Q&A

2026 JULAC Libraries Forum

User-Centric Libraries: A Sustainable Future through INNOVATION, TECHNOLOGY, and COLLABORATION

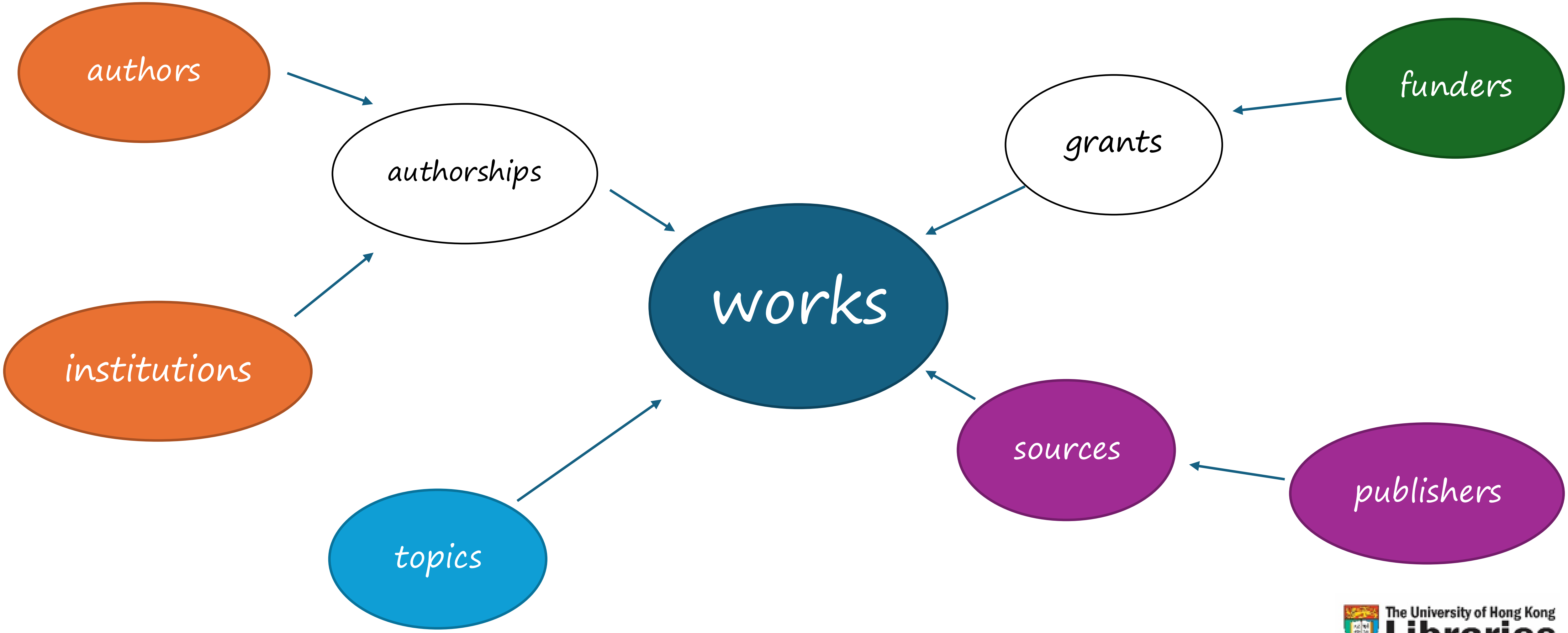
Leveraging Data Analytics for Strategic Library Decision-Making

Mahboobani Vanessa Ramesh *(Data Services & Information Research Librarian, HKUL)*
Katie Kwong *(IT Manager, HKUL)*

The University of Hong Kong Libraries

Trend analysis of research topics

OpenAlex Schema



Q&A

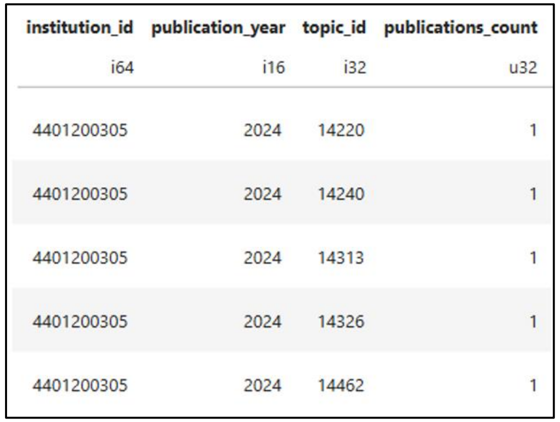
Data Processing Workflow

Extraction → Transform & Load → Insight

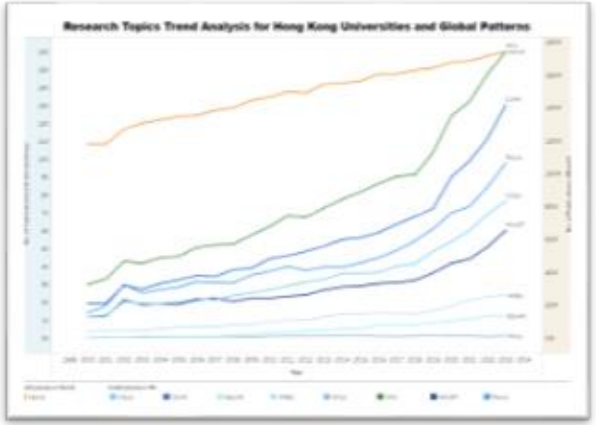
OpenAlex Snapshot from Amazon S3

OpenAlex dataset
works
institutions
topics

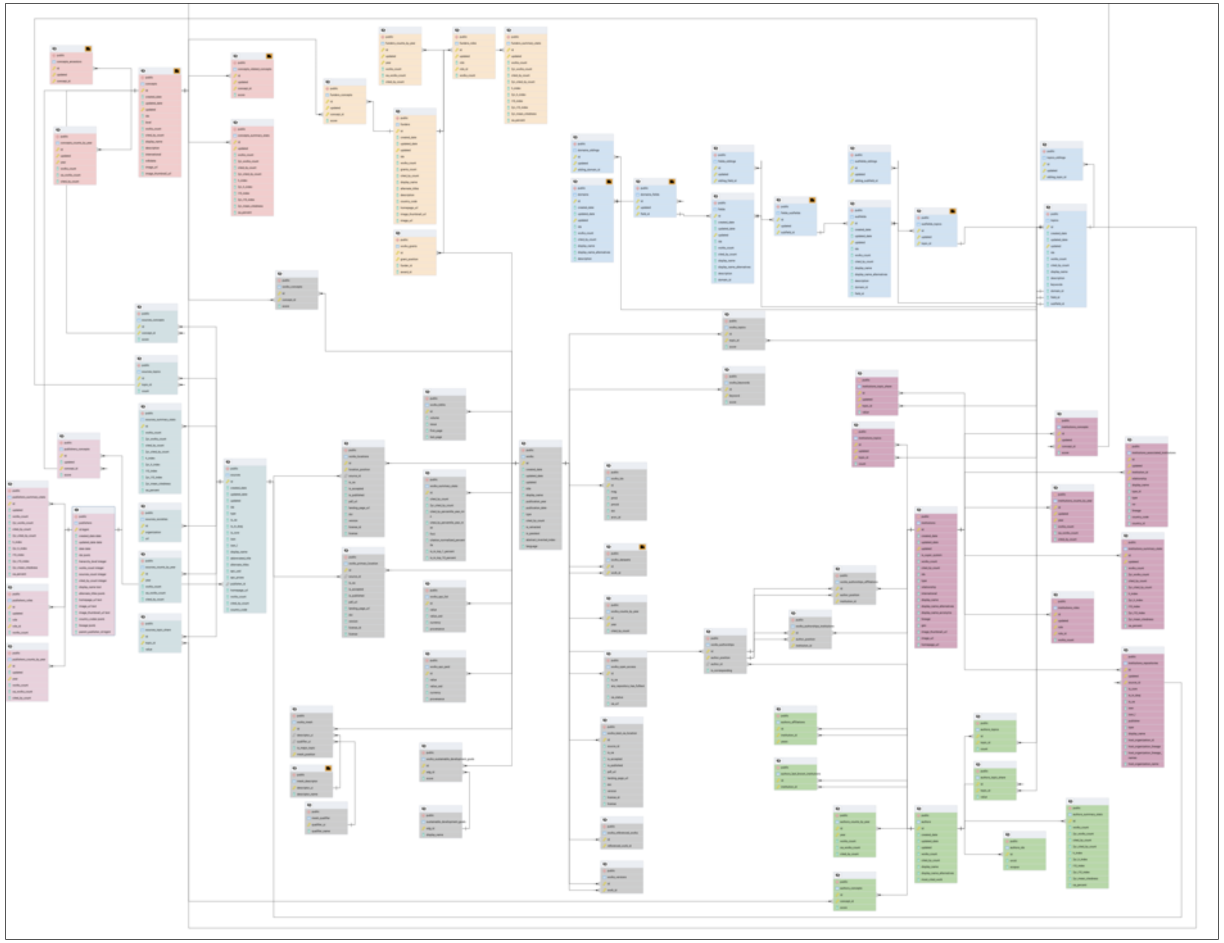
.
.
.



institution_id i64	publication_year i16	topic_id i32	publications_count u32
4401200305	2024	14220	1
4401200305	2024	14240	1
4401200305	2024	14313	1
4401200305	2024	14326	1
4401200305	2024	14462	1



Extract



Works
463 million

Authors
115 million

Institutions
102 thousand

Topics
4.5 thousand

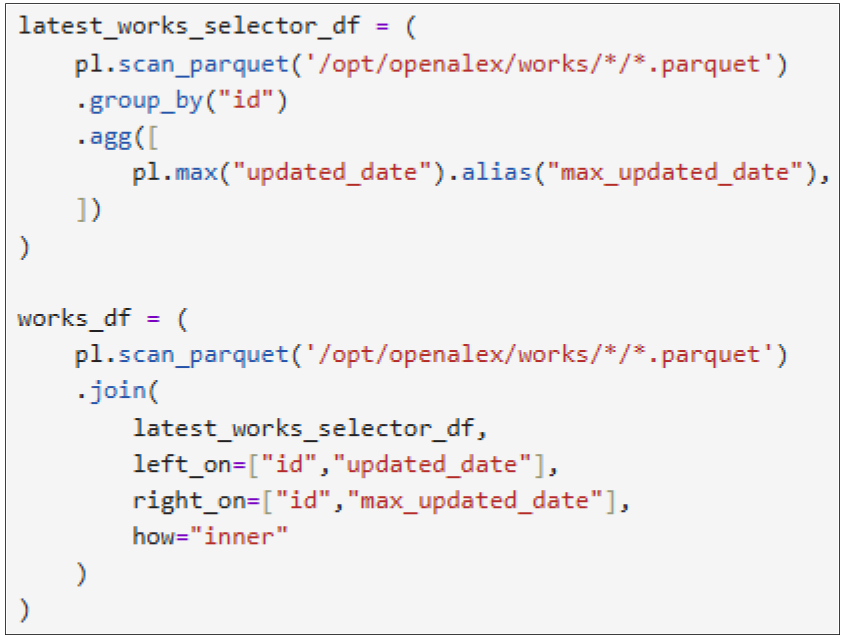
Subfields
252

Fields
26

Domain
4

Transform - Data Deduplication

- OpenAlex files contain historical snapshots
- Identifies the most recent record for every works, institutions, and topics
- Ensuring updated_date types match across different sources

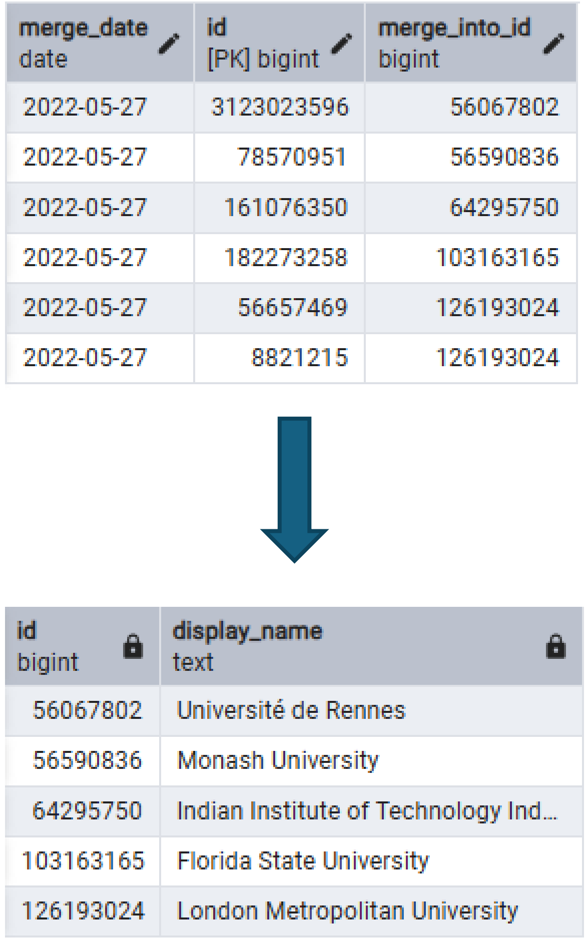


```
latest_works_selector_df = (
    pl.scan_parquet('/opt/openalex/works/*/*.parquet')
    .group_by("id")
    .agg([
        pl.max("updated_date").alias("max_updated_date"),
    ])
)

works_df = (
    pl.scan_parquet('/opt/openalex/works/*/*.parquet')
    .join(
        latest_works_selector_df,
        left_on=["id","updated_date"],
        right_on=["id","max_updated_date"],
        how="inner"
    )
)
```

Transform – Entity Resolution & Merging

- Data often merge, data might still be tagged with the old ID

- Create list of "all possible ids" to connect historical data with their current counterparts, ensuring that all records are linked to the correct current entity



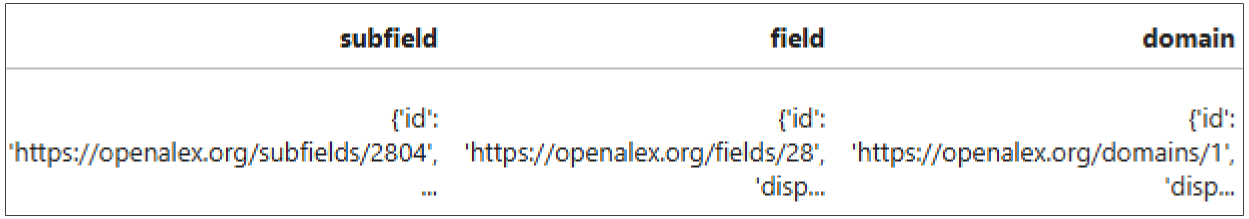
merge_date date	id [PK] bigint	merge_into_id bigint
2022-05-27	3123023596	56067802
2022-05-27	78570951	56590836
2022-05-27	161076350	64295750
2022-05-27	182273258	103163165
2022-05-27	56657469	126193024
2022-05-27	8821215	126193024

id bigint	display_name text
56067802	Université de Rennes
56590836	Monash University
64295750	Indian Institute of Technology Ind...
103163165	Florida State University
126193024	London Metropolitan University

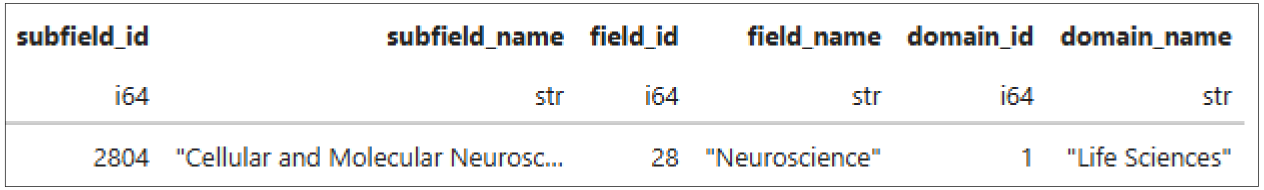
Q&A

Transform - Data Normalization & Standardization

- Some data is nested, "flattens" the data
- Ensure data types of corresponding columns are consistent across different tables to enhances interoperability between datasets



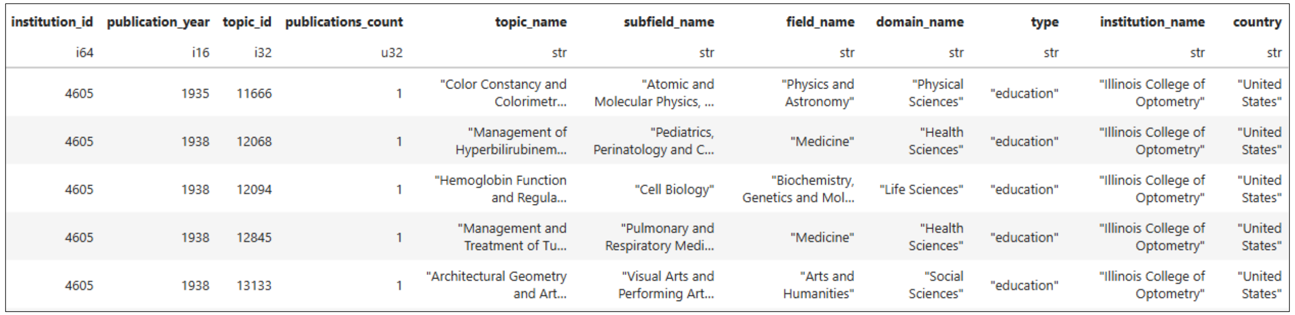
subfield	field	domain
{'id': 'https://openalex.org/subfields/2804', ...	{'id': 'https://openalex.org/fields/28', 'disp...	{'id': 'https://openalex.org/domains/1', 'disp...



subfield_id	subfield_name	field_id	field_name	domain_id	domain_name
i64	str	i64	str	i64	str
2804	"Cellular and Molecular Neurosc...	28	"Neuroscience"	1	"Life Sciences"

Transform – Aggregation

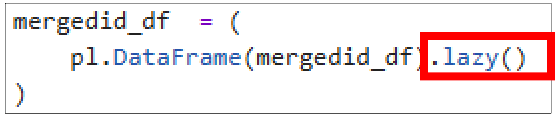
Final join and aggregate all data to produce a clean, multi-dimensional dataset



institution_id	publication_year	topic_id	publications_count	topic_name	subfield_name	field_name	domain_name	type	institution_name	country
i64	i16	i32	u32	str	str	str	str	str	str	str
4605	1935	11666	1	"Color Constancy and Colorimetr…	"Atomic and Molecular Physics, …	"Physics and Astronomy"	"Physical Sciences"	"education"	"Illinois College of Optometry"	"United States"
4605	1938	12068	1	"Management of Hyperbilirubinem…	"Pediatrics, Perinatology and C…	"Medicine"	"Health Sciences"	"education"	"Illinois College of Optometry"	"United States"
4605	1938	12094	1	"Hemoglobin Function and Regula…	"Cell Biology"	"Biochemistry, Genetics and Mol…	"Life Sciences"	"education"	"Illinois College of Optometry"	"United States"
4605	1938	12845	1	"Management and Treatment of Tu…	"Pulmonary and Respiratory Medi…	"Medicine"	"Health Sciences"	"education"	"Illinois College of Optometry"	"United States"
4605	1938	13133	1	"Architectural Geometry and Art…	"Visual Arts and Performing Art…	"Arts and Humanities"	"Social Sciences"	"education"	"Illinois College of Optometry"	"United States"

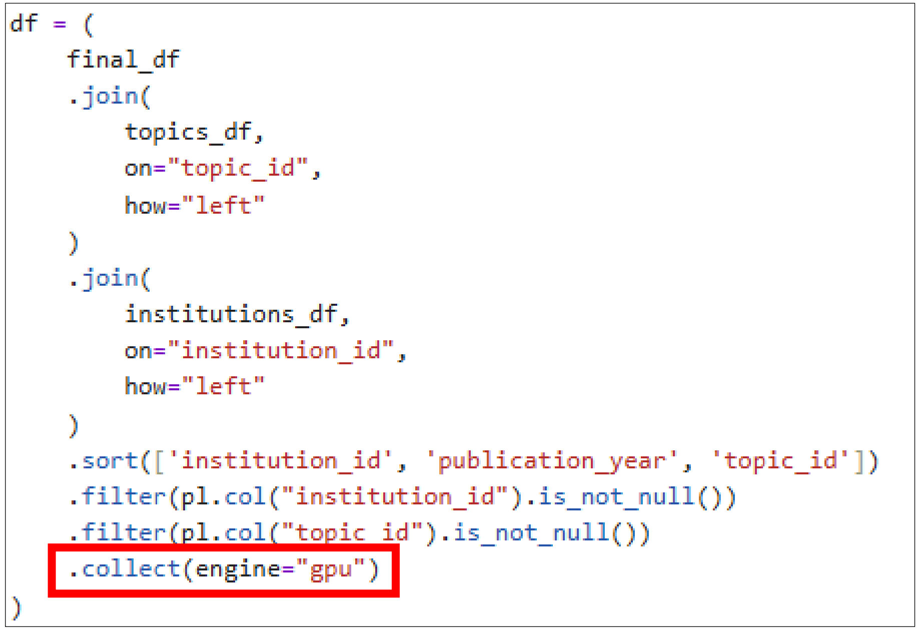
Overcoming "Out of Memory" errors

- Leverage data with Polars' Lazy API (via *.lazy()*)



```
mergedid_df  = (
    pl.DataFrame(mergedid_df).lazy()
)
```

- Collect the data with *engine ="gpu"* for the final materialization



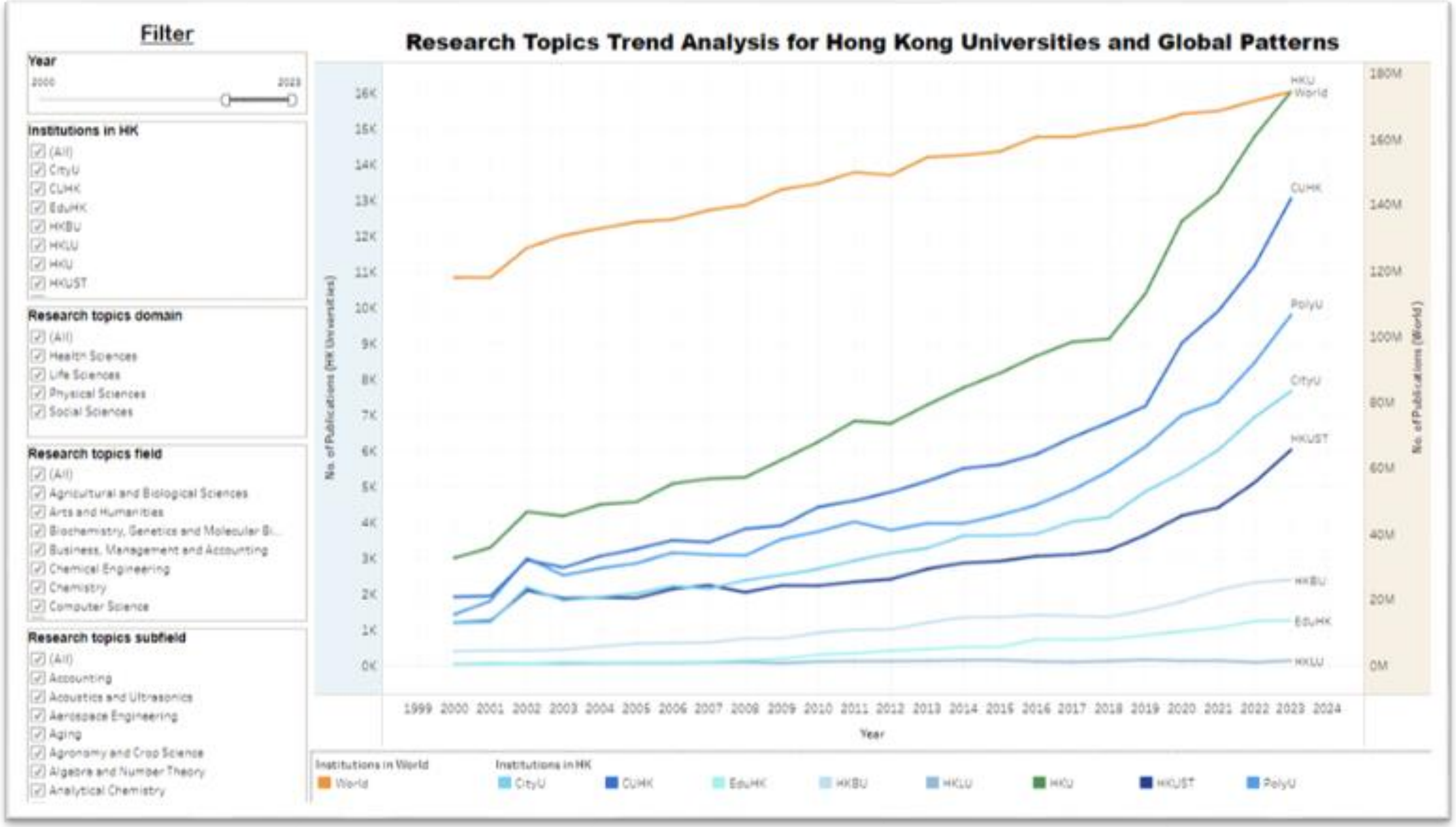
```
df = (
    final_df
    .join(
        topics_df,
        on="topic_id",
        how="left"
    )
    .join(
        institutions_df,
        on="institution_id",
        how="left"
    )
    .sort(['institution_id', 'publication_year', 'topic_id'])
    .filter(pl.col("institution_id").is_not_null())
    .filter(pl.col("topic_id").is_not_null())
    .collect(engine="gpu")
)
```

- Physical GPU hardware: NVIDIA GeForce RTX 4080 x 2

Q&A

Research Topics Analysis – Regional Level

Examine trends in research contributions from UGC Universities at HK over the years



Filter

Year
2000 – 2023

Institutions in HK
- ☑ (All)
- ☑ CityU
- ☑ CUHK
- ☑ EdUHK
- ☑ HKBU
- ☑ HKLU
- ☑ HKU
- ☑ HKUST

Research topics domain
- ☑ (All)
- ☑ Health Sciences
- ☑ Life Sciences
- ☑ Physical Sciences
- ☑ Social Sciences

Research topics field
- ☑ (All)
- ☑ Agricultural and Biological Sciences
- ☑ Arts and Humanities
- ☑ Biochemistry, Genetics and Molecular Bi...
- ☑ Business, Management and Accounting
- ☑ Chemical Engineering
- ☑ Chemistry
- ☑ Computer Science

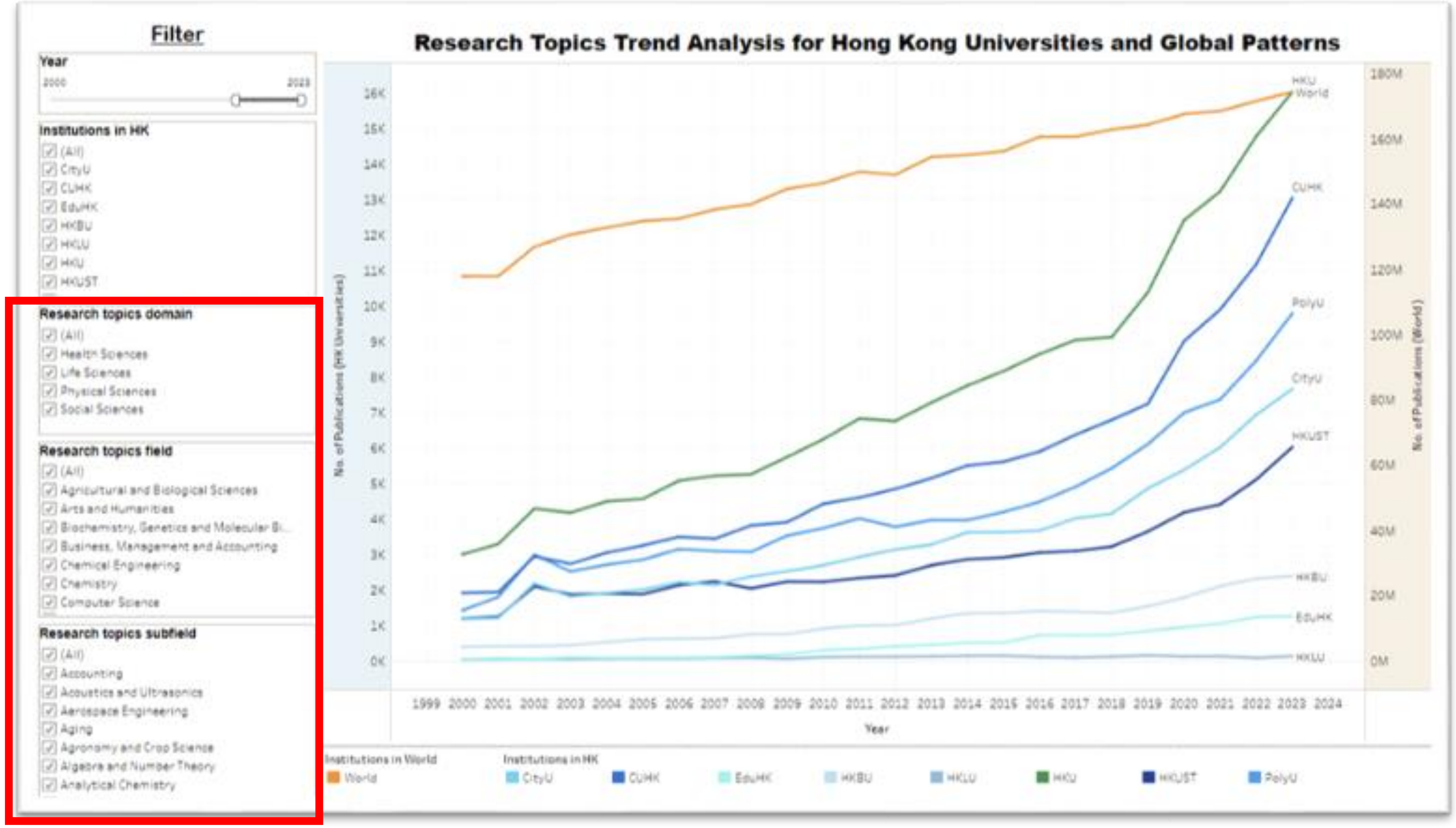
Research topics subfield
- ☑ (All)
- ☑ Accounting
- ☑ Acoustics and Ultrasonics
- ☑ Aerospace Engineering
- ☑ Aging
- ☑ Agronomy and Crop Science
- ☑ Algebra and Number Theory
- ☑ Analytical Chemistry

Research Topics Trend Analysis for Hong Kong Universities and Global Patterns

Research Topics Analysis – Regional Level

- Identify the key research topic, interest across various UGC universities in HK

- Analyze the development and evolution of research topics to understand their changing patterns across different institutions



Filter

Year
2000 2023

Institutions in HK
☑ (All)
☑ CityU
☑ CUHK
☑ EduHK
☑ HKBU
☑ HKLU
☑ HKU
☑ HKUST

Research topics domain
☑ (All)
☑ Health Sciences
☑ Life Sciences
☑ Physical Sciences
☑ Social Sciences

Research topics field
☑ (All)
☑ Agricultural and Biological Sciences
☑ Arts and Humanities
☑ Biochemistry, Genetics and Molecular Bi...
☑ Business, Management and Accounting
☑ Chemical Engineering
☑ Chemistry
☑ Computer Science

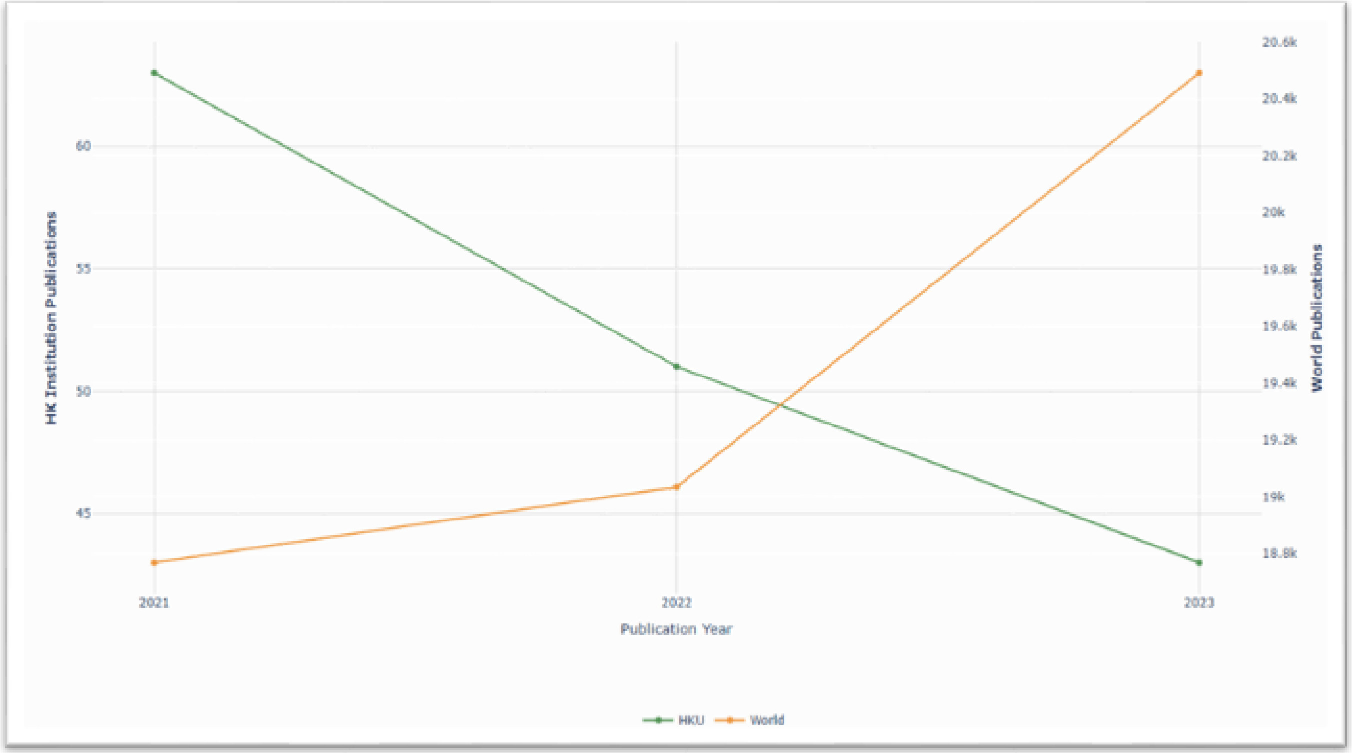
Research topics subfield
☑ (All)
☑ Accounting
☑ Acoustics and Ultrasonics
☑ Aerospace Engineering
☑ Aging
☑ Agronomy and Crop Science
☑ Algebra and Number Theory
☑ Analytical Chemistry

Research Topics Trend Analysis for Hong Kong Universities and Global Patterns

Research Topics Analysis – International Level

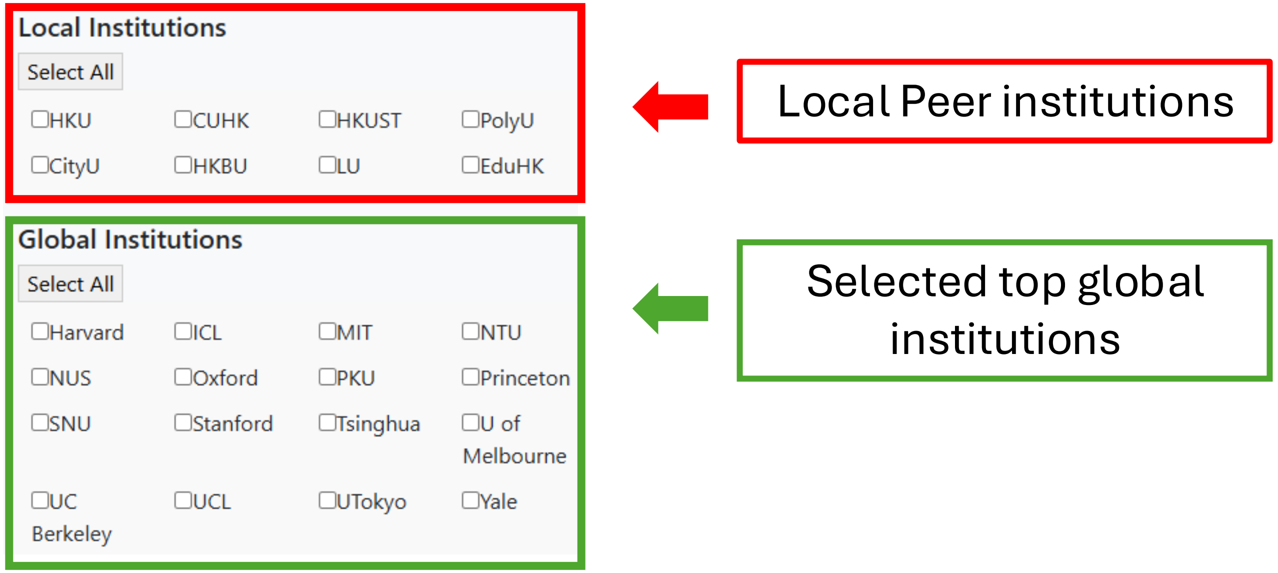
- Identify emerging research interests to align strategic initiatives with global academic trends

- Gaps: specific research topics that have continued growth in the number of publications for global institutions while showing a continued decrease for HKU



Research Topics Analysis – International Level

Assesses HKU's research against local peers and selected global top institutions to identify strengths, areas for improvement, and opportunities for collaboration in key research areas



Local Institutions

Select All

□HKU □CUHK □HKUST □PolyU
□CityU □HKBU □LU □EduHK

Global Institutions

Select All

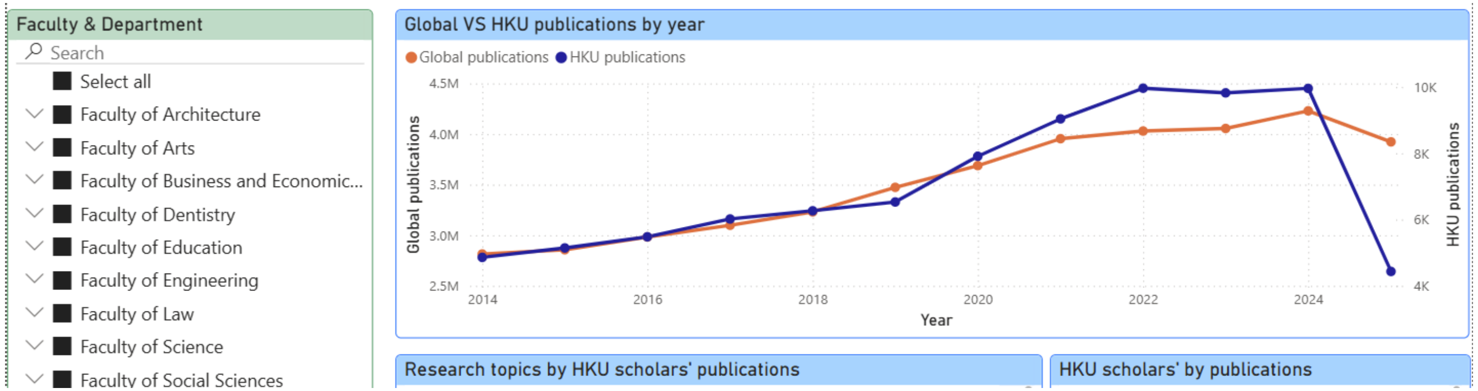
□Harvard □ICL □MIT □NTU
□NUS □Oxford □PKU □Princeton
□SNU □Stanford □Tsinghua □U of Melbourne
□UC Berkeley □UCL □UTokyo □Yale

Local Peer institutions

Selected top global institutions

Research Topics Analysis – Departmental Level

- Assess the performance of different internal units
- Identify faculty, department and scholars involved in top emerging research topic



Faculty & Department

Search

■ Select all

■ Faculty of Architecture

■ Faculty of Arts

■ Faculty of Business and Economic...

■ Faculty of Dentistry

■ Faculty of Education

■ Faculty of Engineering

■ Faculty of Law

■ Faculty of Science

■ Faculty of Social Sciences

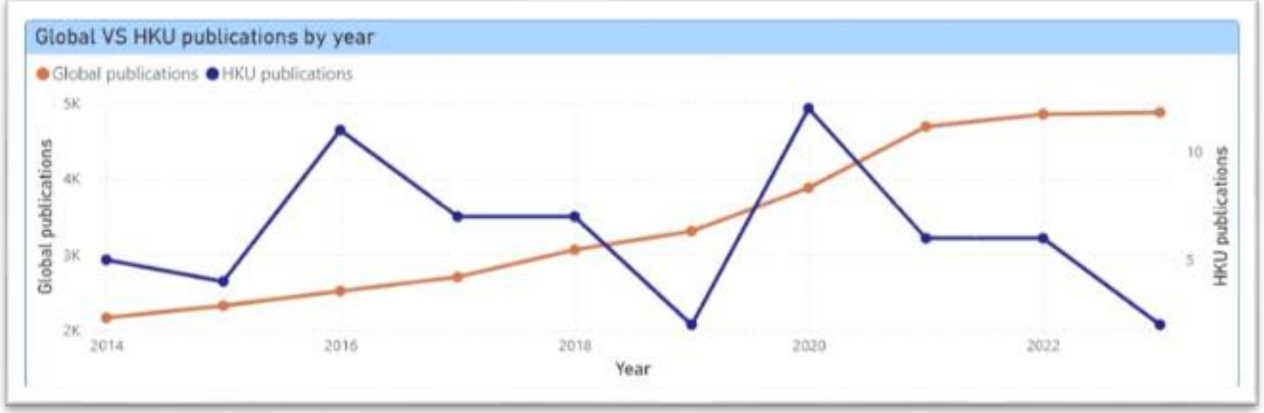
Global VS HKU publications by year

Research topics by HKU scholars' publications

HKU scholars' by publications

Research Topics Analysis – Departmental Level

- Identify gaps where faculty / department lag behind global trend which help management in making data-centric decisions on future research directions, funding, and talent recruitment



Enhance Research Support and Collection Development

- Precision Recommendation
  - Analyze impact within emerging and niche fields
  - Recommend the latest scholarly resources to the research community to maximize their visibility and impact
- Anticipating Collection Needs
  - Track trending research topics, evidence-based collection investment decisions to ensure effective resource allocation
- Strategic Alignment
  - Collections are current and comprehensive, aligning with institutional research priorities
  - Deeper understanding of contemporary research landscapes, ultimately enhancing resource allocation and strategic planning

Integration of linear programming in evaluating database subscriptions

Evaluating database subscriptions

- Involves assessments using a variety of criteria
  - Cost and Budget Considerations
    - **Cost-Effectiveness**
      - Analyze potential cost-per-use, particularly when considering renewals
    - **Usage Statistics**
      - Utilize standardized metrics (like COUNTER-compliant data) to track the number of sessions, full-text downloads, and refused access events to understand actual use
      - EZProxy logs capture user activity and access to library resources. They offer invaluable insights into user behavior and resource utilization.

Linear Programming (LP) optimization

- Linear Programming (LP) is a powerful business analytics tool that uses math to find the best outcome (such as, maximize profit, minimize cost) from choices, given limited resources (time, money, materials)
- Solve complex optimization problems for business cases
- Stock market employs linear programming (LP) to create a balanced investment portfolio that provides both protection and opportunities
  - Maximize expected profits while minimizing associated risks.

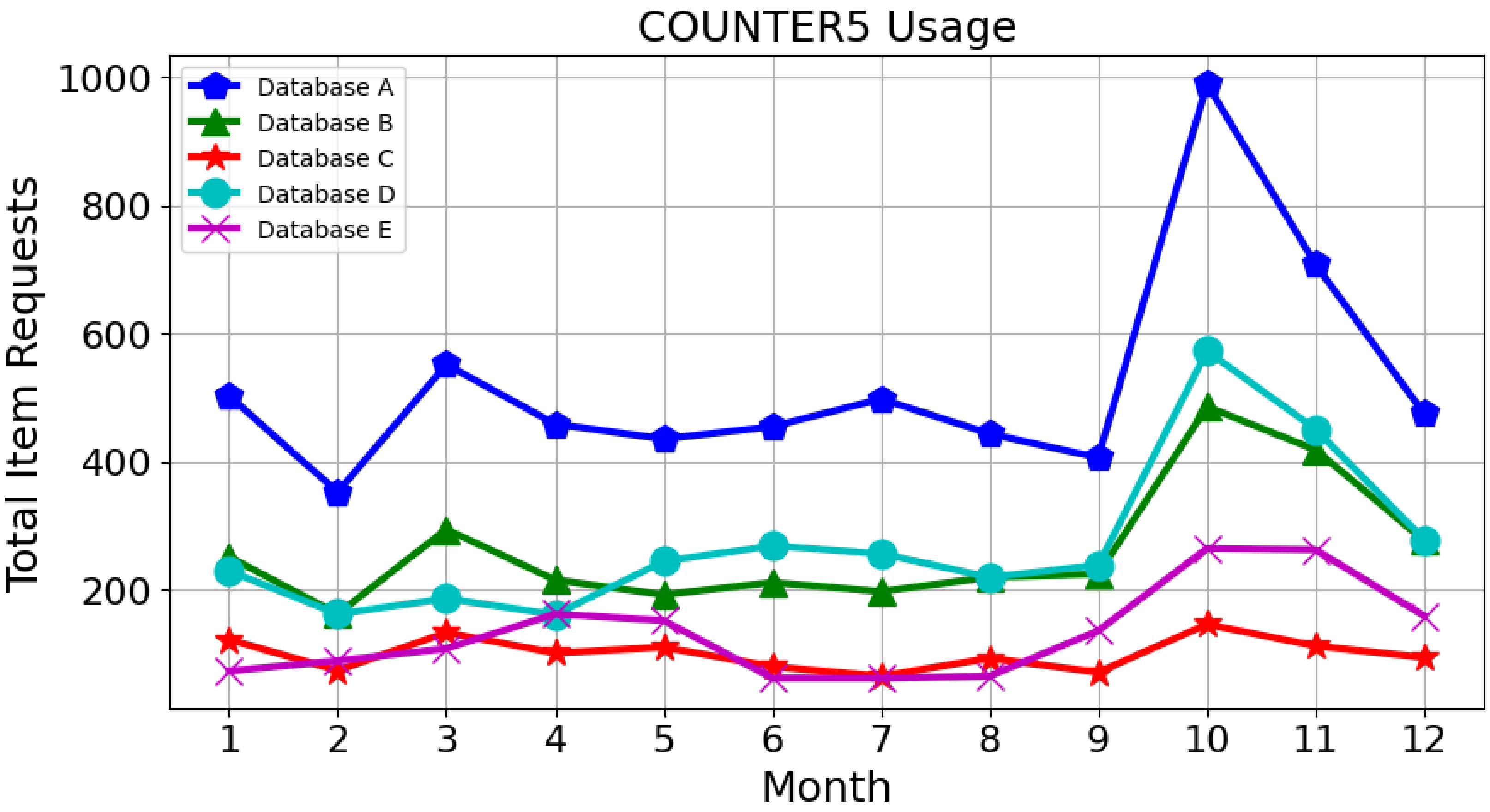
Problem Scenario

Peter aims to conduct a usage analysis on a database package comprising five databases. His goal is to develop a strategic framework that optimizes the cost per usage to enhance performance.

To achieve this, Peter plans to apply economic planning principles by employing linear programming to maximize expected profits while minimizing associated risks.

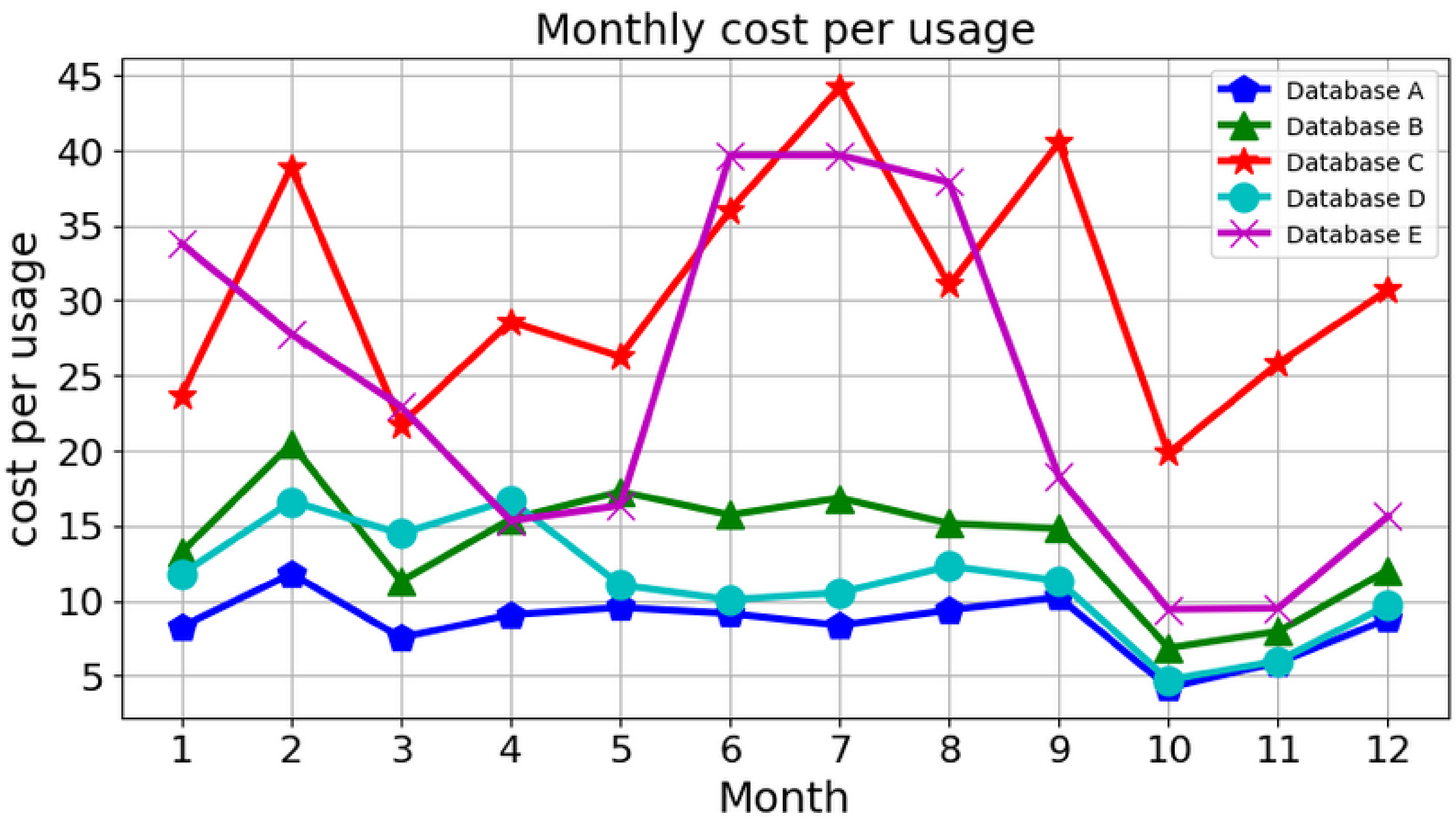
Gather and Analyze Data

- Usage Data
  - COUNTER5- TR_J1: Journal Requests (Excluding OA_Gold)



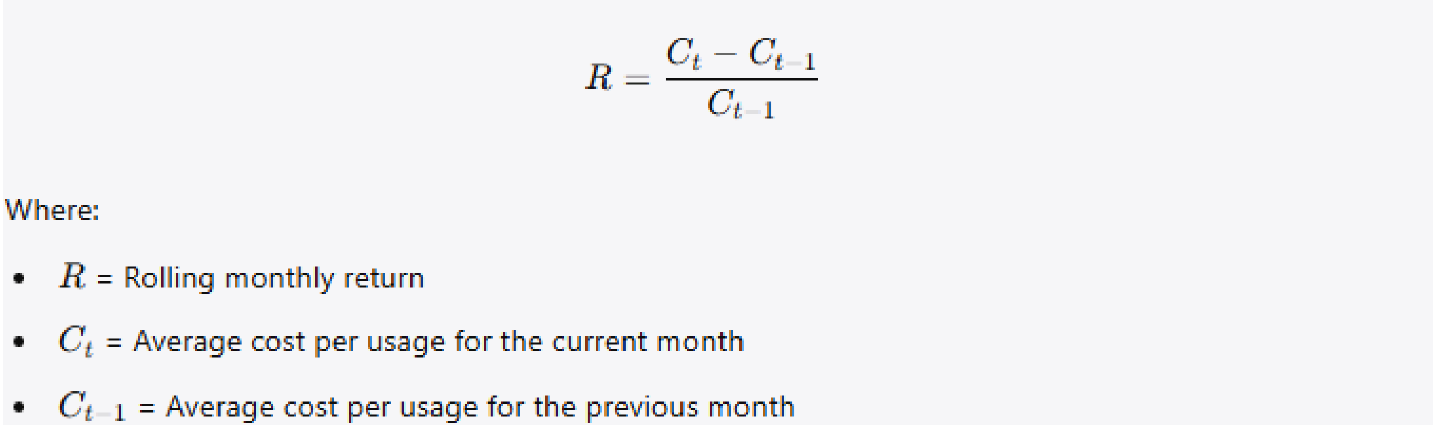
Gather and Analyze Data

$$\text{Cost per Usage} = \frac{\text{Total Monthly Costs}}{\text{Total Monthly Usage}}$$



Gather and Analyze Data

- Calculate the rolling monthly return

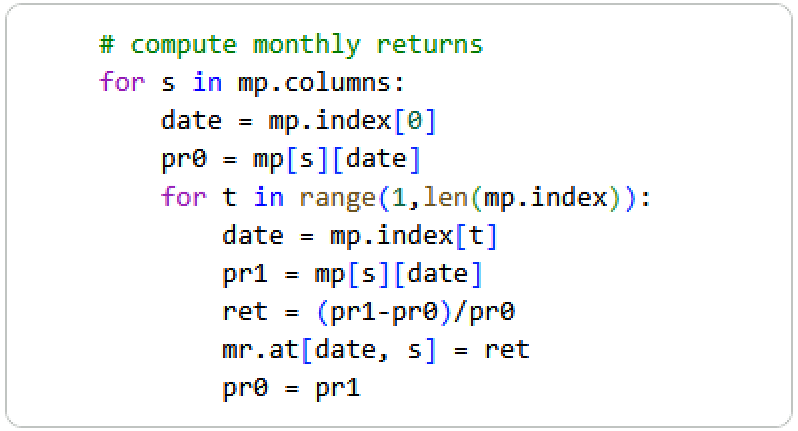


$$R = \frac{C_t - C_{t-1}}{C_{t-1}}$$

Where:

- R = Rolling monthly return
- C_t = Average cost per usage for the current month
- C_{t-1} = Average cost per usage for the previous month

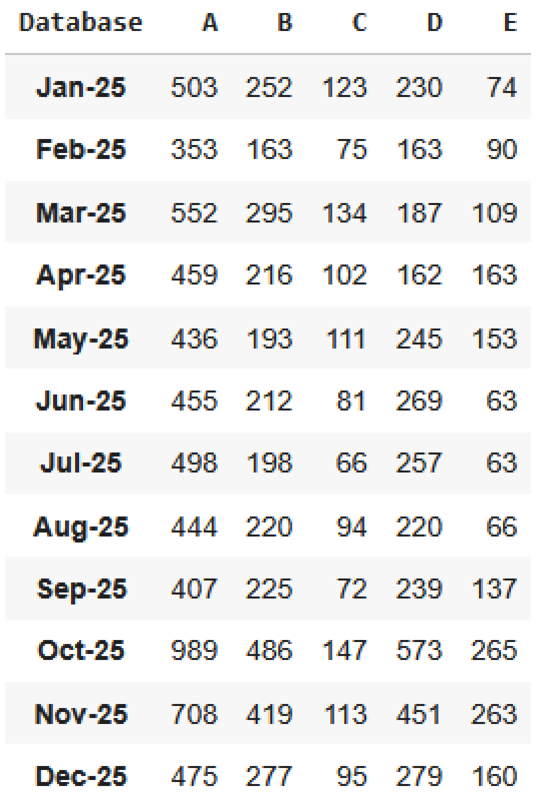
Compute monthly returns



```
# compute monthly returns
for s in mp.columns:
    date = mp.index[0]
    pr0 = mp[s][date]
    for t in range(1,len(mp.index)):
        date = mp.index[t]
        pr1 = mp[s][date]
        ret = (pr1-pr0)/pr0
        mr.at[date, s] = ret
        pr0 = pr1
```

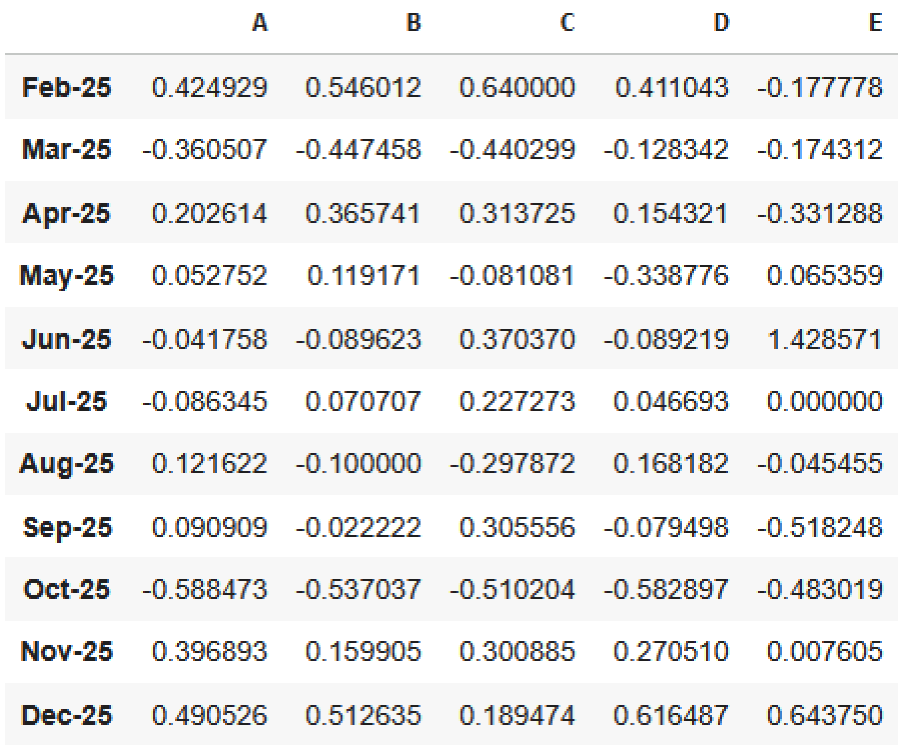
Gather and Analyze Data

Raw data table



Database	A	B	C	D	E
Jan-25	503	252	123	230	74
Feb-25	353	163	75	163	90
Mar-25	552	295	134	187	109
Apr-25	459	216	102	162	163
May-25	436	193	111	245	153
Jun-25	455	212	81	269	63
Jul-25	498	198	66	257	63
Aug-25	444	220	94	220	66
Sep-25	407	225	72	239	137
Oct-25	989	486	147	573	265
Nov-25	708	419	113	451	263
Dec-25	475	277	95	279	160

Return matrix



	A	B	C	D	E
Feb-25	0.424929	0.546012	0.640000	0.411043	-0.177778
Mar-25	-0.360507	-0.447458	-0.440299	-0.128342	-0.174312
Apr-25	0.202614	0.365741	0.313725	0.154321	-0.331288
May-25	0.052752	0.119171	-0.081081	-0.338776	0.065359
Jun-25	-0.041758	-0.089623	0.370370	-0.089219	1.428571
Jul-25	-0.086345	0.070707	0.227273	0.046693	0.000000
Aug-25	0.121622	-0.100000	-0.297872	0.168182	-0.045455
Sep-25	0.090909	-0.022222	0.305556	-0.079498	-0.518248
Oct-25	-0.588473	-0.537037	-0.510204	-0.582897	-0.483019
Nov-25	0.396893	0.159905	0.300885	0.270510	0.007605
Dec-25	0.490526	0.512635	0.189474	0.616487	0.643750

Other steps:

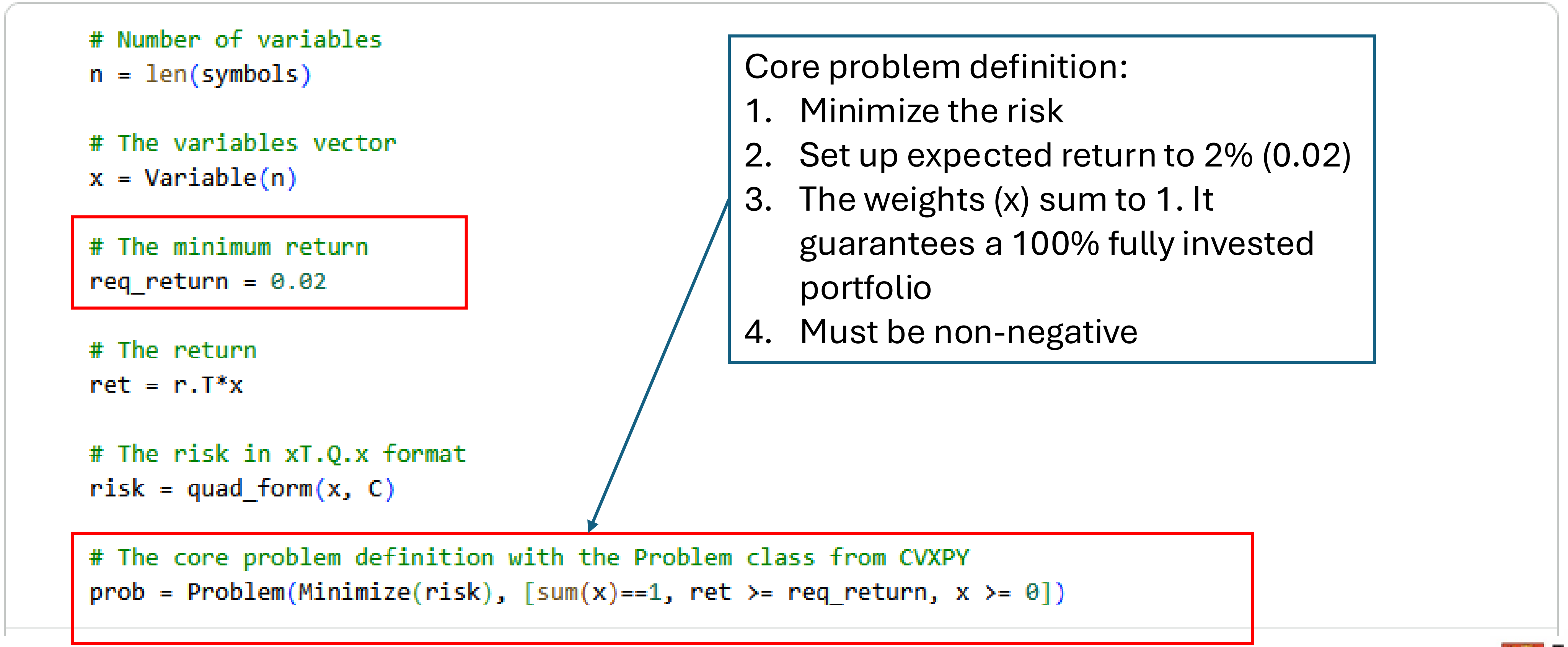
1. Compute mean return
2. Compute covariance matrix
3. Convert data frame to a numpy matrix
4. ………………….

Use Python to solve the optimization problem

- Python library used: CVXPY
  - Python-embedded modeling language for convex optimization problems
  - Express the problem in a natural way that follows the mathematical model

Use Python to solve the optimization problem

Set up the optimization model



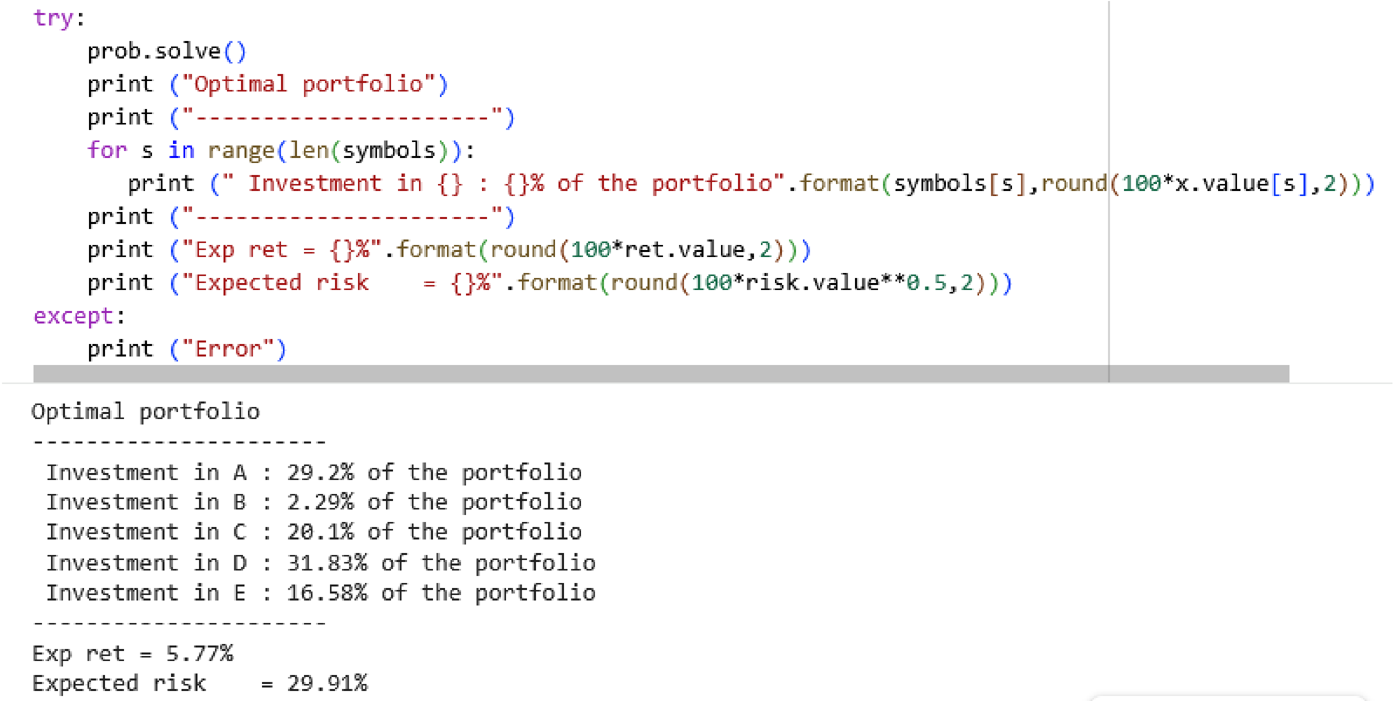
```
# Number of variables
n = len(symbols)

# The variables vector
x = Variable(n)

# The minimum return
req_return = 0.02

# The return
ret = r.T*x

# The risk in xT.Q.x format
risk = quad_form(x, C)

# The core problem definition with the Problem class from CVXPY
prob = Problem(Minimize(risk), [sum(x)==1, ret >= req_return, x >= 0])
```

Core problem definition:

1. Minimize the risk
2. Set up expected return to 2% (0.02)
3. The weights (x) sum to 1. It guarantees a 100% fully invested portfolio
4. Must be non-negative

Solve the optimization problem using Python

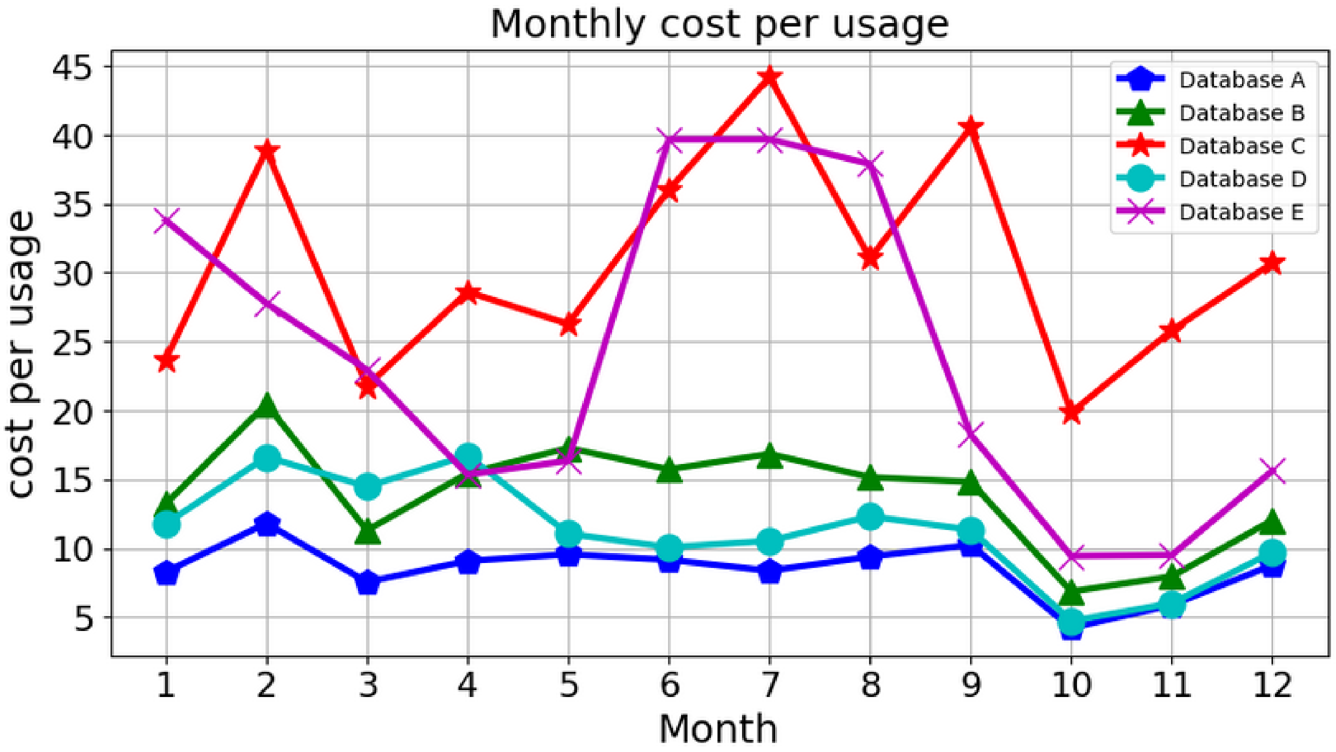
Optimal portfolio	
Database	**Investment Percentage**
A	29.2%
B	2.29%
C	20.1%
D	31.83%
E	16.58%



```
try:
    prob.solve()
    print ("Optimal portfolio")
    print ("----------------------")
    for s in range(len(symbols)):
       print (" Investment in {} : {}% of the portfolio".format(symbols[s],round(100*x.value[s],2)))
    print ("----------------------")
    print ("Exp ret = {}%".format(round(100*ret.value,2)))
    print ("Expected risk     = {}%".format(round(100*risk.value**0.5,2)))
except:
    print ("Error")
```

```
Optimal portfolio
----------------------
 Investment in A : 29.2% of the portfolio
 Investment in B : 2.29% of the portfolio
 Investment in C : 20.1% of the portfolio
 Investment in D : 31.83% of the portfolio
 Investment in E : 16.58% of the portfolio
----------------------
Exp ret = 5.77%
Expected risk     = 29.91%
```

Result Summary



Optimal portfolio	
Database	Investment Percentage
A	29.2%
B	2.29%
C	20.1%
D	31.83%
E	16.58%

Limitations

- Complexity with Large Problems
  - Computational complexity can increase significantly with the number of variables and constraints, leading to longer solve computational time
- Assumptions About Data
  - LP assumes all parameters are known with certainty.
  - Many real-world situations involve uncertainty in parameters (e.g., demand, costs, resource availability)

Q&A

Thank You