
BIG
DATA

Walter Sosa Escudero
Universidad de San Andrés

BIG DATA
Oportunidades y desafíos para las ciencias sociales

Difícilmente exista hoy algún ámbito del conocimiento ajeno a la revolución de *big data*. Las ciencias sociales no solo no son una excepción a lo dicho, sino que parecen haber encontrado una bendición en el virtual diluvio de datos provenientes de dispositivos interconectados como los teléfonos celulares o GPS, cuyo análisis requiere enfoques estadísticos novedosos. Esas disciplinas solían estar tradicionalmente limitadas, a la hora de obtener y analizar datos, a los relevamientos estadísticos acostumbrados, sean censos o encuestas por muestreo, o a experimentos controlados con alguna semejanza a los de las ciencias naturales, y a los métodos estadísticos clásicos.

Los más acérrimos defensores de *big data* llegaron a hablar de un verdadero cambio de paradigma en el sentido definido en 1962 por el físico Thomas Kuhn (1922-1996) en *La estructura de las revoluciones científicas*. Chris Anderson, en su momento columnista en *The Economist* y editor de la revista *Wired*, incluso se refirió a un auténtico 'fin del método científico' en un controvertido artículo publicado hace diez años que se cita entre las lecturas sugeridas. Estos entusiastas, sin embargo, se ven contrapesados por posturas más cautas, como la de Stephen Stigler, historiador de la estadística y profesor de la Universidad de Chicago, quien en su último libro, igualmente citado, afirmó: 'Funes es *big data* sin estadística'. La frase se refiere al memorioso Ireneo Funes, un curioso personaje de Jorge Luis Borges que podía recordar los menores detalles, al punto tal que relatar los acontecimientos de un día le tomaba veinticuatro horas. Stigler sugirió que la auténtica ventaja de *big data* no se debe a la profusión de datos sino a las técnicas de análisis provistas por una disciplina clásica como la estadística, con lo cual relativizó la importancia del fenómeno.

¿DE QUÉ SE TRATA?

La combinación de *big data* con aprendizaje automático por computadoras abre nuevos caminos para las ciencias sociales.

Este artículo adopta una visión intermedia con relación a las posibilidades de *big data* en las disciplinas sociales, y sopesa las enormes oportunidades que brinda con las dificultades propias de estas ciencias.

Cuatro casos

Cualquier definición de *big data*, incluida la que aparece en el sumario de la página del índice, es arbitraria y hasta imprecisa. Así y todo, la idea engloba tanto la masividad de datos proveniente del uso de dispositivos interconectados, la variedad y rapidez de acceso a ellos, como a las técnicas utilizadas para obtenerlos, organizarlos, sistematizarlos y estudiarlos. Naturalmente, la ciencia de la computación desempeña un papel crucial en lo anterior, lo mismo que la matemática, la estadística y hasta disciplinas aparentemente lejanas, como el diseño y la comunicación, a las que se agregan todas las ciencias que se benefician del uso de datos. El fenómeno de *big data* requiere una auténtica visión multidisciplinar.

Los cuatro casos siguientes intentan ilustrar el gran potencial de *big data* en el ámbito de las cuestiones sociales.

Medición de la pobreza en Ruanda

Ruanda es un país de África afectado por múltiples factores que lo colocan entre los más pobres del mundo. No es necesario argumentar demasiado acerca de la relevancia de medir la pobreza en países como este, tanto para su diagnóstico como para el diseño, la ejecución y la evaluación de políticas públicas destinadas a aliviar las penurias extremas de sus poblaciones.

El enfoque tradicional para medir la pobreza se basa en el *método de línea*, por el cual una persona es considerada pobre si su ingreso mensual no alcanza para comprar determinada canasta de bienes de primera necesidad. El método requiere realizar encuestas periódicas de ingresos y relevamientos regulares de precios, para computar el valor de la canasta a partir del cual se deja de ser pobre (límite llamado 'línea de pobreza'). Se trata de una tarea costosa, que requiere un considerable esfuerzo estadístico y logístico, y a la vez un fuerte compromiso institucional, todo lo cual excede claramente las capacidades de países como el de nuestro ejemplo.

Un artículo publicado por académicos norteamericanos (Blumenstock J, Cadamuro G & On R, 'Predicting poverty and wealth from mobile phone metadata', *Science*, 350, 6264: 1073-1076, DOI 10.1126/*science.aac4420*) describe un inteligente enfoque que cae en el campo de *big data* para encarar la medición de la pobreza así definida de los habitantes de Ruanda. Los autores trabajaron con datos de una pequeña encuesta sobre consumo y con información acerca de la intensidad de uso de teléfonos celulares, y diseñaron un modelo matemático que relaciona esos datos. Luego de un largo proceso de evaluación y ajustes de la capacidad predictiva del modelo, lo usaron para extrapolar esa información y predecir el nivel de pobreza en todas las regiones del país. La elección del modelo óptimo se basó en el *aprendizaje automático*, un proceso por el cual un algoritmo computacional elige la mejor modalidad de predicción sobre la base de algún criterio estadístico.

Construcción de índices de costo de vida

Un segundo caso es un proyecto llevado a cabo por el Instituto de Tecnología de Massachusetts (MIT) de construcción de índices de precios con información tomada de internet, una técnica que recibió el nombre de *scrapping*. Verificar los precios de los bienes y servicios es crucial para determinar el costo de vida en un país. El enfoque tradicional, usado por las oficinas gubernamentales de estadística, se basa en definir una canasta de productos, diseñar un complejo mecanismo de muestreo y relevar una gran cantidad de precios tanto en negocios pequeños como en supermercados. Los precios constatados luego se sistematizan y se agregan, con el fin de producir un índice de precios representativo de una región y un período. Naturalmente se trata de una tarea delicada y engorrosa, que requiere realizar considerable esfuerzo económico y operativo.

El proyecto Billion Prices Project del MIT (*http://www.thebillionpricesproject.com/*) construye índices simulando lo que haría un consumidor con una computadora conectada a internet: buscar precios en tiendas online. Un programa de computación toma los precios, los sistematiza, los agrega y construye un índice. El procedimiento evita el complejo sistema de muestreo usado en el método tradicional. Además, proporciona resultados inmediatos y registra sus cambios a medida que se producen, es decir, en tiempo real. Permite así conocer la evolución del costo de vida en distintas regiones y períodos sin demora y con la frecuencia que se desee, algo antes impensado.

Identificación multidimensional de la clase media

El tercer caso es un estudio en el que participó el autor de esta nota y cuyo propósito era encontrar quiénes pertenecen a la clase media. Una seria dificultad de plantearse tal estudio es que no existe ninguna definición obvia del concepto de clase media. Una simplificación usada con frecuencia consiste en identificar ese grupo sobre la base de los ingresos de los hogares, lo que tiene la indudable ventaja operativa de concentrar el análisis en una sola variable. Pero existe consenso en que el ingreso resulta insuficiente para cuantificar el bienestar, lo cual da lugar a enfoques *multidimensionales*, que toman en cuenta otras variables además del ingreso. El estudio en cuestión usa información sobre múltiples determinantes del bienestar, entre ellos desempeño laboral, educación y salud.

Con el proceso de aprendizaje automático empleado resultó posible no solo identificar a quienes pertenecen a la clase media en un sentido multidimensional, sino también aislar un pequeño grupo de variables que permite caracterizar a dicha clase. Es decir, el algoritmo computacional encontró el *mínimo* número de variables que puede representar adecuadamente el bienestar.

El efecto del impuesto a las ventas

Un último caso se refiere a un estudio sobre los efectos del impuesto a las ventas, encarado por la Universidad de Harvard y la empresa de comercio electrónico Ebay. Un problema básico de los mercados es que, de acuerdo con la teoría económica clásica, las personas deciden sus compras considerando el total que pagan por los productos, sin preocuparles cuánto es el precio que cobra el vendedor y cuánto se agrega por impuestos. Para descubrir el efecto de los impuestos sobre la demanda, un experimento bien diseñado debería aislar ambos componentes y considerar las decisiones del comprador ante cada uno. El estudio en cuestión usa el hecho de que en los Estados Unidos quienes compran en Ebay deben proceder por pasos: primero eligen qué comprar y en qué cantidad sobre la base del precio del vendedor, y luego, de proceder con la compra, se enteran del impuesto a pagar (este es variable según los estados de residencia del comprador). Esta secuencia permite identificar el efecto del impuesto a las ventas en las decisiones de compra, es decir, permite saber cuántos desisten de una compra al enterarse del impuesto. El estudio explota el enorme caudal de información contenida en Ebay, que en 2017 tuvo registrados en sus bases de datos entre 162 y 170 millones de compradores activos domiciliados en los Estados Unidos, los que a lo largo de ese año realizaron compras por valor de más de 80.000 millones de dólares. Este inteligente procedimiento explota el enorme caudal de información disponible en Ebay para reemplazar un experimento con datos observacionales.

Los casos anteriores muestran el enorme potencial de *big data* en los estudios sociales, por ejemplo, para predicción (pobreza en Ruanda), medición (índices de precios), clasificación (clase media) o evaluación de políticas públicas (impuesto a las ventas). Los casos también ilustran el potencial de *big data* para

reemplazar métodos tradicionales, que son costosos –como las encuestas oficiales– o impracticables –como establecer el efecto del impuesto a las ventas o medir la pobreza en un país como Ruanda–.

Cinco desafíos

Naturalmente, ninguna tecnología está libre de limitaciones, y *big data* no es una excepción. La postura extrema de sus propulsores lleva a ignorarlas, y la de sus detractores a blandirlas como muestra de la inutilidad del método. Una actitud honesta y cauta lleva a verlas como auténticos desafíos, para extraer todo el potencial del método sin caer en sus limitaciones. Lo que sigue es una breve lista de desafíos que plantea el uso de *big data* en las ciencias sociales.

- *La estructuración de los datos*. Para obtener conclusiones confiables no necesariamente se necesita mucha información. Los mecanismos tradicionales de muestreo complejo subyacentes a todas las encuestas sociales, como la Encuesta Permanente de Hogares de la Argentina, intentan garantizar que la información contenida en unos pocos datos pueda ser extrapolada a una población amplia. Así, el bienestar de los aproximadamente cuatro millones de hogares del Gran Buenos Aires es captado mediante una muestra de 3039 observaciones. Que esta pequeña muestra sea representativa de esa gran población es uno de los grandes logros de la estadística científica, alcanzado porque las pocas observaciones se seleccionan con claro patrón probabilístico, lo cual permite relacionarlas con la población a la que pertenecen en forma confiable.

  Por el contrario, los datos de *big data* no tienen esa característica, pues provienen de dispositivos o sensores cuyo uso no obedece a un plan sistemático. Así, los usuarios de relojes inteligentes que monitorean la actividad física y proveen copiosa cantidad de información no son una muestra representativa de una población más amplia, de modo que es peligroso extrapolar las conclusiones extraídas de los datos de ese grupo a quienes no usan tales relojes. En otras palabras, los millones de datos de *big data* no son directamente comparables con los de una encuesta sistemática, y es posible que unas pocas observaciones bien estructuradas contengan información más útil que una enorme cantidad de datos no estructurados. Un desafío relevante es estructurar los datos de *big data* para su uso representativo en ciencias sociales.
- *Big data no es todos los datos*. La evaluación de políticas sociales requiere comparar las situaciones de beneficiarios y de no beneficiarios, como lo hacen los estudios de nuevos medicamentos, con pacientes que los reciben y otros que no los reciben y constituyen un grupo

de control. Por ejemplo, la medición de la efectividad de un programa de asistencia social como la Asignación Universal por Hijo requiere comparar la situación de familias que la cobran con la de familias que no la cobran (con la condición de que sean elegidas estrictamente al azar). No es lícito comparar familias no elegidas al azar, pues las de un grupo podrían tener características distintas de las del otro. *Big data* no necesariamente permite identificar los grupos que, para cierta variable, puedan servir de control. Un enorme desafío de *big data* es explotar la profusión de datos para encontrar cómo identificar grupos de control y así evitar comparaciones fútiles que resultan de un uso ingenuo de los datos, por muchos que sean. El citado caso de Ebay, sin embargo, señala una forma por la que *big data* puede ayudar a identificar los mencionados grupos comparables.

- *Predecir no es explicar.* La utilidad paradigmática de *big data* es la cuantificación y el pronóstico. Pero medir o pronosticar no necesariamente implica explicar, ni mucho menos encontrar o descartar causas. En el caso de Ruanda, dada la información colectada por la encuesta sobre la que se construyó el modelo, *big data* proveyó una herramienta útil para medir y pronosticar la pobreza, pero poco aportó al esfuerzo de entender sus causas, de encontrar medidas para remediarla y, menos aún, de evaluar la eficacia de políticas para paliarla. Si bien en ese caso concreto el uso de teléfonos celulares pudo ayudar a predecir y medir la pobreza, es falaz argumentar que eso justifica subsidiar el uso de tales adminículos para bajarla (aunque pudiese haber otras razones que justificasen tal subsidio). Dicho en términos genéricos, la naturaleza inductiva de *big data* pone de relieve correlaciones que permiten medir y predecir, pero téngase en cuenta que correlación sola no indica causa, ni la excluye.
- *Transparencia versus privacidad.* En pos de la transparencia, hace unos años el gobierno de Noruega dispuso la difusión online de los ingresos de todos sus habitantes. Muy rápidamente aparecieron episodios de acoso o manifestaciones de envidia social que indujeron a las autoridades a poner límites a la práctica, para preservar la privacidad de los ciudadanos. Como se aprecia en este ejemplo, la espontaneidad y aparente anarquía de *big data* puede crear un conflicto entre valores igualmente deseables para la gente. Cómo manejar esas situaciones es otro de los grandes desafíos del uso de *big data* en el contexto social.
- *El desafío de la comunicabilidad.* Los beneficios que se pueden obtener de *big data* se basan en procedimientos complejos que priorizan la capacidad predictiva. Con este objetivo, muchas veces prevalecen métodos no lineales e iterativos, claros para los expertos pero incomprensibles para los legos. En el ámbito de la política social, sin embargo, la adopción de una tecnología depende tanto de su performance como de la posibilidad de convencer a legos que respalden su uso. El desafío no es simplificar los métodos sino promover la participación de grupos interdisciplinarios entre los que se pueda crear una cadena de confianza. Tal cadena garantizaría que los métodos de *big data* alcancen tanto los fines predictivos como los de consenso social que demanda la información sobre la que se apoyen las políticas públicas.

Comentarios finales

Adoptar ciegamente las técnicas de *big data* por el hecho de que están de moda es tan necio como negarse a hacerlo por las mismas razones. La profusión de datos que ofrece *big data* brinda alternativas antes impensadas para explorar empíricamente varios aspectos de la sociedad. Con todo, la naturaleza sistémica, interactiva y compleja de las ciencias sociales requiere algunos cuidados particulares. Si las limitaciones de la técnica son entendidas como auténticos desafíos y no como limitaciones infranqueables, las oportunidades son enormes. CH

LECTURAS SUGERIDAS

ANDERSON C, 2008, 'The end of theory: The data deluge makes the scientific method obsolete', *Wired*, 26, 6, disponible en *https://www.wired.com/2008/06/pb-theory/*.

JAMES G. et al., 2013, *An Introduction to Statistical Learning with Applications in R*, Springer, Nueva York, accesible en *http://www-bcf.usc.edu/~gareth/ISL/*.

MAZZOCCHI F, 2015, 'Could Big Data be the end of theory in science? A few remarks on the epistemology of data-driven science', *EMBO Reports*, 16, 10: 1250-1255, accesible en *http://embor.embopress.org/content/16/10/1250.long*.

MURPHY K, 2012, *Machine Learning: A probabilistic perspective*, MIT Press, Cambridge MA.

SOSA ESCUDERO W, 2017, '*Big data* y aprendizaje automático: ideas y desafíos para economistas', en *Una nueva econometría*, en prensa, versión técnica de este artículo disponible en redacción preliminar en *http://waltersosa.weebly.com/uploads/2/2/1/8/22189288/aaep2017bigdatawse.pdf*

STIGLER S, 2016, *The Seven Pillars of Statistical Wisdom*, Harvard University Press, Cambridge MA.

Walter Sosa Escudero

Doctor (PhD) en economía, Universidad de Illinois en Urbana-Champaign.
Investigador principal del Conicet.
Profesor plenario, UDESA.
Director del Departamento de Economía, UDESA.
Miembro de número, Academia Nacional de Ciencias Económicas.
wsosa@udesa.edu.ar