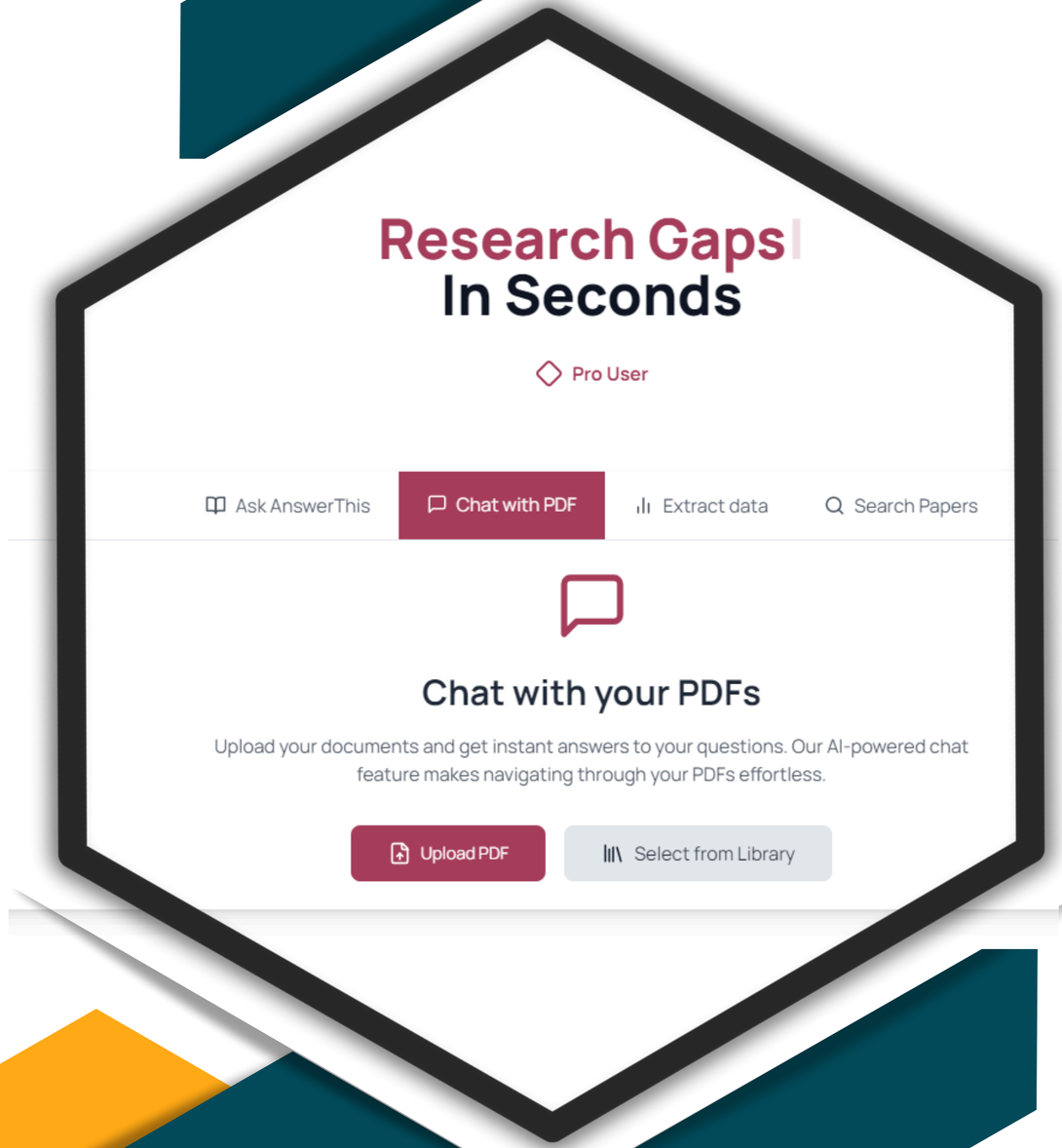
Fakultas Ilmu Terapan
School of Applied Science
Telkom University

idSPORA

by : Ko+ Lab

Membuat Literature Review dengan AI: Langkah Mudah & Efektif!



idSPORA

by : Ko+Lab

Dr. Heru Nugroho, S.Si., M.T.

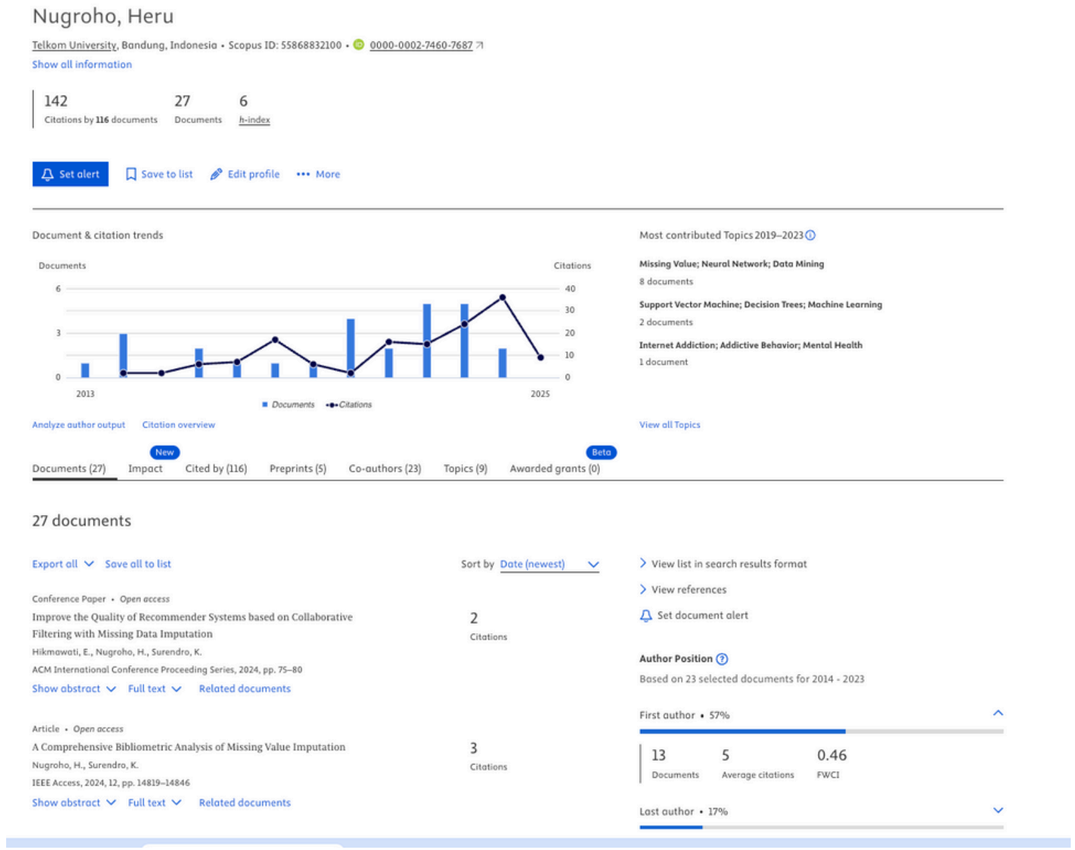
Head of Research & Development, Research Aliance Ko+Lab

S1: Matematika - UPI (2001 - 2005)
S2: Sistem Informasi - ITB (2011 - 2013)
S3: Teknik Informatika - ITB (2018 - 2023)

Publikasi

- 2024 : **A Comprehensive Bibliometric Analysis of Missing Value imputation, IEEE Access (Q1)**
- 2023:
  - Smoothing target encoding and class center-based firefly algorithm for handling missing values, Journal of Big Data (Q1)
  - **Trends in IT Strategy Implementation: A Systematic Review Across Education and Industry, Ingenierie des Systemes d'Information (Q4)**
  - Designing a Public API-Based Order Delivery Service System for the Food and Beverage Industry, Ingenierie des Systemes d'Information (Q4)
  - An ID3 decision tree algorithm-based model for predicting student performance using comprehensive student selection data at Telkom University, Ingenierie des Systemes d'Information (Q4)
- 2022: Data prediction for cases of incorrect data in multi-node electrocardiogram monitoring. International Journal of Electrical & Computer Engineering (Q2)
- 2021:
  - Normalization and outlier removal in class center-based firefly algorithm for missing value imputation, Journal of Big Data (Q1)
  - Class center-based firefly algorithm for handling missing data, Journal of Big Data (Q1)

Scopus and Google Scholar Profile



Heru Nugroho

MENGIKUTI

Telkom University

Email yang diverifikasi di telkomuniversity.ac.id - Beranda

Information System IT Governance Data Science Data Mining Data Analytics

Journals' rankings BETA

Report a bug

Q1	Q2	Q3	Q4	NA
7%	1%	7%	7%	78%

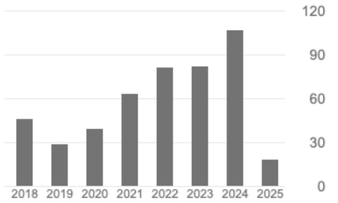
— based on 57 most cited papers in the last 10 years

ExCITATION beta

JUDUL	DIKUTIP OLEH	TAHUN
Conceptual Model of IT Governance for Higher Education Based on COBIT 5 Framework. H Nugroho 0.174 Q4 NA Journal of Theoretical & Applied Information Technology 60 (2)	87	2014
Normalization and outlier removal in class center-based firefly algorithm for missing value imputation H Nugroho, NP Utama, K Surendro 2.068 Q1 NA Journal of Big Data 8, 1-18	41	2021
Matematika SMP dan MTs Kelas VIII H Nugroho, L Meisaroh NA Pusat Perbukuan Departemen Pendidikan Nasional 53 (9)	40	2009
Missing data problem in predictive analytics H Nugroho, K Surendro Proceedings of the 2019 8th international conference on software and …	37	2019
Aplikasi Pendaftaran dan Transaksi Pasien Klinik Hewan di Bandung Berbasis Web N Rizkita, E Rosely, H Nugroho	33	2018

Dikutip oleh LIHAT SEMUA

	Semua	Sejak 2020
Kutipan	544	392
indeks-h	12	10
indeks-i10	14	10



Pengarang bersama EDIT

- Kridanto Surendro, Institut Teknologi Bandung, Indo…
- Nugraha Priya Utama, Institut Teknologi Bandung
- Robbi Hendriyanto, Telkom University
- LISDA MEISAROH, Universitas Telkom
- Agus Maolana Hidayat, Telkom University

Introduction

- Literature review adalah salah satu langkah terpenting dalam proses penelitian.
- Catatan tentang apa yang sudah diketahui/dilakukan tentang fenomena tertentu.
- Tujuannya untuk menyampaikan kepada pembaca tentang pekerjaan yang telah dilakukan dan pengetahuan serta ide-ide yang telah ditetapkan pada topik penelitian tertentu.

idSPORA

by : Ko+ Lab

Posisi Literature Review pada sebuah article?

...ssing value is a co...
...rate, or unreasonable conclu...
...When there is no data value for a variable ...
... missing. The phenomenon of these values is pervas...
large amounts of data (big data). Meanwhile, big data are extr...
pose storage and analysis challenges using conventional managem...

Springer Open

Currently, available analytical methods are only capable of working with complete data [12–15]. Therefore, missing value-related issues present the opportunities of obtaining the right technique as a solution [16].

In classification problems, missing values is a general weakness with the capacity to produce results of an ineffective prediction system [12, 17, 18]. Ignoring these data affects analysis [1, 8, 19–21] or learning outcomes, as well as prediction results on collaborative prediction problems [22]. Furthermore, it has the potential to weaken results and conclusion validities [3, 21]. In the predictive model, incorrect selection of the missing data handling method tends to affect the model's [8, 23] and classifiers' accuracy, as well as performance [24].

According to previous studies, feature normalization has an important effect on classi... ...cation accuracy [25–28]. Furthermore, in a dataset with numeric feature attributes, th... ...lization and processing of missing values are regarded as the main problem... ...g stage [29]. Also, a normalized mean interpolation method w... ...g value in numerical data sets [30]. Several studi... ...ious normalization techniques and s... ...erformance. However, o... ...plying data ...

Introduction

In most studies, missing value is a common and serious problem that often leads to biased, inaccurate, or unreasonable conclusions in cases of inappropriate handling [1–10]. When there is no data value for a variable in an observation, it is considered to be missing. The phenomenon of these values is pervasive in clinical studies involving large amounts of data (big data). Meanwhile, big data are extremely large datasets that pose storage and analysis challenges using conventional management techniques [11].

Currently, available analytical methods are only capable of working with complete data [12–15]. Therefore, missing value-related issues present the opportunities of obtaining the right technique as a solution [16].

In classification problems, missing values is a general weakness with the capacity to produce results of an ineffective prediction system [12, 17, 18]. Ignoring these data affects analysis [1, 8, 19–21] or learning outcomes, as well as prediction results on collaborative prediction problems [22]. Furthermore, it has the potential to weaken results and conclusion validities [3, 21]. In the predictive model, incorrect selection of the missing data handling method tends to affect the model's [8, 23] and classifiers' accuracy, as well as performance [24].

According to previous studies, feature normalization has an important effect on classification accuracy [25–28]. Furthermore, in a dataset with numeric feature attributes, the normalization and processing of missing values are regarded as the main problems in the pre-processing stage [29]. Also, a normalized mean interpolation method was developed to solve the missing value in numerical data sets [30]. Several studies have separately analyzed the effects of various normalization techniques and strategies for dealing with missing value on classification performance. However, only a few rated the effects combining the two [29]. Furthermore, applying data normalization has a significant effect on classification and greatly improves the performance of the KNN imputation method [31]. Previous studies have also shown that combining normalization and imputation using the mean is more accurate than traditional mean and median methods [30].

idSPORA

by : Ko+ Lab

Literature Review in Article – in Introduction

Related work

The three mandatory technical issues that should be considered in the process of inputting missing data, selecting experimental data sets, methods, and evaluating imputation results are shown in Fig. 1 [44].

The choice of experimental data set is related to the problem area, the filling and the type of test data, missing data mechanism (MCAR, MAR, MNAR) as well as a percentage (missing rate). According to Lin and Tsai (2020), the normalization and outlier detection's consideration was not discussed in the review paper "Missing value imputation: a review and analysis of the literature (2006–2017)". The effect of normalization and various techniques for handling the missing value strategy on classification performance separately was extensively conducted in previous studies. However, only a few assessed the simultaneous combination effect of standardization and missing data handling methods [29]. Some also showed that combining normalization and imputation techniques produced better accuracy values [30, 31, 45].

In addition to normalization of pre-processing, outliers significantly influence the statistical estimation process (for instance, the sample mean and standard deviation), resulting in either excessively high or low values [46]. Several missing data imputation methods including mean, linear regression, multiple, and class center-based, utilize the mean value. Generally, the training data contains noisy data or outliers with the ability to reduce the learning model's final performance [33, 34]. Therefore, it is necessary to select instances in the observed data set for missing values imputation and to determine the selection performance of instances from the observed data set before the imputation [32]. According to other studies, outliers play an important role in the imputation method's performance. In cases where a dataset contains these data points, mixed models with high flexibility can produce deviations from the true data pattern [47].

It was also reported that imputation results were strongly influenced by the presence of outliers [35–37]. Therefore, outlier handling should be conducted before imputation [36, 37]. Currently, the classical method is unable to perform imputation accurately in

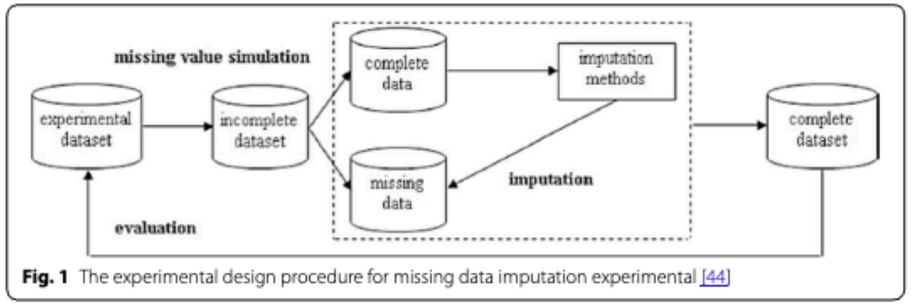


Fig. 1 The experimental design procedure for missing data imputation experimental [44]

idSPORA

by : Ko+ Lab

Literature Review in Article - in Related Work

IEEE Access

Received 21 November 2023, accepted 12 January 2024, date of publication 23 January 2024, date of current version 1 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3357533

TOPICAL REVIEW

A Comprehensive Bibliometric Analysis of Missing Value Imputation

HERU NUGROHO[1,2] AND KRIDANTO SURENDRO[1], (Member, IEEE)

[1]School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Bandung 40132, Indonesia
[2]School of Applied Science, Telkom University, Bandung 40257, Indonesia

Corresponding author: Heru Nugroho (heru@tass.telkomuniversity.ac.id)

This work was supported by the Institut Teknologi Bandung (ITB).

ABSTRACT Data quality plays a crucial role in tasks, such as enhancing the accuracy of data analytics and avoiding the accumulation of redundant data. One of the significant challenges in data quality is dealing with missing data, which has been extensively explored by the scholarly community and has resulted in a significant increase in related publications. It is important to recognize that the landscape of missing data in computer science offers numerous opportunities for further research. However, upon closer examination of existing studies, it becomes evident that many have not fully utilized bibliometric analysis tools and software for comprehensive literature reviews. Therefore, this study aims to explore the essential characteristics, trends, and prevailing themes in the field of missing data imputation. Through a thorough bibliometric analysis, this study demonstrated the evolution of knowledge and key focal points in the field of missing data imputation. The analysis consisted of 352 journal papers in computer science published between 2012 and 2023, all centered on missing data imputation. Among these publications, "IEEE Access" has become a highly respected source. To systematically explore various aspects of missing data imputation, a conceptual framework was used to uncover potential research directions and underlying themes. Ultimately, a thematic map serves as a valuable tool for providing a comprehensive understanding, categorizing significant concepts into basic or overarching, developing, or declining, central, highly developed, and isolated themes. These overarching and underlying themes offer valuable insights and pave the way for prospective directions and critical areas of study.

INDEX TERMS Missing data, imputation, literature review, bibliometric analysis.

I. INTRODUCTION

Data is an essential asset in manufacturing, and a well-executed data-mining strategy increases productivity [1], [2]. Various factors lead to missing values in databases, which adversely affects the quality of algorithms used in data mining [1], [3]. In real-world datasets, the presence of missing values is a significant challenge, impairing data analytics [4], [5], [6], impeding efficient data use, and diminishing the effectiveness of data-driven models [7]. Learning from data with missing values is a widespread issue across disciplines [8]. It is crucial to acknowledge that missing values introduce challenges when identifying data patterns such as clustering and classification [9]. In addition, missing values in tabular data negatively affect the application and efficacy of machine learning, necessitating imputation [10]. While large datasets may allow for value disregarding, removing records with missing values in smaller datasets can lead to inaccurate classification or predictions using DM algorithms [1]. Addressing missing values during data observation or recording is a significant concern [11], [12].

Current data imputation methods are primarily created for specific types of missing values, such as Missing Not at Random (MNAR) or Missing Completely at Random (MCAR) [13]. The extent of missing data, the mechanism causing missing values, and the chosen handling method affect statistical inference [14]. The main suggested technique for resolving the problems caused by incomplete datasets is

The associate editor coordinating the review of this manuscript and approving it for publication was Walter Didimo.

IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 28, No. 6, December, 2023, pp. 1643-1651
Journal homepage: http://lieta.org/journals/isi

Trends in IT Strategy Implementation: A Systematic Review Across Education and Industry (2000–2022)

Varuliantor Dear[1,2], Nandang Dedi[2], Annis Siradj Mardiani[2], Heru Nugroho[3*]

[1] School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Bandung 40116, Indonesia
[2] Space Research Center, National Research and Innovation Agency, Bandung 40135, Indonesia
[3] School of Applied Science, Telkom University, Bandung 40257, Indonesia

Corresponding Author Email: heru@tass.telkomuniversity.ac.id

https://doi.org/10.18280/isi.280622

Received: 9 September 2023
Revised: 29 November 2023
Accepted: 5 December 2023
Available online: 23 December 2023

Keywords:
IT strategy, implementation, education, industry, bibliometric

ABSTRACT

The rapid and global development of digital technology or digital transformation has encouraged various business sectors to adapt to the field of information technology. The education and industrial sectors are interrelated and need to adapt to developments in digital technology where the eduacton sector is the supplieer of human resources and the industrial sector is one of the purposes of the education sector. The balancing prepareness of thos both sector on IT starategy whoulld be conducted for the optimu[illegible] process otuput. One of the indicators of the digital transformation adaptation p[illegible] the information technology (IT) strategy plan. In this p[illegible] carried out from database Scopus on IT strategy implementation in the industry and education sectors published within the last two decades to get an overview of the responses of these two sectors to the dynamics of digital transformation. Bibliometric analysis" refers to the quantitative assessment of scholarly publications and research a[illegible] specific field or discipline. The analysis is based on the papers' growt[illegible] the thematic maps' evolution over each decade. The compilation comprises a total of 41 journals and 23 proceedings. To enhance our comprehension of thematic evolution, the two articles have been categorized into distinct decade periods: 2000–2010 and 2010–2022. The results show that the industrial sector has more publications with an earlier productivity peak than the education sector. The peak productivity of paper in the two sectors occurred before the COVID-19 pandemic. The productivity rate of papers during and after the COVID-19 pandemic was at a reasonably low value, which can be interpreted as an indicator of readiness for the pandemic events that occurred and their effects. The distribution of the thematic maps of the two sectors is different, with the industrial sector having more variables than the education sector. Industrial thematic objects are scattered in all quadrants, while the education sector has been concentrated in quadrants 2 and 3 in the last decade. The thematic objective's distribution indicates the dynamics of the challenges in implementing the information technology strategy for both sectors over the next two decades.

1. INTRODUCTION

Industry must overcome several technical issues to deliver renewable energy in significant quantities. Control is one of the key enabling technologies for the deployment of renewable energy systems. Solar and wind power require the effective use of advanced control techniques. In addition, smart grids cannot be achieved without extensive use of control technologies at all levels.

The rapid development of digital technology demands an adaptation of digital transformation in various business sectors [1-7]. This needs to be done to maintain the continuity of business processes and survive the dynamics of changes that occur. Furthermore, the COVID-19 pandemic phase, which had an impact on drastic and global changes in various activities, forced the adaptation process to the development of digital technology to be carried out more quickly [8-10]. During the COVID-19 pandemic phases, all sectors were forced to adopt digital transformation to maintain their business process due to the limitation of physical interaction.

The education sector and the industrial sector have a reasonably close relationship. The education sector is the central pillar that produces the human resources that the industrial sector needs [11]. For the education sector to produce adequate human resources in the industrial sector, adaptation to changes in information technology is a must, both in business processes and in the substance of education [12]. The inability to adapt to the developments in information technology implies that the output products from the education sector do not meet the qualifications needed in the industrial sector. On the other hand, the industrial sector cannot adapt to developments in information technology, causing business

Apa itu **Literature Review?**

- Rangkuman dari beberapa penelitian yang terkait dengan topik penelitian
- Kajian yang relevan secara mendalam dan kritis tentang topik tertentu
- Kritik (membangun maupun menjatuhkan) dari penelitian yang sedang dilakukan terhadap topik khusus atau pertanyaan terhadap suatu bagian dari keilmuan.

idSPORA

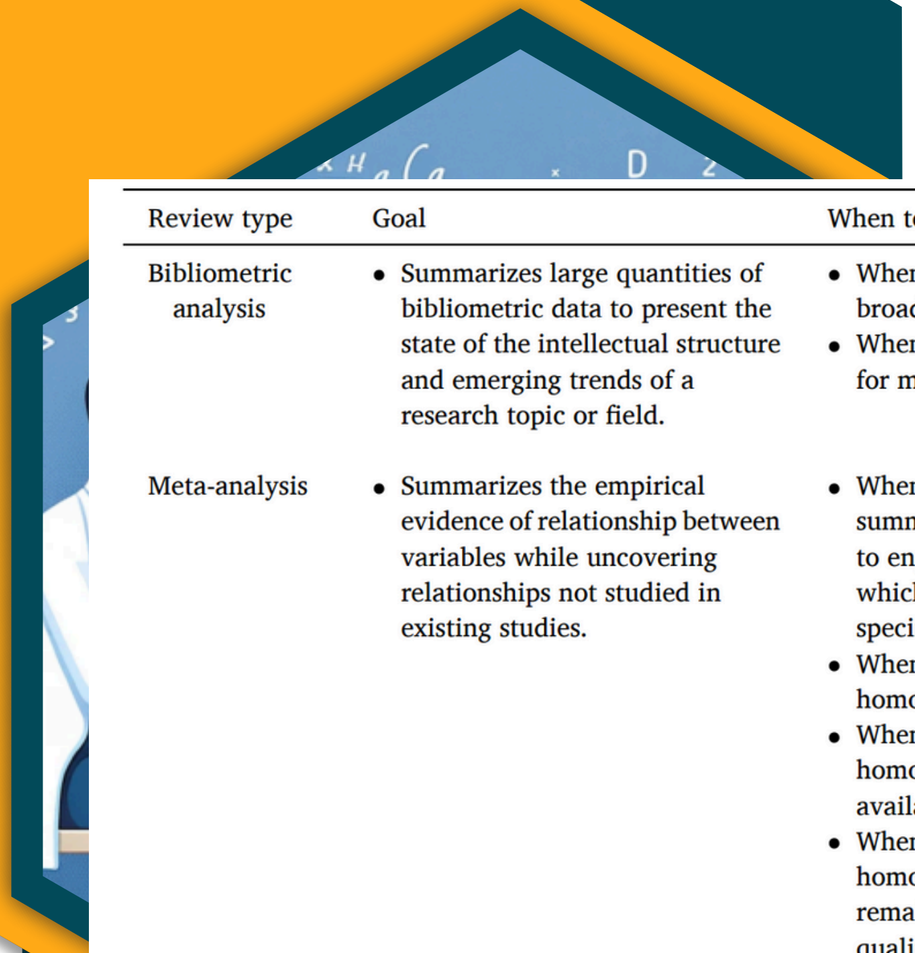
by : Ko+ Lab

Tujuan Literature Review

Sebuah literature review bukan sekadar daftar makalah. Literatur review yang baik harus:

- Mengidentifikasi kesenjangan penelitian dalam bidang yang dikaji.
- Menempatkan riset dalam konteks diskusi ilmiah yang lebih luas.
- Menunjukkan bagaimana penelitian yg dilakukan membangun dan berkontribusi pada pengetahuan yang sudah ada.

Metode Literature Review



Review type	Goal	When to use	When not to use	Scope	Dataset	Analysis
Bibliometric analysis	• Summarizes large quantities of bibliometric data to present the state of the intellectual structure and emerging trends of a research topic or field.	• When the scope of review is broad. • When the dataset is too large for manual review.	• When the scope of review is specific. • When the dataset is small and manageable enough that its content can be manually reviewed.	• Broad	• Large	• Quantitative (evaluation and interpretation) • Qualitative (interpretation only)
Meta-analysis	• Summarizes the empirical evidence of relationship between variables while uncovering relationships not studied in existing studies.	• When the focus of review is to summarize results rather than to engage with content, which may be broad or specific. • When studies in the field are homogenous. • When the number of homogeneous studies available is sufficiently high. • When the number of homogeneous studies remaining after removing low quality studies is sufficiently high.	• When studies in the field are heterogeneous. • When the number of homogenous studies is relatively low. • When the number of high-quality homogeneous studies is relatively low.	• Broad • Specific	• Large • Small but adequate	• Quantitative (evaluation and interpretation)
Systematic literature review	• Summarizes and synthesizes the findings of existing literature on a research topic or field.	• When the scope of review is specific. • When the dataset is small and manageable enough that its content can be manually reviewed.	• When the scope of review is broad. • When the dataset is too large for manual review.	• Specific	• Small	• Qualitative (evaluation and interpretation)

https://www.sciencedirect.com/science/article/pii/S0148296321003155#b0275

idSPORA

by : Ko+ Lab

Perbandingan Metode LR

- Bibliometric Analysis
  - Cocok untuk melihat gambaran umum dan tren dalam suatu bidang penelitian dengan menggunakan data kuantitatif dan kualitatif.
- Meta-analysis
  - Lebih fokus pada analisis statistik gabungan dari beberapa studi untuk mendapatkan kesimpulan yang lebih kuat tentang hubungan variabel tertentu.
- Systematic Literature Review
  - Metode yang sistematis untuk mengumpulkan dan mengevaluasi semua literatur yang relevan pada suatu topik tertentu, biasanya dengan pendekatan kualitatif.

idSPORA

by : Ko+ Lab

Perkembangan AI dalam Riset Akademik

- Dalam dua tahun terakhir, banyak AI tools berkembang untuk membantu peneliti.
- Tools seperti Research Pal, Perplexity, dan Jaspar AI menjadi populer.
- Namun, Elicit dan AnswerThis muncul sebagai inovasi terbaru yang mengubah cara riset dilakukan.

idSPORA

by : Ko+ Lab

Researchers Get Wrong

- **Meringkas tanpa sintesis** → Literatur review bukan sekadar ringkasan, tetapi harus menghubungkan temuan dari berbagai studi.
- **Mengandalkan sumber usang** → Pastikan menggunakan referensi terbaru, terutama di bidang yang berkembang pesat.

idSPORA

by : Ko+ Lab

How to Structure Your Review

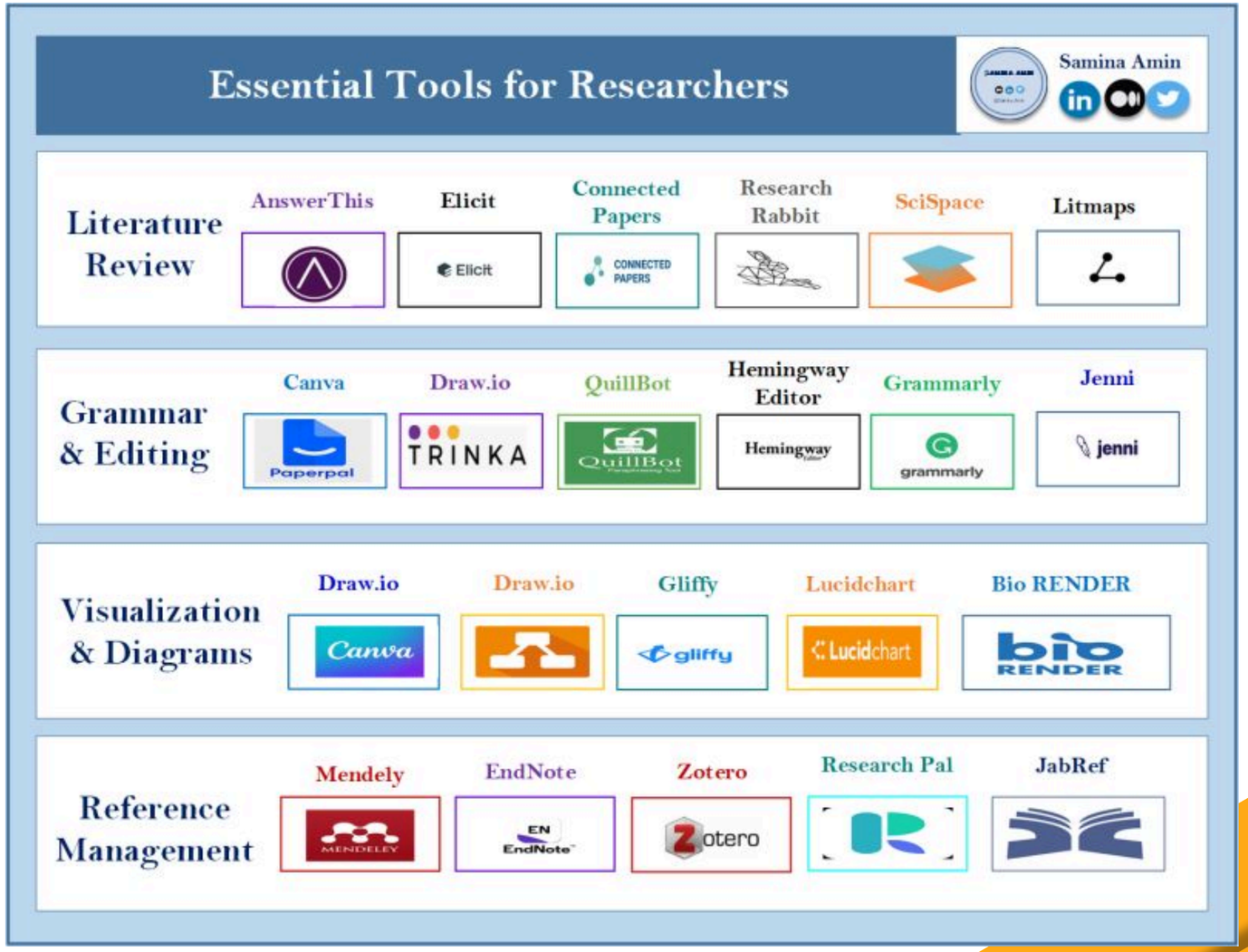
- **Introduction**: Definisikan fokus penelitian Anda dan jelaskan mengapa topik ini penting.
- **Thematic Mapping**: Kelompokkan studi berdasarkan tema, bukan hanya urutan waktu.
- **Critical Analysis**: Soroti kelebihan, kekurangan, dan kontradiksi dalam literatur yang ada.
- **Conclusion**: Tunjukkan kesenjangan penelitian yang ingin Anda jawab.

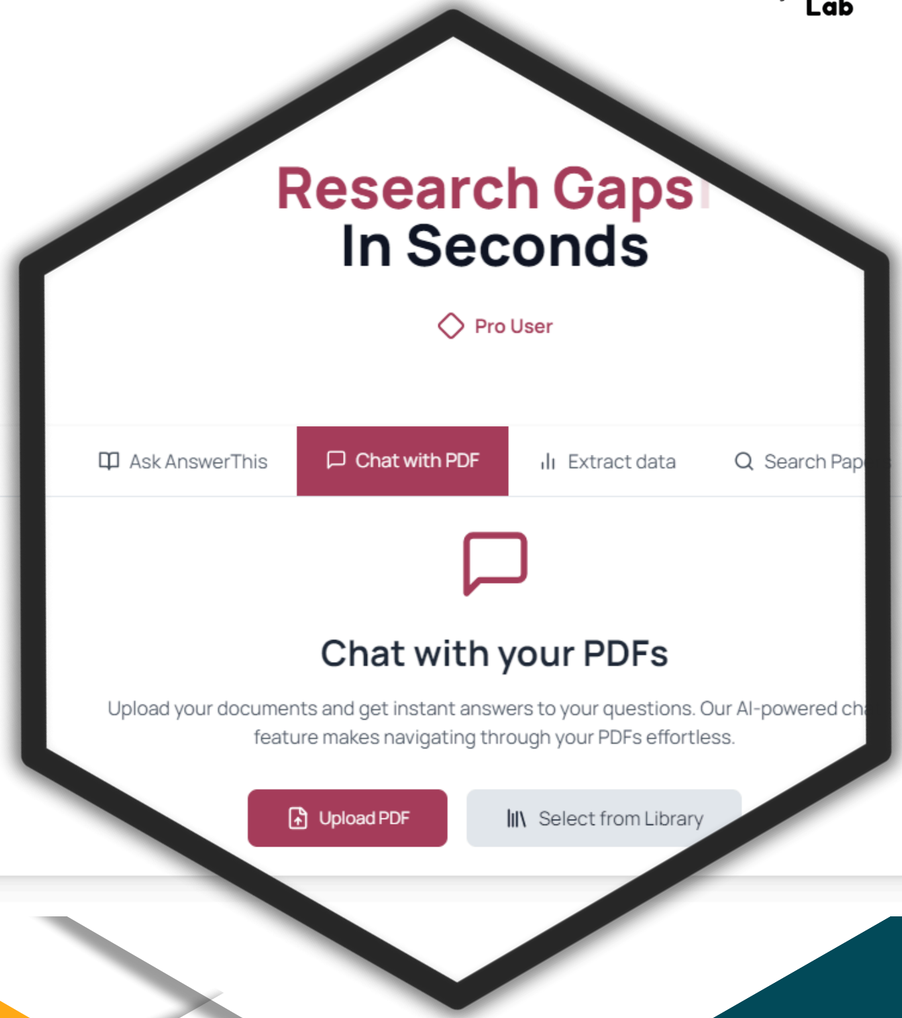
Essential Tools for Researchers:

Boost Your Research Productivity

idSPORA

by : Ko+ Lab





idSPORA

by : Ko+ Lab

Elicit – https://elicit.com/

- Elicit digunakan untuk mencari dan menganalisis jurnal secara cepat dengan bantuan AI.
- Elicit membantu menyelesaikan tugas penelitian yang memakan waktu seperti merangkum jurnal, mengekstraksi data, dan mensintesis temuan

Subscription & billing

Your plan: Basic

Monthly Annual

	Basic	Plus	Pro	Team
	For students and casual exploration	For literature reviews and deeper research	For small systematic reviews and professional research	For large systematic reviews with 2 or more collaborators
	Free	**$10** / month $120 billed annually	**$42** / month $499 billed annually	**$65** per user / month $780 billed annually
	Your current plan	Choose plan	Choose plan	Choose plan
Search across more than 125 million papers	Unlimited	Unlimited	Unlimited	Unlimited
Research Reports	✓	✓	✓	✓
Systematic Review Workflow	✗	✗	✓	✓
Summarization	Unlimited summaries, up to 4 papers at once	Unlimited summaries, up to 8 papers at once	Unlimited summaries, up to 8 papers at once	Unlimited summaries, up to 8 papers at once
Chat with papers	Unlimited chat, up to 4 papers at once	Unlimited chat, up to 8 papers at once	Unlimited chat, up to 8 papers at once	Unlimited chat, up to 8 papers at once
Extract data from PDFs	20 PDFs per month	600 PDFs per year	2,400 PDFs per year	3,600 PDFs per year per user, pooled across your team

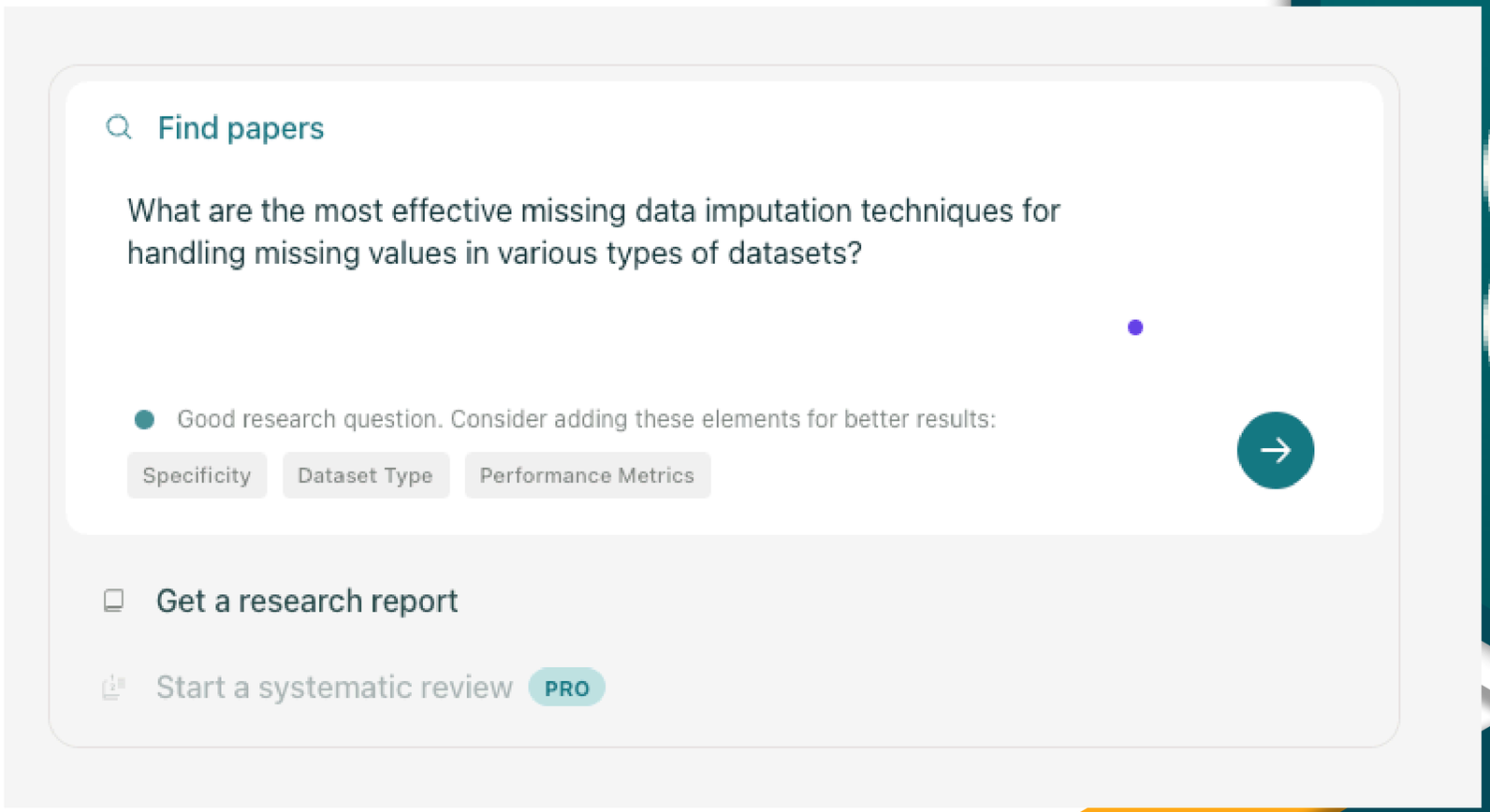
idSPORA

by : Ko+ Lab

idSPORA

by : Ko+ Lab

Elicit – Find Papers



idSPORA

by : Ko+ Lab

Elicit – Find Papers

Effective Missing Data Imputation Techniques

Share

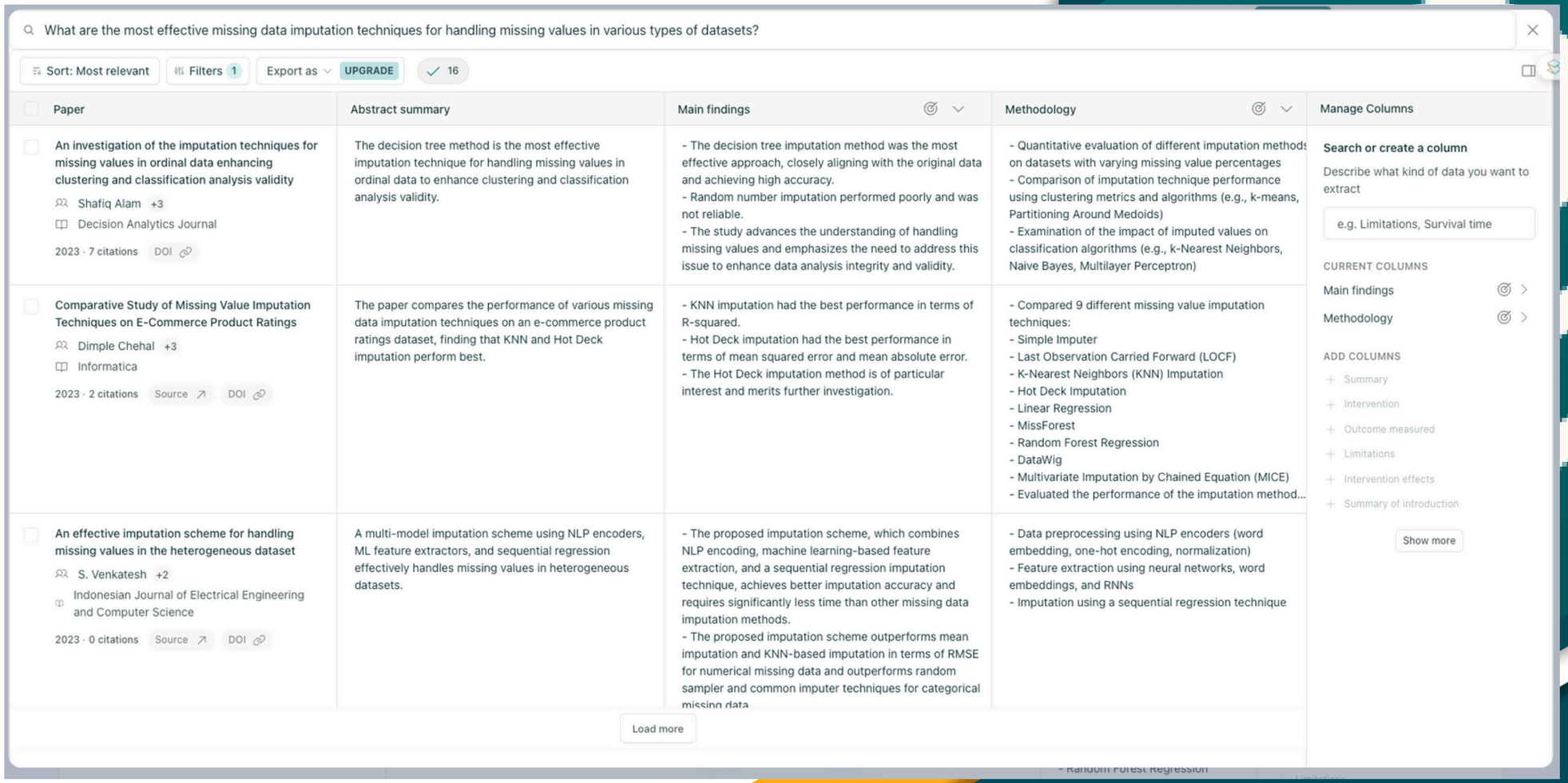
What are the most effective missing data imputation techniques for handling missing values...

Summary of top 4 papers

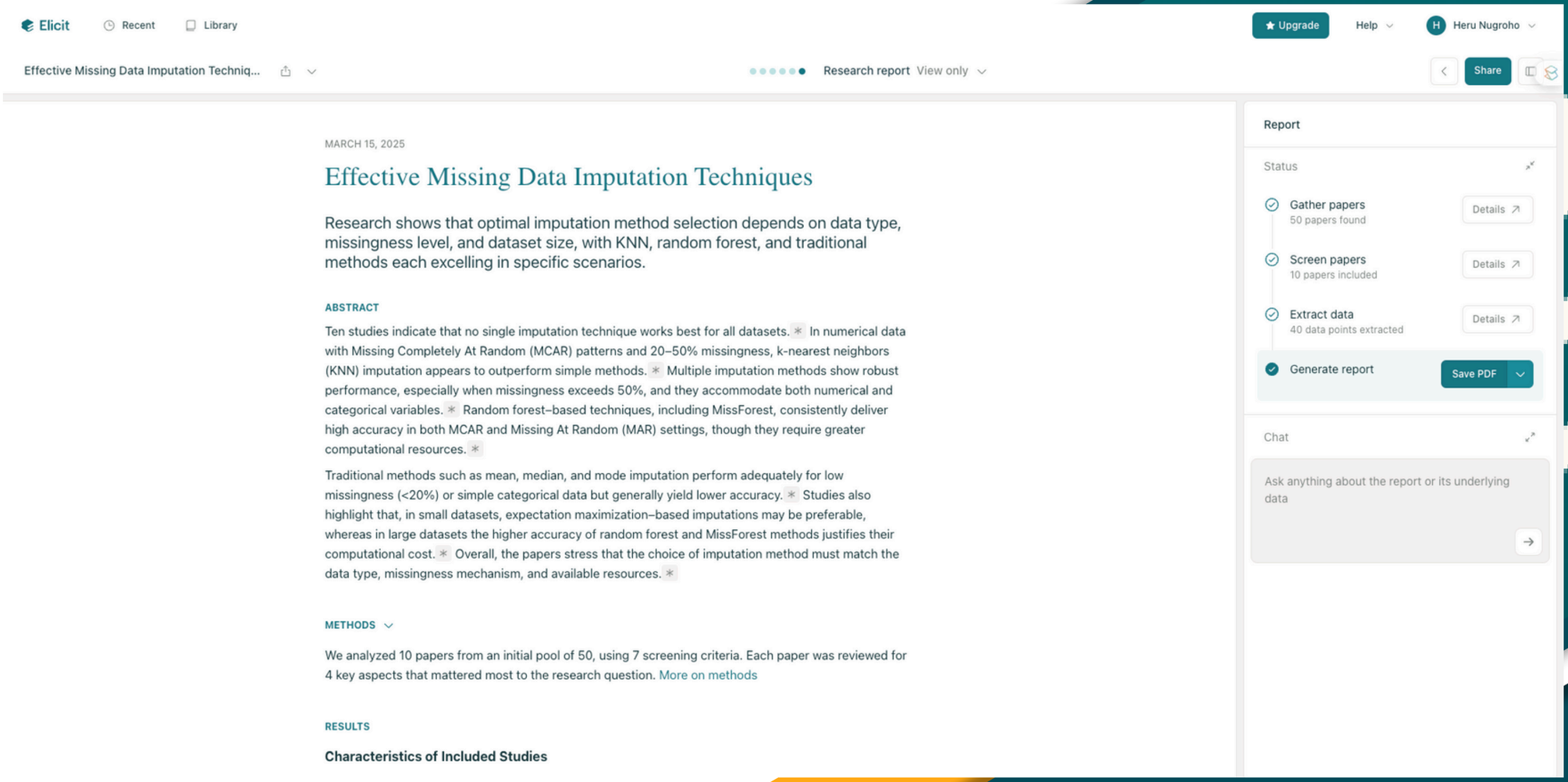
Copy

Missing data imputation is crucial for maintaining dataset integrity and ensuring reliable analysis. Various techniques have been evaluated for different data types and scenarios. For ordinal data, decision tree methods have shown superior performance in clustering and classification tasks (Shafiq Alam et al., 2023). In e-commerce product ratings, KNN imputation yielded the best R-squared values, while hot deck imputation performed well for mean squared and absolute errors (Chehal et al., 2023). For heterogeneous datasets, a multi-model approach combining NLP encoders, machine learning feature extractors, and sequential regression imputation has demonstrated improved accuracy and efficiency (Venkatesh et al., 2023). In numeric datasets, kNN imputation consistently outperformed other methods across various datasets and missing value percentages (Jadhav et al., 2019). These findings highlight the importance of selecting appropriate imputation techniques based on data type and analysis requirements to enhance the validity and reliability of subsequent data analysis tasks.

Elicit – Find Papers



Elicit – Research Report



idSPORA

by : Ko+ Lab

Elicit – Research Report

Effective Missing Data Imputation Techniques

Research shows that optimal imputation method selection depends on data type, missingness level, and dataset size, with KNN, random forest, and traditional methods each excelling in specific scenarios.

Abstract

Ten studies indicate that no single imputation technique works best for all datasets. In numerical data with Missing Completely At Random (MCAR) patterns and 20–50% missingness, k-nearest neighbors (KNN) imputation appears to outperform simple methods. Multiple imputation methods show robust performance, especially when missingness exceeds 50%, and they accommodate both numerical and categorical variables. Random forest–based techniques, including MissForest, consistently deliver high accuracy in both MCAR and Missing At Random (MAR) settings, though they require greater computational resources.

Traditional methods such as mean, median, and mode imputation perform adequately for low missingness (<20%) or simple categorical data but generally yield lower accuracy. Studies also highlight that, in small datasets, expectation maximization–based imputations may be preferable, whereas in large datasets the higher accuracy of random forest and MissForest methods justifies their computational cost. Overall, the papers stress that the choice of imputation method must match the data type, missingness mechanism, and available resources.

Paper search

Using your research question "What are the most effective missing data imputation techniques for handling missing values in various types of datasets?", we searched across over 126 million academic papers from the Semantic Scholar corpus. We retrieved the 50 papers most relevant to the query.

Screening

We screened in papers that met these criteria:

- **Empirical Evaluation**: Does the study include empirical evaluation (using real or simulated data) of at least one imputation method for handling missing values?
- **Performance Metrics**: Does the study report specific quantitative performance measures (e.g., RMSE, MAE, bias) to assess imputation accuracy?
- **Dataset and Mechanism**: Does the study use real-world or simulated datasets AND clearly specify the type of missingness (MCAR, MAR, or MNAR) being addressed?
- **Comparative Analysis**: Does the study compare multiple imputation techniques or compare at least one imputation technique against a baseline method?
- **Methodology Description**: Does the study provide sufficient detail about the imputation methods used and evaluation procedures to enable reproducibility?
- **Study Type**: Is the study either (a) primary research with empirical evaluation OR (b) a systematic review/meta-analysis of imputation techniques?
- **Research Focus**: Does the study focus on evaluating imputation methods (rather than solely on missing data detection or prevention)?

We considered all screening questions together and made a holistic judgement about whether to screen in each paper.

idSPORA

by : Ko+ Lab

AnswerThis – https://answerthis.io/

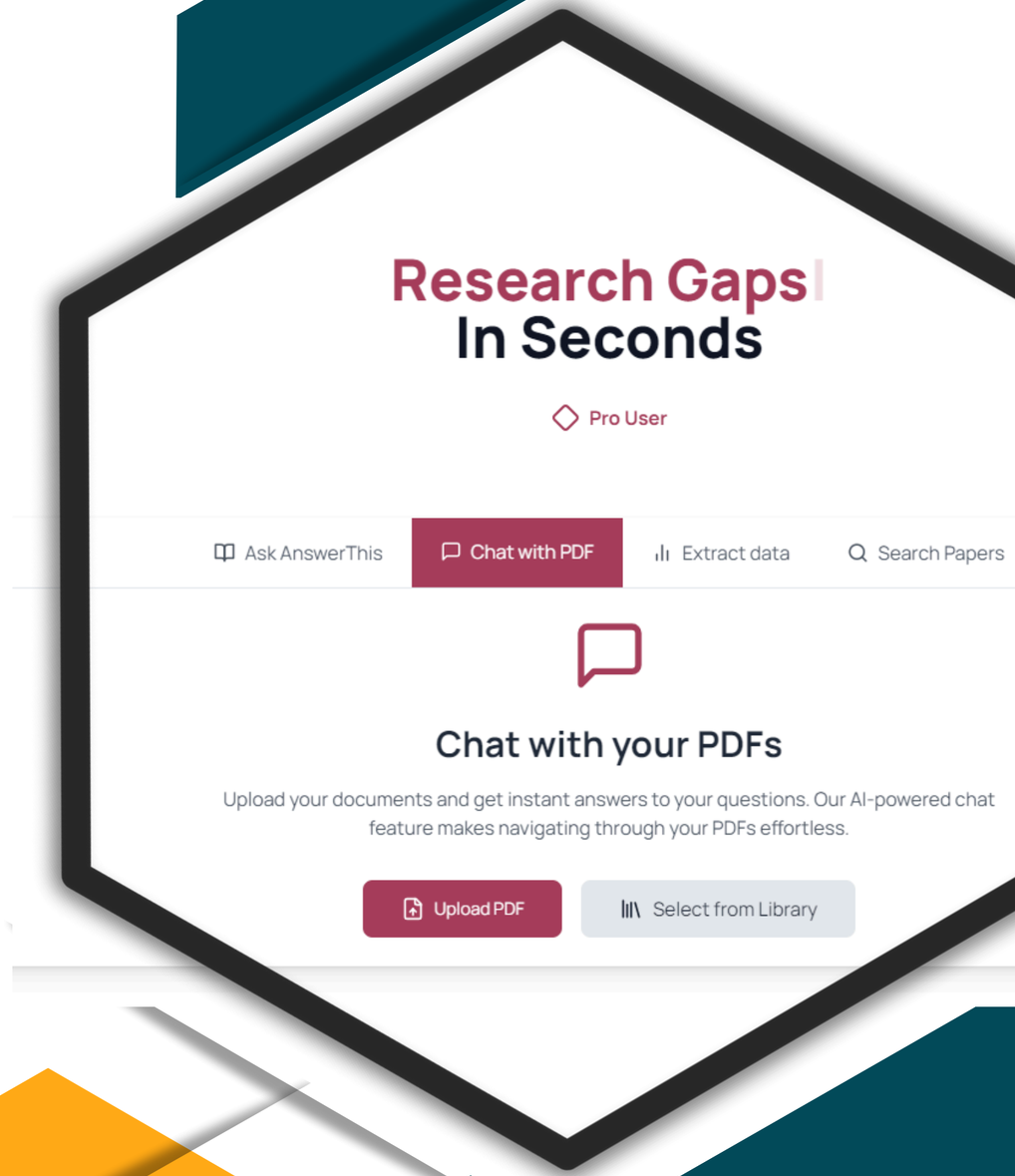
- AnswerThis adalah alat AI yang dirancang untuk menyederhanakan proses literature review.
- Dengan menggunakan AnswerThis, peneliti dapat dengan mudah menemukan, menganalisis, dan mengorganisir sumber-sumber akademik secara efisien

idSPORA

by : Ko+ Lab

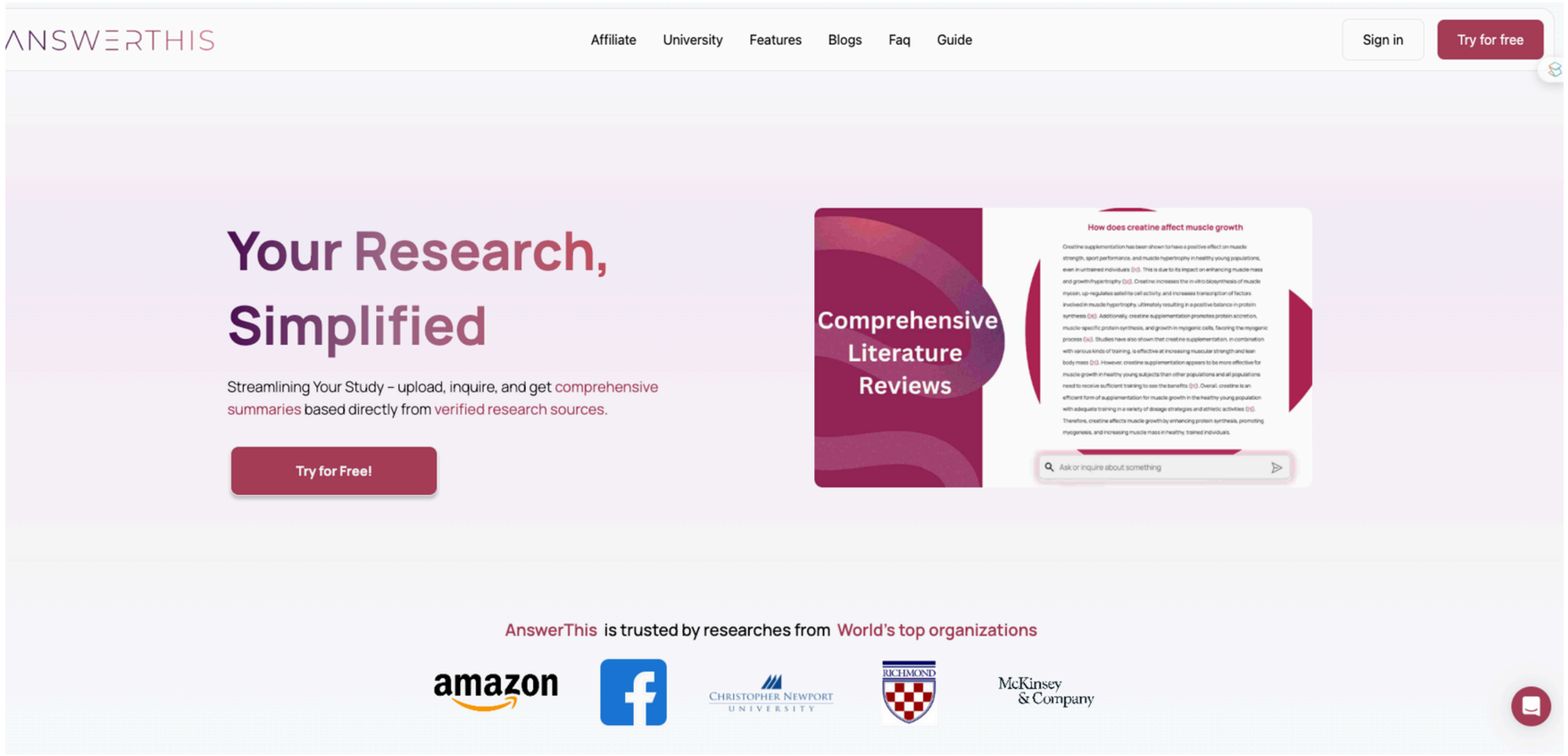
Keunggulan AnswerThis

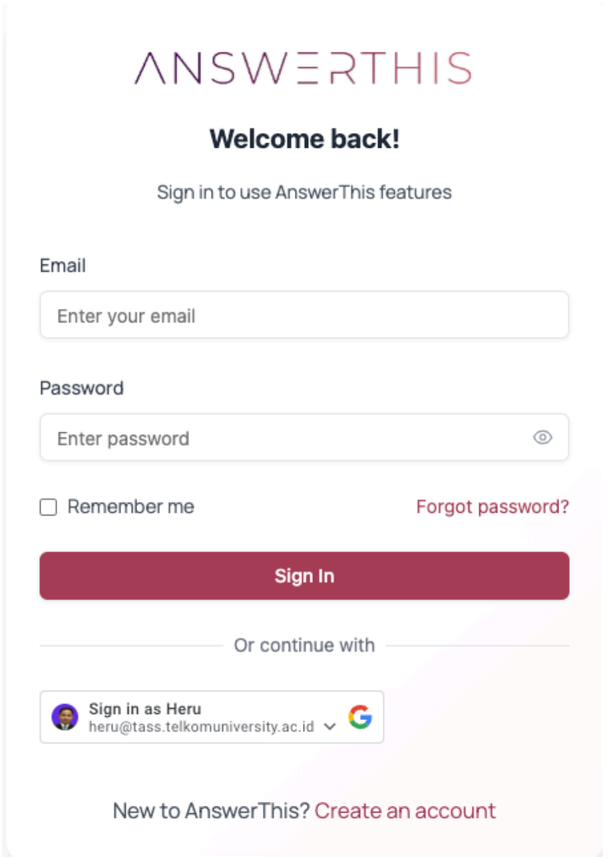
- Evolusi dari literature review hingga pembuatan paper akademik secara lengkap.
- Memiliki library & editing tools untuk mendukung workflow penelitian.
- Akurasi lebih baik dan lebih komprehensif dibandingkan tools lain.
- Konsisten menghasilkan penelitian yang berwawasan dan berbasis kutipan.



Masuk ke Platform AI Research

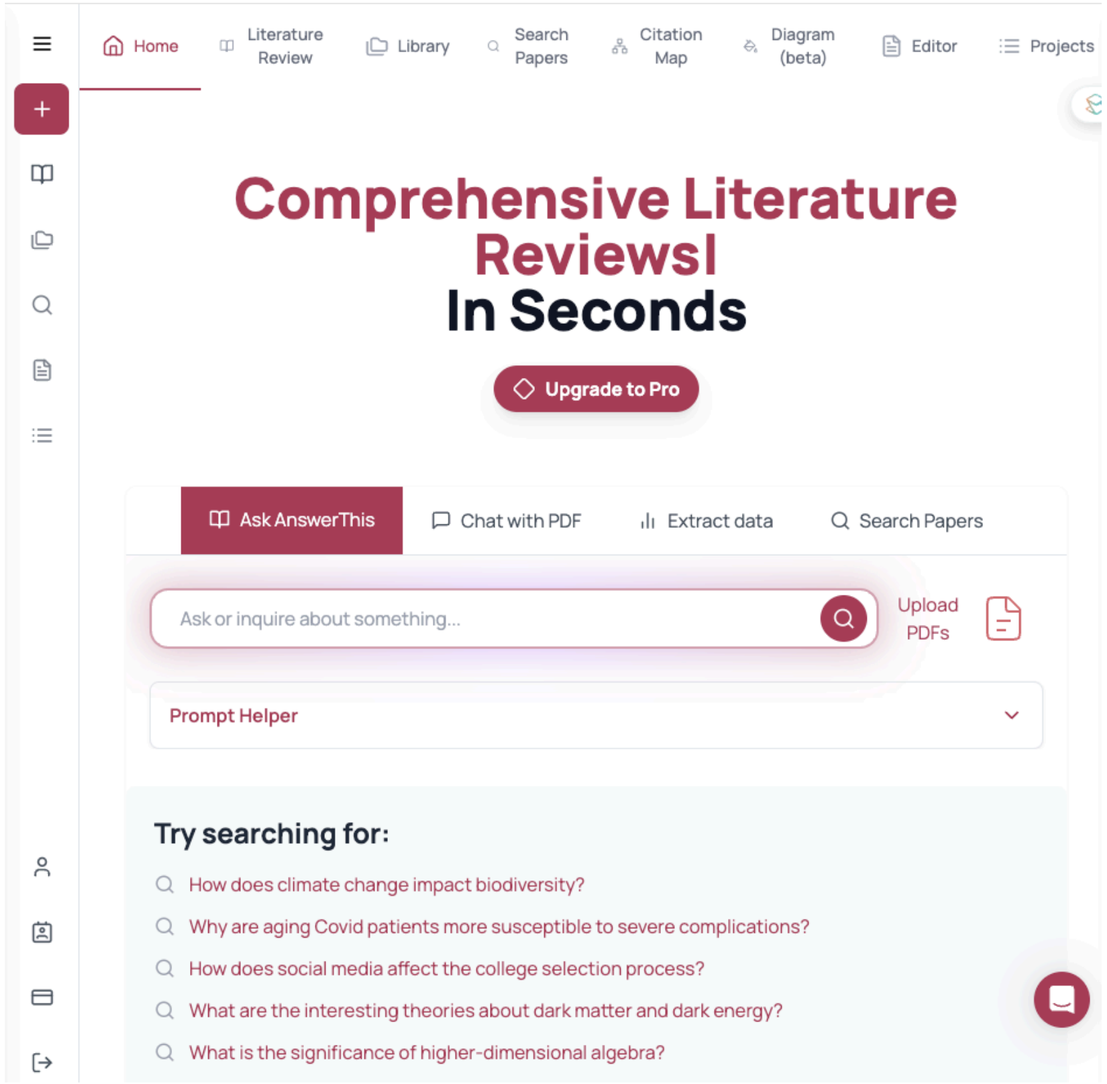
https://answerthis.io

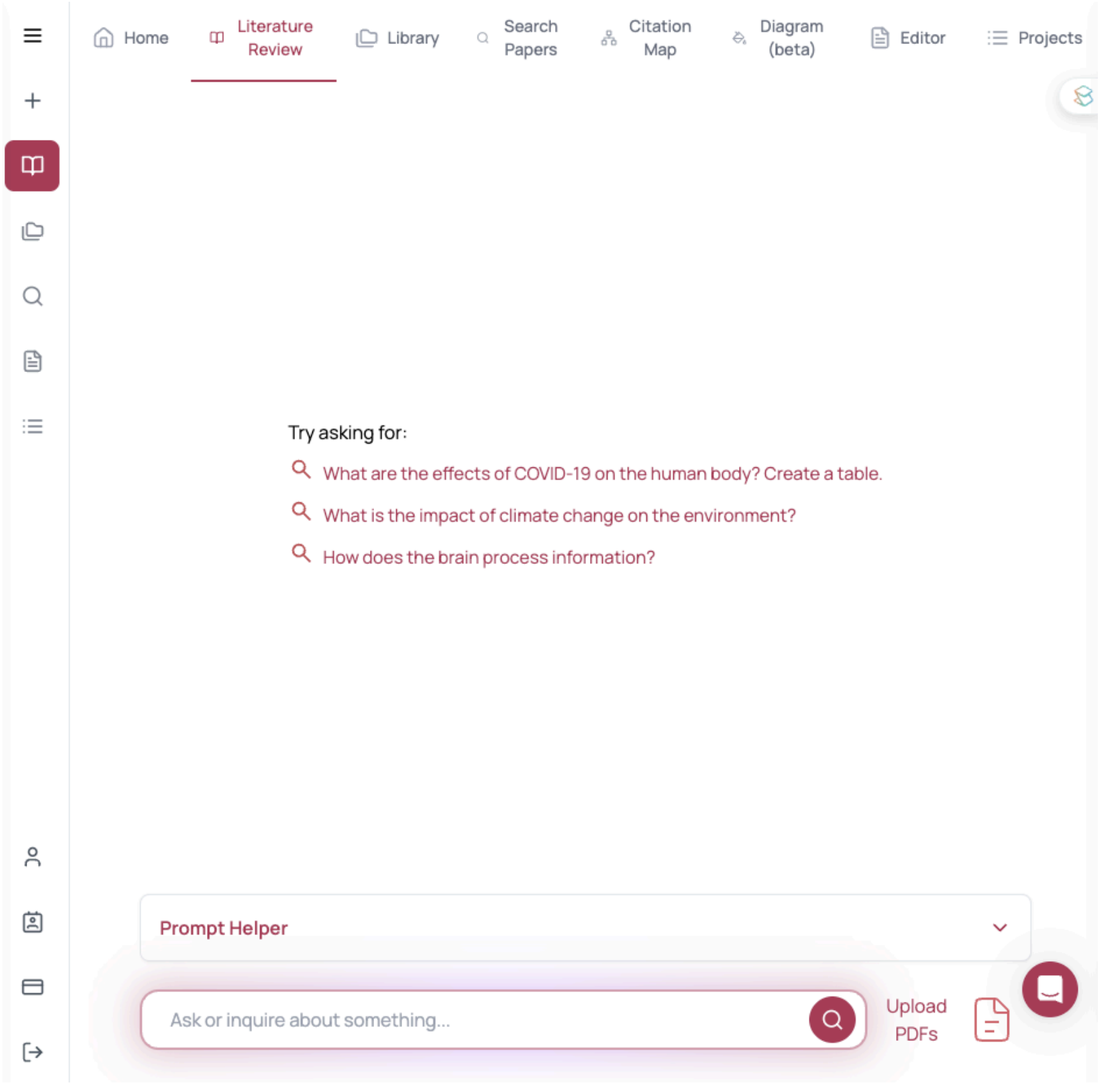




Pilih Mode Literature Review

idSPORA
by : Ko+ Lab



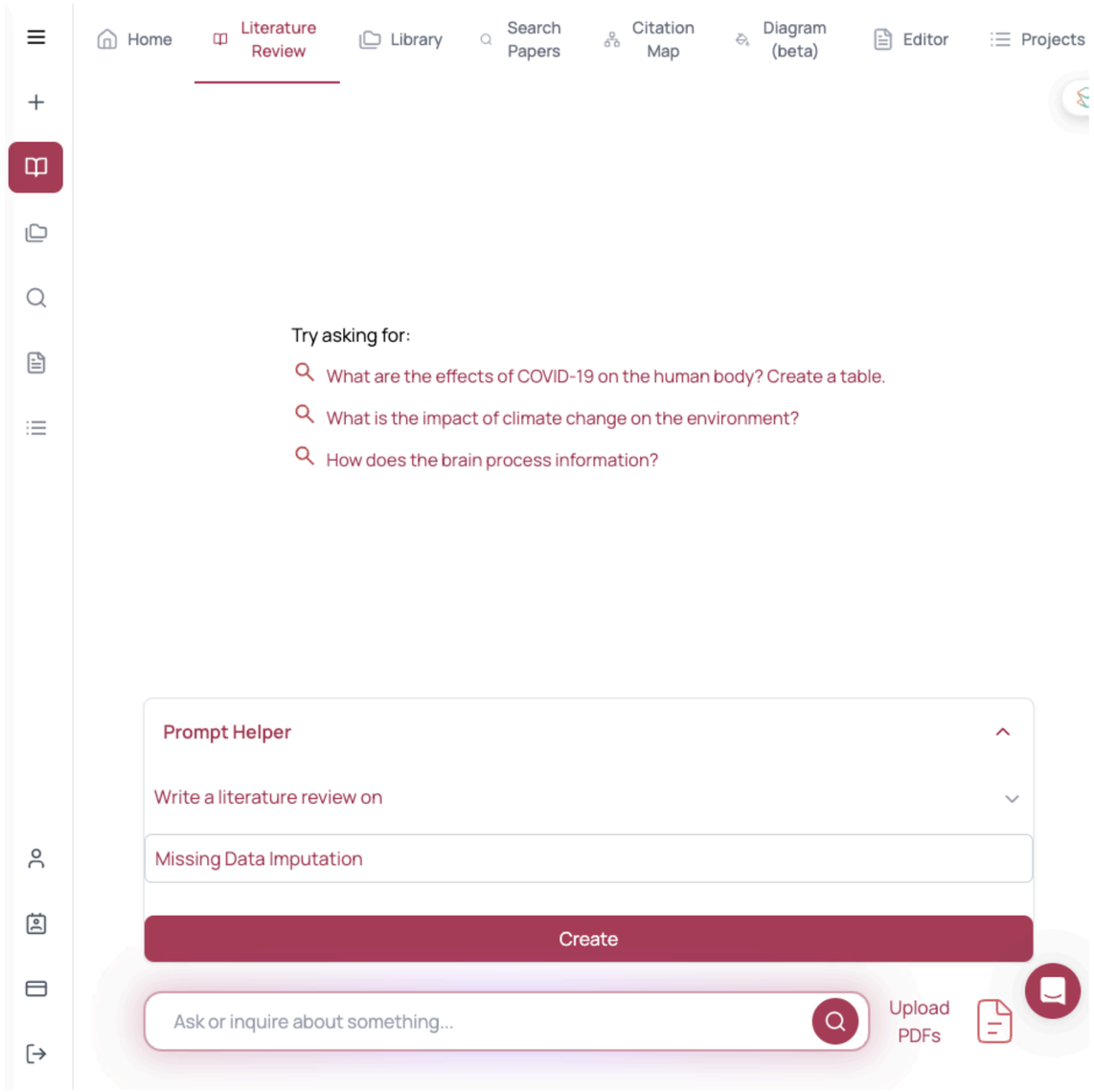


Gunakan Prompt Helper

✍️ Pilih "Write a literature review on..." dan masukkan topik riset
🔹 Contoh: Missing Data Imputation

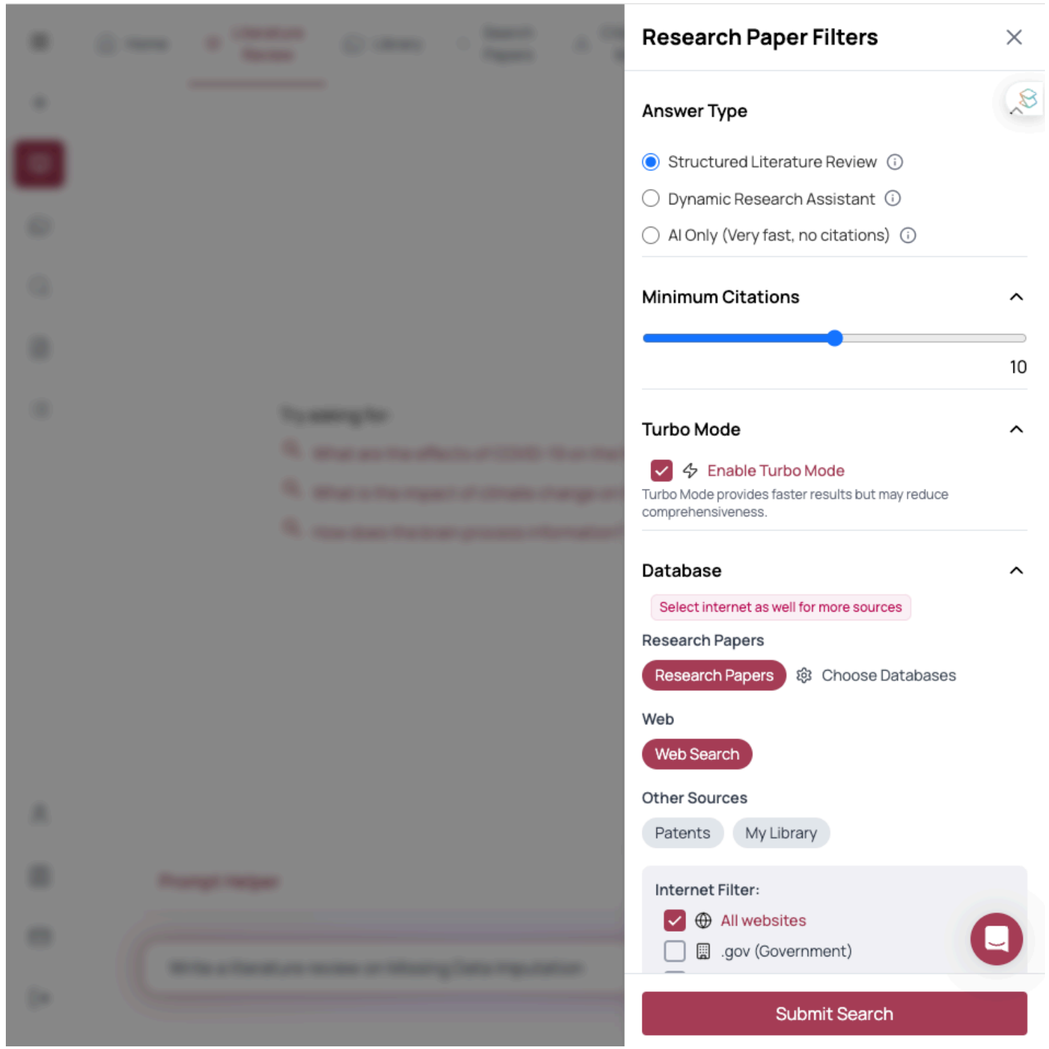
📌 Klik Create untuk menghasilkan prompt awal.

🔍 Tekan Enter untuk menampilkan opsi filter riset.



Atur Jumlah Kutipan Minimum

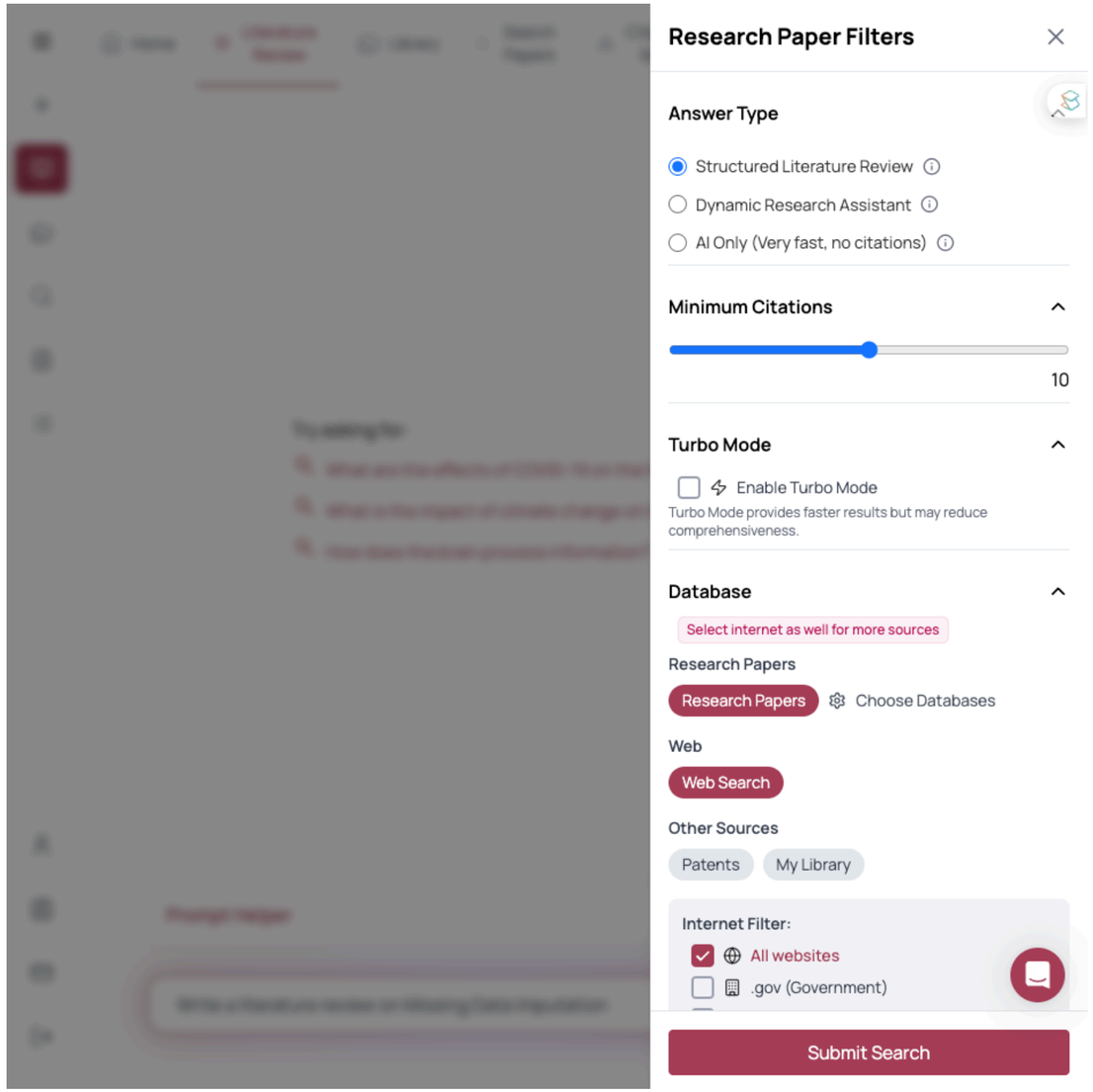
📌 Pilih minimal 10 kutipan agar hasil lebih komprehensif.



Aktifkan atau Nonaktifkan Turbo Mode

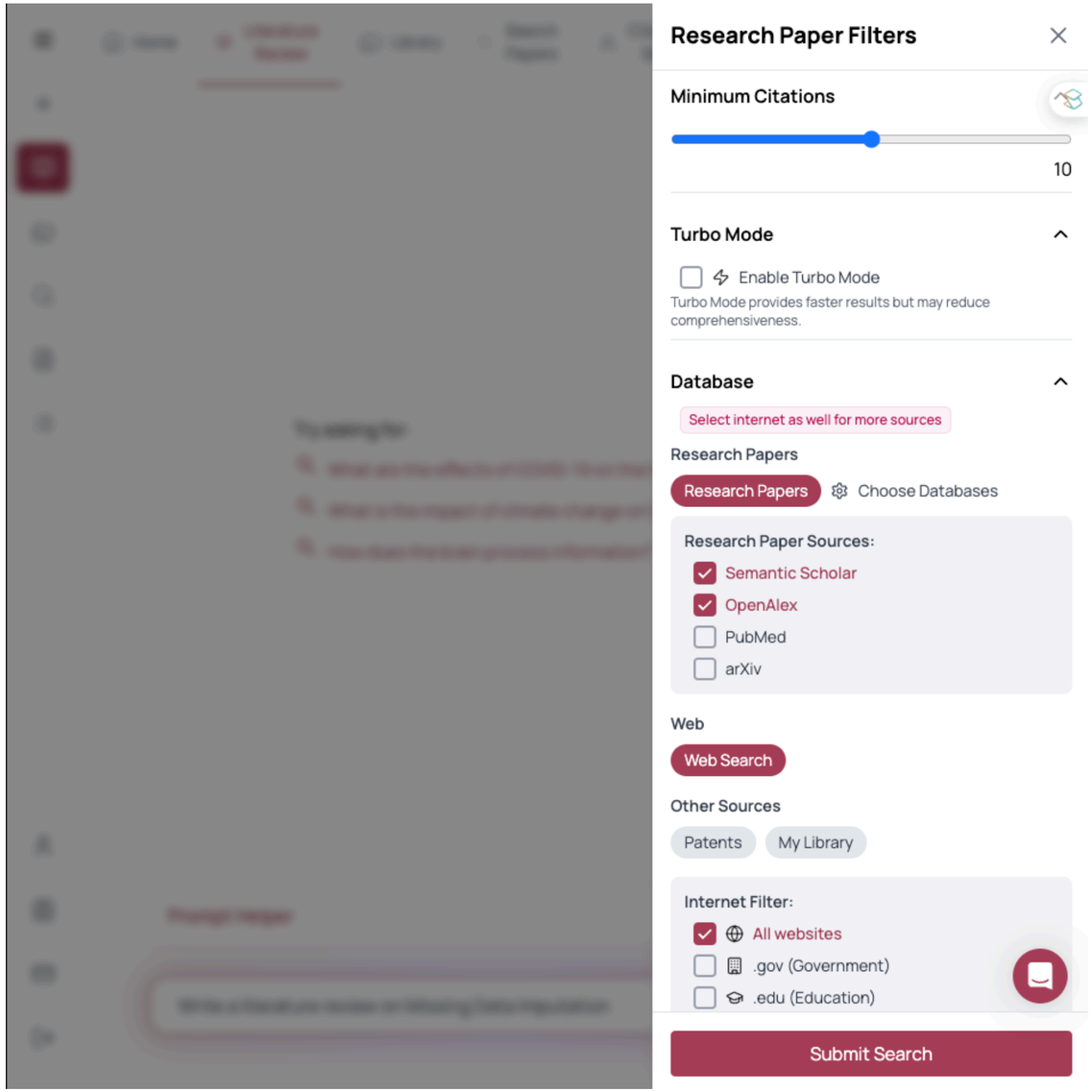
⚡ Turbo Mode ON → Hasil lebih cepat.

🔍 Turbo Mode OFF → Jawaban lebih mendalam & detail.



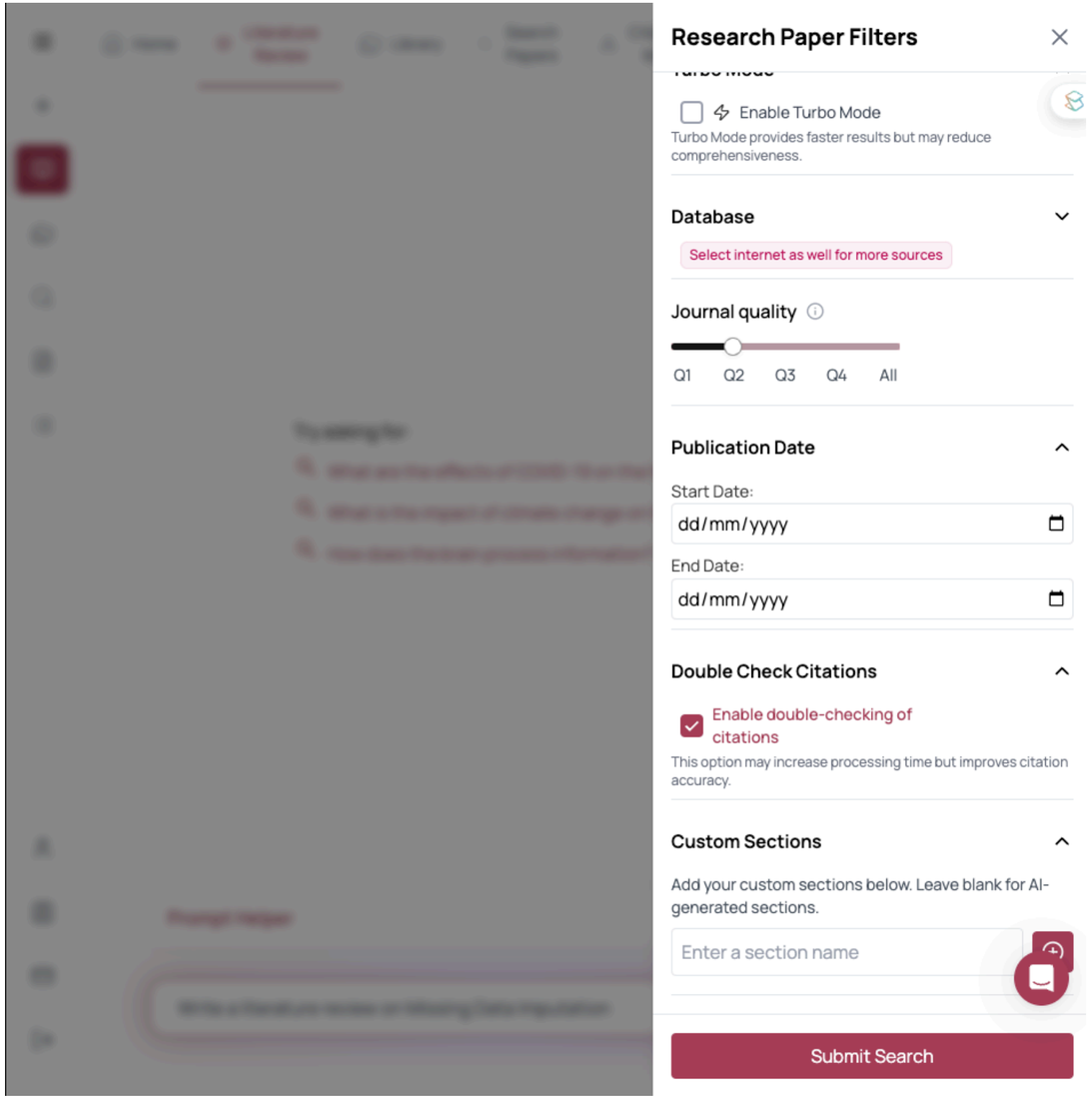
Pilih Sumber Pencarian

🌐 Web + Database untuk cakupan riset yang lebih luas.



Pilih Jurnal Quality dan Rentang Waktu

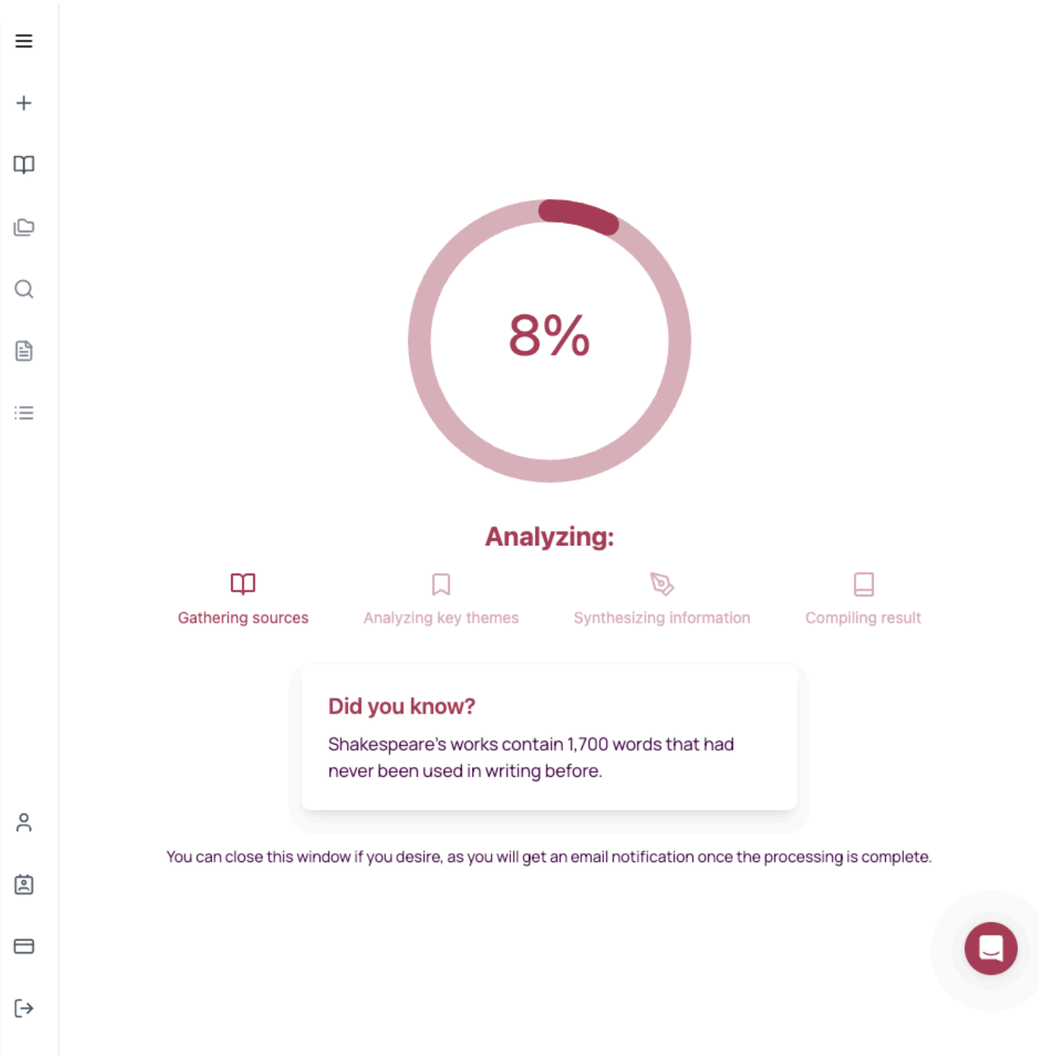
📅 Pilih rentang waktu publikasi agar mendapatkan paper terbaru (5/10 tahun terakhir)



Submit & Tinjau Hasil

📌 Klik Submit Search dan tunggu hingga review literatur dihasilkan.

📜 Sumber dan kutipan akan muncul di bagian kanan.



idSPORA
by : Ko Lab

New Search

Your Projects

Default project

Recent Queries

Write a literature r ...
2/18/2025

Chat with Papers

Heru Nugroho
herupipo@gmail.com

Trial Status

6 literature reviews left

8 paper uploads left

Upgrade to PRO

Log Out

Answer

Write a literature review on Missing Data Imputation

A Literature Review on Missing Data Imputation

Introduction

Missing data is a pervasive problem in various fields, including healthcare [1], [2], [3], environmental science [4], [5], and engineering [6], [7], [8]. The presence of missing values can significantly compromise the validity and reliability of research findings [9], [10], [11], leading to biased estimates and inaccurate conclusions. This literature review systematically examines various missing data imputation techniques, analyzing their strengths, weaknesses, and applicability across different domains. We will explore both traditional and advanced methods, considering their underlying assumptions and the impact on downstream analyses. The review also identifies research gaps and suggests future directions for improving the accuracy and efficiency of missing data imputation.

Types of Missing Data

Before delving into imputation techniques, it's crucial to understand the different mechanisms of missing data. The most common classification distinguishes between Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR) [12], [9]. MCAR implies that the probability of missingness is unrelated to any observed or unobserved variables [13]. This is the most desirable scenario, as it simplifies the imputation process. However, MCAR is rarely encountered in practice. MAR, on the other hand, suggests that the probability of missingness depends on observed variables but not on the missing values themselves [13], [10].

Ask follow up question

Sources

All | PDF | Search

Table View | Sort | Extract Data | CHICAGO | Export

Select papers to perform actions like Chat, Summarize, or Export.

Effects of Different Missing Data Imputation Techniques on the Performance of Undiagnosed Diabetes Risk Prediction Models in a Mixed-Ancestry Population of South Africa

Katya L. Masconi, T. Matsha, R. Erasmus, A. Kengne

PLoS ONE

Background Imputation techniques used to handle missing data are based on the principle of replacement. It is widely advocated that multiple imputation is superior to other imputation methods, however studies have suggested that simple methods for filling missing data can be just as accurate as complex methods. The objective of this study was to implement a number of simple and more complex imputa ...

Insight Related to Question:

Multiple imputation is often deemed superior but simpler methods can perform comparably well in certain contexts, as shown in undiagnosed diabetes risk prediction models. Deletion of missing data resulted in the poorest model performance, highlighting the importance of imputation methods. Stochastic regression imputation performed well, with a significant number of individuals imputed, but simpler methods often yielded similar predictive performances. The study utilized systematic reviews to identify various imputation and predictive modeling techniques, reinforcing a comprehensive approach to missing data handling.

PDF Available DOI: https://doi.org/10.1371/journal.pone.0139210 Number of citations: 24

Date Of Publication: 2015-09-25

Masconi, Katya L., Matsha, T., Erasmus, R., and Kengne, A.. 2015. "Effects of Different Missing Data Imputation Techniques on the Performance of Undiagnosed Diabetes Risk Prediction Models in a Mixed-Ancestry Population of South Africa". PLoS ONE. https://doi.org/10.1371/journal.pone.0139210

Feedback

idSPORA

by : Ko Lab

New Search

Your Projects

Default project

Recent Queries

Write a literature r ...
2/18/2025

Chat with Papers

Heru Nugroho
herupipo@gmail.com

Trial Status

6 literature reviews left

8 paper uploads left

Upgrade to PRO

Log Out

Answer

Write a literature review on Missing Data Imputation

A Literature Review on Missing Data Imputation

Introduction

Missing data is a pervasive problem in various fields, including healthcare [1], [2], [3], environmental science [4], [5], and engineering [6], [7], [8]. The presence of missing values can significantly compromise the validity and reliability of research findings [9], [10], [11], leading to biased estimates and inaccurate conclusions. This literature review systematically examines various missing data imputation techniques, analyzing their strengths, weaknesses, and applicability across different domains. We will explore both traditional and advanced methods, considering their underlying assumptions and the impact on downstream analyses. The review also identifies research gaps and suggests future directions for improving the accuracy and efficiency of missing data imputation.

Types of Missing Data

Before delving into imputation techniques, it's crucial to understand the different mechanisms of missing data. The most common classification distinguishes between Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR) [12], [9]. MCAR implies that the probability of missingness is unrelated to any observed or unobserved variables [13]. This is the most desirable scenario, as it simplifies the imputation process. However, MCAR is rarely encountered in practice. MAR, on the other hand, suggests that the probability of missingness depends on observed variables but not on the missing values themselves [13], [10].

Ask follow up question

Sources

All | PDF | Search

List View | Sort | Extract Data | CHICAGO | Export

Select papers to perform actions like Chat, Summarize, or Export.

Source	Source Information	Insight Related to Question	Relevant Extracts
1	Effects of Different Missing Data Imputation Techniques on the Performance of Undiagnosed Diabetes Risk Prediction Models in a Mixed-Ancestry Population of South Africa PDF: Available Author: Katya L.	Multiple imputation is often deemed superior but simpler methods can perform comparably well in certain contexts, as shown in undiagnosed diabetes risk prediction models. Deletion of missing data resulted in the poorest model performance, highlighting the importance of imputation methods. Stochastic	[1] Extracted Source Background Imputation techniques used to handle missing data are based on the principle of replacement. It is widely advocated that multiple imputation is superior to other imputation methods, however studies have suggested that simple methods for filling
2	A Survey on Data Imputation Techniques: Water Distribution System as a Use Case PDF: Not Available Author: M. S. Osman,	The paper reviews various techniques for missing data imputation from traditional methods to advanced algorithms. Traditional methods discussed include deletion and single imputation, while	[1] Extracted Source The presence of missing data is problematic in most quantitative research studies. Water distribution systems (WDSs) are not immune t this problem. In fact, missing dat

Feedback

A Literature Review on Missing Data Imputation

Introduction

Missing data is a pervasive problem in various fields, including healthcare [1], [2], [3], environmental science [4], [5], and engineering [6], [7], [8]. The presence of missing values can significantly compromise the validity and reliability of research findings [9], [10], [11], leading to biased estimates and inaccurate conclusions. This literature review systematically examines various missing data imputation techniques, analyzing their strengths, weaknesses, and applicability across different domains. We will explore both traditional and advanced methods, considering their underlying assumptions and the impact on downstream analyses. The review also identifies research gaps and suggests future directions for improving the accuracy and efficiency of missing data imputation.

Types of Missing Data

Before delving into imputation techniques, it's crucial to understand the different mechanisms of missing data. The most common classification distinguishes between Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR) [12], [9]. MCAR implies that the probability of missingness is unrelated to any observed or unobserved variables [13]. This is the most desirable scenario, as it simplifies the imputation process. However, MCAR is rarely encountered in practice. MAR, on the other hand, suggests that the probability of missingness depends on observed variables but not on the missing values themselves [13], [10]. Finally, MNAR, the most challenging scenario, indicates that the probability of missingness is related to the missing values [12], [10]. Determining the missing data mechanism is crucial for selecting an appropriate imputation technique, as different methods are better suited for different scenarios [12]. The generation of synthetic missing data, often used to evaluate imputation algorithms, needs to accurately reflect these mechanisms to yield valid results [12]. Failing to consider the missing data mechanism can lead to biased estimates and flawed conclusions [10].

Traditional Imputation Methods

Traditional methods for handling missing data often involve simple techniques with limited statistical sophistication. These include complete case analysis, where observations with any missing values are discarded [11], [14], and mean/mode imputation, where missing values are replaced with the mean or mode of the observed values for that variable [13], [15]. While these methods are straightforward to implement, they suffer from several limitations. Complete case analysis can lead to substantial loss of information and reduced statistical power, particularly when the missing data rate is high [11], [14]. Moreover, it can introduce bias if the missingness is not MCAR [10]. Mean/mode imputation can distort the distribution of the variable and underestimate the variance, leading to inaccurate standard errors and potentially misleading inferences [13], [15]. Other simple imputation methods such as hot-deck imputation (replacing missing values with observed values from similar cases) and regression imputation (predicting missing values based on a regression model) also have their limitations [15], [16]. While these techniques are easier to implement than more sophisticated methods, their simplicity often comes at the cost of accuracy and robustness [17].

Advanced Imputation Methods

In response to the limitations of traditional methods, more advanced techniques have been developed. Multiple imputation (MI) is a widely used approach that generates multiple plausible imputed datasets, each reflecting the uncertainty associated with the missing values [13], [1], [10]. This approach accounts for the uncertainty inherent in the imputation process, leading to more accurate and reliable inferences [13], [1], [10]. Multiple imputation by

idSPORA

by : Ko Lab

idSPORA

by : Ko+ Lab

Hatur Nuhun

- email: heru@tass.telkomuniversity.ac.id
- WA: 081394322043