Revista Digital Universitaria
Vol. 18, Núm. 7, septiembre-octubre 2017

Científico de datos: codificando el valor oculto e intangible de los datos

José Gerardo Moreno Salinas

Resumen

La ciencia de datos es una disciplina emergente y de gran pertinencia para todas las organizaciones que deseen codificar el valor oculto e intangible de sus datos. Hoy más que nunca estamos más conectados con personas y dispositivos, tenemos acceso a más redes y servicios, y sin duda consumimos y producimos mayores cantidades de datos e información. Por lo que requerimos contar con las habilidades, conocimientos, experiencias y técnicas de los científicos de datos para procesar, analizar y visualizar de formas más inteligentes los datos en información, promoviendo así, más y mejores conocimientos de nuestra realidad en sus contextos. En este artículo se explican las principales áreas en las que desarrolla un científico de datos (*Big Data*, minería y visualización de datos) y las intersecciones entre éstas; se incluyen ejemplos de proyectos desarrollados por científicos de datos y del gran valor que han sabido codificar. Además, se presenta una interpretación de los elementos que constituyen al científico de datos.

Palabras clave: científico de datos, ciencia de datos, *Big Data*, minería de datos, visualización de datos.

Data scientist: encoding the hidden and intangible value of the data

Abstract

Data science is an emerging discipline of great relevance to all companies that wishing to encode the hidden and intangible value of data. Today more than ever we are connected to more people and devices, we have access to more networks and services, and not there are doubt that we consume and produce greater amounts of data and information. So we require the skills, knowledge, experiences and techniques of data scientists to process, analyze and visualize the data toward information in smarter ways, promoting more and better knowledge of their reality in their contexts. This article explains the main areas in which a data scientist develops (Big data, data mining and data visualization) and the intersections between these, including examples of projects developed by data scientists and the great value that they have known how to code it. In addition, to present an interpretation of the elements that constitute the scientific data.

Key words: data scientist, data science, Big Data, data mining, data visualization.

Recepción: 04/09/17
Aprobación: 12/09/17
DOI: http://dx.doi.org/10.22201/codeic.16076079e.2017.v18n7.a2
Universidad Nacional Autónoma de México, Coordinación de Desarrollo Educativo e Innovación Curricular (CODEIC)

José Gerardo Moreno Salinas gerardo_moreno@cuaed.unam.mx

Twitter: @jgmorenos

Maestro en Ingeniería de Sistemas por la Universidad Nacional Autónoma de México (UNAM) e Ingeniero en Sistemas Computacionales por el Instituto Tecnológico de Ciudad Guzmán (ITCG). Cabe señalar que mientras estudió la maestría fue seleccionado para representar a nivel nacional a los alumnos del posgrado de Ingeniería en la quinta edición de la Cumbre de Negocios; su proyecto de tesis de licenciatura recibió el reconocimiento de servicio a la comunidad por la Universidad del Norte de Texas. Ha cursado diplomados sobre tecnologías de la información en la UNAM y cursos de actualización relacionados con la ciencia de datos. Se certificó en el manejo del software Tableau en Philadelphia. Es experto en el manejo de técnicas cuantitativas para la toma de decisiones, matemáticas, estadística y ciencia de datos; el diseño de instrumentos de evaluación y la puesta en línea de bases de datos interactivas. Al respecto, ha publicado en revistas de difusión y divulgación de la ciencia, así como en revistas especializadas y capítulos de libros. Ver más...

Introducción

Actualmente todos los que navegamos en Internet somos, en cierta medida y de acuerdo a las proporciones, consumidores y/o productores de datos e información. En éste mismo esquema podemos concebir a las compañías que tienen presencia en Internet en dos grandes grupos, están las que apostaron por la democratización de sus servicios a través de Internet y posibilitaron que los usuarios pasaran de ser sólo consumidores a también ser productores de información y contenidos, entre las más representativas están Google, YouTube, Facebook, Twitter, Instagram, Waze, Airbnb, entre otras tantas. También están las que ofrecen principalmente el servicio de consumo de información y contenidos, como es el caso de Netflix, Amazon, Spotify y YouTube por mencionar algunas. Sin importar de qué compañía se trate, se dieron cuenta que estaban almacenando grandes cúmulos de datos y lejos de verlo como un problema de escalabilidad —por sí mismo significa un reto e inversiones millonarias—, identificaron un problema todavía mayor: cómo dotar de significado a los datos que registran los usuarios en sus sistemas, ya sea de manera consciente o inconsciente, y no sólo eso, sino cómo obtener un conjunto procesado y estructurado de datos que posibilitan una mejor comprensión teórica o práctica de la realidad en menor tiempo.

Con el registro de la huella digital[1] de los usuarios se pueden identificar sus preferencias e intereses, deseos de compra, tendencias y frecuencia de consumo, horarios de interacción, redes sociales, ubicación de conexión, dispositivos utilizados, entre otros tantos identificadores.

Las organizaciones deben saber aprovechar al máximo la información y explorar de manera inteligente cómo pueden beneficiarse del análisis de los

1. Para mayor información sobre el concepto de huella digital, revisar: https://www.internetsociety.org/es/tu-huella-digital

datos que generan sus usuarios, operaciones, productos o servicios. No hay que olvidar que ahora, más que nunca, el recurso intangible más valioso en nuestros tiempos es el poder de la información y del conocimiento que obtengamos de éste.

Un mundo más conectado

Estamos en tiempos donde la conexión a múltiples sistemas de información es innegable, cada vez nos conectamos a más servicios y somos más dependientes de éstos. El paradigma ha cambiado en pocos años, tal como lo advierten Hilbert y Lopez (2011), hemos pasado de ser analógicos a ser digitales, lo que ha propiciado que estemos conectados desde diferentes dispositivos, a toda hora y desde cualquier lugar. Como resultado vivimos en un mundo cada vez más conectado, donde la inmediatez de la información se ha convertido en una necesidad de primer orden para hacer negocios, establecer relaciones sociales, consumir contenidos multimedia e incluso, estudiar en modalidades no tradicionales.

Figura 1.
¿Qué sucede en línea cada 60 segundos?
Fuente:
Smart Insigts. Recuperado de: https://goo.gl/jiaDX2.

La siguiente infografía hace un recuento del crecimiento que han tenido algunas de las principales aplicaciones y servicios en Internet en los tres últimos años. Nos ayuda a tener un referente de la magnitud de datos que llegan a manejar estas grandes compañías, por ejemplo: en 2016, YouTube reporta que en su plataforma cargan 500 horas de vídeo cada 60 segundos, por lo que al término del año suman 262.8 millones de horas de vídeo, es decir, para poder ver todo el contenido cargado en un año en YouTube se requerirían 30 000 años. Y los datos siguen creciendo año tras año.[2]

2. Para mayor información sobre lo que actualmente está sucediendo en Internet, se recomienda visitar el sitio: http://www.internetlivestats.com/.

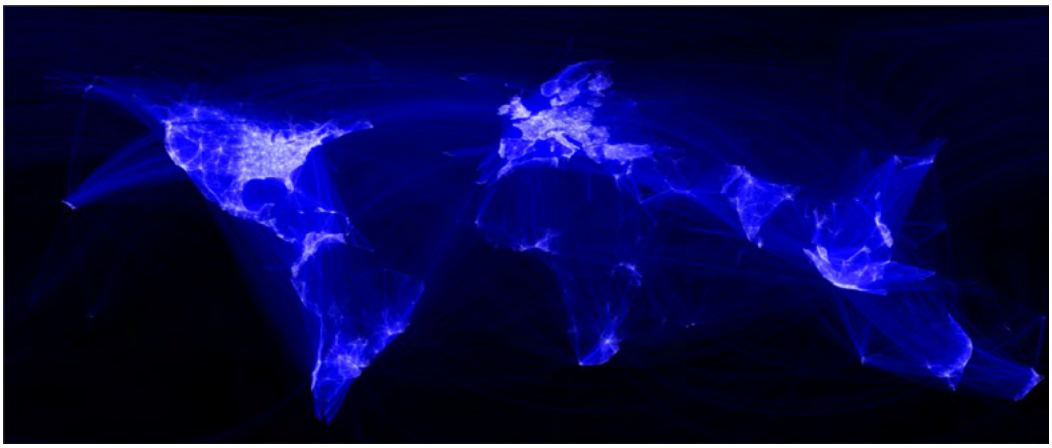

Figura 2.
Interconexión de Facebook en el mundo
Fuente:
http://fbmap.bitaesthetics.com/.

El ingeniero de datos en Facebook, Paul Butler (2010) interpretó muy bien el refrán "una imagen vale más que mil palabras", ya que a finales del 2010 desarrolló el ejemplo más claro que tenemos hasta el momento sobre la visualización de un mundo más conectado. Butler tomó una muestra de 10 millones de pares de amigos en Facebook y los combinó con sus datos de ubicación (latitud y longitud), generando así la siguiente visualización de datos:

En el artículo "40 maps that explain the internet" de Timothy B. Lee (2014), publicado en el sitio *Vox*, podrán consultar diferentes mapas e información sobre la evolución y conexión que ha tenido Internet desde sus inicios.

El valor subestimado de los datos

Las compañías como Google, Facebook y Twitter gastan increíbles cantidades de dinero para mantener sus sistemas, sin embargo, los usuarios finales no son quienes pagan directamente esos gastos, en lugar de ello proveen contenido a la vez que son objeto de ambiciosas campañas publicitarias, lo que significa que otras compañías están pagando los costos de infraestructura a cambio de obtener datos de los usuarios (Van der Aalst, 2014).

Para Twitter existen aplicaciones web donde se calcula el valor que tiene una cuenta, lo cual es un estimado con base al número de seguidores que ten-

gas, la cantidad de personas que te siguen, los tweets que escribes y la velocidad con la que ganas seguidores. Por ejemplo, al hacer la prueba en los sitios twalue.com y tweetvalue.com reportaron que mi cuenta en Twitter (@jgmorenos) está valuada en $18.47 y $44 dólares, respectivamente. Recientemente el analista Cakmak (2017), analizó el valor que tiene para Twitter la cuenta de Donald Trump (37.4 millones de seguidores con más de 35 mil Tweets) y la calculó en 2 mil millones de dólares. Hay que considerar que estos valores son estimaciones y habrá que tomarlos con reserva, pero al menos son una invitación para reflexionar y no subestimar el valor que tienen los datos.

Son varios los casos de éxito en donde las compañías se han beneficiado por codificar el valor oculto que tienen sus datos, para así mejorar sus productos y servicios, principalmente. Por ejemplo, Netflix ha sabido utilizar bien sus datos, pues tiene como objetivo principal: "ayudar a sus suscriptores a encontrar el contenido que realmente disfrutan, maximizando así su satisfacción y retención" (Elahi, 2015, p. 4). Desde sus inicios en 1997, con el servicio de renta y envío de DVD por correo postal, le dio una gran importancia a los datos de sus usuarios y en 2000 comenzó a desarrollar lo que sería su primer algoritmo (Cinematch) para crear un sistema que permitiera recomendar contenido de alto interés para cada uno de sus suscriptores. En el 2006, Netflix abrió su algoritmo a la comunidad científica y ofreció una recompensa de 1 millón de dólares para quién(es) lograran mejorar en un 10% su capacidad predictora, tuvieron que pasar tres años para que el grupo *BellKor's Pragmatic Chaos* logrará resolverlo. En 2007 comenzó con su servicio de descarga y reproducción (*streaming*) de películas y series, y después de seis años lograron recopilar suficientes datos para predecir con seguridad el éxito de su primera producción original "House of Cards". Éste es un claro ejemplo de cómo ser exitosos codificando datos y lograr que una serie obtenga alto interés de parte de los usuarios.

El sitio statista.com reportó que en el segundo cuatrimestre de 2017 Netflix tiene 103.9 millones de suscriptores a nivel mundial, de los cuales procesa en promedio 695 mil millones de eventos por día, es decir, una base de datos de 1.8 Petabytes diarios.[3] Algunos de los eventos registrados por Netflix son:

- ¿Desde dónde se conectan?
- ¿A través de qué dispositivo?
- ¿En qué horarios se conectan?
- ¿El tipo de contenido (película, serie) varia con el dispositivo?
- ¿Ven los créditos?
- ¿Cuánto tardan en ver el contenido?
- ¿Cuáles son sus actores y directores favoritos?

3. Cálculos realizado con base en lo publicado en: https://goo.gl/eFhQQa

- ¿Qué y cómo califican?
- ¿Qué buscan?
- Etcétera.

Sin una gran cantidad de datos, no hubiera sido posible que Netflix siguiera entrenando sus sistemas de recomendación. Se necesita contar con una gran serie de datos históricos para poder analizar todas las posibles combinaciones, y así identificar patrones y tendencias que permitan tener algoritmos más robustos al momento de hacer las recomendaciones a sus suscriptores. Y tal como la misma compañía advierte "alrededor del 75% de la visualización en Netflix es impulsada por el algoritmo de recomendación" (Vanderbilt, 2013).

Científico de datos

El considerado padre del "management", Peter Drucker (2004), reconoció que la sociedad postcapitalista es una sociedad basada en el conocimiento, donde el centro de la producción de la riqueza es el saber y no el capital. Los protagonistas claves en esta economía del conocimiento serán los "trabajadores del conocimiento", es decir, los que posean las capacidades, las habilidades, el pensamiento creativo y la tecnología para procesar, analizar y visualizar las grandes bases de datos.

> **El científico de datos viene a dar solución a las preguntas, ¿cómo almacenar los datos?, ¿cómo analizar y obtener valor de los datos? y ¿cómo visualizar y comunicar lo que nos quieren decir los datos?**

Los "trabajadores del conocimiento" que menciona Drucker, son los que ahora ya tienen un perfil más claro y se les conoce como *científicos de datos*, en ellos recae la responsabilidad de entender en su máxima expresión los datos y sus relaciones, con el objetivo de tomar decisiones más informadas a la vez que mejoran los productos y servicios de las organizaciones.

Davenport y Patil en su artículo "Data Scientist: The sexiest job of the 21st century" (2012), definieron por primera vez el concepto de científico de datos y con ello generaron una gran revolución. De acuerdo con las estadísticas obtenidas de scholar.google.com, el artículo ha sido citado 568 veces y se han producido 15 versiones diferentes. Además de definir quién es un científico de datos, presentan un decálogo para encontrar el científico de datos correcto, explican cuáles son los intereses del profesional y de los cuidados que deberán tener las empresas para conservarlos.

En términos generales, el científico de datos combina estadística, matemáticas, programación y solución de problemas, con la captura datos de forma ingeniosa y la capacidad de mirar las cosas de manera diferente (encontrar patrones), además de hacer las actividades propias de limpieza, preparación e integración de datos (Monnapa, 2017).

De acuerdo con la encuesta que realizó la compañía Crowd Flower (2017) a 179 científicos de datos seleccionados en todo el mundo, identificó la distribución de las actividades que les toma mayor tiempo en su quehacer, las cuales se distribuyen de la siguiente manera:

- 51% colectar, etiquetar, limpiar y organizar los datos.
- 19% construir y modelar los datos.
- 10% el modelado de datos para patrones.
- 9% refinar algoritmos.
- 8% otras actividades.

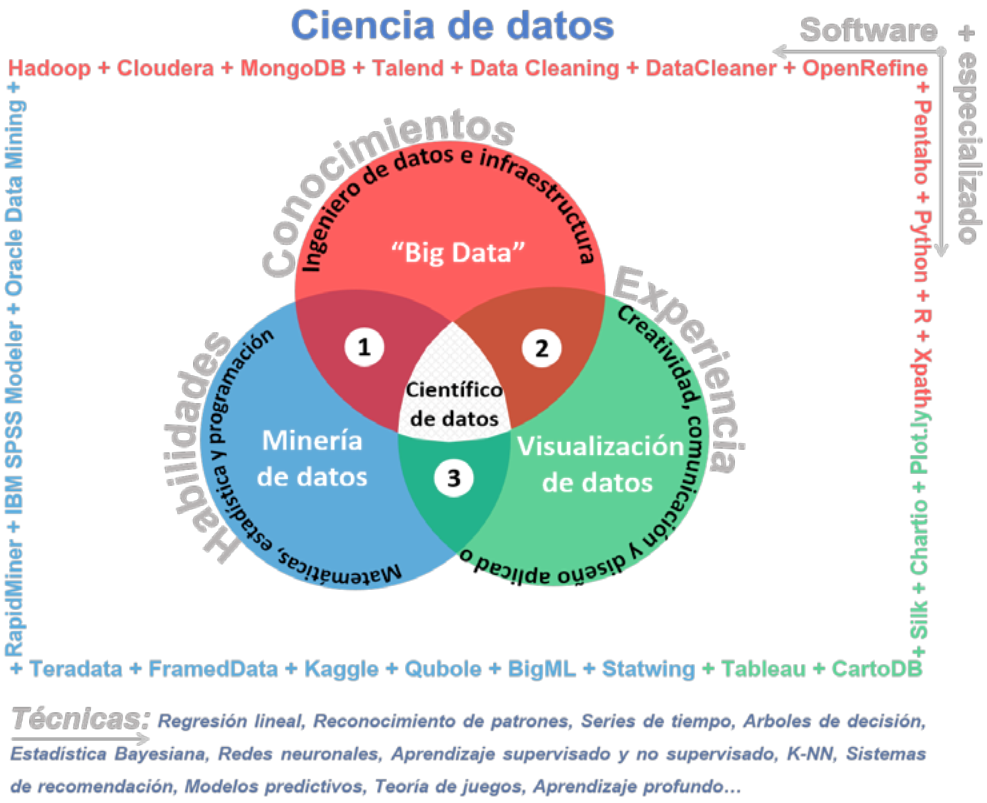

Figura 3.
El científico de datos y su entorno
Fuente:
elaboración propia.

Entre las actividades que más disfrutan, están: construir y modelar los datos, aplicar minería de datos para encontrar patrones y el refinar algoritmos. Entre las que menos gustan, están: limpiar y organizar datos, etiquetarlos y colectarlos. El 51% de los encuestados reportó que trabajan con datos no estructurados. Los datos con los que trabajan provienen principalmente de los sistemas internos de las compañías en las que trabajan, seguido de los que colectan de forma manual y, por último, de los conjuntos de datos disponibles públicamente.

Son tres las áreas en las que se desarrollan principalmente los científicos de datos:

- *Big Data* para procesar datos,
- Minería de datos para analizar e identificar relaciones ocultas, patrones y tendencias,
- Visualización de datos para explicar y socializar mejor la información obtenida.

También existe una amplia gama de técnicas y software especializado que el científico de datos utiliza para desarrollarse en cada una de las áreas, de las cuáles se pueden clasificar por herramientas de extracción, almacenamiento, limpieza, minería, visualización, programación, análisis e integración de datos. En la siguiente figura se presentan las habilidades, los conocimientos y la experiencia que debe poseer el científico de datos, así como una muestra del software especializado y técnicas que existe por sus áreas de desarrollo.

A continuación, se hace una breve descripción de cada una de las áreas e intersecciones que se muestran en la figura, a la vez que se mencionan algunos ejemplos de proyectos que realizan los científicos de datos.

Big Data

Los investigadores Cox y Ellsworth (1997) de la Administración Nacional de la Aeronáutica y del Espacio (NASA por sus siglas en inglés), fueron los primeros en utilizar el término *Big Data* en un artículo científico, en el que señalaron el problema al que se enfrentaban al visualizar y el procesar grandes cantidades datos, así como las limitantes técnicas de las computadoras (gráficos, memorias y almacenamiento) que tenían en esos tiempos.

Ha sido un término que, al igual que su nombre, ha tenido una gran aceptación en todas las industrias y son muchas las definiciones que existen al respecto,[4]

4. La escuela de información de Berkeley, reunió 40 definiciones provenientes de los líderes en las diferentes industrias, ver en: https://goo.gl/P69x3v.

en particular me gusta definir el concepto de *Big Data* como: el gran cúmulo de datos compuesto por diferentes tipos, estructuras y relaciones de datos, que a su vez tienen veloces tasas de generación y dispersión, y el procesarlos con tecnologías convencionales para su posterior análisis es parte del problema (*Big Problem*).

De acuerdo con Van der Aalst (2014), utiliza el término *Internet of Events* (IoE) para referir a todos los datos disponibles en Internet. Y los clasifica de la siguiente manera:

- ***Internet of the Content*** (IoC). Es toda la información creada por los seres humanos para aumentar el conocimiento sobre temas particulares. Incluye páginas web tradicionales, artículos, enciclopedias como Wikipedia, YouTube, libros electrónicos, noticias, etcétera.
- ***Internet of the People*** (IoP). Son todos los datos relacionados con la interacción social. Es decir, correo electrónico, Facebook, Twitter, foros, LinkedIn, etcétera.
- ***Internet of the Things*** (IoT). Son todos los objetos físicos conectados a la red. Son todas las cosas que tienen una identificación única y una presencia en una estructura similar a Internet. Las cosas pueden tener una conexión a Internet o estar etiquetados usando identificación por radio frecuencia (RFID por sus siglas en inglés), proximidad a campos de comunicación (NFC por sus siglas en inglés), etcétera.
- ***Internet of the Locations*** (IoL). Refiere a todos los datos que tienen una dimensión espacial. Con la adopción de dispositivos móviles (por ejemplo, teléfonos inteligentes) cada vez más eventos tienen atributos geoespaciales.

Es importante contar con este marco de referencia, ya que es una buena forma para clasificar la generación del *Big Data* por tipos de eventos.

Minería de datos

En términos sumamente prácticos la minería de datos la podemos definir como a la extracción de conocimientos de grandes cantidades de datos. Han y Kamber (2006), hacen una interesante crítica al concepto de minería de datos: "la extracción de oro de las rocas o la arena se conoce como minería de oro en lugar de minería de roca o arena. Por lo tanto, la minería de datos debería haber sido más apropiadamente llamada minería del conocimiento a partir de datos. Sin embargo, la minería es un término vívido que caracteriza al proceso

de encontrar un pequeño conjunto de preciosas pepitas en una gran cantidad de materia prima" (p. 5).

La minería de datos tiene dos referentes principales. Al primero se le conoce como el proceso de descubrimiento de conocimiento en bases de datos (mejor conocido por sus siglas en inglés, KDD), que fue promovido en 1989 por iniciativa de Shapiro y Smyth (1996) y está definido por cinco etapas (selección, pre-procesamiento, transformación, minería de datos e interpretación/evaluación). El segundo es el proceso estándar de la industria para la minería de datos (mejor conocido por sus siglas en inglés, CRISP-DM), el cual fue concebido en 1996 y define seis fases en su proceso (comprensión del negocio, comprensión de datos, preparación de datos, modelado, evaluación e implementación) (Wirth y Hipp, 2000).

La minería de datos en principio trabaja sobre todo tipo de datos. Los estructurados se refieren a las bases de datos relacionales (filas y columnas claramente identificadas); los semiestructurados son los que tienen un tipo de estructura implícita, pero no como para ser automatizada como la estructurada (datos espaciales, temporales y textuales); los no estructurados son los que principalmente provienen de sitios en Internet y son del tipo multimedia (imágenes, audio y videos). Los dos últimos se identifican con la minería de textos y la minería web, respectivamente.

Entre las principales técnicas que se utilizan en la minería de datos, están: regresión lineal, estimación de densidad, reconocimiento de patrones, series de tiempo, árboles de decisión, estadística Bayesiana, redes neuronales, aprendizaje supervisado y no supervisado, vecinos más próximos (K-NN), sistemas de recomendación, modelos predictivos, teoría de juegos, aprendizaje profundo, entre otros más. Para mayor información de cada una de las técnicas, se sugiere revisar a Granville (2016).

Visualización de datos

La visualización de datos es considerada por algunos como una ciencia y hay quienes la clasifican como un arte, cuando en realidad es una combinación de ambas. Sus principales precursores justo provienen de las ciencias exactas, que han tenido la necesidad de recurrir al campo de la creatividad y del arte, con el propósito de representar con fines estéticos algún aspecto de la realidad.

La visualización de datos sólo tendrá éxito en la medida que nuestros ojos codifiquen la información para poder discernirla y nuestros cerebros la pueden entender. El objetivo es traducir de maneras fáciles, eficientes, precisas y decodificadas la información abstracta en representaciones visuales significativas (Few, 2013).

La visualización de datos ayuda al usuario a examinar una gran cantidad de datos e identificar patrones o tendencias con la ayuda de gráficas o representaciones. Una sola gráfica puede codificar mucha más información que la que se puede presentar en varias hojas de texto (Pujari, 2001, p. 48).

La visualización de datos que ahora conocemos ha sido desarrollada a lo largo de la humanidad, siempre ha existido la necesidad de abstraer y comunicar información. Desde siglo II d. C. se han organizado los datos en tablas (columnas y filas), pero la idea de representar gráficamente la información cuantitativa surgió hasta el siglo XVII, cuando el filósofo y matemático francés René Descartes desarrolló un sistema de coordenadas bidimensional para mostrar valores. A finales del siglo XVIII, el ingeniero y economista William Playfair encontró el potencial de los gráficos para la comunicación de datos cuantitativos, definió muchos de los gráficos que se utilizan actualmente (barras y líneas en función del tiempo), incluso inventó el gráfico circular (pastel). Cabe señalar que este tipo de gráfico ha sido objeto de muchas críticas por parte de los especialistas en el área de visualización de datos y percepción. Por ejemplo, Few (2013) ha demostrado que es ineficaz, ya que codifica los valores como atributos visuales (áreas y ángulos), lo que impide percibir y comparar fácilmente.

El trabajo del cartógrafo Jacques Bertin fue fundamental, pues descubrió que la percepción visual opera según reglas que se pueden seguir para expresar visualmente la información de manera intuitiva, clara, precisa y eficiente. El profesor de estadística en Princeton, John W. Tukey,[5] dio forma a un nuevo enfoque estadístico llamado análisis exploratorio de datos, y fue quien realmente introdujo el poder de la visualización de datos como un medio para explorar y dar sentido a los datos cuantitativos. El estadístico y artista Edward R. Tufte publicó en 1983 el libro *The Visual Display of Quantitative Information*, mismo que revolucionó las formas eficaces de mostrar los datos visualmente. El matemático William S. Cleveland con la publicación de sus libros *The Elements of Graphing Data y Visualizing Data* hizo grandes aportaciones en cuanto a las técnicas que utilizan los estadísticos para la visualización de datos. En 1999, los investigadores Stuart Card, Jock Mackinlay y Ben Shneiderman acuñaron una nueva especialidad "visualización de la información" y publicaron el libro *Information Visualization, Using Vision to Think*, en el que recopilan mucho del trabajo académico que se había realizado hasta ese momento, y a nuestros días es uno de los principales referentes de la visualización de datos e información (Few, 2013).

5. Considerado como uno de los investigadores más importantes de la estadística moderna.

Actualmente el término visualización de datos es el más aceptado entre la comunidad científica de datos,[6] pero en lo particular prefiero el término visualización de la información, ya que las actuales visualizaciones se hacen en al menos dos dimensiones con múltiples atributos y relaciones, por lo que no podemos estar hablando de visualización de datos, ya que de ser así se limitaría a mostrar numerosos elementos en una sola dimensión.

6. Comparativo entre "data visualization" e "information visualization", en Google Trends: https://goo.gl/1xoNHi.

Relación 1. Big Data y minería de datos

Debe existir una estrecha relación entre ambas áreas, ya que los algoritmos y modelos de entrenamiento y prueba desarrollados por el área de minería da datos deberán ser implementados en los grandes cúmulos de datos (*Big Data*), sobre todo cuando se tiene una amplia serie de datos históricos. Un ejemplo de lo anterior, es el artículo: "Applying Data Mining Techniques to Identify Success Factors in Students Enrolled in Distance Learning: A Case Study" (Moreno y Stephens, 2015). El artículo analiza, con técnicas de minería de datos, los perfiles de ingreso, antecedentes académicos y matrícula de los alumnos del Sistema Universidad Abierta y Educación a Distancia (SUAyED), de la Universidad Nacional Autónoma de México (UNAM), con el propósito de determinar los factores clave que impulsan el éxito y el fracaso de los alumnos, así como la creación de su respectivo modelo predictivo usando el algoritmo de clasificación Naive Bayes.

Relación 2. Big Data y visualización de datos

La relación entre el área de *Big Data* y la visualización de datos es la que busca definir la mejor interpretación y visualización de grandes cúmulos de datos y sus relaciones, de forma que al usuario le resulte más fácil entenderlos. En la mayoría de los casos se queda en la descripción de los datos en diferentes dimensiones y en algunos, incluyen interactividad a sus relaciones. A continuación, se presenta una muestra de ejemplos:

- *Referencias cruzadas de la biblia* (Harrison, 2007). Es un proyecto en el que se analiza, en la parte inferior de la visualización, todos los capítulos de la biblia en un gráfico de barras, los cuales alternan entre los colores blanco y gris claro. La longitud de cada barra denota el número de versos en el capítulo. Cada una de las 63 779 referencias encontradas en la biblia está representada por un solo arco y el color corresponde a la distancia entre los dos capítulos, creando un efecto similar al arco iris.
- *Temperaturas del clima* (Carli, 2013). Al superponer datos meteorológicos históricos, ésta visualización muestra cómo la temperatura evoluciona durante el año en diferentes ciudades. Al establecer una banda de zona de confort, es posible ver cuando la temperatura está por encima, dentro o debajo de la zona; tanto a lo largo del año como a través de los días de cada temporada. Es un claro ejemplo de una visualización interactiva.
- *Población mundial* (Carli, 2014). Es un proyecto de visualización interactiva realizado para el Banco Mundial donde cada usuario al digitar su fecha de nacimiento puede compararse con los datos de la población

mundial (7.2 mil millones de personas al 2014), y saber cuántas personas nacieron el mismo día y a la misma hora. Al final presenta de manera ordenada la posición por la edad que tenga el usuario y la compara con la población mundial.

Relación 3. Minería de datos y visualización de datos

La minería y la visualización de datos también pueden trabajar en una dimensión donde no necesariamente se procesen un gran cúmulo de datos (*Big Data*), se pueden implementar proyectos para una cantidad mesurada de datos, donde se apliquen algoritmos y con éstos obtengamos un producto. A continuación, se muestran algunos ejemplos:

- *FLEET: Distribución e Inventario de Unidades* (Moreno, 2017a). Es un proyecto realizado para una empresa especializada al arrendamiento de vehículos, en el que se analiza el número de unidades que ha comprado por estado, municipio y concesionaria (agencia). Además, la visualización permite seleccionar los modelos de los vehículos, por segmento y rango de precios. Cabe señalar que todos los datos están relacionados y permiten la interacción.
- *Gasolid: Identificador de Gasolineras en México* (Moreno, 2017b). Es un proyecto que permite medir el nivel de confianza en las gasolineras mexicanas. Para lograrlo, se relacionaron bases de datos de la Procuraduría Federal del Consumidor (PROFECO) y Petróleos Mexicanos (PEMEX), además de correr procesos de geolocalización para identificar las direcciones de las estaciones de servicio. De acuerdo a la selección aplicada (estado, municipio y código de la estación de servicio), es posible ubicar a las gasolineras e identificar en nivel de confianza en éstas, así como el historial que ha tenido en los últimos cuatro años sobre el número de mangueras verificadas e inmovilizadas.

Relación 1, 2 y 3. Big Data, minería y visualización de datos

El científico de datos puede interactuar en cualquiera de las tres áreas (***Big Data***, minería y visualización de datos) y en sus respectivas intersecciones (1, 2 y 3), siempre y cuando posea las habilidades, conocimientos, técnicas y experiencia para lograr dar respuesta a las necesidades de las organizaciones en términos de codificar mejor sus datos.

Drinking Data (Ratti, 2014). Es un proyecto que integra todas las áreas de la ciencia de datos y sus intersecciones, fue desarrollado por investigadores del

Instituto Tecnológico de Massachusetts (MIT por sus siglas en inglés). La pregunta que detonó el desarrollo del proyecto fue: ¿podemos encontrar algún valor en la enorme cantidad de información registrada por las máquinas dispensadoras de refrescos enlatados? Considerando que en los Estados Unidos existen 15 000 máquinas dispensadoras y que cada una puede despachar 150 latas únicas. En cada transacción se registra una cadena de datos: hora, ubicación y preferencias del usuario (*Big Data*). Con éstos datos se visualizaron patrones de consumo total y se encontraron altos consumos de bebidas los fines de semana y lo contrario, durante los días entre semana (Visualización de datos). Con toda la explotación de los datos pudieron entender mejor lo que sucede: el analizar los dispensadores individualmente mostró características inesperadas (Minería de datos).

Conclusiones

El inicio de la era Web 2.0 fue un momento crucial para las organizaciones que supieron beneficiarse de los datos que dejaban registrados los usuarios en sus plataformas (huella digital), por lo que hubo un incremento en la oferta de aplicaciones web —y sus respectivas versiones para teléfonos inteligentes—, ya que reconocieron el valor oculto e intangible que podían codificar de los datos de sus usuarios, tan sólo hay que recordar: "si no pagas por el producto, tú eres el producto" (Van der Aalst, 2014, p. 19).

El sitio statista.com[7] reportó, a marzo del 2017, que Google Play y Apple Store, administran 2.8 y 2.2 millones de aplicaciones, respectivamente. Por lo anterior, es que los científicos de datos se han convertido en especialistas de gran importancia para cualquier industria, ya que ofrecen servicios y productos hechos a la medida para cada uno de sus usuarios. La advertencia es: las compañías que no utilicen sus datos inteligentemente no serán competitivas y en el peor de los casos, no sobrevivirán.

7. Número de aplicaciones para teléfonos inteligentes, ver en: https://goo.gl/tCnPXW

Los científicos de datos se pueden especializar en cualquiera de las tres principales áreas (*Big Data*, minería de datos y visualización de datos), lo que ha propiciado que todas las industrias tengan una alta demanda por especialistas en estas áreas. IBM (2017) recientemente reportó que para el 2020 el número de empleos para todos los profesionales de datos en EUA aumentará en 364 mil, alcanzando un total de 2.7 millones de empleos.

El científico de datos viene a dar solución a las preguntas, ¿cómo almacenar los datos?, ¿cómo analizar y obtener valor de los datos? y ¿cómo visualizar y comunicar lo que nos quieren decir los datos? Todo lo anterior en términos de eficacia, eficiencia y veracidad. Por otro lado, Van der Aalst (2014), sugiere que el científico de datos responderá a las preguntas ¿qué paso?, ¿por qué paso?, ¿qué sucederá?, ¿qué es lo mejor que puede pasar?, las cuales, en mi opinión, sólo corresponden a el área de la minería de datos.

El *Big Data* ayuda a dar solución al gran problema que existe respecto al volumen, la variabilidad y la velocidad de los datos. El propósito de la minería de datos es pasar del volumen de datos hacia la información, para después el conocimiento y, por último, llegar al valor de la decisión.

Concluyo que debemos tener mucho cuidado con las diferentes visualizaciones de datos que ofrecen los diferentes softwares especializados, ya que existe el riesgo de hacer representaciones ineficaces de la información e incluso, nos podemos dejar llevar por gráficos muy elaborados presentados en diferentes dimensiones y que a simple vista parecen ser muy atractivos, cuando en la realidad dejan de lado la exploración y comunicación útil de la información y sólo presentan una estética superficial. La visualización no tiene límites, se puede representar cualquier tipo de dato con sus atributos, elementos y relaciones, todo dependerá de la creatividad para abstraer dicha información y presentarla de formas más inteligentes.

Referencias

- Butler, P. (2010). *Visualizing Friendships*. Facebook: Facebook Enginering. Recuperado de https://goo.gl/AaHLN.
- Cakmak, J. (2017). What Is Trump Worth to Twitter? One Analyst Estimates $2 Billion. *Fortune.com*. Recuperado de https://goo.gl/AdjDQz.
- Carli, L. (2013). *Weather Temperatures*. Vis.design. Recuperado de https://goo.gl/sgDNsS.
- Carli, L. (2014). *The World Population Project*. Vis.design. Recuperado de https://goo.gl/jSgpEx.
- Cox, M. y Ellsworth, D. (1997). *Managing Big Data for Scientific Visualization*. ResearchGate.net. Recuperado de https://goo.gl/DLj8sd
- Crowd Flower (2017). *Data Scientist Report 2017*. Crowdflower.com. Recuperado de https://goo.gl/4XsUKD
- Davenport, T. H. y Patil, D. J. (2012). Data Scientist: The Sexiest Job of the 21st Century**.** *Harvard Business Review*, 90 (10). Recuperado de https://goo.gl/65lMw1
- Drucker, P. F. (2004). *La sociedad postcapitalista*. Medellín, Colombia: Norma
- Elahi, E. (2015). *Spark and GraphX in the Netflix Recommender System*. SlideShare. Recuperado de https://goo.gl/LUQqx4.
- Few, S. (2013). Data Visualization for Human Perception. En M. Soegaard (2nd Ed.), *The Encyclopedia of Human-Computer Interaction*. Aarhus, Denmark: The Interaction Design Foundation. Recuperado de: https://goo.gl/7uYrrp.
- Granville, V. (2016). *40 Techniques Used by Data Scientists*. datasciencecentral.com. Recuperado de https://goo.gl/UmCb9M.

- Han, J. y Kamber, M. (2006). *Data Mining Concepts and Techniques*. Recuperado de https://goo.gl/jmXNua.
- Harrison, C. (2007). *Bible Cross-References*. Chrisharrison.net. Recuperado de https://goo.gl/21DsY.
- Hilbert, M. y Lopez, P. (2011). The world's technological capacity to store, communicate, and compute information. *Science*, 332(6025), 60–65.
- IBM (2017). *The Cuant Crunch: How the Demand for Data Science Skills is Disrupting the Job Market*. Ibm.com. Recuperado de https://goo.gl/yJ6GtL.
- Monnappa, A. (2017). *Data Science vs. Big Data vs. Data Analytics.* Simplilearn.com. Recuperado de https://goo.gl/EAYQRc.
- Moreno, G. S., Stephens, C. R. (2015). Applying Data Mining Techniques to Identify Success Factors in Students Enrolled in Distance Learning: A Case Study. *Advances in Artificial Intelligence and Its Applications*: Springer. Recuperado de https://goo.gl/zFLHtJ.
- Moreno, G. S. (2017a). *FLEET: Distribución e Inventario de Unidades*. Publictableau.com. Recuperado de https://goo.gl/Db5eXr.
- Moreno, G. S. (2017b). *Gasolid: Identificador de Gasolineras en México*. Publictableau.com. Recuperado de https://goo.gl/26iX9C.
- Pujari, A. K. (2001). *Data Mining Techniques*. Hyderabad, India: Universities Press.
- Ratti, C. (2014). *Drinking Data.* senseable.mit.edu: MIT Senseable City Lab. Recuperado de https://goo.gl/QkJ4iw.
- Shapiro, G. P., Smyth P. (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Communications of the ACM*. Volumen (39, 11), 27-34. Recuperado de https://goo.gl/L67PYa.
- Van der Aalst, W. M. P. (2014). Data Scientist: The Engineer of the Future. *Enterprise Interoperability*, volume 7, 13-28. Springer. Recuperado de https://goo.gl/yiaE9F.
- Vanderbilt, T. (2013). The Science Behind the Netflix Algorithms That Decide What You'll Watch Next. *Wired*. Recuperado de https://goo.gl/aqJqA7.
- Wirth, R., Hipp, J. (2000). *CRISP-DM: Towards a Standard Process Model for Data Mining*. ResearchGate.com. Recuperado de https://goo.gl/XevGEB

Cómo citar este artículo

- Moreno Salinas, José Gerardo (2017). Científico de datos: codificando el valor oculto e intangible de los datos, *Revista Digital Universitaria (RDU)*, vol. 18, núm. 7, septiembre-octubre. DOI: http://dx.doi.org/10.22201/codeic.16076079e.2017.v18n7.a2.