Science of science研究会
チュートリアル

2024 03/16

東京大学工学系研究科 浅谷公威

ICSSI 2023

翻訳中です

監訳：三浦崇寛，神楽坂やちま，松井暉，浅谷公威，坂田一郎
（順序は適当です）

Science of science研究のデータ

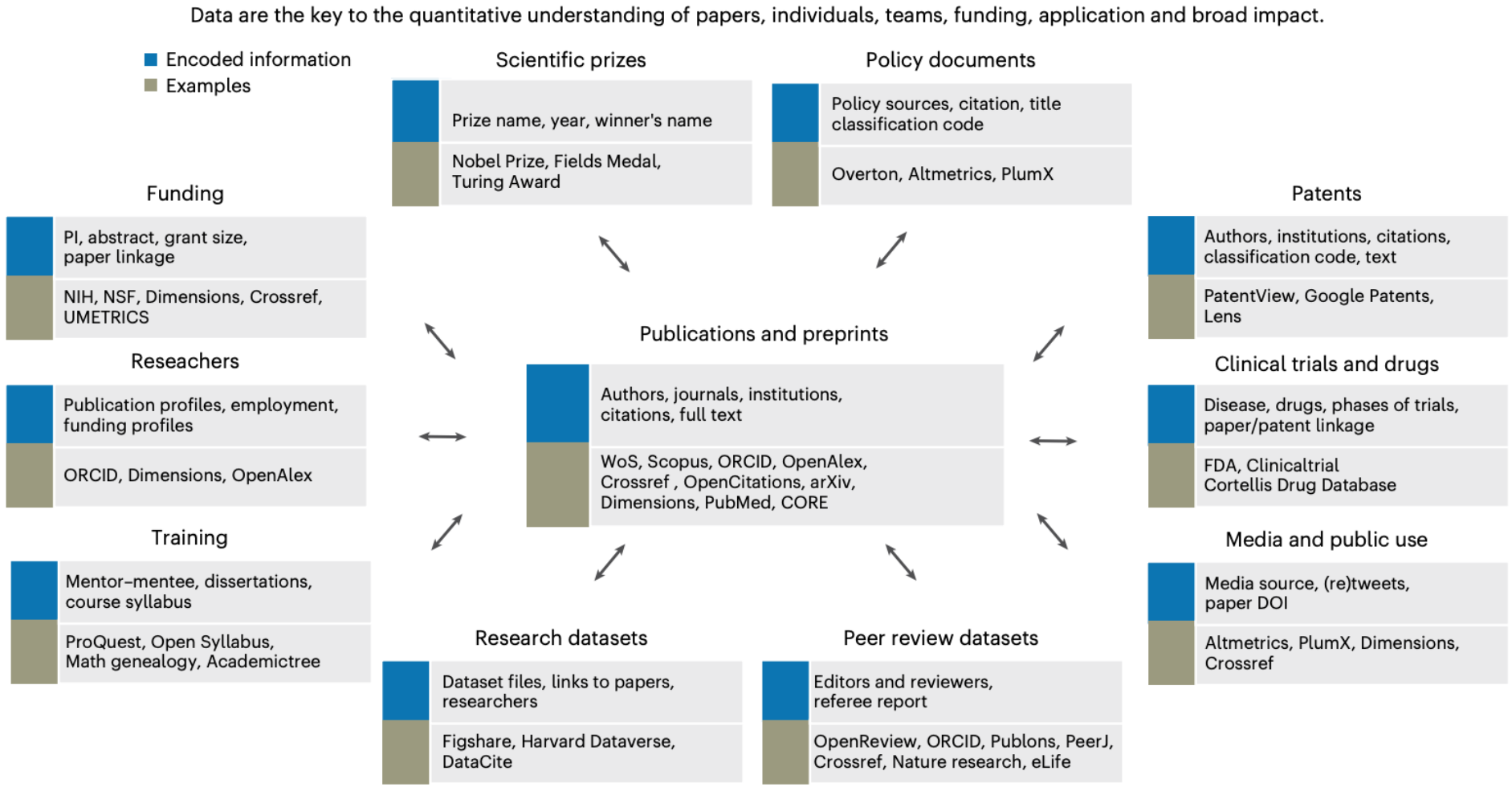


Fig. 1 | Science of science data and linkages. This figure presents commonly used data types in science of science research, information contained in each data type and examples of data sources. Datasets in the science of science research have not only grown in scale but have also expanded beyond publications to integrate upstream funding investments and downstream applications that extend beyond science itself.

Liu, L., Jones, B. F., Uzzi, B., & Wang, D. (2023). Data, measurement and empirical methods in the science of science. *Nature human behaviour*, *7*(7), 1046-1058.

研究に使われるデータセット

- オープンデータ
  - OpenAlex
    - よく使われている
    - 論文数も多い
  - Dimensions
    - こちらも使われる
- 商用データ
  - Web of Science
    - 精度が高い
  - Scopus
    - 名寄の精度が特に高い
    - 十分ではない
- その他
  - Twitter
  - Funding data
  - Overton
  - Award data

	Number of works	Open Access works	Citations	Price	Data Openness	Org structure
OpenAlex	240M	43.8M	1.9B	Freemium	Fully open, CC0 license	Non-profit
Scopus	87M	20.5M (ref)	1.8B	Subscription	Closed	For Profit
Web of Science (core)	87M (ref)	12M (ref)	1.8B	Subscription	Closed	For Profit
Dimensions	135M	29M (ref)	1.7B	Freemium	Partly open, personal use	For Profit
Google Scholar	389M (estimated)	?	?	Free	Closed	For Profit
Crossref	145M	20M	1.45B	Free	Fully open, CC0 license	Non-profit

https://www.igroupjapan.com/contents/openalex/

データの質

- Author Identification
  - これはどのデータベースでも十分でない

  ⇒著者の分析、共著分析、学術機関移動などで問題
  - Scopusが一番高い[1]
    - 6000人のランダムサンプルの著者について分析
    - 99%のPrecision, 94%以上のRecall

- 所属機関など
  - 結構表記揺れや階層の違いがある
    - University of Tokyo, Graduate school of Engineering, University of Tokyo

  ⇒それぞれの研究で補正を行っている

[1]Ioannidis, J. P., Baas, J., Klavans, R., & Boyack, K. W. (2019). A standardized citation metrics author database annotated for scientific field. *PLoS biology*, *17*(8), e3000384.

SciSciJP 2024 Tutorial

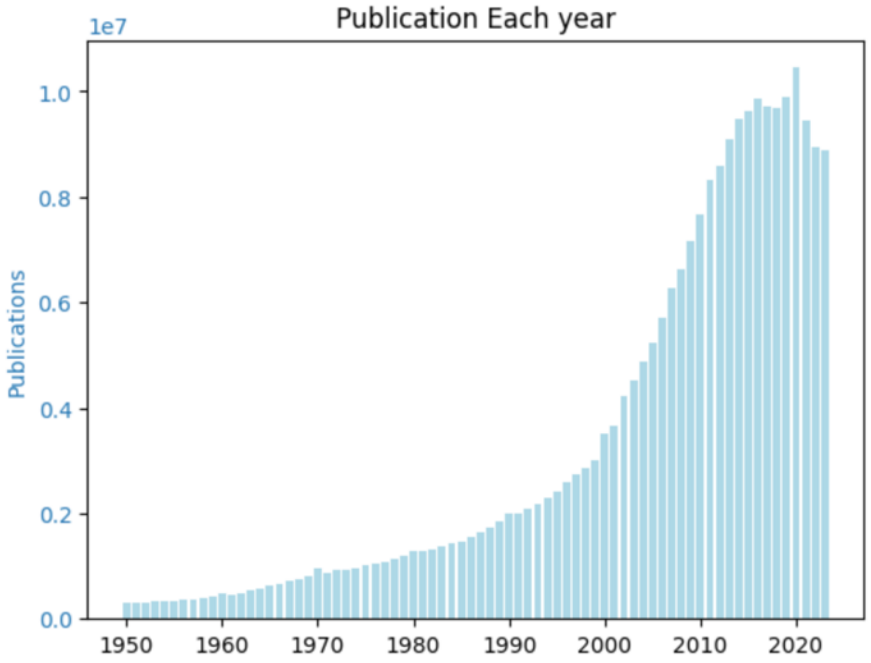
Scisci分野の論文およびコードのチュートリアルを、2024年の3/17開催のScience of Science研究会の前日イベントとして開催します。

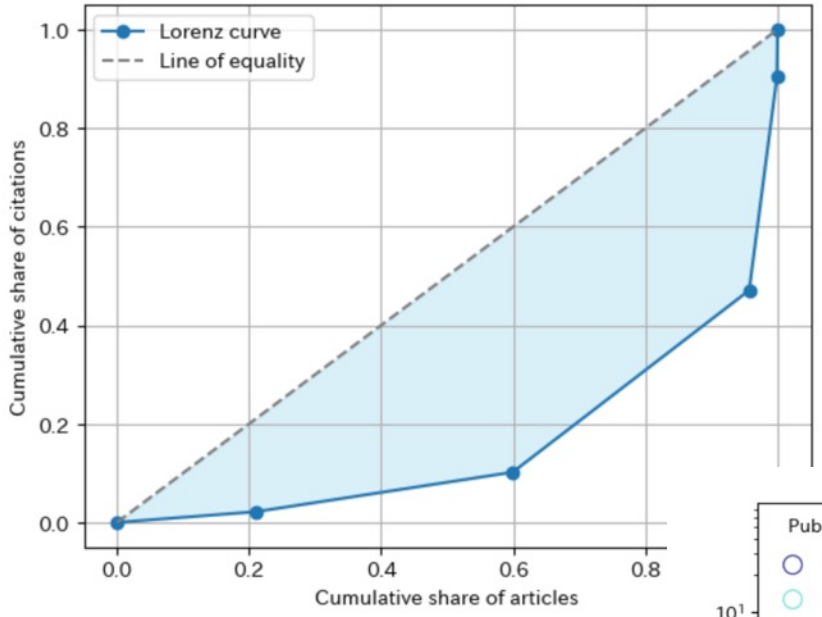
実際のデータを分析することで、SciSciの研究内容と着眼点や手法の限界について学びます。OpenAlexという無料のデータベースを利用し、Google Colabでコードを実装しますので、ブラウザがあればどのような環境でも動作します。

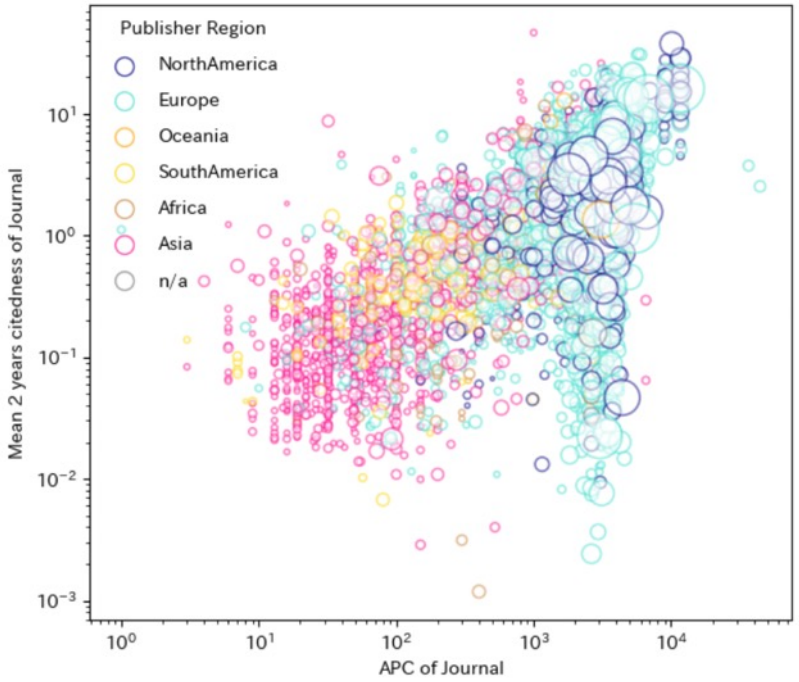
- 1. Getting Started:データへのアクセスと基礎分析
  - Science of science でどのようなデータを取り扱うのか、"Open Alex"というサービスを使っていくつか分析する中で確認しましょう。
  - colab Link
- 2. Visualizing Science:特定の学術分野の分類と可視化による理解
  - 特定の学術分野の論文を抽出し、引用関係を使って論文のネットワークを作り、論文が密に疎に繋がっている様子から分野を抽出できるか検証してみましょう。
  - 様々な分野で分析した結果をサンプルに格納しています。
  - colab Link
- 3. Research Evaluation by Disruptiveness index:トップ論文のD指標の再現
  - 研究者評価の指標として、科学におけるインパクトを定量化する手法の一つとして現在ホットな"Disruption Index"を学びます。対象論文の引用ネットワークにおける親と子を含めた3世代間の引用比で論文の革新性を表現した論文を再現してみましょう。
  - colab Link
- 4. Researcher Evaluation: 被引用による研究者の評価と将来予測、その他の諸々の自由タスク
  - 研究者の評価指標として、"h-index"を算出し、さらに個々の研究者のh-indexを他の変数から推定してみます。様々な交絡を考慮してより精度高く推定してみてください。
  - colab Link

※それぞれの章は概ね独立していますので、好きな章から試してください。

1. Getting Started:データへのアクセスと基礎分析







2. Visualizing Science:特定の学術分野の分類と可視化による理解

Cluster 3.0

Num paper: 1322

NameGPT: Gender & Peer Review

TFIDF: peer, review, gender, research, science, scientific, productivity, publication, reviewers, grant, journals, funding,

Average_year: 2018.1

Top Journals: Scientometrics ,PLOS ONE ,Journal of Informetrics

Top Authors: Flaminio Squazzoni ,Arturo Casadevall ,Francisco Grimaldo

Top Instititutions: Department of Microbiology and Immunology, University of Illinois at Chicago, Chicago, Illinois, USA ,Howard Hughes Medical Institute, Chevy

Top Cited Papers:

[2013] Nature (cited:969): Bibliometrics: Global gender disparities in science

[2012] Journal of the Association for Information Science and Technology (cited:623): Bias in peer review

[2013] PLOS ONE (cited:583): The Role of Gender in Scholarly Authorship

3. Research Evaluation by Disruptiveness index: D指標(Nature論文)の再現

nature

Explore content ∨ About the journal ∨ Publish with us ∨

nature › articles › article

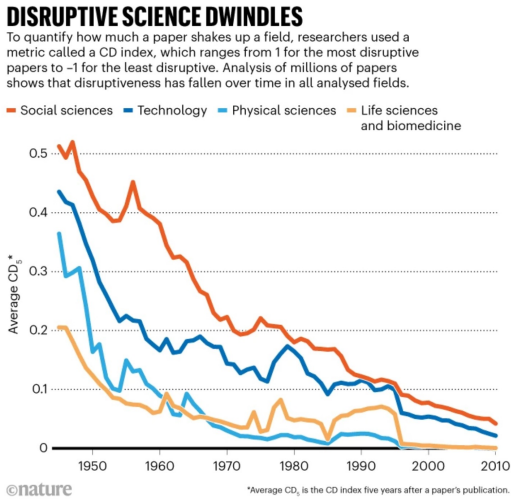
Article | Published: 04 January 2023

Papers and patents are becoming less disruptive over time

Michael Park, Erin Leahey & Russell J. Funk ✉

Nature **613**, 138–144 (2023) | Cite this article

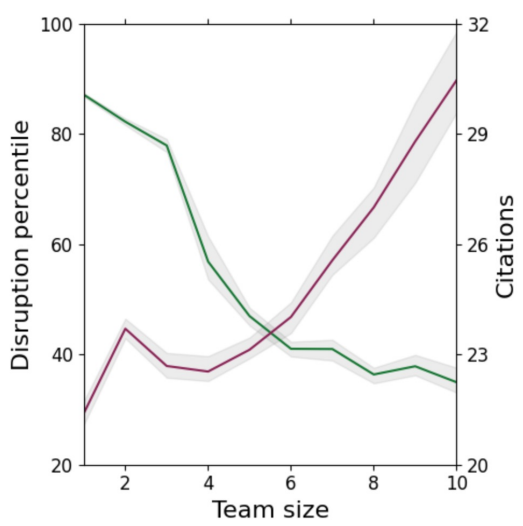
351k Accesses | **165** Citations | **4599** Altmetric | Metrics




```
host.plot(xdisPaper,ydisPaper, label="Disruption",color='#117733')
par1.plot(ximpPaper,yimpPaper, label="Impact",color='#882255')
host.fill_between(range(1,11), ydisPaperCIa,ydisPaperCIb,color='gray',alpha=0.15)
par1.fill_between(range(1,11), yimpPaperCIa,yimpPaperCIb,color='gray',alpha=0.15)

host.set_xlim(1, 10)
host.set_ylim(20,100)
host.set_yticks([20,40,60,80,100])
par1.set_ylim(20,32)
par1.set_yticks([20,23,26,29,32])
host.set_xlabel('Team size',size=16)
host.set_ylabel('Disruption percentile',size=16)
par1.set_ylabel('Citations',size=16)
host.tick_params(axis='both', which='major', labelsize=12)
par1.tick_params(axis='both', which='major', labelsize=12)

plt.tight_layout()
# 横軸が著者数、緑がDIのパーセンタイル、赤紫が被引用数、灰色が95%信頼区間
```



nature

Explore content ∨ About the journal ∨ Publish with us ∨

nature › letters › article

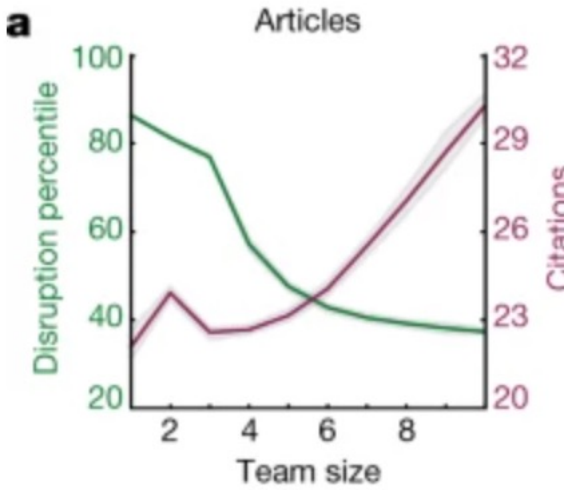
Letter | Published: 13 February 2019

Large teams develop and small teams disrupt scie and technology

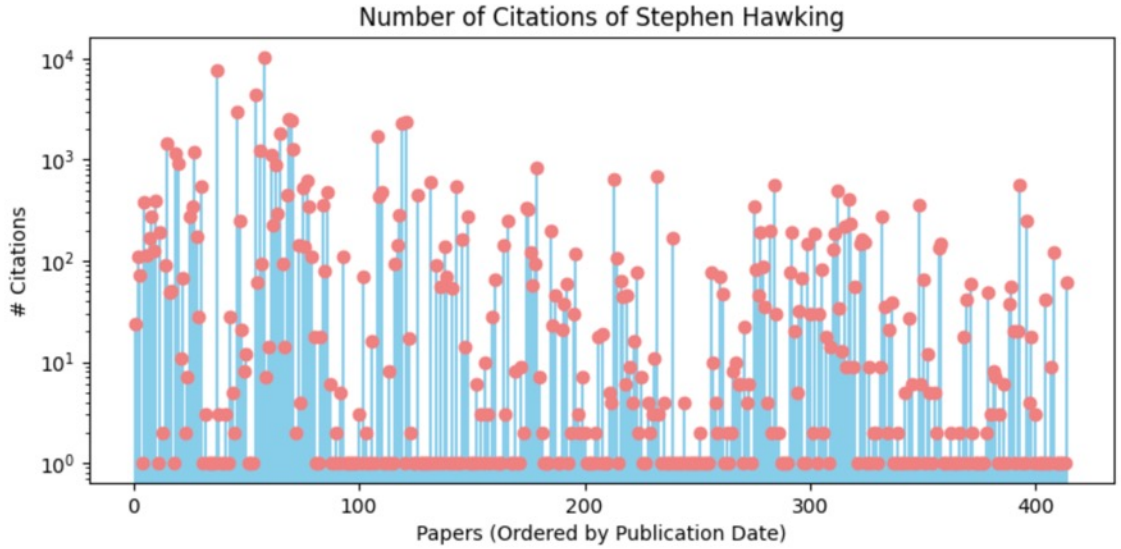
Lingfei Wu, Dashun Wang & James A. Evans ✉

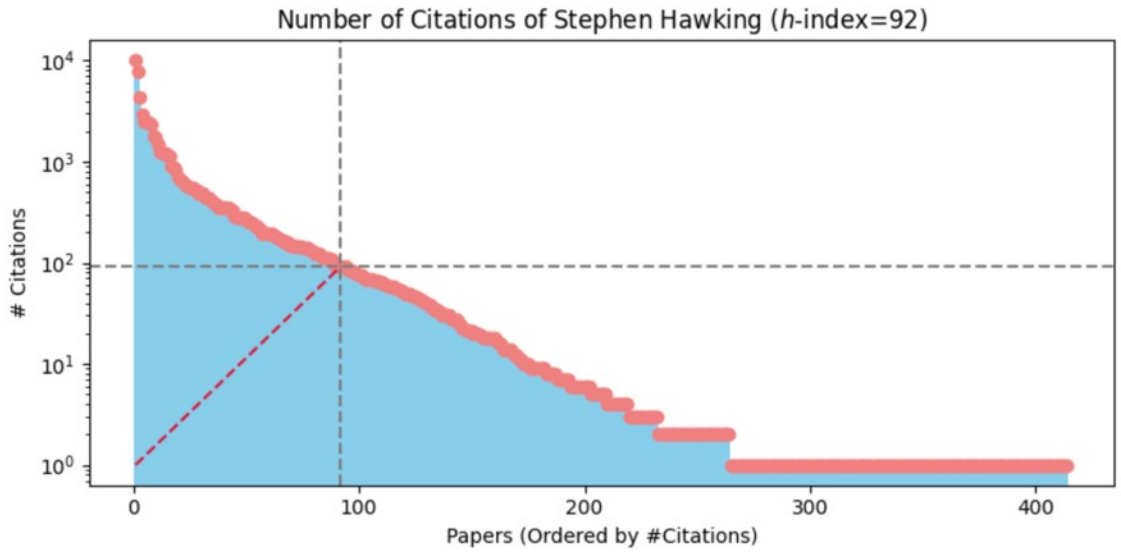
Nature **566**, 378–382 (2019) | Cite this article

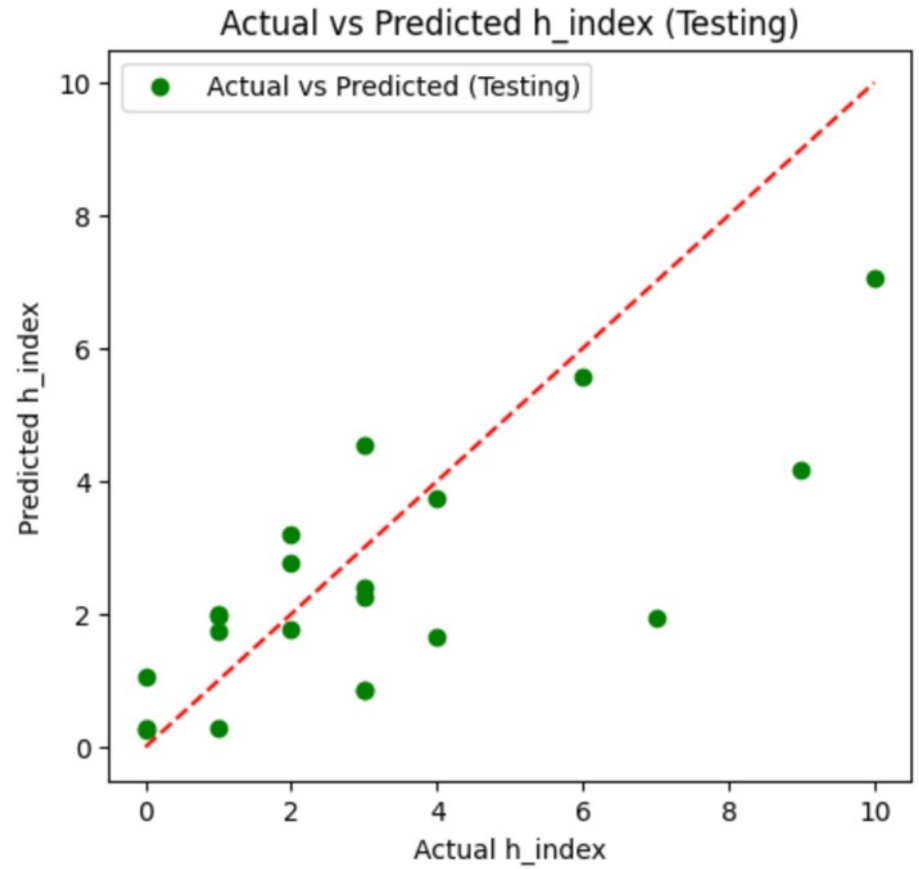
131k Accesses | **404** Citations | **2583** Altmetric | Metrics



4. Researcher Evaluation: 被引用による研究者の評価と将来予測、その他の諸々の自由タスク







チュートリアル委員

- 三浦 千哲 (東京大学)
- 神楽坂 やちま (東京大学)

コーディング

- 1章: 三浦 千哲 (東京大学)
- 2章: 浅谷 公威 (東京大学), 三浦 千哲 (東京大学), 原田 啓矢 (東京大学)
- 3章: 神楽坂 やちま (東京大学)
- 4章: 三浦 千哲(東京大学),神楽坂 やちま (東京大学), 王 思源 (東京大学)

チュートリアルの実施方法（ワンクリック）

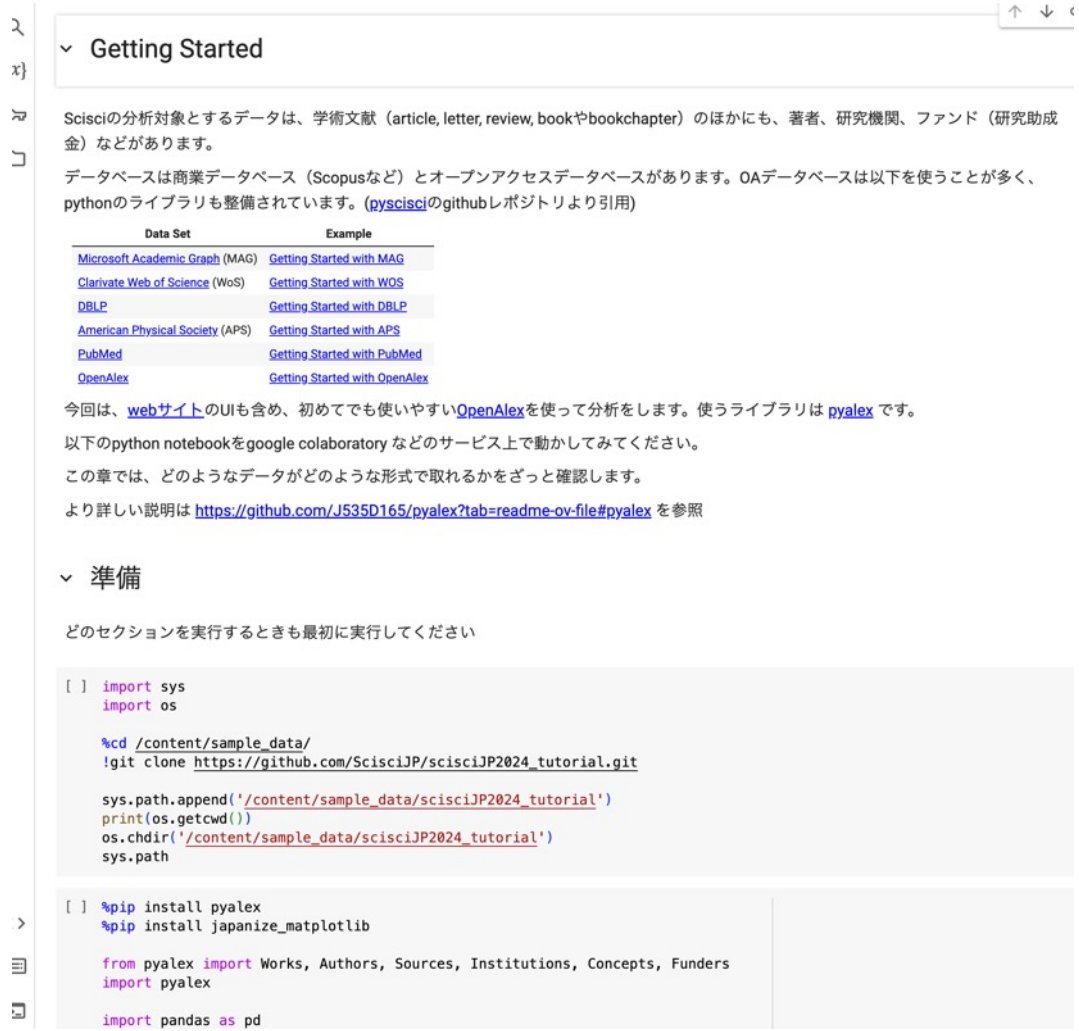
- https://github.com/ScisciJP/scisciJP2024_tutorial

ScisciJP 2024 Tutorial

Scisci分野の論文およびコードのチュートリアルを、2024年の3/17開催のScience of Science研究会の前日イベントとして開催します。

実際のデータを分析することで、SciSciの研究内容と着眼点や手法の限界について学びます。OpenAlexという無料のデータベースを利用し、Google Colabでコードを実装しますので、ブラウザがあればどのような環境でも動作します。

- 1. Getting Started:データへのアクセスと基礎分析
  - Science of science でどのようなデータを取り扱うのか、"Open Alex"というサービスを使っていくつか分析する中で確認しましょう。
  - colab Link クリック
- 2. Visualizing Science:特定の学術分野の分類と可視化による理解
  - 特定の学術分野の論文を抽出し、引用関係を使って論文のネットワークを作り、論文が密に疎に繋がっている様子から分野を抽出できるか検証してみましょう。
  - 様々な分野で分析した結果をサンプルに格納しています。
  - colab Link



Getting Started

Scisciの分析対象とするデータは、学術文献（article, letter, review, bookやbookchapter）のほかにも、著者、研究機関、ファンド（研究助成金）などがあります。

データベースは商業データベース（Scopusなど）とオープンアクセスデータベースがあります。OAデータベースは以下を使うことが多く、pythonのライブラリも整備されています。(pysciscのgithubレポジトリより引用)

Data Set	Example
Microsoft Academic Graph (MAG)	Getting Started with MAG
Clarivate Web of Science (WoS)	Getting Started with WOS
DBLP	Getting Started with DBLP
American Physical Society (APS)	Getting Started with APS
PubMed	Getting Started with PubMed
OpenAlex	Getting Started with OpenAlex

今回は、webサイトのUIも含め、初めてでも使いやすいOpenAlexを使って分析をします。使うライブラリは pyalex です。

以下のpython notebookをgoogle colaboratory などのサービス上で動かしてみてください。

この章では、どのようなデータがどのような形式で取れるかをざっと確認します。

より詳しい説明は https://github.com/J535D165/pyalex?tab=readme-ov-file#pyalex を参照

準備

どのセクションを実行するときも最初に実行してください

```
[ ] import sys
    import os

    %cd /content/sample_data/
    !git clone https://github.com/ScisciJP/scisciJP2024_tutorial.git

    sys.path.append('/content/sample_data/scisciJP2024_tutorial')
    print(os.getcwd())
    os.chdir('/content/sample_data/scisciJP2024_tutorial')
    sys.path
```

```
[ ] %pip install pyalex
    %pip install japanize_matplotlib

    from pyalex import Works, Authors, Sources, Institutions, Concepts, Funders
    import pyalex

    import pandas as pd
```