

# 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



- 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

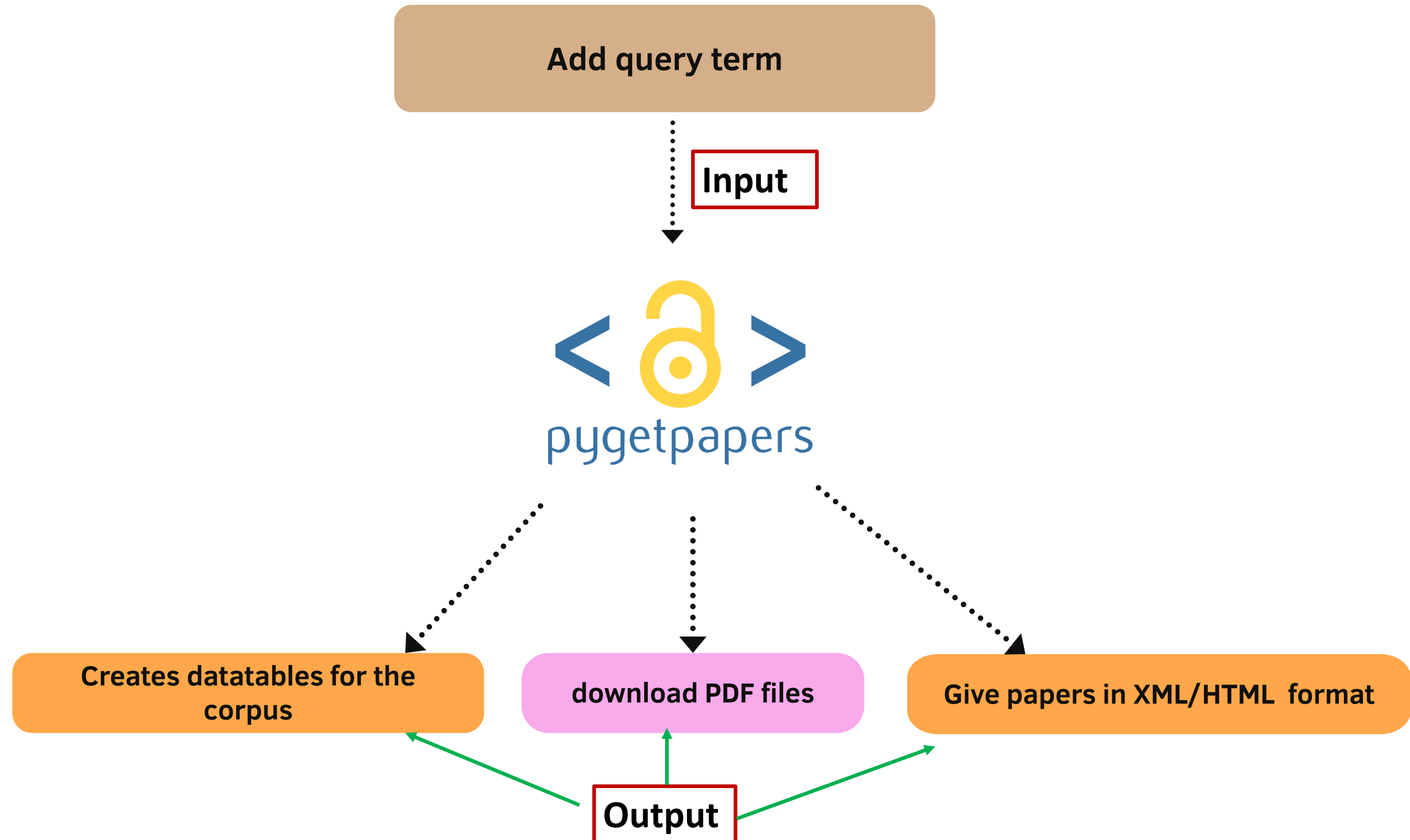
❖ **pygetpapers** is a tool to assist text miners.



Developed by: **Ayush Garg** and  
Peter Murray-Rust

**GitHub repo:** <https://github.com/petermr/pygetpapers>

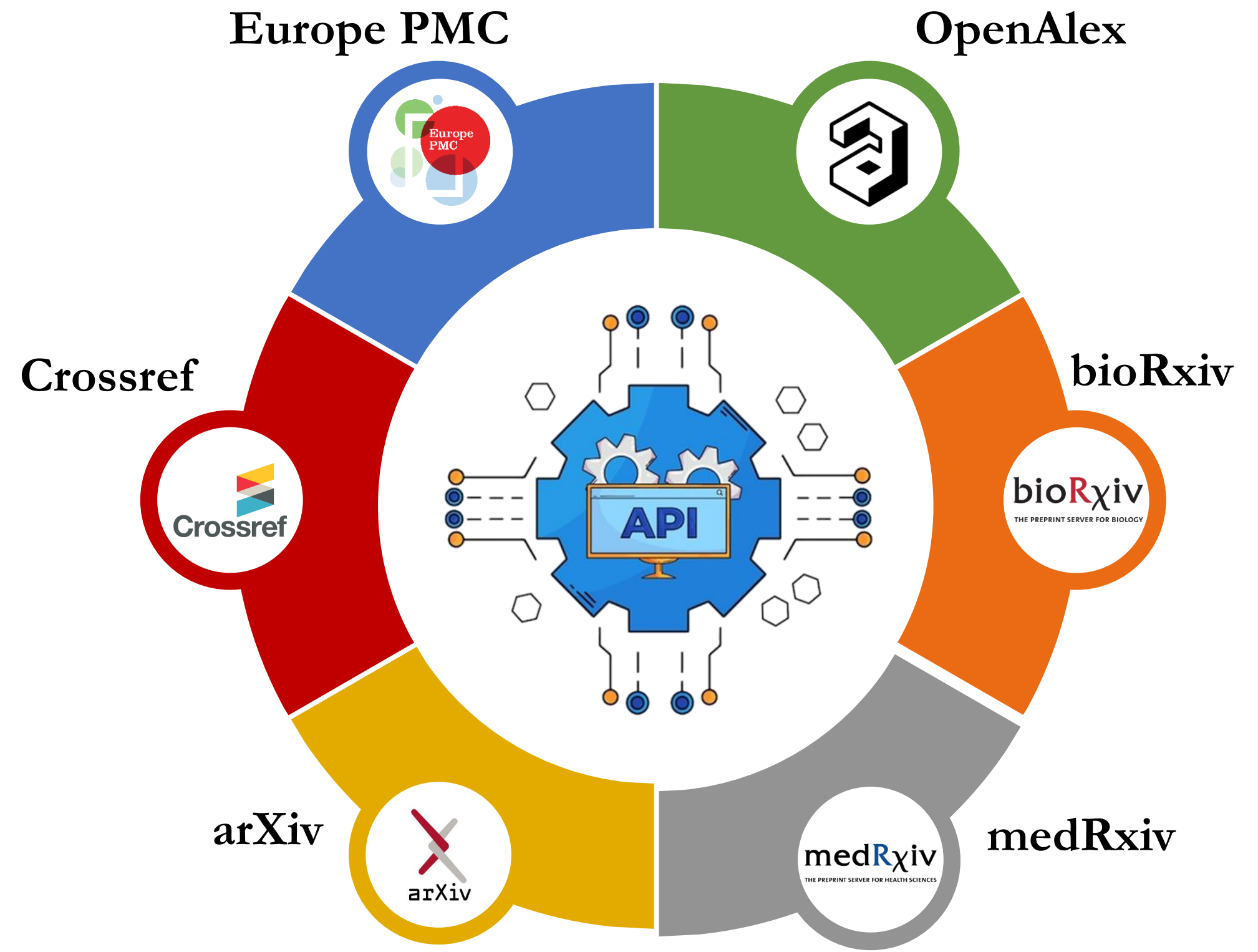
# Workflow



# Query Builder and API support

## Create Query:

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



# Platform to run the tool

**Google Colab Notebook**



- 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)

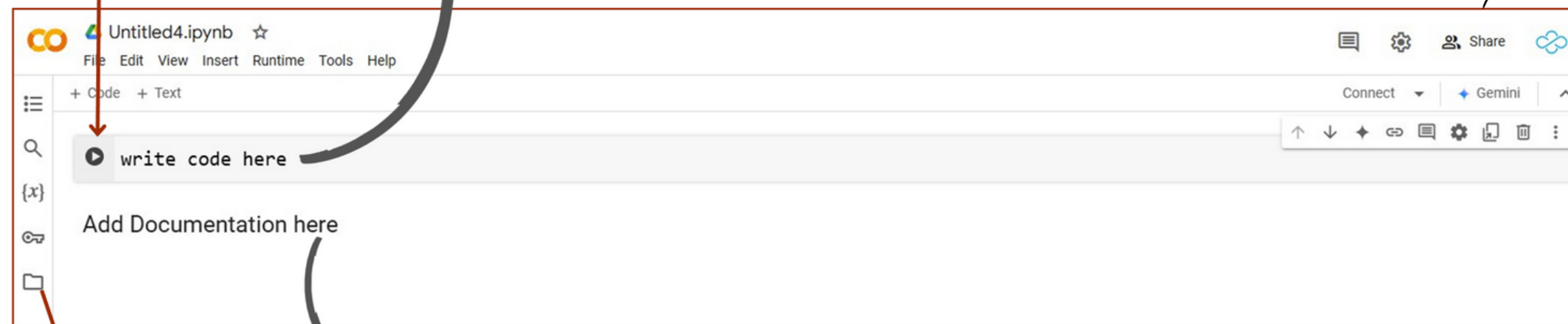


*free, cloud-based platform to write and execute python-based codes*

Click here to run  
the code

Code cell: for writing code

share the notebook



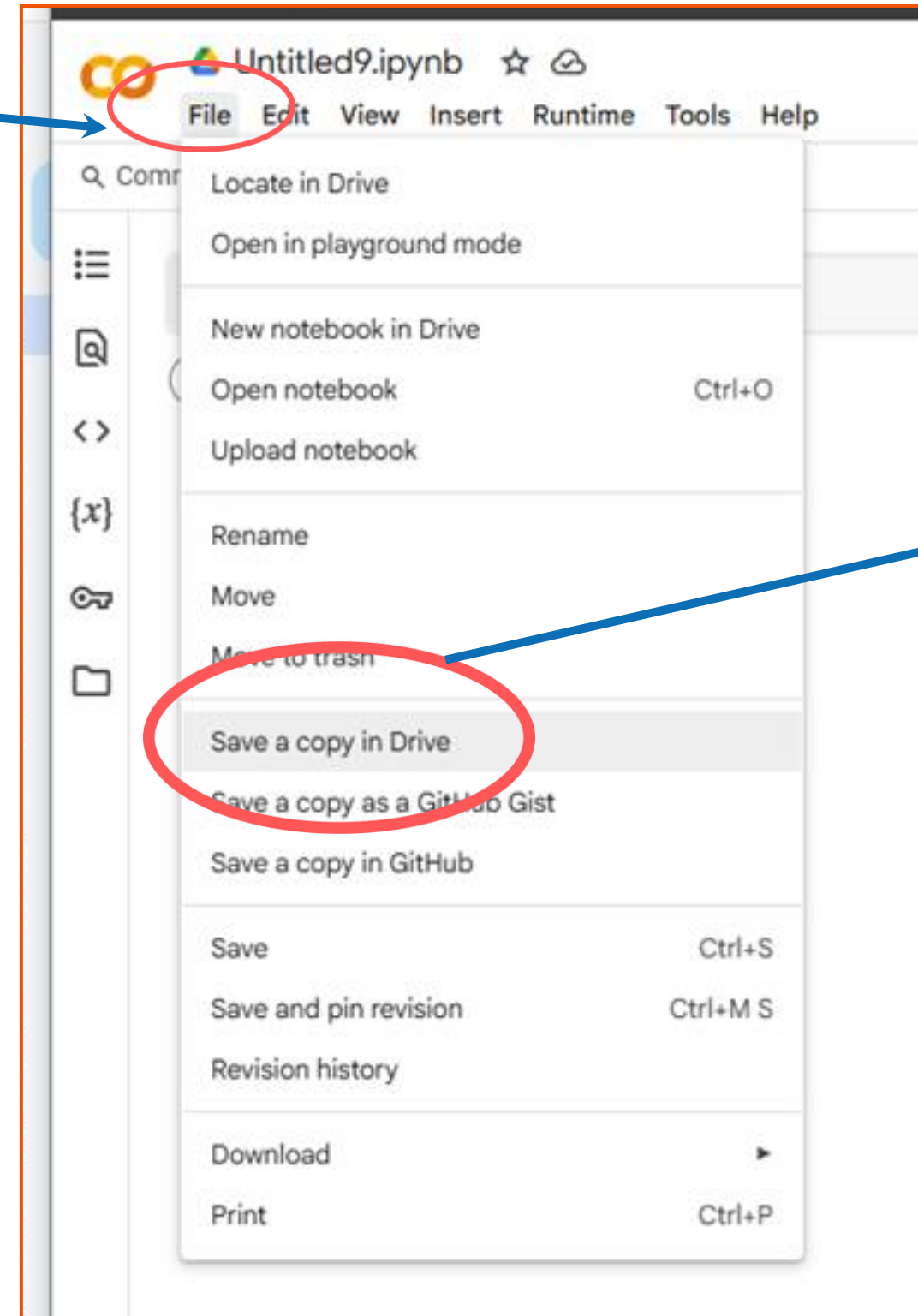
content folder: for saving output

Text cell: for adding documentation

*Need Google account to get started!*

# Google Colab (Collaboratory)

Click to File

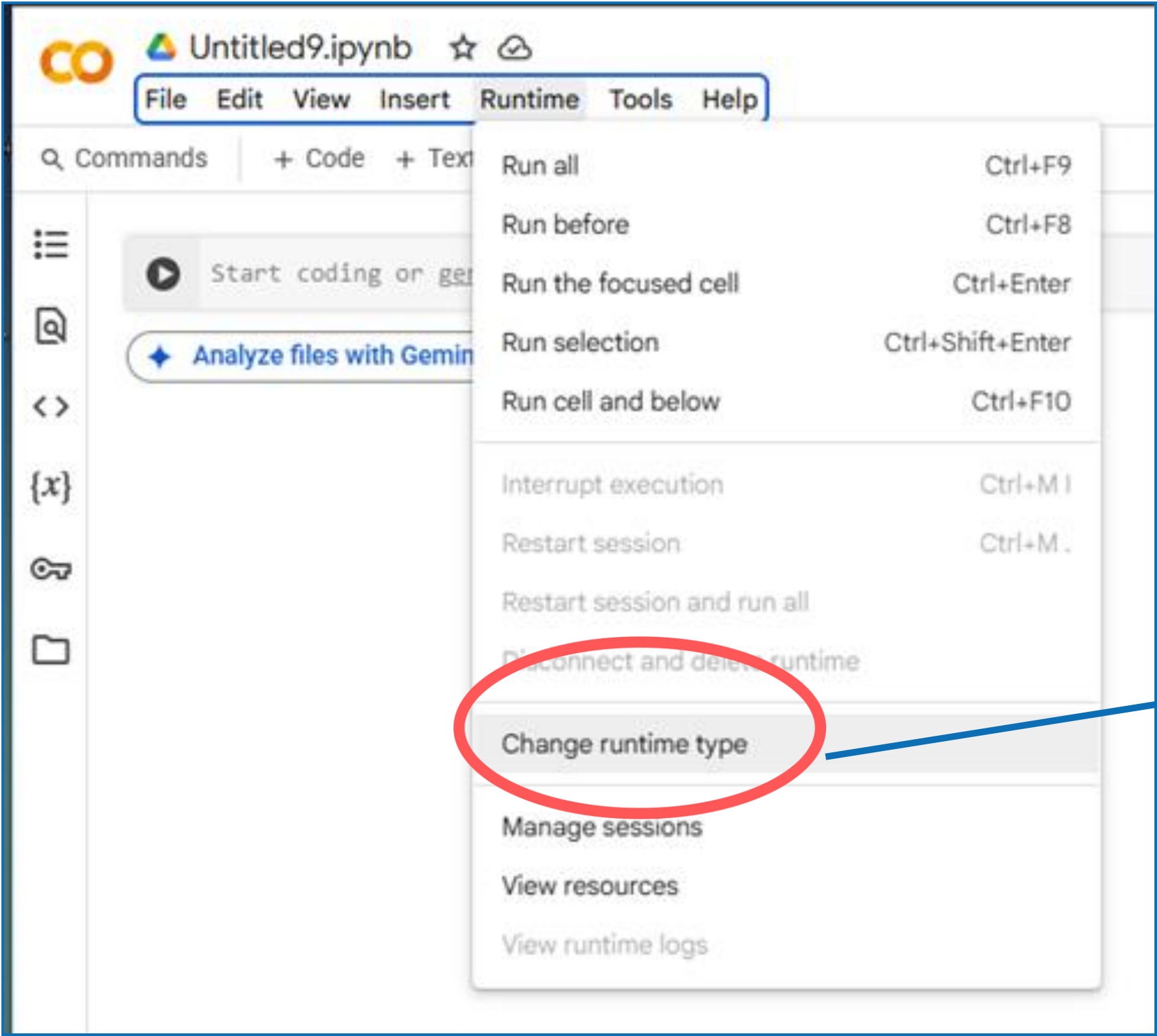


Save a copy in Drive

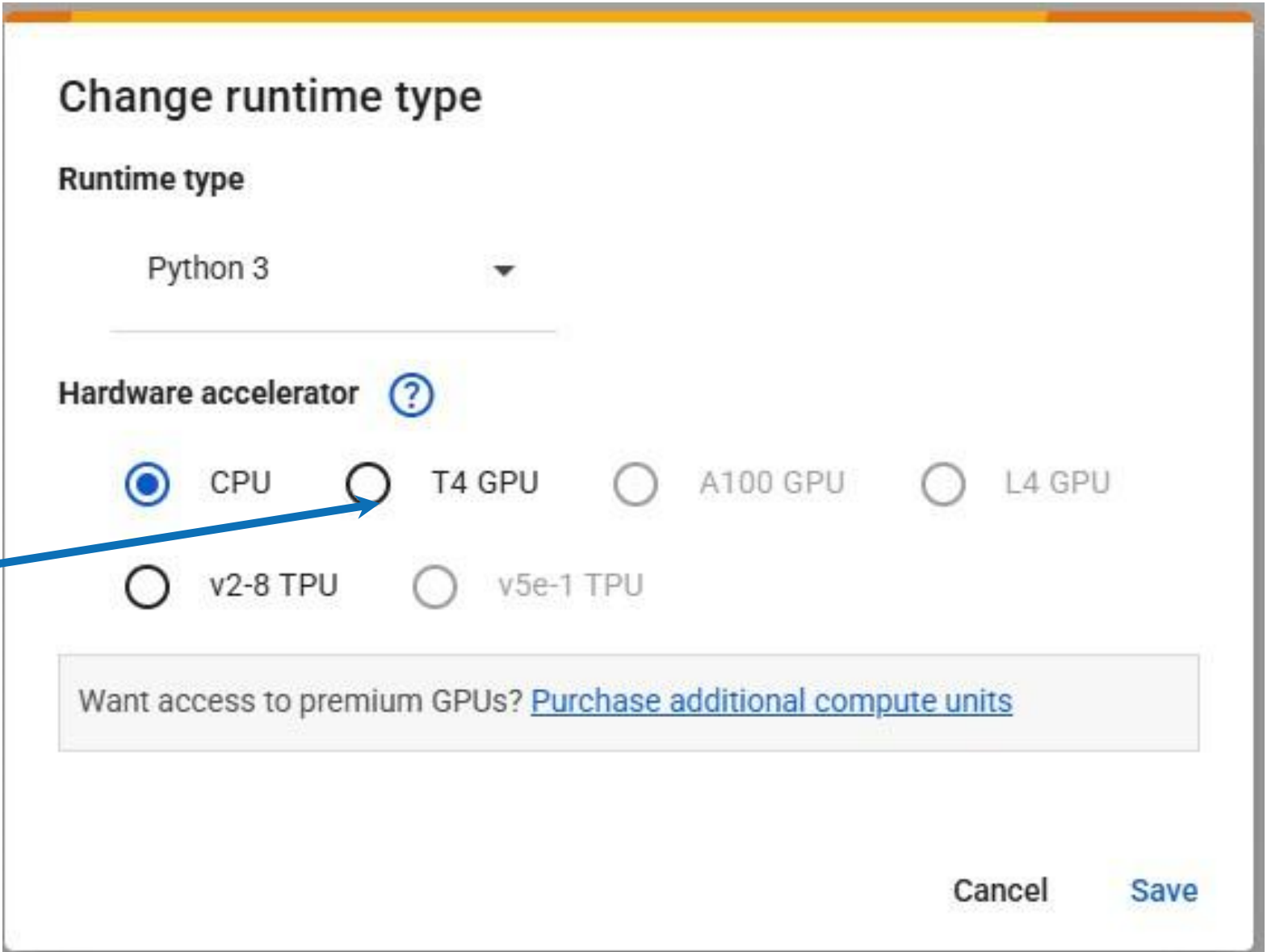
Before using this Colab, Save a copy to your own Google Drive:  
Click on “File” > “Save a copy in Drive”



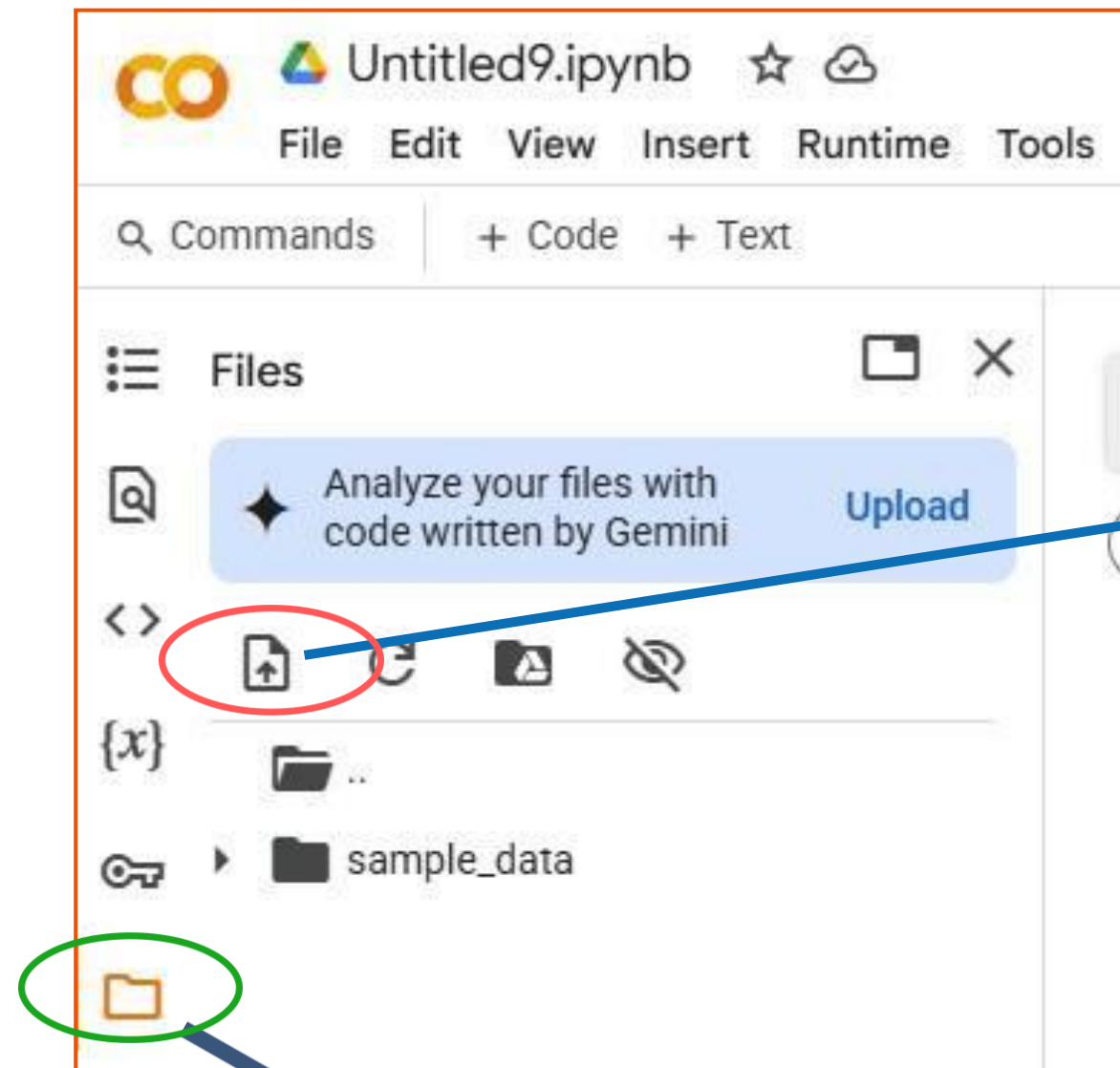
# Change Runtime



T4 GPU



# Input and Output files

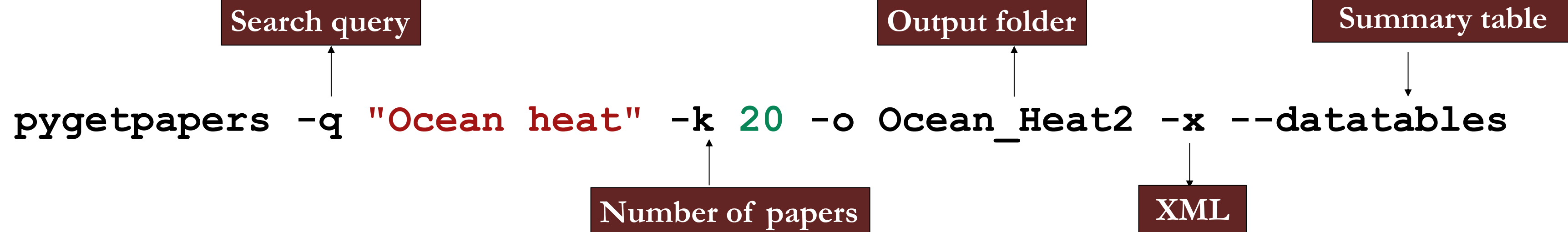


**Upload the files**

**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
```

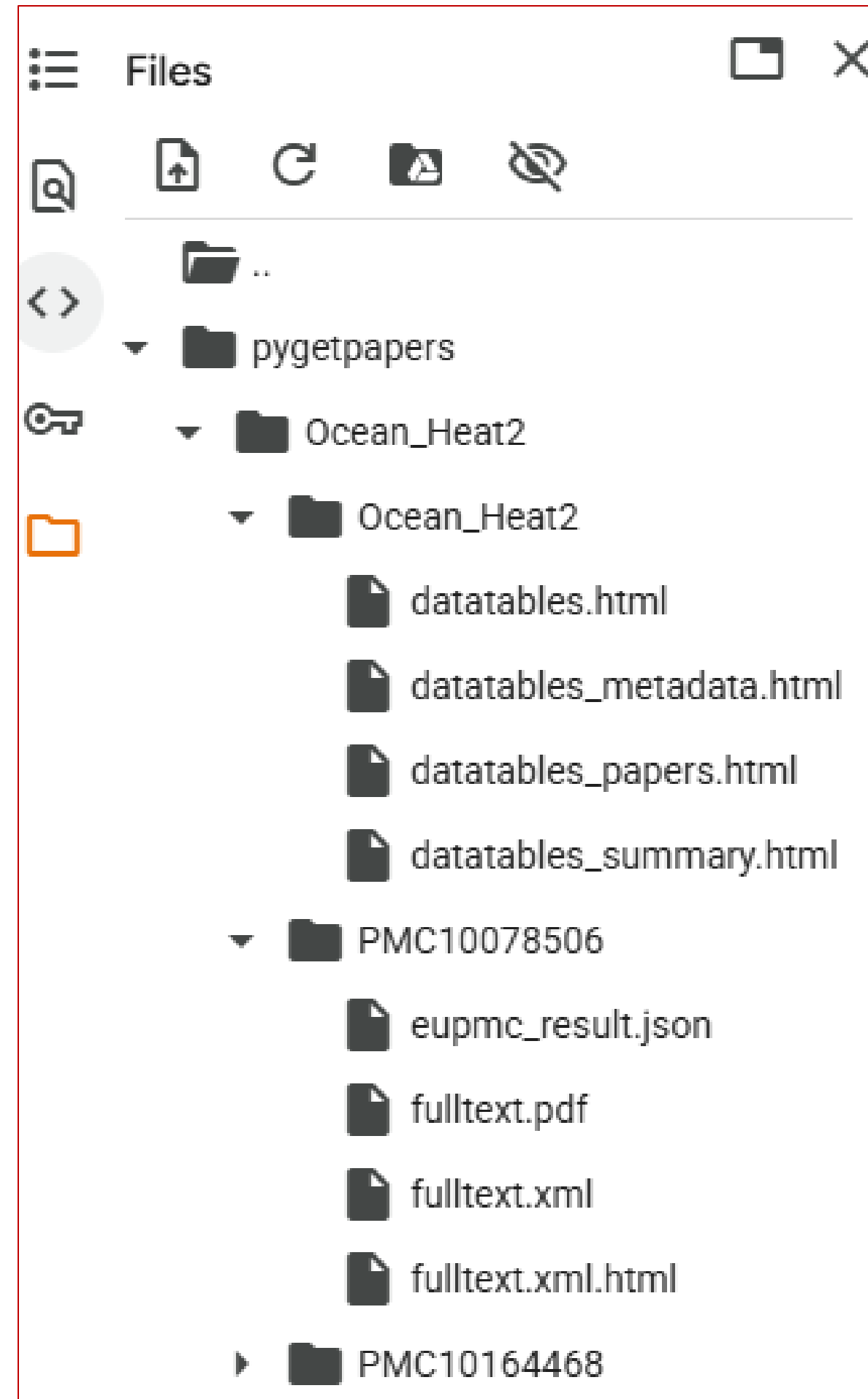
The diagram illustrates the components of the `pygetpapers` command. Arrows point from descriptive labels to specific parts of the command:

- Search query** points to `-q "Ocean heat"`.
- Number of papers** points to `-k 20`.
- Output folder** points to `-o Ocean_Heat2`.
- XML** points to `-x`.
- Summary table** points to `--datatables`.

**Link to colab notenook**

<https://colab.research.google.com/drive/1stqd9YxRda2SmSR-r40LBAGhabJi0vbkq?usp=sharing>

# Output:



# 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
<input type="checkbox"/>	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	<a href="https://doi.org/10.1007/s10712-024-09865-5">10.1007/s10712-024-09865-5</a>	<a href="https://pubmed.ncbi.nlm.nih.gov/39734426/">39734426</a>	<a href="https://europepmc.org/article/PMC/PMC11671432">PMC11671432</a>	2024-10-24	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	3
<input type="checkbox"/>	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	<a href="https://doi.org/10.1098/rsta.2022.0070">10.1098/rsta.2022.0070</a>	<a href="https://pubmed.ncbi.nlm.nih.gov/37150199/">37150199</a>	<a href="https://europepmc.org/article/PMC/PMC10164468">PMC10164468</a>	2023-05-08	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	3
<input type="checkbox"/>	PMC11306100	Highest ocean heat in four centuries places Great Barrier Reef in danger.	Henley BJ, McGregor HV, King AD, Hoegh-Guldberg O,...	Nature	<a href="https://doi.org/10.1038/s41586-024-07672-x">10.1038/s41586-024-07672-x</a>	<a href="https://pubmed.ncbi.nlm.nih.gov/39112620/">39112620</a>	<a href="https://europepmc.org/article/PMC/PMC11306100">PMC11306100</a>	2024-08-07	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	3
<input type="checkbox"/>	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	<a href="https://doi.org/10.1126/sciadv.adf7209">10.1126/sciadv.adf7209</a>	<a href="https://pubmed.ncbi.nlm.nih.gov/36888715/">36888715</a>	<a href="https://europepmc.org/article/PMC/PMC9995037">PMC9995037</a>	2023-03-08	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	3

Link to the output:

[https://github.com/semanticClimate/ai-automated-literature-review/tree/main/Output\\_pygetpapers](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-r40LBAGhabJi0vbkq?usp=sharing>

## QR for colab notebook





# THANK YOU



**#semanticClimate**  
Transforming information into actionable knowledge