

Fundamentos Epistemológicos y Estructurales de la Ciencia de Datos

I. Introducción General

- La Ciencia de Datos ha emergido como una disciplina multifacética e interdisciplinaria, impulsada por la necesidad de transformar grandes volúmenes de datos en conocimiento útil y accionable. A través del tiempo, desde John Graunt en 1592 hasta el auge del big data, la humanidad ha buscado convertir datos en información. Hoy, con tecnologías modernas y abundancia de datos, se requiere una combinación de estadística, computación, pensamiento algorítmico y comprensión ética para dar sentido a los datos.

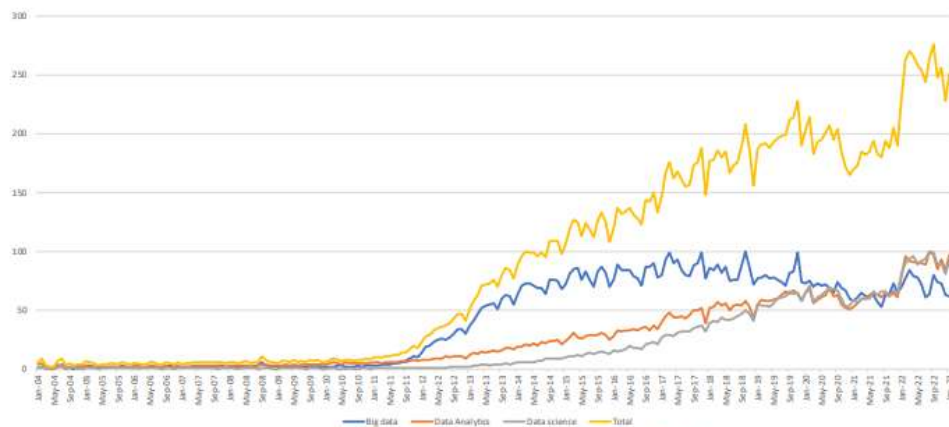


Figure 1: Trending of data science-relevant terms
Source: Google Trends on January 2023

II. Definiciones y fundamentos disciplinarios

2.1 Definiciones diversas

Las definiciones de Ciencia de Datos varían considerablemente según el enfoque disciplinario, el contexto institucional y el objetivo del análisis.

- Para Desai et al. (2022), existe una tensión entre enfoques **minimalistas**, que conciben la ciencia de datos como una mera extensión de la estadística orientada al descubrimiento de patrones (por ejemplo, aprendizaje automático sin interpretación), y enfoques **maximalistas**, que consideran el campo como un ecosistema complejo que abarca infraestructura tecnológica, modelado, visualización, análisis de decisiones, comunicación de resultados y consideraciones éticas. Estas posiciones representan visiones distintas sobre qué constituye conocimiento válido y útil en el siglo XXI.
- Özsü (2024) ofrece una definición más estructurada y funcional: la Ciencia de Datos es la disciplina que estudia, desarrolla y aplica principios para gestionar y analizar el ciclo de vida completo de los datos, desde su generación hasta su utilización para la toma de decisiones. Esta definición destaca su carácter transversal, integrador y computacional, y proporciona un marco metodológico claro para su institucionalización académica.
- Jugal Kalita (2024) define la Ciencia de Datos como el estudio de métodos computacionales y estadísticos necesarios para recolectar, transformar, analizar e interpretar datos provenientes de diversas fuentes, con el propósito de resolver problemas

del mundo real. Subraya además la naturaleza aplicada del campo, su proximidad con la ingeniería de software, y su papel central en la automatización del descubrimiento de conocimiento.

En conjunto, estas perspectivas muestran que la Ciencia de Datos no tiene aún una definición única ni establecida, sino que representa un campo en consolidación que oscila entre lo técnico, lo científico y lo filosófico, entre lo pragmático y lo normativo, y entre lo disciplinar y lo interdisciplinario., es el estudio de métodos computacionales para procesar, analizar y sintetizar datos de múltiples fuentes con el fin de resolver problemas del mundo real.

2.2 Ciencia de datos y otras disciplinas

La Ciencia de Datos es inherentemente interdisciplinaria, pero también mantiene tensiones y distinciones fundamentales con respecto a disciplinas establecidas como la estadística, la informática, la inteligencia artificial (IA), la minería de datos y la ingeniería de software.

- **Con la estadística**, comparte el interés por el análisis de datos y la inferencia, pero se diferencia por su énfasis en la escalabilidad computacional, el uso de datos masivos (big data), y su orientación práctica y aplicada. La estadística tradicional se centra en modelos matemáticos formales, mientras que la ciencia de datos privilegia la predicción útil y operativa incluso con modelos complejos y no interpretables (black box models).
- **Con la inteligencia artificial**, comparte técnicas (como redes neuronales o árboles de decisión) y objetivos como la automatización y la toma de decisiones inteligentes. Sin embargo, la IA tiende a centrarse en la emulación de procesos cognitivos humanos (percepción, razonamiento, lenguaje), mientras que la ciencia de datos se enfoca más en el análisis de datos del mundo real para generar conocimiento práctico o apoyo a decisiones.
- **Con la minería de datos**, hay superposición metodológica, especialmente en algoritmos de descubrimiento de patrones, clustering, y reducción de dimensionalidad. Pero la ciencia de datos abarca un ciclo más amplio que incluye la recolección, limpieza, integración, visualización, interpretación, comunicación y gobernanza de los datos.
- **Con la informática**, coincide en herramientas, lenguajes de programación, algoritmos y estructuras de datos. Sin embargo, la ciencia de datos es más empírica, dirigida por los datos y orientada a responder preguntas aplicadas, mientras que la informática tiende a enfocarse en la abstracción, la eficiencia algorítmica y la arquitectura de sistemas.
- **Con la ingeniería de software**, hay coincidencia en la construcción de sistemas robustos y reproducibles. Pero mientras la ingeniería se enfoca en la funcionalidad, escalabilidad y mantenimiento del software, la ciencia de datos está guiada por el ciclo de descubrimiento de conocimiento, análisis crítico de resultados, y la generación de inferencias o predicciones útiles.

Raíces históricas

La Ciencia de Datos no surge de manera espontánea en el siglo XXI. Tiene antecedentes históricos importantes:

- **John Tukey (1962)**: promovió el análisis exploratorio de datos como un enfoque independiente de la inferencia estadística tradicional.
- **Peter Naur (1974)**: acuñó el término “Data Science” en su libro sobre metodología del procesamiento de datos.
- **Décadas de 1990-2000**: auge del KDD (Knowledge Discovery in Databases) y la minería de datos, que sentaron las bases para el enfoque moderno centrado en grandes volúmenes de datos heterogéneos.

Así, la ciencia de datos ha evolucionado como una síntesis crítica y técnica de múltiples disciplinas, con identidad propia, aunque aún en proceso de consolidación epistemológica.

- Se diferencia de la estadística en su escala computacional y orientación práctica.
- Integra y se apoya en IA, minería de datos, visualización, ética, y gestión de bases de datos.
- Tiene raíces históricas: desde Tukey (1962), Peter Naur (1974) y el surgimiento de la minería de datos y el KDD.

III. Dimensiones epistemológicas

(según Desai et al.)

Las dimensiones epistemológicas de la Ciencia de Datos, tal como las analizan Desai et al., abordan las bases del conocimiento que se produce, los métodos que se emplean y la legitimidad científica de sus prácticas. Estas dimensiones no solo son fundamentales para la consolidación académica del campo, sino también para su uso responsable en la sociedad. A continuación, se detallan cinco ejes centrales de debate epistemológico:

1. **Constitución disciplinaria:** ¿Es la Ciencia de Datos una ciencia formal, una rama de la ingeniería, una práctica empírica o una mezcla de todas? Este eje problematiza la falta de consenso sobre si estamos ante una disciplina emergente o ante una superposición técnica. Algunos argumentan que, al no contar con teorías propias ni un objeto formal definido, la Ciencia de Datos sería más bien un campo técnico-aplicado. Otros señalan que su capacidad de generar conocimiento y su marco metodológico la hacen merecedora de estatus científico.
 2. **Naturaleza del conocimiento:** Se debate si la Ciencia de Datos tiene como objetivo principal la **exploración** (descubrimiento de correlaciones), la **explicación** (identificación de causas) o la **predicción** (anticipación de eventos futuros). La práctica actual parece priorizar el valor predictivo, lo que plantea dilemas sobre la comprensión profunda de los fenómenos modelados. Esta orientación hacia la utilidad, en lugar de la explicación, contrasta con la epistemología tradicional basada en causalidad.
 3. **Modelos opacos y cajas negras:** Muchos de los modelos utilizados (por ejemplo, redes neuronales profundas) tienen alto poder predictivo pero escasa interpretabilidad. Esto crea una tensión entre eficiencia técnica y comprensión científica. La pregunta clave es: ¿podemos considerar legítimos resultados que no podemos explicar en términos comprensibles para humanos? Este punto ha motivado el auge de subcampos como la inteligencia artificial explicable (XAI) y el análisis de interpretabilidad.
 4. **Relación con la filosofía de la ciencia:** La Ciencia de Datos entra en conflicto con algunos postulados del empirismo clásico. La centralidad de los datos, el aprendizaje automatizado y el uso de inferencia estadística a gran escala han llevado a proponer marcos **post-teóricos**, donde las teorías no preceden al análisis sino que emergen de los datos. Esta postura ha sido criticada por ignorar el papel de la construcción teórica y la validación externa del conocimiento.
 5. **Legitimidad científica:** Se plantea si la Ciencia de Datos puede ser considerada ciencia cuando muchas de sus conclusiones no son verificables causalmente ni refutables en términos tradicionales. La confiabilidad de los resultados, la reproducibilidad de los procesos y la transparencia metodológica se presentan como elementos clave para sostener su legitimidad epistemológica. De este modo, la Ciencia de Datos debe desarrollarse no solo como un conjunto de herramientas técnicas, sino como un campo comprometido con estándares científicos rigurosos. ¿puede confiarse en inferencias sin comprensión causal?
-

IV. Fundamentos estructurales

4.1 Ciclo de vida de los datos

El ciclo de vida de los datos representa el núcleo operativo de la ciencia de datos, abarcando desde la generación o captura de datos hasta su transformación en conocimiento accionable. Según Özsu, Kalita y los textos introductorios, este ciclo puede dividirse en seis fases integradas:

1. **Recolección e integración:** Esta etapa implica la adquisición de datos desde múltiples fuentes (sensores, formularios, redes sociales, registros administrativos, bases científicas), asegurando calidad, consistencia y formatos compatibles. La integración puede requerir la combinación de datos estructurados y no estructurados.
2. **Limpieza y preparación (wrangling):** Los datos reales suelen ser incompletos, redundantes o erróneos. En esta fase se realiza imputación de valores faltantes, corrección de errores, normalización, transformación de variables y eliminación de duplicados. Esta es una de las tareas más intensivas y cruciales en todo proyecto de ciencia de datos.
3. **Almacenamiento y acceso:** Los datos deben almacenarse en sistemas que garanticen seguridad, disponibilidad y eficiencia de consulta. Esto incluye desde bases de datos relacionales (SQL) hasta arquitecturas distribuidas (Hadoop, NoSQL, cloud storage). La forma en que se almacena influye en la posibilidad de análisis posterior.
4. **Procesamiento y modelado:** Aquí se aplican técnicas de análisis exploratorio y modelado estadístico y computacional. Incluye algoritmos de clasificación, regresión, clustering, reducción de dimensionalidad, y técnicas avanzadas de machine learning y deep learning. Esta fase transforma datos en modelos que representan relaciones y patrones.
5. **Evaluación, interpretación y visualización:** Una vez entrenados los modelos, se evalúa su desempeño utilizando métricas como precisión, recall, F1-score o AUC. También se analiza la estabilidad, el sobreajuste y la generalización. La visualización permite comunicar hallazgos de forma clara e intuitiva.
6. **Presentación, toma de decisiones y acción:** Finalmente, los resultados del análisis deben ser traducidos en narrativas comprensibles para los tomadores de decisiones. Se construyen dashboards, informes y sistemas interactivos para respaldar decisiones estratégicas, operativas o científicas.

Este ciclo no es necesariamente lineal; las fases pueden ser iterativas y superpuestas, especialmente en entornos ágiles o de análisis continuo.

4.2 Componentes clave de la ciencia de datos

La Ciencia de Datos no es solo un conjunto de técnicas; es también una práctica basada en principios que garantizan calidad, robustez y utilidad del conocimiento generado. Algunos de sus componentes clave son:

- **Wrangling:** Conocido también como data munging, implica la limpieza, transformación, normalización, combinación y filtrado de datos crudos para hacerlos adecuados para el análisis. Esta fase requiere habilidades estadísticas y de programación, así como conocimiento del dominio de aplicación.
- **Visualización:** La representación gráfica de datos es tanto una herramienta exploratoria como una forma de comunicación. Técnicas como histogramas, diagramas de dispersión, mapas de calor o gráficos interactivos ayudan a identificar patrones, anomalías o tendencias ocultas.
- **Modelado:** Consiste en construir representaciones matemáticas o algorítmicas de los datos que permiten describir, predecir o explicar fenómenos. Puede ser supervisado (clasificación, regresión), no supervisado (clustering, reducción de dimensionalidad) o

basado en modelos causales y probabilísticos. Incluye técnicas desde modelos lineales hasta redes neuronales profundas.

- **Evaluación:** La validación rigurosa de los modelos es esencial. Esto incluye el uso de métricas cuantitativas (exactitud, recall, precisión, F1-score), validación cruzada, y tests de generalización. También se considera la interpretabilidad del modelo y su utilidad para el usuario final.
- **Ética y gobernanza:** La ciencia de datos no es neutral. Es crucial incorporar principios de equidad, justicia algorítmica, privacidad y seguridad. Se requiere el diseño de políticas de gobernanza de datos y la evaluación de sesgos potenciales en los datos y modelos.
- **Reproducibilidad:** Una práctica científica sólida exige que los resultados puedan ser replicados por otros. Esto se logra mediante flujos de trabajo bien documentados, uso de entornos controlados (Docker, notebooks), control de versiones (Git) y almacenamiento de conjuntos de datos y parámetros usados.

Estos componentes interactúan a lo largo del ciclo de vida y son fundamentales para garantizar la credibilidad, eficiencia y responsabilidad de los proyectos de ciencia de datos (flujo de trabajo, documentación, control de versiones.)

V. Tareas y objetivos de la Ciencia de Datos

5.1 Clasificación de tareas

Según Kalita y otros autores, las tareas de la Ciencia de Datos pueden clasificarse en cuatro grandes categorías funcionales, que corresponden a distintos niveles de análisis y profundidad en el uso de los datos:

- **Descriptiva:** Responde a la pregunta "¿Qué está ocurriendo?". Implica el análisis exploratorio de datos, la identificación de patrones generales, frecuencias, distribuciones y correlaciones simples. Se basa principalmente en estadísticas resumidas, visualizaciones y dashboards. Es útil para comprender el estado actual de una situación o sistema.
- **Diagnóstica:** Responde a la pregunta "¿Por qué ocurre?". Implica identificar relaciones causales, interacciones entre variables o eventos, y análisis multivariado. Se utilizan herramientas como análisis de regresión múltiple, técnicas causales y comparación de grupos. Esta tarea permite generar hipótesis sobre mecanismos subyacentes y explicar anomalías.
- **Predictiva:** Responde a la pregunta "¿Qué va a pasar?". Utiliza modelos estadísticos y algoritmos de machine learning para anticipar resultados futuros basándose en datos históricos. Incluye modelos supervisados como regresión, clasificación, árboles de decisión, redes neuronales y modelos de series temporales. Es central en aplicaciones como scoring crediticio, mantenimiento predictivo y pronóstico de ventas.
- **Prescriptiva:** Responde a la pregunta "¿Qué debo hacer?". Integra predicciones con modelos de optimización y simulación para recomendar acciones o políticas. Puede involucrar inteligencia artificial, investigación operativa y técnicas de decisión multicriterio. Es común en logística, planificación estratégica, medicina personalizada y sistemas de recomendación.

Estas categorías no son mutuamente excluyentes; en muchos proyectos de ciencia de datos se combinan varias de ellas en un mismo flujo de trabajo analítico.

5.2 Objetivos fundamentales

Los objetivos de la Ciencia de Datos reflejan tanto las capacidades técnicas del análisis como su propósito práctico y social. Entre los más destacados se encuentran:

- **Descubrimiento de patrones:** Identificación de regularidades, tendencias y relaciones significativas dentro de conjuntos de datos grandes y complejos. Esto puede realizarse de forma automática (minería de datos) o mediante análisis exploratorio guiado por hipótesis. Es fundamental en marketing, biología, astronomía, y ciencias sociales.
- **Predicción de resultados:** Anticipación de valores futuros o categorización de eventos, basada en el entrenamiento de modelos con datos históricos. Ejemplos comunes son la predicción de abandono estudiantil, enfermedades, fallas técnicas o fluctuaciones de mercado.
- **Agrupación de observaciones (clustering):** Técnica no supervisada que permite segmentar datos según su similitud. Se aplica en estudios de mercado, medicina (fenotipos), redes sociales (comunidades) y sistemas de recomendación. Permite encontrar estructuras latentes y reducir dimensionalidad.
- **Generación de conocimiento accionable:** Más allá del análisis descriptivo, la ciencia de datos busca generar información que pueda integrarse en procesos de toma de decisiones. Esto implica traducir hallazgos complejos en estrategias concretas, recomendaciones políticas, mejoras operativas o innovaciones tecnológicas.

En conjunto, estas tareas y objetivos posicionan a la Ciencia de Datos como un puente entre el análisis riguroso y la resolución efectiva de problemas reales.

- Predicción de resultados
- Agrupación de observaciones (clustering)
- Generación de conocimiento accionable

VI. Ciencia de Datos como ciencia aplicada y multidisciplinar

La Ciencia de Datos se presenta hoy como una de las áreas más relevantes de aplicación científica, caracterizada por su interdisciplinariedad y su capacidad de integrar marcos teóricos y metodológicos de muy diversas disciplinas para resolver problemas reales y complejos. Su enfoque aplicado implica tanto la comprensión profunda de los datos como su vinculación con necesidades sociales, industriales, científicas o tecnológicas.

- **Interdisciplinariedad estructural:** La Ciencia de Datos no se limita a una sola tradición académica. Integra conocimientos y métodos provenientes de la **estadística** (inferencias, pruebas, modelos), la **matemática aplicada** (optimización, álgebra lineal, teoría de la información), la **informática** (algoritmos, programación, estructuras de datos), la **ingeniería** (procesamiento de señales, control de sistemas), las **ciencias sociales** (análisis de comportamiento, estudios poblacionales) y la **filosofía de la ciencia** (teoría del conocimiento, ética, inferencia científica).
- **Herramientas y enfoques metodológicos:** La práctica de la Ciencia de Datos requiere competencias en **visualización de datos**, **aprendizaje automático**, **análisis de redes**, **simulación computacional**, **modelos bayesianos**, **procesamiento del lenguaje natural**, **sistemas interactivos** y técnicas de **optimización**. Estas herramientas permiten no solo analizar datos, sino también representar, comunicar y simular fenómenos de forma accesible y útil.
- **Conocimiento de dominio:** Una característica esencial es la necesidad de involucrar **expertos del dominio** para interpretar los datos y validar los hallazgos. Esto se traduce en colaboraciones estrechas con áreas como la **biología** (genómica, biodiversidad), la **salud pública** (epidemiología, diagnóstico clínico), la **astronomía** (detección de

exoplanetas, clasificación estelar), la **educación** (análisis de trayectorias estudiantiles), el **comercio y marketing** (análisis de clientes y segmentación), entre muchas otras.

Esta naturaleza aplicada y multidisciplinar convierte a la Ciencia de Datos en una herramienta transformadora, capaz de articular soluciones complejas en un mundo orientado por datos, donde los desafíos son globales y requieren enfoques colaborativos e innovadores.

- Utiliza visualización, aprendizaje automático, simulación y sistemas interactivos.
- Involucra conocimiento experto del dominio (biología, salud, astronomía, educación, negocios).

VII. Aplicaciones

La Ciencia de Datos ha demostrado ser una disciplina clave para resolver problemas complejos en una amplia gama de dominios. Su impacto se manifiesta no solo en la mejora de procesos técnicos, sino también en la toma de decisiones estratégicas, la innovación científica y el diseño de políticas públicas basadas en evidencia. A continuación se detallan algunas de las aplicaciones más relevantes:

- **Salud y biomedicina:** Se aplica al análisis de imágenes médicas (radiografías, resonancias), predicción de enfermedades crónicas (diabetes, cáncer), medicina de precisión basada en genómica, monitoreo de epidemias, desarrollo de fármacos mediante modelos predictivos y personalización de tratamientos. La integración de datos clínicos, genéticos y de comportamiento permite una atención médica más eficaz, preventiva y centrada en el paciente.
- **Negocios y marketing:** Se utiliza para el análisis del comportamiento del consumidor, segmentación de clientes, optimización de precios dinámicos, diseño de estrategias de fidelización, análisis de sentimiento en redes sociales, recomendadores de productos y evaluación del desempeño de campañas publicitarias. Ayuda a transformar grandes volúmenes de datos de ventas y transacciones en conocimiento de valor estratégico.
- **Finanzas y seguros:** Se emplea para modelar riesgos crediticios, detectar fraudes financieros, optimizar portafolios de inversión, predecir fluctuaciones de mercado y establecer precios dinámicos en seguros. Los modelos predictivos permiten reducir pérdidas, incrementar márgenes y mejorar la gestión de activos bajo condiciones de incertidumbre.
- **Internet de las cosas (IoT) y movilidad:** A través de sensores y dispositivos conectados, se generan grandes volúmenes de datos que pueden analizarse en tiempo real para optimizar el tráfico urbano, prever accidentes, mejorar el rendimiento energético en viviendas inteligentes, y permitir el funcionamiento de vehículos autónomos. La Ciencia de Datos permite gestionar esta información compleja para lograr ciudades más eficientes y sostenibles.
- **Educación:** Se aplica para la predicción de abandono escolar, análisis de rendimiento académico, diseño de programas de estudio personalizados, detección de necesidades educativas especiales y mejora de la gestión institucional. También contribuye a la evaluación de políticas públicas y a la generación de métricas más precisas de calidad educativa.

Estas aplicaciones reflejan el potencial transformador de la Ciencia de Datos cuando se combina conocimiento técnico con sensibilidad contextual. Su impacto positivo depende de una implementación ética, una interpretación crítica y una colaboración efectiva entre expertos de datos y del dominio.:** diagnóstico por imágenes, predicción de enfermedades, medicina de precisión.

- **Negocios:** análisis de comportamiento del consumidor, optimización de precios, estrategias de marketing.
- **Finanzas:** análisis de riesgo, detección de fraude, recomendaciones de inversión.
- **IoT y movilidad:** vehículos autónomos, optimización de rutas, predicción de tráfico.
- **Educación:** predicción de abandono, diseño curricular adaptativo.

VIII. Retos y advertencias

A medida que la Ciencia de Datos se expande y consolida, enfrenta importantes retos epistemológicos, técnicos, sociales y éticos. Estos desafíos deben ser comprendidos y abordados para garantizar su práctica responsable y sostenida. A continuación se detallan los principales riesgos y advertencias:

- **Hiperautomatización:** Existe un creciente riesgo de delegar decisiones críticas a sistemas automatizados sin comprender plenamente su funcionamiento. Cuando los modelos son opacos o las variables relevantes no son claramente entendidas, la automatización puede conducir a decisiones injustas, erróneas o peligrosas, especialmente en áreas como justicia penal, salud, crédito o política pública.
- **Fetichismo técnico:** Se refiere a la tendencia a valorar los resultados de un modelo simplemente por su sofisticación matemática o complejidad algorítmica, sin un análisis crítico de los datos subyacentes, su contexto o sus implicaciones. Esto puede conducir a un enfoque superficial del análisis, centrado en herramientas en lugar de en preguntas relevantes.
- **Ética algorítmica:** Incluye desafíos como el sesgo en los datos (y su reproducción en los modelos), la falta de transparencia en algoritmos complejos (problema de la caja negra), la vigilancia masiva a través de big data, y la manipulación de preferencias (como ocurre con los algoritmos de redes sociales o publicidad segmentada). Se requiere una ética de la ciencia de datos basada en principios de justicia, equidad, responsabilidad y consentimiento informado.
- **Fragmentación disciplinaria:** A pesar de su carácter interdisciplinario, la Ciencia de Datos aún carece de una base epistemológica consensuada y de una comunidad unificada. Esta fragmentación dificulta su institucionalización académica, la elaboración de currículos coherentes y la evaluación de su impacto científico. La falta de una teoría unificadora limita también su consolidación como disciplina científica madura.
- **Confusión de términos:** En el discurso público y académico es común confundir Ciencia de Datos con Inteligencia Artificial o con Aprendizaje Automático. Aunque estos campos se solapan, no son equivalentes: la IA busca replicar funciones cognitivas humanas, el Machine Learning desarrolla modelos predictivos y la Ciencia de Datos es un marco más amplio que incluye el ciclo de vida de los datos, interpretación, gobernanza y ética. Esta confusión puede generar expectativas erróneas y decisiones equivocadas de política científica y educativa.

En conjunto, estos retos no niegan el valor de la Ciencia de Datos, pero exigen una práctica más crítica, reflexiva y orientada al bien común.:** riesgo de tomar decisiones sin comprensión del modelo.

- **Fetichismo técnico:** exceso de confianza en herramientas sin comprender los datos.
- **Ética algorítmica:** problemas de sesgo, transparencia, vigilancia y manipulación.
- **Fragmentación disciplinaria:** falta de una base epistemológica consensuada.
- **Confusión de términos:** machine learning \neq ciencia de datos \neq inteligencia artificial.

IX. Conclusiones integradoras

La Ciencia de Datos se consolida como una intersección estratégica entre el pensamiento computacional, el análisis científico y la reflexión ética. Su papel es fundamental en un mundo donde los datos han dejado de ser un subproducto de la actividad humana para convertirse en un insumo primario del conocimiento y la acción.

El análisis realizado a partir de autores como Desai, Özsu y Kalita permite reconocer que, si bien la Ciencia de Datos aún se encuentra en construcción teórica, ya exhibe elementos característicos de una disciplina científica madura: posee una problemática propia, un enfoque metodológico diferenciado, herramientas técnicas potentes, y un marco ético emergente.

A nivel educativo, debe ser abordada como una ciencia reflexiva que combina habilidades estadísticas, pensamiento algorítmico, comprensión de dominio y sensibilidad ética. No se trata simplemente de entrenar modelos predictivos, sino de formar profesionales capaces de interpretar críticamente los datos, comprender su impacto y tomar decisiones informadas y responsables.

Desde el punto de vista epistemológico, la Ciencia de Datos desafía muchas de las fronteras clásicas de la ciencia: promueve la integración en lugar de la segmentación disciplinaria, enfatiza la utilidad sin renunciar al rigor, y redefine la relación entre teoría, datos y computación.

Por ello, su consolidación exige no solo infraestructura técnica y volumen de datos, sino también claridad conceptual, solidez metodológica y compromiso social. La Ciencia de Datos del futuro será, necesariamente, una ciencia crítica, abierta y orientada al bien común. una intersección poderosa entre la técnica, la ciencia y la ética. Su desarrollo requiere tanto infraestructura computacional como claridad conceptual. Debe ser enseñada y practicada como una ciencia reflexiva, con visión interdisciplinaria y responsabilidad social. Las definiciones de Desai, Özsu, Kalita y otros revelan una disciplina en evolución, que busca consolidarse no solo como una práctica eficaz, sino como una ciencia madura, crítica y transformadora.

Referencias completas:

- Desai, J., Watson, D., Wang, V., Taddeo, M., & Floridi, L. (2022). *The epistemological foundations of data science: a critical review*. Synthese, 200(469). <https://doi.org/10.1007/s11229-022-03933-2>
- Özsu, M. T. (2024). *Foundations and Scoping of Data Science*. Communications of the ACM (arXiv:2301.13761)
- Baumer, B., Kaplan, D., & Horton, N. J. (2021). *Introduction to Data Science*. CRC Press.
- Kalita, J. (2024). *Fundamentals of Data Science*. Elsevier Inc.