

# Hacia una exploración de las representaciones sociales en torno al big data

Gastón Becerra <sup>1,2</sup>, Juan Pablo López-Alurralde <sup>1</sup>

<sup>1</sup> Universidad Abierta Interamericana – Centro de Altos Estudios en Tecnología Informática, Ciudad Autónoma de Buenos Aires, Argentina

<sup>2</sup> Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET, Argentina

`gaston.becerra@sociales.uba.ar`

**Resumen.** Compartimos los primeros resultados de una investigación en curso cuyo objetivo es explorar las representaciones sociales del big data. Para ello se tomó un test de evocación a una muestra de 247 estudiantes y graduados universitarios argentinos de distintas carreras. Entre las preguntas que exploramos se encuentran: ¿qué ideas se evocan en la representación social del big data y cómo se estructuran? ¿qué temas se pueden distinguir, por la coocurrencia de evocación en una respuesta, y cuál es la orientación o valoración, en tanto indicador de actitud hacia el big data? ¿cuáles son las fuentes de consulta acerca del big data?

## 1 Introducción

En este trabajo compartimos los primeros resultados de una investigación en curso, cuyo objetivo es explorar las representaciones sociales en torno del big data, junto a otros fenómenos complejos en los que confluyen cuestiones tecnológicas, sociales y epistémicas, tales como la inteligencia artificial, la irrupción de las ciencias de datos, y los fenómenos 2.0-5.0.

Este interés surge como una línea complementaria a una investigación más amplia que busca problematizar el fenómeno del big data desde la perspectiva de los sistemas sociales y el constructivismo de Niklas Luhmann [1, 2]. En dicho enclave teórico habíamos asumido el objetivo de buscar el significado social del big data en las diversas formas en que se lo tematiza desde distintos sistemas, tales como la política, la economía, la ciencia, los medios de comunicación, entre otros.

Aquí consideramos un espacio complementario: el sentido común. Nos orientamos por desarrollos de la psicología social, particularmente, por la teoría de las representaciones sociales [3, 4]. El concepto de representación social designa una forma de conocimiento de carácter práctico, orientado hacia la comunicación, la comprensión y el dominio del entorno social, material e ideal [5]. Las representaciones se construyen activamente ante la necesidad de dar sentido a fenómenos novedosos que nos interpelan diferencialmente en vistas de nuestra posición en el entramado de relacio-

nes sociales [6–8]. En este sentido, la representación social se vuelve particularmente relevante para el estudio de desarrollos científicos que pudieran tener un impacto social, siendo el estudio de Moscovici [9] sobre el psicoanálisis y su recepción en Francia un ejemplar paradigmático.

Aquí presentamos al big data como un fenómeno interesante para ser tratado en términos de representación social. Su relevancia se desprende de que es uno de esos fenómenos tecnológicos, sociales y epistémicos en torno a los cuales se ha desarrollado una “retórica” que augura una profunda transformación de todos los aspectos de la vida social. Por mencionar sólo uno de los espacios alcanzados, podemos referirnos a la economía, donde actores como *The Economist* afirman que los datos constituyen el “nuevo petróleo”<sup>1</sup>, o como cuando *IBM* señala que “hoy todo está hecho de datos” (*IBM*, 2014)<sup>2</sup>. Como señalamos en otro trabajo centrado en el tratamiento del big data por parte de la prensa digital argentina [10], en el centro de este discurso se encuentra una “promesa” de carácter epistémico: que los grandes datos habilitarían formas superiores de inteligencia y conocimiento, que su exploración permitiría arribar a verdades más objetivas o con mayor grado de precisión y previsión, y que la cuantificación de todos los aspectos del comportamiento humano es clave para desarrollar soluciones novedosas [11–13]. Se debe hacer notar que si bien esta promesa se encuentra claramente expresada en los discursos “promocionales” del big data, también hay algo de ella en los discursos más pesimistas que ven en el big data un nuevo “gran hermano”, con un renovado mecanismo de control. En cualquier caso, la promesa no tendría asidero si no estuviera sustentada en una “premisa” que pudiera presumirse evidente y trivial: que vivimos en un mundo donde hay vastas cantidades de datos disponibles, y que los mismos pueden ser técnicamente manipulados para darles sentido. Diferentes estudios sociales críticos del big data han puesto de manifiesto los variados mecanismos retóricos en la comunicación usual de esta premisa y esta promesa, tales como esconder el carácter de producto humano de los datos al equipararlos con recursos naturales que pueden ser “minados” o “colectados”, o esconder las diferencias en las posibilidades reales en torno al acceso y la explotación de los datos, o incluso negar su carácter propietario [14, 15].

Hasta donde conocemos, no hay trabajos empíricos que hayan explorado la representación social del big data. En este trabajo adelantamos algunos resultados preliminares de una investigación en curso con dicho objetivo. Nuestras preguntas son:

**RQ1.** ¿Qué ideas se evocan en la representación social del big data? ¿Cómo se organizan y estructuran?

**RQ2.** ¿Qué temas se pueden distinguir en la representación del big data? ¿Cuál es su orientación o valencia?

**RQ3.** ¿Cuáles son las fuentes de consulta acerca del big data? ¿Cómo lo representan estudiantes de distintas carreras universitarias?

---

<sup>1</sup> <https://www.economist.com/briefing/2017/05/06/data-is-giving-rise-to-a-new-economy> (consultado Junio/2020)

<sup>2</sup> [https://www.youtube.com/watch?v=QCgZrOUd\\_Dc](https://www.youtube.com/watch?v=QCgZrOUd_Dc) (consultado Junio/2020)

## 2 Método

En este primer estudio exploramos la representación social del big data por parte de estudiantes y graduados universitarios de Argentina a través de la técnica de asociación de palabras [16–18].

### 2.1 Instrumentos

La recolección se realizó a través de un formulario online<sup>3</sup>, que incluía un test de evocación con 5 campos (se podía agregar otros) para introducir palabras espontáneas y un control deslizante para indicar una valoración para cada término introducido. El estudio se difundió por redes sociales. La consigna fue la siguiente: *Por favor, indicanos qué palabras o frases te vienen a la mente cuando pensás en "[estímulo]". También te pedimos que, por favor, nos indiques si estas ideas que acabas de introducir se corresponden con algo que valoras positivamente (algo que te agrada) o negativamente (algo que te desagrada), utilizando el deslizador debajo de cada palabra.*

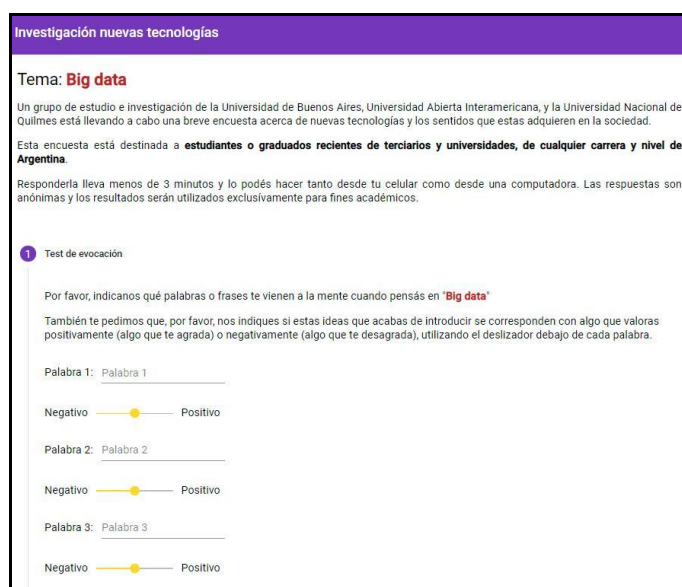


Fig. 1. Captura del instrumento (test de evocación).

Como uno de los objetivos de la investigación es ubicar al big data en un campo de fenómenos de relevancia tecnológica, social y epistemológica más amplia, este formulario fue programado para rotar los términos estímulos en la siguiente colec-

<sup>3</sup> Disponible en <https://evoc-67321.web.app/>

ción (por orden de probabilidad de aparición): *big data*, *inteligencia artificial*, *ciencia de datos*, *conocimiento*. En cada consulta el formulario muestra sólo 1 de estos términos para evitar condicionamientos. En este trabajo, nos centramos sólo en los resultados relativos al *big data*, no haciendo comparaciones entre corpus.

Finalmente, también se incluían preguntas sociodemográficas, de formación (carrera), y referidas a las fuentes consultadas para informarse sobre el tema en cuestión.

## 2.2 Muestra

La muestra fue de tipo intencional, compuesta por estudiantes y graduados universitarios de Argentina ( $N=247$ ), con edades comprendidas entre 17 y 70 años ( $M=28,6$ ;  $SD=7,5$ ); 74,7% de mujeres, 24% de hombres y 1,3% de otros. Más del 55,5% son estudiantes o graduados de ciencias sociales y empresariales (incluyendo psicología, la carrera que más estudiantes aporta en la muestra), el 17,5% de humanidades, el 15% de naturales y exactas, el 6% de ingenierías y tecnologías, y finalmente el 6% restante de médicas y de la salud.

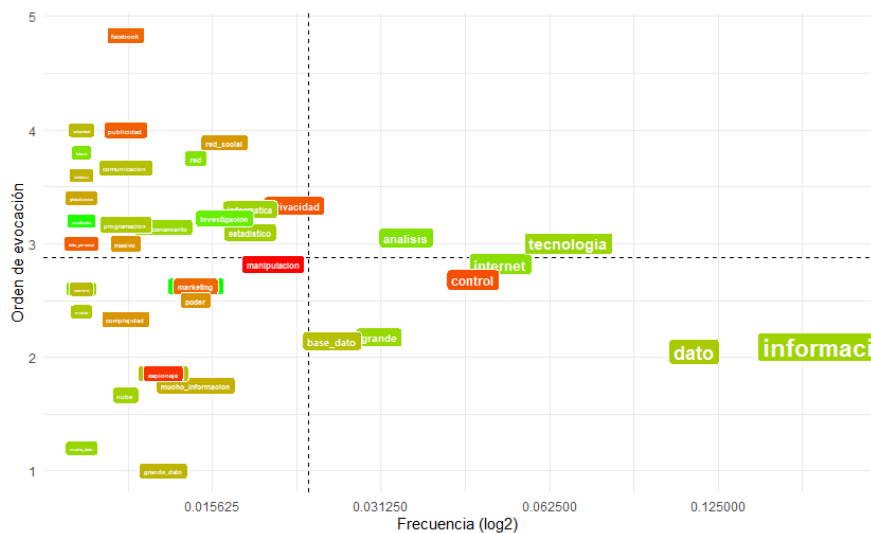
Antes de continuar, advertimos que se trata de una investigación recién iniciada y que el estado de la muestra nos fuerza a que la interpretación de los resultados no busca más que elaborar hipótesis tentativas y conjeturas para futuros análisis.

## 2.3 Preprocesamiento

El pre-procesamiento del corpus de respuestas (todos los estímulos) consistió en varios pasos: 1) se anotaron los términos con UPOS, lo que permitió trabajar con lemmas, remover símbolos, puntos y números, así como también adverbios; 2) se reemplazaron caracteres especiales y acentos; 3) se convirtió el corpus a minúscula. En el caso de que la respuesta fuese una frase o varios términos, se unieron (los que no fueron filtrados en pasos anteriores) por el símbolo “\_”.

Al comenzar el pre-procesamiento se contaba con un vocabulario de 1.655 términos únicos (para todos los estímulos), con cerca de 1.300 términos que aparecían una sola vez, y cuyo término más frecuente tenía 70 repeticiones. Al terminar, el vocabulario se redujo a poco más de 1.200 términos únicos, con cerca de 800 términos que aparecían una sola vez, y cuyo término más frecuente tenía 155 repeticiones. La base de términos específicos del *big data* incluye 590 términos únicos, donde el término más frecuente tiene 110 repeticiones, y hay cerca de 400 términos mencionados una sola vez. Para todos los análisis siguientes se fijó un umbral mínimo de 5 repeticiones.

Todas las tareas de preprocesamiento, análisis, y visualizaciones fueron programas en R [19]. Los datos se estructuraron, en la medida de lo posible, siguiendo principios “tidy”, y fueron manipulados con herramientas del Tidyverse [20]. Para los análisis factoriales se reutilizó código del proyecto DTM-VIC [21].



data a un público masivo por parte de la prensa digital argentina (Becerra, 2019). Sin embargo, también se debe aclarar que algunas de estas ideas están presentes en varias palabras compuestas de menor frecuencia, y que otro pre-procesamiento las podría haber puesto en un lugar más predominante. Todos estos términos están asociados con una valoración positiva, cercana a 7.5/10.

El único término disonante con esta idea es el de *control*, siendo además el único en este segmento con una valoración negativa (2.39/10). Podemos conjeturar que control es una idea central en la promesa del big data: es la que lo ubica en relaciones que vinculan el conocimiento y el poder. Y si bien la idea puede remitir a la capacidad de manipular cualquier objeto, la valoración negativa permite conjeturar que la imagen que informa a la representación es la del sometimiento y el condicionamiento de grupos humanos. Esta conjetura se puede validar con un análisis de correlación de palabras, como se propone para la **RQ2**.

El segundo segmento comprende a las palabras que se encuentran en la periferia del núcleo central, aquí definidas por tener una frecuencia superior a la media, pero también un orden de evocación mayor a la media, indicando que fueron enunciados entre los últimos términos de cada respuesta. Aquí se cuenta con sólo 2 términos: *análisis* y *tecnología*. Si bien son ideas muy generales, podemos conjeturar que refieren al mismo anclaje del big data, que lo vincula con su costado más mundano: el *análisis* refiere al proceso de transformación activa de los datos para darles sentido por parte de aquellos grupos abocados al big data, mientras que la *tecnología* podría referir al campo material y físico que soporta al big data. En ambos casos, se trata también de ámbitos o perfiles laborales distintos en el big data. No casualmente, ambas ideas suelen enfrentarse en la retórica del big data con la idea de *magia*, que encubren esfuerzos y costos en ambos campos [24].

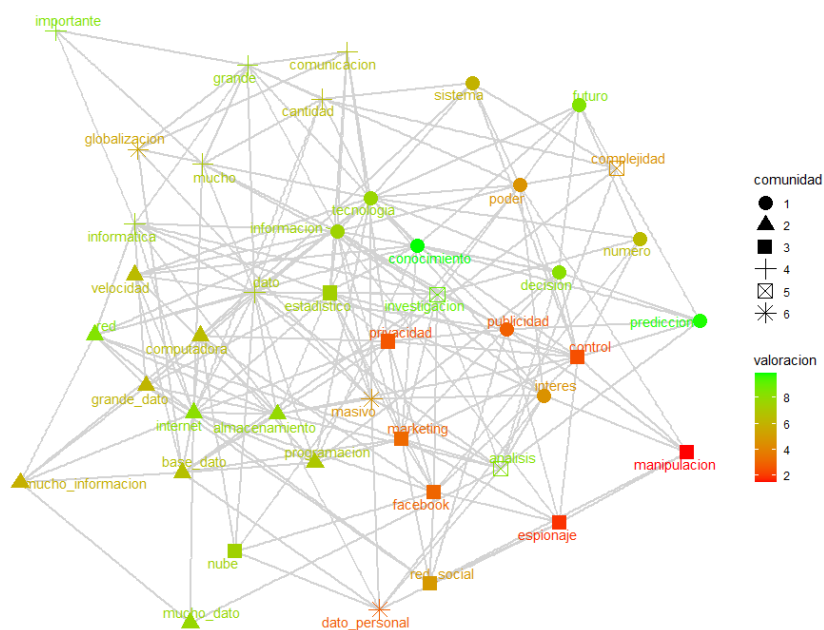
El tercer segmento incluye palabras que se evocaron rápidamente pero que no tienen una frecuencia suficiente como para ser consideradas parte del sentido común. En términos de la teoría, podrían remitir a los sentidos propios de grupos específicos: *conocimiento*, *manipulación*, *marketing*, *computadora*, *mucha información*, *espionaje*, *grandes datos*, o *poder*.

### 3.1 Temas en la representación social del big data

El análisis anterior, tributario del modelo de núcleo y periferia, puede ser útil para dar una idea de la organización jerárquica de la representación, pero no nos permite aclarar de qué manera se relacionan entre sí las evocaciones. Así, uno podría preguntarse si acaso los términos de valoración más negativa en nuestro corpus (por ejemplo, *control*, *manipulación*, *espionaje*) no se dan generalmente juntos o, si por el contrario, se observan distribuidos con otros términos de valoración más positiva. Si algunos términos se dan juntos se podría conjeturar que esto responde a su pertenencia a uno o más “temas”, es decir, a una estructura recurrente e iterativa en la comunicación [25]. Este es el objeto de nuestra **RQ2**.

Para empezar a explorar esta pregunta se analizaron las correlaciones entre las palabras en el nivel de las respuestas, es decir, palabras que se evocaron juntas. Con

este dato se construyó una red, y luego se detectaron automáticamente comunidades o agrupamientos (ilustradas con la forma del punto) (K=6). Para poder evaluar si estos subgrupos se corresponden con valoraciones distintas, hemos incluido la valoración media de cada término evocado, dato que los participantes aportaron siguiendo la consigna de “[indicar] si estas ideas que acabas de introducir se corresponden con algo que valoras positivamente (algo que te agrada) o negativamente (algo que te desagrada)”. En este sentido, la valoración se nos presenta como un posible indicador de la actitud hacia (el recorte de) un objeto de representación.



**Fig. 3.** Correlación entre palabras en la misma respuesta. La forma del punto remite al grupo o comunidad. El color del punto y el término denota la valoración media, en una escala de rojo (valoración baja de 1/10) a verde (valoración alta de 10/10).

Si leemos el gráfico siguiendo el ranking de valoraciones medias, observamos un primer grupo de valoración muy alta, cercana a 7.5/10, con los términos *análisis*, *complejidad*, *investigación*. Son términos que parecen ubicar al big data en su reclamo epistémico.

Los próximos dos grupos de términos mezclan las ideas del núcleo central con otros de su primera y segunda periferia. El primero de estos grupos incluye los términos *informática*, *grande*, *dato*, *cantidad*, *mucho*, *importante*, *comunicación*. La valoración media de este grupo es casi tan alta como la de su promesa, cerca de 7.3/10. Luego, aparecen los términos *red*, *grandes datos*, *internet muchos datos*, *almacenamiento*, *base de datos*, *programación*, *computadora*, *velocidad*, *mucha*

*información*. La valoración media de este grupo se mantiene en una orientación positiva (7.1/10). En ambos grupos conviven nociones cercanas a las mencionadas 3 V's, con otros relativos a la computación, mixtura que no debería sorprendernos, en tanto que, como recuerda Diebold [26], la noción de "big data" se popularizó originalmente dentro del sector informático como un término para referir a los desafíos que se comenzaban a plantear en materia de almacenamiento y consulta de grandes volúmenes de información, y que luego se plasmaron en las 3 V's popularizadas por los medios de comunicación masivas.

El siguiente grupo incluye términos como *predicción, futuro, conocimiento, número, decisión, interés, sistema, publicidad, poder, tecnología, información*. Estos son todos términos que podrían referir a la "promesa" del big data, especialmente en lo que respecta a su utilización. La valoración media de este grupo es apenas inferior a las del anterior (casi 7/10).

Los últimos dos grupos incluyen términos que podrían referir al trasfondo "social" del big data. El primero incluye a los términos *datos personales, masivo y comunicación*, los cuales denotan la naturaleza social y personal de los datos. Su valoración media tiende a una connotación negativa (4.3/10). Esta tendencia se profundiza más en la última comunidad (con una valoración media cercana a 3.8/10), donde los términos se vinculan directamente con los riesgos del big data, tanto en un sentido más inespecífico como en aplicaciones particulares: *espionaje, manipulación, control, marketing, privacidad, nube, Facebook, estadístico, red social*.

### 3.3 Fuentes de difusión y consulta del big data

De acuerdo con Moscovici [9], algunas representaciones sociales han tenido difusores que, con diferentes lógicas, han podido adaptar el objeto de representación a una cierta imagen de su público. Los medios de comunicación masivos como, por ejemplo, la prensa con sus líneas editoriales y secciones especializadas, suelen ser canales de difusión de estas representaciones. En otro trabajo, referido al tratamiento del big data por parte de la prensa digital argentina [10], describimos de qué manera su presentación difiere en contextos tan distintos como su presentación a un público masivo e inespecífico, su tematización dentro de las noticias relativas al trabajo, el empleo o la formación, o incluso a su incorporación en la agenda política.

En el presente trabajo nos limitamos a indagar acerca de los medios consultados para informarse acerca del big data, y nos preguntamos específicamente por el aporte de las carreras universitarias.

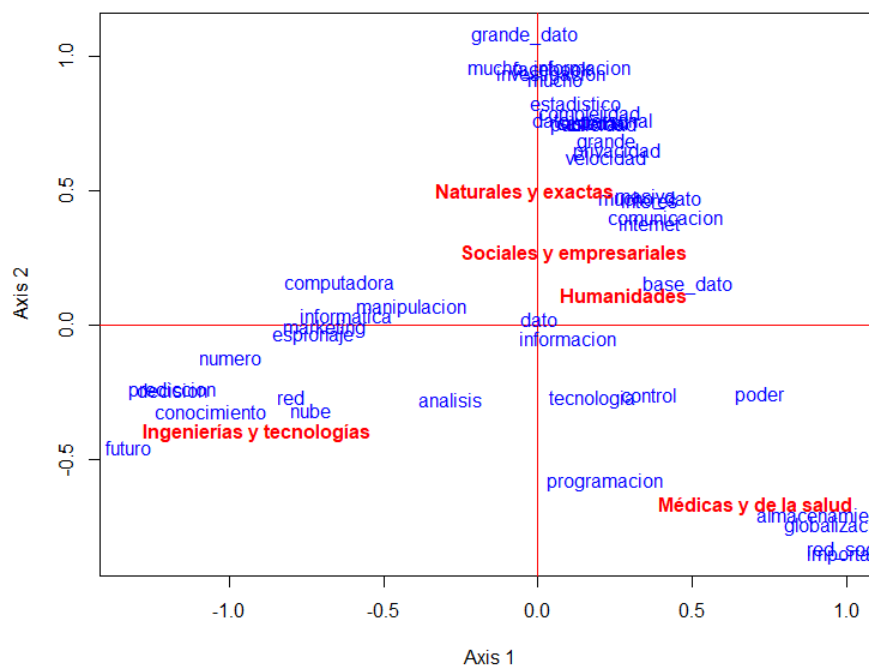
Con respecto a las fuentes, y como se observa en la siguiente tabla, los participantes de todas las carreras consultan mayormente sitios web; quienes más se informan del big data por medio de su carrera son los estudiantes y graduados de naturales y exactas, y de ingenierías y tecnologías, aunque muchos más consultan posteos y sitios; quienes más consultan diarios y sitios de noticias son los ciencias médicas y de la salud, y sociales y empresariales, ambos en un porcentaje apenas menor que a la consulta de posteos y sitios.



**Tabla 1.** % de casos que afirmaron consultar un medio, por grupo de carreras. \* Cursos refiere a formación por afuera de la carrera; \*\* Formación refiere a alguna instancia de la carrera. Hum = Humanidades; Ing. = Ingenierías y tecnologías; Med. = Médicas y de la salud; Nat. = Naturales y exactas; Soc. = Sociales y empresariales.

	Hum.	Ing.	Med.	Nat.	Soc.
Cursos*	35%	43%	9%	39%	43%
Diarios y noticias	33%	29%	43%	34%	46%
Formación**	42%	50%	29%	45%	39%
Posteos y sitios	63%	79%	43%	71%	49%

Una forma posible de indagar la incidencia de estos medios en la representación social es por medio de un análisis de correspondencia que permita representar en forma simultánea los términos y los grupos de carrera.



**Fig. 4.** Análisis factorial de términos y grupos de carreras.

En el centro del gráfico observamos a los términos más salientes del núcleo central (*dato e información*). El grupo de términos más cercano coincide con el grupo de carreras vinculadas a las humanidades, las ciencias sociales y empresariales, y las naturales y exactas. Las principales diferencias se encuentran entre las carreras vin-

culadas a las ingenierías y tecnologías, y las médicas y de la salud. Si volvemos a la tabla de fuentes consultadas, vemos que entre estas dos carreras se dan los contrastes más fuertes en tanto a la relevancia de la carrera para formarse en big data, o al interés dado a los diarios y sitios de noticias. Si observamos la superposición con los términos podemos ver además que las ingenierías y tecnológicas se vinculan mayormente con los términos del discurso de mayor valoración del big data, tales como *futuro*, *conocimiento* o *predicción*, mientras que las de las ciencias médicas y de la salud tienen una valoración menor y refieren mayormente a una de sus dimensiones sociales.

## 4 Conclusiones

En este trabajo presentamos resultados muy preliminares de una nueva línea de investigación, dentro de las coordenadas de la psicología social, que tiene por centro la representación social del big data. Dado el estado muy inicial del relevamiento, aquí nos limitamos a compartir algunos primeros análisis y algunas conjeturas para exploración más profunda en trabajos siguientes.

También hemos ensayado una primera combinación de análisis clásicos de esta tradición de psicología social, con algunas técnicas menos exploradas, como el uso de grafos para construir discursos, sin perder de vista el análisis estructural. Todos estos análisis se han implementado en lenguaje R, y se espera ponerlos a disposición de la comunidad por medio de un *package*.

## Referencias

1. Becerra G (2018) Big data como objeto de estudio y método para la investigación empírica en sociología y psicología social. In: 47 Jornadas Argentinas de Informática & Simposio Argentino de Tecnología y Sociedad. Sociedad Argentina de Informática, Buenos Aires
2. Becerra G (2018) Interpelaciones entre el Big data y la Teoría de los sistemas sociales. Propuestas para un programa de investigación. Hipertextos 6:41–62
3. Wagner W, Hayes N (2011) El discurso de lo cotidiano y el sentido común. La teoría de las representaciones sociales. Anthropos, Barcelona
4. Moscovici S (2001) Why a theory of social representations?
5. Jodelet D (1985) La representación social: fenómenos, conceptos y teoría. In: Moscovici S (ed) Psicología Social II. Paidós, Barcelona, pp 17–40
6. Marková I (2017) The making of the theory of social representations. Cad da Pesqui 47:358–374
7. Wagner W (1998) Social Representations and beyond: Brute Facts, Symbolic Coping and Domesticated Worlds. Cult Psychol 4:297–329. <https://doi.org/10.1177/1354067X9800400302>
8. Howarth C (2006) A social representation is not a quiet thing: Exploring the critical potential of social representations theory. Br J Soc Psychol 45:65–86. <https://doi.org/10.1348/014466605X43777>

9. Moscovici S (1979) *El psicoanálisis, su imagen y su público*. Huemul, Buenos Aires
10. Becerra G (2019) La construcción del big data en la prensa digital argentina. In: XIII Jornadas de Sociología. Universidad de Buenos Aires, Buenos Aires
11. boyd D, Crawford K (2012) Critical Questions for Big Data. *Information, Commun Soc* 15:662–679
12. van Dijck J (2014) Datafication, dataism and dataveillance: Big data between scientific paradigm and ideology. *Surveill Soc* 12:197–208
13. Kitchin R (2014) *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. Sage, Los Angeles
14. Portmess L, Tower S (2015) Data barns, ambient intelligence and cloud computing: the tacit epistemology and linguistic representation of Big Data. *Ethics Inf Technol* 17:1–9. <https://doi.org/10.1007/s10676-014-9357-2>
15. Puschmann C, Burgess J (2014) Metaphors of Big Data. *Int J Commun* 8:1690–1709
16. Dany L, Urdapilleta I, Lo Monaco G (2014) Free associations and social representations: some reflections on rank-frequency and importance-frequency methods. *Qual Quant* 49:489–507. <https://doi.org/10.1007/s11135-014-0005-z>
17. Lo Monaco G, Piermattéo A, Rateau P, Tavani JL (2017) Methods for Studying the Structure of Social Representations: A Critical Review and Agenda for Future Research. *J Theory Soc Behav* 47:306–331. <https://doi.org/10.1111/jtsb.12124>
18. Barreiro A, Gaudio G, Mayor J, et al (2014) La justicia como representación social: Difusión y posicionamientos diferenciales. *Rev Psicol Soc* 29:319–345. <https://doi.org/10.1080/02134748.2014.918821>
19. Team R Core (2018) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna
20. Wickham H (2019) Welcome to the tidyverse. *J Open Source Softw* 4:. <https://doi.org/https://doi.org/10.21105/joss.01686>
21. Lebart L, Piron M (2013) *Práctica del análisis de los datos numéricos y textuales con Dtm-Vic*. INDEC - Instituto Nacional de Estadística y Censos, Buenos Aires
22. Abric J-C (1993) Central System, Peripheral System: Their functions and roles in the dynamics of social representations. *Pap Soc Represent* 2:75–78
23. Laney D (2001) 3D Data Management: Controlling Data Volume, Velocity, and Variety
24. Elish MC, Boyd D (2018) Situating methods in the magic of Big Data and AI. *Commun Monogr* 85:57–80. <https://doi.org/10.1080/03637751.2017.1375130>
25. Luhmann N (2000) *The reality of the mass media*. Stanford University Press, California
26. Diebold FX (2012) The Origin(s) and development of “ Big Data ”: the phenomenon , the term , and the discipline. *Penn Econ Work Pap*. <https://doi.org/10.2139/ssrn.2202843>