Natalia Manola

University of Athens, Department of Informatics

Athena Research & Innovation Center

Explore, model, analyze and visualize systematic research in OpenAIRE

… via text and data mining (topic modeling)

A bird's eye view

Big volumes of data

Publications ARE data in TDM.

Should abide to the FAIR principles.

Meta research: Research analytics

OpenAIRE

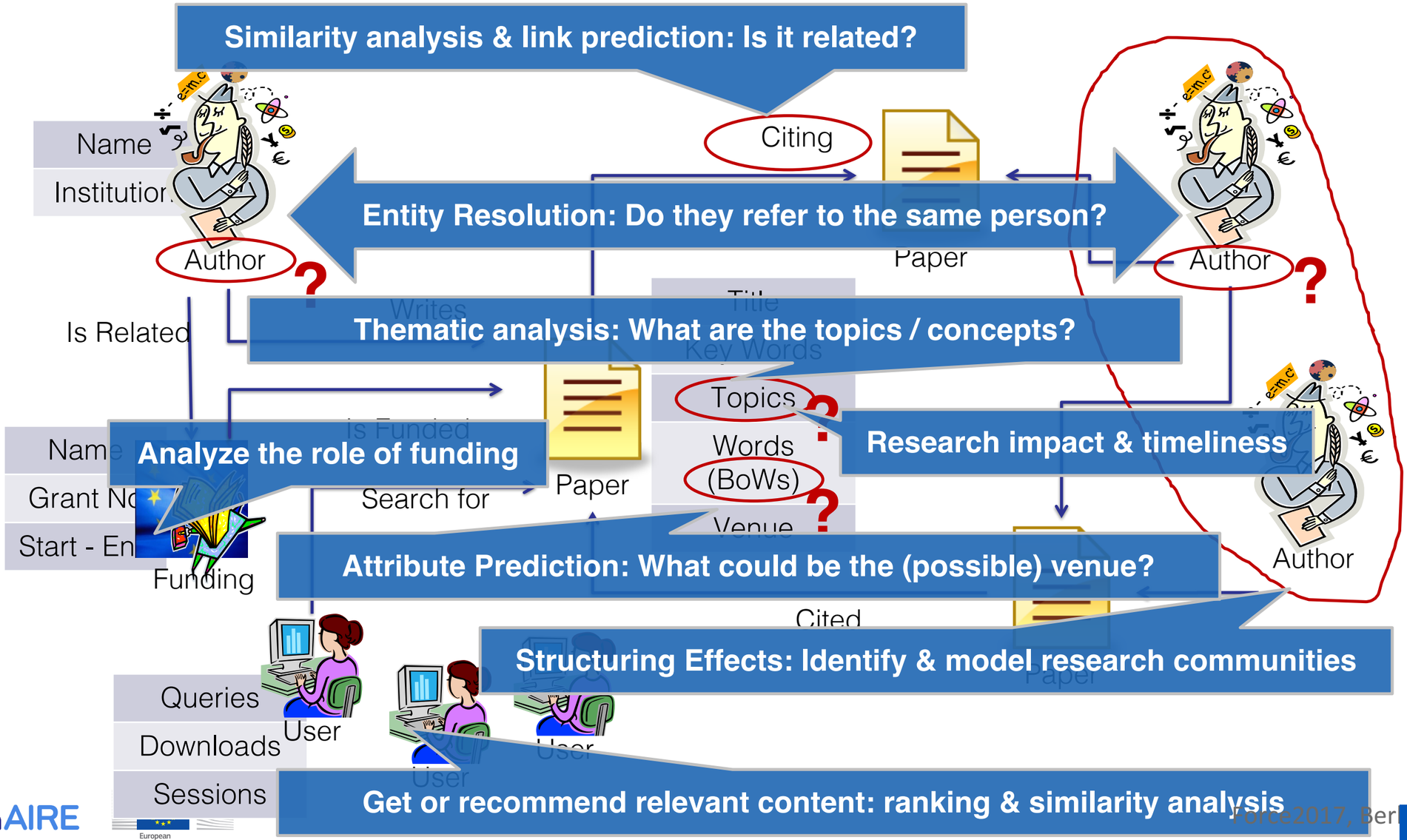
European Commission

openMINTED
Open Mining Infrastructure for Text & Data

Force2017, Berlin, 27 Oct, 2017

Mining scientific/scholarly literature

WHY



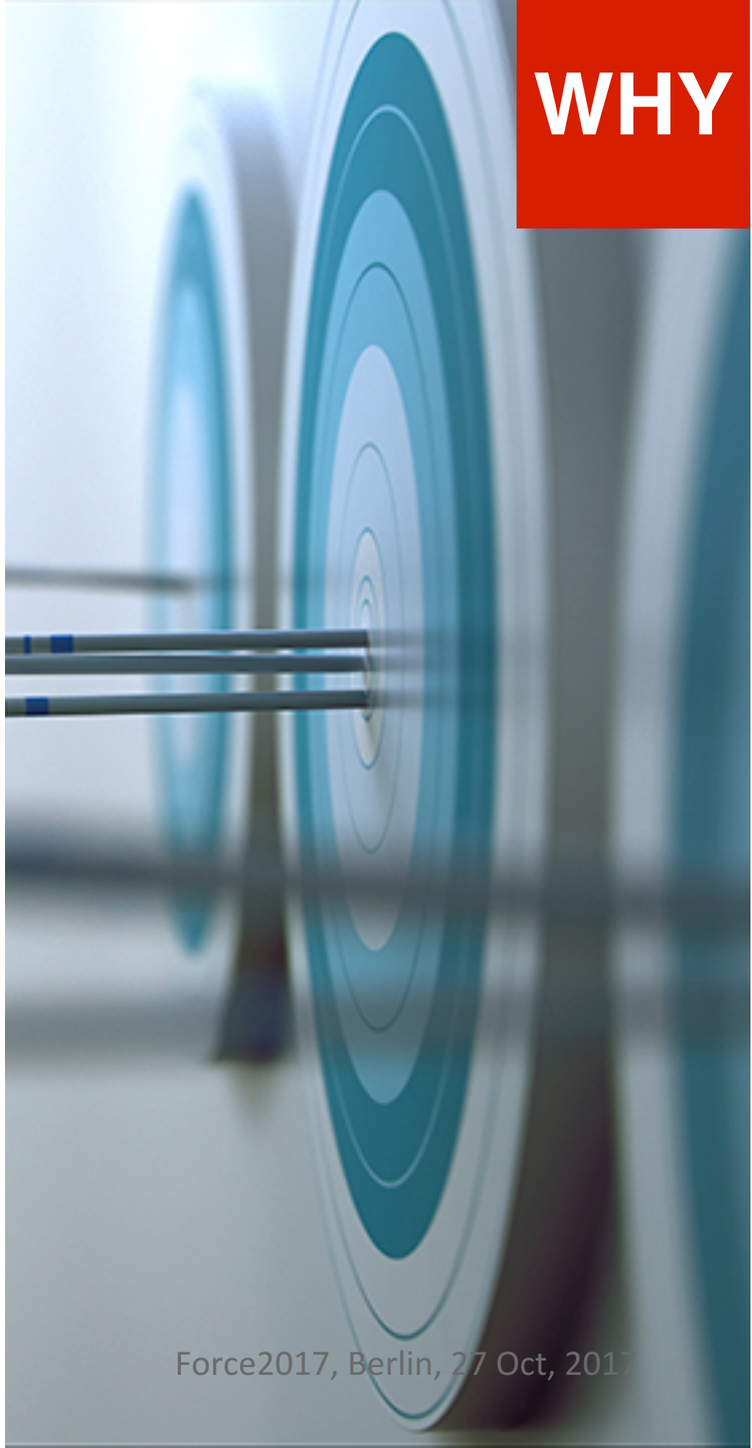
Mining scientific/scholarly literature

New models → new insights → better decisions

Real Output vs. project & call descriptions

Analyze large collections of documents, and meta-data to:

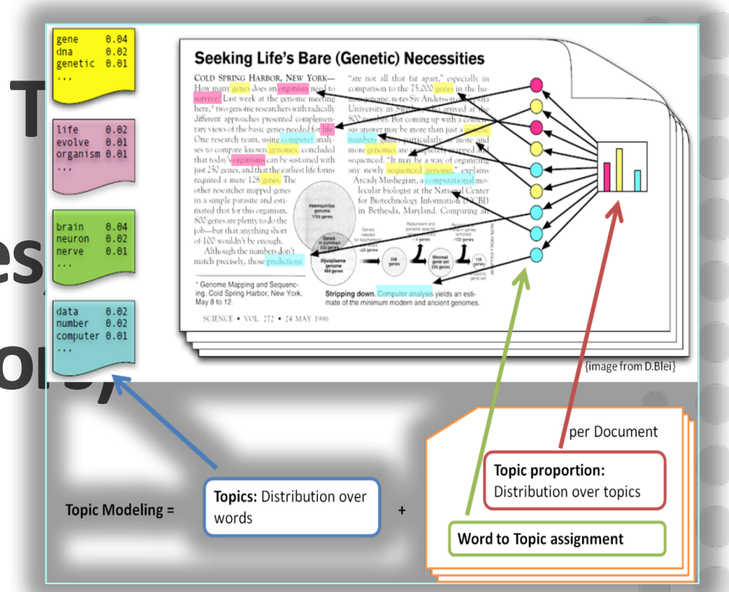
- Assess research collaboration: **authorship network analysis**
- Identify active areas of research: **discover hidden themes** (topics)
- Understand what is actually produced
- Discover clusters and communities
- Identify emerging research areas
- Assess coverage, identify gaps or new challenges

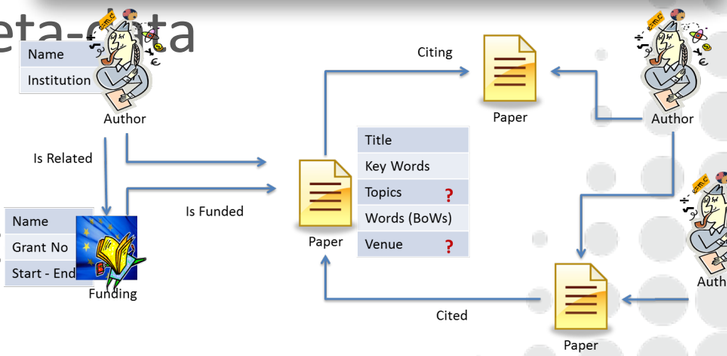


Probabilistic Multi-View Topic Modeling of Text-Augmented Heterogeneous Information Networks

HOW

- **Interconnected** (linked) entities characterized by T
- **Related side** information & links (e.g., taxonomies, venues, projects / research areas, citations, author
- **Side-information**
  - structured or unstructured attributes, links / relations and meta-data
  - form networks: e.g., authorship network, citation network, …
  - incomplete or missing, noisy or not related to textual attribut





OpenAIRE

European Commission

openMINTED Open Mining Infrastructure for Text & Data

Multi-View vs Text only: interoperability and coverage

MV_HDP	topic: "Topic Modeling"	"Cloud/Distributed computing & Big Data Analytics"
Text	topic, latent, lda, document, dirichlet, probabilistic, mining, semantic, allocation, generative, word, mixture, topical, corpus, plsa, bayesian, unsupervised,..	mapreduce, big, hadoop, analytics, cluster, map, scalable, datasets, queries, cloud, intensive, jobs, databases, massive, google, job, scalability, node, computations, mining, hdfs, hive, machine, workloads, volume,…
Citations (ranked list of citation net nodes)	"Dynamic topic models", "Topics over time" "Joint latent topic models for text and citations" "Topic modeling", "Probabilistic topic models" "Probabilistic latent semantic indexing",…	" A comparison of approaches to large-scale data analysis", "Pig latin", "Mesos", "DryadLINQ", "PREGEL", "CIEL", "Improving MapReduce performance in heterogeneous environments", "MapReduce Online", "MapReduce Merge" ,..
Taxonomy	H.3.3 IR: Information Search and Retrieval, H.3.1 IR: Content Analysis and Indexing, H.2.8 DB MNGMT: Database Applications, I.2.6 AI: Learning, I.2.7 AI: Natural Language Processing, I.5.1 PAT.REC.: Models	H.2.4 DB MNGMT: Systems D.1.3 PROGR TECHNIQUES: Concurrent Programmin Distributed Systems, Applications, H.3.4 INF and Software
Keywords	topic modeling, latent dirichlet allocation, latent semantic analysis, generative model, text mining	big data, Map-Redu distributed comput learning, parallel pr
Venues	SIGKDD, WSDM, CIKM	SIGMOD, BigSystem, CloudCP, EUROSEC, EUROSYS,..

Good metadata is important

What is involved?

1 ENRICH & PRE-PROCESS	Extract **features** and **annotate (enrich)** content using NLP, Named Entity Recognition & Semantic Annotation Tokenize, remove stop words	Refine stop words for specific domain
2 FIND TOPICS	Identify topics: distribution over words & “side” information Automatic topic curation & entitling Assign topics to publications	Evaluate & categorize topics Assess topic labels
3 CALCULATE TRENDS & SIMILARITIES	Calculate **topic proportions & trends** of objects based on their publications Calculate **similarity** among different entities based on various metrics	Analyze & Validate the results
4 VISUALIZE	Create **WEB interactive** visualization with **data driven graphs**, charts and layouts	Design optimal views Validate modeling results

What is the result?

1. Linked information

MUSCLE W49 : A Multi-Scale Continuum and Line Exploration of the Most Luminous Star Formation Region in the Milky Way. I. Data and The Mass Structure of the Giant Molecular Cloud

UNKNOWN, ARTICLE, PREPRINT | ENGLISH | OPEN

R. Galván-Madrid ; Liu, H. B. ; Zhang, Z-Y ; Pineda, J. E. ; Peng, T-C ; Q. Zhang ; Keto, E. R. ; Ho, P. T. P. ; Rodríguez, L F ; Zá
De Pree, C. G. (2013)

Publisher: IOP Publishing

doi: 10.5167/uzh-90740, doi: 10.1088/0004-637X/779/2/121

Subject: Astrophysics - Astrophysics of Galaxies | Astrophysics - High Energy Astrophysical Phenomena | 530 Physics
Computational

Classified b
arxiv: Astrop

REFERENCES 105 METRICS

105 REFERENCES, PAGE 1 OF 11

The Multi
(GMC) of
different s

Aguirre, J. E., Ginsburg, A. G., Dunham, M. K., et al. 2011, ApJS, 192, 4

Alves, J., & Homeier, N. 2003, ApJ, 589, L45

Baobab Liu, H., Ho, P. T. P., Zhang, Q., et al. 2010, ApJ, 722, 262

Bastian, N., & Goodwin, S. P. 2006, MNRAS, 369, L9

Belloche, A., Mu&#x7f;ller, H. S. P., Menten, K. M., Schilke, P., & Comito, C. 2013, ArXiv e-prints

Download from
Zurich Open Repository and Archive
org e-Print Archive
ublishing

ed in
án-Madrid, R; Liu, H B; Zhang, Z-
da, J E; Peng, T-C; Zhang, Q;
íguez, L F;
ree, C G
multi-scale
oration of the
ation region
and the mass
olecular
rnal,

STAR FORMATION FROM THE MILKY WAY TO THE UNIVERSE: OBSERVATIONS AND SIMULATIONS IN THE ERA OF INFRARED AND MILLIMETER ASTRONOMY

Project Code: CRSII2_141880

PROGRAMS ON CRITICAL PROBLEMS IN PHYSICS, ASTROPHYSICS AND BIOPHYSICS AT THE ASPEN CENTER FOR PHYSICS

Project Code: 1066293
Funder: National Science Foundation (NSF)
Funding: Directorate for Mathematical & Physical Sciences | Division of Physics

Inferred by Algorithm

om the Milky Way
ations and
infrared and

al Problems in Physics, Astrophysics and Biophysics at the Aspen Center for Physics

How often is "Topic Modeling" encountered?

Association of Computing Machinery Corpus

Rank	TopicId	Title	Weight
230	18	Data management & file systems	0.0028
231	132	Image processing: Face & emotion recognition, facial animation	0.0027
232	373	Project management & software development	0.0027
233	[illegible]	Self-adaptive systems & autonomic computing	0.0027
[illegible]	[illegible]	[illegible]ance	0.0026
[illegible]	[illegible]	[illegible]	0.0026
236	271	Haptic technology, feedback & multimodal user interaction	0.0025
237	322	Information extraction, Named entity recognition, disambiguation, cleaning	0.0025
238	348	cognitive psychology, cognitive and mental models	0.0025
240	74	HCI: Touch screen interaction & interactive surfaces	0.0025
241	**382**	**Topic Modelling**	**0.0025**
242	230	Trust & reputation analysis and management (IOT, Web, recom. systems)	0.0025
243	2	Wikipedia & collaborative editing	0.0025
245	15	Crowdsourcing & human computation	0.0025
246	273	Automatic programming, refactoring & transformations	0.0024
248	323	Reliability, fault tolerance and recovery	0.0024
249	113	Online / computational advertising	0.0024

Out of 382

Is it trendy?

TopicId	Title	Weight	Trend	Journal	Confer
15	Crowdsourcing & human computation	0.003	**27.89**	0.068	0.035
194	Cloud Computing, Storage & Virtualization	0.004	**23.56**	0.077	0.011
201	Social network analysis: influence, info diffusion, communities	0.004	**10.82**	0.119	0.066
350	Distributed (Big) Data analytics (cloud, MapReduce)	0.006	**10.54**	0.057	0.022
41	Mobile applications	0.005	**9.86**	0.135	0.019
68	Social media analysis (twitter, blogs, news feed)	0.004	**9.72**	0.078	0.049
366	Persuasive technologies, gamification, user engagement	0.003	**8.65**	0.126	0.070
61	Wearable computing, technology & activity recognition	0.003	**8.24**	0.135	0.044
40	ICT in developing countries (India)	0.002	**7.72**	0.096	0.100
341	GPU computing	0.004	**6.78**	0.120	0.029
133	Recommendation, personalization and collaborative filtering	0.006	**6.27**	0.096	0.085
134	Flash memory structures, storage & systems	0.002	**6.2**	0.144	0.077
22	HCI: Organic & Flexible user interfaces	0.001	**6.04**	0.123	0.101
74	HCI: Touch screen interaction & interactive surfaces	0.003	**5.87**	0.205	0.118
2	Wikipedia & collaborative editing	0.003	**5.33**	0.079	0.083
52	HCI design & user experience	0.013	**5.15**	0.156	0.082
266	Sentiment analysis & opinion mining	0.002	**4.95**	0.057	0.047
10	Image retrieval & object recognition	0.006	**4.91**	0.082	0.048
382	**Topic Modelling**	**0.003**	**4.57**	**0.111**	**0.069**
228	Software product line engineering	0.003	**3.92**	0.128	0.094
100	Social tagging, annotation & tag recommendation	0.005	**3.88**	0.115	0.037
294	Robotics, human-robot interaction, anthropomorphism	0.005	**3.34**	0.066	0.170

Top 20

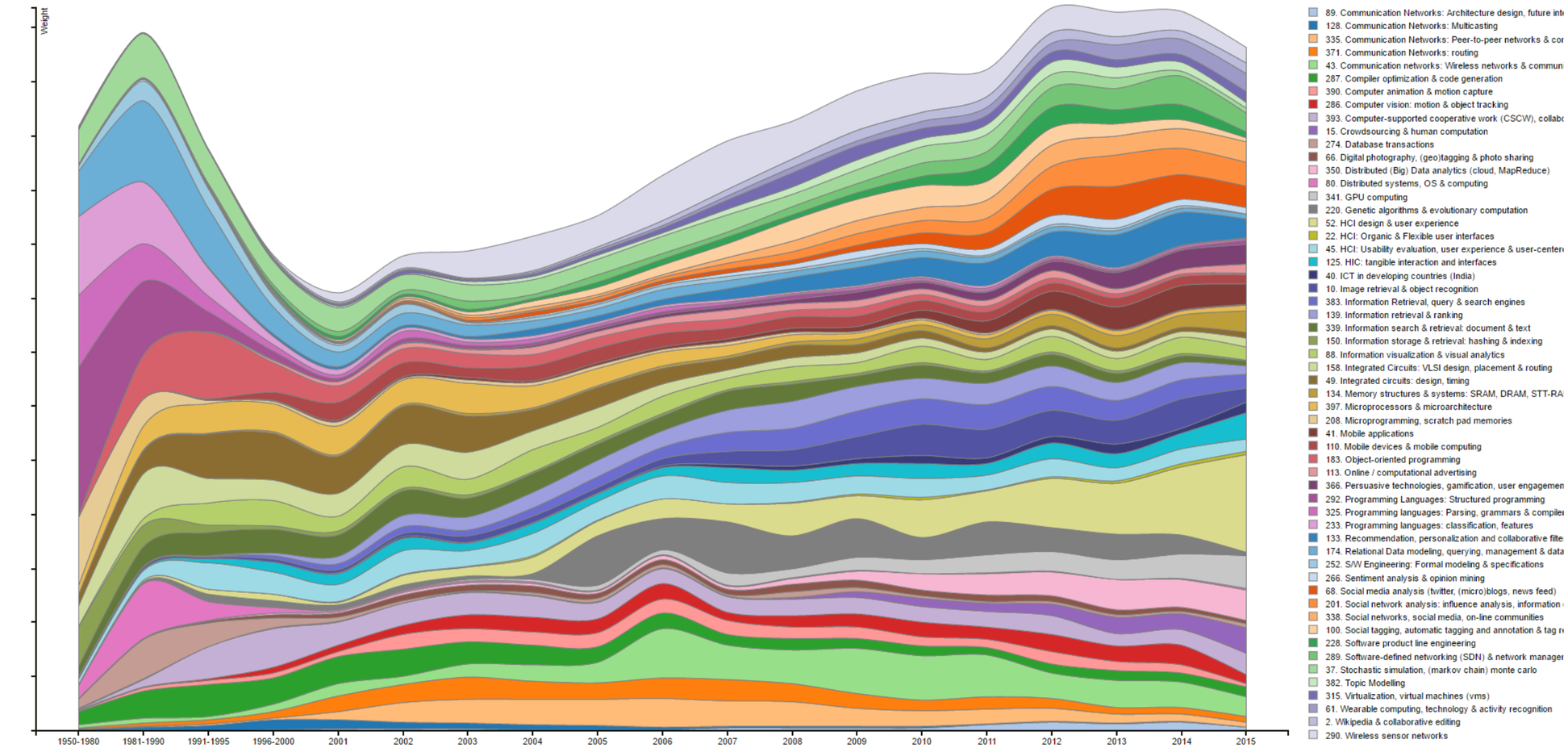
Concept driven search

View top 23 most related publications to “Topic Modeling”

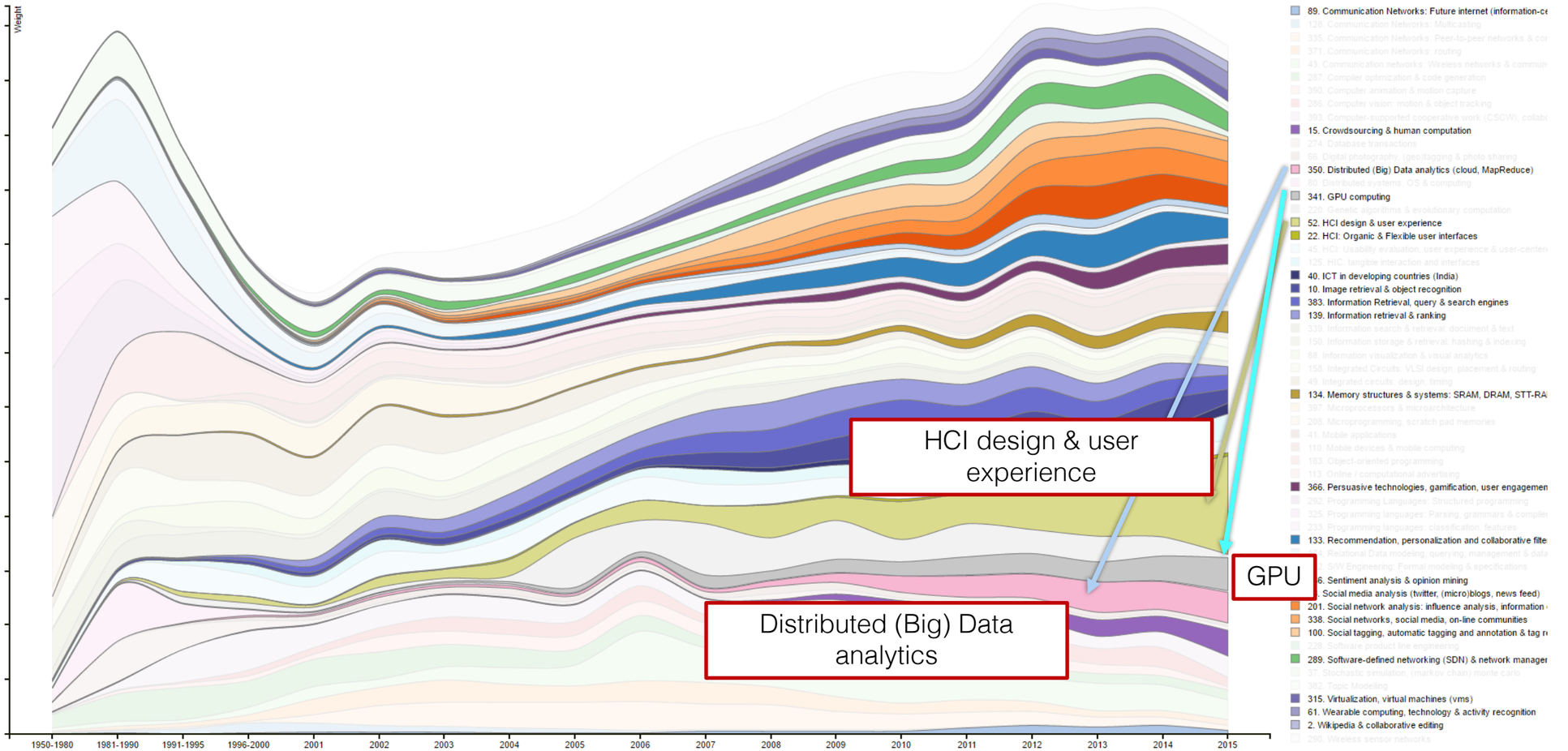
PubId	Weight	Title
1646242	0.72	Dynamic hyperparameter optimization for bayesian topical trend analysis
		Latent interest-topic model
		ing
1458337	0.63	Combining concept hierarchies and statistical topic models
2348335	0.63	Group matrix factorization for scalable topic modeling
2009977	0.63	Mining topics on participations for community discovery
1835890	0.62	Topic models with power-law using Pitman-Yor process
2398483	0.61	Hierarchical topic integration through semi-supervised hierarchical topic modeling
1150482	0.60	A mixture model for contextual text mining
1963244	0.60	Investigating topic models for social media user recommendation
1281249	0.60	Multiscale topic tomography
2086739	0.59	Sequential Modeling of Topic Dynamics with Multiple Timescales
1572095	0.59	A latent topic model for linked documents
2188143	0.59	Latent contextual indexing of annotated documents
1859210	0.58	Topic models vs. unstructured data
1487045	0.58	Linked Topic and Interest Model for Web Forums
2609471	0.58	Probabilistic text modeling with orthogonalized topics
2396861	0.57	Modeling topic hierarchies with the recursive chinese restaurant process
2433438	0.57	Group sparse topical coding
1935880	0.57	Trend analysis model
1390546	0.56	Improving text classification accuracy using topic modeling over an additional corpus
1553410	0.55	Accounting for burstiness in topic models

Visualization

Trendy, old-fashion, common topics

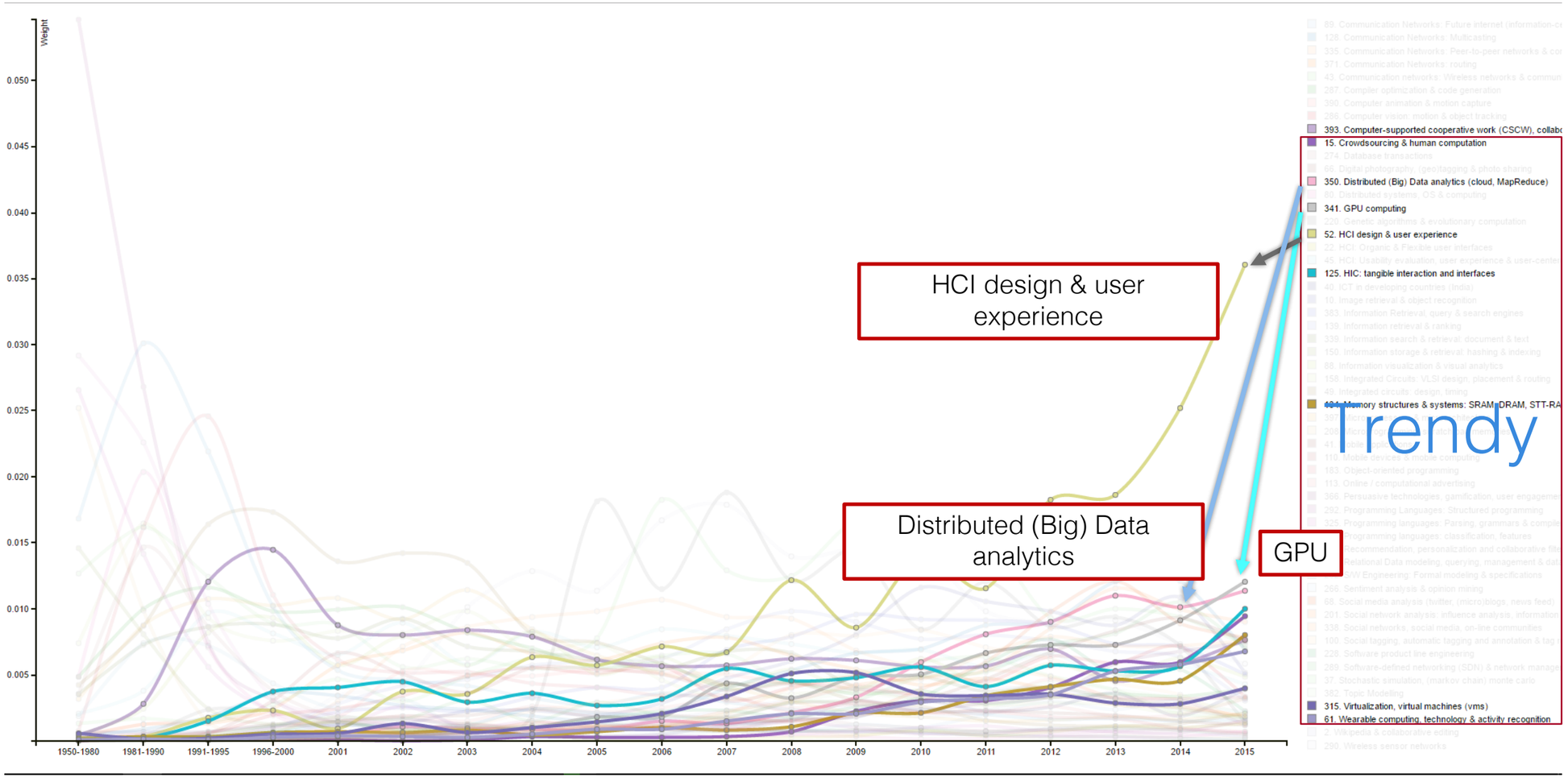


Trendy topics

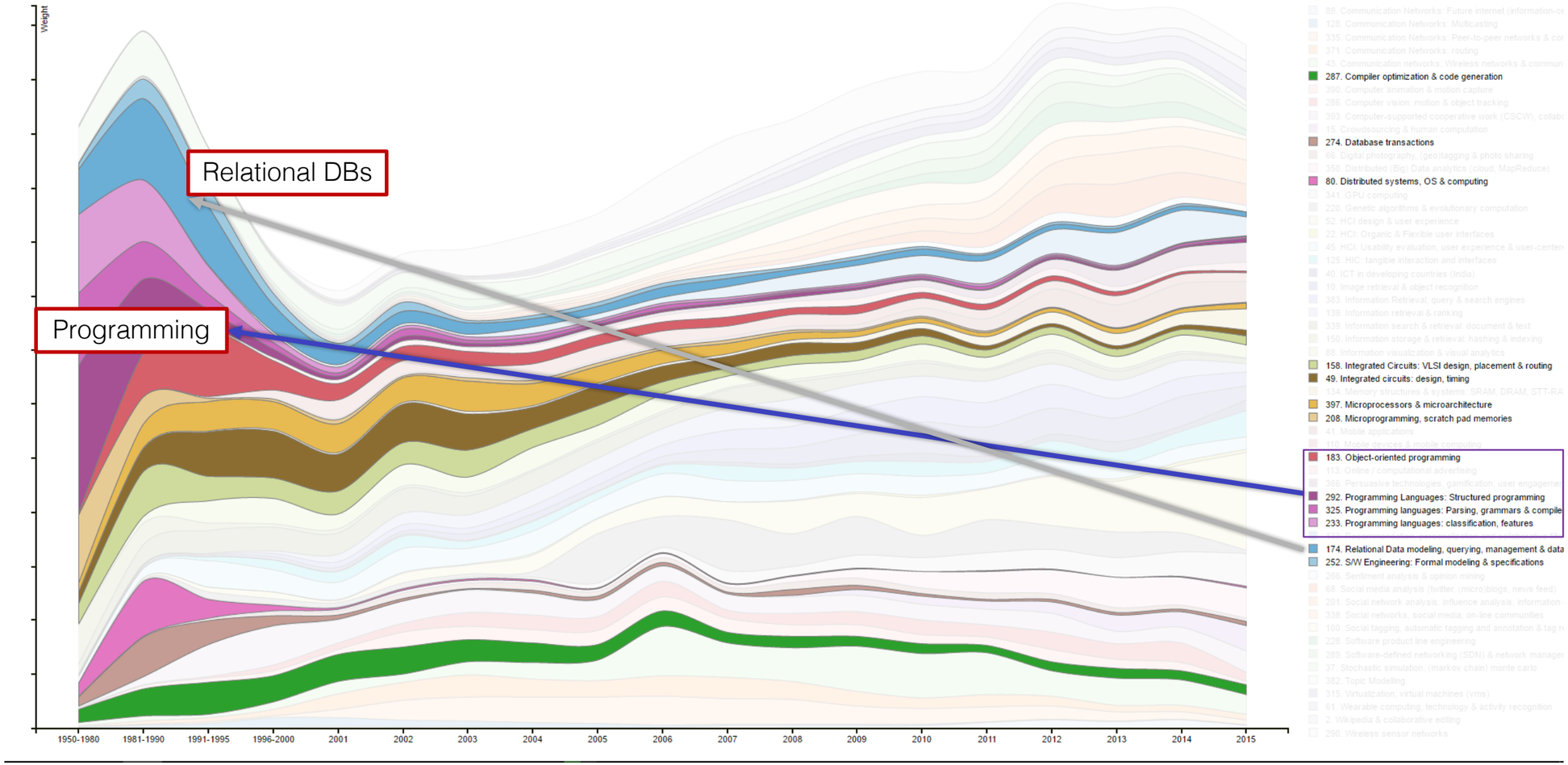


Trendy topics

Compare topics



Old-fashion topics

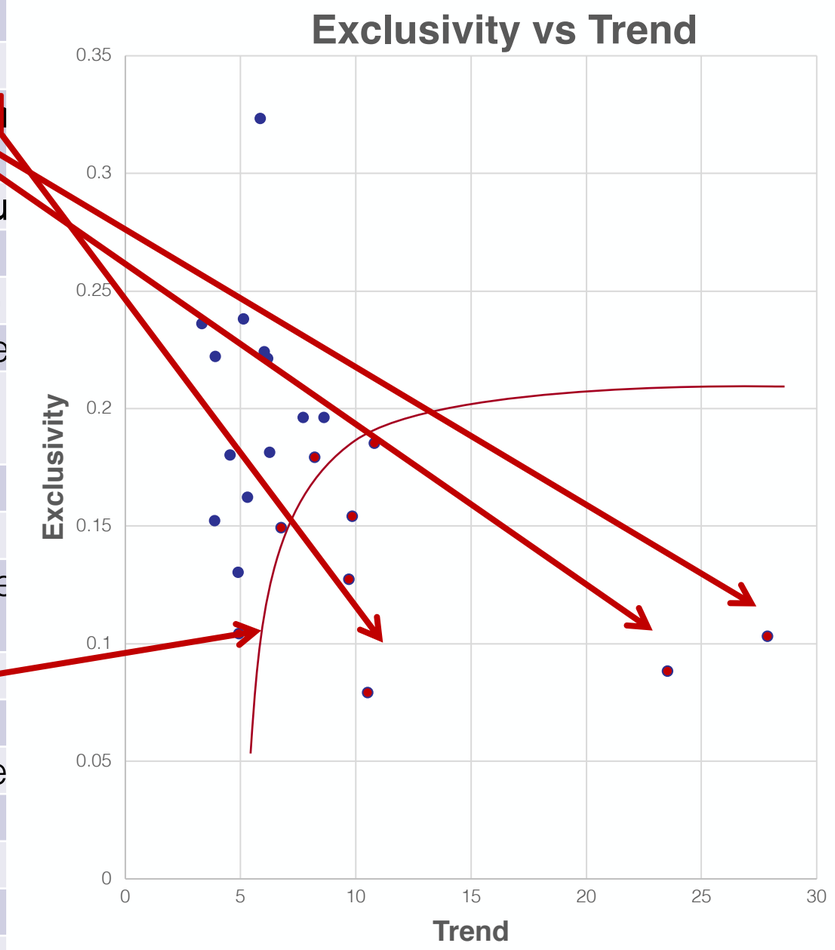


Do we need another venue?

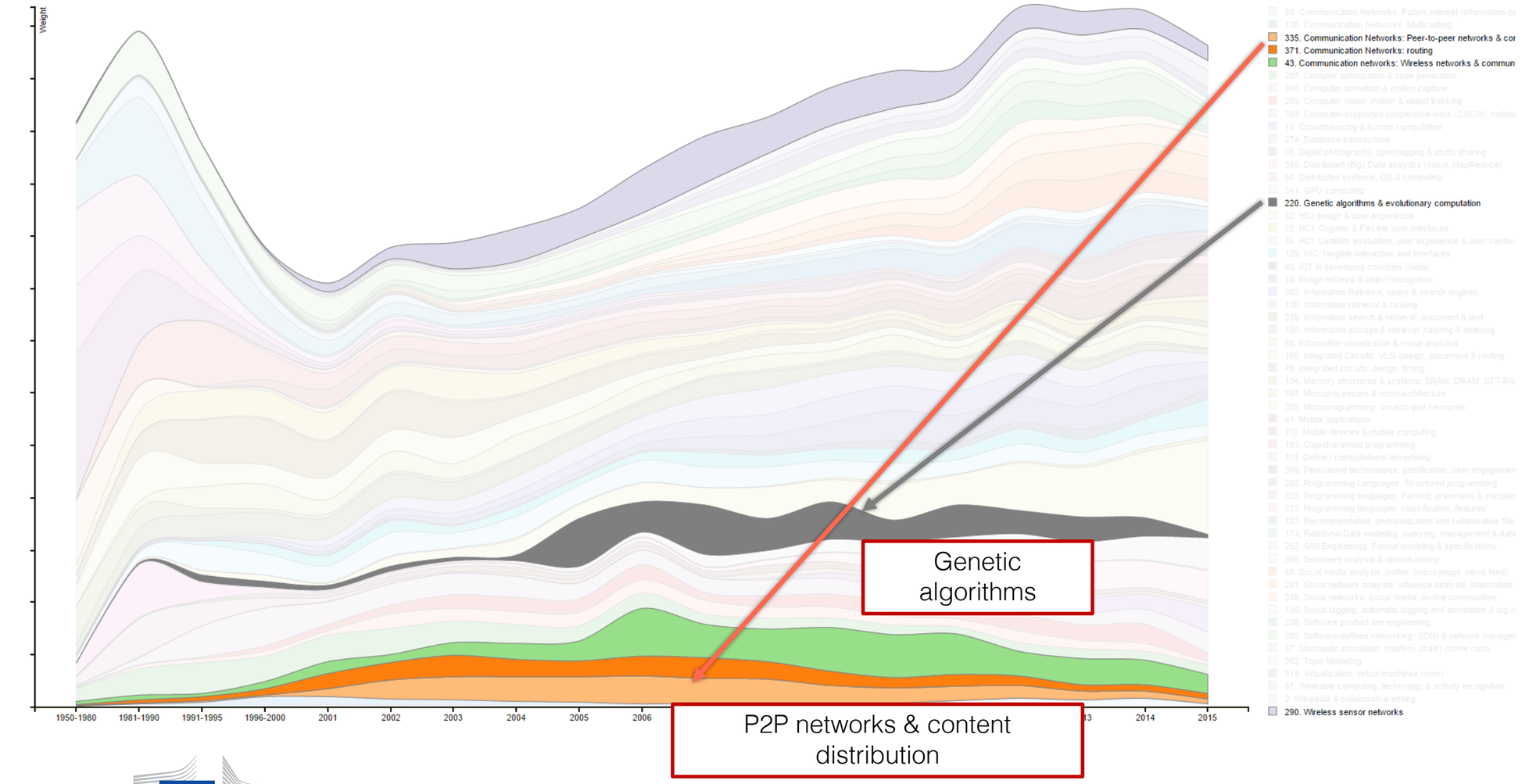
Trendy, but evenly spread across many journals AND conferences

Exclusivity + Journa

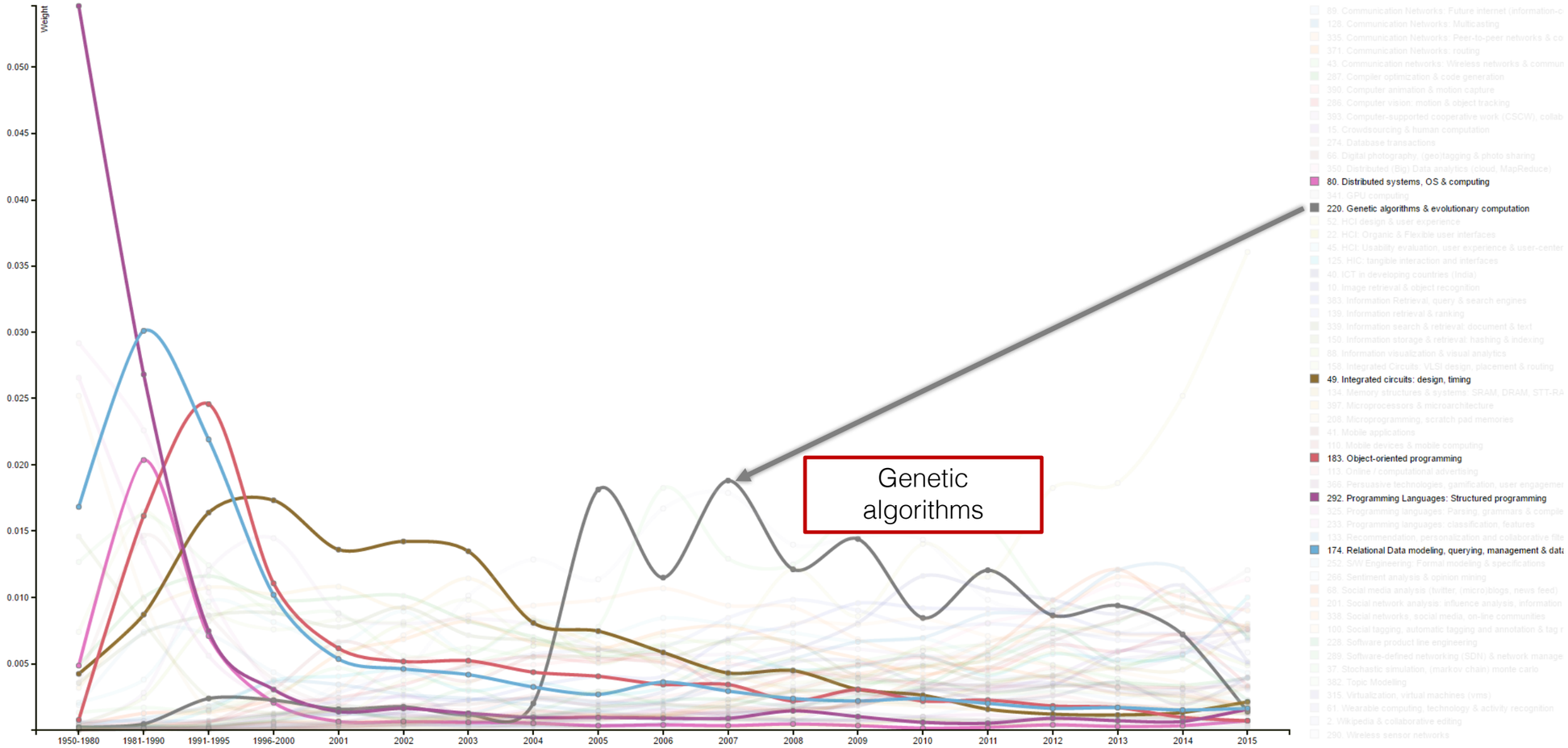
TopicId	Title				
15	**Crowdsourcing & human computation**				
194	**Cloud Computing, Storage & Virtualization**				
	Social network analysis: influence, info diffusion				
201	**communities**				
350	**Distributed (Big) Data analytics (cloud, MapRedu**				
41	**Mobile applications**				
68	**Social media analysis (twitter, blogs, news feed)**				
366	Persuasive technologies, gamification, user engage				
61	**Wearable computing, technology & activity recognition**				
40	ICT in developing countries (India)				
341	**GPU computing**				
133	Recommendation, personalization and collaborative filtering				
134	Flash memory structures, storage & systems				
22	HCI: Organic & Flexible user interfaces				
74	HCI: Touch screen interaction & interactive surface				
2	Wikipedia & collaborative editing				
52	HCI design & user experience				
266	**Sentiment analysis & opinion mining**				
10	Image retrieval & object recognition				
382	Topic Modelling	0.003	4.57	0.111	0.069
228	Software product line engineering	0.003	3.92	0.128	0.094



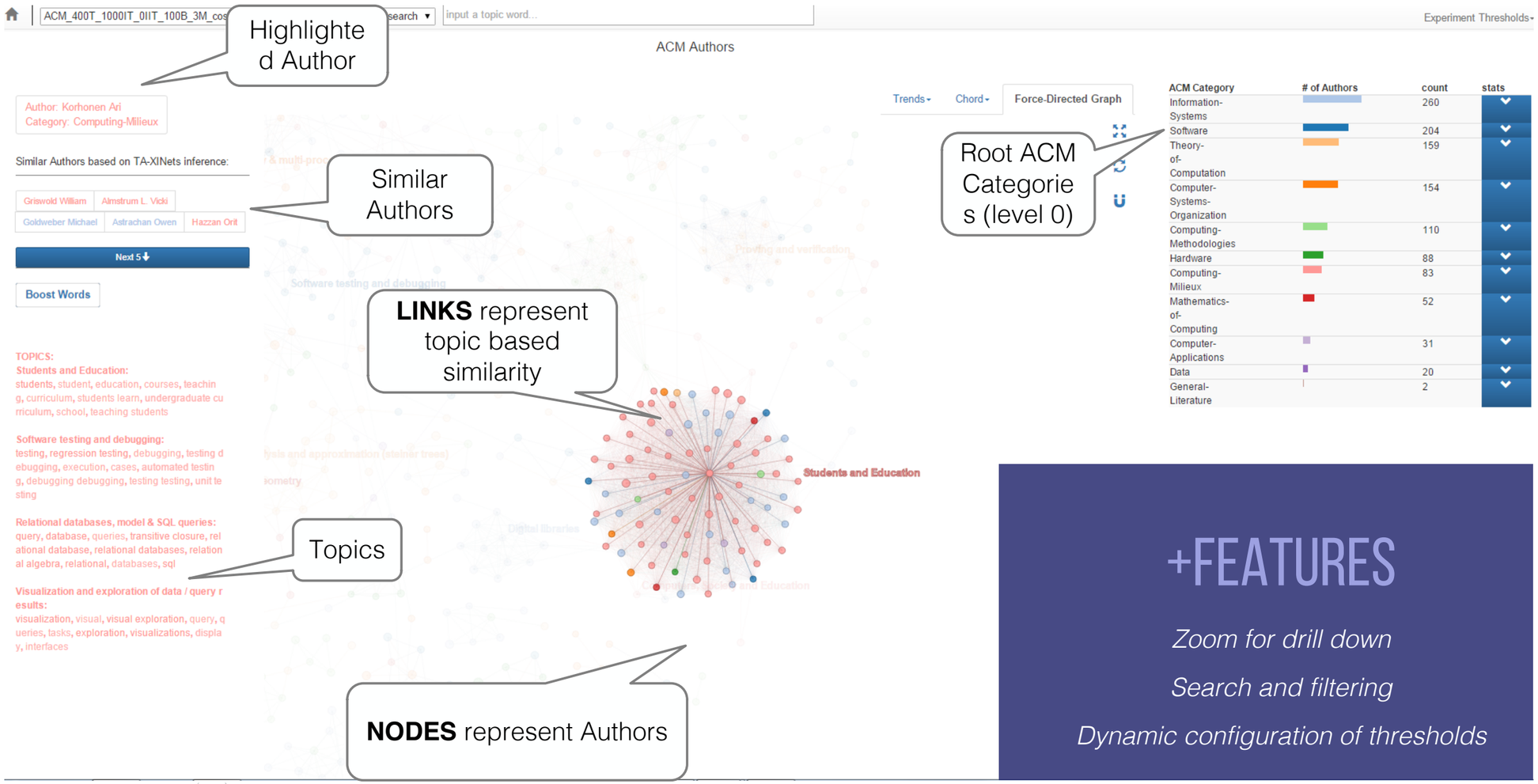
Important but declining (?)



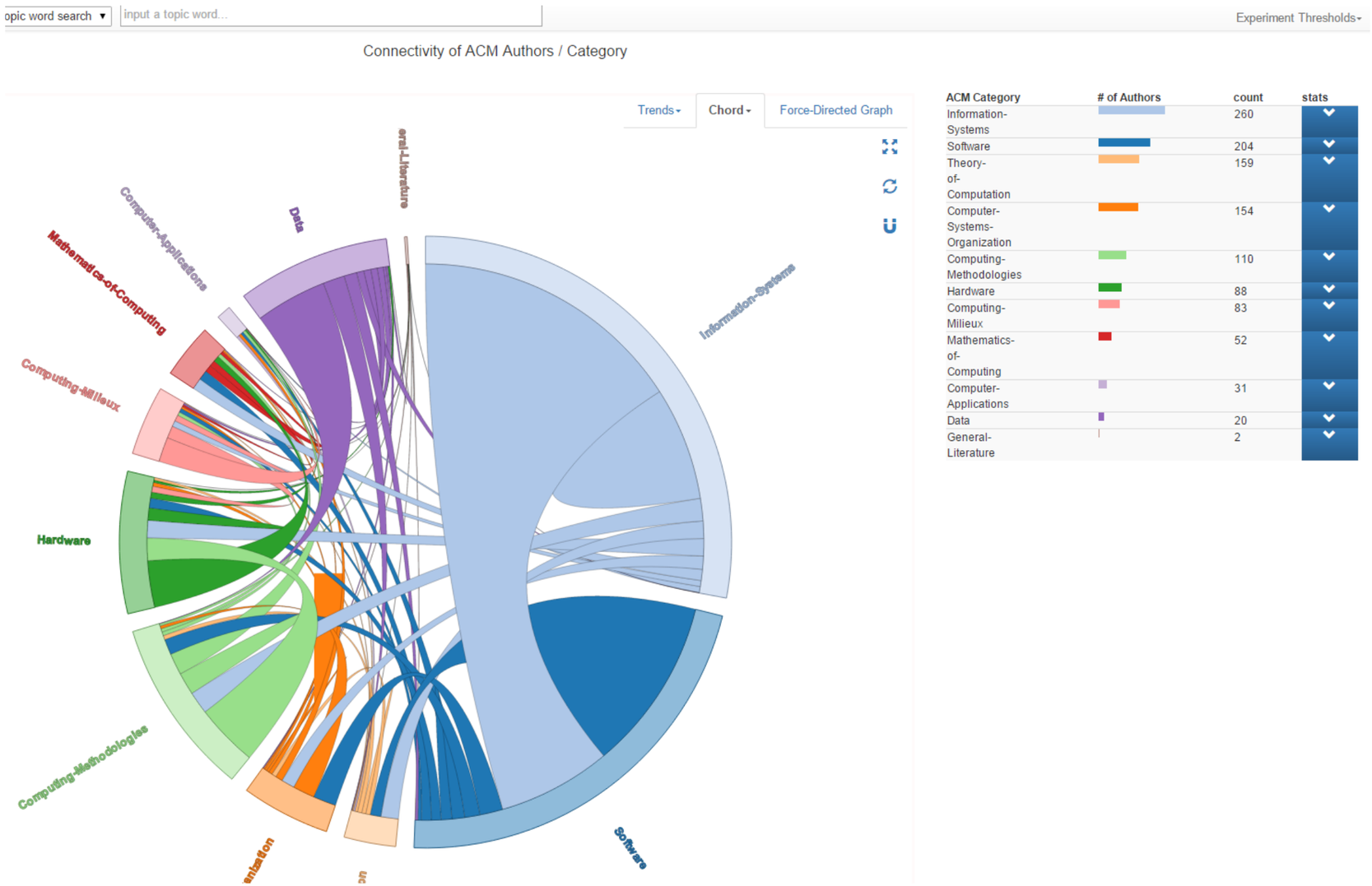
Topic birth, death & fluctuation over time



Authors Similarity Analysis



Categories correlations



ACM Category	# of Authors	count	stats
Information-Systems		260	
Software		204	
Theory-of-Computation		159	
Computer-Systems-Organization		154	
Computing-Methodologies		110	
Hardware		88	
Computing-Milieux		83	
Mathematics-of-Computing		52	
Computer-Applications		31	
Data		20	
General-Literature		2	

What is the potential?

Scratching the surface…

- **Funders and institutions to assess research impact over time**
  - Especially useful when combined with non-research data
  - **OpenAIRE data and services already used by EC for ex-post FP7 evaluation**
- **Policy makers**
  - Binding research to societal policy decisions
- **Scholarly societies**
  - Determine new conferences/merge existing ones. Introduce new themes…
  - New portal services (concept search)
- **Publishers (incl. institutional publications)**
  - Create, adapt journals…

Thank you!

Natalia Manola

natalia@di.uoa.gr
+30 210 9876 432
Skype: natalia.manola