Sophia

Volumen 21, número 1, 2025, p: e1440 | págs. 1-13
ISSN 2346-0806
DOI: https://doi.org/10.18634/sophiaj.21v.1i.1440

ARTÍCULO DE REFLEXIÓN

Una oteada reflexiva hacia el contexto epistemológico de la analítica de datos

A reflective look at the epistemological context of data analytics

Um olhar reflexivo sobre o contexto epistemológico da análise de dados

Ximena Cifuentes Wchima*, John Edward Herrera Quintero[1], Luis Miguel Mejía Giraldo[1*], Luis Fernando Restrepo Betancur[2], Bibiana Vélez Medina[1]

[1]Universidad La Gran Colombia. Armenia, Colombia.
[2]Universidad de Antioquia. Medellín, Colombia.
***Autor para la correspondencia:** mejiagluismiguel@miugca.edu.co

RESUMEN

La abundancia y prominencia moderna de los datos ha llevado al desarrollo de la "ciencia de datos" como un nuevo campo de investigación, junto con un cuerpo de reflexiones epistemológicas sobre sus fundamentos, métodos y consecuencias. Este artículo es derivado del ejercicio investigativo sobre los fines de la educación donde el análisis del conocimiento proporciona una dinámica sistemática y una revisión crítica de importantes problemas y debates abiertos en la epistemología de la analítica y la ciencia de datos, proponiendo una división de la epistemología de la ciencia de datos en los siguientes cinco aspectos: Caracterizaciones maximalistas y minimalistas, taxonomías descriptivas, el conocimiento generado por la ciencia de datos, problemas de la caja negra y la ciencia en un paradigma intensivo en datos, aspectos que brindan un ejercicio reflexivo frente a la comprensión y abordaje de aspectos esenciales de la interpretación de datos y la comprensión de patrones ocultos en ellos, siendo esto el reto de la analítica como tal.

Palabras clave: ciencia de datos; campo de investigación; educación e investigación; epistemología.

ABSTRACT

The modern abundance and prominence of data has led to the development of "data science" as a new field of research, along with a body of epistemological reflections on its foundations, methods, and consequences. This article is derived from the research exercise on the purposes of education where the analysis of knowledge provides a systematic dynamic and a critical review of important problems and open debates in the epistemology of analytics and data science, proposing a division of epistemology of data science in the following five aspects: Maximalistic and minimalist characterizations, descriptive taxonomies, the knowledge generated by data science, black box problems and science in a data-intensive paradigm, aspects that provide a reflective exercise against to understanding and addressing essential aspects of data interpretation and understanding hidden patterns in them, this being the challenge of analytics as such.

Keywords: data science; field of research; education and research; epistemology.

RESUMO

A abundância e a proeminência modernas dos dados levaram ao desenvolvimento da "ciência dos dados" como um novo campo de investigação, juntamente com um conjunto de reflexões epistemológicas sobre os seus fundamentos, métodos e consequências. Este artigo deriva do exercício de investigação sobre os propósitos da educação onde a análise do conhecimento proporciona uma dinâmica sistemática e uma revisão crítica de problemas importantes e debates abertos na epistemologia da análise e da ciência de dados, propondo uma divisão da epistemologia da ciência de dados em os seguintes cinco aspetos: caracterizações maximalistas e minimalistas, taxonomias descritivas, o conhecimento gerado pela ciência de dados, problemas de caixa negra e ciência num paradigma intensivo de dados,

Recibido: julio de 2024 | Aceptado: octubre de 2025 | Publicado: noviembre de 2025

Como citar:
Cifuentes Wchima, X., Herrera Quintero, J.E., Mejía Giraldo, L. M., Restrepo Betancur, Luis F. Vélez Medina, B. (2025). Una oteada reflexiva hacia el contexto epistemológico de la analítica de datos. Sophia, 21(1). https://revistas.ugca.edu.co/index.php/sophia/article/view/1440
Conflicto de intereses: Los autores declaran no tener conflictos de intereses.

Sophia-Education
Copyright 2025. Universidad La Gran Colombia
Esta obra está bajo una Licencia Attribution-ShareAlike 4.0 International

aspetos que proporcionam um exercício reflexivo contra a compreensão e abordagem de aspetos essenciais da interpretação e compreensão dos dados padrão neles ocultos, sendo este o desafio da análise enquanto tal.

Palavras-chave: Ciência de dados; área de pesquisa; educação e pesquisa; epistemologia.

Introducción

El contexto mundial actual, caracterizado por la hiperconectividad y la generación de altos volúmenes de datos ha conllevado la necesidad del estudio de estos y llamarlo como "ciencia de datos", el cual se ha convertido en tiempo reciente en un campo de investigación, desarrollo e innovación profundo e impulsado por la proliferación de datos y una infraestructura informática cada vez más compleja. Sin embargo, aunque existen diversos trabajos sobre los problemas filosóficos de la dinámica analítica de datos, dichos problemas rara vez se han unificado en una "epistemología de la analítica y ciencia de datos" con una visión holística. Ahora, para comprender la naturaleza de dichos datos, se requiere de una caracterización de los mismos, abarcando alternativas de carácter descriptivo e incluso normativo como son las caracterizaciones maximalistas y minimalistas.

Caracterizaciones maximalistas y minimalistas

Cuando se remonta a los orígenes de la llamada hoy ciencia de datos comenzó a desarrollarse con base en el estudio sistemático de la estadística y la probabilidad, en primera instancia a través del análisis de juegos de azar hacia finales de la era del Renacimiento (Hacking, 1975) y posteriormente a través de análisis sociológicos en torno a la Revolución Industrial, tal como lo resaltan Gigerenzer et al. (2013) y culminando con el surgimiento y consolidación del análisis estadístico de la genética en Gran Bretaña en su época victoriana tardía (MacKenzie, 1984), esto conllevó que los incentivos económicos fueran primordiales en todo momento, ya fueran estos a través de tablas actuariales más precisas, mejores estrategias de juego o incluso reportes de rendimientos agrícolas. Ya en los albores del siglo XX, la estadística llegó a ser reconocida como una disciplina académica soportada y argumentada desde de sus propias redes académicas, publicaciones científicas y departamentos universitarios. A su vez, los avances tecnológicos de las décadas siguientes marcaron una ruptura con las estadísticas clásicas basadas en la inferencia y soportadas estrictamente en teorías, abriendo espacio a nuevos enfoques como las simulaciones Monte Carlo, el *Bootstrapping* y el análisis con base en cadena de Markov, cuestionando y reemplazando los supuestos paramétricos sólidos con un poder computacional bruto (Chambers, 1993), dinamizando -desde esta perspectiva- los algoritmos de aprendizaje autónomo, llevando a la detección y explotación automáticamente de patrones ocultos en grandes conjuntos de datos, siendo así, el siguiente paso lógico en la progresión de métodos de análisis a través del tiempo hacia las formas cada vez más automatizadas de razonamiento empírico como las conocemos en la actualidad. La cuestión de cuándo precisamente estas incursiones en métodos cuantitativos de análisis llevaron a lo que ahora se llama "ciencia de datos" presupone que la disciplina tiene posiblemente un carácter esencial aún no especificado y aunque existe escepticismo ante cualquier potencial "solución" al llamado "problema" de la denominación en esta área, como en la ciencia en general, se aprecian dos grandes tendencias en la literatura sobre este tema, que se consideran como los de tipo "minimalista" y "maximalista", respectivamente.

Por un lado, los minimalistas apuntan hacia las condiciones necesarias, lo menos restrictivas posible, pero aun así creando un espacio para comprender la ciencia de datos, mientras que los maximalistas propenden por lograr condiciones suficientes con ejercicios ontológicos detallados, así como taxonomías metodológicas. Es aquí, donde los enfoques minimalistas se caracterizan por los primeros debates sobre la naturaleza de la ciencia de datos, mientras los análisis contemporáneos tienden a abrir espacio a enfoques maximalistas, identificando en la ciencia de datos un medio para desarrollar conocimiento causal directamente conectado con el objeto de análisis (Carmichael y Marron, 2018).

Es de agregar que las concepciones minimalistas no comprometen la ciencia de datos con ningún

método o tema específico y no hacen inferencias y afirmaciones específicas sobre el tipo de disciplina que es la ciencia de datos y solo se enfocan en los aspectos pedagógicos y su dependencia de la información a partir de los datos; al respecto, Chambers (1993) presenta una visión de "Estadísticas mayores" de la ciencia de datos, caracterizadas como "*Todo lo relacionado con el aprendizaje a partir de datos*" y Carmichael y Marron (2018) afirman que la ciencia de datos es "*El negocio de aprender de los datos*" y que un científico de datos es alguien que usa datos para resolver problemas.

Cabe agregar que las explicaciones maximalistas son más detalladas, como lo resalta Breiman (2001), quien caracteriza la ciencia de datos por el tipo de conocimiento que esta genera y es aquí donde los estadísticos (analistas y los científicos de datos) que pueden estar interesados en hacer ejercicios de predicción correlativa a partir de datos y extraer información sobre cualquier mecanismo causal natural asociado y subyacente, ya que el conocimiento correlativo-predictivo y la causalidad se han convertido en algo implícitamente valorado. Además, se supone que el conocimiento causal se asocia directamente con mecanismos "naturales" y "subyacentes" de la realidad sea de manera directa como indirecta a través del análisis de asociación con la causalidad y es aquí, como la explicación maximalista proporcionada por Mallows (2006) resalta que se trata de un fin práctico donde la estadística se enfoca a la relación de los datos cuantitativos con un problema del mundo real, a menudo en presencia de la variabilidad y la incertidumbre e intenta hacer explícito y preciso el respectivo análisis de datos para dar respuesta a un problema de interés, enfatizando la resolución de problemas sobre la pedagogía general, aspecto que también destacan Blei y Smyth (2017) y Vélez (2018), quien a su vez destaca la identidad inmersa en las universidades, lo cual es una cuestión que implica el retorno hacia una serie de principios fundacionales; así como el reconocimiento de su historia, el impacto de su trayectoria y la reafirmación de sus ideales con base en tres principios fundentes como son el análisis de contexto (esencial para la comprensión de "*lo que se analiza*"), el conocimiento (en este caso el análisis estadístico) y la perspectiva de humanidad (lectura de las realidades desde el fundamento ético), ratificando aún más el ejercicio pedagógico.

Cabe resaltar que un aspecto interesante del planteamiento de Mallows cuando hace mención explícita de la variabilidad y la incertidumbre sobre los cuales se deben enfrentar los métodos estadísticos y científicos de datos, constituyendo un compromiso y un reto implícito de separación del mundo real y las construcciones idealizadas que se hacen familiares a las ciencias naturales y sociales. Por otro lado, Donoho (2017) apoya un enfoque maximalista haciendo referencia al "código de conducta profesional" del científico de datos, ya que es un profesional que utiliza métodos científicos para explorar y crear significado a partir de datos sin procesar, sean estos estructurados o no. Esto conlleva una estrecha conexión entre el análisis de datos y la investigación científica, no solo en los aspectos metodológicos, sino también en sus supuestos (para comprender la tipología y estructura del dato mirado como variable) y los objetivos específicos donde la estadística se subordina a estos y no al contrario, tal como lo resaltan Cifuentes et al (2016), lo cual refuerza lo planteado por Donoho (2017) con respecto a que los datos se originan a partir de procesos susceptibles de estudio y comprensión sistemáticos y como lo resalta el mismo autor, que desde la era del Big Data existen una serie de "datos brutos" que se convierten en una base adecuada para análisis a partir de unos objetivos claramente estructurados con el fin de destilar y crear significado, aspecto que puede ser una consecuencia de la actitud contemporánea de que los datos pueden ser registrados con suficiente profundidad, amplitud y calidad para dar respuesta a cualquier dominio problemático, tal como lo afirman Blei y Smyth (2017), quienes a su vez resaltan al caracterizar entre el minimalismo y el maximalismo: "*La ciencia de datos combina el pensamiento estadístico y computacional y se asocia a modelos estadísticos y métodos con enfoque computacional para la resolución de problemas específicos de una disciplina*", visión que restringe a la ciencia de datos únicamente con métodos estadísticos y computacionales, enfatizando una prioridad práctica más que pedagógica. No obstante, esta caracterización no especifica la información (concebida en sentido amplio) como objeto de interés de la ciencia de datos, ni señala disciplinas específicas que se puedan integrar como partes complementarias que permitan la profundización del estudio de los datos como ciencia.

Desarrollo

Taxonomías descriptivas

En el contexto del análisis estadístico, algunos autores han intentado caracterizarlo proporcionando una serie de taxonomías procedimentales de carácter descriptivo de la disciplina; sin embargo, tres relatos descriptivos, escritos en diferentes momentos a lo largo de las últimas seis décadas, brindan una perspectiva diacrónica, empezando por Tukey (1962), quien dio la primera taxonomía descriptiva de "análisis de datos" centrándose en procedimientos para analizar datos y técnicas para interpretar los resultados de tales procedimientos y las formas de planificar la recopilación de dichos datos para hacer su análisis más fácil y más preciso, con todo el proceso operativo así como los resultados que surgen del ejercicio estadístico-matemático que se aplica al "*analizar datos*", como lo asevera Tukey (1962), con base en la descripción diáfana de lo que realmente ocurre en el análisis de datos.

Posteriormente, Wu (1997) presentó una taxonomía descriptiva triple centrada en recopilación de datos (con base en encuestas por muestreo y diseño y análisis de experimentos), ¿el modelado y análisis de datos; así como la comprensión y la búsqueda de soluciones a problemas específicos y la posterior toma de decisiones -y al igual que el planteamiento de Tukey- esta descripción formaba parte de un proyecto más amplio para llevar la estadística matemática en una dirección científica. De manera más recientemente, está el relato de Donoho (2017), quien ha proporcionado una taxonomía extensa que cita el concepto de "*Iniciativa de ciencia de datos*" de la Universidad de Michigan, quien afirma que la ciencia de datos implica la recopilación, gestión, procesamiento, análisis, visualización e interpretación de grandes cantidades de datos heterogéneos. asociado con una amplia gama de aplicaciones científicas, traslacionales e interdisciplinarias (Donoho, 2017).

A su vez, Donoho (2017) ha proporcionado una taxonomía estadística integral para satisfacer las necesidades actuales emulando la terminología de Chambers (1993), donde la "*ciencia mayor de datos*" contrasta con algunas de las taxonomías descriptivas como se mencionó con antelación y que él llama "*ciencia menor de datos*".

Para Donoho, una ciencia mayor de datos consiste en seis fases que son:

1. Recopilación, preparación y exploración de datos.
2. Representación y transformación de datos.
3. Computación con datos.
4. Modelado de datos.
5. Visualización y presentación de datos.
6. Ciencia sobre ciencia de datos.

A la luz de estas consideraciones previas, Donoho (2017) propone la definición de ciencia de datos como el estudio de sistemas de información (naturales o artificiales), mediante razonamiento probabilístico (inferencia y predicción) implementado con herramientas computacionales (bases de datos y algoritmos), en concordancia con las dinámicas actuales que conciernen a la analítica de datos y a su respectiva construcción de conocimiento como se aborda a continuación.

El conocimiento generado por la ciencia de datos

Cuando se aborda el conocimiento generado por la ciencia de datos, el análisis se estructura en dos partes relacionadas: el proceso, o cómo, (en cuanto a los modos de inferencia) y el producto (refiriéndose a los productos epistémicos) de la ciencia de datos, así:

Modos de inferencia

Diferentes medios de investigación tienen diferentes afinidades con los tres modos típicos de inferencia: deducción, inducción y abducción. La epistemología de la ciencia de datos reflexiona sobre hasta

qué punto los científicos de datos implementan estos diversos modos.

En primera instancia, las inferencias deductivas están presentes en la ciencia de datos por medio de una gran dependencia del razonamiento matemático y lógico, donde el cálculo diferencial, la teoría de la probabilidad, el análisis funcional y la informática teórica son disciplinas puramente deductivas que son ampliamente utilizadas para aplicar las propiedades de los algoritmos y diseñar nuevos procedimientos de aprendizaje con bajo enfoque por el comportamiento empírico, como en el caso del algoritmo de retropropagación utilizado para optimizar parámetros en redes neuronales que combina elementos de álgebra lineal y cálculo multivalente en aras de la convergencia demostrable en un óptimo local de una función objetivo y donde no se requieren conjuntos de datos para obtener este resultado.

Por otro lado, están las inferencias inductivas basadas en el criterio de que los datos son una muestra finita del mundo, planteamiento basado en el criterio que el ejercicio analítico de la ciencia de datos identifica estructuras en los datos y las destila en información que se aplica más allá de los datos mismos, lo cual se logra proyectando los patrones y estructuras encontrados en los datos a nuevos contextos. Esta proyección es lo que se conoce como una inferencia inductiva, que representa una solución falible mediante la cual las pruebas estadísticas pueden proporcionar evidencia más fuerte o más débil a favor de hipótesis particulares (Mayo, 1996; 2018). Por esta razón, Harman y Kulkarni (2007) sostienen que la teoría estadística del aprendizaje representa una defensa sofisticada y basada en principios de la inducción, lo cual es corroborado por Frické (2015), quien observa que "*los algoritmos inductivos son un pilar central de la proyección del Big Data*", lo cual se suma a lo planteado por Breiman (2001) quien plantea que uno de los propósitos de la ciencia de datos es identificar estructuras y mecanismos subyacentes.

En cuanto a la inferencia abductiva, Alemany Oliver y Vayre (2015) han enfatizado la importancia del razonamiento abductivo en los métodos de la analítica y ciencia de datos, particularmente en cómo la ciencia de datos se integra en la práctica científica de manera más amplia argumentando que las herramientas de la ciencia de datos tienen interés en primera instancia en la exploración de los datos mismos para determinar su estructura interna y, en segunda instancia, en la identificación de las mejores hipótesis para explicar esta estructura convirtiendo esta inferencia en la estructura de una hipótesis explicativa y, por lo tanto, una inferencia abductiva (Harman, 1965; Lipton, 1991; Niiniluoto, 2018). El estatus de la abducción en un contexto intensivo en datos se eleva aún más por la virtud teórica de la unificación explicativa tal como lo resalta Kitcher (1989), quien resalta que una virtud común de una teoría es su poder explicativo, y algunos autores sostienen que ese poder es motivo para elegir teorías empíricamente equivalentes como lo resalta van Fraassen (1980) sobre las virtudes pragmáticas de este tipo de inferencias, abriendo espacio hacia una dimensión del poder explicativo es el alcance de la diversidad y la heterogeneidad de los fenómenos que una teoría puede explicar simultáneamente (Kitcher, 1976, Vélez, 2018) y si los métodos de la ciencia de datos permiten la identificación de patrones en una gama diversa y heterogénea de fenómenos (Nash, 1950, 1951), entonces se debe desarrollar una imagen más amplia y matizada del poder explicativo de las teorías que son fundantes en el ámbito estadístico y para aquellas teorías que pueden unificar diversos fenómenos, el razonamiento abductivo les confiere un apoyo más sólido considerando diversas técnicas de ciencia de datos (Kitcher, 1989).

Finalmente, las demostraciones matemáticas se formulan deductivamente, pero dada la importancia de las inferencias no deductivas en la ciencia de datos, es necesario reconocer como las ciencias naturales utilizan una combinación de deducción, inducción y abducción en su práctica diaria y las ciencias más formales hacen un uso más frecuente de la deducción y las ciencias aplicadas dependen más de la abducción, mientras que otras ciencias asignan diferentes ponderaciones a diferentes modos de inferencia, como es el caso de la abducción en las ciencias sociales, políticas y económicas, así como en las ciencias cognitivas que se basan en la abducción, dada la frecuencia de teorías indeterminadas y empíricamente equivalentes (Harman, 1965; Lipton, 1991; Niiniluoto, 2018).

Productos epistémicos

La tricotomía del aprendizaje automático (que abarca algoritmos de aprendizaje supervisados, no supervisados y de refuerzo) ayuda a delinear el tipo de conocimiento generado por la ciencia de datos y sus técnicas y los modelos de aprendizaje supervisado predicen resultados basándose en asociaciones observadas que automatizan el proceso de razonamiento inductivo a escalas y resoluciones que superan significativamente la capacidad de los humanos (Spirtes et al., 2000; Pearl, 2009; Imbens & Rubin, 2015; Peters et al., 2017). Sin embargo, los grandes conjuntos de datos y algoritmos potentes no son suficientes para superar los desafíos fundamentales inherentes a este modo de inferencia. Un modelo que funciona bien en un entorno puede fracasar en otro si los datos ya no se ajustan a los patrones observados (El error de la extrapolación de modelos sin comprender las especificidades de las variables) y es aquí donde el aprendizaje no supervisado es un conjunto de métodos más heterogéneo, ampliamente unidos por su tendencia a inferir estructuras sin ninguna variable de resultado predefinida, tal como lo resalta Pearl (2000), como en el caso de los codificadores automáticos, algoritmos de agrupamiento y modelos generativos, que son herramientas que pueden arrojar luz sobre las propiedades latentes como las muestras o características que reflejan algunos hechos subyacentes sobre el proceso de generación de datos tal como lo resalta Spirtes et al. (2000).

Por sí solos, estos métodos no necesariamente proporcionan conocimiento causal; sin embargo, algunas de las investigaciones más importantes sobre Inteligencia Artificial de los últimos 20 años se han centrado en el razonamiento causal tal como lo resaltan Spirtes et al. (2000), Pearl (2000), Imbens & Rubin (2015) y Peters et al., (2017), demostrándose cómo los supuestos probabilísticos pueden combinarse con datos observacionales provenientes de procesos experimentales para inferir la estructura causal y los efectos del tratamiento, lo cual conlleva un manejo delicado para no alterar la expresión de dichos datos. Es de agregar que, trabajos recientes en aprendizaje supervisado han demostrado cómo los principios causales pueden mejorar el rendimiento fuera de la distribución, como lo aseveran Arjovsky et al. (2019), mientras que los algoritmos complejos como las redes neuronales y los bosques impulsados por gradientes se utilizan cada vez más para inferir los efectos del tratamiento en una amplia gama de entornos o escenarios (Nie y Wager, 2021). El descubrimiento causal (la tarea de aprender asociaciones causales a partir de datos observacionales) es un problema de aprendizaje no supervisado por excelencia y es un reto esencial en el ejercicio de la analítica de datos con soporte estocástico, siendo un área de investigación activa desde al menos la década de 1990 y lo sigue siendo hoy como lo aseveran Glymour et al. (2019), ya que el aprendizaje por refuerzo ha sido objeto de intensas investigaciones en los últimos años (Bareinboim et al., 2021).

Varios autores han demostrado cómo la información causal puede mejorar el rendimiento de estos algoritmos, lo que a su vez, ayuda a revelar la estructura causal con base en métodos que pueden -en principio- utilizarse para inferir leyes naturales, como es el caso de Schmidt y Lipson (2009) quienes han propuesto lo que parecen ser las leyes de la mecánica clásica obtenidas algorítmicamente y su método implicó analizar los datos de movimiento de varios sistemas dinámicos utilizando algoritmos que no tenía conocimientos físicos previos de mecánica, proporcionando un punto de datos atractivo para aquellos que tienen esperanzas en la posibilidad del descubrimiento autónomo de las leyes naturales y donde los roles de la correlación y la causalidad en la ciencia y de la ciencia autónoma y libre de teoría, se consolidan desde un enfoque epistémico.

Problemas de la caja negra

Las herramientas de la ciencia de datos se han vuelto muy sofisticadas y complejas y esto se debe en parte a que la ciencia de datos siempre ha respondido a motivaciones prácticas, ya que cualquier desarrollo que produzca un resultado más exitoso se adopta en virtud de su utilidad, a menudo sin hacer una pausa para reflexionar sobre cómo integrarlo en nuestros esquemas conceptuales más amplios, generando dudas sobre la opacidad de estas herramientas, siendo estos los llamados problemas de caja negra, a lo cual Burrell (2016) ha propuesto que hay tres formas en que los algoritmos de la ciencia de

datos se vuelven opacos, el primero es su ocultamiento intencional para beneficio comercial o personal, el segundo es la opacidad que surge del hecho de que la alfabetización y el dominio tecnológico son una condición necesaria para comprender algoritmos sofisticados y el tercero es la complejidad inherente que surge de los procedimientos de optimización algorítmica que exceden la capacidad de la cognición humana, aquí, los dos primeros de estos problemas son problemas pragmáticos que ocurren cuando la ciencia de datos está integrada en la sociedad en general, lo cual es corroborado por Tsamados et. al (2021), quienes resaltan que se pueden presentar disimilitudes o casos en los que problemas supuestamente diferentes pueden colapsar en uno solo (problemas de caja negra).

Por otro lado, Burrell (2016) y Tsamados et. al (2021) resaltan que los problemas de caja negra pueden fragmentarse en conceptuales y no conceptuales, siendo los problemas conceptuales aquellos que se refieren a los límites de los conceptos que se emplean al analizar las cajas negras, mientras que los problemas no conceptuales, por el contrario, no tienen que ver con la naturaleza, los límites y coherencias de los conceptos empleados en sí, sino los problemas más amplios que resultan del uso de estos conceptos, como en la epistemología; no obstante son conceptos que ampliaremos a continuación:

Problemas conceptuales

Algunos problemas de caja negra surgen porque los conceptos ordinarios son de alguna manera inadecuados o poco claros cuando se proyectan en contextos de aprendizaje automático, tal como lo resalta Lipton (2018), quien ha reconocido esta imprecisión sobre el uso de la interpretación y observa que la tarea de interpretar parece poco especificada y los artículos brindan motivaciones diversas y a veces no superpuestas para la interpretabilidad y ofrecen innumerables nociones sobre qué atributos hacen que los modelos sean interpretables (Lipton, 2018). De manera similar, Doshi-Velez y Kim (2017) han señalado la falta de acuerdo sobre una definición de "*interpretabilidad*" y (además-) sobre cómo debe evaluarse, identificando dos usos paradigmáticos de "*interpretabilidad*" en la literatura. Interpretabilidad en el contexto de una aplicación e interpretabilidad a través de un proxy cuantitativo, aunque ha habido intento de refinar estos conceptos como el planteado por Doshi-Velez y Kim (2017), quienes participan en este tipo de proyectos, sentando las bases para la posterior definición y evaluación rigurosa de la interpretabilidad, refinando los conceptos de la misma y haciendo distinciones detalladas dentro de ellos, contribuyendo a su estructura. Doshi-Velez y Kim (2017) distinguen entre interpretabilidad local y global para evitar confusiones, lo cual es aplicable a predicciones individuales, y en este último a todo el límite de decisión, como es en el caso de las superficies de respuesta y modelos de regresión. Es de agregar que Watson y Floridi (2020) hacen una distinción similar entre explicaciones locales (simbólicos) y globales (tipo), aunque en un contexto matemático más formal y otros trabajos sobre las representaciones desplegadas en los problemas de caja negra hacen referencia a la relación entre varios términos aproximadamente sinónimos, palabras como "interpretabilidad", "explicabilidad", "comprensibilidad", entre otros.

A su vez, es de interés filosófico si alguno o todos estos términos se superponen total o parcialmente, tal como lo plantea Krishnan (2020), quien los considera insignificantemente diferentes, argumentando que todos estos términos se definen entre sí de una manera que contribuye poco a aclarar conceptos imprecisos. Otros autores adoptan un enfoque más detallado como es el caso de Tsamados et al. (2021) que enfatizan en la diferencia entre explicabilidad e interpretabilidad, donde la explicabilidad se aplica tanto a expertos como a no expertos como es el caso del científico de datos experto que podría necesitar explicar la mecánica de algún algoritmo a su cliente no experto. Por el contrario, este último está restringido a los expertos (Interpretabilidad como interpretabilidad, en principio). Así, en su opinión, la explicabilidad presupone la interpretabilidad, pero no al contrario, lo cual lo convierte en un reto intelectual para el asesor y analista de información (Krishnan, 2020).

Problemas no conceptuales

Para el concepto de caja negra se derivan una serie de problemas no conceptuales y sus soluciones no abordan las deficiencias de las representaciones en sí mismas y giran en torno a cuatro problemas epistemológicos que han recibido menos atención, tal como lo resaltan Ratti y López-Rubio (2018), quienes han argumentado que la interpretabilidad es crucial para destilar explicaciones causales a partir de las correlaciones identificadas por técnicas de analítica de datos, como puede ser el caso en un contexto científico abundante en datos, como es el caso del paradigma de los modelos biológicos mecanicistas, donde se observa que, para que los biólogos conviertan modelos correlativos de datos científicos en modelos causales con poder explicativo, los modelos correlativos deben ser interpretables, surgiendo una compensación general de relación inversa basada en que cuánto más complejo es un modelo, menos explicativo es y dado que el ejercicio predictivo de los modelos científicos de datos están correlacionados positivamente con su complejidad, tienden a concluir que existe un verdadero problema epistemológico de caja negra.

Por otro lado, Watson y Floridi (2020) han interpretado el sobreajuste como un tipo diferente de problema epistemológico de caja negra, donde el sobreajuste ocurre cuando un modelo de aprendizaje automático hace predicciones correctas en el *corpus* de entrenamiento del mismo, pero no logra predecir correctamente en los datos de prueba, como es el caso de los resultados de Lapushkin et al. (2016), en el que las imágenes de un caballo compartían una marca de agua sutil y distintiva y el clasificador de imágenes resultante asoció fuertemente esa marca de agua con la etiqueta "caballo" y, por lo tanto, no pudo clasificar correctamente los caballos en un conjunto de prueba cuando la marca de agua estaba ausente. Es aquí donde Watson y Floridi (2020) proponen que uno se puede formar accidentalmente una creencia verdadera a través de mecanismos poco fiables de generación de conocimiento y resaltan la necesidad de un marco para interpretar las cajas negras y así reducir el sobreajuste.

Sumado a lo anteriormente expuesto, Krishnan (2020) ha destacado el punto epistemológico más amplio de que, en la medida en que los algoritmos de aprendizaje automático puedan tener una dimensión pedagógica (Que podemos aprender de los errores que puedan cometer los algoritmos), estos deben ser interpretables o comprensibles para que podamos aprender algo, a lo cual Lipton (2018), hace una observación similar sobre el carácter informativo de los algoritmos; por lo tanto, una mayor transparencia algorítmica conlleva importantes beneficios de carácter epistémico.

Con base en la discusión anterior, se da la impresión de que estos problemas son sustanciales y vale la pena resolverlos; no obstante, no todos los autores mencionados están de acuerdo debido a la existencia de dos tipos principales de objeciones: Algunos admiten que las cajas negras son opacas, pero niegan que la forma correcta de proceder sea tratar de explicar o interpretar su funcionamiento interno y en cambio, argumentan que las cajas negras deberían ser reemplazadas por cajas no negras igualmente capaces. En contraste, otros autores niegan que las cajas negras sean problemáticas en absoluto como lo plantean Zerelli et al. (2019) al argumentar que la opacidad de las cajas negras no es un problema genuino en absoluto, y de manera similar, Krishnan (2020) ha argumentado que las preocupaciones sobre la interpretabilidad y sus afines están innecesariamente "infladas" porque "*La interpretabilidad y sus afines son nociones poco claras..., Todavía no tenemos una idea de qué conceptos se supone que captura cualquier definición técnica o, de hecho, si existe cualquier concepto de interpretación o interpretabilidad que deba ser capturado técnicamente*" (Krishnan, 2020).

La ciencia en un paradigma intensivo en datos

Habiendo considerado hasta ahora cuestiones fundamentales en la epistemología de la ciencia de datos, ahora se puede ampliar la reflexión para considerar cómo la analítica de datos podría moldear la ciencia y la filosofía de la ciencia en general. Kuhn (1970), propuso que la ciencia atraviesa ciclos de normalidad, crisis y, en última instancia, revolución. Aquí, la fase normal, presenta a practicantes comprometidos en la búsqueda de resolver acertijos utilizando las herramientas del paradigma predominante. Sin embargo, recientemente se ha propuesto que la proliferación de datos ha inaugurado una

nueva era de la ciencia donde se puede generar conocimiento científico y desplegar métodos matemáticos y científicos de datos sin ningún conocimiento o comprensión previa de los fenómenos o sus interrelaciones, tal como lo resaltan Kitchin (2014) y Hey et al. (2009) y para dilucidar la naturaleza de este nuevo paradigma y ubicarlo en la historia de la ciencia Tabla 1.

Tabla 1. - Paradigmas científicos

Paradigma	Naturaleza	forma	Cuándo
Primero	Experimental	Empirismo: Descripción de fenómenos naturales	Renacimiento
Segundo	Teórico	Modelación y generalización	Previa a computadores
Tercero	Simulación Computacional	Fenómenos complejos	Antes del Big Data
Cuarto	Exploratorio	Intensivo en Datos: Exploración estadística	Ahora

Fuente: (Tomado de Kitchen (2014), compilado de Hey et al. 2009)

Aquí, el científico puede permanecer en gran medida desinformado sobre cualquier teoría científica subyacente y la estructura de sus datos, ya que, con las herramientas de la ciencia de datos contemporánea, los datos sin procesar se pueden analizar y explotar así la estructura de forma más o menos automática.

Después de observar que esta parece ser una dirección importante en la práctica científica, Napoletani et al. (2018) plantean la pregunta de por qué las matemáticas y los datos tienen una sinergia tan efectiva, apelando entonces al concepto de "*Efectividad irrazonable*" corroborando lo planteado por Wigner (1960) y por Anderson (2008), quien a su vez ha sostenido que la ciencia basada en la teoría clásica se está volviendo obsoleta debido a la densidad y pluralidad de correlaciones producidas por el análisis de cantidades extraordinariamente grandes de datos que serán más útiles que las generalizaciones causales proporcionadas por la ciencia clásica, argumento que es corroborado por otros autores como Prensky (2009) y Steadman (2013), aunque Kitchin (2014) ofrece una caracterización más formal de esta visión, a la que llama un nuevo tipo de empirismo, donde "*Todos los datos proporcionan visiones oligópticas del mundo*" y siguiendo a Leonelli (2014), quien señala que incluso el análisis y comprensión de estructuras y patrones a partir de datos no puede ocurrir al vacío de toda teoría científica y debido a su arraigo en la sociedad, las teorías y la formación científicas siempre proporcionan el soporte en torno a la recopilación y el análisis de datos.

Finalmente, en la medida en que la ciencia sea acumulativa, sostiene que los resultados individuales de las investigaciones científicas de datos siempre requerirán interpretación y encuadre por parte de científicos que estén equipados con el conocimiento de las teorías científicas y si los datos y los resultados de su análisis se interpretan sin ninguna teoría subyacente corren el riesgo de volverse infructuosos y les resultará difícil contribuir a cualquier comprensión fundamental de la naturaleza de los fenómenos, ya que "carece de integración en un conocimiento más amplio" (Kitchin, 2014). Aquí, la tarea será entonces la integración de prácticas científicas de datos en la metodología científica y para ello, Kitchin (2014) propone una explicación de esta integración, llamándola "*Ciencia impulsada por datos*" y toma la forma de un reequilibrio de los tres modos de inferencia mencionados con antelación y sostiene que la ciencia contemporánea tiene una dimensión experimental en unos casos, deductiva en otros, donde las hipótesis que se formulan surgen a partir de hipótesis más fundamentales y luego se ofrecen para su confirmación o refutación mediante experimentos, lo cual conlleva a plantearnos que la ciencia hoy se fundamenta en un paradigma basado en datos que eleva el estatus de la lógica inductiva en este proceso de formación de hipótesis, con hipótesis experimentales generadas a partir de correlaciones identificadas mediante métodos científicos de datos y donde es probable que el futuro de la ciencia intensiva en datos siga estando basado en la teoría, aunque a veces se utilizarán métodos alternativos y científicos de datos para ayudar en la generación de teorías.

La ciencia y analítica de datos se distingue por su capacidad de integrar enfoques estadísticos y computacionales para abordar problemas específicos, resaltando la importancia del conocimiento cau-

sal y práctico en su aplicación, todo ello dentro de un contexto que combina la evolución histórica con la innovación tecnológica y metodológica.

A lo largo de las últimas seis décadas, se han desarrollado diversas taxonomías descriptivas para caracterizar el análisis estadístico y la ciencia de datos. Tukey (1962) introdujo la primera taxonomía enfocada en los procedimientos y técnicas para analizar datos, así como en la planificación de la recopilación de datos. Wu (1997) amplió esta visión con una taxonomía centrada en la recopilación, modelado y análisis de datos, y la toma de decisiones basada en estos análisis. Más recientemente, Donoho (2017) propuso una taxonomía integral que abarca la recopilación, gestión, procesamiento, análisis, visualización e interpretación de grandes volúmenes de datos heterogéneos. Donoho define la ciencia de datos como el estudio de sistemas de información mediante razonamiento probabilístico y herramientas computacionales, reflejando las dinámicas actuales de la analítica de datos y su construcción de conocimiento.

La ciencia de datos integra diferentes modos de inferencia para generar conocimiento, utilizando herramientas y métodos avanzados para inferir estructuras y relaciones en los datos, lo que permite tanto la predicción como la explicación de fenómenos. Esta integración facilita la creación de conocimiento causal y empírico, impulsando avances significativos en múltiples disciplinas.

Las herramientas de la ciencia de datos se han vuelto muy sofisticadas y complejas y esto se debe en parte a que la ciencia de datos siempre ha respondido a motivaciones prácticas, ya que cualquier desarrollo que produzca un resultado más exitoso se adopta en virtud de su utilidad, a menudo sin hacer una pausa para reflexionar sobre cómo integrarlo en nuestros esquemas conceptuales más amplios, generando dudas sobre la opacidad de estas herramientas, siendo estos los llamados problemas de caja negra, a lo cual Burrell (2016) ha propuesto que hay tres formas en que los algoritmos de la ciencia de datos se vuelven opacos, el primero es su ocultamiento intencional para beneficio comercial o personal, el segundo es la opacidad que surge del hecho de que la alfabetización y el dominio tecnológico son una condición necesaria para comprender algoritmos sofisticados y el tercero es la complejidad inherente que surge de los procedimientos de optimización algorítmica que exceden la capacidad de la cognición humana.

Los problemas de caja negra pueden fragmentarse en conceptuales y no conceptuales, siendo los problemas conceptuales aquellos que se refieren a los límites de los conceptos que se emplean al analizar las cajas negras, mientras que los problemas no conceptuales, por el contrario, no tienen que ver con la naturaleza, los límites y coherencias de los conceptos empleados en sí, sino los problemas más amplios que resultan del uso de estos conceptos, como en la epistemología.

La analítica de datos está transformando la ciencia y la filosofía de la ciencia. La ciencia tradicional sigue un ciclo de normalidad, crisis y revolución, donde los practicantes trabajan dentro de un paradigma predominante. Sin embargo, actualmente se puede generar conocimiento científico y utilizar métodos matemáticos y científicos de datos sin tener conocimiento previo de los fenómenos. Esto ha llevado a un nuevo paradigma en la ciencia, donde los científicos pueden analizar y explotar datos sin procesar de manera automática, sin tener una comprensión profunda de la teoría subyacente o la estructura de los datos.

La sinergia entre las matemáticas y los datos ha sido llamada "efectividad irrazonable". Esto se debe a que el análisis de grandes cantidades de datos produce correlaciones que son más útiles que las generalizaciones causales proporcionadas por la ciencia clásica. Sin embargo, algunos argumentan que el análisis de datos debe estar respaldado por teorías científicas y que la interpretación de los resultados debe ser realizada por científicos equipados con conocimiento teórico. El análisis y comprensión de datos requiere el soporte de teorías y formación científica.

Para integrar la práctica científica de datos en la metodología científica, se propone la "ciencia impulsada por datos". Esto implica un reequilibrio de los modos de inferencia en la ciencia contemporánea, con una combinación de enfoques experimentales, deductivos e inductivos. La formación de hipótesis se basa en correlaciones identificadas mediante métodos científicos de datos.

Conclusiones

Aunque es probable que la ciencia intensiva en datos siga basándose en la teoría, también se utilizarán métodos alternativos y científicos de datos para generar nuevas teorías en el futuro. La integración de estas prácticas es crucial para avanzar en la comprensión de la naturaleza de los fenómenos. En resumen, la analítica de datos está cambiando la forma en que se hace ciencia, permitiendo nuevas formas de generar conocimiento y reequilibrando los enfoques de inferencia.

Reseña de los autores:

Ximena Cifuentes Wchima: Magister en Desarrollo Sostenible y Medio Ambiente. Decana -Facultad de Ingenierías, Universidad La Gran Colombia. Armenia, Colombia. Integrante del Grupo de Investigación Gerencia de la Tierra. Correo electrónico: defingenieria@ugca.edu.co

John Edward Herrera Quintero: Magister en Sistemas Integrados de Gestión de la Calidad. Profesor Asociado - Facultad de Ingenierías, Universidad La Gran Colombia. Armenia, Colombia. Correo electrónico: herreraquijohn@miugca.edu.co

Luis Miguel Mejía Giraldo: Magister en Desarrollo Sostenible y Medio Ambiente. Profesor Asociado - Facultad de Ingenierías, Universidad La Gran Colombia. Armenia, Colombia. Líder del Grupo de Investigación GIDA. Correo electrónico: mejiagluismiguel@miugca.edu.co

Luis Fernando Restrepo Betancur: Especialista en Estadística y Biomatemáticas. Profesor Titular de la Universidad de Antioquia. Medellín, Colombia. Correo electrónico: lfernando.restrepo@udea.edu.co

Bibiana Vélez Medina: Ph.D. en Ciencias de la Educación. Mg. en Educación. Rectora delegataria de la Universidad La Gran Colombia. Armenia, Colombia. Líder del grupo de investigación PAIDEIA. Correo electrónico: rectoraugca@ugca.edu.co

Contribución de los autores:

Los autores han participado en la redacción del trabajo y análisis de los documentos.

Referencias Bibliográficas

Alemany Oliver, M. and Vayre, J.-S. (2015). Big data and the future of knowledge production in marketing research: Ethics, digital traces, and abductive reasoning. Journal of Marketing Analytics, 3(1), pp. 5-13. doi: 10.1057/jma.2015.1. https://link.springer.com/article/10.1057/jma.2015.1

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

Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Pad, D. (2019). Invariant risk minimization. arXiv preprint, 1907.02893. https://www.researchgate.net/publication/334288906_Invariant_Risk_Minimization

Bareinboim, E., Lee, S., & Zhang, J. (2021). An introduction to causal reinforcement learning. Columbia CausalAI Laboratory, Technical Report (R-65). https://ics.uci.edu/~dechter/courses/ics-295cr/2024-25_Q2_Winter/presentations/P1%20-%20Jiapeng%20Zhao%20-%20An%20Introduction%20to%20Causal%20Reinforcement%20Learning.pdf

Blei, D. M. and Smyth, P. (2017). Science and data science. Proceedings of the National Academy of Sciences, 114(33), 8689-8692. doi: 10.1073/pnas.1702076114. https://www.pnas.org/doi/10.1073/pnas.1702076114

Breiman, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). Statistical Science, 16(3), 199-231. doi: 10.1214/ss/1009213726. https://projecteuclid.org/journals/statistical-science/volume-16/issue-3/Statistical-

Modeling--The-Two-Cultures-with-comments-and-a/10.1214/ss/1009213726.full

Burrell, J. (2016). How the machine "thinks: Understanding opacity in machine learning algorithms. Big Data & Society. doi: 10.1177/2053951715622512. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2660674

Carmichael, I. and Marron, J. S. (2018). Data Science vs. Statistics: Two Cultures? Japanese Journal of Statistics and Data Science, 1(1), 117-138. doi: 10.1007/s42081-018-0009-3. https://arxiv.org/abs/1801.00371

Chambers, J. (1993). Classes and Methods in S.I: Recent Developments Computational Statistics, 8:3, 167-184.

Chambers, J. M. (1993). Greater or lesser statistics: a choice for future research. Statistics and Computing, 3(4), 182-184. doi: 10.1007/BF00141776. https://link.springer.com/article/10.1007/BF00141776

Cifuentes et al. (2016). Métodos de análisis para la investigación, desarrollo e innovación (I+D+i) de procesos agrícolas y agroindustriales. En https://www.ugc.edu.co/sede/armenia/files/editorial/metodos_de_analisis_para_la_investigacion.pdf

Donoho, D. (2017). 50 Years of Data Science. doi: https://doi.org/10.1080/10618600.2017.1384734745-766 https://www.tandfonline.com/doi/full/10.1080/10618600.2017.1384734

Doshi-Velez, F. and Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. http://arxiv.org/abs/1702.08608

Frické, M. (2015). Big data and its epistemology. Journal of the Association for Information Science and Technology. 66(4), pp. 651-661. doi: 10.1002/asi.23212. https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.23212

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J. & Krüger, L. (2013). The empire of chance: How probability changed science and everyday life. New York: Cambridge University Press. https://books.google.com.cu/books/about/The_Empire_of_Chance.html?id=Bw2yKfpvts8C&redir_esc=y

Glymour, C., Zhang, K., & Spirtes, P. (2019). Review of Causal Discovery Methods Based on Graphical Models. Frontiers in genetics, 10, 524. https://doi.org/10.3389/fgene.2019.00524

Hacking, I. (1975). The emergence of probability: A philosophical study of early ideas about probability, induction, and statistical inference. New York: Cambridge University Press. https://www.cambridge.org/core/books/emergence-of-probability/9852017A380C63DA30886D25B80336A7

Harman, G. (1965). The inference to the best explanation. Philosophical Review, 74(1), 88-95. https://www.jstor.org/stable/2183532

Harman, G. & Kulkarni, S. (2007). Reliable reasoning: Induction and statistical learning theory. Cambridge, MA: The MIT Press. https://direct.mit.edu/books/monograph/2565/Reliable-ReasoningInduction-and-Statistical

Hey, T., Tansley, S. and Tolle, K. (2009). The Fourth Paradigm: Data-Intensive Scientific Discovery. pp 287. https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. Big Data & Society, 1(1), p. 2053951714528481. doi: 10.1177/2053951714528481. https://www.researchgate.net/publication/271525133_Big_Data_New_Epistemologies_and_Paradigm_Shift

Kitcher, P. (1976). Explanation, Conjunction, and Unification. The Journal of Philosophy, 73(8), 207-212. doi:10.2307/2025559. https://www.jstor.org/stable/2025559

Kitcher, P. (1989). Explanatory unification and the causal structure of the world. In P. Kitcher & W. Salmon (eds.), Scientific Explanation, 410-505. Minneapolis: University of Minnesota Press. https://conservancy.umn.edu/server/api/core/bitstreams/8f6f9fe7-b511-43cd-8d75-5c8570fefd59/content

Krishnan, M. (2020). Against Interpretability: a Critical Examination of the Interpretability Problem in Machine Learning. Philosophy & Technology, 33(3), 487-502. doi: 10.1007/s13347-019-00372-9. https://www.researchgate.net/publication/335148516_Against_Interpretability_a_Critical_

Examination_of_the_Interpretability_Problem_in_Machine_Learning

Kuhn, T. S. (1970). The structure of scientific revolutions. 2nd Edition. Chicago: University of Chicago Press. https://www.lri.fr/~mbl/Stanford/CS477/papers/Kuhn-SSR-2ndEd.pdf

Leonelli, S. (2014). What difference does quantity make? On the epistemology of Big Data in biology. Big Data & Society, 1(1), 2053951714534395. doi: 10.1177/2053951714534395. https://journals.sagepub.com/doi/10.1177/2053951714534395

Lipton, P. (1991). Inference to the best explanation. London: Routledge. https://books.google.es/books?id=WIfYNExpSC0C

Lipton, Z. C. (2018). The mythos of model interpretability. Communications of the ACM, 61(10), 36-43. doi: 10.1145/3233231. https://arxiv.org/abs/1606.03490

MacKenzie, D. (1984). Statistics in Britain, 1865-1930: The social construction of scientific knowledge. Edinburgh: Edinburgh University Press. https://gwern.net/doc/statistics/1981-mackenzie-statisticsinbritain18651930.pdf

Mallows, C. (2006). Tukey's Paper After 40 Years. Technometrics, 48, pp. 319-325. doi: 10.1198/004017006000000219. https://www.researchgate.net/publication/238879758_Tukey%27s_Paper_After_40_Years

Mayo, D. (1996). Error and the growth of experimental knowledge. Chicago: University of Chicago Press. https://errorstatistics.com/wp-content/uploads/2020/10/egek-pdf-red.pdf

Mayo, D. (2018). Statistical inference as severe testing: How to get beyond the statistics wars. New York: Cambridge University Press. https://www.cambridge.org/core/books/statistical-inference-as-severe-testing/

D9DF409EF568090F3F60407FF2B973B2

Napoletani, D., Panza, M. and Struppa, D. (2018). The Agnostic Structure of Data Science Methods. p. 17. https://arxiv.org/abs/2101.12150

Nash, J. (1950). Non-Cooperative Games. PhD thesis, Princeton University.

Nash, J. (1951). Non-Cooperative Games. The Annals of Mathematics,54(2):286-295. https://www.jstor.org/stable/1969529

Nie, X. and Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. Biometrika, 108(2), pp. 299-319. doi:10.1093/biomet/asaa076. https://arxiv.org/abs/1712.04912

Niiniluoto, I. (2018). Truth-seeking by abduction. Cham, Switzerland: Springer. https://link.springer.com/book/10.1007/978-3-319-99157-3

Pearl, J. (2000). Causality: Models, reasoning, and inference. Cambridge, England: Cambridge University Press. https://bayes.cs.ucla.edu/BOOK-2K/neuberg-review.pdf

Pearl, J. (2009) Causality. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511803161. https://www.cambridge.org/core/books/causality/B0046844FAE10CBF274D4ACBDAEB5

F5B

Peters, J., Janzing, D., & Schölkopf, B. (2017). The elements of causal inference: Foundations and learning algorithms. Cambridge, MA: The MIT Press. Pietsch, W. (no date) 'Big Data, The New Science of Complexity. https://books.google.com.cu/books/about/Elements_of_Causal_Inference.html?id=XPpFDwAAQBAJ&redir_esc=y

Prensky, M. (2009). H. Sapiens Digital: From Digital Immigrants and Digital Natives to Digital Wisdom, p. 11. https://eric.ed.gov/?id=ej834284

Ratti, E. and López-Rubio, E. (2018). MECHANISTIC MODELS AND THE EXPLANATORY LIMITS OF MACHINE LEARNING. Machine Learning, p. 18. https://philsci-archive.pitt.edu/14452/1/manuscript%20philsci%20-%20Ratti%20%26%20Lopez-Rubio.pdf

Schmidt, M. and Lipson, H. (2009). Distilling Free-Form Natural Laws from Experimental Data. Science, 324(5923), 81-85. doi: 10.1126/science.1165893. https://www.science.org/doi/10.1126/science.1165893

Spirtes, P., Glymour, C., & Scheines, R. (2000). Causation, prediction, and search. Cambridge, MA: The MIT Press. https://direct.mit.edu/books/monograph/2057/Causation-Prediction-and-Search

Steadman, I. (2013). Big data and the death of the theorist. Wired UK, 25 January. https://www.wired.co.uk/article/big-data-end-of-theory

Tukey, J. W. (1962). The Future of Data Analysis. Ann. Math. Statist. 33(1): 1-67 (March, 1962). DOI: 10.1214/aoms/1177704711. https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-33/issue-1/The-Future-of-Data-Analysis/10.1214/aoms/1177704711.full

Van Fraassen, B. C. (1980) The Scientific Image. Oxford University Press. https://epistemh.pbworks.com/f/2.+Oxford.University.Press.USA.The.Scientific.Image.Okt.1980.pdf

Vélez, B. (2018). Fines y estrategias de un modelo de universidad socialmente responsable. Sophia-Educación, 4 (2). https://dialnet.unirioja.es/servlet/articulo?codigo=6996273

Wigner, E.P. (1960). The unreasonable effectiveness of mathematics in the natural sciences. Richard Courant lecture in mathematical sciences delivered at New York University, May 11, 1959, Communications on Pure and Applied Mathematics, 13(1), 1-14. doi:10.1002/cpa.3160130102. https://www.researchgate.net/publication/227990770_The_unreasonable_effectiveness_of_mathematics_in_the_natural_sciences_Richard_Courant_lecture_in_mathematical_sciences_delivered_at_New_York_University_May_11_1959

Wu, C. F. J. (1997). Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence', Philosophy and Technology, 1-24. doi: 10.1007/s13347-019-00382-7. https://arxiv.org/pdf/1903.04361/1000