

# **Introducción al Data Science**

## **Pre-procesamiento y**

## **Visualización de Datos**

**INF-396**

**Prof: Juan G. Pavez S.**

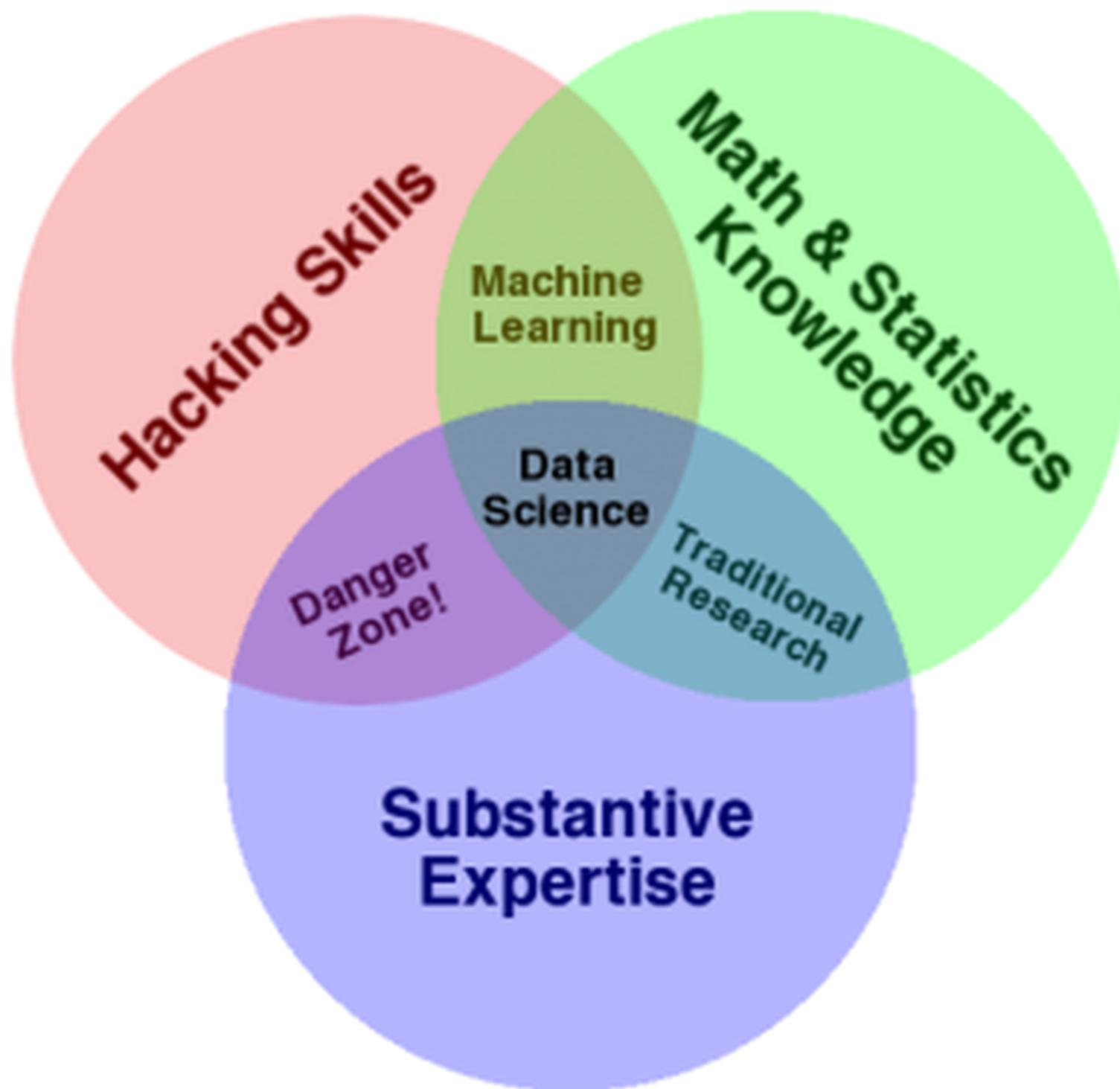
# Introducción a la Ciencia de Datos

“Data Scientists:  
The Definition of Sexy  
Forbes, 2012

“Data Scientists:  
The Sexiest Job of the 21st  
Century

Harvard Business Review, 2012

# Introducción a la Ciencia de Datos



# Introducción a la Ciencia de Datos

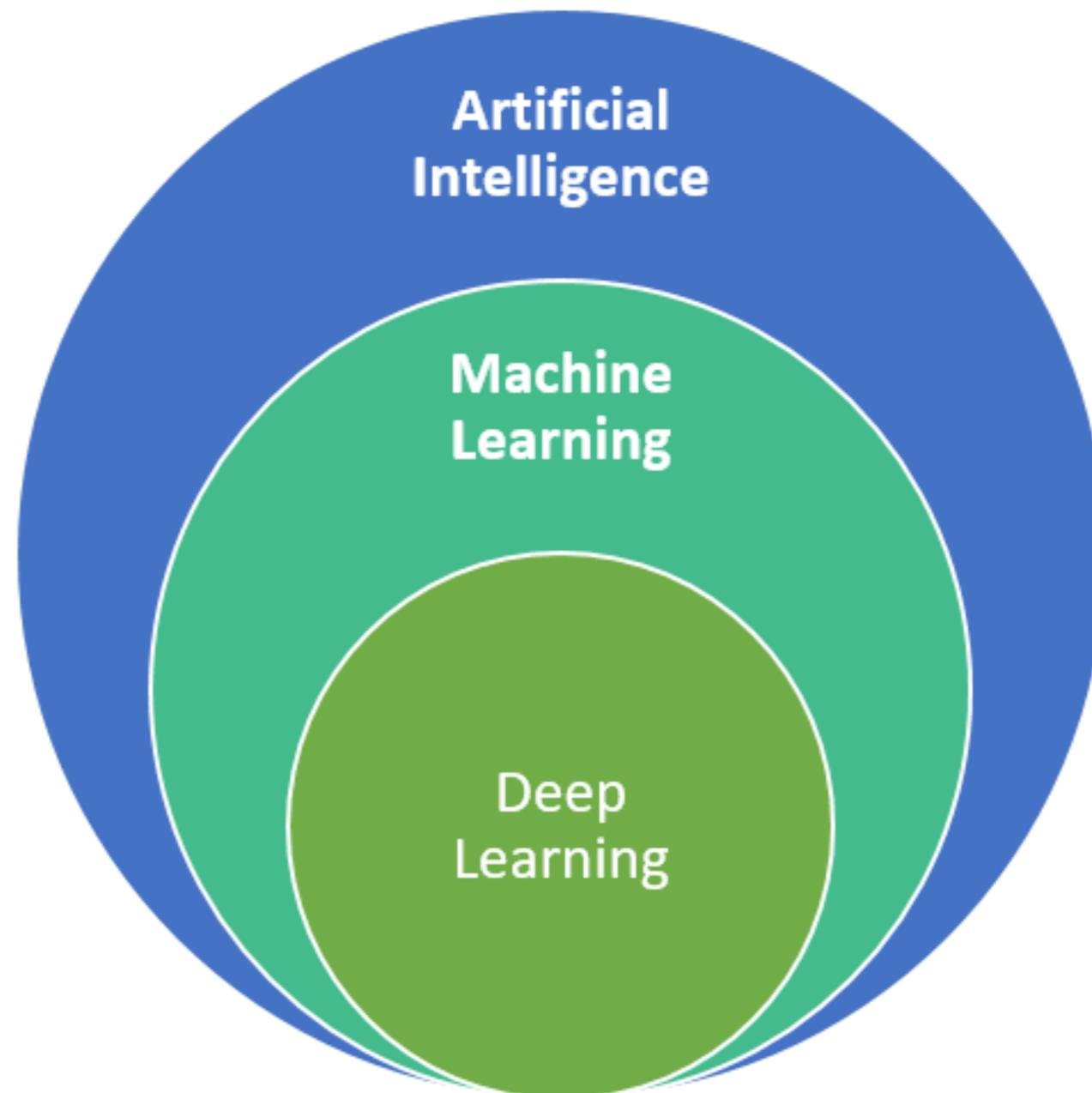
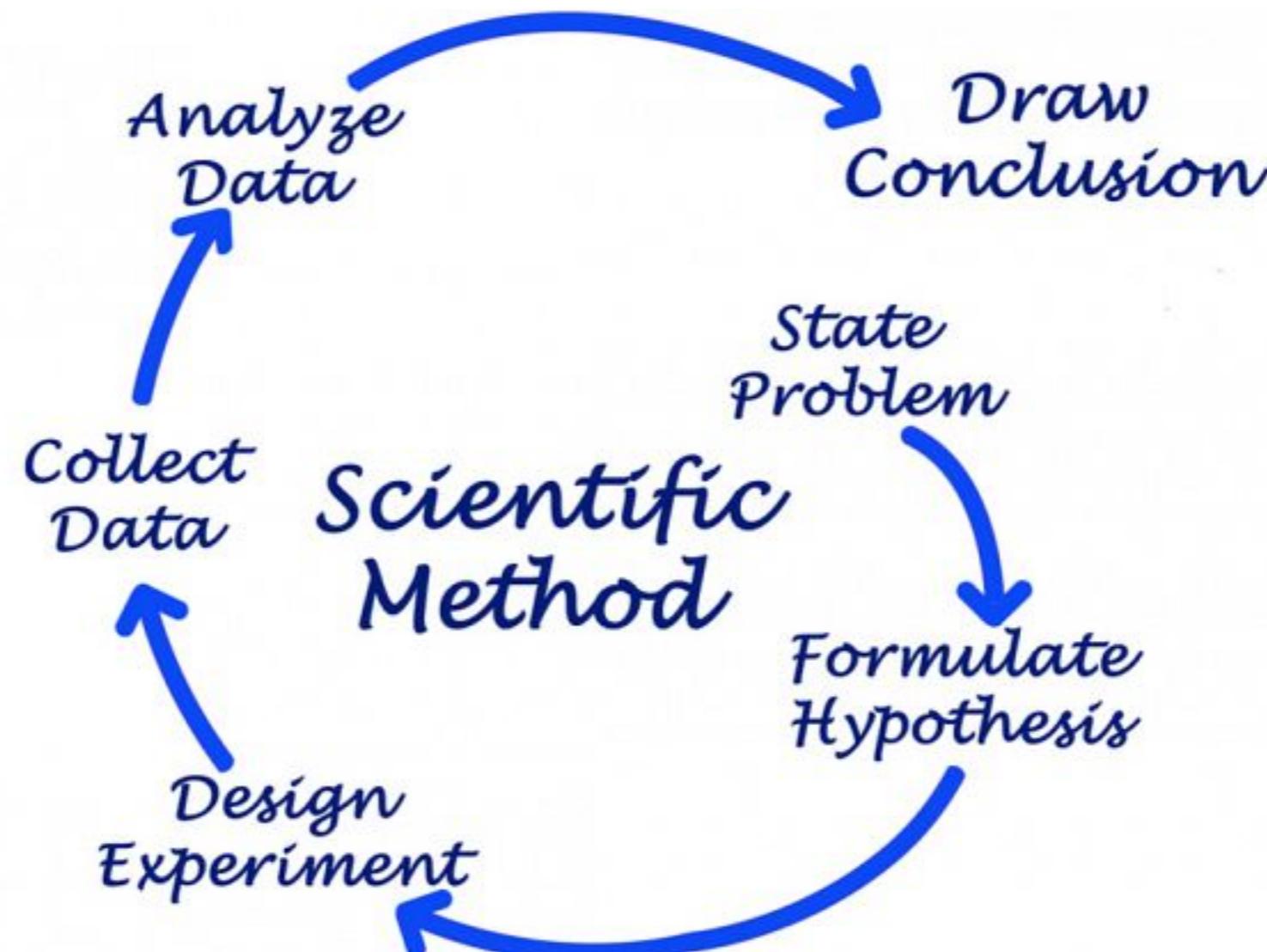
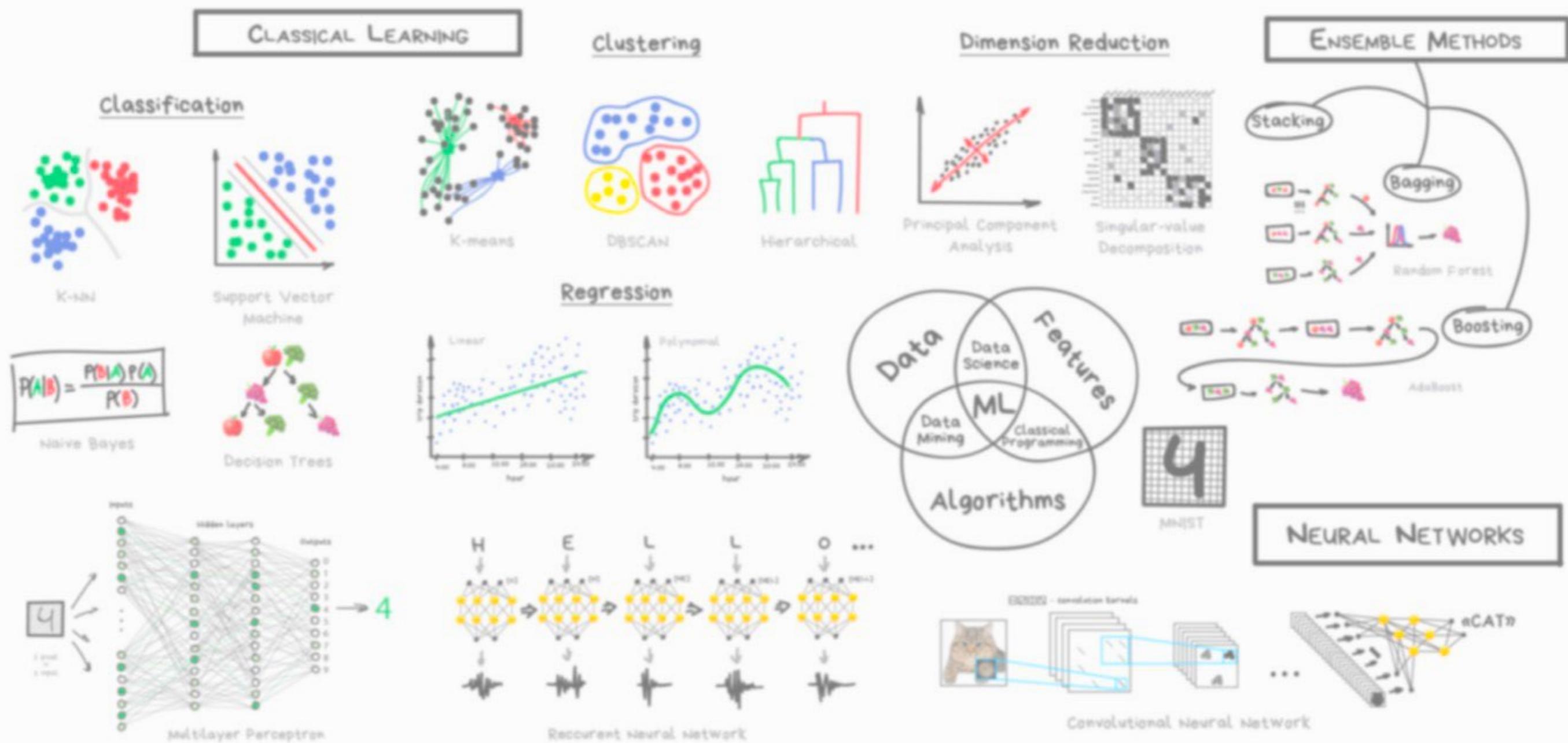


Figure 1: artificial intelligence, machine learning and deep learning Source: Nadia BERCHANE (M2 IESCI, 2018)

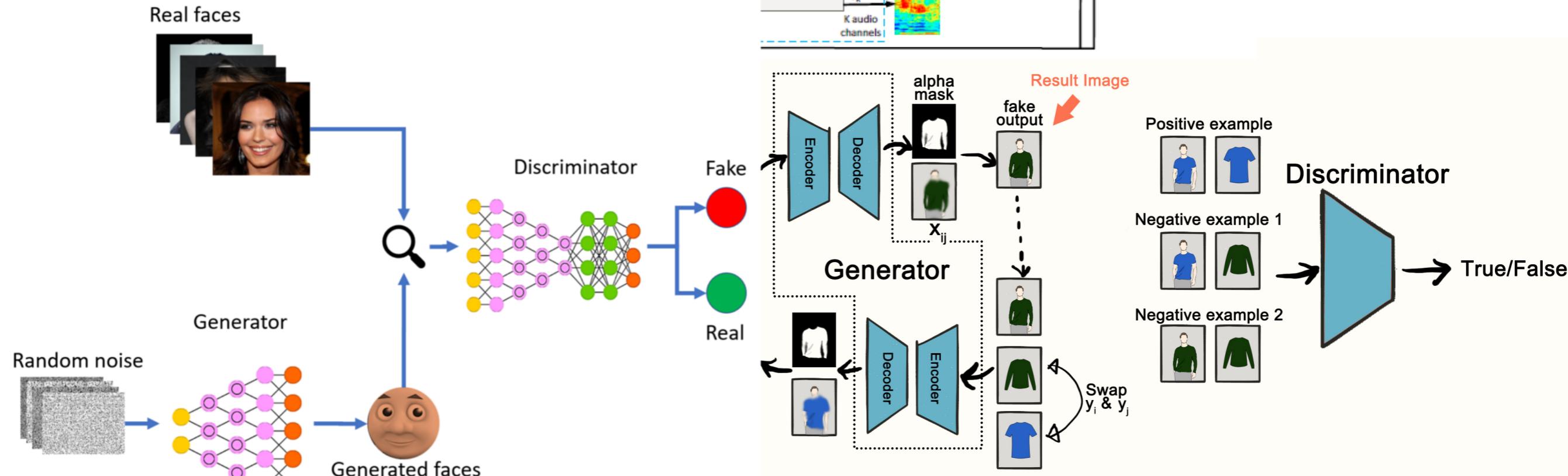
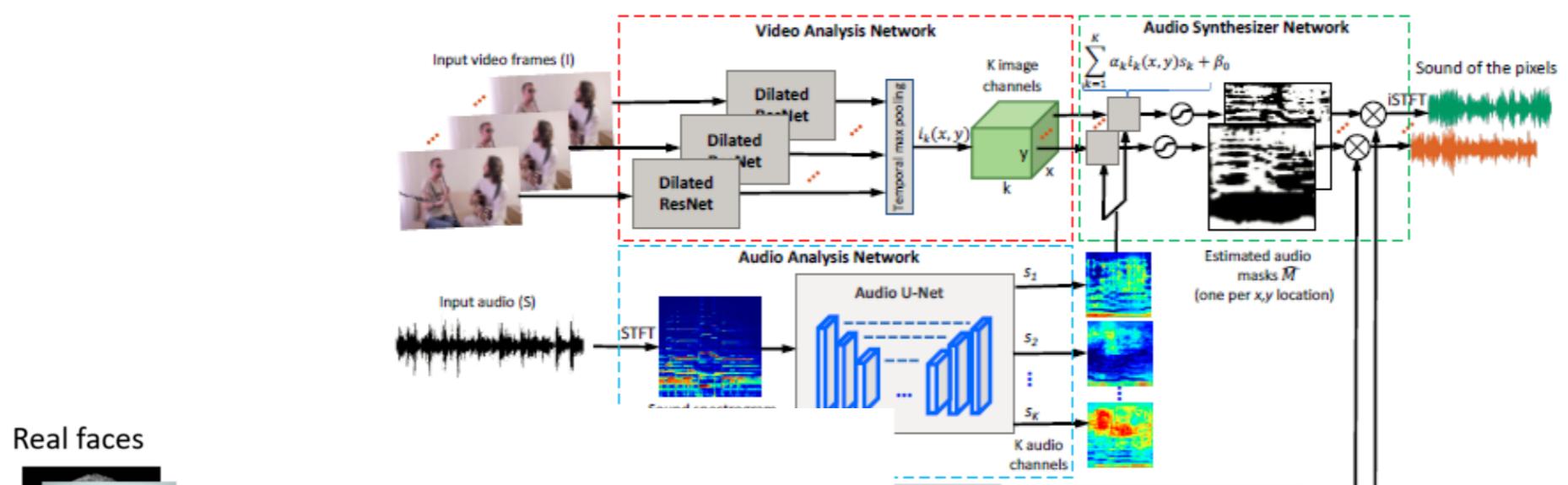
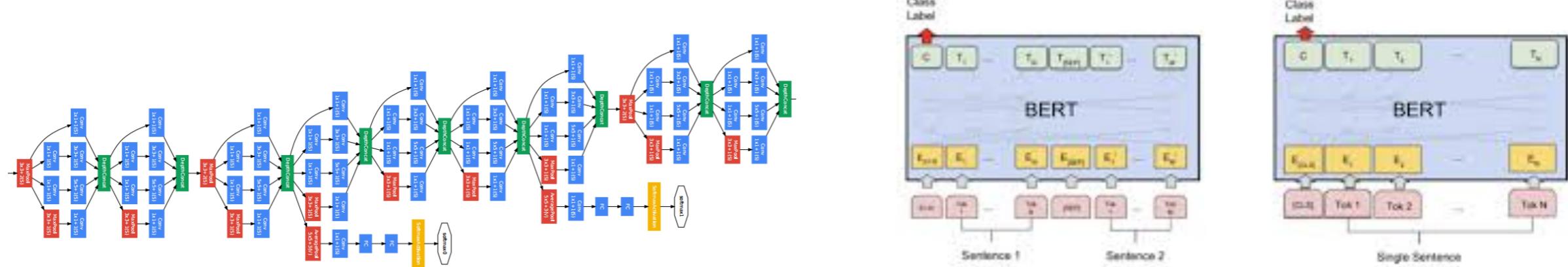
# Introducción a la Ciencia de Datos



# Introducción a la Ciencia de Datos



# Introducción a la Ciencia de Datos



# Introducción a la Ciencia de Datos



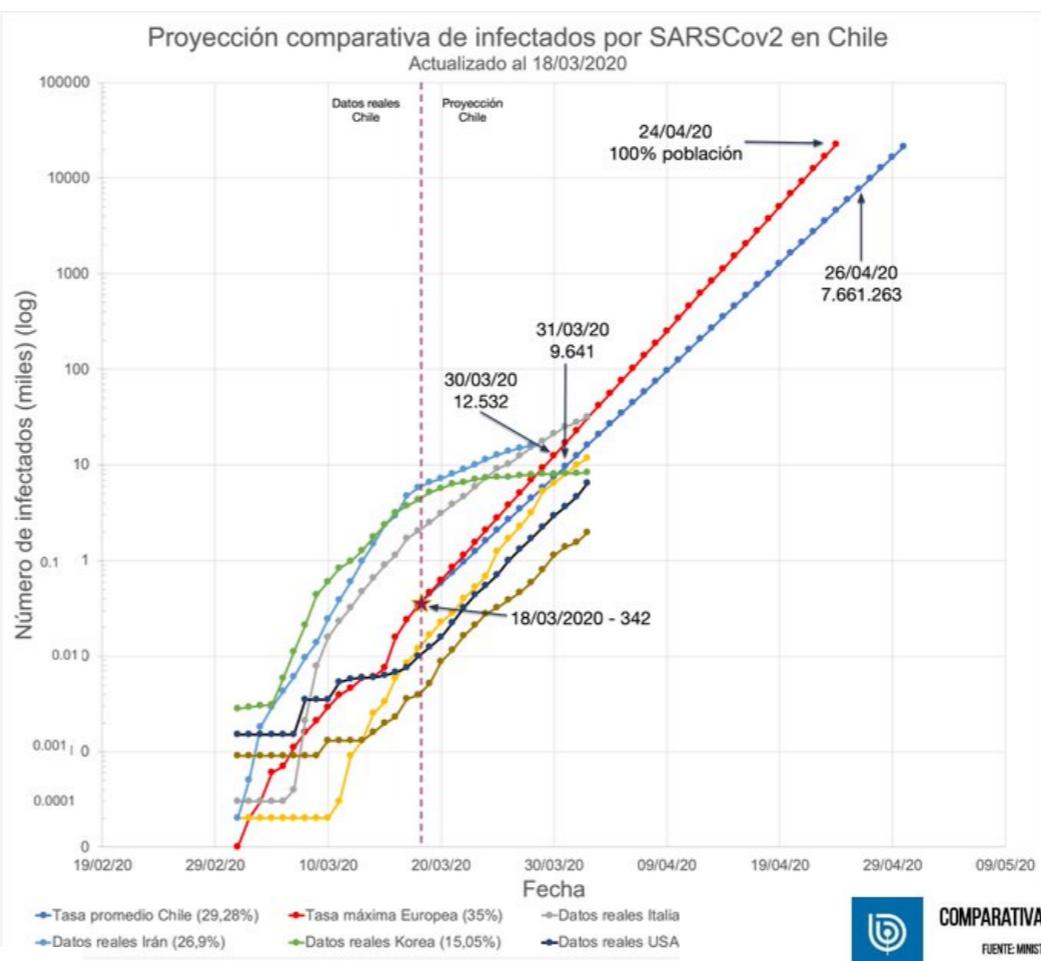
**Conclusión 4:** *El resto del país verá un escenario similar al de la Región Metropolitana.*

Así dado que la otra mitad del país vive fuera de la Región Metropolitana, uno espera ver los mismos 5,000 muertos durante los próximos 30 días.

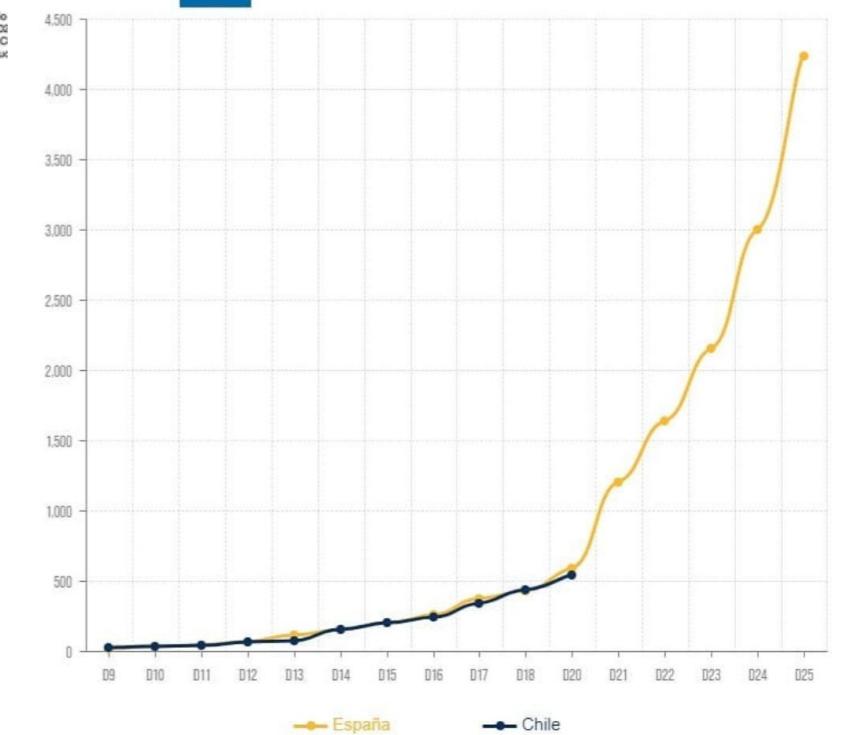
Así uno llega al macabro  $5,000 + 5,000 = 10,000$  muertos en los próximos 30 días.

Agregando un margen a mi pronóstico, **veo 10,000 muertos en Chile en los próximos 3 meses.**

Este pronóstico me da mucha rabia y tristeza. He indagado por lado y lado para ver qué se puede hacer para cambiar esto, pero las decisiones claves ya fueron tomadas.



COMPARATIVA CASOS CORONAVIRUS ESPAÑA - CHILE  
FUENTE: MINISTERIO DE SANIDAD - MINISTERIO DE SALUD - BIOBIOCHILE.CL



# El Pipeline del análisis de datos

## 1. Plantea una pregunta o problema para resolver.

¿Que harías si tuvieras todos los datos?.

¿Que se quiere predecir o estimar?.

¿Que valor entregará al negocio?.



## 2. Obtén los datos

¿Donde están los datos? (página gobierno, base de datos, notas escritas,...).

¿Qué datos son importantes?.

¿Cómo fueron obtenidos o muestreados?.

¿Políticas de privacidad a considerar?.



## 3. Explorar los datos - Limpiar los datos.

Mirar una submuestra de los datos.

Graficar los datos.

¿Hay patrones?, ¿Hay anomalías?.



## 4. Modelar los datos

Proponer un modelo y programarlo.

Ajustar el modelo a los datos.

Validar el modelo en datos de validación.



## 5. Comunicar los resultados y visualizar

¿Como funciona el modelo en datos fuera de muestra?

¿Que sentido podemos obtener de los resultados?

¿Que historia podemos contar con los resultados?

Construir visualizaciones, prototipos de prueba, tablas.

# El Pipeline del análisis de datos

## 1. Plantea una pregunta o problema para resolver.

¿Que harías si tuvieras todos los datos?.

¿Que se quiere predecir o estimar?.

¿Que valor entregará al negocio?.

Tiempos

?



## 2. Obtén los datos

¿Donde están los datos? (página gobierno, base de datos, notas escritas,...).

¿Qué datos son importantes?.

¿Cómo fueron obtenidos o muestreados?.

¿Políticas de privacidad a considerar?.

~ 50-70%

## 3. Explorar los datos - Limpiar los datos

Mirar una submuestra de los datos.

Graficar los datos.

¿Hay patrones?, ¿Hay anomalías?.



## 4. Modelar los datos

Proponer un modelo y programarlo.

Ajustar el modelo a los datos.

Validar el modelo en datos de validación.

~El resto

## 5. Comunicar los resultados y visualizar

¿Como funciona el modelo en datos fuera de muestra?

¿Que sentido podemos obtener de los resultados?

¿Que historia podemos contar con los resultados?

Construir visualizaciones, prototipos de prueba, tablas.

# Garbage In, Garbage Out



# Obtén los Datos

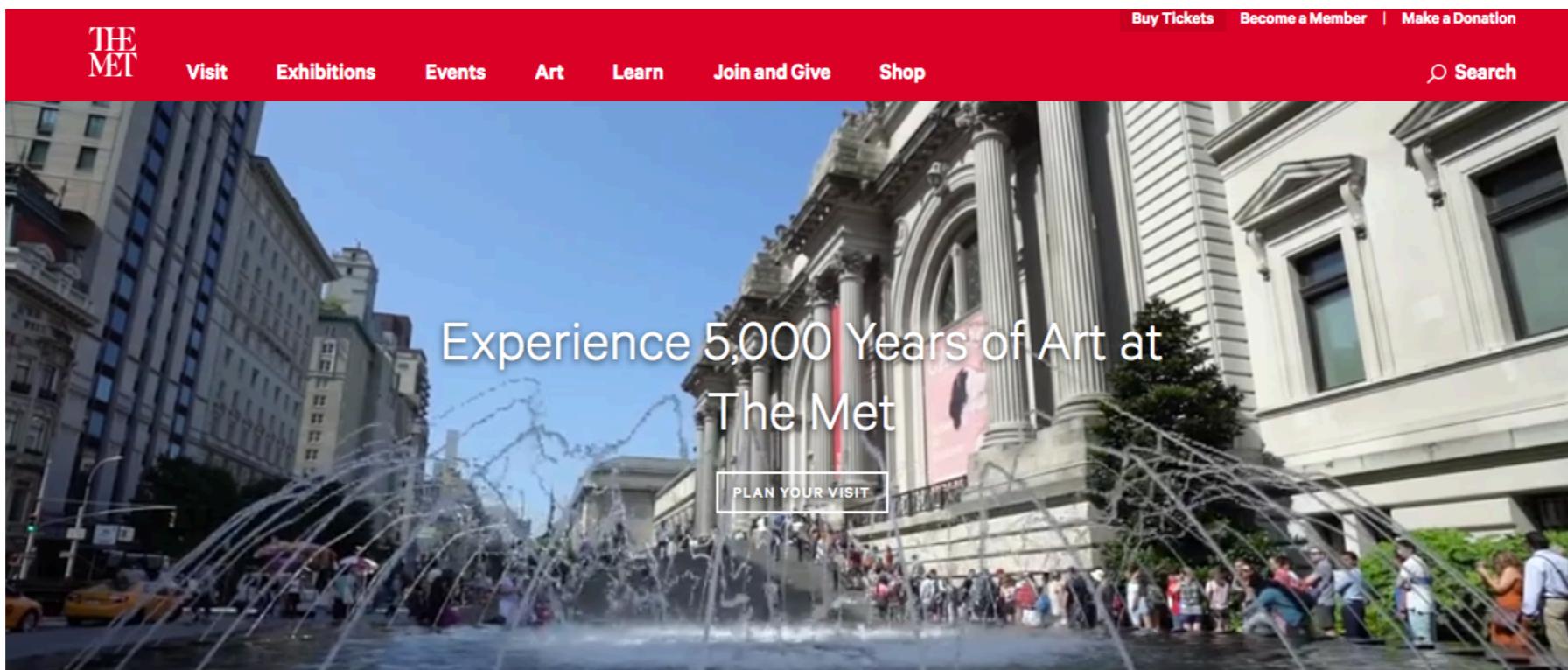
- ¿De donde vienen los datos?
  - **Fuentes internas:** Ya colecciónadas por la organización, por ejemplo, en data warehouses
  - **Fuentes externas existentes:** Ya disponibles en un formato fácil de leer en una fuente interna, por pago o gratis. Por ejemplo, fuentes gubernamentales ([datos.gob.cl](http://datos.gob.cl), [es.datachile.io](http://es.datachile.io)), base de datos de datasets ([toolbox.google.com/datasetsearch](http://toolbox.google.com/datasetsearch) , [kaggle.com](http://kaggle.com)), datos del mercado de valores, etcétera.
  - **Fuentes externas que requieren recolección:** Están en fuentes externas pero no procesadas, por ejemplo datos en web (requieren scrapping), datos en pdf (requieren ocr), datos impresos.

# Obtén los Datos

- ¿De donde obtener datos online?
  - **APIs (Application Programming Interface)**: Algunos sitios presentan sus datos a la web en forma de un conjunto de funciones que se pueden acceder por http. Por ejemplo: Google Map API, Facebook API, Twitter API.
  - **RSS (Rich Site Summary)**: Algunos sitios muestran su contenido resumido en este formato, comúnmente gratuito. Común en blogs y sitios de noticias.
  - **Web Scraping**: Muchos sitios no presentan sus datos de manera estructurada. Se pueden usar scripts o librerías para extraer automáticamente los datos analizando el HTML de la página web.

# Obtén los Datos

- APIs (Application Programming Interface):
- Ejemplo MET (NY)



<https://metmuseum.github.io/>

<https://collectionapi.metmuseum.org/public/collection/v1/objects/719664>

# Obtén los Datos

- APIs (Application Programming Interface):
  - Jupyter Notebook APIs cells

# Obtén los Datos

- **Web Scraping**
- La gran mayoría de los sitios no poseen APIs u otro método estructurado para acceder a sus datos.
- Aunque hacer scraping no es siempre ilegal, es necesario tomar en consideración las políticas y términos de servicio de cada sitio web:
  - ¿Estás evitando una API de pago?.
  - ¿Estás violando sus términos de servicio?. Es mejor revisar que se menciona sobre web scraping en los términos del sitio web.

# Obtén los Datos

- **Web Scraping**
  - **Jupyter Notebook Web Scraping cells**

# Obtén los Datos

- Formatos de datos comunes
  - **CSV (Comma-Separated Values)**: Los valores están separados en columnas separadas por comas (u otro delimitador) y filas separadas por salto de linea. Comúnmente la primera es la cabecera. Útil porque puede ser exportado a excel (bueno para comunicarse con no programadores).

```
Title,Author,ISBN13,Pages
1984,George Orwell,978-0451524935,268
Animal Farm,George Orwell,978-0451526342,144
Brave New World,Aldous Huxley,978-0060929879,288
Fahrenheit 451,Ray Bradbury,978-0345342966,208
Jane Eyre,Charlotte Brontë,978-0142437209,532
Wuthering Heights,Emily Brontë,978-0141439556,416
Agnes Grey,Anne Brontë,978-1593083236,256
Walden,Henry David Thoreau,978-1420922615,156
Walden Two,B. F. Skinner,978-0872207783,301
"Eats, Shoots & Leaves",Lynne Truss,978-1592400874,209
```

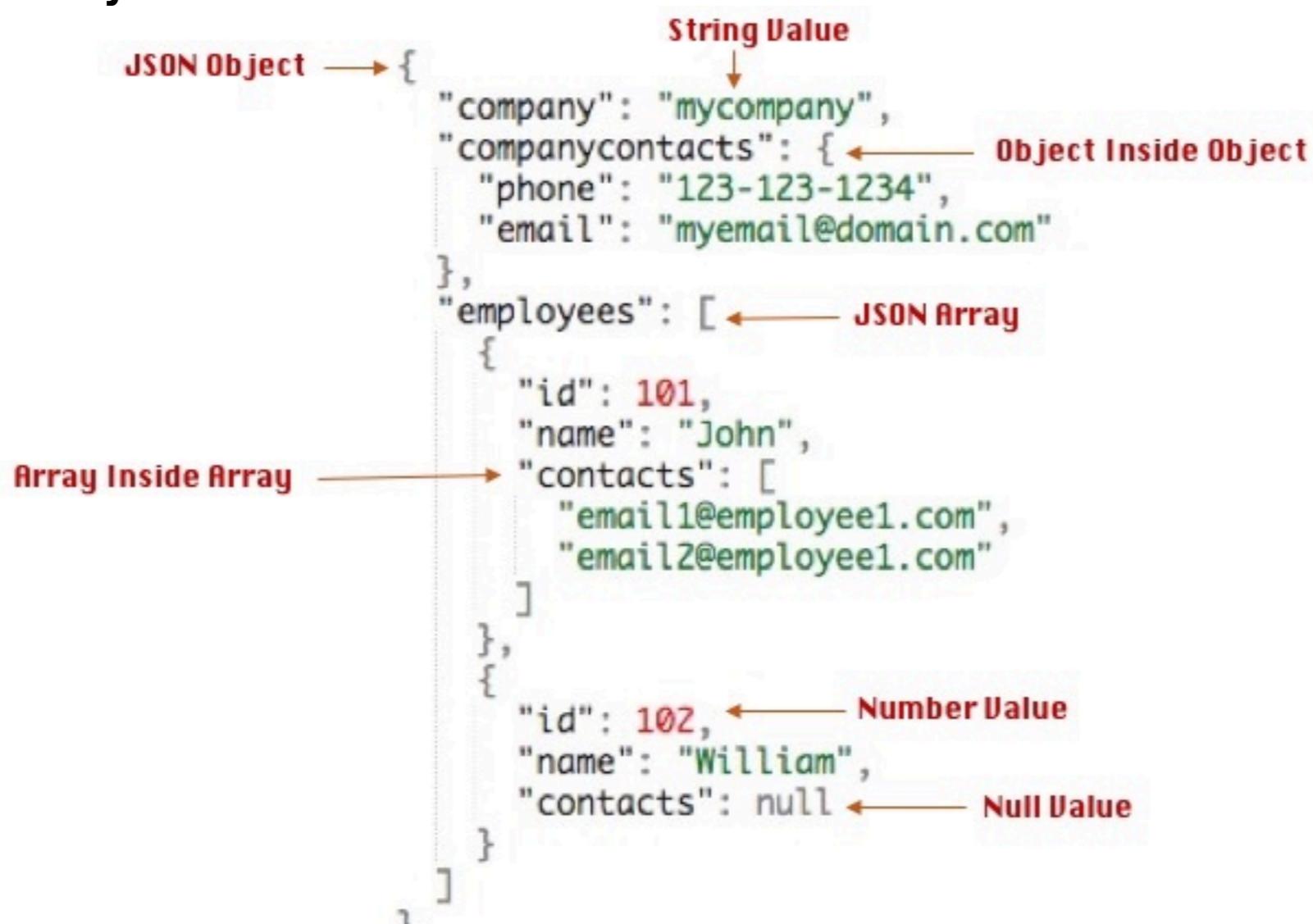
# Obtén los Datos

- Formatos de datos comunes
  - **XML (Extensible Markup Language)**: Lenguaje de etiquetas, cada dato esta delimitado por etiquetas. No es muy recomendable para grandes cantidades de datos debido a todo el espacio que ocupan las etiquetas. Un caso importante es el **HTML (xHTML)**.

```
<?xml version="1.0" encoding="UTF-8"?>
<events>
    <event>
        <week>-mtwtf-</week>
        <starttime>06:00</starttime>
        <endtime>11:59</endtime>
    </event>
    <event>
        <week>SMTWTFS</week>
        <starttime>12:00</starttime>
        <endtime>20:59</endtime>
    </event>
</events>
```

# Obtén los Datos

- Formatos de datos comunes
  - **JSON (Javascript Object Notation)**: Más ligero que XML y permite definir arreglos. En **JSONL** cada linea es un objeto JSON.



# Explorar - Limpiar los Datos

- En la práctica, los datos casi nunca vienen listos para el análisis.
- Hay varios problemas que se pueden originar de varias fuentes y que afectan el correcto procesamiento de los datos.
  - Valores perdidos.
  - Valores erróneos
  - Valores no usables.
  - Formatos difíciles de analizar.
  - Formatos dispares.

# Explorar - Limpiar los Datos

- Es difícil definir reglas para realizar la exploración y limpieza de datos, ya que depende mucho de cada caso, algunas guidelines:
  - **Considerar el contexto del negocio:** ¿Cómo se recolectaron los datos?, ¿Se puede entender en base a esto los datos que faltan?, ¿y los datos erróneos?.
  - **Mirar los datos directamente:**
    - **Mirar algunos datos individuales:** Tomar una submuestra y mirar uno por uno ¿Se puede identificar un patrón de error recurrente? (Excel sirve para esto).
    - **Mirar propiedades globales:** Analizar estadísticas de sumario como la media y la varianza por columna.
    - **Mirar propiedades por grupo:** Analizar estadísticas de sumario en grupos. Por ejemplo, estadísticas para compradores y para no compradores.

# Explorar - Limpiar los Datos

- Es difícil definir reglas para realizar la exploración y limpieza de datos, ya que depende mucho de cada caso, algunas guidelines:
  - **Graficar los datos:** Graficar los datos ya sea a nivel global o por grupos. Los gráficos pueden ayudar a identificar outliers, identificar variables dependientes, entre otros.
    - Histogramas
    - Gráficos de dispersión (scatter plots).
    - Gráficos de series de tiempos para datos temporales.

# Explorar - Limpiar los Datos

Para explorar los datos Python es de gran ayuda. En especial la librería pandas.

VERB	dplyr	pandas	SQL
QUERY/SELECTION	filter() (and slice())	query() (and loc[], iloc[])	SELECT WHERE
SORT	arrange()	sort()	ORDER BY
SELECT-COLUMNS/PROJECTION	select() (and rename())	(and rename())	SELECT COLUMN
SELECT-DISTINCT	distinct()	unique(), drop_duplicates()	SELECT DISTINCT COLUMN
ASSIGN	mutate() (and transmute())	assign	ALTER/UPDATE
AGGREGATE	summarise()	describe(), mean(), max()	None, AVG(), MAX()
SAMPLE	sample_n() and sample_frac()	sample()	implementation dep, use RAND()
GROUP-AGG	group_by/summarize	groupby/agg, count, mean	GROUP BY
DELETE	?	drop/masking	DELETE/WHERE

# **Explorar - Limpiar los Datos**

**Jupyter Notebook EDA cells**

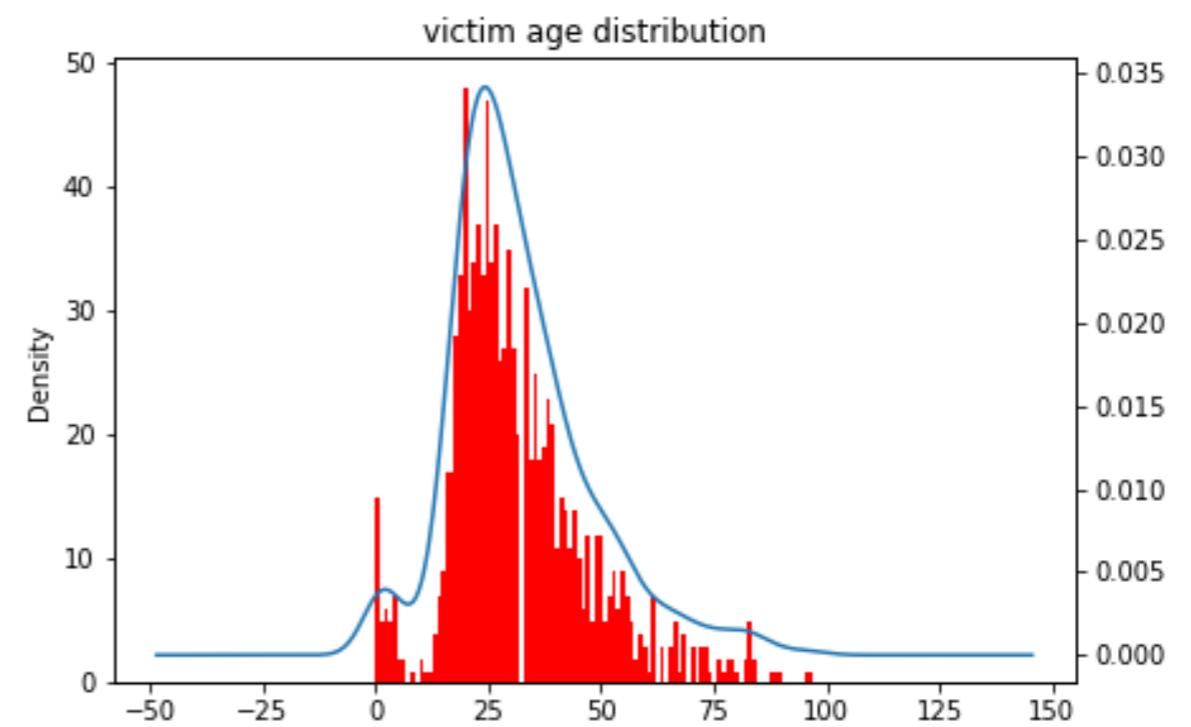
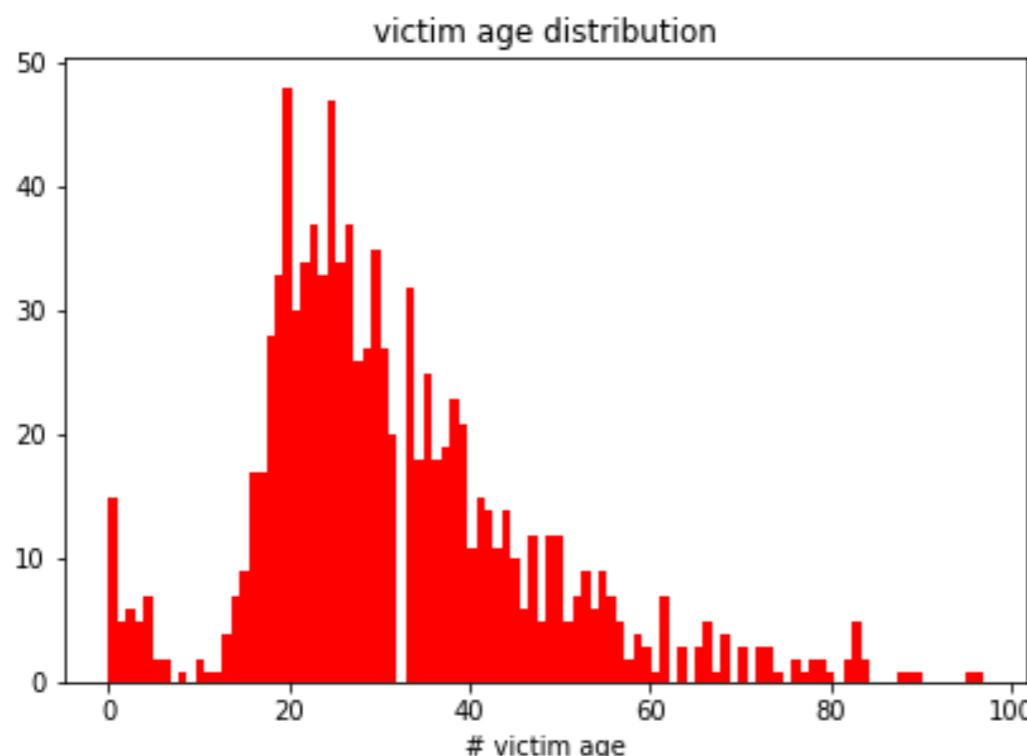
# Explorar - Limpiar los Datos

- La visualización de datos puede tener dos propósitos:
  - **Explorativo:**
    - Explora los datos.
    - Entiende la situación.
    - Determina cómo proceder.
    - Determinar que hacer.
  - **Explicativo** (veremos esto en más profundidad luego):
    - Presentar datos e ideas.
    - Explicar e informar.
    - Proveer evidencia.
    - Influir y persuadir.

# Explorar - Limpiar los Datos

- **Histogramas**

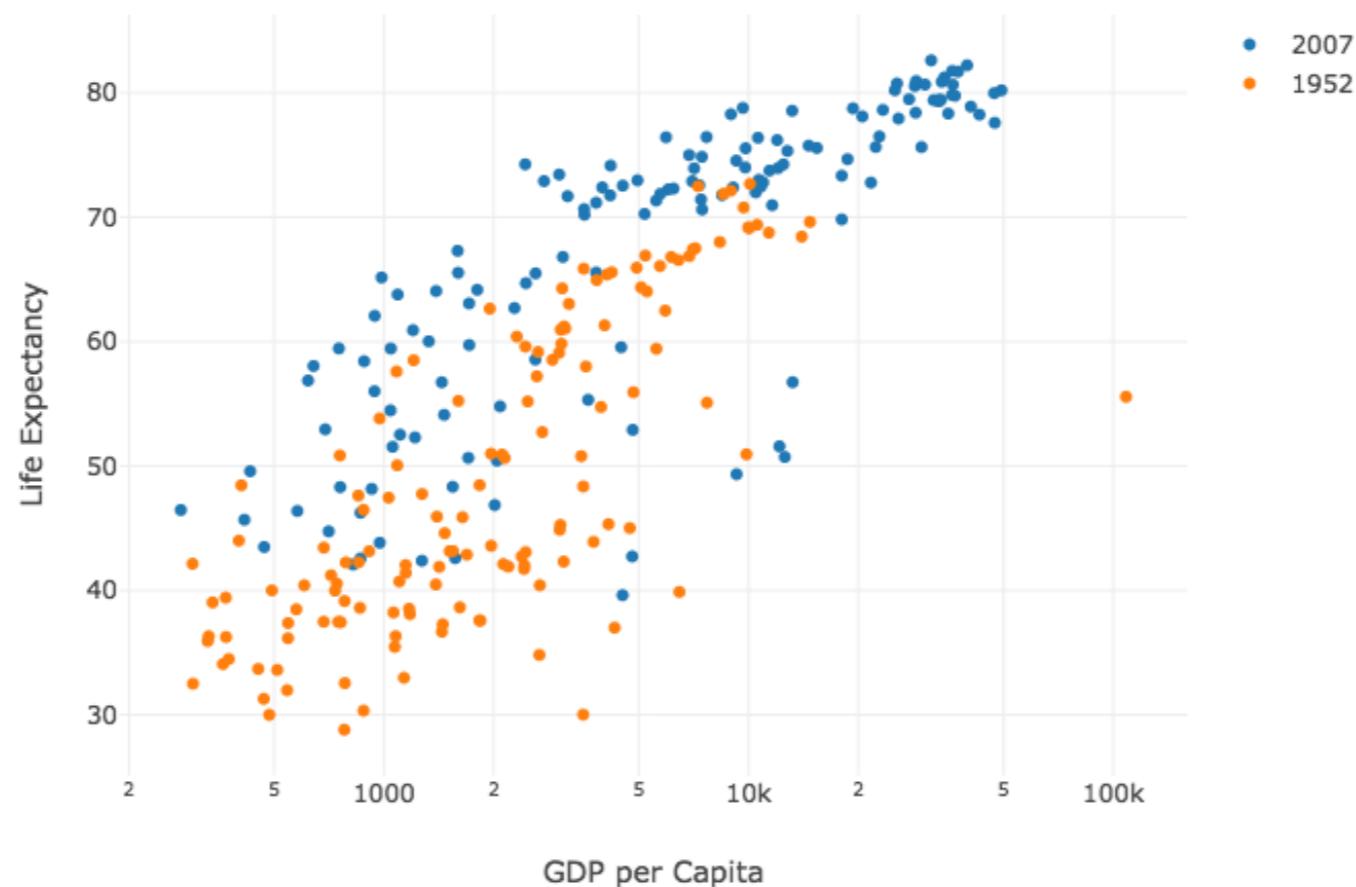
- Útiles para visualizar distribuciones unidimensionales. Permite hipotetizar distribuciones, encontrar outliers, identificar valores erróneos.
- KDE (kernel density estimation) permite ajustar un aproximador simple al histograma, para entregar gráficos más suaves.



# Explorar - Limpiar los Datos

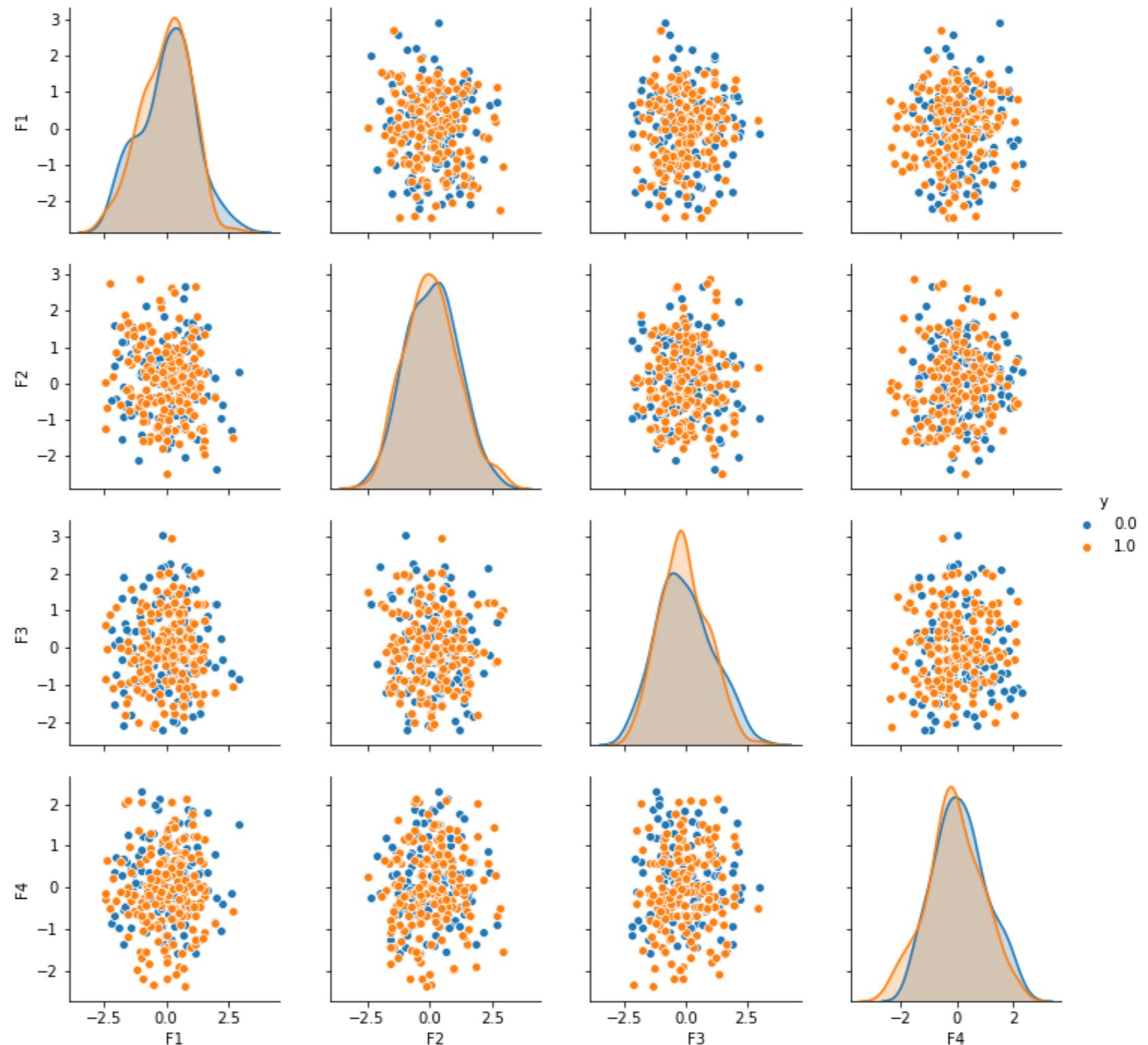
- **Scatter plots**

- Útiles para visualizar pares de variables. Permite identificar variables correlacionadas.



# Explorar - Limpiar los Datos

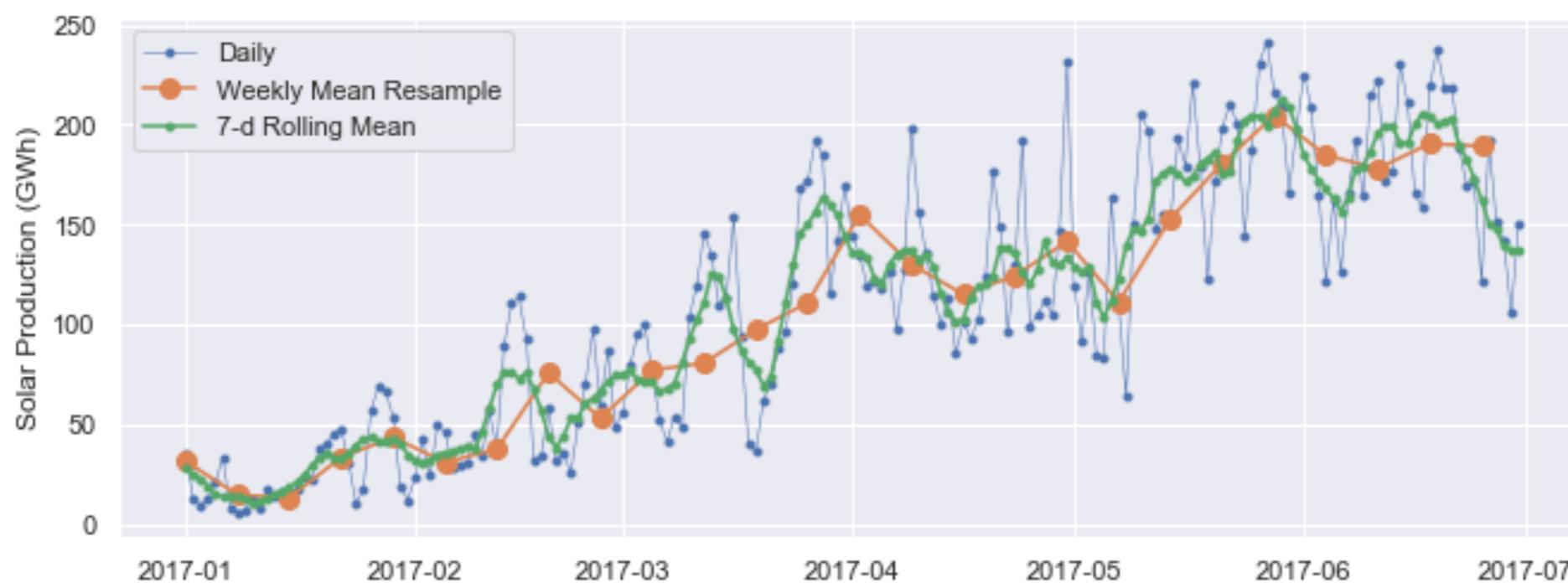
- **Pairplots**



# Explorar - Limpiar los Datos

- **Gráficos de Series de Tiempo**

- Muy útiles si se trabaja con datos temporales (datos del clima, datos de acciones, consumos eléctricos, etcétera).
- Permiten identificar tendencias y estacionalidades.
- Otros gráficos muy útiles para estos datos: **Correlaciones y Correlaciones Parciales.**



# **Explorar - Limpiar los Datos**

**Jupyter Notebook EDA Visualization cells**

# Explorar - Limpiar los Datos

- La siguiente fase corresponde a limpiar los datos. Es una fase **extremadamente importante**. Malos datos producen malos resultados in importar el modelo.
- Nuevamente no hay una formula general, ya que depende del **tipo de datos** (texto, valores categóricos, valores reales, imágenes), del **negocio** (datos tomados a mano, datos en formulario, planillas excel), del **modelo usado** (no todos los modelos necesitan transformación de escalas) y de varios factores más.
- Hay algunas cosas que se repiten:
  - Imputación de datos.
  - Transformación de escalas.
  - One-hot encoding de variables categóricas.
  - Selección de features (lo veremos al revisar **PCA**).
  - Traslaciones y rotaciones de imágenes, cropping de imágenes, one hot encoding de texto, tf-idf encoding de texto, entre muchas otras.

# Explorar - Limpiar los Datos

- **Datos perdidos**
- **Missing completely at Random (MCAR):** Los valores aparecen perdidos de manera completamente aleatoria (Sólo en este caso podemos usar las técnicas que estudiaremos)
- **Missing at Random (MAR):** (el nombre confunde un poco) Los valores que aparecen perdidos dependen de alguna otra variable.
- **Not Missing at Random or Nonignorable (NMAR):** Los valores perdidos dependen del valor de la variable.

# Explorar - Limpiar los Datos

- **Eliminación de casos:** Una opción es eliminar entradas en caso de que existan valores perdidos:
  - **Listwise:** Eliminar los ejemplos que tengan al menos una variable perdida.
  - **Pairwise:** Eliminar el ejemplo sólo si la variable se usará en el análisis.
- Otra opción es eliminar una **columna o feature completa** del análisis en caso de que existan muchos casos perdidos.

# Explorar - Limpiar los Datos

Nombre	D o I	Avg	Fecha Nac.	Fecha Muerte	Altura	Peso
Alexis Vidal	D	0.33	12/03/1991	NA	1.8	80
Arturo Cuevas	I	0.1	03/09/1989	NA	NA	75
Gary Vidal	D	0.03	05/04/1992	03/07/1990	1.75	72
Eduardo Bravo	D	0.2	03/12/1980	NA	1.82	NA
Guillermo Vargas	I	0.01	02/02/1970	17/12/2009	1.70	68

# Explorar - Limpiar los Datos

Listwise

Nombre	D o I	Avg	Fecha Nac.	Fecha Muerte	Altura	Peso
Alexis Vidal	D	0.33	12/03/1991	NA	1.8	80
Arturo Cuevas	I	0.1	03/09/1989	NA	NA	75
Gary Vidal	D	0.03	05/04/1992	03/07/1990	1.75	72
Eduardo Bravo	D	0.2	03/12/1980	NA	1.82	NA
Guillermo Vargas	I	0.01	02/02/1970	17/12/2009	1.70	68

# Explorar - Limpiar los Datos

Pairwise (queremos predecir el peso con la altura)

Nombre	D o I	Avg	Fecha Nac.	Fecha Muerte	Altura	Peso
Alexis Vidal	D	0.33	12/03/1991	NA	1.8	80
Arturo Cuevas	I	0.1	03/09/1989	NA	NA	75
Gary Vidal	D	0.03	05/04/1992	03/07/1990	1.75	72
Eduardo Bravo	D	0.2	03/12/1980	NA	1.82	NA
Guillermo Vargas	I	0.01	02/02/1970	17/12/2009	1.70	68

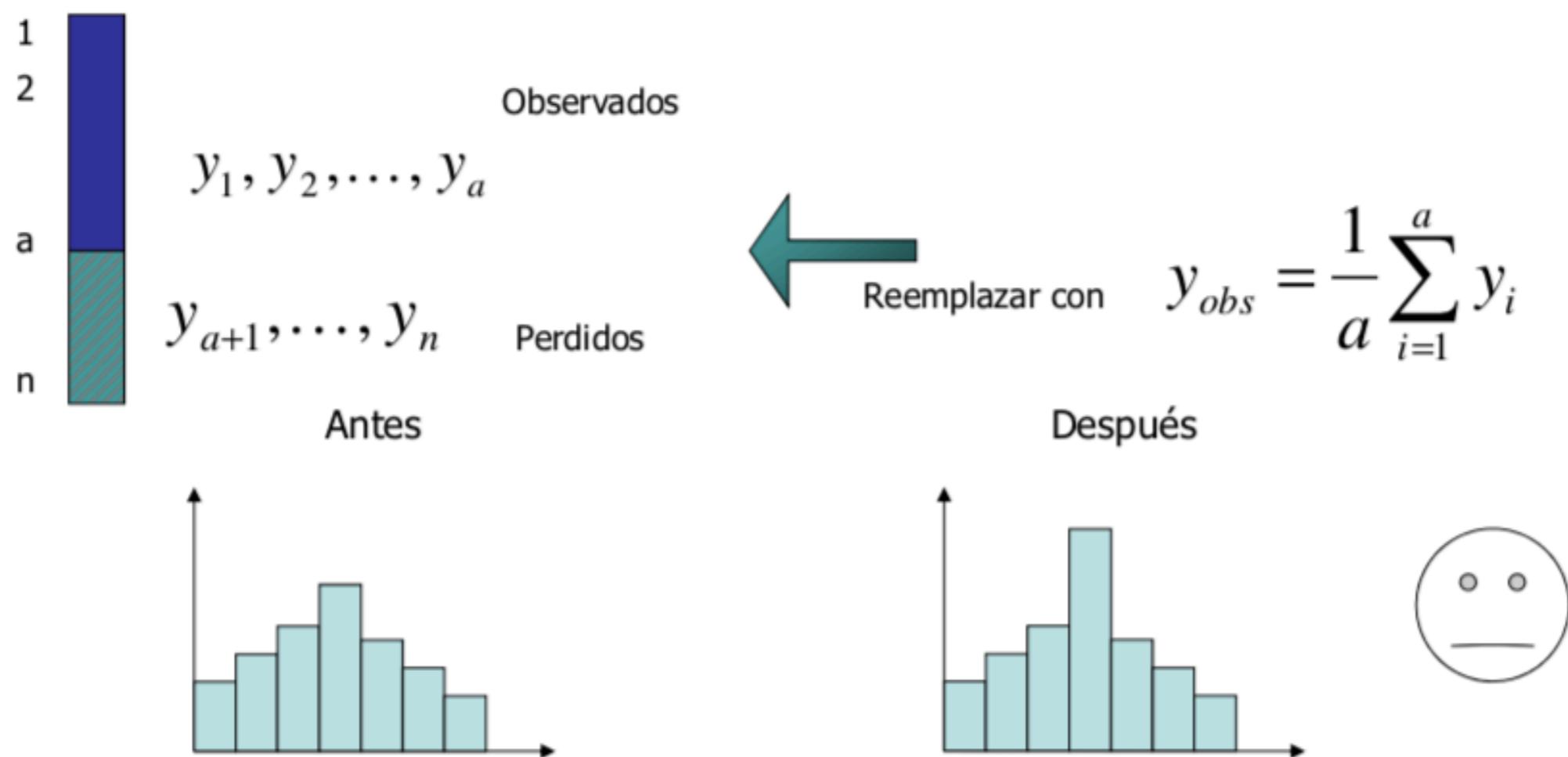
# Explorar - Limpiar los Datos

Columna

Nombre	D o I	Avg	Fecha Nac.	Fecha Muerte	Altura	Peso
Alexis Vidal	D	0.33	12/03/1991	NA	1.8	80
Arturo Cuevas	I	0.1	03/09/1989	NA	NA	75
Gary Vidal	D	0.03	05/04/1992	03/07/1990	1.75	72
Eduardo Bravo	D	0.2	03/12/1980	NA	1.82	NA
Guillermo Vargas	I	0.01	02/02/1970	17/12/2009	1.70	68

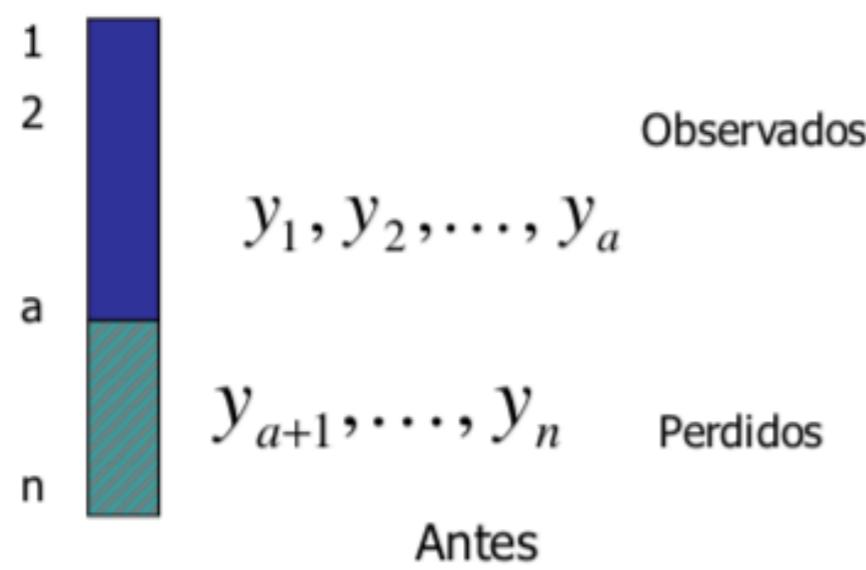
# Explorar - Limpiar los Datos

- **Imputación:** Otra opción es reemplazar los datos desconocidos en base al conocimiento que podemos derivar de los datos presentes.
- **Imputación por la media (mediana o moda):**



# Explorar - Limpiar los Datos

- **Imputación:** Otra opción es reemplazar los datos desconocidos en base al conocimiento que podemos derivar de los datos presentes.
- **Imputación por la media (mediana o moda):**



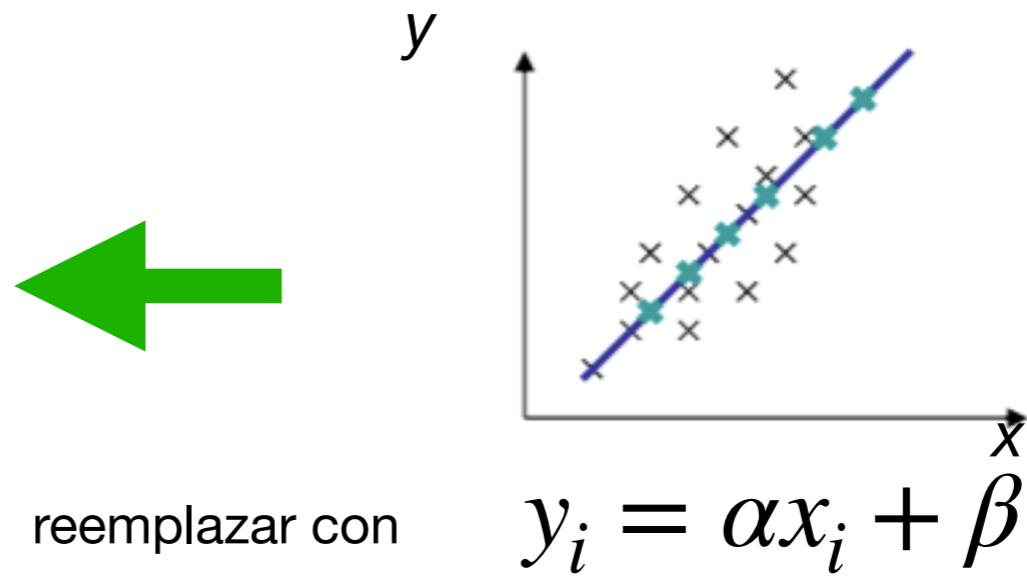
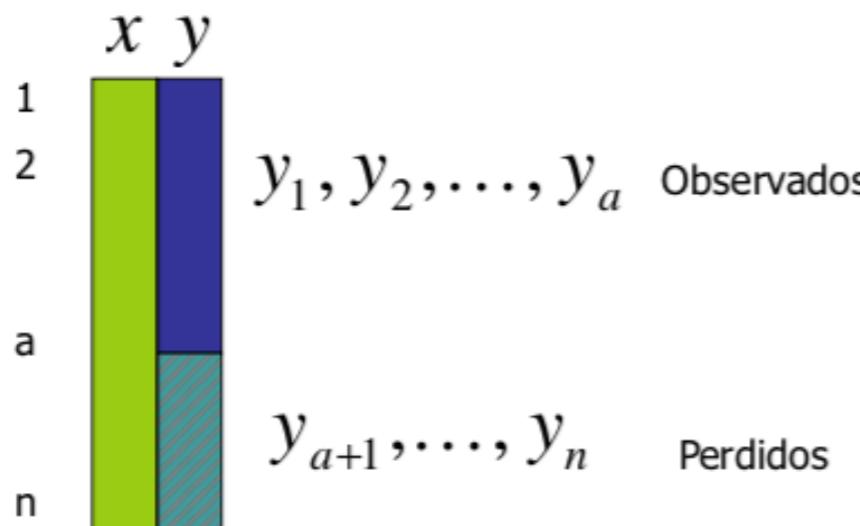
- Corrompe distribución
- Cambia correlaciones

$$y_{obs} = \frac{1}{a} \sum_{i=1}^a y_i$$



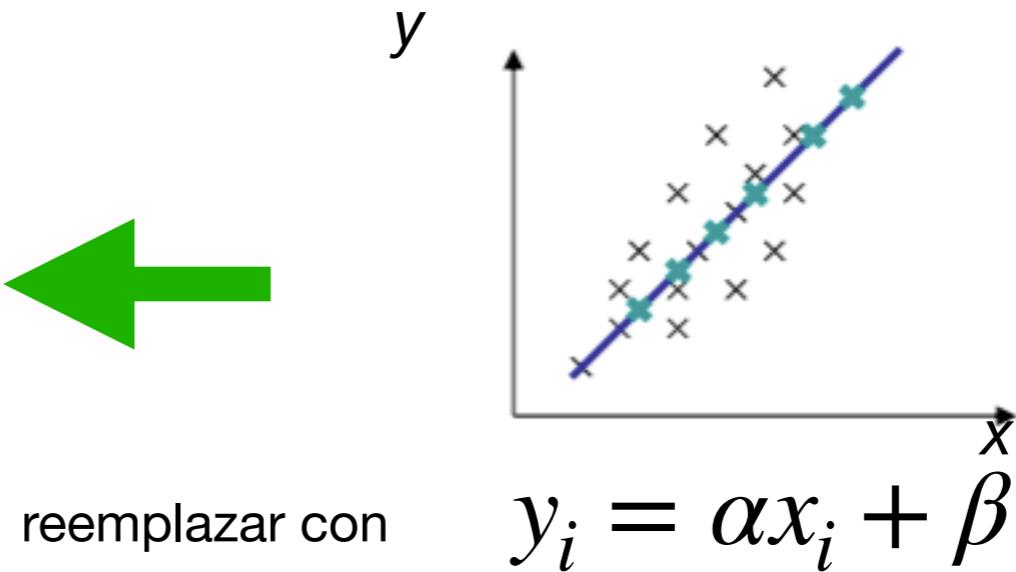
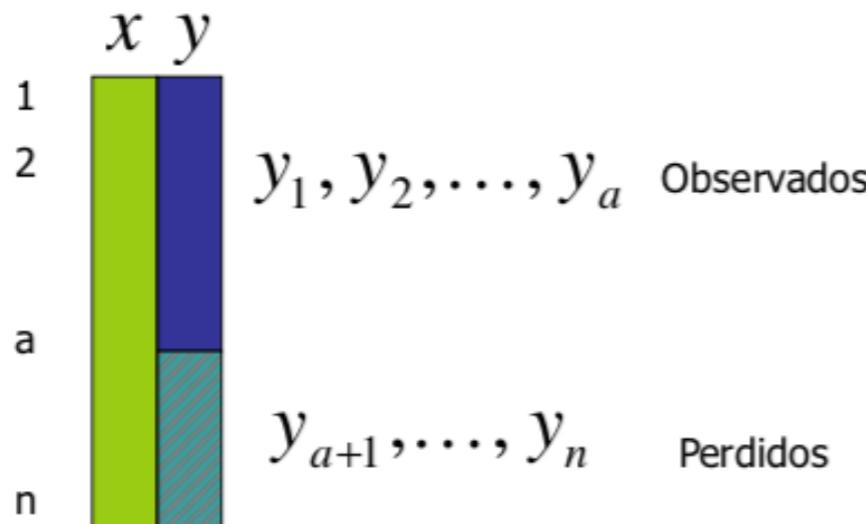
# Explorar - Limpiar los Datos

- Imputación:
- Imputación por regresión lineal:



# Explorar - Limpiar los Datos

- Imputación:
- Imputación por regresión lineal:



- Aumenta correlaciones

# Explorar - Limpiar los Datos

- **Imputación**
- Vimos sólo los métodos **más simples de imputación**. El problema en sí es muy difícil por lo que hay mucho trabajo en el problema de imputación.
- Otros métodos de imputación incluyen:
  - Imputación basada en arboles de decisión.
  - Imputación por EM (Expectation Maximization).
  - Imputación usando Modelos Gráficos.
- Muchas veces conviene entender el **proceso de producción** de los datos para proponer métodos mas ad-hoc de imputación.
- No todos los modelos necesitan imputación de datos, por ejemplo, los **modelos basados en árboles** pueden trabajar con datos perdidos.

# Explorar - Limpiar los Datos

- **Transformación**
  - Para algunos modelos ayuda tener las diferentes features en rangos similares (por ejemplo SVM, Redes Neuronales - Lo veremos en mayor profundidad al analizar estos métodos).
  - Para lograr esto se pueden transformar las variables.
- 
- **MinMax Scaling:**

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

- donde  $x_{min}$  y  $x_{max}$  pueden ser conocidos desde antes o obtenidos desde el dataset.

# Explorar - Limpiar los Datos

- Transformación
  - Standardization:

$$x' = \frac{x - \bar{x}}{\sigma}$$

- Los nuevos datos quedarán con media zero y varianza unitaria.
- La media y la desviación estándar pueden ser obtenidas del dataset. Notar que los valores deben ser guardados para ser usados luego.

# Explorar - Limpiar los Datos

- **One-hot encoding o Variables Dummy**
- Considere una variable categórica como color de ojo {1: marrón, 2: azul, 3: verde, 4: gris} para predecir altura (0mt, 2mt).
- De ninguna forma le queremos decir a nuestro modelo que hay alguna correlación entre los números asignados al color (1 a 4) y la altura.
- Para incluir estas variables en el modelo, usamos variables dummy.
  - **color** -> **es\_marrón** {no:0, si:1}, **es\_azul** {no:0, si:1}, **es\_verde** {no:0, si:1}
- Para una variable categorica de **K** valores se necesitan **K-1** variables dummy.
- No es necesario incluir variables dummy en modelos basados en **árboles de decisión**.

# **Explorar - Limpiar los Datos**

**Jupyter Notebook EDA Preprocessing cells**

# Visualización de Datos

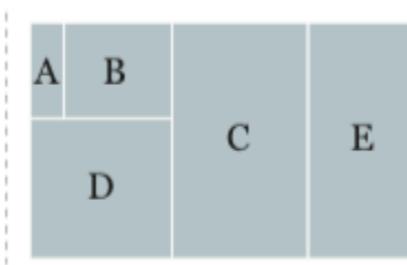
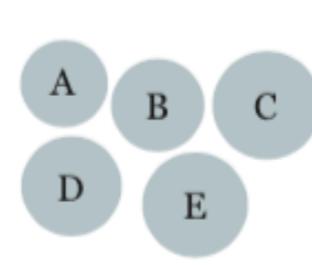
## *Length or height*



## *Position*



## *Area*



## *Angle/area*



## *Line weight*



## *Hue and shade*

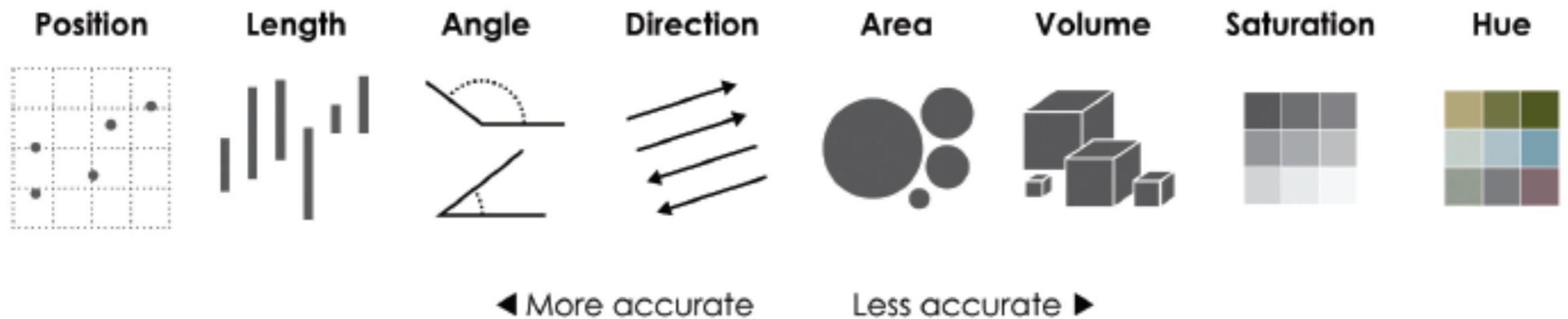


Figures represented  
in all these graphics:  
22%, 25%, 34%, 29%, 32%

Data visualization  
and visual encoding

# Visualización de Datos

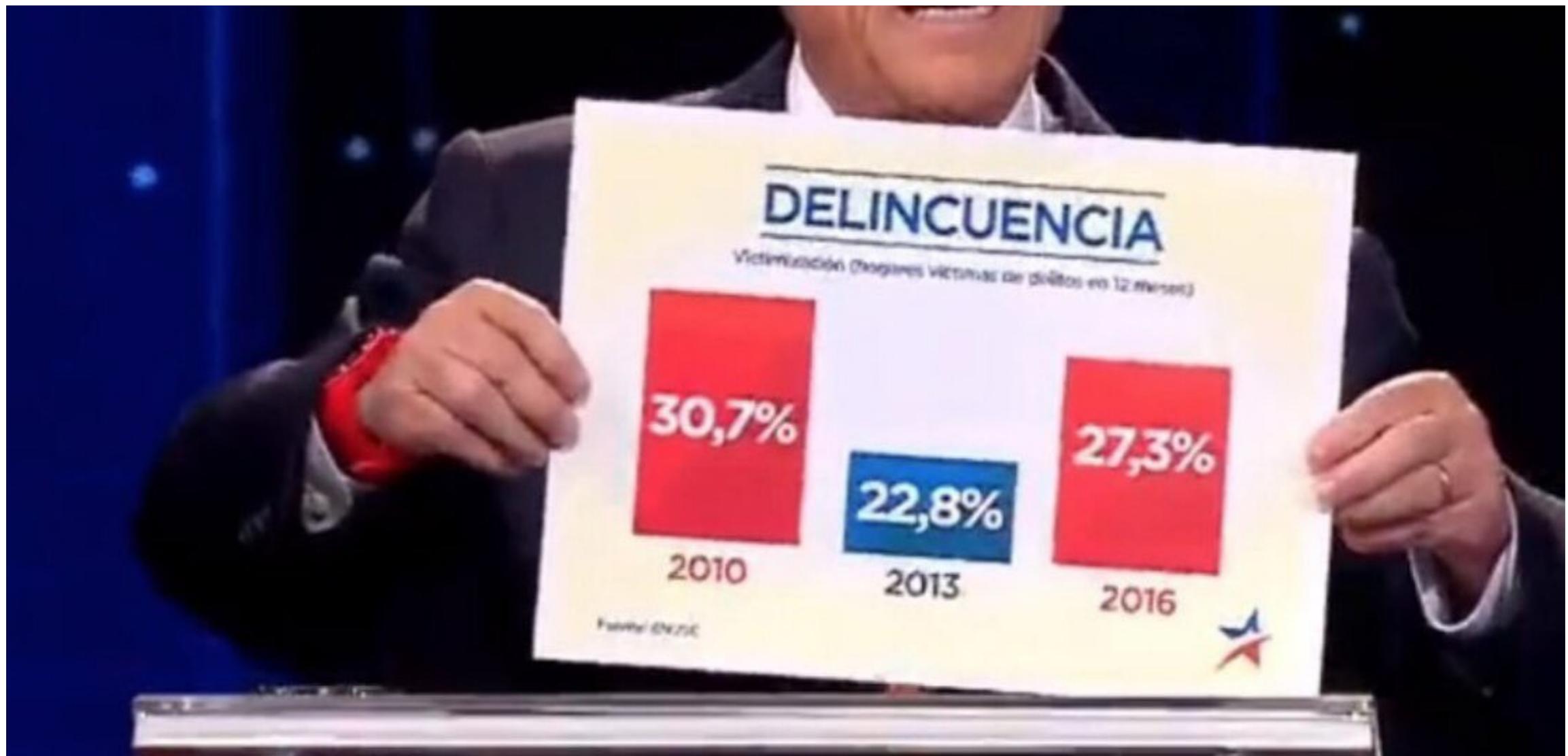
# Codificación de información



Graphical Perception and Graphical Methods for Analyzing Scientific Data, Cleveland W. & McGill R.

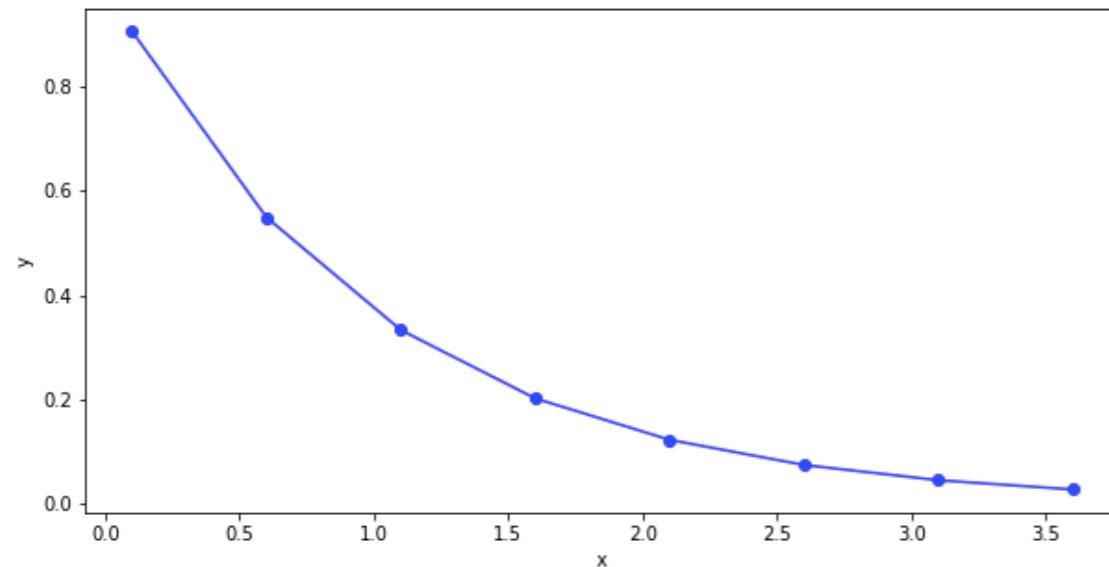
# Visualización de Datos

## Integridad Gráfica



# Visualización de Datos

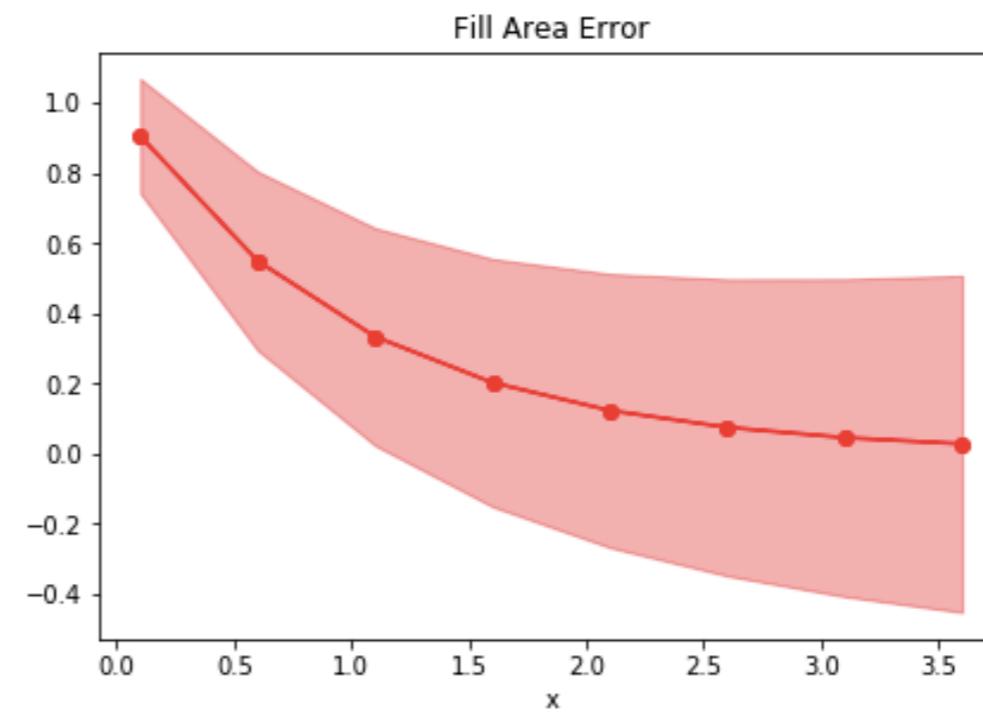
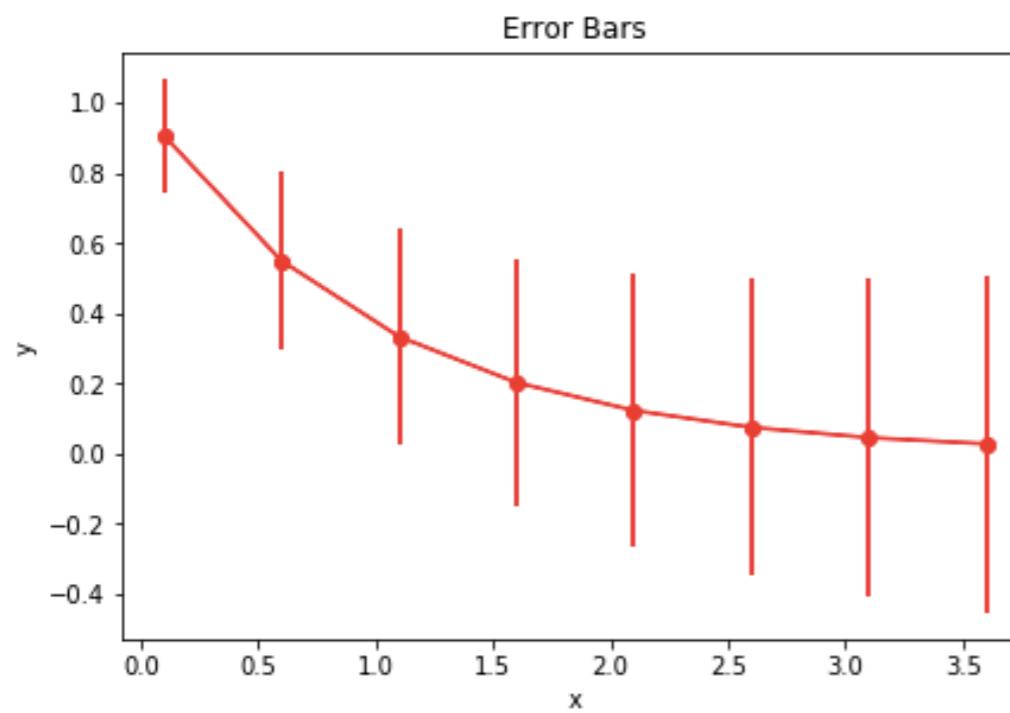
## Incluir Incerteza



$$SD = \sigma$$

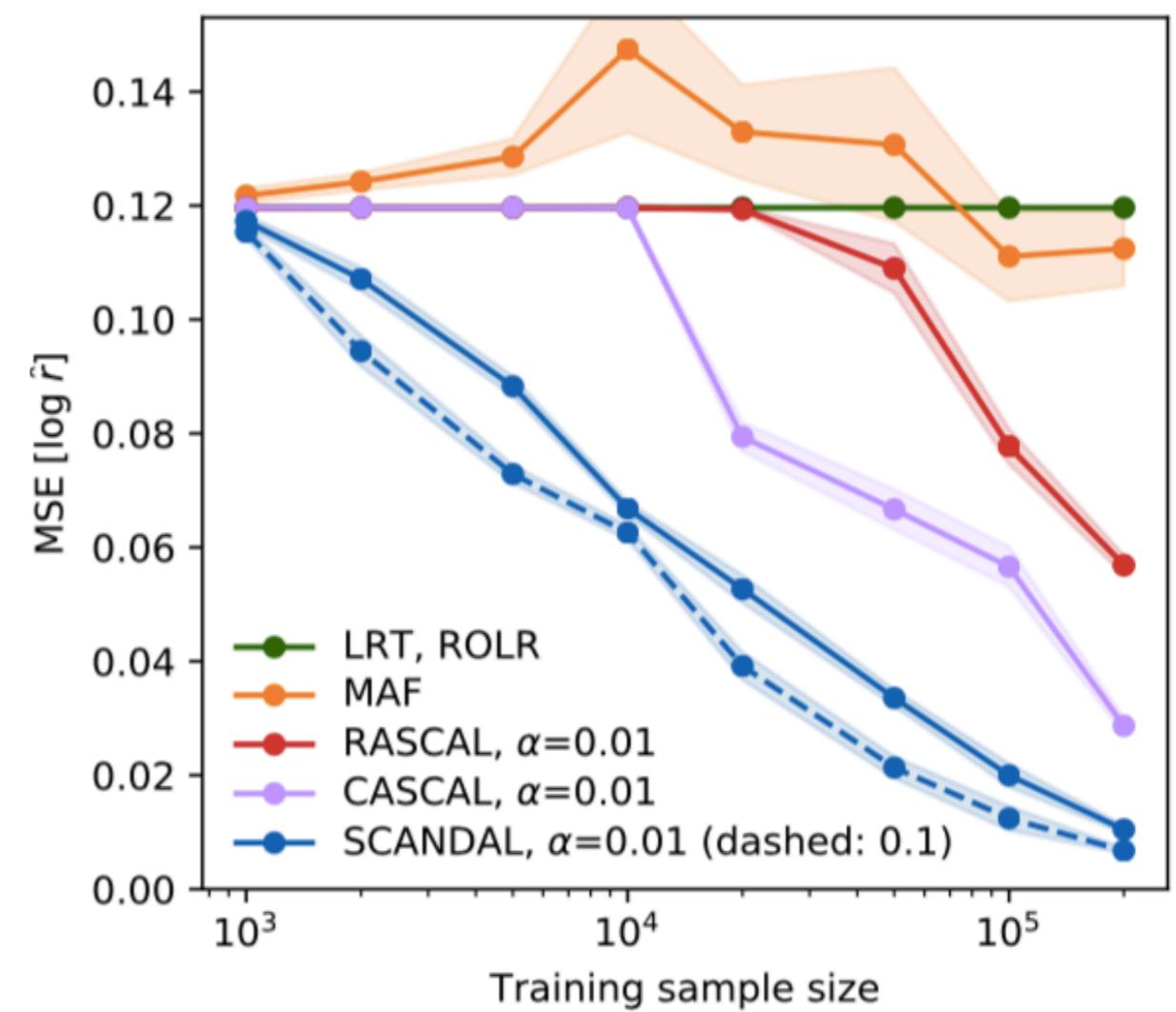
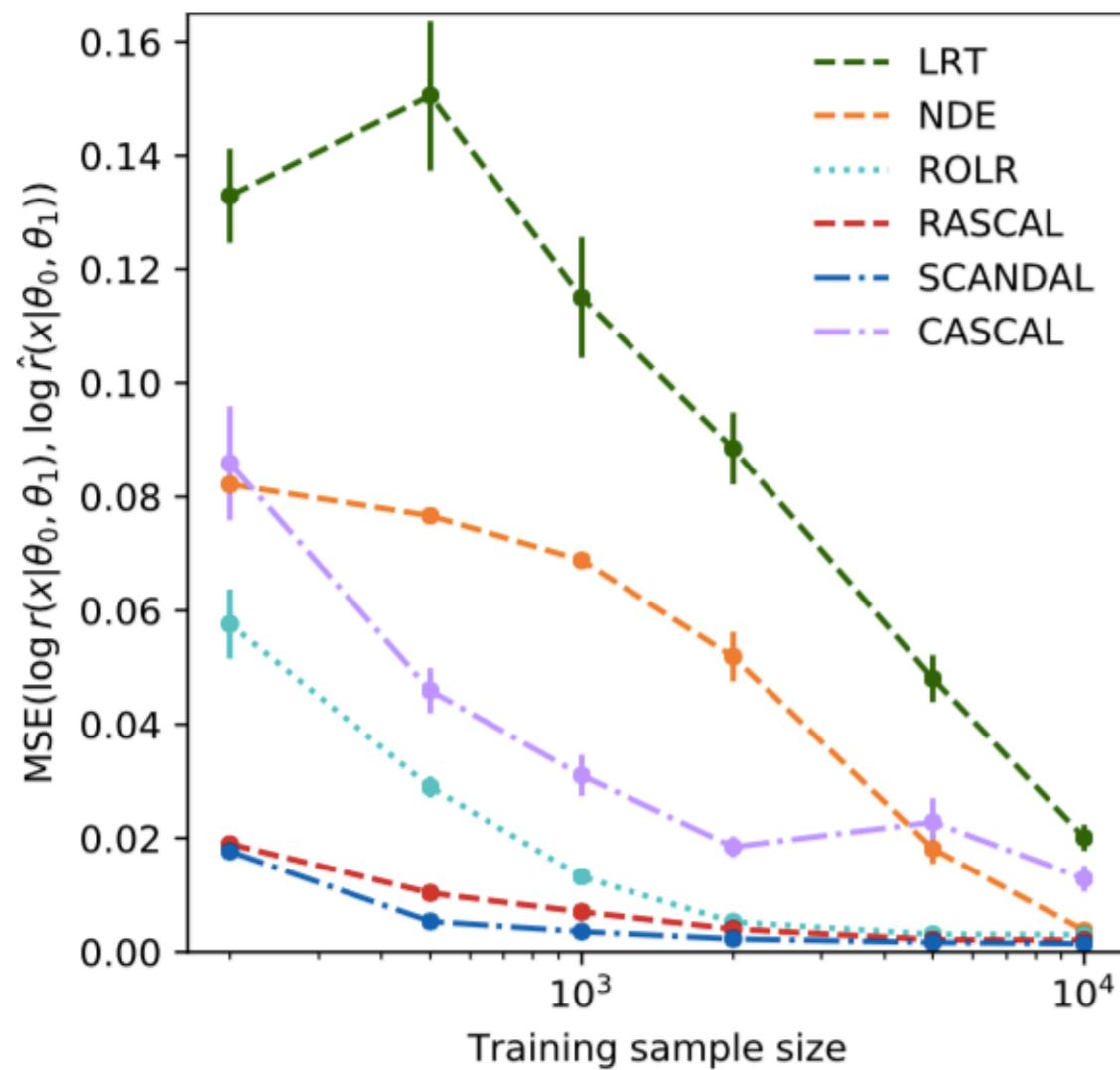
$$SE = \sigma/\sqrt{n}$$

$$95\% CI = x \pm 1.96 \times SE$$



# Visualización de Datos

## Incluir Incerteza



# Visualización de Datos

## Colormap: Viridis

<https://www.youtube.com/watch?v=xAoljeRJ3IU>  
(A Better Default Colormap for Matplotlib)

