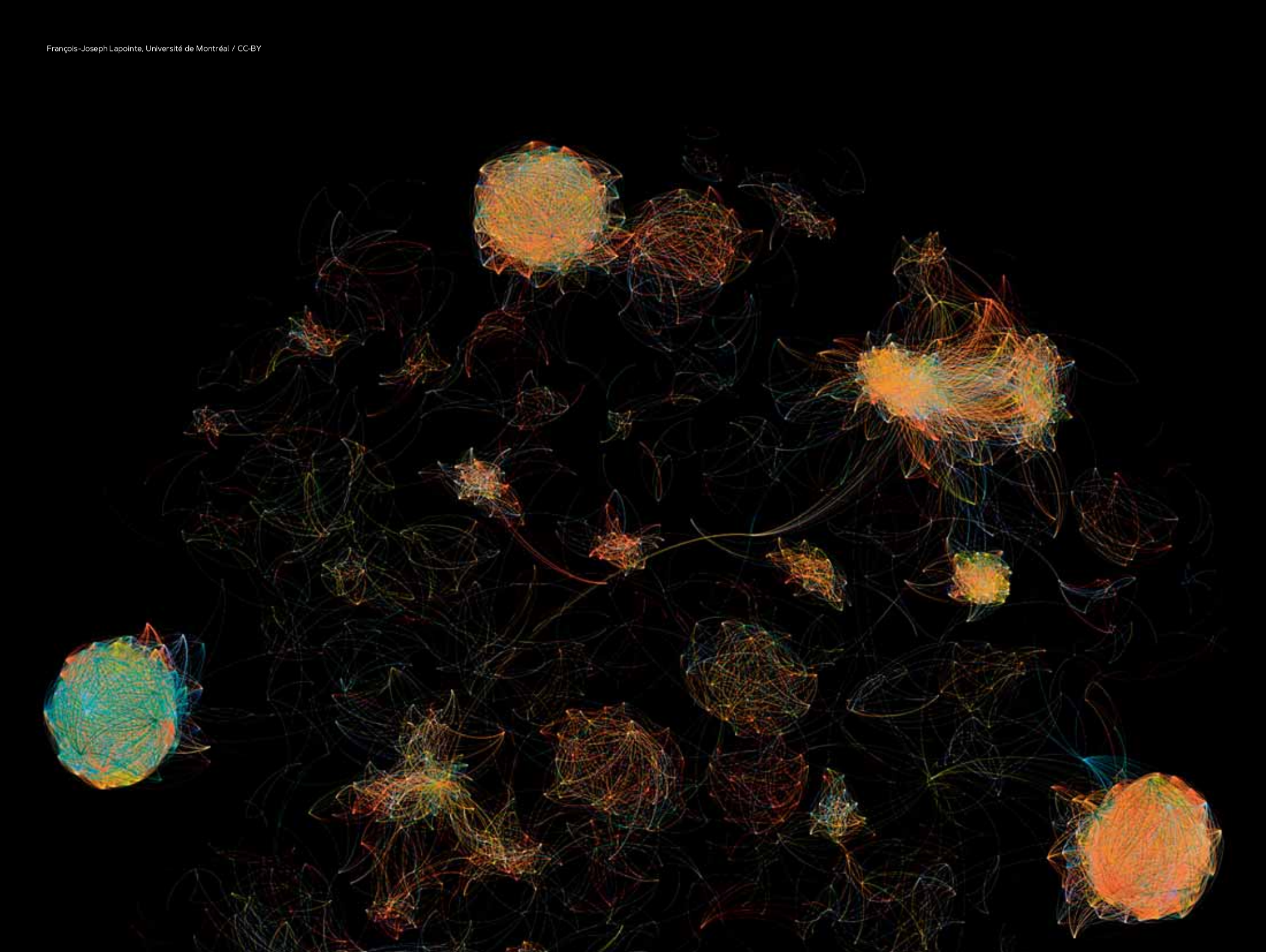


DOCUMENT
Mètode Science Studies Journal (2020). Universitat de València.
DOI: 10.7203/metode.11.15258
Enviat: 23/05/2019. Acceptat: 27/09/2019.

SOBRE EL "BIG DATA"

Com podríem donar sentit a les dades massives?

Fulvio Mazzocchi

Actualment, existeix un intens debat sobre la qüestió del *big data* o dades massives, no solament per raons tècniques. Aquest és degut també al fet que se solen presentar les dades massives com un element que comporta un canvi de paradigma epistemològic en la investigació científica que podria reemplaçar el mètode tradicional, basat a plantejar hipòtesis. En aquest article faig un escrutini crític de dues afirmacions clau associades habitualment amb aquest enfocament; concretament, que les dades parlen per si mateixes –un argument que menysprea el paper de teories i models– i la prioritat de la correlació sobre la causalitat. La meua intenció és, d'una banda, reconèixer el valor de l'anàlisi de les dades massives com una eina heurística innovadora i, d'una altra, explicar detalladament què podem esperar de les dades massives i què no.

Paraules clau: dades massives, ciència basada en dades, epistemologia, fi de la teoria, causalitat, opacitat dels algorismes.

Les imatges que il·lustren aquest article formen part de la sèrie «Selfies del microbioma» realitzada per l'artista i biòleg François-Joseph Lapointe en la seua *performance 1.000 encaixades*. L'artista va encaixar la mà de més de 1.000 persones, canviant gradualment la comunitat microbiana invisible del palmell de la seua mà. Cada 50 encaixades, es van mostrejar i analitzar els microbis de la seua mà per a revelar com el nostre contacte amb uns altres dona forma a nostra microbiota comuna. Aquest projecte en curs s'ha realitzat en diferents ciutats al voltant del món (incloent-hi Copenhaguen, Mont-real, San Francisco, Perth, Berlín i Baltimore) com una manera de mapar el nostre microbioma col·lectiu amb dades científiques. La producció de «Selfies del microbioma» va implicar moltes fases diferents. Després de la recollida de mostres, es va extraure, es va amplificar i es va seqüenciar l'ADN per a crear les dades bioinformàtiques que es mostren en aquesta sèrie. Els nodes de les xarxes representen les seqüències d'ADN bacterià, i dos nodes apareixen connectats per una línia quan les seues seqüències d'ADN bacterià tenen més d'un 95% de similitud. Els diferents colors corresponen a diferents mostres recollides cada cinquanta encaixades, des de 0 fins a 1001.

En aquestes pàgines, *Selfie del microbioma*, de François-Joseph Lapointe, després d'encaixar 550 mans durant la seua *performance 1.000 encaixades*.

LA «FI DE LA TEORIA» I ALTRES AFIRMACIONS SOBRE LA INNOVACIÓ DE LES DADES MASSIVES

Segons alguns experts (per exemple, Anderson, 2008), el mètode científic basat en hipòtesis no té futur. Hi ha qui ha proclamat la «fi de la teoria», indicant que estem en el punt de partida d'una nova etapa en la investigació científica, una etapa basada en petabytes d'informació i en les supercomputadores. El futur pertany a una nova forma d'empirisme basada en la tecnologia i les seues potents eines, incloent-hi algorismes i tècniques estadístiques molt perfeccionats. Aquestes eines són capaces de recercar en enormes quantitats de dades i recopilar informació que es puga transformar en coneixement.

Els partidaris de les dades massives defensen que aquest enfocament és revolucionari i apunten principalment dues innovacions clau. La primera és que és possible extraure patrons significatius a partir de l'anàlisi de dades. Aquests patrons s'originen directament en les dades. A conseqüència d'això, es postula un caire ateòric segons el qual no seria necessari plantejar hipòtesis, teories ni models previs. En segon lloc, en el regne de les dades massives, «la correlació és suficient» (Anderson, 2008), i no és necessari investigar els vincles causals entre variables associades. Per tant, la correlació substitueix la causalitat.

«Hi ha qui ha proclamat la "fi de la teoria", una nova etapa en la investigació científica basada en petabytes d'informació i en les supercomputadores»

El cert és que l'arribada de les dades massives comporta vertaderes novetats de tipus tecnològic. Aquestes no es caracteritzen només pel seu volum, velocitat i varietat, sinó també pel seu abast exhaustiu i resolució detallada, i per ser molt relacionals, a més de flexibles i escalables en producció (Kitchin, 2014). Les tècniques d'aprenentatge automàtic poden extraure dades i detectar regularitats sota el supòsit que «molt del que es genera no respon a cap pregunta en particular o és un subproducte d'una altra activitat» (Kitchin, 2014, p. 2). Utilitzant un enfocament col·lectiu, es poden aplicar diversos algorismes als conjunts de dades amb l'objectiu d'optimitzar-ne el rendiment predictiu. El que s'afirma en aquest cas és que està sorgint «un enfocament epistemològic totalment nou per a donar sentit al món». De fet, «en lloc de provar una teoria analitzant dades rellevants, les noves anàlisis de dades tracten d'obtenir informació "que naix en les dades"» (Kitchin, 2014, p. 2).

No hi ha dubte que l'enfocament de dades massives està contribuint a canviar el panorama epistèmic actual. A més, les tècniques de mineria de dades també estan creant noves oportunitats per a la investigació científica. Per exemple, existeix la possibilitat de comparar centenars de genomes del càncer i, gràcies a la seqüenciació d'ADN, establir la freqüència de moltes mutacions potencialment significatives per a diferents tipus de càncer, juntament amb les seues conseqüències funcionals: això pot fins i tot contribuir al desenvolupament de noves teràpies (Golub, 2010). En termes més generals, mitjançant aquestes tècniques és possible descobrir patrons potencialment significatius en grans volums de dades, alguns dels quals haurien passat desapercebuts anteriorment a causa de la seua complexitat.

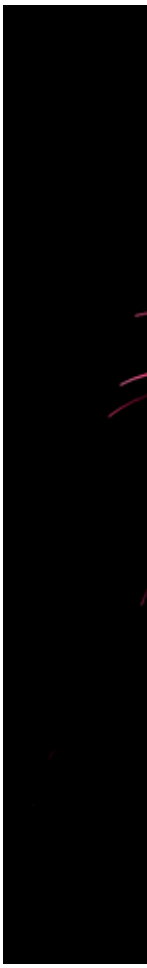
No obstant això, suposar que les dades massives representen un vertader canvi de paradigma epistemològic (almenys en el sentit que indicàvem anteriorment) és una qüestió completament diferent. De fet, no hi ha raó per a pensar que les dades massives permeten crear una nova manera de producció de coneixement en la qual els supòsits teòrics i les hipòtesis no complisquen cap paper i es puga ignorar la idea de causalitat.

Totes dues afirmacions sobre les dades massives han despertat fortes reaccions. Per exemple, atenent tant la generació com l'anàlisi de dades, observem que difícilment podem trobar una manera de crear coneixement sense necessitat de formular hipòtesis (és a dir, una manera que depenga únicament de la manipulació estadística i la inducció).

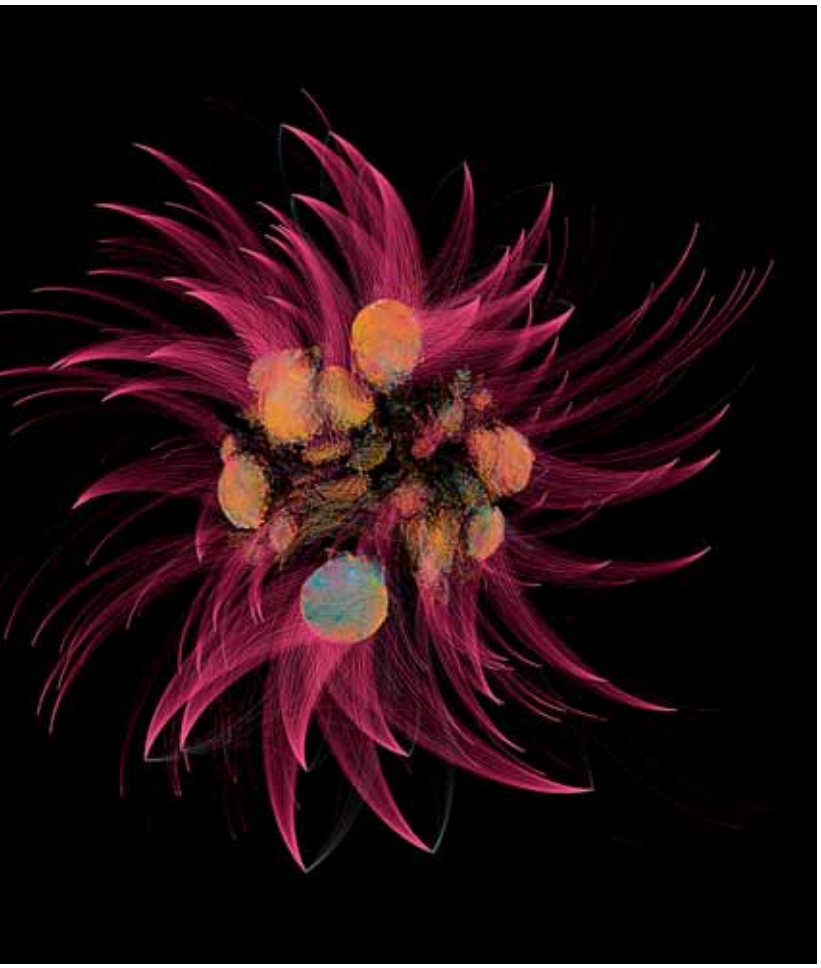
En primer lloc, les dades no sorgeixen del no-res. La filosofia de la ciència del segle XXI ha discutit extensament el paper que representen les nocions preconcebudes, començant per Karl Popper (1959, per exemple). En la seua opinió, les hipòtesis compleixen un paper essencial en la investigació científica, ja que ens indiquen què buscar i quines dades recopilar. Un altre argument conegut és la «saturació teòrica» de les dades i l'observació, és a dir, el fet que aquestes estiguen «contaminades» per presumpcions teòriques.

En realitat, la naturalesa no s'investiga a l'atzar. El que s'arriba a inspeccionar i mesurar està influït pel coneixement de fons, els interessos i les estratègies de l'investigador. Fins i tot el disseny d'experiments depèn de limitacions teòriques, metodològiques i tècniques específiques. Per tant, les dades sempre són el resultat de la interacció entre l'investigador (que pertany a una determinada escola de pensament) i el món, sempre que es complisquen les condicions materials adequades (Leonelli, 2015; Mazzocchi, 2015).

En segon lloc, les dades o les xifres no parlen per si mateixes. S'hi poden trobar regularitats significatives mitjançant computadores, però la qüestió és trobar-los



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



Selfie del microbioma, de François-Joseph Lapointe, després d'encaixar 650 mans durant la seua *performance 1.000 encaixades*.

una explicació. Això pressuposa l'existència d'un «marc d'anàlisi», una lent teòrica de la qual depèn com s'interpreten les dades: és ací on el paper del coneixement específic de domini resulta crucial. Boyd i Crawford (2012, p. 667) van indicar que «tots els investigadors són intèrprets de dades […]. Un model pot ser sòlid en termes matemàtics, un experiment pot semblar vàlid, però el procés d'interpretació comença així que l'investigador intenta entendre el que significa».

Diversos científics de dades, així com molts bioinformàtics en la disciplina de la biologia, pensen que entendre les estadístiques pot ser suficient per a donar sentit a les dades. Es pressuposa que els patrons són significatius per si mateixos, és a dir, que el seu significat transcendeix el context o domini, i no és necessari buscar fora de les dades. En la seua opinió, el coneixement teòric «depèn de generalitzacions reduccionistes que s'abstrauen de la realitat de manera problemàtica» (Chandler, 2015, p. 847). Per contra, l'enfocament computacional ens permetria accedir a conjunts de dades interconnectades i aconseguir una comprensió més holística –més enllà dels obstacles disciplinaris– de fenòmens complexos. No obstant això, és una mica paradoxal esperar que les dades, que s'han produït en un context concret (per exemple, la biologia), es puguen interpretar fàcilment exemptes de qualsevol context. Permeteu-me subratllar de nou aquest concepte: els coneixements específics de domini són importants.

A més, fins i tot els algorismes d'aprenentatge automàtic estan impregnats de suposicions particulars, com per exemple, què considerem un patró regular: cada algorisme té la seua pròpia manera de desenvolupar estratègies per a trobar relacions entre els conjunts de dades, i és probable que diferents algorismes troben diferents tipus de patrons (Hales, 2013). Això ho reconeixen fins i tot alguns especialistes en dades massives.

La segona afirmació –la idea que «la correlació és suficient»– exagera el valor de les prediccions realitzades a partir de correlacions. Potser hi ha circumstàncies particulars, com la publicitat, en les quals aquesta idea podria tenir sentit. No obstant això, probablement no és cert en el cas de la investigació científica.

Les correlacions poden suggerir connexions potencialment interessants. Poden fins i tot ser útils per a generar o avaluar noves hipòtesis, encara que aquesta tasca sempre estarà guiada per alguns supòsits teòrics subjacents i pel coneixement disponible (Kitchin, 2014). Però les correlacions no ens informen sobre la causa que subjau a aquestes relacions.

«Les dades o les xifres no parlen per si soles. La qüestió és trobar-los una explicació»

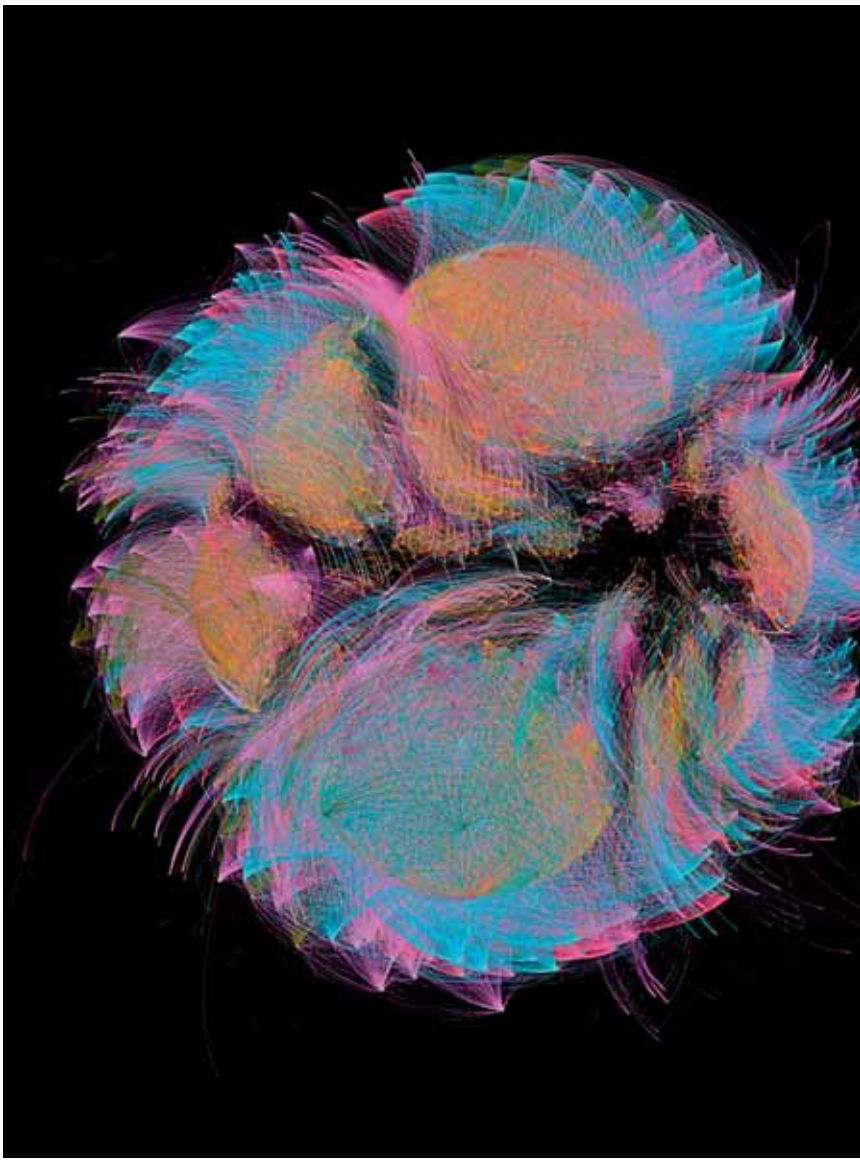
En la ciència, establir connexions causals és essencial, fins i tot per a saber com intervenir de manera efectiva en situacions de màxima prioritat, com per exemple per a curar una malaltia. Per tant, la investigació científica no es deté en les correlacions. Existeix la necessitat de realitzar anàlisis i proves addicionals: les correlacions s'han de «validar» d'alguna manera. El coneixement fiable tan sols es pot aconseguir al final d'aquest procés. Això depèn també del fet que, especialment en les bases de dades de grans dimensions, la majoria de correlacions són falsos positius (Calude i Longo, 2017). A causa del gran volum de dades, el problema és com afrontar la presència de moltes relacions de correlació, i distingir les connexions significatives de les confuses (les falses).

EL CAS D'EXPOsOMICS

Analitzem ara la història d'un cas de dades massives biomèdiques: el projecte EXPOsOMICS. Aquest projecte investiga les associacions entre exposició i malaltia en relació amb la nova idea d'exposoma, és a dir, la quantitat total d'exposició que afecta els individus durant tota la seua vida. Aquest concepte inclou l'exposició interna i externa: per exemple, en el seu estudi sobre el càncer de mama i de còlon, Chadeau-Hyam et al. (2011) van analitzar tant la dieta com l'estil de vida dels pacients (exposició externa), així com la seua resposta metabòlica (exposició interna).

Els biomarcadors exerceixen un paper clau en aquest tipus d'investigació. Aquests són elements mesurables de l'entorn i l'organisme que mostren processos biològics. De fet, es busquen les associacions entre biomarcadors d'exposició i biomarcadors de malaltia. Resulta significatiu que la investigació en biomarcadors se sol dur a terme utilitzant dades massives, habitualment obtingudes per mitjà de tecnologies d'alt rendiment: l'òmica en el cas de l'exposició interna, i els sensors, satèl·lits i altres recursos en el de l'exposició externa. Com va assenyalar Canali (2016, p. 4):

> EXPOsOMICS és un projecte de dades massives en el qual els científics busquen biomarcadors associats que puguen rastrejar l'exposició i la malaltia. Qui defense un enfocament basat en les dades podria dir que aquest projecte és l'exemple perfecte de la manera com la investigació en dades massives consisteix a recopilar grans quantitats de dades, analitzar-les, buscar correlacions entre biomarcadors d'exposició i biomarcadors de malaltia i realitzar prediccions. Això demostraria que les correlacions són suficients per si soles i que el coneixement causal no és necessari.



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

Selfie del microbioma, de François-Joseph Lapointe, després d'encaixar 850 mans durant la seua *performance 1.000 encaixades*.

Però això no és així. De fet, en l'estudi abans esmentat sobre el càncer de mama i de còlon, la cerca d'associacions en les dades per a identificar llistes de supòsits biomarcadors que relacionen exposició i malaltia és tan sols el punt de partida. Una correlació entre biomarcadors també es pot considerar significativa en termes estadístics, però el que s'intenta trobar és una relació causal entre l'exposició i la malaltia (Canali, 2016).

Per a això, existeix la necessitat de buscar biomarcadors «intermedis», que es pensa que poden estar involucrats com a causants de malalties. Es troben en la intersecció entre els biomarcadors d'exposició i els de malaltia. En el cas del càncer de còlon, la ingesta

«La naturalesa no s'investiga a l'atzar. El que s'arriba a inspeccionar i mesurar està influït per l'investigador»

de fibra s'identifica com un possible biomarcador intermedi. Tot aquest procés està guiat per una combinació de dades, proves estadístiques, principis teòrics, experiments previs i coneixement causal disponible sobre els mecanismes de la malaltia, per exemple utilitzant la Base de Dades del Metaboloma Humà, que conté informació sobre els mecanismes metabolòmics (Chadeau-Hyam et al., 2011).

En realitat, aquesta descripció del projecte EXPOsOMICS, i de molts més com ENCODE (vegeu Mazzocchi, 2015), mostra la invalidesa de les afirmacions de les dades massives sobre el final de la teoria i la prioritat de la correlació sobre la causalitat. Encara que a vegades la investigació científica puga començar a partir d'unes dades i, per tant, sense la participació de models o hipòtesis sòlids a priori, el coneixement teòric i experimental continua sent necessari immediatament després. A més, les consideracions metodològiques, així com l'elecció d'un tipus específic de model estadístic, exerceixen un paper essencial per a donar forma a la investigació i assegurar-se que l'anàlisi de dades és realment eficaç.

MÉS ENLLÀ DEL MITE DE LES DADES BRUTES I L'OBJECTIVITAT

Podem considerar l'afirmació que «les dades parlen per si mateixes» des d'un altre punt de vista. Si tenim en compte l'etimologia, el terme *data* és la forma plural de *datum* en llatí, que significa "alguna cosa donada", referit a "allò que es dona abans d'un argument" i no és necessari qüestionar. Es conceptualitzen les dades com a elements de naturalesa «preanalítica» i imparcial, se'ls té en compte com una reflexió directa o com una representació «nua» d'un aspecte en concret de la naturalesa, com si foren fotografies (Gitelman, 2013). Aquesta concepció queda encapsulada en el terme «dades brutes». Les dades massives compliquen la situació perquè l'objectivitat de les dades (com a elements concrets) es combina amb l'objectivitat o neutralitat dels patrons que naixen directament d'aquestes dades.

«Diversos científics de dades creuen que entendre les estadístiques pot ser suficient per a donar sentit a les dades»

No obstant això, hauríem d'entendre la naturalesa epistemològica de les dades d'una manera més sofisticada. Com ja s'ha assenyalat, les dades no estan determinades i mai estan nues; en certa manera, es «fabriquen». Com recorda Leonelli (2015, p. 820), «el que s'entén per dades sempre està relacionat amb una determinada investigació en la qual es busquen proves per a respondre, o fins i tot formular, una pregunta». Per tant, s'han de veure les dades com a artefactes socioculturals. A més, per a ser utilitzables i funcionar com a prova, sol ser necessari manipular-les i organitzar-les mitjançant una estructura de dades, i malgrat tot, fins i tot aquest procés està impulsat per consideracions teòriques i, per tant, està lluny de ser neutral (vegeu Gitelman, 2013).

El procés de generació i gestió de dades implica, en efecte, realitzar diverses eleccions i judicis –cadascun dels quals en certa manera esbiaixat– sobre, per exemple, què és significatiu o fiable i què no ho és. Aquestes consideracions es poden comparar, per exemple, amb la idea d'«oclusió ontològica» (Knobel, 2010), un mecanisme segons el qual la representació d'un objecte bloqueja qualsevol altra representació possible. A conseqüència d'això, els elements bloquejats no es tenen en compte i no «donen forma a la narrativa» en absolut. A la llum d'aquest enfocament, el procés d'admissió de dades, per exemple per a un arxiu, és un procés de bloqueig d'altres possibilitats. Com que no es pot superar la finitud de l'arxiu, es deixaran de considerar o representar diversos aspectes de la realitat (Bowker, 2014).

La percepció i la cognició humanes, que funcionen projectant límits en la realitat, també són al mateix temps mitjans per al descobriment i per al bloqueig. Es pot ordenar el món de diferents maneres basant-se en diferents formes de projectar els límits. No obstant això, el mecanisme comú que subjau en tots els casos és que per a crear un ordre concret o «visualitzar» alguna cosa és necessari excloure la resta d'opcions. En altres paraules, la nostra percepció i cognició són intrínsecament «perspectivistes».

Els filòsofs de la ciència contemporanis com Ronald Giere (2006) també han destacat el caràcter perspectivista de la ciència, és a dir, el fet que fins i tot l'observació i la teorització científiques només poden descriure el món natural a la llum d'una perspectiva determinada.

En aquest sentit, l'enfocament de dades massives, que postula un model d'objectivitat «no perspectivista», representa un pas arrere. Les declaracions d'objectivitat que defensen que l'anàlisi algorítmica de les dades garantiria la veritat i la neutralitat reflecteixen *de facto* la immaduresa filosòfica de la disciplina. Com va indicar Bollier (2010, p. 13):

> Com que és una gran massa d'informació sense processar, les dades massives no s'expliquen per si mateixes. I no obstant això, les metodologies específiques per a interpretar les dades estan obertes a tota mena de debat filosòfic. Poden les dades representar una

> «veritat objectiva», o qualsevol interpretació està necessàriament esbiaixada per un filtre subjectiu o per la forma en què es «netegen» les dades?

Per tant, hauríem de considerar que fins i tot la naturalesa mateixa de les dades és perspectivista. Ni les dades ni les xifres parlaran mai per si soles, només donen compte de les suposicions que porten incorporades. A més, pressuposar la neutralitat de les dades ja és, per si mateix, una posició no neutral.

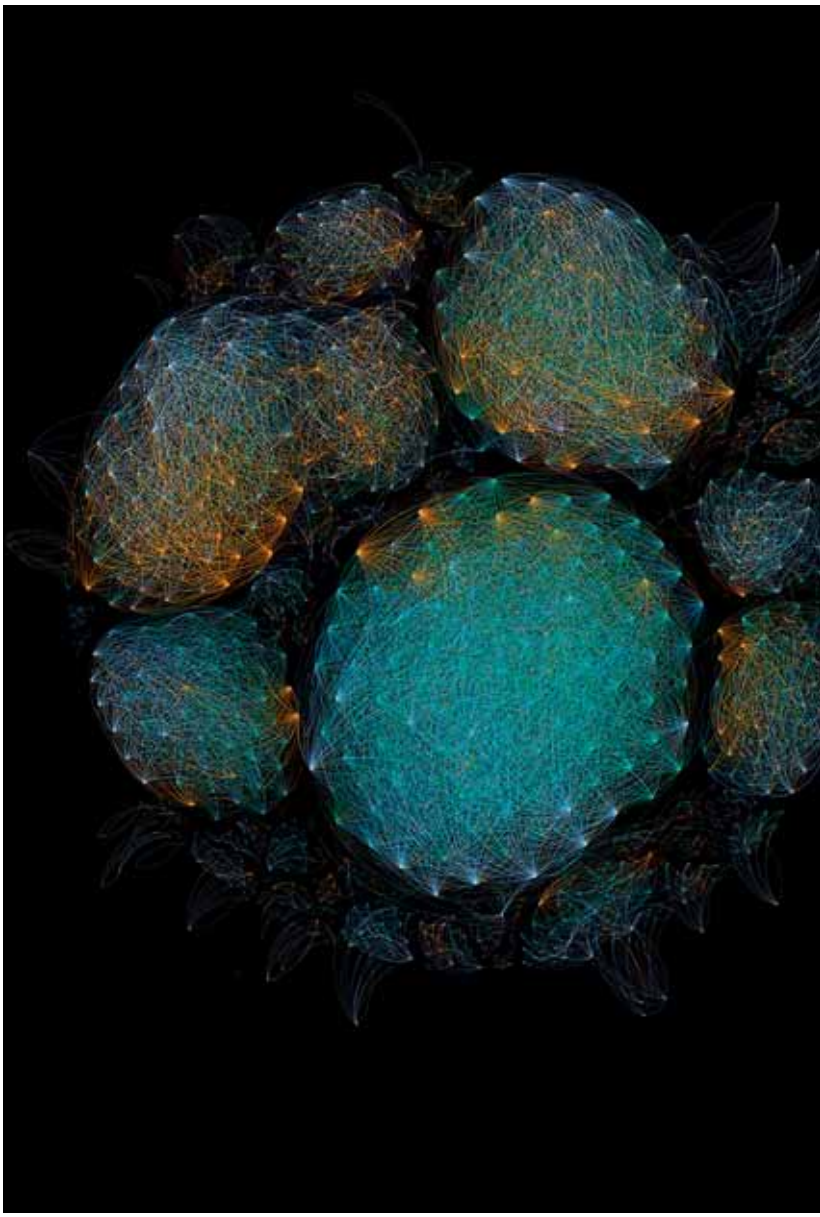
EL VALOR DE LES DADES MASSIVES COM A HEURÍSTICA I L'OPACITAT DELS ALGORISMES

Podem fer un bon ús de les novetats que ens ofereix l'anàlisi de dades massives. No obstant això, no podem esperar reemplaçar principis i procediments que s'han emprat i refinat al llarg de molts segles d'investigació científica. La ciència d'avui dia continua estant fonamentada en la teoria i l'experimentació, i és molt probable que ho continue estant en el futur. El valor de les dades massives és més aviat que representa un instrument heurístic molt potent i innovador.

Les dades massives i l'enfocament computacional ajuden a completar la caixa d'eines dels investigadors. La paraula clau ací és el *pluralisme*, perquè en augmentar el nombre d'eines heurístiques disponibles, és possible desenvolupar múltiples estratègies d'investigació que es complementen entre si. Per exemple, existeix la possibilitat de comparar i establir sinergies entre un enfocament d'hipòtesi i un de basat en dades. Potser en el futur arribem fins i tot a explorar noves maneres de desenvolupar teories. En qualsevol cas, un gran nombre de projectes de dades massives com EXPOsOMICS mostren que les dades i els elements teòrics s'«influeixen mútuament» i que tots dos participen repetidament en el cicle de la investigació científica (Canali, 2016, p. 8).

En conclusió, cal reafirmar que és necessari no acceptar de manera acrítica la cultura de l'algorisme que subjau a les dades massives. En cas contrari, fins i tot eines molt útils poden contribuir a crear una realitat no desitjada. De fet, els algorismes més refinats no són sols eines per a extraure informació. Cada vegada afecten més el mateix entramat de vides públiques i individuals i contribueixen en gran manera a donar-los forma:

> Avui vivim en un món en el qual els algorismes (i les dades amb què els alimentem) s'ocupen d'una gran varietat de decisions relacionades amb les nostres vides: no es tracta només de motors de cerca i sistemes personalitzats de notícies en línia, sinó que afecten



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

Selfie del microbioma, de François-Joseph Lapointe, després d'encaixar 350 mans durant la seua *performance 1.000 encaixades*.

«La ciència continua estant fonamentada en la teoria i l'experimentació. El valor de les dades massives és que representa un instrument heurístic molt potent i innovador»

també les avaluacions educatives, el funcionament dels mercats i les campanyes polítiques, el disseny dels espais urbans públics i fins i tot la manera com es gestionen serveis com la seguretat pública i les prestacions socials. Però es pot argumentar que els algorismes cometen errors i funcionen sobre la base de determinats biaixos. L'opacitat dels algorismes tècnics complexos que operen a gran escala fa difícil examinar-los, la qual cosa comporta una falta de claredat envers el públic, en relació amb la manera com exerceixen el seu poder i influència (Diakopoulos, 2015, p. 398).

Els algorismes, especialment els d'aprenentatge, són molt performatius i influents. No obstant això, és difícil comprendre'n el funcionament i implicacions. Ni tan sols els especialistes en treball de camp poden explicar completament què passa realment quan la màquina processa grans quantitats de dades per a obtenir informació nova, o la raó per la qual tria una manera de conducta en lloc d'una altra en determinades situacions (vegeu Burrell, 2016). Per aquesta raó, es descriuen com a «caixes negres».

«Hauríem de considerar que fins i tot la naturalesa mateixa de les dades és perspectivista. Ni les dades ni les xifres parlaran mai per si soles»

Aquesta opacitat per a la comprensió humana, que fins i tot reforça el «poder» dels algorismes, és deguda a qüestions tècniques i a la complexitat del seu funcionament. Una manera d'expressar-ho és en termes d'«opacitat epistèmica», és a dir, no és possible entendre tots els factors amb rellevància epistèmica implicats en les seues operacions (Humphreys, 2009).

En qualsevol cas, la creixent opacitat dels algorismes és una qüestió sobre la qual hem de meditar amb prudència. Avui dia, se celebra la performativitat de les eines de dades massives fins i tot amb triomfalisme. La potència epistèmica i la suposada neutralitat dels algorismes, que poden realitzar funcions inassolibles per a la ment humana, s'oposen a la fal·libilitat de la interpretació i la presa de decisions humanes. No obstant això, no hauríem d'usar la performativitat com a raó per a cedir l'autoritat i el control a les màquines.

En lloc de limitar-nos a elogiar l'enfocament de dades massives i els seus algorismes, hauríem de preguntar-nos una sèrie de qüestions. Per exemple, quin tipus de situació és aquella en la qual utilitzem eines per a realitzar determinades tasques complexes però no som capaços d'explicar com fan possible fer aquestes tasques? Ningú dubtaria que dispositius tecnològics com aquests influeixen en gran manera en la nostra representació del món. Per tant, una altra pregunta seria: quin tipus de situació és aquella en la qual hi ha eines capaces de donar forma a la manera com experimentem la realitat, però a la lògica subjacent de la qual i models de representació som incapaços d'accedir completament?

REFERÈNCIES

Anderson, C. (2008, 23 de juny). The end of theory: The data deluge makes the scientific method obsolete. *Wired.* Consultat 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 febrer). Lies, damned lies and big data. Consultat 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* (Tesi 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. Biòleg i filòsof. Investigador de l'Institut de Ciències del Patrimoni del CNR (Roma, Itàlia). La seua activitat d'investigació se centra en l'epistemologia (pluralisme epistèmic, perspectivisme), els problemes filosòfics de la investigació científica (com el debat reduccionisme-holisme en la biologia, la validació de models climàtics o els problemes epistemològics de les dades massives) i l'organització del coneixement. ✉ **fulvio.mazzocchi@cnr.it**