Las ciencias sociales en los intersticios del big data: guía práctica para no ser conservadores ni caer en el fin de la teoría.

The social sciences in the interstices of big data: a practical guide to avoid being conservative and avoid falling into the end of theory.

Autora: Abarzúa Cutroni, Anabella.

Citar: Abarzúa Cutroni, A. (2022) Las ciencias sociales en los intersticios del big data: guía práctica para no ser conservadores ni caer en el fin de la teoría.
Revista *Intersticios* 2, pp. 155-163.

Recibido: octubre 2022
Aceptado: julio 2023

Insterticios en acción.

Resumen:

Cuando comenzamos estudiar la problemática del Big data asociado a las ciencias sociales, nuestra primera reacción como profesoras de metodología fue pensar que estamos ante una nueva forma de empirismo donde la disponibilidad de datos primaría más que nunca por sobre la teoría. Eramos herederas - sin saberlo - de las alarmas que se encendieron cuando la revista Wired en 2008 publica "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete" de Chris Anderson. Luego de unas cuantas lecturas comprendimos las texturas y matices que conllevaba esta problemática y que como cientistas sociales estábamos ante lo que empezaba a perfilarse como un nuevo cambio en las formas de interpretar y representar el mundo (Gindin y Busso 2021), cambio que creemos no necesariamente implica abdicar ante el cúmulo de datos y sí implica sumar nuevos métodos y técnicas a los ya tradicionales. En las ciencias sociales el volumen de los datos más que una virtud en si misma debe ser motivo de reflexión e relación a la calidad, el acceso, los contextos de construcción de los datos y los problemas éticos asociados a la privacidad de las personas que los producen.

Palabras clave: big data; ciencias sociales computacionales; metodología.

Abstract:

When we began studying the problem of Big data associated with the social sciences, our first reaction as methodology tea-

chers was to think that we are facing a new form of empiricism where the availability of data would prevail more than ever over theory. We were heirs - without knowing it - to the alarms that went off when Wired magazine in 2008 published "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete" by Chris Anderson.

After a few readings we understood the textures and nuances that this problem entailed and that as social scientists we were facing what was beginning to emerge as a new change in the ways of interpreting and representing the world (Gindinand Busso 2021), a change that not necessarily implies abdicating before the accumulation of data and it does imply adding new methods and techniques to the already traditional ones. In the social sciences, the volume of data, more than a virtue in itself, should be a reason for reflection in relation to quality, access, the contexts construction of data and the ethical problems associated with the privacy of the people who produce them.

Keywords: *big data; computational social science; methodology*

Introducción.

En nuestra tarea como docentes hemos detectado la creciente curiosidad por el *big data.* Esta especie de "indagaciones pragmáticas" de las y los estudiantes universitarias/os proviene muchas veces de su propia inmersión digital cotidiana. La misma es cada día más profunda y significativa. A través de *smarthphones* – móviles que llevamos en nuestras manos casi como una extensión de nosotros/as mismos/as – y otros dispositivos electrónicos interactuamos cada vez con mayor regularidad y con fines cada vez más diversos mediante aplicaciones o plataformas. Estas interacciones se dan tanto entre humanos como entre seres humanos con dispositivos y sensores de diverso tipo. Así generamos datos precisos sobre nuestra ubicación, nuestras preferencias – como consumidores/as, ciudadanos/as, amantes, amigos/as – y nuestros sentires – que van desde reacciones a notas periodísticas a reflexiones íntimas compartidas con seres queridos/as. En las redes sociales expresamos expectativas, opiniones, representaciones tanto políticas como sociales y sentimientos que hace más de dos décadas hubieran sido considerados íntimos. La navegación en internet es rastreada dando cuenta de nuestras prácticas virtuales. Todo esto queda registrado y almacenado de manera automática, es lo que se denomina huella o rastro digital.

Sin embargo, los perfiles virtuales no funcionan como reflejos de la identidad de un indi-

viduo. Según Fernanda Bruno (2013) el perfil virtual es un conjunto de trazos que no conciernen tanto a un individuo específico como a las relaciones que se establecen entre individuos. Por lo cual el perfil es más bien interpersonal que intrapersonal. Estos datos forman un perfil complejo de nuestros atributos, intereses y preferencias y también incluyen información de nuestro contexto y de las personas que forman parte del mismo (Suárez-Gonzalo 2019).

Ante la irrupción del *big data* en los pasillos de la Facultad de Ciencias Políticas y Sociales de la UNCuyo, nuestra primera reacción como profesoras de metodología fue pensar que estamos ante una nueva forma de empirismo donde la disponibilidad de datos primaría más que nunca por sobre la teoría. Éramos herederas - sin saberlo - de las alarmas que se encendieron cuando la revista Wired en 2008 publica "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete" de Chris Anderson. El autor se preguntaba qué podían las ciencias aprender de Google y afirmaba que ya no era necesario construir modelos para testear hipótesis debido a que con suficientes datos, los números - las matemáticas - hablaban por sí solos, ya no era necesario buscar causalidades ya que con las correlaciones bastaba.

Luego de unas cuantas lecturas comprendimos las texturas y matices que conllevaba esta problemática y que como cientistas sociales estábamos ante lo que empezaba a perfilarse como un nuevo cambio en las formas de interpretar y representar el mundo (Gindin y Busso 2021), cambio que creemos no necesariamente implica abdicar ante el cúmulo de datos y sumar nuevos métodos y técnicos a los ya tradicionales. En la actualidad y ante esta coyuntura, nuestras disciplinas están en un punto de inflexión, ante el cual, como docentes-investigadoras, nos encontramos frente a la creciente necesidad de problematizar este fenómeno y sus repercusiones para la investigación en ciencias sociales sin adoptar por esto una posición conservadora.

Pero ¿qué es el big data? ¿Por qué suscita fascinación entre expertos/as y la opinión pública en general? Hace más de una década, el *big data* está plenamente instalado en el ámbito de los negocios como un recurso indispensable para el gerenciamiento y el aumento de las ganancias de cualquier tipo de empresa o industria, es lo que se conoce comúnmente como *business intelligence* y *data science*. Además, cada vez es más común escuchar demandas hacia los Estados en relación de que deberían invertir en tecnologías de gobierno abierto, plataformas de e-democracia y complejas formas de gobernanza electrónica, que harían a dichos Estados más inteligentes, por la primacía de los datos en la toma de decisiones para el diseño de políticas públicas por ejemplo.

La fascinación radica en que estamos ante un hecho inaudito: la proliferación incesante de datos sobre la humanidad que tienen la posibilidad de "capturar", en mayor o menor medida, el conjunto de las relaciones sociales a gran escala (Scasserra y Sai 2020) y conjugarlas con tecnológicas inéditas, aunque largamente imaginadas, provenientes de los desarrollos de inteligencia artificial. Dichos desarrollos han sido llevados a cabo a partir del entrenamiento de algoritmos en base a esa ingente masa de datos a partir de diversas técnicas de aprendizaje automático (Gualda 2022).

En principio el *big data* puede ser pensado como un dispositivo técnico relativamente novedoso, asociado al almacenamiento y procesamiento algorítmico de una cantidad de datos ingente y diversa. Sin embargo, el *big data*, como parte de una serie de innovaciones tecnológicas, trasciende lo meramente técnico para transformarse en un fenómeno social, político y cultural que tiene como denominador común la datificación de las sociedades contemporáneas.

En la actualidad, las relaciones sociales - al menos una parte de estas - no solo se han virtualizado o digitalizado sino que están mediadas por algoritmos, en los que hemos delegado parte del trabajo de la sociedad y la cultura. Rodriguez (2018) plantea que la algoritmización de la sociedad es posible hoy "no es sólo porque los usamos para cualquier cosa, sino también y sobre todo porque todos ellos se encuentran conectados a través de sistemas que son incesantemente alimentados por nuestros usos, de manera tal de poder procesar los registros de diferentes actividades (una solicitud de amistad, la visión de una serie televisiva, la frecuencia cardíaca de un running en el parque, la búsqueda de un dato cualquiera en internet) en un suelo común que permita luego la "personalización", la asignación de esa masa de datos a un individuo, la definición de un perfil. De eso se tratan los metadatos, que constituyen el alimento de los algoritmos" (p. 22).

La proliferación de datos no tradicionales y la ciencias sociales computacionales.

Usualmente, sobre todo en los manuales para la formación de *managers* y expertos en *marketing*, para caracterizar el *big data* se debate en torno a las "tres V": Volumen, Velocidad y Variedad para luego pasar a las promesas de la elaboración de modelos de negocios cada vez más complejos y, a su vez más herméticos. Estos atributos, que presentan cierto consenso en la literatura mercadotécnica, pueden resultar insuficientes para la reflexión en el campo de las ciencias sociales, ya que para este ámbito se requiere también revisar la calidad de los datos, su representatividad y los dilemas éticos asociados.

Entonces, desde la perspectiva de las ciencias sociales y más allá de las provocaciones de Anderson (2008), la cuestión no radica tanto en la cantidad de datos sino en el contexto y en los modos en que se construyen estos datos (Boyd y Crawford 2012), dado que esto es fundamental para establecer en qué medida estos datos pueden propiciar o no, lo que podríamos denominar una nueva heurística para la construcción de conocimiento científico.

En materia de investigación en ciencias sociales y humanidades, el *big data* se presenta como un gran corpus de información, que, a primera vista, se diferencia sustancialmente -debido al contexto y a cómo se generan esos datos- de los datos construidos tradicionalmente por las ciencias sociales y que permite fundamentalmente aumentar las escalas de observación. A raíz de esto se han desarrollado nuevas técnicas de recolección de datos y se ensayan nuevas formas de interpretación de los mismos. Todo esto implica repensar la relación entre lo micro y lo macro y el desarrollo de métodos mixtos de investigación (Gualda 2022, Parra Saiani y Piovani 2021).

Intersticios, núm. 2, 2023: 155-163 3
e-ISSN: 2796-9045

Es importante destacar que la mayoría de estos datos son extraídos y almacenados por corporaciones privadas (Google, Facebook, Amazon, Apple, Microsoft) que: 1) desarrollaron las plataformas mediante las cuales los/as usuarios/as generan los datos, generalmente sin tener en cuenta los costos que esto implica en relación a su privacidad; y 2) cuentan con las capacidades materiales que requiere el almacenamiento y procesamiento de este volumen de datos. El valor comercial de los mismos y el costo de su producción y almacenamiento, los torna prácticamente inaccesibles para la investigación científica, al menos en el sur global. Contrariamente a lo que se cree, en la mayoría de los casos no son datos públicos, ni de acceso libre (Boyd y Crawford 2012, Parra Saiani 2016).

Esta caracterización nos brinda un elemento fundamental para pensar el *big data* en relación a las ciencias sociales. No se trata de datos construidos a partir de la teorización y problematización de determinado objeto/sujeto de estudio mediante técnicas largamente puestas a prueba intensivamente en los últimos 70 años y validadas tradicionalmente para el abordaje de determinadas problemáticas, como por ejemplo las entrevistas en profundidad y las encuestas. Son datos privatizados en su mayoría, generados sin el sentido explícito de generar conocimiento científico por usuarios/as de internet. Esto acarrea severos problemas éticos vinculados con la intimidad de las personas. Que sea factible acceder a un dato, no implica que sea ético hacerlo (Boyd y Crawford 2012). Estamos ante la "paradoja de la privacidad", en la cual las personas saben que sus datos son utilizados por estas corporaciones pero son incapaces de distinguir qué datos exponen y cuáles no cuando utilizan internet, lo que implica la imposibilidad de resguardar individualmente su privacidad (Suárez-Gonzalo, 2019).

Según Gindin y Busso (2021) las discusiones sobre la pertinencia del big data para la investigación en ciencias sociales, se vuelven a inscribir en la clásica disyuntiva cualitativo-cuantitativo. Esto se debe, según las citadas autoras, no solo a una reactualización de este debate, si no al cuestionamiento de "la manifestación de la fuerza misma de verdad adjudicada a la cantidad de datos" dejando de lado cuestiones fundamentales como "la generación misma de los datos, las relaciones que estos establecen entre sí, el contexto en el que son producidos, entre otras variables" (p. 50). En síntesis, el *big data* como dispositivo técnico despierta gran atracción dado que el volumen de los datos, que asociado a la posibilidad de predecir las preferencias de los/as consumidores/as e inclusive fenómenos político-sociales y el vértigo con el que se producen los parece brindarles una veracidad *ipso facto* (p. 50).

La capacidad de predicción de fenómenos sociales y humanos en las ciencias sociales constituye un debate epistemológico de larga data entre distintas corrientes de pensamiento. A lo largo de dicho debate incluso las perspectivas más próximas al positivismo plantean limitaciones a la misma. El post-positivismo del Círculo de Viena, por ejemplo, señalaba que la predicción sólo era posible en términos de probabilidad. Esto requería de enunciados o hipótesis que pudieran ser comprobadas empíricamente. K. Popper entendía que la predicción era una de las potenciales formas de falsar una hipótesis, procedimiento mediante el cual las teorías se consolidaban a medida que ganaban poder explicativo y de predicción. A diferencia del positi-

vismo, el *big data* - al menos en sus usos como herramienta para negocios o para el diseño de campañas políticas - parece estar frente a un empirismo ciego y un abandono de la teoría, dado que el volumen de datos y la complejidad de los modelos de correlaciones garantizaría, por sí sola, la precisión de las predicciones (Deviani 2018).

Así, según Bruno (2013) los perfiles no buscan identificar a un individuo promedio, desde el punto de vista estadístico, sino que buscan establecer principalmente taxonomías o clasificaciones de grupos de individuos a partir de la manifestación de un factor producto de la correlación de un conjunto de variables. Estas clasificaciones tienen como objetivo anticipar conductas de futuro inmediato. Para dichas anticipaciones no sería necesario establecer causalidades, ya que bastarían las correlaciones establecidas - muchas veces automáticamente - entre un volumen cada vez más cuantioso de datos.

Sin embargo, el volumen, más que una virtud en sí misma, es una característica del *big data* que debería incrementar la vigilancia sobre la representatividad de los datos y la validez de sus conclusiones y predicciones. Esta particularidad impone fuertes restricciones para determinar una muestra representativa porque no pueden establecerse los límites de este corpus de datos que representa el universo del *big data*. La generación constante de datos, es decir, la permanente expansión de este universo, hace muy dificultoso el establecimiento del grado de representatividad de los datos (Gindin y Busso, 2018). A la hora de su interpretación, la generación espontánea, anárquica y amorfa de los mismos por parte de usuarios/as de internet (Sosa Escudero, 2019) presenta algunas dificultades tales como: identificar a dichos usuarios/as - y diferenciales de bots y trolls por ejemplo -, distinguirlos según las regularidades de uso de estas plataformas, el tipo de actividad que desempeñan y diferenciar su condición individual o institucional, entre otras (Gindin y Busso, 2021).

Quienes programan el código introducen de esta manera, a partir de su interpretación y los datos que utilizan para entrenar algoritmos, asimetrías del poder (Gindin y Busso, 2021; Boyd y Crawford, 2012). Es importante identificar entonces los sesgos que introducen esta masa de datos a los algoritmos, que a su vez dan forma a nuevos datos a partir del procesamiento de los mismos. Por un lado, existen sesgos, que podríamos denominar materiales, que son los provenientes del acceso limitado de determinado sector de la población a internet (por problemas de conectividad, de acceso a la tecnología, costos, etc.) y por otro, hay sesgos introducidos durante el procesamiento del *big data* mediante algoritmos que a su vez son entrenados por más datos que implican sesgos. Acá radica, por ejemplo, la problemática de los sesgos de género y raciales entre otros (Failero, 2021; Suárez-Gonzalo 2019; D'Ignazio y Klein, 2020).

De este modo, una de las contribuciones centrales procede de las teorías feministas, en general y de la perspectiva del feminismo de datos (D'Ignazio y Klein, 2020) que se posiciona en una reflexión interseccional sobre los usos y limitaciones de los datos. Estas corrientes y tantas otras muestran que los datos nunca son datos "crudos", sino que los mismos están situados cultural, social e institucionalmente.

Intersticios, núm. 2, 2023: 155-163 3
e-ISSN: 2796-9045

Entonces, reflexionar sobre la validez y representatividad de estos datos es algo ineludible, hoy más que nunca, para los estudios de ciencias sociales que pretendan construir conocimiento científico a partir de este tipo de datos, o elaborar una crítica fundada hacia este fenómeno y sus consecuencias políticas y sociales.

En síntesis, como cientistas sociales, primero debemos interrogarnos acerca del acceso restringido/privatizado en relación a esa masa de datos, de las controversias éticas y políticas que implica para la privacidad de las personas y los dispositivos de control y vigilancia social que pueden construirse a partir de estos datos (Bruno 2013, Rodriguez 2018). Para esto debemos comprender las lógicas que subyacen al fenómeno del *big data* y conocer cómo funcionan las tecnologías asociadas al mismo.

Segundo, debemos interrogarnos acerca de si a partir de estos datos - que por sus características, distan técnica y epistemológicamente, de los datos construidos tradicionalmente para las ciencias sociales - es plausible y cabal generar conocimientos relevantes sobre nuestras sociedades contemporáneas, y por tanto qué oportunidades y limitaciones se presentan para la implementación novedosa de metodologías tanto cuantitativas como cualitativas.

A grandes rasgos cuando nos referimos a potencialidades pensamos en la posibilidad de estudiar nuevos fenómenos y construir otra clase de objetos de investigación, con nuevas escalas, por ejemplo; y cuando hablamos de limitaciones nos referimos a las dificultades y complejidades que implica la construcción de datos mediante el dispositivo del *big data* y la necesaria reflexividad para la construcción de conocimiento científico. En esta línea, coincidimos con la aseveración de que "los datos no existen fuera de las ideas, de los instrumentos, las prácticas y el contexto que enmarcan su creación e interpretación" (Rocha Meneses 2018, p. 424).

Guía práctica a modo de cierre.

En los últimos años en torno al *big data* o a partir de reflexiones sobre los entornos digitales se han generado nuevos campos de conocimiento caracterizados por su interdisciplinaridad. Campos dinámicos, en plena construcción, las humanidades digitales, las ciencias sociales computacionales y la ciencia de datos son ejemplos de esto. En este sentido, una de las tareas y desafíos de las ciencias sociales no solo es la articulación o diálogo con las ciencias computacionales, sino la necesidad de adquirir conocimientos novedosos, habilidades y lenguajes específicos adecuados para este nuevo contexto.

Para esto nos parece fundamental y recomendamos:

1) Escuchar a nuestras/os estudiantes. Prestar atención a sus preguntas e inquietudes ya que "en sus nervios hay mucha más información de futuro..." (Solari, 1997) que la que tenemos hoy las personas que tenemos la responsabilidad de abrirles camino en el campo académico y profesional.

2) No dejar de estudiar estadística. Comprender en profundidad los modelos -, sus

correlaciones y clasificaciones - con los que operan los algoritmos son nuestra principal responsabilidad como cientistas sociales. La clave para abrir las cajas negras algorítmicas mediante la que se procesan cantidades cada vez más grandes de *inputs* para generación de cada vez más complejos *outputs* es identificar y revelar sesgos y esto solo es posible desandando la construcción estadística que se hace de las taxonomías.

3) Estudiar programación no con la idea de que nuestros saberes han caducado si no con la idea de entablar un profundo diálogo con programadores/as. Las ciencias sociales tenemos mucho que aportar. Un acervo centenario de problemáticas, teorías y análisis empíricos y sobre todo nuestra capacidad de interpretación es invaluable para la construcción colectiva del futuro que anhelamos como seres humanos.

Seamos creativos, no nos conformemos con presionar "botones", casi todo lo demás, lo hacen las máquinas.

Bibliografía

Anderson, C. (2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired.* 23 de Junio. https://www.wired.com/2008/06/pb-theory/

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

Bruno, F. (2013). *Máquinas de ver, modos de ser: vigilância, tecnologia e subjetividade.* Sulina.

D'Ignazio, C y Klein, L. (2020). The Numbers Don't Speak for Themselves. *Data Feminism*. Cambridge MA, MIT Press.

Deviani, R. (2018). Consideraciones epistemológicas, teóricas y críticas en relación al big data. En Biselli, R. y Maestri, M. (Eds.) *La mediatización contemporánea y el desafío del big data.* Pp. 11-34. UNR Editora.

Faliero, J. C.(2020). Limitar la dependencia algorítmica. Impactos de la inteligencia artificial y sesgos algorítmicos. *Revista Nueva Sociedad (NUSO*). N° 294, julio- agosto de 2021, pp. 120-129. ISSN: 0251-3552.

Gindin, I.L. y Busso, M.P. (2021) El Big Data bajo la lupa: notas sobre el retrato de una época. En Actis, E.; Berdondini, M; Castro Rojas S.R. (Comps.) *Ciencias Sociales y Big Data*. Pp. 49-64. UNR Editora.

Gindin, I.L. y Busso, M. (2018). Investigaciones en comunicación en tiempos de Big Data:

sobre metodologías de análisis y temporalidades en el abordaje de redes sociales. *Revista adComunica*, nº15, pp. 25-43.

Gualda, E. (2022). Social big data, sociología y ciencias sociales computacionales. *Empiria. Revista de Metodología de Ciencias Sociales,* (53), 147–177.

Meneses Rocha, M.E. (2018). Grandes datos, grandes desafíos para las ciencias sociales. *Revista mexicana de sociología.* Vol. 80, n. 2, pp. 415-444.

Parra Saiani, P. y Piovani, J. I.(2021). Triangulación metodológica y big data. *PRACS: Revista Eletrônica de Humanidades do Curso de Ciências Sociais da UNIFAP* . v. 14, n. 2, p. 157-167, maio/jun. 2021. ISSN 1984-4352.

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, pp. 221- 240.

Rodriguez, P. (2018). Gubernamentalidad algorítmica. Sobre las formas de subjetivación en la sociedad de los metadatos. *Revista Barda.* Año 4 - Nro. 6 - Junio 2018. Pp. 14-35.

Scasserra, S. y Sai, L. (2020) *La cuestión de los datos. Plusvalía de vida, bienes comunes y Estados inteligentes.* Friedrich Ebert Stiftung Argentina.

Sosa Escudero, W. (2019). *Big Data. Breve manual para conocer la ciencia de los datos que ya invadió nuestras vidas.* Siglo XXI.

Solari, C. (1997). *Histórica conferencia de prensa de los Redondos en Olavarría* (16-08-1997). Redondos subtitulados. https://youtu.be/xWfHUiqeCEU. Consultada el 28/09/2022.

Suárez-Gonzalo, S. (2019). Personal data are political. A feminist view on privacy and big data. *Recerca. Revista de Pensament i Anàlisi,* 24 (2), pp. 173-192.