



machine learning  
excellence delivered

# Context Search with Advanced Topic Analysis

# Why search?



- Knowledge workers spent up to **2.5 hour per day\*** to find the information
  - Equals to **€25 million per year** for 1000-workers corporation
  - Extra cost of **not finding** important data

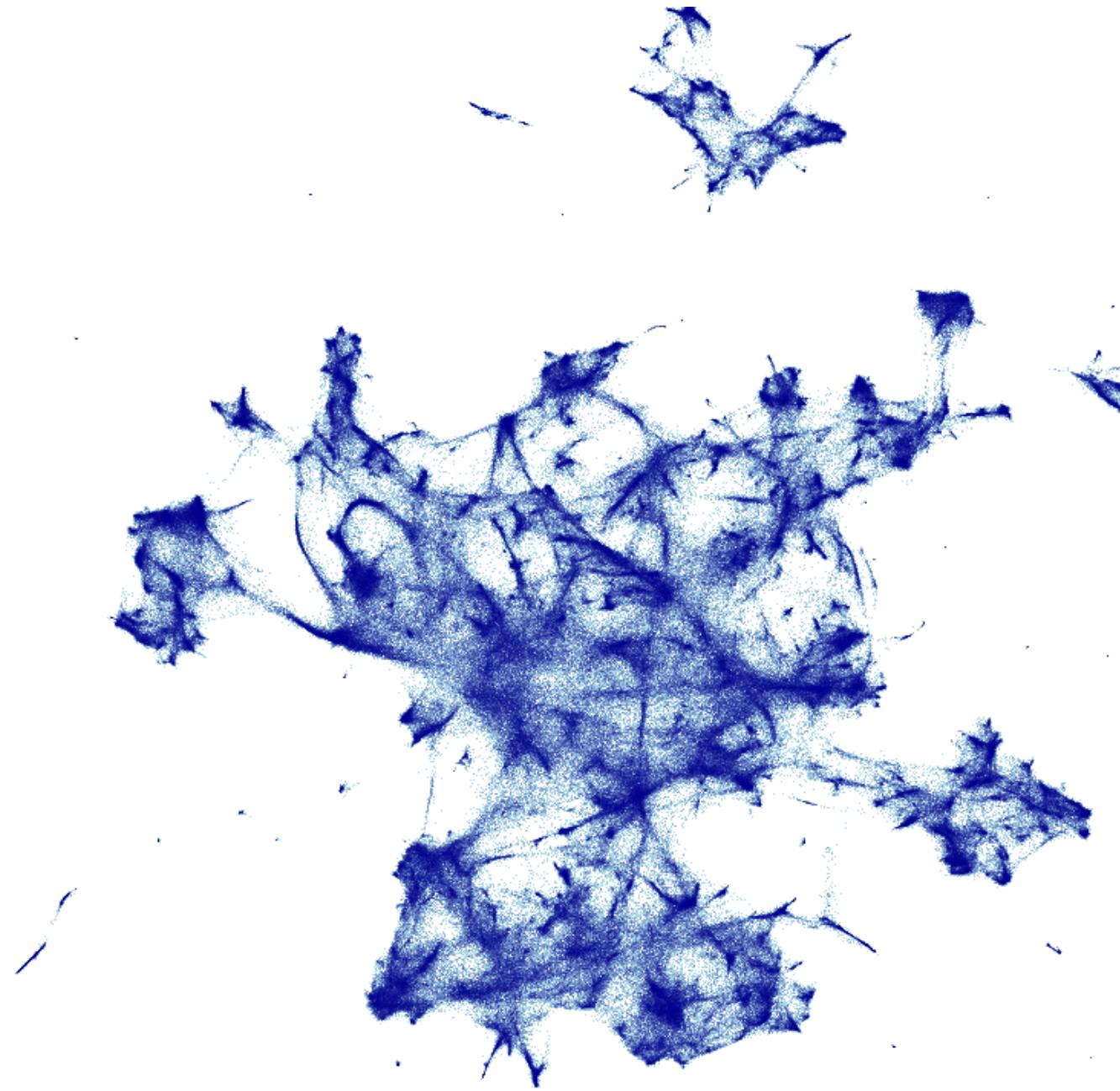
\* Source: <http://intranetfocus.com/2-5-hours-a-day-spent-searching-really/>

# Common search engines view

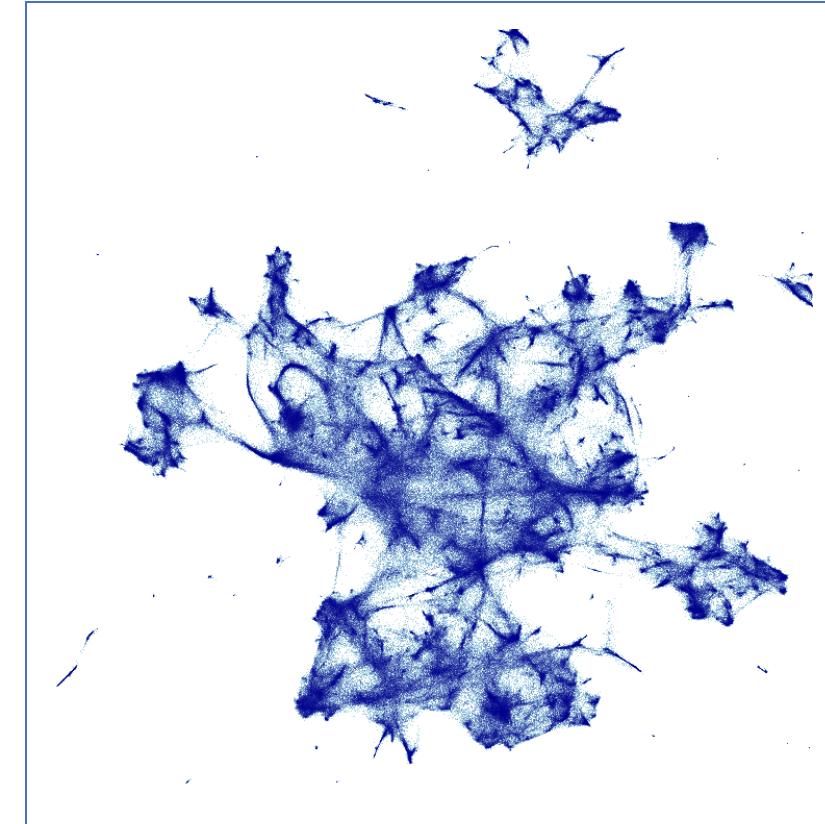
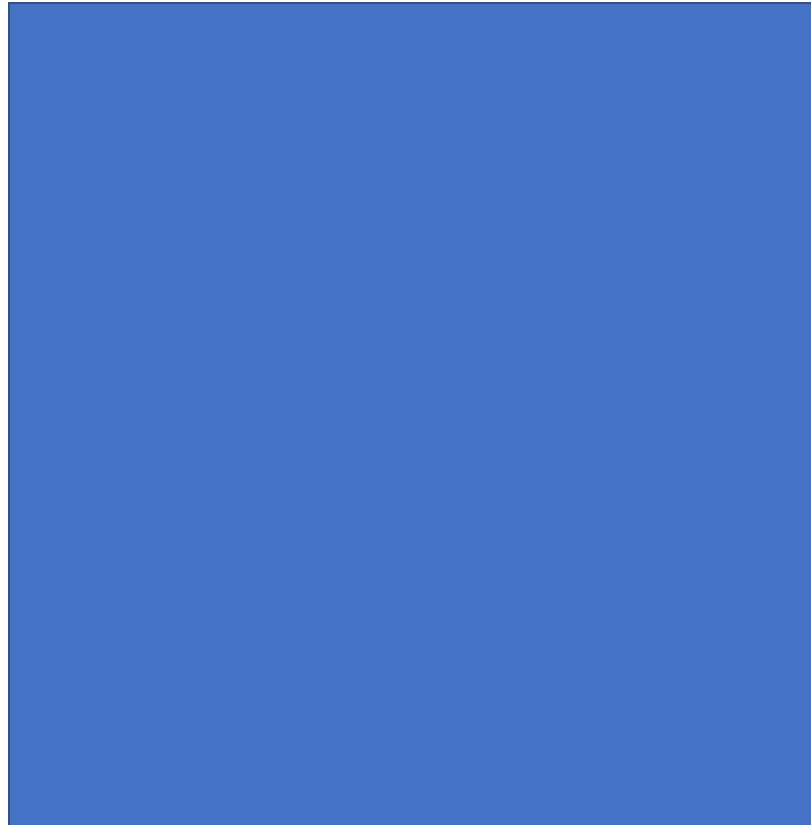


.... as usual search sees the documents

# What is it?



# Our view



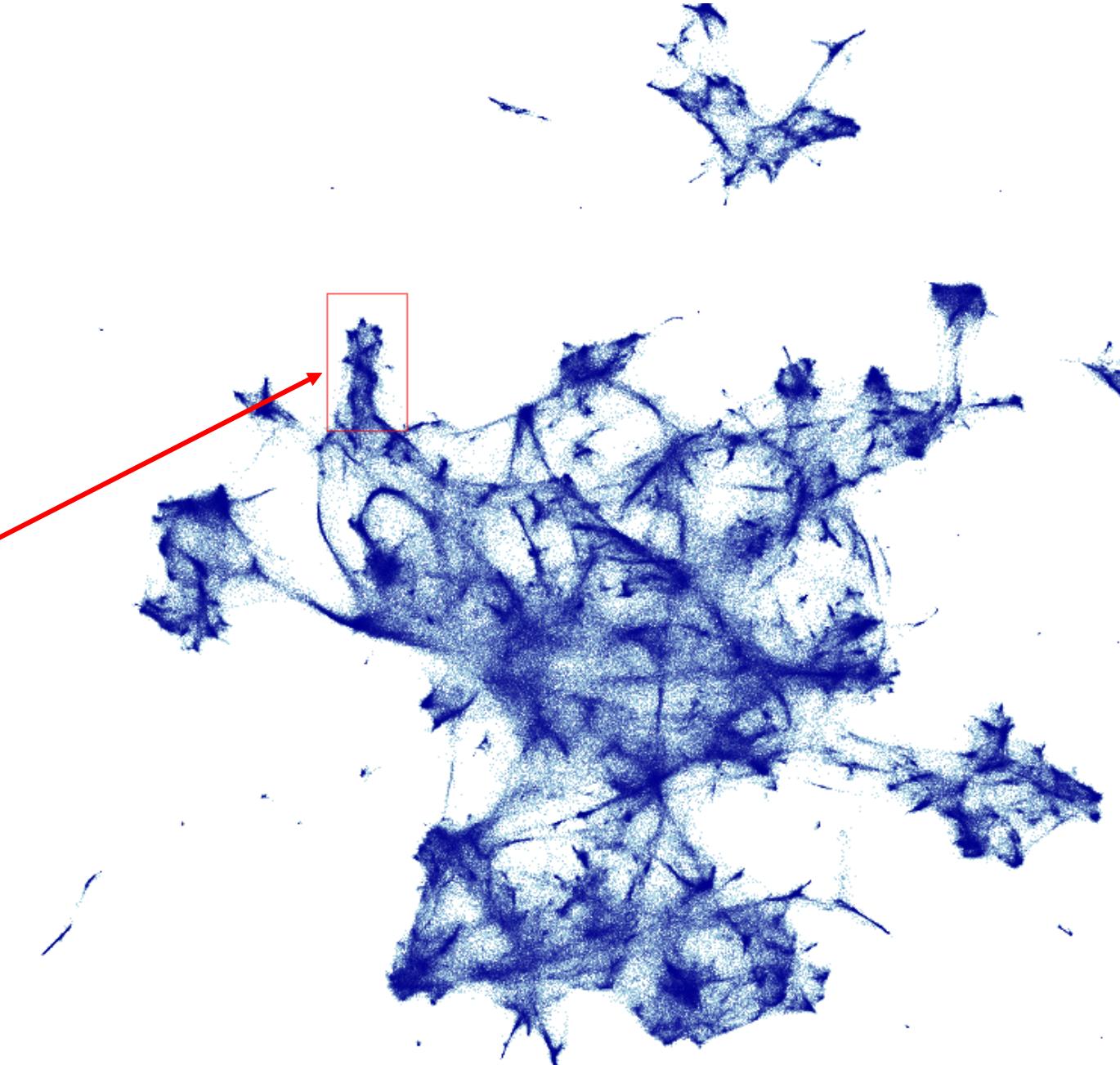
.... as usual search sees the documents

.... as our search sees text collection

# Map of the whole Wikipedia

TV broadcasting:

- Fox, NBC, CBS
- Soap opera
- NFL (on TV)



# Goals



- Help people with information overload
  - **VUCA**-world (**V**olatility, **U**ncertainty, **C**omplexity and **A**mbiguity)
- Analysis of large collections of unstructured documents

# Market sectors

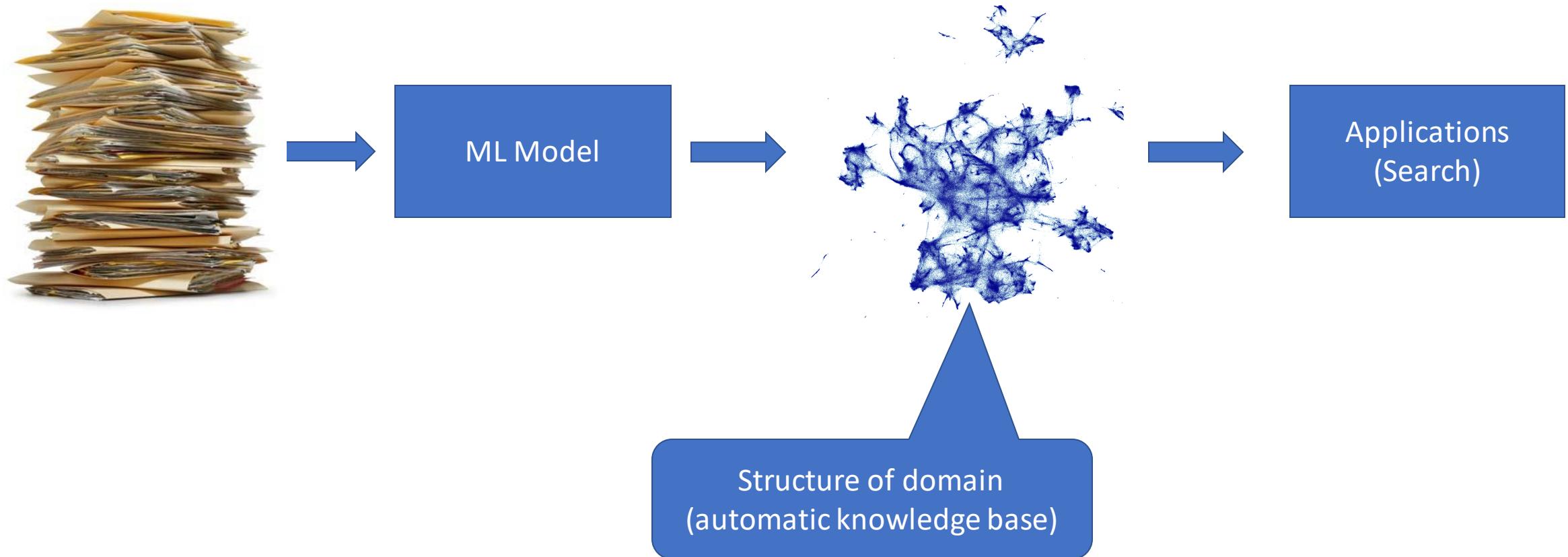
- Research
  - Market
  - Patent
  - ...
- Legal companies
- Banks
- Publishers

# Unstructured Documents

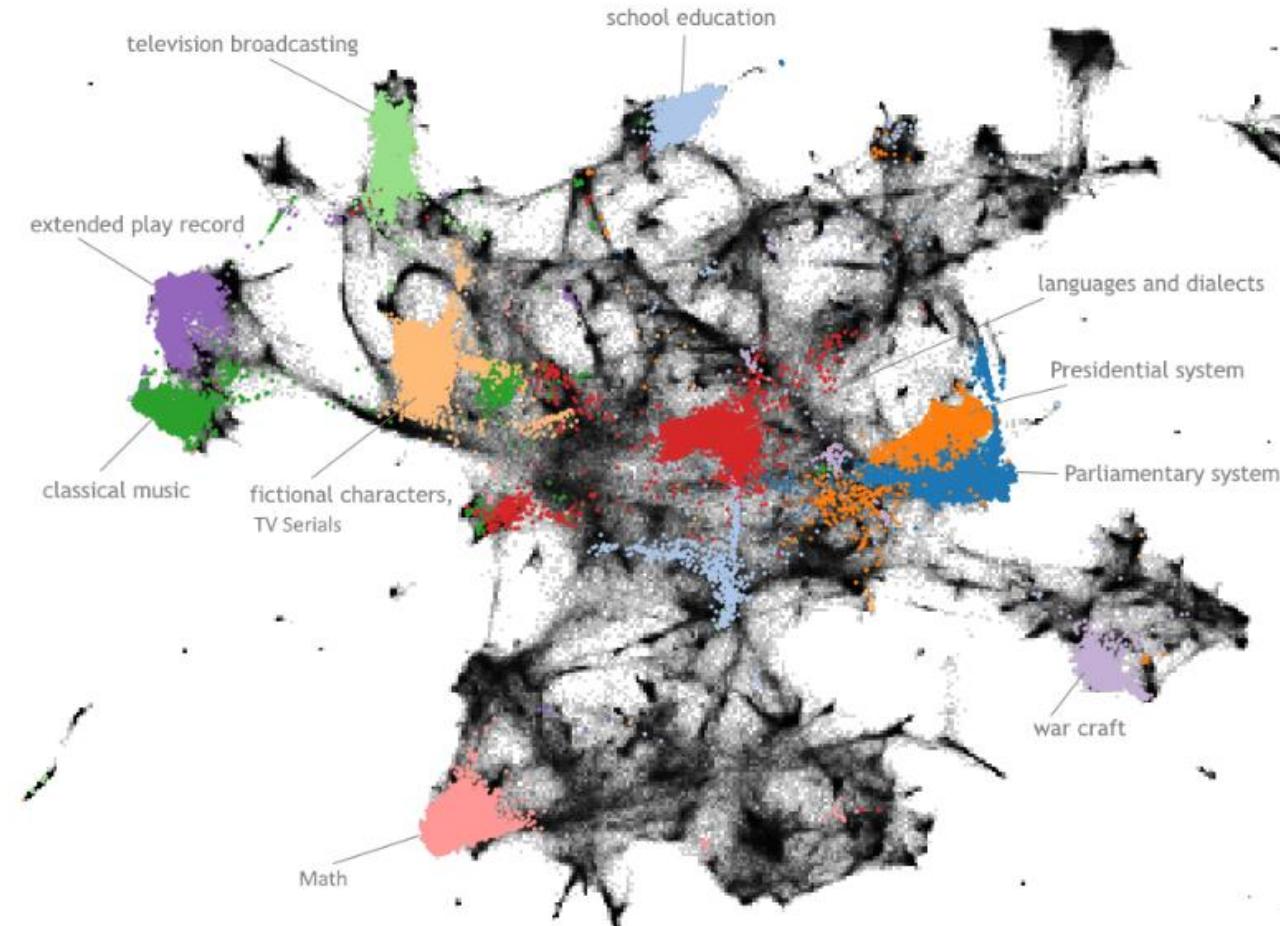


- Legal texts
  - Contracts
  - Due diligence
- Emails
- Biomedical articles (PubMed)
- Corporate Wiki

# Semantic map

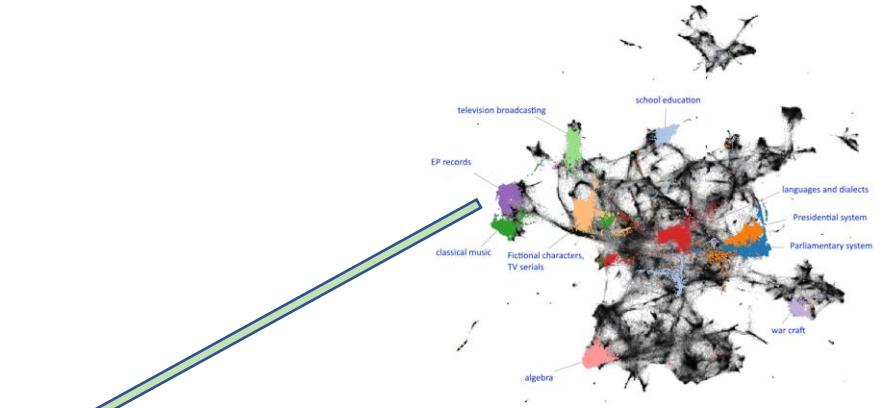
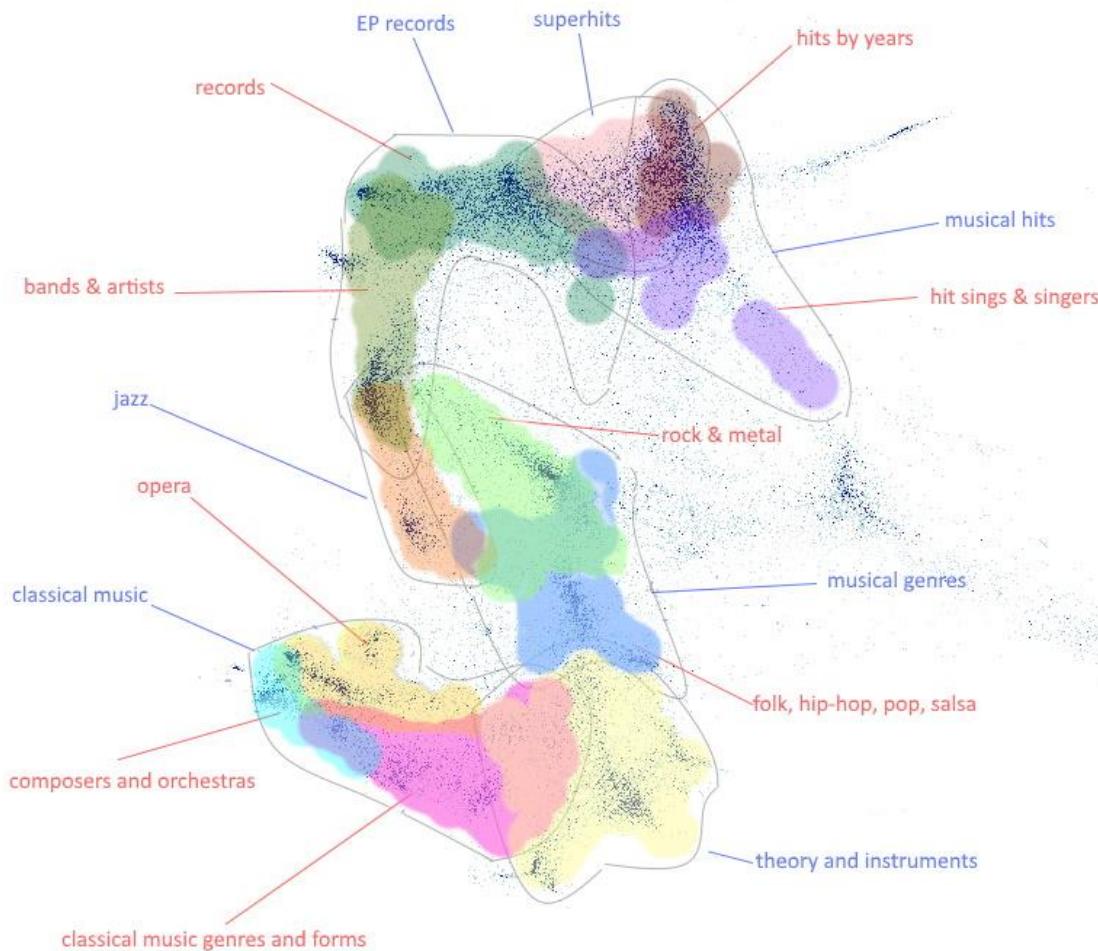


# Wikipedia Topics



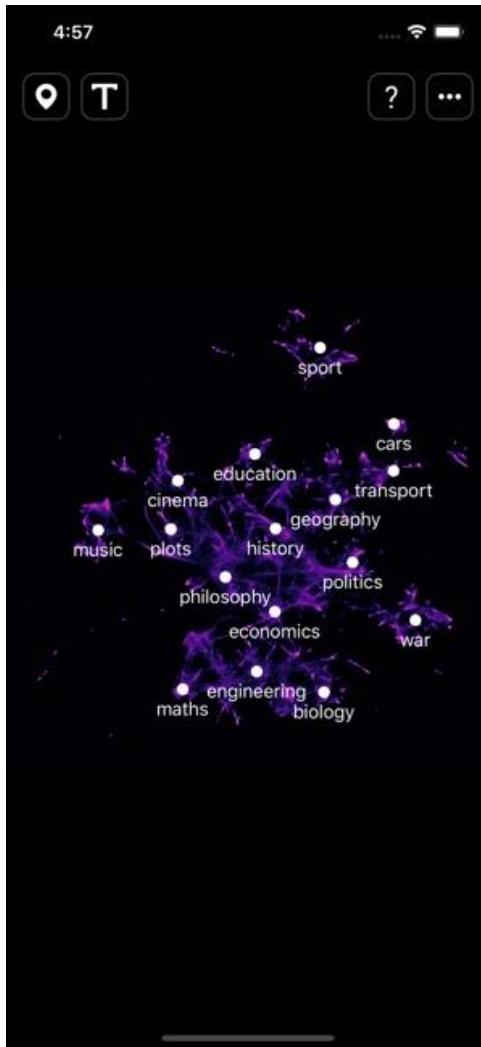
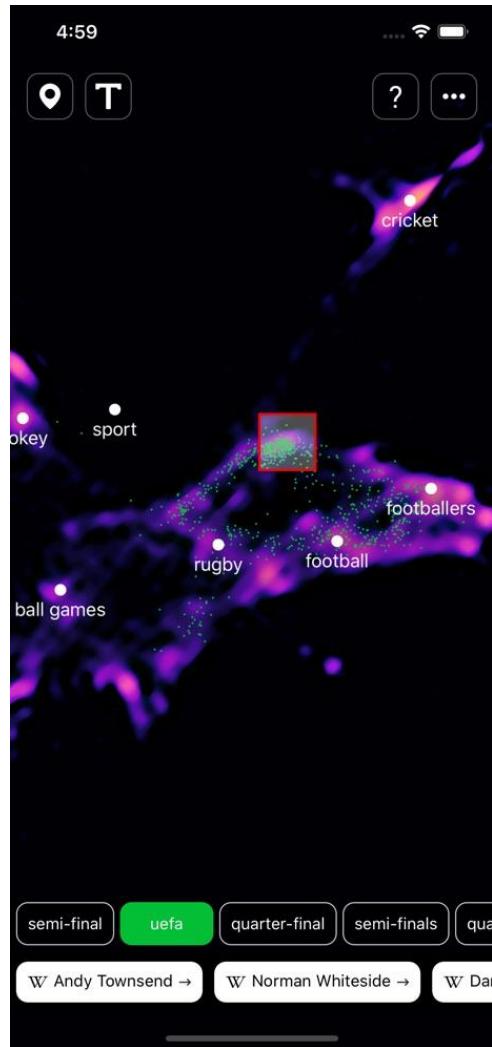
Example available: <https://pegasus.silkcodeapps.de/modeling>

# Music Topics

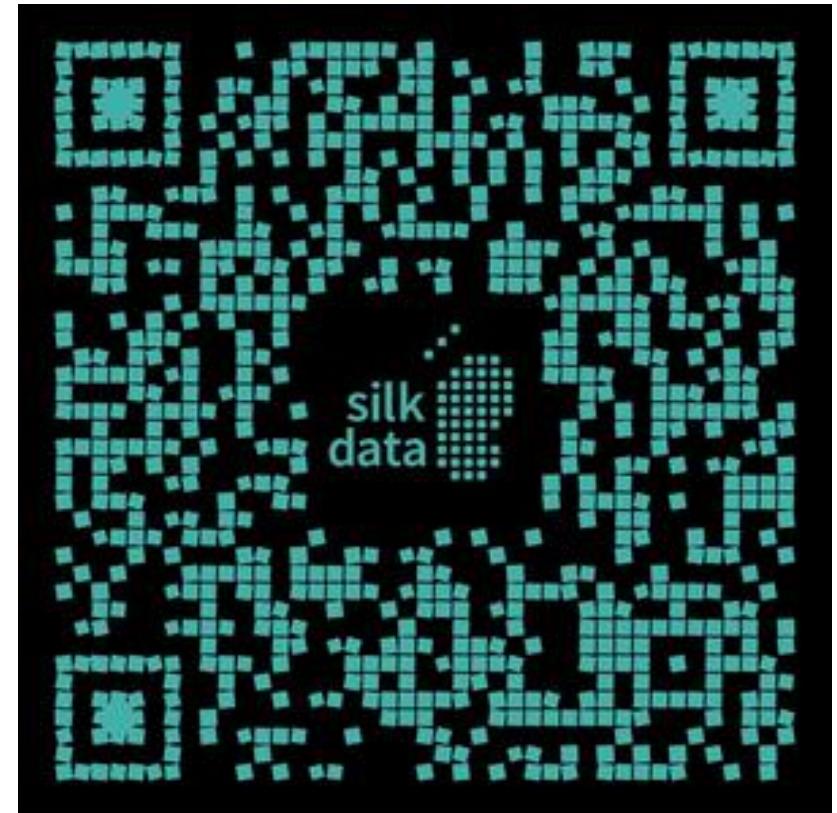


Classes (topics) identified automatically  
We can find up to 200-300 classes

# iOS app



Play now:



# Applications / Products



- **Context search**
- Navigation over book or document collection
- Document classification
- Automatic reading (Semantic segmentation)

# Search: example

 Topic search

---

frankfurt

Topic search    Show relevance   [Filters](#)

---

[frankfurt rhein-main regional authority](#)

- frankfurt rhein-main regional authority the regional authority frankfurtrheinmain be the cooperation body of the administratively fragment

---

[frankfurt school of finance & management](#)

- frankfurt school of finance amp ; management frankfurt school of finance management be a private ,

---

[1. ffc frankfurt](#)

- 1 ffc frankfurt 1 ffc frankfurt be a german woman association football club base in frankfurt , hesse

---

[frankfurt](#)

- frankfurt frankfurt , officially frankfurt be main ( ' frank ford at the main ' )

---

[frankfurt–hanau \(south bank\) railway](#)

- frankfurt hanau ( south bank ) railway the south main line connects frankfurt be main hauptbahnhof

# Search: example

 Topic search

---

frankfurt

[To search map](#)  Topic search  Show relevance [Filters](#)

---

**Category:** Transport ; **Topic:** suburban (local) trains

[capital of germany](#)

- their sovereign . subsequently , frankfurt briefly become the official german capital during the short-lived revolution of 1848 in the german

---

**Category:** Transport ; **Topic:** railway infrastructure

[frankfurt–hanau \(south bank\) railway](#)

- frankfurt hanau ( south bank ) railway the south main line connects frankfurt be main hauptbahnhof

---

**Category:** Transport ; **Topic:** commercial aviation

[frankfurt airport](#)

- frankfurt airport frankfurt airport ( also know a ' rhein-main-flughafen ' ) be a major international

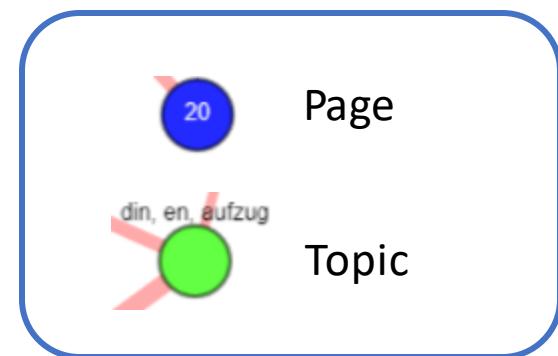
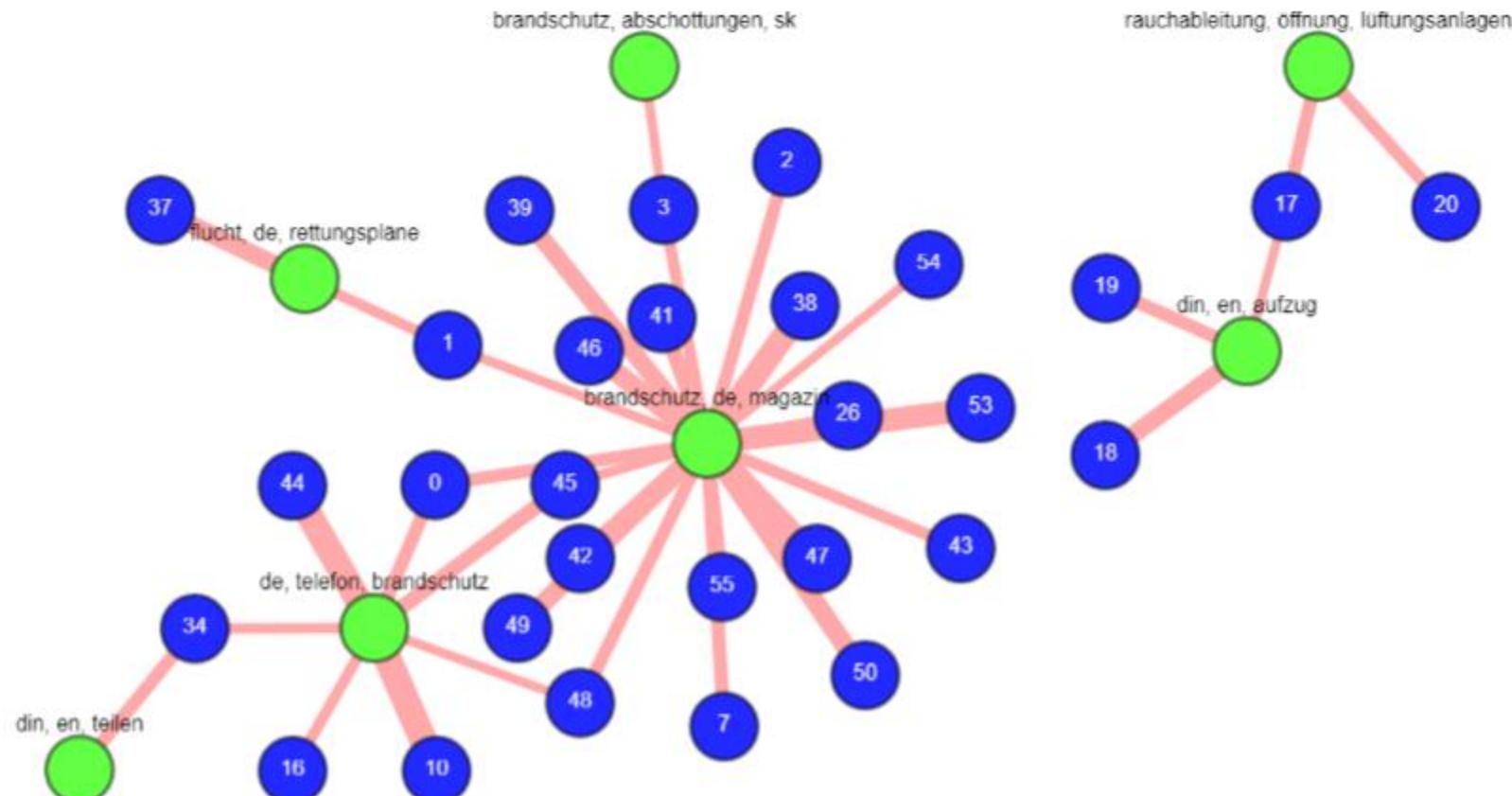
---

**Category:** Sport ; **Topic:** footballers

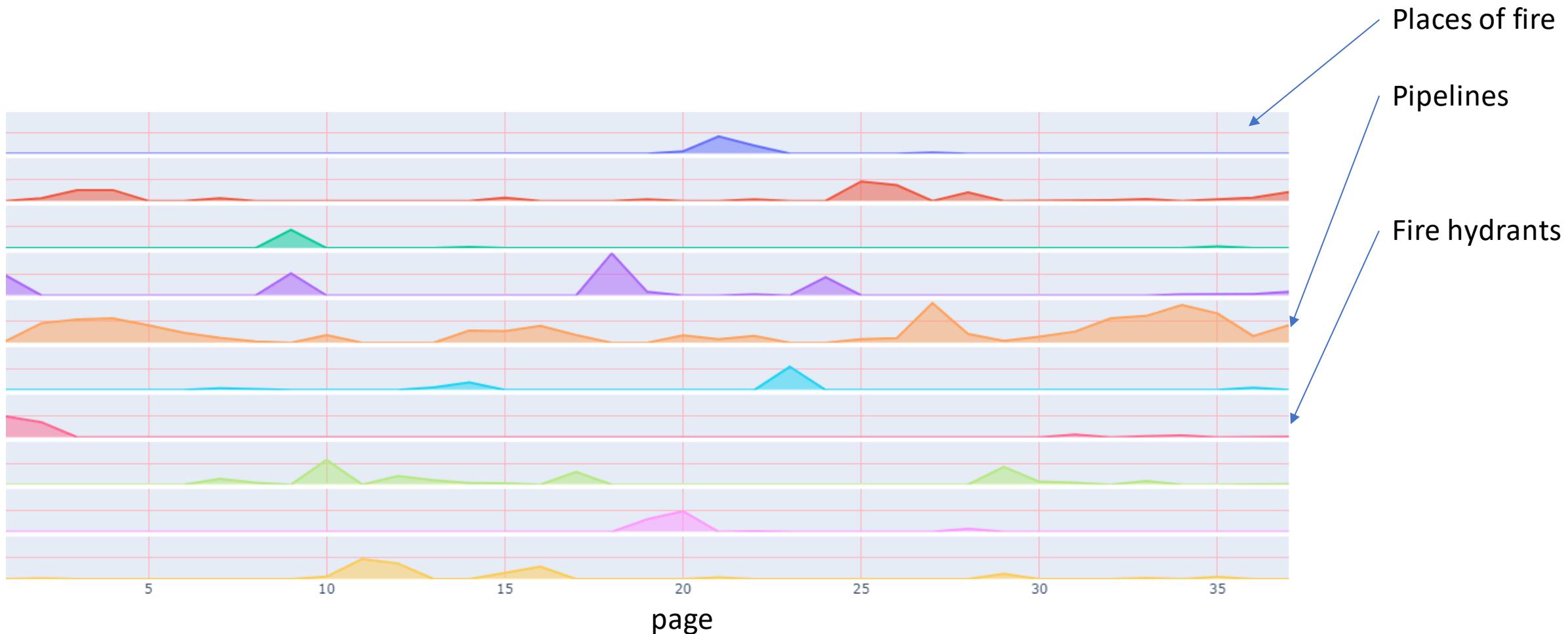
[christopher faust](#)

- december 1968 in frankfurt main , know a chris faust ) be a german hockey coach and trainer .

# Topic graph



# Book map



# Our offer:

- Context search
- Information navigation: books, documents, laws, ...
- Document classification
- Integration with your product

Format-independent: PDF, XML, Word, ...

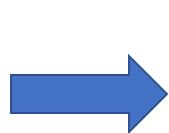
**Initial data analysis: FREE!!!**

(text analysis using Wikipedia topics)

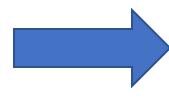


# The workflow

Input documents



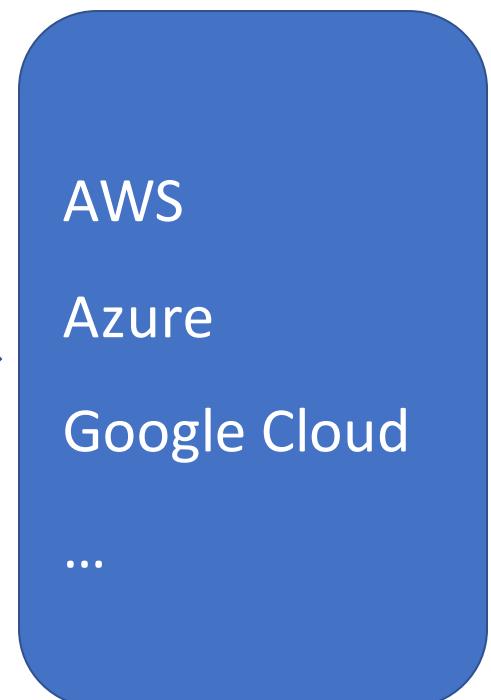
Text extraction



Find structure



Deployment



# Thank you for attention!



visit: [silkdata.ai](https://silkdata.ai)

mail to: [hello@silkdata.ai](mailto:hello@silkdata.ai)



[www.silkdata.ai](http://www.silkdata.ai)  
[hello@silkdata.ai](mailto:hello@silkdata.ai)  
+49 (2151) 387 3531

Headquarters:  
Silk AI lines LLC

Hikaly 3-7  
220005 Minsk  
Belarus

Partner in EU:  
SilkCode GmbH

Luisenstraße 62  
D-47799 Krefeld  
Germany

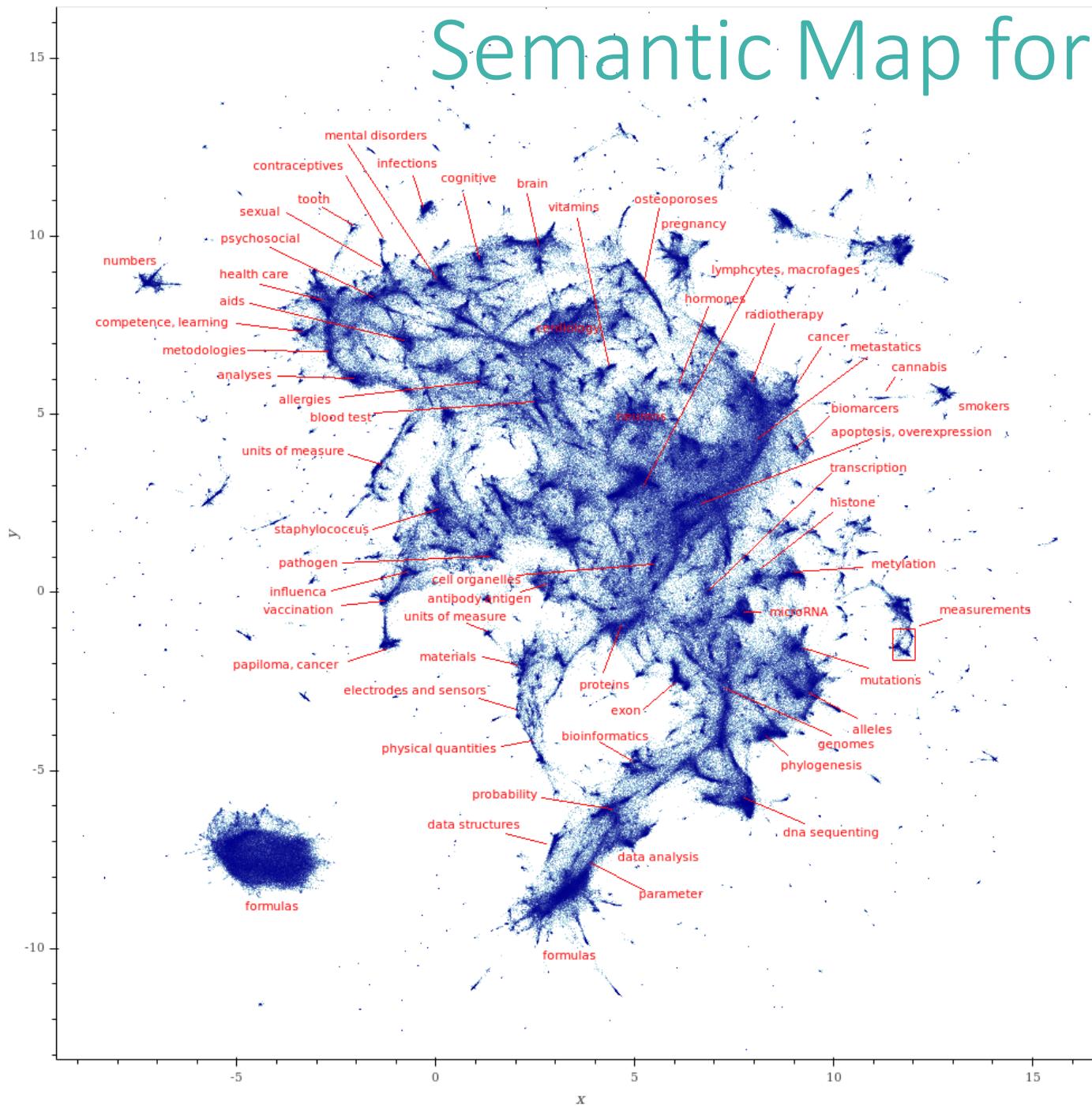
# Extra slides

# Semantic Map Properties



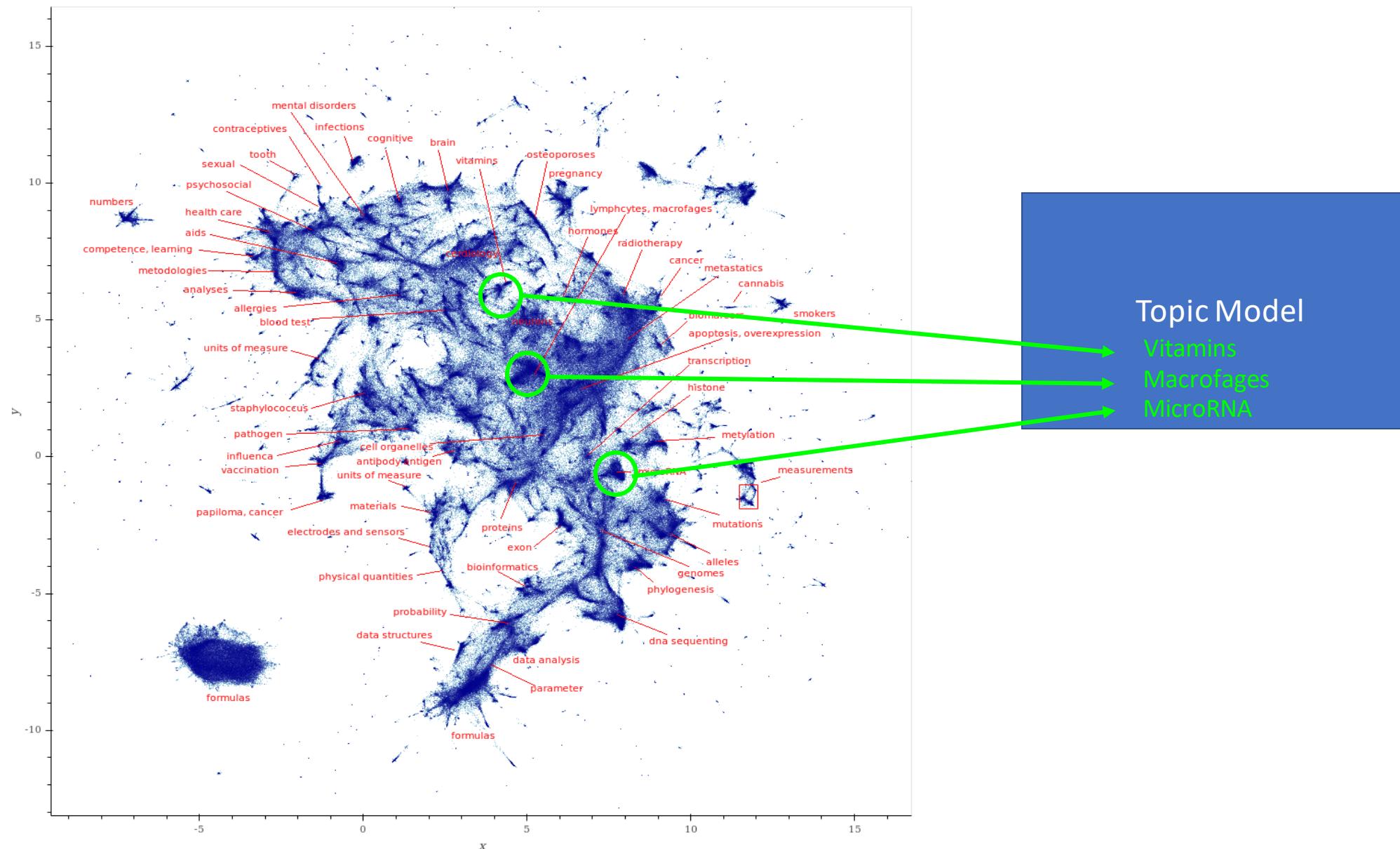
- 1) Semantically similar points are in the same map region
- 2) Point clusters form topics
  - Dense areas on map
  - Examples: rock music, car racing, universities
- 3) Similar topics are located close to each other
- 4) Hierarchical topics
  - Can be combined or split
  - Examples: physics, sport, literature

# Semantic Map for PubMed Articles

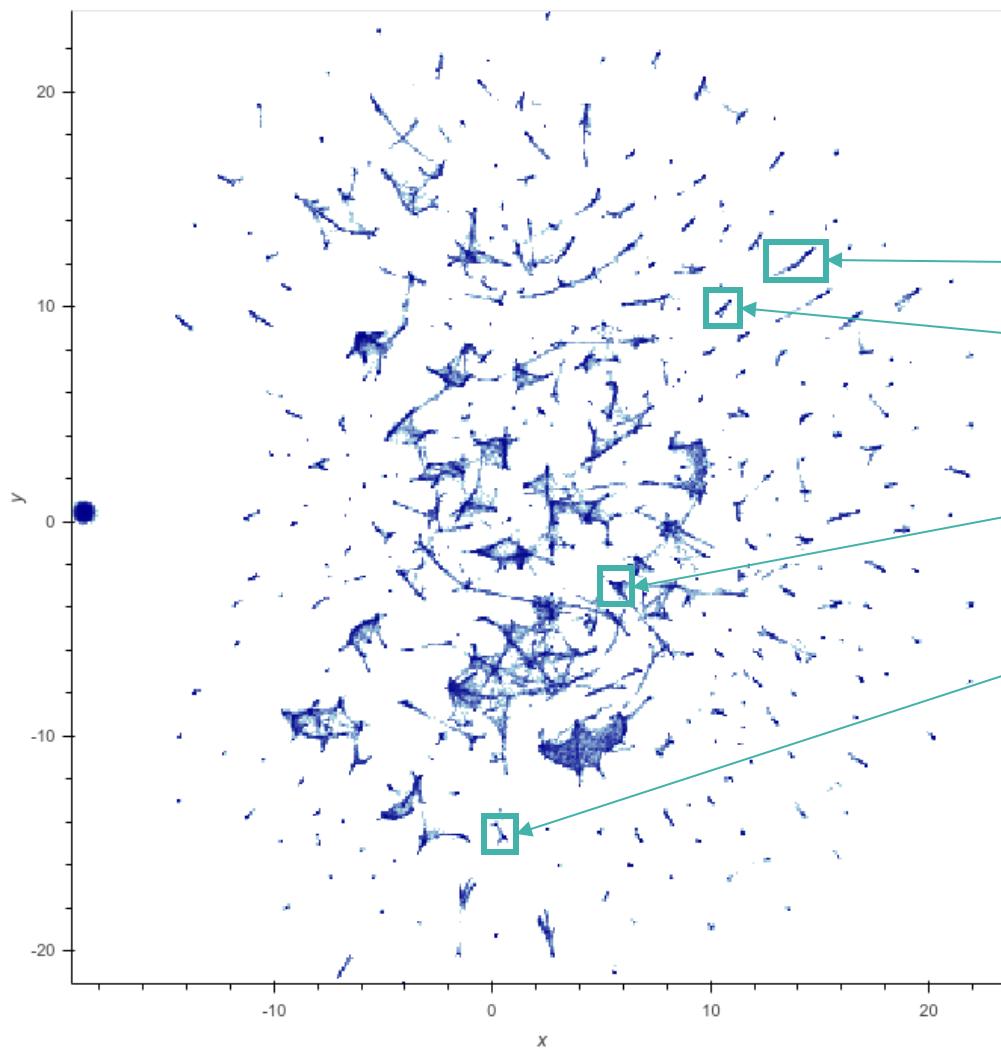


Dense areas on map are syntactically homogeneous and form a topic

# Find Topics



# Document classification



## Identified document classes:

Hospitals

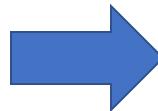
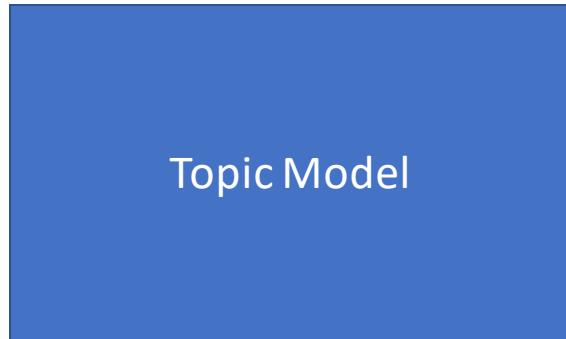
Humanities and libraries

Philharmonic music

Stadiums

Classes (topics) identified automatically  
We can find up to 200-300 classes

# Topics Segmentation



Durch Betriebsvereinbarung kann für einzelne Arbeitnehmer, Arbeitnehmergruppen, Betriebsteile oder für den gesamten Betrieb eine Wochenarbeitszeit zwischen 33 und 43 Stunden vereinbart werden. Der Arbeitgeber soll die Lage einer geänderten Wochenarbeitszeit den Arbeitnehmern jeweils 2 Tage im voraus ansagen.

Der Durchschnitt von 38 Stunden pro Woche muß in zwölf Monaten erreicht werden. Aus betrieblichen Gründen kann der Ausgleichszeitraum um 3 Monate - bis Ende März des Folgejahres - ausgedehnt werden.

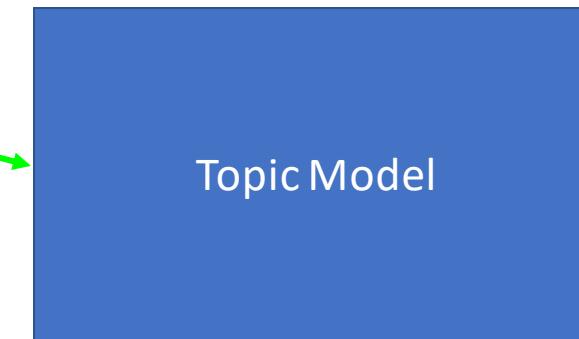
Wenn der Betrieb wegen Materialmangel oder Betriebsstörungen die Arbeit morgens nicht aufnehmen kann oder die Arbeit im Laufe des Tages ruhen muß, wird der Lohn bis zu 8 Stunden einschließlich der an diesem Tage geleisteten Stunden bezahlt. Bestehen Ersatzansprüche gegenüber Dritten, hat der Arbeitnehmer diese dem Arbeitgeber abzutreten.

Arbeitszeit

Lohn

# Topics recognition

microRNA genes are transcribed by RNA polymerase II in the cell nucleus as long single-stranded RNAs (ssRNA) referred to as the “primary microRNA”, or “pri-miRNA”. The ssRNA transcript forms a hairpin loop structure, which signals for RNA nuclease cleavage by a nuclear protein complex called Drosha/Dgcr8. The resulting short hairpin RNA is termed “precursor-microRNA” or “pre-miRNA”. Pre-miRNAs are released into the cytoplasm by the nuclear export protein Exportin 5 and undergo cleavage by the enzyme Dicer into ~22 nucleotide-long mature microRNAs. After strand separation, the single-stranded mature microRNAs become incorporated into the microRNA induced silencing complex (miRISC). Through the RISC, the microRNA binds to the 3'UTR of mRNA molecules by direct base pairing. Binding of miRISC to a mRNA results in post-transcriptional gene silencing either through inhibition of mRNA translation or mRNA destabilization. The 5'UTR region of microRNAs is also known as seed region (nucleotides 1 through 8) and has the most crucial impact on targeting and function. MicroRNAs do not require perfect complementarity for target recognition and a single microRNA is capable of regulating up to hundred or more mRNA species.

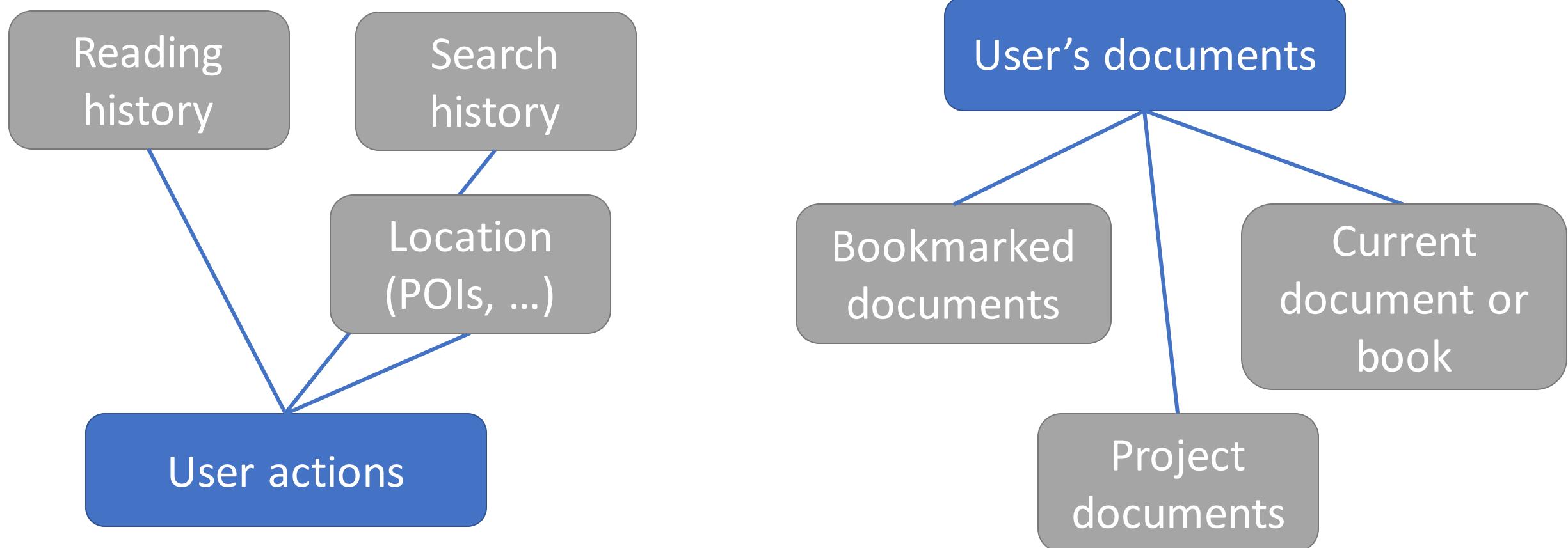


Topic Model



Topics List  
microRNA  
cytoplasm  
protein

# Search: context



Where to get context?