ACCESS2010 Applications and Acceleration Track

# Cognitive Space and Information Space

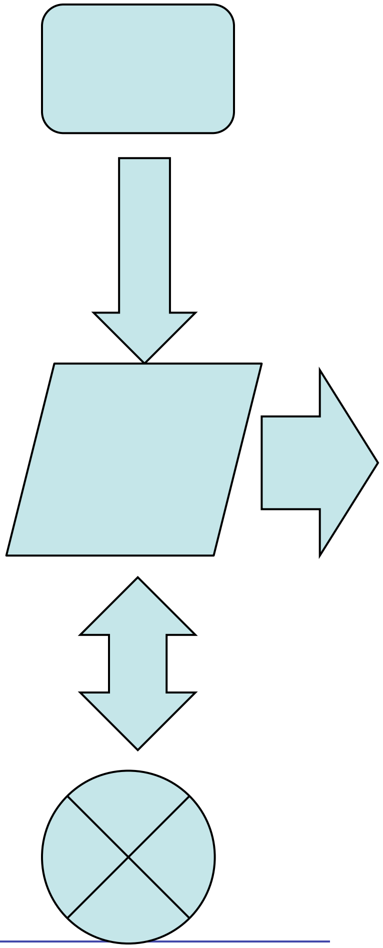
Gregory B. Newby

Arctic Region Supercomputing Center

newby@arsc.edu

# Presentation Outline

- Abstract ("for the record")
- About your speaker's institution, ARSC
- High level overview of where we are going: engineering progress towards a more data-intensive and personalized world
- High level conceptual and scientific challenges
- Some practical ongoing efforts
  - Information retrieval research
  - Large-scale eigensystems
- Future plans



# Abstract

Computational clusters in excess of 100TF have begun to approach the estimated computational capacity of the human brain.  Such systems theoretically could, with the right software, emulate human cognitive processes such as thought and memory.  In this talk, we will look at the specific topic of information search and retrieval, in which similarity among items may be considered as a fundamental human cognitive phenomenon.  From the perspective of data intensive computing systems, we would like to have information spaces (of systems) that are consistent with cognitive spaces (of humans).  In such systems, searching for information with a system would be akin to a human remembering that information.  However, modern document retrieval systems take shortcuts that greatly reduce needed computation, but at the expense of consonance between information spaces and cognitive spaces.  These include shortcuts on data representation (such as Boolean matching of search terms to documents) and shortcuts on the human information seeking experience (by not considering prior knowledge or temporal aspects of search).  This talk will look at challenges and benefits of enhanced scale in size, throughput, representation, and user customization that, when applied, greatly expand data intensive aspects of information search and retrieval.  We envision information systems of the future where system information spaces are modeled more closely after human cognitive spaces.

# Arctic Region Supercomputing Center

**Helping scientists to seek understanding of our past, present and future by applying computational technology to advance discovery, analysis and prediction**

**ARSC's state of the technology supercomputers provide 24/7 accessibility for high-performance computing needs**

**ARSC's storage infrastructure provides redundant back-up, swift access and retrieval of archival data**

HPC DoD Modernization Program SOLVING THE HARD PROBLEMS

UAF UNIVERSITY OF ALASKA FAIRBANKS

# What happens…

## When these…

## Are part of these?

# What happens...

When this...

Is directly connected to this...

# Answer: Human intellect is augmented by machine

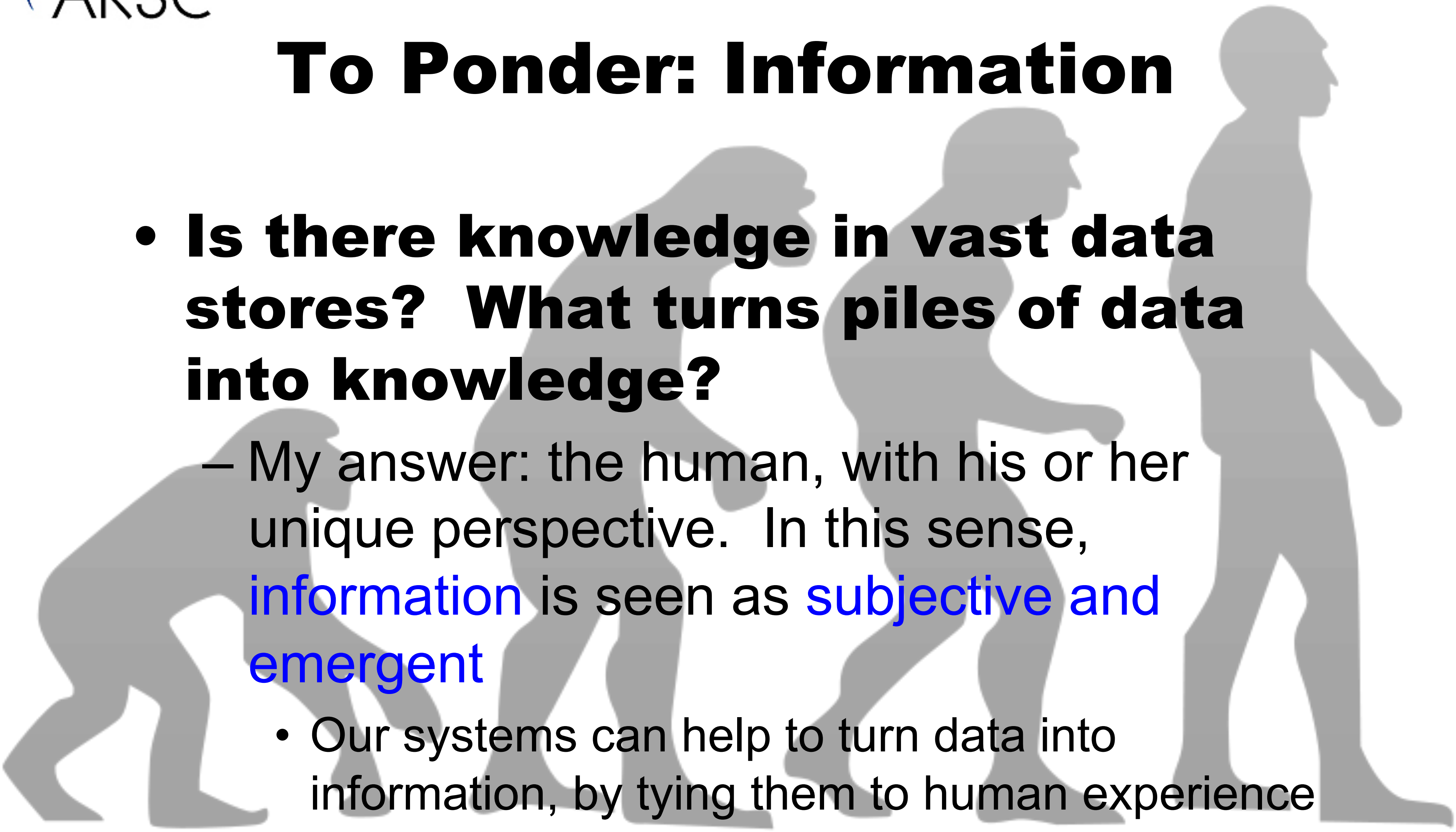
- … With **personalized** and **contextual** data
- … With the ability to store and record our life experiences
  - Location-based (GIS ties), temporally connected
  - Audio/video experiences
  - Our information experiences: search history, information encountered

# Sidebar: Recording human experience

- Audio/Visual bandwidth: TB/day
- Location tracking: where we are; tying experience to location in time and space
- Life's information experiences
  - Content we encounter through everyday interaction
  - Documents viewed
  - Our conversations, notes, calendars, etc. – already being digitized, but not yet integrated well with external data sources
- This engineering challenge quickly becomes an information storage and retrieval problem

Steve Mann

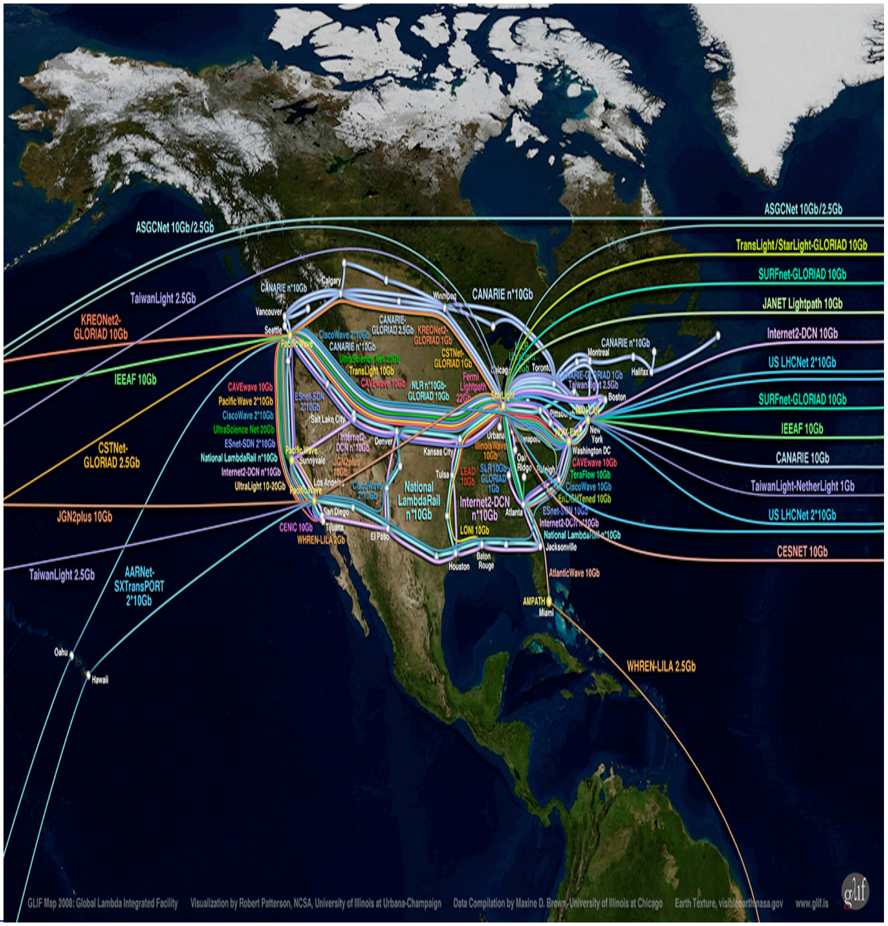
# To Ponder: Information



- **Is there knowledge in vast data stores? What turns piles of data into knowledge?**
  - My answer: the human, with his or her unique perspective. In this sense, information is seen as subjective and emergent
    - Our systems can help to turn data into information, by tying them to human experience

# The good news

- To make all this happen,
  - Scientists don’t need to invent artificial intelligence
  - We don’t need to deploy infrastructure: it’s already there (telephone & networking)
  - We don’t need to miniaturize devices (that’s happening already)
  - There are commercial and social forces at work, heading in a useful direction for us



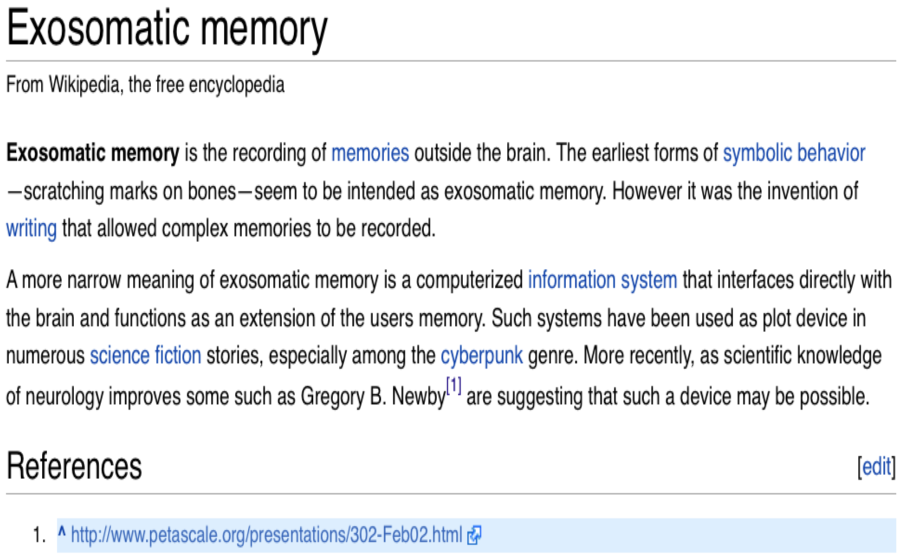
GLIF Map 2008: Global Lambda Integrated Facility   Visualization by Robert Patterson, NCSA, University of Illinois at Urbana-Champaign   Data Compilation by Maxine D. Brown, University of Illinois at Chicago   Earth Texture, visibleearth.nasa.gov   www.glif.is

# Building blocks

- **We already have conceptual frameworks for augmenting human intelligence by machine**

**Some examples,**

- "Alternative neuromorphic computing architectures," "brain state in a box" – with thanks to Richard Linderman of AFRL
- Minsky's "Society of Mind," which talks about mind as emergent, from many small parts. Very interested in commonsense knowledge (i.e., data)



## Exosomatic memory

From Wikipedia, the free encyclopedia

**Exosomatic memory** is the recording of memories outside the brain. The earliest forms of symbolic behavior —scratching marks on bones—seem to be intended as exosomatic memory. However it was the invention of writing that allowed complex memories to be recorded.

A more narrow meaning of exosomatic memory is a computerized information system that interfaces directly with the brain and functions as an extension of the users memory. Such systems have been used as plot device in numerous science fiction stories, especially among the cyberpunk genre. More recently, as scientific knowledge of neurology improves some such as Gregory B. Newby[1] are suggesting that such a device may be possible.

## References

[edit]

1. ^ http://www.petascale.org/presentations/302-Feb02.html

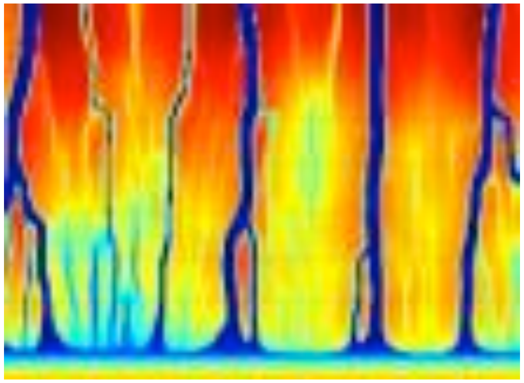
# Transition: Efforts by Your Humble Narrator

- Topic area: **Information retrieval** (i.e., Web search). Large datasets, natural language
- **Research goal: How can we make information retrieval so effective, it's like remembering things you didn't know?**
- Part of my approach is to look at statistical ways we can have "information space" of systems match the "cognitive space" of human information seekers

# Recent Results

- ARSC systems used for data-intensive computing, and for large-scale eigensystems. These eigensystems are a numerical approach to build cognitive-like spaces from raw data input
  - Data-intensive: 25TB dataset, 1.2G Web documents. Work with MySQL, Tomcat, Lucene. This is mostly for testing information retrieval algorithms and “divide and conquer” federated search
  - Eigensystems: Using PETSc/SLEPc to look at large-scale relationships among terms and document

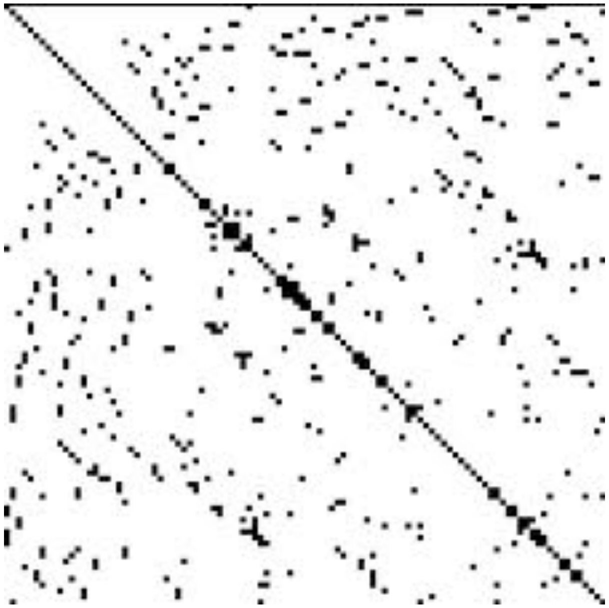
Lucene logo



PETSc logo

# Computational Methods

- Software used: Mostly custom. But based on toolkits/libraries. For eigensystems, lots of C/C++ (PETSc). For search, Java (Tomcat, Lucene), HTML, PHP, Perl; some MATLAB
- Computational performance: PETSc scaling beyond 256 cores. Large sparse matrices (over 100Kx2M, various submatrices)



sparse matrix

# Our Application Area: Text Retrieval

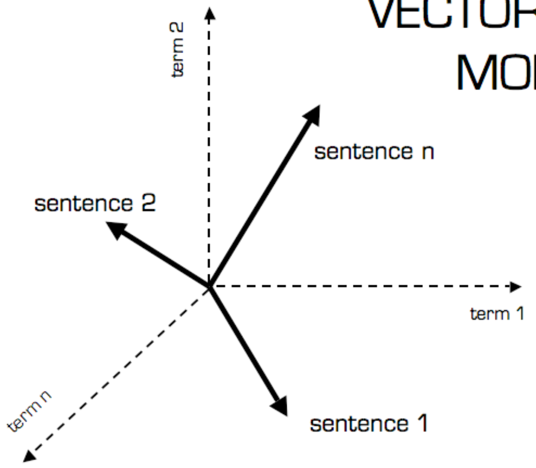
- **We work with large sparse matrices (some symmetric, some rectangular)**
- **Representing collections of textual documents**
  - In our matrices, rows correspond to unique terms in a collection of documents. Each document is represented by a column. Thus, column cells are > 0 for terms that occur in that document and zero otherwise
  - Since most terms do not occur in most documents, the matrices are sparse
  - By solving eigensystems for the matrices, we are able to identify statistical relationships among terms (words) and documents

Text REtrieval Conference

# Text Retrieval Goals for use of Eigensystems

- Assess term-term similarity. Useful for identifying synonyms (for search term expansion); helpful for recommending related terms
- Assess document-document similarity. Useful for "more like these" queries
- Assess query-document similarity to generate and rank search results. An alternative to Boolean search sets (though much more computationally expensive)
  - All these techniques are said to supplement (for example) Google's methods, and have been found in various commercial and experimental text retrieval systems
  - However, they are computationally expensive

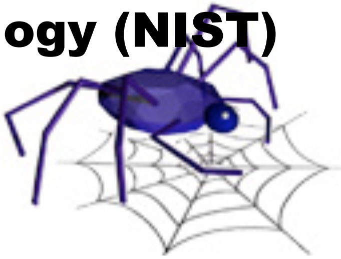
VECTOR SPACE MODEL



The eigenvectors are dense matrices, versus sparse term-by-document matrices. This reduces the utility of computational shortcuts used for simple Boolean query-document matching

# Our Matrices, the Data

- **Derived mainly from the "gov2" test collection of the National Institute of Standards and Technology (NIST) Text REtrieval Conference (TREC)**
  - 2 million Web documents from .gov
  - 20GB collection size



- **We use the terms (i.e., vocabulary) from Google Gigaword collection of n-grams (1, 2, 3-grams; 4 and 5-grams are also available)**
  - Downselected to 645271 single-word terms (from 13.5M candidates)
  - Gigaword vocabulary is available from the Linguistic Data Consortium (LDC2006T13)
  - Over 1T tokens (words, phrases, sentences) total (this is very helpful for searching for multi-word queries)

# Our Matrices, the Numbers

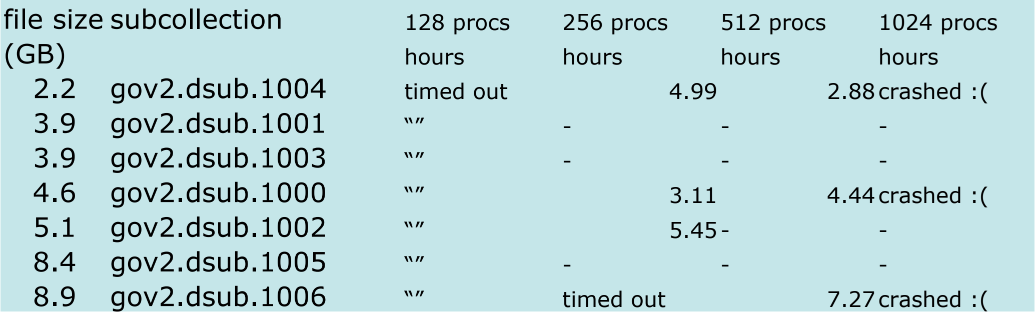
- **From 2M documents, we built 996 subcollections based on hostname**
- **For today's talk, we looked at some of the largest subcollections**
  - We can combine subcollections, too

| file size (GB) | subcollection | nnz (non-zeros) | # columns (documents) |
|---|---|---|---|
| 2.2 | gov2.dsub.1004 | 193,537,001 | 792,735 |
| 3.9 | gov2.dsub.1001 | 341,192,859 | 1,076,476 |
| 3.9 | gov2.dsub.1003 | 339,858,097 | 1,349,402 |
| 4.6 | gov2.dsub.1000 | 408,094,751 | 835,162 |
| 5.1 | gov2.dsub.1002 | 451,306,104 | 1,449,770 |
| 8.4 | gov2.dsub.1005 | 749,287,610 | 1,911,638 |
| 8.9 | gov2.dsub.1006 | 791,164,201 | 2,799,172 |

.94 correlation
NNZ with # cols

# Large Eigensystems: Run Times (hours)

- **All runs on ARSC-DSRC’s *pingo*, a 3456-core Cray XT5**



| file size (GB) | subcollection | 128 procs hours | 256 procs hours | 512 procs hours | 1024 procs hours |
|---|---|---|---|---|---|
| 2.2 | gov2.dsub.1004 | timed out | 4.99 | 2.88 | crashed :( |
| 3.9 | gov2.dsub.1001 | “” | - | - | - |
| 3.9 | gov2.dsub.1003 | “” | - | - | - |
| 4.6 | gov2.dsub.1000 | “” | 3.11 | 4.44 | crashed :( |
| 5.1 | gov2.dsub.1002 | “” | 5.45 | - | - |
| 8.4 | gov2.dsub.1005 | “” | - | - | - |
| 8.9 | gov2.dsub.1006 | “” | timed out | 7.27 | crashed :( |

  - – These matrices are big enough to *require* over 128 processors to complete an eigensystem in 8 hours

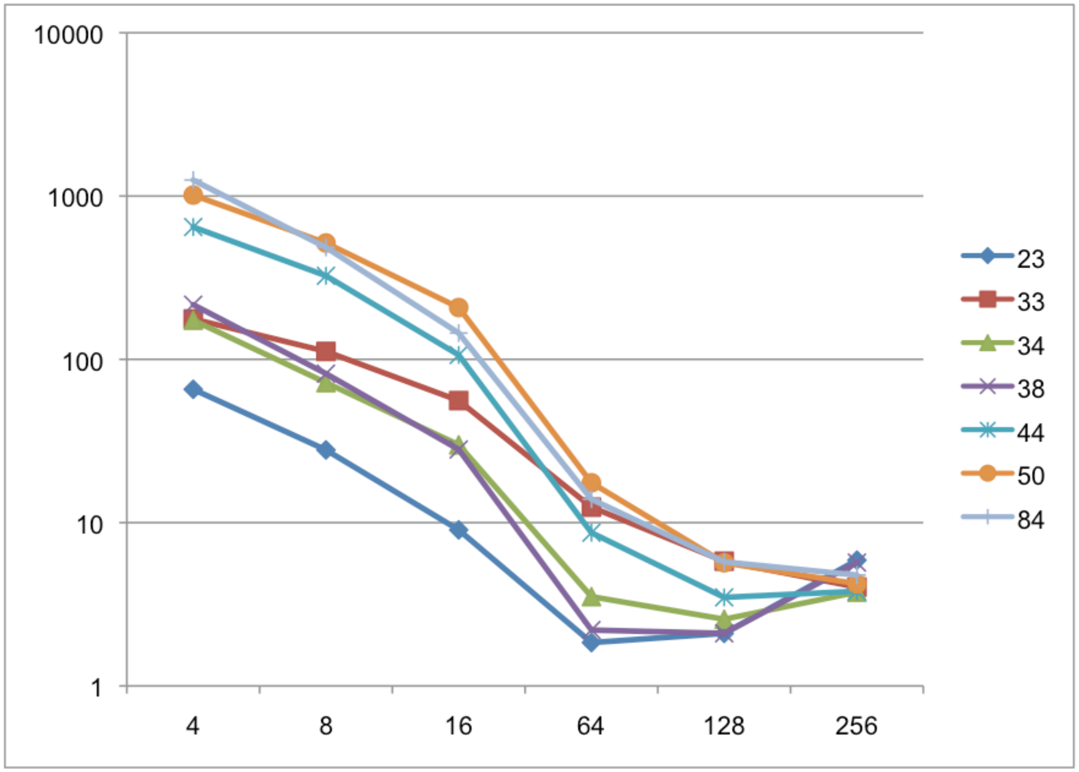
# Smaller Eigensystems: Run Times (Values)

- **These were far smaller subcollection matrices we ran to get timing curves. No problems (MPI or otherwise)**

| | Matrix Size (mb) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| # Processors | 23 | 33 | 34 | 38 | 44 | 50 | 84 | 1700 |
| 4 | 65.539 | 176.715 | 173.555 | 216.100 | 645.971 | 1013.416 | 1252.171 | |
| 8 | 27.870 | 111.796 | 71.965 | 81.732 | 324.939 | 515.944 | 484.852 | |
| 16 | 9.028 | 55.896 | 29.956 | 28.049 | 106.32 | 207.523 | 144.804 | |
| 64 | 1.843 | 12.441 | 3.523 | 2.200 | 8.678 | 17.539 | 13.912 | 4065.824 |
| 128 | 2.098 | 5.805 | 2.562 | 2.100 | 3.487 | 5.679 | 5.747 | 1006.609 |
| 256 | 5.890 | 4.028 | 3.751 | 5.694 | 3.801 | 4.222 | 4.773 | 271.789 |

- **Conclusion: PETSc/SLEPc scales linearly to 64 processors for these small matrices, but very short run times at large processor counts yield flattened performance curves**

# Smaller Eigensystems: Run Times



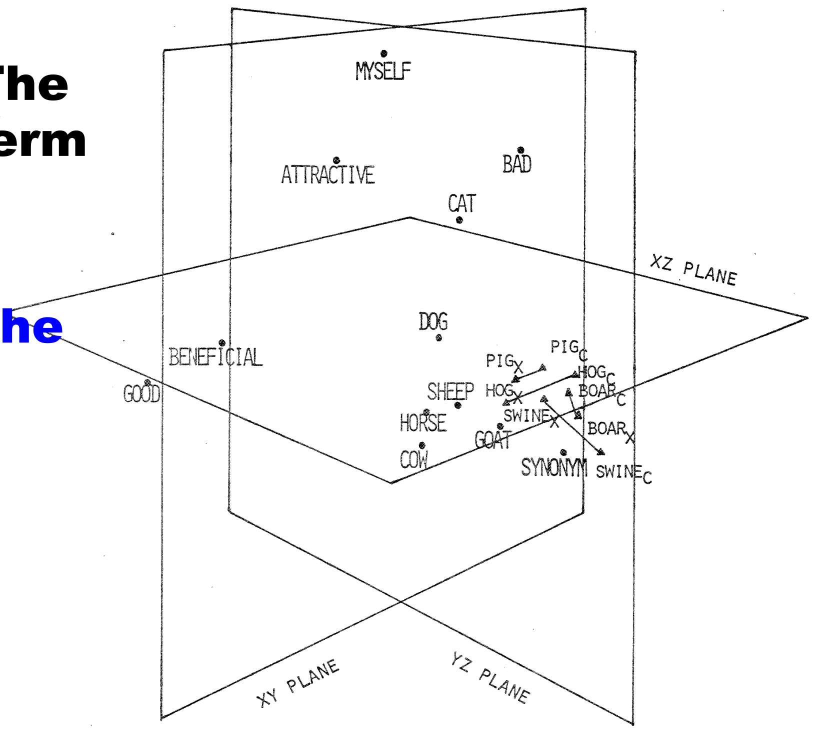
**X=nprocs, y=log(cpu time in seconds), Lines are matrix size in MB**

# Visualizing Text Relations: Historical

- **Joseph Woelfel and colleagues measured concept relations using paired-comparison surveys given to humans. The output was akin to a term-term matrix**
- **In this example, note that synonyms are not actually the same (pigs, boar, hogs and swine)**
- **Data collection was labor intensive; analysis was essentially the same as our eigensystems analysis**

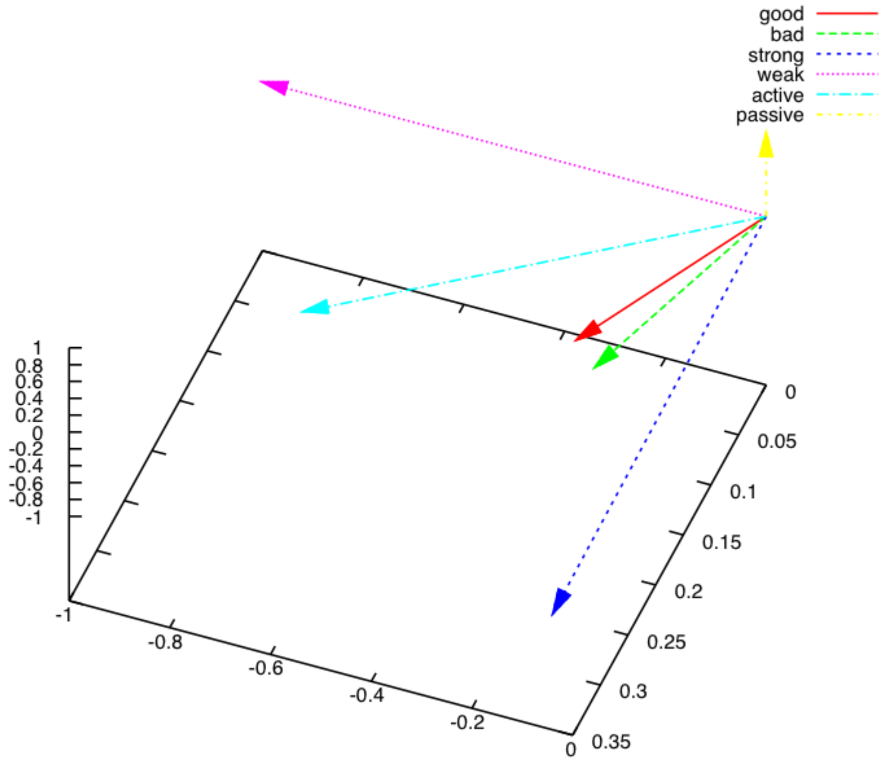
FIGURE 1

3 DIMENSIONAL PLOT OF BARNYARD ANIMAL DOMAIN



# Visualizing Text Relations: Current Results

- After Charles Osgood's (1957) "semantic differential" scale for human concept space, we looked at relations among: good, bad, active, passive, strong, weak
- Plot shows the 1st three eigenvectors for subcollection 1322
- Indeed, active/passive and strong/weak appear nearly orthogonal, in this space



# Future Plans: Computational

- **Continue efforts towards larger matrices with larger processor counts**
  - Once these eigensystems are computed, they can be efficiently represented for text retrieval. Computing eigensystems for very large collections will enable further R&D
  - How long until these content refactorizations and integrations are happening on our own personal devices, with streamed live contextual data?

# Future Plans: Conceptual

- **Integrate "cognitive space" approaches, with the goal of presenting document collections as elements of exosomatic memory**
  - Potentially relevant for intelligence community and defense environments.
  - Semantic Web concepts and efforts will help.
- **Keep working on content and personalization, through Project Gutenberg and other forums**

# “Thanks” to many

- In **ARSC** and **UAF**
  - Don Bahls and other consultants, co-workers
  - Kylie McCormick, multi-summer intern
  - Science staff & students, including Chris Fallen, John Styers, many others
- **HPCMP**/DoD scientists, users & support
- **ACCESS10** & **NCSA**: Looking forward to collaborations and discussions