FSCI 2025

#semanticClimate
Transforming information into actionable knowledge

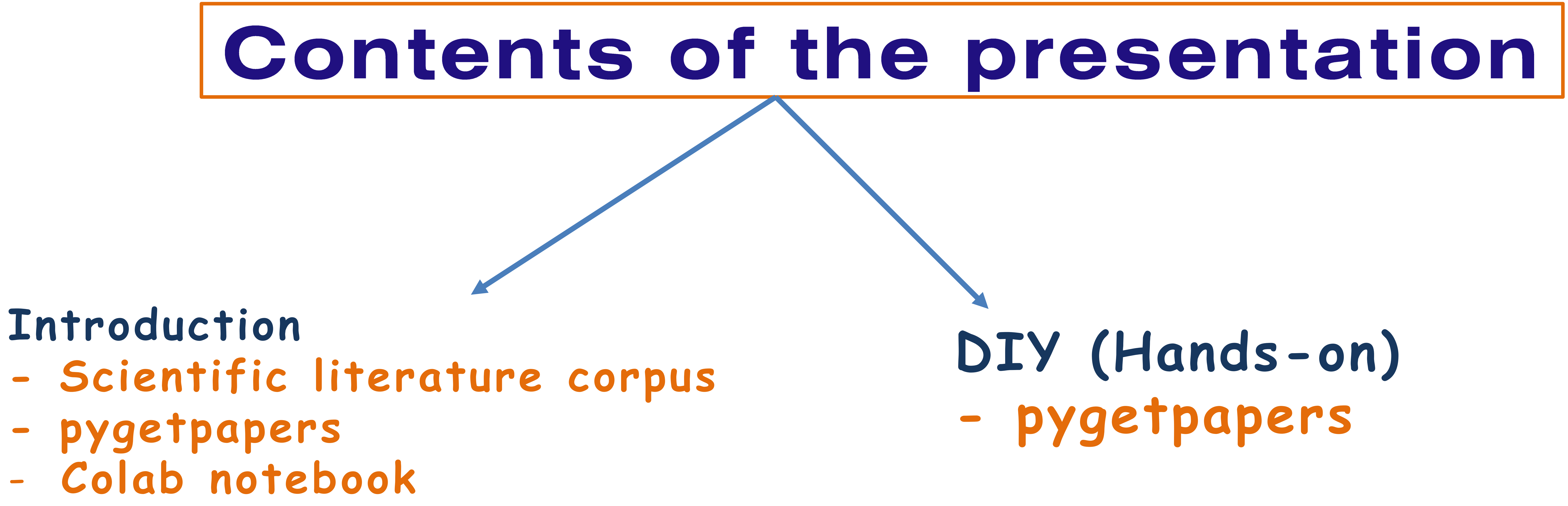
# Corpus creation: Open Access repository retrieval and analysis

22 July, 2025

Presented by: Dr. Renu Kumari

Project Scientist, NIPGR

Program manager, #semanticClimate



# Contents of the presentation

Introduction
- Scientific literature corpus
- pygetpapers
- Colab notebook

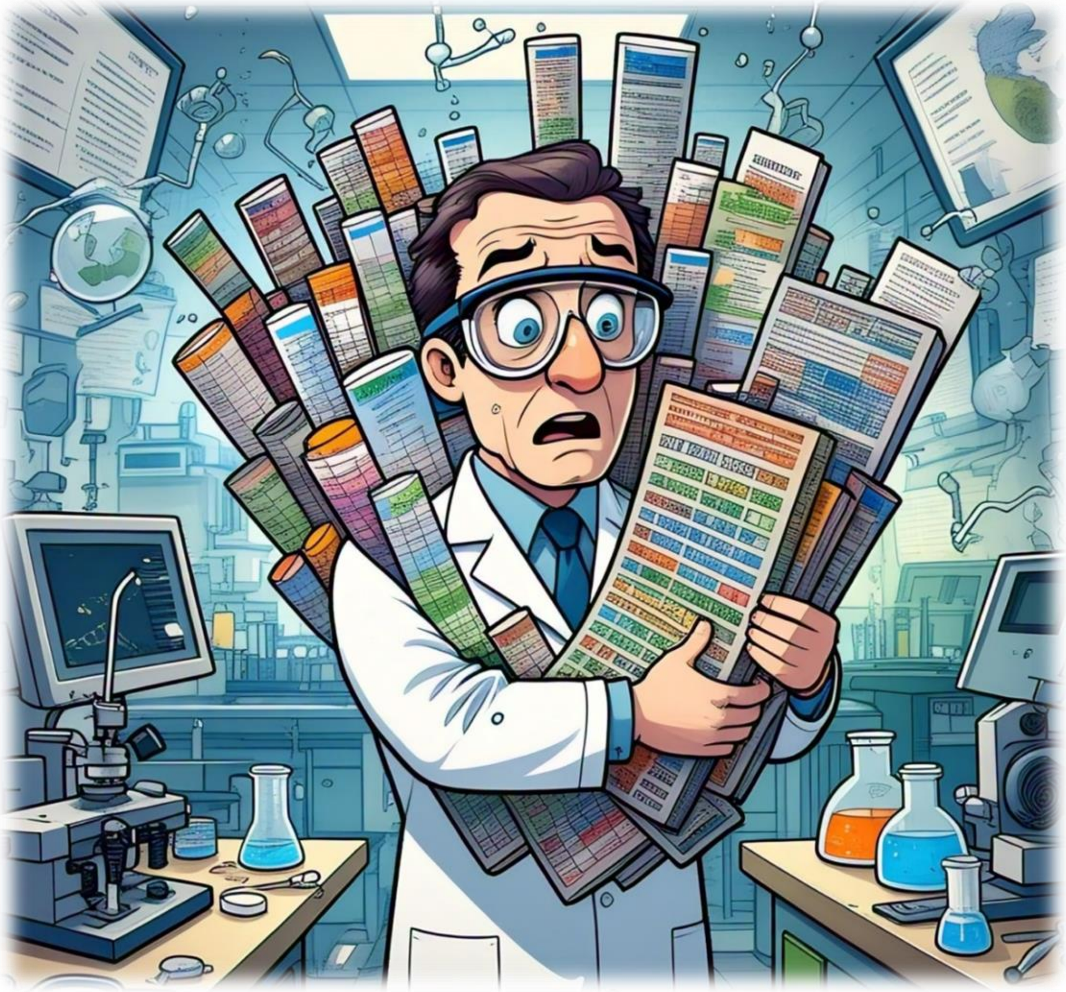
DIY (Hands-on)
- pygetpapers

# What is Scientific Literature corpus?

It is a structured collection of scholarly articles and research papers that can be used for further analysis.

# Why there is a need of Scientific Literature corpus?

# Challenges with Scientific literatures and analysis



Source: Meta AI

- Exponential growth of publications
- Difficult to keep up with the latest developments
- Literature exists in various formats: mainly PDFs
- Not machine-readable or structured formats
- Limited Access to the repositories and journals
- No single platform for getting access to all research outputs
- Bulk downloading is often restricted
- Technical Barriers to automate article retrieval for people with no coding

# Applications of Scientific Literature corpus?

To train **Natural Language Processing (NLP)** models for the following:

- **Named Entity Recognition (NER)**
- **Automated summarization**

**Facilitate literature reviews:**

- **identifying gaps**
- **formulating hypotheses**

So, we need a tool which can create curated corpus in a machine readable format……………….

# #semanticClimate tool used to create corpus

pygetpapers

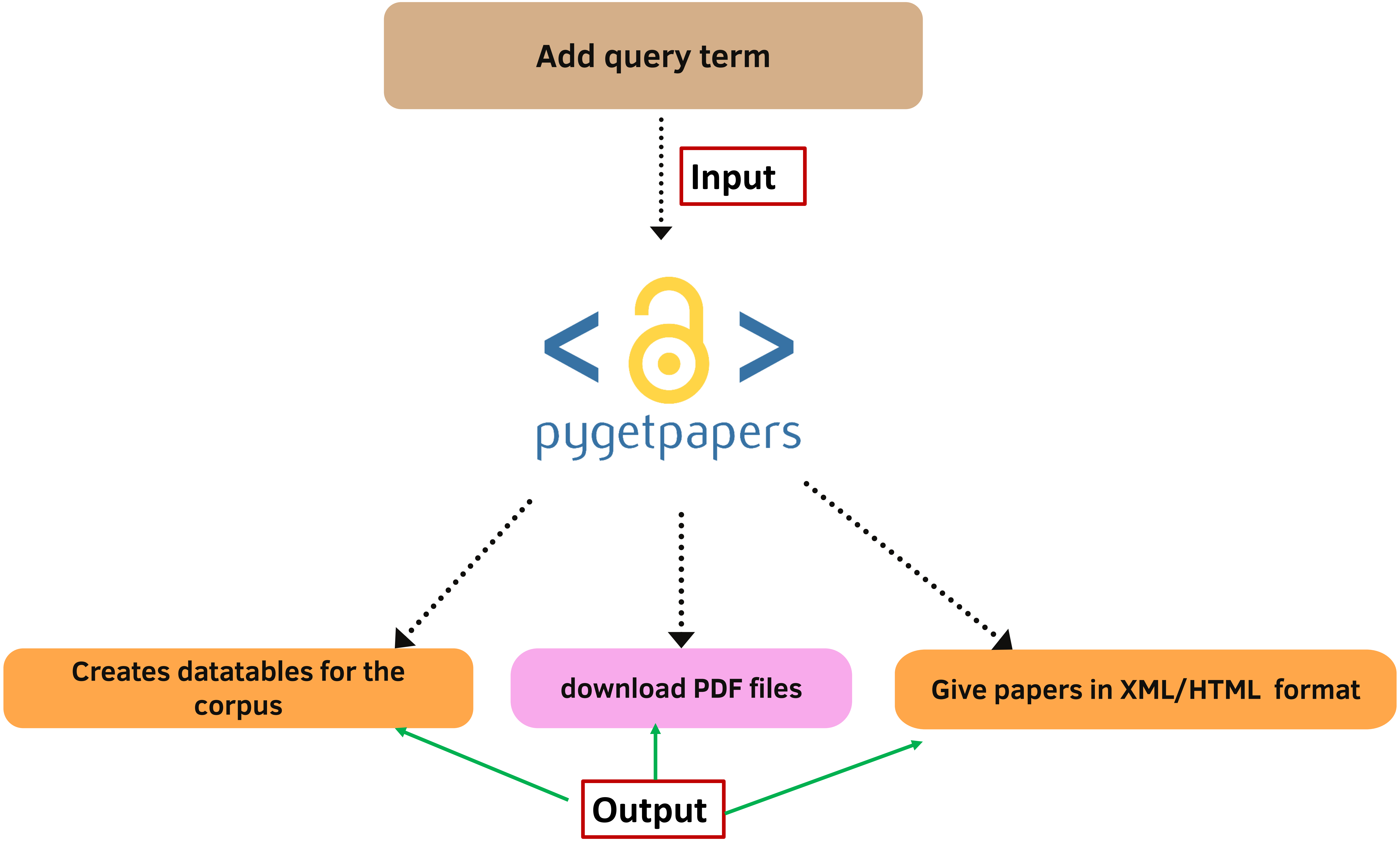
- **pygetpapers** is a tool to assist text miners.

pygetpapers

Developed by: Ayush Garg and Peter Murray-Rust

GitHub repo: https://github.com/petermr/pygetpapers

# Workflow



Add query term

Input

pygetpapers

Creates datatables for the corpus

download PDF files

Give papers in XML/HTML  format

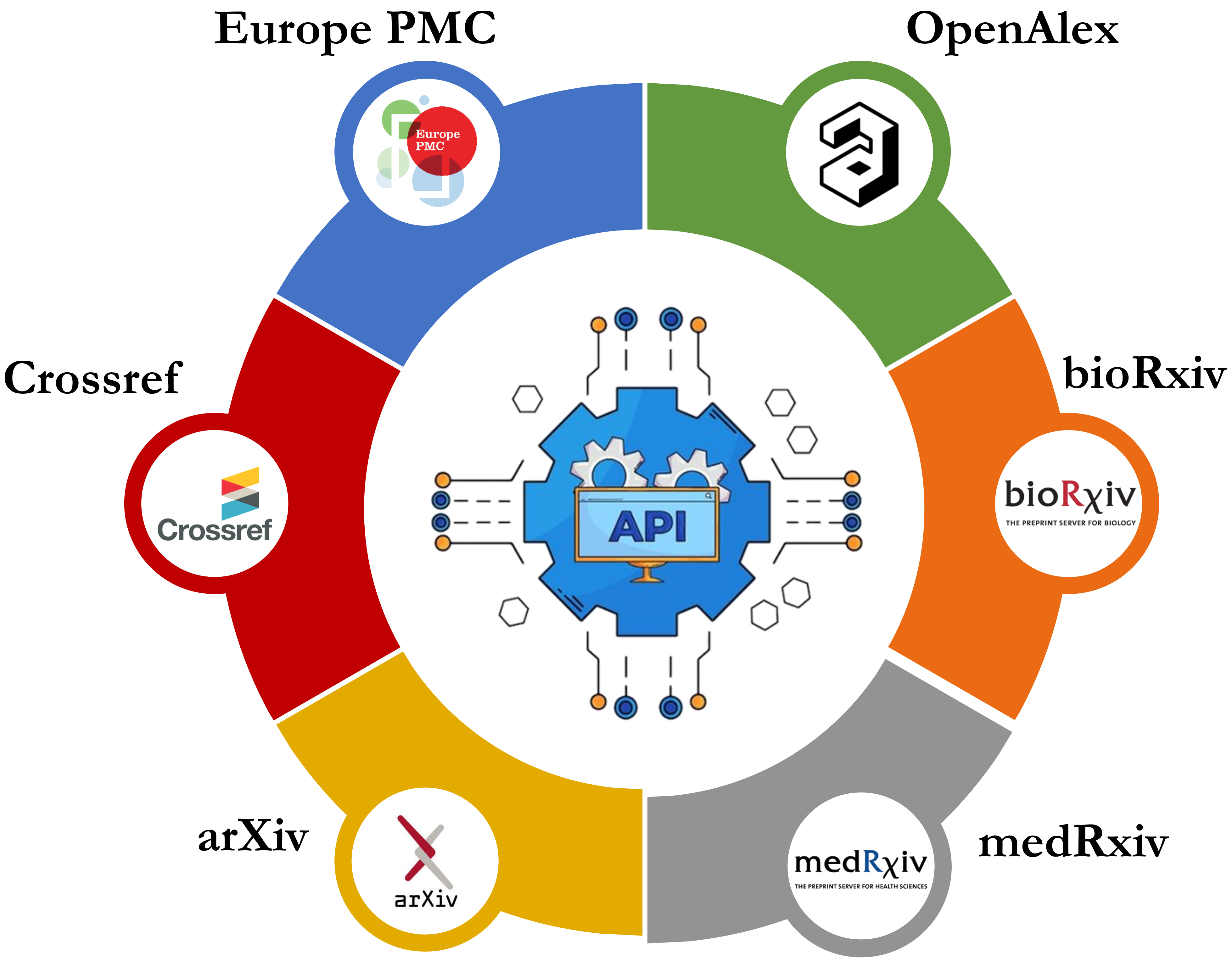
Output

# Query Builder and API support

#semanticClimate
Transforming information into actionable knowledge

**Create Query:**

- Search within a **date range**
- Query with **terms in text file**
- Compound Queries (**AND, OR, NOT**)



**Different API**

# Platform to run the tool

## Google Colab Notebook

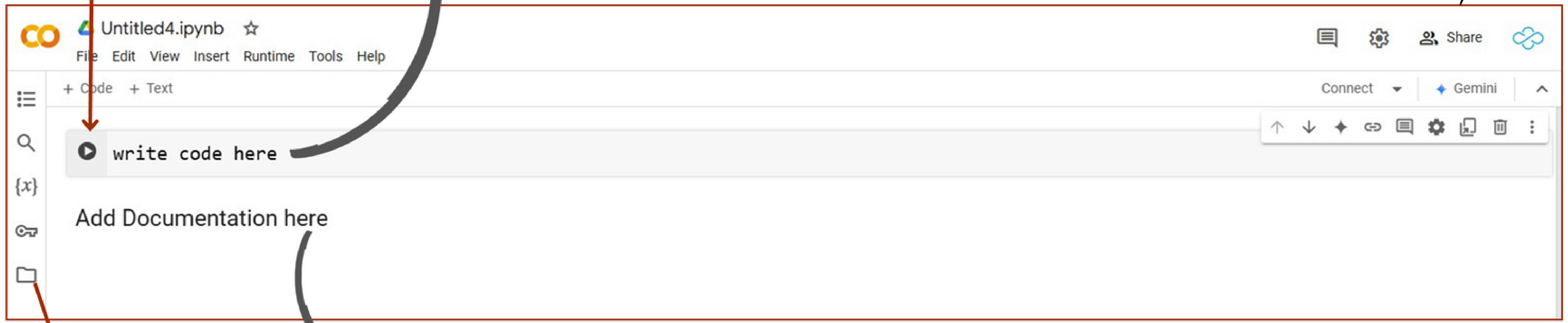
# Importance of Google Colab Notebook

#semanticClimate
Transforming information into actionable knowledge

- Open Jupyter E-Notebook environment
- No pain with setups, versions
- Human Machine friendly
- Supports interactive programming
- Easy learn and explore new tools

# Google Colab (Collaboratory)

*free, cloud-based platform to write and execute python-based codes*



```
write code here
```

Add Documentation here

Click here to run the code

Code cell: for writing code

share the notebook

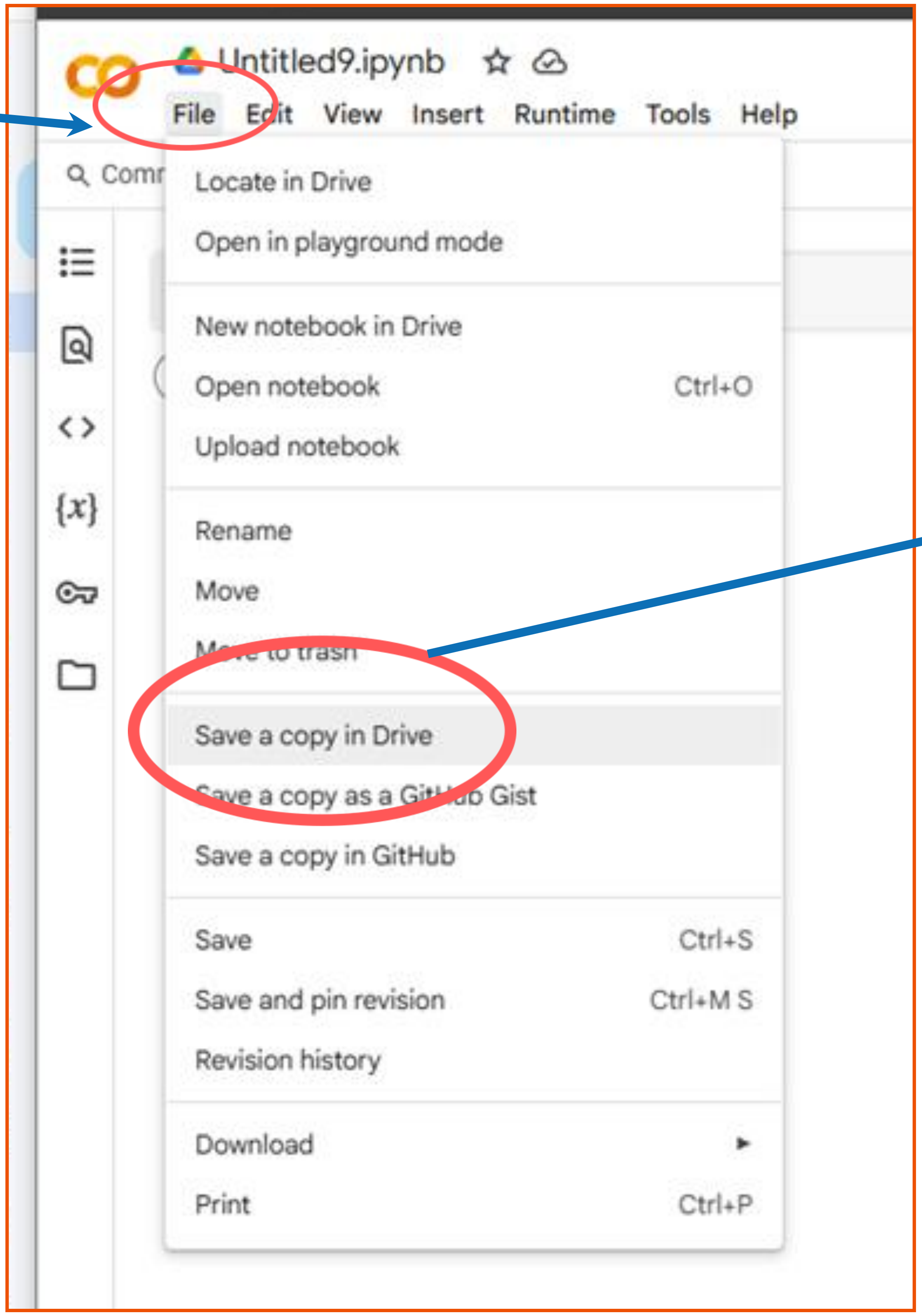
Text cell: for adding documentation

content folder: for saving output

*Need Google account to get started!*

# Google Colab (Collaboratory)

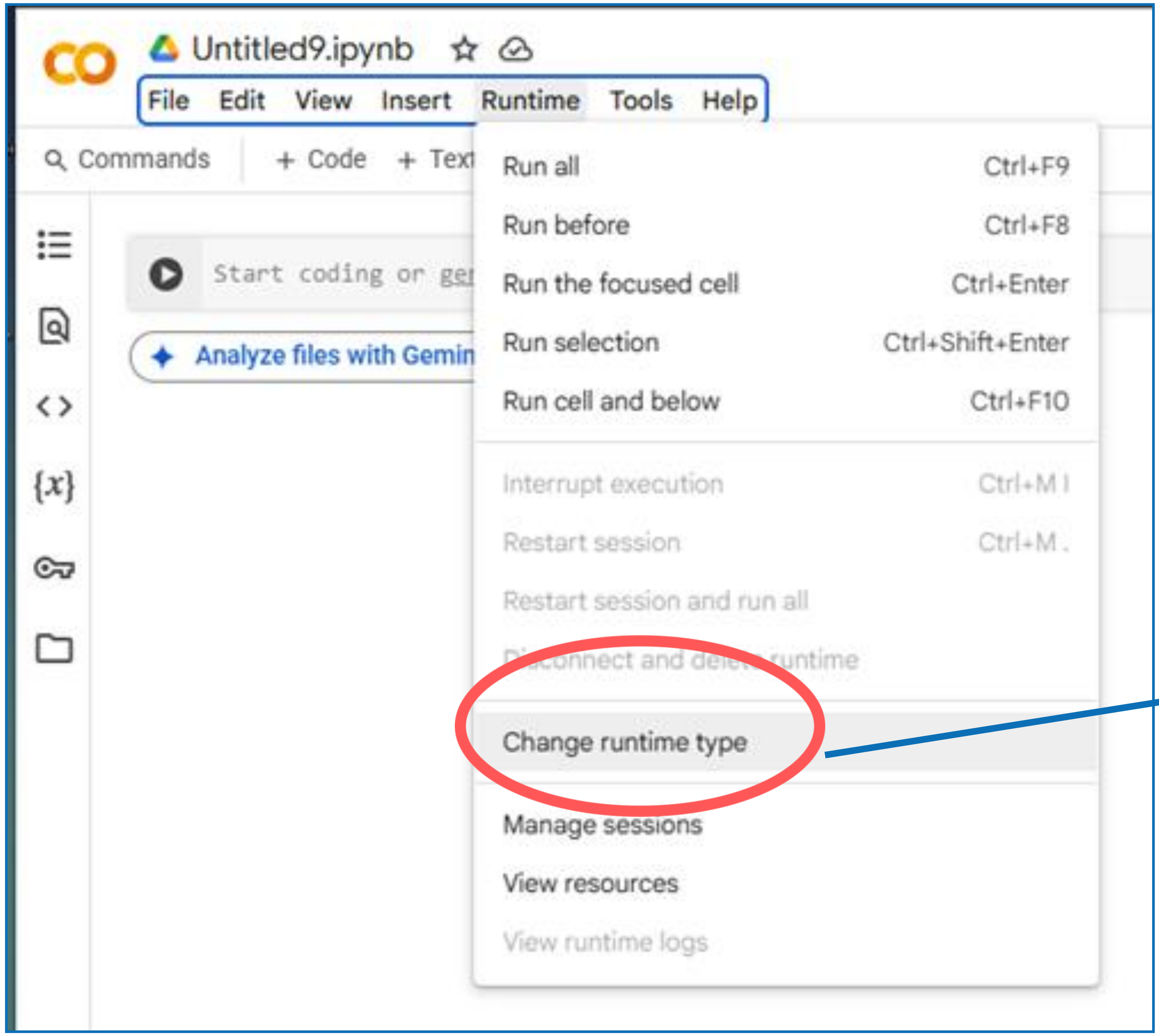
Click to File



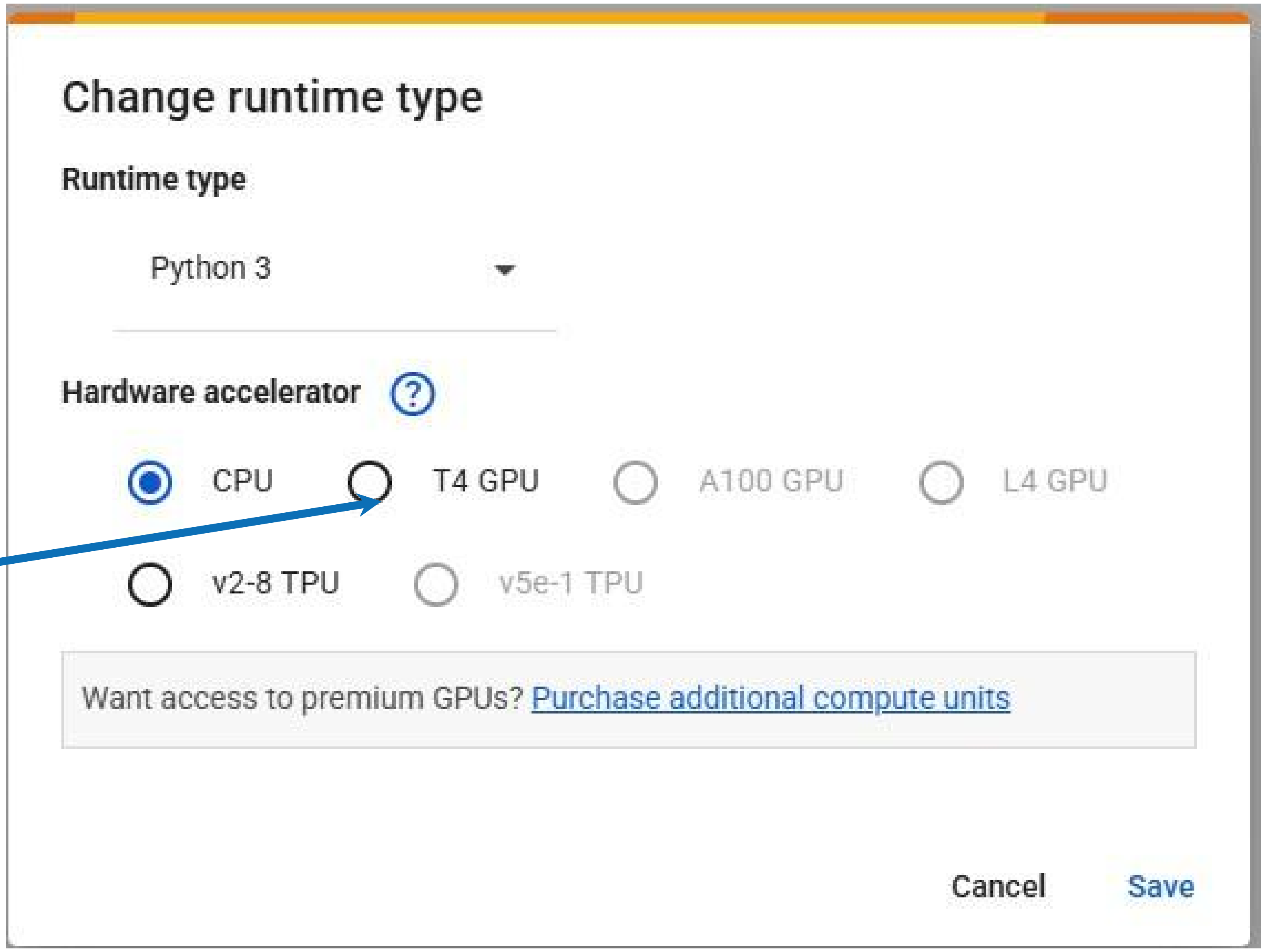
Save a copy in Drive

Before using this Colab, Save a copy to your own Google Drive:
Click on “File” > “Save a copy in Drive”

# Change Runtime



## T4 GPU



Change runtime type

Runtime type: Python 3

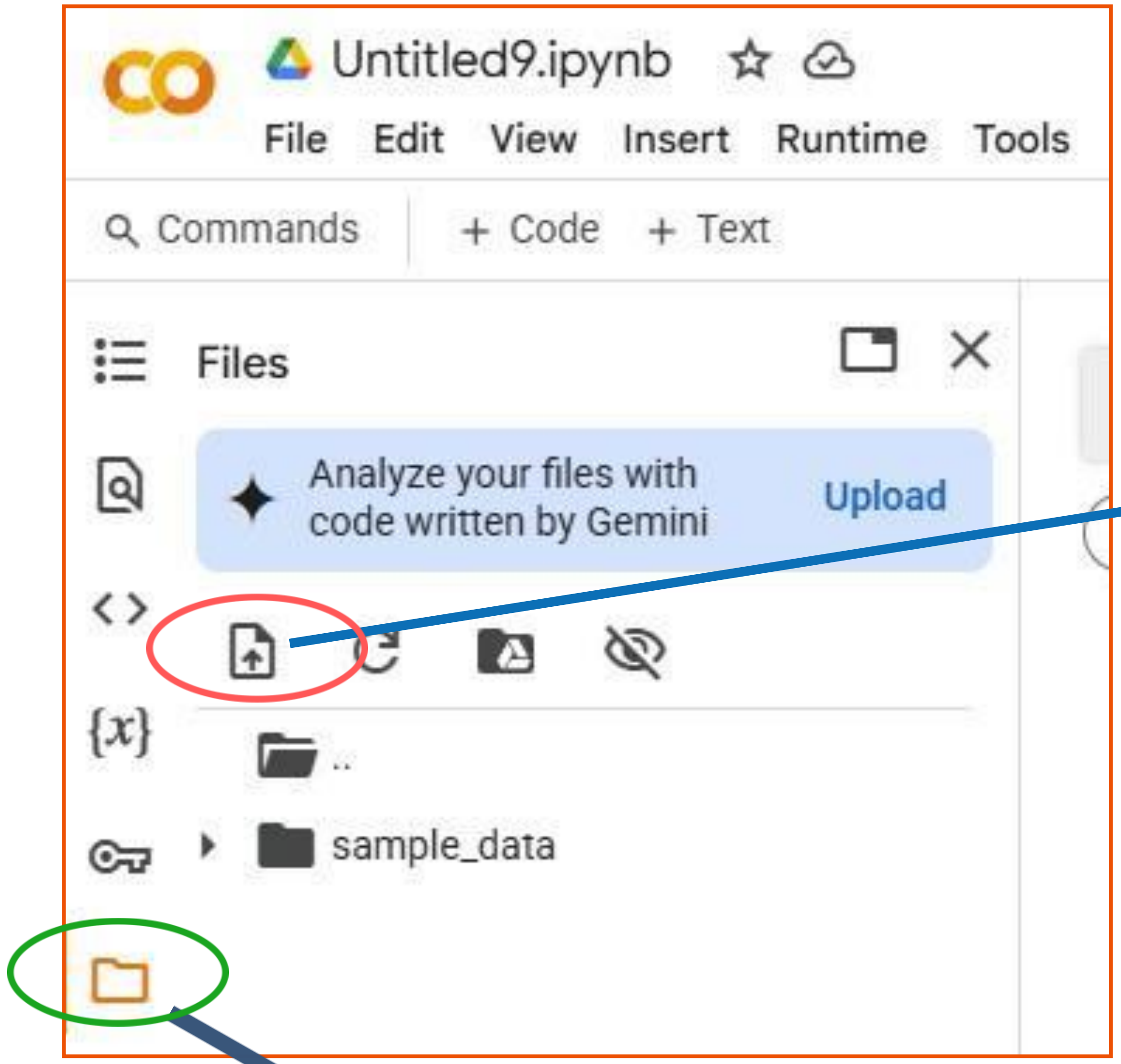
Hardware accelerator

● CPU ○ T4 GPU ○ A100 GPU ○ L4 GPU

○ v2-8 TPU ○ v5e-1 TPU

Want access to premium GPUs? Purchase additional compute units

# Input and Output files

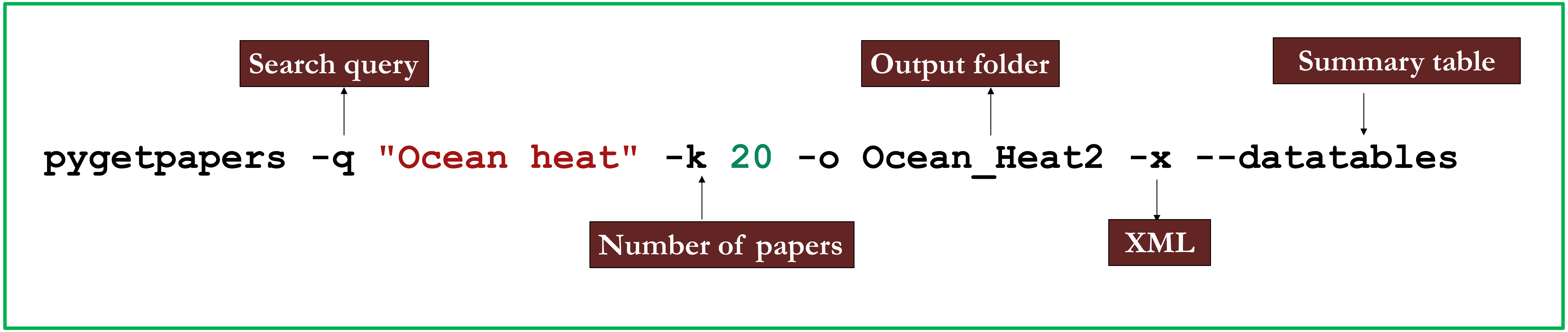

CO
Untitled9.ipynb
File Edit View Insert Runtime Tools
Commands
+ Code + Text
Files
Analyze your files with code written by Gemini
Upload
..
sample_data

Upload the files

For all the results and input files

# Using pygetpapers in colab notebook

# Querying pygetpapers

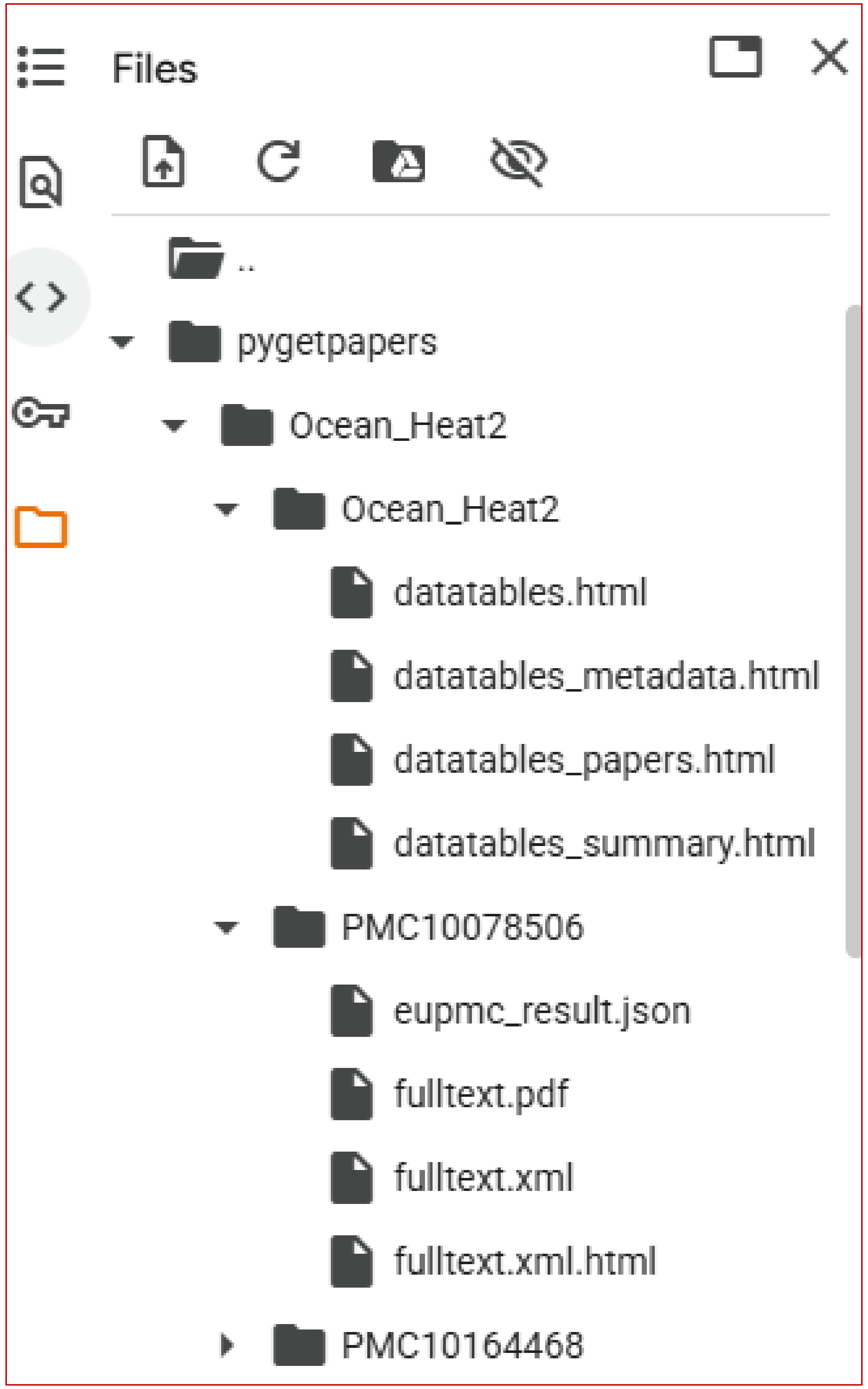


```
pygetpapers -q "Ocean heat" -k 20 -o Ocean_Heat2 -x --datatables
```

**Link to colab notenook**

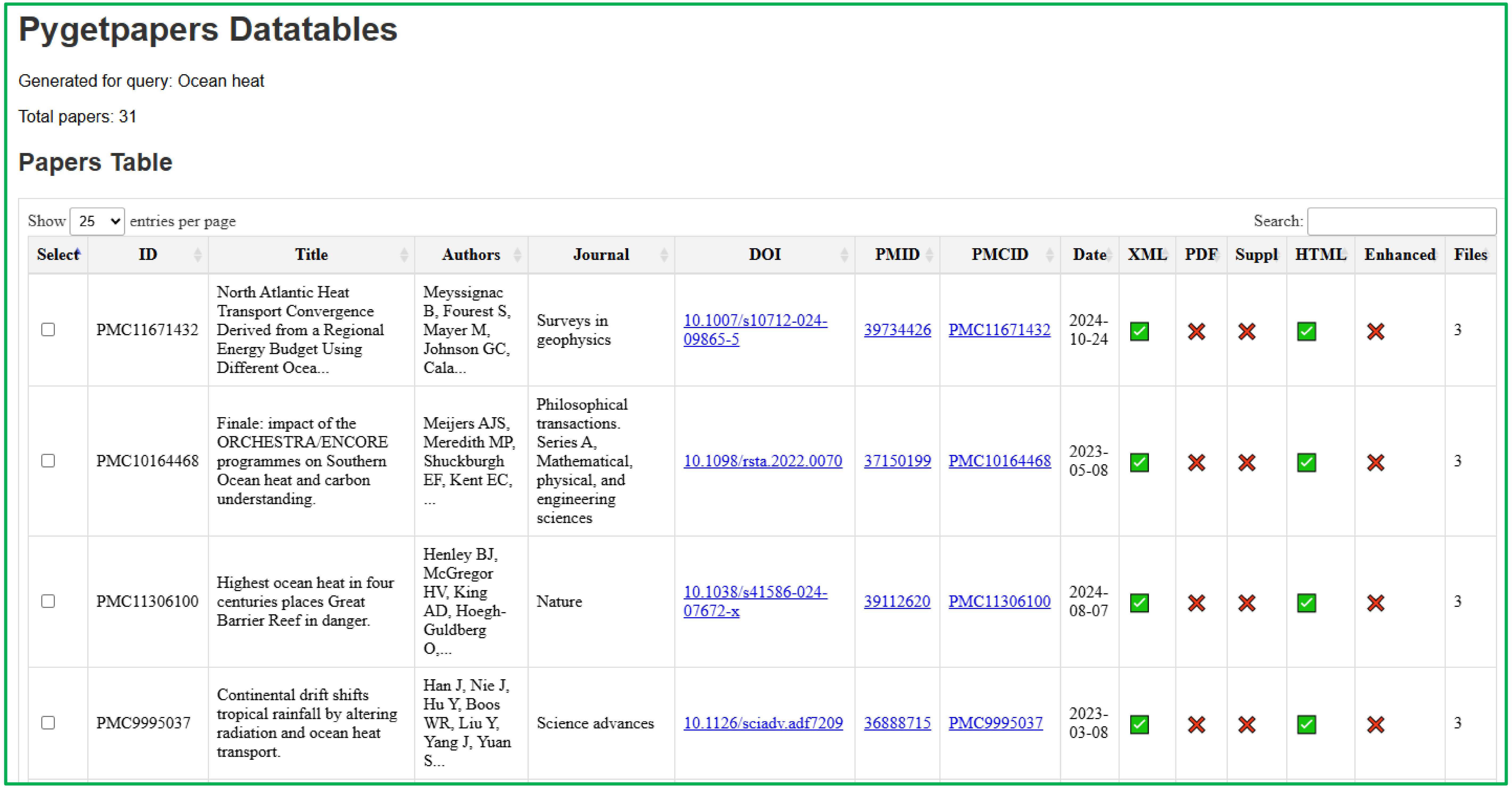
https://colab.research.google.com/drive/1stqd9YxRda2SmSR-r4OLBAGhabJi0vkq?usp=sharing

# Output:



```
..
pygetpapers
  Ocean_Heat2
    Ocean_Heat2
      datatables.html
      datatables_metadata.html
      datatables_papers.html
      datatables_summary.html
    PMC10078506
      eupmc_result.json
      fulltext.pdf
      fulltext.xml
      fulltext.xml.html
    PMC10164468
```

# Datatables



## Pygetpapers Datatables

Generated for query: Ocean heat

Total papers: 31

### Papers Table

Show 25 entries per page

Search:

| Select | ID | Title | Authors | Journal | DOI | PMID | PMCID | Date | XML | PDF | Suppl | HTML | Enhanced | Files |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| □ | PMC11671432 | North Atlantic Heat Transport Convergence Derived from a Regional Energy Budget Using Different Ocea... | Meyssignac B, Fourest S, Mayer M, Johnson GC, Cala... | Surveys in geophysics | 10.1007/s10712-024-09865-5 | 39734426 | PMC11671432 | 2024-10-24 | ✓ | ✗ | ✗ | ✓ | ✗ | 3 |
| □ | PMC10164468 | Finale: impact of the ORCHESTRA/ENCORE programmes on Southern Ocean heat and carbon understanding. | Meijers AJS, Meredith MP, Shuckburgh EF, Kent EC, ... | Philosophical transactions. Series A, Mathematical, physical, and engineering sciences | 10.1098/rsta.2022.0070 | 37150199 | PMC10164468 | 2023-05-08 | ✓ | ✗ | ✗ | ✓ | ✗ | 3 |
| □ | PMC11306100 | Highest ocean heat in four centuries places Great Barrier Reef in danger. | Henley BJ, McGregor HV, King AD, Hoegh-Guldberg O,... | Nature | 10.1038/s41586-024-07672-x | 39112620 | PMC11306100 | 2024-08-07 | ✓ | ✗ | ✗ | ✓ | ✗ | 3 |
| □ | PMC9995037 | Continental drift shifts tropical rainfall by altering radiation and ocean heat transport. | Han J, Nie J, Hu Y, Boos WR, Liu Y, Yang J, Yuan S... | Science advances | 10.1126/sciadv.adf7209 | 36888715 | PMC9995037 | 2023-03-08 | ✓ | ✗ | ✗ | ✓ | ✗ | 3 |

**Link to the output:**

**https://github.com/semanticClimate/ai-automated-literature-review/tree/main/Output_pygetpapers**

# DEMO SESSION

## How to use pygetpapers?

# Link to colab notebook

https://colab.research.google.com/drive/1stqd9YxRda2SmSR-r4OLBAGhabJi0vkq?usp=sharing

# QR for colab notebook

#semanticClimate
Transforming information into actionable knowledge

# THANK YOU

Website : [https://semanticclimate.github.io/p/en/]

email : semanticclimate@gmail.com

X : [@semanticClimate]

LinkedIn : [@semantic Climate]

Git hub : [https://github.com/semanticclimate]