Tipos de big data y análisis sociológico: usos, críticas y problemas éticos

Types of big data and Sociological Analysis: Uses, Criticisms and ethics

Biagio Aragona

Universidad de Nápoles Federico II
aragona@unina.it (ITALIA)

Recibido: 14.07. 2020
Aceptado: 03.12.2021

RESUMEN

Solo con un conocimiento más consciente de los diferentes tipos de big data y sus posibles usos, límites y ventajas la sociología se beneficiará realmente de estas bases empíricas. En este artículo, a partir de una clasificación de los diversos tipos de big data, se describen algunas áreas de uso en la investigación social destacando cuestiones críticas y problemas éticos. Los límites se vinculados a cuestiones fundamentales relativas a la calidad de los big data. Otra cuestión clave se refiere al acceso. Otro aspecto metodológico a tener en cuenta es que los datos digitales en la web deben considerarse no intrusivos. Los métodos de investigación encubiertos han desafiado la práctica de evaluación ética establecidas adoptadas en la mayoría de las instituciones de investigación: el consentimiento informado. Las pautas éticas digitales no pueden ser universales y estar establecidas de una vez por todas.

PALABRAS CLAVE

Big data, Cualidad de los datos, Postdemografia, Consentimiento informado, Search as research.

ABSTRACT

Only through expert knowledge of the different types of big data and their possible uses, limits and advantages will sociology benefit from these empirical bases. In this article, based on a classification of the various types of big data,

some areas of use in social research are described, highlighting critical questions and ethical problems. The limits are related to fundamental questions regarding the quality of big data. Another paramount issue concerns access. A further methodological aspect is that digital data on the web should be considered non-intrusive. Covert research methods have challenged the established ethical evaluation practice adopted in most research institutions: informed consent. Digital ethical guidelines cannot be universal and established once and for all.

KEY WORDS

Big Data, Data Quality, Postdemography, Informed consent, Search as research.

1. INTRODUCCIÓN

Actualmente el término *big data* se considera una palabra de moda, que se usa ampliamente, pero con poca precisión. En realidad, *big data* se refiere a un conjunto de datos muy diferentes en relación con actores sociales específicos y áreas de aplicación potenciales. Sólo con un conocimiento más consciente de los diferentes tipos de *big data* y sus posibles usos, límites y ventajas la sociología se beneficiará realmente de estas nuevas bases empíricas. Rechazar el uso de grandes datos porque no es familiar, o porque no se producen de acuerdo con los estándares de calidad a los que estamos acostumbrados, o aceptarlos sin crítica, sin reflexionar sobre qué actores sociales los están produciendo, sin embargo, no permitirá que la sociología se beneficie de los aportes del conocimiento que podrían derivarse de las "revolución de los datos" (Kitchin, 2014). En cambio, es necesario adoptar una postura creativa y al mismo tiempo crítica con respecto a su uso en los diseños de investigación social. Esta postura debe tener como objetivo probar las características de los diferentes tipos de *big data*, identificando lo que pueden ofrecer a la investigación social y lo que no pueden ofrecerles. En este artículo, a partir de una clasificación de los diversos tipos de *big data*, se describen algunas áreas de uso en la investigación social destacando cuestiones críticas y problemas éticos.

Pero, aunque el término *big data* se utiliza cada vez más, sigue identificando un conjunto muy amplio de datos digitales almacenados con fines administrativos, comerciales o científicos, que en realidad tienen características muy diferentes. La principal fuente de *big data* es definitivamente Internet. El ejemplo más emblemático e inmediato está representado por los medios de comunicación y todos los contenidos generados por el usuario (Boccia Artieri, 2015) pueden ser de diferentes formas: textos, imágenes, vídeos, url, etc.. Se trata de datos producidos voluntariamente y que se insertan en un sistema o en una plataforma digital. Entre ellos figuran los contenidos de blogs y sitios web que, a partir de datos en forma no estructurada, generalmente en formato HTML, se transforman mediante

técnicas de *web scraping* (Turland, 2010)– que permiten extraer datos de un sitio web mediante programas informáticos que simulan la navegación humana en la World Wide Web– en metadatos que pueden almacenarse y analizarse localmente en una base de datos.

Otro efecto de la extensión de la web son los datos producidos por la *internet de las cosas*. La internet de las cosas (o, más propiamente, la *internet de los objectos*) es un neologismo que se refiere a la extensión de la *red* al mundo de los objetos y los lugares concretos. Las cosas comunican datos sobre sí mismas y acceden a datos de otros dispositivos. Los objetos pueden adquirir un papel activo gracias a su conexión a la red. Por ejemplo, los que tienen etiquetas (identificación por radiofrecuencia (Rfid) o códigos QR) comunican los datos a través de la red o a los dispositivos móviles. Entre los ejemplos se incluyen los datos de los sensores, tanto fijos–como los sistemas de automatización del hogar (contadores de calorías, presencia, fuego, gas), sensores médicos (frecuencia cardíaca, presión sanguínea, glucosa en sangre, etc.) o los de vigilancia de fenómenos externos (clima, contaminación, tráfico)–como móviles–como los sistemas de localización montados en los teléfonos inteligentes, o los sensores (de aparcamiento, de frenado, etc.) instalados en los coches que intervienen para ayudar al conductor.

Una última categoría de datos producidos en la web son los rastros que se dejan en Internet cada vez que se navega en un sitio o una página web en particular. Un ejemplo son los registros (logs), es decir los archivos que registran los eventos o mensajes que ocurren en un sistema operativo, software o aplicación– como los que se crean cuando se realiza una búsqueda en un motor de búsqueda por ejemplo en Google. El procesamiento de estos datos con técnicas estadísticas permite analizarlos en tiempo real, produciendo informes específicos sobre un determinado grupo de eventos, que se registran por separado y se presentan directamente en el sistema. Por ejemplo, en el entorno de los MOOC (*Massive Open On-line Courses*), los datos de registro se utilizan para producir informes específicos sobre la inscripción en los cursos de los estudiantes, de modo que el sistema genera informes automáticos sobre los flujos de inscripción en tiempo real y alerta cuando el número de participantes en un curso alcanza un determinado umbral.Este tipo de datos son rastros, un subproducto de otras acciones, enfocadas y orientadas de manera diferente, que quedan impresas empíricamente en la realidad. La diferencia entre la información que se comunica (da) intencionalmente, como la de los usuarios de las redes sociales, y la información que se deja (emite) accidentalmente dentro de los ecosistemas digitales, parece ser profunda, y esto también desde el punto de vista del análisis sociológico.

La información que está en las pistas, es decir, la información que no tenía la intención de ser informativa, es muy buscada. La oportunidad de interpretar esas huellas que las personas dejan es el verdadero gran desafío de la sociología digital. Históricamente, el origen del interés de los investigadores sociales por las trazas (Webb et al, 1966) empezó de la necesidad de contrastar las grandes encuestas muestrales con una forma de hacer investigación más abierta al uso integrado de diferentes dispositivos, con el fin de compensar las inevitables

distorsiones introducidas con la solicitud directa de información a los sujetos involucrados en las investigaciones. El uso de este material empírico en particular requiere reflexionar sobre al menos dos aspectos del método relevantes para cada diseño de investigación. En primer lugar, reactividad. Molestar o no el tema de nuestro análisis es una elección importante que debe hacerse pensando en las posibles consecuencias que nuestra intrusión puede tener sobre el comportamiento de las personas que vamos a estudiar. En un escenario en el que la captación de sujetos para la administración de cuestionarios será cada vez más difícil --como lo demuestra el aumento de no respuestas incluso en las grandes encuestas internacionales de oficinas internacionales de estadística (Eurostat, 2020) -- las huellas dejadas en la red parecen ser fundamentales para seguir estudiando los comportamientos de los individuos. Otro aspecto a considerar es la profundidad. ¿Qué se puede entender del comportamiento social a partir de las huellas presentes en la red? ¿Qué tan profundo es el análisis que se puede hacer al respecto? A menudo, es posible rastrear una serie inconmensurable de comportamientos, pero que muchas veces no es posible vincular a información de contexto fundamental que serviría para aumentar su profundidad informativa.

Otro conjunto de datos se produce por las innumerables transacciones que tienen lugar entre los dispositivos digitales conectados. Un ejemplo son los datos que se crean cada vez que realizamos cualquier comunicación (llamada, mensaje de texto, mensaje de vídeo) en nuestro teléfono móvil o *smartphone*. Para cada evento de comunicación se genera un conjunto de datos que indica, por ejemplo, la fecha, la hora, el número de teléfono del destinatario, la duración de la llamada o el número de caracteres utilizados en el mensaje. Este tipo de datos son también los datos de los movimientos realizados con tarjetas de crédito y débito, o los datos generados por los lectores ópticos de códigos de barras utilizados en los supermercados. Las oficinas de estadística de diversos países han llevado a cabo interesantes experimentos sobre estos últimos para utilizarlos en la elaboración de estadísticas oficiales. Por ejemplo, el Instituto Italiano de Estadística (ISTAT) utiliza los datos de los escáneres de las principales cadenas de supermercados para equilibrar los resultados de la encuesta de precios al consumidor. La actualización de los datos de los escáneres permite la vigilancia de los precios en tiempo real y también hace que el estudio de los precios sea más preciso que los datos de la investigación tradicional. De hecho, los datos de los lectores ópticos permiten detectar con precisión el número de ventas. Sin embargo, a pesar de la mejora de la precisión, los datos del escáner no son adecuados para responder a preguntas específicas; por ejemplo, no ayudan a explicar por qué los consumidores prefieren determinados productos. En resumen, pueden carecer de validez. El mismo riesgo se aplica también a los datos de los teléfonos móviles. Siguen existiendo algunas limitaciones importantes al tratar de utilizar esos datos para explicar los movimientos de la población. Aunque permiten el análisis en tiempo real, estos datos no ofrecen la posibilidad de obtener más información sobre las características de los individuos vigilados, y sobre las motivaciones que impulsan sus movimientos (turismo, trabajo, familia, etc.), los medios de transporte

utilizados y muchas otras informaciones que serían pertinentes para estudiar la movilidad de los individuos.

Una última categoría de datos es la generada por las denominadas infraestructuras de datos: catálogos, archivos, sistemas de información y portales que, a través de diversos procedimientos de agregación, fusión y unión, forman bases de datos mucho más voluminosas que antes. Ejemplos de ello son los datos de los sistemas de información que fusionan los datos administrativos con los de las *encuestas* o los censos, o los datos gestionados por archivos de datos de ciencias sociales como el CESSDA (Central European Social Science Data Archives). Este tipo de grandes datos es tradicional en muchos sentidos. Se trata de datos numéricos estructurados que son el resultado de definiciones operacionales establecidas previamente por los fabricantes. También se construyen los metadatos necesarios para las fusiones e integraciones entre las diferentes bases de datos. Debido a sus características tradicionales, con excepción del volumen, también se han definido los *data that are getting bigger* (Aragona, 2016) o *small big data* (Gray *et al.* ,2015).

2. DILUVIO DE DATOS Y ANÁLISIS SOCIOLÓGICO

La primera consecuencia del «diluvio de datos» (The Economist, 2010) sobre la sociología es lo que se ha definido como la crisis del análisis sociológico empírico (Savage y Burrows, 2007). En primer lugar, a medida que aumentaba la cantidad de datos, las tradicionales bases empíricas cuantitativas de las ciencias sociales (encuestas y experimentos) fueron sustituidas por grandes análisis de datos. Por ejemplo, la investigación de mercado se lleva a cabo ahora en comunidades en línea, en lugar de mediante encuestas por muestreo, y *network analysis* y *sentiment analysis* están sustituyendo a las tradicionales encuestas electorales; por no mencionar lo mucho que ha cambiado el análisis documental con la llegada de los *big corpora* (Amaturo y Aragona, 2017). Además, nuestra disciplina ha perdido la propiedad de la recopilación y análisis de datos sociales. Nuevos estudiosos, principalmente informáticos con poco o ningún conocimiento sociológico, han comenzado a trabajar en el análisis social y a sacar conclusiones de estas enormes bases de datos.

Una primera reacción a esta crisis fue un notable cambio de perspectiva en comparación con el método que se utilizaba generalmente en la investigación social cuantitativa, basado en el modelo hipotético-deductivo (Hempel, 1942). Extraer conocimientos de un «diluvio» de datos no es simplemente un problema técnico (Floridi, 2012), sino que cambia la forma en que se formulan las preguntas de la investigación, y cómo se buscan las respuestas a estas preguntas. Por lo tanto, ha surgido una visión de las ciencias sociales basada en los datos, disminuyendo el valor y el papel de las hipótesis en el proceso de investigación. Concretamente, Lazer, en el artículo *Life in the network, the coming age of computational social science (La vida en la red, la nueva era de las ciencias sociales computacionales*) apareció en *Science in* 2009– un artículo que tuvo mucho

éxito en la comunidad científica (más de 3403 citas)– identificó los grandes datos en Internet como el núcleo de las *Ciencias Sociales Computacionales*, una disciplina que a través las técnicas de *data mining* y *machine learning* aplicadas a enormes bases de datos produce análisis sociales a gran escala, y casi en tiempo real.

Aunque no se puede dejar de observar cierto reduccionismo en la perspectiva de las *ciencias sociales computacionales,* no deja de ser cierto que una propuesta científica más basada en los datos puede abrir nuevas oportunidades para la sociología, porque hay menos limitaciones debido a las estructuras de referencia teórica de las distintas disciplinas, y favorece la interdisciplinariedad. La forma de la ciencia social computacional que se ha establecido es a menudo criticada por no ser capaz de explicar la complejidad social, sino sólo de describirla. Chris Anderson, de manera provocativa, escribió en 2008: «Los petabytes nos permiten decir: 'La correlación es suficiente'... La correlación reemplaza a la causalidad, y la ciencia puede avanzar incluso sin modelos coherentes y teorías unificadas». Pero una cosa es identificar las regularidades dentro de los datos, y otra es descubrir los mecanismos que las generan. Esta última operación no puede hacerse sin una teoría y un conocimiento profundo y contextualizado de su objeto de investigación. Es a partir de esta convicción que se están desarrollando nuevas propuestas de ciencias sociales intensivas en datos, como la ciencia de datos sociales (Lauro et al., 2017) que vinculan las habilidades informáticas y estadísticas con el dominio del conocimiento (sociológico) en el que se emplean, lo que trae consigo sus teorías y visiones de la realidad. De esta manera, los conocimientos típicos de una determinada ciencia se integrarían y combinarían con las disciplinas técnicas formales necesarias para atravesar la era de los grandes datos.

La razón por la que se pensó que el big data ponía en crisis las encuestas es doble. Primero, la encuesta estaba obteniendo tasas de no respuesta cada vez más altas. Los porcentajes de no respuesta han aumentado enormemente (en el caso de las encuestas telefónicas comerciales incluso hasta el 90%). El principal problema es la carga estadística (Struijs, Braaksma, Daas, 2014), la exasperada solicitud de información de la población para encuestas estadísticas. En los países estadísticamente más avanzados, el cuestionario y la encuesta por muestreo se han convertido en métodos de investigación generalizados utilizados por empresas privadas y organismos públicos, lo que ha generado molestias en la población así como desconfianza ante la posibilidad de un uso indebido de la información (Amaturo, Aragona , 2012). La carga estadística ha sido un problema importante para las oficinas de estadística nacionales e internacionales porque puede afectar la calidad de los datos recopilados a través de cuestionarios y entrevistas (Machin, 1998).

La segunda razón por la que se pensó que los macrodatos reemplazaban a las encuestas es que se creía ingenuamente que las redes sociales y las plataformas digitales involucran mundos sociales enteros, que podrían estudiarse rápidamente y a bajo costo. Esta concepción epistemológicamente algo ingenua de Big Data, sin embargo, no logró problematizar algunas cuestiones metodológicas

cruciales. La idea de que Big Data permite observar el desarrollo «natural» de las actividades humanas desde arriba es consecuencia de un positivismo ingenuo que olvida la mediación socio-técnica de los datos digitales Los datos digitales se generan en el curso de actividades (por ejemplo, comprar un producto en Amazon) o interacciones comunicativas (por ejemplo, chatear en WhatsApp), en situaciones sociales públicas, semipúblicas o aparentemente privadas, mediadas por arquitecturas específicas y los algoritmos del plataformas que los albergan. El hecho de que la generación socio-técnica de datos digitales no sea controlable por el investigador puede provocar errores estadísticos y su descontrol.

Obviamente, los límites de la *ciencia social computacional* son muchos y algunos de ellos están vinculados a cuestiones fundamentales relativas a la calidad de los big data. Por ejemplo, la cuestión de la representatividad estadística y el muestreo en general es una de las primeras cuestiones que se plantean. Entre esos datos hay problemas generalizados de insuficiencia de cobertura, que se producen cuando algunas unidades de la población son excluidas sistemáticamente de las investigaciones. Los investigadores sociales, al utilizar datos de la Internet, deben recordar que parte de la población sigue siendo inalcanzable por definición; porque tal vez no tengan acceso a la Internet, o porque muchos son simplemente consumidores pasivos de la información contenida en la Internet, en lugar de usuarios que participan activamente en la Web 2.0. Además, el acceso a la red puede segmentarse en relación con variables sociodemográficas como la nacionalidad, la edad, el sexo, el nivel de educación y los ingresos, lo que lleva a una subestimación sistemática de estratos enteros de la población.

Otra cuestión clave se refiere al acceso. El acceso a los *big data* puede concederse a unos y no a otros; en relación con la influencia, el presupuesto y los objetivos que tenga el investigador. Boyd y Crawford (2012) observaron que en el sector privado algunas empresas limitan el acceso a todos; otras venden derechos por una cuota y otras ofrecen pequeños conjuntos de datos creados específicamente para la investigación académica: «esto produce una considerable desigualdad en el sistema: los que tienen dinero -o los que trabajan en ciertas empresas- pueden acceder a una base empírica diferente de los que tienen poco dinero o están fuera de la empresa». (Boyd y Crawford, 2012: 674). Pero lo más importante de todo es que cuando se usan datos que no son producidos personalmente por el investigador, las posibilidades de redirigirlas a los objetivos de la investigación son limitadas. Estos datos son construidos por actores sociales específicos con objetivos muy diferentes de la investigación, por lo tanto, puede haber grandes diferencias entre lo que le gustaría al investigador y lo que tiene en su lugar. Es precisamente la estimación de estas diferencias, y cómo pueden impactar en la consecución de los objetivos del investigador social lo que le permite ganar conciencia en el uso de estas bases empíricas, superando los límites de validez intrínsecos a estos datos que se producen para objetivos distintos de ‘actividad de investigación. Sociólogos deberían desarrollar más investigaciones sobre l análisis de los ensambles de datos (Aragona y Felaco, 2019), abriendo las cajas negras (*black box*) (Pasquale, 2015) en las que se producen los big data.

Otro aspecto a tener en cuenta es la postdemografía. El hecho de que los conjuntos de datos digitales usualmente consistan en conjuntos de «eventos» determina un cambio de perspectiva: del individualismo metodológico, dominante en la investigación social clásica, a un enfoque holístico y postdemográfico, orientado al análisis de datos en forma agregada. Incluso cuando los individuos usuarios son los casos estudiados a gran escala, la dificultad (práctica como la ética) de rastrear sus rasgos sociodemográficos hace que la respuesta a las preguntas de investigación canónica en las ciencias sociales sea problemática, por ejemplo, en relación con la movilidad y los flujos electorales.

Sin embargo, uno no puede pensar en los grandes datos solo como datos malos, de hecho. Existen estrategias de investigación en las que estos datos pueden ser realmente útiles para la investigación social (Salganik, 2018), y estas estrategias, incluso si no son mutuamente excluyentes o integrales, se refieren a diferentes tipos de big data. El primero es el conteo de algunos fenómenos sociales. Si el conteo puede parecer una simple pregunta de investigación, en realidad la posibilidad de contar con precisión ciertos fenómenos puede ser extremadamente interesante para la sociología. Botta, et al. (2015) han utilizado con éxito los datos transaccionales de las actividades de los teléfonos móviles para predecir los espectadores de un partido de fútbol en el estadio de San Siro. Los resultados coincidieron perfectamente con el número de espectadores que fue calculado por el personal después de contar todos los accesos al estadio. Además, el análisis de estos datos es mucho más rápido que el de los datos oficiales. Los mismos objetivos se pueden perseguir con los datos de internet de las cosas. Los sensores conectados a la red pueden detectar entradas y salidas de lugares públicos y privados, presencias en museos y otros lugares de interés.

Otra pregunta de investigación que se puede cumplir con big data es la predicción del comportamiento a través de las pistas en Internet. Un ejemplo es la *search as research* (la búsqueda como investigación), es decir, el uso de motores de búsqueda para hacer investigación. Un modelo es el análisis que ha desarrollado Google para investigar a través de búsquedas realizadas en su motor de búsqueda. Mediante el uso de estos análisis, como Google Trends, Google llevó a cabo un famoso estudio sobre la capacidad de predecir la propagación de la influenza ante los Centros para el Control de Enfermedades (CDC) de EE. UU., que recopila de forma regular y sistemática datos de médicos cuidadosamente muestreados en los EE. UU. Ese trabajo (Ginsberg et al. 2009), a pesar de ser complicado en algunos contextos y eventos específicos (Goel et al., 2010), se utilizó para afirmar que las técnicas nativas digitales permitieron formas de análisis que antes no se podían realizar (Mayer-Schonberger y Cuckier, 2013, Rogers, 2013).

En realidad, más que ofrecer una nueva forma de hacer análisis sociológico, el *big data* ofrece una nueva forma de integrar encuestas con otras técnicas de investigación social, que se pueden adoptar sucesivamente o al mismo tiempo crear diseños mixtos. La mezcla puede ser con respecto a los métodos de recolección de datos o para la inclusión de un estudio piloto preliminar, o incluso para realizar encuestas multinivel que mantengan juntas unidades de análisis

individuales, contextuales y relacionales. Tomando por ejemplo las formas de integración entre métodos cuantitativos y cualitativos propuestas por Creswell y Plano Clark (2007), adoptando un enfoque pragmático para la integración entre la investigación por encuestas y el Big Data. Existen sobre todo tres formas de las posibles combinaciones de las dos:

a) Lo que se denomina integración exploratoria, es decir, en la que la investigación de *big data* tiene como objetivo afinar la encuesta;
b) Integración complementaria, en la que los dos enfoques se integran en la fase de recopilación conjunta de datos;
c) Integración interpretativa, que ve el uso de big data para profundizar y validar los resultados de la encuesta.

En las formas de integración exploratoria, el enfoque de big data precede al enfoque de encuesta; en la integración complementaria los dos están al mismo nivel, mientras que en la interpretativa, el Big Data soporta la interpretación y validación de los resultados de una encuesta.

Finalmente, los grandes datos se pueden usar para llevar a cabo experimentos, o más bien cuasi-experimentos. El control de las hipótesis causales siempre ha sido la base de importantes investigaciones sociales y cuestiones de política, pero la construcción de grupos experimentales es complicada y a menudo es difícil identificar grupos grandes. Las plataformas web, por otro lado, ofrecen la posibilidad de llevar a cabo cuasi-experimentos en los que el grupo experimental y el grupo de control se crean espontáneamente, accediendo a un número muy grande de usuarios. Por ejemplo, dos perfiles profesionales idénticos en todos los aspectos, excepto la variable experimental (por ejemplo, etnia o género) se pueden cargar en la misma red social profesional para detectar formas de discriminación en la solicitud de empleo.

Con el objetivo de evitar posiciones ideológicas sobre el papel de los grandes datos en la investigación social, deberíamos promover una participación activa de los sociólogos en la prueba de las diferentes capacidades de los grandes datos. Para hacerlo, no debemos centrarnos en la teoría social abstracta ni en la ciencia social computacional cuantitativa. La mejor manera de hacerlo es mediante la construcción de diseños de investigación que hagan un uso efectivo de las diferentes fuentes de *big data*.

3. LÍMITES Y PROBLEMAS ÉTICOS

A diferencia de los datos de transacciones y los datos de Internet de las cosas, en el uso de datos en la web hay dos problemas típicos de la metodología de investigación social que deben tenerse en cuenta: el efecto Hawthorne y el efecto de deseabilidad social. El efecto Hawthorne (Mayo, 1949) se refiere al hecho de que los individuos cambian su comportamiento cuando saben que están siendo observados. Si en el pasado los usuarios consideraban que sus actividades en

línea estaban privadas o al menos compartidas por otros usuarios, han surgido numerosos casos de análisis ilegal del comportamiento en línea. Ya existen ejemplos en este sentido, como las actividades de vigilancia masiva que realiza la Agencia de Seguridad Nacional de los Estados Unidos (NSA) con los servicios de inteligencia de otros países, tanto hacia los ciudadanos e instituciones estadounidenses como hacia los extranjeros. La NSA ha recopilado metadatos sobre las llamadas telefónicas realizadas a través de todos los operadores de los Estados Unidos y una división especial de la agencia Follow the Money recopila datos sobre las transacciones financieras de las principales instituciones internacionales como Visa, Mastercard y SWIFT. A través del programa de vigilancia PRISM, la NSA tiene acceso directo a los servidores de muchas de las principales empresas de TI de EE.UU. como Microsoft, Google, Yahoo!, Facebook, Apple, YouTube y Skype. El organismo supervisa entonces las actividades de los usuarios, incluidos los intercambios de mensajes, fotografías y vídeos, y en particular maneja listas de direcciones de usuarios utilizadas en los servicios de correo electrónico y de mensajería instantánea. El tratamiento de esta enorme cantidad de datos está justificado por razones de seguridad y relacionado con la lucha contra el terrorismo y la protección del Estado.

Incluso los escándalos más recientes como Cambridge Analytica relacionados con las redes sociales y la privacidad han aumentado la conciencia de las personas de que su comportamiento en línea es observado y registrado. El caso de Cambridge Analytica llamó la atención del público gracias al trabajo de investigación de la periodista de *The Guardian*, Carole Cadwalldr. La encuesta de Cadwalldr reveló que «las capacidades aparentes de análisis de datos y orientación psicográfica de los votantes de Cambridge Analytica, basadas en particular en los datos de redes sociales obtenidos por Facebook con la ayuda de una aplicación de» prueba de personalidad «que recopila información de hasta 87 millones de usuarios» (Bruns , 2019: 1547).

Estrechamente vinculado a aumentar la conciencia de los usuarios de la red de ser observados está la deseabilidad social de sus comportamientos. La deseabilidad social es la tendencia de algunos individuos a dar una respuesta de la manera que encuentran socialmente más aceptable. Lo hacen para dar una imagen positiva de sí mismos y para evitar recibir valoraciones negativas de los demás. Las redes sociales están particularmente influenciadas por el prejuicio de la deseabilidad social porque las personas administran su presencia en línea para proyectar una imagen positiva de sí mismas. Esto lleva a lo que los expertos denominan sesgo de positividad en el contenido de las redes sociales (Veltri, 2021).

Otro aspecto metodológico a tener en cuenta es que los datos digitales en la web deben considerarse no intrusivos, en el sentido de que los sujetos a los que pertenecen los datos no saben que serán utilizados con fines de investigación o al menos no lo saben. saber quién los utilizará y con qué objetivos. Las técnicas de recopilación de datos no intrusivas han aumentado en comparación con el pasado predigital. Los investigadores a menudo pueden recopilar información de páginas web sin que sus propietarios realicen ninguna acción, especialmente

cuando interactúan con plataformas de redes sociales para acceder a sus datos, las denominadas API (interfaces de programas de aplicación), que establecen protocolos para consultar una plataforma y sus datos. Un importante paso adelante en la investigación de API se produjo a principios de la década de 2010, con el uso generalizado de plataformas de redes sociales (como Facebook, Twitter e Instagram). Los verdaderos cambios en el juego aquí fueron las API públicas lanzadas por las principales plataformas de redes sociales en el mercado, que dieron a los investigadores acceso a un rico conjunto de datos digitales, tanto cuantitativa como cualitativamente (Russell, 2013). Un ejemplo simple de esto es la API de Instagram, que antes de 2015 permitía a los desarrolladores obtener datos de hashtags y / o perfiles de usuario sin límites en términos de cantidad y tiempo, junto con un rico conjunto de metadatos, como: ID de publicaciones, comentario. recuento, como recuento, posición, enlace de publicación, hashtag, título de menciones, tipo, imagen de URL, autor de ID de usuario, autor de nombre de usuario, fecha de publicación). El acceso a las API de redes sociales ha iniciado una pequeña revolución en el campo de los métodos digitales, ya que las plataformas de redes sociales han permitido a los investigadores explorar no solo las estructuras socio-técnicas que dan forma a la comunicación en línea (por ejemplo, la lógica de Google PageRank), sino también la cultura procesos que surgen de las prácticas digitales diarias de los usuarios. No es una coincidencia que en la última década los estudiosos de los métodos digitales hayan producido una notable línea de investigación que ha arrojado luz sobre los fenómenos socioculturales más cruciales y convincentes que caracterizan la era digital contemporánea, como las cámaras de eco (Colleoni, Rozza, Arvidsson, 2014), bots políticos (Bessi, Ferrara, 2016), cultura algorítmica (Airoldi, Beraldo, Gandini, 2016), fake news (Gray, Bounegru, Venturini, 2020).

El acceso a los datos de las redes sociales a través de API se ha reducido progresivamente. Las cosas han comenzado a cambiar desde 2018, cuando se produjo el infame escándalo de Cambridge Analytica. En respuesta al escándalo y con el fin de proteger mejor la privacidad de sus usuarios, Facebook (junto con otras plataformas) ha iniciado una política de cierre y restricción de sus API previamente abiertas. Axel Bruns (2019) sostiene que el escándalo de Cambridge Analytica ha sido un pretexto conveniente para que empresas de redes sociales como Facebook y Twitter hagan que sus datos sean progresivamente inaccesibles. Un movimiento que solo aumenta el valor comercial de esos datos (dado que el modelo de negocio de las plataformas de redes sociales consiste precisamente en vender los datos de los usuarios a terceros (Srnicek, 2017), en lugar de aumentar la privacidad del usuario. Las API siguen siendo accesibles, por una tarifa, para las empresas privadas, que las utilizan con fines comerciales y de marketing. Por lo tanto, no es una coincidencia que los académicos hayan estado entre los más afectados por la reducción de las API de redes sociales donde el acceso a las plataformas y sus datos se está volviendo cada vez más difícil. Especialmente después de la reducción de las redes sociales provocada por el escándalo de Cambridge Analytica, se ha vuelto cada vez más difícil investigar las redes sociales. Esta condición trae nuevos desafíos metodológicos y éticos

que requieren repensar los métodos digitales para un entorno de investigación post-API (Caliandro, 2021).

Un tema estrechamente relacionado con el problema de las API es que el uso indiscriminado de técnicas en línea no intrusivas ha generado preocupaciones sobre la búsqueda encubierta, ya que la búsqueda en línea puede representar un riesgo para la privacidad y la confidencialidad de las personas porque a menudo los métodos impiden que los sujetos sepan que su los comportamientos y las comunicaciones se observan y registran. La difusión de métodos de investigación encubiertos ha desafiado las prácticas de evaluación ética establecidas adoptadas en la mayoría de las instituciones de investigación. La parte principal de la ética de la investigación ha sido el consentimiento informado durante años, o el procedimiento mediante el cual el investigador informa a los participantes sobre la naturaleza y el procesamiento de los datos, asegurándose de que serán entendidos y aceptados antes de comenzar con la recolección de datos. El punto clave del consentimiento informado obviamente no tiene posibilidad de aplicarse a los datos digitales recopilados con métodos no intrusivos. Un punto clave es establecer si las fuentes digitales que se utilizan están diseñadas específicamente para producir y compartir contenidos como sitios web, blogs, grupos de noticias, etc., y redes sociales que, por el contrario, pueden recoger contenido que quiera ser compartido con diferentes grados de publicidad. Por lo tanto, se adoptó como práctica considerar el primer tipo de fuentes como públicas, con el supuesto de que, dado que la investigación académica es sin fines de lucro, era un uso justo de este material. Obviamente, este sistema no se puede aplicar a los datos de las redes sociales. El Reglamento General de Protección de datos (RGPD) europeo exige que el consentimiento informado se extienda también a este tipo de datos, considerando que se debe buscar el consentimiento cada vez que se utilicen algunos datos para fines distintos a aquellos para los que fueron recopilados, una condición muy extendida para los datos de las redes sociales. Aún más complicado pedir el consentimiento informado para bases de datos muy grandes, a menudo longitudinales, aquí la posibilidad de obtener el consentimiento de cada participante es casi imposible. El debate sobre la propiedad y la privacidad de los datos digitales no terminará pronto. Si bien algunos piensan que los usuarios se han vuelto menos conscientes de la privacidad a cambio de servicios gratuitos, existe una cierta paradoja de que existe una brecha entre la idea que tienen los usuarios de las redes sociales sobre cuánto se divulgarán sus datos y cuánto efectivamente se divulgarán.

La no intrusividad de los *big data* no solo concierne a aspectos puramente metodológicos, sino a una discusión muy amplia sobre las implicaciones éticas de la investigación y, a nivel social, sobre las normas sociales y legales de privacidad y propiedad de los datos. Desde el punto de vista metodológico, cuando se trata de investigación social digital, no es posible establecer pautas unívocas. Las pautas éticas digitales no pueden ser universales y estar establecidas de una vez por todas. En cambio, deben estar orientados a ser contextuales, es decir, elaborados y adaptados de acuerdo con las plataformas digitales específicas que se están estudiando, el tipo de datos recopilados, los tipos de dispositivos

utilizadas para recopilar los datos, los objetivos de la investigación, la pregunta de investigación. En un momento en el que el acceso a los datos a través de las API se vuelve cada vez más difícil, los investigadores deben encontrar un equilibrio (muy complicado) entre ser un activista de datos y una preocupación por proteger a los participantes de daños, asegurarse de que tengan un equilibrio entre el beneficio y la carga derivados de su participación en la investigación. En este sentido, probablemente, en el caso de *big data*, el punto no es simplemente salvaguardar la privacidad de las personas de quienes se extraen los datos, sino también tratar de redistribuir a los participantes el valor extraído de sus datos, al menos en parte. En el caso de la investigación social, esto puede traducirse en compartir los resultados de la investigación tanto como sea posible con los participantes.

4. CONCLUSIONES

Los *big data* han afectado a nuestras sociedades y a nuestra disciplina. Después de una década en la que los grandes datos fueron vistos alternativamente como el nuevo oro de las ciencias sociales (Lazer, 2009; Mayer-Schonberger y Cuckier, 2013) o como una nueva y peligrosa forma de cuanfrenia (Boyd y Crawford, 2012), un grupo ahora numeroso de estudiosos (Ruppert, 2015; O'Sullivan, 2017) han identificado la posibilidad de construir un terreno intermedio en el que sea posible tanto asumirlos retos como aprovechar las oportunidades que estos datos presentan.

Los *big data* han contribuido a reiterar de una vez por todas, si aún existía una necesidad, el pluralismo que distingue el método de nuestra disciplina. Muchas técnicas que se han desarrollado en otros contextos disciplinarios han llevado al desarrollo de nuestra disciplina y su método (por ejemplo, solo por nombrar algunas, las técnicas de escala en psicología, la comparación en ciencias políticas, el método biográfico en la historia), Lo mismo pasa con los *big data*. Es a partir de esta convicción que se están desarrollando nuevas propuestas de ciencias sociales intensivas en datos, como la *social data science* (Lauro et al., 2017) e la *symphonic social science* (Halford y Savage, 2017) que vinculan las habilidades informáticas y estadísticas con el dominio del conocimiento (sociológico) en el que se emplean, lo que trae consigo sus teorías y visiones de la realidad. De esta manera, los conocimientos típicos de una determinada ciencia se integrarían y combinarían con las disciplinas técnicas formales necesarias para atravesar la era de los grandes datos. Aunque con perspectivas diferentes, todos estos autores están de acuerdo en que los grandes datos son una gran fuente de innovación para las ciencias sociales, y la sociología en particular. No sólo porque amplían el panorama de las bases empíricas que nuestra disciplina puede utilizar, sino también porque pueden promover la interdisciplinariedad entre los diferentes campos científicos, aumentando las posibilidades de integración de datos y técnicas. Sólo mezclando la teoría sociológica y la computación, los enfoques explicativos y descriptivos, los aspectos técnicos y sociales, los datos

tradicionales y los nuevos datos de manera innovadora, los científicos sociales podrán contribuir a la integración de los grandes datos y sus técnicas de recopilación y análisis con los enfoques tradicionalmente practicados en la investigación social. El éxito dependerá de cómo los científicos sociales puedan abordar los límites de calidad y los problemas éticos de los *big data*.

5. BIBLIOGRAFÍA

AIROLDI M., BERALDO D., GANDINI A. (2016): "Follow the algorithm: An exploratory investigation of music on YouTube", Poetics, 57, pp.1-13.

AMATURO, E., ARAGONA, B. (2012): "La costruzione della documentazione empirica" en Metodologia della ricerca sociale, Torino, Utet, pp.52-77.

AMATURO, E, ARAGONA, B., (2017): "Introduction" en Data Science and Social Research: Epistemology, Methods, Technology and Applications, Heidelberg, Springer, pp.1-8.

ANDERSON. C. (2008): "The End of Theory: The Data Deluge makes the Scientific Method Obsolete", Wired, June 23, disponible en http://www.wired.com/science/discoveries/magazine/16-07/pb_theory [consulta: 9/1/2016].

ARAGONA, B. (2016): "Big Data or data that are getting bigger?", Sociologia e ricerca sociale, 109(3), pp.42-53.

ARAGONA, B., FELACO, C. (2019): "Big data from below: researching data assemblages" Tecnoscienza, 10, pp.51-70.

BESSI A., FERRARA E. (2016): "Social bots distort the 2016 US Presidential election online discussion", First Monday, 21, 11/7, disponible en https://firstmonday.org/article/view/7090/5653 [consulta: 4/8/2018].

BOCCIA ARTIERI, G. (2015): Gli effetti sociali del web. Forme della comunicazione e metodologie della ricerca on-line, Milano, Franco Angeli.

BOTTA, F., MOAT, H. S., PREIS, T. (2015): "Quantifying crowd size with mobile phone and Twitter data", Royal Society open science, *2*(5), pp.150-162.

BOYD, D., CRAWFORD, K. *(*2012*)*: "Critical Questions for Big Data", Information, Communication and Society, XV(5), pp.662-79.

BRUNS A. (2019): "After the 'APIcalypse': social media platforms and their fight against critical

scholarly research", Information, Communication and Society, 22(11), pp.1544–1566.

CALIANDRO, A. (2021): "Repurposing Digital Methods in a Post-API Research Environment: Methodological and Ethical Implications", **Italian Sociological Review**, 11(4S), pp. 225-237.

COLLEONI E, ROZZA A, ARVIDSSON A. (2014): "Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data", Journal of communication, 64(2), pp.317-332.

CRESWELL, J. W., HANSON, W. E., CLARK PLANO, V. L., MORALES, A. (2007): "Qualitative research designs: Selection and implementation", The counseling psychologist, *35*(2), pp.236-264.

EUROSTAT (2020): Annual Quality Report, Luxemburg, Eurostat:

FLORIDI L. (2012): "Big data and their Epistemological challenges", Philosophy and technology, 25(4), pp.435-7;

GINSBERG J., MOHEBBI M. H., PATEL R. S., BRAMMER L., SMOLINSKI M. S., BRILLIANT L. (2009): "Detecting influenza epidemics using search engine query data", Nature, CDLVII, 7232, p.1012.

GOEL S., HOFMAN J. M., LAHAIE S., PENNOCK D. M., WATTS D. J. (2010): "Predicting consumer behavior with Web search", Proceedings of the National academy of sciences, 107(41), pp. 17486-17490.

GRAY, E., JENNINGS, W., FARRALL, S., AND HAY, C. (2015): "Small Big Data: Using multiple data-sets to explore unfolding social and economic change", Big Data & Society, *2*(1).

GRAY J., BOUNEGRU L., VENTURINI T. (2020): 'Fake news' as infrastructural uncanny", New Media & Society, 22(2), pp. 317-341.

HALFORD S., SAVAGE, M. (2017): "Speaking Sociologically with Big Data: symphonic social science and the future of big data analytics", Sociology, 51(6), pp.1132–1148.

HEMPEL, C. G. (1942): "The function of general laws in history", The Journal of Philosophy, *39*(2), pp. 35-48.

KEIM, D., KOHLHAMMER, J., ELLIS, G., MANSMANN, F. (2010): Mastering the information age solving problems with visual analytics, Brussels, Eurographics Association.

KITCHIN, R (2014): The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences. London, Sage.

LAURO, C. (2017): Preface, in Lauro et al (eds) *Data Science and Social Research: Epistemology, Methods, Technology and Applications*, Heidelberg: Springer-Verlag.

LAZER, D. (2009): Life in the Network: the Coming Age of Computational Social Science, Science, CCCXXIII, 5915, 721-3.

MACHIN, A. (1998): *Reducing statistical burdens on business*, vol. 9, London, Office for National Statistics.

MAYER-SCHÖNBERGER V., CUKIER K. (2013), Big Data: A Revolution That Will Transform How We Live, Work, and Think, New York, Houghton Mifflin Harcourt.

MAYO, E. (1949): "Hawthorne and the western electric company", en The social problems of an industrial civilisation, Chicago, UCPress, pp.1-7.

O'SULLIVAN D. (2017): "Big Data: why (oh why?) this computational social science?", disponible en https://escholarship.org/uc/item/0rn5n832. [consulta: 1/2/2018].

PASQUALE, F. (2015): The Black Box Society: The Secret Algorithms that Control Money and Information. Cambridge(MA), Harvard University.

ROGERS R. (2013): Digital methods, Cambridge (MA), MIT press.

RUPPERT, E. (2015), "*Socialising Big Data: From concept to practice"*. CRESC Working Paper Series, 138.

RUSSELL M. (2013): "Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More", Sebastopol, O'Reilly Media.

SALGANIK M.J., (2018): Bit By Bit: Social Research in the Digital Age, London, Princeton.

SAVAGE M, BURROWS R *(2007), "*The coming crisis of empirical sociology*"*, *Sociology,* 41*(5), pp.* 885–899.

SRNICEK N. (2017): *"Platform capitalism"*, Cambridge, Polity Press.

STRUIJS, P., BRAAKSMA, B., DAAS, P.J., (2014): "Official statistics and Big Data", Big Data & Society, 1(1), pp.46-61.

THE ECONOMIST (2010): "The data deluge: Businesses, governments and society are only starting to tap its vast potential", Feb 25th, print edition.

TURLAND, M. (2010): *php*, Milan, Marco Tabini & Associates.
VELTRI G., (2021), La ricerca sociale digitale, Milano, Mondadori.
WEBB, E. J., CAMPBELL, D. T., SCHWARTZ, R. D., SECHREST, L. (1966): Unobtrusive measures: Nonreactive research in the social sciences, Chicago, Rand McNally.