Los *gatekeepers* y los recursos de la investigación. Viejos desafíos y nuevas perspectivas en el tiempo de los *big data*[1]

Paolo Parra Saiani

Università degli studi di Genova, Italia

Correo electrónico: paolo.parra.saiani@unige.it

Resumen

Los recortes en la investigación, en particular en la social, generan cada tanto peticiones y advertencias de las principales asociaciones científicas internacionales. Al respecto, surgen las siguientes preguntas: ¿Cuáles son las consecuencias y los desafíos que plantea la reducción de los recursos disponibles para fines de investigación? ¿Los grandes volúmenes de datos (*big data*) pueden ser una respuesta, una solución, una forma de acceder a la información de una manera rentable? ¿O aumentarán la brecha entre universidades ricas y pobres y el nivel de desigualdad entre los investigadores? ¿Los grandes volúmenes de datos son una manera de sentar las bases de la sociedad de la información y del conocimiento, vislumbrado en las últimas décadas? ¿La importancia creciente de los grandes

[1] Este borrador está basado sobre (pero no es igual a) Parra Saiani (2016), y forma parte de una reflexión más amplia sobre las relaciones entre las técnicas de investigación y los vínculos exteriores, sean los del mercado o de la política.

volúmenes de datos en la investigación se acompaña de problemas que aún no están totalmente claros?

Estas preguntas llevan a plantear el objetivo de este trabajo, orientado a profundizar, por un lado, en el tema del acceso a la información. Los sociólogos estaban acostumbrados a recolectar datos preguntando y observando. Ahora enfrentan el surgimiento de un nuevo tipo de *gatekeeper*, privado, que no tiene ningún vínculo con la distribución de la información y que puede determinar el acceso en consideración del tipo de investigación y de sus preguntas. Por otro lado, en el artículo se considera el tema de los recursos disponibles para la investigación social hoy: a menudo se habla de los *big data* como si su característica intrínseca fuera su gratuidad, pero casi nunca se mencionan sus costos y los fondos necesarios para el acceso o para instalar un centro de conservación de estos. Como último punto, se analiza uno de los recursos más importantes disponibles para la investigación social: la capacidad de hacer (buenas) preguntas y la importancia de contar con un aparato teórico, componente que sigue vigente en la época de los *big data*.

Palabras clave: *big data*, desigualdades en la ciencia, investigación social, recursos.

Introducción

La expresión *big data* ha sido definida de distintas maneras: desde la observación banal por la cual los datos son grandes cuando son demasiados para entrar en un archivo excel o para ser archivados en un solo *server*; hasta descripciones más refinadas basadas en las características intrínsecas que diferencian los *big data* de otros tipos de datos (Kitchin, 2014, p. 1). Los tipos de información que siempre están más disponibles no son solamente números, sino también imágenes, videos, textos digitalizados. Por esta razón, Kitchin propone un conjunto de características de los *big data*: a) grandes dimensiones (terabytes o petabytes); b) rapidez y volatilidad, o sea creados y modificables en tiempo real; c) alto nivel de heterogeneidad; d) exhaustivos, porque derivan de una población o de un sistema considerado como unidad; e) alto nivel de detalle; f) flexibilidad. Entonces, los datos que respetan esas

características son las grandes bases de datos administrativos, que hacen que informaciones sobre temas distintos sean reconducidas al nivel de cada unidad de análisis: las huellas electrónicas dejadas para las personas que adquieren un servicio o un producto; las comunicaciones electrónicas —*social networks*, *blogs*, etc.—, datos sobre tipo y nivel de uso de objetos electrónicos —*smartphones*, *tablets*, etc.— (Kitchin, 2014, p. 2).

De Leonardis y Neresini (2015, p. 373) afirman que, en el estudio de "datificación" de la vida social, podemos distinguir tres trayectorias de investigación: a) el carácter performativo de los números, que con sus presencia omnipresente están asumiendo una carga normativa considerable, así como las formas y los significados específicos en diferentes contextos, sobre todo en la propia investigación sociológica; b) el control de los números, sobre su producción y uso —basta pensar en sus concentraciones en grandes cantidades en las bases de datos— como una forma de poder crucial, en la actualidad; c) las cifras del gobierno, especialmente cuando esconden la dimensión política de las elecciones, dando "objetividad mecánica" —según la expresión acuñada por Porter— a las más diversas tareas de gobierno. En otros trabajos he analizado el tema de la importancia de los números para el gobierno (Parra Saiani, 2009, 2011, 2012) y el carácter performativo de los números (Parra Saiani, 2015a, 2015b). Aquí propongo empezar a reflexionar sobre el control de los números y, en particular, de las consecuencias para la investigación social de la aparición de un nuevo tipo de *gatekeeper* para acceder a la información.

El acceso a la información

Muchas críticas se han hecho a los *big data*: la información podría ser utilizada para manipular a los consumidores, con el fin de competir de forma desleal en el mercado; aun más graves, se denuncian abusos que ponen en peligro la privacidad personal y las libertades civiles —basta pensar en el caso Prism que implica a la nasa—. En lugar de poner mi atención a *los números que controlan*, voy a considerar los problemas relacionados con *el control de los números*, es decir, el acceso a la información, por parte de los investigadores.

En el tiempo de los *big data*, la información que se supone interesante no es de los productores —el sujeto que, por ejemplo, accede a Facebook y pone un *like*

Mendoza, FCPYS-UNCUYO, 16 al 18 de noviembre de 2016
sitio web: http://elmecs.fahce.unlp.edu.ar - ISSN 2408-3976

sobre un post—: el mismo productor cede su derecho —en este caso, a Facebook—, a menudo sin imaginar su valor (Grimaldi, Gallina y Cavagnero, 2016), participando, de esta manera, en la perpetuación de un proceso casi invisible (Han, 2012, 2014). Entonces, la información no se recoge directamente por las personas, hablando, entrevistando, observando, etc., sino por parte de terceros: instituciones públicas o privadas que la recogen para sus fines. Estas instituciones terceras —con respecto a la universidad—, decidirán cómo y cuándo los datos serán disponibles para fines de investigación. Nos enfrentamos a verdaderos *gatekeepers*, que en lugar de filtrar la distribución del *output* (Hirsch, 1972) tendrán la facultad de regular el *acceso* a la información. Por su parte, boyd y Crawford (2012, pp. 674-675) señalan que en el sector privado algunas empresas limitan el acceso, otros venden los derechos y otros ofrecen una búsqueda limitada a la investigación académica, lo que produce injusticia sustancial en el sistema, ya vinculada con los desequilibrios en la asignación de recursos a disposición de las diferentes universidades.

Esta no es la primera vez en la historia que la información puede ser obtenida por terceros —basta pensar en el censo— o simplemente gracias a la acción de un intermediario —baste pensar en Doc de *Street Corner Society*—. No es la primera vez en la historia que la diferencia en el acceso a los recursos económicos afecta la posibilidad de realizar una investigación, así como su tipo y su extensión, etc. Pero, con los *big data* se está afirmando un sistema en el que los titulares/propietarios de la información —aparte el Estado y sus instituciones— son las compañías telefónicas, los comerciantes online, etc.: entidades sin obligación de publicarla, que pueden decidir a quién —y, especialmente, sobre la base de las preguntas de la investigación— permitir el acceso.

> The typical university or independent researcher is increasingly locked out […] The result, we fear, is a two-tiered system of research. Scientists working for or with big Internet companies will feast on humongous data sets – and even conduct experiments – and scholars who do not work in Silicon Valley (or Alley) will be left with proverbial scraps. (Conley et ál., 2015)

Mendoza, FCPYS-UNCUYO, 16 al 18 de noviembre de 2016
sitio web: http://elmecs.fahce.unlp.edu.ar - ISSN 2408-3976

Esta advertencia no viene de autores a la periferia del sistema de financiación, sino de un grupo de investigadores que ya se ha beneficiado de fondos de la Fundación Nacional de Ciencia. Conley es el primer sociólogo —y hasta ahora único— en recibir en el 2005 el Premio Alan T. Waterman, un premio de la misma National Science Foundation (NSF). A la advertencia siguió un silencio total: ni una institución reaccionó (Dalton Conley, comunicación personal).

Strong (2015) plantea preocupaciones similares, aunque en un texto que aborda la cuestión de los *big data* con entusiasmo. Si, entre las muchas ventajas identifica la posibilidad de acceder a un mundo de información extremadamente detallada imposible de conseguir de otras maneras, Strong —preocupado por la necesidad de interpretar los múltiples detalles— recuerda la oportunidad de involucrar a los científicos sociales y al mundo académico en general en el trabajo de análisis de las grandes empresas:

> Widespread access to data in organizations has been relatively limited for a number of good reasons, including commercial confidentiality and privacy concerns, along with a worry that spreading data too widely can lead to erroneous conclusions. But the downside of this is that too few people —and too few people with the right knowledge and understanding—are making sense of the data. What we now need to consider are models for managing the extraction of value from big data from both within the organization and externally. (Strong, 2015, p. 134)

Además, en un artículo publicado por quince distinguidos profesores de prestigiosas universidades, leemos que:

> Computational social science could become the exclusive domain of private companies and government agencies. Alternatively, there might emerge a privileged set of academic researchers presiding over private data from which they produce papers that cannot be critiqued or replicated. Neither scenario will serve the long-term public interest of accumulating, verifying, and disseminating knowledge. (Lazer et al., 2009, p. 721)

Mendoza, FCPYS-UNCUYO, 16 al 18 de noviembre de 2016
sitio web: http://elmecs.fahce.unlp.edu.ar - ISSN 2408-3976

La posesión de informaciones por parte de entidades terceras que no tienen alguna obligación de ponerlas a disposición hace más difícil y problemática la transparencia y la réplica de los estudios: uno de los pilares de la misma actividad científica que, habitualmente y de por sí, en las ciencias humanas es problemático. Si la ciencia es una actividad acumulativa y si todos estamos a hombros de gigantes, necesitamos saber cuáles son las bases sobre las que se está construyendo.

La disponibilidad de computadoras cada vez más rápidas y el consiguiente aumento de la capacidad de computación, además del fácil acceso a grandes cantidades de datos ya recogidos por otros, hace que sea más fácil hoy adoptar estilos de investigación diferentes a los del pasado, esto escribió Platt (1996, p. 139) cuando la velocidad de las computadoras y la cantidad de datos disponible no eran ni remotamente comparables a las de hoy. Y Platt, por supuesto, no se refería a los *big data*. Para contrarrestar la progresiva "privatización" de la información relevante han surgido muchas iniciativas, como el programa federal estadounidense *Open Government Initiative*, que establece la obligación de compartir los datos como una cláusula que se inserta en muchas propuestas de financiación; otras iniciativas del sector público y privado, a través de revistas científicas y fundaciones, han dado lugar a la disponibilidad sin precedentes de datos para el análisis secundario (Shaikh, Butte, Schully, Dalton, Khoury y Hesse, 2014).

La plataforma web Data.gov, por ejemplo, permite el acceso público a 191.711 *datasets*[2] organizados en 14 temas; otras bases de datos públicas como Gene Expression Omnibus (geo) permiten archivar y redistribuir datos útiles para la investigación genómica: 3.848 conjuntos de datos, 1.668.684 muestras[3]. Está claro que estas iniciativas hacen el acceso a las fuentes extremadamente fácil y barato. En general, el surgimiento y la consolidación de los *open data* hacen posible utilizar de manera gratuita muchos datos para mejorar la calidad de vida. Un ejemplo de estos son los servicios "gis-based" que pueden ayudar a gestionar la movilidad urbana en las grandes ciudades.

[2] Cfr. http://www.data.gov/, online 17/1/2016.
[3] http://www.ncbi.nlm.nih.gov/geo/, online 17/1/2016.

La investigación y su financiamiento: "Había una vez..."

Philip Hauser nos cuenta que, durante su investigación para el doctorado, pidió dos asistentes para conducir un estudio sobre las diferencias entre la fertilidad y mortalidad en Chicago; cuando terminó su estudio, era responsable de 150 empleados (Hauser, 1982, citado en Platt, 1996, p. 153). La anécdota es ilustrativa del clima que reinaba en esos años en los Estados Unidos, y hace menos extraño a nuestros ojos modernos el pedido de Whyte, recién licenciado, para diez asistentes por sus investigaciones sobre Cornerville, durante la Junior Fellowship en la Universidad de Harvard (1936-1940). Su solicitud fue rechazada, no por falta de fondos, sino porque L. J. Henderson, bioquímico y secretario de la Society of Fellows, no consideró oportuno asignar la coordinación de un grupo de diez investigadores a una persona que, desde su punto de vista, no conocía el campo de su investigación[4] (Whyte, 1993, p. 284).

¿Un momento feliz para la sociología? Sin duda, un período en el que —disipadas las primeras objeciones— logró establecerse en las universidades públicas, a partir de las presiones de ilustres financiadores: la resistencia del *board* de Columbia, por ejemplo, fue vencida gracias a una donación sustancial (Turner, 2014, p. 9). La Fundación Russell Sage financió en primer lugar la "Pittsburgh survey" y luego siguió alentando a las encuestas sociales (Bannister, 1987, p. 179). Las diversas fundaciones de Rockefeller invirtieron 40 millones de dólares en las ciencias sociales en la década de 1920, y continuaron con grandes sumas de dinero durante los años treinta (Ross, 2003, p. 227). Incluso la famosa escuela o tradición de Chicago es un producto de la generosidad de los Rockefeller y, de alguna manera, es un ejemplo típico (Chapoulie, 2001, p. 169; Turner, 2014, p. 28). Standard Oil financió un proyecto interdisciplinario ("Industrial Accidents") en Harvard y la publicación del libro de Parsons *La Estructura de la Acción Social* (1937). La investigación *An*

[4] Henderson propuso formar un equipo de investigación poco a poco, durante el trabajo en el terreno: "Henderson poured cold water on the mammoth beginning, told me that I should not cast such grandiose plans when I had done hardly any work in the field myself. It would be much sounder to get in the field and try to build up a staff slowly as I went along. If I should get a ten-man project going by fall, the responsibility for the direction and co-ordination of it would inevitably fall upon me, since I would have started it. How could I direct ten people in a field that was unfamiliar to me? Henderson said that, if I did manage to get a ten-man project going, it would be the ruination of me, he thought. Now, the way he put all this it sounded quite sensible and reasonable" (Whyte, 1993, p. 284).

American Dilemma por Gunnar Myrdal (1944) estuvo financiada por la Carnegie Corporation de Nueva York "que empleaba (y por lo tanto cooptó) un pequeño ejército de investigadores para escribir informes que fueron consultados en la redacción del libro final" (Turner, 2014, pp. 29-30).

El periodo actual es, sin duda, diferente. Sobre la financiación de las ciencias sociales y políticas de la nsf, leemos titulares como *Las ciencias sociales están bajo ataque en los Estados Unidos* (Boyle, 2013, p. 719), o *La guerra contra la ciencia política* (Zaino, 2013). Los republicanos estadounidenses presentan con regularidad leyes para reducir la inversión en ciencias económicas y sociales —en 2014, por ejemplo, "Frontiers in Innovation, Research, Science and Technology (First) Act", también conocida como HR4186—, para trasladarla a las ciencias físicas y biológicas. Además, la nsf debe justificar públicamente cómo los proyectos financiados pueden promover la seguridad nacional o los intereses económicos de los Estados Unidos.

Los recortes en la investigación y, en particular, en la investigación social, son comunes al otro lado del océano; las principales asociaciones científicas europeas lanzan periódicamente peticiones y alarmas para limitar los recortes y denunciar la supresión total de los programas de investigación. ¿Cuáles son las consecuencias y los desafíos que plantea la reducción de los recursos disponibles? ¿Los *big data* pueden ser una respuesta, una solución, una manera de acceder a la información de manera rentable? ¿O aumentará la brecha entre universidades ricas y universidades pobres, así como el nivel —ciertamente endémico; véase, por ejemplo, Price (1963) y Xie (2014)— de la desigualdad entre los investigadores? ¿Es una manera de sentar finalmente las bases de la sociedad de la información y del conocimiento que hemos visto solamente vislumbradas en las últimas décadas? ¿Acaso la importancia creciente de los *big data* en la investigación traerá consigo problemas que aún no son totalmente claros?

La expresión *big data* está referida muchas veces a la gran cantidad de datos —el *data deluge*, la avalancha de datos— disponible para los investigadores de cada área y disciplina en tiempo casi real: una grande, enorme matriz de datos lista para el análisis y la restitución de correlaciones estadísticamente significativas. Pero, habitualmente, no nos preguntamos cuáles son las condiciones económicas y

técnicas que hacen posible este logro: en particular, cuáles son los costes de los *big data* y las estrategias para el acceso a las bases de datos.

Las instituciones que financian la investigación han sido siempre una parte importante de su contexto social. La mayor parte del trabajo empírico no puede hacerse sin financiación y pocos sociólogos son capaces de recaudar fondos (Platt, 1996, p. 142). Por lo tanto, es particularmente importante aclarar en qué medida la investigación está condicionada por la disponibilidad de fondos. Mucho se ha escrito sobre el papel de las fundaciones en el nacimiento y desarrollo de la sociología (Brown, 1979; Arnove, 1980; Fisher, 1980, 1983, 1984, 1993; Bulmer, 1984; Alchon, 1985; Platt, 1996; Picó, 2001); pero mucho menos se ha estudiado su importancia en la actualidad.

Tampoco, el acceso a grandes volúmenes de datos es gratuito. La ambigüedad en relación con el costo del acceso a los recursos probablemente se debe al hecho de que, aunque la expresión *big data* sugiera consistencia y uniformidad dentro de sí mismo, es más bien un conjunto de aspectos diferenciados entre sí por contenido, estructura, propiedades y disponibilidad (Tinati, Halford, Carr y Pope, 2014, p. 664). Twitter, por ejemplo, hace que la información sea accesible a cualquier persona a través de la interfaz de los canales dedicados —*Application Programming Interface* (api)—, que permite el acceso a una pequeña selección de *tweets*, a una muestra de 10 % —*garden-hose*, o a todo el corpus (*firehose*)[5] —.

Pero Twitter —cuyos usuarios generan más de 400 millones de tweets cada día (Kumar, Morstatter y Liu, 2014, p. 1)— es una excepción en el paisaje de grandes volúmenes de datos, y no podemos sorprendernos, entonces, si ha despertado tanto interés entre los científicos sociales: entre el 2008 y el 2012 más de 110 publicaciones científicas sobre Twitter (Tinati et al., 2014, p. 664).

Pero, ¿es todo realmente gratuito? Con el aumento de tamaño de las bases de datos están empezando a aumentar los costos: Philip Bourne —First associate director for data science de la National Institutes of Health (NIH)— prevé que las

[5] Como recuerdan Tinati et ál., la elección de analizar un subconjunto así determinado del corpus disponible compromete las cualidades que hacen los mismos *big data* tan interesantes, en particular, el tamaño y dinamismo: "Rendering Big Data manageable in this way overrides its nature as 'big' data, bypassing the scale of the data for its availability or imposing an external structure by sampling users or tweets according to a priori criteria, external to the data themselves" (2014, p. 665).

primeras 50 bases de datos en orden de tamaño —sin contar el GenBank y otros de la National Library of Medicine (NLM)— necesitan aproximadamente 110 millones de dólares al año; otras nueve bases de datos grandes, financiadas por el National Human Genome Research Institute (NHGRI), con unos 30 millones de euro para el año 2015 solo tienen cuatro años para encontrar nuevas vías de financiación (Kaiser, 2016, p. 14); otras organizaciones privadas han establecido tarifas de acceso.

Por otra parte, Leonelli recuerda que todas las actividades de análisis de las bases de datos —de-contextualización, re-contextualización y reutilización (2014, pp. 3-5)— requieren una financiación considerable de mano de obra, además de la colaboración de la comunidad científica. Pero esto es poco frecuente, debido a las restricciones de confidencialidad a la que muchos investigadores están vinculados; de igual forma, el mismo sistema que recompensa la cantidad de publicaciones anuales haría que los científicos sean más competitivo y menos propensos a compartir las bases de sus trabajos (Leonelli, 2014, p. 6). A esto, se añade que los datos disponibles a través de la base de datos en la investigación genómica o biológica son solo una selección, y la mayor parte del trabajo de los científicos queda excluido (Leonelli, 2014, p. 6). Lo que es peor, esta selección no es el resultado de decisiones científicas, sino de factores técnicos, económicos, políticos y sociales. Grandes consorcios ganan fondos cada vez más importantes, en detrimento de las áreas más especializadas de investigación, pero menos "centrada en los datos" (Leonelli, 2014, p. 10), y esto profundiza la distancia entre países ricos y pobres, del norte y del sur, entre países que siempre están en el centro del proceso de investigación, y otros que están en la periferia, estos resultan útiles para externalizar el trabajo menos interesante.

El problema no reside únicamente en la biología o en la investigación sobre los genes. Mazzonis y Cini (1981, p 10) nos recuerdan que antes el físico construía —solo o con la ayuda de algunos técnicos— instrumentaciones diseñadas especialmente para poner de relieve un fenómeno considerado de especial interés; adaptaba los medios de investigación a un objetivo de conocimiento predeterminado. En el laboratorio que se encuentra alrededor de un gran acelerador, sin embargo, la relación entre el ser humano y la máquina se invierte. Ahora es la máquina que determina lo que puede y debe ser estudiado. Lo que podemos llamar el "availabilty

Mendoza, FCPYS-UNCUYO, 16 al 18 de noviembre de 2016
sitio web: http://elmecs.fahce.unlp.edu.ar - ISSN 2408-3976

bias" (Mahrt y Scharkow, 2013, p. 25) no es un fenómeno nuevo: por ejemplo, ya Pintaldi (2003, p. 19) afirmaba —hablando de los datos ecológicos— que la inmediata disponibilidad de datos, a menudo ya disponibles en matriz, lleva consigo el riesgo de suscitar en el investigador una actitud de negligencia en la selección de los indicadores, y a invertir el orden temporal entre la fase del proyecto de la investigación y la fase de la recolección de datos.

De esa manera, el investigador se contenta con la información a su disposición, sin averiguar si hay fuentes alternativas con informaciones más útiles, aunque menos accesibles. En un contexto internacional que ve cómo se reducen los recursos disponibles de los investigadores, es necesario preguntarse cuáles son las consecuencias del desarrollo de grandes volúmenes de datos y su difusión dentro de la investigación social. ¿Cuán tentador será estudiar todo a través de twitter o de la próxima herramienta de comunicación de moda?

En este sentido, es importante el cuestionamiento de Edward J. Hackett (2014, p. 637) sobre la necesidad de (re?)lanzar estudios para examinar la influencia del capital sobre la estructura, la conducta y el contenido de la investigación, sus prioridades y —debemos añadir— el método en general: "El sistema de recursos afecta significativamente el ciclo de la investigación social, la percepción de la situación del problema hasta la re-definición del problema, a través de todas las etapas técnicas de la investigación" (Cannavò, 2007, p. 38). Sin duda, el acondicionamiento asociado con la (no) disponibilidad de fondos siempre ha estado presente, no es una novedad, pero casi nunca se reconoce ni se menciona en la mayoría de los manuales de metodología de la investigación.

Daza (2012, p. 773) utiliza la expresión "cientificismo neoliberal" para describir la convergencia de la retórica comercial y de las orientaciones pre-kuhnianas a la ciencia. A medida que la actividad académica se vuelve más dependiente de los recursos financieros externos, "cede la libertad, el propósito y la capacidad de actuar como una fuerza moral independiente en la sociedad" (Hackett, 2014, p. 637). ¿Es esta una posición excesivamente fuerte? Puede ser, pero, seguramente, son muy ingenuas las posiciones de quienes afirman que actualmente "los científicos, aunque indirectamente expuestos a estrés externo sobre los programas experimentales a

seguir, pueden comunicar los resultados de su profesión sin restricciones" (Beretta, 2002, p. 62). Stephan (2012), por ejemplo, nos recuerda que la financiación de la investigación, específicamente la que está vinculada con la industria y los farmacéuticos, debe llevar a cabo acuerdos restrictivos sobre las publicaciones, con mayor frecuencia que sus colegas otras áreas.

Sin embargo, los programas de investigación reflejan lo que está financiado, por lo que no es sorprendente ver grandes sumas que se invierten en estudios criminológicos o gerontológicos y no en estudios sobre la pobreza o el racismo (Agger, 2000, p. 89). En estos y muchos otros casos, las presiones externas han tenido un papel clave —y también positivo, al menos para los beneficiarios de los fondos—. Pero la cuestión crucial, debido a sus implicaciones políticas, es: ¿cómo distinguir entre presiones legítimas e ilegítimas? (Mazzonis y Cini, 1981, p. 51). Reconocer la estrecha relación entre la ciencia como actividad social y el contexto social en que se desarrolla, nos conduce a reconocer la relación entre el nivel epistemológico y el nivel pragmático de la ciencia, así como, los cambios de los objetivos que regulan y caracterizan una sociedad.

Un último recurso: la teoría

La enorme cantidad de datos disponibles ha llevado a la propuesta de eliminar la teoría, demasiado abstracta y general para ser "medida objetivamente", bastante inútil, cuando se tienen millones de datos de los que extraer cientos de correlaciones estadísticamente significativas. "Forget taxonomy, ontology, and psychology […] With enough data, the numbers speak for themselves", escribe Anderson (2008) en Wired. Nótese la similitud con lo que escriben algunos autores a menudo considerados vecinos —con o sin razón— a Lundberg, como Stewart, que abogaba por la prioridad de la búsqueda de regularidades empíricas antes de la formulación de las hipótesis teóricas:

> When celestial mechanics was being developed, in the 16th and 17th Centuries, the order of advance was: (1) the collection of quantitative observations (Tycho Brahe); (2) their condensation into empirical mathematical regularities (Kepler); and (3) theoretical interpretation of the

Mendoza, FCPYS-UNCUYO, 16 al 18 de noviembre de 2016
sitio web: http://elmecs.fahce.unlp.edu.ar - ISSN 2408-3976

> latter (Newton). If there is to be a social physics, its beginnings must follow the same standard pattern. (1947, p 179)

En los mismos años, Zipf recogió grandes cantidades de datos, a menudo ya disponibles de otras fuentes, tales como los registros de las compañías telefónicas y del telégrafo, de los ferrocarriles y autobuses (1949, p. 383 y ss.), buscando regularidades empíricas:

> We may in all modesty to have increased the number of our observations to such a point that they may be viewed as empiric natural laws, regardless of the correctness of any of our theoretical interpretations. In other words, by means of the accepted methods of the exact sciences, we have established an orderliness, or natural law, that governs human behavior. (Zipf, 1949, p. 543)

Es cierto que muchos autores se distancian de esas afirmaciones tan extremas. Después de todo, Anderson era solamente el *editor-in-chief* de una revista, aunque influyente. Sin embargo, palabras que habrían podido ser consideradas solamente como una broma, a menudo se mencionan por parte de muchos investigadores y científicos para ensalzar la importancia de los big data, como hizo Atul Butte (2012), en una de las más seguidas conferencias: tedmed. En la opinión de Euan Ashley, cardiólogo y genetista entrevistado por la revista Stanford Medicine,

> We've been so focused on generating hypotheses […] but the availability of big data sets allows the data to speak to you. Meaningful things can pop out that you hadn't expected. In contrast, with a hypothesis, you're never going to be truly surprised at your result. (Conger, 2012, p. 11)

Mayer-Schönberger y Cukier, autores de un libro muy exitoso, traducido en muchos idiomas, escriben: "There is a treasure hunt under way, driven by the insights to be extracted from data and the dormant value that can be unleashed by a shift from causation to correlation" (Mayer-Schönberger y Cukier, 2013, p. 15); "No longer do we necessarily require a valid substantive hypothesis about a phenomenon

to begin to understand our world [...] In place of the hypothesis-driven approach, we can use a data-driven one" (p. 55).

Una nueva especie de *grounded theory* que despierta entusiasmo unánime, con el regreso de la metáfora de la mina de la que extraer datos, para celebrar el triunfo del método inductivo. Este ciertamente no es el lugar para tratar la importancia de la teoría en la investigación social, pero solo falta pensar en los muchos errores —incluso en los tiempos de *big data*— solo porque los datos habían sido mal interpretadas o leídos a la luz de una teoría improvisada. Sin embargo, si el tamaño de las bases de datos aumenta, mayor es la probabilidad de encontrar regularidades empíricas recurrentes y toparse con correlaciones espurias (Calude y Longo, 2016). Y agregamos que, en este sentido, la teoría es inevitable:

> The alternative for the scientist in the social or any other field is not as between theorizing and not theorizing, but as between theorizing explicitly with a clear consciousness of what he is doing with the greater opportunity that gives of avoiding the many subtle pitfalls of fallacy, and following the policy of the ostrich, pretending not to theorize and thus leaving one's theory implicit and uncriticized, thus almost certainly full of errors. (Parsons, 1938, p. 15)

Con esto, no se quiere negar la importancia de la posibilidad de utilizar una cantidad de información inimaginable hasta hace unos pocos años, pero no comparto (aún) el entusiasmo de Golder y Macy cuando dicen que las nuevas técnicas desarrolladas para observar la actividades en línea tendrán sobre las ciencias sociales un efecto similar al del microscopio electrónico, el telescopio espacial, el acelerador de partículas y de resonancia magnética, en sus respectivas disciplinas, que permiten analizar los objetos de estudio como nunca antes había sido posible (2014, p. 130). Las fuentes de información aumentarán, podrá haber más datos, pero habrá siempre la necesidad de una teoría, de un modelo interpretativo que explique las relaciones y los mecanismos causales.

De una ciencia siempre queremos respuestas, pero hay que recordar que cada respuesta está vinculada a una pregunta y parte de la tarea para las ciencias sociales

es favorecer la formulación de las preguntas correctas. La continua reducción de fondos a disposición de la investigación —social y de otras áreas de conocimiento— puede fortalecer la tendencia a estudiar *solamente* lo que ya está disponible, "ya listo". Hay que evitar que la disponibilidad de información sea el factor principal para determinar la elección del tema, en detrimento de su relevancia teórica, como se ha visto con el caso de Twitter. No se puede negar la utilidad de la abundancia de datos, pero tampoco se puede reducir todo el proceso de investigación a un mero proceso inductivo, descuidando deducción y abducción.

Algunas conclusiones

El *data deluge* del que mucho se habla es, sin duda —en comparación con hace solo diez años— real, concreto. Pero el uso de la metáfora del diluvio, como siempre cuando se utiliza una metáfora, puede ser engañoso. Engañoso por tres razones: a) asume implícitamente que es una cosa mala —a menos que tengamos un arca, un diluvio sigue siendo un problema bastante grave—; b) sugiere estar en presencia de un nivel más allá del cual no se puede ir; c) asume, implícitamente, que los datos son iguales, al menos en términos de fiabilidad de la información —cuando estás en medio de una inundación, no se va a controlar la diferencia entre una gota y la otra, de qué nube provienen, etc.—.

Una anécdota puede ayudar a ilustrar los tres aspectos. Hace algún tiempo, estuve presente en la entrevista de una joven médico a un paciente en un hospital en el norte de Italia, el paciente fue entrevistado con el fin de obtener informaciones sobre él y su esposa, así como de sus familiares; el objetivo era poner en contexto una enfermedad bastante rara, determinar antecedentes en familia, investigar los posibles orígenes genéticos, etc. Como sociólogo me gustó el intento de llevar a un contexto social los orígenes y las causas de la enfermedad; como metodólogo, no pude dejar de notar la aproximación con la que se obtuvo la información: puedo afirmar con seguridad que todas las "buenas reglas" que enseñamos —o tratamos de enseñar— a nuestros estudiantes han sido brutalmente ignoradas.

Mendoza, FCPYS-UNCUYO, 16 al 18 de noviembre de 2016
sitio web: http://elmecs.fahce.unlp.edu.ar - ISSN 2408-3976

Ignorados los efectos adversos de las respuestas "en caliente", o sea, inmediatas sobre la fiabilidad de las respuestas, la simpática interna hizo preguntas sobre las enfermedades del paciente, de sus padres y sus abuelos, así como de su propia esposa, sobre las enfermedades en la adolescencia y la pubertad; siempre para todos los componentes de la familia extendida.

En todo ello, olvidaba la necesidad de contextualizar temporalmente la pregunta, pidió el tipo de ciudad donde había vivido el paciente —y parientes, etc.—, "¿urbano o rural?". Cuando se le pidió una aclaración sobre cómo responder si "nació en una comuna urbana, vivió durante mucho tiempo en una zona rural, se transfirió de nuevo a un ambiente urbano", la respuesta fue "bueno, ponemos urbano". Y luego: ¿las áreas urbanas —o rurales—, son todas iguales entre sí? ¿Nacer —aunque solo el lugar de nacimiento sea importante y no tanto dónde pasó la mayor parte de su vida— en una gran metrópolis altamente contaminada, porque las industrias violan todas las leyes, es comparable con nacer en una ciudad de tamaño promedio, de vanguardia en la sostenibilidad del ambiente? Sin embargo, estos datos se alimentan a través de bases de datos en conexión constante y continua entre ellas, en un contenedor virtual más grande, formando, de hecho, los *big data*15[6].

Esta historia nos recuerda que: a) todavía existe la posibilidad de tener millones y millones de datos —en una inundación podemos refugiarnos en el arca, perdidos sin agua en un desierto —; b) aunque ya se hable de *big data*, los datos todavía podrían aumentar, afinando el proceso de recopilación de información, mejorando la conexión en red, a través bases de datos, que incluso en las ciencias sociales son cada vez más comunes; c) a menos que alguien desee desempolvar el viejo lema de sentido común acerca de los errores que se compensan —precisamente debido a la disponibilidad de un conjunto de datos que los investigadores de hace unas décadas no podían ni siquiera imaginar—, debemos prestar aún más atención a la calidad de los datos, al proceso de obtención de las informaciones. El viejo dicho aprendido en

[6] El problema de la fidelidad de la información es aún más relevante si tenemos en cuenta los cambios en los algoritmos de Google (86 cambios solo entre junio y julio del 2012), o con las plataformas Twitter o Facebook, que a menudo hacen que sea imposible comparar la información obtenida únicamente un año después, por no hablar de la manipulación fraudulenta con fines comerciales o financieras (Lazer, Kenndy, King y Vespignani, 2014, p. 1204).

el primer año de los cursos de estadística nos enseña que *garbage in, garbage out*. Puede ser *big*, pero, aun así, quedará *garbage.*

Referencias

Agger, B. (2000). *Public sociology. From social facts to literary acts*. Lanham: Rowman & Littlefield Publishers.

Agnoli, S. y Parra Saiani, P. (eds.) (2016). *Sulle tracce dei big data. Questioni di metodo e percorsi di ricerca*. *Sociologia e ricerca sociale* (número monográfico), XXXVII(109).

Alchon, G. (1985). *The invisible hand of planning*. Princeton: Princeton University Press.

Anderson, C. (2008). The end of theory: the data deluge makes the scientific method obsolete. *Wired*, (23). Consultado el 15 de enero del 2016 en http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory

Arnove, R. F. (ed.) (1980). *Philanthropy and cultural imperialism: the foundations at home and abroad*. Bloomington: Indiana University Press.

Baker, M. (2012). Gene data to hit milestone. *Nature*, (487), 282-283. doi:10.1038/487282a.

Bannister, R.C. (1987). *Sociology and scientism. The american quest for objectivity, 1880-1940*. Chapel Hill: University of North Carolina Press.

Beretta, M. (2002). *Storia materiale della scienza. Dal libro ai laboratori*. Milan: Bruno Mondadori.

boyd, D. y Crawford, K. (2012). Critical questions for big data. Provocations for a cultural, technological and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662-679. doi:10.1080/1369118X.2012.678878

Boyle, P. (2013). A U.K. View on the U.S. attack on social sciences. *Science*, (341), p. 719. doi:10.1126/science.1242563

Brown, E.R. (1979). *Rockefeller medicine men: medicine and capitalism in America*. Berkeley: University of California Press.

Bulmer, M. (1984). Philanthropic foundations and the development of the social sciences in the early twentieth century: a reply to Donald Fisher. *Sociology*, (18), 572-579. doi:10.1177/0038038584018004008

Butte, A. (2012). *What if you outsource three double-blind mice?* Washington: tedmed. Consultado el 20 de febrero del 2016 en http://www.tedmed.com/talks/show?id=7340

Calude, C. S. y Longo G. (2016). The deluge of spurious correlations in big data. *Foundations of science* (en prensa).

Cannavò, L. (2007). Introduzione. En L. Cannavò y L. Frudà (eds.), *Ricerca sociale. Dal progetto dell'indagine alla costruzione degli indici* (pp. 13-50). Roma: Carocci.

Chapoulie, J. M. (2001). *La tradition sociologique de Chicago. 1892-1961*. París: Seuil.

Conger, K. (2012). Big Data. What It Means for Our Health and the Future of Medical Research. *Stanford Medicine*, (Summer), 6-15.

Conley, D., Lawrence, J., Brady, H., Cutter, S., Eckel, C., Entwisle, B., … Scholz, J. (2015, 2 de febrero). Big Data. Big Obstacles. *The chronicle of higher education*, . Consultado el 20 de abril del 2016 en http://chronicle.com/article/Big-Data-Big-Obstacles/151421/

Daza, S. L. (2012). Complicity as infiltration: the (im)possibilities of research with/in nsf engineering grants in the age of neoliberal scientism. *Qualitative Inquiry*, *18*(9), 773-786. doi:10.1177/1077800412453021

De Leonardis O. y Neresini, F. (2015). Introduzione. *Rassegna Italiana di Sociologia*, *lvi*(34), 371-378. doi:10.1423/81796

Fisher, D. (1980). American philanthropy and the social sciences in Britain, 1919-1939: the reproduction of a conservative ideology. *Sociological Review*, (28), 277-315. doi:10.1111/j.1467-954X.1980.tb00366.x

Fisher, D. (1983). The role of philanthropic foundations in the reproduction and production of hegemony: Rockefeller foundations and the social sciences. *Sociology*, (17), 206-233. doi:10.1177/0038038583017002004

Fisher, D. (1984). Philanthropic foundations and the social sciences: a response to Martin Bulmer. *Sociology*, (18), 580-587. doi:10.1177/0038038584018004009

Fisher, D. (1993). *Fundamental development of the social sciences*. Ann Arbor: University of Michigan Press.

Golder, S. A. y Macy, M. W. (2014). Digital footprints: opportunities and challenges for online social research. *Annual Review of Sociology*, (40), 129-152. doi:10.1146/annurev-soc-071913-043145

Grimaldi R., Gallina A. y Cavagnero S. (2016). Uso della rete e consapevolezza delle tracce digitali. En S. Agnoli y P. Parra Saiani (eds.), *Sulle tracce dei big data. Questioni di metodo e percorsi di ricerca* (número monográfico). *Sociologia e ricerca sociale*, *xxxvii*(109) (en prensa).

Hackett, E. J. (2014). Academic capitalism. *Science, Technology, & Human Values*, *39*(5), 635-638. doi:10.1177/0162243914540219

Han, B. C. (2012). *Transparenzgesellschaft*. Berlín: Matthes & Seitz.

Han, B. C. (2014). *Psychopolitik: Neoliberalismus und die neuen Machttechniken*. Frankfurt: Fischer.

Hauser, P. M. (1982). Interview. En J. Platt (1996), *A history of sociological research methods in America. 1920-1960*. Cambridge: Cambridge University Press.

Hirsch, P. M. (1972). Processing fads and fashions: an organization-set analysis of cultural industry systems. *American Journal of Sociology*, *77*(4), 639-659. doi:10.1086/225192

Kaiser, J. (2016). Funding for key data resources in jeopardy. *Science*, *351*(6268), 14. doi:10.1126/science.351.6268.14

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

Kumar, S., Morstatter, F. y Liu, H. (2014). *Twitter Data Analytics*. Nueva York: Springer.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., … Van Alstyne, M. (2009). Computational social science. *Science*, (323), 721-723. doi:10.1126/science.1167742

Lazer, D. Kenndy, R., King, G. y Vespignani, A. (2014). Big data. The parable of Google Flu: traps in big data analysis. *Science*, (343), 1203-1205. doi:10.1126/science.1248506

Leonelli, S. (2014). What difference does quantity make? On the epistemology of Big Data in biology. *Big Data and Society*, (abril-junio), 1-11. doi:10.1177/2053951714534395.

Mahrt, M. y Scharkow M. (2013). The value of big data in digital media research. *Journal of Broadcasting & Electronic Media*, *57*(1), 20-33. doi: 10.1080/08838151.2012.761700

Mayer-Schönberger, V. y Cukier, K. (2013). *Big data. A revolution that will transform how we live, work, and think*. Boston: Houghton Mifflin Harcourt.

Mazzonis, D. y Cini, M. (1981). *Il gioco delle regole. L'evoluzione delle strutture del sapere scientifico*. Milan: Feltrinelli.

Myrdal, G. (1944). *An american dilemma. The negro problem and modern democracy*. Nueva York: Harper & Row.

NSF. Alan T. Waterman Award. Consultado el 25 de abril del 2016 en https://www.nsf.gov/od/waterman/waterman.jsp

Parra Saiani, P. (2009). *Gli indicatori sociali*. Milano: FrancoAngeli.

Parra Saiani, P. (2011). Knowledge and participation: which democracy? *Revista Latinoamericana de Metodología de las Ciencias Sociales*, *1*(2), 112-140.

Parra Saiani, P. (2012). Democracy and public knowledge. An issue for social indicators. En F. Maggino y G. Nuvolati (eds.), *Quality of life in Italy: researches and reflections* (pp. 225-242). Amsterdam: Springer-Social Indicators Book Series. doi:10.1007/978-94-007-3898-0_12.

Parra Saiani (2015a). La sociología frente a los nuevos ataques cientificistas. *Revista Latinoamericana de Metodología de las Ciencias Sociales*, *5*(1), 1-22.

Parra Saiani (2015b). Sobre la retórica cientificista: algunas consecuencias metodológicas y políticas del debate epistemológico. En A. Marradi (comp.), *Las ciencias sociales ¿seguirán imitando a las ciencias duras? un simposio a distancia* (pp. 185-201). Buenos Aires: Editorial Antigua.

Parra Saiani, P. (2016). Los gatekeepers y los recursos de la investigación. Viejos desafíos y nuevas perspectivas en el tiempo de los big data. *Revista Colombiana de Sociología*, 39(2), 221-240. doi: http://dx.doi.org/10.15446/rcs.v39n2.58973

Parsons, T. (1937). *The structure of social action. A study in social theory with special reference to a group of recent European writers*. Nueva York: McGraw-Hill.

Parsons, T. (1938). The role of theory in social research. *American Sociological Review*, *3*(1), 13-20.

Picó, J. (2001). El protagonismo de las fundaciones americanas en la institucionalización de la sociología (1945-1960). *Papers*, (63/64), 11-32.

Pintaldi, F. (2003). *I dati ecologici nella ricerca sociale. Usi e applicazioni*. Roma: Carocci.

Platt, J. (1996). *A history of sociological research methods in America. 1920-1960.* Cambridge: Cambridge University Press.

Price, D. J. (1963). *Little science, big science*. Nueva York: Columbia University Press.

Ross, D. (2003). Changing contours of the social science disciplines. En T. M. Porter y D. Ross (eds.), *The Cambridge history of science*. Vol. 7. The Modern Social Sciences (pp. 205-237). Cambridge: Cambridge University Press.

Shaikh, A. R., Butte, A, Schully, S., Dalton, W., Khoury, M. y Hesse, B. (2014). Collaborative biomedicine in the age of big data: the case of cancer. *Journal of Medical Internet Research*, *16*(4), e101. doi:10.2196/jmir.2496

Stephan, P. E. (2012). *How economics shapes science*. Harvard: Harvard University Press.

Stewart, J. Q. (1947). Suggested principles of social physics. *Science*, *106*(2748), 179-180. doi: 10.1525/aa.1947.49.4.02a00300

Strong, C. (2015). *Humanizing big data. Marketing at the meeting of data, social science and consumer insight*. Londres: Kogan Page.

Tinati, R., Halford, S., Carr, L. y Pope, C. (2014). Big data: methodological challenges and approaches for sociological analysis. *Sociology*, *48*(4), 663-681. doi:10.1177/0038038513511561

Turner, S. P. (2014). *American sociology from pre-disciplinay to post-normal*. Basingstoke: Palgrave Macmillan.

Vigen, T. (2015). *Spurious correlations*. Nueva York: Hachette Books.

Whyte, W. F. [1943] (1993). *Street corner society. the social structure of an italian slum*. Chicago: Chicago University Press.

Xie, Y. (2014). "Undemocracy": inequalities in science. *Science*, (344), 809-810. doi:10.1126/science.1252743

Zaino, J. (2013). The war against political science. *Inside Higher Ed.*. Consultado el 25 de abril del 2016 en https://www.insidehighered.com/blogs/university-venus/war-against-political-science

Zipf, G. K. (1949). *Human behavior and the principle of least effort. an introduction to human ecology*. Cambridge: Addison-Wesley Press.