9. Cartografías de la culturómica

Mario Kiektik ~ Rodrigo Oshiro

Letra. Imagen. Sonido **L.I.S.** Ciudad Mediatizada
Año VI, # 12, Segundo semestre 2014
Buenos Aires ARG | Págs. 121 a 127

El artículo propone un acercamiento a la compleja utilización de la BigData articulando academia e industria. De este cruce surgen tantas posibilidades como escollos que dividen el campo en dos: entusiastas y optimistas ven en el análisis de enormes bases de datos el gran salto paradigmático de la investigación, mientras que escépticos descansan sobre el muestreo representativo y desconfían de la *determinación*. A tal fin se exponen los principales avatares que atravesó la utilización de la *BigData* y la minería de datos en el proyecto Flatiron Health, iniciado en el 2012 en Estados Unidos, donde siguiendo los preceptos del *netvertising*, se intenta sistematizar y agrupar la información de los pacientes del cáncer de colon, para optimizar el posterior tratamiento.

Palabras clave: BigData ~ culturómica ~ patrones ~ cáncer

The purpose of this article is to approach to the still complex uses of *BigData* by the academy and the industry. On this intersection emerge a lot of possibilities and problems, that divides the field in two: the enthusiast and optimists see in this enormous analysis of databases the big paradigmatic jump of the investigation, while the skeptics still rest in the representative sampling and refuse to think in a *determination*. With this finality, in this article are exposed the main avatars of the project Flatiron Health, its relationship with the *BigData* and the attempt to systematize the patients information for a better treatment of the colon cancer.

Keywords: BigData ~ culturomics ~ patterns ~ cancer

Culturoma

Hace unos años empezó a rondar en algunas mentes febriles el sueño del *culturoma*, algo así como un genoma de la cultura, un conjunto de matrices invisibles estructurantes, un código difícil de descifrar pero del que podían desprenderse las tendencias y dis-posiciones instituyentes. A la vista de todos, al menos de los investigadores e inversores coherentes, el asunto parecía una locura. Pero la empresa Google tomó el guante, escaneó millones de libros que fueron cargados en bases de datos y, tras ser macerados en el caldo de las trillonarias búsquedas de su web, los puso a disposición de todos los que quisieran jugar con ellos: los resultados no fueron solamente experiencias lúdicas, sino que fueron publicados varios papers en journals reconocidos.

Hoy, aún con buena parte de la academia y la industria en contra, con contradicciones, tropiezos y grandes atajos descubiertos en la espesura de la hipertrofia de datos, nos acercamos poco a poco a la culturómica: un modo cuantitativo de abordar las cuestiones sociales, que aún corriendo el riesgo de desdibujarlas, ha comenzado a impactar en la lexicografía, la gramática evolutiva, la inteligencia colectiva, la difusión de la tecnología, los estudios sobre la moda, la teoría de la censura o la epidemiología, entre otros. Grandes conjuntos de datos, que hasta hace poco no podían ser capturados por los métodos de recolección ortodoxos, son ahora gestionados y procesados de manera inmediata por computadoras de diferentes potencias, solas o en redes. ¿Es la culturómica -o la *BigData*, el término más en boga actualmente- un abordaje basado exclusivamente en lo cuantitativo y en la velocidad? En términos de *bytes*, estamos hablando de *petabytes* (1015) y *exabytes* (1018). Este volumen inmenso de datos es sólo uno de los pilares de la *BigData*. La variedad de los mismos -dispositivos móviles, audio, video, sistemas GPS, incontables sensores digitales en equipos industriales, automóviles, medidores eléctricos, veletas, anemómetros y tantos otros- es también esencialmente inherente a este nuevo abordaje. Sea como fuere la cultorómica puesta a disposición de la academia y la industria constituye todavía un conjunto problemático: desde el polo más entusiasta o integrado al polo más escéptico o apocalíptico (la vieja metáfora de Eco continúa sobrevolando con fuerza estos debates) encontramos en el medio las interacciones, los comportamientos y las opiniones *backupeadas* en escalas desconocidas y nunca antes imaginadas por los sociólogos, que se habían conformado durante generaciones con el muestreo representativo. Ellos advierten, y con buenas razones, que no van a dejarse seducir por las sirenas de grandes volúmenes de datos hasta que no los vean funcionar y, por otro lado, señalan que todo esto no resuelve la vieja cuestión: causalidad, correlación o pronóstico, cuando no determinación.

De Sillicon Valley a Hollywood

Como si esto fuera poco, además de Google otros peces gordos están detrás del asunto enturbiando las aguas: las ecuaciones del gigante IBM han calculado que la miríada de búsquedas en Google, tweets y otras actividades en la web destilan diariamente 2.5 billones de gigabytes de datos por día. La *BigData* vendría a recoger lo posible de los medios efímeros, caracterizados por lo transitorio, evanescente y periférico, o de los viejos medios escriturarios como las ondas de 500 años de guerras regulares antes de la caída de Constantinopla o las frecuencias de temáticas en la novela inglesa del siglo XIX.

De esta minería se destila un orden y una coherencia singular. Nos encontramos frente a una situación inédita gracias al procesamiento computacional: lo efímero, caracteri-

zado por su temporalidad limitada y su constitución residual, es ahora posible de ser recolectado, *reciclado* y reubicado en una matriz.

Podemos problematizar la materia enfrascándonos en las megacorporaciones de buscadores y de sistemas operativos, pero la cuestión va mucho más allá: desde el costado de la industria del entretenimiento se estrenó hace poco Transcendence[1], una película de ciencia ficción y suspenso, que entre otras cosas presenta como personaje central un dispositivo maquínistico capaz de aprender a partir de grandes cúmulos de información. En el filme, el software investiga y navega en bases de datos hechas con otras bases de datos y se vuelve capaz de crear y aplicar tratamientos médicos basados en *BigData* y nanotecnología para traumas, parálisis, ceguera y prácticamente todos los males del mundo conocidos. El film es un paso más en la dirección que ya había dado otra película dedicada a esta temática, como fue el caso de *Her*[2], en la que el argumento se centra en la historia de un sistema operativo ubicuo que aprende a relacionarse emocionalmente con los usuarios de Internet, también nutriéndose de enormes bases de datos.

Todas estas historias tienen el común denominador de la *detección de patrones* para pronosticar comportamientos, cuando no inducirlos. El tema no es nuevo: la geoeconomía, la arquitectura, o la política misma han inventado/construido, a fuerza de ensayo y error, patrones compartidos socialmente y luego los han puesto a funcionar por ejemplo en el reconocimiento que se realiza a partir de imágenes satelitales con fines recaudatorios, para el pronóstico meteorológico o el reconocimiento de rostros.

Ingeniería contra el cáncer

El asunto que nos convoca en este artículo se originó en la biología, que es la disciplina que más se benefició con su versión *bioinformática* del procesamiento de grandes volúmenes de datos. Así nacieron las llamadas *ómicas*, concepto originario del inglés y que los biólogos utilizan como sufijo para referirse al estudio de la totalidad o del conjunto de algo: ramas de las ómicas como la genómica, la metabolómica, la fluxómica, la regulómica o la signalómica empezaron a dar frutos concretos, en modelado de proteínas para desarrollo de medicamentos anticancerosos o de enfermedades degenerativas neurológicas. Evidentemente, de ahí proviene la metáfora de la comprensión cuantitativa de la cultura como una x-ómica más.

La inversión en Transcendence fue de aproximadamente 100 millones de dólares, un poco más de los 80 que Nat Turner[3] y Zach Weinberg consiguieron por la venta de Invite Media, una empresa de *netvertising* —publicidad *online*— basada en *BigData* que compró Google en 2010. ¿En qué se basaba Invite Media? Básicamente en tomar decisiones correctas escuchando el murmullo de enormes bases de datos en tiempo real, orientando grandes paquetes publicitarios hacia dónde anunciar y cómo hacerlo.

Hay, sin embargo, un elemento que enriquece la historia. Turner comenzó a interesarse en el cáncer en 2009 cuando su primo de 7 años de edad, Brennan Simkins, se enfermó, y después de una serie de pruebas, se le diagnosticó una leucemia mieloide aguda. Cuatro trasplantes de médula ósea hicieron retroceder la enfermedad, pero en el pro-

1 *Transcendence* (2014) 120 min http://www.transcendencemovie.com/. USA, CHINA.

2 *Her* (2013) Director: Spike Jonze. USA.

3 Kimberley Ong, Wharton Journal. 14 de octubre de 2013.

ceso Turner detectó algo que había visto trabajando con la publicidad online: diferentes plataformas, herramientas y bases de datos que, como instrumentos musicales de una orquesta mal dirigida no terminaban de sonar coordinadamente.

Turner creyó que podían encontrar la forma de producir coherencia en otros campos, además del *netvertising*, y convenció a Zach de usar parte de los fondos de la venta de Invite Media para crear Flatiron Health, una empresa que tuviese como objetivo curar el cáncer con ingeniería de sistemas.

La idea resultó ser interesante hasta para el mismísimo Google, que lo demostró invirtiendo más de 100 millones de dólares a través de Google Ventures, su división de capital de riesgo —en total, Flatiron Health recaudó unos 140 millones de dólares—.

La tesis de Flatiron Health es simple: si en Estados Unidos sólo el 4 % de los datos del tratamiento de pacientes de cáncer es recolectado de manera sistemática, la sistematización de la información del 96% restante podría ayudar a la medicina a contar con mejores opciones de tratamiento: el procesamiento de datos podría decir qué funcionó mejor, detectar deficiencias rápidamente, buscar puntos de derroche y acelerar el desarrollo y la aprobación de nuevos medicamentos. Con esta especie de Babel cancerológica Turner y Weinberg han estimado conseguir un impacto sobre el 5% de los casi 1,7 millones de estadounidenses diagnosticados con cáncer al año, lo que equivaldría a salvar decenas de miles de vidas.

Analógicos y digitales

La anécdota nos ilustra sólo una porción del enorme abanico de posibilidades de la *BigData*. Lo primero que hicieron Turner y Weinberg fue ponerse de acuerdo en cómo organizar la montaña de datos clínicos que estaban dispersos en los sistemas de archivo de los centros de tratamiento oncológico en Estados Unidos. Se propusieron como objetivo el recopilar los datos digitales y analógicos para clasificarlos y agruparlos, y luego ofrecerlos nuevamente a los médicos con el fin ayudarlos a tomar decisiones más sistemáticas para tratar a los pacientes.

Para Turner y Weinberg el problema de organizar los datos —de oncología clínica en este caso— era algo familiar. Uno de los principales inconvenientes que enfrentaron fue el hecho de que gran cantidad de médicos le rehuían a los registros digitales y, por otro lado, que los datos de un sólo paciente pueden provenir de decenas de fuentes no digitales: informes de internación, otros oncólogos, radiólogos, cirujanos, laboratorio y patología. Luego, incluso cuando están digitalizados, los datos, en lugar de estar perfectamente organizados en bases de datos, aparecen en diferentes formatos: muchos escritos a mano, en grabaciones de audio o en archivos PDF -formato clásico de almacenamiento de textos digitales- de baja resolución de los equipos de fax.

Viendo los primeros archivos, los datos que se habían obtenido mediante procesamiento de lenguaje natural, una metodología por medio de la cual los ordenadores *leen* los documentos analógicos y extraen datos de ellos, estaban llenos de errores. Ante esto, Flatiron Health creó un sistema de aprendizaje híbrido para capturar los datos y corregirlos. La compañía contrató a un equipo de cincuenta enfermeras para introducir datos de los primeros quinientos pacientes a mano, lo que permitió crear algoritmos para detectar errores en la recopilación de forma automática. Las discrepancias fueron agregadas de nuevo en el sistema informático para ayudar a perfeccionar el proceso de recolección automatizado.

A estos problemas se les deben sumar una amplia gama de regulaciones sobre la privacidad del acto médico que rigen la información de los pacientes, lo que hace que sea aún más difícil compartir y ensamblar miles y miles de prácticas de oncología.

Turner y Weinberg pasaron más de dos años desarrollando esa forma de organizar la información clínica en las categorías ordenadas que se habían propuesto. Se centraron en el cáncer de colon y descartaron los demás. Visitaron cientos de clínicas y entrevistaron decenas de especialistas. Basados en los ensayos clínicos publicados, extrajeron más de 350 categorías de datos, desde la demografía a las etapas del cáncer, marcadores biológicos de la enfermedad, respuestas a los tratamientos, en fin, todo lo clasificable. Actualmente el proyecto trabaja con 210 centros de cáncer que aportan colectivamente alrededor de 300.000 pacientes nuevos cada año.

La experiencia de Flatiron Health nos permite observar, con todos sus avances y retrocesos, un primer modo de proceder en la aplicación de la *BigData*, articulando miradas en superficie con otras en profundidad. El software requiere *inputs múltiples* —oncólogos y enfermeros en este caso, pero también podemos pensar en antropólogos, biólogos, etc.— para el ensamblado de atributos, clases y patrones. Turner y Weinberg apostaron a que su *expertise* residía en su capacidad para generar arquitecturas para la información.

BigData al servicio

Flatiron Health no es el primer proyecto en embarcarse en este tipo de misión. Desde la Sociedad Americana de Oncología Clínica al Instituto Nacional del Cáncer de Estados Unidos, ya se llevan destinados millones de dólares para dejar el camino preparado a iniciativas como Flatiron Health.

El Dr. Robert Weinberg, miembro fundador del Instituto Whitehead para la Investigación Biomédica del MIT se mantiene con el escepticismo propio de los que conocen del tema del que hablan. ¿La explosión de datos podría abrumar a los médicos más que orientarlos? Según su opinión ya tienen suficientes problemas con muchísimos menos datos y los beneficios podrían no ser lo suficientemente importantes como para que los médicos cambiaran sus prácticas habituales.

Idas, venidas y futuro

Por lo que vemos, la *BigData* aplicada que hemos ejemplificado en este artículo está todavía en una zona de intensa polémica, creando esas corrientes de admiración y repudio que producen los adolescentes cuando ingresan como bárbaros en campos constituidos, arrasando con todo, o repitiéndolo sin saberlo. Sin embargo, la *BigData* ya juega un papel esencial para los expertos en salud dedicados a la prevención.

Lo de Flatiron es un ejemplo de lo que está por venir. Idas y venidas, intertextualidad, resultados blandos, inversiones millonarias, pérdidas, preguntas nuevas y sospechas industriales. Es también, de algún modo, un sacudón: lo efímero se está convirtiendo en significativo; las bases de datos proporcionan un compromiso con el momento y el lugar, que a su vez otorga *un cualquier momento* y *un cualquier lugar* accesible para nuevas preguntas y respuestas.

Este burbujeo dejará cosas atrás y cosas adelante, un sistema de filtrado nuevo, autónomo, donde la subjetividad parece quedar destilada en pliegues, si es que queda, atra-

vesada por una distinción entre el flujo de datos enredándose en patrones emergentes o como un archivo tensionado entre temporalidades flexibles, entre la permanencia y la transitoriedad. Según parece vislumbrarse, será cada vez más difícil mantener a la oncología como una artesanía exquisita.

Por último queremos agregar una nota sobre la relevancia de estos conceptos, relacionada con los medios locativos, es decir con los medios capaces de inteligir respuestas de acuerdo a la ubicación territorial.

Actualmente existen decenas de aplicaciones de las principales plataformas móviles (iOS, Android, Windows Phone) diseñadas expresamente para recopilar datos de los usuarios de *smarthphones* que posteriormente pueden ser procesados con técnicas de *BigData*. Se avanzaría así desde el polo de los grandes datos por un lado, junto con la geolocalización —información de la posición de un objeto en un sistema de coordenadas determinado— de cada usuario de medios locativos desde el otro, abriendo la posibilidad de una *BigData* singularizada.

En definitiva, volviendo a la culturómica, —de la que hemos intentado retratar una de sus facetas con un ejemplo concreto— parece ser que estos modos cuantitativos han llegado para quedarse, aunque todavía estén dando los primeros pasos (y tropiezos) y los costos sean altos. En todo caso, se explora un conjunto tangible de prácticas, estrategias y formatos ayudando a profesionales y al público a diseñar, negociar e inventar un paisaje cultural cada vez más fragmentado y fugaz, pero instituido por datos relacionados.

En los patrones que surgen del procesamiento de datos, de esa relación entre grandes números y singularidad, es donde se quiebra la "perspectiva subjetivista" de las ciencias sociales, se encuentra la potencia y también la incertidumbre de este abordaje epistemológico. Al mismo tiempo la culturómica introduce el escalpelo en otro punto delicado para los cientistas sociales: la imposibilidad del anhelo moderno de una delimitación del objeto a estudiar, de su reducción a un conjunto definible y bordeado, ya que las formas que surgen de la sobreabundancia de datos lo hacen siempre y por definición desde el exceso, desde la hipertrofia, la sobreabundancia.

Lo que queda por verse es si se podrá hacer algo por fuera de la interdisciplina, porque los viejos moldes parecen estar caducos frente a los nuevos materiales de trabajo.

Referencias bibliográficas

Andrews, M. (2013) "Crime, punishment and the BigData revolution". http://techpageone.dell.com/technology/data-center/crime-punishment-and-the-big-data-revolution/ (visitado 20/09/2014)

Grainge, p. (2011) "Ephemeral Media", Londres, British Film Institute, 2011.

"Big Data Now" (2012) by O'Reilly Media, Inc. USA.

Hey, T.; Tansley, Stewart & Tolle (2009) "The Fourth Paradigm: Data-Intensive Scientific Discovery", en Kristin (eds.) *Redmond: Microsoft Research*.

Michel, J., et. al. (2010) "Quantitative Analysis of Culture Using Millions of Digitized Books" en *Science*, Nueva York, Vol. 331 N° 6014.

Moretti, F. (2005) "Gráficos", en *La literatura vista desde lejos*. Barcelona, Marbot, 2007.

Preiser-Kapeller, J (2010) "Calculating Byzantium? Social Network Analysis and Complexity Sciences as tools for the exploration of medieval social dynamics" en Austrian Academy of Science's Working Papers. http://www.oeaw.ac.at/byzanz/repository/Preiser_WorkingPapers_Calculating_I.pdf (visitado 10/09/2014)

Cukier K. y Mayer-Schonberger V. (2013) "Big Data: A Revolution That Will Transform How We Live, Work and Think".