ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

EDUCACIÓN PROFESIONAL

Programación en R para ciencia de datos

DBDC

Educación Profesional

Escuela de Ingeniería

Profesor:

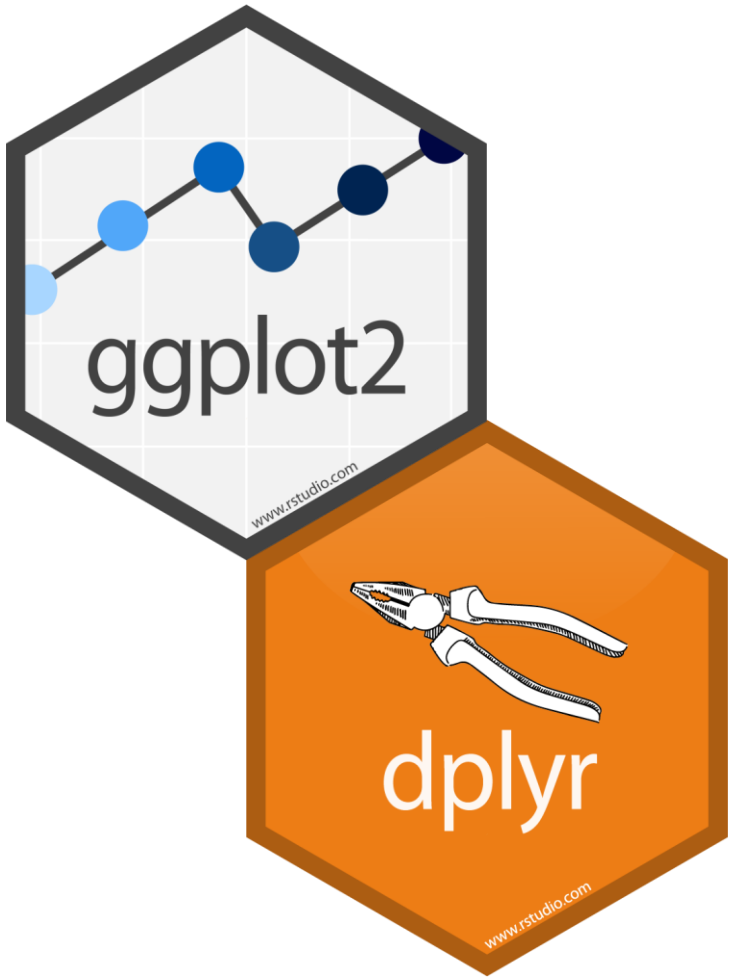
Miguel Jorquera Viguera

RESUMEN

Profesor: **Miguel Jorquera Viguera**

Manipulación de tablas

- Análisis exploratorio de datos
  - `dplyr`:
    - Manipulación de tablas
  - `ggplot2`:
    - Gramática de gráficos

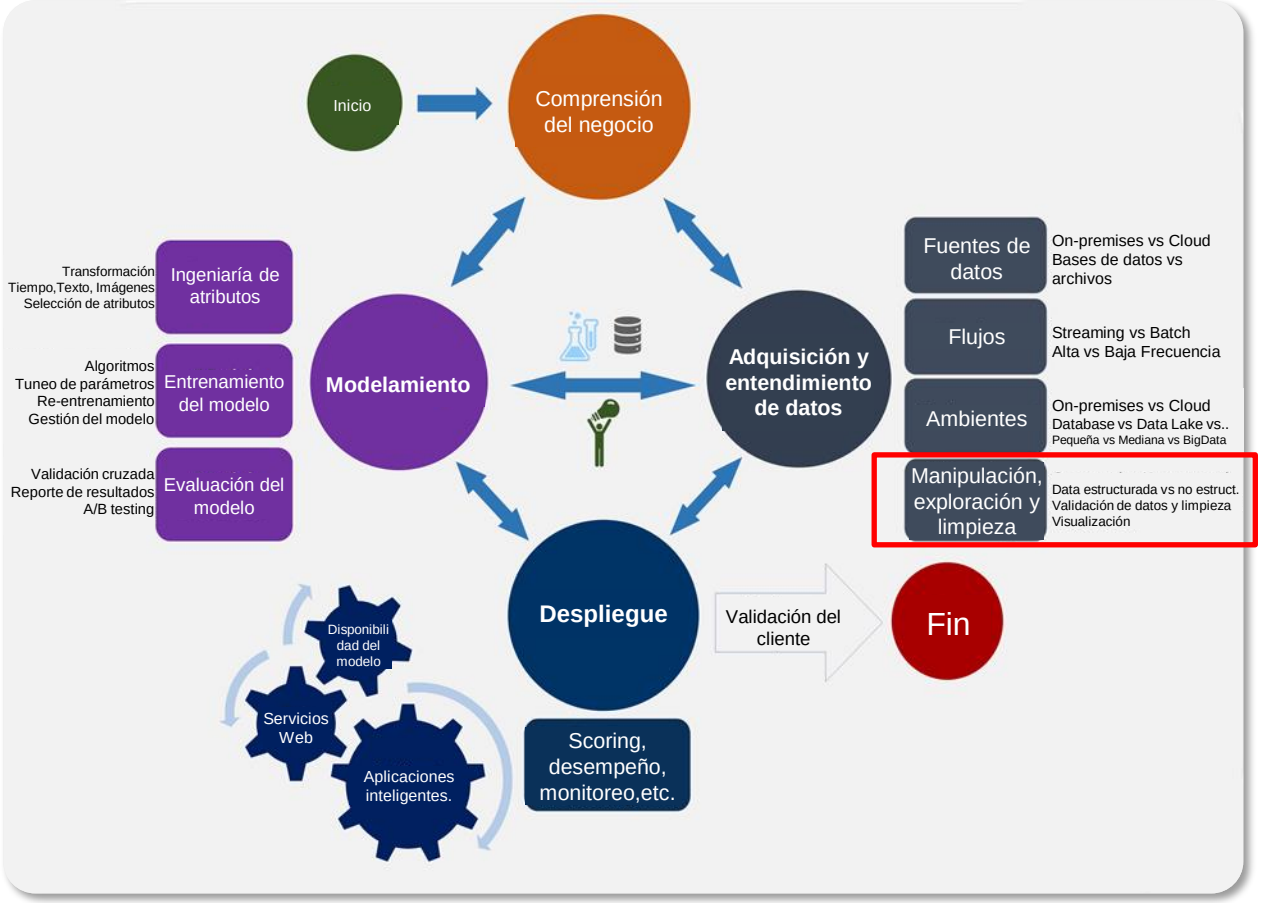


TEMAS PARA HOY

TDSP

La metodología TDSP (Team Data Science Process) propuesta por Microsoft es una metodología ágil, iterativa y eficiente, que promueve la colaboración entre los distintos miembros del equipo de desarrollo así como la interacción permanente con el cliente.

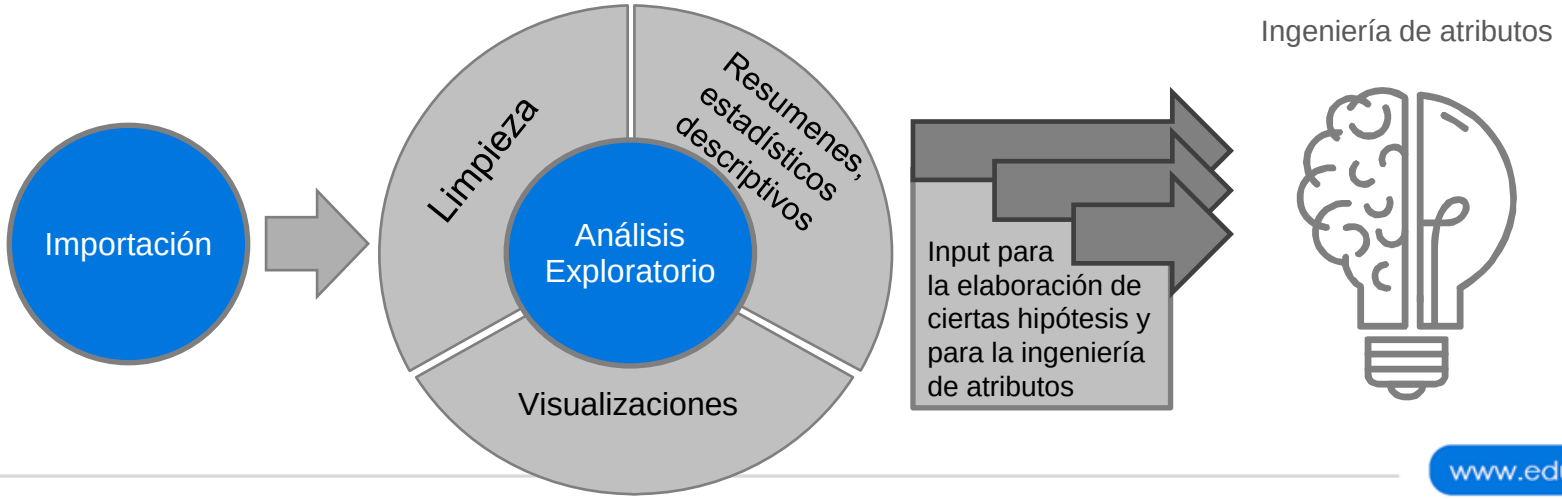
Flujo de trabajo en Data Science



Profesor: **Miguel Jorquera Viguera**

Manipulación, Exploración y Limpieza

- Esta fase la denominaremos como fase exploratoria. En ella se llevarán a cabo los siguientes procesos
  - Se valida la consistencia de los datos proporcionados
  - Se describen los datos importados a nivel estadístico y visual.
  - Se plantean hipotesis sobre las relaciones entre las variables existentes.
  - Se da paso a la ingeniería de atributos



Tareas usuales

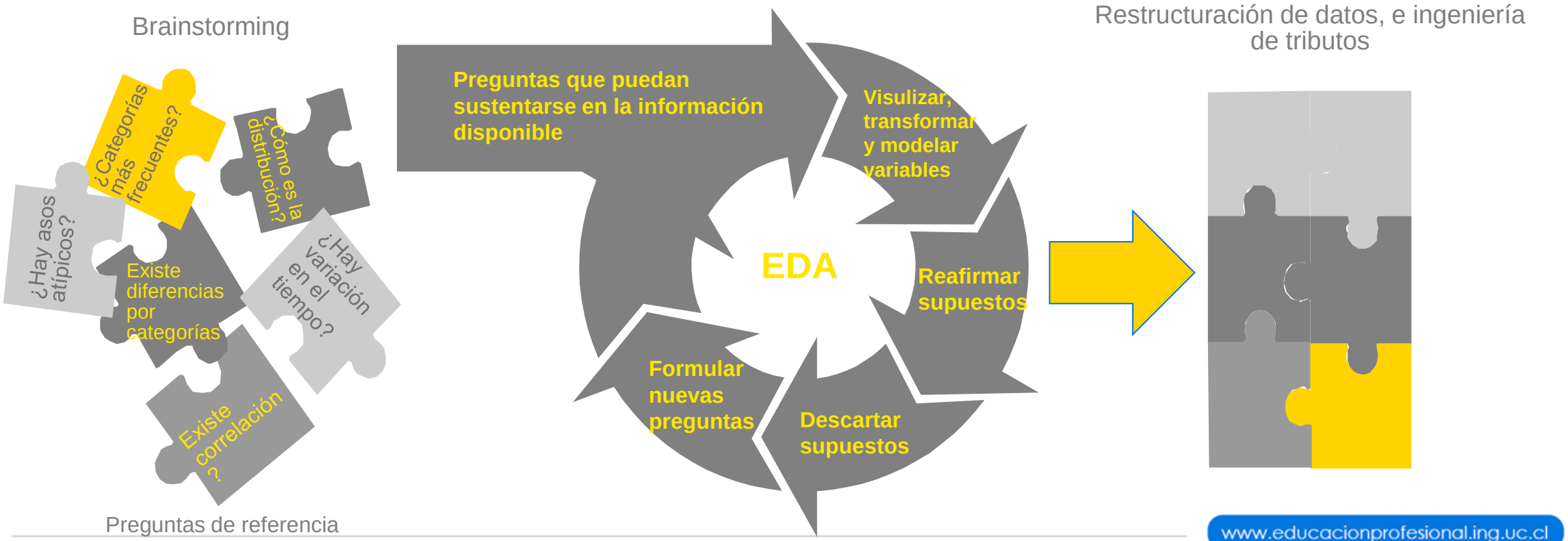
- Al importar datos a un nuevo ambiente (R en nuestro caso), es de utilidad chequear los siguientes aspectos
  - Nombres de las columnas en formato estándar.
  - Validar la consistencia de los tipos de dato de cada campo.
  - Verificar total de registros importados.
  - Analizar e imputar datos faltantes

Herramientas disponibles

- Si bien hay variedad de herramientas para llevar acabo la fase exploratoria, nosotros nos centraremos en la utilización de dos packages principalmente
  - `dplyr` para consultas
    - Generación de información agregada.
    - Tablas de frecuencia.
    - Facilita el cálculo de estadísticos descriptivos en general
  - `ggplot2` para visualización
    - Gráficos univariados y bivariados.
    - Visualización de variables continuas y categóricas
    - Gráficos de dispersión
    - Gráficos temporales
    - Visualización de distribuciones.

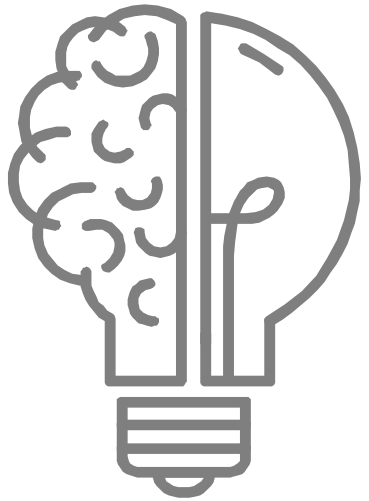
Proceso Iterativo

- La fase de análisis exploratorio, es un proceso iterativo que se caracteriza por la formulación de preguntas de interés que permitan guiar el análisis cuyas respuestas tenga sustento en la información disponible

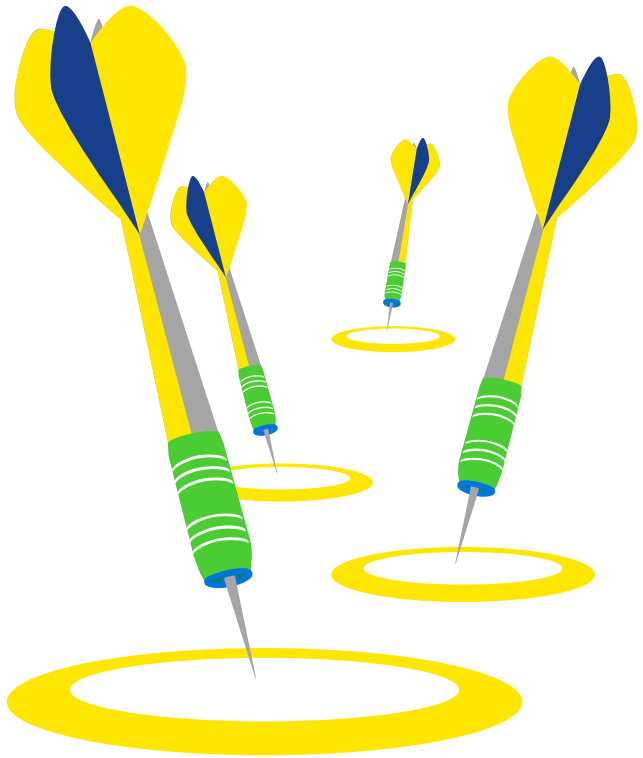


Características

- EDA no es un proceso formal con reglas estrictas. Es un **proceso creativo**!
  - Hay libertad de explorar toda idea inicial. Algunas llegará a punto muerto, otras serán dignas de ser comunicadas.
  - La limpieza es parte de la exploración. Podemos **visualizar**, **transformar** e incluso **modelar** en esta fase!.



Cómo guiar el análisis



- ¿Cómo generar buenas preguntas?
  - **Generando muchas preguntas!**
- ¿Cómo comenzar? Dos focos iniciales:
  - **¿Qué tipos de variaciones presentan mis variables?**
  - **¿Qué tipo de co-variación existe entre mis variables?**

Cómo guiar el análisis

- Para medir variaciones usualmente es de utilidad:
  - Histogramas y gráficos de barra.
  - Tablas de frecuencia
  - Polígonos de frecuencia
  - Boxplots y gráficos de violín

→
- Se caracterizan los valores típicos.
- Se observan posibles casos atípicos y anomalías.

- Para el caso de la co-variación
  - Medidas de correlación
  - Gráficos de dispersión
  - Tablas de contingencia

→
- Búsqueda de patrones
  - Tendencias
  - Clusters

→
- ¿Coincidencia?
- ¿Es posible describir el posible patrón?
- ¿Qué tan ”fuerte” es la relación de dependencia dada por el patrón?
- ¿Otras variables podrían afectar al patrón observado?
- ¿Cambia la relación si se observan subgrupos idividuales?

Algunos conceptos y funciones

- En R contamos con funciones para lo anterior:
  - `summary()`: por defecto entrega estadísticos de posición (cuartiles), min, max y media.
  - `quantiles()`: podemos calcular uno o varios percentiles de interés.
  - `mean()`: Calcula la media de un vector numérico.
  - `median()`: Calcula la mediana de un vector numérico.
  - `sd()`: Retorna la desviación estándar muestral.
  - `var()`: Retorna la varianza muestral.

- En particular, recordemos que la varianza se estima como:

$$Var(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Mientras que la desviación estándar corresponde a la raíz cuadrada de la varianza.
- Una manera de visualizar el grado de dispersión de un conjunto de datos, son los gráficos de cajas (boxplots) y los gráficos de dispersión.
- En particular, a través de un boxplot se pueden visualizar los cuartiles, los valores máximos y mínimos de un set de datos. (ya vimos ejemplos en el notebook de la clase 5)

Vamos!

Profesor: **Miguel Jorquera Viguera**