walter sosa escudero

borges, big data y yo

guía nerd (y un poco rea) para perderse en el laberinto borgeano

siglo veintiuno
editores

Índice

Este libro (y esta colección) **9**

Introducción **15**

1. Borges y la kryptonita de la ciencia de datos. "Funes el memorioso", "Del rigor en la ciencia" **29**
Funes es big data sin estadística 32
¡Mamá, mamá, mi promedio armónico es 5,8065! 36
La leyenda de Keith Richards: la parte por el todo 41
El mapa de Tomi y el de John Snow 44
Borgesdata 50

2. Cinco problemas para Jorge Luis Borges. "Evaristo Carriego", "Pierre Menard, autor del Quijote", "Emma Zunz" **55**
¿Qué tienen en común Gauss, Borges y Evaristo Carriego? 58
¿Qué edad tenía Funes a su muerte? 62
¿Quién escribió *El Federalista*? ¿Hamilton o Madison? 67
¿Quién escribió el Quijote? ¿Cervantes o Pierre Menard? 70
¿Culpable o inocente?: Emma Zunz 74
Borgesdata 78

3. Big databorges. "La biblioteca de Babel" **81**
Tolstoi va a Calcuta: traducciones automáticas como "cortar y pegar" 84
Vaciando el *Titanic* con una cuchara: buscar y encontrar en "La biblioteca de Babel" 88
Seño, seño, Borges escribió "culo" 92
Correlaciones espurias en la biblioteca de Babel 95
Borgesdata 99

4. Al infinito y más acá. "El jardín de senderos que se bifurcan" 103
Perdidos en un laberinto de correlaciones espurias 106
Experimentos: los científicos entran al jardín de senderos que se bifurcan 111
¿Qué es más grande: un censo o una muestra? 114
Probabilidades: un jardín de datos que se bifurcan 121
Borgesdata 127

5. El cerebro mágico de Borges. "El idioma analítico de John Wilkins", "El Golem", "Ajedrez" 133
Recuerdos del futuro: Borges y Herbert Simon, a comienzos de los setenta 136
Mariekondomanía y progreso: los *clusters* y la arbitrariedad del orden inalterable 139
Golems y algoritmos 144
Jaque mate al golem 147
Inteligencia artificial: el Golem aprende a jugar al ajedrez 149
Borges y la divulgación científica 153
Borgesdata 159

Epílogo. Un Aleph de 3 cm 163

Referencias 169

Agradecimientos 173

Nube de palabras del libro: el tamaño de la letra representa la cantidad de veces que cada término aparece

1. Borges y la kryptonita de la ciencia de datos

"Funes el memorioso",
"Del rigor en la ciencia"

"Una idea, un *paper*, dos ideas, dos *papers*", me dijo una vez Rolf Mantel, uno de mis mentores. "*Paper*" es la forma rápida de llamar a un artículo científico de investigación, de pocas páginas y contenido directo. Un error de investigador novato es atiborrar al lector con varias ideas que compiten entre sí, cuando posiblemente lo mejor sea separarlas en distintos escritos. Y como esta es nuestra primera incursión en el universo de Borges y los datos, les propongo una suerte de danza iniciática alrededor de una sola idea, central en el mundo de la estadística y en el de Borges: la tensión entre la realidad y sus representaciones. ¿Y qué implica ese tema, esa danza? El gran desafío del "análisis de datos" no es dar cuenta de los datos sino de lo que los datos quieren decir; los datos son meras manifestaciones de las verdades que los generan. Es decir, si en la radio dicen "son las 10.15 hs y la temperatura es de 23 ºC", cualquier mortal intenta entender qué quiere decir ese dato, qué pretende implicar, para decidir cómo vestirse o si pintar una pared al aire libre. Lo mismo ocurre con las cifras de pobreza o de la cantidad de infectados durante la pandemia de Covid-19: el desafío está en lo que los datos significan, más allá de ellos mismos.

En esta primera incursión en el universo de Jorge Luis Borges exploraremos "Funes el memorioso" y "Del rigor en la ciencia", dos relatos centrales de su obra. Como dijimos en la Introducción, Funes es un muchacho con una memoria prodigiosa pero incapaz de "ver" a través de los datos: para Funes los datos son tanto un punto de partida como de llega-

da y, por consiguiente, él es algo así como la negación misma de la estadística y de la ciencia, que pretenden ir de los datos a lo que estos quieren significar. Para Funes los datos son tan solo datos, que no implican ni son implicados por nada. Amigos de Funes parecen ser los cartógrafos de "Del rigor en la ciencia" que, en búsqueda de la perfección, puestos a hacer un mapa de un imperio terminan haciendo uno ¡del mismo tamaño que el imperio! Un mapa tan preciso como inútil.

Entonces, empezamos de modo simple. Estos dos relatos cumplirán el doble propósito de meternos en el universo de Borges y en el de los datos. Ya más en confianza, en los próximos capítulos abriremos el abanico de ideas y de relatos de Borges. Todo a su tiempo.

Funes es big data sin estadística

> Nosotros, de un vistazo, percibimos tres copas en una mesa; Funes, todos los vástagos y racimos y frutos que comprende una parra. Sabía las formas de las nubes australes del amanecer del treinta de abril de mil ochocientos ochenta y dos y podía compararlas en el recuerdo con las vetas de un libro en pasta española que solo había mirado una vez y con las líneas de la espuma que un remo levantó en el Río Negro la víspera de la acción del Quebracho.

Como ya hemos mencionado, Ireneo Funes es el protagonista de "Funes el memorioso", un cuento central en la obra de Borges. Poseedor de una memoria prodigiosa, Funes podía (y quería) recordar detalles insignificantes para cualquier otro mortal, tanto que reproducir los eventos de un día le tomaba… ¡veinticuatro horas! Lo llamativo en Funes es su capacidad para recordar detalles pero también su necesidad de hacerlo, y su postura terca y escéptica ante cualquier intento de abstracción. Funes opina que "pensar es

olvidar diferencias, es generalizar, abstraer". Sigue Borges: "En el abarrotado mundo de Funes no había sino detalles, casi inmediatos".

El siguiente ejemplo ilustra la reticencia de Funes a la abstracción y da una contundente muestra del finísimo humor de Borges, que repetidas veces encontraremos en este libro y del que ya les advertí en la Introducción:

> En lugar de siete mil trece, decía (por ejemplo) *Máximo Pérez*; en lugar de siete mil catorce, *El Ferrocarril*; otros números eran *Luis Melián Lafinur*, *Olimar*, *azufre*, *los bastos*, *la ballena*, *el gas*, *la caldera*, *Napoleón*, *Agustín de Vedia*. En lugar de quinientos, decía nueve. […] Yo traté de explicarle que esa rapsodia de voces inconexas era precisamente lo contrario de un sistema de numeración. Le dije que decir 365 era decir tres centenas, seis decenas, cinco unidades: análisis que no existe en los "números" *El Negro Timoteo* o *manta de carne*.

La necesidad de lidiar con datos es tan vieja como la información y las sociedades y, en consecuencia, tanto lo es la estadística. Pero en el siglo XX es cuando alcanza su mayoría de edad, de la mano de Ronald Fisher, uno de sus padres fundadores. La opinión pública y la economía en las disciplinas sociales, y la experimentación en las ciencias naturales, avanzan rápidamente al finalizar la Segunda Guerra Mundial, y en la década de 1950 se afianza la estadística como disciplina científica. Los enormes progresos tecnológicos que bajan los costos de recolección y procesamiento de datos, el surgimiento de la computación personal y, en nuestros días, la internet, tuvieron un fuerte impacto en esa trayectoria. Así y todo, la estadística como ciencia ha vivido aislada, a un costado de la matemática, valiéndose de los avances computacionales e interactuando con las disciplinas que la utilizan más enfáticamente, como la biología, la psicología, la economía, la agronomía o la meteorología.

Pero en los últimos años las cosas han cambiado radicalmente, producto de las interacciones electrónicas que dejan "huellas digitales" que pueden ser usadas como datos. A modo de ejemplo, escribo este capítulo en un bonito café de Buenos Aires en mi computadora personal conectada a internet. Sin que yo haga nada, el "gran hermano" se dio cuenta de dónde estoy, y, a través de las redes sociales a las cuales pertenezco, está enviando información sobre mi edad, ubicación geográfica, gustos musicales y otros hábitos. Datos, datos y más datos.

"Big data" es la designación, cada vez más usual, para esta profusión de datos generados por dispositivos interconectados, como las computadoras, los teléfonos celulares, la tecnología de localización geográfica (GPS), las tarjetas de crédito o cualquier cosa que por su operatoria tenga que interactuar electrónicamente con otra.

Esta "piñata de datos" excede el continente de la estadística, que, en relación con los datos mismos, y abusando de la imagen anterior, parece un párvulo que con un cucharón intenta atrapar los cientos de caramelos que caen del techo. Así, la computación, la matemática, la comunicación, la ingeniería, el diseño y muchas otras disciplinas se disputan el dominio de big data, como niñitos en un cumpleaños, desesperados y agitados debajo de la lluvia de golosinas.

Pero a la larga, el enorme desafío de la estadística es uno de "señal y ruido", en el cual más datos son la mejor opción solo si contribuyen a mejorar la calidad de la señal y no a aumentar el barullo. Un día, Daniel Heymann, mi entrañable profesor de macroeconomía, en el medio de una considerable crisis económica me dijo: "'Cuchame, pibe, ¿te imaginás el quilombo que se armaría si hubiese una medición diaria del PBI?" ("'cuchame", y no "escuchame", era la señal que anticipaba la irrupción de una de las máximas épicas por las que Heymann, un sabio, es admirado por todos los economistas). En casi todos los países, el PBI se mide una vez al año, porque es costosísimo hacerlo y porque lo que se pretende medir se

mueve con relativa lentitud. Si en la Argentina cada vez que se publica ese indicador se arma un escándalo mediático de proporciones, cuesta imaginar qué sucedería si se pudiese monitorear en tiempo real, como si fuese la cotización del dólar o la temperatura de una ciudad. La cuestión, entonces, es que lo que se pretende del PBI es que sea un *resumen* confiable y estable, a lo cual los movimientos de cortísimo plazo solo agregan ruido. En el mismo sentido, a las personas hipertensas les recomiendan que se midan la presión arterial con cierta frecuencia (tal vez diaria) y no cada cinco minutos, o a las personas sanas se les pide un análisis de sangre anual.

Y en este contexto, el vendaval de datos de big data es una buena noticia solo si contribuye a mejorar la señal sin aumentar el ruido, o si es capaz de proveer información útil allí donde antes había solo silencio y oscuridad. En definitiva, las ventajas de big data no vienen de la masividad de datos *per se* sino de que se los observe a través de alguna tecnología o modelo que permita usarlos y notar si es cierto que contribuyen más al orden que al caos.

"Big data es Funes sin estadística", escribió Stephen Stigler –profesor de la Universidad de Chicago, y máxima autoridad en la historia de la materia–, en un reciente libro titulado *Los siete pilares de la sabiduría estadística*. La frase, casi un tuit, tuvo un impacto inmediato y causó una enorme polémica en la profesión.

Puesta en contexto, la frase de Stigler tiene un mensaje claro, escéptico y socarrón en relación con big data: a contramano de la postura de Funes, los datos en sí mismos no tienen entidad y su masividad se traducirá en una mejora solo a través de algún tipo de procesamiento que, como anticipamos, pueda ver más allá de ellos. Stigler relativiza el fenómeno de big data y lo subsume a la estadística, actitud esperable de quien milita en una disciplina histórica, que ve amenazada su hegemonía sobre el análisis de datos.

Ni lerdo ni perezoso, y a raíz de los dichos de Stigler, Xiao Li Meng, director del departamento de Estadística de la

Universidad de Harvard (quizás el más prestigioso del mundo), dictó un curso cuyo título es, precisamente, "Ireneo Funes y Big Data", para revisar las relaciones existentes entre el personaje de Borges, el reciente libro de Stigler y el análisis de datos. El programa del curso es fascinante, ya que invita a la discusión tanto técnica como filosófica del fenómeno de big data en la sociedad. A contramano de las propuestas de éxito cortoplacista de algunas sospechosas instituciones educativas (muchas de ellas en el ámbito de ciencia de datos), ese programa de Meng dice que "este curso está destinado a aquellos alumnos cuya felicidad pasa por entender los fundamentos filosóficos del razonamiento bajo incertidumbre", promesa que habría hecho huir despavorido al Funes de Borges, suspicaz de las elucubraciones de todo tipo.

¡Mamá, mamá, mi promedio armónico es 5,8065!

"¿Qué hora son, Ireneo?", pregunta Bernardo, primo del narrador, en el cuento. "Faltan cuatro minutos para las ocho", responde "el cronométrico Funes" sin consultar reloj alguno, provocando el asombro de Borges. Habilidad característica de este pobre muchacho (en el momento del relato tiene unos 20 años), capaz de percibir detalles ínfimos donde el resto de los mortales solo ve trazos gruesos.

> Una circunferencia en un pizarrón, un triángulo rectángulo, un rombo, son formas que podemos intuir plenamente; lo mismo le pasaba a Ireneo con las aborrascadas crines de un potro, con una punta de ganado en una cuchilla, con el fuego cambiante y con la innumerable ceniza, con las muchas caras de un muerto en un largo velorio.

Hay dos rasgos llamativos en la actitud de Funes. El primero es la capacidad de medir el tiempo sin reloj y con precisión

asombrosa. Y el segundo es la necesidad de comunicarlo con exactitud similar. Aun munido de un reloj con agujas –como los que circulaban allá por 1887, el año del relato–, cualquiera de nosotros habría redondeado a "las ocho menos cinco"; sin embargo, y sin reloj, Funes elige decir la hora con una precisión de minutos. Pero si bien sorprendente, llama la atención que la precisión del joven "cronométrico" no parece infinita ni bastante menos, por lo menos si nos avenimos a las puntillosas descripciones de Borges en el cuento.

El tiempo transcurre en forma continua, de modo que la precisión de un reloj podría ser infinitamente pequeña, al menos en teoría. En 2005, un poco más de cien años después de los eventos que involucran a Funes y Borges, la reconocida empresa Tag Heuer lanzó con bombos y platillos el modelo de reloj pulsera más preciso de su historia, el "Calibre 360", que permite monitorear el tiempo con una precisión de una centésima de segundo, muy lejos del "Chaque heure pour la minorie", el pretencioso reloj sobre el que bromeaba Les Luthiers. De acuerdo con el genial grupo humorístico-musical argentino, ese "¡flor de relós!" venía "para la dama y el caballero, / con minutero y con segundero" y garantizaba el ascenso social y las conquistas amatorias. Ahora bien, nuestro Funes no consulta ningún reloj, de modo que es una incógnita si puede percibir el tiempo con precisión de minutos o más finamente. Puede que sea cierto que el memorioso perciba el tiempo en forma continua (acorde a sus asombrosas habilidades), pero elige comunicarlo en forma discreta (en minutos), aviniéndose a una convención social, lo cual requiere un mínimo ejercicio de abstracción (un redondeo), actividad de la cual reniega manifiestamente. O tal vez su grado de precisión no es infinito, y su percepción del tiempo en minutos y sin un reloj, aun sorprendente, solo lo deja un poco más allá de lo que podría hacer cualquier persona, pero lejos de la performance del "Calibre 360".

En definitiva, no queda claro cuánto es que Funes no abstrae porque no quiere o porque no puede. El propio Borges