Frameworks para el desarrollo de proyectos de minería de datos

Prof. Lorena Zúñiga S.

Febrero, 2020

Frameworks

- Knowledge Discovery in Databases (KDD)
- Cross Industry Standard Process for Data Mining (CRISP-DM)

Metodología KDD

1. **Data Cleaning**
   - Missing values, ruido, transformaciones
2. **Integración de datos**
   - Extracción, Transformación, Carga
3. **Selección de datos**
4. **Transformación de datos**
   - Para el proceso de minería

Metodología KDD

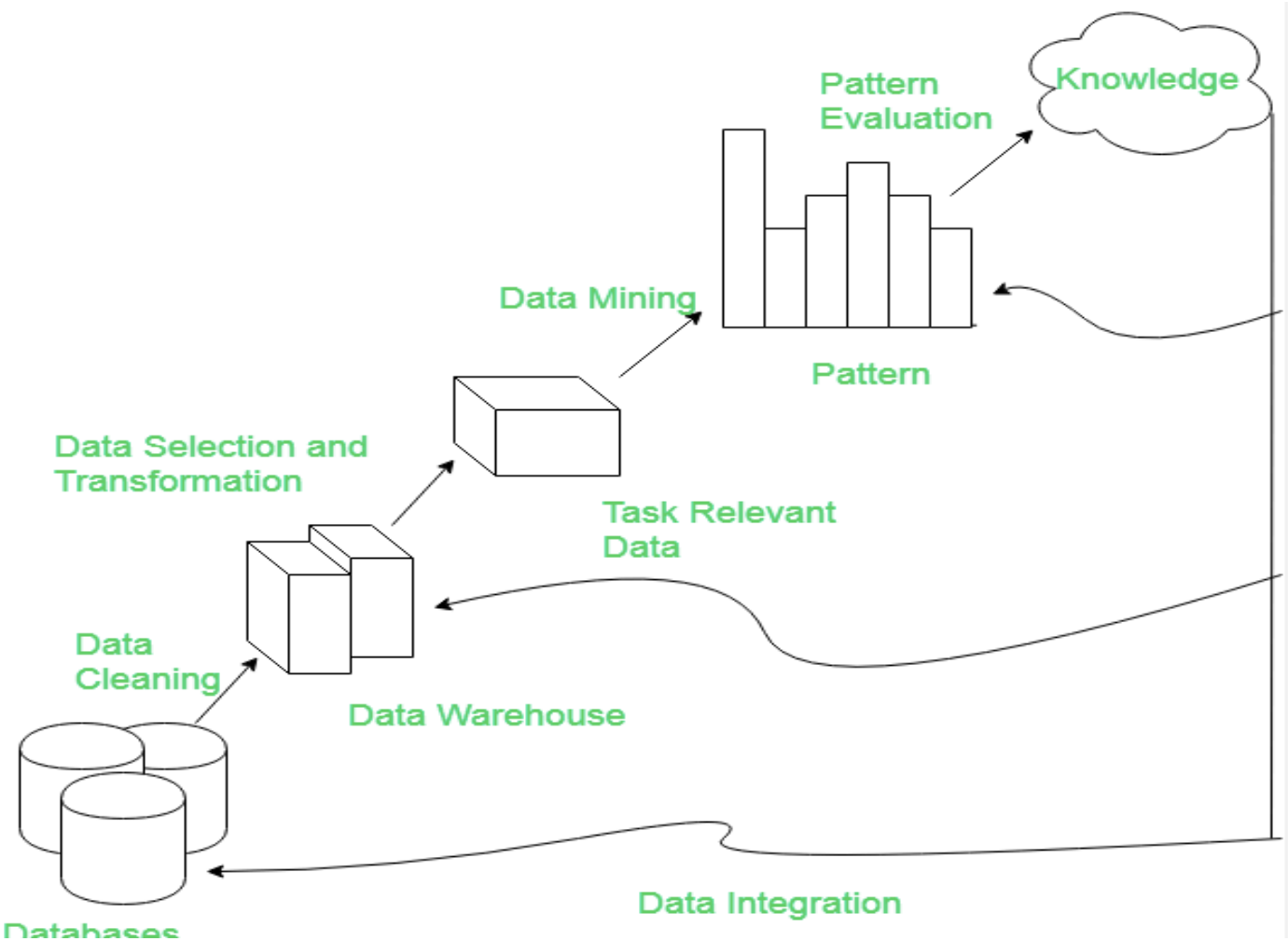
5. Minería de datos

- Propósito del modelo - Clasificación vs descripción
- Transformar datos en patrones

6. Evaluación de patrones

- Visualización, sumarización

7. Representación del conocimiento


Knowledge
Pattern Evaluation
Data Mining
Pattern
Data Selection and Transformation
Task Relevant Data
Data Cleaning
Data Warehouse
Data Integration
Databases

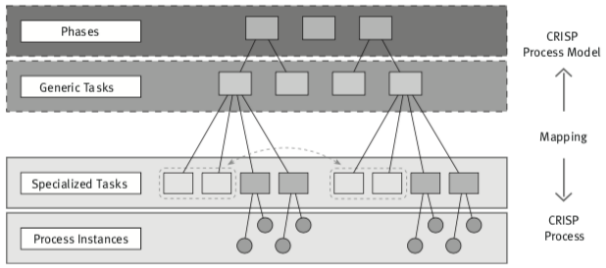
CRISP-DM

CRISP-DM

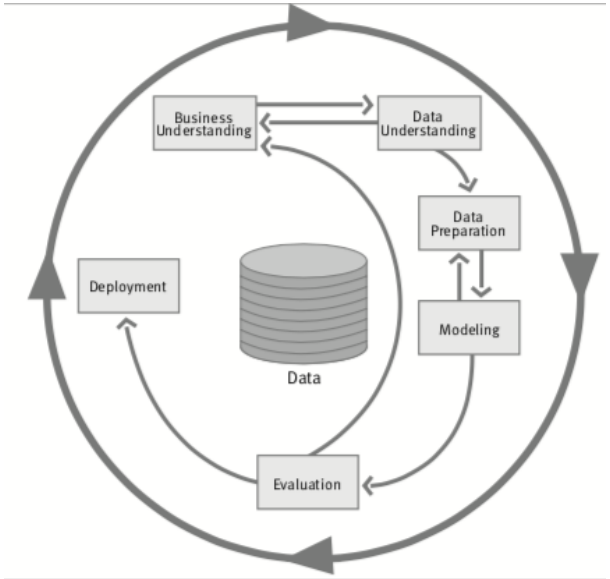
- Cross Industry Standard Process for Data Mining
- Propuesto por Daimler-Benz, SPSS, NCR Corporation en 1999
- Para cualquier software, datos

Framework CRISP-DM

- Organizada en fases
  - tareas genéricas
  - tareas especializadas (cómo)
  - instancias de proceso
    - registro de acciones, decisiones resultados



Fases del modelo



Fase Entendimiento del negocio

- Business understanding.
- Comprender los objetivos del negocio.
- Comprender los requerimientos, perspectiva de negocio.

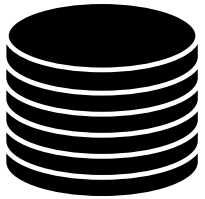
Fase Entendimiento del negocio

Definición de un problema de minería de datos

Plan preliminar

Fase Entendimiento de los Datos

- Recopilación de datos.
- Familiarizarse con los datos.
- Identificar problemas de calidad de datos.
- Primeros *insights.*



Fase Preparación de los datos

- Construcción del conjunto de datos final.
- Tareas de preparación de datos no tienen un orden fijo y pueden repetirse.
- Ejemplos

Fase de Modelado

- Selección y aplicación de las primeras técnicas de modelado.
- Calibración de los parámetros.
- Posible regreso a la fase de Preparación de datos.

Fase de Evaluación

- Evaluar los pasos que se llevaron a cabo para la construcción de los modelos.
- Asegurarse de que el modelo logra los objetivos del negocio.
- Decidir si el modelo se usará o no.

Fase de Deployment

- Aplicar el modelo en los procesos de toma de decisiones
- Puede ser simple o compleja, según los requerimientos.
- Por lo general, el despliegue no lo ejecuta el analista de datos.

Fases y tareas genéricas

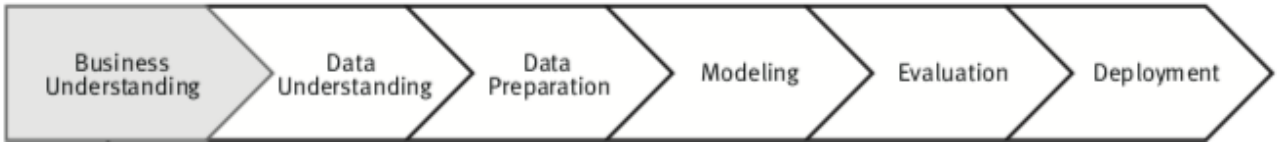
1. Tarea: Determinar los objetivos del negocio

- **Salidas:**
  - Registro de información de lo que se conoce sobre la situación del negocio (*background*).
  - Objetivos del negocio
  - Criterios de éxito del negocio

2. Tarea: Evaluación de la situación

Salidas:

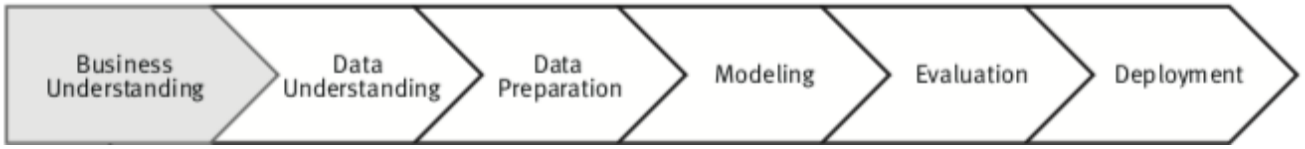
- Inventario de recursos (personal, datos, TI)
- Requerimientos, restricciones, supuestos
- Riesgos y contingencias
- Terminología
- Costos y beneficios



3. Tarea: Determinar los objetivos de minería de altos

Salidas:

- Objetivos de minería de datos
- Criterios de éxito desde la perspectiva de minería de datos



4. Tarea: Generar el plan del proyecto

Salidas:

- Plan del proyecto.
- Evaluación inicial de las técnicas y herramientas a utilizar.

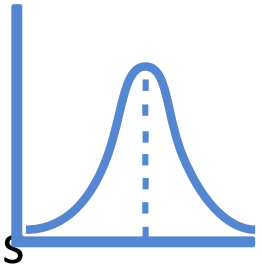
1. Tarea: Recopilación inicial de datos

- **Salida:**
  - Reporte de recopilación inicial de los datos

2. Tarea: Describir los datos

- **Salida:**
  - Reporte de descripción de los datos

3. Tarea: Explorar los datos



Salidas:

- Reporte de exploración de los datos

4. Tarea: Verificar la calidad de los datos

- **Salidas:**
  - Reporte de calidad de datos

1. Tarea: Seleccionar los datos

Salidas:

- Criterios para incluir/excluir datos

2. Tarea: Limpiar los datos

Salidas:

- Reporte de limpieza de datos

3. Tarea: Construcción de datos

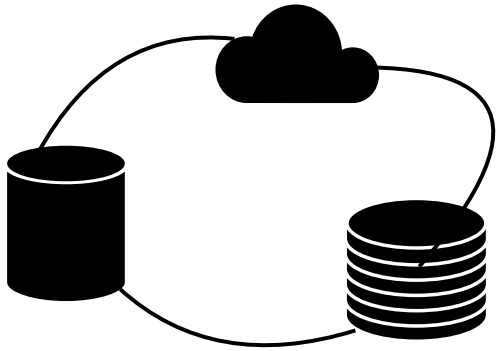
Salidas:

- Atributos derivados
- Registros generados

4. Tarea: Integrar datos

Salidas:

- Datos integrados



5. Tarea: Formatear datos

Salidas:

- Datos con nuevo formato

1. **Tarea: Seleccionar la técnica de modelado**
   - **Salidas:**
     - Técnica de modelado
     - Supuestos del modelado

2. Tarea: Generar diseño de testing

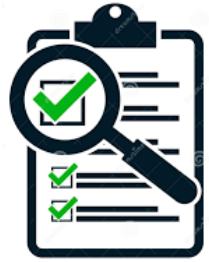
- **Salidas:**
  - Plan para testing, training y evaluación

3. Tarea: Construir el modelo

- **Salidas:**
  - Configuración de los parámetros y su justificación
  - Modelos
  - Descripción de los modelos

4. Tarea: Evaluar el modelo

- **Salidas:**
  - Evaluación de los modelos
  - Revisión de la configuración de los parámetros



1. Tarea: Evaluar los resultados

- **Salidas:**
  - Evaluación de los resultados de data mining
  - Modelo(s) aprobado(s)

3. Tarea: Determinar pasos a seguir

- Salidas:
  - Lista de posibles acciones
  - Decisión

1. Tarea: Plan de despliegue

- **Salidas:**
  - Plan de despliegue

2. Tarea: Plan de monitoreo y mantenimiento

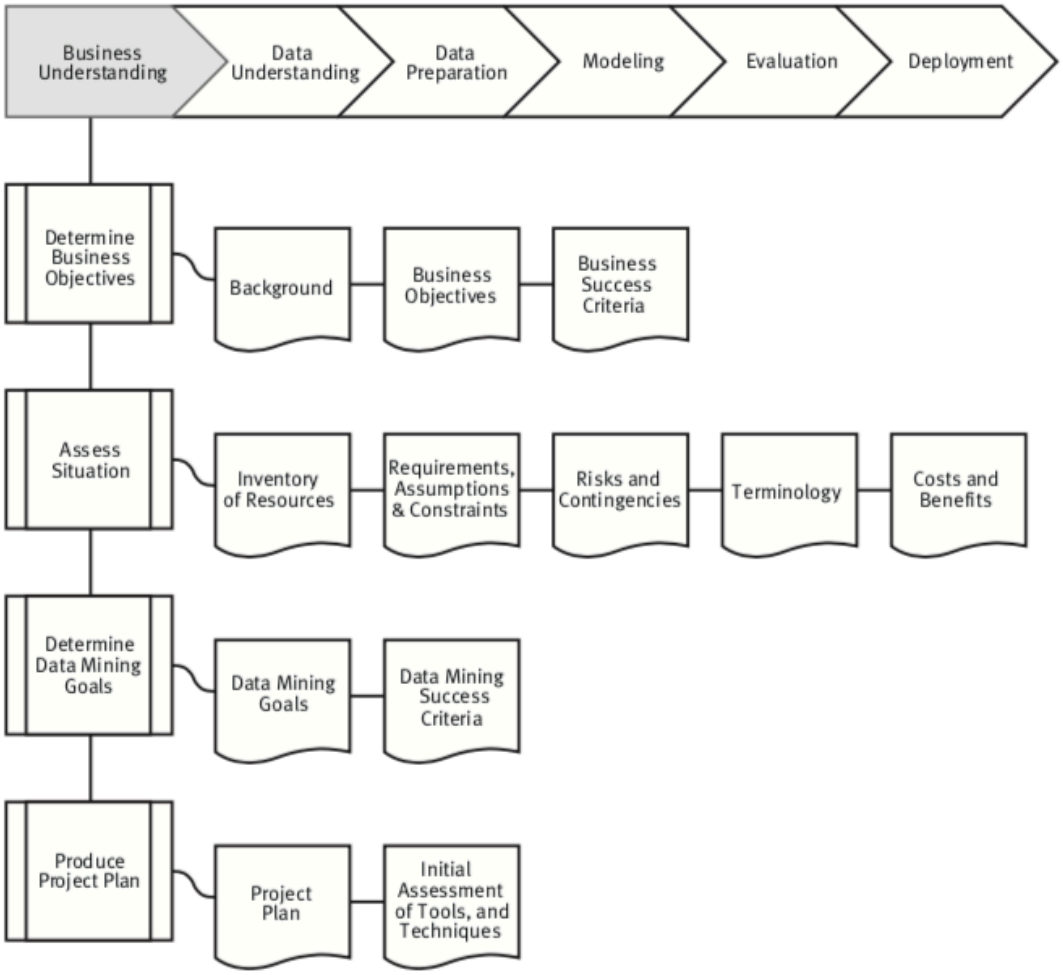
- Salidas:
  - Plan de monitoreo y mantenimiento

3. Tarea: Producción del reporte final

- **Salidas:**
  - Reporte final
  - Presentación final

4. Tarea: Revisión del proyecto

- Salidas:
  - Documentación de la experiencia generada


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Determine Business Objectives
Background
Business Objectives
Business Success Criteria
Assess Situation
Inventory of Resources
Requirements, Assumptions & Constraints
Risks and Contingencies
Terminology
Costs and Benefits
Determine Data Mining Goals
Data Mining Goals
Data Mining Success Criteria
Produce Project Plan
Project Plan
Initial Assessment of Tools, and Techniques