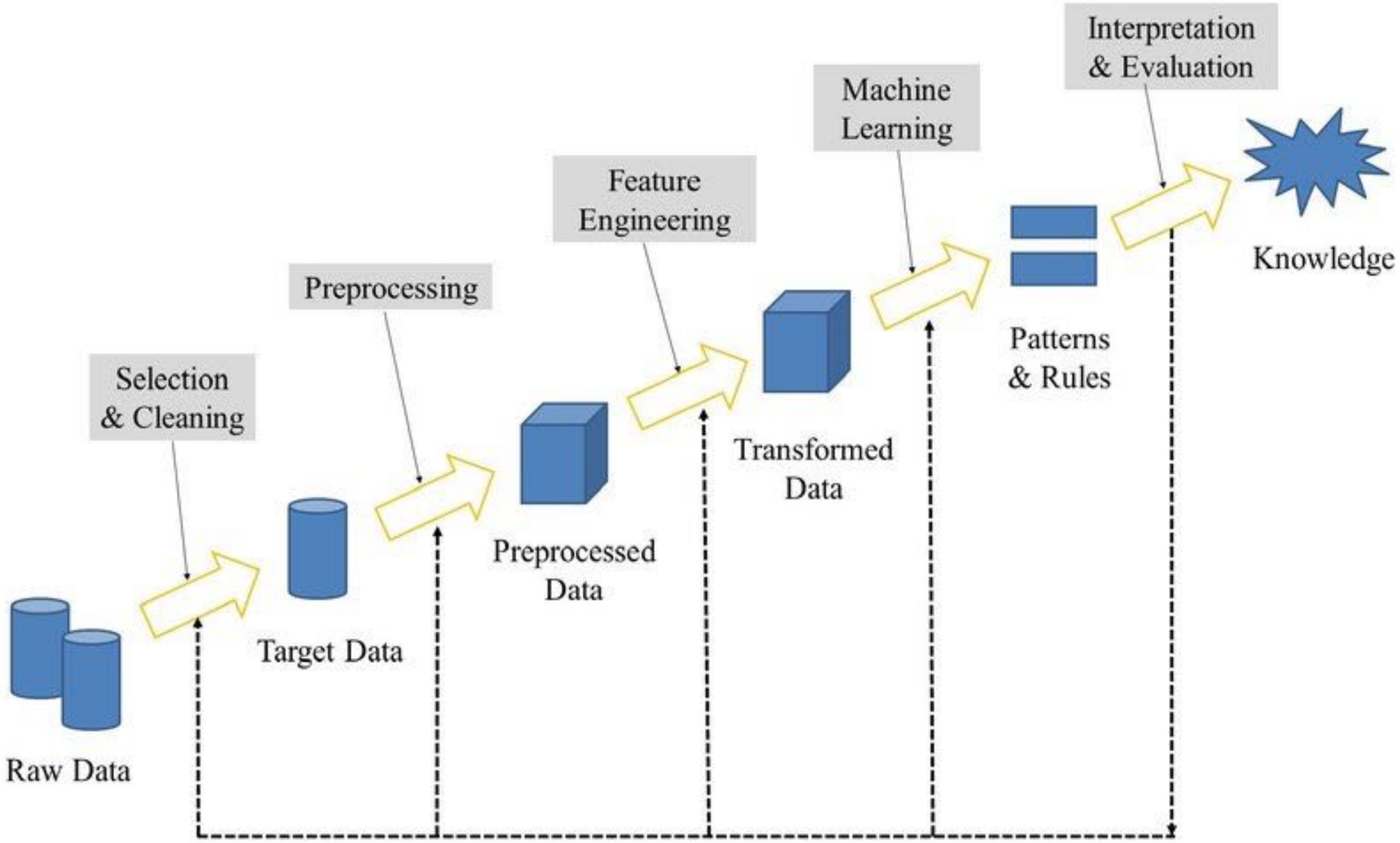
# Extract, Transform and Load

## TEL-354: Minería de Datos

Mauricio Araya

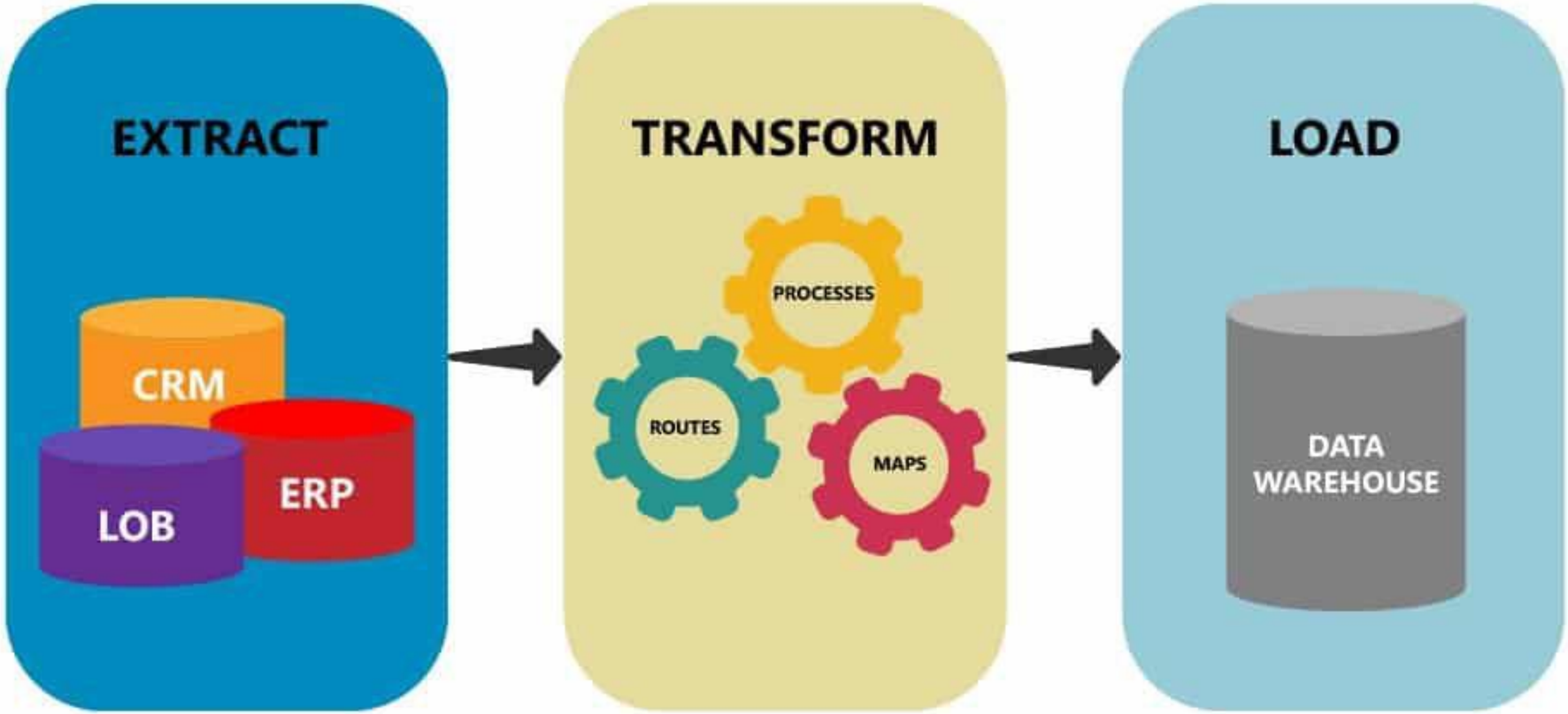
# Data Mining Process

**Cool, but raw data come from where?**



# To have raw data you need to slaughter the animal

**Databases, version control systems, web scrapping, file formats, etc.**




ETL - Extract, Transform, Load

# All you need is Data

## Extracting the data is always your first step

- Find your data and know the steps to access it completely
  - How many **sources** are? Are they reliable?
  - How **easy** is to get them?
  - Need **authorization?**
  - Technical **access** (user/pass)
  - Can I **host** it?
  - If not. How **fast** I can access it?
  - Get as much **metadata** you can...

- **Ejercicio:**
  - Busque por datos **oficiales** del COVID-19 en Chile
  - Hay más de una fuente **oficial**?
  - Que **volumen** tiene?
  - Podemos **bajarla**?
  - Cómo se **baja**?
  - **Baje** la data!

# Own your data: Transforming

## Put all your stuff in order before data mining

- Removing fat, make the cuts, complement, etc.
  - What can I **eliminate**?
  - How can I **update** the data if needed?
  - Data come in thee **file formats** I want?
  - Are the **semantics** of the data fine for my institution?
  - Can I **join** some data from different sources?
  - Can I **split** some data that is in one file/DB?

- **Ejercicio:**
  - Objetivo: imprimir gráficos.
  - ¿Qué datos queremos usar?
  - ¿Están en el formato adecuado para Jupyter notebooks?
  - ¿Pandas puede leerlo?
  - Saquelo del repo y pongalo en su propio directorio (contexto).
  - Opcional: Haga un script de actualización o use enlaces simbólicos

# Loading your data for the first time

### Check if you are ready to rumble...

- Can load the data in your system?
- This could be very complex or simple depending on:
  - Your **systems** characteristics
    - Personal vs Institutional
    - Production vs Testing
    - Small vs Big
  - Your **data** characteristics
    - Big V's: Volume, Velocity, Variety, Veracity, Value
    - Batch vs Streaming

- **Ejercicio:**
  - Cargue sus datos ocupando pandas (pd.read_csv)
  - Describa sus datos (df.describe)
  - Qué datos son útiles?
  - Grafique algo
  - Puede identificar Samples y Features?

# Cloud Revolution

## Just a Change of Letter