Next-Generation Patent Searching

James Ryley, Ph.D.

Recall and Precision

- “Recall” is making sure that the answers you want are contained in the result set.
- “Precision” is making sure that ONLY the answers you want are contained in the result set.

Recall and precision are poor in conventional patent search engines.

Recall: The Problem of Synonymy

- Many words can express similar concepts (synonyms).
- For example, alternatives to “cell phone” include cellular phone, mobile phone, hand phone.

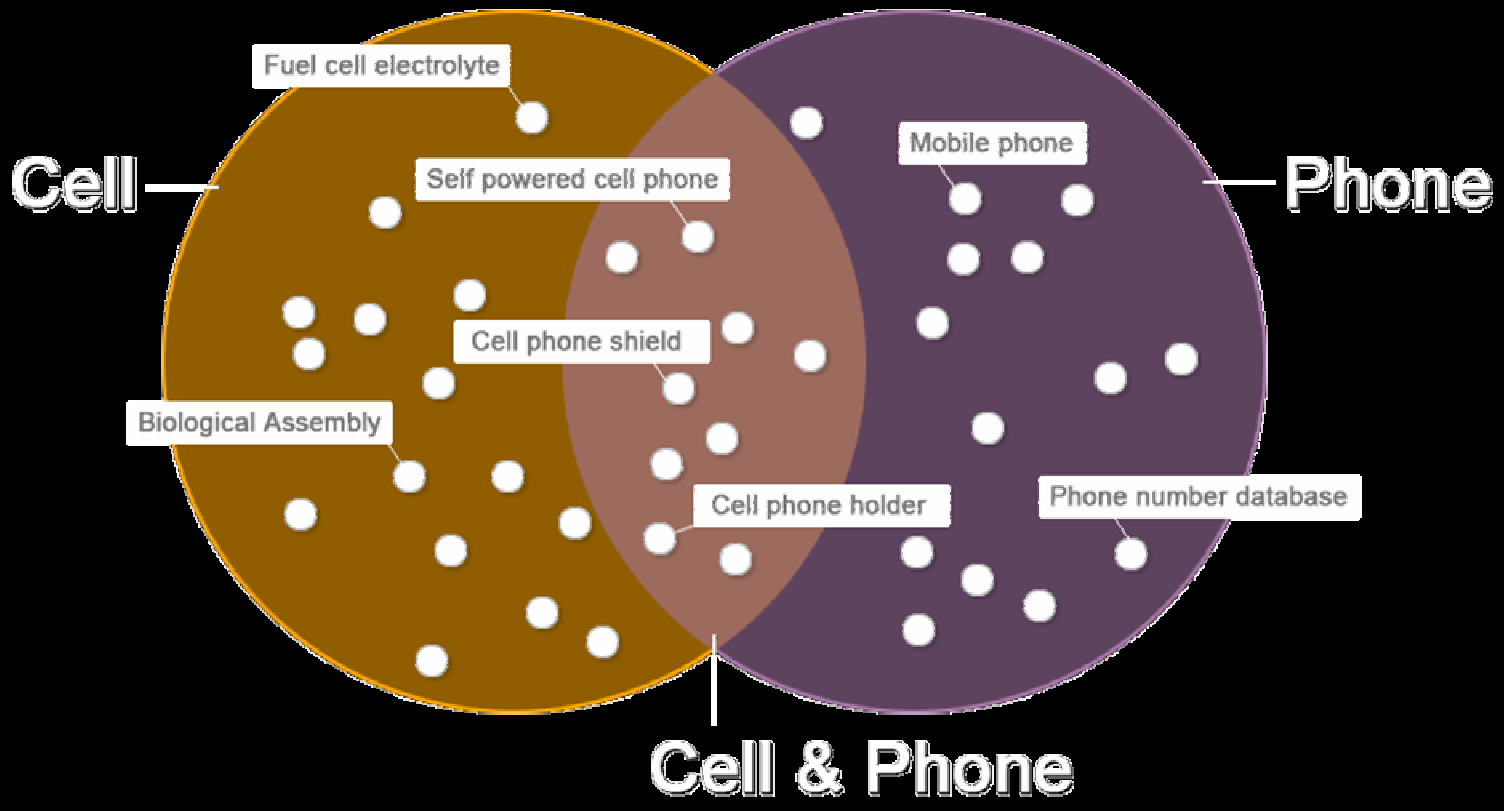
Precision: The Problem of Polysemy

- In addition to one concept being expressed by many words, one word can express many concepts:

Does the word “cell,” refer to a fuel cell, a jail cell, an electrolytic cell, a stem cell, a terrorist cell, a cell phone, or the Cell microprocessor in the PlayStation 3?

Conventional (Boolean) Search

- Search terms are simple filters. They are not “understood.”



The Ideal Search

- Automatically ignores irrelevant (polysemous) uses of input terms
- Understands concepts and synonyms, rather than being limited to literal input terms
- Has accurate ranking so that highly-relevant documents can be examined first

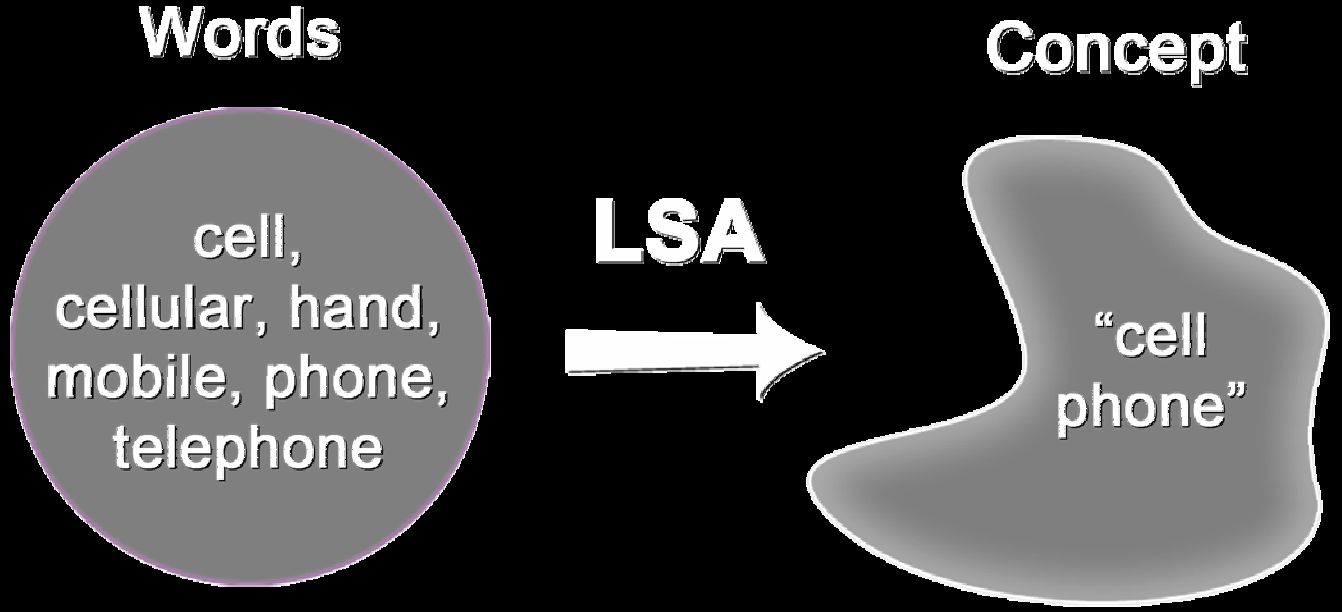
Latent Semantic Analysis addresses these needs.

Latent Semantic Analysis

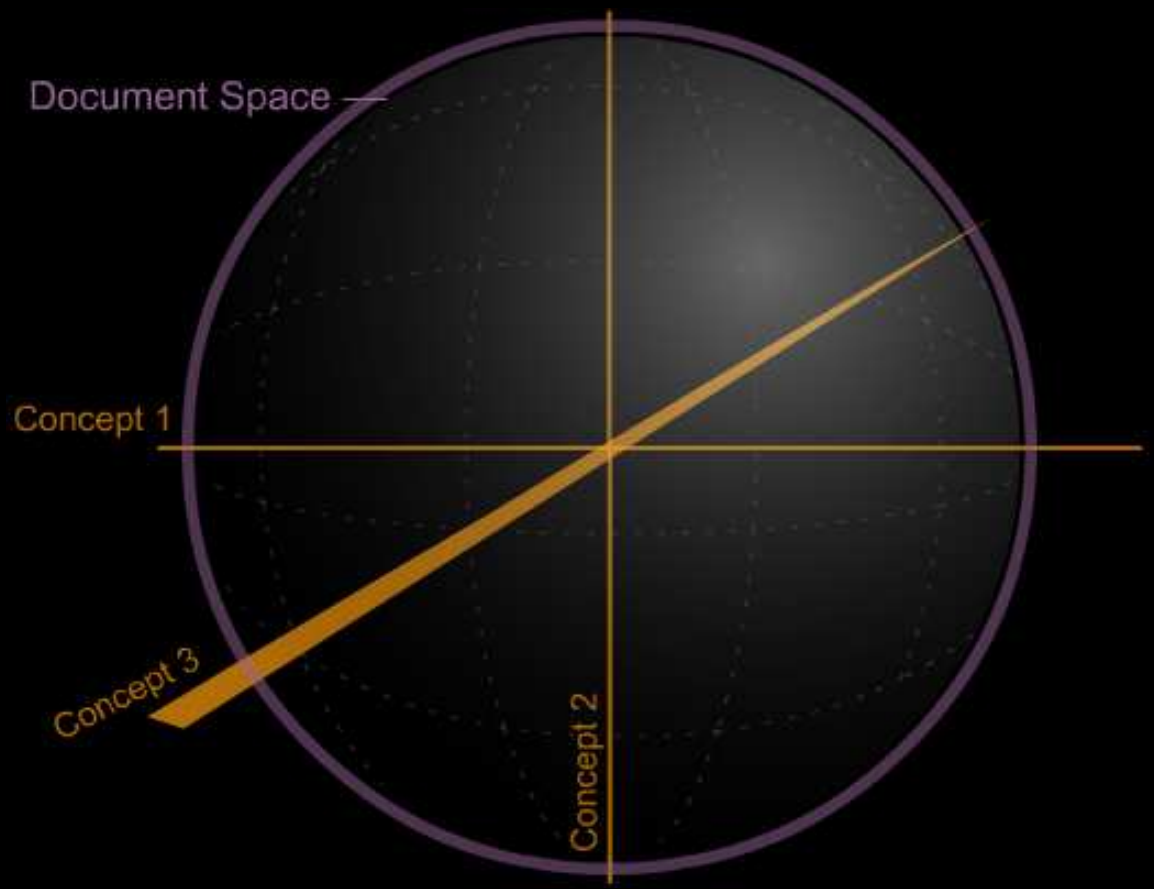
- Understands words conceptually, enhancing recall and precision
- Uses a spatial representation of documents to enhance precision through accurate ranking

Concept Mapping

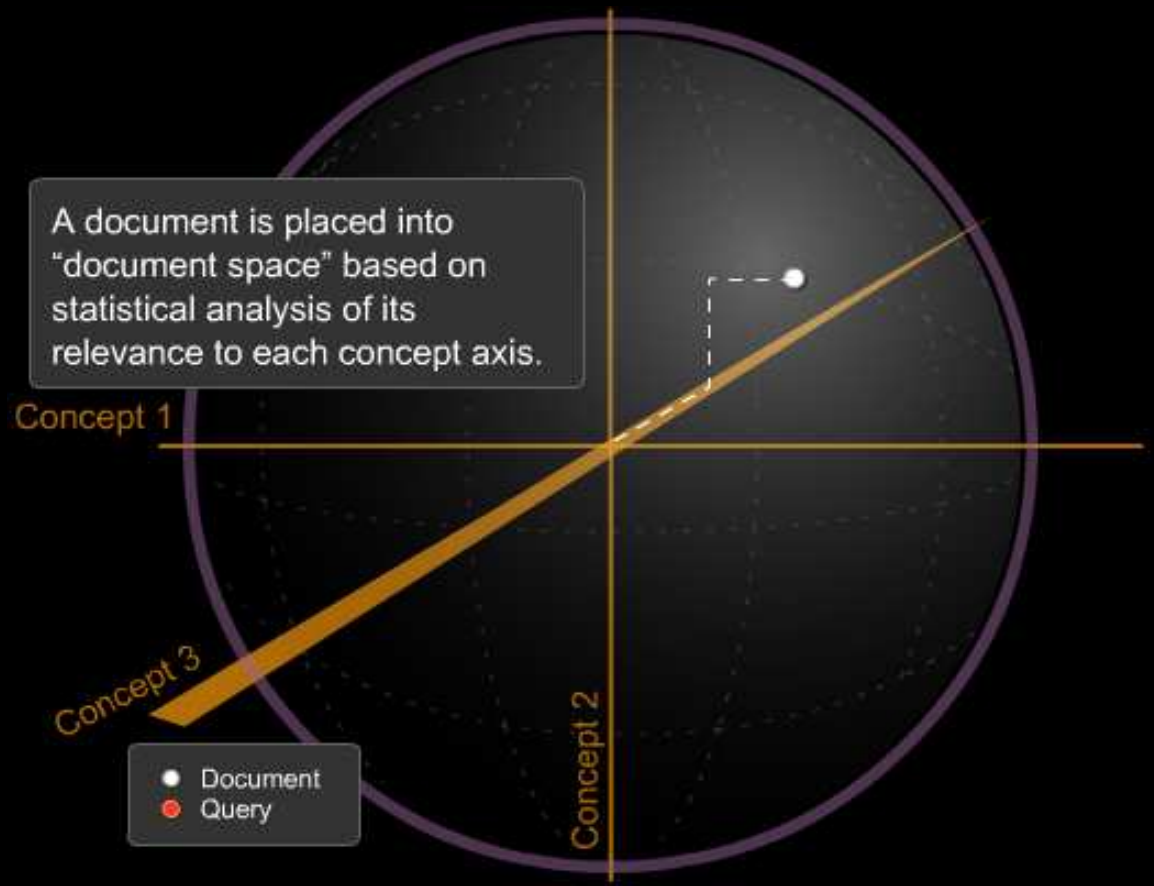
- Statistical techniques can be used to build a thesaurus of synonyms.



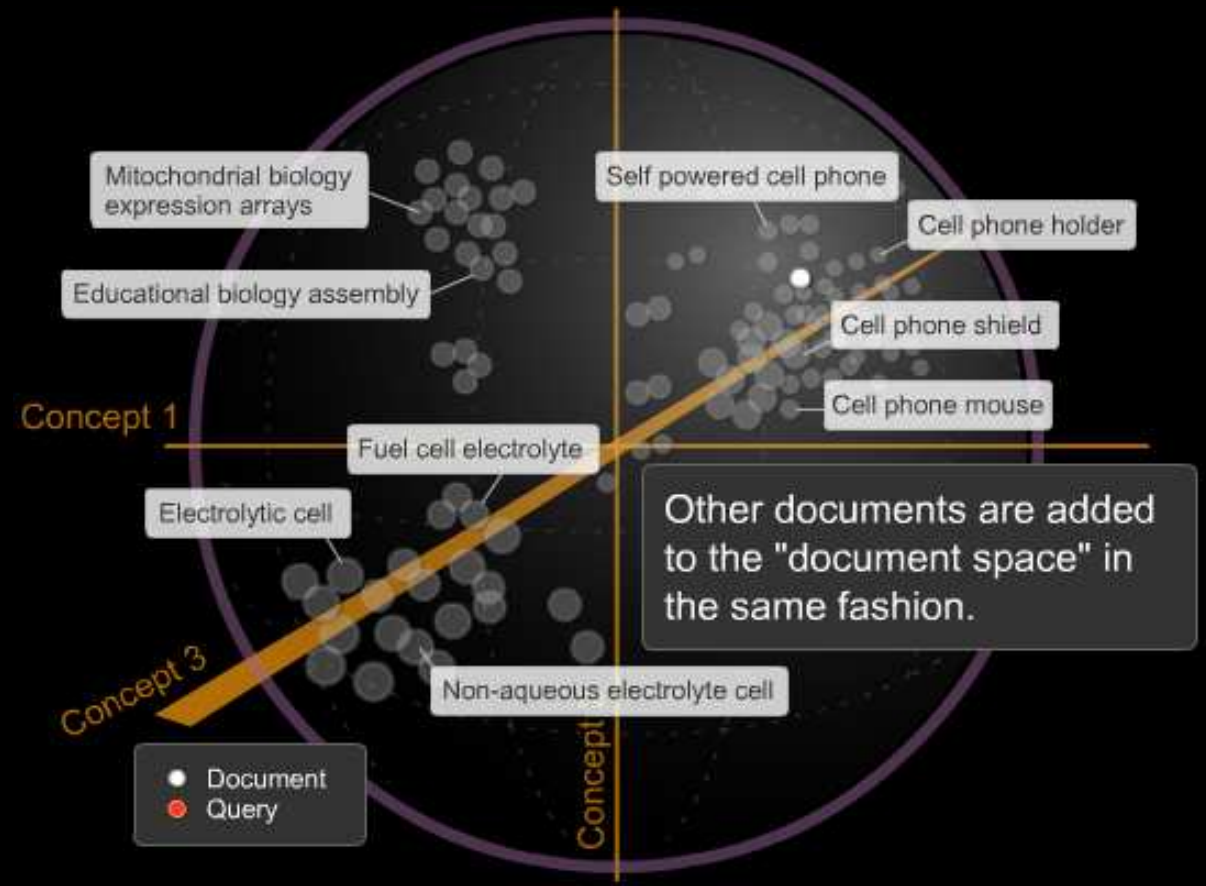
The Spatial Document Model



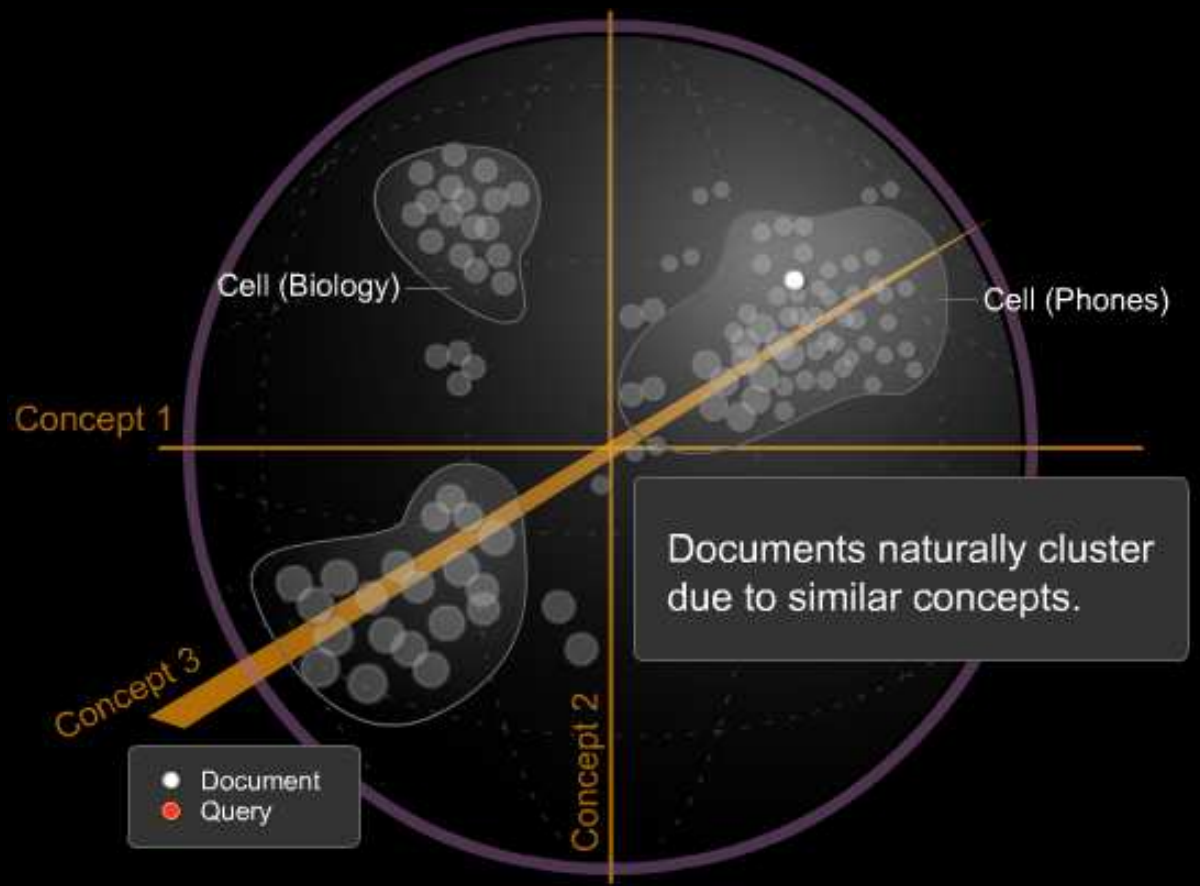
Locating a Document in Space



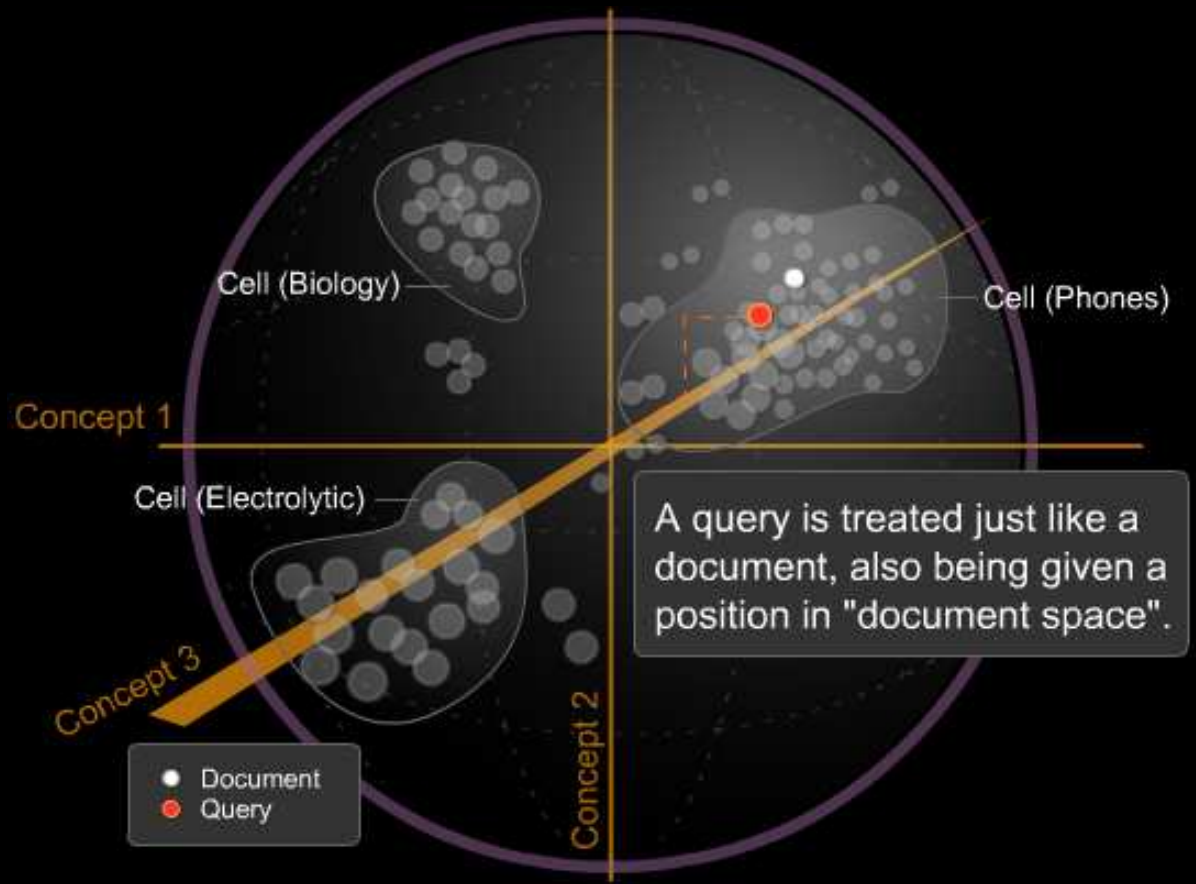
Filling Up Document Space



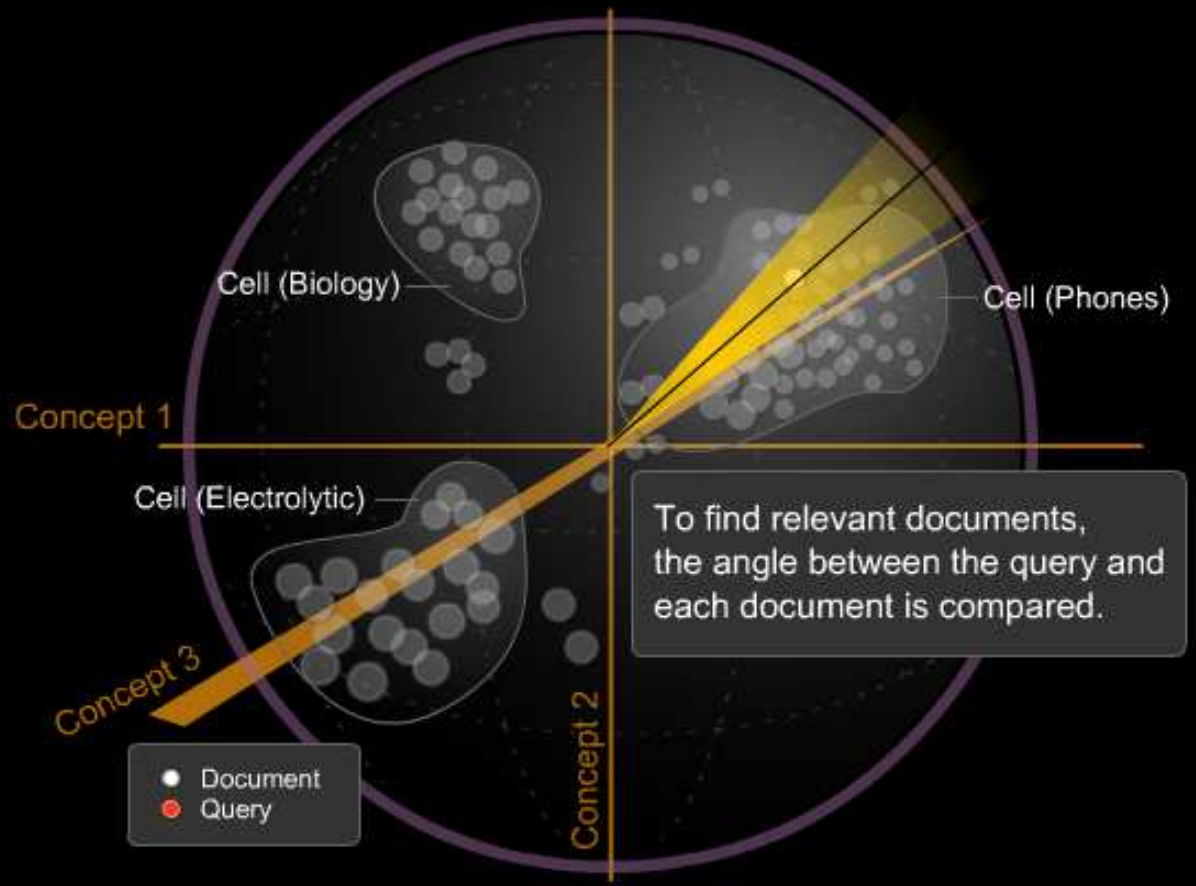
Spatial Clustering



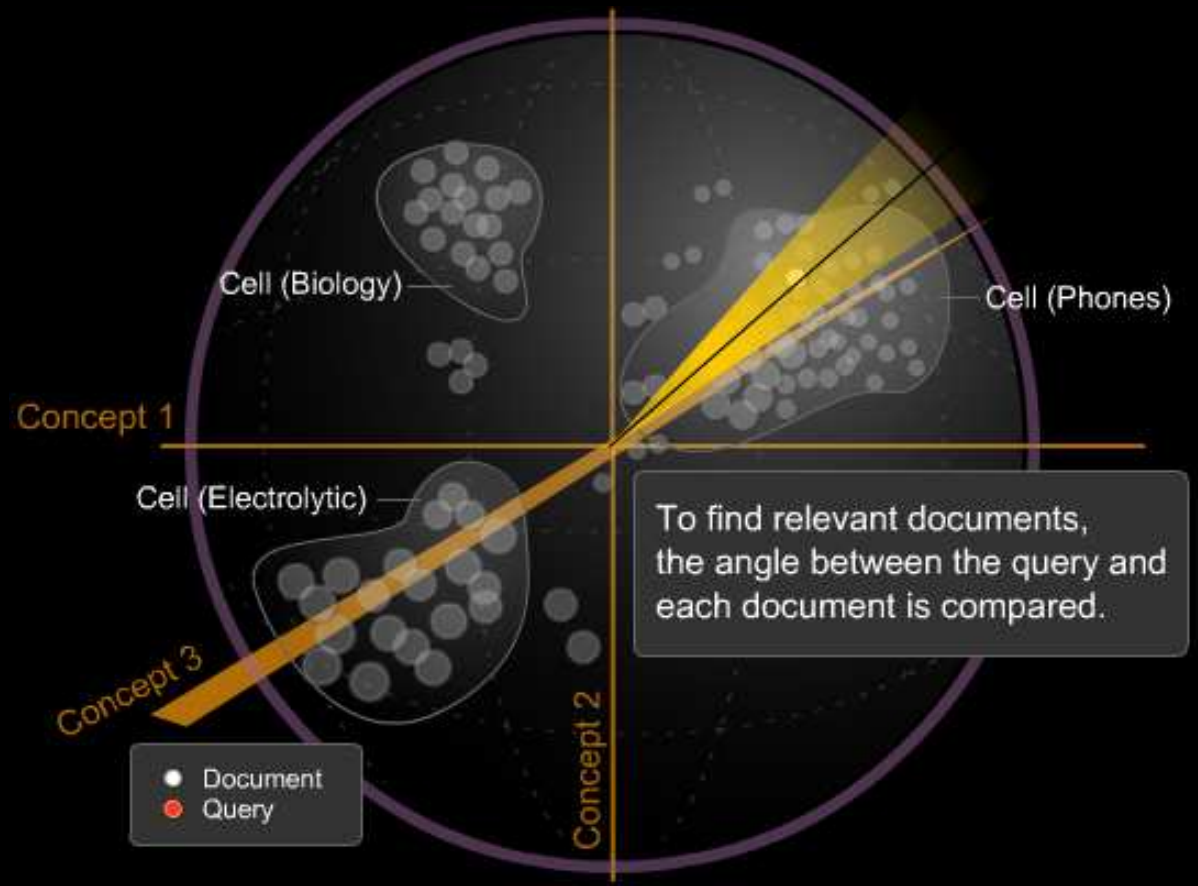
Queries in Document Space



Ranking Using Angles



Angles and Polysemous Terms



LSA Summary

- The system “understands” which words are synonymous and finds documents accordingly
- Polysemous words (“cell” in “cell phone” vs. “stem cell”), are appropriately ignored
- Results are ordered accurately by relevance

LSA has high recall, and high precision.

Other LSA Benefits

- Arbitrarily long queries can be used, without the need for Boolean logic (AND, OR, NOT)
  - Allows passages of text to be used as queries
  - Allows “more like this” queries where entire documents are used as the input query
- Allows automated grouping of documents by topic

If LSA is so Great, Why Isn't Everyone Using It?

- Mathematical and logical complexity
- Large computational demands
- *Partial* or *hybrid* solutions do exist, using parts of LSA grafted onto a traditional search, to reduce computational demands

Pure LSA Solution

- Computational issues have been solved, making LSA computationally tractable on huge document collections.
- Mathematical issues have been addressed, advancing the state of the art in accuracy.

A pure LSA solution, with high recall and precision, has been created.

Thanks!

- Contact info:
  - James Ryley, james.ryley@sumobrain.com
  - Phone: 410.977.5716
  - www.SumoBrain.com