vantage point®
Search Technology

Alan Porter & Nils Newman | June 23, 2022

What Knowledge to Extract from “Tech Mining”?

Empowering analysts for over 20 years

TheVantagePoint.com

Alan Porter

- "Interdisciplinary" background
  - B.S. in Chemical Engineering (Caltech)
  - PhD in Engineering/Psychology (UCLA)
- Research focus: Analysis of Emerging Technologies
- Georgia Tech – Co-director, Program in Science, Technology & Innovation Policy (STIP) &
  Professor Emeritus, ISyE & PubP
- Director, R&D, Search Technology
  - Develop & apply text mining software (VantagePoint; Derwent Data Analytics)

Nils Newman

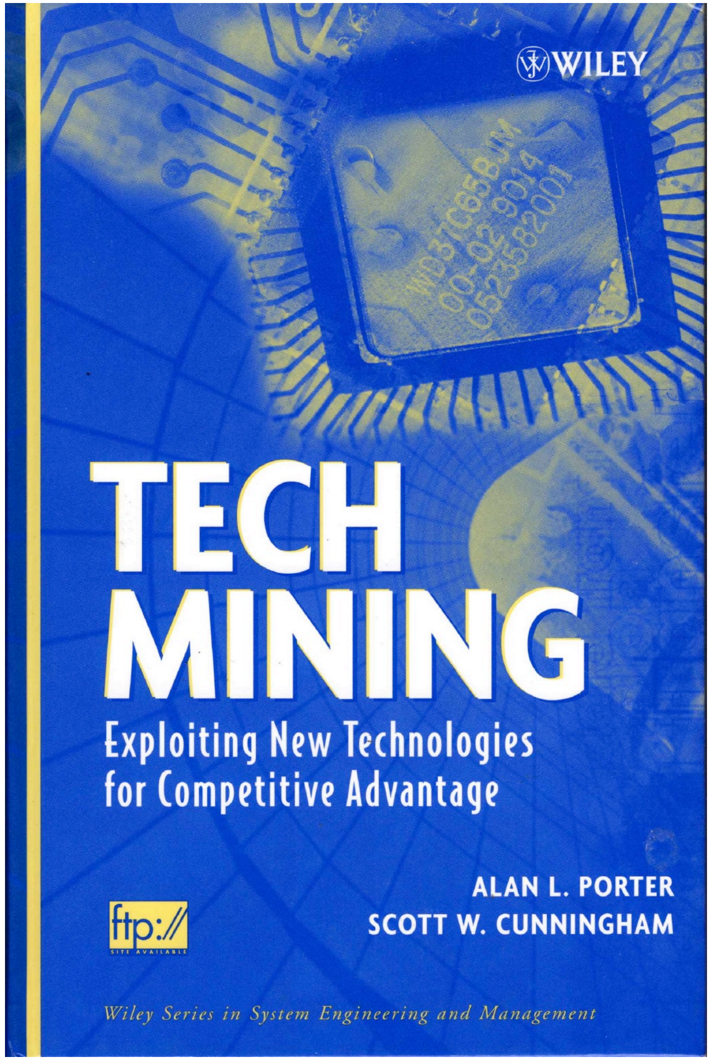
- “Interdisciplinary” background
  - B.S. in Mechanical Engineering (Georgia Tech)
  - M.S. in Technology and Science Policy (Georgia Tech)
- Research Focus: Creating analytical tools for the management of technology
- President, Search Technology

Agenda

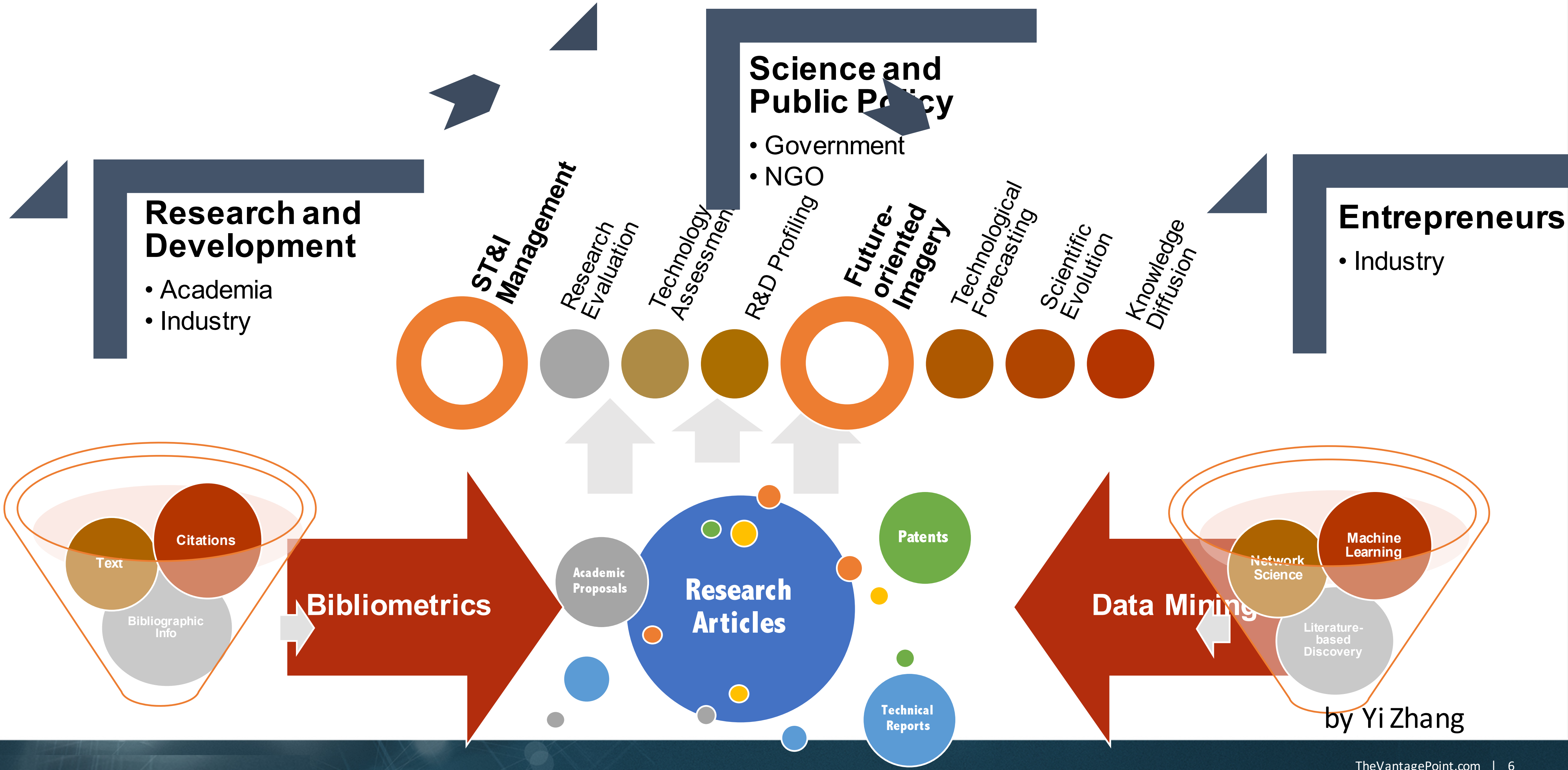
- Overview of Tech Mining
- Some Samples:
  - Multi-Generational Citation Analysis with Indications
  - Tech Emergence Scores
- Comparison of Tech Mining to EEKE
- Lessons?

How to find the “missing” R&D knowledge to enable innovation?

One approach to find that “missing” science is to use the “Tech Mining” process to generate R&D intelligence systematically



The Scope of Tech Mining

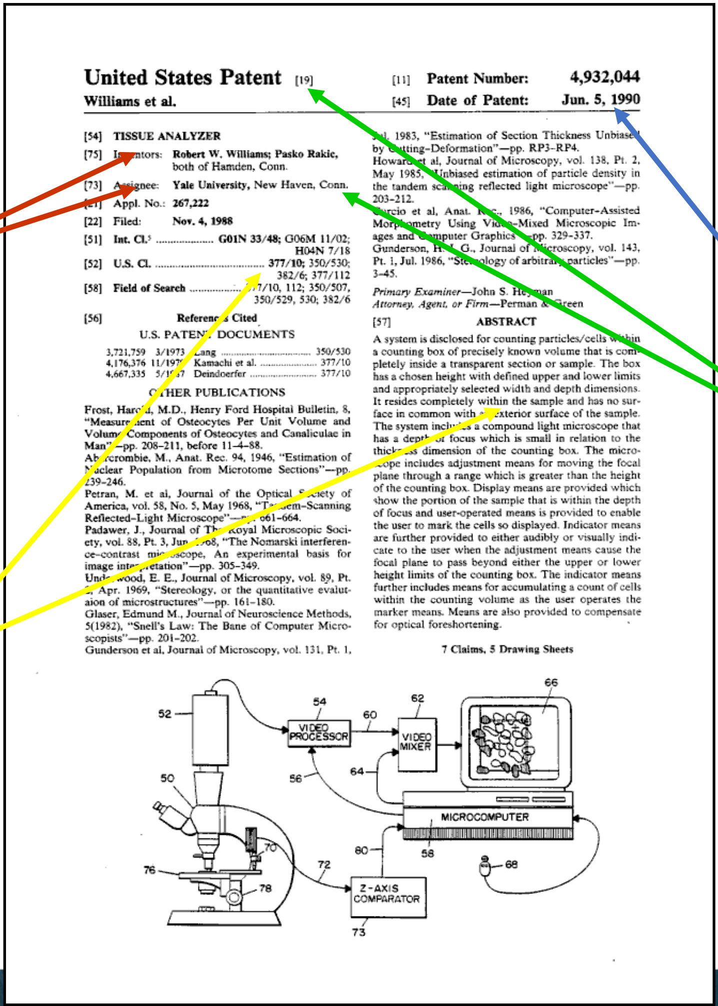


by Yi Zhang

What are we trying to find?
Entities that can answer questions.

Tech Mining Questions to Answer
Primarily from field-structured data

Who?



United States Patent [19]
Williams et al.

[11] Patent Number: 4,932,044
[45] Date of Patent: Jun. 5, 1990

[54] TISSUE ANALYZER
[75] Inventors: Robert W. Williams; Pasko Rakic, both of Hamden, Conn.
[73] Assignee: Yale University, New Haven, Conn.
[21] Appl. No.: 267,222
[22] Filed: Nov. 4, 1988
[51] Int. Cl.[5] G01N 33/48; G06M 11/02; H04N 7/18
[52] U.S. Cl. 377/10; 350/530; 382/6; 377/112
[58] Field of Search 377/10, 112; 350/507, 350/529, 530; 382/6
[56] References Cited

U.S. PATENT DOCUMENTS

3,721,759	3/1973	Lang	350/530
4,176,376	11/1979	Kamachi et al.	377/10
4,667,335	5/1987	Deindoerfer	377/10

OTHER PUBLICATIONS

Frost, Harold, M.D., Henry Ford Hospital Bulletin, 8, "Measurement of Osteocytes Per Unit Volume and Volume Components of Osteocytes and Canaliculae in Man"—pp. 208–211, before 11–4–88.
Abercrombie, M., Anat. Rec. 94, 1946, "Estimation of Nuclear Population from Microtome Sections"—pp. 239–246.
Petran, M. et al, Journal of the Optical Society of America, vol. 58, No. 5, May 1968, "Tandem-Scanning Reflected-Light Microscope"—pp. 661–664.
Padawer, J., Journal of The Royal Microscopic Society, vol. 88, Pt. 3, Jun. 1968, "The Nomarski interference-contrast microscope, An experimental basis for image interpretation"—pp. 305–349.
Underwood, E. E., Journal of Microscopy, vol. 89, Pt. 2, Apr. 1969, "Stereology, or the quantitative evaluation of microstructures"—pp. 161–180.
Glaser, Edmund M., Journal of Neuroscience Methods, 5(1982), "Snell's Law: The Bane of Computer Microscopists"—pp. 201–202.
Gunderson et al, Journal of Microscopy, vol. 131, Pt. 1, Jul. 1983, "Estimation of Section Thickness Unbiased by Cutting-Deformation"—pp. RP3–RP4.
Howard et al, Journal of Microscopy, vol. 138, Pt. 2, May 1985, "Unbiased estimation of particle density in the tandem scanning reflected light microscope"—pp. 203–212.
Curcio et al, Anat. Rec., 1986, "Computer-Assisted Morphometry Using Video-Mixed Microscopic Images and Computer Graphics"—pp. 329–337.
Gunderson, H. J. G., Journal of Microscopy, vol. 143, Pt. 1, Jul. 1986, "Stereology of arbitrary particles"—pp. 3–45.

Primary Examiner—John S. Heyman
Attorney, Agent, or Firm—Perman & Green

[57] ABSTRACT

A system is disclosed for counting particles/cells within a counting box of precisely known volume that is completely inside a transparent section or sample. The box has a chosen height with defined upper and lower limits and appropriately selected width and depth dimensions. It resides completely within the sample and has no surface in common with an exterior surface of the sample. The system includes a compound light microscope that has a depth of focus which is small in relation to the thickness dimension of the counting box. The microscope includes adjustment means for moving the focal plane through a range which is greater than the height of the counting box. Display means are provided which show the portion of the sample that is within the depth of focus and user-operated means is provided to enable the user to mark the cells so displayed. Indicator means are further provided to either audibly or visually indicate to the user when the adjustment means cause the focal plane to pass beyond either the upper or lower height limits of the counting box. The indicator means further includes means for accumulating a count of cells within the counting volume as the user operates the marker means. Means are also provided to compensate for optical foreshortening.

7 Claims, 5 Drawing Sheets

When?

Where?

What?

Field structured bibliographic data make some of the work easy and large data repositories are accessible.
But finding "What" is still a challenge.

How? & Why? –
Need human analyst or more advances in EEKE!

“How to”: Ten-Step Tech Mining “Framework”

1. Spell out your Science, Technology & Innovation (ST&I) questions and how to answer them
2. Get suitable data
3. Search (iterate) & retrieve ~abstract records
4. Import into text mining software
5. Refine (clean; consolidate) the data
6. Analyze
7. Visualize (Map)
8. Integrate with Internet analyses & expert opinion
9. Summarize; Interpret; Communicate (multi-dimensionally)!
10. Standardize and semi-automate where possible

Major Components of Tech Mining

Import

- Moving the text from the source to an analytical environment.
- Data can be in a variety of formats: Text, XML, etc…
- Scale is an issue – many text sources do not make it easy to move large numbers of records.

Refine

- Text analytics is very much "Garbage in: Garbage out".
- The most time consuming part of an analysis.

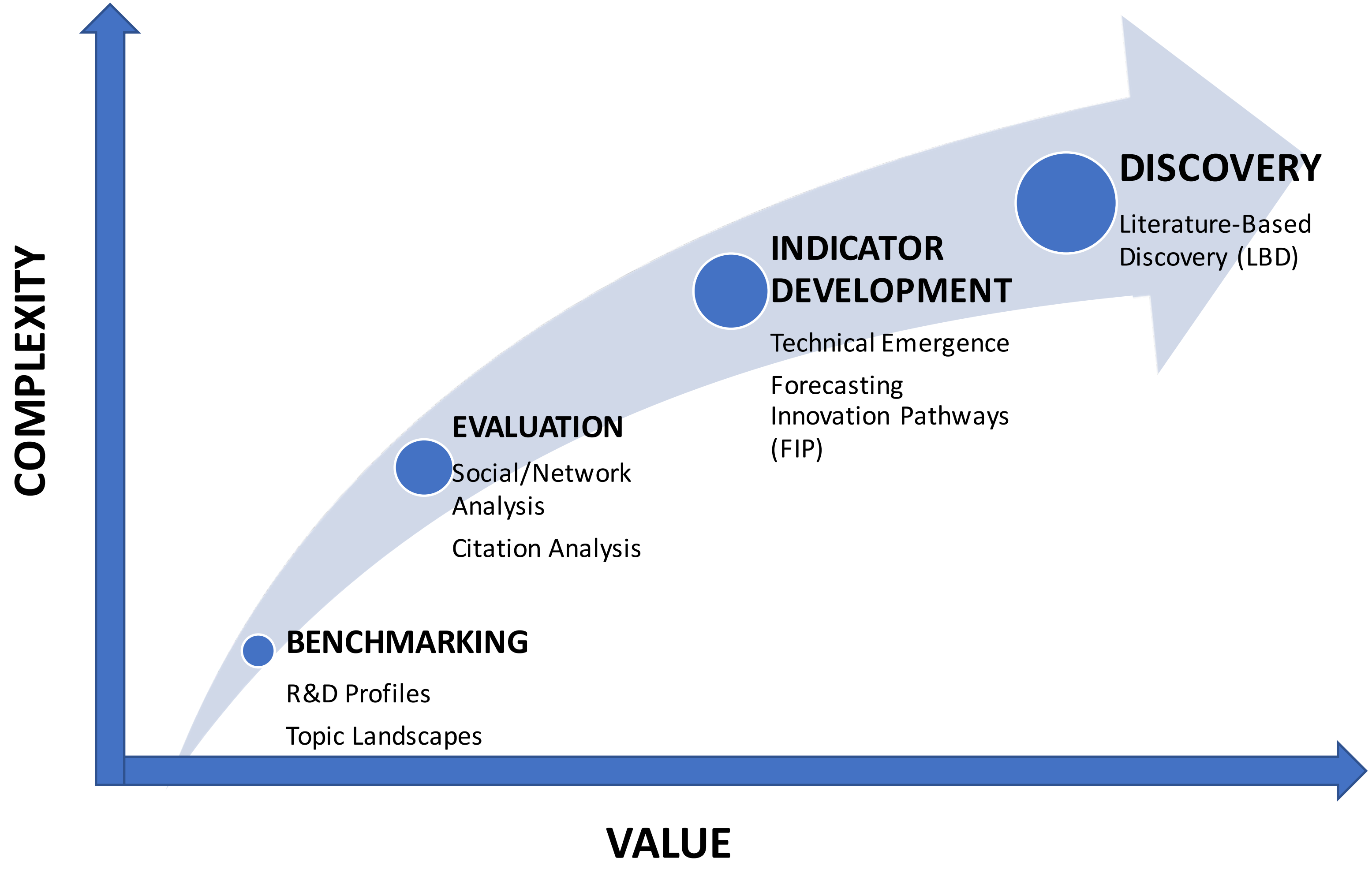
Analyze

- Patterns identification and extraction are key.
- Many approaches possible but ultimately co-occurrence is at the root of most techniques.

Report

- Text mining is an alternative to expert opinion in decision making, but it is relatively new.
- Getting decision makers to trust text analysis can be a challenge so effective communication is key.

Value Chain of Tech Mining (Science/Innovation Discourse Analysis)



Some Samples

Sample #1: Citation Analysis

Tracking multi-generational research knowledge transfer with:

- ***Interdisciplinarity metrics***
- ***Science overlay mapping***

#1: Papers Citing Level #2 Papers – Citing Paper Overlay Maps [Knowledge Diffusion]

- Diffusion scores
- Science Citing Overlay Maps
- Relative engagement by ISI Subject Categories

#2: Main Level (e.g., research outputs of a target program) – publication overlay maps

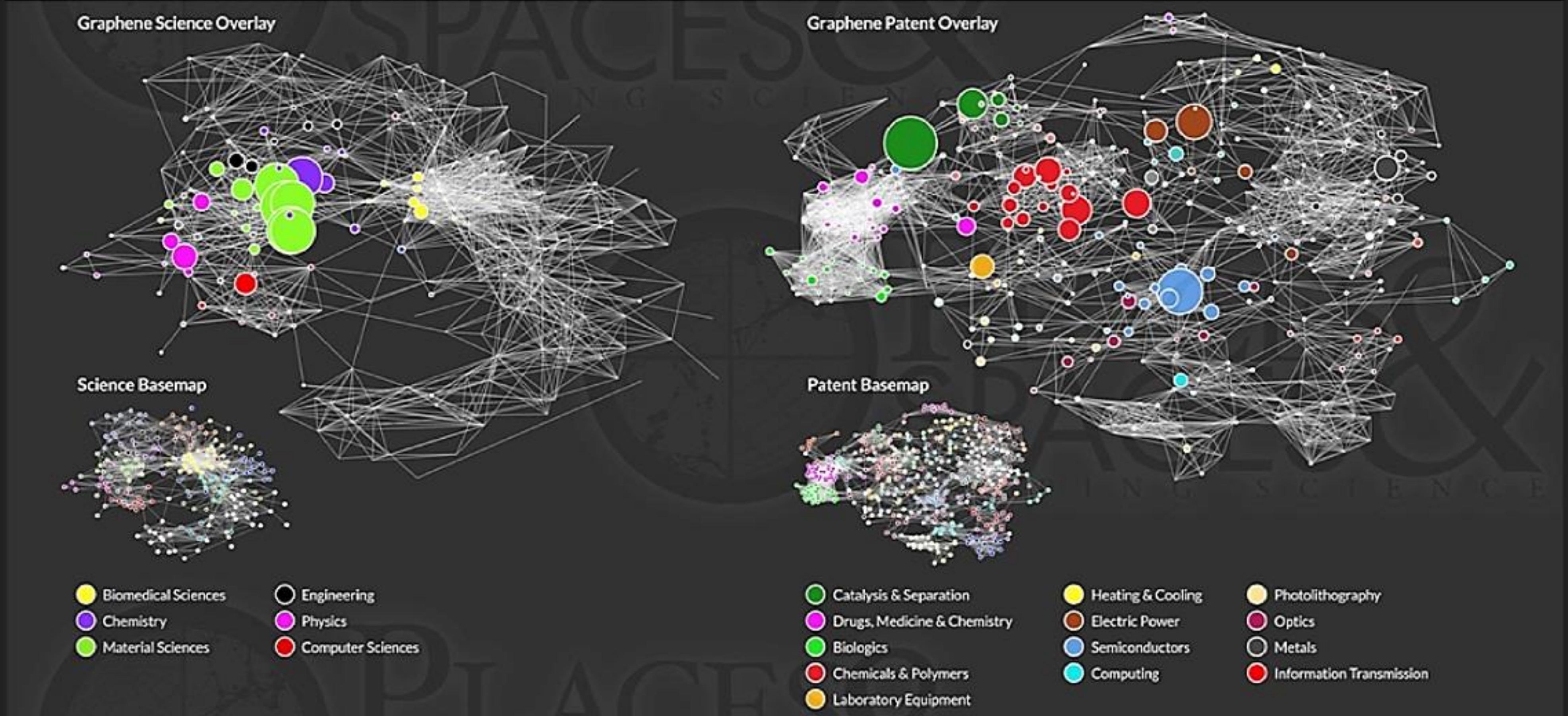
- "Specialization" scores (Diversity of areas of publication)
- Science overlay maps (Location of publications among ISI Subject Categories)

- Integration scores (Average diversity of areas of citation)
- Science citation maps
- Bibliographic coupling

#3: Papers cited by #2

- Coherence measures (do #3 papers draw upon distinct topics?)
  - ["Bibliographic Coupling" measures available – e.g., % shared references]

#4: Papers cited by #3



About the Basemaps

The science basemap was created using publication citation index data for 2009. It features 222 Web of Science category nodes that are grouped into 18 color-coded macro-science disciplines. The patent basemap was computed by mining patents for the time period 2000-2006. It consists of 466 IPC technology nodes grouped into 35 color-coded macro-patent categories. In each network map, edges are drawn between nodes that have a threshold above the median similarity value. The Kamada & Kawai layout algorithm in Pajek was used to layout the networks in a two-dimensional space—the closer two nodes are the higher the similarity between them. The basemaps show the structure of science and technology landscape respectively and serve as a reference system for data overlays.

How to Read the Overlay Maps

The data overlays show the number and placement of publications and patents that match "graphene" on the respective basemap. Node size indicates the number of matching publications or patents per node. Labels and colors of the six macro-science disciplines and the 13 macro-patent categories that contain matching publications and patents are given below the network maps. There exists no relation between the colors in the two maps.

Acknowledgments: This work is based on research undertaken largely at Georgia Tech drawing on support from the U.S. National Science Foundation (NSF) through the Center for Nanotechnology in Society, Arizona State University (Award No. 0531194) and the NSF project "Revealing Innovation Pathways: Hybrid Science Maps for Technology Assessment and Foresight" (Award No. 1064146). Part of this research was also undertaken in collaboration with SPRU, University of Sussex, and the Center for Nanotechnology in Society, University of California Santa Barbara (NSF Awards No. 0938099 and No. 0531184). The observations contained in this work are those of the authors and do not necessarily reflect the views of NSF.

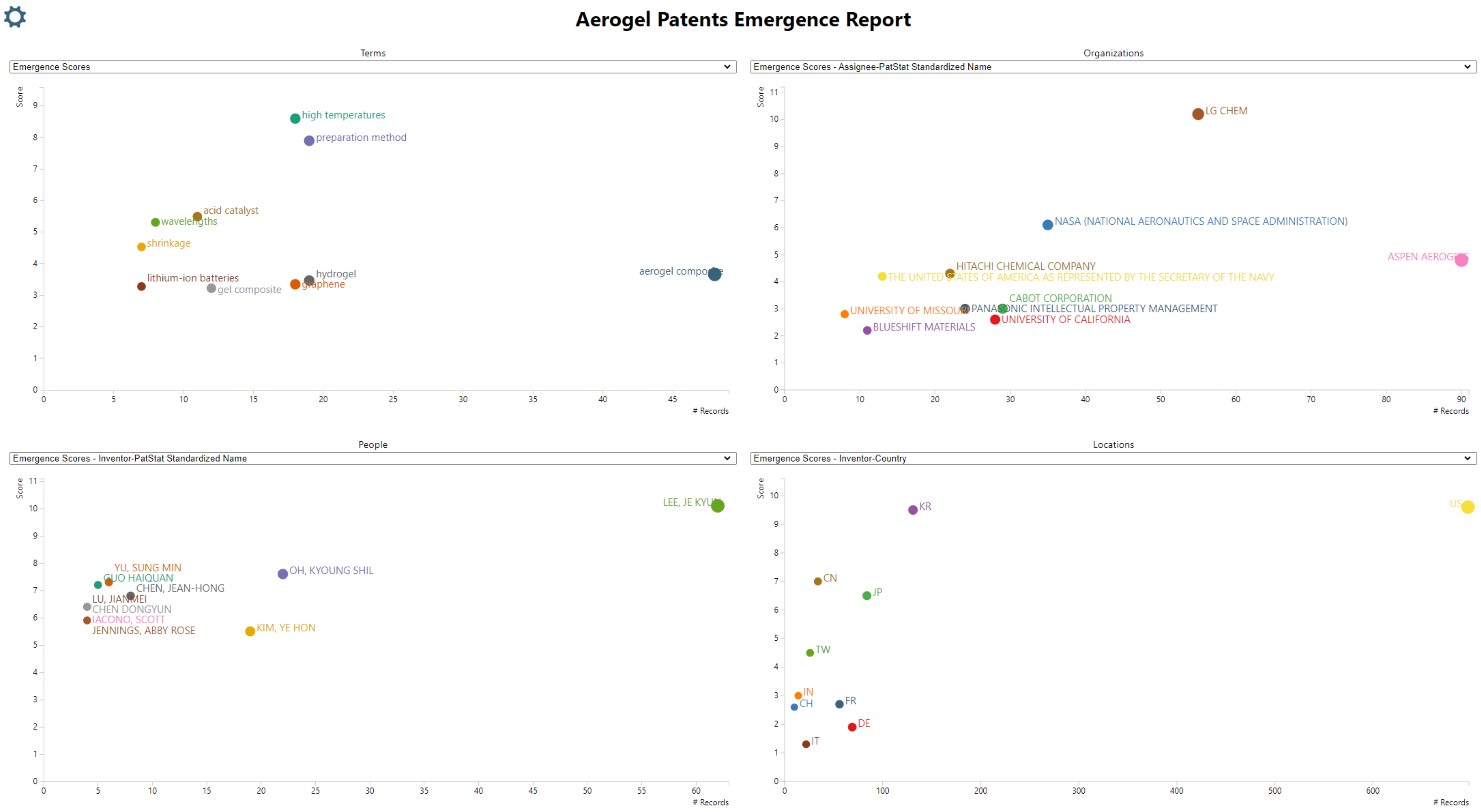
X.8
Mapping Graphene Science and Development: Focused Research with Multiple Application Areas, by Luciano Kay, Alan L. Porter, Ismael Rafols, Nils Newman, and Jan L. Youtie

SANTA BARBARA, CA, ATLANTA, GA & VALÈNCIA, SPAIN, 2014
Courtesy of Center for Nanotechnology in Society (UCSB), Georgia Tech, and Intelligent Information Services Corporation, Ingenio (CSIC-UPV)

Sample #2: How We Generate Tech Emergence Indicators

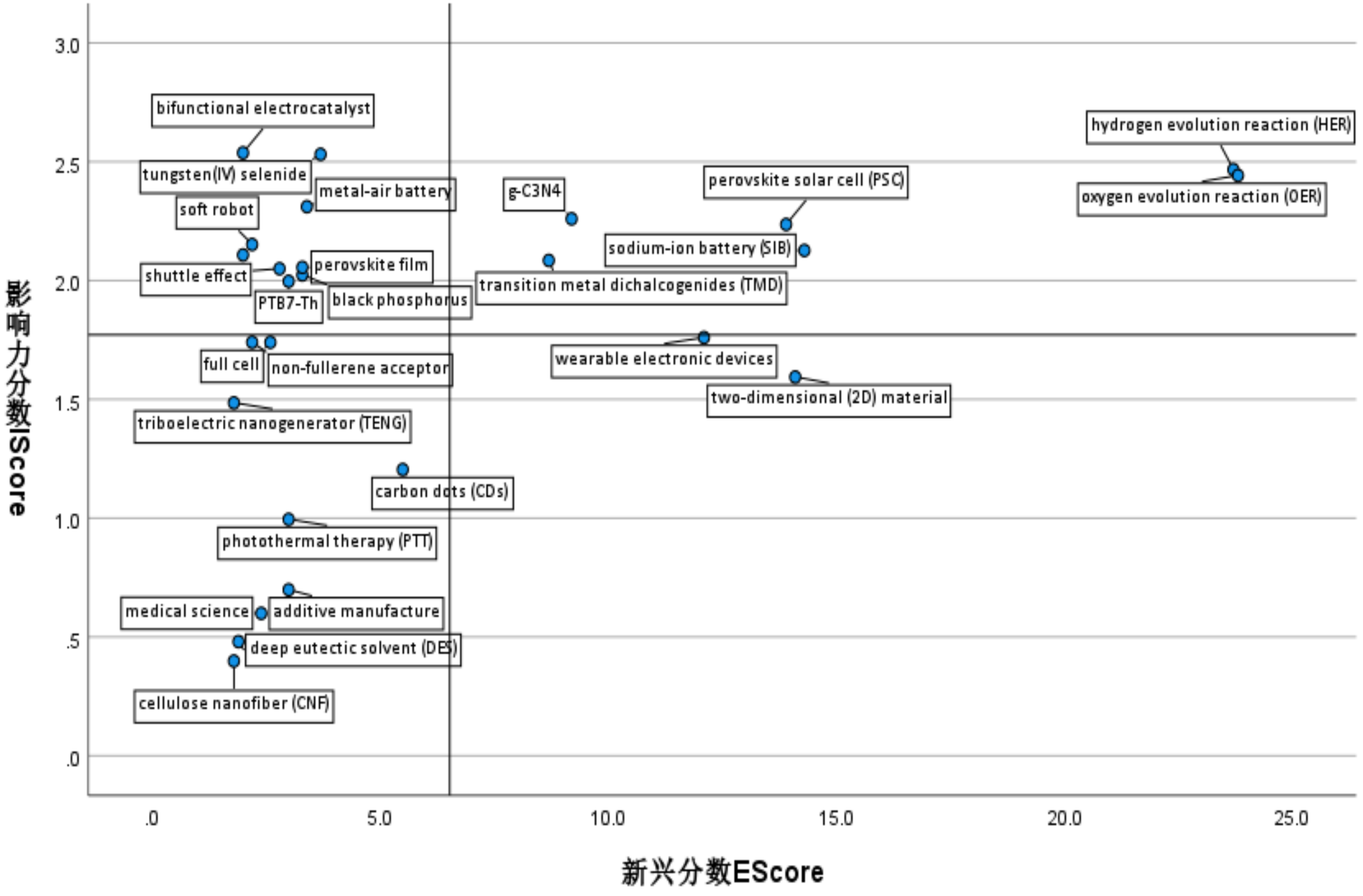
1. Specify the Science or Technology domain & suitable database source (e.g., Web of Science)
2. Search & retrieve abstract records
3. Select topical fields (e.g., Title & Abstract Natural Language Processing phrases)
4. Refine terms
5. Apply thresholds
6. Generate emergence scores (**EScores**) for terms
7. Generate EScores for players

Emergence Output: Aerogel Patent Sample



Building on Emergence:

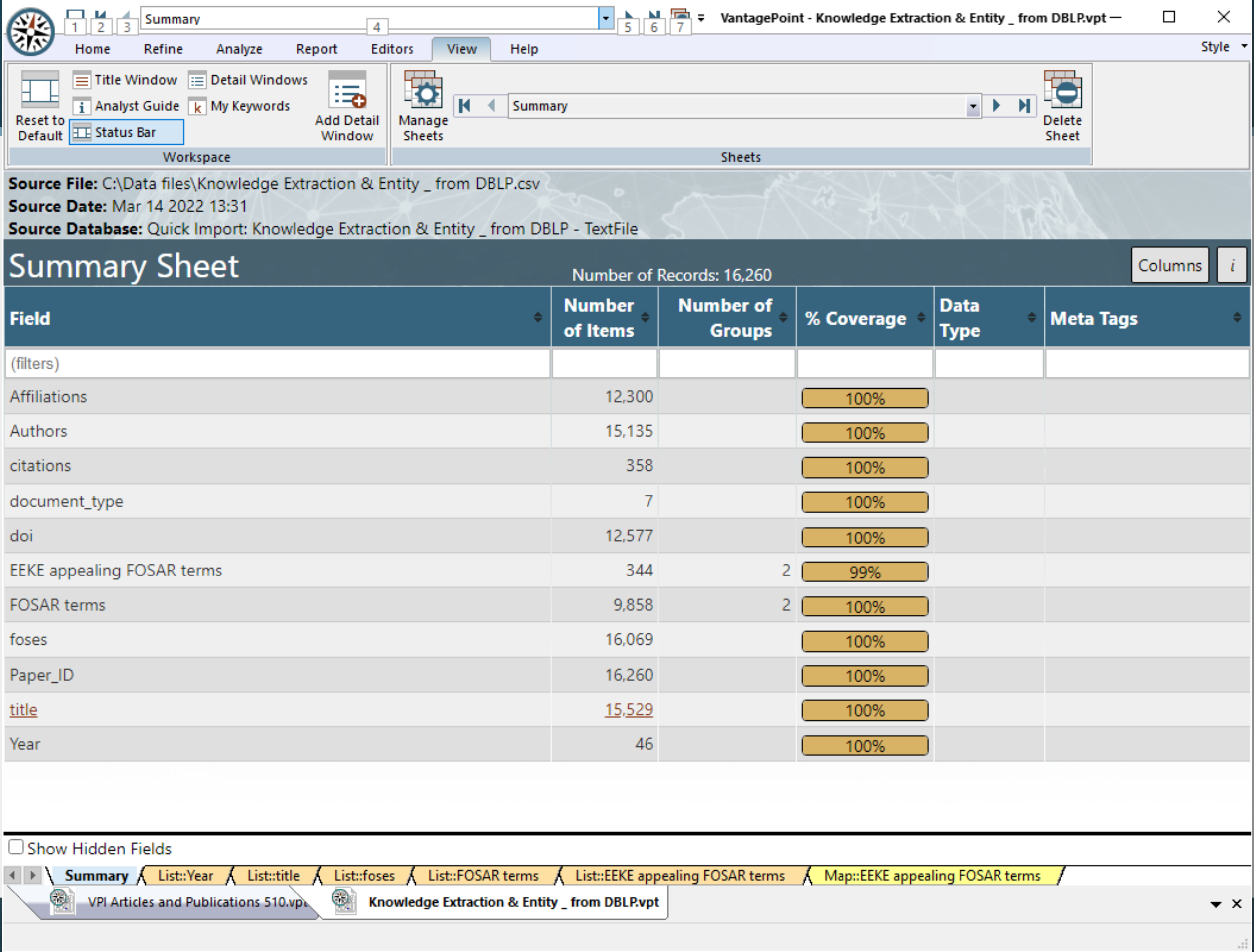
Does research addressing emerging topics within a domain have greater scientific impact? ***YES → for Nanotechnology (Wang, 2021)***



Sample 3: Tech Mining – Applied to itself and EEKE

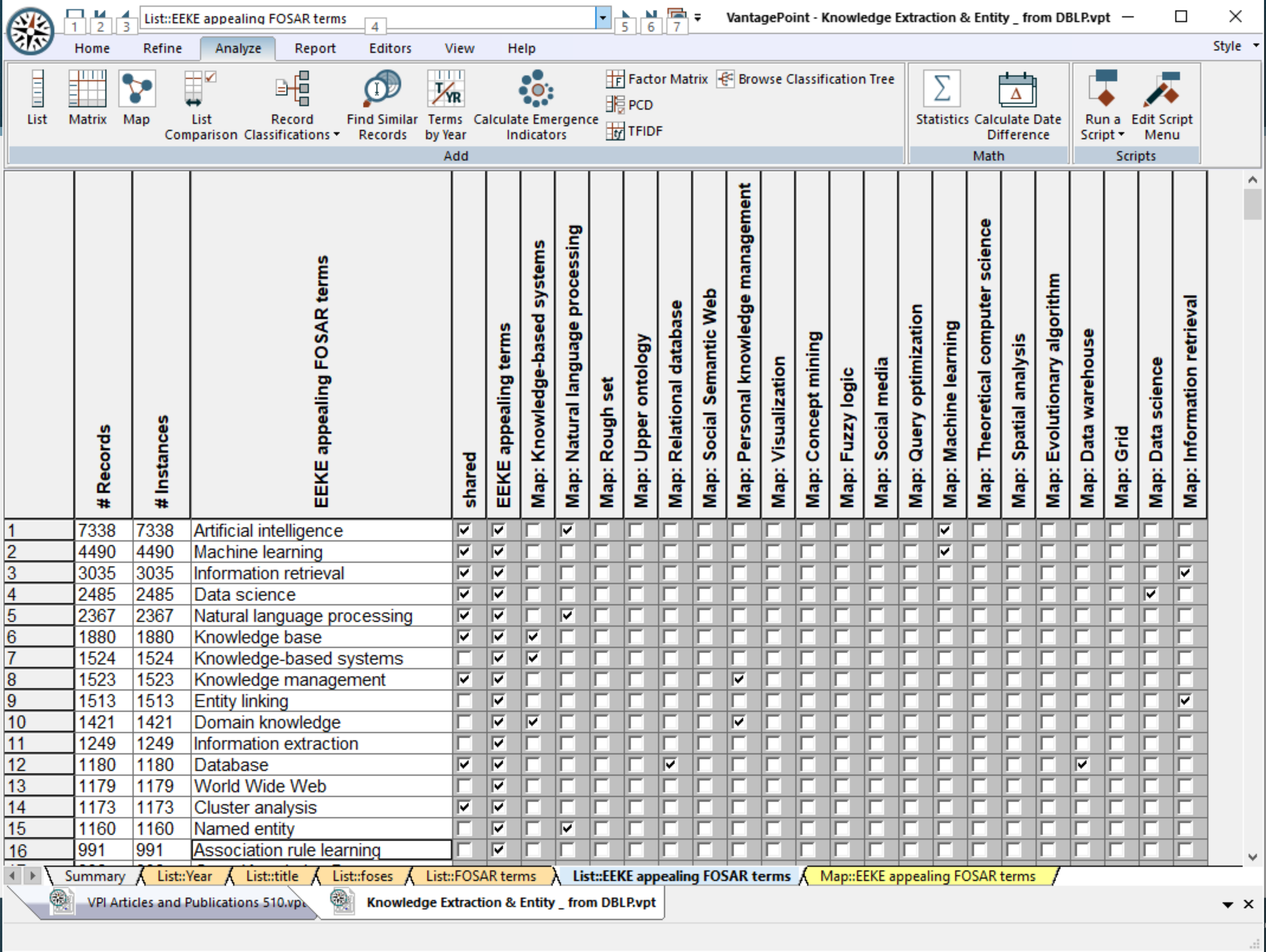
- Analyze 2 datasets
  - 510 abstract records related to Tech Mining and/or *VantagePoint* software from the VP Institute site [https://vpinstitute.org] – articles from researchers around the world who use our technology
  - 16,260 abstract records on "Knowledge Extraction & Entity" from DBLP database
- Compare emphases

VantagePoint Summary Page



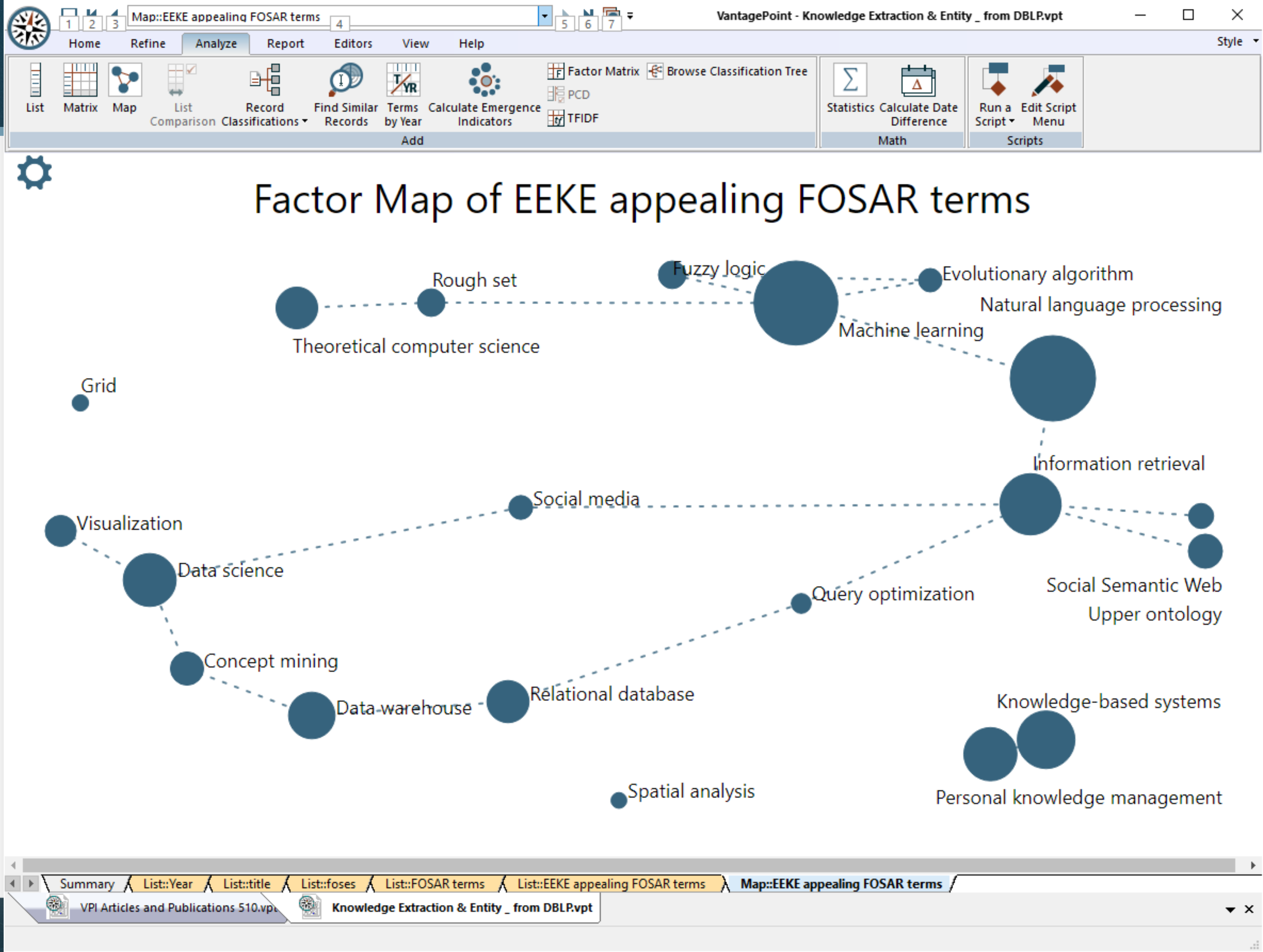
Field	Number of Items	Number of Groups	% Coverage	Data Type	Meta Tags
(filters)					
Affiliations	12,300		100%		
Authors	15,135		100%		
citations	358		100%		
document_type	7		100%		
doi	12,577		100%		
EEKE appealing FOSAR terms	344	2	99%		
FOSAR terms	9,858	2	100%		
foses	16,069		100%		
Paper_ID	16,260		100%		
title	15,529		100%		
Year	46		100%		

List, with Groups



	# Records	# Instances	EEKE appealing FOSAR terms	shared	EEKE appealing terms	Map: Knowledge-based systems	Map: Natural language processing	Map: Rough set	Map: Upper ontology	Map: Relational database	Map: Social Semantic Web	Map: Personal knowledge management	Map: Visualization	Map: Concept mining	Map: Fuzzy logic	Map: Social media	Map: Query optimization	Map: Machine learning	Map: Theoretical computer science	Map: Spatial analysis	Map: Evolutionary algorithm	Map: Data warehouse	Map: Grid	Map: Data science	Map: Information retrieval
1	7338	7338	Artificial intelligence	✓	✓		✓											✓							
2	4490	4490	Machine learning	✓	✓													✓							
3	3035	3035	Information retrieval	✓	✓																				✓
4	2485	2485	Data science	✓	✓																			✓	
5	2367	2367	Natural language processing	✓	✓		✓																		
6	1880	1880	Knowledge base	✓	✓	✓																			
7	1524	1524	Knowledge-based systems		✓	✓																			
8	1523	1523	Knowledge management	✓	✓							✓													
9	1513	1513	Entity linking		✓																				✓
10	1421	1421	Domain knowledge		✓	✓						✓													
11	1249	1249	Information extraction		✓																				
12	1180	1180	Database	✓	✓					✓												✓			
13	1179	1179	World Wide Web		✓																				
14	1173	1173	Cluster analysis	✓	✓																				
15	1160	1160	Named entity		✓		✓																		
16	991	991	Association rule learning		✓																				

PCA Factor Map



EEKE Factors – Commonality with Tech Mining

EEKE dataset PCA factors	# hi-load terms	# shared w TechMining
Knowledge-based systems	8	2
Natural language processing	5	2
Rough set	4	0
Upper ontology	5	1
Relational database	9	1
Social Semantic Web	4	1
Personal knowledge management	9	4
Visualization	4	2
Concept mining	3	1
Fuzzy logic	5	1
Social media	3	2
Query optimization	4	0
Machine learning	4	4
Theoretical computer science	6	1
Spatial analysis	2	0
Evolutionary algorithm	4	0
Data warehouse	5	1
Grid	2	0
Data science	2	2
Information retrieval	2	1
SUM	90	26

Most Frequent & Interesting Phrases in common: TechMining + EEKE

Method-oriented Terms

- Bibliometrics
- Text Mining
- Citation
- Big Data
- Network Analysis
- Patent Analysis

Substantive Terms

- China
- Collaboration
- Emerging Technologies
- Climate Change
- Interdisciplinarity

Summary

Tech Mining: Lessons Learned

- Analysts needs vary
  - Want analytical software ranging from:
    - "Hands on" full control
    - "Easy button" simple
- End-users want different things
  - Researchers: a few novel papers to read
  - Research managers: "10,000 foot perspective" & opportunities
  - Executive Suite: Want answers – simple visuals and clear options
- End-users relate to different presentations
  - What the content?
    - What is the "Right" amount of information
  - How is it presented?
    - Multiple modes
    - Familiar means

Everyone wants an easy “What”

- Finding the “WHAT” in records is a real challenge
- One of the fundamental issues is there are two major types of “What’s”

➢ The Analyst’s What

- An analyst with Subject Matter Expertise has an expected “What” in mind when they look at data based on their own knowledge. So their “What” is sometimes not represented in the data.

➢ The Data What

- Algorithms let the data speak for itself. The “What” is not predetermined or often based on information outside of the data
  - The AI Exception – Training can bring in external information

- The two “What’s” often do not agree…
  - If the “What” an algorithm finds does not match what an expert expects, the expert will not trust the algorithm and will rely on expert opinion (even if that opinion is wrong or obsolete).

Discussion/Questions

Resources

- Porter, A.L., and Cunningham, S.W. (2005), *Tech mining: Exploiting new technologies for competitive advantage*, Wiley, New York.
- Porter, A.L., Zhang, Y., Huang, Y., and Wu, M. (2020), Tracking and mining the COVID-19 research literature, *Frontiers in Research Metrics and Analytics* 5: 594060, 1-18; doi: 10.3389/frma.2020.594060.
- Zhang, Y., Zhang, G., Zhu, D., & Lu, J. (2017), Scientific evolutionary pathways: Identifying and visualizing relationships for scientific topics. *Journal of the Association for Information Science and Technology*, 68(8), 1925-1939.
- Porter, A.L., Garner, J., Carley, S.F., and Newman, N.C. (2018), Emergence scoring to identify frontier R&D topics and key players, *Technological Forecasting and Social Change*, 146, 628-643; doi.org/10.1016/j.techfore.2018.04.016.
- Kwon, S,, Liu, X., Porter, A., & Youtie, J. (online, 2019), Research addressing emerging technological ideas has greater scientific Impact), *Research Policy*, 48 (9); https://www.sciencedirect.com/science/article/pii/S0048733319301544.

vantage point®
Search Technology

Alan Porter & Nils Newman | June 23, 2022

Empowering analysts for over 20 years

TheVantagePoint.com