Garbero, M. N. (2020). Nota crítica de libro. *Big data. Breve manual para conocer la ciencia de datos que ya invadió nuestras vidas*, de Walter Sosa Escudero. *Cuyonomics. Investigaciones en Economía Regional*, 4(6), 83-89. https://revistas.uncu.edu.ar/ojs3/index.php/cuyonomics/article/view/4327

Nota crítica de libro

Big data. Breve manual para conocer la ciencia de datos que ya invadió nuestras vidas

Walter Sosa Escudero

Siglo XXI Editores
2019
208 páginas

Por:

María Noelia Garbero
Universidad de Mendoza y Universidad del Aconcagua
Mendoza, Argentina
noeliagarbero@gmail.com

Este libro vino a traer un poco de luz al mundo revolucionario del *big data*, tanto para aquellos que algo conocemos del universo de la estadística y los datos como para quienes comienzan a escuchar este lenguaje por primera vez. No llegó solo, sino de la mano del prestigioso economista Walter Sosa Escudero.

El autor es un experto en estadística, estudió Economía en la Universidad de Buenos Aires y se especializó en Econometría. Obtuvo su doctorado en la Universidad de Illinois, en Urbana-Champaign, Estados Unidos. También es docente e investigador de tiempo completo en la Universidad de San Andrés e investigador y profesor invitado en la Universidad Nacional de La Plata –donde tuve la suerte de tenerlo como profesor–. Complementa sus actividades como investigador del CONICET.

Dentro de sus obras de divulgación científica, podemos encontrar los libros *Qué es (y qué no es) la Estadística*, de la colección Ciencia que ladra (Siglo XXI Editores, Buenos Aires), y *Pobreza y desigualdad en América Latina*, que escribió junto a los reconocidos economistas Leonardo Gasparini y Martin Cicowiez (Temas, Buenos Aires).

¿Pero qué es *big data*? ¿De dónde surge? ¿Cómo está revolucionando a la ciencia del análisis de los datos? ¿Cuál es su futuro? ¿Llegó para condenar a la estadística tradicional? El libro responde estas y otras preguntas más. Con un lenguaje coloquial, y a la vez con rigurosidad científica, el autor nos introduce al revolucionario campo de este concepto novedoso. En palabras del Walter Sosa Escudero, *big data* se refiere al volumen y tipo de datos provenientes de la interacción con dispositivos interconectados, como teléfonos celulares, tarjetas de crédito, cajeros automáticos, relojes inteligentes, computadoras personales, dispositivos de GPS y cualquier objeto capaz de producir información y enviarla electrónicamente a otra parte.

Es fácil darse cuenta de que cada uno de nosotros creamos este *big data* día a día, en cada oportunidad en la que abrimos una página web, buscamos algo, realizamos una compra, usamos nuestra tarjeta de crédito, visitamos nuestra red social favorita para poner un *like* a esa foto, película o comentario que nos agradó, comentamos una nota del diario, utilizamos el GPS para llegar a destino o simplemente usamos internet. Con estas acciones estamos generando información acerca de nuestro comportamiento. A nivel agregado, el volumen de información que se genera diariamente toma una gran dimensión y abarca diversos aspectos.

En el libro se concilian dos posturas extremas: los "talibanes de los datos", que creen que *big data* vino para remplazar todo tipo de conocimiento y hablan de una especie de muerte anunciada de la estadística, y los "escépticos", que creen que es una moda pasajera. El autor nos guía en este nuevo camino de los datos masivos,

presentándonos los algoritmos y las técnicas estadísticas y computacionales que permiten procesarlos y hacerlos capaces de brindar información realmente útil. Nos permite recorrer ese camino de una manera amigable para el lector no experto, mediante ejemplos cotidianos o el repaso de investigaciones de colegas y ex alumnos, introduciendo conceptos formales con el fin de mostrarnos que solo datos, por muchos (o *big*) que sean, no son suficientes. Obviamente, y fiel a Sosa Escudero, todo esto está hecho con un poco de *rock and roll*.

La publicación se encuentra dividida en siete capítulos. A lo largo de esta reseña, abordaré las principales ideas de cada uno y haré mención a los tópicos que, en mi opinión, resultan ser los más relevantes.

Con el primer capítulo empezamos a entender de qué se habla cuando se habla de *big data*, aunque en él se recalca la dificultad de intentar dar una definición formal a dicho concepto. De forma anecdótica, el autor nos cuenta que, en el 2001, Doug Laney, analista de la consultora Gartner, escribió un influyente artículo en el que resumió esta discusión diciendo que la revolución del *big data* tenía que ver con las "tres v de *big data*": volumen, velocidad y variedad, conceptos a los que posteriormente se les agregó una cuarta v por veracidad y que luego se trasformaron en "las 42 v de *big data*", según un artículo de Tom Shafer.[1] Esta historia es suficiente para mostrar al lector lo complicado que es dar una definición precisa de este término, por lo que el autor se conforma con decirnos que "se refiere a la copiosa cantidad de datos producidos espontáneamente por la interacción con dispositivos interconectados" (p. 19).

Los avances informáticos han sido responsables de la gran creación de datos que existe en la actualidad. Actualmente, lejos de conformarse con los 30 datos que muchos estudiantes de estadística, erróneamente, consideraban grandes, es posible obtener muchos más y de manera más rápida. Si bien esta revolución comienza con el tamaño, su relevancia se relaciona con el hecho de que los datos no son más de los mismos.

La utilidad de *big data* depende de los métodos empleados para su análisis. Con respecto a esto, el autor nos habla de *machine learning*, nombre que reciben las técnicas computacionales, matemáticas y estadísticas asociadas a este fenómeno. El objetivo del método es explotar los datos pasados para construir un modelo que prediga, de la mejor manera, una variable de resultado. Sosa Escudero nos cuenta que, en la vieja visión de la estadística, los datos se usaban para estimar un modelo que venía de afuera (teoría o experiencia previa), pero el *machine learning* cambia esta estrategia. La profusión de datos permite construir, estimar y reevaluar el modelo a medida que se lo usa. De esta forma, el rol de *big data* se relaciona con que los modelos complejos son altamente demandantes, en términos de datos. Cuanto

1 Shafer, Tom (2017). The 42 V's of Big Data and Data Science. https://www.kdnuggets.com/2017/04/42-vs-big-data-data-science.html

más flexible sea el modelo y cuanto menos se conozca de él, más datos se necesitarán para construirlo de forma confiable.

Finalizando el capítulo y con una referencia poética al entrañable Borges y su cuento "Funes el Memorioso", el libro habla de la importancia de la estadística en el análisis de datos; los datos por sí solos son cacofonía pura o, según el autor, "Funes es *big data* sin estadística". No obstante, también se aclara la necesidad que tiene la estadística, y en especial su enseñanza, de adaptarse a los nuevos tiempos.

En el segundo capítulo se nos presentan cuatros casos de la sociedad moderna respecto a la información masiva e instantánea y las diferentes maneras de usarla. En primer lugar, la investigación de Laura Trucco[2] –una alumna argentina del doctorado de Harvard– y su aplicación a la lentitud de los iPhone nos enseña que uno de los usos de *big data* y los algoritmos es el reconocimiento de patrones o correlaciones, aunque no necesariamente de causalidad. Una segunda cualidad de *big data* se relaciona con la sistematización de los datos. Si los datos están, pero no están convenientemente sistematizados, es lo mismo que nada. La ventaja de esta sistematización es mostrada en el caso anecdótico del programador Manuel Aristán, de Bahía Blanca, Argentina, y sus esfuerzos por crear un sistema funcional para analizar el gasto público de la municipalidad de dicha ciudad. En tercer lugar, se puede leer sobre el trabajo de Nicolás Bottan y Ricardo Pérez Truglia,[3] que estudian el efecto de la difusión mediática de los abusos sexuales cometidos por sacerdotes sobre la religiosidad. El trabajo sugiere que uno de los roles de *big data* es funcionar como fuente de información cruda, que, con el debido procesamiento, puede producir datos limpios y ordenados, como si fueran un experimento, de forma tal que permita hablar de causalidad. El cuarto y último ejemplo se refiere al caso de medición de la pobreza en Ruanda,[4] el cual apunta a que *big data* y sus algoritmos pueden complementar, y quizás sustituir, los mecanismos tradicionales de relevamiento estadístico.

El capítulo tres nos presenta las herramientas, técnicas y algoritmos para trabajar con este aluvión de datos. En él se repasan algunos conceptos ya conocidos en la estadística tradicional y se introducen otros nuevos. Comienza hablando de la importancia del agrupamiento de datos y de las posibilidades que nos brinda *big data* al respecto. De esta manera, se nombra al análisis de *clusters* como una herramienta crucial para el aprendizaje automático y al algoritmo de las *k-medias*. De forma obligada, se menciona el conocido y popular análisis de regresión y es el mismo autor el que lo bautiza como madre del resto de las estrategias más sofisticadas, resaltando

2 Trucco, Laura (s/f). On slow iPhones and conspiracy theories. En prensa.

3 Bottan, Nicolás y Pérez Truglia, Ricardo (2015). Losing my Religion: The Effects of Religious Scandals on Religious Participation and Charitable Giving. Journal of Public Economics, (129), 106-119.

4 Blumenstock, Joshua; Cadamuro, Gabriel y On, Robert (2015). Predicting Poverty and Wealth from Mobile Phone Metadata. Science, 350(6264), 1073-1076.

el rol de *big data* en mejorar la precisión de los modelos para predecir, dado el mayor volumen de datos disponible, lo que, a su vez, permite trabajar con modelos más complejos. Finalmente, el capítulo nos cuenta acerca de CART (Classification and Regression Trees), un concepto no tan conocido en la estadística tradicional, pero que es nombrado como la técnica más popular del aprendizaje automático.

El cuarto capítulo vuelve a la historia de los datos, pero desde la perspectiva de los algoritmos. Se recalcan dos características: la primera es que la revolución de datos tiene que ver con lo que se hace con ellos, la segunda gira sobre la idea de que una parte importantísima de la revolución de datos se relaciona con ampliar radicalmente el tipo de información o dato que es susceptible de análisis por un método sistemático. Las herramientas, técnicas y algoritmos mencionados previamente permiten operar con números, palabras, fotos, canciones, olores, frases y dibujos. A través de entretenidos relatos de hechos reales, el autor ejemplifica el avance exponencial del aprendizaje automático en tareas complejas como traducir textos, reconocer manuscritos, analizar estadísticamente millones de conversaciones o reconocer nuestros patrones de consumo. No obstante, esto no es posible sin la intervención creativa de las personas que estuvieron detrás. Es la interacción entre datos, computadoras, estadística e investigador lo que nos asegura un correcto y ventajoso uso de la nueva y copiosa información.

En el capítulo cinco el autor nos habla de la complejidad de un modelo en el mundo de *big data*, un concepto que luego relaciona con los de sobreajuste, regulación y *cross validation*. La predicción es uno de los objetivos que tiene un investigador cuando desarrolla un modelo. Es común caer en la tentación de plantear un modelo complejo, que ajusta muy bien a los datos que se tienen pero que presenta un mal desempeño fuera de ellos. Acá estamos en presencia de un sobreajuste de los datos. En consecuencia, una tarea del investigador es elegir la complejidad del modelo y la técnica que se usa; por esto se la llama *regularización*, en relación con métodos que intentan *negociar* entre el objetivo de ajustar bien respecto de los datos disponibles, pero penalizando o regularizando el uso de modelos demasiado complicados. La manera de elegir entre los distintos modelos es evaluando su capacidad predictiva, siendo *cross validation* la estrategia que el autor nos presenta como la más empleada para este fin.

Una forma alternativa de pensar el problema de sobreajuste es que se trata de una consecuencia de que la complejidad del modelo creció tanto o más rápido que los datos. Si bien el lector podría pensar que *big data* es una solución a ese problema del sobreajuste, Sosa Escudero nos recuerda la "maldición de la dimensionalidad", la cual nos dice que la cantidad de datos necesarios para estimar confiablemente un modelo crece mucho más rápido que su complejidad. Como una posible solución al problema de la dimensionalidad, aparece la técnica de *deep learning*. Esta es una forma progresiva de construir modelos complejos no lineales a partir de muchos modelos simples. Un punto central de este capítulo es que existe un grado óptimo

de complejidad y que *big data* ayuda al avance de modelos más complejos, aunque debe hacerse con cautela.

El capítulo seis nos invita a pasear por algunas limitaciones o inquietudes que surgen con el uso de estos datos masivos. Con ejemplos ilustrativos, el autor nos muestra cómo, en algún punto, el inadecuado uso de *big data* puede pasar los límites de la privacidad y la ética al confundirse con transparencia. Otra limitación se relaciona con los datos empleados erróneamente, tanto por hablar de una correlación espuria o por estar en presencia de sesgo. Como dice el autor, "la profusión de datos de *big data* no logra tapar los viejos sesgos de la estadística sino todo lo contrario: los amplifica y les da una nueva vida" (p. 116). El sesgo ocurre porque la intensidad de datos se relaciona con el fenómeno a estudiar; ya estamos lejos de ese muestreo aleatorio del que siempre nos hablan en estadística. Por último, el capítulo hace hincapié en que, muchas veces, la dificultad de la técnica se enfrenta con lo que puede comprender la sociedad. Este punto es un tema crucial, ya que las sociedades necesitan entender los mecanismos que producen cifras sensibles como la pobreza o el desempleo. La falta de entrenamiento o de "educación algorítmica" es un problema a sortear con el fin de poder beneficiarse de las oportunidades de aplicar métodos más complejos.

El capítulo final nos habla del futuro de los datos. En los párrafos iniciales se menciona un artículo publicado en el 2008 en la revista Wired y escrito por Chris Anderson, que se titula "El fin de la teoría: el diluvio de datos hará que el método científico sea obsoleto".5 Entre sus oraciones se pueden leer algunas palabras que resumen muy bien la idea del artículo y uno de los debates más polémicos de big data: "Basta de una vez con la teoría del comportamiento humano... Con suficientes datos los números hablan por sí mismos" (Anderson, 2008, p. 2). La publicación augura el fin de la teoría y la estadística en pos de un reinado de los datos. En esta parte del libro, el autor nos presenta el punto de vista del sector más escéptico del big data y el de sus fundamentalistas. También nos invita a reflexionar al respecto.

Entre las razones para pensar que los datos no bastan, el autor le enseña al lector que N (tamaño de la muestra) no es del todo como argumentan los talibanes. Mediante un ejemplo didáctico nos trata de convencer, y yo creo que lo logra, de que por más datos que genere *big data*, no hay forma de tenerlos todos. ¿Por qué? No hay que olvidarse de la información contrafáctica (sí...esa que aparece en los experimentos). Entonces, desde el punto de vista de la determinación de causas y efectos, solo se observan nuestras acciones y no nuestros contrafácticos: *big data* nunca es todos los datos. Aun así, ayuda considerablemente al diseño de experimentos, a la construcción de contrafácticos, o a la detección de datos que, aun siendo de origen observacional, se comporten como si hubiesen sido generados por un experimento

5 Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. Wired magazine, 16(7), 16-07.

y sirvan para entender canales causales. Pero es el investigador el que deberá usar el potencial de los datos masivos para explorar esas cuestiones causales.

Una aclaración respecto a la contribución de *big data* es que no se relaciona con el volumen de información, sino que estos datos iluminan aspectos de la sociedad que raramente son captados por los mecanismos tradicionales, como las encuestas. Por ejemplo, el nivel de racismo en una sociedad es difícil de detectar en una encuesta o experimento, pero puede verse reflejado en los datos provenientes de las interacciones digitales.

Terminando el capítulo, el autor nos habla del futuro de la estadística y resalta que sigue teniendo un rol central, a pesar de la revolución de datos y los nuevos algoritmos. Lo malo de *big data* aparece cuando desvaloriza a la teoría, pero lo peor de la estadística clásica surge al no apreciar la revolución algorítmica. Y, como nos dice el autor, dado que no podemos tener todos los datos, la estadística y la ciencia tienen un presente y un futuro asegurados, interactuando con la información masiva y los algoritmos y no compitiendo con ellos.

En resumen, este libro es una buena lectura para comenzar a sumergirse en la literatura de *big data*. Sosa Escudero no solo nos introduce un nuevo lenguaje, conceptos, técnicas y herramientas, sino que nos invita a reflexionar acerca de un fenómeno que ya forma parte cotidiana de nuestra vida y, quizás, estábamos ignorando. Aunque ciertamente, la ciencia del análisis de datos lo tiene muy presente.

Para terminar la reseña me gustaría recordar una idea final que plantea el libro, la cual, considero, es de lo más acertada. El verdadero desafío de la lluvia de datos es para el sistema educativo, el cual debería apuntar a acercar la enseñanza de la matemática a la computación, a la estadística y a casi todas las disciplinas científicas. Este diluvio de datos ofrece un futuro promisorio para todas las ciencias, pero se necesitan investigadores formados para poder sacarle provecho.