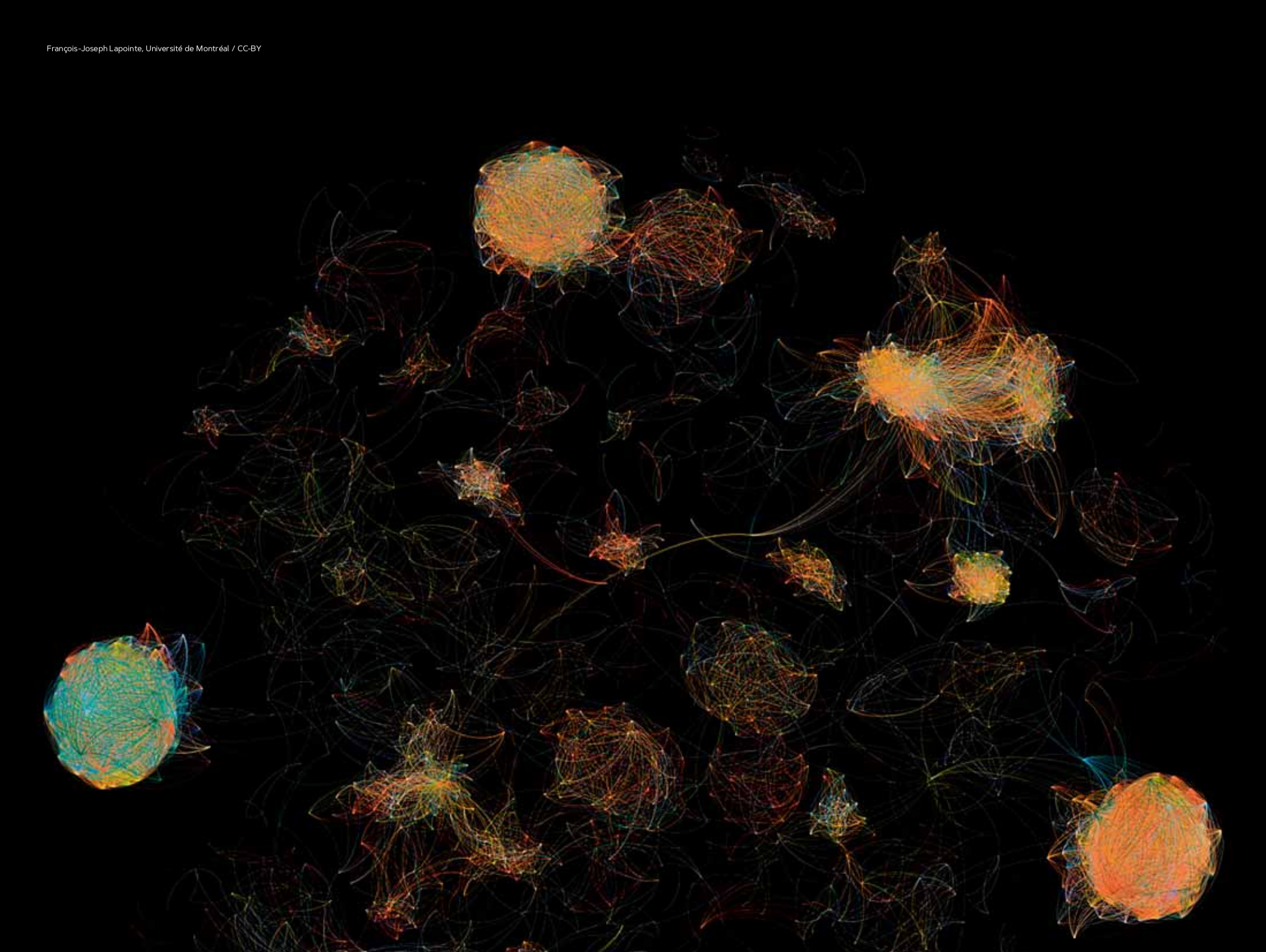


DOCUMENTO
Mètode Science Studies Journal (2020). Universitat de València.
DOI: 10.7203/metode.11.15258
Enviado: 23/05/2019. Aceptado: 27/09/2019.

SOBRE EL "BIG DATA"

¿Cómo podríamos dar sentido a los macrodatos?

Fulvio Mazzocchi

Actualmente, existe un intenso debate acerca de la cuestión del *big data*, no solo por razones técnicas. Esto se debe también a que se suele presentar el *big data* como un elemento que conlleva un cambio de paradigma epistemológico en la investigación científica que podría reemplazar al método tradicional, basado en plantear hipótesis. En este artículo realizo un escrutinio crítico de dos afirmaciones clave asociadas habitualmente con este enfoque, concretamente que los datos hablan por sí solos –un argumento que menosprecia el papel de teorías y modelos– y la prioridad de la correlación sobre la causalidad. Mi intención es, por una parte, reconocer el valor del análisis del *big data* como una herramienta heurística innovadora y, por otro, explicar detalladamente qué podemos esperar de los macrodatos y qué no.

Palabras clave: *big data*, ciencia basada en datos, epistemología, fin de la teoría, causalidad, opacidad de los algoritmos.

Las imágenes que ilustran este artículo forman parte de la serie «Selfis del microbioma» realizada por el artista y biólogo François-Joseph Lapointe en su *performance 1.000 apretones de manos*. El artista estrechó la mano de más de 1.000 personas, cambiando gradualmente la comunidad microbiana invisible de la palma de su mano. Cada 50 apretones, se muestrearon y analizaron los microbios de su mano para revelar cómo nuestro contacto con otros da forma a nuestra microbiota común. Este proyecto se ha realizado en varias ciudades alrededor del mundo (incluyendo Copenhague, Montreal, San Francisco, Perth, Berlín y Baltimore) como una manera de mapear nuestro microbioma colectivo con datos científicos. La producción de «Selfis del microbioma» implicó muchas fases diferentes. Tras la recogida de muestras, se extrajo, se amplificó y se secuenció el ADN para crear los datos bioinformáticos que se muestran en esta serie. Los nodos de las redes representan las secuencias de ADN bacteriano, y dos nodos aparecen conectados por una línea cuando sus secuencias de ADN bacteriano tienen más de un 95% de similitud. Los diferentes colores corresponden a diferentes muestras recogidas cada cincuenta apretones de manos, desde 0 hasta 1001.

En estas páginas, *Selfi del microbioma*, de François-Joseph Lapointe, después de estrechar 550 manos durante su *performance 1.000 apretones de manos*.

EL «FIN DE LA TEORÍA» Y OTRAS AFIRMACIONES SOBRE LA INNOVACIÓN DEL "BIG DATA"

Según algunos expertos (por ejemplo, Anderson, 2008), el método científico basado en hipótesis no tiene futuro. Hay quien ha proclamado el «fin de la teoría», indicando que estamos en el punto de partida de una nueva etapa en la investigación científica, una etapa basada en *petabytes* de información y en las supercomputadoras. El futuro pertenece a una nueva forma de empirismo basada en la tecnología y sus potentes herramientas, incluyendo algoritmos y técnicas estadísticas muy perfeccionados. Estas herramientas son capaces de rebuscar en enormes cantidades de datos y recopilar información que se pueda transformar en conocimiento.

Los partidarios del *big data* defienden que este enfoque es revolucionario y apuntan principalmente a dos innovaciones clave. La primera es que es posible extraer patrones significativos a partir del análisis de datos. Estos patrones se originan directamente en los datos. Como consecuencia de ello, se postula un giro ateórico según el cual no sería necesario plantear hipótesis, teorías ni modelos previos. En segundo lugar, en el reino del *big data*, «la correlación es suficiente» (Anderson, 2008), y no es necesario investigar los vínculos causales entre variables asociadas. Por lo tanto, la correlación sustituye a la causalidad.

«Hay quien ha proclamado el "fin de la teoría", una nueva etapa en la investigación científica basada en *petabytes* de información y en las supercomputadoras»

Lo cierto es que la llegada del *big data* conlleva verdaderas novedades de tipo tecnológico. Este no se caracteriza solo por el volumen, la velocidad y la variedad de los datos, sino también por su alcance exhaustivo y detallada resolución, y por ser muy relacionales, además de flexibles y escalables en producción (Kitchin, 2014). Las técnicas de aprendizaje automático pueden extraer datos y detectar regularidades bajo el supuesto de que «mucho de lo que se genera no responde a ninguna pregunta en particular o es un subproducto de otra actividad» (Kitchin, 2014, p. 2). Utilizando un enfoque colectivo, se pueden aplicar varios algoritmos a los conjuntos de datos con el objetivo de optimizar el rendimiento predictivo. Lo que se afirma en este caso es que está surgiendo «un enfoque epistemológico totalmente novedoso para dar sentido al mundo». De hecho, «en lugar de probar una teoría analizando datos relevantes, los nuevos análisis de datos tratan de obtener información "que nace en los datos"» (Kitchin, 2014, p. 2).

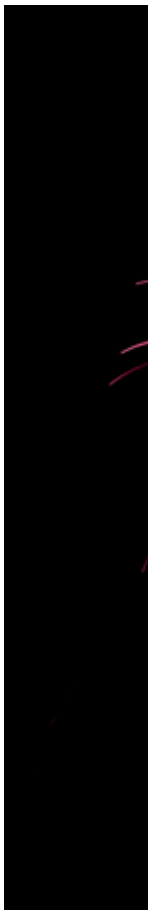
No cabe duda de que el enfoque de *big data* está contribuyendo a cambiar el panorama epistémico actual. Además, las técnicas de minería de datos también están creando nuevas oportunidades para la investigación científica. Por ejemplo, existe la posibilidad de comparar cientos de genomas del cáncer y, gracias a la secuenciación de ADN, establecer la frecuencia de muchas mutaciones potencialmente significativas para diferentes tipos de cáncer, junto con sus consecuencias funcionales: esto puede incluso contribuir al desarrollo de nuevas terapias (Golub, 2010). En términos más generales, mediante estas técnicas es posible descubrir patrones potencialmente significativos en grandes volúmenes de datos, algunos de los cuales habrían pasado desapercibidos anteriormente debido a su complejidad.

Sin embargo, suponer que el *big data* representa un verdadero cambio de paradigma epistemológico (al menos en el sentido que indicábamos anteriormente) es una cuestión completamente diferente. De hecho, no hay razón para pensar que los macrodatos permitan crear un nuevo modo de producción de conocimiento en el que los supuestos teóricos y las hipótesis no cumplan ningún papel y se pueda ignorar la idea de causalidad.

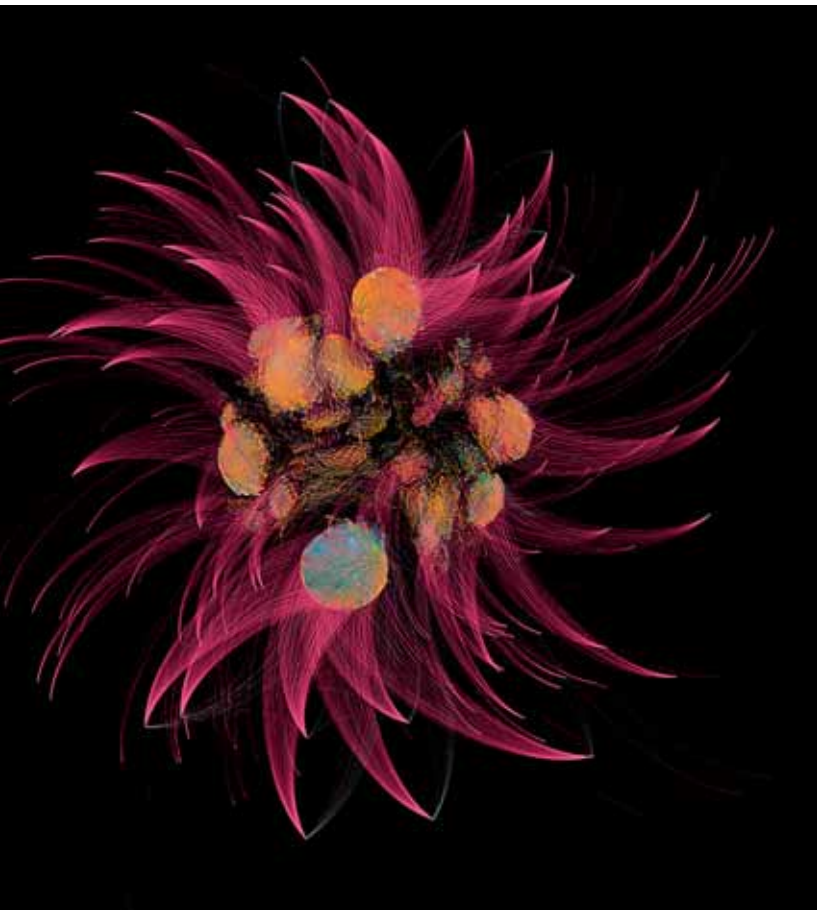
Ambas afirmaciones sobre el *big data* han despertado fuertes reacciones. Por ejemplo, atendiendo tanto a la generación de datos como a su análisis, observamos que difícilmente podemos encontrar una forma de crear conocimiento sin necesidad de formular hipótesis (es decir, una forma que dependa únicamente de la manipulación estadística y la inducción).

En primer lugar, los datos no surgen de la nada. La filosofía de la ciencia del siglo XXI ha discutido extensamente el papel que representan las nociones preconcebidas, comenzando por Karl Popper (1959, por ejemplo). En su opinión, las hipótesis cumplen un papel esencial en la investigación científica, ya que nos indican qué buscar y qué datos recopilar. Otro argumento conocido es la «saturación teórica» de los datos y la observación, es decir, el hecho de que estos estén «contaminados» por presunciones teóricas.

En realidad, la naturaleza no se investiga al azar. Lo que se llega a inspeccionar y medir está influenciado por el conocimiento de fondo, los intereses y las estrategias del investigador. Hasta el diseño de experimentos depende de limitaciones teóricas, metodológicas y técnicas específicas. Por lo tanto, los datos siempre son el resultado de la interacción entre el investigador (que pertenece a una determinada escuela de pensamiento) y el mundo, siempre que se cumplan las condiciones materiales adecuadas (Leonelli, 2015; Mazzocchi, 2015).



François-Joseph Lapointe, Université de Montréal / CC-BY



Selfi del microbioma, de François-Joseph Lapointe, después de estrechar 650 manos durante su *performance 1.000 apretones de manos*.

En segundo lugar, los datos o las cifras no hablan por sí solos. Se pueden encontrar regularidades significativas mediante computadoras, pero lo importante es encontrarles una explicación. Esto presupone la existencia de un «marco de análisis», una lente teórica de la que depende cómo se interpretan los datos: es aquí donde el papel del conocimiento específico de dominio resulta crucial. Boyd y Crawford (2012, p. 667) indicaron que «todos los investigadores son intérpretes de datos […]. Un modelo puede ser sólido en términos matemáticos, un experimento puede parecer válido, pero el proceso de interpretación comienza en cuanto el investigador intenta entender lo que significa».

«Los datos o las cifras no hablan por sí solos. Lo importante es encontrarles una explicación»

Varios científicos de datos, así como muchos bioinformáticos en la disciplina de la biología, creen que entender las estadísticas puede ser suficiente para dar sentido a los datos. Se presupone que los patrones son significativos por sí mismos, es decir, que su significado trasciende el contexto o dominio, y no es necesario buscar fuera de los datos. En su opinión, el conocimiento teórico «depende de generalizaciones reduccionistas que se abstraen de la realidad de forma problemática» (Chandler, 2015, p. 847). Por el contrario, el enfoque computacional nos permitiría acceder a conjuntos de datos interconectados y alcanzar una comprensión más holística –más allá de los obstáculos disciplinarios– de fenómenos complejos. Sin embargo, es un poco paradójico esperar que los datos, que se han producido en un contexto concreto (por ejemplo, la biología), se puedan interpretar fácilmente exentos de cualquier contexto. Permítanme subrayar de nuevo este concepto: los conocimientos específicos de dominio son importantes.

Además, incluso los algoritmos de aprendizaje automático están impregnados de suposiciones particulares, como por ejemplo, qué consideramos un patrón regular: cada algoritmo tiene su propia manera de desarrollar estrategias para encontrar relaciones entre los conjuntos de datos, y es probable que diferentes algoritmos encuentren diferentes tipos de patrones (Hales, 2013). Esto lo reconocen incluso algunos especialistas en *big data*.

La segunda afirmación –la idea de que «la correlación es suficiente»– exagera el valor de las predicciones realizadas a partir de correlaciones. Tal vez existan circunstancias particulares, como la publicidad, en las que esta idea podría tener sentido. Sin embargo, probablemente no es cierto en el caso de la investigación científica.

Las correlaciones pueden sugerir conexiones potencialmente interesantes. Pueden incluso ser útiles para generar o evaluar nuevas hipótesis, aunque esta tarea siempre estará guiada por algunos supuestos teóricos subyacentes y por el conocimiento disponible (Kitchin, 2014). Pero las correlaciones no nos informan sobre la causa que subyace a estas relaciones.

En la ciencia, establecer conexiones causales es esencial, incluso para saber cómo intervenir de manera efectiva en situaciones de máxima prioridad, como por ejemplo para curar una enfermedad. Por lo tanto, la investigación científica no se detiene en las correlaciones. Existe la necesidad de realizar análisis y pruebas adicionales: las correlaciones se han de «validar» de alguna forma. El conocimiento fiable solo se puede alcanzar al final de este proceso. Esto depende también del hecho de que, especialmente en las bases de datos de gran tamaño, la mayoría de correlaciones son falsos positivos (Calude y Longo, 2017). Debido al gran volumen de datos, el problema es cómo hacer frente

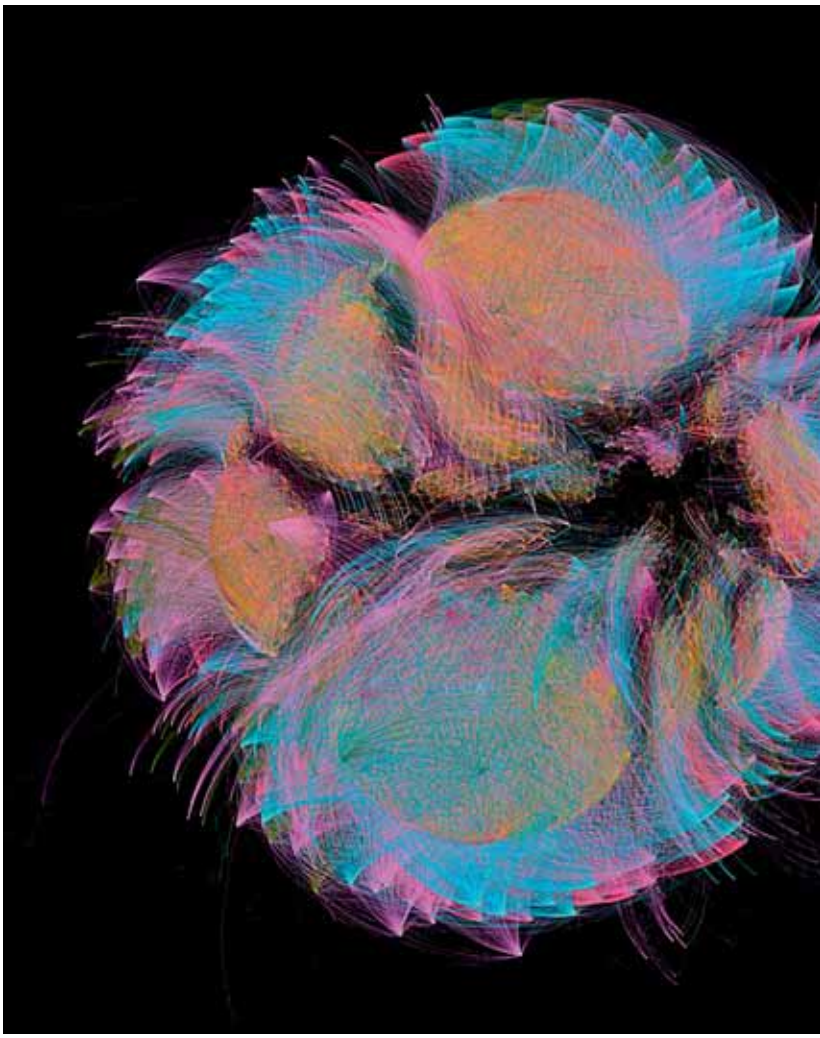
a la presencia de muchas relaciones de correlación, y distinguir las conexiones significativas de las confusas (las falsas).

EL CASO DE EXPOsOMICS

Analicemos ahora la historia de un caso de macrodatos biomédicos: el proyecto EXPOsOMICS. Este proyecto investiga las asociaciones entre exposición y enfermedad en relación con la novedosa idea de exposoma, es decir, la cantidad total de exposición que afecta a los individuos durante toda su vida. Este concepto incluye la exposición interna y externa: por ejemplo, en su estudio sobre el cáncer de mama y de colon, Chadeau-Hyam et al. (2011) analizaron tanto la dieta como el estilo de vida de los pacientes (exposición externa), así como su respuesta metabólica (exposición interna).

Los biomarcadores desempeñan un papel clave en este tipo de investigación. Estos son elementos medibles del entorno y el organismo que muestran procesos biológicos. De hecho, se buscan las asociaciones entre biomarcadores de exposición y biomarcadores de enfermedad. Resulta significativo que la investigación en biomarcadores se suele llevar a cabo utilizando *big data*, habitualmente obtenidos mediante tecnologías de alto rendimiento: la ómica en el caso de la exposición interna, y los sensores, satélites y otros recursos en el de la exposición externa. Como señaló Canali (2016, p. 4):

> EXPOsOMICS es un proyecto de *big data* en el que los científicos buscan biomarcadores asociados que puedan rastrear la exposición y la enfermedad. Quien defienda un enfoque basado en los datos podría decir que este proyecto es el ejemplo perfecto de cómo la investigación en *big data* consiste en recopilar grandes cantidades de datos, analizarlos, buscar correlaciones entre biomarcadores de exposición y biomarcadores de enfermedad y realizar predicciones. Esto demostraría que las correlaciones son suficientes por sí solas y que el conocimiento causal no es necesario.



François-Joseph Lapointe, Université de Montréal / CC-BY

Selfi del microbioma, de François-Joseph Lapointe, después de estrechar 850 manos durante su *performance 1.000 apretones de manos*.

«La naturaleza no se investiga al azar. Lo que se llega a inspeccionar y medir está influenciado por el del investigador»

Pero esto no es así. De hecho, en el estudio antes mencionado sobre el cáncer de mama y de colon, la búsqueda de asociaciones en los datos para identificar listas de supuestos biomarcadores que relacionen exposición y enfermedad es solo el punto de partida. Una correlación entre biomarcadores también se puede considerar significativa en términos estadísticos, pero lo que se intenta encontrar es una relación causal entre la exposición y la enfermedad (Canali, 2016).

Para ello, existe la necesidad de buscar biomarcadores «intermedios», que se piensa que pueden estar involucrados como causantes de enfermedades. Se encuentran en la intersección entre los biomarcadores de exposición y los de enfermedad. En el caso del cáncer de colon, la ingesta de fibra se identifica como un posible biomarcador intermedio. Todo este proceso está guiado por una combinación de datos, pruebas estadísticas, principios teóricos, experimentos previos y conocimiento causal disponible sobre los mecanismos de la enfermedad, por ejemplo utilizando la Base de Datos

del Metaboloma Humano, que contiene información sobre los mecanismos metabolómicos (Chadeau-Hyam et al., 2011).

En realidad, esta descripción del proyecto EXPOsOMICS, y de otros muchos como ENCODE (véase Mazzocchi, 2015), muestra la invalidez de las afirmaciones del *big data* sobre el final de la teoría y la prioridad de la correlación sobre la causalidad. Aunque a veces la investigación científica pueda comenzar a partir de unos datos y, por tanto, sin la participación de sólidos modelos o hipótesis a priori, el conocimiento teórico y experimental sigue siendo necesario inmediatamente después. Además, las consideraciones metodológicas, así como la elección de un tipo específico de modelo estadístico, desempeñan un papel esencial para dar forma a la investigación y asegurarse de que el análisis de datos es realmente eficaz.

MÁS ALLÁ DEL MITO DE LOS DATOS BRUTOS Y LA OBJETIVIDAD

Podemos considerar la afirmación de que «los datos hablan por sí mismos» desde otro punto de vista. Si tenemos en cuenta la etimología, el término *data* es la forma plural de *datum* en Latín, que significa "algo dado", referido a "aquello que se da antes de un argumento" y no es necesario cuestionar. Se conceptualizan los datos como elementos de naturaleza «preanalítica» e imparcial, se les tiene en cuenta como una reflexión directa o como una representación «desnuda» de un aspecto en concreto de la naturaleza, como si fueran fotografías (Gitelman, 2013). Esta concepción queda encapsulada en el término «datos brutos». El *big data* complica la situación porque la objetividad de los datos (como elementos concretos) se combina con la objetividad o neutralidad de los patrones que nacen directamente de ellos.

No obstante, deberíamos entender la naturaleza epistemológica de los datos de una manera más sofisticada. Como ya se ha señalado, los datos no están determinados y nunca están desnudos; en cierto modo, se «fabrican». Como recuerda Leonelli (2015, p. 820), «lo que se entiende por datos siempre está relacionado con una determinada investigación en la que se buscan pruebas para responder, o incluso formular, una pregunta». Por lo tanto, se han de ver los datos como artefactos socioculturales. Además, para ser utilizables y funcionar como prueba, suele ser necesario manipularlos y organizarlos mediante una estructura de datos, y pese a todo, incluso este proceso está impulsado por consideraciones teóricas y, por lo tanto, está lejos de ser neutral (véase Gitelman, 2013).

El proceso de generación y gestión de datos implica, en efecto, realizar varias elecciones y juicios –cada uno de ellos en cierto modo sesgado– sobre, por ejemplo, qué es significativo o fiable y qué no lo es. Tales consideraciones se pueden comparar, por ejemplo, con la idea de «oclusión ontológica» (Knobel, 2010), un mecanismo según el cual la representación de un objeto bloquea cualquier otra representación posible. Como consecuencia de ello, los elementos bloqueados no se tienen en cuenta y no «dan forma a la narrativa» de manera alguna. A la luz de este enfoque, el proceso de admisión de datos, por ejemplo para un archivo, es un proceso de bloqueo de otras posibilidades. Como no se puede superar la finitud del archivo, se dejarán de considerar o representar diversos aspectos de la realidad (Bowker, 2014).

La percepción y la cognición humanas, que funcionan proyectando límites en la realidad, también son al mismo tiempo medios para el descubrimiento y para el bloqueo. Se puede ordenar el mundo de diferentes maneras basándose en diferentes formas de proyectar los límites. Sin embargo, el mecanismo común que subyace en todos los casos es que para crear un orden concreto o «visualizar» algo es necesario excluir el resto de opciones. En otras palabras, nuestra percepción y cognición son intrínsecamente «perspectivistas».

«Varios científicos de datos creen que entender las estadísticas puede ser suficiente para dar sentido a los datos»

Los filósofos de la ciencia contemporáneos como Ronald Giere (2006) también han destacado el carácter perspectivista de la ciencia, es decir, el hecho de que incluso la observación y la teorización científicas solo pueden describir el mundo natural a la luz de una determinada perspectiva.

En este sentido, el enfoque de *big data*, que postula un modelo de objetividad «no perspectivista», supone un paso atrás. Las declaraciones de objetividad que defienden que el análisis algorítmico de los datos garantizaría la verdad y la neutralidad reflejan *de facto* la inmadurez filosófica de la disciplina. Como indicó Bollier (2010, p. 13):

> Al ser una gran masa de información sin procesar, el *big data* no se explica por sí mismo. Y sin embargo, las metodologías específicas para interpretar los datos están abiertas a todo tipo de debate filosófico. ¿Pueden los datos representar una «verdad objetiva», o cualquier interpretación está necesariamente sesgada por un filtro subjetivo o por la forma en que se «limpian» los datos?

Por lo tanto, deberíamos considerar que incluso la naturaleza misma de los datos es perspectivista. Ni los datos ni las cifras hablarán nunca por si solos, solo dan cuenta de las suposiciones que llevan incorporadas. Además, presuponer la neutralidad de los datos ya es, de por sí, una posición no neutral.

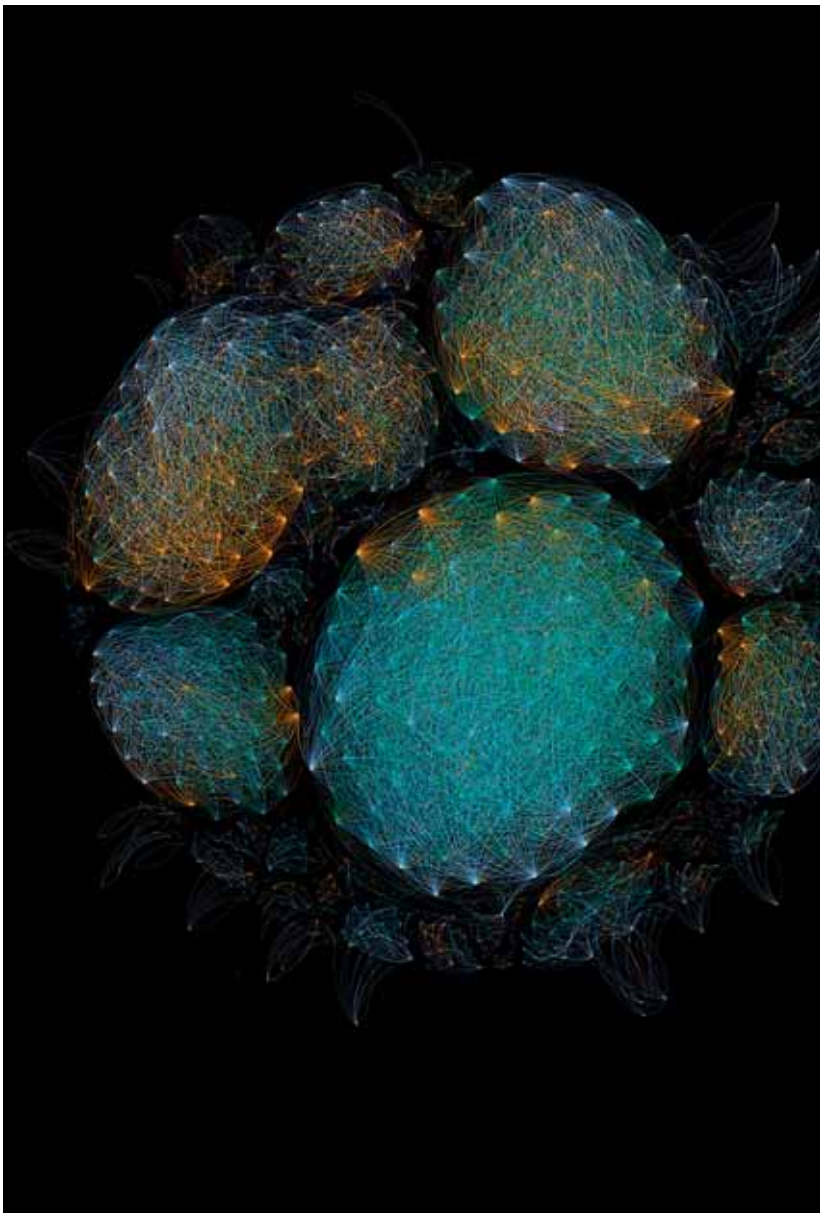
EL VALOR DEL "BIG DATA" COMO HEURÍSTICA Y LA OPACIDAD DE LOS ALGORITMOS

Podemos hacer un buen uso de las novedades que nos ofrece el análisis de macrodatos. Sin embargo, no podemos esperar reemplazar principios y procedimientos que se han empleado y refinado a lo largo de muchos siglos de investigación científica. La ciencia de hoy en día sigue estando fundamentada en la teoría y la experimentación, y es muy probable que lo siga estando en el futuro. El valor del *big data* es más bien que representa un instrumento heurístico muy potente e innovador.

El *big data* y el enfoque computacional ayudan a completar la caja de herramientas de los investigadores. La palabra clave aquí es el *pluralismo*, porque al aumentar el número de herramientas heurísticas disponibles, es posible desarrollar múltiples estrategias de investigación que se complementen entre sí. Por ejemplo, existe la posibilidad de comparar y establecer sinergias entre un enfoque de hipótesis y uno basado en datos. Tal vez en el futuro lleguemos incluso a explorar nuevas maneras de desarrollar teorías. En cualquier caso, un gran número de proyectos de *big data* como EXPOsOMICS muestran que los datos y los elementos teóricos se «influyen mutuamente» y que ambos participan repetidamente en el ciclo de la investigación científica (Canali, 2016, p. 8).

En conclusión, resulta necesario reafirmar que es necesario no aceptar de forma acrítica la cultura del algoritmo que subyace a los macrodatos. De lo contrario, incluso herramientas muy útiles pueden contribuir a crear una realidad no deseada. De hecho, los algoritmos más refinados no son solo herramientas para extraer información. Cada vez afectan más al propio entramado de vidas públicas e individuales y contribuyen en gran medida a darles forma.

> Hoy vivimos en un mundo en el que los algoritmos (y los datos con los que los alimentamos) se ocupan de una gran variedad de decisiones relacionadas con nuestras vidas: no se trata solo de motores de búsqueda y sistemas personalizados de noticias en línea, sino que afectan también a las evaluaciones educativas, el funcionamiento de los mercados y las campañas políticas, el diseño de



François-Joseph Lapointe, Université de Montréal / CC-BY

Selfi del microbioma, de François-Joseph Lapointe, después de estrechar 350 manos durante su *performance 1.000 apretones de manos*.

«La ciencia sigue estando fundamentada en la teoría y la experimentación. El valor del *big data* es más bien que representa un instrumento heurístico muy potente e innovador»

> los espacios urbanos públicos e incluso la manera en que se gestionan servicios como la seguridad pública y las prestaciones sociales. Pero se puede argumentar que los algoritmos cometen errores y funcionan en base a determinados sesgos. La opacidad de los algoritmos técnicos complejos que operan a gran escala hace difícil examinarlos, lo que conlleva una falta de claridad para con el público, en relación con cómo ejercen su poder e influencia (Diakopoulos, 2015, p. 398).

Los algoritmos, especialmente los de aprendizaje, son muy performativos e influyentes. Sin embargo, es difícil comprender su funcionamiento e implicaciones. Ni siquiera los especialistas en trabajo de campo pueden explicar completamente qué pasa realmente cuando la máquina procesa grandes cantidades de datos para obtener información nueva, o la razón por la cual elige una manera de proceder en lugar de otra en determinadas situaciones (véase Burrell, 2016). Por esta razón, se los describe como «cajas negras».

Esta opacidad para la comprensión humana, que incluso refuerza el «poder» de los algoritmos, se debe a cuestiones técnicas y a la complejidad de su funcionamiento. Una forma de expresarlo es en términos de «opacidad epistémica», es decir, no es posible entender todos los factores con relevancia epistémica implicados en sus operaciones (Humphreys, 2009).

«Deberíamos considerar que incluso la naturaleza misma de los datos es perspectivista. Ni los datos ni las cifras hablarán nunca por si solos»

En cualquier caso, la creciente opacidad de los algoritmos es algo sobre lo que debemos meditar con prudencia. Hoy en día, se celebra la performatividad de las herramientas de *big data* incluso con triunfalismo. La potencia epistémica y la supuesta neutralidad de los algoritmos, que pueden realizar funciones inalcanzables para la mente humana, se oponen a la falibilidad de la interpretación y la toma de decisiones humanas. No obstante, no deberíamos usar la performatividad como razón para ceder la autoridad y el control a las máquinas.

En lugar de limitarnos a elogiar el enfoque de *big data* y sus algoritmos, deberíamos preguntarnos una serie de cuestiones. Por ejemplo, ¿qué tipo de situación es aquella en la que utilizamos herramientas para realizar determinadas tareas complejas pero no somos capaces de explicar cómo hacen posible realizar dichas tareas? Nadie dudaría de que dispositivos tecnológicos como estos influyen en gran medida en nuestra representación del mundo. Por lo tanto, otra pregunta sería: ¿qué tipo de situación es aquella en la que hay herramientas capaces de dar forma a cómo experimentamos la realidad, pero a cuya lógica subyacente y modelos de representación somos incapaces de acceder completamente?

REFERENCIAS

Anderson, C. (2008, 23 de junio). The end of theory: The data deluge makes the scientific method obsolete. *Wired.* Consultado en **https://www.wired.com/2008/06/pb-theory/**

Bollier, D. (2010). *The promise and peril of big data*. Washington, DC: The Aspen Institute.

Bowker, G. (2014). The theory/data thing. Commentary. *International Journal of Communication, 8*(2043), 1795–1799.

Boyd, D., & Crawford, K. (2012). Critical questions for big data. *Information, Communication and Society, 15*(5), 662–679. doi: **10.1080/1369118X.2012.678878**

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1), 1–12. doi: **10.1177/2053951715622512**

Calude, C. S., & Longo, G. (2017). The deluge of spurious correlations in big data. *Foundations of Science*, *22*(3), 595–612. doi: **10.1007/s10699-016-9489-4**

Canali, S. (2016). Big data, epistemology and causality: Knowledge in and knowledge out in EXPOsOMICS. *Big Data & Society*, *3*(2), 1–11. doi: **10.1177/2053951716669530**

Chadeau-Hyam, M., Athersuch, T. J., Keun, H. C., De Iorio, M., Ebbels, T. M., Jenab, M., … Vineis, P. (2011). Meeting-in-the-middle using metabolic profiling – A strategy for the identification of intermediate biomarkers in cohort studies. *Biomarkers, 16*(1), 83–88. doi: **10.3109/1354750X.2010.533285**

Chandler, D. (2015). A world without causation: Big data and the coming age of posthumanism. *Millennium: Journal of International Studies*, *43*(3), 833–851. doi: **10.1177/0305829815576817**

Diakopoulos, N. (2015). Algorithmic accountability: Journalistic investigation of computational power structures. *Digital Journalism, 3*(3), 398–415. doi: **10.1080/21670811.2014.976411**

Giere, R. (2006). *Scientific perspectivism.* Chicago, IL: University of Chicago Press.

Gitelman, L. (Ed.). (2013). *'Raw data' is an oxymoron.* Cambridge, MA: The MIT Press.

Golub, T. (2010). Counterpoint: Data first. *Nature, 464*(7289), 679. doi: **10.1038/464679a**

Hales, D. (2013, 1 de febrero). Lies, damned lies and big data. Consultado en **https://aidontheedge.wordpress.com/2013/02/01/lies-damned-lies-and-big-data/**

Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese, 169*(3), 615–626. doi: **10.1007/s11229-008-9435-2**

Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 1–12. doi: **10.1177/2053951714528481**

Knobel, C. (2010). *Ontic occlusion and exposure in sociotechnical systems* (Tesis doctoral), University of Michigan, USA.

Leonelli, S. (2015). What counts as scientific data? A relational framework. *Philosophy of Science*, *82*(5), 810–821. doi: **10.1086/684083**

Mazzocchi, F. (2015). Could big data be the end of theory in science? A few remarks on the epistemology of data-driven science. *EMBO Reports*, *16*(10), 1250–1255. doi: **10.15252/embr.201541001**

Popper, K. R. (1959). *The logic of scientific discovery.* Londres: Hutchinson.

FULVIO MAZZOCCHI. Biólogo y filósofo. Investigador del Instituto de las Ciencias del Patrimonio del CNR (Roma, Italia). Su actividad de investigación se centra en la epistemología (pluralismo epistémico, perspectivismo), los problemas filosóficos de la investigación científica (como el debate reduccionismo-holismo en la biología, la validación de modelos climáticos, o los problemas epistemológicos del *big data*) y la organización del conocimiento.

✉ **fulvio.mazzocchi@cnr.it**