A non-technical introduction to *Text Mining*

Tom De Schryver
Information specialist for BMS/MB
t.deschryver@utwente.nl

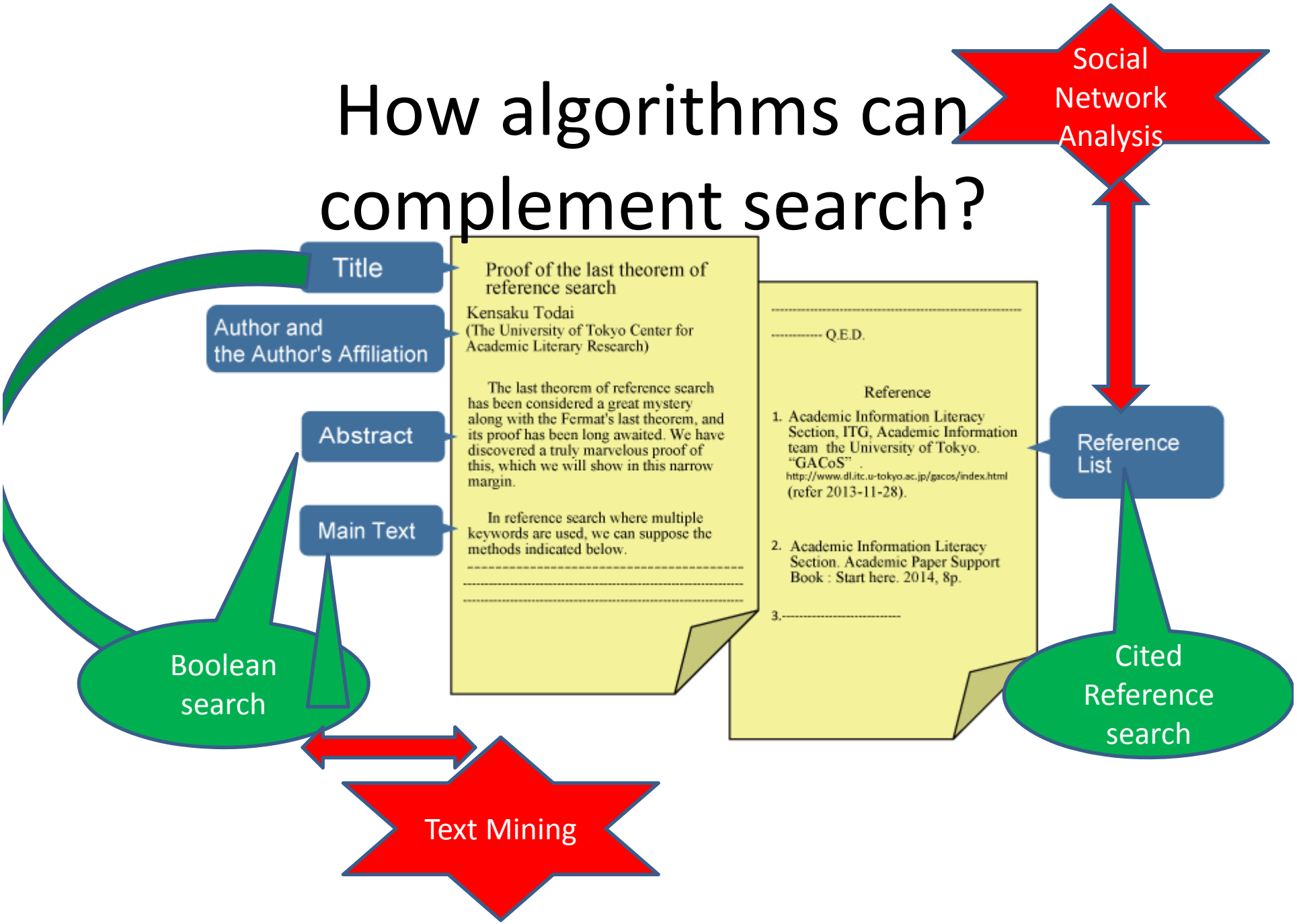
Embedded Information Services
Library & Archive - University of Twente

Propositions

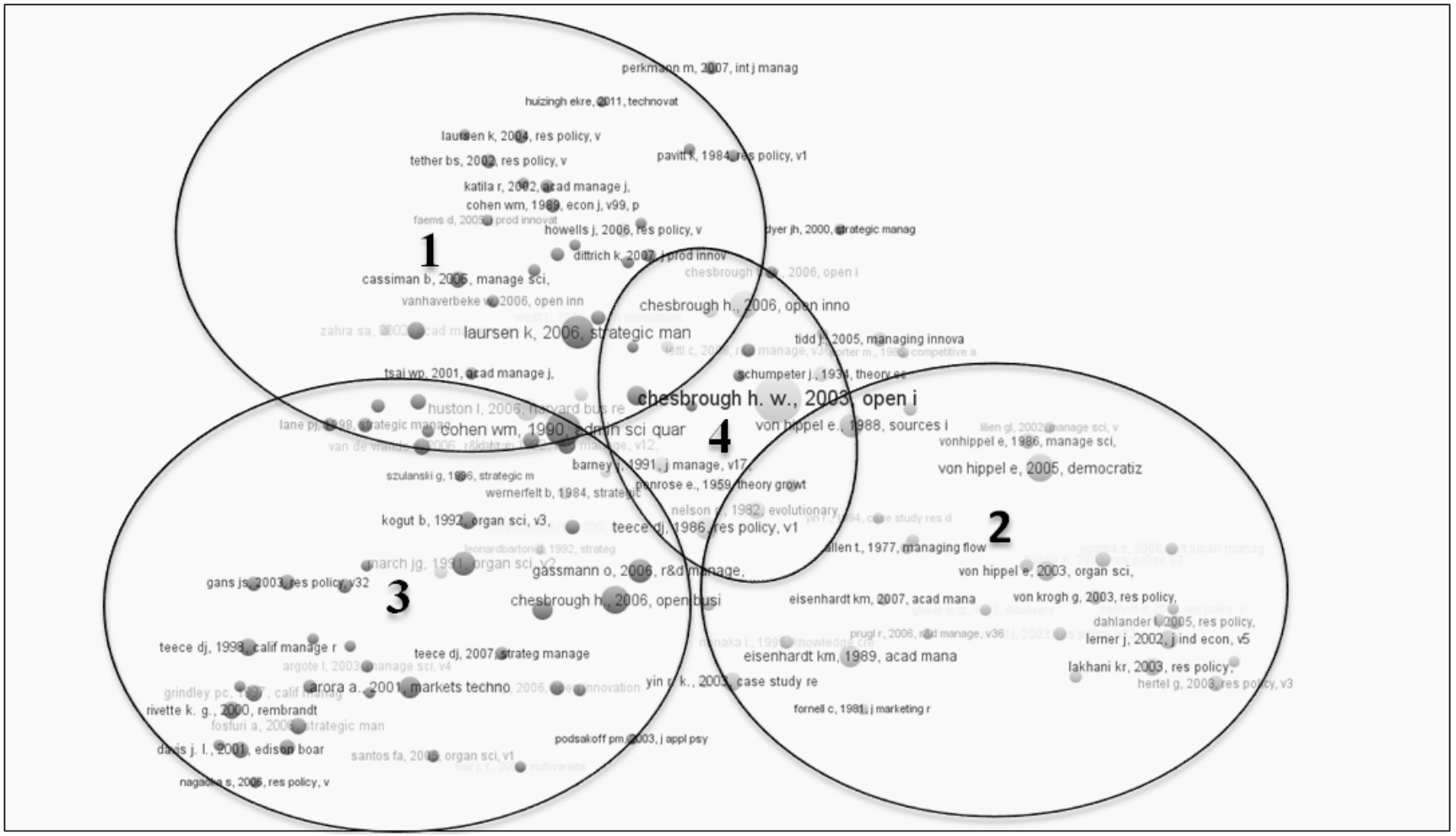
Text mining

- is a **threat** for the information specialist.
- is a **new tool** for the information specialist.
- requires **new skills** from the information specialist.
- can be a **great opportunity** to collaborate with researchers/ clients.

How algorithms can complement search?

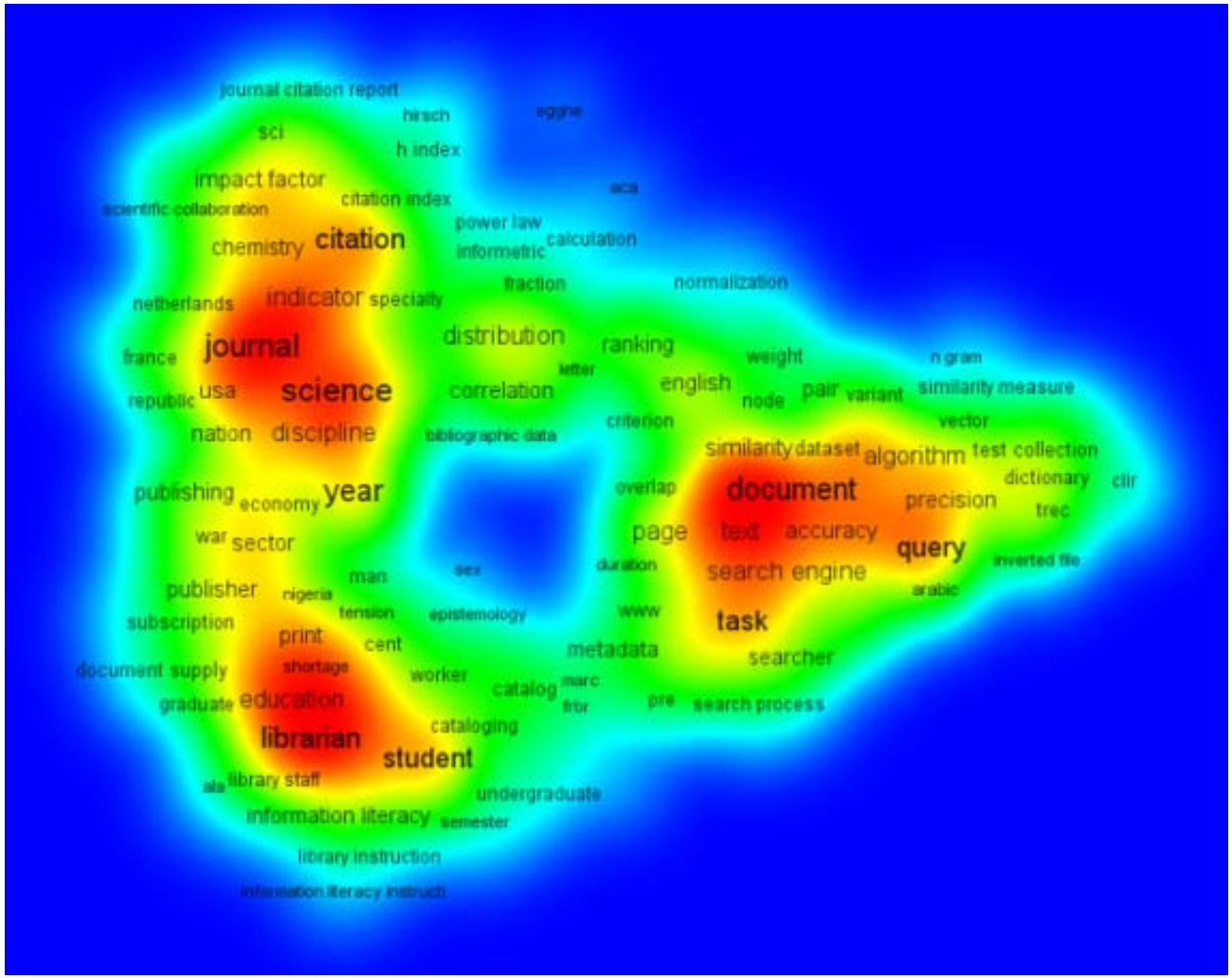


“Social” network analysis



Source: Kovács, A., Van Looy, B., & Cassiman, B. (2014). Exploring the scope of open innovation: A bibliometric review of a decade of research. *FEB Research Report MSI_1401*.https://lirias.kuleuven.be/bitstream/123456789/439281/1/MSI_1401.pdf

Text mining example



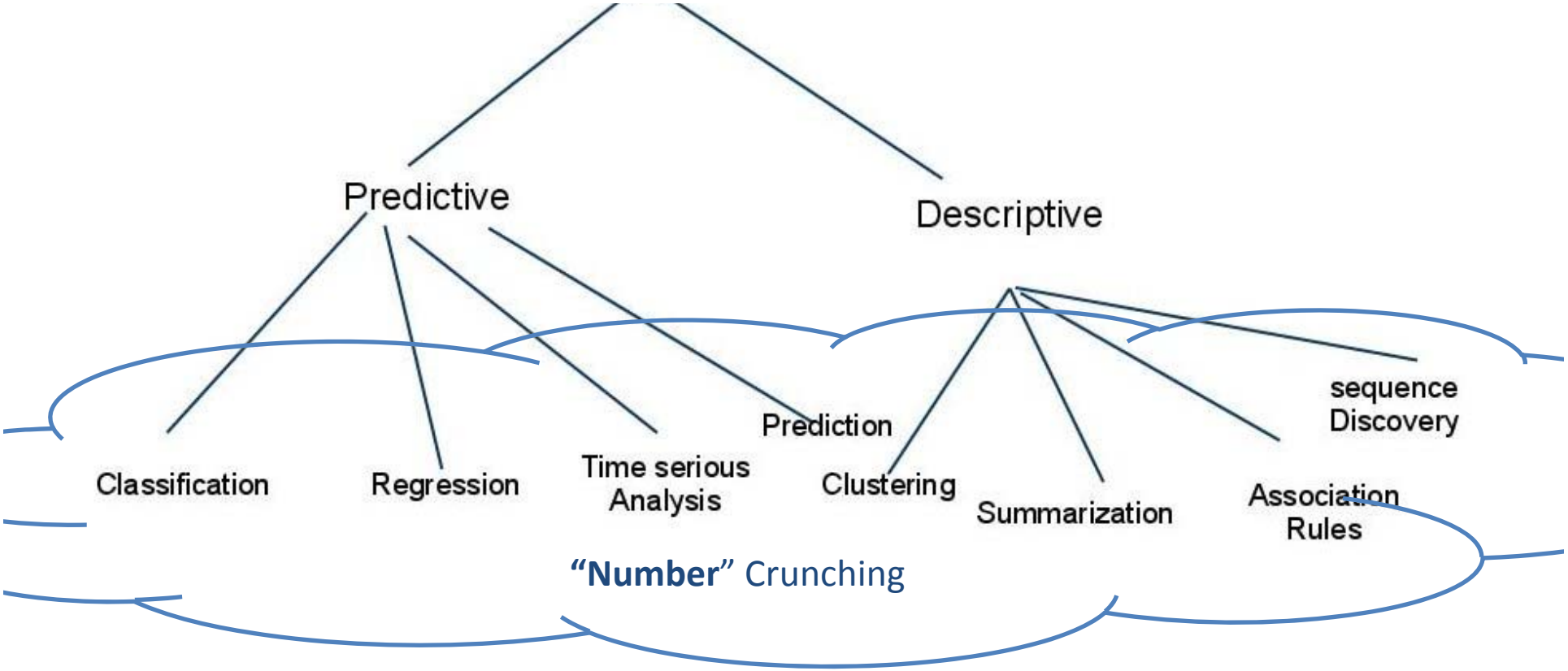
Source: Van Eck, N.J., & Waltman, L. (2011). Text mining and visualization using VOSviewer. *ISSI Newsletter*, *7*(3), 50-54

Based on the two graphs

- What are the **differences** between bibliometric network analysis and text mining?
- What are the **similarities** between bibliometric network analysis and text mining?

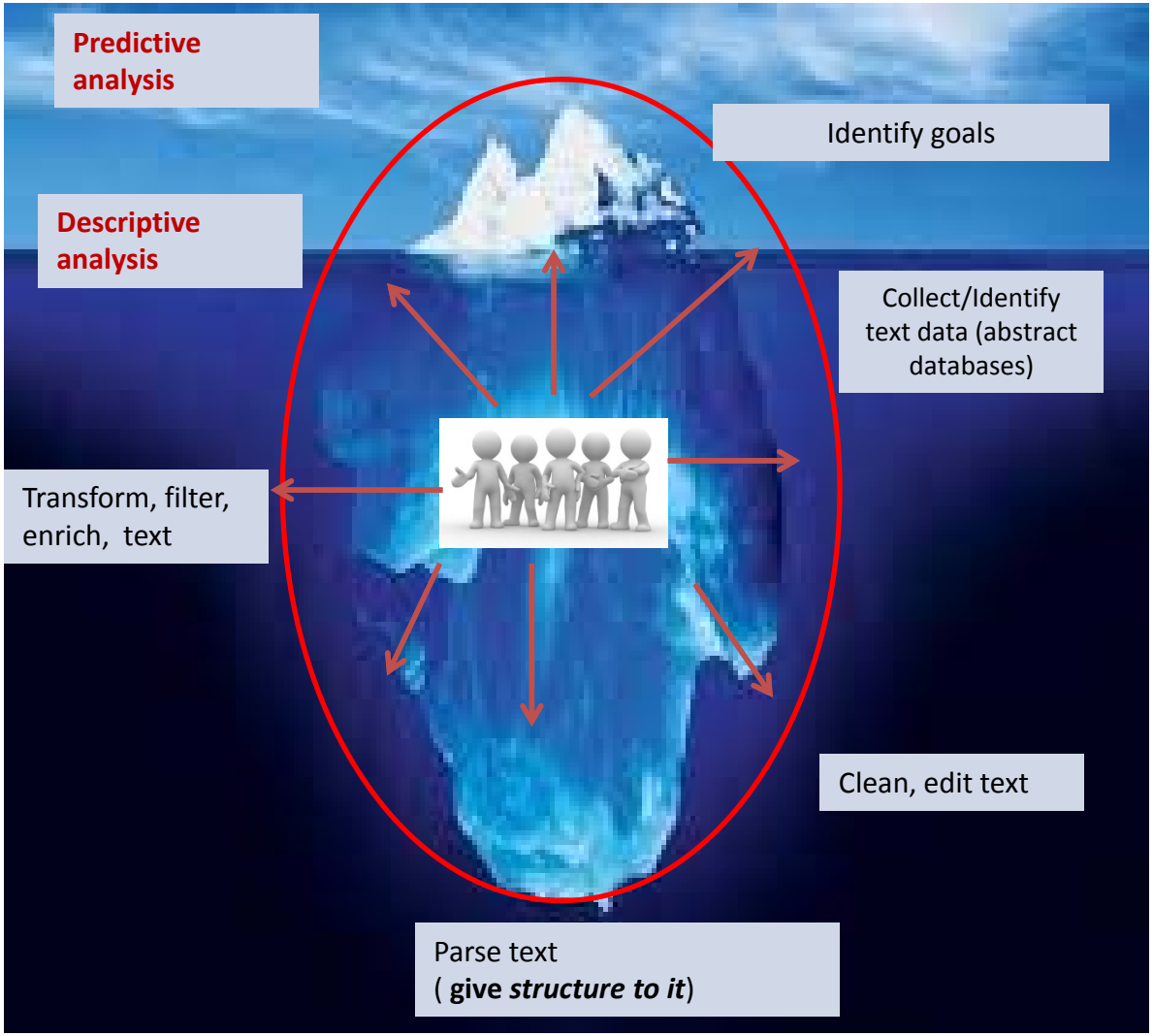
Text mining/ analysis?

= Descriptive or predictive analysis of **text**



Text? = unstructured data !

TexT analytics cycle



Which data for text mining?

- for text mining you need a lot of data (separate documents/ paragraphs)
  - –the longer the documents, the more documents you need without them introducing new words
  - –(the curse of dimensionality problem: see appendix)

Example corpus

document	Text
D1	I love IPAD.
D2	IPAD is great for kids.
D3	Kids love to play soccer.
D4	I play soccer at UT.

Fits with literature review searches

- Scientific articles: Boolean search often on
  - Titles
  - abstracts
  - Keywords

- Data can easily be taken from abstract databases and exported for analysis.

Parsing

Raw data

document	Text
D1	I love IPAD.
D2	IPAD is great for kids.
D3	Kids love to play soccer.
D4	I play soccer at UT.

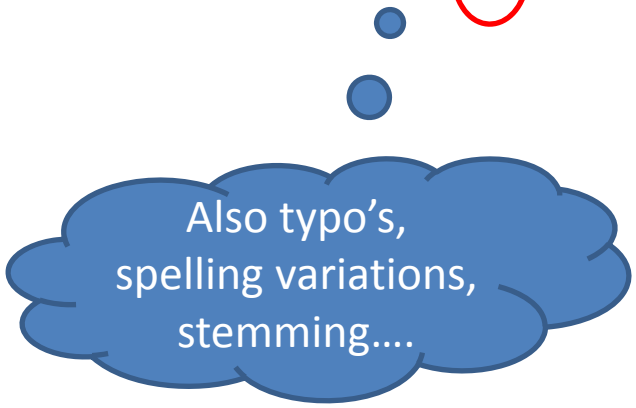
Term by document matrix

term	d1	d2	d3	d4
I	1	0	0	1
love	1	0	1	0
Ipad	1	1	0	0
Is	0	1	0	1
great	0	1	0	0
kids	0	1	1	0
play	0	0	1	1
soccer	0	0	1	1
ut	0	0	0	1

Filtering

Raw data

document	Text
D1	I love IPAD.
D2	IPAD is great for kids.
D3	Kids love to play soccer.
D4	I play soccer at UT.



Term by document matrix

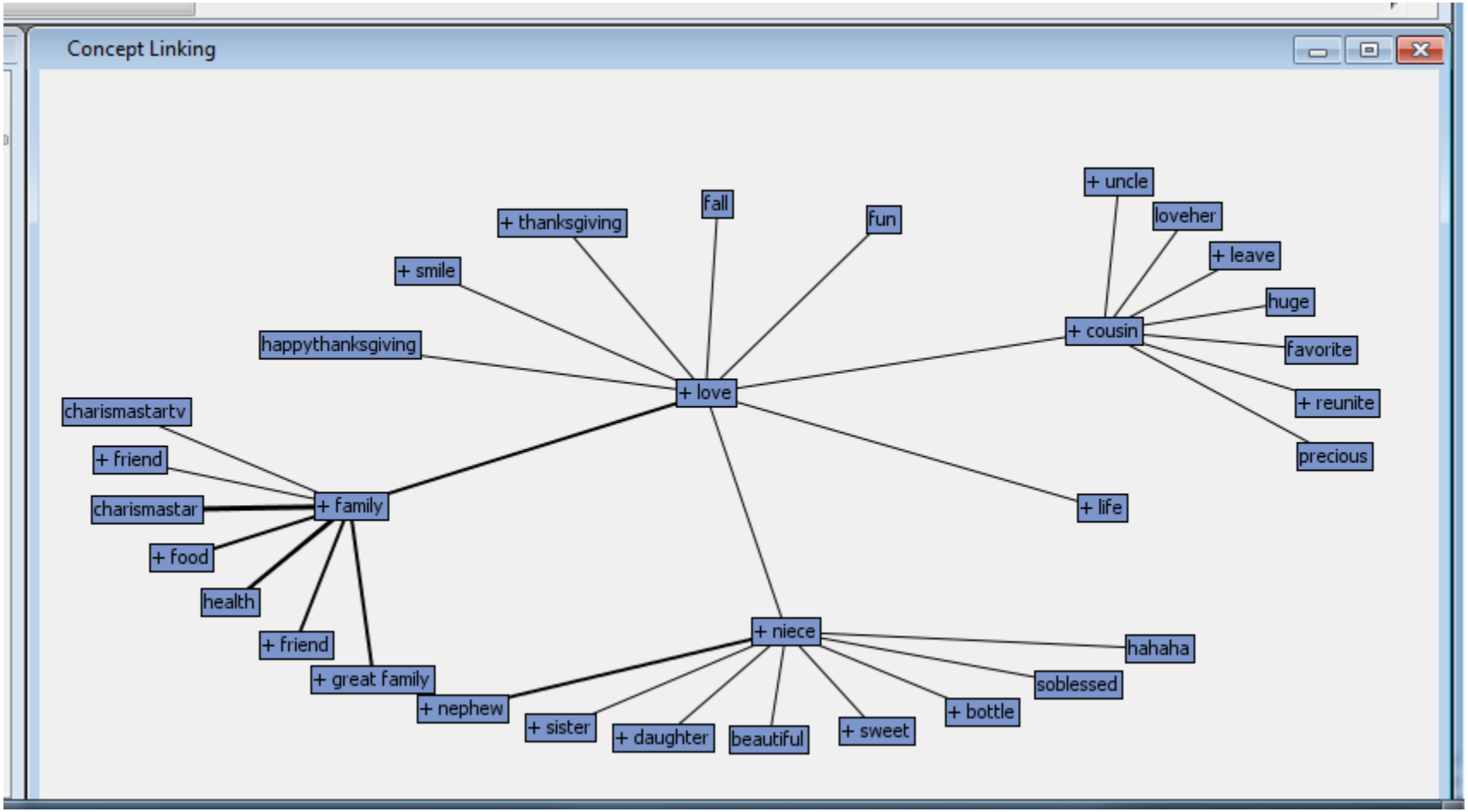
term	d1	d2	d3	d4
I	1	0	0	1
love	1	0	1	0
Ipad	1	1	0	0
Is	0	1	0	1
great	0	1	0	0
kids	0	1	1	0
play	0	0	1	1
soccer	0	0	1	1
ut	0	0	0	1

An enriched term by doc matrix

term	d1	d2	d3	d4	POS tag	dft	sentiment
I	1	0	0	1	Noun	2	☺
love	1	0	1	0	Verb	2	☺
Ipad	1	1	0	0	Noun	2	😐
Is	0	1	0	1	Verb	2	😐
great	0	1	0	0	Adjective	1	☺
kids	0	1	1	0	Noun	2	☺
play	0	0	1	1	Verb	2	☺
soccer	0	0	1	1	Noun	2	☹
ut	0	0	0	1	Noun	1	😐
Yr (20..)	15	14	09	13			

Much more attributes of terms and doc's can/should be added

First example: Descriptive analysis of **terms**

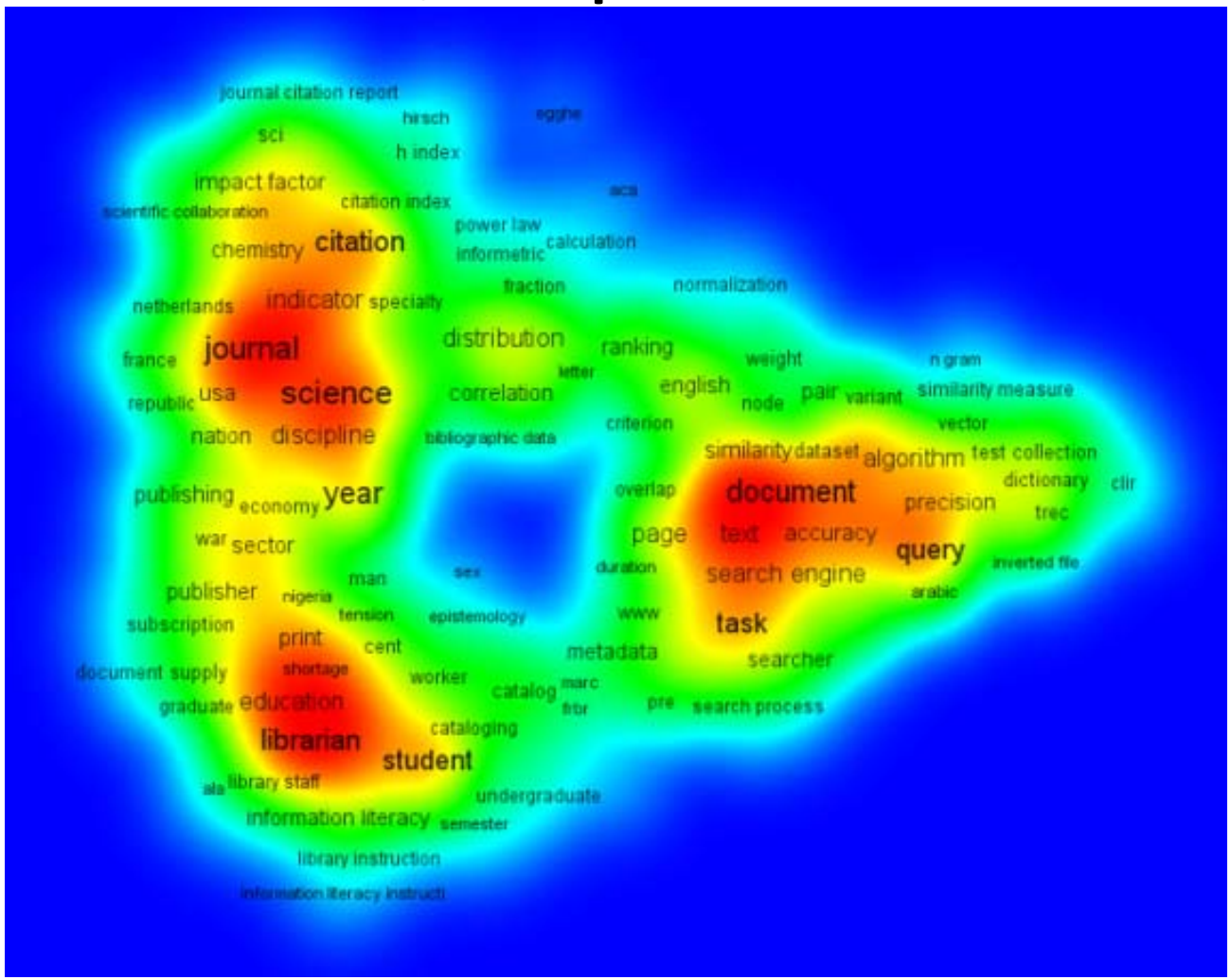


Source: SAS text mining software / Chakraborty et al. (2013) see also https://support.sas.com/resources/papers/proceedings14/1288-2014.pdf

descriptive analysis of **terms**

- **related**- broader – narrower terms?
  - conventional tool: thesaurus search
  - textmining tool: which terms co-occur in corpus of text?
  - See also:
  - http://pushaqa.blogspot.nl/2014/12/presenting-keywords-eh-which-keywords.html
-

Second example : Descriptive analysis of "document/corpus"-structure

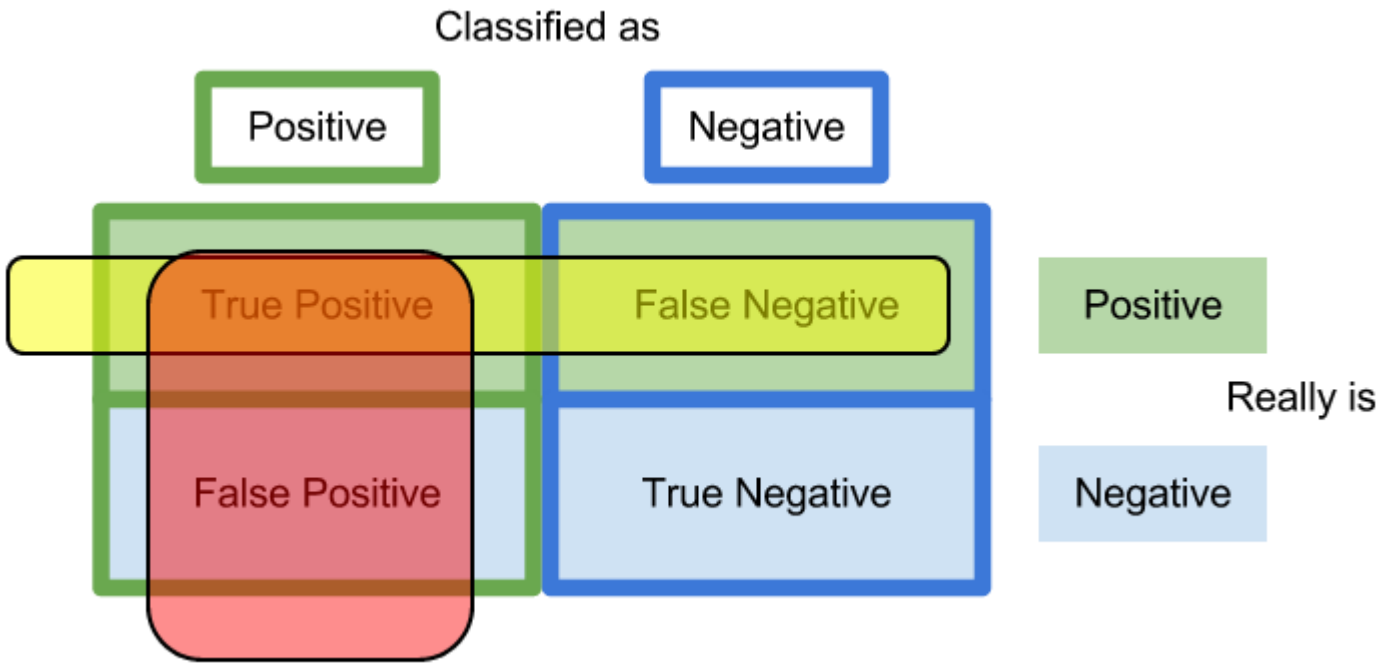


Source: Van Eck, N.J., & Waltman, L. (2011). Text mining and visualization using VOSviewer. *ISSI Newsletter*, *7*(3), 50-54

Example of a Predictive analysis: literature review

Example of Predictive analysis:

- The information specialist / text mining suggests a shortlist of relevant literature



First results?: with Text mining faster and better

	Student		Included / excluded Correctly	Included / excluded incorrectly
Manual	1	85 min	25	12
Manual	2	54 min	22	15
Text mining	3	30 min	27	10
Text mining	4	58 min	28	9

Source: (Felizardo et al. 2011) See also Hausner et al. (2015)

Already: your today's reality!

- Mailbox: spam filtering
- Literature research: ham filtering
  - See also http://www3.nd.edu/~steve/computing_with_data/20_text_mining/text_mining_example.html#/

Different free software programs available

- bibliometric networks
  - Gephi/Pajek
  - Vosviewer/Citespace

Source : http://www.vosviewer.com/download/f-x2.pdf

- Text mining (free software)
  - Python textminer
  - R- tm

  Source: https://en.wikipedia.org/wiki/List_of_text_mining_software

References

- Ananiadou, S., Rea, B., Okazaki, N., Procter, R., & Thomas, J. (2009). Supporting systematic reviews using text mining. *Social Science Computer Review*.
- Felizardo, Katia R., et al. "Using visual text mining to support the study selection activity in systematic literature reviews." *Empirical Software Engineering and Measurement (ESEM), 2011 International Symposium on*. IEEE, 2011.
- Hausner, E., Guddat, C., Hermanns, T., Lampert, U., & Waffenschmidt, S. (2015). Development of search strategies for systematic reviews: validation showed the noninferiority of the objective approach. *Journal of clinical epidemiology*, *68*(2), 191-199.
- Van Eck, N.J., & Waltman, L. (2014). Visualizing bibliometric networks. In Y. Ding, R. Rousseau, & D. Wolfram (Eds.), *Measuring scholarly impact: Methods and practice* (pp. 285–320). Springer

APPENDIX

The curse of dimensionality problem (step 1)

Raw data

document	Tekst	rowi nexcel
D1	I love IPAD.	1
~~D2~~	~~IPAD is great for kids.~~	~~2~~
~~D3~~	~~Kids love to play soccer.~~	~~3~~
~~D4~~	~~I play soccer at UT.~~	~~4~~

Term by document matrix

Obs	term	d1
1	I	1
2	love	1
3	Ipad	1

The curse of dimensionality problem (step 2)

Raw data

document	Tekst	rowinexcel
D1	I love IPAD.	1
D2	IPAD is great for kids.	2
~~D3~~	~~Kids love to play soccer.~~	~~3~~
~~D4~~	~~I play soccer at UT.~~	~~4~~

Term by document matrix

Obs	term	d1	d2
1	I	1	0
2	love	1	0
3	Ipad	1	1
4	Is	0	1
5	great	0	1
6	kids	0	1

The curse of dimensionality problem (step3)

Raw data

document	Tekst	rowinexcel
D1	I love IPAD.	1
D2	IPAD is great for kids.	2
D3	Kids love to play soccer.	3
~~D4~~	~~I play soccer at UT.~~	~~4~~

Term by document matrix

Obs	term	d1	d2	d3
1	I	1	0	0
2	love	1	0	1
3	Ipad	1	1	0
4	Is	0	1	0
5	great	0	1	0
6	kids	0	1	1
7	play	0	0	1
8	soccer	0	0	1

The curse of dimensionality problem (step4)

Raw data

document	Tekst	rowinexcel
D1	I love IPAD.	1
D2	IPAD is great for kids.	2
D3	Kids love to play soccer.	3
D4	I play soccer at UT.	4

Term by document matrix

Obs	term	d1	d2	d3	d4
1	I	1	0	0	1
2	love	1	0	1	0
3	Ipad	1	1	0	0
4	Is	0	1	0	1
5	great	0	1	0	0
6	kids	0	1	1	0
7	play	0	0	1	1
8	soccer	0	0	1	1
9	ut	0	0	0	1

A word versus a term

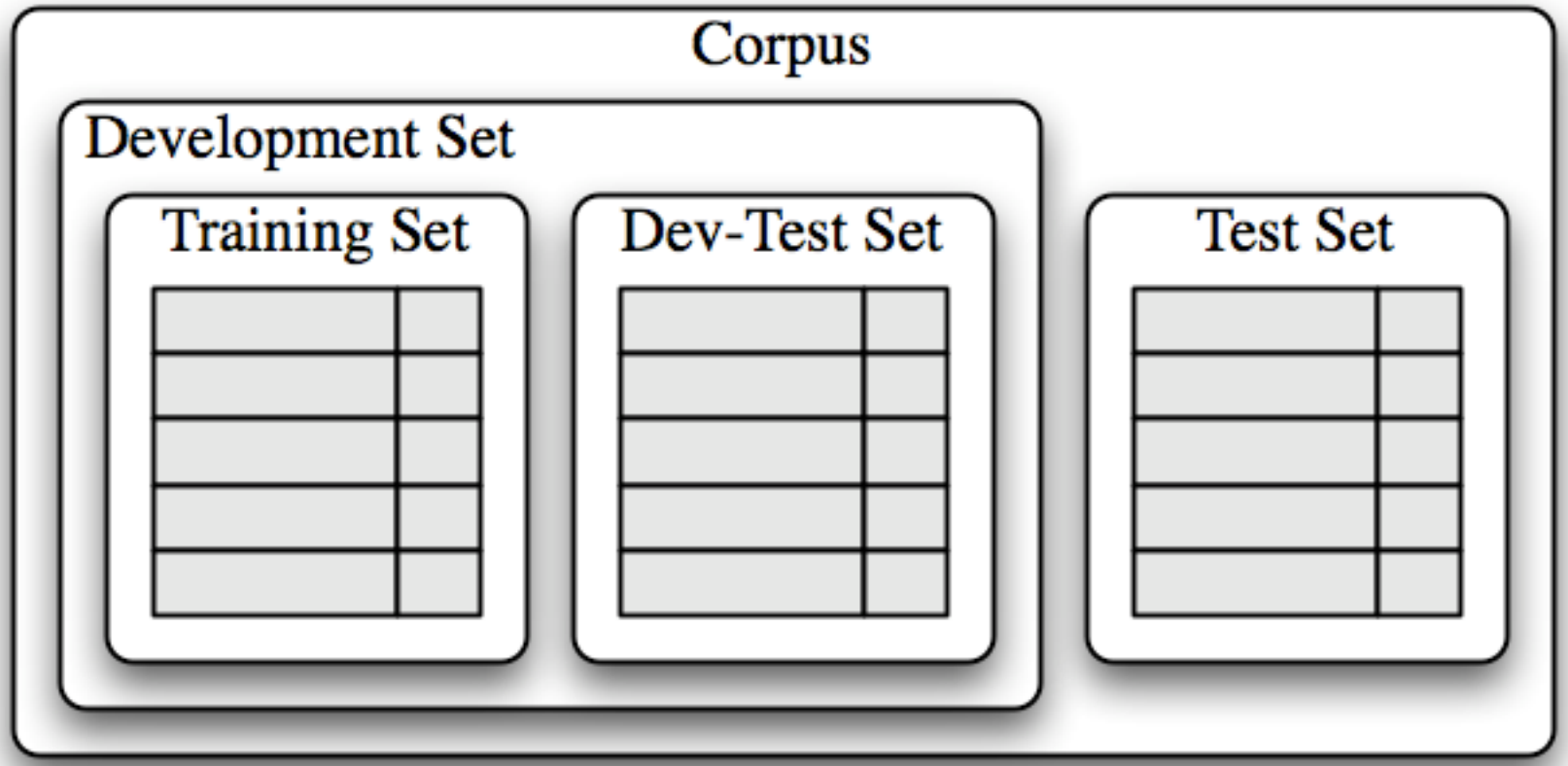
- Example: white is one word but stands for 3 terms!
  - The car is white.
  - Mr. White is coming to town.
  - Where is The White House?

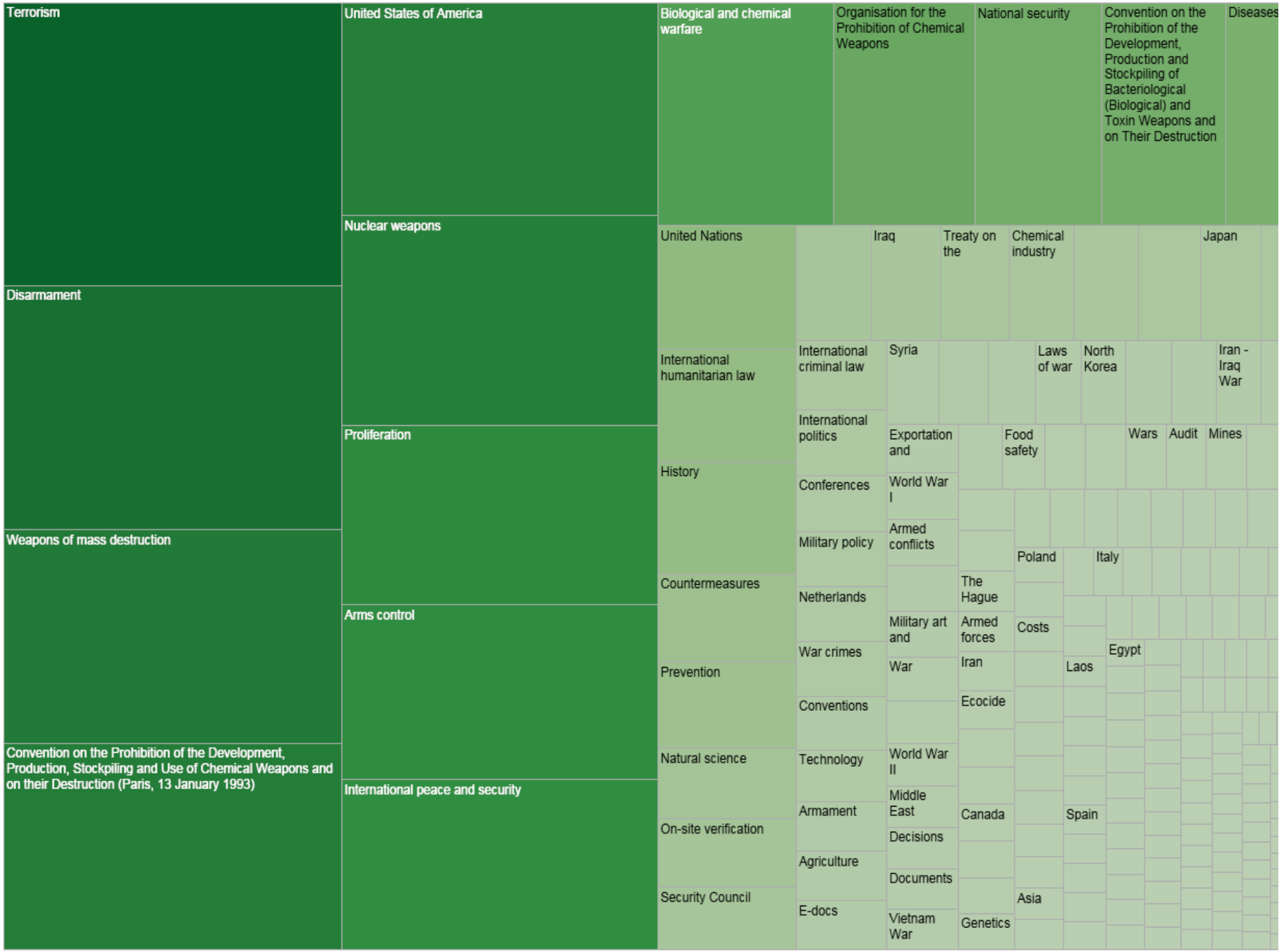
- Similar words can be reduced to one term
  - Stemming: grammatical variations (e.g. single/ plural)
  - spelling mistakes or variations (UK/ US English)

Term reduction

- Curse of dimensionality
  - By experts
  - Frequency based (e.g. UT)
  - language based (e.g stop words)
  - According to the contribution to the variance (SVD)

Partition the data?




Terrorism
Disarmament
Weapons of mass destruction
Convention on the Prohibition of the Development, Production, Stockpiling and Use of Chemical Weapons and on their Destruction (Paris, 13 January 1993)
United States of America
Nuclear weapons
Proliferation
Arms control
International peace and security
Biological and chemical warfare
Organisation for the Prohibition of Chemical Weapons
National security
Convention on the Prohibition of the Development, Production and Stockpiling of Bacteriological (Biological) and Toxin Weapons and on Their Destruction
Diseases
United Nations
Iraq
Treaty on the
Chemical industry
Japan
International humanitarian law
International criminal law
Syria
Laws of war
North Korea
Iran - Iraq War
International politics
Exportation and
Food safety
Wars
Audit
Mines
History
Conferences
World War I
Military policy
Armed conflicts
Poland
Italy
Countermeasures
Netherlands
The Hague
Military art and
Armed forces
Costs
War crimes
War
Iran
Egypt
Prevention
Laos
Conventions
Ecocide
Natural science
Technology
World War II
Middle East
Armament
Canada
Spain
On-site verification
Decisions
Agriculture
Documents
Security Council
E-docs
Asia
Vietnam War
Genetics