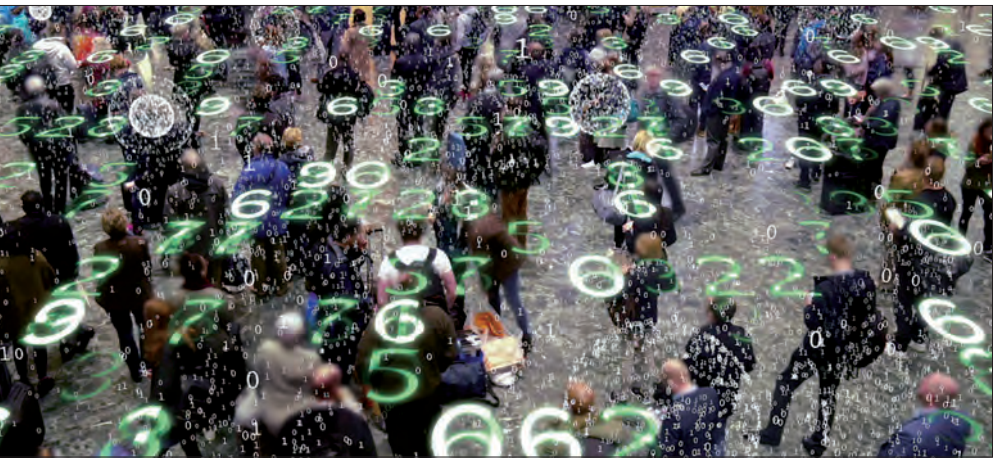
BIG DATA

CADA vez nos extraña menos escuchar o leer que organizaciones, empresas e incluso gobiernos emplean técnicas de Big Data para alcanzar sus objetivos. Sin embargo, es menos habitual conocer con detalle a qué se refiere este rimbombante nombre. En este reportaje trataremos de explicar qué es el Big Data, cuáles son sus principales características, para qué sirve y cuáles son algunos de los peligros que pueden derivarse de su uso.

Si tenemos en cuenta que la capacidad de los medios de almacenamiento de datos va creciendo a lo largo de los años, no parece tener demasiado sentido llamar a una tecnología Big Data o «datos masivos», puesto que lo que hoy parece masivo, en unos pocos años puede resultar manejable e incluso minúsculo (todavía recuerdo una imagen en la que varios operarios descargan un enorme disco duro de 5 MB capacidad y una tonelada de peso, cuando hoy en día cualquier tarjeta de memoria almacena 1000 veces más datos y pesa unos pocos gramos). Sin embargo, consideramos masivo a un conjunto de datos que no puede tratarse desde un solo ordenador (o unos pocos) y requiere ser capturado, almacenado, filtrado y procesado por muchos ordenadores colaborando entre sí.



Entonces, ¿podríamos haber hablado de Big Data hace 20 años? Sí... y no. La razón por la que este término se ha hecho tan popular últimamente es porque por primera vez en la historia disponemos de una capacidad de almacenamiento de datos (discos duros y similares) y de procesamiento (CPU) lo suficientemente grandes como para enfrentarnos a problemas irresolubles para la mayoría de humanos.

Pensemos, por ejemplo, en el ajedrez. Está claro que desde hace bastantes años es posible enfrentarse a máquinas que despliegan un juego más que aceptable. Pero, en la actualidad existen máquinas tan buenas jugando al ajedrez que ningún humano es lo suficientemente bueno como para resolver los problemas ajedrecísticos que plantean en sus partidas. Como humano, me da cierta rabia admitir que, en este dominio concreto, hemos sido totalmente superados por las máquinas gracias a técnicas de análisis masivo de partidas previas.

Algo parecido ocurre en otros ámbitos: si bien es cierto que las máquinas llevan más de un siglo ayudándonos a procesar mejor grandes cantidades de información, por primera vez han dejado de ser meros asistentes y nos permiten ver relaciones que hasta ahora eran invisibles a nuestros ojos.

LAS «UVES» DEL BIG DATA

Cuando se habla de datos masivos, se suelen mencionar las 3, 4, 5 o incluso 6 «uves». Las tres iniciales eran Volumen, Variedad y Velocidad. A estas, pronto se le sumó Veracidad, y más adelante hay quien habla también de Valor, Variabilidad y otras. Veamos a qué se refiere cada una de ellas.

Volumen se refiere a la enorme cantidad de datos generados y almacenados. De los miles de caracteres (kilobytes), hemos pasado a los megabytes (millones), gigabytes (miles de millones) y terabytes (billones) a nivel doméstico, pero incluso superamos los petabytes (miles de billones), exabytes (millones de billones) o zottabytes (miles de millones de billones) si nos referimos al gigantesco caudal de datos que se generan en las redes sociales, los sensores de sistemas conectados o la información recolectada de experimentos científicos como el colisionador de hadrones del CERN o radiotelescopios astronómicos.

Tan enormes son estos conjuntos de datos que, si los pudiéramos imprimir y poner cada hoja una encima de la otra, podríamos recorrer miles de veces la distancia de la Tierra al Sol. Las técnicas de Big Data han evolucionado para ser capaces no solo de almacenar esa

información, sino también de tratarla en un tiempo razonable.

Variedad se refiere a la naturaleza de los datos. En los conjuntos de datos anteriores al Big Data, lo normal es que cada fila en una tabla tuviera siempre los mismos campos. Si, por ejemplo, almacenamos lecturas del contador de la luz, todas las filas almacenan una información similar (nº del titular, fecha, valor de la lectura, etc.), y lo mismo ocurre para cualquier otro ámbito. En el caso del Big Data, sin embargo, se combinan tantas fuentes de datos heterogéneas que no es posible almacenarlos de forma homogénea en una sola tabla. Los sistemas de Big Data precisan ser más flexibles para poder almacenar y procesar esta variedad de datos.

Velocidad se refiere al imparable ritmo de generación de estos datos masivos. En unos pocos segundos de un experimento en el CERN se generan terabytes de información que es necesario capturar y almacenar a una velocidad suficiente como para que no se pierdan los próximos datos (habitualmente esto implica tener que hacerlo en tiempo real). En los sistemas previos al Big Data era habitual tener momentos para poder realizar volcados de datos en determinados momentos puntales, pero actualmente resulta inviable detener los sistemas de captura y almacenamiento si queremos no perder ningún dato.

Veracidad se refiere al tremendo reto que supone dilucidar cuáles de esos datos que estamos capturando son veraces y cuáles no. Dado que muchos de los sistemas de poder más importantes del mundo dependen de análisis proporcionados por técnicas de Big Data, hay muchos intereses por falsear algunos de estos datos. Pensemos, por ejemplo, en la cantidad de comentarios falsos que hay en TripAdvisor o sistemas de recomendación similares: muchos dueños de restaurantes convencen a sus conocidos para que hagan una reseña muy benévola de sus locales, al mismo tiempo que restaurantes de la competencia tratan de llenar de reseñas negativas ese mismo perfil. Un sistema de Big Data que no quiera ser engañado fácilmente, tendrá que incluir mecanismos para determinar la veracidad de sus fuentes de datos.

La aplicación de técnicas de Big Data ha influido positivamente en ámbitos como la predicción meteorológica, la salud y la agricultura.

Valor se refiere a la capacidad de detectar qué fuentes de datos son más valiosas que otras y pasar de la mera acumulación de conjuntos de datos masivos a la priorización o valoración de determinados subconjuntos frente a otros. Hay quienes hablan aquí de Smart Data, datos inteligentes, en contraposición a los datos meramente masivos. A pesar de que el valor no siempre es fácil de identificar por parte de un programa informático, hay casos claros como el del algoritmo de clasificación automática de páginas web de Google (PageRank) que demuestran claramente cómo es posible determinar automáticamente la existencia de enlaces entre webs que tienen más valor que otros.

Variabilidad (no confundir con Variedad) se refiere a la volatilidad de muchos de los datos que se capturan para conformar un conjunto de datos masivos. Gracias a la Internet de las Cosas o a la presencia de miles de cámaras conectas a Internet, continuamente se están generando cantidades enormes de información que no se están almacenando en ningún sitio. Si un sistema de Big Data dependiera de estas fuentes de datos, debería emplear técnicas para evitar que esta falta de persistencia pudiera afectar al resultado de sus predicciones (almacenando temporalmente parte de esta información en sistemas intermediarios, calculando valores agregados como la media o la desviación típica para poder resumir esos periodos en los que no ha sido posible enviar o almacenar los valores en tiempo real, etc.).

Y así podríamos seguir determinando más «uves», pero la idea general es que un sistema de Big Data no solamente es complejo por el masivo volumen de datos que tiene que gestionar, sino también porque tiene que extraer valor de datos que se generan a toda velocidad desde fuentes de datos muy diversas. Algunas de las cuales no son veraces o no almacenan su información para poder consultarla en el futuro.

AVANCES GRACIAS AL BIG DATA

Son numerosos los ámbitos en los que la aplicación de técnicas de Big Data han logrado avances muy notables. La meteorología ha mejorado espectacularmente su capacidad →

predictiva, la biología ha resuelto problemas relacionados con la genética, el enrollamiento de proteínas o la simulación de procesos muy difíciles de replicar in vitro, la logística ha sido capaz de prever patrones de demanda latentes que han mejorado muchísimo el abastecimiento y reducido los tiempos de entrega, entre muchos otros ejemplos.

La combinación de conjuntos de datos aparentemente inconexos ha permitido mejorar procesos de formas inesperadas. Por ejemplo, la cadena de supermercados estadounidense Walmart descubrió que, en los días previos a la llegada de un huracán a una zona, la demanda de unas galletas precocinadas para microondas se disparaba debido a que la gente suele hacer acopio de productos no perecederos y con alto valor calórico. Así, la combinación de los datos logísticos con los meteorológicos permitió a Walmart atender perfectamente la demanda de este tipo de productos.

Algo parecido emplean muchas organizaciones contra el crimen al combinar datos masivos de compañías aéreas, servicios de telefonía, compras por Internet, etc. para identificar a criminales o terroristas, resolver casos que habían permanecido años sin respuesta o desmontar redes de tráfico de sustancias ilegales, personas o armas.

Puede estar amenazada nuestra intimidad y muy extendido el prejuicio.

EL REVERSO TENEBROSO DEL BIG DATA

Pero estos sistemas de predicción automática de crímenes me recuerdan mucho a la sección de «pre-crimen» de la película *Minority Report*. Puede darse el caso de que hayamos heredado una vieja escopeta y compremos por Internet una caja de cartuchos para probar si funciona. Además, aprovechamos las ofertas de una tienda del barrio para comprarle unas medias a nuestra hija que pagamos con VISA. Y semanas más tarde buscamos vídeos en Youtube de cómo cortar metal con una sierra porque queremos hacer una chapuza en casa. Un algoritmo de Big Data podría estar constantemente analizando fuentes de datos heterogéneas como los historiales de venta de Internet, los pagos de tarjetas de crédito y los historiales de navegación en Internet y concluir que somos un potencial ladrón de bancos que va a cortar el metal de una escopeta para hacer una recortada y ponerse una media en la cabeza para perpetrar su robo.

Un ejemplo muy famoso en el que un sistema de Big Data ocasionó un problema, incluso acertando en su predicción, fue el de otra cadena de supermercados estadounidense, Target, que envió unos cupones de descuento para productos de premamá a una adolescente en Minneapolis. Al recibirlos, su padre se puso como una furia y fue a protestar al supermercado. Sin embargo, a los pocos días la chica confesó que efectivamente estaba embarazada. El sistema de Big Data simplemente analizó su historial de compra y vio que los complementos vitamínicos y el resto de productos que había comprado últimamente coincidían bastante bien con su perfil de una futura mamá.

Pero no solamente está amenazada nuestra intimidad, sino también aspectos que afectan mucho más directamente a nuestra subsistencia. En su libro *Weapons of Math Destruction* (*Armas de Destrucción Matemática*), Cathy O'Neil nos explica cómo el Big data puede incrementar la desigualdad y amenazar a la democracia a través de ejemplos claros como el acceso a la universidad, a un crédito, a una vivienda, a un seguro médico, etc.

La tesis de O'Neil es que muchos de los algoritmos de Big Data están plagados de prejuicios y de asunciones que no siempre se cumplen. Además, suelen apoyarse más en los datos que son fáciles de conseguir que en los datos que realmente necesitarían para hacer una buena predicción. Dada la complejidad de su programación y los modelos matemáticos subyacentes, mucha gente que cuestionaba los prejuicios de las personas que tenían que tomar decisiones difíciles ha dejado de cuestionar esas mismas decisiones si vienen de una máquina.

Como hemos visto, los sistemas de Big Data pueden suponer una mejora muy notable de la calidad de vida de mucha gente, pero al mismo tiempo pueden usarse para justificar decisiones políticas bajo una pátina de supuesta objetividad. Por esta razón, es muy importante conocer no solo sus potencialidades sino también sus debilidades y las implicaciones derivadas de que aspectos importantes de nuestras vidas estén regidos exclusivamente por máquinas.

PABLO GARAIZAR