Máster en Medios, Comunicación y Cultura
Departamento de Medios, Comunicación y Cultura
Facultad de Ciencias de la Comunicación
Universitat Autònoma de Barcelona

El debate académico en curso sobre 'big data' y su incidencia en la comprensión de la comunicación mediática contemporánea

Trabajo de Fin de Máster

Autor: Yuri Cesarotto
Director: Prof. Albert Chillón

Bellatera (Cerdanyola del Vallès)
Junio de 2018

Universitat Autònoma de Barcelona

SUMARIO

Resumen

Los datos masivos ('big data') están en el centro de importantes debates por parte de investigadores y académicos de distintos campos científicos, a la vez que su normalización ha generado un gran número de preocupaciones sociales que van desde la vigilancia masiva hasta la predicción de hábitos de consumo.

La minería de datos y el análisis de correlaciones entre grandes conjuntos de datos, producto de la digitalización y la datificación de nuestro entorno, generan un gran impacto en la forma en que entendemos y extraemos conocimientos de eventos naturales y sociales. El 'big data', fenómeno emergente en la sociedad hiperdigitalizada, trae desafíos y consecuencias que apuntan a un cambio de paradigma en las estructuras de las sociedades contemporáneas, sea en la dimensión social, tecnológica o científica.

Esta investigación pretende establecer un *status quaestionis* sobre el importante debate académico en curso en este momento. Así, el presente trabajo tiene como principal objetivo identificar y cartografiar los conceptos, materiales relevantes e investigaciones actuales del fenómeno en las disciplinas del campo de la comunicación mediática, como forma de establecer un documento propedéutico para futuras investigaciones sobre el tema.

Palabras clave: 'big data' – datificación – minería de datos – comunicación mediática

Resum

Les dades massives ('big data') estan al centre d'importants debats per part d'investigadors i acadèmics de diverses àrees científiques, alhora que la seva normalització ha generat un gran nombre de preocupacions socials que van des de la vigilància massiva fins a la predicció d'hàbits de consum.

La mineria de dades i l'anàlisi de correlacions entre grans conjunts de dades, fruit de la digitalització i la datificació del nostre entorn, generen un gran impacte en la manera en què entenem i extrèiem coneixement d'evsdeveniments naturals i socials. El "big data ", fenomen emergent a la societat hiperdigitalitzada, comporta desafiaments i conseqüències que apunten a un canvi de paradigma a les estructures de les societats contemporànies, sigui a la dimensió social, tecnològica o científica.

Aquesta investigació pretén establir un status questionis sobre l'important debat acadèmic en cus en aquest moment. Així, el present treball té com a objectiu principal identificar i posar en relleu els conceptes, materials rellevants i investigacions actuals del fenomen a les disciplines de la comunicació mediàtica, com a manera d'establir un document propedèutic per a futures investigacions sobre el tema.

Paraules clau: 'big data' – datificació – mineria de dades – comunicació mediàtica

Abstract

The massive data ("Big Data") is the centerpiece of important debates among researchers and academics from different scientific fields, as its normalization has generated a great number of social concerns, from mass surveillance to the prediction of consumption habits.

Data mining and the analysis of correlations between large data sets, product of digitization, and the datification of our environment cause a great impact on the way we understand and absorb knowledge from natural and social events. Big Data - an emerging phenomenon in the hyperdigitalized society – brings challenges and consequences that indicate a paradigm change in the structures of modern society, whether it is in the social, technological, or scientific aspect.

This research aims to establish a status quaestionis about the significant current academic debate. Consequently, the main purpose of this work is to identify and map concepts, relevant materials, and present research on the big data phenomenon in the disciplines of media communication, as a way of establishing a propaedeutic document for future research on the matter.

Key words: 'big data' – datification – data mining – media communication

1. Introducción

En el siglo XVII, el científico y matemático inglés Isaac Newton proclamó: "Lo que sabemos es una gota, lo que ignoramos es un océano" (Pou, 2017). De hecho, su pensamiento apuntaba a la percepción de que el conocimiento humano sobre su entorno era aún incipiente, en el momento en que la revolución científica despuntaba y pretendía responder a los principales cuestionamientos no a partir de la visión teológica dominante, sino a partir de la vía empírica de la experiencia humana con el mundo.

Esta frase revela dos aspectos que los científicos de la época encontraban al desarrollar sus experimentos. En primer lugar, era laborioso y costoso recoger grandes cantidades de datos acerca de un evento o un fenómeno determinado. Segundo, el conocimiento generado en ese proceso no agotaba sus búsquedas de la determinación de leyes generales para explicar el funcionamiento del mundo. Estos aspectos eran característicos de la era analógica del conocimiento humano, también marcada por la restricción tecnológica para captar grandes cantidades de datos[1].

En este comienzo del siglo XXI, el ascenso de un nuevo paradigma se ha puesto de manifiesto, no sólo por el hecho de determinar que las correlaciones entre diversos acontecimientos sean más importantes que sus relaciones causales, sino por la enorme capacidad que demuestra para expandir el conocimiento sobre dichos acontecimientos, ya que permite compilar y procesar todas las unidades de datos, y no una muestra o pequeña parte (N = todo).

A partir de la digitalización exponencial del mundo captada, mediada y reproducida por los dispositivos mediáticos y de vigilancia, acumulamos tal volumen de datos que su almacenamiento y procesamiento sólo pueden ser operados a partir de máquinas, ya que esta tarea ya no es posible a nivel humano.

[1] Algunos autores han llamado como la era del *small data*, en oposición a la actual del *big data*. Por su vez, Mayer-Schoenberger & Cukier (2013) prefieren usar el concepto de "era de los datos escasos" (p. 31).

A este paradigma se ha convenido en llamarlo 'big data', o la era de los datos masivos.

Sin embargo, la idea de que todo pueda transformarse en datos accesibles es un sueño antiguo del ser humano que nos acompaña desde las primeras etapas del desarrollo de las sociedades actuales. Fue esa premisa la que llevó al primer gran paso en ese sentido: el surgimiento de la escritura.

Hoy se considera que entre las civilizaciones mesopotámicas pre-cristianas la escritura surgió a partir de la necesidad de registrar bienes materiales y transacciones comerciales de la época, como una especie de método de contabilidad rudimentaria. La escritura, por lo tanto, era esencial para el control y la organización de esas sociedades que comenzaban a estructurarse alrededor de la agricultura y del comercio.

La importancia de la escritura para la historia de nuestra especie y para la conservación de nuestros registros viene del hecho de que permitió el almacenamiento y la propagación de informaciones no sólo entre individuos, algo posible sólo en virtud del lenguaje verbal, sino también a través del tiempo, de generación en generación.

El segundo gran paso hacia ese ideal fue dado dos milenios después con el surgimiento de la máquina de prensa móvil, inventada por Gutenberg en el siglo XV. Este sistema mecánico de reproducción permitió superar la difusión de los saberes por medio de los manuscritos, que eran los soportes hegemónicos en ese momento. Esta fue la semilla para el establecimiento de una cultura basada en el registro de datos de cualquier naturaleza, así como para nuestra actual economía, basada en el conocimiento.

El tercer gran paso fue posible gracias al desarrollo, ya en el siglo XX, de las máquinas de cálculos computacionales, o simplemente, computadoras. En un primer momento, estaban basadas en procesos electromecánicos que, tras la

incorporación de las nuevas técnicas de cálculo binario, evolucionaron hacia los actuales modelos de procesamiento totalmente digitales.

El asombro causado por las primeras aplicaciones de esas máquinas nuevas aturdía e inspiraba a intelectuales y escritores de entonces, que vislumbraban el desarrollo de una supercomputadora con inteligencia avanzada, capaz de interactuar con las personas y reunir todos los datos posibles sobre el conocimiento humano.

H. G. Wells[2], por ejemplo, sugirió que los científicos deberían preocuparse por construir una especie de enciclopedia mundial que indexaría todo libro, escritura, documento, etcétera, escrito por la humanidad hasta el momento. Este sistema debería ser una red de centros científicos que actualizarían constantemente el conocimiento humano producido, posibilitando que cualquier persona accediese desde su casa en el momento que fuera.

Esta mística *wellesiana* a propósito de la posibilidad de concentrar todo el conocimiento humano sobre la realidad en una única plataforma inspiró diversos proyectos que hoy tratan de hacer realidad ese sueño, como es el caso del Proyecto Gutenberg, Google Print o incluso de la Wikipedia, la más popular de todas.

Como se ve, todos estos pasos a lo largo de nuestra historia fueron saltos cuantitativos en base a la creencia de que los fenómenos del mundo y nuestra relación con ellos pueden ser observados, registrados y cuantificados en forma de datos accesibles. La acumulación exponencial y el perfeccionamiento de estos procesos, además de la omnipresencia de los nuevos dispositivos tecnológicos capitaneados por el *smartphone*, desembocaron en la delta del rio informacional en que hoy nos encontramos.

Llegar a este escenario de abundante oferta de datos acerca de cualquier actividad humana fue posible, por un lado, por el predominio del proceso de digitalización

[2] Estas ideas se exponen en el libro *World Brain*, cuyo contenido recompila escritos y discursos de ese autor sobre el tema (Wells, 1938).

de nuestro entorno; por otro, por la creciente normalización de la datificación que atravesamos.

Alan Turing, en su texto *Computing Machinery and Intelligence* (1950) ya preveía que el avance del procesamiento digital nos llevaría a un estado de completa normalización de las máquinas tecnológicas. Esta capacidad de penetración e integración en muchos aspectos de nuestra sociedad, ya sean económicos, políticos, militares, etc., podría llevar al estado de invisibilidad de esos dispositivos; tanto, que las personas ni siquiera se darían cuenta.

Este tipo de reflexión llevaría a Turing a proponer una prueba[3] para identificar si estas computadoras, acercándose a un estado de inteligencia pensante, podrían pasar por humanos en actividades simples, como una conversación. Si un juez humano no puede distinguir quién es hombre y quién es máquina en esa conversación, el ordenador superaría la prueba, y sería considerado una máquina pensante (Turing, 1950).

Resulta interesante notar que las preocupaciones planteadas por Turing en la ya lejana década de los cincuenta estarían hoy más presentes que nunca. También sorprende saber que computadoras inteligentes son capaces de superar la prueba, cada vez con más naturalidad[4].

Por medio de la aparición de nuevos dispositivos y métodos que permiten captar, procesar y almacenar datos de forma cada vez más rápida y barata, este proceso de datificación se vuelve más normalizado dentro de nuestro mundo social. El nivel de aceptación y de credulidad que este proceso suscita es tan elevado que algunos autores han llamado la atención sobre el surgimiento de una nueva ideología: el

[3] En ese artículo Turing no llama su idea "Prueba de Turing", sino "Juego de la Imitación". Con el objetivo de responder a la pregunta "¿Pueden las máquinas pensar?", él propone el siguiente juego: un hombre y una máquina serían los jugadores A y B y deberían responder a preguntas hechas por un interrogador humano que sólo tiene contacto con ellos por medio de una interfaz en la que se establecen las preguntas y respuestas. Si este interrogador, después de varias preguntas, no sabe distinguir qué jugador, A o B, es humano, la máquina superaría la prueba, y así podríamos afirmar que, de alguna forma, esa máquina piensa.

[4] La tecnología Duplex de la empresa Google causó gran repercusión al ser presentada recientemente pues puede establecer diálogos simples y dinámicos con seres humanos sin que éstos sepan que interactúan con una máquina (Coutinho, 2018).

dataísmo. La creencia en el dataísmo consiste en entender el universo como un flujo constante de datos, donde cualquieras entidades y los agentes sociales aportan elementos a este flujo por medio de sus acciones (Harari, 2015).

Con ese punto de vista, no es difícil suponer que el ser humano, a partir del momento en que dispone de métodos para la cuantificación total y análisis de esos datos, podría tener acceso profundo a la realidad humana, en cuanto construcción histórica y social, comprendiéndola por la investigación de las correlaciones que en ella se operan. Este es el principal punto de apoyo para aquellos que ven en el 'big data' la solución para muchos problemas de nuestras sociedades.

En la serie de documentales *All watched over by machines of loving grace*[5], producida por Adam Curtis asociado con la BBC, se nos muestra como en la década de los sesenta se gestó la utopía de que el poder de las máquinas nos conduciría a una sociedad en red que prescindiría, entre otras cosas, de políticos, y que podría auto gestionarse con la ayuda de los inmensos ordenadores que surgían ya por entonces.

Así, grupos de jóvenes visionarios de la costa oeste de Estados Unidos, alimentados por la ideología de una sociedad auto-organizada, interconectada y constituida a nivel global, dejaron de lado el desarrollo de grandes máquinas computacionales y pasaron a invertir en nuevas tecnologías que pudieran conectar pequeños ordenadores personales en redes. Surgió así el *Silicon Valley*, el polo industrial y tecnológico más importante e influyente de nuestro tiempo, en el que se gestan las principales empresas y centros de investigación de las tecnologías que nos rodean (Curtis, 2011). De hecho, en las listas de las marcas y compañías más valiosas del mundo actual, seis de las 10 primeras son del sector de tecnología y están, en cierta forma, ligadas a esta revolución fermentada en ese contexto [6].

[5] Este título hace referencia a un poema de Richard Brautigan publicado en 1967 bajo el mismo nombre, en el cual el autor describe una sociedad donde los hombres estaban libres de trabajo y la naturaleza había alcanzado su estado de equilibrio, todo gracias al avance de los robots (Madrigal, 2011).

[6] La lista de Forbes se actualiza constantemente siguiendo el valor de mercado de las marcas (Forbes, 2018)

Es un hecho que el avance y perfeccionamiento de las técnicas de registro y las nuevas tecnologías de procesamiento de datos, o sea, la combinación entre la expansión de los mecanismos de obtención y almacenamiento y de la velocidad de procesamiento, permitan la solidificación de ese estado híper-masivo de información, que es la característica del 'big data'.

El carácter exponencial de ese desarrollo puede ser entendido en esta anécdota contada por Negroponte (1995):

> ¿Conoce ese acertijo en el que se pregunta cuánto dinero se tendría al cabo de un mes, si se ganara sólo un centavo por día durante ese período, pero duplicando el salario cada día? Si se comenzara ese fantástico programa salarial el día de Año Nuevo, el último día de enero se estaría ganando más de 10 millones de dólares por día. Ésta es la solución que recuerdan todos los que conocen el acertijo. Lo que casi nadie analiza, es que, con el mismo esquema de pago, sólo se ganaría 1,3 millones de dólares si enero tuviese tres días menos (como, por ejemplo, febrero). Planteado de otra forma, el ingreso acumulativo por todo el mes de febrero sumaría alrededor de dólares 2,6 millones en lugar de los 21 millones de dólares que se hubieran ganado en enero. Al tener un efecto exponencial, esos últimos tres días significan muchísimo. En computación y telecomunicaciones digitales nos estamos acercando a esos últimos tres días. (p. 01)

Esto muestra cómo el desafío de la humanidad en querer cuantificar aspectos de nuestra realidad parece estar limitado apenas por la extensión del desarrollo y la capacidad de procesamiento, almacenamiento y uso de los datos disponibles. Y los límites están cada vez más extendidos.

En 2016, IBM estimó que diariamente se producían $2{,}5 \times 10^{18}$ de bytes de datos digitales; dicho de otra forma, algo alrededor de 2.500.000 de discos duros que se llenan todos los días por los más diversos tipos de datos que se puedan generar, capturar o medir (IBM, 2016). Así, cada dos años, aproximadamente, la cantidad de datos generados y almacenados por la humanidad se duplicaría.

Si tomamos en cuenta sólo los números generados por el uso de servicios de comunicación instantánea y medios sociales, las cifras pueden ser consideradas obscenas. En tan solo una hora, trocamos 3.000 millones de mensajes en Whastapp

(Moreno, 2018), se suben 2 millones de fotografías a Instagram, se dan 9.375.000 likes en Facebook, se lanzan 20 millones de *tweets*, se visualizan 183 millones de videos en Youtube y se genera un tráfico total de 113.400 terabytes en internet (Llorens, 2017).

Resulta evidente que una nueva forma de organización social va surgiendo. La normalización del 'big data', junto con un paquete de otras tecnologías, como la inteligencia artificial, la automatización, el almacenamiento de datos en la nube, la nanotecnología, la impresión 3D y la *Internet de las cosas*, son la base de lo que está siendo considerada como la 4ª Revolución Industrial (Salesforce, 2018).

A grandes rasgos, podría entenderse como un conjunto de cambios en los procesos y en la forma en que se fabricarán los productos que consumimos y los servicios de que disponemos. En este contexto, el mundo físico y el virtual (datos generados durante la fabricación, transporte y consumo) están unidos a través de la conexión en red, de modo que ambos pueden alimentarse de este proceso.

Pero no es sólo en la industria y en el mundo empresarial donde los datos masivos se están convirtiendo en una cuestión central. Esto impacta también en toda la cadena de producción científica y el universo académico, que lo ven como una revolucionaria oportunidad de investigación para entender el comportamiento humano.

Esta nueva forma de comprender y explorar nuestro entorno tiene una potencialidad tal que generó un importante debate después de que Anderson (2008), en su artículo *The End of Theory*, proclamase que las teorías científicas no serían más fundamentales para la construcción del conocimiento, ya que el análisis de correlaciones dentro de un conjunto masivo de datos sería suficiente para entender lo que los datos *quieren decir*, sin la necesidad de aplicar una teoría general previa para ello.

En respuesta, muchos académicos negaron tal entendimiento. En este sentido, cabe destacar la posición taxativa de Mayer-Schoenberger & Cukier (2013):

> Esto es ridículo. El propio enfoque de los datos masivos está basado en la teoría. Por ejemplo, emplea teorías estadísticas y matemáticas y, en ocasiones, recurre también a la teoría de las ciencias informáticas. Sí, éstas no son teorías acerca de las dinámicas causales de un fenómeno determinado como la gravedad, pero no por eso dejan de ser teorías. Y los modelos basados en ellas tienen un poder de predicción muy útil. De hecho, los datos masivos pueden ofrecer una perspectiva fresca y enfoques nuevos precisamente porque no están lastradas por el pensamiento convencional ni por el sesgo inherente implícito en las teorías de un campo determinado. (p. 93)

La gran euforia que inicialmente se levantó con las afirmaciones de Anderson comenzó a dar lugar a un rico debate acerca de los efectos sociales que los enormes conjuntos de datos y los algoritmos utilizados para correlacionarlos y analizarlos están generando.

Como en una especie de actualización del combate entre apologistas y apocalípticos sobre la cultura y los medios de masa, expuesto por Umberto Eco en *Apocalípticos e Integrados* a finales delos años sesenta, hoy se nota una misma polarización cuando se observa la producción académica sobre las NTICs[7].

Por un lado, muchos investigadores parecen hoy mirar la tecnología, y más precisamente el 'big data', como instrumentos eficaces del progreso humano, capaces de acelerar el desarrollo social, cultural, científico, e incluso como forma de estimular la participación democrática.

Por otro, una parte de autores más críticos demuestran cierto rechazo ante las actuales tecnologías, las cuales pueden mostrarse como instrumentos maléficos de nuestras sociedades, capaces de alterar negativamente el tejido social, como prácticas de hipervigilancia de los individuos o la hipermediatización de las relaciones humanas.

[7] El concepto de Nuevas Tecnologías de Información y Comunicación (NTIC) adquiere relevancia en la década de los 90 con la popularización de los ordenadores personales y está asociada a las nuevas formas y medios de comunicación posibilitados por la revolución digital.

Este debate revela que el tema genera aún muchas preocupaciones y desconfianzas por la aplicación de las nuevas tecnologías en diversos ámbitos. También es bastante actual, ya que el ritmo de transformaciones al que asistimos es cada vez más acelerado y plantea cuestiones inéditas, muchas veces atropellando a otras sobre las que aún no hemos sido capaces de reflexionar. Con ello, se percibe que en la mayoría de los sectores económicos, el 'big data' exige nuevas prácticas y habilidades de los profesionales, lo que supone un cambio en sus perfiles, algo que, como veremos con más detalles, está ya ocurriendo en el sector de las comunicaciones mediáticas.

Así, en los próximos apartados se espera desentrañar los aspectos formales de este trabajo, para explicar el enfoque de la investigación, su justificación y las hipótesis y objetivos pretendidos con este estudio. También se presenta una breve fundamentación que nos presentará algunos de los principales autores que articularemos en el cuerpo del trabajo y que nos ayudarán a entender cómo los macro datos irrumpen actualmente en nuestro entorno y en las investigaciones del campo académico.

Aquí vale aclarar que la utilización del término 'big data' deriva de su origen anglófono y es la forma más común de ser utilizada y encontrada, independiente de la región o lengua. En español, el término fue traducido como *datos masivos* o *macro datos* y aparecen de ambas formas en muchos estudios. En este trabajo, se utilizarán las dos formas (inglés y español) de referirse al mismo fenómeno, a fin de evitar la repetición de los términos.

Siguiendo, en el cuerpo de la investigación (propuesta y condensada a partir del capítulo 2), se articulará el *status quaestionis* en tres diferentes partes. En la primera, se mostrarán los fundamentos estructurantes del 'big data', las dimensiones características y otras bases conceptuales involucradas.

Posteriormente actualizaremos, a la luz del tiempo presente, las grandes discusiones afectadas por el cambio de paradigma en nuestra relación con los datos: de la escasez a la abundancia; de la probabilidad a la totalidad; de la

causalidad a las correlaciones; entre lo público y lo privado; finalmente, del pasado al futuro.

Una vez trillado ese camino, partiremos hacia la tercera parte, en la que se presentarán las implicaciones para la comprensión de las ciencias de las comunicaciones, con especial énfasis en las disciplinas y prácticas adoptadas como grandes campos de estudio de las comunicaciones mediáticas.

Por último, serán explicitados los resultados y conclusiones después del recorrido del *status quaestionis*, así como la bibliografía utilizada para erigir esta investigación.

1.1 Delimitación del objeto de estudio y justificación

El objeto de estudio de esta investigación es el debate académico en curso sobre el 'big data' y sus implicaciones en las disciplinas académicas que estudian acerca de la comunicación mediática.

Este planteamiento pretende dar respuesta a la creciente normalización del uso del 'big data', sea por organizaciones y empresas, o sea en el ámbito científico, además de su reconocimiento como un fenómeno tecno-social reciente y poderoso. Su repercusión es amplia, dado el hecho de que permea diversas áreas de estudio e impacta en disciplinas académicas muy distintas, como la genética o la lingüística.

Soy consciente de que este ámbito de aplicación y, en consecuencia, los estudios derivados generan debates e investigaciones en diversos centros académicos del mundo, adoptando tantas formas como las lenguas y enfoques investigativos parecen permitir. Por lo tanto, el primer corte de este estudio se refiere a la limitación natural de tan sólo abordar los trabajos y materiales en lenguas ibéricas (portugués y español) y en inglés, donde está la mayor parte del cuerpo de la producción actual sobre el tema.

La importancia de este estudio estriba sobre el hecho de que hay un fuerte impacto de muchas aplicaciones e investigaciones sobre y mediante el uso de 'big data' en el campo de las humanidades y las ciencias sociales y, en este caso, en los estudios en comunicación mediática y periodismo. Como se pretende demostrar, este impacto inicial tiene potencial de desdoblamiento en diversas facetas, lo que genera falta de consensos, conceptos frágiles o inexistentes, además de diversos desafíos de orden epistemológico y técnico, a pesar de que esta última no es el enfoque de la disertación propuesta. Muchos otros trabajos ya se ocupan de eso.

Así, asumiendo la importancia que muchos investigadores han atribuido al tema y de común acuerdo con el director de este estudio, la muestra de delimitación con el que se diseña esta investigación se centra en el material académico sobre el 'big data', en especial aquellos textos – libros, artículos, investigaciones diversas – que tratan de los impactos en los estudios de la comunicación mediática contemporánea.

1.2 Objetivos

A partir del progreso de la investigación que se inicia ahora, este trabajo de Fin de Máster tiene como principal objetivo identificar y mapear los conceptos, materiales relevantes e investigaciones actuales del fenómeno del 'big data' en estas disciplinas académicas como forma de establecer un documento propedéutico sobre el tema, un primer paso para la profundización del estudio en un proyecto de doctorado.

Puesto que el cuerpo de conocimiento construido a partir de estudios, aplicaciones y prácticas en muchas disciplinas académicas encuentra paralelismos interesantes en torno a algunos conceptos ya establecidos, también parecen indicar que hay un gran trecho para recorrer antes de llegar a una base epistemológica común, en vista de normas mejor definidas.

En nuestro caso, apuntamos la mirada para el lado de las comunicaciones mediáticas, un campo más amplio que nos permitirá aglomerar producciones e

investigaciones en redes sociales, análisis del discurso, periodismo de datos, entre otros, además de, muchas veces, también explorar preocupaciones de estudios más críticos, como los recientes debates sobre escrutinio de datos, vigilancia, privacidad, etcétera. Ellos también tienen repercusiones en el campo de la filosofía del conocimiento.

Por lo tanto, como objetivo secundario, se busca explorar y revelar las características principales del objeto de estudio a fin de organizar una base de consulta que también sea útil para otros trabajos e investigadores del tema propuesto, en especial dentro de España, además de otros países de Latino América.

1.3 Preguntas de Investigación (hipótesis)

En los estudios de carácter hipotético-deductivos es fundamental la demostración de una conjetura propositiva que se juzga verdadera mediante hipótesis. Someter éstas a pruebas, experimentos y testes, es decir, probar su potencial de falseamiento, es una forma consistente para determinar si las iniciales se confirman o no, pavimentando la construcción del conocimiento científico riguroso.

Sin embargo, dentro de las ciencias sociales y humanas, muchas veces la\s hipótesis ganan un peso mayor en la función de orientar y delimitar el camino del estudio propuesto, además de complementar otros trabajos sobre el mismo tema, ya que los objetos de estudio en dichas disciplinas son pasibles de recibir diversos enfoques acerca de la resolución de los problemas de investigación y dado además que un número significativo de investigaciones no son del tipo analítico-deductivo sino argumentativo e interpretativo.

Hecha la aclaración, el presente trabajo asume que las hipótesis presentadas abajo y apuntadas al objeto de estudio deberán servir de guía para que, a partir de la perspectiva metodológica adoptada, podamos validar lo que parece una cierta

euforia sobre nuevas posibilidades que el uso del 'big data' permite. Así, las principales hipótesis que se plantean son:

(i) La ascensión del 'big data' como nuevo fenómeno tecnológico y social altera significativamente la forma que se construye el conocimiento científico, una vez que permite encontrar correlaciones entre conjuntos de datos por medio de técnicas de minería;

(ii) El salto cualitativo en los análisis, a partir del salto cuantitativo de datos, permite nuevas formas de extracción de información de los fenómenos sociales y naturales que son posibles por medio de la datificación masiva de dichos fenómenos;

(iii) Desafíos y cuestionamientos señalados por recientes estudios sobre 'big data' no se refieren solo a cuestiones de orden técnico, sino también de orden ético, pues la obtención y procesamiento de datos personales están en la raíz de problemas asociados con privacidad y control de información;

(iv) Un cambio de paradigma, en lo que se refiere a los estudios académicos y en el desarrollo de la comprensión del mundo social, ya está ocurriendo a partir de las correlaciones permitidas por el 'big data', pues provoca nuevas creencias, habilidades, técnicas y desafíos para las comunidades académicas de dichas disciplinas;

(v) Al igual que otros campos académicos, los estudios sobre la comunicación mediática también sufren implicaciones con la ascensión de este nuevo paradigma, sea por las nuevas estructuras de medios que se organizan a partir de los datos, sea por la masiva datificación de contenidos, hábitos de las audiencias y nuevas plataformas de difusión surgidas en este contexto.

1.4 Método

El propósito de sedimentar un trabajo propedéutico, de carácter notoriamente exploratorio, es la guía de esta investigación. Así, este trabajo gana rasgos de lo que Umberto Eco llama de *tesis de compilación*, o sea, que pueda ofrecer una visión panorámica del tema y que se muestre útil como documento informativo para

aquellos que no estén al tanto del asunto que se pretende abordar. (Eco, 2001, p. 20)

Por la limitación de recursos y tiempo necesario para el desarrollo de una metodología propia y efectiva en el abordaje de un tema que, por su importancia e impacto, atrae para sí grandes interrogantes, se ha acordado con el director de este trabajo que la investigación debe arrojar luz sobre el conocimiento en cuestión, el estado del arte, principalmente cuando se busca familiaridad con el objeto de estudio, alimentando la ambición de profundizar el tema en futuras investigaciones o para un proyecto de doctorado.

De esta manera, para llegar a los resultados esperados, me propuse efectuar una exploración lo más amplia posible del objeto de estudio. La construcción de la comprensión sobre el tema se da en el levantamiento exhaustivo, dentro de sus posibilidades, de la bibliografía y *docugrafía* disponible, además de *webgrafía*, cuando necesario. En diversos momentos, los asuntos se precipitaron como una "bola de nieve", en la que la lectura de los artículos y libros iniciales fueron llevando a otras lecturas y citas internas que ampliaron la investigación.

De este modo, durante el texto se identifican y citan innumerables artículos, trabajos, tesis, libros, noticias, centros y grupos de investigación, revistas especializadas, institutos, etc. que surgen como referencia en los debates académicos en curso y acaban reflejados en la lista bibliográfica final.

Por lo tanto, y teniendo en cuenta que muchos de los principales trabajos de referencia surgieron en la última década, no hay una limitación temporal en el enfoque, pero hay un cuidado de situarse, cuando sea posible, lo más cerca de la fecha de publicación de esta tesis.

1.5 Fundamento Teórico

Como adelantado en los puntos anteriores, el cuerpo teórico relativo a este gran campo llamado 'big data' es reciente y aún se busca fortalecer conceptos y buenas

prácticas comunes a los distintos usos y enfoques atribuidos. De esta forma, dejando de lado los libros y documentos técnicos sobre programación y *softwares*, no hay de momento un sólido material que se pueda reivindicar como característico y exclusivo de estos conceptos.

En principio, este trabajo propone establecer el *status quaestionis* articulando los trabajos relevantes y recientes que orbitan el tema y que sirven, hasta el momento, como un punto de apoyo en los debates en curso. Este planteamiento hace hincapié en la importancia de entender el estado de arte de los temas que se pretenden explorar y se inspira en la tradición académica germánica de retratar el *zeitgeist* en lo cual esta inserido nuestro objeto de estudio.

Sin embargo, todo estudio académico se articula y sitúa dentro de un contexto temporal y conceptual, sin quedar a la deriva, aislado y cerrado, por los amplios campos del conocimiento humano. Por eso, el propio desarrollo del trabajo consiste en explorar y articular las ideas y autores que considero relevantes en el debate en curso para ayudar en la construcción de este marco teórico. Ahora, entonces, cabe adelantar lo que se contempla en los próximos capítulos de esta investigación.

La principal obra hasta el momento es el libro *Big data. La revolución de los datos masivos*, de los autores Mayer-Schoenberger & Cukier (2013) con lo cual introduce importantes conceptos, como la 'datificación', y explica otros elementos como la 'minería de datos', el uso de algoritmos y, principalmente, el análisis de correlaciones para extraer conocimiento de los datos masivos. Estos aportes son fundamentales para la evolución del recorrido que se propone este trabajo y son la base de la articulación conceptual de la primera parte de la investigación.

Terminaremos esta parte explorando las definiciones de las *dimensiones Vs* que suelen utilizarse para la comprensión de la dinámica de los procesos con 'big data'. Dichas dimensiones son así conocidas pues sus características son explicadas por términos que se inician con la letra V, como volumen, velocidad, etcétera. Para eso,

se utiliza el modelo integrado de In Lee que considero de fácil aprensión para nuestro propósito.

En la segunda parte, se trata de identificar las principales características que sostienen la creencia de que hay un importante cambio de paradigma científico en curso. Esta creencia es fundamentada en las teorías de Thomas Khun sobre las revoluciones científicas y sus paradigmas y que son aclaradoras como marco para entender cómo la tecnología juega un rol fundamental en ese proceso.

El cambio de paradigma ya se va identificando en algunas reflexiones que hacen Mayer-Schoenberger & Cukier (2013) y con las cuales articularemos con otros autores, como Boyd & Crawford (2012) que, a partir de su primer texto intitulado *Critical Questions for Big Data*, exploran estos temas y lanzan algunas provocaciones para el presunto debate.

La tercera y última parte del cuerpo de trabajo se centra en cartografiar los principales debates en curso de los impactos del 'big data' con sus implicaciones específicamente en las disciplinas académicas que tratan de comprender el campo de la comunicación mediática contemporánea.

Ahí se articula una amplia lista de autores y artículos para buscar dónde existen relaciones y paralelismos entre estas diversas líneas académicas y las diferentes perspectivas teóricas. Pero, dentro de este mapa, hay una importante contribución de las ideas de Manovich (2002; 2011) a partir de su línea de investigación de los estudios culturales en software y visualizaciones de datos que nos presenta útiles para tejer la trama de estas implicaciones.

Una vez situados en el contexto, con algunas referencias básicas presentadas como semillas del tema, pasaremos ahora al ejercicio exploratorio que podrá aclarar el estado del arte de la discusión en curso sobre 'big data'.

2. El debate en curso sobre el 'big data' y su incidencia en la comprensión de la comunicación mediática contemporánea. Propuesta de un s*tatus quaestionis*

En su libro *Big data. La revolución de los datos masivos*, Mayer-Schoenberger & Cukier (2013) trazan el camino de cómo esta nueva tecnología está alterando nuestro entorno y cuál será el impacto en nuestra sociedad en el futuro próximo.

La datificación, importante concepto descripto por estos autores, es el proceso de transformación de cualquier fenómeno social o físico en un formato cuantificable, permitiendo así que sean ordenados de alguna forma para posterior análisis.

> La palabra latina data significa 'dado', en el sentido de 'hecho'. Este término se convirtió en el título de una obra clásica de Euclides, que explica la geometría a partir de lo que se sabe. Hoy en día, por dato se entiende una descripción de algo que permite ser registrado, analizado y reorganizado. 'Datificar' un fenómeno es plasmarlo en formato cuantificado, para que pueda ser tabulado y analizado. (Mayer-Schoenberger & Cukier, 2013, p. 100)

El paso siguiente de exploración y manipulación en este proceso se denomina 'minería de datos', que podemos entender como un conjunto de herramientas y técnicas que permiten buscar patrones consistentes, determinar reglas generales o específicas, identificar factores y asignar relación entre los conjuntos de datos, como explicaremos con más detalles.

La obra de Mayer-Schoenberger & Cukier (2013) también trae puntos de vista críticos provenientes del gran proceso de acatamiento del entorno social, planteando importantes cuestiones; por ejemplo, el invasivo acumulo de información de los gobiernos sobre sus ciudadanos, o modernos sistemas policiales de seguridad y prevención de crímenes basados en datos masivos, según lo previsto en la película *Minority Report*. Estas preocupaciones son compartidas por otros investigadores que buscan comprenderlas y analizarlas por distintos ángulos, para ver con mejor luz las perspectivas.

Pero, antes de profundizar las inquietudes de lo que parece ser este nuevo paradigma retratado por Mayer-Schoenberger & Cukier (2013), cabría preguntar: ¿Cómo y por qué llegamos hasta aquí? Para entender y responder a estas preguntas, volveremos al final del siglo XX, cuando la era digital o, más precisamente, la emergencia de una sociedad informacional[8], comenzaba a ganar un cuerpo más sólido.

Estos mismos autores nos acercan a aquel tiempo, apuntando:

> (…) cuando Nicholas Negroponte, del laboratorio de medios del MIT, publicó su sobresaliente libro titulado *Ser digital*, uno de sus principales temas era el paso de los átomos a los bits. En la década de 1990, fundamentalmente nos dedicamos a digitalizar textos. Más reciente, puesto que la capacidad de almacenaje, la potencia de procesamiento y el ancho de banda han aumentado, lo hemos hecho también con otros formatos de contenido, como las imágenes, los videos y la música. (Mayer-Schoenberger & Cukier, 2013, p. 101)

En ese momento, las sociedades occidentales comenzaron a observar la popularización del ordenador, que salió de un contexto empresarial y científico para ser un objeto personal dentro de nuestras casas. En los Estados Unidos, donde ese movimiento se inició con más vigor, el salto fue de algunos miles de unidades en los comienzos de los ochenta, a más de cien millones de unidades vendidas a finales de la década de noventa (Dediu, 2012).

En ese contexto, Castells (1997) publicó una obra esencial, *La Era de la Información*, en la que ya apuntaba el surgimiento de una nueva estructura social, a partir del acelerado desarrollo de las tecnologías de la información que estaban redefiniendo las condiciones laborales junto con las organizaciones de las empresas, dentro de una nueva división internacional del trabajo.

[8] Para Castells (1997), es importante distinguir entre la sociedad de la información y la sociedad informacional. En la primera terminología, se trata del papel destacado que la información adquiere dentro de las sociedades en un sentido más amplio, como comunicación del conocimiento. Ya sociedad informacional "indica el atributo de una forma específica de organización social en la que la generación, el procesamiento y la transmisión de la información se convierten en las fuentes fundamentales de productividad y poder, debido a las nuevas condiciones tecnológicas que surgen en este periodo histórico" (p. 56).

El autor sostiene que el rápido avance de las tecnologías informacionales resultó en un nuevo modo de desarrollo, informacional, cuya fuente matriz de producción es la propia tecnología de generación de conocimiento, de procesamiento de información y de comunicación masiva de símbolos. Así, aunque el conocimiento e información son elementos cruciales en todos los modos de producción que hemos visto hasta el momento, lo que es característico del modo informacional de desarrollo es la retroalimentación de extracción de conocimientos sobre los propios conocimientos, como principal fuente de productividad (Castells, 1997, p. 35).

Entre los impactos que ello genera en los diversos medios de producción, llama la atención la transformación que incide en los procesos de comunicación, jugando un papel decisivo en esta nueva estructura.

> Un nuevo sistema de comunicación, que cada vez habla más un lenguaje digital universal, está integrando globalmente la producción y distribución de palabras, sonidos e imágenes de nuestra cultura y acomodándolas a los gustos de las identidades y temperamentos de los individuos. Las redes informáticas interactivas crecen de modo exponencial, creando nuevas formas y canales de comunicación, y dando forma a la vida, a la vez que ésta les da forma a ellas. (Castells, 1997, p.27)

Empero, si por un lado Castells (1997) arrojó luz sobre la profunda transformación de las sociedades, que culminaría en un nuevo sistema económico y tecnológico entendido como "capitalismo informacional" (p. 36), Negroponte se preocupó con la transformación en la naturaleza de esos nuevos sistemas de comunicación, hegemónicos en los días actuales.

En *Ser Digital*, Negroponte (1995) se empeña en elucidar algunos cambios causados en las tecnologías de información y en la sociedad como un todo por la revolución digital, argumentando que ya no estaríamos más en la era de la información relacionada con la era tradicional de los *mass media*. Así, estaríamos en la era de la 'postinformación', en la cual los límites de tiempo y espacio son minimizados y las relaciones sociales mediadas por esas NTICs son asincrónicas y sin fronteras, conectadas en redes (p. 06).

Dos fenómenos interdependientes marcan esta era: la digitalización de nuestro entorno y la creciente convergencia de las tecnologías digitales. No sería exagerado decir, que la emergencia del 'big data' es el proceso más radical de convergencia de ambos en la revolución digital.

La digitalización, cabe aclarar, no es el primer proceso tecnológico de nuestra era informática. El primer paso fue dado a partir de los rústicos sistemas computacionales, o sea, cuando se utilizaron máquinas automatizadas para efectuar cálculos de forma más rápida que por métodos anteriores[9], pero todavía utilizando soportes de dominios analógicos.

El paso siguiente, con el advenimiento de los transistores y, posteriormente, de los microprocesadores, desencadenó una transformación profunda en la naturaleza de la información, ya que su procesamiento se basa predominantemente en la tecnología digital, binaria. Pasaje, por lo tanto, de la información basada en átomos (analógica), para la información basada en bits (digital).

Toda esta transformación es resumida así por Negroponte (1995):

> La era industrial, básicamente una era de átomos, nos legó el concepto de la producción en masa, con economías basadas en una producción realizada con métodos uniformes y repetitivos, en cualquier espacio y tiempo dado. La era de la información nos mostró la misma economía de escala, pero con menor énfasis en el espacio y en el tiempo (p. 05) (…) La era de la postinformación tiene que ver con la relación a través del tiempo: máquinas que comprenden al individuo con el mismo grado de sutileza (o con un grado mayor aún) que esperamos de otro ser humano, incluyendo manías y hechos en todo aleatorios, buenos y malos, en la narrativa que constituye nuestras vidas. (p. 07)

De tal manera, el cambio en la naturaleza de la información, ahora plasmada en bits, permite la compatibilidad de diferentes fuentes de datos, ya que el factor delimitante de la naturaleza analógica de cada soporte desaparece. Así, se abre

[9] Estos sistemas fueron utilizados, por ejemplo, por Alan Turing durante la Segunda Guerra Mundial para descifrar los códigos encriptados del ejército alemán, como fue retratado en la película *The Imitation Game*.

camino para un fenómeno que ahora parece alcanzar en su máxima expresión: la convergencia digital.

Este punto es relevante para la epistemología de los datos masivos, ya que una de las grandes dimensiones que lo definen trata justamente de la variedad de datos cruzados que se computan dentro de uno o más bloques masivos de datos. El cambio de paradigma de la causalidad para la correlación de factores que caracteriza el 'big data' se da por el cruce de datos de diferentes fuentes y eso sólo es posible por compartir la misma base, para convergir cada vez más entre sí dentro de una la misma estructura tecnológica y social.

Entonces, si nuestra actual era informacional es sostenida por la presencia hegemónica de estructuras digitales, omnipresentes y convergentes, la propia relación que mantenemos con nuestro entorno, mediado por esas estructuras, deberá ser pautada por una nueva lógica. Este es el punto de vista sostenido por Manovich (2002) en *El lenguaje de los nuevos medios* y otros ensayos posteriores.

En este sentido, el autor apunta que los *softwares* (y también los algoritmos) vienen sustituyendo un gran número de tecnologías analógicas, mecánicas y electrónicas utilizadas en el siglo XX para la creación, almacenamiento, distribución y acceso a las informaciones que nos rodean. El *software*, entendido como la herramienta primordial de este universo computacional digital, organiza nuestra interfaz con el mundo, con nuestro entorno, con nuestra memoria y nuestra imaginación, como Manovich (2002) explica:

> Si, en la física, el mundo está hecho de átomos y, en la genética, está hecho de genes, la programación computacional encapsula el mundo de acuerdo con su propia lógica. El mundo se reduce a dos tipos de objetos de software, que son complementarios entre sí: estructura de datos y algoritmos. Cualquier proceso o tarea se reduce a un algoritmo, una secuencia final de operaciones simples que un equipo puede realizar para alcanzar una tarea dada. Y cualquier objeto en el mundo -sea la población de una ciudad, o la temperatura a través del curso de un siglo, o una silla, o un cerebro humano- es modelado como una estructura de datos, es decir, datos organizados de un modo particular para permitir la búsqueda eficiente y su recuperación. (p. 05)

Ahora bien, si cada nueva tecnología de información genera prácticas y lenguajes propios, la decisión de establecer bases epistemológicas para el 'big data' gana relevancia. En ese sentido, durante los años noventa ya se trabajaba con grandes volúmenes de datos, pero su alcance estaba restringido a algunas empresas o grupos de investigadores.

Aunque su surgimiento se remonta al último cuarto del siglo pasado, los macro datos todavía son considerados algo nuevo, tanto en su uso como en su concepción, ya que se está normalizando tan sólo en los últimos años. Debido a eso, todavía no hay un concepto exacto que defina lo que es, de hecho, el 'big data'.

Sin embargo, hay un punto de partida en común de diversos trabajos que consideran que el concepto de datos masivos fue solidificado, por primera vez, por Laney (2001), en una investigación que analizaba los desafíos enfrentados por las empresas en la gestión de los datos. Por esta vía, se definieron tres importantes dimensiones asociadas a los datos que debían tenerse en cuenta para determinar si eran masivos: volumen, velocidad y variedad.

Como veremos más detenidamente en el final del siguiente capítulo, en los años posteriores, en la medida en que se intensificaron los usos de estas dimensiones en investigaciones y trabajos dentro de los diferentes campos en los que era posible su aplicación, fueron añadidas otras más que apuntan para ampliar su utilización.

A continuación, desarrollaremos la primera parte del trabajo con foco en entender las bases de lo que se ha convenido llamar de 'big data', en la que hablaremos de los conceptos y elementos estructurantes que forman parte de este fenómeno, además de revisar la conceptualización en curso de las *dimensiones Vs* que surgieron a lo largo de la expansión del uso por parte de investigadores, académicos y profesionales.

3. Elementos estructurantes del 'big data'

El 'big data', ante todo, se viene revelando como un fenómeno amplio, no sólo tecnológico, que se expande por nuestra cotidianidad y que transforma nuestro entorno, como nos recuerdan Mayer-Schoenberger & Cukier (2013):

> Los 'big data', los datos masivos, se refieren a cosas que se pueden hacer a gran escala, pero no a una escala inferior, para extraer nuevas percepciones o crear nuevas formas de valor, de tal forma que transforman los mercados, las organizaciones, las relaciones entre los ciudadanos y los gobiernos, etcétera. (p. 17)

Por lo tanto, considerando el 'big data' como un conjunto de elementos, técnicas y prácticas que permiten el salto cualitativo en la forma como se producen, articulan y analizan los datos, debemos tener en cuenta que algunos de estos elementos juegan un papel principal en este escenario: la datificación, la minería de datos y los algoritmos.

Para explicar la estructura de este fenómeno, el presente trabajo propone el camino inverso de lo que se suele encontrar en otros textos, en los que primero se definen los datos masivos a partir de las *dimensiones Vs* y luego se trata de los puntos adyacentes a ese tema.

En esta configuración, comenzamos con lo que llamo elementos estructurantes del fenómeno, sus orígenes y lógicas, para así pasar a la explicación de cada *dimensión V* tomando en cuenta la perspectiva de los usos y sus definiciones académicas.

3.1 La datificación

En la introducción de esta investigación, retratamos brevemente cómo a lo largo de la historia fuimos desarrollando sistemas y prácticas que nos permitieron medir la realidad, registrar y poder acceder a ella más tarde. Apuntamos también que determinados eventos históricos relacionados con el registro de información

fueron esenciales para dicho desarrollo, entre ellos el surgimiento de la escritura y la invención de la prensa de Gutenberg.

No hay duda de que el progreso humano está íntimamente ligado con nuestra capacidad de registro y desarrollo de análisis de la información registrada. Es en este punto que Mayer-Schoenberger & Cukier (2013) son categóricos en afirmar que "la capacidad de archivar información es una de las líneas que separan las sociedades primitivas de las avanzadas" (p. 101).

Sin embargo, otro punto importante sostenido por los autores es que no sólo los actos de medir y registrar fueron necesarios para ese progreso, sino también el hecho de que esos registros pudieran ser efectivamente analizados y organizados para cálculo matemático. Así, hay que destacar que la diseminación y popularización de los dígitos arábigos en las civilizaciones europeas, a partir del siglo XII, representó un gran cambio cualitativo y proporcionó el despegue de las matemáticas, que serían la base para la revolución científica siglos más tarde. "Las matemáticas dieron un nuevo sentido a los datos: ahora éstos podían ser analizados, no solo registrados y recuperados." (Mayer-Schoenberger & Cukier, 2013, p. 103)

Así, a pesar de que durante nuestra era analógica hayamos de hecho desarrollado diversos sistemas y lenguajes que nos ayudaron en ese incansable trabajo de interpretación de la realidad, fue en estas últimas décadas que pudimos presenciar un salto cualitativo en la forma en que medimos eso. Como ya se ha explicado, el creciente poder de procesamiento, junto con el abaratamiento de los costos de almacenamiento y la popularización de dispositivos y herramientas de captación de datos, permitió que hoy podamos extraer y guardar una cantidad exponencial de información de nuestro entorno natural o social.

Así, en la visión moderna que supone este trabajo de cuantificación y almacenamiento de datos de cualquier acción humana o de los fenómenos naturales, se denomina como datificación. De acuerdo con Mayer-Schoenberger & Cukier (2013), podemos entender este proceso como la transformación de las

acciones humanas (sociales) o de la naturaleza en datos digitales cuantificados, que podrán ser monitoreados y analizados a partir de una óptica de valoración de éstos, o sea, que resulte en la generación de un nuevo tipo de información o que ayude en la elaboración de nuevas prácticas para el conocimiento científico.

De esta forma, muchos científicos e investigadores encuentran en esta nueva perspectiva una otra manera de entender el mundo que nos rodea. Economistas utilizan grandes cantidades de datos para predecir tendencias en mercados financieros (Sabouni, 2018); astrónomos, para entender y calcular eclipses pasados (Stanley, 2013); científicos sociales y de humanidades, como forma inédita de unir metodologías cuantitativas y cualitativas (Manovich, 2011).

Esa última posibilidad se plasma en la idea de que grandes cantidades de datos, por lo tanto, aspectos de matriz cuantitativa, permiten entender con más profundidad los *qualia* de los fenómenos sociales, lo que se refiere a aspectos cualitativos. O sea, ciertas calidades sólo son accesibles cuando se disponen de muchos o de todos los datos disponibles de determinado fenómeno.

Los ejemplos actuales permean casi todas las disciplinas académicas y áreas del conocimiento científico. Sin embargo, hay que tener en cuenta que "toda disciplina e institución tiene sus propias normas y estándares para la manipulación de los datos, así como muchos campos tienen sus propias metodologías aceptadas y sus estructuras de prácticas" (Gitelman, 2013, p. 03).

En nuestro foco de investigación, que es el impacto en los estudios sobre comunicación mediática, aparecen cuestiones, muchas veces sensibles, cuando las fuentes de datos involucran prácticas sociales donde individuos y sus acciones son la base de la datificación.

En la era del 'big data', por medio del acelerado proceso de datificación permitido por los actuales sistemas y dispositivos digitales, empresas e instituciones pasan a cuantificar aspectos sociales que antes eran imposibles, como el deseo (Google), las

amistades (Facebook), relaciones profesionales (Linkedin), gustos musicales (Spotify), consumo de contenido (Netflix), acceso a la cultura (Youtube), etcétera.

Además, la universalización del *smartphone* permitió que cada individuo fuera una fuente constante de datos y metadatos; numerosas empresas recopilan esa información a través de aplicaciones y servicios instalados en esos dispositivos, como el uso de GPS. Este volumen de información también ha sido de importante valor para los académicos por la riqueza de detalles que antes no era posible obtener.

Así, los usuarios proporcionan información personal a las compañías y pueden utilizar estos servicios, casi siempre de forma gratuita. Este modelo de intercambio de (meta) datos por el acceso a servicios de todo tipo de finalidad se ha convertido en el estándar en la actual estructura de internet y es importante pues, por un lado, modela las nuevas formas de uso de los medios de comunicación y, por otro, genera un lastre de actividades en esas plataformas que posteriormente son útiles para el trabajo de minería de datos con fines comerciales, políticos y científicos.

> Su uso como moneda para pagar por los servicios *online* ha convertido a los metadatos en una especie de activo invisible, procesados, la mayoría de las veces, fuera de su contexto original y sin que la gente tenga conciencia. Las compañías de medios sociales monetizan los metadatos al reprocesarlos y venderlos a anunciantes o compañías de datos. (Van Dijck, 2017, p. 45)

Es que el valor de capturar y almacenar datos de cualquier cantidad y naturaleza reside en el hecho de que su utilización no genera desgaste y depreciación, como en el caso de los bienes materiales. Por el contrario, la posibilidad de ser usados más allá de su finalidad original agrega más valor en la medida en que pasan a ser analizados y recombinados con otros datos, generando usos secundarios *ad infinitum* y ampliando el universo de aplicaciones y posibilidades para todo tipo de trabajo, inclusive académico.

Esto lleva al surgimiento de una cadena extensa de repositorios de datos que pueden ser utilizados para fines comerciales, científicos o gubernamentales[10]. En el caso de cómo se comparte, y el acceso a estos repositorios para fines científicos y académicos, hay un precioso debate en curso sobre las formas en que esto debe ser conducido, que trataremos más adelante.

En cierto sentido, esto amplía mucho las posibilidades de investigación en ciencias sociales y humanidades, ya que la datificación del conjunto de prácticas sociales, antes imposibles, permite a los investigadores una base cuantificable y, en cierto punto, objetiva, de un determinado enfoque. Esta posición, defendida por Gitelman (2013), propone "que los estudiantes y académicos en humanidades se preocupen por los datos, en términos generales, en la medida en que sus objetos de estudio han sido asumidos y discernidos. ¿No harían las preguntas presuponer o delimitar sus respuestas en algún grado?" (p. 03)

La propia existencia de un dato ya determina, en cierto punto, la forma y el tipo de información que podrá extraerse de él. En ese sentido, esto levanta una discusión antigua y profunda sobre la manera en que interpretamos nuestra realidad, o mejor, de cómo tenemos acceso a los fenómenos *de lo real*. Soy consciente de la importancia de estas discusiones y entiendo que aquí no es el momento de tocar a fondo estas cuestiones, con las que podrán ser exploradas en otro momento.

Así, la provocación de Gitelman es importante. En cierto nivel, la datificación puede suponer la interpretación que se dará a estos datos. Los datos no existirían por sí; primero, deben ser *imaginados*, para ser captados e interpretados; aún, la *imaginación* de los datos ya implica una base interpretativa (Manovich, 2011).

Aquí, se apunta que la datificación de las prácticas sociales aporta un carácter de objetividad y neutralidad; no siendo un proceso consciente de los usuarios, no estaría impregnado de juicios y valores humanos preexistentes (Anderson, 2008).

[10] Algunos ejemplos de esto son los bancos de datos ofrecidos por las Naciones Unidas, el Banco Mundial, Google y otras fuentes gubernamentales, como el Data.gov de los EE.UU. (Ching, 2017).

En otro sentido, como veremos más adelante, una parte de los autores críticos sostiene que ese proceso de datificación es la raíz de la mayoría de los cuestionamientos éticos, aliados a prácticas dudosas. Por más que el lastre de actividades que las personas dejan en el uso de las plataformas digitales sea valioso como fuente de estudios, hay problemas éticos importantes y no resueltos que sobrevuelan el tema.

De facto, la datificación, además de un medio para acceder, entender y monitorear el comportamiento social, "se está convirtiendo en un principio central, no sólo entre los adeptos de la tecnología, sino también entre los académicos, que la ven como una revolucionaria oportunidad de investigación para comprender el comportamiento humano" (Van Dijck, 2017, p. 41).

3.2 La minería de datos

En el proceso de datificación, cualquier dato generado, captado y almacenado tiene el potencial de convertirse en una materia prima preciosa. Por este motivo, no faltan metáforas para referirse al acto de procesamiento o de transformación por la cual esa *commodity* pasa, del estado bruto, al de valor agregado.

La idea de datos y metadatos como recursos brutos que pueden ser procesados para muchas finalidades encaja perfectamente en el concepto más popularizado y recurrente de datos como materia prima. Los desechos, el reciclaje de información o los subproductos digitales son términos frecuentemente usados para referirse a los (meta) datos y su utilización.

La extracción, reciclaje, minería, exploración, enriquecimiento, procesamiento y muchas otras acciones que se aplican en la manipulación de datos nos muestran cómo estos se van convirtiendo en el principal recurso de investigación para algunas disciplinas académicas. Así, este conjunto de procesos y prácticas en la manipulación de los bloques de datos puede ser asociado al concepto de minería de datos, ya que

> (...) la minería es un término vívido que caracteriza el proceso de encontrar un pequeño conjunto de pepitas preciosas de una gran cantidad de materia prima (...) la minería de datos, entonces, es el proceso de descubrir patrones interesantes y conocimiento a partir de grandes cantidades de dados. Las fuentes pueden incluir bases de datos, *data warehouses*, Internet, otros repositorios de información o datos que se transmiten a un sistema dinámicamente. (Han, Kamber & Pei, 2012, p. 08)

Este conjunto de técnicas puede ser entendido como etapas sucesivas que se adoptan para la manipulación de estos datos masivos brutos, entre ellos: limpieza, integración, selección, agrupación y transformación de datos; descubrimiento, visualización y análisis de patrones; por último, generación y consolidación del conocimiento.

En el caso de las investigaciones académicas, todo este conjunto de técnicas expanden las metodologías posibles para realizar dichas investigaciones, apuntando a un salto cualitativo en la obtención de resultados. "La metáfora de la minería de datos se basa en una racionalidad que permite a los emprendedores, a los académicos y las agencias estatales en la búsqueda de un nuevo paradigma social-científico" (Van Dijck, 2017, p. 45).

Veremos en el próximo apartado cuáles son las bases de ese nuevo paradigma. Antes, sin embargo, vale destacar que la práctica de minería de datos posibilita, por un lado, descubrir y analizar patrones y correlaciones ocultas que pueden contener un valor intrínseco, aún no revelado (Mayer-Schoenberger & Cukier, 2013). Por otro lado, permite también aplicar usos a los cuales los conjuntos de datos iniciales no habían sido preparados para ello.

En este punto, hay que tomar un cierto cuidado en cómo se extraen nuevas correlaciones de datos obtenidos para determinado uso, ya que un conjunto de datos se refiere a un "corpus de conocimiento en particular" (Arcila-Calderón, Barbosa-Caro & Cabezuelo-Lorenzo, 2016, p. 627). Tirarlos del contexto original puede llevar a conclusiones y correlaciones que distorsionan el resultado de una investigación o aplicación práctica.

Pero, de manera acertada, Gitelman (2013) afirma que precisamente porque los datos “son un hecho, pueden tomarse para construir un modelo suficiente en sí mismo: ciertos datos llevan a ciertas conclusiones que pueden ser probadas o argumentadas a seguir. Teniendo en cuenta otros datos, uno llegaría a diferentes argumentos y conclusiones” (p. 07).

La misma lógica podemos utilizar para entender los cuidados en la aplicación de las técnicas de minería. De un mismo conjunto de datos, podemos aplicar una técnica y extraer ciertos argumentos; cambiando la técnica, se llega a otros argumentos y posteriores aplicaciones a partir de los iniciales.

Para dar un ejemplo de reutilización de datos, podemos acreditar la gran efectividad de la diseminación de *fake news* en los recientes procesos del Brexit, o en la elección de Donald Trump, a las técnicas de minería de datos. Después de la recolección y tratamiento de los (meta) datos personales de millones de usuarios de Facebook para una investigación académica sobre perfiles psicológicos de estos usuarios, hubo la aplicación de minería de datos para otros usos, en tales casos, para campañas políticas.

Así, por medio del descubrimiento, la visualización y el análisis de patrones del comportamiento de estos usuarios, la empresa Cambridge Analytica pudo agrupar a los individuos en distintos *clusters* y así, vender ese conocimiento a los partidos políticos que distribuían anuncios efectivos para esos grupos de usuarios.

Es la misma lógica que está por detrás de los servicios de publicidad y personalización de contenidos en las plataformas digitales: identificación, selección y agrupamiento de individuos; descubrimiento, visualización y análisis de patrones de comportamiento; por último, la generación y consolidación del conocimiento que revierte en nuevos negocios.

En resumen, cualquier sistema o dispositivo utilizado para datificar una acción o un conjunto de eventos dentro del contexto de 'big data', generará un gran bloque masivo de datos que quedará almacenado para que sea utilizado *a posteriori*. Este

bloque no nos revela en un primer momento su tramado de correlaciones internas que se supone existir, ni mucho menos el camino a seguir para la extracción de información relevante y útil. Los investigadores deben entonces aplicar las técnicas de minería de datos para extraer valor.

Para eso, es necesario pulir este bloque (minería) para encontrar sus correlaciones internas; en un segundo momento, crear sentido en esas correlaciones para que sea posible generar información, para llevar a la formación de un nuevo conocimiento.

Sin embargo, si por un lado la minería es un conjunto de métodos y, por otro, sabemos que esta operación no se da a nivel humano por el tamaño del bloque, ¿cómo entonces se opera la extracción de la información? Para ello, es necesario que el propio sistema computacional sea instruido en el sentido de hacer posible la manipulación. Por lo tanto, en esta etapa serán empleados los algoritmos para modelar y extraer la información de este bloque masivo de datos.

3.3 Los algoritmos

Los algoritmos son, actualmente, los motores de la mayoría de los sistemas y servicios en Internet. Sin embargo, es importante destacar que su surgimiento antecede a la invención de los ordenadores modernos. El surgimiento de este término está ligado al desarrollo de un conjunto de técnicas matemáticas antiguas, consideradas manualmente. "Algorismus" era, en su origen, el proceso de calcular los dígitos hindú-arábigos[11].

Muchas definiciones se pueden encontrar para describir lo que es un algoritmo moderno. En la concepción básica, se puede decir que es un método para resolver un problema específico (Finn, 2017). Sin embargo, dentro de la lógica computacional, los algoritmos pueden ser entendidos como procesos y fórmulas de computadora que transforman las preguntas en respuestas (Google, 2011).

[11] A pesar de que su origen se remite a un antiguo matemático persa, fue sólo en 1971 que fue utilizado por primera vez en una aplicación práctica por parte del economista Leontief (Fernández, 2013).

En el contexto de las operaciones con macro datos, podemos decir que el algoritmo es una serie lógica de instrucciones matemáticas que se utiliza para resolver un determinado problema dentro de un período de tiempo limitado. En este caso, la efectividad en el uso de un determinado algoritmo reside en el equilibrio y efectividad en procesar datos teniendo en cuenta las dimensiones de volumen, variedad y velocidad como veremos a seguir.

En muchas investigaciones con grandes volúmenes de datos textuales, los algoritmos ayudan a clasificar, organizar y separar en conjuntos para posterior análisis. "Esto es muy útil para la clasificación automática de noticias. Lo mismo sucede para la generación de predicciones de comportamiento de la opinión pública a partir del uso de perfiles creados con encuestas históricas" (Arcila-Calderón et al., 2016, p. 629).

Sin embargo, así como las técnicas de minería de datos no son epistemológicamente buenas o malas, lo mismo vale para la concepción sobre el uso de los algoritmos.

> A diferencia de las cosas concretas en el mundo, como las partículas u organismos, los algoritmos no se pueden observar como objetos de estudios en sistemas experimentales directamente. Solo se les pueden evaluar en su funcionamiento como componente de un conjunto de tareas computacionales; por su propia cuenta, son inertes. (Lowrie, 2017, p. 01)

Así como en el caso de la datificación, investigadores han señalado que la forma de concepción y el uso que se hace de los algoritmos debería atraer más atención de la comunidad académica, ya que a despecho de la idea de neutralidad objetiva que se les puede atribuir, existe una lógica ideológica por detrás de su construcción, que a menudo induce o interfiere en los resultados esperados (Mager, 2014). Incluso, el uso de algoritmos en las investigaciones académicas se ha expandido de tal forma que permean prácticamente todas las disciplinas de las diversas áreas científicas.

Es por este camino que Manovich viene desarrollando su línea de investigación sobre algoritmos y *softwares* a través de una perspectiva cultural. El *Cultural Analytcs Lab* (anteriormente *Software Studies Initiative*), dónde es director, indica que entender el uso de los algoritmos y las informaciones que ellos generan es fundamental para entender las actuales implicaciones en la investigación académica.

> El software es aún invisible para la gran mayoría de los académicos, artistas y profesionales de la cultura interesados en TI (Tecnología de la Información), sus efectos culturales y sociales. Pero si seguimos limitando las discusiones críticas a las nociones de "ciber, digital, nuevos medios" o "Internet", estaremos en peligro de lidiar sólo con los efectos y no con las causas. Se corre el riesgo de quedarse observando solamente los resultados que aparecen en la pantalla del ordenador en lugar de analizar los programas y las culturas sociales que producen esos efectos. (Software Studies Initiative, 2008)

Así, una vez repasados estos elementos (datificación, minería y algoritmos) que considero importantes para entender el fenómeno de los datos masivos, repasaremos ahora las *dimensiones Vs* que los investigadores, principalmente de las ciencias computacionales, han utilizado para definir lo que puede ser considerado como 'big data'. Después de eso, partiremos para la segunda sesión en la cual pasaremos por los principales puntos que se muestran en el horizonte: la configuración de los elementos del fenómeno 'big data' sugiere un cambio de paradigma en la conducción de las investigaciones académicas.

3.4 Las *dimensiones Vs*

Sumado a los elementos que componen la estructura conceptual de lo que se entiende como datos masivos, hay todavía una corriente que se propone definir a partir de dimensiones asociadas tanto a su estructura, como a los problemas donde la tecnología actúa a partir de sus propias potencialidades.

El uso de las *dimensiones Vs* para definir y entender el concepto de 'big data' se inicia con Laney (2001) en una investigación sobre los desafíos que las empresas enfrentan para trabajar con bancos de datos cuya capacidad y tamaño exigen

formas innovadoras de procesamiento de información para una mejor percepción y toma de decisión.

En ese momento, Laney apuntó que la popularización del comercio electrónico exigía que las empresas solucionasen los desafíos que los grandes volúmenes de datos imponían a partir de tres dimensiones: volumen, velocidad y variedad (Laney, 2001).

Con el pasar de los años, otras dimensiones fueron propuestas, de acuerdo con la evolución de los sistemas computacionales y de los usos y nuevos desafíos que las empresas e investigadores encontraron en la manipulación de grandes volúmenes de datos: veracidad, valor, volatilidad, visualización, validez, viralidad, etcétera (Lee, 2017).

El *National Institute of Standarts and Technology* (NIST) de EE.UU., un importante órgano de métricas de ese país, considera en su sección de términos y definiciones las siguientes ocho dimensiones: volumen, velocidad, variedad, validez, valor, variabilidad, veracidad y volatilidad (NIST, 2015).

En este trabajo, como forma de entender cómo esas dimensiones impactan en las investigaciones académicas sobre el tema, atribuimos diez dimensiones que ayudan en la conceptualización de ese fenómeno: volumen, velocidad, variedad, veracidad, valor, volatilidad, visualización, viralidad, variabilidad y viscosidad. Cada dimensión corresponde a un enfoque:

Volumen: La primera y más importante característica del 'big data'. Supone el uso de un volumen considerablemente grande de datos[12], no manejables a nivel humano y que sólo pueden ser procesados y minados por lenguajes computacionales específicos, los algoritmos.

[12] No hay consenso de cuán grande debe ser este volumen, pero Anderson (2008) dice que eso sólo es posible con volúmenes de datos en la casa de los *terabytes* o *petabytes*.

Velocidad: Se trata entonces del flujo temporal continuo que posibilita que los datos sean generados, almacenados, recuperados y procesados (etapas de accionamiento y minería) para la generación de nuevas informaciones en tiempo real o en un tiempo determinado no demasiado largo.

Variedad: Se refiere a la naturaleza de la fuente de datos[13], a la naturaleza de los propios datos y al estado bruto en que se presentan. Este último punto trata de cómo estos datos están disponibles para su manipulación, pudiendo ser presentados como datos estructurados, no estructurados y semiestructurados.

Veracidad: Esta dimensión de los datos masivos trata de la veracidad de las fuentes y es un importante punto para la reflexión del concepto de verdad que etiquetamos a esas fuentes. Puede ser afectada por la imperfección, inexactitud, inconsistencia o subjetividad de la naturaleza de los datos.

Valor: Indica que la manipulación de grandes volúmenes de datos sólo tiene sentido si podemos extraer nuevos conocimientos, identificar oportunidades, resolver problemas, reforzar o renegar hipótesis. En suma, generar valor positivo al conocimiento humano.

Volatilidad: Esta dimensión apunta al tiempo necesario en que se debe almacenar determinados tipos de datos. Empero, debido a la velocidad y el volumen de 'big data', su volatilidad debe ser llevada en consideración en determinadas condiciones; si no, podrí llevar a un desequilibrio entre el volumen y la velocidad, impidiendo el procesamiento de bloques voluminosos.

Visualización: Se refiere a la forma cómo las herramientas computacionales permiten visualizar los datos. Muchos de los instrumentos actuales enfrentan desafíos técnicos debido a las limitaciones de procesamiento y la baja escalabilidad. También es un punto sensible, pues diferentes formas de

[13] Con respecto a las fuentes de datos, se pueden dividir de acuerdo con la clasificación de *Open Knowledge International*: 1) Culturales; 2) Científicas; 3) Financieras; 4) Estadísticas; 5) Meteorológicas; 6) Ambientales; 7) Medios de transporte (OKI)

visualización llevan a diferentes formas de interpretación del mismo conjunto de datos.

Viralidad: Esta perspectiva está relacionada con la preocupación por estructurar y almacenar los datos de investigación de forma que otros estudios e investigadores tengan acceso y facilidad en compartimentos. Cuanto mejor organizado y más abierto los datos, más viral será.

Viscosidad: Esta dimensión guarda relación con otras dos dimensiones: variedad y velocidad. Hace referencia al nivel de dificultad para manipular ciertos datos en determinadas condiciones; esta resistencia puede ocurrir cuando se trabaja con diferentes fuentes de datos con demasiada incompatibilidad entre sí (variedad) o en la fricción encontrada en el flujo de integración y procesamiento necesarios para transformar los datos en *insights* (velocidad).

Variabilidad: La cuestión de la variabilidad puede referirse a diversas situaciones que dependen del contexto. Se puede entender como posible inestabilidad cuando los datos manipulados presentan muchas anomalías, como la variación del flujo cuando se aumenta el volumen, o puede referirse a la variación que ocurre cuando se manipula un mismo bloque de datos por medio de metodologías o *softwares* distintos, impactando directamente en la calidad de la información extraída.

De todas maneras, más importante que señalar cuáles dimensiones están involucradas en la concepción del 'big data' es la forma en que interactúan cuando se combinan en su procesamiento y en las técnicas de minería. En este sentido, Lee (2017) propone una visión integrada de esas dimensiones, asumiendo que en un proceso dinámico ellas se alteran e impactan unas a las otras, en una conexión dependiente. Esta relación puede ser mejor expresada por la Figura 1.

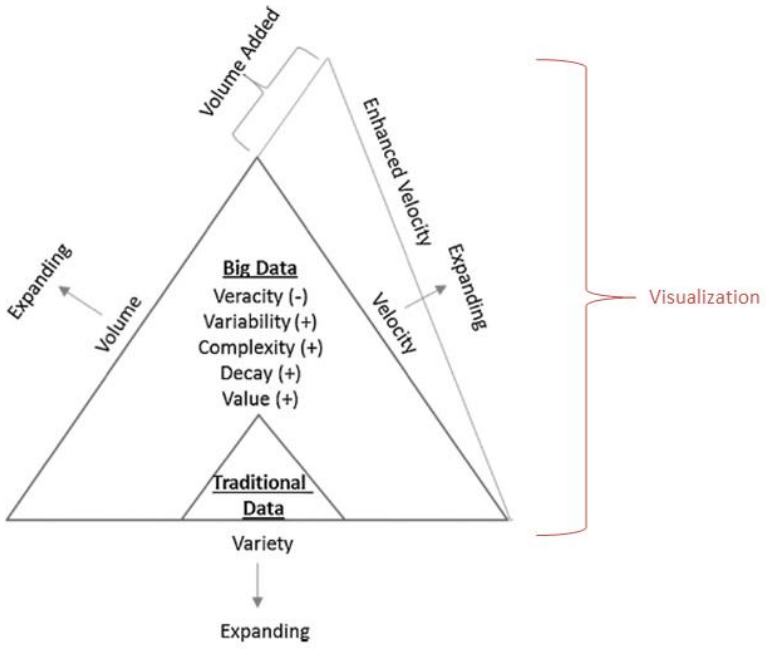

Figura 1. An integrated view of big data (Lee, 2017)

A partir de la unidad básica de las tres dimensiones (volumen, velocidad y variedad), hay una interdependencia de cambios cuando se altera una de esas dimensiones. Lee (2017) explica que

> los tres bordes de la vista integrada de los 'big data' representan tres dimensiones de los datos masivos: volumen, velocidad y variedad. Dentro del triángulo están las cinco dimensiones del 'big data' que se ven afectadas por el crecimiento de las tres dimensiones triangulares: veracidad, variabilidad, viscosidad (*complexity*), volatilidad (*decay*) y valor. (p. 295)

En este esquema, hay una contraposición a lo que no puede ser considerado como 'big data' que Lee (2017) llama 'tradicional data' (o 'small data', como fue antes identificado). Vemos que, en ese escenario, no hay interdependencia de las dimensiones básicas (volumen, velocidad y variedad) y la variación de éstas no impacta en la dinámica de las otras (que quedan fuera del triángulo).

En nuestro caso, optamos por agregar un elemento más, externo al diagrama, que indica la importancia de la visualización de todo ese esquema dentro de un software o como resultado de la acción de un algoritmo. Esto se muestra relevante, pues la forma que esta dinámica es visualizada impacta en la extracción de las

correlaciones en muchas aplicaciones e investigaciones académicas, principalmente en estudios culturales y análisis de textos.

En este trabajo, se considera que las *dimensiones Vs* ayudan a delimitar el concepto, a pesar de no haber consenso entre los investigadores en cómo cuantificar o precisar, dentro de cada dimensión, patrones precisos que encuadren un determinado universo de datos en 'big data'.

4. Un nuevo paradigma científico-social y sus características

Una vez que esperamos haber aclarado algunos elementos que considero importante en el entendimiento y la concepción del fenómeno 'big data', tales como la datificación y la minería de datos, además de señalar las principales dimensiones que inciden en la dinámica de la manipulación de datos masivos, pasamos a la etapa de aclarar los puntos que indican un cambio de paradigma en curso.

En ese sentido, pautado en las ideas de Kuhn (1962) de paradigma científico, desarrolladas en su libro *La estructura de las revoluciones científicas*, hay que comprender el "papel desempeñado por el progreso tecnológico o por las condiciones externas, sociales, económicas e intelectuales, en el desarrollo de las ciencias" (p. 15).

En la concepción de Kuhn (1962), un paradigma científico es constituido por lo que se comparte dentro de una comunidad académica a partir de sus prácticas científicas, incluyendo conjuntamente leyes, teorías, aplicaciones, instrumentación, creencias y metodologías que proporcionan modelos de problemas y soluciones coherentes a esa comunidad para la investigación científica (p. 13).

Así, durante la etapa de una revolución científica, hay un proceso de transformación en el conjunto de esos elementos que pueden determinar un nuevo paradigma dentro de esa comunidad, cambiando lo que se considera como problemas admisibles y soluciones legítimas para esos problemas. O, como prefiere Kuhn "como una transformación del mundo en que se llevaba a cabo el trabajo científico" (Kuhn, 1962, p. 28).

A pesar de que se ha señalado que estos cambios van solidificando un nuevo paradigma científico, como indica algunos autores citados (Anderson, 2008; Mayer-Schoenberger & Cukier, 2013), este trabajo no se propone investigar la cuestión en profundidad, sino indicar algunos cambios permitidos por el 'big data' que parecen apuntar en esa dirección, con la necesidad de más estudios para ello.

Los puntos siguientes giran en torno a los cambios ocurridos con el salto cuantitativo y cualitativo posibilitados por el uso del 'big data' y cómo afecta nuestra relación con los datos; consecuentemente, cómo extraemos conocimiento de ello.

4.1 De la escasez a la abundancia

En vista que la datificación ya era recurrente durante toda la historia de nuestra sociedad, el cambio profundo en ese sentido fue permitido por la hegemonía de la tecnología digital que elevó exponencialmente el volumen de información. “Pese a que los conceptos de *revolución de la información* y *era digital* existen desde la década de 1960, apenas acaban de convertirse en realidad de acuerdo con ciertas medidas” (Mayer-Schoenberger & Cukier, 2013, p. 20).

Siguiendo esta pista, se estima que pasamos de 2,6 exabytes de datos almacenados en 1986 a más de 300 exabytes en 2007 (Hilbert & López, 2012). En 2017 la estimación era que estábamos cerca del nivel de los 20.000 exabytes (Gantz & Reinsel, 2012), por lo tanto un salto de 67 veces en sólo en los últimos 10 años.

Interesante notar que en 1986 prácticamente no había volumen de datos digitales significativos almacenados, siendo que casi el 100% eran analógicos. Sin embargo, en 2007 los datos digitales respondían por el 93% del total (Hilbert & López, 2012). Se estima que en los próximos años estaremos cerca del 100% de los datos almacenados en dispositivos digitales, con la parcela analógica también replicada digitalmente en ese total (Gantz & Reinsel, 2012).

Por eso, no hace mucho tiempo, se tenía más control y sabíamos las formas y los métodos para colectas de datos personales. Estaban restringidos a procedimientos como censo demográfico, emisión de documentos de identificación, encuestas de opinión, medición de audiencias, etc.

Hoy, estamos constantemente en contacto y bajo el control de todo tipo de máquina. Por medio de aparatos "amigables" para los usuarios y con interfaces inteligentes, estamos cada vez más rodeados de dispositivos de datificación para todo y cualquier propósito (Gitelman, 2013).

En un día común y viviendo en una gran ciudad, una persona emerge en un entorno cohabitado por todo tipo de dispositivos: tarjetas de crédito, tarjeta de la empresa o escuela, códigos de barras en productos, cámaras de vigilancia, tarjetas magnéticas, tarjetas de transporte, etcétera. Además, tenemos nuestro propio dispositivo personal de datificación, el teléfono móvil. Con él dejamos nuestras huellas digitales, ya sea navegando por Internet, interactuando en las redes sociales, escuchando música, leyendo e-mails, utilizando aplicaciones o simplemente dejando el GPS activo.

En 2016, se estimaba que por día se creaban datos suficientes para llenar 2,5 millones de discos duros (IBM, 2016). Considerando el total de habitantes de la Tierra[14], podemos decir que un solo individuo es responsable, en promedio, por generar algo en torno a 30 megabytes de datos diarios.

A pesar de que el momento actual de hiperdigitalización y de datificación constante impacta en todos nuestros aspectos cotidianos, es en la dimensión de los datos personales que vemos generar promesas de nuevas posibilidades de entender las dinámicas sociales, pero, por otro lado, también la preocupación en torno a la privacidad y seguridad.

El mercado de datos personales es cada vez más relevante en la actual sociedad de la información y puede entenderse como los flujos económicos basados en la compra y venta de la información relativa a las personas.

> Generados por las identidades y comportamientos, por los individuos y sus acciones en redes digitales, los datos personales son la moneda que se paga por el uso gratuito de plataformas, sitios y servicios en línea. Los datos personales se han

[14] En 2016 se estimaba que éramos 7.444 millones de habitantes en nuestro planeta (Banco Mundial, 2016).

> convertido en un importante bien económico. (Avelino, Silveira & Souza, 2016, p. 220)

Los recientes casos de vigilancia de la *National Security Agency* (NSA)[15] o de la captura ilegal de datos de millones de usuarios de Facebook por la empresa Cambridge Analytica evidencian que la datificación, almacenamiento y reutilización de los datos personales pueden ser utilizados para control y manipulación social, además de colocar en el centro del debate la importancia del control de datos por parte de los propios usuarios.

Un importante avance, en este momento, es la reglamentación de *la General Data Protection Regulation* (GDPR) por la Unión Europea, que trata de establecer un conjunto de normas para ser adoptadas por todas las empresas que operan utilizando datos de ciudadanos europeos. En este escenario, las nuevas prácticas sugeridas por la reglamentación intentan velar por la libertad y los derechos de los usuarios (GDPR, 2016).

En la misma sintonía, dentro del debate académico, parte de los investigadores apunta a los temas de la privacidad, de la datificación cotidiana de las personas y de la transformación de los datos personales en monedas de cambio, como los problemas más sensibles desde el punto de vista ético y práctico.

Van Dijck (2017), por ejemplo, nos recuerda que la responsabilidad de mantener la credibilidad del ecosistema de datos "también es de los académicos. El desenfrenado entusiasmo de muchos investigadores por la datificación como un paradigma neutro, reflejando una creencia en una comprensión objetiva, cuantificada de lo social, debe ser analizada de manera más rigurosa" (p. 54).

Las personas no están necesariamente consciente de todos los múltiples usos, beneficios y otras ganancias que vienen de la información que proporcionan a

[15] En 2013, Edward Snowden reveló las prácticas rutinarias de vigilancia de la NSA con las que él y muchos otros funcionarios de ese órgano vigilaban a las personas a través de los metadatos de llamadas telefónicas y de interacciones registradas en Internet.

empresas y gobiernos o de los contenidos que generan en su cotidiano cada vez más datificado (Boyd & Crawford, 2012).

Por eso, más que números, la abundancia de datos nos muestra la transformación por la cual nuestra estructura social está pasando al condicionarse por el uso y análisis de esos datos para la toma de decisiones (negocios), fabricación de productos (industria), administración pública (gobierno), vigilancia (militar) y generación de conocimiento (ciencia).

> No es sólo una cuestión de escala y tampoco es suficiente considerar en términos de capacidad de análisis. Por el contrario, es un cambio profundo en los niveles de la epistemología y de la ética. 'Big data' reformula cuestiones claves sobre la constitución del conocimiento, la privacidad, los procesos de investigación, cómo debemos comprometernos con la información, y la naturaleza y la categorización de la realidad. (Boyd & Crawford, 2012, p. 665)

De hecho, vemos que hay un salto cualitativo evidente cuando se analizan grandes conjuntos de datos, pero no siempre ese salto es necesariamente positivo.

4.2 De la probabilística a la totalidad

Este salto cualitativo gana cuerpo justamente cuando, al analizar los datos de un fenómeno, podemos tener acceso al conjunto completo de esos datos y disponemos de poder de procesamiento adecuado para ello. Al mirar al pasado cuando las tecnologías analógicas eran hegemónicas, nos encontrábamos con dos problemas. O no teníamos forma de captar y almacenar grandes conjuntos de datos o, si pudiéramos traspasar esa barrera, no teníamos herramientas adecuadas para su procesamiento y análisis.

La solución entonces era utilizar muestras y cálculos probabilísticos para extraer la información relevante de gran cantidad de datos como, por ejemplo, en los análisis posteriores a un censo demográfico. "El muestreo es como una copia fotográfica analógica. A cierta distancia, se ve muy bien, pero cuando se mira más de cerca,

enfocando algún detalle particular, se vuelve borrosa" (Mayer-Schoenberger & Cukier, 2013, p. 39).

En cambio, la combinación de procesos y dispositivos de datificación con la ampliación de la capacidad de procesamiento y almacenamiento de datos posibilitada por el 'big data', nos lleva a la capacidad de extender la amplitud de la muestra y que nos permite, muchas veces, llegar al N = todo y renunciar a cálculos y métodos estadísticos tradicionales.

Este es el camino recorrido por los investigadores dentro del área de ciencias sociales que encuentran en el análisis de las plataformas de medios, a partir de un conjunto completo de determinados datos, un método eficaz de extraer informaciones y tendencias.

Así, por ejemplo, en un estudio reciente, Sabouni (2018) sugiere que a partir del análisis semántico de las canciones más populares en listas como la *Billboard Hot 100*[16], es posible extraer el sentimiento general de las personas y que, de esa forma, pueden ser usadas como previsiones económicas y estadísticamente significativas de varios índices financieros.

Al cruzar estos datos con las herramientas de procesamiento de lenguaje desarrolladas por Spotify sobre la base de su acervo de más de 70 millones de canciones, "se extraen las siguientes métricas: *danzabilidad*, energía, progreso, ruido y cadencia de las canciones para proporcionar información sobre las características culturales y perceptivas de esas 100 mejores canciones" (Sabouni, 2018, p. 07).

Entonces, analizar y entender cuáles son las canciones más escuchadas del momento a partir de métricas construidas en base a un gran volumen de datos (acervo completo del Spotify) se mostró un índice efectivo para predecir el

[16] Uno de los más importantes rankings de popularidad musical y que es editado semanalmente por la revista Billboard de EE.UU.

sentimiento económico e incluso cómo eso influye en los movimientos del mercado de acciones a corto plazo, concluye el autor (Sabouni, 2018).

Esta es una ilustración de cómo se puede obtener un salto de calidad de análisis a partir de un conjunto completo de datos (todas las canciones de Spotify), imposible si se hiciera a partir de muestras menores o subgrupos.

Por otro lado, analizar las huellas digitales de las personas a partir de las redes sociales para entender los hábitos y comportamientos puede ser impreciso o llevar a los investigadores a conclusiones erróneas, por diversos motivos. En un primer punto, se cuestiona el poder de representatividad de estos conjuntos y cierta incredulidad en el rigor necesario para llegar a conclusiones científicas relevantes.

> Twitter no representa a “todas las personas” y es un error asumir que "personas" y "usuarios de Twitter" son sinónimos: estos son un subconjunto muy particular. Tampoco podemos considerar la población usando Twitter como representante de la población global. No podemos suponer que las cuentas y los usuarios sean equivalentes. Algunos usuarios tienen varias cuentas, mientras que algunas cuentas son utilizadas por varias personas. Algunas personas nunca abrieron una cuenta y simplemente acceden a Twitter vía web. (Boyd & Crawford, 2012, p. 669)

Este análisis vale para cualquier red social que se tome como ejemplo. Además, existe la preocupación por la alta presencia de *bots*[17] que operan en esas redes para los más diversos fines (Grassi et al., 2017); también para el número de cuentas inactivas o poco activas que aumentan significativamente el margen de error en cualquier investigación hecha a partir de un conjunto que contenga estos grupos en gran cantidad.

En esa misma línea, Manovich (2011) sugiere también que es necesario tener cierta cautela con la dimensión de veracidad de los datos extraídos de plataformas de medios sociales:

[17] Los *bots* son cuentas controladas por software que generan artificialmente contenido y establecen interacciones con otros usuarios.

> Necesitamos tener cuidado al leer las comunicaciones en las redes sociales y en las redes digitales tasadas como "auténticas". Los mensajes de las personas, *tweets*, fotos cargadas, comentarios y otros tipos de participación *online* no son ventanas transparentes por sí mismas; en su lugar, a menudo se curan y se administran sistemáticamente. (p. 06)

4.3 De la causalidad a las correlaciones

En el ejemplo mostrado en la sección anterior, Sabouni (2018) concluyó que los índices construidos por su investigación estaban en equilibrio de efectividad con otros índices tradicionales usados para medir el sentimiento económico de los consumidores.

Volviendo al escenario de predominancia de las tecnologías analógicas y de bajo índice de datificación de los fenómenos, tan característicos de la era del 'small data', probablemente el estudio de Sabouni (2018) no sería posible. Hasta entonces las ciencias buscaban explicar los fenómenos naturales y sociales principalmente por medio de sus relaciones de causalidad. Es decir, la ocurrencia de un determinado tipo de evento sólo podría haber ocurrido a partir de su relación directa con un evento anterior.

En ese punto, explicar las relaciones de causalidad lleva a un ejercicio de entender el *porqué* de un determinado fenómeno ocurrir, analizando las causas y consecuencias. Al concentrarnos en encontrar correlaciones permitidas por los datos masivos, pasamos a entender *lo que* ocurre, dejando de lado así las causas y pasando a prestar atención a la interacción de las variables.

Las correlaciones aluden, de cierto modo, al pensamiento analógico, o sea, la búsqueda de relación por medio de semejanzas entre los datos a partir de una operación de vinculación analógica, y no deductiva.

"En esencia, una correlación cuantifica la relación estadística entre dos valores de datos. Una correlación fuerte significa que, cuando cambia uno de los valores de

datos, es altamente probable que cambie también el otro" (Mayer-Schoenberger & Cukier, 2013, p. 72).

Esta es una de las grandes ventajas apuntadas por aquellos que abogan que el salto cualitativo en el uso de grandes conjuntos de datos se da justamente en la posibilidad de descubrir pautas y conexiones que nos ofrezcan perspectivas nuevas, no posibles con otras tecnologías o con pocos datos. Aun cuándo, las correlaciones encontradas no siempre puedan ser evaluadas como verdaderas.

De la misma forma que Sabouni (2018) vinculó dos conjuntos de datos de distintas naturalezas para sacar correlaciones útiles, Leinweber (2012) demostró que muchas veces estas técnicas pueden encontrar correlaciones fuertes, pero sin sentido. Confrontando varios conjuntos de datos no financieros con la variación de las acciones en el índice S&P500[18], vio que había una sólida relación entre el volumen de producción de mantequilla en Bangladesh y ese índice de acciones. De hecho, los dos eventos en nada se influencian, pero el nivel de precisión llegó al 99%.

Este ejemplo ilustra un problema que los análisis de datos pueden generar: el fenómeno de la apofenia, un término surgido de estudios cognitivos que indica la tendencia del cerebro humano de ver patrones o conexiones a partir de objetos aleatorios. Es común que ocurra a partir de la observación de imágenes y está asociado al surgimiento de los primeros *emoticons*. También puede ocurrir cuando se analizan datos brutos.

En el contexto del 'big data', esta tendencia se puede agravar, ya que enormes cantidades de datos pueden ofrecer conexiones para todas las direcciones, por sí solas (Boyd & Crawford, 2012, p. 668). Este es un problema que hay que tener en cuenta cuando se trabaja con la minería de diversas fuentes distintas, es decir, un problema inherente a la dimensión de variedad.

[18] Es el índice calculado por la empresa Standard & Poor y compuesto por las 500 acciones financieras más relevantes de las bolsas de EE.UU.

Otro fenómeno sensible que surge frecuentemente con los recientes usos de los macro datos y la extracción de conocimiento por medio de correlaciones es el llamado efecto de la caja negra, en el que no se sabe las reglas y principios que llevan a un determinado proceso a ocurrir. Aquí, Mayer-Schoenberger & Cukier (2013) nos alertan "el riesgo de que las predicciones basadas en datos masivos, y en los algoritmos y conjuntos de datos que tienen detrás, se conviertan en cajas negras que no nos ofrecen ninguna rendición de cuentas, trazabilidad o confianza" (p. 220).

En el caso de aplicaciones de 'big data' en procesos de *deep learning*[19] se acentúa, ya que a veces los investigadores no saben cómo la máquina fue capaz de llegar a nuevos aprendizajes por medio de las correlaciones.

Sin embargo, a pesar de estos aspectos, analizar una correlación potente entre dos o más variables distintas permite capturar detalles dinámicamente que nos abren innumerables posibilidades; en particular, en lo referido a predecir que la correlación de las variables en el presente tiende a repetirse en el futuro.

4.4 Del pasado al futuro

La extracción de información de datos aleatorios o articulando distintas fuentes de datos por medio de correlaciones es relevante para dos aspectos: explicar un hecho ocurrido (análisis descriptivo) y predecir tendencias futuras (análisis predictivo). Aunque no se tiene total control o entendimiento de qué manera operan esas correlaciones, como en el caso del efecto de caja negra, el valor de tal análisis está en saber que determinado evento A predispone que ocurra otro evento B, con menor o mayor probabilidad. Cuanto mayor es la probabilidad, más fuerte es la vinculación de la correlación y más asertiva es la previsión de un evento futuro. "Con las correlaciones no existe la certeza, sólo la probabilidad. Pero si una correlación es fuerte, la probabilidad de que exista un vínculo es elevada" (Mayer-Schoenberger & Cukier, 2013, p. 72).

[19] *Deep learning* es un conjunto de métodos de inteligencia artificial que permite a una maquina modelar, de forma simplificada, el funcionamiento de las neuronas del cerebro humano y así perfeccionar constantemente sus cálculos matemáticos.

Es a partir de este punto que surgió la idea de que no hay necesidad de aplicar una teoría previa al analizar un gran conjunto de datos, ya que las correlaciones dentro de ellos se mostrarían espontáneamente, mostrando el camino a ser trillado para extraer información (Anderson, 2008).

En cierta forma, esta metodología parece ser menos problemática cuando se aplica el análisis *a posteriori*, o sea, cuando se analiza datos de un evento pasado sólo para interpretarlo y explicarlo. Este es el caso de algunos ejemplos de éxito utilizados en el campo del periodismo de datos, como en los *Panama Papers*.

Cuando se utiliza el análisis por correlación para poder predecir tendencias futuras con más precisión y muchas veces en tiempo real, como en el caso de las recomendaciones de contenido o de productos[20], esta etapa es hecha por algoritmos que, como vimos, deben ser instruidos *a priori* de cómo deben interpretar esas correlaciones.

Y esta etapa de operación del algoritmo no es necesariamente neutra ni objetiva (Gitelman, 2013). Muchas veces, se corre el riesgo de que valores subjetivos puedan estar impregnados en esas operaciones haciendo que el resultado refleje algún tipo de ideología o concepción de mundo predeterminada.

Muchas denuncias de malas prácticas u operaciones cuestionables de estos algoritmos fueron reportadas, incluyendo preconceptos raciales y discriminaciones de género. Google Photos, por ejemplo, etiquetaba automáticamente personas negras como gorilas en fotos o cámaras Nikon preguntaban si las personas asiáticas parpadeaban en las fotos (Crawford, 2016). En otro estudio, investigadores concluyeron que las mujeres eran menos propensas a ser impactadas por las ofertas de empleo de alta remuneración en anuncios de *Google Ads* debido a cómo se programaron estos algoritmos (Datta, Datta & Tschantz, 2015).

[20] Estas recomendaciones se hacen, por ejemplo, por los algoritmos de Youtube para sugerir nuevos videos después de asistir a un contenido determinado o por empresas como Amazon para sugerir otros productos basados en sus compras anteriores.

Esto se agrava cuando la complejidad de las operaciones de estos algoritmos no es muy bien entendida, ya que el valor final de las correlaciones está en su resultado más que en su estructura. Ver muchos vídeos sobre Donald Trump en YouTube, por ejemplo, puede llevar a que el algoritmo le recomiende vídeos de contenido xenófobo o de extremistas blancos (Tufecki, 2017). Esto plantea serias preocupaciones éticas, pero también llevan a impactos económicos negativos cuando los anuncios de empresas en esas plataformas se asocian a estos contenidos de forma no intencional[21].

¿Hasta qué punto la cantidad de datos de un fenómeno – como ver determinados contenidos en Youtube – puede determinar o explicar la calidad positiva de dicho fenómeno? Esta cuestión merece ser planteada y no está resuelta satisfactoriamente, con la cual se puede acometer con más atención en futuras investigaciones.

Otro fenómeno reciente asociado a las operaciones de algoritmo en la recomendación de contenido se refiere a la exposición de las líneas de tiempo de los usuarios en las redes sociales. Muchas veces, están diseñadas para mostrar más contenido de lo que ellos identifican como de nuestro interés, así que el algoritmo es un curador invisible que lleva a reforzar nuestras creencias preexistentes (Viner, 2016).

Esto resulta en la creación de filtros que alimentan la propia burbuja de intereses del usuario y, por lo tanto, crean un ambiente propicio para la proliferación de contenidos que entienden y siguen la lógica del algoritmo. Esta es apuntada como una de las principales estrategias exitosas en la circulación orgánica de noticias y que, como consecuencia, permite también la proliferación de las llamadas *fake news*.

[21] En 2017 empresas como Unilever y P&G amenazaron con suspender su publicidad en plataformas como Youtube y Facebook, ya que algunos anuncios de sus marcas aparecían relacionados con contenidos ofensivos (Agencias RTVE, 2018).

> "Con la ayuda de los algoritmos se les presenta a los usuarios la información y las noticias que corresponden a sus preferencias e intereses, creando universos o burbujas aisladas de información que continuamente auto-refuerzan las opiniones propias. Por la personalización del uso de internet y la adaptación a los gustos de los cibernautas, los usuarios se encuentran cada vez menos con información o ideas ajenas de su propia visión." (Mittermeier, 2017, p. 35)

En general, dentro de las investigaciones en comunicación se suele utilizar análisis descriptivos para ayudar en la solución de un problema, para limpieza y clasificación de datos para posterior análisis o análisis automatizado de contenido, por ejemplo (Arcila-Calderón et al., 2016). Sin embargo, investigaciones con análisis predictivos son comúnmente asociadas a las industrias que las utilizan para el desarrollo de nuevos servicios o generación de negocios.

4.5 De lo privado al público, del abierto al cerrado

Por último, se llega al centro de otra importante cuestión con respecto a la naturaleza de los datos, que puede verse por dos grandes perspectivas: su origen y la disponibilidad de acceso a los mismos.

En cuanto al origen, en realidad hay que colocar la mira en el agente responsable por el proceso de datificación de un determinado fenómeno. Este agente puede ser privado (empresas, individuos, etcétera) o órganos e instituciones de la esfera estatal, principalmente en lo que se refiere a la administración pública. En este punto, podemos decir si los datos tienen un origen *privado* o *público* a partir de la naturaleza del agente que ejecuta la correspondiente datificación.

Sin embargo, otro aspecto es si estos datos son accesibles amplia y abiertamente o si existe algún tipo de restricción o protección legal que limite el acceso a ellos. Así, podemos decir si los datos son *abiertos* o *cerrados*.

Una clasificación no necesariamente determina la otra. Una empresa privada puede generar datos y compartirlos libremente, como en el ejemplo de Twitter que

permite el acceso vía API[22] del contenido de sus publicaciones (privado y abierto). En el mismo caso, Twitter no ofrece algunos (meta) datos que sus algoritmos usan para fines publicitarios (privado y cerrado).

En el caso de datos referentes a la esfera pública, a menudo vemos que en las democracias modernas hay un aprecio por la transparencia del mayor número posible de informaciones generadas por sus agentes y órganos. Un censo demográfico, por ejemplo, es público y abierto en la medida en que proporciona sus resultados después de la medición. Pero puede ser público y cerrado una vez que algunas informaciones sean clasificadas con algún tipo de restricción por motivos militares o de seguridad nacional.

Este tema asumió gran importancia cuando, en los últimos años, varios informes y órganos empezaron a abogar por la máxima posibilidad de que agentes, empresas e investigadores deberían ofrecer los datos abiertamente para que otras personas y otros usos fueran les asignados. Cuando se ve específicamente dentro del contexto de la producción académica, este tema es más visible.

> Lo recomienda la OCDE (2015) y lo exige el gobierno de EUA desde 2013 a través de las diversas agencias de financiación: *National Science Foundation* (NSF, 2014) y el *National Institutes of Health* (NIH, 2015), entre otros. En Europa, el acceso abierto a los datos de investigación ha sido, hasta ahora, sólo un piloto (ORD Pilot) para nueve áreas de proyectos financiados en el marco de Horizon 2020. (Gómez, Méndez & Hernández-Pérez, 2016, p. 546)

Hay una percepción general de que en una sociedad de información efectiva es necesario que la mayor cantidad de datos esté disponible para todos, pues sólo de esa forma sería posible la innovación y el desarrollo de estas sociedades. Esta es la idea que está detrás de lo que entonces se denomina *open data*[23].

[22] Las *Application Programming Interfaces* (API) son un conjunto de protocolos computacionales que permiten a los programadores informáticos crear soluciones específicas para ciertos sistemas o software.
[23] Se puede entender el *open data* como un subgrupo de 'big data', pero va más allá, ya que se enmarca cualquier tipo de dato que esté en formato abierto, público y accesible.

Sin embargo, a pesar del incentivo de órganos e instituciones en democratizar y perfeccionar las prácticas de intercambio de datos de cualquier naturaleza, persiste una serie de problemas relacionados con cuestiones culturales, técnicas y de definición y estandarización de normas.

En la dimensión cultural, desde empresas a investigadores todavía se recusan en poner a disposición sus datos por temor a consecuencias legales, mal uso por parte de terceros o simplemente desconocimiento de dónde y cómo compartirlos (Gómez, Méndez & Hernández-Pérez, 2016, p. 547).

Por las cuestiones técnicas, existen problemas que surgen por la enorme complejidad del ecosistema de códigos, lenguajes, programas, repositorios, normas, etcétera que impiden la estandarización y que deriva también de los diversos usos que se dan y se recompilan entre las distintas disciplinas. A todo ello, se suma el hecho de que a nivel de 'big data' se multiplican los problemas también en las cuestiones de procesamiento y almacenamiento que, a menudo, necesitan una estructura específica.

En este sentido, hay una preocupación de que este escenario acentúe la división e intensifique la desigualdad de oportunidades entre profesionales, instituciones e investigadores que no llegan a tener medios o habilidades de acceder a ese material o tecnología para procesarlo y analizarlo.

Así, muchos datos y metadatos procedentes de plataformas de redes sociales se utilizan para fines publicitarios y mejoras del propio sistema, lo que lleva a estas compañías a no liberar el acceso o restringir su uso por medio de patentes u otros tipos protección legal.

De esta forma, los investigadores que son contratados por estas empresas pueden tener acceso irrestricto a estos datos, mientras que el resto de la comunidad

académica, no[24]. Aún se puede discutir si estos investigadores tendrían la independencia de plantear cuestiones que podrían ser sensibles a estas mismas empresas (Manovich, 2011).

> Además de cuestiones de acceso, existen cuestiones de habilidades. *APIs, scraping* o el análisis de grandes cantidades de datos son habilidades generalmente restringidas para aquellos con conocimientos computacionales. Cuando las habilidades computacionales se colocan como las más valiosas, surgen cuestiones sobre quién es favorecido y quién es desfavorecido en tal contexto. Esto configura nuevas jerarquías alrededor de quién puede *leer los números*. (Boyd & Crawford, 2011, p. 674)

Por lo tanto, esto nos hace señalar que hay una transformación acelerada del tipo de profesional o investigador necesario para ello. Por un lado, está el desarrollo de proyectos multidisciplinarios, incorporando profesionales de comunicación que agreguen conocimiento y habilidades para contar historias o modos de visualización a partir de los datos. En el otro, cada vez se exige más que comunicadores en un aspecto general desarrollen habilidades como programación o estadísticas, con frecuencia exigidas a su entorno profesional, cada vez más digital y datificado.

> Lo que ya no está tan claro es que los profesionales de la información, bibliotecarios o comunicadores, puedan considerar que los big data serán algo con lo que tratarán habitualmente; más bien, no. Muy pocos podrán integrarse, ni siquiera a medio plazo, en equipos multidisciplinares que trabajen con datos masivos y habrá que distinguir entre *data scientist*, *data analyst* y otras nuevas denominaciones que seguro irán apareciendo. (Hernández-Pérez, 2016, p. 518)

Es decir, vemos que diversos elementos apuntan a un cambio significativo en las prácticas y habilidades, o sea, un cambio paradigmático en la forma que se hace ciencia del campo social. No obstante, parece haber un punto sensible: ciertas comunidades académicas no comparten o no tienen acceso a dicha infraestructura.

[24] El debate en torno al acceso a los datos de plataformas de redes sociales por parte de la comunidad académica no es nuevo, pero con los recientes escándalos de filtración de informaciones involucrando a Facebook hubo una nueva ola de críticas en favor de la apertura (Brooks, 2018).

Como detallaremos más adelante en el capítulo 5.3.3, estos cambios apuntan para nuevas formas de obtener y manipular estos datos para la investigación, sea por su disponibilidad de acceso para los investigadores, sea por la habilidad y herramientas que estos necesitan para compilar, procesar y analizar dichos conjuntos de datos.

En este escenario, veremos como está ocurriendo ese impacto en las disciplinas dentro del campo de la comunicación; principalmente, desde el punto de vista de la producción académica, pero también en algunas prácticas profesionales en esas áreas que son objetos de estudio para tales investigaciones.

5. Repercusión en las investigaciones sobre las comunicaciones mediáticas

En los dos apartados iniciales de este capítulo se trató de identificar los elementos estructurales que se relacionan con el "big data", así como explorar los principales cambios epistemológicos detrás de ese fenómeno, que nutren los debates académicos e investigaciones en curso dentro de esa perspectiva de cambio de paradigma.

En este apartado, nos dedicaremos a explorar con más detalle la repercusión de esos cambios en los debates sobre la comunicación mediática contemporánea y sus implicaciones para la investigación académica en las disciplinas de esa área.

De principio, se identifica que hay una importante implicación del fenómeno en la forma en que las industrias de medios se están reorganizando, cuando se observa por la perspectiva de los datos pasando a ser una importante materia prima, a veces más que la producción de contenido. Este impacto está en el orden de la estructura de los agentes y de su relación con los datos masivos dependiendo de su posición frente a ellos.

Las implicaciones estructurales en los sistemas de medios, tanto locales como globales, no son el foco principal de este trabajo, pero entiendo que es importante echar un vistazo a esas transformaciones que están en curso, pues también influyen en la forma como las investigaciones académicas sobre y mediante los datos masivos de estos medios analizan tal cambio.

En este sentido, Mayer-Schoenberger & Cukier (2013) sostienen que las empresas, instituciones e individuos que trabajan con 'big data' pueden agruparse en tres amplias categorías, de acuerdo con el énfasis de actuación dentro del ciclo de datificación, almacenamiento, procesamiento, minería y utilización de los conjuntos de datos (p. 156).

En el primer grupo están aquellos que generan los datos o, de alguna forma, permiten el acceso a ellos. Son empresas, instituciones, órganos e individuos de las más variadas industrias y sectores, académicos o no, públicos o privados, posicionados en el centro de los flujos de información, que logran así capturar y datificar a ese flujo para extraer valor y oportunidades (Mayer-Schoenberger & Cukier, 2013, p. 159).

En el segundo, están empresas e individuos que son especialistas en datos, es decir, presentan las herramientas y el conocimiento para minar y analizar los grandes conjuntos de datos, muchas veces oriundos de los agentes del primer grupo (Mayer-Schoenberger & Cukier, 2013, p. 160).

En la tercera categoría, Mayer-Schoenberger & Cukier (2013) colocan a los que poseen *mentalidad de datos masivos*. "En el caso de este grupo, la fuerza radica en que pueden ver las oportunidades antes que los demás, aun cuando carezcan de los datos o de las aptitudes para actuar en función de esas oportunidades" (p. 162).

Manovich (2011) también sostiene que se establecen tipos diferentes de clases dentro de la estructura de una sociedad basada en datos masivos. "En concreto, las personas y las organizaciones se dividen en tres categorías: aquellas que crean datos (tanto conscientemente como dejando huellas digitales), aquellos que tienen los medios para recogerlos y aquellos que tienen experiencia para analizarlos" (p. 10).

Muchas veces, empresas, instituciones e individuos se encajan en más de un grupo e incluso pueden actuar en las tres etapas simultáneamente. Las grandes empresas de tecnología que están en el centro de este fenómeno, como Google y Facebook, tienden a dominar las tres etapas como forma de ser más competitivas e innovadoras en sus servicios.

En el caso de la amplia área referente a la comunicación mediática, hay una convergencia en curso de empresas de diversos sectores que, por un lado, actúan en servicios ligados a los medios, como las empresas de tecnología o de

telecomunicaciones, además de las propias empresas tradicionales de medios. Por otro lado, hay aquellas que interactúan o se nutren de los datos provenientes de los procesos de comunicación mediática para el desarrollo de sus negocios o investigaciones.

De esta forma, el desarrollo de un ecosistema de empresas e instituciones alrededor y dentro de ese flujo de datos está alterando significativamente la configuración estructural de los sistemas de medios de comunicación a nivel global y local.

Como sostiene Arsenault (2017), a partir de la digitalización y de la datificación masiva de los servicios de medios, las empresas de comunicación dejan de ser sólo productoras de contenido y pasan a ser también productoras y gestoras de datos. Y toda la cadena de producción de las industrias de medios pasa a estar condicionada por la extracción de valor de esos datos (el contenido también es datificado). "No es más que limitarse a apostar en el nuevo y mejor sistema de productos de medios; es sobre acumular los datos para saber quién va a consumir lo qué, en cuál plataforma, cuándo y cuánto" (Arsenault, 2017, p. 20).

El ejemplo más bien acabado de esto es NetFlix. Su apuesta en la producción de contenido propio, además de su servicio de *streaming*, está basada en los datos de comportamiento de sus usuarios a partir del consumo de todos los contenidos dentro de la plataforma[25].

Pero, volviendo el foco nuevamente a las cuestiones relativas a la producción académica, la casi totalidad de las investigaciones dentro del campo de comunicación mediática y se encaja en el tercer grupo, a partir de la clasificación de Manovich. Esto se revela importante pues, como diremos más adelante, hay cuestiones sensibles en accesos a los datos y habilidades para analizarlos que pueden aumentar la brecha entre investigadores.

[25] A partir del análisis de datos de consumo y audiencia, Netflix pasó a monitorear los comportamientos de sus clientes en 190 países y redefinió la forma de distribución, visualización y producción de su contenido original basado en esos datos (Agencia Efe, 2017).

Así, en lo que se refiere específicamente al campo de los estudios sobre periodismo y comunicación mediática, parece haber un creciente interés por parte de los investigadores en ese fenómeno y que se evidencia en los números dedicados al tema en revistas de comunicación y también en el surgimiento de revistas especializadas en 'big data' y comunicación, como *Big Data & Society*, *Social Media + Society*, *Big Data Quarterly* o *Big Data Innovation Magazine*, como ejemplos.

Por este ángulo, Shahin (2016) argumenta que hay dos grandes categorías expansivas con respecto a estos estudios en las disciplinas de ciencias sociales y de humanidades: investigaciones sobre 'big data' y las investigaciones mediante el uso de 'big data'.

La distinción propuesta está en sintonía con otros autores, como Beer (2016), que entiende que los datos masivos deben ser estudiado como un fenómeno material y conceptual, o como Veltri (2017), que articula el 'big data' entre contexto social y metodología innovadora para investigaciones sociales.

Así, partiendo de esa división señalada por Shain (2016), exploraremos algunas corrientes de investigaciones que se efectúan en el campo de las comunicaciones, separadas en investigaciones sobre el 'big data' (como fenómeno, concepto, discurso o estructura) e investigaciones mediante el uso de "big data" como metodología para los más diversos fines.

5.1 Investigaciones sobre el 'big data'

Los estudios sobre 'big data' son aquellos que suponen que la idea de los datos masivos como un fenómeno social va más allá de las cuestiones técnicas de volúmenes de datos y velocidades de procesamiento. Tratan de analizar sus impactos en el tejido social, independiente de usar metodologías de 'big data' para eso.

Podemos también separarlos en dos grandes corrientes. La primera, de tradición hermenéutica, trata de comprender y explicar el fenómeno por medio de sus más

variados aspectos, sean ellos culturales, sociales, conceptuales, estructurales, fenomenológicos, etcétera. Las próximas dos secciones pretenden profundizar estos estudios.

La segunda corriente, a partir de las tradiciones de la teoría crítica, busca interpretar el fenómeno dentro de las relaciones sociales, a fin de contextualizarlo como una herramienta de mantenimiento del *status quo*, proponiendo así una reflexión crítica sobre el tema. Abordaremos estos estudios en la secuencia.

5.1.1 Investigaciones conceptuales

Algunos de los primeros estudios realizados en estos últimos diez años lanzaban una mirada sobre aspectos característicos e intentaban definir el 'big data' como un concepto, con la finalidad de ver rasgos marcados y comunes que fueran propios para establecer una epistemología amplia.

Después de la primera contribución de Laney (2001) con la definición a partir de las tres *dimensiones Vs* originales (volumen, velocidad y variedad), otros estudios fueron siguiendo en esa línea para ampliar la misma definición con otras dimensiones, en la medida que fueron evolucionando sus usos. En general, son estudios ligados directamente al uso interno de las ciencias computacionales.

Como explica Lee (2017), los estudios de IBM agregaron la dimensión de veracidad, mientras que los de SAS introdujeron la de variabilidad y Oracle sugirió la dimensión de valor (p. 294). En ese campo, se siguieron otros aportes y se espera que para los próximos años ese número aumente considerablemente[26].

Otros, a pesar de apuntar que las dimensiones son un referente, siguieron una línea de carácter hermenéutico con intención de analizar el 'big data' como un fenómeno más amplio. En este sentido, fueron esenciales las contribuciones de Boyd & Crawford (2012) y de Manovich (2011).

[26] Shafer (2017) presenta la evolución de esas dimensiones y llega a elaborar una lista de 42 Vs que deben ser considerados en 'big data' y 'data science' hoy en día.

En la visión de Boyd & Crawford (2012), los datos masivos deben ser definido como un fenómeno al mismo tiempo cultural, tecnológico y académico, resultante de la interacción de características tecnológicas, analíticas y mitológicas (p. 663).

De la otra parte, los estudios de Manovich se centran más en las aplicaciones tecnológicas y académicas para las ciencias sociales y de humanidades, permitiendo a los investigadores entender el 'big data' como una práctica que une métodos cuantitativos y cualitativos. “No es necesario elegir entre el tamaño de datos y la profundidad de los datos. El conocimiento detallado y las percepciones que antes sólo podían alcanzarse sobre algunas personas, ahora se pueden alcanzar sobre muchas más” (Manovich, 2011, p. 03).

En la misma concepción están Mayer-Schoenberger & Cukier (2013), para los que no era tan importante estudiar a partir de las *dimensiones Vs*, ya que estas definiciones fueron útiles en un primer momento, pero hoy se muestran improductivas e imperfectas (p. 247). De esta forma, los autores también siguieron en la línea de identificar las principales características del fenómeno en una perspectiva holística y su impacto en nuestro cotidiano, lo que hizo su libro un importante referente.

Otros siguieron proponiendo alternativas a estas clasificaciones, como Lupton (2015), que sugiere una clasificación por las *dimensiones Ps* (Productivo, Perverso, Polimorfo, Político, entre otras). Otros trataron de analizar y dejar más precisa la clasificación según sus cualidades, como Kitchin & McArdle (2016) quienes, tras el análisis de veinte y seis conjuntos de datos masivos distintos, sugieren que "el 'big data', como una categoría analítica, precisa ser desmembrada, con los géneros dentro de ese universo mejor delineados y sus varias especies identificadas” (p. 01).

5.1.2 Investigaciones estructurales

Además del trabajo de conceptualización, hay un gran interés en lanzar luz sobre el fenómeno social que de él se constituye y cómo eso impacta en la producción académica y, al mismo tiempo, en las propias estructuras de los medios de comunicación y de las industrias creativas.

A pesar de no haber todavía tantos estudios buscando analizar la forma en que el impacto del 'big data' se da en macro estructuras sociales, como instituciones y sectores industriales, algunos investigadores empiezan a preocuparse por ese movimiento. En la búsqueda de comprender la cuestión, Arsenault (2017) examinó cómo las grandes redes de datos están solidificando las relaciones entre grandes empresas de Internet y empresas de medios, para concluir que los cambios en las estructuras de los mercados de comunicación, cada vez más globales, se dan básicamente de dos formas:

> En primer lugar, la datificación de los medios de comunicación y la aplicación de servicios de 'big data' a partir de estos datos están facilitando la consolidación de redes de competencia y de fusiones anteriormente vistos en la formación de las *joint ventures* y en los mercados de bienes y servicios. En segundo lugar, 'big data' representa un formato global emergente, teóricamente análogo a la proliferación global de formatos de televisión. (Arsenault, 2017, p. 08)

Los conglomerados empresariales se forman dentro de cierta lógica. En el caso de las industrias de comunicación[27], el tradicional enfoque en la producción de contenido como matriz de estas industrias está cediendo espacio para servicios de obtención de datos que pautarán no sólo la producción de contenido, proporcionando análisis predictivos, sino también los servicios y estructuras de distribución (Arsenault, 2017).

Si, por un lado, aún hay pocos estudios en ese sentido, por otro, hay un esfuerzo creciente en entender el impacto en las estructuras de producción e investigación

[27] Hay un importante movimiento de formación de oligopolios en las industrias de telecomunicación y medios que se evidencia tras la compra de Time Warner por parte de AT&T y en el reciente avance de Comcast y Disney por adquirir a Fox (Zeitchik, 2018).

científica que ya señalamos en diversos momentos en este trabajo. En la introducción del volumen especial de la revista *Big Data & Society* intitulado *Conceiving the Social with Big Data: A Colloquium of Social and Cultural Scientists*, Wagner-Pacifici, Mohr & Breiger (2015), después de analizar todos los diez y seis textos que constituyen ese volumen, afirman en su conclusión que las ciencias sociales y las humanidades parecen estar en camino de transformaciones significativas, cada vez más involucradas con las ciencias computacionales.

En este sentido, el artículo de Kitchin (2014) para esa misma revista describe los indicios de lo que parece ser el camino para el surgimiento de nuevos programas de investigación que caracterizan la llamada *Data Driven Science*. Estos nuevos programas se caracterizan por sus esfuerzos "en mantener los principios del método científico tradicional, pero más abiertos a utilizar una combinación híbrida de enfoques abductivos, inductivos y deductivos para llegar a la comprensión de un fenómeno" (Kitchin, 2014, p. 05).

Así, una implicación importante parece ser la convergencia entre estructuras sociales y estructuras informáticas normalizadas en la vida cotidiana de nuestras sociedades, que ocurre también dentro de los centros académicos, moldeando la forma como se hace ciencias sociales y las habilidades de los propios investigadores.

> Aunque los investigadores de comunicación están bien posicionados para explorar estos ámbitos debido a su larga historia de evaluación cuidadosa del contenido de la comunicación, algunas reformulaciones serán inevitables, así como la necesidad de investigación colaborativa con científicos de la computación e ingenieros, y la redefinición de la enseñanza para estudiantes de graduación que contemplen las más recientes y poderosas técnicas para el análisis de materiales textuales en formato electrónico. (Shah, Cappella, & Neuman, 2015, p. 12)

Por lo tanto, entender mejor el 'big data' como fenómeno más amplio y complejo se hace imprescindible para entender cómo reordena las relaciones entre los agentes en un sector como el de la industria de la comunicación. En simultáneo, cómo redefine la forma y las habilidades necesarias para realizar las investigaciones en las disciplinas relacionadas con ese campo.

5.1.3 Investigaciones críticas

Pero, volviendo a las promesas iniciales de que el 'big data' permite un acceso profundo y objetivo a los fenómenos de la realidad hasta llegando a los recientes casos de vigilancia por parte de gobiernos, de cercenamiento de la privacidad de los ciudadanos e, incluso, de la filtración de informaciones personales por empresas que deberían velar por la integridad de esos datos, hay un volumen creciente de investigadores que proyectan una mirada crítica sobre cómo estas tecnologías pueden impactar en la vida social cotidiana. De acuerdo con Shahin (2016), la perspectiva crítica

> desafía el clima de tecno-utopía generado y constantemente revitalizado en los discursos convencionales sobre 'big data'. Cuestiona la "normalidad" de la visión de mundo neoliberal, en que grandes corporaciones y su búsqueda de lucro son vistas como el camino natural del progreso. También discute la apropiación capitalista de la actividad humana y de las democracias sociales, y expone la relación entre *Big Data, Big Business* y *Big Government* que hace que dicha apropiación sea posible. (p. 984)

El enfoque crítico apunta al fenómeno de los datos masivos entendido como discurso e ideología y, a partir de eso, debemos entender cómo el término 'big data' ha sido utilizado conceptualmente dentro de los ámbitos comerciales, económicos, científicos, organizacionales y de la esfera política. Beer (2016) también sigue en esa línea y sostiene que "hasta ahora hemos centrado prácticamente toda nuestra atención sobre el fenómeno y hemos dado muy poca atención al poderoso concepto que define, aprueba e introduce aquello que es aparentemente considerado como 'big data'" (p. 09).

Parte de este discurso está asentada sobre la retórica de objetividad de los datos y, en nuestro caso, también del acceso profundo a la realidad "precisa" que los datos proporcionan a los investigadores en su práctica científica. Es por eso también que hay cierto tono mitológico en las potencialidades de los datos masivos para todas sus aplicaciones.

Esto moldea, por un lado, cierto tipo de discurso ideológico que va siendo solidificado en el mercado y en la industria y que lleva, como consecuencia, a construir una noción popular de que 'big data' es una fuente de verdad objetiva y neutra. Sin embargo, posicionar de esa forma oscurece las muchas maneras por las cuales los datos -masivos o no- son socialmente construidos e interpretados (Puschmann & Burgess, 2014).

Estos autores apuntan que los datos masivos como concepto y como fenómeno están interrelacionados; hay otras líneas de investigaciones críticas que lo utilizan también como método, explorando enormes conjuntos de datos textuales con la ayuda de las herramientas de minería de datos para hacer análisis de discurso.

> Al analizar las aplicaciones del 'big data', sus métodos y suposiciones, tienen como objetivo mejorar la forma en que se realiza la investigación social y cultural. El ya superado binario del 'big data' - es bueno o malo? - descuida una realidad mucho más compleja que se está desarrollando. Hay una infinidad de configuraciones disciplinares diferentes, a veces completamente opuestas, técnicas y prácticas que se reúnen (aunque incómodamente) bajo la bandera del 'big data'. Campos involucrados en investigaciones de medios y comunicación que utilizan datos masivos para resolver dilemas o plantear nuevas cuestiones nos llevan a considerar cuidadosamente las maneras en que el término y las técnicas se emplean. (Crawford, Miltner & Gray, 2014, p. 1665)

En general, son investigaciones que examinan grandes volúmenes de datos generados por las redes sociales o cualquier otro conjunto de textos extraídos de internet y que pueden ser utilizados para análisis interpretativos variados o dentro de una perspectiva de contextualización y encuadramiento.

5.2 Estudios mediante 'big data'

Aquí es donde se encuentra la mayor parte de las investigaciones académicas (a menudo de contenido técnico) que utilizan volúmenes de datos masivos generados por los medios de comunicación y técnicas de minería para el desarrollo de sus estudios. Son los que Shahin (2016) identifica como pertenecientes a una línea hegemónica administrativa, o sea, "investigaciones metodológicas que permiten a

los administradores - gubernamentales y corporativos - descubrir nuevas fuentes de datos, nuevas formas de minería y nuevas técnicas de análisis" (p. 982).

Por eso, en el caso de los usos del 'big data' como metodología, hay una gran amplitud de perspectivas, actuaciones y enfoques. Este trabajo no tiene el objetivo de analizar exhaustivamente todas las posibilidades, que son innumerables y de los más variados tipos. Como parte del objetivo de explorar e identificar los principales debates en curso, apuntamos abajo las principales líneas de investigación en las disciplinas de comunicación que ilustran nuestro propósito.

A pesar que algunos estudios críticos y conceptuales, como dijimos, utilizan técnicas de 'big data' como parte del desarrollo de sus investigaciones (Shahin, 2016b; Kitchin & McArdle, 2016), es en la línea empírico-analítica donde generalmente encontramos la mayor parte de los usos de datos masivos, permeando todas las áreas de estudios.

Así, en las disciplinas de comunicación, se popularizaron muchos estudios basados en textos y discursos nativos de la llamada web 2.0, principalmente de aquellos venidos de las redes sociales y otras plataformas colaborativas.

5.2.1 Estudios sobre redes sociales

Con el surgimiento y popularización de las plataformas de redes sociales entre 2005 y 2015 y su hegemonía frente a otros servicios populares de comunicación *online*, como el correo electrónico (Lee, 2017), éstas se convirtieron en la principal fuente de (meta) datos para que empresas, políticos e investigadores pudieran analizar las huellas digitales dejadas por las personas.

Sólo Facebook, la mayor plataforma de las redes sociales, tiene actualmente más de 2.200 millones de usuarios en todo el mundo (Facebook, 2018). A partir de su modelo de negocios en el que los usuarios pueden usarlo gratis dando a cambio los datos personales de su navegación, fue posible la datificación masiva y minería del comportamiento de esos usuarios, preciosos para los estudios sociales y culturales.

El hecho de que todas estas plataformas permiten de alguna manera dar voz a sus usuarios, permitiendo e incentivando que publiquen sus opiniones y pensamientos[28], ha colocado a los investigadores en una posición privilegiada de análisis. Como sostiene Crawford (2009), "estas plataformas permiten entender los espacios divergentes de la modernidad en un solo lugar, crean simultánea una división entre este ideal y aquello que es humanamente manejable" (p. 526).

Mientras la visión de Crawford va en el sentido de una diferencia entre cómo las personas se comportan en el cotidiano y cómo es reflejado y representado en su vida *online*, están aquellos que ven en esas plataformas un vasto territorio a ser explorado de formas antes nunca posibles.

> El surgimiento de las redes sociales a mediados de los años 2000 creó oportunidades para estudiar procesos y dinámicas sociales y culturales de nuevas maneras. Por primera vez, podemos rastrear imaginaciones, opiniones, ideas y sentimientos de cientos de millones de personas. Podemos ver las imágenes y los vídeos que crean y comentan, monitorear las conversaciones en las que están involucradas, leer sus mensajes y *tweets*, navegar en sus mapas, escuchar sus listas de canciones y seguir sus trayectorias en el espacio físico. Y no necesitamos pedir su permiso para hacerlo, ya que ellos mismos nos permiten hacer esto dejando todos estos datos públicos. (Manovich, 2011, p. 02)

Para Manovich y otros autores citados en este estudio, como Mayer-Schoenberger & Cukier (2013) y Gitelman (2013), los estudios sociológicos y comunicacionales basados en el análisis de plataformas sociales aportan a las investigaciones un marco cuantitativo y "objetivo" importante, muchas veces faltante en los estudios de esas disciplinas.

Entonces, el salto cualitativo que se obtiene a partir del salto cuantitativo de datos, es decir, datos completos de un individuo sumados a los datos completos de un colectivo, sería positivo para las ciencias sociales y de humanidades, al permitir a los investigadores trabajar conjuntamente con tamaño y profundidad de datos,

[28] En la línea de tiempo de los usuarios está la famosa pregunta de Facebook "¿Que estás pensando?"

para así estudiar muestras exactas formadas por miles de millones de expresiones culturales, experiencias, textos y enlaces (Manovich, 2011).

5.2.2 Análisis lingüísticos y análisis de discurso

Pero no sólo las experiencias y el comportamiento de las personas fueron volcados en datos cuantificables. En lo que Manovich antes determinó como expresiones culturales podemos incluir todos los contenidos en cualquier formato y lenguaje que también pasan a cohabitar en un entorno común, el del ambiente digital.

Por lo tanto, si por un lado las opiniones de los individuos y de los colectivos sobre los asuntos cotidianos de cualquier naturaleza pueden ser leídos como discurso, por otro, la comunicación y digitalización de expresiones culturales que antes no eran posibles generan amplias muestras de texto y datos sobre canciones, películas, juegos o cualquier expresión cultural, también interesantes para los investigadores como forma de discurso cultural.

De tal modo, los métodos tradicionales de análisis de discurso pueden ser combinados con técnicas de procesamiento de lenguaje natural y otras técnicas de minería de datos que expanden las aplicaciones y alternativas para estudios en semántica y disciplinas en el campo de la lingüística.

Este punto sostienen muchas investigaciones innovadoras en este campo que mediante el uso de los macro datos, sea por grandes volúmenes o por técnicas de minería posibles en esas escalas, permitiría ver patrones que antes no eran posibles. Este es el camino que orienta algunas líneas de investigación, como las de Manovich (2011) en las que indica que

> (...) podemos utilizar ordenadores para explorar rápidamente conjuntos de datos visuales masivos; a continuación, seleccionar los objetos para un análisis manual más profundo. Mientras que la investigación asistida por ordenador de grandes conjuntos de datos culturales revela nuevos patrones en esos datos que incluso el mejor "lector atento" podría dejar escapar -y, por supuesto, incluso un ejército de científicos en humanidades no sería capaz de "leer con cuidado" estos conjuntos de

> datos masivos- un ser humano todavía es necesario para dar sentido a estos patrones. (p. 09)

Este ejercicio de *dar sentido* es lo que todavía orienta a los investigadores en sus estudios, por más que los algoritmos y las técnicas de minería ayudan a estructurar los datos y posteriormente visualizar patrones. Sin embargo, el descubrimiento de patrones recurrentes en estos discursos permite a los investigadores predecir hábitos y comportamientos y así desarrollar análisis predictivos para las más variadas finalidades, como en el estudio citado anteriormente de Sabouni (2018) sobre el mercado financiero.

> Estos métodos predictivos y técnicas analíticas pueden proporcionar una visión, si resuelven directamente problemas sociales significativos. La disponibilidad de grandes cantidades de datos de comunicación social sobre la vida cotidiana - reacciones en tiempo real sobre eventos mediáticos, políticos, ambientales y sociales- y la evaluación de esos datos por los grupos productores del contenido plantea la posibilidad de acceso inmediato a los discursos culturales. (Shah et al., 2015, p. 09)

A continuación, veremos cómo los estudios que utilizan datos masivos desarrollan métodos predictivos, bastante utilizados por investigadores en áreas como publicidad y marketing, mientras que las técnicas analíticas están siendo explotadas por disciplinas como el periodismo de datos para la construcción de narrativas y forma de desvelar tramas ocultas sobre problemas sociales significativos, como afirmado.

5.2.3 Análisis predictivos

Los estudios a partir de análisis predictivos son un gran campo de aplicación de las técnicas de minería usadas para encontrar patrones, principalmente en comportamientos de navegación *online*, ya que muchas de esas técnicas fueron desarrolladas y mejoradas a partir de la popularización del comercio electrónico y de los análisis de navegación de las páginas web en los años 90 y 2000 (Lee, 2017).

Como hemos visto en el apartado sobre las correlaciones en el segundo capítulo, la minería de datos es utilizada para descubrir los patrones de correlaciones que permitan a investigadores y empresas la aplicación de análisis predictivos. Por cuenta de eso, "encontrar aproximaciones en contextos sociales es una de las aplicaciones de las técnicas predictivas de datos masivos. Igualmente poderosas resultan las correlaciones con nuevos tipos de datos para solucionar necesidades cotidianas" (Mayer-Schoenberger & Cukier, 2013, p. 78).

Solucionar problemas o necesidades cotidianas, en las palabras de los autores, abre un campo amplio de aplicaciones partiendo de estudios de análisis predictivos. En ese punto, hay un importante impacto en estudios en la gran área de investigación ligada a las audiencias. En el caso de Netflix, como apuntado en el inicio de este capítulo, entender la forma cómo su audiencia consume los contenidos está en la base de toda su estructura de servicios, definiendo hasta la forma como produce su contenido propio.

Este ejemplo nos muestra, por estos motivos, las formas en que muchas de estas aplicaciones predictivas están vinculadas a investigaciones del orden administrativo, es decir,

> es por eso que las técnicas como la minería de opinión y el análisis de los sentimientos se están volviendo tan populares, porque hacen que los administradores entiendan mejor cómo sus consumidores se sienten sobre productos específicos y personalizan la visualización de productos y contenido con más eficiencia. (Shaim, 2016, p. 982)

5.2.4 Análisis descriptivos

Antagónicamente a los estudios con análisis predictivos, los estudios y la práctica periodística, por medio de lo que se ha convenido llamar el periodismo de datos, vienen apoyándose en las nuevas técnicas de minería de grandes conjuntos de datos para análisis descriptivos en la identificación de elementos para sus respectivas narrativas.

Los estudios en periodismo de datos apuntan que esta modalidad no es nueva, porque el periodismo, de una forma u otra, en su trabajo tradicional siempre analizó datos para poder contextualizar e informar (Crucianelli, 2012).

La evolución de este tipo de periodismo remite a lo que anteriormente era posicionado como periodismo de investigación o periodismo de precisión. Con las técnicas computacionales más evolucionadas, a partir de los años 90, se comenzó a hablar en *Computer Assisted Journalism* (CAJ) o de *Computer Assisted Reporting* (CAR). Eran "modalidades del periodismo de precisión en la que se utilizan los ordenadores para examinar las bases de datos y descubrir asociaciones o correlaciones estadísticas en todo tipo de documentos informatizados" (Rodríguez, 2016, p. 260).

Con la normalización de la datificación y del fenómeno de los macro datos que hemos presenciado, hay una nueva realidad en la que el trabajo periodístico debe apoyarse para renovar su esencia y seguir buscando nuevas historias ocultas en los datos. No es muy diferente de lo que los académicos buscan cuando se acercan al mismo fenómeno, como ya visto.

En esencia, las nuevas técnicas de minería de datos, el uso de algoritmos específicos y las nuevas formas de visualización han sido elementos y habilidades recurrentes en la rutina de algunas redacciones contemporáneas que vemos así estructuradas (Renó & Renó, 2015). Esto, en concreto, ha sido la base para investigaciones largas y profundas de destaque de los últimos años, como el escándalo de los *Panama Papers*.

En resumen, el debate de estas cuestiones en el medio académico se comprende como

> producto de los cambios culturales y tecno-comunicativos que distinguen nuestra época. Así el desarrollo del periodismo de datos resume los procesos de digitalización; la filosofía del *Open Data* y la transparencia y el acceso público a la información; herramientas estadísticas y de visualización; y las habilidades investigativas del periodista de toda la vida para compilar, filtrar, contextualizar,

> contrastar, organizar, jerarquizar y contar una historia de forma atractiva. (Brito & Chico, 2013, p. 03)

Así, la misma clase de impactos que estamos identificando en muchos de los textos académicos mediante el uso del 'big data' parecen surgir en el trabajo periodístico que utiliza también grandes volúmenes de datos, en los límites de sus propias características. Es decir, el deseo de encontrar historias interesantes permanece, pero en ese sentido hay un cambio en la composición de las fuentes, en las formas de narrativas y en las nuevas habilidades que desafían esa disciplina.

Al ampliar la mirada a los debates que son seguidos por muchos de los profesionales que trabajan o hacen investigación, no sólo en periodismo, sino en el amplio campo de la comunicación mediática contemporánea, se puede apuntar cuáles son actualmente estos importantes desafíos que el 'big data' levanta para nosotros.

5.3 Desafios

En la introducción de su principal obra *El lenguaje de los nuevos medios de comunicación*, Manovich (2002) hace una interesante reflexión: "¿Tiene sentido teorizar sobre el presente cuando parece estar cambiando tan rápido?" (p. 51)

De hecho, con el acelerado ritmo de cambios que presenciamos debemos siempre tener ese cuestionamiento en mente.

Por eso, considero importante destacar algunos cuestionamientos que están en pauta en los temas de desarrollo del 'big data', siempre con la esperanza que sean también de interés para otros investigadores dentro del campo de la comunicación mediática.

5.3.1 Visualización

En el capítulo sobre los elementos estructurantes del concepto de los datos masivos, tratamos de explorar y aclarar las *dimensiones Vs* que, hasta el momento, los investigadores y empresas que tratan de datos masivos utilizan para orientar sus trabajos.

Este trabajo optó por seguir el modelo de Lee (2017), para desarrollar un concepto integrador respecto a las diez principales dimensiones. Aún, otros estudios consultados sugieren la inclusión de nuevas dimensiones de acuerdo con los desafíos y problemas que van surgiendo con el nuevo fenómeno.

Así, el V relativo a la visualización viene cobrando importancia dentro de los usos que se encuentran en el orden del día en las debates, ya que a menudo la forma de visualización condiciona el análisis de los investigadores.

De hecho, las formas de visualización de datos no son un atributo nuevo o exclusivo de los macro datos. Desde que las sociedades modernas empezaron a registrar la realidad, apareció la preocupación de cómo presentar estos datos de la mejor manera. Muchas veces, la única forma de entenderlos es por medio de formas gráficas. Otras veces, la forma gráfica elegida determina lo que los datos quieren decir.

Las formas tradicionales de visualización de datos pueden ser utilizadas, pero no se muestran tan efectivas para ser aprovechadas en el contexto del 'big data', sean gráficos, tablas, diagramas, mapas, infografías, etcétera. Son útiles en imágenes estáticas. "Los métodos de visualización se utilizan para crear tablas y diagramas para entender los datos. La visualización de 'big data' es más difícil que la visualización tradicional de pequeños conjuntos de datos debido a la complejidad de las dimensiones Vs" (Yaqoob et al., 2016, p. 1240).

Por estos motivos, hay una serie de problemas técnicos que derivan las complejidades de las herramientas y usos de técnicas de minería, al exigir un

proceso dinámico y requerir nuevas soluciones basadas en las particularidades de los Vs estructurantes (volumen, variedad y velocidad). Se desprenden, sobre todo, del poder de procesamiento y escalabilidad.

Estos problemas de orden técnico son importantes y no están dentro del foco de nuestro trabajo, pero hay una preocupación de orden ontológico sobre la cuestión de la visualización de datos que sobresale en la medida que observamos algunos debates académicos, como el propuesto por Gitelman (2013) u observado en la mayoría de los trabajos de análisis cultural de Manovich.

Esta preocupación reside en el hecho de que los datos por sí solos a menudo no pueden decir nada y, en el contexto del 'big data', son tan masivos y a veces tan desestructurados, que su acceso y análisis apenas son posibles por las formas de visualizaciones disponibles para los investigadores.

> No sólo los datos son abstractos y agregados, pero también se manipulan gráficamente. Es decir, para ser utilizados como parte de una explicación o como base para la argumentación, los datos normalmente requieren representación gráfica y a menudo implican una cascada de representaciones. Cualquier interfaz es una especie de visualización de datos - piense en cuántas pantallas que usted encuentra todos los días- así como hojas de cálculo, gráficos, diagramas y otras formas gráficas. La visualización de datos amplifica la función retórica de los datos, ya que diferentes visualizaciones son claramente eficaces, bien o mal diseñadas, y cualquier conjunto de datos puede ser visualizado de múltiples formas y por lo tanto ser persuasivo de distintas maneras. (Gitelman, 2013, p. 12)

Por lo tanto, con la adición de una nueva dimensión en nuestra lista, ahora el V de visualización, los investigadores deben considerar que diferentes herramientas, diferentes técnicas de minería y, por consiguiente, diferentes formas de visualizar un mismo conjunto de datos puede llevar al reconocimiento de diferentes patrones; de allí, a distintas conclusiones.

5.3.2 Estructura de los datos

Otro punto importante vinculado con el tema de la visualización que acabamos de discurrir se refiere a la calidad de los datos disponibles para los estudios, pues cuanto mayor sea su desestructuración, más problemática se convierte la correcta visualización de los mismos.

Por eso, además de la preocupación que ya se debe tener con la cuestión de las fuentes de datos, relacionada justamente con la dimensión de veracidad, también puede suceder que muchas veces esos datos estén de tal forma desestructurados que no permitan su manipulación o que puedan llevar a conclusiones erróneas.

Se estima que alrededor del 80% de los datos digitales generados y almacenados están de alguna manera desestructurados. Pueden ser, por ejemplo, documentos diversos, (meta) datos de medios sociales, fotos y vídeos digitales, transmisiones de audio, sensores usados para recopilar información sobre aparatos y, principalmente, contenidos disponibles en la web (Shacklett, 2017).

> Con gran parte de los principales datos sociales ahora en formato de texto y con un cambio central en cómo los datos son adquiridos y reducidos, los académicos necesitarán llegar a nuevos acuerdos sobre lo que constituye descripciones confiables y válidas sobre éstos; las categorías utilizadas para organizar, con las herramientas necesarias para acceder, procesar, y estructurarlos. (Shah, Cappella, & Neuman, 2015, p. 12)

Lee (2017) también va por ese camino y apunta la necesidad de que un proceso de "control de la calidad de los datos necesita ser establecido para que sea posible desarrollar métricas, evaluar la calidad, reparar errores en esos datos, y evaluar un equilibrio entre garantía de calidad y los resultados costos y ganancias" (p. 301).

Esto se convierte en un enorme desafío actualmente para los investigadores, pues de un lado faltan métricas y controles para la estructuración de datos provenientes de fuentes tan diversas como documentos de investigación o la web. Por otro lado,

puede haber un impacto de costos y de tiempo en las investigaciones para estructurar correctamente los bloques de datos.

No sorprende, por lo tanto, que haya una intensa atención centrada principalmente en las preocupaciones de mantener los datos abiertos, estandarizados, fácilmente accesibles y compartidos por la comunidad académica.

5.3.3 Acceso a los datos

Los principales autores citados en este trabajo (Manovich, 2011; Mayer-Schoenberger & Cukier, 2013; Boyd & Crawford, 2012; Lee, 2017) expresan su preocupación que las investigaciones mediante el uso de 'big data' puedan llevar a un nuevo tipo de división de clases entre los investigadores y los diferentes tipos de estudio. Esta división tiende ser vista por dos perspectivas: una cuestión relativa al acceso a los datos y otra relativa a la capacidad para analizarlos (habilidades e infraestructura).

En el primer caso, sobre el acceso a los datos, ya hemos introducido esta cuestión en la sección sobre los datos abiertos/cerrados y púbicos/privados.

Los estudios que utilizan grandes bases de datos procedentes de las plataformas de medios sociales a menudo no logran acceder a muestras completas de un determinado objeto, sea por la restricción al tamaño de la muestra, sea por la restricción a acceder a determinados metadatos sensibles para el desarrollo de estas plataformas.

Manovich (2011) lo confirma, cuando señala que "el acceso limitado a cantidades masivas de datos sociales transaccionales que se están recogiendo es una de las razones por las que gran parte de las ciencias sociales contemporáneas orientadas a datos no son fáciles de hacer en la práctica" (p. 12).

El propio autor, en algunos de sus estudios (Manovich & Yazdani, 2015; Manovich et al., 2015), utilizó grandes cantidades de datos facilitados por esas plataformas

de medios sociales, sea por el acceso, sea por el tratamiento y estructuración de esos datos para posterior análisis en sus estudios.

En el otro lado, surge la cuestión de las habilidades y herramientas para que los investigadores puedan desarrollar investigaciones mediante datos masivos. Las herramientas siempre fueron una cuestión importante para disparidades de este tipo en investigaciones de tecnología, pero las habilidades se deben encontrar naturalmente dentro de las aptitudes del investigador, pues son puntos en común en una comunidad científica.

> Las ciencias sociales computacionales, con su enfoque en datos a gran escala y datos de medios sociales, precipitar otros cambios en los modos de producción y entrenamiento de los investigadores, algunas obvias y otras no tanto. Muchos de los datos de las ciencias sociales computacionales son y serán textuales, y requieren habilidades de perfeccionamiento en el procesamiento de lenguaje natural. Los científicos sociales cuantitativos están acostumbrados a datos numéricos, recogidos a través de respuestas autorreferidas o en evaluaciones de instrumentos formales. (Shah, Cappella, & Neuman, 2015, p. 12)

Por eso, vemos que la cuestión del cambio de paradigma, desarrollada en el capítulo 4, también se muestra en este caso, cuando hay una mudanza en el conjunto de habilidades que una comunidad necesita compartir.

5.3.4 Privacidad y seguridad

Por último, permanecen cuestiones críticas que desde siempre acompañan las discusiones sobre la manipulación de datos personales y surgen nuevas cuestiones dentro del contexto del uso de 'big data'; en especial, relativas a la privacidad y seguridad de individuos y sus (meta) datos.

Los métodos y aparatos de datificación se han tornado más sofisticados y personales desde la popularización de los *smartphones*, como ya visto. Muchas veces, es difícil separar los datos que se obtienen de un aparato de los datos personales de aquellos que usan ese aparato, que conduce a un proceso delicado desde el punto de vista de la privacidad.

Por otro lado, se trata de cuestiones relativas a la seguridad de estos datos por medios de las instituciones y empresas que los posee. Esta preocupación ha aumentado en la medida en que nuevos casos de violación y fuga de datos personales son publicados con cada vez con mayor frecuencia[29].

Muchos de los servicios y aplicaciones que hacemos *online* se basan en el análisis de los datos de los clientes para el desarrollo de sus productos. Lee (2017) apunta que "proteger la privacidad es a menudo contraproducente para ambos, tanto empresas como clientes, ya que el 'big data' es una clave para mejorar la calidad de los servicios y reducir los costos" (p. 301).

En el campo de las investigaciones académicas, principalmente las mediadas por el uso de los datos masivos y técnicas de minería a partir de datos 'públicos' de blogs, páginas de internet o plataformas de medios sociales, también son objetivas las reflexiones por la forma que se exploran. Es en ese punto que cuestionan Boyd & Crawford (2012):

> ¿Por qué alguien debería ser incluido como parte de un gran conjunto de datos? ¿Y si la entrada de un blog público es tomada del contexto y analizada de una manera que el autor nunca imaginó? ¿Qué significa para alguien ser analizado sin saberlo? ¿Quién es responsable de cerciorarse de que los individuos y las comunidades no se vean perjudicados por la investigación en curso? ¿Cómo se informa el consentimiento a esas personas? (p. 672)

Esas y muchas otras cuestiones que van siendo planteadas por investigadores y que aún no se han señalado respuestas y prácticas convincentes para tratar el tema. En este sentido, Mayer-Schoenberger & Cukier (2013) apuntan la necesidad de individuos competentes y métodos para auditar la autorización y estructuración de los datos, las técnicas de minería y el funcionamiento de los algoritmos como forma de prevenir problemas de naturaleza ética.

[29] En los últimos años estos problemas de seguridad han alcanzado a todo tipo de empresas y datos (Information is Beautiful, 2018).

Pero, otros autores son ascéticos a respecto de estas soluciones. Van Dijck (2017) así alerta que

> las relaciones entre las empresas de datos y las agencias estatales de inteligencia muestran cómo los expertos técnicos circulan entre empleos en la academia y la industria y se transfieren de empresas de datos a servicios financieros o agencias de inteligencia. Los intereses de las corporaciones, de la academia y de las agencias estatales convergen de varios modos. (p. 50)

Debido a esto, las reutilizaciones por los investigadores pueden ser problemáticas, pues retiran los datos de su contexto original por los que se generaron. Por ese motivo, "la gestión del contexto a la luz de la gran fecha será un desafío continuo" (Boyd & Crawford, 2012, p. 671).

6. Conclusiones

Esta investigación se propuso el reconocimiento de las características del universo del 'big data' y cómo la normalización de este fenómeno es debatida en la esfera académica frente a sus consecuencias dentro de las disciplinas relativas al estudio de las comunicaciones mediáticas.

Teniendo en mente la pregunta de Manovich sobre el porqué de escribir a respecto de fenómenos que parecen cambiar aceleradamente, intentamos concluir este recorrido también con el objetivo de preparar un documento propedéutico que retrate el debate en curso sobre dicho fenómeno. Es que todo estudio en tecnologías de comunicación, en los días actuales, se vuelve en pocos años un documento histórico, por causa de los cambios acelerados que ocurren.

Así, el primer paso fue determinar y articular los elementos que considero estructurantes de ese fenómeno, para pasar a la exploración del modelo de conceptualización por medio de las *dimensiones Vs*. A pesar de que éstas son útiles para delimitación de características y aspectos importantes en la manipulación de los datos masivos en las aplicaciones prácticas, no hay consensos en esas dimensiones que permitan establecer una epistemología única para la variedad de prácticas y enfoques en las más diversas disciplinas.

El segundo paso, por lo tanto, fue observar el 'big data' como un fenómeno mucho más amplio, con aspectos culturales, tecnológicos e ideológicos que necesitan una mirada integral sobre sus impactos en nuestra sociedad, al ser apuntada como la base de un cambio de paradigma tecno-científico profundo por algunos autores.

Esta mudanza de paradigma parece ocurrir en la medida en que profundizamos los puntos conceptuales de ese cambio, contextualizado como resultante de un período histórico marcado por el auge del proceso de digitalización y del alto nivel de convergencia de esas tecnologías digitales, ejemplificada por la soberanía del *smartphone*, que lleva, entre otras cosas, al actual estado de normalización de la datificación de nuestras vidas.

A su vez, se suman las nuevas técnicas de análisis de datos a través de la minería y el uso de algoritmos que permiten la extracción de conocimiento basados en patrones que nos posibilitan ver y explicar los fenómenos por medio de sus correlaciones, abriendo espacio para análisis descriptivas y, principalmente, predictivas, más complejas y profundas.

Este conjunto de cambios lleva a un salto cualitativo en la forma de producir conocimiento; algunos autores lo llaman de *Data Driven Science* o 'ciencias orientadas por datos', permeando todas las disciplinas, alterando la forma en que estas comunidades académicas pueden concretizar sus problemas y soluciones para sus objetos de estudio.

Así, en el debate en curso que tratamos de retratar en este trabajo hay una línea ideológica *tecnoutopista*, que ve en estas modificaciones una forma más completa de analizar fenómenos sociales y naturales; una serie de investigaciones empíricas se está aplicando con éxito en las más diferentes áreas.

Por otro lado, investigadores de corrientes *tecnofóbicas*, principalmente los alineados por la visión crítica, identifican serios problemas conceptuales en la aplicación de esas tecnologías, cuando ven en el fenómeno del 'big data' un discurso ideológico que tiende a perpetuar las estructuras jerárquicas de nuestras sociedades.

En este sentido, el principal punto de crítica se refiere a que toda forma de aplicación de 'big data' que tenga su origen en un proceso de datificación de aspectos sociales involucrando datos personales, plantea cuestiones éticas y epistemológicas inescapables que deben ser objeto de preocupación por parte de los investigadores.

Identificamos cómo estos aspectos detonan cuestionamientos importantes sobre la objetividad y neutralidad de los datos obtenidos de personas en su cotidiano, sobre el consentimiento de esas para con la obtención de esos datos y la falsa

representatividad de muestras en las investigaciones que llevan a reflexiones sobre privacidad que están en el orden del día.

¿Hasta qué punto la disponibilidad de grande cantidad de información al respecto de una práctica social cotidiana puede determinarla y explicarla positivamente? ¿La cantidad necesariamente determina la calidad? Son planteamientos aún no resueltos satisfactoriamente y que no debemos perder de vista.

Estas cuestiones, por lo tanto, impactan contradictoriamente las investigaciones en el campo de las comunicaciones mediáticas, especialmente en lo que se refiere a los estudios de las plataformas de redes sociales que, por una parte, proporcionan información valiosa para los investigadores, lo que antes no era posible. Por otra, son habilitadas por estos procesos de datificación y minería de datos que no logran solucionar estos cuestionamientos éticos.

Todo este escenario revela una serie de implicaciones en los estudios en comunicaciones mediáticas, mejor visibles en la línea de los estudios empíricos, pero que también impactan estudios de naturaleza hermenéutica y crítica.

Otras dos implicaciones reveladas en este camino se refieren, primero, a los cambios instrumentales que plantean una actualización y dominio de nuevas habilidades por parte de los investigadores, como conocimientos informáticos, estadísticos, analíticos y narratológicos que exigen cada vez más de ellos en el contexto de la 'ciencia orientada por datos'.

En segundo, de manera más amplia, es tendencia el establecimiento de una nueva forma de estructura de los medios de comunicación, a nivel local y global, pautada no más por la producción de contenido exclusivamente, sino también por el análisis y valor de los (meta) datos involucrados en transmisión y consumo de ese contenido por parte de las audiencias. Este fenómeno es reciente y pide nuevas investigaciones para que se entienda mejor su funcionamiento, ya que no era el enfoque principal de este trabajo.

Por último, identificamos algunos importantes desafíos actuales que deben ser mejor explorados y resueltos por los investigadores, como la forma de visualización de los datos en contexto de 'big data', la estructuración de los datos para optimización de todo este proceso, además de cuestionamientos de orden ético, de privacidad y de acceso a estos datos que aún no presentan una solución satisfactoria.

Así, a pesar de las deficiencias y de algunas limitaciones encontradas en el camino, tales como el escaso tiempo de investigación que se nos impone y la imposibilidad de trabajar empíricamente con macro datos (que nos podría aportar más claridad sobre sus impactos), nuevas posibilidades e inquietudes surgen.

Hay caminos y aspectos poco explorados o dejados en abierto que deben ser tenidos en cuenta, abriendo la puerta para posibles estudios; por ejemplo, el 'big data' como nuevo fundamento de las industrias creativas, quizás la base para un futuro proyecto de doctorado.

Dicho esto, por lo tanto, el fenómeno del 'big data' es complejo, contemporáneo y trae innumerables desafíos que deberán ser aún más investigados, para extender sus potencialidades sin dañar el tejido social en el que estamos inmersos.

Referencias Bibliográficas

Agencia Efe. (2017, 08 de Diciembre). "Netflix es un 50 % análisis de datos y un 50 % creatividad": Greg Peters. *El Espectador.* Recuperado de https://www.elespectador.com/entretenimiento/medios/netflix-es-un-50-analisis-de-datos-y-un-50-creatividad-greg-peters-articulo-727361

Agencias RTVE. (2018, 12 de Febrero). Unilever amenaza con retirar su publicidad de plataformas como Facebook y Twitter si no combaten las noticias falsas. *RTVE.* Recuperado de http://www.rtve.es/noticias/20180212/unilever-amenaza-retirar-su-publicidad-plataformas-digitales-no-combatan-noticias-falsas-discursos-odio/1677540.shtml

Anderson, C. (2008, 23 de Junio). The end of theory: the data deluge makes the scientific method obsolete. *WIRED.* Recuperado de https://www.wired.com/2008/06/pb-theory/

Arcila-Calderón, C., Barbosa-Caro, E. & Cabezuelo-Lorenzo, F. (2016). Técnicas big data: análisis de textos a gran escala para la investigación científica y periodística. *El profesional de la información*, v. 25 (n. 4), pp. 623-631. doi: 10.3145/epi.2016.jul.12

Arsenault, A. H. (2017). The datafication of media: Big data and the media industries. *International Journal of Media & Cultural Politics*, v. 13 (n. 1&2), pp. 07-24. doi: 10.1386/macp.13.1-2.7_1

Avelino, R., Silveira, S. A. & Souza, J. (2016). A privacidade e o mercado de dados pessoais. *Liinc em Revista*, v. 12 (n. 2), pp. 217-230. doi: 10.18617/liinc.v12i2.902

Banco Mundial. (2016). Datos. Recuperado de https://datos.bancomundial.org/indicador/sp.pop.totl/

Beer, D. (2016). How should we do the history of Big Data?. *Big Data & Society*, v. 03 (n. 1), pp. 01-10. doi: 10.1177/2053951716646135

Boyd, D. & Crawford, K. (2012). Critical questions for Big Data: provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, v. 15 (nº 05), pp. 662-679. doi: 10.1080/1369118X.2012.678878

Brito, A. R. & Chico, L. G. (2013). Big Data y periodismo en el continente americano. Cinco casos de estudio. *Revista TELOS.* nº 95, pp. 01-10. Recuperado de https://telos.fundaciontelefonica.com/url-direct/pdf-generator?tipoContenido=articuloTelos&idContenido=2013062110110002&idioma=es

Brooks, C. F. (2018, 15 de Marzo). It´s time for Facebook to share more data with researchers*. Wired.* Recuperado de https://www.wired.com/story/its-time-for-facebook-to-share-more-data-with-researchers/

Castells, M. (1997). *A Era da Informação: Economia, Sociedade e Cultura Vol. 1 - O Poder da Identidade*. São Paulo, Brasil: Ed. Paz e Terra.

Ching, A. (2017, 05 de Enero). 8 Useful Databases to Dig for Data (and 100 more). *Pik To Chart.* Recuperado de https://piktochart.com/blog/8-useful-databases-to-dig-for-data/

Coutinho, I. (2018, 10 de Junio). Google Duplex soa quase a humano*. Publico.* Recuperado de https://www.publico.pt/2018/06/10/tecnologia/noticia/google-duplex-soa-quase-a-humano-1833839

Crawford, K. (2009). Following you: Disciplines of listening in social media. *Continnum Journal of Media & Cultural Studies.* Vol. 23 (nº 01), pp. 525-535. doi: 10.1080/10304310903003270

Crawford, K. (2016, 25 de Junio). Artificial Intelligence's White Guy Problem. *New York Times.* Recuperado de https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html

Crawford, K., Miltner, K. & Gray, M. L. (2014). Critiquing Big Data: Politics, Ethics, Epistemology. *International Journal of Communicatio*n, Vol. 8, pp. 1663-1672.

Crucianelli, S. (2012). Introducción al Periodismo de Datos. *Knight Center for Journalism in the Americas.* Recuperado de https://knightcenter.utexas.edu/es/00-13538-inscribete-ahora-al-curso-introduccion-al-periodismo-de-datos-en-espanol

Curtis, A. (2011). *All watched over by machines of loving grace* [serie de television]. Reino Unido: BBC Two

Datta, A., Datta, A. & Tschantz, M. C. (2015). Automated Experiments on Ad Privacy Settings.

A Tale of Opacity, Choice, and Discrimination. *Proceedings on Privacy Enhancing Technologies*, v. 2015 (n. 1), pp. 92-112. doi: 10.1515/popets-2015-0007.

Dediu, H. (2012, 17 de Enero). The rise and fall of personal computing. *Asymco*. Recuperado de: http://www.asymco.com/2012/01/17/the-rise-and-fall-of-personal-computing/

Eco, U. (2001). *Cómo se hace una tesis*. Barcelona, España: Gedisa Editorial

Facebook (2018). *Company Info.* Recuperado de https://newsroom.fb.com/company-info/

Fernández, A. (2013, 29 de Septiembre). La verdadera (y fascinante) historia del algoritmo de Google*. El Diario.* Recuperado de

https://www.eldiario.es/turing/algoritmo-google-pagerank-matematicas-nodos_0_179882073.html

Finn, E. (2017). *What Algorithms Want. Imagination in the Age of Computing*. Londres, UK: The MIT Press.

Forbes (2018). The World's Most Valuable Brands. *Forbes*. Recuperado de https://www.forbes.com/powerful-brands/list/#tab:rank

Gantz, J. & Reinsel, D. (2012). *THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East*. IDC View. Recuperado de: https://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf

GDPR (2016). Principles relating to processing of personal data. *Intersoft Consulting.* Recuperado de https://gdpr-info.eu/art-5-gdpr/

Gitelman, L. (Ed.) (2013). *"Raw Data" is an Oxymoron*. London, Inglaterra: The MIT Press

Gómez, N. D., Méndez, E. & Hernández-Pérez, T. (2016). Social sciences and humanities research data and metadata: A perspective from thematic data repositories. *El profesional de la información*, v. 25 (n. 4), pp. 545-555. doi: 10.3145/epi.2016.jul.04

Google (2011). *Inside Search.* Recuperado de https://www.google.co.in/insidesearch/howsearchworks/algorithms.html

Grassi, A., Freitas, A., Contarato, A., Taboada, C., Carvalho, D., Ferreira, H.,...Traumann, T. (2017). Robôs, redes sociais e politica no Brasil*. FGV DAPP.* Recuperado de http://dapp.fgv.br/robos-redes-sociais-e-politica-estudo-da-fgvdapp-aponta-interferencias-ilegitimas-no-debate-publico-na-web/

Han, J., Kamber, M & Pei, J. (2012). *Data Minig: Concepts and technique. 3th Edition*. EE.UU: Ed. Morgan Kauffmann

Harari, Y. N. (2015). *Homo Deus*. São Paulo, Brasil: Editora Companhia das Letras

Hernández-Pérez, T. (2016). En la era de la web de los datos: primero datos abiertos, después datos masivos. *El profesional de la información*, v. 25 (nº 04), pp. 517-525.

Hilbert, M. y López, P. (2012). How to Measure the World's Technological Capacity to
Communicate, Store, and Compute Information Part I: Results and Scope. *International Journal of Communicatio,* vol. 06, pp. 956–979

IBM (2016). Entenda porque o Big Data é o petróleo do século 21. *IBM*. Recuperado de https://www.ibm.com/blogs/robertoa/2016/03/entenda-porque-o-big-data-e-o-petroleo-do-seculo-21/

Information is Beautiful (2018). World's Biggest Data Breaches*. Information is Beautiful.* Recuperado de http://www.informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, v. 01 (n. 1), pp. 01-12. doi: 10.1177/2053951714528481

Kitchin, R. & McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological
characteristics of 26 datasets. *Big Data & Society*, v. 03 (n. 1), pp. 01-10. doi: 10.1177/2053951716631130

Kuhn, T. (1962). *La estructura de las revoluciones científicas*. Argentina: FCE. 8ª Edición [2004]

Laney, D. (2001). 3D Data Management: Controlling data volume, velocity, and variety. *META Group*. Application Delivery Strategies, file 949. Recuperado de ttp://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Lee, I. (2017). Big data: Dimensions, evolution, impacts, and challenges. *Business Horizons.* Vol. 60 (nº 01), pp. 293-303. EE.UU., Elsevier. doi: 10.1016/j.bushor.2017.01.004

Leinweber, D. (2012, 24 de Julio). Stupid Data Miner Tricks: How Quants Fool Themselves And The Economic Indicator In Your Pants*. Forbes.* Recuperado de https://www.forbes.com/sites/davidleinweber/2012/07/24/stupid-data-miner-tricks-quants-fooling-themselves-the-economic-indicator-in-your-pants/#22a196fa3c40

Llorens, A. (2017, 03 de Abril). Sólo el 1% de los datos que se generan en internet pueden ser analizados*. Futuro a Fondo.* Recuperado de http://www.futuroafondo.com/es/noticia/solo-1-de-datos-que-se-generan-en-internet-pueden-ser-analizados

Lowrie, I. (2017). Algorithmic rationality: Epistemology and efficiency in the data sciences. *Big Data & Society*, v. 04 (n. 1), pp. 01-13. doi: 10.1177/2053951717700925

Lupton, D. (2015, 11 de Mayo). The thirteen Ps of big data*. This Sociological Life.* Recuperado de https://simplysociology.wordpress.com/2015/05/11/the-thirteen-ps-of-big-data/

Madrigal, A. C. (2011, 17 de Septiembre). Weekend Poem: All Watched Over by Machines of Loving Grace*. The Atlantic.* Recuperado de https://www.theatlantic.com/technology/archive/2011/09/weekend-poem-all-watched-over-by-machines-of-loving-grace/245251/

Mager, A. (2014). Defining Algorithmic Ideology: Using Ideology Critique to Scrutinize Corporate Search Engines. *TripleC*, v. 12 (n. 1), pp. 28-39.

Mayer-Schoenberger, V. & Cukier, K. (2013). *Big Data. La revolución de los datos masivos*. Madrid, España: Turner Noema Publicaciones

Manovich, L. (2002). *El lenguaje de los nuevos medios*. Madrid, España: Paidos Iberica.

Manovich, L. (2011). Trending: The Promises and the Challenges of Big Social Data. *Manovich Net.* Recuperado de http://manovich.net/content/04-projects/067-trending-the-promises-and-the-challenges-of-big-social-data/64-article-2011.pdf

Manovich, L., Ushizima, D., Margolis, T. & Douglass, J. (2015) Cultural Analytics of Large Datasets from Flickr. *AAAI Technical Report.* Vol. 12 (nº 03), pp. 30-34

Manovich, L. & Yazdani, M. (2015). Predicting social trends from non-photographic images on Twitter. *2015 IEEE International Conference on Big Data (Big Data).* doi: 10.1109/BigData.2015.7363935

Mittermeier, J. (2017). *Desmontando la posverdad. Nuevo escenario de las relaciones entre la política y la comunicación* (Tesis de Máster). Universitat Autonoma de Barcelona, Cerdanyola del Valles, España

Moreno, G. (2018, 08 de Mayo). Se envían 65.000 millones de mensajes de WhatsApp al día. *Statista.* Recuperado de https://es.statista.com/grafico/13779/se-envian-65000-millones-de-mensajes-de-whatsapp-al-dia/

Negroponte, N. (1995). *Ser Digital*. Buenos Aires, Argentina: Editorial Atlántida

NIST (2015). Big data interoperability framework: Vol. 1, Definitions. Big Data Public Working Group, Definitions and Taxonomies Subgroup. *NIST*. Recuperado de http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-1.pdf

OKI (n.d.). What is Open?. *Open Knowledge International.* Recuperado de https://okfn.org/opendata/

Pou, J. A. (2017). *Paseando por una parte de la Historia: Antología de citas*. Madrid, España: Penguin Random House Grupo Editorial.

Puschmann, C. & Burgess, J. (2014). Metaphors of Big Data. *International Journal of Communicatio*n, Vol. 8, pp. 1690-1709.

Renó, D. & Renó, L. (2015). Las nuevas redacciones, el 'Big Data' y los medios sociales como fuentes de noticias. *Estudios sobre el Mensaje Periodístico.* Vol. 21 (nº 01), pp. 131-142. Madrid, Ediciones Complutense. doi: 10.5209/rev_ESMP.2015.v21.51135

Rodríguez, E. M. F. (2016). El periodismo de Datos en España. *Estudios sobre el Mensaje Periodístico.* Vol. 22 (nº 01), pp. 255-272. Madrid, Ediciones Complutense.

Sabouni, H. (2018). The Rhythm of Markets. *Research Gate*. doi: 10.13140/RG.2.2.31484.64646

Salesforce (2018). O que é a 4ª Revolução Industrial?. *Salesforce*. Recuperado de: https://www.salesforce.com/br/blog/2018/Janeiro/O-que-e-Quarta-Revolucao-Industrial.html

Shacklett, M. (2017, 14 de Julio). Unstructured data: A cheat sheet. *TechRepublic.* Recuperado de https://www.techrepublic.com/article/unstructured-data-the-smart-persons-guide/

Shafer, T. (2017, 01 de Abril). The 42 V's of Big Data and Data Science. *Elder Research.* Recuperado de https://www.elderresearch.com/blog/42-v-of-big-data

Shah, D. V., Cappella, J. N. & Neuman, W. R. (2015). Big Data, digital media and computational social science: Possibilities and perils. *The ANNALS of the American Academy of Political and Social Science*, vol. 659, pp. 6-13. doi: 10.1177/0002716215572084

Shahin, S. (2016). A critical axiology for Big Data studies. *Palabra Clave*, vo. 19 (nº 04), pp. 972-996. doi: 10.5294/pacla.2016.19.4.2

Shahin, S. (2016b). When scale meets depth: Integrating natural language processing and textual analysis for studying digital corpora. *Communication Methods and Measures*, Vol. 10 (nº 01), pp. 28-50. doi: 10.1080/19312458.2015.1118447

Software Studies Initiative. (2008). Sobre. *Software Studies.* Recuperado de http://lab.softwarestudies.com/2008/05/sobre-software-studies.html

Stanley, M. (2013). Where Is That Moon, Anyway? The Problem of Interpreting Historical Solar Eclipse Observations. En Lisa Gitelman (Ed.) *"Raw Data" is an Oxymoron* (pp. 77-88). London, Inglaterra: The MIT Press

Tufecki, Z. (2017, Septiembre). We're building a dystopia just to make people click on ads [archivo de video]. *TED Global Talks: NYC.* Recuperado de https://www.ted.com/talks/zeynep_tufekci_we_re_building_a_dystopia_just_to_make_people_click_on_ads#t-1363291

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind, New Series*, *Vol. 59* (nº 236), pp. 433-460. Recuperado de http://phil415.pbworks.com/f/TuringComputing.pdf

Van Dijck, J. (2017). Confiamos nos dados? As implicações da datificação para o monitoramento social. *MATRIZes*, v. 11 (nº 01), pp. 39-59. doi: 10.11606/issn.1982-8160.v11i1p39-59

Veltri, G. A. (2017). Big Data is not only about data: The two cultures of modelling. *Big Data & Society*, v. 04 (n. 1), pp. 01-06. doi: 10.1177/2053951717703997

Viner, K. (2016, 12 de Julio). How technology disrupted the truth*. The Guardian.* Recuperado de https://www.theguardian.com/media/2016/jul/12/how-technology-disrupted-the-truth

Wells, H. G. (1938). *World Brain*. London: Methuen & Co., Ltd.

Yakoob, I., Hashem, I. A. T., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N. B. & Vasilakos, A. V. (2016). Big data: From beginning to future. *International Journal of Information Management.* Vol. 36 (nº 01), pp. 1231-1247. doi: 10.1016/j.ijinfomgt.2016.07.009

Zeitchik, S. (2018, 13 de Junio). Comcast makes $65 billion offer for 21st Century Fox, setting up bidding war with Walt Disney*. The Washington Post.* Recuperado de https://www.washingtonpost.com/news/business/wp/2018/06/13/comcast-prepares-to-bid-for-fox-setting-up-a-media-smackdown/?utm_term=.5b68dd7f621a