Mahidol University
Faculty of Information
and Communication Technology

SERU

TYPHON

Automatic Recommendation of Relevant Code Cells in Jupyter Notebooks

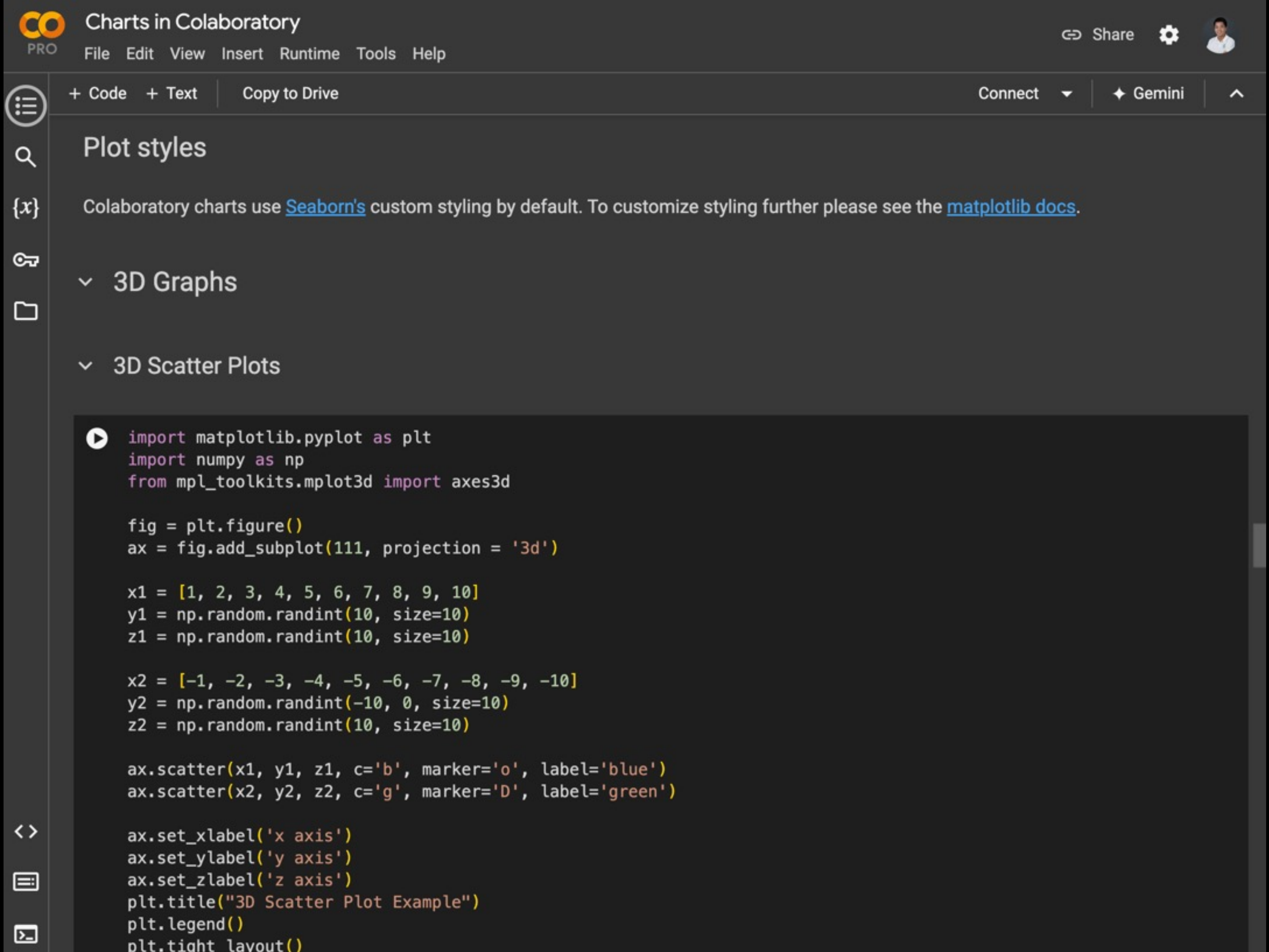
Chaiyong Ragkhitwetsagul, Veerakit Prasertpol,
Natanon Ritta, Paphon Sae-Wong,
Thanapon Noraset and Morakot Choetkiertikul
Faculty of ICT, Mahidol University

JCSSE 2024, 19-21 July 2024

COMPUTATIONAL NOTEBOOK

Computational notebook is a well-known and well-adopted technology in tasks related to data analysis

A Jupyter notebook can be a central place for collaborative data analysis.



Plot styles

Colaboratory charts use Seaborn's custom styling by default. To customize styling further please see the matplotlib docs.

3D Graphs

3D Scatter Plots

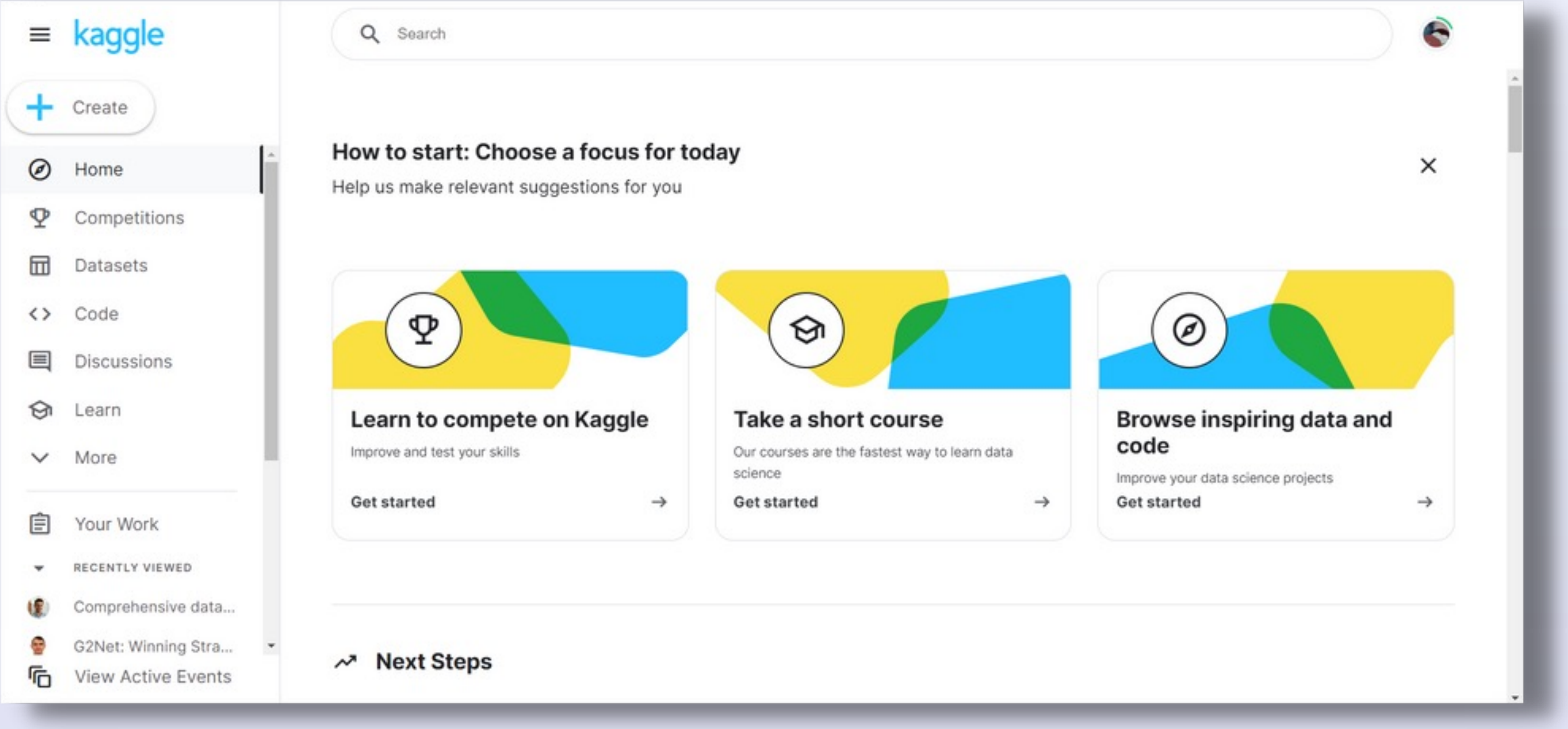
```
import matplotlib.pyplot as plt
import numpy as np
from mpl_toolkits.mplot3d import axes3d

fig = plt.figure()
ax = fig.add_subplot(111, projection = '3d')

x1 = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
y1 = np.random.randint(10, size=10)
z1 = np.random.randint(10, size=10)

x2 = [-1, -2, -3, -4, -5, -6, -7, -8, -9, -10]
y2 = np.random.randint(-10, 0, size=10)
z2 = np.random.randint(10, size=10)

ax.scatter(x1, y1, z1, c='b', marker='o', label='blue')
ax.scatter(x2, y2, z2, c='g', marker='D', label='green')

ax.set_xlabel('x axis')
ax.set_ylabel('y axis')
ax.set_zlabel('z axis')
plt.title("3D Scatter Plot Example")
plt.legend()
plt.tight_layout()
```

kaggle

A CLOUD-BASED COLLABORATIVE PLATFORM INVOLVING DATA ANALYTICS TASKS USING A COMPUTATIONAL NOTEBOOK IN PRACTICES

FOR EXAMPLE:

- MACHINE LEARNING
- COMPETE IN REAL-WORLD PROBLEMS
- FIND OR PUBLISH DATASET
- DISCUSS WITH OTHERS



kaggle

Search

Create

Home

Competitions

Datasets

Code

Discussions

Learn

More

Your Work

RECENTLY VIEWED

Comprehensive data...

G2Net: Winning Stra...

View Active Events

How to start: Choose a focus for today

Help us make relevant suggestions for you

Learn to compete on Kaggle

Improve and test your skills

Get started →

Take a short course

Our courses are the fastest way to learn data science

Get started →

Browse inspiring data and code

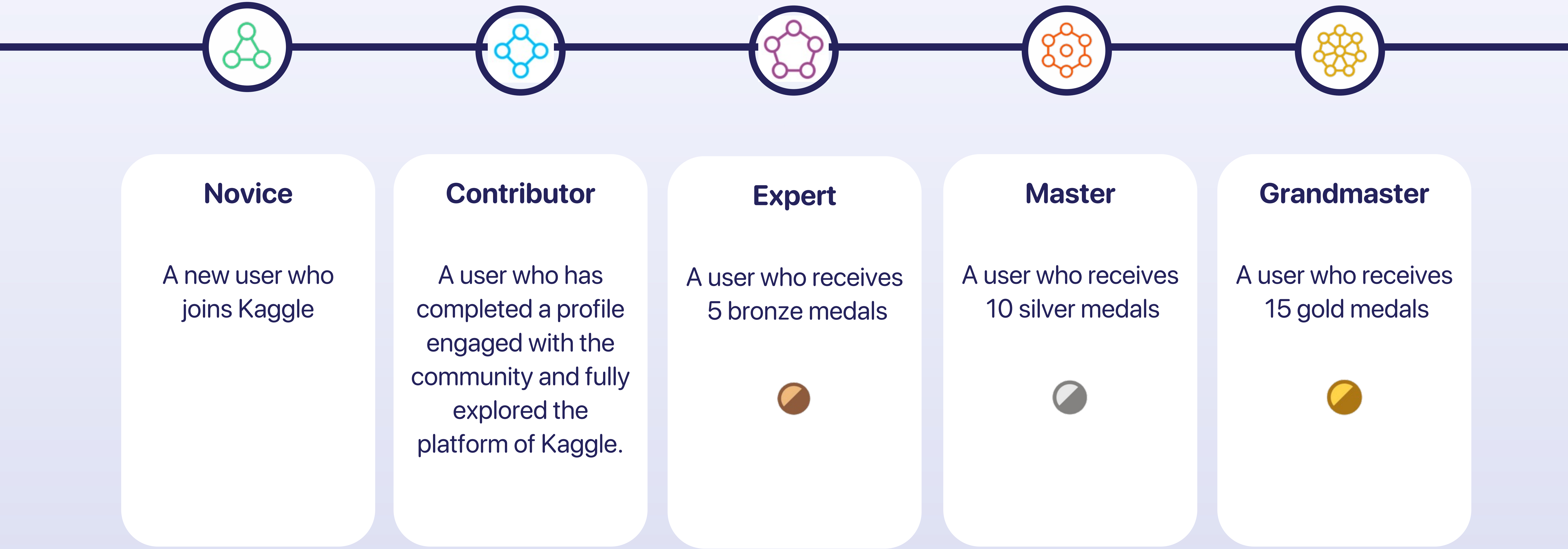
Improve your data science projects

Get started →

Next Steps

Kaggle Web Page

Kaggle's User Tier



Novice

A new user who joins Kaggle

Contributor

A user who has completed a profile engaged with the community and fully explored the platform of Kaggle.

Expert

A user who receives 5 bronze medals

Master

A user who receives 10 silver medals

Grandmaster

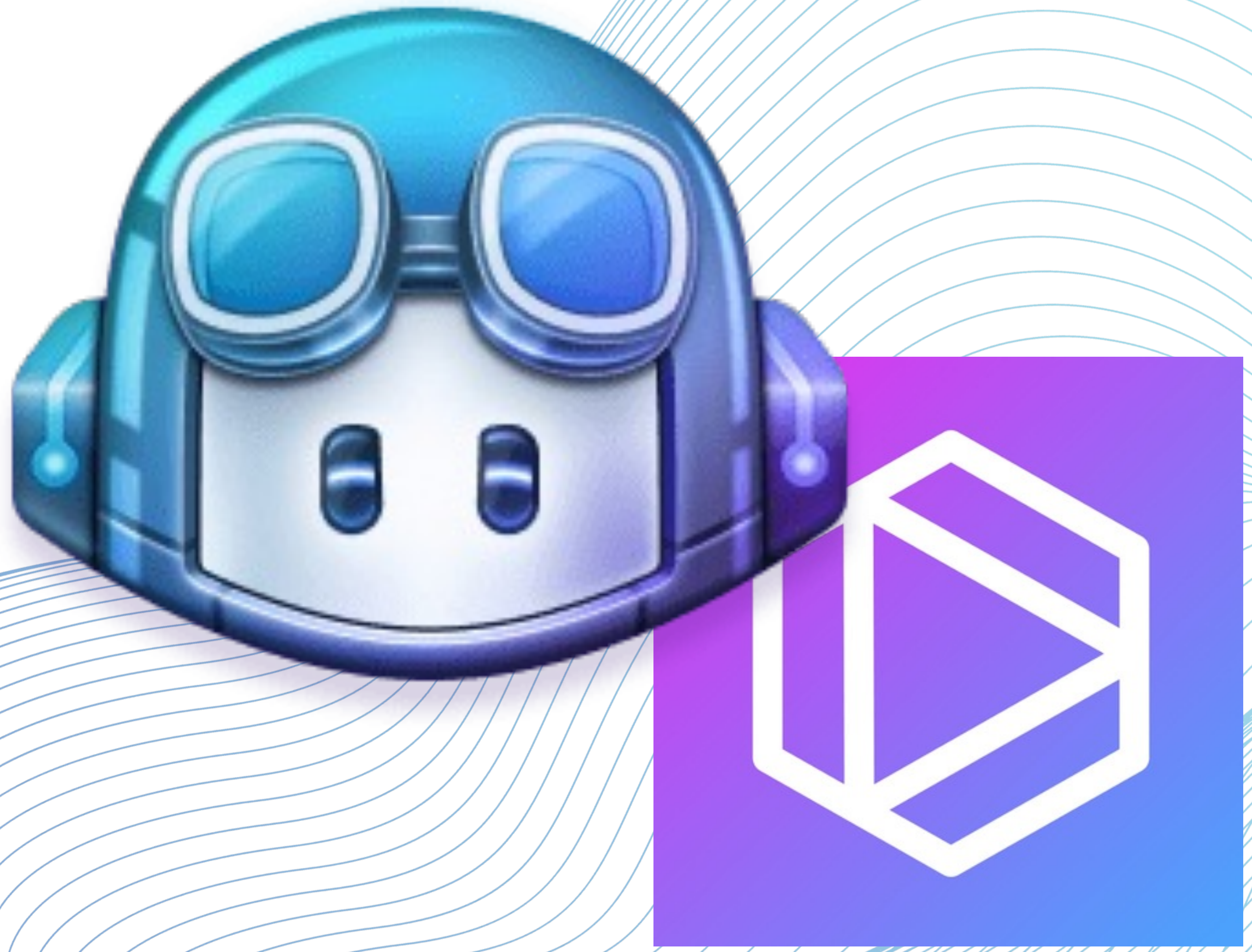
A user who receives 15 gold medals

CODE RECOMMENDATION

Code recommendation helped improve developer productivity significantly.

Such tools have been gaining a lot of attention.

Many new tools are AI-powered

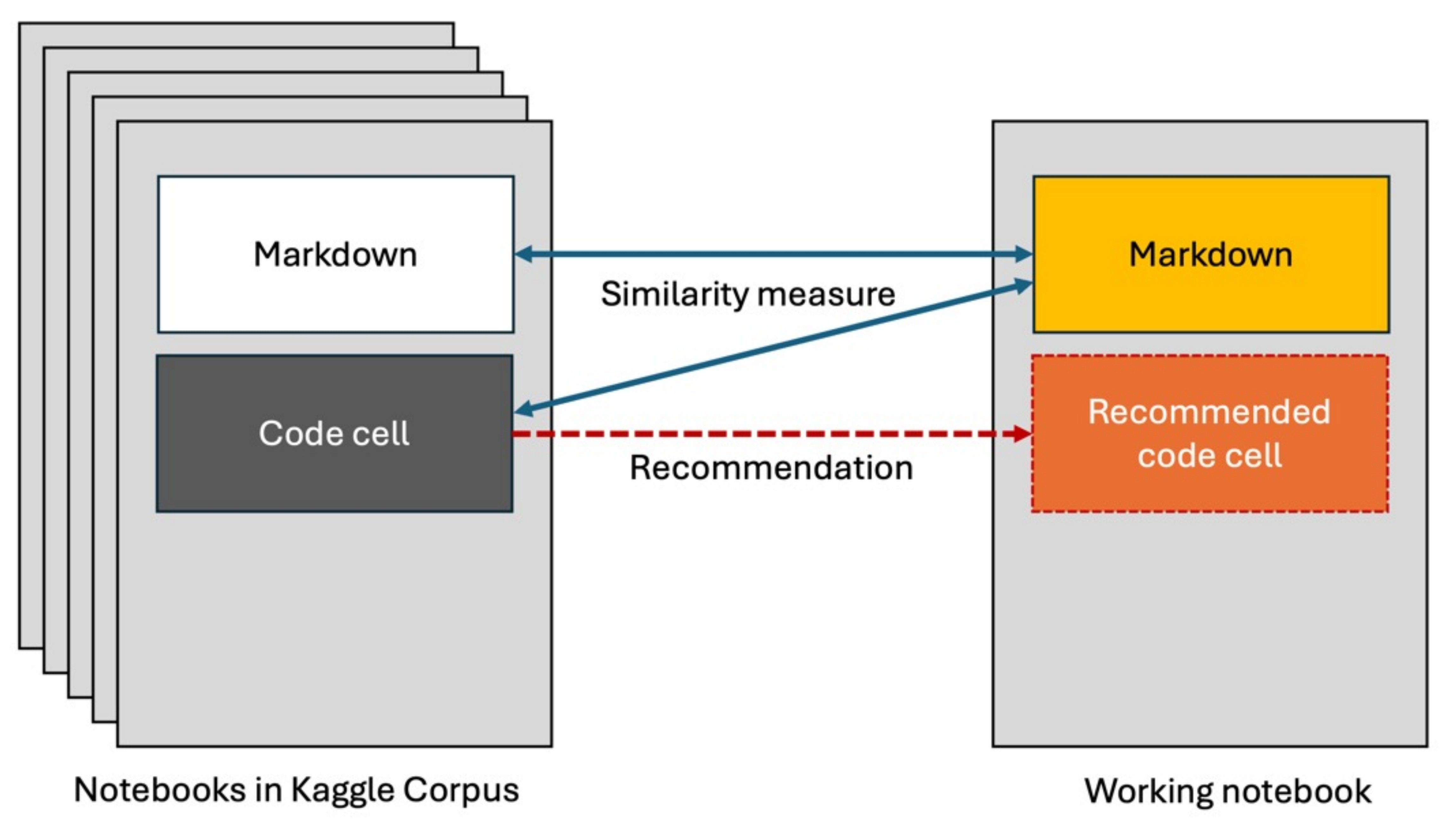


EXISTING CODE RECOMMENDATION TECHNIQUES

Name	Input	Approach
Example Overflow	Text	Similarity of keywords search on database using fine-tuned TF-IDF weight
Copilot	Text/Code	OpenAI Codex model trained on large open-source projects in GitHub
Tabnine	Text/Code	AI-based proprietary algorithm
Aroma	Code	Similarity distance from code using clustering and ranking code snippets
Strathcona	Code	Similarity from user's local structural detail and code structural detail in repositories
Senatus	Code	Similarity of query input and indexed code using Minhash-LSH technique

Can we recommend a code cell based on the given markdown cell by searching from existing Jupyter notebooks?

OUR APPROACH

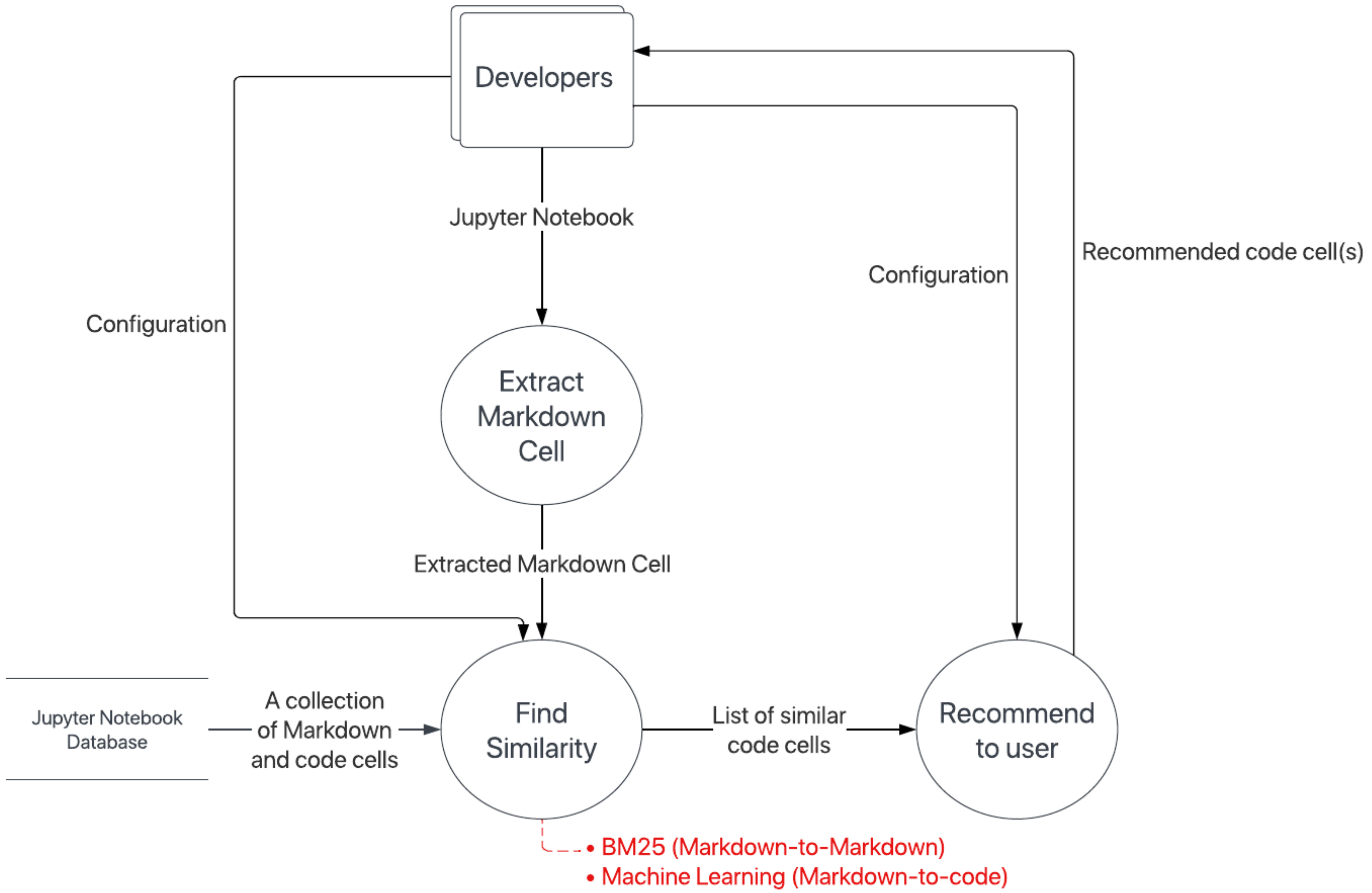


TYPHON

An approach for **recommending code cells** based on **existing code cells from Jupyter notebooks** in the Kaggle dataset.

Typhon suggests a code cell by, given a markdown cell, searching through the Jupyter Notebook corpus for the markdown cell with **the most similarity** and making recommendations

TYPHON ANALYSIS PROCESS


Developers
Jupyter Notebook
Configuration
Configuration
Recommended code cell(s)
Extract Markdown Cell
Extracted Markdown Cell
Jupyter Notebook Database
A collection of Markdown and code cells
Find Similarity
List of similar code cells
Recommend to user
• BM25 (Markdown-to-Markdown)
• Machine Learning (Markdown-to-code)

SIMILARITY TECHNIQUES

BM25

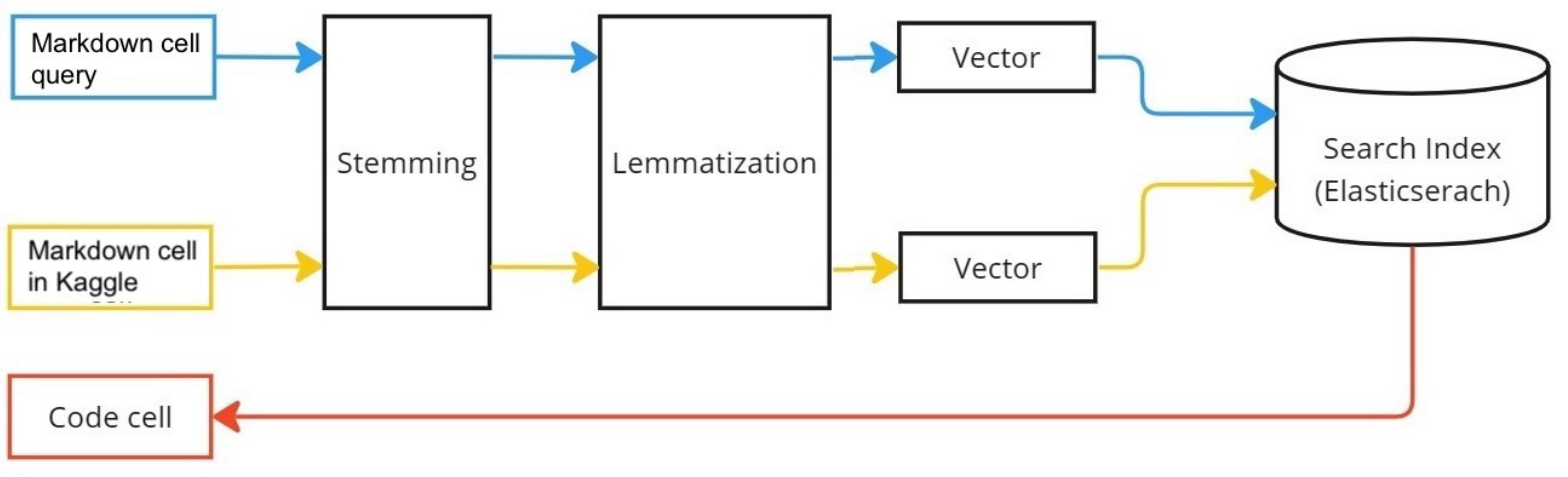
BM25 algorithm is a bag-of-words retrieval function, which ranks a set of documents regarding the query terms appearing in each document regardless of the proximity within the document.

BM25 has been widely used in search engines, such as Elasticsearch, as it is a robust and effective way to rank documents by relevance.

$$\sum_{i}^{n} IDF(q_i) \frac{f(q_i, D) \, * \, (k1+1)}{f(q_i, D) \, + \, k1 \, * \, (1 \, - \, b \, + \, b \, * \, \frac{fieldLen}{avgFieldLen})}$$

https://www.elastic.co/blog/practical-bm25-part-2-the-bm25-algorithm-and-its-variables

Typhon with BM25



Text < -- > Text ----→ Code

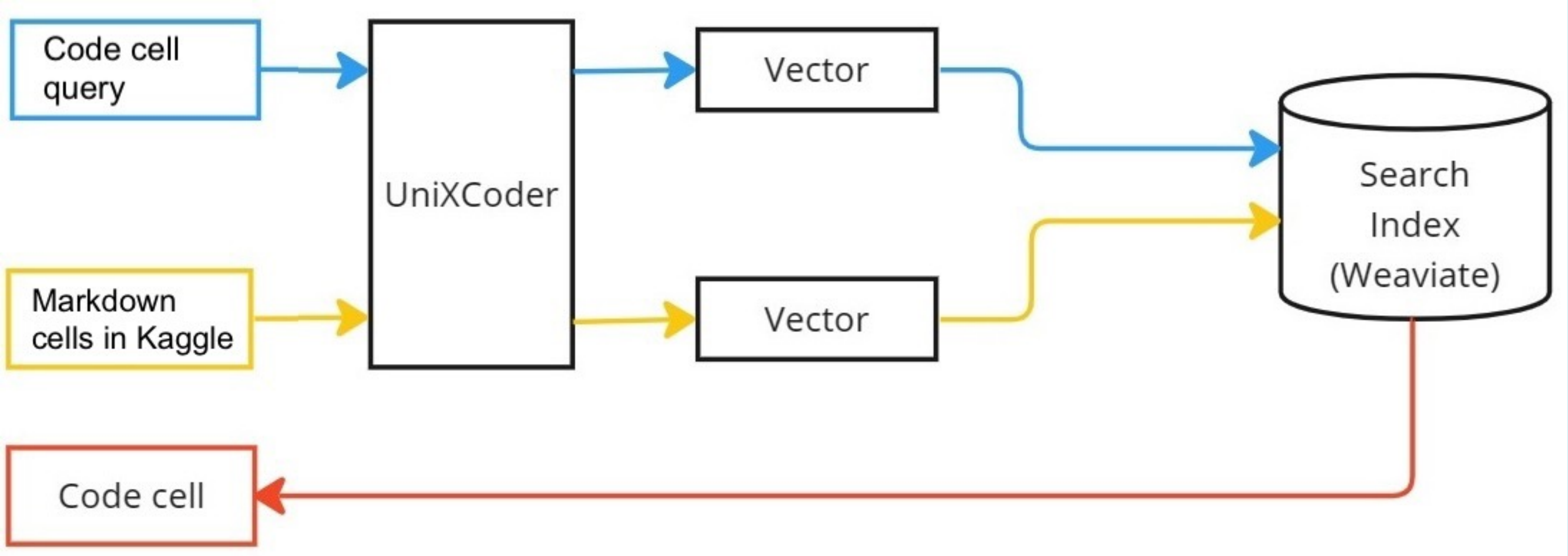
UniXcoder

A unified **cross-modal pre-trained model for the programming language**, which specializes in code understanding and code generation tasks.

It facilitates the usage of natural language and programming languages for code-related tasks.

Guo et al. (2022). UniXcoder: Unified Cross-Modal Pre-training for Code Representation. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1, 7212–7225.

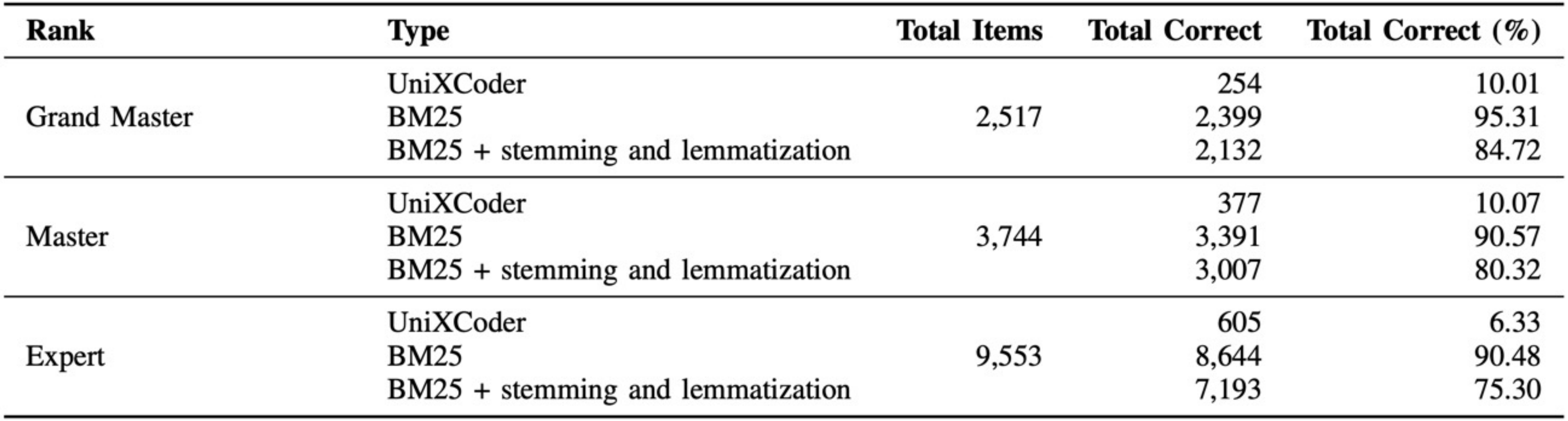
Typhon with UniXcoder



Code < -- > Text ----→ Code

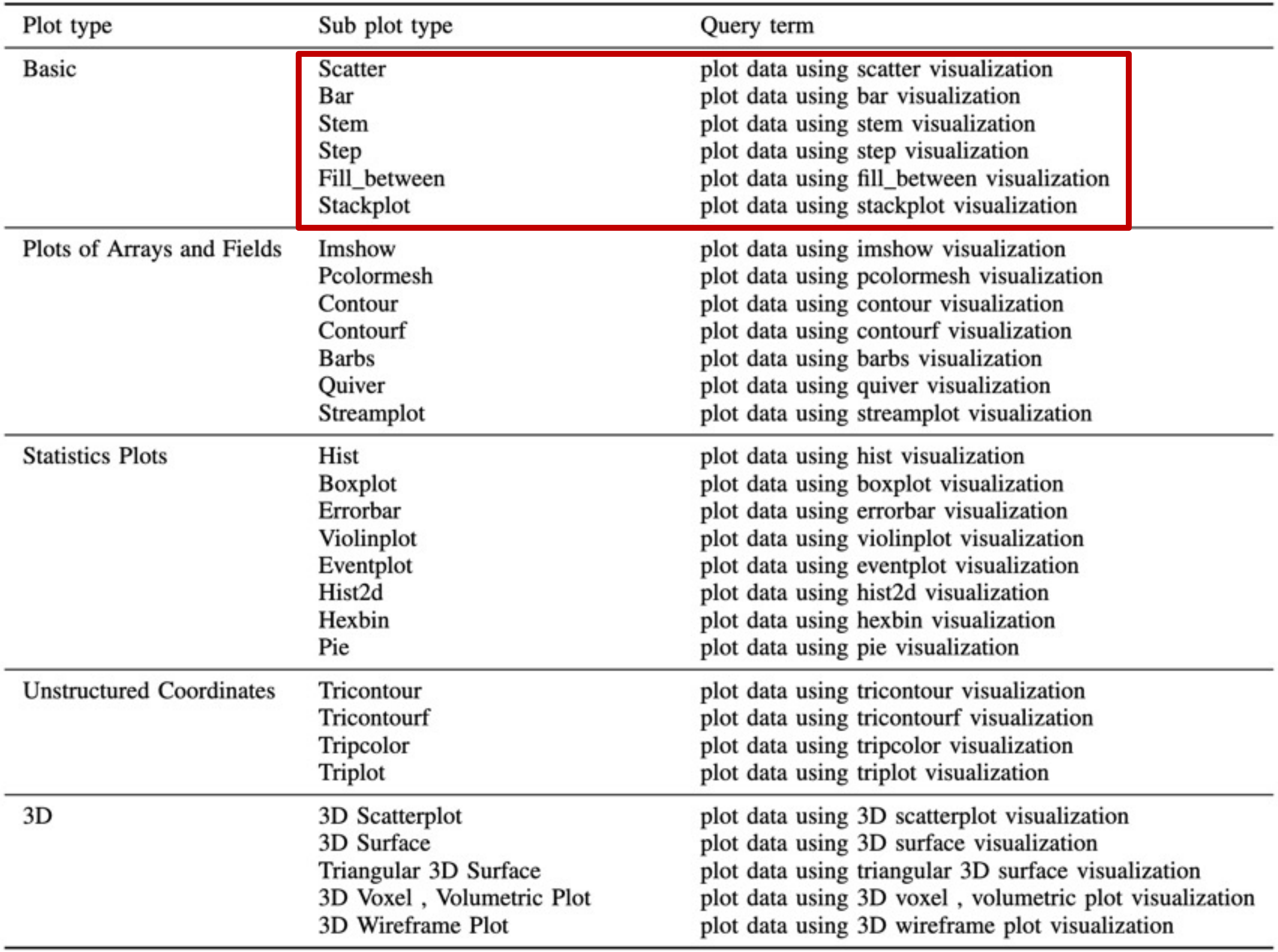
EVALUATION

SANITY CHECK



Rank	Type	Total Items	Total Correct	Total Correct (%)
Grand Master	UniXCoder	2,517	254	10.01
	BM25		2,399	95.31
	BM25 + stemming and lemmatization		2,132	84.72
Master	UniXCoder	3,744	377	10.07
	BM25		3,391	90.57
	BM25 + stemming and lemmatization		3,007	80.32
Expert	UniXCoder	9,553	605	6.33
	BM25		8,644	90.48
	BM25 + stemming and lemmatization		7,193	75.30

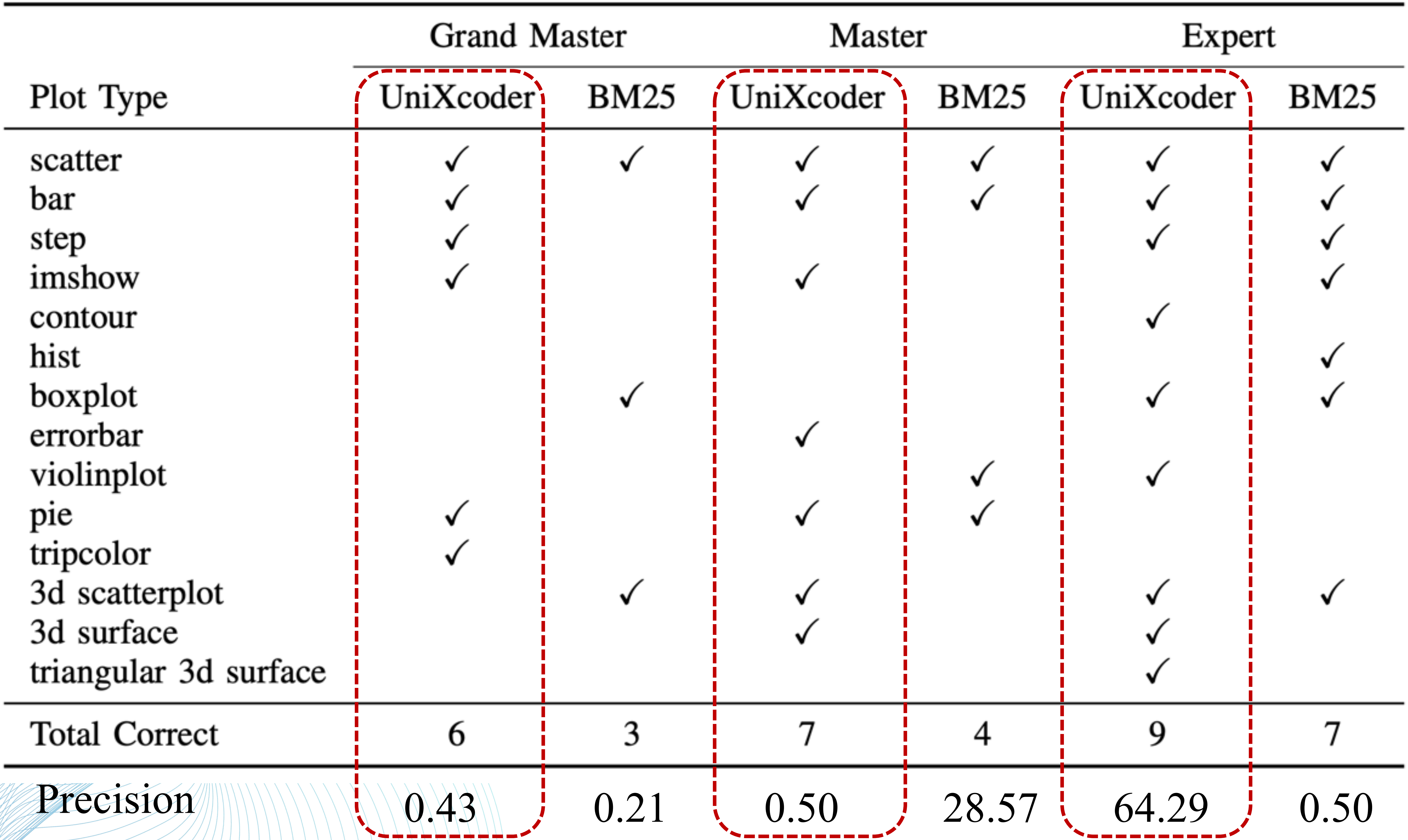
Matplotlib Visualization Code Cell Recommendation



Plot type	Sub plot type	Query term
Basic	Scatter	plot data using scatter visualization
	Bar	plot data using bar visualization
	Stem	plot data using stem visualization
	Step	plot data using step visualization
	Fill_between	plot data using fill_between visualization
	Stackplot	plot data using stackplot visualization
Plots of Arrays and Fields	Imshow	plot data using imshow visualization
	Pcolormesh	plot data using pcolormesh visualization
	Contour	plot data using contour visualization
	Contourf	plot data using contourf visualization
	Barbs	plot data using barbs visualization
	Quiver	plot data using quiver visualization
	Streamplot	plot data using streamplot visualization
Statistics Plots	Hist	plot data using hist visualization
	Boxplot	plot data using boxplot visualization
	Errorbar	plot data using errorbar visualization
	Violinplot	plot data using violinplot visualization
	Eventplot	plot data using eventplot visualization
	Hist2d	plot data using hist2d visualization
	Hexbin	plot data using hexbin visualization
	Pie	plot data using pie visualization
Unstructured Coordinates	Tricontour	plot data using tricontour visualization
	Tricontourf	plot data using tricontourf visualization
	Tripcolor	plot data using tripcolor visualization
	Triplot	plot data using triplot visualization
3D	3D Scatterplot	plot data using 3D scatterplot visualization
	3D Surface	plot data using 3D surface visualization
	Triangular 3D Surface	plot data using triangular 3D surface visualization
	3D Voxel , Volumetric Plot	plot data using 3D voxel , volumetric plot visualization
	3D Wireframe Plot	plot data using 3D wireframe plot visualization

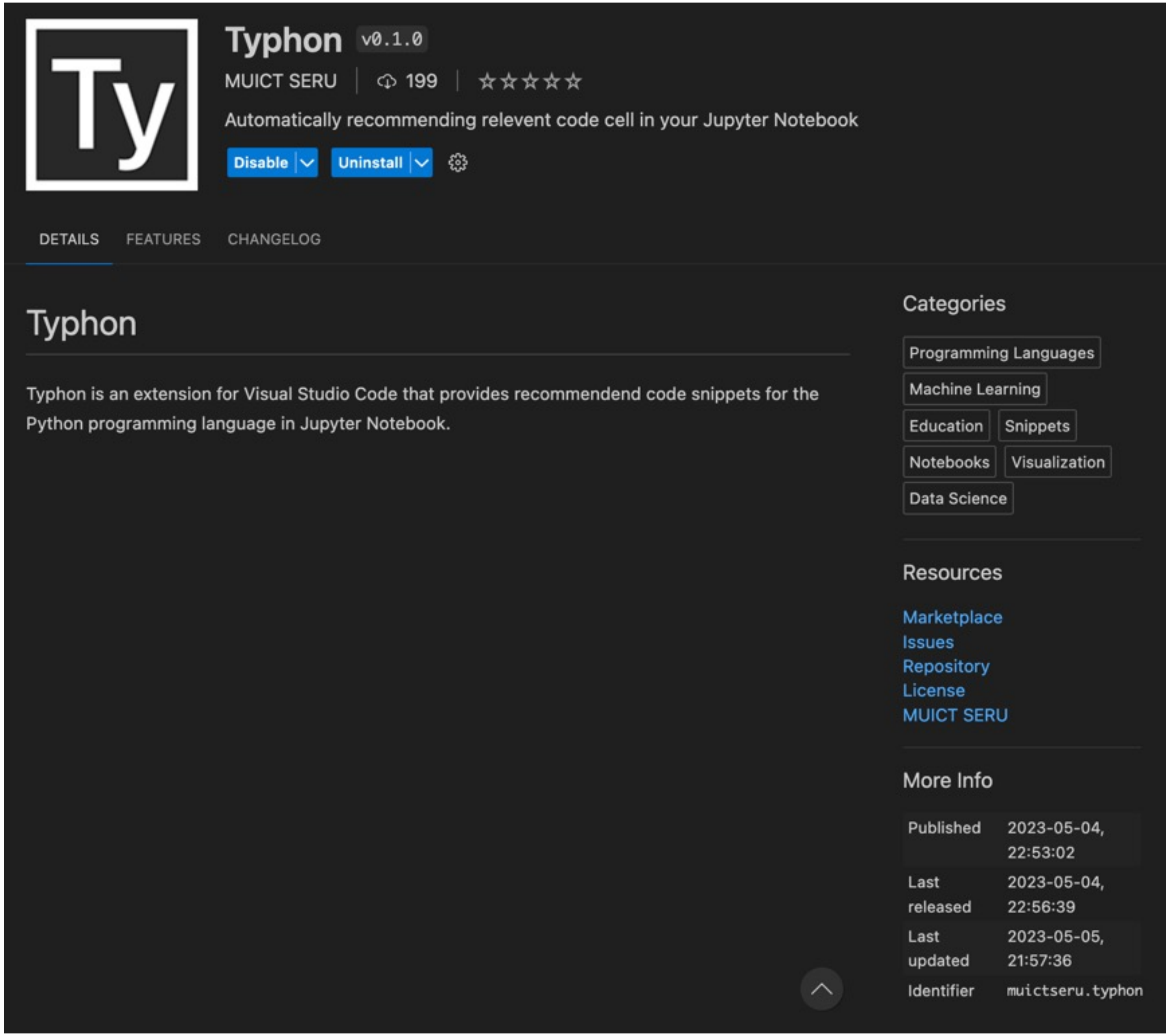
https://matplotlib.org/stable/plot_types/index

RESULTS



	Grand Master		Master		Expert	
Plot Type	UniXcoder	BM25	UniXcoder	BM25	UniXcoder	BM25
scatter	✓	✓	✓	✓	✓	✓
bar	✓		✓	✓	✓	✓
step	✓				✓	✓
imshow	✓		✓			✓
contour					✓	
hist						✓
boxplot		✓			✓	✓
errorbar			✓			
violinplot				✓	✓	
pie	✓		✓	✓		
tripcolor	✓					
3d scatterplot		✓	✓		✓	✓
3d surface			✓		✓	
triangular 3d surface					✓	
Total Correct	6	3	7	4	9	7
Precision	0.43	0.21	0.50	28.57	64.29	0.50

TYPHON VS CODE EXTENSION



Typhon v0.1.0

MUICT SERU | 199 | ☆☆☆☆☆

Automatically recommending relevent code cell in your Jupyter Notebook

Typhon

Typhon is an extension for Visual Studio Code that provides recommendend code snippets for the Python programming language in Jupyter Notebook.

Categories

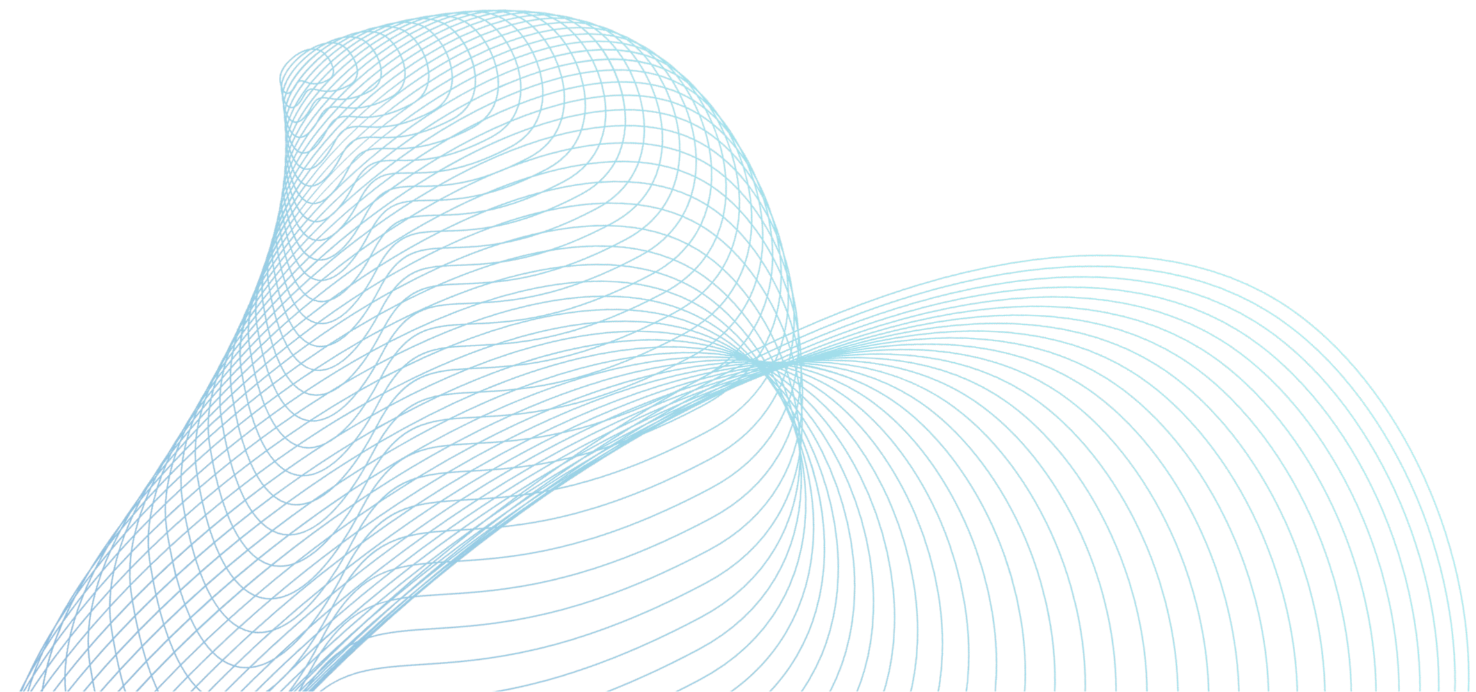
Programming Languages
Machine Learning
Education
Snippets
Notebooks
Visualization
Data Science

Resources

Marketplace
Issues
Repository
License
MUICT SERU

More Info

Published	2023-05-04, 22:53:02
Last released	2023-05-04, 22:56:39
Last updated	2023-05-05, 21:57:36
Identifier	muictseru.typhon



CONCLUSION Ty

We propose **Typhon**

- An approach for recommending code cells based on existing code cells from Jupyter notebooks in the Kaggle dataset.

- We investigate using BM25 and UniXcoder for code and text similarity measurement.

- We performed an evaluation based on matplotlib visualizations and found moderate accuracy in recommendations with UniXcoder outperforming BM25.

- Our Typhon VS code extension is available in the marketplace.