A Measure of Staying Power: Does the Persistence of Emergent Concepts Significantly Vary by Technology Space?

STEPHEN CARLEY (IISC)

ALAN PORTER (GEORGIA TECH, SEARCH TECHNOLOGY)

NILS NEWMAN (IISC, SEARCH TECHNOLOGY)

JON GARNER (SEARCH TECHNOLOGY)

What is Technical Emergence?

Technical Emergence is a concept that has attributes of:

- Novelty
- Persistence
- Community
- Growth

The four dimensions

To be emergent, a concept must have all four attributes.

All four attributes exist as traits in the scientific, technical, and patent literature.

All four traits can be measured using bibliometric and ‘tech mining’ techniques.

This combination means we might have a chance to do effective forecasting.

Novelty

- One cannot really predict the appearance of a concept that does not yet exist; but one can analyze the past rate at which new concepts have emerged within a specified technical area.
- One can track and forecast progressions of incremental change.
- One can also use past activity to determine a probability for future radical change, but with a higher degree of uncertainty.

Persistence

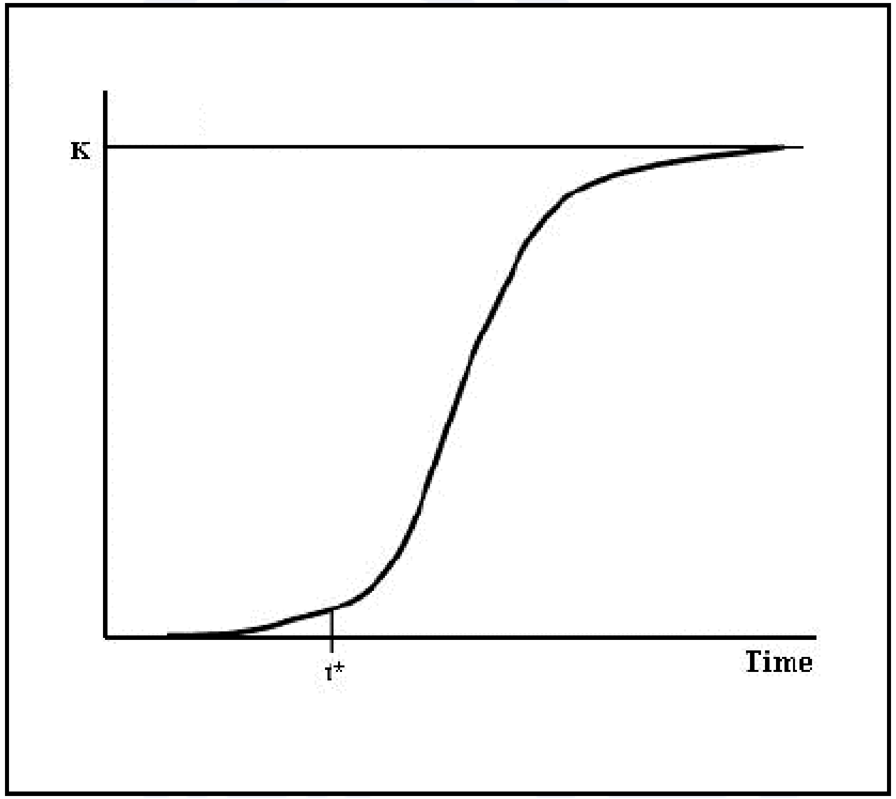
- Persistence is easy to measure and allows for effective forecasts.
- However, there is still a great deal of uncertainty as to how persistent a concept must be to be considered persistent.
- The scale of the concept and the technical area influence the behavior of persistence.
- Persistence can be a source of noise in forecasting process.

Community

- Community can be difficult to measure.
- Complete and accurate information on all documents' contributors is not always available.
- Web of Science/Scopus work reasonably well. EI Compendex/INSPEC are more problematic. Patents vary, depending on authority.
- Cleaning of organizational and personal names, and accurate matching of people with organizations, is a challenge.
- Once cleaned, the social network analysis needed to measure and forecast community is reasonably well understood.
- However, there is still debate on the level of analysis to apply when looking to ascertain "a community."

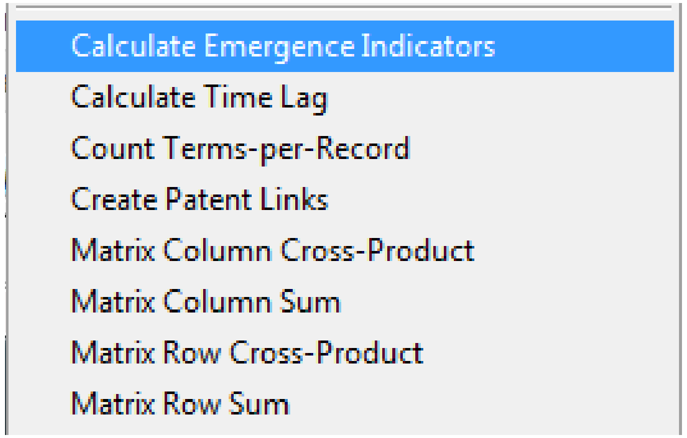
Growth

- Growth is tricky in that it comprises multiple dimensions:
  - Growth within the concept's technology space.
  - Growth into other technology spaces.
  - Growth within a community.
- Once identified, there are a variety of ways to forecast future growth.
- One technique involves curve fitting to logistics curves.

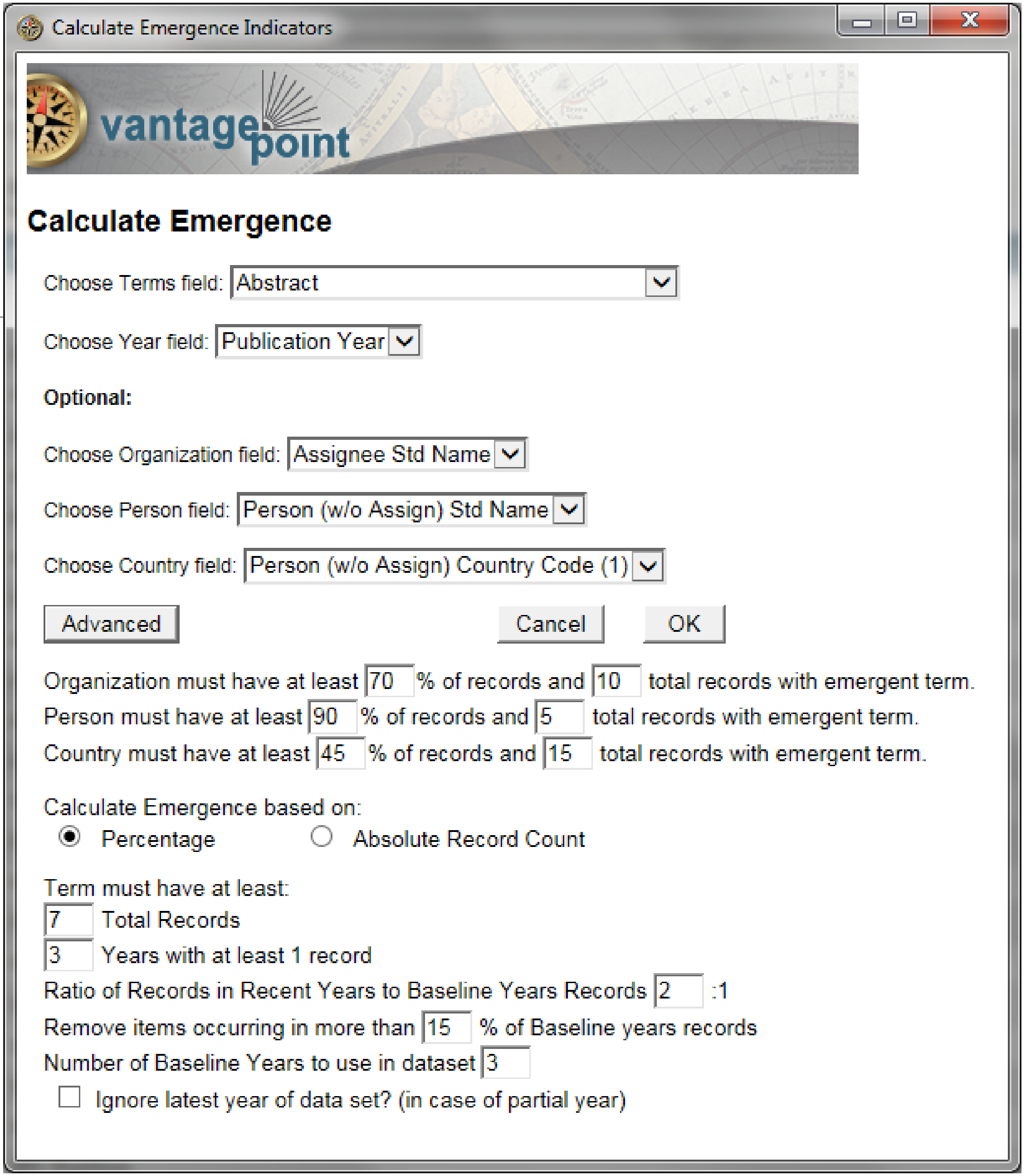


The Details

- The April 2015 release of VP9 includes an emergence script.
- This script calculates emergence for a target field and then identifies organizations, people, and countries with high concentrations of emergent terms.
- Script works best with 10 years of data.



Script Settings



Calculate Emergence

Choose Terms field: Abstract

Choose Year field: Publication Year

Optional:

Choose Organization field: Assignee Std Name

Choose Person field: Person (w/o Assign) Std Name

Choose Country field: Person (w/o Assign) Country Code (1)

Organization must have at least 70 % of records and 10 total records with emergent term.
Person must have at least 90 % of records and 5 total records with emergent term.
Country must have at least 45 % of records and 15 total records with emergent term.

Calculate Emergence based on:
● Percentage ○ Absolute Record Count

Term must have at least:
7 Total Records
3 Years with at least 1 record
Ratio of Records in Recent Years to Baseline Years Records 2 :1
Remove items occurring in more than 15 % of Baseline years records
Number of Baseline Years to use in dataset 3
□ Ignore latest year of data set? (in case of partial year)

Today’s Focus: Persistence
Why is Staying Power Important?

- Emergent Research is Preferable to non-Emergent Research (for identifying high impact research).
- Recurring, or Persistent, Emergence is Preferable to Short-Lived Emergence (for identifying high impact research of lasting value) (e.g. NBA championships).
- Given that Emergent Research is - by definition - persistent (within a given 10-year time period), persistent emergence can be referred to as research which is persistently persistent.
- This is research unique in that it distinguishes itself from a corpus of research that’s already emergent.

Research Questions

- How is the behavior of persistence influenced by (i) technical domain and (ii) scale?
- Which of these has the greater impact?

Case Study: Dye-Sensitized Solar Cells (DSSCs)

- DSSCs provide an ideal example of an emergent technology which has attracted considerable attention in recent years
- The dataset used in this study, which comes from the Web of Science, was developed by Alan Porter (Georgia Tech) and Ying Guo (BIT) using a multi-step Boolean search algorithm
- The entire dataset spans 1991 to 2014
- It can be deconstructed into 15 10-year datasets, each of which represents a time period to measure persistence in this study

Top 10 Persistently Emergent Authors

Emergent Author	# Times Emergent
Durrant, James R	13
Gratzel, Michael	12
Yanagida, Shozo	12
Hara, Kohjiro	12
Hagfeldt, Anders	12
Sugihara, Hideki	12
Zakeeruddin, Shaik Mohammed	11
Sayama, Kazuhiro	10
Nazeeruddin, Mohammad Khaja	10
Dai, Songyuan	10

Top 10 Persistently Emergent Affiliations

Emergent Affiliations	# Times Emergent
Ecole Polytech Fed Lausanne	15
Natl Renewable Energy Lab	14
Univ London Imperial Coll Sci Technol & Med	14
Uppsala Univ	13
Natl Inst Adv Ind Sci & Technol	13
Osaka Univ	12
Chinese Acad Sci	12
Univ Bath	10
Peking Univ	10
Univ Tokyo	10

Emergent Terms

The term field in this data is generated by:

- Combining Abstract and Title phrases into a single field
- Removing those terms with fewer than 2 instances
- Applying 5 thesauri from ClusterSuite (O’Brien et al., 2013)
- Running a general list cleanup in VantagePoint
- Dividing the remaining field into unigrams + multigrams
- Processing the former using a WOS stopwords thesaurus and the latter via a Folding NLP Terms algorithm
- Recombining the processed unigram and multigram fields into a single terms field

Top 10 Persistently Emergent Terms

Emergent Terms	# Times Emergent
impedance spectroscopy	9
power conversion efficiency	8
photovoltaic performance	8
power conversion	7
hydrothermal method	7
electrochemical impedance	6
electrochemical impedance spectroscopy	6
efficient dye	6
electron microscopy	6
dye adsorption	6

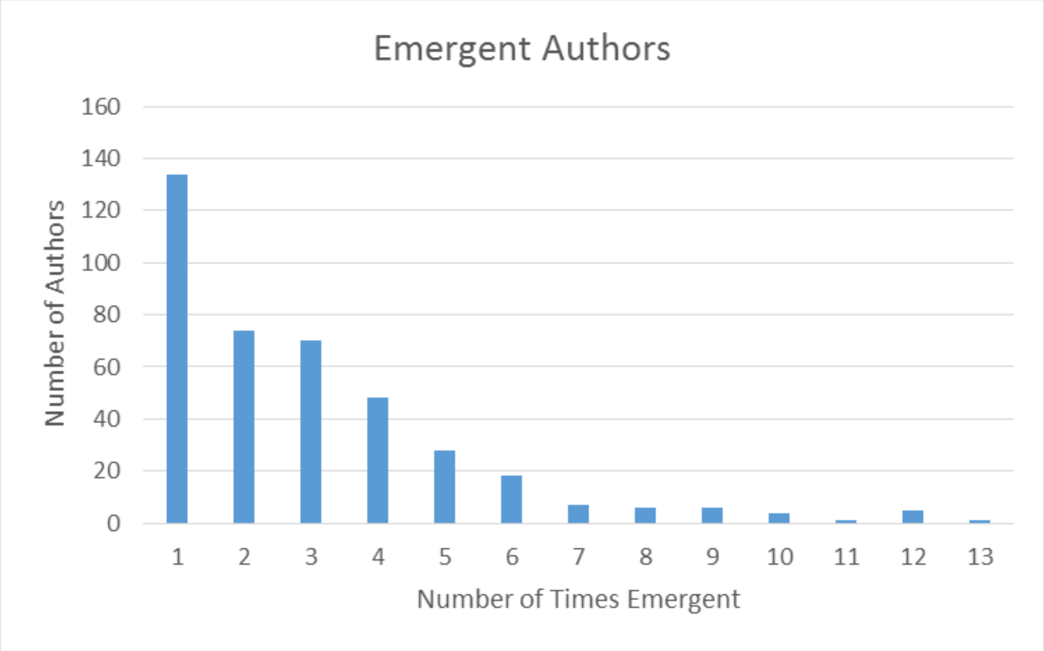
Persistence Trends Among Emergent Authors, Affiliations and Terms

Number of Terms	# Times Emergent
1	9
2	8
2	7
10	6
11	5
39	4
64	3
119	2
335	1

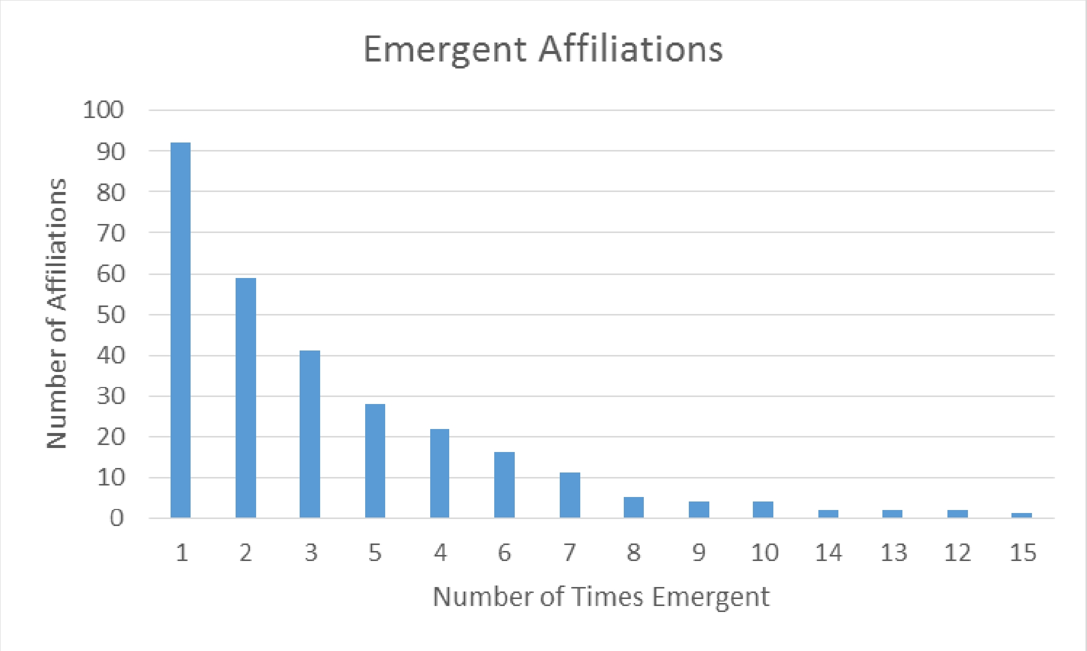
Number of Authors	# Times Emergent
1	13
5	12
1	11
4	10
6	9
6	8
7	7
18	6
28	5
48	4
70	3
74	2
134	1

Number of Affiliations	# Times Emergent
1	15
2	14
2	13
2	12
4	10
4	9
5	8
11	7
16	6
28	5
22	4
41	3
59	2
92	1

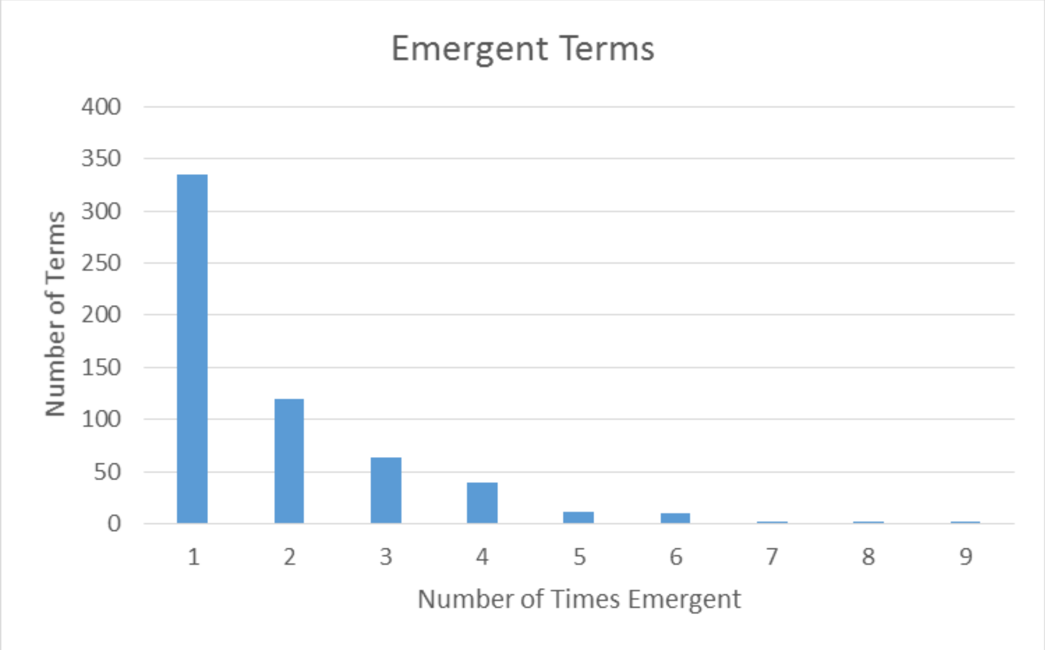
Charting Persistence Trends for Emergent Authors, Affiliations and Terms



Emergent Authors



Emergent Affiliations



Emergent Terms

- It’s interesting how smooth the rate of change is for Emergent Affiliations looks given it’s the smallest (3k) of all datasets (Authors is 24k and Terms is 30k)

What is the influence of domain on persistence?

- DSSCs can be deconstructed into sub-domains, or clusters, using a thesaurus (developed by Ismael Rafols) which groups Web of Science Categories into their respective clusters
- It is hypothesized that within these domains the behavior of persistence (as well as other emergence outputs) is likely to vary
- The clusters used in this analysis are: Physical Science and Engineering, Biology and Medicine, Environmental S&T and Psychology and Social Sciences

In which cluster to we observe the most emergence variance?

Field	Variance in # Emergent Entities Across all 15 Datasets (All Clusters) (N=13,196)	Variance in # Emergent Entities Across all 15 Datasets (Physical Science and Engineering) (N=12,893)	Variance in # Emergent Entities Across all 15 Datasets (Environmental S&T) (N=324)	Variance in # Emergent Entities Across all 15 Datasets (Biology and Medicine) (N=368)	Variance in # Emergent Entities Across all 15 Datasets (Psychology and Social Sciences) (N=16)
Authors	9,770.65	9,827.02	0.00	0.00	0.00
Affiliations	5,692.92	5,571.26	0.06	0.00	0.00
Countries	74.46	84.53	0.16	0.06	0.00
Terms	959.42	1,178.69	60.29	20.38	0.00

What is the Influence of Scale on Persistence?

- Both very small and very large datasets tend to produce unexpected emergence (and persistence) results.

- But how small is small and how large is large?

- DSSCs can be deconstructed into numerous sub-datasets using a random sample script.

- Drawing random samples the DSSC is divided into thirds for comparative purposes.

The Impact of Scale on Emergent Terms

	1/3rd Random Sample	2/3rd Random Sample	Entire Population
Average Emergent Term Growth Rate	43%	32%	20%
Emergent Term Variance Across the 15 Datasets	1,757	1,398	959

- Here we observe decrease in emergent term growth rate as well as emergent term variance as scale increases

The Impact of Scale on Emergent Affiliations

	1/3rd Random Sample	2/3rd Random Sample	Entire Population
Average Emergent Affiliation Growth Rate	58%	57%	41%
Emergent Affiliation Variance Across the 15 Datasets	546	2,288	5,693

- Here the trends for emergent affiliation growth rate and emergent affiliation variance go in opposite directions: while emergent affiliation growth rate declines with sample size, emergent affiliation variance noticeably increases with the same.
- It would seem natural to suppose that the latter would move in step with emergent variance for authors (but we next observe a that emergent variance for authors increases are a remarkably faster rate).

The Impact of Scale on Emergent Authors

	1/3rd Random Sample	2/3rd Random Sample	Entire Population
Average Emergent Author Growth Rate	57%	78%	77%
Emergent Author Variance Across the 15 Datasets	150	1,995	9,771

- The impact of scale on emergent authors is a different picture: here we observe a general increase in average emergent author growth rate and a remarkable increase in variance in emergent author variance as scale increases.
- Authors outperform affiliations and terms in the spread between average and emergent growth rates.
- The fact that the spread for authors significantly outpaces the spread for terms seems to indicate that an increasing number of authors are gravitating to the DSSC field and focusing attention on preexisting emergent concepts.

Discussion

- Which influences the behavior of persistence more between domain and scale?

- While scale shows more impact in this particular dataset we leave room for the possibility of different results in a different dataset.

- Across all scales emergent authors show not only the strongest growth rate, but also the most variance across the 15 datasets, which can be taken as evidence of an increasing number of scholars gravitating toward a field with preexisting emergent concepts.

Further Questions

- Does persistent emergence in one domain translate into or increase likelihood of an emergent presence in other domains?

- Concerning the effect of scale - what sized dataset produces the most robust set of results? Does this vary by domain? For massively large datasets does a random sample suffice?

- The previous discussion centers on the effects domain and scale have on persistence, but are there other influential factor(s) that are being overlooked?

References

O'Brien, J.J., Stephen Carley and Alan L. Porter (2013). ClusterSuite [computer software], Atlanta, GA [available via VPInstute.org].

Foresight and Understanding from Scientific Exposition (FUSE). 2014. Available: http://www.iarpa.gov/index.php/research-programs/fuse. Accessed: 2016 Sep 13.

Search Technology. 2012. Available: www.thevantagepoint.com. Accessed: 2016 Sep 13.