

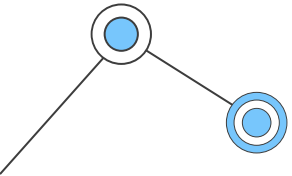
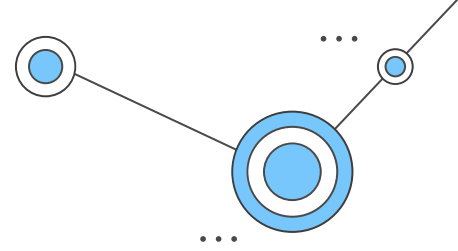
# Introducción a la minería de datos



**UNIVERSIDAD  
CATÓLICA**  
DE CÓRDOBA  
JESUITAS

Dr. Francisco Arduh  
2023

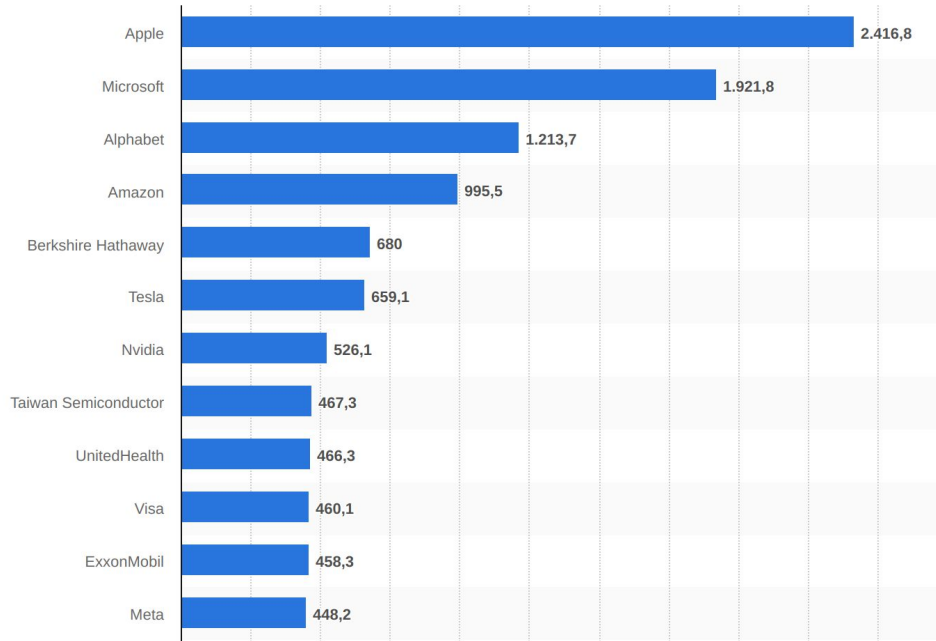
“Los datos son el nuevo petróleo”



# “Los datos son el nuevo petróleo”



# “Los datos son el nuevo petróleo”



Fuente: <https://es.statista.com/>

# Pirámide D-I-K-W

Sabiduría (toma de decisiones)

Conocimiento (reglas)

Información (estructura)

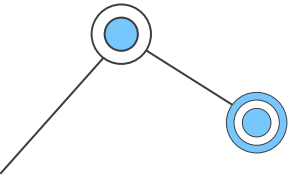
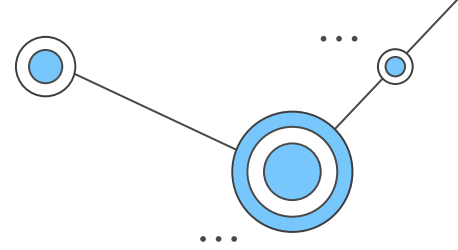
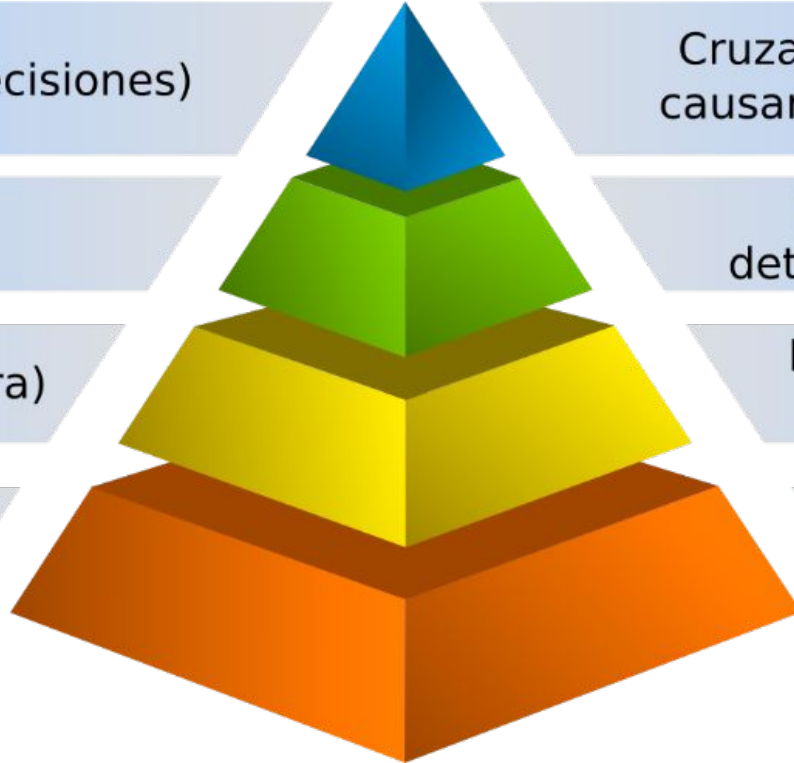
Datos (mediciones)

Cruzar con luz roja puede  
causar accidentes y multas

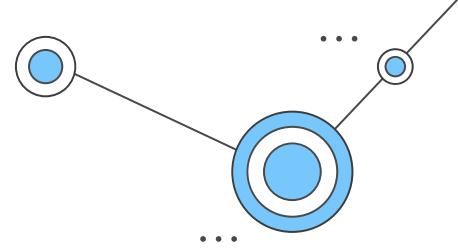
El semáforo indica  
detener el cruce de calle

La luz en el semaforo al  
frente cambió a rojo

Luz roja  
#FF0000

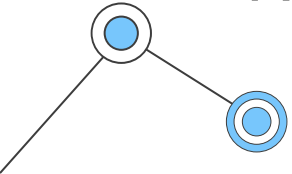


# Nuevas necesidades

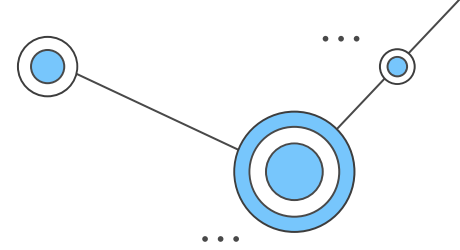


Convertir datos en conocimiento:

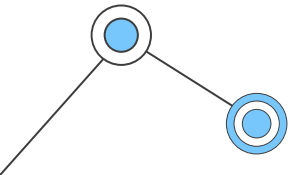
- Expertos
- Almacenes de datos -> OLAP (Online analytical process)
- Necesidad de nuevas herramientas



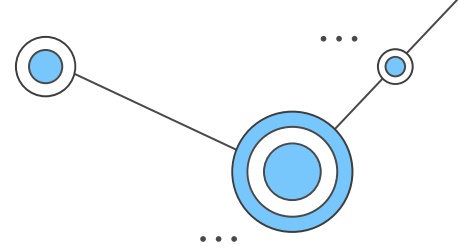
# Minería de datos



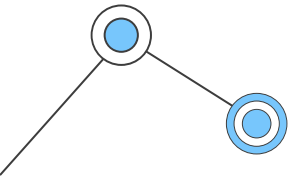
Definición: Proceso de extraer conocimiento **útil y comprensible**, **previamente desconocido**, desde grandes cantidades de datos almacenados en distintos formatos.[Witten & Frank 2000]



# Retos de la minería de datos

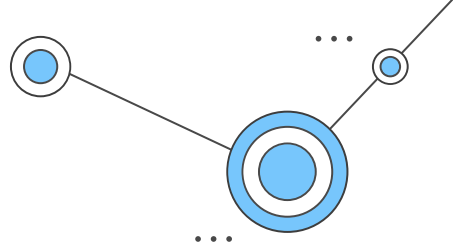


- Trabajar con grandes volúmenes de datos.
  - Implica tratar con ruidos, datos ausentes, volatilidad de datos, etc
- Técnicas adecuadas para analizar los mismos.
  - Técnicas en función del objetivo, del volumen de datos, de la complejidad del modelo que se espere.



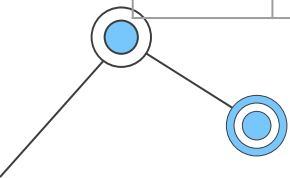


# Ejemplo

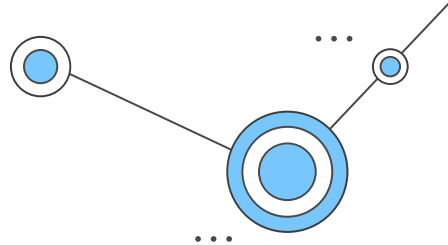


Un banco desea obtener reglas para predecir qué personas solicitan créditos y no los devuelven. Se cuenta con datos de créditos concedidos con anterioridad a sus clientes.

IDC	D-Créditos (años)	C-Créditos (euros)	Salarios (euros)	Casa propia	Cuentas morosas	...	Devuelve crédito
101	15	60000	2200	si	2	...	no
102	2	30000	3500	si	0	...	si
103	9	9000	1700	si	1	...	no
104	15	18000	1900	no	0	...	si
105	10	24000	2100	no	0	...	no
...	...	...	...	...	...	...	...

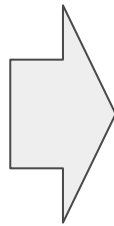


# Ejemplo

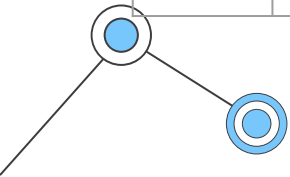




Un banco desea obtener reglas para predecir qué personas solicitan créditos y no los devuelven. Se cuenta con datos de créditos concedidos con anterioridad a sus clientes.

IDC	D-Créditos (años)	C-Créditos (euros)	Salarios (euros)	Casa propia	Cuentas morosas	...	Devuelve crédito
101	15	60000	2200	si	2	...	no
102	2	30000	3500	si	0	...	si
103	9	9000	1700	si	1	...	no
104	15	18000	1900	no	0	...	si
105	10	24000	2100	no	0	...	no
...	...	...	...	...	...	...	...



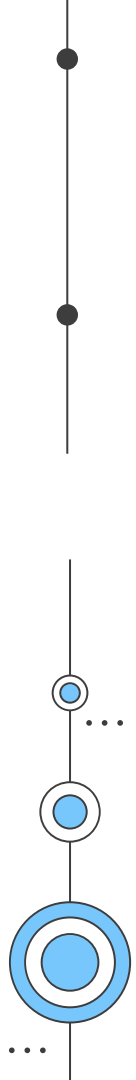
- **SI** cuentas morosas > 0 **ENTONCES** Devuelve créditos = no  
- **SI** cuentas morosas = 0 y [Salario > 2500 o (D-creditos)> 10] **ENTONCES** Devuelve crédito = si





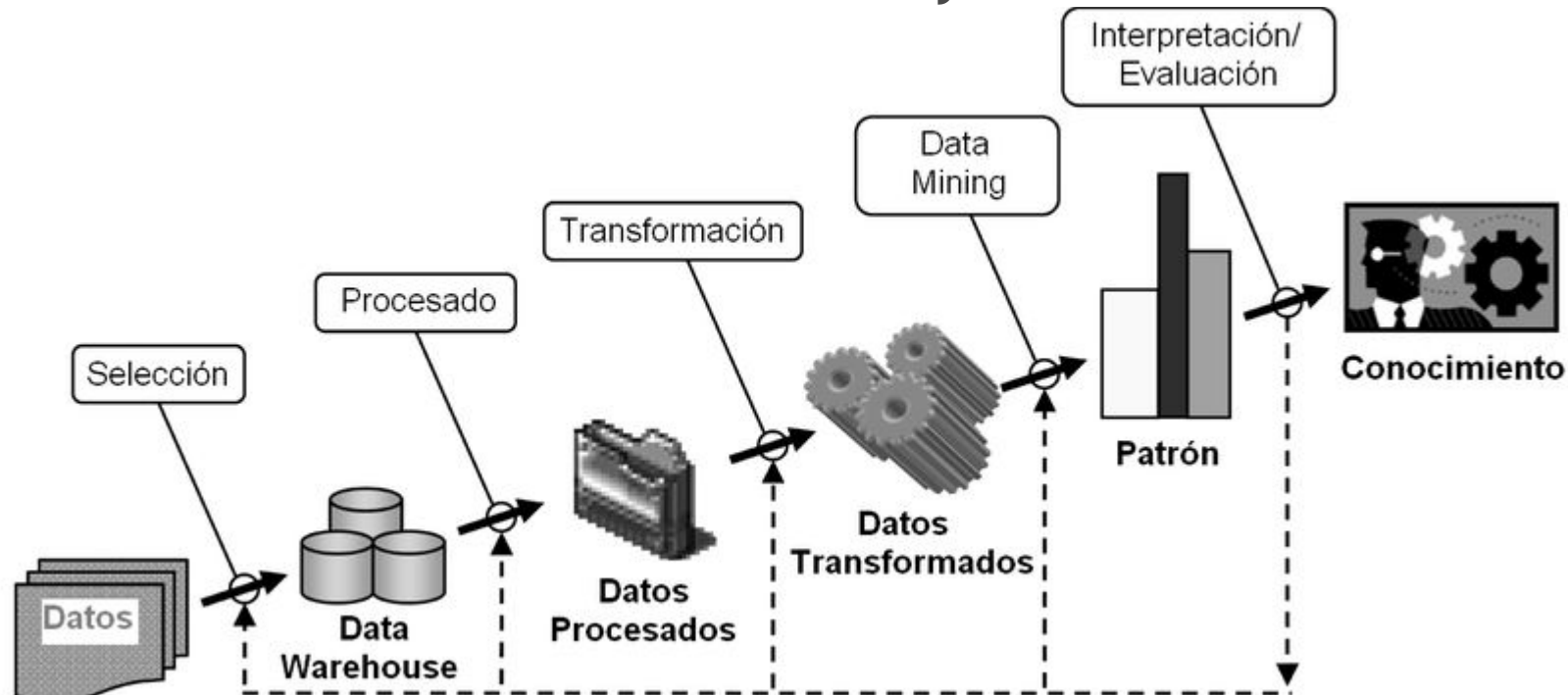
# Proceso KDD

(Proceso de descubrimiento  
de conocimiento en bases de  
datos)



# KDD

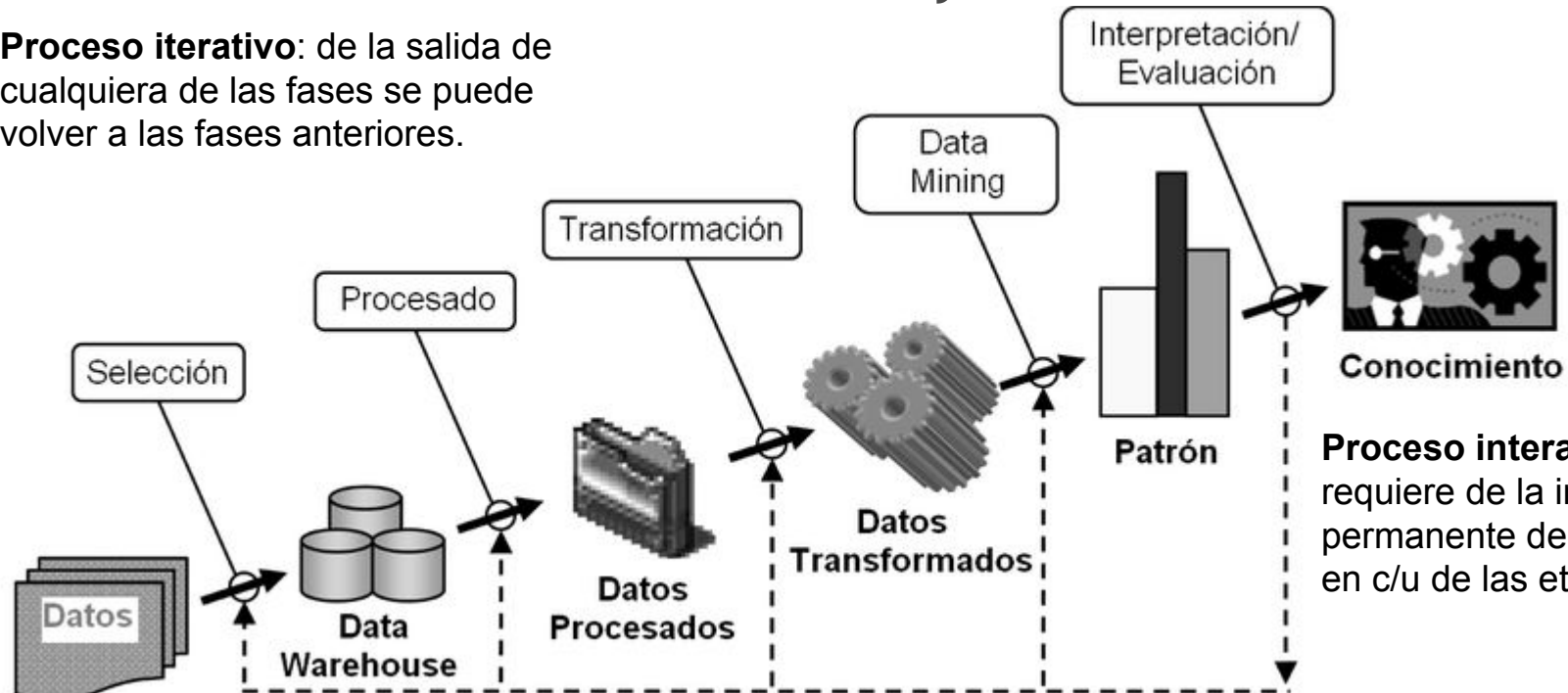
(Proceso de descubrimiento de conocimiento en bases de datos)



# KDD

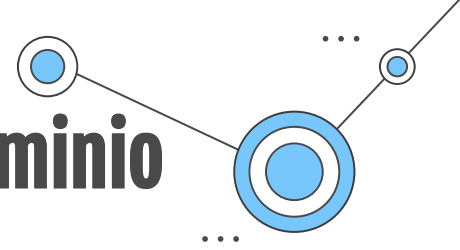
## (Proceso de descubrimiento de conocimiento en bases de datos)

**Proceso iterativo:** de la salida de cualquiera de las fases se puede volver a las fases anteriores.

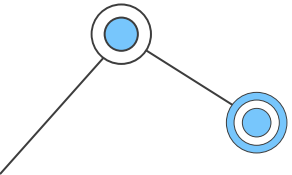


**Proceso interactivo:** se requiere de la interacción permanente del experto en c/u de las etapas.

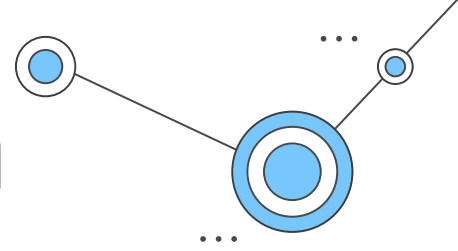
# Etapas del proceso KDD: Entender el dominio



- ¿En donde quiero aplicar el proceso de KDD?
- ¿Cuál es el problema a resolver?
- ¿Cuáles son los objetivos?

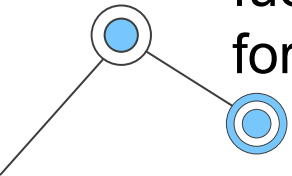


# Etapas del proceso KDD: Selección

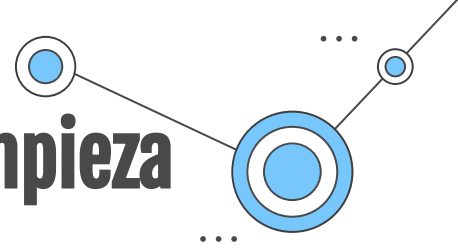


Para realizar el descubrimiento de conocimiento, es necesario contar un conjunto de datos (dataset). Este se obtiene a partir de un subconjunto de variables obtenido a través de la selección, extracción o muestreo de diversas fuentes (app, web, APIs).

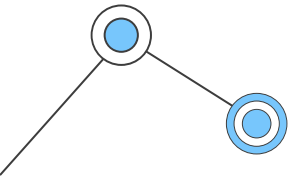
Es recomendable trabajar con un almacén de datos. Si los datos no son muy grandes se podría trabajar con la fuente de los datos directamente o con distintos formatos.



# Etapas del proceso KDD: Procesado y limpieza



- Se trata de mejorar la fiabilidad de los datos.
- Aquí se incluye limpieza de datos, tales como el manejo de los **datos faltantes** y la **eliminación de ruido o valores atípicos**.
  - Hay que ser cauto y entender sobre el dominio del problema (no todo es anomalía, no todo es ruido)



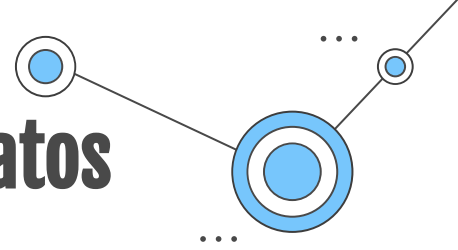




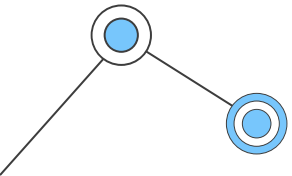
# Etapas del proceso KDD: Transformación

- **Reducción de dimensionalidad:** PCA, Correlación,  $\chi^2$ , etc.
- **Suavizados:** Discretización, Binning Methods, medias móviles, etc.
- **Agregación:** *Group By*. Ejemplo ventas diarias agrupadas en ventas mensuales o anuales. (Granularidad)
- **Generalización:** Datos de Bajo Nivel (raw data) son reemplazados por conceptos de Alto Nivel a través del concepto de jerarquía.
- **Normalización:** Escalado de los atributos para unificar dominios. Puede ser llevar a un rango: -1 a 1 ó 0 a 1, también restar la media y dividir por el desvío estándar (z-score).
- **Construcción de Atributos:** Se construyen nuevas variables que aportan mayor variabilidad favoreciendo al proceso de mining.

# Etapas del proceso KDD: Minería de datos



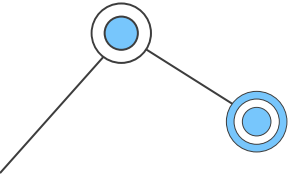
- Construir el modelo en base a los datos recopilados.
- Es necesario determinar:
  - Qué tipo de información queremos obtener: agrupación, clasificación, regresión, asociación.
  - Qué modelo/modelos utilizar, por ejemplo: árboles de decisión, redes neuronales, etc.
- Para el proceso de construcción de modelos predictivos es necesario tener bien definido los conjuntos de entrenamiento y de evaluación.



# Etapas del proceso KDD: Evaluación y interpretación

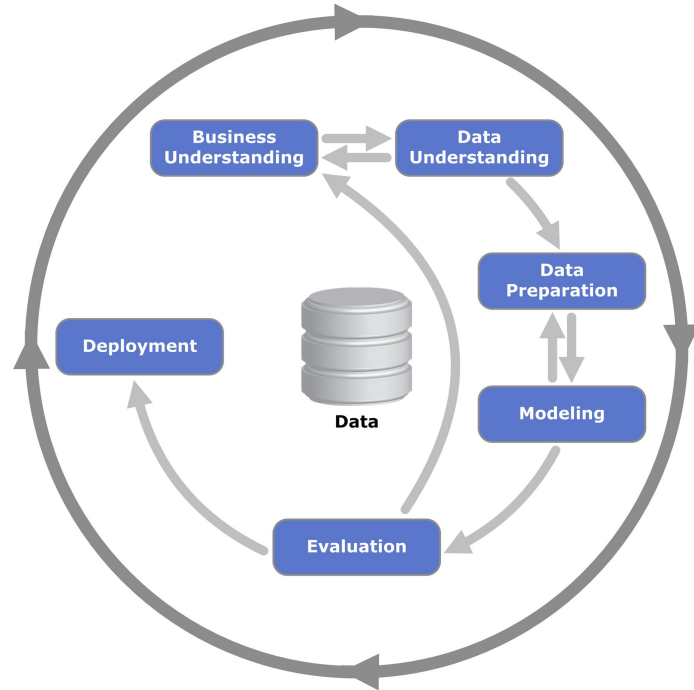


- Idealmente los patrones descubiertos deben ser:
  - Precisos, comprensibles e interesantes.
- Para la evaluación es necesaria haber hecho la separación de datos nombrada anteriormente.
- La evaluación depende del tipo de modelo qué se esté construyendo (clasificación, regresión, etc) y del objetivo del trabajo.




# CRISP-DM

## (Cross Industry Standard Process for Data Mining)



Qué es CRISP DM (Metodología de Data Mining):

<https://www.youtube.com/watch?v=UyKkSsEbXkw>



# CRISP-DM

## (Cross Industry Standard Process for Data Mining)

Comprensión del negocio	Comprensión de los datos	Preparación de los datos	Modelado	Evaluación	Despliegue
Determinar los objetivos del negocio	Recolectar los datos	Seleccionar los datos	Selección de la técnica de modelado	Evaluar resultados	Planificar despliegue
Cuadro de situación	Describir los datos	Limpiar los datos	Generar el esquema de prueba	Revisión	Planificar mantenimiento y monitoreo
Objetivos de la minería de datos	Análisis exploratorio	Construir los datos	Construir el modelo	Determinar futuros cursos de acción	Reporte final
Plan del proyecto	Verificar la calidad de los datos	Integrar los datos	Validación		Revisión
		Formateo			

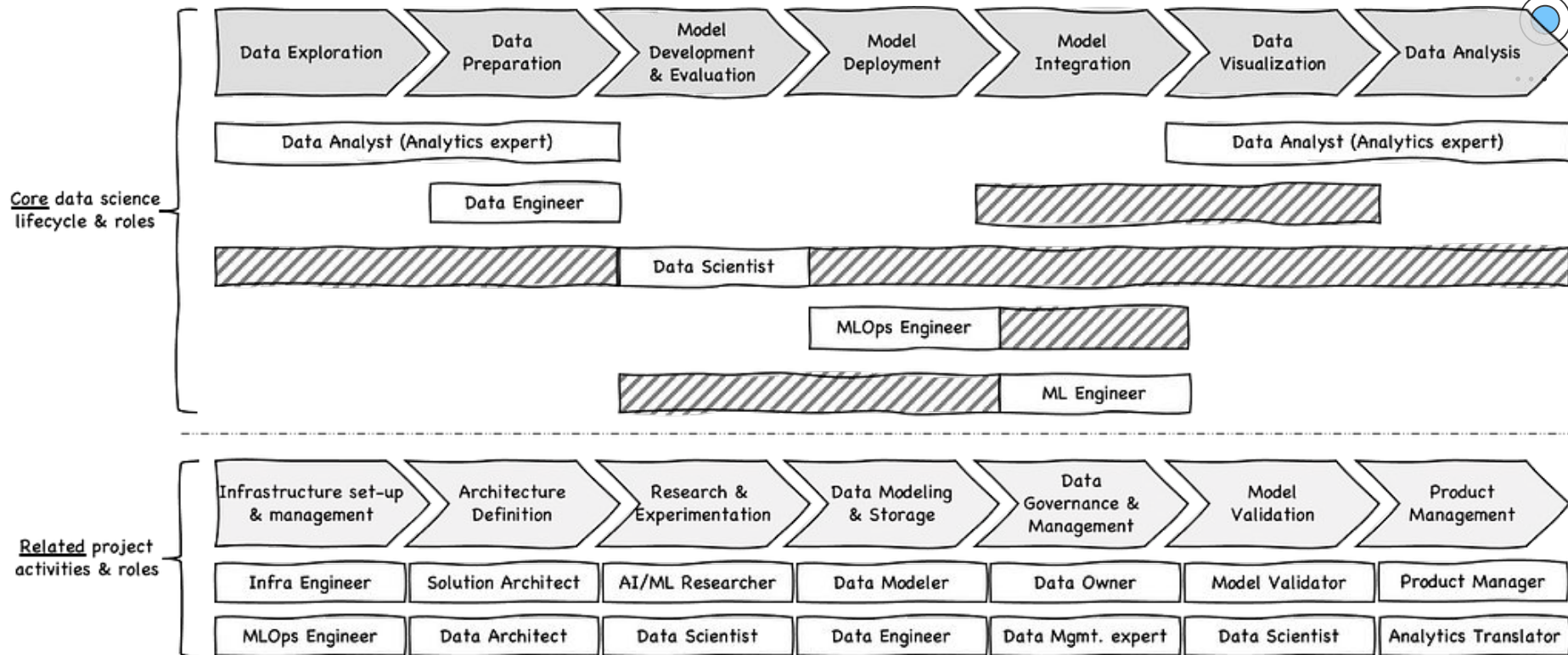


Qué es CRISP DM (Metodología de Data Mining):

<https://www.youtube.com/watch?v=UyKkSsEbXkw>

# Roles en datos

# Roles en datos





# Bibliografía principal



- “Introducción a la minería de datos” 1/e. Hernández Orallo, José/y otros
- “Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems” Aurélien Geron