7. BIG DATA

Autor: Lic. Sebastián Schanz

7.1 INTRODUCCIÓN

Cuando hablamos de Big Data debemos tener en claro dos conceptos básicos que suelen usarse como sinónimos pero que en realidad no lo son. Estos conceptos son Datos e Información.

Datos es un término que se refiere a eventos, hechos, transacciones, etc., que han sido registrados. Es la entrada sin procesar desde la cual se produce la información.

Por otra parte, la información se refiere a datos que han sido procesados y comunicados de tal manera que pueden ser entendidos e interpretados por algún receptor.

Figura 7.1

Si vamos particularmente al concepto de Big Data, podemos decir que es un término que proviene del inglés y que se traduciría como “datos masivos” o “grandes datos”. Muchas son las definiciones que entidades y organizaciones han dado para el término big data, pero todas ellas se pueden resumir en el conjunto de datos cuyo tamaño supera considerablemente la capacidad de captura, almacenamiento, gestión y análisis del software convencional de bases de datos.

Volviendo al tema de los datos, los mismos se pueden categorizar de diferentes maneras.

La primera categorización que podemos encontrar es la de datos estructurados, datos semiestructurados o datos no estructurados.

7.2 TIPOS DE DATOS

7.2.1 Datos estructurados

La mayoría de las fuentes de datos tradicionales son datos estructurados, datos con formato que poseen campos fijos. En estas fuentes, los datos vienen en un formato bien definido que se especifica en detalle, y que conforma por lo general las bases de datos relacionales.

Algunos ejemplos son los datos de las bases de datos relacionales, las hojas de cálculo y los archivos. Los datos estructurados se componen de piezas de

información que se conocen de antemano, vienen en un formato especificado, y se producen en un orden especificado. Estos formatos facilitan el trabajo con dichos datos. Formatos típicos son: fecha de nacimiento (DD, MM, AA); documento nacional de identidad o pasaporte (por ejemplo, 8 dígitos y una letra); número de la cuenta corriente en un banco (20 dígitos), CBU (22 dígitos) etcétera.

7.2.2 Datos Semiestructurados

Los datos semiestructurados tienen un flujo lógico y un formato que puede ser definido, pero no es fácil su comprensión por el usuario. Estos datos no tienen formatos fijos, pero contienen etiquetas y otros marcadores que permiten separar los elementos.

La lectura de datos semiestructurados requiere el uso de reglas complejas que determinan cómo proceder después de la lectura de cada pieza de información. Un ejemplo típico de datos semiestructurados son el texto de etiquetas de lenguajes XML y HTML.

7.2.3 Datos no estructurados

Los datos no estructurados son datos sin tipos predefinidos. Se almacenan como “documentos” u “objetos” sin una estructura uniforme, y se tiene poco o ningún control sobre ellos. Datos de texto, video, audio, fotografía son claros ejemplos de datos no estructurados.

Por lo general los datos no estructurados no tienen campos fijos; ejemplos típicos son: audio, video, fotografías, documentos impresos, cartas, mensajes de correo electrónico y de texto, formatos de texto libre que se ingresan por medio de algún formulario, mensajes instantáneos SMS, Whatsapp, Telegram. Se calcula que al menos, el 80% de la información de las organizaciones no se encuentra cargado en las bases de datos, sino que se encuentran esparcidos a lo largo y ancho de la organización. Todos estos datos se conocen como datos no estructurados. Sin duda, los datos más difíciles de dominar por los analistas son los datos no estructurados, pero su continuo crecimiento ha provocado el nacimiento de herramientas para su manipulación.

7.2.4 Otra Categorización

Los datos también se pueden categorizar como datos privados, datos anónimos, datos agregados e Insights

Los ***datos privados***, son datos propios de una persona en particular, los mismos pueden servir para identificar a la persona, posicionarla geográficamente, o perfilar sus gustos y preferencias. Un ejemplo de dato privado podría ser el siguiente:

Me llamo Pedro, tengo 33 años y viajo a Rawson todos los días a las siete de la mañana.

Luego tendríamos los ***datos anónimos o anonimizados***, que son datos a los cuales se les quita la información personal, es decir la información que apunta unívocamente a un individuo.

También están los que se denominan ***datos agregados***, que no son ni más ni menos que datos relacionados por medio de algún patrón. Como ejemplo podría ser:

Somos cien personas de menos de 30 años que viajamos de lunes a viernes a Rawson a las 7 de la mañana

Por último, existen los ***Insights***, que son datos anonimizados, agregados y extrapolados al total de la población para dar una respuesta a una pregunta en base a datos.

Podríamos decir que todas las empresas que no estén dando respuestas a sus necesidades de negocio por medio de datos, están perdiéndose una posibilidad de obtener grandes ventajas competitivas sobre empresas similares del mismo rubro, porque el mundo al que vamos, y la competencia a la que nos dirigimos, utiliza datos para tomar decisiones.

Obviamente que los datos por sí solos, no nos van a brindar ningún tipo de valor agregado. Es por este motivo, que aparece otro concepto estrechamente relacionado con el de Big Data, que es el Machine Learning.

El **Machine Learning** es una disciplina del campo de la Inteligencia Artificial que busca a través de algoritmos, dotar a las computadoras con la capacidad de identificar patrones en datos masivos y elaborar predicciones. Esta elaboración de predicciones es la que se conoce como análisis predictivo, el cual es otro concepto de gran importancia a la hora de hablar de Big Data y Machine Learning.

El **análisis predictivo** es muy útil para las empresas, ya que permite adelantarse a las demandas de sus clientes. El análisis predictivo es una forma de análisis avanzado que examina datos o contenidos para responder a la pregunta: ¿qué es probable que ocurra en el futuro? Gracias a la tecnología Big Data, los datos obtenidos a través de todos los sistemas conectados pueden interpretarse para obtener predicciones sobre cómo se va a comportar una persona o un grupo de personas

Nuevamente si vamos a un ejemplo práctico, el análisis predictivo funciona de la siguiente forma:

Un día estás pensando en comprarte una Notebook, por lo cual entras a Mercado Libre y buscas varias opciones. Un instante después recibís un correo electrónico con una oferta, y cuando entras a leer el diario, ves que todas las

publicidades que te aparecen en la página, están relacionadas con ofertas de notebook de diversos locales comerciales. Cuando algo así ocurre, seguro que en más de una ocasión te has preguntado: ¿me leen el pensamiento?, ¿Me están espiando? De alguna manera podría decirse que sí, porque acontecimientos de este tipo no suceden por casualidad, detrás se encuentra lo que es el ***análisis predictivo***.

Así como existe el análisis predictivo, existen otros tipos de análisis de datos, que se pueden utilizar. El análisis descriptivo, el diagnóstico, el predictivo y el prescriptivo.

El análisis predictivo, como se dijo previamente utiliza los datos recopilados para realizar predicciones sobre un supuesto, por su parte el análisis prescriptivo toma esos datos y profundiza en ellos hasta encontrar la manera de lograr que ese supuesto ocurra. En cuanto al análisis descriptivo y el diagnóstico, podríamos decir que el análisis descriptivo examina el hecho en sí, mientras que el análisis diagnóstico atiende a las causas.

7.3 DISTINTOS ALGORITMOS DE MACHINE LEARNING

Los algoritmos de Machine Learning se dividen en tres categorías, siendo las dos primeras las más comunes:

7.3.1 Aprendizaje supervisado

Estos algoritmos cuentan con un aprendizaje previo basado en un sistema de etiquetas asociadas a unos datos que les permiten tomar decisiones o hacer predicciones. Un ejemplo es un detector de spam que etiqueta un e-mail como spam o no dependiendo de los patrones que ha aprendido del histórico de correos (remitente, relación texto/imágenes, palabras clave en el asunto, etc.). Se denomina aprendizaje supervisado, porque de alguna forma, nosotros como usuarios tenemos que indicarle al sistema cuando un correo es spam, o cuando un remitente envía spam, y en base a eso, el algoritmo aprende.

7.3.2 Aprendizaje no supervisado

Estos algoritmos no cuentan con un conocimiento previo. Se enfrentan al caos de datos con el objetivo de encontrar patrones que permitan organizarlos de alguna manera. Por ejemplo, en el campo del marketing se utilizan para extraer patrones de datos masivos provenientes de las redes sociales y crear campañas de publicidad altamente segmentadas. También suelen utilizarse mucho en la política, para realizar

el perfilamiento o la segmentación de personas, basándose en sus publicaciones estando a favor o en contra de ciertos temas sensibles.

7.3.3 Aprendizaje por refuerzo

El objetivo es que un algoritmo aprenda a partir de la propia experiencia. Esto es, que sea capaz de tomar la mejor decisión ante diferentes situaciones de acuerdo a un proceso de prueba y error en el que se recompensan las decisiones correctas. En la actualidad se está utilizando para posibilitar el reconocimiento facial, hacer diagnósticos médicos o clasificar secuencias de ADN.

7.4 DATA MINING O MINERÍA DE DATOS

El Data Mining o Minería de datos es otro concepto estrechamente relacionado con lo que es la tecnología del Big Data. La minería de datos es el conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos de forma automática con el objetivo de encontrar patrones, tendencias o reglas que expliquen los datos en un determinado contexto. Esto intenta ayudar a comprender el contenido de un repositorio de datos. Utiliza prácticas estadísticas e incluso en algunos casos usa algoritmos próximos a la inteligencia artificial.

El nombre surge a raíz de la cuarta revolución industrial y su nombre viene de la propia minería, ya que ahora los profesionales en vez de buscar carbón intentan ofrecer información interesante para las empresas a través de minar datos.

El proceso de Data Mining tiene cuatro etapas principales:

- La determinación de objetivos
- El procesamiento de datos
- La determinación del modelo
- El análisis de resultados

El primero de estos procesos trata de dar información relevante de la base de datos proporcionada por el cliente, después se seleccionan y enriquecen los datos proporcionados por la empresa, a continuación, se realizan análisis estadísticos de los datos y después se lleva a cabo una visualización gráfica de los mismos. Por último, el profesional del Data Mining verificará si los resultados son coherentes y los cotejará con los obtenidos por los análisis estadísticos y de visualización gráfica.

El cliente comprobará si los datos son novedosos y si aportan información relevante para tomar decisiones.

7.5 ¿QUÉ ES BIG DATA?

Hasta aquí se han nombrado conceptos sueltos, que son importantes conocer para comprender mejor qué es Big Data. Hablamos de datos e información, de cómo se categorizan los datos. También hablamos de datos estructurados, semiestructurados y no estructurados, y como en base a esos datos estructurados y por medio del Machine Learning se pueden realizar diversos tipos de análisis, entre los que los análisis predictivos son quizás los más importantes. También nombramos el concepto de Data Mining, que nos sirve para analizar y tratar de entender que tiene para ofrecernos un repositorio de datos. Habiendo hecho una breve introducción de estos temas, podemos profundizar un poco en lo que es la tecnología de Big Data.

En Big Data confluyen una gran cantidad de tecnologías que venían madurando desde la primera década del siglo XXI, y que se han consolidado durante los años 2010 en adelante.

La aparición de las redes sociales, las mejoras de velocidad de procesamiento, la reducción de los costos del hardware, el aumento de la velocidad de internet, la creación de dispositivos móviles con sensores que capturan posicionamiento o la aceleración, y que además traen cámaras de fotos integradas, internet de las cosas, entre otras tantas, han generado una explosión de datos.

Se calcula que, en los últimos dos años, se ha generado el 90 % de los datos que se encuentran en circulación. Y cada año se producen más y más datos. Esto hace que el Big Data, sea una tecnología que no para de crecer, y cada día toma un papel más preponderante dentro de las empresas.

No existe una única definición de Big Data, aunque sí hay un cierto consenso en la mayoría de las definiciones que podemos encontrar, en donde el gran volumen de información, y la necesidad de capturarla, almacenarla y analizarla aparecen en todas ellas.

La primera definición que pondremos es la de Adrian Merv, vicepresidente de la consultora Gartner, que dice:

“Big Data excede el alcance de los entornos de hardware de uso común y herramientas de software para capturar, gestionar y procesar los datos dentro de un tiempo transcurrido tolerable para su población de usuarios”

Otra definición muy significativa es del McKinsey Global Institute

“Big Data se refiere a los conjuntos de datos cuyo tamaño está más allá de las capacidades de las herramientas típicas de software de bases de datos para capturar, almacenar, gestionar y analizar”

Otra fuente de referencia es la consultora tecnológica IDC, que define a Big Data de la siguiente manera:

"Big Data es una nueva generación de tecnologías, arquitecturas y estrategias diseñadas para capturar y analizar grandes volúmenes de datos provenientes de múltiples fuentes heterogéneas a una alta velocidad con el objeto de extraer valor económico de ellos"

Como podemos ver, todas las definiciones de Big Data, apuntan al gran volumen de datos, y al capturar, almacenar y analizar estos datos.

7.5.1 Las "5V" del Big Data

El concepto de Big Data no hace referencia únicamente al tamaño de la información, sino también a la variedad del contenido, y a la velocidad con la que pode

Sin embargo, el concepto no hace referencia simplemente al **V**olumen de la información, sino también a la **V**ariedad del contenido y a la **V**elocidad con la que los datos se generan, almacenan y analizan. Estas dimensiones son las principales y junto con la **V**eracidad y el **V**alor de los datos, son las que se conocen como las "5V" de Big Data, es decir volumen, velocidad, variedad, veracidad y valor de los datos.

7.5.1.1 Volumen: Como su propio nombre indica, big data corresponde al gran volumen de datos que se generan diariamente en las empresas y organizaciones de todo el mundo. Como ejemplos de empresas que hacen uso de Big Data, para enriquecer las experiencias de los clientes, o para mejorar sus ventas podemos citar, Netflix, Spotify, Mercado Libre, Facebook, entre otras tantas.

7.5.1.2 Velocidad: La velocidad se refiere a flujos de datos, la creación de registros estructurados y la disponibilidad para el acceso y la entrega de los mismos. Es decir, que tan rápido se están produciendo los datos, así como la velocidad en la que se trata de satisfacer la demanda de éstos. La tecnología Big Data es capaz de y trabajar en tiempo real con las fuentes generadoras de información. Estas fuentes pueden ser sensores, redes sociales, blogs, páginas webs, dispositivos móviles, etc. Básicamente son fuentes que generan millones y millones de datos por segundo. Por otro lado, la tecnología Big Data, tiene que tener la capacidad de analizar dichos datos con la suficiente rapidez, reduciendo así los largos tiempos de procesamiento que presentaban las herramientas tradicionales de análisis.

7.5.1.3 Variedad: Big Data debe tener la capacidad de combinar una gran variedad de información digital en los diferentes formatos en los que se puedan presentar. Las empresas líderes en tecnología siempre han tenido un problema para traducir grandes volúmenes de información transaccional en decisiones. Debido a que ahora hay más tipos de información para analizar, provenientes principalmente de las redes sociales, la complejidad aumenta. Esta variedad en los datos incluye datos

estructurados (bases de datos) y no estructurados, datos jerárquicos, documentos, correo electrónico, datos de medición, vídeo, imágenes fijas, audio, datos de cotizaciones, transacciones financieras, etc., entre otras clases de fuentes generadoras de diferentes tipos de información.

7.5.1.4 Veracidad: La tecnología Big data debe de ser capaz de tratar y analizar inteligentemente este gran volumen de datos con la finalidad de obtener una información verídica y útil que nos permita mejorar la toma de decisiones basada en los datos más exactos.

7.5.1.5 Valor: Este concepto hace referencia a los beneficios que se desprenden del uso de Big Data (reducción de costes, eficiencia operativa, mejoras de negocio). Lo ideal es conseguir todo esto de manera eficiente.

7.5.2 Fuentes de Datos

Hoy en día los datos proceden de numerosas fuentes, que pueden ir desde videojuegos hasta las innumerables cantidades de datos de operaciones comerciales en supermercados, bancos, la administración pública, los sensores, los teléfonos inteligentes, relojes inteligentes, el internet de las cosas, etcétera.

Todos estos datos procedentes de fuentes tradicionales han ido constituyendo los grandes volúmenes de datos, y crecen de modo exponencial. Las bases de datos de organizaciones y empresas han ido creciendo y pasando de volúmenes de datos de terabytes a petabytes.

Sin embargo, son los datos de la Web los que hoy día constituyen la mayor cantidad de información vinculada a lo que es Big Data, ya que, probablemente, es la fuente de datos más ampliamente utilizada y reconocida en la actualidad.

Hay muchas otras fuentes que añaden grandes cantidades de datos, algunos de los orígenes más usuales son:

- Datos de la Web.
- Datos de los medios sociales (redes sociales, blogs, wikis).
- Datos de Internet de las cosas.
- Datos de interconexión entre máquinas.
- Datos industriales de organizaciones y empresas
- Datos de la industria del automóvil.
- Datos de redes de telecomunicaciones.
- Datos de medios de comunicación (prensa, radio, televisión, cine)
- Datos procedentes de sensores en los más diferentes campos de la industria y la agricultura.
- Datos de videojuegos.

- Datos procedentes de posiciones geográficas y de telemetría.
- Datos procedentes de chips NFC, RFID
- Datos procedentes de servicios de telefonía móvil inteligente: texto, datos, audio, video, fotografía.
- Datos personales, datos de texto
- Cientos de datos y documentos no estructurados
- Otros.

Cada uno de estos puntos arriba expuesto, podría constituir en sí una categorización de las fuentes de datos, que, a su vez, pueden contener un gran número de fuentes diversas, que recolectan, almacenan, procesan y analizan.

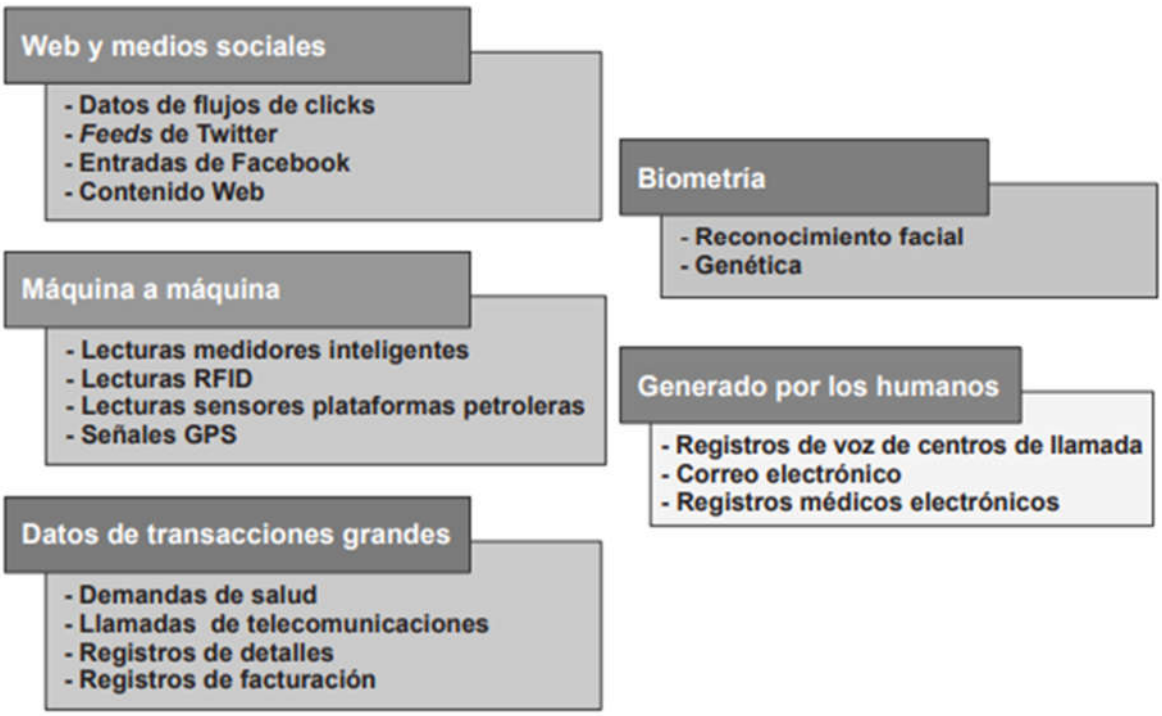

Figura 7.2 - Fuentes de datos de Big Data.
Fuente: Big Data Análisis de grandes volúmenes de datos en organizaciones Luis Joyanes Aguilar

7.5.3 Tomar decisiones en base a datos

Las personas estamos a gusto, cuando estamos cerca de otras personas que nos conocen mucho, como por ejemplo estando cerca de nuestra familia, nuestros amigos, o si vamos a un restaurante y el mozo que nos atiende, es el que los hace habitualmente.

Uno se siente cómodo con estas personas, porque justamente estas personas nos conocen y tienen una gran cantidad de datos nuestros, y por lo general usan esos datos para que estemos mejor.

Con los servicios nos ocurre algo muy similar, nos sentimos a gusto consumiendo servicios que utilizan nuestros datos para hacer el entorno más cómodo. Los principales ejemplos que vienen a la mente en estos casos son Netflix, o Spotify, que usan nuestras búsquedas anteriores para generar un perfil de nuestros gustos, y luego nos recomiendan series o canciones en base a dicho perfil.

También empresas como Mercado Libre hacen uso de Data Mining o de Machine Learning, para recomendarnos productos en base a nuestras búsquedas anteriores, o nuestro historial de favoritos.

Para cualquier empresa que quiera ser competitiva en los tiempos que corren, tomar decisiones, basándose en los datos puede ser una gran opción.

Hay que tener en cuenta que la tecnología Big Data, no es aplicable solo a grandes empresas. Ya que, si bien el volumen de datos variará mucho entre una Pyme y una Multinacional con cientos de sucursales alrededor del mundo, las bondades que puede brindar el análisis de datos, y la toma de decisiones en base a ellos, pueden ser igualmente beneficiosas para empresas de cualquier tamaño.

El primer paso a la hora de saber si se puede aplicar o no algún método de análisis de datos dentro de una Pyme, será relevar cuales son las fuentes de datos con la que la empresa trabaja.

7.5.3.1 Relevar Fuentes de Datos

Una buena forma de realizar el relevamiento de las fuentes de datos seria buscando bases de datos, y hojas de cálculo que se usan para intercambiar información internamente dentro de la organización.

Esta sería la forma más sencilla de buscar datos estructurados que ya están siendo utilizados por la empresa.

Luego habría que buscar datos semiestructurados o no estructurados que se estén utilizando, y plantearse si de alguna forma se los puede estructurar para obtener algún tipo de valor agregado.

Por ejemplo, si nos llegan pedidos de clientes por correo electrónico (datos no estructurados), quizás se podría generar algún tipo de formulario, en donde el cliente pueda seleccionar de alguna forma desde una lista cuales son los productos que me solicita (datos estructurados). Esto a futuro podría ayudar a conocer las preferencias del cliente, evitar errores en los pedidos, agilizar el proceso de preparación del pedido, entre otros muchos beneficios.

7.5.3.2 Analizar Datos

Una vez que se conocen cuáles son las fuentes de datos con las que se va a trabajar, se deberían analizar las mismas, para saber qué tipo de información se

podría sacar de ellas. Básicamente sería hacer una minería de datos en forma casi artesanal.

Esto probablemente nos va a clarificar mucho en cuanto a lo que se puede lograr a futuro, y que nos haría falta para potenciar el negocio.

7.5.3.3 Plantear Objetivos

La empresa debe tener en claro cuáles son sus objetivos, para ver a donde hay que apuntar las armas con las que contamos. Cuando decimos armas, nos referimos a los datos que ya fueron relevados y analizados.

Si una empresa quiere por ejemplo mejorar sus ventas, el foco debería estar principalmente en ver cuáles son sus productos estrella, como mejorar la captación de clientes, cómo obtener algún tipo de feedback en cuanto a la relación del cliente para con la empresa, y cómo mejorar esa experiencia, para que el cliente se encuentre cómodo y encuentre lo que está buscando de forma fácil y rápida.

Si una empresa quiere mejorar su producción, debería ver donde ocurren los mayores retrasos, o en donde se está generando la mayor cantidad de reprocesos o fallos. De esta forma podría intentar controlar mejor la línea de producción en esos sectores para mejorar los tiempos de producción

7.5.3.4 Análisis de Herramientas Disponibles

En cuanto a herramientas para hacer análisis predictivo, debemos plantearnos si necesitamos aplicar Machine Learning o Data Mining. Hay gran cantidad de herramientas Open Source en la web. Algunas de las más conocidas son:

- PyTorch.
- Weka.
- H2O.
- KNIME.

7.5.3.5 Planificar recolección

En caso de que no se cuenten con datos, se debería planificar la recolección de los mismos.

Los datos a recolectar dependen en gran medida del tipo de empresa y sus objetivos. Datos de clientes, de transacciones, de ventas, de producción, de proveedores, de uso de materiales o insumos, de ingresos y egresos de dinero, son solo algunos ejemplos útiles a tener en cuenta.