Mapping the Minds of Many:

Methodological Principles for Analysing Large Bodies of the Research Literature through Unsupervised Text Analytics

Rens Scheepers
Jacob Cybulski

AACSB ACCREDITED

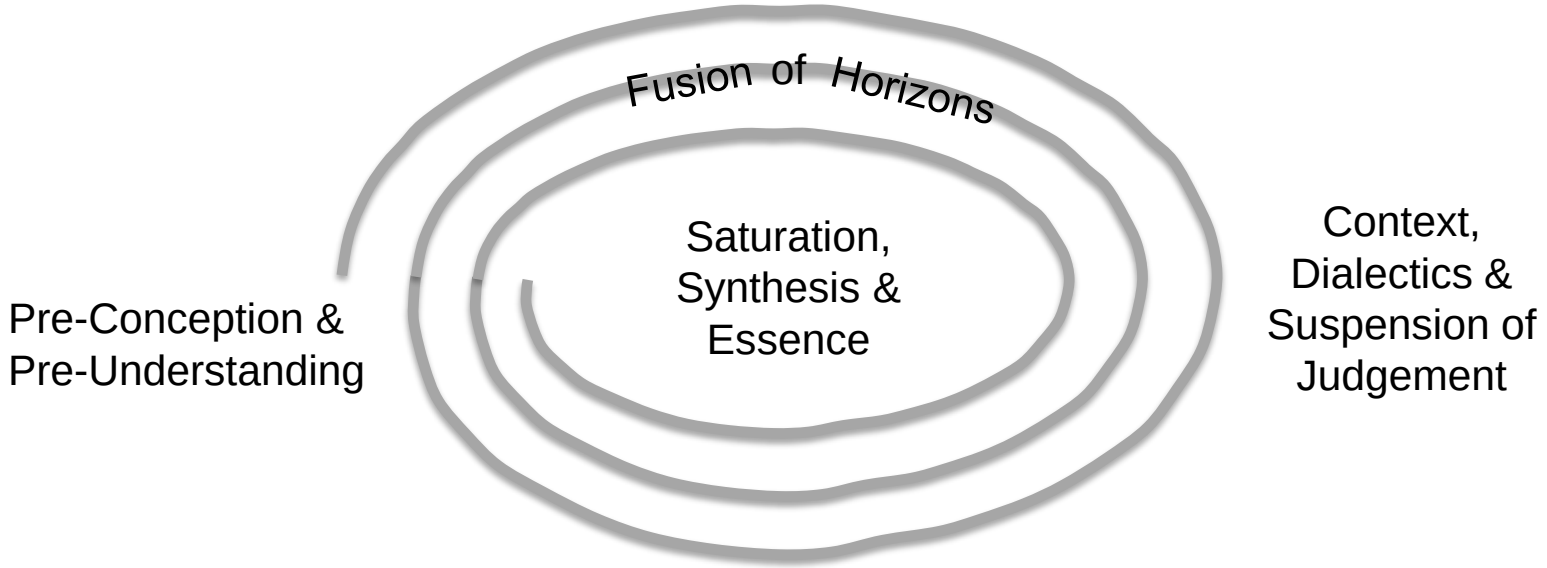
EQUIS ACCREDITED

Introduction

- Text analytics are enabling analyses of large bodies of research literature.
- Unsupervised text analytics can yield results beyond human-directed coding by “letting the invisible college speak”, rather than imposing a researcher’s pre-conceived understanding on the process.
- Based on our experience of this process, advice in the research methods literature, and hermeneutic principles, we are currently developing our learning into a more generic method.

Principles of Hermeneutics

- Origin of hermeneutics (interpretation of ancient texts across different time/cultural horizons)
- Modern hermeneutics adapted for interpretive analysis (Klein & Myers, 1999)
- *"..The critical task of hermeneutics then becomes one of distinguishing between "true prejudices, by which we understand, from the false ones by which we misunderstand""* (Gadamer)
- The hermeneutic circle is utilized as a methodological device (Klein & Myers, Sarker & Lee, 1999)



Hermeneutics in Action

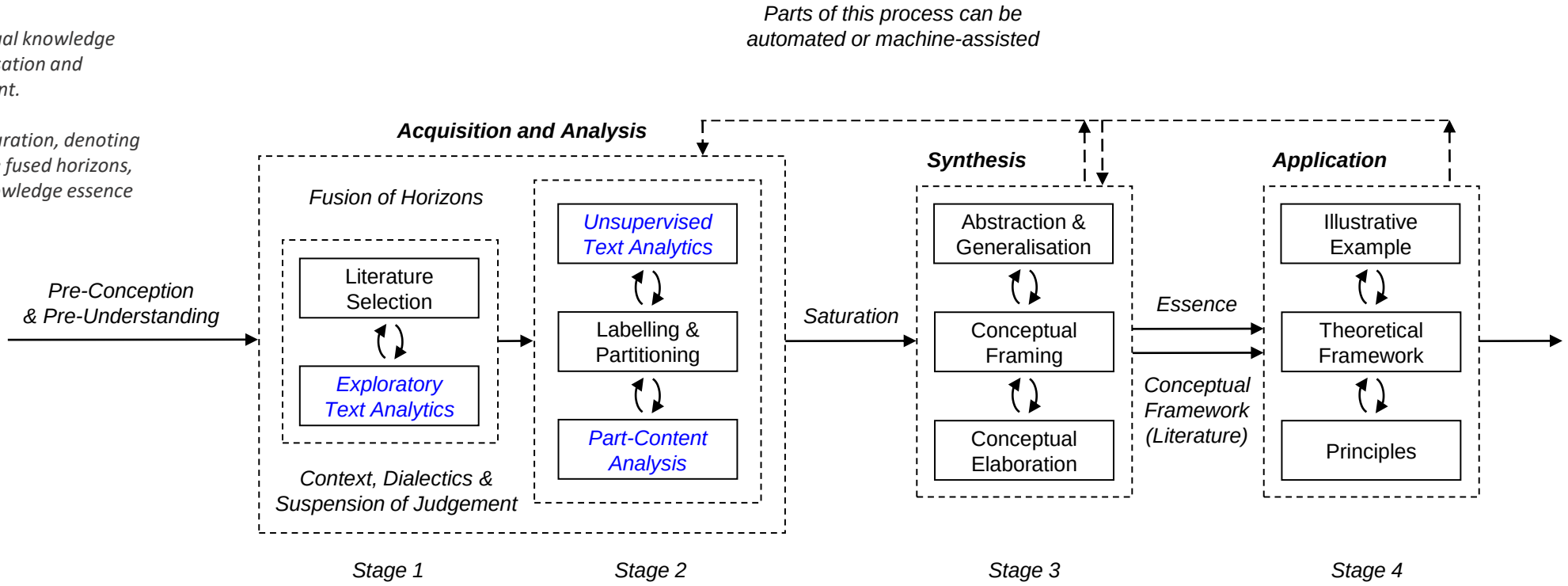
Literature Review Process

This literature review process enacts the principles of hermeneutics.

The aim of this process is to develop shared knowledge between co-researchers by fusing highly contextualised views and opinions collected from different sources and from differing vantage points.

The process is cyclical to allow gradual knowledge acquisition, its questioning, crystallisation and reflection, without any pre-judgement.

The process ends at the point of saturation, denoting consistency and completeness of the fused horizons, thus allowing formulation of the knowledge essence and the potential application.



Pre-understanding & Pre-conception

Data science in an organisational context entails *understanding phenomena via the analysis of data with the ultimate goal of improving decision making* (Provost and Fawcett 2013, p53)

So, it involves obtaining ***data*** pertinent to the phenomenon of interest, applying data processing ***methods***, interpreting results typically by means of visualisation and interaction via ***interfaces***, in support of human ***cognition*** to inform problem-solving or decision-making.

Domain	Concepts	References
Data	Processing a variety of data of different structure, size, granularity and dynamics.	(van der Aalst 2016, 10; Donoho 2015)
Method	Adopting systematic methods	(Pierson 2015, ch 1)
	Using quantitative and qualitative methods	(Waller & Fawcett 2013)
	Interdisciplinary field of statistics, informatics, computing, comms, management, and sociology	(Cao 2017a; 2017b; Aalst 2016, 10; Cleveland 2014)
	Developing new methods of dealing with very large volume of data	(Karpatne et al. 2017)
	Developing novel algorithms	(Berman et al. 2018)
	Extracting, preparing, exploring, transforming, storing, retrieving and applying data	(Aalst 2016, 10; Berman et al. 2018; Provost & Fawcett 2013a)
	Extracting information and knowledge from data in a generalised / principled fashion	(Dhar 2013; Provost & Fawcett 2013a)
	Facilitating data collection, organisation, stewardship and preservation	(Berman et al. 2018)
	Deriving insights, models and inferences	(Cleveland 2014; De Veaux et al. 2017; Provost & Fawcett 2013b, 2)
Interface	Visualising data and delivering insights	(Aalst 2016, 10)
	Understanding a problem, comm of results, and facilitating human input into data analysis	(Blei & Smyth 2017)
Cognition	Supporting people in predicting, making decisions, modelling and generating insights	(Aalst 2016, 10; Pierson 2015, ch 1; Ozdemir 2016, ch 1)
	Acquiring knowledge, intelligence and wisdom from data	(Dhar 2013; Cao 2017a; Pierson 2015, ch 2; Bostrom 2014)
	Viewing business problems from a data perspective	(Provost & Fawcett 2013a)
	Analysing and understanding real world phenomena with data	(C. Hayashi 1998; Pierson 2015, ch 1 and 2; Provost & Fawcett 2013a)
	Offering a systematic discipline-specific problem-solving framework	(Provost & Fawcett 2013a; Blei & Smyth 2017)
	Formulating and evaluating solutions in their business context	(Provost & Fawcett 2013b, ch 2)
	Improving decision making	(Provost & Fawcett 2013a)
	Using “data science thinking”, i.e. by blending statistical and computational thinking	(Cao 2017b; Blei & Smyth 2017)
	Human is required to understand a domain, select data, models and methods	(Blei & Smyth 2017)

Stage 1A: Acquisition

Literature Selection

In total, 41 academic journals and professional magazines were selected:

- Basket of 8 IS journals
- 8 top IT / Business journals
- Plus journals recommended by EBSCOhost

Publications spanning 7 years 2014-2020 (Big Data 2.0: data science in business)

Top 16 journals were systematically scanned issue-by-issue for papers.

The journal database was further searched by selected keywords.

Journals Selected for Review of Data Science Papers	
Basket of 8 IS Journals	***Number of Papers***
European Journal of Information Systems	7
Information Systems Journal	5
Information Systems Research	1
Journal of the Association for Information Systems	9
Journal of Information Technology	13
Journal of Management Information Systems	5
Journal of Strategic Information Systems	9
Management Information Systems Quarterly	14
Journals with Organizational Context Emphasis	
Communications of the ACM	36
Communications of the Association for Information Systems	32
Decision Support Systems	63
Harvard Business Review	30
Human-Computer Interaction	1
Management Science	9
MIT Sloan Management Review	93
Organization Science	9
Journals focusing on Data Science Topics with Organizational Relevance	
Such as ACM and IEEE journals and transactions, EJOR, ES with Apps,…	88
TOTAL Number of selected papers for review	**424**
TOTAL Selected for automatic analysis after preliminary review	**294**

EBSCOhost Keywords Searched:

- Data Science
- Data Analytics
- Data Analysis
- Business Analytics
- Business Analytics
- Predictive Analytics
- Decision Analytics
- Machine Learning
- Deep Learning
- Data Mining
- Data Engineering
- Big Data
- Data Visualisation
- Information Visualisation
- Interactive Visualisation
- Visual Analytics
- Decision Support
- Group Decision Support
- Human-Computer Interaction
- Intelligent Interfaces
- Distributed Cognition
- Artificial Intelligence
- Expert Systems
- Knowledge-Based Systems

Text Analytics and Visualisation

Text analytics is often used to analyse a large body of text.

Typical applications involve:
- Document classification and clustering
- Information retrieval from text
- Pattern matching and text mining
- Text summarisation
- Understanding of documents

There are many software tools that can assist text analytics, e.g.
tm in R, nltk in Python, SAS Viya, RapidMiner, Leximancer, etc.

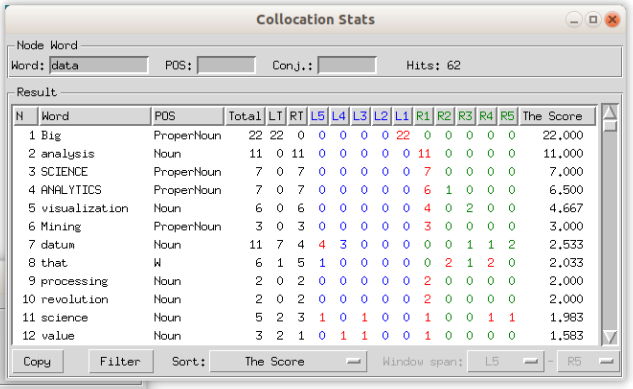
Our aim is to extract the key knowledge from the collected literature, and to create a *conceptual model of its domain of discourse*.

To this end article abstracts are *lexically analysed*, nouns and verbs are extracted and *lemmatised*. Terms of special importance are identified (*start words*) and “noise terms” removed (*stop words*).

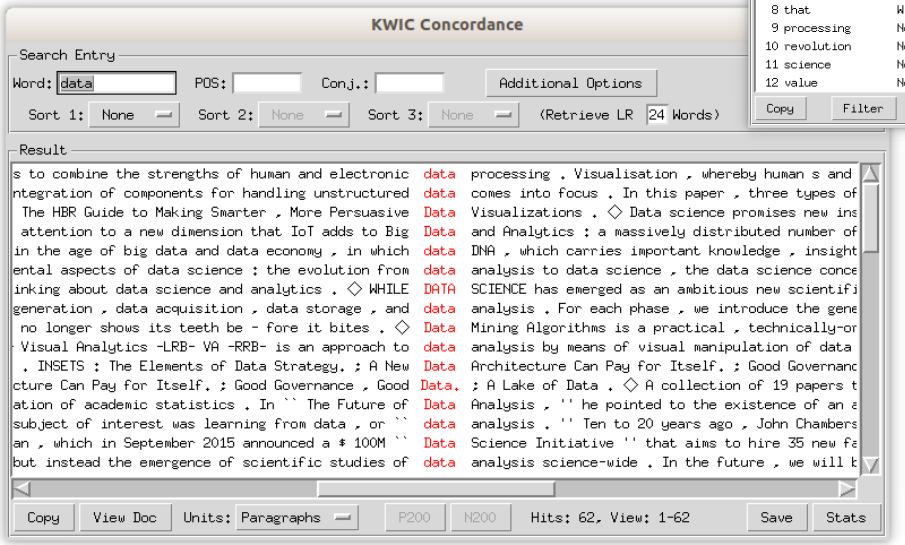
Text of abstracts can then be represented in a suitable form.

The representation helps identifying word-to-word, document-to-word, document-to-document relationships across abstracts.

Several text analysis techniques are then possible, e.g. *self-organising maps* and *multi-dimensional scaling* which both revealed *concepts proximity* and the emergence of *clusters*; *correspondence analysis* and *co-occurrence analysis* allowed finding *relationships between terms* and *identification of term communities*.



Collocation Index



KWIC Concordance

Various forms of document / text representation

We used *KH Coder*

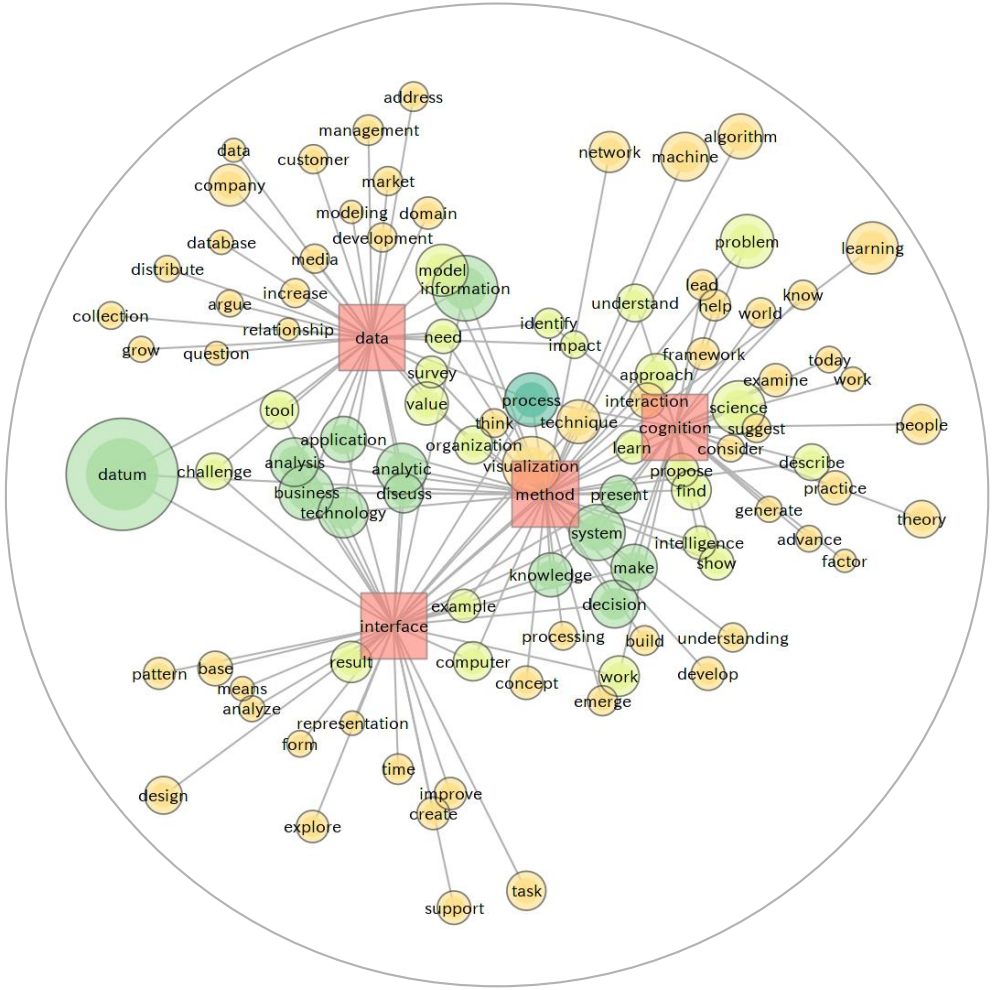
Used to conduct lexical analysis of abstracts
Deals with start and stop words
Represents text in documents-word matrix,
KWIC concordance, word association, etc.
Determines association between words
Analyses text using variety of techniques
Allows text visualisation using layout algorithms,
Such as Fruchterman and Reingold algorithm
Term communities and clusters can be identified
Minimum spanning trees can simplify relationships

docid	datum	information	science	system	visualization	process	business	analytic	model
1	4	0	2	0	0	12	3	1	0
2	0	0	0	0	0	0	0	0	0
3	2	1	1	1	0	0	0	1	0
4	0	0	0	0	0	0	0	0	0
5	1	0	1	0	0	0	0	0	0
6	0	1	1	0	0	0	0	0	0
7	0	1	0	0	2	0	0	1	1
8	1	7	0	0	6	0	3	0	0
9	0	0	0	0	0	3	0	0	9
10	0	0	0	0	0	0	0	0	0

Document-Word Frequency Matrix

Stage 1B: Acquisition

Exploratory Text Analytics



Preliminary insights into the domain of discourse related to Data Science in Organisations were obtained from *text analytics* using *co-occurrence analysis*.

Term co-occurrences in a text unit (abstract) define their association or similarity, which can be measured using the *Jaccard coefficient of similarity*. The larger the coefficient between two terms, the higher association.

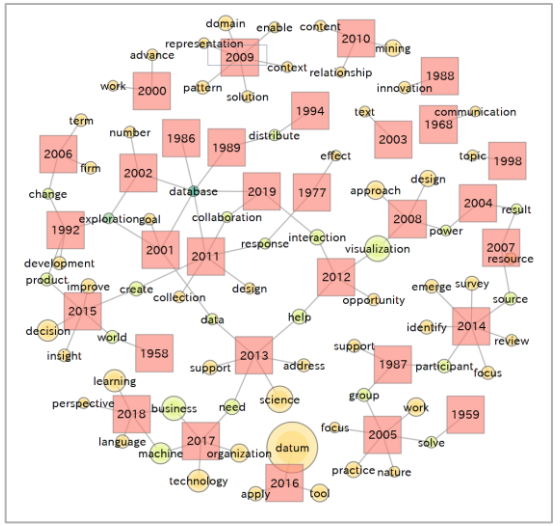
The associated terms form a *co-occurrence network*.

Such a network usually consists of many sub-networks, called *communities* (see next slide), which represent groups of strongly associated terms, which can be interpreted in aggregate as a *higher-level concept*.

The co-occurrence network is a 3D structure, where terms distance represents strength of association (short=high). This can be plotted using a *layout algorithm*, such as the Fruchterman and Reingold algorithm, so that the relative spatial proximity in 3D is preserved in 2D (when possible).

It is also possible investigating association between text terms and *additional variables*, such as data science domain (left) or the year of publication (right).

To test researchers' intuition regarding Data Science domains we manually coded 25% of abstracts (supervised process) – the codes were later discarded.

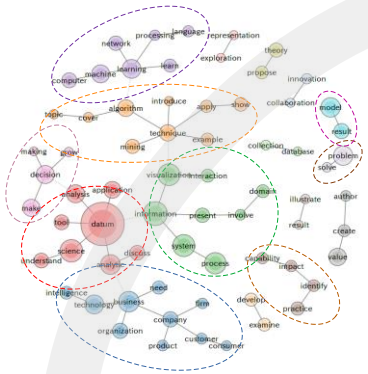


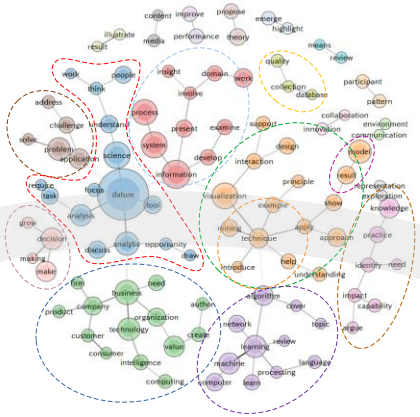
Note that the process of exploratory analysis assisted in refining search keywords, stop-list and start-list.

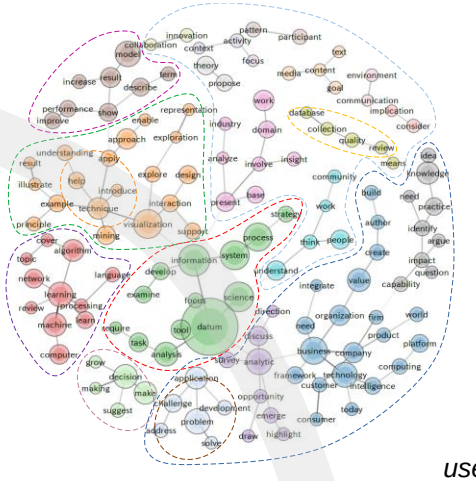
Stop words			Start Words
allow	give	role	cognition
area	have	see	method
article	include	set	interface
author	integrate	study	hardware
be	issue	take	software
become	level	thing	experience
book	literature	type	human
call	manager	use	
case	paper	user	
do	provide	variety	
exist	research	way	
field	researcher	year	

Stage 2A: Analysis
Unsupervised Text Analytics

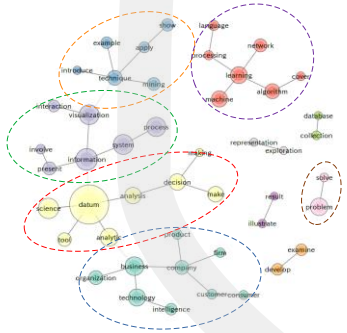
Context, Dialectics
Suspension of Judgement
(passive observation by researchers)



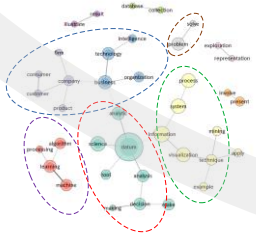




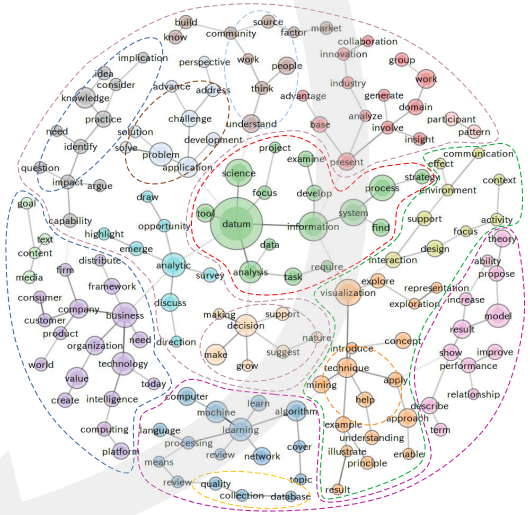
Fusion of Horizons
shared learning by co-researchers:
Invisible College
Co-Authors
Reviewers
Journal Editors
Machine (KH Coder)



Unsupervised
Text Analytics



Pre-understanding
(from exploratory analytics)



Saturation
(community cohesion + no fragmentation)

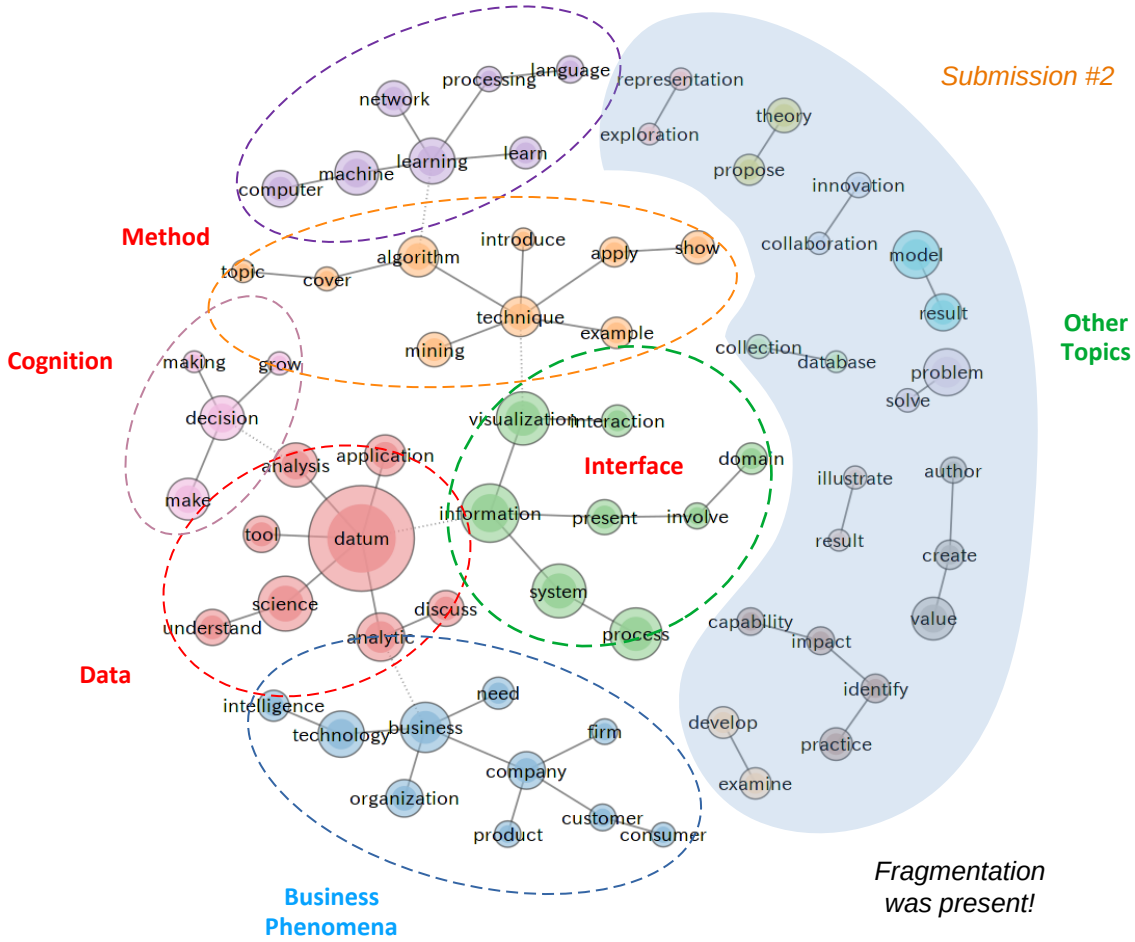
KH Coder

Evolved co-occurrence network over
40-**50**-**80**-100-**150**-200-**250**-300-400-**450** links
Layout algorithms broke at 500 links
Final network consisted of 450 links
Strongly associated nodes formed communities depicted with colour

Tracking evolution of communities was a manual process, which was passive as it did not interfere with any analytic tasks

Hyperparameters
used to control the amount of information used in the analysis and visualisation
(the number of high-frequency links)

80 Links

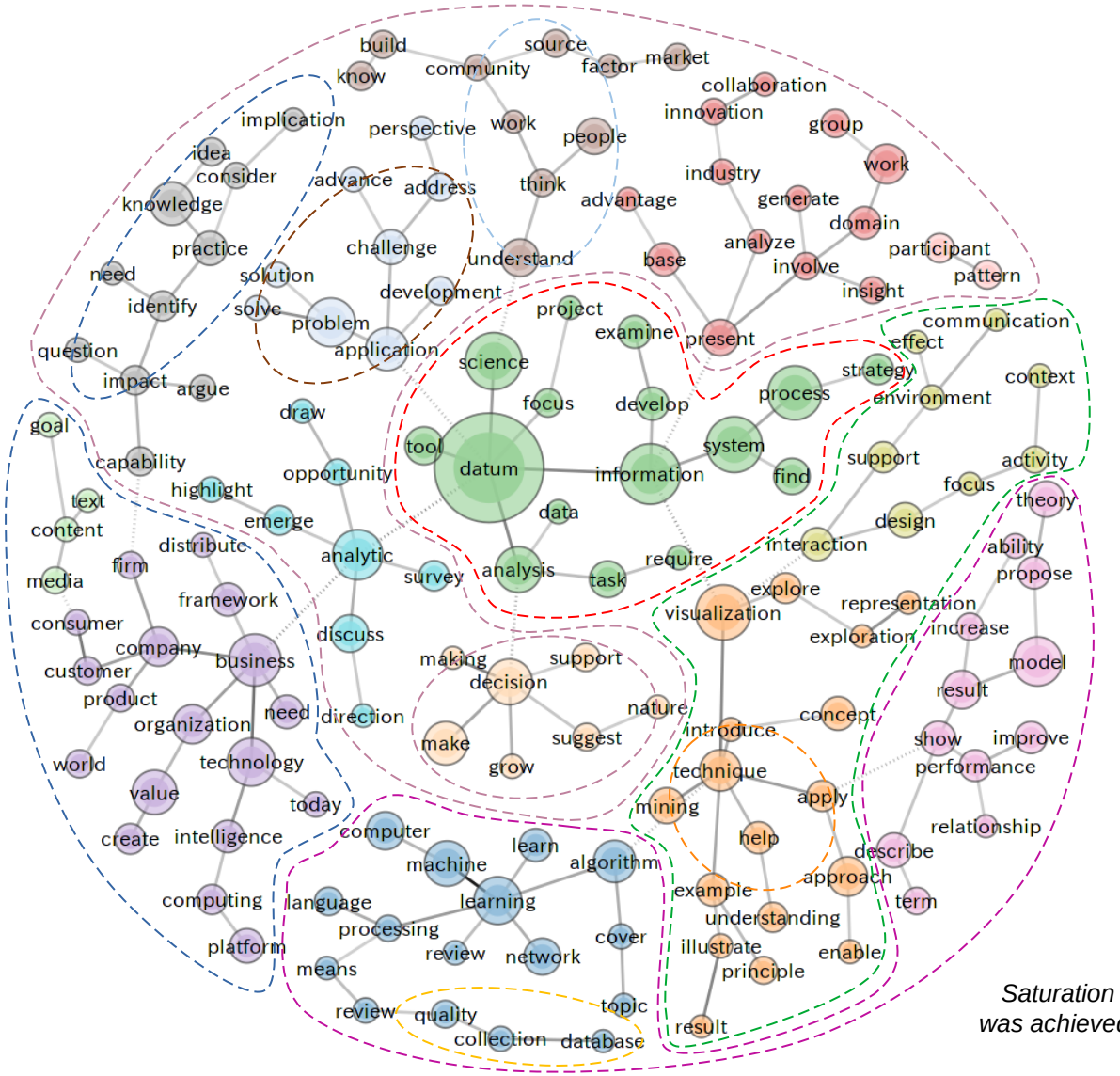


Our pre-conceptions were evident

Fragmentation was present!

80 Links

450 Links



Saturation was achieved!

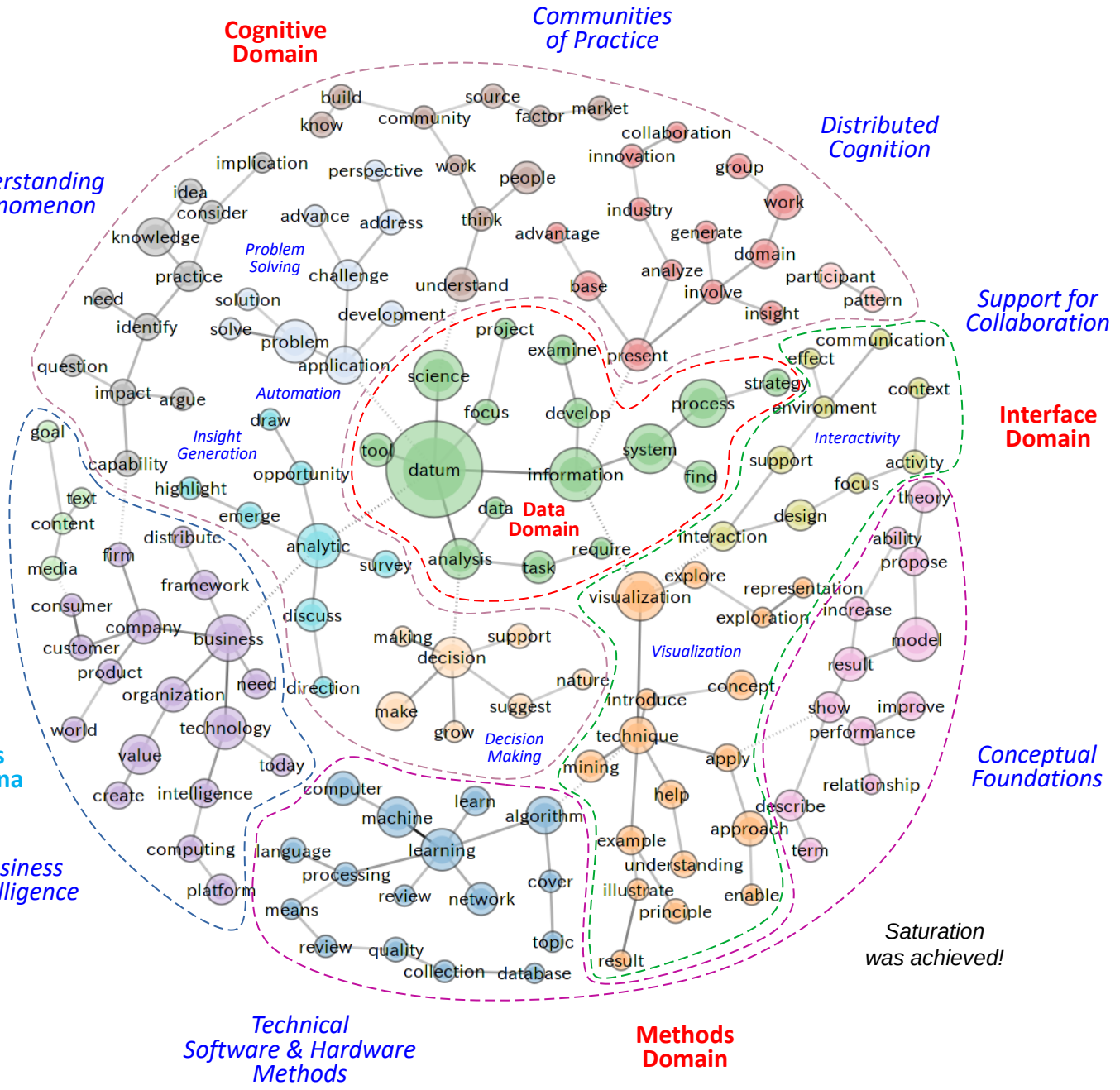
450 Links

Stage 2B-C: Analysis

Labelling & Partitioning

Subjective Labelling & Partitioning

Manual / Machine-Aided Content Analysis



Saturation was achieved!

Submission #3

Network

By subjective labelling and partitioning of the network, which is guided by insights retrieved (machine-aided) from the literature, the network gains its final interpretation.

Until this point, subjectivity and judgement were suspended.

450 Links
(cleaned and reinterpreted)

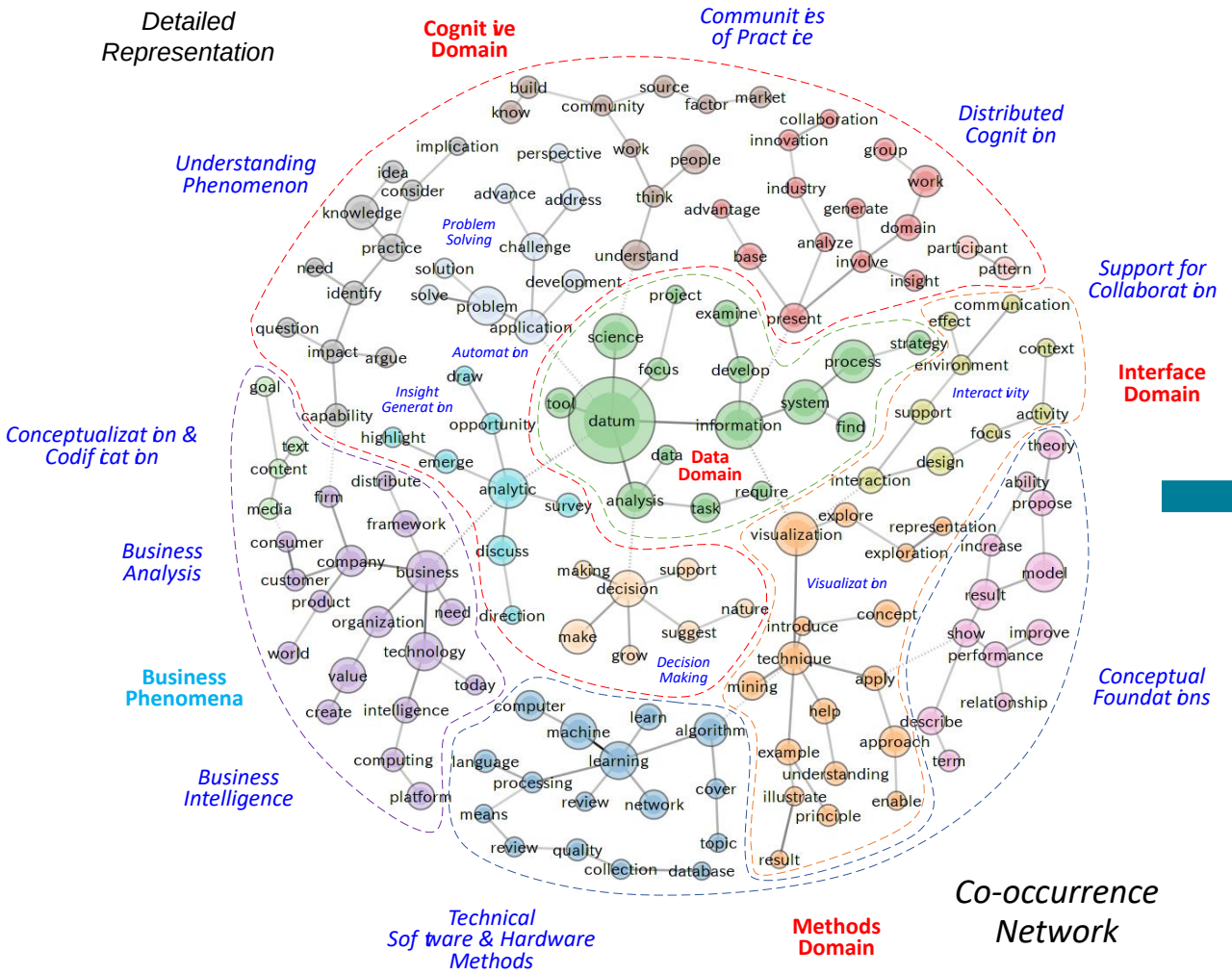
Stage 3A: Synthesis

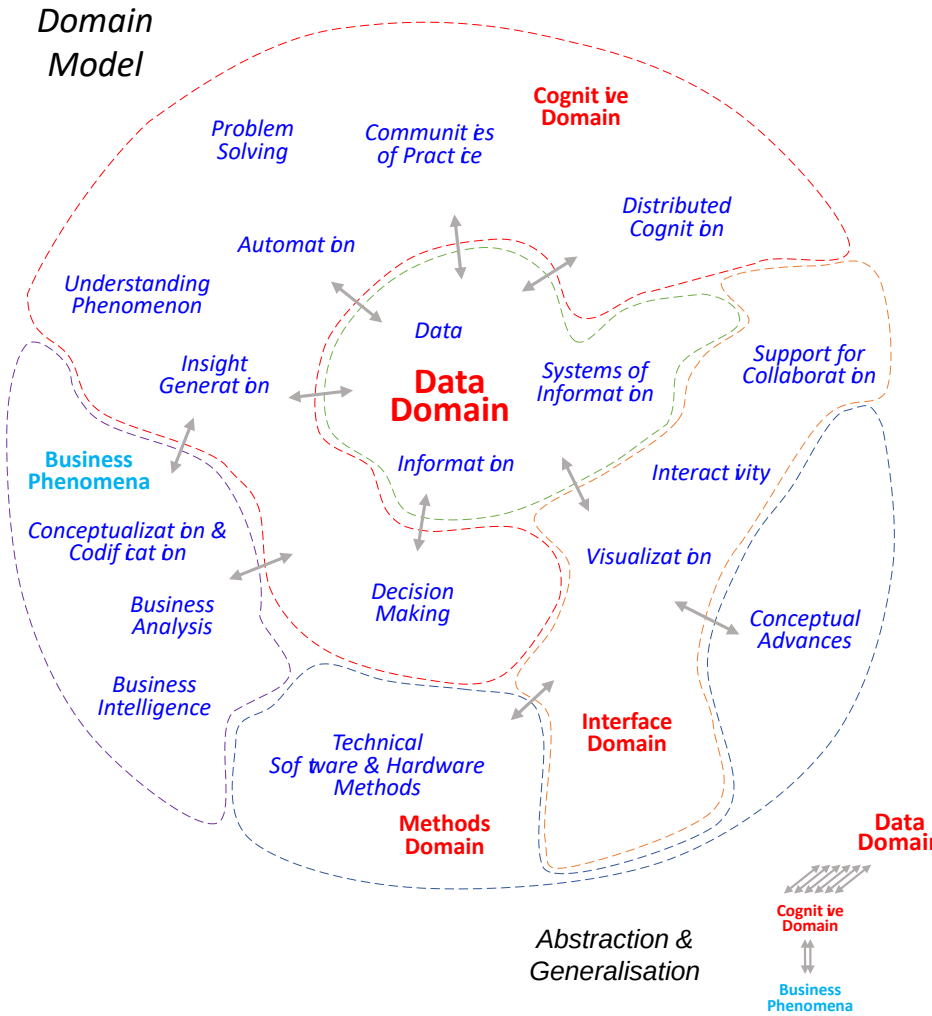
Abstraction and Generalisation

Network to Model

The partitioned and labelled co-occurrence network has now been simplified by removing all network nodes and links. Only major domain partitions, their relationships and their labels remained.

This created a general domain model, which provided structure for the manual analysis of literature content, which both were subsequently used to formulate the domain conceptual framework (including its essence).





Stage 3B-C: Synthesis

Essence & Conceptual Elaboration

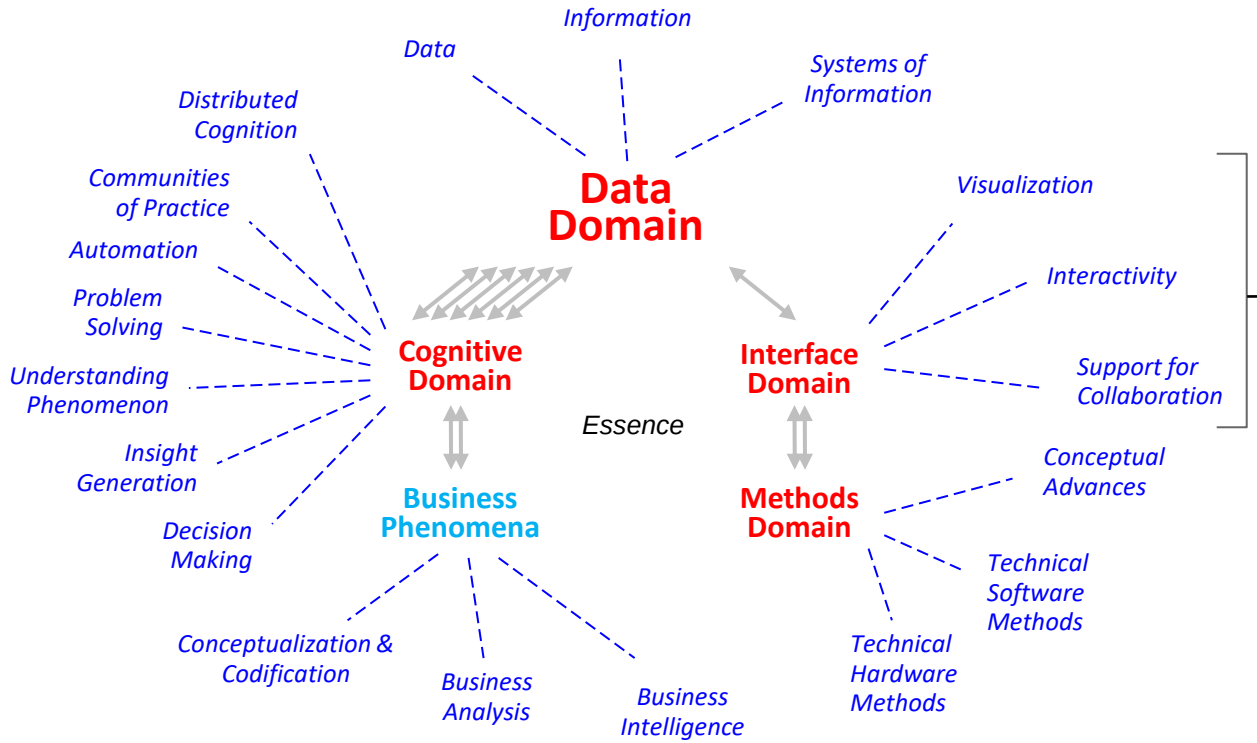


Table 8 - Key limitations and advances in *interfaces* pertinent to data science

Limitations:

- **Abstract, unstructured and complex business data hard to visualize**
  - Visualization of large and abstract business data is difficult (Dilla et al., 2010)
- **Lack of visual metaphors fit for business applications**
  - Difficulty selecting effective visual metaphor (Chen, Guo, et al., 2015)
  - Difficulty in visualizing business data (Al-Kassab et al., 2014; Edmunds and Morris, 2000)
  - Behavioral habits and cognitive biases lower effectiveness of visualization (Sacha et al., 2016)
- **Difficulty to represent some business phenomena with the available visual methods**
  - Using visualization requires special skills and training (Berinato, 2016; Wall et al., 2018)
  - Difficulties collaborating via visual forms (Martinez-Maldonado et al., 2019)

Advances in interfaces pertinent to data science:

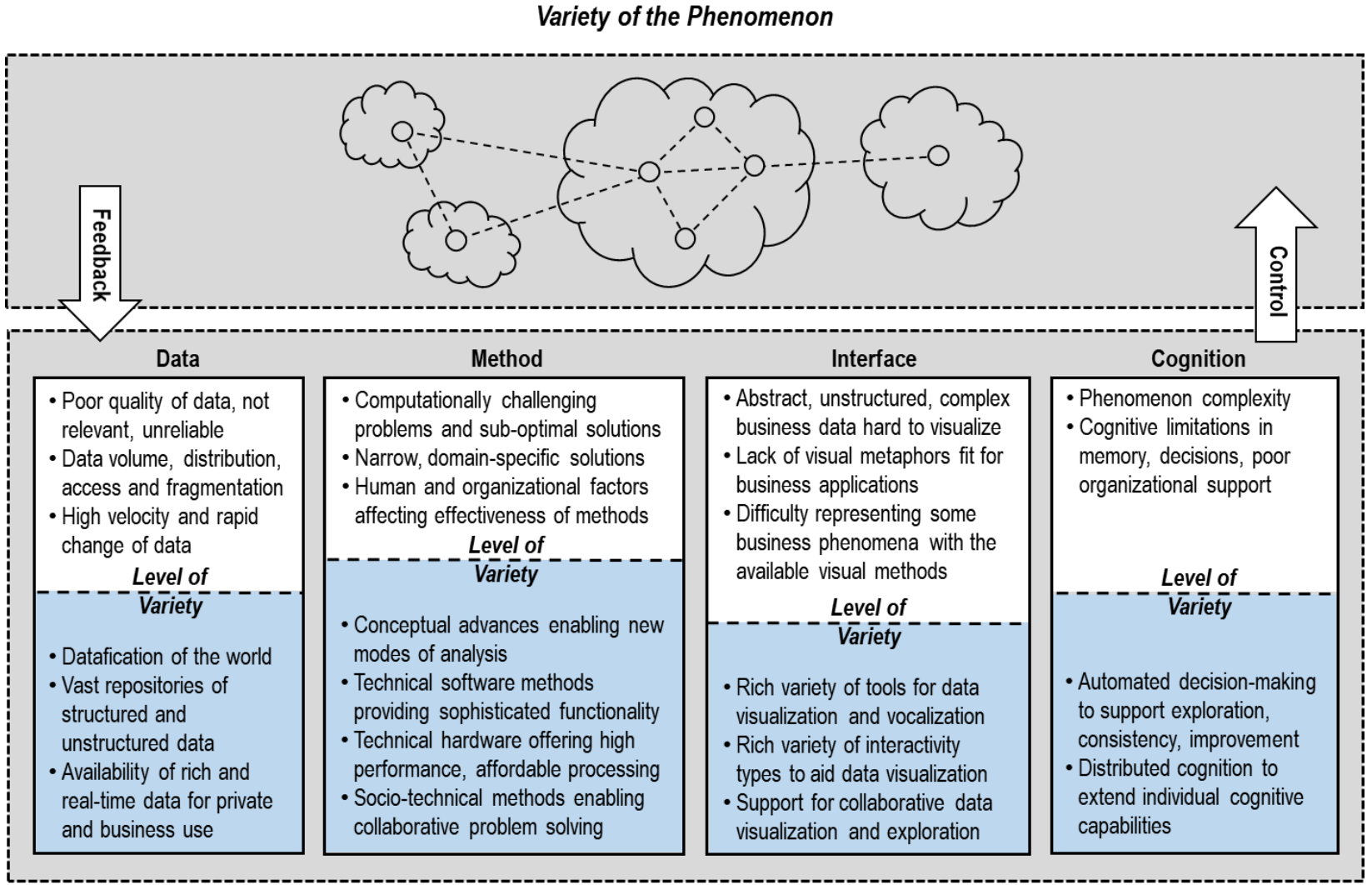
- **Rich variety of tools for data visualization and vocalization**
  - Data visualization amplifies cognition (Fekete et al., 2008; Keenan and Jankowski, 2019)
  - Visual exploration simplifies analysis of very large data sets (Keim, 2001, 2002)
  - Visual metaphors make data easier to understand (Cybulski et al., 2015)
  - Physical representations of data are now possible, e.g. 3D printing (D'Aveni, 2015)
- **Rich variety of interactivity types to aid data visualization**
  - Interactivity improves visualizations (Heer and Shneiderman, 2012)
  - Detecting emotions possible (Hibbeln et al., 2017; Kratzwald et al., 2018; Kurzhals et al., 2015)
  - Augmented reality for businesses (Olshannikova et al., 2015; Porter and Heppelmann, 2017)
  - Hybrid human-machine systems (Buxbaum-Conradi et al., 2016)
- **Support for collaborative data visualization and exploration**
  - Collaboration via interactive visual analytics (Isenberg et al., 2011)
  - Interactive visualization for distributed cognition (Liu et al., 2008)

Elaboration

Conceptual Framework

Stage 4: Application

Theoretical Framework



Reflection and Future Work

Novelty*:

- Hermeneutics + text analytics to conceptualise a large, complex body of knowledge
- The machine represents voice of the invisible college
- Unsupervised text analytics reduces researchers' pre-conceived biases in the process

Validity*:

- Demonstrated as method in action (JIT paper)
- Can it be replicated?

Performance Utility* (e.g. benchmarks)

- Dependent on specific text analytics method
- We used co-occurrence networks, hyper-parameterisation
- Saturation achieved when there is no longer fragmentation (at 450 links)

Application* (blind spots, caveats, how would scholars use this, research opportunities for use, further method development):

- Depends on the size and complexity of the body of knowledge
- Blind spots revealed as part of scholarly discourse (need engaged reviewers)
- Caveats, limitations (abstracts, requires pre-understanding of domain)

*Arun Rai's criteria for methodological papers…