big data

breve manual para conocer la ciencia de datos que ya invadió nuestras vidas

walter sosa escudero

siglo veintiuno
editores

grupo editorial
siglo veintiuno

siglo xxi editores, méxico
CERRO DEL AGUA 248, ROMERO DE TERREROS, 04310 MÉXICO, DF
www.sigloxxieditores.com.mx

siglo xxi editores, argentina
GUATEMALA 4824, C1425BUP, BUENOS AIRES, ARGENTINA
www.sigloxxieditores.com.ar

anthropos
LEPANT 241, 243 08013 BARCELONA, ESPAÑA
www.anthropos-editorial.com

Sosa Escudero, Walter
Big data / Walter Sosa Escudero.- 1ª ed.- Ciudad Autónoma de Buenos Aires: Siglo XXI Editores Argentina, 2019.
208 p.; 21x14 cm.- (Ciencia que ladra… serie Mayor / dirigida por Golombek, Diego)

ISBN 978-987-629-899-5

1. Estadísticas. 2. Ciencia de la Información. I. Título.
CDD 519.5

Diseño de cubierta: Pablo Font

ISBN 978-987-629-899-5

Impreso en Master Graf SA // Mariano Moreno 4794, Munro
en el mes de mayo de 2019

Hecho el depósito que marca la Ley 11 723
Impreso en Argentina // Made in Argentina

Índice

Este libro (y esta colección)

Cubriéndonos, cegándonos, matándonos /
desde las mesas, desde los bolsillos, /
los números, los números, / los números.
Pablo Neruda, "Una mano hizo el número"

Si viene la lluvia, / ellos corren y esconden
sus cabezas.
Los Beatles, "Rain"

Hay conceptos que duran un día, y pueden ser buenos. Hay otros que están de moda, y no sabemos qué son. Y hay, claro, los que duran toda la vida, los que son imprescindibles, los que nos cuentan de tal manera que se nos enciende un "ajá" en el cerebro y de pronto la vida cambia. Entre estos, seguro escucharon hablar de "big data", grandes datos, datos masivos, datos hasta en la sopa. Llueven datos y no siempre tenemos las cucharas para recibirlos y degustarlos.

Vamos a los datos, a los números, entonces. Según un estudio de la consultora Cumulus Media, en un minuto de internet 900 000 personas se conectan a Facebook, 3,5 millones de usuarios realizan búsquedas en Google, se envían 452 000 tuits, se reproducen 4,1 millones de horas de video en YouTube, se miran 70 000 horas de contenido de Netflix y se suben unas 46 200 fotos a

Instagram. Sí, en un minuto de internet. Esto, por supuesto, genera una cantidad de información inusitada, inaudita… imposible. Pero a estas tres "I" se les oponen las tres "V" de esta nueva ciencia de los datos: volumen, velocidad y variedad. En otras palabras: a grandes datos, grandes métodos para analizarlos y grandes memorias para guardarlos. La cantidad de información da tortícolis: se dice que un exabyte alcanzaría para registrar todas las palabras pronunciadas por todos los humanos que hayan existido. Más aún: la mayor parte de esta catarata de datos se crea porque sí, por generación espontánea, cada vez que hacemos algo que involucra una transacción, registro o aparatito digital. En el medio, predicciones de epidemias o cambios climáticos, datos sociales y hombres de la bolsa.

Entre tal maraña lo más obvio (quizá hasta lo indicado) es perderse, como Tony y Douglas en el *El túnel del tiempo* (*millennials* abstenerse) o Neo dentro de la Matrix. Pero cual mago del orden en nuestras cajoneras, por fortuna aparece el mejor guía de este infierno encantador: el inigualable Walter Sosa Escudero nos lleva de la mano entre números y estadísticas, entre algoritmos y computadoras que aprenden sobre nosotros. Pero este no es solo un libro de datos; como no podía ser de otra manera tratándose de Walter, es además un libro de *rock and roll*. Por sus páginas viajamos de gobiernos abiertos a Elvis y Bill Haley, de la gran epidemia de gripe A (y sus huellas digitales) a Jimi Hendrix y Eric Clapton. Aunque hay para todos los gustos: también tenemos historias de inteligencia artificial regadas por Air Supply, A-ha o Rubén Blades.

Es que en esta nueva ciencia de datos (de muchos datos) entra todo. El análisis de la personalidad extraído de una minuciosa búsqueda de millones de usuarios en

Twitter. Mapas detallados del cerebro basados en los billones de conexiones de las neuronas. Planos del comportamiento criminal en las grandes ciudades (que ayudan a combatir y reducir esos crímenes de manera que, por una vez, la caballería ya no llegue tarde). Manejo de crisis y catástrofes naturales sobre la base de la información que se genera "sola" cuando millones de personas comparten opiniones y anuncios. Y, en el medio, nosotros, hormiguitas en el mundo de los datos tratando de encontrarle algún sentido a esta inundación que amenaza con taparnos los ojos y marearnos el futuro.

Pero no: Walter lo logra, una vez más, y nos rescata justo a tiempo para entender, nada menos, dónde estamos, adónde vamos y, quizá, adónde queremos ir. Llueven datos, sí, pero en estas páginas están las cucharas, los paraguas y las plantas para aprovechar la lluvia.

Esta colección de divulgación científica está escrita por científicos que creen que ya es hora de asomar la cabeza por fuera del laboratorio y contar las maravillas, grandezas y miserias de la profesión. Porque de eso se trata: de contar, de compartir un saber que, si sigue encerrado, puede volverse inútil.

Ciencia que ladra… no muerde, solo da señales de que cabalga.

Diego Golombek

A mi esposa Mercedes, fuente inagotable de energía y generosidad

Agradecimientos

Mi principal agradecimiento es para Sebastián Campanario, periodista, economista y divulgador de la tecnología y la creatividad. Valoro su permanente voto de confianza, y que haya visto en mí la dimensión de divulgador que mantuve latente durante muchos años. En particular, le agradezco el espacio que con frecuencia me brinda en "Alter eco", su notable columna en *La Nación*, donde aparecieron publicadas algunas de las historias que dieron origen a este proyecto.

El ámbito de los datos es marcadamente multidisciplinar. Una gratísima sorpresa es haber encontrado un clima amistoso y cooperativo en este ambiente tan diverso. A todos les estoy muy agradecido, sin involucrarlos en ninguno de los desaciertos que este libro pueda tener y son de mi exclusiva responsabilidad.

Ernesto Mislej y Manuel Aristarán me proveyeron información muy provechosa y una eterna palabra de aliento. María Edo, Leonardo Gasparini, Marcela Svarc, Mercedes Iacoviello, Mariana Marchionni, Javier Alejo, Ignacio Sarmiento, Luján Stasevicius, Ricardo Bebczuk, Andrés Ham y Laura Ación leyeron todo o parte del texto original y me hicieron llegar útiles comentarios. Noelia Romero aportó su pericia y entusiasmo en varias etapas de la elaboración de este trabajo. Marina Navarro leyó el manuscrito y me hizo valiosas sugerencias de estilo.

María Sagardoy me asistió en cuestiones de diseño gráfico. Diego Pando, Eugenia Mitchelstein, Federico Bayle, Fernando Zerboni y Edmundo Szterenlicht aportaron material relevante para varias de las historias que incorporé en los capítulos. Mariel Romani y Moira Guppy, de la biblioteca de la Universidad de San Andrés (UdeSA), encontraron todos mis exóticos pedidos bibliográficos con asombrosa eficiencia. UdeSA apoyó enfáticamente mis actividades de divulgación, en particular el proceso de elaboración de este libro. Agradezco también a todos mis alumnos del curso de Big Data y Aprendizaje Automático de UdeSA, que ha sido la contraparte técnica y docente de esta obra.

En Siglo XXI Editores, Marisa García Fernández hizo una gran tarea de edición que contribuyó a mejorar sustancialmente este libro. En especial, agradezco a Carlos Díaz y a Diego Golombek por confiar en mí y por cuidar celosamente la colección Ciencia que Ladra.

Buenos Aires, noviembre de 2018

Introducción acuífera

–Buen día, pase y tome asiento. ¿Cómo le va? Cuénteme, ¿qué lo trae por aquí?
–¡Doctor, veo datos por todas partes! Que si doy "me gusta" a una foto en Facebook, que si busco una dirección en Google, que no sé cuántos kilómetros corrió un jugador de fútbol en el último partido, que si volví en tren en vez de volver en auto, ¡datos, datos, datos y más datos!
–Tranquilícese, esta cuestión de la ciencia de datos, los algoritmos, las computadoras y las estadísticas se nos ha ido de las manos a todos.
–Ah, ¿mal de muchos, consuelo de tontos? ¿No le parece otro argumento estadístico? ¡¡Ayúdeme!!

Llueven datos. De redes sociales, tarjetas de crédito, teléfonos celulares, páginas web y sus buscadores, dispositivos de GPS, relojes inteligentes, rastreadores satelitales, análisis clínicos, cámaras de fotos y cualquier otro objeto interconectado electrónicamente. Y ante tanta lluvia, las reacciones son dispares. Hay quienes buscan guarecerse; algunos quieren recoger el agua con una cuchara, mientras otros piensan en enormes tanques; algunos, en extraños dispositivos para transformarla en otra cosa y otros simplemente no hacen nada, fieles a eso

de que "siempre que llovió, paró". Desde la perspectiva de los datos, las cucharas, los pilotos, los paraguas, los contenedores y los procesadores químicos de lluvia son las técnicas utilizadas para analizarlos y convertir este diluvio en conocimiento relevante, y juegan un rol tan importante o más que la información.

Este libro ofrece un paseo guiado por el aguacero de datos y algoritmos. No presupone ninguna formación técnica, tan solo la curiosidad por saber qué promete esta batalla de información, fórmulas y computadoras, que unos ven como el comienzo de una nueva era y otros como una moda pasajera. Al respecto, adoptaremos una postura optimista y a la vez sincera: destacaremos tanto el enorme potencial de esta tormenta de datos como sus dificultades. Mojaremos nuestros pies en el mar de big data, surfearemos sus olas con innovadores algoritmos y navegaremos a bordo del poderoso buque de la estadística. Además de presentar un muestrario de casos, los invitamos a pensar acerca de si esta catarata de información será capaz de cambiar radicalmente nuestra forma de ver el mundo y cómo convivirá con los métodos tradicionales de la ciencia.

Terminaremos el recorrido empapados de historias de exitosos navegantes de datos y también de naufragios épicos. Regresaremos chorreando aprendizaje automático, petas, yottas, clasificación, regularización, Python, R, validación cruzada, árboles, redes neuronales, clústers y otros esoterismos de la jerga de los valientes marineros de la información. Escucharemos las historias de los jóvenes científicos de datos trepados a sus veloces motos de agua, y nos deleitaremos con las anécdotas de los capitanes de la estadística, aferrados al timón de sus navíos.

El plan de acción es el siguiente. El capítulo 1 arranca con una breve "ducha" en la que intentamos acla-

rar qué es esto de big data y los algoritmos, y qué rol juega la estadística en esta historia. El crucero sigue, en el capítulo 2, con algunas experiencias de análisis de datos en la sociedad moderna. El capítulo 3 es una primera "clase de natación" sobre algoritmos y métodos. Habiendo aprendido algunas maniobras básicas, el capítulo 4 invita a recorrer nuevas historias de datos con más detalle, como si fuésemos a nadar con los delfines que en el capítulo 2 apenas veíamos desde la cubierta. El capítulo 5 indaga en las técnicas más recientes de aprendizaje. Luego, para ir secándonos de tanto remojón, tomaremos distancia y discutiremos las limitaciones del análisis de datos en la sociedad moderna, en el capítulo 6. El capítulo 7 reflexiona sobre el futuro de los datos y los algoritmos. Cada capítulo comienza con un breve diálogo que tal vez remita al lector a *Karate Kid*, *Kung Fu* o a las sesiones semanales con su analista, y concluye con una breve sección titulada "Da capo" (una instrucción musical que indica al interprete volver al principio de la partitura) que ofrece alguna reflexión a modo de resumen.

Una aclaración: evité tercamente las notas al pie y las citas bibliográficas que usualmente pueblan los libros de texto y los *papers* científicos para no interrumpir el flujo de la lectura. El apéndice contiene todas las referencias y fuentes utilizadas en este libro.

"¿Quién va a parar la lluvia?" cantaba John Fogerty en los sesenta. La lluvia de big data, parece que nadie, por eso los invito a unirse a esta humilde arca de Noé.