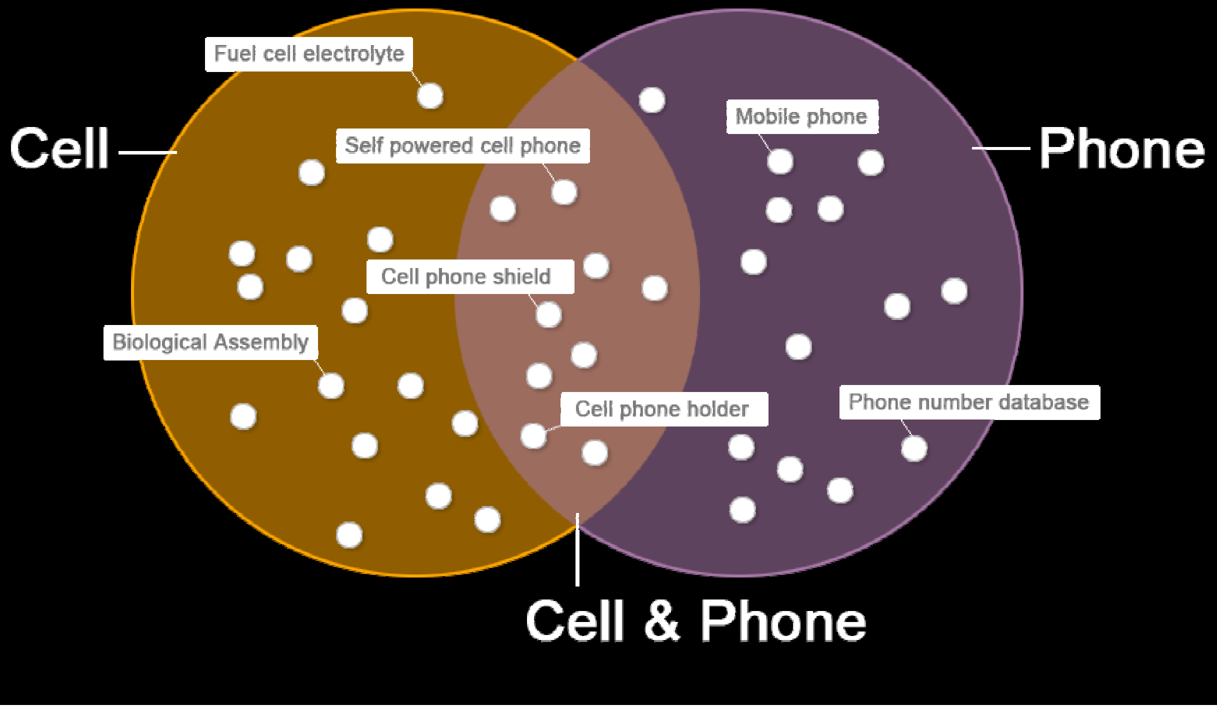
Using Conceptual Search in Scientific, Financial and Intellectual Property Databases

James Ryley, Ph.D.

Conventional (Boolean) Search

- Works on the basis of set intersection. There is no "understanding" of document language.



Recall and Precision

Recall and precision suffer when language is not understood.

Consider that, for the search query “cell AND phone”:

Documents containing the terms “mobile phone” or “hand phone” (common in Asia) will be missed.

Recall and Precision

Switching to “cell OR phone” leads to a result set containing documents about:

Terrorist cells
Jail cells
Electrolytic cells
Sony Playstation Cell^{TM} MicroProcessor
Stem Cells

and more…

Improving Recall and Precision in Boolean Search

- The solution is "simple," (though creates ranking nightmares) just create a query like this:

  (phone OR telephone OR radio) AND (mobile OR cell OR cellular OR hand OR wireless OR radio OR satellite OR portable OR … plus the terms for your specific interest in cell phones

The Ideal Search

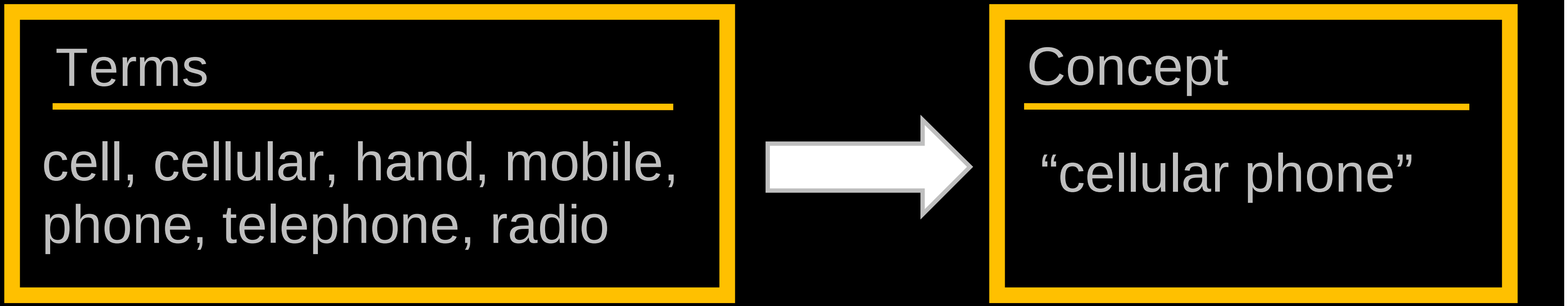
- Uses knowledge about word usage to:
  - Ignore irrelevant (polysemous) terms
  - Identify related concepts not specified in the query
- Has accurate ranking so that highly-relevant documents can be examined first.

Latent Semantic Analysis (LSA)

- Searches conceptually, not literally, enhancing recall and precision
- Uses a spatial representation of documents to enhance precision through accurate ranking

LSA Concept Mapping

- Statistical techniques are used to relate terms together.
- The result is the ability to find documents that do not contain the literal search terms.



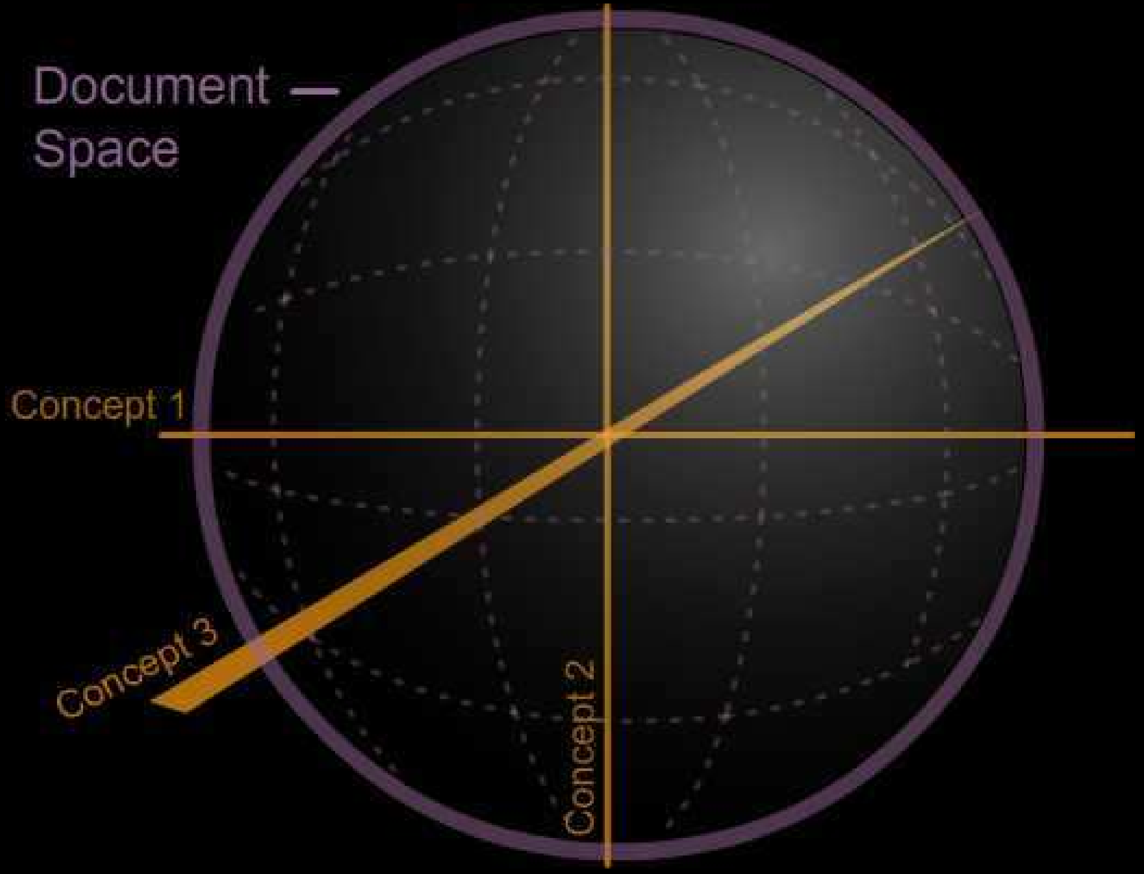
Terms

cell, cellular, hand, mobile, phone, telephone, radio

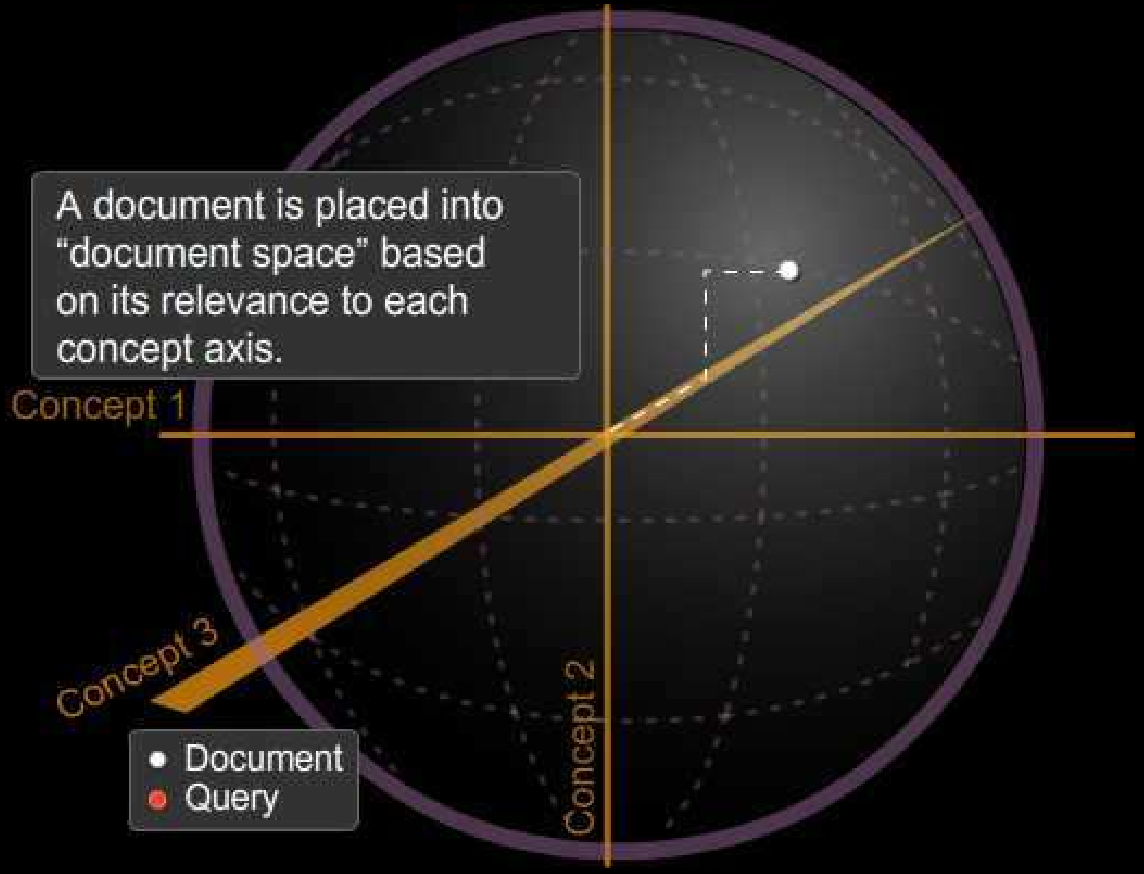
Concept

“cellular phone”

The Spatial Document Model



Locating a Document in Space

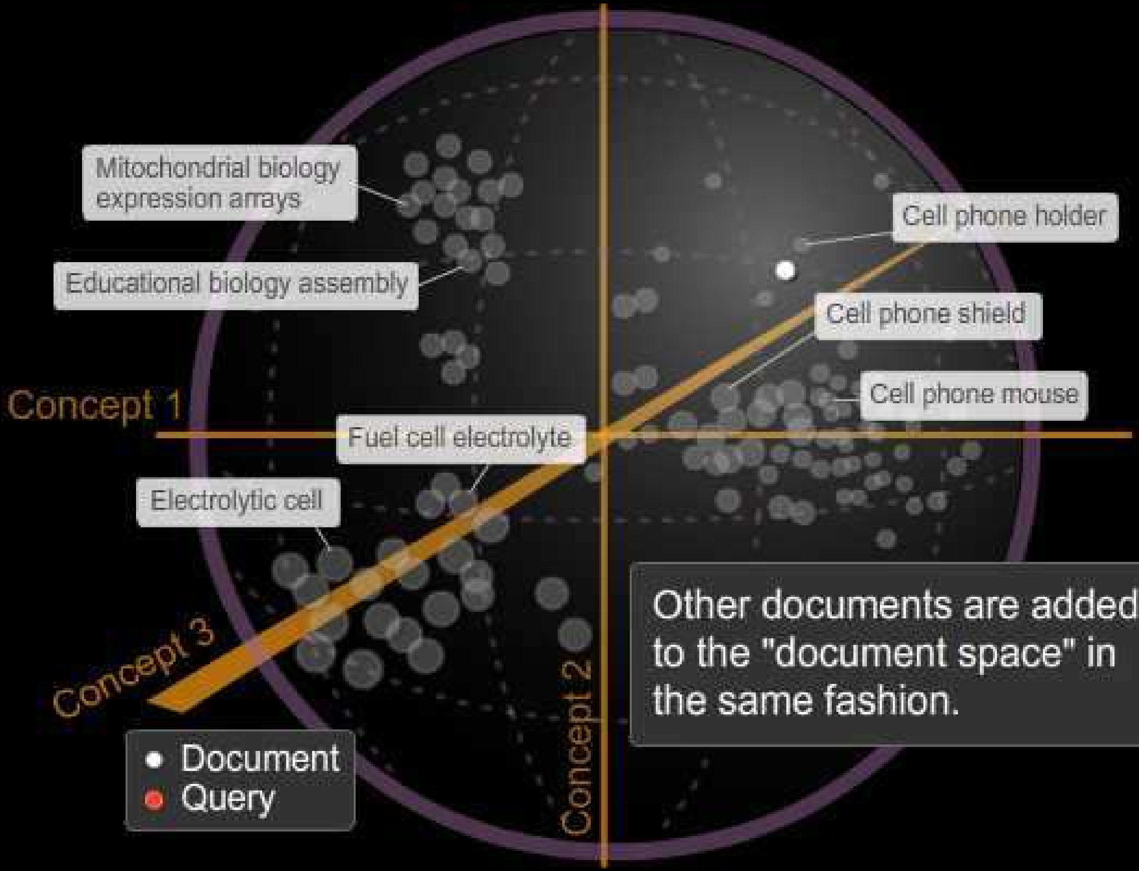


A document is placed into "document space" based on its relevance to each concept axis.

SUMOBRAIN
Big, powerful, smart searching

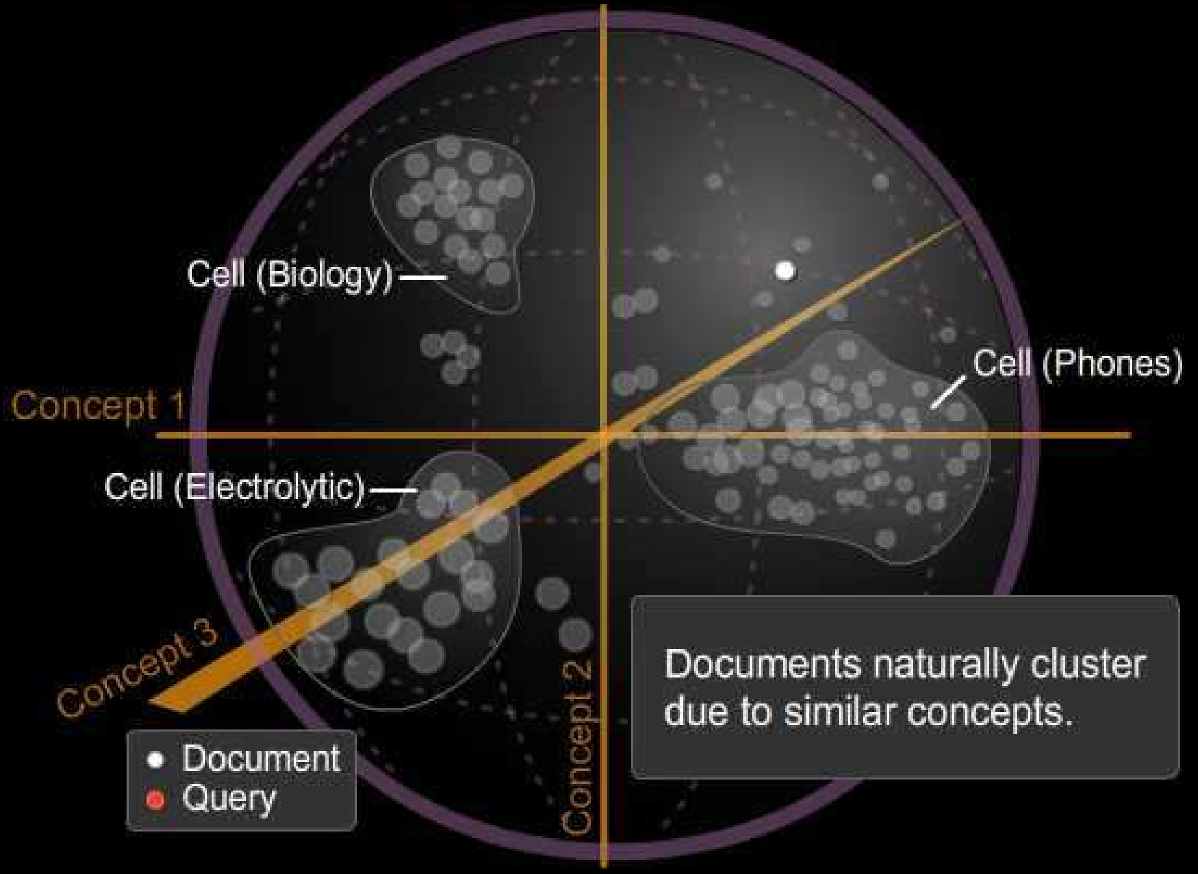
Next-Generation Patent Searching

Filling Up Document Space



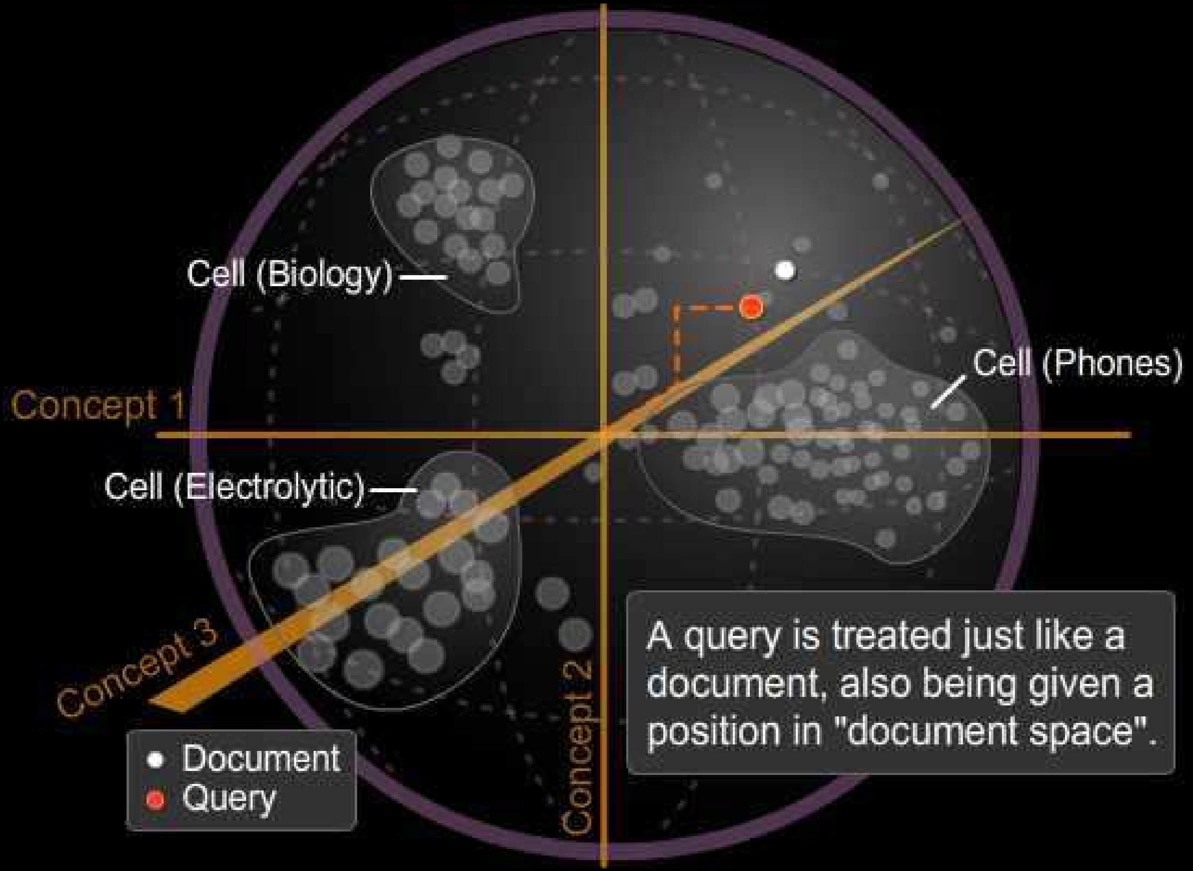
Other documents are added to the "document space" in the same fashion.

Spatial Clustering



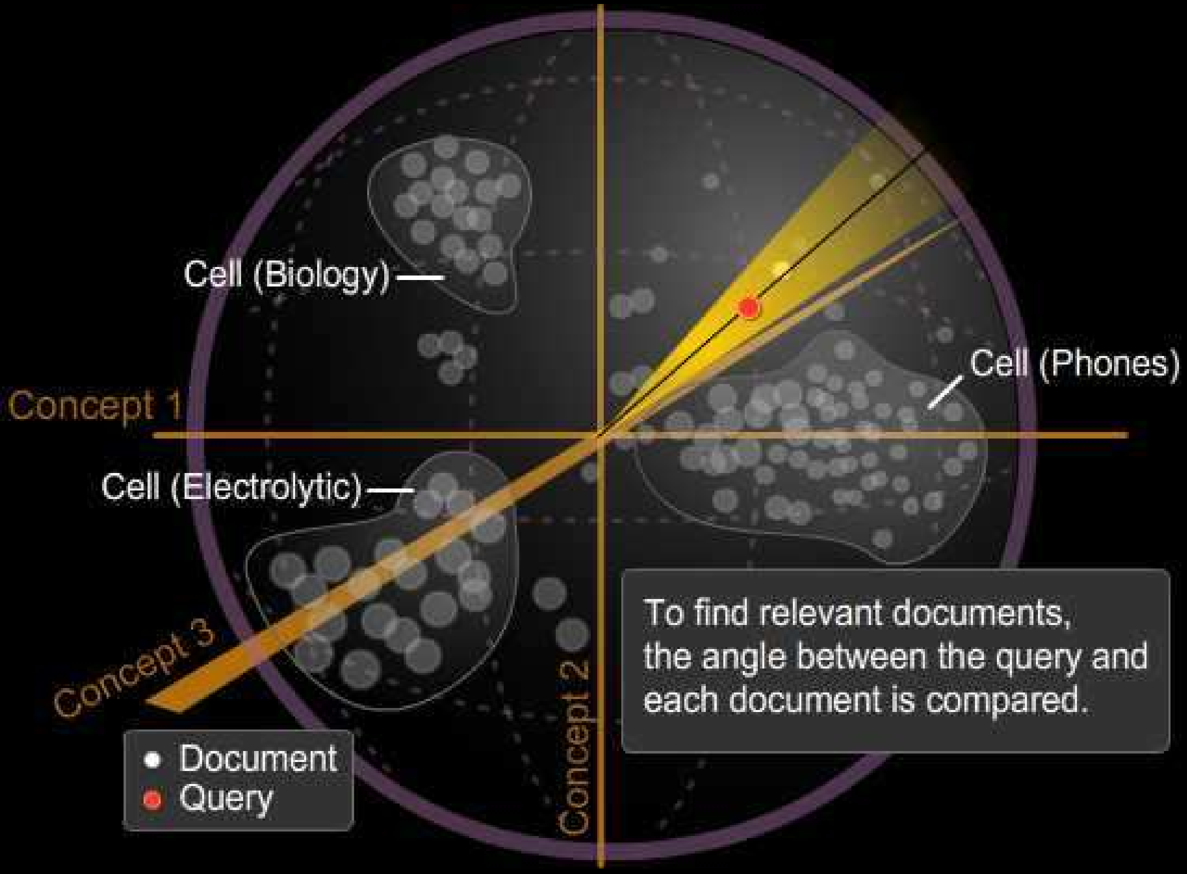
Documents naturally cluster due to similar concepts.

Queries in Document Space



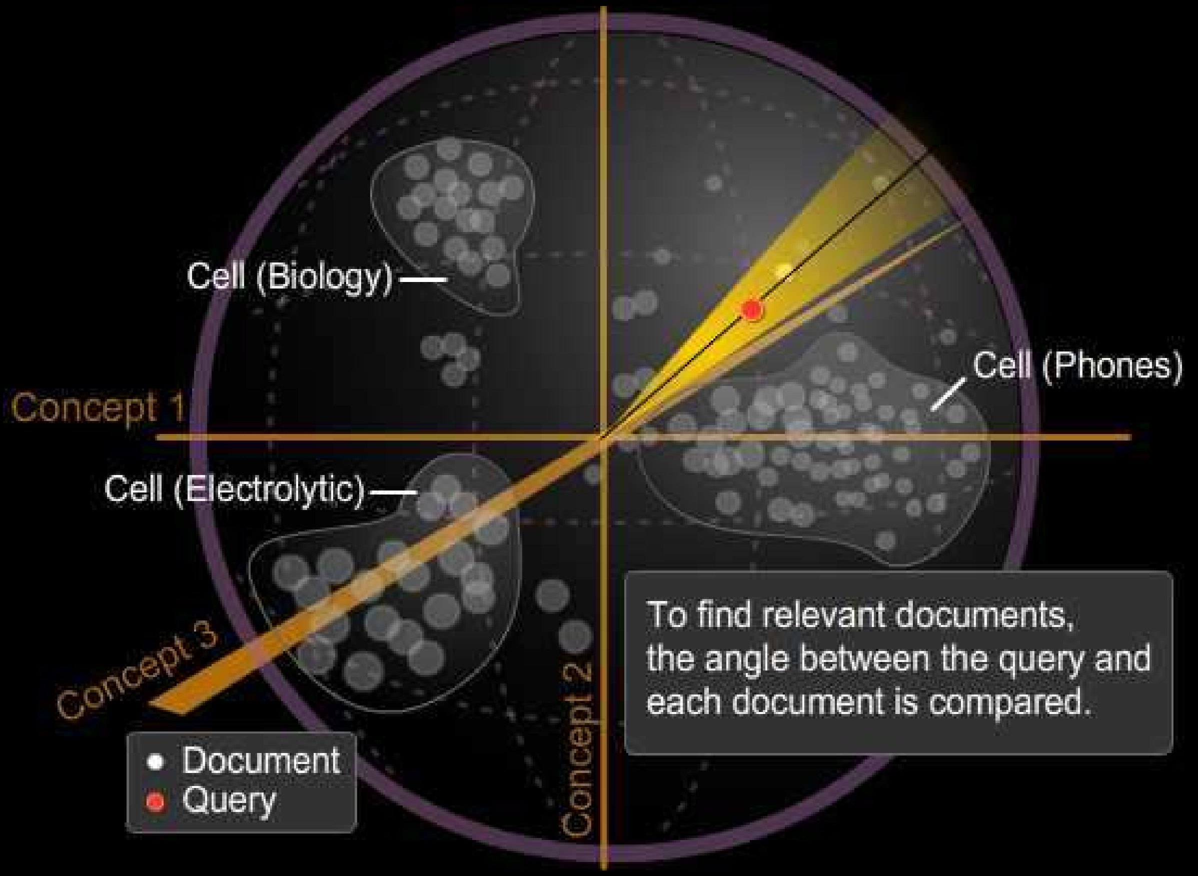
A query is treated just like a document, also being given a position in "document space".

Ranking Using Angles



To find relevant documents, the angle between the query and each document is compared.

Angles and Polysemous Terms



To find relevant documents, the angle between the query and each document is compared.

LSA Summary

- The system “understands” concepts, not just words
- 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.

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

- Mathematical and logical complexity
- Large computational demands
- *Lack of transparency: Can we get by this??*

Practical Applications

- Currently performing research on:
  - World Patent Literature
  - MedLine/PubMed database
  - SEC's EDGAR database (public company filings)
  - Cross-collection searching

Thanks!

Contact info:

james.ryley@sumobrain.com

www.SumoBrain.com