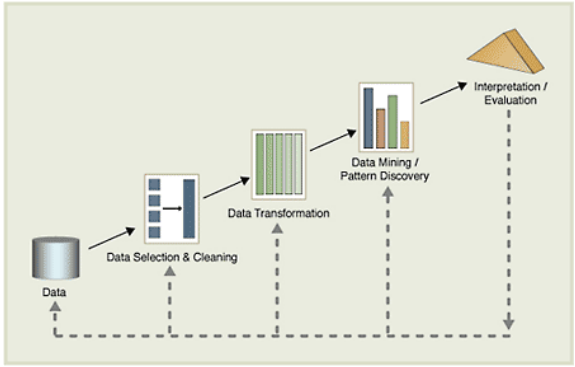
Módulo
Minería de Datos
Diplomado

Por
Elizabeth León Guzmán, Ph.D.
Profesora
Ingeniería de Sistemas
Grupo de Investigación MIDAS

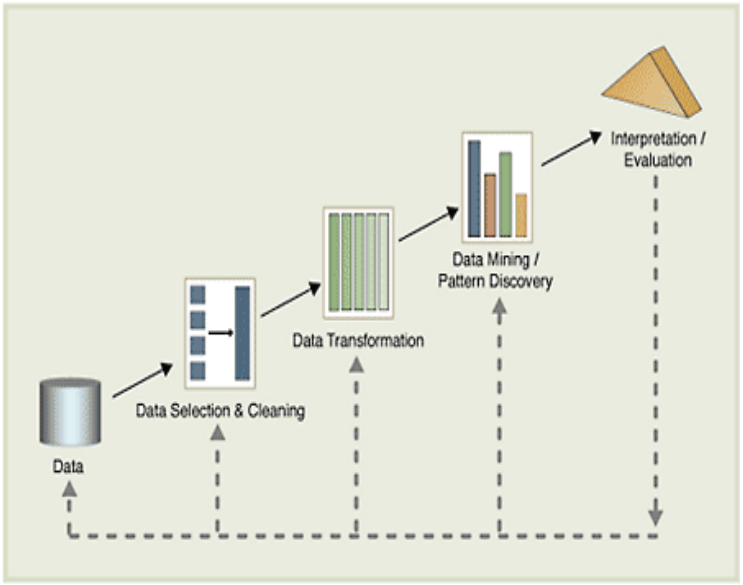
Metodologías

Las tres metodologías dominantes para el proceso de la minería de datos son: KDD, CRISP-DM y SEMMA.

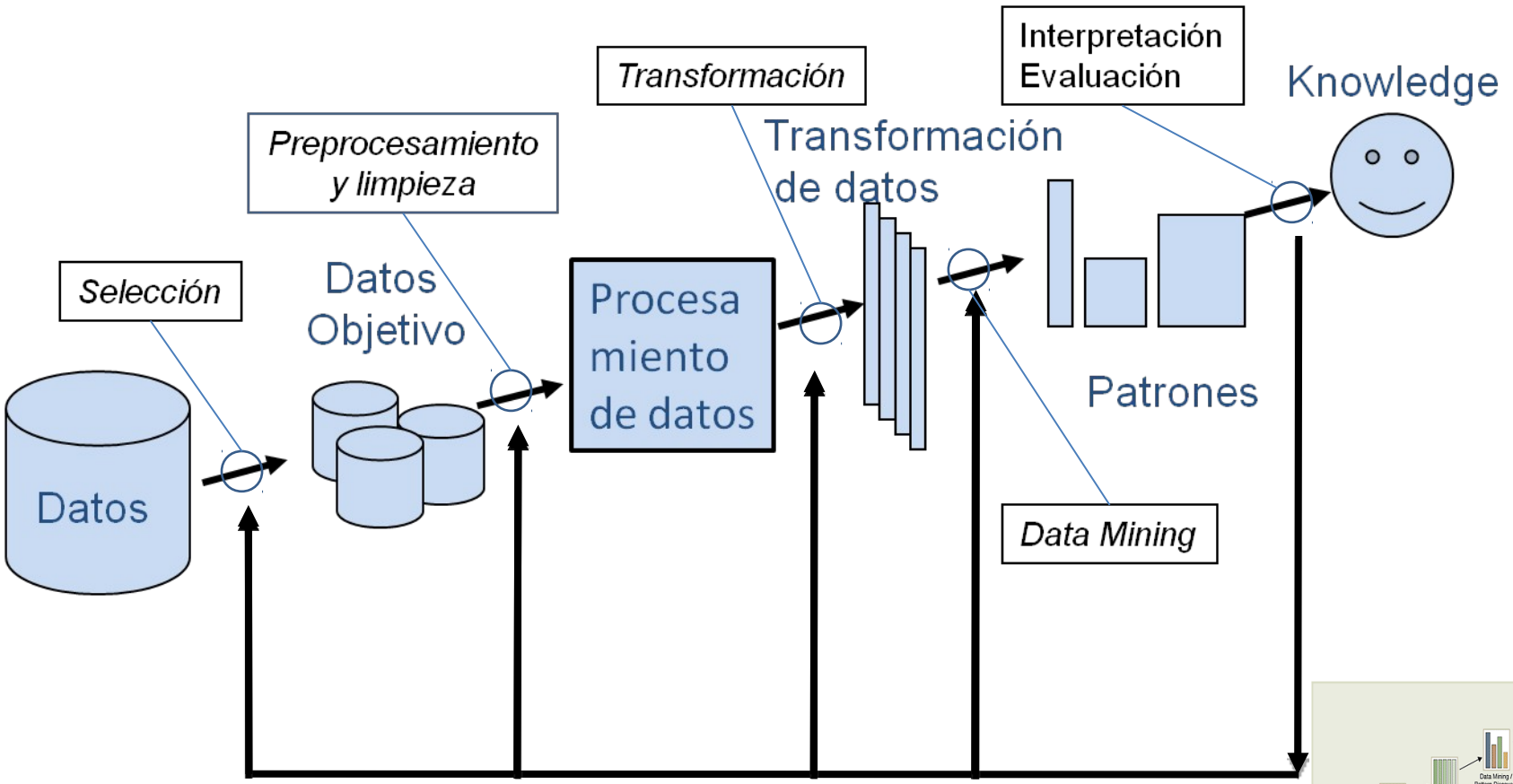


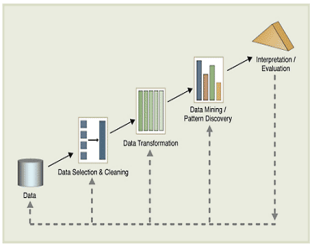
KDD

Es una metodología propuesta por Fayyad [3] en 1996, propone 5 fases: Selección, preprocesamiento, transformación, minería de datos y evaluación e implantación. Es un proceso iterativo e interactivo.



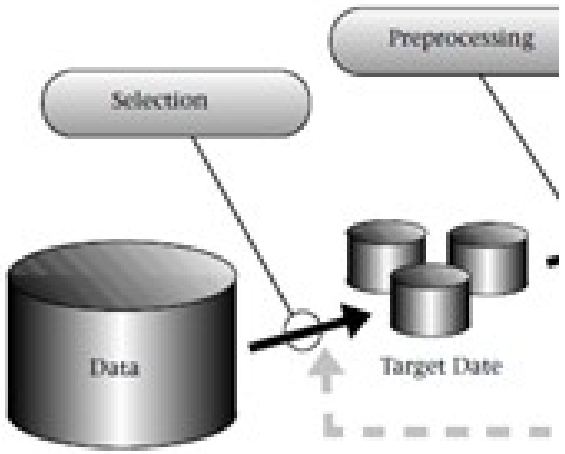
KDD

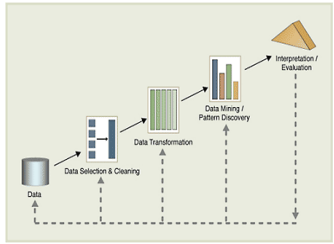

Interpretación
Evaluación
Knowledge
Transformación
Preprocesamiento
y limpieza
Transformación
de datos
Selección
Datos
Objetivo
Procesa
miento
de datos
Datos
Patrones
Data Mining



1 - 2 Pasos KDD

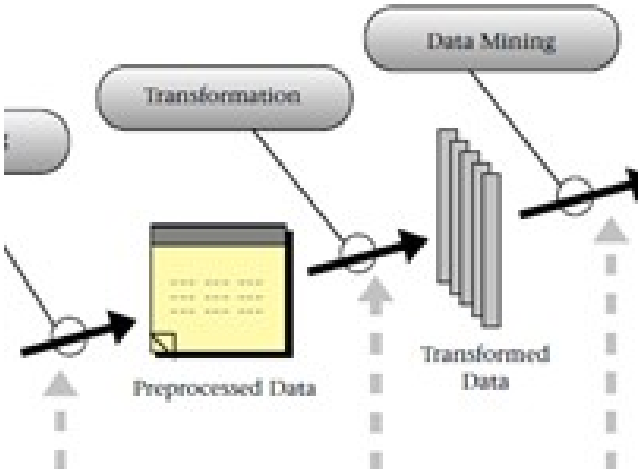
1. Desarrollar un entendimiento de la aplicación de dominio y los conocimientos previos y la identificación de la meta del proceso de KDD desde el punto de vista del cliente.
2. Crear un conjunto de datos objetivo: la selección de un conjunto de datos, o que se centren en un subconjunto de variables o datos de muestras, el descubrimiento que se llevará a cabo.





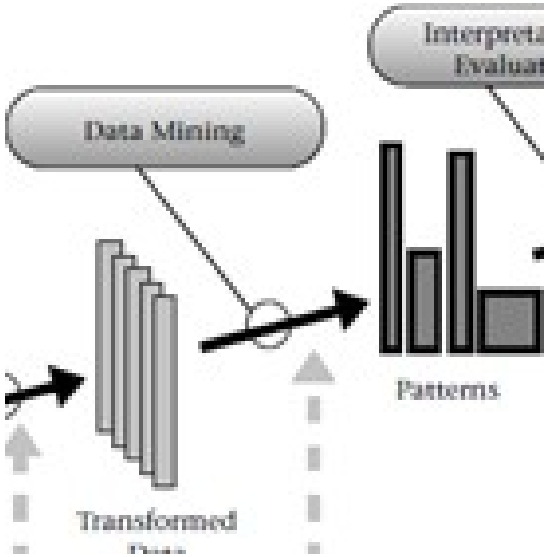
3 - 4 Pasos KDD

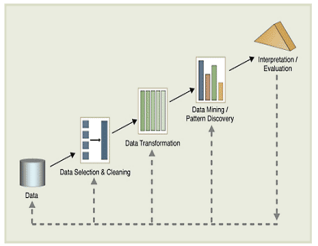
1. Limpieza y preprocesamiento de datos. Operaciones básicas incluyen la eliminación de ruido campos de datos vacios, etc.
2. reducción de datos y la proyección: la búsqueda de características útiles para representar los datos en función del objetivo de la tarea. (Reducción de la dimensionalidad)



5 – 7 Pasos KDD

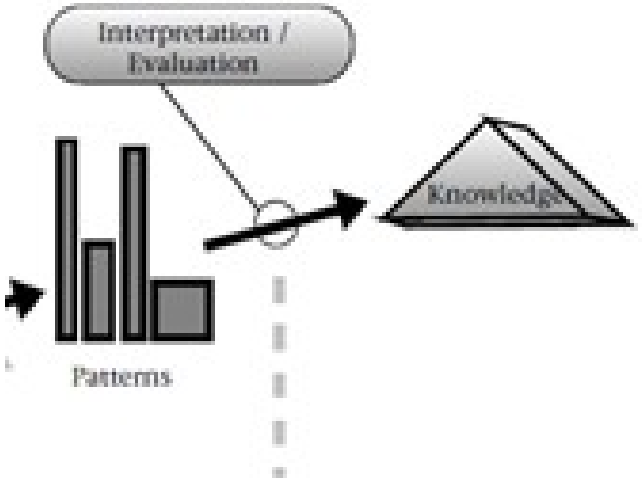
1. Es colocar el objetivo del KDD (paso 1) a un método de minería de datos.
2. Es el análisis exploratorio y de hipótesis y el modelo de selección: la elección del algoritmo de minería de datos que se utilizará para la búsqueda de patrones de datos.
3. Séptimo es la minería de datos: la búsqueda de patrones de interés en una determinada forma de representación o de un conjunto de tales representaciones.

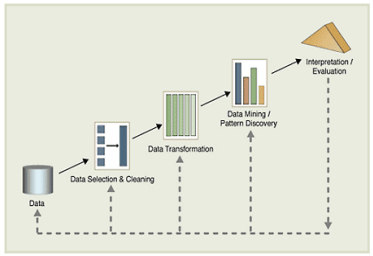




8 – 9 Pasos KDD

1. Interpretación de los patrones minadas, posiblemente se puede regresar a cualquiera de los pasos 1 a 7 para más iteración. Este paso puede implicar también la visualización de los patrones y modelos extraídos o visualización de los datos que figuran extraído modelos.
2. Está actuando sobre el conocimiento descubierto: el uso del conocimiento directamente, incorporando el conocimiento en otro sistema para la adopción de nuevas medidas o, simplemente, documentación y presentación de informes a las partes interesadas. Este proceso también incluye la comprobación y la solución de posibles conflictos con creían (o extrae) los conocimientos.





SEMMA

SEMMA es el acrónimo a las cinco fases: (Sample, Explore, Modify, Model, Assess) La metodología es propuesta por SAS Institute Inc, la define como: "… proceso de selección, exploración y modelamiento de grandes cantidades de datos para descubrir patrones de negocios desconocidos…"[2].

Fases y actividades SEMMA

Entrada de datos,
Ejemplos,
Partición de datos

Exploración distribuida,
Múltiples particiones,
Intuición, Asociación,
Selección de Variables

Transformación de variables, Filtros a los datos fuera de rango, Agrupamiento, ruido

Regresión,
Arboles,
Redes neuronales
Etc.

Evaluación,
Medidas,
Reportes

CRISP- DM

- Cross-Industry Standard Process for Data Mining (CRISP-DM)
- Iniciativa financiada por la Comunidad Europea ha unido para desarrollar una plataforma para Minería de Datos.
- Objetivos:
  - Fomentar la interoperabilidad de las herramientas a través de todo el proceso de minería de datos
  - Eliminar la experiencia misteriosa y costosa de las tareas simples de minería de datos.

CRISP-DM Proceso Estandar

- Plataforma para almacenar experiencia:
  - Permite que los proyectos sean replicados.
- Ayuda a la planeación y gerencia del proyecto.
- “Factor de Comodidad” para nuevos usuarios:
  - Demuestra la madurez de la minería de datos.
  - Reduce la dependencia en “estrellas”

CRISP-DM VENTAJAS

- ✓ No - propietario
- ✓ Independiente de la aplicación o la industria.
- ✓ Neutral con respecto a herramientas
- ✓ Enfocado en problemas de negocios así como en el análisis técnico.
- ✓ Plataforma guía
- ✓ Experiencia Base
- ✓ Plantillas para Análisis

CRIPS - DM

CRoss Industry Standard Process for Data Mining (CRISPDM)

Inicia en Sept. de 1996

Financiado por la Comisión Europea:

- Red mundial de aproximadamente 200 miembros

CRISPDM: Fabricantes de herramientas DW:

SPSS · SAS · DaimlerChrysler · OHRA · Teradata a division of NCR · sgi · DD Informatica Begrijpelijke Software · IBM

Proveedores de Sistemas / consultores:

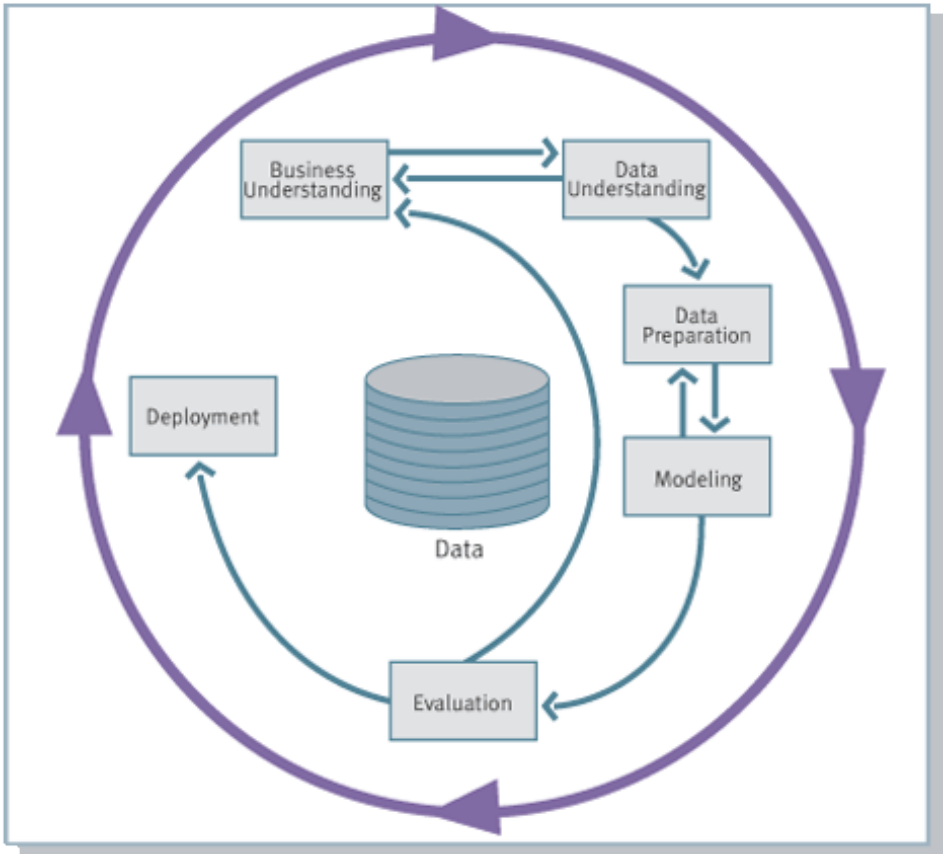
Capgemini Consulting. Technology. Outsourcing · ICL a Fujitsu company · Deloitte.

LLOYDS BANKING GROUP

Usuarios Finales:

AirTouch Paging · ABB · experian

Proceso CRISP-DM



CRISP-DM: Fases

1. **Comprensión del negocio:**
   - ✓ Entendimiento de los objetivos y requerimientos del proyecto.
   - ✓ Definición del problema de Minería de Datos
2. **Comprensión de los datos**
   - ✓ Obtención conjunto inicial de datos.
   - ✓ Exploración del conjunto de datos.
   - ✓ Identificar las características de calidad de los datos
   - ✓ Identificar los resultados iniciales obvios.
3. **Preparación de Datos**
   - ✓ Selección de datos
   - ✓ Limpieza de datos
4. **Modelamiento**
   Implementación en herramientas de Minería de Datos
5. **Evaluación**
   - ✓ Determinar si los resultados coinciden con los objetivos del negocio
   - ✓ Identificar las temas de negocio que deberían haberse abordado
6. **Despliegue**
   - ✓ Instalar los modelos resultantes en la práctica
   - ✓ Configuración para minería de datos de forma repetida ó continua

Fases y Actividades

Comprensión del Negocio	Comprensión de Datos	Preparación de Datos	Modelamiento	Evaluación	Despliegue
•**Determinar los Objetivos del Negocio** ✓*Antecedentes* ✓*Objetivos del Negocio* ✓*Criterio de Éxito* • **Evaluar la situación** ✓*Inventario de requerimientos de Recursos, Hipótesis y Limitaciones* ✓ *Riesgos y Contingencias* ✓*Terminología* ✓*Costos y Beneficios* • **Determinar el objetivo de Minería de Datos** ✓*Objetivos de Minería de Datos* ✓*Criterio de Éxito de Minería de Datos* • **Desarrollar el Plan de Proyecto** ✓ *Plan de proyecto* ✓ *Evaluación inicial de*	•**Obtener los datos iniciales** ✓*Reporte de la obtención de los datos iniciales* • **Describir los Datos** ✓*Reporte con la descripción de los datos* • **Explorar de Datos** ✓*Reporte de la Exploración de Datos* • **Verificar de la calidad de los Datos** ✓ *Reporte de la calidad de los datos*	• *Conjunto de Datos* • *Descripción de los Datos* • **Seleccionar los Datos** ✓*Justificación de la inclusión / Exclusión* • **Limpiar Datos** ✓*Reporte de Limpieza de Datos* •**Construir Datos** ✓*Atributos Derivados* ✓*Registros Generados* • **Integrar Datos** ✓ *Datos Combinados* • **Dar formato a los Datos** ✓*Datos Formateados*	•**Seleccionar Técnica de Modelamiento** ✓ *Técnica de Modelamiento* ✓*Modelamiento* ✓*Hipótesis* • **Generar el Diseño de Prueba** ✓*Diseño de Prueba* • **Construir el Modelo** ✓*Configuración de los parámetros del Modelo* ✓*Descripción del Modelo* • **Evaluar el Modelo** ✓*Evaluación del Modelo* ✓*Revisión de la configuración de los parámetros del modelo*	• **Evaluar Resultados** ✓*Hipótesis de Minería de Datos* ✓*Resultados* ✓*Criterio de éxito del negocio* ✓ *Modelos aprobados* • **Revisar el Proceso** ✓*Revisión del Proceso* • **Determinar los siguientes pasos** ✓*Lista de Posibles Acciones* ✓*Decisión*	• **Desplegar el Plan** ✓ *Plan de Despliegue* • **Monitorear y Mantener** ✓*Plan de monitoreo y Mantenimiento* • **Desarrollar el reporte final** ✓*Reporte Final* ✓*Presentación Final* •**Revisión del Proyecto** ✓*Documentación de las experiencias*

Comparación entre KDD, SEMMA y CRISP-DM

KDD	SEMMA	CRISP-DM
Pre KDD	xxxxx	Conocimiento del negocio
Selección	muestra	Conocimiento de los datos
Preprocesamiento	exploración	
Transformación	Modificación	
Minería de datos	Modelo	
interpretación / evaluación	evaluación	
Post KDD	xxxxx	

Bibliografía

1. **From Data Mining to Knowledge Discovery in Databases,** *Usama Fayyad, Gregory Piatetsky-Shapiro,* From Data Mining to Knowledge Discovery in Databases*and Padhraic Smyth,* American Association for Artificial Intelligence, 1996
2. http://www.crisp-dm.org/CRISPWP-0800.pdf
3. http://www.sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html