# Final Project

2110531
Data Science and Data Engineering Tools

# Project Team

**Peeraporn Rimdusit**

**Tipakorn Phaengnet**

**Kingrak Phairoh**

# Agenda

01 Objective

02 Project Flow

03 Data Engineering

04 Machine learning

05 Visualization

# Objectives

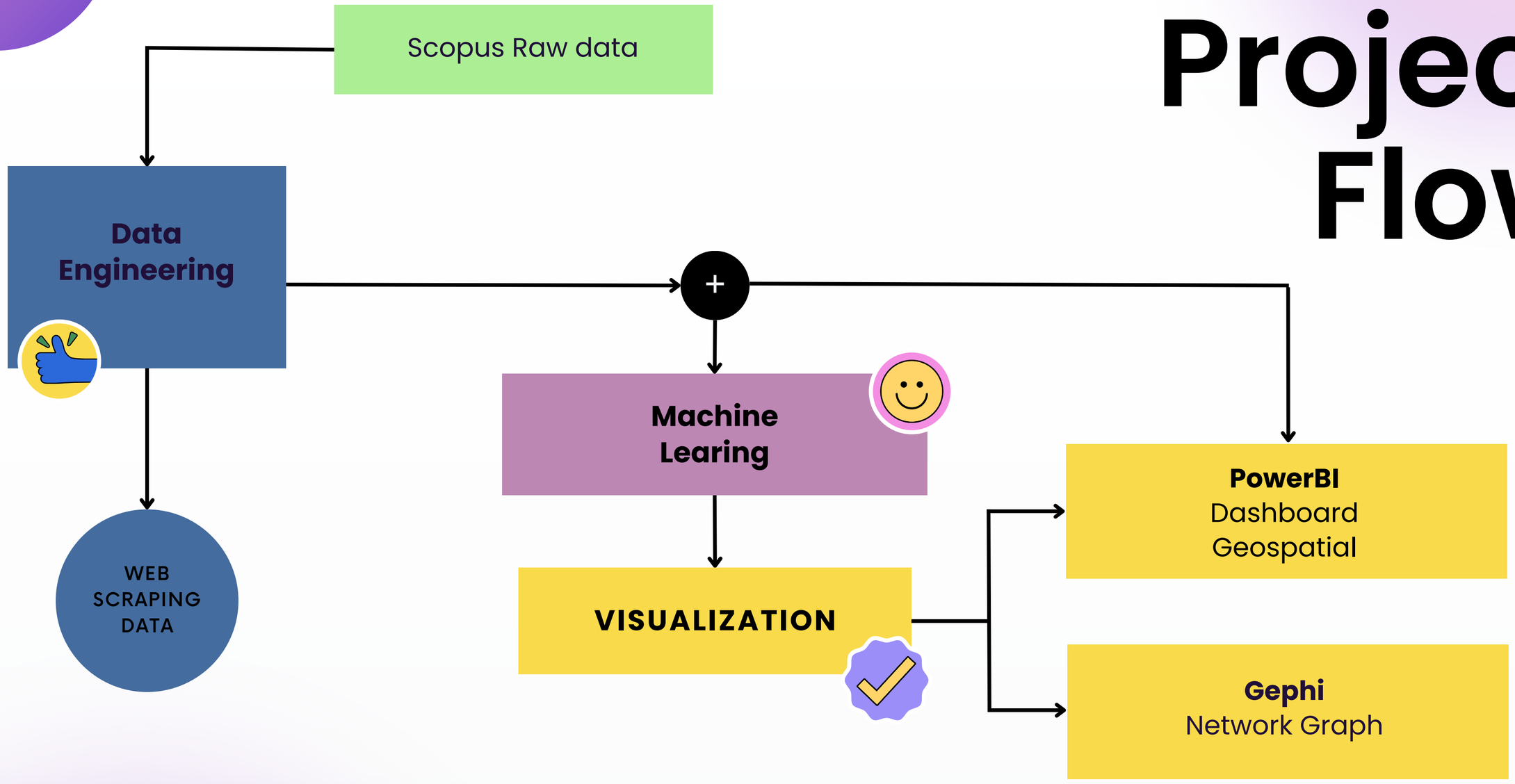
**Objective 1**
**Data Engineering**

Applying data engineering tools for streamlined and efficient data transmission

**Objective 2**
**Machine Learning Visualization**

Utilize the data to derive benefits and present the findings in an aesthetically pleasing manner

# Project Flow

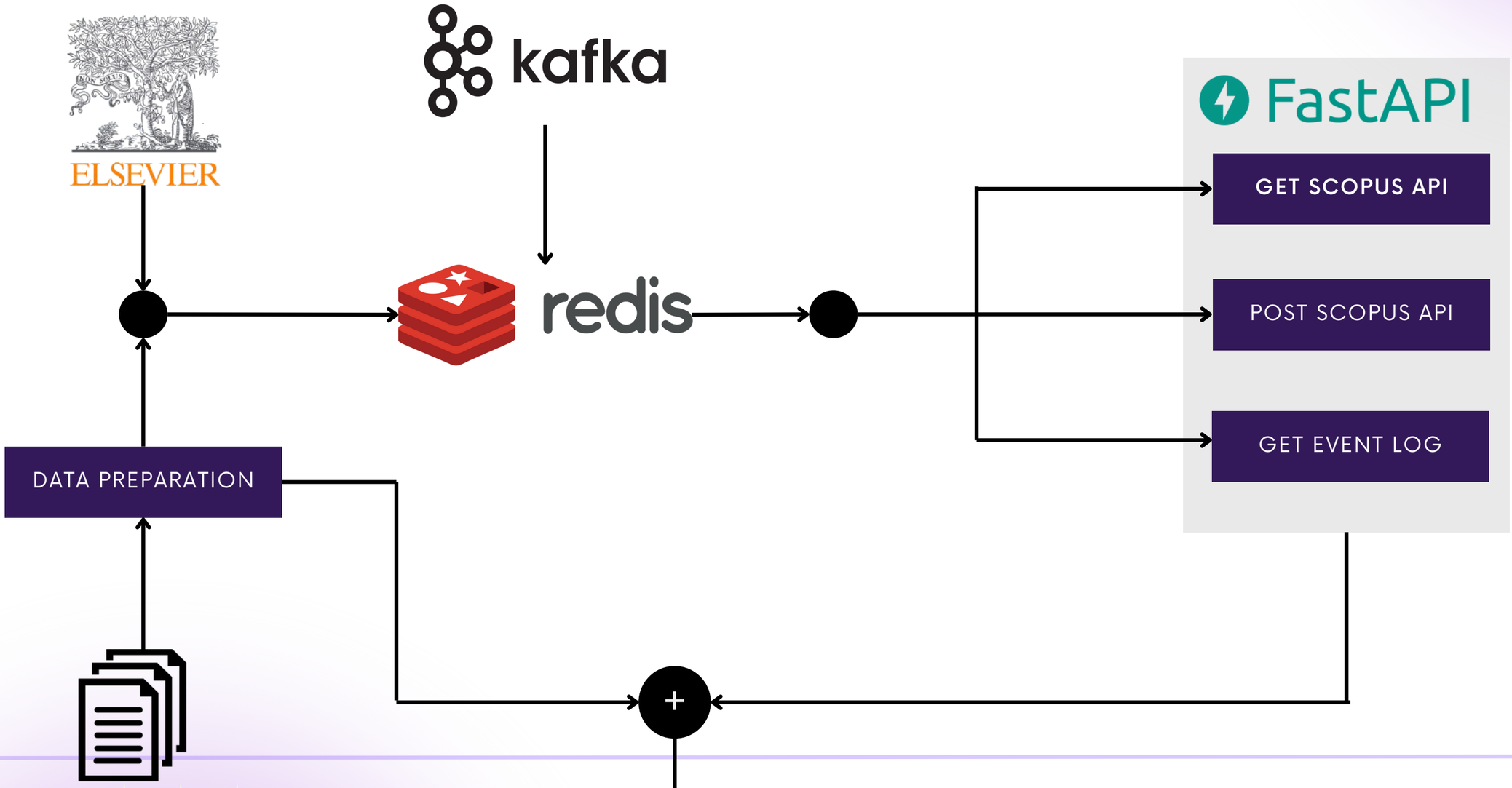


# Data Engineering

Prepare a data pipeline to facilitate seamless extraction and utilization of data for end-users.

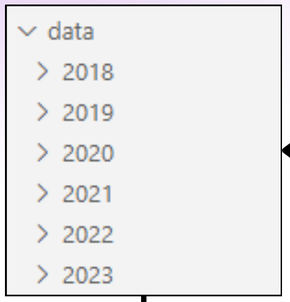
Next Page

# Data Pipeline

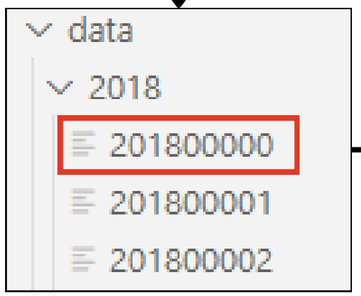


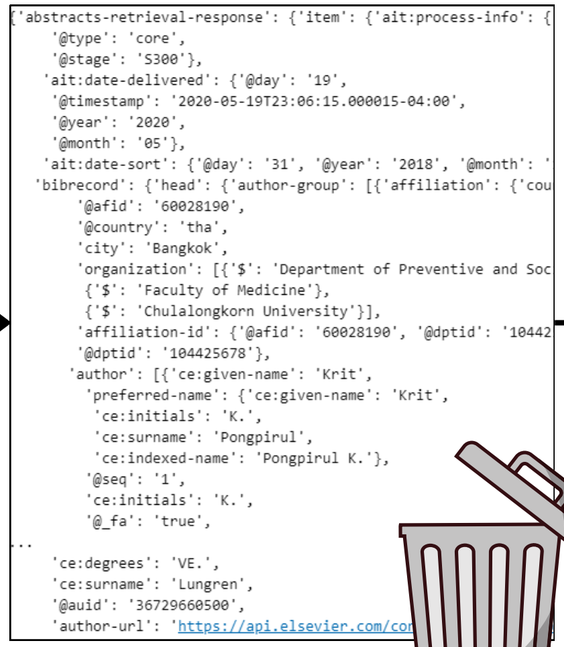
# DATA ENGINEERING

## 1. PREPARING RAW SCOPUS DATA







```
{'abstracts-retrieval-response': {'item': {'ait:process-info': {
    '@type': 'core',
    '@stage': 'S300'},
   'ait:date-delivered': {'@day': '19',
    '@timestamp': '2020-05-19T23:06:15.000015-04:00',
    '@year': '2020',
    '@month': '05'},
   'ait:date-sort': {'@day': '31', '@year': '2018', '@month': '
  'bibrecord': {'head': {'author-group': [{'affiliation': {'cou
       '@afid': '60028190',
       '@country': 'tha',
       'city': 'Bangkok',
       'organization': [{'$': 'Department of Preventive and Soc
        {'$': 'Faculty of Medicine'},
        {'$': 'Chulalongkorn University'}],
       'affiliation-id': {'@afid': '60028190', '@dptid': '10442
       '@dptid': '104425678'},
      'author': [{'ce:given-name': 'Krit',
        'preferred-name': {'ce:given-name': 'Krit',
         'ce:initials': 'K.',
         'ce:surname': 'Pongpirul',
         'ce:indexed-name': 'Pongpirul K.'},
        '@seq': '1',
        'ce:initials': 'K.',
        '@_fa': 'true',
...
    'ce:degrees': 'VE.',
    'ce:surname': 'Lungren',
    '@auid': '36729660500',
    'author-url': 'https://api.elsevier.com/co
```



**27**

**Neccessary Attributes**

```
['Id',
 'Abstract',
 'Classification',
 'Year',
 'Reference_number',
 'Reference_title',
 'Reference_title_str',
 'Number_mainterm',
 'Mainterms',
 'Affiliation',
 'University',
 'Country',
 'Title',
 'Publisher',
 'PublicationName',
 'Type',
 'Description',
 'Openaccess',
 'Coppyrigth',
 'Citation_Number',
 'Language',
 'Number_Keywords',
 'Keywords',
 'Subject_Area',
 'Subject_Area_Code',
 'Number_authors',
 'Index_authors']
```

| | Id | Abstract | Classification | Year | Reference_numbe |
|---|---|---|---|---|---|
| 0 | 201800000 | None | [2700, MEDI] | 2018 | 76 |
| 1 | 201800001 | © 2018 The Institute of Electronics, Informati... | [[{'$': '2208'}, {'$': '2504'}], [{'classifica... | 2018 | 4 |
| 2 | 201800002 | © 2018 Elsevier LtdComputational fluid dynamic... | [[{'classification-code': '522', 'classificati... | 2018 | 42 |
| 3 | 201800003 | © 2018 Elsevier B.V. A superhydrophobic/supero... | [[{'classification-code': '641.1', 'classifica... | 2018 | 45 |
| 4 | 201800004 | © 2018 Elsevier B.V. A label-free electrochemi... | [{'classification-code': '4', 'classification-... | 2018 | 55 |

5 rows × 27 columns

DATA ENGINEERING

## 2. WEB SCRAPING USING SCOPUS API

# SCRAPE MORE 1,200 PAPERS

- USING SCOPUS API TO QUERY PAPERS IN THE LATEST 6 YEARS (DEFAULT)

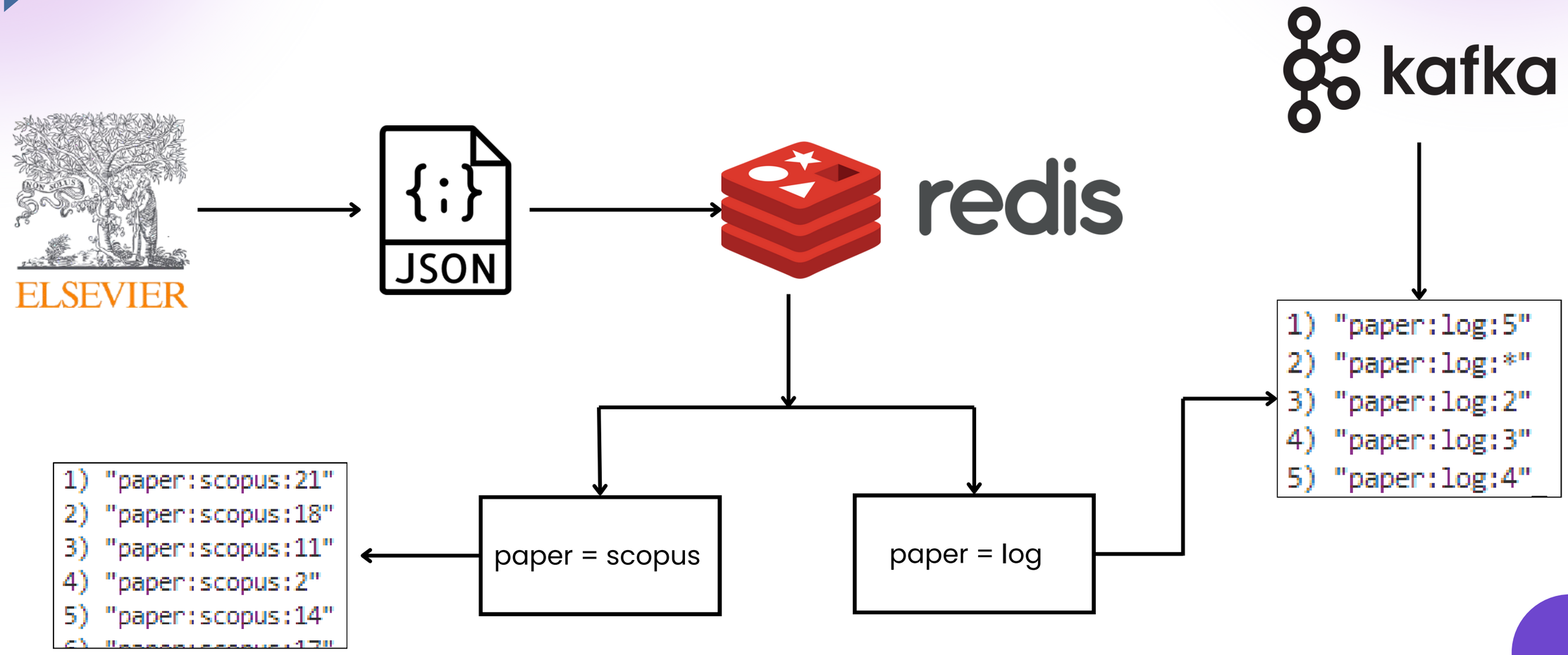
**15 Necessary Atrributes**

```
['Title',
 'Abstract',
 'University',
 'Country',
 'Index_authors',
 'Keywords',
 'PublicationName',
 'Publisher',
 'Subject_Area',
 'Subject_Area_CodeType',
 'Citation_Type',
 'Citation_Number',
 'Reference_number',
 'Year',
 'Reference_title']
```

| | Title | Abstract | University | Country | Index_authors | Keywords | Pub |
|---|---|---|---|---|---|---|---|
| 0 | Solid Waste Management Awareness, Attitude, an... | © 2017, University of San Jose-Recoletos. All ... | ['Silliman University', 'University of Negros ... | ['Philippines'] | ['Oracion E.G.', 'Madrigal D.V.'] | ['Attitude', 'Awareness', 'Practices', 'Solid ... | M Re |
| 1 | Economic Decision across Regions of the Philip... | © 2017, University of San Jose-Recoletos. All ... | ['Southern Leyte State University'] | ['Philippines'] | ['Ordiz J.E.G.'] | ['economic disparity', 'multivariate analysis'... | M Re |
| 2 | Evaluating Philippine Students' Class Particip... | © 2017, University of San Jose-Recoletos. All ... | ['University of the Philippines Cebu'] | ['Philippines'] | ['Yu S.Q.'] | ['class participation', 'token economy'] | M Re |
| 3 | Energy and society | NaN | NaN | NaN | NaN | NaN | Case |
| 4 | Stimulating economic recovery through euro are... | © Institute of Economic Research.Research back... | ['Slovak Academy of Sciences', 'University of ... | ['Slovakia'] | ['Siranova M.', 'Kotlebova J.'] | ['Growth pole', 'Network analysis', 'Unconvent... | Quart E |

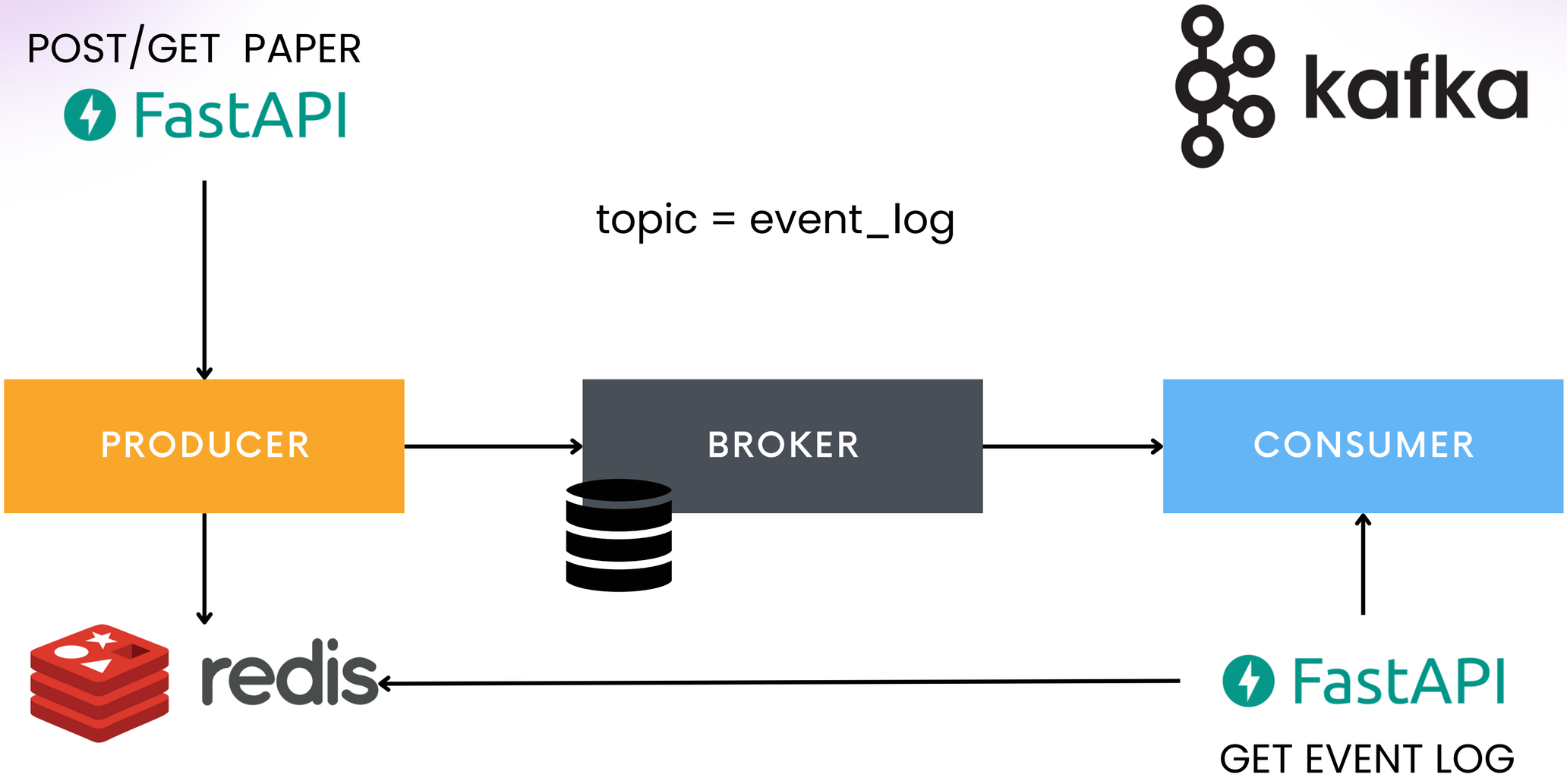
DATA ENGINEERING

# 3. STORING DATA IN REDIS



DATA ENGINEERING

# 4. EVENT STEAMING WHEN ACCESSING API



DATA ENGINEERING

## 5. API GATEWAY

FastAPI

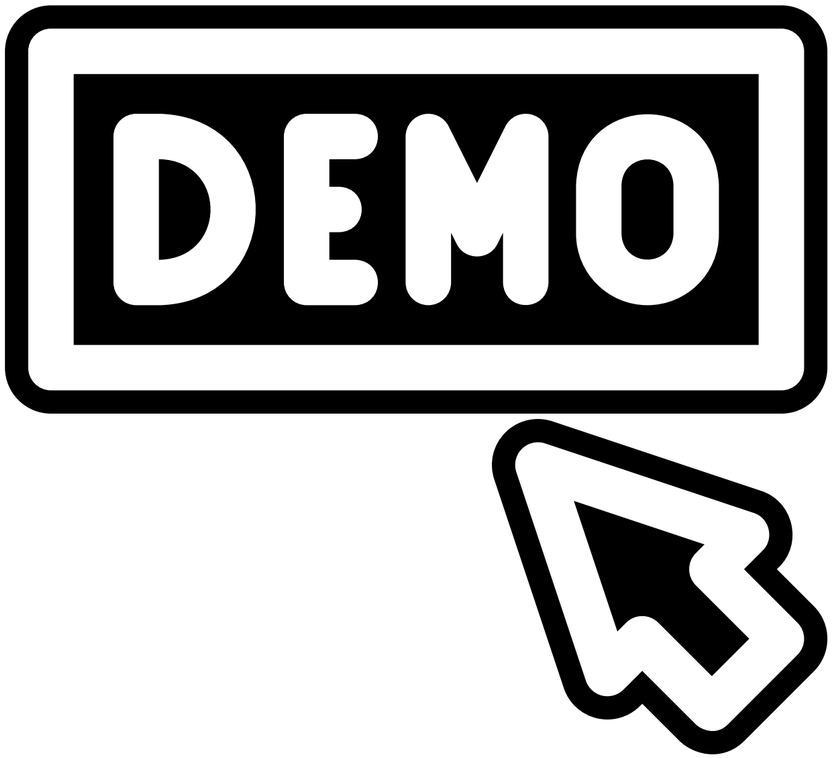
GET SCOPUS API

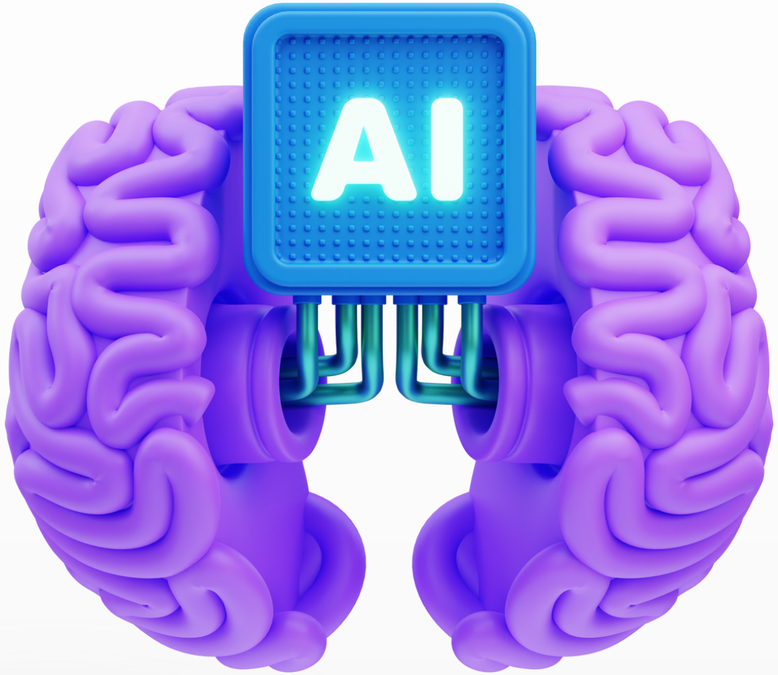
GET EVENT LOG

POST SCOPUS API

| (request path) | | (response) |
|---|---|---|
| GET /scopus_api | ⟶ | all scorpus papers stored in redis |
| GET /event_log | ⟶ | all events stored in redis |
| POST /scopus_api | ⟶ | papers that queried from api scopus |

```
body {
  “api_key”: “xxxxxxxxxxxx”,
  “query”: “Yolov8”,
  “count”: 10
}
```


DEMO




# Machine Learning

**Objective:**
Group research papers related to their titles, abstracts, and keywords for recommending about research papers within each group. This serves as a **guideline for further exploration and investigation of research papers.**

Next Page

# DATA PREPROCESSING

## 1. CONCAT DATA + CHOOSE ESSENTIAL COLUMNS & DROPNA

RAW SCOPUS DATA + DATA SCOPUS API = **17253 PAPERS**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20216 entries, 0 to 20215
Data columns (total 28 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Unnamed: 0           20216 non-null  int64
 1   Id                   20216 non-null  int64
 2   Title                20215 non-null  object
 3   Publisher            20205 non-null  object
 4   PublicationName      20216 non-null  object
 5   Type                 20216 non-null  object
 6   Description          19551 non-null  object
 7   Openaccess           18548 non-null  float64
 8   Coppyrigth           19478 non-null  object
 9   Citation_Number      20212 non-null  float64
 10  Number_authors       20216 non-null  int64
 11  Index_authors        20216 non-null  object
 12  Affiliation          20216 non-null  object
 13  University           20216 non-null  object
 14  Country              20216 non-null  object
 15  Subject_Area         20216 non-null  object
 16  Subject_Area_Code    20216 non-null  object
 17  Number_Keywords      20216 non-null  int64
 18  Keywords             16454 non-null  object
 19  Number_mainterm      20216 non-null  int64
 20  Mainterms            11572 non-null  object
 21  Language             20096 non-null  object
 22  Abstract             19551 non-null  object
 23  Classification       20216 non-null  object
 24  Year                 20216 non-null  int64
 25  Reference_number     19805 non-null  float64
 26  Reference_title      19805 non-null  object
 27  Reference_title_str  19805 non-null  object
dtypes: float64(3), int64(6), object(19)
memory usage: 4.3+ MB
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1200 entries, 0 to 1199
Data columns (total 15 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Title                  1200 non-null   object
 1   Abstract               1129 non-null   object
 2   University             1147 non-null   object
 3   Country                1147 non-null   object
 4   Index_authors          1186 non-null   object
 5   Keywords               939 non-null    object
 6   PublicationName        1200 non-null   object
 7   Publisher              1200 non-null   object
 8   Subject_Area           1200 non-null   object
 9   Subject_Area_CodeType  1200 non-null   object
 10  Citation_Type          1200 non-null   object
 11  Citation_Number        1200 non-null   int64
 12  Reference_number       1110 non-null   float64
 13  Year                   1103 non-null   float64
 14  Reference_title        1110 non-null   object
dtypes: float64(2), int64(1), object(12)
memory usage: 140.8+ KB
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 17253 entries, 0 to 17264
Data columns (total 9 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Title              17253 non-null  object
 1   Abstract           17253 non-null  object
 2   Keywords           17253 non-null  object
 3   Subject_Area       17253 non-null  object
 4   Subject_Area_Code  17253 non-null  object
 5   Type               17253 non-null  object
 6   Citation_Number    17249 non-null  float64
 7   Year               17230 non-null  float64
 8   Reference_title    17253 non-null  object
dtypes: float64(2), object(7)
memory usage: 1.3+ MB
```

- CHOOSE ESSENTIAL COLUMNS
- DROPNA (TITLE,ABSTRACT,KEYWORDS)
- DROP DUPLICATE TITLE

# DATA PREPROCESSING

## 2. PREPARE TEXT FOR CLUSTERING

**COMBINE TEXT**

- TITLE
- ABSTRACT
- KEYWORDS

**Text preprocessing**

- Remove some special characters not found in the paper
- Lower characters
- Remove English stopword

**Hugging Face**

**Encode by Embedder**

- SentenceTransformer('distilbert-base-nli-stsb-mean-tokens')
- Maps sentences & paragraphs to a 768 dimensional dense vector space

**FEED TO CLUSTERING MODEL**

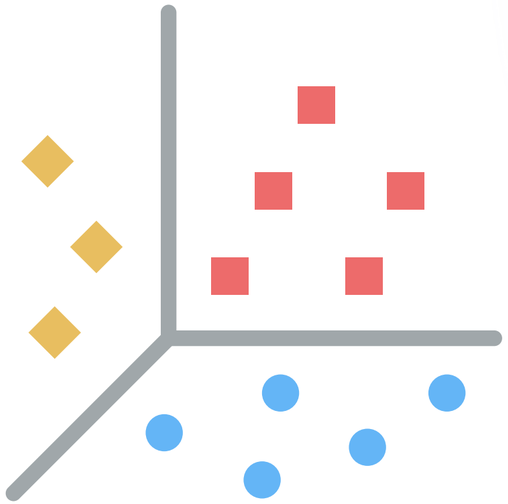
## TRAINING & EVALUATION

### USE KMEANS CLUSTERING VARY K = [2, 15]

Preprocessing Data

**KMEANS**

N_CLUSTER = [2,15]
RANDOMSTATE = 42

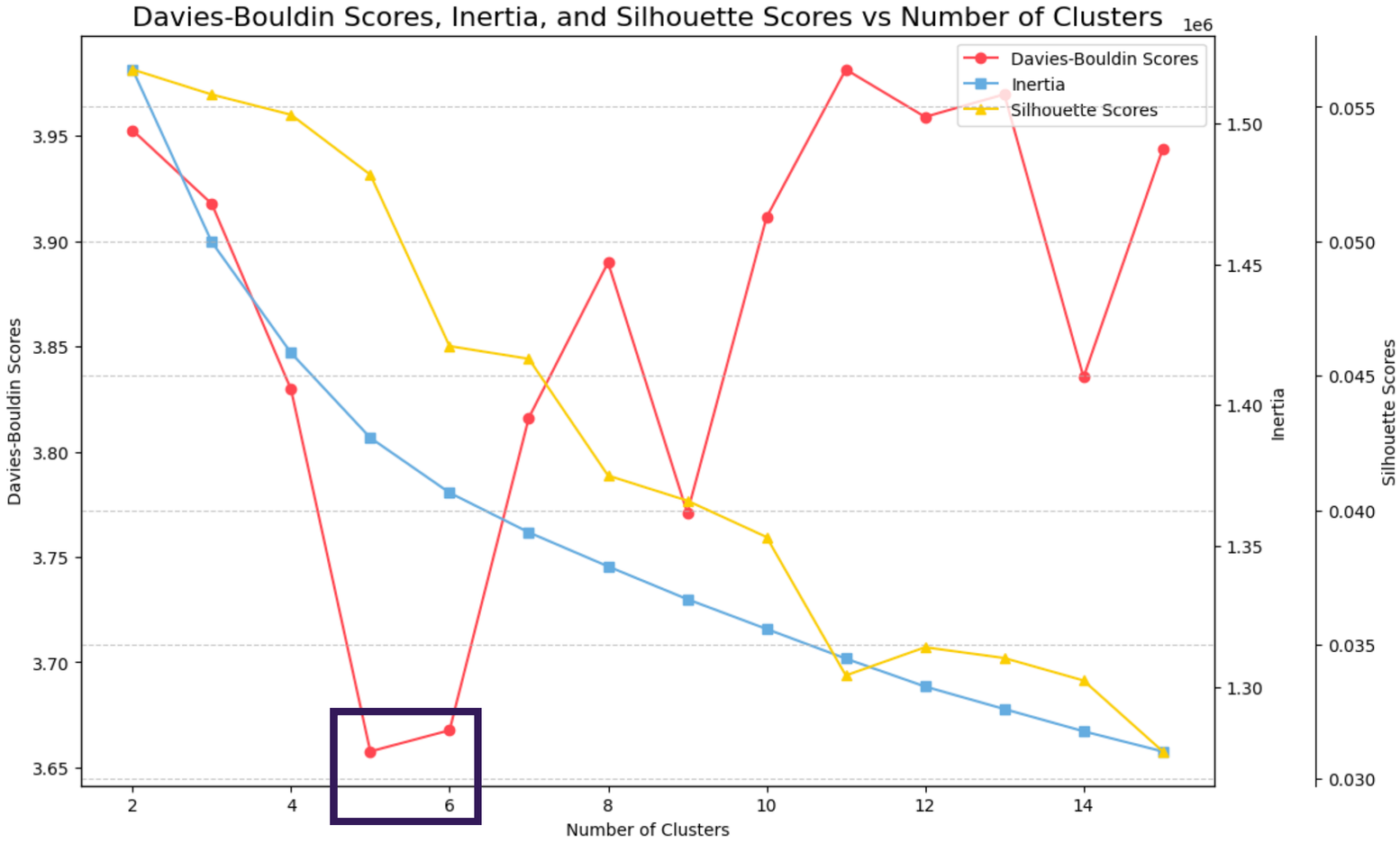


**Evaluated Score**

- Inertia
- Silhouette Scores
- Davies-Bouldin Scores

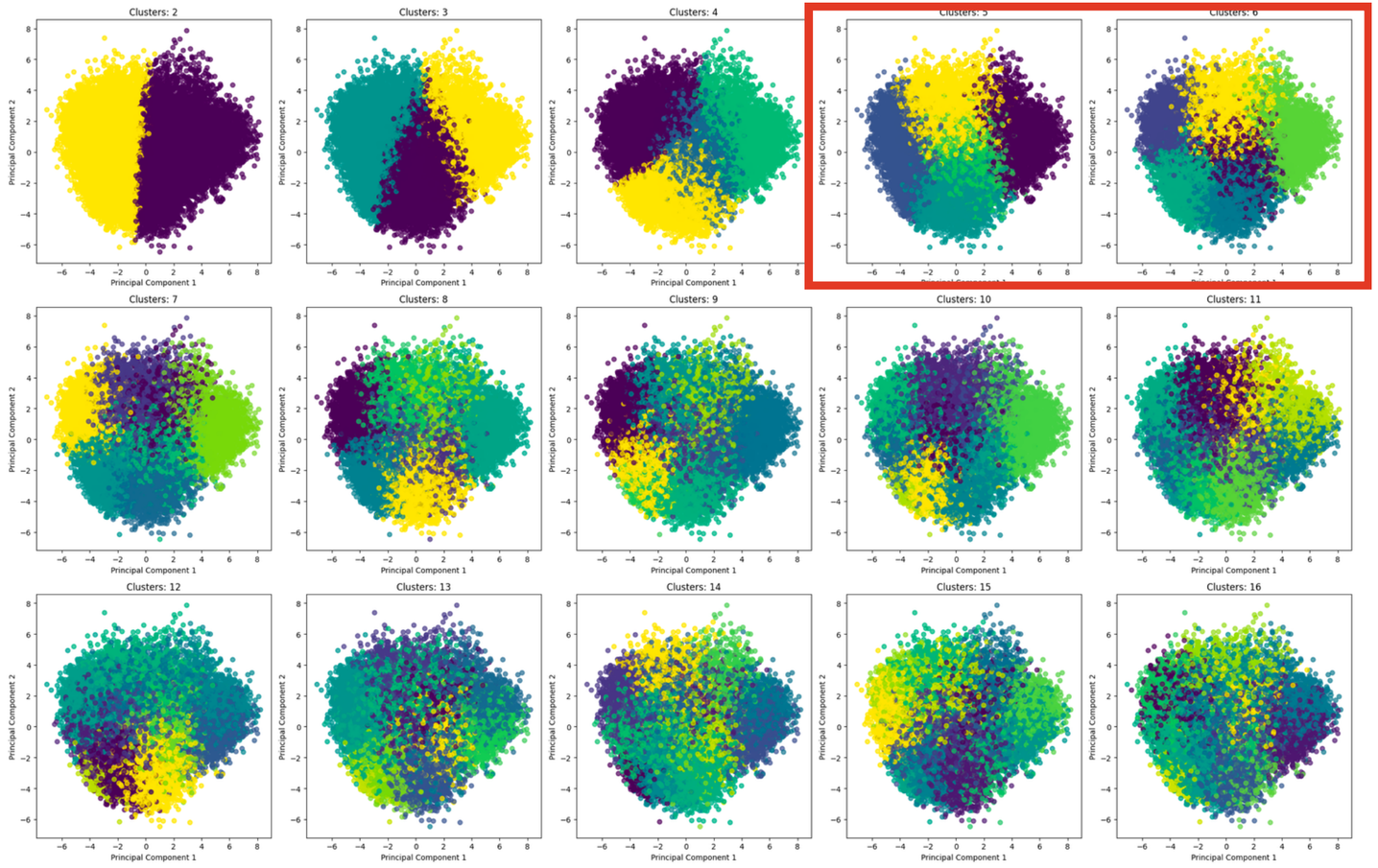
TRAINING & EVALUATION

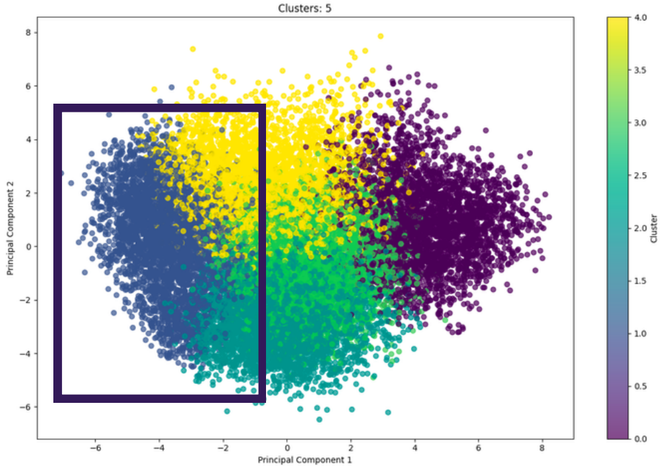
# RESULT OF EVALUATE SCORE

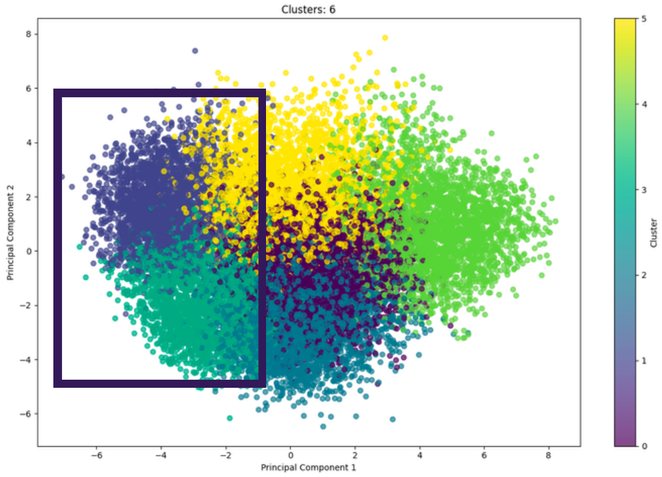


# TRAINING & EVALUATION

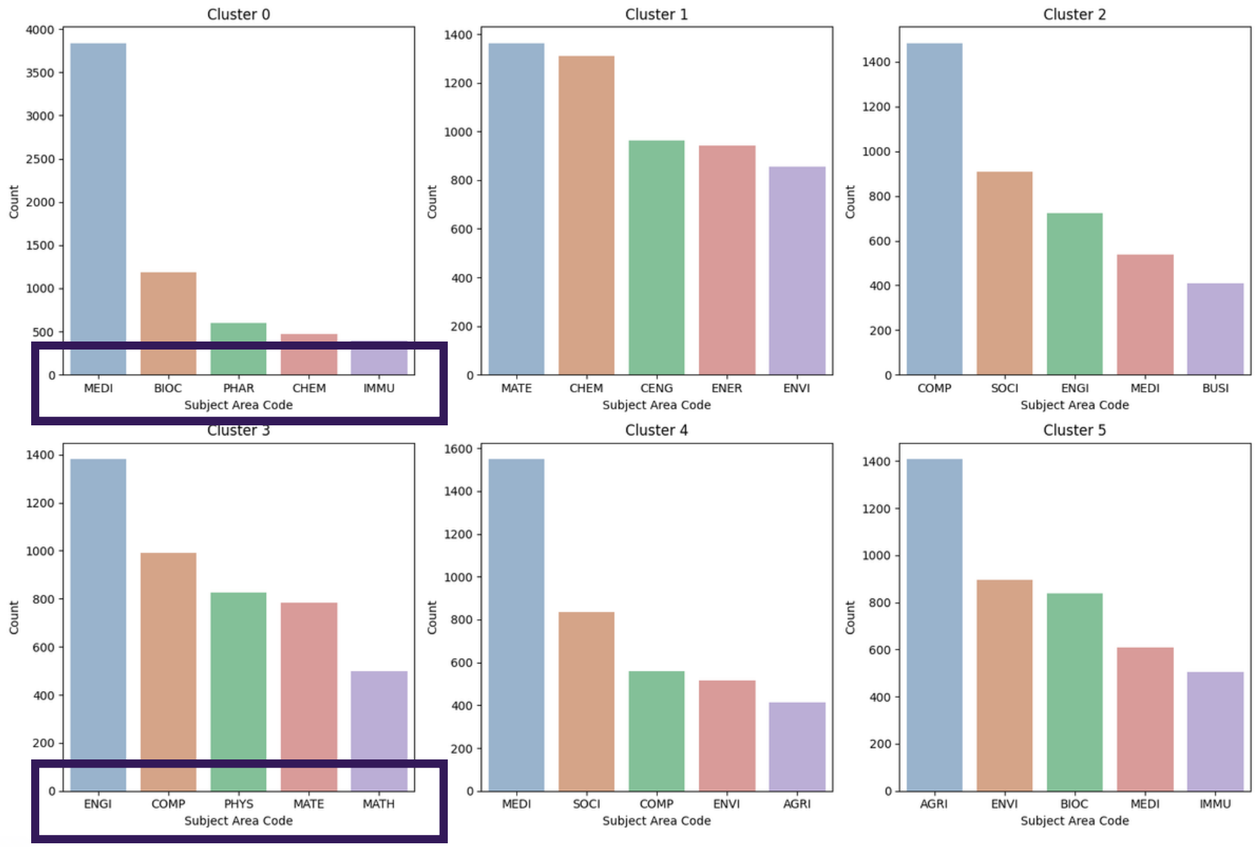
## PLOT PCA (COMPNENT = 2)



# SELECT K = 6 AND PLOT SUBJECT AREA EACH CLUSTER




Use PowerBI to create more insightful visualizations.

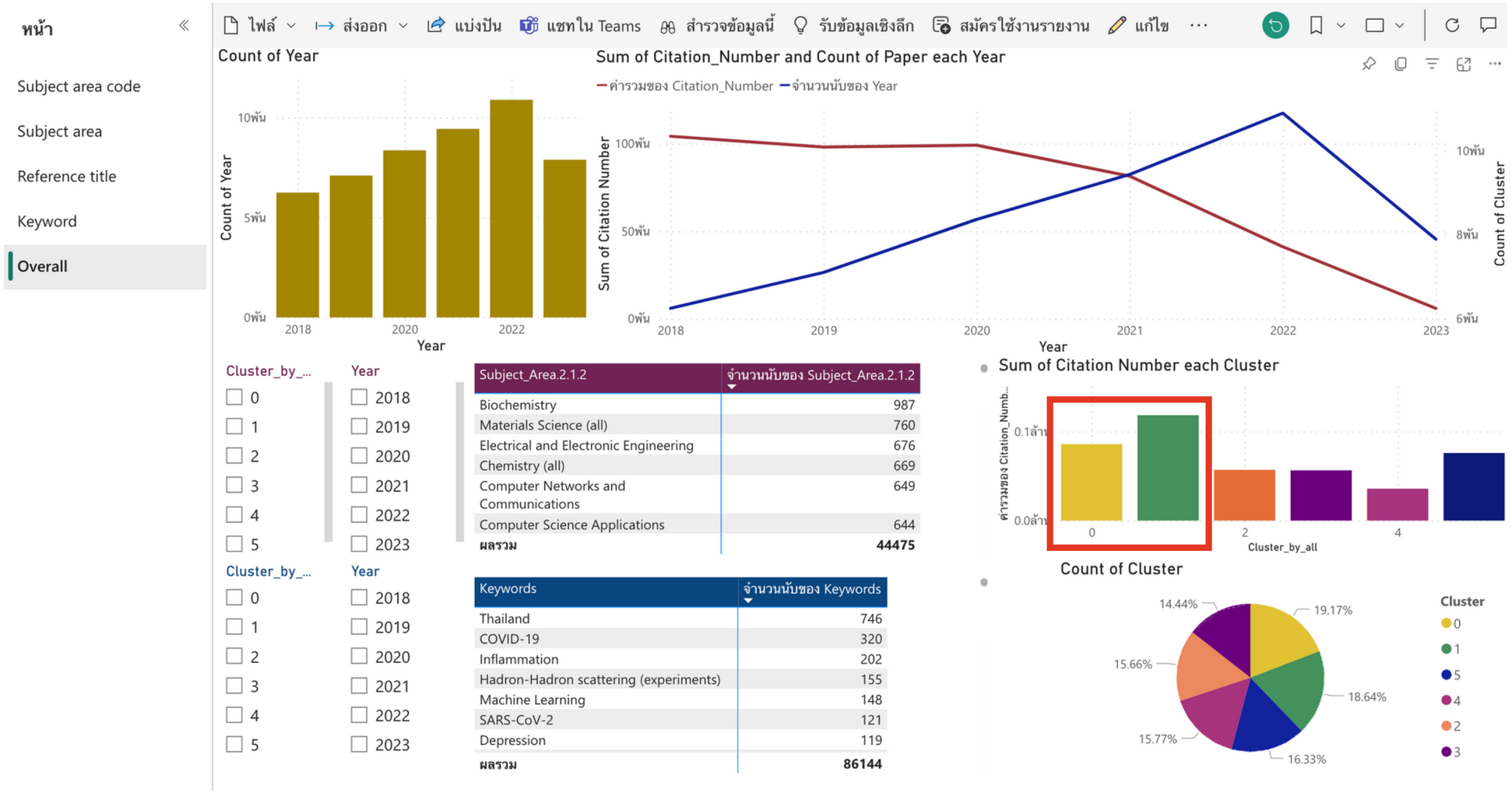
# Visualization

Using PowerBI and Gephi for visualization to uncover insightful findings.

Next Page

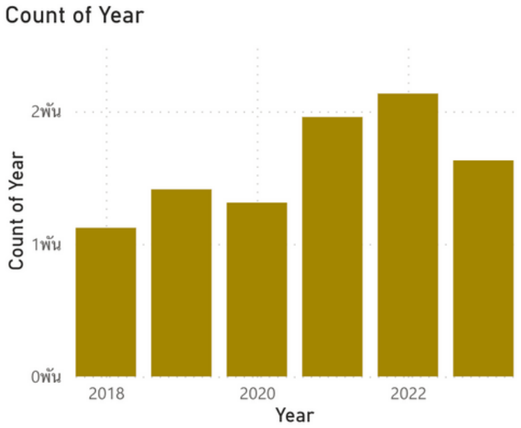
Power BI

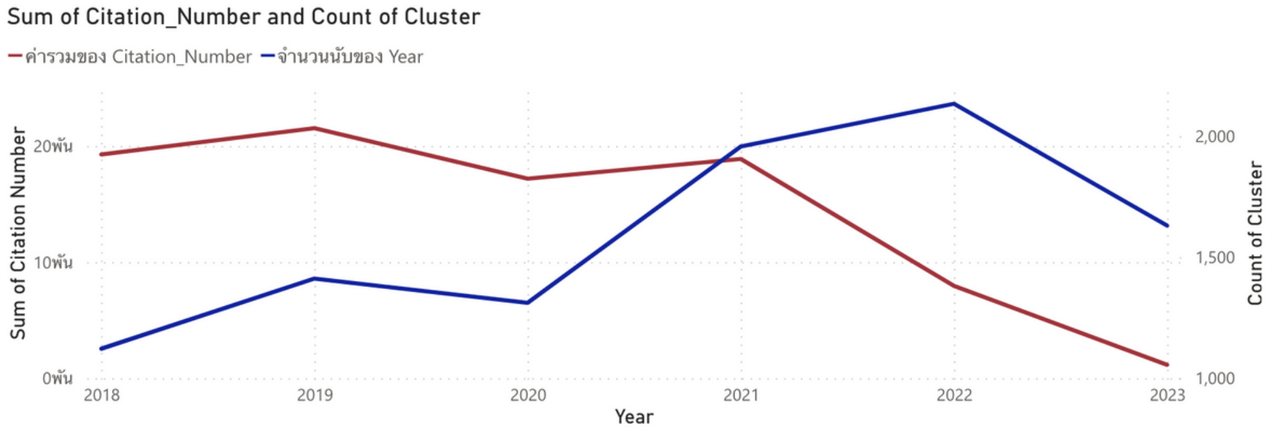
# OVERALL

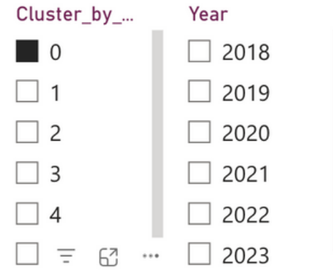


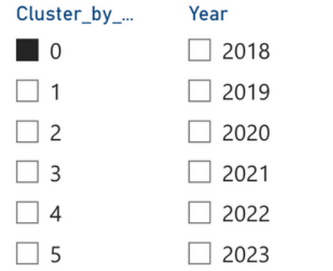
Power BI

# CLUSTER 0



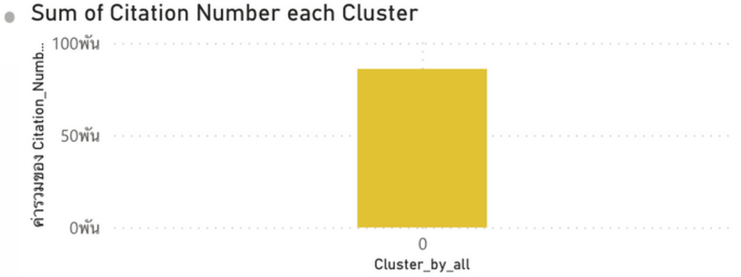



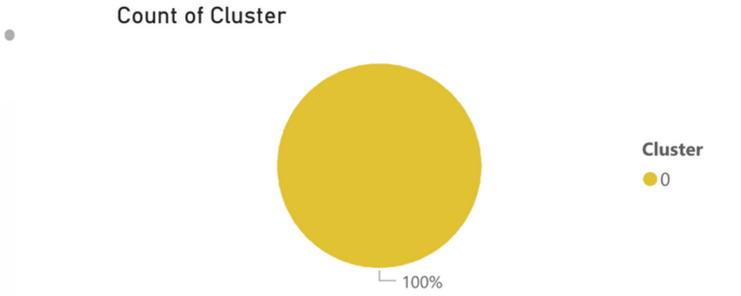




| Subject_Area.2.1.2 | จำนวนนับของ Subject_Area.2.1.2 |
|---|---|
| Infectious Diseases | 305 |
| Biochemistry | 284 |
| Pharmacology | 282 |
| Medicine (all) | 248 |
| Surgery | 217 |
| Veterinary (all) | 203 |
| Oncology | 199 |
| **ผลรวม** | **8770** |

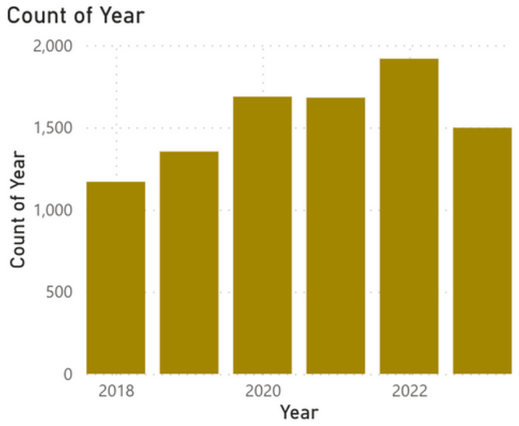
| Keywords | จำนวนนับของ Keywords |
|---|---|
| Inflammation | 153 |
| COVID-19 | 151 |
| SARS-CoV-2 | 85 |
| HIV | 72 |
| Oxidative stress | 72 |
| Meta-analysis | 69 |
| Cytokines | 63 |
| **ผลรวม** | **20485** |

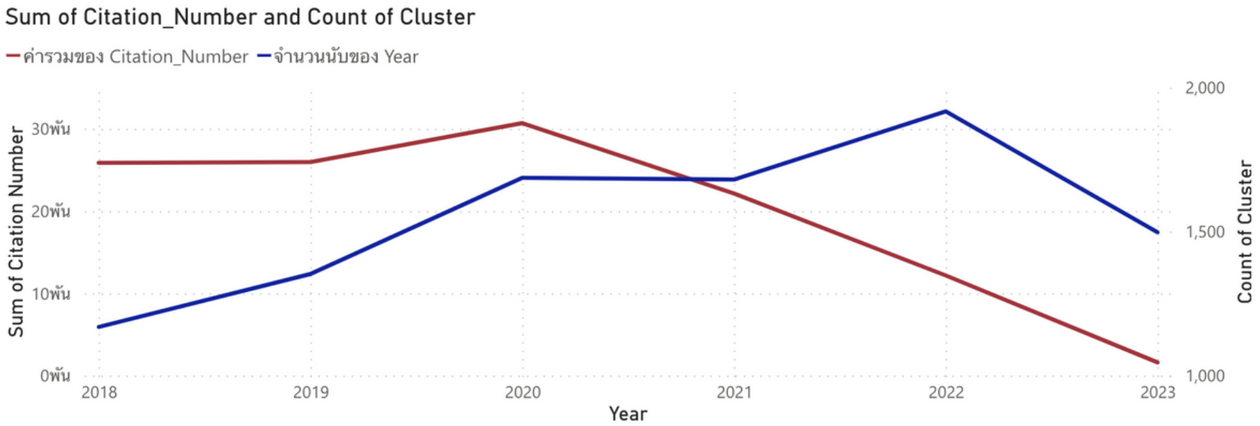


Power BI

# CLUSTER 1



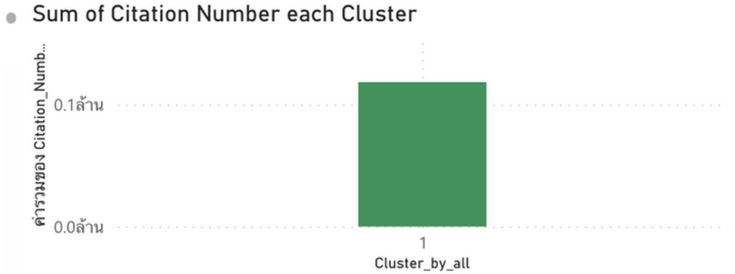



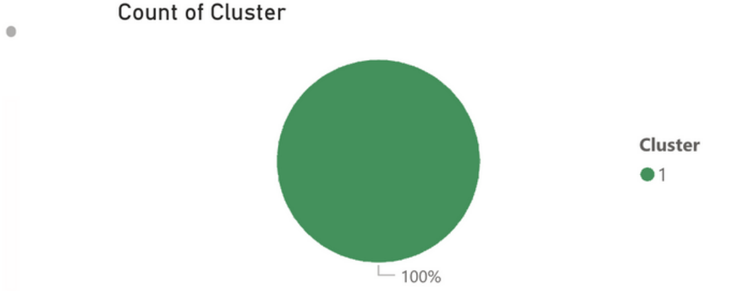





| Subject_Area.2.1.2 | จำนวนนับของ Subject_Area.2.1.2 |
|---|---|
| Chemistry (all) | 444 |
| Chemical Engineering (all) | 365 |
| Materials Science (all) | 329 |
| Renewable Energy | 321 |
| Condensed Matter Physics | 313 |
| Materials Chemistry | 303 |
| Organic Chemistry | 253 |
| **ผลรวม** | **8394** |

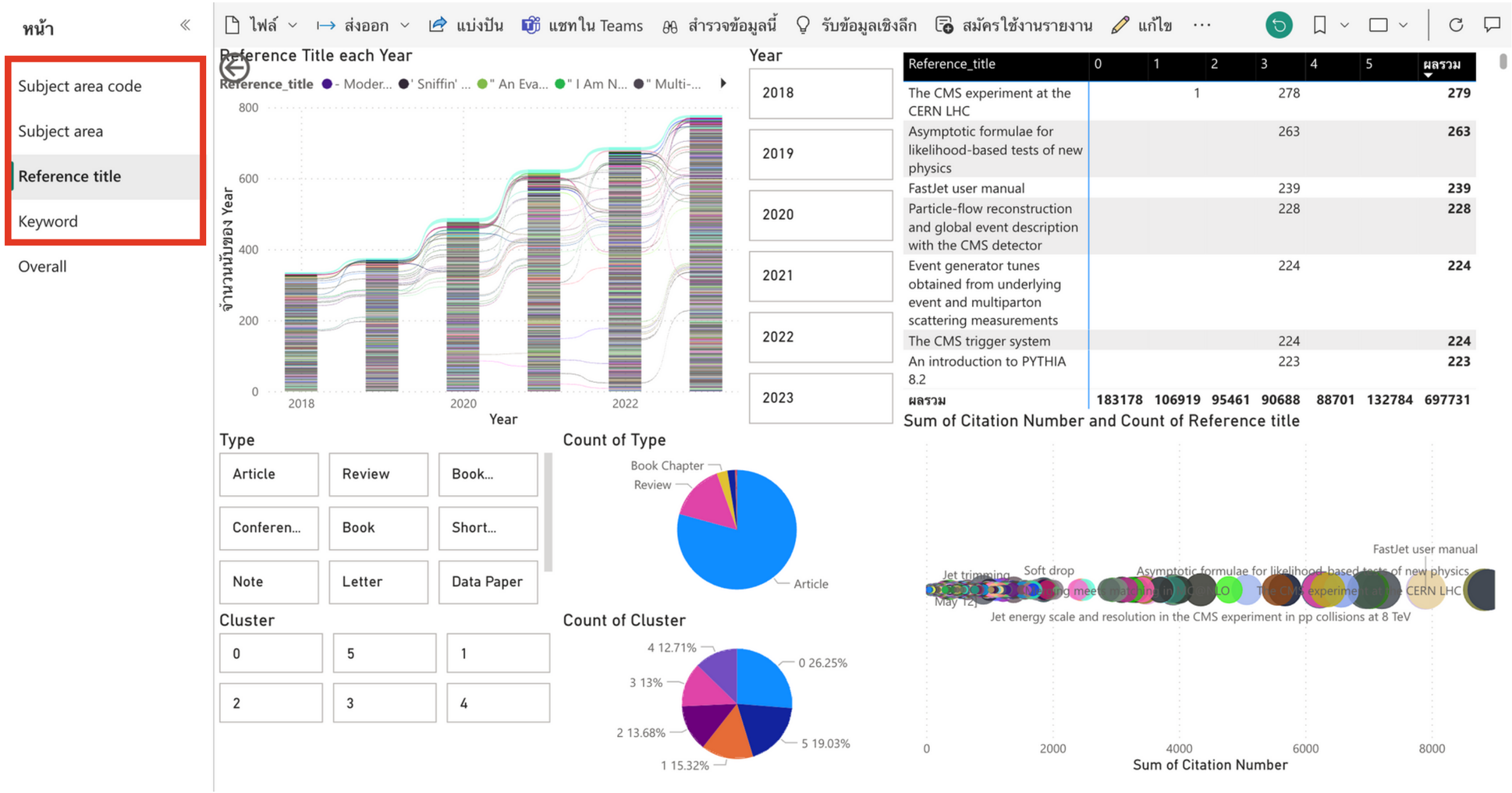
| Keywords | จำนวนนับของ Keywords |
|---|---|
| Adsorption | 64 |
| Biodiesel | 47 |
| Natural rubber | 32 |
| Photocatalyst | 32 |
| Chitosan | 30 |
| Biomass | 28 |
| Glycerol | 27 |
| **ผลรวม** | **13241** |



# Power BI

## ADDITIONAL




https://app.powerbi.com/groups/me/reports/12821efe-75fc-492e-ab6f-23bc47ea027f?ctid=271d5e7b-1350-4b96-ab84-52dbda4cf40c&pbi_source=linkShare

Power BI

# DEMO

**SAMPLE DATA**

**Title:** Graph Neural Network for Molecular Property Prediction

encode and clustering

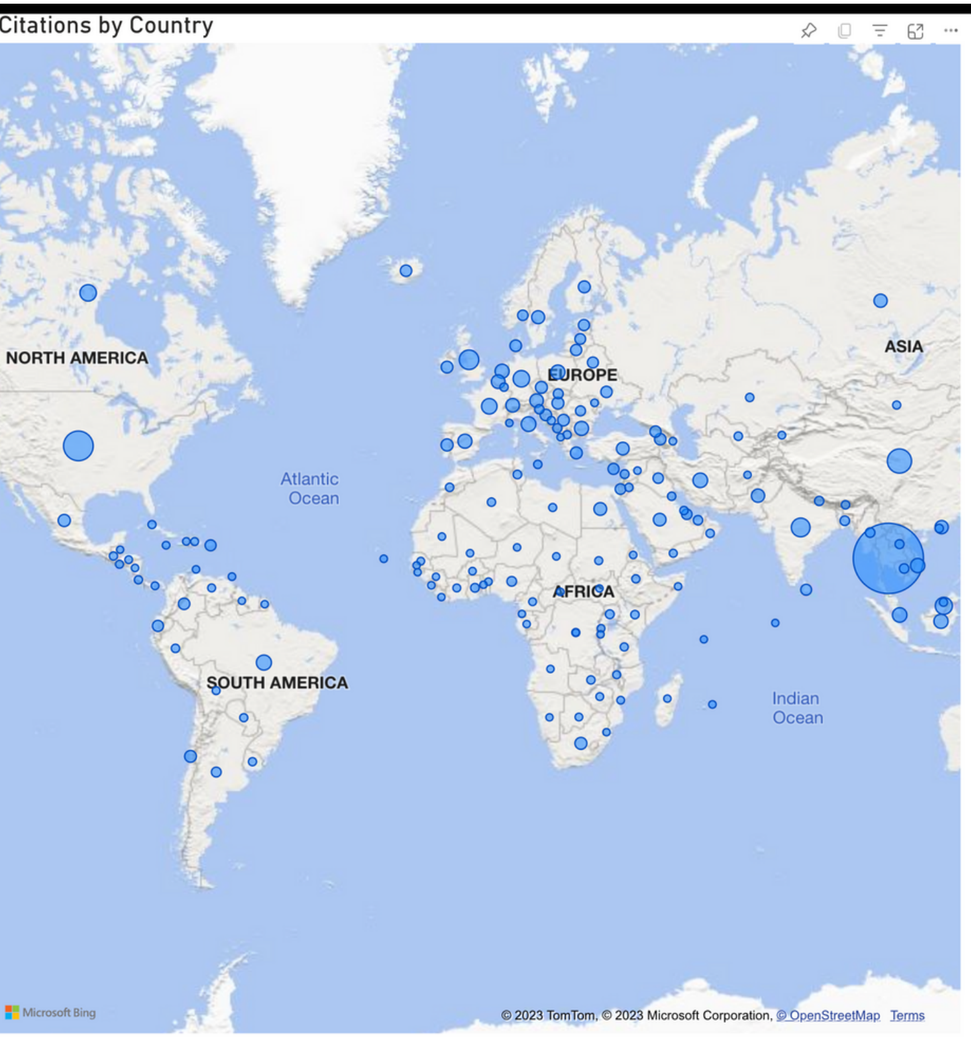
**MODEL**

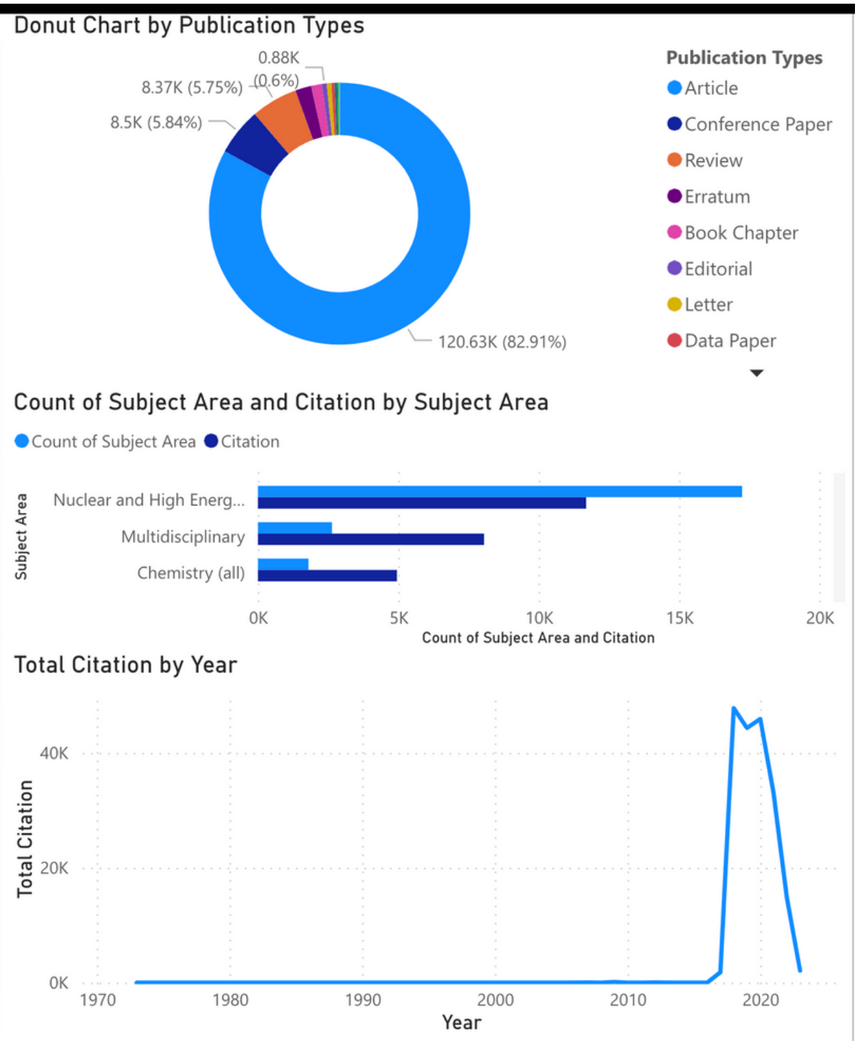
use number clustering to find somthing

**POWERBI**

GET INSIGHT!

# Power BI

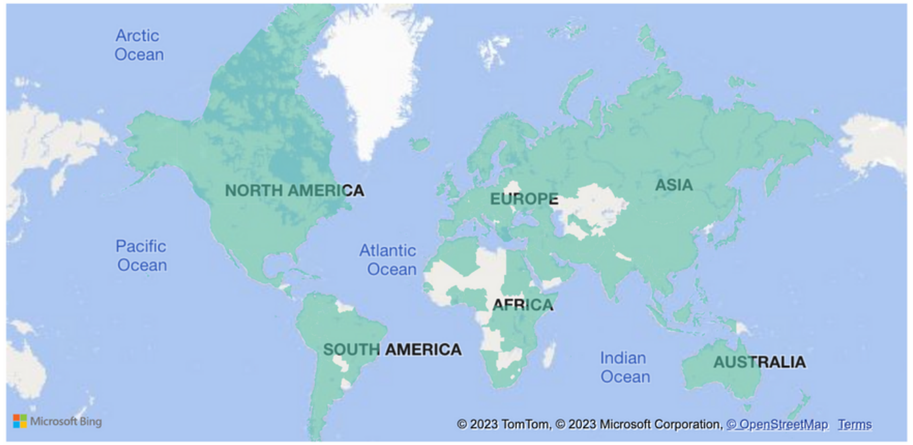

Citations by Country
NORTH AMERICA
EUROPE
ASIA
AFRICA
SOUTH AMERICA
Atlantic Ocean
Indian Ocean
Microsoft Bing
© 2023 TomTom, © 2023 Microsoft Corporation, © OpenStreetMap Terms


Donut Chart by Publication Types
0.88K (0.6%)
8.37K (5.75%)
8.5K (5.84%)
120.63K (82.91%)
Publication Types
Article
Conference Paper
Review
Erratum
Book Chapter
Editorial
Letter
Data Paper
Count of Subject Area and Citation by Subject Area
Count of Subject Area
Citation
Subject Area
Nuclear and High Energ...
Multidisciplinary
Chemistry (all)
0K
5K
10K
15K
20K
Count of Subject Area and Citation
Total Citation by Year
40K
20K
0K
Total Citation
1970
1980
1990
2000
2010
2020
Year

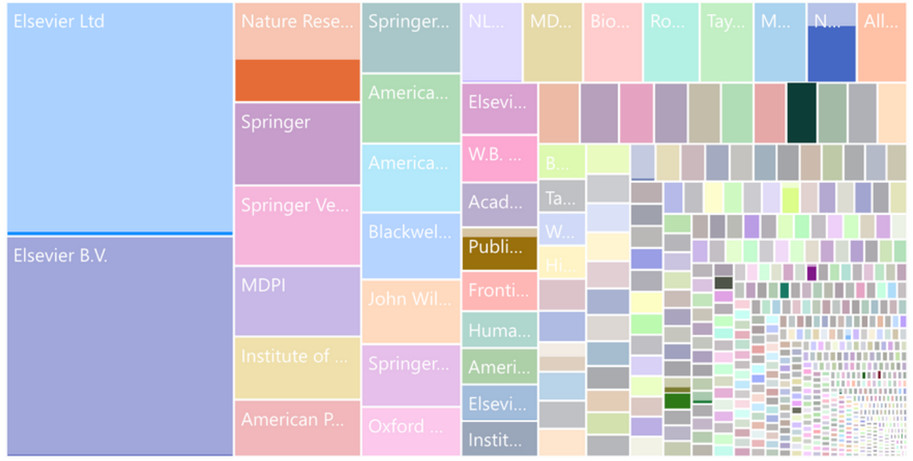
# Power BI

## Country and Subject Area

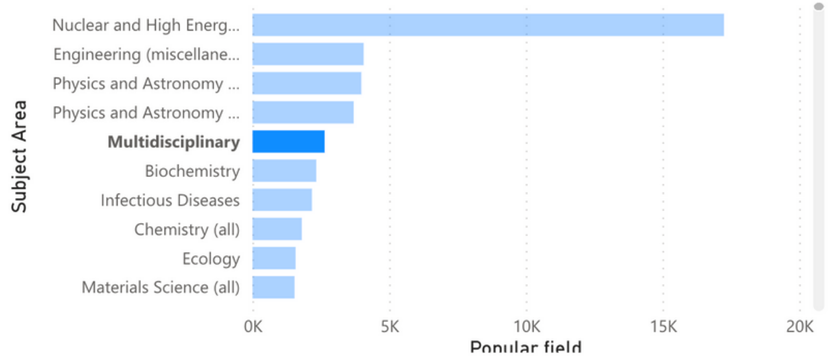
**Subject Area** Multidisciplinary




## Citations by Publisher




## Popular field by Subject Area



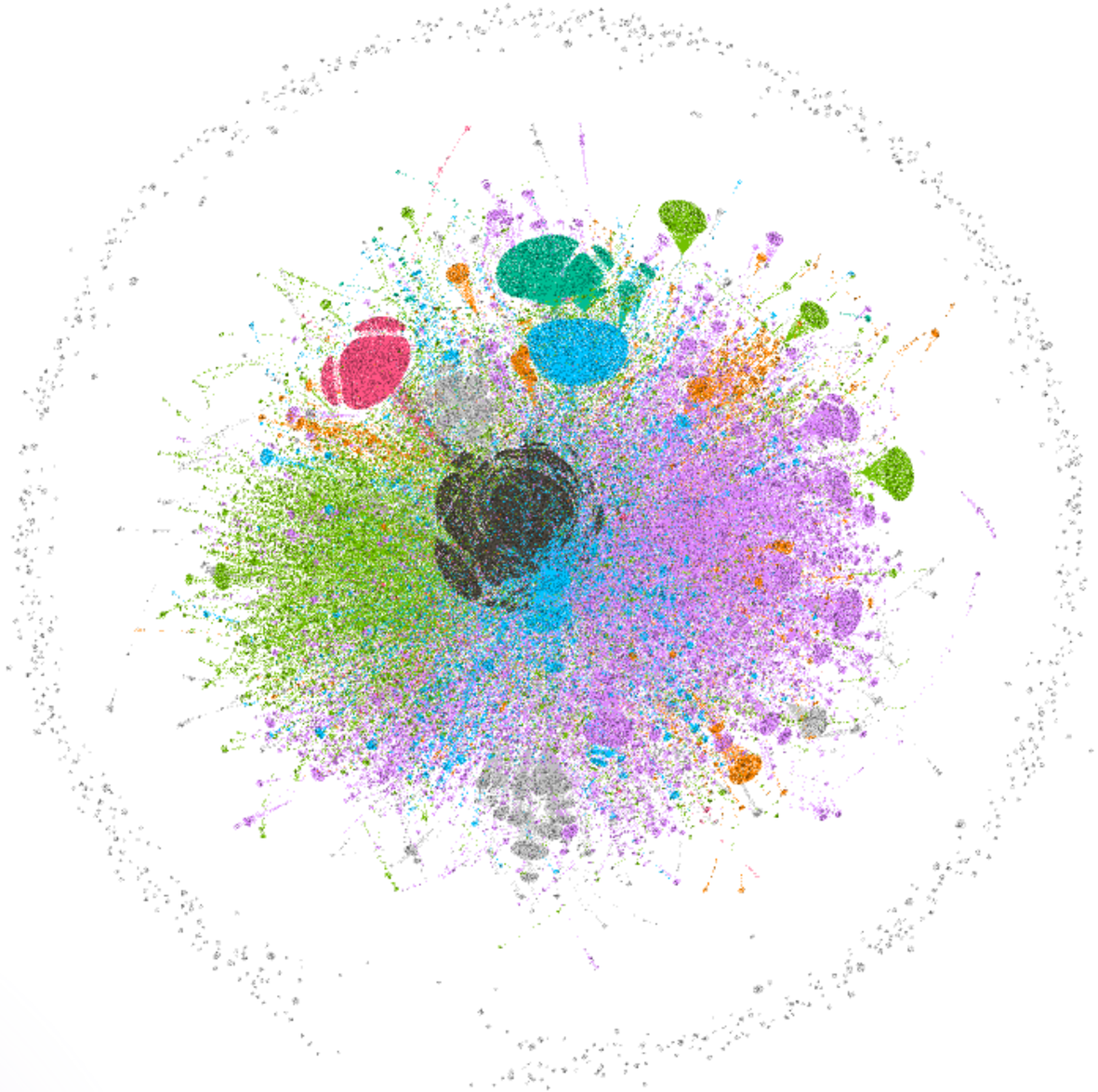

| First Subject_Area | Publisher |
|---|---|
| Multidisciplinary | Nature Research |
| Multidisciplinary | Nature Publishing GroupHoundmillsBasingstoke, HampshireRG21 6XS |
| Multidisciplinary | Public Library of Science |
| **Multidisciplinary** | |

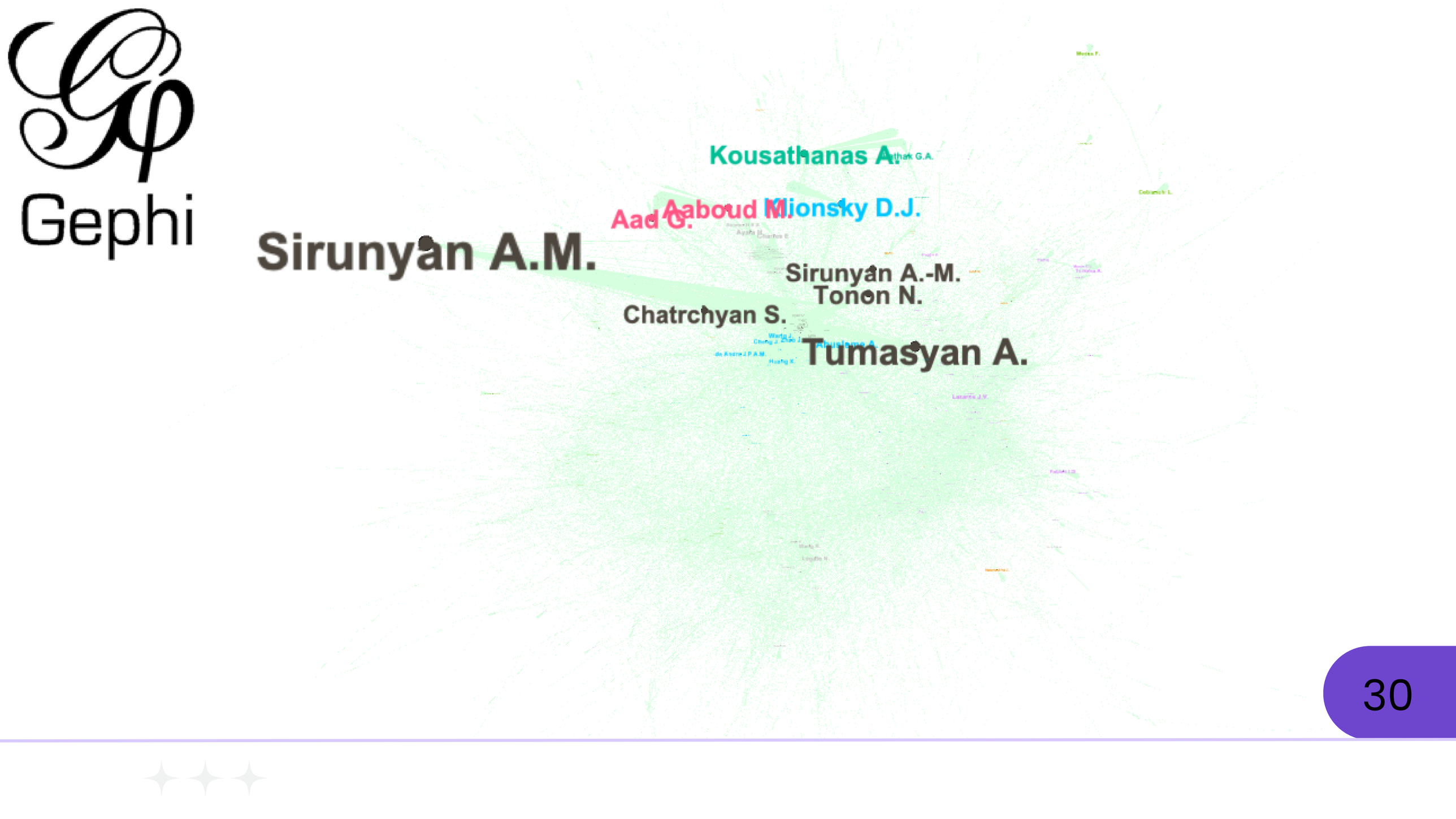
**Choose Subject Area**

- [ ] Accounting
- [ ] Acoustics and Ultrasonics
- [ ] Advanced and Specialized Nursing
- [ ] Aerospace Engineering
- [ ] Aging

1973 2023

Gephi




Gephi
Sirunyan A.M.
Kousathanas A.
Aad G.
Aaboud M.
Mionsky D.J.
Sirunyan A.-M.
Tonon N.
Chatrchyan S.
Tumasyan A.

Gephi

Google

Sirunyan A.M.

รูปภาพ | วิดีโอ | ช็อปปิ้ง | ข่าวสาร | Maps | หนังสือ | เที่ยวบิน | การเงิน

ตัวกรองทั้งหมด | เครื่องมือ | ฟีเจอร์ค้นหาปลอดภัย

ผลการค้นหาประมาณ 248,000 รายการ (0.50 วินาที)

sirunyan.am
http://www.sirunyan.am › ...

## Sirunyan.am

Welcome to Law Office “Sirunyan” – one of the market-leading legal services in Armenia. We provide a full range of counsel and legal services on various ...

ResearchGate
https://www.researchgate.net › scientific-contributions

## A. M. Sirunyan's research works | Yerevan Physics Institute ...

A. M. Sirunyan's 502 research works with 32229 citations and 10986 reads, including: Erratum to: Search for new physics in dijet angular distributions using ...

sirunyan.am
http://www.sirunyan.am

## Sirunyan.am

am ru en · Գրասենյակի մասին · Մենք TV-ում · ԶԼՄ ... www.sirunyan.am © 2014 Բոլոր իրավունքները պաշտպանված են

### ค้นรูป


### อัลเบิร์ต เอ็ม. ซีรุนยาน

นักวิจัย

**เกิด:** 13 มกราคม 2491

**เสียชีวิตเมื่อ:** 15 กันยายน 2561

**การศึกษา:** Yerevan State University (2513)

ข้อมูลเพิ่มเติมเกี่ยวกับ อัลเบิร์ต เอ็ม. ซี…

ความคิดเห็น

Top journals

Journal of High Energy Physics (233)
The European Physical Journal C (96)
Physics Letters B (72)
Physical Review Letters (48)
Physical Review D (16)

Affiliations

**Yerevan Physics Institute**

**Institut für Hochenergiephysik Wien**

**Imperial College London**

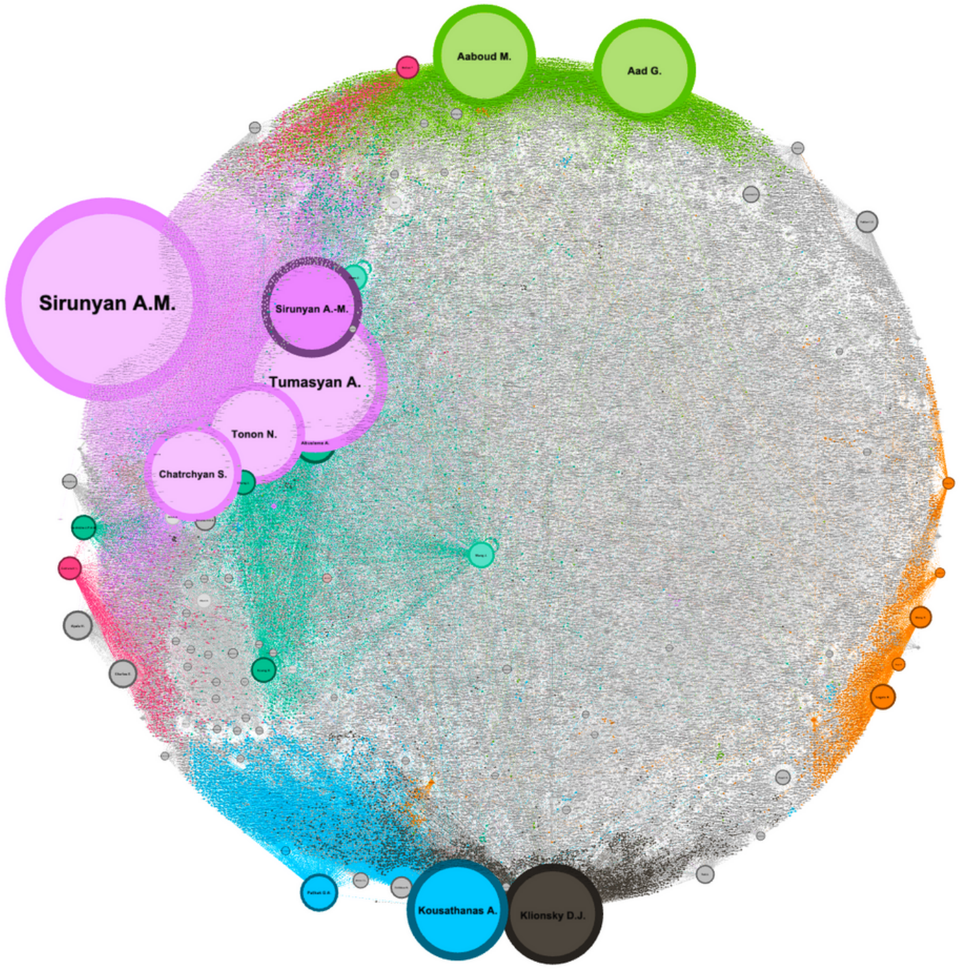
Disciplines

**Physics**
Topic
Experimental Physics · Theoretical Physics · Elementary Particle Physics · Nuclear Physics · Cosmology · Accelerator Physics · Mathematical Physics · Computational Physics

Publication Stats

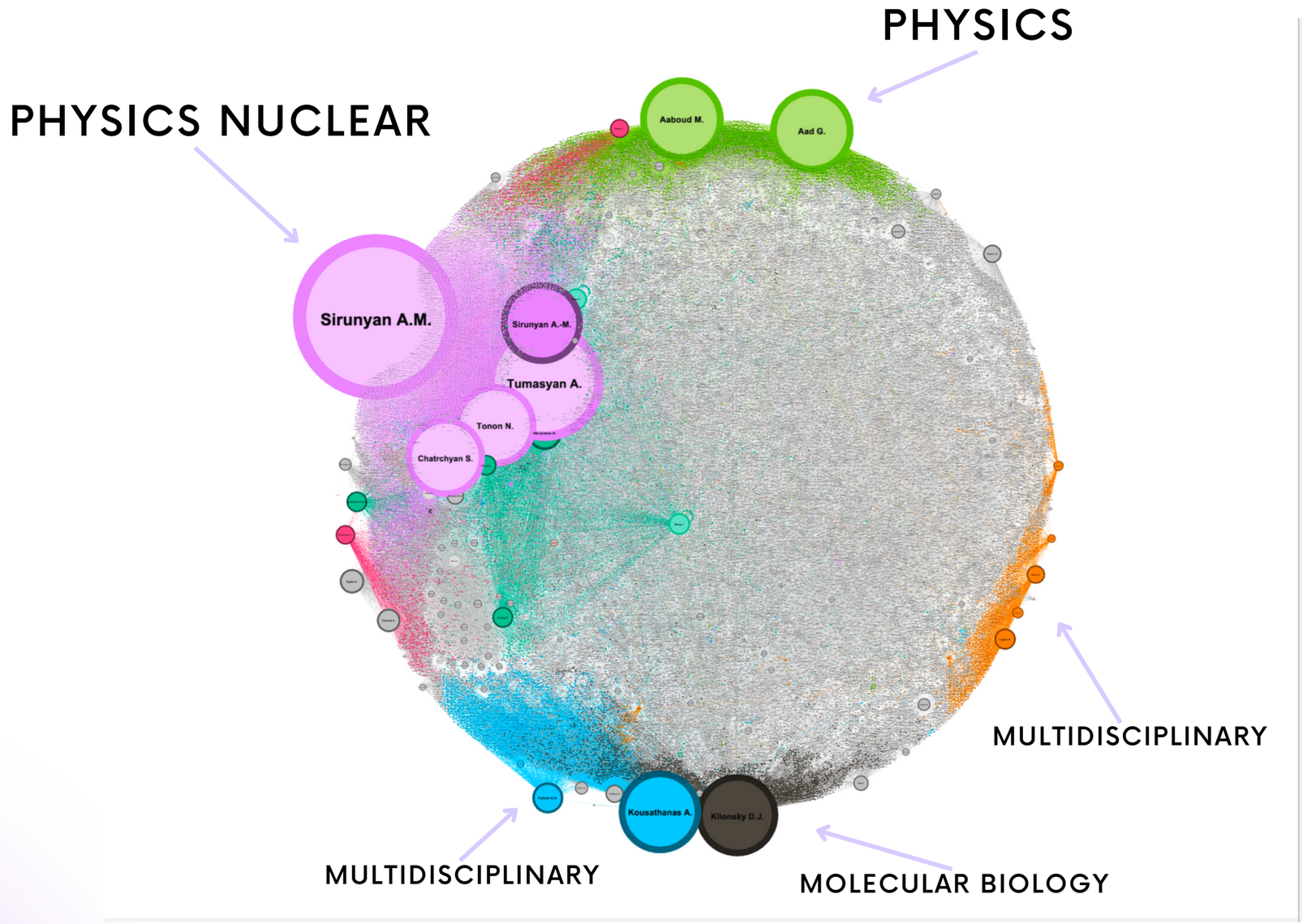
Citations 32,277

Gephi



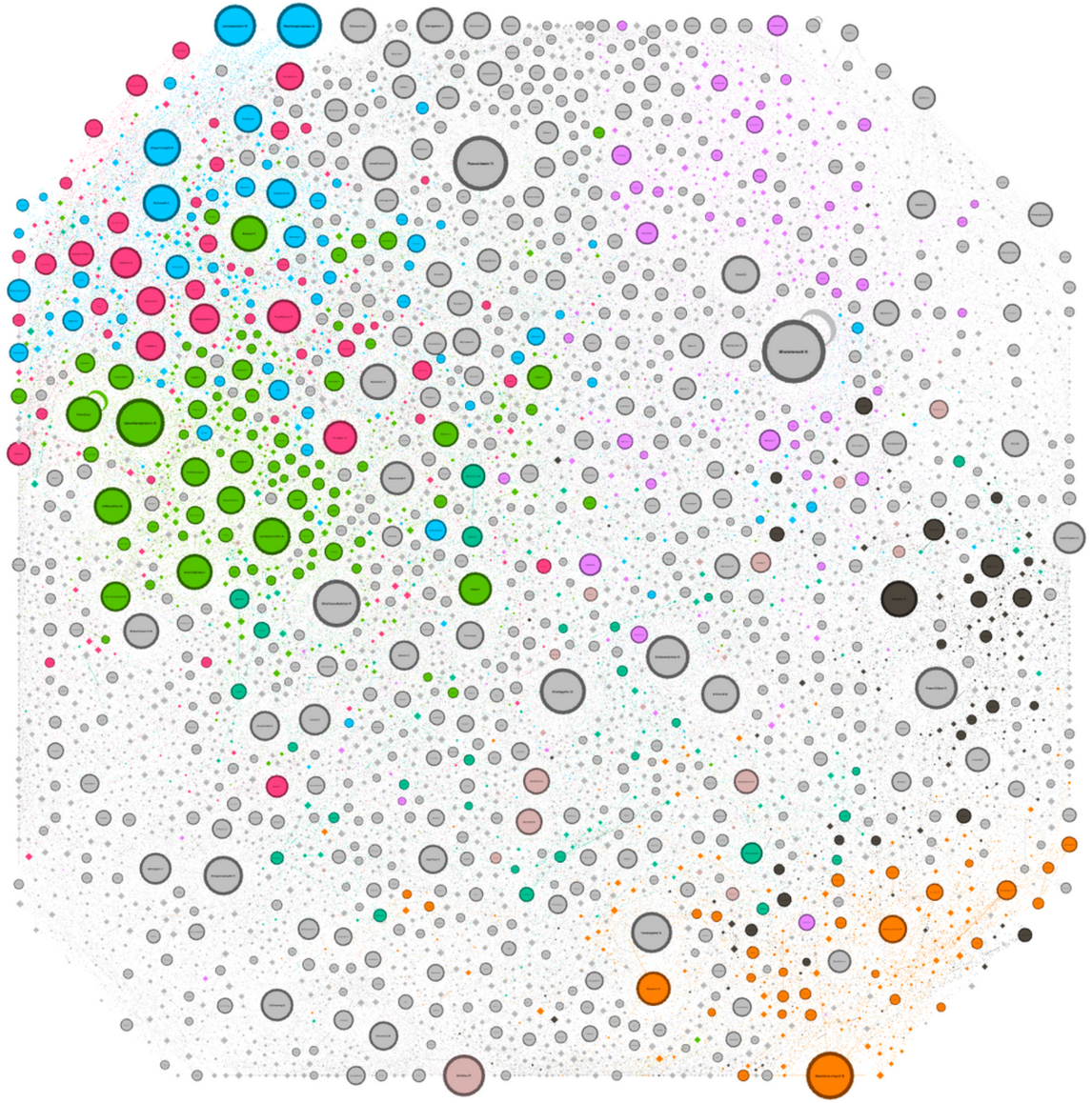

CHULALONGKORN CO-AUTHOR

Gephi




CHULALONGKORN CO-AUTHOR

Gephi

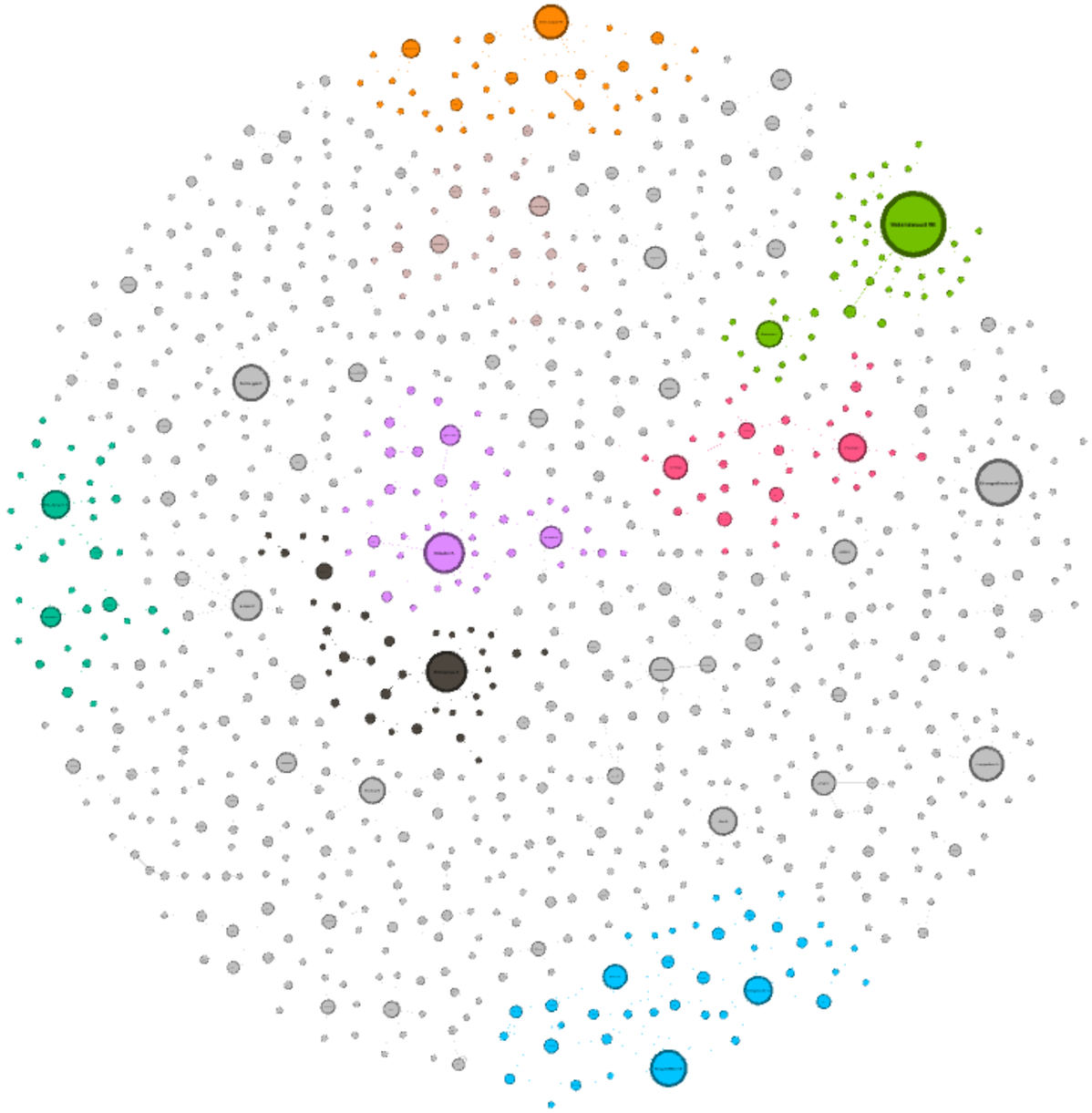


| | |
|---|---|
| Modularity | 0.881 |
| Statistical Inference | 239514.177 |



THAILAND CO-AUTHOR

Gephi

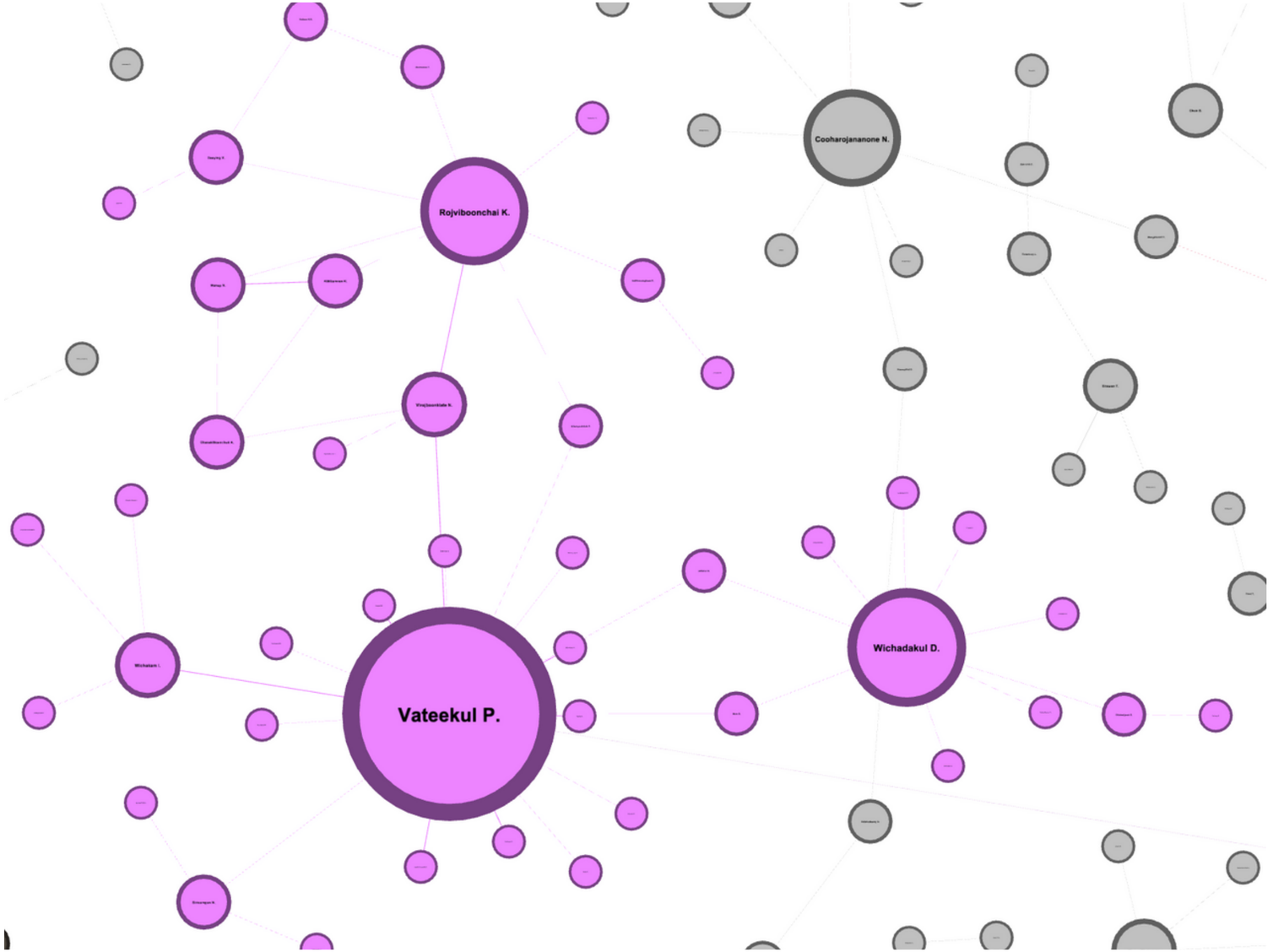


| | |
|---|---|
| Modularity | 0.963 |
| Statistical Inference | 7449.408 |



CHULALONGKORN CO-AUTHOR
COMPUTER

Gephi






CHULALONGKORN CO-AUTHOR
COMPUTER

# Thank You