# *Aprendizaje automático*

Departamento de Ingeniería en Informática
ITBA

# Trabajo Práctico 4

- Para cargar datos desde un archivo excel, por ej. los datos del ej. 4.

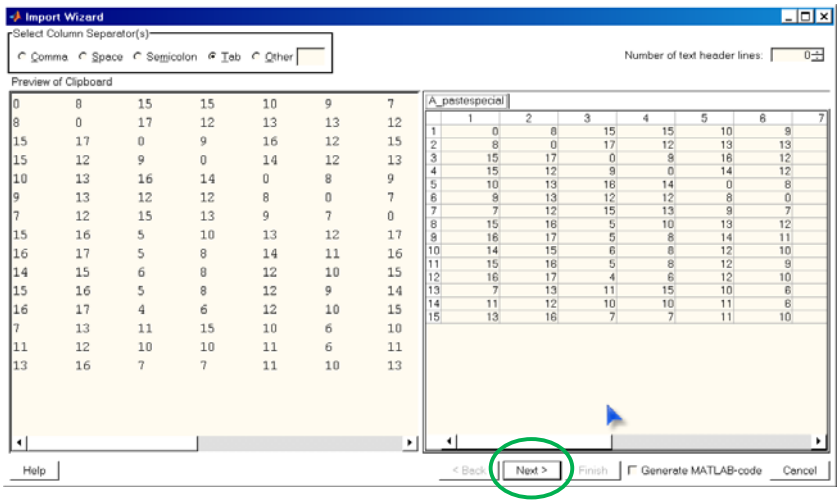
1. Abrir el Excel y seleccionar D2:R16 con CTRL-C para que quede en guardado en el portapapeles.

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | Nombre | Partido | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | |
| 2 | 1 | Hunt | R | 0 | 8 | 15 | 15 | 10 | 9 | 7 | 15 | 16 | 14 | 15 | 16 | 7 | 11 | 13 | |
| 3 | 2 | Sandman | R | 8 | 0 | 17 | 12 | 13 | 13 | 12 | 16 | 17 | 15 | 16 | 17 | 13 | 12 | 16 | |
| 4 | 3 | Howard | D | 15 | 17 | 0 | 9 | 16 | 12 | 15 | 5 | 5 | 6 | 5 | 4 | 11 | 10 | 7 | |
| 5 | 4 | Thompson | D | 15 | 12 | 9 | 0 | 14 | 12 | 13 | 10 | 8 | 8 | 8 | 6 | 15 | 10 | 7 | |
| 6 | 5 | Freylinghuyseb | R | 10 | 13 | 16 | 14 | 0 | 8 | 9 | 13 | 14 | 12 | 12 | 12 | 10 | 11 | 11 | |
| 7 | 6 | Forsythe | R | 9 | 13 | 12 | 12 | 8 | 0 | 7 | 12 | 11 | 10 | 9 | 10 | 6 | 6 | 10 | |
| 8 | 7 | Widnall | R | 7 | 12 | 15 | 13 | 9 | 7 | 0 | 17 | 16 | 15 | 14 | 15 | 10 | 11 | 13 | |
| 9 | 8 | Roe | D | 15 | 16 | 5 | 10 | 13 | 12 | 17 | 0 | 4 | 5 | 5 | 3 | 12 | 7 | 6 | |
| 10 | 9 | Heltoski | D | 16 | 17 | 5 | 8 | 14 | 11 | 16 | 4 | 0 | 3 | 2 | 1 | 13 | 7 | 5 | |
| 11 | 10 | Rodino | D | 14 | 15 | 6 | 8 | 12 | 10 | 15 | 5 | 3 | 0 | 1 | 2 | 11 | 4 | 6 | |
| 12 | 11 | Minish | D | 15 | 16 | 5 | 8 | 12 | 9 | 14 | 5 | 2 | 1 | 0 | 1 | 12 | 5 | 5 | |
| 13 | 12 | Rinaldo | R | 16 | 17 | 4 | 6 | 12 | 10 | 15 | 3 | 1 | 2 | 1 | 0 | 12 | 6 | 4 | |
| 14 | 13 | Maraziti | R | 7 | 13 | 11 | 15 | 10 | 6 | 10 | 12 | 13 | 11 | 12 | 12 | 0 | 9 | 13 | |
| 15 | 14 | Daniels | D | 11 | 12 | 10 | 10 | 11 | 6 | 11 | 7 | 7 | 4 | 5 | 6 | 9 | 0 | 9 | |
| 16 | 15 | Patten | D | 13 | 16 | 7 | 7 | 11 | 10 | 13 | 6 | 5 | 6 | 5 | 4 | 13 | 9 | 0 | |
| 17 | | | | | | | | | | | | | | | | | | | |

2. En Matlab, seleccionar

**Edit** > **Paste to Workspace**

3. Aparece una ventana como la siguiente:




4. Hacer click en **Next** y después en **Finish** en la ventana siguiente

5. Este proceso crea una matriz de 15x15 llamada **A_pastespecial** que contiene los datos.

6. Es conveniente renombrar esa matriz, por ej. hacer

   ```
   Datos = A_pastespecial;
   ```

- K-medias

`[idx,C] = kmeans(x,k)`
Devuelve las coordenadas de los k centroides en la matriz $C$ de $k \times p$, siendo $p$ el número de variables y en `idx` el número de cluster al que pertenece cada observación.

`[ ... ] = kmeans(..., 'param1',val1, 'param2',val2, ...)`
Algunos de los parámetros que se pueden dar como input son:

- 'Distance': función de distancia. Algunas posibilidades son:
  - 'sqEuclidean': cuadrado de la distancia euclídea (default)
  - 'cityblock': distancia de Manhattan
- 'Options': opciones para el algoritmo iterativo que se usa para minimizar. Se pueden especificar:
  - 'Display': nivel del display del output
    - 'off' (default)
    - 'iter'
    - 'final'
  - 'MaxIter': número máximo de iteraciones permitidas (default 100).

- Clusters jerárquicos

`Y = pdist(X,distance)`
Devuelve un vector `Y` con las distancias entre cada uno de los pares de observaciones de la matriz `X` de $n \times p$. Las filas de `X` corresponden a las observaciones, las columnas corresponden a las variables.
`Y` es un vector fila de dimensión $n(n-1)/2$, que corresponde a los $n(n-1)/2$ pares de observaciones en `X`. El output de `Y` sigue el siguiente orden $(2,1), (3,1), \ldots, (n,1), (3,2), \ldots (n,2), \ldots . (n,n-1)$.
Para encontrar la distancia entre las observaciones $i$ y la $j$ ($i < j$), se puede usar la fórmula

$$Y((i-1) * (n - i/2) + j - i)$$

Algunas de las funciones de distancia que se pueden usar son:

- 'euclidean' : distancia Euclídea
- 'cityblock': distancia de Manhattan
- 'minkowski': distancia de Minkowski con exponente 2

`Z = linkage(Y, method)`
Crea el árbol de clusters jerárquicos. El input es `Y` , un vector de distancias como el generado por `pdist`. También puede ser una matriz de similitud que debe reformatearse para tener el formato de output de `pdist`.
Algunas de los métodos que se pueden usar son:

- 'single': distancia mínima
- 'complete': distancia máxima
- 'average': promedio de las distancias

Devuelve una matriz `Z` de $m \times 3$, donde $m$ es el número de observaciones. Las columnas 1 y 2 de `Z` contienen los pares de índices que se unen para formar el árbol. Las hojas se numeran de 1 a $m$

`T = cluster(Z,'MaxClust',N)`
Construye un máximo de `N` clusters a partir del árbol de clusters jerárquicos dado por `Z`.
`Z` es una matriz de $(m-1) \times 3$ generada por `linkage`, donde $m$ es el número de observaciones.
La altura de cada nodo del árbol representa la distancia entre los dos subnodos que se unen en ese nodo.
Devuelve `T` que es un vector de largo $m$ que contiene el número de cluster al que pertenece cada una de las $m$ observación.

`dendrogram(Z)`
Genera el plot del dendrograma del del árbol de clusters jerárquicos dado por `Z`.
`Z` es una matriz de $(m-1) \times 3$ generada por `linkage`, donde $m$ es el número de observaciones