

GENE_629
GENE_641 GENE_700 GENE_504
GENE_731 GENE_737 GENE_631 GENE_457
NE_896 GENE_768 GENE_861 GENE_314 GENE_581 GENE_438
NE_816 GENE_944 GENE_426 GENE_434 GENE_301 GENE_236 GENE_4
NE_416 GENE_797 GENE_625 GENE_471 GENE_906 GENE_199 GENE_2
NE_108 GENE_830 GENE_678 GENE_419 GENE_872 GENE_219 GENE_1
NE_305 GENE_929 GENE_597 GENE_914 Disease_50 GENE_592 GENE_1
NE_935 GENE_601 GENE_200 GENE_277 GENE_660 GENE_917 GENE_1
NE_807 GENE_958 GENE_359 GENE_253 Disease_23 GENE_832 GENE_1
NE_436 GENE_105 GENE_255 Disease_9 GENE_327 Disease_47 Disea
E_892 GENE_349 GENE_608 GENE_386 GENE_688 GENE_828 GENE_1
GENE_466 GENE_155 Disease_87 GENE_248 GENE_525
GENE_682 Disease_52 Disease_29
Disease_52 Disease_70
570 GENE_775
313 GENE_550 Disease_40 GENE_129 GENE_463 GENE_779 GEN
NE_406 GENE_781 GENE_909 GENE_462 GENE_742 GENE_334 GENE_1
NE_393 GENE_908 GENE_727 GENE_560 GENE_519 GENE_789 GENE_8
ease_57 GENE_451 GENE_197 Disease_6 GENE_812 GENE_421 GENE_9
NE_716 GENE_256 GENE_712 GENE_405 GENE_278 GENE_744 GENE_1
NE_211 GENE_494 GENE_808 GENE_875 GENE_622 GENE_884 GENE_3
NE_500 GENE_395 GENE_748 GENE_901 GENE_561 GENE_829 GENE_9
NE_865 GENE_798 GENE_614 Disease_89 GENE_632 GENE_510 GENE_1
NE_181 GENE_928 GENE_799 GENE_479 GENE_138 GENE_370 GENE_2
GENE_195 GENE_150 GENE_539 GENE_168 GENE_254

Case Studies:

Support of Target Discovery Research by KNIME Textprocessing and Networkmining

Frank Dullweber

Is KNIME a useful tool to support target discovery research?

Our Goal:

We want to process and visualise information from external resources ...

- a) ... to structure and group data for colleagues.
- b) ... to support knowledge generation by a scientist responsible for a new indication.

This presentation:

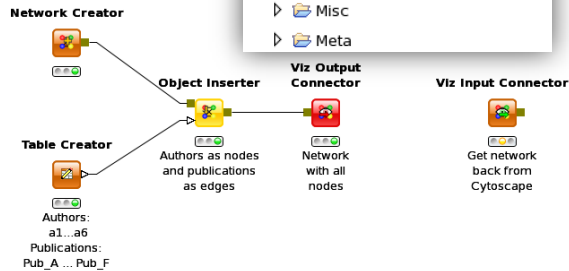
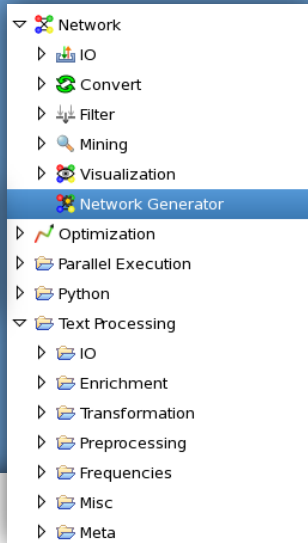
I will discuss four use cases during this presentation.

KNIME Networking & Textprocessing is used as a tool.

What are the technical prerequisites ?



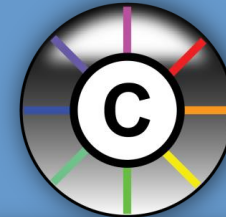
KNIME Labs Nodes



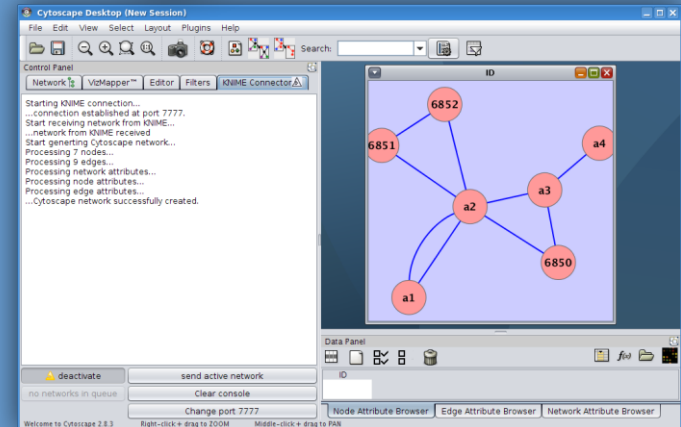
KNIME
Plugin

for

Cytoscape



Cytoscape
V2.8.3



Data Processing

Data Visualisation

1. Pubmed Query:
“my disease term”[all fields])OR “my abbreviation”[Title/Abstract])
2. Download results via PubMed webpage OR load results directly in KNIME
3. Attention: Each reader has its advantages and disadvantages !

Document Grabber




Read directly from
PubMed

File Reader




Read downloaded
CSV file

XML Reader




Read downloaded
XML file

1. Gene Dictionary

We use our own dictionary with about 2 mio. entries.

Synonyms and identifiers for 7 species are mapped to (human) EntrezGeneIDs and (human) StandardGeneNames.

2. Disease Dictionary

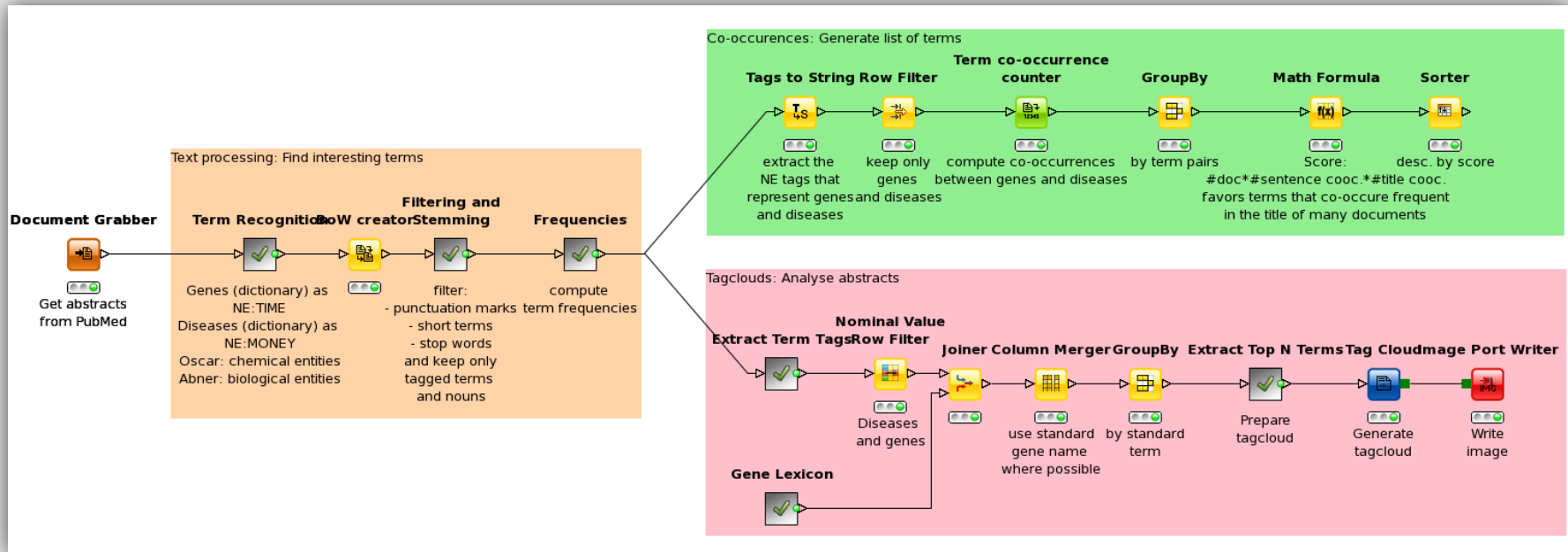
There are several free resources available:
MeSH, ICD-10, Others:

http://www.vistainformatics.com:8000/ben_disease_list.html

Real example: About 3700 (20.000) articles were found in PubMed.

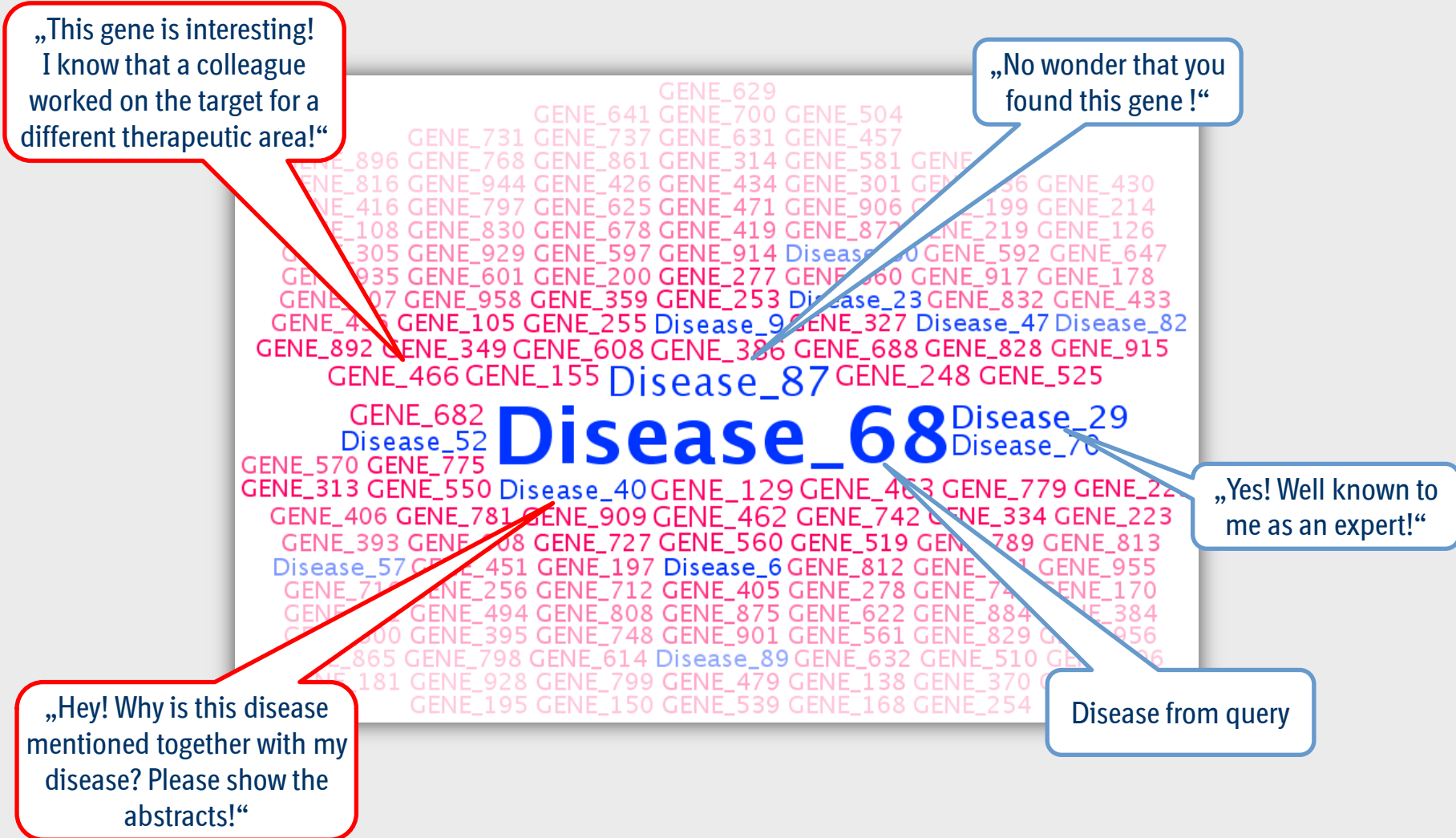
These articles are the basis of all the following analyses, but are blinded in this presentation.

Use Case 1: How can we process the information with the KNIME Labs nodes?

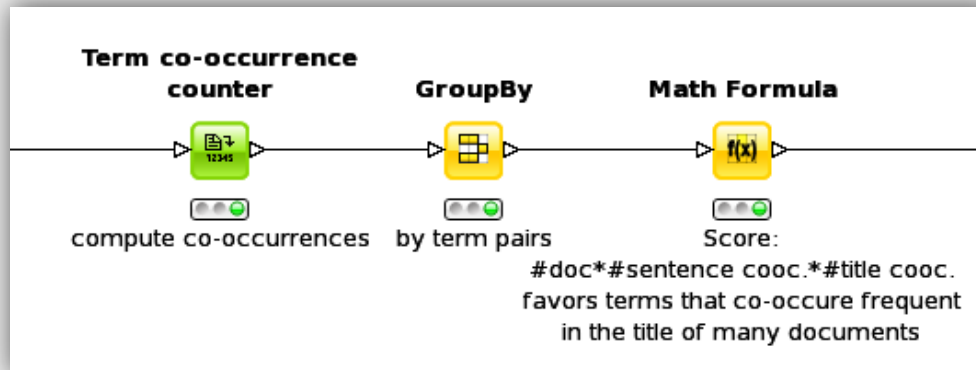


1. *Document Grabber* – to get PubMed article
2. *Term Recognition* – to find terms in articles
3. Postprocessing & Frequencies
4. Data analysis
 - a. *Term co-occurrence counter*
 - b. *Tag Cloud*

Use Case 1: How can a tagcloud structure information to support knowledge generation ?



Use Case 1: What kind of information can be derived from co-occurrences of terms in PubMed articles?



Term1	Term2	Sentence cooccurrence	Neighbour count	Title cooccurrence	Orig Document	Score *
.....	my disease	202	0	24	185	934'250
carcinoma	my disease	69	0	12	58	52'026
..	obesity	53	0	9	48	20'352

* Score = Sentence cooccurrence * (1+Title cooccurrence) * Orig Document

„Hey! Why is this disease mentioned together with my disease? Please show the abstracts!“

„No wonder! “

1. Construct a suitable PubMed query
2. Download hits as CSV file



Read CSV
from pubmed

 minimum 2
(connected with at least 2
authors)



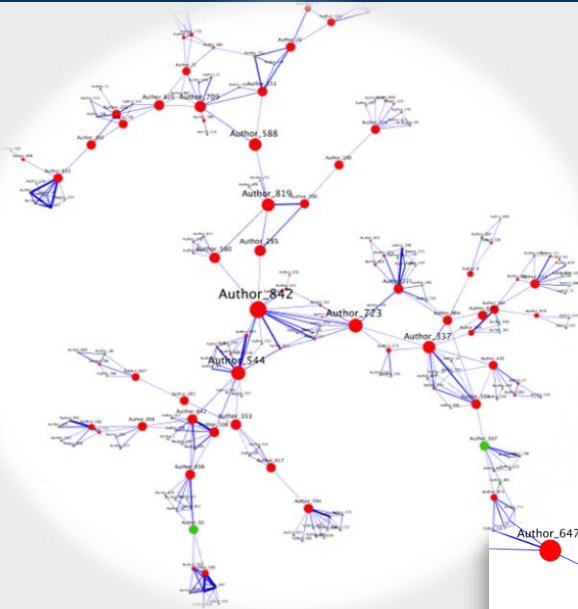
Cytoscape

1. Use CSV file from PubMed, not *Document Grabber* node
2. Review extracted author names (!)
3. Insert features like *Max.Year*, *#publications*,

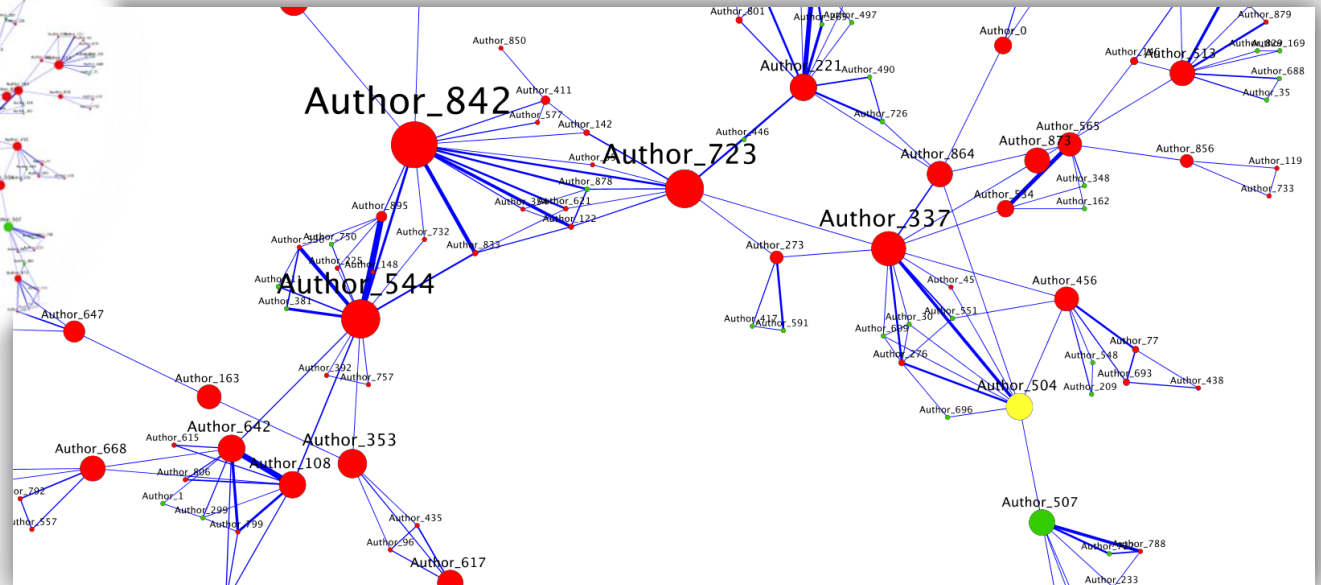


Use Case 2: What are the „most important“ authors for our new indication?

Starting with already known authors ! Who are the key opinion leader?



edge width: #publications
node size: betweenness centrality
node color: last year of publication



Shortest Path

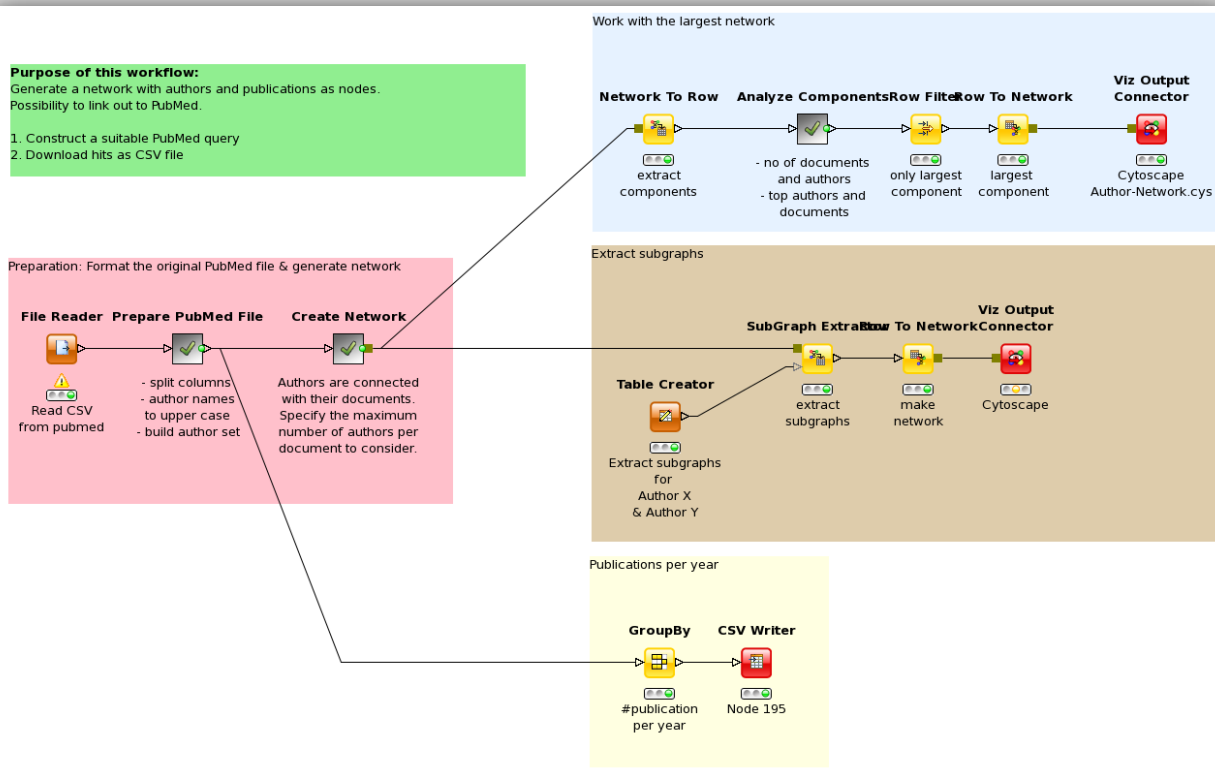


Shortest
path

Author_842 Author_723 Author_337 Author_504

6 2 9

Use Case 3: We want to visualise the authors together with all their publications !

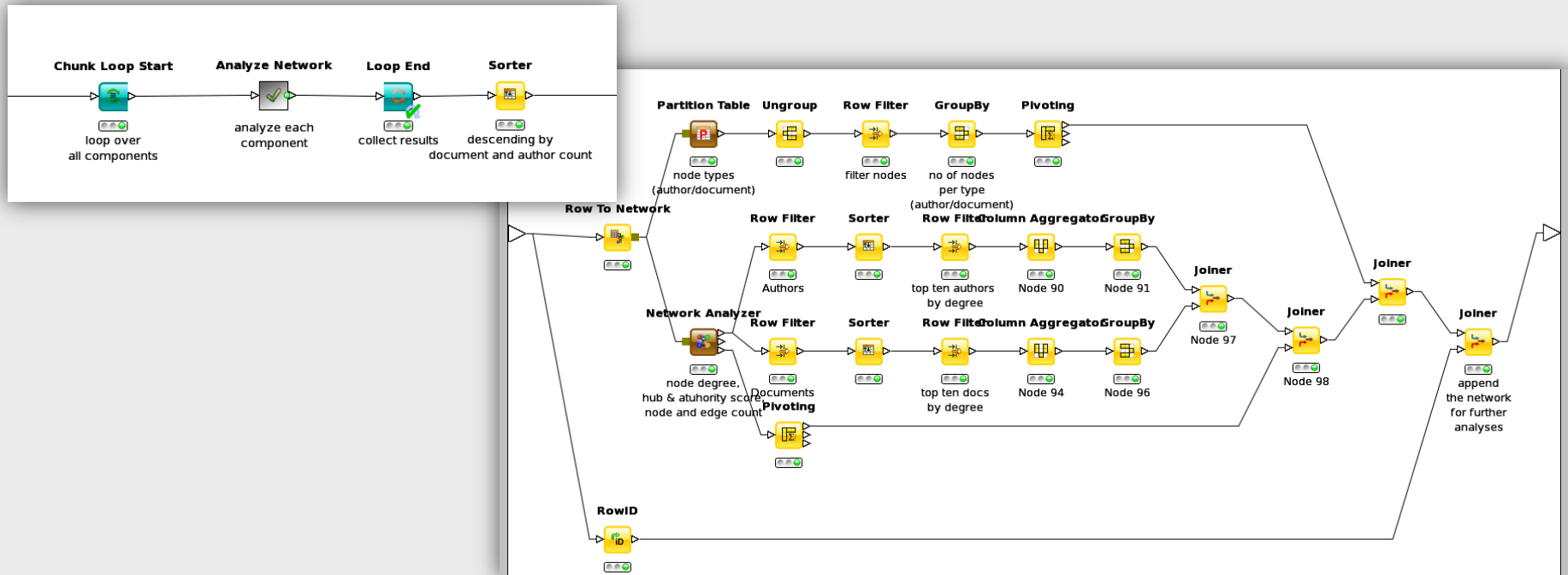


Analyse network components and find top10 authors for each component !

Extract subgraphs for your already known authors!

Authors as nodes & PubMed articles as nodes in a network

Use Case 3: How can we analyse the different components of the network ?

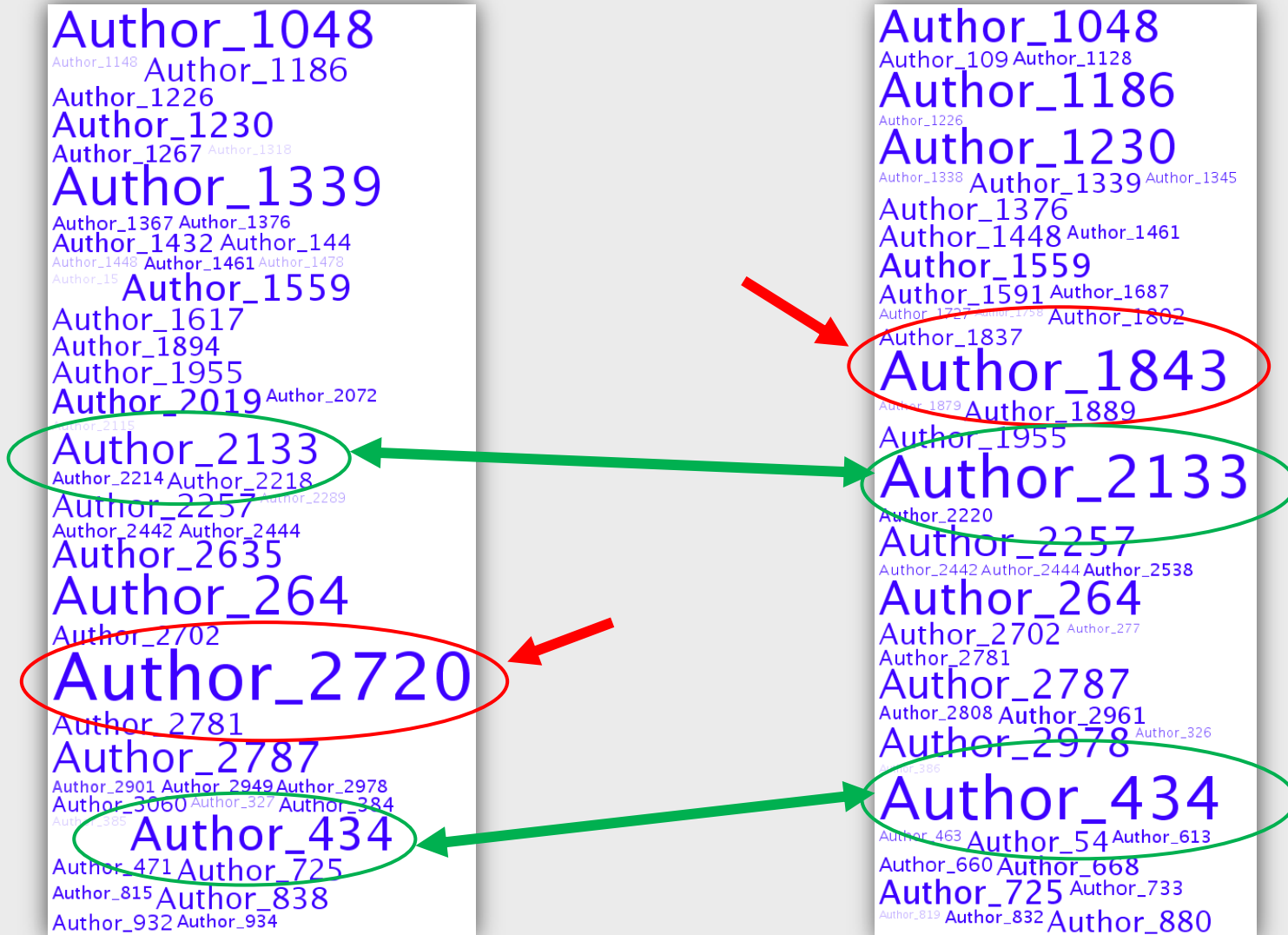


#author	#publications	Top 10 authors	Network statistics	Network in row
4089	2279	[Meyer: 30, ...]	...	...
45	24	[Smith: 20]	...	...
(834 nets altogether)				

Use Case 4: Are author with high numbers of publications are the „most important“ authors?

Node Degree: „#publication“

Closeness Centrality: „networking“



- Networks can integrate your information and can add a new perspective to your data („*not just tables*“).
- KNIME *Networkmining & Textprocessing* can support your work together with Cytoscape due to the existing algorithms, **but ...**
- ... you will have to bring your own dictionaries („*BYOD*“).
(Commercial products will include dictionaries – but for a lot of money.)
- Not perfect but no big issues, only some minor problems (errors) in *KNIME Labs* nodes that were fixed by the developers on a new KNIME version.

Computational Chemistry:

Nils Weskamp

IS Department:

Johannes Koppe, Andreas Eberlein

Computational Biology:

Stephen Gelling, Peter Leonhardt

KNIME:

Tobias Kötter (developer of NetworkMining nodes)

Thank you!

Questions?