Mapping Variation in the Phraseology of Student Writing

Phil Durrant

p.l.durrant@exeter.ac.uk

Motivation

Variation as a key ‘problem’ for EAP

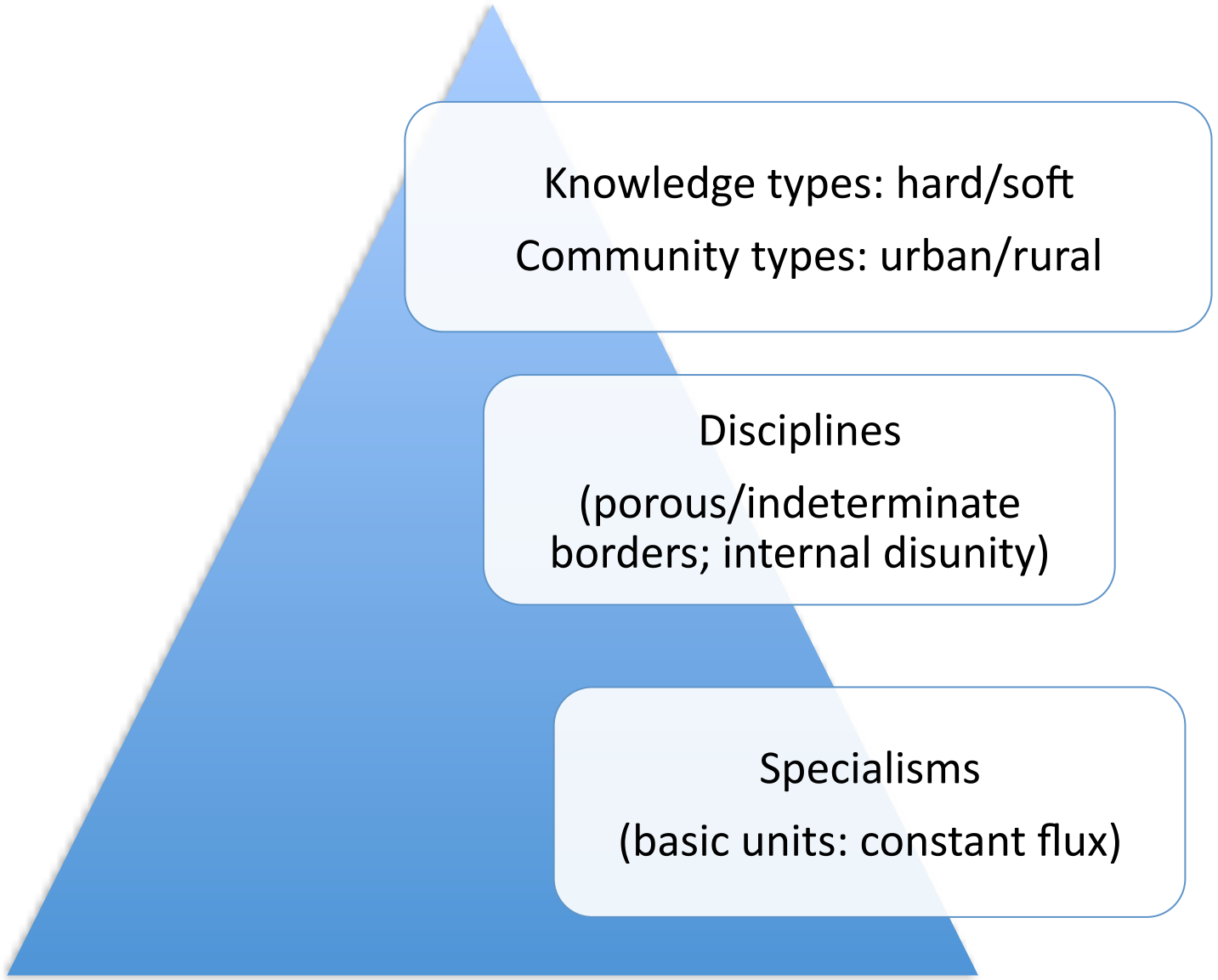
Motivation

How can corpora help our understanding?

More specifically:

What are ‘disciplines’?

How do they relate to each other?



Becher & Trowler 2001

Defining Discipline

- Disciplinary identities are negotiable, permeable and transient (Brew, 2008)
- Different definitions for different purposes:
  - Research vs. teaching (Manathunga & Brew, 2012)
  - Administrative vs. linguistic (Durrant, 2009)

Disciplines in corpus linguistics

Coxhead, 2000

Arts	Commerce	Law	Science
Education	Accounting	Constitutional	Biology
History	Economics	Criminal	Chemistry
Linguistics	Finance	Family and medicolegal	Computer science
Philosophy	Industrial Relations	International	Geography
Psychology	Management	Pure commnercial	Geology
Sociology	Marketing	Quasi-commercial	Mathematics
	Public Policy	Rights and remedies	Physics

Coxhead, 2000

Arts	Commerce	Law	Science
Education History Linguistics Philosophy Psychology Sociology	Accounting Economics Finance Industrial Relations Management Marketing Public Policy	Constitutional Criminal Family and medicolegal International Pure commnercial Quasi-commercial Rights and remedies	Biology Chemistry Computer science Geography Geology Mathematics Physics

Biber et al, 2006

Business	Education	Engineering	Humanities	Natural Science	Social Science

Coxhead, 2000

Arts	Commerce	Law	Science
Education History Linguistics Philosophy Psychology Sociology	Accounting Economics Finance Industrial Relations Management Marketing Public Policy	Constitutional Criminal Family and medicolegal International Pure commnercial Quasi-commercial Rights and remedies	Biology Chemistry Computer science Geography Geology Mathematics Physics

Biber et al, 2006

Business	Education	Engineering	Humanities	Natural Science	Social Science

Simpson-Vlach & Ellis, 2010

Humanities & Arts	Social Sciences	Natural Sciences/ Medicine	Technology and Engineering

Coxhead, 2000

Arts	Commerce	Law	Science
Education History Linguistics Philosophy Psychology Sociology	Accounting Economics Finance Industrial Relations Management Marketing Public Policy	Constitutional Criminal Family and medicolegal International Pure commnercial Quasi-commercial Rights and remedies	Biology Chemistry Computer science Geography Geology Mathematics Physics

Biber et al, 2006

Business	Education	Engineering	Humanities	Natural Science	Social Science

Simpson-Vlach & Ellis, 2010

Humanities & Arts	Social Sciences	Natural Sciences/ Medicine	Technology & Engineering

Gardner & Davies, 2014

Education	Humanities	History	Social Science	Philosophy, religion, psychology
Law & Political Science	Science & technology	Medicine & health	Business & finance	

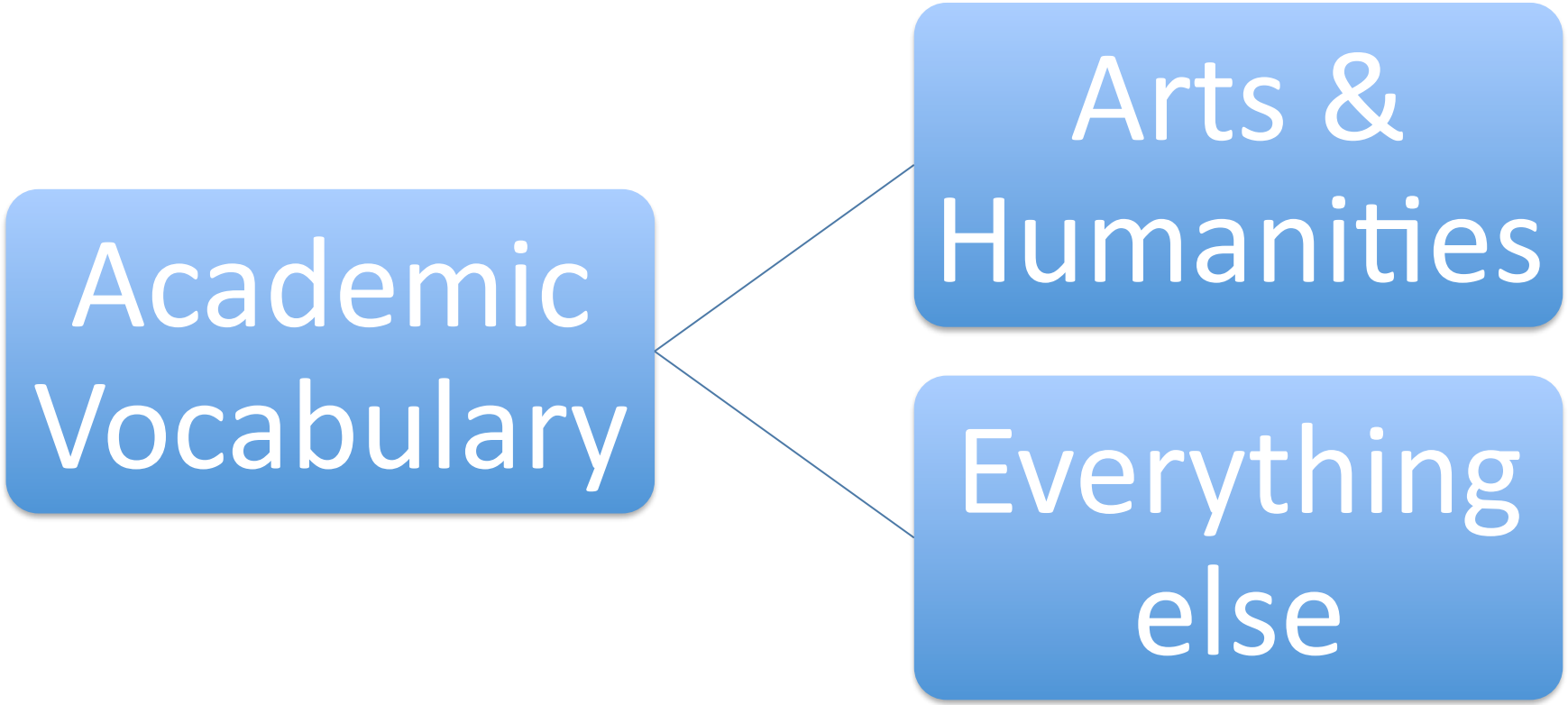
Defining disciplinary groups through clusters (1)

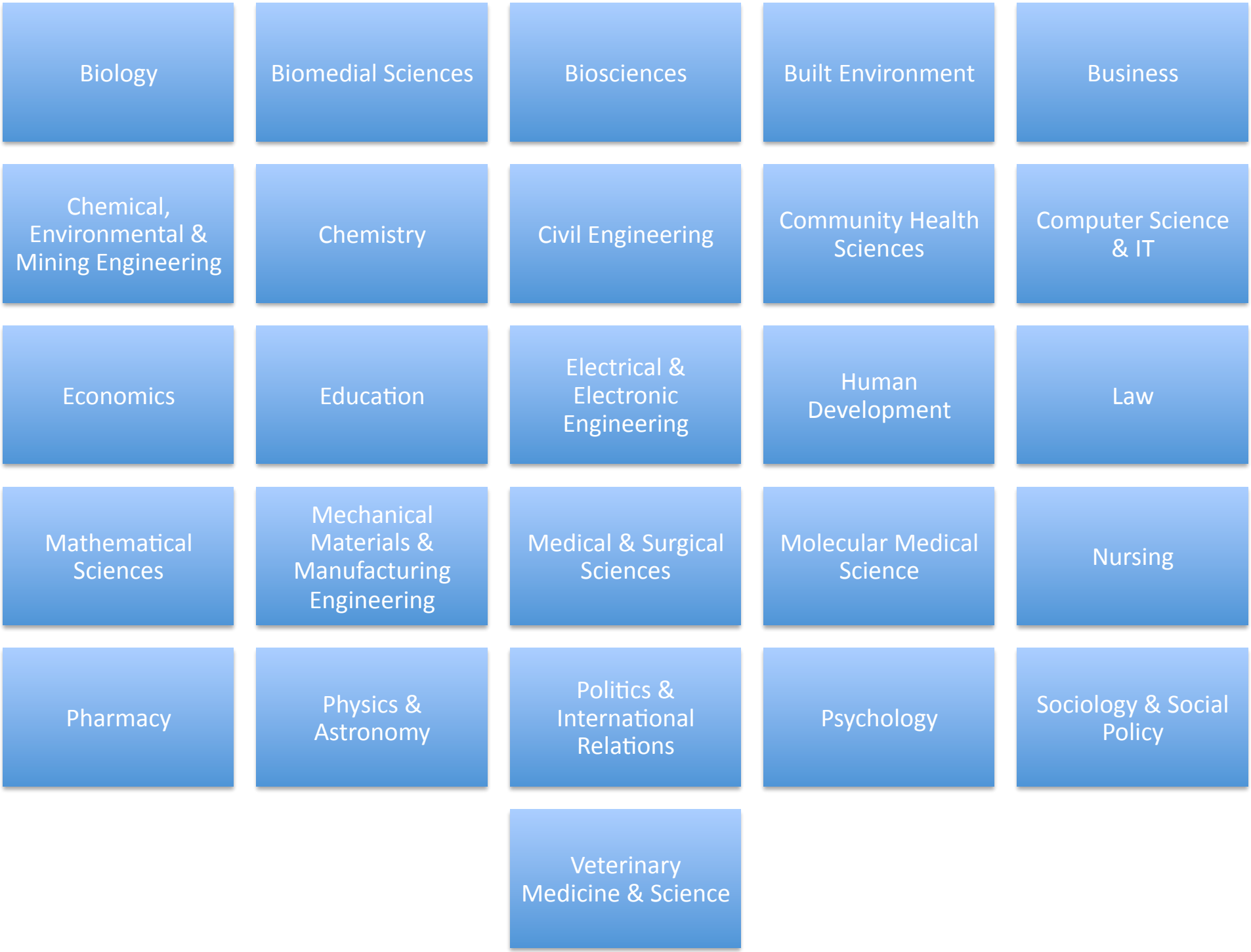
- Find ‘keywords’ in 31 disciplinary units
- Create ‘overlap matrix’ for all units
  - *Overlap = total shared keywords/total keywords*
- Use hierarchical cluster analysis to identify groups

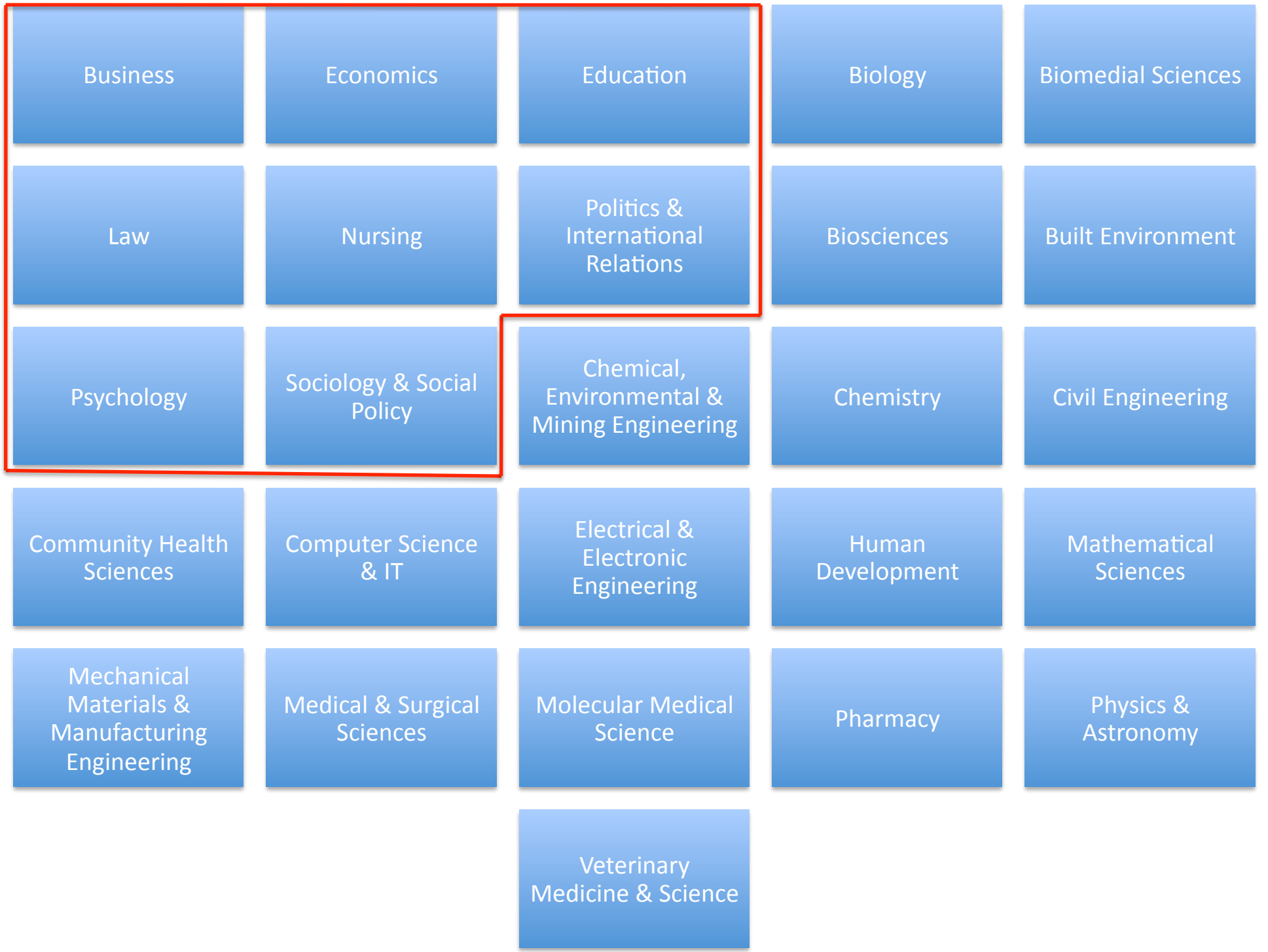
Durrant, 2009

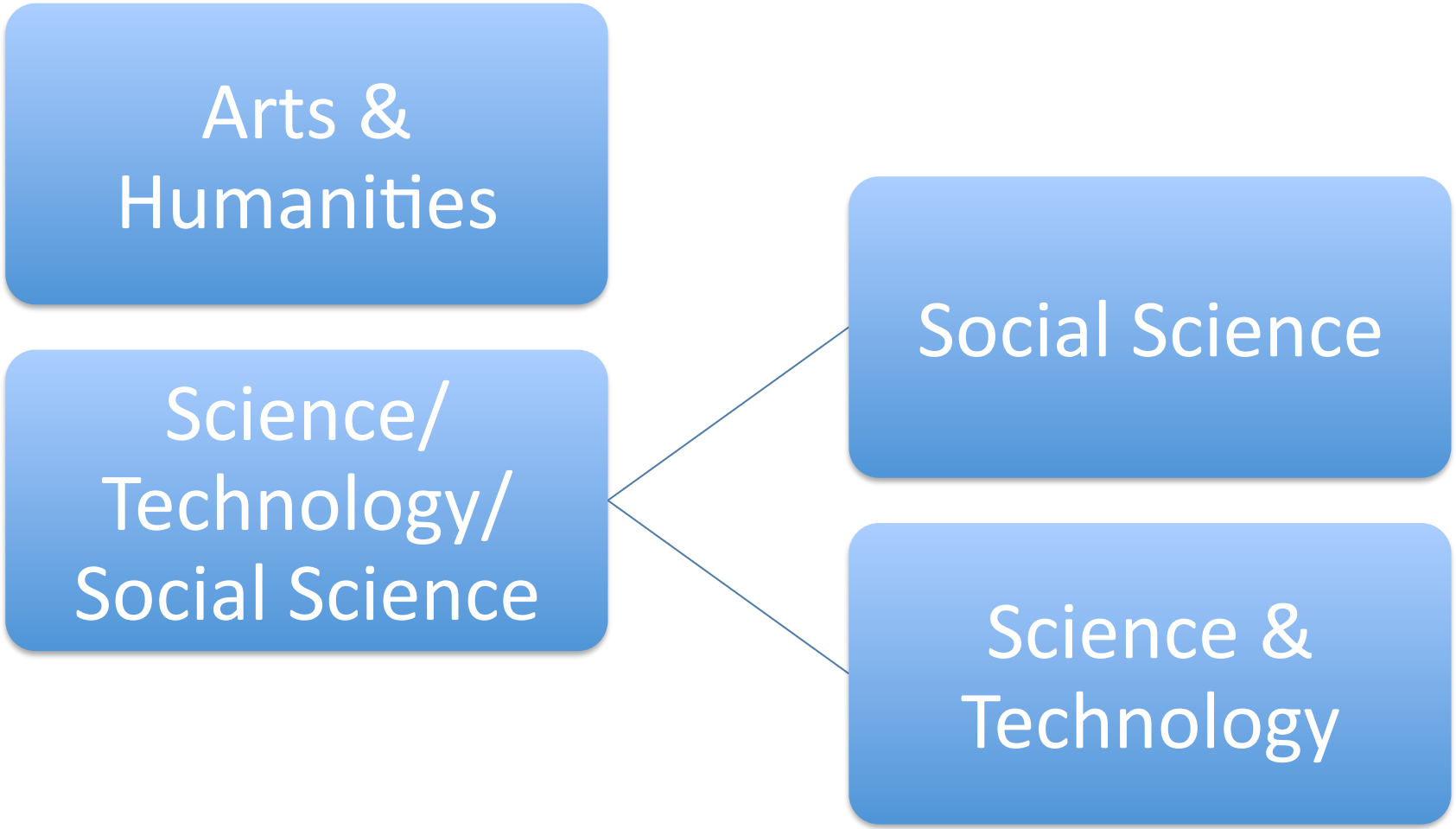

American & Canadian Studies
Biology
Biomedial Sciences
Biosciences
Built Environment
Business
Chemical, Environmental & Mining Engineering
Chemistry
Civil Engineering
Community Health Sciences
Computer Science & IT
Economics
Education
Electrical & Electronic Engineering
English Studies
History
Human Development
Law
Mathematical Sciences
Mechanical Materials & Manufacturing Engineering
Medical & Surgical Sciences
Modern Language & Cultures
Molecular Medical Science
Nursing
Pharmacy
Physics & Astronomy
Politics & International Relations
Psychology
Sociology & Social Policy
Veterinary Medicine & Science

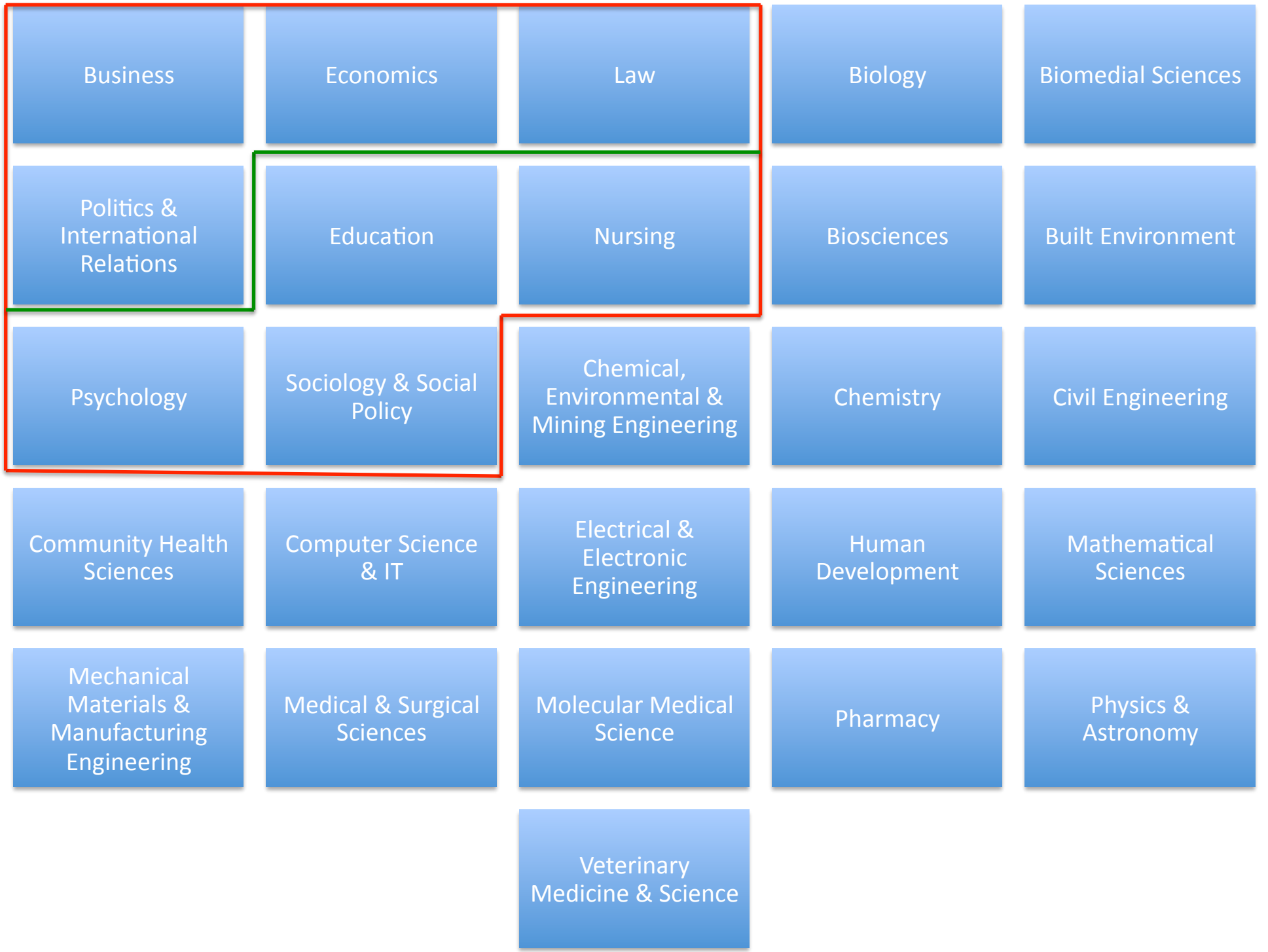

American & Canadian Studies
English Studies
Biology
Biomedial Sciences
Biosciences
History
Modern Language & Cultures
Built Environment
Business
Chemical, Environmental & Mining Engineering
Chemistry
Civil Engineering
Community Health Sciences
Computer Science & IT
Economics
Education
Electrical & Electronic Engineering
Human Development
Law
Mathematical Sciences
Mechanical Materials & Manufacturing Engineering
Medical & Surgical Sciences
Molecular Medical Science
Nursing
Pharmacy
Physics & Astronomy
Politics & International Relations
Psychology
Sociology & Social Policy
Veterinary Medicine & Science

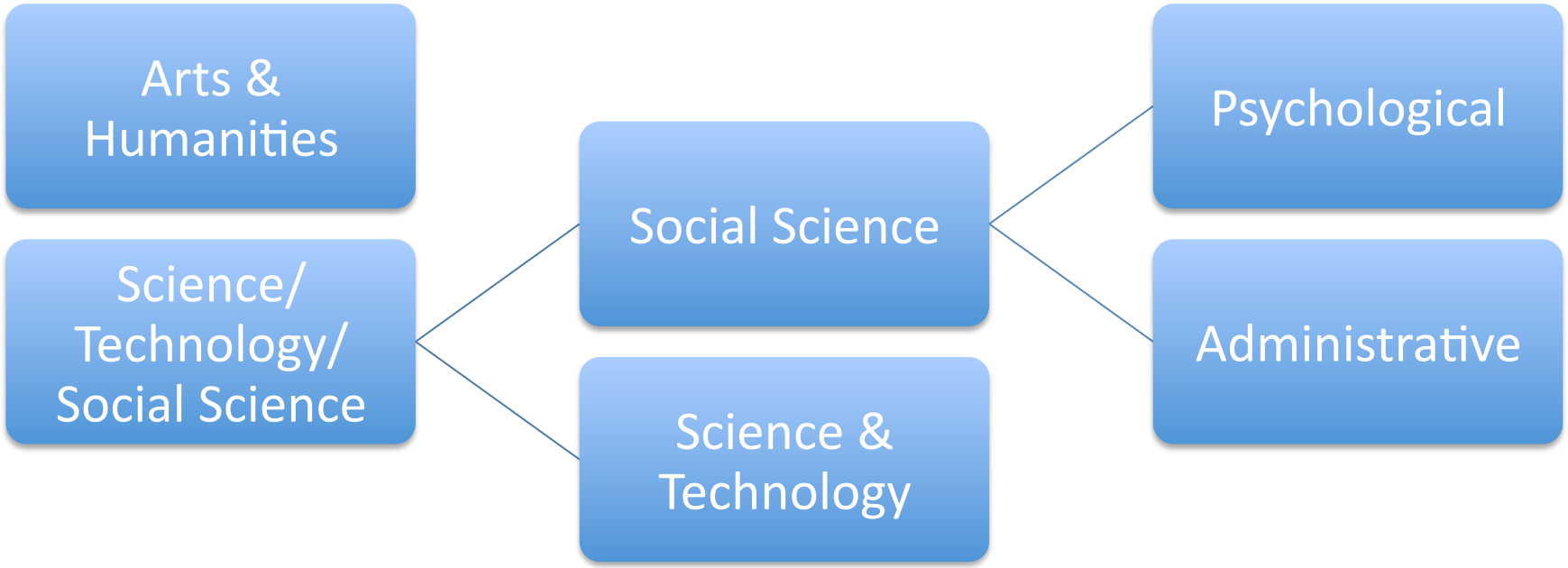

Academic Vocabulary
Arts & Humanities
Everything else

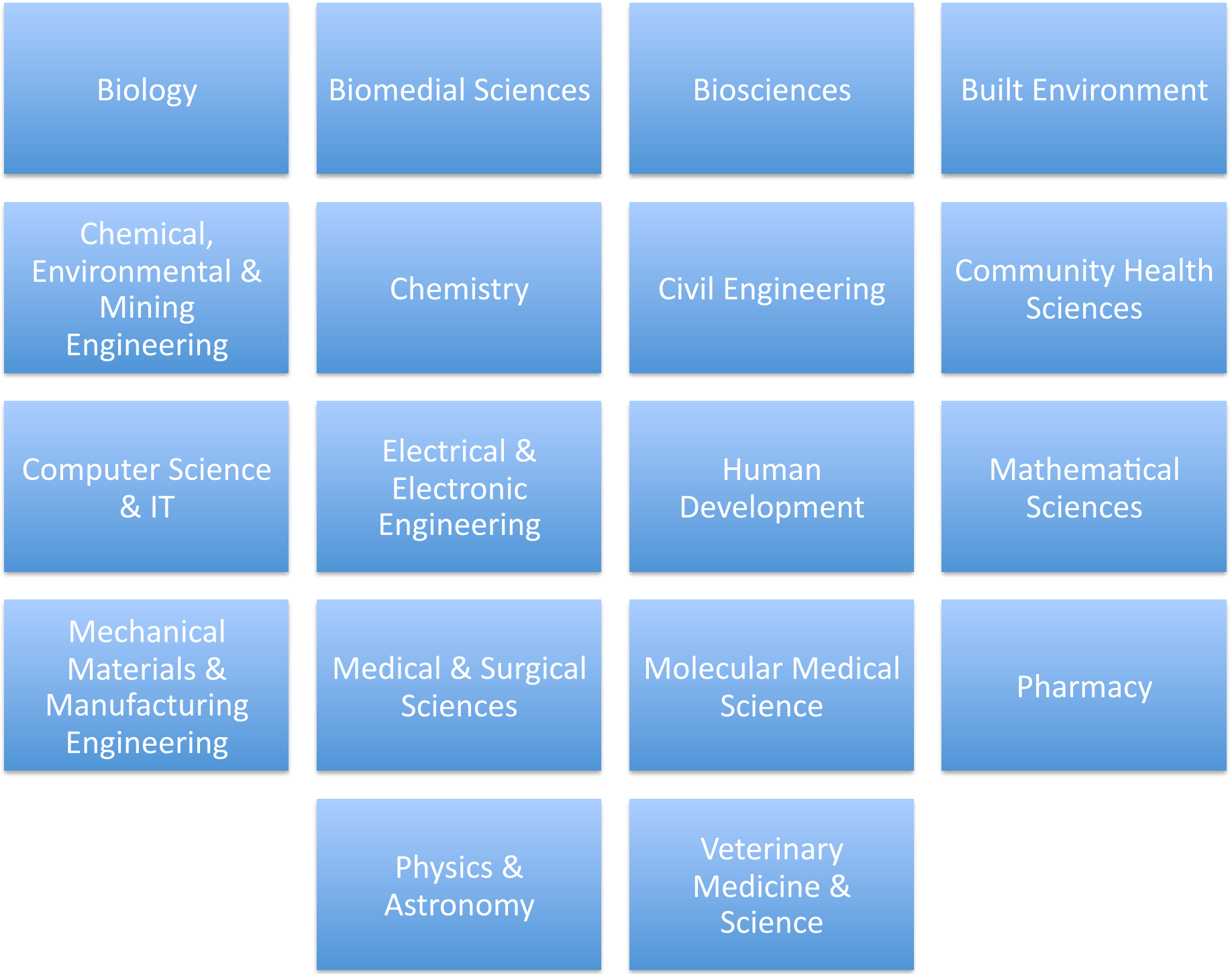

Biology
Biomedial Sciences
Biosciences
Built Environment
Business
Chemical, Environmental & Mining Engineering
Chemistry
Civil Engineering
Community Health Sciences
Computer Science & IT
Economics
Education
Electrical & Electronic Engineering
Human Development
Law
Mathematical Sciences
Mechanical Materials & Manufacturing Engineering
Medical & Surgical Sciences
Molecular Medical Science
Nursing
Pharmacy
Physics & Astronomy
Politics & International Relations
Psychology
Sociology & Social Policy
Veterinary Medicine & Science

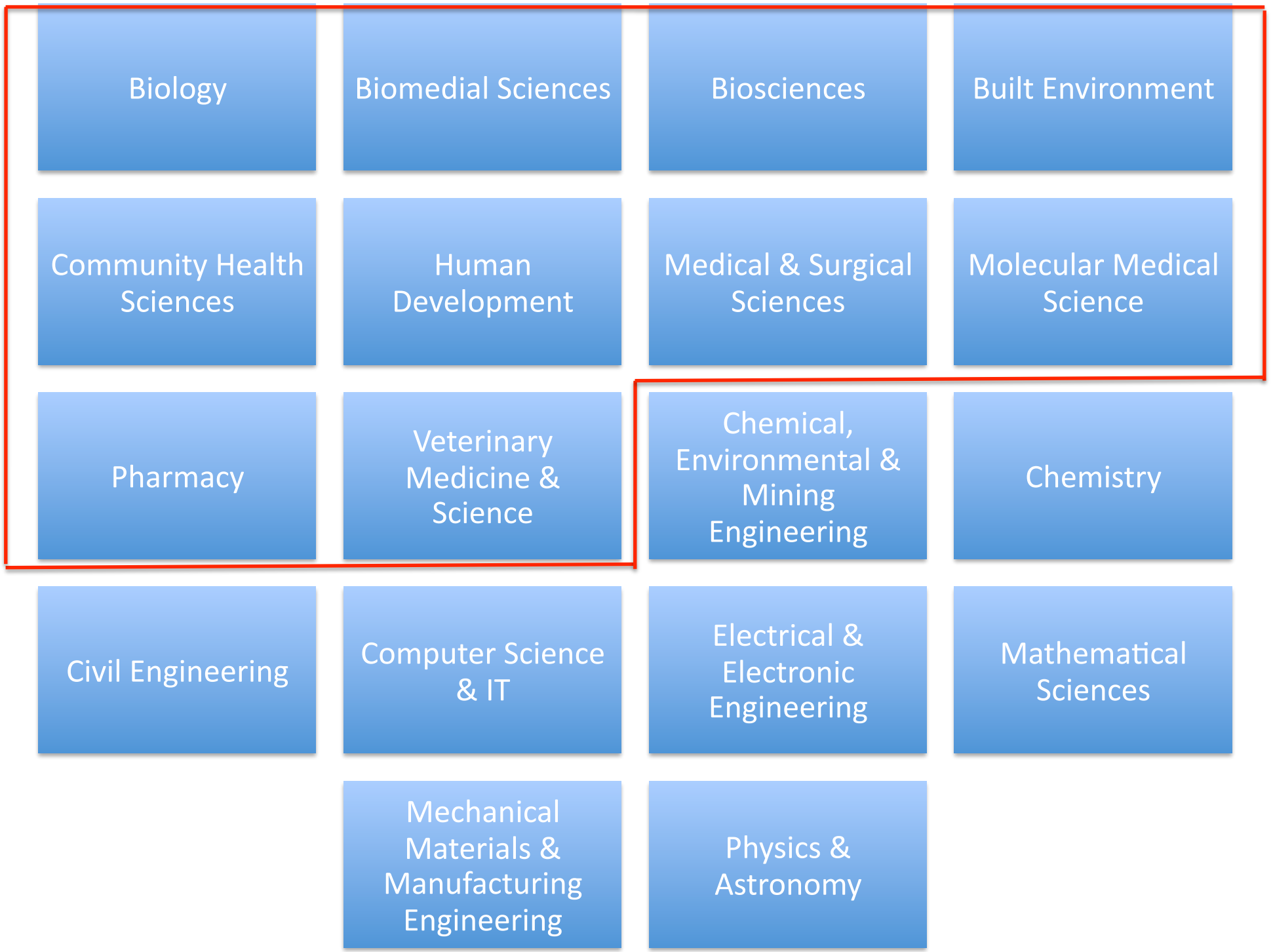

Business
Economics
Education
Biology
Biomedial Sciences
Law
Nursing
Politics & International Relations
Biosciences
Built Environment
Psychology
Sociology & Social Policy
Chemical, Environmental & Mining Engineering
Chemistry
Civil Engineering
Community Health Sciences
Computer Science & IT
Electrical & Electronic Engineering
Human Development
Mathematical Sciences
Mechanical Materials & Manufacturing Engineering
Medical & Surgical Sciences
Molecular Medical Science
Pharmacy
Physics & Astronomy
Veterinary Medicine & Science

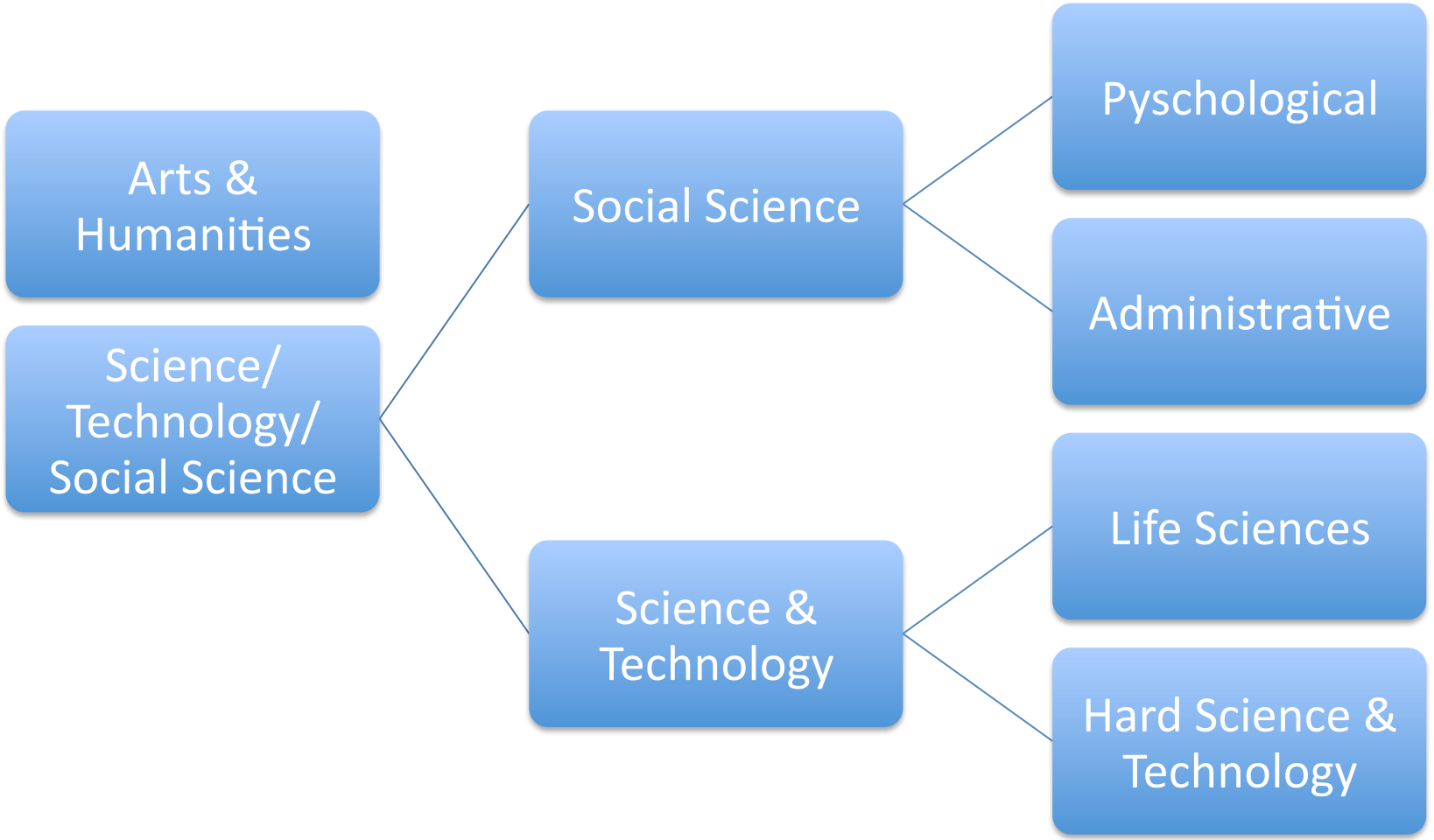

Arts & Humanities
Science/ Technology/ Social Science
Social Science
Science & Technology

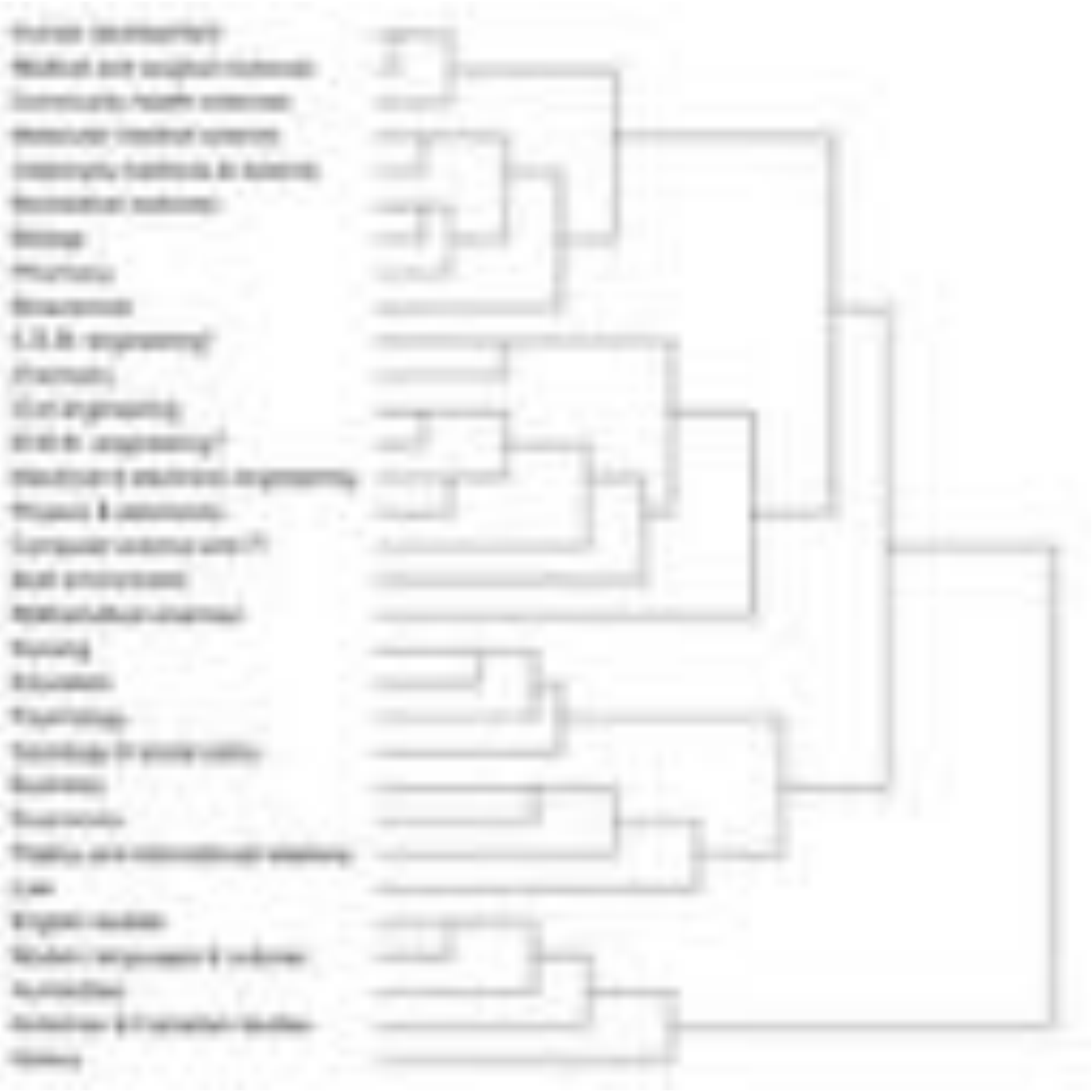

Business
Economics
Law
Biology
Biomedial Sciences
Politics & International Relations
Education
Nursing
Biosciences
Built Environment
Psychology
Sociology & Social Policy
Chemical, Environmental & Mining Engineering
Chemistry
Civil Engineering
Community Health Sciences
Computer Science & IT
Electrical & Electronic Engineering
Human Development
Mathematical Sciences
Mechanical Materials & Manufacturing Engineering
Medical & Surgical Sciences
Molecular Medical Science
Pharmacy
Physics & Astronomy
Veterinary Medicine & Science


Arts & Humanities
Science/ Technology/ Social Science
Social Science
Science & Technology
Psychological
Administrative


Biology
Biomedial Sciences
Biosciences
Built Environment
Chemical, Environmental & Mining Engineering
Chemistry
Civil Engineering
Community Health Sciences
Computer Science & IT
Electrical & Electronic Engineering
Human Development
Mathematical Sciences
Mechanical Materials & Manufacturing Engineering
Medical & Surgical Sciences
Molecular Medical Science
Pharmacy
Physics & Astronomy
Veterinary Medicine & Science


Biology
Biomedial Sciences
Biosciences
Built Environment
Community Health Sciences
Human Development
Medical & Surgical Sciences
Molecular Medical Science
Pharmacy
Veterinary Medicine & Science
Chemical, Environmental & Mining Engineering
Chemistry
Civil Engineering
Computer Science & IT
Electrical & Electronic Engineering
Mathematical Sciences
Mechanical Materials & Manufacturing Engineering
Physics & Astronomy


Arts & Humanities
Science/ Technology/ Social Science
Social Science
Science & Technology
Pyschological
Administrative
Life Sciences
Hard Science & Technology



Durrant, 2009

Defining disciplinary groups through clusters (2)

- For 86 discipline x level groups in BAWE, make lists of words with frequency > 100/m and appearing in > 10% of texts
- Make matrix of overlaps between each group
- Use hierarchical cluster analysis to identify groups

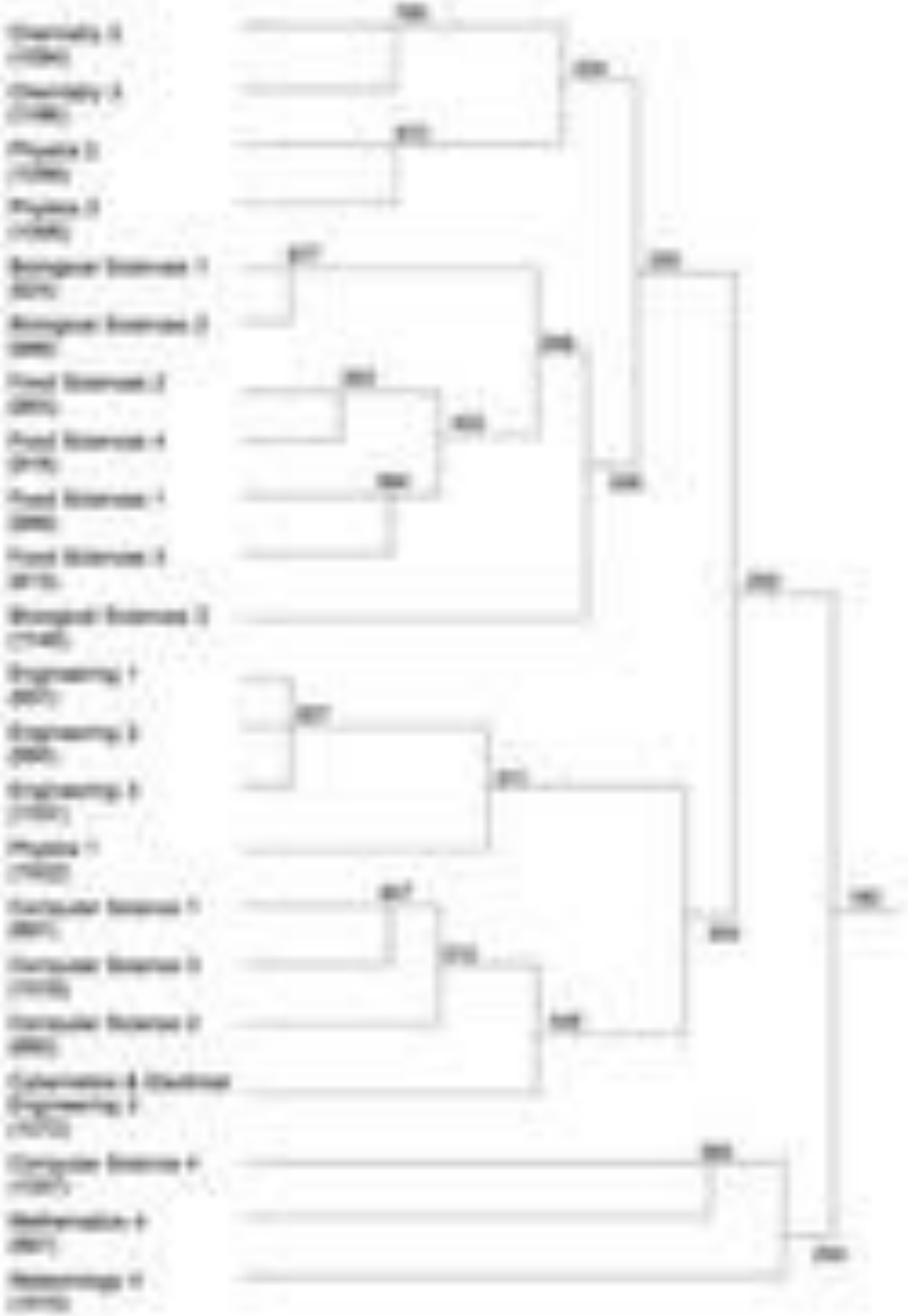
Durrant, 2014

Durrant, 2014

Durrant, 2014

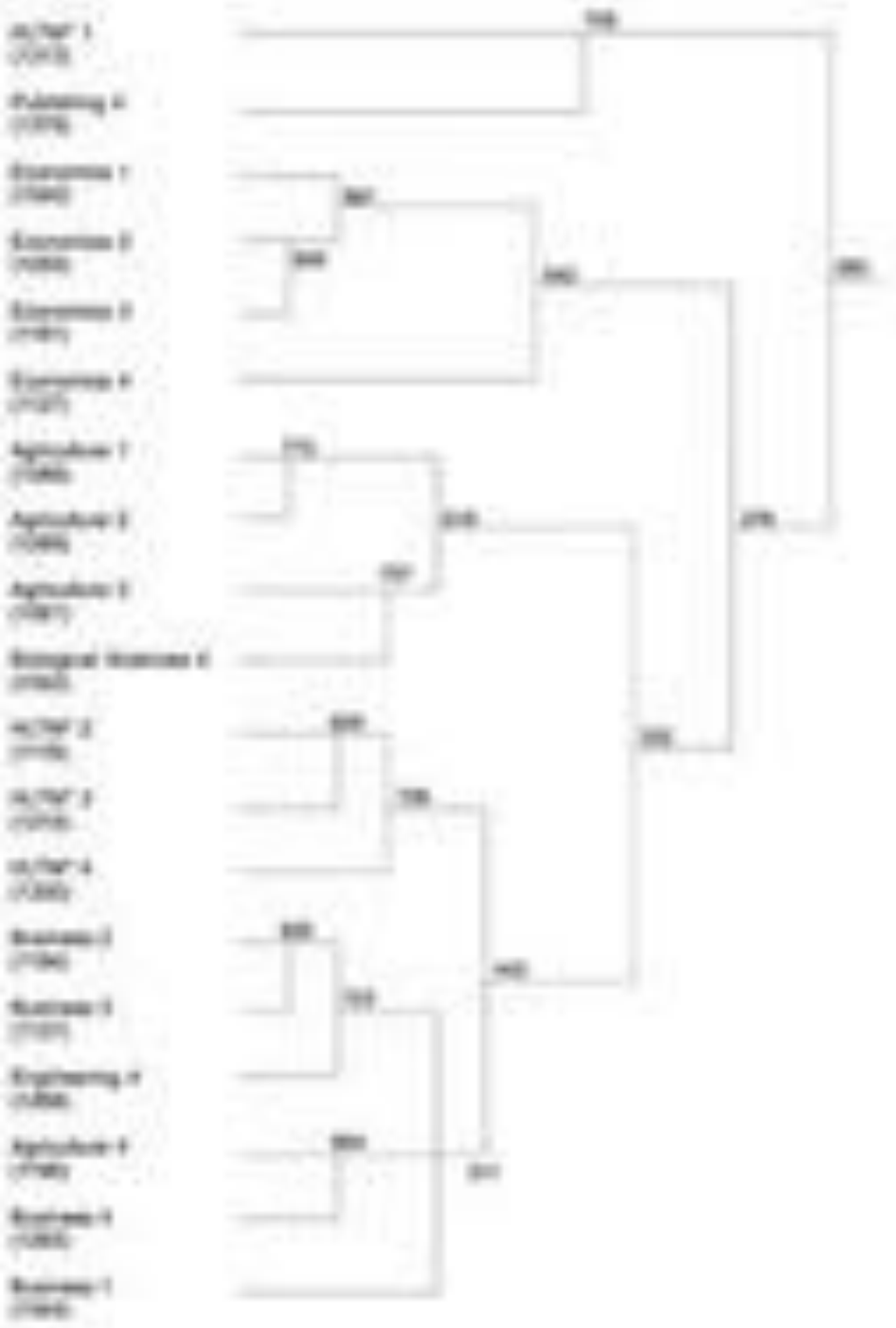
Durrant, 2014

Durrant, 2014



Durrant, 2014

Durrant, 2014



Durrant, 2014

The present…

- To what extent are disciplines good units for categorising student writing?
- Can we find alternative, emergent groupings?
- How can we characterize differences between disciplines/groups?

Why 4-grams?

- Easy to identify across large numbers of texts
- High-frequency 4-grams (*lexical bundles*) tend to have distinctive functions, allowing characterisation.

Caveats…

- Linguistic analysis does NOT give a ‘definitive’ chatacterization of disciplines.
- 4-gram analysis is NOT a ‘definitive’ characterization of linguistic variation.
- Analysis of student writing is NOT a ‘definitive’ analysis of disciplinary writing in general.
- This analysis conflates all texts written by each individual. Texts from different genres/levels/ assignment topics are sometimes combined.

Corpus

- BAWE
- Divided into text written by each author
- Authors with < 3 texts/6000 words excluded
- Authors with multiple disciplines excluded
- 285 authors; 24 disciplines

Analysis

- Writers assigned IDs: e.g. *agriculture_6001*
- Each writer's texts broken down into 4-grams:
  - *poland is predominantly an*
  - *is predominantly an agricultural*
  - *predominantly an agricultural country*
- Gives 285 lists of 4-grams (one per writer)

Analysis

- Find overlaps between each writer:

  Overlap = total shared 4-grams/total 4-grams

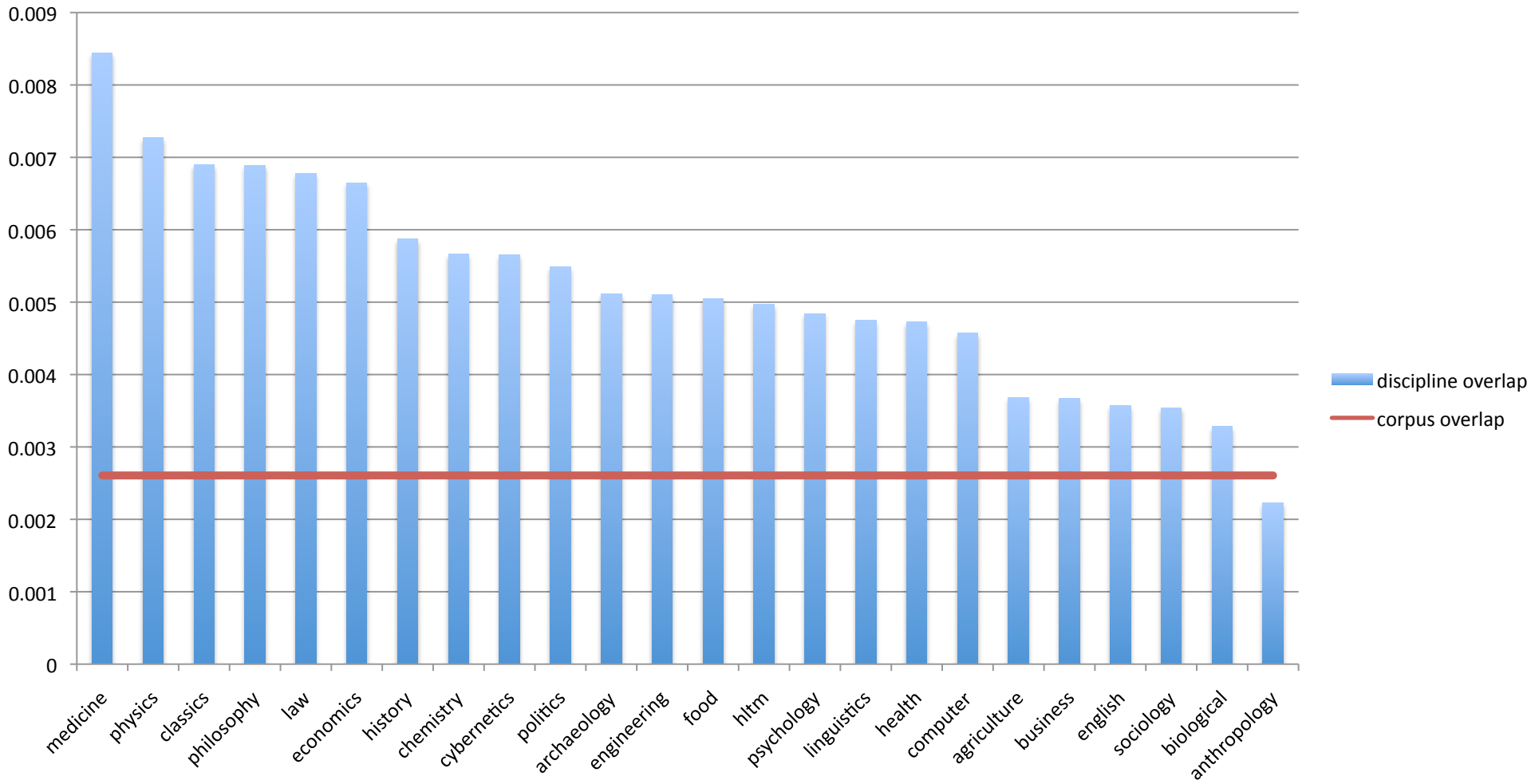
- Create matrix of overlaps:

	A	B	C	D	E
1	id	agriculture_6001.txt	agriculture_6015.txt	agriculture_6031.txt	agriculture_6034.txt
2	agriculture_6001.txt	1	0.001482321	0.002013648	0.003024453
3	agriculture_6015.txt	0.001482321	1	0.001717892	0.001718565
4	agriculture_6031.txt	0.002013648	0.001717892	1	0.003095975
5	agriculture_6034.txt	0.003024453	0.001718565	0.003095975	1

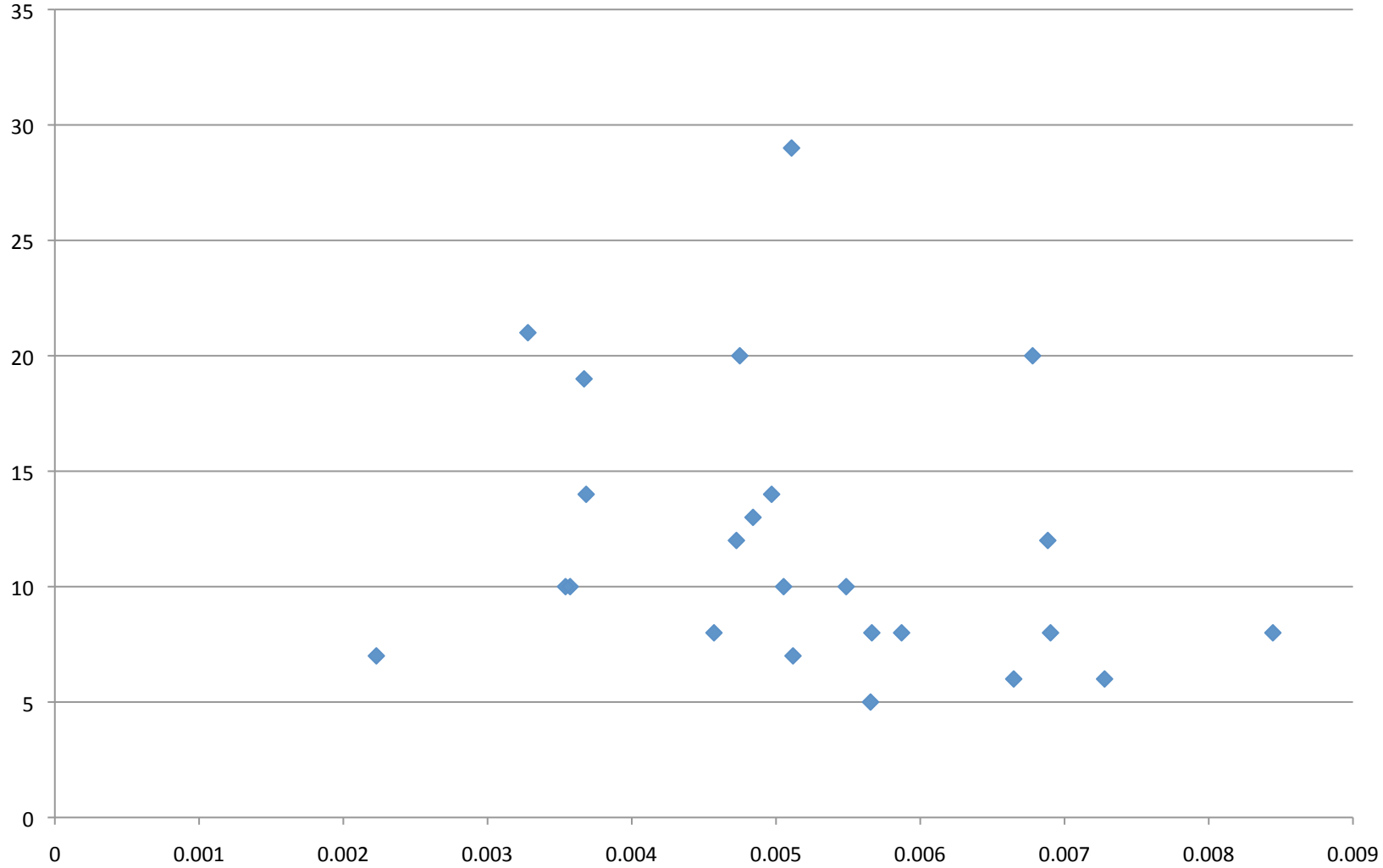
4-grams vs. *lexical bundles*

- The analysis is not restricted to *lexical bundles*
- *Lexical bundles* carry a great deal of weight in the analysis

How homogenous are disciplines?



Corpus size vs. overlap



r= -.24 p > .05

How do disciplines relate to each other?

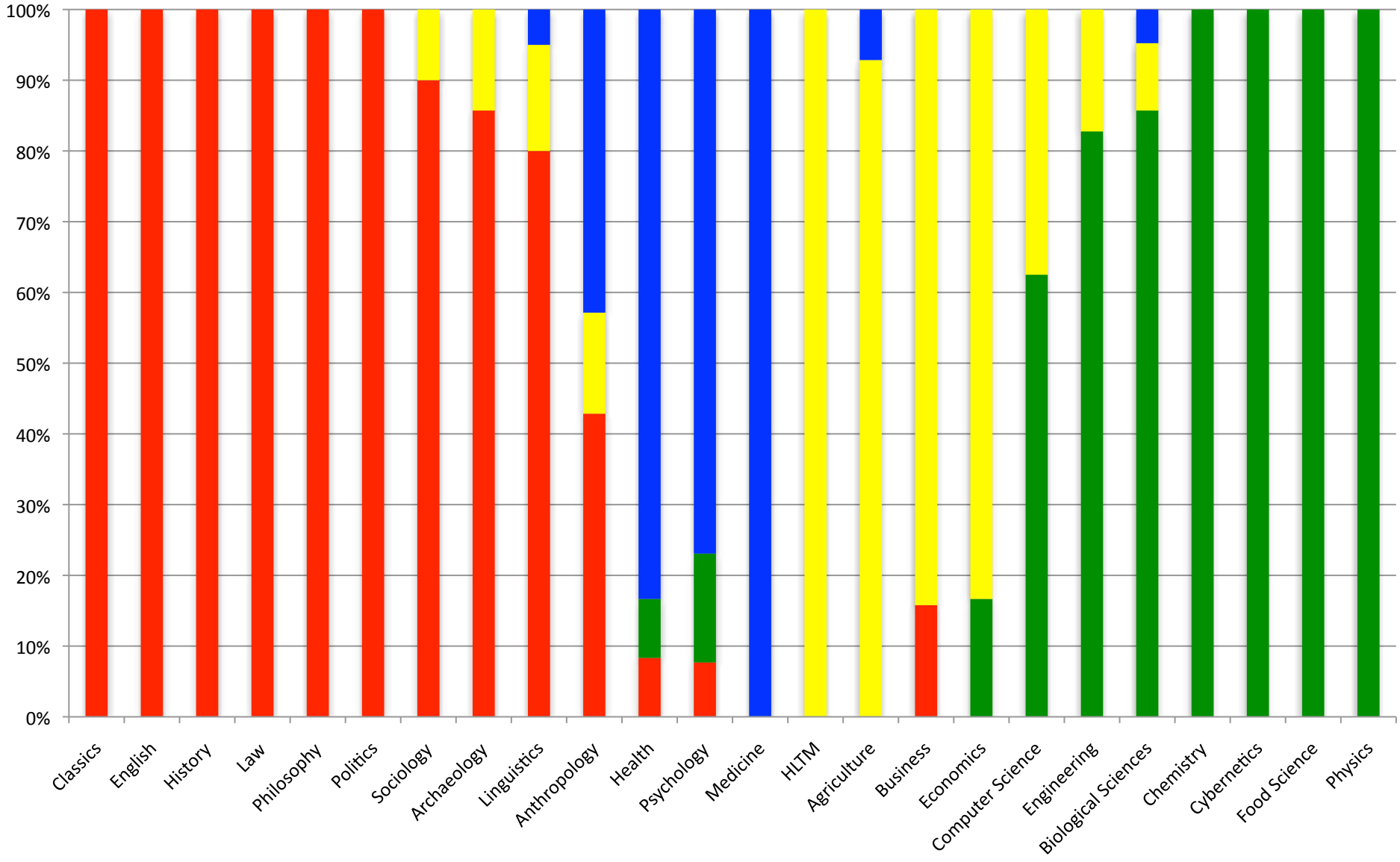
Are there emergent groups?

VOS Viewer

- Bibliometric tool designed for co-citation analysis
- Each unit (author) is represented by a dot
- Distances between dots represent degree of similarity:
  - each pair of units is given a weighting
  - the more similar the pair, the higher the weighting
  - map determined such that weighted sum of distances is minimized
  - average distance must = 1
- Cluster analysis represented by colours

van Eck & Waltman 2010

Clusters


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Classics
English
History
Law
Philosophy
Politics
Sociology
Archaeology
Linguistics
Anthropology
Health
Psychology
Medicine
HLTM
Agriculture
Business
Economics
Computer Science
Engineering
Biological Sciences
Chemistry
Cybernetics
Food Science
Physics

Differences in Discourse

- 4-grams which are:
  - used by >20% writers in the cluster
  - used by > twice the percentage of writers inside the cluster as outside

Differences in Discourse

Epistemic Bundles

	Red	Green
Hedged	*can be seen as* *it can/could be argued that* *appears to be a*	*be due to the*
Unhedged	*it is not a* *does not mean that*	*was found to be* *it was found that* *(this) is due to the* *the reason for this* *this is because the* *(it) can be seen (that) (the)* *this means that the*

Differences in Discourse

Directives

Blue	Yellow
there is a need *should be able to* *it is important for*	*should be carried out* *also need to be* *it is (also) important (to)*

- *…customers **should be able to** receive their buffet food swiftly* (Business)
- *the stigma of having major surgery and resulting in a stoma bag may **also need to be** addressed* (Health)

Differences in Discourse

Text structuring bundles

	Red	Blue	Yellow	Green
Topic signalling		*to look at the*		*the aim of this*
Text deixis				*as can be seen (from)* *it/this can be seen* *as shown in Figure N* *is shown in Figure N*

Differences in Discourse

Concept structuring

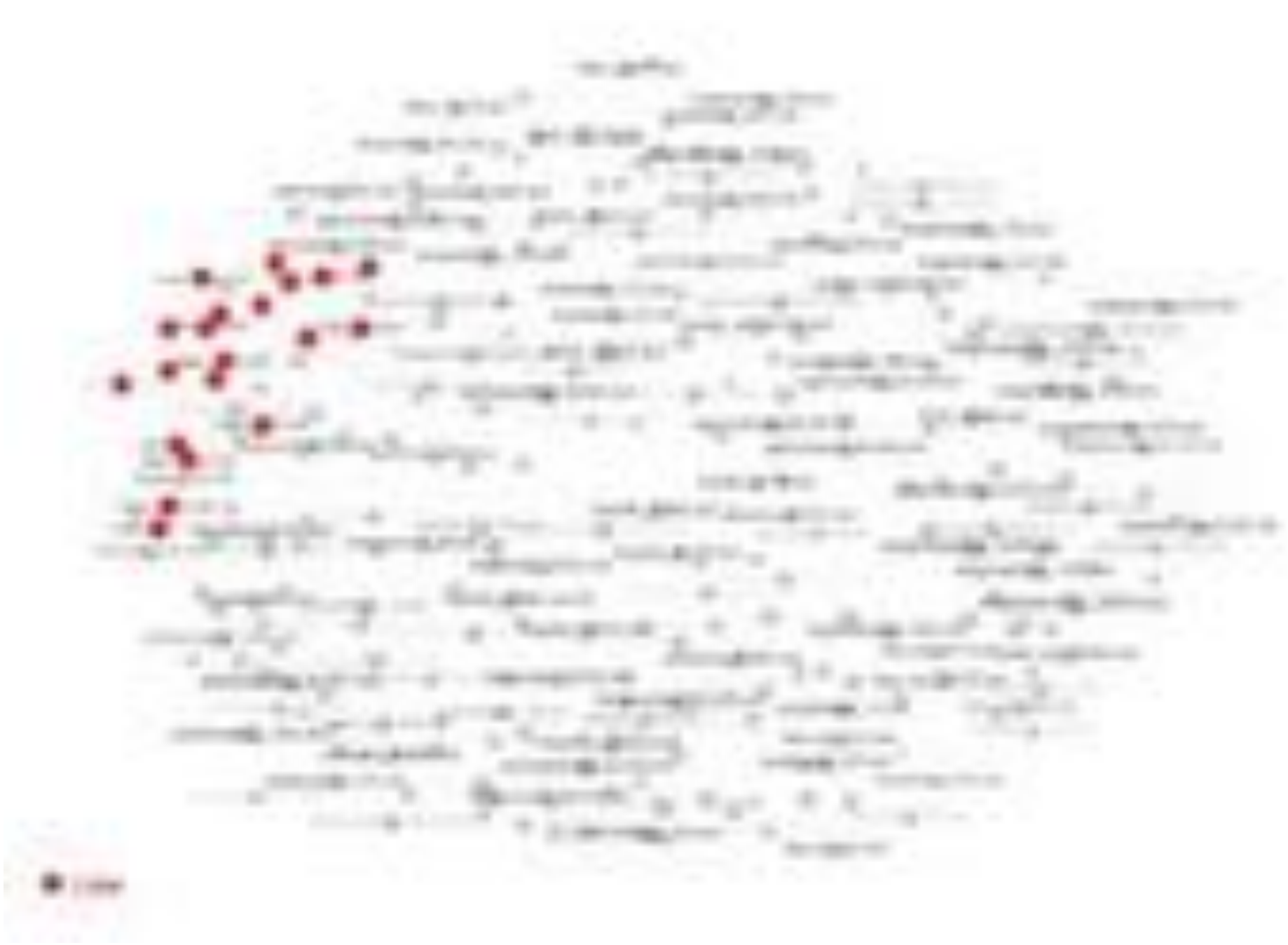
	Red	Green
Contrast	*as opposed to the* *in contrast to the* *despite the fact that* *in favour of the* *on the other hand* *in relation to the*	*than that of the*
Equivalence/consistency		*is the same as (the)*
Specifying a scope	*in relation to the* *for the purposes of* *in the context of*	*with respect to the*
Specifying an interpretive/ explanatory context	*in the light of the* *in relation to the* *in the context of*	*is defined as the* *an example of a*

Differences in Discourse

Referential

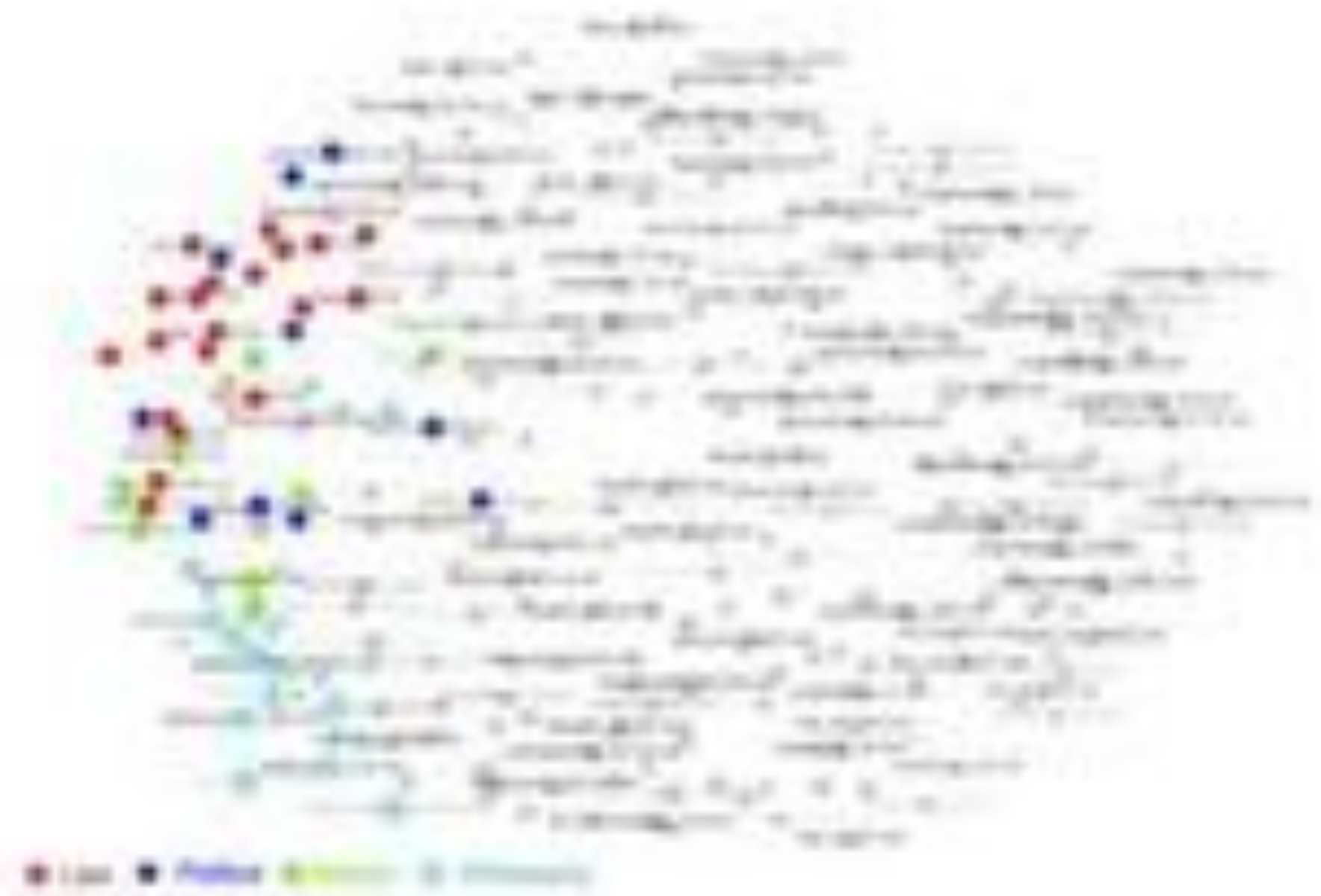
	Red	Blue	Green
Quantification	the the extent that	*that the majority of* *is the most common* *of a number of* *there was no significant*	*a large number of* *the difference between* *the*
Tangible framing			*the length of the* *the shape of the* *the size of the*
Intangible framing	*the power of the* *the nature of the* *the role of the* *the form of the* *of the concept of* *the idea of a/the* *the existence of the* *the validity of the* *the fact that*	*the meaning of the* *the presence of a* *the need for a* *and the use of* *in the treatment of*	*the presence of a/the* *the efficiency of the*

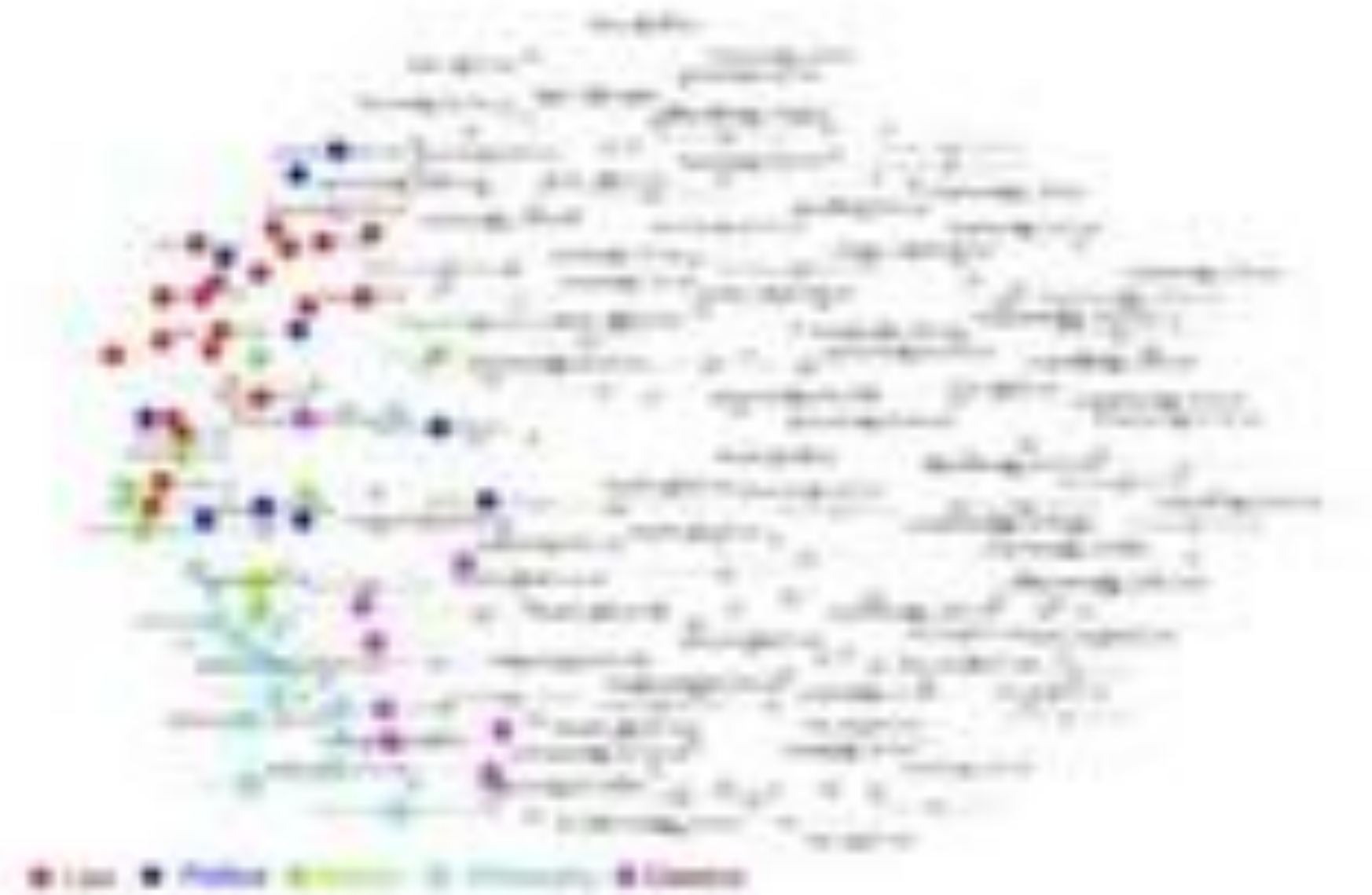
Exploring the map



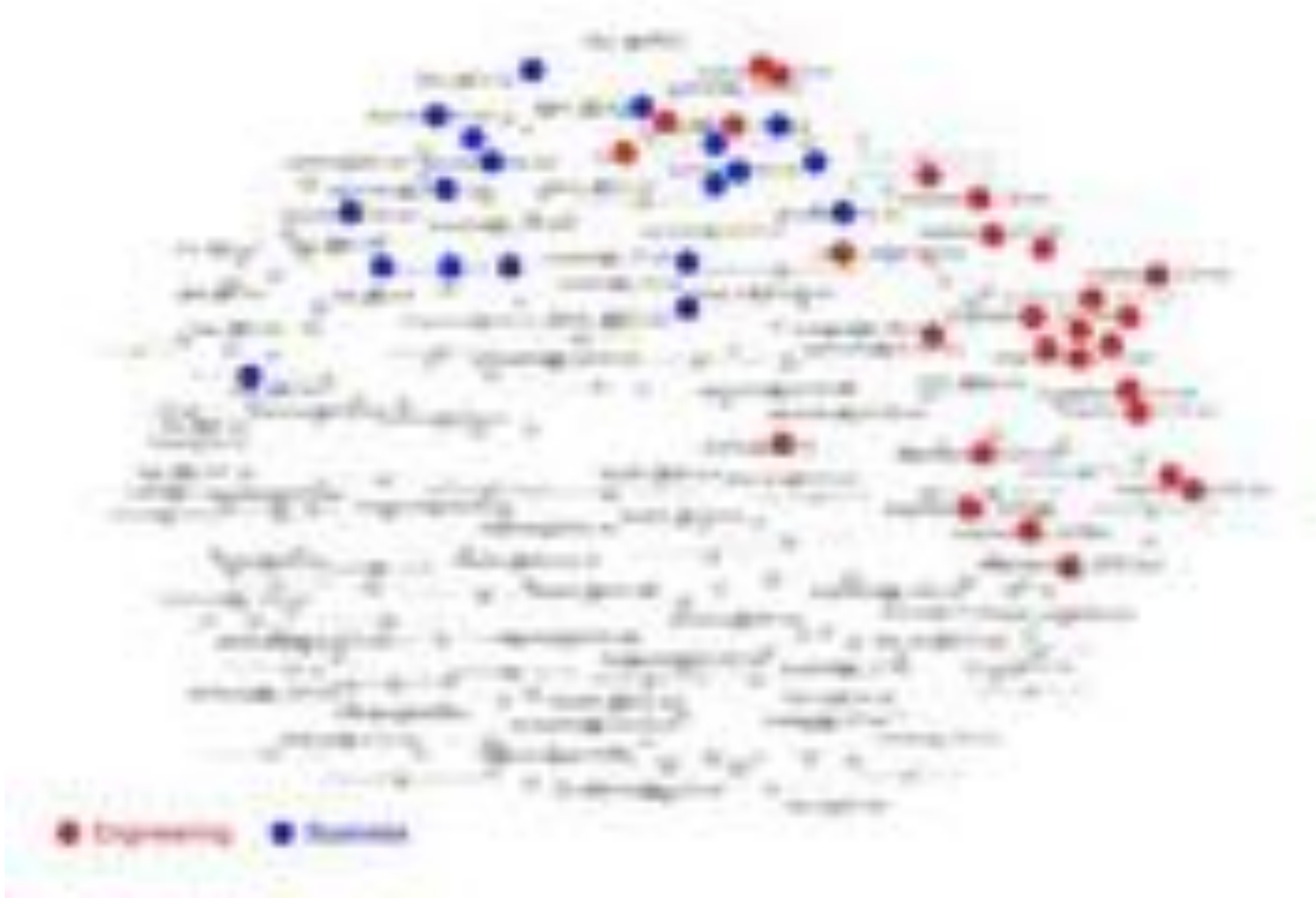

Liver
Politics

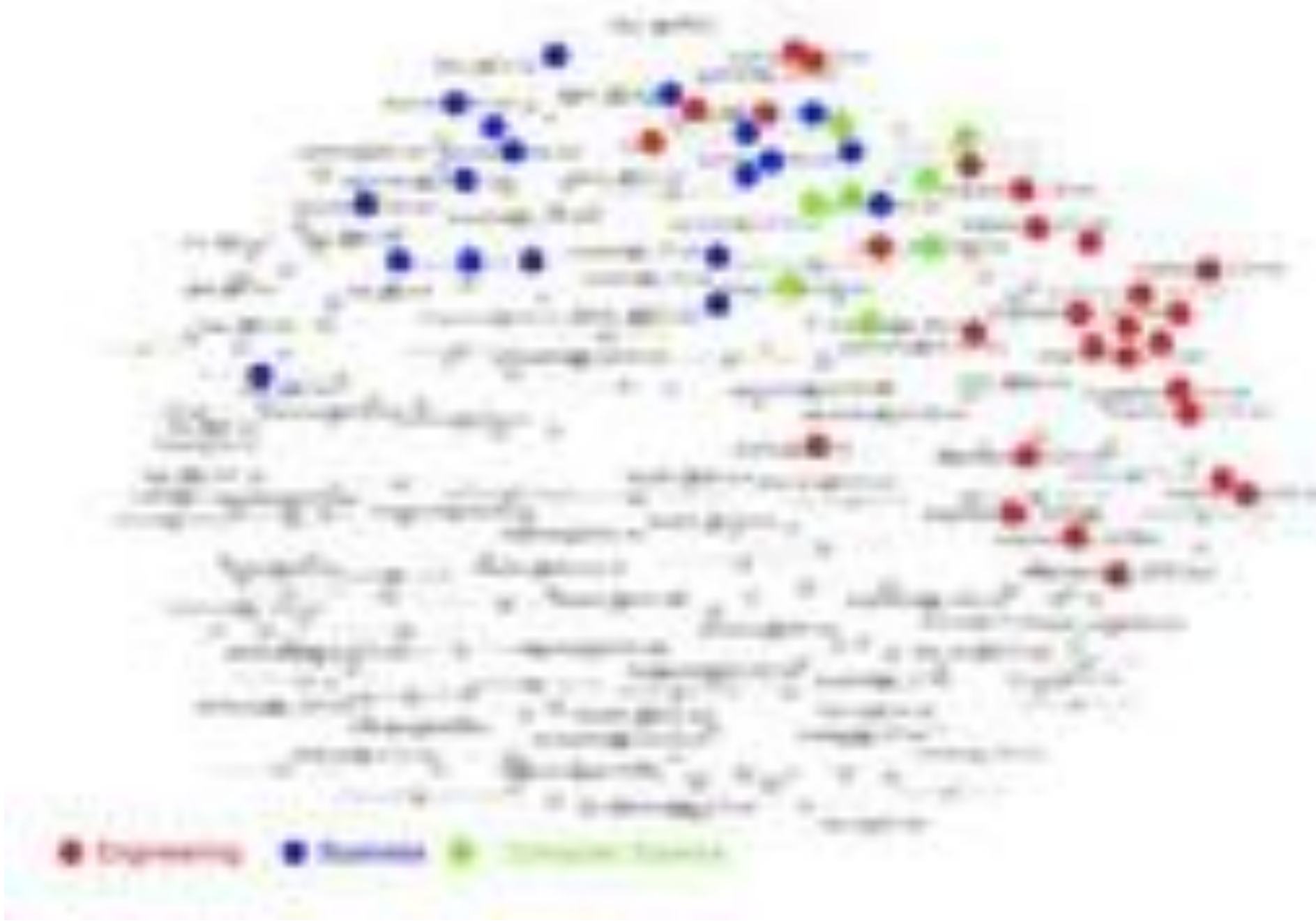

Liver
Pistils

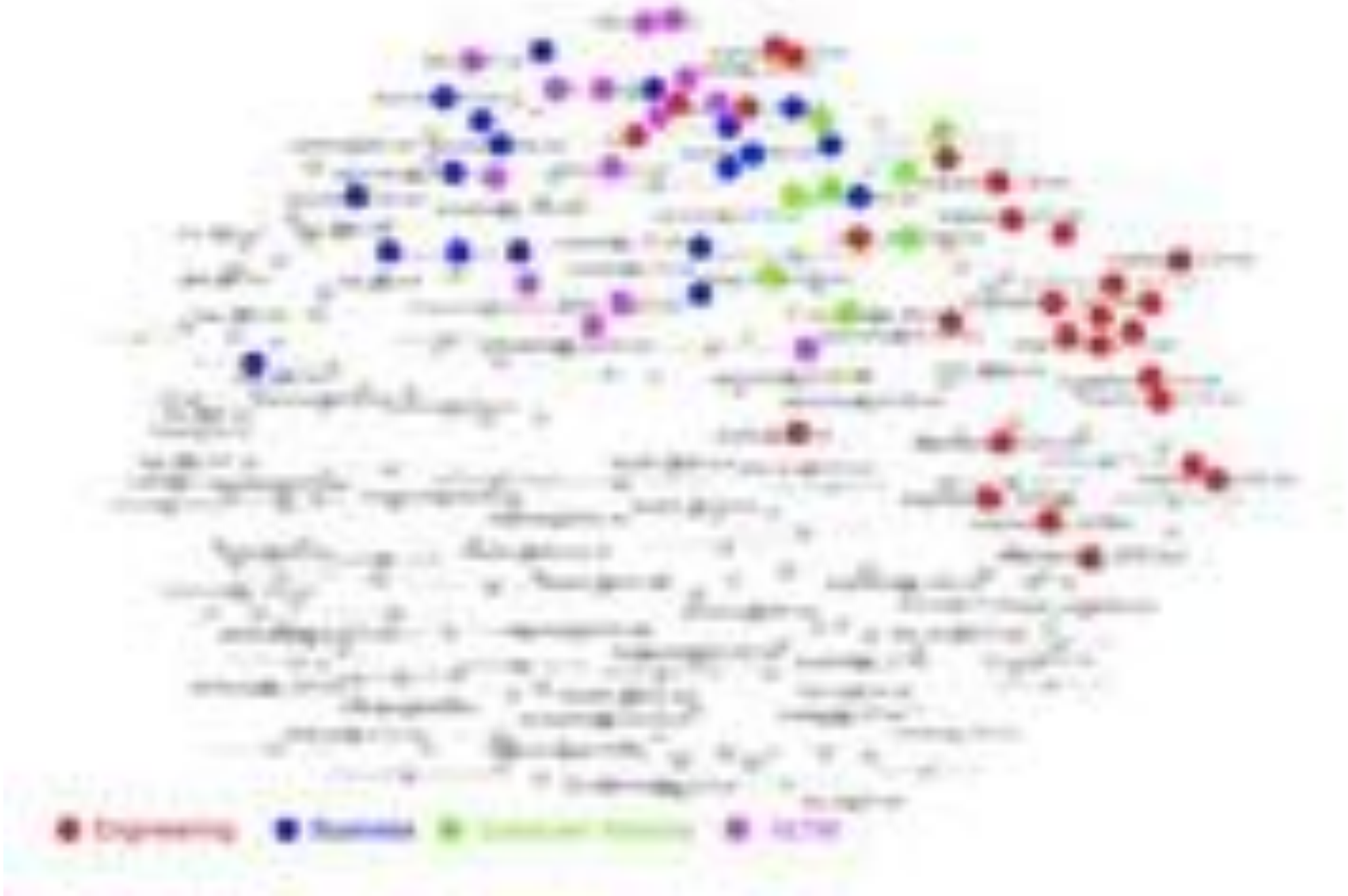

Loss
Positive

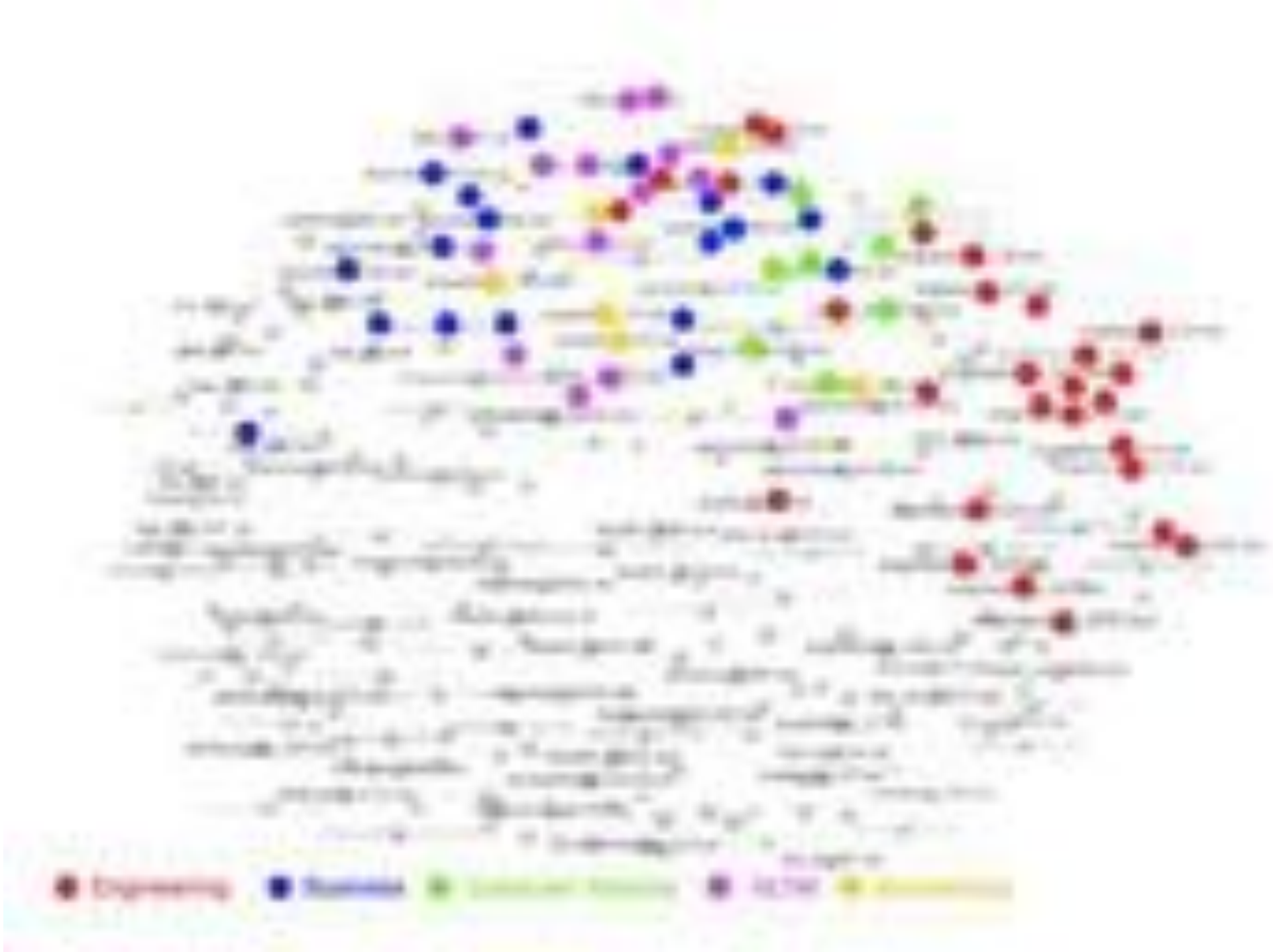


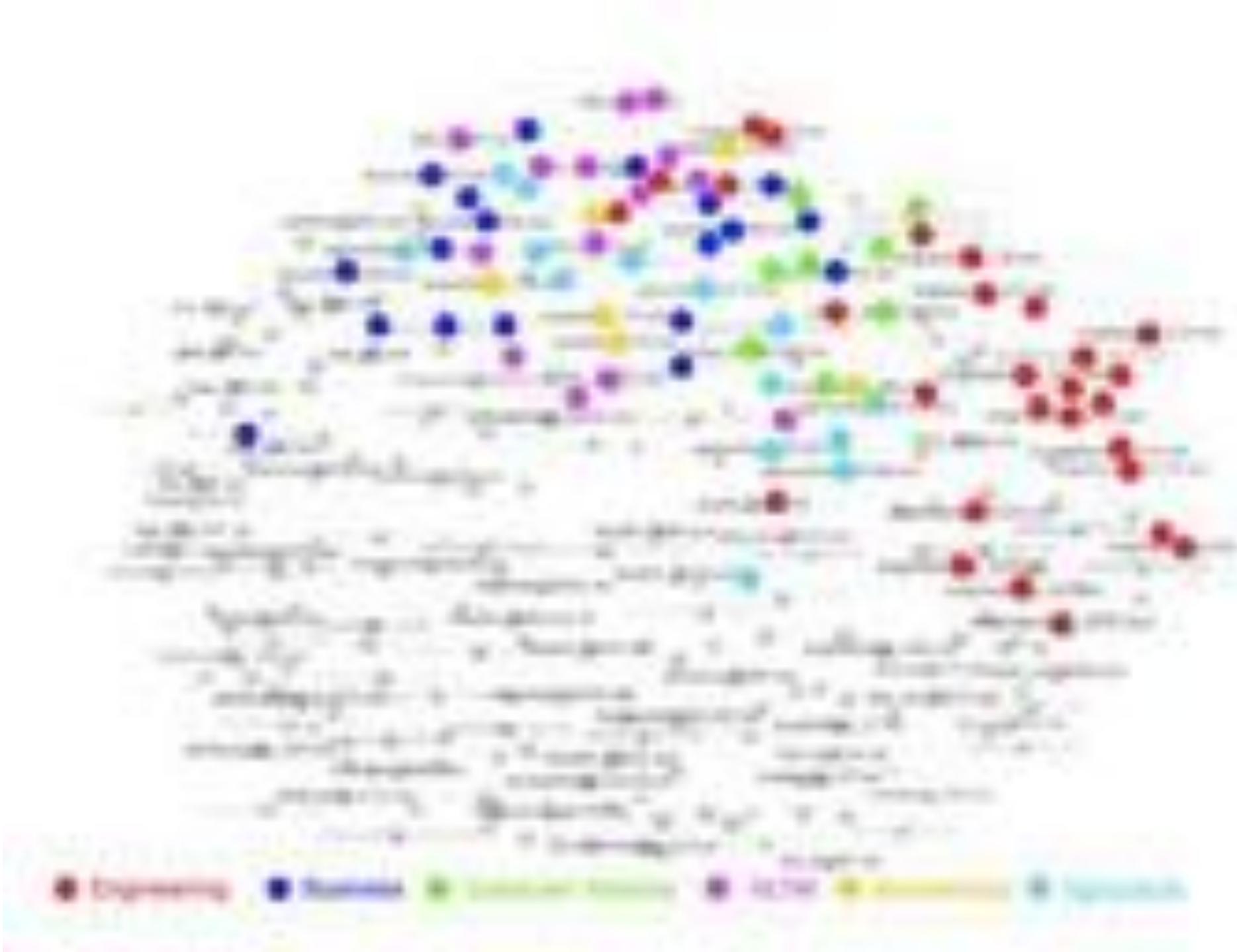

Engineering


Engineering
Statistics

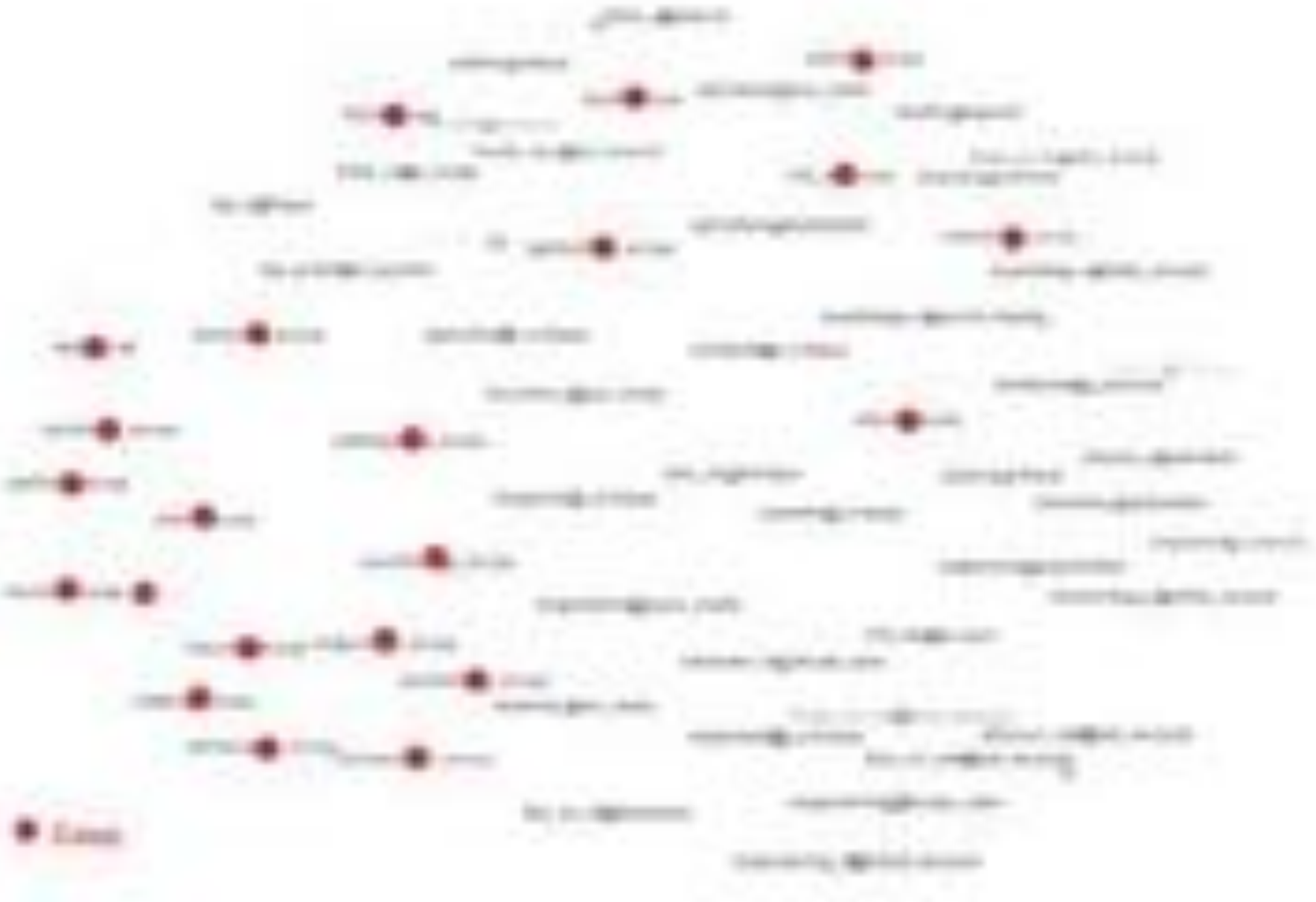

Engineering
Statistics

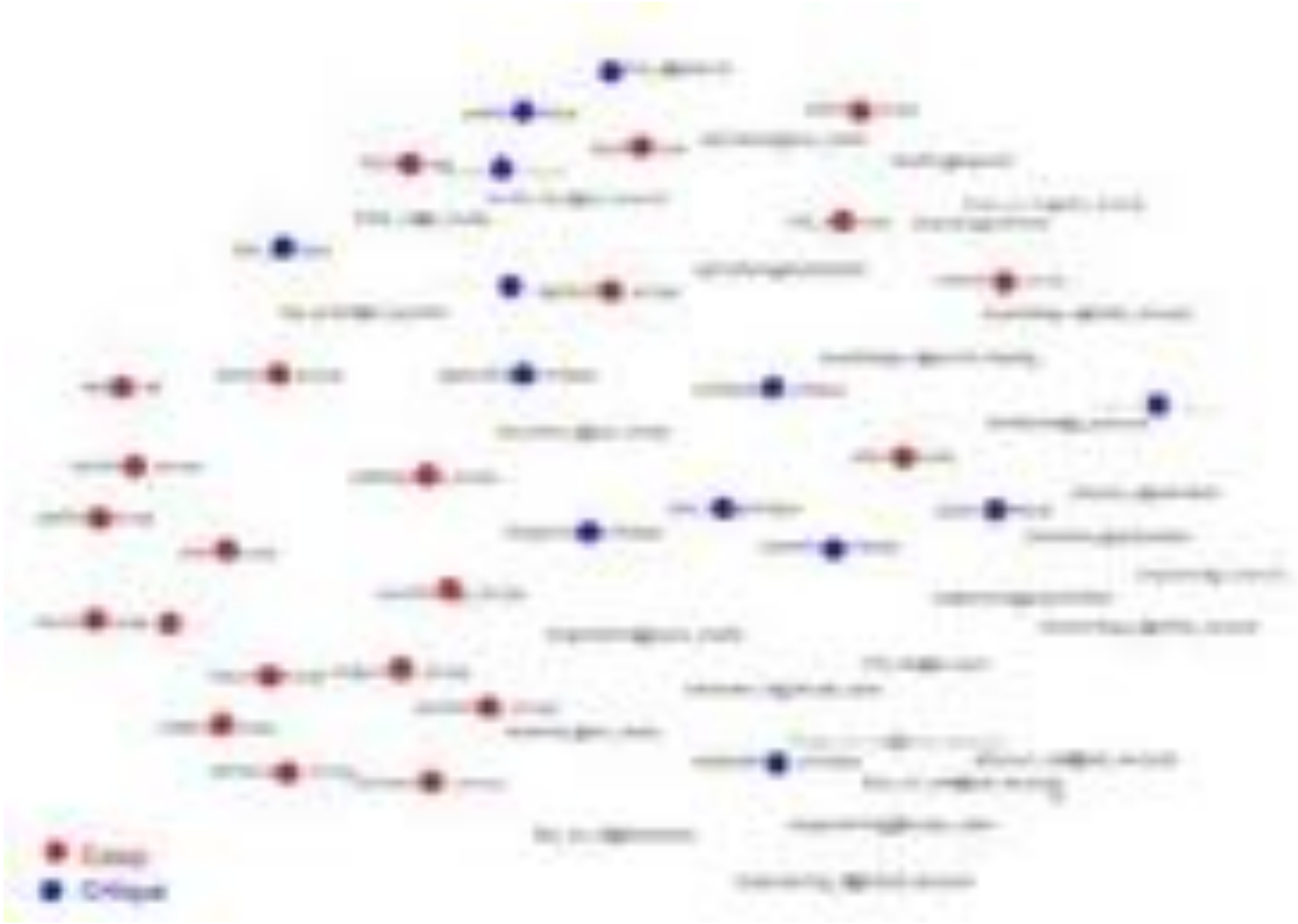

Engineering
Statistics

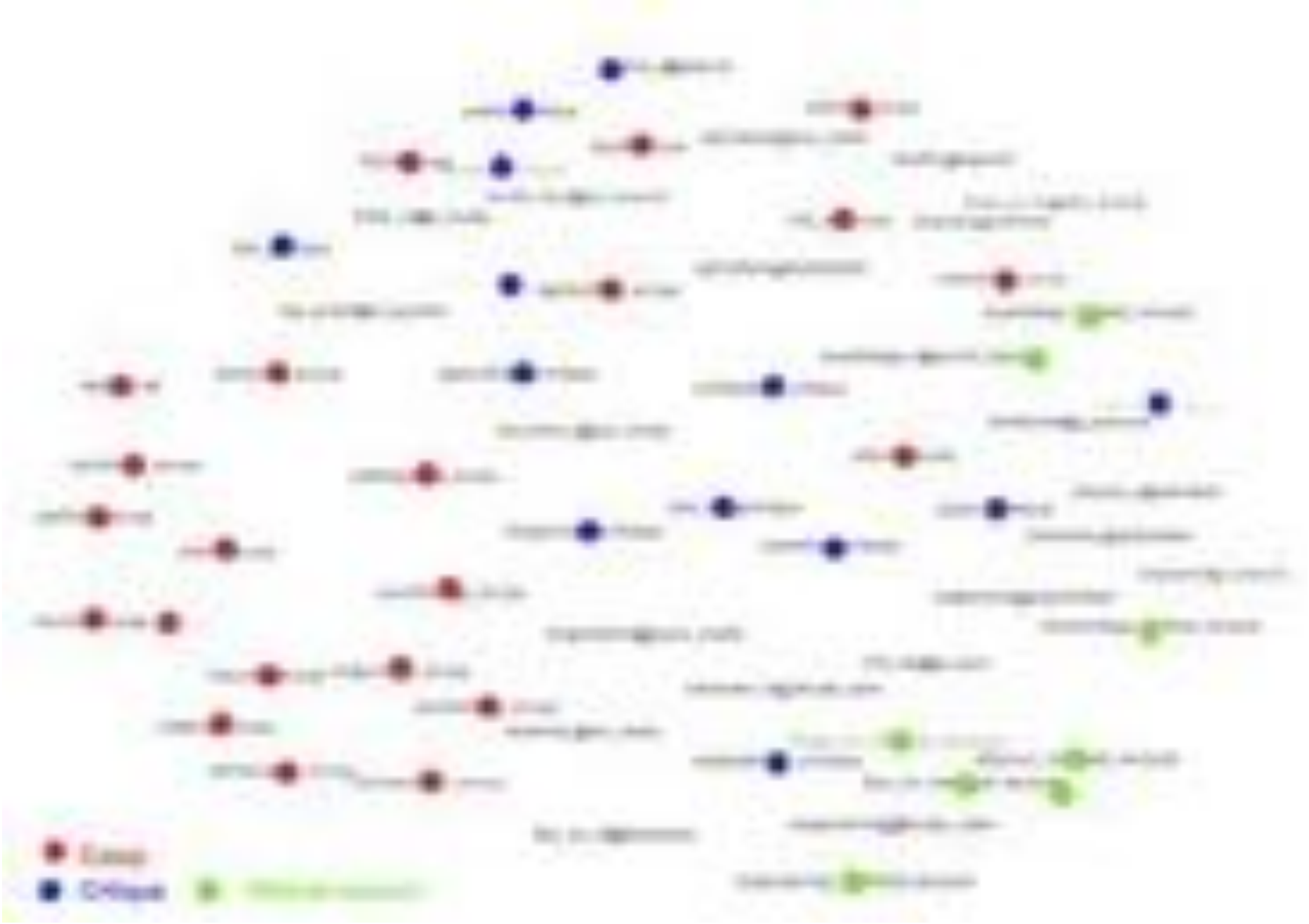

Engineering
Statistics

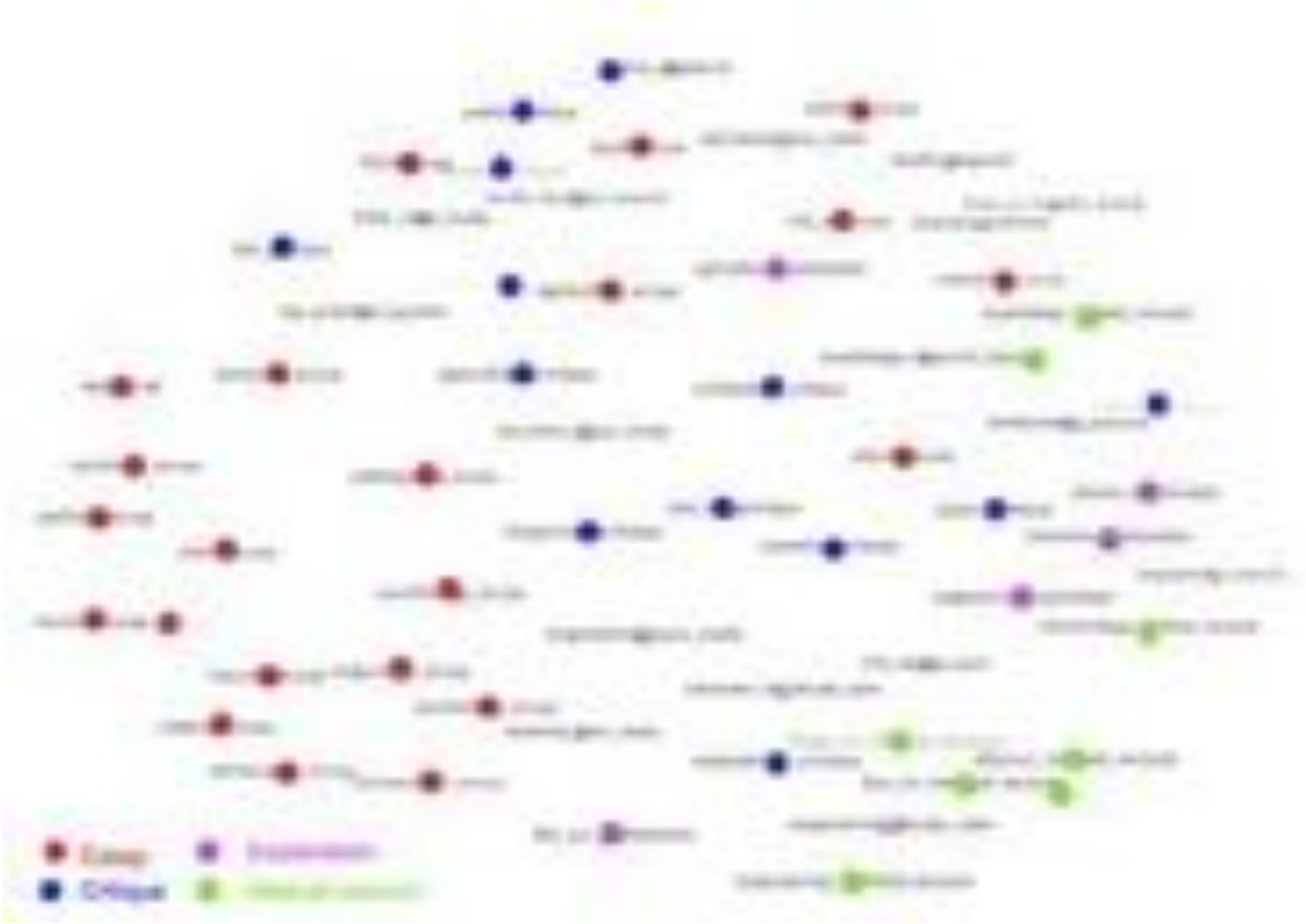

Engineering
Statistics

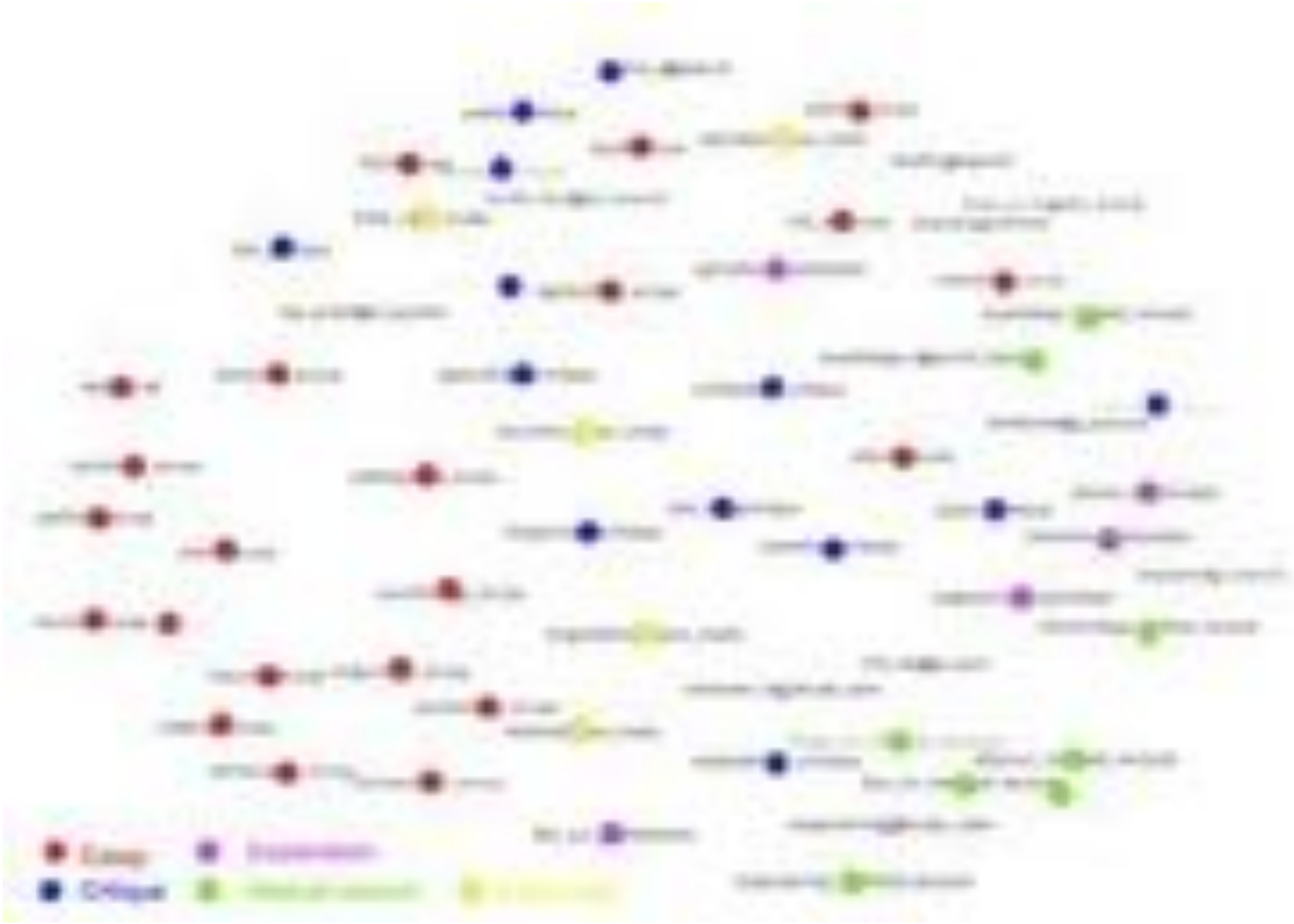
Discipline x Genre


Events


Essay
Critique


Loop
CHIqua


Loop
Chirps


Loop
CHique

Conclusions

- *Most* disciplines (as defined in BAWE) are relatively coherent units
- The hard vs. soft knowledge distinction is highly relevant for grouping disciplines
- This is a 'middle ground' of *Commerce* and *Health*-related disciplines which can draw on either side (but not both!)
- Maps can help us:
  - understand relationships between disciplines
  - identify heterogeneity within disciplines

References

Becher, T., & Trowler, P. (2001). *Academic tribes and territories: Intellectual enquiry and the cultures of disciplines*. Buckingham: SRHE and Open University Press.

Brew, A. (2008). Disciplinary and interdisciplinary affiliations of experienced researchers. *Higher Education*, 56: 423-438.

Biber, D. (2006). *University Language: A corpus-based study of spoken and written registers.* Amsterdam: John Benjamins.

Coxhead, A. (2000). A new academic wordlist. *TESOL Quarterly*, 34(2): 213-238.

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for Academic Purposes. *English for Specific Purposes*, 28(3): 157-169.

Durrant, P (2014). Discipline- and level-specificity in university students' written vocabulary. *Applied Linguistics*, 35(3): 328-356.

Gardner, D. & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics*, 35(3): 305-327.

Manathunga, C., & Brew, A. (2012). Beyond Tribes and Territories: *New Metaphors for New Times*. In P. Trowler, M. Saunders & V. Bamber (Eds.), *Tribes and Territories in the 21st Century* (pp. 44-56). London: Routledge.

Simpson-Vlach, R. & Ellis, N.C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, 31(4): 487-512.