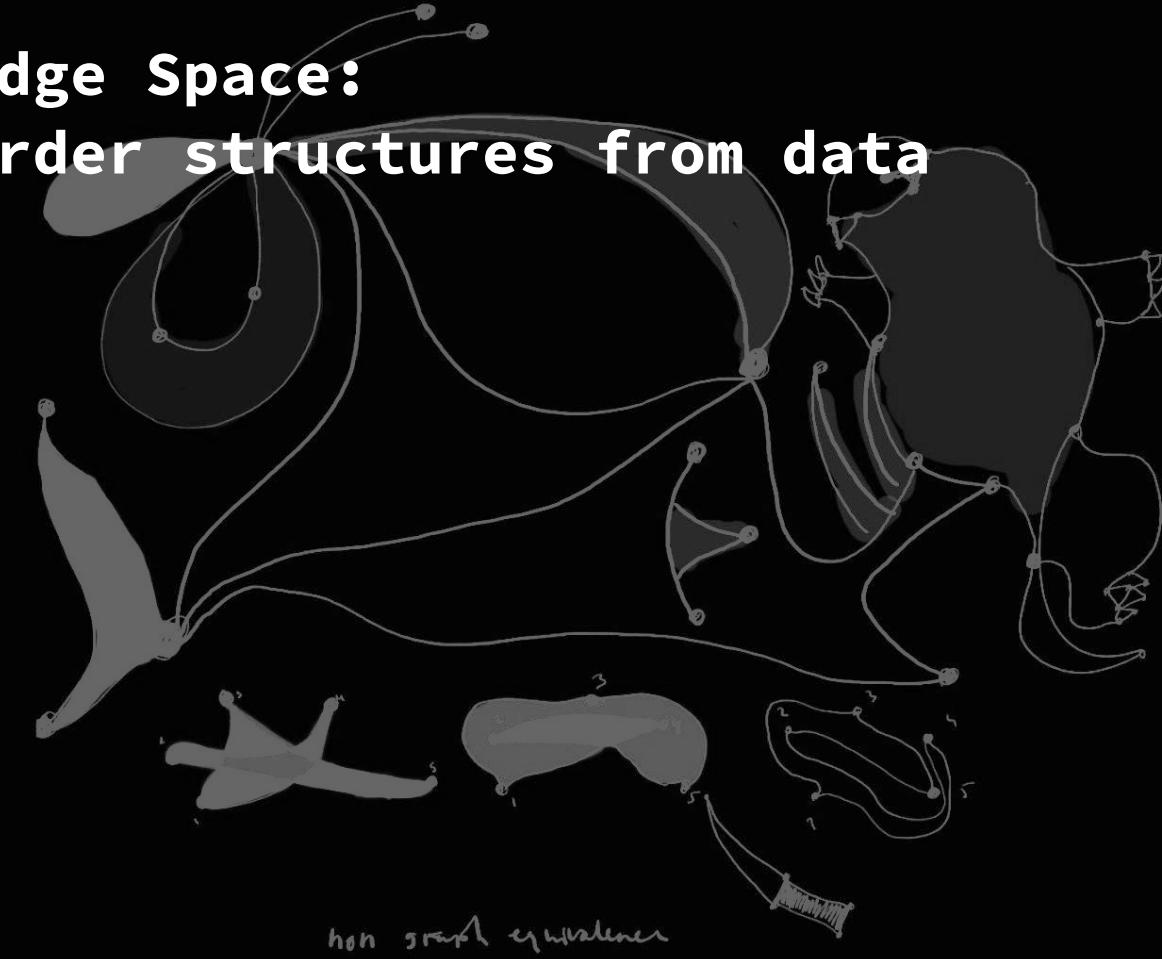


Navigating Knowledge Space: learning higher-order structures from data

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

The screenshot shows a dark-themed documentation page for 'Mixedbread'. On the left, there's a sidebar with links like 'General', 'Overview', 'Quickstart', 'Glossary', and 'Embeddings'. The main content area has a heading 'Suitable Scoring Methods' with a list of three bullet points. A red rectangular box highlights a tooltip in the top right corner.

Mixedbread

Search K

Documentation Search & AI

General

Overview

Quickstart >

Glossary

Embeddings

Overview

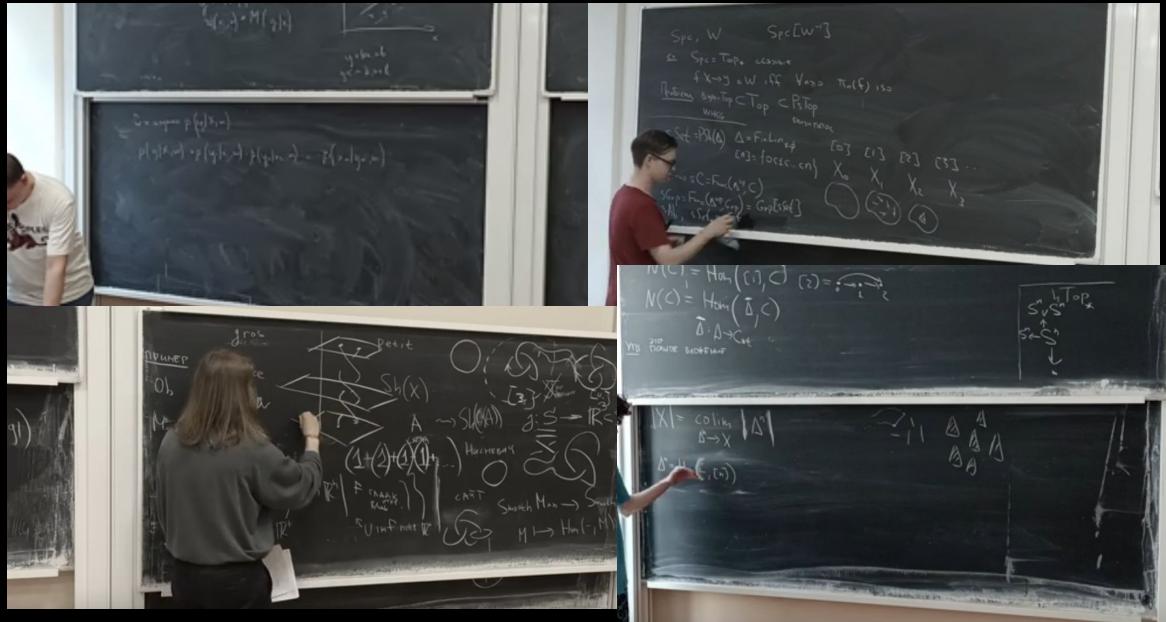
Suitable Scoring Methods

- **Cosine Similarity:** Ideal for measuring the similarity between text vectors, commonly used in tasks like semantic textual similarity and information retrieval.
- **Euclidean Distance:** Useful for measuring dissimilarity between embeddings, especially effective in clustering and outlier detection.
- **Dot Product:** Appropriate when embeddings are normalized; used in tasks where alignment of vector orientation is critical.

The tooltip contains the following text:
The **prompt** parameter is available via our [/embeddings endpoint](#), [SDKs](#), and some third-party integrations, to automatically prepend the prompt to the texts for you. By default, we calculate the embeddings using the provided text directly.

'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:)

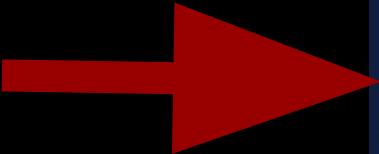


- Science of science analysis collaboration w/Marc (LPI), Hritika (Bell labs), Paris [network seminar](#) (2019-2024)
- [CUDAN lab conference, Estonia](#) 2023
- [Hypermatrix workshop 2024](#) (Wolfram Institute, Paris, France 2024)
- Hypermatrix workshop (Wolfram Institute, Snt.Petersbourg, Moscow,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

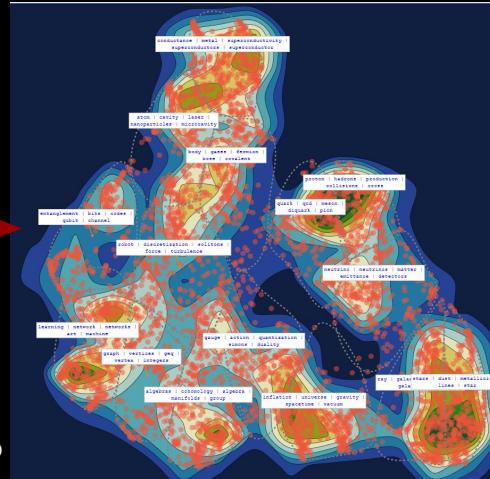


The embedding definition

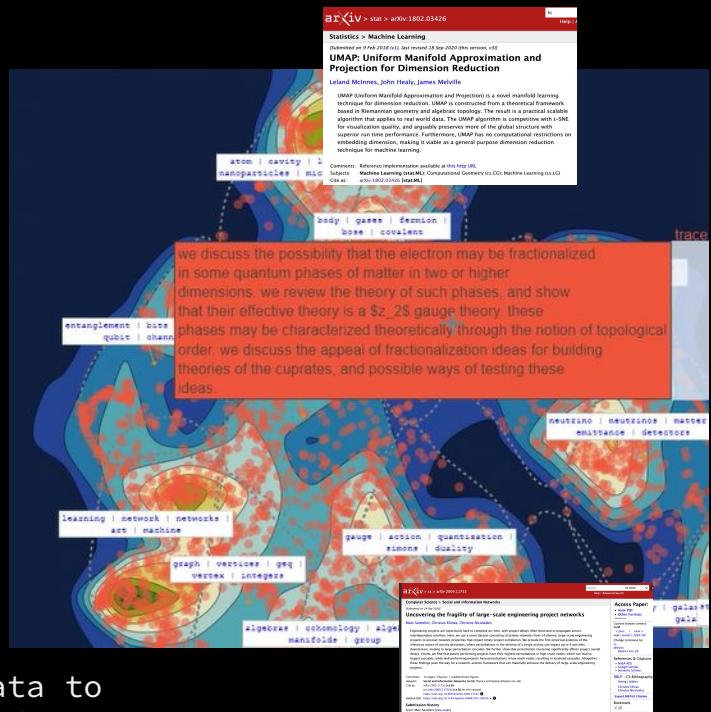
	topic_id	topic_name	size	percent
0	bt-6	conductance metal superconductivity supe...	462	9.24
1	bt-5	learning network networks art machine	408	8.16
2	bt-1	ray galaxies redshift galaxy radio	394	7.88
3	bt-7	inflation universe gravity spacetime v...	375	7.50
4	bt-12	stars dust metallicity lines star	373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



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

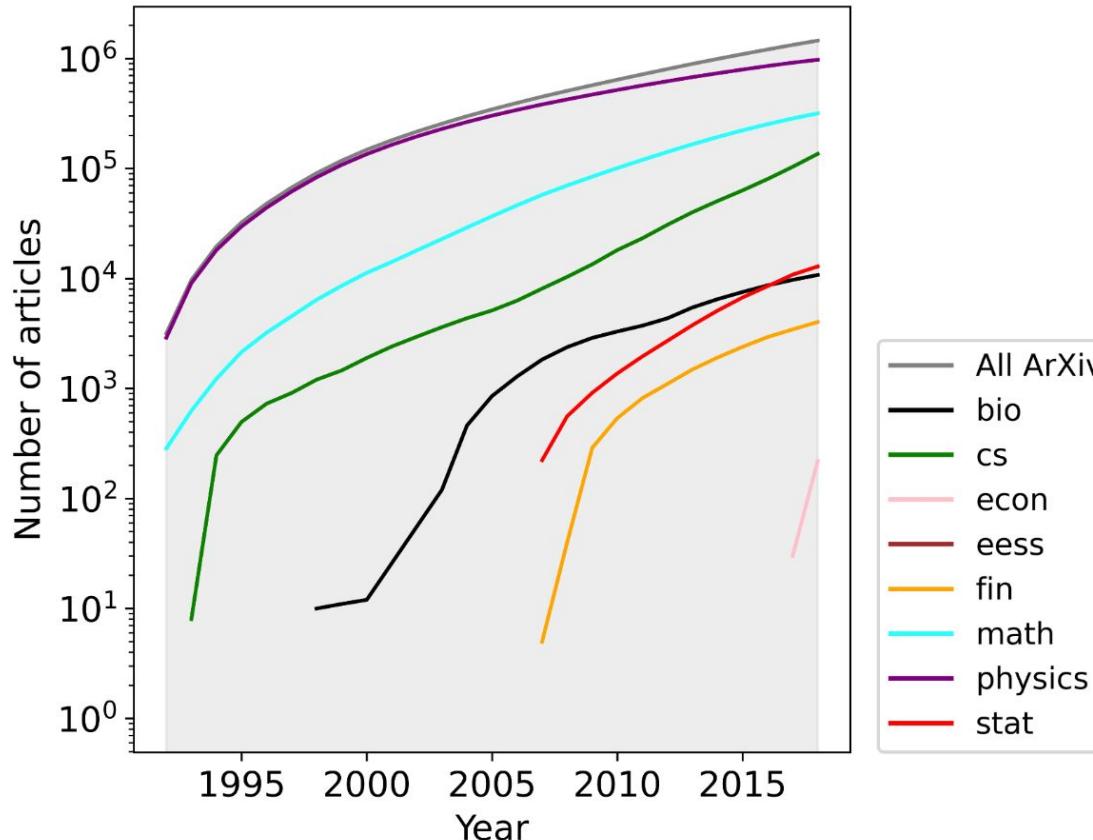


	topic_id	topic_name	size	percent
0	bt-6	conductance metal superconductivity supe...	462	9.24
1	bt-5	learning network networks art machine	408	8.16
2	bt-1	ray galaxies redshift galaxy radio	394	7.88
3	bt-7	inflation universe gravity spacetime v...	375	7.50
4	bt-12	stars dust metallicity lines star	373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



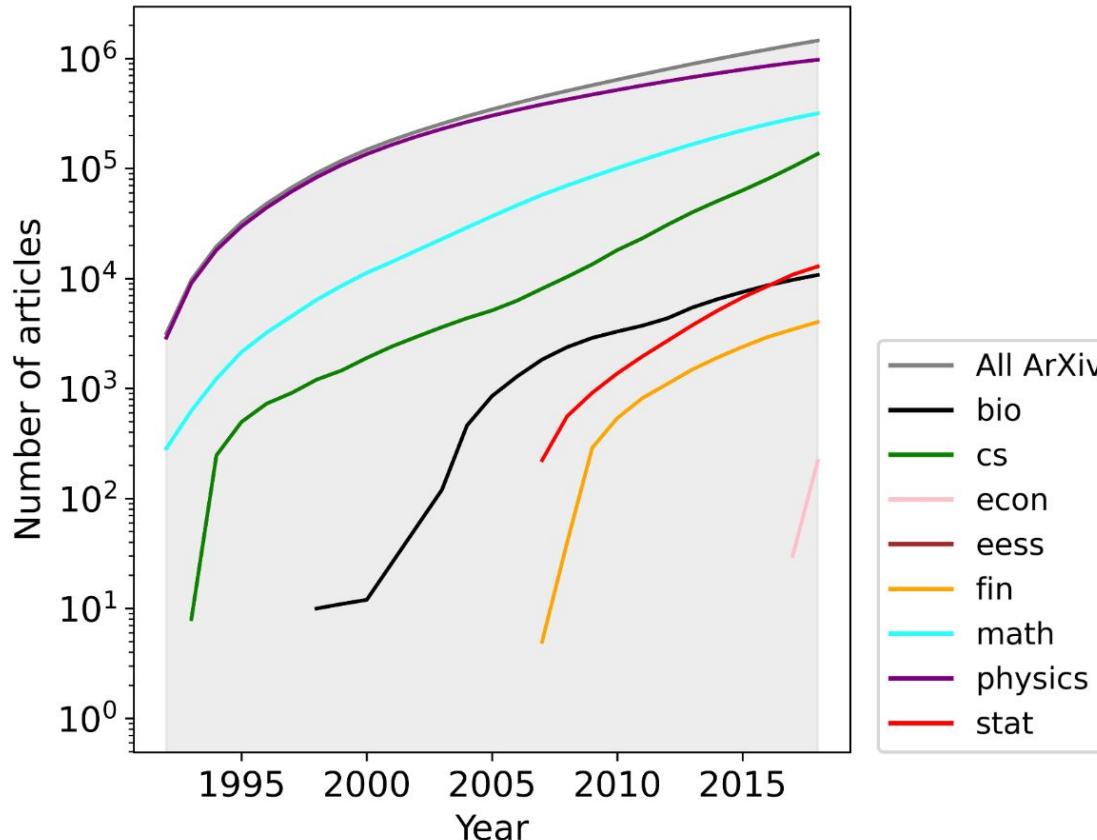
Embedding of large data to
(lower-dimensional)
'knowledge' space

Science of science open data

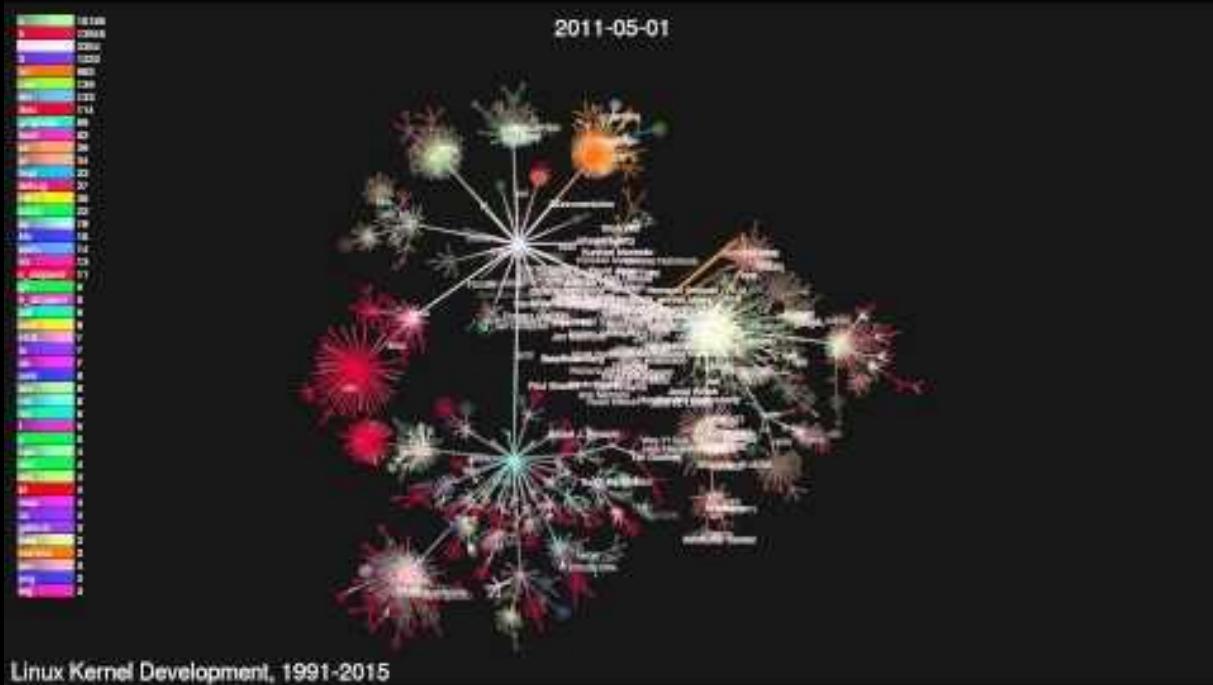


<https://github.com/mattbierbaum/arxiv-public-datasets>

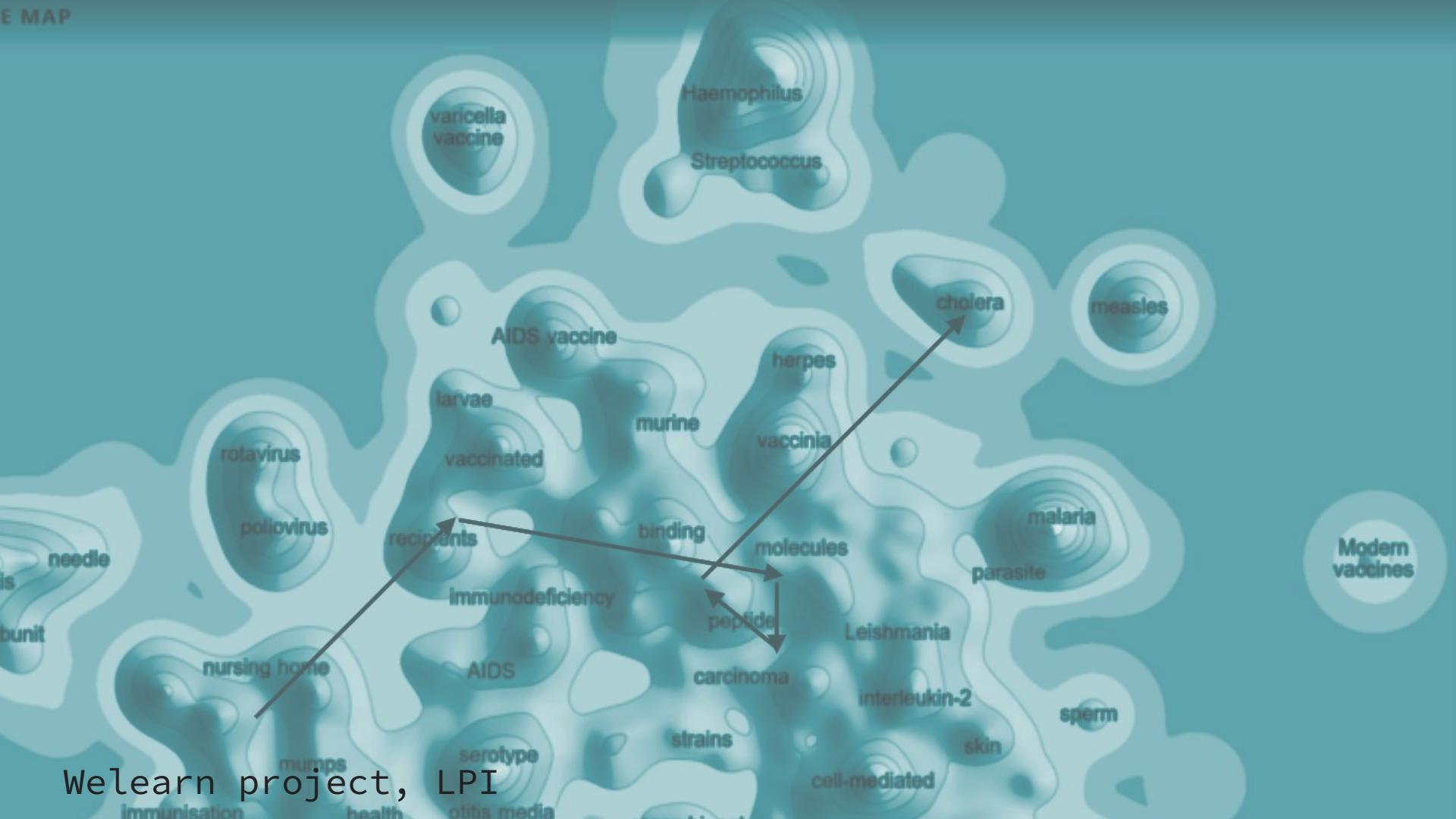
Other science of science open datasets



The scale of open collaborations



Linux
17,000+ contributors
~30 years
3 **billion** users



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, \ell)\$-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...

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 $\$(k,\ell)\$$ -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 viterbi...

```

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))

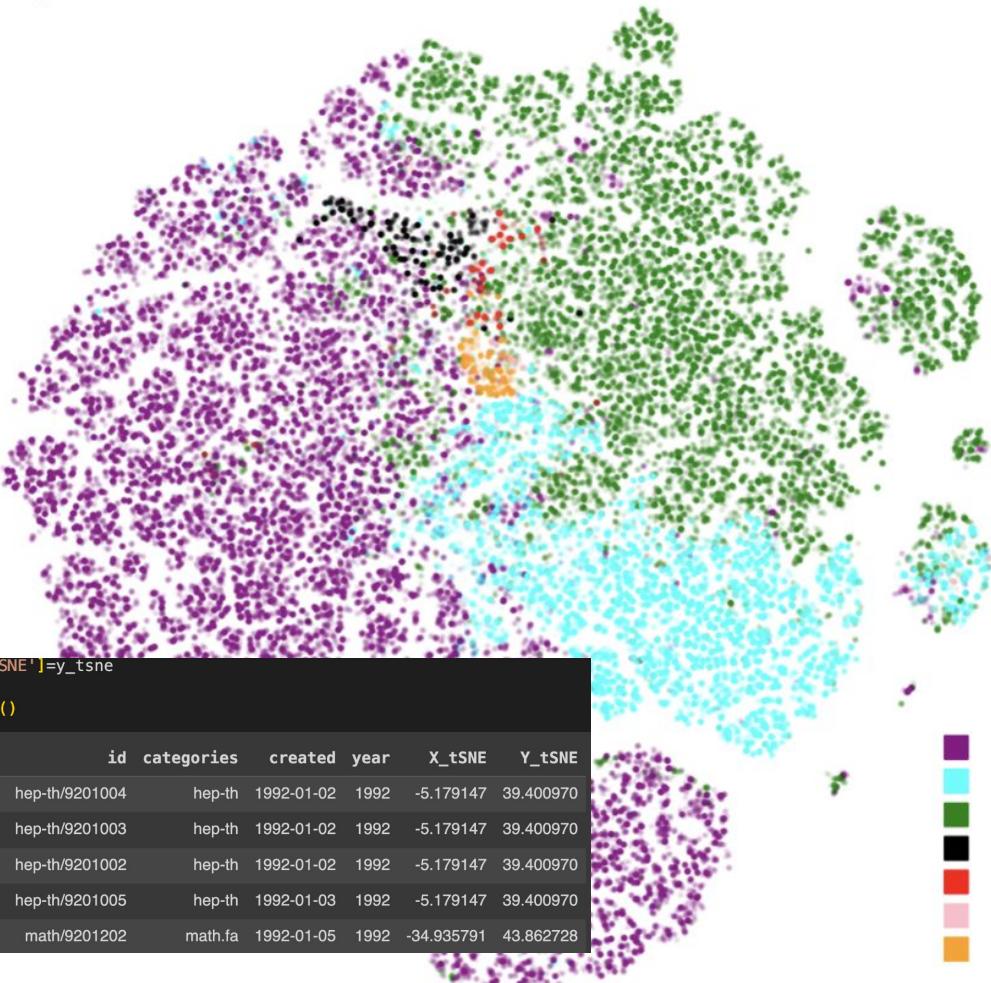
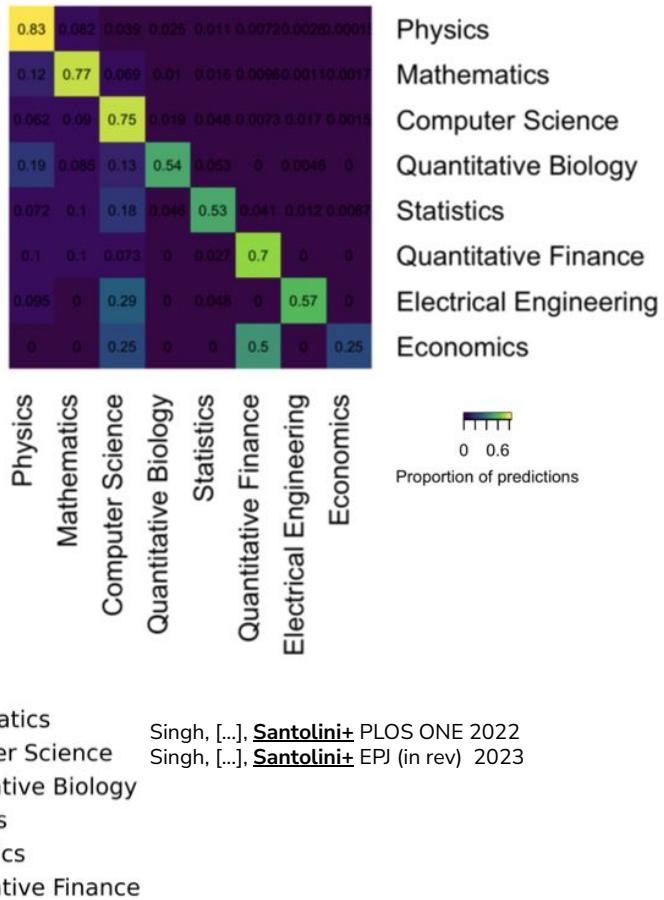
    return zscore

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)
    
```

Modern vaccines

a**b**

Collective level: Rise and Fall of Research Fields



arXiv.org > cs > arXiv:1905.00075

Computer Science > Information Retrieval

[Submitted on 30 Apr 2019]

Time Stamp

On the Use of ArXiv as a Dataset

Colin B. Clement, Matthew Bierbaum, Kevin P. O'Keeffe, Alexander A. Alemi

Authors

The arXiv has collected 1.5 million pre-print articles over 28 years, hosting literature from scientific fields including Physics, Mathematics, and Computer Science. Each pre-print features text, figures, authors, citations, categories, and other metadata. These rich, multi-modal features, combined with the natural graph structure—created by citation, affiliation, and co-authorship—makes the arXiv an exciting candidate for benchmarking next-generation models. Here we take the first necessary steps toward this goal, by providing a pipeline which standardizes and simplifies access to the arXiv's publicly available data. We use this pipeline to extract and analyze a 6.7 million edge citation graph, with an 11 billion word corpus of full-text research articles. We present some baseline classification results, and motivate application of more exciting generative graph models.

Subjects: Information Retrieval (cs.IR), Machine Learning (cs.LG), Social and Information Networks (cs.SI), Physics and Society (physics.soc-ph)

(or arXiv:1905.00075v1 [cs.IR] for this version)

Tags - research fields

Bibliographic data

[Enable Bibex (What is Bibex)?]

Submission history

From: Colin B Clement [view email]

[v1] Tue, 30 Apr 2019 19:43:53 UTC (217 KB)

We gratefully acknowledge support from
the Simons Foundation and member institutions.

Search All fields Search

Help | Advanced Search

Download:

- PDF
- PostScript
- Other formats

(some)

Current browse context:

cs.IR

< prev | next >

new | recent | 1805

Change to browse by:

cs

cs.LG

cs.SI

physics

physics.soc-ph

References & Citations

- NASAADS
- Google Scholar
- Semantic Scholar

DBLP + CS Bibliography

listing | bibtex

Colin B. Clement

Matthew Bierbaum

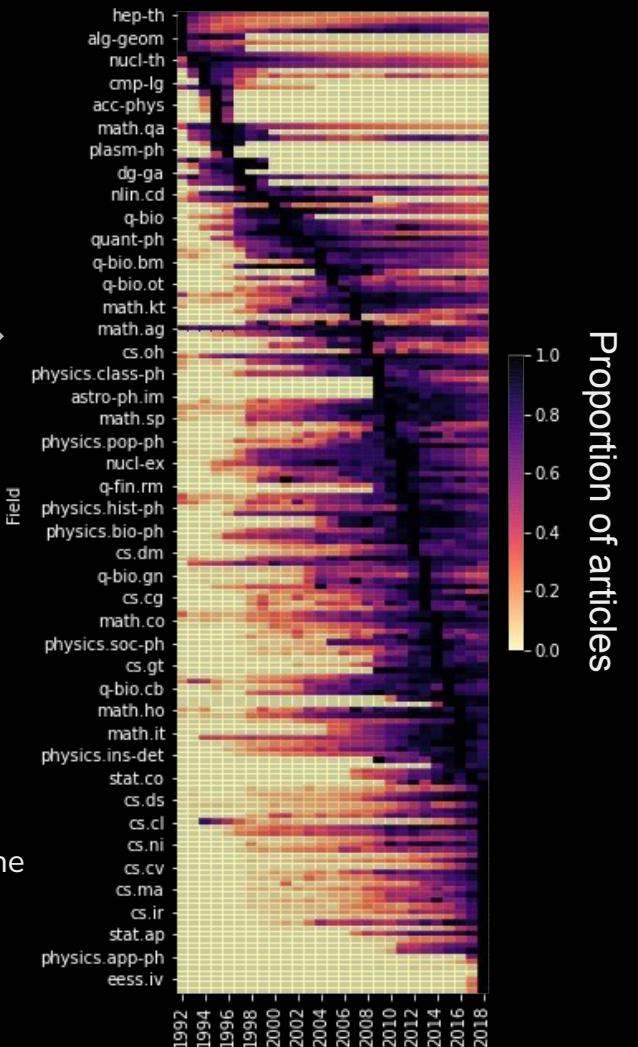
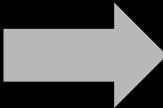
Kevin P. O'Keeffe

Alexander A. Alemi

Export citation

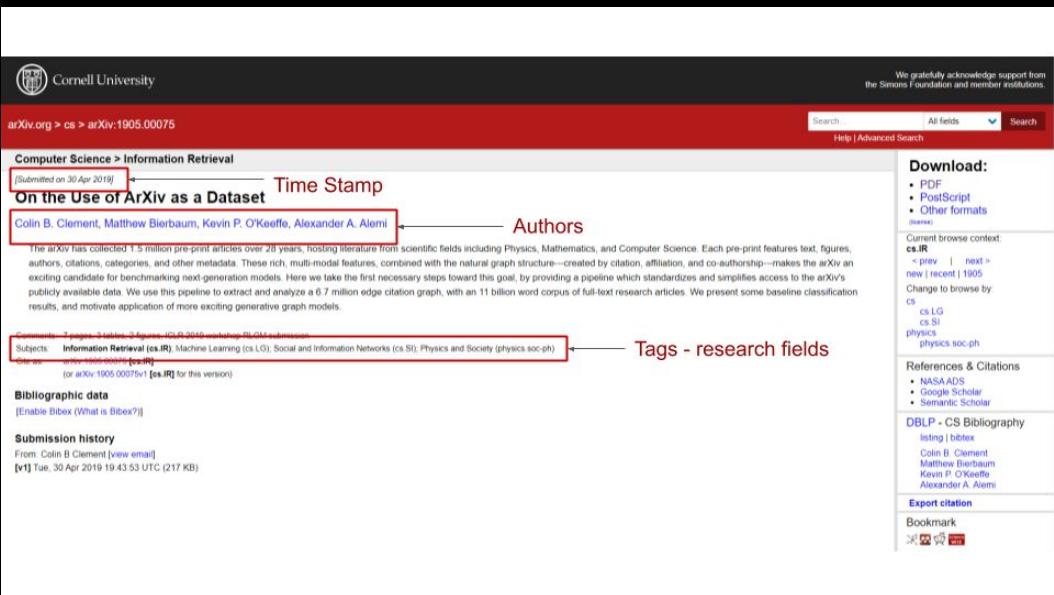
Bookmark

View

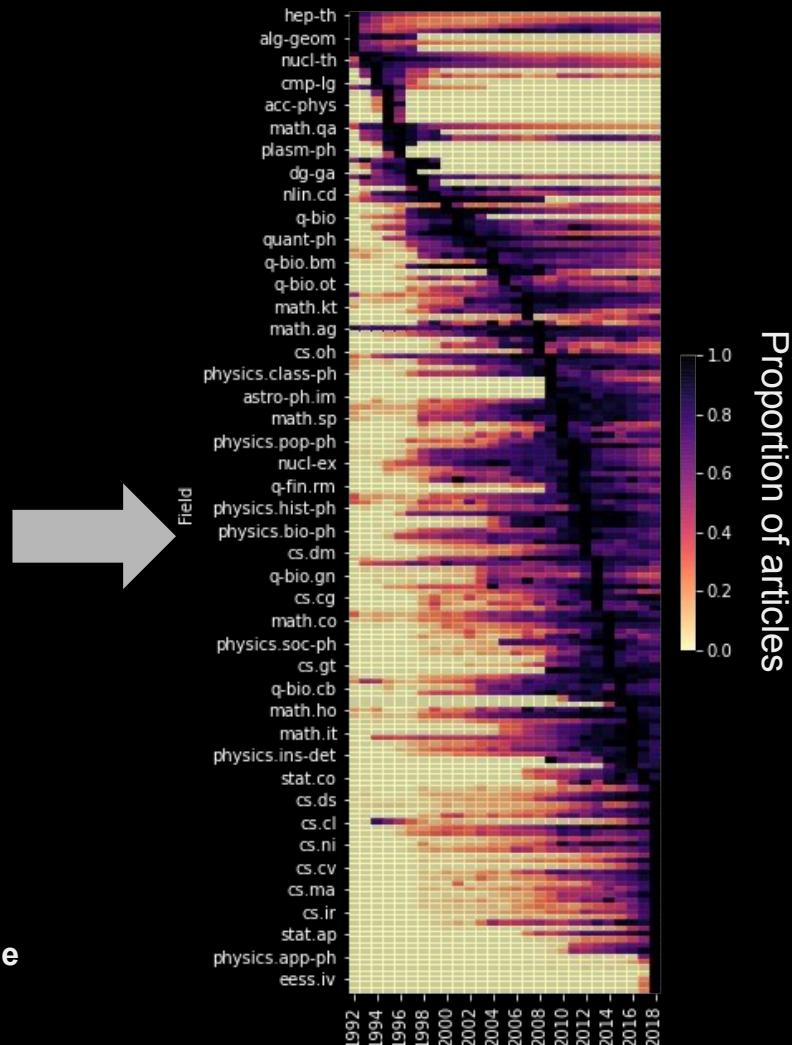


Clement et al. **On the Use of ArXiv as a Dataset**. arXiv:1905.00075 (2019),
Singh (...) Tupikina, Santolini **Quantifying the rise and fall of scientific fields**", Plos One
(2022)

Collected using the arXiv API

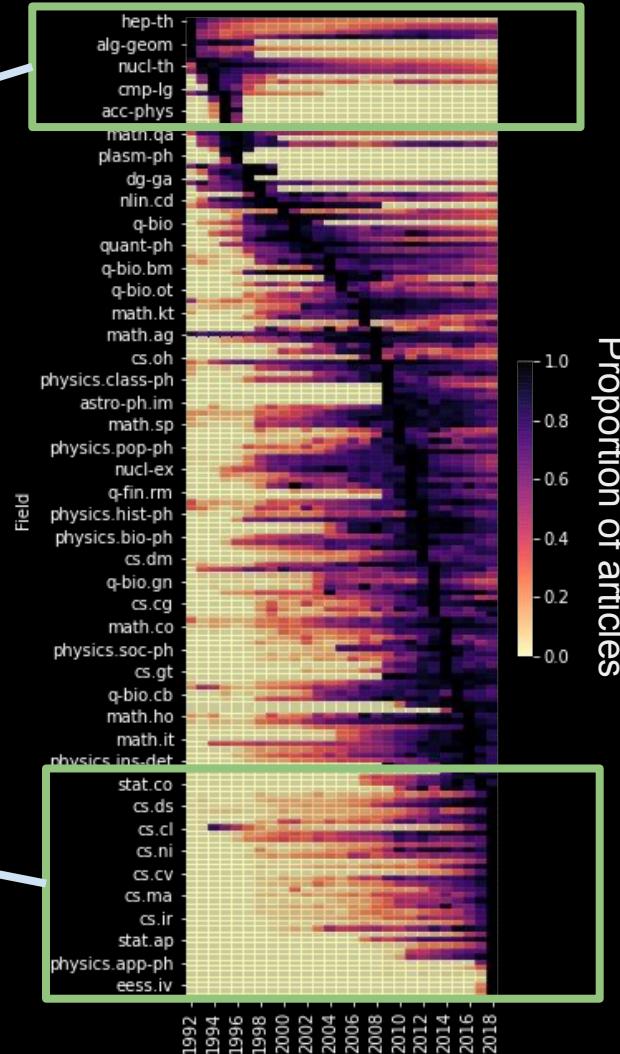


Clement, C. B., Bierbaum, M., O'Keeffe, K. P., & Alemi, A. A. (2019). **On the Use of ArXiv as a Dataset**. *arXiv preprint arXiv:1905.00075*.



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

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



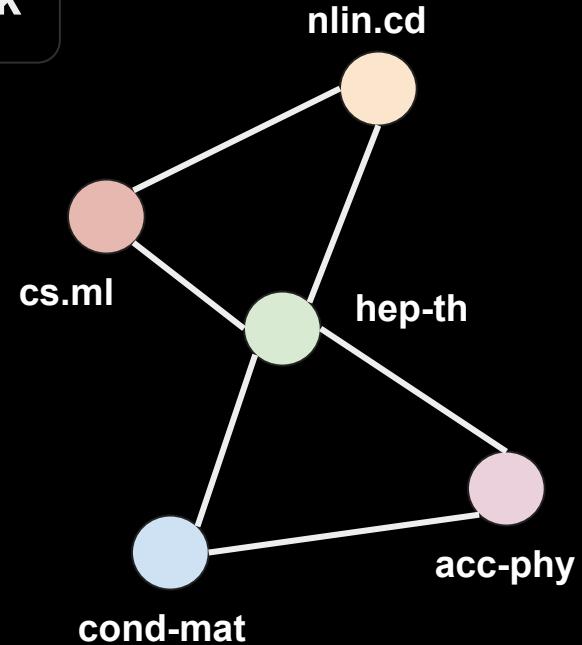
Building the Co-Tag Network

Article i

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

Article j

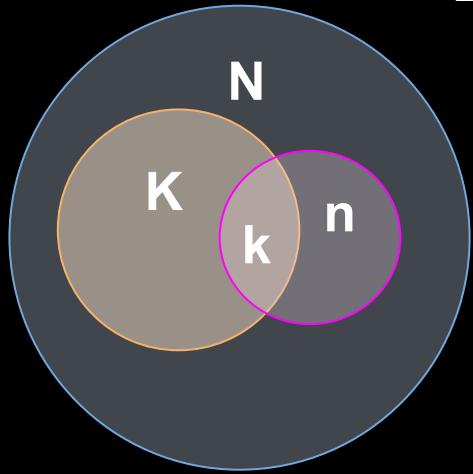
{ hep-th; cond-mat; acc-phy }



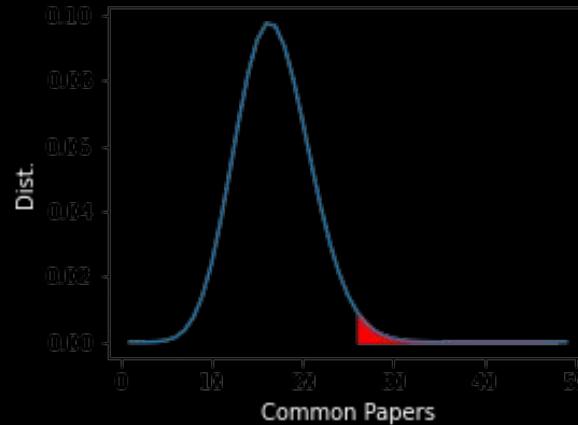
Edge Weight is defined as a function of
common papers b/w two field tags

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

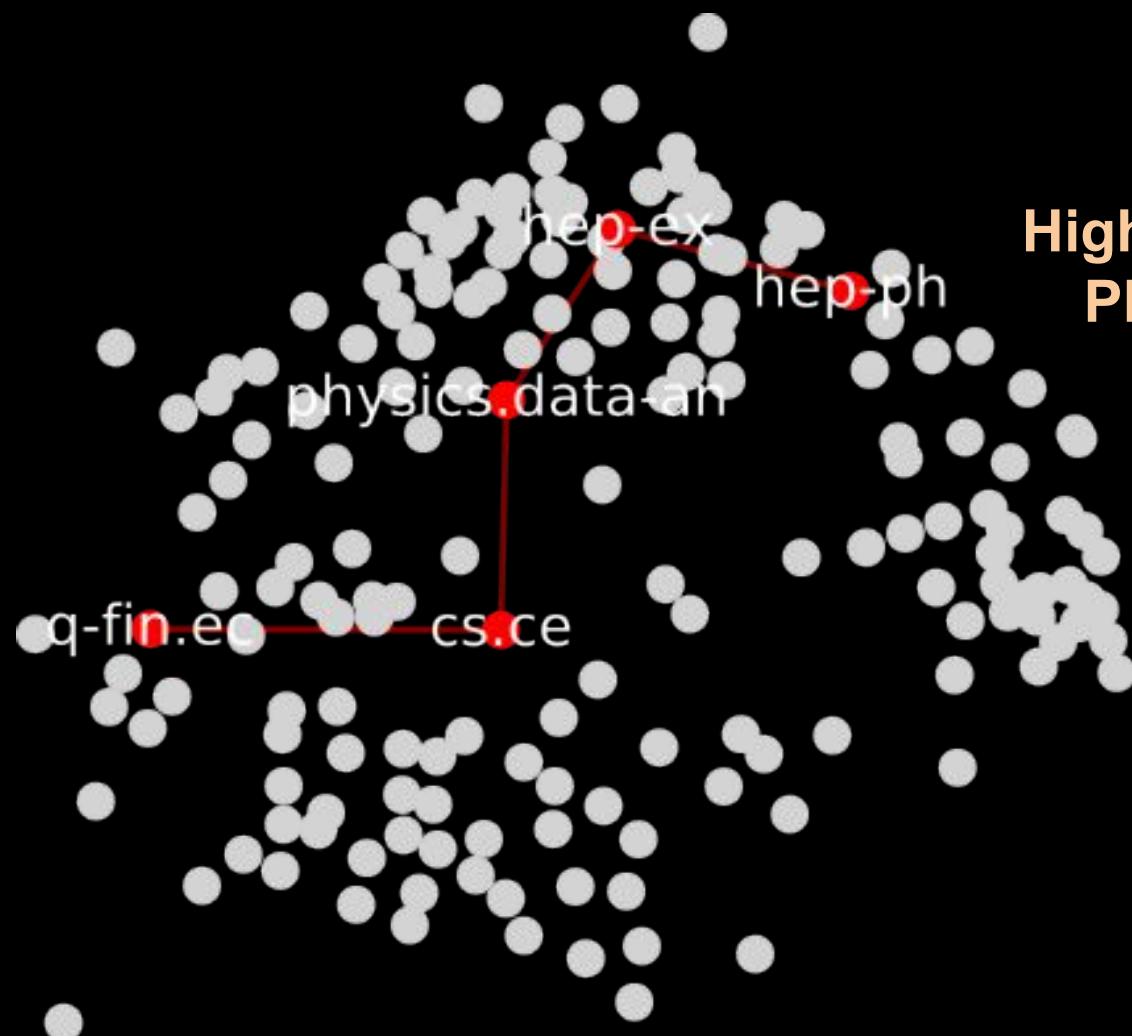
Cognitive Distance

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

e : edge on shortest path

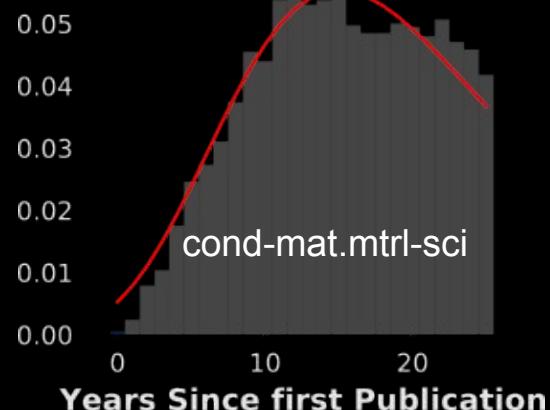
High Energy Physics

Economics

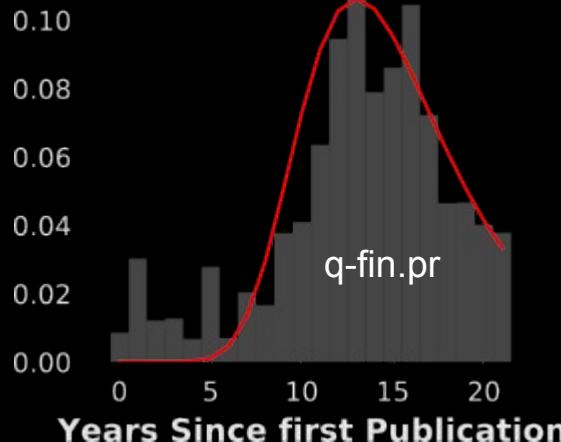
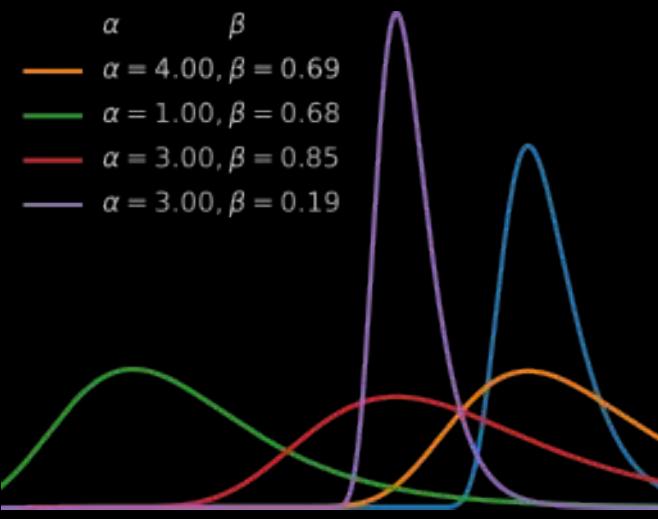


Gumbel distribution function

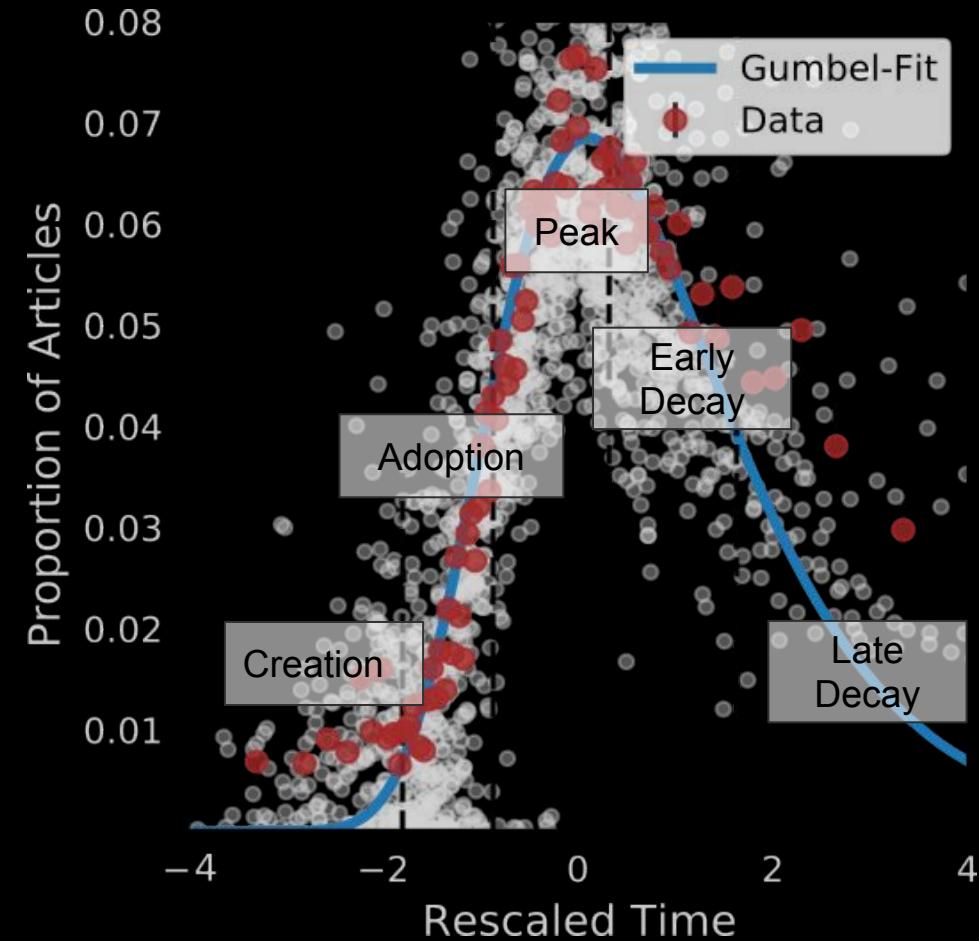
$$G = \frac{1}{\beta} e^{\frac{-(x-\alpha)}{\beta}} e^{-e^{\frac{-(x-\alpha)}{\beta}}}$$



Material Sciences



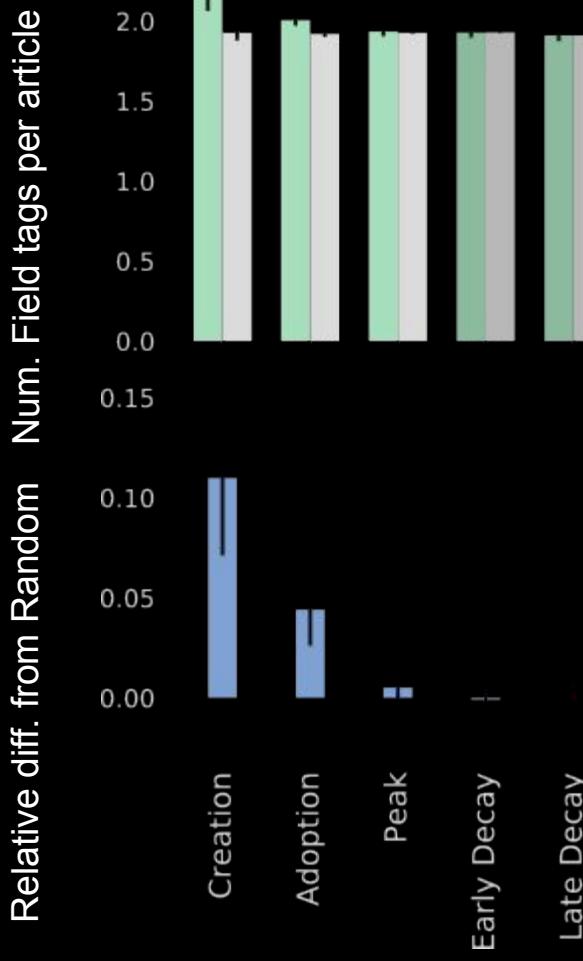
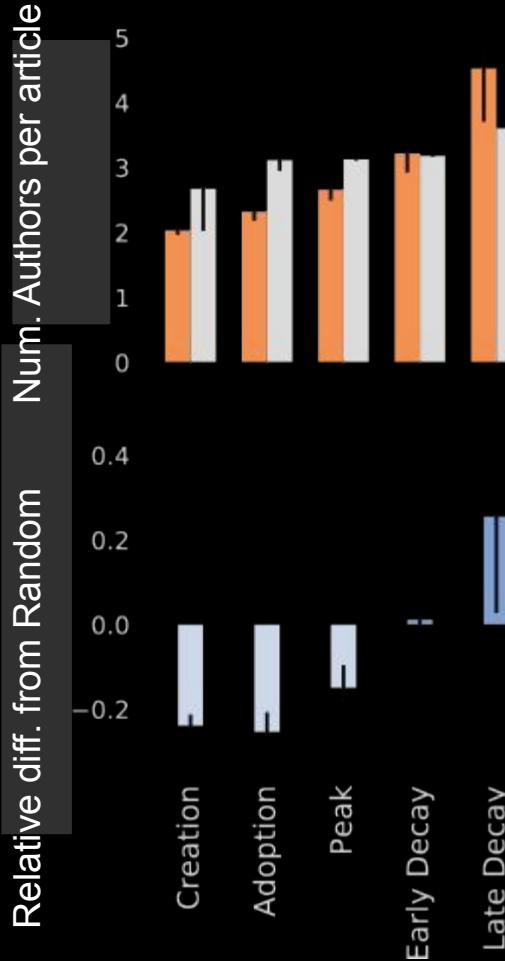
Quantitative Finance (Pricing)

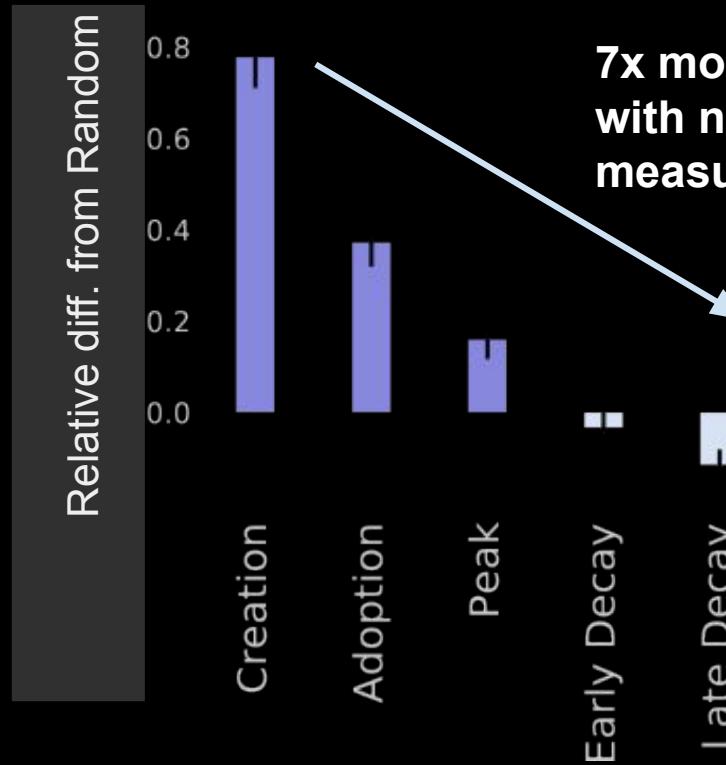
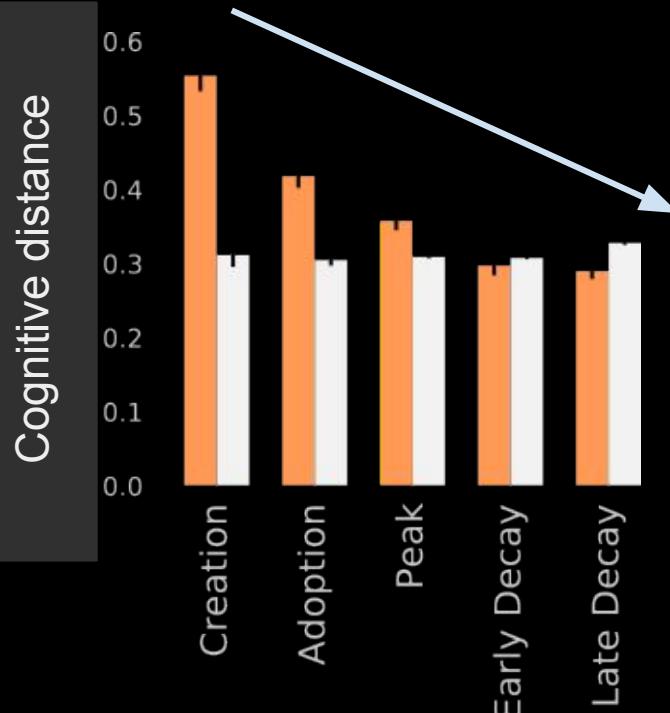


Rescaled time

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

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

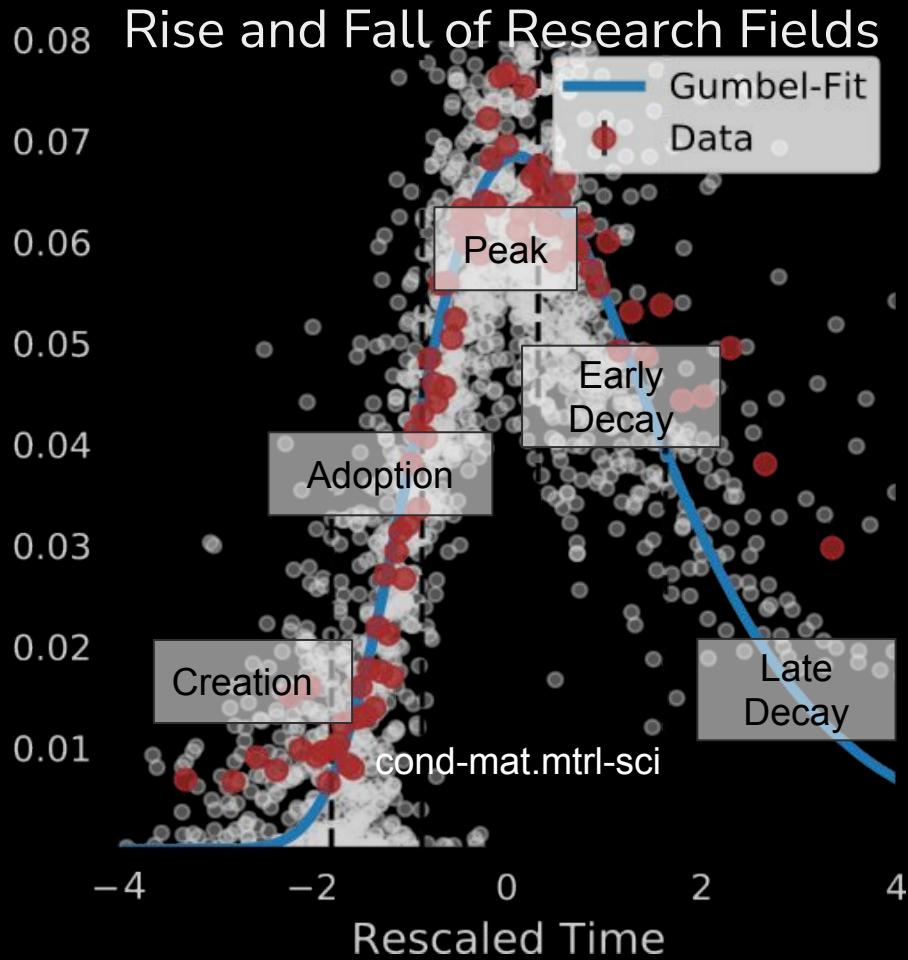




Researchers in creation phase connect distant fields together (**explore**) whereas in later phases they focus on closely related fields (**exploit**)

Rise and Fall of Research Fields

Proportion of Articles



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

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

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