Transformación de Datos

Sesión 01

Ing. Gómez Marín, Jaime[1]

Módulo 3 : Análisis de Datos con Python
Departamento de TdG

September 2019

Índice

Introducción

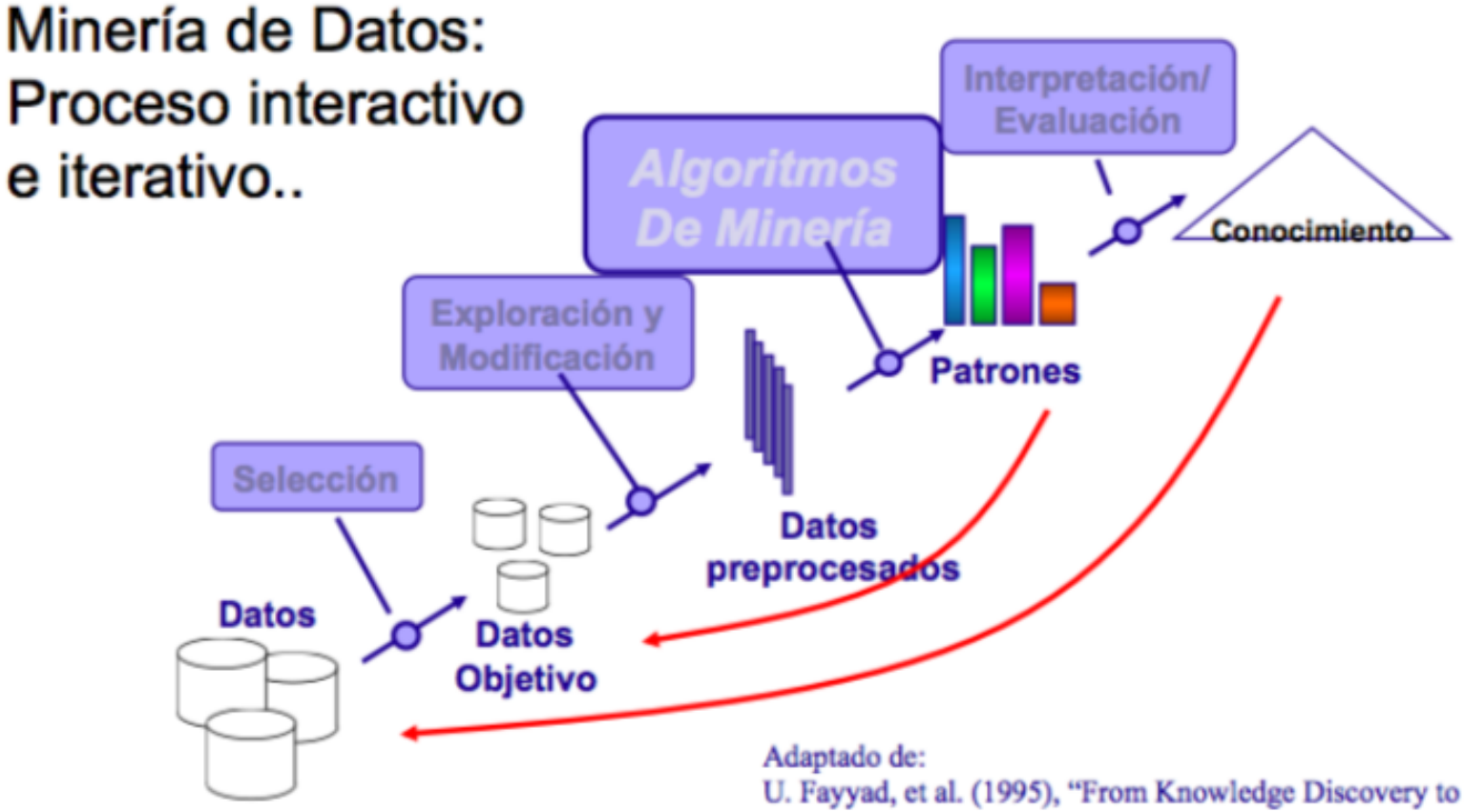
¿ Porque extraer datos ?

- Gran cantidad de datos coleccionada y almacenada
- ILas computadoras se vuelven más baratas y poderosas
- La presión de la competencia es fuerte
- Data recolectada y almacenada a enorme velocidad (GB/hora).

Motivación

Las Herramientas y nuevas tecnologías automatizadas de almacenamiento de datos conllevan a que grandes cantidades de datos sean almacenados en bases de datos, data warehouses y otros repositorios de información.

KDD : Proceso de extraccción del Conocimiento

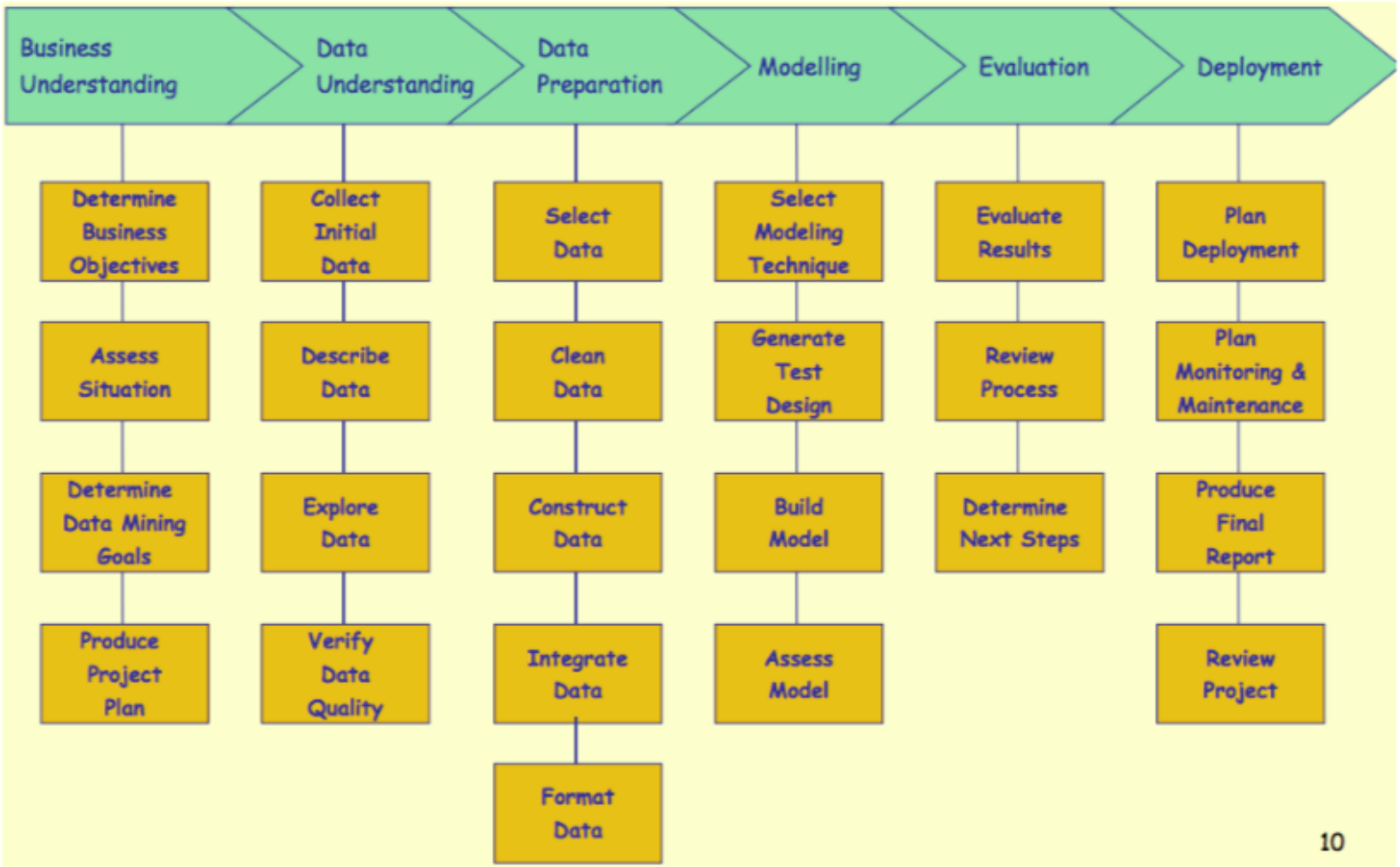


Adaptado de:
U. Fayyad, et al. (1995), "From Knowledge Discovery to Data Mining: An Overview," Advances in Knowledge Discovery and Data Mining, U. Fayyad et al. (Eds.), AAAI/MIT Press

CRISP-DM

CRISP-DM es una metodología completa de Minería de Datos y modelamiento de procesos que proporciona a cualquiera, desde principiantes a expertos, un programa completo para la realización de un proyecto de minería de datos. La metodología enumera los pasos para reproducir el éxito

CRISP-DM : Fases



Preparar los datos

El propósito de la preparación es transformar los conjuntos de datos de tal forma que la información que contienen esté mejor expuesta para la herramienta de minería de datos que se utilizará.

Datos Perdidos

- Identificar y manejar valores perdidos
  - Identificar valores perdidos
  - Evaluar datos perdidos
  - Corregir el formato de los datos
- Estandarizar datos
- Normalzacion de datos
- Binning

Datos perdidos : Impacto

Impacto de los valores faltantes:

- 1 % datos faltantes – trivial.
- 1-5 % – manejable
- 5-15 % – requiere métodos sofisticados
- Más del 15 % – interpretación perjudicial

Web Scraping

Web scraping es una técnica utilizada mediante programas de software para extraer información de sitios web.
En python se usan las siguiente librerias

- request
- beautifulsoup4

Conclusion

En esta session se han usado la forma de como poder obtener y limpiar los datos usando tecnicas de Mineria de Datos.

Bibliography

 Naomi Ceder. The Quick Python Book - Manning Publications, 2018.