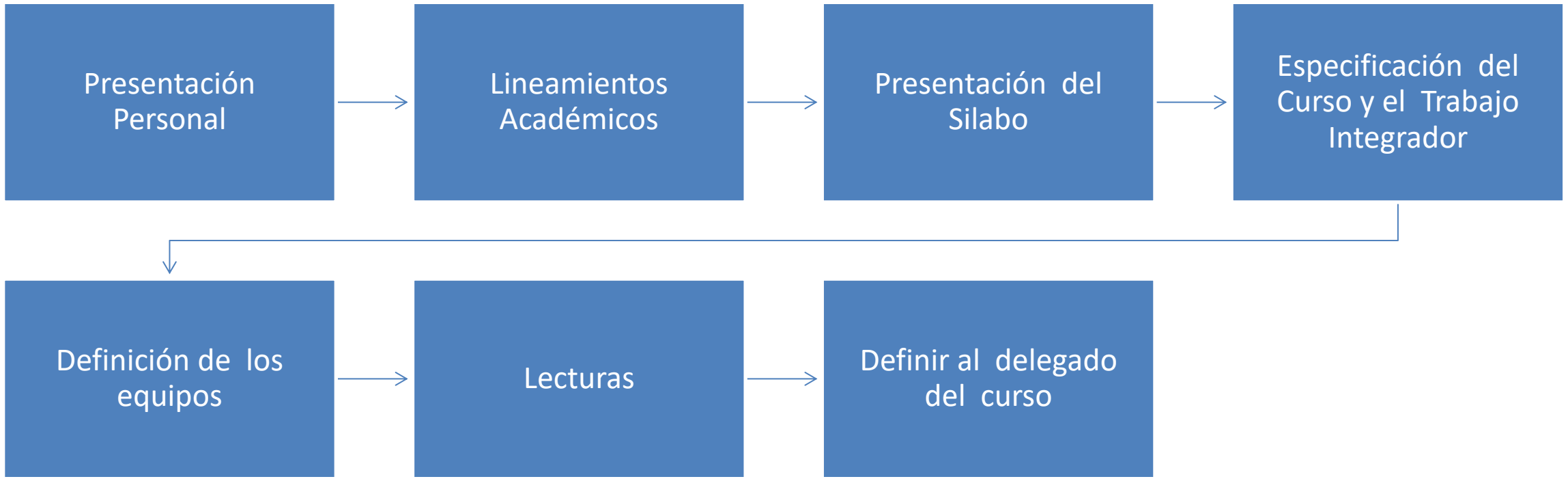
Minería de Datos

SESIÓN 1


Presentación Personal
Lineamientos Académicos
Presentación del Silabo
Especificación del Curso y el Trabajo Integrador
Definición de los equipos
Lecturas
Definir al delegado del curso

Presentación Personal

Expectativa del curso de Minería de Datos (Aprendizaje automático)

Conocimiento en herramientas Software (Python nivel)

PEDRO MARTIN LEZAMA GONZALES

DR, MENG, Ingeniero de Sistemas, PMP, PMI-ACP, SAMC, SPOC, SMC, SSGB

EXPERIENCIA PROFESIONAL

Experto en BI y Analítica, proyectos en el MEF, COPEINCA, ABACO, STEFANINI, BANCO DE LA NACION, RENIEC.

Sistemas OLTP. Edelnor, Inca Kola Mail:

pedrolezamagonzales@gmail.com Mob: 9-45473135

Lineamientos Académicos:

- Lista – asistencia
- Dentro de un marco de respeto.

Presentación del Sílabo

Especificación del Curso y el Trabajo Integrador

1.- Aplica el análisis crítico y pensamiento creativo para la identificación y solución de problemas empresariales con minería de datos.

2.- Aplicación práctica en un caso de negocio real, (datos abiertos)

Definición de los Equipos

- Examen Parcial: Definición de los Equipos de trabajo (x integrantes)

Lecturas

Papers y Libros de lectura obligatoria, ver el detalle en el sílabo

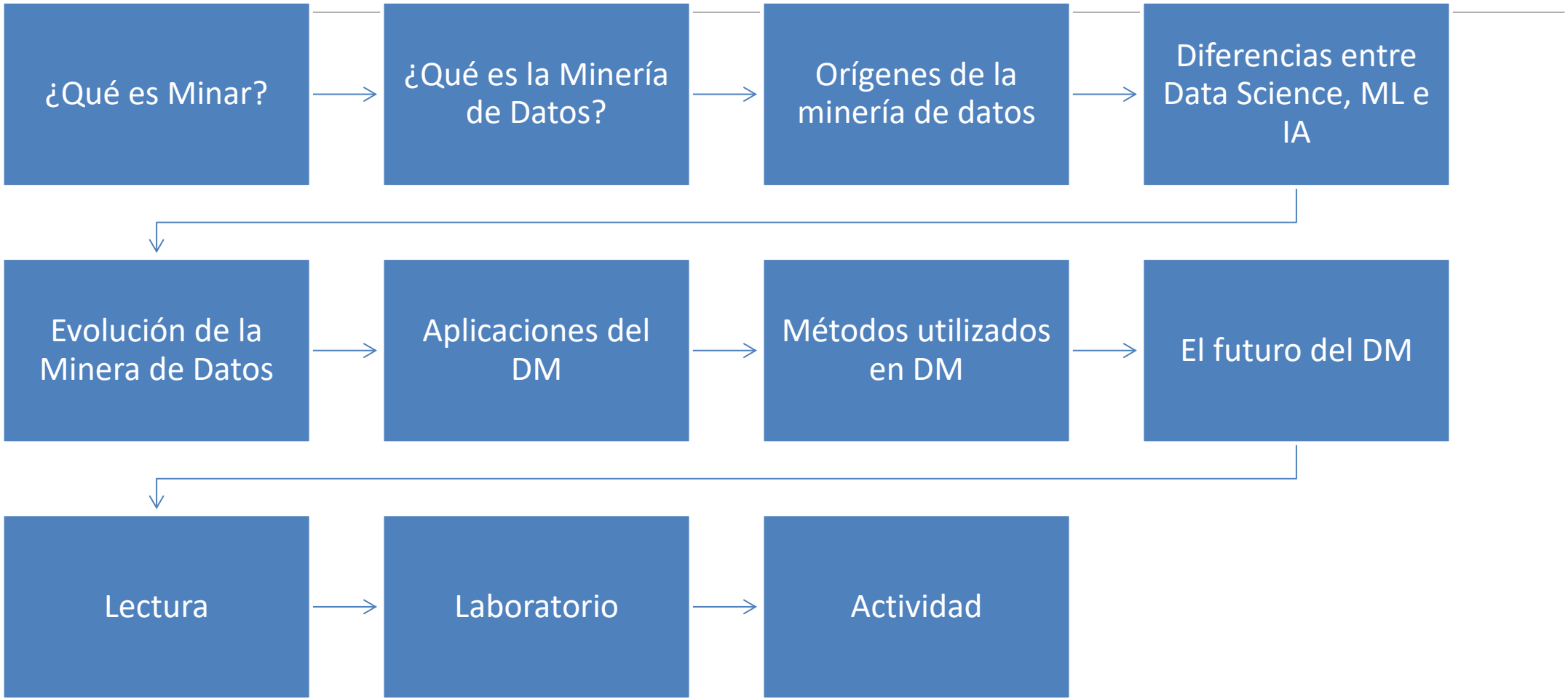
Definir al delegado del curso

Delegado del curso

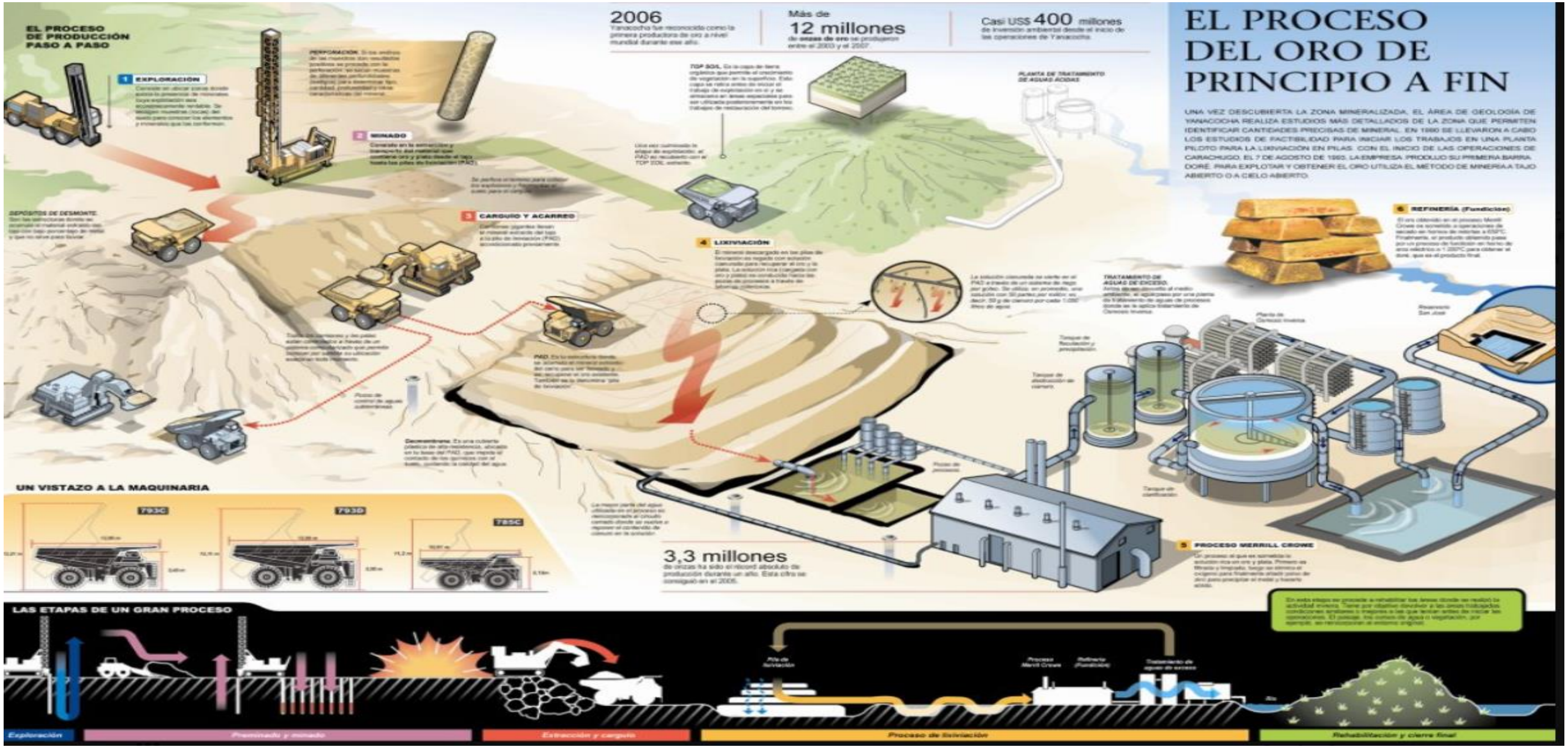
Temas a Tratar

Sem	Contenido	Actividades
1	Fundamento y Conceptos de Minería de Datos	• **Exposición:** El profesor explica los Fundamento y Conceptos de Minería de Datos • **Lecturas:** Using data mining to improve customer engagement: A case study https://medium.com/@manishbhujel22/using-data-mining-to-improve-customer-engagement-a-case-study-633e6bfd334f • **Programación:** Evaluación preparatoria. Propuesta de la mejora del caso del laboratorio, parte 1.

Temas a Tratar



¿Qué es Minar?


EL PROCESO DEL ORO DE PRINCIPIO A FIN
EL PROCESO DE PRODUCCIÓN PASO A PASO
2006
Más de 12 millones
Casi US$ 400 millones
3,3 millones
UN VISTAZO A LA MAQUINARIA
LAS ETAPAS DE UN GRAN PROCESO

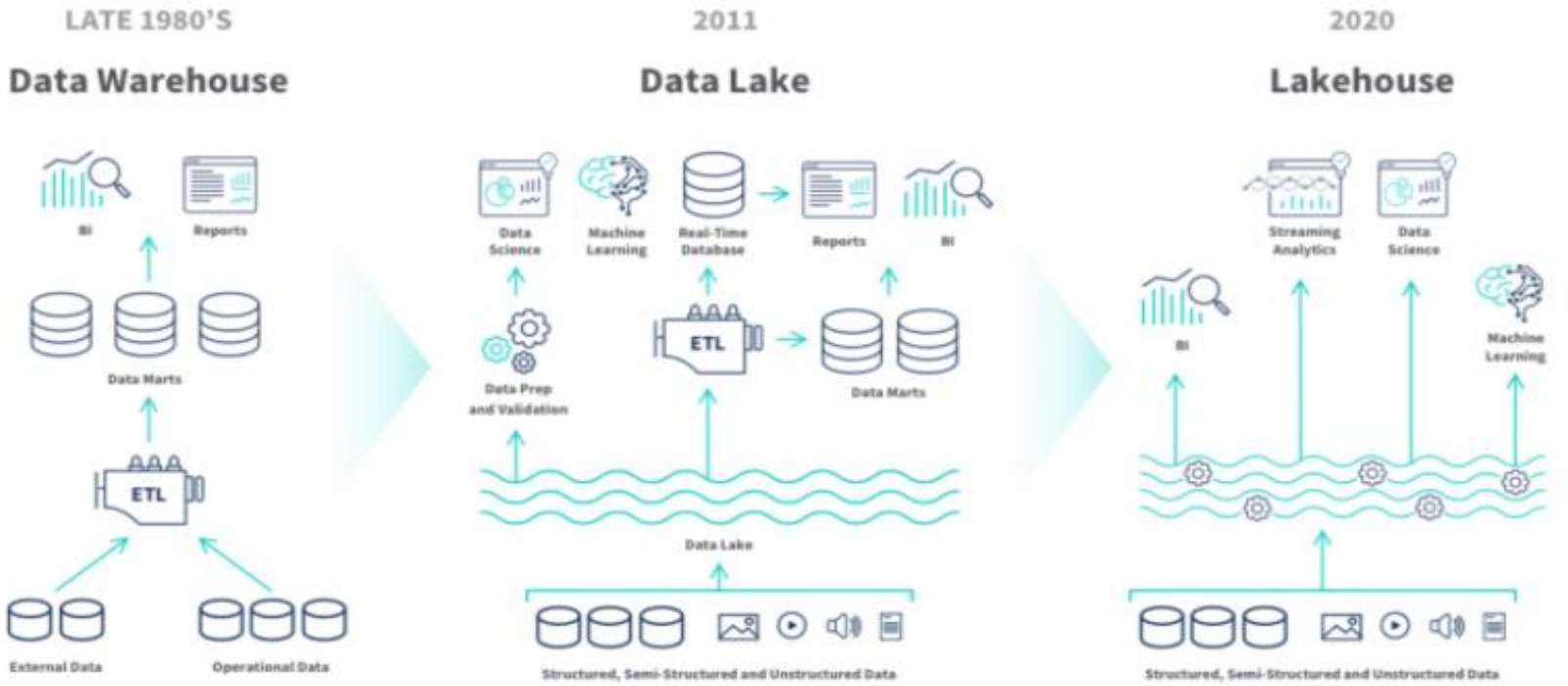
¿Qué es Minar?

RAE

"Hacer grandes diligencias para conseguir algo"

¿Qué es la Minería de Datos?

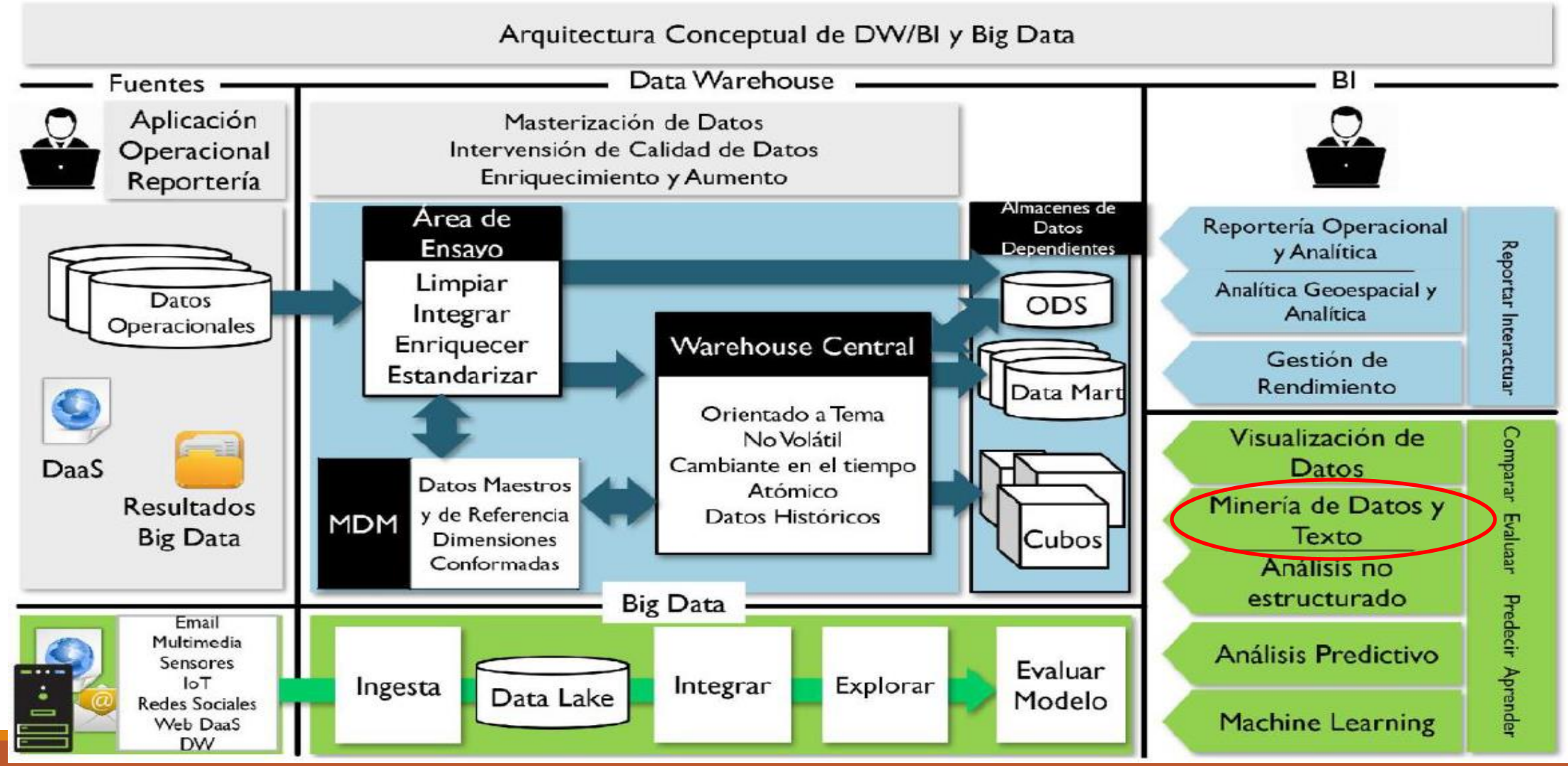
*"Descubrir **automáticamente** información **útil** (información de **valor**), en **grandes repositorios** de datos"*



¿Qué es la Minería de Datos?

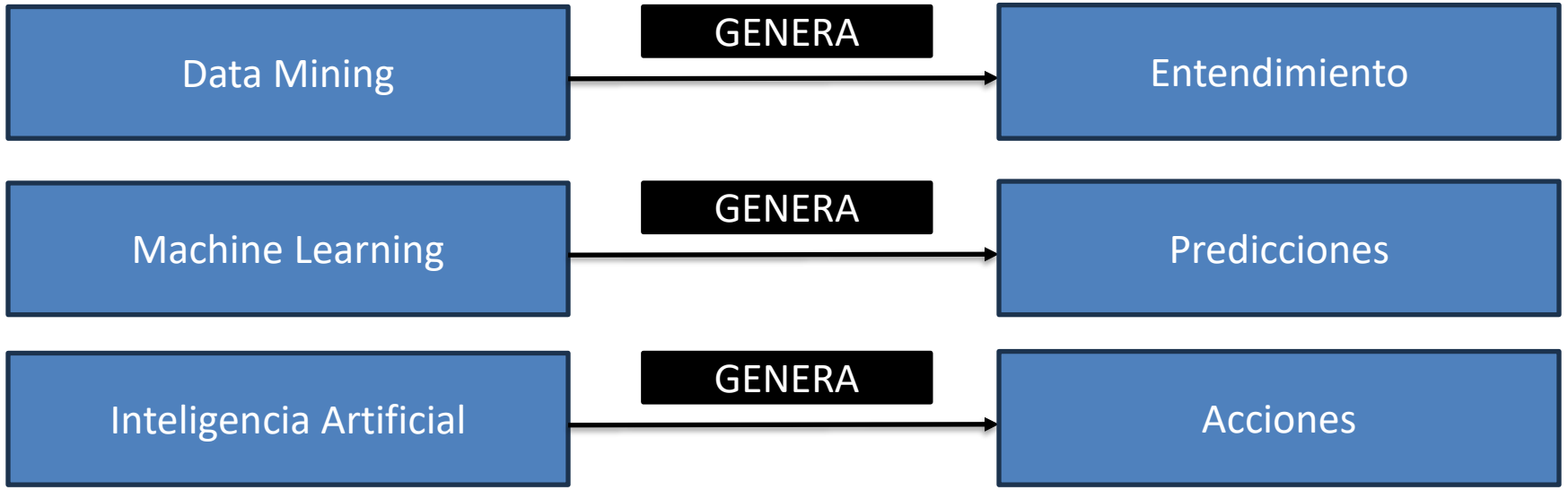
La minería de datos es un ***proceso que implica descubrir patrones, tendencias, relaciones y conocimientos útiles*** a partir de grandes conjuntos de datos y complejos.
Se basa en el ***análisis de datos*** para extraer ***información valiosa*** y ***tomar decisiones informadas.***

DM y su relación con DW/BI y Big Data



Diferencias entre Data Science, Machine Learning e Inteligencia Artificial

- Están de moda, pero no son los mismo, ni son intercambiables.
- **Data Science** es el nombre reciente para algo mucho mas antiguo: **Data Mining (90's).**
- **Definición (sobre) simplista:**



Diferencias entre Data Science, Machine Learning e Inteligencia Artificial

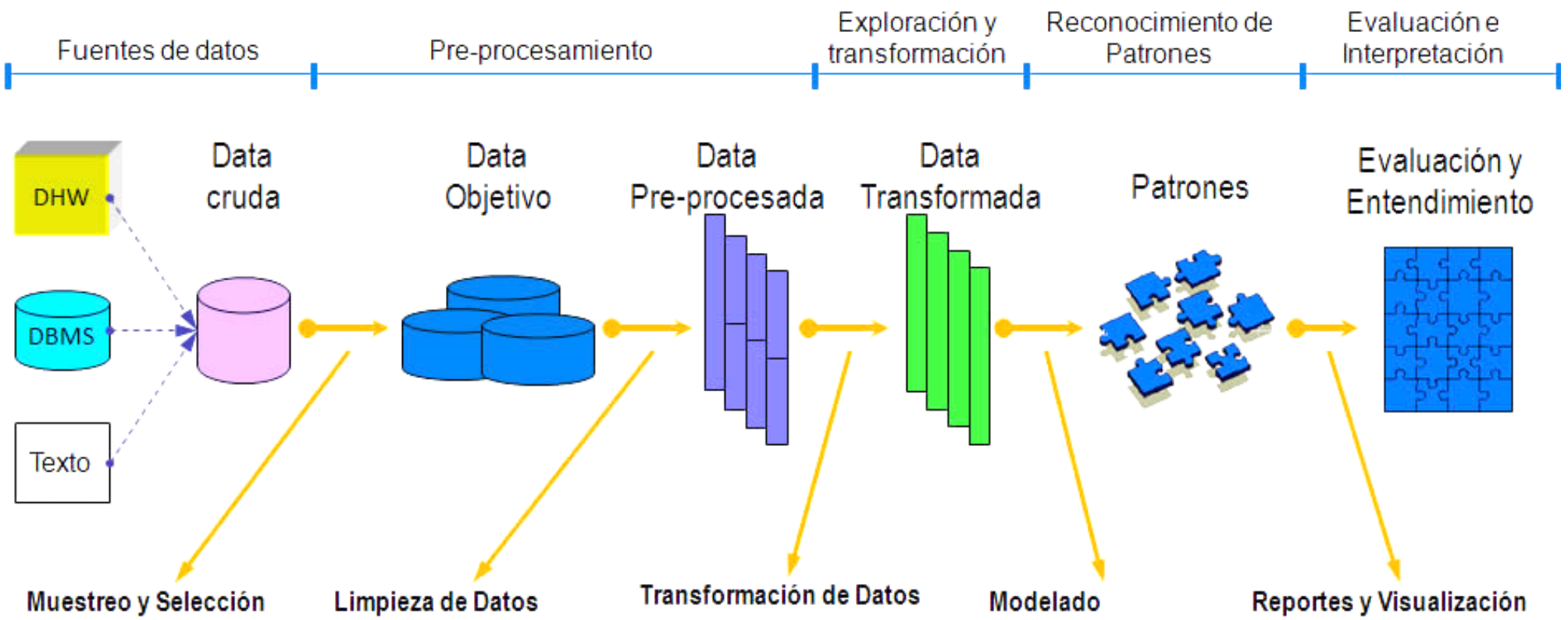
- Inteligencia Artificial: auto reconoce una señal de STOP y toma la ***acción*** de frenar.
- Machine Learning: Señala STOP usando cámaras y ***predice*** en base a entrenamiento.
- Data Mining: Auto transita por las calles y evaluamos que su rendimiento no es el esperado. Luego ***entendemos*** que se debe a varios factores externos

¿Por que es importante entender estas diferencias?

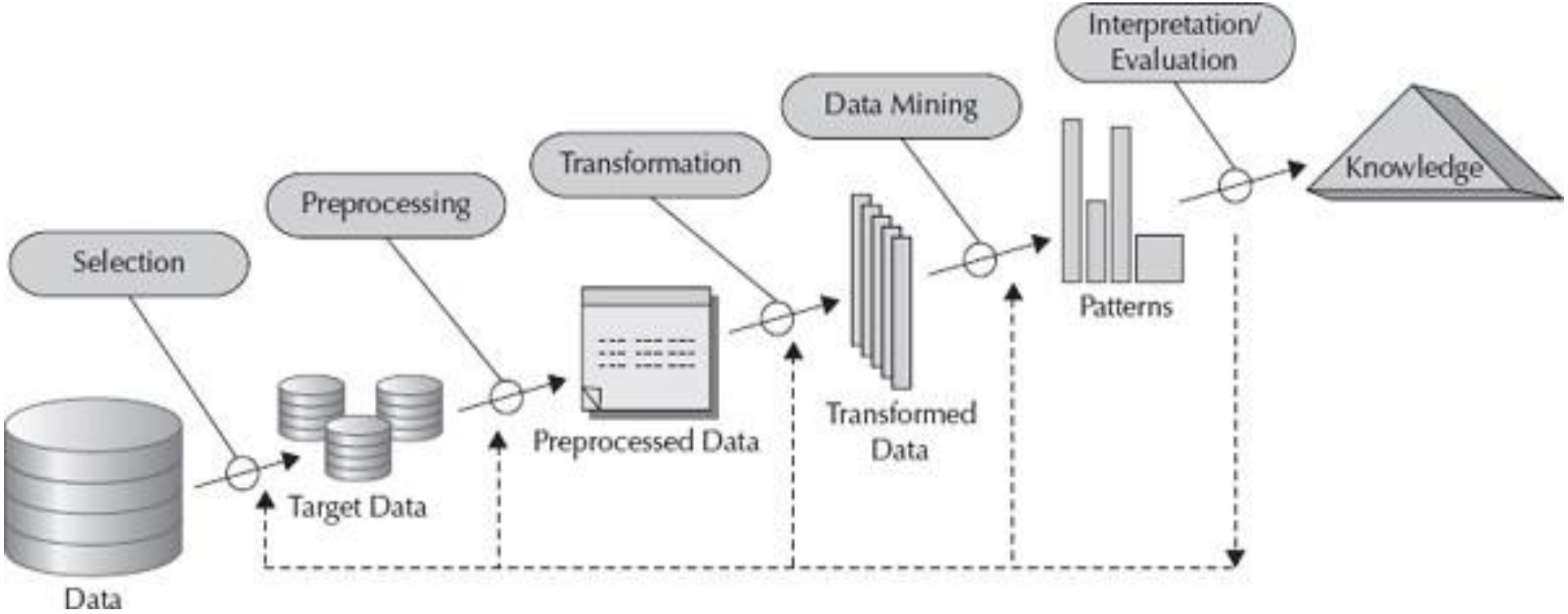
- ***Porque este no es un curso de ML, es un curso de DM***
- **ML:** ***Estudia, diseña y desarrolla algoritmos*** que permiten a las ***computadoras aprender*** sin ser explícitamente programados (Arthur Samuel). ***Técnicas genéricas, aplicables a varios dominios.***

- **DM:** El enfoque está en ***extraer conocimiento, o patrones previamente desconocidos, a partir de grandes volúmenes de datos.*** Para esto se pueden ***utilizar técnicas de ML, entre otras.*** Requiere conocimiento de los datos mismos y su dominio

Proceso de minería de datos



KDD (Descubrimiento de Conocimiento en Bases de Datos)



Orígenes de la minería de datos

Extrae ideas del ***aprendizaje automático ML/IA, reconocimiento de patrones, estadísticas y sistemas de bases de datos***

Debe atender:

- **Enormidad de datos**
- **Alta dimensionalidad de los datos**
- **Naturaleza heterogénea y distribuida de los datos**

Evolución de la Minera de Datos

1960 - Colección de Datos

- ***Los estadísticos manejaban temas como data fishing, data mining, o data archaelogy con la idea de encontrar correlaciones sin una hipótesis previa en base de datos con ruido.***

1989 Descubrimiento de conocimientos en bases de datos

- El término "Descubrimiento de conocimiento en bases de datos" ***(KDD) es acuñado por Gregory Piatetsky-Shapiro***. También en este momento cofunda el primer taller también llamado KDD

1990 Aparición del termino "Minería de datos" en B.D.

- El término ***"minería de datos" apareció en la comunidad de la base de datos***. Las empresas minoristas y la comunidad financiera están utilizando la minería de datos para ***analizar datos y reconocer las tendencias para aumentar su base de clientes, predecir las fluctuaciones en las tasas de interés, los precios de las acciones y la demanda de los clientes.***

2001 Introducción del término "Ciencia de datos" como un disciplina independiente.

- Aunque el término ciencia de los datos ha existido desde la década de 1960, no fue hasta 2001 que William S. Cleveland lo introdujo como una disciplina independiente. ***Según Build Data Science Teams, DJ Patil y Jeff Hammerbacher utilizaron el término para describir sus roles en LinkedIn y Facebook.***

2015 Minería de datos extendida:

- ***DJ Patil se convirtió en el primer científico jefe de datos en la Casa Blanca. Hoy en día, la minería de datos está muy extendida en los negocios, la ciencia, la ingeniería y la medicina, por nombrar solo algunos. La minería de las transacciones de tarjetas de crédito, los movimientos bursátiles, la seguridad nacional, la secuenciación del genoma y los ensayos clínicos son solo la punta del iceberg para las aplicaciones de minería de datos.***

Aplicaciones del DM

En Internet	El Mundo de los Negocios	En Mundo de la Ciencias
• **E-bussines:** Perfiles de clientes, publicidad dirigida, fraude. • **Buscadores Inteligentes:** Generación de jerarquías, bases de conocimiento web. • **Gestión del Tráfico de la Red:** Control de errores.	• **Banca:** Grupos de clientes, préstamos, oferta de productos. • **Compañías de Seguros:** Detección de fraude, administración de recursos. • **Marketing:** Publicidad dirigida, estudios de competencia.	• **Meteorología:** Teleconexiones (asociaciones espaciales), predicción. • **Física:** Altas energías, datos de colisiones de partículas (búsqueda de patrones). • **Bio-Informática:** Búsqueda de patrones en ADN, proyectos científicos como genoma humano, datos geofísicos, altas energías.

Casos de Éxito

35% de ventas anuales debido a la recomendación de productos

El 85% del consumo en Netflix es generado a través de big data y machine learning, lo que contribuye significativamente a su alta tasa de retención de usuarios del 93%

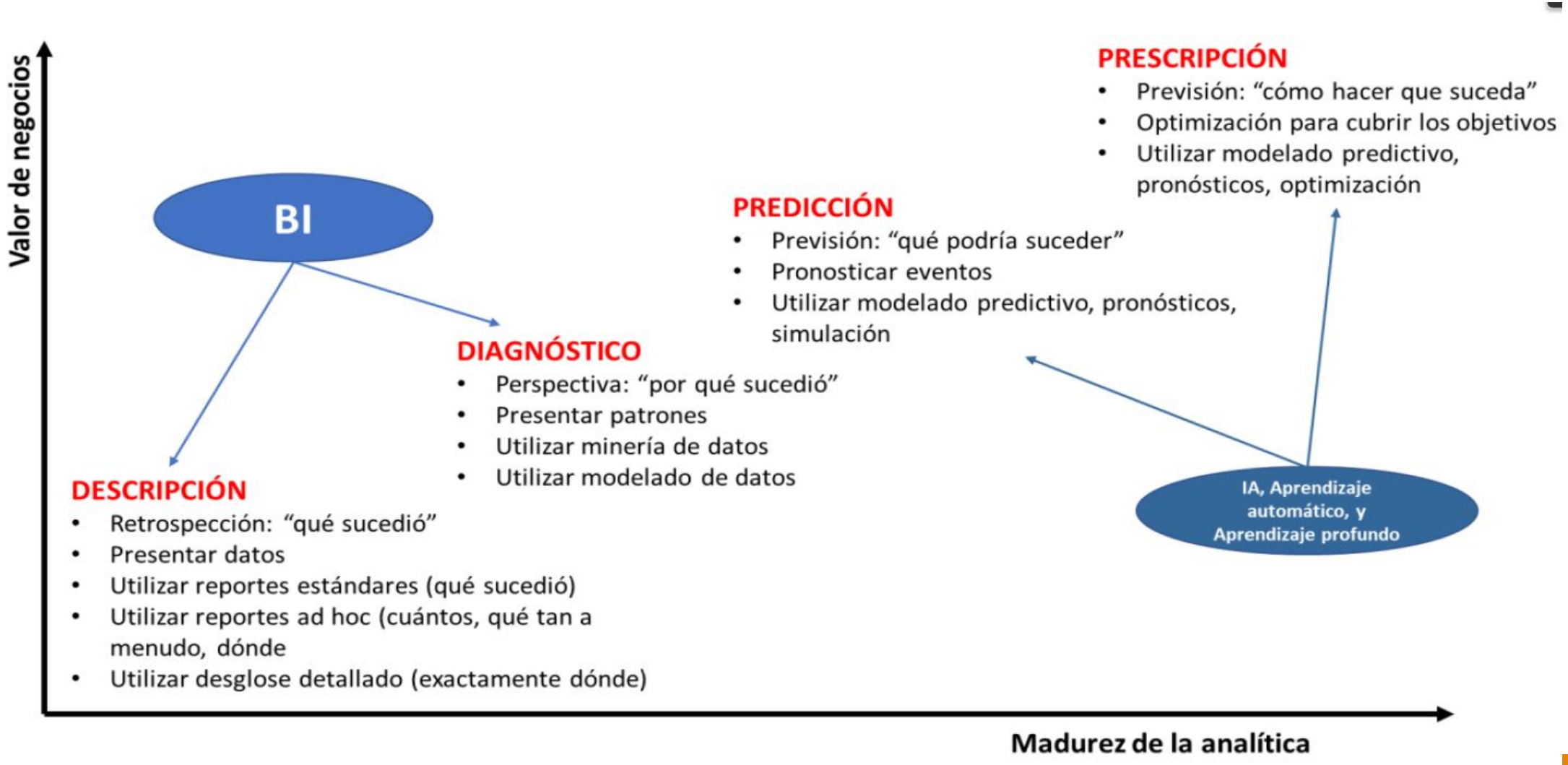
Analiza la información geográfica, demográfica y el tráfico, esta empresa logra hacer estimados sobre el posible éxito de una nueva tienda; es decir, usan el Big Data para elegir sus nuevas ubicaciones.

Métodos utilizados en DM

- **Métodos descriptivos:** Encontrar patrones interpretables por humanos que permitan describir los datos.

- **Métodos predictivos:** Usar variables para predecir variables desconocidas o valores futuros de otra variables

Valor de negocios y madurez analítica



Métodos utilizados en DM

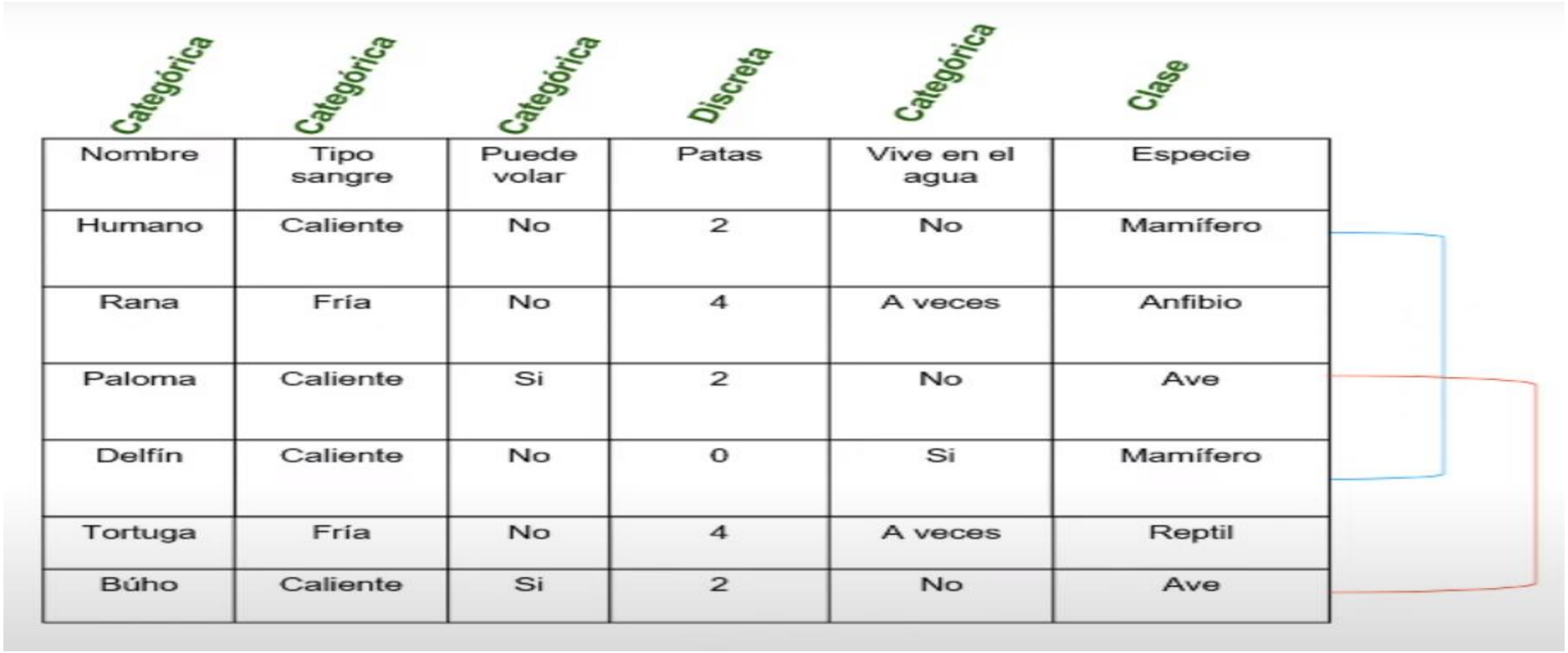
- Clustering (Descriptivo)
- Descubrimiento de Reglas de asociación (Descriptivo)
- Descubrimiento de Patrones Secuenciales (Descriptivo)
- Clasificación (Predictivo)
- Regresión (Predictivo)
- Detección de Desviación (Predictivo)

Clustering (Descriptivo)

- Conjunto de puntos(datos), cada uno con un set de atributos y una medida de similitud

- Encontrar conjunto tales que
  - Puntos en un cluster sean mas similares entre si
  - Puntos en conjuntos diferentes sean menos similares entre si
  - **Se crean grupos de elementos similares**

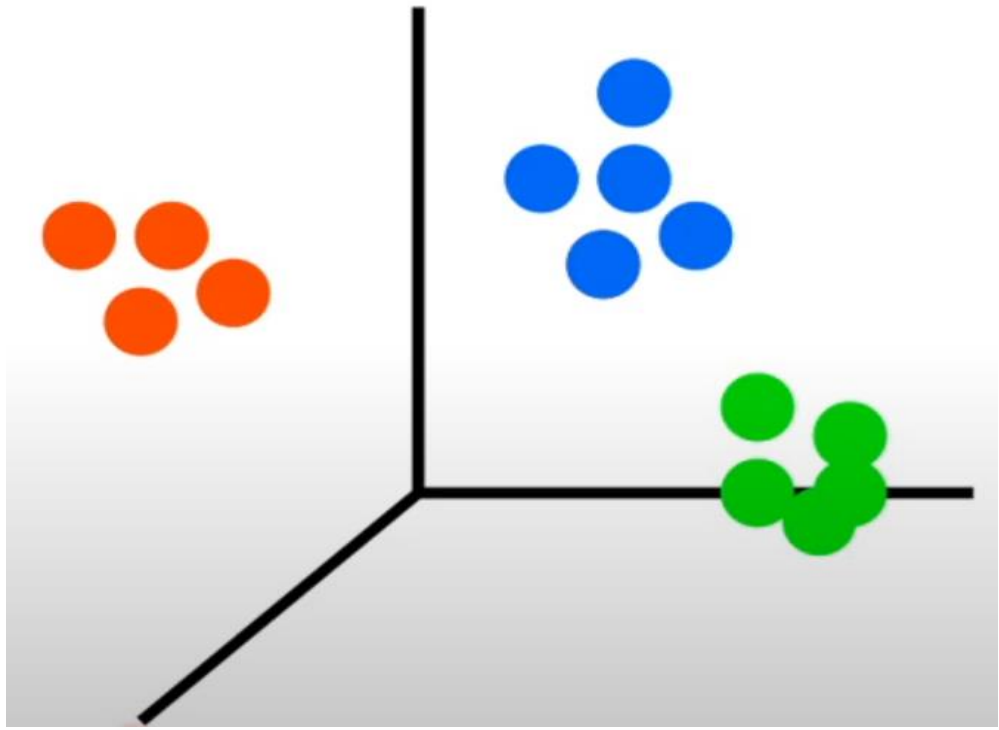
Clustering (Descriptivo)



Categórica	Categórica	Categórica	Discreta	Categórica	Clase
Nombre	Tipo sangre	Puede volar	Patas	Vive en el agua	Especie
Humano	Caliente	No	2	No	Mamífero
Rana	Fría	No	4	A veces	Anfibio
Paloma	Caliente	Si	2	No	Ave
Delfín	Caliente	No	0	Si	Mamífero
Tortuga	Fría	No	4	A veces	Reptil
Búho	Caliente	Si	2	No	Ave

Visualización de clustering (Descriptivo)

- Clustering 3D basado en distancias Euclidianas
- Distancia intra-cluster es minimizada
- Distancia inter-cluster es maximizada



Clustering – Aplicación 1 (Descriptivo)

- Segmentación de mercado
  - Meta: Subdividir un mercado en subconjuntos de clientes en donde cualquier conjunto es un potencial objetivo de marketing (ej: Netflix, Amazon)
  - ¿Cómo?
  - Edad, Fecha de Nacimiento, grado académico ----> productos que consumen.

Clustering – Aplicación 2 (Descriptivo)

- Clustering de documentos
  - Meta: Encontrar grupos de documentos que son similares entre sí, basándose en las palabras más importantes que contienen. (Directorios, Wiki)
  - ¿Cómo?
  - Agrupando artículos por ejemplo en políticos, deportes, entretenimiento entre otros

Clustering – Aplicación 2 (Ejemplo) (Descriptivo)

- Clustering de puntos: 3204 artículos del L.A. Times
- Medida de similitud: cuantas palabras tienen en común estos documentos (después de filtrar algunas palabras).

Category	Total Articles	Correctly Placed
Financial	555	364
Foreign	341	260
National	273	36
Metro	943	746
Sports	738	573
Entertainment	354	278

Reglas de Asociación (Descriptivo)

- Dado un conjunto de records, cada uno contiene un número de elementos de una colección determinada
- Objetivo: Producir reglas de dependencia que predecirán la ocurrencia de un elemento (item) basándose en ocurrencias de otros ítems

Reglas de Asociación

TID	Items
1	Pan, Coca-cola, Pañales, Leche
2	Cerveza, Pan
3	Cerveza, Coca-cola, Pañales, Leche
4	Cerveza, Pan, Pañales, Leche
5	Coca-cola, Pañales, Leche

Reglas de Asociación (Descriptivo)

- Promoción de Marketing y Ventas
- Sea la regla encontrada del tipo
- {Queso,.....} ------> {PapasFritas}

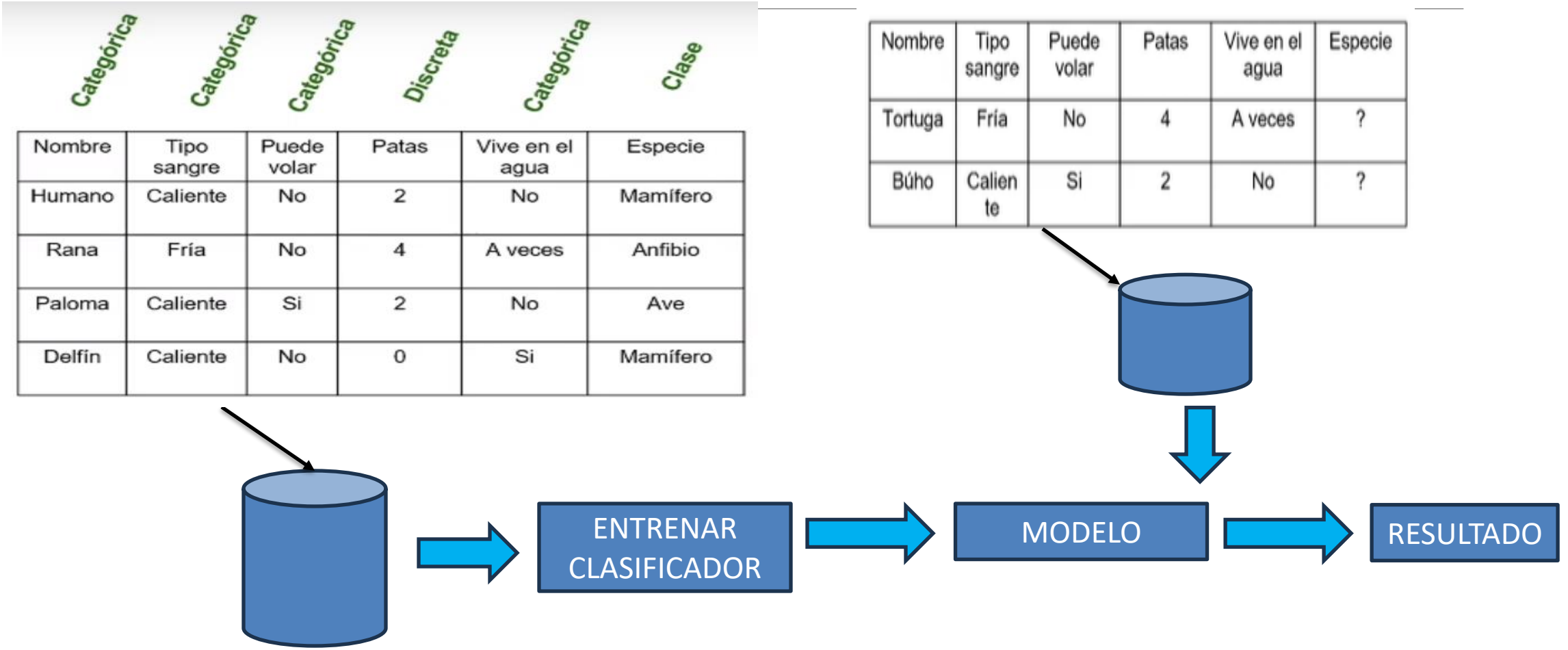
Clasificación (Predictivo)

- Set de entrenamiento (atributos incluyendo clase)
- Busca modelar en atributo clase
- Objetivo: asignar la clase mas correcta a records nuevos
- Set de Evalueación

Clasificación (Predictivo)

Categórica	Categórica	Categórica	Discreta	Categórica	Clase
Nombre	Tipo sangre	Puede volar	Patas	Vive en el agua	Especie
Humano	Caliente	No	2	No	Mamífero
Rana	Fría	No	4	A veces	Anfibio
Paloma	Caliente	Si	2	No	Ave
Delfín	Caliente	No	0	Si	Mamífero

Nombre	Tipo sangre	Puede volar	Patas	Vive en el agua	Especie
Tortuga	Fría	No	4	A veces	?
Búho	Caliente	Si	2	No	?



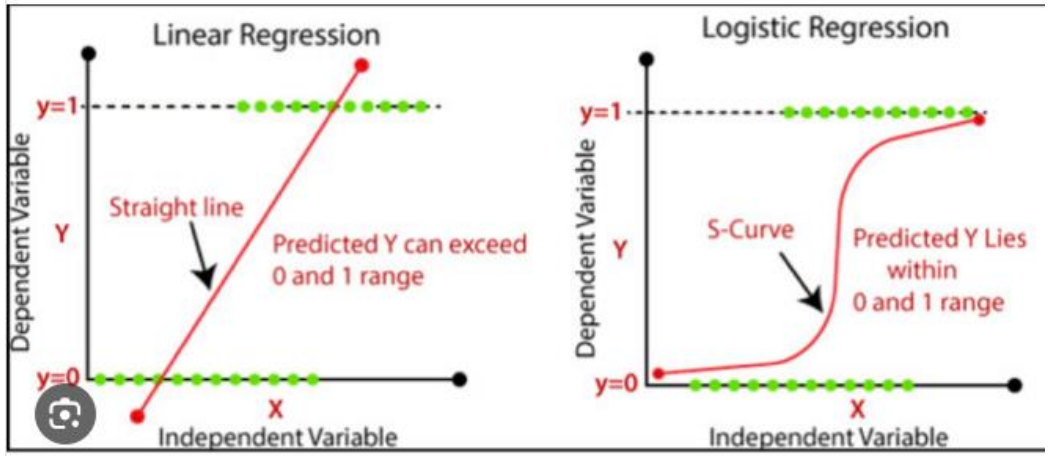
Clasificación (Predictivo)

- Marketing Directo
- Meta: Reducir costos de publicidad apuntando directamente a potenciales compradores
- ¿Cómo?
- Ejemplo: Relación de ventas de productos entre compradores o clientes (Amazon, Netflix)

Regresión (Predictivo)

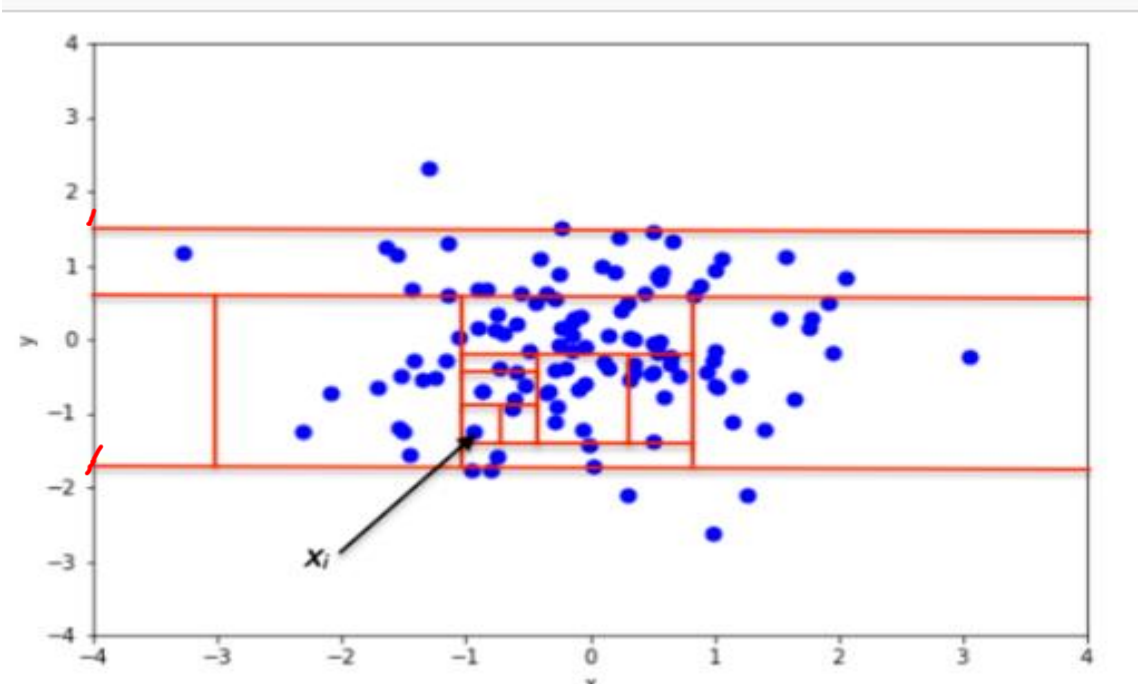
- Predecir el valor de una variables continúa, en base a valores de otras variables, asumiendo modelos de dependencia lineal o no-lineal

- Estadística y Redes Neuronales

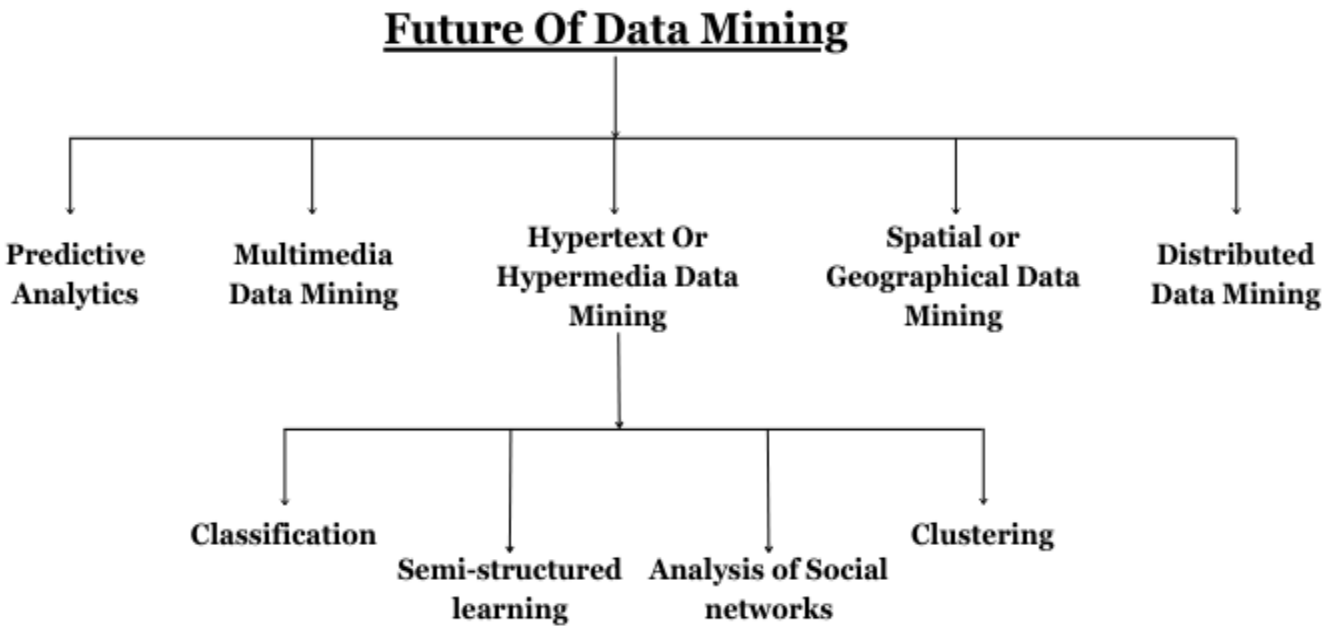


Detección de desviación/anomalía (Predictivo)

- Detectar desviaciones significativas de los valores normales



El futuro del DM



https://www.bizprospex.com/understanding-data-mining-with-the-help-of-case-studies-on-data-mining-in-market-analysis/

Laboratorio

- Práctica Introductoria (https://forms.gle/gQF9YfGsJfTPGiSa6)
- Evaluación preparatoria (Caso Covid)
- Evaluación preparatoria – Propuesta de la mejora del caso del laboratorio

Lectura

Using data mining to improve customer engagement: A case study
https://medium.com/@manishbhujel22/using-data-mining-to-improve-customer-engagement-a-case-study-633e6bfd334f

Actividades Realizadas

- Revisión del Sílabo
- Desarrollo de la clase
- Práctica de bienvenida
- Lectura: Using data mining to improve customer engagement: A case study
- Laboratorio: Evaluación preparatoria (Caso Covid)
- Actividad: Evaluación preparatoria – Propuesta de la mejora del caso del laboratorio

Gracias