Text-representing Centroids as Instruments of Document Analysis and Classification

Herwig Unger
Mario Kubek

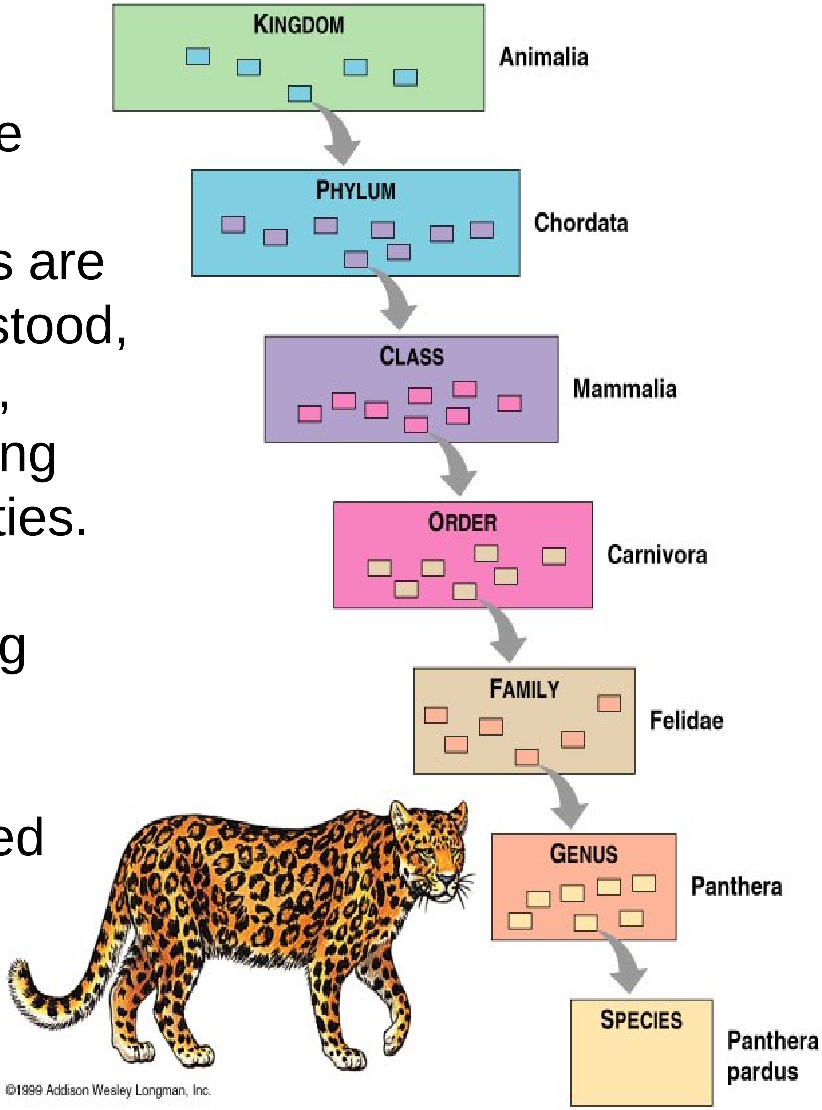
Categorisation is needed

Categorisation

- **….** is a cognitive process to sort objects and entities, and to make the chaos of the world understandable. It is
  - ✓ a process, in which ideas and objects are recognised, differentiated and understood,
  - ✓ requiring significant prior knowledge,
  - ✓ basing on abstraction, i.e. term building and disregarding insignificant properties.

- Plato introduced the approach of grouping objects based on their *similar properties*.

- Aristotle further explored and systematised this approach by introducing *classes* and *objects*.



©1999 Addison Wesley Longman, Inc.

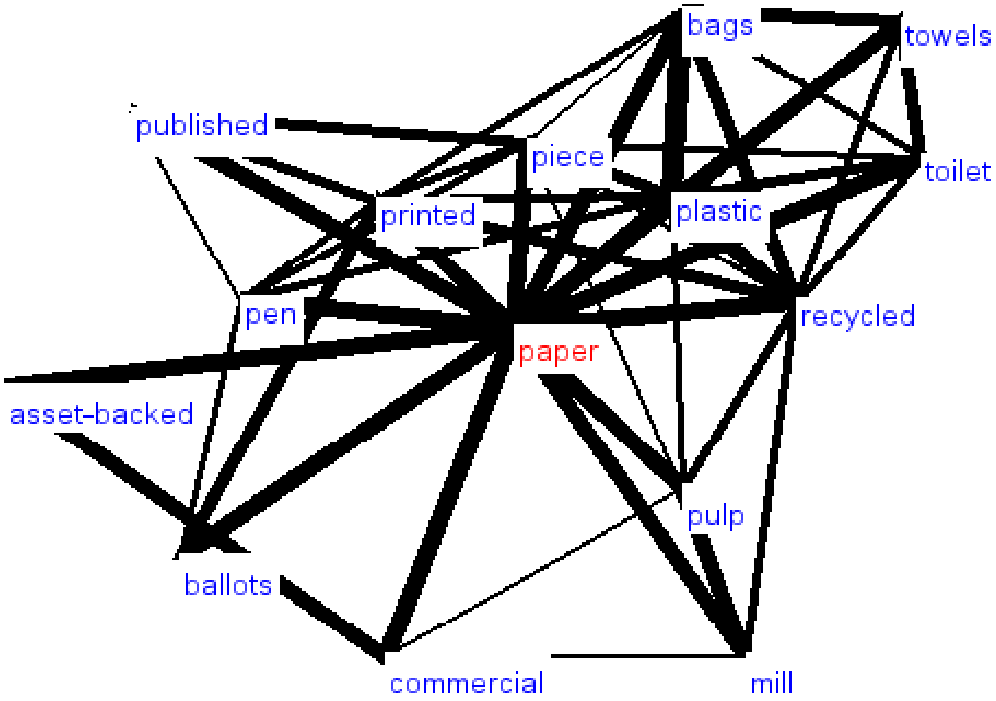
Jeff Hawkins: “On Intelligence”

Preliminaries

The Basics: Co-occurrence Analysis

- □ Significant co-occurrences appear with probabilities above a specific threshold in sentences (sentence level), in paragraphs (paragraph level) or in whole texts (document level).

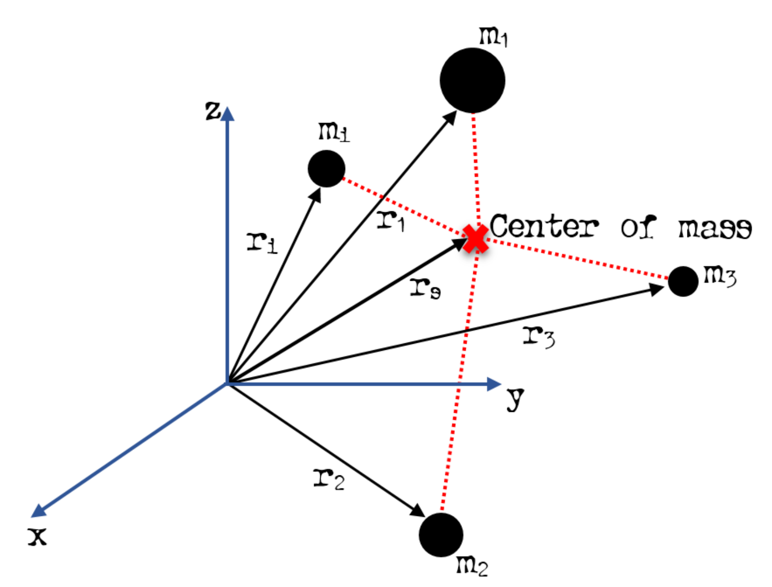
- □ The set of all significant co-occurrences can be represented by a co-occurrence graph (usually undirected): nodes-terms, edges-relations



Source: *corpora.uni-leipzig.de*

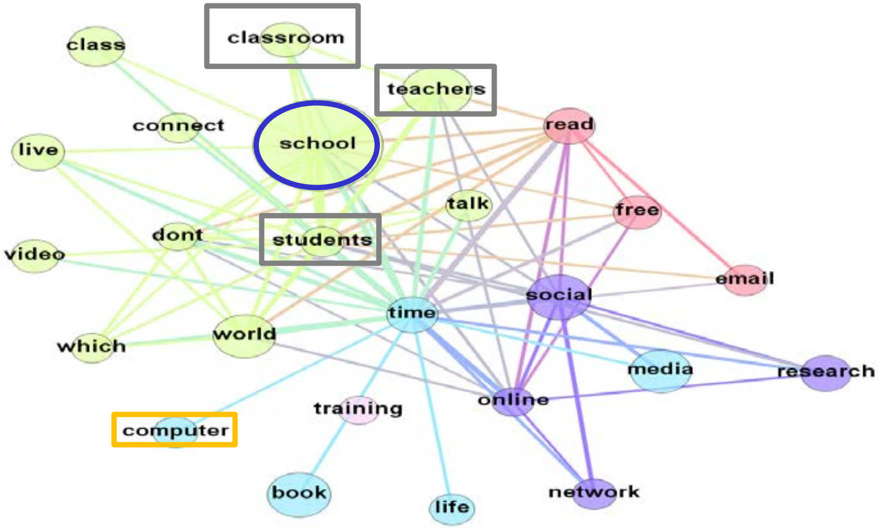
Document Centroids

The physical analogon:
→ centre of mass



•	words	= mass point
•	distance vector	= distance in co-occ. graph

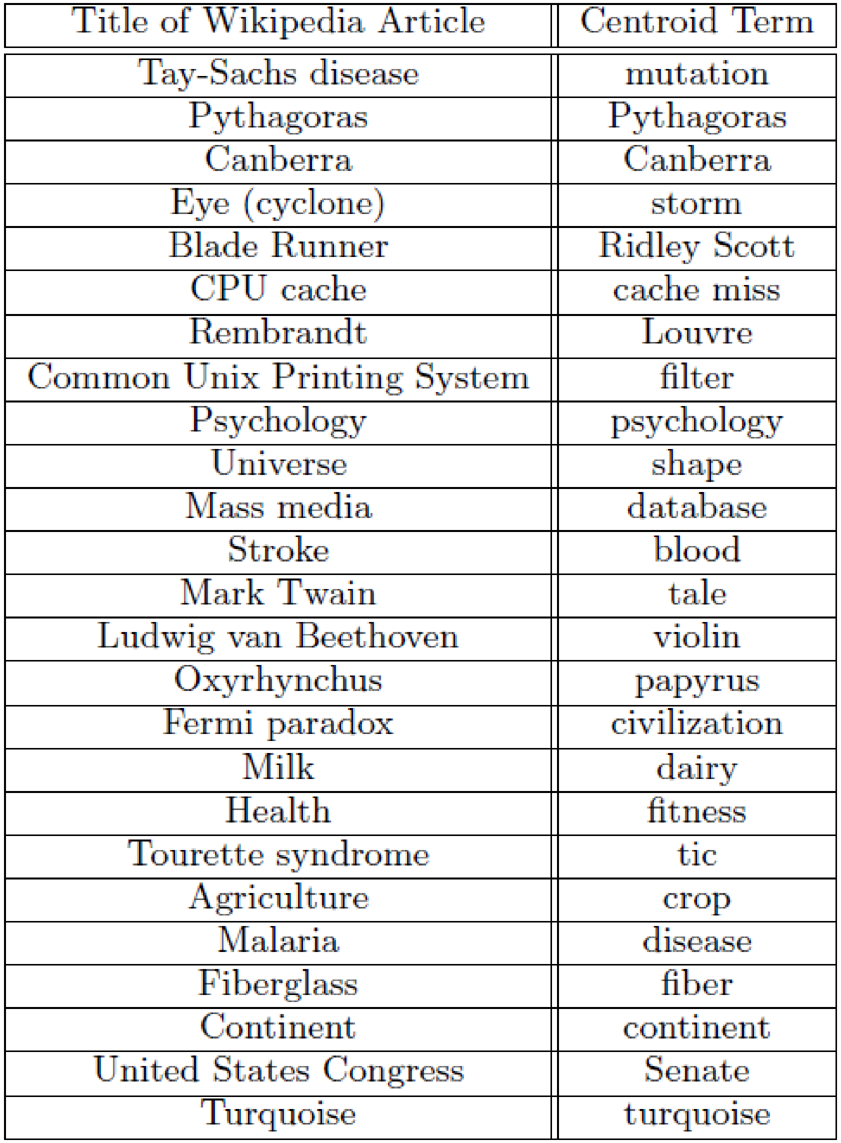
→ e.g. school is the centroid of a document containing classroom, students, teacher but also computer



→ The centroid of a document is the term with the minimum average distance to all words of the respective document in the co-occ. graph.

Properties of Centroids

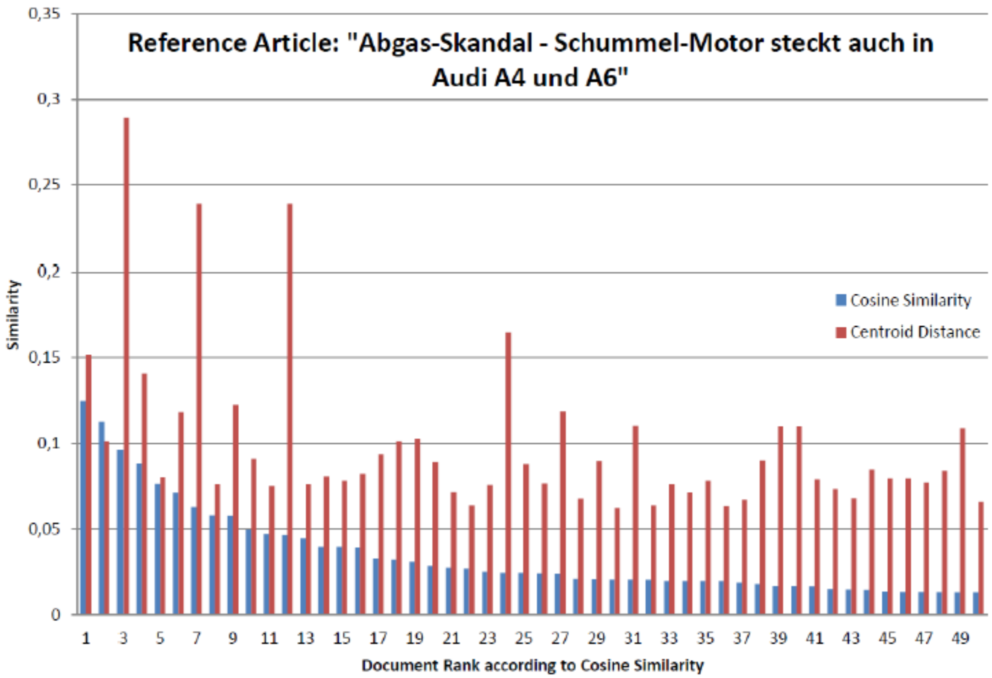
1. Expressivity



Title of Wikipedia Article	Centroid Term
Tay-Sachs disease	mutation
Pythagoras	Pythagoras
Canberra	Canberra
Eye (cyclone)	storm
Blade Runner	Ridley Scott
CPU cache	cache miss
Rembrandt	Louvre
Common Unix Printing System	filter
Psychology	psychology
Universe	shape
Mass media	database
Stroke	blood
Mark Twain	tale
Ludwig van Beethoven	violin
Oxyrhynchus	papyrus
Fermi paradox	civilization
Milk	dairy
Health	fitness
Tourette syndrome	tic
Agriculture	crop
Malaria	disease
Fiberglass	fiber
Continent	continent
United States Congress	Senate
Turquoise	turquoise

- ✓ A centroid may be a word, which is not contained in any of the documents.
- ✓ Often, generalising terms will be found.
- ✓ Theoretically, a document may have more than one centroid.
- ✓ Centroid terms can be assigned to long texts as well as to short queries with only a few words.

2. Similarity

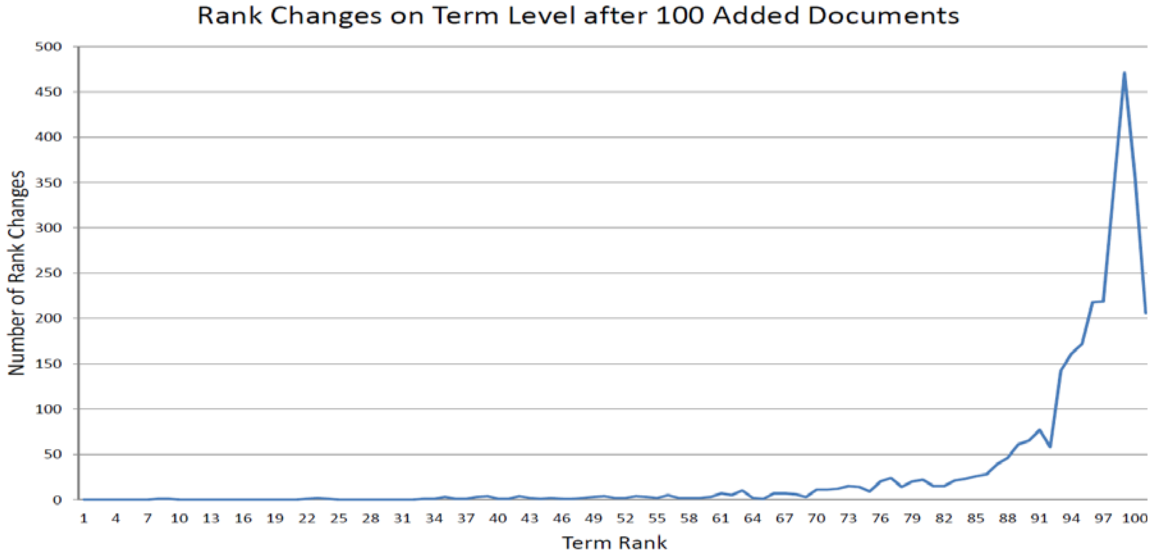


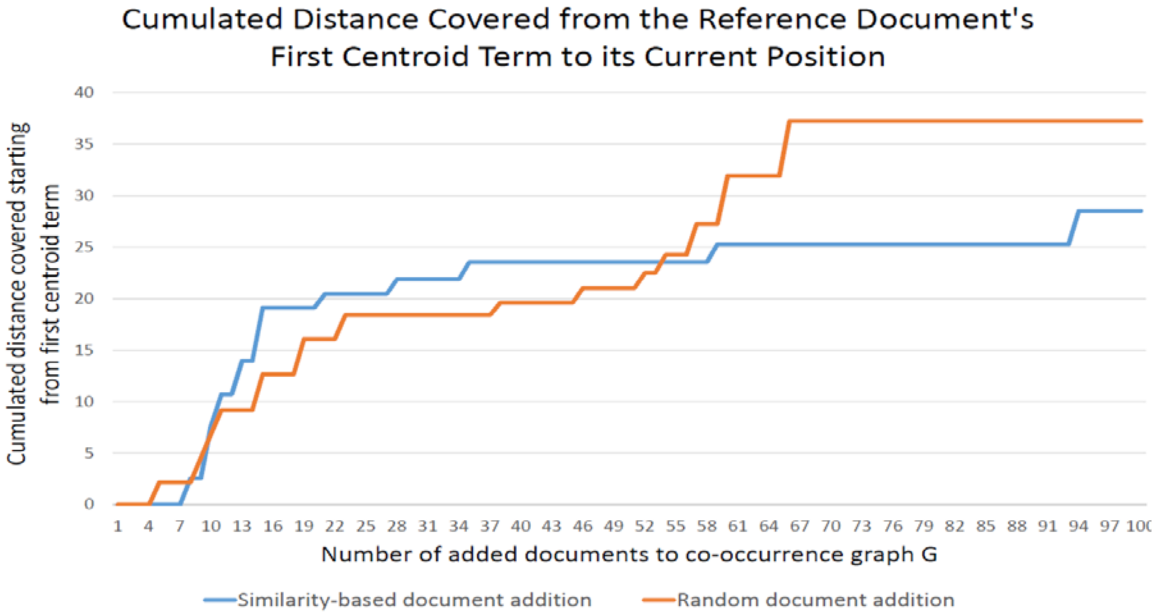
- ✓ The distance of two centroids in the co-occurrence graph can be used to determine the similarity of two documents.
  The smaller the distance is, the more similar the two documents are.

- ✓ Also, texts from different authors using a different wording may be compared (successfully).

- ✓ The centroid similarity can be distinguished (sometimes) from other similarity measures (i.e. cosine similarity).
  Usually, it better reflects content aspects, especially of multidisciplinary texts.

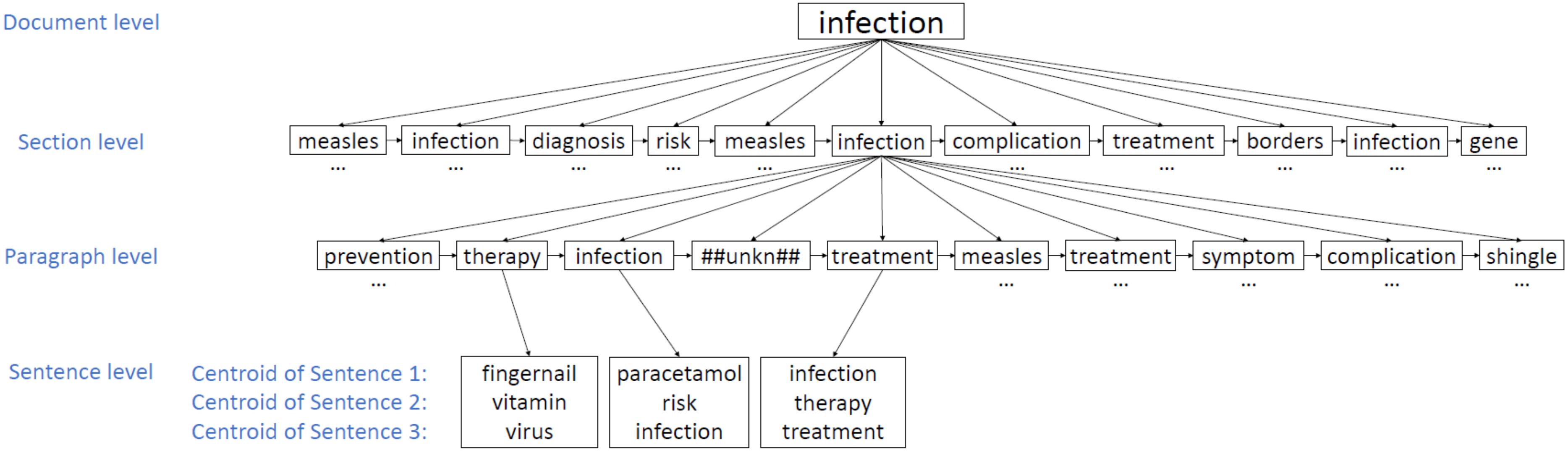
3. Stability





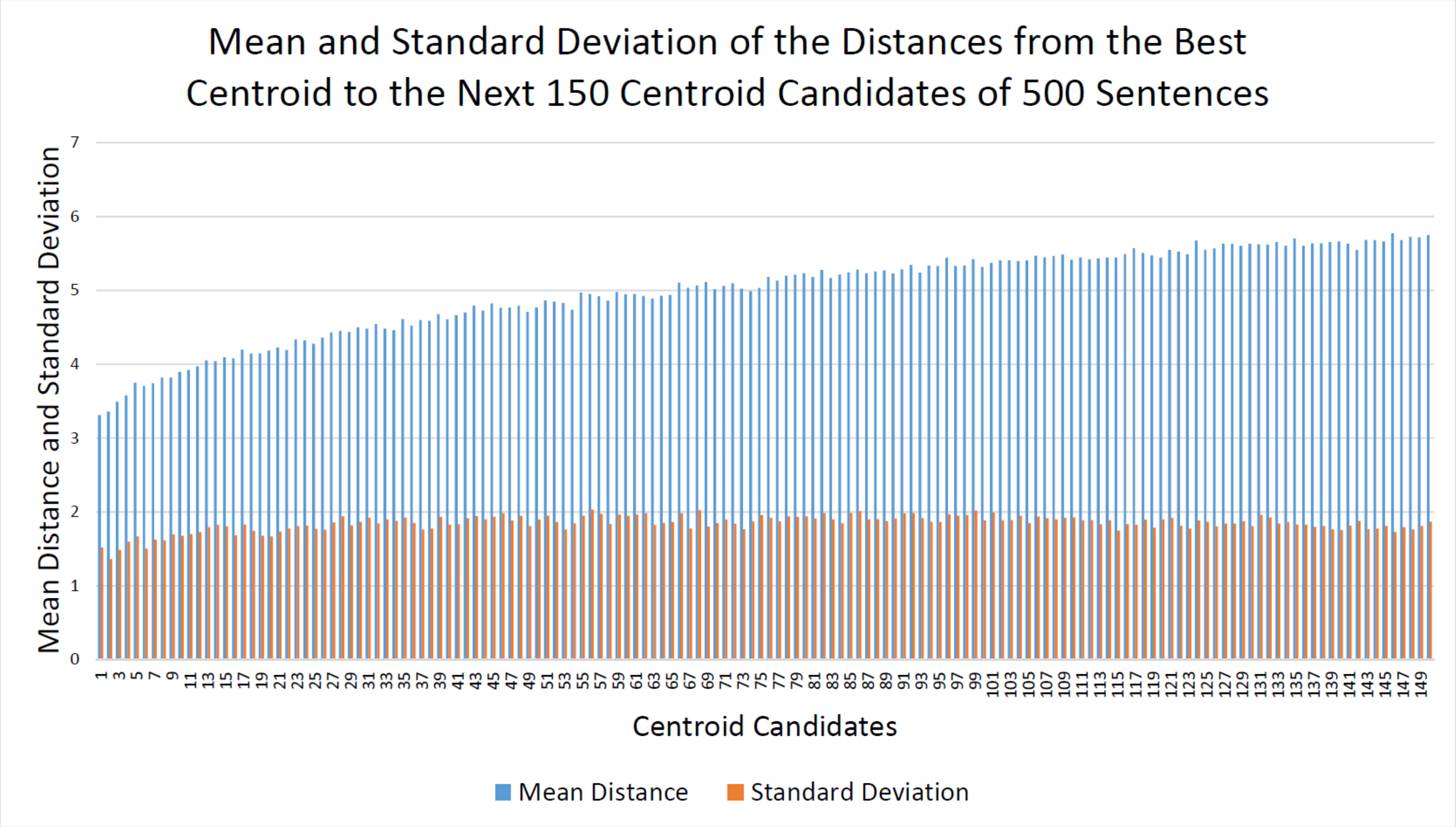
- ✓ Changes are considered when new documents are added.
- ✓ The importance of words (i.e. their rank in a majority list) hardly varies for the most frequently occurring ones.
- ✓ Calculated centroids do not rapidly change their positions, and are also stable after a small number of documents (<100) read (computed).

4. Hierarchies



- Centroids allow to investigate text structures.
- Text structures may be another criterion to compare texts.

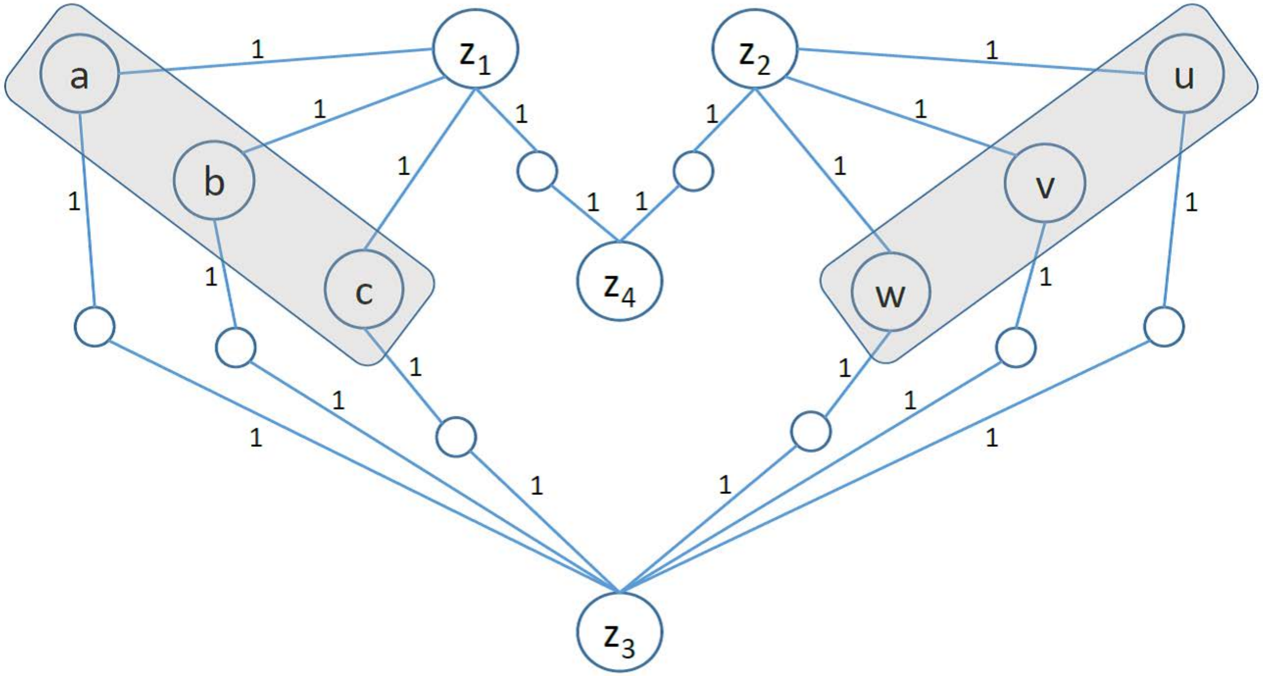
5. Uniqueness



- Although formally possible, it was never observed that two centroids could be found for a document.

6. Combination: Centroid of Centroids

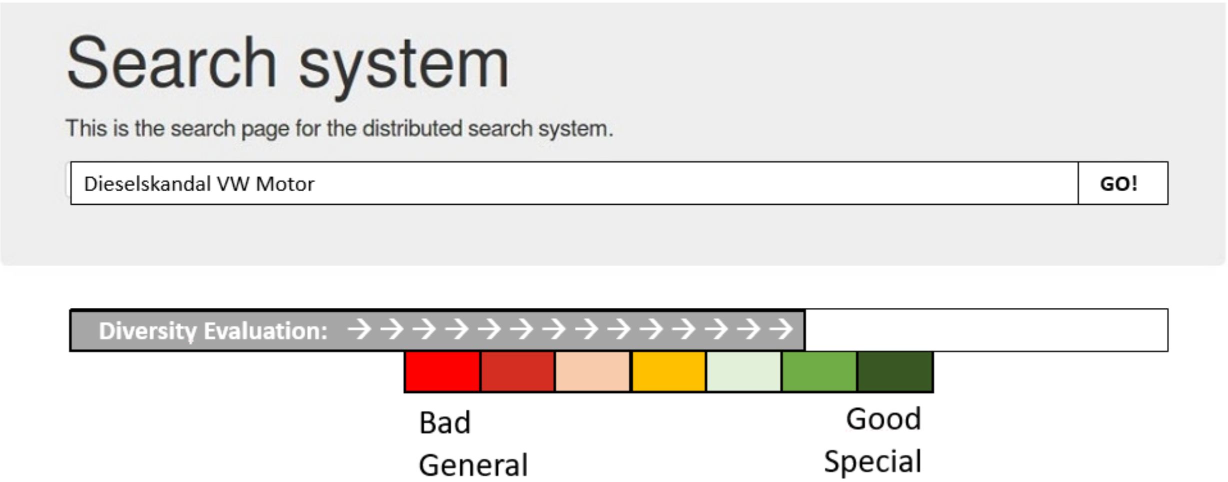
- Let A and B be two documents with the centroids $\chi(A)$ and $\chi(B)$
- Is $\chi(\chi(A) + \chi(B)) = \chi(A + B)$, i.e. can the centroid of two documents be calculated from their centroids only?



- The answer is no, but very often it works, or the distance is not significant.

7. Diversity

- ✓ is a property of the set of words used to calculate a centroid.
- ✓ is defined as the maximum distance of any pair of words of the query or text in the respective co-occurrence graph.
- ✓ The smaller the diversity is, the more a query targets a designated, tight topic area, while high values of the diversity mark a more general, common request.
- ✓ useful in…

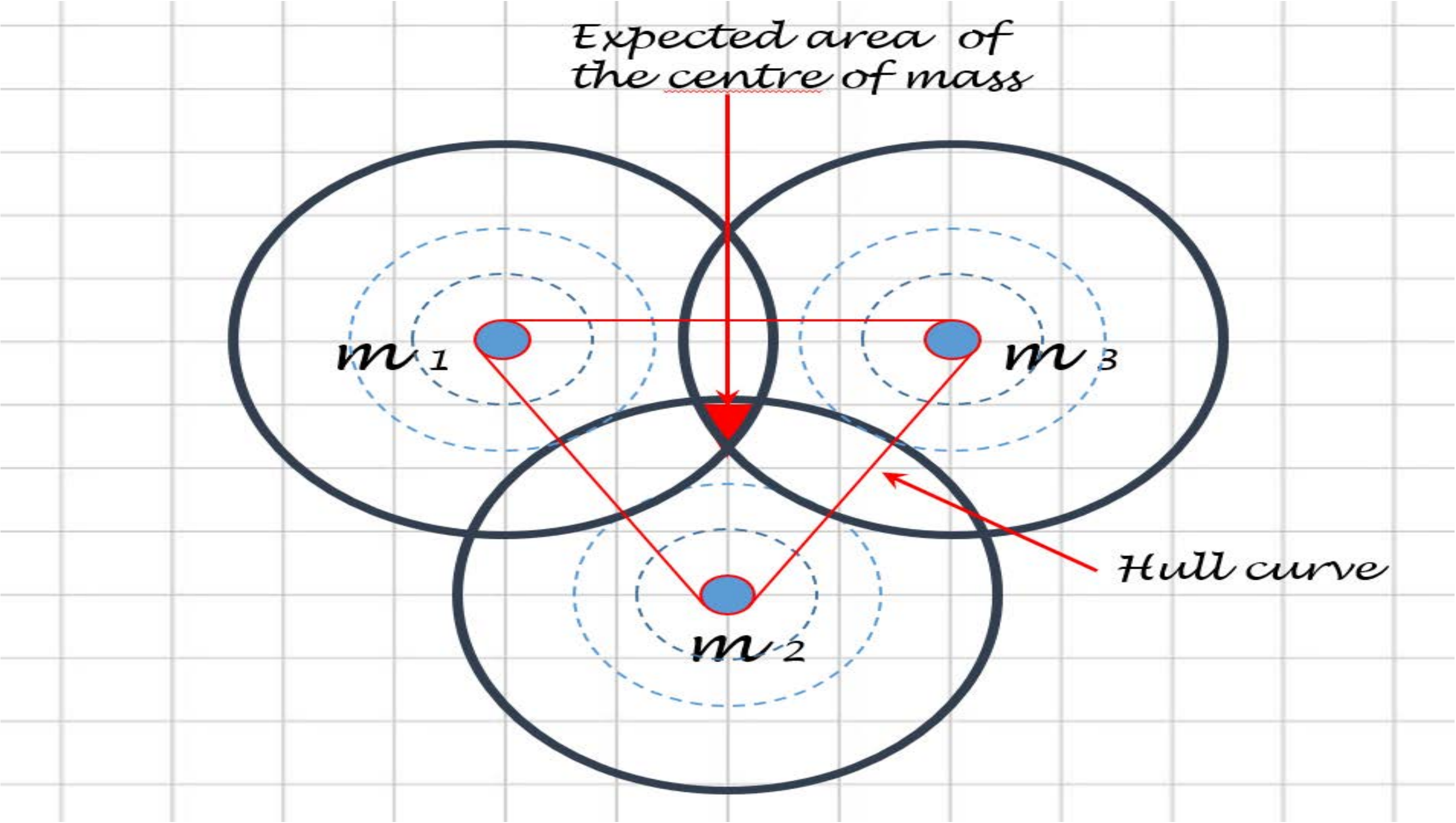


Search system

This is the search page for the distributed search system.

Dieselskandal VW Motor

8. Fast Calculation. Remember Physics!



→ The centre of mass is always to be found within the convex hull curve.

→ … also in the discrete case !

Algorithm on Co-occurrence Graph

I. Determine the diversity Δ of all query terms, i.e maximum distance between any two of th

II. Activate all nodes within a di with r= ½ Δ + ε.

III. If there is all query ter goto II.

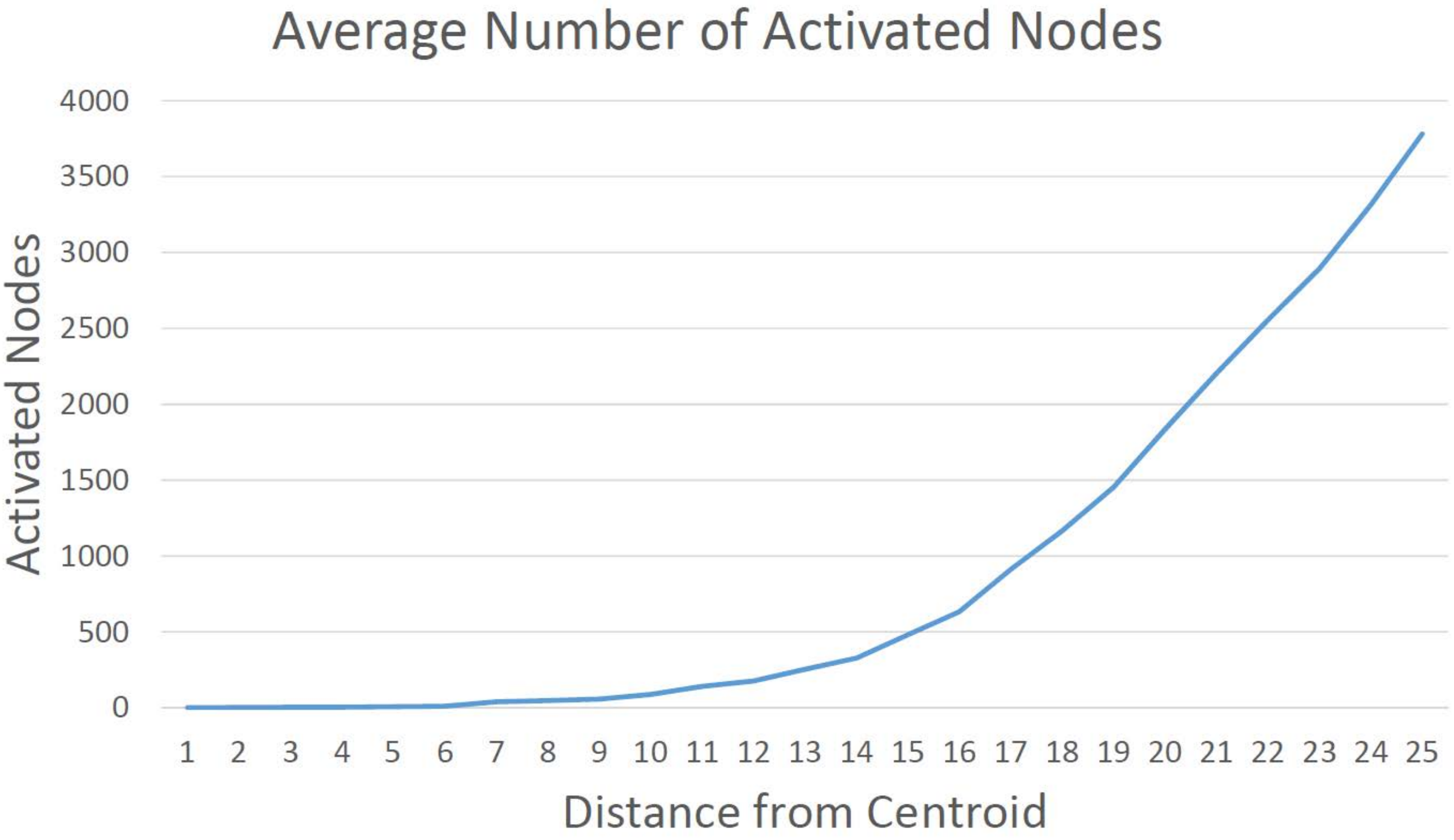
term among the activated

has been activated by all query terms **and**

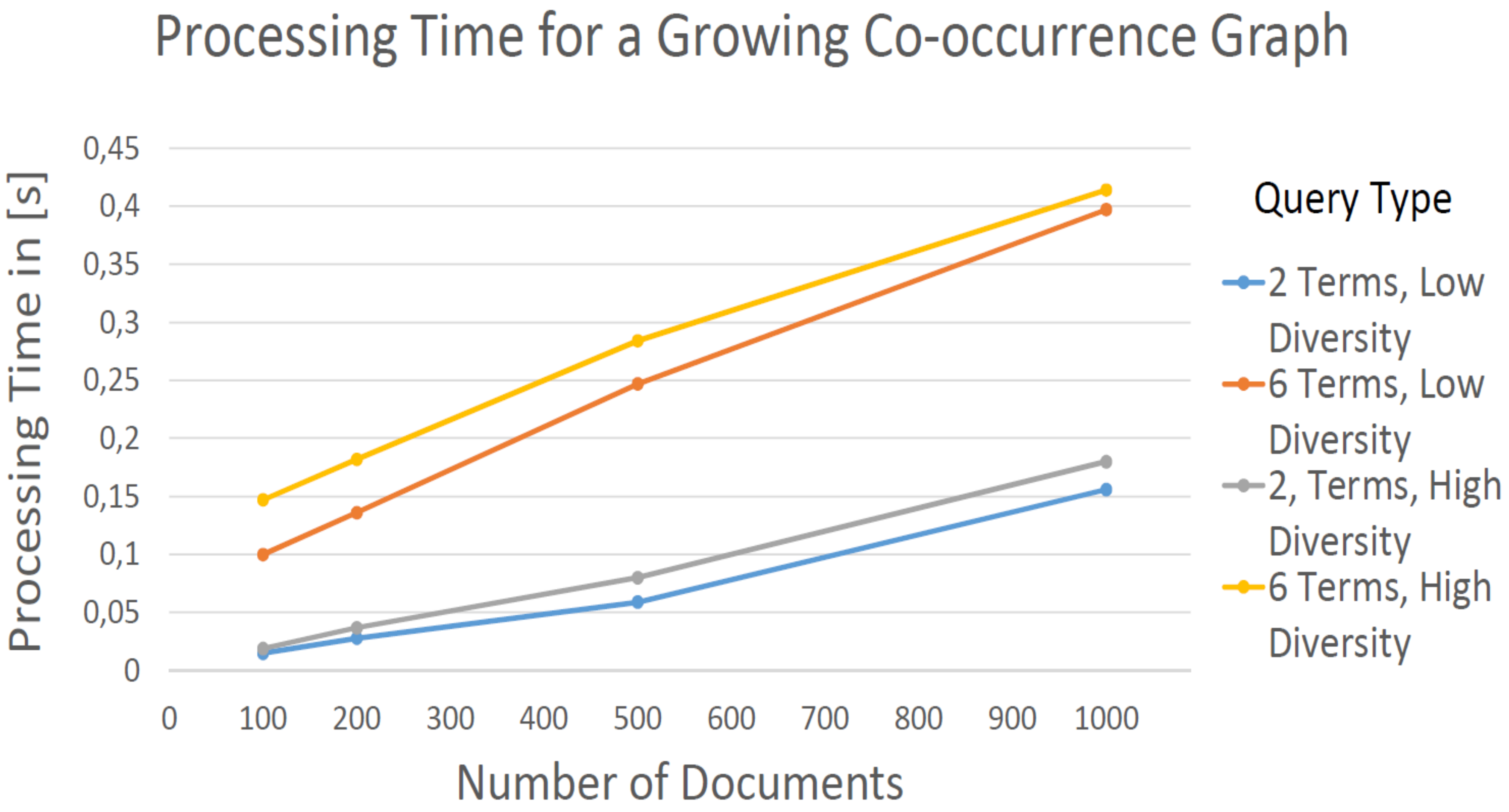
➤ has the lowest average distance to all of them.

The algorithm works locally on a limited subgraph of the giant co-occurrence graph!

Simulation Results 1



Simulation Results 2



Dynamic Centroids

Reading Process

- None of the known methods is able to ‘read’ and consider documents as ordered sequences of words.
- However, different sequences may significantly determine a text's contents and meaning - as well as its quality e.g.

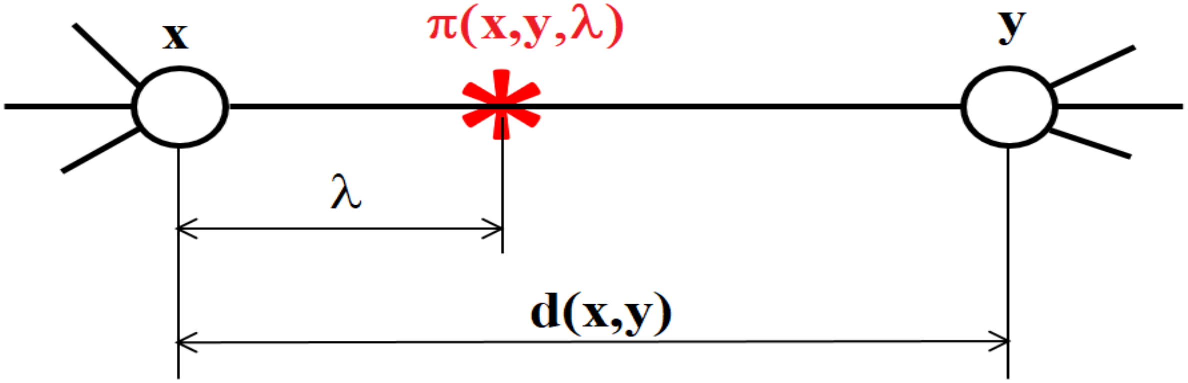
  THE BEAUTIFUL LADY ALWAYS WEARS UGLY DRESSES.

  THE UGLY LADY ALWAYS WEARS BEAUTIFUL DRESSES.
- Texts are usually categorised by human thinking depending on
  - → the already existing general knowledge,
  - → the sequence of words read.

Positions

- ✓ So far a centroid is associated with a node of the co-occurrence graph.
- ✓ This limits the numbers of categories significantly, and may make classification difficult.
  e.g.: ~~THIS IS SOME~~ TOILET PAPER.

→ **A generalisation is needed. Positions π.**



Algorithm

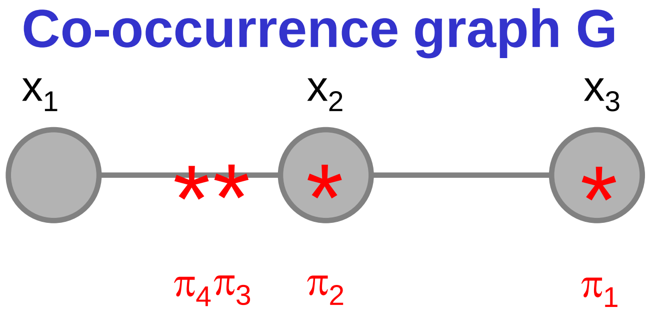
I. Take a connected co-occurrence graph G and a document D={w_1,w_2,..w_n} (some side condition applies)
Set i=1 and the first centroid χ_i = $w_{1.}$

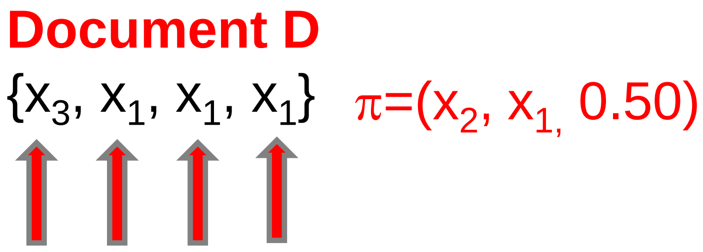
II. Consider w_{i+1}.
Determine the shortest path P between χ_i and w_{i+1}.

III. Find a new position χ_{i+1} on P such that χ_{i+1} partitions P starting from χ_i by the ratio 1 / i.

IV. Increase i:=i+1.

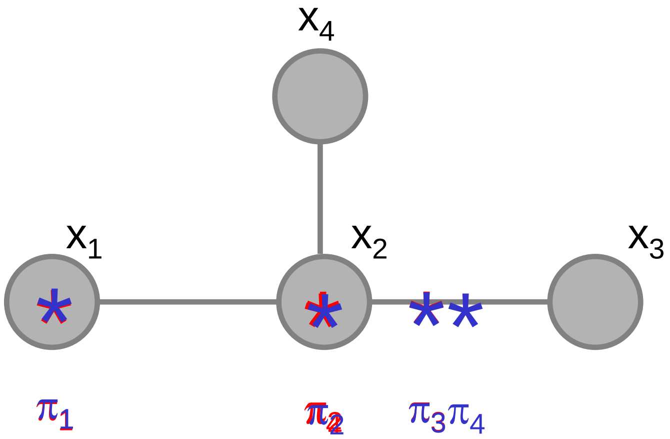
V. GoTo II, while i<n, otherwise STOP.





Another Example

Co-occurrence graph G



Document D

$\{x_1, x_3, x_3, x_4\}$

Document D

$\{x_1, x_4, x_3, x_3\}$

→ DIFFERENT SEQUENCE, ONLY

Centroid-Trace

$(x_1,x_1,0)$ → $(x_2,x_2,0)$ → $(x_2,x_3,1/3)$

→ $(x_2,x_2,0)$

≠

Centroid-Trace

$(x_1,x_1,0)$ → $(x_2,x_2,0)$ → $(x_2,x_3,1/3)$

→ $(x_2,x_3,1/2)$

Experimental Results

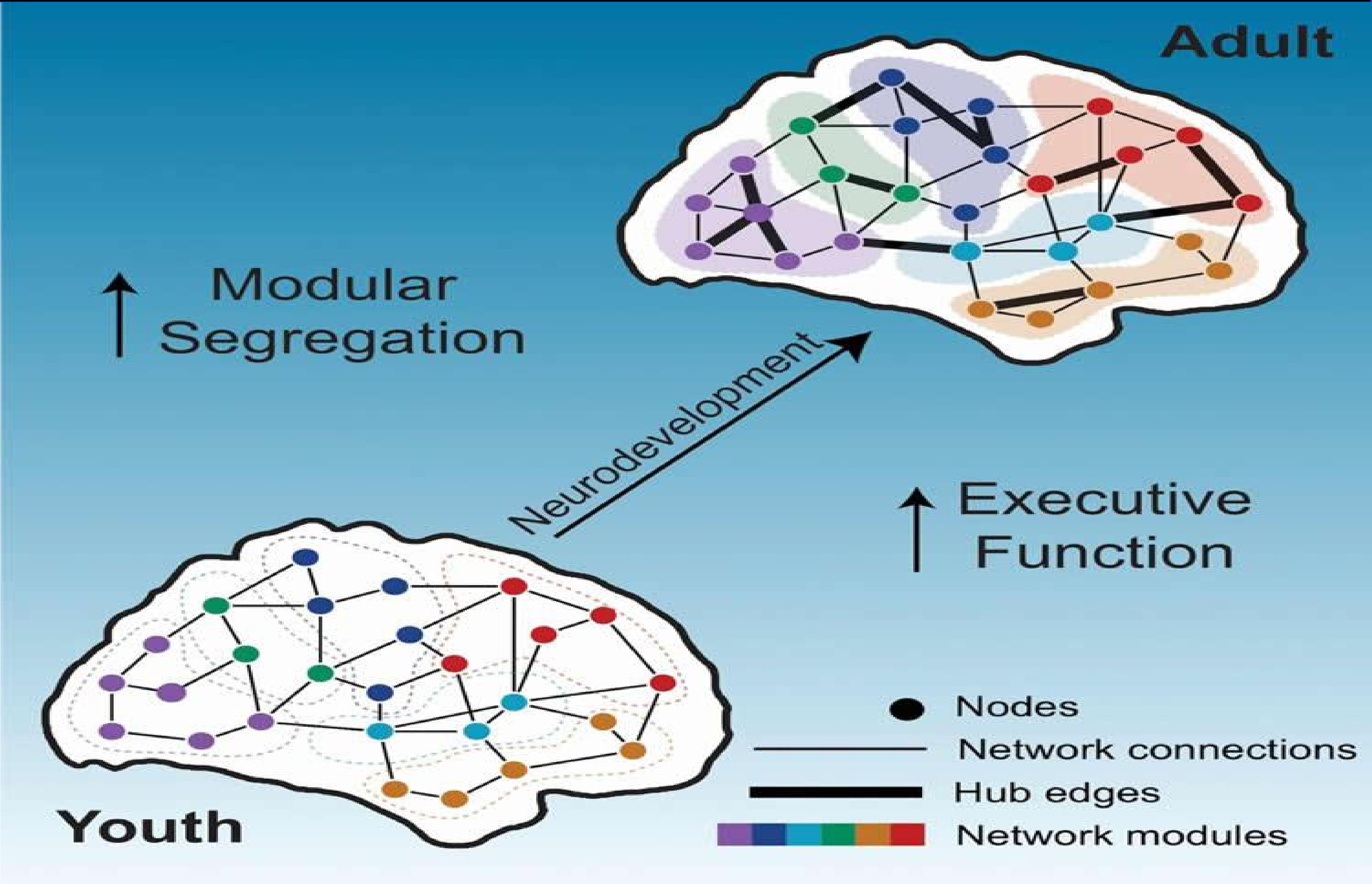
- Wikipedia article **'Fermi paradox'**:
  → classic centroid: civilization

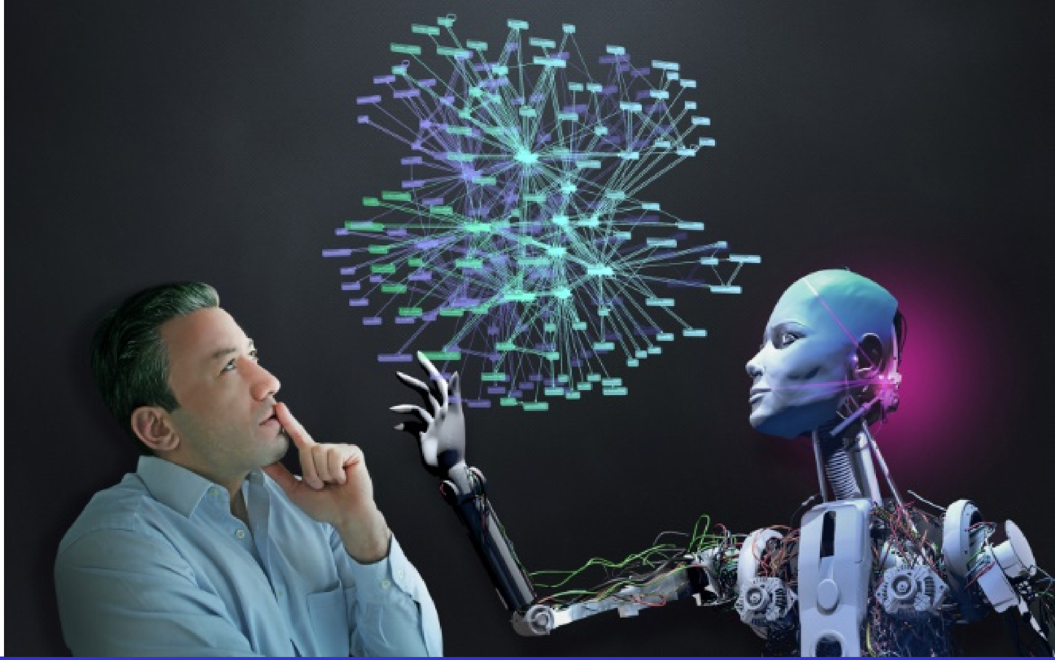
- **Dynamic Centroid Trail:**

```
[[Fermi, Fermi, 0.0], [Fermi, paradox, 0.78], [Fermi,
paradox, 0.33], [Fermi, paradox, 0.71], [paradox,
civilization, 1.99], [paradox, civilization, 0.73],
[paradox, civilization, 0.45], [paradox, civilization,
2.04], [paradox, civilization, 0.14], [paradox, argument,
1.51], [argument, star, 0.10], [argument, paradox, 1.97],
[argument, paradox, 2.55], [argument, paradox, 2.97],
[paradox, civilization, 1.18], [paradox, civilization,
0.01], [paradox, artifact, 1.26], [paradox, emission,
0.75], [paradox, emission, 0.33], [paradox, life, 0.68],
[paradox, life, 0.30], [paradox, life, 0.23], [paradox,
evidence, 0.55], [paradox, civilization, 0.54]]
```

→ Centroid trails as - document fingerprints?
- chaotic systems?

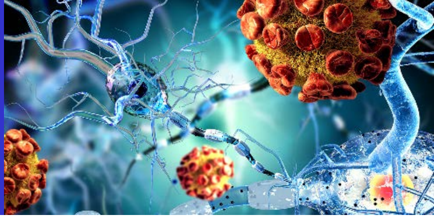
Summary: Deep learning?

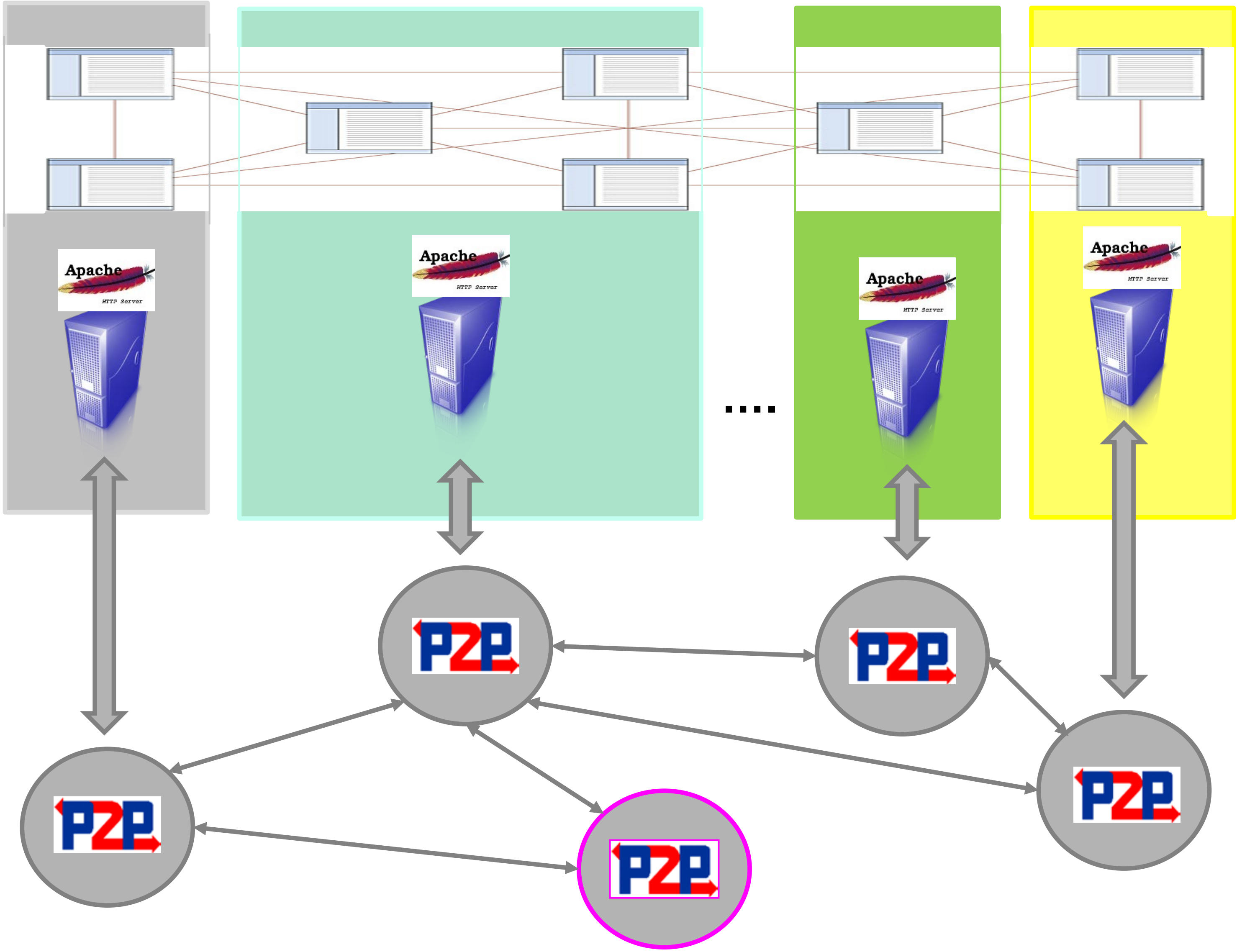




Application: The Librarian of the Web

Decentralised Search Engines (see also YaCy and Faroo)





WWW Pages

+

Webserver in Internet

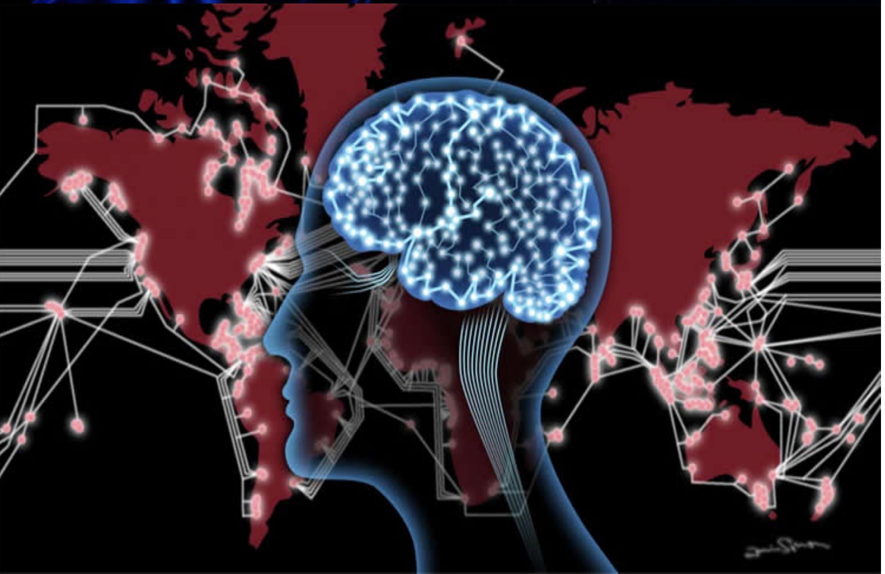
+

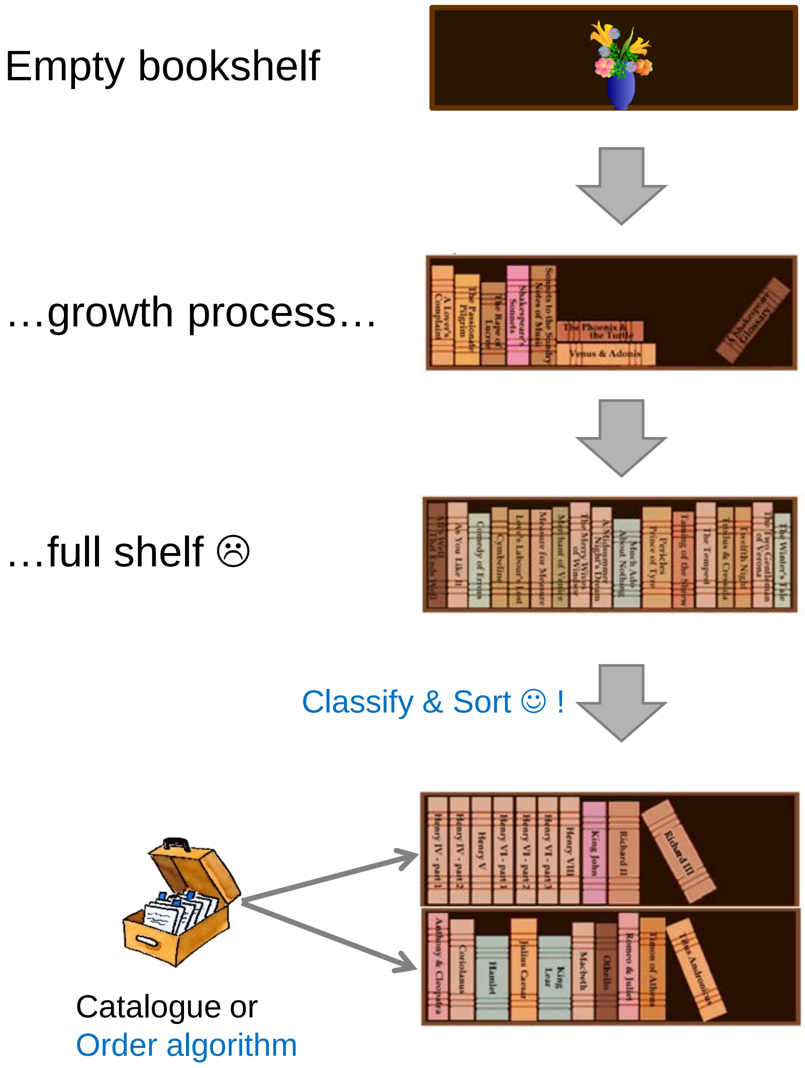
P2P-Level

===

NEW SYSTEM

The Librarian of the Web





Empty bookshelf

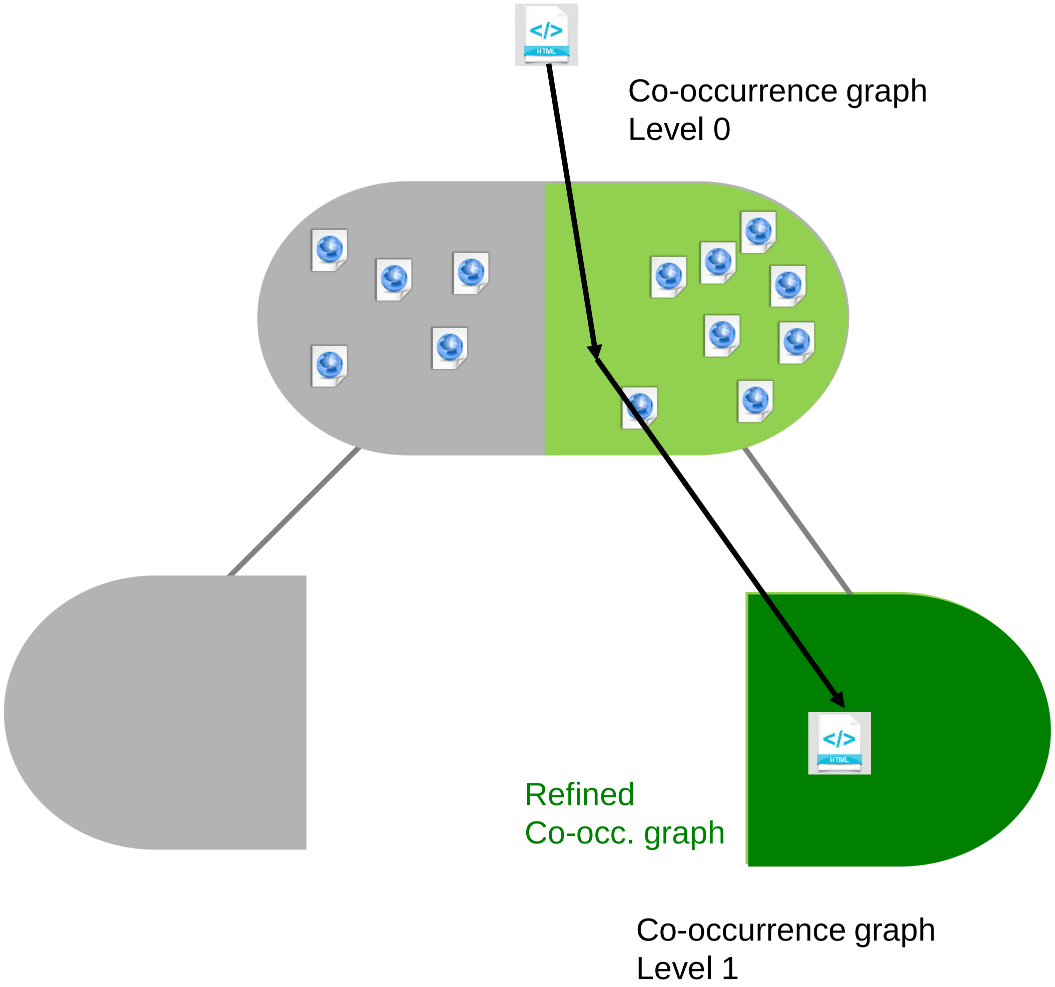
…growth process…

…full shelf ☹

Classify & Sort ☺ !

Catalogue or
Order algorithm

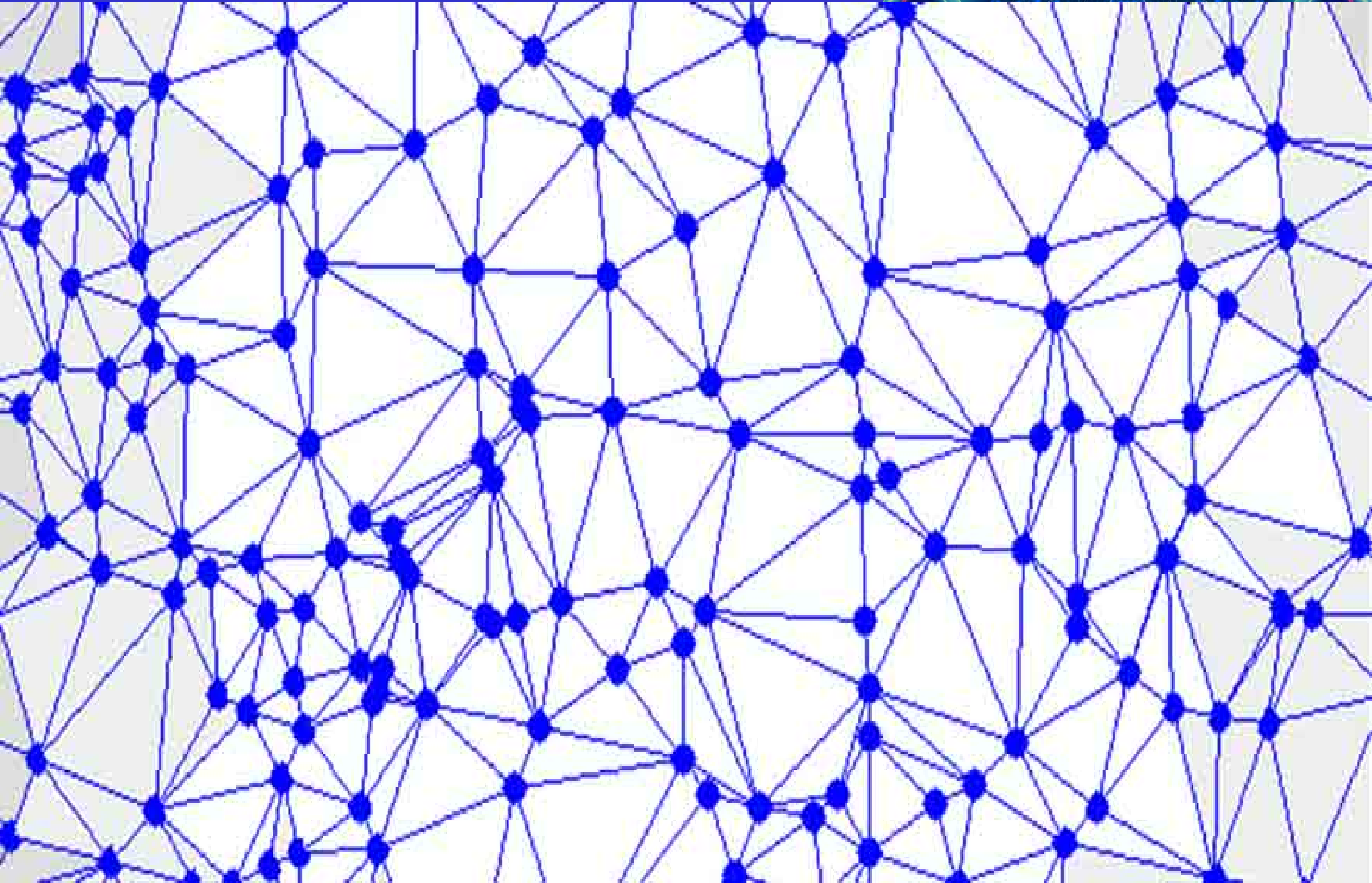
Top-down: Building a Self-specialising Hierarchy



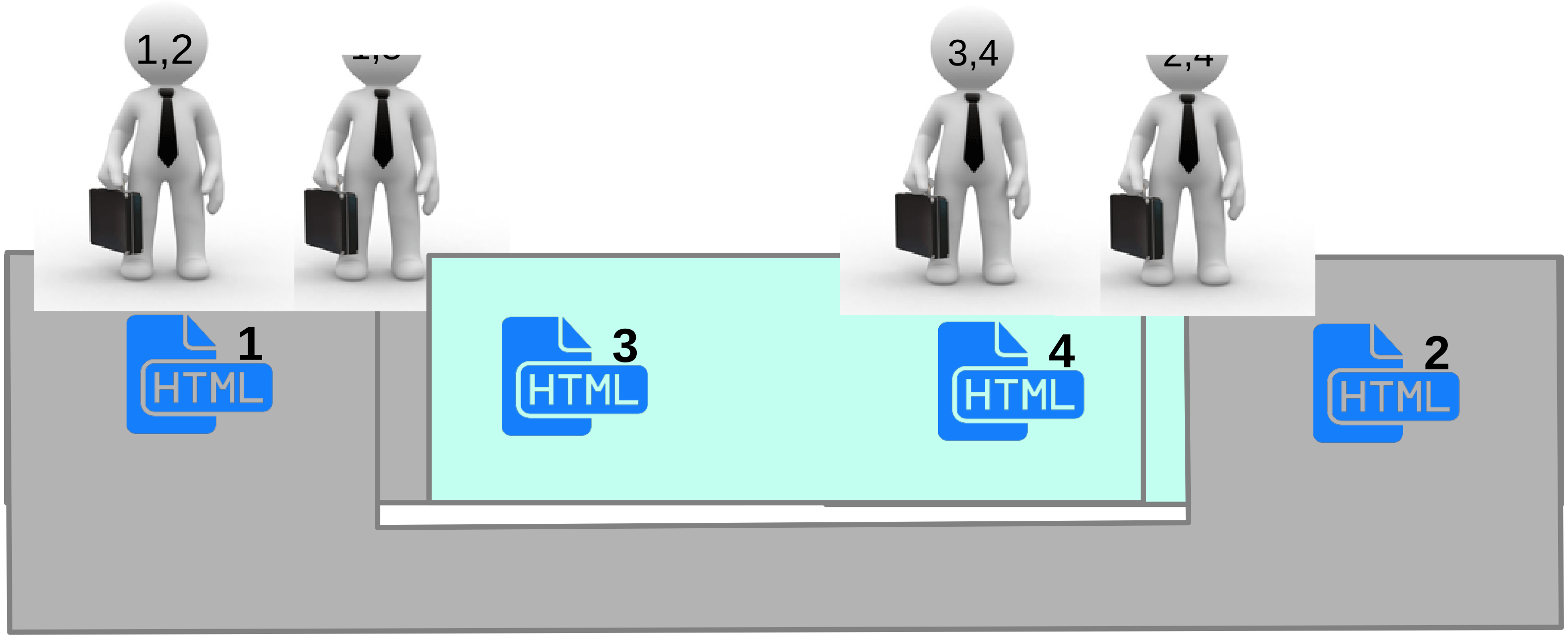
Rules of the game

- ✓ If a level is full, the local co-occ. graph is partitioned.
- ✓ Document links are moved to one node of the lower level depending on the location of their centroids. (some words of a document may be in the other partition, however)
- ✓ The upper levels persist as a chunky classification of newly arriving documents or queries which are later refined.
- ✓ The co-occ. graph in the lower level will be refined by documents assigned to the respective node.
- ✓ In case the next node is full, the game is repeated in a successive manner.

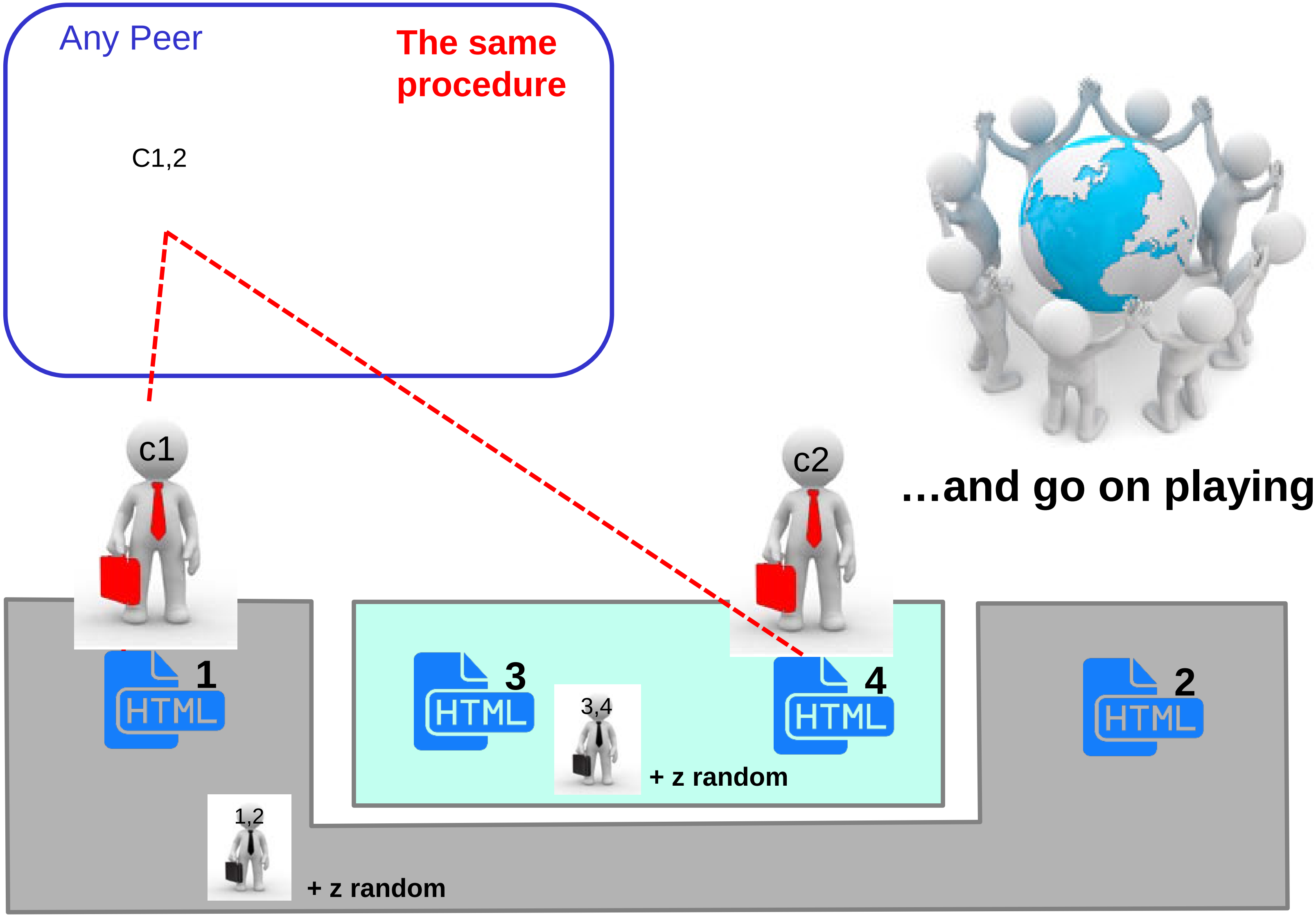
Bottom-up: Agent Game



Bottom-up: Agent Game



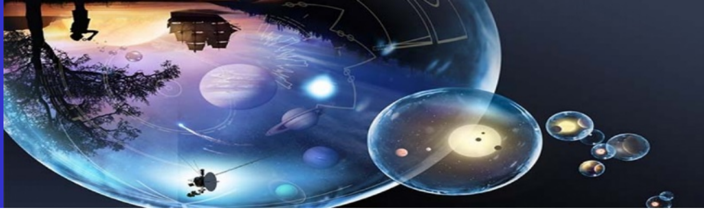
Bottom-up: Agent Game



The same procedure

…and go on playing

Properties of the Agent Game



- New peers will automatically be included.
  If needed, new agents and peers will be added.

- Peers leaving the community will be tolerated.

- Agent faults are no problem.
  A lost agent may be replaced and included
  without any bigger problem to the remaining community.

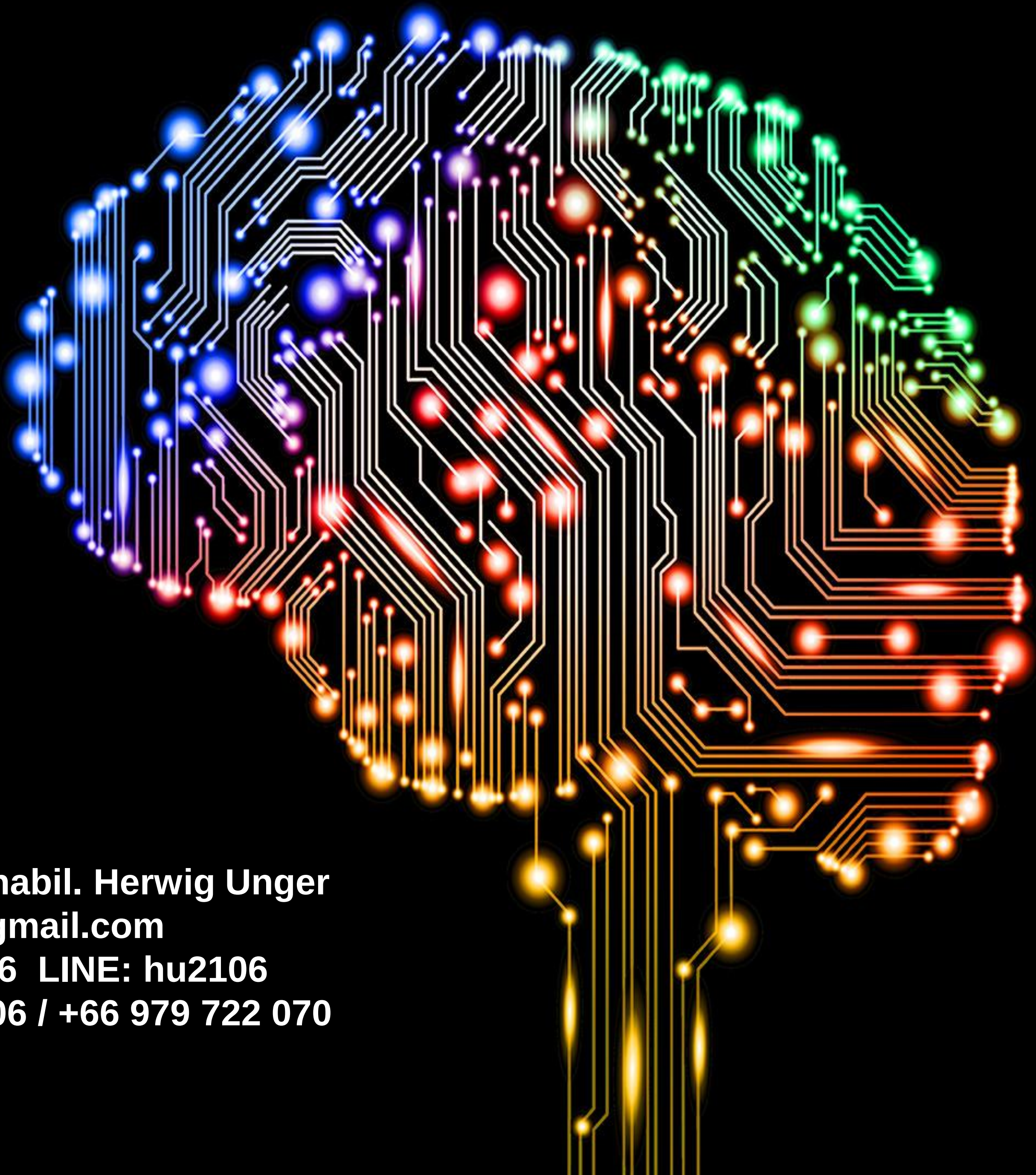
- Fully connected clusters make the system more fault-tolerant.
  Also, several peers may fulfil the task as surrogate of the whole (local) sub-cluster, increasing fault tolerance even more.

- The structure size automatically adapts to changing needs.

- Search requests may be routed – even if not arriving at the root node – within predictable time.

Summary

- ✓ Today text analysis and classification are two major problems in NLP.
- ✓ Ontologies, statistic methods, annotation based methods and semantical analyses often fail.
- ✓ The centroid approach is a formal classification method neglecting human meanings.
- ✓ Dynamic centroids mimic human reading and understanding, …
- ✓ … thus showing similarities with the work of the human brain.
- ✓ Our future work will be to investigate this in more detail.



Contact:

→ Prof. Dr.-Ing. habil. Herwig Unger
Herwig.Unger@gmail.com
WeChat: pdu1966 LINE: hu2106
+49 176 8183 2106 / +66 979 722 070