COVID-19 knowledge deconstruction and retrieval: Intelligent bibliometric solutions

Mengjia Wu[1*], Yi Zhang[1], Mark Markley[2], Caitlin Cassidy[2], Nils Newman[2], Alan Porter[2,3]

1 University of Technology Sydney, Australia
2 Search Technology, Inc., United States
3 Georgia Institute of Technology, United States

1. Research motivation

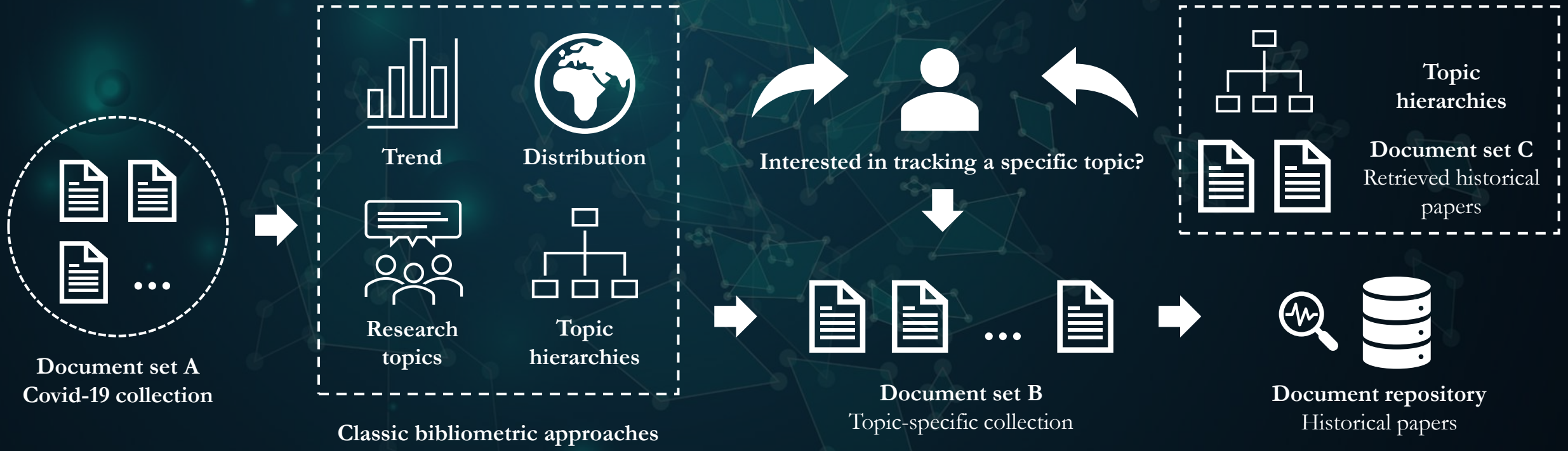
- **COVID-19 knowledge flood brings a severe information crisis.**

CORD-19	NCBI SARS-CoV-2 literature	World Health Organization Collection	Web of Science	Social media
• WHO • Preprints (biorxiv, medrxiv) • PubMed Central	• PubMed • PubMed Central • LitCovid	• Research database • News/updates	• Research articles • Reviews • Conference proceedings • …	• Twitter • Facebook • Weibo • …
1 million +	269k+	614k+	304k+	*

- Facing nearly a million of publications, how can a researcher obtain knowledge that falls in his/her specific interests? (Knowledge mining)

1. Research motivation

- **Our aim: To build a systematic Covid-19 knowledge search and retrieval framework for researchers.**



2. Data and methods

- We decided to use PubMed collection as our data source
  - Peer-reviewed articles
  - More curated metadata (MeSH terms, biomedical entities) to support our project analysis need

Covid-19 collection

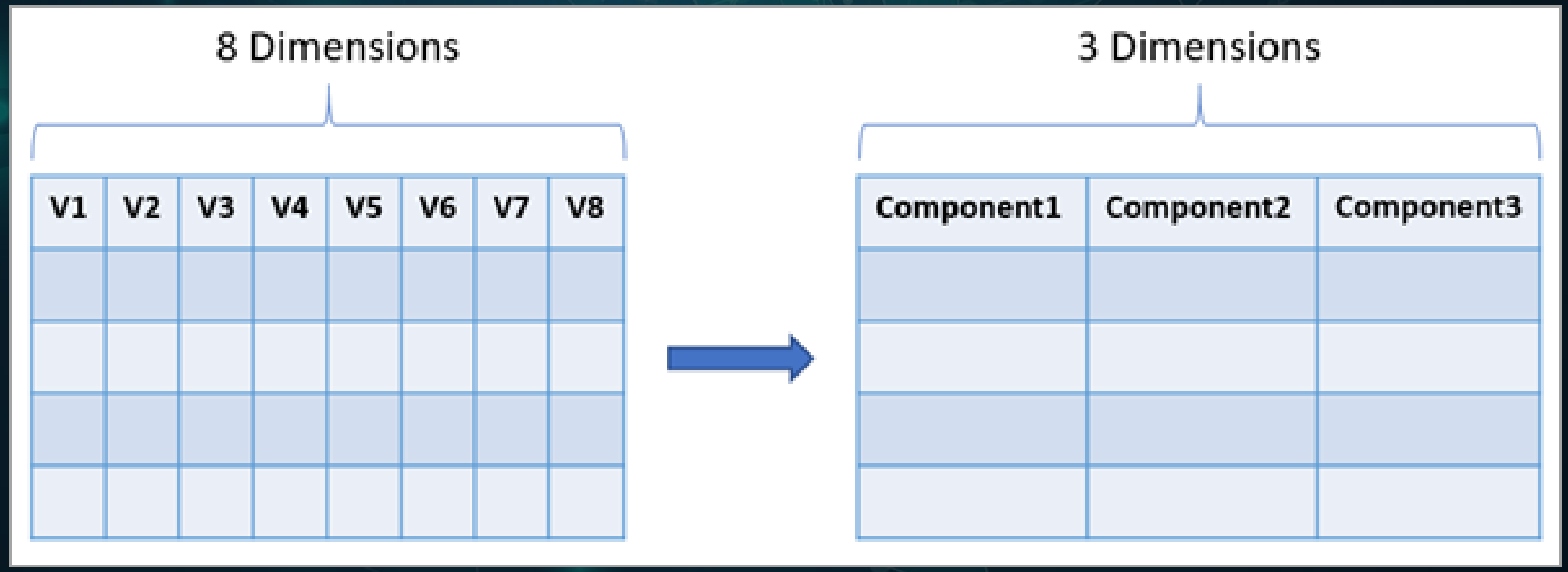
- 127,971 papers (before 2022)
- With titles and abstracts
- Remove editorial, letters

Historical collection (pre-2020)

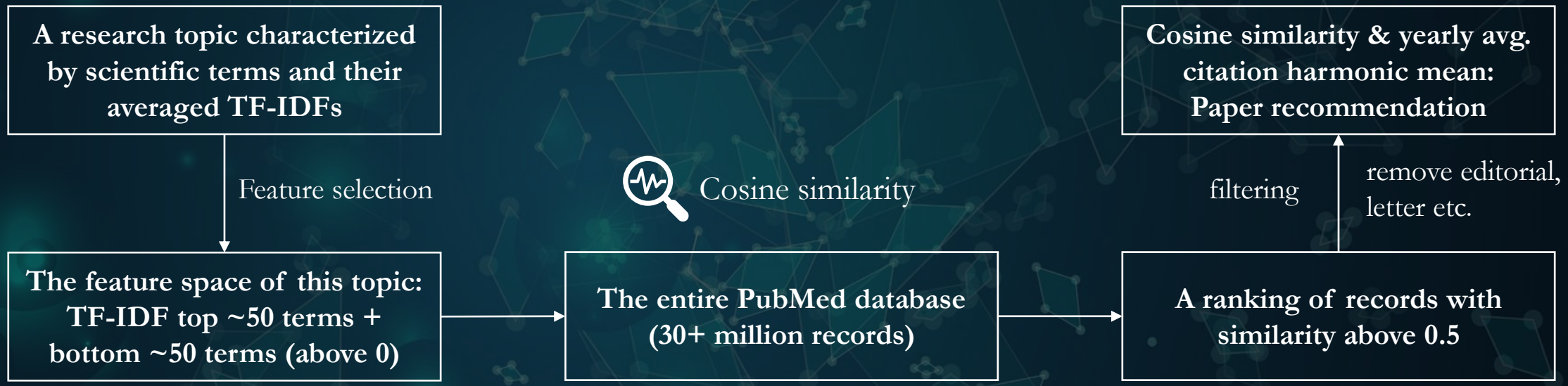
- 30+ million historical papers

2. Research methodology – PCD

- A principal component analysis (PCA) variant
- Works on term-document matrix and reduces term factors
- PCD standardizes the number of factors by minimizing the entropy and maximizing the cohesiveness of the derived factor groups.

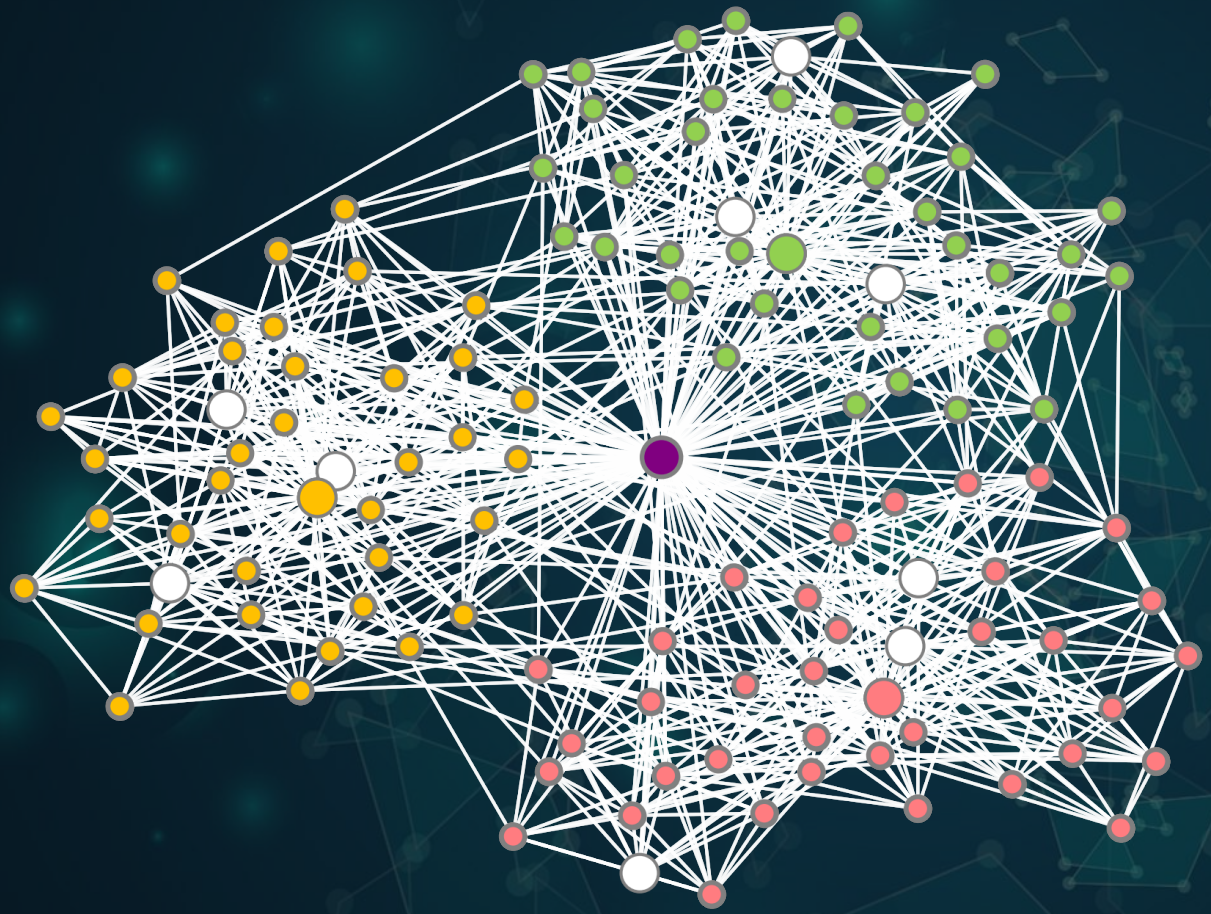


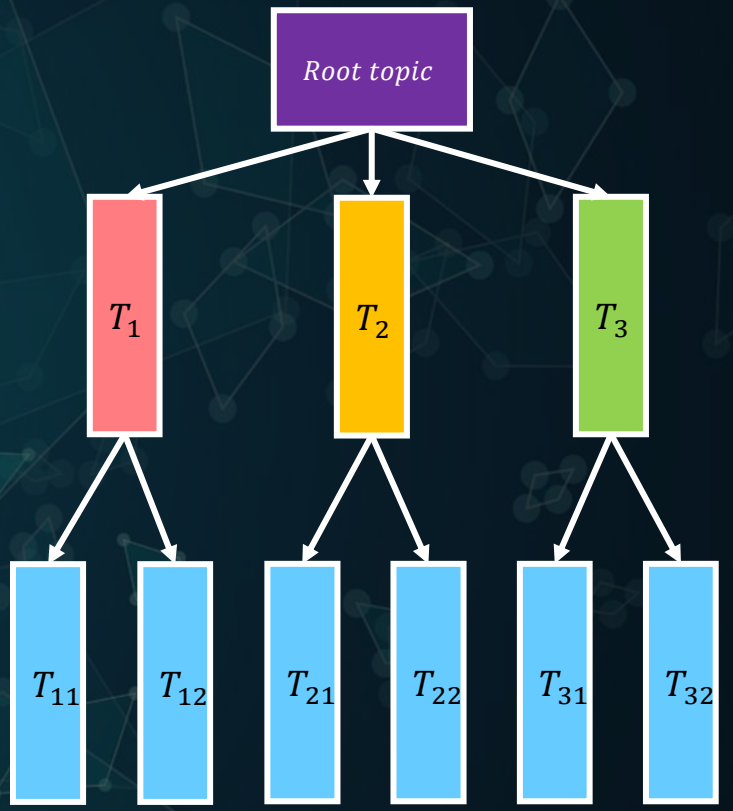
2. Method 2 – Knowledge model search



Why TF-IDF model? - Robustness - Scalability

2. Method 3 – Hierarchical topic tree (HTT)





Co-term network → Density peak search → Node allocation → The ith layer of topics → HTT result

New density peak node in the ith layer found

No density peak found

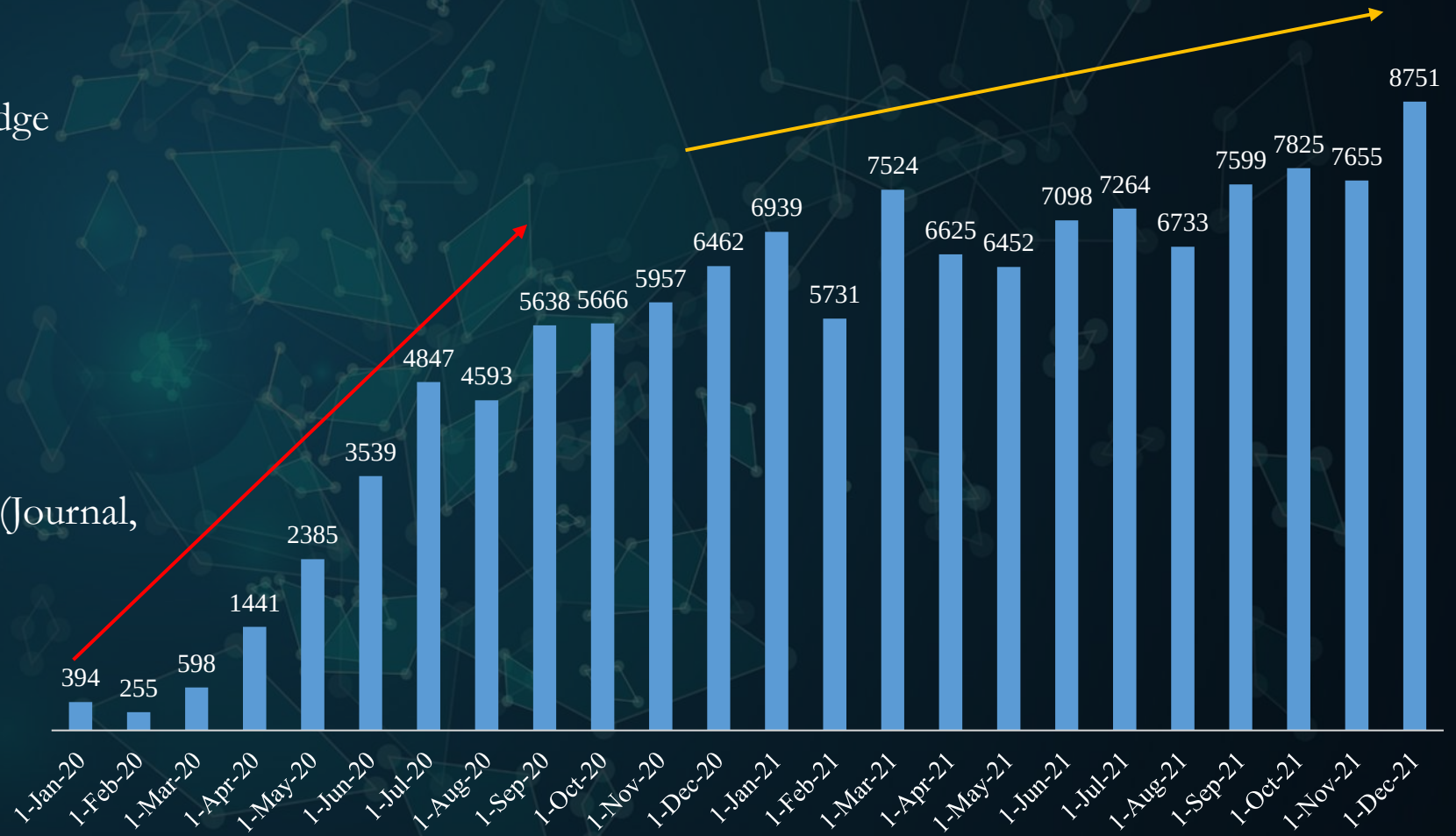
4. Results - Trend & Distribution

Knowledge burst

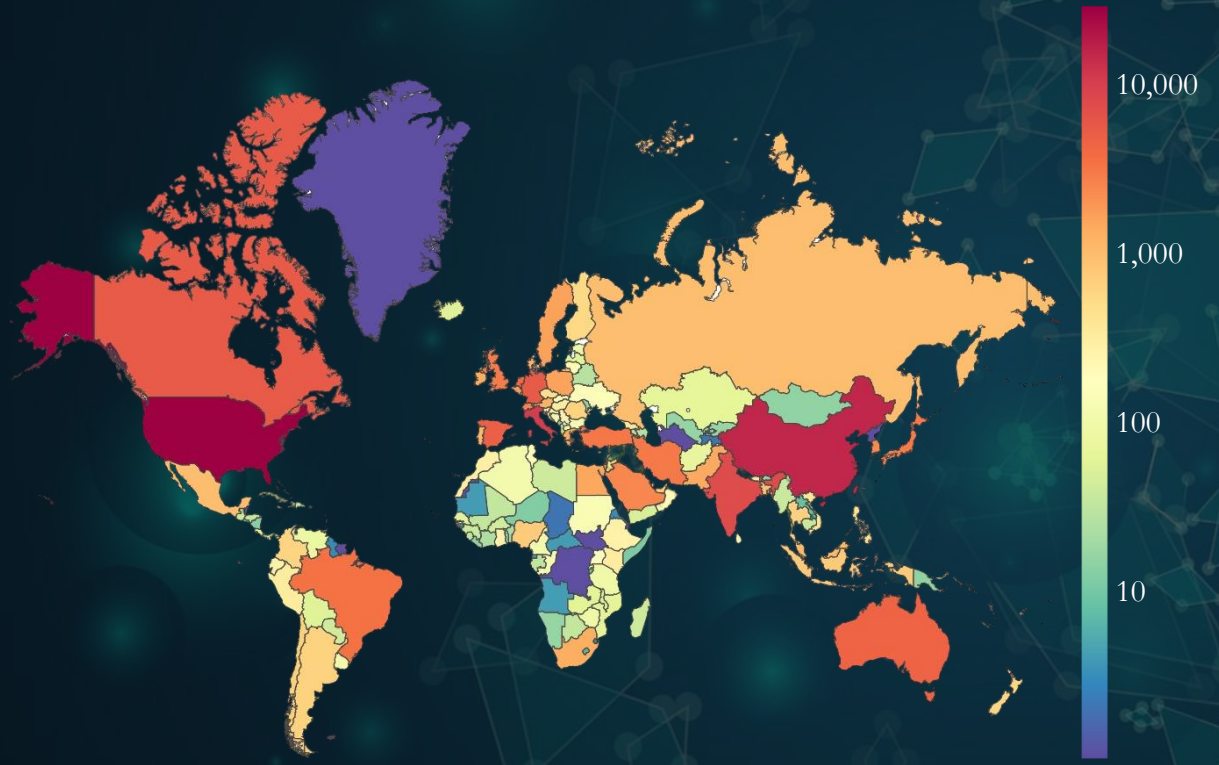
- Massive novel/disruptive knowledge

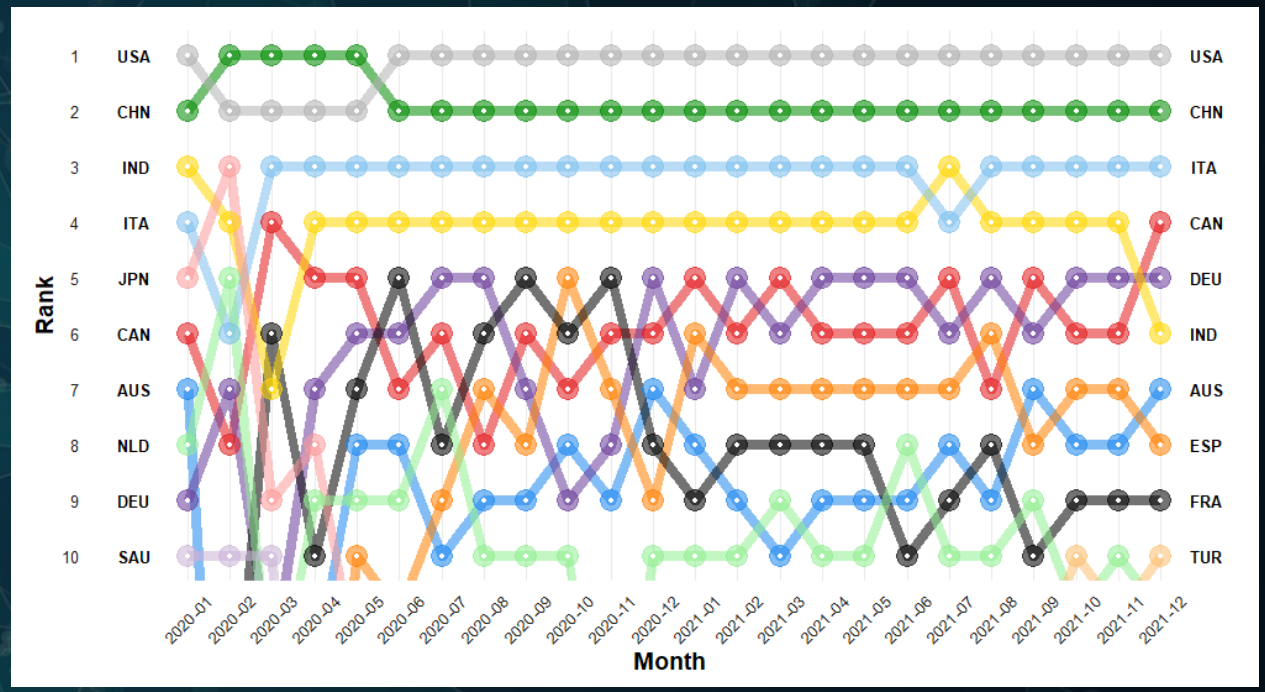
Slowing down rate

- Knowledge convergence?
- Research resource capacity limit? (Journal, review resources, funding limit)



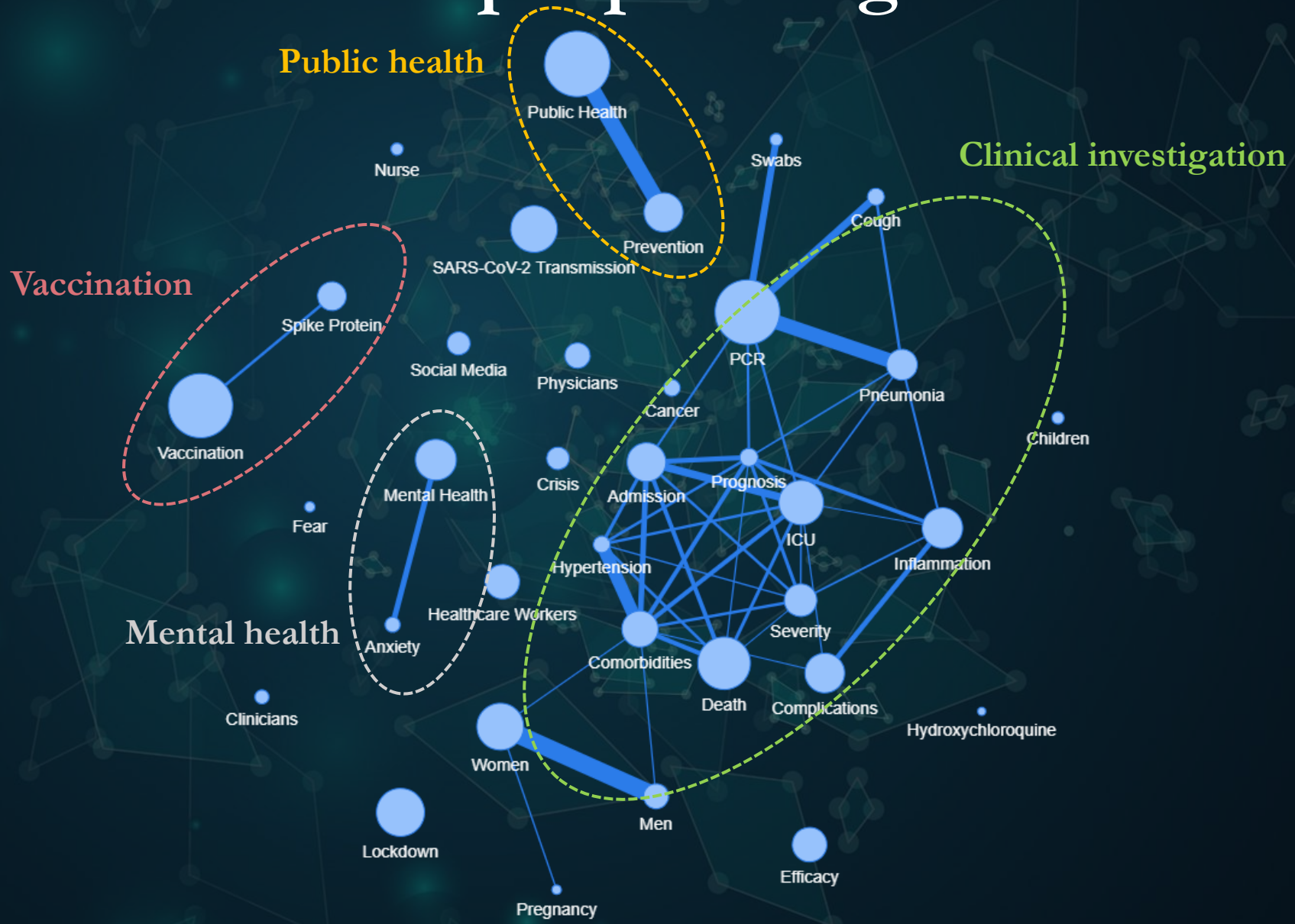
4. Results - Trend & Distribution



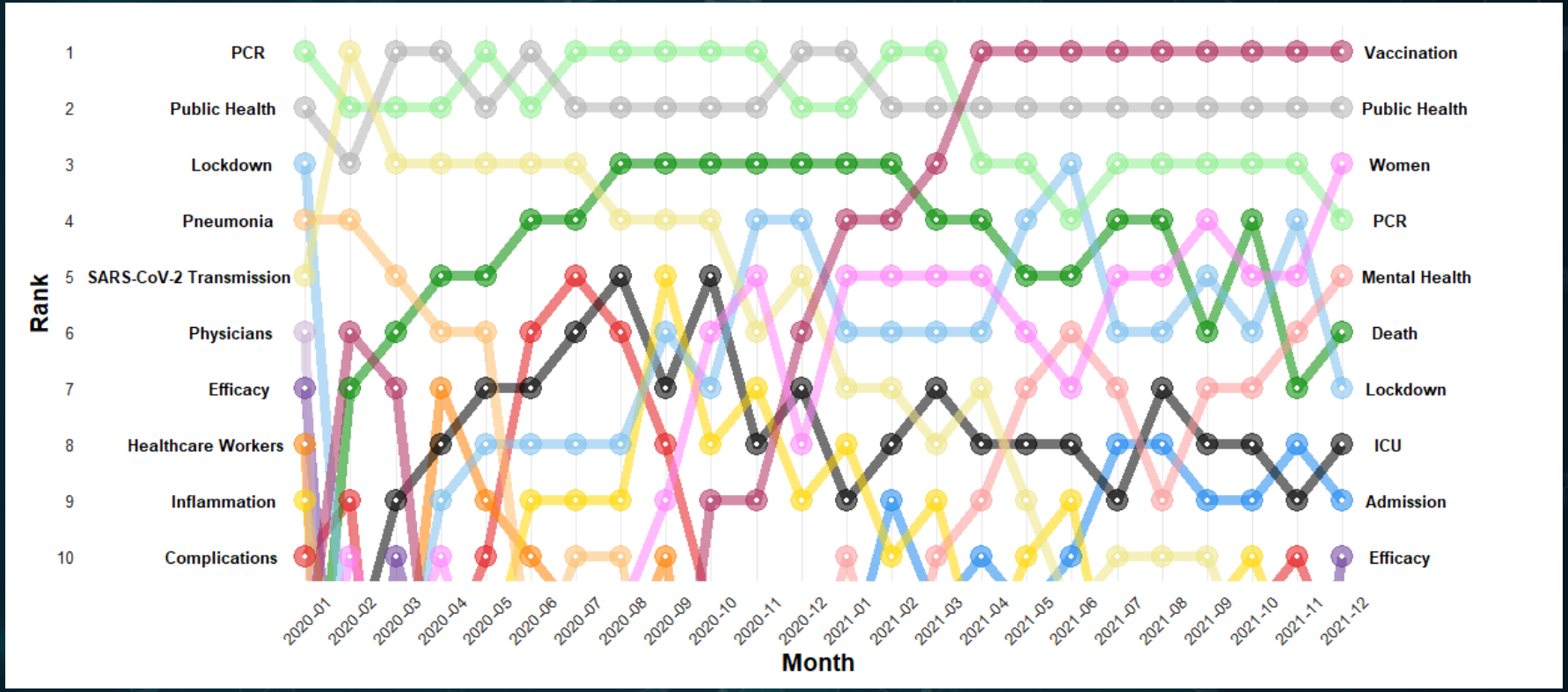


- China-led vs. US lead
- A possible positive correlation between article productivity and Covid-19 severity

4. Results - PCD topic profiling

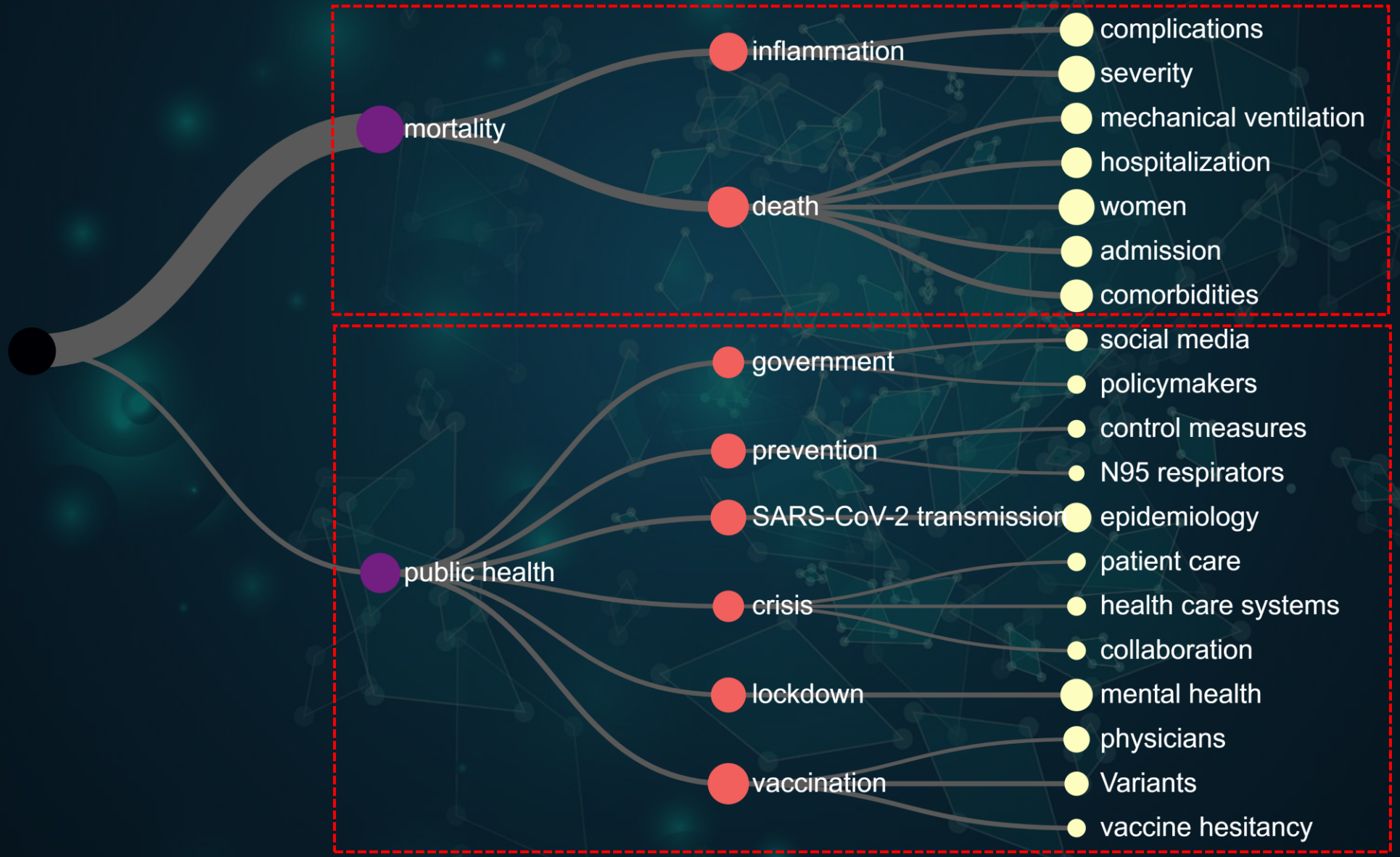


4. Results – PCD topics change



- Public health
- Vaccination
- Women
- SARS-CoV2-transmission
- ICU, PCR, death
- Lockdown
- Mental health

4. Results – Overall HTT



Clinical branch

Public health branch

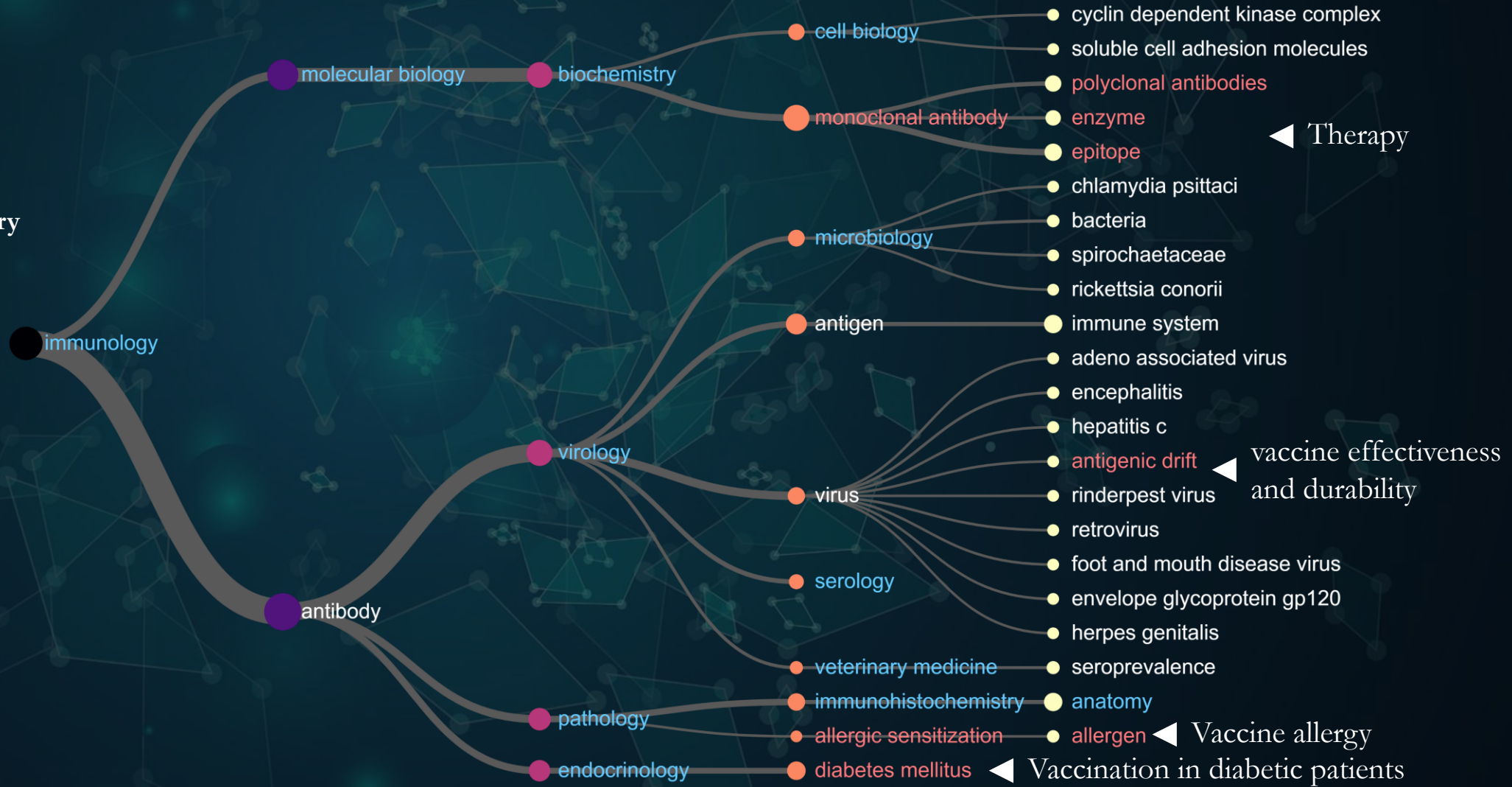
4. Results – Knowledge retrieval (Vaccination)

Top				Bottom			
Stem	TF-IDF (avg.)	Stem	TF-IDF (avg.)	Stem	TF-IDF (avg.)	Stem	TF-IDF (avg.)
vaccin	0.0711	bnt162b2	0.0054	synchronis	4.13E-07	e31del	4.82E-07
antibodi	0.0289	case	0.0054	africain	4.26E-07	f888l	4.82E-07
immun	0.0138	protect	0.0053	d253g	4.38E-07	formerlygr	4.82E-07
neutral	0.0121	individu	0.0053	q954h	4.38E-07	h69del	4.82E-07
dose	0.0101	diseas	0.0052	s373p	4.38E-07	havevaccin	4.82E-07
variant	0.0097	popul	0.0051	s375f	4.38E-07	k848	4.82E-07
infect	0.0093	coronaviru	0.005	y505h	4.38E-07	l212i	4.82E-07
respons	0.0091	antigen	0.0049	voor	4.44E-07	n211del	4.82E-07
cell	0.009	efficaci	0.0049	andb	4.54E-07	n417	4.82E-07
mrna	0.0089	titer	0.0048	d796y	4.54E-07	n969k	4.82E-07
spike	0.0087	receiv	0.0047	f157l	4.54E-07	namelyepsilon	4.82E-07
protein	0.0082	report	0.0047	l981f	4.54E-07	q1071h	4.82E-07
test	0.0076	posit	0.0047	r190	4.54E-07	q19e	4.82E-07
patient	0.0075	serolog	0.0047	severityof	4.54E-07	r32del	4.82E-07
anti	0.0074	epitop	0.0047	t1027i	4.54E-07	s33del	4.82E-07
develop	0.0069	human	0.0046	thebeta	4.54E-07	s929i	4.82E-07
hesit	0.0065	syndrom	0.0046	v70del	4.54E-07	spathogen	4.82E-07
assai	0.0064	clinic	0.0045	variantwa	4.54E-07	tegallyet	4.82E-07
viru	0.0062	respiratori	0.0045	1092k	4.82E-07	theb	4.82E-07
bind	0.0059	trial	0.0044	156del	4.82E-07	thebind	4.82E-07
effect	0.0057	influenza	0.0044	157del	4.82E-07	thedelta	4.82E-07
viral	0.0057	mutat	0.0044	2020in	4.82E-07	theepsilon	4.82E-07
sever	0.0056	receptor	0.0043	351also	4.82E-07	thefus	4.82E-07
detect	0.0055	base	0.0043	7variantwa	4.82E-07	theheptad	4.82E-07
specif	0.0055	group	0.0043	a63t	4.82E-07	thep	4.82E-07

4. Results – Visualizing retrieved knowledge

Document repository
Historical papers

89,951 records



5. Conclusions

1. A methodology framework for Covid-19 knowledge profiling & retrieval
2. Two research trajectories: Clinical research and public health
3. The social impacts of Covid-19 are attracting more attention than the disease and virus themselves
4. Highlight four specific knowledge foundations from historical articles for future vaccination research reference

5. Future directions

1. Quantitative validation for knowledge model search & HTT (Being improved in our methodology paper)
2. How can we measure the quality of Covid-19 publications? (next goal in our project)
3. Quantitative indicators to measure knowledge disruption caused by Covid-19.
4. How do we exploit the results to facilitate novel knowledge discovery?

Thank you

Q&A

Email address: Mengjia.Wu@uts.edu.au