BORDES, MAYO-JULIO DE 2019. REVISTA DE POLÍTICA, DERECHO Y SOCIEDAD
ISSN 2524-9290 http://revistabordes.com.ar

Big Data y ciencias sociales

JULIÁN TAGNIN (UNLZ/UNPAZ)
21 DE JUNIO DE 2019

La recopilación, selección y análisis de grandes cantidades de datos (Big Data) despierta notables cuestiones epistemológicas, lógicas y metodológicas para las ciencias. Específicamente para las ciencias sociales se abre un nuevo paradigma de investigación y se reviven históricos debates sobre los criterios para fundamentar los dominios de objetos legítimos en el discurso científico.

El Big Data ha reavivado las posiciones del empirismo y el realismo ingenuos, desde el mainstream de promoción de la ciencia y la tecnología se ha vuelto a hablar de objetos sin teoría y de la obsolescencia de los métodos en ciencias sociales. La nueva ola de escepticismo teórico, que asume la naturalización de los discursos montados sobre datos extraídos a partir de Big Data, ignora que desde el empirismo científico mismo se

ha abandonado la idea de que pueda existir una realidad desideologizada. No hay que confundir la realidad ni lo social con los datos que tengamos, que necesariamente son filtrados por criterios constituyentes de una selección con implicancias éticas y políticas.

Pero más interesante que detenernos en esta crítica, es pensar hasta dónde nos permiten estas nuevas técnicas ver fenómenos sociales imposibles de estudiar con otros métodos. En parte, de modo intuitivo, creo que las ciencias sociales deberían recibir ciertas técnicas con la misma expectativa que pudo traer, por ejemplo, la invención del telescopio en astronomía. Con la minería de datos, los aprendizajes automatizados, las simulaciones, los análisis de clústeres y demás procesos asociados al Big Data nos encontramos ante una multiplicación magnífica de las entidades observables en el mundo social, a escalas nunca antes vistas.

Predictibilidad o explicación en las ciencias sociales

El modelo del positivismo lógico y el racionalismo crítico, ceñidos al empirismo, supone poder encontrar leyes universales del comportamiento social. La ausencia de un fundamento último que sostenga este modelo, que en filosofía de la ciencia es conocido como la concepción estándar, dio entonces paso a la teoría hermenéutica o interpretativista, desde donde se argumentaba que el interés de las ciencias sociales está en comprender más que en explicar o predecir. El Big Data entra en esta discusión con la implementación de aplicaciones, modelos y simulaciones computarizadas que pueden predecir distintas trayectorias futuras para una mejor toma de decisión en el presente. Los trabajos conocidos como *future studies* sostienen que los dominios referidos por la explicación son diferentes a los de la predicción, y que la relación entre la corrección de una explicación y el éxito de una predicción no es lógico, orgánico ni necesario. Por medio de la correlación y la inferencia estadística pueden predecirse cosas que quizá no se puedan explicar. Sin embargo, estos modelos y simulaciones son importantes para las ciencias sociales porque el modo en que la gente piensa sobre su futuro impacta en las decisiones que toma. Por ejemplo, puede decirse que quienes pueden ver más lejos en el futuro están menos comprometidos en tomas de decisiones riesgosas.[1]

1 Thorstada, R. y Wolffa, P. (2018). A big data analysis of the relationship between future thinking and decision-making. *Proceedings of the National Academy of Sciences, 115*(8) E1740-E1748; DOI: 10.1073/pnas.1706589115.

Metodología de la investigación en Big Data

Montados sobre el valor de la interpretación de la información en cualquier investigación, podemos señalar la síntesis entre métodos cualitativos y cuantitativos que posibilita el Big Data, más que contraponer, como se hizo históricamente, los métodos de cada abordaje.[2] Hoy es posible unir inteligentemente el dato de superficie de grandes cantidades de entidades que ya existía en el siglo XX (en la estadística poblacional, por ejemplo), con el dato profundo de pequeños grupos o pocos entes, como en la descripción densa propuesta en antropología por Clifford Geertz.

Existen actualmente esfuerzos para aunar los análisis de datos de las ciencias computacionales con métodos socio-semióticos, siempre asociados al recorte de pequeños universos. De hecho, hay quienes auguran que esta síntesis habilitará a los sociólogos culturalistas a alcanzar progresos teóricos en áreas hasta ahora pensadas como inmensurables. A su vez, los analistas computacionales y estadísticos podrían aprender a perfeccionar sus herramientas para mapear los contornos de los campos culturales y semióticos, clasificar sus elementos y trazar la evolución de estas entidades diacrónicamente.[3]

El problema para los científicos sociales reside muchas veces en la complejidad de algunos de esos métodos, que los obliga a conocer y saber usar técnicas y programas no sólo ajenos a la mayoría de sus planes de estudio, sino también en constante renovación. También, por supuesto, existen los grupos inter o transdisciplinarios de trabajo, que aúnan esfuerzos de distintos investigadores y profesionales.

Inferencias por correlación

Si bien siempre hubo investigaciones y producción de hipótesis sobre estadísticas y grandes datos, nunca tuvimos la velocidad y variabilidades para el análisis computacional que tenemos hoy y, más importante, nunca hubo información tan detallada sobre las relaciones, discursos, localización espacial, consumos, problemáticas sanitarias y demás

2 Manovich, L. (2011). Trending: The Promises and the Challenges of Big Social Data. En M. K. Gold (ed.), *Debates in the Digital Humanities* (pp. 460-476). Minneapolis: The University of Minnesota Press.

3 Bail, C. A. (2014). The cultural environment: Measuring culture with big data. *Theory and Society, 43*(3-4), 465-482.

dimensiones para el análisis social. En los análisis de Big Data existe en la actualidad un conjunto de técnicas englobadas en los procesos de la minería de datos (*Machine learning, Cluster analysis, Association rule learning*, clasificación y regresión, entre otros), que permiten descubrir relaciones entre variables, valiéndose de inferencias bayesianas y de la capacidad de cómputo de los procesadores. De esta manera, se habilitan distintos métodos para alcanzar índices de correlación (como el de Pearson, Spearman o Kendall) que dan lugar a la formulación o validación de hipótesis.

Pero cualquiera sea el método y el proceso algorítmico de búsqueda y procesamiento de la información, sin dudas estamos ante la habilitación de un método históricamente cuestionado en las ciencias sociales porque, naturalmente y como se repite de memoria, la correlación no implica causación. En este sentido, es necesario ser cautelosos en el uso de estas inferencias, pero no se puede desconocer que el trabajo con Big Data podría permitir, propongo a modo de hipótesis, descubrir entidades relevantes para las ciencias sociales que son invisibles para la deducción o la inducción experimental, simplemente porque pueden existir entidades sociales que escapen a esos tipos de razonamiento.

Modelización y selección de corpus

Los desafíos epistémico-políticos tienen su correlato en el abordaje del material de estudio como problemas metodológicos. Las investigaciones en ciencias sociales deben ajustarse al formato de las plataformas, pero además deben seleccionar qué recorte harán de la cuantiosa masa de los datos. En este sentido, es válida la pregunta de los investigadores Gastón Cingolani y Mariano Fernández: "¿con qué criterio construir representatividad si no sabemos dónde toca fondo el universo?[4] y ¿qué técnicas de recolección emplear de acuerdo a las particularidades y a las exigencias de cada interfaz?".[5]

4 Cingolani, G. y Fernández, M. (2018). Objeto, objetos, corpus, sistemas. Lo que se deshace y lo que hacemos cuando analizamos discursividad en la crisis de mediatizaciones contemporáneas En R. Biselli y M. Maestri, *La mediatización contemporánea y el desafío del big data*. Rosario: UNR Editora.

5 Gindin, I. L. y Busso, M. P. (2018). Investigaciones en comunicación en tiempos de big data: sobre metodologías y temporalidades en el abordaje de redes sociales. *adComunica. Revista Científica de Estrategias, Tendencias e Innovación en Comunicación*, 15.

El alcance del corpus tendrá que ver entonces también con la disponibilidad de datos, y la modelización con las categorías estructurantes de los data sets, más las categorías propias del investigador. Las discusiones que están llevando y tienen por delante lxs investigadores hablan de innovaciones técnicas pero también de criterios políticos en el cómo y por qué modelizar y seleccionar un corpus. Aquí sugiero que debería primar el sentido social de las investigaciones, es decir que deberíamos preguntarnos para qué buscamos los datos, qué ciencia queremos y en qué áreas queremos tener mejores políticas públicas o más discusión cívica sobre los problemas más acuciantes de la sociedad. Esto sin desatender los avances en investigación básica que permitan descubrir cómo trabajar mejor con las bases existentes y qué nuevo tipo de entidades pueden ser observables e incidir en el análisis social. Aquí también surgen inconvenientes para la investigación relacionados con la privacidad y la dificultad en el acceso a esas bases.

Transparencia y alcance de los datos

Las investigaciones estadísticas en ciencias sociales han sido por lo general muy costosas y muchas veces también imprecisas. Especialmente cuando hablamos de gustos, consumos, preferencias políticas, etc., se convive con cierto margen de discrecionalidad y con la reticencia de las poblaciones en estudio a entregar ciertos datos. Hoy es igualmente costoso para investigadores que no cuenten con los permisos de las empresas que sistematizan los data sets, pero para estas empresas o para los investigadores que pueden comprar esos datos, es muy sencillo medir y comparar tendencias, si tienen las preguntas bien orientadas. En este contexto, se elimina el factor de opacidad de los sujetos, que es reemplazado por la opacidad de los algoritmos que condicionan la accesibilidad de los datos.

Los Estados que promueven el gobierno abierto asumen políticas de transparencia que permiten una mejor accesibilidad en las investigaciones, pero aún no hay reservorios públicos significativos de bases, ni políticas de acceso abierto en las mayores plataformas de socialidad digital.

Otros asuntos relevantes a considerar son las cuestiones de la privacidad y el alcance de los datos. La legislación sobre privacidad aún no se ajusta a la socialidad digital. Para

los Estados, por ejemplo, es muy difícil configurar digitalmente la accesibilidad restringida, y en la práctica termina pasando que por cuestiones operativas muchos terminan abriendo universalmente a su personal el acceso a los expedientes digitales. También los problemas de memoria y archivo son un desafío para las investigaciones de Big Data, pero todo tiende a sugerir que mientras los estudios mantengan el anonimato en las entradas de datos no deberíamos tener problemas con la violación de datos personales.

En limpio

Después de este breve recorrido llegamos a algunas consideraciones generales sobre los cambios que aporta el Big Data a las ciencias sociales. En principio podemos decir que si bien notamos cierto sesgo de un empirismo ingenuo asociado con el uso de estas técnicas y herramientas (lo cual significa en cierto modo un retroceso dado que se desconocen ciertos argumentos de la historia de la ciencia que habían superado esa posición), podemos valorar los avances que pueden darse gracias al arsenal renovado para estudiar lo social. Entiendo que estamos ante la posibilidad de que todo un espectro de nuevas emergencias significativas para el entendimiento social surjan como objetos de estudio. También vemos, en la misma línea, remozadas las inferencias por correlación. Será interesante ver además hasta dónde la capacidad de procesamiento de datos, la capacidad de predicción, el auto e-learning y la simulación computarizadas podrán aportarnos al entendimiento de nuestro comportamiento y al diseño de políticas públicas.

Sin dudas, la carga axiológica y epistémico-política de las investigaciones seguirá siendo un factor considerable en cualquier representación llevada adelante a partir de métodos de Big Data, pues la selección de los corpus es siempre relativa a una perspectiva y posición específica, pero es indudable también que el Big Data podría permitir avanzar mucho más dentro del desarrollo normal de cada paradigma multidisciplinar de las ciencias sociales.