big data

walter sosa escudero

¡tercera edición!

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

siglo veintiuno editores

ciencia que ladra...
serie mayor

1. Perdidos en el océano de datos

Big data, aprendizaje automático, ciencia de datos, estadística y otras yerbas

> –Doctor, escúcheme, esta gente está muy mal. Me dicen que tengo que hacer un curso de Hadoop, me hablan de modelos obesos, de riesgo de Bayes, de matrices de confusión y de la curva ROC. No, doctor, *rock*, no, ¡ROC! Bueh, no sé, en algún momento nombraron a Reproducing Kernel Hilbert Space, y yo creí que era grupo de *rock* psicodélico de los setenta, como Pink Floyd… Doctor, no entiendo nada. ¡Socorrooooooo!

¿Así que no entendieron nada? No se preocupen, no están solos. Los datos son tierra de todos y de nadie. Y como en la Buenos Aires de comienzos del siglo XX, en el ambiente del análisis de datos se escucha hablar ese cocoliche propio de quien intenta decir en castellano lo que los años le enseñaron en otro idioma.

Este capítulo es nuestra primera visita a la políglota metrópolis de los datos. Fracasaremos en nuestro primer intento de definir qué es big data, pero saldremos airosos diciendo que, hasta ahora, todos los intentos han sido fallidos. Visitaremos los bodegones nobles de la estadística y nos deleitaremos en los nuevos restobar del aprendizaje automático. Nos detendremos a apreciar el monumental edificio de datos que construye big data y seremos testigos de algunas disputas entre los

viejos cocineros de la estadística y los nuevos chefs de la ciencia de datos. Y al finalizar el recorrido tal vez ya no les resulten tan raros algunos de los términos esotéricos del comienzo.

El Elvis Presley de la ciencia de datos (vida, muerte, resurrección y nueva muerte de Google Flu Trends)

El 4 de julio de 2009 fue un sábado de sol radiante en Buenos Aires. Cinco días antes las autoridades habían decidido cerrar todas las escuelas por temor a la propagación de la pandemia de gripe A, medida que afectó a casi 11 millones de estudiantes, incluyendo a mi hijo, que en ese entonces tenía 6 años. Y tras cuatro días de encierro, concluimos con mi esposa que salir a dar una vuelta en auto no podía ser mucho más peligroso que la PlayStation, que tenía atrapado a mi hijo en su confinamiento. La ciudad nos devolvió un panorama desolador: las calles vacías, los negocios tristes, los carteles oportunistas de venta de alcohol en gel, y alguno que otro transeúnte con barbijo, como esos que hasta la fecha solo habíamos visto por televisión, en el aeropuerto de algún lejano país, como Japón.

Epidemias como la gripe A son un serio desafío para la salud pública, y es crucial monitorear con precisión y rapidez su evolución, tanto en el espacio (por dónde se reproduce) como en el tiempo (a qué velocidad). Se trata de una tarea compleja, aun para naciones desarrolladas como los Estados Unidos. En 2009, la forma de llevar a cabo el monitoreo en ese país era a través de un sistema de reportes estadísticos coordinados por el Centro para el Control y la Prevención de las Enfermedades (CDC). Las unidades hospitalarias (clí-

nicas, salas, hospitales, etc.) recababan información de las consultas por síntomas de gripe A, sus tratamientos y algunas características demográficas de los pacientes (género, edad, etc.). Estos reportes eran agregados a nivel de ciudad, condado, estado y región, y finalmente condensados en un informe a nivel nacional. Todo este proceso tomaba unos diez días: demasiado tiempo para una epidemia peligrosa como la gripe A.

En la antesala de la pandemia, la empresa Google propuso un ingenioso mecanismo –Google Flu Trends– que prometía bajar el rezago informativo de diez días a tan solo uno: un gol de media cancha de big data. El punto de partida del método fue una base de datos pequeña, de la cantidad semanal de visitas por gripe A a las unidades hospitalarias de las nueve regiones en las que el CDC divide a los Estados Unidos, entre 2003 y 2007, y medidas como porcentaje del total de visitas. Nueve regiones por cinco años, por 52 semanas da 2340 datos. Por ejemplo, uno de los datos diría que en la región 3, en la semana 12 de 2005, 1,2% de las personas que visitaron hospitales o clínicas lo hicieron con síntomas de gripe A. Estos datos miden cómo se distribuye la enfermedad por región y en el tiempo, o sea, es "la" variable que se precisa para monitorear la pandemia y que, según dijimos, tomaba unos diez días en elaborarse.

Estas localizaciones de datos no nos resultan tan extrañas. Ahora, por ejemplo, mientras espero aburrido que mi hijo salga de un cumpleaños, descubro en el celular una simpática opción en Google Maps que se llama "tus rutas". Con pasmoso detalle me muestra todos los lugares en los que estuve durante el día: mi ruta al trabajo, la bicicleteada junto al río, las tres cuadras que me desvié para comprar leche en el supermercado, etc., etc. Además de nuestra localización geográfica, Google

ve y atesora las canciones, libros, colegas, restaurantes, zapateros, vendedores de heladeras, direcciones, teléfonos de *delivery* y todo, absolutamente todo, lo que hemos buscado. Y también cuando una mamá atemorizada escribió "mi hijo tiene gripe A" en el buscador, y cuando otra persona puso "tengo fiebre, tos y estoy fatigado", y cuando otro tipeó "remedio influenza".

Aquí interviene el análisis de datos. Los expertos de Google cruzaron los 2340 datos de porcentaje de visitas a hospitales con la proporción de búsquedas relacionadas con la gripe A en cada período y región. Fácil no es: hay que empezar por definir qué significa "búsquedas relacionadas con la gripe A", lo que requiere un delicado trabajo de "curación", es decir, decidir qué términos y frases se relacionan estrictamente con esta enfermedad y cuáles no. Concretamente, poner en Google "tengo frío" puede ser tan compatible con síntomas de gripe A como con la mera llegada del invierno. Luego de concluida esta delicada tarea, Google disponía de 2340 pares de datos: la intensidad de visitas a hospitales por gripe A –provenientes de la información oficial– y las búsquedas en Google de términos relacionados con la enfermedad –proporcionadas por la misma empresa–, para cada región, año y semana. Con estos datos, los científicos de Google construyeron un modelo para predecir la intensidad de gripe A sobre la base de la intensidad de búsquedas.

Típicamente, para aprender a manejar alguna técnica, en una clase de estadística los alumnos estiman algún modelo simple usando datos reales; "modelo" entendido no como un ideal, sino como una representación matemática o computacional de la realidad. Los científicos de Google estimaron 450 millones de modelos alternativos para elegir el que mejor predice la gripe A sobre la base

de la intensidad de búsqueda. Un punto importante es que todo este proceso de estimación (que más adelante definiremos como "de aprendizaje") se basó solo en 2340 pares de datos, de intensidad de consultas y búsquedas semanales y a nivel de región, es decir, sobre la base de la desagregación más fina posible; a nivel de hospital en una región, para una semana en particular. Pero una vez construido el modelo, podría usarse para predecir la intensidad de la epidemia a partir de cualquier información disponible sobre intensidad de búsqueda.

Y en esta parte de la historia Google saca a relucir su monstruosa base de datos. A diferencia de la agencia de control estadounidense, que solo ve datos semanales y por región, Google puede observar la intensidad de búsquedas en cualquier parte, en tiempo real y con un nivel de precisión tan fino como sea necesario. Es decir, Google puede medir, por ejemplo, la intensidad de búsquedas sobre gripe A en Monticello, un minúsculo pueblito del estado de Illinois y, a partir del modelo estimado previamente, predecir la intensidad de la enfermedad en ese lugar. Y también puede hacerlo de forma diaria, semana o mensual, tanto para Monticello como para la ciudad de Nueva York, el estado de California o cualquiera de las nueve regiones en las que el CDC divide a los Estados Unidos.

En definitiva, a Google le toma solo un día hacer lo que al sistema público de una de las naciones más ricas del planeta le toma diez, y con una capacidad predictiva mucho más microscópica. Es David dándole una contundente paliza a Goliat.

De ser big data *rock and roll*, Google Flu Trends sería Elvis: el abanderado insignia de la revolución de datos y algoritmos, entendidos como procedimientos y reglas sistemáticas para hallar la solución a un problema. Éxito

rotundo, resultados publicados en la prestigiosísima revista *Nature*, "aplauso, medalla y beso", como se decía en un vetusto programa de televisión argentino. Pero los aficionados al *rock* sabemos que luego del éxito masivo a Elvis le sobrevino el ostracismo y una inoportuna convocatoria para hacer el servicio militar en 1958. Derrotero similar sufrió Google Flu Trends, cuyos éxitos predictivos se transformaron rápidamente en preocupantes desaciertos. En particular, para varios períodos el algoritmo predice intensidad de gripe A muy por arriba de la realidad. Varios analistas dicen que este error se debe a que Google alteró sus motores de búsqueda para retener a los que entran al buscador con consultas relacionadas con la gripe A, como si escribiesen "síntomas gripe A" y Google les sugiriese buscar términos como "tos" o "jarabe", reteniéndolos en el buscador para ofrecerles publicidad. Es decir, los cambios en los procesos de búsqueda de Google indujeron espuriamente a más búsquedas sobre la gripe A, lo que implicó que se sobredimensionara su intensidad y, por lo tanto, la epidemia. Sin embargo, como Elvis a fines de los sesenta y su exitosísimo *comeback* ya en épocas de Los Beatles, Google Flu Trends fue resucitado por la comunidad científica, que logró reparar algunos de sus errores y restablecer parte de su credibilidad. No obstante, en agosto de 2015 Google dio de baja el acceso público al servicio, si bien sigue recolectando información que es enviada para su análisis a la Universidad de Columbia y otras instituciones científicas.

En su momento, Google Flu Trends fue el "chico de tapa" de big data: los algoritmos contra la burocracia, los datos versus la teoría. Y todavía no sabemos si hemos visto su final definitivo, como con Elvis, quien más allá de su regreso glorioso terminó sus días prematuramente cuando ya era una cruel caricatura de sí mismo, o

si surfeará exitosamente el paso del tiempo cual Keith Richards, a quien en los setenta, por sus excesos, nadie le daba más de un par de años de vida.

Hace unos treinta años que me dedico profesionalmente a la estadística. Y cada cinco años emerge una tecnología destinada a barrer con todo lo existente, para luego desvanecerse con la misma intensidad. Entonces, hago mías las palabras de Charly García: "mientras miro las nuevas olas, yo ya soy parte del mar", tanto en lo que se refiere a la actitud suspicaz de quien vio ir y venir las modas, como a la de quien –como el propio García– no dudó en reemplazar su larga melena *hippie* por uno de esos "raros peinados nuevos" y abrazar la nueva tecnología musical de los ochenta para mantener intacta su creatividad de los setenta.

En tecnología y en ciencia, quien se cierra a las innovaciones porque cree que van a pasar de moda recuerda al adolescente que no se baña porque "total me voy a volver a ensuciar", y a la larga termina viviendo en escasas condiciones de higiene. El derrotero de Google Flu Trends es una linda alegoría de lo que sucede actualmente. Los talibanes de los datos creen que big data reemplazará a todo tipo de conocimiento y solo ven su parte exitosa. Los escépticos, por el contrario, creen que es una moda pasajera y únicamente relatan su costado negativo. A nosotros nos toca contar toda la historia, de éxitos y fracasos, de aciertos y aprendizajes, de revoluciones y fiascos, de muertes y resurrecciones. E inferir la que todavía no hemos visto.

¿De qué hablamos cuando hablamos de big data?

Si un habitante del futuro pudiese viajar en el tiempo a septiembre de 2016, le llamaría la atención ver a un montón de personas en la calle haciendo movimientos extra-

ños con sus teléfonos celulares: era el inicio de la histeria de la caza de Pokemones. Se trataba de ubicar, perseguir y atrapar a esas criaturas virtuales –los Pokemones– de esotéricos nombres como Rowlet, Dartix o Decidueye. Para la misma época, la revolución de big data vino acompañada de términos como "Seahorse" (un entorno visual), "Hadoop" (un sistema de código abierto) o "Summingbird" (una biblioteca virtual de programación). No tardó mucho en aparecer un hilarante sitio web llamado "¿Es Pokemon o big data?", que proponía un jueguito virtual que consistía en adivinar si un término pertenecía a la jerga de big data o de Pokemon.

Uno de los enormes problemas de cualquier tecnología de moda es que viene acompañada de jerga: un catálogo de extraños términos, muchos en inglés e intraducibles, que sirve tanto a los efectos de designar objetos nuevos e imposibles de nombrar con los viejos términos, como de crear una innecesaria barrera a la entrada, al solo efecto de impresionar a los novatos en las reuniones de amigos como si realmente fuese necesaria una nueva palabra para referirse al agua tibia. La propia expresión "big data" es jerga. Cualquiera que haya permanecido durante quince minutos en una clase de inglés se da cuenta de que "big" significa "grande" y que "data" son "datos". No intentaremos ninguna traducción del término, porque no hay ninguna comúnmente aceptada (he visto "gran dato", que parece provenir de las frases del Tarzán de Ron Ely en *Sábados de superacción*), y porque tampoco está claro que "big data" tenga un significado preciso.

Este libro debería comenzar aclarando entonces qué es big data, en el mismo sentido y con la misma dificultad con que un libro de *jazz* debería decir qué es el *swing*. Pregúntenle a un avezado jazzero qué es el *swing*

y es probable que reciban como respuesta la que dio Louis Armstrong cuando alguien lo interrogó sobre qué era el *jazz*: "Desde que me lo preguntas, me di cuenta de que nunca lo entenderás". Lo más obvio es decir que big data son "datos masivos". Pero en realidad 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.

Piensen en lo que hicieron en las últimas dos horas. Si caminaron con su celular, muy posiblemente hayan generado datos de su ubicación geográfica, y ni hablar si activaron el GPS para viajar en auto. Lo mismo si salieron a correr con su reloj inteligente que les cuenta el ritmo cardíaco y los pasos. O si usaron la tarjeta de crédito, viajaron en subte, se entretuvieron con una serie en Netflix, le pusieron "me gusta" a una foto de su tía en Facebook, si mandaron o recibieron un *e-mail* o si buscaron un par de zapatos en Amazon. Todo generó datos.

Más adelante hablaremos acerca de que la cantidad de datos que se produce a través de estos medios desafía cualquier concepto de inmensidad que hayamos considerado nunca. Pero el volumen (*big*) es solo una parte de la historia. A diferencia de una encuesta sistemática, como una encuesta política o esas que todavía funcionan por teléfono de línea, los datos de big data son anárquicos y espontáneos. Toda vez que abrieron su celular para que una *app* de GPS los guíe hacia algún lugar, han generado datos, no con el propósito de contribuir a ninguna encuesta ni estudio científico, sino con el de evitar el tráfico o perderse. Es decir, los datos

no fueron generados por el propósito de crearlos, como en las respuestas a una encuesta tradicional, sino como resultado de otra acción: ir a una reunión, pagar con una tarjeta de crédito, entrar a un sitio web, etc.

Entonces, los datos de big data no son más de los mismos viejos datos (de encuestas, registros administrativos, etc.), sino un animal completamente distinto. En 2001, Doug Laney, analista de la consultora Gartner, escribió un influyente artículo en el que resumió esta discusión diciendo que la revolución de big data tenía que ver con las ahora archifamosas "tres V de big data": volumen, velocidad y variedad. La primera de las V hace referencia a "big" –mucho–. La segunda se refiere a que los datos de big data se generan a una velocidad que los hace disponibles a una tasa prácticamente virtual, en tiempo real. Y la tercera –variedad– remite a la naturaleza espontánea, anárquica y amorfa del objeto que ahora llamamos "dato": un tuit, una posición geográfica de un GPS o una foto, todo constituye un dato, muy lejos de los datos tradicionales, esos que uno imagina prolijamente ordenados en una planilla de cálculo. El truco comunicacional de las tres V es efectivo para decir que big data es bastante más que muchos datos. Pronto fue necesario agregar una cuarta V: veracidad, término que se refiere a que la naturaleza ruidosa y espontánea de los datos de big data contrasta con la de los datos burocráticos o de encuestas tradicionales, usualmente sometidos a puntillosos ejercicios de validación.

Pero en algún momento lo de las V se desmadró, y añadir una más a la lista original se transformó en algo no muy distinto de la caza de Pokemones: otra tontera social. En un jocoso artículo reciente, Tom Shafer habla de "las 42 V de big data": las tres iniciáticas propuestas por Laney, las dos o tres que juiciosamente se agrega-

ron en años posteriores, como "veracidad", y la insólita lista que se añadió recientemente, que incluye "vudú", "vainilla" o "varifocal" (no, no les miento).

Chanzas aparte, una definición de big data que tenga que referirse a 42 ideas es inoperante y oximorónica, como cuando un conocido peinador estilista se ufanaba de sus desfiles "con más de 200 top models", como si "top" no se contradijese con "200". Una definición que abarque 42 conceptos es cualquier cosa menos una definición.

Este libro no adopta ninguna definición precisa de big data. Porque es seguro que entre las tres V iniciales y las 42 del chiste de Shafer hay dimensiones relevantes por abarcar, y no querríamos pecar ni por omisión ni por inclusión innecesaria. Nos conformaremos diciendo que big data se refiere a la copiosa cantidad de datos producidos espontáneamente por la interacción con dispositivos interconectados.

Los amplificadores de big data van hasta 11

¿De cuántos datos hablamos cuando hablamos de big data? Allá por los años setenta, los viejos disc-jockeys (ahora DJ) clasificaban los temas musicales en "lentos" y "movidos". Y una irritante práctica de novato era, ante un tema, preguntar: ¿es lento o movido? Y algo igualmente molesto ocurre con quien pregunta si por arriba de cierto número de datos estamos hablando de big data.

A pesar de ser, por lejos, el más popular de los instrumentos musicales, la guitarra es muy ineficiente: demasiado grande para el bajo volumen que produce. El instrumento favorito de los fogones y las serenatas puede ser fácilmente tapado por un violín o una trompeta, de mucho menor tamaño. La guitarra eléctrica nace como

solución a este problema. Pero en la década del sesenta los músicos notaron que la conjunción de una guitarra y un amplificador producen más que "más volumen". Eric Clapton y Jimi Hendrix transformaron en ventajas lo que la tradición guitarrística veía como una contra de la amplificación. Y así es como, para espanto de los ingenieros de sonido de la época, aberraciones sonoras como el *feedback* y la distorsión se volvieron parte del lenguaje del *rock*. Y así también nació la carrera por el volumen, desde los pequeños parlantes de los bluseros de los cincuenta a las paredes de amplificadores de The Who o Kiss. Esta alocada carrera por el volumen es parodiada en *This is Spinal Tap*, la desopilante película que se mofa de los excesos del *rock*. En una memorable escena, el guitarrista Nigel Tufnel muestra orgulloso su amplificador a un periodista y le dice que "nuestros amplificadores suenan más fuerte porque el volumen va hasta 11". El atónito reportero pregunta: "¿Y por qué suenan más fuerte?", a lo que Nigel responde: "Los amplificadores de cualquier estúpido van hasta 10. Y una vez que estás en 10, ¿adónde podés ir? ¡A 11! ¡Uno más fuerte!".

Y cual Nigel Tufnel, "¿ahora adónde podemos ir?" fue la pregunta que nos hicimos allá en los ochenta, cuando arribaron los *diskettes* de 3½ y sus entonces asombrosos 720 Kb de capacidad de almacenamiento de información, máxime cuando ya habíamos explotado la "K" para referirnos a miles de bytes. Y así es que no tardó en aparecer "mega" (millón). Y después "giga". Y "tera", "peta", "exa", "zetta" y "yotta". Y ahora se habla de "hella": 1 000 000 000 000 000 000 000 000 000 de bytes.

"¡Hella más 1!", gritarían Tufnel y cualquier niñito de la primaria, conscientes del "segundo axioma de Peano", ese que dice que todo número natural tiene un

sucesor y que no hay tal cosa como un número entero más grande que todos los otros. Y así como a los pequeños les gusta que les cuenten cientos de veces la misma historia, a nosotros nos agrada que nos repitan hasta el hartazgo la saga del arrollador avance del volumen de información, la que justifica esos extraños términos como peta, zetta y yotta. Y como todo autor se debe a su público, vayan aquí algunos ejemplos:

- En los últimos dos años, en todo el mundo hemos creado más datos que en toda la historia de la humanidad.
- Cada segundo se crean 1,7 megabytes de información nueva.
- Los usuarios de Facebook envían 31,25 millones de mensajes y miran 2,77 millones de videos por minuto.
- En 2015 se sacaron 1 000 000 000 000 fotos.

Ejemplos que aun condenados a una prematura obsolescencia hemos escuchado cientos de veces, y nos producen una aparatosa sensación de falso asombro como la tía sorprendida por "lo grande que está el nene" cada vez que nos visita.

Un hito de la estadística fue el descubrimiento de "la t de Student", en 1908. La historia es archiconocida para quienes hayan tomado un curso de estadística. Para los que no, el tal "Student" era en realidad William Sealy Gosset, que descubrió una importantísima fórmula mientras trabajaba en la empresa Guiness (sí, la de la cerveza), que prohibía a sus empleados publicar resultados con sus nombres, de ahí el uso del seudónimo. Sin entrar en detalles, "la t de Student" es una mejora con respecto a la famosa "campana de Gauss" para tamaños

de muestra pequeños. De hecho, la tabla de valores de la t de Student se corta abruptamente en 30 datos, porque de ahí en más no hay mucha ganancia en usar la fórmula de Gosset en vez de la de Carl Friedrich Gauss. Sobre la base de esta apreciación, muchos estudiantes de estadística responden, erróneamente, "30" cuando les preguntan cuán grande es una muestra grande; tal vez uno de los disparates más grandes de la práctica estadística.

Treinta observaciones era una cifra normal para los estudios estadísticos de la época de Gosset, cifra que resulta irrisoria junto a números como el uno seguido de 27 ceros del hella, y comparado con el volumen de datos que hoy un estudiante de la escuela secundaria puede bajar con un clic para hacer un trabajo práctico. ¿Es cierto que desde Gosset a la actualidad el conocimiento relevante creció en una proporción similar a la del volumen de datos? No, pero estamos mejor. Entonces, la pregunta clave es cuánto mejor estamos.

Cuando hace poco alguien preguntó en las redes sociales cuán grande debía ser una base de datos para considerarla "de big data", un reconocido programador respondió: "Si no entra en Excel, es big data", afirmación que muchos tomaron literalmente, y el resto como una chanza al estilo de Louis Armstrong cuando le preguntaron qué era el *jazz*.

Hella, peta o lo que sea, las ventajas de big data no necesariamente vienen de "big". No seremos los primeros en decir que, en muchas circunstancias, el tamaño no importa. La revolución de big data empieza por el tamaño, pero muy rápido va por otros caminos mucho más interesantes, porque los datos de big data no son más de lo mismo.

El derrotero de los *Spinal Tap* y su obsesión por el volumen y los excesos fue triste: el final de la película los mues-

tra decadentes y ridículos, intentando acomodar en vano sus masivos amplificadores y su aparatosa escenografía en los pequeños teatros donde terminaron tocando, porque la calidad de su música había crecido muchísimo menos que su volumen. Por el contrario, la guitarra no ha caído un ápice en su popularidad, todavía infaltable en cualquier reunión de amigos. Será la conjunción del copioso volumen de datos, los métodos de análisis y procesamiento y las ideas lo que garantizará que big data siga el derrotero del noble instrumento de Paco de Lucía y Andrés Segovia.

La máquina de aprender

Así como hacen falta dos para bailar el tango, la contracara de la explosión de big data son los métodos utilizados para su análisis. *Machine learning* es el nombre que reciben las técnicas computacionales, matemáticas y estadísticas asociadas al fenómeno de big data. Y si es difícil traducir big data, *machine learning* es todavía más delicado, pero, afortunadamente, la práctica parece haber convergido en "aprendizaje automático".

Todo parece sugerir que hay una máquina que aprende automáticamente, y que la cuestión de los muchos datos cumple un rol importante. Aclarar esta cuestión es uno de los temas centrales de este libro, a tal punto que el capítulo 5 estará dedicado por completo a esta cuestión. Los métodos de aprendizaje automático caen en la frontera entre la computación y la estadística: ambas reclaman su paternidad, aunque en estadística se habla de "aprendizaje estadístico". Si bien existen diferencias entre "automático" y "estadístico", la línea divisoria es difusa, y en este libro los tomaremos como sinónimos, enfatizando las diferencias cuando sea relevante.

A fin de entender qué es esto de *machine learning*, pensemos en un ejemplo simple. La señora Manfredi entra a un banco a solicitar un crédito, y la institución debe decidir si se lo concede. El banco dispone de información pasada de los créditos que ya otorgó y si estos fueron pagados o no. Así, en su base de datos se encuentra el caso del señor Averastain, quien en el momento de solicitar el crédito tenía 32 años, con un trabajo estable como abogado, una casa y un auto, y que pagó el crédito asignado en tiempo y forma. Y también el del señor Vattuone, de 43 años, soltero, sin hijos y de profesión jugador internacional de póker, que nunca pagó su crédito.

Llamemos "Y" al hecho del pago del crédito, donde Y es igual a 1 si fue pagado e igual a 0 si no se pagó. Entonces, para Averastain Y vale 1, y 0 para Vattuone, que no pagó. Usemos "X" para referirnos a toda la información disponible para cada persona a la cual se le otorgó un crédito en el pasado, que en nuestro ejemplo incluye la edad, si trabaja o no, los bienes de que dispone como garantía, etc. Como es de esperar, que una persona pague un crédito depende de factores observables por el banco (reunidos en X) y también de cuestiones azarosas o inobservables, que llamaremos "u". Así, tal vez una cruel enfermedad llevo a Vattuone a abandonar el país y dejar el crédito impago, o quizás una oportuna herencia le permitio a Averastain enfrentar su deuda. Todos estos factores que el banco no ve conforman u. Con esta información, el banco construye un modelo matemático que de forma diagramática funciona de la siguiente manera:

$$f(X,u) \rightarrow Y$$

La forma de leer este objeto (que todavía no es ninguna fórmula) es la siguiente: para una persona cualquiera con información previa X y cuyo azar (o cuestiones inobservables) fueron u, el modelo (f) predice que ocurrirá Y. Que si vale 1 significa que se predice que la persona pagará el crédito y que no lo hará si vale 0.

El principal objetivo de *machine learning* es explotar los datos pasados para contruir el modelo f, que predice de la mejor manera Y. "Construir el modelo" significa dar con una suerte de fórmula matemática que funcione para la predicción. Una vez que el modelo esté construido, podríamos "alimentarlo" con la información de la señora Manfredi y ver si el modelo predice que hay que darle el crédito (Y = 1) o no (Y = 0).

En la vieja visión de la estadística, la idea era *estimar* el modelo f, propuesto por una teoría o tal vez por la experiencia previa. El modelo venía de afuera del problema y los datos se usaban solo para estimarlo. La revolución de *machine learning* cambia por completo esta estrategia. La profusión de datos permite construir, estimar y reevaluar el modelo a medida que se lo usa. Esta es la idea de *aprender*, en vez de estimar. En términos de nuestro ejemplo, los datos iniciales de créditos y características de clientes se utilizan para construir una enorme variedad de modelos prototípicos, uno de los cuales se elige para predecir la capacidad de repago de clientes nuevos como la señora. Manfredi, tal como contamos que hizo Google con Google Flu Trends para predecir la intensidad de la gripe A. Con posterioridad, estos nuevos datos se usan para evaluar la *performance* del modelo y reconstruirlo adaptativamente.

Lo de *automático* tiene que ver con que una parte de (y a veces toda) la tarea de reconstrucción del modelo puede relegarse a un procedimiento computacio-

nal, que sobre la base de algún criterio puede ajustar de forma automática el modelo a la luz de nuevos datos e iterativamente hasta dar con un modelo con la mejor *performance*.

¿Dónde aparece big data en esta historia? La construcción automática de modelos complejos es altamente demandante en términos de datos. Cuanto más flexible sea el modelo y cuanto menos se conozca de él, más datos se necesitan para construirlo de forma confiable. Y es aquí donde la revolución de datos juega un rol crucial. Big data le permite a la estadística liberarse de su mero rol de estimar los modelos que otra disciplina le propone, y pasa a asumir la tarea de construirlos, evaluarlos y rediseñarlos, a través de la conjunción de algoritmos y datos masivos.

Ireneo Funes va a Harvard

Volviendo a las analogías rockeras, ¿fue realmente Elvis el padre del *rock*? Algunos expertos señalan a Bill Haley, otros a Ike Turner y su "Rocket 88" o, más atrás en el tiempo, a los viejos *bluesmen*. Y sin un criterio obvio, la búsqueda de las raíces del *rock* puede llevarnos al hombre de Cromañón, al menos a juzgar por su naturaleza rústica, afín a la cultura de un estilo musical que se refiere a un pedazo de piedra en su mismísima denominación. Y de forma análoga, el análisis de datos es tan viejo como la humanidad; es solo cuestión de imaginar a un antiguo nómade intentando predecir la lluvia mientras observa los movimientos de las nubes. Pero en los últimos dos siglos la estadística se estableció como una disciplina concreta, más allá de la matemática, nutriéndose de los progresos en el cálculo de probabilidades –en las

épocas de Gauss o Laplace– o, más recientemente, del arrollador avance de la computación.

Como dijimos, de ser una región, la estadística remite a una metrópolis políglota como Londres o Nueva York, y también a esas "tierras de nadie y de todos", como Ciudad del Este o Kabul, atestadas de lugareños, periodistas, militares, turistas, científicos, mercenarios, diplomáticos, trabajadores rurales y sospechosos banqueros trajeados. Yo mismo me sentí un forastero cuando llegué a la estadística desde la ciencia social, y durante años cargué con esa culpa de haber "entrado por la ventana". Hasta que varios años después me di cuenta de que se trata de una disciplina casi sin puertas, en la que muchos nos habíamos colado por ventanas, chimeneas o rendijas. Llamativamente, existen muy pocas carreras de grado en estadística, lo que contrasta con la relevancia de una disciplina presente en todas las ramas de la ciencia y la vida cotidiana. El "corpus" de la estadística mundial se compone de estadísticos propiamente dichos, y también de quienes vienen de la matemática, la computación, la ingeniería y de todas las disciplinas que usan datos de manera activa, desde las naturales, como la biología o la agronomía, hasta las humanísticas, como la lingüística, pasando por las sociales, como la economía o la sociología. A modo de ejemplo de esta auténtica Babilonia disciplinar que es la estadística, hace poco dicté una conferencia para estadísticos profesionales y cuando pregunté: "¿Cuántos de ustedes tienen un título de grado en estadística?", muy pocos levantaron la mano.

Como en una mala película de acción de sábado a la tarde, hace unos quince años la aparente calma de la estadística se vio alterada por la irrupción de una extraña tribu, atraída por el aluvión de datos de big data. Gente

de pantalones chupines de colores fuertes y modismos extraños comenzó a sacudir el ecosistema de los estudiosos de datos. Y así es como en vez de estadística se empezó a hablar de análisis, ciencia o minería de datos.

La historia de la ciencia es una historia de revoluciones. Los propios estadísticos plantaron su bandera en terrenos otrora de fulleros, adivinos y burócratas apiladores de datos: fue una disputa por un juego de dados lo que convocó a mentes brillantes como Pascal o Fermat a sentar las bases de la probabilidad. Y como tal, el espíritu revolucionario –fundamental para la ciencia– choca con su innata necesidad de autodefensa. Así es como ante la explosión de big data, la estadística clásica se siente como los protagonistas de "Casa tomada", el brillante y alegórico cuento de Julio Cortázar que relata las peripecias de dos hermanos que habitan una enorme casa y que, ante la supuesta presencia de extraños, día a día se recluyen en espacios cada vez más pequeños de la casona.

El profesor Stephen Stigler, de la Universidad de Chicago, ha dedicado toda su vida profesional al estudio de la historia de la estadística. En 2016 escribió un interesantísimo libro titulado *Los siete pilares de la sabiduría estadística* en el que pasa revista a importantes hitos históricos de esta disciplina. Con respecto al fenómeno que nos ocupa, dijo: "Funes es big data sin estadística", frase que tuvo un inmediato impacto en la profesión y causó una gran polémica.

El Funes en cuestión es Ireneo Funes, un extraño personaje del universo de Jorge Luis Borges, el insigne escritor argentino. Se trata de un muchacho con una memoria prodigiosa, que podía (y quería) recordar detalles insignificantes para cualquier otro mortal, a tal punto que reproducir los eventos de un día le tomaba…

¡24 horas! Lo llamativo en Funes es tanto su capacidad para recordar pormenores como su necesidad de hacerlo y su postura tercamente escéptica ante cualquier intento de abstracción. Según Borges, Funes opina que "pensar es olvidar diferencias, es generalizar, abstraer. En el abarrotado mundo de Funes no había sino detalles, casi inmediatos".

En relación con la invasión al campo de la estadística, la reacción de Stigler no deja de ser "pasivo-agresiva", como la del cumpleañero que una vez debajo de la piñata primero empuja a sus amiguitos para luego poder quedarse con más caramelos. Actitud esperable de una disciplina histórica que ve amenazada su hegemonía sobre el análisis de datos. Tal fue el revuelo, que el profesor Xiao Li Meng, director del prestigiosísimo Departamento de Estadística de la Universidad de Harvard, organizó un seminario llamado "Funes y big data" para discutir estas cuestiones de la pertinencia de la estadística a la luz de la potencial invasión de otros practicantes.

Por cierto, Funes es big data sin estadística; los datos por sí solos son cacofonía pura. Pero si la estadística –y, fundamentalmente, su enseñanza– no es capaz de avenirse a los nuevos tiempos, se quedará como el cumpleañero solitario recogiendo unos pocos caramelos del piso de su patio, y viendo cómo los otros niños corren a otros cumpleaños. Porque llover, llueve en todos lados. Y datos, ni hablar.

Da capo

"Algo viejo, algo nuevo y algo prestado" dice una máxima de los casamientos, que se aplica de manera idéntica a la tecnología. En este libro veremos que lo mejor de

big data aparece cuando confluyen los conocimientos de la tradición de la ciencia con las innovaciones.

Este capítulo intentó delimitar el ámbito de acción de tres ideas: big data, estadística y aprendizaje automático. Un punto central es que el concepto de big data va mucho más allá de lo que su etimología sugiere, en relación con el tamaño. *New* (nuevo), *more* (más) o *right* (correcto) data son términos que quizás describan mejor la naturaleza disruptiva del fenómeno. La estadística es la disciplina del aprendizaje a partir de datos, "culpable" de la mayoría de los avances científicos de los últimos dos siglos y omnipresente en la vida cotidiana. La nueva ciencia de datos no arregla un problema de la estadística, sino que explota los más recientes avances computacionales para aprovechar la oportunidad única que brinda la irrupción de datos masivos, producto de la interacción con dispositivos interconectados.

Hace poco armamos un grupo de estudio virtual para leer un reciente texto de estadística moderna, que por su aproximación cae decididamente en el ámbito de la ciencia de datos. El "catálogo" de disciplinas de las que provienen quienes acudieron a la convocatoria incluye: sistemas; biología; economía; estadística; matemática; física; ingeniería electrónica, mecánica e industrial; neurociencia; ciencias actuariales; contabilidad; derecho; farmacia; ciencia política; sociología; economía; medicina; urbanismo; periodismo; negocios e historia.

Este libro abona la idea de que la nueva ciencia de datos ofrece una oportunidad única de interacción entre disciplinas aparentemente disímiles, que tienen en común la necesidad de lidiar con información masiva. A los científicos honestos nos toca desarmar el Boca-River

de los datos, y crear una suerte de Beatles-Rolling Stones, que contrariamente a lo que muchos creen, eran buenos amigos, se admiraban y compartían la pasión por el *blues* y el *rock* de raíces. Y ajenos a las discusiones de sus fanáticos extremos, crearon música memorable.

2. Livin' la vida data

Historias de datos y algoritmos en la sociedad

> –Bien, no tenga miedo, vamos despacio. Lo que a Ud. le pasa es que se espanta con las cuestiones técnicas. Si no entiende chino y le empiezan a decir tonterías en chino, no tiene forma de saber cuánto de lo que no entiende tiene que ver con que le hablan en otro idioma y cuánto con que le hablan de cosas sofisticadas. Hagamos lo siguiente. Ud. parece tenerle miedo a la autopista de big data y a los algoritmos. Entonces, le propongo primero dar una vuelta por su barrio, y ver cómo estas cuestiones aparecen asociadas a cosas que Ud. conoce. Y una vez que les haya perdido el miedo, el tratamiento sigue con algunas experiencias más fuertes.
> ¿Qué opina?

La mayoría de los cursos sobre alguna técnica usan una curiosa estrategia pedagógica. Empiezan por la solución y luego pasan al problema, al revés de como ocurren las cosas en la práctica. Posiblemente ustedes noten que una silla está medio inestable, y, luego de darse cuenta de que se debe a un tornillo flojo, van en busca de un destornillador, y no al revés. Casi todos se inician en el análisis de datos a través de los algoritmos, los modelos, los lenguajes de programación, etc. Y la típica clase de una técnica comienza explicándola –muchas

veces en términos formales– y luego viene el ejemplo o la aplicación a la realidad. O sea, la clase sigue el proceso inverso al modo como ocurren las cosas en la vida diaria: primero el destornillador y después la silla y el tornillo. Hay dos problemas con esta estrategia pedagógica. El primero es que pone a la solución por encima del problema, y así muchos recién llegados al análisis de datos se "sobreentrenan" en herramientas y no en la detección de problemas relevantes, pasando por alto la habilidad más compleja e importante: elegir qué herramienta usar para cada problema. El segundo es que la técnica es formal y plantea una barrera alta a la entrada de los recién llegados.

La propuesta de este capítulo es hacer un breve recorrido por algunos usos relevantes de big data, algoritmos y estadísticas en problemas de la sociedad moderna. Será nuestro primer encuentro con las enormes ventajas de big data y aprendizaje automático, focalizando en el tipo de problema que pide a gritos la intervención de datos y algoritmos, y no tanto en los aspectos técnicos acerca de cómo se implementó la solución (de lo que nos ocuparemos en el próximo capítulo). "Cuando se tiene un martillo, todos los problemas parecen clavos", decía el psicólogo Abraham Maslow. El objetivo de este capítulo es inducirlos a pensar en los clavos, los tornillos y las tuercas del análisis de datos y no tanto en las herramientas. ¡No sea cosa que terminen martillando tornillos!

¡Que vuelvan los (iPhones) lentos!

Allá lejos y hace tiempo cualquier baile incluía un interludio de canciones románticas: los "lentos". Para muchos, la señal de que si algo tenía que pasar, era en ese

instante. Pero el tiempo, que todo lo puede, arrasó con ese *medley* de canciones pegajosas que nos recuerdan a Air Supply, A-ha y otros artistas que nos visitan con recurrencia, cada vez que se atrasa el tipo de cambio. Y cada tanto, en alguna reunión de egresados de la secundaria, se escucha a un nostálgico reclamar "¡que vuelvan los lentos!", grito de guerra de los que añoran un pasado que no volverá. O no tanto. *Boom* demográfico mediante, tal vez la masa de nostálgicos no sea tan pequeña, y quizás haya espacio para un pingüe negocio dirigido a este público. Y como no logro reprimir mis instintos analíticos, escribo "que vuelvan los lentos" en Google, y veo que en Chile más de 35 000 personas firmaron una convocatoria auspiciada por una empresa de *snacks*, pidiendo el regreso del empalagoso género musical. En definitiva, el reclamo parece tener una base sólida, lejos de ser el lamento aislado de algún pelado beodo en una reunión de egresados.

Bastante más acá en el tiempo, la empresa Apple parece encargarse puntualmente de que vuelvan los lentos. O al menos así opinan los usuarios, que cada vez que sale un nuevo modelo de celular sienten que el viejo se vuelve sospechosa y automáticamente más lento, como reclamando a gritos comprar el nuevo.

Hay varios puntos a dirimir en relación con esta cuestión. En primer lugar, verificar si la masa de gente que siente esta "lentificación" es lo suficientemente grande como para darle entidad al fenómeno, en línea con quien llama a todas las mamás del curso para ver si la descompostura estomacal de su hijo es producto de un atracón o de una intoxicación masiva en el cumpleaños del día anterior. En segundo lugar, es relevante dilucidar si en efecto los iPhones se vuelven más lentos o si solo se trata de una mera sensación, tal vez provocada

por la envidia y la necesidad de autojustificar la compra de un nuevo celular. Y, por último, es importante evaluar si este fenómeno se da también para otras marcas de teléfonos inteligentes.

La joven economista argentina Laura Trucco tradujo estas cuestiones en acciones concretas, cuando todavía era estudiante del doctorado en Harvard. Inteligentemente, Trucco notó que la percepción de lentitud de los iPhones debería reflejarse en búsquedas en Google que contuviesen conjunciones de palabras como "iPhones" y "lento", y construyó una base de datos de intensidad de búsquedas de estos términos.

Los datos muestran clarísimos picos cada vez que sale un nuevo iPhone. Es decir, la sensación de que el iPhone se vuelve más lento cada vez que sale uno nuevo no es una mera leyenda urbana, sino que tiene un correlato verificable en datos concretos que pueden ser estudiados sistemáticamente. Más aún, Trucco encuentra que este fenómeno es inexistente para marcas alternativas, como Samsung, que no presentan ningún pico de intensidad de búsqueda sobre lentificación tras la introducción de un modelo nuevo.

Este caso ilustra las posturas extremas que muchos tienen acerca del fenómeno de big data. Los que únicamente ven la mitad vacía del vaso señalan que, big data mediante, solo se ha podido avanzar un poquito en estas cuestiones. De hecho, varias preguntas relevantes quedan todavía sin respuesta, a saber:

1. si realmente los iPhones se vuelven más lentos o es una mera sensación, tal vez provocada por la envidia;
2. si aun existiendo "lentificación" se trata de una maniobra inescrupulosa de Apple o de una mera

> consecuencia del hecho de que la empresa actualiza de manera simultánea su *hardware* y también su sistema operativo, lo que, naturalmente, torna a los modelos anteriores más lentos.

Los que ven la mitad llena del vaso señalan que establecer que se trata de un fenómeno considerable, que afecta a Apple y no a Samsung, es un notable primer paso, que pudo ser estudiado de forma científica y reproducible sobre la base de algoritmos y datos inexistentes no hace mucho tiempo.

Estos "primeros pasos" ocupan un lugar central en la historia de la ciencia. Salvando las diferencias, el gran paso adelante en la historia de la epidemiología lo dio John Snow, cuando en 1823 mostró contundentemente que la transmisión del cólera por el agua no era una habladuría, sino una apreciación fundamentada en datos concretos, echando por tierra la hipótesis de transmisión por el aire. De forma análoga al fenómeno de los iPhones, la evidencia de Snow no puede ir más allá de establecer que el cólera se transmite por el agua y no por el aire (como se creía hasta entonces), pero sin explicar por qué. Las explicaciones causales y microbiológicas vinieron por otro carril, alentadas por el descubrimiento de Snow. Big data permite dar muchos "primeros pasos", en términos de descubrir patrones que posibilitan ir más allá de la evidencia anecdótica, sobre la base de un análisis preciso y ordenado como el de Laura Trucco. La falacia de la correlación dice que es imposible inferir causalidad de meras correlaciones, y a tal efecto el análisis de Trucco no es suficiente para distinguir entre explicaciones alternativas de la sensación de lentificación. Y así como preocupa caer en la confusión entre correlación y causalidad, también inquieta la

reciente instalación de una suerte de "metafalacia de la correlación" consistente en creer que ninguna correlación sirve para nada. Big data y *machine learning* se mueven en este delicado terreno intermedio, de correlaciones que no implican causalidad, pero que pueden sugerir patrones interesantes para el análisis.

Hace muy pocos días la empresa Apple reconoció públicamente el proceso de lentificación de sus iPhones y ofreció un sustancial descuento en el reemplazo de las baterías de los modelos viejos, en apariencia las causantes del problema. Nos encantaría creer que la reacción de Apple fue provocada por los resultados de Laura Trucco, pero hacerlo implicaría caer en la falacia de la correlación, y no hacerlo, en pensar que las correlaciones son inútiles. Tal vez sea interesante encarar una suerte de metainvestigación para ver cuál fue el efecto de la investigación de Trucco en el cambio de política de Apple.

Y un último ejercicio empírico es verificar si efectivamente las visitas oportunistas de grupos como A-ha o Air Supply se correlacionan con los atrasos cambiarios de nuestros países. Aunque no faltará quien proponga "que vuelvan los lentos" como medida de política para hacer subir el dólar.

Datactivismo, orden y progreso

En *On writing well* [Acerca de escribir bien], el escritor William Zinsser dice que la frase más importante de un texto es la primera. Todos reconocemos el comienzo de obras como *El Quijote* ("En un lugar de la Mancha, de cuyo nombre no quiero acordarme…") o *Cien años de soledad* ("Muchos años después, frente al pelotón de fu-

silamiento, el coronel Aureliano Buendía había de recordar aquella tarde remota en que su padre lo llevó a conocer el hielo."). Así, una linda recomendación para el escritor novato es releer los comienzos de 100 buenos libros de su biblioteca, buscando esa magia a la que se refiere Zinsser.

No parece algo difícil. A dos líneas por libro, si en una página entran unas 40 líneas, los comienzos de 100 libros ocupan cinco páginas de material. El problema es que si uno quisiera trabajar en un bar debería cargar los 100 libros elegidos o desarmarlos para quitarle a cada uno su primera página.

A nuestros fines, lo que importa es que la ejecución de esta tarea se choca con la forma en que la información está organizada: libros impresos. De tener la versión electrónica de los 100 libros, podríamos cargarlos en una *notebook*. Y de contar con los servicios de un programador, podríamos pedirle que nos arme un archivo de cinco páginas con los dos primeros renglones de cada libro; una tarea simple para cualquier profesional de la informática.

Esta cuestión ilustra un importante desafío para big data: si los datos están, pero no sistematizados convenientemente, es casi lo mismo que si no están. En nuestro ejemplo, si bien el material relevante entra en cinco páginas (las que contendrían los comienzos de los 100 libros), a menos que podamos apelar a la versión electrónica no hay forma de enfrentar la tarea sugerida sin tener que lidiar con los 100 libracos.

Creer que la información está por el mero hecho de que los datos existen es un serio error de principiante. Un punto crucial en la revolución de big data es que la falta de sistematización es la regla más que la excepción. Por su naturaleza espontánea, los miles de millones de

datos de big data jamás vienen ordenados en una tabla prolija como una planilla de cálculo. Por el contrario, la sistematización previa es muchas veces la más importante de las tareas, a tal punto que una buena plataforma de visualización –en forma de tablas o gráficos– puede llegar a reemplazar el análisis. Y es justamente una inspirada tarea de sistematización de información pública lo que le valió al programador Manuel Aristarán el ingreso al Laboratorio de Medios del Massachusetts Institute of Technology (MIT), tal vez la meca de la tecnología, los medios de comunicación y el diseño.

En 2010 Aristarán notó que la municipalidad de Bahía Blanca, su ciudad natal, usaba un sistema *online* en el que se podía consultar la ejecución del presupuesto público. Y que, si bien la información estaba disponible, el formato usado hacía muy difícil, cuando no imposible, responder preguntas muy elementales y útiles tales como "cuánto se gastó en publicidad en el último trimestre" o "qué reparticiones del municipio gastaron más". En no más de una semana creó "Gasto Público Bahiense" (GPB), una versión amigable del sistema oficial que permitía realizar consultas sistemáticas acerca de la forma en la que se ejecutaba el presupuesto de la ciudad del básquetbol y de Manu Ginóbili.

"Los datos son materia prima para la discusión. Con GPB, tengo la esperanza de que algún bahiense se alarme o se alegre por el dinero invertido en una cosa u otra y que se produzca la discusión que motive acciones transformadoras", escribió Aristarán en su blog en relación con los objetivos de GPB, consciente de que la transparencia abre una peligrosa ventana, tanto a lo más claro como a lo más oscuro de la gestión pública. Y el escándalo no tardó en llegar. Primero los insultos. "Aristarán es un tremendo mentiroso además de malin-

tencionado, un ignorante tecnológico", dijo un oscuro burócrata que se sintió afectado por las mismas acciones que condujeron a Manuel al prestigiosísimo MIT y a una exitosa carrera como programador, comunicador y actor social. Y a continuación, un rediseño de la página web del municipio, que ahora incluía un molesto CAPTCHA que complicaba la tarea de Aristarán.

La sigla se refiere a Completely Automated Public Turing test to tell Computers and Humans Apart (prueba de Turing completamente automática y pública para diferenciar ordenadores de humanos). Es decir, una serie de preguntas o acciones muy simples para un humano, pero casi imposibles para un robot como el programado por Aristarán. Por su lógica, el CAPTCHA es la kriptonita de los robots. Tareas tontas para una persona, como señalar en un conjunto de fotografías cuáles tienen partes de automóviles, son muy difíciles de sistematizar, y un ejemplo de las que se usan en el CAPTCHA para evitar que los robots computacionales accedan a cuentas de bancos u otros sitios con información delicada.

Poner un CAPTCHA para acceder a la información pública es tan contradictorio como pedir un certificado de buena conducta para ingresar a la mafia. Pero la mejor ilustración de la naturaleza oximorónica de esta situación es relatada por el propio Aristarán en su muy recomendable charla TED, cuando menciona que en una repartición oficial alguien se negó a brindarle información a un colega diciendo: "Ah, no, esa información no te la puedo dar porque es pública". Afortunadamente, luego de un par de años del lanzamiento de GPB, una nueva gestión municipal tuvo una actitud más colaborativa. Y así es como en 2012 –y por la iniciativa de Aristarán– nació en la lejana ciudad del sur bonaerense un pionero y exitoso caso de gobierno abierto.

Sin un sistema funcional que permita cruzar la información desde múltiples perspectivas, la disponibilidad *online* de datos de gestión es un gesto valioso pero inoperante, tanto como tener que cargar 100 libros en una carretilla al solo efecto de poder consultar los dos primeros renglones de cada uno.

Se habla con preocupación de la posibilidad de que varios trabajos sean reemplazados por algoritmos. Sin embargo, el episodio que involucra a Manuel Aristarán sugiere una sana convivencia con robots, que llevarán a cabo tareas que ningún humano podría, como la de sistematización de la información relatada en este capítulo, tan importante como su análisis. A nosotros nos tocará devolverles la gentileza a los amigos automáticos, enseñándoles a reconocer dibujitos en una pantalla, a cantar apasionadamente un bolero y a que se animen al *Martín Fierro* luego de leer "Aquí me pongo a cantar, al compás de la vihuela". Y a alzarse ante las arbitrariedades a las que nos someten varias instituciones, esas que ningún robot tendría ganas de integrar.

Un oasis de agua dulce en medio del mar de datos

En su canción "El padre Antonio y el monaguillo Andrés", Rubén Blades –el icónico salsero panameño– cuenta la trágica historia del cura pacifista Antonio Tejeira, asesinado a balazos junto a su monaguillo de 10 años mientras daba misa en un pueblito de El Salvador. Blades describe tierna y pícaramente al niño Andrés diciendo: "Le han dado el puesto en la iglesia de monaguillo, / a ver si la conexión compone al chiquillo; / y su familia está muy orgullosa, porque a su vez se cree / que con Dios conectando a uno, conecta a diez".

La premisa parece ser que ir a la iglesia hace bien. Y la pregunta obvia es si practicar una religión vuelve mejores a las personas, o si simplemente se trata de que aquellas con más inclinaciones por el bienestar comunitario son más propensas a las prácticas religiosas.

A fin de dilucidar esta cuestión, un experimentalista designaría aleatoriamente a un grupo de personas para que practiquen una religión y a otro no, de forma de ver cuál es el efecto causal de la religiosidad. Tanto por razones éticas como operativas, las cuestiones humanas tienen vedada esta vía de análisis, habitual en las ciencias biológicas. La alternativa más factible consiste en apelar a datos observacionales en vez de experimentales, es decir, datos que no fueron generados por ningún experimento explícito, sino que surgen de observar las acciones de las personas, como los que brotan a borbotones de big data. Un primer problema es que la comparación entre personas religiosas con las que no lo son no aporta mayor información, en el mismo sentido en que comparar el peso entre personas que hacen dieta y las que no puede llevar a la conclusión errónea de que hacer dieta engorda. ¿O acaso no es cierto que quienes hacen dieta son más gordos? Tampoco funciona comparar actitudes sociales antes y después de que una persona se vuelva religiosa; tal vez un tercer factor (la muerte de un familiar, por ejemplo) la ha llevado tanto a mejorar su relación con sus pares como a acercarse a alguna religión. A la luz de estas cuestiones, resulta difícil pensar en una arquitectura de datos no experimentales que permita echar luz sobre si es cierto que la religiosidad mejora a las personas.

Ahora bien, ¿qué garantiza un experimento en relación con los datos? Una separación clara entre causa y efecto. Por ejemplo, si dividiéramos en dos a un grupo de per-

sonas, le asignásemos al azar una droga a una mitad y a la otra no y luego les midiéramos a todos la temperatura corporal, la diferencia de temperaturas entre quienes tomaron la droga y quienes no debería ser ocasionada por los efectos de la droga, bajo ciertas condiciones simples. Es decir, en este contexto lo único que distingue a las personas de los dos grupos es que unas tomaron la droga y las otras no, de ahí que las diferencias de temperatura deberían atribuirse a la única cosa que ha cambiado entre ambos grupos. Por el contrario, si en vez de una elección al azar, la droga es asignada a personas que manifestaron tener fiebre, la comparación de temperaturas no es válida ya que se confunden los efectos de la droga con los de la enfermedad, que hace subir la temperatura corporal.

Técnicamente, cuando la causa se mueve de forma ajena al resultado del experimento se dice que se mueve exógenamente. A modo de ejemplo, la asignación de droga al azar implica una variación exógena, mientras que si asignamos la droga solo a las personas que sienten fiebre, en este caso la causa se mueve endógenamente, es decir, en relación con el efecto que se quiere medir. Usando esta jerga, la principal contribución de un experimento es garantizar que la causa se mueve exógenamente. En el experimento es el azar lo que garantiza que la droga es la única causa de variación en la temperatura. En la práctica, la gente no anda tomando ibuprofeno porque sí, sino cuando tiene fiebre, de ahí que los datos observacionales no permiten aislar con claridad las causas de los efectos. De hecho, un día cualquiera, la relación entre la temperatura corporal de las personas y la cantidad de ibuprofeno que tomaron es positiva, no porque el ibuprofeno haga subir la temperatura sino porque quienes tenían fiebre lo tomaron y quienes tenían temperatura normal no.

Una de las enormes oportunidades que brinda el paradigma de big data es que, si se es muy cuidadoso, el enorme océano de datos tal vez permita aislar un subconjunto de información que, si bien no proviene de ningún experimento explícito, puede comportarse como si lo fuese. Y esta fue la compleja tarea que encararon los investigadores argentinos Nicolás Bottan y Ricardo Pérez Truglia a fin de estudiar si es realmente cierto que la práctica religiosa hace mejores a las personas.

El objetivo, entonces, consiste en buscar en el océano de datos información de actitudes sociales y religiosidad cuya correlación pueda interpretarse como causal, es decir, como si proviniese de un experimento. Una tarea bien difícil, como veremos.

El 23 de agosto de 2003 el exsacerdote y abusador serial John Geoghan murió en una cárcel de máxima seguridad del estado de Massachusetts, estrangulado y golpeado por otro preso. Geoghan cumplía una condena por abusar de más de 130 niños durante sus casi treinta años de servicio. Cualquiera se horroriza ante este tipo de noticia, de hecho, los sucesos en los que se vio involucrado Geoghan devinieron en un escándalo mediático mayúsculo en 2002.

Bottan y Pérez Truglia parten de la premisa razonable de que la difusión mediática de estos aberrantes episodios tiene un impacto negativo sobre la religiosidad de las personas. Y también notan que el *timing* de los abusos no sigue ningún patrón temporal obvio, es decir, ocurren en fechas más o menos azarosas. Consecuentemente, y en los términos antes descriptos, la difusión mediática de escándalos sexuales que involucran a curas debería producir una variación exógena en la religiosidad. Supongamos, por ejemplo, que en un barrio ocurre un escándalo sexual que involucra a sacerdotes: esto provocaría una caída en

las prácticas religiosas (la gente va menos a misa o desiste de anotar a sus hijos en colegios religiosos), y también un cambio en los comportamientos sociales promovidos por los valores religiosos (donaciones, participación en actividades caritativas, etc.). Esta caída en la religiosidad debería interpretarse como si proviniese de un experimento. ¿Por qué? Porque los abusos no ocurren siguiendo un patrón temporal, es como si ocurriesen al azar a lo largo del tiempo. ¿Qué rompería la naturaleza exógena de estos cambios? Por ejemplo, que los casos de abuso ocurriesen como reacción a la falta de compromiso social de la gente, es decir, si (macabramente) los religiosos saliesen a cometer abusos como reacción a la falta de religiosidad. En esta circunstancia la causa (el cambio en la religiosidad) no se mueve de forma exógena sino endógena, es decir, en relación con el efecto que se quiere medir (el comportamiento social), como cuando un analgésico no es asignado al azar (exógenamente), sino solo a las personas que tienen fiebre (endógenamente). Es la naturaleza azarosa del *timing* de los casos de abuso justamente lo que garantiza que estos son una auténtica causa y no una consecuencia. Es importante aclarar que lo de "azaroso" se refiere a *cuándo* ocurren los abusos, no hay nada azaroso en que se comentan abusos.

Ahora, hay dos problemas. Uno es cómo medir variables tan difusas como "grado de religiosidad" o "actitudes sociales". El otro es reunir una cantidad de eventos suficiente como para que los resultados tengan cierta credibilidad estadística. Bottan y Pérez Truglia encararon esta ciclópea tarea. En primer lugar, construyeron una masiva base de datos que permite identificar 3024 episodios de abuso sexual en los Estados Unidos, para el período 1980-2010. Un delicadísimo trabajo de campo les permitió identificar con precisión la fecha y también el código postal del

barrio donde ocurrió cada uno de estos aberrantes episodios, acudiendo a registros administrativos, diarios *online* y herramientas electrónicas como Google Maps.

Los autores exploran muchas formas de medir "religiosidad" o "actitudes sociales o comunales". A modo de ejemplo, uno de los indicadores de religiosidad se construye sobre la base de la matriculación de niños en colegios católicos, información obtenida de un censo de escuelas regularmente implementado por las autoridades educativas de los Estados Unidos. Variables "prosociales" como la contribución a entidades benéficas pueden medirse también sobre la base de encuestas. En síntesis, un milimétrico trabajo conceptual permite construir una base de 3024 casos en que se observa el grado de religiosidad y la intensidad de las actitudes "prosociales", antes y después de cada uno de estos eventos lamentables. Una vez más, es la naturaleza azarosa del *timing* de estos episodios lo que permite aseverar que estos cambios en la religiosidad fueron provocados por sucesos aberrantes de forma completamente exógena al evento que se trata de medir: las actitudes prosociales.

Los resultados del estudio son sorprendentes. En primer lugar, Bottan y Pérez Truglia encuentran que, efectivamente, los escándalos sexuales tienen un impacto negativo en la participación en actividades religiosas. Sin embargo, esta caída en la participación activa en la religiosidad no tiene efectos considerables sobre las actitudes prosociales ni sobre las creencias religiosas de las personas. Es decir, contra lo que la hermosa canción de Blades sugiere, la religiosidad per se no parece cambiar las actitudes prosociales de las personas. Los autores someten estas conclusiones a una enorme batería de tests y mediciones alternativas, como puede consultarse en el puntilloso estudio científico publicado en una prestigiosa revista internacional.

Esta historia ilustra elocuentemente una de las principales oportunidades del paradigma de big data: es un trabajo inteligente y meticuloso el que permite ir de una masiva cantidad de datos anárquicos y en apariencia inconexos (de censos, encuestas, diarios *online*, etc.) a un subconjunto pequeño pero que puede ser estudiado como si hubiese provenido de un experimento, aun cuando dicho experimento jamás fue implementado. Los 3024 datos del estudio parecen ínfimos en relación con los peta o zetta bytes mencionados en el capítulo 1, si bien provienen de una copiosa cantidad de información (que nadie dudaría en calificar como "de big data") que luego de un meticuloso trabajo de sistematización pudo ser "curada" para medir un fenómeno concreto. Entonces, una de las principales ventajas de big data es que el océano caótico de datos puede contener alguna dosis de datos puros y cristalinos que pueden echar luz sobre cuestiones complejas como las aquí relatadas.

La canción de Blades concluye diciendo que "Antonio cayó, hostia en mano y sin saber por qué; / Andrés se murió a su lado sin conocer a Pelé". Por el contrario, Geoghan supo exactamente por qué murió. Y fue la naturaleza aleatoria de la frecuencia de sus aberrantes hechos lo que le permite al analista detectivesco encontrar en la masividad de big data un canal preciso que conecta la causa (la religiosidad) con el efecto (las prácticas sociales, los valores).

Big data y la medición de la pobreza en Ruanda

¿Qué es la pornografía? "No sé, pero la reconozco cuando la veo", respondió en 1964 Potter Stewart, entonces juez de la Corte Suprema de los Estados Unidos. La frase

de marras es repetida hasta el hartazgo en relación con fenómenos fáciles de percibir pero elusivos en cuanto a definiciones precisas, como la obscenidad, la inteligencia o la pobreza. Y respecto de esta última, que no exista una forma obvia de medirla es una consecuencia directa de que no hay ninguna manera trivial de definir qué significa ser pobre. Pero más allá de esta apreciación, y en línea con los pensamientos de Stewart, lo que está fuera de discusión es que la pobreza existe y es un flagelo persistente que afecta la vida de muchísimas personas, y que es (o debería ser) la preocupación central de la política social.

La medición moderna de la pobreza surge de una suerte de acuerdo social, técnico y comunicacional, que sopesa las ventajas y desventajas de miles de formas de definir qué es ser pobre. "Miles" no es una exageración. Un estudio de Miguel Székely y Nora Lustig reporta que, aun restringiendo el análisis a las mediciones sobre la base de ingresos, una simplificación decididamente grosera, existen unas 6000 (sí, 6000) formas de medir la pobreza, que surgen de considerar las distintas alternativas involucradas en las fórmulas de pobreza.

El monitoreo del bienestar es un problema complejo y urgente, máxime para las regiones del mundo más castigadas por este flagelo. En países de desarrollo intermedio, como los de América Latina, la medición de la pobreza se hace sobre la base del llamado "enfoque de líneas", que consiste en cotejar el ingreso de un hogar contra una *línea de pobreza*: el valor monetario de una canasta de bienes que una familia tiene que poder comprar para dejar de ser pobre. Entonces, este enfoque, llamado "de incidencia", requiere relevar los ingresos de las familias y los precios de los bienes de la canasta. En la Argentina se usa la Encuesta Permanente de Hogares, que se realiza cuatro veces por año y, entre

otra información socioeconómica, pregunta cuáles son los ingresos por hogar. Los precios de los bienes de la canasta se relevan a través de encuestas de precios. Sin entrar en detalles, se trata de una costosa tarea que requiere un gran aparato institucional a fin de garantizar tanto que las cifras resultantes sean creíbles y representativas de la población, como comparables entre regiones y a lo largo del tiempo.

África central enfrenta una situación compleja, no solo por la severidad y persistencia de la pobreza extrema, sino también por su debilidad institucional, que hace imposible pensar en encuestas sistemáticas. En 2015, Joshua Blumenstock, Gabriel Cadamuro y Robert On publicaron un estudio en la prestigiosa revista *Science*, que ilustra el enorme potencial que tiene big data con respecto a estas cuestiones. Ruanda es un país extremo en términos de pobreza, azotado por un pasado de terribles guerras. Así y todo, Blumenstock y sus coautores notaron que el uso de teléfonos celulares en este país está bastante más extendido que lo que muchos inferirían de su historia de privaciones. O por lo menos lo suficiente como para que exista una relación relevante entre la intensidad de uso de celulares y el bienestar.

La tarea que encararon es la siguiente: "maridaron" una extensísima base de datos de llamados de teléfonos celulares con una pequeña encuesta a 856 personas, a las cuales se les realizaron varias preguntas a partir de las cuales se construyó un índice de bienestar para cada persona. Posteriormente apelaron a sofisticados métodos de aprendizaje automático a fin de construir un modelo que permite predecir el bienestar para cada una de las 856 personas, sobre la base de la intensidad de uso de celulares. Luego de una extensa evaluación, el modelo fue utilizado para predecir el bienestar del resto

de los ruandeses, para quienes se observa información de uso de celulares, pero no de su bienestar.

Esta estrategia permite construir un mapa de pobreza para todo el país africano con una elevada "granularidad", o sea que la medición puede realizarse a nivel individual, ya que predice el bienestar de una persona sobre la base de su consumo de telefonía celular. Asimismo, esta estrategia permitirá monitorear la evolución temporal de la pobreza en Ruanda y también recalibrar el modelo a la luz de mejores encuestas de bienestar o con más datos.

Este caso muestra claramente las ventajas de big data en estas delicadas cuestiones sociales, y también la relevancia de la interacción entre datos masivos (como los provenientes del uso de celulares) con encuestas tradicionales, como la implementada en este caso para medir el bienestar. Desde un punto de vista técnico, la estrategia empleada en este caso es virtualmente idéntica a la usada por Google Flu Trends: una pequeña base de datos (la encuesta de bienestar en Ruanda, en el primer caso, y las estadísticas oficiales sanitarias, en el segundo) es puesta a interactuar con una masiva fuente de información: el uso de celulares en Ruanda y la búsqueda de términos relacionados con la gripe A.

Mucho se habla en los medios de que big data reemplazará a la estadística tradicional. El caso de la pobreza en Ruanda y el de Google Flu Trends sugieren lo contrario: que hay mucho por ganar de la interacción entre la disponibilidad de datos masivos y las encuestas sistemáticas implementadas con medios tradicionales. Ya dijimos que el futuro no es de David contra Goliat, sino de ambos interactuando en pos de un objetivo común.

Da capo

Los cuatro casos de este capítulo tienen algo en común: directa o indirectamente se basan en información masiva y espontánea, propia de big data. Pero difieren en la forma en la que extraen o usan esa información, ilustrando distintas facetas y potencialidades del análisis de datos. En el caso de los iPhones lentos, el principal uso del combo big data/algoritmos es como herramienta de reconocimiento de patrones, en el límite de la tecnología y lo social. Se trata de una crucial tarea dentro de la ciencia de datos; retomaremos esta idea, con más ejemplos, en el capítulo 4. El ejemplo de gobierno abierto en Bahía Blanca destaca el hecho de que una importantísima tarea en big data es la sistematización de datos, muchas veces tanto o más relevante que el propio análisis estadístico. El caso de los efectos sobre la religiosidad sugiere que un crucial rol de big data es funcionar como una suerte de "fuente primaria" de información cruda, que, con el debido procesamiento, puede producir datos limpios y ordenados como los que obtendría un agrónomo a través de un experimento. Finalmente, el caso de medición de la pobreza sugiere que big data y sus algoritmos pueden complementar y quizás sustituir los mecanismos tradicionales de relevamiento estadístico.

La mayoría de los escritos laudatorios sobre big data se refieren a sus logros en términos de predicción o reconocimiento de patrones. Este capítulo muestra que el potencial de big data va mucho más allá de lo meramente descriptivo, pues resulta útil en roles clásicos del análisis de datos, como la evaluación de relaciones causa-efecto, o la elaboración de estadísticas públicas.

Habrán notado que nuestros cuatro casos aluden muy tangencialmente a métodos estadísticos o algoritmos.

Por el contrario, hemos dicho cosas elípticas tales como "usando sofisticados algoritmos…" o "sobre la base de un modelo" cada vez que la historia pasó cerca de alguna técnica, porque prometí que focalizaríamos en los problemas antes que en los métodos. Y esa será la tarea de nuestro próximo capítulo. No se me van a achicar ahora, ¿no?

Índice

Este libro (y esta colección) **11**

Agradecimientos **17**

Introducción acuífera **19**

1. Perdidos en el océano de datos. Big data, aprendizaje automático, ciencia de datos, estadística y otras yerbas **23**
El Elvis Presley de la ciencia de datos (vida, muerte, resurrección y nueva muerte de Google Flu Trends) 24
¿De qué hablamos cuando hablamos de big data? 29
Los amplificadores de big data van hasta 11 33
La máquina de aprender 37
Ireneo Funes va a Harvard 40
Da capo 43

2. Livin' la vida data. Historias de datos y algoritmos en la sociedad **47**
¡Que vuelvan los (iPhones) lentos! 48
Datactivismo, orden y progreso 52
Un oasis de agua dulce en medio del mar de datos 56
Big data y la medición de la pobreza en Ruanda 62
Da capo 66

3. Una nueva ferretería para el aluvión de datos. Herramientas, técnicas y algoritmos **69**
Ordenando el "segundo cajón de la cocina" (análisis de clústers) 70
Los Rolling Stones del análisis de datos (regresión) 76
Nadie zafó del hundimiento del *Titanic* (árboles decisorios) 85
Da capo 94

4. Gran Hermano, gran data. Datos y algoritmos hasta en la sopa **95**
El desafío Netflix del millón de dólares 97
Letra de médico (OCR) 104
Revoleando piedrazos con la mano invisible 109
Nga këto plazhe të bukura 114
Da capo 119

5. Cajas negras para magia blanca. Más herramientas para el aprendizaje automático **121**
Pescar en una pecera (complejidad y regularización) 122
El test de Chuck Norris (validación cruzada) 128
La leyenda de Ícaro (la maldición de la dimensionalidad) 130
Aprendizaje profundo (redes neuronales) 133
Da capo 137

6. No todo lo que brilla es oro. La letra chica de los datos y los algoritmos **139**
Señor, su hija está un poquito embarazada: datos y privacidad 140
Porno impuestos en Noruega: datos y transparencia 144
Millones de moscas no pueden estar equivocadas: big data y poca información 148
El "efecto Styx": datos y sesgos de uso 155

La datamanía cada tanto encuentra hombres embarazados: big data y la falacia de la correlación 159
Revoleando *bitcoins* para dirimir cuestiones sociales: datos, algoritmos y comunicabilidad 164
Da capo 168

7. Puedo ver crecer el pasto. El futuro del futuro de los datos 171
Big data no es todos los datos 172
¿Quiero tener un millón de amigos? 176
Right data 181
Titanes en el ring de los datos 185
Da capo 190

Comentarios finales, ya sobre tierra firme 193

Referencias comentadas 197

Bibliografía comentada 201