Introducción a Big data ¡Bienvenidos todos!

Dra. Elvia Ruiz Beltrán
Tecnológico Nacional de México
Instituto Tecnológico de Aguascalientes

1. ¿Qué es Big Data?

- ***Big data*(macrodatos, datos masivos, inteligencia de datos** o **datos a gran escala)** es un término que hace referencia a una cantidad de datos tal que **supera** la capacidad del **software convencional** para ser capturados, administrados y procesados en un tiempo razonable.
- **Big Data** no es fácil de definir ya que en mi opinión, es un término «inventado para el marketing».
- El término de moda en el mundo de la informática y de la administración de negocios (Management Business Administration).

2. El Significado de Big Data

- “**Volumen masivo de d**atos, tanto estructurados como no-estructurados, los cuales son demasiado grandes” y difíciles de procesar con las bases de datos y el software tradicionales (ONU, 2012).

3. Breve Historia de Big Data

La popularización del término Big Data viene, sin duda, ligada al documento del concepto publicado por **McKinsey Global Institute en Junio de 2011**, en el cual se define como "**conjuntos de datos cuyo tamaño va más allá de la capacidad de captura, almacenamiento, gestión y análisis de las herramientas de base de datos tradicionales**".

Evolución de <<Big Data>>

1943 -	1945 -	1954 -	1964 -
The UK developed the first data-processing machine to decipher Nazi codes.	ENIAC, the first electronic general purpose computer, was completed.	The first fully transistorised computer used all transistors and diodes and no vacuum tubes.	The IBM System/360 family of mainframe computer systems was launched.

1971 -	1973 -	1977 -	1981 -
Intel's 4004 became the first general purpose programmable processor.	Xerox unveiled the first desktop system to include a graphical user interface and internal memory storage.	ARCnet introduced the first LAN at Chase Manhattan Bank, connecting 255 computers.	The PC era began.

Puedes visitar esta fuente también:
https://www.verdict.co.uk/big-data-timeline/

Evolución de «Big Data

1989 -

Implementation of the Python programming language began.

1998 -

Carlo Strozzi developed NoSQL, an open-source relational database.

1999 -

VMware began selling VMware Workstation, allowing users to set up virtual machines.

2002 -

Amazon Web Services (AWS) launched as a free service.

2006 -

AWS started offering web-based computing infrastructure services, now known as cloud computing.

2007 -

Apple launched the first iPhone, creating the mobile internet as we know it today.

2010 -

The first solutions for 100 Gigabit Ethernet were introduced.

2011 -

Facebook launched the Open Compute Project to share specifications for energy efficient data centers.

Evolución de «Big Data

2013 -

Docker introduced open-source OS container software.

2015 -

Google and Microsoft lead massive build outs of data centers.

2017 -

Huawei and Tencent joined Alibaba in major data centre build-outs in China.

2018 -

Leading data center operators started the migration to 400G data speeds.

2018 -

Silicon photonics technology started to positively impact data center networking architectures..

2020 -

Edge computing will revise the role of the cloud in key sectors of the economy.

2021 -

Data center speeds expected to exceed 1,000G.

2025 -

Data centers will be increasingly on-device.

Evolución de <<Big Data Analytics>>

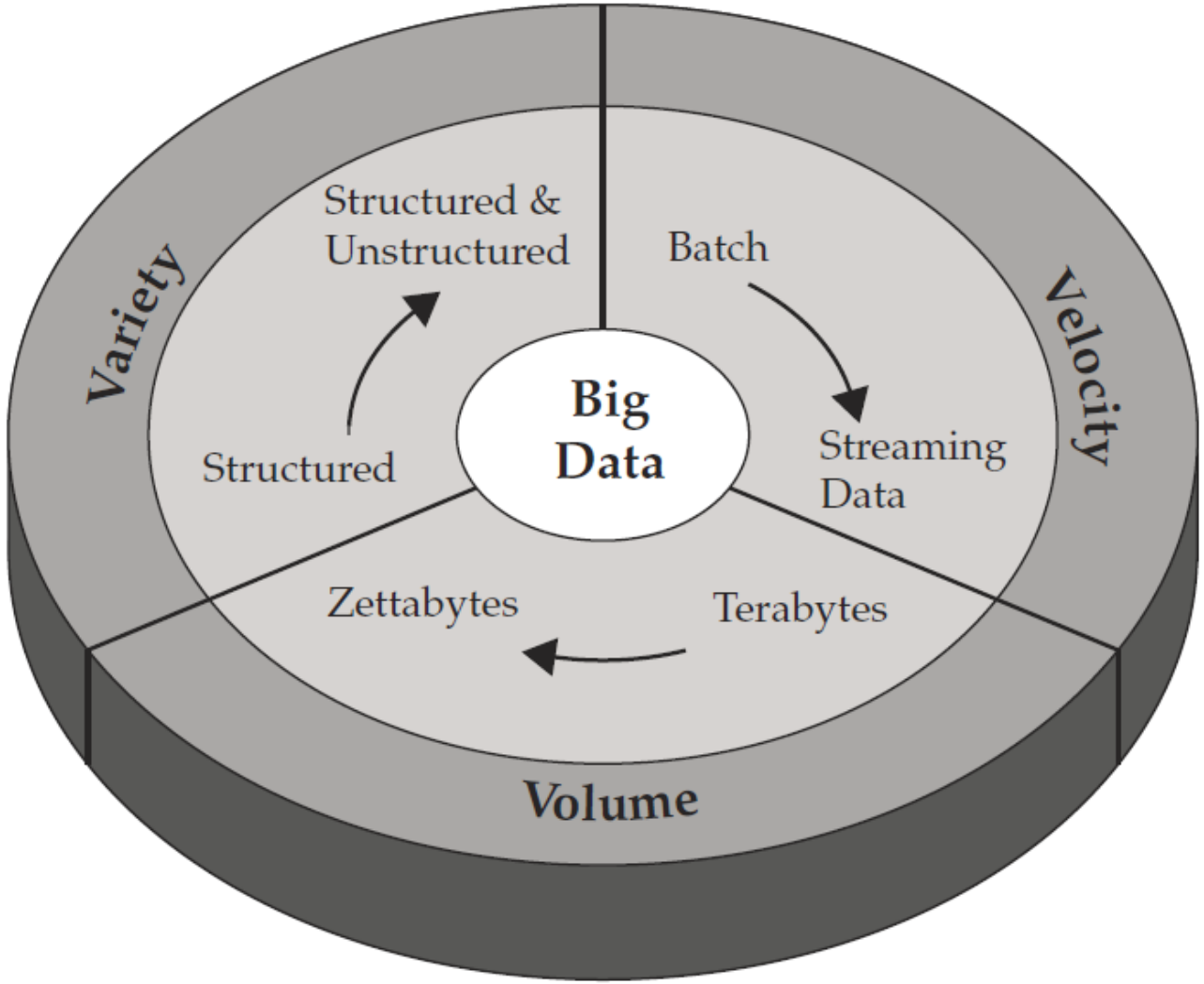
1970-1990	1990-2000	2000-2005	2005-2011	2011-presente
Análisis de datos (Exploratory Data Analysis), SQL, Object-oriented programming (OOP) languages	OLAP(Online Analytical Processing) The World Wide Web, using HyperText Transfer Protocol (HTTP) and the HyperText Markup Language (HTML), a TCP/IP application layer protocol for distributing, searching and retrieving documents	Minería de datos, Business Intelligence	Modelos predictivos, Analytics	"Big Data" , Big Data Analytics

Puedes visitar esta fuente también:
https://www.verdict.co.uk/big-data-timeline/

3 V's El significado de Big Data- 3V´s Características(atributos) de «Big Data»

Big Volume
Big Velocity
Big Variety

Big data son activos de información caracterizados por su volumen, velocidad y variedad, que demandan soluciones innovadoras y eficientes de procesado para la mejora del conocimiento y toma de decisiones de las organizaciones.



4V´S Características(atributos) de «Big Data»

Volumen

Datos a escala

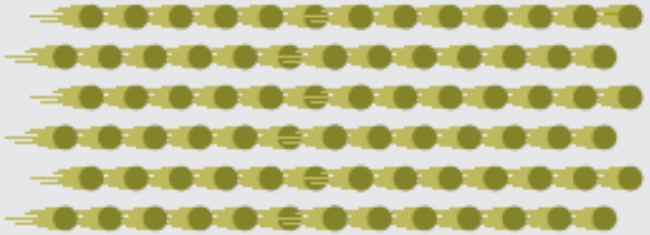
De terabytes a petabytes de datos

Variedad

Datos en muchas formas

Estructurados, no estructurados, texto, multimedia

Velocidad



Datos en movimiento

Análisis de datos en streaming para una toma de decisiones en cuestión de segundos

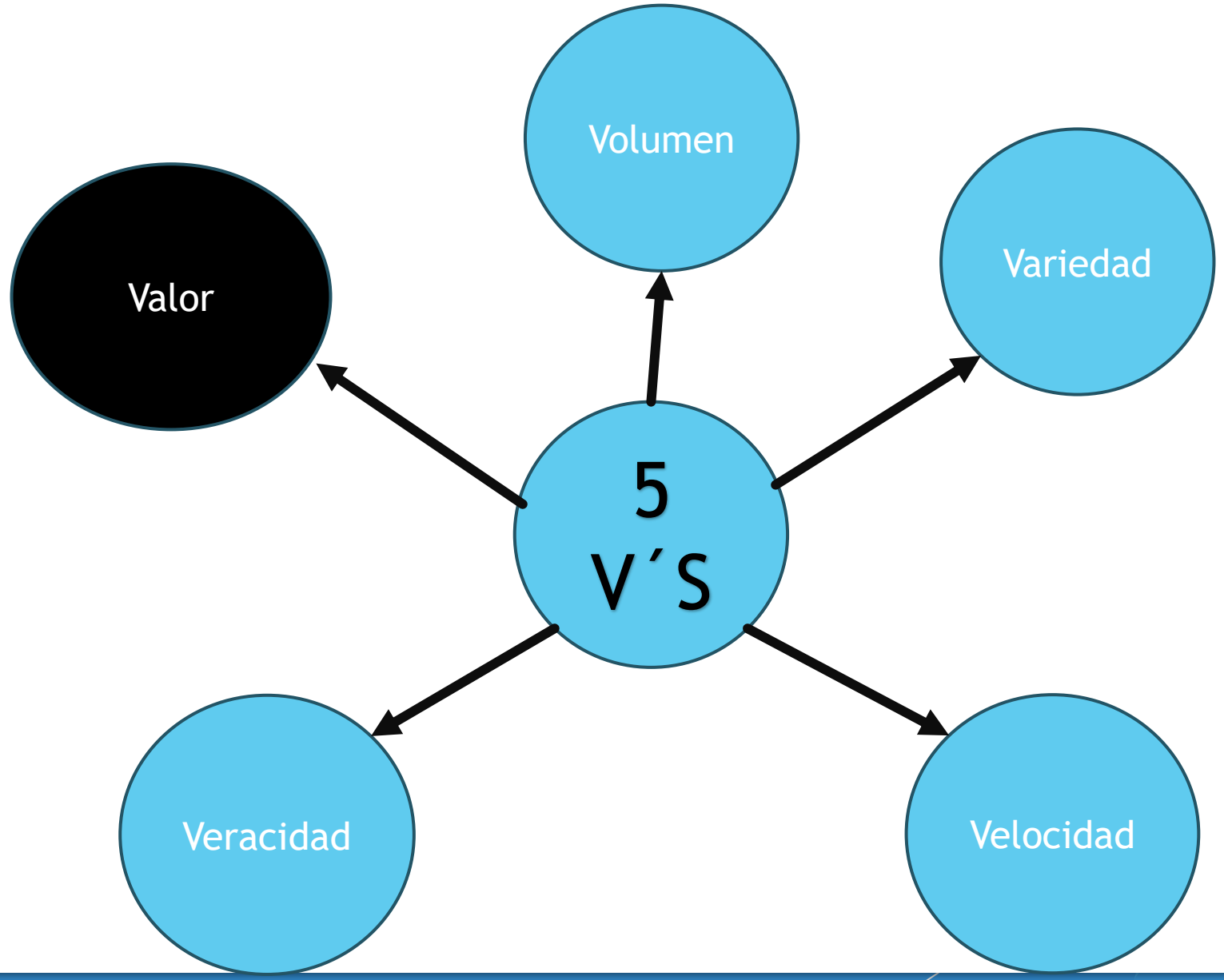
Veracidad

Incertidumbre de datos

Gestionar la fiabilidad y previsibilidad de tipos de datos intrínsecamente imprecisos

Características de «Big Data»

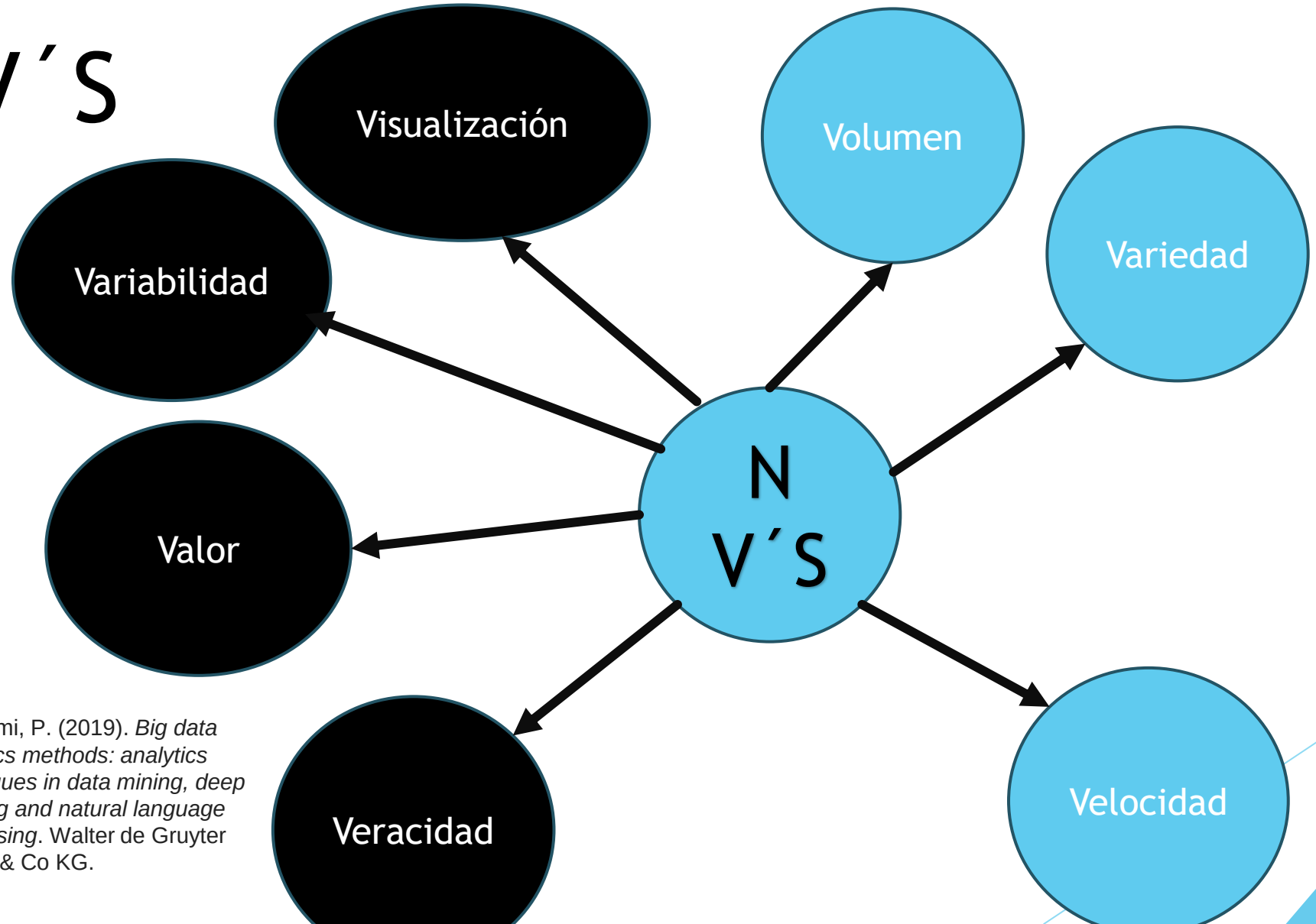
5V´S



Cuatro V’s más se agregan a la lista:

Veracidad, Valor, Variabilidad y Visualización.

Características de «Big Data»



Ghavami, P. (2019). *Big data analytics methods: analytics techniques in data mining, deep learning and natural language processing*. Walter de Gruyter GmbH & Co KG.

Volumen

Redes Sociales

Instrumentos científicos

Dispositivos Móviles

Bancos

Comercio

Tecnología de sensores y redes

Características de «Big Data»

- ¿Qué tan grande es el Big Data?
  - Lo que es grande hoy en día tal vez no lo sea mañana.

Ghavami, P. (2019). *Big data analytics methods: analytics techniques in data mining, deep learning and natural language processing*. Walter de Gruyter GmbH & Co KG.

Unida de Memoria	Tamaño	Tamaño binario
kilobyte(Kb/KB)	10^3	2^{10}
megabye(MB)	10^6	2^{20}
gigabyte(GB)	10^9	2^{30}
terabyte(TB)	10^{12}	2^{40}
petabyte(PB)	10^{15}	2^{50}
exabyte(EB)	10^{18}	2^{60}
zettabyte(ZB)	10^{21}	2^{70}
yottabyte(YB)	10^{24}	2^{80}

Table I.1: Storage units of measure.

Data Volume	Size
Bytes – 8 Bits	1 byte: a single character
Kilobyte – 1000 Bytes	A very short story
Megabyte – 1000 KiloBytes	A small novel
Gigabyte – 1000 MegaBytes	A movie at TV quality
Terabyte – 1000 GigaBytes	All X-ray films in a large hospital
Petabyte – 1000 TeraBytes	Half of all US academic research libraries
Exabyte – 1000 PetaBytes	Data generated from SKA telescope in a day
Zettabyte – 1000 ExaBytes	All worldwide data generated in 1^{st} half of 2012
Yottabyte – 1000 ZetaBytes	1 YB = 1000^8 bytes – 10^{24} byes

Variedad

Tipos de datos del Big Data

Estructurados	No-estructurados	Semi-estructurados
Datos que tienen bien definidos su longitud y su formato, como las fechas, los números o las cadenas de caracteres.	Datos en el formato tal y como fueron recolectados, carecen de un formato específico.	Datos que no se limitan a campos determinados, pero que **contiene marcadores para separar los diferentes elementos**.
Se almacenan en tablas. Un ejemplo son las bases de datos relacionales y los almacenes de datos.	No se pueden almacenar dentro de una tabla ya que no se puede desgranar su información a tipos básicos de datos. Algunos ejemplos son los PDF, documentos multimedia, correos electrónicos o documentos de texto.	Es una información poco regular como para ser gestionada de una forma estándar. Estos datos poseen sus propios metadatos semiestructurados que describen los objetos y las relaciones entre ellos, y pueden acabar siendo aceptados por convención. Como ejemplos tenemos los archivos tipo **hojas de cálculo**, **CSV**, **HTML**, **XML** o **JSON**.

Velocidad

La velocidad “designa la **rapidez con que se generan** los datos y con la que **deben procesarse** para satisfacer la demanda”.

La **tecnología Big Data** ha de ser capaz de **almacenar** y **trabajar en tiempo real** con las **fuentes generadores** de información como sensores, cámaras de videos, redes sociales, blogs, páginas webs, que son fuentes que general millones y millones de datos al segundo.

Por otro lado la capacidad de análisis de dichos datos han de ser rápidos, reduciendo los largos tiempos de procesamiento que presentaban las herramientas tradicionales de análisis.

Veracidad

- Veracidad: por último el Big Data ha de ser capaz de **tratar y analizar inteligentemente** este inmenso volumen de datos con la finalidad de **obtener información verídica y útil** que nos permite la toma de decisiones en las organizaciones.

Valor

- **Valor:** toman un valor dependiendo de la información que se tenga y que se desea hacer con la misma.

¿Quién genera los datos en «Big Data»?

- Catalogamos la procedencia de los datos según las siguientes categorías:
  - **Generados por las propias personas.**
  - **Obtenidas a partir de transacciones.**
  - ***Marketing* electrónico y web.**
  - **Obtenidos a partir de las interacciones máquina a máquina.**
  - **Datos biométricos recolectados.**

Empresas bien metidas con Big Data

- Facebook-Mega
- Twittter
- LinkedIn
- NetFlix
- eBay
- Amazon
- Google
- Youtube
- DHL
- Fórmula 1
- VW
- Telemetría

Algunas Experiencias Internacionales a Nivel Gubernamental

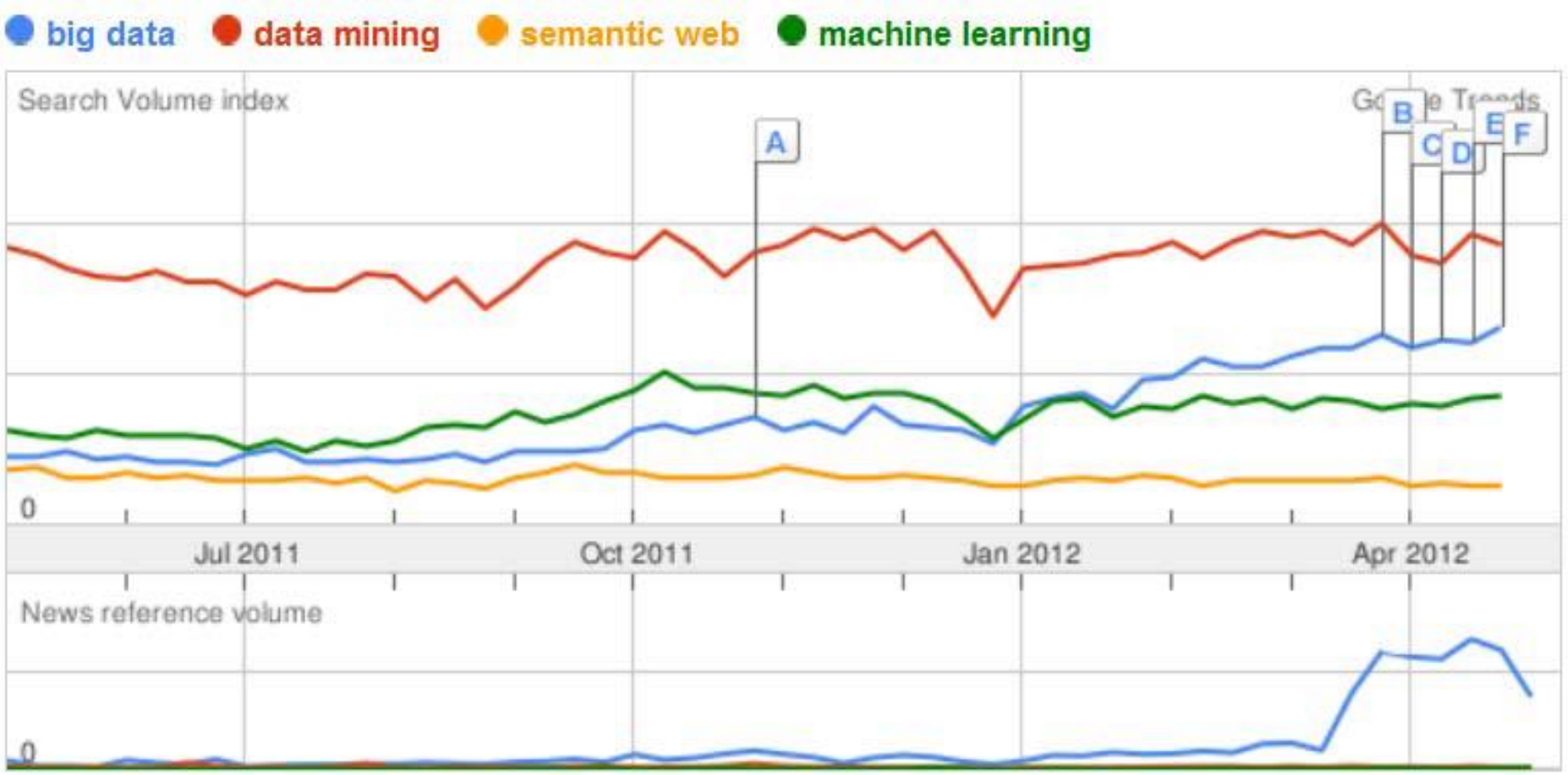
Corea del Sur: "Plan Maestro de Big Data para la Implementación de una Nación Inteligente" (2013), del gobierno coreano.

Estados Unidos: "Iniciativa de I+D en Big Data" (2012), propuesta de la administración Obama, dirigido por la Oficina para la Ciencia y la Tecnología de la Casa Blanca.

Japón: Dentro de la primera estrategia de crecimiento del Japón del gobiernode Shinzo Abe ("Desatar el poder del sector privado hasta su máxima extensión"), se encuentra un plan básico para aprovechar Big Data" (Mayo 2012).

Comisión Estadística de Naciones Unidas: Seminario de Asuntos Emergentes en la 44°Sesión de la Comisión: Big Data para la Política, el Desarrollo y las Estadísticas Oficiales

Big-Data popularity on the Web



A [Spectra Logic Delivers ExaScale Storage for 'Big Data'; Announces Series of Products and Advancements and Unveils World's Highest Capacity Storage System](#)
MarketWatch - Nov 1 2011

B [Webcast: Obama Goes Big on Big Data](#)
Wired News - Mar 27 2012

C [Cisco Joins Forces with EMC to Advance IT Skills in Cloud, Big Data and Data Center Technologies](#)
Justmeans - Apr 3 2012

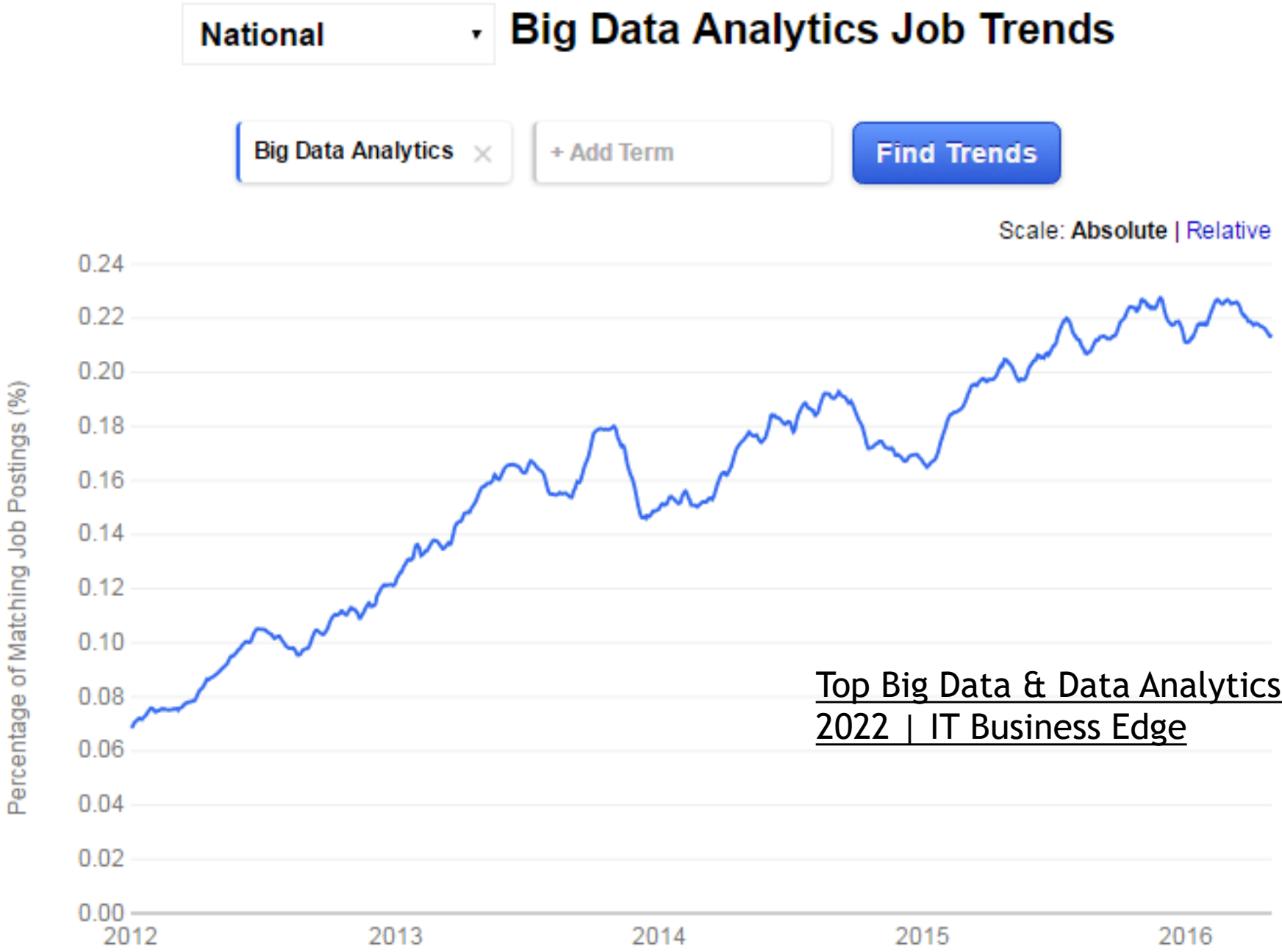
D [Ferranti Unveils its MECOMS™ "Big Data" Strategy for Utility Meter Data Management and Real Time Billing](#)
Victoria Times Colonist - Apr 10 2012

E [Deconstructing Big Data - BuildZoom Launches an Article Series that Reveals the Hype and Substance Behind Big Data](#)
Houston Chronicle - Apr 17 2012

F [Harvard Releases Big Data for Books](#)
New York Times - Apr 21 2012

Big Data Analytics Applications

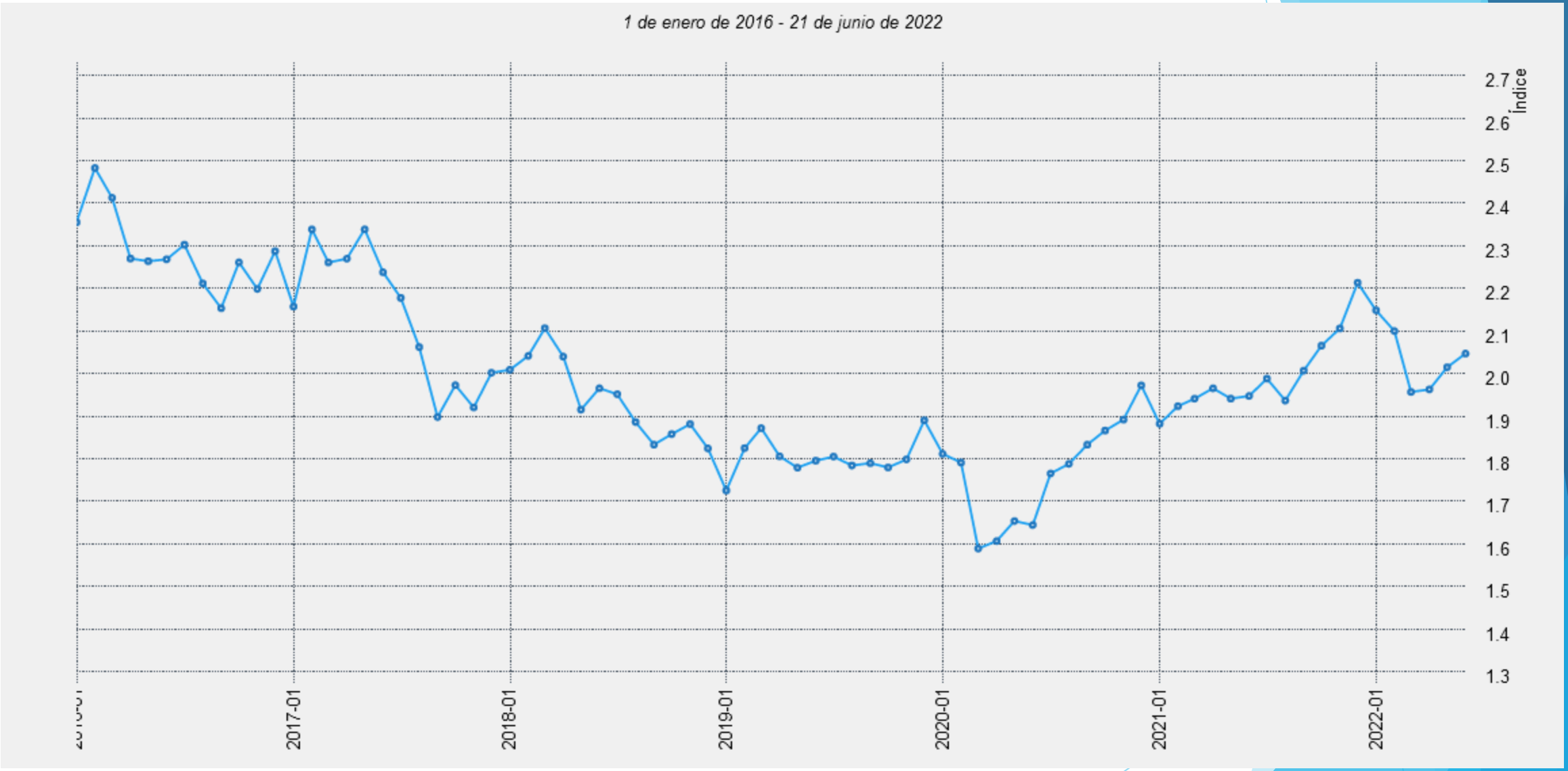
- Identification of unwanted spam messages in e-mail.
- Segmentation of customer behavior for targeted advertising.
- Forecasts of weather behavior and long-term climate changes.
- Reduction of fraudulent credit card transactions.
- Actuarial estimates of financial damage of storms and natural disasters.
- Prediction of popular election outcomes.
- Development of algorithms for auto-piloting drones and self-driving cars.
- Optimization of energy use in homes and office buildings.
- Projection of areas where criminal activity is most likely.
- Discovery of genetic sequences linked to diseases.
- Social media relationships.



Top Big Data & Data Analytics Jobs in 2022 | IT Business Edge

Estado de ánimo de los tuiteros en México

1 de enero de 2016 - 21 de junio de 2022



https://www.inegi.org.mx/app/animotuitero/

Proyectos regionales

CASO DE ÉXITO

En colaboración con el Instituto Nacional de Estadística y Geografía (INEGI) se construye el mapa de estado de ánimo los tuiteros

PROCEDIMIENTO

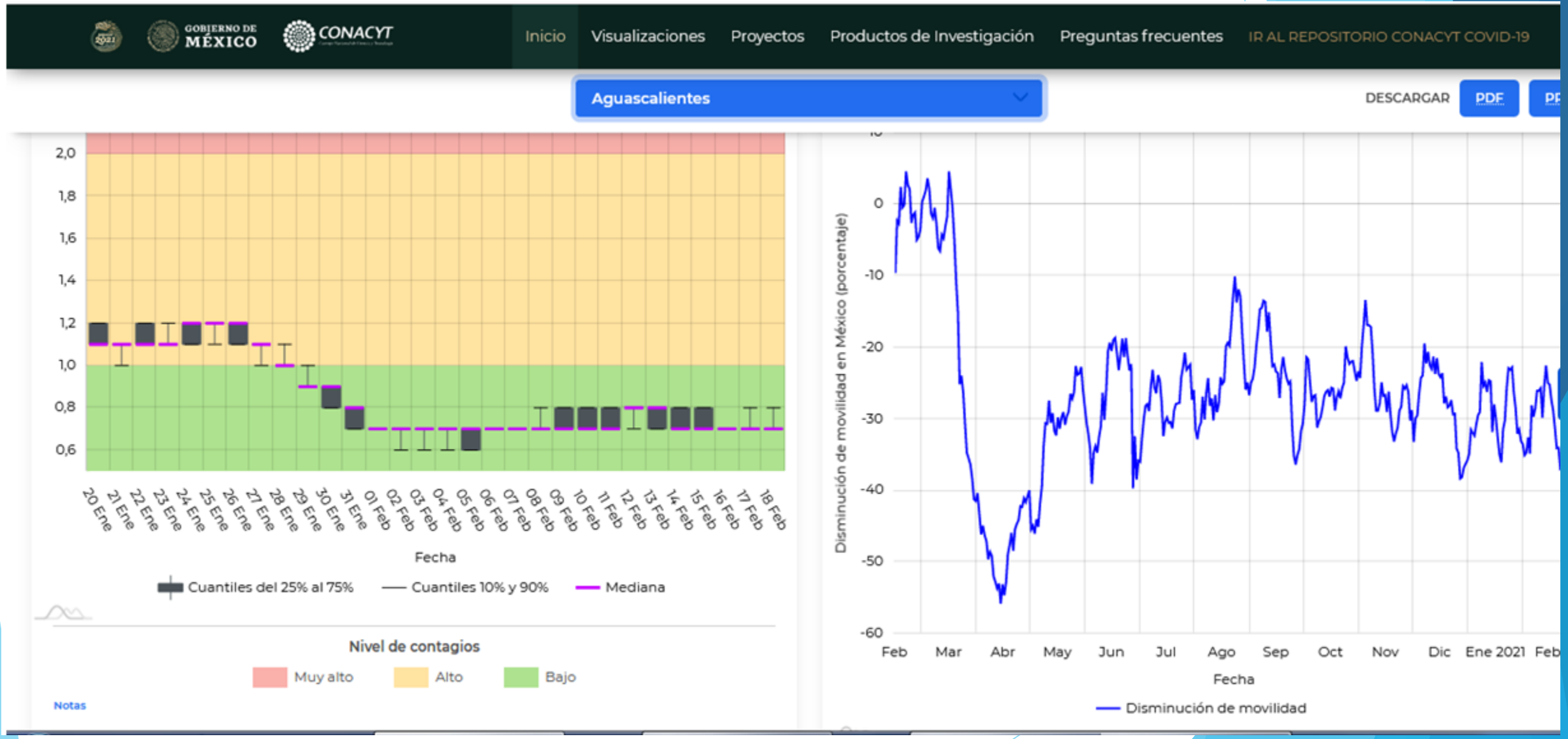
1. 60 Millones de tuits
2. Depuración
3. Normalización
4. Análisis de sentimientos
5. Geolocalización

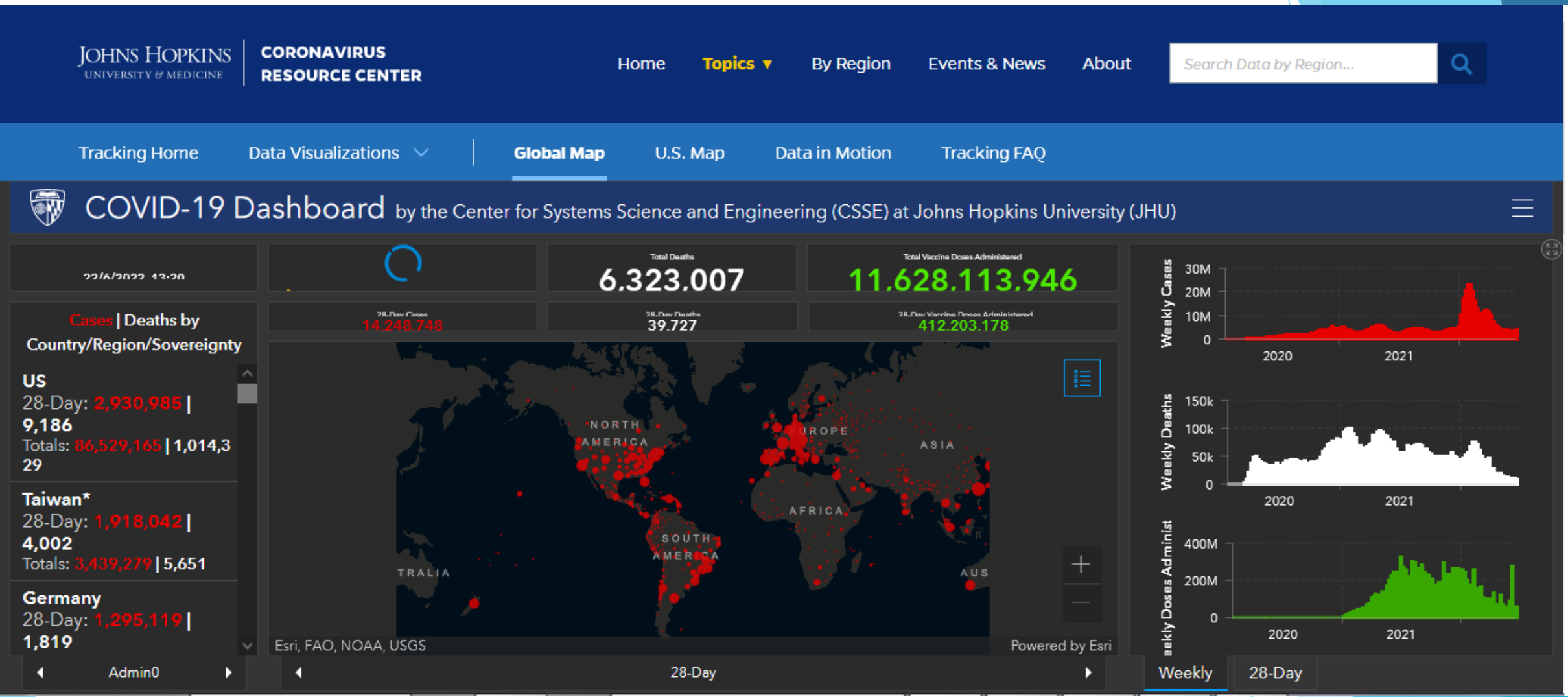
Ecosistema Nacional Informático COVID-19 (ENI/COVID-19)

Entrar a: https://coronavirus.conacyt.mx/

Repositorio CONACYT COVID-19

https://covid-19.conacyt.mx/jspui/





https://coronavirus.jhu.edu/map.html

https://coronavirus.jhu.edu/map.html

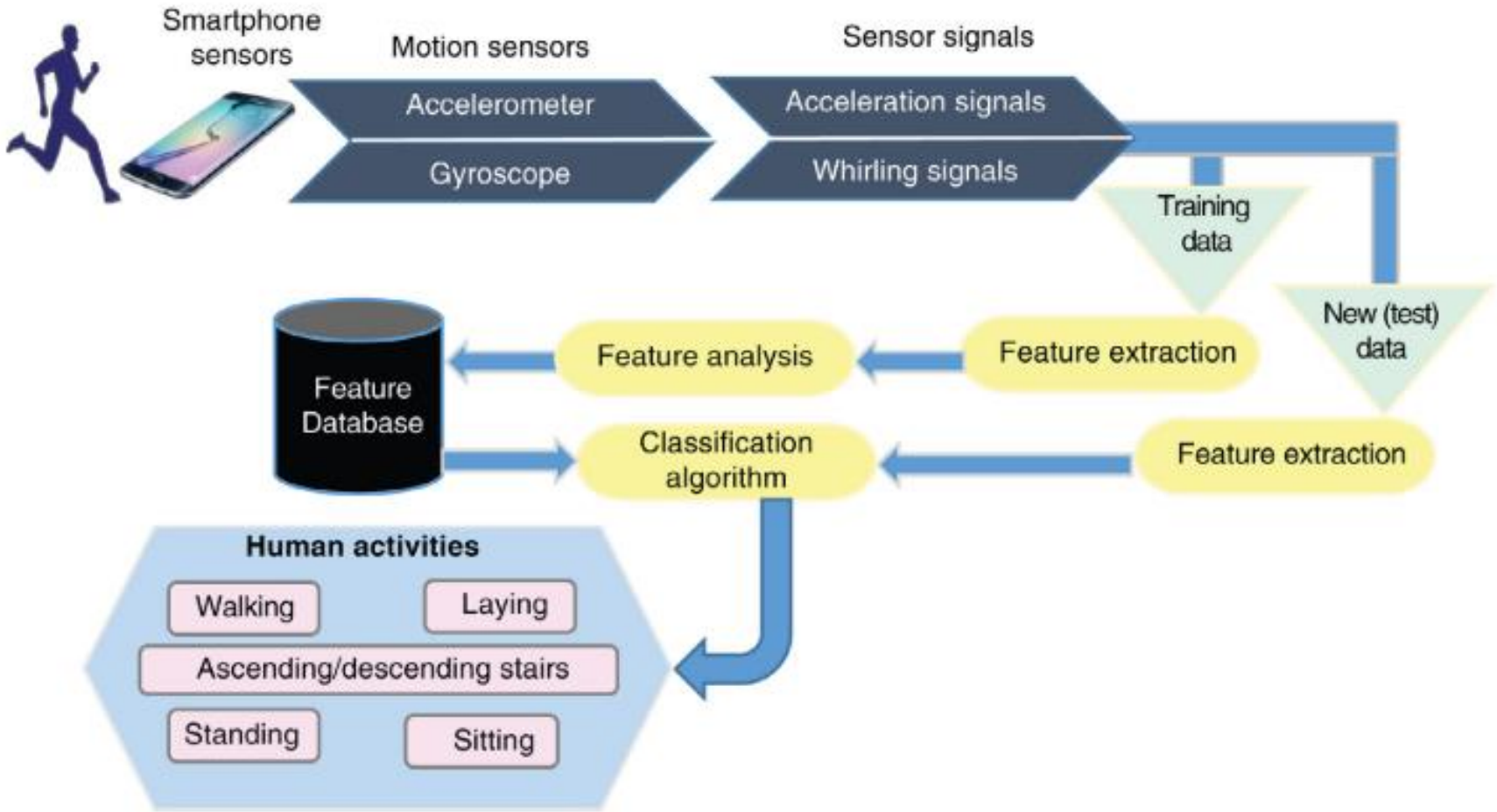


FIGURE 4.10 **A general framework for smartphone-based human activity recognition.** *Source: Adapted from Subasi et al. (2020b).*

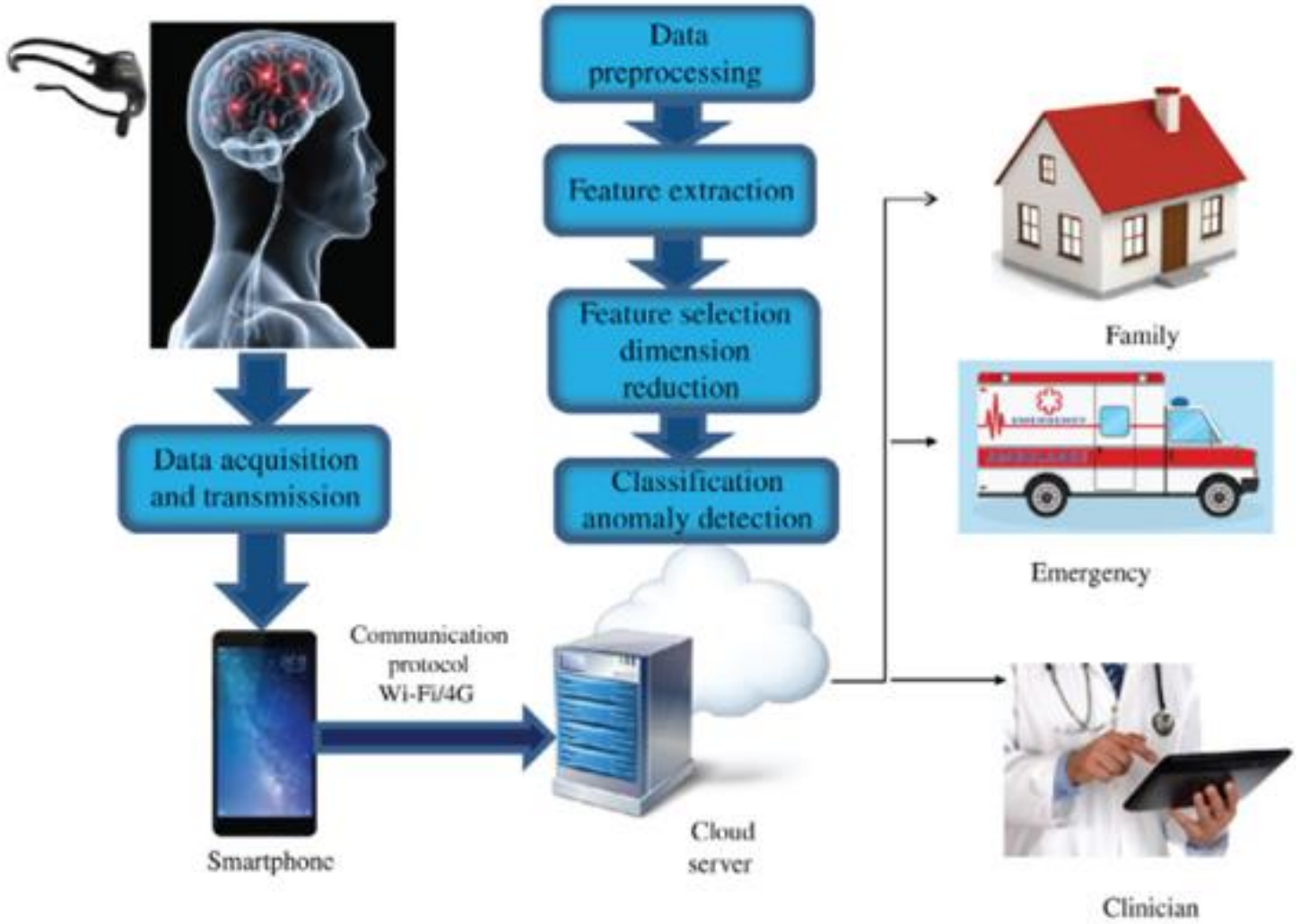


FIGURE 4.3 A framework for cloud-based mobile epileptic patient monitoring. *Source: Adapted from (Subasi et al. 2020a).*

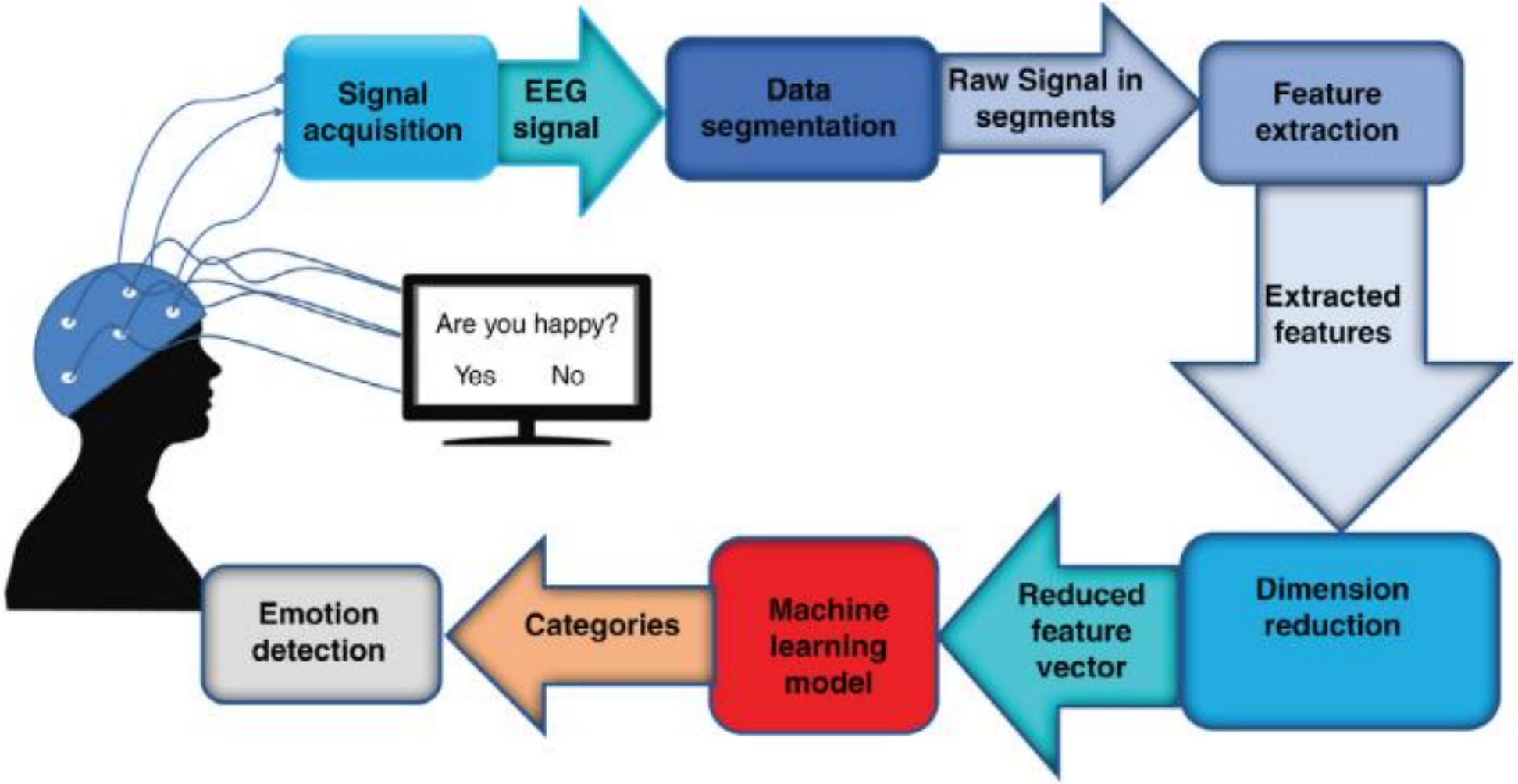


FIGURE 4.4 A general framework for emotion recognition.

CASOS DE ESTUDIO

Industria hotelera
¿Cómo puede saber una plataforma de reservaciones de un hotel si la publicidad en su sitio web funciona?

Netflix
¿Cómo utilizó Netflix una competencia para mejorar su algoritmos de búsqueda y recomendación?

Arte
¿Cómo se utilizan las redes neuronales para entrenar computadora para replicar los estilos de diferentes artistas?

Carter Racing
¿Cómo decide un equipo de carreras de autos si participar o no en una carrera?

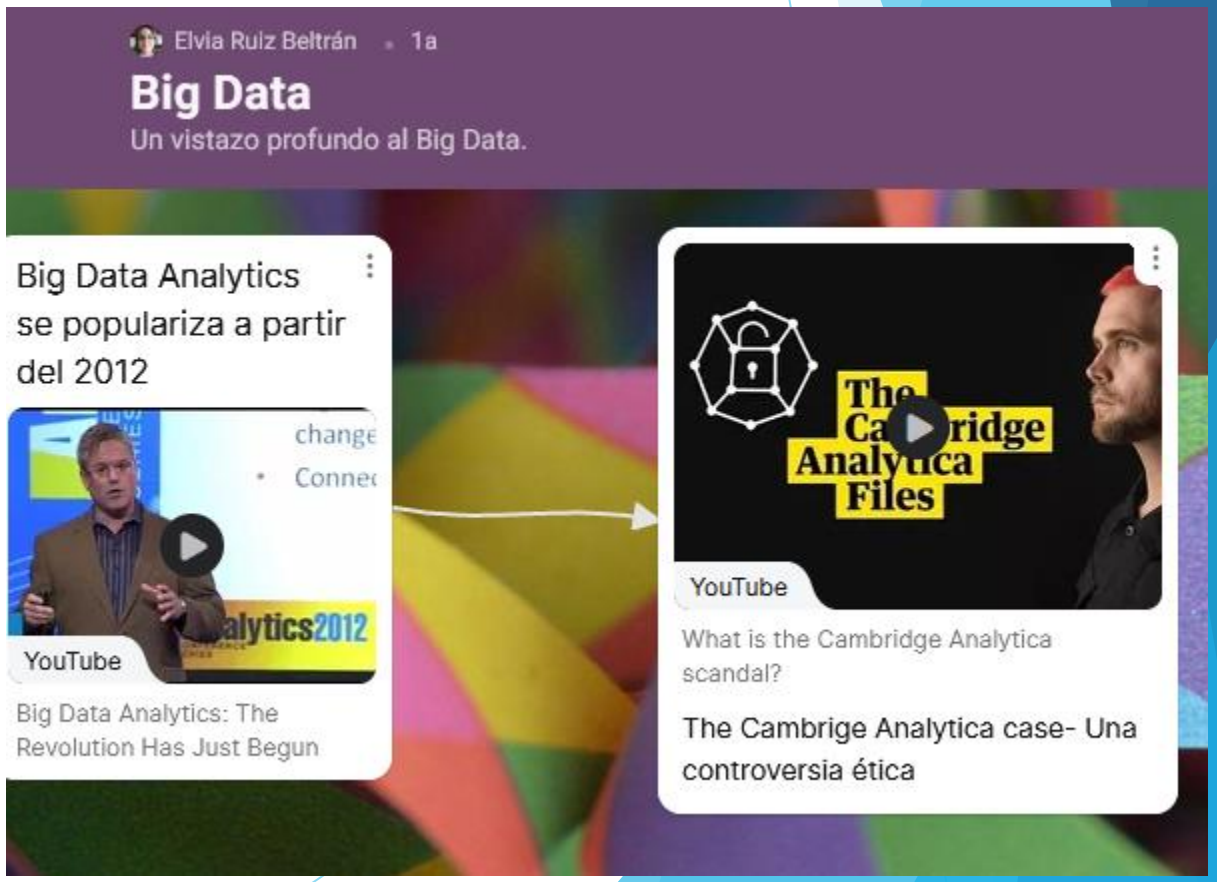
SmartService (Compañía de Instrumentos científicos)
¿Cómo utiliza SmartService un árbol de decisión para calcular su oferta para un nuevo contrato?

Talk Talk
¿Qué debería haber hecho Talk Talk de manenera diferentes después de haber sido hackeado?
¿Cómo debería haber protegido los datos de sus clientes?

PRIMERA ACTIVIDAD- Fecha de entrega se indica en Moodle

- Actividades de aprendizaje (complementarias)
  - **Ver videos que se indican en el documento de texto "VIDEOS EN PADLET"**(EN EL PADLET DE LA DRA. ELVIA RUIZ).
  - Participar en el foro creado en Moodle.

https://padlet.com/doctoraruizbeltran/hegmmwsqwvkvmjmk



SEGUNDA ACTIVIDAD-Fecha de entrega se indica en Moodle

- Este actividad es **INDIVIDUAL**, ya que se requiere que cada estudiante instale Anaconda (Pyhton e IDE) para realizar todas las prácticas que se propongan posteriormente.
- Se podría considerar Anaconda Distribution la cual nos brinda una gran cantidad de funcionalidades que permitirán que desarrollemos aplicaciones de una manera más eficiente, rápida y sencilla en el mundo de los proyectos de Ciencia de datos.