Collaboration Spotting: Big Data Visual analytics

A. Agocs, D. Dardanis, R. Forster, M. Gazzari, J.-M. Le Goff, X. Ouvrard, D. Proios
CERN

Wigner, Budapest, Hungary, 12 August 2016

Collaboration Spotting

- Vision
- Concepts
- Modus Operandi
- Targeted performances
- Computational needs & optimization
  - Graph DB management and operations (A. Agocs)
  - Interactive visual graph processing (R. Forster)
- Future work

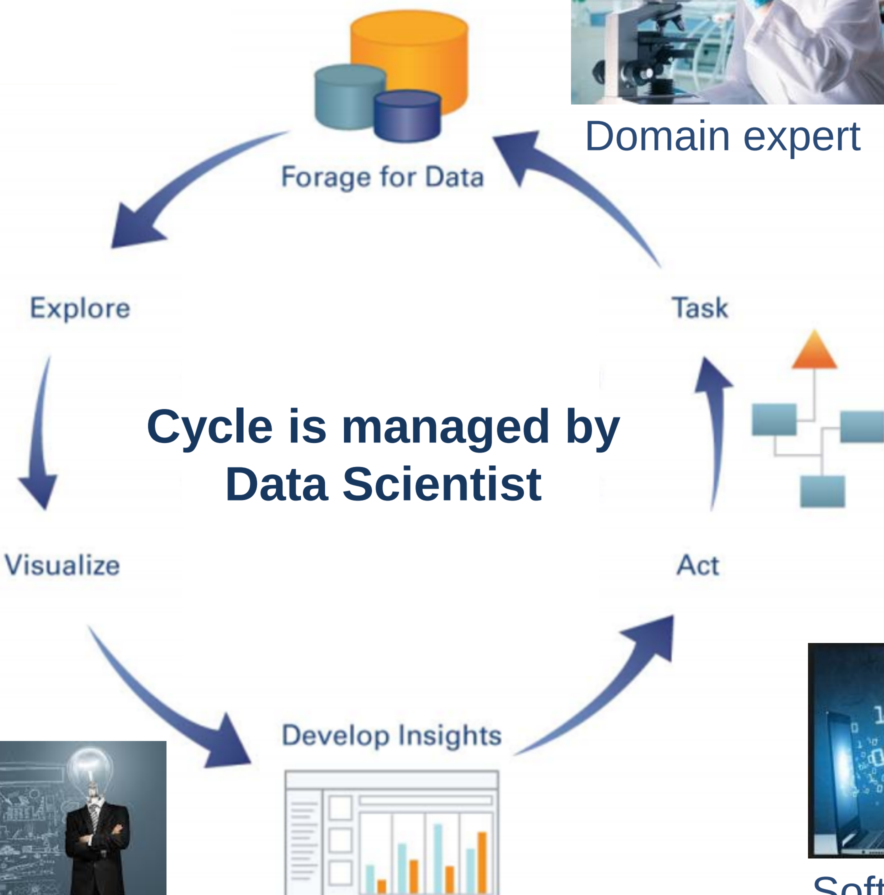
A suite to support the Visual Analytics Process

Vision

Big Data Analytics Cycle (Today)

Software developer

Domain expert



Cycle is managed by Data Scientist

Source: Tableau Software

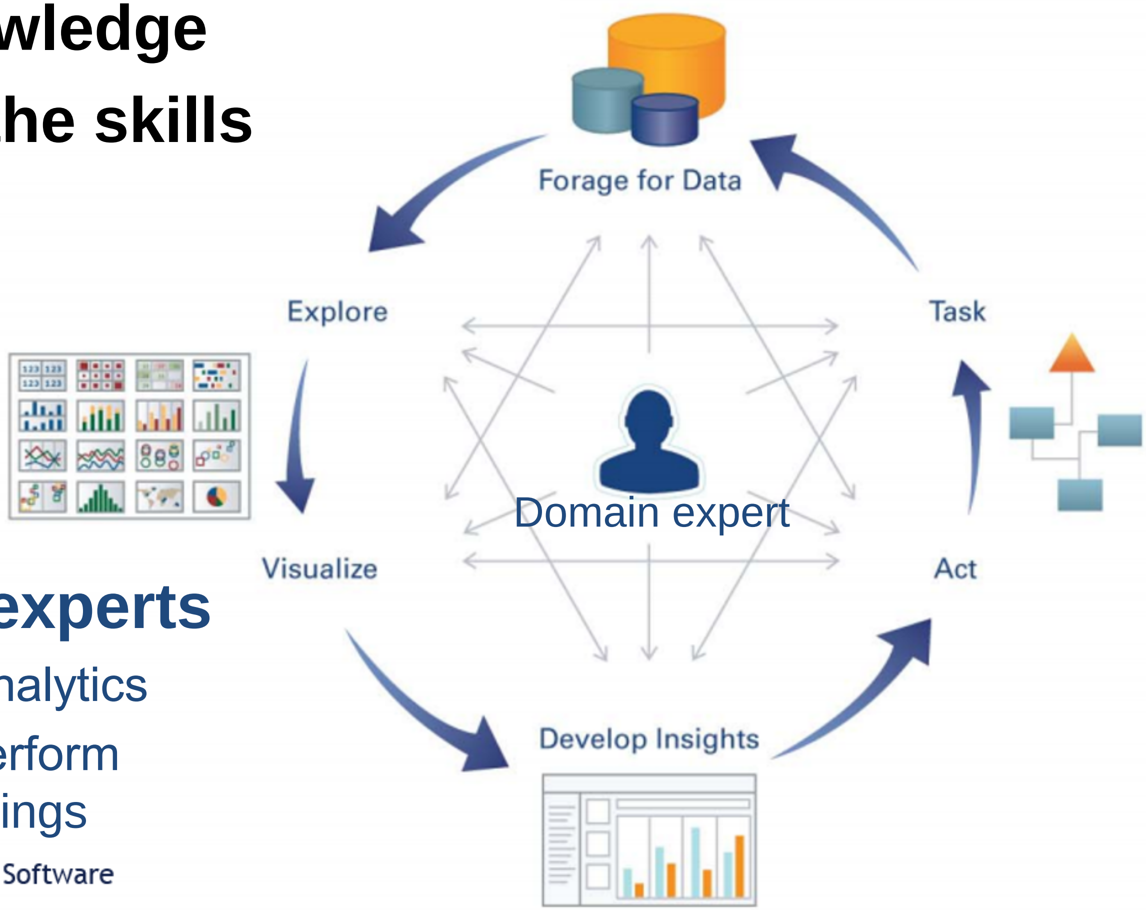
Domain expert

Software developer

Source: JIOX: Intelligence Tradecraft & Analysis

VISION➔ Expert at the centre of the cycle

- **Experts have the knowledge**
- **Data scientists have the skills**



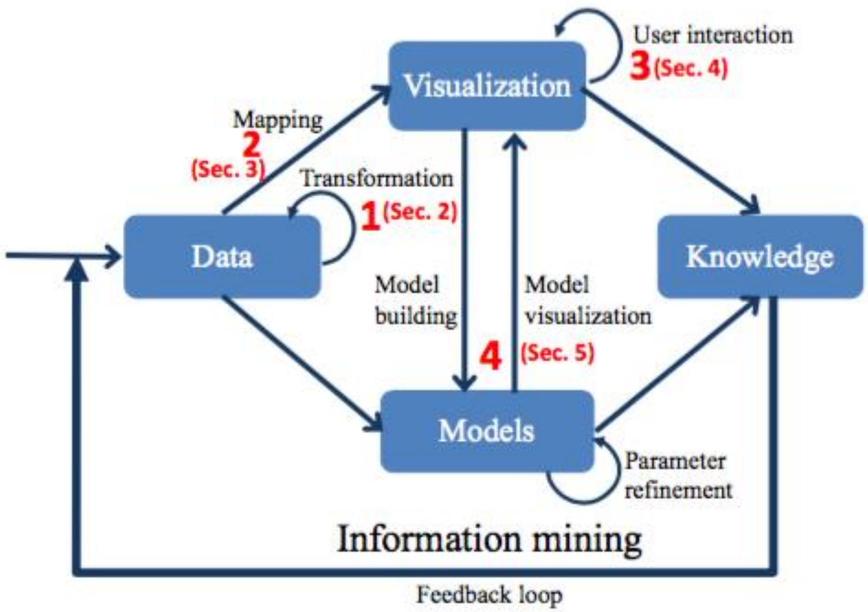
- **➔Bring analytics to experts**
  - “Understand” results of analytics
  - “Instruct” computers to perform analytics according to findings

Source: Tableau Software

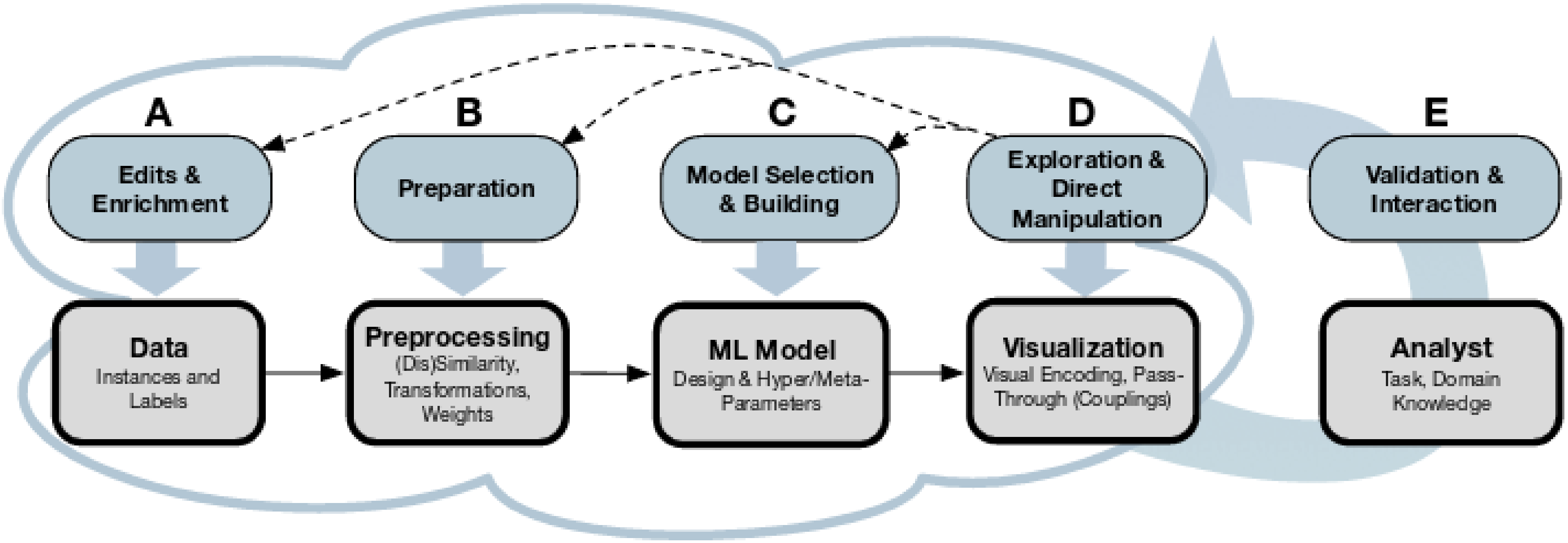
Data scientists to enable experts to perform analytics by themselves

Collaboration Spotting to support the Visual Analytics Process*

1. Data Pre-processing
2. Mapping/Layout
3. Visual user interactions
4. Model-based analysis



Collaboration Spotting Framework



The project follows the proposed conceptual framework of D. Sacha et al*.

*Human-Centered Machine Learning Through Interactive Visualization: Review and Open Challenges
Dominik Sacha, Michael Sedlmair, Leishi Zhang, John Aldo Lee, Daniel Weiskopf, Stephen North, Daniel Keim

A suite to support visual analytics

- **Long-term project**
- **Iterative approach**
  - **First release**: Publications and Patents metadata for Technology Innovation Monitoring
  - Applied to other data sources:
    - LHCb data processing
    - CERN procurement data
  - **Second release:** (under construction):
    - Data source pre-processing tools
    - User self-defined visual framework
    - Contextual analysis
  - **Third release**: The full suite

Big Data

Multi-dimensional networks

Graph database

Interactive graph visual analysis

CS Concepts

Characteristics of Big Data

- **Huge quantity**
- Processing and storage

- **Distributed sources**
- Access rights, security

- **Complexity**
- Valuable information may be hidden behind complexity

- **Interconnectivity**
- Unravelling new knowledge

➔ Data scientists are instrumental to analytics

➔ Domain experts are at the heart of the reasoning process

Big Data is organised in networks

Big Data is distributed

- Document systems with data and metadata in Database
- Database tables with metadata in schema

Big Data is strongly interconnected

- Networks are **not materialised** due to the distributed nature of data sources
- Ex: Publications and patents metadata

Building Data network

Metadata in source contains provider specific information and format

- Ex: The Web of Knowledge (Thomson Reuters)
  - Title, Abstract, Authors and Affiliation, DOI, Citations, etc.

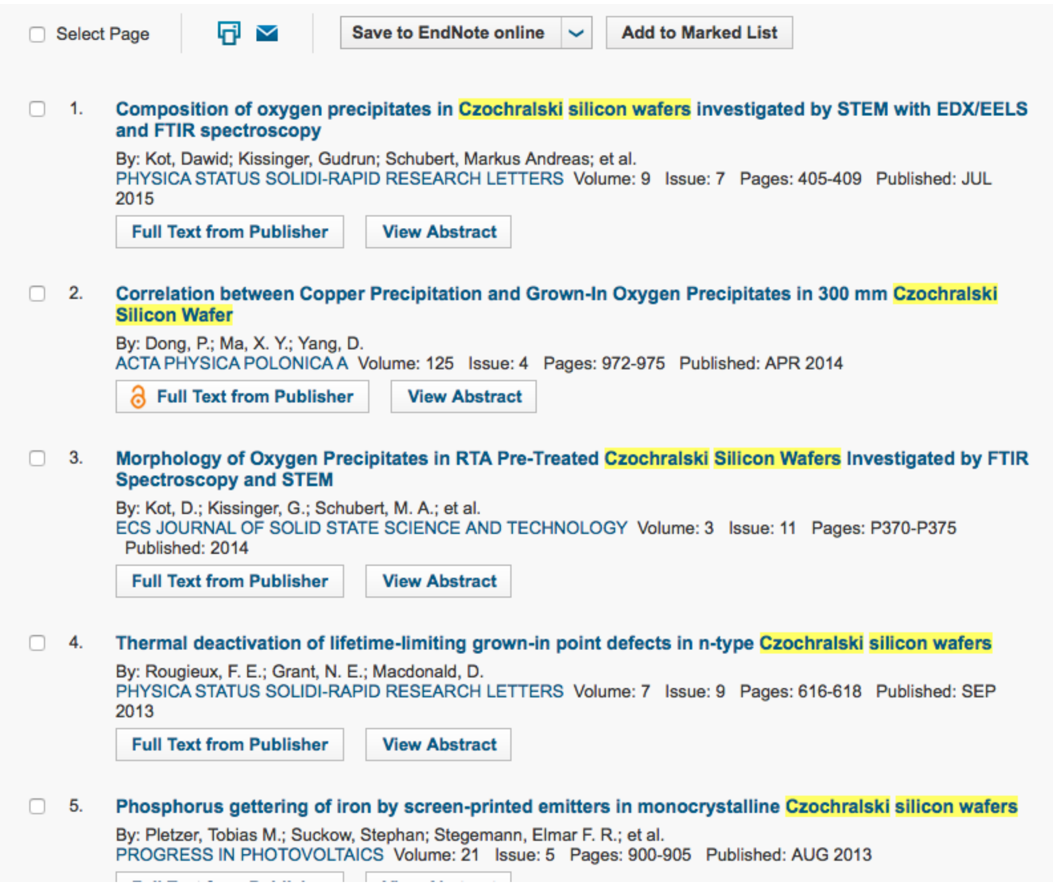
Some is of interest for analysis

- Title, Abstract, Citations
- → Each publication metadata becomes a vertex with attributes

Some is of interest for visualisation

- Organisations, cities, keywords, journal categories, Citations, etc.
- → Each of the above becomes a vertex with attribute that is link to its publication vertices

EX: Building a Network from publications metadata


Select Page | Save to EndNote online | Add to Marked List

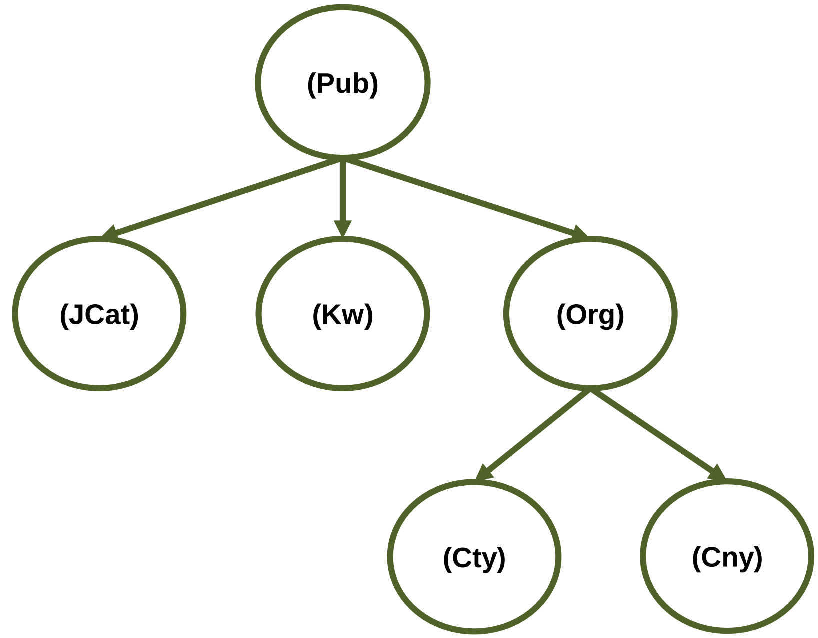
1. **Composition of oxygen precipitates in Czochralski silicon wafers investigated by STEM with EDX/EELS and FTIR spectroscopy**
By: Kot, Dawid; Kissinger, Gudrun; Schubert, Markus Andreas; et al.
PHYSICA STATUS SOLIDI-RAPID RESEARCH LETTERS Volume: 9 Issue: 7 Pages: 405-409 Published: JUL 2015
Full Text from Publisher | View Abstract

2. **Correlation between Copper Precipitation and Grown-In Oxygen Precipitates in 300 mm Czochralski Silicon Wafer**
By: Dong, P.; Ma, X. Y.; Yang, D.
ACTA PHYSICA POLONICA A Volume: 125 Issue: 4 Pages: 972-975 Published: APR 2014
Full Text from Publisher | View Abstract

3. **Morphology of Oxygen Precipitates in RTA Pre-Treated Czochralski Silicon Wafers Investigated by FTIR Spectroscopy and STEM**
By: Kot, D.; Kissinger, G.; Schubert, M. A.; et al.
ECS JOURNAL OF SOLID STATE SCIENCE AND TECHNOLOGY Volume: 3 Issue: 11 Pages: P370-P375 Published: 2014
Full Text from Publisher | View Abstract

4. **Thermal deactivation of lifetime-limiting grown-in point defects in n-type Czochralski silicon wafers**
By: Rougieux, F. E.; Grant, N. E.; Macdonald, D.
PHYSICA STATUS SOLIDI-RAPID RESEARCH LETTERS Volume: 7 Issue: 9 Pages: 616-618 Published: SEP 2013
Full Text from Publisher | View Abstract

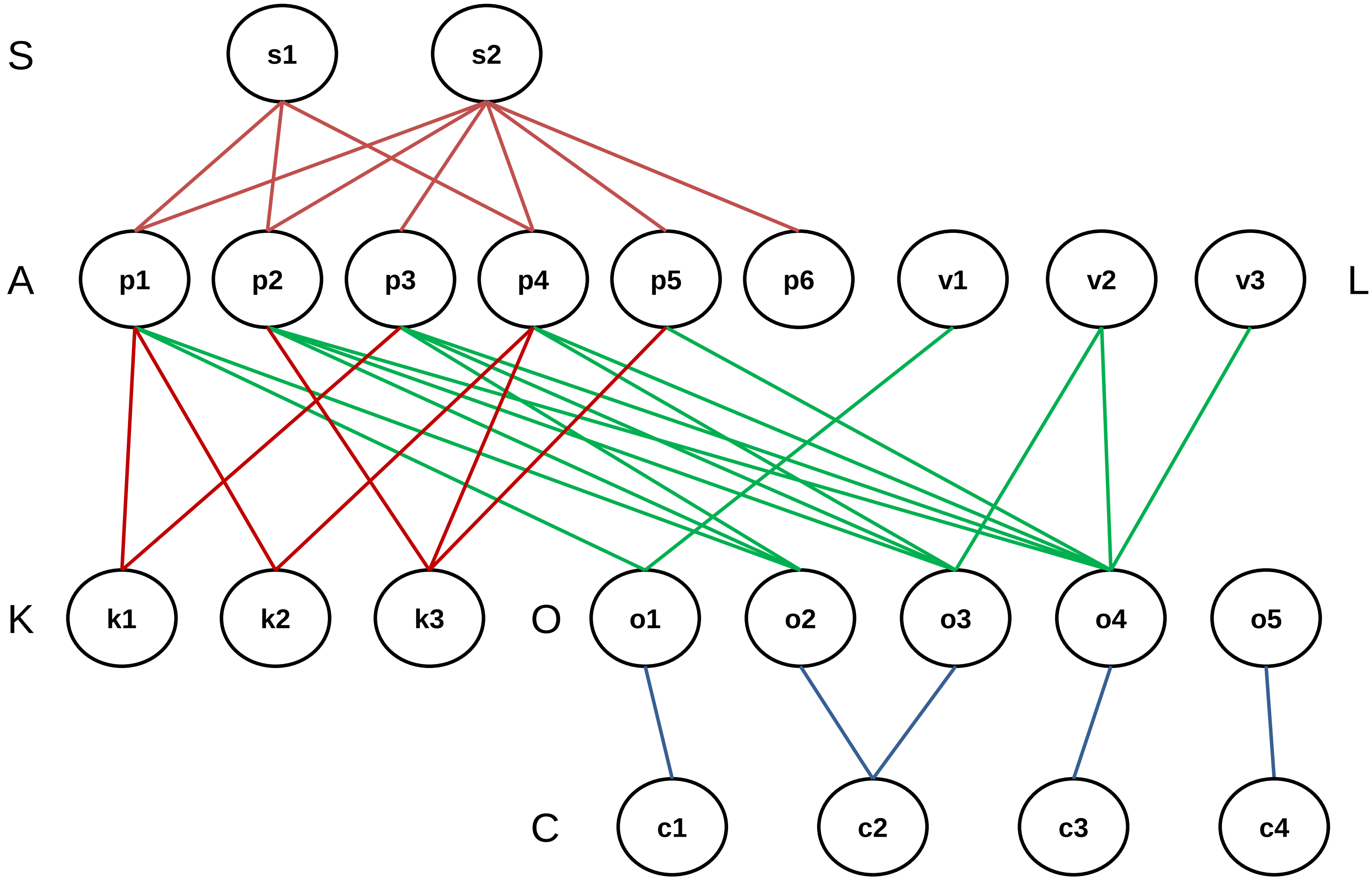
5. **Phosphorus gettering of iron by screen-printed emitters in monocrystalline Czochralski silicon wafers**
By: Pletzer, Tobias M.; Suckow, Stephan; Stegemann, Elmar F. R.; et al.
PROGRESS IN PHOTOVOLTAICS Volume: 21 Issue: 5 Pages: 900-905 Published: AUG 2013



Document metadata | **Reachability Graph: Graph of data types**

JCat: Journal category, Kw: Keyword, Org: Organisation, Cny: Country, Cty: City

Graph of Metadata / Data



S: Categories, A: Pub/Pat, K: keywords, O: Organisations, C: Countries, L: Cities

Storing network data as graphs

Graphs are natural representations of networks

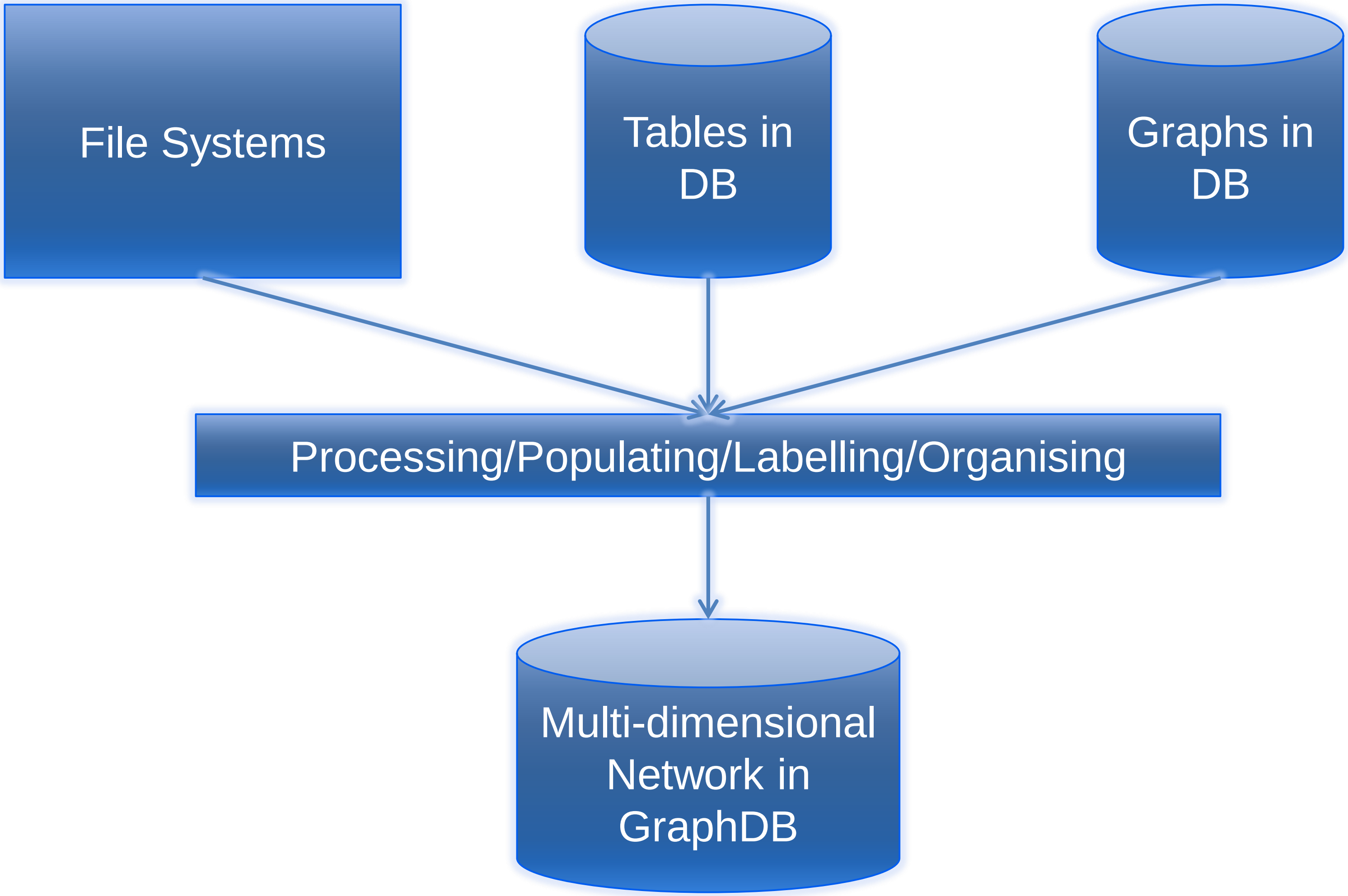
- Complexity
- Interconnectivity
- Scalability
- Multi dimensional

Schema is embedded in the data

- Nodes' labels
- Compact graph structure
- Graph query language
- No schema evolution

➔ **Graph Databases offer a natural support for storing network information**

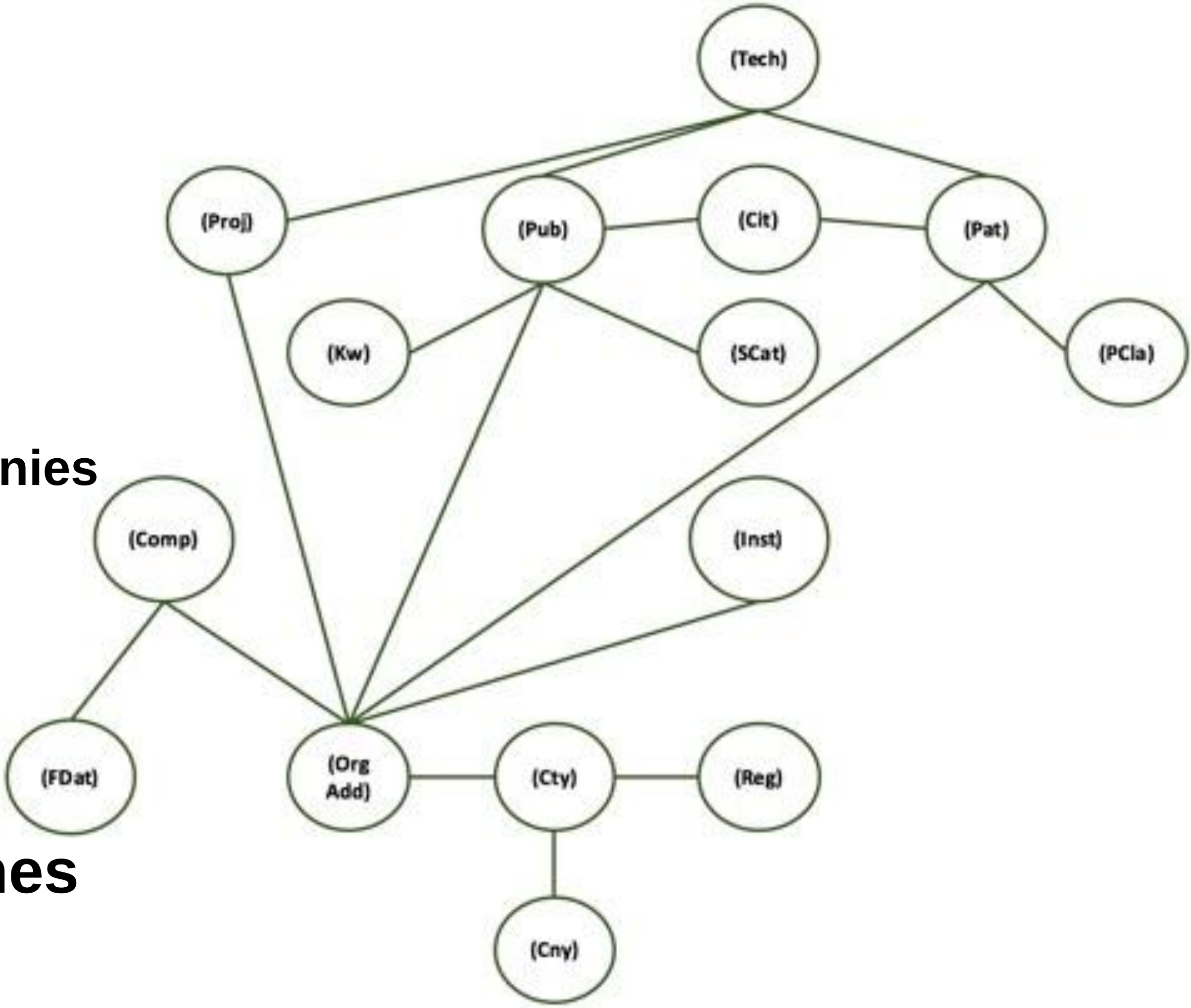
Building multi-dimensional networks from various data sources



➔ No limit on the sixe of a network!

Example: Enriching Network

- **Data sources**
  - Publications/**Patents**
    - **Citations**
    - **Institutions/Companies**
- **Data sources**
  - **EU projects**
  - **Financial data**
  - **Geolocation data**
- **Technology searches**
  (resulting from processing)



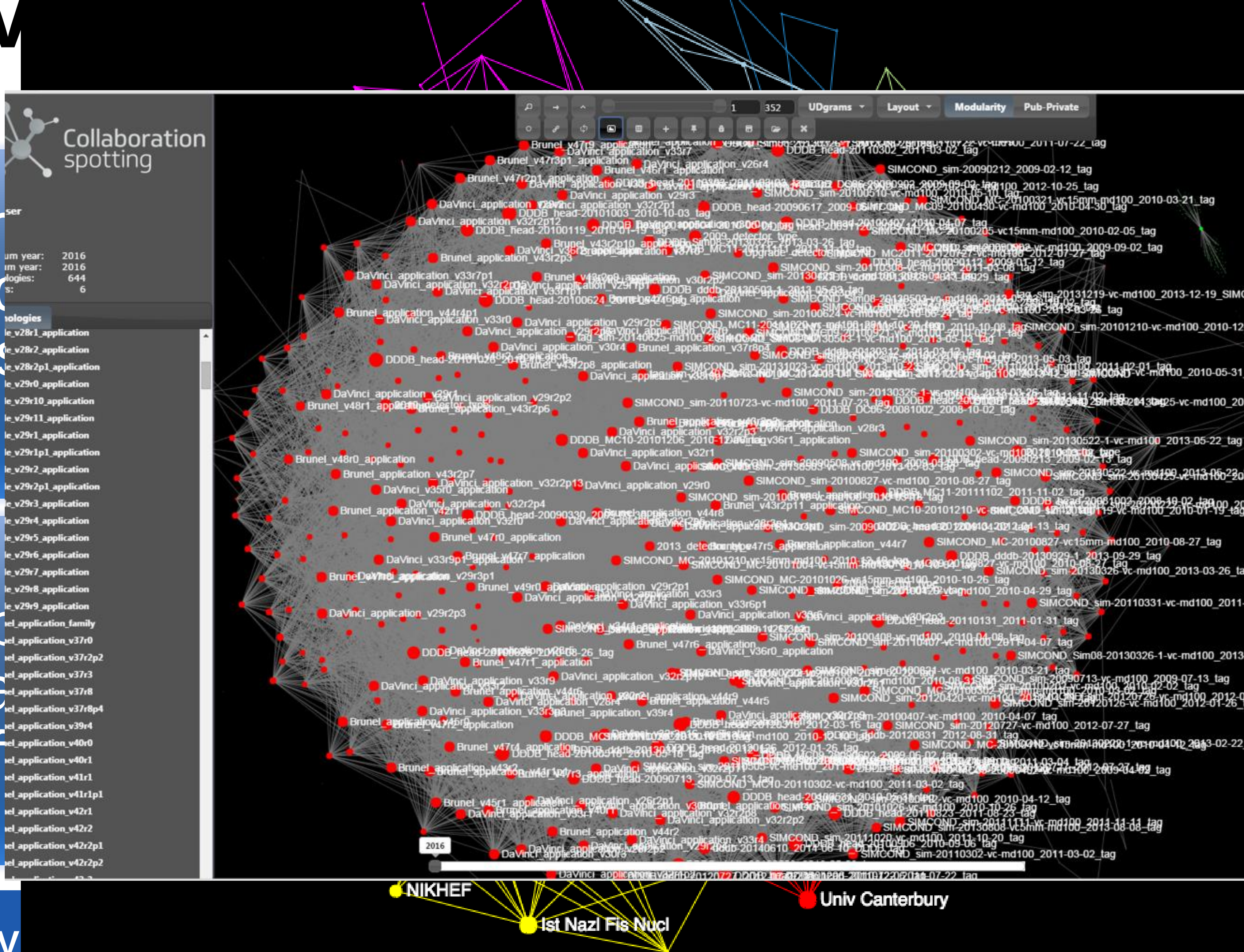
Reachability Graph (Schema): Graph of linked datatypes
Dimension: a node in the Reachability Graph (a datatype)

Graph of Organisations

V

Gr
exc
vis
n

Grap
ma
inf



→ v

Setting up user analysis environment

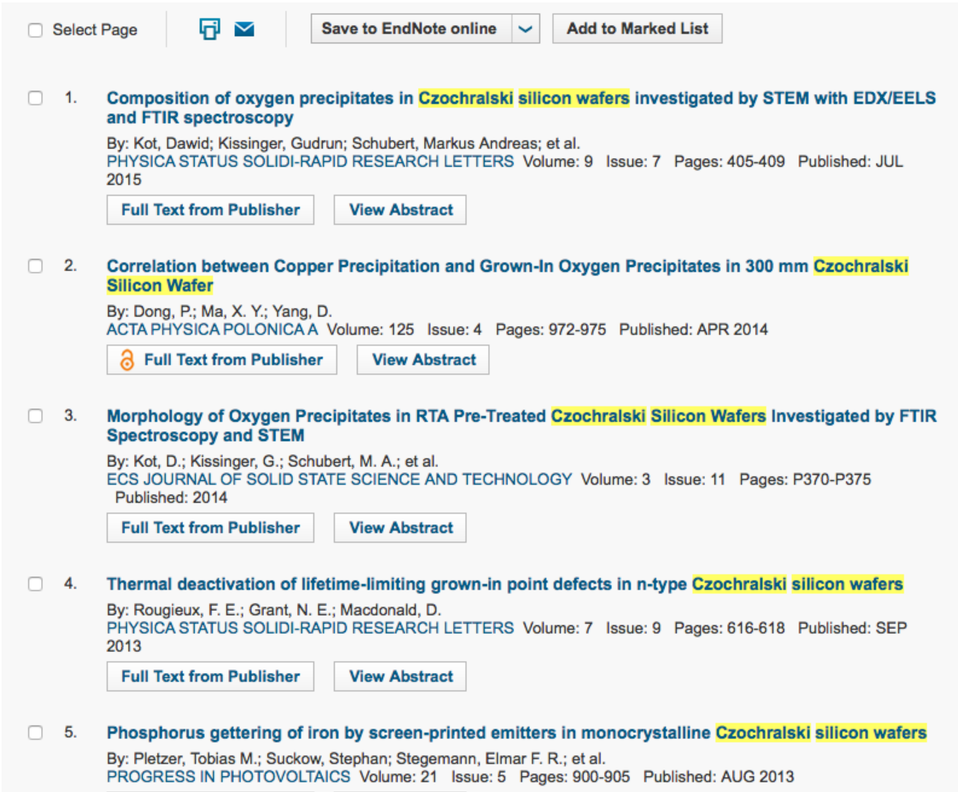
Reachability Graph

- Graph of connected dimensions
- Data analysis dimensions (user selection)
  - Ex. 1: Publications, Patents
  - Ex. 2 : Compatibility and require relationships in LHCb
- Visual analysis dimensions (user selection)
  - Ex. 1: Organisations, cities, countries, keywords, categories, etc.
  - Ex. 2: Components, Environment, conditions, etc.

Entry graph (user specified)

- Visual dimension of the front graph
- Ex. 1: Technology
- Ex. 2: Processing Pass Description → Connects to top applications

Selecting dimensions (Ex. 1)


Select Page | Save to EndNote online | Add to Marked List

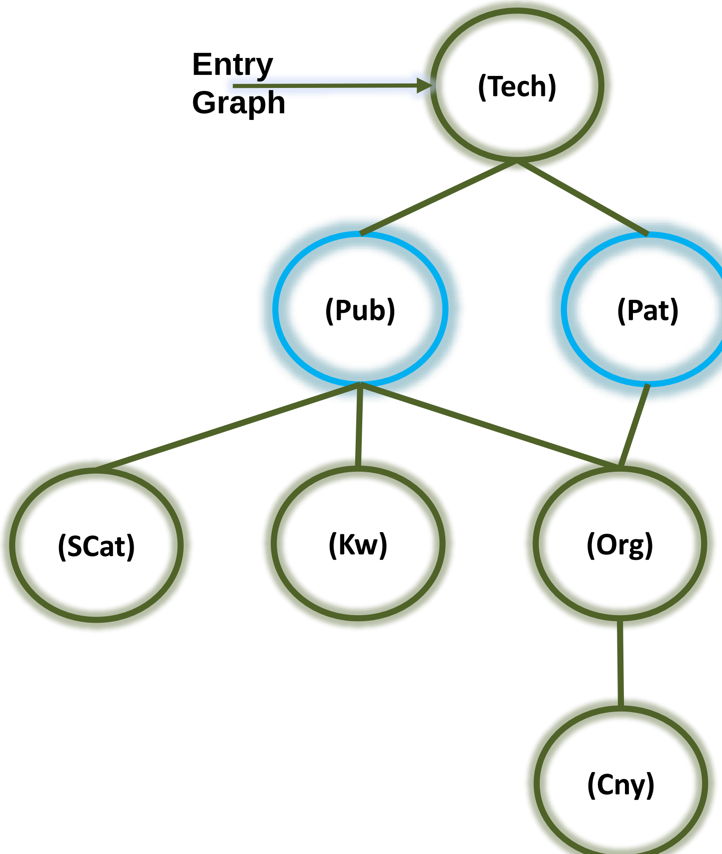
1. Composition of oxygen precipitates in Czochralski silicon wafers investigated by STEM with EDX/EELS and FTIR spectroscopy
By: Kot, Dawid; Kissinger, Gudrun; Schubert, Markus Andreas; et al.
PHYSICA STATUS SOLIDI-RAPID RESEARCH LETTERS Volume: 9 Issue: 7 Pages: 405-409 Published: JUL 2015
Full Text from Publisher | View Abstract

2. Correlation between Copper Precipitation and Grown-In Oxygen Precipitates in 300 mm Czochralski Silicon Wafer
By: Dong, P.; Ma, X. Y.; Yang, D.
ACTA PHYSICA POLONICA A Volume: 125 Issue: 4 Pages: 972-975 Published: APR 2014
Full Text from Publisher | View Abstract

3. Morphology of Oxygen Precipitates in RTA Pre-Treated Czochralski Silicon Wafers Investigated by FTIR Spectroscopy and STEM
By: Kot, D.; Kissinger, G.; Schubert, M. A.; et al.
ECS JOURNAL OF SOLID STATE SCIENCE AND TECHNOLOGY Volume: 3 Issue: 11 Pages: P370-P375 Published: 2014
Full Text from Publisher | View Abstract

4. Thermal deactivation of lifetime-limiting grown-in point defects in n-type Czochralski silicon wafers
By: Rougieux, F. E.; Grant, N. E.; Macdonald, D.
PHYSICA STATUS SOLIDI-RAPID RESEARCH LETTERS Volume: 7 Issue: 9 Pages: 616-618 Published: SEP 2013
Full Text from Publisher | View Abstract

5. Phosphorus gettering of iron by screen-printed emitters in monocrystalline Czochralski silicon wafers
By: Pletzer, Tobias M.; Suckow, Stephan; Stegemann, Elmar F. R.; et al.
PROGRESS IN PHOTOVOLTAICS Volume: 21 Issue: 5 Pages: 900-905 Published: AUG 2013



Data dimensions for Analytics

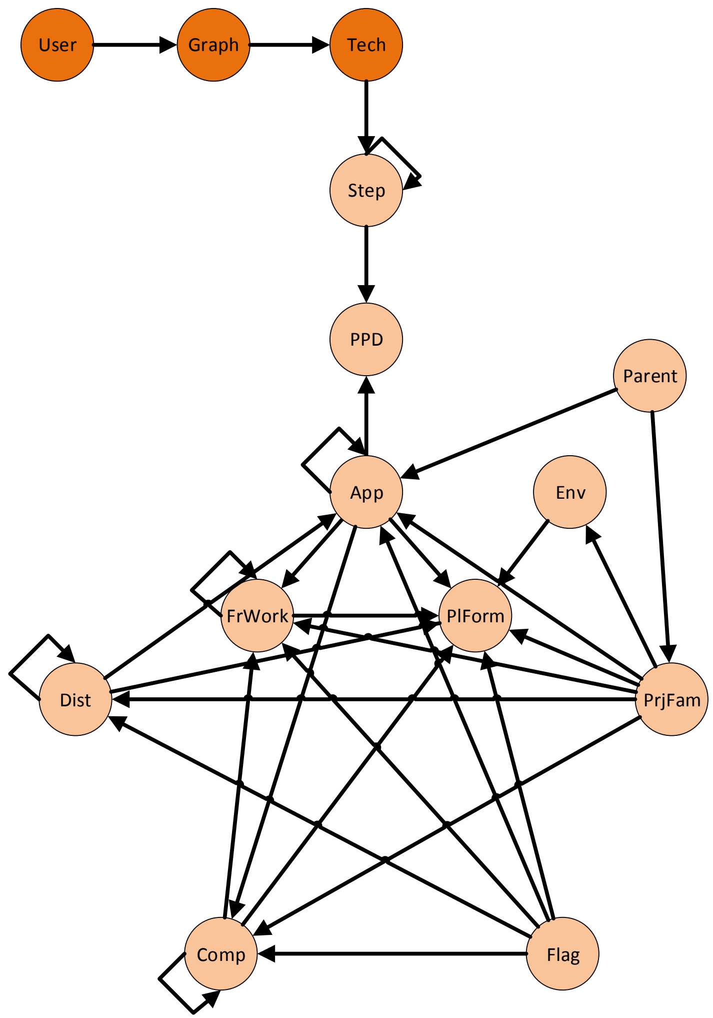
Pub: Publications, Pat: Patents (Attributes: Title and abstract are used for semantic searches)

Visualisation dimensions of Analytics results:

SCat: Journal category, Kw: Keyword, Org: Organisation and Cny: Country)

<u>Technology Search</u>: Czochralski Silicon wafer
<u>Pub/Pat</u>: documents found in search results

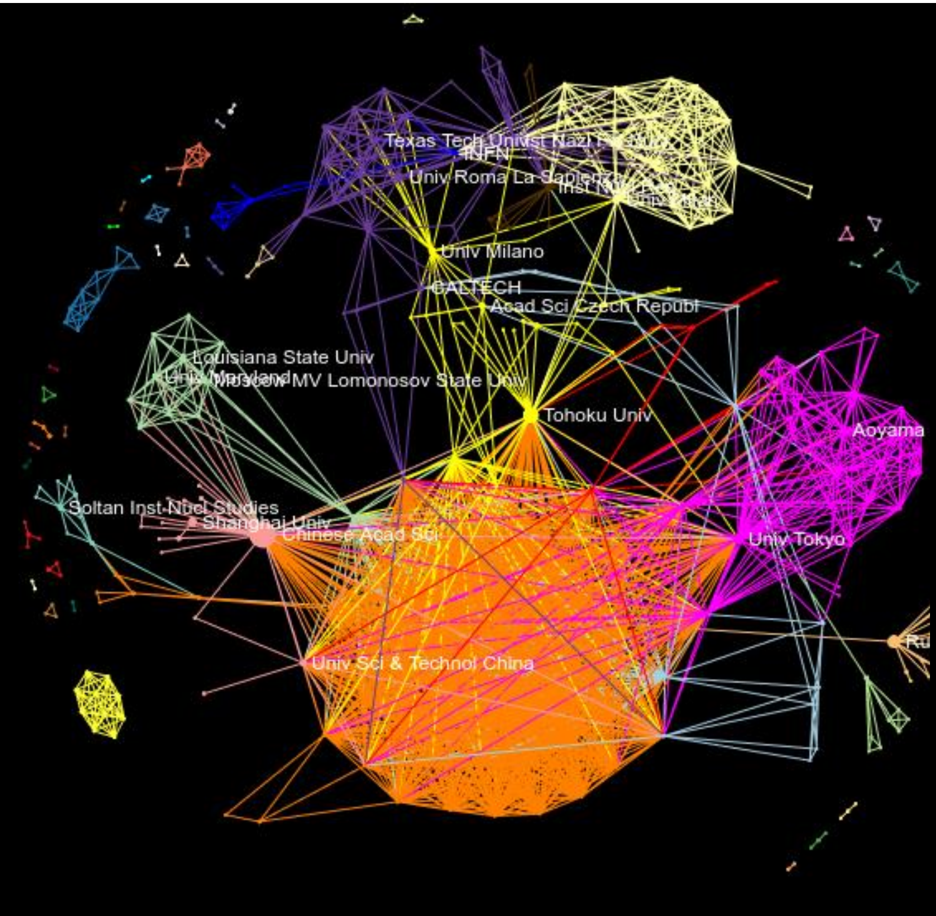
Selecting dimensions (Ex. 2)



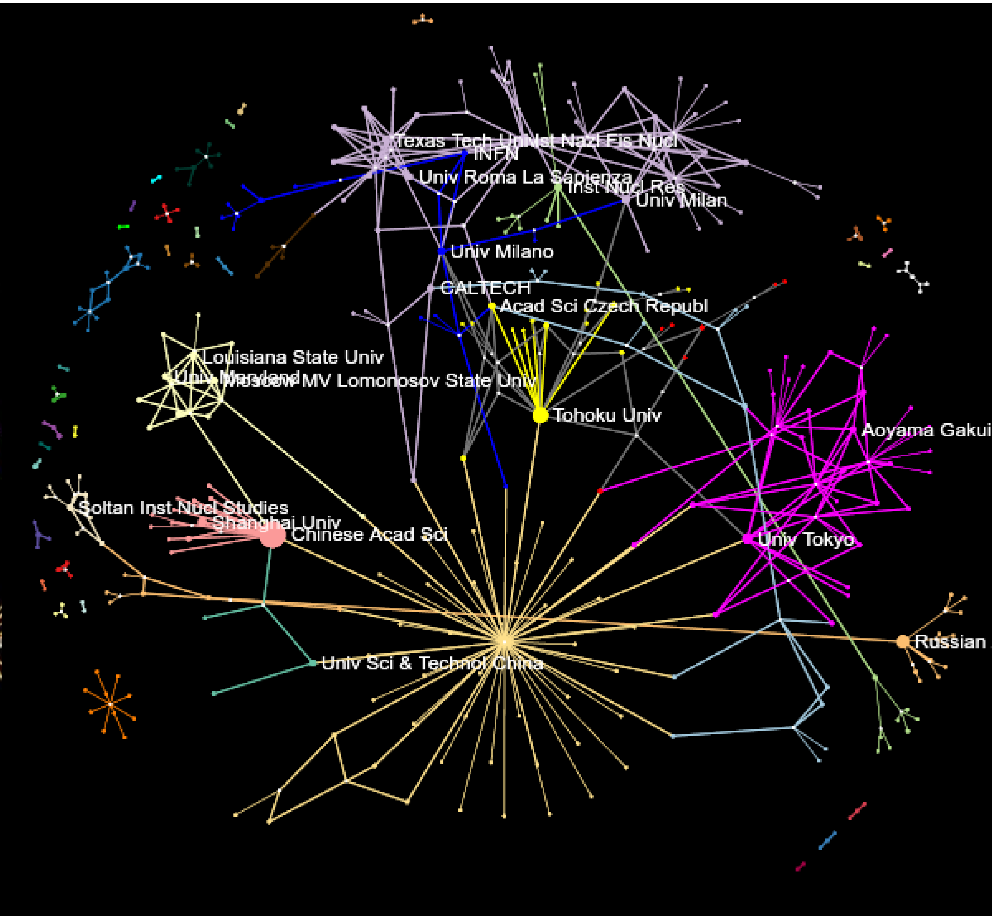
CollSpotting supported Graph Visual representations

- Static graph with timeline window
- Node-link using different layout techniques
  - Clique representation (default)
    - Force Atlas (default)
    - Circular representation
  - Extra node representation (hyper-graph)
    - Force Atlas
    - Circular representation

Clique vs Extra node representations (ForceAtlas)

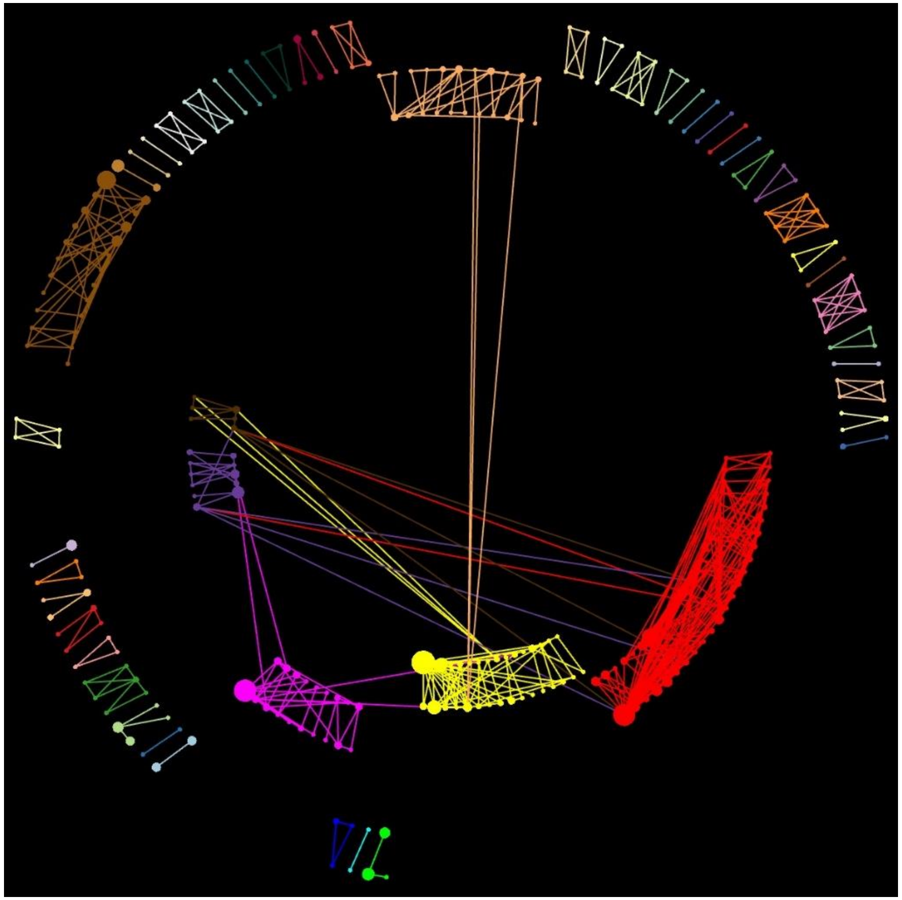


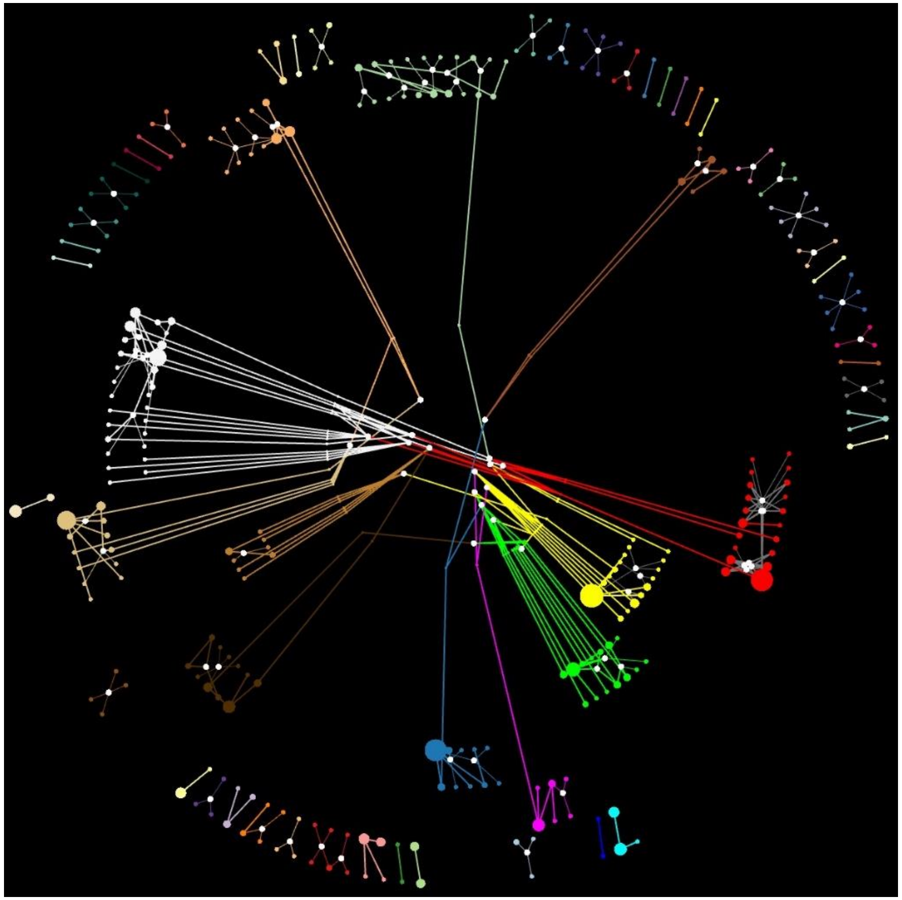
Organisation landscape graph view



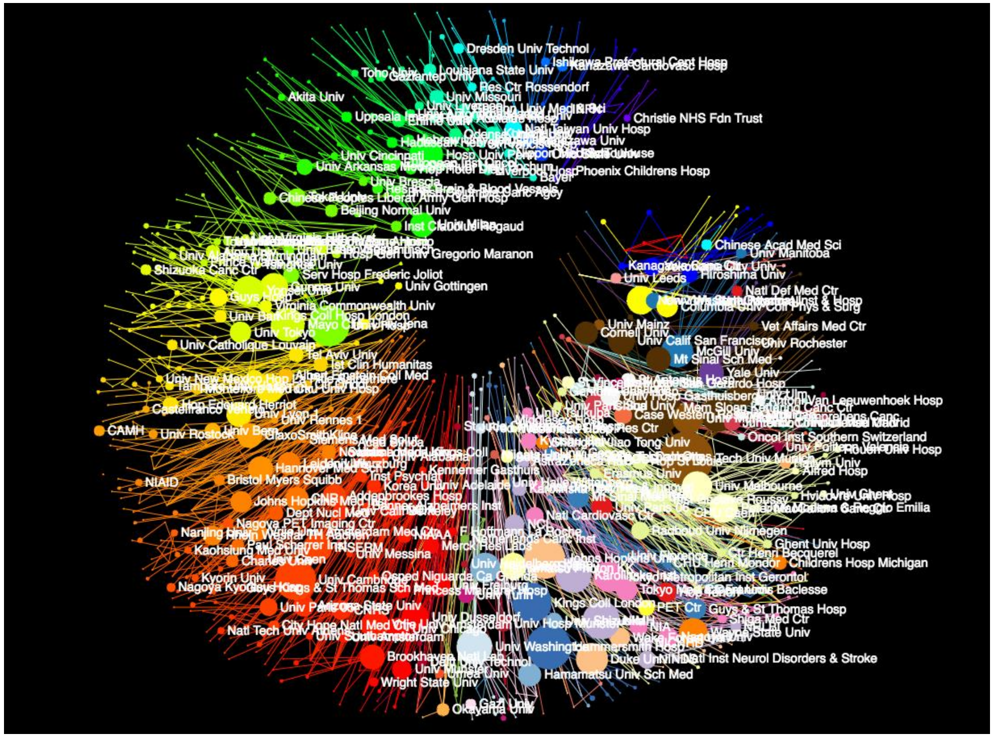
Organisation landscape hypergraph view

Clique vs Extra Node circular representations

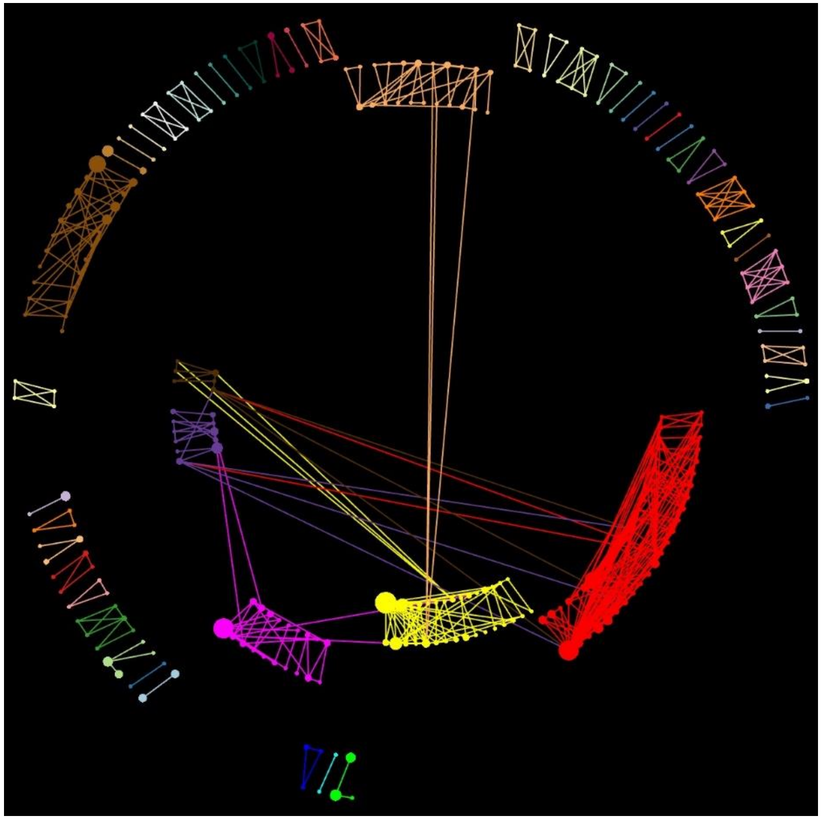




ForceAtlas vs Circular



Clique view with ForceAtlas2



Cluster circular distribution according to cluster interconnectivity

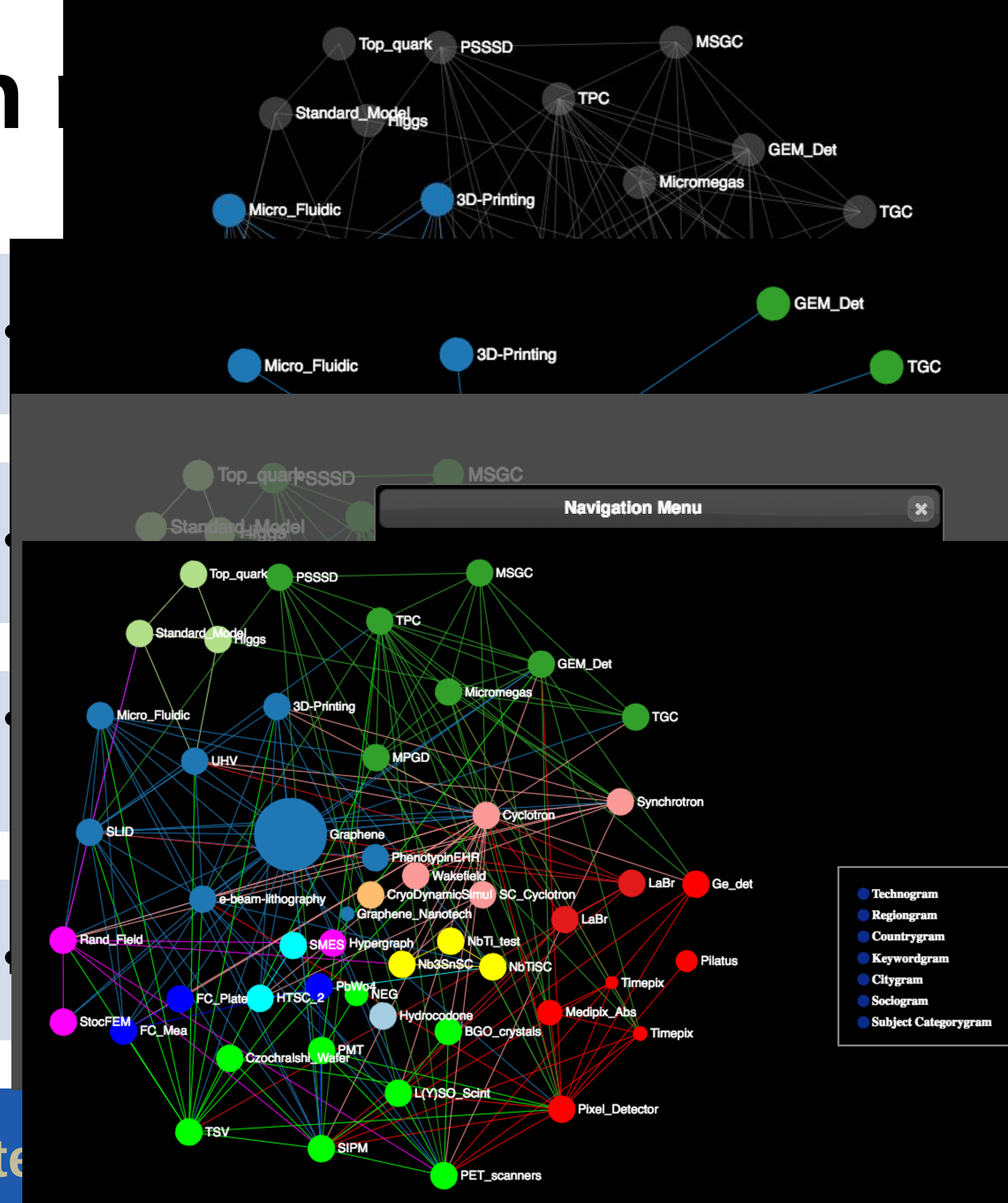
Visual graph

Hovering:

Left click:

Right click:

Right pane



Dataset: Publications and Patents

Illustration of visual interactive graph features

CS Graph visualisation features

Maximizing human understanding

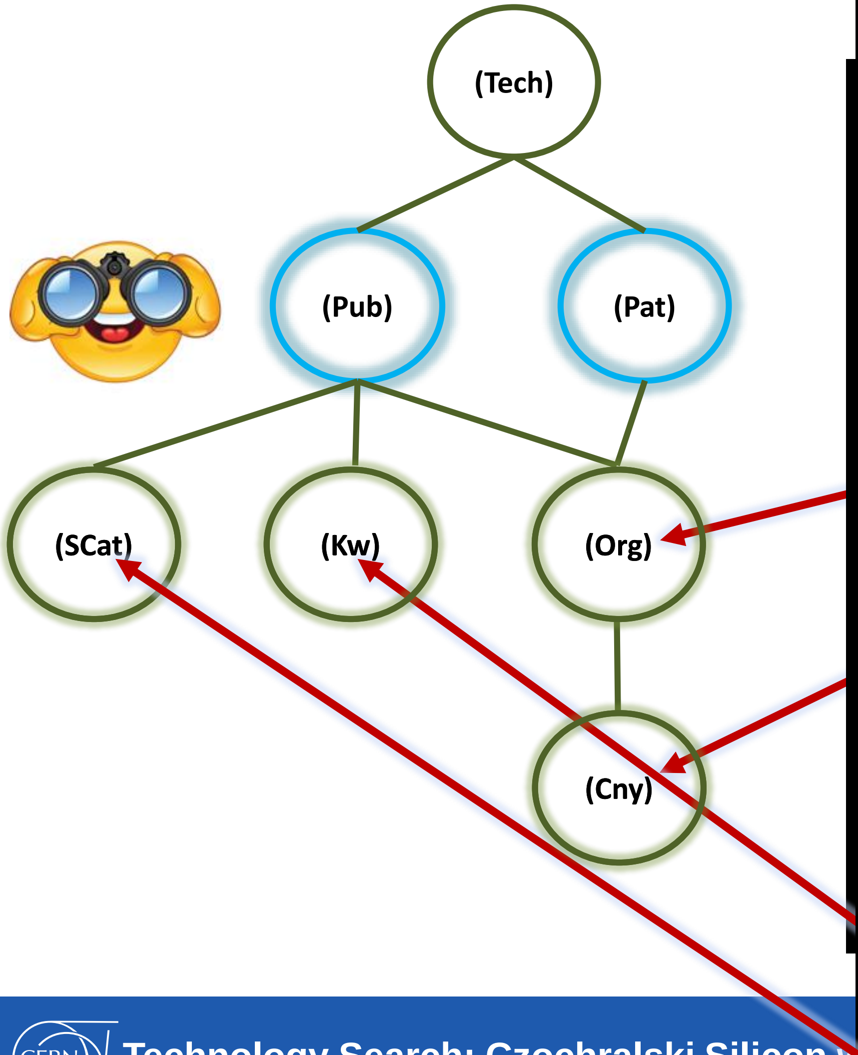
- Selecting network dimensions
- Traversing network dimensions
- Graphical queries
- Time/Frequency evolution

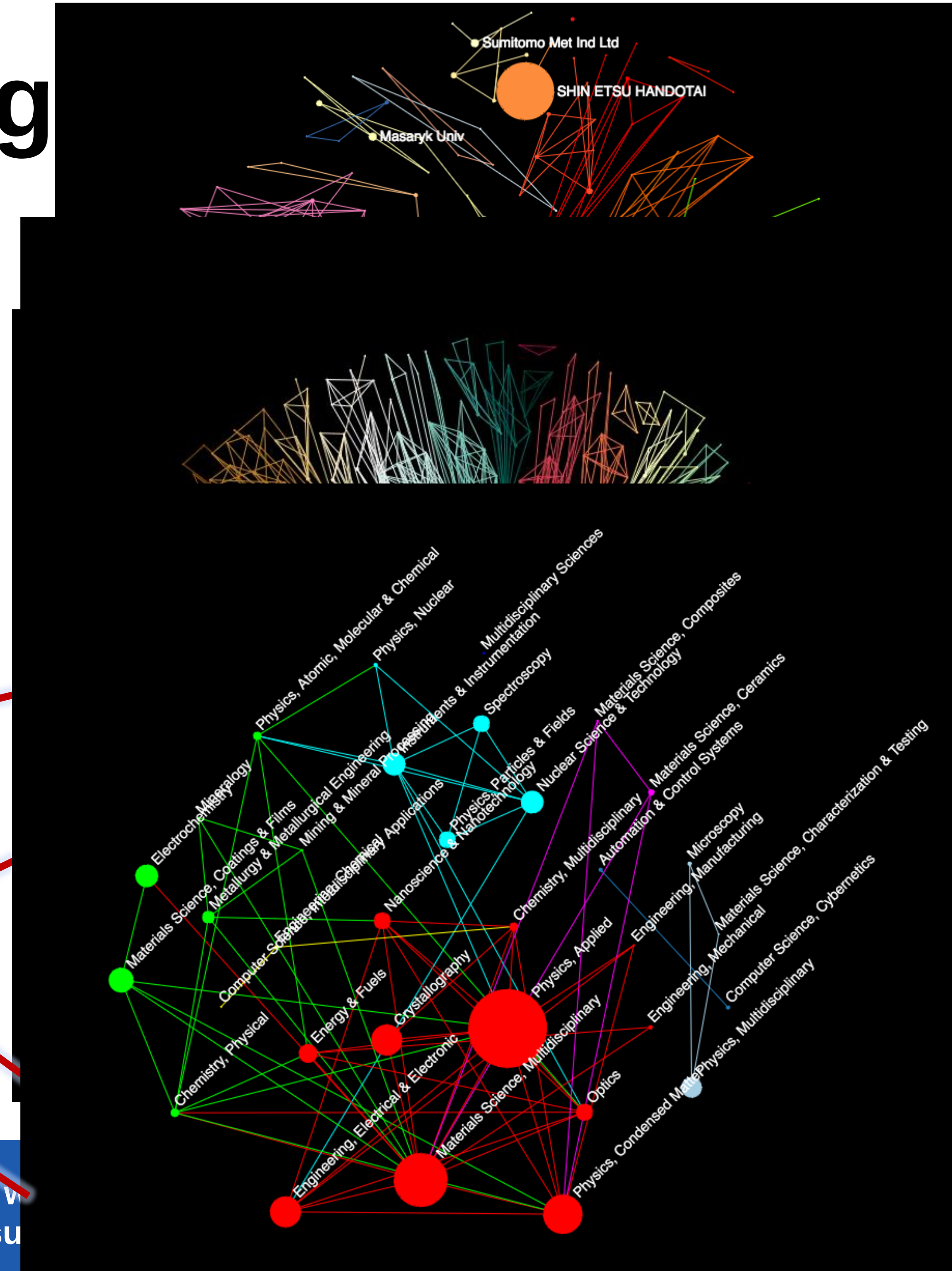
Enhancing reasoning

- Viewing multiple data sources
- Looking for collaborations
- Sorting data
- Contextual visualisation & analytics

➔Features illustrated in the following slides using publications/patents

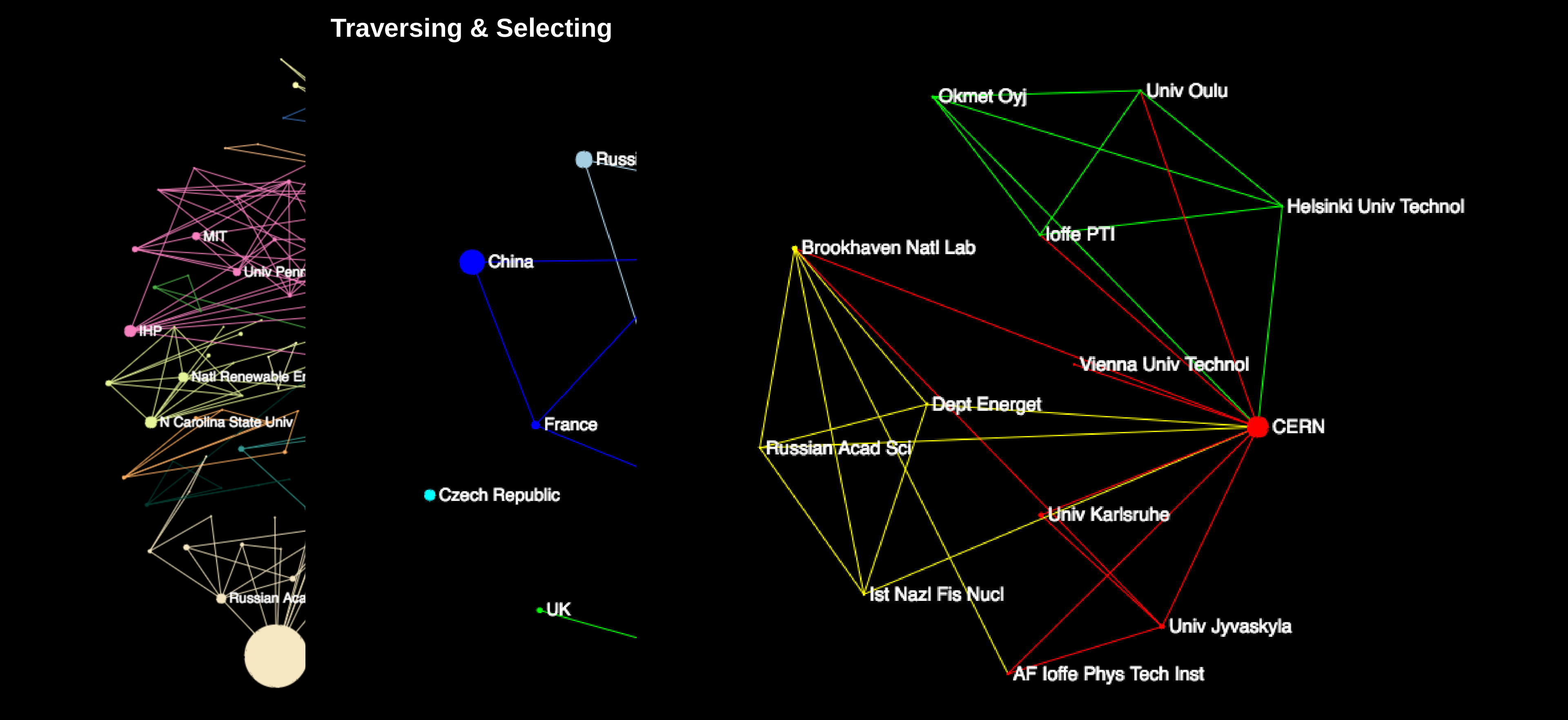
Traversing





Technology Search: Czochralski Silicon w
Pub/Pat: documents found in search resu

Graphical Queries



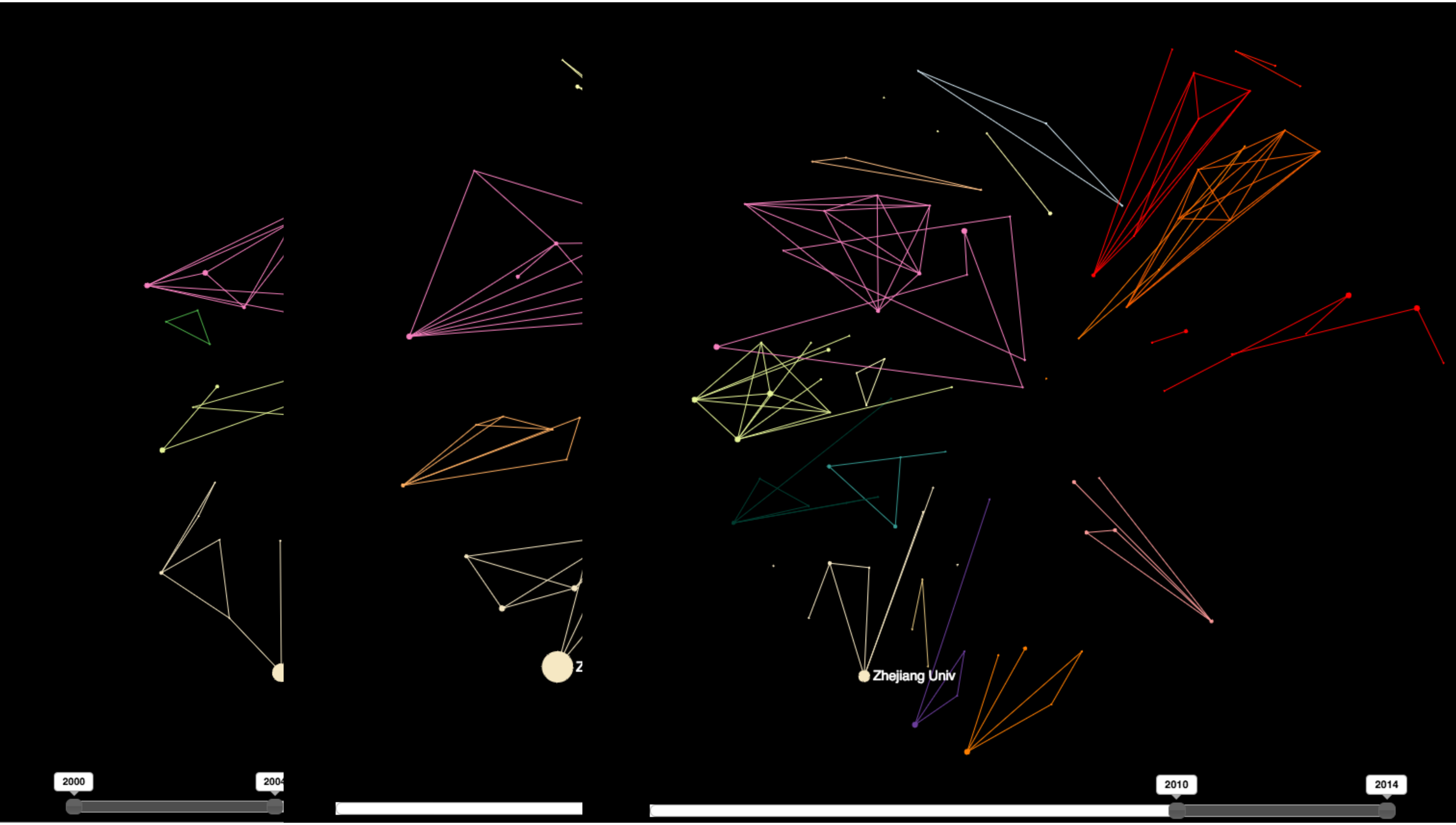
Organisation landscape

Country landscape

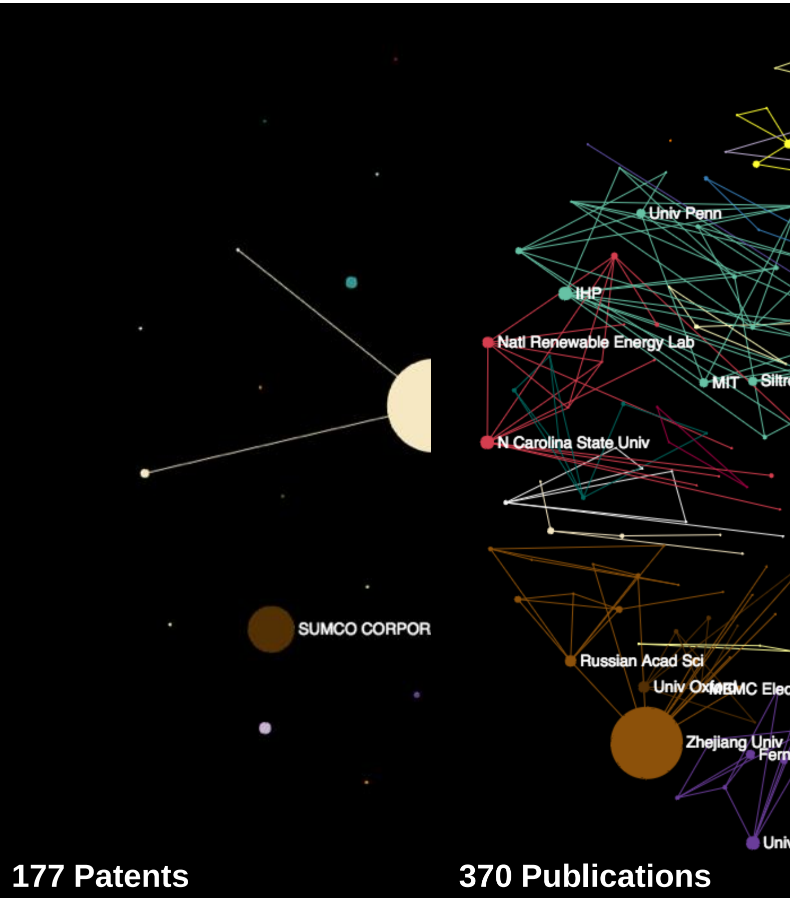
Organisation landscape for Switzerland

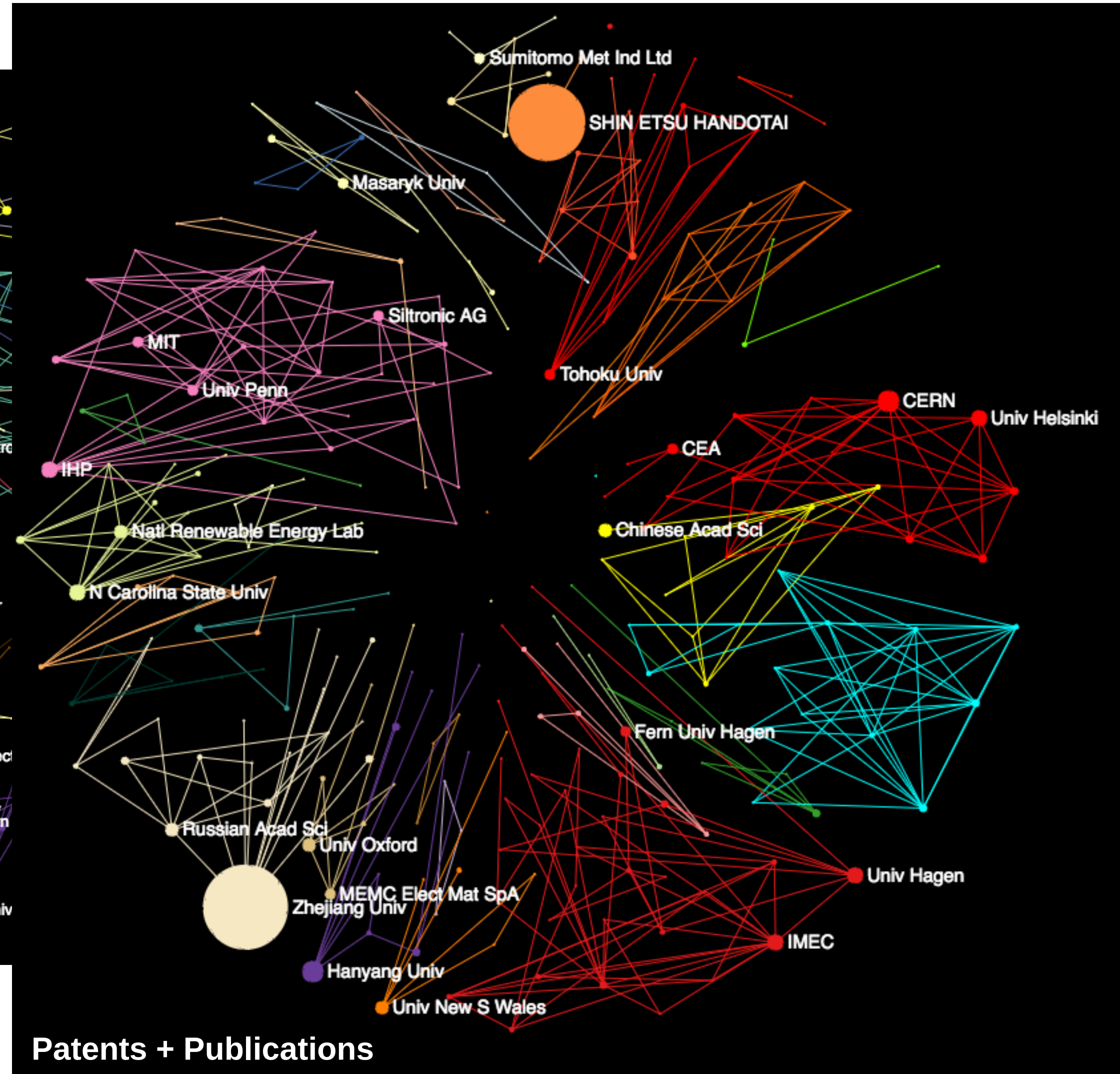
Technology Search: Czochralski Silicon wafer
Pub/Pat: documents found in search results

Time/Frequency evolution

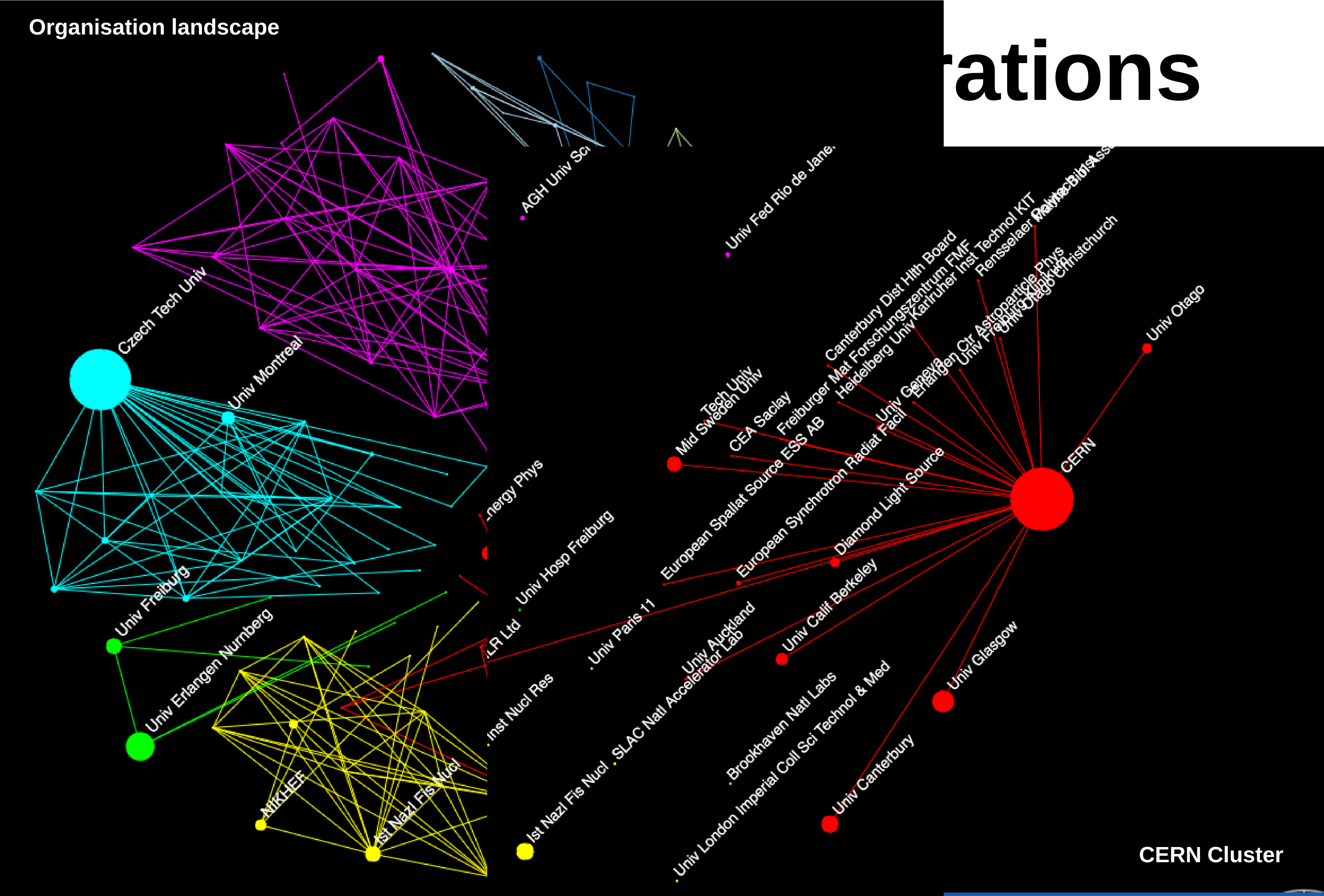


Combining data sources





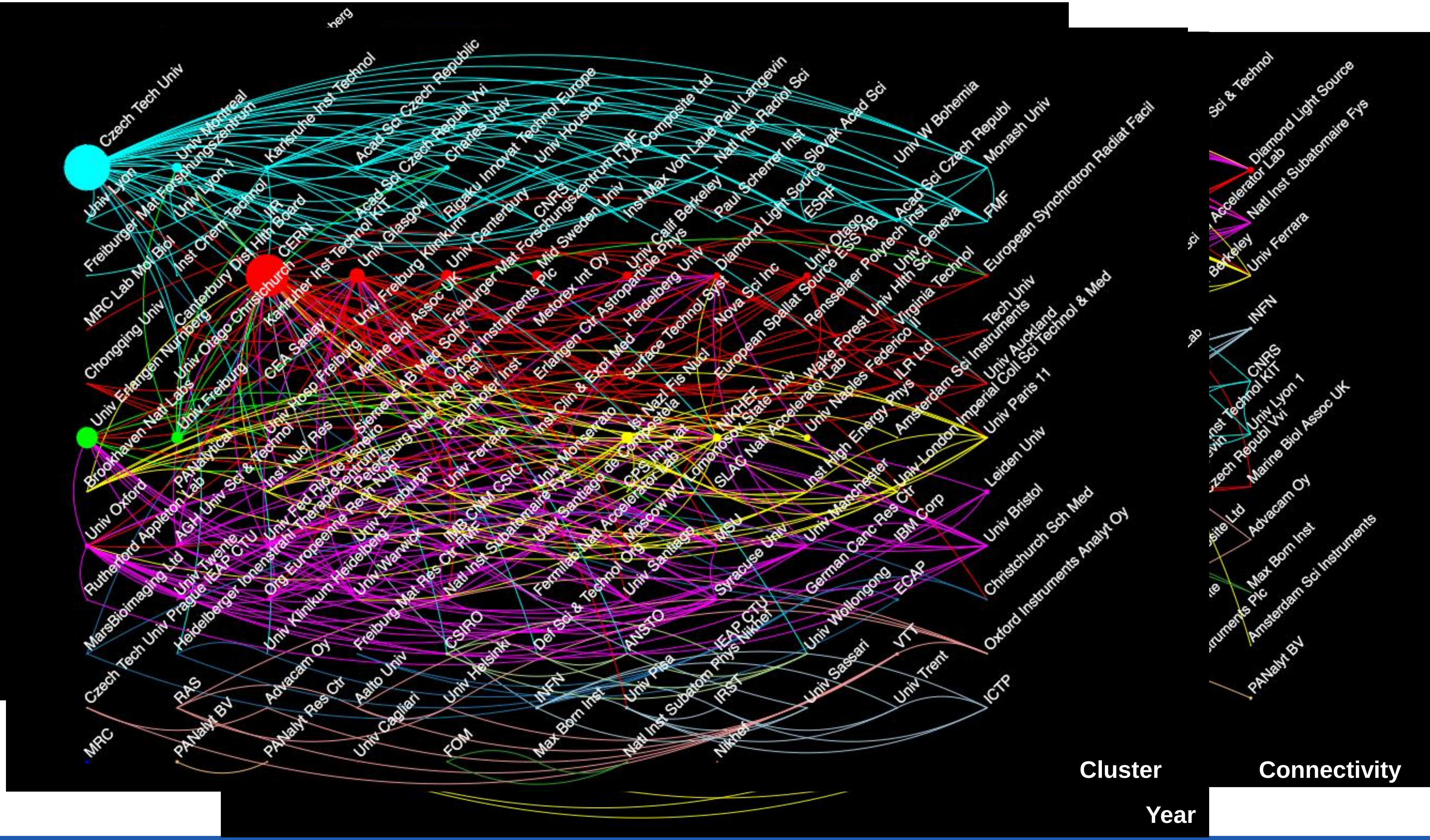
Technology Search: Czochralski Silicon wafer
Pub/Pat: documents found in search results



Technology Search: Pixelated detector (Medipix)
Pub/Pat: Documents found in search results

Cluster: Organisations that tend to collaborate more together than with others

Sorting information



Technology Search: Pixelated detector (Medipix)
Pub/Pat: Documents found in search results

Contextual visualisation

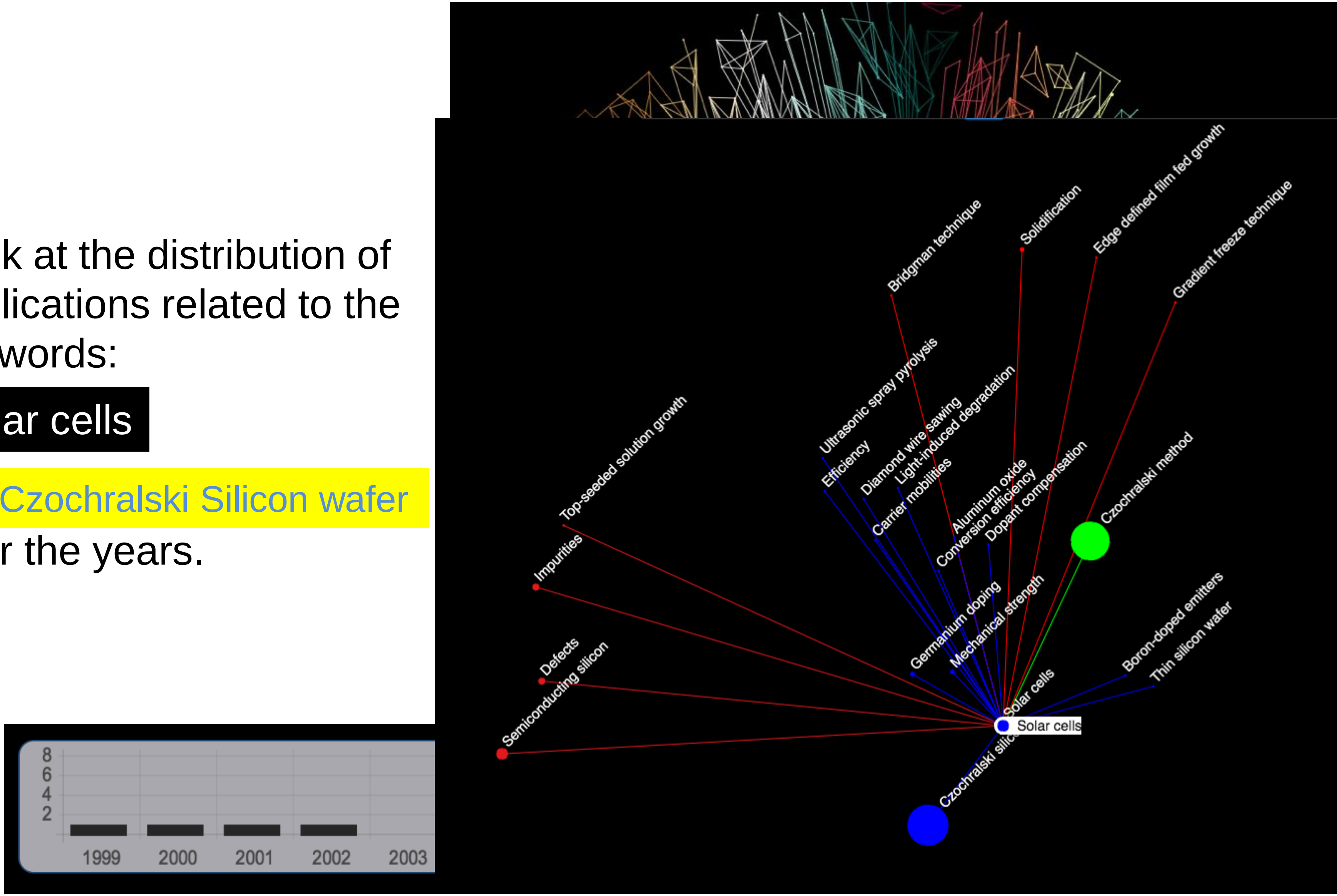
Technology search: 3-D Printing
Pub/Pat: documents found in search results

Contextual Analytics (egocentric view)

Look at the distribution of publications related to the keywords:

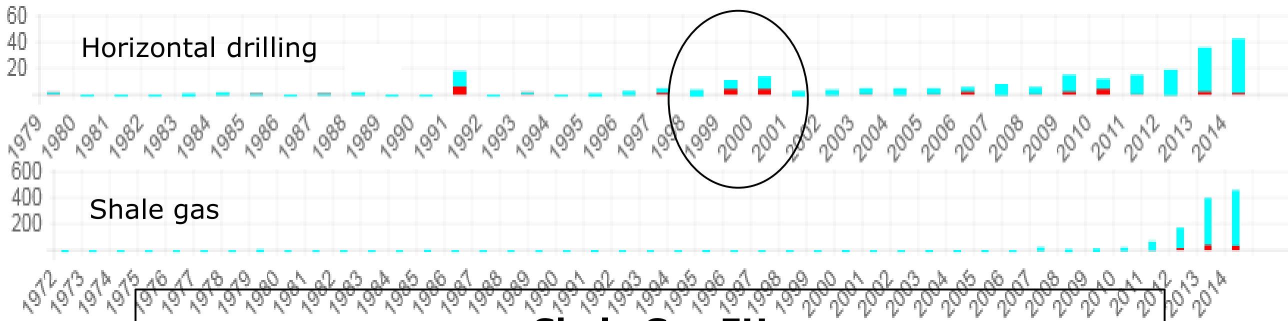
Solar cells

In Czochralski Silicon wafer over the years.



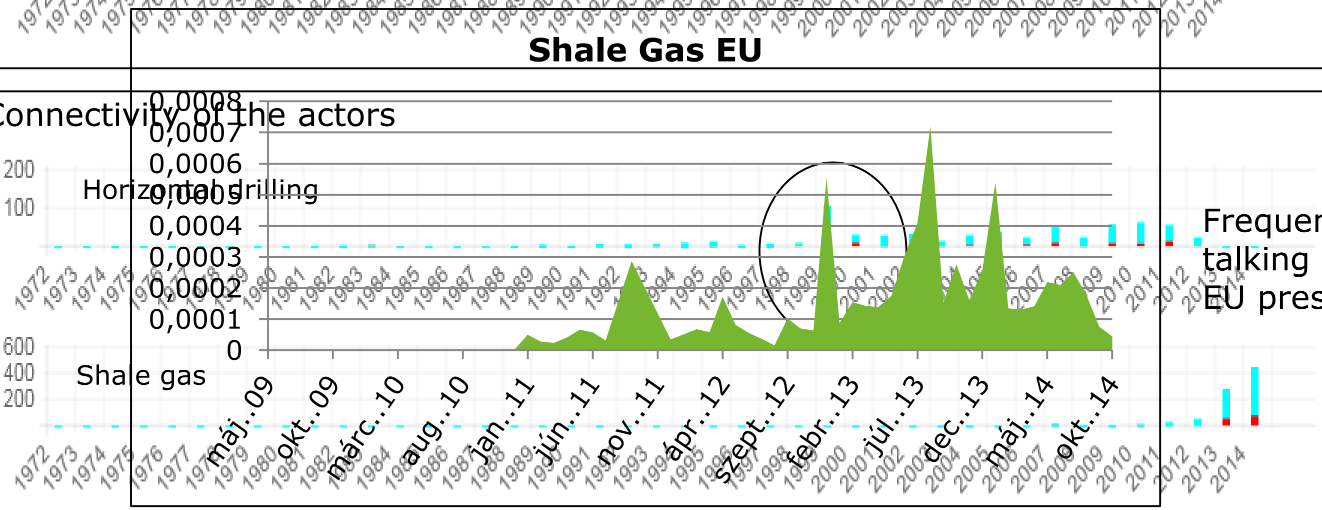
Early signals – Horizontal drilling Vs shale gas

Volume of publications



Volume of publications (blue: academic; red: industry)

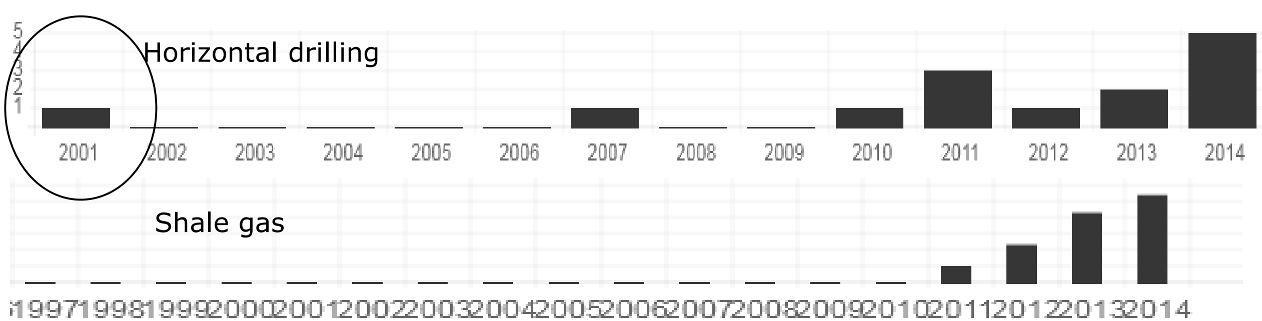
Connectivity of the actors



Frequency of articles talking about shale gas in EU press

Connectivity of the actors (blue: ind-ind or univ-univ; red: univ-ind)

New keywords or subject categories



Evolution of occurrences of subject category "Environment"

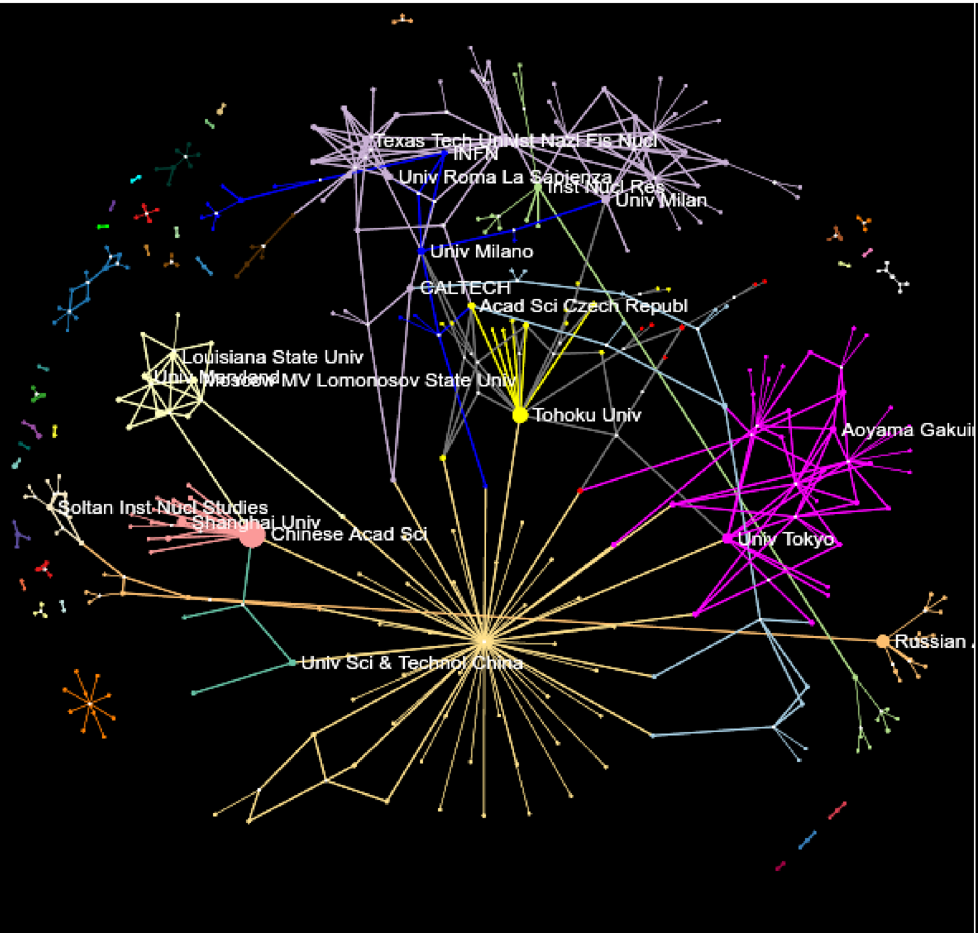
Courtesy: O. Eulaert/EC-JRC

Filtering

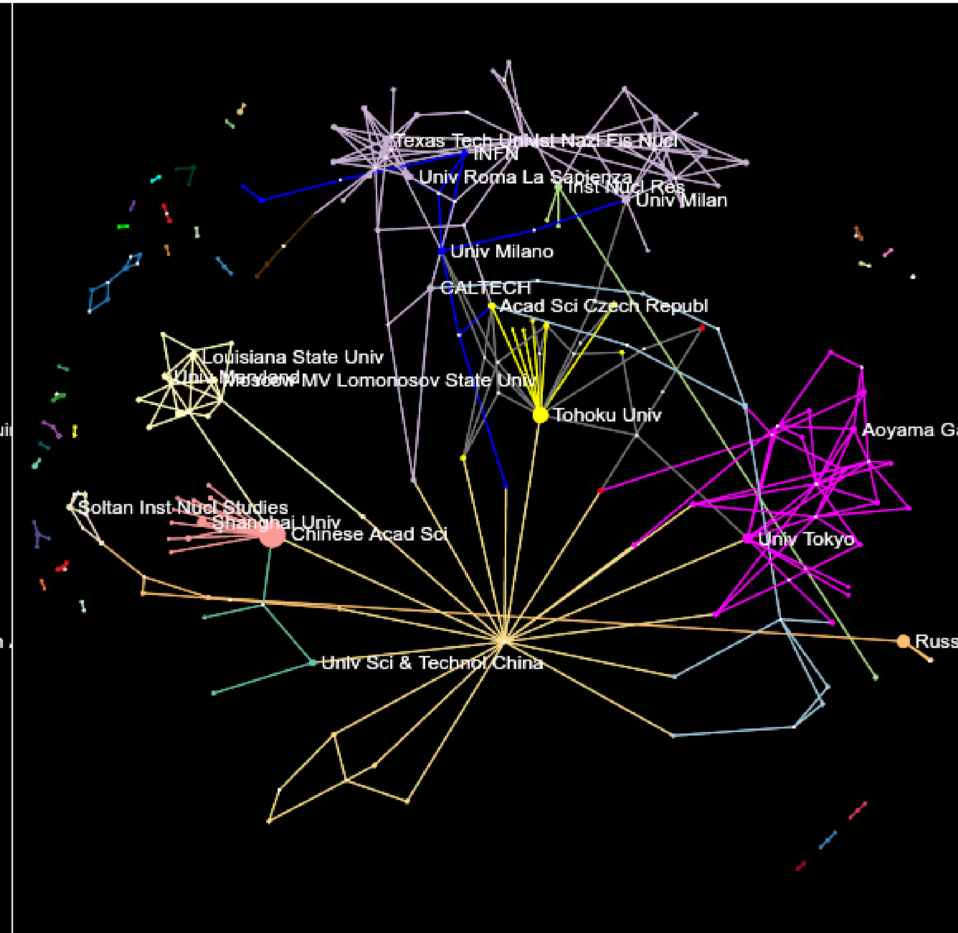
User-triggered operations

Operations on visual graphs

Filtering

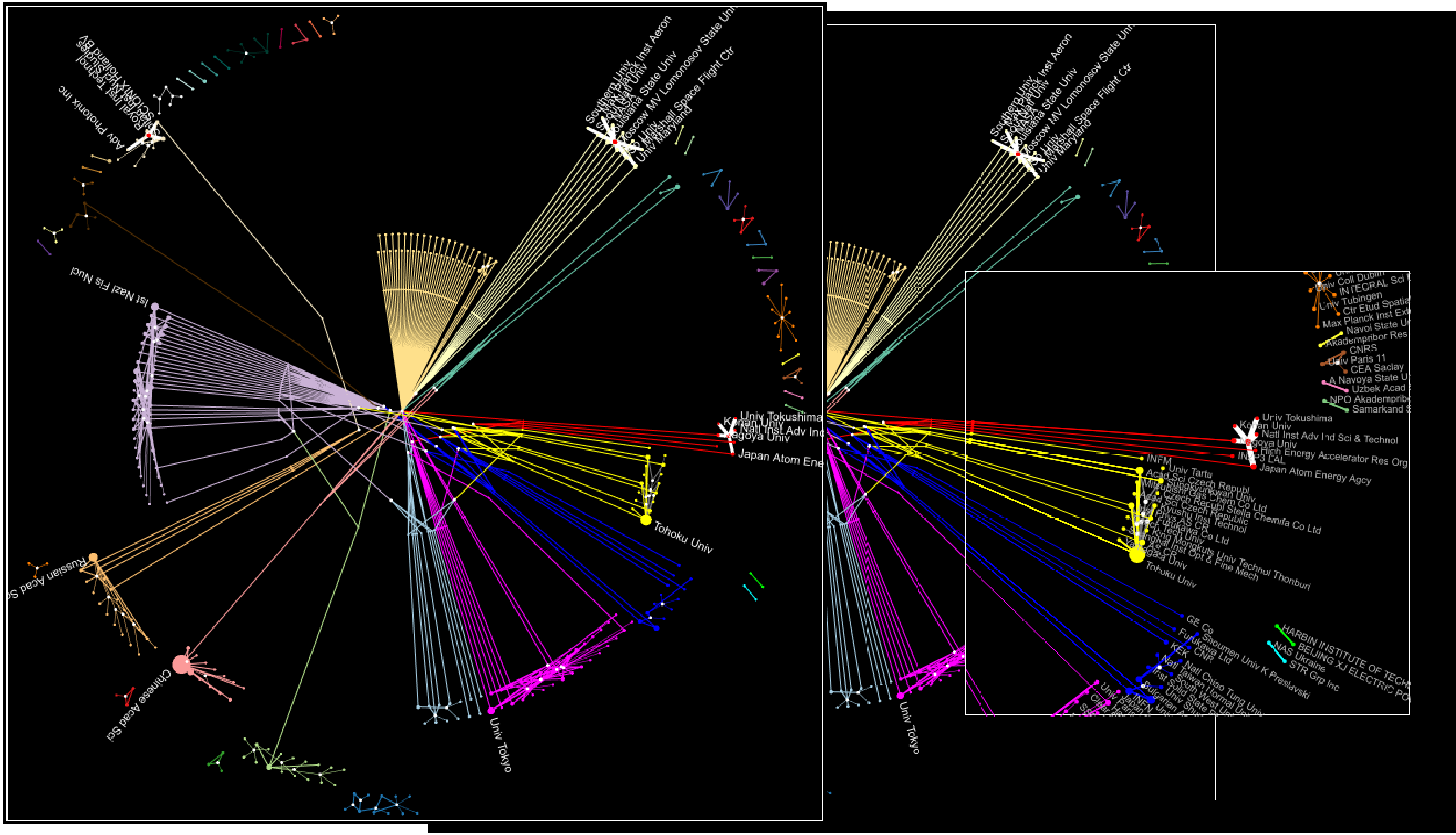


Organisation Landscape hypergraph view

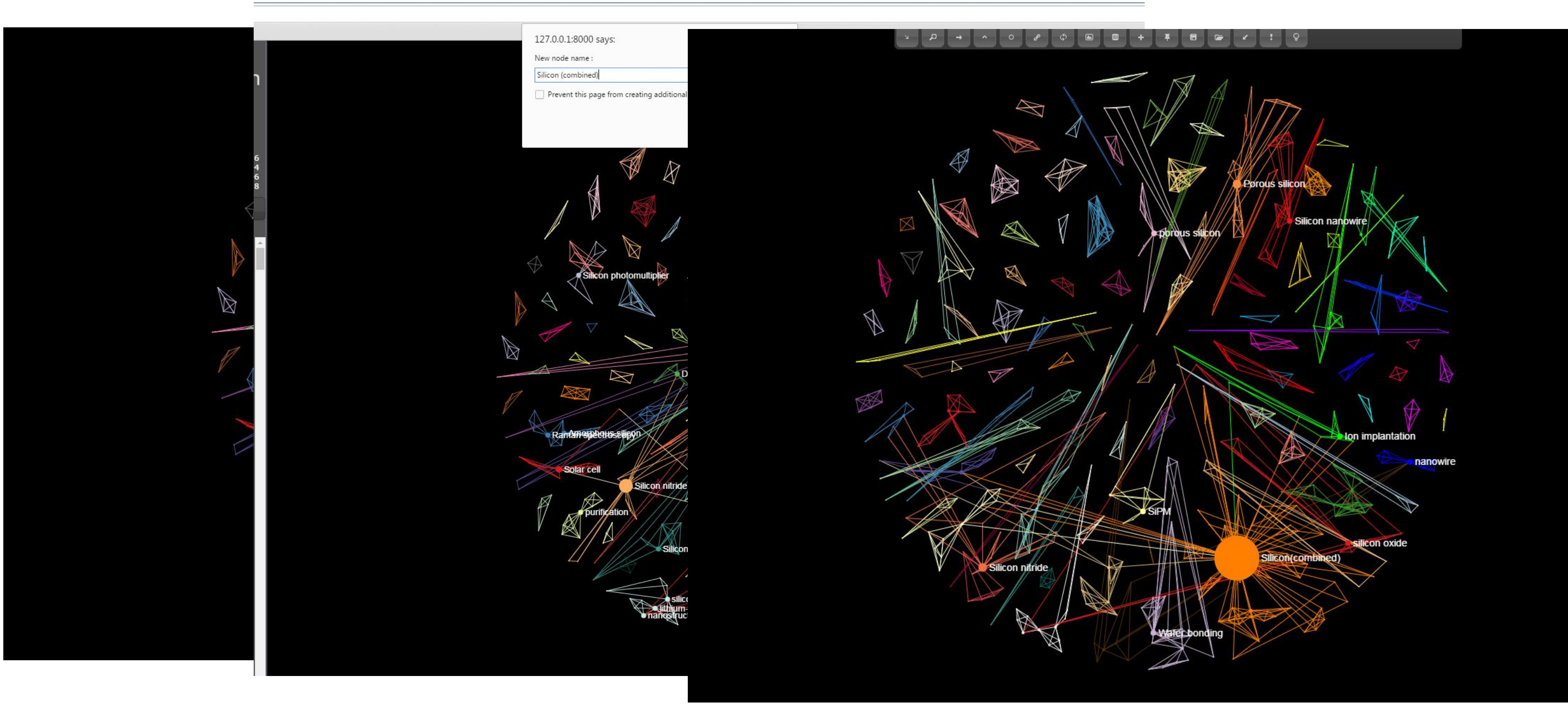


Organisation Landscape hypergraph view with Filter

Zooming



Merging vertices



User triggered operations are not affecting the data in the Graph DB!

Reachability Graph

Analytics dimensions

Visualisation dimensions

CS Modus Operandi

Processing Steps

I: Processing Search

- Retrieve list of matching publications & patents

II: Build collaborations for all visualisation graphs

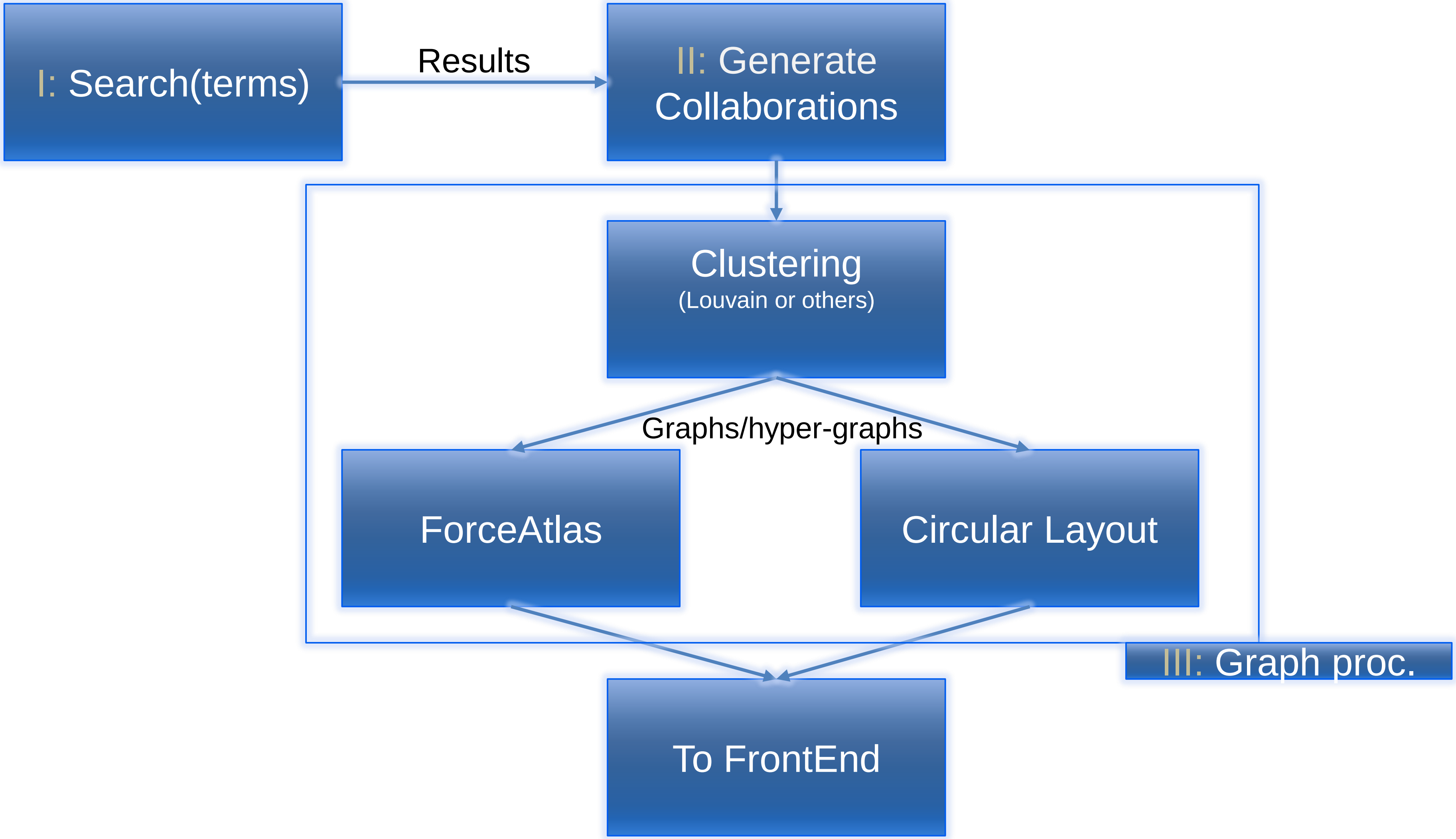
- List of sets of Vertices

III: Visual graph processing

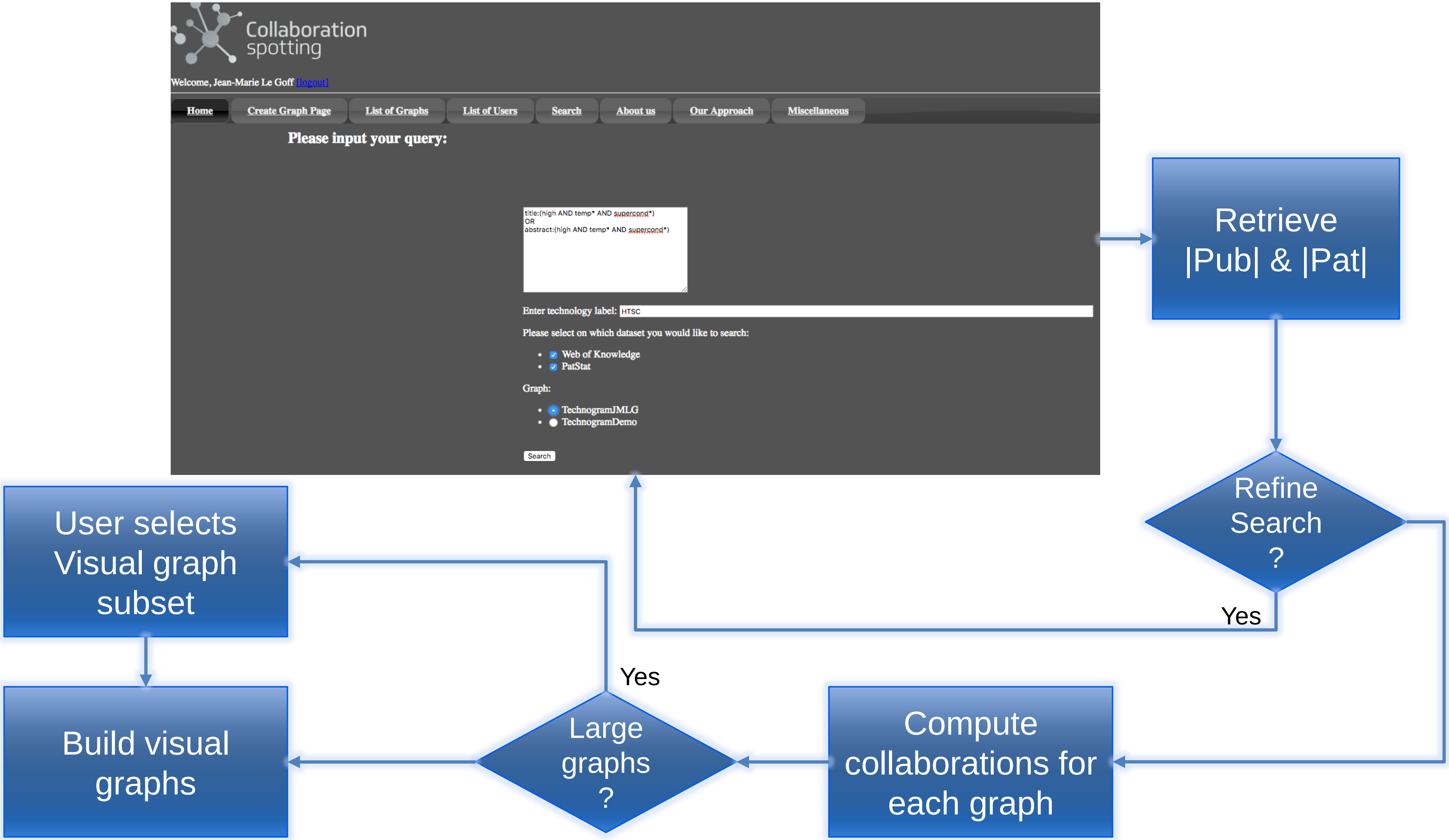
- Louvain
- ForceAtlas/Circular representations
- Clique Expansion View
- Extra Node Views

Processing sequence

(Building new visualization graphs from data analysis)

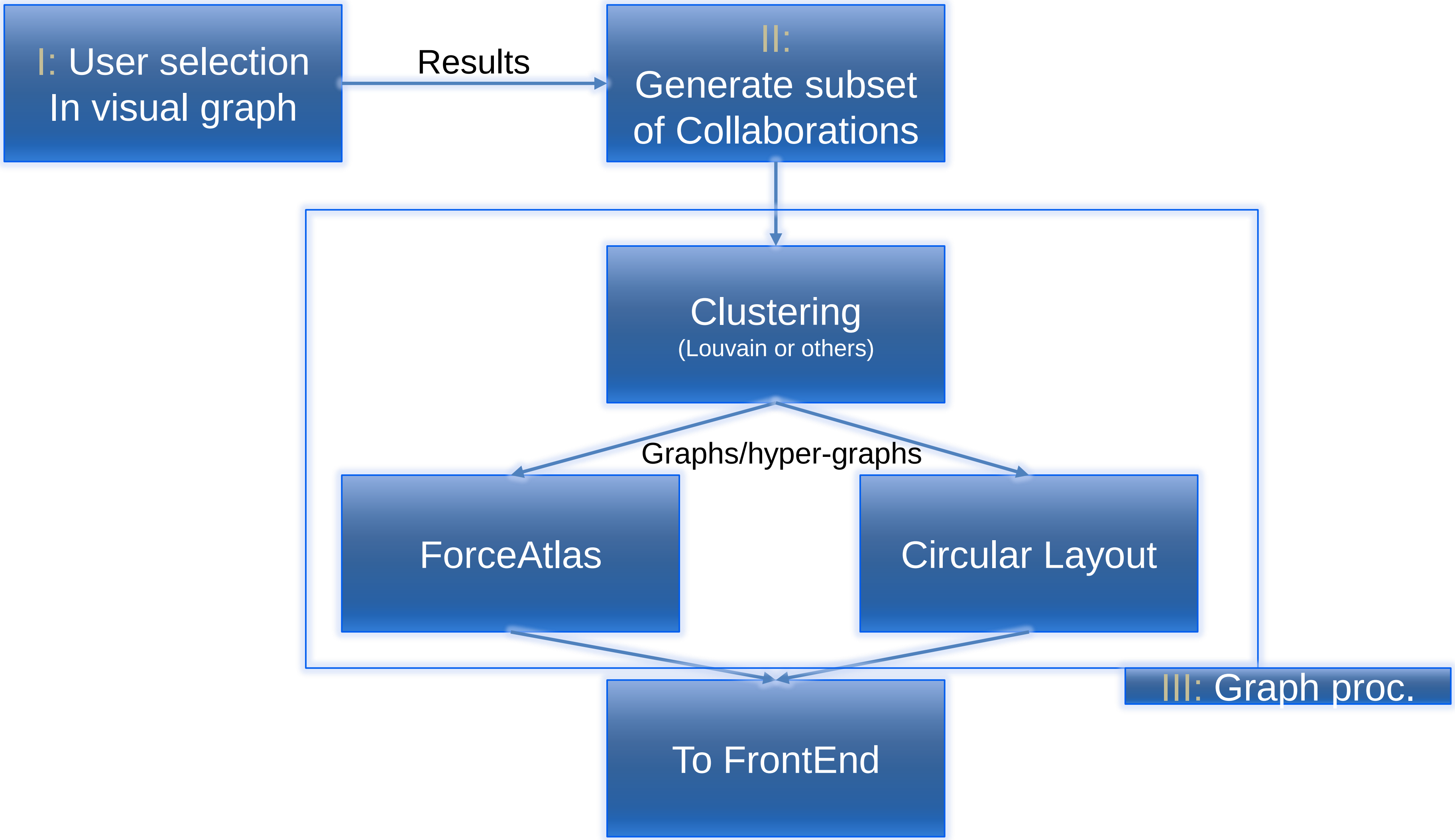


Processing Sequence



Processing sequence

(During user exploration)



Concurrent users

Data Analysis

Visual Interactions and Analysis

Performance Target

Performance target

- **Concurrent users**
  - Ex. 1: 100 interactive users
  - Ex. 2: TBD
- **User-triggered analysis** (~ seconds)
  - Ex. 1: Keyword- based search
  - Ex. 2: Compatibility search for a given data analysis process.
- **Visual interactions** (~ seconds, special provisions for large graphs)
  - Graphical queries (deterministic graph filtering)
  - Analysis of query results (incl. other visualisation modalities)

Neo4j graph DB

Cypher queries

Reachability Graph

Longer paths vs additional edges

Data records examples

Computational needs & optimization

Graph DB management and operations (A. Agocs)

Overview

- Neo4j DB
- Cypher + Visual operations
- Multi-navigation on the reachability graph
- Longer paths vs. additional edges

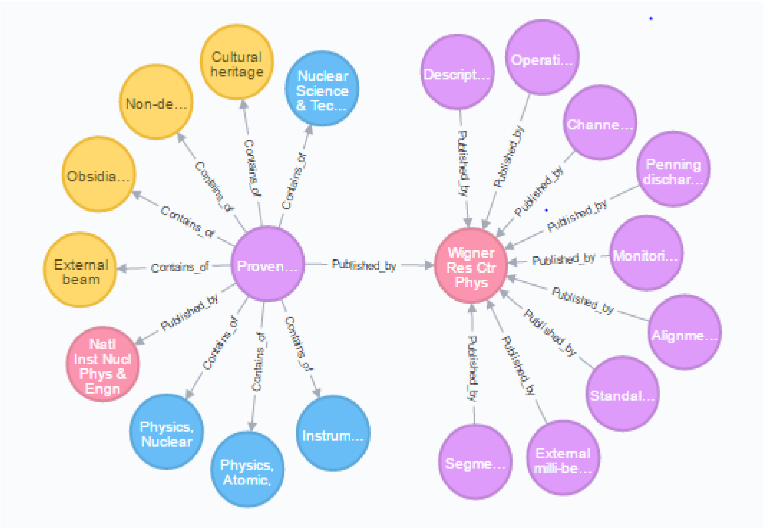
Neo4j database 1.

- Uses **Labelled Property Graph Data model**
- Nodes and relationships (with labels and properties) instead of tables, tuples and attributes
- Two big advantages:
  - Easier to adapt to another graph models (Feynman diagram)
  - Relationships can represent one-to-one, many-to-one, one-to-many and many-to-many relationships

Neo4j database 2.

Short example:

- Red: Organisations
- Purple: Publications or Patents
- Orange: Author Keywords
- Blue: Subject Categories



Neo4j database 3.

Type of nodes	Number of nodes
Patents	15.000.442
Publications	20.087.904
Organisations	2.918.060
Author Keywords	8.193.604
Subject Categories	230
Cities	7.741
Regions	946
Countries	128
$\sum$	46.209.055

	Patents	Publications	$\sum$
Organisations	12.440.903	36.672.677	49.113.580
Author Keywords	-	48.941.098	48.941.098
Subject Categories	-	32.566.806	32.566.806
Cities	3.193.709	8.826.222	12.019.931
Regions	265.421	2.504.441	2.769.862
Countries	3.156.449	8.020.648	11.177.097
$\sum$	19.056.482	137.531.892	156.588.374

Statistics on Patent & Publication database (2000-2014): Nodes (left) and edges (right)

Cypher

- Inspirited by SparQL (SQL based query language for semantic data, stored in RDF)
- The biggest advantages:
  - Pattern matching:
    - Define a subgraph via Cypher
    - Database finds all occurrences of it in the graph.
- Question: How should we create a pattern?

Cypher – Operation from GUI

- Support:
  - Selection on a graph view (GUI solves it)
  - Extension on a graph view (DB call)
  - Navigation from one graph view to another one. (DB call)
- We got from GUI: list of selected nodes and navigation purpose

Cypher – Pattern builder

- Goal 1: the pattern has to contain labels of the selected nodes + basic label + navigation label
- Goal 2: create a pattern which use the minimal amount of labels
- Solution: Steiner tree problem – NP-hard →use minimal spanning tree

Challenges on database level

- Data: 46M nodes + 156M edges (Pats&Pubs)
  - It can be increased by:
    - Using different sources
    - Using more dimensions (example: authors, journals, etc.)
    - Making time interval wider (from 2000-2014 to 1970-2016)
- Complex patterns (spec. with extra dimensions)
- Solution: Neo4j Cluster (Master-Clients)

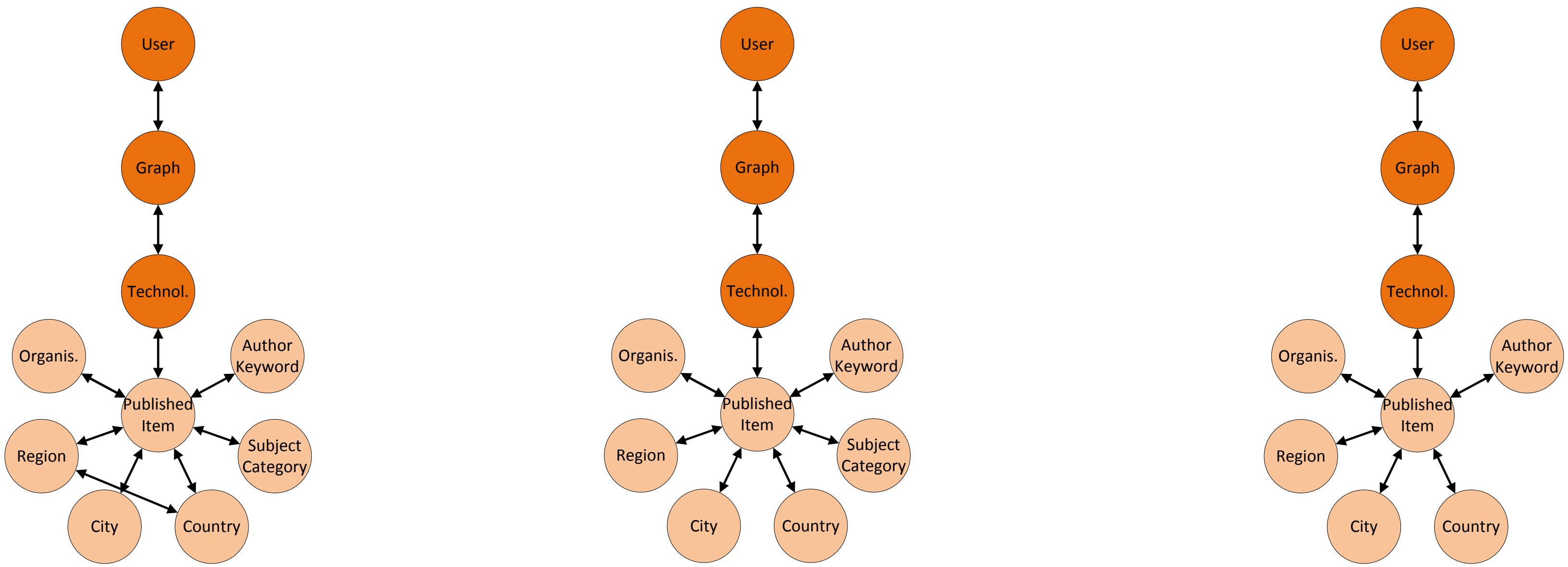
Multi-nav. on the reachability graph

Reachability graph:

- “Schema” of the graph based on node labels
- Pb1: Different paths from one node to another can mean different things:
  - Example: Publication & Patent; (EU) Region; Country
- Pb2: Sensitive data

Solution: Use multiple navigations

Multi-nav. on the reachability graph



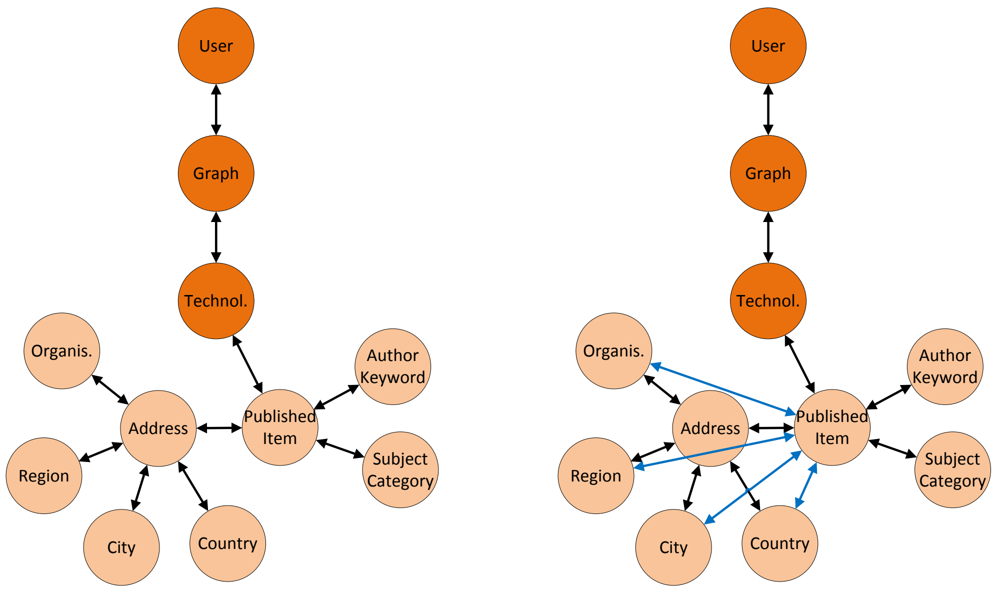
Ex: Modified Pub&Pat reachability graph with two navigation options

Long paths vs. additional rel.ships

- Long Paths:
    - Disadvantage: Costly
- Additional (generated) relationships:
    - Advantage: Solve execution time problem.

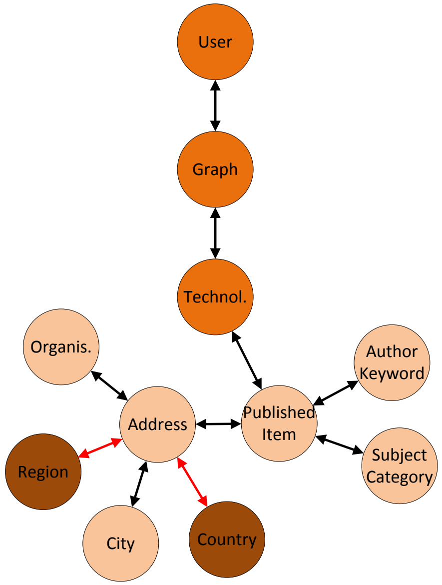
Problem: They answer two different questions.

Long paths vs. additional rel.ships



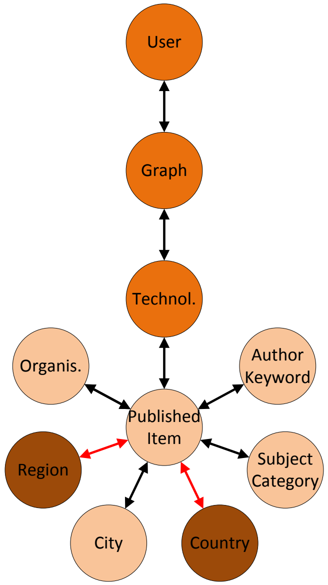
Ex: The future reachability graph (Pats&Pub DB) (left); Adding additional relationships (middle); Navigation with additional relationships (right)

Example: Long paths vs. add. rels.



- Question: How many regions does Hungary has?
- Answer: 11 (NUTS3)

Example: Long paths vs. add. rels.



- Question: How many EU regions does Hungary publish with?
- Answer: 198 (NUTS3)

Support user visual interactions
Collaborations
Louvain
ForceAtlas
Circular layout
Graph and hypergraph visual representations

Computational needs & optimization

Interactive visual graph processing (R. Forster)

AGENDA

- Graph generation
- Community Detection
- ForceAtlas
- Performance results
- Future work

GRAPH GENERATION

GRAPH GENERATION

- This is a process required for every single graph by any user
- First, database returned data needs to be transformed
- Have to generate:
  - Collaborations
  - Nodes
  - Edges

GRAPH GENERATION

	Silicon	Database	3D	CT
Collaborations	33,45	13,85	18,64	12,65
Nodes	0,69	0,65	2,48	0,38
Edges	1,48	1,87	0,71	0,69

Computation time for specific parts of the graph generation in seconds

COMMUNITY DETECTION

COMMUNITY DETECTION

- Used to reveal groups in real world data
- Louvain method
- Parallel heuristics

LOUVAIN METHOD

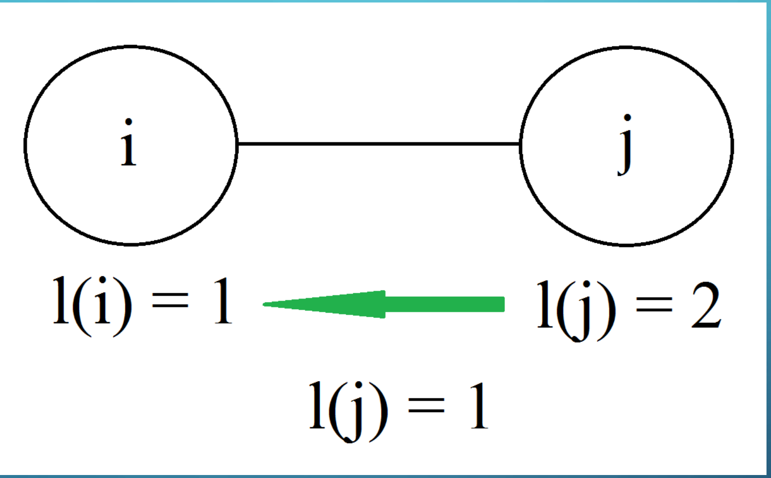
- Multi-phase, iterative, greedy algorithm
- Monotonically increasing modularity
- Inherently sequential

LOUVAIN PARALLEL HEURISTICS

- **Singlet minimum label heuristic**
- **Generalized minimum label heuristic**

SINGLET MINIMUM LABEL HEURISTIC

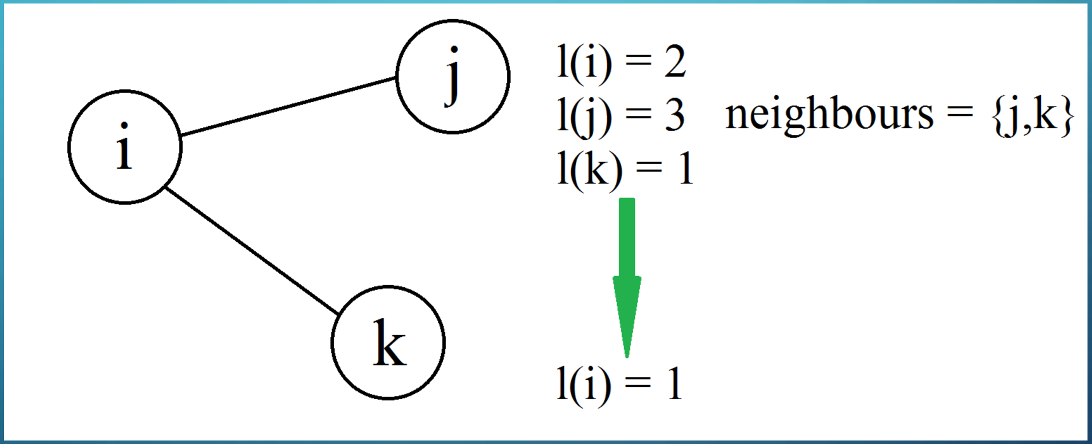
- Move node (j) only if: $l(C(i)) < l(C(j))$

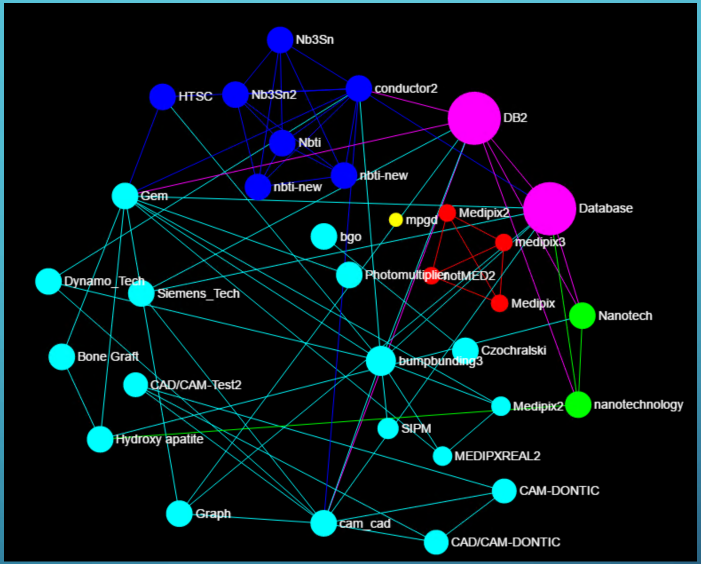


GENERALIZED MINIMUM LABEL HEURISTIC

- Move node (i) to neighbour community only if:

$$\min_{n \in neighbours} l\big(C(n)\big) < l(C(i))$$




Nb3Sn
HTSC
Nb3Sn2
conductor2
DB2
Nbti
nbti-new
nbti-new
Gem
Database
mpgd
Medipix2
bgo
medipix3
Dynamo_Tech
Siemens_Tech
Photomultiplier
notMED2
Medipix
Nanotech
Bone Graft
Czochralski
bumpbunding3
CAD/CAM-Test2
Medipix2
nanotechnology
Hydroxy apatite
SIPM
MEDIPXREAL2
CAM-DONTIC
Graph
cam_cad
CAD/CAM-DONTIC

FORCEATLAS

FORCEATLAS

- Force-directed layout based on n-body simulation
- Repulsion-attraction
- Makes visual interpretation easier
- Result depends on starting state

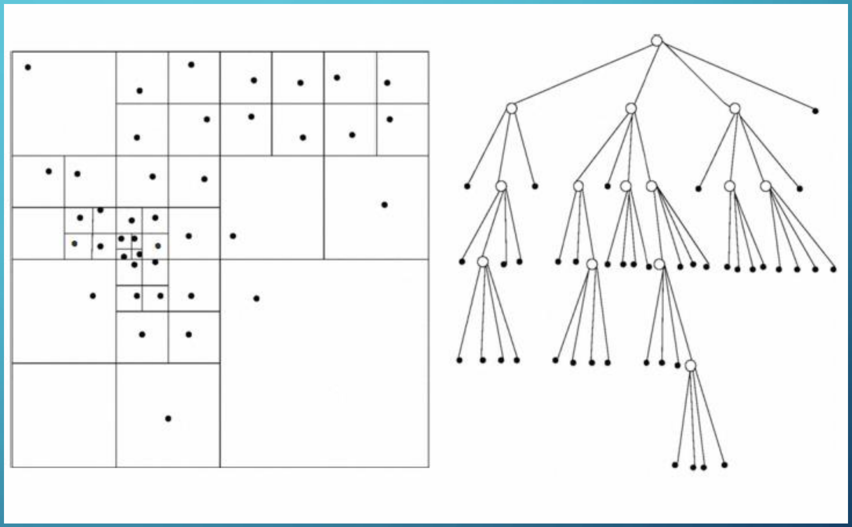
PARALLEL FORCEATLAS

- Repulsion in parallel
- Attraction in parallel

PARALLEL REPULSION

- Barnes-Hut algorithm
- Repulsion on regions
  - Complexity:

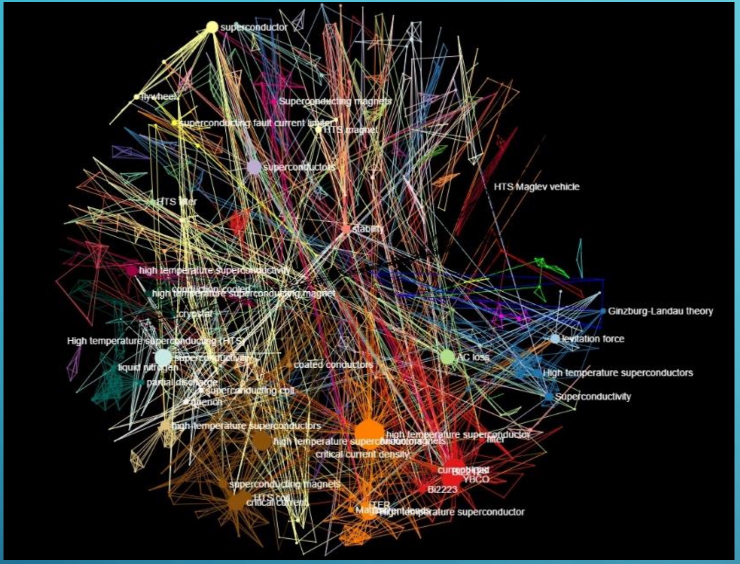
    O(N*LOG(N)) instead of O(N^2)



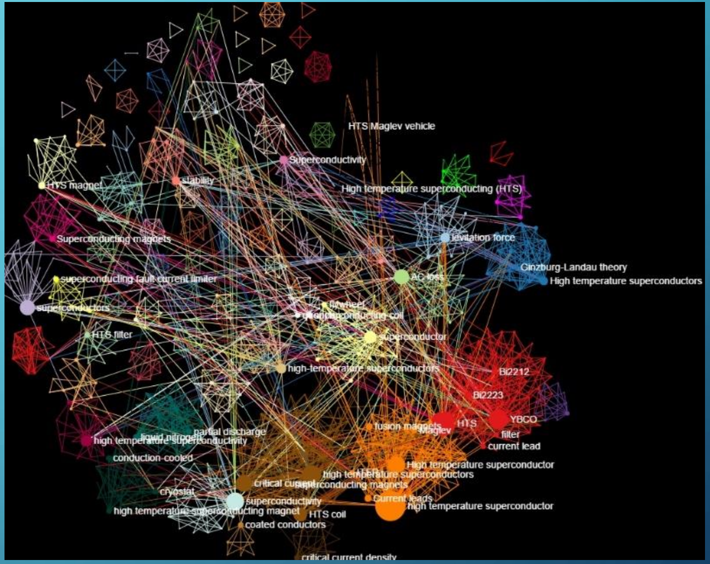
PARALLEL ATTRACTION

- Connected nodes are attracted to each other
- Attraction intensity controlled by the user

FORCEATLAS LAYOUT TYPES



Original layout



Community based layout

Search: in title or abstract (high AND temp* AND supercond*)

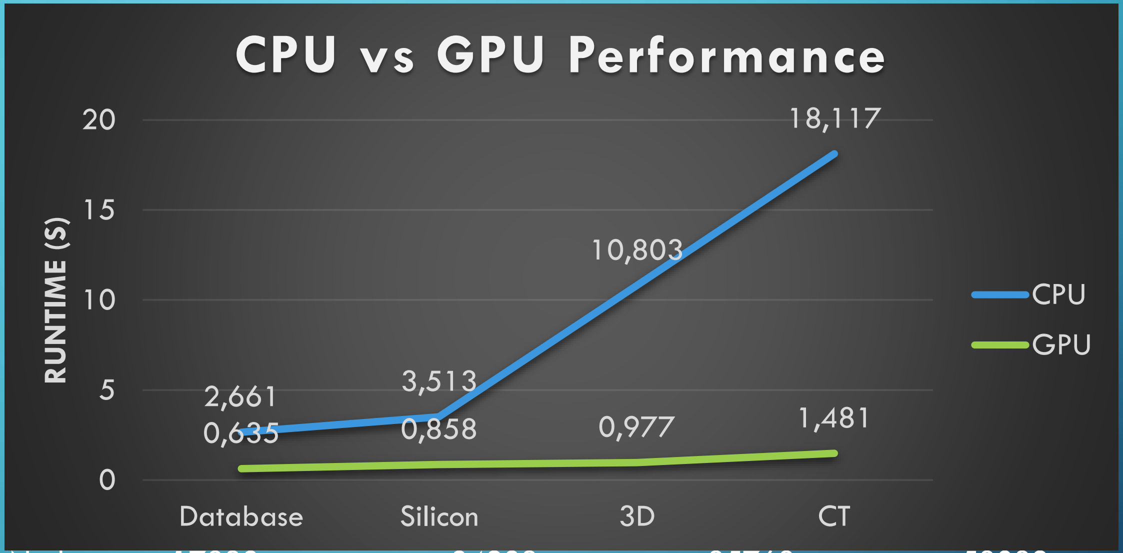
PERFORMANCE RESULTS

PERFORMANCE RESULTS

- Graphs used: *3D, CT, Database* and *Silicon* technologies
- Implementation in C++
- 8 threads in parallel execution
- System:
  - Intel Core i7 4710HQ
  - 24 GB DDR3
  - GeForce GTX 980M

RESULTS
(COMMUNITY DETECTION)

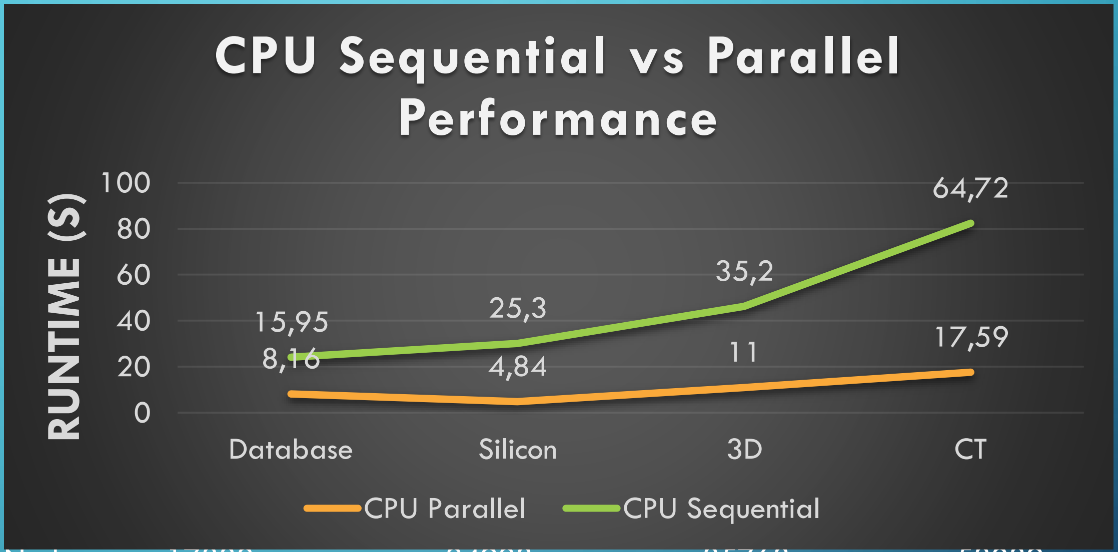
- Database: 4x
- Silicon: 4x
- 3D: 11x
- CT: 12x



Nodes	17832	24923	35763	53039
Edges	113622	185594	192336	325490

RESULTS
(FORCEATLAS)

- Database: 1,95x
- Silicon: 5,22x
- 3D: 3,2x
- CT: 3,68x



Nodes	17832	24923	35763	53039
Edges	113622	185594	192336	325490

FUTURE WORK

- Further optimization of the graph data generation
- GPU implementation of ForceAtlas

Explore other data sources

Advanced analytics

Future Work

Deployment & collaborations

- CollSpotting on Publications and Patents under tests on Open Stack @ CERN
  - Limited performance; could be improved with GPUs
  - Cores with max RAM to store graph of DB in memory needed
- System will be available to HEPTech members
  - CERN login required (Licence issues)
  - Additional data sources according to HEPTech needs
- Exploration of collaborations with CERN
  - EPFL → Big Data to be identified
  - UN-UNICRI (Interregional Crime and Justice Research Institute) → Big Data analytics to reinforce security

Advanced visual analytics

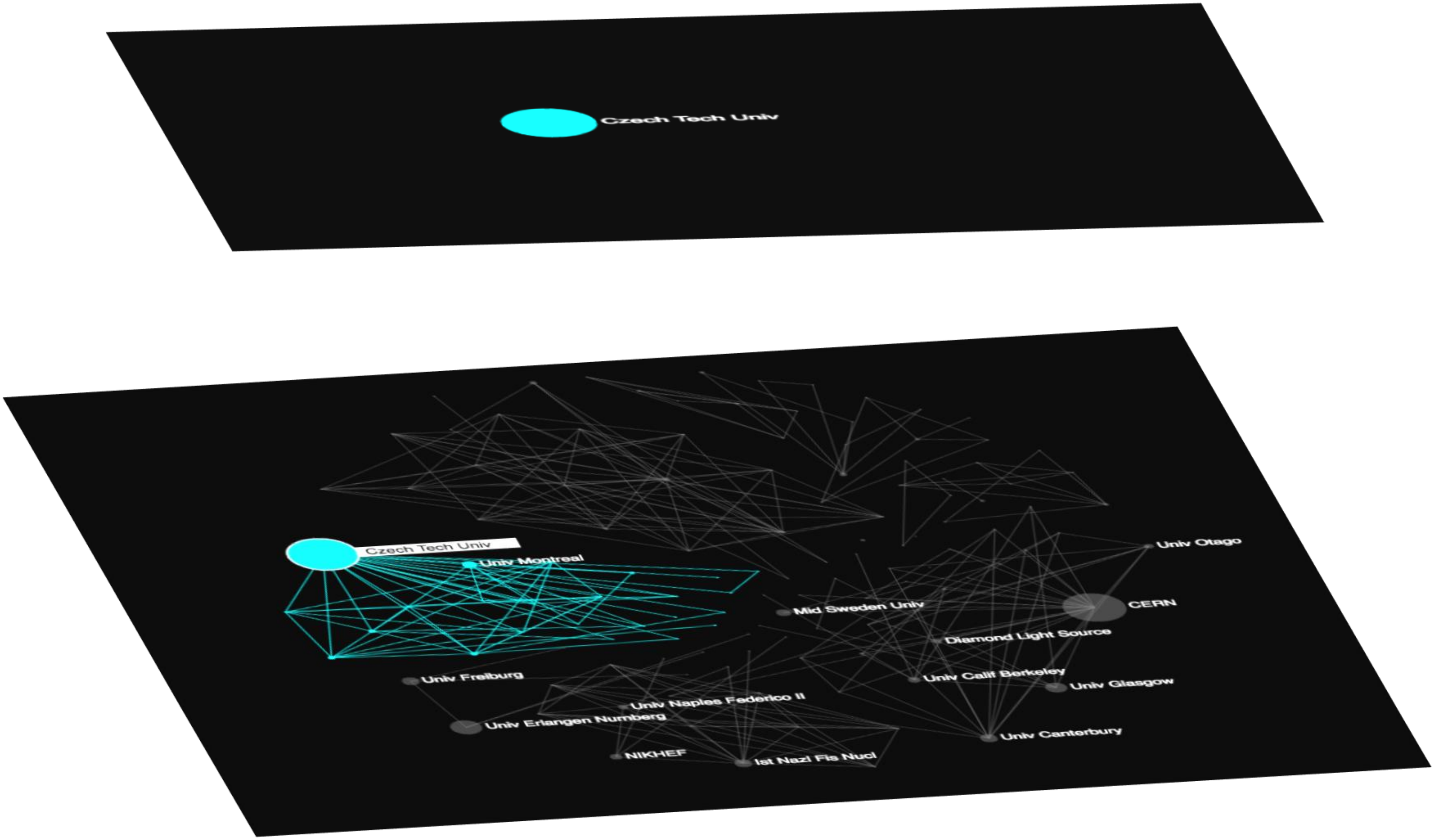
- **➔More value to users when analysing Clusters**
- **Contextual analytics**
  - **While navigating across dimensions**
  - **While performing operations on visual graphs**
- Support to the visualisation of very large graphs in the node-link representation
- Multi-dimensional visual graphs and cluster optimization
- Dynamic visual dimensions selection to optimize cluster visualisation
- Major performance optimization efforts to maintain acceptable processing time for users, irrespective of the graph's size.

Compound graphs(*)

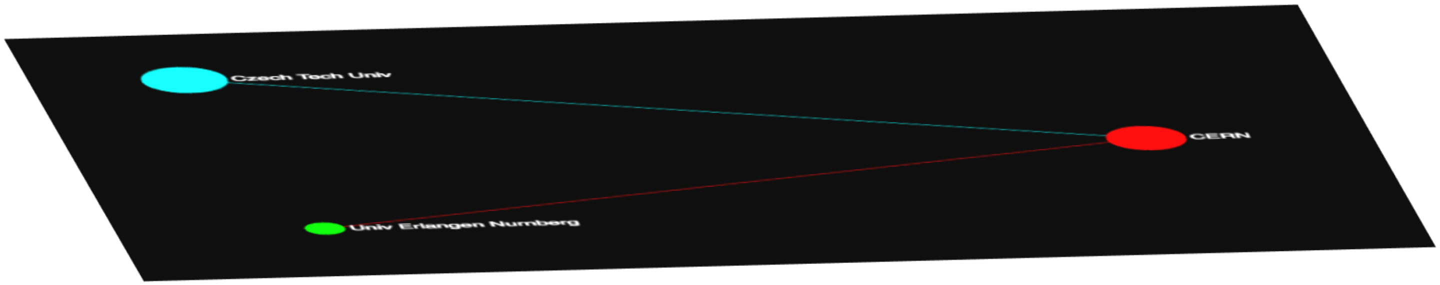
- **<u>Compound graphs</u>**: C=(G,T) is defined as a graph $G=(V, E_G)$ and a rooted tree $T=(V, E_T, r)$ that share the same set of vertices such as:
- $\forall e = (v_1, v_2) \in E_G, v_1 \notin \ path_T(r, v_2) \land v_2 \notin \ path_T(r, v_1)$
- Relationships between vertices are expressed by T:
- Vertices sharing a common parent in T belong to the same group.
- When two vertices sharing a common parent are connected in G, they share a generic relationship.

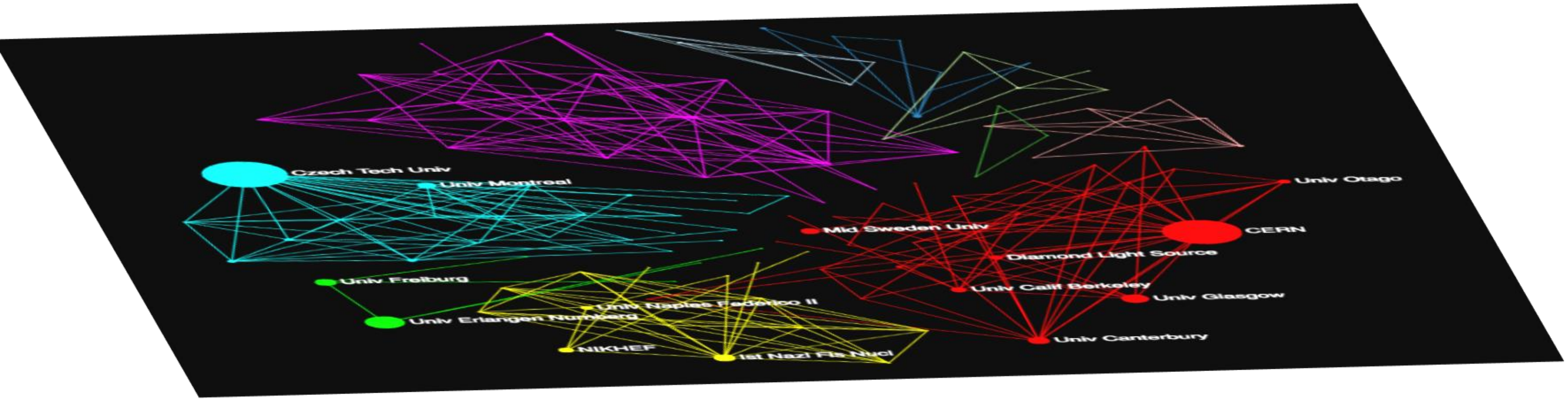
(*)KEIM D., ANDRIENKO G., FEKETE J.-D., GÖRG C., KOHLHAMMER J., MELANCON G.: *Information Visualiza- tion*, vol. 4950 of *Lecture Notes in Computer Science*. Springer, 2008, ch. Visual Analytics: Definition, Process, and Challenges, pp. 154–175.

Map Cluster to single vertex

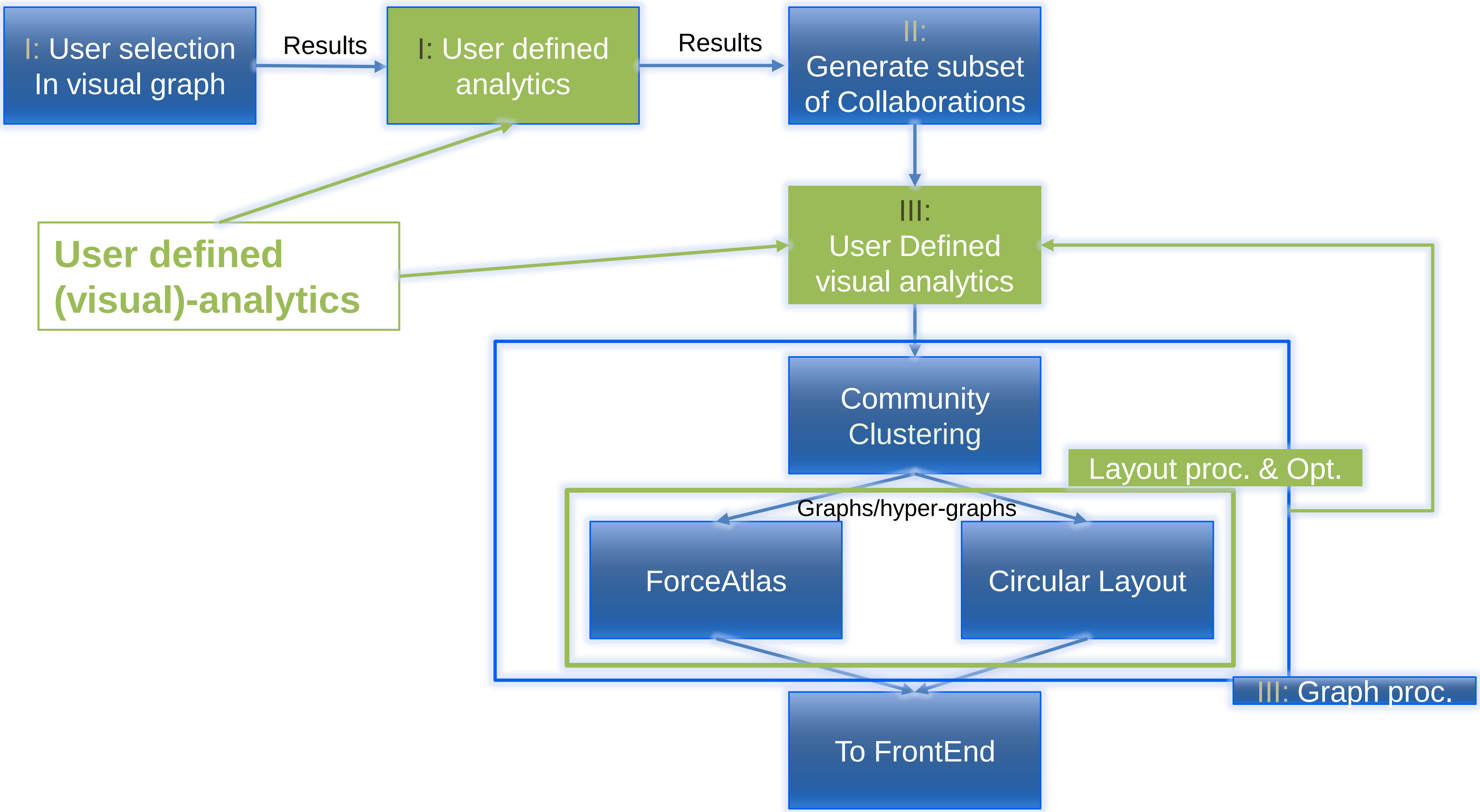


Replacing clusters with vertices





Introducing Analytics in the visualisation process



Explore other datasets

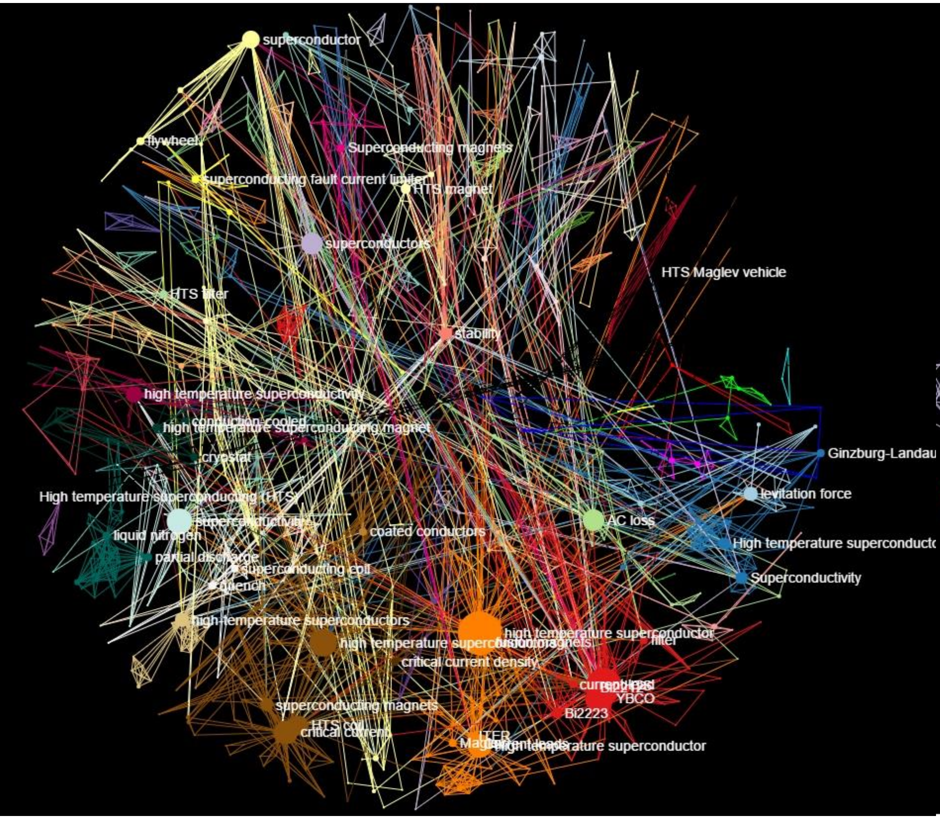
- To validate development and support future work we need to team with domain experts to gather **Use Cases,** **in particular** **to explore multi-dimensional visual analytics**
- Possible sources (Collaboration required!)
  - Particle Physics
  - Biotech
    - Ex: Phenotyping

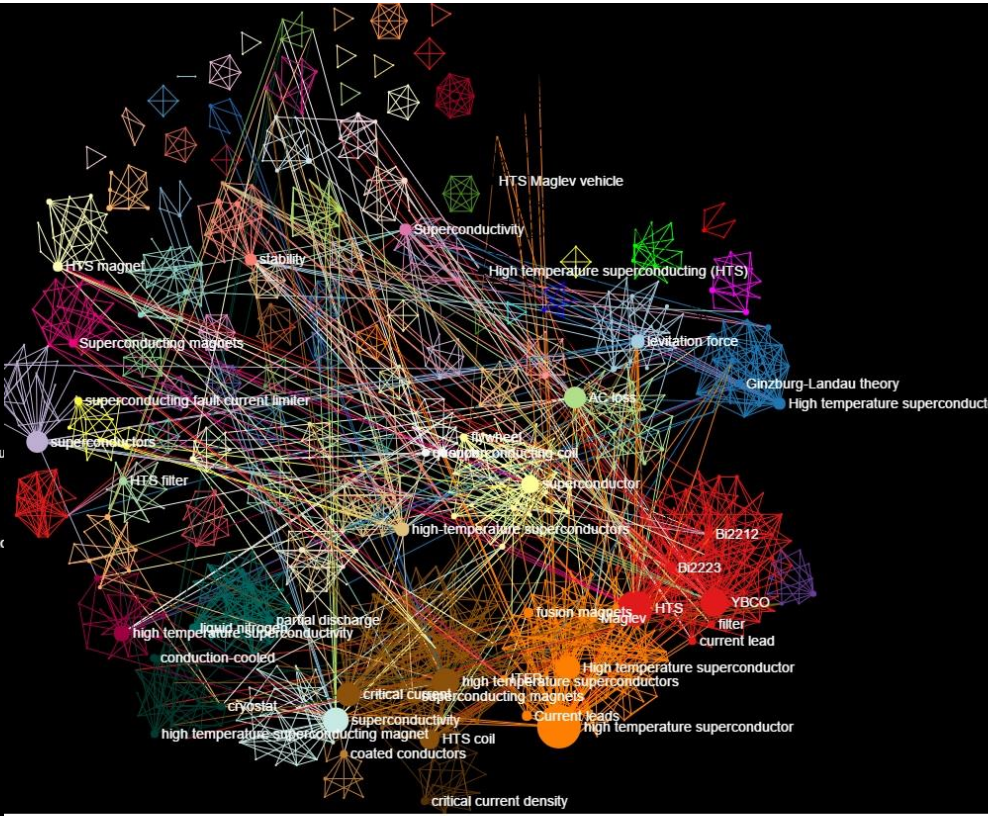
Conclusion

- Big Data visual Analytics bring the expert at the centre of the analytics cycle
- Interactive and multi-dimensional graph-based tools provide strong support to analytics
- **Techniques applicable to any kind of data**!

Thank you for your attention!

ForceAtlas is computing savvy





Different computation strategies on graphs of keywords

CERN

www.cern.ch

Exploiting graph relationships