PRÁCTICAS DE LIBRERÍAS PYTHON Y MACHINE LEARNING PARA INFORMACIÓN BIBLIOGRÁFICA

Marcelo Lorca González

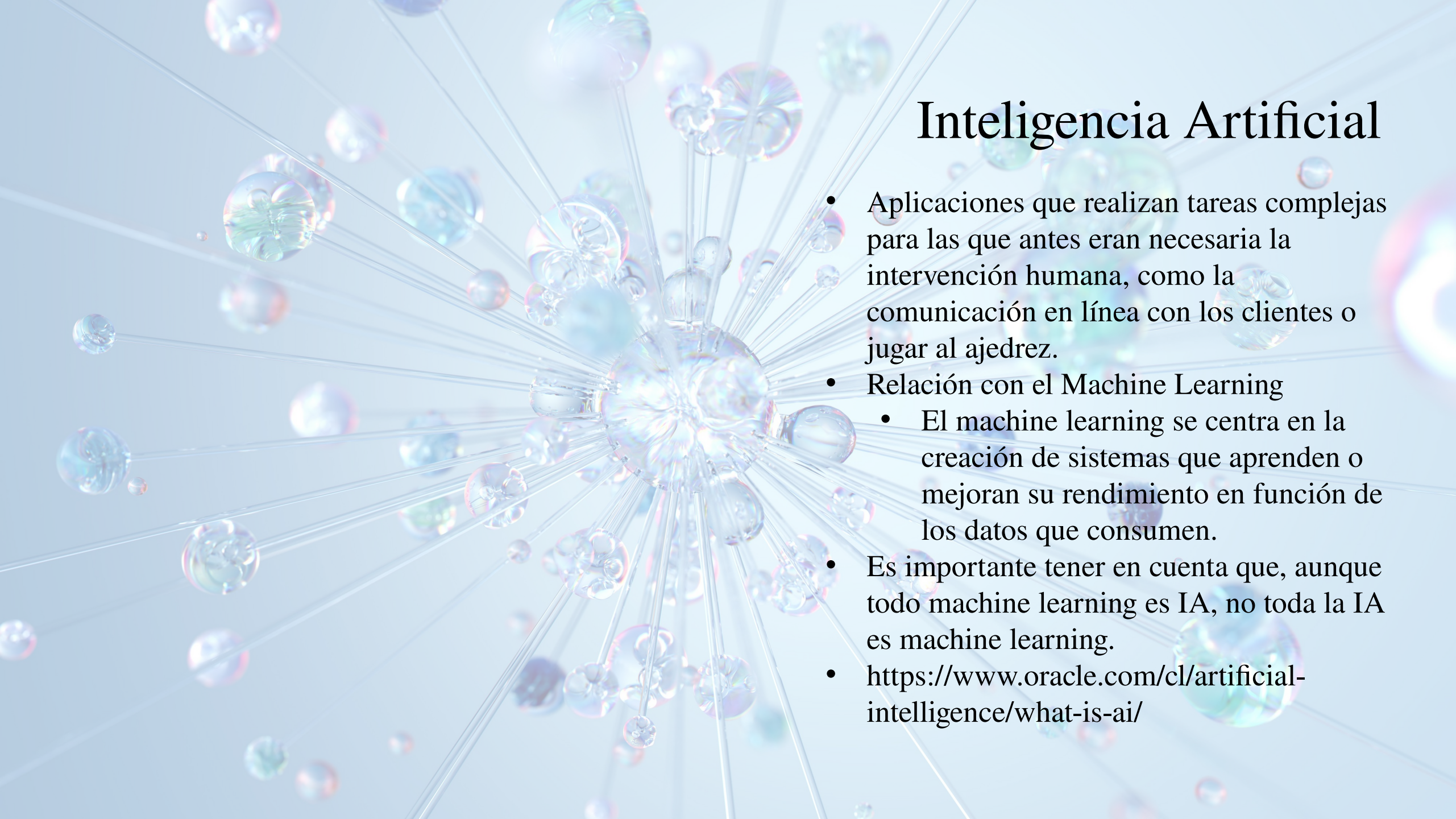
Biblioteca del Congreso
Nacional de Chile / BCN

Machine Learning

Capacidad de las Máquinas de generar aprendizajes automáticos a partir de los datos.

Identifican patrones complejos a partir de los datos. Generan predicciones a partir de los mismos.

Utilizan elementos de estadística.



Inteligencia Artificial

- Aplicaciones que realizan tareas complejas para las que antes eran necesaria la intervención humana, como la comunicación en línea con los clientes o jugar al ajedrez.
- Relación con el Machine Learning
  - El machine learning se centra en la creación de sistemas que aprenden o mejoran su rendimiento en función de los datos que consumen.
- Es importante tener en cuenta que, aunque todo machine learning es IA, no toda la IA es machine learning.
- https://www.oracle.com/cl/artificial-intelligence/what-is-ai/

Lenguaje de Programación Python

Python es un lenguaje de programación potente y fácil de aprender. Tiene estructuras de datos de alto nivel eficientes y un simple pero efectivo sistema de programación orientado a objetos.

Características

- **Lenguaje Interpretado**
- Integración con varios sistemas operativos
- Gran cantidad de librerías en la versión standard
- Comunidad de desarrollo muy grandes
- Poderosas librerías para IA

Librerías

- Recursos de métodos y atributos que permiten extender las potencialidades y funcionalidades de un lenguaje de programación, relacionado con aspectos específicos.

Librerias Python

Numpy

- permite a los desarrolladores de Python realizar en forma rápida una **amplia variedad de cálculo numéricos**.

Pandas

- Herramienta para manipulación de datos de alto nivel desarrollada por Wes McKinney. **Es construido sobre Numpy y permite el análisis de datos que cuenta con las estructuras de datos que necesitamos para limpiar los datos en bruto y que sean aptos para el análisis** (por ejemplo, tablas). /https://joserzapata.github.io/courses/python-ciencia-datos/pandas

MatplotLib

- Matplotlib es una biblioteca completa para crear visualizaciones estáticas, animadas e interactivas en Python. https://matplotlib.org/

Scikit-learn

- Herramientas para el análisis predictivo de datos. Accesible para todos y reutilizable en diversos contextos. Construido sobre NumPy, SciPy y matplotlib Código abierto, utilizable comercialmente: licencia BSD. https://scikit-learn.org/stable/

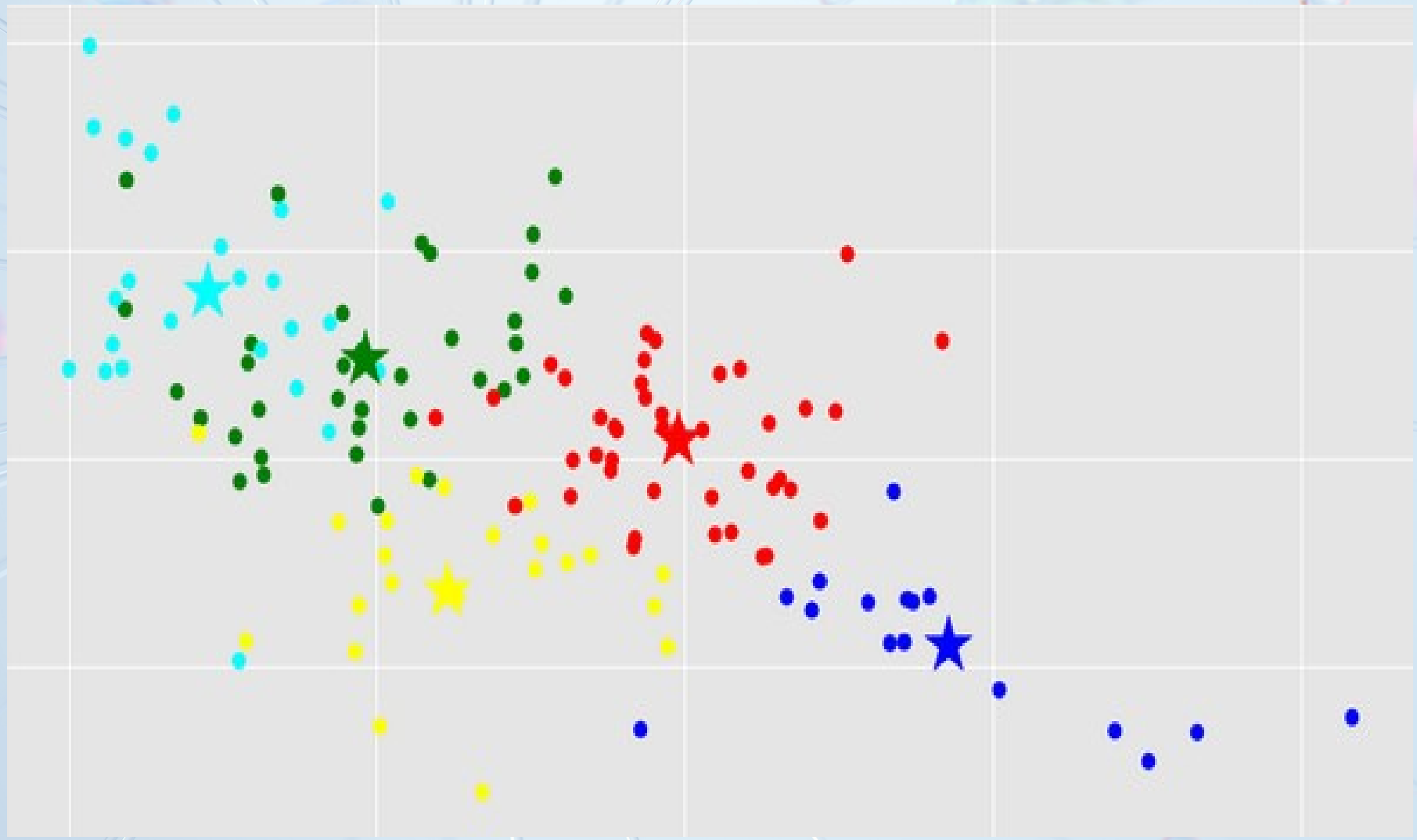
Machine Learning: Tipos de Análisis

Análisis Supervisado: datos etiquetados	• Predecir
Análisis no supervisado: datos no etiquetados	• Clustering • KMEANS
Sobreajuste	• Se produce cuando el modelo de aprendizaje automático proporciona predicciones precisas para los datos de entrenamiento, pero no para los datos nuevos.

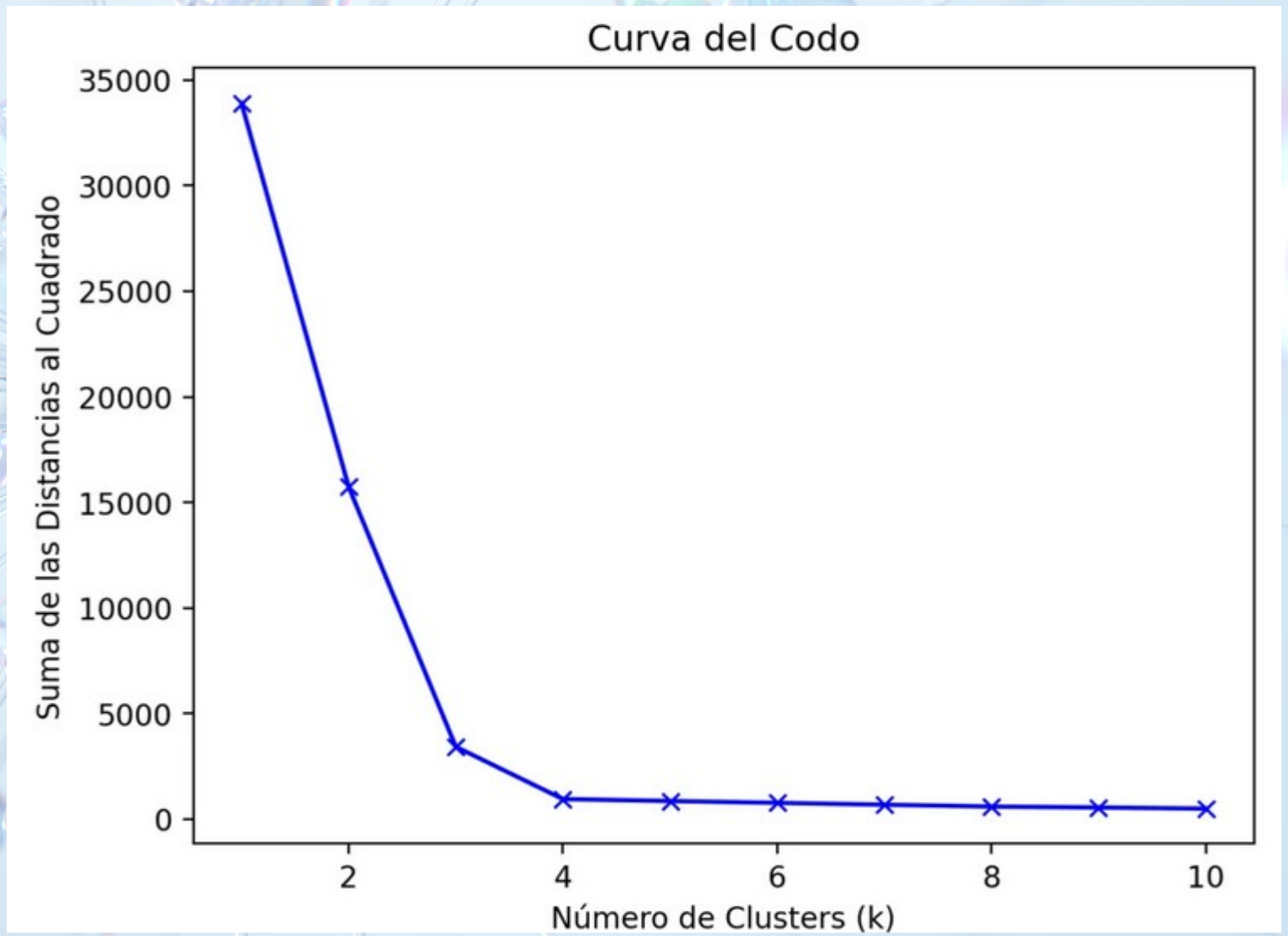
K-MEANS

- Método de Análisis: Clustering o agrupamiento.
- Elementos: K=Centroides. Distancia de los elementos a los centroides.
- Obtener patrones, anomalías o particularidades a partir de los datos

K-means: gráfico



Cuántos centroides necesito?



- **Problema a resolver**: Definir las áreas temáticas más relevantes de la colección de Monografías de Valparaíso.
- Se analizan los datos bibliográficos asociados a dicha colección.
- (Datos en archivos Excel)

Colección Monografías Valparaíso

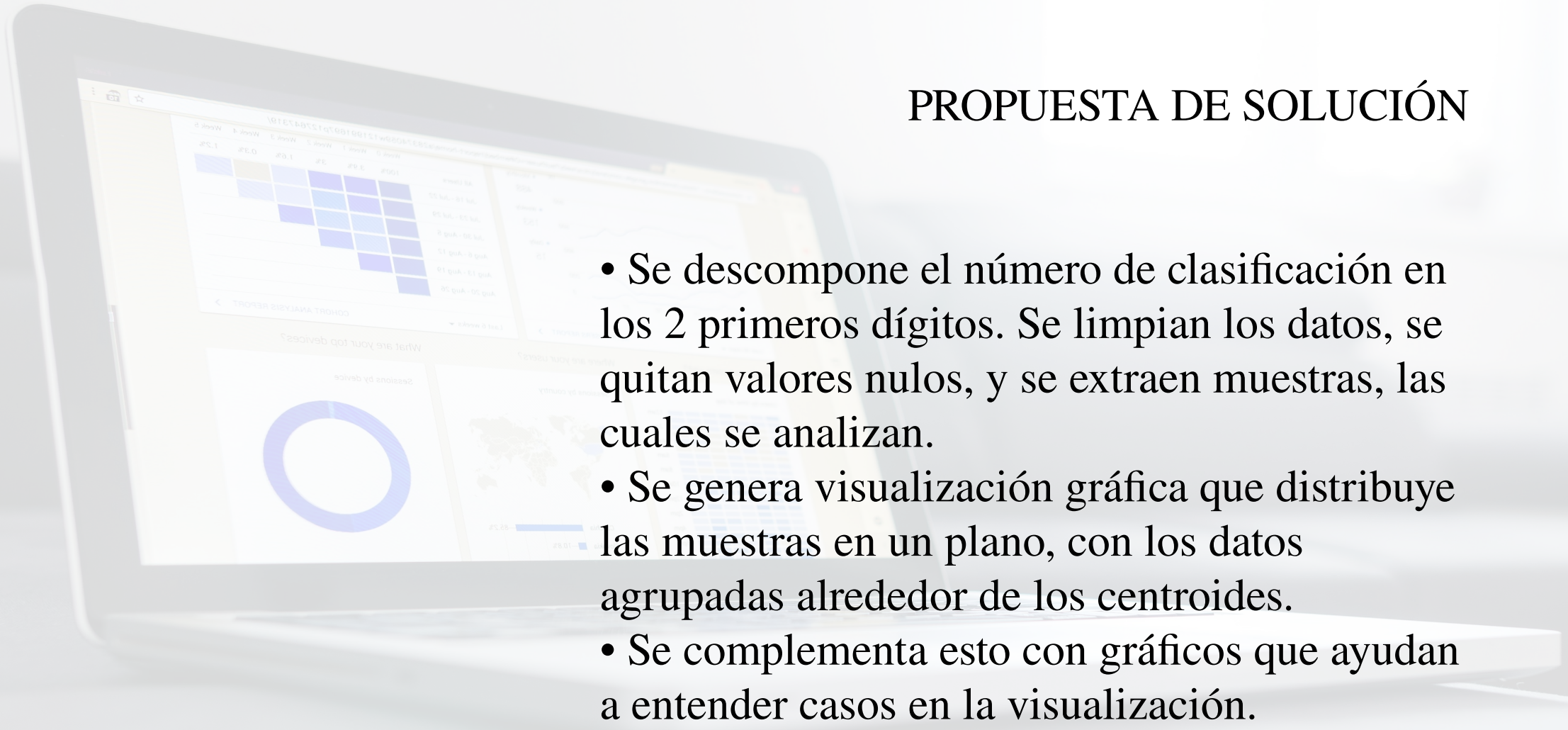
	Número Clasificación	Materia (tag 650)
0	<NA>	<NA>
1	343.2(83) Ch537c 1988	Derecho penal Chile
2	347.249(83) Ch537c 1983	Derecho minero Chile;Concesiones Chile
3	159.964 J95sg.E 1962	Psicoanálisis;Simbolismo (Psicología);Diablo...
4	396(8=6) M953e 1984	Mujeres América Latina Condiciones sociales C...
5	32(8=6) C146h 1991	Filosofía política América Latina;América ...
6	347.23(83) F475p 1991	Patrimonio Chile
7	371.13(83) P979c 1991	Profesores Formación profesional Chile;Calida...
8	327"1989" G531r.E 1989	Política mundial 1989-
9	32:316 C387l 1979;32:316 C387l 1979	Ciencias políticas Reseñas de libros;Ciencia...
10	342.7 B638v 1990	Derechos humanos;Derecho internacional;Delitos...
11	347.96/.97 C577m 1990	Derecho;Abogados;Jueces;Administración de jus...

Detalle de la data

Df.info()

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30683 entries, 0 to 30682
Data columns (total 15 columns):
 #    Column                  Non-Null Count  Dtype
---   ------                  --------------  -----
 0    bib#                    30683 non-null  int64
 1    autor                   20323 non-null  object
 2    título                  30683 non-null  object
 3    Editor                  28081 non-null  object
 4    fecha publicación       28134 non-null  object
 5    Nota contenido          28201 non-null  object
 6    Resumen                 7479 non-null   object
 7    Materia (tag 650)       28922 non-null  object
 8    otros autores           10871 non-null  object
 9    Número Clasificación    28704 non-null  object
 10   Sede                    28796 non-null  object
 11   colección               28796 non-null  object
 12   item#                   28796 non-null  object
 13   estado                  28796 non-null  object
 14   Tipo item               28796 non-null  object
dtypes: int64(1), object(14)
memory usage: 3.5+ MB
```



PROPUESTA DE SOLUCIÓN

- Se descompone el número de clasificación en los 2 primeros dígitos. Se limpian los datos, se quitan valores nulos, y se extraen muestras, las cuales se analizan.
- Se genera visualización gráfica que distribuye las muestras en un plano, con los datos agrupadas alrededor de los centroides.
- Se complementa esto con gráficos que ayudan a entender casos en la visualización.

Clasificación Decimal Universal. Áreas Principales

0 CIENCIA Y CONOCIMIENTO. ORGANIZACIÓN. CIENCIAS DE LA COMPUTACIÓN. INFORMACIÓN. DOCUMENTACIÓN. INSTITUCIONES. PUBLICACIONES

1 FILOSOFÍA. PSICOLOGÍA

2 RELIGIÓN. TEOLOGÍA

3 CIENCIAS SOCIALES

4 [NO USADO]

5 MATEMÁTICAS. CIENCIAS NATURALES

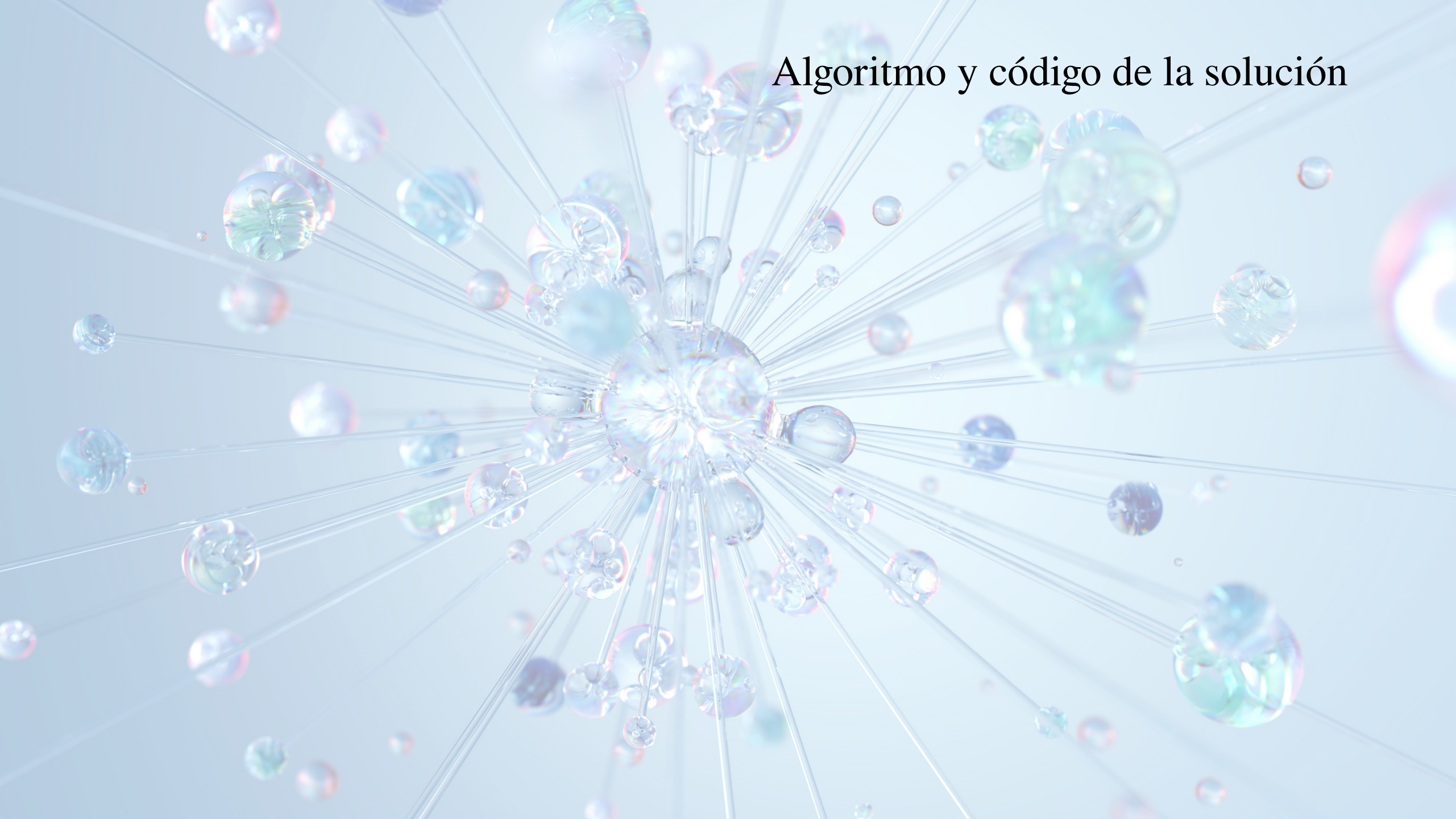
6 CIENCIAS APLICADAS. MEDICAMENTO. TECNOLOGÍA

7 LAS ARTES. RECREACIÓN. ENTRETENIMIENTO. DEPORTE

8 IDIOMA. LINGÜÍSTICA. LITERATURA

9 GEOGRAFÍA. BIOGRAFÍA. HISTORIA

Algoritmo y código de la solución



Algoritmo y código de la solución

Abrir Colab,

Cargar librerías,

Conectar a google drive

```
[1]  # Configuración warnings
     # ========================================
     import warnings
     warnings.filterwarnings('ignore')
```

```
[2]  import numpy as np
     import matplotlib.pyplot as plt
```

```
     from google.colab import drive
     drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
[4]  import matplotlib.pyplot as plt
```

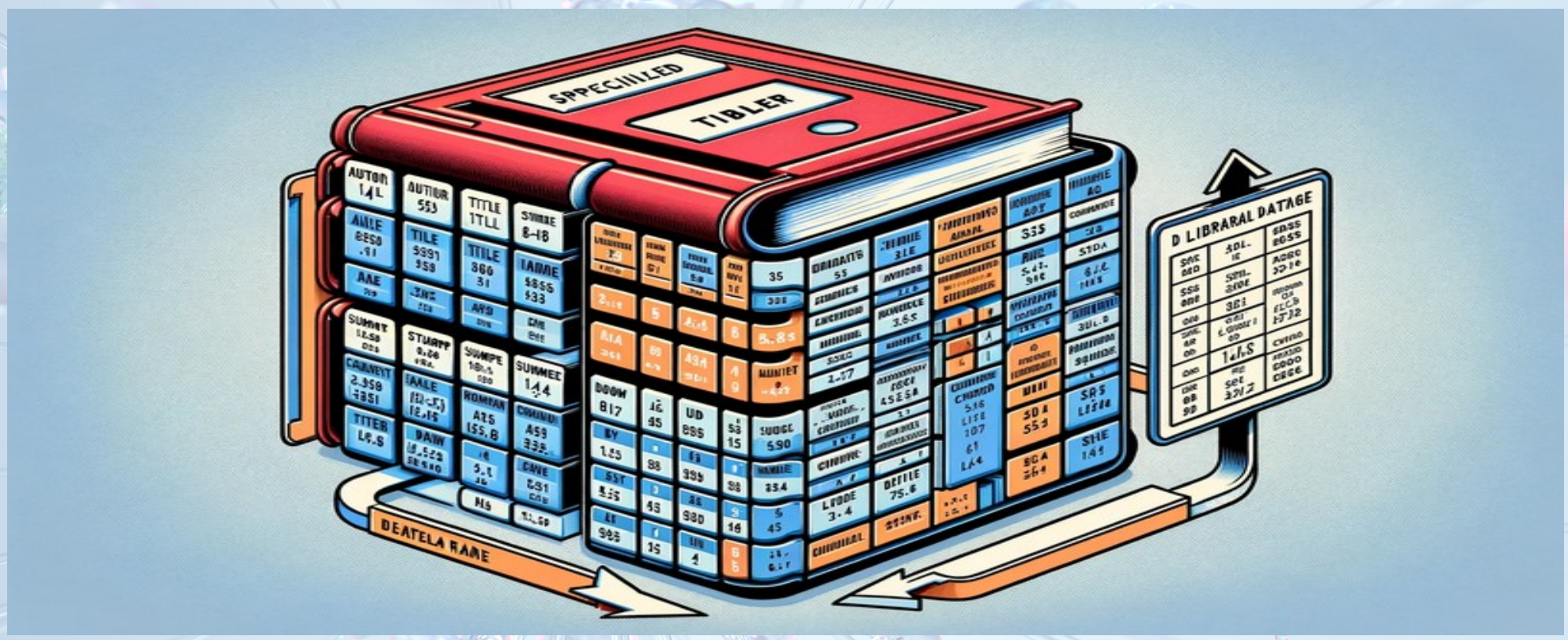
Carga de archivos al ambiente Colab

```
from google.colab import
files

uploaded = files.upload()

for fn in uploaded.keys():
  print('User uploaded file
"{name}" with length {length}
bytes'.format(
      name=fn,
length=len(uploaded[fn])))
```

Carga de datos



```
df = pd.read_excel('listadoParaSimposium-SV.xlsx')

df.head(100)
```

Preparación de los datos

- Limpieza de datos
  - (filtrar datos)
- Tratamiento de datos nulos
- Tratamiento de datos en blanco
- Ajustar los tipos de datos

```
df_tematicas_clasif_is_sig
no_mas =
df_tematicas[df_tematicas[
"Número
Clasificación"].notnull()]

df_tematicas_not_null["cl
asif3"]=
df_tematicas_not_null['Nú
mero
Clasificación'].str.slice
(0, 2)
```

Kmeans: preparación y ejecución

```
import matplotlib.pyplot as plt
from sklearn.preprocessing import MinMaxScaler
from sklearn.cluster import KMeans
import numpy as np


escalar=MinMaxScaler().fit(df_clasif_or.values)
#escalar
clientes =
pd.DataFrame(escalar.transform(df_clasif_or.valu
es),
                                columns=["Total"])


kmeans =
KMeans(n_clusters=3).fit(clientes.values)

clientes["cluster"] = kmeans.labels_
```

Visualización de Datos

```
plt.figure(figsize=(6, 5), dpi=100)

colores = ["red", "blue", "orange", "black", "purple", "pink", "brown"]

for cluster in range(kmeans.n_clusters):
    plt.scatter(clientes[clientes["cluster"] == cluster]["clasif_ind"],
                clientes[clientes["cluster"] == cluster]["Total"],
                marker="o", s=100, color=colores[cluster], alpha=0.2)

    plt.scatter(kmeans.cluster_centers_[cluster][0],
                kmeans.cluster_centers_[cluster][1],
                marker="P", s=280, color=colores[cluster])

plt.title("Distribución Clasif CDU en Colección Valpo", fontsize=20)
plt.xlabel("Clasif CDU", fontsize=15)
plt.ylabel("Total", fontsize=15)
plt.text(1.15, 0.2, "K = %i" % kmeans.n_clusters, fontsize=25)
plt.text(1.15, 0, "Inercia = %0.2f" % kmeans.inertia_, fontsize=25)
plt.xlim(-0.1, 1.1)
plt.ylim(-0.1, 1.1)
plt.show()
```

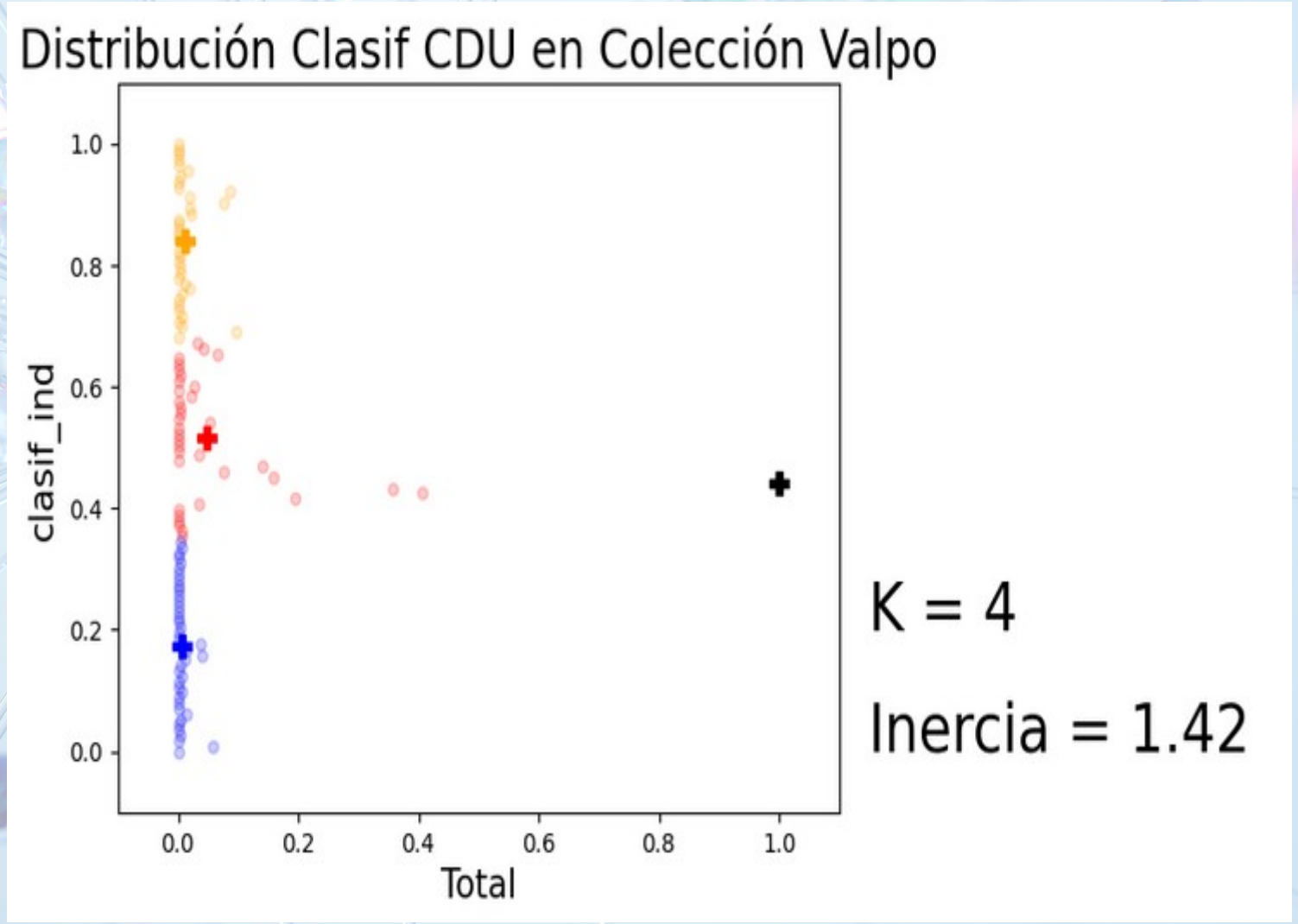
Data K-MEANS

clasif3	Total	clasif_ind
(8	1	0
00	496	1
01	5	2
02	37	3
03	1	4
05	3	5
06	43	6
07	123	7
08	3	8
1	5	9
1"	5	10

Datos Normalizados

clasif_ind	Total
0.000000	0.000000
0.058003	0.008850
0.000469	0.017699
0.004218	0.026549
0.000000	0.035398
0.000234	0.044248
0.004921	0.053097
0.014296	0.061947
0.000234	0.070796
0.000469	0.079646
0.000469	0.088496
0.006093	0.097345
0.000000	0.106195
0.000937	0.115044

Gráfico K-MEANS

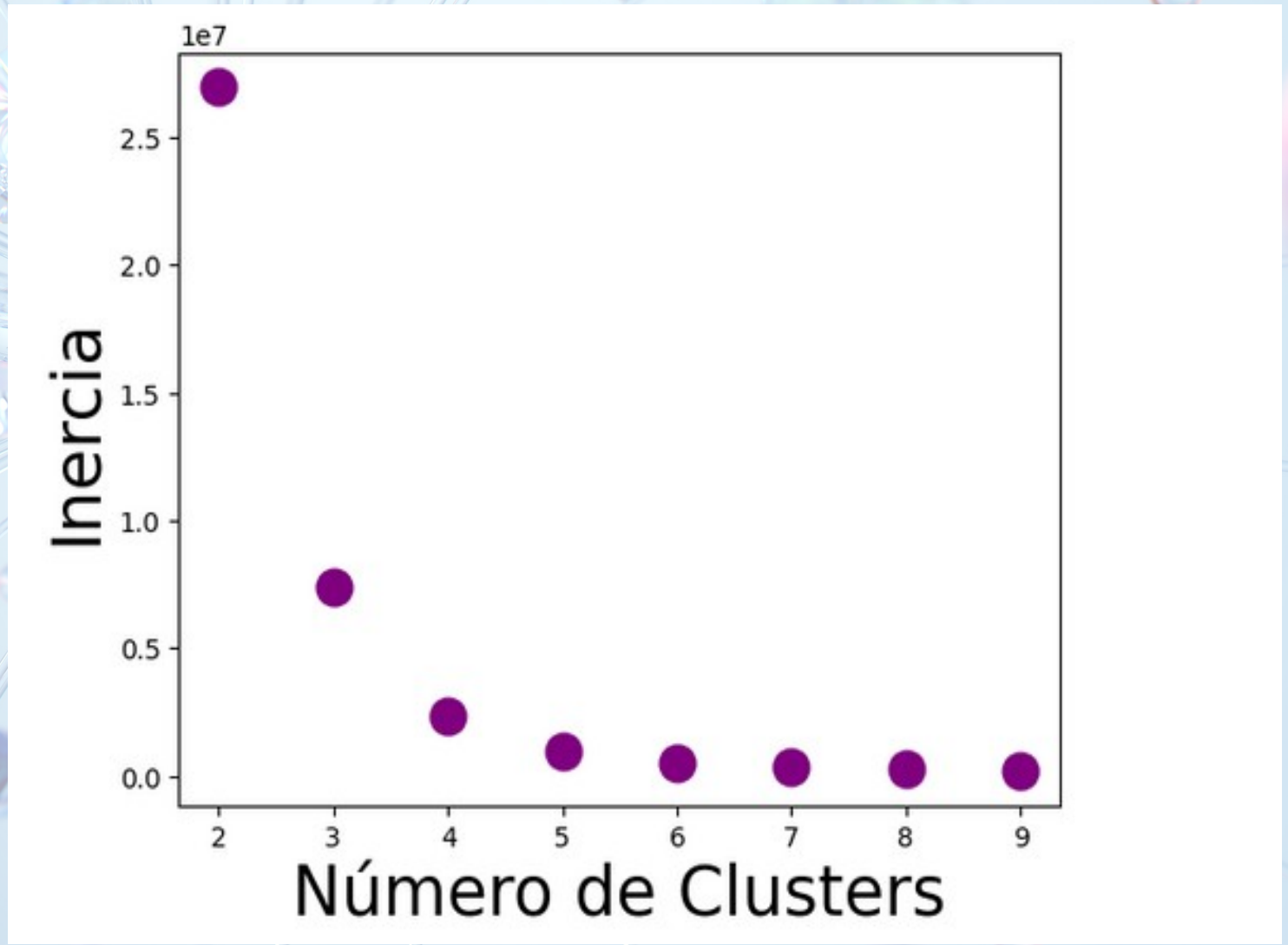


Método del Codo

```
inercias = []
for k in range(2, 10):
    kmeans =
KMeans(n_clusters=k).fit(df_clasif_or.values)
    inercias.append(kmeans.inertia_)

plt.figure(figsize=(6, 5), dpi=100)
plt.scatter(range(2, 10), inercias, marker="o",
s=180, color="purple")
plt.xlabel("Número de Clusters", fontsize=25)
plt.ylabel("Inercia", fontsize=25)
plt.show()
```

Gráfico del Codo



Conclusiones

- El número de centroides entre 3 y 4 corresponde a un buen valor, pues los grupos que genera son bastante homogéneos.
- Es posible potenciar mucho más este análisis con un trabajo colaborativo con el área de procesos técnicos, para una adecuada limpieza de datos.

Conclusiones

- Existen grupos dispersos del resto (rojo) y que corresponden a las temáticas con bastantes títulos en las colecciones. Existe además un grupo de un integrante (negro), debido a que su cantidad de títulos es ostensiblemente más alto al resto.

- Data ordenada por el total de títulos. Nos muestra que el área temática asociada al 34 (derecho) tiene un valor que sobresale de los demás.
- Los otros 5 datos sobresalen en menor medida de los demás.

Temáticas príncipes CDU (300 -399)

310 Estadísticas
320 Ciencia política
330 Economía
340 Derecho
350 Administración pública y ciencia militar
360 Problemas y servicios sociales, asociaciones
370 Educación

clasif3	Total	clasif_ind
34	8535	0
32	3463	1
33	3061	2
31	1672	3
35	1361	4
37	1208	5
65	836	6
94	743	7
92	658	8
36	644	9
61	569	10
00	496	11
50	443	12
62	362	13

Bibliografía

IBM. ¿Qué es el aprendizaje no supervisado?. https://www.ibm.com/es-es/topics/unsupervised-learning

Zapata, José R. PANDAS - Manipulacion de Datos con Python. https://joserzapata.github.io/courses/python-ciencia-datos/pandas

Matplotlib: Visualization with Python. https://matplotlib.org/

Scikit-learn: Machine Learning in Python. https://scikit-learn.org/stable/

El tutorial de Python. https://docs.python.org/es/3/tutorial/

La biblioteca estándar de Python. https://docs.python.org/es/3/library/index.html#library-index

Datos de Contacto:
Marcelo Lorca González
Email: mlorca@bcn.cl