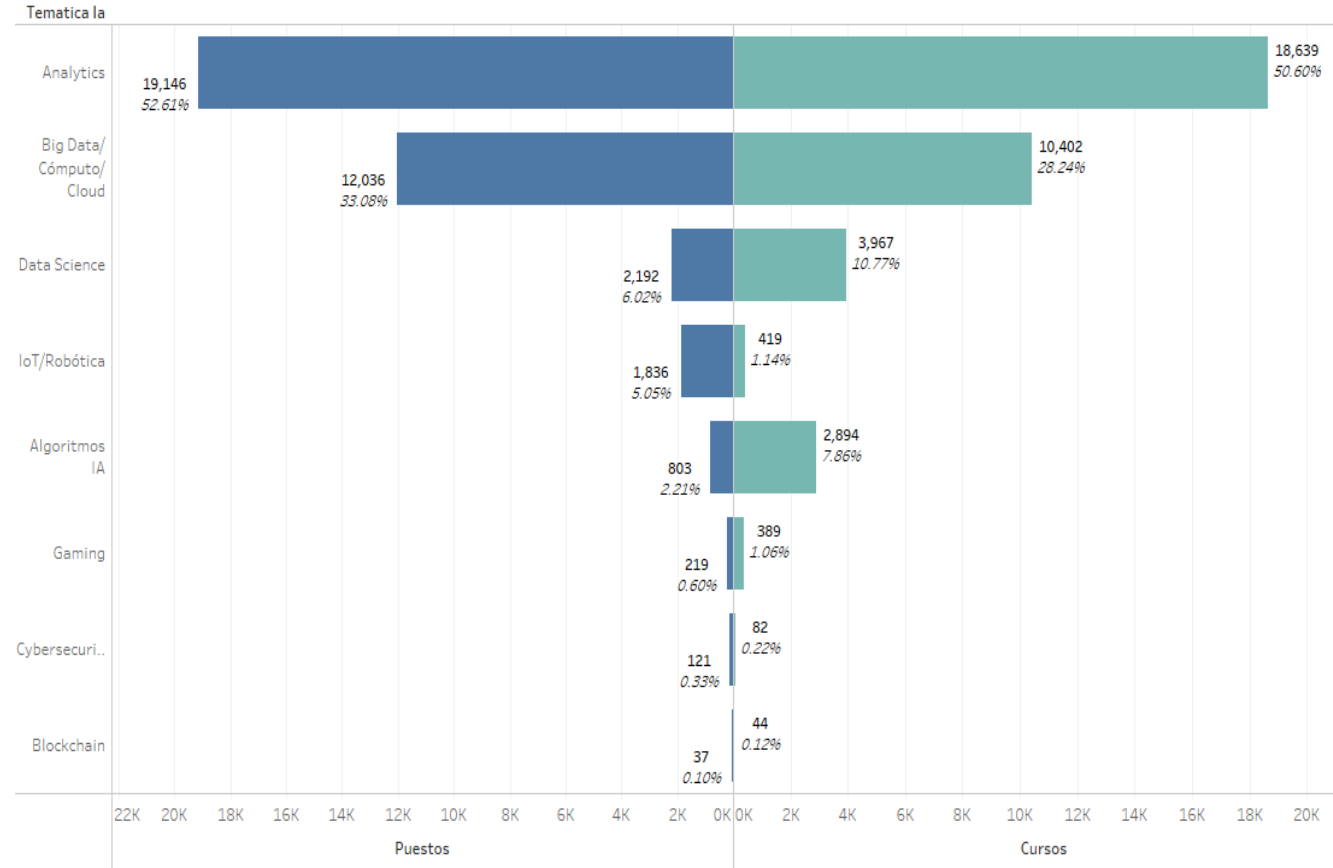


# MERCADO LABORAL & CAPACITACIÓN EN TEMÁTICAS “IA”

## Incidencia de búsquedas en Redes Sociales

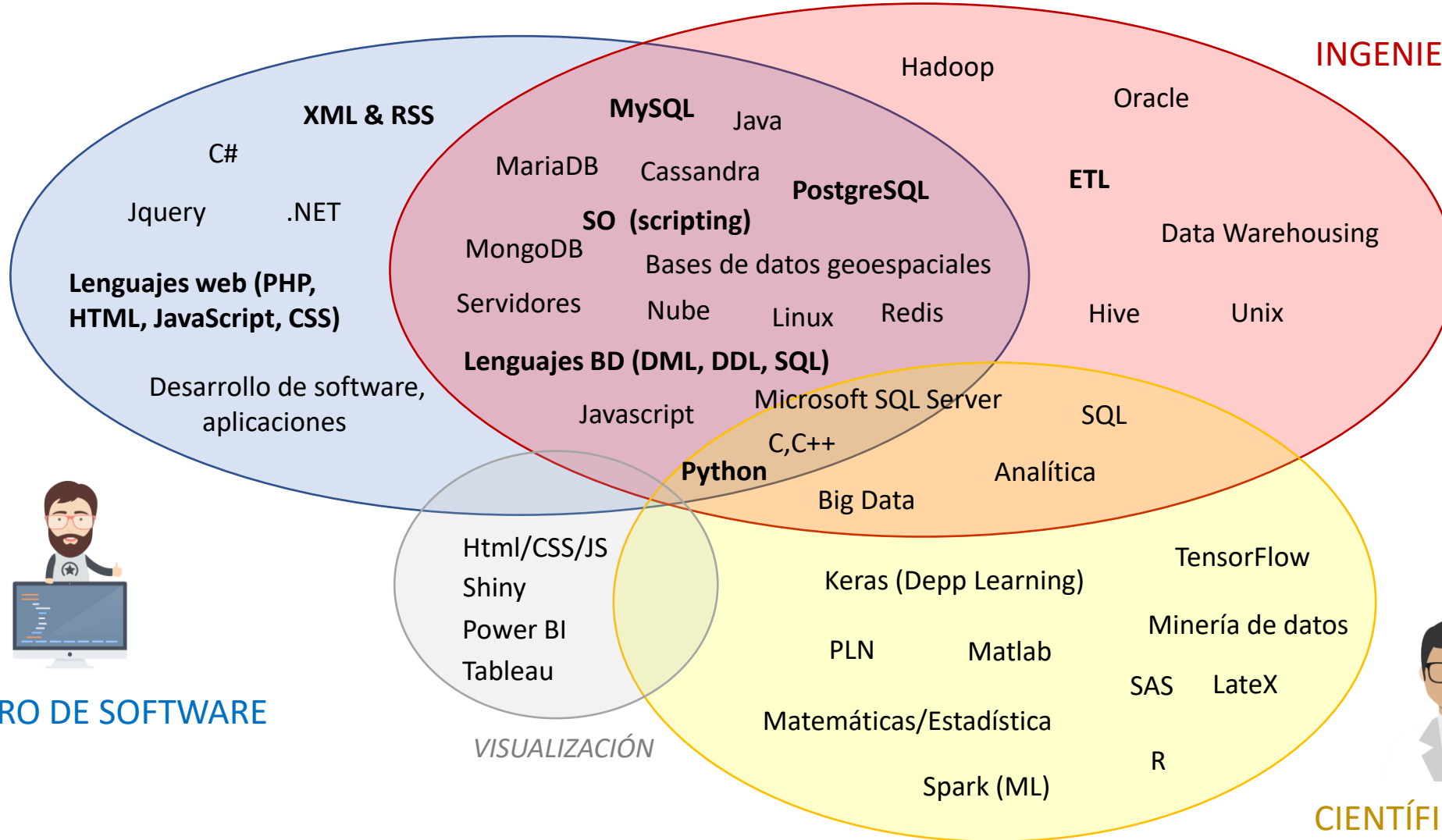
Método de Extracción: “Scraping”

(OCC Mundial, LinkedIn, Coursera)



# PERFILES

INGENIERO DE BASES DE DATOS



INGENIERO DE SOFTWARE

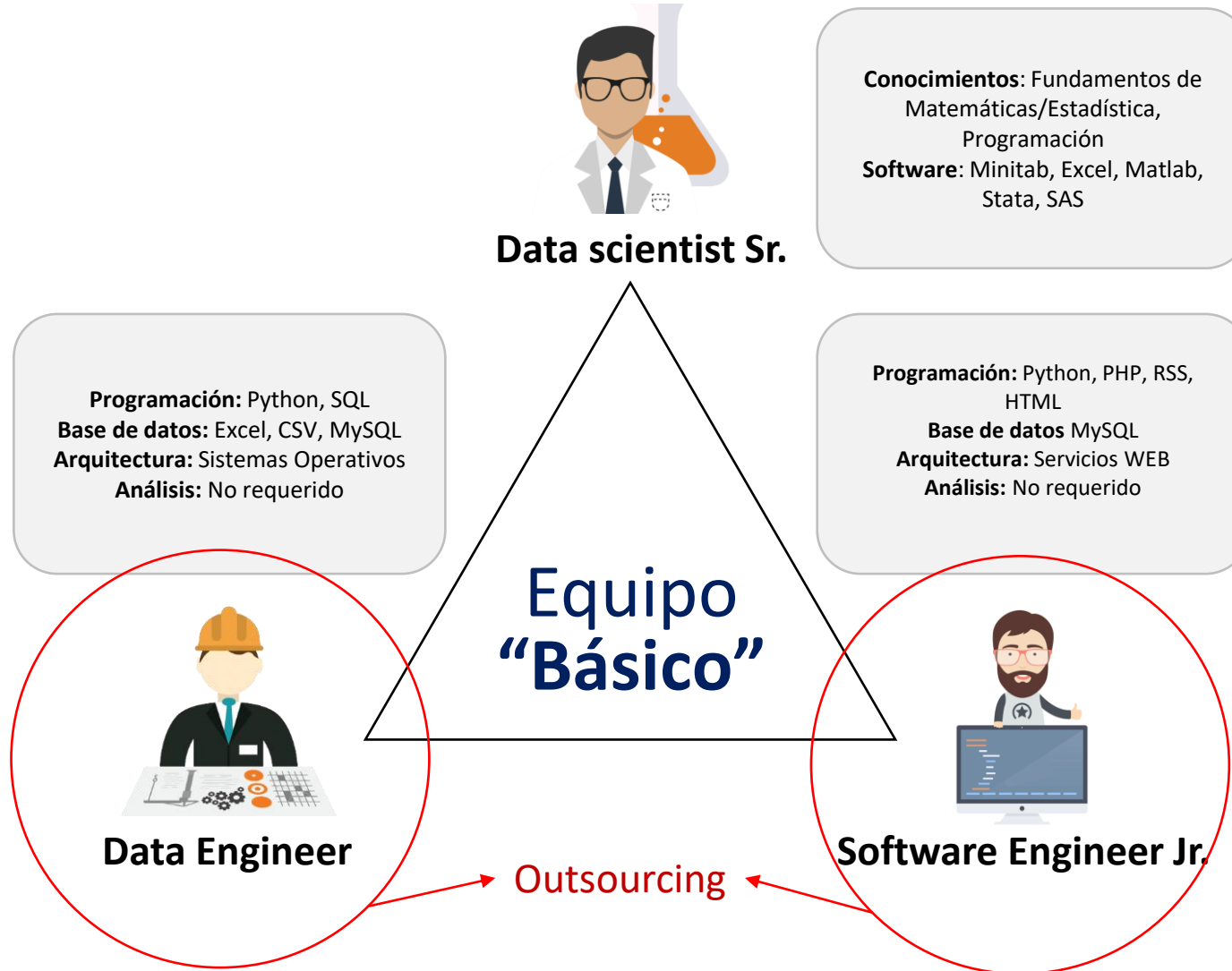


VISUALIZACIÓN

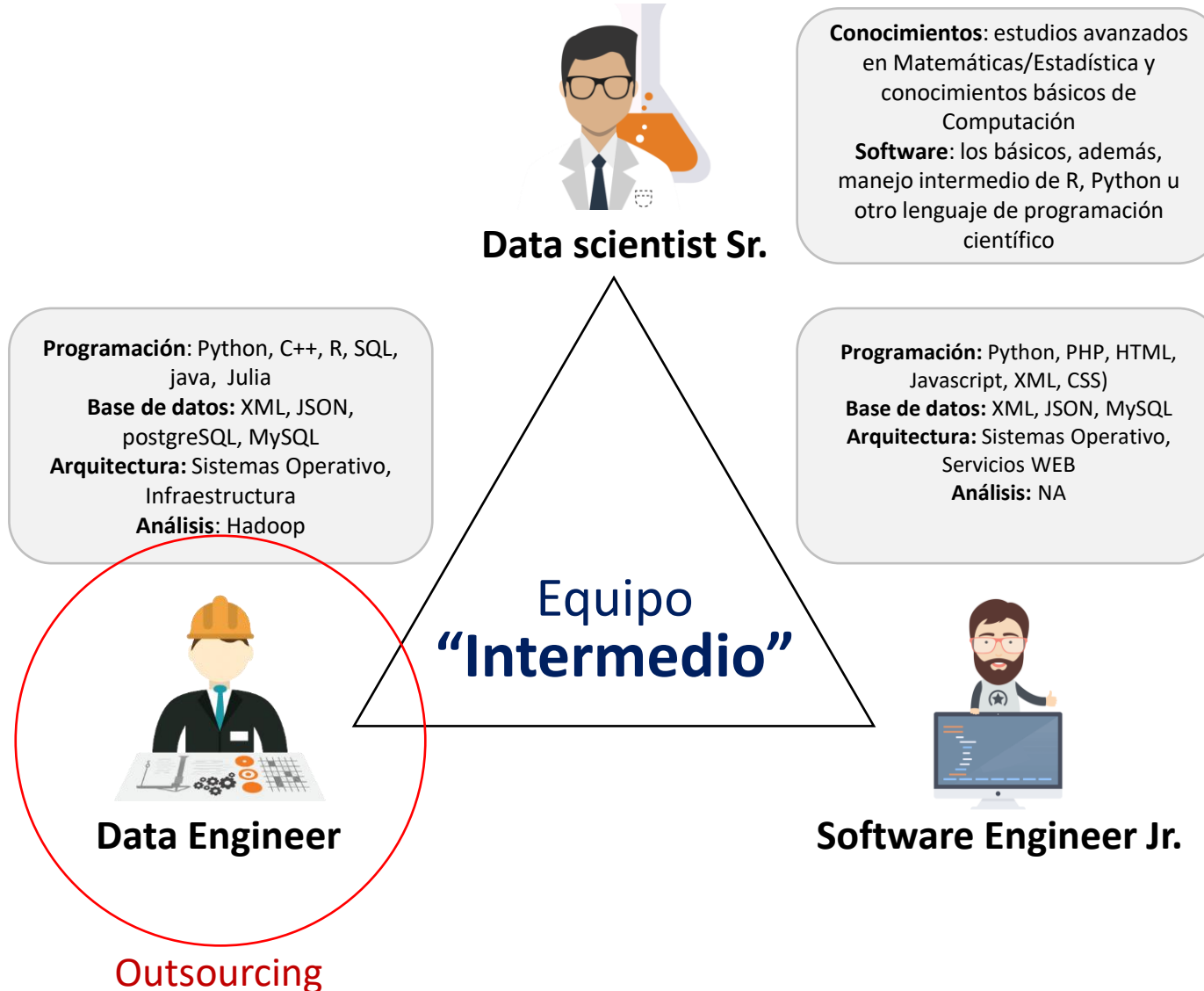
CIENTÍFICO DE DATOS



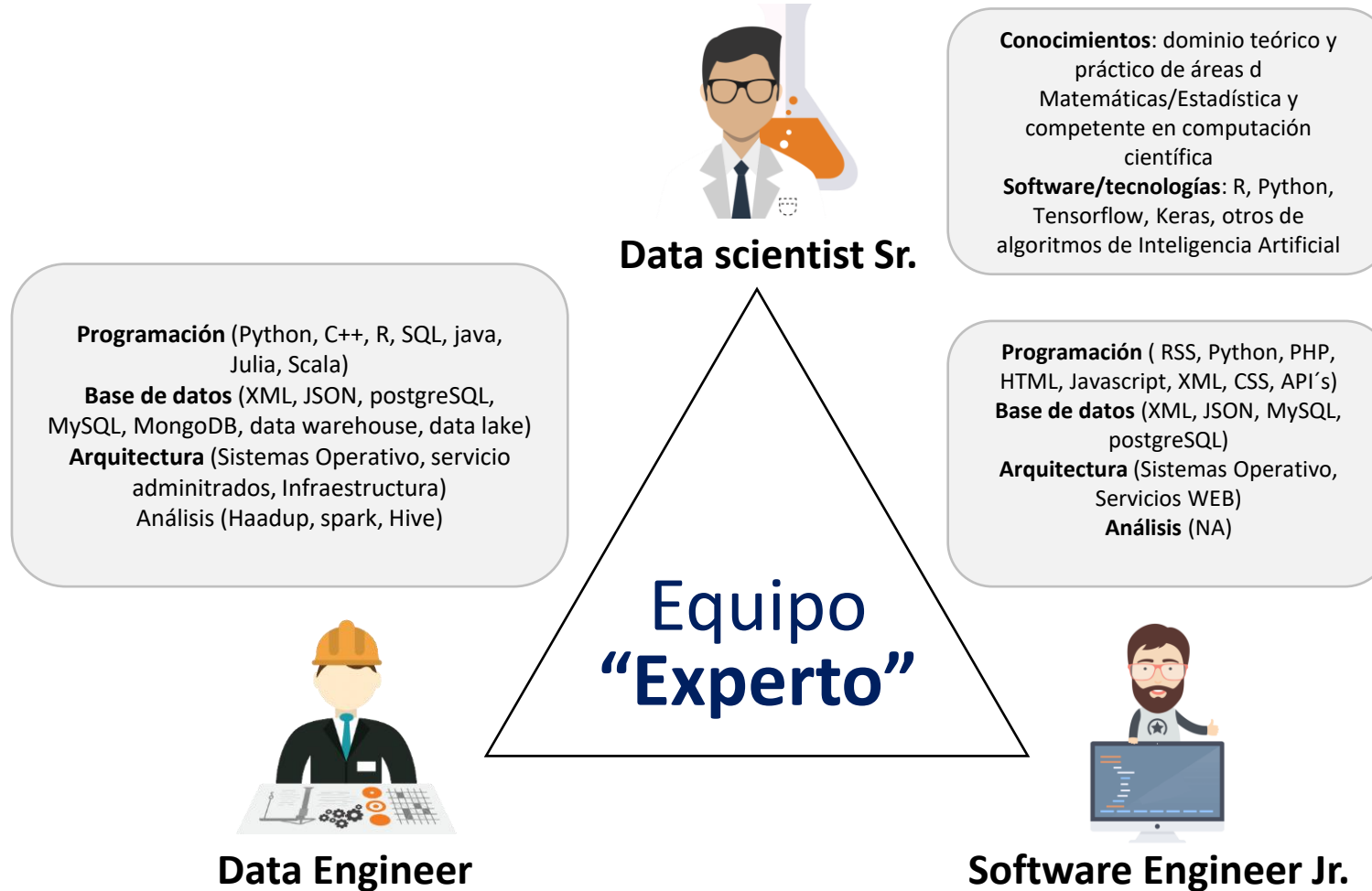
# PERFILES - *SEGÚN NIVEL DE COMPETENCIAS*



# PERFILES - *SEGÚN NIVEL DE COMPETENCIAS*



# PERFILES - *SEGÚN NIVEL DE COMPETENCIAS*





## IV. PLATAFORMAS **BIG DATA**





# PLATAFORMAS LÍDERES

## TECNOLOGÍAS INTEGRALES (NUBE + BD + ANALÍTICA)



### CLOUD ML SERVICES COMPARISON

	Amazon	Microsoft	Google	IBM
Automated and semi-automated ML services				
	Amazon ML	Microsoft Azure ML Studio	Google Prediction API	IBM Watson ML Model Builder
Classification	✓	✓	deprecated	✓
Regression	✓	✓		✓
Clustering	✓	✓		✗
Anomaly detection	✗	✓		✗
Recommendation	✗	✓		✗
Ranking	✗	✓		✗
Platforms for custom modeling				
	Amazon SageMaker	Azure ML Services	Google ML Engine	IBM Watson ML Studio
Built-in algorithms	✓	✗	✗	✓
Supported frameworks	TensorFlow, MXNet, Keras, GlueN, Pytorch, Caffe2, Chainer, Torch	TensorFlow, scikit-learn, Microsoft Cognitive Toolkit, Spark ML	TensorFlow, scikit-learn, XGBoost, Keras	TensorFlow, Spark MLlib, scikit-learn, XGBoost, PyTorch, IBM SPSS, PMML

# PLATAFORMAS LÍDERES



***Ej. “Cognitive Services”***

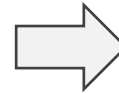
El director general del Instituto Mexicano del Seguro Social (IMSS), Tuffic Miguel, supervisó la obra del Hospital General Regional (HGR) No. 2, en el municipio de El Marqués, la cual presenta un avance de 99 por ciento.

De esta forma, la infraestructura será inaugurada en las siguientes semanas, informó el director del IMSS.

El nuevo hospital, que representa una inversión de mil 500 millones de pesos en obra y equipamiento, beneficiará a más de 800 mil derechohabientes con 53 especialidades, más de las que ofrece el HGR No. 1 del IMSS, en la capital del estado.

Entre las nuevas especialidades, de acuerdo con el organismo, se encuentran la atención en inmunología y gineco-oncología; mientras que en el área de pediatría se otorgarán endocrinología, cardiología, nefrología, neumología, neonatología, gastroenterología, hematología y oncología.

Las autoridades del Instituto han destacado la importancia de este proyecto, toda



**i FRASES CLAVE:**

IMSS, Seguro Social, director general, HGR, derechohabientes de Querétaro, Hospital General Regional, obra, gobernador de Querétaro, oncología, Campaña de Vacunación, nuevo hospital, organismo, Tuffic Miguel, secretario de Salud nacional, Instituto Mexicano, millones de acciones, millones de pesos, Semana Nacional, nuevas especialidades, distribución de vida suero oral, Manuel Ruiz López, neumología, neonatología, aplicación de vacunas, ácido fólico, José Narro, enfermeras, endocrinología, cardiología, nefrología, gastroenterología,

# PLATAFORMAS LÍDERES



*Ej. "Cognitive Services"*

El director general del Instituto Mexicano del Seguro Social (IMSS), Tuffic Miguel, supervisó la obra del Hospital General Regional (HGR) No. 2, en el municipio de El Marqués, la cual presenta un avance de 99 por ciento.

De esta forma, la infraestructura será inaugurada en las siguientes semanas, informó el director del IMSS.

El nuevo hospital, que representa una inversión de mil 500 millones de pesos en obra y equipamiento, beneficiará a más de 800 mil derechohabientes con 53 especialidades, más de las que ofrece el HGR No. 1 del IMSS, en la capital del estado.

Entre las nuevas especialidades, de acuerdo con el organismo, se encuentran la atención en inmunología y gineco-oncología; mientras que en el área de pediatría se otorgarán endocrinología, cardiología, nefrología, neumología, neonatología, gastroenterología, hematología y oncología.

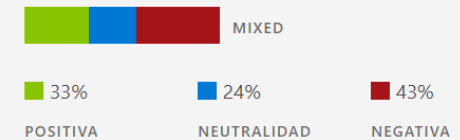
Las autoridades del Instituto han destacado la importancia de este proyecto, toda

## FRASES CLAVE:

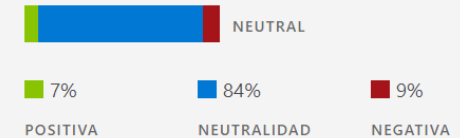
IMSS, Seguro Social, director general, HGR, derechohabientes de Querétaro, Hospital General Regional, obra, gobernador de Querétaro, oncología, Campaña de Vacunación, nuevo hospital, organismo, Tuffic Miguel, secretario de Salud nacional, Instituto Mexicano, millones de acciones, millones de pesos, Semana Nacional, nuevas especialidades, distribución de vida suero oral, Manuel Ruiz López, neumología, neonatología, aplicación de vacunas, ácido fólico, José Narro, enfermeras, endocrinología, cardiología, nefrología, gastroenterología,

## OPINIÓN:

### DOCUMENTO



### FRASE 1



## ENTIDADES CON NOMBRE:

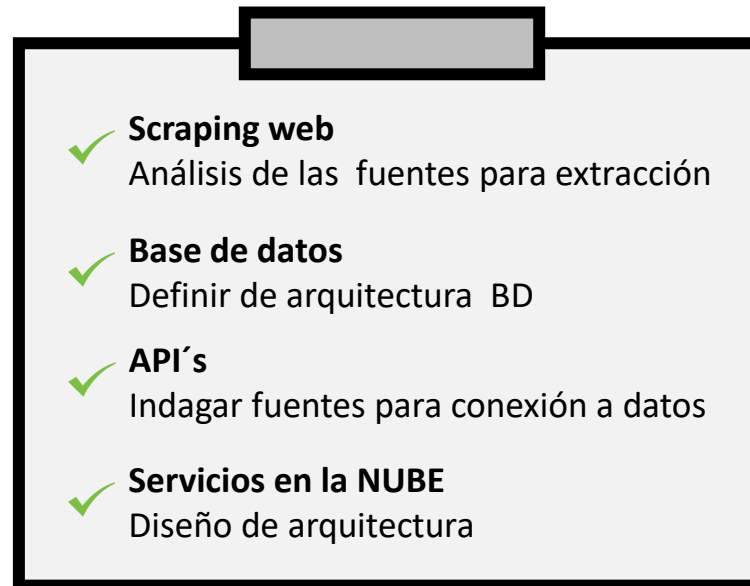
Instituto Mexicano del Seguro Social [Organization]  
IMSS [Organization]  
Tuffic Miguel [Person]  
2 [Quantity-Number]  
un [Quantity-Number]  
99 por ciento [Quantity-Percentage]  
IMSS [Organization]  
una [DateTime-Time]  
mil [Quantity-Number]  
500 millones de pesos [Quantity-Currency]  
800 mil [Quantity-Number]

## IV. RECOMENDACIONES



## IV. RECOMENDACIONES

- 1 Identificación de los **procesos críticos** y el levantamiento de los **requerimientos técnicos específicos** bajo una inspección exhaustiva de las fuentes y los métodos de transferencia del conocimiento, considerando los siguientes criterios:



# IV. RECOMENDACIONES

- 1 Identificación de los **procesos críticos** y el levantamiento de los **requerimientos técnicos específicos** bajo una inspección exhaustiva de las fuentes y los métodos de transferencia del conocimiento, considerando los siguientes criterios:

1) MONITOREO DIARIO DE MEDIOS	
Dimensión	Descripción
Periodicidad	Diario
Fuentes	Periódicos digitales Presidencia (Mañaneras) Agendas de gobierno
Temáticas	<ul style="list-style-type: none"> <li>✓ Gobierno Seguridad</li> <li>✓ Social y Laboral</li> <li>✓ Economía y Negocios</li> <li>✓ Conferencias matutinas</li> <li>✓ Columnas</li> </ul>
Tipo de resultado	Descubrimiento y seguimiento a temas de interés

2) MONITOREO SEMANAL DE NEGOCIOS	
Dimensión	Descripción
Periodicidad	Semanal
Fuentes	Periódicos digitales
Temáticas	<ul style="list-style-type: none"> <li>✓ Política</li> <li>✓ Economía</li> <li>✓ Energía</li> <li>✓ Logística</li> <li>✓ Global</li> </ul>
Tipo de resultado	Descubrimiento y seguimiento a temas de interés

3) REPORTE DEL CONTEXTO INTERNACIONAL	
Dimensión	Descripción
Periodicidad	Semanal
Fuentes	Periódicos digitales globales
Temáticas	✓ Temas de vanguardia
Tipo de resultado	Descubrimiento y seguimiento a temas de interés

**NOTA:** no se dispone de la información suficiente para el levantamiento de requerimientos técnicos específicos

# IV. RECOMENDACIONES

**1** Identificación de los **procesos críticos** y el levantamiento de los **requerimientos técnicos específicos...**

## 4) MONITOREO DE ECONOMÍA COAHUILA

Dimensión	Descripción
Periodicidad	Mensual
Fuentes	Periódicos digitales
Temáticas	✓ Economía
	✓ Negocios
Tipo de resultado	Descubrimiento y seguimiento a temas de interés

## 5) VERIFICACIÓN DE ANTECEDENTES

Dimensión	Descripción
Periodicidad	A petición
Fuentes	Websites de medios digitales, portales institucionales/oficiales, redes sociales, buscadores,
Temáticas	✓ Información general
	✓ Formación académica
	✓ Experiencia profesional
	✓ Vínculos
	✓ Imagen pública
	✓ Legal
	✓ Historia
	✓ Socios y colaboradores
	✓ Información legal
Tipo de resultado	Reporte de hallazgos

## 6) INFORMES DE CONTEXTO

Dimensión	Descripción
Periodicidad	A petición
Fuentes	Diversidad
Temáticas	✓ Partidos políticos
	✓ Resumen estadístico municipios
	✓ Actores de interés
	✓ Presencia de grupos criminales
	✓ Eventos de alto impacto
	✓ Monitoreo de sindicatos
	✓ Actores de influencia
Tipo de resultado	Descubrimiento y seguimiento a temas de interés

**NOTA:** *no se dispone de la información suficiente para el levantamiento de requerimientos técnicos específicos*

## IV. RECOMENDACIONES

2

Contar con una **Base de Datos estructurada y robusta** permite:

- Mayor **velocidad** (ejecución y desarrollo)
- Mayor **control**
- Evitar **duplicidad y redundancia**
- Implementación rápida de **estudios experimentales**
- **Proyectos de innovación** (nuevos indicadores, nuevos enfoques, nuevos fenómenos)

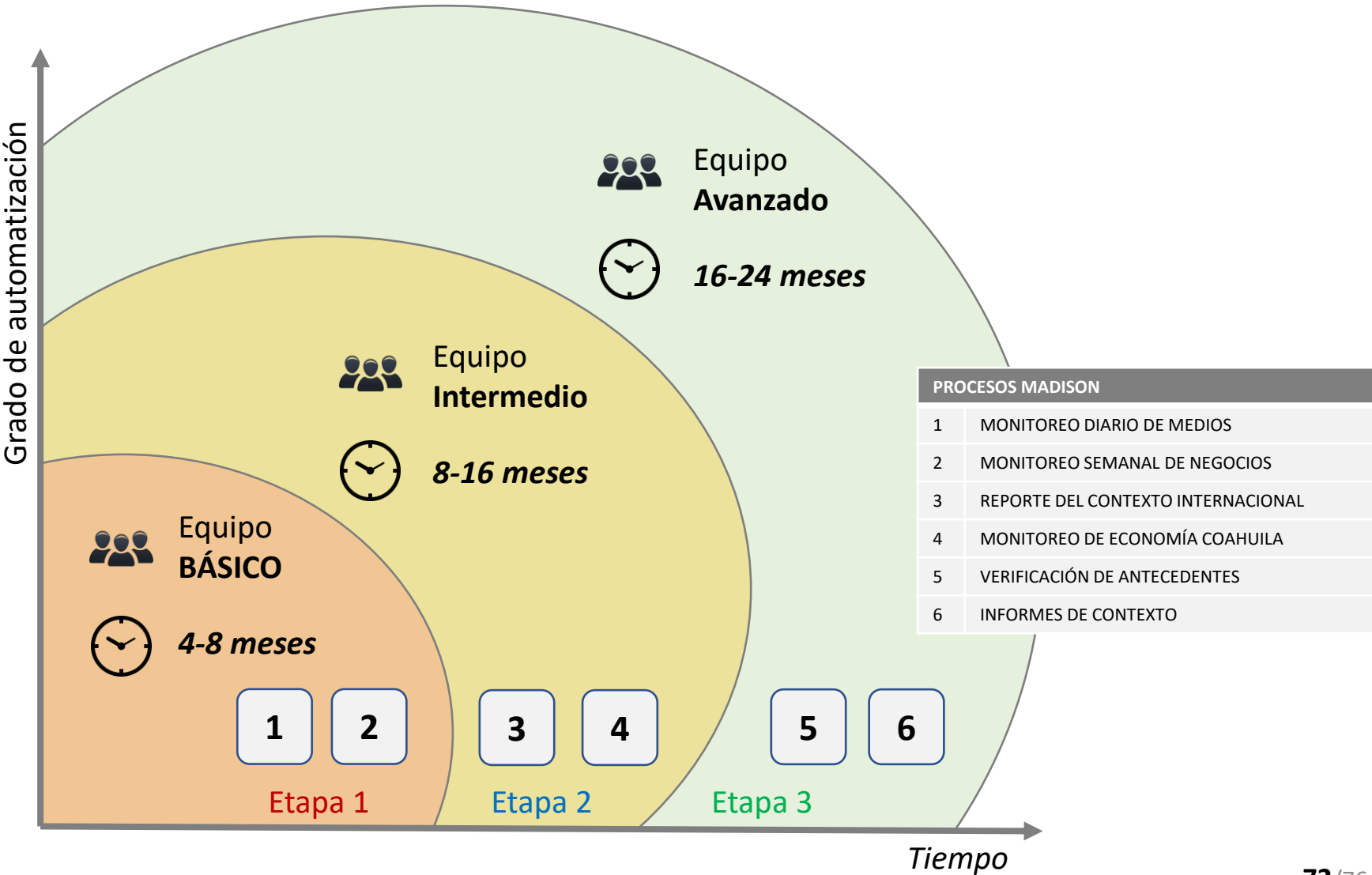




# IV. RECOMENDACIONES

## 3 Un programa de Implementación Tecnológica de corto, mediano y largo plazo:

- Grado de automatización **óptima y ad hoc** (códigos en distintos niveles de prioridad)
  - Automatización integral en el proceso de **almacenamiento** basada en tecnología **robusta y escalable**
  - **Automatización integral** de la amplia **variedad y volumen** de datos, BIG DATA.
  - Se desarrollan **algoritmos complejos** de Procesamiento de Lenguaje Natural (**PLN**) y de Inteligencia Artificial (**IA**)
- Se pueden analizar y **extraer** datos de algunas fuentes con **mayor grado de complejidad**
  - **Semi-automatización** en el desarrollo de **infraestructura de almacenamiento** de datos
  - Se define un **modelo de bases de datos** con el potencial de ser **escalado y robusto** en el largo plazo
  - Se desarrollan **algoritmos automatizados de normalización y clasificación** con técnicas de nivel intermedio y avanzadas
- Automatización en la extracción de fuentes con **características similares**.
  - **Almacenamiento local y modelo de bases de datos de uso común** con preparación para un futuro escalamiento (robusto, rápido)
  - Técnicas básicas de **minería de textos**
  - Levantamiento de **requerimientos técnicos** ("scorecards" y diccionarios) para los objetivos y procesos de negocio prioritarios



## IV. RECOMENDACIONES

**4** Para incrementar el valor de las soluciones, se recomienda el desarrollo de **“autómatas de internet”** como complemento e incremento al portafolios de soluciones actuales:

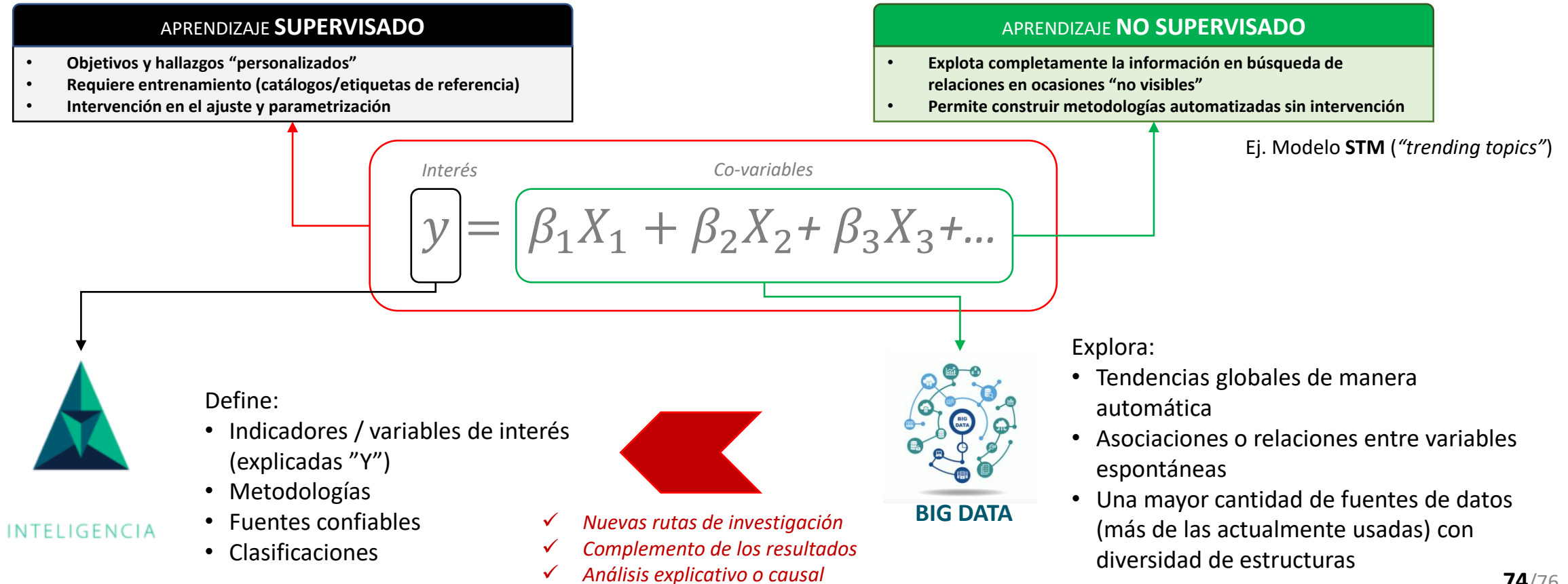
### PROCESOS MADISON

- 1 MONITOREO DIARIO DE MEDIOS
- 2 MONITOREO SEMANAL DE NEGOCIOS
- 3 REPORTE DEL CONTEXTO INTERNACIONAL
- 4 MONITOREO DE ECONOMÍA COAHUILA
- 5 VERIFICACIÓN DE ANTECEDENTES
- 6 INFORMES DE CONTEXTO



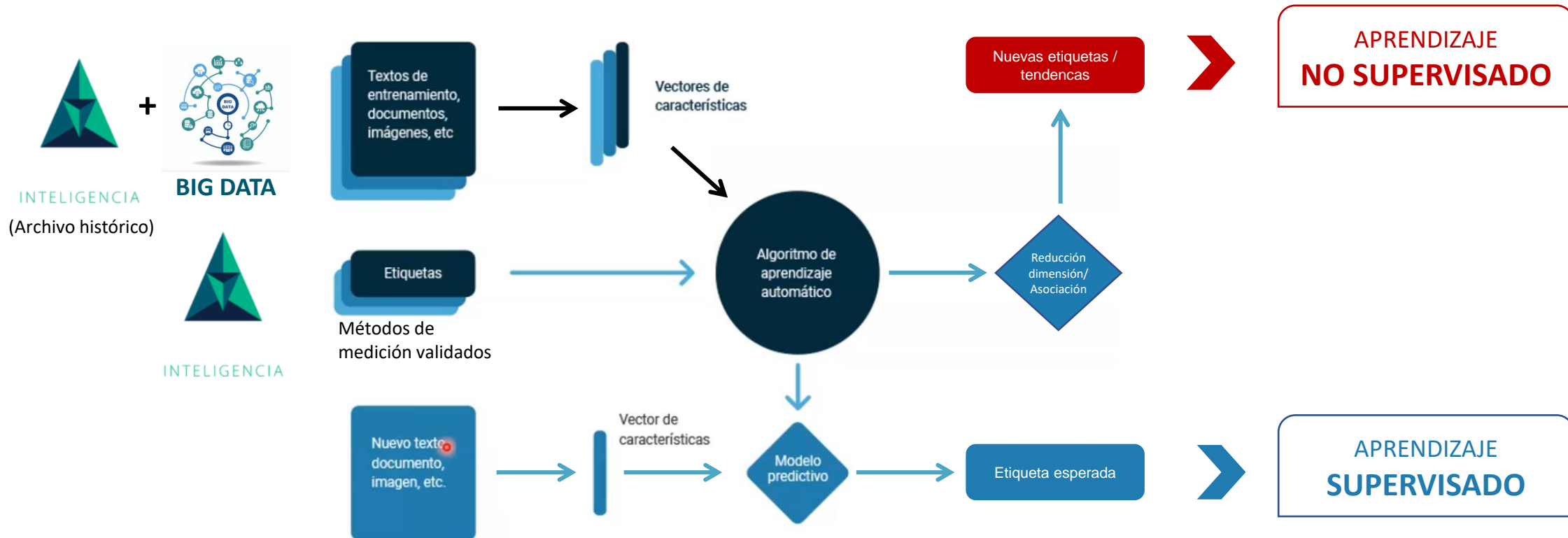
## IV. RECOMENDACIONES

- 4 Para incrementar el valor de las soluciones, se recomienda el desarrollo de **“autómatas de internet”** como complemento e incremento al portafolios de soluciones actuales:



## IV. RECOMENDACIONES

4. Para incrementar el valor de las soluciones, se recomienda el desarrollo de **“autómatas de internet”** como complemento e incremento al portafolios de soluciones actuales:

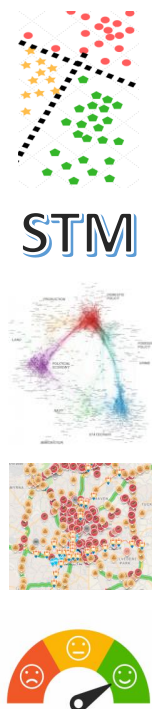


Tomado de: [https://cdn-images-1.medium.com/max/1600/1\\*0\\_fuxcGj6FL0Yqsu0rt4gQ.png](https://cdn-images-1.medium.com/max/1600/1*0_fuxcGj6FL0Yqsu0rt4gQ.png)

Dr. Saúl Domínguez Isidro. Curso: “Reconocimiento de patrones con algoritmos de Aprendizaje Supervisado”. Mayo 2020.

# IV. RECOMENDACIONES

**5** Con base en la información compartida, se recomienda la implementación total o parcial de lo siguientes métodos analíticos a los procesos de negocio de MADISON:



	1) MONITOREO DIARIO DE MEDIOS	2) MONITOREO SEMANAL DE NEGOCIOS	3) REPORTE DEL CONTEXTO INTERNACIONAL	4) MONITOREO DE ECONOMÍA COAHUILA	5) VERIFICACIÓN DE ANTECEDENTES	6) INFORMES DE CONTEXTO
CLASIFICACIÓN	✓	✓	✓	✓		✓
MOIDELO STM	✓	✓	✓	✓	✓	✓
RECONOCIMIENTO DE ENTIDADES	✓	✓	✓	✓	✓	✓
GEORREFEREN- CIACIÓN			✓	✓		✓
ANÁLISIS DE SENTIMIENTOS					✓	✓



# MIETRIKA

Business Analytics