Ciencia de Datos: un primer acercamiento

Curso CICADA 2022

Gestión de datos

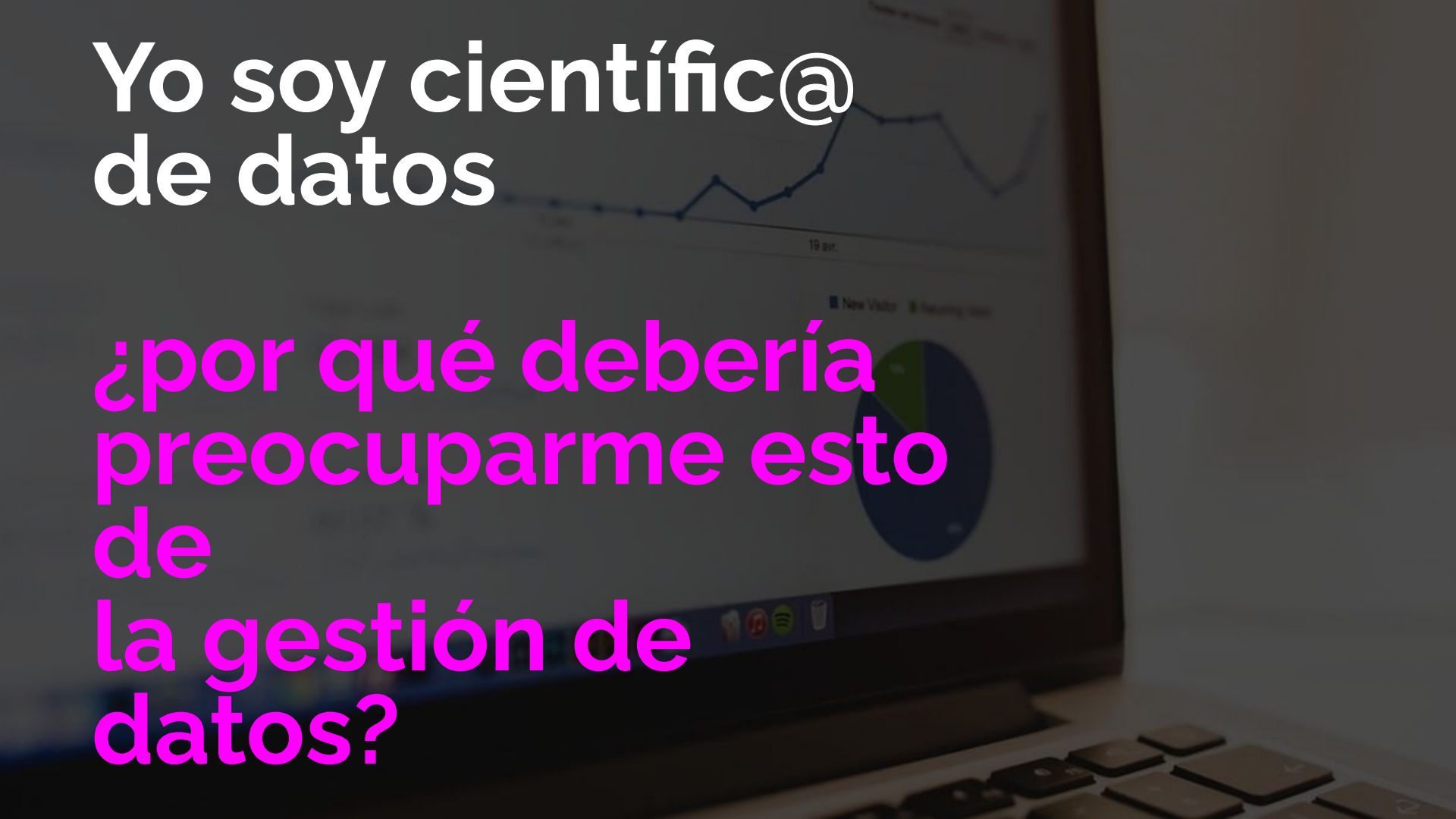
UNIVERSIDAD DE LA REPUBLICA URUGUAY

Acerca de mi

Lorena Etcheverry Venturini

Dra. en Informática (PEDECIBA Informática - UdelaR), Ingeniera en Computación (FING- UdelaR)
Trabajo en el Instituto de Computación de la FING dentro del grupo GEMA (Gestión de Datos, Modelado y Análisis).
Mi área de interés es la gestión de datos, en particular las bases de datos, la web semántica y los sistemas de análisis de datos.

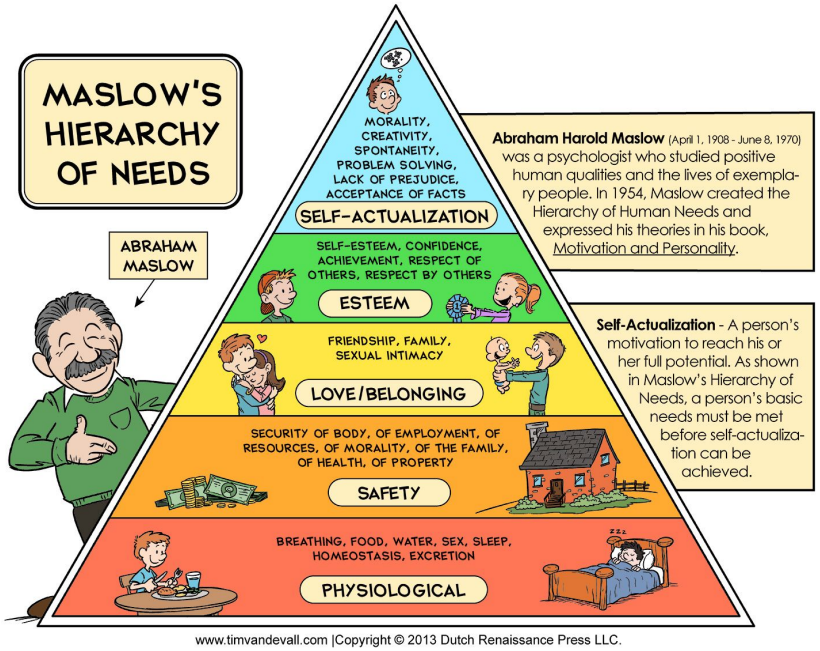
Les invitamos a presentarse en el foro de “Consultas e intercambio”

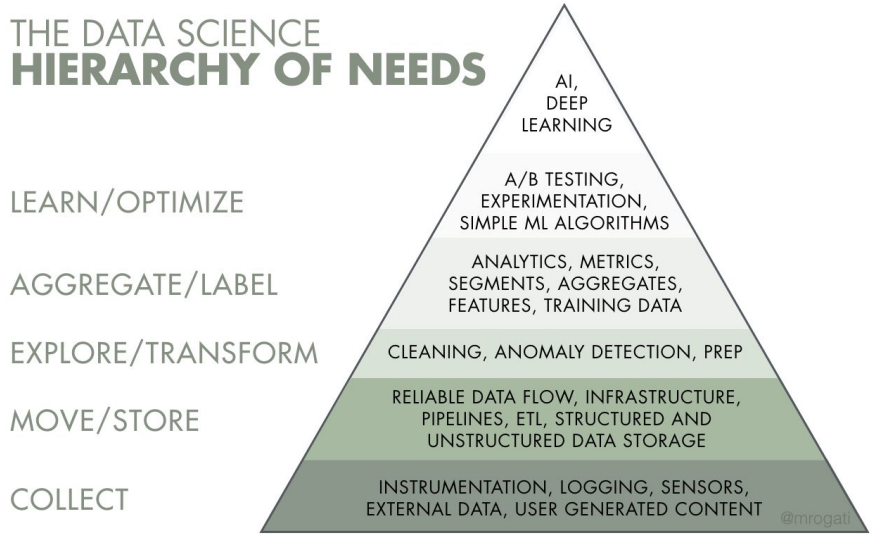


Yo soy científic@ de datos

¿por qué debería preocuparme esto de la gestión de datos?

Es imposible hacer ciencia de datos sin datos 😉





Mónica Rogatti, The AI hierarchy of needs

LAS EXPECTATIVAS EN CIENCIA DE DATOS

Ezequiel Orbe DATA ENGINEERING (A Survival Kit For Data Scientists), FaMAF - UNC

LA REALIDAD EN CIENCIA DE DATOS

¿ENTONCES?

Muchas veces, las empresas/organizaciones intentan comenzar por la punta de la pirámide debido a diferentes razones

1. la presión por obtener resultados ya!!
2. la poca madurez y/o la gestión de datos desorganizada (a veces bajo nivel de *data literacy*)
3. la falta de conocimiento

Los científicos de datos quedan en el lugar de ☞

The Data Scientist Unicorn

Joel Quesada Follow
Jan 10, 2019 · 3 min read

A Data Scientist Unicorn during his interview

Joel Quesada, The Data Scientist Unicorn

Pero, ¿de dónde salen los datos?

Se puede decir que casi siempre provienen de un Sistema de Información

Conclusión

La capacidad de extraer valor de los datos está directamente relacionada con la madurez de la **plataforma de datos** de la organización/empresa.

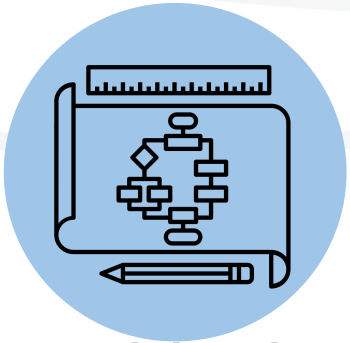
Si la organización no tiene una infraestructura de datos madura, uds van a necesitar habilidades de gestión de datos (a.k.a data engineering).

Objetivos de esta parte del curso

Brindarles definiciones y conceptos generales sobre gestión de datos y *data engineering*.

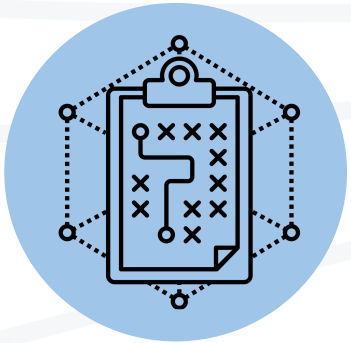
Desarrollar capacidades básicas, y proveer referencias para que uds. puedan profundizar.

Temas que abordaremos



Modelos de datos y modelado

Por que no todo es una planilla de cálculo



Calidad de datos y limpieza

Data profiling y
Data wrangling
Conceptos de Calidad de Datos

Aspectos éticos, privacidad y sesgo

Privacidad y anonimización
Modelos y sesgo

¿de qué hablamos cuando hablamos de datos?

Datos, información y conocimiento

Datos

- Un parámetro o hecho, un número, una afirmación, una imagen
- Representan algo en el **mundo real**
- Son la materia prima para la producción de información

Información

- Datos con un significado en cierto contexto
- Datos relacionados
- Datos luego de su manipulación

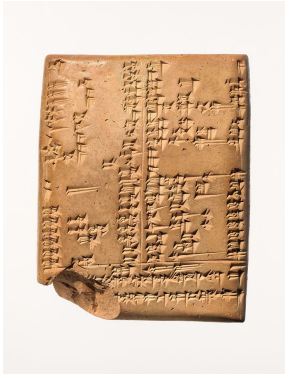
Conocimiento

- Experiencia e información acumulada

Los Sistemas de Información

Conjunto de componentes que interactúan con el objetivo de **almacenar**, **recuperar** y **procesar datos** e información para crear nueva información.
Los componentes de un Sistema de Información son software y hardware, pero es fundamental el rol de las personas.

Un ejemplo de los primeros Sistemas de Información: Censos (de personas y/o bienes) babilonios año 3800 a.c. !!!!



Los Sistemas de Información Informáticos

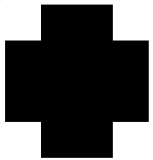
Utilizan tecnologías informáticas para realizar algunas de sus tareas.

Cumplen con tres funciones principales:

- **Memoria:** mantienen una representación del estado de cierto **dominio**
- **Informativa:** proveen información acerca del estado de cierto **dominio**
- **Activa:** realizan acciones que cambian el estado de cierto **dominio**

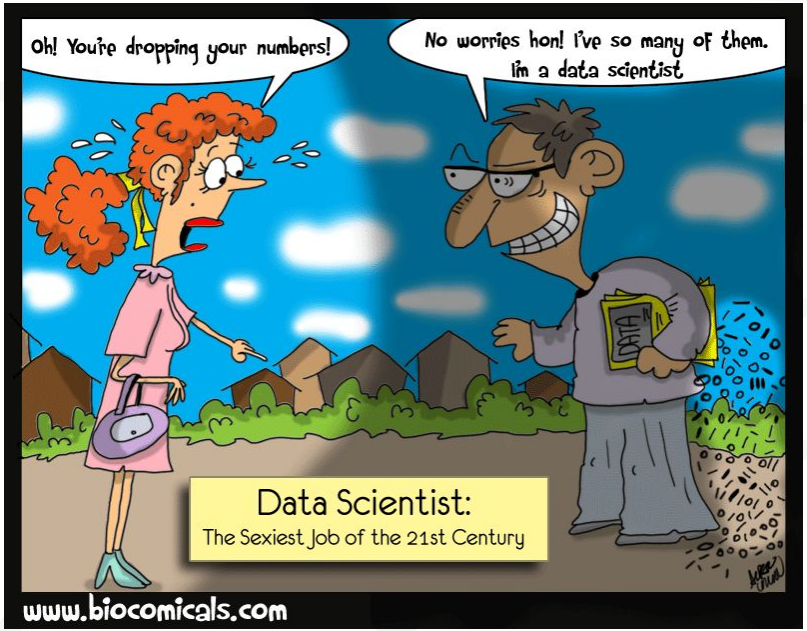
Las Bases de datos cumplen un rol central en los Sistemas de Información

DATOS

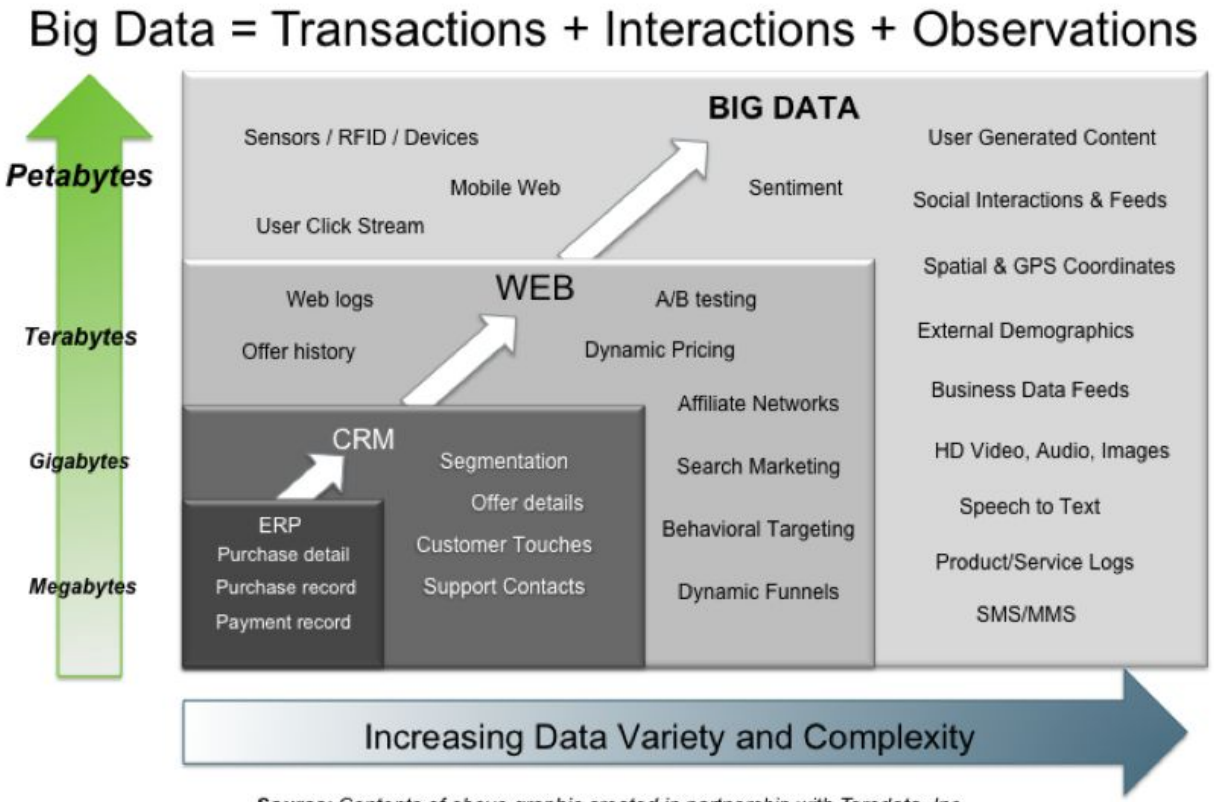


METADATOS

En los últimos 20 años el panorama es más complejo



La naturaleza de los datos y su volumen



***Source**: Contents of above graphic created in partnership with Teradata, Inc.*

¿Qué es un modelo de datos?

Modelo de datos: definición

Los modelos de datos son lenguajes usados para especificar y manipular Bases de Datos (BD)

Un Modelo de Datos permite expresar:

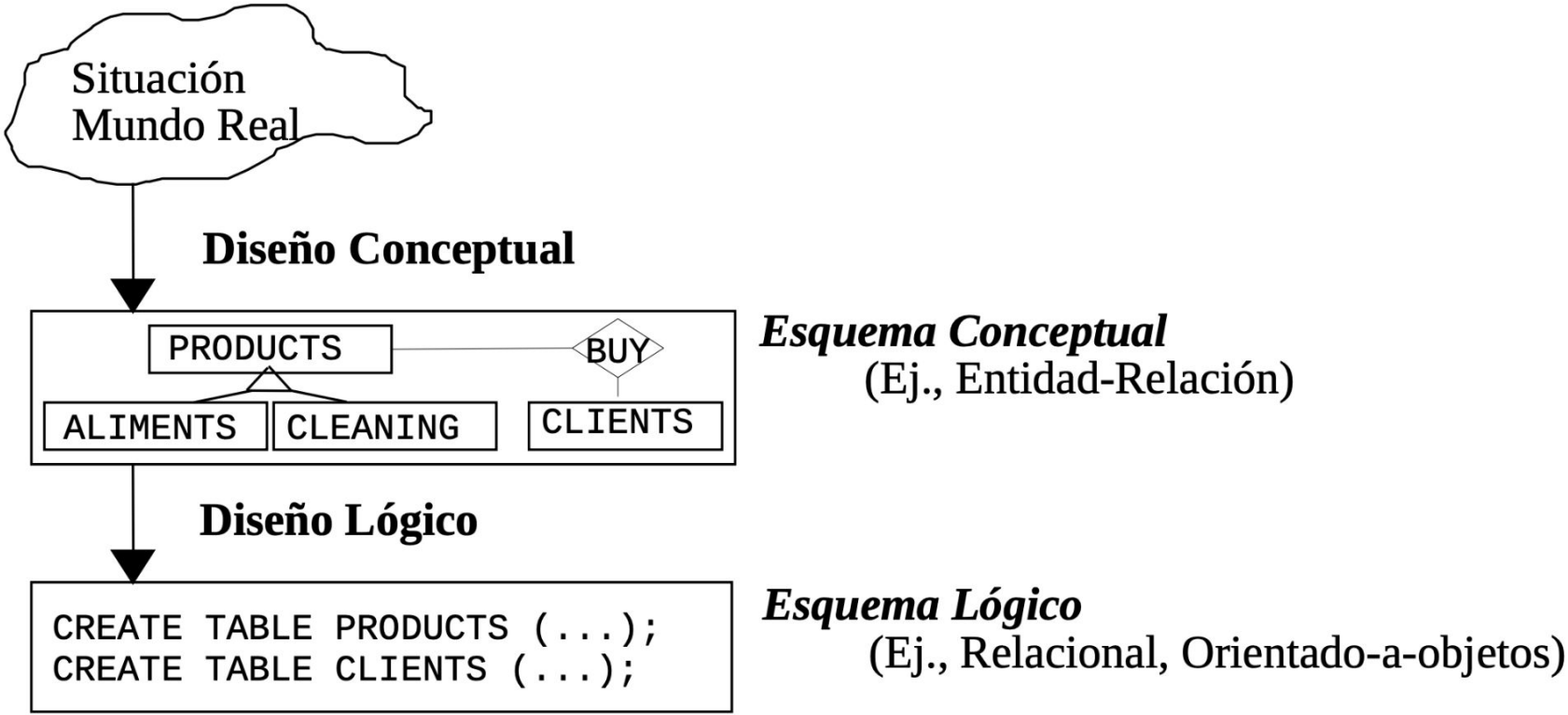
- **Estructuras:** Elementos de los problemas.
- **Restricciones:** Reglas que deben cumplir los datos para que la base sea considerada válida.
- **Operaciones:** Insertar, borrar y consultar la BD.

Clasificación de los modelos de datos por nivel de abstracción

Conceptuales: Representan la realidad independientemente de cualquier implementación de BD.

Lógicos: Son implementados en un manejador de bases de datos particular.

Físicos: Corresponden a cómo está implementado el manejador de bases de datos (estructuras de datos)



Esquema Conceptual
(Ej., Entidad-Relación)

Esquema Lógico
(Ej., Relacional, Orientado-a-objetos)

Modelos conceptuales

Se usan en las primeras etapas del diseño de una BD.
Actividad en la cual se construyen esquemas conceptuales de una realidad.

- Estudio del problema real.
- Especificación usando un lenguaje de muy alto nivel.
- Validar resultado.

Resultado: Esquema Conceptual

El **modelo Entidad-Relación** es un ejemplo de modelo Conceptual.

Modelado conceptual : ejemplo

En un hospital se tiene un registro de pacientes, un registro de personal y uno de salas con funcionarios que trabajan en esas salas y con pacientes internados en esas salas.
Del personal nos interesa el número de empleado, el nombre, la dirección y el teléfono.
Sabemos que dos empleados no tienen el mismo número.
De los pacientes nos interesa el número de registro (le es asignado cuando ingresa) y el nombre mientras que de las salas nos interesa el nombre y la cantidad de camas que tiene.
También se sabe que un empleado trabaja en una única sala y que en una sala trabajan varios empleados. Lo mismo ocurre con los pacientes.

Conjuntos de elementos de la realidad: Pacientes, Salas, Personal

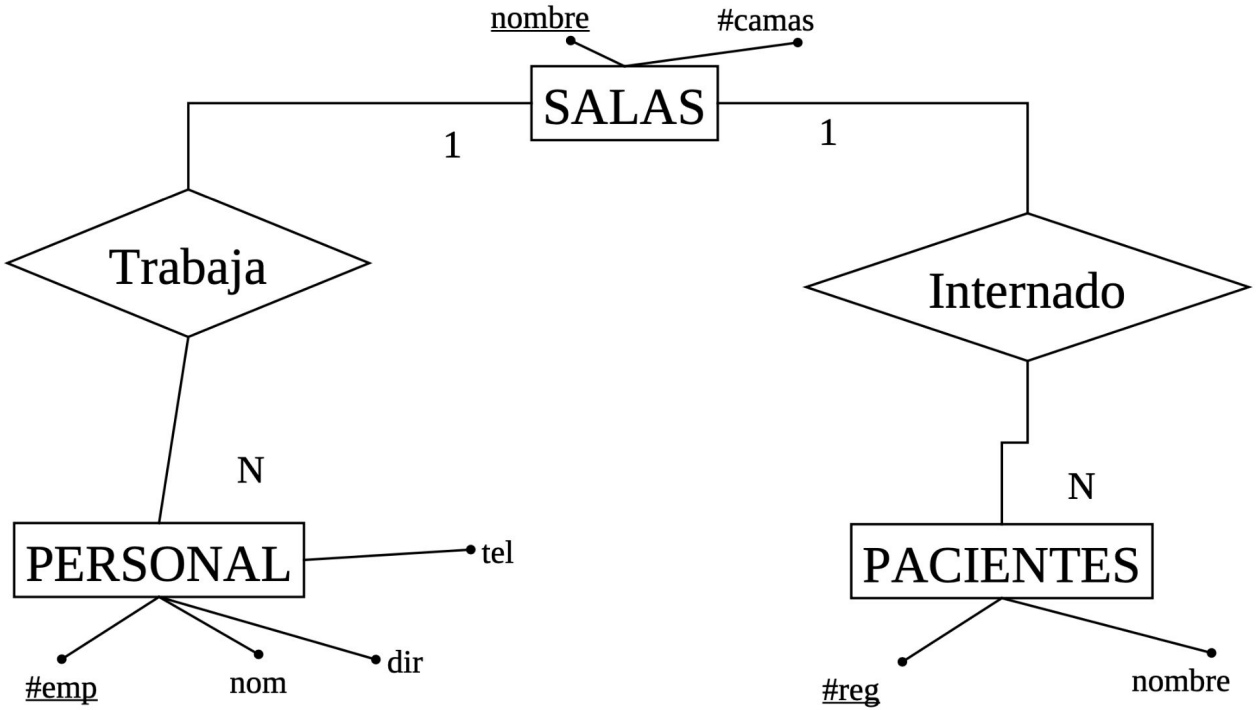
Relaciones entre esos conjuntos:
Los Pacientes están Internados en las Salas y el Personal Trabaja en las Salas.

Características que interesan de los objetos:

Personal: nro. de funcionario, nombre, direccion y telefono Pacientes: nro. de registro, nombre
Salas: nombre, cantidad de camas

Restricciones:
Un empleado trabaja en una unica sala y en una sala trabajan varios empleados. Un paciente está internado en una sola sala pero en una sala hay varios pacientes.

Esquema conceptual resultante



Modelos lógicos

El **modelo Relacional** es un ejemplo famoso y exitoso de modelo lógico.

Las estructuras consisten en TABLAS, cuyas columnas corresponden a ATRIBUTOS de tipo atómico y las filas corresponden a registros de datos.

Las operaciones manejan las TABLAS, como conjuntos de registros. Es un modelo de datos extremadamente simple y claro, que también ha resultado potente para la mayor parte de las aplicaciones de BDs.

Un esquema relacional del ejemplo

PERSONAL (#emp, nom, dir, tel)
PACIENTES (#reg, nombre)
SALAS (nombreSala, #camas)

TRABAJA(#emp, nombreSala)
INTERNADO(#reg, nombreSala)

Esquema vs Instancia

Describe qué datos hay en la base, cómo se relacionan esos datos entre sí y qué restricciones de integridad deben cumplir

Estructuras + Restricciones

Conjunto de datos almacenados en una base.

Es el valor de la base en un instante de tiempo.
Si respetan todos las restricciones, se considera que la instancia es correcta.

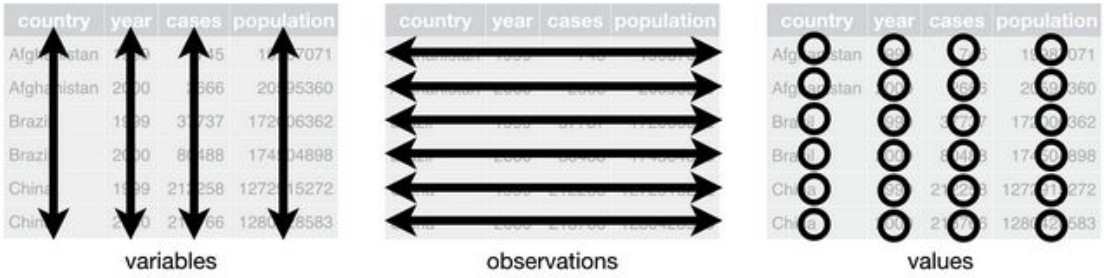
Muy volátiles.
Una instancia es un conjunto de elementos

Datos tabulares

Usualmente trabajamos con planillas, muchas veces aisladas.
Contienen instancias de BDs o datos propios

Hay ciertas propiedades deseables (*tidy data*):

- Cada columna un atributo
- Cada fila una observación o registro
- Valores atómicos en las celdas



Ciencia de Datos: un primer acercamiento

Next: Taller (Martes 5/7) - Sala Udelar A/B