

Introducción a la ciencia de datos

Módulo 1

Mauricio Rosales-Rivera



UNIVERSIDAD AUTÓNOMA DEL
ESTADO DE MORELOS

¿Qué es la ciencia de datos?

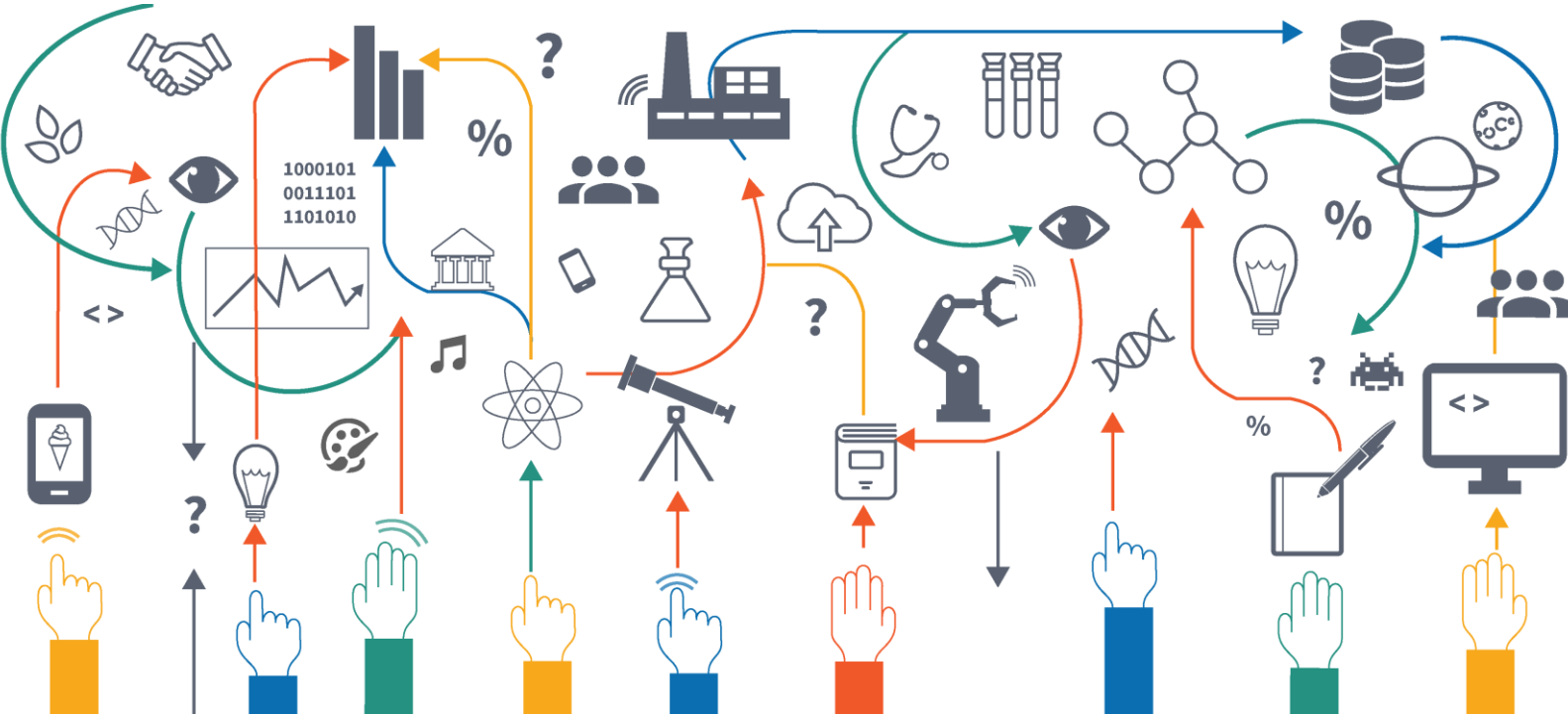
La ciencia de datos es un campo interdisciplinario que se ocupa de la extracción de conocimiento y la toma de decisiones a partir de datos.

Esto incluye:

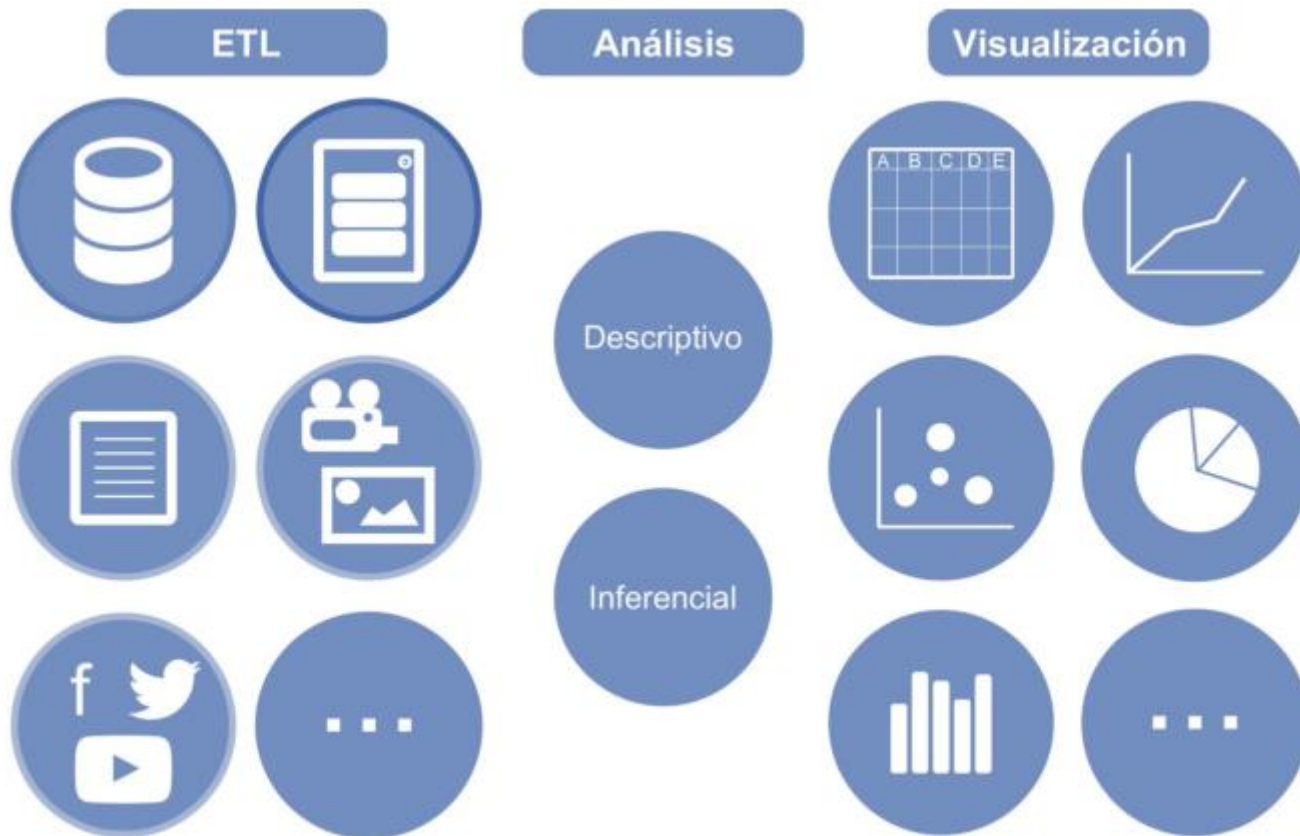
- Recolección,
- Almacenamiento,
- Análisis,
- Visualización y
- Presentación de datos

así como la aplicación de técnicas estadísticas, de aprendizaje automático y de inteligencia artificial para extraer información valiosa.

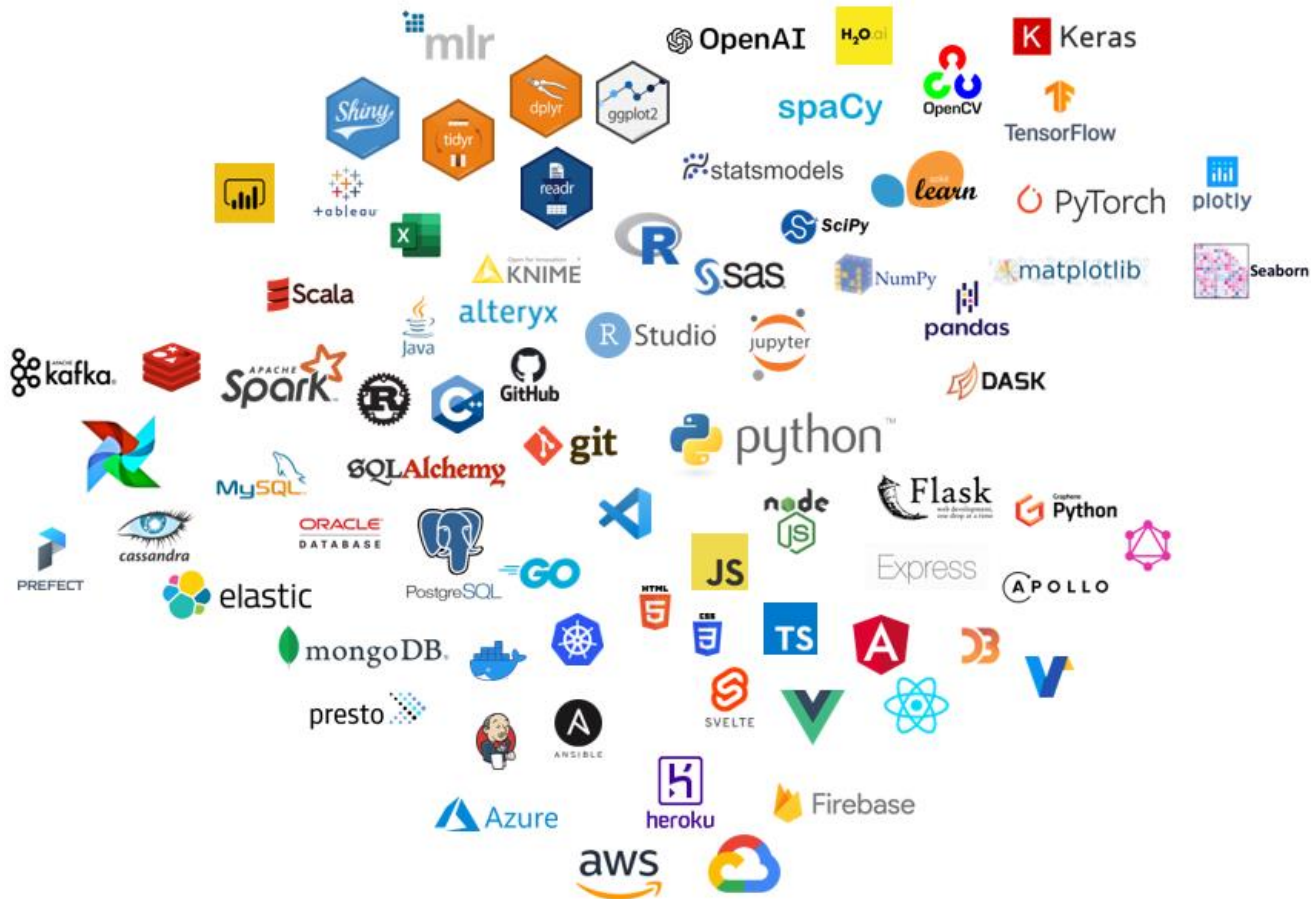
¿Por qué ciencia de datos?



Flujo de datos



Herramientas y tecnologías



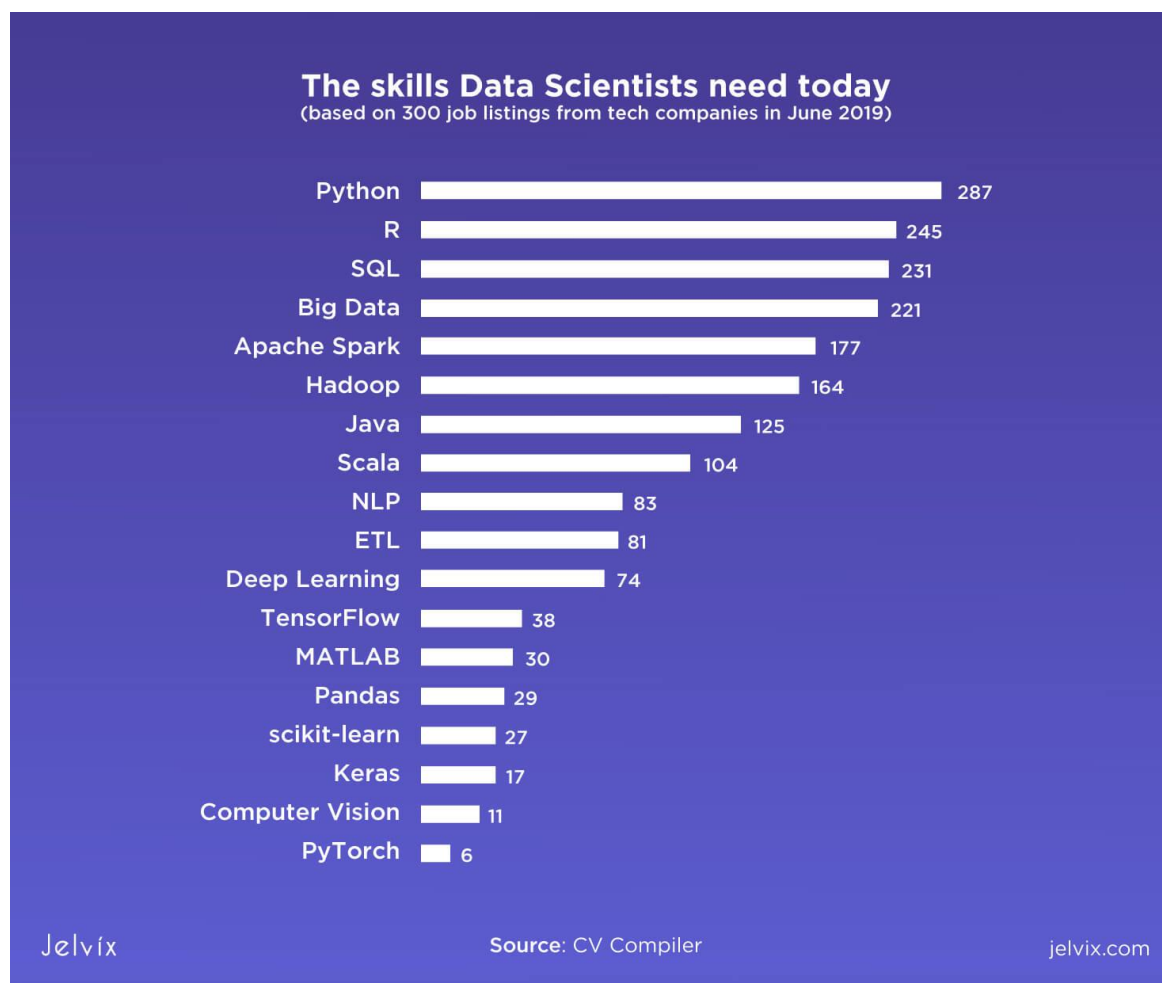
¿Qué es Python?

- Es un lenguaje de programación popular debido a su sintaxis legible y sencilla, lo que lo hace fácil de aprender y usar.
- Python se utiliza en una variedad de aplicaciones, como el desarrollo web, la automatización de tareas, la ciencia de datos, el aprendizaje automático, la inteligencia artificial, la visualización de datos, y mucho más. Es ampliamente utilizado en la industria y en la educación.

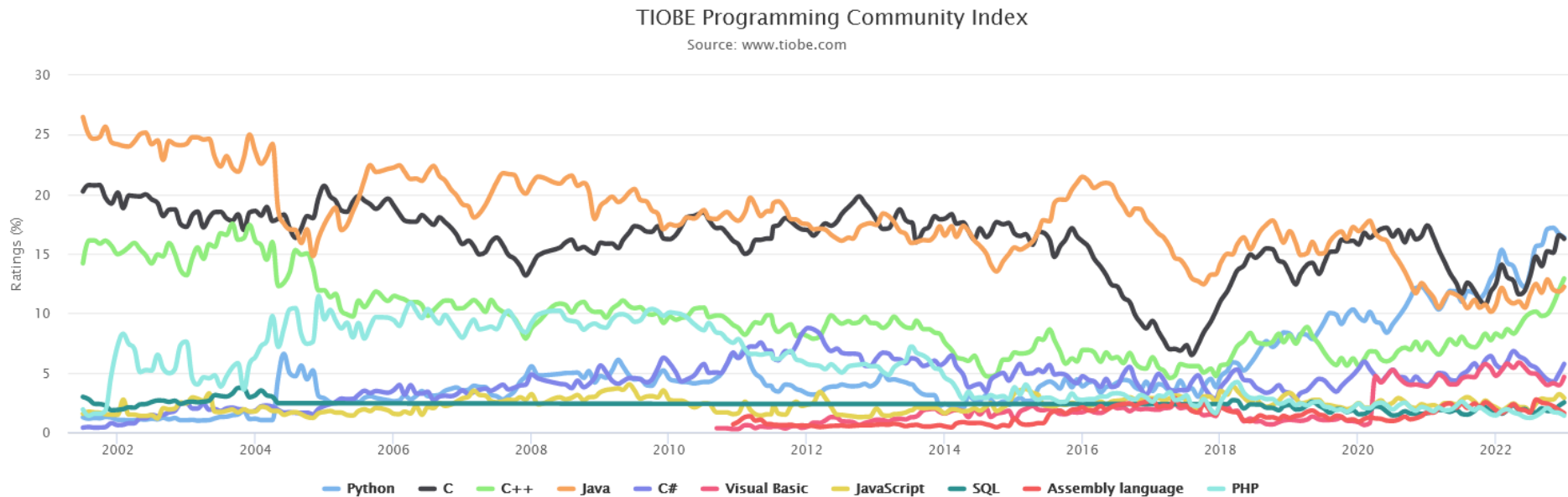
Ventajas

- Código fácil de leer.
 - Hello world: `print("Hello, world!")`
- Propósito general (back-end, front-end, análisis de datos, ...).
- Múltiples paradigmas de programación.
 - Imperativo, Orientado a objetos y funcional.
- Código abierto.
- Extensible (extensiones en C y C++).
- Muchas librerías disponibles para todo tipo de problemas.

Python en ciencia de datos



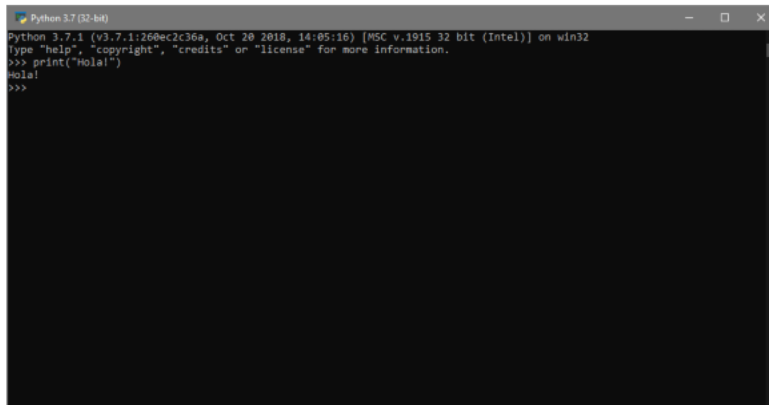
PL en ciencia de datos



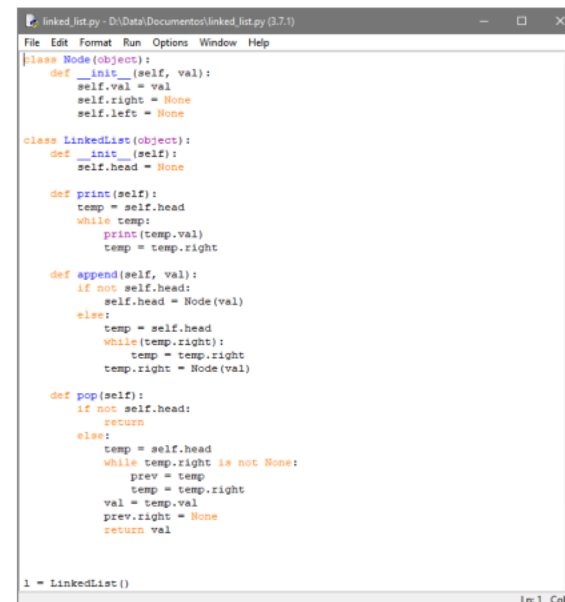
<https://www.datacamp.com/blog/top-programming-languages-for-data-scientists-in-2022>

Formas de utilizar Python

- Desde la línea de comandos de manera interactiva.
- Desde un editor de texto para crear programas ejecutables.
- Desde un IDE para crear programas ejecutables.
- Desde notebook para la ejecución interactiva de código.



```
Python 3.7.1 (v3.7.1:260ec2c36a, Oct 20 2018, 14:05:16) [MSC v.1915 32 bit (Intel)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>> print("Hola!")
Hola!
>>>
```



```
linked_list.py - D:\Data\Documentos\linked_list.py (3.7.1)
File Edit Format Run Options Window Help

class Node(object):
    def __init__(self, val):
        self.val = val
        self.right = None
        self.left = None

class LinkedList(object):
    def __init__(self):
        self.head = None

    def print(self):
        temp = self.head
        while temp:
            print(temp.val)
            temp = temp.right

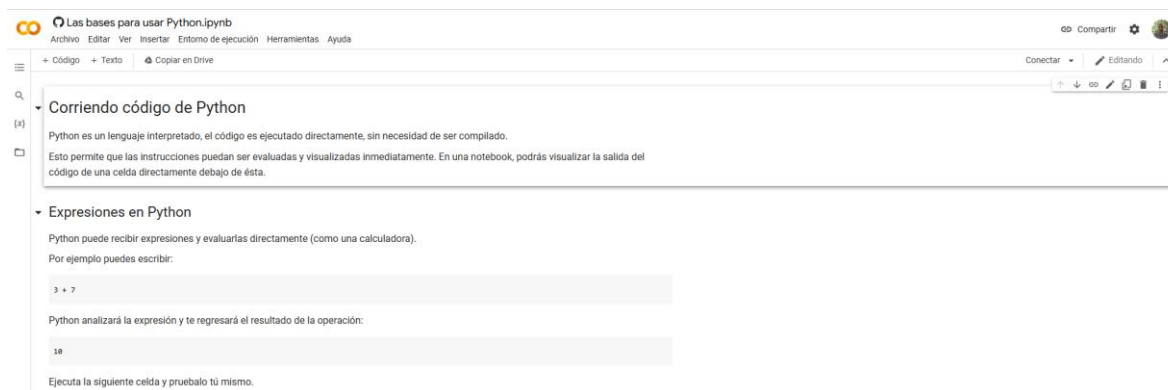
    def append(self, val):
        if not self.head:
            self.head = Node(val)
        else:
            temp = self.head
            while temp.right:
                temp = temp.right
            temp.right = Node(val)

    def pop(self):
        if not self.head:
            return
        else:
            temp = self.head
            while temp.right is not None:
                prev = temp
                temp = temp.right
            val = temp.val
            prev.right = None
            return val

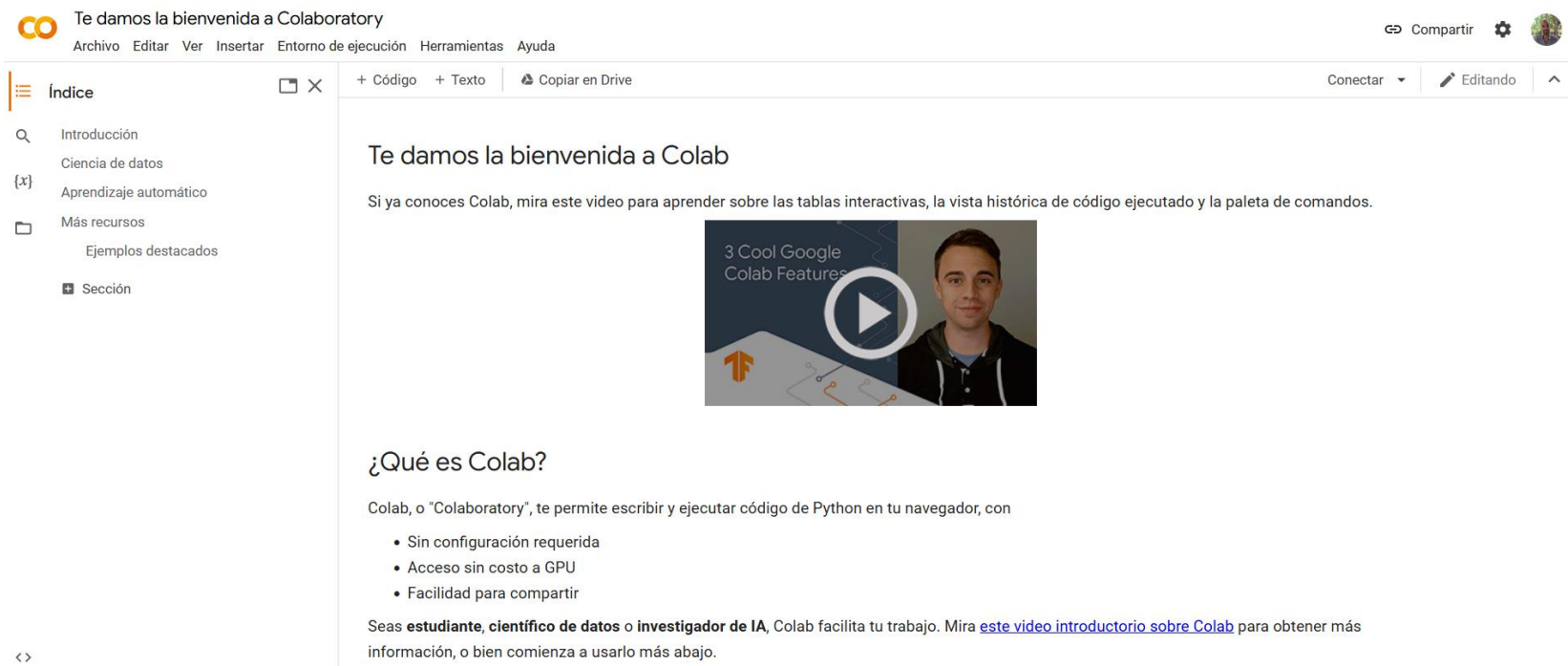
l = LinkedList()
```

Notebooks

- Las notebook son una interfaz de programación que permite integrar procesamiento de texto y la capacidad de ejecutar código del lenguaje de programación.
- Son particularmente útiles para:
 - Análisis de datos (se puede documentar, analizar y visualizar en el mismo entorno)
 - Cursos (capacidad de combinar texto y código en un mismo entorno)



Notebooks en la nube con Colaboratory



The screenshot displays the Google Colaboratory web interface. At the top, a header bar contains the Colab logo, the text "Te damos la bienvenida a Colaboratory", and a menu with options: Archivo, Editar, Ver, Insertar, Entorno de ejecución, Herramientas, and Ayuda. On the right of the header are links for "Compartir", a settings gear, and a user profile icon. Below the header, a sidebar on the left shows an "Índice" (Index) with a search icon and a list of items: "Introducción", "Ciencia de datos", "Aprendizaje automático", "Más recursos", "Ejemplos destacados", and "Sección". The main content area has a title "Te damos la bienvenida a Colab" and a paragraph: "Si ya conoces Colab, mira este video para aprender sobre las tablas interactivas, la vista histórica de código ejecutado y la paleta de comandos." Below this text is a video player showing a man speaking, with the title "3 Cool Google Colab Features" overlaid. Further down, a section titled "¿Qué es Colab?" explains that Colab allows writing and running Python code in a browser. It lists three features: "Sin configuración requerida", "Acceso sin costo a GPU", and "Facilidad para compartir". At the bottom, it encourages students, scientists, data analysts, or IA researchers to use Colab, providing a link to an introductory video.

Te damos la bienvenida a Colaboratory

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda

Compartir

Conectar Editando

Índice

- Introducción
- Ciencia de datos
- Aprendizaje automático
- Más recursos
- Ejemplos destacados
- Sección

Te damos la bienvenida a Colab

Si ya conoces Colab, mira este video para aprender sobre las tablas interactivas, la vista histórica de código ejecutado y la paleta de comandos.

3 Cool Google Colab Features

¿Qué es Colab?

Colab, o "Colaboratory", te permite escribir y ejecutar código de Python en tu navegador, con

- Sin configuración requerida
- Acceso sin costo a GPU
- Facilidad para compartir

Seas **estudiante, científico de datos o investigador de IA**, Colab facilita tu trabajo. Mira [este video introductorio sobre Colab](#) para obtener más información, o bien comienza a usarlo más abajo.

Python en tu computadora

- Además de la opción en la nube, se puede instalar Python y su paquetería de manera local:

- www.python.org



- pypi.org/Project/pip/
- (paquetería)



- www.anaconda.com



- O mediante el manejador de paquetes de tu Sistema Operativo.

Propósito del módulo 1

Python y estadística de datos

- Familiarizarnos con Python como lenguaje de programación
- Conocer las herramientas para análisis de datos utilizando Python
- Entender la importancia del uso eficiente de recursos computacionales para el procesamiento de datos
- Que cuenten con un “cheat-sheet” de los paquetes esenciales para el procesamiento científico y visualización de Python.

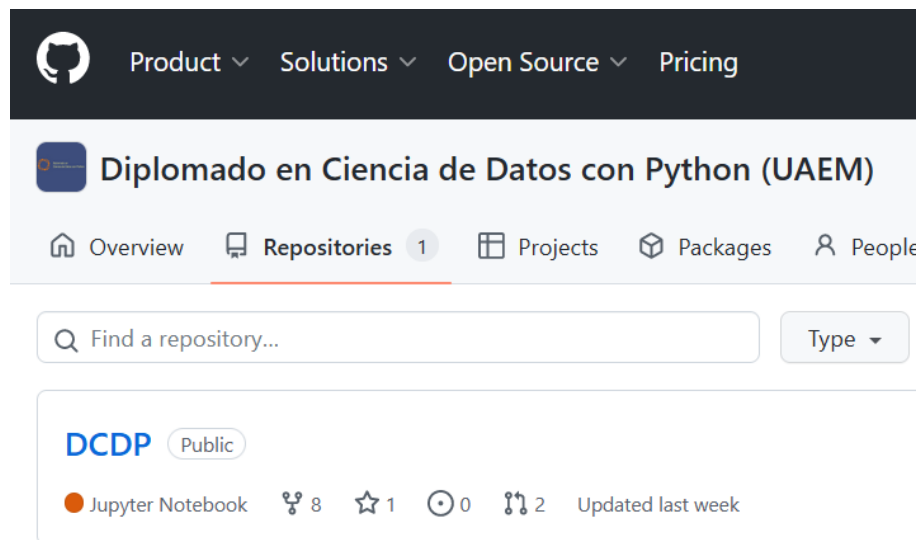
Contenido – Módulo 1

- Bases de Python
- Paquetes para ciencia de datos (Numpy, Pandas, Matplotlib...)
- Estadística descriptiva
- Teoría muestral



Material del diplomado

- El material estará disponible en un repositorio de GitHub. Desde dicho repositorio podrás acceder directamente a las notebooks desde Google Colab.



<https://github.com/DCDP-UAEM/DCDP>

En los siguientes módulos

- Machine learning
 - ¿Qué es ML?
 - Preprocesamiento de datos
 - Clasificadores básicos
 - Agrupamiento
- Deep Learning
 - ¿Qué es DL?
 - Redes MLP
 - Redes de convolución
 - Auto-encoders
 - LSTM