Polytechnic University of Catalonia
BDMA Joint Project

Scholaria

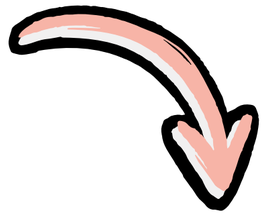
Your long-term academic assistant

June 11th, 2024

OUTLINE

- Business Side
  - Our Team & Product & Innovation
  - Market Analysis & Our Marketing
  - Ethical Analysis
  - Financials & Our Business Plan
- Demo
- Tech & Implementation
  - Integration with BDM
  - Integration with SDM
- Q&A

We are looking for



€150,000

For

10% Equity

Manual Keyword Search

Snowballing

Manual Organization

We value your time to focus on your:

Research

Creativity

Personal time

SCHOLARIA: ALL-IN-ONE

What we offer you?

A software application

- Organized reading with automated categorization
- Reading assistant
- Automated recommendations based on your research history

No more:

- Need to change between several tools to manage your papers
- Effort to find similar and/or newly released papers
- Waste of time to skim through the paper manually

LEARN FROM EXPERIENCE

Pere Codina - UPC and VBP class alumni

Acquired for 10M$
2022

1- FIND THE BEST COFOUNDERS WHO CAN DELIVER RESULTS!

4 different continents

5 different backgrounds

Computer Engineer

Furkan

loves basketball

Software Engineer

Louai

loves perfumes

Statistician

Maria

loves dancing

Computer Engineer&Scientist

Rana

loves talking

Computer Scientist &Mathematician

Simon

loves trying new food

1 vision

1 mission

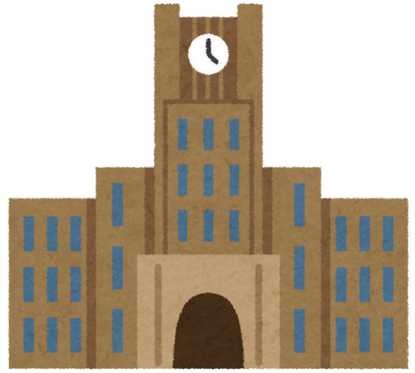
2- FOCUS ON THE CUSTOMER FIRST, NOT THE PRODUCT!

~3M

Researchs Students

~4M

Professors



~2,700

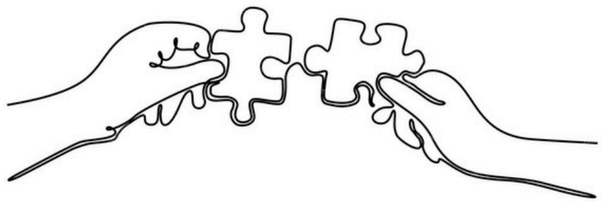
Universities in Europe

Caren A. Arbeit and Michael Yamaner. Trends for graduate student enrollment and postdoctoral appointments in science, engineering, and health fields at u.s. academic institutions between 2017 and 2019. https://ncses.nsf.gov/pubs/nsf21317. Published: March 31, 2021

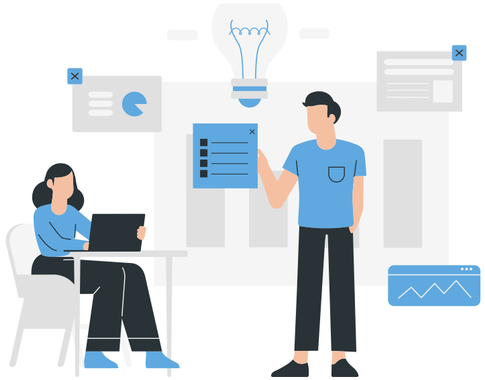
RICHARD PRICE. The number of academics and graduate students in the world. https://richardprice.io/post/12855561694/the-number-of-academics-and-graduate-students-in. Published: Nov 15, 2011.

N.a.Snapshot of higher education in europe. https://www.4icu.org/Europe/#:~:text=European%20Universities%20World%20Representation&text=According%20to%20the%20uniRank%20database,higher%2Deducation%20institutions%20in%20Europe. Accessed: June 5, 2024

3- THINK AS BIG AS YOU CAN!



Recommender algorithm using semantic intersection of any paper under your workspace!

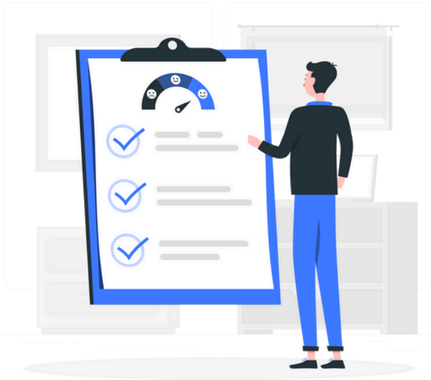


All-in-one research assistant that automizes most of your needs!

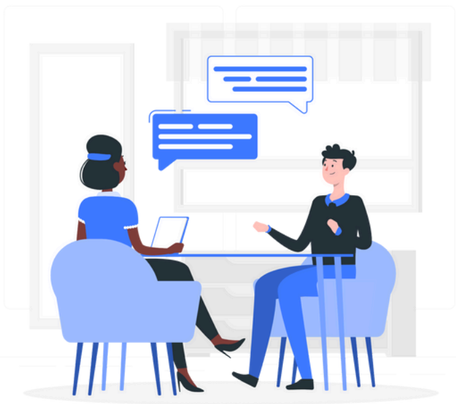
Market Analysis & Our Marketing Strategies

MARKET RESEARCH

Competitor Analysis



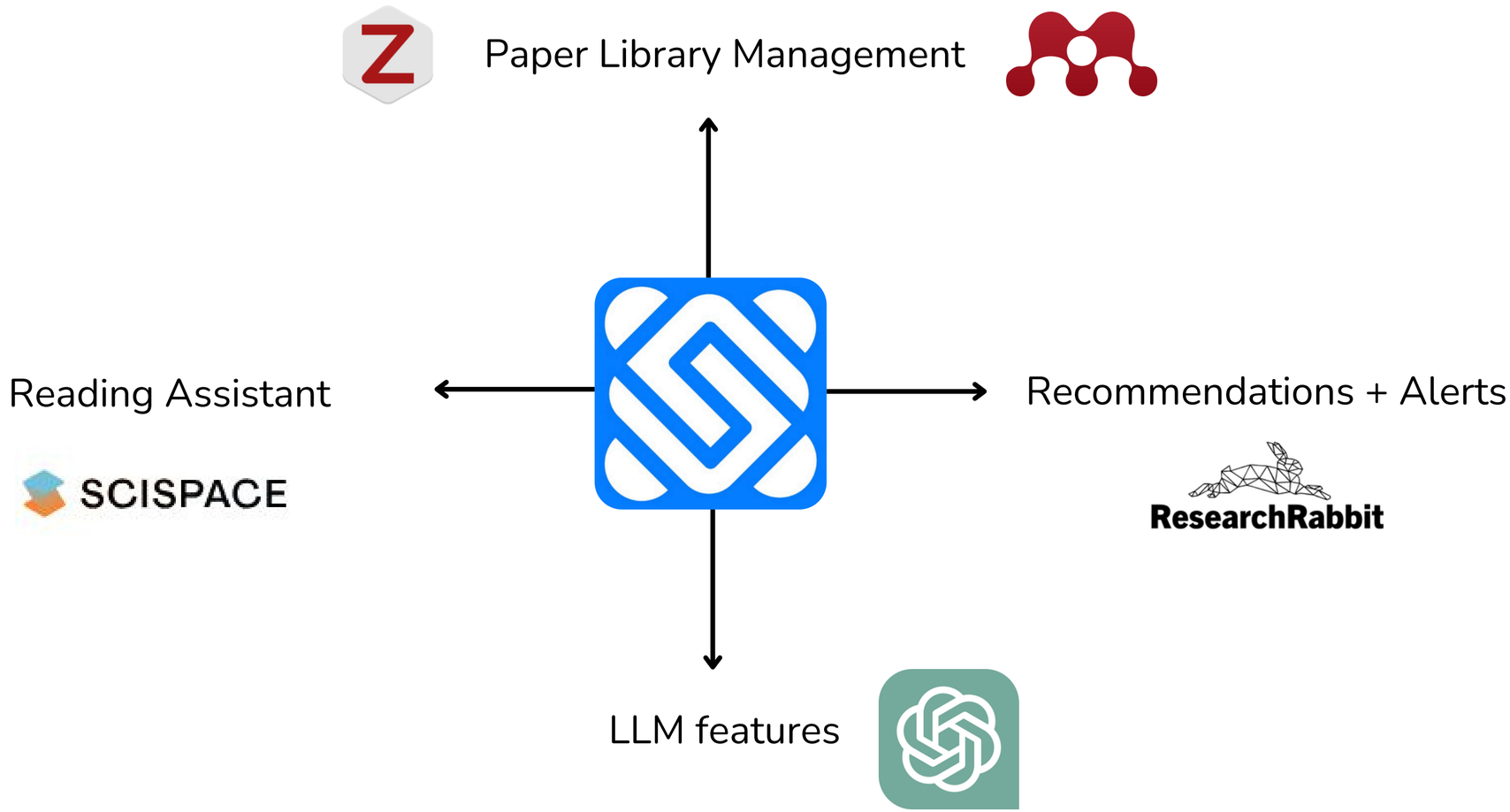
Surveys



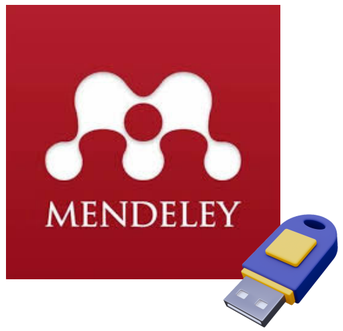
Interviews

Consulting Circle

COMPETITORS: POSITIONING

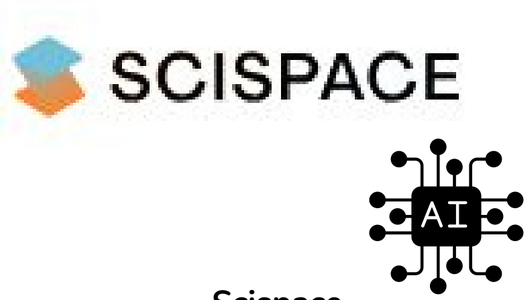


COMPETITORS: PRICING



Mendeley

Free
PLUS $55/year
PRO $110/year
MAX $165/year



Scispace

Free
INDIVIDUAL $144/year
UNIVERSITIES $8/user/month

Zotero

Free
2GB $20/year
6GB $60/year
UNLIMITED $120/year

SURVEYS

23

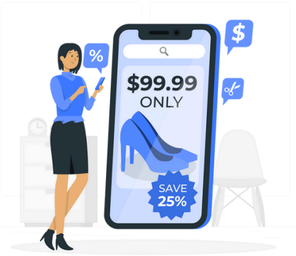
Respondents:
57% Master/PhD students
26% Professors
17% Others

78% Europe
22% LATAM

Articles (monthly reading):
22% Less than 5
70% Between 5 -20
9% More than 20

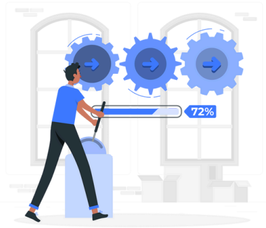
70% use a software
43% Mendeley, +
30% Google Scholar, +

CONSULTING CIRCLE: FEEDBACK



Price:

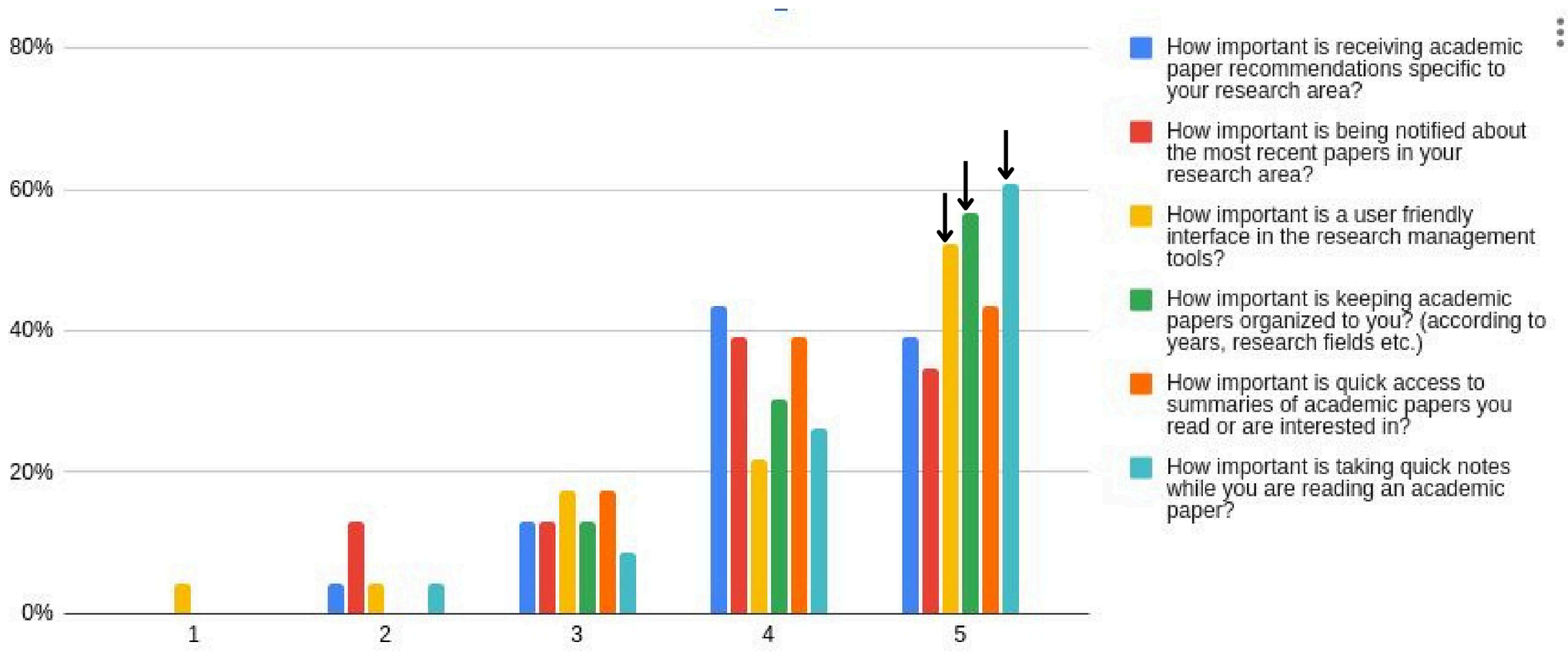
- Majority willing to pay: < $8/month
- Price for the universities: cheap
- Price based on: number of features



Development:

- Pay attention: ChatGPT like features
- Focus in develop one by one
- Recommendation of new papers

SURVEYS + INTERVIEWS



QUICK NOTE TAKING, PAPER ORGANIZATION, FRIENDLY USER-INTERFACE

An **all-in-one** software application:

- paper storage
- regular recommendations based on your library
- paper summarization
- note-taking on papers
- intuitive and user-friendly design

Online Distribution: accessible in the EU (later globally)
Professional Access: corporate packages for industry professionals -> easy integration into their existing systems
University Partnerships: direct integration into university systems for seamless access by students and faculty
(BDMA Partners: 5 Universities)

MARKETING MIX: 4P'S

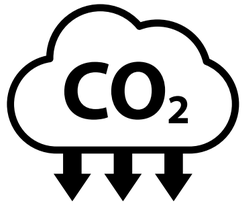
Pricing model (#paper uploads):
Freemium model (free -> subscription)
INDIVIDUAL $5/month
STUDENT $3/month
UNIVERSITIES $2/month
* university bulk subscription: need 12,500 users at €2 to reach the monthly revenue target of €25,000, feasible with only 12 university agreements

Digital Marketing: major channel -> Google Ads
Conference Sponsorships: sponsor & participate in academic conferences to showcase ScholarIA directly to potential users.
Promotional Offers: freemium model -> initial free access followed by discounted rates

Ethical Analysis

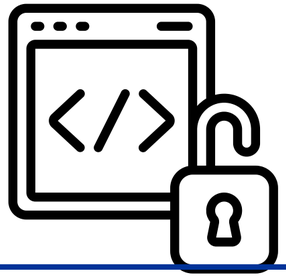
SCHOLARIA EFFECTS

Environmental awareness



Innovation via leveraging cutting-edge technologies

Collaboration

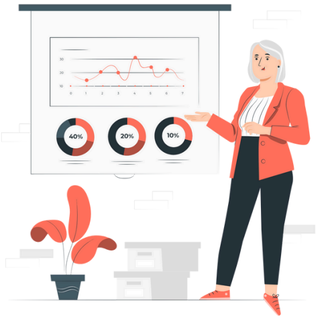


Financials

SHOPPING LIST

Shopping list	**One-off investment**	**Anual expenses**
Four laptops	€ 4.000,00	
Cloud service		€ 5.000,00
Notary, lawyers, deed of incorporation and other	€ 3.000,00	
Initial branding and communication expenses	€ 5.000,00	
marketing expenses	€ 10.000	€ 10.000
Salaries (including Social Security)		€ 300.000,00
TOTAL	**€ 22.000,00**	**€ 315.000,00 €**

120.000 - founders salaries
90.000 - Freelancer



Fixed Assets	€ 4.000,00	Equity	**€ 50.000,00**
Current Assets	€ 0	LT Liabilities	€ 287.000,00
Treasury	€ 333.000,00	ST Liabilities	
Total	**€ 337.000,00**		**€ 337.000,00**

SALES

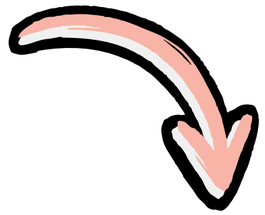
Year/Month	1	2	3	4	5	6	7	8	9	10	11	12	€Total
Year 1	-	-	-	-	-	-	4K	8K	12K	16K	2K0	24K	€168K
Year 2	24K 500	24K 500	24K 500	24K 500	28K 500	24K 500	28K 500	28K 500	28K 500	28K 500	28K 500	28K 500	€317K € 15K

107%

PROFIT AND LOSS

Sales	€168.000,00	
Cost of Goods Sold	€0	
Gross Margin	€ 168.000,00	
Expenses	€315.000,00	
Amortizations	€1.333,33	
Total	€316.333,33	
Operational Margin	€- 148.333,33	
Financial Expenses	€11.480,00	
Profit Before Tax	€- 159.813,33	
Tax	€0	
Net Profit	€- 159.813,33	
Pro Memoria		
Gross Cash Flow	€-158.480,00	PBT+Amort
Net Cash Flow	€-158.480,00	NP+Amort
EBITDA	€-147.000,00	NP+Tax+FE+Amort

We are looking for



€150,000

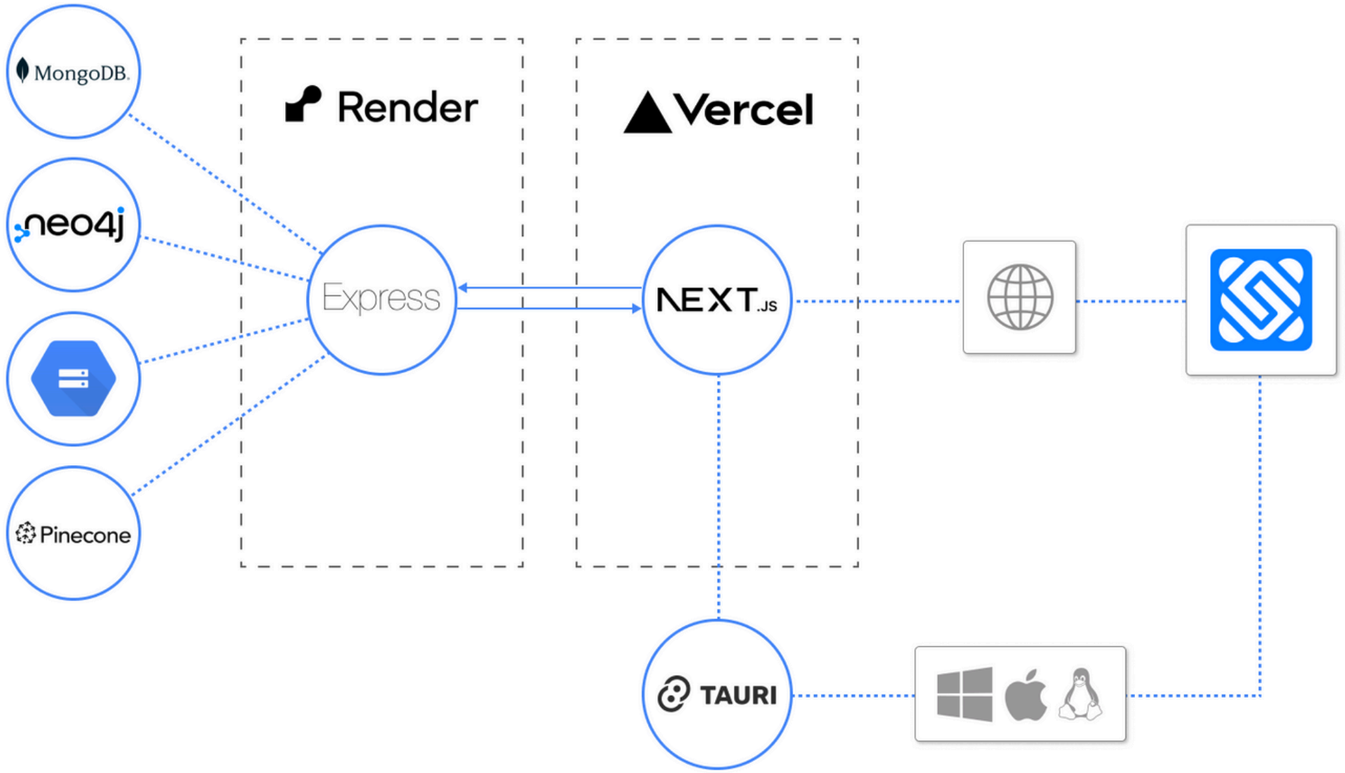
For

10% Equity

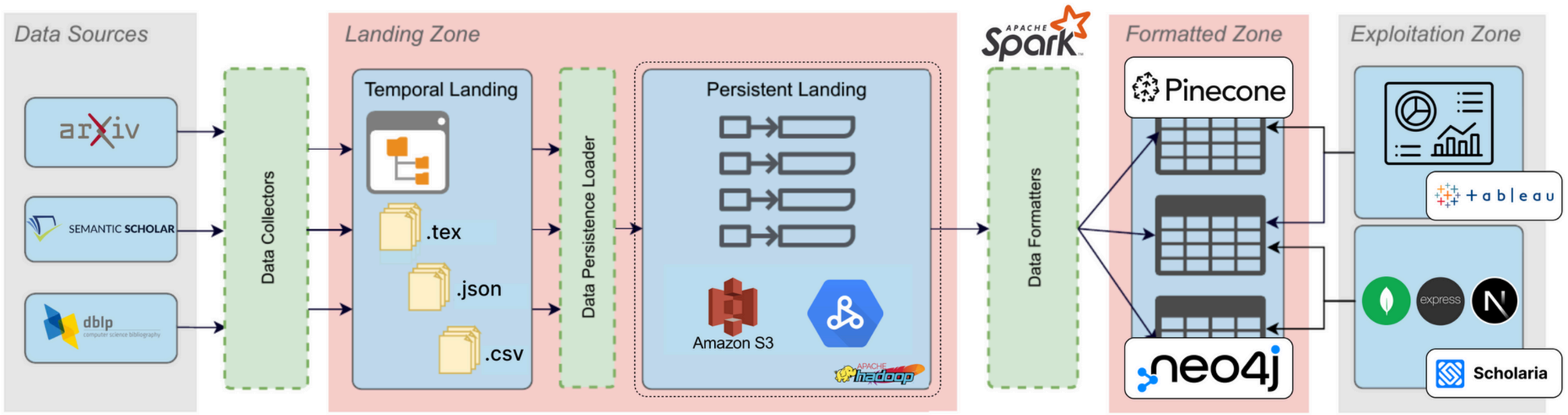
Scholaria: Demo

Tech & Implementation

FUNCTIONAL ARCHITECTURE



BDM PIPELINE



DATA SOURCES

arXiv supports the OAI protocol for metadata harvesting to provide access to metadata for all articles, updated daily with new articles.

DBLP Computer Science bibliography is an open library providing bibliographic information on major computer science journals and proceedings

Semantic Scholar's Graph API provides a free and on-demand source of data about authors, papers, citations, venues, and more utilizing its AI-powered academic search engine

LANDING ZONE

Google Dataproc is a fast, easy-to-use, fully managed cloud service for running Apache Spark and Apache Hadoop clusters

S3 is a cloud storage solution provided by **AWS**. Datalakes require additional features such as catalog management on top of a storage place.

Data was stored and distributed under different formats:
.CSV & .JSON: In Parquet Format
.PDF: As raw files

FORMATTED ZONE

Graph database management system used to store nodes and relationships about papers metadata, authors, keywords, categories

Vector database platform used to store papers' SPECTER2 Embeddings, and to query recommendations.

EXPLOITATION ZONE

Search

The graph serves as a backbone to our application search

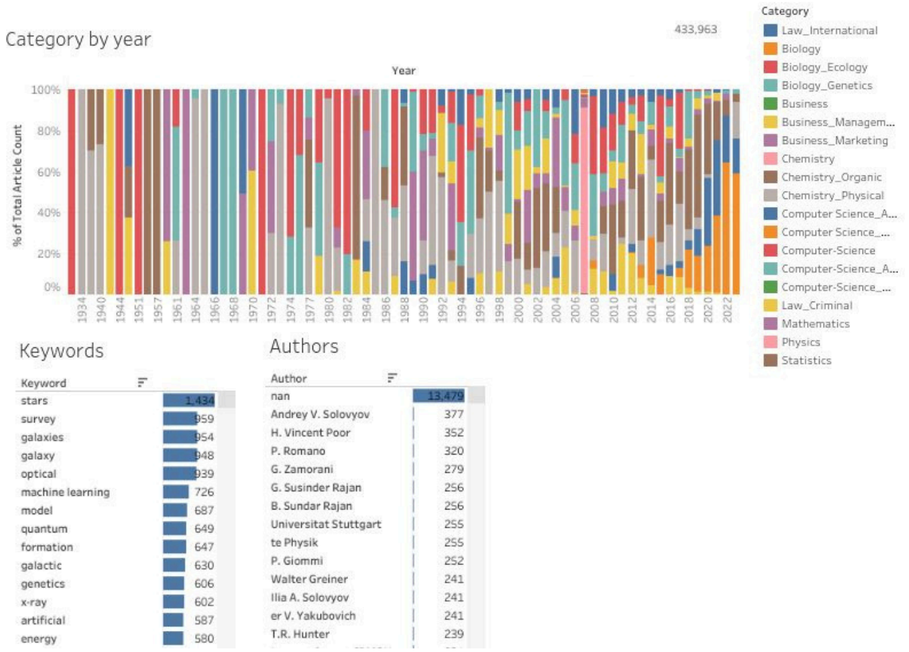
- Textual Data
- Relationships with Categories and Keywords
- Relationships with Authors

EXPLOITATION ZONE

Analytics

We leveraged the graph to extract insights related to our data and in order to visualize it we are using **Tableau**.

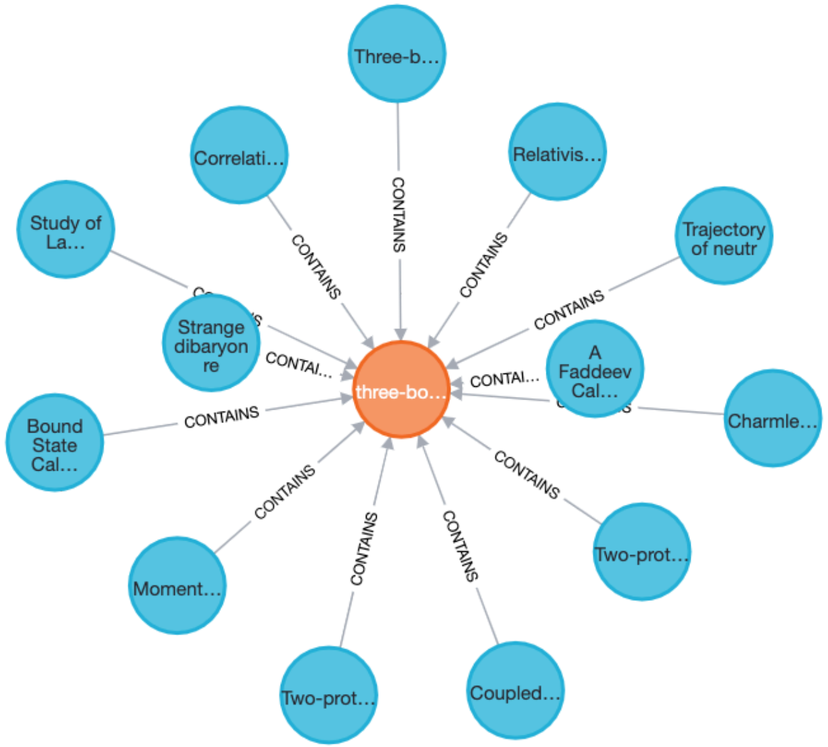
- Total number of papers
- Distribution of papers across different fields
- Distribution of papers across the years
- Statistics about authors works and collaborations
- Ranking of most important keywords



EXPLOITATION ZONE

Recommendation system requirements

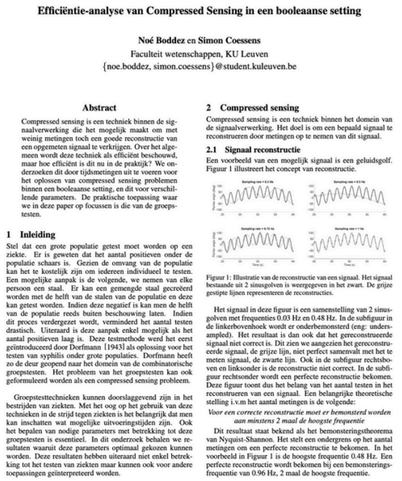
- Based on papers that the user uploads we want to recommend new papers
- This has to be done efficiently and correctly



EXPLOITATION ZONE

Vector embeddings

- Each paper is mapped to a vector
- Then stored in a vector database
- Recommendation retrieval done in O(n.d) time
  (d = number of dimensions)



Efficiëntie-analyse van Compressed Sensing in een booleaanse setting

Noé Boddez en Simon Coessens
Faculteit wetenschappen, KU Leuven
{noe.boddez, simon.coessens}@student.kuleuven.be

Abstract

Compressed sensing is een techniek binnen de signaalverwerking die het mogelijk maakt om met weinig metingen toch een goede reconstructie van een opgemeten signaal te verkrijgen. Over het algemeen wordt deze techniek als efficiënt beschouwd, maar hoe efficiënt is dit nu in de praktijk? We onderzoeken dit door tijdsmetingen uit te voeren voor het oplossen van compressed sensing problemen binnen een booleaanse setting, en dit voor verschillende parameters. De praktische toepassing waar we in deze paper op focussen is die van de groepstesten.

1 Inleiding

Stel dat een grote populatie getest moet worden op een ziekte. Er is geweten dat het aantal positieven onder de populatie schaars is. Gezien de omvang van de populatie kan het te kostelijk zijn om iedereen individueel te testen. Een mogelijke aanpak is de volgende, we nemen van elke persoon een staal. Er kan een gemengde staal gecreëerd worden met de helft van de stalen van de populatie en deze kan getest worden. Indien deze negatief is kan men de helft van de populatie reeds buiten beschouwing laten. Indien dit proces verdergezet wordt, verminderd het aantal testen drastisch. Uiteraard is deze aanpak enkel mogelijk als het aantal positieven laag is. Deze testmethode werd het eerst geïntroduceerd door Dorfmann [1943] als oplossing voor het testen van syphilis onder grote populaties. Dorfmann heeft zo de deur geopend naar het domein van de combinatorische groepstesten. Het probleem van het groepstesten kan ook geformuleerd worden als een compressed sensing probleem.

Groepsteststechnieken kunnen doorslaggevend zijn in het bestrijden van ziekten. Met het oog op het gebruik van deze technieken in de strijd tegen ziekten is het belangrijk dat men kan inschatten wat mogelijke uitvoeringstijden zijn. Ook het bepalen van nodige parameters met betrekking tot deze groepstesten is essentieel. In dit onderzoek behalen we resultaten waaruit deze parameters optimaal gekozen kunnen worden. Deze resultaten hebben uiteraard niet enkel betrekking tot het testen van ziekten maar kunnen ook voor andere toepassingen geïnterpreteerd worden.

2 Compressed sensing

Compressed sensing is een techniek binnen het domein van de signaalverwerking. Het doel is om een bepaald signaal te reconstrueren door metingen op te nemen van dit signaal.

2.1 Signaal reconstructie

Een voorbeeld van een mogelijk signaal is een geluidsgolf. Figuur 1 illustreert het concept van reconstructie.

Figuur 1: Illustratie van de reconstructie van een signaal. Het signaal bestaande uit 2 sinusgolven is weergegeven in het zwart. De grijze gestipte lijnen representeren de reconstructies.

Het signaal in deze figuur is een samenstelling van 2 sinusgolven met frequenties 0.03 Hz en 0.48 Hz. In de subfiguur in de linkerbovenhoek wordt er onderbemonsterd (eng: undersampled). Het resultaat is dan ook dat het gereconstrueerde signaal niet correct is. Dit zien we aangezien het gereconstrueerde signaal, de grijze lijn, niet perfect samenvalt met het te meten signaal, de zwarte lijn. Ook in de subfiguur rechtsbovenven en linksonder is de reconstructie niet correct. In de subfiguur rechtsonder wordt een perfecte reconstructie bekomen. Deze figuur toont dus het belang van het aantal testen in het reconstrueren van een signaal. Een belangrijke theoretische stelling i.v.m het aantal metingen is de volgende:

Voor een correcte reconstructie moet er bemonsterd worden aan minstens 2 maal de hoogste frequentie

Dit resultaat staat bekend als het bemonsteringstheorema van Nyquist-Shannon. Het stelt een ondergrens op het aantal metingen om een perfecte reconstructie te bekomen. In het voorbeeld in Figuur 1 is de hoogste frequentie 0.48 Hz. Een perfecte reconstructie wordt bekomen bij een bemonsteringsfrequentie van 0.96 Hz, 2 maal de hoogste frequentie.

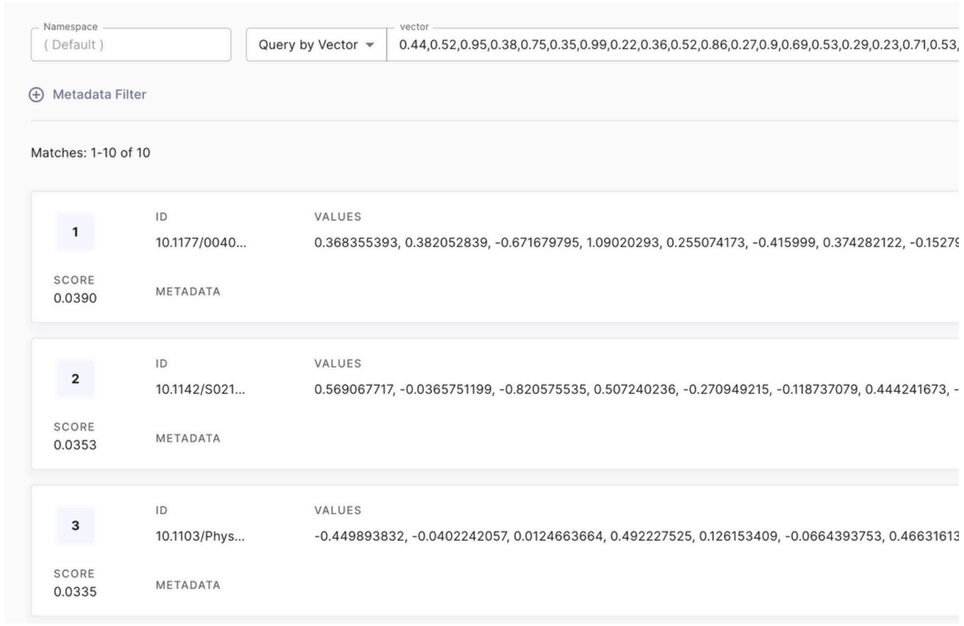
$$\mathbf{x} = \begin{bmatrix} 3.14 \\ -1.57 \\ 0.56 \\ 2.65 \\ -0.23 \\ 1.76 \\ -3.45 \\ 0.99 \\ -2.11 \\ 1.03 \\ 0.87 \\ -1.24 \\ 2.98 \\ -0.91 \\ 3.31 \\ -0.55 \\ 1.45 \\ -2.76 \\ 0.65 \\ -1.34 \end{bmatrix}$$

EXPLOITATION ZONE

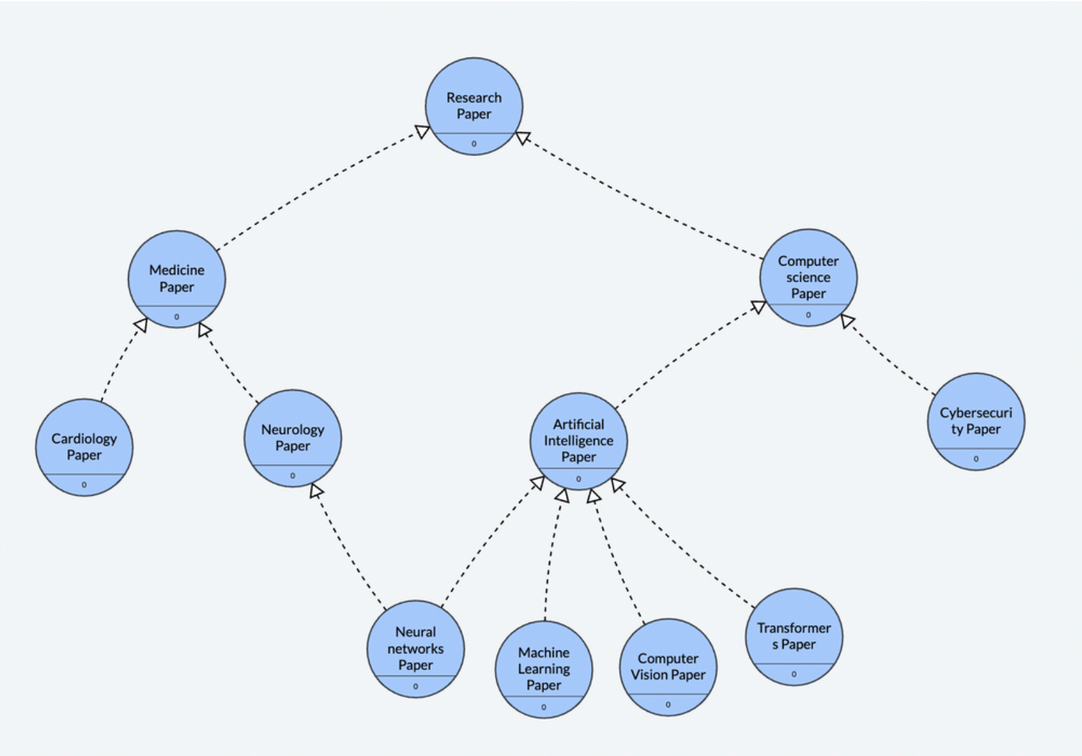
Recommendation System

- Store SPECTER2 Embeddings and create an index using cosine similarity for measuring vector closeness.
- When searching for recommendations based on multiple papers, we create a mean vector from the given papers.
- Query Pinecone Indexed Collection for recommendations based on this mean vector to get top K papers based on score. This uses K-nn.

Pinecone



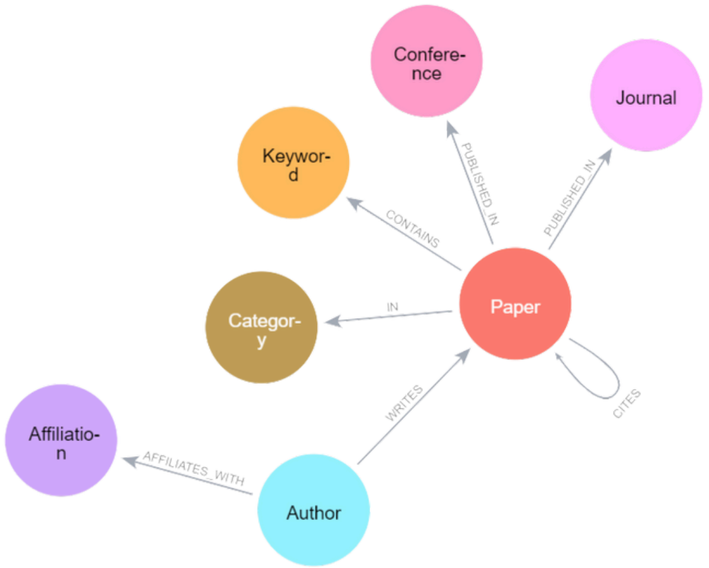
GRAPH REPRESENTATION OF THE DATA



Property Graphs vs Knowledge Graphs

- Knowledge graphs are standardized
- Knowledge graphs allow for complex representations of data
- Property graphs are quicker to implement and more efficient
- Final choice was made to implement a Property Graph

GRAPH REPRESENTATION OF THE DATA

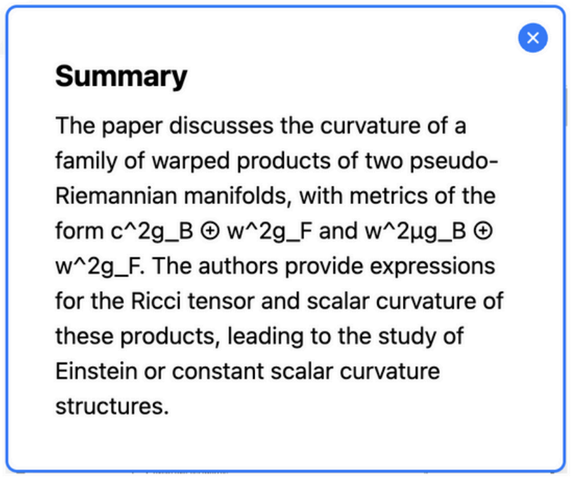


Property Graphs vs Knowledge Graphs

- Knowledge graphs are standardized
- Knowledge graphs allow for complex representations of data
- Property graphs are quicker to implement and more efficient
- Final choice was made to implement a Property Graph

USE OF ARTIFICIAL INTELLIGENCE

Scholaria AI



Considerations

- Currently using OpenAI's GPT3-turbo model
- For copyright reasons in the future we will switch to an open-source model
- For example BERT

READY TO ANSWER QUESTIONS

Thank you!