

**descubrimiento de conocimiento  
a partir de bases de datos (KDD)**

# KDD Knowledge Discovery from Databases

## KDD Knowledge Discovery from Databases

- el KDD es el proceso completo de extracción del conocimiento a partir de bases de datos
- término acuñado en 1989 para enfatizar que el conocimiento es el producto final de un proceso de descubrimiento guiado por datos
- Minería de Datos es sólo **un paso** en el proceso de KDD
- Informalmente, Minería de Datos =  $\sim$  KDD

## El proceso KDD

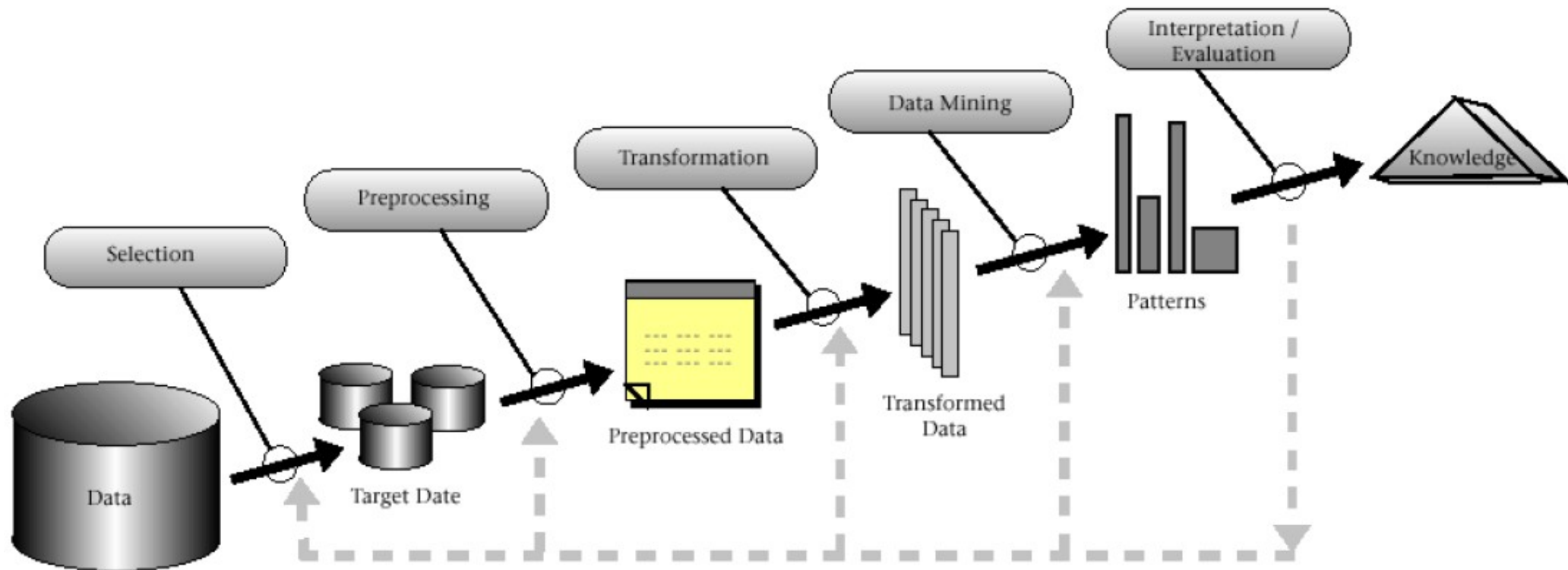
- KDD está enfocado al proceso global de descubrimiento de conocimiento a partir de bases de datos. Entre otros aspectos incluye:
  - cómo son almacenados y accedidos los datos,
  - cómo pueden escalarse los algoritmos para trabajar con cantidades de datos enormes y seguir siendo eficientes,
  - cómo pueden interpretarse y visualizarse los resultados,
  - cómo modelar y dar soporte a la interacción hombre-máquina durante todo el proceso.
- KDD hace especial énfasis en la búsqueda de modelos/patrones comprensibles
- También es importante la robustez frente a grandes conjuntos de datos ruidosos

## El proceso KDD

El proceso del KDD contiene:

- El uso de la base de datos junto con cualquier operación de selección, preprocesamiento, (sub)muestreo y transformación de la misma
- Algoritmos para obtener patrones/modelos a partir de los datos
- Evaluación del resultado de los algoritmos y selección de aquellos modelos que puedan considerarse conocimiento

# El ciclo KDD



## Etapas del proceso KDD

- 1) Comprensión del dominio de aplicación del problema, identificación de conocimiento a priori y creación del datawarehouse
- 2) Selección de datos, limpieza y preprocesamiento
- 3) Refinación de datos: reducción y proyección
- 4) Selección de la técnica de MD y aplicación de algoritmos concretos de MD
- 5) Evaluación, interpretación y presentación de los resultados obtenidos
- 6) Difusión y uso del nuevo conocimiento

## Fase 1: dominio del problema y creación del DW

- ¿Realmente es un problema adecuado para KDD?
- La familiarización con el dominio y la obtención de conocimiento a priori disminuirá el espacio de soluciones posibles => más eficiencia en el resto del proceso
- Unificación de la información en un Datawarehouse a partir:
  - Información interna: distintas BD diseñadas para trabajo transaccional y de otro tipo (hojas de cálculo, informes,...)
  - Estudios publicados (demografía, catálogos, páginas,...).
  - Otras bases de datos (compradas, industrias/empresas afines, ...)
- El resto del proceso será más cómodo si la fuente de datos está **unificada**, es **accesible** (interna) y **dedicada** (desconectada del trabajo transaccional).

## Fase 2: selección, limpieza y preprocesamiento

- A partir del resultado de la fase anterior, mediante exploración del datawarehouse y a partir de análisis y visualizaciones previas, **seleccionar el conjunto de datos adecuado** para el resto del proceso.
- **Limpieza de datos** (data cleaning), consiste en rellenar valores perdidos, identificar y/o eliminar valores anómalos (outliers), suavizar el ruido, eliminar inconsistencias (DW)
- Preprocesamiento: transformación de los datos, variables, valores, ...



## Fase 2: selección, limpieza y preprocesamiento

### Limpieza de datos (data cleaning)

- Datos perdidos (missing): rellenarlos manualmente, ignorarlos, eliminar la fila/columna, usar un valor especial p.e. unknown, inferirlos usando técnicas estadísticas, ...
- Datos anómalos (outliers): primero identificarlos y después el tratamiento es parecido al caso anterior, sólo que el valor puede darnos alguna idea.
- Ruido: error aleatorio o siguiendo una varianza en los datos. El tratamiento básico es suavizar mediante técnicas estadísticas (binning, regresión, ...)
- Inconsistencias: registros duplicados, datos inconsistentes, ... normalmente ya tratado en la elaboración del DW.

## Fase 2: selección, limpieza y preprocesamiento

### Procesamiento/transformación

- Redefinición de los atributos mediante agrupamiento o separación.
- Transformación de los atributos: fecha nacimiento => edad, apellidos => etiquetas separadas, ...

En ocasiones => almacenar meta-información sobre la información realmente almacenada por cada campo.

- Discretización. Pasar atributos continuos (o discretos con muchos valores) a casos discretos manejables.

Diversas técnicas.

Es imprescindible para muchos algoritmos de MD.

## Fase 3: refinación de datos

- Reducción de casos/filas: Las técnicas usadas van desde la compresión al muestreo de los datos, pasando por la elección de representantes (clustering)
- Proyección: Seleccionar el conjunto de atributos adecuado para la tarea específica a realizar.

Suele conocerse también como **selección de variables** (feature subset selection o feature selection)

Las técnicas a emplear son: estadísticas, basadas en búsqueda combinadas con métodos empíricos, ...

## Fase 4: minería de datos

- ¿Qué tipo de conocimiento buscamos?  
predictivo o descriptivo
- ¿Qué técnica es la más adecuada?  
clasificación, regresión (predicción numérica),  
clustering/agrupamiento/segmentación, asociaciones, ...
- ¿Es necesario considerar la incertidumbre en el modelo  
resultante?  
certeza, probabilidad, lógica difusa, ...
- ¿Qué algoritmo es el más adecuado?  
clustering duro, difuso, jerarquizado, k-means, iterativo, EM  
(maximización expectación), ...

## Fase 5: evaluación e interpretación

- La fase de MD puede producir varias hipótesis de modelos
- Será necesario establecer qué modelos son los más válidos (técnicas habituales son el uso de conjuntos de tests independientes, ...)
- La interpretación de los mejores modelos (visualización, simplicidad, posibilidad de integración, ventajas colaterales, ...) ayudará a la selección del modelo(s) final(es)

## Fase 6: difusión y uso del nuevo conocimiento

- Elaboración de informes para su distribución
- Usar el nuevo conocimiento de forma independiente
- Incorporarlo a sistemas ya existentes (verificar con el conocimiento ya usado para evitar inconsistencias y posibles conflictos)

La monitorización del sistema en acción dará lugar a nuevos casos que realimentarán el ciclo del KDD

Las condiciones iniciales pueden variar, invalidando el modelo adquirido.