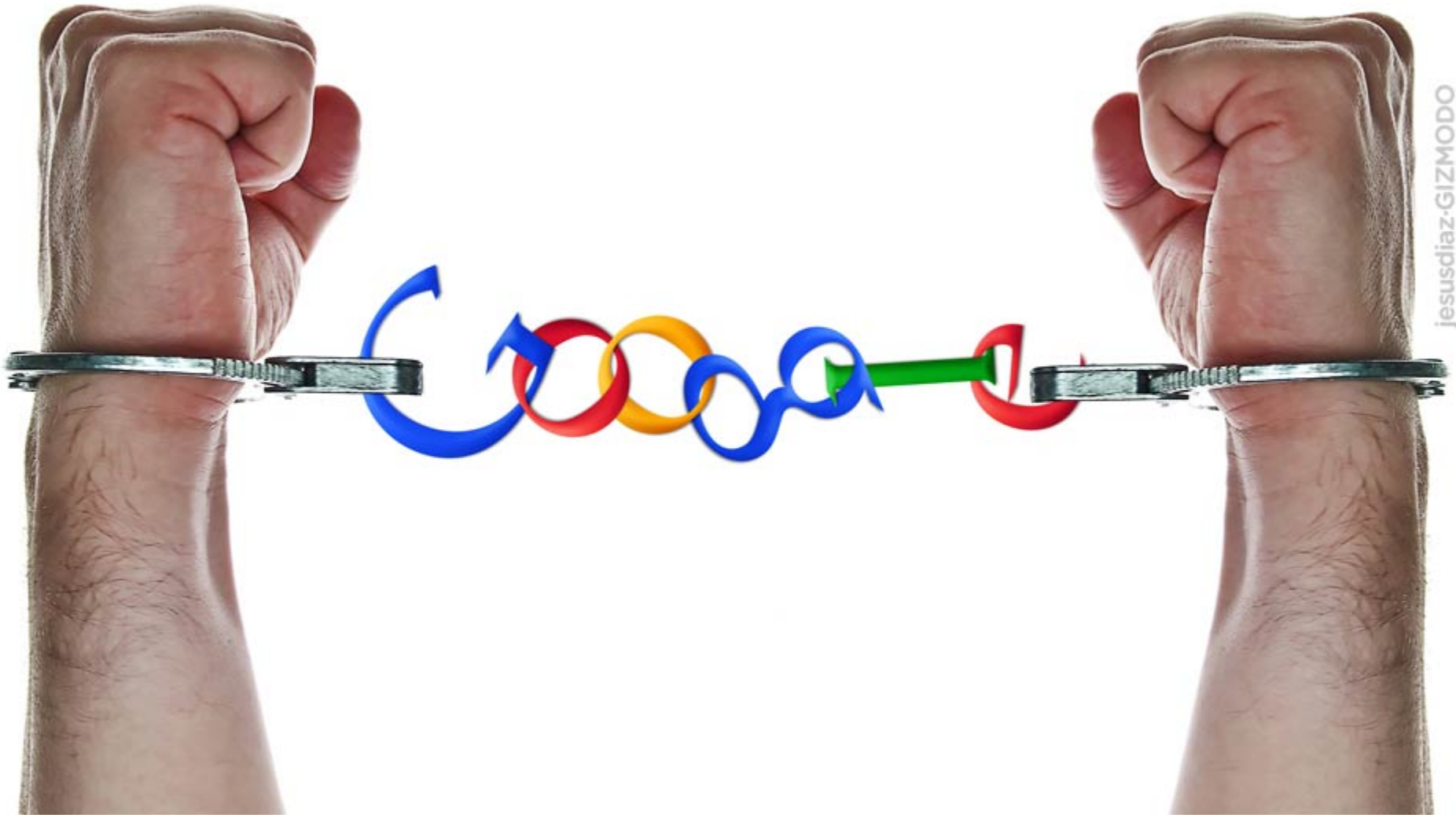
THE FUTURE OF INTERNET SEARCH

Herwig Unger

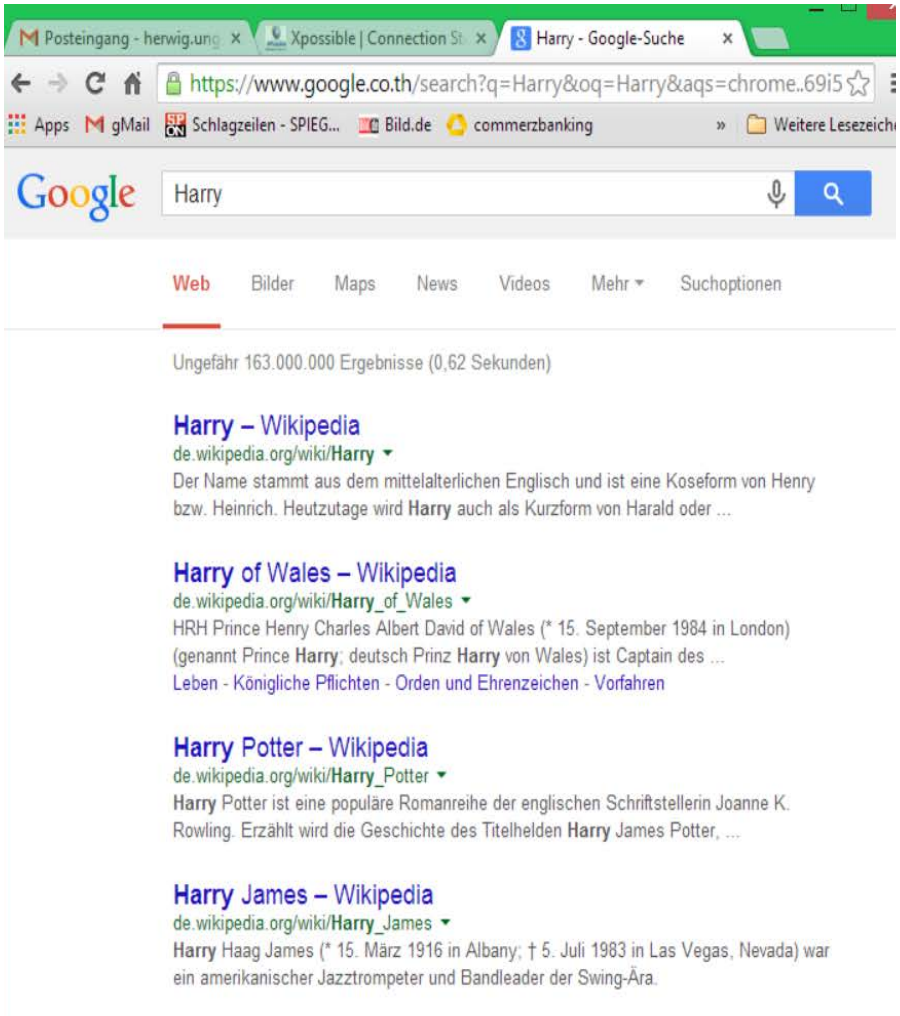
The perfect world of google…???



MI-6, KGB and Stasi were yesterday.

Zombies everywhere!

A more direct comparison



But search is ….

- … an iterative process
- … need consider many alternatives
- … having its own (very personal) context and history
- … coming along with learning effects derived from positive and negative feedbacks
- … also an influence to the objects being searched and retrieved
- … possibly a cooperative activity
- … in the nature seldomly centralised ☺

(think about foraging, ants, partner search)


042
TESSERACT

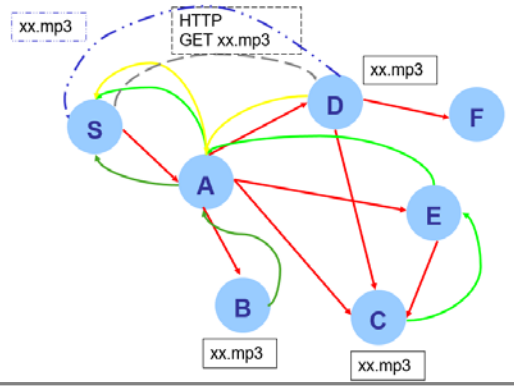
Today is Google. And tomorrow?

- Search is understood as process with several participants in a given context and a history
- Learning and adaptation, which are caused by multiple feedback sources
- Decentralisation and emergence of structures
- No central control or knowledge
- Brain like structures and processes where connections between different instances are the most important things

Jeff Hawkins: “On Intelligence”

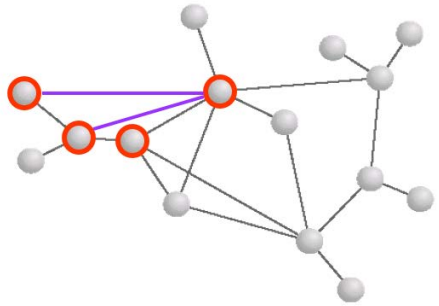
Going decentralised….

Alternatives



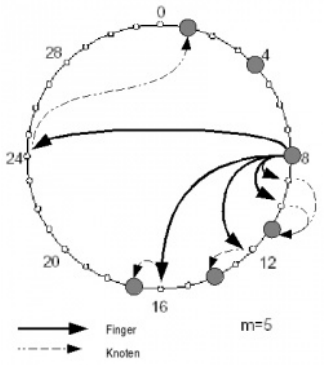
GNUTELLA

- Broadcast-based, i.e. many messages
- Keep anonymity of user until download
- Relatively fast
- Less overhead since simple protocol



FREENET

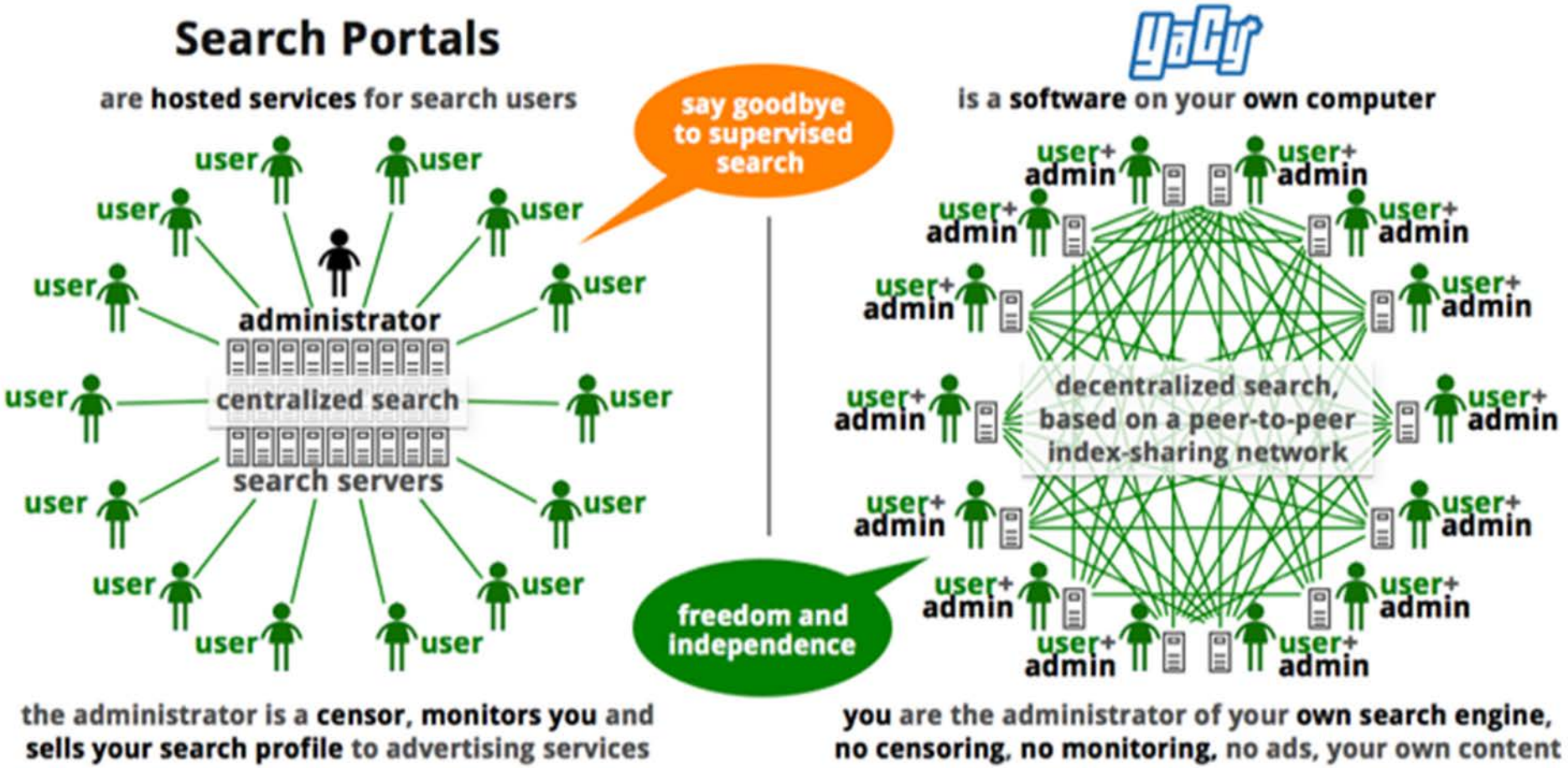
- No broadcasts, single search messages
- Keep anonymity fully
- Get faster over time by copies and new links
- Still a simple protocol



Dynamic Hash Tables (DHT)

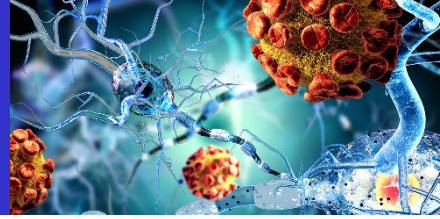
- Scalable with logarithmic expenses
- Just a content directory
- Fast
- Many overhead due to complex protocol

YaCy



Source: www.yaci.net

Decentralised search engines (see also YaCy and Faroo)



WWW
Pages

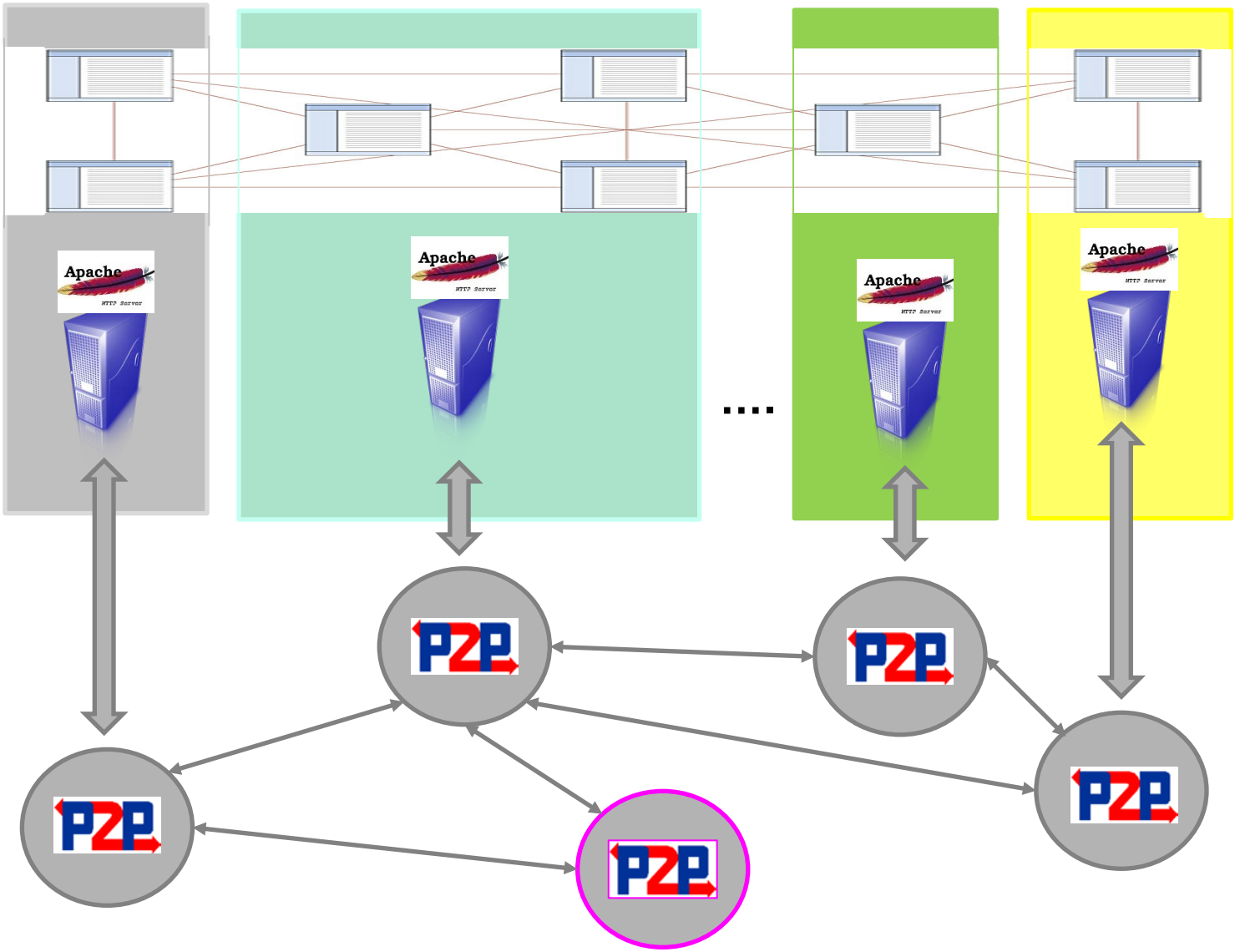
+

Webserver
Im Internet

+
P2P-Level

===

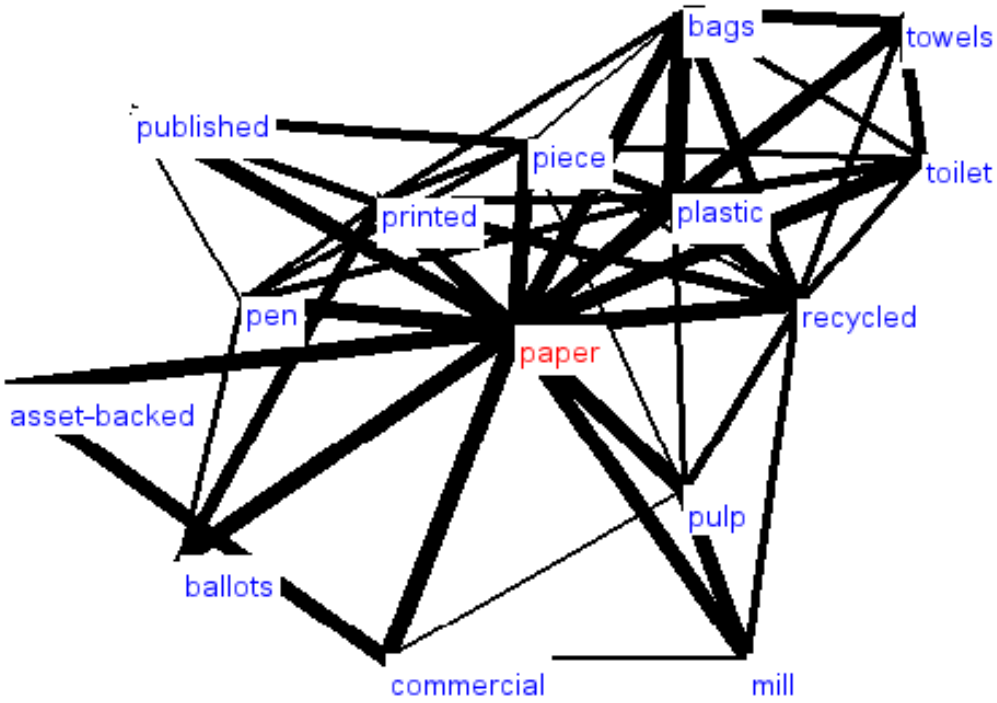
NEW
SYSTEM



“Our” Preliminaries…

The basics: co-occurrence analysis

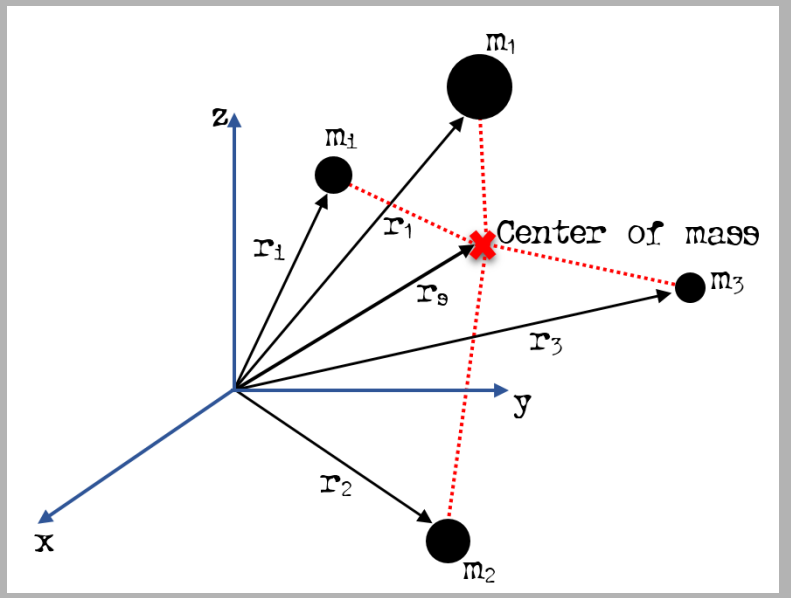
- Significant co-occurrences appear with a probability above a specific threshold in sentences (sentence level), in paragraphs (paragraph level) or in the whole text (document level).
- The set of all significant co-occurrences can be presented as a co-occurrence graph (usually undirected): nodes-terms, edges-relations



Source: *corpora.uni-leipzig.de*

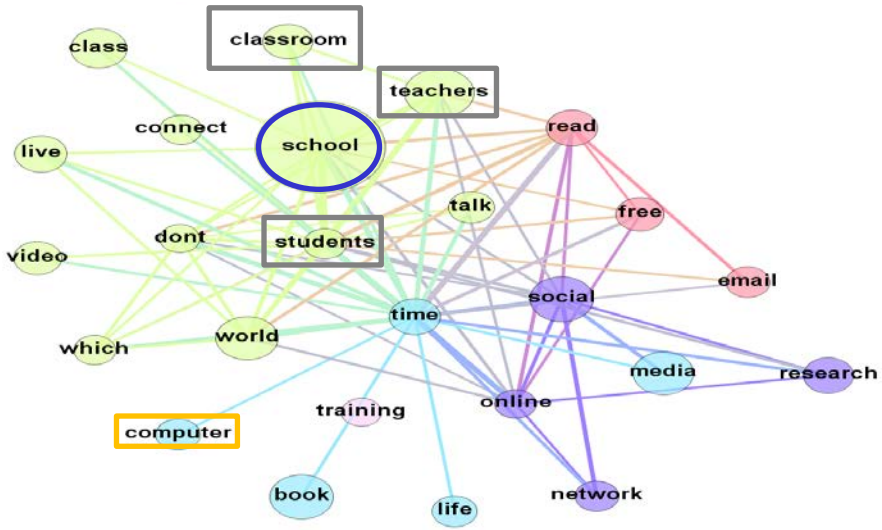
Document centroids

The physical analogon:
→ **the 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 minimal average distance to all words of the respective document in the co-occ. graph.

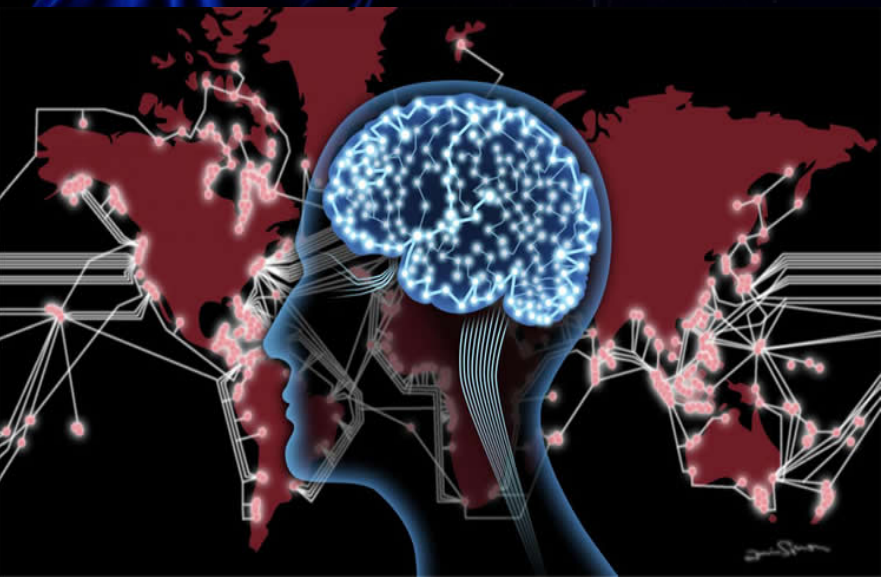
Properties of centroids

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

- ✓ The centroid can 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.
- ✓ The distance of two document centroids in the co-occurrence graph can be used to define the similarity of the documents.
- ✓ Even to short queries may a centroid term may be assigned.


Towards a
Librarian
of the
WWW

The librarian of the web

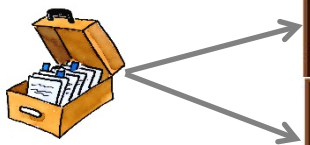


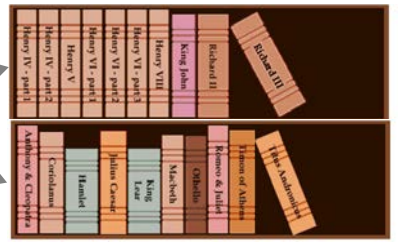
Empty bookshelf

…growth process…

…full shelf ☹

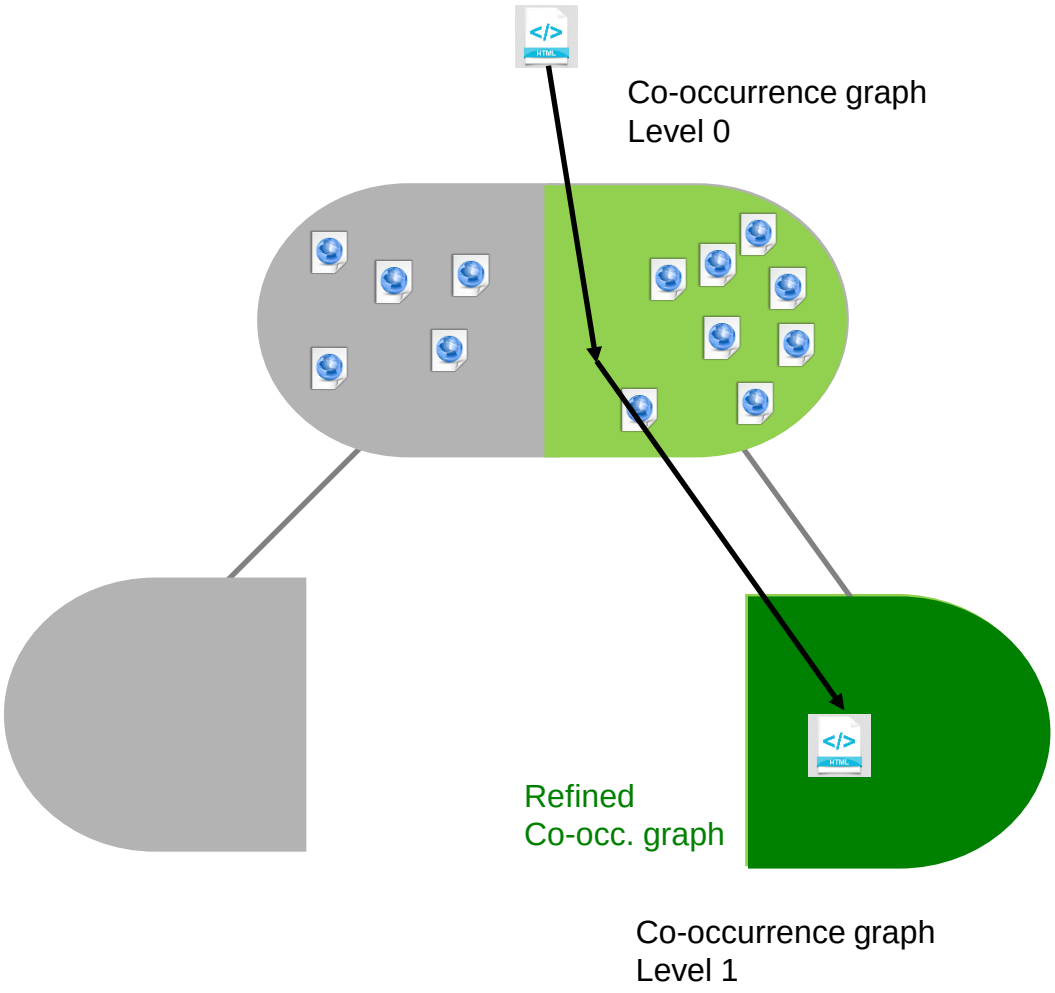
Classify & Sort ☺ !





Catalogue or
Order algorithm

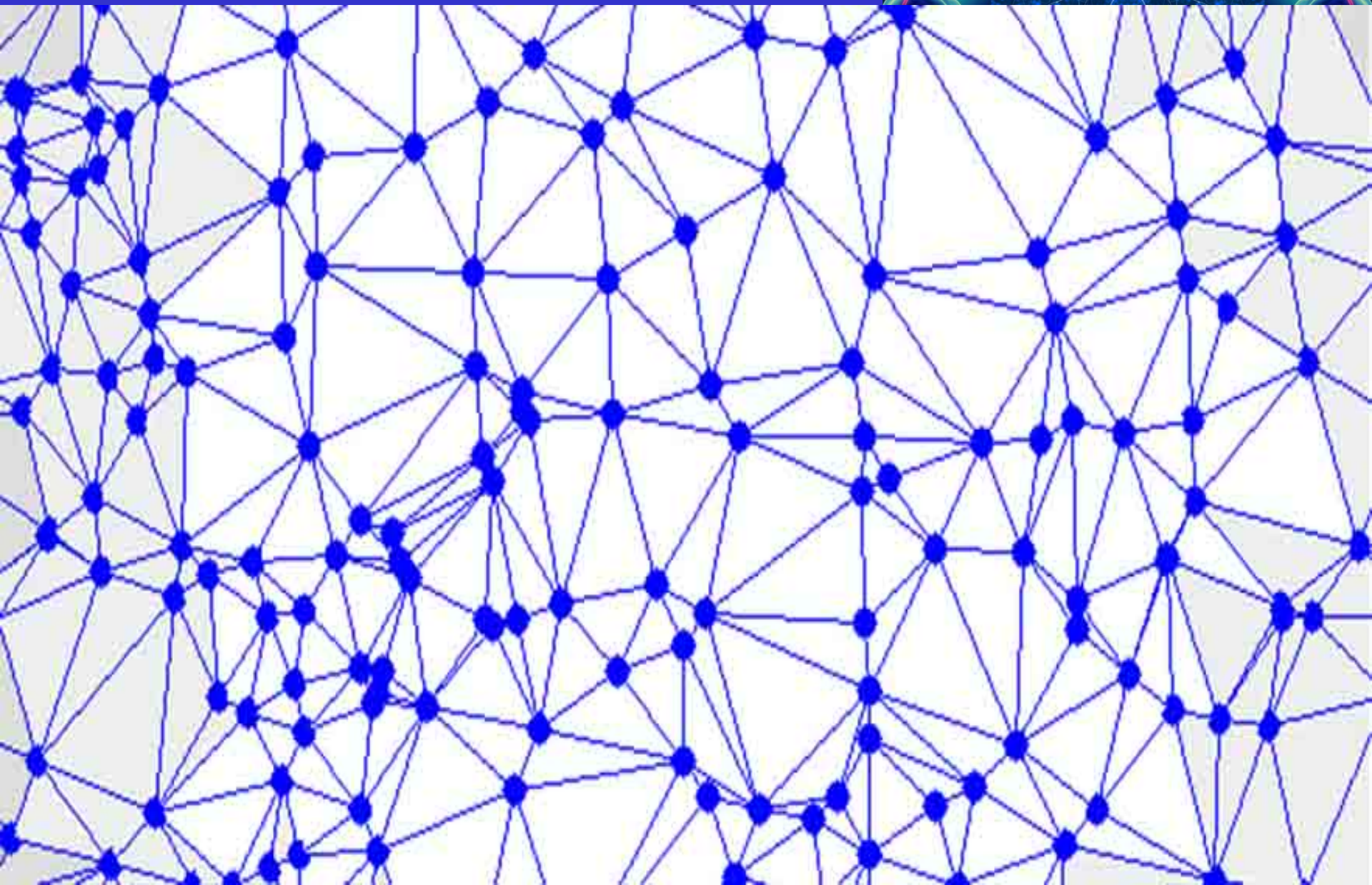
Top Down: Building a self-specialising hierarchy



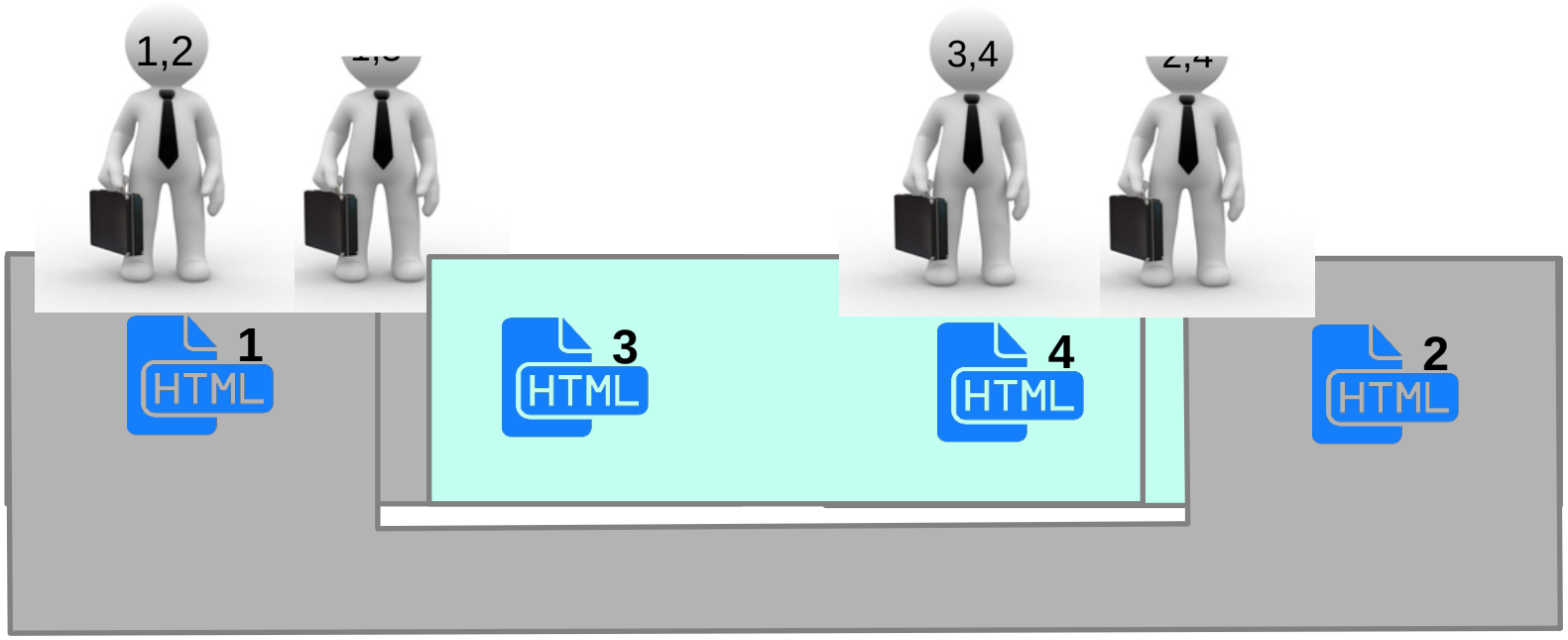
Rules of the game

- ✓ If a level is full, the local co-occ. graph is partitioned.
- ✓ Document links are given to one node of the lower level depending on the location of its centroids. (some words of a document may be in the other partition, however)
- ✓ The upper levels remain as a chunky classification of new 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.

Button-up: Agent play



Button-up: Agent play



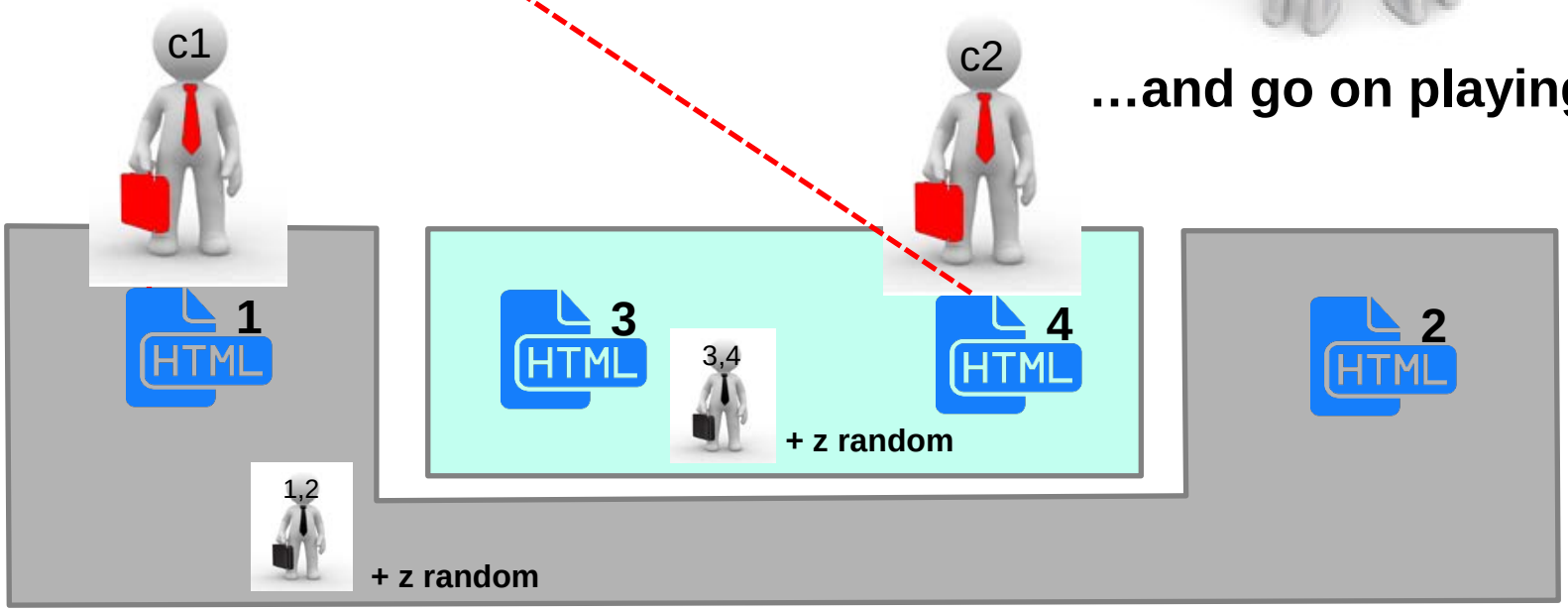
Button-up: Agent play

Any Peer

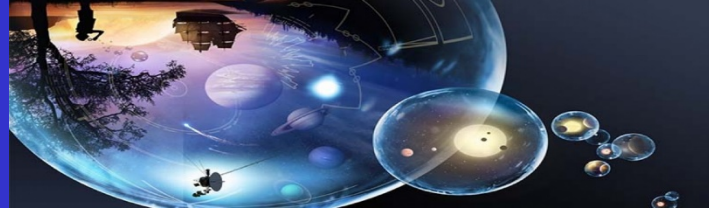
The same procedure

C1,2

…and go on playing

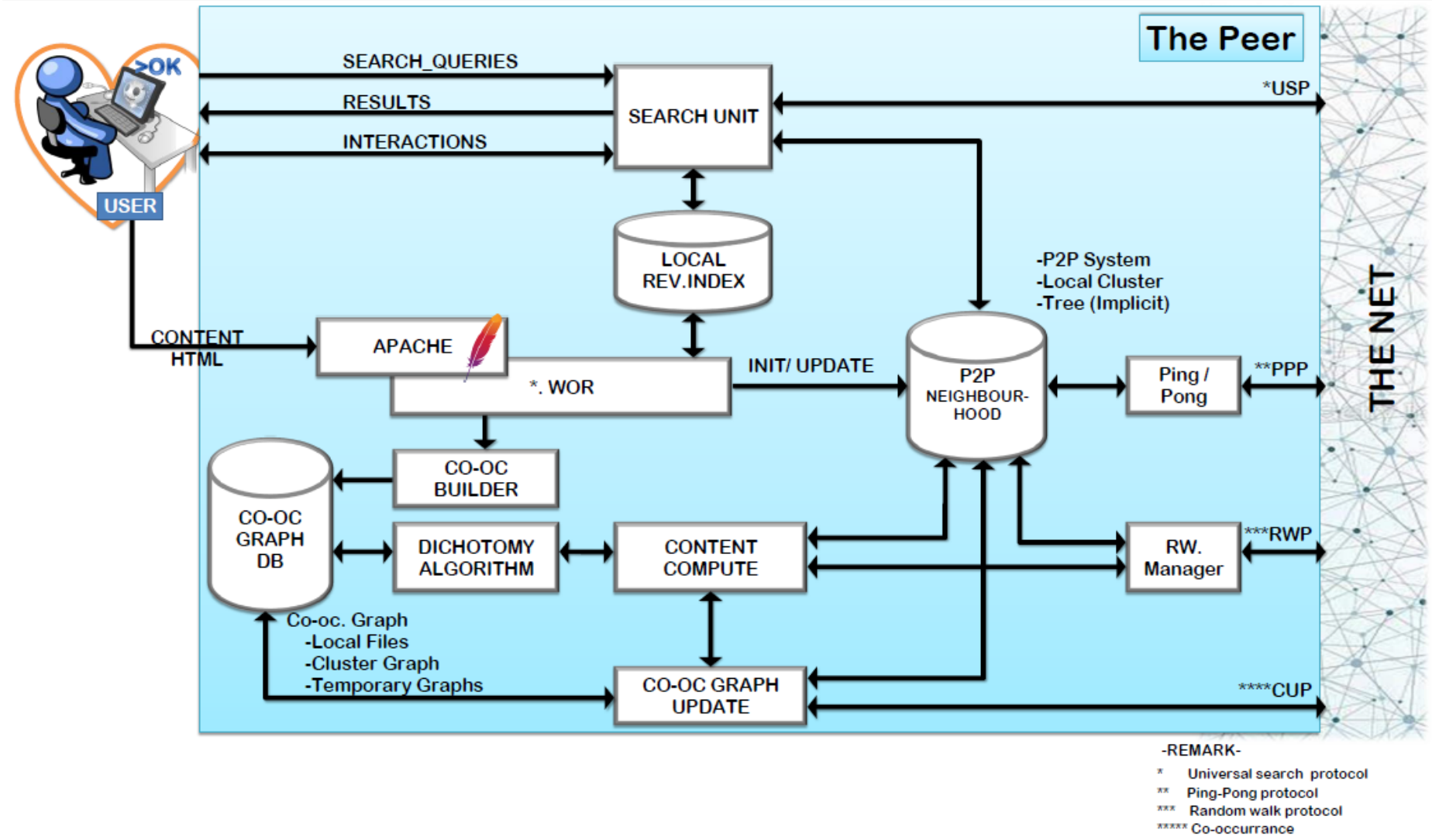


Properties of the agent play

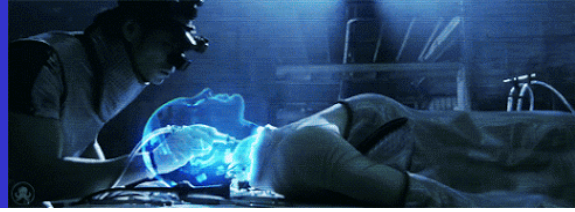


- New peers will be automatically included. If needed, new agents and peer will be added.
- Peers leaving the community will be tolerated.
- Agent faults are no problem. A lost agent maybe replaced and included without any bigger problem to the remaining community.
- Fully connected cluster makes the system more fault tolerant. Also, several peers may fulfil the task as surrogate of the whole (local) sub-cluster, what increase fault tolerance once more.
- The size of the structure automatically adapts to changing needs.
- Search request may be routed --even if not coming in to the root node– in a calculatable time.

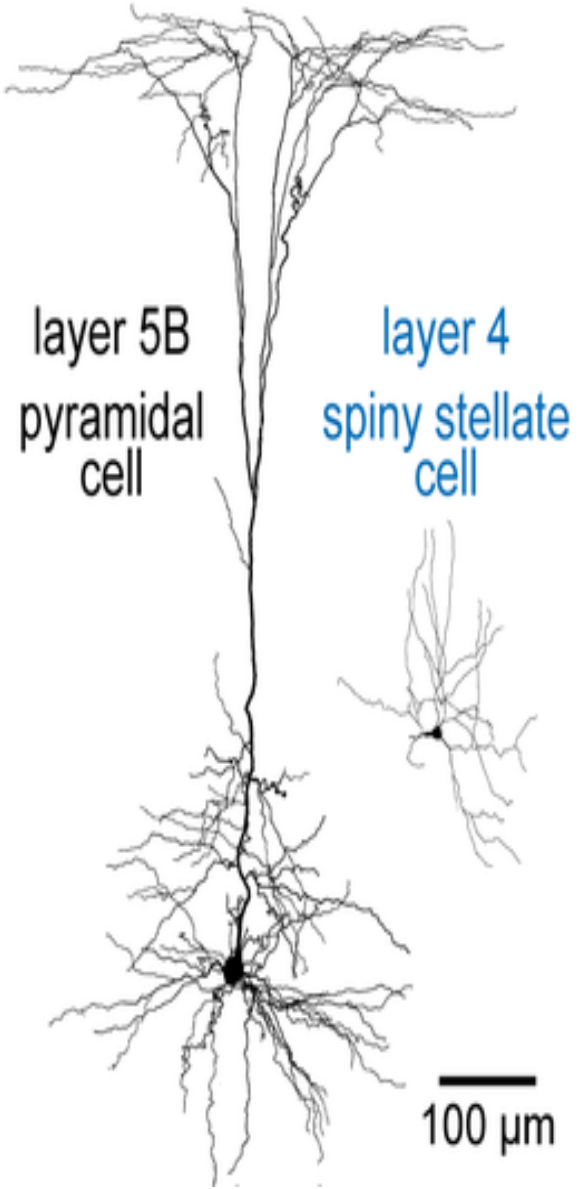
Peer architecture



Summary



- ✓ Today‘s search engines are far away to replace a librarians functionality.
- ✓ Small, decentralised systems are more flexible and competitive.
- ✓ Business models exist also for P2P.
- ✓ Copying the WWW is not a good approach, except for the NSA and secret services.
- ✓ An new, fully decentralised concept of search is investigated, offering new services, interfaces and ahigh degree of privacy. This approach is scalable and may adapt to changing needs in the WWW.
- ✓ It shows similarities to the work of the human brain. This must be considered more detailed in the future.



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