



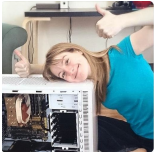
Searching for Research Fraud in OpenAlex with Graph Data Science

Adam Day - SAGE Publishing

Ebru Cucen - OpenCredo



OC - Sage R&D Project



Fahran Wallace



Sebastian Margineanu



Helen King



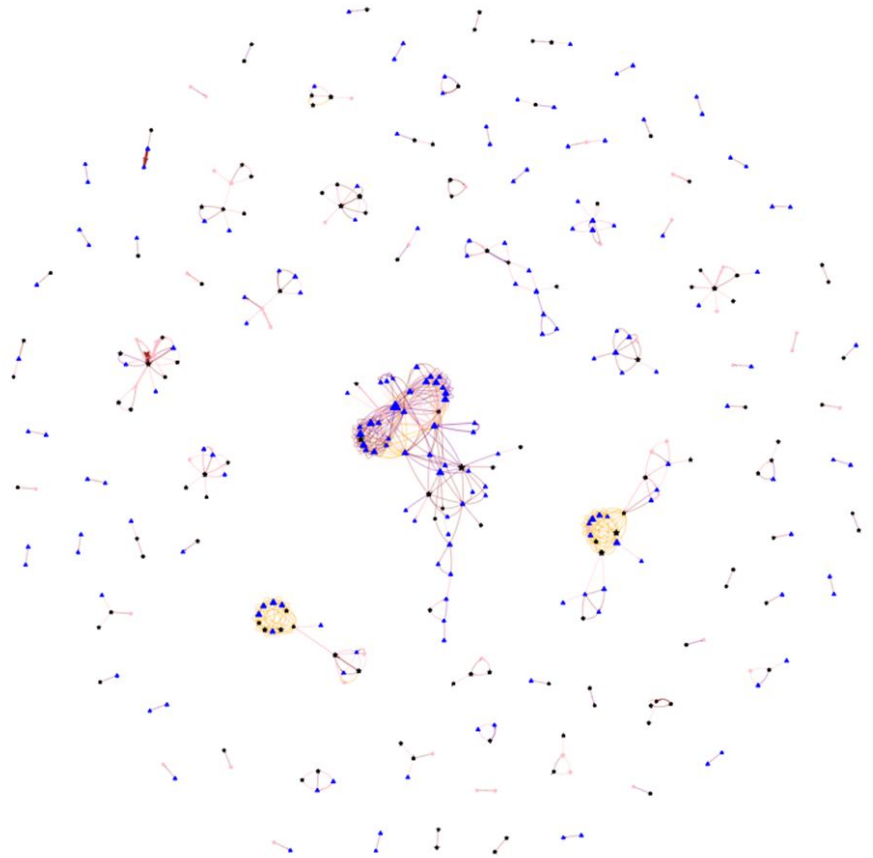
Key Takeaways

1. Graph analytics is a good approach to identify organised misconduct
2. To do this efficiently, we need reliable data structures, such as a knowledge graph built on well-curated data
 - OpenAlex is a good data set for this
3. We can detect unusual co-authorships. Either with machine-learning, or a simple rule:
 - It is unusual for an author to co-author with a different institution on their first paper

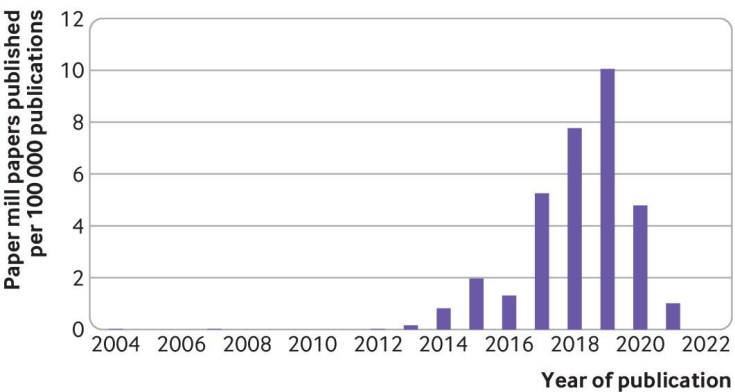
The value of graphs

Imagine what happens when we link these accounts to:

- Authors who recommended them
- Coauthors
- Past papers....



The different kinds of paper mills



Candal-Pedreira et al *BMJ* 2022; 379

DOI: <https://doi.org/10.1136/bmj-2022-071517>

(Published 28 November 2022)

AI isn't really helping!

- Already a number of AI-based tools to detect paper mills.
 - But these don't solve the problem on their own
- Already examples of paper mills using generated text from easy-to-detect models like GPT-2
 - Interestingly, NO examples of detections of GPT-3 or larger models
- Generating fake scientific images with Stable Diffusion is trivial
- Meta recently released a model *specifically for generating scientific text*.
 - Swiftly pulled!

AI BUSINESS







ML ▾

NLP ▾

Data ▾

Automation ▾

Verticals ▾

Responsible AI ▾

Companies ▾

More ▾



NLP

Language models

Meta

Update: Meta's Galactica AI Criticized as 'Dangerous' for Science

Model quickly pulled after renowned experts said results were 'statistical nonsense' and 'wrong'



Ben Wodecki
November 17, 2022

 3 Min Read

Latest News

Retraction Watch

Tracking retractions as a window
into the scientific process

**Exclusive: Russian site says it has
brokered authorships for more
than 10,000 researchers**



A company in Russia hawks its wares

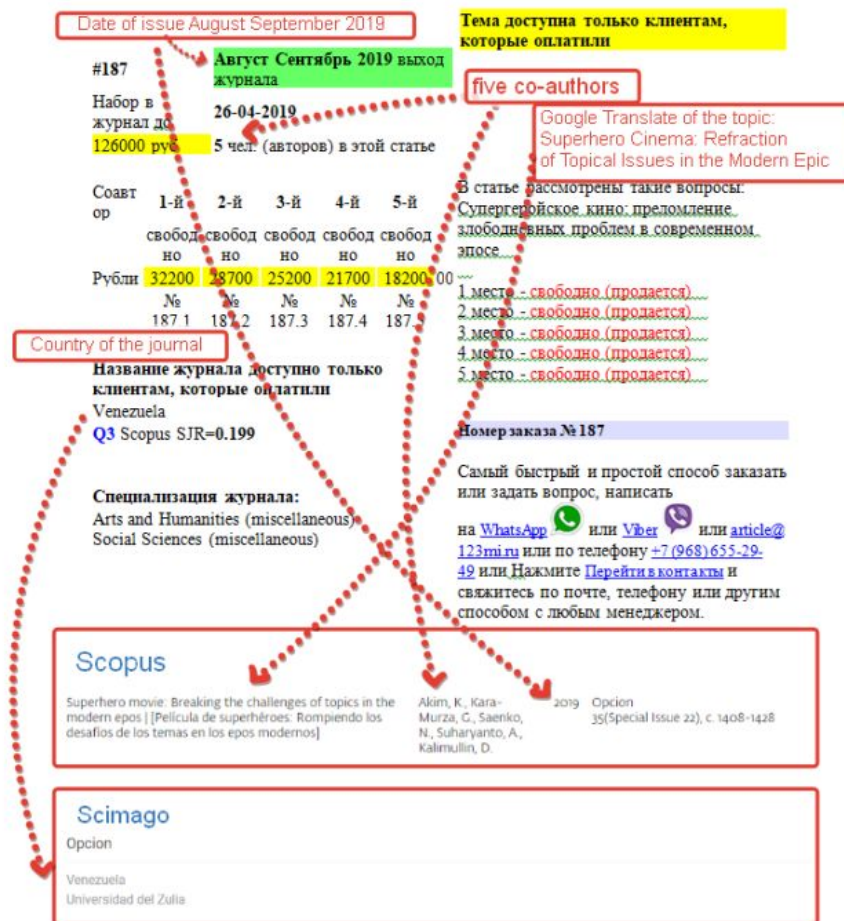
Example of the identification of the offer

← Tweet



Anna Abalkina
@AbalkinaAnna

This is my new preprint on @arxiv "Publication and collaboration anomalies in academic papers originating from a paper mill: evidence from Russia". At least 303 papers were identified. and a set of predictors of a Russian paper mill were proposed.
arxiv.org/abs/2112.13322



Questions:

- Can we take the publicly-available data for these papers and model the co-authorships?
- Can we work out when a co-authorship is unusual and worth looking into more closely?

Data



Data



- Free, Hundreds of millions of entities
- Inspired by the ancient Library of Alexandria
- Launched at Sep '22
- Updated regularly
- 4 ways to access:
 - Web
 - API for ad-hoc queries
 - AWS Snapshot : 1.7 TB
 - Snowflake Marketplace

 Authors : people who create works

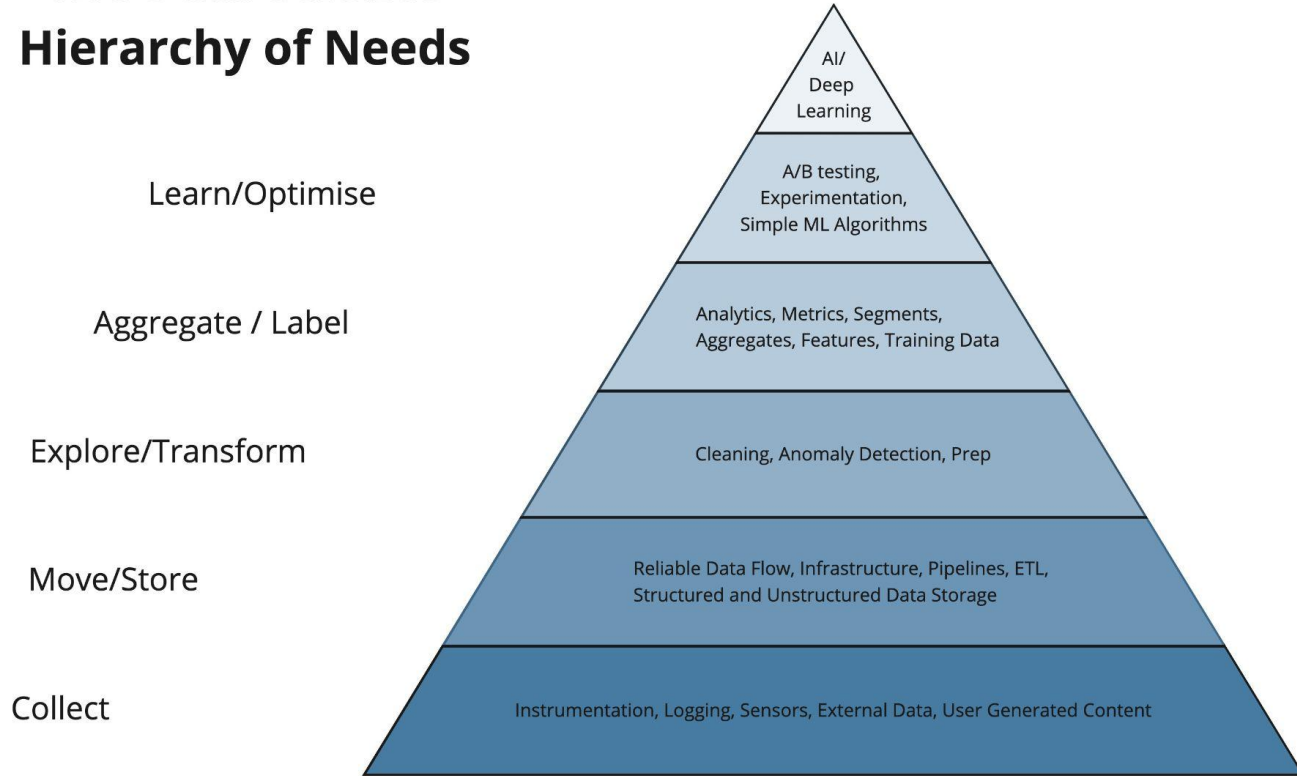
 Works : papers, books, datasets, etc; they cite other works

 Concepts : keywords associated with works

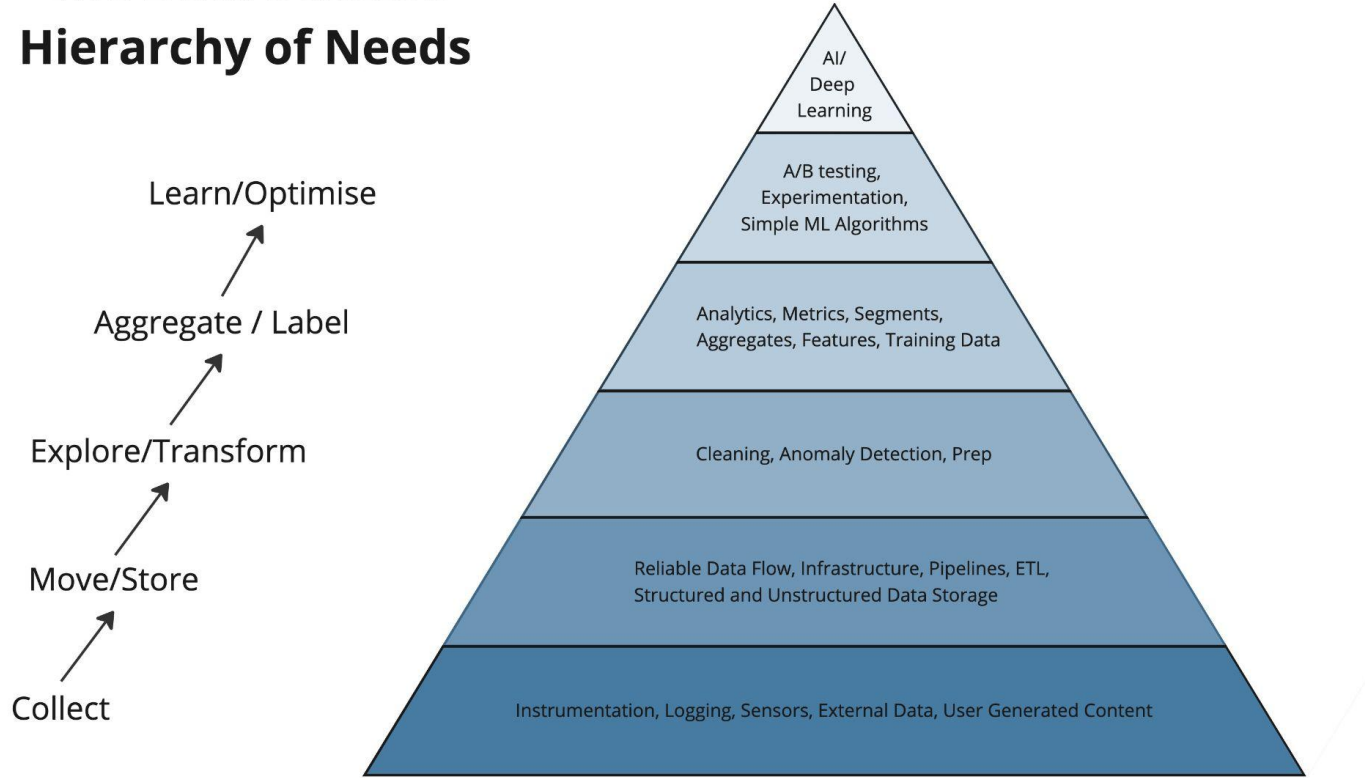
 Venues : journals and repositories that host works

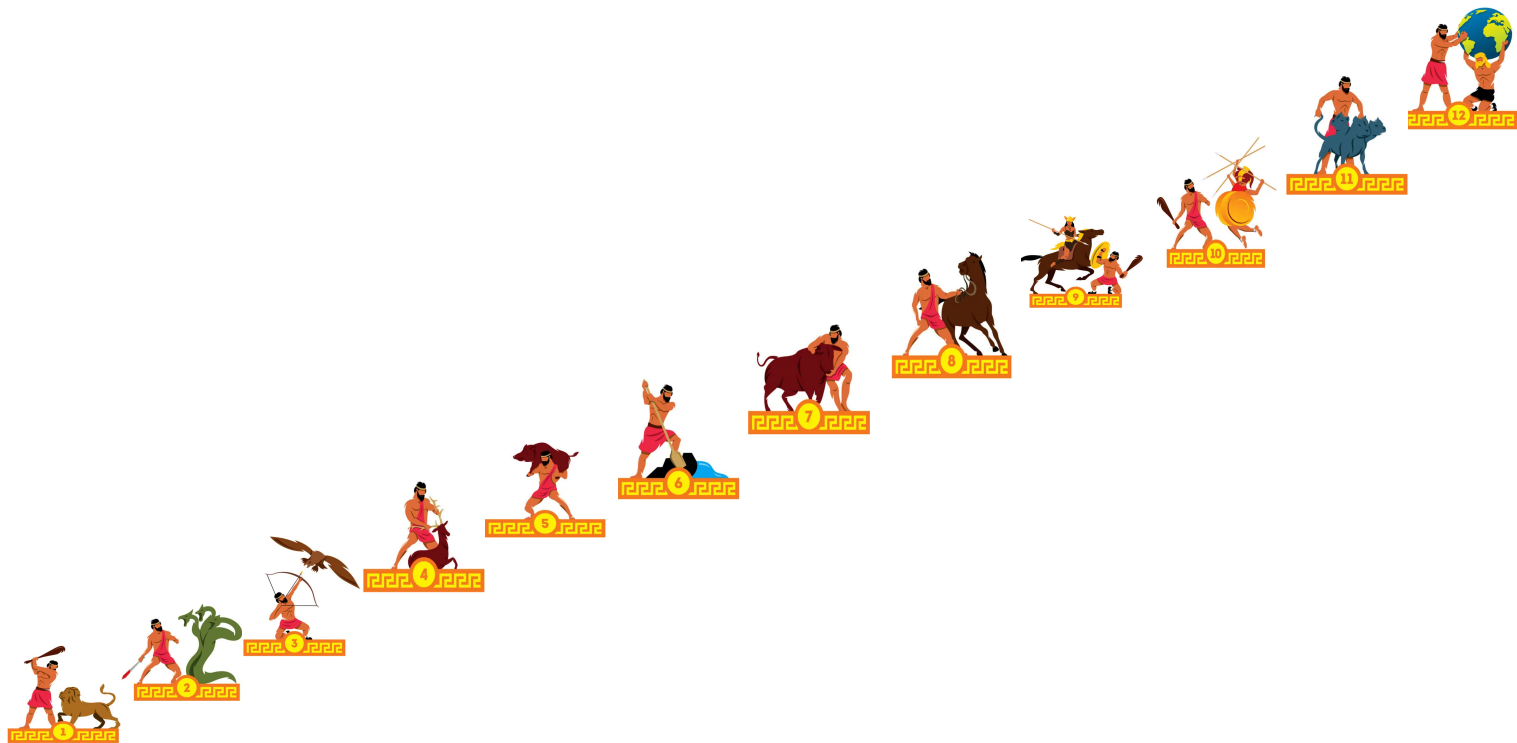
 Institutions : universities and other orgs that are affiliated with works (via authors)

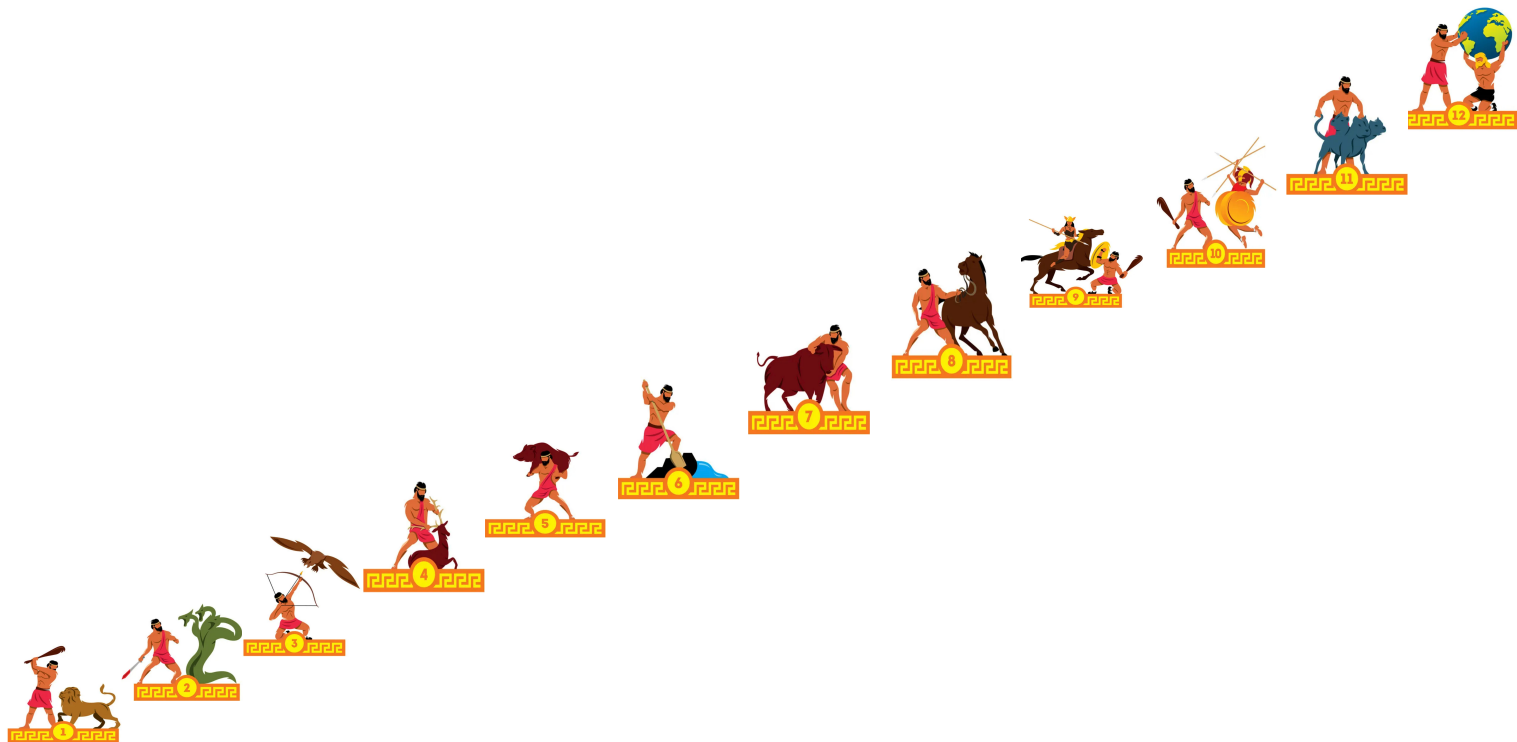
The Data Science Hierarchy of Needs



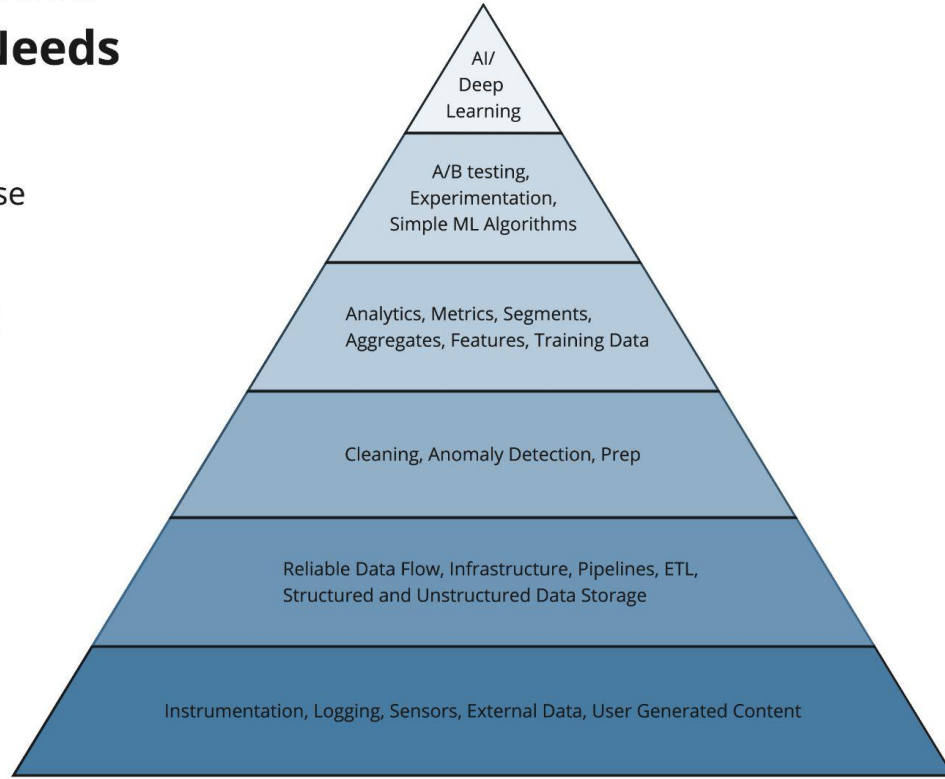
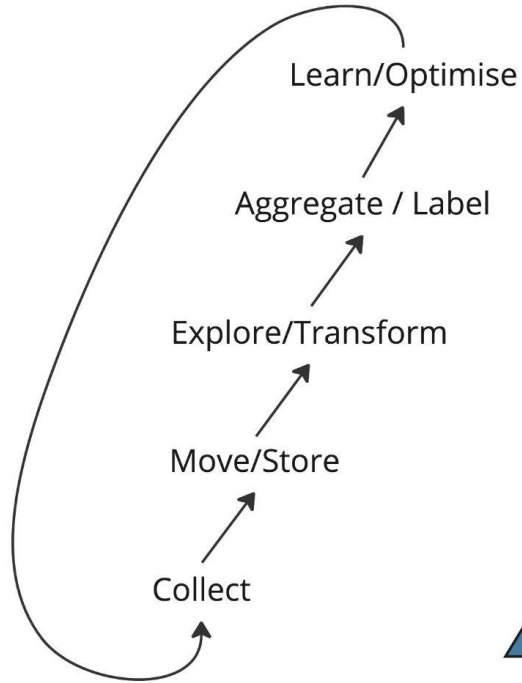
The Data Science Hierarchy of Needs

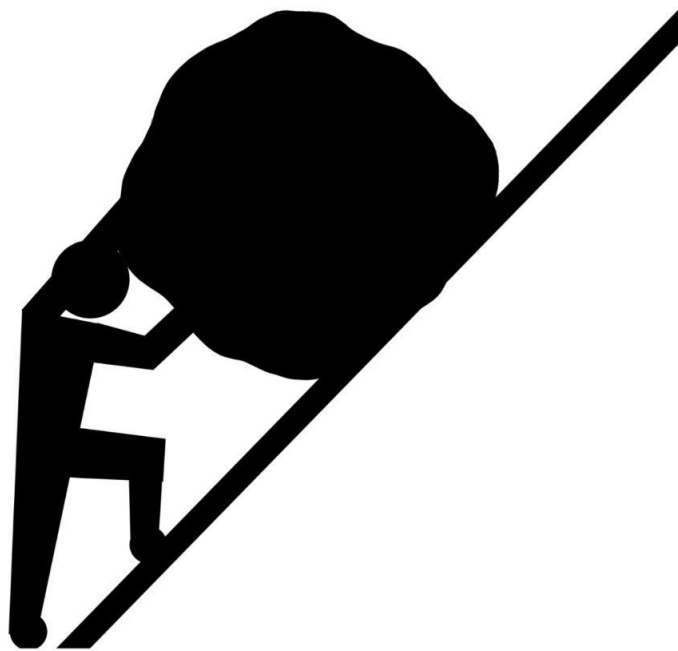
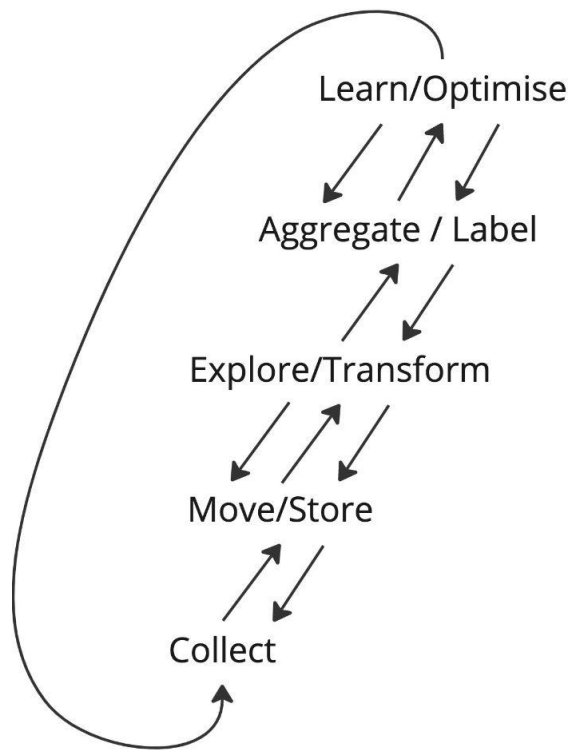




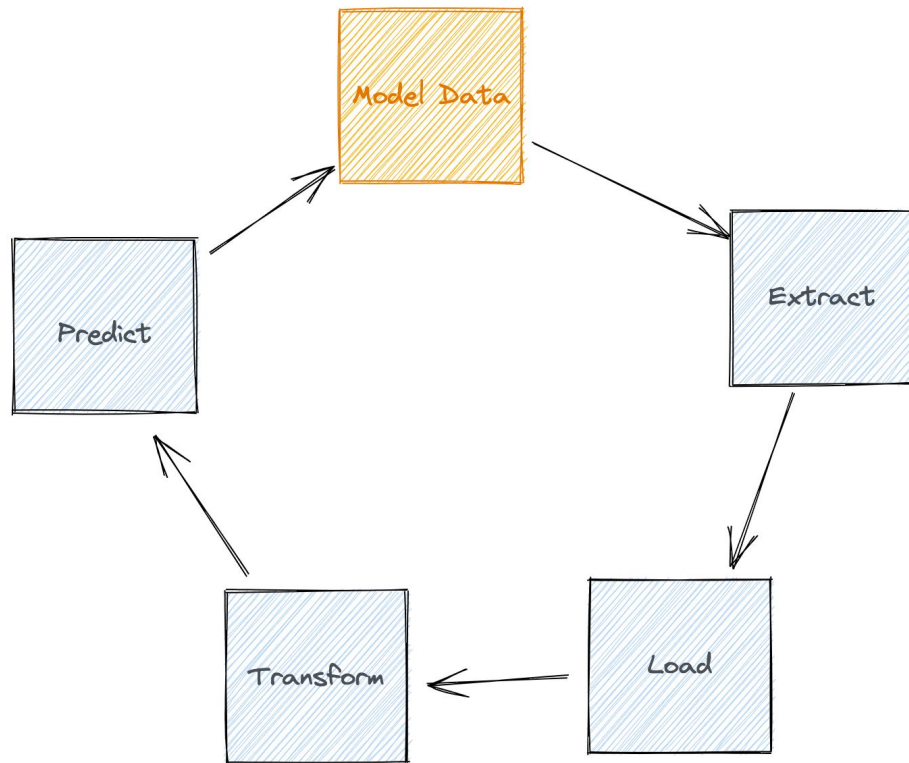


The Data Science Hierarchy of Needs





Data LifeCycle



Concepts

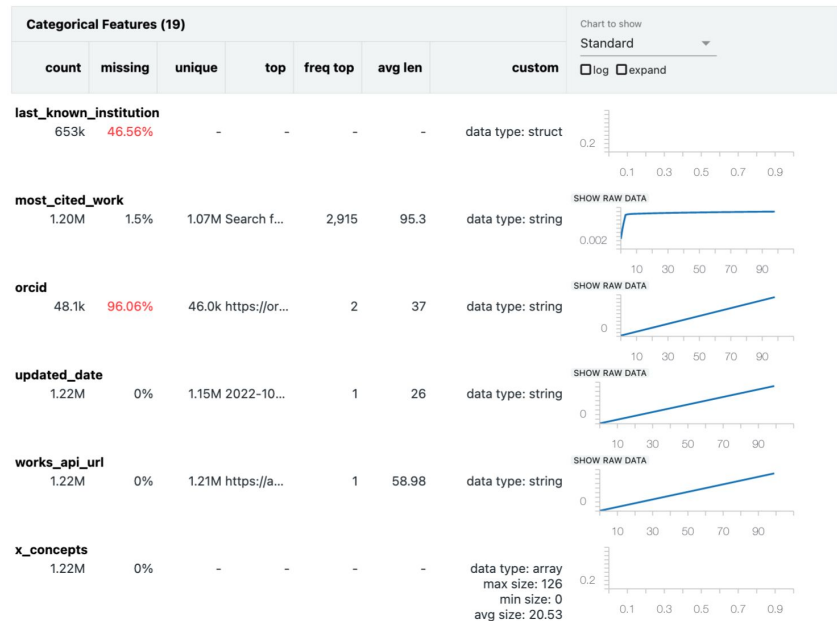
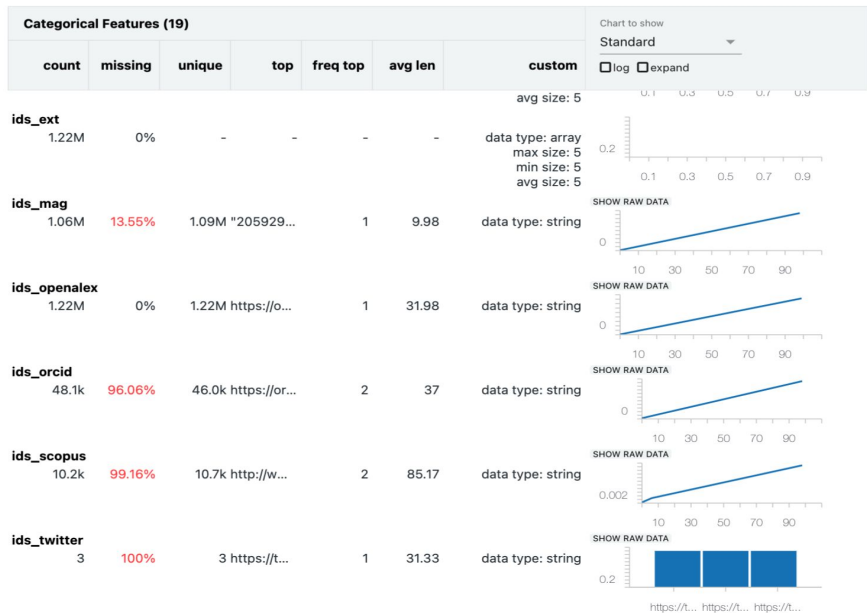
```
1  {'ancestors': [],
2   'cited_by_count': 391428928,
3   'counts_by_year': [{'cited_by_count': 26388154, -
4   'created_date': '2016-06-24',
5   'description': 'theoretical study of the formal foundation enabling the ' -
6   'display_name': 'Computer science',
7   'id': 'https://openalex.org/C41008148',
8   'ids': {'mag': '41008148',
9   'openalex': 'https://openalex.org/C41008148',
10  'uris_cui': ['C4099726'],
11  'wikidata': 'https://www.wikidata.org/wiki/Q21198',
12  'wikipedia': 'https://en.wikipedia.org/wiki/Computer%20science'},
13  'image_thumbnail_url': 'https://upload.wikimedia.org/wikipedia/commons/thumb/6/6a/Sorting_quicksort_anim.gif/180px-Sorting_quicksort_anim.gif',
14  'image_url': 'https://upload.wikimedia.org/wikipedia/commons/6/6a/Sorting_quicksort_anim.gif',
15  'international': {'description': {'ar': 'دراسة الخوارزميات التي تتعامل مع البيانات',
16  'و التي يمكن تمثيلها كجوانبات في شكل',
17  'والترتيب',
18  'az': 'elə sahəsi',
19  'bn': 'সম্পর্কিত ধর্ম সম্বন্ধে আলোচনা ' -
20  'bs': 'naučna disciplina koja se bavi svim ' -
21  'ca': 'estudi teòric de la base formal que ' -
22  'cs': 'věda studující principy strojového ' -
23  'da': 'disciplin med rødder i matematik ' -
24  'de': 'Wissenschaft von der systematischen ' -
25  'en': 'theoretical study of the formal ' -
26  'en-gb': 'study of the theoretical ' -
27  'eo': 'scienco pri teoria fonda de ' -
28  'es': 'ciencia dedicada a la computación y ' -
29  'fa': 'مطالعه الالگوات و پردازش اطلاعات و ' -
30  'fi': 'tieteenala, joka tutkii ' -
31  'fr': 'activité scientifique, technique, et ' -
32  'gl': 'ciencia dedicada á computación e ao ' -
33  'he': 'מדע המחשב', 'המדע של המידע', 'מדע המחשב',
34  'hsb': 'wědomosć wo systematiskim ' -
35  'id': 'kajian dasar teoretis dari informasi ' -
36  'it': 'scienza che studia i fondamenti ' -
37  'ja': '情報と計算の理論的基礎、及びそのコンピュータ上への実装と応用に関する研究分野',
38  'kn': 'ಸಾಕಾ ಸಂಖ್ಯೆ ಸೂತ್ರದ ಅಧ್ಯಯನ ' -
39  'ko': '전산 이론, 소프트웨어의 운영을 연구하는 학제적 학제학의 한 ' -
40  'nb': 'fagfelt med røtter i matematisk ' -
41  'nl': 'de studie van informatie en ' -
42  'pl': 'nauka zajmująca się przetwarzaniem ' -
43  'pt': 'estudo dos fundamentos teóricos da ' -
44  'pt-br': 'estudo dos fundamentos teóricos ' -
45  'ro': 'ştiinţă care se ocupă cu prelucrarea ' -
46  'ru': 'racunnnnna oō wopopuauu',
47  'scn': 'scienza chi studia l'elaburazzioni ' -
48  'sco': 'scienteeific an practical approach ' -
49  'sl': 'veda o računalniški obdelavi ' -
50  'sq': 'studimi i teknologjisë së ' -
51  'tr': 'Belirli evren kurallarna dayalı, ' -
52  'uk': 'haykosa racunnnnna',
53  'vec': 'Sienza che la stùdia i fondaninti ' -
54  'vi': 'ngành học nghiên cứu về cơ sở lý ' -
55  'wa': 's'yince ki studeye les coplufreces',
56  'zh': '研究信息和計算的理論基礎'},
57  'display_name': {'ar': 'Informatika', -
58  'Level': 8,
59  'related_concepts': [{'display_name': 'Mathematics', -
60  'updated_date': '2022-11-29T07:48:02.45438',
61  'wikidata': 'https://www.wikidata.org/wiki/Q21198',
62  'works_api_url': 'https://api.openalex.org/works?filter=concepts&id=C41008148',
63  'works_count': 7657665}
```

Concepts

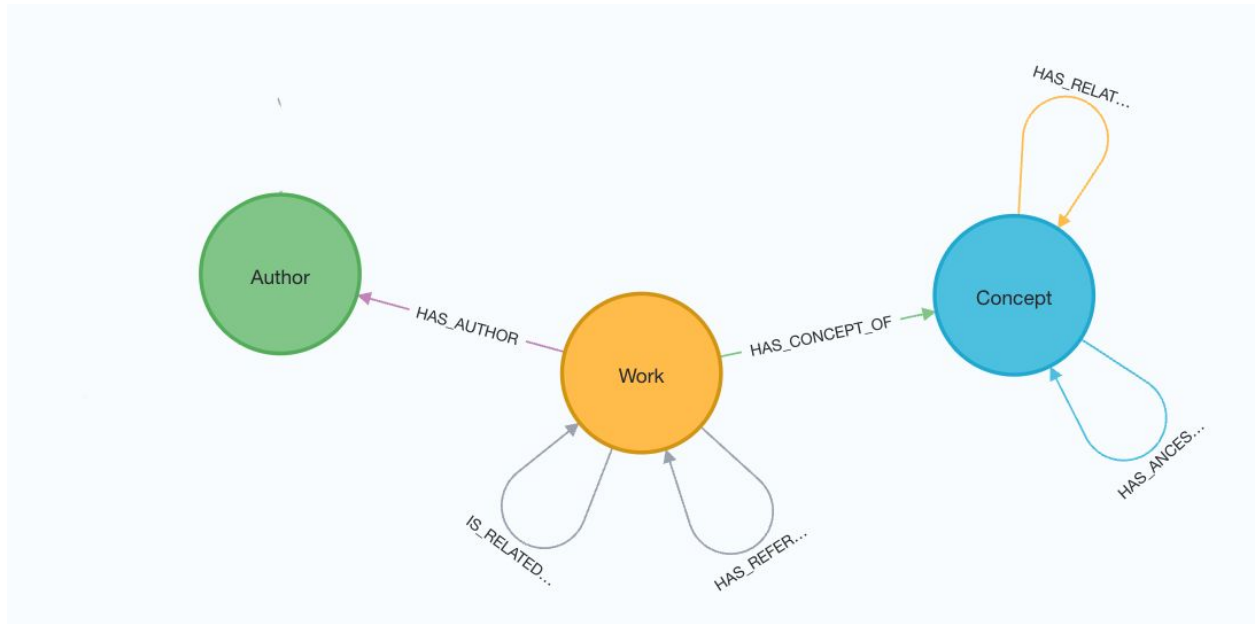
```
1 {^ancestors': [],
2 ^cited_by_count': 391428928,
3 ^counts_by_year': [{^cited_by_count': 26388154,-
36 ^created_date': '2016-06-24',
37 ^description': 'theoretical study of the formal foundation enabling the
40 ^display_name': 'Computer science',
41 ^id': 'https://openalex.org/C41008148',
42 ^ids': {'mag': '41008148',
43 ^openalex': 'https://openalex.org/C41008148',
44 ^uris_cui': ['C0099726'],
45 ^wikidata': 'https://www.wikidata.org/wiki/Q21198',
46 ^wikipedia': 'https://en.wikipedia.org/wiki/Computer%20science'},
47 ^image_thumbnail_url': 'https://upload.wikimedia.org/wikipedia/commons/thumb/6/6a/Sorting_quicksort_anim.gif/180px-Sorting_quicksort_anim.gif',
48 ^image_url': 'https://upload.wikimedia.org/wikipedia/commons/6/6a/Sorting_quicksort_anim.gif',
49 ^international': {'description': {'ar': 'دراسة الخوارزميات التي لتتضمن مع بيانيات
50 ' و التي يمكن تمثيلها كجوانبات في شكل
51 ' و'arabii',
52 ^az': 'ela sabsat',
53 ^bn': 'গণিতের নথি সংগ্রহ',
54 ^bs': 'naučna disciplina koja se bavi svim '-
55 ^ca': 'estudi teòric de la base formal que '-
56 ^cs': 'věda studující principy strojového '-
57 ^da': 'disciplin med rødder i matematik '-
58 ^de': 'Wissenschaft von der systematischen '-
59 ^en': 'theoretical study of the formal '-
60 ^en-gb': 'study of the theoretical '-
61 ^eo': 'scienco pri teoria fonda de '-
62 ^es': 'ciencia dedicada a la computación y '-
63 ^fa': 'مطالعه الالوارتم و پردازش الالوارتم و'
64 ^fi': 'tieteenala, joka tutkii '-
65 ^fr': 'activité scientifique, technique, et '-
66 ^gl': 'ciencia dedicada á computación e ao '-
67 ^he': 'מדע המחשב ופרדוקט המחשב'
68 ^hsb': 'wědomosć wo systematiskim '-
69 ^id': 'kajian dasar teoretis dari informasi '-
70 ^it': 'scienza che studia i fondamenti '-
71 ^ja': '情報と計算の理論の基礎、及びそのコンピュータ上への実装と応用に関する研究分野',
72 ^kn': 'ದೌಡಾ ದೌಡಾ ರೂಪದ ಸಂಗ್ರಹ',
73 ^ko': '전산 이론, 소프트웨어의 운영을 연구하는 학제적 연구의 한 '-
74 ^nb': 'fagfelt med retter i matematisk '-
75 ^nl': 'de studie van informatie en '-
76 ^pl': 'nauka zajmująca się przetwarzaniem '-
77 ^pt': 'estudo dos fundamentos teóricos da '-
78 ^pt-br': 'estudo dos fundamentos teóricos '-
79 ^ro': 'ştiinţă care se ocupă cu prelucrarea '-
80 ^ru': 'наука о вычислениях',
81 ^scn': 'scienza chi studia l'elabburazzioni '-
82 ^sco': 'scienteeif an practical approach '-
83 ^sl': 'veda o računalniški obdelavi '-
84 ^sq': 'studimi i teknologjise se '-
85 ^tr': 'Belirli evren kurallarına dayalı, '-
86 ^uk': 'наукова дисципліна',
87 ^vec': 'Sienza che la stùdia i fondaminti '-
88 ^vi': 'ngành học nghiên cứu về cơ sở lý '-
89 ^wa': 'syince ki studeye les copluftrees',
90 ^zh': '研究信息和計算的理論基礎'},
91 ^display_name': {'ar': 'Informatika',
92 ^level': 8,
93 ^related_concepts': [{'^display_name': 'Mathematics',-
94 ^updated_date': '2022-11-29T07:48:02.45438',
95 ^wikidata': 'https://www.wikidata.org/wiki/Q21198',
96 ^works_api_url': 'https://api.openalex.org/works?filter=concepts.id:C41008148',
97 ^works_count': 76576650}
```

Property	Value
cited_by_count	231013792
description	theoretical study of the formal foundation enabling the automated processing or computation of information, for example on a computer or over a data transmission network
display_name	Computer science
id	https://openalex.org/C41008148
image_thumbnail_url	https://upload.wikimedia.org/wikipedia/commons/thumb/6/6a/Sorting_quicksort_anim.gif/100px-Sorting_quicksort_anim.gif
image_url	https://upload.wikimedia.org/wikipedia/commons/6/6a/Sorting_quicksort_anim.gif
level	0
updated_date	2022-04-12
wikidata	https://www.wikidata.org/wiki/Q21198
works_api_url	https://api.openalex.org/works?filter=concepts.id:C41008148
works_count	27478212

Authors



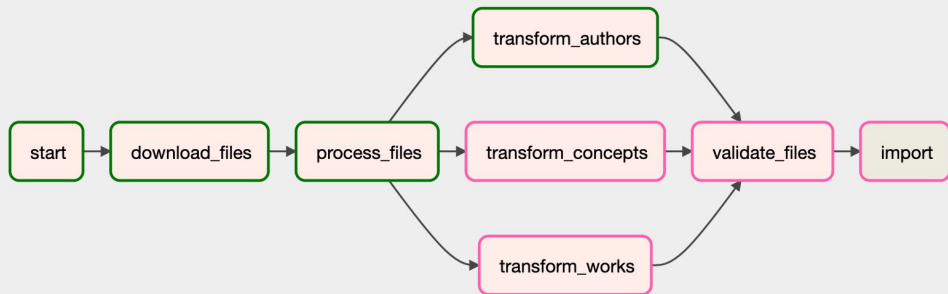
Modelling



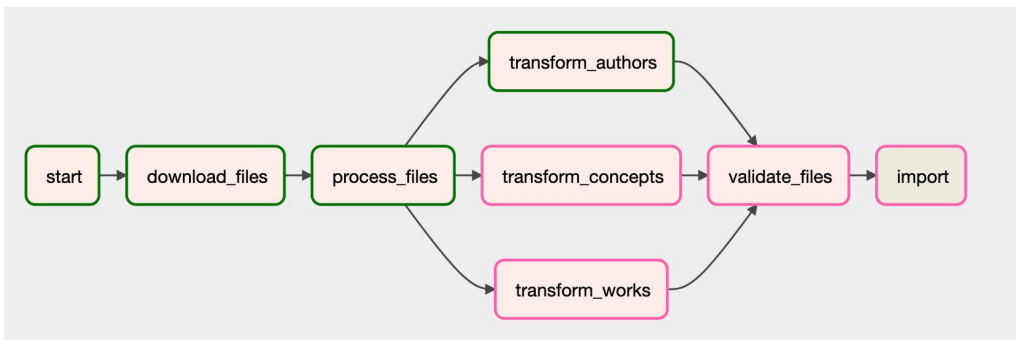
ETL / ELT

ELTTLTLTL ...

- Challenges
 - Database provisioning is easy
 - Ingestion is still expecting you transformation*
 - Storage is cheap
 - Not the database licensing model
 - Not the compute power you will need
 - Pipeline as code helps next iteration
 - Data is oil vs data should be curated



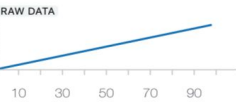
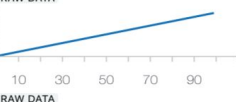
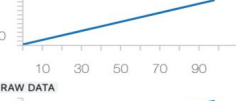
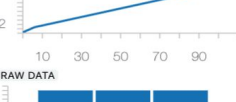
ETL/ ELT ... ETLTLTL



<https://datacreation.substack.com/p/why-data-contracts-are-obviously>



Categorical Features (19)							Chart to show	
count	missing	unique	top	freq top	avg len	custom	Standard	
							☐ log	☐ expand

							avg size: 5	
ids_ext								
1.22M	0%	-	-	-	-	data type: array		
							max size: 5	
							min size: 5	
							avg size: 5	
ids_mag								SHOW RAW DATA
1.06M	13.55%	1.09M	"205929...		1	9.98	data type: string	
ids_openalex								SHOW RAW DATA
1.22M	0%	1.22M	https://o...		1	31.98	data type: string	
ids_orcid								SHOW RAW DATA
48.1k	96.06%	46.0k	https://or...		2	37	data type: string	
ids_scopus								SHOW RAW DATA
10.2k	99.16%	10.7k	http://w...		2	85.17	data type: string	
ids_twitter								SHOW RAW DATA
3	100%	3	https://t...		1	31.33	data type: string	

Numeric Features (2)								Chart to show	
count	missing	mean	std dev	zeros	min	median	max	custom	
								☐ log ☐ expand	
cited_by_count									
1.22M	0%	94.75	715.61	15.97%	0	8	120k	data type: bigint	
works_count									
1.22M	0%	5.17	22.95	0%	1	1	2,548	data type: bigint	

Categorical Features (19)							Chart to show	
count	missing	unique	top	freq top	avg len	custom	Standard	
							☐ log ☐ expand	

last_known_institution						data type: struct		
653k	46.56%	-	-	-	-			
most_cited_work						data type: string		SHOW RAW DATA
1.20M	1.5%	1.07M	Search f...	2,915	95.3			
orcid						data type: string		SHOW RAW DATA
48.1k	96.06%	46.0k	https://or...	2	37			
updated_date						data type: string		SHOW RAW DATA
1.22M	0%	1.15M	2022-10...	1	26			
works_api_url						data type: string		SHOW RAW DATA
1.22M	0%	1.21M	https://a...	1	58.98			
x_concepts						data type: array max size: 126 min size: 0 avg size: 20.53		
1.22M	0%	-	-	-	-			
								☐ log ☐ expand
counts_by_year						data type: array max size: 11 min size: 0 avg size: 4.14		
1.22M	0%	-	-	-	-			
created_date						data type: string		SHOW RAW DATA
1.22M	0%	279	2016-06...	286k	10			
display_name						data type: string		SHOW RAW DATA
1.22M	0.01%	1.10M	John B. ...	183	13.33			
display_name_alternatives						data type: array max size: 25 min size: 0 avg size: 0.01		
1.22M	0%	-	-	-	-			
id						data type: string		SHOW RAW DATA
1.22M	0%	1.22M	https://o...	1	31.98			
ids						data type: struct		
1.22M	0%	-	-	-	-			

Similarity

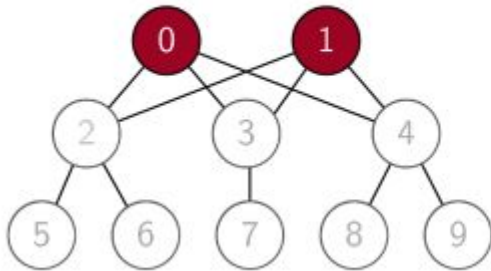
Node level similarity :

Regular

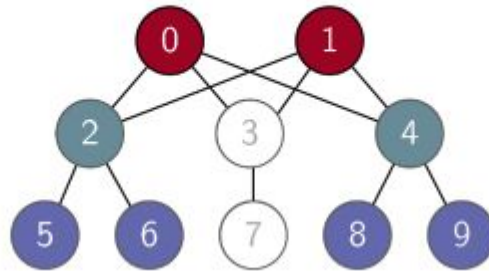
Edge level similarity

Automorphic

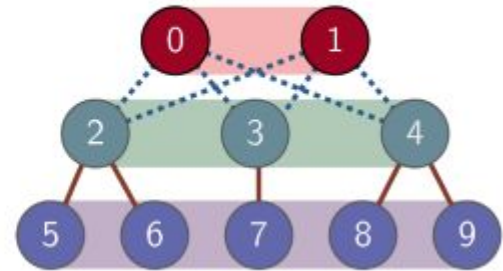
Structural

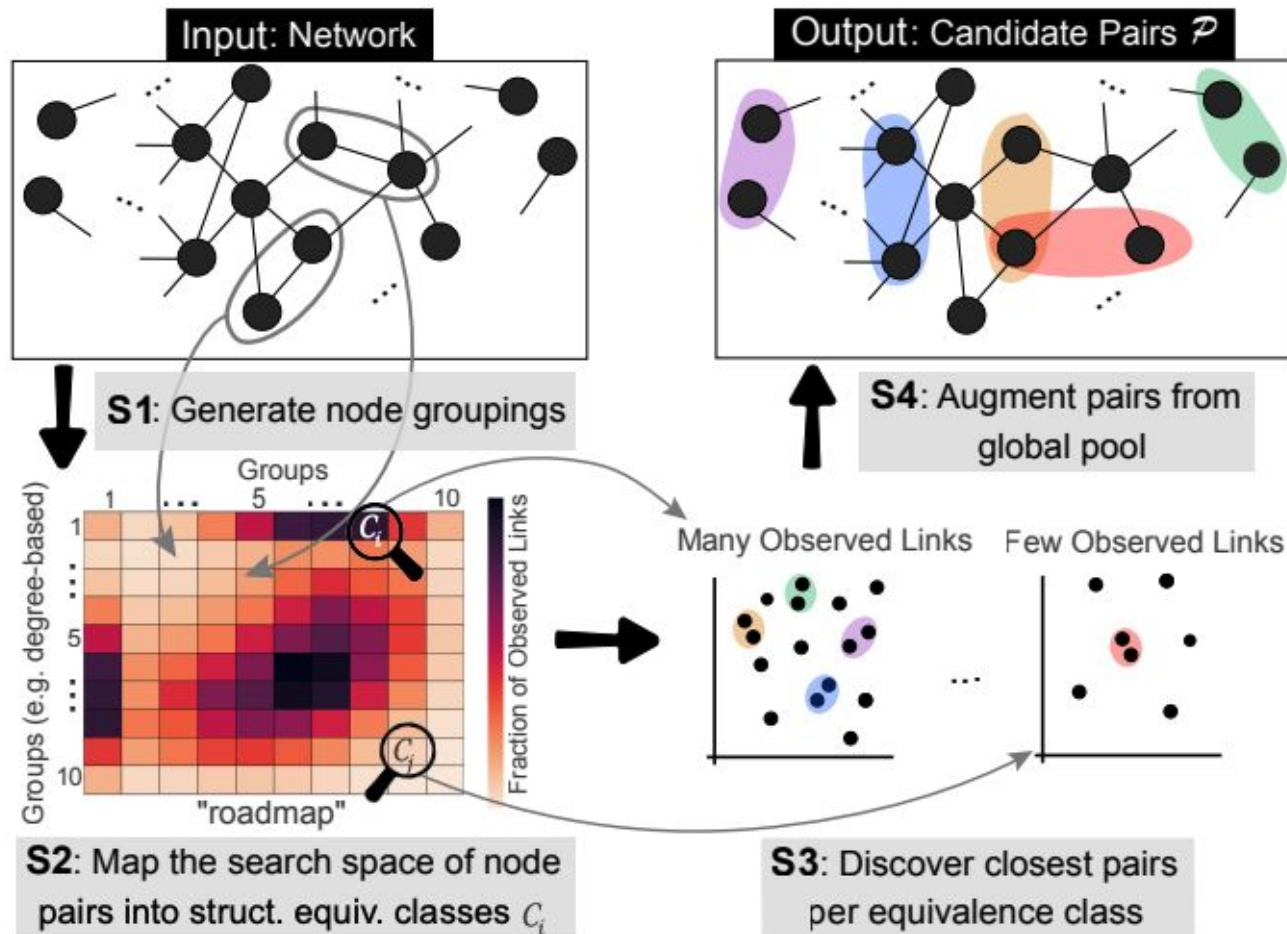


Automorphic



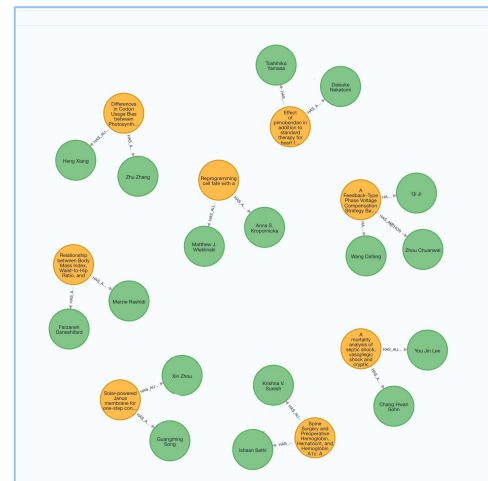
Regular





Co-authorship

```
MATCH p=(a:Author)←[:HAS_AUTHOR]-(w:Work)-[:HAS_AUTHOR]→(a2:Author)
WHERE a◇a2
RETURN p
```

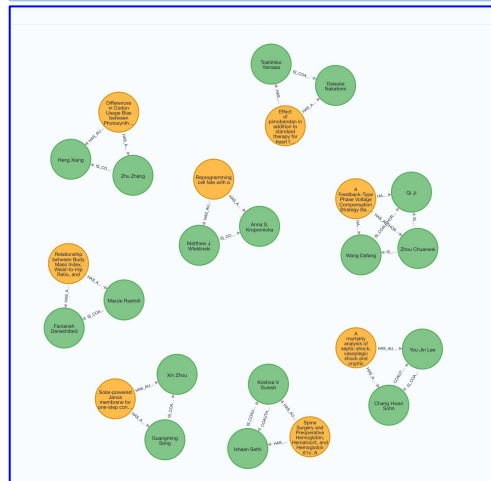
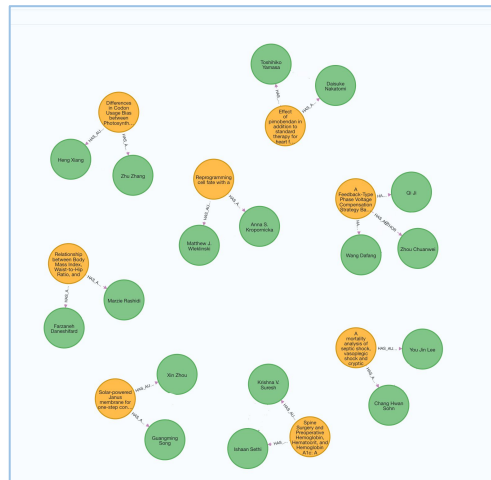


Co-authorship

```
MATCH p=(a:Author)←[:HAS_AUTHOR]-(w:Work)-[:HAS_AUTHOR]→(a2:Author)
WHERE a<>a2
RETURN p
```

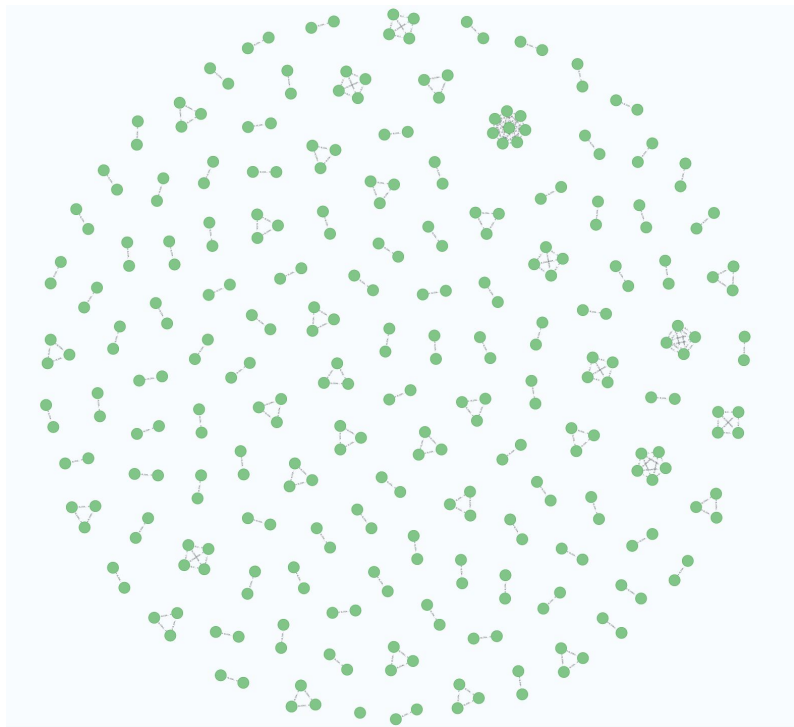
How many times have they co-authored before?

```
MATCH (a1:Author)←[:HAS_AUTHOR]-(work:Work)-[:HAS_AUTHOR]→(a2:Author)
WHERE id(a1) < id(a2)
WITH a1, a2, count(distinct work) as worksCount
CREATE (a1)-[:IS_COAUTHOR_WITH {works_count: worksCount}]->(a2)
```



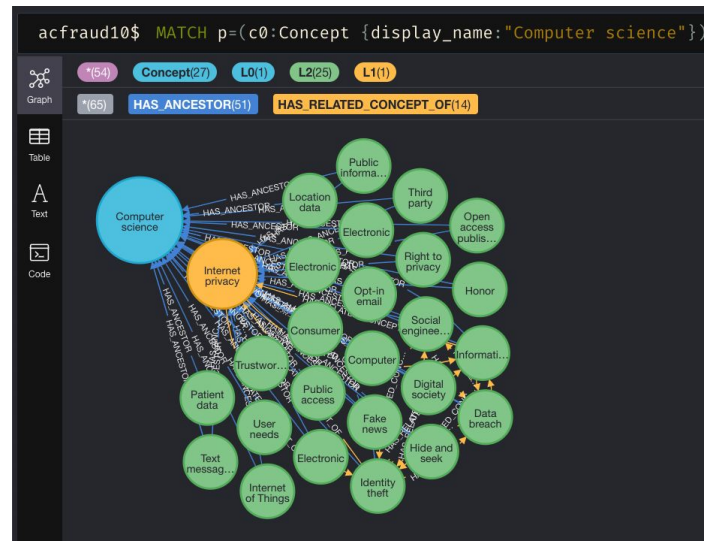
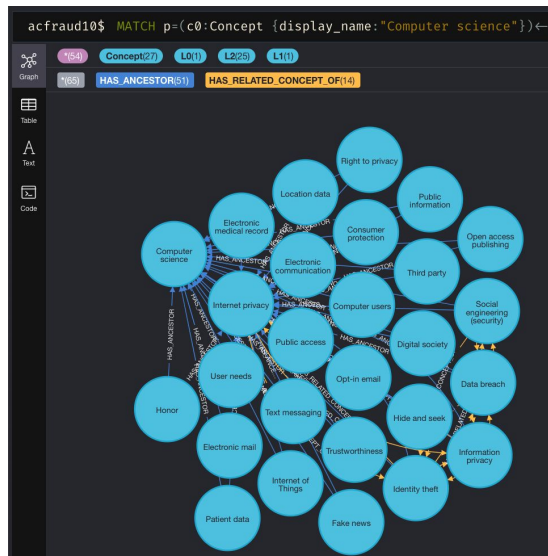
Subgraph

```
CALL gds.graph.project(  
  'coauthors',  
  'Author',  
  {IS_COAUTHOR_WITH: {  
    type: 'IS_COAUTHOR_WITH',  
    orientation: 'UNDIRECTED'  
  }})  
}}
```



New Labels

```
1 MATCH (c:Concept)
2 CALL apoc.create.addLabels (id(c), ['L'+c.level])
3 YIELD node
4 RETURN node
```



Concept Similarity

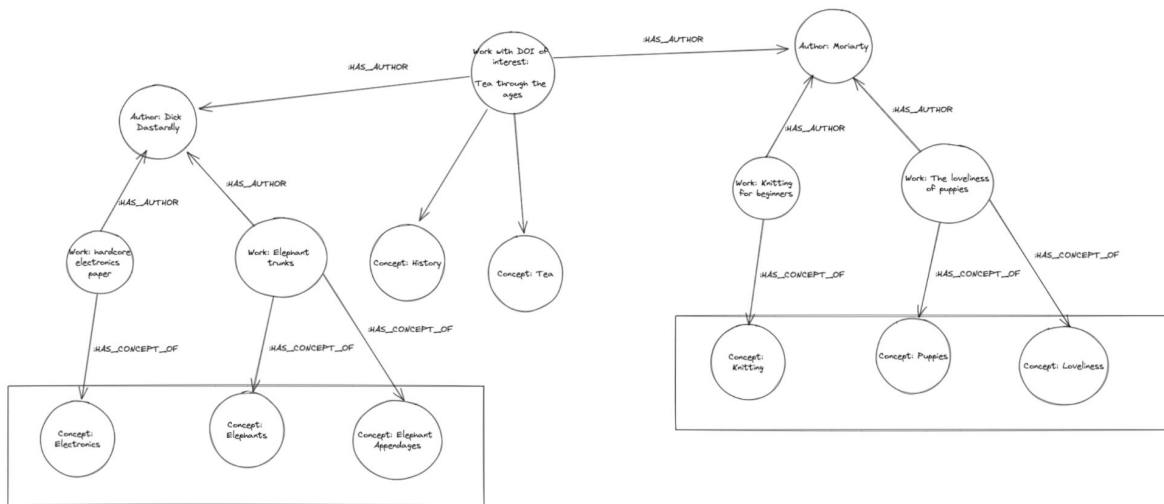
Explored the concepts of the work co-authors published

The similarity between the concepts gave us the idea how much concepts they share:

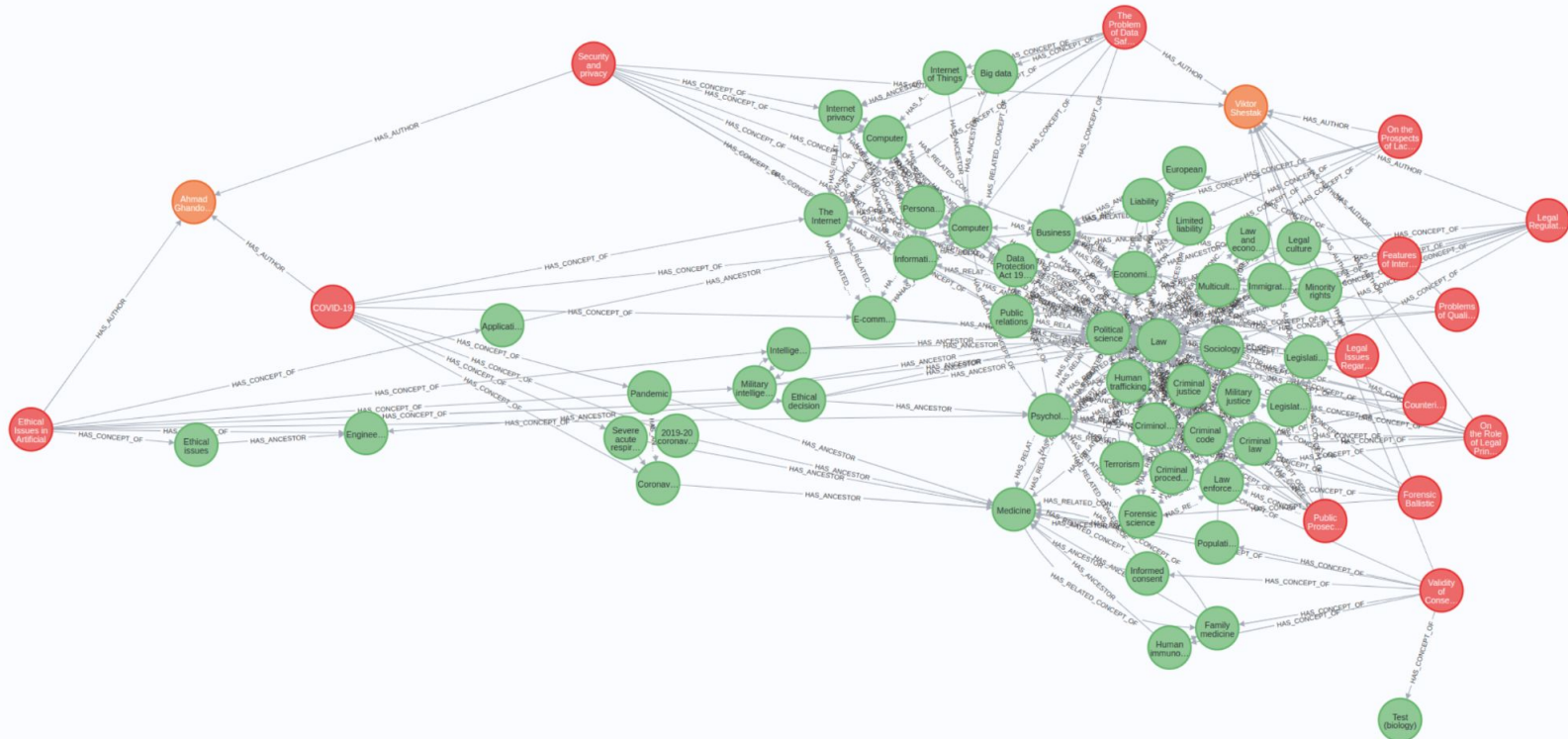
$$\text{Set}(\text{AuthorA}) \cap \text{Set}(\text{AuthorB})$$

$$\text{Set}(\text{AuthorA}) \cup \text{Set}(\text{AuthorB})$$

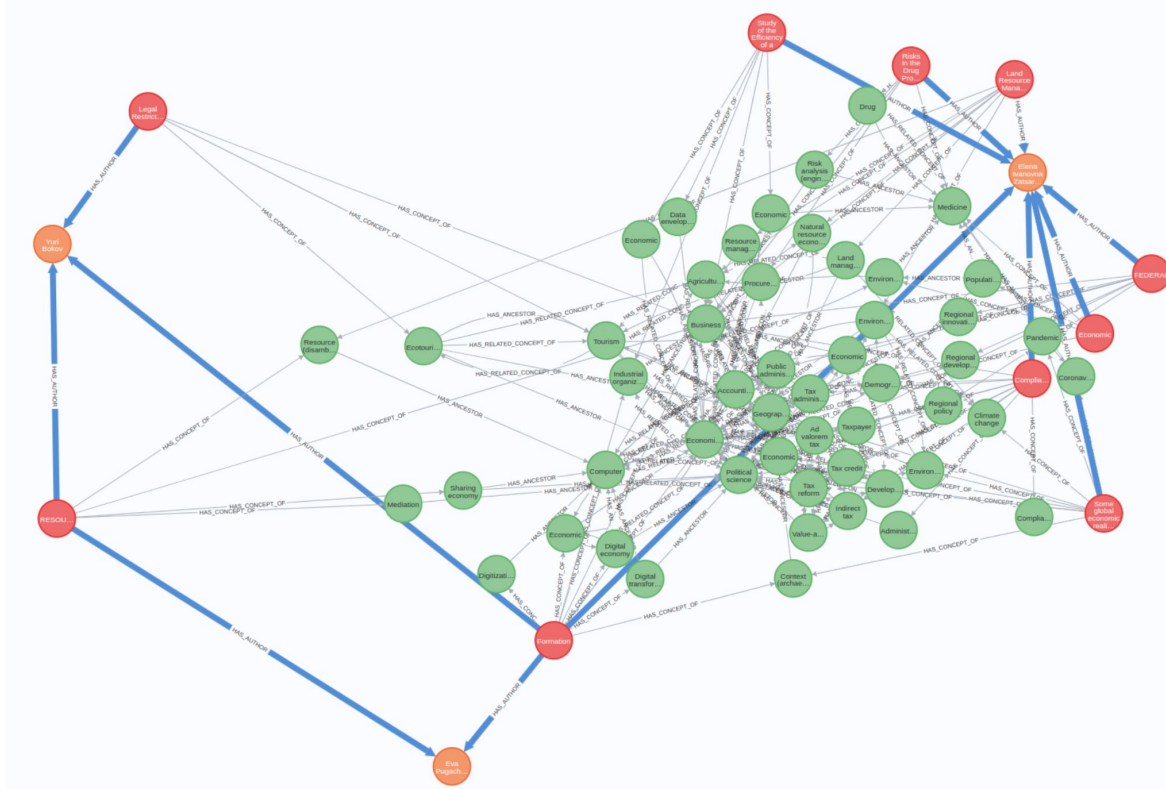
The example pic
has low similarity (0!),
so possibly fake authors: 0/6



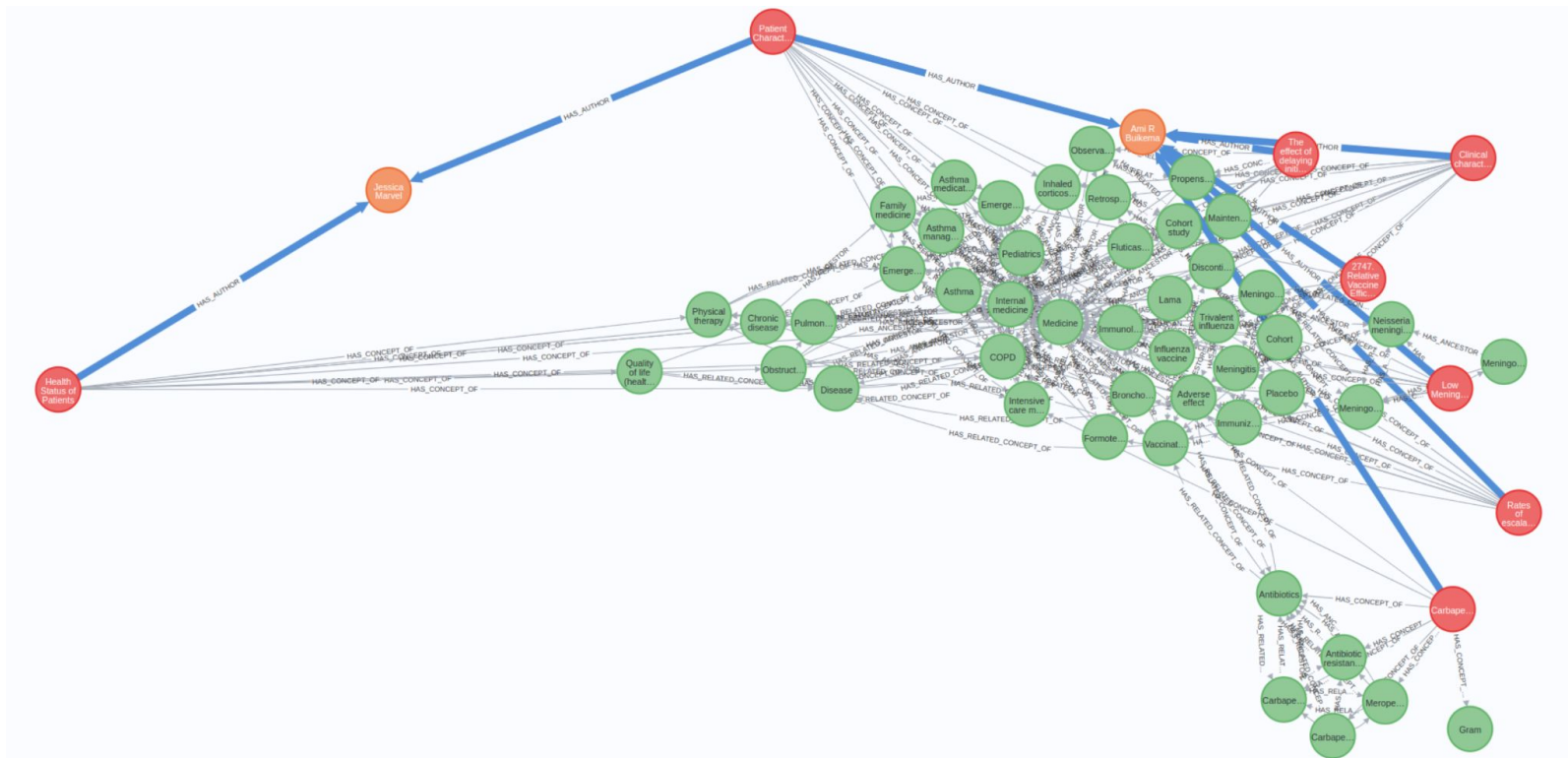
Fake Authors: Igor & Simon



Fake Authors: Ben & Alex



Fake Authors: Charles & Tom

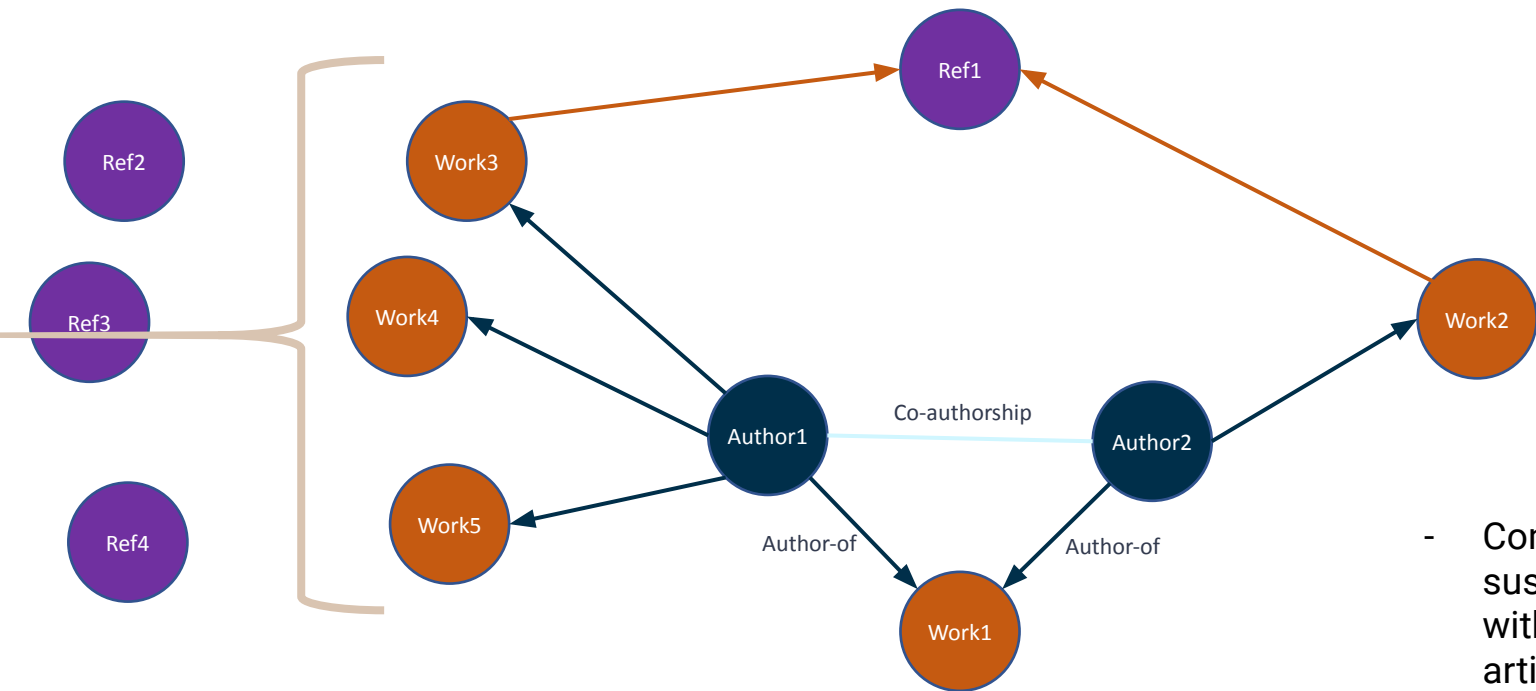


Lessons learned...

- Explore using API first
- Work on small dataset first to prove data flow
- Work as Data product team
- Communicate data constraints as early as possible
- Change model rapidly when there is no hope
- Scale infrastructure when needed is cost-effective

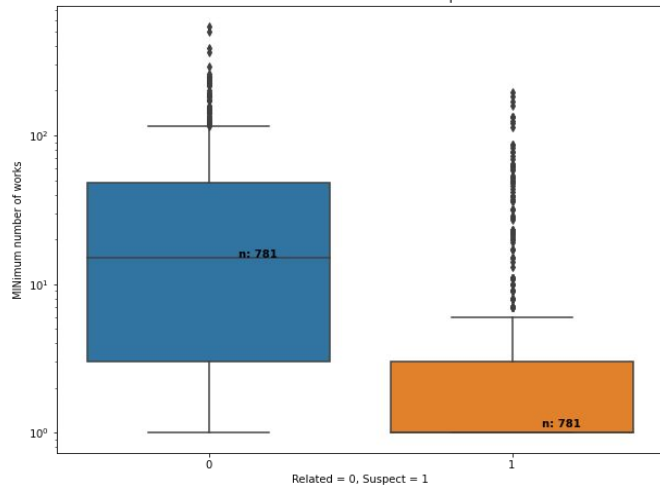
Data Analysis

The background image is a photograph of a modern library interior. It features long, curved white bookshelves that create a sense of depth and perspective. The shelves are filled with books, some of which are standing upright while others are lying flat. The lighting is soft and even, highlighting the clean lines of the shelving system. A dark blue gradient is applied over the entire image, starting from a lighter blue at the top and fading into a darker blue at the bottom. The text 'Data Analysis' is written in a large, white, serif font, positioned in the upper left quadrant of the image.

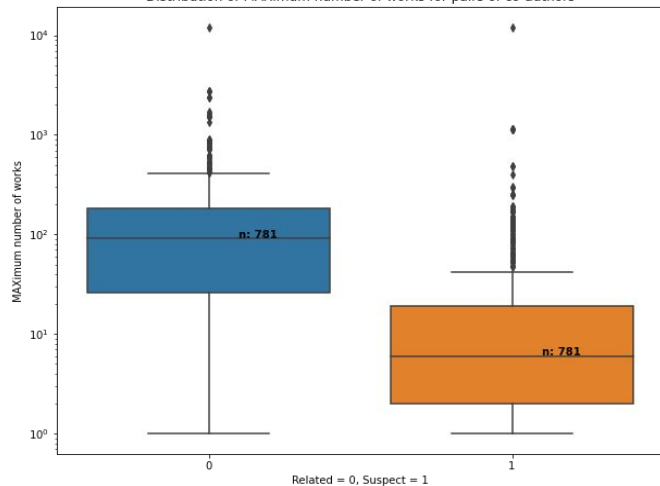


- Compare large set of suspect co-authorships with large set of related articles.
- Limit to coauthors at different institutions

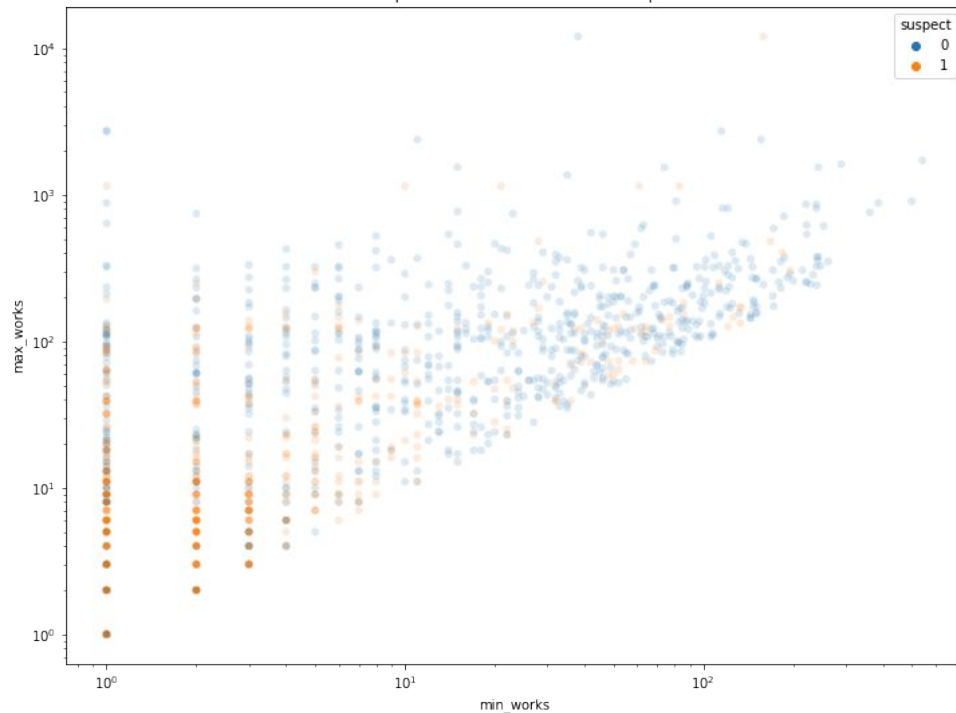
Distribution of MINimum number of works for pairs of co-authors



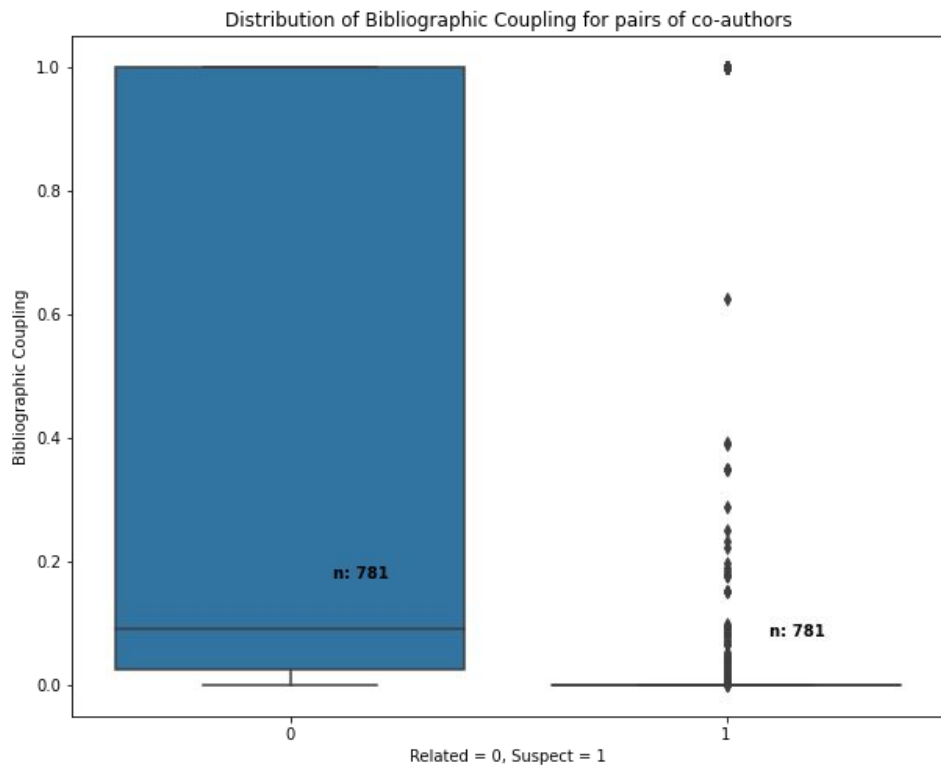
Distribution of MAXimum number of works for pairs of co-authors



Co-authorship minimum and maximum works for
'suspect' co-authorships on works with brokered authorships
compared with control co-authorships



Findings



Machine learning

- Does work well...
- **However,**
 - a simple rule is almost as good
 - more data and testing would be required to make something usable in the real world.

Conclusions

1. Graph analytics are a good approach to identifying organised misconduct
2. To do this efficiently, we need graph data structures, such as a knowledge graph built on well-curated data.
 - OpenAlex is a good data set for this
3. We can detect unusual co-authorships. Either with machine-learning, or a simple rule:
 - It is unusual for an author to co-author with a different institution on their first paper.

References

- Graph Machine Learning ebook by Ebru Cucen
- Publication and collaboration anomalies in academic papers originating from a paper mill: evidence from a Russia-based paper mill Anna Abalkina | @AnnaAbalkina | ArXiv: 2202.03310,
- Exploratory analysis of text duplication in peer-review reveals peer-review fraud and paper mills Adam Day | @AdamSci12 | ArXiv: 2202.03310
- For more on Papermills, see Adam's blog: <https://publisherad.medium.com/>
- Retraction Watch story: <https://retractionwatch.com/2019/07/18/exclusive-russian-site-says-it-has-brokered-authorships-for-more-than-10000-researchers/>
- Candal-Pedreira et al *BMJ* 2022; 379 doi: <https://doi.org/10.1136/bmj-2022-071517> (Published 28 November 2022)