UNITED STATES
National Library of Medicine

Integrating Semantic Search, Federated Search and Biomedical Knowledge Base Technologies for Improved Services by Medical Libraries

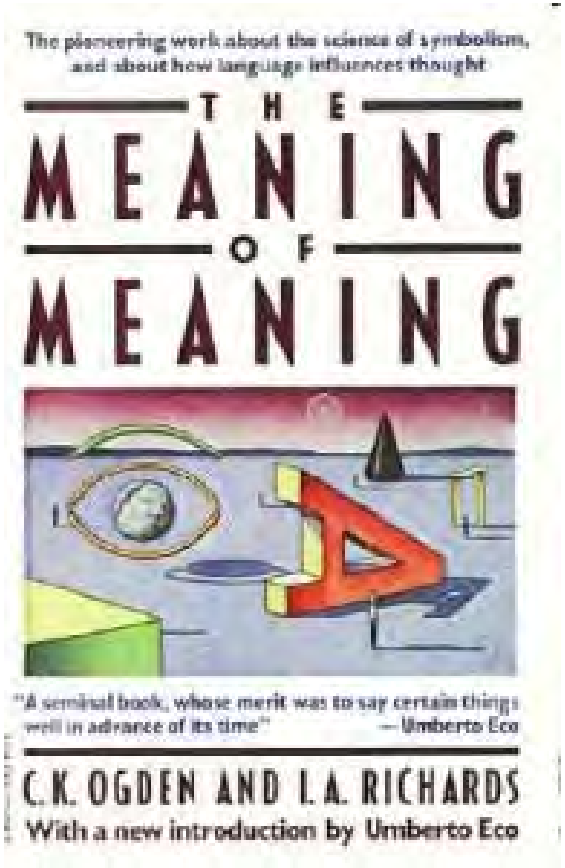
EAHIL 2010

Tamas Doszkocs, Ph.D.
Computer Scientist
doszkocs@nlm.nih.gov

“What is semantic search?”

or

The Meaning of Meaning



“He who knows does not speak, he who speaks does not know”

Lao Tse

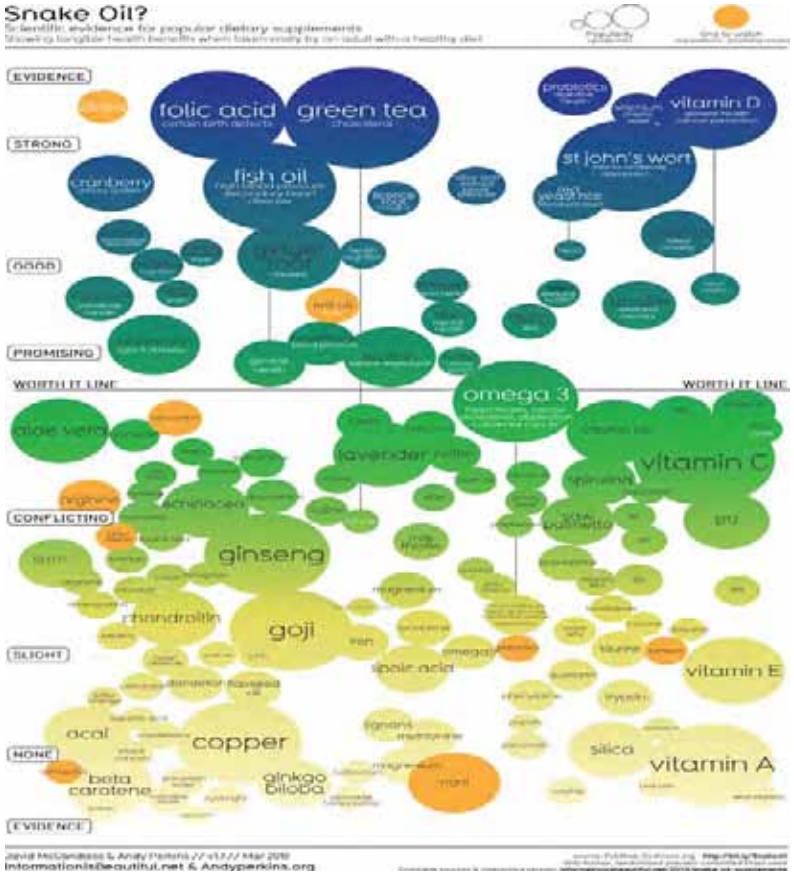
"It depends on what the meaning of the word 'is' is”

Bill Clinton

“the answer, my friend, is blowing in the wind”

Bob Dylan

What is semantic search?



"**semantic search** is **a search** or **a question** or **an action** that produces **meaningful results**, even when the retrieved items contain **none of the query terms**, or the search involves **no query text** at all "

(my definition)

scientific evidence for popular dietary supplements

Thinking of the Web
and
Semantic Search

"Wholly new forms of **encyclopedias** will appear, ready made with a **mesh** of **associative trails** running through them, ready to be dropped into the **MEMEX** and there **amplified**"

Vannevar Bush **"As We May Think"** Atlantic Magazine

JULY 1945

Trends in Searching 2009-2010

- ***The Web*** *is the* ***Memex***
- *Vertical Search*
- *Universal Search*
- *Discovery Search*
- *Social Search*
- *Real Time search*
- ***Semantic Search***
- *Semantic Web*

- *Connected Mobility*
- *Focus on Consumers*
- *Information Democracy*
  - *Social content*
  - *Social search*
  - *Social interaction*
- *Information Monopolies*
- *manipulating text, context, meaning, people*

Semantic Search Engines Mean Well

- *A long history …*
- *At a loss for words*
- *Ranking Results by Relevance*
- *Searching for Meaning*
- *Google and the Rest*

- *“Power to the People”*
- *“Pure” Semantic Search Engines*
- ***Specialized Semantic Search Engines***
- *Demos*

Semantic Searching: a long history

- *Libraries as Knowledge Bases*
- ***Librarians as Search Engines***
- *The Web as the Knowledge Base*
- ***Search Engines as Librarians***

Toward Meaning-Based Search

- ***Understanding content***
- ***Understanding context***
- *Understanding people*
- *Semantic Search vs. The Semantic Web*

- ***Meaningful results***
- *Web 1.0 linking pages*
- ***Web 2.0*** *linking content and people*
- ***Web 3.0*** *linking data and people and applications*
- ***Web 4.0*** *¿ inferences, conjectures, the pursuit of happiness ☺?*

Search Engines: at a loss for words

"In the beginning was the Word"

- *WebCrawler (1994)*
- *AltaVista (1996)*
- *InfoSeek (1996)*
- *"bag of words"*
- *Stopwords*
- *Boolean logic*
- *No linguistics*

- *Meaningless function words?*
  - *"to be or not to be"*
  - *What is it all about?*
  - *Why me?*
- *Ignoring syntax*
  - *"state police" vs.*
  - *"police state"*

Ranking Results by Relevance

"all animals are equal but some animals are more equal than others"

Science Citation Index
Eugene Garfield (1955, 1961)

Google Page Rank
Larry Page (1996)

Semantic Ranking
Colucci et al. (2006)

Searching for Meaning

the difference between “sciatica” the word and “sciatica” the pain

- ***understanding language***
- *polysemy*
- *synonymy*
- *intensions*
- *context*
- *disambiguation*
- *meaning*
- *personalization*

- ***Natural Language Processing and contextual understanding in searching***
- *Semantic Resources*
- *Metacrawler* (1994)
- ***WebLine*** (1994)
- *NorthernLight* (1996)
- ***TotalAccess*** (1999)
- ***AllPlus*** (2004)

Semantics in the Search Engines

- evolution by *trial and error*
- *incremental gains*
- *internal research*
- *personalization*
- *mobilty*
- *Interactivity*
- *acquisitions*
- *competition*
- ***Bingoogle***

- *Related Searches*
- *Categories*
- *Context*
- *Mashups*
- ***Answers***
- *Timelines*
- *“Something different”*
- *Visualizations*

Of the People
By the People
For the People

- ***web 2.0***
- *social content*
- *tagging*
- *folksonomies*
- *blogging*
- *tweeting*
- *chatroulette*

- ***Community***
- ***Creating***
- *Indexing*
- *Cataloging*
- *Searching*
- *Sharing*

“Pure” Semantic Search Engines

- ***semantics from the ground up***
- *understanding the Web*
  - *people and their intent*
  - *content and quality*
  - *data and attributes*
  - *entities and relationships*
  - *Context and meaning*
- *creating*
  - *actionable information*
  - *dynamic applications*

- *meaningful search results*
- ***better quality results***
- *better relevance*
- ***better presented results***
- ***better currency of results***
- *better tailored results*
- *better info streams*
- *better explore/discover*
- *better learning*
- *meaningful interactions*

Specialized Semantic Search Engines

- *domain knowledge*
- *matching people and needs*
- *improving diverse applications*
  - *real time news*
  - *finding jobs*
  - *recommendations for movies, music, goods, etc.*
  - *trend trackers*
  - *health*
  - *computable knowledge*
  - *semantic Facebook*

- ***Semantic resources and tools***
- ***Related searches/topics***
- ***Related items***
- ***Semantic mapping***
- ***Semantic synthesis***
- ***Semantic/linguistic annotations***
- ***Clustered search***
- ***Faceted search***
- ***Answers***

Semantic Search Engines for Health

- *Rich Semantic Resources in Biomedicine*
- *Small Number of Important Facets*
  - *Diseases and conditions*
  - *Causes*
  - *Tests and diagnosis*
  - *Treatments and procedures*
  - *Drugs and side effects*
  - *Alternative medicine*
  - *Life style and prevention*

- *Participatory Medicine*
- *E-patients*
- *first, second and third opinions* *on and off the Web*
  - *Health Professionals*
  - *Friends*
  - *Family*
  - *The "long tail" of health concerns*
- ***Informed Personal Health Decisions***

Trusted Health Information is vital for

- dealing with health problems,
- promoting healthy behavior,
- making healthy decisions and
- for overall well being
- However, as Mark Twain put it:
  Be careful of reading health books.
  You may die of a misprint!

examples of health search engines with Semantic Search Capabilities

- HealthLine
  - uses own taxonomy of > 250,000 health terms
  - thousands of Indian doctors and pharmacists
- http://www.everydayhealth.com/
  - Meta-data clusters
  - Topical clusters
  - Second most popular site after WebMD
- **http://righthealth.com/**
  - federated search engine
  - taxonomy of several million nodes
  - organized into a graph by
  - using a combination of human operators and algorithms
- http://MedStory.com
  - high-level categorizations or popular URLs
  - Purchased by Microsoft
  - http://health.msn.com

examples of health search engines with Semantic Search Capabilities

- http://www.semanticmedline.com/
  - Semantic NLP "understands" word and phrase meanings within context
- http://skr3.nlm.nih.gov/SemMedDemo/
  - Research prototype
  - summarizes MEDLINE citations returned by a PubMed search
  - Natural language processing is used to analyze salient content
- **http://healthbase.netbase.com/**
  - Based on language understanding
  - Surfaces facts, events, behaviors and connections among them"
  - combines classical keyword-based Web search with text-mining and ontologies
- http://healthmash.com/
  - Best-of-class semantic search engine
  - Powered by an automatically generated
    - Health Knowledge Base

the best health search engines

offer information that is

- **Reliable**
- **Relevant,**
- **Recent** and
- **Related meaningfully** to the search topic or user action

HealthMash

anatomy of a semantic health search and discovery engine

- **http://healthmash.com/** is an innovative next generation **semantic health search** engine, currently in public beta

- HealthMash developers have been working on NLM and NIH R & D projects for over 5 years
- HealthMash utilizes a pragmatic mix of **natural language processing tools**, semantic engineering techniques and **multiple knowledge sources**, including a proprietary **Health Knowledge Base**, to achieve both high precision and relevancy in its search results

HealthMash Demo

The **Explore and Discover** data comes from the **Health Knowledge Base**

- The ***Health Knowledge Base*** is automatically generated from trusted health content sites and diverse knowledge sources, such as **MeSH** and **UMLS**. HealthMash also utilizes the **Web** itself as a **data base**.
- The ***Health Knowledge Base*** contains explicit knowledge about
  - Health Concerns, Causes, Signs and Symptom, Tests, Procedures, Treatments, Drugs and Substances and Alternative, Complementary and Integrative Medicine
- The ***Health Knowledge Base*** is available via a web service API
- The ***Health Knowledge Base*** facilitates exploration and discovery

UMLS: Strengths, Weaknesses and Potential

- MeSH vs. UMLS Scale
- Missing concepts .
  - Aichmophobia; pca3 test
- Missing semantic types
  - Alternative and complementary medicine
- Missing NLM applications **):**
- Robust R&D: NLM Semantic MEDLINE

Semantic Knowledge Bases and Tools in HealthMash

- Semantic knowledge sources utilized by HealthMash
  - MESH (the Medical Subject Headings Thesaurus of the NLM) ,
  - UMLS (the Unified Medical Language System of NLM/NIH) and other semantic data repositories and
  - The Web as a knowledge base
  - **The Health Knowledge Base**
    - the **Health Knowledge Base** is the most important **semantic resource** in **HealthMash**
- Proprietary Natural Language Processing tools
  - Lexical/morphological and orthographic tools
  - Syntactic tools and
  - **Semantic tools**

behind the scenes in
Explore and Discover

- **Semantic Engineering**
- **Heuristics**
- Proprietary **Semantic Knowledge Base**
- Proprietary **Semantic Search Algorithms**
- Proprietary **Semantic Ranking**

HealthMash utilizes the following NLP resources

(these tools can also be licensed)

- **PolyDictionary** (English, medical and scientific dictionary)
- **PolySpell** (English, medical and scientific spell checkers)
- **PolyTagger** (part-of-speech tagger)
- **PolyPhraser** (phrase parser)
- **PolySearch** (intelligent concept search engine)
- **PolyCluster** (search result clustering engine)
- **PolyMeta** (federated search and discovery engine)

Putting it all together

- the **automatically generated** and automatically enhanced **Health Knowledge Base** is the key value-added semantic component of **HealthMash**

- the proprietary **semantic search**, **semantic processing** and **semantic ranking** technologies utilized in HealthMash contribute to **better search results**
- HealthMash can be integrated into any search engine via a **web service** (or **API**)

IN SUMMARY

HealthMash combines **vertical** semantic search of Trusted Health Information, **mashups** (Health News, Videos etc), Semantic Relations in context for Exploration and Discovery and **Table of Contents** and **Topic Clusters** for *drill down* in search results and *dynamic query modification*

Semantic Knowledge: HealthMash

Semantic Knowledge:
MedStory

Semantic Knowledge: Google Health

Search Trends

People who searched for **Warts** also searched for:

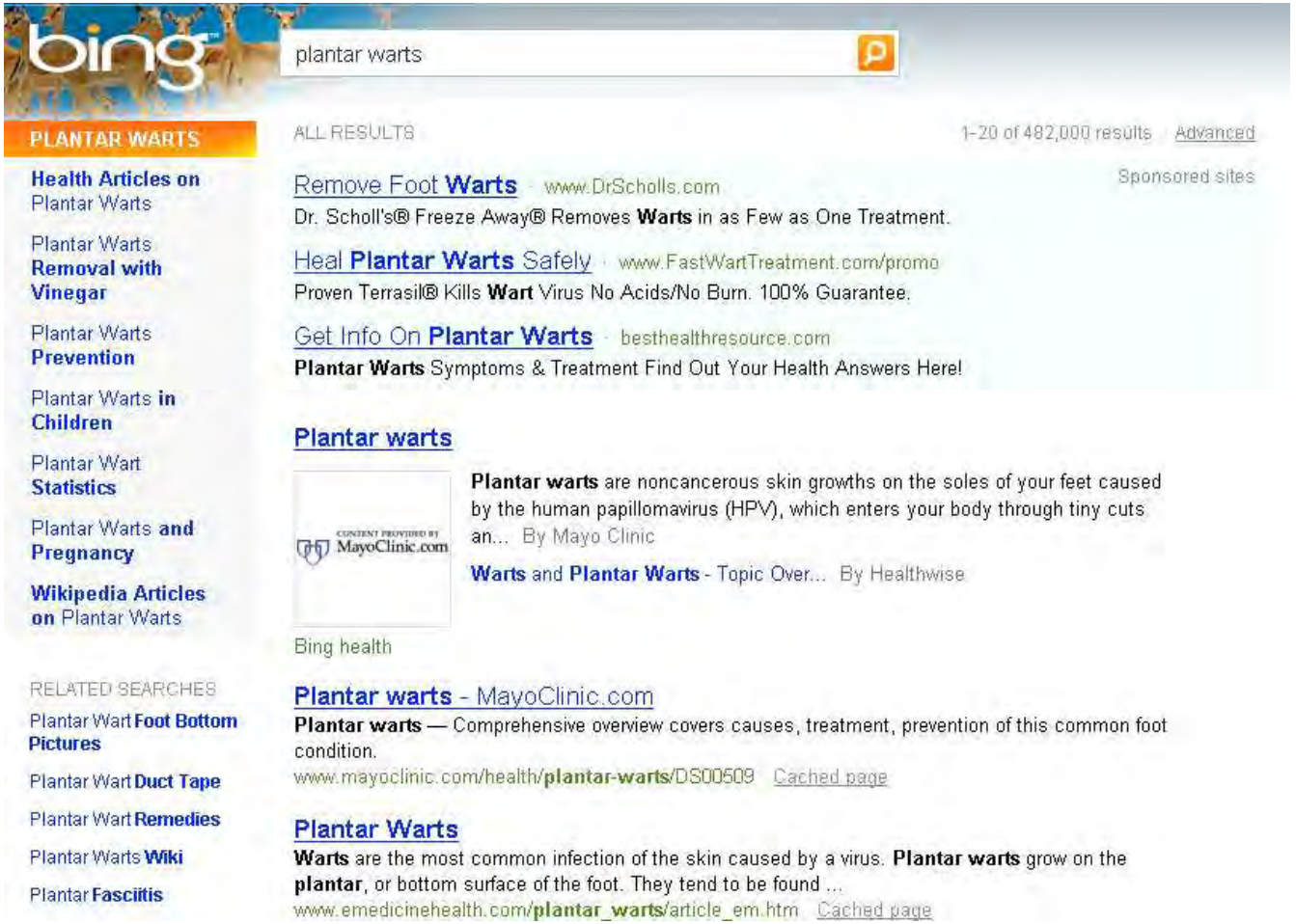
Conditions

1. genital warts
2. human papillomavirus infection
3. plantar warts
4. acute lymphocytic leukemia
5. pregnancy
6. moles
7. infection
8. acne

Symptoms

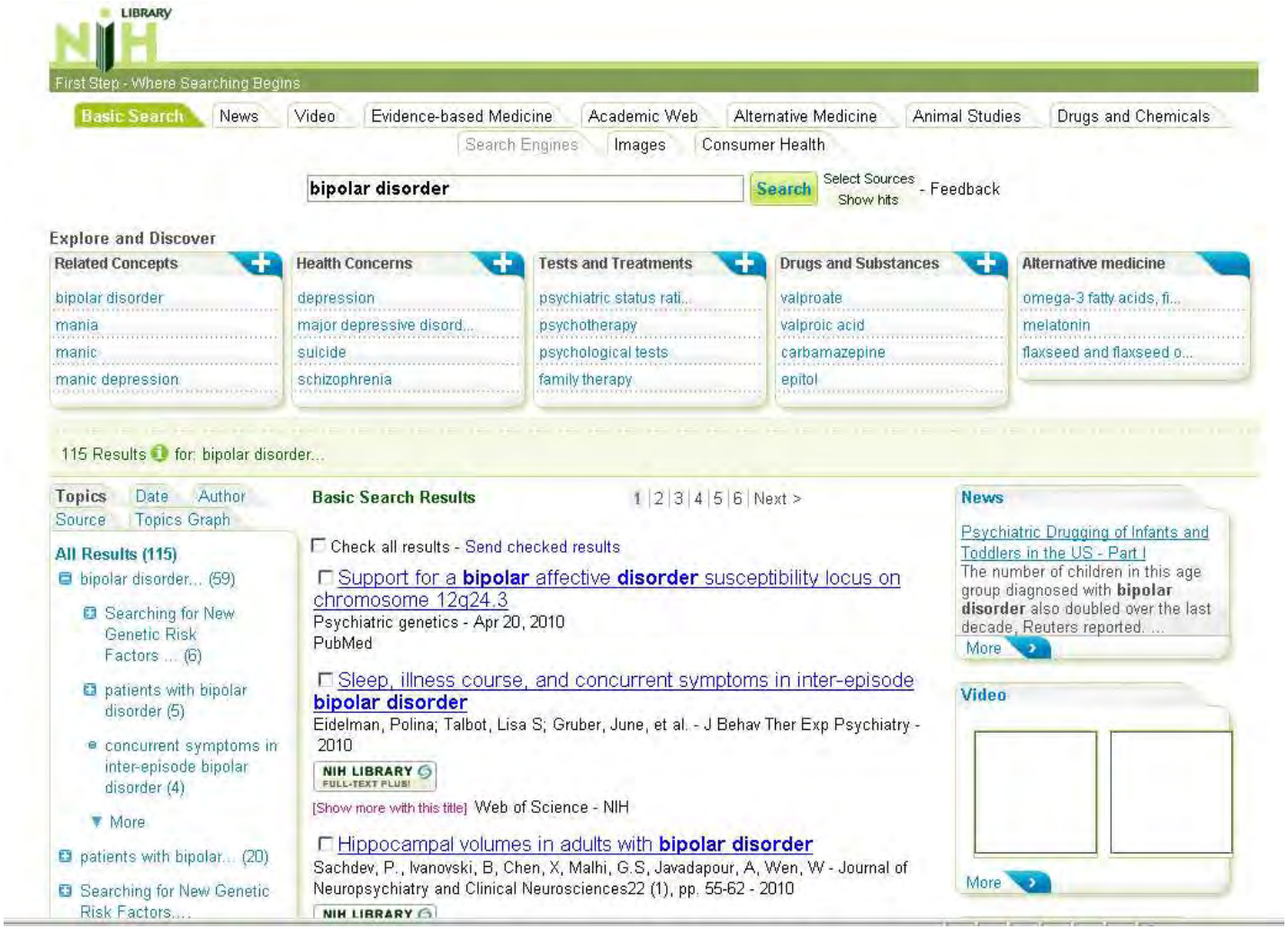
1. headache
2. weight loss
3. fever
4. cough
5. fatigue
6. back pain
7. sore throat
8. diarrhea

Semantic Knowledge: Bing



Integrating the Health Knowledge Base

via a **web service** into a search or federated search application



Distributing **content**, **search** and **librarians**

- From **ad hoc medicine** to **evidence based medicine** to **translational medicine** to **personalized medicine**
- From **point of need** to **point of care**
- The **NIH Clinical Center** WebLib Federated Search application
- The **Welsh Medical Library** at Johns Hopkins University

Next Steps for HealthMash

and the

Health Knowledge Base

- implement comprehensive semantic search
- expand the Health Knowledge Base
  - approximately 60,000 concepts in consumer health
  - millions of relationships in context
- into a Biomedical Knowledge Base
  - more 4,000,000 concepts
  - tens of millions of relationships **in context**
  - tens of millions of **semantic triples** and **n-tuples**
  - **Inferencing** and **question answering**
- from data to information to knowledge to wisdom
- leverage the wisdom of health professionals, the wisdom of patients, the wisdom of the masses and the wisdom of the ages

Improved health information services

- *Libraries as Knowledge Bases*
- ***Librarians as Search Engines***
- *The Web as the Knowledge Base*
- ***Search Engines as Librarians***
- ***Librarians** as **Personal Information Trainers** and **Team Coaches***

so **semantic search** **is** **what** **semantic technologies** can do **today**

but

what the …

is

semantics

?

Professor Irwin Corey explains

Professor Irwin Corey at the Cutting Room NYC

UNITED STATES
National Library of Medicine

Integrating Semantic Search, Federated Search and Biomedical Knowledge Base Technologies for Improved Services by Medical Libraries

EAHIL 2010

Tamas Doszkocs, Ph.D.
Computer Scientist
doszkocs@nlm.nih.gov