



Universidad Nacional Mayor de San Marcos

Universidad del Perú. Decana de América

Dirección General de Estudios de Posgrado
Facultad de Letras y Ciencias Huamanoas
Unidad de Posgrado

**Epistemología de la Ciencia de Datos. Una revisión
crítica del concepto de evidencia empírica en la
investigación con Metadatos**

TESIS

Para optar el Grado Académico de Magíster en Filosofía con
mención en Epistemología

AUTOR

José Carlos CHAHUARA QUISPE

ASESOR

Mg. Álvaro Arturo REVOLLEDO NOVOA

Lima, Perú

2024



Reconocimiento - No Comercial - Compartir Igual - Sin restricciones adicionales

<https://creativecommons.org/licenses/by-nc-sa/4.0/>

Usted puede distribuir, remezclar, retocar, y crear a partir del documento original de modo no comercial, siempre y cuando se dé crédito al autor del documento y se licencien las nuevas creaciones bajo las mismas condiciones. No se permite aplicar términos legales o medidas tecnológicas que restrinjan legalmente a otros a hacer cualquier cosa que permita esta licencia.

Referencia bibliográfica

Chahuara, J. (2024). *Epistemología de la Ciencia de Datos. Una revisión crítica del concepto de evidencia empírica en la investigación con Metadatos*. [Tesis de pregrado, Universidad Nacional Mayor de San Marcos, Facultad de Letras y Ciencias Humanas, Unidad de Posgrado]. Repositorio institucional Cybertesis UNMSM.

Metadatos complementarios

Datos de autor	
Nombres y apellidos	José Carlos Chahuara Quispe
Tipo de documento de identidad	DNI
Número de documento de identidad	40604930
URL de ORCID	https://orcid.org/0000-0002-8003-8804
Datos de asesor	
Nombres y apellidos	Álvaro Arturo Revolledo Novoa
Tipo de documento de identidad	DNI
Número de documento de identidad	08148444
URL de ORCID	https://orcid.org/0000-0002-9170-9968
Datos del jurado	
Presidente del jurado	
Nombres y apellidos	Óscar Augusto García Zárate
Tipo de documento	DNI
Número de documento de identidad	06688834
Miembro del jurado 1	
Nombres y apellidos	Cristian Santos Barturén Castillo
Tipo de documento	DNI
Número de documento de identidad	70435333
Miembro del jurado 2	
Nombres y apellidos	Gian Franco Sandoval Mendoza
Tipo de documento	DNI
Número de documento de identidad	44977168
Datos de investigación	

Línea de investigación	E.2.9.3. Epistemología
Grupo de investigación	No aplica
Agencia de financiamiento	Sin financiamiento
Ubicación geográfica de la investigación	Edificio: Universidad Nacional Mayor de San Marcos País: Perú Departamento: Lima Provincia: Lima Distrito: Cercado de Lima Calle: Av. Carlos Germán Amezaga #375 Latitud: -12.0562533 Longitud: -77.0888471 Se requieren coordenadas, no colocar enlaces. Puedes obtener las coordenadas GD de sitios como https://www.google.com.pe/maps/ https://www.coordenadas-gps.com/ https://www.mapsdirections.info/
Año o rango de años en que se realizó la investigación	Noviembre 2021 – Noviembre 2023
URL de disciplinas OCDE	Filosofía https://purl.org/pe-repo/ocde/ford#6.03.01 Historia y filosofía de la ciencia y la tecnología https://purl.org/pe-repo/ocde/ford#6.03.02

UNIDAD DE POSGRADO
ACTA DE SUSTENTACIÓN DE TESIS DE
GRADO ACADÉMICO DE MAGISTER

Siendo las 10.00 horas del día 15 de junio de dos mil veinticuatro, en la sala de grados, el Jurado de Tesis conformado por los siguientes docentes:

Presidente: Dr. Óscar Augusto García Zárate

Miembro : Mg. Cristian Santos Barturen Castilla

Miembro : Mg. Gian Franco Sandoval Mendoza

Asesor : Mg. Álvaro Arturo Revolledo Novoa

Se reunieron para la sustentación de la tesis titulada **Epistemología de la Ciencia de Datos. Una revisión crítica del concepto de evidencia empírica en la investigación con Metadatos**, presentada por el bachiller José Carlos Chahuara Quispe egresado del programa de posgrado de la maestría en Filosofía con mención en Epistemología.

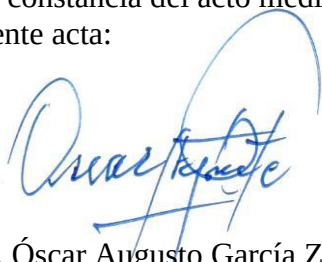
Concluida la sustentación, los miembros del Jurado de Tesis procedieron a formular sus preguntas las que fueron absueltas por el graduando; acto seguido se procedió con la evaluación correspondiente. En forma colegiada, el jurado asignó el calificativo:

Excelente (19) Diecinueve

Aprobada la sustentación de la Tesis, el jurado evaluador recomienda al Consejo de Facultad que se apruebe el otorgamiento del grado académico de magíster en Filosofía con mención en Epistemología a don **José Carlos Chahuara Quispe**

Siendo las 12:15 horas, se levantó la sesión.

Se deja constancia del acto mediante las firmas del jurado de sustentación y asesor de la tesis en la presente acta:



Dr. Óscar Augusto García Zárate
Presidente
Profesor Principal T.C.



Mg. Cristian Santos Barturen Castilla
Informante
Profesor Contratado



Mg. Gian Franco Sandoval Mendoza
Informante
Profesor Contratado



Mg. Álvaro Arturo Revolledo Novoa
Asesor
Profesor Asociado T.C.



Universidad Nacional Mayor de San Marcos

Universidad del Perú. Decana de América

Vicerrectorado de Investigación y Posgrado



CERTIFICADO DE SIMILITUD

Yo Álvaro Arturo Revolledo Novoa en mi condición de Asesor acreditado con el Dictamen N° 015-UPG-FLCH-2023, de la **tesis/monografía/informe de investigación/trabajo académico**, cuyo título es **Epistemología de la Ciencia de datos. Una revisión crítica del concepto de evidencia empírica en la investigación con Metadatos**, presentada por el **bachiller/ magíster/egresado/licenciado/estudiante José Carlos Chahuara Quispe** para optar el grado/título/especialidad de **magíster** en Filosofía con mención en Epistemología, **CERTIFICO** que se ha cumplido con lo establecido en la Directiva de Originalidad y de Similitud de Trabajos Académicos, de Investigación y Producción Intelectual. Según la revisión, análisis y evaluación mediante el software de similitud textual, el documento evaluado cuenta con el porcentaje de 8 % de similitud, nivel **PERMITIDO** para continuar con los trámites correspondientes y para su **publicación en el repositorio institucional**.

Se emite el presente certificado en cumplimiento de lo establecido en las normas vigentes, como uno de los requisitos para la obtención del **grado/ título/ especialidad** correspondiente.

Firma del Asesor

DNI: 08148444

Nombres y apellidos del Asesor

Álvaro Arturo Revolledo Novoa



Huella digital

Dedicatoria

Para Aurelia y José Gabriel, mis padres.

Resumen

El gran desarrollo tecnológico impulsado por las ciencias de la computación ha abierto posibilidades ilimitadas para la recolección de información y para su procesamiento por parte de industrias, gobiernos y científicos. Sin embargo, la gran disponibilidad de insumos informáticos y la flexibilidad en el uso de las técnicas de los metadatos y sus algoritmos, ha creado una realidad en la que la verdad científica parece subordinarse a una serie de opiniones fundadas en el uso más conveniente que las herramientas de la Ciencia de Datos ponen a disposición de los investigadores para encontrar la respuesta más ‘adecuada’ a determinadas hipótesis, creando una ilusión de credibilidad científica fundamentada principalmente en la abundancia de los datos procesados y no en la evidencia empírica. El presente trabajo tiene como principal interés estudiar el concepto de evidencia empírica en las investigaciones científicas que hacen uso de la *Big Data* o Metadatos. Para ello, se presenta inicialmente una definición de *Big Data* aceptada dentro de la comunidad científica, se describen los principales problemas epistemológicos derivados del uso de esta tecnología, para posteriormente ahondar en la naturaleza de la evidencia empírica y en cómo las prácticas de la investigación que utiliza Big Data - tales como la visualización de agrupaciones de datos - parece llevar la noción de evidencia científica a un plano de discusión secundario. Finalmente, se planteará una respuesta al problema epistemológico de la evidencia empírica en Ciencia de Datos a través de la propuesta de una formulación conceptual que considere las características de Big Data en la evidencia empírica para investigación científica.

Palabras Clave: Epistemología, evidencia, evidencia empírica, metadatos, ciencia de datos

Abstract

The great technological development of computer science has opened unlimited possibilities for the collection of information and its processing by industries, governments and scientists. However, the flexibility in the use of metadata and algorithms has created a reality in which scientific truth seems to be subordinated to a series of opinions based on the most convenient use that Data Science tools make available to researchers. to find the most 'appropriate' response to certain hypotheses, creating an illusion of scientific credibility based mainly on the abundance of processed data and not on empirical evidence. The main interest of this work is to study the concept of empirical evidence in scientific research that makes use of Big Data or Metadata. For this, a definition of Big Data accepted within the scientific community is initially presented, the main epistemological problems derived from the use of this technology are described, to later delve into the nature of empirical evidence and how research practices in Big data, such as visualization of data pools, seems to take the evidence to a secondary plane of discussion. Finally, a response to the epistemological problem of empirical evidence in Data Science will be proposed through the proposal of a conceptual formulation that considers the characteristics of Big Data in empirical evidence for scientific research.

Keywords: Epistemology, evidence, empirical evidence, metadata, data science

Índice General

Dedicatoria	V
Resumen	VI
Abstract	VII
Índice General	VIII
Lista de Figuras	X
Lista de Tablas.....	XI
INTRODUCCIÓN.....	1
CAPÍTULO I CIENCIA DE DATOS Y SUS DESAFÍOS EPISTEMOLÓGICOS.....	7
1.1 Información y ciencia de datos	7
1.2 Conceptualización de <i>Big Data</i> y Ciencia de Datos	13
1.3 Características de la <i>Big Data</i> (Volumen, Velocidad, Variedad, Veracidad, Valor)	18
1.4 Impacto de Big Data y la Ciencia de Datos en la Sociedad.....	24
1.5 Impacto de <i>Big Data</i> y de la Ciencia de Datos en la Comunidad Científica.....	27
1.6 Problemas epistemológicos del uso de Big Data en investigación.....	32
1.7 El Problema de la Relación entre Dato y Evidencia Empírica en Big Data	37
CAPÍTULO II EVIDENCIA EMPÍRICA Y METADATOS	41
2.1 ¿Qué es la evidencia empírica?.....	41
2.2 Datos y Evidencia Empírica	48
2.2.1. De la medición a los datos: el cambio epistemológico	51
2.2.2. Un nuevo paradigma: De los datos a la visualización	54
2.3 Teorías y Modelos de la Evidencia.....	58
2.4 Evidencia empírica y <i>Big Data</i> en la ciencia contemporánea	66
2.4.1 Explicacionismo.....	69
2.4.2 Bayesianismo o Experimentalismo	69

CAPÍTULO III DISCUSIÓN Y CONCLUSIONES.....	73
3.1 Consideraciones para una nueva formulación conceptual de ‘evidencia empírica’	73
3.2 La autoridad de los métodos y algoritmos en la Ciencia de Datos	79
3.3 Un enfoque epistemológico para la Ciencia de Datos	81
3.3.1 La investigación con Ciencia de Datos y la hipótesis de investigación.....	82
3.3.2 El Rol de las Teorías en la Ciencia de Datos	85
3.3.3 Causalidad en la ciencia de datos.....	86
3.4 Hacia un nuevo tipo de conocimiento. La epistemología de los datos.	86
BIBLIOGRAFÍA.....	89

Lista de Figuras

Figura 1. La pirámide de conocimiento.....	10
Figura 2 - Diagrama de Venn de Conway	14

Lista de Tablas

Tabla 1: Técnicas de procesamiento de datos	56
---	-----------

INTRODUCCIÓN

El desarrollo tecnológico de las computadoras, los algoritmos y los metadatos ha abierto posibilidades casi ilimitadas para la recolección y procesamiento de información en el mundo actual; posibilidades que eran inimaginables hace tan sólo 50 años. La aplicación de estas tecnologías se enmarcan dentro del *corpus* científico como parte de las Ciencias de la Computación, y dentro de ellas, en la disciplina de la Ciencia de Datos, la cual se ha convertido no sólo en un campo de interés por mérito propio, sino también en una herramienta fundamental para realizar investigación en otras diversas áreas de la ciencia.

Así mismo, la gran cantidad de lenguajes de programación, metodologías, procedimientos y enfoques para la solución de problemas, han permitido que exista gran flexibilidad en cuanto a la forma de uso de los datos y de los algoritmos por parte de los investigadores, creando una instancia de la realidad en la que, de forma similar a como se argumenta en el postmodernismo, no hay una verdad, sino una serie de opiniones fundadas en el uso más conveniente de la Ciencia de Datos, y sus herramientas, para encontrar las respuestas más ‘adecuadas’ a determinadas hipótesis de investigación, creando una ilusión de credibilidad que estaría basada sólo en la abundancia de los datos (Raoult, 2020) que fueron utilizados para realizar la investigación y en el uso de la propia tecnología, pero no necesariamente teniendo como sustrato algún claro elemento teórico que tenga como fundamento la epistemología.

Hay casos muy destacables en cuanto al mal uso que se puede dar a las técnicas de la Ciencia de Datos como medio para expresar un determinado nivel de certeza respecto a una noción de conocimiento, ya sea científico y no científico. Por ejemplo, hace pocos años salió a la luz un caso de estudio que pretendía demostrar que la sustancia conocida como hidroxiclороquina era un elemento eficaz para el tratamiento del Covid, sin embargo, este estudio tuvo que ser retirado por la prestigiosa revista científica Lancet pues los investigadores no permitieron que su trabajo, basado en técnicas de *Big Data*, fuese replicado, al negarse a compartir las bases de datos utilizadas (Mehra, Ruschitzka, & Patel, 2020). Otro ejemplo relevante se dio en el Perú, con el caso del criptoanalista Arturo Arriagán quien presentó, en televisión nacional, un modelo matemático basado en Ciencia de Datos para demostrar la supuesta existencia de fraude en la última elección

presidencial peruana, modelo que fue calificado por varios representantes de la comunidad científica como insuficiente y mal elaborado. Este caso tuvo mucha relevancia porque en ese momento la atención nacional estaba completamente centrada en la posibilidad de fraude, y una presentación tan pobre sólo logró desprestigiar no sólo al propio criptoanalista, como a las técnicas de las Ciencias de Datos.

Es por motivos similares a los expuestos que la epistemología de la Ciencia de Datos se ha convertido en un tema de interés en la comunidad científica y filosófica desde hace algo más de una década. Los investigadores, ya sean filósofos o científicos, que se han interesado por este tema han tenido diversos tipos de aproximaciones para su estudio, algunos muy relevantes son:

- La posibilidad de crear teorías científicas universales a través de los metadatos (Fricke, 2013).
- La generación de valor económico de los metadatos en organizaciones empresariales (Günther, Rezazade Mehrizi, Huysman, & Feldberg, 2017), puesto que el valor económico se suele generar como resultado de un proceso de transformación de conocimiento científico y tecnológico de las organizaciones empresariales hacia el mercado (Fontes, 2005).
- El posible fin del método científico a partir del uso de la Ciencia de Datos (Succi & Coveney, 2019) (Mazzocchi, 2015).
- El impacto de la Ciencia de Datos en la sociedad actual (Lee & Cook, 2020) (Ahonen, 2015), cuyo progreso ha estado impulsado en gran medida por el conocimiento científico.
- La influencia de la Ciencia de Datos en el proceso de toma de decisiones (Hirschheim, 2021).
- La poca transparencia que hay al interior de los procedimientos que utilizan Metadatos (Obar, 2020) (Driscoll & Walker, 2014).
- El análisis de técnicas de investigación que sugieren que los estudios con Metadatos cumplen con características de estudios con volumen de datos convencionales (Hekler et al., 2019).
- El análisis de definiciones respecto a qué son los Metadatos dentro del campo de estudios científicos (Favaretto, de Clercq, Schneble, & Elger, 2020) (Jones, 2019).

- El análisis de procedimientos para la aplicación de Metadatos en las investigaciones científicas (Cheung, Leung, & Seto, 2019), y la epistemología de los Metadatos (Canali, 2016).

Es relevante, para la presente discusión, resaltar que conceptos propios de la Ciencia de Datos – que son de uso muy extendido en ámbitos en apariencia tan disímiles como pueden ser los ámbitos científico o el empresarial - no tienen un consenso dentro de la comunidad científica (Balazka & Rodighiero, 2020). El propio concepto de *Big Data*, o Metadatos en idioma castellano, no ha logrado obtener aceptación unánime, pues el mismo término es utilizado para describir la gran magnitud en cuanto al volumen que puede alcanzarse sólo para el almacenamiento de información, pero también es utilizado por otros profesionales dentro de la definición de lo que se llama “procesamiento de datos” que implica un esfuerzo mayor comparado al simple almacenamiento, e incluso el concepto se puede relacionar a las propias diferentes tecnologías con que se realizan aquellos procedimientos, que fungirían como elemento transversal aglutinante de la técnica.

Sin embargo, aunque los términos de Ciencias de Datos o *Big Data* sean calificados de forma vaga y utilizados indistintamente por muchos profesionales, científicos y filósofos, en la práctica, el uso de la tecnología está produciendo cambios en la forma en que se obtiene el conocimiento. Algunos de esos cambios que son de interés destacar (Balazka & Rodighiero, 2020), se listan a continuación:

- 1) El cambio que implica pasar de un ámbito de tener pocos datos al de tener una gran abundancia de datos para realizar investigaciones, lo cual podría tener como consecuencia que los procedimientos de muestreo que son tan relevantes para la investigación científica pasen a ser poco eficientes, e incluso obsoletos y poco útiles.
- 2) El cambio que implica pasar de tener una muestra *n* de datos a tener *todos* los conjuntos de datos disponibles para cualquier investigación, que tiene como consecuencia que las consideraciones metodológicas acerca de la exactitud de los datos sea innecesaria. Exactitud en términos que describen los datos que se usan como aquellos que son los más apropiados para la investigación.

Si se tiene acceso a todos los datos, se podría afirmar que la totalidad de los datos es la relevante para la investigación y no sólo parte de ella.

- 3) El cambio que implica dejar de lado la búsqueda de la causalidad y asumir en su lugar la búsqueda de la correlación, que tiene como consecuencia una modificación muy relevante en el entendimiento del proceso explicativo de la ciencia.

A partir de estas consideraciones iniciales, en términos del interés de los científicos e investigadores, y en términos de los posibles cambios que la Ciencia de Datos y la *Big Data* han empezado a generar en los procedimientos formales de la búsqueda de conocimiento, se puede afirmar con certeza que existe un problema muy actual y relevante en cuanto a la definición de los aspectos epistemológicos de la Ciencia de Datos, independientemente de su aún imprecisa definición, y que este problema merece ser estudiado.

A efectos de la presente investigación, se hace necesario precisar los conceptos de *Big Data* (Metadatos) y Ciencia de Datos. Para tal efecto, a pesar de que autores como Favaretto et al. (2020) consideran que *Big Data* es un fenómeno tanto cultural como técnico, se mantendrá la disertación central dentro del límite técnico en el que *Big Data*, o Metadatos en castellano, es el término que se usa para representar conceptualmente conjuntos masivos de datos de estructura variada y compleja que presenta dificultades para ser almacenada, procesada y analizada (Sagiroglu & Sinanc, 2013). Así mismo, por su característica de ser la intersección de diversas habilidades, conocimientos y técnicas (Conway, 2010), y para no desviar el foco del problema hacia la científicidad de la propia Ciencia de Datos sino mantener la discusión en la científicidad del conocimiento obtenido por las investigaciones que usan Ciencia de Datos, en la presente investigación, el concepto de Ciencia de Datos debe entenderse como el *corpus* de herramientas, procedimientos y tecnologías que permiten realizar la analítica del *Big Data*, es decir, aquellas que permiten el procesamiento, la visualización y la interpretación de los patrones presentes en los datos. Por extensión, las investigaciones basadas en Ciencia de Datos serán aquellas investigaciones que hacen uso tanto de *Big Data*, como de las técnicas que permiten el procesamiento especializado de esos conjuntos de datos. Estas técnicas son propias de la Ciencia de Datos pues requieren el conocimiento de diversas habilidades que pertenecen a los campos de la computación, las matemáticas y la

experiencia del ámbito de investigación en el que se trabaja (i.e. astronomía, biología, medicina, etc.), y se hace referencia a ellas como técnicas de Ciencia de Datos. Es importante entonces hacer la diferenciación entre *la ciencia como el saber obtenido mediante investigación metódica, ordenado y preservado en su propia forma sistemática* (Störig, 2016), y la Ciencia de Datos como conjunto de técnicas y no como ciencia *per se* en su sentido tradicional metodológico.

Dentro de las oportunidades de investigación y reflexión filosófica que abren la muy extendida aplicación de la Ciencia de Datos y la *Big Data* en la actual generación de conocimiento, es de mucho interés en la filosofía de la ciencia, aquella que atañe a la evidencia empírica, puesto que la existencia y la calidad de la evidencia empírica es un tópico ampliamente discutido en epistemología.

En la búsqueda de conocimiento a través de la ciencia es importante la capacidad de confirmación del conocimiento adquirido, incluso a pesar de que no haya certeza respecto a si esa capacidad de confirmación es algo real o no, dados los múltiples casos que se estudian mediante la ciencia. Así, muchos de los objetos de estudio científico son analizados de forma indirecta, a partir de sus efectos, debido a la imposibilidad de estudiarlos a ellos mismos, por ejemplo, a objetos físicamente inalcanzables como el Sol o a personajes inexistentes como Napoleón (Russell, 1988). Sin embargo, a pesar de no tener una experiencia directa con los objetos de estudio, se tiende a creer en la veracidad de las conclusiones obtenidas a través de la ciencia y que a su vez han sido aceptadas por la comunidad científica. De forma más general, más allá de la ciencia, las razones por las que se cree en algo que no se conoce por experiencia propia suelen estar, en muchos casos, relacionadas a la autoridad de la que emana aquello en lo que se cree. Es conocida la anécdota del matemático Isaac Todhunter quien siempre objetó la presencia de estudiantes mientras se realizaban los experimentos en los laboratorios de los cursos que dictaba en la universidad de Cambridge, ya que, en su opinión y por su autoridad, que los resultados hayan sido atestiguados por los profesores era condición necesaria y suficiente para que fuesen considerados irrefutables e infalibles (Russell, 1988).

En la presente tesis se plantea que el actual poder de procesamiento y la ubicuidad de los sistemas computacionales, la *Big Data* y la Ciencia de Datos podrían estar llevando a la ciencia, y a la discusión científica, a un estado en el que la existencia de evidencia y validez empírica pueda quedar subordinada sólo al uso de los algoritmos y técnicas ‘correctas’ para definir si una investigación es rigurosa y válida o no, ignorándose características propias que la evidencia tiene, o debe tener, para la corroboración de algún

evento o fenómeno. En este mismo sentido, se hace interesante también discutir si los datos informáticos son algún tipo de evidencia empírica, incluso a pesar de que su naturaleza física sea la de ser sólo impulsos electromagnéticos almacenados en servidores en alguna parte del mundo, y que, además, son procesados múltiples veces antes de ser utilizados por distintos investigadores, que usan a su vez distintos algoritmos, para usarlos de insumo y generar algún tipo de conocimiento que en muchos casos suele ser aceptado como científico al ser publicado en revistas aceptadas por la comunidad científica. La ventaja de la unicidad que permite que los datos puedan ser utilizados en cualquier parte del mundo y por cualquier científico tiene la capacidad de permitir que la información se modifique tanto que es válido dudar de la existencia o permanencia de algún tipo de evidencia empírica en los datos procesados una y otra vez.

Postulaba el gran filósofo británico Hume (1980) que existen dos clases de conocimiento, el propiamente racional y el empírico, y que el entendimiento de ambos requiere evidencia, o al menos requiere de grados de evidencia que puedan dar solidez al razonamiento demostrativo o al razonamiento probable que concluya efectivamente en conocimiento (Loeb, 1995). El avance tecnológico ha generado entonces el siguiente problema que se puede formular como pregunta: ¿es posible que el uso de la *Big Data* y algoritmos degraden la calidad de evidencia empírica y perjudiquen la capacidad de generar conocimiento fiable a partir de ella?

La respuesta *a priori* es que la ciencia de datos sí debe hacer uso de una nueva concepción del concepto de evidencia empírica que envuelva no sólo el grado de racionalidad o de validez empírica, sino también características respecto al procesamiento que esa información sufre a través de algoritmos que muchas veces son ininteligibles para el investigador o profesional que hace uso de los datos.

CAPÍTULO I CIENCIA DE DATOS Y SUS DESAFÍOS EPISTEMOLÓGICOS

1.1 Información y ciencia de datos

En la actualidad, es muy común escuchar, leer o hablar acerca de la *sociedad de la información o de la sociedad del conocimiento*, etiqueta utilizada para resaltar la importancia, y el impacto, que la revolución tecnológica impulsada por las tecnologías de la información ha ocasionado en el mundo contemporáneo. El uso del término *sociedad* es claro en cuanto a que se refiere al conjunto de actividades mediante las que las personas interactúan y a las particularidades culturales de los días presentes, sin embargo, el término *información* que se utiliza tan frecuentemente, no está tan bien esclarecido, puesto que se suele confundir con otros términos, en apariencia similares, como son *dato*, y *conocimiento*; además, en muchos ámbitos, la palabra *información* suele usarse también como un término que engloba por sí mismo a los otros dos, e incluso como un concepto que describe la conexión entre los datos sin procesar (*raw data*) y el conocimiento que se alcanza a partir de ellos luego de realizar algún tipo de análisis numérico (Davenport & Prusak, 1997), por lo que definir qué es información no es una tarea sencilla.

Peter Drucker (1988), hizo el ejercicio de definir información como un *conjunto de datos provistos de propósito y relevancia*, poniendo al ser humano en el lugar del único ente con la capacidad de proveer a los datos de ambas características. De esta manera, logró vincular a los datos con la inteligencia humana y convirtió ese vínculo en un requisito para que los datos se conviertan en información, independientemente del medio que se utilice para registrar los datos. Por ejemplo, ya sea que se use la computadora más moderna para almacenar datos y registrarlos en una hoja de cálculo, por ejemplo del programa *Excel*, para posteriormente representar esos datos mediante un gráfico que permita mostrarlos visualmente como un histograma o un *pie*, sólo a través de la inteligencia de *alguien* es que se puede tomar decisiones respecto a cómo hacer para ordenar los datos de la mejor manera y cómo elegir la representación visual más adecuada para que la información se pueda transmitir a otras inteligencias de forma clara y útil. De esta manera, la inteligencia de las personas toma el rol y se convierte en el elemento activo que es capaz de transformar los datos en información, y toma parte tanto en la construcción de la información, como en la interpretación de la misma a lo largo del proceso de pasar de inteligencia a inteligencia.

Sin inteligencia no habría información sino únicamente datos, puesto que la información, a diferencia de los datos, requiere que existan y se identifiquen importantes cualidades que permitan su análisis y discusión, por ejemplo, al tratar con un dato simple relacionado a la inflación económica como puede ser el dato de la variación del precio anual de un producto expresado en forma de un porcentaje que compara las tasas de precios del año 2023 y 2022, este dato o número no tiene significado fuera del ámbito de la inteligencia, es necesario y no sólo suficiente que alguien piense algo o diga algo respecto de ese número para que recién ahí el simple dato se convierta en información.

Los datos no son más que símbolos que representan las propiedades de objetos y eventos, y pueden ser números o letras si están expresados en lenguaje humano, o sólo *bits*, impulsos eléctricos que representan unos y ceros si están expresados en algún código o lenguaje computacional. La información está en un nivel más elevado, pues consiste en datos procesados por una inteligencia; en palabras de Ackoff (2005), el procesamiento está dirigido a incrementar la utilidad de los datos, y en ese mismo sentido, es importante considerar que no es de mucha utilidad un simple número o una palabra si una inteligencia no la contextualiza y toma algún tipo de decisión a partir de ese procesamiento. Es decir, el procesamiento no sólo incrementa la utilidad de los datos, sino que hace tangible que éstos sean útiles en el sentido de que pueden ser utilizados de alguna manera por una inteligencia para alcanzar un determinado propósito. Por ejemplo, los gobiernos organizan censos, para ello capturan y registran los datos de la población. En el caso peruano, el INEI (Instituto Nacional de Estadística e Información) contrata personas que recolectan los datos, los ingresan en un sistema computacional y posteriormente procesan estos datos, para convertirlos en información que se representa mediante gráficos y tablas que publica en su página web. Esta información es utilizada luego por el gobierno para tomar decisiones en cuanto a los ámbitos que le competen como niveles de educación, necesidad de inversión en salud, o creación de programas de vivienda o infraestructura, también es utilizada por empresas para tomar decisiones de inversión, por organizaciones internacionales para definir políticas de ayuda social, o también es utilizada por investigadores y estudiosos para responder hipótesis relevantes.

Es importante considerar que tanto los datos como la información representan características y propiedades de objetos y eventos, pero lo hacen de forma distinta, los datos son números sueltos y la información contiene no sólo a los datos sino razonamiento embebido que permite hacer la relación con la realidad de una forma más compacta, legible y útil que como lo hacen los datos. De esta manera a partir de lo expresado en los

últimos párrafos, se puede afirmar con relativa certeza que, en efecto, existe una diferencia relevante entre los conceptos de dato e información, y que esta diferencia es principalmente funcional, no estructural, en cuanto al uso que se obtiene, y se da, por y desde ambos.

Un nivel más allá de la información se encuentra el conocimiento, que también sería información, pero información con el mayor valor, ya que la inteligencia de alguna persona tendría no sólo que haber puesto contexto, significado, interpretaciones particulares en la información de la que dispone, sino que esta información tendría que ajustarse a una definición epistemológica como la que se encuentra en Otero & Gibert (2016) y que explica el conocimiento como lo opuesto a la conjetura o a la opinión, y con la característica de ser una creencia verdadera justificada y donde cada uno de los elementos que la conforman, considerados de forma individual, sean una condición necesaria pero no suficiente para ser verdad. Es decir, la información puede generar creencias y estar contenida en descripciones, respuestas a preguntas que empiezan con palabras como quién, qué, dónde, cómo y cuántos, pero en el conocimiento, alguien tiene que haber reflejado y adicionado no sólo su propia sabiduría en la información, sino que esta debe ser verdadera y justificada.

De esta manera se hace muy claro el postulado respecto a que la cantidad de participación del ser humano, y su compromiso con la verdad, se incrementa conforme se vaya avanzando en la línea ascendente de dato-información-conocimiento. Es así que se logra entender la forma en que las computadoras resultan muy útiles para manejar y procesar los datos, menos útiles para manejar información y aún menos útiles para manejar conocimiento, al menos por ahora, ya que se está invirtiendo esfuerzo en el campo de la inteligencia artificial para dotar de conciencia a las máquinas, lo cual podría posibilitar la creación de conocimiento por parte de las mentes artificiales.

La Figura 1 muestra una referencia a la jerarquía: datos, información y conocimiento que es uno de los modelos más fundamentales y ampliamente reconocidos de la literatura sobre información y conocimiento (Jones, 2019). Parafraseando a Ackoff (2005), se podría afirmar que un gramo de sabiduría equivale a un kilo de conocimiento, este a 10 kilos de información, y este a su vez a 100 kilos de datos.



Figura 1. La pirámide de conocimiento (Jones, 2019)

La importancia de recordar y considerar la inclusión del factor humano como elemento clave cuando se tratan asuntos relacionados al término *información*, se debe a que la fascinación por la tecnología ha hecho que se olvide el propósito clave de toda información, que es *informar a alguien sobre algo*, de forma análoga a como cuando las personas se concentran en ver sólo el dedo que señala algo, y olvidan aquello que se está señalando. No se debe olvidar que todas las computadoras que existen en el mundo dejarían de tener utilidad si es que los usuarios no estuviesen interesados en generar y utilizar información mediante ellas. Por ejemplo, todos los sistemas de tecnologías de la información en los que invierten las empresas no adicionarían ningún tipo de valor a su desempeño si los empleados no generasen y compartiesen la información que generan en sus funciones con sus colegas de las distintas áreas de la organización, así como con sus clientes y sus proveedores. La recolección de datos, la organización de información y la generación de conocimiento son actividades humanas de gran importancia, que sólo pueden ser definidas, entendidas y gestionadas correctamente si las personas tienen el rol primario en su ejecución (Davenport & Prusak, 1997).

Puesto que, a partir de lo descrito, parece lógico que la creación de información a partir de datos es una actividad inherentemente humana e inherentemente contemporánea en la medida en que existen computadoras, es de interés enriquecer tal afirmación recordando nuestra historia, puesto que desde las más antiguas figuras rupestres, utilizadas por los primeros hombres primitivos para expresar algún tipo de conocimiento que lograron desarrollar acerca de la geografía, la fauna, la flora de su entorno cercano

como se ve en los restos artísticos de las cuevas de Altamira o Lascaux, pasando por los jeroglíficos egipcios, los ideogramas chinos y los kipus incas hasta la actualidad, el ser humano independientemente de la ubicación geográfica en la que haya habitado, del tiempo en que haya vivido y del tipo de organización social que haya construido, ha requerido a lo largo de su evolución social, de la acumulación de datos, información y conocimiento que pueda relacionar, procesar y utilizar para sus actividades cotidianas.

Así mismo, a partir de la invención de la imprenta, la mayor cantidad de información y conocimiento que ha sido registrada y transmitida hasta antes de la revolución de las computadoras, se ha conseguido a través del uso de la tinta y el papel, los cuáles, aunque han cumplido un rol muy importante, no dejan de ser un medio poco flexible, de difícil gestión y actualización, y frecuentemente expuesto al deterioro por el paso del tiempo y por los insectos y otras alimañas que se alimentan de este material. Más allá de estas características negativas o poco prácticas, el papel generó una revolución cultural que permitió compartir conocimiento e ideas de forma exponencial comparada con el alcance que se tenía antes. Mucho tiempo después, con la invención de las computadoras, casi 500 años luego de la invención de la imprenta, en la segunda mitad del siglo XX se ha logrado contar con una poderosa herramienta de almacenamiento y procesamiento de información que funciona de forma eficiente y que ha permitido crear sistemas informáticos complejos que han permiten integrar el manejo de los datos, los métodos y la tecnología necesaria para procesar información, generar conocimiento y tomar decisiones a partir de ella.

En la actualidad, el muy extendido uso de computadoras y de los sistemas de tecnología de la información; su constante evolución tecnológica, y su cada vez mayor influencia en la vida de las personas han despertado gran interés filosófico tanto por propia la tecnología de los sistemas de información, como por las consecuencias de su uso. Algunos de los factores que han impulsado tal interés entre los filósofos son (Floridi, 2005):

- i. El factor sociológico, expresado a partir de la existencia de la sociedad post-industrial moderna que está basada en el intercambio y capitalización de la información (*sociedad de la información*), la cual es generada a partir de la masificación de la informática, la democratización del uso de las computadoras, el alcance global de internet, los medio de información masivos globales, las redes

de comunicación móviles y la creación exponencial de aplicaciones (*apps*) que las personas usan en diversos ámbitos de sus vidas.

- ii. El factor científico, expresado a partir del indispensable uso de la informática para el desarrollo de la ciencia y la generación de conocimiento científico. En la actualidad, resultaría muy difícil realizar investigaciones científicas en ámbitos como la astronomía, bioingeniería, la genética, las neurociencias, la física de materiales, la arqueología, entre otras ciencias, sin el soporte computacional que ofrecen los sistemas informáticos, tanto por la capacidad de procesamiento de información, como por la cantidad de datos que son necesarios para llevar la investigación a término, así como por la existencia de *software* especializado para trabajar en cada rama de estudio científico especializada.
- iii. El factor cultural, expresado a través de las nuevas formas de experiencia personal, de comunicación y de socialización que están a disposición de la sociedad a nivel global. Nunca antes se han tenido tantos medios y formas de conseguir todo tipo de información (entretenimiento, ciencia, noticias, artes, etc.), y se han creado tantas posibilidades y espacios para dar a conocer las propias opiniones en foros especializados o abiertos en cualquier área de interés, y de interactuar con amigos, familiares y desconocidos alrededor del mundo, a través del espacio virtual que ha sido creado mediante el uso de las redes de computadoras, las aplicaciones y las redes sociales.

El segundo factor es el de mayor relevancia para la presente disertación, y sobre el que se tratará en los apartados siguientes. Este segundo factor está asociado a otro aspecto de interés que se da en cuanto al avance explosivo de las capacidades de procesamiento de datos de los sistemas computacionales, que ha seguido lo expresado en la Ley de Moore, postulada por Gordon E. Moore, quien predijo en 1965 la duplicación de la capacidad de procesamiento en los microprocesadores cada dos años. Esta mejora tecnológica ha tenido como consecuencia, una cada vez mayor especialización en el área de conocimiento de la electrónica y la informática, así como la aparición de nuevos conceptos y métodos que permitan sistematizar el estudio y desarrollo de las ciencias informáticas. La Ciencia de Datos y los Metadatos (*Big Data* en inglés) hacen parte de ese nuevo corpus de conocimiento, y se han convertido en su conjunto, en un factor que

ha cambiado, a lo largo de la última década, el manejo de datos e información en el mundo, generando, como se ha explicado previamente, gran interés filosófico tanto desde el aspecto tecnológico, ético, como epistemológico.

Finalmente, la Figura 1 muestra gráficamente como los datos están superpuestos al mundo, es decir como una forma de abstracción de lo que sucede en la realidad, por lo que *a priori*, el conocimiento que se encuentra en un nivel superior se fundamenta en la correcta representación que esos datos hagan del mundo, es decir de la realidad empírica, que es el tema principal que trata la presente investigación.

1.2 Conceptualización de *Big Data* y Ciencia de Datos

La Ciencia de Datos y la *Big Data* han encontrado campos de aplicación en diversos sectores de la actividad humana actual como el científico, industrial, financiero, político y el de entretenimiento; creando las condiciones para procesar y analizar datos, información, y generar conocimiento (*insights*) para permitir que la toma de decisiones sea más acertada en cuanto a competencia empresarial, pruebas de hipótesis científicas, estrategias de campaña política, seguridad ciudadana y nacional, análisis de riesgo financiero, rendimiento deportivo, e incluso el posible éxito de una producción discográfica o filmográfica, entre otras actividades sociales y académicas. Sin embargo, a pesar del amplio uso del término, no existe consenso respecto a su definición, principalmente en cuanto al término *Big Data*, que en español se suele representar como Metadatos, y que en primera impresión da a entender que hace referencia a un gran conjunto de información.

Entrando en el problema de definición, es posible entender el concepto de *Big Data* como una tendencia tecnológica con impacto social, o como la posibilidad de existencia, registro y organización de grandes cantidad de datos, *Big Data* es entonces un fenómeno tanto cultural como técnico, pero a efectos de la presente disertación la definición se debe mantener dentro del límite técnico en el que *Big Data*, o Metadatos en castellano, es el término que se usa para representar conceptualmente conjuntos masivos de datos de estructura variada y compleja que presenta dificultades para ser almacenada, procesada y analizada (Sagiroglu & Sinanc, 2013).

Hay una dificultad mayor cuando se quiere entender lo que es la Ciencia de Datos, en principio la asociación con la palabra ciencia podría dar a entender que en efecto se está tratando con una actividad de la disciplina científica en su concepción tradicional,

pero no es necesariamente así. Una forma de aclarar el concepto, es a través del diagrama de Venn propuesto por Conway (2010) (Figura 2)



Figura 2 - Diagrama de Venn de Conway (2010)

El diagrama muestra que la Ciencia de Datos (*Data Science* en inglés), y su ámbito de acción, se encuentran en el espacio que intersecta el dominio de distintas habilidades y conocimientos de uso extendido en la actualidad, como son:

- i) La habilidad de hackeo o programación, es decir el dominio técnico de los sistemas computacionales, la internet y sus lenguajes.
- ii) El conocimiento de ciencias formales de gran utilidad, y fundamento, dentro de las ciencias como las matemáticas y la estadística.
- iii) La experiencia relevante en investigación científica que permita entender el procedimiento que sigue el descubrimiento científico en cualquiera de sus ramas de aplicación.

Esta representación gráfica de lo que la ciencia de datos es, se puede describir y defender de forma más detallada y a partir de su uso práctico considerando que:

- i) Los datos informáticos se almacenan y transfieren de forma electrónica, y se procesan mediante algoritmos, por lo que es necesario tener habilidades de programación para diseñar y trabajar con herramientas especializadas. Incluso si el investigador no es experto en informática y tiene su campo de

investigación en un área tan disímil como la antropología, es condición necesaria que sepa un mínimo de programación para usar las herramientas más elementales de la informática que hacen parte de la Ciencia de Datos.

- ii) En la práctica, debe existir algún tipo de rigurosidad aceptada por la comunidad científica, los datos, una vez procesados, deben interpretarse utilizando técnicas matemáticas y estadísticas estándar, por lo que el uso y entendimiento de regresiones, por ejemplo, y la interpretación de sus resultados son requisitos obligatorios para hacer análisis detallados que permitan conclusiones sólidas.
- iii) Las investigaciones formales, y en particular las científicas, siguen procesos rigurosos, y las más relevantes de entre ellas tienen como punto de partida preguntas de interés que se construyen sobre respuestas científicas previas para corroborarlas o refutarlas. La experiencia del investigador es un factor clave para el éxito de la investigación, debido a que un investigador experimentado hará mejores preguntas de investigación que un investigador no experimentado, y además tendrá mejores herramientas para realizar la investigación. De esta forma, la investigación será recibida favorablemente por la comunidad científica y será insertada en el *corpus* científico.

De acuerdo a Conway (2010), las habilidades de *hackeo* y sólidos conocimientos de matemáticas y estadística conforman elementos suficientes para crear diversos tipos de algoritmos con aplicaciones diversas en el campo de la investigación, como por ejemplo los algoritmos de *machine learning*, que conforman algunos de los casos de uso más populares sobre los que se construye la inteligencia artificial. Sin embargo, el simple conocimiento de los algoritmos, aunque es un requisito necesario, no es suficiente para realizar investigaciones sólidas y relevantes en el ámbito de la Ciencia de Datos, pues un trabajo de este tipo, podría requerir un nivel de entendimiento del proceso de manejo de información mayor al que se consigue a partir de la simple aplicación de un algoritmo, por ejemplo, de *machine learning*, pues la simple aplicación del algoritmo sin mayor comprensión del mismo por parte del investigador resultaría en hacer un uso de la herramienta similar al de una caja negra a la que se ingresa información, y devuelve otra información de forma mágica. Este problema de la transparencia de los algoritmos de ciencia de datos se tratará en apartados posteriores.

Conway (2010) también considera, al igual que la mayor parte de científicos y epistemólogos, que la ciencia tiene como objetivo descubrir y organizar el conocimiento, proceso que en el ámbito científico requiere seguir etapas que se inician al identificar problemas, hacerse preguntas acerca de lo que sucede en el mundo para luego crear hipótesis que puedan ser contrastadas de forma empírica a través de los datos y del uso de herramientas estadísticas, de esta forma el trabajo con los datos los llevará a convertirse en información y en conocimiento, más propiamente, conocimiento científico. Es en esta área en la que podemos identificar el lugar en el que se encuentra el proceso de realización de los trabajos científicos de investigación tradicionales, que requieren para su correcta ejecución, el contar con la combinación correcta de experiencia en el uso del método científico y de la aplicación del conocimiento de matemáticas y estadística que el investigador haya desarrollado a lo largo de su formación. Al ser esta la forma tradicional en la que se han realizado las investigaciones a lo largo de los años, desde los albores de la formalización del método científico en el S. XVII por parte de Descartes en su Discurso del Método, es natural que muchos científicos e investigadores, hayan logrado desarrollar mucha experiencia para la realización de investigaciones científicas de gran calidad, pero sin contar necesariamente con suficiente conocimiento técnico respecto a la tecnología de los algoritmos, la Ciencia de Datos y la *Big Data* que les permita hacer un mejor uso de estas técnicas para potenciar sus investigaciones y para ser más productivos.

Es muy interesante notar que Conway (2010) destaca el área que se encuentra entre las habilidades de hackeo y las habilidades de experiencia científica (Figura 2) como una zona peligrosa, a la que describe como un área en la que es posible *saber lo suficiente como para hacer algo peligroso*. Esta expresión se puede interpretar como una advertencia frente a lo problemático - en el sentido de generar dificultades epistemológicas - que es la existencia de situaciones en las que una persona pueda ser capaz de obtener y estructurar datos provenientes de un área de aplicación que conoce bien, pero sin tener un conocimiento estadístico y matemático suficiente para interpretar y explicar el significado de la información generada a partir de los datos de tal forma que se pueda generar conocimiento científico fiable, por ejemplo, en una situación en la que sea necesario realizar un correcto análisis de los coeficientes de una regresión lineal para obtener la confirmación o el rechazo de una hipótesis ya como parte final de un trabajo de investigación, o incluso, en una etapa mucho más inicial, cuando se deben utilizar las técnicas que se deben utilizar para realizar una actividad higiénica inicial de los

datos, e identificar cuáles de todos los conjunto de *Big Data* pueden ser utilizados para realizar la investigación y cuáles no serían apropiados para llevarla a cabo. En muchos casos esta es el área en la que por desconocimiento o mala intención, algunas investigadores dan a conocer análisis supuestamente legítimos pero que quedan desvirtuados, o son fácilmente refutables desde el aspecto metodológico por haber sido conocimiento creado a partir de una falta de entendimiento de las características de los algoritmos usados y de sus productos informáticos generados.

A pesar del esfuerzo realizado por Conway, y otros expertos en el tema, en cuanto a dar una definición certera y aceptada de forma consensuada por al comunidad científica respecto a lo que es la Ciencia de Datos, quedan aún muchas dudas y controversias que impiden el alcanzar una definición estandarizada, sin embargo, en donde si parece existir consenso, es en las tres etapas de procesamiento que sigue la Ciencia de Datos, las cuales se describen como:

- a) Etapa de recolección de datos, que es la etapa que contiene los procesos de captura y registro de una gran cantidad de mediciones, números, palabras, imágenes, sonidos, entre otros elementos, por lo general de forma no estructurada y a partir de fuentes muy diversas.
- b) Etapa de análisis de datos, que es la etapa que involucra el uso de herramientas computacionales como los algoritmos de, por ejemplo, *machine learning*, ecuaciones estructurales, correlaciones estadísticas, herramientas de visualización, entre otros, que permiten generar información a partir de los datos. En esta etapa la experiencia del investigador le permitirá discriminar entre diversos tipos de algoritmos y elegir aquellos que sean más apropiados para el fin que desea conseguir.
- c) Etapa de generación de inferencias predictivas, que se realiza a partir del uso de la información generada por los datos procesados mediante los algoritmos (Van Der Sloot, Broeders, & Schrijvers, 2016) de una forma similar a como se concluye la validez de una hipótesis para que haga parte de la teoría científica. Esta sería la etapa de la creación de conocimiento, en la que el investigador se arriesga a hacer una predicción con respecto al futuro y aseverar que *conoce algo*. Como se explicaba en una sección anterior, en esta etapa la inteligencia del investigador que utiliza Ciencia de Datos interactúa con la información, usualmente creada a

partir de técnicas de visualización gráficas, para mostrar y discutir sus conclusiones respecto a la investigación que ha realizado.

De esta manera, se podría complementar la definición gráfica de Ciencia de Datos que se presentó en la Figura 2, para nominarla, de forma similar pero diferente al concepto de ciencia, como una forma de proceder para encontrar conocimiento a partir de los datos, es decir para realizar en la práctica y con herramientas computacionales correctas, la realización metodológica de la línea que una los conceptos dato-información-conocimiento, bajo estándares establecidos por la comunidad que hace uso de la Ciencia de Datos, sea científica o no.

Para efectos de la presente investigación se considerará entonces que la Ciencia de Datos es la *ciencia de la Big Data*, es decir, Ciencia de Datos debe entenderse como el *corpus* de herramientas, procedimientos y tecnologías que facilitan realizar la analítica de la *Big Data*; aquellas técnicas que permiten el procesamiento, la visualización y la interpretación de los patrones presentes en los datos, pues a través de su tecnología y sus herramientas, los investigadores logran procesar grandes volúmenes de datos para obtener información, e interpretar sus resultados para obtener conocimiento. La obtención, la clasificación y estructuración de la *Big Data* para llevar a cabo una investigación se hace a partir del uso de la *Ciencia de Datos*. En capítulos posteriores se verá cómo el problema de la existencia de evidencia empírica recae sobre la información y conocimientos generados a partir de la *Big Data* y su tratamiento con la Ciencia de Datos. Sin embargo, para llegar a ese punto de la discusión se hace necesario también caracterizar a la *Big Data*, puesto que es otro concepto que, aunque fue definido previamente para efectos de la presente disertación como conjuntos masivos de datos de estructura variada y compleja que presentan dificultades para ser almacenados, procesados y analizados (Sagiroglu & Sinanc, 2013), realmente no existe una definición por consenso en las comunidades científicas y filosóficas, y su naturaleza requiere ser evidenciada a partir del entendimiento de las características que la *Big Data* tiene.

1.3 Características de la *Big Data* (Volumen, Velocidad, Variedad, Veracidad, Valor)

Las primeras menciones acerca de la etimología de lo que se entiende como *Big Data* pueden rastrearse a partir de la década de los 90s del siglo XX, sin embargo el

término alcanzó mucha mayor popularidad a partir de la publicación de un artículo de Mashey (1999) titulado *Big Data and the Next Wave of Infrastrress* (Big Data y la próxima ola de *Infrastrress*). La palabra *infrastrress* no tiene una traducción literal al español, pero se puede entender como aquellos elementos físicos, cibernéticos o computacionales, que son críticos para la infraestructura de las plantas industriales modernas sobre las que se desarrolla la producción fabril que sostiene la economía mundial. De ahí la importancia que John Mashey da a la *Big Data* en cuanto a su significancia para la sociedad actual, ya que hace referencia al estrés que sufrirían las infraestructuras de la sociedad, tan dependiente actualmente de la informática, debido al *tsunami* imparable de datos creados a cada segundo. Hacer de la *Big Data* un fenómeno similar a un *tsunami* es bastante gráfico, pero no es necesariamente la mejor de las definiciones posibles.

Por su parte, Floridi (2014), concluyó que la mejor forma de entender la *Big Data* es como un fenómeno contemporáneo que ha permitido a la humanidad entrar en *la era del zettabyte*, la cual describe, de forma similar a Mashey, como el *tsunami* de datos e información en el que se encuentra sumergida la sociedad mundial actual.

En base al trabajo realizado tanto por Luciano Floridi como por John Masher, y otros investigadores de talla similar, podemos concluir que definir el significado del término *Big Data* de una forma distinta a un fenómeno es complicado, y esto explica la afirmación ya establecida previamente respecto a que no ha sido posible aún alcanzar un acuerdo a ese respecto.

Sin embargo, para alcanzar el objetivo de la presente investigación, que es lograr una revisión crítica del concepto de evidencia en el ámbito de la Ciencia de Datos y la *Big Data*, se hace necesario tomar la definición que la explica como conjuntos masivos de datos de estructura variada y compleja que presenta dificultades para ser almacenada, procesada y analizada (Sagiroglu & Sinanc, 2013), y también caracterizar el fenómeno de la *Big Data*, para de esta manera relacionar sus características a los conceptos de evidencia científica que se deben estudiar en los siguientes apartados. Una forma aceptada de hacer la caracterización, y que no afecta de forma negativa la naturaleza de ser un fenómeno a la que se refieren Floridi y Masher, es mediante el uso de las 5 Vs del paradigma penta-dimensional (Gupta & Saxena, 2016) de la *Big Data*, es decir, de forma más apropiada, las 5 Vs que caracterizan el fenómeno de la *Big Data*:

- a) *Volumen*: Es la característica asociada a la gran cantidad de datos que se generan y almacenan en el mundo actual, no sólo por las personas durante su interacción

social, si no también por la interacción entre las propias máquinas con capacidades computacionales. El internet de las cosas (*IoT*) está integrando millones de aparatos que se comunican entre sí, por ejemplo, televisores, refrigeradoras, alarmas, relojes, automóviles, etc. y todos ellos tienen la capacidad de producir e intercambiar datos unos con otros mientras realizan las tareas para las que fueron creados. Así mismo, ya en el ámbito humano, el masivo uso de aplicaciones de mensajería avanzadas que incluyen características multimedia como *Whatsapp* o *Telegram*, aplicaciones de correo electrónico, y de redes sociales como *Twitter*, *Facebook*, *Instagram*, de videojuegos, de entretenimiento como *Spotify*, *Netflix*, entre otras plataformas computacionales sin las que es posible imaginar la vida actual, permiten la creación, difusión y recepción de datos en forma de texto, videos, audio, fotografías a niveles que hasta hace pocas décadas eran inimaginables. Hasta el año 2016, sólo *Facebook* generaba una cantidad de 10 billones de posts y mensajes al día creados por usuarios dispersos geográficamente a lo largo de todos los continentes del planeta, pero que se encuentran reunidos en una enorme comunidad virtual, que ha logrado generar más de 5 billones de ‘*me gusta*’ en posts publicados por los usuarios, quienes suben más de 400 millones de fotografías al día, entre otras variables que podrían identificarse para medir la gran cantidad de datos, información y conocimiento que se produce a diario. El almacenamiento de tal cantidad de datos es un problema tecnológico en sí mismo, pues no es posible almacenarlo en las bases de datos tradicionales, sino que se hace necesario utilizar nuevas tecnologías de arquitecturas computacionales distribuidas lógicamente y geográficamente, y que estén específicamente diseñadas para conseguir un manejo óptimo de este tipo de información, así mismo que sean eficientes en cuanto a su consumo de energía. No es irreal pensar que el volumen de datos seguirá creciendo, y que la única forma de enfrentar el desafío del almacenamiento y procesamiento de la información es a través de la creación de sistemas computacionales físicos y algoritmos cada vez más rápidos, eficientes energéticamente e inteligentes, y de la utilización de cada vez mejores técnicas para buscar y seleccionar los datos más apropiados para cualquier necesidad requerida; este es uno de los espacios de desarrollo de la ingeniería informática y la ciencia de datos en los que se concentra una mayor cantidad de recursos humanos y materiales por parte de la industria tecnológica.

- b) *Velocidad*: Esta es la característica referida a la velocidad con la que se crean y se propagan los nuevos datos cada día. En el mundo actual no sólo hay una gran abundancia de datos, sino que se crean, distribuyen y almacenan muchos más a cada segundo. Cada usuario de smartphone, por ejemplo, no sólo es un consumidor de datos e información, sino es un potencial creador de ambos. Una publicación en redes sociales puede hacerse viral en sólo minutos sin importar la zona geográfica o incluso el idioma en la que se haya creado, generando reacciones positivas y negativas entre millones, o incluso, miles de millones de personas. La explotación empresarial permite ilustrar esta característica de la *Big Data*, pues es cada vez más un estándar la existencia de analistas de negocios especializados en el uso de esta tecnología, que pueden trabajar en crear información, a partir del análisis de datos, y otra información, de las redes sociales para generar *insights* relevantes de oferta y demanda que generen rentabilidad en sus empresas, en muy poco tiempo al aplicar algoritmos especializados para correlacionar la información obtenida en las redes sociales de los potenciales clientes y correlacionarlos con datos de variables que sean de interés para la operación de las empresas. Esta información procesada como conocimiento o, en términos más del ámbito empresarial, *know how*, puede ser luego vendida a otras empresas para campañas de mercadeo y ventas específicas en determinados segmentos de clientes, generando una reacción en cadena en cuanto la generación de nuevos datos a partir de esas nuevas campañas y ventas, por ejemplo, al realizarse una compra en el mercado, donde cualquier comprador puede hacer la transacción a través de una tarjeta de crédito, que es validada por el banco permitiendo al vendedor despachar la compra en cuestión de segundos. La *Big Data* es entonces un insumo poderoso para analizar datos, tomar decisiones y ejecutar acciones en muy poco tiempo.
- c) *Variedad*: Esta característica determina que los conjuntos de datos de la *Big Data* pueden carecer de cualquier orden definido en su estructura y que sus elementos pueden ser muy diversos en sus tipos, es decir, son agnósticos en cuanto a órdenes, estructuras y tipos. En la recolección de datos tradicional, como la usada para investigación durante cientos de años, los datos recogidos son obtenidos cuidadosamente por el investigador quien los clasifica y ordena en tablas de bases de datos para su posterior procesamiento haciendo de este conjunto de datos un elemento de información, pero en el mundo actual de la *Big Data*, la mayor parte

de los datos generados no están necesariamente estructurados y son de diferentes tipos, por ejemplo, gráficos como fotos o videos, texto como actualizaciones de estados en redes sociales, datos en forma de unos y ceros obtenidos por sensores en alguna fábrica automatizada, audio como grabaciones de voz o música, y conversaciones de chats y correos electrónicos son almacenadas en bases de datos como grupos de bits incomprensibles para el ser humano, que se hacen comprensibles sólo luego de realizar un procesamiento previo para transformar los datos en información, la diferencia con respecto al investigador es que mientras aquel usa su inteligencia para formar la información, para la *Big Data* se usa la Ciencia de Datos. Las herramientas de análisis de la *Big Data*, que conforman la Ciencia de Datos, permiten el análisis de todos los tipos diferentes de datos sin importar su origen y formato, ya que existen algoritmos especializados para procesar uno o más tipos de información, y la inteligencia o experiencia del investigador se hace patente en cuanto a la selección de los algoritmos más apropiados para el alcance de sus objetivos.

- d) *Veracidad*: Esta característica se refiere al nivel de corrección (o la capacidad de ser no tener fallos), y a la fiabilidad que tienen los conjuntos de datos. En la era de la *Big Data* las técnicas utilizadas para la gestión de las bases de datos, y los conceptos subyacentes que esas técnicas deben garantizar, como la calidad y la capacidad de gestionar los datos, parecen verse afectadas por la cantidad de fuentes de datos que deben gestionarse y supervisarse constantemente, lo cual crea una sensación de inseguridad respecto a la fiabilidad y calidad de los datos que son generados y almacenados por las personas, organizaciones y máquinas. Es por ello que los profesionales en Ciencia de Datos desarrollan con frecuencia algoritmos cada vez más óptimos y complejos que permitan evaluar la calidad de los datos con los que se trabaja, incluso a pesar de la creencia tan extendida que indica que frente a una mayor cantidad de datos analizados en un trabajo científico, es mucho más probable que los errores estadísticos se haga poco significativos de forma lineal al volumen de datos.
- e) *Valor*: Esta es la característica referida al valor que tienen los datos, entendiéndose valor en el sentido de insumo para generar riqueza económica de algún tipo. A diferencia de las otras ‘V’, esta no está relacionada con aspectos técnicos, sino puramente económicos. El uso de la *Big Data* ofrece oportunidades para que diversas industrias, y de diferentes sectores económicos, puedan

organizar su información de formas más eficientes, y generar conocimiento sistematizado para ofrecer nuevos productos y servicios dentro de sus mercados, a partir de realizar acciones muy especializadas y con un nivel de granularidad tan alto, que hasta hace poco parecerían imposibles, por ejemplo, determinar el perfil del segmento de población que podría ser el más propenso a fallar en el pago de un préstamo, o, en el campo de las industrias creativas, el crear una serie televisiva con los elementos argumentales que la hagan más exitosa comercialmente en una plataforma de *streaming* con espectadores de determinados perfiles socioculturales. Así mismo, a partir de su uso científico para la investigación y el desarrollo, la *Big Data* permite generar valor a partir de aplicaciones diversas, por ejemplo cuando se usa para mejorar el desempeño de la agricultura y la agroindustria a partir de la información obtenida respecto a las características de los campos de cultivo y la experiencia de cultivo en miles de campos similares con cientos de productos distintos, para encontrar las semillas y fertilizantes más adecuados para la productividad. Hay también aplicaciones de uso común que generan valor sin que los usuarios sean completamente conscientes de ellas, por ejemplo al encontrar formas de enfrentar el tráfico a partir del uso de sensores de geolocalización de los *smartphones* de millones de personas que se desplazan a diario por las vías de las ciudades se consigue incrementar la productividad de una sociedad, o en un ejemplo más reciente, al probar los efectos de vacunas que enfrenten pandemias mundiales como la de la COVID-19 que afectó severamente a la economía de muchos países.

Al igual que en el caso de la conceptualización de la Ciencia de Datos, es importante tomar en cuenta que la caracterización de *Big Data* a partir de las 5 Vs no ha logrado acuerdo en cuanto a ser suficiente para definir completamente el término, y esto es debido a la relación que estos mismos términos tienen con otros desarrollos tecnológicos como la computación en la nube (*cloud computing*), o el internet de las cosas (*IoT*) que se encuentran relacionados como habilitadores tecnológicos del fenómeno de la *Big Data*, e incluso como elementos de un sistema de *Big Data*, pero que no son la *Big Data* con todo lo que la palabra conlleva. Por ejemplo, el concepto de *cloud computing* se refiere al proceso de transformación que ha sufrido la arquitectura de las tecnologías de la información, desde los sistemas empresariales propios que eran almacenados en grandes salas de servidores de computadoras y que requerían de personal también propio

encargado de su gestión y mantenimiento, hacia la compra de espacio virtual de procesamiento y almacenamiento de datos provisto por alguna empresa, cuyos servidores se encuentran en cualquier parte del mundo sin que las empresas o las personas que contratan el servicio requieran saber de, ver o comprobar la existencia del elemento físico en el que se almacenan los datos, y a pesar de eso, utilizar la información proporcionada por las técnicas de *Big Data* para enfocar esfuerzos en cambiar y potenciar los procesos de toma de decisiones a partir de los datos almacenados en algún lugar del mundo, en lugar de utilizar recursos para gestionar máquinas y personas especialistas en computación.

Es por las diferencias en las acepciones ilustradas en el párrafo anterior que existen muchas veces confusiones respecto al uso apropiado del término *Big Data*, y se hace necesario y de mucho interés continuar con el esfuerzo de encontrar una definición apropiada que permita clarificar y asociar a otros conceptos relacionados para lograr un mismo entendimiento epistemológico entre investigadores y profesionales del área que permita una discusión filosófica más clara y debatible.

1.4 Impacto de Big Data y la Ciencia de Datos en la Sociedad

Los datos, la información y el conocimiento, así como los sistemas de informática que permitan a las personas generar y hacer uso de esos insumos, nunca han estado y difícilmente llegarán a un punto en el que existan y tengan relevancia en un ámbito distinto al ámbito social (Lee & Cook, 2020), a tal punto que se puede afirmar que el gran progreso de la tecnología que sustentan a la Ciencia de Datos y a la *Big Data* ha creado las bases de la nueva *sociedad de la información* en la que vivimos. En la actualidad casi todo fenómeno natural o social, incluyendo los comportamientos fenoménicos de cualquier tipo que hagan parte de esos fenómenos, pueden ser capturado fácilmente en forma de datos, a través de dispositivos recolectores como sensores y memorias electromagnéticas, que permiten que los datos sean medidos, registrados, procesados, analizados y visualizados.

En la actualidad, tanto las personas como las máquinas, están conectadas entre sí mediante la internet y generan gran cantidad de información a partir de su interacción constante (Mayer-Schönberger & Cukier, 2013), interacción que hace parte de las dinámicas sociales propias de la vida humana. El impacto en la sociedad a partir de esta revolución tecnológica es enorme, por ejemplo, en el campo empresarial, muchas

compañías están haciendo uso de la tecnología de *Big Data* para encontrar patrones y correlaciones estadísticas que les permitan optimizar sus servicios, atender de mejor manera a sus clientes, e incluso personalizar los contenidos que les ofrecen. Por otro lado, a nivel del gobierno de los países, en todo el mundo, las instituciones gubernamentales estatales, están aplicando la Ciencia de Datos con distinto nivel de amplitud y profundidad para mejorar sus políticas y acciones en los ámbitos de la seguridad, de justicia, salud, infraestructura y transporte (Sloot, Broeders, & Schrijvers, 2016). Incluso en el Perú el uso reciente de la *Big Data* ha permitido hacer seguimiento y caracterizar el progreso en la lucha contra la COVID-19, al igual que en países más adelantados tecnológicamente como China o Corea del Sur, lo que significa que esta tecnología ha logrado un rápido alcance de nivel global, y no está restringido sólo a su uso en sociedades del hemisferio norte, que se caracterizan por tener los mayores adelantos tecnológicos al servicio de sus ciudadanos.

Se han abierto interesantes discusiones de tipo filosófico a partir del impacto de la Ciencia de Datos y la *Big Data* en la sociedad, abriendo campo al desarrollo de la Filosofía de la Información, que ha sido definida por Floridi (2005) como una disciplina encargada de:

- a) La investigación crítica sobre la naturaleza conceptual y los principios básicos de la información, de su dinámica en cuanto a los procesos de obtención, almacenamiento, procesamiento e intercambio de datos por medio de la computación, y de la utilización ética de los datos y los métodos que se usan para procesarlos.
- b) La elaboración de metodologías teórico-informacionales y computacionales susceptibles de ser aplicables a los problemas filosóficos.

Puesto que se ha planteado que los datos son los insumos tanto para la información como para el conocimiento, la *Big Data* tiene impacto en los avances que realiza el campo de la Filosofía de la Información, y se hace cada vez más relevante por cuatro razones de interés que son derivadas de su propia naturaleza tecnológica (Succi & Coveney, 2019):

- i) El explosivo crecimiento en la producción, adquisición y procesamiento de datos.
- ii) La capacidad de realizar análisis de patrones en grandes bases de datos utilizando algoritmos, permitiendo que obtener información sea un

proceso más rápido, y más revelador, que modelar el comportamiento de las variables utilizando alguna teoría.

- iii) La capacidad de ser aplicada a cualquier disciplina científica, incluso a aquellas que por su naturaleza compleja no son adecuadas para el tratamiento matemático usual (e.g. ciencias de la vida o el clima).
- iv) La capacidad de ser aplicada de forma inmediata y sencilla a ámbitos que no son científicos, o en todo caso, que no son parte de la ciencia natural, sino de las ciencias sociales, como la política y los negocios, utilizando algoritmos que permitan el análisis de opiniones y sentimientos, abriendo campo a estudios científicos con evidencia empírica, incluso en áreas como el estudio de la ética. De esta manera, la tecnología de la *Big Data*, está generando posibilidad ilimitadas en cuanto a aplicaciones sociales, y que son aprovechadas por millones de personas que buscan generar conocimiento sistematizado alrededor del mundo.

El impacto de las tecnologías de Ciencia de Datos y la *Big Data* en los campos de la creación y gestión del conocimiento en organizaciones de la sociedad es también muy importante. Se sabe que el conocimiento es el activo más importante de la sociedad, el simple hecho de que se pueda contar con la capacidad de obtener y procesar cada vez más datos, significa claramente que, en la sociedad actual se logra *saber* o *conocer* más que en sociedades del pasado. Como ejemplo práctico del uso de los datos, la información y el conocimiento en una organización cuyo funcionamiento es crítico en la sociedad actual como son los bancos, utilizando un procedimiento muy usual dentro de la gestión de servicios financieros, si una persona roba una tarjeta de crédito a otra, o roba los datos de la tarjeta de crédito, puede utilizar esa tarjeta para hacer compras suplantando la identidad del legítimo propietario. La forma en que la banca comercial se protege de esos fraudes, es a través del monitoreo y seguimiento de datos de sus millones de clientes, mediante el almacenamiento y procesamiento de las transacciones realizadas en las compras, permitiendo generar información que puede ser analizada, a través de patrones de consumo, por ejemplo, que deberían identificar patrones inusuales en determinados clientes, y revelar todo posible fraude instantáneamente, permitiendo que los responsables de los sistemas de seguridad financiera obtengan conocimiento práctico de forma casi inmediata de lo que está sucediendo, y permitiendo que las personas responsables por los procesos de calidad, seguridad y atención a los clientes de los bancos

tomen decisiones adecuadas para atenuar el daño que los ladrones de tarjetas de crédito podrían ocasionar tanto a los legítimos propietarios de las tarjetas, como a la institución bancaria y a los comercios que podrían sufrir las estafas.

Conocimiento práctico como el ilustrado en el párrafo anterior, tal vez no logre alcanzar un lugar trascendental dentro del campo del corpus del conocimiento filosófico, en el sentido en que no nos da un mejor entendimiento del ser, de la vida en la Tierra, de la expansión del Universo u otro similar, pero es conocimiento relevante para la sociedad, cuyo origen parte desde una generalización estadística acerca de una persona, su tarjeta de crédito extraviada o robada, un comportamiento de compra anormal, y la alarma de seguridad asociada, y que podría permitir la inferencia de un comportamientos similares más adelante, dificultando que se concreten acciones que puedan generar grandes pérdidas a personas e instituciones, por lo que es también conocimiento de gran relevancia para la sociedad. La *Big Data* se ha convertido entonces en una fuente para la creación de conocimiento, incluso si este no es necesariamente científico, por lo que la justificación de ese conocimiento a través de la existencia de una evidencia es fundamental.

Finalmente, los productos de la *Big Data* se pueden asociar con conceptos relacionados al valor social, en el sentido de responsabilidad social, que se basa en la creación de bienestar en áreas de interés comunitario como son la educación, la salud, la seguridad pública, entre otros. Por ejemplo, los gobiernos pueden usar *Big Data* para mejorar y hacer más transparente la información de interés público como salarios de funcionarios, gastos de inversión, recaudación de impuestos, permitiendo que los ciudadanos participen más activamente del escrutinio y análisis de esta información para prevenir fraudes y crímenes relacionados con la corrupción, mejorar la seguridad nacional, y la calidad de vida de los ciudadanos tomando las mejores decisiones para mejorar necesidades en áreas altamente prioritarias como son la educación y los sistemas de salud (Günther et al., 2017). Este proceso también puede ser tratado como un proceso de creación de conocimiento, y debe ser justificado mediante evidencia para sustentar su veracidad.

1.5 Impacto de *Big Data* y de la Ciencia de Datos en la Comunidad Científica

Los Megadatos (*Big Data*) no sólo impactan la sociedad en sus ámbitos financieros o de entretenimiento, sino que constituyen también un cambio significativo

en la forma en la que se hace la investigación para obtener conocimiento científico del mundo y generar productos que puedan ser llevados al mercado para generar valor económico. De acuerdo a Bachimont (2015), hemos pasado de una epistemología de la medición, sobre la que se fundan las ciencias naturales, hacia una epistemología de los datos que permite abordar de mejor manera los problemas del mundo y la cultura actuales. Si bien es cierto que se puede argumentar que las mediciones son datos, como se ha explicado y mostrado en la Figura 1, los datos en el fenómeno de la *Big Data* cuentan con características particulares (las 5 Vs) que les brindan una ontología propia.

Para Mosterín (2016), cuando se reflexiona respecto a las mediciones y a los modelos matemáticos que se hacen acerca de la realidad para utilizarlos en los procedimientos científicos, se busca responder la pregunta respecto al *cómo pasar las características de los sistemas cualitativos hacia los numéricos y cómo se introduce el uso de las magnitudes en este procedimiento*. Metrizar - que es como llama Mosterín al proceso de tomar una medición - cualquier tipo de característica cualitativa, consiste en encontrar la forma de representar esa característica numéricamente, por ejemplo, la longitud de un cuerpo, o la determinación de qué tan largo es algo, requiere que se utilicen números para poder determinar esa característica completamente. Esta representación numérica de la cualidad de longitud, es la medición, y requiere el uso de una escala que se convierta en la representación equivalente desde el sistema cualitativo empírico (el qué tan largo es algo en el mundo) hacia su forma de sistema numérico (el largo como algo claramente definido e inequívoco a partir del uso de un número único). Una de las actividades que se realiza con mayor frecuencia en los laboratorios científicos del mundo, es tomar medidas una y otra vez, a niveles que alcanzarían los millones de medidas cada día si consideramos el conjunto de laboratorios que hay en el mundo, pues medir es una parte importante del trabajo cotidiano que se realiza en la ciencia experimental.

La actividad de realizar mediciones no ha dejado de existir con la extensión del uso de las técnicas de la *Big Data* y la ciencia de datos en el campo científico, por el contrario, la capacidad de medición se ha potenciado enormemente no sólo en cuanto a obtener nuevos datos en diversos tipos de formatos, sino en cuanto a su almacenamiento como dato puro o como dato con algún tipo de procesamiento. Lo que sí ha cambiado significativamente, es que esta actividad ha dejado de ser una de las que más esfuerzo requiere por parte de los científicos, pues en comparación a la experiencia previa al fenómeno de la *Big Data*, los investigadores que tuviesen alguna dificultad para tomar sus propias mediciones del fenómeno que quieren estudiar – lo cuál es raro, por la

cantidad de dispositivos y técnicas con los que se pueden obtener datos – tienen la posibilidad de acceder a mediciones en bases de datos ya existentes y avaladas por la comunidad científica, también su interpretación y análisis se ha facilitado de forma gigantesca, por los diversos algoritmos que se han puesto a disposición de los investigadores.

Por otro lado, sin el lenguaje no habría lugar para el conocimiento, del tipo que este fuese, incluyendo por supuesto, al conocimiento científico, esto es, conocimiento metódico y con base empírica sobre el comportamiento de la naturaleza (Artola & Sánchez Ron, 2012), y en este punto, es importante destacar que los investigadores y científicos que hacen uso de la Ciencia de Datos, requieren hacer uso de un lenguaje muy especializado para referirse a los propios datos y a las técnicas que se requieren para procesarlos. Por la cada vez mayor extensión en cuanto a su uso, diversos conceptos y definiciones de las técnicas que conforman la *Big Data* y la Ciencia de Datos se han convertido en el estándar aceptado del lenguaje en diversos ámbitos de la nueva sociedad de la información, y no hay duda de que esos términos están teniendo un impacto significativo en la comunidad científica actual, por ejemplo, con palabras ampliamente conocidas como: *machine learning*, *pattern analysis*, librerías de programas, *big data*, entre otras.

También es relevante considerar dentro del espectro de aplicaciones de la *Big Data* con impacto en la comunidad científica, los usos que se dan en algunas ramas de la ciencia contemporánea y que hacen uso intensivo de datos desde siempre; por ejemplo, la astronomía, la genética, la física de partículas y las ciencias del océano. En el caso de la astronomía, por ejemplo, se sabe que existe un número infinito de cuerpos celestiales cuyos fenómenos se estudian a partir del análisis de los espectros electromagnéticos que se logran registrar gracias a los grandes telescopios dedicados a explorar el cosmos, en conjunto con instrumentos y sistemas computacionales muy especializados que registran y almacenan la abundante información generada en bases de datos (Frické, 2015). El gran volumen de información es una característica del fenómeno de la *Big Data*, pero no es la única característica – como se explicó con las 5 Vs-, por lo que el sólo registro y almacenamiento de los datos generados por los cuerpos celestes, aunque muy abundantes, no necesariamente formarían parte de un proceso de estudio de *Big Data*, es necesario además que estos datos generen información a partir del uso de la Ciencia de Datos, y que los científicos utilicen el lenguaje de este campo de estudio para dar a conocer sus descubrimientos en la comunidad científica.

Es posible utilizar el mismo ejemplo de la astronomía del párrafo anterior para ilustrar una línea de críticas negativas que se hacen al uso de *Big Data* en la ciencia, considérese el uso de la base de datos llamada *Sloan Digital Sky Survey*, que contiene un enorme listado de todos los cuerpos celestes que los astrónomos van descubriendo, y para hacer el ejemplo algo más extensivo, considérese también la base de datos de la *Ocean Observatories Initiative* que almacena millones de observaciones respecto al comportamiento marino a nivel mundial. En ambos casos, como grandes listados de datos, las observaciones son importantes, pero lo que los sistemas hacen en la práctica es sólo recoger y registrar datos y cada vez más datos. Los listados de datos no explican por sí mismos algún fenómeno, no crean las bases para hipótesis ni para teorías, ni resuelven algún problema científico, son sólo datos, no información y mucho menos conocimiento.

Es por ello que, de forma equivocada y mostrando desconocimiento más profundo respecto al fenómeno de la *Big Data*, muchos detractores del uso de la Ciencia de Datos y la *Big Data*, partiendo únicamente de la características ley volumen e ignorando como mínimo las otras 4 Vs que caracterizan el fenómeno, consideran que, en general, en muchos de los campos científicos, esta tecnología tiene la utilidad de habilitar la posibilidad de crear grandes listas, catálogos y clasificaciones de datos, pero no las de generar conocimiento científico (Frické, 2015) o, como mínimo, conocimiento verdadero justificado. Esta crítica no debiese ser considerada como válida seriamente, puesto que lo que se debe verificar es si realmente se pueden hacer descubrimientos científicos *a partir* de la gran cantidad de datos *utilizando las técnicas adecuadas*, o si se puede aprender algo nuevo a partir de la gran cantidad de datos *utilizando las técnicas adecuadas*, las críticas a la *Big Data* y la Ciencia de Datos, no pueden considerarse seriamente si no incluyen además de la gran cantidad de datos, las tecnologías computacionales asociadas, y el proceso de trabajo deductivo que debe realizarse a partir de la información obtenida, más allá del volumen de datos que se utilizan como insumo. Sólo este análisis integral, en el que confluyen los datos, la tecnología y la inteligencia humana, hacen posible generar, a partir de los datos, nuevas teorías, explicaciones o predicciones. Para evitar este error categórico, es importante recordar que el volumen es sólo una de las características del fenómeno de la *Big Data*.

Asimismo, el uso de la *Big Data* en el campo científico ha sido criticado muy acertadamente desde un frente más elaborado que el del ejemplo anterior por Succi & Coveney (2019) a partir del estudio de sistemas complejos. Para ilustrar esta crítica, es muy relevante notar que gran parte de las áreas de investigación científica hacen parte de

lo que se conocen como sistemas complejos. Existen cuatro puntos clave que dan pie a una serie de críticas de interés y discusiones respecto a los problemas epistemológicos de la *Big Data* a partir de los sistemas complejos:

- i) Los sistemas complejos son sistemas con variables fuertemente correlacionadas, por lo que generalmente no obedecen a las leyes de las estadísticas Gaussianas que se suelen usar en Ciencia de Datos.
- ii) Ningún conjunto de datos es lo suficientemente grande como para evitar errores, puesto que en los sistemas complejos, que son sistemas con gran sensibilidad a los errores de datos, el mínimo error puede generar resultados impredecibles.
- iii) La correlación no implica causalidad, el lazo entre ambas se debilita conforme se incrementa el volumen de datos.
- iv) En un mundo con capacidad de procesamiento y análisis finito, tener muchos datos es tan bueno como no tener ningún dato, la discusión no pasa por tener más o menos datos, sino por utilizar los recursos apropiados para su procesamiento y análisis, de esta manera se trata de desmitificar el poder de la *Big Data* para su aplicación en problemas científicos reales.

Estas críticas, aunque válidas, son incompletas, pues como se ha expuesto anteriormente *Big Data* es un fenómeno que no se caracteriza sólo por el volumen de información generado o por la única posibilidad de registro de información, sino también tanto por las otras características del fenómeno, como por las técnicas y algoritmos de la Ciencia de Datos que permiten el procesamiento de esa información para facilitar el análisis y la posibilidad de crear inferencias y descubrimientos científicos. En respuesta a cada uno de los puntos por Succi & Coveney (2019), se puede argumentar que:

- i) Los algoritmos de las ciencias de datos pueden diseñarse con capacidad para enfrentar estadísticas más allá de las gaussianas y con mayor nivel de complejidad. Las ciencias de la complejidad han alcanzado mayor desarrollo justamente a partir del uso de algoritmos computacionales especializados cada vez más sofisticados y eficientes.

- ii) Aunque existen errores en los conjuntos de datos, los algoritmos permiten automatizar no sólo la identificación de errores, sino también la corrección de los mismos.
- iii) Aunque la causalidad ha sido una característica del proceder científico, por la propia naturaleza de orientación al progreso de la ciencia, siempre podría caber la posibilidad de preguntarse si ha o no ha llegado el tiempo de considerar otras alternativas.
- iv) La propia búsqueda de mediciones por parte de los científicos y el deseo de tener cada vez más datos y de mayor calidad, niega la posibilidad de que tener muchos datos sea similar no tener ninguno. Mientras haya y se mantengan los niveles de calidad de los datos, siempre será mucho mejor tener más datos para realizar una investigación.

Por otro lado, la propia naturaleza digital de los datos, la información y el conocimiento, y su constante crecimiento así como las posibilidades de comunicación casi ilimitadas entre profesionales alrededor del mundo, facilita la posibilidad de que los investigadores trabajen juntos en equipos multidisciplinarios para enfrentar de una mejor manera los desafíos técnicos, éticos y legales que la investigación con *big data* requiere (Favaretto et al., 2020). Por lo que dificultades o problematizaciones que pongan en duda la solidez epistemológica de estas técnicas, pueden ser discutidas abiertamente en la comunidad científica, favoreciendo su crecimiento, y el de la propia comunidad, e impulsando su alcance a una mayor parte de investigadores ya sea que estén formados o que se encuentren en proceso de formación.

1.6 Problemas epistemológicos del uso de Big Data en investigación

Como sucede con muchas revoluciones tecnológicas, la importancia de la *Big Data* y la Ciencia de Datos se ha extendido más allá del ámbito científico, generando discusiones de segundo orden de tipo filosófico y epistemológico (Succi & Coveney, 2019). Es así que la epistemología de la Ciencia de Datos se ha convertido en un tema de interés en la comunidad científica con interés filosófico y en la propia comunidad filosófica con interés científico desde hace algo más de una década, permitiendo que los investigadores definan diversos tipos de aproximaciones para su estudio. Algunas de esas aproximaciones son:

- Reflexiones respecto a la posibilidad de crear teorías científicas universales, o inferencias inductivas, a partir únicamente del procesamiento de los datos utilizando las técnicas de la *Big Data* (Frické, 2015), es decir que a partir sólo de realizar un análisis estructurado de la gran cantidad de datos, se puedan generar explicaciones, predicciones e incluso nuevas teorías que den respuesta a problemas científicos de interés. Es fundamental para este problema entonces, tener consideraciones de evidencia que permitan que las explicaciones sean correctamente justificadas dentro del entorno empírico en el que se desenvuelve la ciencia.
- Reflexiones acerca de la posibilidad de generar valor económico a partir del uso de la *Big Data* en las organizaciones empresariales (Günther et al., 2017) y la implicancias éticas de la aplicación de esta tecnología en la industria y la ciencia, ya que al implementar esta tecnología, las empresas pueden incrementar sus ingresos utilizando aplicaciones como comercio electrónico, gobernanza electrónica, seguridad de la información, toma de decisiones, entre otras. Es importante considerar que el entendimiento de valor o valor económico para cada organización puede variar, y depende de los objetivos estratégicos identificados y relacionados a la adopción de la tecnología de *Big Data*. Este punto es de interés para filósofos de la tecnología y filósofos de la información, y filósofos de la moral. Como buena práctica de gestión, se puede considerar que la toma de decisiones se debe seguir tomando en cuenta evidencia que sustente una determinada decisión.
- Reflexiones respecto a la posible llegada del fin del método científico tradicional a partir del uso de la Ciencia de Datos para resolver preguntas de interés científico, considerando que de acuerdo al uso actual que se tiene de estas técnicas y de *big data*, casi cualquier conclusión puede inferirse acertadamente detectando patrones en las enormes bases de datos que tienen a su disposición los usuarios de la ciencia de datos. Llevada a un extremo, esta idea es semejante a creer que con suficientes datos, *los números hablarán por sí mismos*, que la correlación reemplazará a la causalidad, y que la ciencia podrá avanzar incluso sin modelos coherentes o teorías unificadas (Succi & Coveney, 2019) o desde otro punto de vista,

considerando que en la actual era de algoritmos sofisticados e información masiva, las teorías, hipótesis, prueba de hipótesis, discusiones respecto a si determinado experimento refuta o confirma una teoría se han innecesarias, creándose un velo de méritos superiores en la investigación basada en ciencia de datos, por sobre la investigación basada en hipótesis (Mazzocchi, 2015). La evidencia sería entonces la existencia de patrones identificables dentro del conjunto de datos, y las características de esta evidencia debería ser clarificada para que el conocimiento generado esté justificado.

- Reflexiones respecto al impacto de la Ciencia de Datos en la sociedad actual a través de todo el ecosistema creado por las tecnologías de la comunicación, los algoritmos, las redes sociales, la comunicación móvil, y la inteligencia artificial que no se encuentran de forma aislada, sino que están insertos, e influyendo constantemente, en la sociedad actual (Lee & Cook, 2020), también en la forma en la que estos nuevos métodos afectan las bases institucionales de la sociedad, y las bases de la política científica, que como actividad social se rige por normas y convenciones que se están modificando desde la aparición de la Ciencia de Datos y la *Big Data* (Ahonen, 2015). Estas reflexiones no tienen una relación directa con el problema de la evidencia en las investigaciones con Ciencia de Datos.
- Reflexiones respecto a la influencia de la ciencia de datos en el proceso de toma de decisiones que realiza una persona, a partir de determinar una definición clara respecto a lo que el entendimiento de una situación puede significar para el tomador de decisiones, es decir, de su capacidad de funcionar como agente epistemológico, así como lo útil o no útil que puede resultar tener gran abundancia de información antes de elegir un mejor rumbo de acción (Hirschheim, 2021), es decir, de la influencia que la tecnología tiene en la agencia epistemológica. De forma similar a uno de los puntos anteriores, la existencia de evidencia es fundamental para sustentar que una decisión es mejor que otra.
- Reflexiones respecto a la poca transparencia que hay al interior de los algoritmos de la *Big Data*, tanto desde la disponibilidad de los detalles de la organización y estructura de los datos, como los de su utilidad y manejo (Obar, 2020), y también del uso de las herramientas de la *Big Data* y la

Ciencia de Datos para el procesamiento de los datos y generación de información y conocimiento, por ejemplo, el entendimiento de como se hace la recolección, limpieza, almacenamiento y presentación de la información en una *app* como *Twitter* para todos los distintos usuarios que hacen uso de esta plataforma, y que al igual que otra plataforma muy utilizada, como es *Facebook*, son entornos tecnológicos digitales que son inestables y ambiguos por el propio uso que hacen los miembros de estas redes sociales (Driscoll & Walker, 2014) de forma casi anárquica, sin mayor regulación, excepto para aquellas áreas con más censura social como la pornografía. La existencia de evidencia válida tendía que partir del supuesto de que los datos que la sustentan han tenido un tratamiento impecable.

- Reflexiones respecto al planteamiento de enfoques que sugieren que los estudios y las conclusiones obtenidas con la *Big Data* son insuficientes si se trabajan de forma aislada, y que requieren de una aproximación colaborativa en conjunto con los métodos científicos tradicionales de tal forma que puedan hacerse análisis, descripción, predicción y control de unidades individuales de estudio (e.g. una persona, una clínica, un hospital, un sistema de salud, etc.). Este uso dentro de la tecnología, dentro de las normas de la comunidad científica, impediría que se dé una situación en la que se haga uso de una gran cantidad de datos recolectados, de un enorme grupo de individuos, pero con el objetivo de hacer una descripción y predicción del comportamiento de otro grupo de individuos, y no necesariamente aquellos que han provisto los datos (Hekler et al., 2019). Esto sin duda es una falta metodológica, o incluso ética, que pondría en tela de juicio el resultado de cualquier investigación. La aproximación colaborativa sólo podría darse entre estudios que contengan evidencia suficiente y similar, puesto que no podría ser posible integrar estudios sin que la evidencia respalde alguno de ellos.
- Planteamiento de definiciones de *Big Data* y Ciencia de Datos, que sean aceptados y utilizados de forma estándar en los campos científicos en los que los investigadores están haciendo uso de esta tecnología (Favaretto et al., 2020) y clarificaciones respecto a la distinción de las definiciones de acuerdo al estado de uso en que se encuentra la información, como *datos*

en principio cuando son registrados, y *datos en la práctica* cuando son utilizados (Jones, 2019), o sólo como *información* que es como se plantea en el presente trabajo. Esta falta de consenso refleja claramente la dificultad que existe entre los investigadores para dar una definición de lo que es *big data*. Las definiciones de *Big Data* y Ciencia de datos se han debido acotar para la presente disertación, pero es una discusión aún abierta.

- Desarrollo de metodologías y procedimientos para la aplicación de *Big Data* en las investigaciones científicas empíricas. Por ejemplo en las ciencias de la salud, se empieza a extender el uso de estas tecnologías bajo la creencia de que la investigación con enfoque en *Big Data* tiene el potencial de mejorar no sólo la salud individual de un determinado paciente a partir de la recolección masiva de datos que describan el estado de su salud, sino también del desempeño completo de los sistemas de cuidado de la salud, tanto dentro de una institución, como el de toda una ciudad, un país o una región continental (Cheung et al., 2019), debido a la posibilidad de capturar grandes cantidades de datos. Es en la generación del conocimiento científico en que la evidencia empírica se hace más relevante, por lo que el uso de *Big Data* requiere que se tengan muy claras las características de la evidencia empírica de los estudios.
- El estudio de la epistemología de *Big Data* desde el punto de vista del conocimiento causal a partir del análisis de investigaciones realizadas en proyectos reales, como pueden ser los estudios que se realizan en el campo de la biomedicina en el que la *Big Data* se está utilizando para estudiar la asociación entre la exposición de una persona a una posible situación de contagio, y compararla con la posibilidad de real adquisición de una enfermedad (Canali, 2016), partiendo sólo desde las posibilidades que tiene la naturaleza de los datos y de su valor intrínseco (Cano, 2016), en este caso no como valor económico, sino como valor operativo para responder a la pregunta de investigación planteada. También en este caso se hace indispensable que el concepto de evidencia empírica esté claro para dar validez a los estudios dentro de la comunidad científica.

Como se ha descrito en las líneas precedentes, existen diversas discusiones epistemológicas que han salido a la luz a partir del uso de la *Big Data* y la Ciencia de Datos, pero es claro que en todos ellos hay un factor que se hace presente y es la existencia de evidencia empírica en los estudios que se realizan con estas técnicas.

1.7 El Problema de la Relación entre Dato y Evidencia Empírica en Big Data

Para dar inicio a la discusión de la presente sección, se puede considerar partir por la clarificación de la naturaleza de los datos, es decir, definir que es lo que son los datos. Para Davenport & Prusak (1997), los datos son ‘simples observaciones acerca de los estados del mundo’, por otro lado, de acuerdo a Frické (2015), los datos son cualquier cosa que pueda ser registrable en una base de datos relacional de forma correcta, es decir reflejando fielmente la realidad, tanto semánticamente como pragmáticamente. El aspecto semántico de la naturaleza de los datos, requiere que los registros de datos puedan ser identificados como verdaderos o falsos, por otro lado, el aspecto pragmático de la naturaleza de los datos sugiere que se debe favorecer el registro de aquellos que parezcan ser hechos concretos, es decir que reflejen la realidad.

Así mismo, los datos también pueden ser definidos como toda aquella posibilidad de *información* que se recoge mediante herramientas físicas especializadas, como instrumentos de medición y sensores, que permitan capturar determinadas características físicas de la realidad para que puedan ser trasladadas a algún tipo de registro. Es importante notar que estas herramientas que permiten la medición de los elementos concretos, o fenoménicos, de la realidad, son contruidos o adoptados en base al conocimiento, principalmente científico, que se haya alcanzado en algún determinado momento, y de las teorías que se tengan aceptadas respecto a qué es lo que se está midiendo y qué es aquello que los instrumentos expresan sobre alguna clase de realidad, más allá de la naturaleza y las características propias de los mismos instrumentos, y de sus indicadores de medición aparentes (Boyd & Bogen, 2021), por lo que, a pesar de que existe una relación entre el dato que se quiere registrar a partir del fenómeno físico, y el instrumento que lo captura, y a pesar de que ambos coexisten en un mismo ámbito teórico, lo cual tiene impacto en la definición e identificación del fenómeno físico, la medición no debe verse en ningún caso afectada por el instrumento de medición.

La forma en la que se definen e identifican los datos a partir de su naturaleza conceptual, o de la conceptualización de aquello que pretenden representar, es también

muy importante para entender el rol del fenómeno de la *Big Data* en la investigación científica. En este respecto, existen dos puntos de vista filosóficos relevantes (Leonelli, 2020): el relacional y el representativo que son compatibles con el desarrollo de *Big Data*.

El punto de vista representativo considera que los datos son una representación confiable de la realidad, generada a partir de la interacción de las personas con el mundo que les rodea, por ejemplo, cuando un biólogo registra en una hoja de *excel* los datos de las células que estudia o cuando un profesor registra la asistencia de sus alumnos a clases. Lo que se consideran datos en estos casos son los objetos que se crean en el proceso de describir o medir el mundo, y pueden ser digitales o físicos, y generan un registro de lo que está sucediendo en un momento dado en el mundo que puede ser utilizado para algún estudio posterior o para extraer información para toma de decisiones. Por esta razón, se puede considerar que los datos son realmente una fuente de evidencia empírica, ya que la producción de los datos equivale a capturar características de la realidad que puede ser utilizada luego para cualquier tipo de estudio posterior. De acuerdo al enfoque representativo, los datos son objetos con contenido fijo y permanente, cuya interpretación, en virtud de ser fieles representaciones de la realidad, necesita ser estudiada siguiendo los métodos inferenciales adecuados.

Por otro lado, el enfoque relacional centra su atención en los aspectos históricos de los datos, resaltando su evolución continua y, en algunos casos, su alteración significativa, y el impacto de estas características en la capacidad de los datos para confirmar o refutar hipótesis, es decir, para ser considerados evidencia. Este enfoque resalta la importancia de documentar los procesos de gestión y transformación o procesamiento de los datos, especialmente con la *big data* que suele almacenarse y transitar por diversos canales digitales alrededor del mundo, y que son agrupados, procesados y analizados en diferentes formas y formatos por distintos investigadores y usuarios. De esta forma se justifica el cada vez mayor reconocimiento profesional de las competencias de aquellos que producen, adecúan, y analizan los datos como pasos indispensables para la correcta interpretación de la información de *Big Data* ya sea en el ámbito científico como en el industrial.

Dependiendo de cual de los dos enfoques se siga, las expectativas respecto a la influencia de la *Big Data* en la ciencia varía considerablemente. El enfoque representativo impulsa la idea de *Big Data* como proveedora de la base de información más completa y confiable que se haya generado nunca antes en la historia de la ciencia, mientras que el enfoque relacional no hace tal afirmación, sino que se centra en los tipos de inferencias o

descubrimientos que se pueden hacer a partir de los datos en un punto dado, así como en la forma en que se hace la inferencia y los fines que persigue, por lo que considerándolos en conjunto, su potencial de convertirse en evidencia requiere no sólo la representación empírica de la realidad, si no, los procesos de inferencia que se siguen para demostrar que los datos son en efecto evidencia empírica de algo.

Otro aspecto que se suele considerar para la evaluación de la esencia de los datos, es el de la conexión que existe entre lo sensible y lo racional, es decir, entre cómo se percibe el mundo y cómo se piensa de éste. De este modo, el resultado de la participación de ambas realidades del ser humano como sujeto cognoscente determinará en gran medida el tratamiento que deba admitirse para los datos (Cano, 2016).

Desde el momento en que el ser humano escoge la variable (establece las componentes principales de la información) y recoge los datos correspondientes, toda información queda conformada por él. Pese a ello, muchos investigadores de *Big Data*, tienden a considerar que, mientras los datos constituyen una entidad libre de subjetividad, el entendimiento humano siempre está expuesto a ella. Se tiende a pensar que resulta más improbable ser engañado por los datos a ser engañado por el propio entendimiento humano, olvidando que los datos, pese a su nombre, distan mucho de ser algo dado sin más, presentado ante los ojos del investigador de forma aséptica y neutral por un ente externo de objetividad infinita. Como explicado por Crawford (2013), *los datos y conjunto de datos no son (per se) objetivos; son creaciones del diseño humano. Es el ser humano el que le da voz a los números, el que extrae inferencias de ellos y el que delimita su significado a través de interpretaciones*. De esta manera, se podría concluir que la pretensión de una ciencia centrada exclusivamente en datos sería una pretensión absurda (Cano, 2016), y permitiría apalancar que la existencia de evidencia empírica en las investigaciones con *Big Data* depende de, además de los datos *per se*, en las inferencias explicativas que hace el investigador.

Por otro lado, a pesar de la actual capacidad de acceso a grandes cantidades de información, se podría decir que la naturaleza de los datos sigue siendo en esencia la misma, y que el problema de la representatividad de una muestra, por citar un aspecto, permanece inalterable, sucediendo que una muestra no es más representativa sólo por el hecho de ser más amplia. No se trataría sólo del *cuánto*, sino del *cómo* y no hay duda de que un adecuado control sobre el origen y la validez de los datos (los cuales pueden ser débiles, sesgados, irrelevantes, ininterpretables, etc.) resultará más difícil conforme el tamaño de la muestra crezca. El mejor conocimiento de un fenómeno no dependerá

únicamente de una mayor disponibilidad de datos, por lo que no debiera aceptarse que sea la mera disponibilidad de datos la que determine su valor; esta no puede convertirse en un criterio de muestreo que termine por convertir los datos masivos en sesgos masivos. Por supuesto que la disponibilidad de datos puede resultar muy positiva, pero no puede hablarse de un progreso científico (una mejora en la comprensión de un fenómeno) basándose exclusivamente en la cantidad de datos que haya sido posible recoger (Cano, 2016). De esta manera se desvirtúa la posibilidad de que sólo la cantidad de datos, incluso obtenidos del mundo empírico, con los que se realiza una investigación sean la evidencia de su certeza.

Es importante recordar y tener presente también, que la cantidad de los datos no es equiparable con la cantidad de verdad (Cano, 2016). La información ha de recabarse en profundidad y no en cantidad. Han de buscarse *smart data* en la *Big Data*. Lo cual significa que más allá de la característica de volumen, existen otras cuatro características (velocidad, valor, etc) a través de las cuáles deben ejercerse tareas de inteligencia sobre los datos y no sólo el acopio y almacenamiento de los mismos. Estos elementos son clave para la formación de la idea de evidencia empírica en las investigaciones con *Big Data*.

CAPÍTULO II EVIDENCIA EMPÍRICA Y METADATOS

2.1 ¿Qué es la evidencia empírica?

En los apartados posteriores se ha descrito el problema de estudio y el contexto tecnológico en el que éste problema existe, así como su relación con problemas epistemológicos existentes. A continuación es necesario describir el concepto de evidencia empírica e incluirlo de manera formal en la presente disertación.

Los logros que el conocimiento ha alcanzado se deben en mucho a una actitud especial hacia la evidencia empírica. El uso de evidencia empírica es especialmente característico del conocimiento científico, aunque no exclusivo. Entonces, lo que identifica a la ciencia y la distingue de la pseudociencia es un tipo de actitud respecto a la evidencia (Sanjuán, 2021).

En la actualidad, la existencia, y el consenso, en cuanto al entendimiento del concepto de evidencia empírica es fundamental para las discusiones filosóficas sobre la ciencia que realizan tanto los propios científicos como los epistemólogos, puesto que existe una amplia aceptación respecto a la naturaleza empírica de las ciencias como medio para alcanzar cada vez un mayor conocimiento de la realidad. Sin embargo, este consenso y entendimiento actual no ha existido siempre así. A lo largo de la historia se han confrontado dos visiones respecto a cual es la mejor forma de entender al mundo y a sus fenómenos, la forma puramente racional y la forma empírica.

En los albores del desarrollo científico, Descartes consideraba la existencia de dos aspectos claramente diferenciados en el ser humano, por un lado la mente humana dotada de capacidad racional y por otro lado, su cuerpo, dotado con la capacidad de actuación mecánica, es decir, la de interactuar físicamente con el mundo, y a partir de esta dualidad se enfrentó al problema de cómo explicar la relación entre ambas partes, para él, siendo ambas tan disímiles. Descartes asimismo, rechazó y negó la posibilidad de que cualquier conocimiento sensible, del mundo físico, obtenido a través de los sentidos del cuerpo humano pudiera llevar a un conocimiento certero de la realidad, argumentando que sólo el conocimiento obtenido, decantado y refinado a través de la capacidad racional de la mente humana, podría alcanzar ese conocimiento de la realidad, partiendo sólo de algunas verdades *a priori* que la mente trae desde el momento en que nace, y generando a partir de ellas, una construcción lógica de la realidad y sus razones.

En contraposición a la visión de Descartes, algunas décadas después, los primeros filósofos empiristas, entre los que destacaron John Locke, George Berkeley y David Hume, presentaron cada uno sus propias propuestas respecto al problema del conocimiento de la realidad desde una óptica completamente opuesta al racionalismo de Descartes. Para los empiristas, el conocimiento tiene su origen únicamente, y siempre, a partir de los datos e información que capturan los sentidos y que son provistos al entendimiento de la mente, es decir el propio cuerpo humano a través de los órganos funcionales de los sentidos es el que permite experimentar la realidad físicamente, y sólo a partir de la experiencia sensible se puede registrar cualquier tipo de conocimiento en la mente pues, el hombre, al nacer es semejante a una “*tábula rasa*” un disco de información vacío que jamás podría enriquecerse sin la interacción de los sentidos con el mundo. Por esta razón, para los empiristas, todo conocimiento humano puede darse sólo *a posteriori* después de que alguien haya logrado vivir y registrar las experiencias sensibles que ofrece el mundo físico (Rendueles Mata & Dreher Grosch, 2007).

Así, John Locke en su famosa obra *An Essay concerning human understanding* establece que “*solo mediante la sensación podemos tener conocimiento de la existencia de cualquier cosa*”; para Locke, el que exista una idea en el intelecto, no prueba de ninguna manera la existencia de esta idea en la realidad. Por ejemplo, nuestra idea de unicornio, no prueba la existencia de unicornios sólo por existir como idea.

George Berkeley por su parte, planteó una línea empirista que se mantuvo a lo largo de su obra, y que sostiene, al igual que como hicieron los otros empiristas, que la experiencia sensible es la única fuente de conocimiento posible, pero la llevó a un punto extremo al argumentar que incluso la existencia material podría no tener independencia respecto a la experiencia, posición que se conoce como *solipsismo*, según la cual el mismo hecho de tener la experiencia afectaría ya a la realidad, cuyo conocimiento se haría imposible.

David Hume, por su parte, se enfrentó también de manera radical al racionalismo de Descartes, rechazando la creencia en el carácter primordial y casi exclusivo de la mente como instrumento para el entendimiento de la realidad. Para Hume, la capacidad cognoscitiva de la mente existe sólo en cuanto logra apropiarse del mundo físico, lo cual logra únicamente a través de la experiencia sensible que se consigue con los sentidos. Por ello puede decirse que para Hume, la mente y su proceso cognoscitivo no es más que un elemento de naturaleza abstracta que no hace referencia a un “algo” en particular. La mente sería por el contrario una especie de película en la cual las experiencias sensibles

hacen su aparición de manera sucesiva, por lo que este tipo de apariciones en un escenario se pueden entender como “representaciones” (Kant, 2002).

La filosofía empirista tuvo gran impacto en la escena intelectual durante sus años de apogeo, sin embargo, es a partir de la *Crítica de la Razón Pura* publicada por Kant en 1791, que la filosofía encontró un área de desarrollo alrededor de la epistemología, en un ámbito que podría estar entre Descartes y los empiristas. Kant tuvo influencias por parte de ambos puntos de vista, el racionalista y el empirista, incluso a pesar de que ambos enfoques plantean tratamientos epistemológicos diferentes. Kant tomó como objeto de su reflexión dos teorías ya bien establecidas en sus días, como eran la geometría euclídea y la mecánica newtoniana y se preguntó qué estructura conceptual subyacente podría explicar la razón por la cuál estas dos teorías, a pesar de ser tan abstractas, lograban ofrecer un conocimiento preciso y eficaz respecto a la realidad empírica (Moulines, 2011). De esta manera, Kant se planteó si era posible que existiese un conocimiento con las características de ser conocimiento necesario y universal, es decir racional o ideal, y que al mismo tiempo pueda existir algún conocimiento de los fenómenos de la existencia que pudiese ser dado a partir de la experiencia. La influencia del racionalismo permitió a Kant postular la existencia de un *yo trascendental* innato en las personas, el cual llegaría al mundo con una serie de categorías *a priori* que permitan tener nociones respecto a la realidad, y que a su vez funcionen como una especie de instrumentos, que modelan y coordinan las sensaciones y las ideas recibidas en la mente, transformando el caos de percepciones que se obtienen de la experiencia, en un conjunto ordenado de conocimiento respecto a la realidad. Por otro lado, de forma complementaria, por influencia del empirismo, Kant postula que este *yo trascendental* sólo podrá conocer la existencia de la realidad del mundo a través de la experiencia sensible, pues la existencia no es un predicado que se aplique a un objeto, sino que es el objeto mismo (Kant, 2002).

Muy posteriormente y construyendo sobre los postulados de los empiristas y de Kant, un grupo de intelectuales, filósofos, economistas y científicos, denominado el “Círculo de Viena” sentó las bases de la epistemología como filosofía de la ciencia, que se convirtió en la forma dominante de entender el proceder científico durante la primera mitad del S. XX: el empirismo lógico. Los empiristas de esta rama del pensamiento filosófico consideraban que los datos obtenidos de la experiencia sensible constituían el único insumo fiable para generar información, también confiable, acerca del mundo natural y rechazaban con la etiqueta de “metafísica” cualquier postulado que expresase

que la estructura de la naturaleza pudiese ser descifrada de cualquier forma alternativa al uso de observaciones sensoriales. (Gordin, 2021)

Transcurrido el devenir histórico descrito brevemente en los párrafos anteriores, es posible volver a la actualidad, en la que el método científico es ampliamente aceptado como herramienta para la obtención de conocimiento del mundo, siguiendo un proceso práctico y especializado que consiste, esencialmente, en el uso de las capacidades de razonamiento humano a partir de evidencia empírica relevante, obtenida de la naturaleza, para elaborar teorías que permitan entender y explicar la realidad de la mejor forma posible, y también para actuar sobre ella creando y utilizando herramientas e instrumentos específicos que permitan hacerlo de forma eficiente. Las teorías obtenidas de esta manera son susceptibles de ser constantemente modificadas y corregidas a partir de la evaluación constante de la propia realidad que permita obtener más y mejor evidencia empírica, y también de otras teorías relacionadas que permitan complementar el trabajo de entender y explicar la teoría bajo estudio (Maletta, 2015).

Es así que la reflexión filosófica respecto a la naturaleza del conocimiento como algo racional o empírico, ha permitido concluir que el conocimiento, y en particular, el conocimiento científico requiere evidencia, y más precisamente, evidencia empírica obtenida de la realidad física para generar conocimiento fiable. Considerando que el método científico utiliza hipótesis para responder preguntas de investigación asociadas a problemas que existen en el mundo real, y que una misma pregunta puede generar diferentes hipótesis como respuesta, la misma evidencia empírica puede ser siempre considerada como información relativa en el sentido de que puede ser evidencia a favor o evidencia en contra de una determinada hipótesis científica u otra (T. Johansson, 2011). Es decir, podría ser que la evidencia empírica no tenga un valor por sí misma, sino que su valor sólo podría ser tangible en relación a la hipótesis que algún científico o investigador haya planteado en su proceso de generación de conocimiento. Como se mencionó previamente, en el caso de la *Big Data*, esta necesidad debería ser aún mayor en cuanto a que la relación que conforma la evidencia no se da únicamente entre dato e hipótesis, sino entre el dato, su procesamiento y las herramientas utilizadas para realizar la investigación.

Desde luego, cualquier evidencia, o conjunto de evidencia acumulada, que podría confirmar o refutar alguna hipótesis que esté bajo cuestión difícilmente podría ser totalmente completa, ya que ningún científico podría contar con los medios físicos para tener disponible toda la evidencia que pueda existir en el mundo real, que le permita

validar una determinada hipótesis. Por lo tanto, incluso cuando la evidencia de apoyo obtenida sea muy sólida y confiable, no será total, y siempre permanecerá algo de duda junto con la única certeza viable, en estos casos, que es la de que la evidencia total es una posibilidad inalcanzable (Copi & Cohen, 2013), lo mismo podría argumentarse en cuanto a una posible necesidad de evidencia total para justificar las investigaciones con *Big Data*. Un punto de interés, en este respecto, sin embargo, es que el advenimiento de las ciencias de la computación, y el desarrollo de la *Big Data* y la Ciencia de Datos podría eventualmente, en un futuro, tal vez no muy lejano, crear las condiciones para que sí se pueda alcanzar un punto muy cercano a *tener toda la evidencia* que pueda ser registrada para validar una hipótesis utilizando la metodología de la ciencia tradicional. La cantidad de información que se registra gracias a la *Big Data* ha alcanzado tal nivel, que no sería extraño, por ejemplo, que en un pueblo de miles de personas en el que todos tuviesen acceso a algún *smartphone*, se puedan realizar investigaciones de ciencias sociales que alcancen a toda la población relevante, que en este caso serían los miles de habitantes del pueblo, en contraste con las investigaciones actuales que requieren que se considere una muestra estadística representativa que casi siempre es menor al 100% de la población. Aunque este punto interesante complementa el estudio de la evidencia empírica que se realiza en esta disertación, no es el punto que se desea tratar por lo que no se continuará el desarrollo de la posibilidad de *evidencia total* en la investigación científica tradicional.

La importancia de la evidencia empírica para el progreso del conocimiento científico humano, radica también en que su ausencia permite que cualquier explicación no científica pueda tomarse como verdadera únicamente a partir de razones imposibles de ser verificadas, por ejemplo, se podría considerar que algo es verdadero sólo por haber sido revelada por los dioses de los cielos, o porque sencillamente “*todo el mundo sabe que algo es así*” o por costumbre, tradición o cualquier otra construcción social subjetiva, respecto a que algo es como se observa que es. El caso de la ayahuasca y sus supuestos poderes curativos físicos y espirituales es un buen ejemplo de creencia tradicional que no tiene relación seria con el conocimiento científico. Creencias pseudocientíficas, y no científicas, se sostienen y mantienen vigentes independientemente de cualquier información que podría considerarse como evidencia a su favor, muchas creencias no necesitan evidencia. La ciencia, por el contrario, sí requiere que una hipótesis sea digna de ser aceptada, o rechazada, sólo en la medida en que exista suficiente evidencia empírica que la respalde. Su verdad o falsedad siempre permanecerá en duda y por ello la búsqueda de evidencia no terminará incluso cuan. La ciencia es empírica porque

sostiene que la prueba de la verdad radica en nuestra experiencia; y por lo tanto, la esencia de la explicación científica es que posee la cualidad de ser comprobable (Copi & Cohen, 2013). En ese sentido, cualquier relación con el conocimiento científico que se haga a partir de la investigación con *Big Data* y Ciencia de Datos, requiere claridad respecto a la existencia de evidencia empírica que sustente sus resultados.

Allen (2001) expresó sus pensamientos acerca de la evidencia de la siguiente manera:

Cuando una observación es consistente con y puede ser explicada por, una determinada hipótesis, entonces nosotros la consideramos evidencia respecto a esa hipótesis, más aún si ninguna hipótesis alternativa puede ser explicada a partir de la misma información, y especialmente cuando existe alguna hipótesis alternativa que predice algo totalmente distinto.

De acuerdo a Popper, esta debiese ser la única base racional a partir de la cual se debe preferir una hipótesis sobre otra, es decir a partir del contraste constante de las hipótesis con evidencia que pueda falsarla. Esta es también la solución que Popper dió al problema psicológico de la inducción que planteó Hume. Una nueva y mejor hipótesis es aquella que puede explicar todo lo que ella consideró previamente, pero también aquella que puede hacer predicciones acerca de cuál hipótesis anterior o competidora podría estar errada o no podría tener respuesta alguna. Estas mismas predicciones son las falsaciones potenciales, o pruebas de las nuevas hipótesis, y el acuerdo que existe entre estas predicciones y las observaciones es lo que cuenta como evidencia en su favor (Allen, 2001).

De esta forma, queda bastante clara la importancia de la evidencia para el desarrollo de investigaciones siguiendo método científico. Pero, ¿cómo se debe entender este concepto en el ámbito de la *Big Data*?

En principio, los datos no son un concepto moderno, parecen serlo porque actualmente, en virtud de la expansión de la informática, por lo general se asocia la palabra ‘*dato*’ con los sistemas computacionales, pero la realidad muestra que el uso del término es mucho más antiguo. En inglés, el uso de la palabra *dato* se remonta a los inicios del S. XVII, y el incremento en el uso de este concepto en los siglos XVII y XVIII está muy vinculado a los conceptos modernos de conocimiento y argumentación (Helmond, 2014), que se empezaron a utilizar con la expansión de la ciencia y las ideas de la ilustración.

Al considerar la forma en la que se han producido los datos a lo largo de la historia, podría parecer razonable asumir que tienen como principal objetivo el ser *referenciales*. Es decir, que con algunas excepciones, los datos son recolectados, registrados y utilizados partiendo de la base de que *nos dicen algo acerca del mundo*, incluso si ese algo tiene una forma incompleta. Es decir, desde el planteamiento inicial de una investigación, los datos son considerados como la representación de un fenómeno de la realidad y potencial evidencia para probar una hipótesis, lo cual conlleva una decisión por parte del investigador respecto a cuál sería la representación más adecuada de los datos para alcanzar su objetivo de investigación (Jones, 2019). Esta decisión no necesariamente tiene una única respuesta, diferentes investigadores pueden tener apreciaciones distintas respecto a los datos como elementos individuales, sin embargo el entendimiento del conjunto de datos, o de los datos como conjunto, genera como resultado productivo de una investigación, la posibilidad de ser informado de algo respecto a lo que sucede en la realidad, y que es lo que se quiere entender o investigar.

Las apreciaciones distintas por parte de los científicos, que se mencionó en el párrafo anterior, tiene como posibilidad que *el origen de los datos puede afectar su valor percibido*. En el campo de la medicina que trabaja con evidencia (*evidence based medicine*), por ejemplo, se han propuesto varias jerarquías para trabajar con datos, a partir de la fortaleza de los diferentes métodos usados para obtenerlos, puesto que la interpretación de los datos de reportes de salud pueden ser influenciados por el conocimiento de quien los registró. Los síntomas que son observados y registrados como evidencia de una condición de salud particular, por ejemplo, tendrán como fundamento, los modelos etiológicos establecidos. Los fenómenos que podrían ser representados por los datos, no son entonces necesariamente claros, objetivos y fijos entre profesionales, sino que podrían ser modelados socialmente (Jones, 2019) o al menos influenciados por tendencias sociales, afectando su capacidad de servir como evidencia empírica para validar objetivamente algún tipo de hipótesis científica. ¿Es acertado entonces afirmar que los datos son por sí mismos evidencia empírica?. Queda bastante claro por lo expuesto hasta el momento que no, los datos por sí mismos no son evidencia empírica de algo.

Así mismo, es necesario retomar un problema planteado a partir de las discusiones respecto a la epistemología de la Ciencia de Datos y del fenómeno de la *Big Data*, y es que las metodologías que hacen uso de esta tecnología no necesariamente requieren la existencia de una hipótesis formal para iniciar la búsqueda de conocimiento. Si es así, y

consideramos, entonces, únicamente los datos en la investigación, sin relacionarlos a una hipótesis, ¿podrían ser estos evidencia de algo?, si es así, entonces, ¿serían evidencia de qué?

2.2 Datos y Evidencia Empírica

Como se describió en el apartado anterior, cada vez que se alcanza algún descubrimiento científico y se piensa que se ha conseguido algo de interés para la comunidad científica a partir de la validación de alguna hipótesis, las personas dentro y fuera de la comunidad científica preguntan de forma natural, ¿cuál es la evidencia que da fundamento a tal descubrimiento?. Este cuestionamiento tan común, abre espacio a una etapa importante del proceso que sigue el método científico, y tiene que ver con la forma en la que las teorías se relacionan con la evidencia empírica, esto es lo que los filósofos de la ciencia como French (2007) llaman: *la justificación*. Es propio de los científicos e investigadores elegir cuál de los objetos producidos a través de su interacción con el mundo, es decir los datos obtenidos ya sea a partir de intervenciones experimentales, observaciones o mediciones merece la mayor atención como evidencia potencial para la justificación del descubrimiento (Leonelli, 2019b).

Desde luego, puesto que el *descubrimiento científico* y la *justificación* son etapas que pueden encontrarse claramente separadas, el proceso de *la justificación* en sí misma es ya una pregunta filosófica frente al problema respecto a cuál es el impacto de los datos experimentales en la generación de teorías. Para esta pregunta, se tienen dos respuestas (French, 2007):

- i) La primera dice que los datos verifican las teorías.
- ii) La segunda dicen que los datos falsan las teorías.

Es relevante notar la forma en la que French utiliza datos y evidencia empírica de forma intercambiable, en las dos afirmaciones previas, se podría cambiar la palabra datos por evidencia empírica, y el mensaje sería muy similar. Sin embargo, recordando los acápites anteriores, la palabra más apropiada no necesariamente sería dato sino evidencia empírica, la cuál deberíamos entender como conocimiento, o por lo menos, como información. Una creencia racional debe estar sustentada en la evidencia disponible para un agente (van Benthem & Pacuit, 2011), y, en el caso del conocimiento científico, esta

evidencia debe ser necesariamente obtenida por medios empíricos, independientemente de si es resultado de una investigación que sigue métodos cuantitativos, o de una investigación que sigue métodos cualitativos. Los datos son los elementos que en su conjunto hacen tangible la evidencia (Jones, 2019), pero se debe adicionar que, en el proceder científico tradicional, esta relación es necesaria, pero no suficiente, ya que se requiere al menos la asociación con una hipótesis para que los datos sean evidencia de algo.

El punto de principal interés para la presente investigación, está en que la investigación que hace uso de la *Big Data* permite obviar el elemento *hipótesis* para conformar la idea de evidencia empírica, haciendo que se argumente, en muchos casos, que la existencia de grandes cantidades de datos ya sea condición suficiente y necesaria para la existencia de evidencia empírica que respalde algún tipo de conocimiento obtenido a partir de las técnicas de la Ciencia de Datos. Asimismo, aunque se puede objetar que los datos en grandes cantidades y sin procesar puedan ocultar información importante para el investigador, justamente por sus características de *Volumen* y *Variedad*, mediante la Ciencias de Datos, es posible transformarlos, reorganizarlos, reducir los errores y transparentarlos de tal forma que sea posible extraer información relevante (*insights*) a partir de su análisis también por medio de algoritmos y técnicas, y a partir de esa información argumentar que se cuenta con evidencia respecto a algo. Esto no significa que los datos no contengan las características que permitan un entendimiento del mundo por sí mismos, sino que es importante tener cuidado al momento de seleccionarlos, y tener muy claro todo lo que se asume al hacer la selección, y que esa selección se hace con algoritmos poco transparentes para la gran mayoría de investigadores científicos. Así mismo, se debe considerar, que la investigación que usa Ciencia de Datos y *Big Data* no es sólo una nueva forma de enfrentar los problemas tradicionales de la ciencia, sino que también permite enfrentar nuevos problemas, y ha creado otros problemas que no existirían en ausencia de esta tecnología, esto no sólo en el ámbito científico, sino en el industrial, y con mayor razón, pues en este último no hay una comunidad como la comunidad científica que supervise o respalde la correcta aplicación de la tecnología, sino que el éxito o fracaso de la solución de algún problema se da como consecuencia de los resultados económicos de la empresa que hace uso de estos métodos.

Como información complementaria para la discusión, es importante notar, que la generación de evidencias puede, entonces, ser cuestionada también en términos de quién determina lo que es evidencia, y también en los términos de la objetividad y

referencialidad de los datos a los que la evidencia está asociada (Leonelli, 2019b). ¿Hasta que punto, o qué tanto, se deben procesar los datos para que sean evidencia?, y, ¿qué pasa si diferentes científicos procesan datos de diferente manera y uno los considera evidencia mientras que el otro no?. La reflexión respecto estas preguntas hace parte del análisis crítico de lo que es la evidencia empírica en la Ciencia de Datos y *Big Data*.

Retomando el análisis de las características semánticas de la palabra dato, en la sociedad actual el término *dato* se usa además de sustantivo, como un adjetivo, por ejemplo, la expresión *basado en datos* (*data-based* o *data-driven* en inglés), se suele utilizar como recurso retórico para dar la impresión de que un determinado análisis de información ha sido realizado con mayor rigurosidad y que es de tipo concreto y bien sustentado, independientemente de que en efecto sea así. Esta expresión *data-based* se suele usar tanto en ámbitos académicos como industriales, y es común encontrar opiniones o argumentos con premisas que contienen la etiqueta de *evidencia basada en datos*, incluso en ámbitos como el político, como una figura retórica para dar mayor importancia a aquello que alguna persona quiere argumentar. De esta manera, el uso de la palabra *dato* ha transcendido en su uso técnico instrumental hacia convertirse en la forma de descripción principal para dar solidez, al menos de forma retórica, a cualquier conocimiento obtenido por medio de programas de computadoras.

Finalmente, se pueden distinguir también, dos formas complementarias de considerar un conjunto de datos que tienen relación con la idea de evidencia; por un lado, como una *entidad* que brinda la posibilidad de hacer cálculos y simulaciones computacionales, a través de las cuales representamos racionalmente el mundo, y por otro lado, como una *utilidad* para las representaciones funcionales, y como interfaces a través de las cuales reportamos los hechos del mundo. La primera consideración es más epistemológica y aborda la cuestión de entender aquello en lo que se convierte la ciencia cuando sus principios científicos son gobernados por los principios del cálculo basado en programas de computadoras, por ejemplo, algoritmos como los de *machine learning*, en lugar de los principios clásicos como son la matematización de las ciencias naturales para crear hipótesis o como la interpretación y crítica de los discursos en las ciencias sociales. La segunda consideración es más fenomenológica en la medida en que se trata de comprender la forma en la que los datos expresan las manifestaciones de la realidad que captamos a través de los sentidos y la forma en la que redefinen el entendimiento sobre el que abordamos la existencia de los objetos del pensamiento, de la percepción e incluso de las imaginaciones que habitan el mundo de las ideas.

2.2.1. De la medición a los datos: el cambio epistemológico

La revolución científica moderna se ha logrado en base a la materialización de experimentos y la realización de cálculos para construir una nueva forma de entender la naturaleza, que permita crear una representación racional y que esté sometida a las leyes de la razón y la lógica. Puesto que se necesita obtener una representación de la naturaleza sobre la cual trabajar, hacer experimentos y hacer cálculos, la ciencia depende entonces, en gran forma, de la medición y del sometimiento de estas mediciones a representaciones y leyes matemáticas, y de cálculo, que permitan deducir un modelo teórico, o ideal, que comprensión y predicción de fenómenos físicos, esta es la forma tradicional de conseguir insumos para la evidencia empírica. Es a partir de esas consideraciones, que Bachimont (2015) define, en el contexto de la epistemología de la medición, las condiciones y los principios bajo los cuáles se justifica el hecho de que el conocimiento científico se construye de forma confiable y razonada sólo a partir de la medición. La medición es entonces el proceso que traduce la realidad del mundo en datos que eventualmente serían la evidencia de un postulado científico en relación a una hipótesis, es decir, la medición sería la base principal para la formación de evidencia empírica.

Para Mosterín (2016), por su parte, los modelos matemáticos de la realidad tienen también utilidad capital en el ámbito científico, puesto que constituyen un procedimiento para pasar de las categorías cualitativas a las numéricas para lo cual es necesario que se introduzca el concepto de la magnitud, que se obtiene a través de la medición, es decir, la traducción de la realidad al dato depende de las matemáticas, por lo cual, la evidencia empírica requiere de herramientas conceptuales matemáticas además de mediciones directas de la realidad. Explica Mosterín (2016) que:

metrizar una entidad cualitativa consiste en representarla numéricamente. La representación numérica toma la forma de una escala que relaciona de la forma más precisa e idéntica posible al sistema cualitativo empírico con algún sistema numérico.

Una instancia medible de la magnitud, permite crear un conjunto de escalas, relacionadas a la misma característica cualitativa que desea representarse numéricamente y que pueden ser transformables unas a otras mediante conversiones matemáticas, por ejemplo al pasar de kilos a libras para medir el peso, o de metros a yardas para medir la longitud.

Es importante distinguir claramente los problemas de *medición* de los de *metrización*, explica también Monsterín (2016), puesto que cuando se dispone de un concepto métrico para un ámbito determinado, y de lo que se trata es de averiguar cuál es el valor el número que una escala de ese concepto asigna a un objeto determinado del dominio, nos encontramos ante una tarea de medida. Sin embargo, cuando por el contrario, se carece de un concepto métrico para un ámbito, que de momento sólo existe cualitativamente, y de lo que se trata es de introducir por primera vez un concepto métrico que lo cuantifique, se tiene un problema de *metrización*.

Metrizar es, entonces, introducir un concepto métrico donde no lo había, lo cual es una tarea relevante, pero que sólo es preciso llevar a cabo en raras ocasiones, puesto que a estas alturas del desarrollo social, se puede decir que la mayor parte de conceptos cualitativos de la realidad ya han sido metrizados. Medir, por el contrario, es hallar el valor que la función métrica asigna un objeto. En todos los laboratorios del mundo se realizan constantemente medidas a veces millones de medidas cada día punto es el trabajo cotidiano de la ciencia experimental. Los datos existen entonces en un ámbito metrizado, y su captura implica algún tipo de medición. De esta manera, se puede afirmar que la evidencia empírica en la ciencia tradicional se obtiene a través de la medición y la representación de esta en un sistema metrizado.

Además de la medición, otra de las actividades científicas más frecuentes durante la captura de datos para la creación de evidencia empírica, es la que consiste en clasificar a los individuos en grupos que se encuentran bajo estudio, de tal modo que sea posible reflexionar y formular leyes e hipótesis sobre esos grupos con más facilidad y relacionarlos para obtener conclusiones a partir de la comparación de sus comportamientos, o de sus respuestas frente a determinados estímulos. De esta manera la evidencia empírica quedaría limitada a las hipótesis obtenidas y aplicables a clasificaciones determinadas, haciéndolas más precisas y robustas.

A partir del fenómeno de la *Big Data* y de la utilización de las herramientas de la Ciencia de Datos, el proceso de medición parece haber perdido la gran relevancia que tenía antaño, dejando paso a la clasificación como el proceso de mayor importancia para iniciar el proceso de investigación, y esta clasificación, a diferencia de como se realizaba en la ciencia tradicional, no es clasificación de individuos sobre los cuales experimentar, sino que es la clasificación de los propios datos, *Big Data*, obtenidos y procesados mediante la Ciencia de Datos. En el pasado, para un investigador, las mediciones requerían de mucho esfuerzo, en la actualidad, usualmente ya han sido realizadas – y

continúan realizándose en cada minuto. Hay bases de datos completas y enormes alrededor del mundo a las cuales no es difícil acceder para obtener datos. En cuanto al proceso de clasificación, aunque el concepto aún puede ser relevante para segmentar a que individuos estudiar dentro de las propias bases de datos, como proceso, es más relevante saber técnicas adecuadas y algoritmos más eficientes y adecuados para clasificar y dar forma a los propios datos, ordenarlos, graficarlos, visualizarlos y obtener conclusiones (*insights*) que son en última instancia, el conocimiento que se desea obtener. *A priori* podría considerarse que al no realizarse mediciones directamente del mundo, no habría la posibilidad de evidencia empírica, sin embargo, se ha mencionado ya que la *Big Data* tiene otras características (5 Vs) a partir de las cuáles se puede conformar una justificación de conocimiento generado por ella.

Es necesario recordar dentro del proceso de medición que se utiliza en el método científico tradicional, que tanto las hipótesis y las teorías de un determinado campo, como las leyes que se obtienen a partir de ellas, fueron parte también del marco teórico que permitió crear los propios aparatos de medición, y las herramientas mentales – conceptos – que permiten la interpretación de la información que utiliza determinado campo de la ciencia. De esta manera, los fenómenos científicos que estudia tal campo de estudio, son capturados por los aparatos de medición que han sido creados y realizados considerando las leyes científicas de ese campo, es decir, tienen la misma coherencia que la interpretación científica de los fenómenos con los que están relacionados: *son las mismas leyes las que permiten construir los aparatos de medición y de interpretar aquello que estos mismos aparatos producen*. En la Ciencia de Datos el proceso no necesariamente se sigue de esta manera, al ser la medición algo que puede considerarse incluso al margen del proceso que sigue el científico de datos, los instrumentos con los que se hayan medidos los datos son irrelevantes al punto de que podrían serles hasta desconocidos, y por la característica de *Variedad* del fenómeno de la Big Data, lo medido podría no estar incluso en el mismo formato que el conjunto de datos con el que se trabajará. Existe entonces un cambio real y perceptible en cuanto a relevancia de la medición en el proceso científico cuando se trabaja en el ámbito de la Ciencia de Datos, y este cambio debe ser considerado para la formación de evidencia.

No sería conveniente cerrar este acápite, sin considerar que existe también otro ámbito que los datos informáticos se han convertido en una rica fuente de generación de conocimiento, y que consecuentemente, es fuente de evidencias que para justificar el entendimiento de algo, incluso dentro del ámbito del conocimiento científico. Este

ámbito, existe en el campo de la simulación computacional, en el que mediante el uso de *software* especializado, se crean situaciones virtuales que se asemejan a las situaciones reales, a partir de modelos matemáticos, y que permite que se obtengan resultados que se suelen considerar cercanos a la realidad, dependiendo de la cantidad de variables utilizadas en contraste con las variables que existan en la realidad. Mientras más variables de la realidad se puedan incluir en el modelo matemático de la simulación, la posibilidad de error en la simulación se minimiza, y se le considera una buena fuente de evidencia plenamente confiable (Symons & Alvarado, 2016). Las características epistemológicas de la simulación, aunque de mucho interés, no hacen parte de los objetivos que se quieren alcanzar con la presente investigación, por lo que no hay necesidad de explicar más la explicación respecto a este tópico, pero sí válida que un conjunto de *Big Data* contenga evidencia que justifique un conocimiento obtenido a través de una simulación en un ámbito puramente virtual, lejano a la realidad empírica.

2.2.2. Un nuevo paradigma: De los datos a la visualización

En el apartado previo, se ha podido explicar que en la Ciencia de Datos, los principios que dan base a la epistemología de la medición, y que son clave para la investigación científica tradicional, no son fundamentales para realizar trabajos de investigación con *Big Data*. La Ciencia de Datos tiene sus propias etapas de proceso fundamentales para realizar investigaciones y generar conocimiento, estas son:

- La recolección de los datos, en la que se siguen procedimientos y metodologías para acceder a bases de datos de origen diverso, ubicados en cualquier del mundo, en múltiples períodos de tiempo, e incluso siendo éstos heterogéneos en cuanto a su naturaleza y a su formato. La recolección, desde luego, no implica medición directa, pues aunque se puede medir, es más eficiente tomar datos ya almacenados y a veces incluso procesados. En esta etapa se hacen evidentes las características de las 5 Vs de los metadatos: Velocidad, Variabilidad, Volumen, Valor y Veracidad, y es necesario que el investigador esté familiarizado con ellas para realizar la selección de datos más apropiada para sus fines.
- El tratamiento y procesamiento de los datos, en el que se utilizan algoritmos y otras herramientas de *software*, en las que subyacen aplicaciones matemáticas y

estadísticas, por medio de las cuáles el cúmulo de datos se transforma en información una vez ordenados e interpretados inicialmente.

- La visualización de los resultados, que requiere el uso de técnicas y software especializado para dar forma gráfica y presentar los datos resultados bajo esquemas de barras, puntos, líneas, cartográficos e incluso una mezcla de ellos, de tal forma que el investigador pueda identificar, de forma visual, patrones de comportamiento en los datos. Así mismo, se pueden crear tablas de resumen que muestren las características estadísticas descriptivas o probabilísticas de los datos para seguir inferencias y obtener conclusiones. Este procedimiento, tiene un área de desarrollo propio y se conoce como proceso de visualización de la información, y es objeto de mucho desarrollo en la ingeniería de *software*.

Las tres etapas descritas previamente permiten obtener algunas conclusiones relevantes respecto a como se puede entender y dar forma a un paradigma epistemológico de evidencia empírica para la Ciencia de Datos. Como explica Bachimont (2015), por un lado, no existe, necesariamente, una hipótesis particular relacionada al conjunto de los datos obtenidos, en especial, cuando se considera que los datos no constituyen una manifestación de conocimiento que se obtiene únicamente dentro de un marco *a priori* de la naturaleza matemática que le da forma. Asimismo, los datos no son tampoco, necesariamente, una medición directa de la naturaleza. Los datos almacenados en bases de datos serían una suerte de información o pre-información, es decir, expresiones de la realidad bajo un formato y según una estructura simbólica arbitraria, que luego de ser recolectada es organizada en un formato ordenado que le haga perder su condición de estructura arbitraria. Entonces se debe contar con información complementaria a los propios datos, por ejemplo, *logs*, texto, transacciones, que no pertenecen al mismo marco *a priori* de la naturaleza de los datos, sino que son complementarios y accesorios a esta. Se puede alcanzar entonces la noción de información en el sentido preciso de contenido de datos formateados, en la medida en que tal formateo permite la manipulación metódica de los datos mediante algoritmos y *software* que hacen posible el proceso de procesamiento o tratamiento de los datos. Vinculando este punto con las secciones iniciales de la presente disertación, queda más claro que el dato sería entonces la parte manipulable más pequeña a partir de la cuál se da forma a la información, que en su más mínima expresión, sería contenido de datos formateados. El dato por sí mismo no es información, el dato bien puede ser la medición, de algo en el mundo, que es registrada

en una base de datos, que se convierte en información sólo dentro de un contexto y de la elección de un formateo particular que puede originar, por ejemplo, una tabulación que puede ser analizable; es decir, puesto que el dato en sí mismo no sería información, es preciso que se tenga la posibilidad de visualizar e interpretar el contenido epistemológico de los datos para obtener la información que se busca.

Así mismo, el tratamiento de los datos se realiza según principios que son totalmente transparentes a la naturaleza de los datos. En la reconstrucción que hace Kant, por ejemplo, de la ciencia tradicional, las mediciones permiten construir el objeto científico de acuerdo a las leyes y conceptos, pues la medición y las leyes científicas pertenecen al mismo marco, se articulan de manera complementaria dentro de una misma concepción y construcción de la ciencia. Por el contrario, en *Big Data*, cualquiera que sea la naturaleza de los datos, su recolección, en el momento en que sean organizados en una tabla, se le podrán aplicar algoritmos que busquen explicar, eliminar los casos aislados, etc., estos algoritmos pueden estar basados en diferentes técnicas y evolucionan constantemente, se adaptan en diferentes lenguajes y funcionan en diversos tipos de plataformas computacionales. Algunas de las técnicas usadas para el tratamiento de los datos se describen en la Tabla 1 (Miller, 2010):

Tabla 1: Técnicas de procesamiento de datos		
Tarea a realizar	Descripción	Técnica
Segmentación o Clusterización	Determinar un conjunto finito de grupos que describan las características de los datos.	• Análisis de clúster
Clasificación	Predecir la clase a la que un conjunto de datos pertenece, en base a un entrenamiento realizado a partir de los datos.	• Clasificación bayesiana • Árboles de decisión • Redes neuronales artificiales
Asociación	Encuentra relaciones entre los datos; predice los valores de algunos atributos basados en los valores de otros atributos	• Reglas de asociación • Redes bayesianas
Desviaciones	Encontrar datos que presentan desviaciones inusuales más allá de las expectativas	• Análisis de clúster • Detección de <i>outliers</i> • Análisis de evolución
Tendencias	Líneas y curvas que resumen la información de los datos, frecuentemente en el tiempo.	• Regresiones • Extracción de patrones secuenciales

Generalizaciones	Descripciones compactas de los datos	• Reglas de resumen • Inducción orientada en atributos
------------------	--------------------------------------	---

Dando un paso más allá en el entendimiento epistemológico del conocimiento generado utilizando la Ciencia de Datos, las técnicas de visualización de los datos crean modos o formas distintas de presentación de la información, en las que la interpretación de las formas puede ser diferente según las preferencias bajo las que se ha permitido recolectar los datos de una parte, y procesarlos de otra. Poco importa lo que se muestra, si es que no hay una organización estética y semiótica de la información que contienen los datos, por ejemplo mediante una representación cartográfica, en la que los principios de comprensión y la forma de dar sentido a la representación, pueden ser arbitrarias con respecto al origen de los datos, en el sentido de la base de datos de los que fueron tomados, y a su tratamiento técnico, que puede ser muy variado. Sin embargo, a pesar de esas aparentes relatividades, esta forma de organizar la información mediante técnicas de visualización, es realmente productiva al nivel del sentido, al nivel de la recepción y percepción de las formas estéticas de la representación visual y del registro semiótico adoptado, que es influido en gran medida por estereotipos culturales, como el uso de los algoritmos de *moda*, o la pertenencia del investigador a alguna determinada industria que tiene preferencia por determinado tipo de representación con respecto a otra.

Lo descrito en los párrafos anteriores, implica que la investigación con Ciencia de Datos ha originado una cuádruple ruptura con la forma en que se entendían los datos en la ciencia tradicional:

- La ruptura de los datos con respecto a su origen y a su naturaleza.
- La ruptura en cuanto a las formas de recolección y la poca relevancia del uso de formatos parecidos debido a su heterogeneidad.
- La ruptura en cuanto al tratamiento de los datos para generar información.
- La ruptura entre lo que se muestra con las técnicas de visualización frente a la transparencia de lo que se calcula con los datos.

Estas rupturas constituyen tanto fortalezas como debilidades de la Ciencia de Datos. Puesto que hay una ruptura entre los datos y su origen, se les puede recolectar de manera arbitraria en relación a su origen y reunir aquello que es diferente y que de acuerdo a criterios clásicos, debería permanecer separado, como imágenes de audio, y texto de música, lo que permite hacer aproximaciones de investigación novedosas, por ejemplo,

unir la información obtenida a partir de entrevistas, grabaciones y mensajería en twitter para determinar el perfil de algún sujeto de investigación. De la misma manera, la ruptura al nivel de la representación de los resultados que con las potentes herramientas *software*, que cuenta ahora con un componente estético relevante, permite superar la complejidad y la masa de tratamientos que son difícilmente comprensibles en sus formas matriciales cuando se busca una representación numérica, e ininteligibles en tanto a la cantidad de cálculos y datos tratados. Sin tal ruptura, el investigador tal vez no estaría perdido, pero sí requeriría contar con un conocimiento matemático de muy alto nivel de complejidad. Finalmente, la ruptura al nivel de tratamiento es también esencial porque es agnóstica a la naturaleza de los datos, y se les puede procesar sea cual sea su naturaleza (Bachimont, 2015).

2.3 Teorías y Modelos de la Evidencia

Es a Aristóteles a quien se le atribuye el planteamiento de una idea de sistema axiomático como objetivo de toda construcción asociada a una teoría científica, y dentro de esa atribución, es también importante recordar que la axiomática ha desempeñado un papel de primer orden en la filosofía de la ciencia durante el Siglo XX (Moulines, 2011). En este sentido, la existencia de una axiomática de la evidencia asociada a las teorías de y modelos de la evidencia, debería tener como objetivo responder la pregunta respecto a qué condiciones deben darse para que algunos hechos sean considerados evidencia en relación con la verdad de determinadas. En el caso particular de la presente disertación, debería estar relacionada a las condiciones en las que se hacen las investigaciones que hacen uso de *Big Data*.

De acuerdo a Achinstein (2001), es usual que los científicos no siempre logren alcanzar algún acuerdo acerca de si cierto hecho es evidencia para justificar la verdad de ciertas hipótesis, o en todo caso respecto a que tan fuerte es esa evidencia respecto a la hipótesis. Muchas veces se dan situaciones en las que los científicos pueden estar de acuerdo en que un hecho ha sido observado y establecido como real, en que un resultado experimental ha sido obtenido, en la descripción del hecho y del resultado, pueden incluso estar de acuerdo en la validez de la hipótesis que se quiere probar; pero su desacuerdo yace en sí es que o en que medida lo que ha sido observado, o el resultado experimental obtenido, provee evidencia para justificar la verdad de la hipótesis. Es decir se presenta un desacuerdo en el proceso de justificación.

Achinstein (2001), también propone un ejemplo muy ilustrativo de este problema a partir del ejemplo del descubrimiento de la naturaleza eléctrica de los rayos catódicos:

En 1883 Hertz realizó experimentos con rayos catódicos para determinar si estaban eléctricamente cargados. En el experimento los rayos catódicos estuvieron dispuestos de tal forma que se pudiese medirse la presencia de electricidad por medio de un electrómetro. En el experimento, la aguja de electrómetro permaneció estática cuando se produjeron los rayos catódicos. En un segundo experimento Hertz introdujo platos electrificados y los dispuso de forma opuesta al rayo catódico. Si el rayo catódico estuviese electrificado los platos deberían deflectarse pero no se dio ningún cambio. A partir de estos resultados, Hertz concluyó que existía evidencia decisiva respecto a que los rayos catódicos no estaban eléctricamente cargados. 14 años después Thompson repitió el segundo experimento de Hertz y obtuvo el mismo resultado. Sin embargo, Thompson rechazó tomar esto como evidencia y menos aún como evidencia decisiva de que los rayos catódicos eran eléctricamente neutrales. Thompson creó la hipótesis según la cual si el rayo catódico está cargado de partículas, entonces cuando las partículas pasan a través del gas, ionizan las moléculas de gas produciendo cargas positivas y negativas que se neutralizan entre sí o que neutralizan la carga en los platos de metal que se encuentran a través de donde se encuentran los rayos catódicos. Entonces si el gas en el tubo catódico no estuviese lo suficientemente evacuado no habría deflexión. A partir de esta hipótesis, en 1897 Thompson fue capaz de remover suficiente cantidad de gas desde el tubo y demostrar la deflexión eléctrica de los rayos catódicos. Como se ve de este ejemplo, Thompson no invalidó el que Hertz no obtuviese deflexión de los rayos catódicos. De hecho, Thompson obtuvo el mismo resultado en sus experimentos iniciales. Lo que decidió desafiar, fue la declaración de Hertz respecto a que estos resultados fuesen evidencia de que los rayos catódicos fuesen eléctricamente neutrales.

En el caso presentado, es claro que hay un acuerdo entre ambos investigadores respecto a que existe algo que ha sido observado, es decir *existen* datos, pero el punto del desacuerdo aparece en cuanto a si esos casos observados constituyen evidencia empírica, en este caso, para validar una hipótesis o no.

A partir de la existencia de la discusión previa Achinstein (2001) logró identificar tres teorías estándar de la evidencia:

- i) La *teoría bayesiana* según la cual, para que un hecho sea evidencia e de que una hipótesis h es verdadera, es necesario que e incremente la probabilidad de h respecto a su probabilidad anterior. Por ejemplo si comprar un ticket dentro de un universo de un millón de tickets de la lotería incrementa la probabilidad de que alguien la pueda ganar, entonces este hecho es evidencia de que se podría ganar. No es mucha evidencia y tampoco es evidencia decisiva pero es innegable que es *alguna* evidencia. Mientras más grande sea el impulso de probabilidad que se obtenga a partir de e en favor de h , más fuerte será en cuanto a su capacidad de ser evidencia. Este tipo de teoría no sería el que se encuentra relacionado con la investigación con *Big Data*.
- ii) La *teoría hipotético deductiva* que dice que para que la evidencia e respecto a la hipótesis h sea válida, es suficiente que e sea derivable deductivamente desde h . Por ejemplo si el hecho de que la luz viaje en forma lineal es derivable de la teoría clásica ondular de la luz, entonces este hecho es evidencia de que la luz tiene un movimiento clásico de onda. Es fácil notar de este ejemplo, que esta es una noción débil de evidencia porque permite que el mismo hecho sea evidencia para un rango de teorías en conflicto entre sí, puesto que la propagación lineal de la luz también es derivable de la teoría clásica de las partículas de luz, entonces el hecho sería evidencia también de que la luz está compuesta de partículas clásicas. Este tipo de teoría sí se podría relacionar al tipo de evidencia que se desea tener para justificar investigaciones con *Big Data*, siempre que pueda evitarse que se sostengan teorías conflictivas entre sí.
- iii) Un tercer enfoque es la *teoría satisfactoria* que dice que un reporte de observación confirma evidencia e para una hipótesis h si la hipótesis es satisfecha por la clase de individuos mencionados en el reporte. Así por ejemplo, si un reporte de observaciones respecto a un cuervo en particular, lo describe como observado de color negro, entonces esto es evidencia de que los cuervos son negros. Entonces si el hecho de que se observe un ave

que no sea de color negro, entonces, esta no sería un cuervo. Este tipo de teoría también podría ser apropiado para justificar investigaciones con Ciencia de Datos, debido a que las técnicas de *clusterización* utilizadas, por ejemplo, divide conjunto de datos en clases y se puede utilizar para clasificar clases similares.

Además de las teorías presentadas previamente, existen modelos de la evidencia. De acuerdo a Brown (2015), el modelo tradicional de la evidencia tiene las siguientes características:

- i) Es esencialista, pues la evidencia se define por alguna propiedad esencial que la hace ser evidencia.
- ii) Es mono funcional, pues la evidencia juega sólo un rol funcional importante que es dar soporte para la justificación de hipótesis o teorías.
- iii) Es no dinámica, pues cualquiera sea la dinámica que siga un descubrimiento científico esta no es relevante para entender la evidencia que refrenda una posible hipótesis. El soporte que da la evidencia es una relación abstracta temporal entre un conjunto de evidencia y una hipótesis
- iv) Es absolutista, pues una pequeña parte de evidencia es evidencia sin importar el contexto. Todo lo que no sirva o encaje como evidencia en cualquier momento y lugar en el que se hace una determinada investigación, no sirve para funcionar como evidencia en ningún otro momento o lugar para esa misma investigación.

La evidencia en *Big Data*, requiere de un modelo más dinámico en el sentido en como lo describe Brown (2015):

Se hace necesario que diferentes tipos de entidades cuenten como evidencia y no sólo una en particular como son los hechos observados, sino también los desarrollos históricos, los análisis estadísticos, las tendencias generales, las leyes fenomenológicas y cualquier otro que pueda servir alguna parte de los roles funcionales de la evidencia y en alguna etapa de la investigación.

Es también necesario notar los diferentes roles que la evidencia juega en el curso de la investigación en Ciencia de Datos. En muchos casos, la evidencia no es mono funcional o esencialista, en el sentido de tener una única utilidad, sino que tiene una

variedad de fuentes y otra variedad de propósitos. Así mismo, la evidencia a partir de *Big Data* que es de naturaleza observacional, puede cumplir diferentes roles, ya que a través de la discriminación de hechos observados en los procesos de visualización de la Ciencia de Datos, provee información acerca de las condiciones de la situación problemática, y permite plantear y replantear el problema de investigación guiando a la especulación y la formación de hipótesis, y guiando también al razonamiento para mejorar la hipótesis original, en el caso de investigaciones de naturaleza científica. De esta manera parece quedar claro que la existencia de evidencia empírica en *Big Data* requiere un desarrollo conceptual mayor que el del modelo tradicional de evidencia que describe Brown (2015).

Vinculando lo expuesto con la explicación previa, y dando un paso más adelante en la discusión de la naturaleza de la evidencia respecto a los datos, es necesario recordar que los datos en sí mismos son unidades que fundamentarán información, y para su correcto manejo y procesamiento, deben almacenarse en algún tipo de medio físico independiente de la memoria humana. Estos medios físicos son de diversa naturaleza y varían desde el cuaderno en el que el científico en psicología anota las respuestas que recibe de un entrevistado, hasta la nube en la que se almacenan los resultados obtenidos de la aplicación de encuestas a 300 personas que podrían conformar el conjunto de datos de una investigación respecto a las características del liderazgo más apropiadas para que una organización empresarial consiga los objetivos que se plantea. Sin embargo, aunque los datos por sí mismos contienen la evidencia, para la ciencia tradicional no son por sí mismos la evidencia, requieren de la interpretación del investigador o científico, y la existencia de un problema al cual el científico quiera dar solución para que los datos conformen una evidencia. Es decir, sólo podría haber evidencia cuando se quiere respaldar una determinanda posición o proposición argumentativa respecto a una posible hipótesis, y en el caso de *Big Data*, cuando se satisfagan condiciones propias de las características de este fenómeno tecnológico.

Esta interpretación de los datos traída al contexto de la Ciencia de Datos y la *Big Data* puede abrir espacio a una discusión respecto al contextualismo epistemológico entre grupos de investigadores de metodologías tradicionales e investigadores de metodologías que usan Ciencia de Datos, que a partir de Lewis (1996) se puede proponer como situaciones en las que “*S sabe que P si y sólo si la evidencia elimina toda posibilidad en la cual no-P ... excepto para aquellas posibilidades que están siendo ignoradas apropiadamente*”.

La discusión se abre paso al considerar que el ignorar posibilidades es un asunto pragmático y contextual, es decir, que depende de la persona que considera que sabe algo. Más aún, si se parte de la idea que indica que la justificación es la característica que distingue el conocimiento de una opinión (incluso verdadera opinión). Entonces, se puede concluir que las adscripciones del conocimiento son dependientes del contexto porque los estándares para una justificación adecuada son dependientes del contexto (Lewis, 1996), y esto debería ser también aplicable al entorno de la investigación científica y a la comunidad de científicos, considerando que lo distintivo de la ciencia es que se preocupa por la evidencia empírica y está dispuesta a modificar sus teorías en función de ella (Sanjuán, 2021).

Contextos diferentes requieren diferentes estándares de evaluación (Lewis, 1996). La actitud científica es un asunto de grupos más que de individuos: “*no es sólo la honestidad o la buena fé del científico individual, sino la fidelidad a la actitud científica como práctica comunitaria lo que hace a la ciencia especial como institución*”. Para ello se analizan y persiguen las buenas prácticas – el método estadístico, el peer review y la puesta en común de los datos – y las malas prácticas – errores deliberados, procedimientos negligentes y errores involuntarios – de la ciencia en tanto institución social a través de algunos casos concretos (Sanjuán, 2021)

Dentro del contexto de la Ciencia de Datos, de acuerdo a Leonelli (2019), existen al menos 8 principios guía para la gestión práctica de los Metadatos (*Big Data*) en la investigación científica, y se relaciona cada uno de esos principios a una consecuencia práctica para diferentes sectores sociales. Aunque la lista está muy simplificada, en las propias palabras de la autora, pueden ser útiles para la construcción axiomática de las teorías de la evidencia en *Big Data*. Estos principios son:

- *Principio 1*- El dato es una categoría relacional: No pueden existir datos sin alguna relación. Sean *Big Data* o no, los datos son interpretables sólo a partir de una red de relaciones conceptuales, materiales y sociales que deben ser expresadas de forma explícita para justificar los resultados de algún análisis. En consecuencia, es indispensable que un investigador esté atento, y tenga claras, las circunstancias en los que se han producido los datos o la forma en que han sido trasladados a lo largo del tiempo, antes de considerar su utilización.

- *Principio 2* - El mantenimiento regular de las infraestructuras es necesario para justificar la confianza en la investigación con *Big Data*: La acumulación y la interoperabilidad de la *Big Data* requiere de un enorme aparato conceptual, material e institucional bajo la forma de infraestructuras, de bancos de datos, de reglamentaciones y programas de de formación adaptados. Estos aparatos requieren ellos mismos, financiamientos específicos dedicados a mantener y a tener siempre actualizados sus elementos a lo largo del tiempo. En consecuencia, las instituciones de investigación deben trabajar con el gobierno, los consorcios y las asociaciones internacionales para desarrollar y mantener sistemas eficaces para el tratamiento de los datos (por ejemplo el trabajo que hace el organismo *European Open Science Cloud*, con el que colaboran diversas instituciones de investigación y financiamiento europeo)
- *Principio 3* – La infraestructura y las competencias en gestión de datos son esenciales para la extracción de conocimiento a partir de la *Big Data*: Es crucial que los investigadores implicados en el análisis de *Big Data* se interesen por el funcionamiento de los bancos de datos y de los algoritmos utilizados para movilizar y analizar los datos, con el fin de poder evaluar de forma crítica el impacto de esos instrumentos, las metodologías que están relacionadas, y los sistemas de clasificación sobre la información extraída a partir de los datos. En consecuencia, los expertos en datos (aquellos a los que se llama *data scientists*) deben ser valorados por las universidades y por el mundo industrial como un nuevo tipo de figura profesional indispensable para la utilización correcta de la *Big Data*. Además, los investigadores científicos de todos los dominios de la ciencia deben recibir una educación mínima para la utilización de tecnologías, infraestructura y métodos relativos al análisis con tecnologías de Ciencia de Datos y *Big Data*.
- *Principio 4* – Debe preservarse el espacio para la investigación exploratoria. Es importante que los investigadores no se queden comprometidos con la utilización de un único tipo de datos por ser los más fáciles a producir o movilizar. En consecuencia, el análisis realizado con *Big Data* debe estar también destinado a la identificación de ámbitos del conocimiento para los cuales no hay datos disponibles, y en consecuencia, estimular el reciclaje de datos antiguos, y al mismo tiempo, producir nuevos datos.

- *Principio 5* – La investigación científica debe enriquecerse de tantas fuentes de datos como sea posible, teniendo en cuenta los riesgos de discriminación y desigualdades relacionadas a la utilización de *Big Data*. La triangulación de los de la *Big Data* ayuda a producir conocimiento más fiable solamente en los casos donde surgieron de fuentes diferentes y separadas (datos que provienen de líneas de investigación diversas). En los casos en los que eso no sea posible, los investigadores deben dejar claras las formas en las que los diferentes tipos de datos pueden o no ser comparados, siguiendo los métodos desarrollados en los distintos dominios científicos de referencia. En consecuencia, la elección de los datos a insertar en un banco de datos debe estar documentada de forma explícita y respaldada por las buenas prácticas utilizadas desde siempre por las ciencias naturales y sociales.
- *Principio 6:* La ética, la seguridad y la responsabilidad social son parte integrante de la investigación centrada en datos: No existe un medio para separar el valor ético de los datos, respecto a su valor científico y epistemológico, y puesto que el análisis de los datos puede estar reglamentado y estandarizado, aquellos que se encargan de organizar y analizar la *Big Data* resultan responsables por la forma en que las reglas son aplicadas en cada caso preciso. Las elecciones fundamentales efectuadas a lo largo de la extracción de conocimiento de la *Big Data*, tiene implicancias decisivas para el nivel de conservación, fiabilidad, parcialidad, corrupción y sensibilidad social presente en los resultados y procesos de la investigación. En la medida en que las responsabilidades son distribuidas entre diferentes personas altamente especializadas, todas las frases del trabajo asociados a la circulación de datos, que comprenden su planificación, creación, movilización y su análisis, implican una serie de responsabilidades y consecuencias sociales. En consecuencia, cada elemento en la gestión de *Big Data* deberá ser evaluado por su valor técnico como por sus implicancias éticas y sociales, como se indica en numerosos códigos de trabajo para científicos de datos creados recientemente para gobernar sus labores.
- *Principio 7* – La utilización de la *Big Data* tiene fines de investigación y está ligado al análisis de hipótesis creadas en diversos contextos de aplicación científica: Debe existir consulta y diálogo social alrededor de los procedimientos de construcción y mantenimiento de infraestructuras y técnicas

de análisis dedicadas a la *Big Data* dentro de la discusión en la comunidad científica.

- *Principio 8* – Es fundamental que cada sector social implicado en la utilización de conocimientos y tecnologías originados en el análisis de *Big Data* se interese por determinar el fundamento empírico de esos conocimientos y de los instrumentos necesarios para interactuar técnicamente con las decisiones tomadas. Se debe promover la creación de diálogo y participación social en los procesos científicos, a través de los gobiernos, las universidades y las entidades que financian las investigaciones

El seguir estos principios desde un aspecto normativo en la investigación que se realiza con Ciencia de Datos y *Big Data*, puede cerrar la brecha del contextualismo epistémico que existe entre investigadores que usan métodos tradicionales, e investigadores que usan las nuevas tecnologías, abriendo paso a una forma de acuerdo axiomático respecto a lo que se debe tomar en cuenta para realizar investigaciones con *Big Data*. Así mismo, el cumplimiento de estos principios puede ser el factor que defina que exista o no evidencia en una determinada investigación que hace uso de la *Big Data*.

2.4 Evidencia empírica y *Big Data* en la ciencia contemporánea

Diversas ramas de la ciencia están haciendo cada vez más uso de la *Big Data* y de sus técnicas. A efectos de ilustración, esta sección describe un campo de investigación que hace un importante uso de la *Big Data* como parte de su evidencia empírica, como es el campo de la medicina sistémica.

En primer lugar, la medicina sistémica es un nuevo paradigma de investigación que tiene como objetivo descubrir asociaciones, relaciones causales y otros mecanismos sistémicos en medicina. Sin embargo, diversas investigaciones demuestran que este nuevo paradigma enfrenta importantes desafíos a partir del uso de la *Big Data* y de sus técnicas, en particular el cómo integrar la evidencia y el cómo estructurar el desarrollo de modelos de análisis a partir de la evidencia. La medicina sistémica utiliza el enfoque de sistemas, de forma similar al enfoque que se usa en biología, con el objetivo de mejorar el tratamiento médico y el progreso de la ciencia médica. Este enfoque de investigación hace uso intensivo de datos y los investigadores los conocen como ‘*data-driven approach*’ (aproximación de investigación basada en datos) porque tienen como finalidad

que se realicen inferencias a partir del gran conjunto de datos que se captura y se tiene disponible (Williamson, 2017).

De acuerdo a Williamson (2017), hay dos tipos de problemas que enfrenta este tipo de investigación en medicina y están asociados a las características propias de la *Big Data* explicadas en un apartado previo en la presente investigación:

1) El problema de la diversidad de evidencia: Además de la gran cantidad de datos que proveen los metadatos, la evidencia tiende a ser heterogénea, y las cuestiones en discusión se atienen a identificar como es que el gran rango de evidencia puede ser integrado de una forma coherente para permitir inferencias confiables.

2) El problema de la diversidad de modelos: la medicina sistémica emplea diferentes modelos de análisis de datos para diferentes propósitos, y está muy lejos de quedar clara la forma en que se esos modelos se relacionan entre sí.

A pesar de esos problemas, al igual que en otras ramas de la ciencia, la medicina sistémica es considerada un paradigma nuevo y excitante para la medicina, principalmente debido a su uso intensivo de la metodología basada en datos. El uso abundante de datos promete conclusiones más robustas, con menos conclusiones atribuibles a artilugios sobre los datos y una gran proporción atribuible a las conexiones genuinas en la población de muestra. Así mismo, el uso de la *Big Data* también ofrece la esperanza de una mayor personalización de los resultados, la gran cantidad de datos podría permitir que el investigador sea capaz de descubrir relaciones causales que se obtengan de pequeñas sub-poblaciones, que podrían de otra manera quedar escondidas dentro de una gran población general. Adicionalmente, los metadatos que utiliza la medicina sistémica normalmente miden un gran número de variables, permitiendo la posibilidad de considerar muchos factores al mismo tiempo, dando a la medicina sistémica el potencial de descubrir mecanismos patofisiológicos más complejos que los que podrían descubrirse sólo con estudios enfocados en causa y efecto, y unas cuantas variables relacionadas (Williamson, 2017).

Los datos que utiliza la medicina sistémica son probablemente más diversos que los utilizados por otras ciencias como la biología, puesto que incluye además de datos moleculares sub-celulares, un mayor nivel de variables clínicas (e.g. tamaño de un tumor, sexo del paciente, etc.) y características ambientales (que describen por ejemplo, el origen de una enfermedad). Además, algunos investigadores que trabajan en medicina sistémica esperan hacer uso de características recolectadas por dispositivos personales y apps –

como el número de pasos dados en un día, peso y presión de sangre – así como una gran cantidad de historias médicas recolectadas por los sistemas TI de hospitales, está claro así que los metadatos juegan un rol importante en la medicina sistémica. Así mismo, el proceso de descubrimiento causal depende de la información acerca de mecanismos, incluyendo los mecanismos sociales y medioambientales adicionalmente a los mecanismos fisiológicos y sus variantes de malfuncionamiento. La evidencia de mecanismos ayuda al descubrimiento causal de diversas formas. Por ejemplo, ayuda a determinar la dirección de la causabilidad y a identificar los intermediarios causales. Adicionalmente, puesto que la medicina sistémica utiliza datos a diverso nivel de escala – ya sea desde el nivel de genoma, hasta el nivel de poblaciones – muchas de estas variables de los conjuntos de datos están relacionadas constitutivamente más que causalmente (Williamson, 2017).

Finalmente, y como complemento al ejemplo de aplicación de la *Big Data* en un campo boyante de la ciencia contemporánea, la aplicación científica de la *Big Data*, no se da sólo en el ámbito de la medicina sistémica, si no en muchos otros, como en el ámbito de las ciencias sociales, disciplinas como la sociología, las ciencias políticas o la antropología tienen algunas características que para los investigadores de las ciencias naturales son problemáticas, como el exceso de teoría, la ausencia de experimentación y replicación, o la renuncia a la causalidad, que pueden amenazar su carácter científico. Para Sanjuán (2021), la solución es clara, si los científicos sociales estuvieran más preocupados, tanto individualmente como en grupo, en confiar en la evidencia y construir mejores procedimientos con los que capitalizarla, las ciencias sociales estarían en mejor situación, y una forma de conseguirlo es utilizando las herramientas de la *Big Data* y de la Ciencia de Datos, que por su propia naturaleza podría enriquecer los aspectos objetivos de la investigación en ciencias sociales.

En este punto, es importante enfrentar la discusión respecto a como participa la evidencia empírica del proceso de justificación de un determinado conocimiento, y su asociación a los tópicos explicados con respecto a la *Big Data* y la Ciencia de Datos. Dos enfoque de mucho interés en este respecto son el explicacionismo y el bayesianismo que se describirán a continuación.

2.4.1 Explicacionismo

En la forma explicacionista de entender los hechos científicos, el estándar de justificación que el científico debe alcanzar no es solamente un umbral de probabilidad matemática sino que requiere además de un juicio cualitativo de la “plausibilidad relativa” de una o más teorías explicatorias (Ward, 2020). La visión explicacionista complementa una de las características de la investigación en *Big Data*, puesto que la mayor parte del trabajo de investigación con estas técnicas se hace y valida utilizando umbrales de probabilidad, clústers de datos y gráficos de visualización. Corresponde al investigador asociar estas características matemáticas con elementos y teorías cualitativas, pero sin la necesidad de requerir a hipótesis. El trabajo se reduciría a explicar lo que los gráficos y resultados estadísticos demuestran a partir del procesamiento en software.

De acuerdo a Appley y Stoutenburg (2017), el explicacionismo señala que una proposición p está justificada por el sujeto s si p hace parte de la mejor explicación disponible para s en base a toda la evidencia e con la que cuenta s .

El problema para el explicacionismo es que las proposiciones acerca del futuro parecen estar siempre justificadas, pero no es totalmente plausible que la justificación sea resultado de ser parte de la mejor explicación que uno pueda tener a partir de la evidencia presente. En el explicacionismo, una persona s con evidencia e en el tiempo t está justificada en creer p , sí solo sí:

- i) p es parte de la mejor explicación disponible para s en el tiempo t para la cual s tiene la evidencia e .
- ii) p está disponible para s como una consecuencia explicatoria de la mejor explicación disponible para s en el tiempo t para el cual s tiene evidencia e .

En Ciencia de Datos, serían los resultados visuales de las técnicas aplicadas a los conjuntos de datos las que generen la mejor explicación posible a la información obtenida de ellos, por lo que el explicacionismo se convertiría en la forma más usada de justificar la validez de la evidencia.

2.4.2 Bayesianismo o Experimentalismo

De acuerdo a Fleck (1979) la observación y la experimentación son sujeto de un mito muy popular: El que quiere conocer es visto como una especie de conquistador. Un

investigador que quiere conocer algo hace su observación, luego su experimento y sólo entonces logra saber. Los científicos que han tenido éxito suelen creer en este mito cuando ven su propio trabajo en retrospectiva. Como máximo admitirán que la primera observación puede haber sido algo imprecisa, mientras que la segunda y tercera se ajustó totalmente a los hechos que buscaban, pero la situación no es tan simple, excepto en ciertos campos muy limitados, como la física en la cual hay conocimiento muy antiguo y amplio sobre los hechos que las sustentan. En las ciencias más modernas y en los campos más complicados, en los cuales primero hay que aprender a observar y hacer las preguntas apropiadas, es muy probable que el trabajo científico probablemente no obtenga resultados hasta que la tradición, la educación y la familiaridad hayan permitido que la comunidad científica se encuentre lista para aplicar las apropiado de percepción de hechos y acción experimental, es decir, hasta que un paradigma haya sido establecido, como lo propugnaba Khun, pues mientras más aspectos desconocidos existan en un campo nuevo de investigación, menos definidas estarán las normas que deben seguir los experimentos. Una vez que un campo de conocimiento se ha trabajado lo suficiente, las posibles conclusiones estarán más o menos limitadas a la existencia o ausencia de una medida que permita determinar de forma cuantitativa, por ejemplo el uso de valores p para determinar si una hipótesis debe ser rechazada o no, permitiendo de esta manera que los experimentos pueden ser cada vez mejor definidos.

Es así que el bayesianismo tiene su fundamento en las teorías de las probabilidades. Concluir que una hipótesis científica está soportada, validada o rechazada por un experimento no es algo que se pueda considerar preciso sólo con la explicación o descripción cualitativa. Es sin duda mucho más deseable que exista alguna forma o medida de credibilidad de la prueba de hipótesis a través de experimentos (L.-G. Johansson, 2016). Se hace entonces posible y necesario calcular la probabilidad de que una cierta hipótesis sea verdadera a través de un cierto experimento. El problema se puede formular de la siguiente manera:

queremos conocer la probabilidad condicional de que siendo una hipótesis verdadera dado el resultado de un experimento calculemos la probabilidad de que sea cierta.

Esta es la característica principal del Bayesianismo. El bayesianismo es un marco que respalda la evidencia y que la sustenta, esta marco representa la mejor forma de dar valor a la evidencia ya que satisface un cierto número de condiciones necesarias con la

normativa lógica científica adecuada, por lo que escoger el bayesianismo suele ser la mejor opción para cuantificar una medida del valor de la evidencia (Taroni, Garbolino, Bozza, & Aitken, 2021).

Como explican Efron & Hastie (2016), una buena definición de argumento estadístico para la evidencia respecto a la verdad de una hipótesis, es aquella en la cual diversas componentes pequeños de evidencia, a veces contradictorios, se combinan para producir una conclusión completa. Por ejemplo, durante la prueba clínica de una nueva medicina, no esperamos que la droga cure a cada paciente o que el placebo siempre falle, pero que eventualmente, luego de muchas pruebas, se obtenga evidencia convincente de la eficacia de la nueva medicina. La prueba clínica recolecta evidencia clínica estadística directa en la cual cada éxito o fallo de la medicina en los sujetos se relaciona directamente con la pregunta de interés respecto a si la medicina será eficaz o no. La evidencia directa interpretada por los métodos que usan la medición de frecuencias, fue el modo dominante de la aplicación estadística durante siglo 20 para la justificación de evidencias, y estaba muy conectada con la idea de objetividad científica. La inferencia bayesiana ha creado bases teóricas para incorporar también evidencia indirecta, por ejemplo la experiencia previa que un doctor puede haber obtenido en el uso de la medicina previamente con grupos relacionados al grupo de estudio con el cual se experimenta, por ejemplo la respuesta a variables como la edad, el sexo u otros. Los cambios en las estadísticas del siglo 21 han respondido a la mayor cantidad de datos disponibles gracias a la modernización de los equipos científicos.

Actualmente, en la investigación científica, la evidencia se cuantifica a través de métodos estadísticos tales como los valores p y los métodos bayesianos de probabilidades de forma rutinaria, sin embargo, existe un alto grado de confusión acerca del uso de estos métodos entre estudiantes como investigadores e incluso estadistas profesionales (Hennig, 2010). Por ejemplo, en una investigación respecto a las preferencias de las hormigas por algún determinado tipo de árbol para construir sus nidos, un investigador podría preguntarse si hace sentido postular que existe una verdad objetiva de acuerdo a las preferencias de las colonias de hormigas por el tipo de acacia en el cual construyen sus nidos lo cual lleva a preguntarnos qué tan fuerte es la evidencia de los datos para llamarles una verdad concebible, cuando lo que se concluye se concluye a partir de un resultado estadístico. La cuantificación de evidencia como problema general así como

también como el problema de asumir el modelo de probabilidad en el que está lista estará basado son instancias del modelamiento matemático de la realidad.

La probabilidad es siempre un tema que trata acerca de lo que pudo haber sucedido, lo cual es diferente de lo que en realidad ha sucedido. En este sentido, siempre involucra lo que es esencialmente inobservable (Hennig, 2010). La probabilidad los enunciados de probabilidad y las cuantificaciones de evidencia nunca pueden ser revisadas sólo mediante la observación experimental.

La Big Data y la ciencia de datos están muy relacionadas a las teorías estadísticas y probabilísticas, y muchas de las conclusiones que se obtienen cuentan con el fundamento que requiere el bayesianismo, aunque este fundamento no esté necesariamente relacionado a una hipótesis, si no a distribuciones y clasificaciones de los datos. Sin embargo, al no necesariamente haber obtenido los datos directamente de la experimentación y medición, el bayesianismo no sería la forma más adecuada para justificar la existencia de evidencia empírica en investigaciones con Ciencia de Datos.

CAPÍTULO III DISCUSIÓN Y CONCLUSIONES

3.1 Consideraciones para una nueva formulación conceptual de ‘evidencia empírica’

Desde el momento en que la ciencia se constituye como una disciplina autónoma en relación con la filosofía, los filósofos iniciaron una reflexión filosófica de segundo orden sobre esta disciplina es decir una reflexión metodológica y metateórica acerca de la ciencia y su proceder (Moulines, 2011), y esto mismo es lo que está sucediendo también en el campo de la Ciencia de Datos y el fenómeno de la *Big Data* durante los últimos años.

De acuerdo a Maletta (2015), los dos tipos principales de investigación empírica que explican la naturaleza del mundo son los que parten de un diseño experimental y los que se basan en diseños no experimentales. Cuando se usa un diseño experimental, el investigador interviene en la realidad, el experimento, para controlar los factores o variables que no le son relevantes, y producir los resultados a partir de acciones deliberadas sobre los factores o variables que sí le son relevantes para lo que quiere probar. En los diseños no experimentales, por el contrario, el investigador sólo observa y registra los hechos que ocurren naturalmente, sin tener la posibilidad, o el deseo, de intervenir directa o indirectamente en el desarrollo del fenómeno, incluso si se tiene la posibilidad de realizar una intervención controlada que no afecte en mucho las variables escogidas para realizar la investigación. En ambos casos el investigador participa activamente en la recolección de los datos y utiliza los instrumentos apropiados dentro del marco teórico en el que ha decidido trabajar. El proceso de sustentación de la evidencia empírica sigue entonces la teoría bayesianista de la evidencia y conforma una justificación, también bayesianista, con el fin de aceptar o rechazar una determinada hipótesis.

En el caso de un diseño no experimental, se puede argumentar que se asemeja de alguna manera a la forma de afrontar una investigación científica utilizando Ciencia de Datos y *Big Data*, puesto que no hay posibilidad de modificar la información obtenida, sino como mucho seleccionar que variables utilizar y que variables no utilizar para realizar el estudio, con la gran diferencia de que el investigador, muchas veces, no toma las mediciones de la realidad empírica, sino que lo hace de una base de datos previamente existente. Su actividad ya no se enmarca en la recolección de datos para justificar la

existencia de evidencia, sino en el procesamiento correcto de los mismos a través de la selección de las herramientas informáticas más adecuadas para realizar su trabajo.

En particular, sólo cuando se trabaja directamente con un diseño experimental, el investigador está en condiciones de manipular la realidad empírica de tal modo que puede mantener constantes los aspectos que no son relevantes para su investigación, y puede hacer variar deliberadamente y en forma controlada sus variables independientes, que operan como causas, para observar cómo esa variación influye en las variables dependientes generando efectos que permitan sacar conclusiones respecto al experimento, y justificar la existencia de evidencia empírica de una forma muy directa. Sin embargo, en el campo de las ciencias humanas lo mismo que en algunas otras ramas de las ciencias naturales como la astronomía, queda poco campo de acción para la experimentación y el científico solamente puede registrar las correlaciones observables entre diversos fenómenos para tratar de hacer inferencias causales a partir de los datos y las fórmulas correlacionales. Estas inferencias son menos concluyentes que las basadas en experimentos, porque el investigador nunca está seguro de que los efectos obtenidos tengan su origen precisamente en la variable que se consideró como variable independiente. Los efectos también podrían deberse a alguna otra variable que no fue considerada, pero que opera al mismo tiempo y que no es posible mantener aislada, o al menos constante, mientras el investigador observa el comportamiento de las variables que ha elegido como más relevantes, debilitando la existencia de evidencia empírica en su investigación. Aún así, dentro de ciertos límites es posible construir y someter a contrastación empírica diversos modelos causales mediante investigaciones no experimentales, es así que en la actualidad la producción científica ha alcanzado niveles gigantescos en cuanto a artículos producidos, y no es la mayor parte de ellos la que hace uso del método experimental. Es decir, la comunidad científica acepta la existencia de evidencia empírica incluso si los datos que la sustentan no han sido capturados directamente del fenómeno físico que se quiere estudiar y en las condiciones de control que sólo permite un experimento.

A partir de los párrafos anteriores, es posible incidir específicamente en cuanto a los diseños experimentales, y notar que existen notorias diferencias entre lo que se conoce como experimento científico y lo que se entiende como la experiencia construida a partir del experimento científico. Mientras que el experimento científico puede ser entendido e interpretado en términos de una pregunta simple, una posible hipótesis y su respuesta obtenida a través del proceso de recolección de datos y su procesamiento matemático, la

experiencia que tiene el científico que realiza el experimento, debe ser entendida como un proceso mental complejo que ha requerido entrenamiento intelectual y que está basada en la interacción entre el científico que científico que busca confirmar o rechazar su hipótesis, todo aquello que ya conoce, es decir, las teorías y experiencias previas que ha tenido, y aquello que tiene aún por conocer, es decir la respuesta a su pregunta (Fleck, 1979).

Sin embargo, todo el proceso de adquisición de habilidades intelectuales, físicas y psicológicas que el científico ha recibido durante su formación, así como la experiencia obtenida luego de realizar cierto número de observaciones y experimentos a lo largo de su carrera, y también su habilidad y creatividad personal para utilizar y vincular conceptos, introducen una serie de factores relevantes en su proceso cognitivo, que no pueden ser totalmente regulados mediante la objetividad de la lógica formal (Fleck, 1979). Todos esos factores dificultan el tratamiento sistemático del proceso cognitivo que sigue el científico mientras trabaja.

Por otro lado, muchos hechos científicos que se encuentran en la actualidad sólidamente establecidos, estuvieron muy relacionados al inicio de su desarrollo con ideas que podrían llamarse pre-científicas que conformaron las hipótesis que los investigadores han buscado confirmar o rechazar y que luego del propio desarrollo de la experiencia científica, terminaron convirtiéndose en hechos científicos verificados como parte del corpus científico. De esta manera, como explica Fleck (1979), es posible considerar que un hecho científico es una relación conceptual que tiene como origen la investigación acerca de la realidad siguiendo el método científico, y que puede ser estudiado desde el punto de vista de la rama de la ciencia a la que pertenece el hecho, así como también de la historia y de la psicología del científico y sus colegas que trabajan en ese campo, tanto de forma individual como de forma colectiva. Así mismo, es importante notar, que el hecho científico, no puede ser reconstruido totalmente a partir únicamente de estos puntos de vista, es decir el psicológico y el histórico, sino que su requiere la existencia de la evidencia empírica que vincule al hecho científico con la realidad, y que sustente su relación conceptual con el resto de las teorías que conforman la rama de la ciencia dentro de la que se hace el estudio.

En la Ciencia de Datos, como en la ciencia en general, la evidencia está contenida de alguna forma en los datos, puesto que están relacionados a valores numéricos obtenidos directamente, ya sea a través de mediciones, textos capturados, imágenes, audios, entre otros, o indirectamente a través de la descarga y procesamiento de datos

almacenados en bases de datos de las comunidades científicas, instituciones gubernamentales, educativas o de forma pública en redes sociales y páginas web. En particular, en la Ciencia de Datos y en la *Big Data*, el valor numérico que contiene el dato, se construye siguiendo un proceso de tratamiento del conjunto de información que define la forma en que se ordenan los datos y los cálculos que se realicen con ellos, en ambos casos, el valor de la información para el investigador no se define de otra manera que a través de su manipulabilidad, es decir de lo que se puede hacer con los datos, independientemente de toda realidad que le sea asociada. Lo numérico de los datos, como elementos que contienen la evidencia se caracteriza entonces por dos situaciones:

- Una característica material, según la cual el símbolo numérico es independiente de la forma en que se almacena materialmente, es decir, independiente de su implementación; esta propiedad es la que nos permite creer que obtendremos la misma información de una archivo a pesar de utilizar sustratos físicos diferentes, como un disco duro, una página web o un DVD; en efecto, mientras que los archivos son físicamente diferentes, por ejemplo, surcos y valles en un DVD con un soporte óptico, y a tensiones magnéticas fuertes o débiles en el caso de un disco duro magnético), son numéricamente idénticos porque ambos codifican la misma cosa unos y ceros.
- Una característica semántica, que determina que el símbolo numérico que contiene el dato no tiene ninguna interpretación particular respecto a lo que significa en términos de origen, por ejemplo un mismo conjunto numérico binario puede ser leído como una imagen o como un sonido dependiendo de los programas y algoritmos que se utilicen para descifrarlo.

Estas características de los datos deben tener un impacto en la forma en la que se entiende el concepto de evidencia. Considerando que la evidencia es una forma de información, y en mayor medida una forma de conocimiento, la evidencia empírica haría parte no sólo de los aspectos puramente sensibles que se obtienen para crear información, sino también de los aspectos mentales y racionales del investigador. En Ciencia de Datos, así mismo, debiese incluirse una instancia de influencia de los algoritmos, ya que su selección depende de la experiencia del investigador, y su uso de la pericia que tenga para aplicarlos e interpretar los resultados. Es posible que dos usuarios de ciencia de datos obtengan conclusiones diferentes del uso de un mismo algoritmo en un mismo conjunto

de datos, con diferentes no se quiere decir opuestas, sino diferentes en cuanto a riqueza y profundidad, ya que la interpretación requiere de procesos mentales que son diferentes en cada investigador.

Como explica Fleck (1979), *un hecho es algo definido permanente e independiente de cualquier interpretación subjetiva por parte del científico. Esto es lo que la mayor parte de disciplinas científicas esperan lograr. La crítica a los métodos utilizados para establecer el hecho constituyen la materia principal de la epistemología. La epistemología normalmente se compromete con un error, o comete un error fundamental, pues casi siempre considera que los hechos realmente establecidos, como los de la física clásica, son los únicos que son confiables y valiosos de ser investigados. Pero existen una diversidad de factores que deben considerarse además del hecho, y son:*

- *Lo individual*
- *Lo colectivo*
- *La realidad objetiva*

Los tres factores en su conjunto dan forma al hecho que aparece. De esta manera, entonces, la conceptualización de evidencia empírica debe seguir consideraciones individuales propias del investigador o científico que realiza el trabajo, de la comunidad científica que acepta la investigación, la valida y la incluye dentro del *corpus* de conocimiento aceptado, y de las propias características técnica de la metodología que se sigue para realizar la investigación y que permiten expresar la realidad objetiva en forma de datos. Se describe a continuación una propuesta de consideraciones que deben seguirse para formular la existencia de evidencia empírica en investigación con *Big Data* por cada uno de esos elementos:

- a) El aspecto individual: En lo individual, es necesario que el investigador demuestre que tiene suficiente conocimiento teórico del tema de investigación que está tratando, y de las particularidades técnicas de la *Big Data* y la Ciencia de Datos, es decir del lenguaje que se usa en este entorno y que lo podrá insertar en el espacio social en el que existe. Entender la naturaleza de este fenómeno y las técnicas que se utilizan permite un uso consciente de su potencial y limitaciones. Esta consideración es relevante puesto que la justificación probablemente seguirá el proceso de justificación

basado en el Explicacionismo, dentro de la teoría hipotético-deductiva de la evidencia, que se sostiene sólo con una argumentación lógica impecable del proceso de justificación y de las premisas que se siguen. Es particularmente importante que se cumplan estas consideraciones considerando que el investigador en *Big Data* no suele participar del proceso de recolección de datos, sin ellas no habría sustento suficiente para hablar de evidencia, y mucho menos de evidencia empírica en el trabajo de investigación.

b) El aspecto colectivo: En lo colectivo, se deben superar las limitaciones que plantea el contextualismo epistémico que podría existir entre investigadores que realizan investigaciones en la forma científica tradicional, ya sea experimental o no experimental, e investigadores que usan Ciencia de Datos y *Big Data*, para ello sería muy recomendable que se sigan los 8 principios planteados por Leonelli (2019), descritos en el apartado 2.3 de la presente disertación. Siendo la existencia de evidencia empírica un asunto de acuerdo social entre los miembros de la comunidad científica, se deben plantear claramente las condiciones que permitan que se eviten posibilidades que sean ignoradas debido al uso de la Ciencia de Datos o *Big Data* en la proposición de Lewis que indica que *“S sabe que P si y sólo si la evidencia elimina toda posibilidad en la cual no-P ... excepto para aquellas posibilidades que están siendo ignoradas apropiadamente”*.

c) En el aspecto de la realidad objetiva: En este aspecto, se hace necesario mantener las características relacional y representativa de los datos, incluso a pesar de haber sido procesados previamente, de tener orígenes diversos y de su abundante cantidad. La documentación constante de todo el trabajo que se realice con los datos, de su origen, de su antigüedad y mantenimiento es de gran utilidad para que mantengan su vínculo con la realidad, y no se conviertan en un elemento etéreo o puramente racional alejado del mundo empírico.

El seguimiento de estas consideraciones debería permitir una justificación explicacionista de las investigaciones con Ciencia de Datos y *Big Data*, dentro de la teoría de la evidencia hipotético-deductiva.

3.2 La autoridad de los métodos y algoritmos en la Ciencia de Datos

Aunque la Ciencia de Datos usa algoritmos que son procedimientos derivados de la lógica de tradiciones deductivas y de las matemáticas, opera de acuerdo a una a una serie conjuntos de estándares, y epistemológicos, diferentes que producen diferentes experiencias respecto a lo que es la verdad. Para el científico de datos se da una situación más pragmática que la del matemático, y a diferencia de éste que alcanza el éxito mediante su propio razonamiento, el científico de datos alcanza el momento más importante de su investigación cuándo su su algoritmo empieza a producir resultados útiles.

En el análisis de datos el objetivo no necesariamente es probar algo. Los algoritmos empleados tienen como objetivo cumplir alguna de las tareas descritas en la Tabla 1, por ejemplo, clasificar, hacer clusters hacer proyecciones de series de tiempo o visualización de redes. El objetivo no es la demostración a partir de deducción lógica de los algoritmos, tampoco es la búsqueda de la congruencia elegantes entre los modelos matemáticos, estructuras físicas o procesos dinámicos que se suelen buscar en matemáticas, incluso si estas formas de razonamiento puedan algunas veces generar discusiones acerca del desarrollo de nuevos enfoques algorítmicos para determinadas clases de problemas, estas discusiones juegan un segundo rol en lo que es la visión más pragmática lógica realizable y práctica de la eficiencia computacional del algoritmo.

La Ciencia de Datos puede ser considerada un esfuerzo colectivo para el desarrollo de algoritmos y manejo de datos, en lugar de una ruta de prueba y refutación lógica como son las matemática. Esta es la razón por la que la eficiencia de los algoritmos ha surgido como un aspecto relevante para las personas que trabajan tanto en el mundo académico como corporativo con ciencia de datos. La eficiencia se convierte en el problema a solucionar para definir si un algoritmo es mejor que otro.

La ejecución de experimentos o procedimientos de investigación científicos se realiza de forma diferenciada en el caso en que se usen técnicas de Ciencias de Datos. Se puede notar que se requiere del uso de un conjunto de expresiones técnicas relacionadas al propio ámbito de la Ciencia de Datos, en particular expresiones características de los lenguajes de programación y los algoritmos, por lo que se requiere una preparación académica adicional y conocimiento experto generado a través de la práctica, ya que utilizar estas técnicas requiere ingresar a un área de conocimiento muy especializada, más

allá del área de conocimiento de la que el investigador se encuentre trabajando una investigación provenga. Al mismo tiempo, puesto que la investigación científica es una actividad social, el investigador entra en mayor contacto con otras personas involucradas y comprometidas con esas técnicas y las reglas que rigen los procesos de pensamiento colectivamente e individualmente. En línea con lo que describía Khun, este es un fenómeno general, pues mientras más se adentra un investigador en un área especializada, más fuerte será el lazo entre el pensamiento colectivo de la comunidad científica a la que pertenece y el trabajo individual que realiza el investigador.

La experiencia del investigador en Ciencia de Datos radica entonces en llevar a la práctica el uso más correcto de los algoritmos, y en estar actualizado en cuanto a la utilización de aquellos desarrollos de *software* que sean los más avanzados tecnológicamente, también que sean los más eficientes, y que sean así reconocidos por la comunidad de investigadores a los que pertenece. De lo contrario, se corre el riesgo de que las soluciones y resultados propuestos sean refutados o rechazados *a priori* tan sólo por el uso del algoritmo que se considere inadecuado por las características técnicas del algoritmo, como por la especialización que podría tener para determinadas aplicaciones.

Puesto que los términos que conforman el léxico que permite la comunicación entre expertos es especializada, cada forma de comunicación escrita u oral respecto a una investigación realizada utilizando ciencia de tiende a hacer cada elemento del conocimiento en este campo, más esotérica y menos popular. Como explica Mosterín (2016) *un lenguaje formal es un instrumento, un tinglado convencional que inventamos y construimos para poder describir más fácilmente las estructuras, permitiéndonos hablar simultáneamente de todos los sistemas de un mismo tipo de semejanza punto su utilidad es restringida sólo nos sirve para hablar de los sistemas de ese tipo de semejanza*. Pero, dentro de esa limitación, permite describir las estructuras que esos sistemas tienen o no tienen con mucha mayor claridad y precisión que el lenguaje ordinario u otros lenguajes. Un lenguaje formal consta de una serie de signos que constituyen su alfabeto, un conjunto de términos y un conjunto de fórmulas, que son ciertas combinaciones o secuencias de signos del alfabeto.

El lenguaje de la *Big Data* y la Ciencia de Datos cuenta con esas características, y su uso correcto autoriza a participar dentro de la comunidad de usuarios y expertos de estas técnicas. Asimismo, como explica Fleck (1979), el carácter social inherente a la naturaleza de la actividad científica tiene consecuencias sustantivas. Las palabras que inicialmente fueron simples términos se convierten en slogans, las oraciones que alguna

vez fueron simple simples declaraciones se convierten en llamados a la batalla. Esto altera completamente el valor socio-cognitivo de una determinada técnica sobre otra. No logra influenciar a las mentes sólo a través de su significado lógico, sino que tienen un poder mágico y ejercen una influencia mental simplemente por el hecho de ser usadas. Como un ejemplo uno puede considerar el efecto de términos como materialismo o ateísmo los cuales en algunas comunidades desacreditan a priori a aquellos que las proponen, pero en otros funcionan como claves esenciales para la aceptación. Existen términos similares cuando se trata de la investigación con Ciencia de Datos, y es importante que el investigador que hace uso de esta metodología no caiga en el sectarismo que da autoridad a la herramienta por sobre el problema a investigar y sus conclusiones. Es menester escapar de esa autoridad, y ponerla en el lugar de herramienta que es donde debe estar.

3.3 Un enfoque epistemológico para la Ciencia de Datos

Para entender la forma en la que se justifica el conocimiento que se genera a través de la Ciencia de Datos, es necesario realizar un abordaje del problema desde dos frentes, uno epistemológico, y el otro fenomenológico (Bachimont, 2015). El primero exige que se comprenda la forma en la que las representaciones de los datos permiten formar un ámbito de conocimiento y no sólo de representación, es decir, se debe entender cómo es que la *Big Data* y las técnicas de la ciencia de datos proponen nuevos conocimientos y nuevos descubrimientos científicos, y cómo se da la justificación de la evidencia, mientras que la segunda corresponde al hecho de articular la explicación generada en respuesta a la pregunta de investigación siempre dentro del marco de tratamientos algorítmicos efectuados sobre la *Big Data* y que deben mantenerse relacionados a las actividades y hechos que se estudian científicamente, es decir, mantener siempre la explicación dentro de los márgenes en los que la *Big Data* debe preservar el significado de los hechos humanos y sociales incluso a través de la mediación del tratamiento matemático de la información y de los algoritmos y procesos de visualización que se utilicen para dar a conocer los resultados obtenidos.

Por otro lado, de acuerdo a Fleck (1979), para la epistemología no es importante investigar como se descubrió una conexión entre fenómenos, sino sólo como legitimizar esa conexión científicamente, cómo probarla objetivamente, y cómo construirla lógicamente. En el caso de la Ciencia de Datos, este enfoque epistemológico debe aplicarse de forma similar, y para efectos de la presente disertación, la existencia de

evidencia empírica cumple un rol fundamental. La legitimización del conocimiento obtenido es muy importante dentro de los límites razonables que las técnicas permiten y con la precisión adecuada, pero no necesariamente, esta verificación debe ser la única o más importante tarea de la epistemología en Ciencia de Datos, o en las mismas formas de examinar la consistencia de conceptos y sus interconexiones dentro de un sistema científico tradicional. Es usual que todo lo que es ya conocido siempre ha parecido sistemático, probado, aplicable y evidente para el que conoce sus procedimientos, técnicas y lenguaje. Sin embargo, todo sistema de conocimiento que se encuentra fuera, ya sea por un cambio de paradigma de estilo Kuhniano o por la aparición de nuevas técnicas como las de la Big Data y la Ciencia de Datos siempre ha parecido en un inicio contradictorio, improbadado, inaplicable, pretencioso y hasta místico. A continuación se discuten algunos problemas epistemológicos presentados por la irrupción de la *Big Data* y la Ciencia de Datos en los cuales la existencia de una concepción de evidencia empírica impacta sustantivamente.

3.3.1 La investigación con Ciencia de Datos y la hipótesis de investigación

Indica Maletta (2015) que en investigación existen diferentes tipos de hipótesis, en las investigaciones descriptivas y explicativas los criterios básicos de éxito radican en la corroboración de hipótesis sustantivas y de hipótesis estadísticas. Las hipótesis sustantivas son las que postulan relaciones entre las variables, generalmente relaciones de tipo causal. Las hipótesis estadísticas o hipótesis de generalización, son las que se refieren a la significatividad o generalizada de los resultados empíricos obtenidos. Las hipótesis sustantivas, y en particular las hipótesis causales, sólo suelen ponerse a prueba en los diseños explicativos.

El objetivo principal de la producción científica es entender el mundo esto implica encontrar una explicación aceptable de lo que ocurre pero el intento de explicar una regularidad empírica no es el único tipo de problema científico. Algunas investigaciones sólo pretenden describir ciertos hechos o clasificarlos o predecir los algunos pueden permanecer en el plano teórico como para formular criticar o reformular una determinada teoría, o bien interpretar el significado de las teorías de diferentes autores. Algunas procuran proponer y justificar nuevos métodos o técnicas de investigación.

Las hipótesis y las teorías no son similares. Lo que vale para las primeras puede no valer para las segundas, ya que estas no son proposiciones sino conjuntos infinitos de

proposiciones (Bunge, 2002). Mientras las hipótesis deben enriquecerse con datos para poderse encontrar ser contrastadas, las teorías deben ser enriquecidas con datos y con hipótesis adicionales (Bunge, 2002). Por ejemplo para poner a prueba una teoría mecánica es preciso añadir la hipótesis sobre la composición del sistema que se quiere estudiar, las fuerzas que actúan entre sí, y sobre sus componentes. Al ser enriquecidas de esta manera las teorías dejan de ser completamente generales y en cambio aumenta su capacidad de ser porque se tornan capaces de formular predicciones precisas (Bunge, 2002). La regla se convierte entonces en que a mayor generalidad hay menor contraste y viceversa. Para que una idea sea considerada científica es necesario que sea contrastable con la experiencia empírica.

Es claro entonces que la investigación científica se ha relacionado tradicionalmente al uso de hipótesis, sin embargo, el proceso que sigue la ciencia tradicional no necesariamente es el que se utiliza cuando se realizan investigaciones en Ciencia de Datos, pues la necesidad no es la de comprobar una hipótesis, sino la de, por ejemplo, identificar patrones de comportamiento a partir de los cuales explicar que es lo que sucede en el mundo. Por lo que la hipótesis pierde el rol central con el que tradicionalmente ha contado.

Para llevar adelante una investigación ya sea científica de metodología tradicional o de Big Data y Ciencia de Datos, es menester apropiarse de los conocimientos relevantes, advertir lo que se ignora, escoger lo que se quiere averiguar, plantear la manera de hacerlo, etcétera. El método que se siga no suple a estos conocimientos, decisiones, planes, etcétera sino que ayuda a ordenarlos, precisarlos y enriquecerlos. El método representa una actitud más que una receta para resolver problemas. Tanto es así que la mejor manera de aprender a plantear o resolver problemas científicos no es estudiar un manual de metodología escrito por algún filósofo, sino estudiar e imitar paradigmas o modelos de investigación exitosa (Kuhn 1970).

En ese sentido, incluso si la investigación con Big Data y Ciencia de Datos no necesariamente requiere de una hipótesis de investigación, para que el trabajo proceda en línea con el método científico sería necesario que cumpla o al menos se proponga cumplir con las siguientes etapas que propone Bunge (2002) para la investigación científica y que pueden ser aplicables también para la investigación con *Big Data*:

- i) Descubrimiento del problema o brecha en un conjunto de conocimientos. Si el problema no está enunciado con claridad, se pasa en la etapa siguiente si lo está se pasa a las subsiguientes.

- ii) Planteamiento preciso del problema, en lo posible en términos matemáticos o bien replanteamiento de un viejo problema a la luz de nuevos conocimientos empíricos o teóricos sustantivos o metodológicos.
- iii) Búsqueda de conocimientos o instrumentos relevantes al problema, por ejemplo, datos empíricos teorías aparatos de medición técnicas de cálculo de medición o sea inspección de lo conocido para ver si puede resolver el problema. En el caso particular de *Big Data* y Ciencia de Datos, se puede considerar la selección de bases de datos especializadas y algoritmos más apropiados.
- iv) Tentativa de solución del problema con ayuda de los medios identificados. Si el intento falla se pasa a la siguiente si no a las subsiguientes.
- v) Generación de nuevas ideas, hipótesis, teorías o técnicas de producción de nuevos datos empíricos que prometan resolver el problema. En el caso de la investigación con *Big Data* y Ciencia de Datos es muy apropiado el término generación de nuevas ideas.
- vi) Obtención de una solución exacta o aproximada del problema con ayuda del instrumental conceptual o empírico disponible. En el caso de la investigación con *Big Data* y Ciencia de Datos se incluyen los algoritmos utilizados.
- vii) Investigación de las consecuencias de la solución obtenida. Si se trata de una teoría, búsqueda de predicciones que puedan hacerse con su ayuda si se trata de nuevos datos, examen de las consecuencias que puedan tener para las teorías relevantes. Es en este paso en el que hace importante la consideración de evidencia, y en particular de evidencia empírica, en la investigación con *Big Data*.
- viii) Puesta a prueba, contrastación de la solución y confrontación de ésta con la totalidad de las teorías y de la información empírica pertinente. Si el resultado es satisfactorio, la investigación se da por concluida hasta nuevo aviso. Si no, se pasa a la siguiente etapa. En el caso de la investigación con *Big Data* y Ciencia de Datos se puede considerar replicar los resultados con otros algoritmos o conjuntos de datos.
- ix) Corrección de las hipótesis, teorías, procedimientos o datos empleados en la obtención de la solución incorrecta. Este es, por supuesto, el comienzo de un nuevo ciclo de investigación.

Es claro entonces que la investigación científica se ha relacionado tradicionalmente al uso de hipótesis, sin embargo, el proceso descrito bien puede adaptar varios de sus pasos a la metodología que sigue la investigación con Ciencia de Datos, aunque la necesidad no es la de comprobar una hipótesis, sino la de identificar patrones de comportamiento a partir de las cuales explicar que es lo que sucede en el mundo, hay aspectos formales del procedimiento que se reflejan entre una metodología y la otra. Al validar entonces que se pueda seguir un proceso similar, se argumenta que la hipótesis está perdiendo el rol central con el que tradicionalmente ha contado en la investigación científica tradicional.

3.3.2 El Rol de las Teorías en la Ciencia de Datos

Como explica Moulines (2011), las teorías son convenciones complejas que se adoptan para organizar conceptualmente la masa de datos que se obtienen de las observaciones y de los experimentos de laboratorio. Sin estas convenciones, el investigador se perdería en los detalles, por lo que necesita realizar este tipo de abstracción. Siguiendo esta proposición se podría decir que la investigación con Ciencia de Datos podría ser una fuente generadora de teorías en la medida en que su conocimiento quede claramente justificado en las condiciones de la investigación científica que se presentaron en el apartado precedente. Sin embargo estas formas de organización de los datos no necesariamente reflejan alguna verdad trascendental ni pueden pretender ser la imagen de la realidad en sí, incluso si se habla de un volumen enorme de datos, pues aunque las características de la *Big Data* y las propuestas de consideraciones para la existencia de evidencia empírica, presentada previamente en esta disertación, se cumplan, al igual que en el caso de la ciencia tradicional, no se podría alcanzar aúncerteza en cuanto a la representación fiel de la realidad o de verdades trascendentales por la naturaleza misma de los datos y su procesamiento.

Poincaré admite que podemos estar frente a teorías divergentes en competencia por el mismo dominio de la experiencia pero el criterio para decidir entre esas teorías no es saber cuál es la verdadera en sentido estricto ni siquiera cuál es la menos falsa sino más bien saber cuál de las convenciones adoptadas resulta más útil es decir la más simple para organizar la experiencia. Esto mismo se puede aplicar al caso de la investigación con *Big Data* y Ciencia de Datos.

3.3.3 Causalidad en la ciencia de datos

De acuerdo a Rosenberg (2005), la explicación científica es una explicación causal. Los científicos buscan causas y lo hacen porque la ciencia busca explicaciones que permitan controlar y predecir fenómenos y esto sólo lo puede lograr el conocimiento de las causas que los generan. Por el contrario, la correlación suele ser muchas veces arbitraria y no permite la secuenciación de eventos de forma empíricamente contrastable.

Es por ello relevante considerar, como explican Balazka & Rodighiero (2020), que la relación entre correlación y causación es causa de un enfrentamiento entre las diversas maneras en que la Ciencia de Datos es utilizada por los investigadores de empresas en la industria y por investigadores de universidades que pertenecen al ámbito académico. Muchas investigaciones del ámbito industrial se centran en la correlación y no en la causalidad, a diferencia de los objetivos del ámbito académico. Estas diferentes perspectivas separan a los expertos en Ciencia de Datos de los científicos que usan Ciencia de Datos, ocasionando que la ciencia de datos presente aplicaciones diferentes en la industria y en la academia. Los expertos en Big Data dicen que producen mejor ciencia que los propios científicos, incluso desafiando los estándares y prácticas del método científico y confiando, de forma tal vez exagerada, en la cantidad, la calidad y la computabilidad de los datos para obtener diversos resultados reproducibles. El factor que debe zanjar esta discusión es el uso y contraste de la evidencia empírica como medio para determinar la legitimidad de una investigación sobre otra.

3.4 Hacia un nuevo tipo de conocimiento. La epistemología de los datos.

La científicidad de las investigaciones realizadas con *Big Data* y Ciencia de Datos no se puede basar en el modelo tradicional del método científico, puesto que no existe un marco que sea común al modelo tradicional y al modelo que hace uso de los Metadatos, de las leyes que gobiernan su procesamiento, y de los aspectos estéticos de su representación visual. Incluso siguiendo las etapas presentadas en la sección 3.3.1 como una forma de alinear una metodología con la otra, no se puede eliminar toda posibilidad de objeción y reparo.

En Ciencia de Datos, a priori, no hay alguna característica que permita comprender en que medida el resultado que se presenta en la investigación es científico, y aún más, no es fácil determinar que estatus darle a la investigación y de que forma interpretarla científicamente. En consecuencia, la Ciencia de Datos pareciera no ser la herramienta más útil para abordar las preguntas tradicionales que se enfrentan utilizando el método científico (Bachimont, 2015), permitiendo resolverlas de una forma diferente, principalmente a través de nuevas técnicas que enfrentan problemas más difíciles que los tradicionales, principalmente por su tamaño y masa de información a procesar. Estas diferencias, evidencian que la Ciencia de Datos representa una ruptura epistemológica, y hace necesario construir una epistemología a su altura, la epistemología de los datos

La Ciencia de Datos pone frente a frente por un lado a la motivación del paradigma clásico de investigación, que fundamenta el trabajo dentro de un cuestionamiento a la realidad, que sigue ciertas leyes y principios que permitirán interpretarla por un lado, y por otro a la arbitrariedad que poseen los datos, la manipulación que se realiza sobre ellos y la técnica de visualización seleccionada para mostrar los resultado, operaciones que no son motivadas por la naturaleza de lo real y las hipótesis que se formulan para ellas, sino por elecciones que realiza el investigador a partir de su experiencia y de la comunidad de investigadores a la que pertenece.

La arbitrariedad que existe en la ciencia de datos, es necesaria e inevitable pues permite aproximar las realidades heterogéneas unas a otras a partir de la recolección de información de fuentes distintas, para superar el desafío que implica trabajar con cantidades gigantescas de datos y los tratamientos matemáticos asociados, y por la complejidad de los cálculos realizados para la visualización y representación estética de los resultados para su presentación y defensa.

La ciencia basada en Big Data y Ciencia de Datos renueva la primacía del razonamiento inductivo en la forma de empirismo fundamentado en la tecnología y ha inspirado una visión de futuro en la que la minería de datos automatizada permitirá generar nuevo conocimiento. De acuerdo a esta perspectiva, la nueva forma de hacer ciencia con una *hipótesis neutral* reemplazará a la forma de hacer ciencia mediante al prueba de hipótesis. El análisis de grandes volúmenes de datos generará nuevas y sorprendentes correlaciones, patrones y reglas, en un proceso emergente basado en la inducción y la manipulación estadística, haciendo innecesaria la teoría. Esos patrones nacerán desde los datos y generarán nuevas hipótesis de investigación en el proceso. De esta manera, el enfoque de ciencia de datos podría ser visto como un generador de

hipótesis, en contraste con el carácter general de la ciencia tradicional que es el de probar hipótesis (Mazzocchi, 2015).

Esta es la razón principal por lo que la discusión que permita esclarecer la relación entre los datos y la evidencia empírica es relevante para la investigación con *Big Data* y Ciencia de Datos, puesto que su potencial de generar conocimiento debe ser igualado con la solidez de su sustento en evidencia, principalmente, de evidencia empírica que conforma una de las bases de la investigación científica que ha impulsado el desarrollo del conocimiento en la humanidad en los últimos siglos.

BIBLIOGRAFÍA

- Achinstein, P. (2001). *The Book of Evidence*. Oxford University Press.
- Ackoff, R. (2005). From Data to Wisdom. *Dialogue and Universalism*, 15(5), 55–71. <https://doi.org/10.5840/du2005155/629>
- Ahonen, P. (2015). Institutionalizing Big Data methods in social and political research. *Big Data and Society*, 2(2), 1–12. <https://doi.org/10.1177/2053951715591224>
- Allen, J. F. (2001). Hypothesis, induction and background knowledge. Data do not speak for themselves. Replies to Donald A. Gillies, Lawrence A. Kelly and Michael Scott. *BioEssays*, 23(9), 861–862. <https://doi.org/10.1002/bies.1125>
- Appley, B. C., & Stoutenburg, G. (2017). Two new objections to explanationism. *Synthese*, 194(8), 3069–3084. <https://doi.org/10.1007/s11229-016-1093-1>
- Artola, M., & Sánchez Ron, J. M. (2012). *Los Pilares de la Ciencia*. Espasa Libros.
- Bachimont, B. (2015). Le numérique comme milieu: enjeux épistémologiques et phénoménologiques. *Interfaces Numériques*, 4. <https://doi.org/10.3166/RIN.4.385>
- Balazka, D., & Rodighiero, D. (2020). Big Data and the Little Big Bang: An Epistemological (R)evolution. *Frontiers in Big Data*, 3(September), 1–13. <https://doi.org/10.3389/fdata.2020.00031>
- Boyd, N. M., & Bogen, J. (2021). Theory and Observation in Science. Retrieved from <https://plato.stanford.edu/archives/win2021/entries/science-theory-observation/>
- Brown, M. J. (2015). The functional complexity of scientific evidence. *Metaphilosophy*, 46(1), 65–83. <https://doi.org/10.1111/meta.12123>
- Bunge, M. (2002). *Epistemología*. Barcelona: Siglo XXI Editores.
- Canali, S. (2016). Big Data, epistemology and causality: Knowledge in and knowledge out in EXPOsOMICS. *Big Data and Society*, 3(2), 1–11. <https://doi.org/10.1177/2053951716669530>
- Cano, I. C. M. (2016). Reflexiones epistemológicas sobre Big Data. *Eikasía Revista de Filosofía*, julio, 451–474.
- Cheung, K. S., Leung, W. K., & Seto, W. K. (2019). Application of Big Data analysis in gastrointestinal research. *World Journal of Gastroenterology*, 25(24), 2990–3008. <https://doi.org/10.3748/wjg.v25.i24.2990>
- Conway, D. (2010). The Data Science Venn Diagram. Retrieved from

- <http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>
- Copi, I. M., & Cohen, C. (2013). *Introducción a la Lógica*.
- Crawford, K. (2013). The Hidden Biases in Big Data.
- Davenport, T. H., & Prusak, L. (1997). *Information ecology*. Oxford University Press.
- Driscoll, K., & Walker, S. (2014). Working within a black box: Transparency in the collection and production of big twitter data. *International Journal of Communication*, 8(1), 1745–1764.
- Drucker, P. (1988). The Coming of the New Organization. Retrieved from <https://hbr.org/1988/01/the-coming-of-the-new-organization>
- Efron, B., & Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press.
- Favaretto, M., de Clercq, E., Schneble, C. O., & Elger, B. S. (2020). What is your definition of Big Data? Researchers' understanding of the phenomenon of the decade. *PLoS ONE*, 15(2), 1–20. <https://doi.org/10.1371/journal.pone.0228987>
- Fleck, L. (1979). *Genesis and Development of a Scientific Fact*. The University of Chicago Press.
- Floridi, L. (2005). Por Una Filosofía De La Información. *Anthropos*, 1–12.
- Floridi, L. (2014). The 4th. Revolution. *Oxford Press*, 248.
- Fontes, M. (2005). The process of transformation of scientific and technological knowledge into economic value conducted by biotechnology spin-offs. *Technovation*, 25(4), 339–347. <https://doi.org/10.1016/j.technovation.2003.08.004>
- French, S. (2007). *Science, key concepts in philosophy*. Manchester: Continuum.
- Fricke, M. (2013). Big Data and Its Epistemology. *Journal of the American Society for Information Science and Technology*, 64(July), 1852–1863. <https://doi.org/10.1002/asi>
- Frické, M. (2015). Big Data and Its Epistemology. *Journal of the Association for Information Science and Technology*, 66(April), 651–661. <https://doi.org/10.1002/asi>
- Gordin, M. D. (2021). *On the Fringe*. New York: Oxford University Press.
- Günther, W. A., Rezazade Mehrizi, M. H., Huysman, M., & Feldberg, F. (2017). Debating big data: A literature review on realizing value from big data. *Journal of Strategic Information Systems*, 26(3), 191–209. <https://doi.org/10.1016/j.jsis.2017.07.003>
- Gupta, S., & Saxena, S. (2016). *Real-Time Big Data Analytics*.
- Hartnack, J. (2021). *La teoría del conocimiento de Kant* (11ma ed.). Madrid: Ediciones Cátedra.
- Hekler, E. B., Klasnja, P., Chevance, G., Golaszewski, N. M., Lewis, D., & Sim, I. (2019). Why we need a small data paradigm. *BMC Medicine*, 17(1), 1–9.

- <https://doi.org/10.1186/s12916-019-1366-x>
- Helmond, A. (2014). 'Raw data' is an oxymoron. *Information, Communication & Society* (Vol. 17). <https://doi.org/10.1080/1369118x.2014.920042>
- Hennig, C. (2010). A constructivist view of the statistical quantification of evidence. *Constructivist Foundations*, 5(1), 39–54.
- Hirschheim, R. (2021). The attack on understanding: How big data and theory have led us astray: A comment on Gary Smith's Data Mining Fool's Gold. *Journal of Information Technology*, 36(2), 176–183. <https://doi.org/10.1177/0268396220967677>
- Johansson, L.-G. (2016). *Philosophy of science for scientists*. Uppsala: Springer.
- Johansson, T. (2011). Hail the impossible: P-values, evidence, and likelihood. *Scandinavian Journal of Psychology*, 52(2), 113–125. <https://doi.org/10.1111/j.1467-9450.2010.00852.x>
- Jones, M. (2019). What we talk about when we talk about (big) data. *Journal of Strategic Information Systems*, 28(1), 3–16. <https://doi.org/10.1016/j.jsis.2018.10.005>
- Kant, I. (2002). *Crítica de la Razón Pura I*. (J. del Perojo & J. Rovira Armengol, Eds.). Barcelona: Ediciones Folio.
- Lee, A. J., & Cook, P. S. (2020). The myth of the "data-driven" society: Exploring the interactions of data interfaces, circulations, and abstractions. *Sociology Compass*, 14(1), 1–14. <https://doi.org/10.1111/soc4.12749>
- Leonelli, S. (2019a). Cinq façons dont les données massive nuisent à la.
- Leonelli, S. (2019b). Data Governance is Key to Interpretation: Reconceptualizing Data in Data Science. *Harvard Data Science Review*, 1–9. <https://doi.org/10.1162/99608f92.17405bb6>
- Leonelli, S. (2020). Scientific Research and Big Data.
- Lewis, D. (1996). Elusive knowledge. *Australasian Journal of Philosophy*, 74(4), 549–567. <https://doi.org/10.1080/00048409612347521>
- Loeb, L. E. (1995). Hume on Stability, Justification, and Unphilosophical Probability. *Journal of the History of Philosophy*, 33(1), 101–132. <https://doi.org/10.1353/hph.1995.0004>
- Lowrie, I. (2017). Algorithmic rationality: Epistemology and efficiency in the data sciences. *Big Data and Society*, 4(1), 1–13. <https://doi.org/10.1177/2053951717700925>
- Maletta, H. (2015). *Hacer Ciencia*. Universidad del Pacífico.
- Mashey, J. (1999). Big Data and the Next Wave of InfraStress Problems, Solutions, Opportunities. Retrieved from <https://www.usenix.org/conference/1999-usenix-annual-technical-conference/big-data-and-next-wave-infrastress-problems>

- Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work and Think*.
- Mazzocchi, F. (2015). Could Big Data be the end of theory in science? *EMBO Reports*, 16(10), 1250–1255. <https://doi.org/10.15252/embr.201541001>
- Mehra, M. R., Ruschitzka, F., & Patel, A. N. (2020). Retraction—“Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis” (The Lancet, (S0140673620311806), (10.1016/S0140-6736(20)31180-6)). *The Lancet*, 395(10240), 1820. [https://doi.org/10.1016/S0140-6736\(20\)31324-6](https://doi.org/10.1016/S0140-6736(20)31324-6)
- Miller, H. J. (2010). The data avalanche is here. Shouldn't we be digging? *Journal of Regional Science*, 50(1), 181–201. <https://doi.org/10.1111/j.1467-9787.2009.00641.x>
- Mosterín, J. (2016). *Conceptos y teorías en la ciencia*. (A. Editorial, Ed.) (Cuarta). Madrid.
- Moulines, C. U. (2011). *El Desarrollo Moderno de la Filosofía de la Ciencia (1890-2000)* (1era ed.). México.
- Obar, J. A. (2020). Sunlight alone is not a disinfectant: Consent and the futility of opening Big Data black boxes (without assistance). *Big Data and Society*, 7(1). <https://doi.org/10.1177/2053951720935615>
- Otero, E., & Gibert, J. (2016). *Diccionario de Epistemología*. Santiago de Chile.
- Raoult, D. (2020). Lancet gate: a matter of fact or a matter of concern. *New Microbes and New Infections*, 38, 100758. <https://doi.org/10.1016/j.nmni.2020.100758>
- Rendueles Mata, M., & Dreher Grosch, M. (2007). La epistemología y los sistemas de información basados en inteligencia artificial. *Revista Electrónica de Estudios Telemáticos*.
- Rosenberg, A. (2005). *Philosophy of science*. New York: Routledge.
- Russell, B. (1988). *El panorama de la ciencia*. Ercilla.
- Sagiroglu, S., & Sinanc, D. (2013). Big data: A review. *Proceedings of the 2013 International Conference on Collaboration Technologies and Systems, CTS 2013*, 42–47. <https://doi.org/10.1109/CTS.2013.6567202>
- Sanjuán, M. (2021). On upholding empirical evidence. McIntyre in the Scientific Attitude. *Daimon*, 0507(82), 189–195. <https://doi.org/10.6018/DAIMON.436061>
- Sloot, B., Broeders, D., & Schrijvers, E. (2016). Exploring the Boundaries of Big Data. *Exploring the Boundaries of Big Data*. <https://doi.org/10.5117/9789462983588>
- Störig, H. J. (2016). *Historia Universal de la Ciencia*. (Editorial Tecnos, Ed.). Madrid.
- Succi, S., & Coveney, P. V. (2019). Big data: The end of the scientific method? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and*

- Engineering Sciences*, 377(2142). <https://doi.org/10.1098/rsta.2018.0145>
- Symons, J., & Alvarado, R. (2016). Can we trust Big Data? Applying philosophy of science to software. *Big Data and Society*, 3(2), 1–17. <https://doi.org/10.1177/2053951716664747>
- Taroni, F., Garbolino, P., Bozza, S., & Aitken, C. (2021). The Bayes' factor: The coherent measure for hypothesis confirmation. *Law, Probability and Risk*, 20(1), 15–36. <https://doi.org/10.1093/lpr/mgab007>
- van Benthem, J., & Pacuit, E. (2011). Dynamic Logics of Evidence-Based Beliefs. *Studia Logica*, 99(1), 61–92. <https://doi.org/10.1007/s11225-011-9347-x>
- Van Der Sloot, B., Broeders, D., & Schrijvers, E. (2016). *Exploring the Boundaries of Big Data*.
- Ward, T. (2020). Explaining and trusting expert evidence: What is a 'sufficiently reliable scientific basis'? *International Journal of Evidence and Proof*, 24(3), 233–254. <https://doi.org/10.1177/1365712720927622>
- Williamson, J. (2017). Models in systems medicine. *Disputatio*, 9(47), 429–469. <https://doi.org/10.1515/disp-2017-0014>