

Maestría en Minería de Datos

UTN - Universidad Tecnológica Nacional, Rosario

Minería de Datos

Abril 2025

Prof. Rodrigo Kataishi, Ph.D.

rkataishi@untdf.edu.ar



Propósito del Curso

Este curso aborda la **minería de datos** como un enfoque analítico con identidad propia y fundamentos sólidos.

El foco está puesto en su capacidad para generar **conocimiento útil, comprensible y no trivial** a partir de datos.

Se explorará cómo descubrir patrones no obvios, que no están a simple vista, pero que poseen sentido intrínseco y aplicabilidad práctica.

“El objetivo es desarrollar la capacidad de identificar señales significativas en los datos, susceptibles de ser utilizadas eficazmente para interpretar fenómenos, anticipar escenarios y fundamentar decisiones.”

Relevancia de la Minería de Datos

La minería de datos se distingue por varias razones clave:

- **Descubre Estructuras Latentes:** Permite encontrar relaciones y patrones ocultos no derivados de la observación directa ni de hipótesis preestablecidas, yendo más allá de lo evidente.
- **Integra Múltiples Fases:** Combina la exploración empírica de los datos, la aplicación de técnicas de automatización y modelado, y una necesaria interpretación contextual de los hallazgos.
- **Fomenta el Criterio Analítico:** Forma profesionales capaces de **comprender profundamente** los métodos aplicados, las razones de su elección y las implicaciones de los resultados.

“El valor no reside solamente en la aplicación mecánica de algoritmos, sino en descubrir conocimiento relevante y accionable a partir de datos reales, situados en un contexto específico y con un propósito práctico definido.”



```
def __init__(self, path=None, debug=False, logduplicates=True):
    self.file = None
    self.fingerprints = set()
    self.logduplicates = True
    self.debug = debug
    self.logger = logging.getLogger(__name__)
    if path:
        self.file = open(os.path.join(path, 'fingerprint.log'), 'a')
        self.file.seek(0)
        self.fingerprints.update(fp for fp in self.file)

    @classmethod
    def from_settings(cls, settings):
        debug = settings.getbool('DEBUG')
        return cls(job_dir(settings), debug=debug)

    def request_seen(self, request):
        fp = self.request_fingerprint(request)
        if fp in self.fingerprints:
            return True
        self.fingerprints.add(fp)
        if self.file:
            self.file.write(fp + os.linesep)

    def request_fingerprint(self, request):
        return request_fingerprint(...)
```

Objetivos de Aprendizaje Específicos

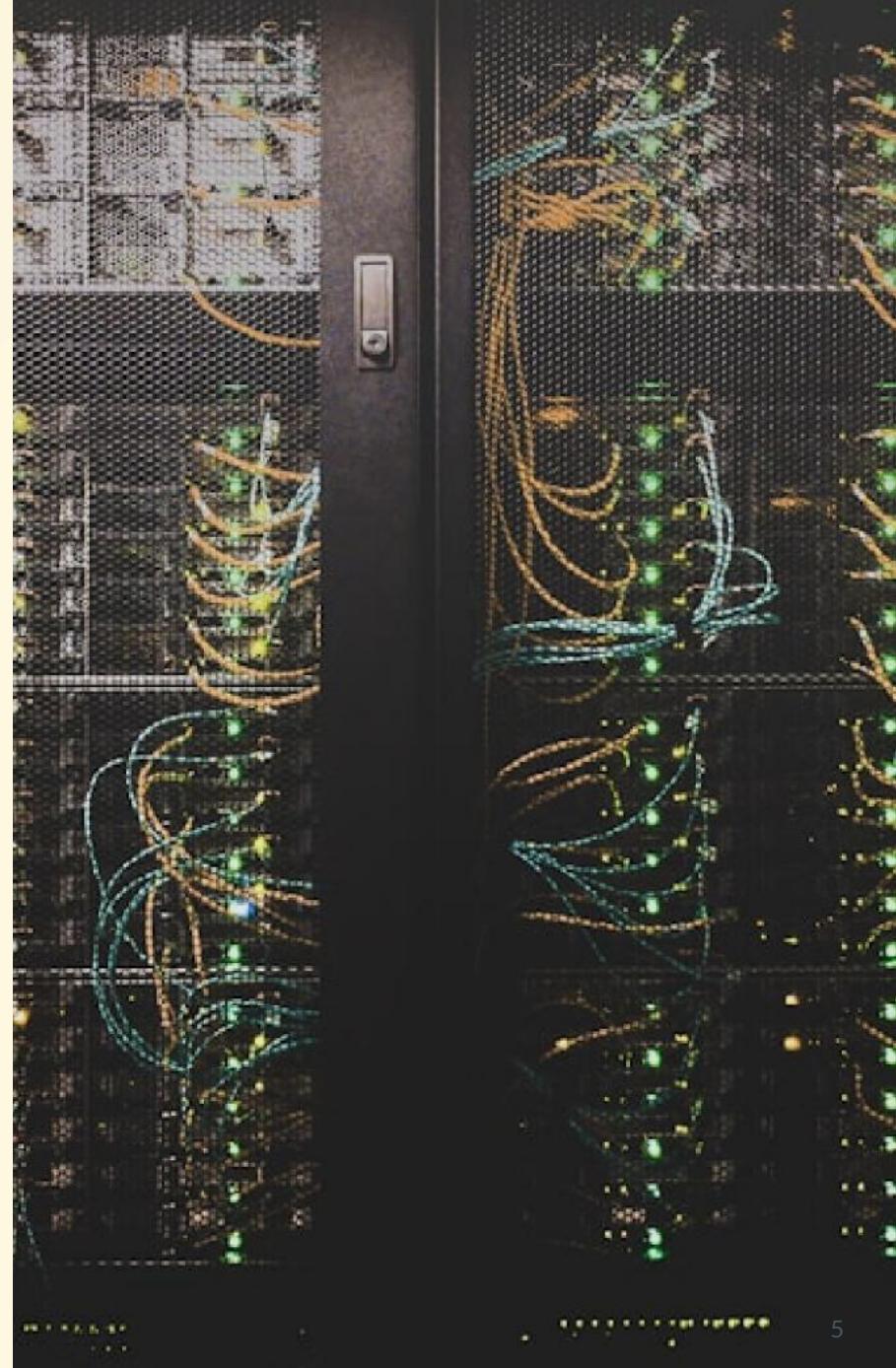
El curso está diseñado para desarrollar competencias integrales en minería de datos:

- **Comprendión Conceptual:** Entender qué es la minería de datos, sus principios fundamentales y su centralidad en el análisis de datos actual, diferenciándola de enfoques puramente predictivos o confirmatorios.
- **Dominio Técnico:** Adquirir destreza en la aplicación de técnicas esenciales como clustering, árboles de decisión, reglas de asociación, y detección de patrones secuenciales o atípicos.
- **Habilidades Prácticas:** Desarrollar la capacidad de trabajar con datos reales utilizando Python, integrando el ciclo completo de análisis: exploración, limpieza, preprocesamiento, modelado e interpretación.
- **Criterio Metodológico:** Fomentar la capacidad de tomar decisiones informadas en cada etapa: transformaciones adecuadas de datos, selección pertinente de algoritmos y evaluación crítica de la calidad y utilidad de los patrones descubiertos.

Estructura y Metodología del Curso

El curso se organiza para facilitar un aprendizaje progresivo y aplicado:

- **Organización Temporal:** Consta de 12 clases, agrupadas en **tres bloques temáticos** de 4 clases cada uno.
- **Dinámica Semanal:** Cada clase se despliega en **dos jornadas complementarias**:
 - **Viernes:** Dedicados generalmente a la **discusión conceptual, fundamentos teóricos** y marcos de referencia.
 - **Sábados:** Enfocados en la **implementación práctica en Python**, resolución de ejercicios, análisis de casos reales y aplicación de algoritmos.
- **Flexibilidad:** Aunque existe esta estructura base, los contenidos teóricos y prácticos **podrán adaptarse y solaparse** según las necesidades y la dinámica del grupo, valorando la interacción.





El Enfoque: Combinación de Teoría y Práctica

La propuesta se basa en la sinergia entre el saber conceptual y el hacer práctico:

“Se combina teoría y práctica para formar criterio analítico. La teoría permite pensar críticamente sobre los procedimientos y sus fundamentos; la práctica permite comprobar empíricamente la validez y aplicabilidad de ideas y modelos.”

Este proceso iterativo facilita el aprendizaje no solo de técnicas de minería de datos, sino también de:

- **Estrategias de Abordaje:** Desarrollo de habilidades generales para el análisis de datos.
- **Anticipación de Problemas:** Capacidad para identificar y sortear obstáculos comunes.
- **Código Efectivo:** Desarrollo de habilidades en Python para análisis de datos.
- **Buenas Prácticas:** Adopción de metodologías de trabajo robustas y reproducibles.
- **Resolución de Problemas:** Identificación y abordaje de desafíos específicos en proyectos de data mining.

Bloque 1: Fundamentos del Análisis con Datos

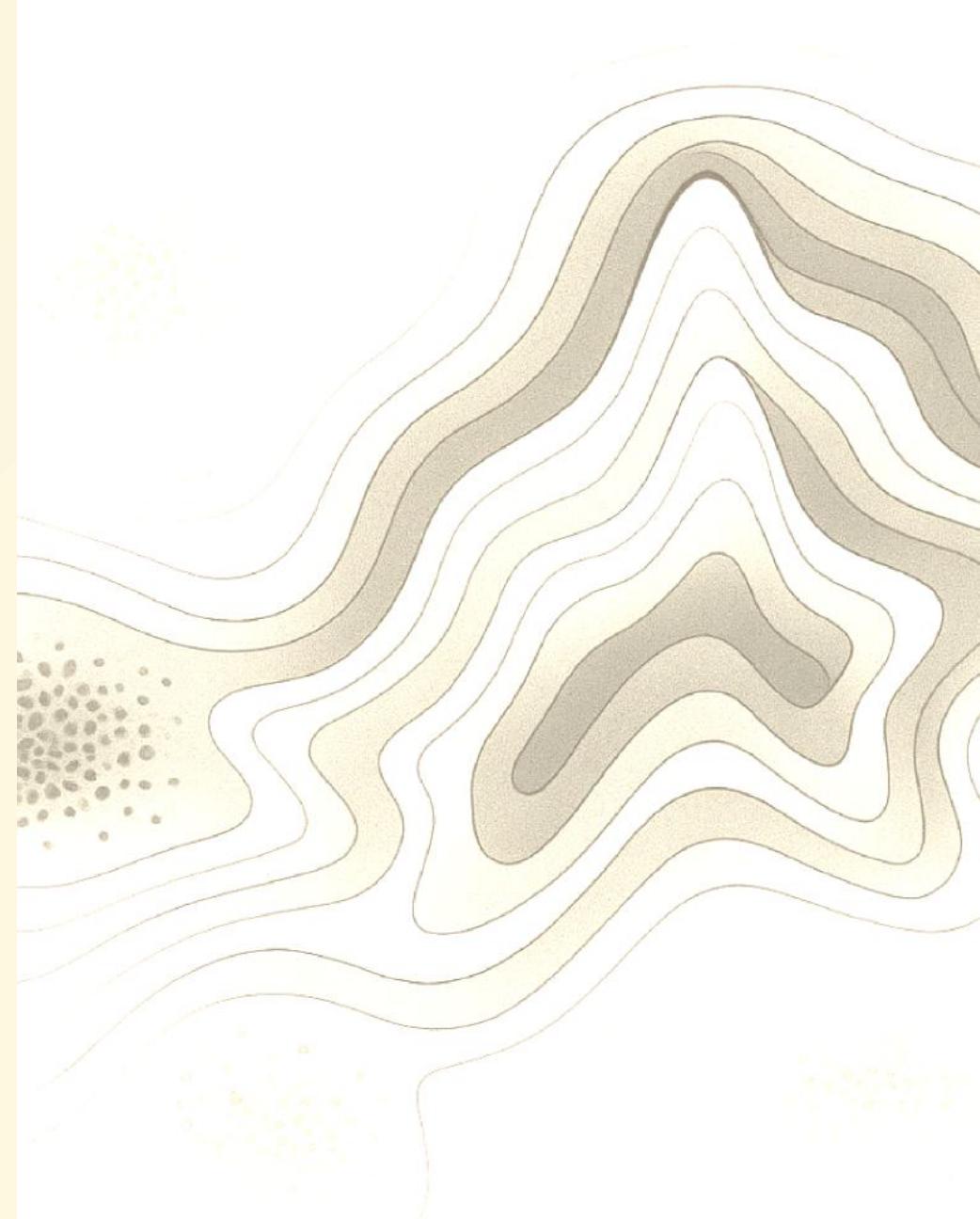
Este primer bloque sienta las bases conceptuales y técnicas del curso:

- **Definición y Relevancia:** Se explorará en profundidad qué es la minería de datos y su relevancia en el contexto actual de sobreabundancia de información.
- **Principios Clave:** El foco se pondrá en el objetivo de descubrir conocimiento **útil, comprensible y no trivial**, diferenciándolo de hallazgos obvios o irrelevantes.
- **Comparación con otros Enfoques:** Se analizarán las diferencias fundamentales con metodologías puramente confirmatorias o exclusivamente predictivas.
- **Herramientas Iniciales:** Se establecerán las bases técnicas en **Python** necesarias para abordar las tareas prácticas de los bloques subsiguientes.

Bloque 2: Preparación y Transformación de Datos

El segundo bloque se enfoca en una etapa crítica: la adecuación de los datos para el análisis:

- **Calidad de Datos:** Se abordará la identificación y corrección de problemas comunes como errores, inconsistencias, valores atípicos (outliers) y datos faltantes.
- **Ingeniería de Atributos:** Se presentarán técnicas para la **reducción de dimensionalidad**, codificación de variables categóricas y limpieza estructural de datasets.
- **Segmentación Exploratoria:** Se introducirá la **segmentación automática** mediante técnicas no supervisadas como K-means, y el análisis de perfiles resultantes para comprender la estructura inherente de los datos.





Bloque 3: Modelado, Evaluación e Integración

El último bloque consolida el aprendizaje mediante la aplicación de modelos y la evaluación de resultados:

- **Técnicas de Modelado:** Se profundizará en **árboles de decisión** para clasificación/predicción y **reglas de asociación** para descubrir relaciones frecuentes.
- **Evaluación Rigurosa:** Se establecerán criterios para determinar si un patrón o modelo es **confiable, interpretable y útil** en su contexto. Se discutirán métricas y estrategias de evaluación.
- **Validación y Generalización:** Se implementarán técnicas como la **validación cruzada** para asegurar la robustez y generalización de los hallazgos. Se revisará el proceso KDD.
- **Herramientas Avanzadas:** Se explorarán conceptos actuales como **interpretabilidad y explicabilidad** de modelos (XAI) y su importancia práctica.

Más Allá de los Datos: Preguntas y Miradas Críticas

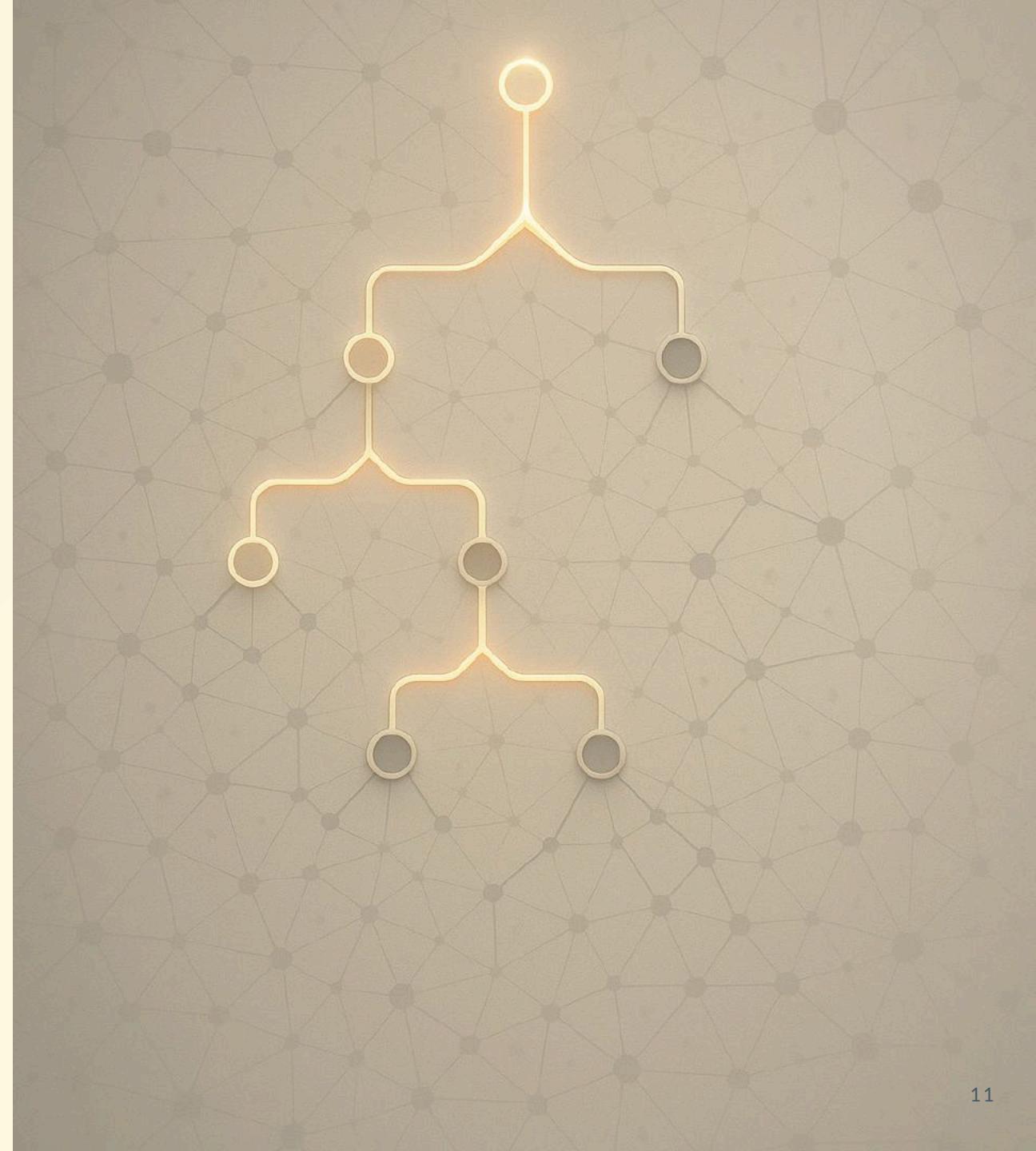
El trabajo con datos es esencial, pero no constituye el fin último del análisis.

“El curso implica un trabajo intenso con datos, pero, sobre todo, aborda el desafío de formular preguntas estructurantes, realizar elecciones metodológicas fundadas y mantener miradas críticas sobre todo el proceso analítico e interpretativo.”

El objetivo es desarrollar criterios sólidos para **descubrir, entre grandes datos, aquel conocimiento que verdaderamente valga la pena comprender, utilizar y eventualmente comunicar**. El foco estará en las **decisiones, las preguntas y las formas de mirar los problemas**.

Definiendo la Minería de Datos

**Conceptos
Fundacionales,
Contexto y
Comparaciones**





Introducción al Campo

El objetivo de esta sección es introducir la minería de datos no como una simple técnica, sino como un **enfoque complejo** que integra métodos computacionales, inferencia y análisis contextual.

Resulta central comprender que **la minería de datos trasciende la aplicación algorítmica ciega**; es un proceso cuyo núcleo reside en el **descubrimiento de patrones que sean útiles, comprensibles y no triviales**.

Este eje conceptual distingue al campo frente a otros enfoques como la estadística inferencial, el análisis exploratorio de datos o el machine learning centrado exclusivamente en la predicción.

Data Mining en KDD: Definición Canónica

Una definición fundacional, propuesta por Fayyad, Piatetsky-Shapiro y Smyth (1996), establece:

*“Data mining is the **nontrivial** process of identifying **valid**, **novel**, **potentially useful**, and ultimately **understandable** patterns in data.”*

(La minería de datos es el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles en los datos).”

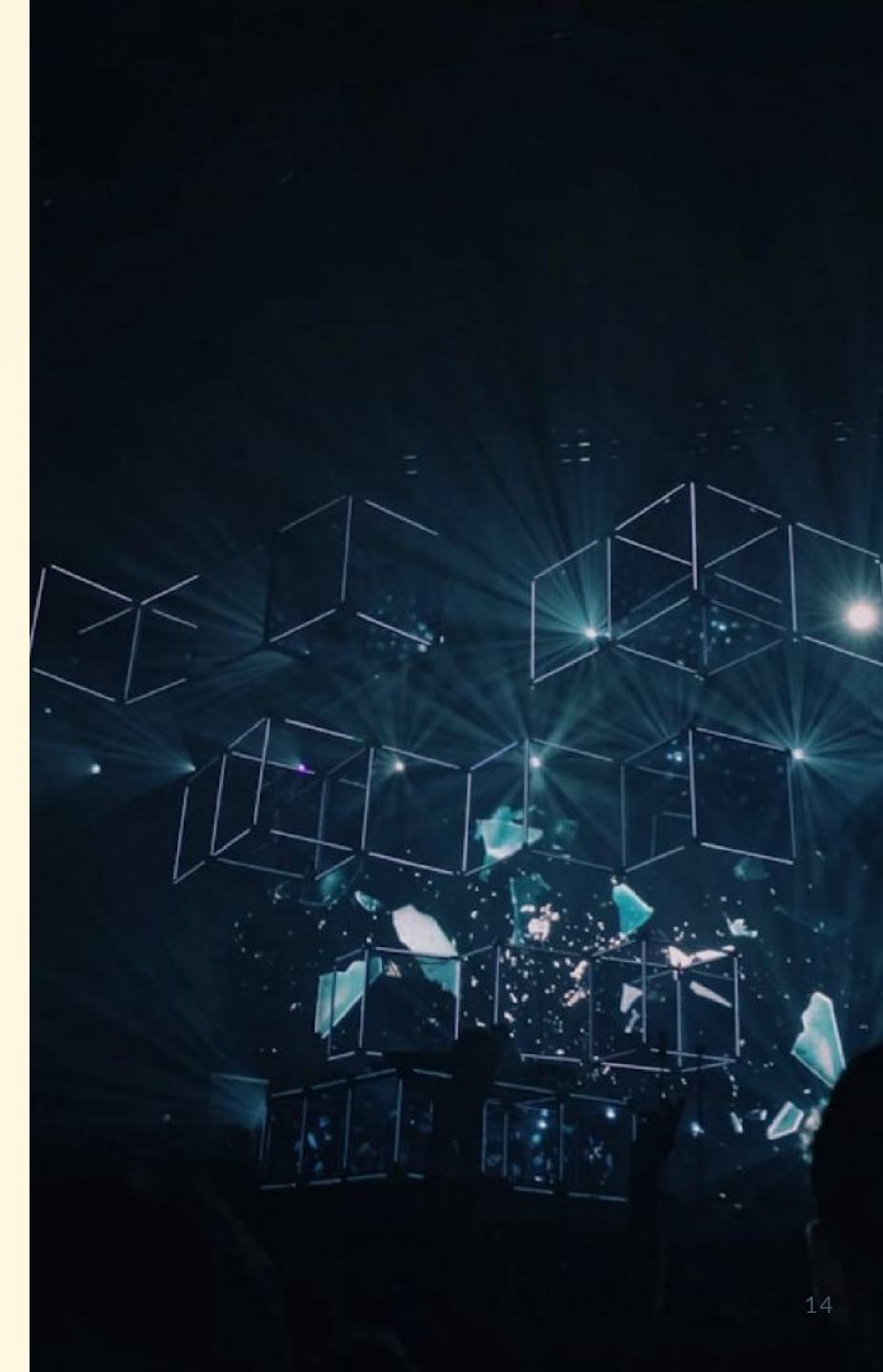
Esta definición marcó un hito por la precisión de sus términos:

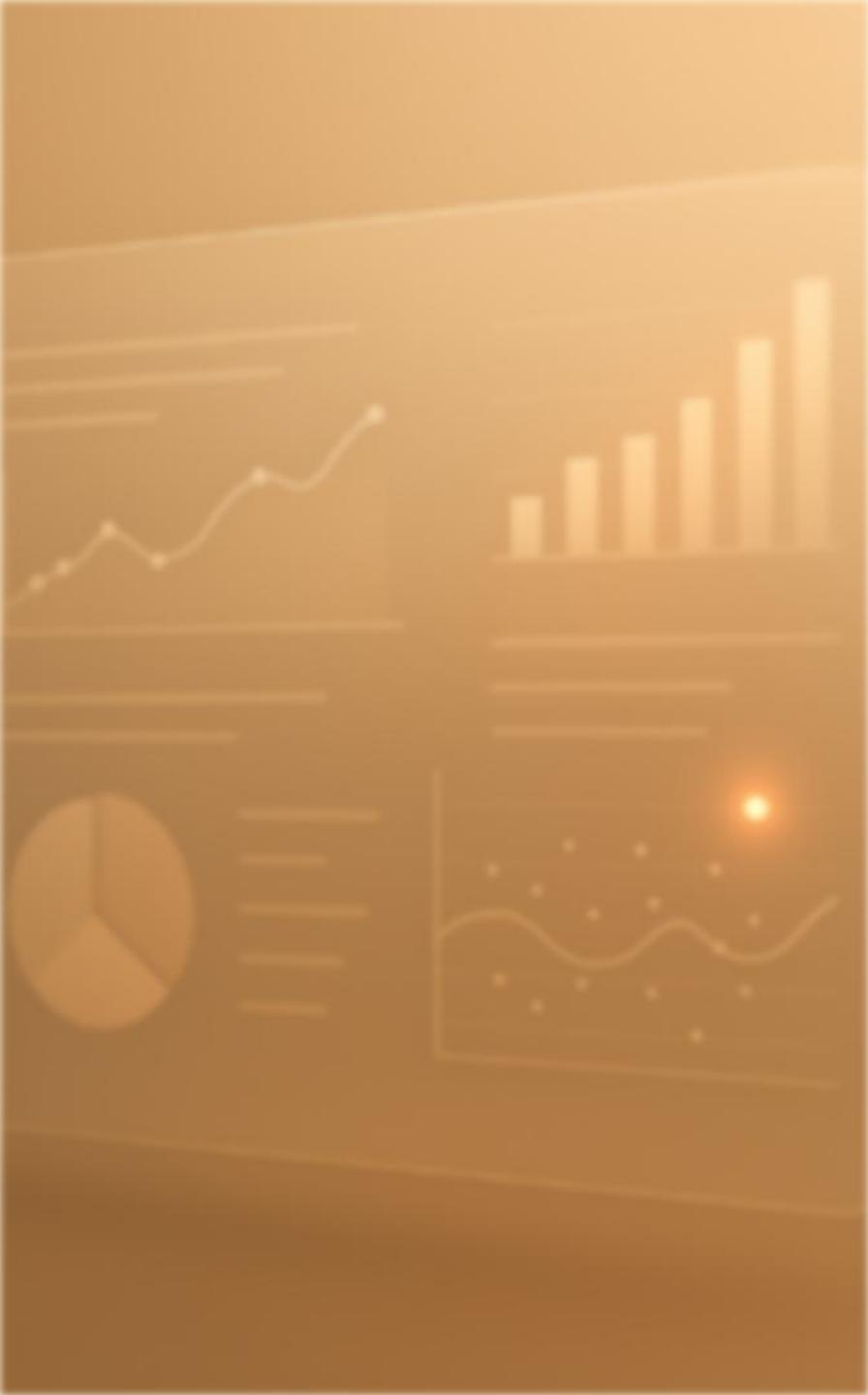
- **No Trivial:** El proceso debe generar conocimiento no evidente. Excluye patrones obvios o deducibles mediante inspección directa o estadísticas simples (medias, modas). Requiere métodos computacionales, heurísticos o inductivos para emergir.
- **Válido:** El patrón identificado no debe ser aleatorio o espurio. Debe poseer consistencia en los datos y ser potencialmente replicable o validable, aunque no necesariamente bajo los estrictos estándares formales de la estadística clásica.
- **Potencialmente Útil:** El conocimiento descubierto debe poseer aplicabilidad práctica. Debe poder guiar acciones, decisiones (empresariales, políticas), predicciones o segmentaciones. Su valor no radica solo en ser interesante o complejo.
- **Comprendible:** Los patrones deben poder ser interpretados, al menos parcialmente, por un experto humano. Esto exige un puente entre la salida algorítmica y el sentido común o la cognición humana, articulando procesamiento automático y capacidad comunicativa.

Marco Histórico y Técnico: El KDD

La definición de Fayyad et al. surge en la década de 1990, un período caracterizado por:

- **Explosión de Datos:** Los avances en la capacidad de almacenamiento superaban ampliamente la capacidad humana o estadística tradicional para analizar la información acumulada.
- **Necesidad de Descubrimiento:** Se hizo imperativa una disciplina que convirtiera el **almacenamiento masivo en conocimiento útil**, no mediante más almacenamiento, sino a través del **descubrimiento automatizado**.
- **Proyecto KDD:** Este contexto dio origen al marco del **Knowledge Discovery in Databases (KDD)**. En este proceso integral, la minería de datos se posiciona como **una de las etapas intermedias cruciales**, no como sinónimo del proceso completo, que incluye fases previas (selección, limpieza, transformación) y posteriores (evaluación, interpretación, despliegue).





Delimitando el Campo: Puntos de Comparación

Para precisar la identidad de la Minería de Datos, es útil abordar tres puntos de comparación:

- 1. Definición y alcance del Análisis Exploratorio de Datos (EDA), y su relación (vínculo o diferencia) con la estadística inferencial.**
- 2. La relación específica entre EDA y Minería de Datos.**
- 3. Las diferencias metodológicas y epistémicas sustantivas entre EDA, la Minería de Datos y la estadística.**

Comparativa: EDA, Estadística Inferencial y Minería de Datos

Dimensión	Análisis Exploratorio de Datos (EDA)	Estadística Inferencial	Minería de Datos
Propósito	Explorar, describir, visualizar, sugerir patrones preliminares	Confirmar/rechazar hipótesis (sobre poblaciones desde muestras)	Descubrir patrones útiles, comprensibles y no triviales en grandes volúmenes de datos
Unidad de análisis	Variables individuales o pares (1-2 variables)	Población (inferida desde muestra representativa)	Estructuras complejas, relaciones multivariadas, patrones latentes
Herramientas	Gráficos (histogramas, boxplots), frecuencias, medidas resumen	Modelos probabilísticos, pruebas de hipótesis, estimadores	Algoritmos (clustering, reglas, árboles, redes), visualización computacional avanzada, heurísticas
Supuestos teóricos	Mínimos o implícitos (flexibilidad)	Fuertes y explícitos (normalidad, independencia, etc.)	Débiles o inexistentes; conocimiento emerge inductivamente de los datos
Tipo de conocimiento	Intuiciones, identificación de patrones simples, detección errores	Validación estadística formal de hipótesis predefinidas	Conocimiento aplicado, contextualizado , empíricamente útil y accionable
Escala de datos	Pequeña a mediana	Pequeña a mediana (foco en representatividad)	Grande a masiva (Big Data)
Automatización	Bajo (manual, guiado por el analista)	Medio (software estadístico, intervención humana clave)	Medio a Alto (procesos semi-automáticos/automáticos con interpretación humana)
Relación con contexto	Alta (interpretación depende del dominio)	Media (supuestos generales, diseño experimental estándar)	Muy Alta (patrones deben ser útiles y comprensibles en dominios específicos)
Epistemología	Inductiva, visual, exploratoria	Deductiva, confirmatoria, basada en modelos teóricos	Inductiva, empírica, orientada al descubrimiento útil y accionable

EDA: Estadística Exploratoria

El **Análisis Exploratorio de Datos (EDA)** es, en efecto, parte del campo estadístico, pero se distingue de su vertiente **inferencial o confirmatoria**.

- **Origen:** Fue desarrollado por John Tukey (1970s) como una reacción al uso excesivo de pruebas de hipótesis sin un conocimiento previo adecuado de los datos.
- **Filosofía:** Propone utilizar gráficos, resúmenes numéricos y herramientas visuales/descriptivas para **"escuchar"** lo que los datos **sugieren**, sin imponerles a priori una estructura formal.
- **Herramientas Típicas:** Histogramas, diagramas de caja (boxplots), tablas de frecuencia, matrices de dispersión (scatter plots).
- **Naturaleza:** Es un enfoque **descriptivo e inductivo**. Su función es **preparar el terreno** para análisis posteriores, generar intuiciones, detectar anomalías o visualizar estructuras básicas. No busca generalizar a poblaciones ni probar hipótesis preestablecidas formalmente.

“👉 EDA representa un enfoque estadístico **inductivo, visual y flexible**, distinto de la lógica confirmatoria tradicional basada en hipótesis y pruebas formales.”

Diferencias entre EDA y Minería de Datos

Aunque ambos enfoques comparten un **espíritu inductivo y exploratorio** (no suelen basarse en hipótesis formales a priori), existen diferencias significativas en **escala, automatización y tipo de descubrimiento**:

Característica	Análisis Exploratorio de Datos (EDA)	Minería de Datos
Unidad de análisis	Variable o par de variables	Estructuras multidimensionales, patrones complejos
Herramientas	Gráficos, estadísticas descriptivas	Algoritmos (clustering, reglas, árboles, etc.)
Grado de automatización	Manual, guiado por el analista	Algorítmico, semi-automático o automático
Volumen de datos	Pequeño a mediano	Grande o masivo (millones de registros, variables)
Profundidad analítica	Explora, resume, detecta anomalías	Descubre relaciones no evidentes, segmenta, predice
Contexto original	Ánalysis científico y académico	Problemas aplicados (negocios, políticas, industria)

- EDA no tiene como objetivo primario encontrar patrones complejos ocultos, sino ofrecer una visión general, generar intuiciones o detectar errores.
- La Minería de Datos, en cambio, sí persigue estructuralmente encontrar patrones no triviales y potencialmente útiles, a menudo mediante algoritmos que superan la capacidad humana de inspección directa.

Validez de la Diferenciación

Si bien la frontera no es absolutamente rígida (ambos son inductivos y no requieren hipótesis fuertes), la distinción entre EDA y DM se justifica por:

- **Escala y Complejidad del Análisis:** EDA opera eficazmente en problemas de tamaño pequeño a mediano; la Minería de Datos está diseñada para **estructuras grandes y complejas (Big Data)**.
- **Grado de Automatización y Tecnificación:** En EDA, el analista guía el proceso visualmente; en Minería de Datos, los **algoritmos desempeñan un rol activo** en la sugerencia de estructuras, incluso aquellas no anticipadas por el analista.
- **El Objetivo Central:** EDA fundamentalmente **resume y visualiza**; Minería de Datos busca activamente **descubrir conocimiento oculto y actionable**.

“En resumen: EDA es estadística de tipo exploratoria. La Minería de Datos retoma y expande ese enfoque exploratorio mediante herramientas computacionales más potentes, orientadas al descubrimiento de patrones más complejos en grandes volúmenes de datos. No son excluyentes; suelen articularse secuencialmente.”





Visiones en Data Mining

No existe una única definición universalmente aceptada. Se pueden identificar dos visiones principales:

1. Visión Técnica/Algorítmica: Considera DM principalmente como una colección de algoritmos computacionales que encuentran patrones automáticamente (e.g., árboles de decisión, clustering, reglas de asociación, redes neuronales, PCA).

- *Tiende a priorizar el rendimiento técnico (velocidad, precisión) y puede desvincular el análisis del contexto o del problema de investigación.* La pregunta clave "**¿PARA QUÉ?**" puede quedar relegada. Sin un propósito claro, juzgar la utilidad o comprensibilidad del patrón se vuelve difícil.

2. Visión Procesual (KDD): Entiende DM como una etapa crucial dentro de un proceso más amplio (**KDD**). El éxito depende no solo del algoritmo, sino también de:

- Una adecuada **preparación** (limpieza, transformación, selección de datos).
- Una **aplicación metodológica** informada.
- Un **uso posterior** riguroso (validación, interpretación, acción).
- *Requiere experticia metodológica y contextual para elegir datos,*

Síntesis de Diferencias Clave

La Minería de Datos se distingue de disciplinas afines en dimensiones específicas:

- **Estadística Inferencial:** Parte de modelos probabilísticos a priori; su objetivo es generalizar de muestra a población mediante pruebas de hipótesis formales. Es **deductiva y confirmatoria**.
- **Análisis Exploratorio de Datos (EDA):** Busca explorar datos sin hipótesis formales; usa resúmenes y visualizaciones para detectar patrones o anomalías. Es **inductivo y preliminar**.
- **Machine Learning (ML):** Se enfoca principalmente en la **predicción**; los algoritmos aprenden reglas automáticamente, a menudo en modelos "caja negra" donde la precisión prima sobre la interpretabilidad.
- **Minería de Datos (DM):** Su foco es **descubrir patrones útiles, novedosos y comprensibles**. Se apoya en técnicas de ML, EDA y estadística, pero requiere un **equilibrio entre potencia técnica, interpretación contextual y aplicación práctica**.

Tensión Epistemológica: Existen debates sobre si DM es una extensión automatizada de la estadística (Hand et al., 2001) o un campo autónomo con lógica inductiva propia (Witten et al., 2016). Esta discusión impacta en criterios de validación y métricas de éxito.





Descubrimiento No Trivial

(I)

Los conceptos de **utilidad** y **no trivialidad** son centrales para definir el valor de la Minería de Datos:

- **Descubrimiento:** Implica encontrar conocimiento **no explícito previamente** en los datos ni en los modelos mentales del analista. Va más allá de la consulta o el cruce simple; es **hallar estructuras o relaciones nuevas**.
- **No Trivialidad:** El patrón descubierto **no debe ser obvio**. Requiere una operación analítica compleja (algorítmica, heurística) que extraiga conocimiento **no visible por simple inspección**.
- **Utilidad:** El patrón debe tener **valor práctico** en un contexto real. Puede manifestarse como una mejor predicción, una segmentación más eficaz, una interpretación más profunda o una guía para la acción. No basta con ser teóricamente interesante; debe **servir a un propósito concreto**.

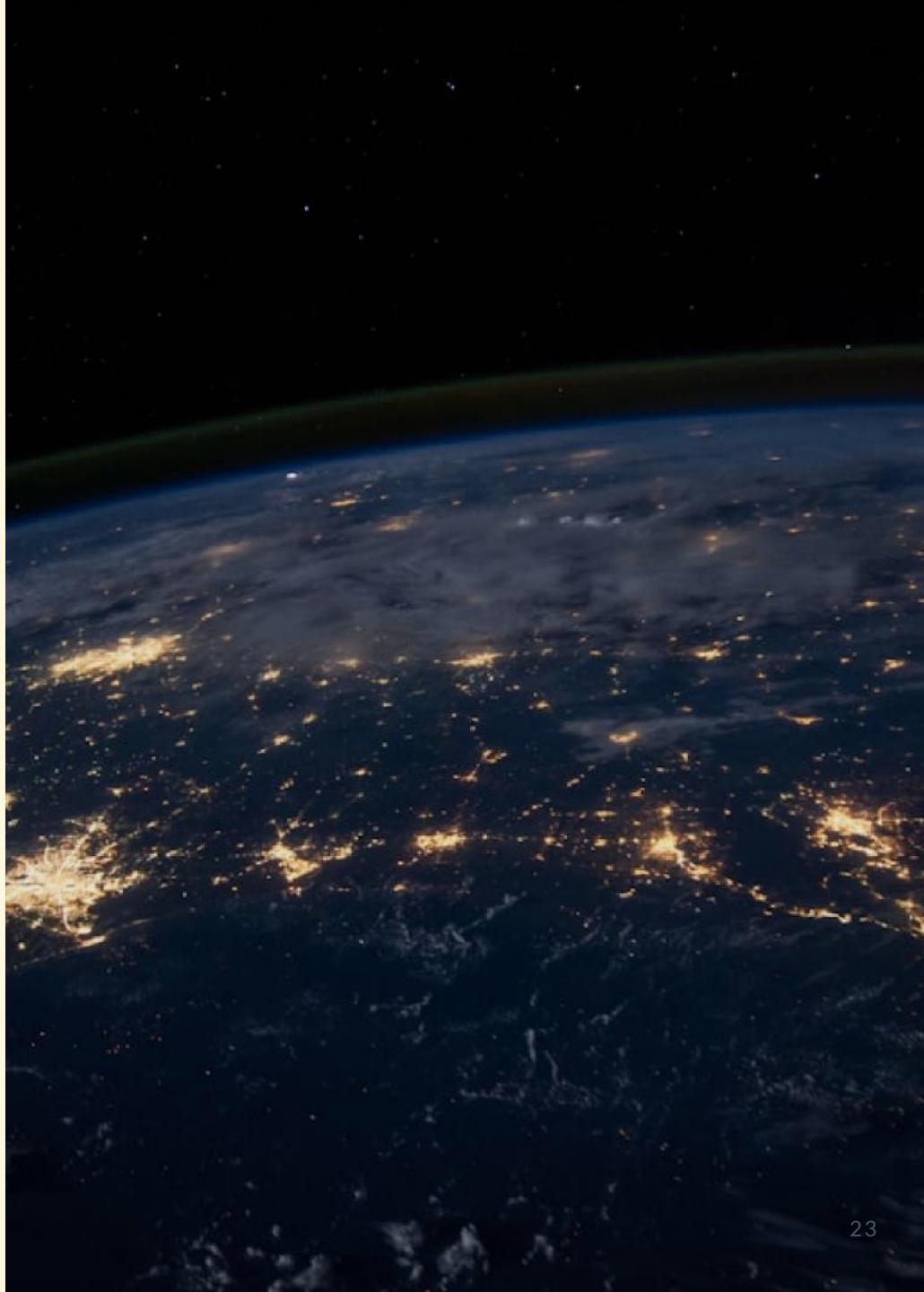
Descubrimiento No Trivial (III)

Estos principios tienen raíces técnicas e históricas:

- **Reacción al "Data Warehouse" Inerte:** Surgieron como crítica al paradigma de los 80s/90s de "almacenar todo esperando que sirva". La acumulación masiva de datos no generaba valor por sí misma.

*“ Se produjo la paradoja: **mucho dato, poco conocimiento**. Era necesario pasar del almacenamiento a la extracción de señales significativas. ”*

- **Combinación de Enfoques:** La Minería de Datos integra:
 - **Heurísticas:** Métodos prácticos que priorizan la efectividad empírica sobre la validación formal estricta, adecuados para datos complejos y voluminosos. Buscan resultados **útiles y suficientemente buenos**. (*Una heurística como "regla de sentido común algorítmico"*).
 - **Técnicas Computacionales Inductivas:** Algoritmos que identifican patrones latentes sin modelos teóricos a priori, aprendiendo de los datos.
- **Aplicaciones Prácticas:** Permiten automatizar el descubrimiento en áreas como marketing (segmentación), bioinformática, industria





Punto Crítico Metodológico: Descubrir vs. Predecir

Una distinción crucial separa a DM de enfoques puramente predictivos:

“La minería de datos *no busca primariamente la predicción perfecta*, sino el *descubrimiento de conocimiento accionable*. ”

- **Foco en ML Puro:** A menudo centrado en **optimizar la precisión predictiva** (minimizar error), incluso a costa de la interpretabilidad (modelos "caja negra" como Deep Learning). Métricas clave: Accuracy, AUC, etc.
- **Foco en Minería de Datos:** Busca un **equilibrio entre descubrimiento, utilidad e interpretación**. Un patrón puede ser valioso si es **comprendible, revela algo nuevo y permite tomar mejores decisiones**, aunque no sea el más preciso predictivamente.

Se valora la construcción de **modelos interpretables orientados a problemas reales**. El objetivo es **balancear descubrimiento, utilidad e interpretación**, desplazando el foco exclusivo de la precisión algorítmica.

Minería de Datos, Ciencia de Datos y Aprendizaje Automático

Diferencias Conceptuales y Metodológicas

Objetivo de la Sección

Esta sección tiene como objetivo establecer **comparaciones conceptuales, metodológicas y técnicas** y, en cualquier caso, desambiguar dudas entre tres campos fundamentales del análisis de datos:

- Minería de Datos (Data Mining - DM)
- Ciencia de Datos (Data Science - DS)
- Aprendizaje Automático (Machine Learning - ML)

Se busca clarificar la definición de cada enfoque, sus diferencias en objetivos, herramientas y escalas de análisis, así como el lugar que cada uno ocupa dentro del ecosistema contemporáneo del análisis de datos.



Diferenciación Conceptual por Niveles

Una forma de visualizar estos campos es a través de su nivel de abstracción y rol en el ecosistema de análisis:

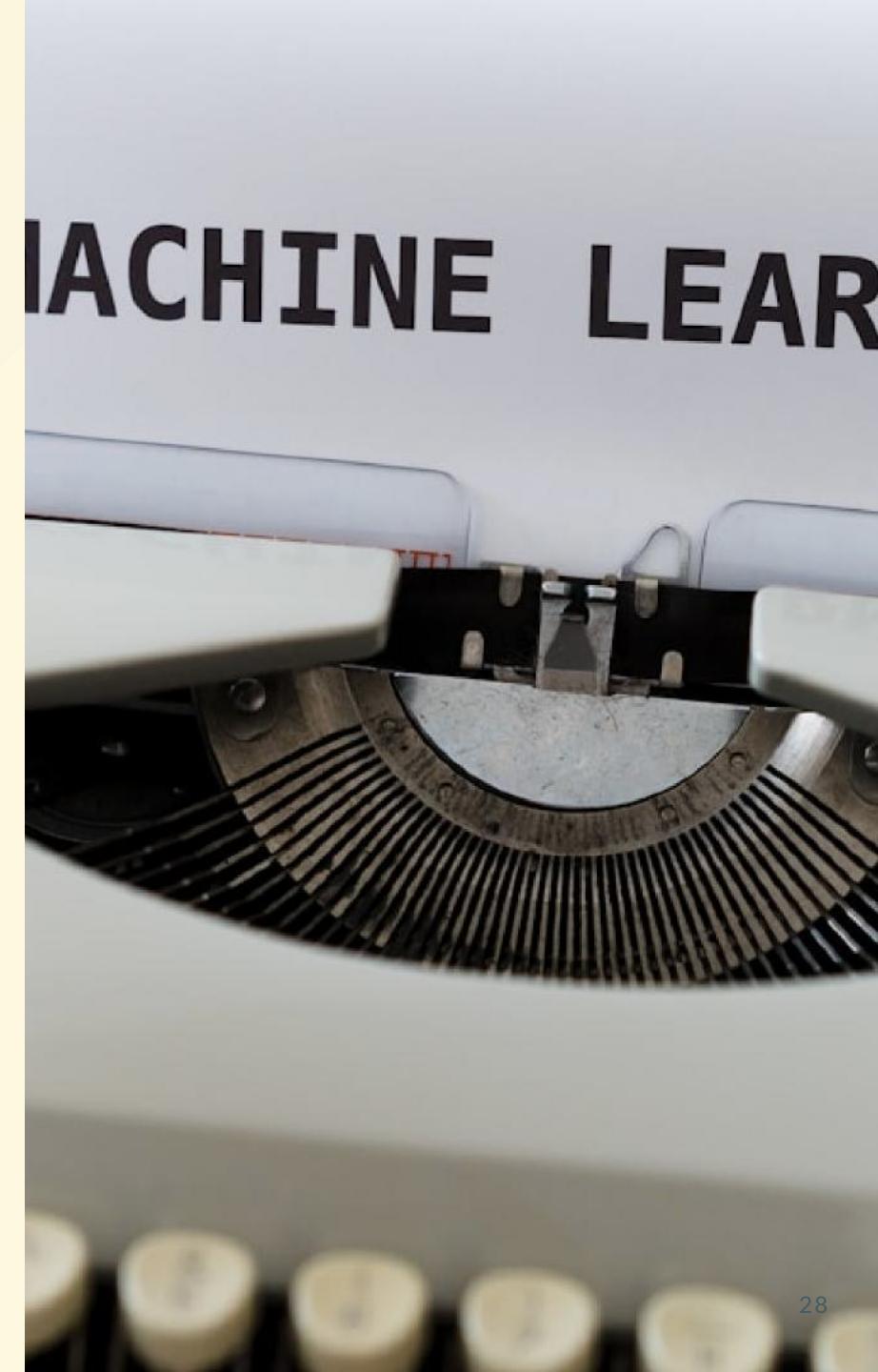
Nivel	Denominación	Naturaleza Principal	Rol en el Ecosistema
Método	Machine Learning (ML)	Conjunto de técnicas algorítmicas para predicción/clasificación	Herramienta especializada para tareas específicas
Enfoque	Minería de Datos (DM)	Proceso inductivo de descubrimiento de patrones útiles/no triviales	Forma de análisis orientada a generar conocimiento aplicable
Paradigma	Ciencia de Datos (DS)	Marco metodológico transversal e integrador	Orquesta métodos y enfoques para resolver problemas con datos

- **ML** constituye el conjunto de herramientas algorítmicas.
- **DM** representa un enfoque metodológico específico para el descubrimiento de conocimiento.
- **DS** opera como el marco amplio que integra el proceso completo.

1.1 Aprendizaje Automático (ML)

El ML se define como un conjunto de **métodos algorítmicos** diseñados para aprender reglas generales a partir de datos de entrenamiento.

- **Objetivo Central:** La optimización de la capacidad predictiva o clasificatoria, buscando minimizar el error en nuevas observaciones.
- **Técnicas Fundamentales:**
 - **Redes Neuronales Profundas:** Modelos bio-inspirados para reconocer patrones complejos (imágenes, texto, voz).
 - **Métodos de Ensamblado (Boosting, Random Forest):** Combinación de modelos simples para mejorar la precisión general.
 - **Otros:** SVM, k-NN, Regresión Logística; técnicas para clasificación, regresión o detección de anomalías.
- **Evaluación:** Se basa en **métricas de desempeño cuantitativas** (Accuracy, F1-Score para clases desbalanceadas, AUC para clasificación binaria, etc.).
- **Interpretabilidad:** Tiende a emplear modelos de "caja negra", donde la lógica interna no es fácilmente interpretable por humanos.
- **Orientación:** Predominantemente **data-driven**, sin un compromiso inherente con la interpretabilidad o el contexto más allá de la tarea predictiva.





1.2 Minería de Datos (DM)

La DM se caracteriza como un **enfoque inductivo** orientado al **descubrimiento de patrones útiles, válidos y no triviales** en grandes volúmenes de datos.

- **Objetivo Principal:** Generar **conocimiento accionable** a partir de los datos. En general, se hace uso de estas herramientas:
- **1. Clustering (e.g., K-means, DBSCAN):** Agrupamiento automático de observaciones similares sin clases predefinidas.
- **2. Árboles de Decisión:** Modelos jerárquicos que clasifican mediante reglas comprensibles.
- **3. Reglas de Asociación:** Identificación de co-ocurrencias frecuentes (e.g., análisis de canasta de compra).
- **4. Detección de Outliers y Patrones Secuenciales:** Para encontrar anomalías o series de eventos relevantes.
- **Metodología:** **No parte de hipótesis formales**, permitiendo que los datos sugieran las estructuras relevantes.
- **Énfasis:** La **interpretabilidad y validación empírica** son cruciales; se valora que el patrón sea comprensible, contextualizable y útil.
- **Diferencia con ML Puro:** No busca solo predecir, sino también **entender e intervenir** sobre el fenómeno estudiado, integrando técnicas con foco en el conocimiento aplicable.

1.3 Ciencia de Datos (DS)

La DS es el **paradigma más reciente, transversal e integrador**, articulando el ciclo completo del análisis de datos, desde la **formulación del problema** hasta la **comunicación de resultados accionables**.

- **Alcance:** Cubre etapas como:
 - **Recolección y Almacenamiento de Datos** (Ingeniería de Datos).
 - **Preparación y Limpieza:** Identificación y corrección de errores, tratamiento de faltantes, etc.
 - **Modelado y Validación.**
 - **Visualización, Interpretación y Comunicación:** Elaboración de gráficos, informes y explicaciones para la toma de decisiones (Storytelling).
- **Competencias Múltiples:** Requiere habilidades en Estadística, Programación, Conocimiento del Dominio, Comunicación y Ética de Datos.
- **Integración:** Utiliza herramientas de DM, ML, Big Data, NLP, visualización avanzada, etc.
- **Rol:** Funciona como un **marco metodológico amplio** que articula **métodos (ML)** y **enfoques (DM)** con otras disciplinas.



Relación entre Ciencia de Datos y Minería de Datos

En los marcos actuales, la **Ciencia de Datos (DS)** se presenta frecuentemente como un **enfoque integrador** o "paraguas" que engloba a la Minería de Datos (DM), al Aprendizaje Automático (ML), la estadística, visualización, ingeniería de datos, etc.

“Bajo esta perspectiva dominante, DM es una de las técnicas o aproximaciones utilizables dentro de un proyecto de DS.”

Existen, no obstante, diferentes lecturas de esta relación:

Enfoque	Relación DS y DM
Instrumentalista (Visión amplia)	DM es una técnica dentro del flujo de trabajo de DS, subordinada a sus necesidades.
Histórica y Metodológica	DS es posterior y se apoya en el desarrollo previo de DM , que tiene su propia identidad, epistemología y campo de estudio consolidado.
Epistemológica Crítica	DM no es simplemente subsumible : posee un proyecto epistémico propio (descubrimiento útil, no trivial), que puede contradecir o complementar lógicas de DS (e.g., automatización total).

- DS puede considerarse el **marco operativo** más amplio.
- DM ocupa un lugar dentro de ese marco, pero manteniendo una **identidad epistémica diferenciada**.
- No es solo "una técnica más", sino una **forma específica de pensar y descubrir conocimiento**. La distinción radica en **el enfoque, el método y el tipo de saber** priorizado.



Justificación del Foco en Minería de Datos

¿Por qué centrar la formación en DM y no en el paradigma global de DS?

Profundidad Metodológica: Mientras DS abarca amplitud, a veces diluye el foco. DM ofrece un **marco coherente y riguroso** para el descubrimiento de patrones. Dominar un enfoque específico es valioso.

Núcleo del Descubrimiento Inductivo: El objetivo central de DM (encontrar **estructuras nuevas, útiles y comprensibles**) sigue siendo clave para análisis significativos, más allá de la predicción.

Formación de Analistas Integrales: En proyectos de DS, las tareas pueden segmentarse. DM promueve **analistas con criterio holístico**, capaces de conducir el proceso completo, no solo ejecutar fases técnicas.

Vínculo con Contexto y Pregunta: DM exige **interpretar patrones en función del problema**, algo que enfoques puramente técnicos pueden omitir.

Articulación Única: DM combina **inductividad, automatización e interpretación** en su núcleo (vs. Estadística: deductiva; ML: automatizado/opaco; EDA: inductivo/manual).

Base para una DS con Sentido: Una DS robusta requiere un **corazón exploratorio y comprensible** proporcionado por DM. Lo técnico sin

DM vs. DS: Alcance Comparado de Competencias

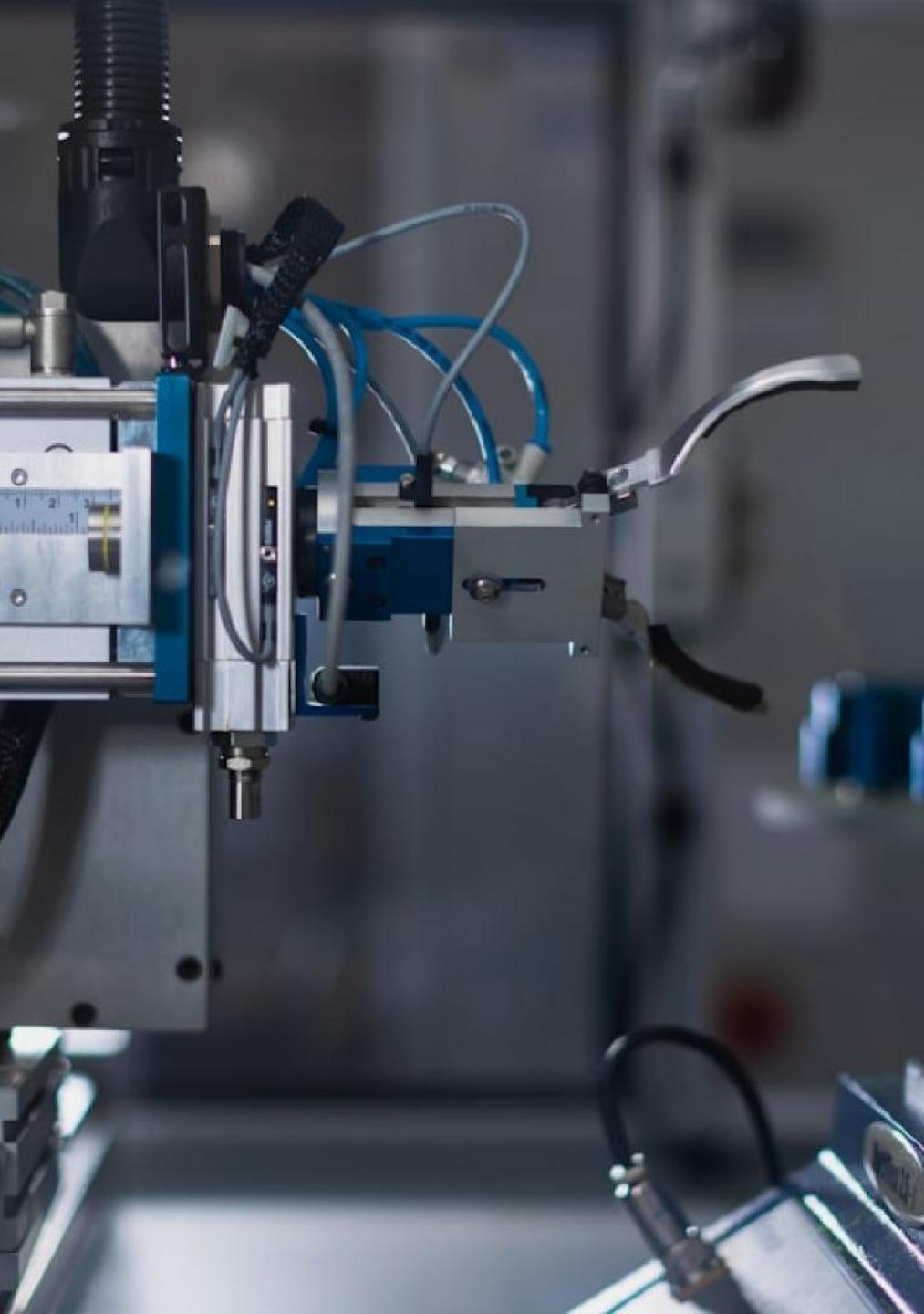
Esta tabla ilustra la distribución de competencias (fronteras porosas):

Etapa / Competencia	Presente en DS?	Presente en DM?	Central / Constitutiva en DM?
Formulación del Problema	Sí	A veces	No siempre. Central en DM guiada por dominio, menos en DM exploratoria pura.
Recolección y Almacenamiento (Ing. Datos)	Sí	No	No. DM generalmente asume datos disponibles.
Preparación y Limpieza de Datos	Sí	Sí	Sí. Fase crucial e inseparable.
Modelado y Análisis (Algorítmico/Estad.)	Sí	Sí	Sí. Núcleo técnico del descubrimiento.
Visualización e Interpretación	Sí	Sí	Sí. La comprensibilidad del patrón es definitoria.
Comunicación de Resultados (Storytelling)	Sí	A veces	No siempre. Fundamental en DM aplicada, menos en DM puramente exploratoria/automatizada.
Conocimiento del Dominio	Sí	Sí	Sí. Necesario para evaluar utilidad y relevancia.
Ética, Regulación, Escalabilidad	Sí	No	No. Consideraciones del marco (DS), no intrínsecas a la definición de DM.

DM se concentra en el **núcleo analítico**: preparación, modelado e interpretación, validado por el conocimiento del dominio.

Vínculos, Solapamientos y Distinciones

ML, DM y DS, aunque con historias y metodologías distintas, **no son campos aislados**. En la práctica, sus herramientas y objetivos **se articulan, superponen y complementan**, pero también pueden generar **tensiones si sus diferencias no se comprenden con claridad**.



3.1 Complementariedad Práctica

En proyectos reales, estos campos interactúan frecuentemente:

- Un proyecto de **Ciencia de Datos** puede **integrar tareas de Minería de Datos** para exploración inicial, segmentación de poblaciones o generación de hipótesis data-driven.
- La **Minería de Datos** a menudo **utiliza algoritmos de Machine Learning** como herramientas (árboles, clustering, redes), pero **priorizando la interpretabilidad y utilidad contextual** sobre la mera precisión.
- El **Aprendizaje Automático** se **beneficia de técnicas previas de Minería de Datos** (preparación de datos, selección de atributos, EDA) para mejorar la calidad y robustez de sus modelos.

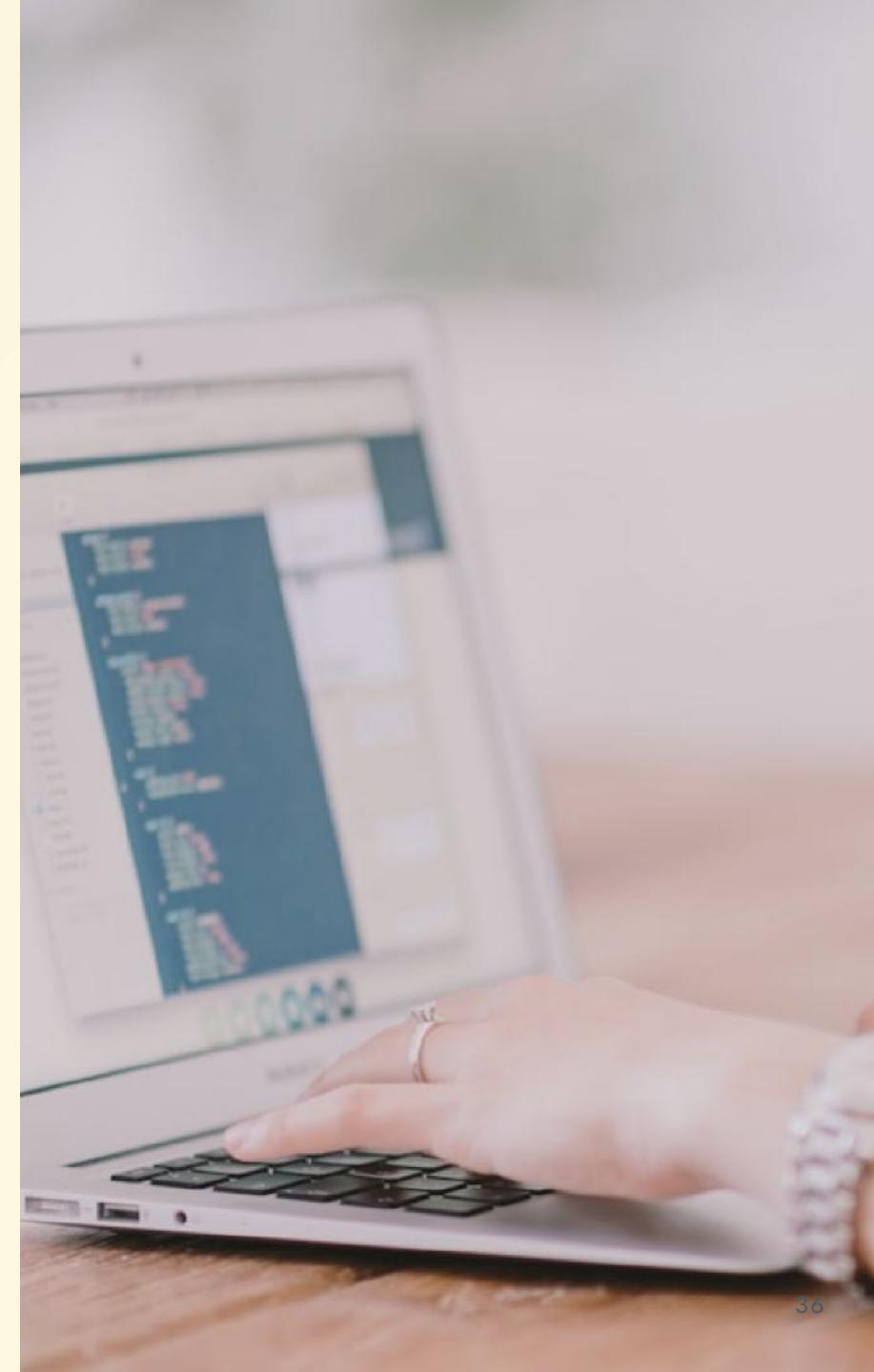
“La cuestión no es elegir uno sobre otro de forma excluyente, sino comprender el aporte específico de cada uno, su momento adecuado de aplicación y los objetivos que persigue.”

3.2 Solapamientos Históricos y Conceptuales

La evolución terminológica puede generar confusión:

- **Orígenes Compartidos:** En la literatura temprana (90s-2000s), *Data Mining* y *Machine Learning* se usaban a menudo como sinónimos o campos estrechamente relacionados.
- **Diferenciación Posterior:** Con la madurez de ambos campos, sus identidades se perfilaron:
 - **ML:** Adoptó un enfoque más **formal, técnico y centrado en el rendimiento algorítmico.**
 - **DM:** Consolidó su identidad como enfoque **inductivo, exploratorio y orientado a la comprensión contextual de patrones.**
- **Irrupción de DS:** La Ciencia de Datos, más reciente, integró elementos de ambos, con el riesgo inherente de **pérdida de claridad conceptual** si no se distinguen adecuadamente los roles de cada componente.

“Reconocer estos solapamientos ayuda a evitar confusiones terminológicas y metodológicas frecuentes.”





3.3 Tensiones Epistemológicas y Lugar Estratégico de DM

Las relaciones entre estos campos implican **diferencias profundas en la concepción del conocimiento, su validación y el rol del analista.**

- **Perspectiva Estadística Clásica:** Puede ver a ML/DM como modelado estadístico automatizado, útil pero limitado por la falta de transparencia o rigor formal.
- **Perspectiva de Ciencia de Datos:** Promueve integración amplia, pero a veces a costa de profundidad metodológica o generando enfoques técnicos desconectados del sentido analítico.

La Minería de Datos ocupa aquí un lugar estratégico:

“**Funciona como núcleo articulador del saber técnico-analítico, al combinar:**

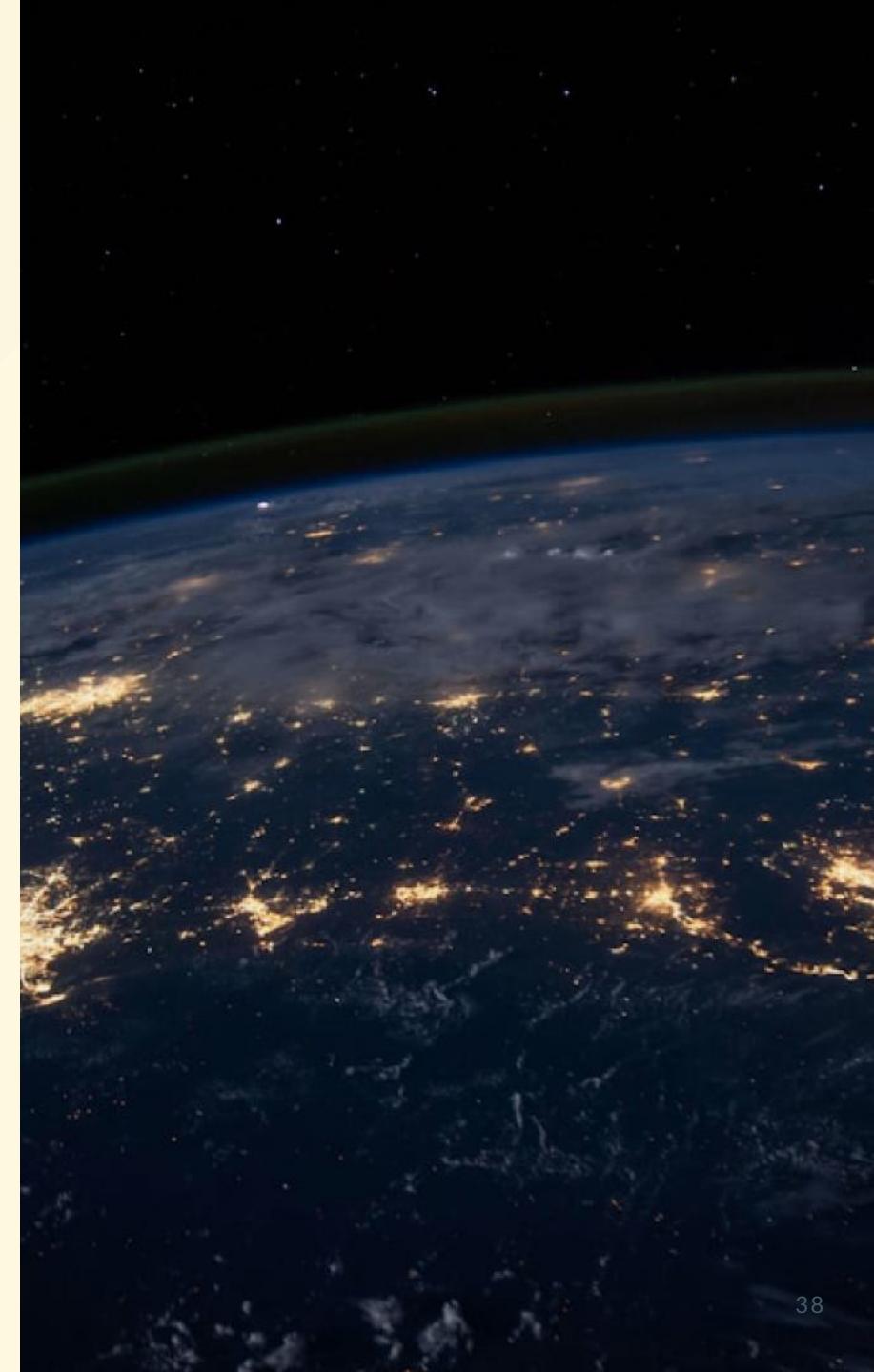
- *La rigurosidad conceptual (heredada de la estadística).*
- *La potencia algorítmica (compartida con ML).*
- *La orientación práctica e interpretativa (central en DS)*

Relevancia del Lugar Central de DM

Reconocer la identidad y el rol específico de DM es importante porque:

- Permite construir proyectos analíticos **con sentido**, equilibrando técnica y criterio.
- Ofrece una **base formativa sólida** desde donde articular otros métodos sin diluir el foco analítico.
- Históricamente, **equilibra automatización e interpretación** como objetivo constitutivo.
- Contribuye a formar **analistas con criterio**, capaces de ir más allá de la ejecución algorítmica.

“En síntesis: La Minería de Datos no es un mero eslabón intermedio, sino el punto de cruce donde se articulan método, herramienta y problema. Por ello, sigue siendo fundamental para formar analistas capaces de pensar críticamente con datos.”



Importancia de Comprender las Distinciones

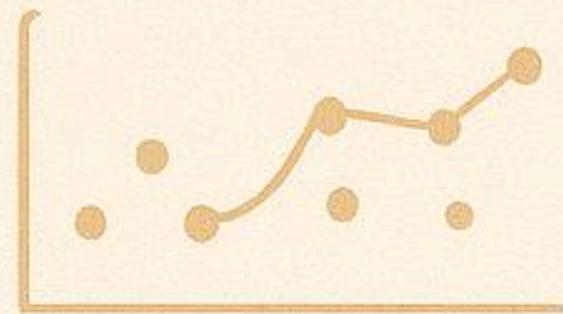
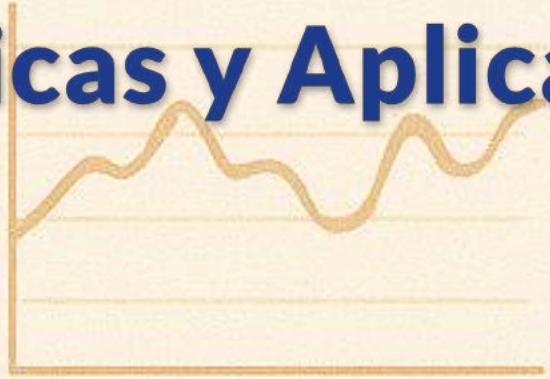
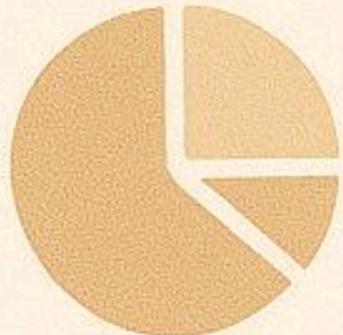
Entender las diferencias entre ML, DM y DS es crucial para:

- **Evitar Identificaciones Simplistas:** No reducir DM a ML, lo que ignora su rol analítico, procesual e interpretativo.
- **Preservar la Profundidad de DM:** Valorar su carácter exploratorio e interpretativo frente a enfoques puramente técnicos.
- **Comprender la Necesidad de DS:** Reconocer que DS requiere enfoques como DM para operar con criterio y producir conocimiento situado, no solo resultados funcionales.
- **Formar Analistas Críticos:** Capacitar profesionales para **elegir con fundamento** el enfoque o herramienta adecuada, según el contexto, los objetivos y los criterios de validación pertinentes.

“El objetivo no es jerarquizar paradigmas, sino reconocer el tipo de saber que produce cada uno y cómo pueden articularse inteligentemente sin perder identidad ni potencia analítica.”

Tareas Principales en Minería de Datos

Objetivos, Técnicas y Aplicaciones



Objetivo de la Sección

Cuáles son las **tareas fundamentales en minería de datos?** Se examinarán sus objetivos específicos, las técnicas algorítmicas asociadas y sus aplicaciones prácticas más relevantes.

El fin es que se comprenda la **diversidad de problemas** abordables mediante la minería de datos, así como los **enfoques y herramientas específicos** para cada tipo de tarea, proporcionando una visión estructurada del campo de acción de esta disciplina.

1. Clasificación General de Tareas en Minería de Datos

Las tareas en minería de datos se pueden agrupar según su objetivo principal y las técnicas empleadas:

Tipo de tarea	Objetivo Principal	Ejemplos de Aplicación	Algoritmos Principales Asociados
Clasificación	Asignar observaciones a categorías predefinidas (supervisado)	Diagnóstico médico, detección de fraude, predicción de abandono (churn)	Árboles de decisión, SVM, Redes Neuronales, Random Forest, Regresión Logística
Regresión	Predecir valores numéricos continuos (supervisado)	Predicción de precios, estimación de demanda, pronóstico de ventas	Regresión Lineal/Polinomial, Árboles de Regresión, Redes Neuronales, SVM para Regresión (SVR)
Clustering	Agrupar observaciones similares sin categorías predefinidas (no sup.)	Segmentación de clientes, detección de comunidades, análisis de documentos	K-means, DBSCAN, Clustering Jerárquico, Gaussian Mixture Models (GMM)
Reglas de Asociación	Descubrir relaciones frecuentes entre ítems (no supervisado)	Ánalisis de canastas de compra, sistemas de recomendación, análisis web	Apriori, FP-Growth, Eclat
Detección de Anomalías	Identificar observaciones inusuales o atípicas (no sup./semi-sup.)	Detección de fraude/intrusiones, monitoreo de sistemas, control de calidad	Isolation Forest, One-Class SVM, Local Outlier Factor (LOF), Métodos Estadísticos
Reducción de Dimens.	Reducir el número de variables preservando información (no sup.)	Visualización, preprocessamiento, compresión de datos, extracción features	PCA (Principal Component Analysis), t-SNE, UMAP, Autoencoders

2. Descripción Detallada de Tareas (I)

2.1. Clasificación

- **Definición:** Tarea **supervisada** que consiste en asignar etiquetas categóricas (clases) a nuevas observaciones, basándose en un modelo aprendido a partir de datos previamente etiquetados.
- **Características Principales:** Requiere datos de entrenamiento con etiquetas conocidas; produce modelos predictivos para categorías discretas; la evaluación se centra en la precisión de la asignación de clases.
- **Métricas de Evaluación Comunes:** Accuracy (Precisión Global), Precision (Exactitud de predicciones positivas), Recall (Sensibilidad, capacidad de detectar positivos), F1-Score (Media armónica de Precision y Recall), Matriz de Confusión, Curva ROC y AUC.
- **Aplicaciones Típicas:** Diagnóstico médico (e.g., benigno/maligno), detección de spam (spam/no spam), clasificación de documentos por tema, predicción de riesgo crediticio (buen/mal pagador).

2.2. Regresión

- **Definición:** Tarea **supervisada** cuyo objetivo es predecir un valor numérico **continuo** para nuevas observaciones, basándose en patrones aprendidos de datos históricos donde la variable objetivo es conocida.
- **Características Principales:** La variable objetivo es continua; busca modelar relaciones funcionales (lineales o no lineales) entre variables predictoras y la variable objetivo; permite interpretar la influencia de los predictores (coeficientes en modelos lineales).
- **Métricas de Evaluación Comunes:** MSE (Error Cuadrático Medio), RMSE (Raíz del Error Cuadrático Medio), MAE (Error Absoluto Medio), R² (Coeficiente de Determinación, proporción de varianza explicada).
- **Aplicaciones Típicas:** Predicción de precios (viviendas, acciones), pronóstico de demanda de productos, estimación de tiempos de

110128881121 901209-60 1301-1

110001 00210020

2. Descripción Detallada de Tareas (III)

2.5. Detección de Anomalías (Outlier Detection)

- **Definición:** Tarea enfocada en la identificación de observaciones, eventos o patrones que se desvían significativamente del comportamiento considerado normal o esperado dentro de un conjunto de datos.
- **Características Principales:** Detección de valores atípicos (outliers); a menudo es no supervisada o semi-supervisada (con pocos ejemplos de anomalías) y puede ser sensible al ruido en los datos.
- **Métricas de Evaluación (si hay etiquetas):** Precision, Recall y F1-Score (especialmente en clases desbalanceadas), Curva ROC y AUC, Tasa de Falsos Positivos.
- **Aplicaciones Típicas:** Detección de fraude en transacciones financieras o seguros, monitoreo de sistemas para identificar fallos o ataques, control de calidad en procesos industriales, seguridad informática (detección de intrusiones).

2.6. Reducción de Dimensionalidad

- **Definición:** Proceso de transformar datos desde un espacio de alta dimensión (muchas variables) a un espacio de menor dimensión, intentando preservar la mayor cantidad posible de información o estructura relevante.
- **Características Principales:** Facilita la compresión de datos; permite la visualización de datos complejos en 2D o 3D; puede ayudar a eliminar ruido y redundancia; a menudo se usa como paso de preprocessamiento para otras tareas.
- **Métricas/Criterios de Evaluación:** Varianza explicada (PCA), minimización de la distorsión o divergencia (t-SNE, UMAP), preservación de la estructura local/global, tiempo de cómputo, rendimiento de modelos posteriores.
- **Aplicaciones Típicas:** Visualización exploratoria de datos de alta dimensión, preprocessamiento para algoritmos sensibles a la "maldición

3. Selección de Tareas y Algoritmos

La elección de la tarea y el algoritmo adecuados es crucial y depende de varios factores.

3.1. Criterios para la Selección

- **Naturaleza del Problema:**
 - ¿Se dispone de etiquetas (Supervisado) o no (No Supervisado)?
 - ¿Se busca predecir una categoría (Clasificación) o un valor numérico (Regresión)?
 - ¿El objetivo es descubrir patrones ocultos (Clustering, Reglas) o realizar predicciones?
- **Características de los Datos:**
 - Tamaño del conjunto de datos (volumen).
 - Dimensionalidad (número de variables).
 - Calidad de los datos (ruido, valores faltantes).
 - Distribución de las clases (en problemas supervisados, ¿están balanceadas?).
- **Requisitos del Modelo:**
 - ¿Es fundamental que el modelo sea interpretable (e.g., árboles, reglas) o prima la precisión (e.g., redes profundas)?
 - ¿Necesita escalar a grandes volúmenes de datos?
 - ¿Debe ser robusto frente a outliers o ruido?
 - ¿Existen restricciones en el tiempo de entrenamiento o predicción?

3. Selección de Tareas y Algoritmos (Continuación)

3.2. Proceso Iterativo de Selección

La selección no es un paso único, sino un proceso que suele involucrar iteraciones:

- 1. Definición Clara del Problema:** Establecer objetivos medibles, métricas de éxito claras y restricciones operativas (tiempo, recursos, interpretabilidad).
- 2. Análisis Exploratorio de Datos (EDA):** Caracterizar los datos, visualizar distribuciones, identificar patrones preliminares, detectar posibles problemas (outliers, faltantes, sesgos).
- 3. Selección Preliminar del Enfoque:** Determinar el tipo de tarea

4. Consideraciones Prácticas Fundamentales

La aplicación exitosa de la minería de datos requiere atención a aspectos prácticos clave:

4.1. Preprocesamiento de Datos

Es una etapa crítica que consume gran parte del tiempo en un proyecto real. Incluye:

- **Limpieza:** Manejo de ruido, corrección de errores.
- **Tratamiento de Valores Faltantes:** Imputación o eliminación.
- **Normalización/Escalado:** Llevar variables a rangos comparables.
- **Tratamiento de Outliers:** Detección y decisión sobre su manejo.
- **Balanceo de Clases:** Técnicas para abordar desbalances en clasificación (oversampling, undersampling, SMOTE).
- **Codificación de Variables:** Transformación de variables categóricas a numéricas (One-Hot Encoding, Label Encoding).
- **Selección/Extracción de Características:** Reducir dimensionalidad o crear nuevas variables relevantes.

4. Consideraciones Prácticas Fundamentales

(II)

4.2. Evaluación Rigurosa de Modelos

No basta con entrenar un modelo; su evaluación debe ser sistemática:

- **Validación Cruzada (Cross-Validation):** Técnica estándar para estimar el rendimiento del modelo sobre datos no vistos y reducir el sobreajuste (overfitting). K-Fold es común.
- **Conjuntos Separados:** División explícita en conjuntos de entrenamiento, validación (para ajuste de hiperparámetros) y prueba (para evaluación final imparcial).
- **Métricas Adecuadas:** Elegir métricas relevantes para la tarea y el objetivo del negocio (e.g., no solo Accuracy en clases desbalanceadas).
- **Interpretación Contextual:** Analizar los resultados (e.g., matriz de confusión, errores específicos) en el contexto del problema.
- **Comparación Justa:** Utilizar las mismas particiones de datos y métricas al comparar diferentes modelos.

4.3. Despliegue y Monitoreo

La minería de datos no termina con el modelo entrenado:

- **Integración:** Incorporar el modelo en sistemas de producción existentes.
- **Actualización (Retraining):** Los modelos pueden degradarse con el tiempo (data drift, concept drift); es necesario reentrenarlos.

5. Casos de Estudio y Aplicaciones Sectoriales

La minería de datos tiene aplicaciones en una vasta gama de sectores:

5.1. Sector Financiero

- **Detección de Fraude:** Identificación de transacciones anómalas (tarjetas de crédito, seguros).
- **Scoring Crediticio:** Evaluación del riesgo de impago de solicitantes de crédito.
- **Análisis de Riesgo de Mercado:** Predicción de volatilidad, evaluación de inversiones.
- **Marketing Dirigido:** Segmentación de clientes para ofertas personalizadas.

5.2. Sector Retail (Comercio Minorista)

- **Análisis de Canasta de Compra:** Descubrimiento de asociaciones de productos (Reglas de Asoc.).
- **Sistemas de Recomendación:** Sugerencia de productos a clientes (filtrado colaborativo, basado en contenido).
- **Predicción de Demanda:** Optimización de inventarios y logística.
- **Segmentación de Clientes:** Creación de perfiles para campañas de marketing.

5. Casos de Estudio y Aplicaciones Sectoriales (II)

5.3. Sector Salud

- **Apoyo al Diagnóstico Médico:** Clasificación de imágenes médicas, predicción de enfermedades basada en historial.
- **Descubrimiento de Fármacos:** Análisis de datos genómicos y moleculares.
- **Medicina Personalizada:** Agrupamiento (Clustering) de pacientes para tratamientos específicos.
- **Monitoreo Remoto de Pacientes:** Detección de anomalías en signos vitales.

5.4. Sector Manufactura e Industria

- **Control de Calidad:** Detección de defectos en productos mediante análisis de sensores o imágenes.
- **Mantenimiento Predictivo:** Anticipación de fallos en maquinaria basada en datos operativos (sensores, logs).
- **Optimización de Procesos:** Identificación de cuellos de botella o inefficiencias.
- **Detección de Anomalías en Producción:** Identificación de desviaciones en parámetros de proceso.

7. Recursos y Herramientas Comunes

El ecosistema de herramientas para minería de datos es amplio y dinámico:

7.1. Algunas de las Bibliotecas Principales (Python)

- **Pandas / NumPy / SciPy:** Manipulación de datos y computación científica.
- **Matplotlib / Seaborn / Plotly:** Visualización de datos.
- **Scikit-learn:** Biblioteca fundamental para ML clásico y preprocesamiento.
- **PyTorch:** Biblioteca líder en Deep Learning, conocida por su flexibilidad.
- **TensorFlow & Keras:** Ecosistema popular para Deep Learning.
- **XGBoost / LightGBM / CatBoost:** Implementaciones eficientes de Gradient Boosting.

7.2. Plataformas de Desarrollo y Cloud

- **Jupyter Notebooks / Google Colab:** Entornos interactivos para desarrollo y experimentación.
- **Kaggle:** Plataforma para competiciones, datasets y notebooks públicos.
- **Marimo:** Notebooks de alto rendimiento y deployment.

¡Gracias!

Prof. Rodrigo Kataishi, Ph.D.
rkataishi@untdf.edu.ar

Maestría en Minería de Datos
UTN - Universidad Tecnológica Nacional, Rosario