LexisNexis™

# Semantic Searching – Challenges & Solutions:

## Making it work in the real world of professional searching

October 2008

Brian Elias
*Director, IP Product Management*

# The Goals of Searching

**"To find what you're looking for…even if you're not sure what you're looking for… even if you don't know what you don't know"**

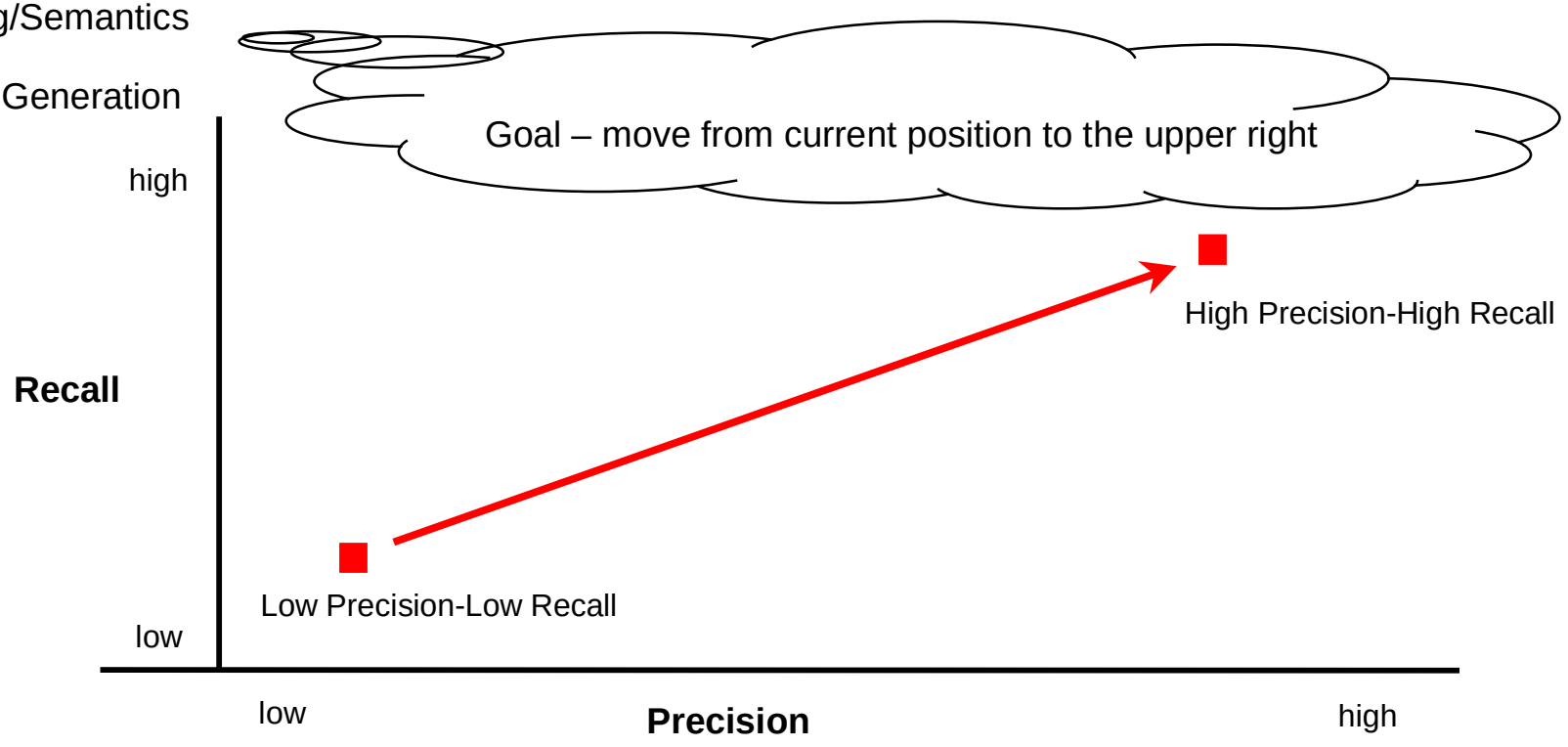
**The Goals Are the Same:**

- **Precision:** Retrieving a high level of accurate results relevant to your search query (a measure of exactness)
- **Recall:** Retrieving a high percentage of relevant documents (a measure of completeness)

**How We Get There is Changing:**

***Distributed Semantic Searching***

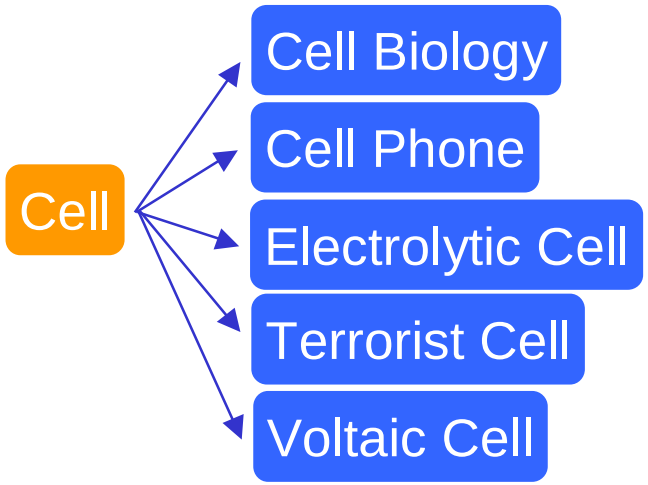
- Machine Learning/Semantics
- Inference
- Automatic Query Generation



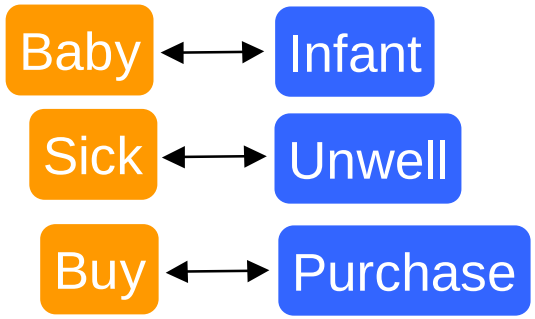
# Why is searching so challenging?

## Challenges with Language

- **Polysemy** – Single words (or phrases) with multiple meanings, like:




- **Synonymy** - Multiple words with the same meaning, like:




**Challenges Specific to Patent Searching:**

- Disparate Nomenclature across art
- Applicants can act as own Lexicographers
- Industry language changes over time
- Technical abbreviations often conflict with common words (Au = Gold, Australia)
- Many technical abbreviations are common noise words (e.g. He = Helium, Be = Beryllium)
- Lengthy compound expressions are often misspelled
- 1000's of COCs a year

- **Human** – Search experience, domain knowledge, accidental mistakes and idiosyncrasies in language

# Why is searching so challenging? (continued)

## Challenges with Search Systems

- **Search is unmanageable:**
  - Too many search engines
  - Too many indexes
  - Too much disparate data
- **Search is Disconnected:**
  - Multiple search partner environment.
  - Disparate partner Indexes are not compatible
- **Search Intelligence is relatively static:**
  - State of the art still involves Lexicons, Synonym Lists and Thesauri.
  - Search suffers from chronic amnesia (user starts over on each search)
- **Search is Time Consuming:**
  - Disparate, disconnected systems with no common language require too much time for one to be exhaustive and comprehensive.

# Why is searching so challenging? (continued)

## Shared Limitations Amongst All Search Vendors

- Most are keyword based
- Few have learning systems, "systemic amnesia"
- None use any sort of true semantic search
- All issue the "lowest common denominator" search logic
- Most are based on 12 year old keyword technology.

Others claim true semantic search, but the results have been mixed…











# What is Semantic Searching?

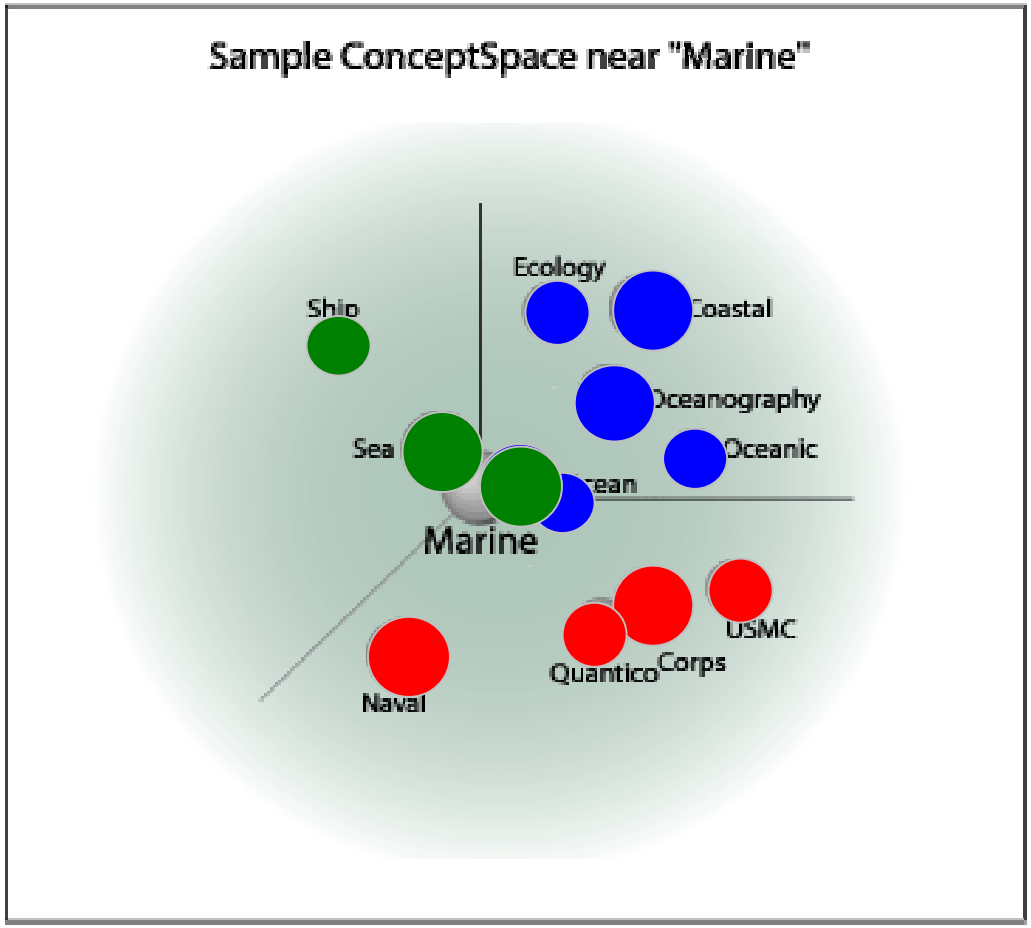
***Semantic Search (SS) or concept search technology provides users the ability to search with not only keywords, but concepts. Concept Search uses machine learning methods to surface meaning in documents while disambiguating semantic search queries.***

- **<u>Semantic Search Methods</u>: Latent Semantic Analysis (or LSA) , Bayesian Inference and Support Vector Machines (SVM)**
  - Analyzing statistical co-occurrence of terms and phrases to surface the relationship between words inside of a hyperspace.
  - Using these relationships to search on concepts and the meaning of words, phrases or larger sets of textual input (sentences, paragraphs, whole documents, etc.)
  - Ranking by putting the corpus of documents into a model that ranks their content in relation to the search query, adjusted by the LSA, and then comparing them to one another.

# What is Semantic Searching? (continued)

**LSA extracts every contextual relation among every term in every word or phrase within a collection (learning document set). It then generates a vector space representation of all terms based on those relations. Within that space, proximity is a strong indicator of conceptual similarity.**

**<u>The result</u>: similarities can be identified based on concepts within the material.**



# What are the Challenges with Semantic Searching?

## The Black Box - Surrendering Transparency, Control and Scale

### TRANSPARENCY

Semantic Search is effective and many times better than straight keyword searching, but thus far has locked users out of understanding **how** results are generated. We are forced to "trust" its algorithms, trust its function and trust its results. Virtually no system transparency.

### CONTROL

There is a surrendering of "control" with semantic search. Users cannot control how a query is constructed, much less the algorithms that create the search logic.

### SCALE

Semantic Searching requires a semantic index. Control of the search corpus via content indexing (and thus control over formatting, storage, etc.). This just does not scale. The web may never be semantically indexed.

**The Result**: **Semantic searching is not a practical search alternative to traditional search methodology. It may even be a liability.**

# We asked ourselves, what if?

- …you could capture and store all of the knowledge contained in the US patent database AND millions of NPL/ STM articles as a collective intelligence?
- …you could utilize that collective intelligence as a bank of virtual topic matter experts to help you with EVERY search you performed on ANY database?
- …there was no more search language - and you could just type as you speak, or cut and paste relevant text and have the system understand.
- …this search system continued to learn from documents and searches - dynamically growing its intelligence as the volume of available learning data increased?
- …you could benefit from the recall of semantic search and the precision of keyword search at the the same time?
- …you had complete transparency into the learning of the system and ultimate control over how queries were generated?

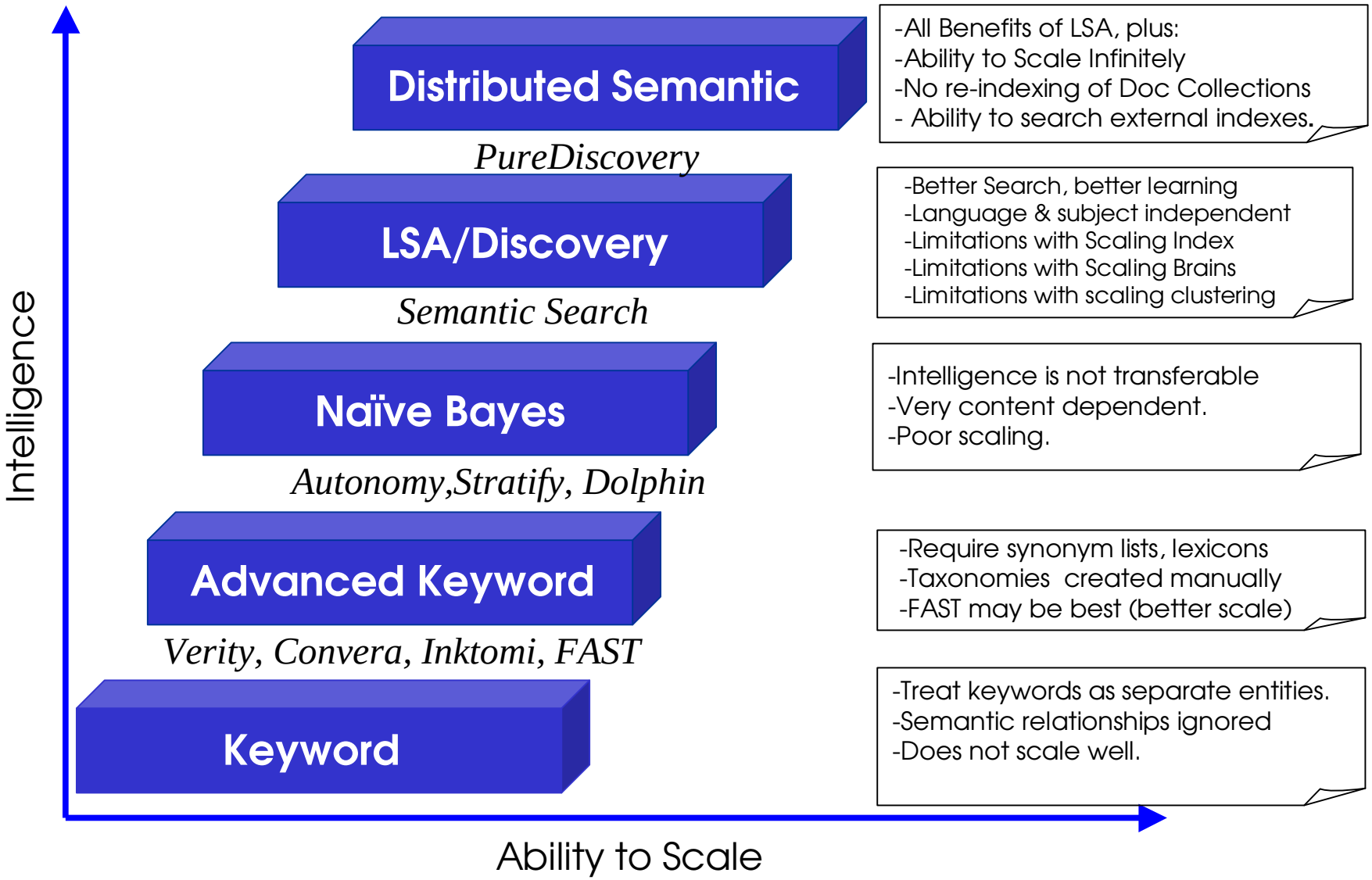
# Semantic Searching at LexisNexis

- Aggregates the intelligence of the entire US patent database
- Semantic processing occurs above the search index (no indices changed)
- Resultant “Queries” are generated / optimized for each search engine
- Professional searcher is given complete control.

**What is the expected result?**

- Significant improvement in “recall”
- Marked improvement in “precision”
- Leverages knowledge, experience and expertise of the professional
- Scalable across multiple content sets

# Distributed Semantic Searching (DSS)



# How is DSS different?

## Comparison Matrix for Lexis / PureDiscovery Semantic Search

| | Capability | LexisNexis/PD | Semantic Search | Notes: |
|---|---|---|---|---|
| Scalability | **Ability to Scale** | Unlimited. Does not require that client rebuild their indexes. | Must re-index all existing indexes that are to be searched. | Client databases are all currently indexed. Why do it again? |
| Scalability | **Ability to access external indexes** | Yes, PD separates the learning from the indexing. So, system learning and available sources can be built and managed separately. | No, clients must index the documents they are searching. | PD has connectors built to create and connect client created brains to all current client indexes. |
| Scalability | **Scalable Intelligence** | Yes. The PD Brain Library Architecture is built to support up to 200 brains per server. | No. Semantic Searches are optimized for single intelligence and very small brain collections. | While LSA is very good at learning from text, brains work best if they are homogenous in context. This requires that more brains be built on content containing multiple topics. |
| Intelligence | **Effectiveness of Learning - Dimensions of a Brain.** | Yes. Brain dimensions are calculated and adjusted independently on each PureDiscovery Brain (there can be 100's of Brains). | No. Brain Dimensionality is normally fixed on all Semantic Search brains. | Proper dimensions of a brain are critical to causing effective learning. Fixed dimensions force learning where it is not needed and fail to provide enough learning when it is needed most. Very rarely does a fixed dimension hit the mark. |
| Intelligence | **Configurable Vocabulary** | Yes. PD Brain optimizer looks at all potential words to be included in a brain build and removes the troublesome vocabulary. This toolset is configurable by client. | No. Intelligence indexes are locked. They do not allow the client to optimize vocabularies. There are no toolsets for optimizing vocabularies in most engines. | Brains are only as good as the words that are used to create them. Meta data, headers, footers and markup can ruin the integrity of the words. Tools for cleaning vocabulary are important. |
| Intelligence | **Supports Phrases** | Recognizes phrases dynamically. Uses phrases in search, even infers phrases, as well as in brain builds. | Straight LSA. Treats documents as a bag of words with no phrase recognition in queries or in the brain. | Phrases are critical in machine learning. Word combinations create specific meanings that are necessary for proper learning. |

# How does DSS work?

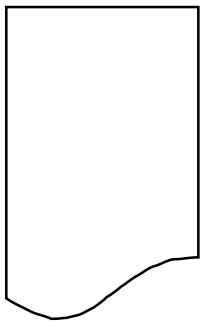
**Learning Document Set**
Tens of Millions of patents and related prior art full text documents

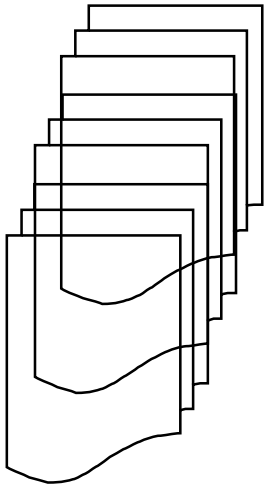
**Dynamic Clustering**
Documents are semantically collated into like categories.

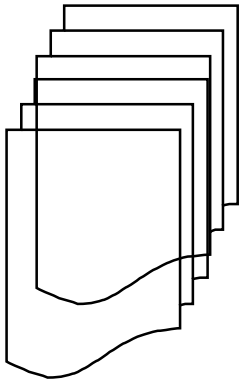
**Near De-Duplication**
Documents are semantically fingerprinted. Documents that are near duplicates are removed.

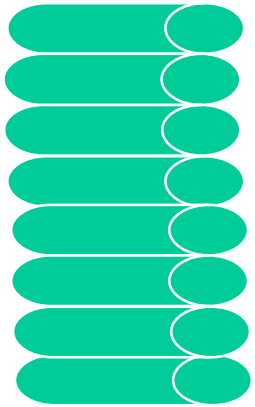
**Brain Library Creation**
Clusters are transformed into machine intelligences or “Brains”, each “tuned” to knowledge in a given area.

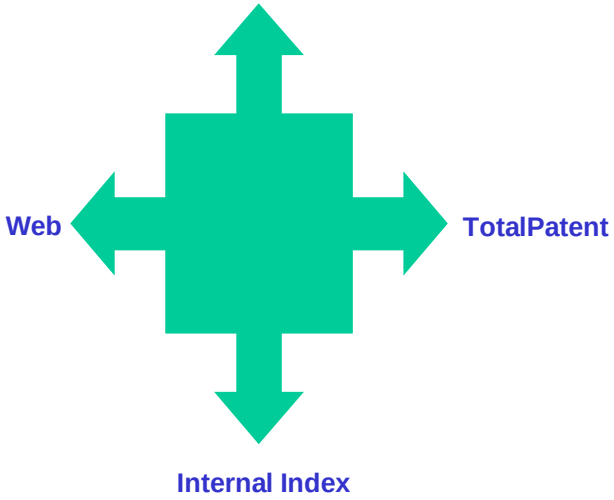
**ConceptBridge**
System uses the Brain Library to extract and infer related terms. Those terms are then crafted Into optimized searches for multiple indexes simultaneously.









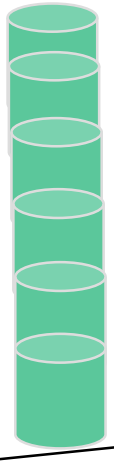


# How does DSS work in practice ?

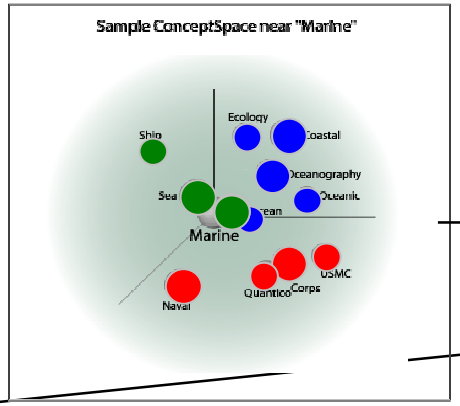
**1) User builds Query**

**Semantic Query**
* Words
* Sentence
* Paragraph
* Document, etc.

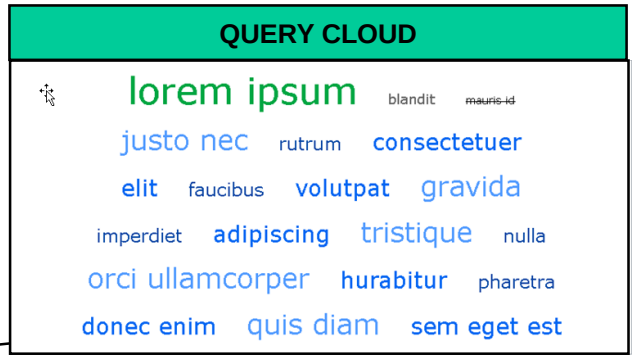
**2) Brain is Selected**



**3) Intelligence:** Query terms are extracted and inferred terms from the brain are added.




**4) Transparency:** QueryCloud is generated. User is shown everything generated by the machine learning.




**5) Automatic Query Generation:** The System uses the words extracted and generated by the brain to formulate fully optimized queries for each index being searched.

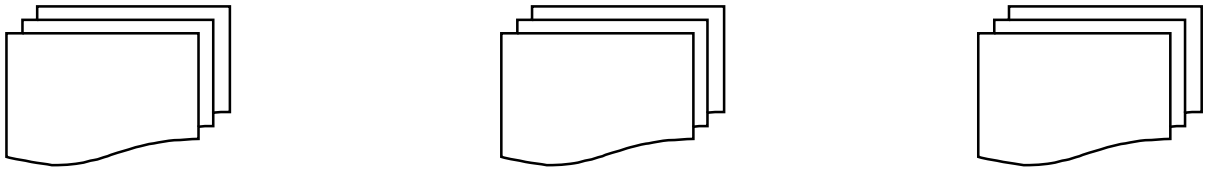
**Query Creation/Optimization**
Boolean
Ranked OR
Term Weights

## Total time elapsed = 1 second

**Query 1**: lorem and ipsum and (rutrum (.8) or volupat (.74) or elit(.63) etc.

**Query 2**: lorem and ipsum and (rutrum or vlupat or elit etc.)

**Query 3**: lorum ipsum rutrum volupat elit etc



**TotalPatent** | **Internal Database** | **Web**

# Why is this approach so promising?

- Full transparency into the results of the Semantic analysis – the Black Box "trust me" factor is minimized.
- Full enablement of Boolean tuning to allow the professional to leverage their experience and knowledge.
- Ability to use the engine and the intelligence to search multiple content sets that might be stored by numerous parties in various databases that are searched using numerous search engine technologies.

# What it doesn't do …

- Replace Boolean searching – instead it couples Boolean with state of the art machine learning techniques to create a unique, powerful and most importantly scaleable solution.

- Disintermediate the professional searcher or their experience/expertise – instead it offers a platform that casts a more comprehensive net from which searchers can work quickly to find the documents that are most relevant.

*Technology is rarely the solution by itself. The best role of technology is to empower the user in ways that make them more effective.*

# Why is this important?

*I've been searching with [Boolean] for many years. I know what I'm doing and know the value of the results. It works for me. I don't see a need to change."*

**Question: Are you really doing the best search you can?**

- shortest amount of time,
- highest precision,
- highest recall

Allowing you to analyze and assess the content in relation to the problem versus spending time getting the content to analyze?

**"How well do you know, what you think you know, and how often are you finding, what you don't know?"**

**OR …are you just staying ahead of expectations and the competition…and are those changing?**

# Which one is you?

*Two hikers are hiking through the woods when they come upon a charging bear. One hiker bends down and starts changing into his running shoes. The other hiker looks at him and says, "what are you doing, you can't outrun a bear!" The other hiker looks up and says, "I don't need to outrun the bear, I just need to outrun you!"*