# Information Rate and Branching Processes of Scientific Fields

FranČesko, Henry, Jan, MJ

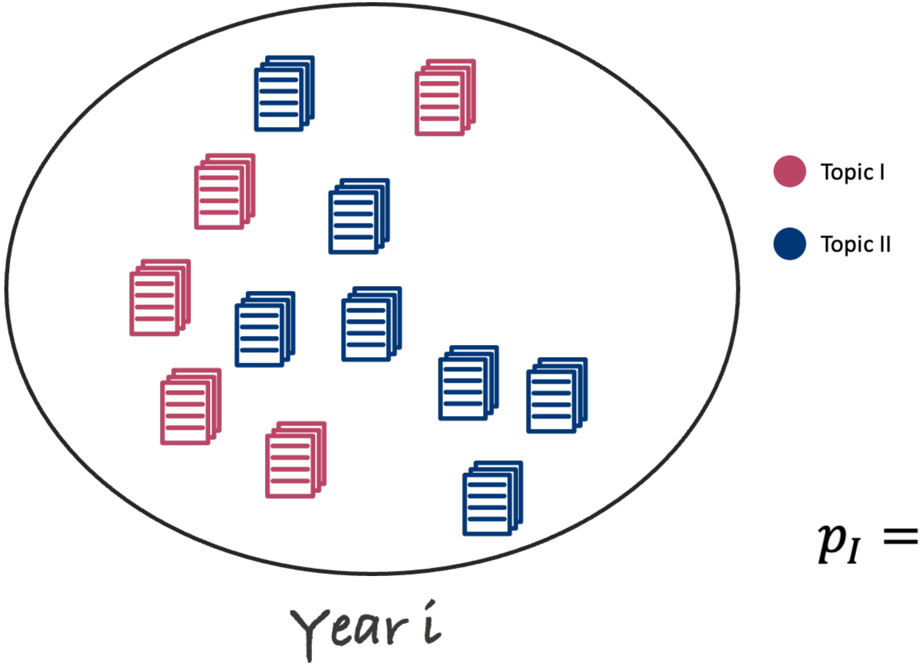
(Advisory Board: Jana)

# *Do Societies have a limited Information processing rate?*

For Sciences:

Is there a correlation between reaching a certain entropy-rate and the emergence of new subfields?

# Field A

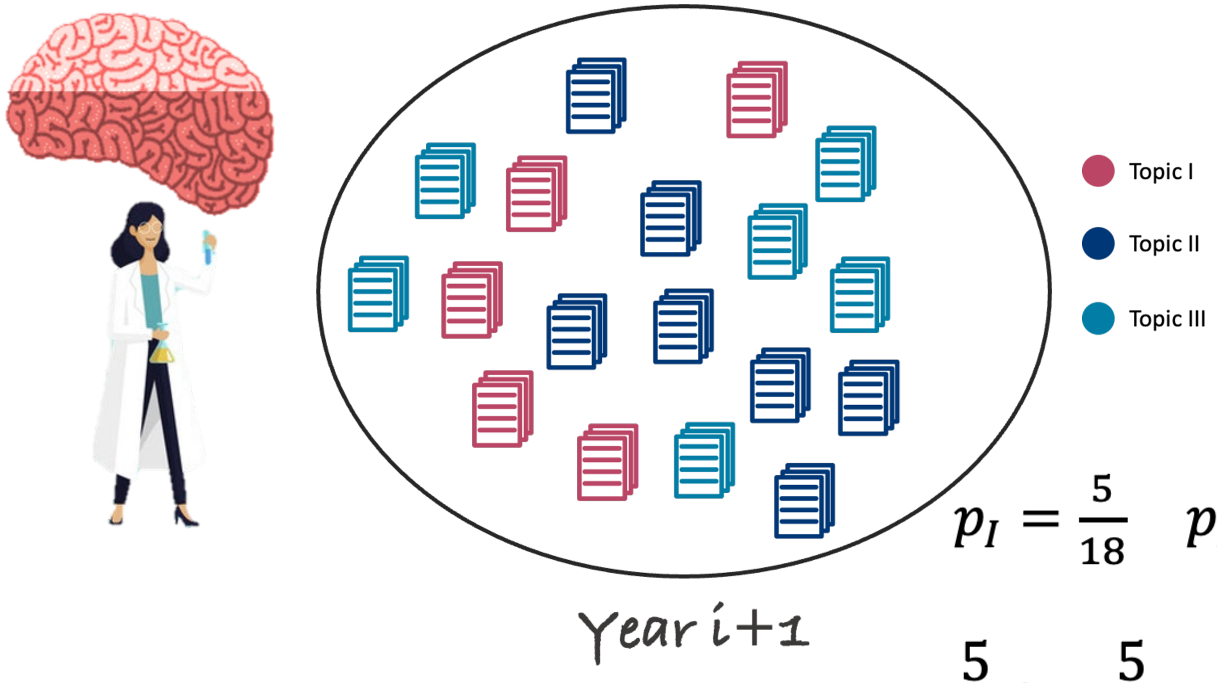


$$p_I = \frac{5}{12} \quad p_{II} = \frac{7}{12}$$

$$H = \sum_i p_i \log p_i = \frac{5}{12}\log\frac{5}{12} + \frac{7}{12}\log\frac{7}{12}$$

$$= 0.68$$

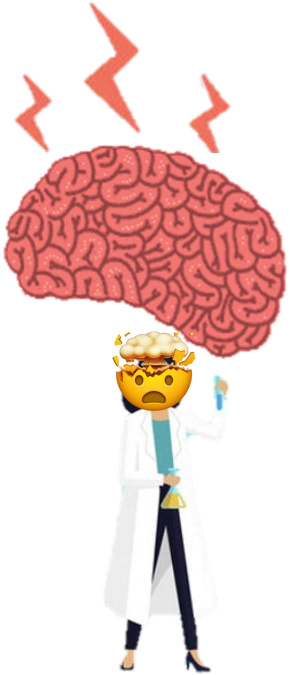
# Field A



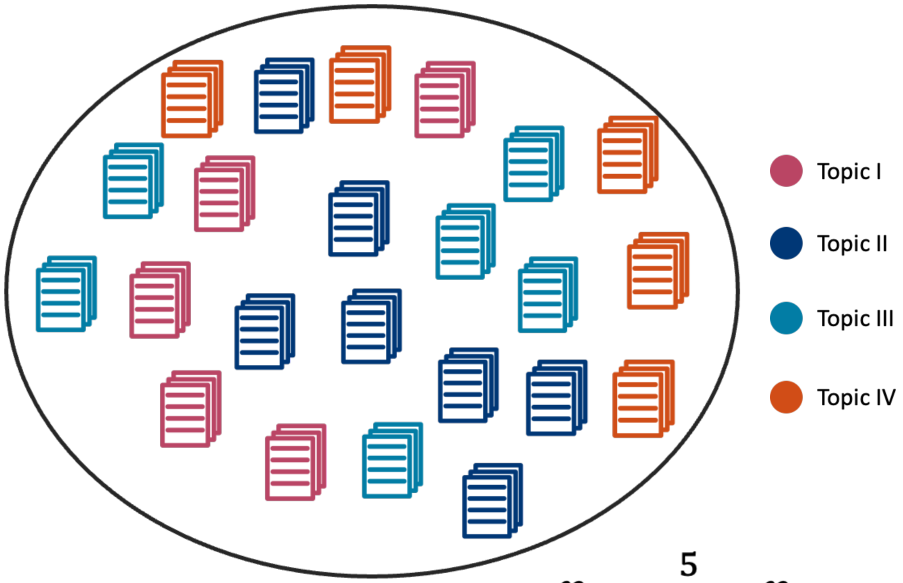
$$p_I = \frac{5}{18} \quad p_{II} = \frac{7}{18} \quad p_{III} = \frac{6}{18}$$

$$H = \frac{5}{18}\log\frac{5}{18} + \frac{7}{18}\log\frac{7}{18} + \frac{6}{18}\log\frac{6}{18} =$$

$$=1$$



# Field A

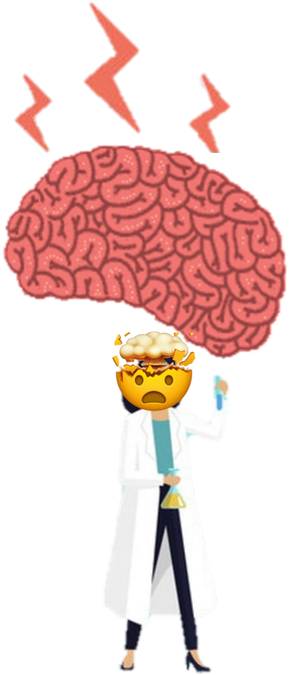


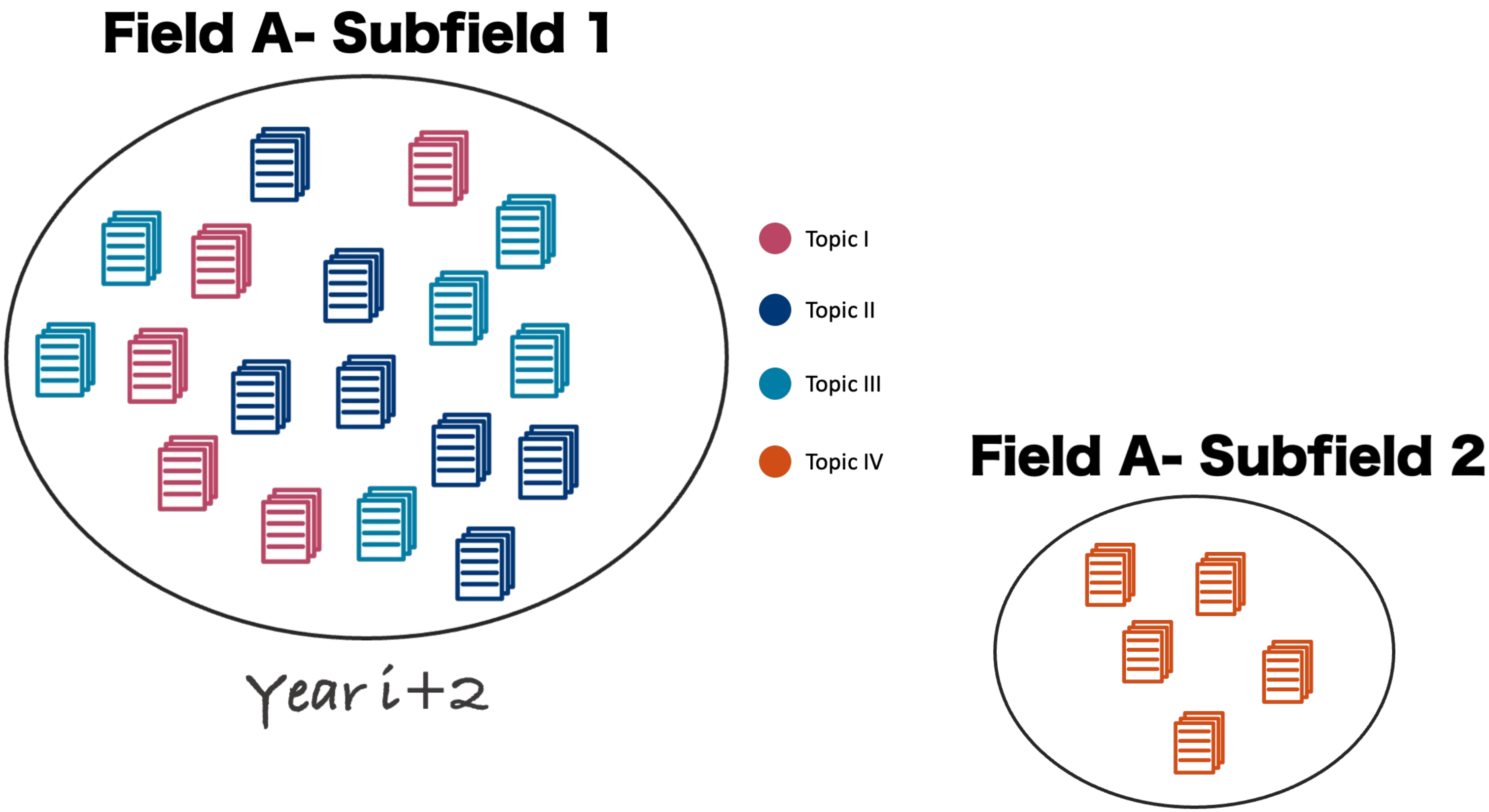
Year i+2

$$p_I = \frac{5}{23} \quad p_{II} = \frac{7}{23} \quad p_{III} = \frac{6}{23} \quad p_{IV} = \frac{5}{23}$$

$$H = \frac{5}{23}\log\frac{5}{23} + \frac{7}{23}\log\frac{7}{23} + \frac{6}{23}\log\frac{6}{23} + \frac{5}{23}\log\frac{5}{23} =$$

$$=1.3$$





## Field A- Subfield 1

Year i+2

● Topic I

● Topic II

● Topic III

● Topic IV

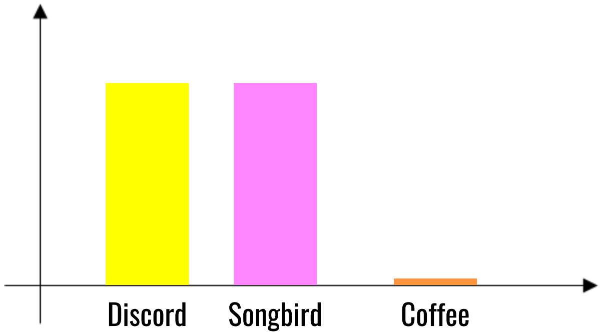
## Field A- Subfield 2

# Method

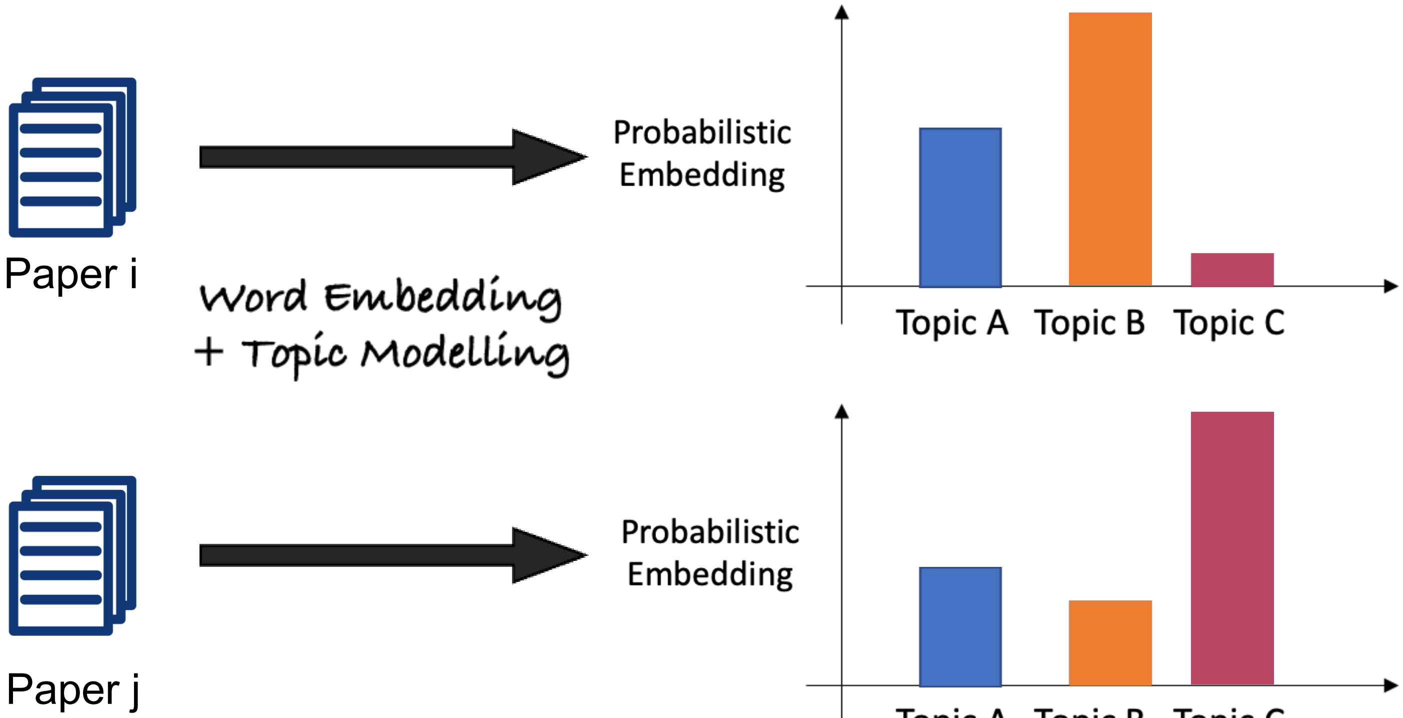
Constantly **checking Discord** has become second nature, like a **songbird** warily scanning the skies for signs of danger before taking flight.

(ChatGPT)

Probabilistic Embedding



# Method



Paper i

Word Embedding
+ Topic Modelling

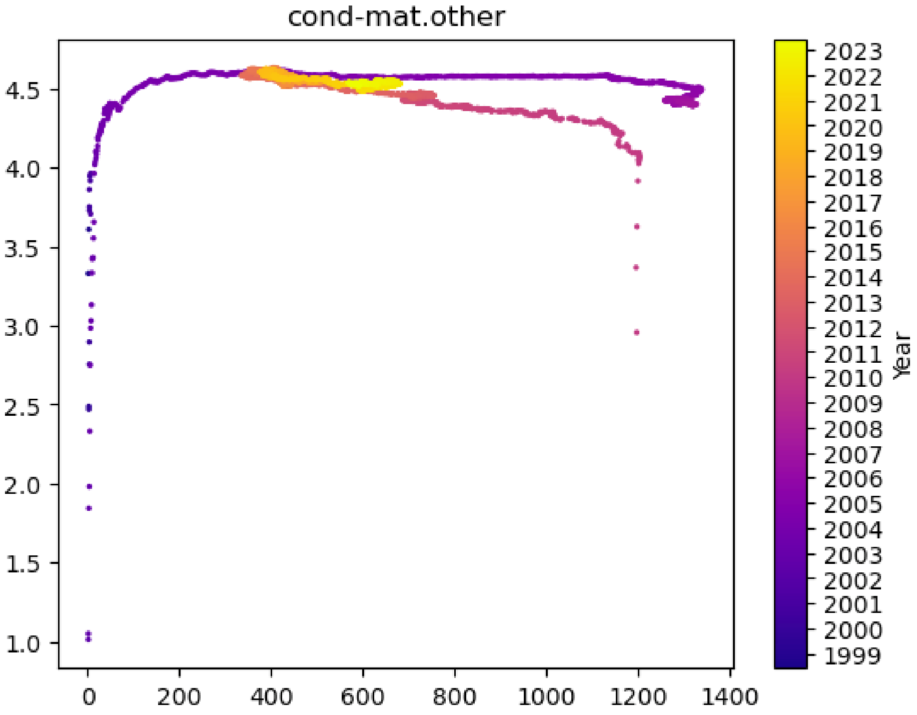
Probabilistic Embedding

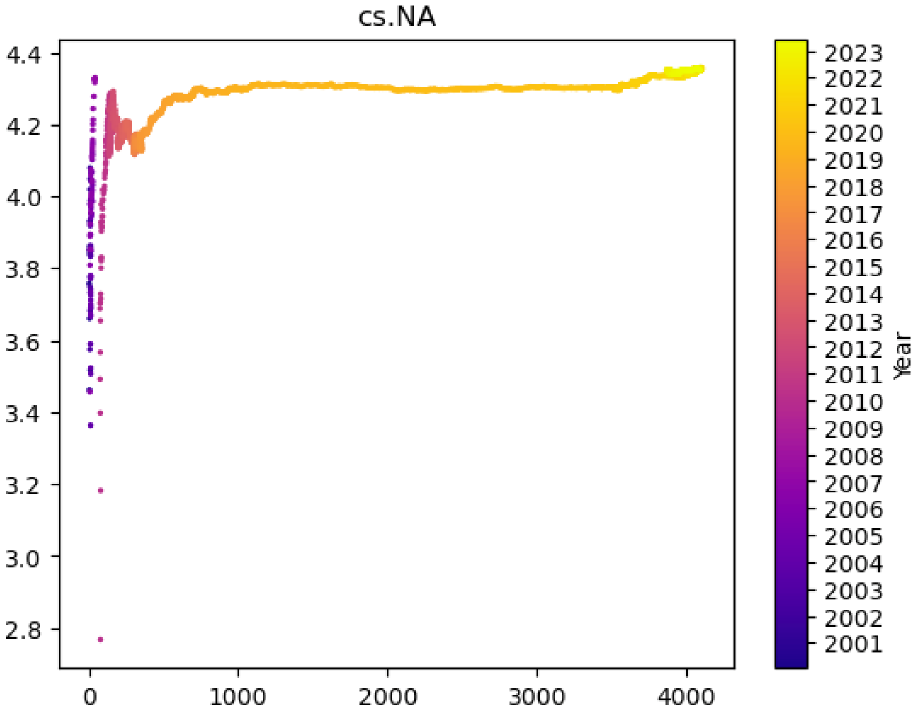
Paper j

Probabilistic Embedding

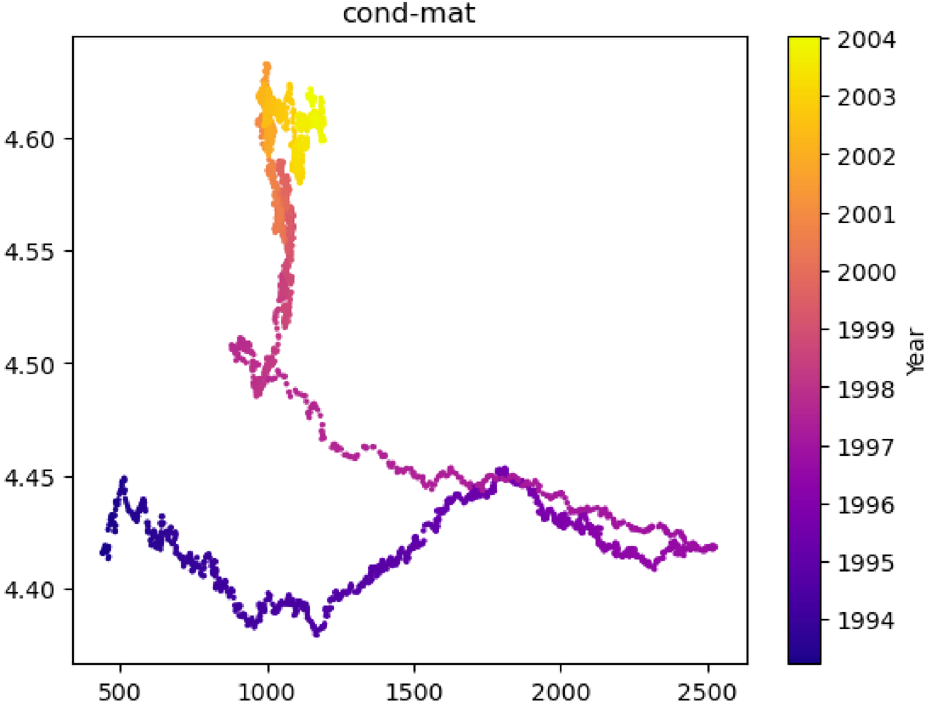
# Results

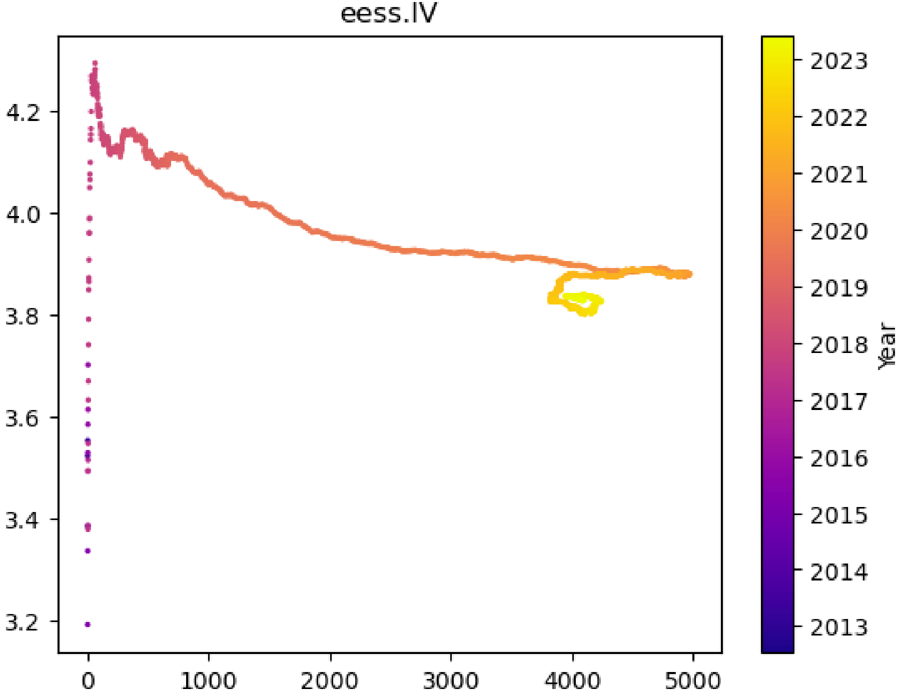
# Entropy rates



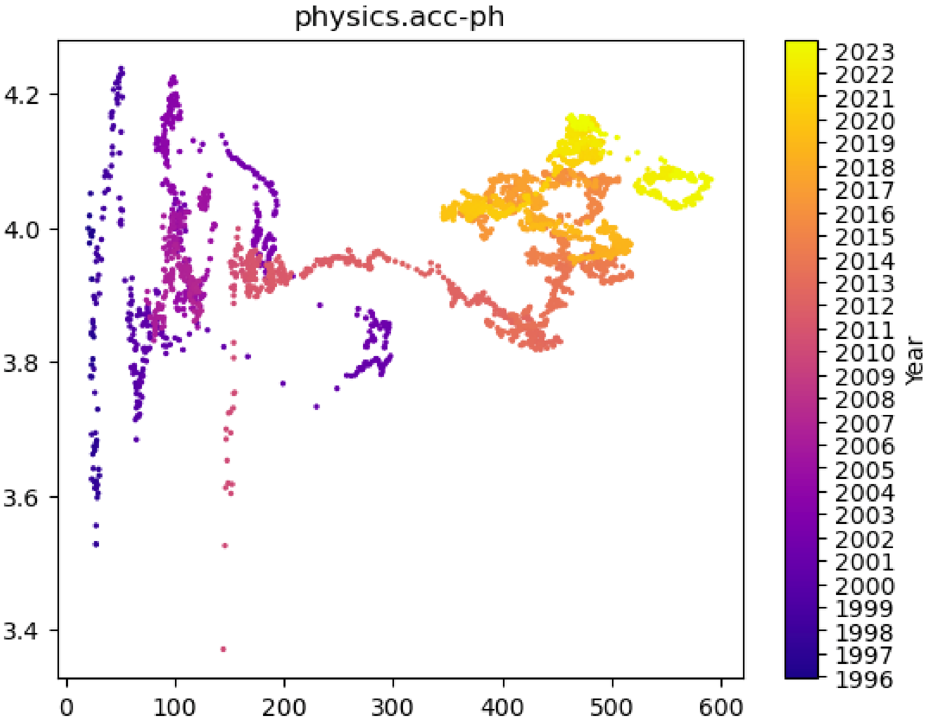


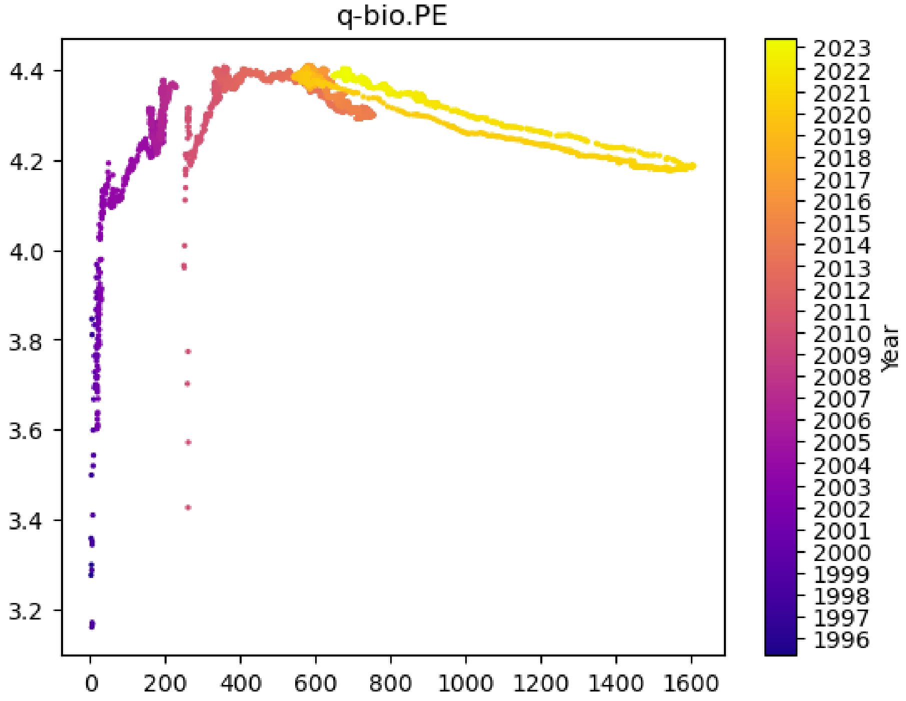
# Entropy rates





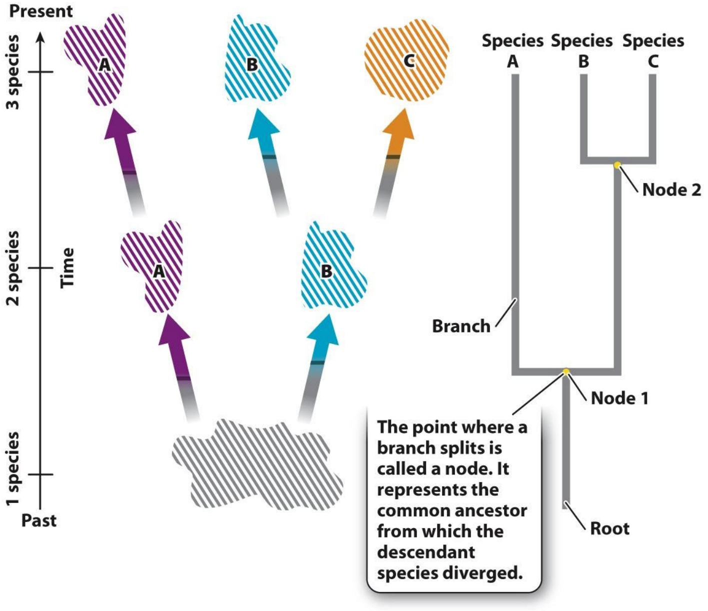
# What does that tell us? 🤷 🤷 🤷


physics.acc-ph


q-bio.PE

# Branching process

## Speciation

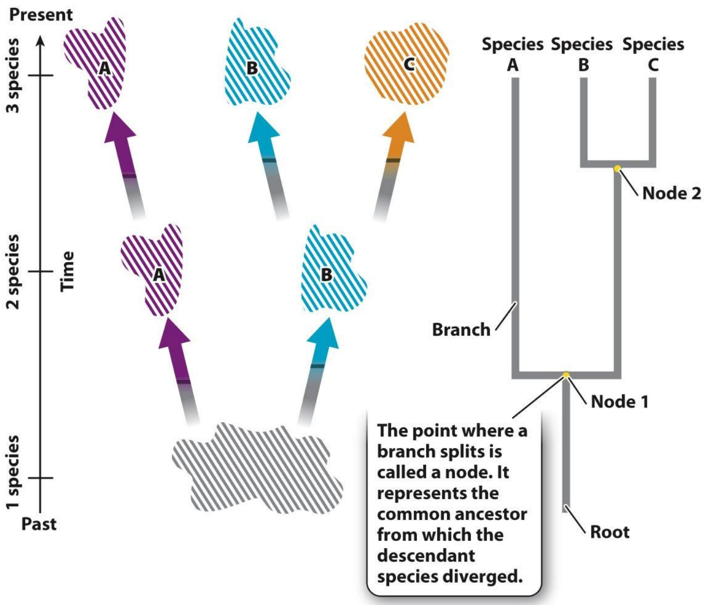


The point where a branch splits is called a node. It represents the common ancestor from which the descendant species diverged.

[Morris, et al., 2023]

# Branching process

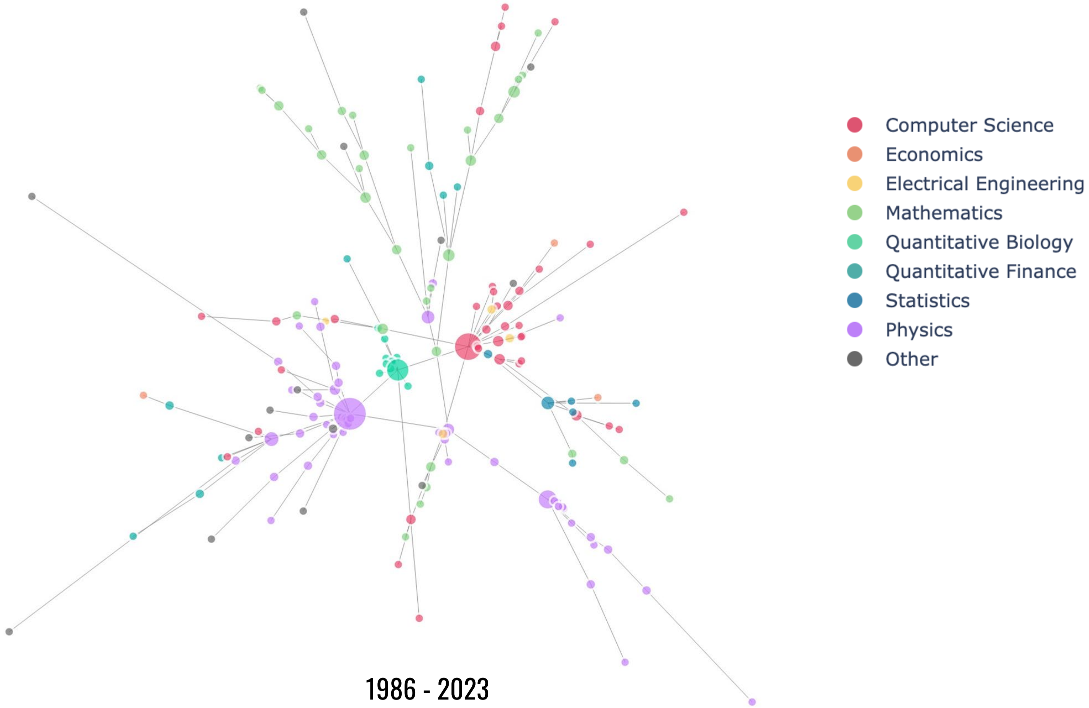
## Speciation



The point where a branch splits is called a node. It represents the common ancestor from which the descendant species diverged.

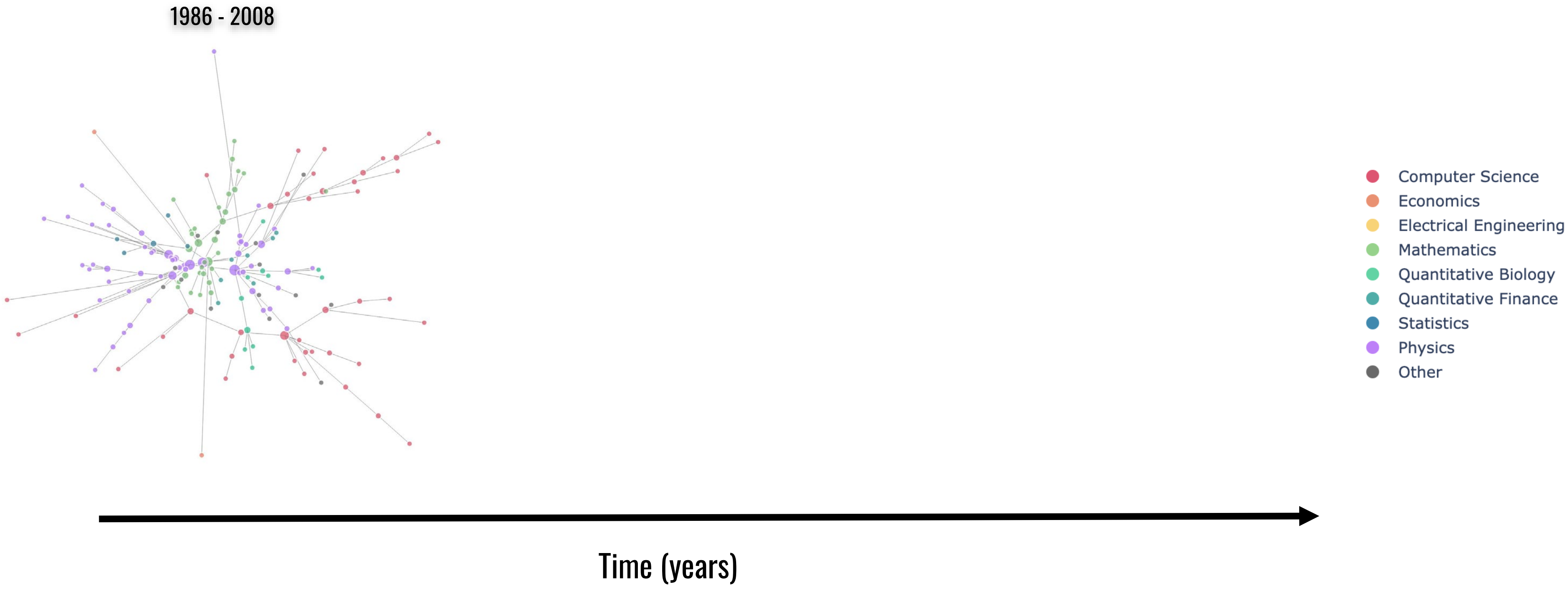
[Morris, et al., 2023]

## Emergence of scientific fields

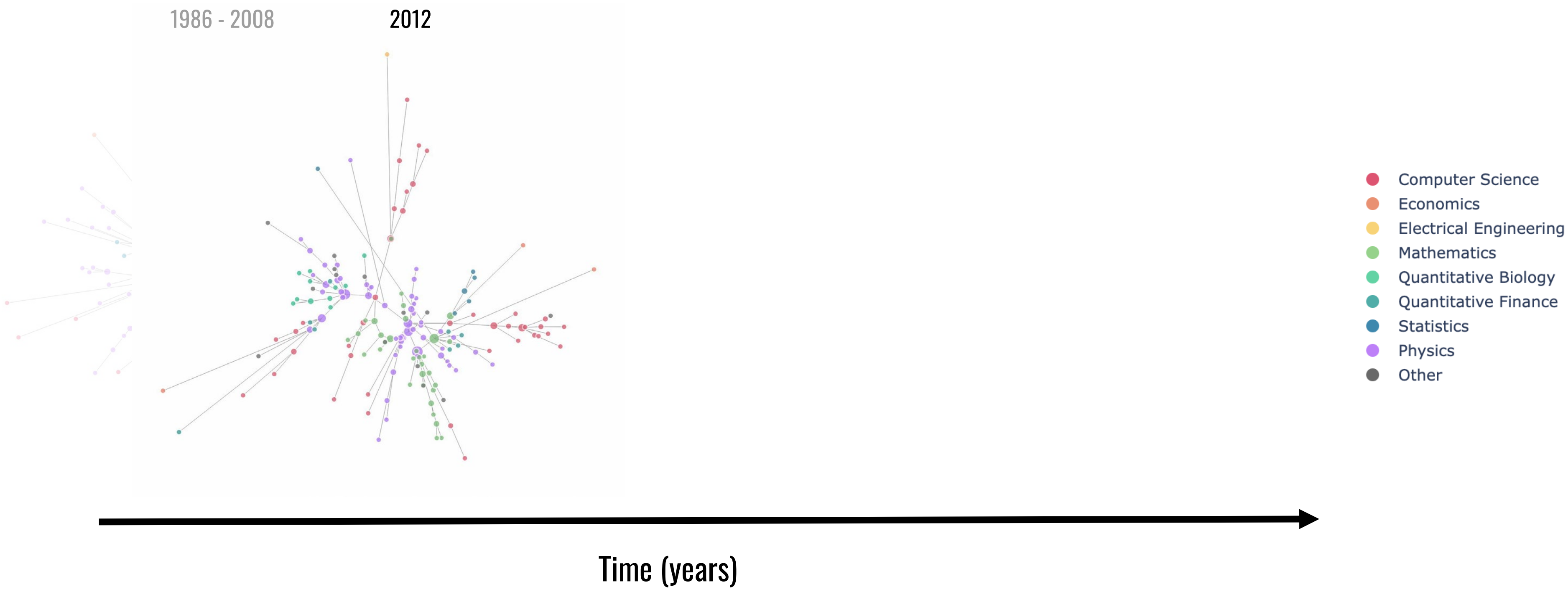


1986 - 2023

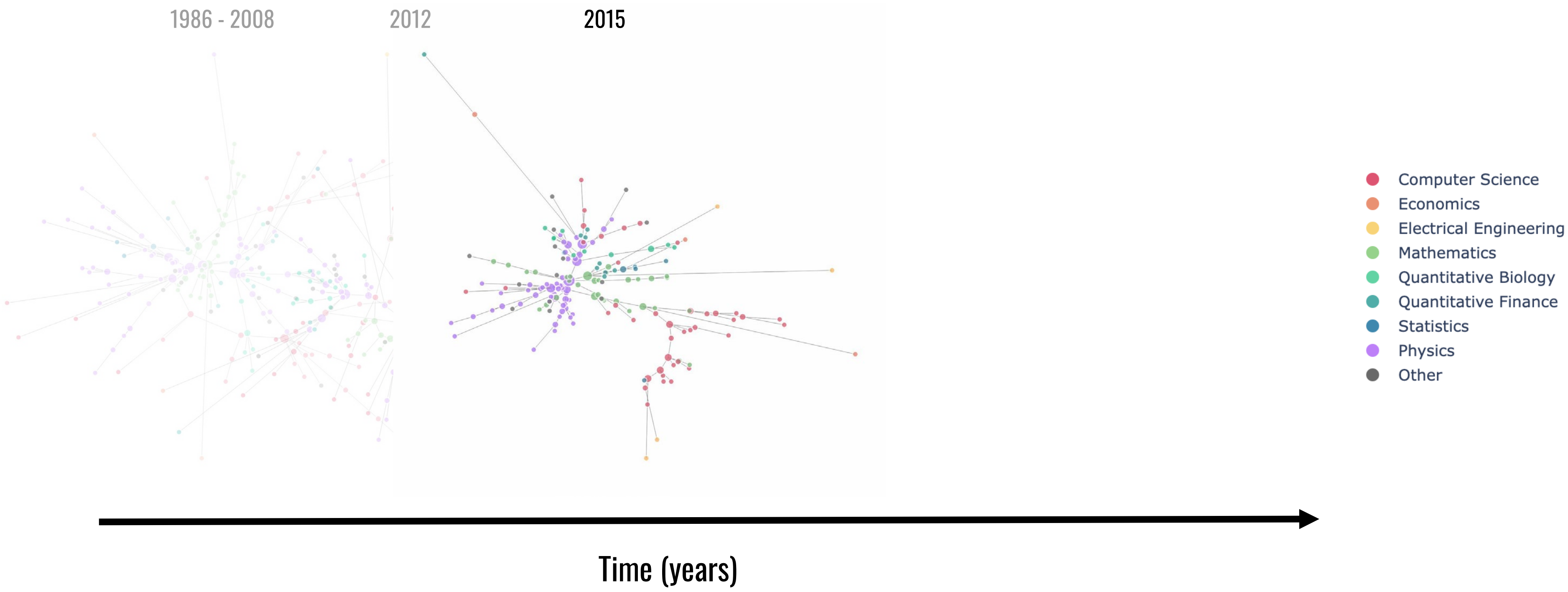
# Branching process



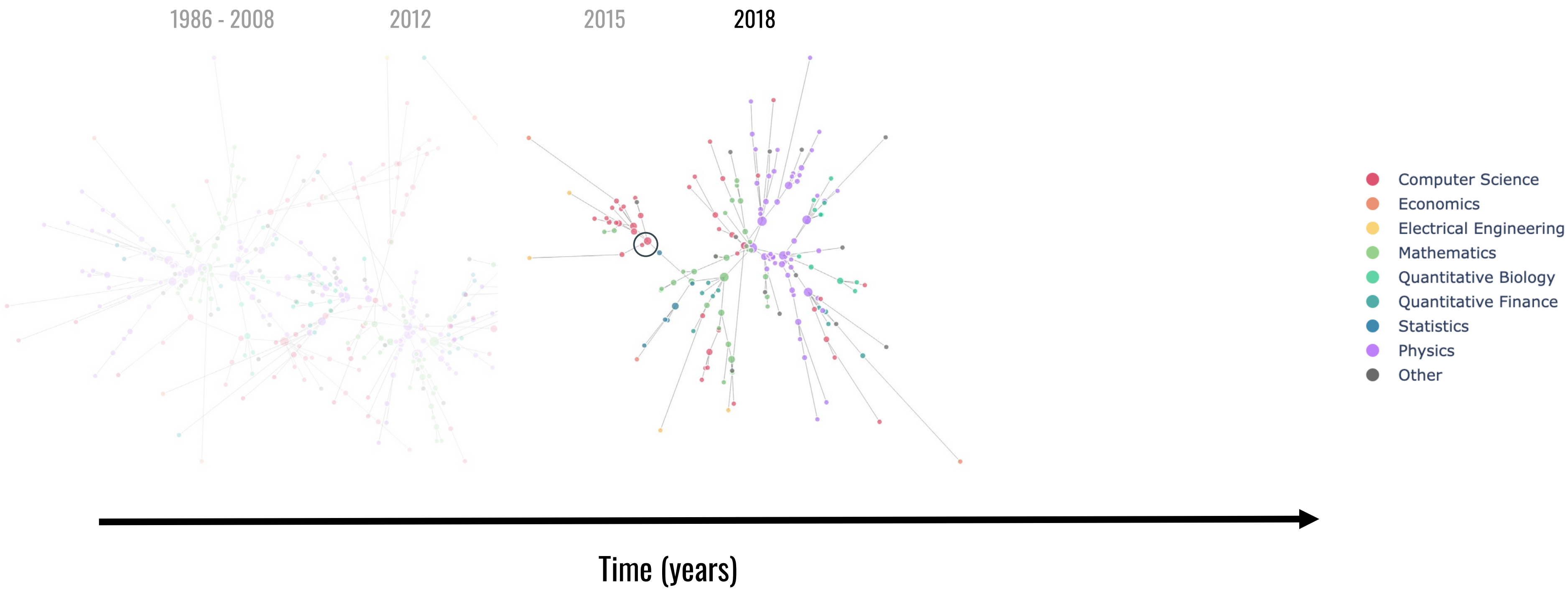
# Branching process



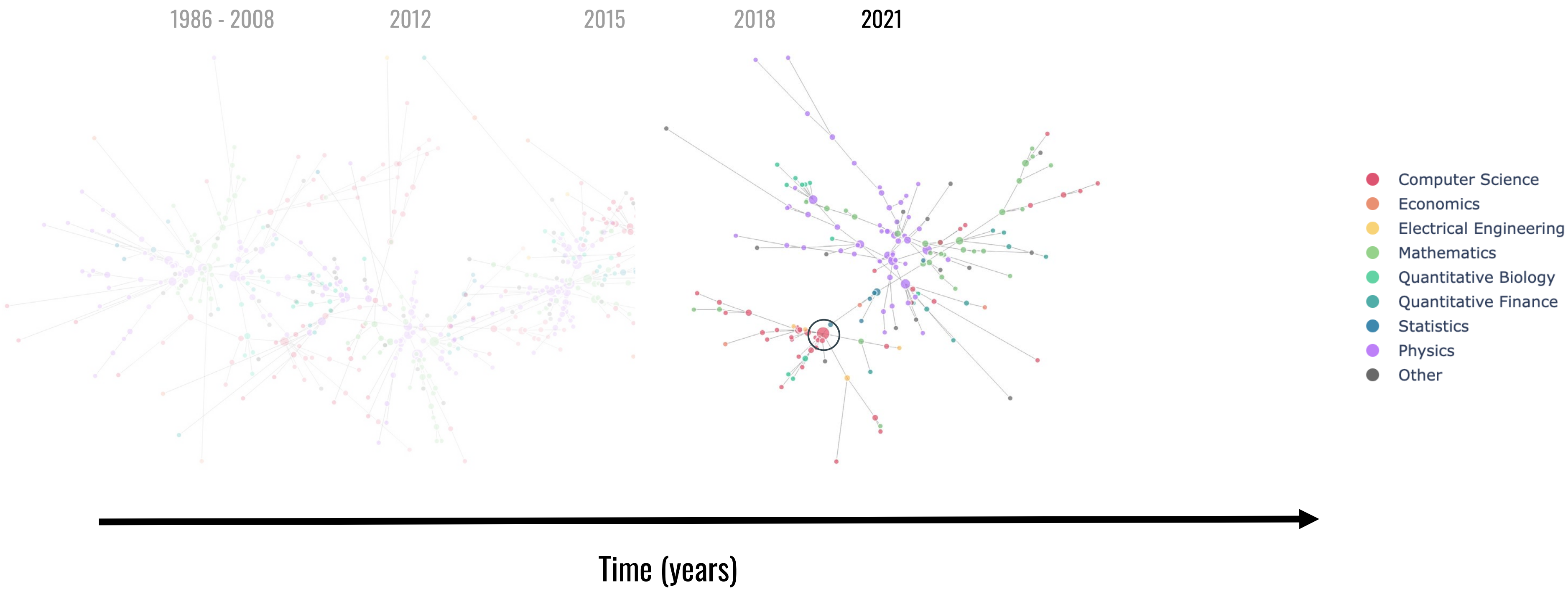
# Branching process



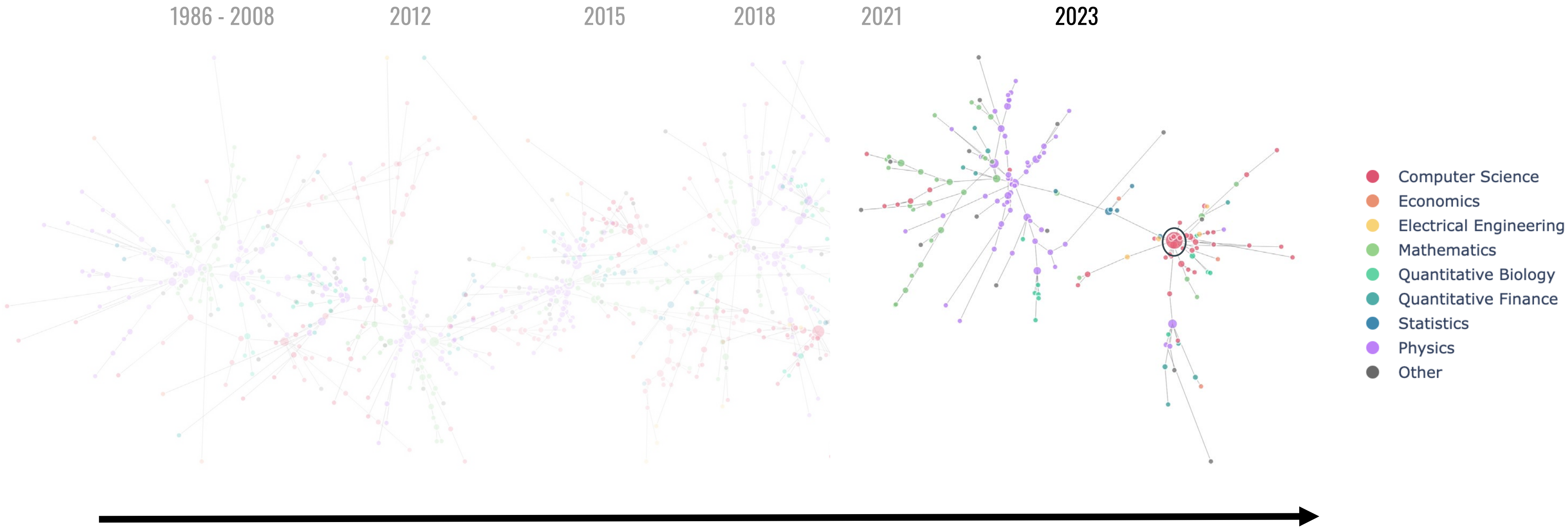
# Branching process



# Branching process

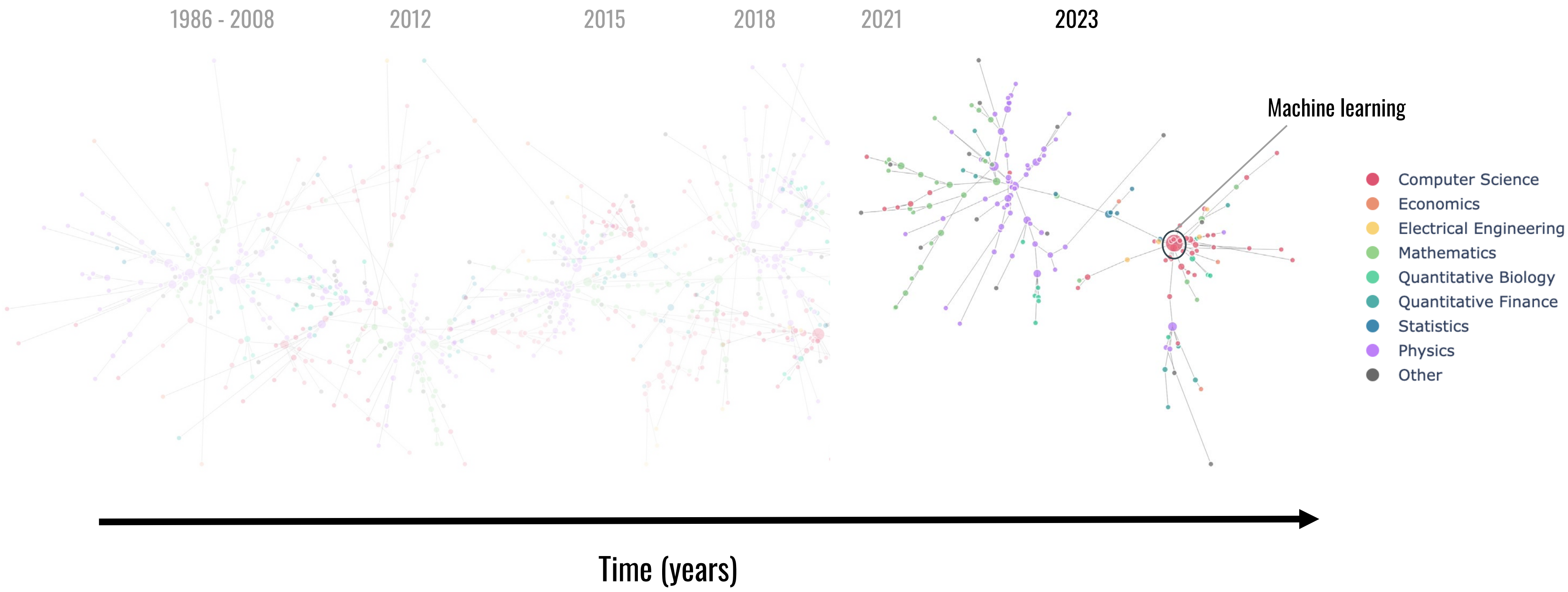


# Branching process

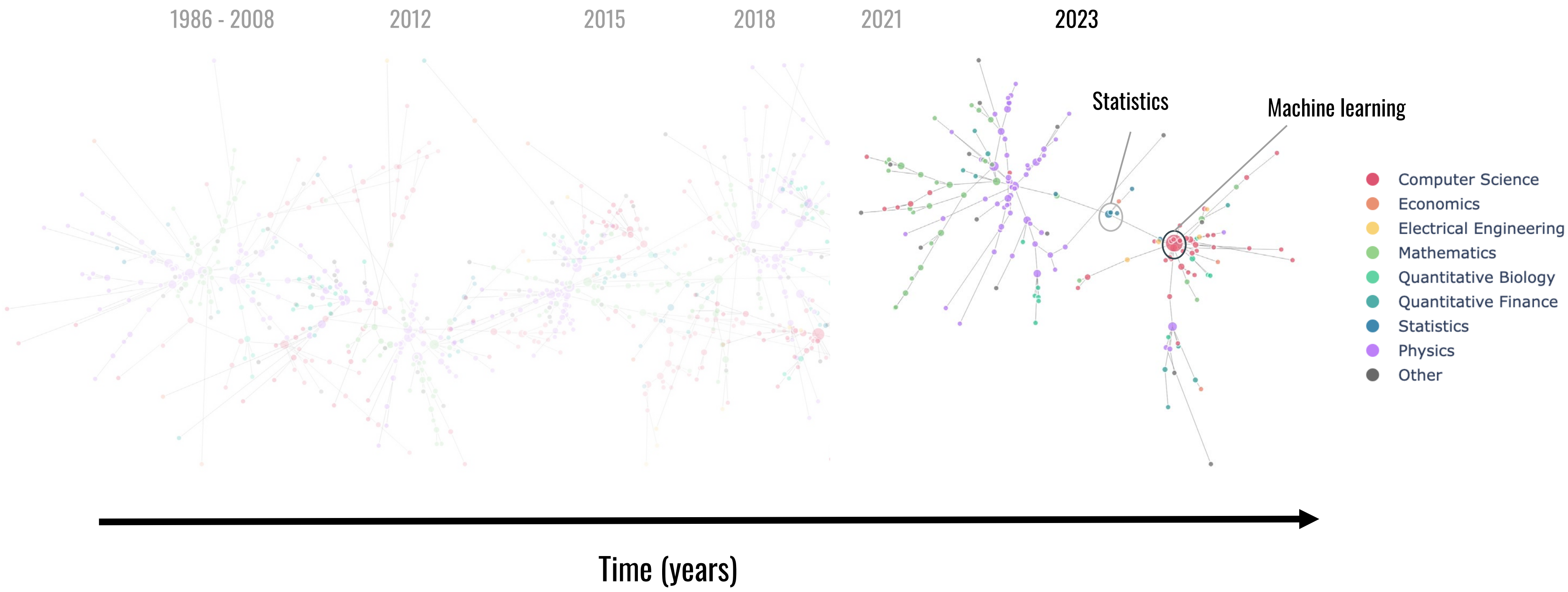


Time (years)

# Branching process



# Branching process



# Future work

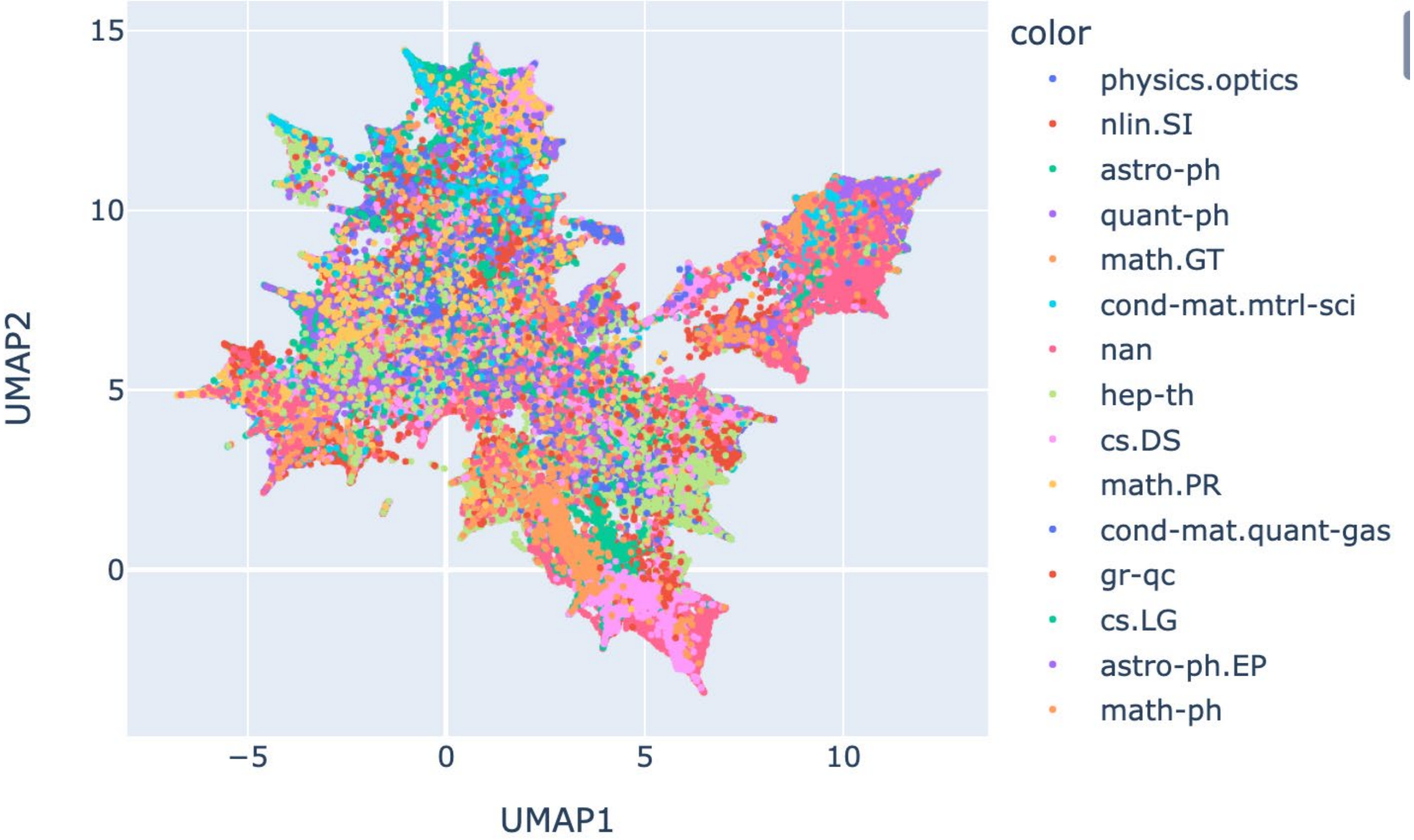
- Correlate the entropy rate against other indicators (e.g.: number of distinct authors,...)
- Expand methodology to other data sets
  - Reddit (subreddit), Patent, and APS dataset

3 Ì !đÌ –!! .†İ π!!†Êđ!Ì ð. .π″ !!
µ†ð†!!′πØÌ Â Â πÊµ†ðÕ Ê↨!!

Hope you liked our presentation!

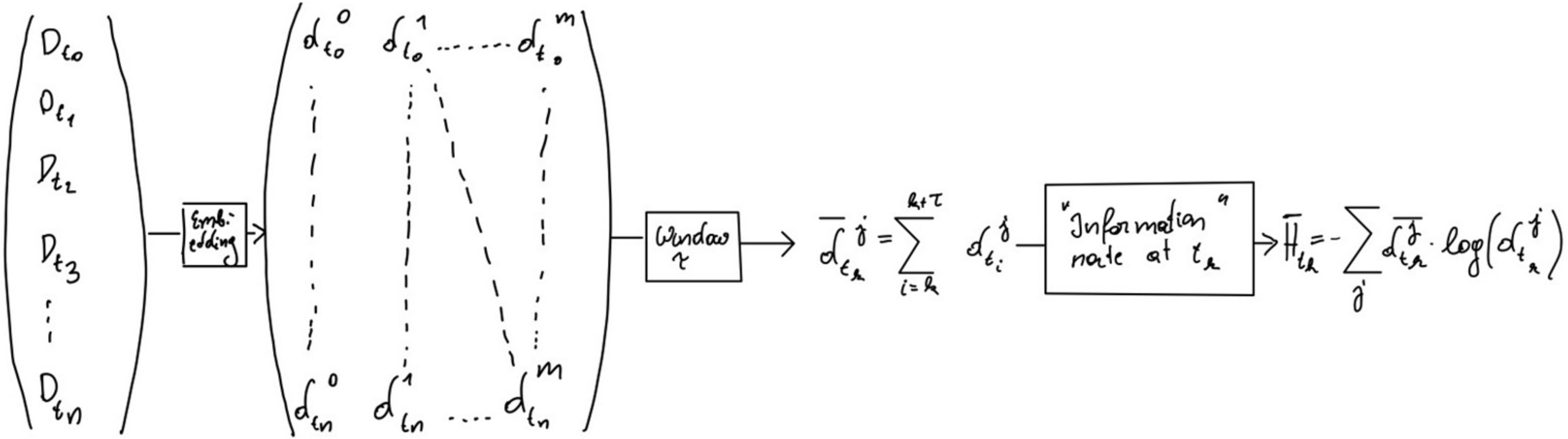
FIN

# Embedding


15
10
5
0
−5
0
5
10
UMAP1
UMAP2
color
● physics.optics
● nlin.SI
● astro-ph
● quant-ph
● math.GT
● cond-mat.mtrl-sci
● nan
● hep-th
● cs.DS
● math.PR
● cond-mat.quant-gas
● gr-qc
● cs.LG
● astro-ph.EP
● math-ph

# Integrating topic modeling and word embedding to characterize violent deaths

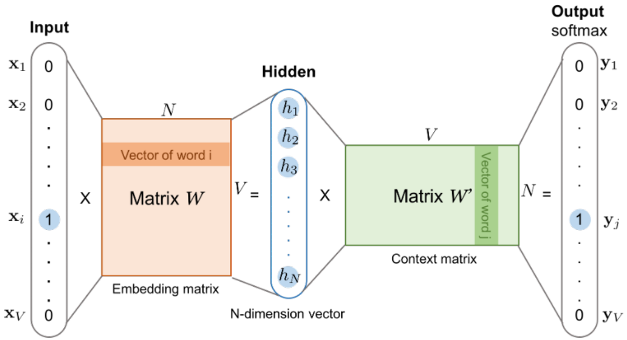
Alina Arseniev-Koehler[a,b,1], Susan D. Cochran[b,c,d], Vickie M. Mays[b,e,f], Kai-Wei Chang[b,g], and Jacob G. Foster[a,b,1]

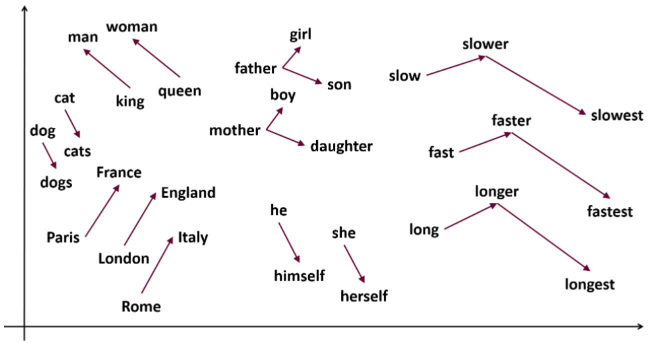

$D_{t_0}$
$D_{t_1}$
$D_{t_2}$
$D_{t_3}$
$D_{t_n}$
Embedding
$d_{t_0}^0$ $d_{t_0}^1$ $d_{t_0}^m$
$d_{t_n}^0$ $d_{t_n}^1$ $d_{t_n}^m$
Window $\tau$
$\bar{d}_{t_k}^j = \sum_{i=k}^{k+\tau} d_{t_i}^j$
"Information" rate at $t_k$
$\bar{H}_{t_k} = -\sum_j \bar{d}_{t_k}^j \cdot \log(d_{t_k}^j)$

$$\forall t_i: \sum_j d_{t_i}^j = 1$$

# Latent embedding space

## WORD2VEC EMBEDDING





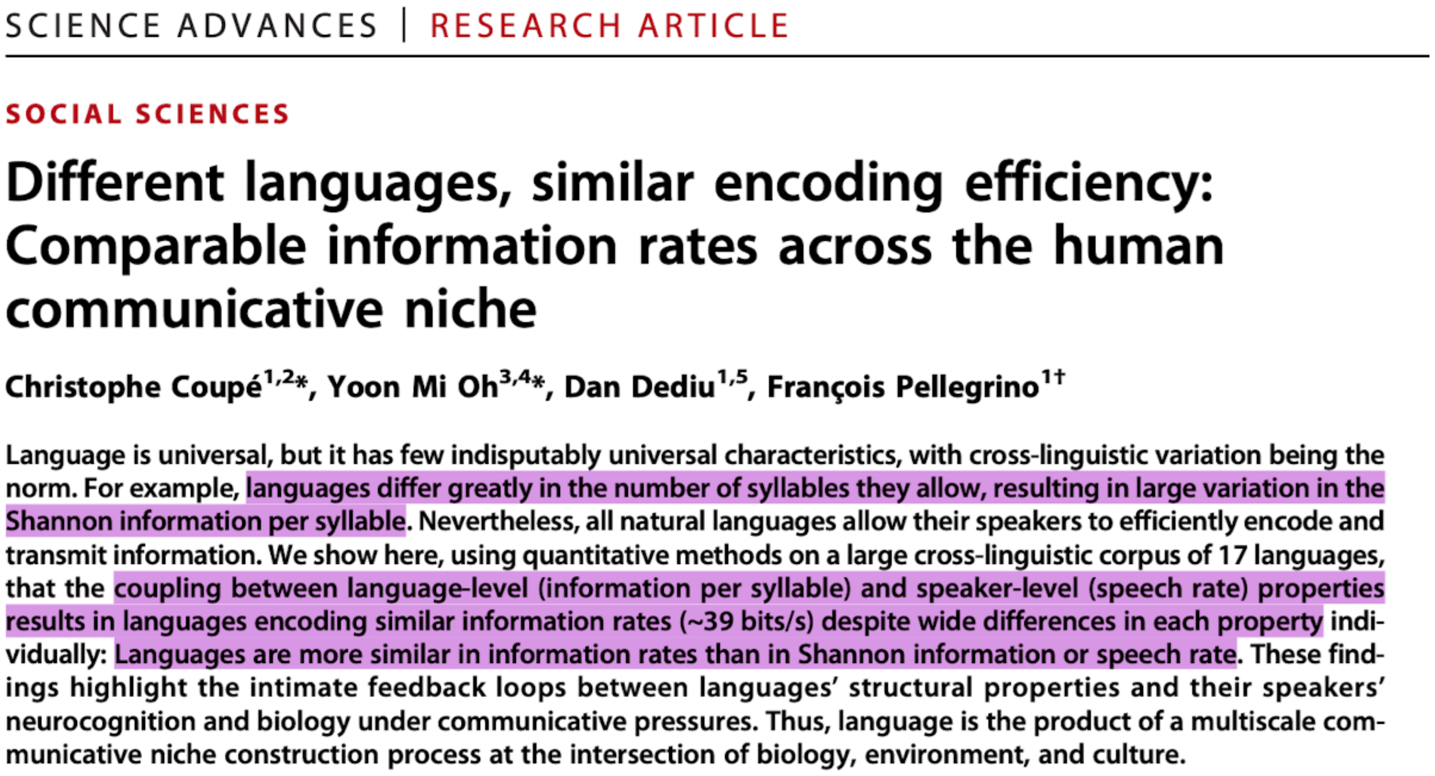
## TOPIC EMBEDDING

$$\Pr[w \text{ emitted at } t \,|\, \mathbf{c}_t] = \frac{\exp(\langle \mathbf{c}_t, \mathbf{w} \rangle)}{Z_{\mathbf{c}_t}}.$$

$$\Pr[w \text{ emitted at } t \,|\, \mathbf{c}_t] = \alpha p(w) + (1-\alpha)\frac{\exp(\langle \widetilde{\mathbf{c}}_t, \mathbf{w} \rangle)}{Z_{\widetilde{\mathbf{c}_t}}},$$

$$(\widetilde{\mathbf{c}}_t)_{\mathrm{MAP}} = \sum_{w \in \mathscr{C}} \frac{a}{p(w)+a}\mathbf{w}, \text{ where } a = \frac{1-\alpha}{\alpha Z}.$$

# Motivation



SCIENCE ADVANCES | RESEARCH ARTICLE

SOCIAL SCIENCES

## Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche

Christophe Coupé[1,2]*, Yoon Mi Oh[3,4]*, Dan Dediu[1,5], François Pellegrino[1]†

Language is universal, but it has few indisputably universal characteristics, with cross-linguistic variation being the norm. For example, languages differ greatly in the number of syllables they allow, resulting in large variation in the Shannon information per syllable. Nevertheless, all natural languages allow their speakers to efficiently encode and transmit information. We show here, using quantitative methods on a large cross-linguistic corpus of 17 languages, that the coupling between language-level (information per syllable) and speaker-level (speech rate) properties results in languages encoding similar information rates (~39 bits/s) despite wide differences in each property individually: Languages are more similar in information rates than in Shannon information or speech rate. These findings highlight the intimate feedback loops between languages' structural properties and their speakers' neurocognition and biology under communicative pressures. Thus, language is the product of a multiscale communicative niche construction process at the intersection of biology, environment, and culture.