# Preprocesamiento para Ciencia de Datos
# Avances de Proyecto; primer entrega.

Daniel Rojo Mata
Fernando Rodrigo Valenzuela García de León

04 de septiembre de 2025

# Introducción

- Construcción de una **base de datos heterogénea federada** en español en el sector salud.
- Enfoque: **trastornos por uso de sustancias** (CIE-10 F10–F19).
- Integración de tres tipos de fuentes:
  - **Relacional (CSV → SQL)**: defunciones y urgencias.
  - **Grafo (CIE-10)**: jerarquía y comorbilidad.
  - **Texto (CoWeSe)**: corpus en español con mapeo a códigos.

## Sustancias psicoactivas

- Cualquier sustancia química que, al ser ingerida, produce un efecto en el cerebro, alterando la percepción, el estado de ánimo, los pensamientos o el comportamiento.

# Fuentes de datos

## Relacional

- `defunciones_uso_sustancias_clean.csv` (12,473 filas).
- `urgencias_uso_sustancias_clean.csv` (25,051 filas).

## Grafo

- `cie10_f10_f19_nodes.csv` y `edges_enriched.csv`.
- 110 nodos, 107 aristas.

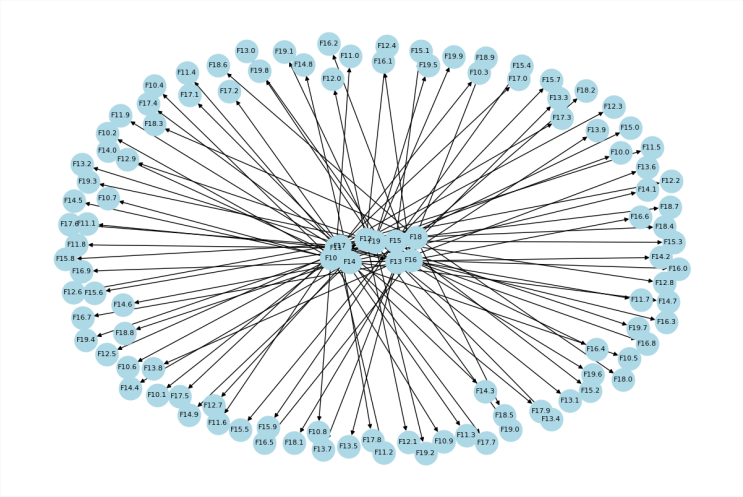
## Texto

- Corpus `CoWeSe.txt` (50k oraciones).
- 770 oraciones mapeadas a F10–F19.

# Integración de fuentes

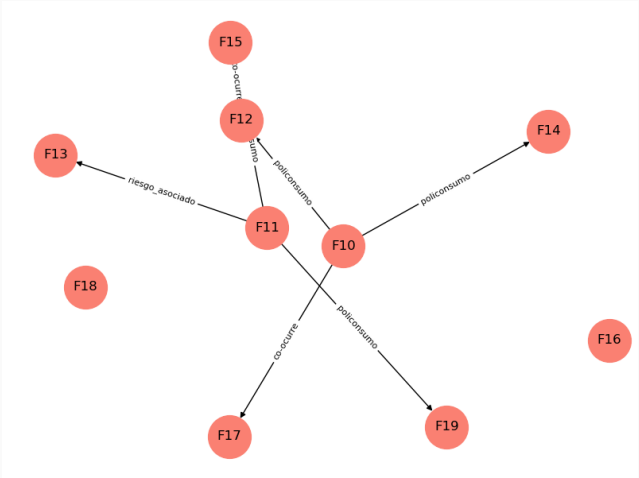
- **Claves comunes:**
  - Códigos CIE-10 (F10–F19).
  - Contexto: año, entidad, sexo, edad.
- **Aristas en el grafo:**
  - Jerárquicas (CIE-10 oficial).
  - Epidemiológicas (comorbilidad/policonsumo).
- **Texto como puente:**
  - Detección exacta con diccionario.
  - Búsqueda semántica*

# Integración de fuentes



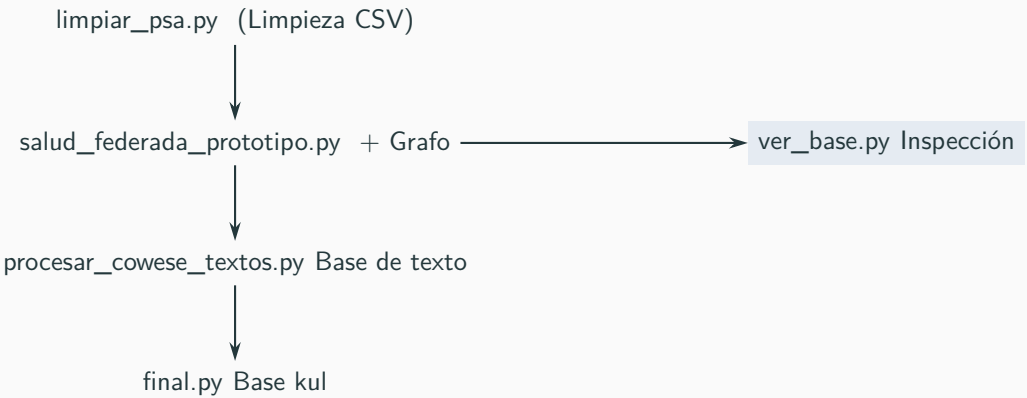

Red completa: nodos y adyacencias (jerarquía + epidemiología).

# Integración de fuentes




Subred de comorbilidad/policonsumo: aristas epidemiológicas.

# Proceso del proyecto



# Ejemplo de consulta federada

**Consulta:** “cocaína”

- Texto: detección de F14 (cocaína) + co-ocurrencias (F10, F11, F12, F15, F17).
- SQL: casos de urgencias en 2018 $\rightarrow$ 89 en Aguascalientes, 21 en Baja California...
- Grafo: subtipos F14.0–F14.9.
- Ejemplo de oración: *“Las drogas ilegales (cocaína, anfetaminas) son...”*

# Resultados

- Tablas SQL cargadas en `salud_federada.db`.
- Grafo con jerarquía y aristas de comorbilidad/policonsumo.
- Corpus procesado: 50k oraciones, 770 matches.
- Integración lograda: (texto $\rightarrow$ SQL + grafo).

# Eso es todo, gracias, :)