Navigating Knowledge Space: learning higher-order structures from data

Series CIRSS

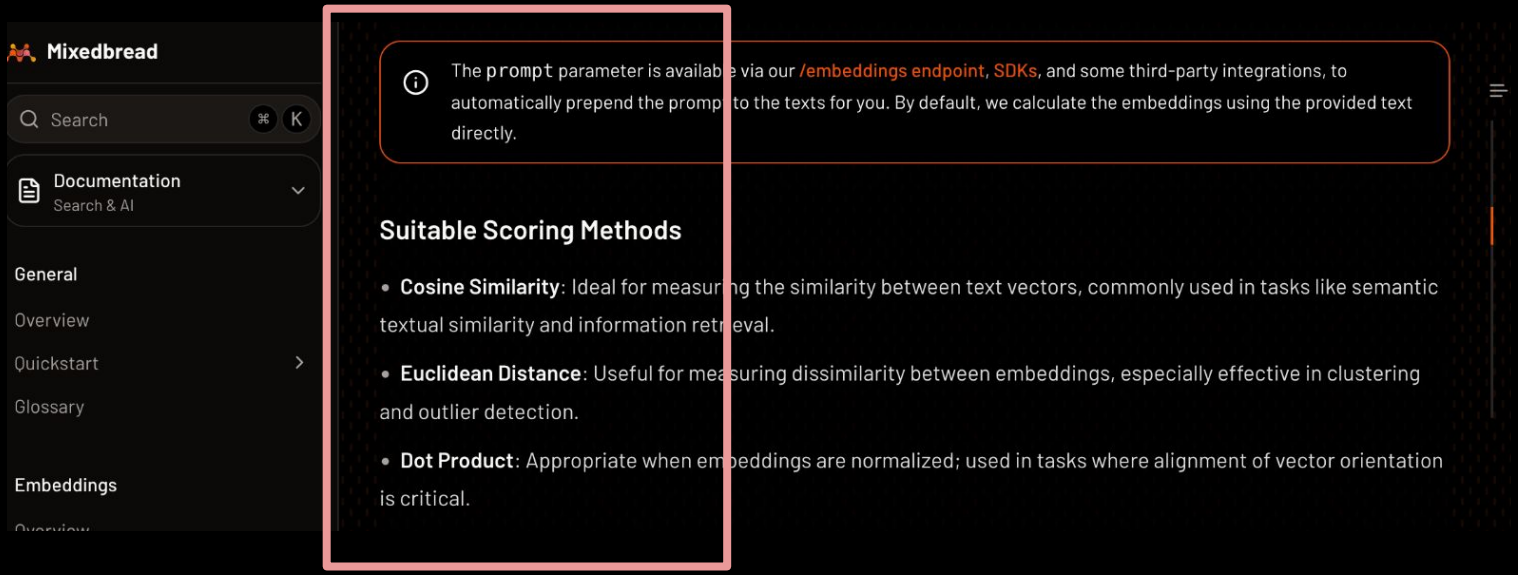
Liubov Tupikina (Bell labs, LPI/CRI, ITMO)

What is it about?

Part 1:
In this talk we will speak about the formalisation of the problem of the embedding of data (text data, …) into low dimensional spaces

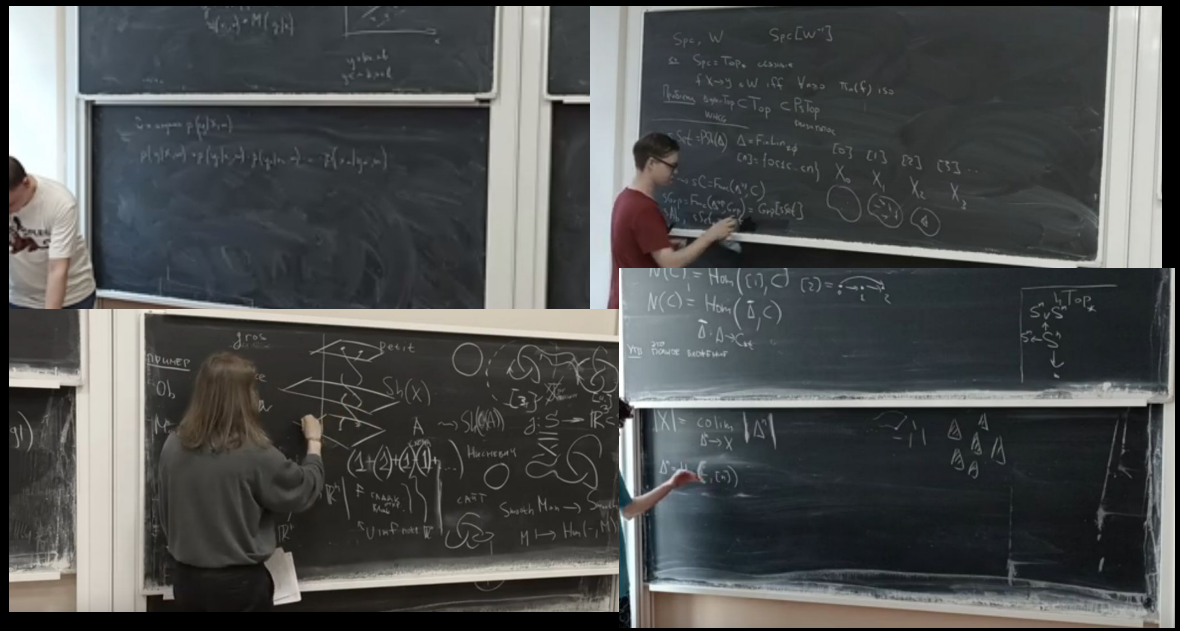
Part 2:
Applications to science of science data

In this talk we will speak about the formalisation of this problem
and about the embedding of data into low dimensional space.
And why some of the embedding techniques are not giving us good results



'All ML analysis is the approximation of one set of functions by another set of functions in some special spaces',

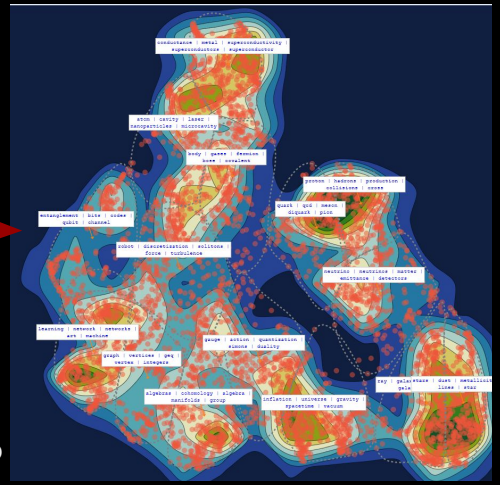
The only issue is that we do not always know what are those spaces:)



- Science of science analysis collaboration w/Marc (LPI), Hritika (Bell labs), Paris network seminar (2019-2024)
- CUDAN lab conference, Estonia 2023
- Hypermatrix workshop 2024 (Wolfram Institute, Paris, France 2024)
- Hypermatrix workshop (Wolfram Institute, Snt.Petersbourg, Moscow, Madrid 2025 TBD)
- NetSci Conferencw presentation (Liubov, Carlos)

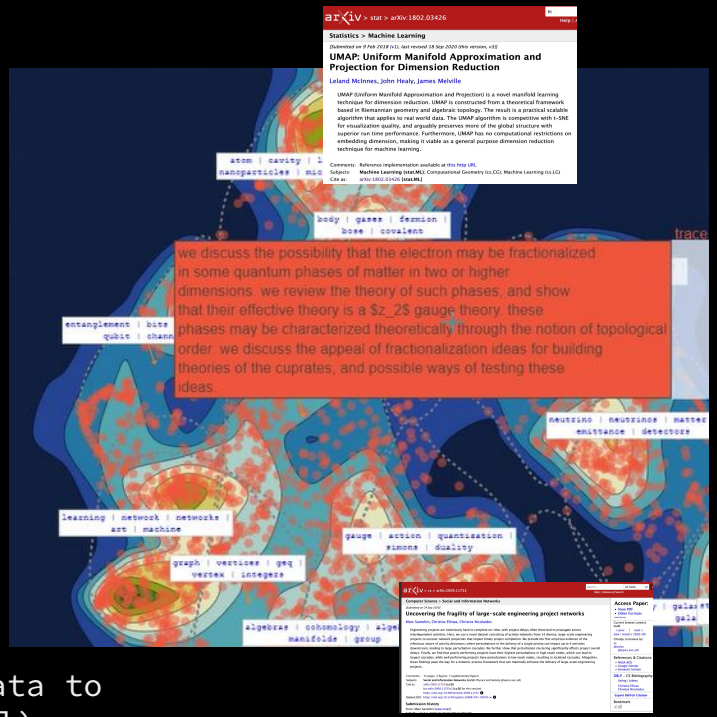
The embedding definition

	topic_id	topic_name	size	percent
0	bt-6	conductance | metal | superconductivity | supe...	462	9.24
1	bt-5	learning | network | networks | art | machine	408	8.16
2	bt-1	ray | galaxies | redshift | galaxy | radio	394	7.88
3	bt-7	inflation | universe | gravity | spacetime | v...	375	7.50
4	bt-12	stars | dust | metallicity | lines | star	373	7.46



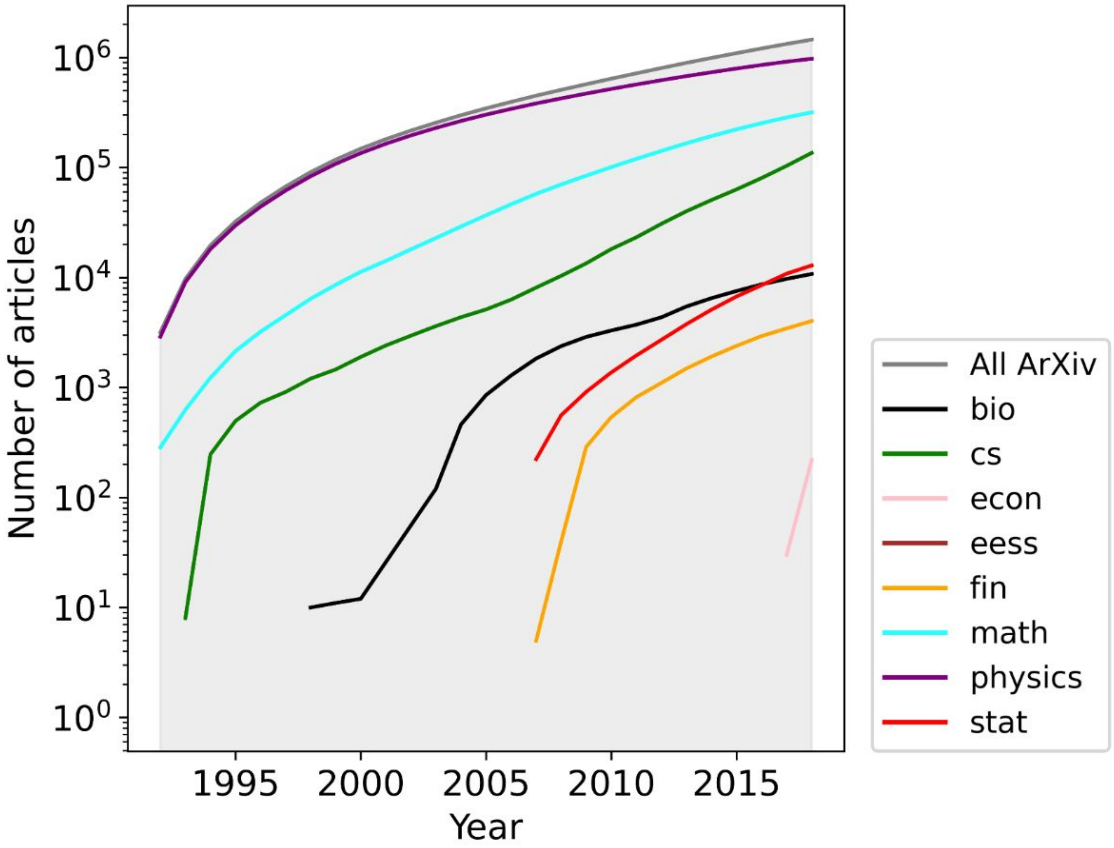
Embedding of large data to
(lower-dimensional)
‘knowledge’space

	topic_id	topic_name	size	percent
0	bt-6	conductance | metal | superconductivity | supe...	462	9.24
1	bt-5	learning | network | networks | art | machine	40[illegible]	[illegible]
2	bt-1	ray | galaxies | redshift | galaxy | radio	394	7.88
3	bt-7	inflation | universe | gravity | spacetime | v...	375	7.50
4	bt-12	stars | dust | metallicity | lines | star	373	7.46



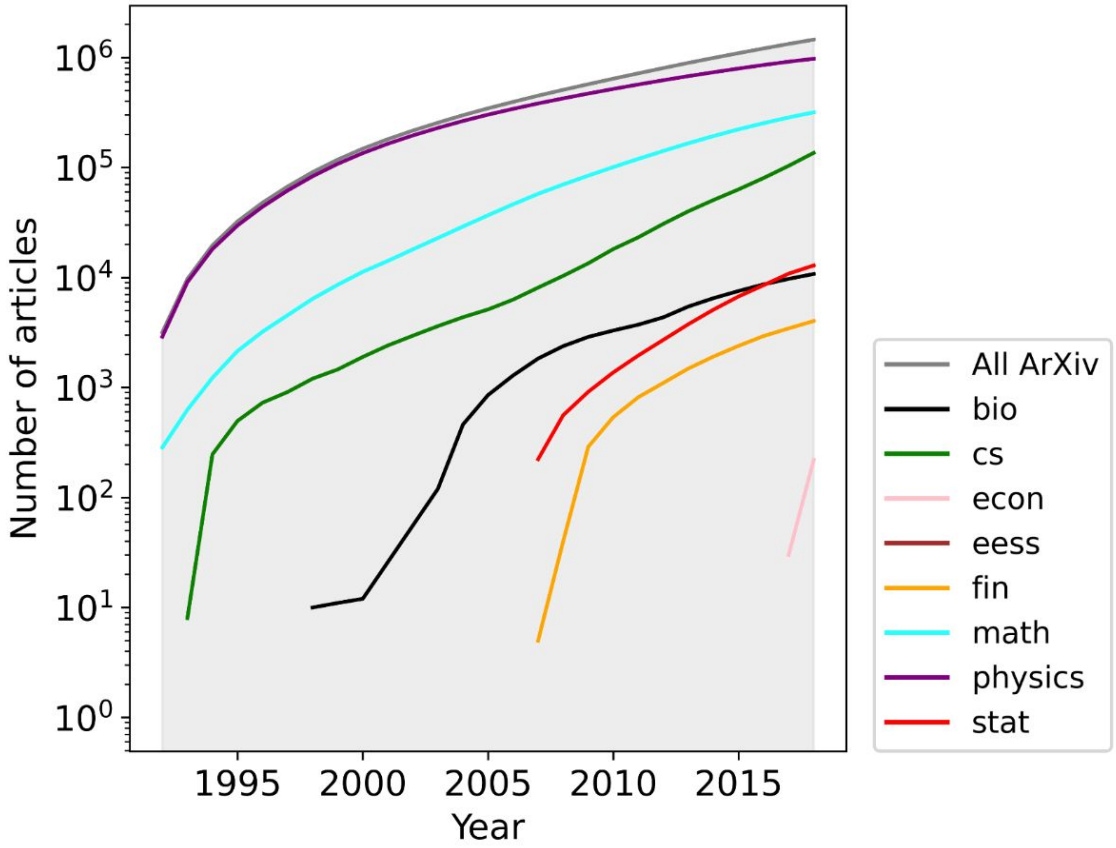
Embedding of large data to
(lower-dimensional)
‘knowledge’ space

Science of science open data

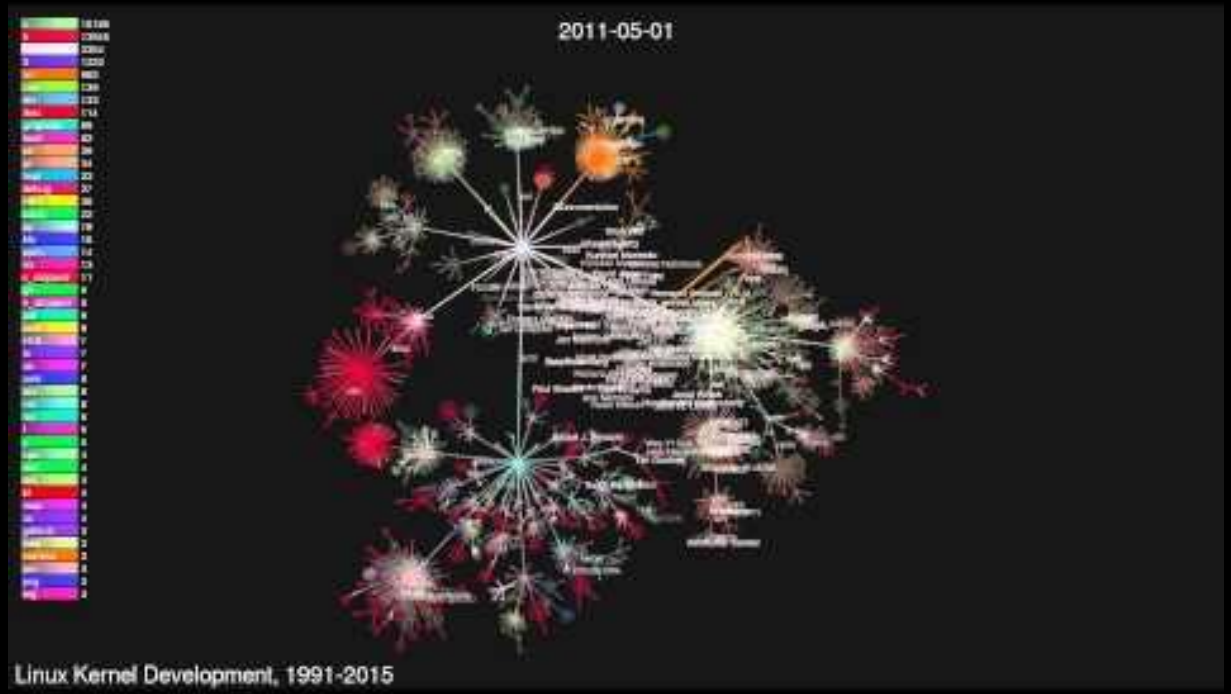


https://github.com/mattbierbaum/arxiv-public-datasets

Other science of science open datasets



The scale of open collaborations



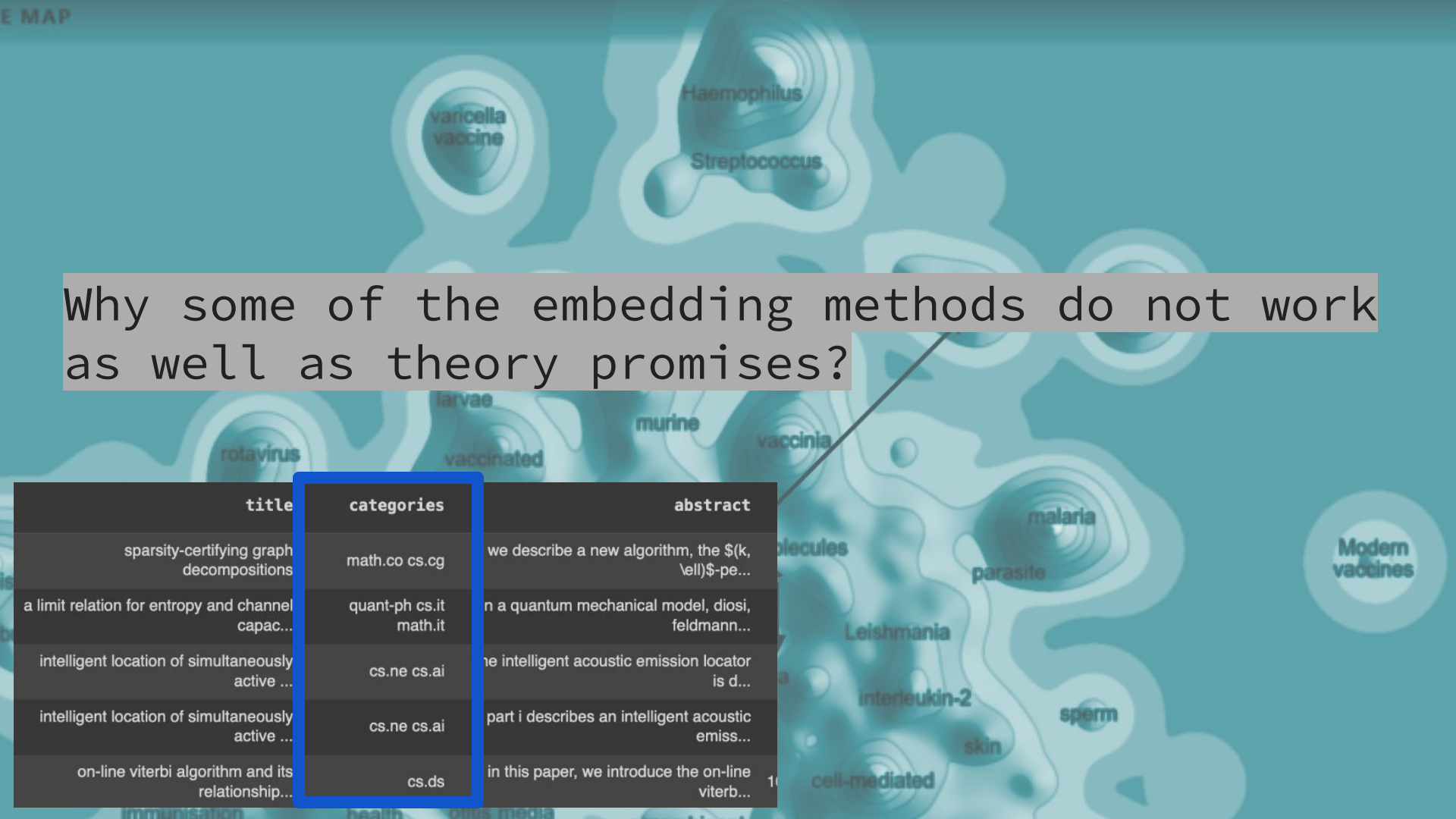
Linux

17,000+ contributors

~30 years

3 ***billion*** users

Welearn project, LPI



Why some of the embedding methods do not work as well as theory promises?

title	categories	abstract
sparsity-certifying graph decompositions	math.co cs.cg	we describe a new algorithm, the (k, ℓ)-pe...
a limit relation for entropy and channel capac...	quant-ph cs.it math.it	n a quantum mechanical model, diosi, feldmann...
intelligent location of simultaneously active ...	cs.ne cs.ai	e intelligent acoustic emission locator is d...
intelligent location of simultaneously active ...	cs.ne cs.ai	part i describes an intelligent acoustic emiss...
on-line viterbi algorithm and its relationship...	cs.ds	in this paper, we introduce the on-line viterb...

Why some of the embedding methods do not work as well as theory promises?

[colab]

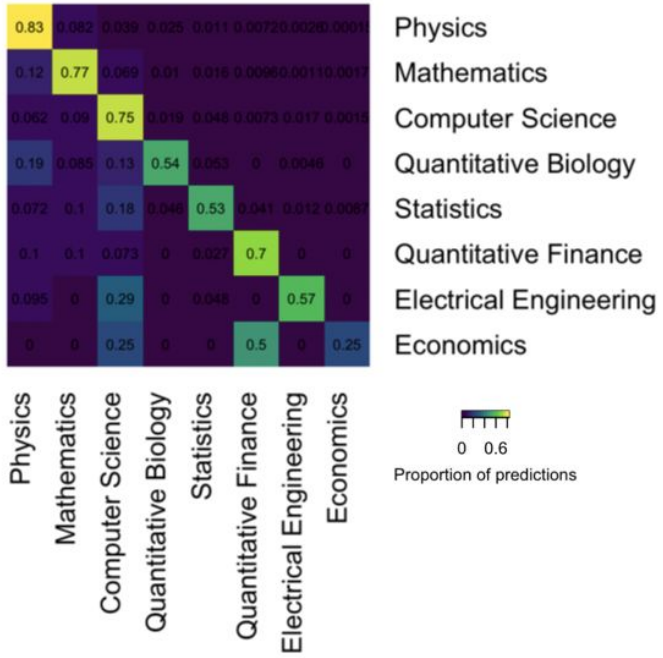
[github]

title	categories	abstract
sparsity-certifying graph decompositions	math.co cs.cg	we describe a new algorithm, the (k, ℓ)-pe...
a limit relation for entropy and channel capac...	quant-ph cs.it math.it	in a quantum mechanical model, diosi, feldmann...
intelligent location of simultaneously active ...	cs.ne cs.ai	the intelligent acoustic emission locator is d...
intelligent location of simultaneously active ...	cs.ne cs.ai	part i describes an intelligent acoustic emiss...
on-line viterbi algorithm and its relationship...	cs.ds	in this paper, we introduce the on-line viterb...

```python
def jump_distance_vector(trajectory):
    j_dist=[]
    for i in range(len(trajectory)-1):
        p1,p2=trajectory[i],trajectory[i+1]
        eucl = np.sqrt(((p2[0]-p1[0])**2) + ((p2[1]-p1[1])**2))
        j_dist.append(eucl)

    return j_dist

def z_scoreD(dist_vec):

    u,sig = np.mean(dist_vec),np.std(dist_vec)
    zscore = list(map(lambda x:(x-u)/sig,dist_vec))

    return zscore

def distance_from0_vector(trajectory):
    dist=[]
    for i in range(1,len(trajectory)):
        p0,p1=trajectory[0],trajectory[i]
        eucl = np.sqrt(((p1[0]-p0[0])**2) + ((p1[1]-p0[1])**2))
        dist.append(eucl)

    return dist

def jump_duration(days):
    time=[]
    for i in range(len(days)-1):
        d0,d1=days[i],days[i+1]
        time.append(d1-d0)
```

a

b



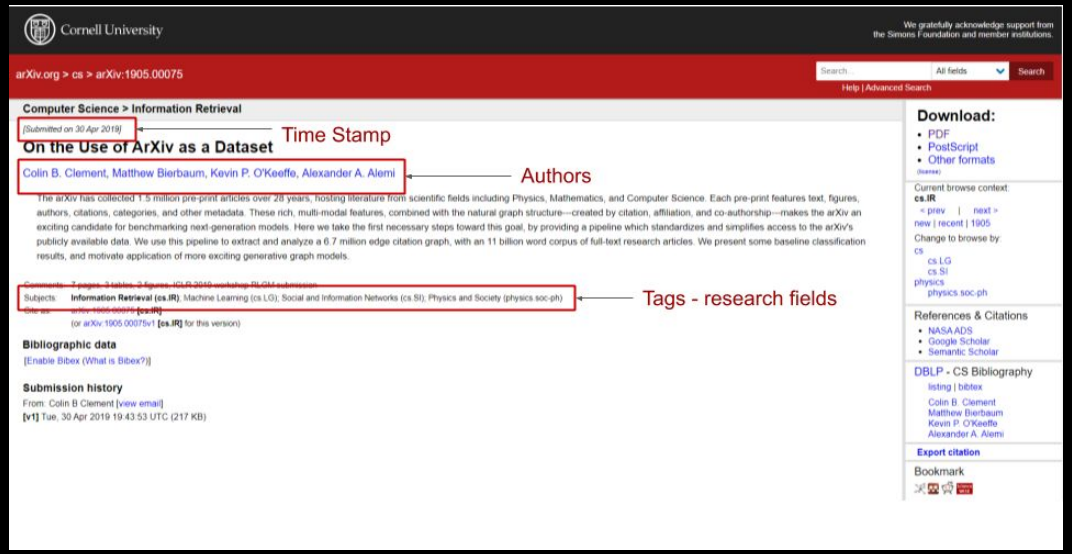

```
DF['Y_tSNE']=y_tsne

DF.head()
```

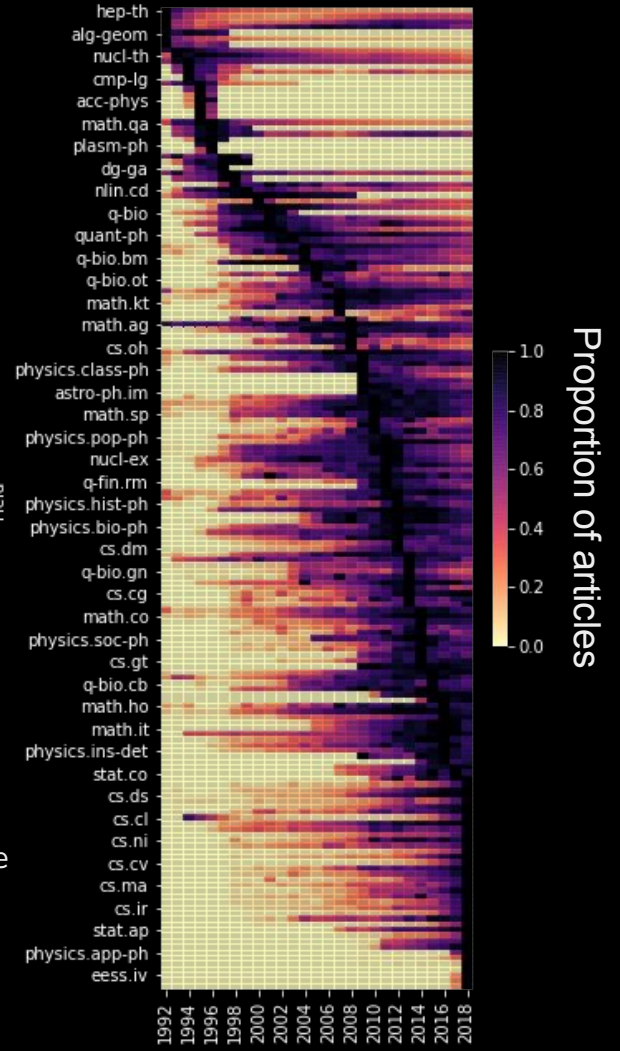
	id	categories	created	year	X_tSNE	Y_tSNE
1251136	hep-th/9201004	hep-th	1992-01-02	1992	-5.179147	39.400970
1251135	hep-th/9201003	hep-th	1992-01-02	1992	-5.179147	39.400970
1251134	hep-th/9201002	hep-th	1992-01-02	1992	-5.179147	39.400970
1251137	hep-th/9201005	hep-th	1992-01-03	1992	-5.179147	39.400970
526877	math/9201202	math.fa	1992-01-05	1992	-34.935791	43.862728

Physics
Mathematics
Computer Science
Quantitative Biology
Statistics
Economics
Quantitative Finance

Singh, [...], **Santolini+** PLOS ONE 2022
Singh, [...], **Santolini+** EPJ (in rev) 2023

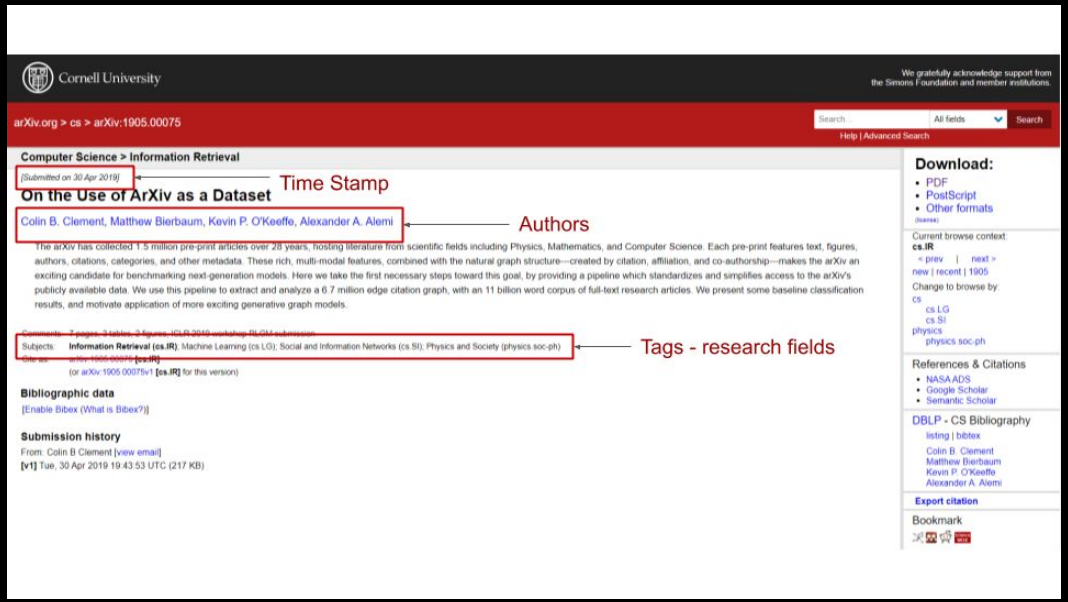
Collective level: Rise and Fall of Research Fields

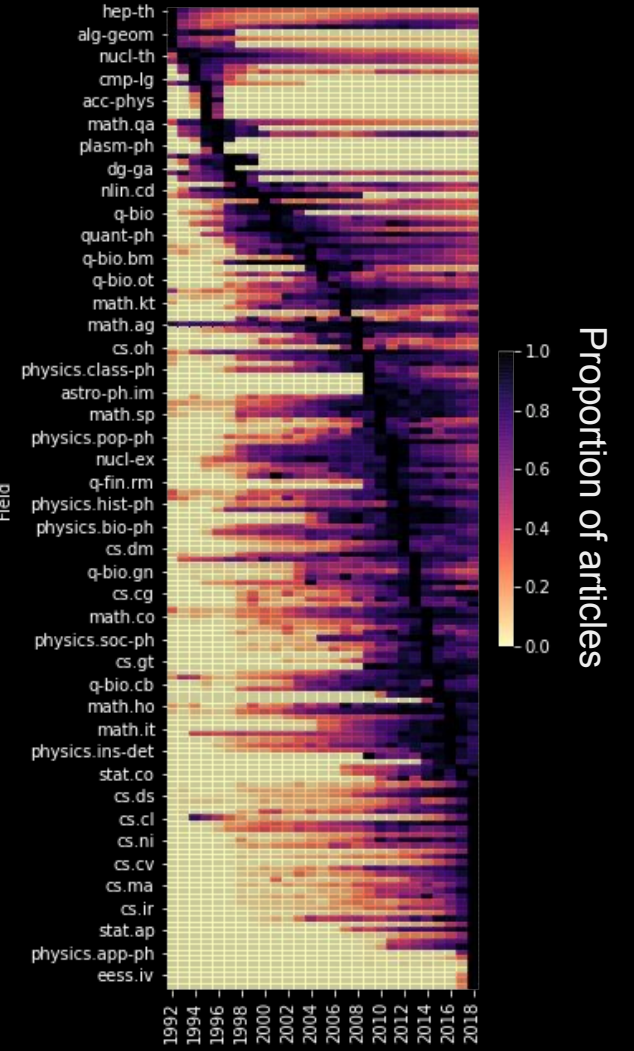




Clement et al. **On the Use of ArXiv as a Dataset**. *arXiv:1905.00075* (2019),
Singh (...) Tupikina, Santolini **Quantifying the rise and fall of scientific fields**", Plos One (2022)

Collected using the arXiv API

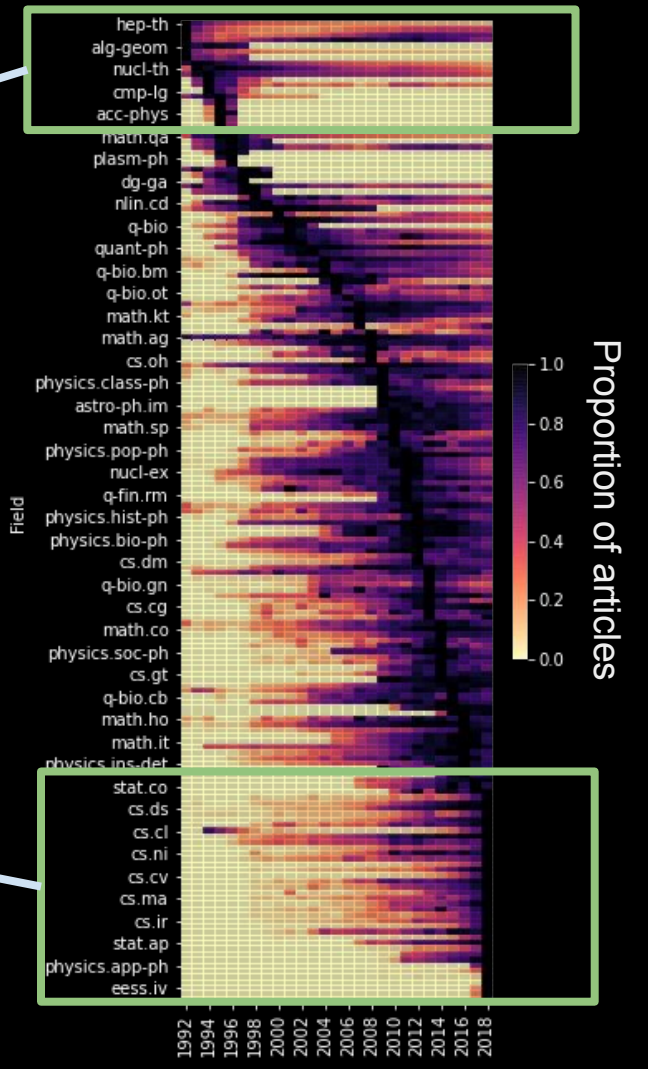




Clement, C. B., Bierbaum, M., O'Keeffe, K. P., & Alemi, A. A. (2019). **On the Use of ArXiv as a Dataset**. *arXiv preprint arXiv:1905.00075*.

High Energy Physics, Accelerator Physics, Nuclear Physics, Algebraic Geometry

Computation and Language, Vision and Pattern Recognition, Data Structures, Computation, Applications, Applied Physics



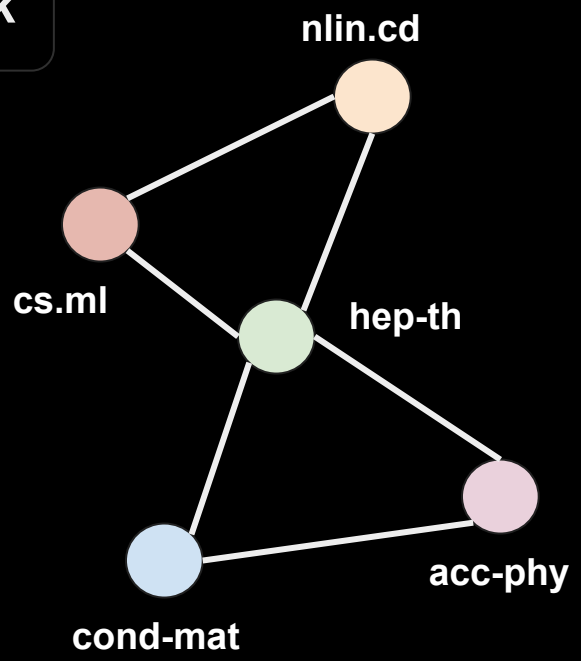
Building the Co-Tag Network

Article i

{ hep-th; nlin.cd; cs.ml }

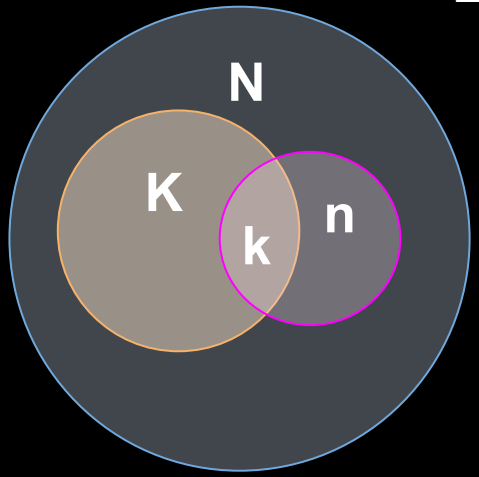
Article j

{ hep-th; cond-mat; acc-phy }



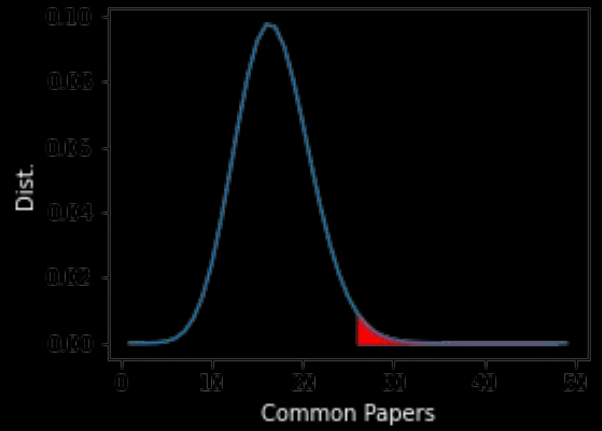
Edge Weight is defined as a function of common papers b/w two field tags

Edge Weight = $-log_{10}(p_{ij})$

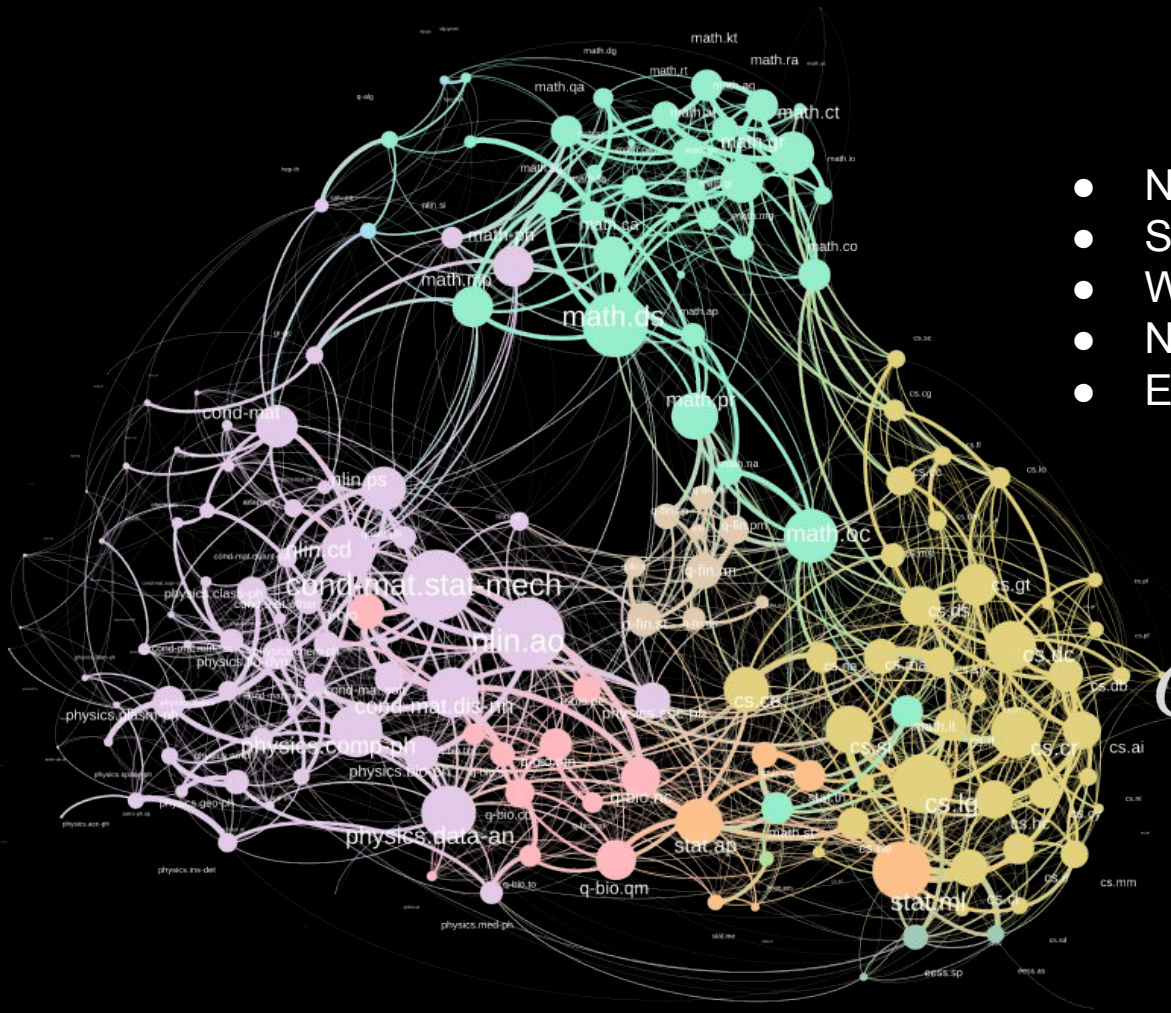


$$p_{ij} = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

- N - Total Articles
- K - Articles in field i
- n - Articles in field j
- k - common articles bw i and j



Note - Here lower p-values are more significant. We eliminate edges with p > 0.01

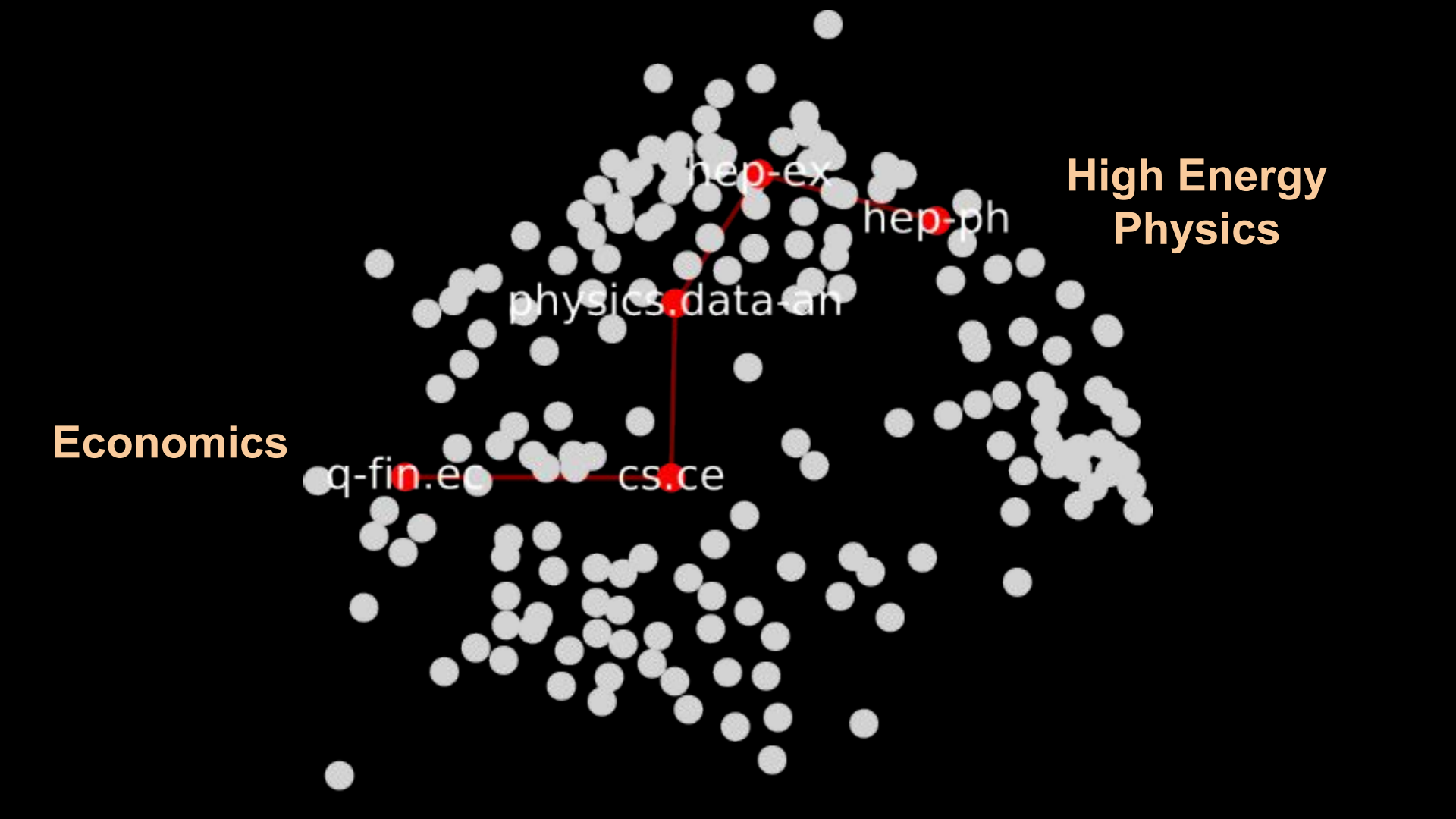


- Nodes are field tags
- Size is proportional to degree
- Weighted and undirected
- Node color is the main research area
- Edge thickness proportional to weight

Cognitive Distance

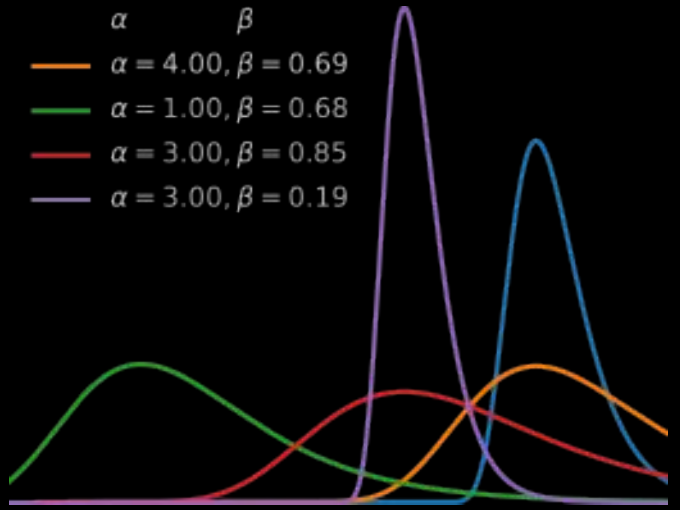
$$CD_{ij} = min(\sum_e \frac{1}{W_e})$$

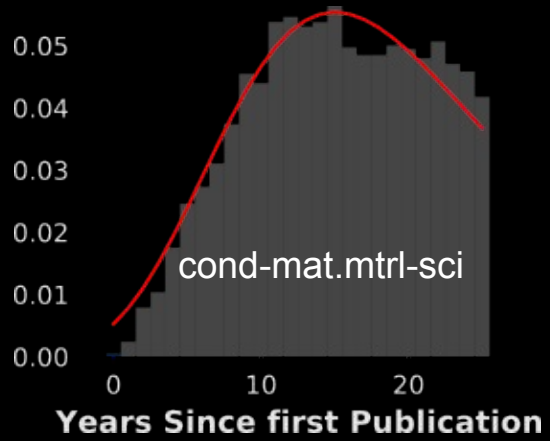
e : edge on shortest path


hep-ex
hep-ph
physics.data-an
q-fin.ec
cs.ce
High Energy Physics
Economics

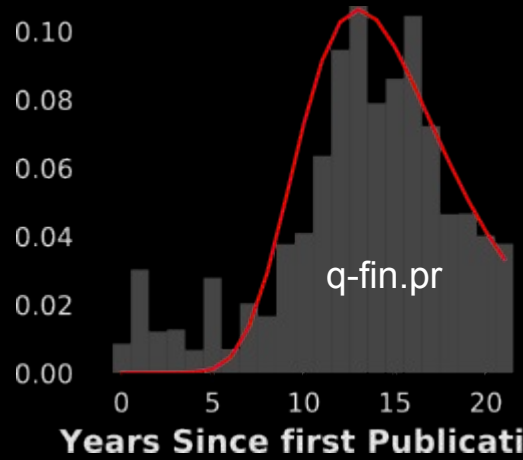
Gumbel distribution function

$$G = \frac{1}{\beta} e^{\frac{-(x-\alpha)}{\beta}} e^{-e^{\frac{-(x-\alpha)}{\beta}}}$$

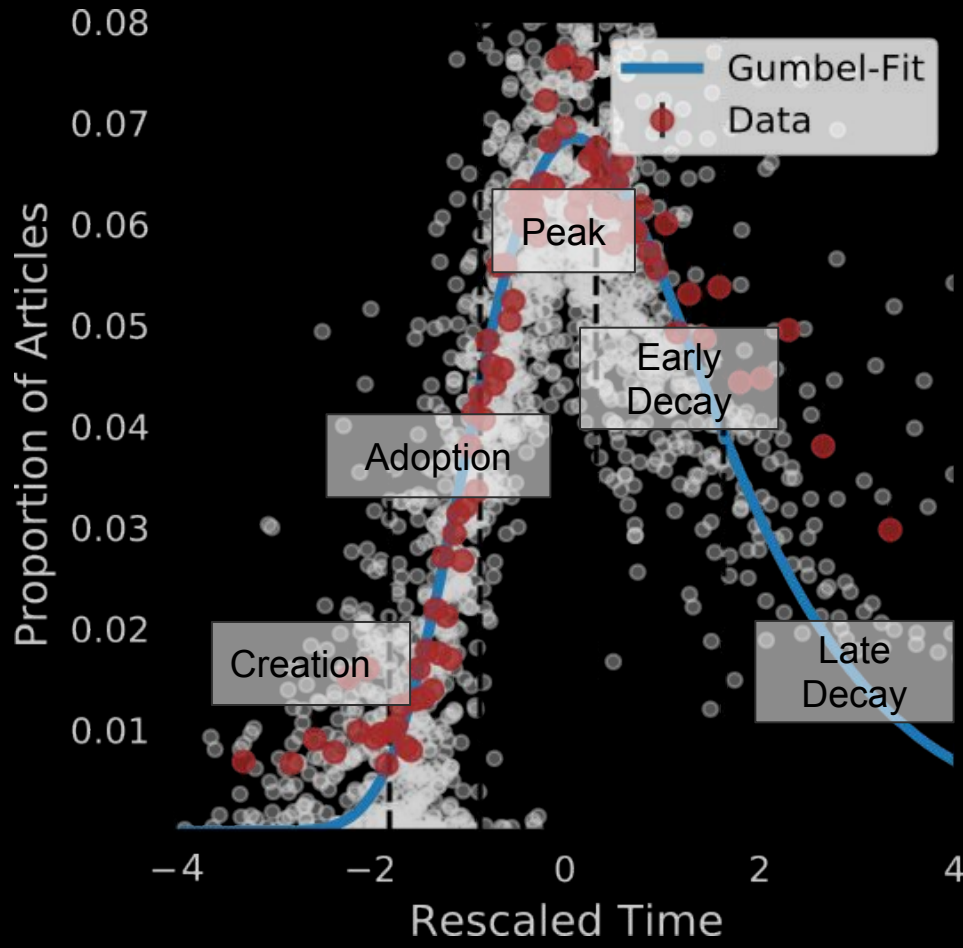




Material Sciences



Quantitative Finance (Pricing)

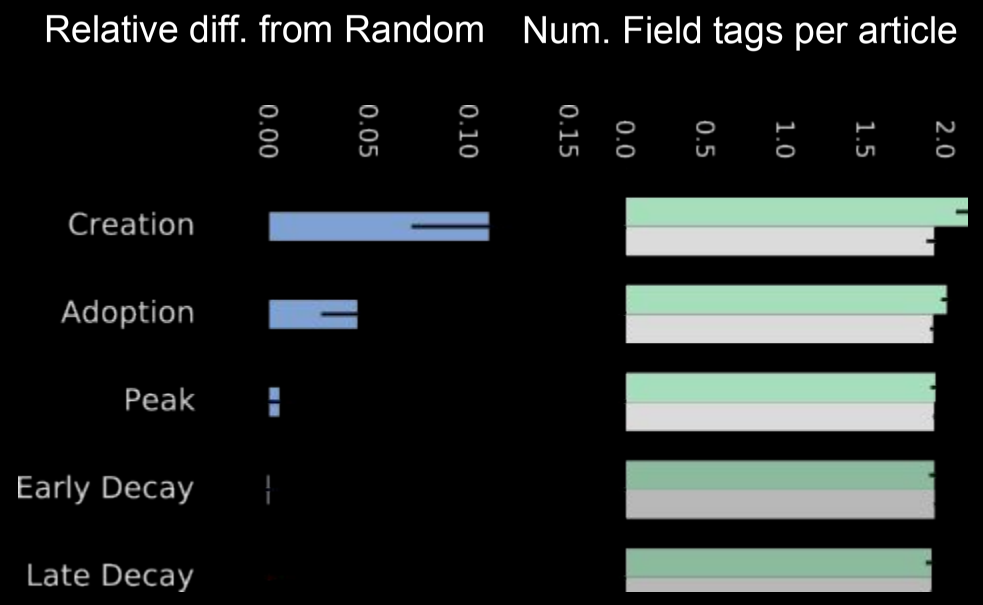


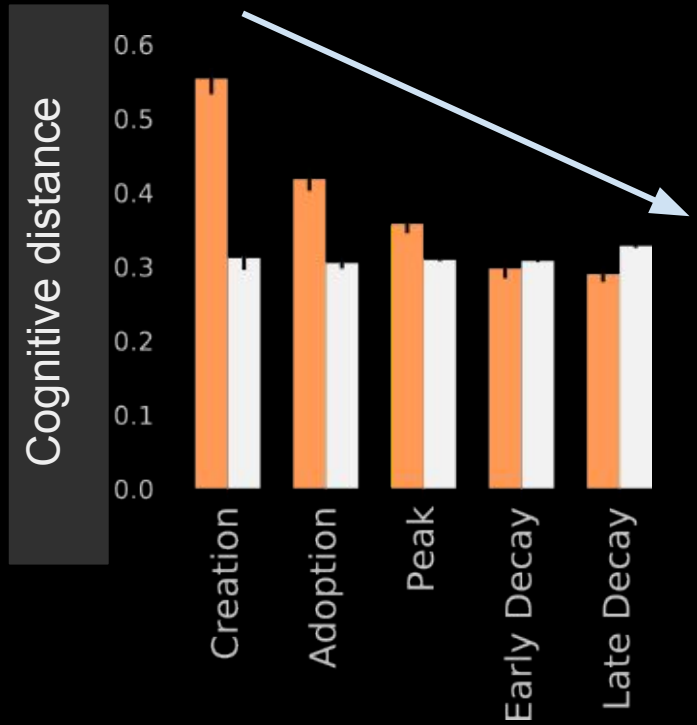
Rescaled time

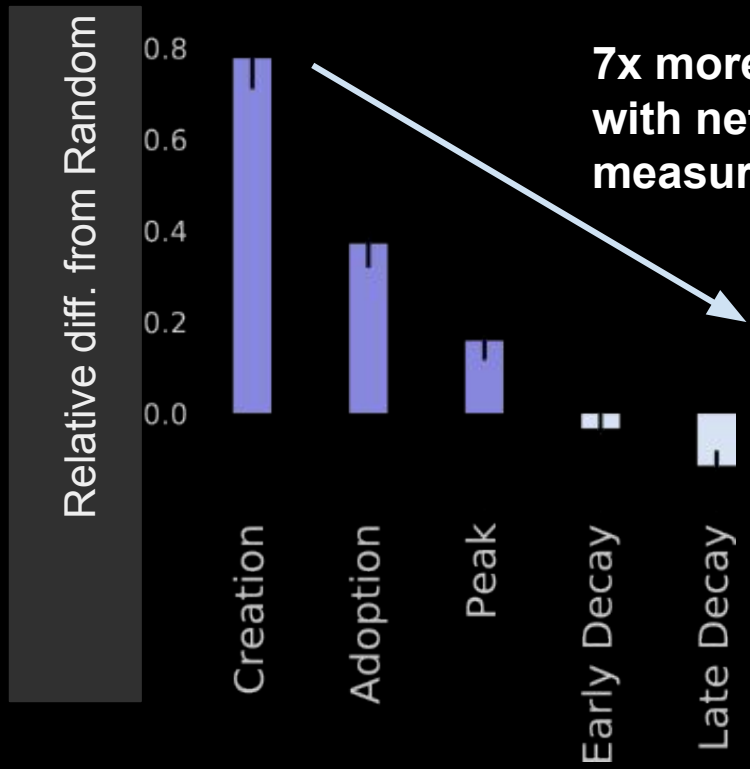
$$t' = \frac{t-\alpha}{\beta}$$

Field stages are defined at 2.5%, 16%,50% and 84% of the fit curve (blue). These number are borrowed from the **diffusion of innovation** literature


Relative diff. from Random
Num. Authors per article
−0.2
0.0
0.2
0.4
0
1
2
3
4
5
Creation
Adoption
Peak
Early Decay
Late Decay


Relative diff. from Random
Num. Field tags per article
0.00
0.05
0.10
0.15
0.0
0.5
1.0
1.5
2.0
Creation
Adoption
Peak
Early Decay
Late Decay

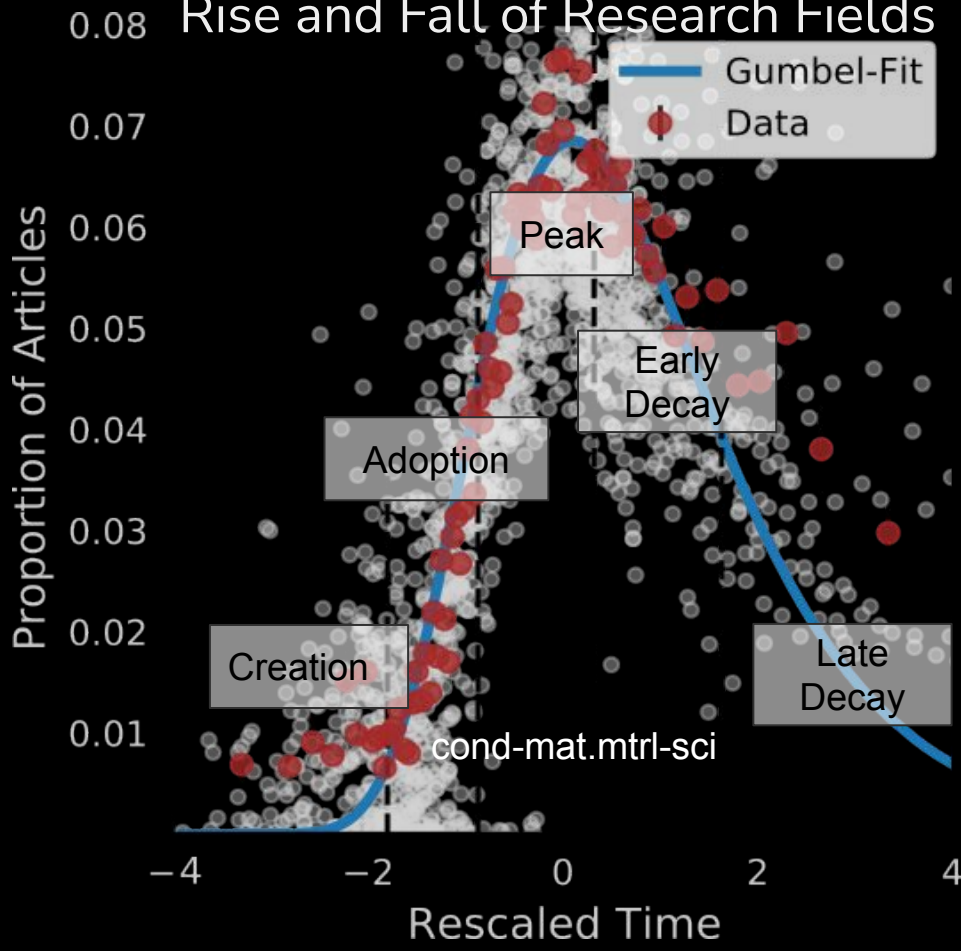




7x more enrichment with network based measure

Researchers in creation phase connect distant fields together (**explore**) whereas in later phases they focus on closely related fields (**exploit**)

Rise and Fall of Research Fields



$$G = \frac{1}{\beta} e^{\frac{-(x-\alpha)}{\beta}} e^{-e^{\frac{-(x-\alpha)}{\beta}}}$$

$$t' = \frac{t-\alpha}{\beta}$$

Field stages are defined at 2.5%, 16%,50% and 84% of the fit curve (blue). These number are borrowed from the **diffusion of innovation** literature