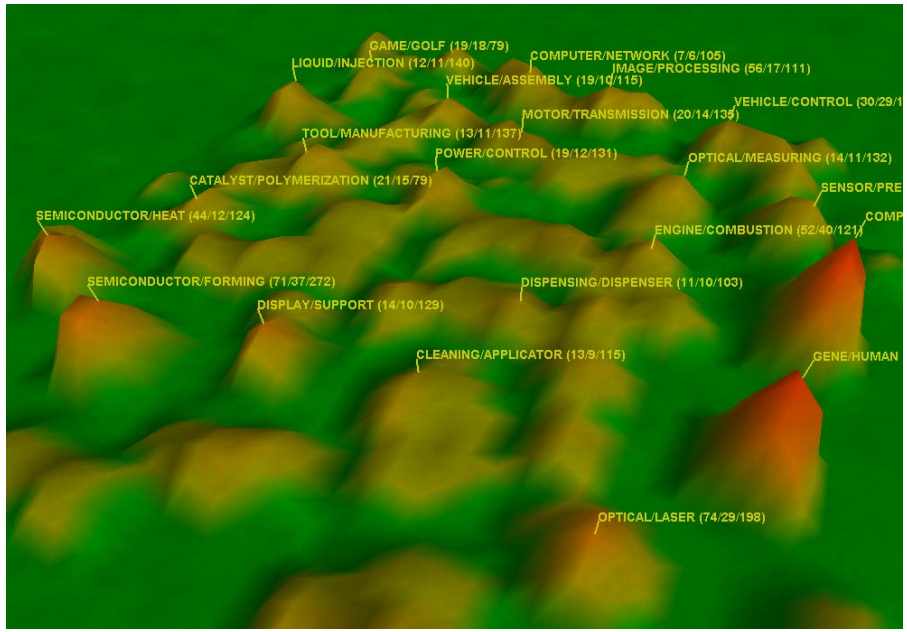
Applications of Data Mining and Information Visualization at SNL

Kevin W. Boyack

Sandia National Laboratories

SIAM International Data Mining Conference
April 23, 2004



Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000.

Data Disclaimer

- Lots of different data sources
- Lots of different format
- Lots of missing or dirty data
- Different levels of detail, styles, writers, objectives
- Different data types (text, signal, numeric, multimedia)
- Everyone has the similar issues, but different applications

Tool Disclaimer

- Lots of tools available
  - Commercial
    - Extraction, indexing, retrieval, portal, personal
    - Visualization
  - Home-grown, particularly in bioinformatics
    - Mostly algorithmic for a single purpose
  - Lots of one-off approaches, no silver bullet
- SNL uses some of our own, some commercial for different projects
  - Use fragmented within company

Outline

- Knowledge domain visualization
  - Overview of process
  - Recent PNAS special issue
- Overview of applications
  - Each has different gotchas

Domain Visualization

- Domain
  - Scientific or technical field
  - Corporate database (projects, requirements, IP ...)
    - Both of the above imply a library of documents
  - Genome
  - Network (social, email, protein ...)
- Visualization
  - Organized visual picture of the domain with both context and content
  - Best ones support interaction and retrieval

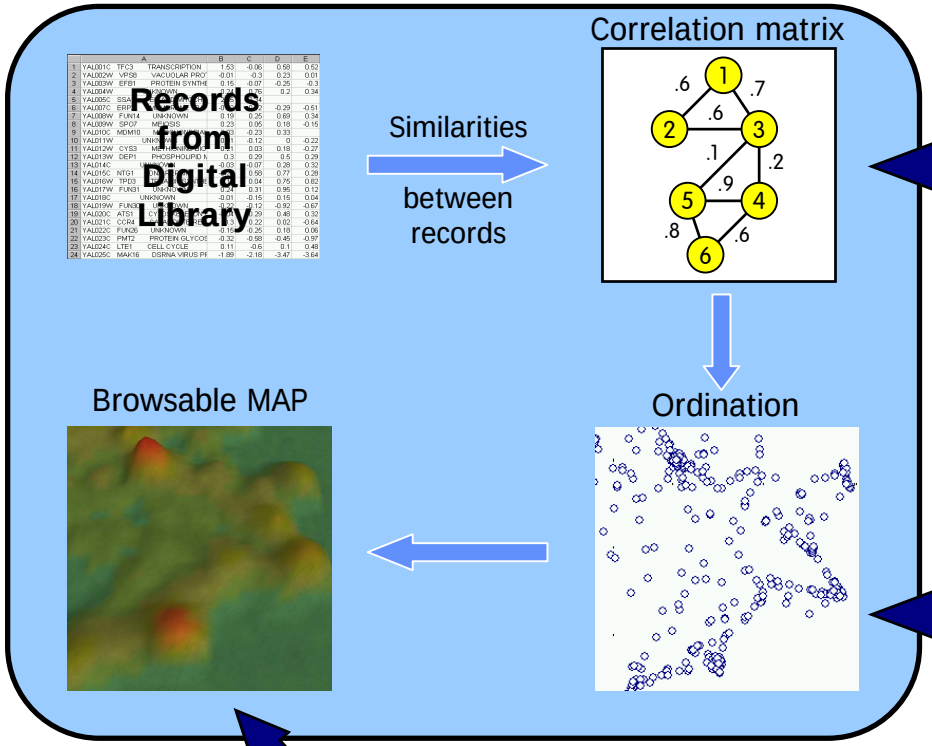
Why Use Domain Visualization?

- Reduces information overload
  - Ensure better coverage of related material
    - Less refined initial searches are sufficient
  - Takes advantage of human cognition and visual perception
    - Effectively screen out unrelated material
- Useful for experts and non-experts
  - Expert – validate paradigm, easy access, surprises
  - Non-expert – quick overview of a field

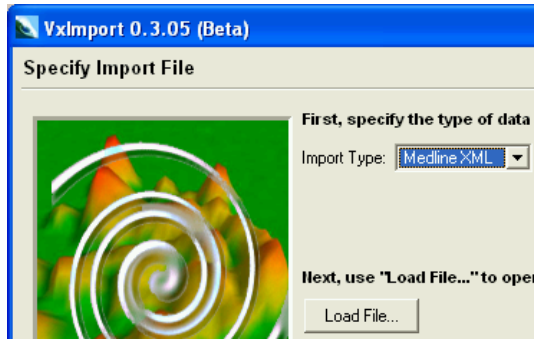
Literature Domain Analyses Uses

		QUESTIONS RELATED TO			
		Fields and paradigms	**Communities and networks**	**Research performance or competitive advantage**	**Commonly used algorithms**
UNIT OF ANALYSIS	Authors		Social structure, intellectual structure, some dynamics	Use network characteristics as indicators	Social network packages, MDS, factor analysis, Pathfinder networks
	Documents	Field structure, dynamics, paradigm development		Use field mapping with indicators	Co-citation, co-term, vector space, LSA, PCA, various clustering methods
	Journals	Science structure, dynamics, classification, diffusion between fields			Co-citation, intercitation
	Words		Cognitive structure, dynamics		Vector space, LSA, LDA (20)
	Indicators and metrics			Comparisons of fields, institutions, countries, etc., input-output	Counts, correlations

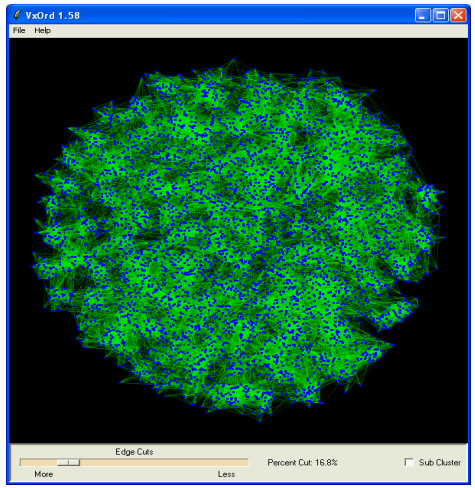
Example Process with SNL Tools



VxImport



VxOrd



Applications

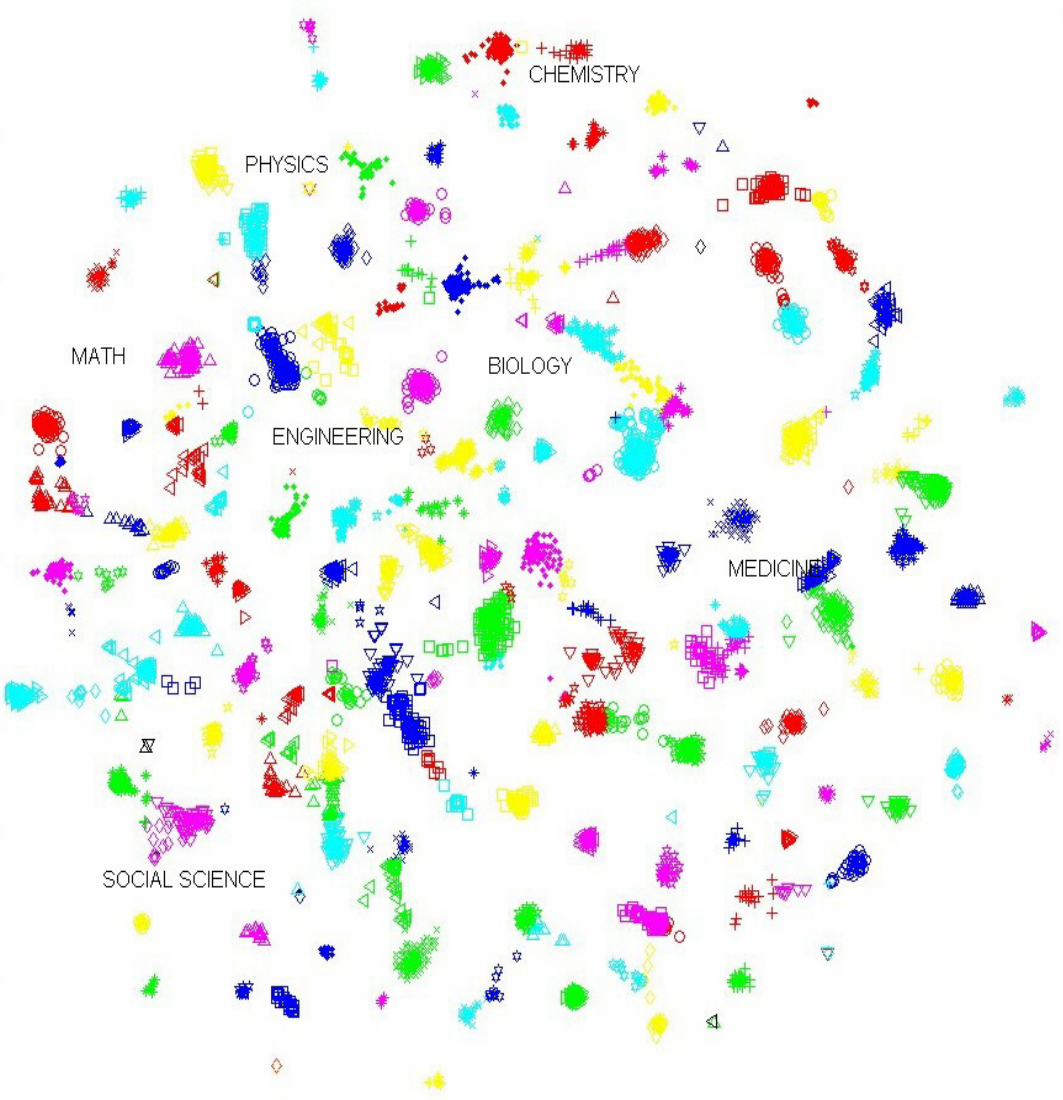
- Research indicators
- Corporate information
- Bioinformatics, networks

Research Indicators

- Models of science
  - Identify recent relevant high impact work
  - Identify possible collaborators
  - Focus new ideas (proposal writing)
  - Evaluate internal work
  - Test theories of scientific progress and innovation

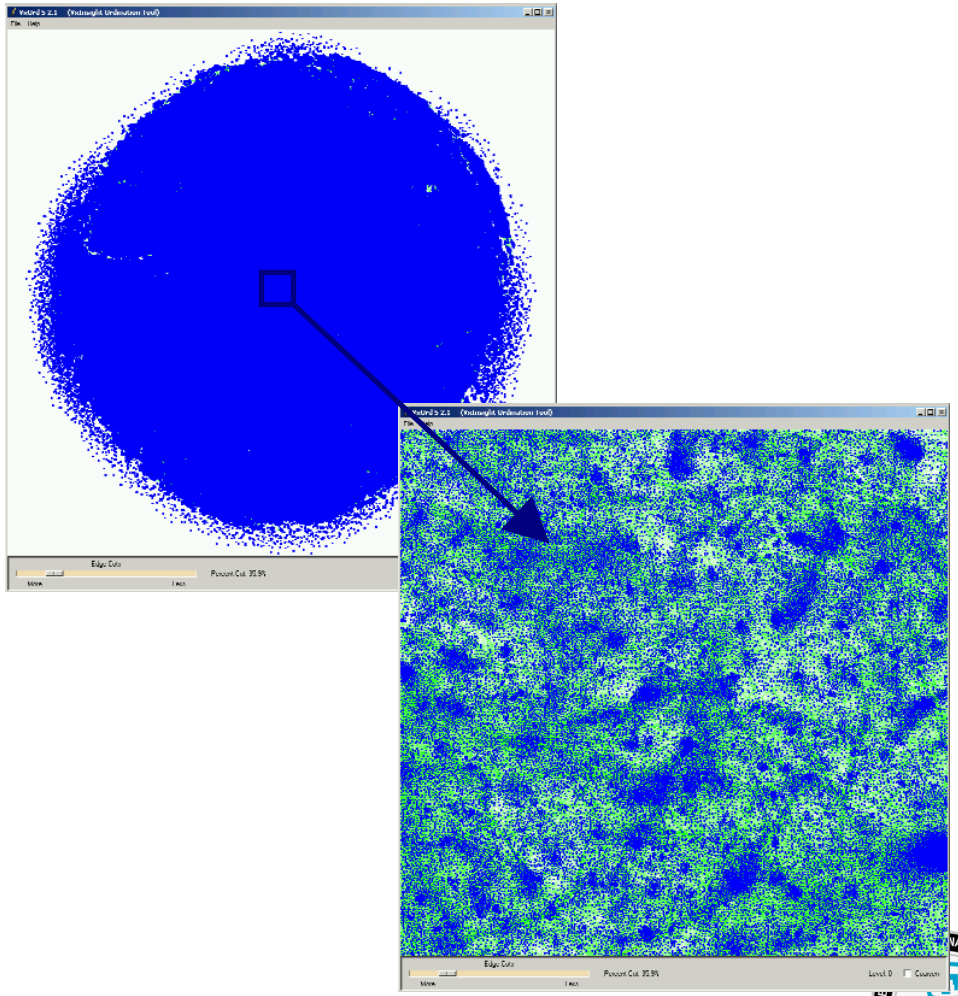
Macromodel (Structure) of Science

- Map of 7000 journals for the year 2000
- Based on citation statistics
- Validates methods so we can construct micromodel

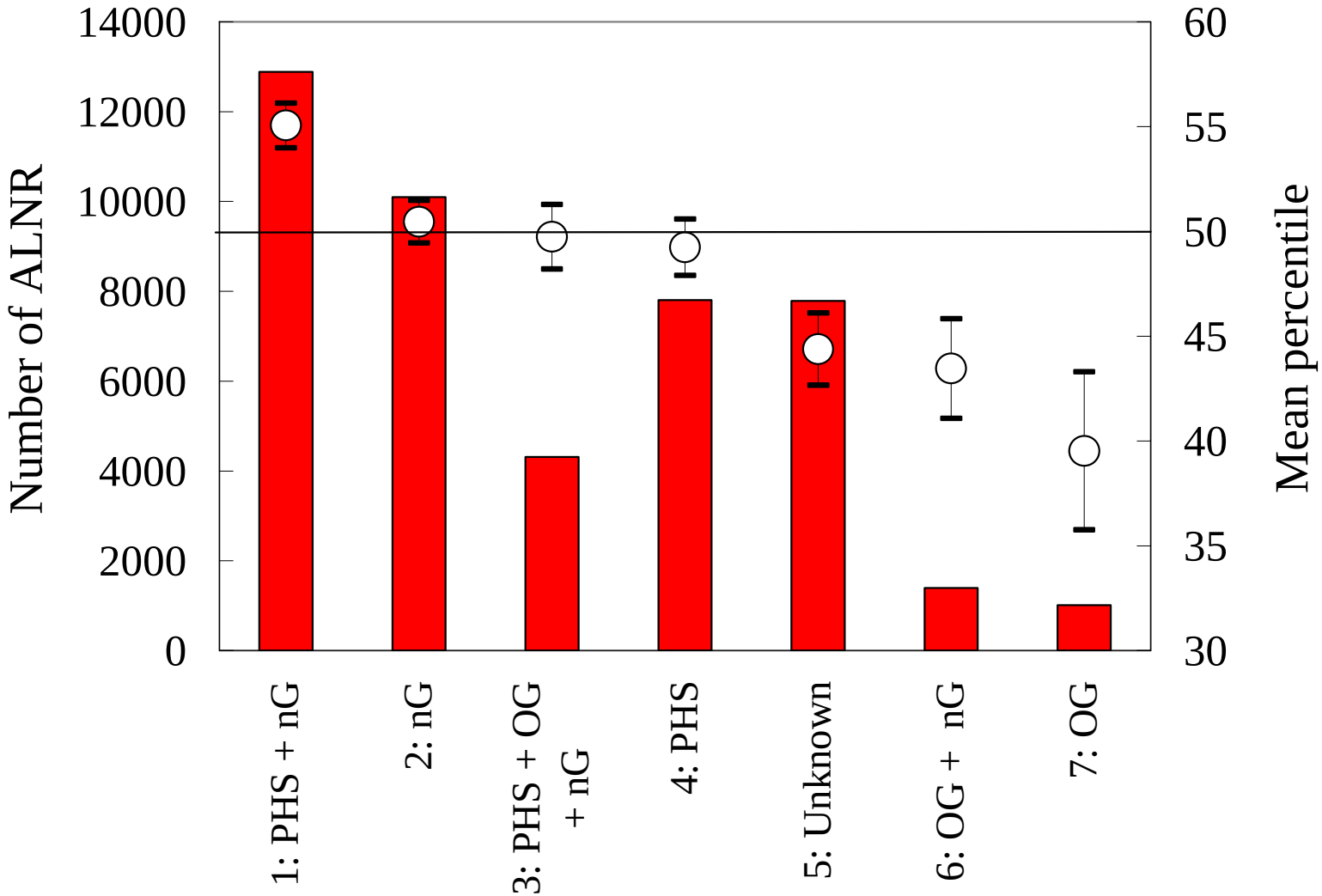


Micromodel of Science

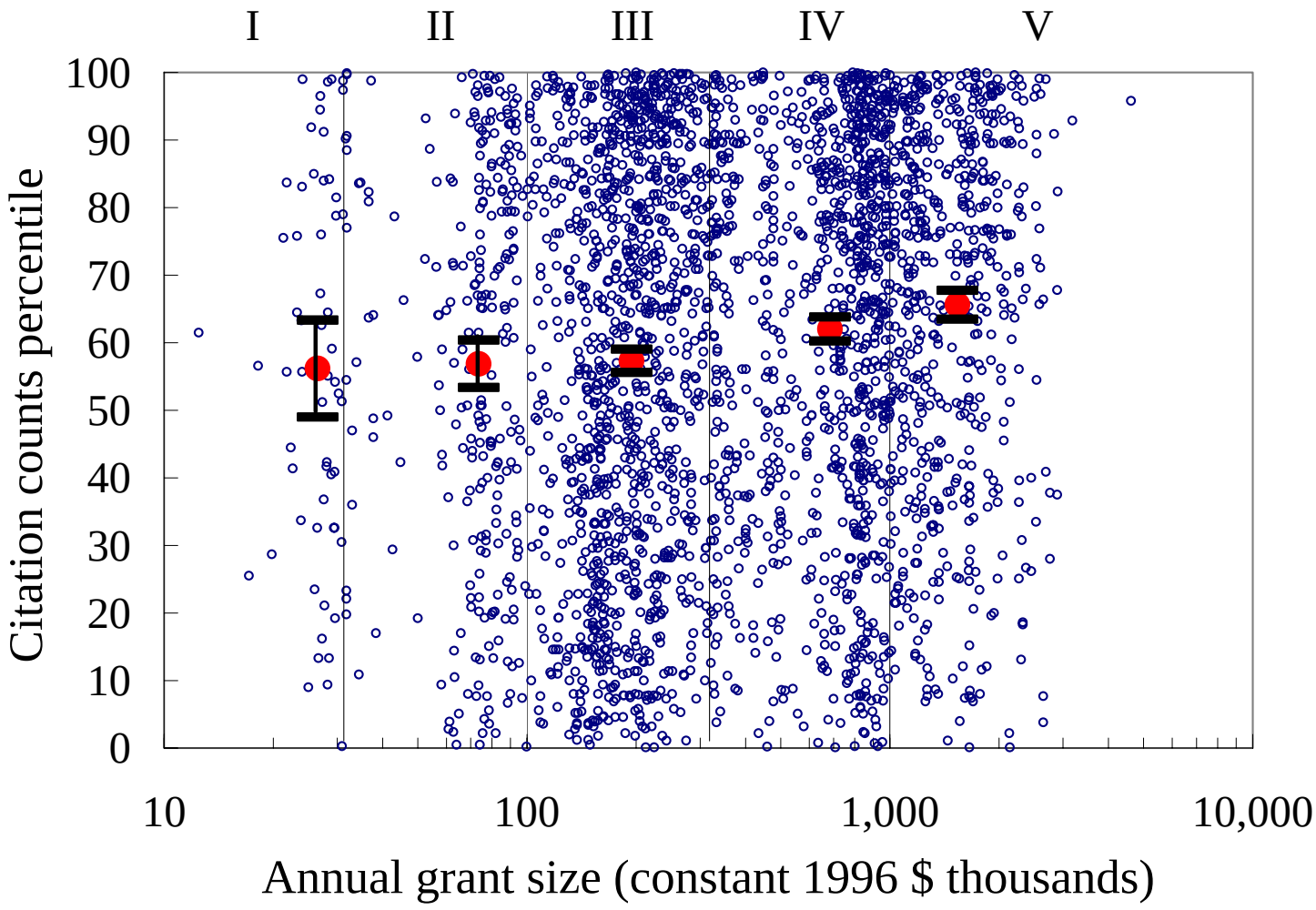
- Models science on scale that allows metric-based assessment to promote innovation
  - 833k nodes (papers), 5.6M edges
  - Local structure preserved
  - Develop local indicators



Impact by Funding Category



Impact by Grant Amount

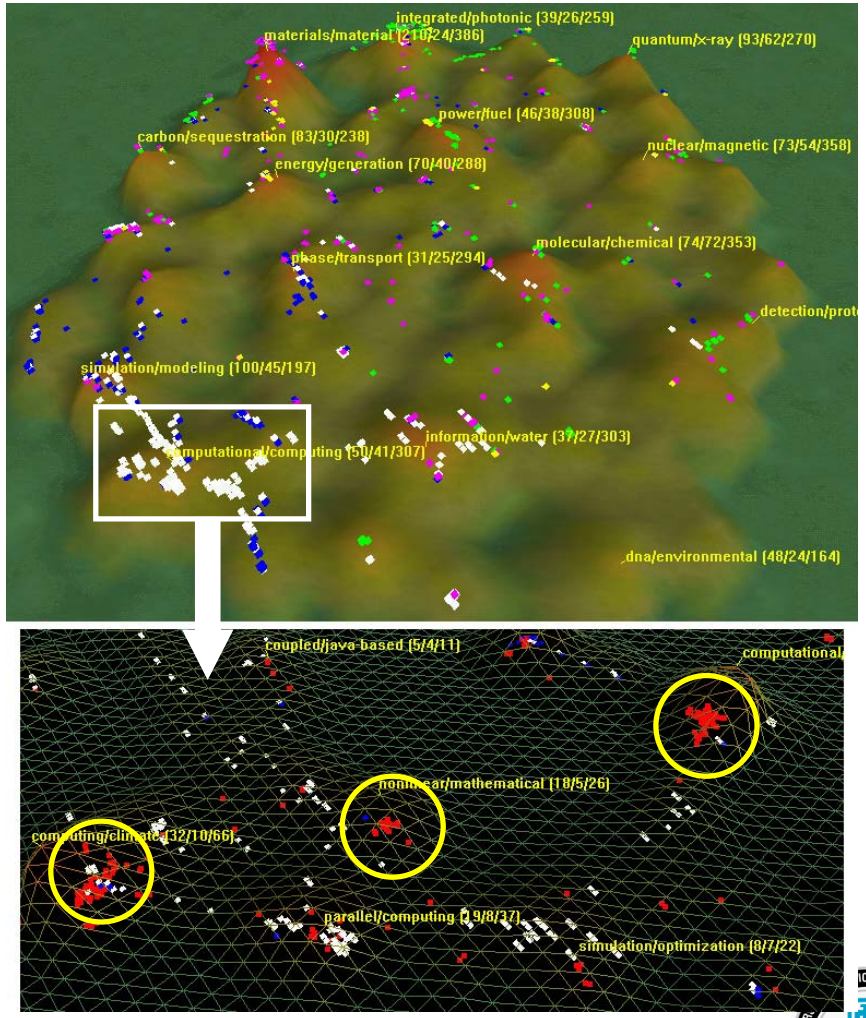


Corporate Information

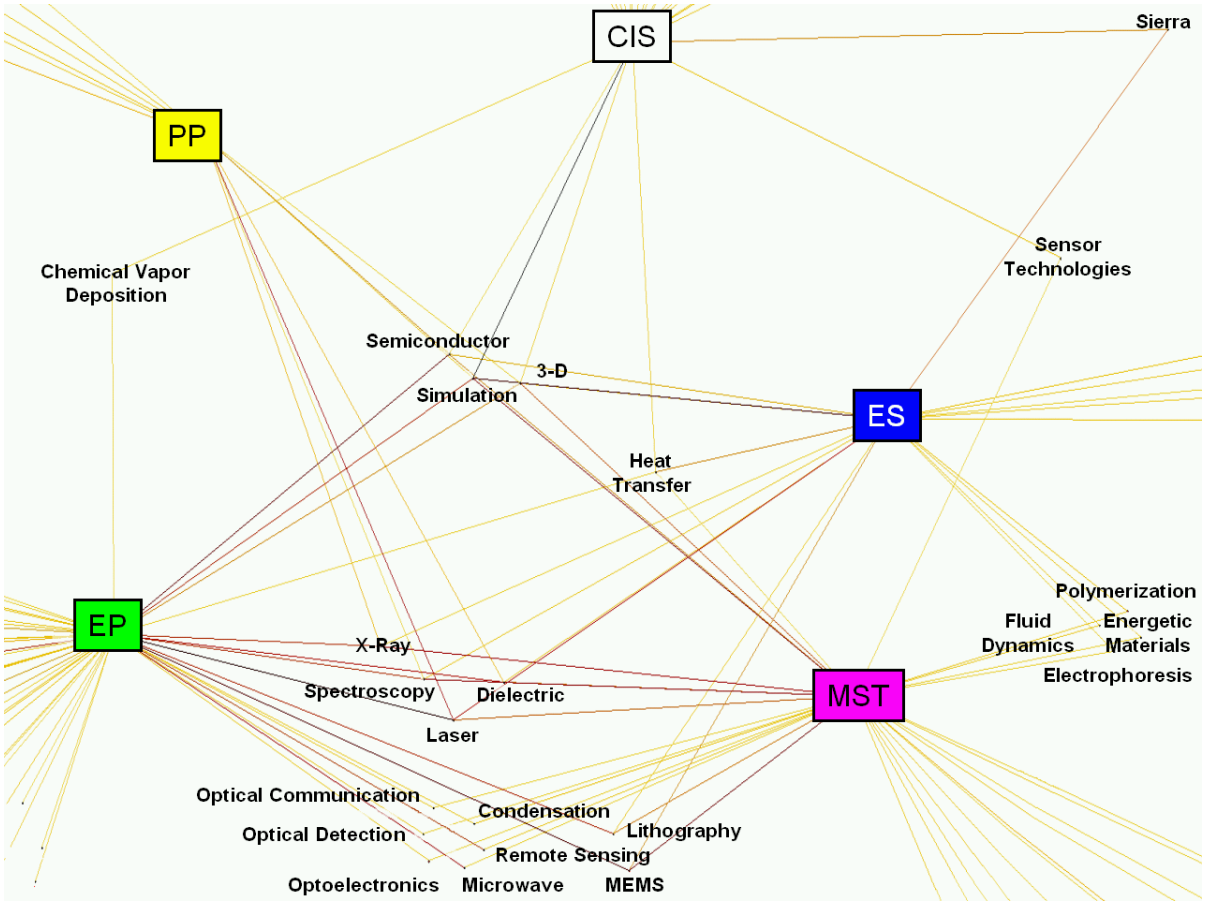
- Internal investment overlaps
- Investment opportunities
- Retrieval
- Parallel computer loganalysis

LDRD Investment Area Analysis

- Maps of Sandia LDTD IA's
  - As good as the mental models of IA leads
  - Trends, overlaps between IA's, identifies internal leverage points
- Maps of all DOE LDRD projects
  - Identify potential opportunities for IA's
  - Identify trends at other labs where they found a new opportunity

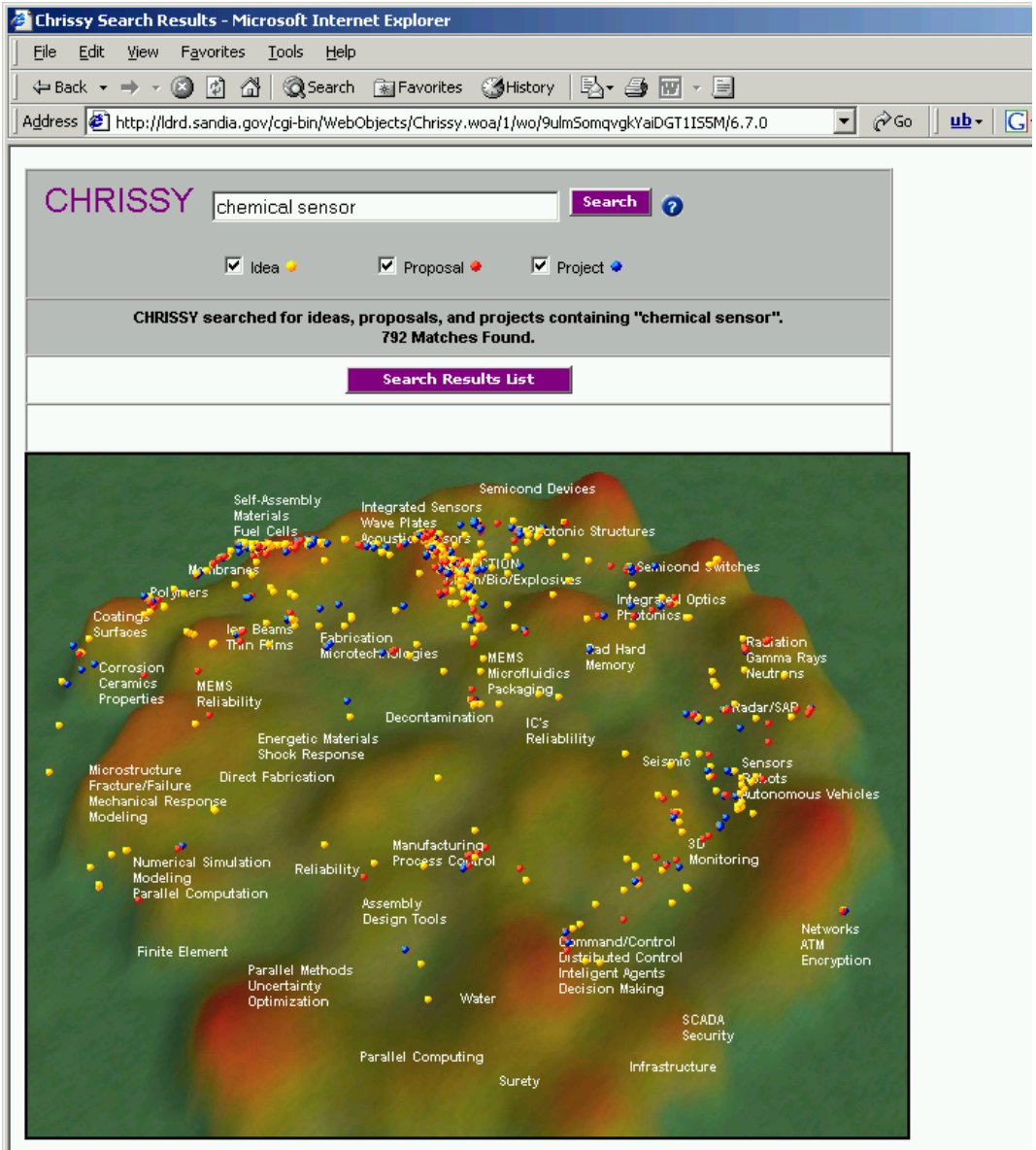


LDRD Investment Area Analysis



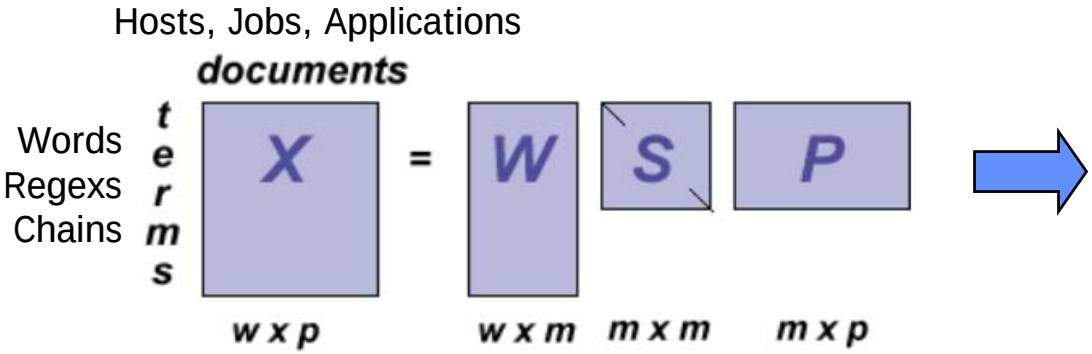
CHRISSY – LDRD Context Search

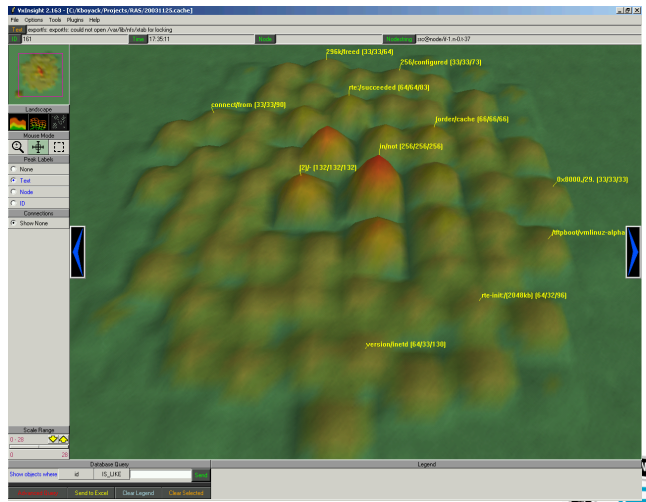
CHRISSY – LDRD Context Search



Loganalysis

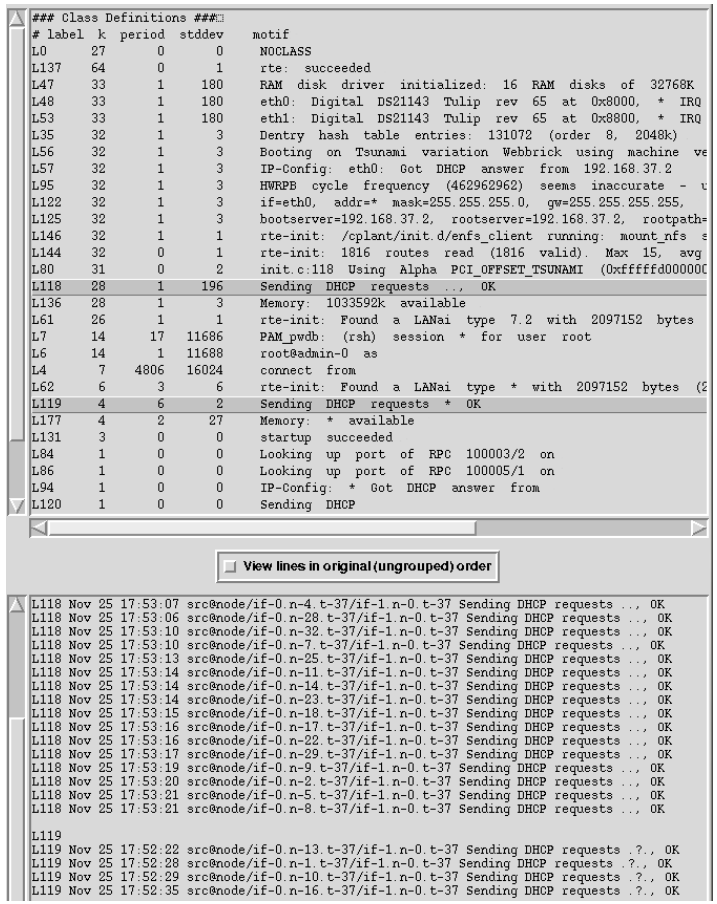
- Established expert-generated annotation of known error messages
- Established log archive of CPlant and ICC logs, in a format that supports precise and flexible subset selection using SQL
- Using multiple analysis types (Teiresias, LSA)





Loganalysis

- Implemented automatic regex generation using Teiresias
- Successes
  - Cplant – TSUNAMI hardware faults on Ross
  - ICC – PBS scheduler problems on Liberty
  - LosLobos (UNM AHPCC) version mismatch and portscan

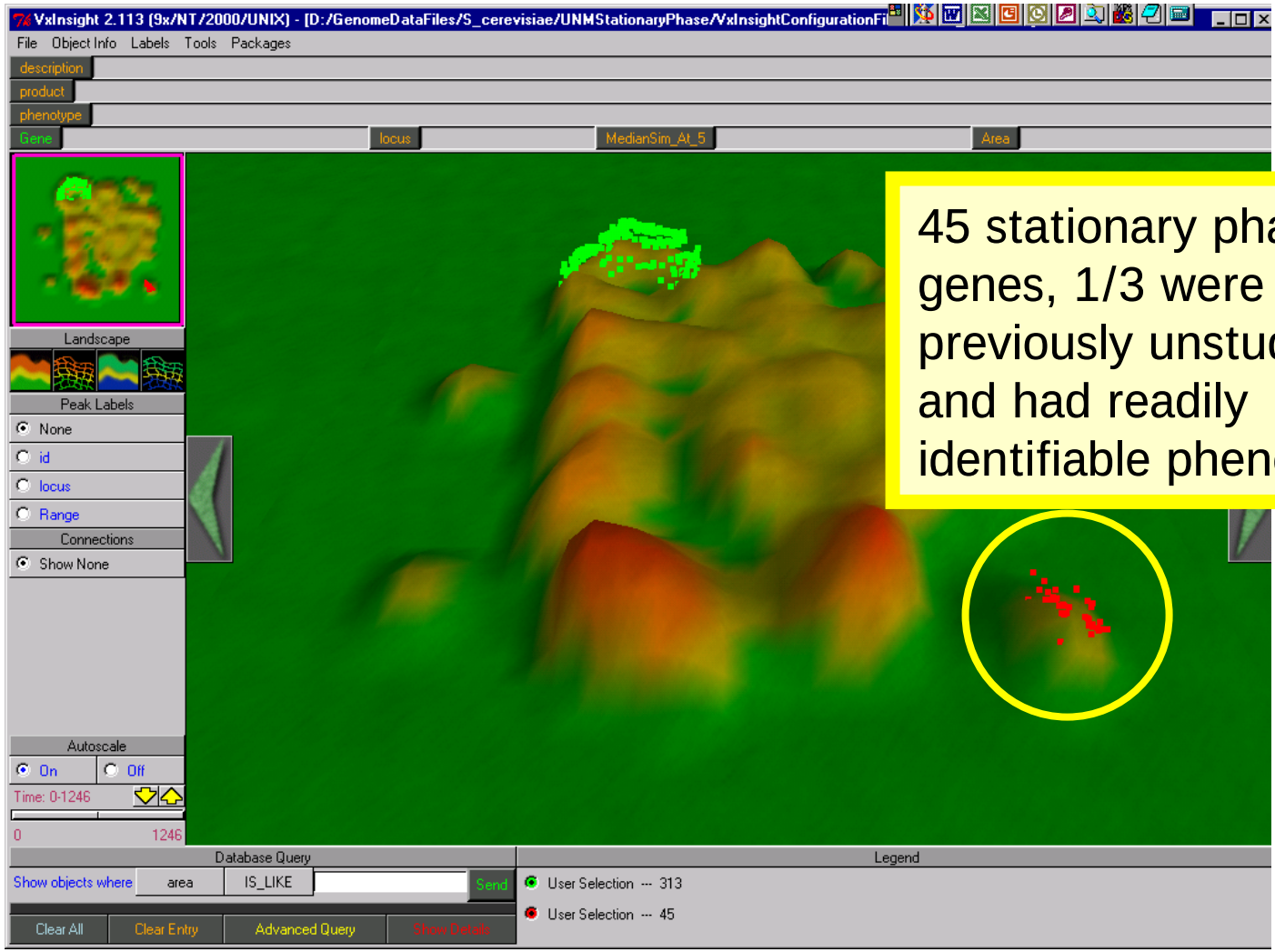


Bioinformatics

- Yeast
- C. elegans
- Infant leukemia
- How do you enable junior researchers to perform at expert level?

Yeast Stationary Phase Data

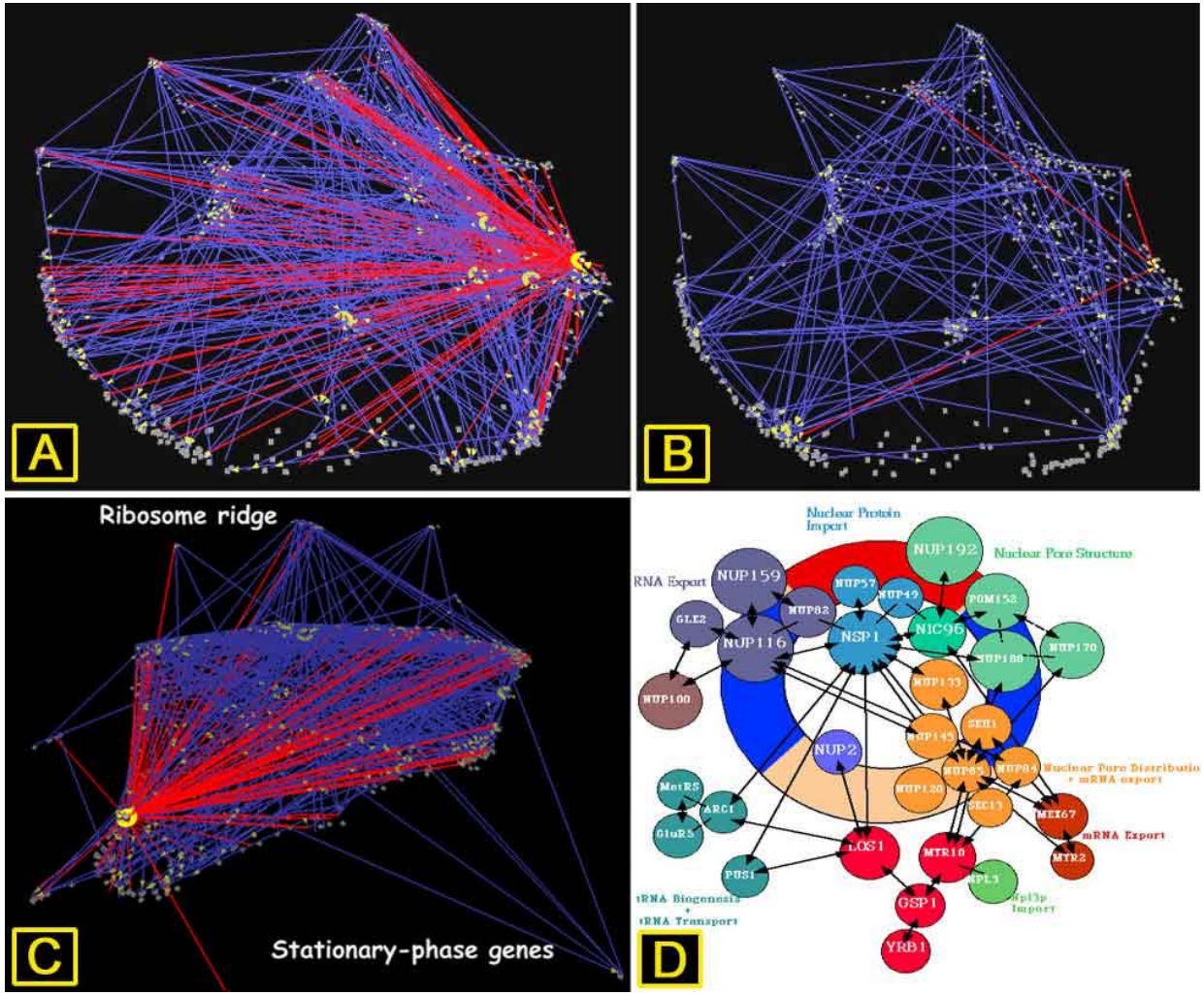
(Werner-Washburne lab)



45 stationary phase genes, 1/3 were previously unstudied and had readily identifiable phenotypes.

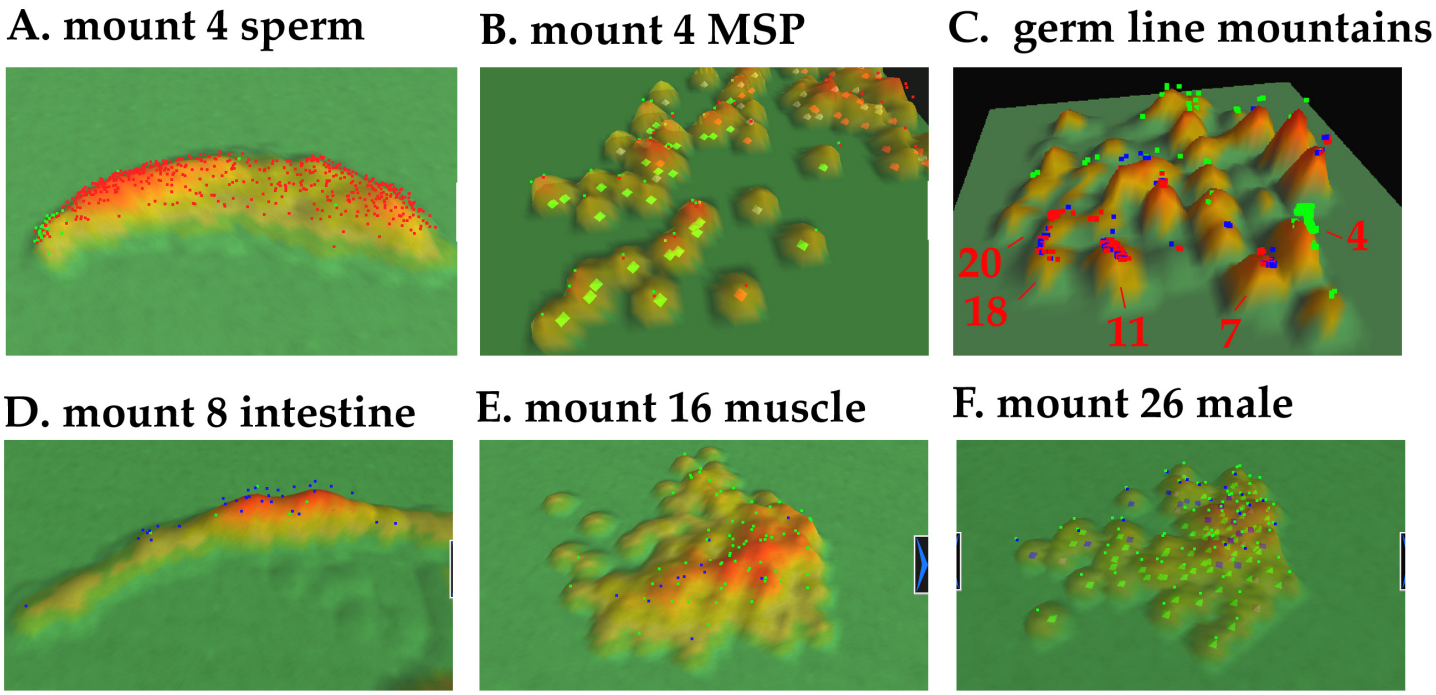
Yeast Protein Interactions on Gene Map

(Werner-Washburne lab)
Genome Research (**12**) 1564-1573, Oct. 2002



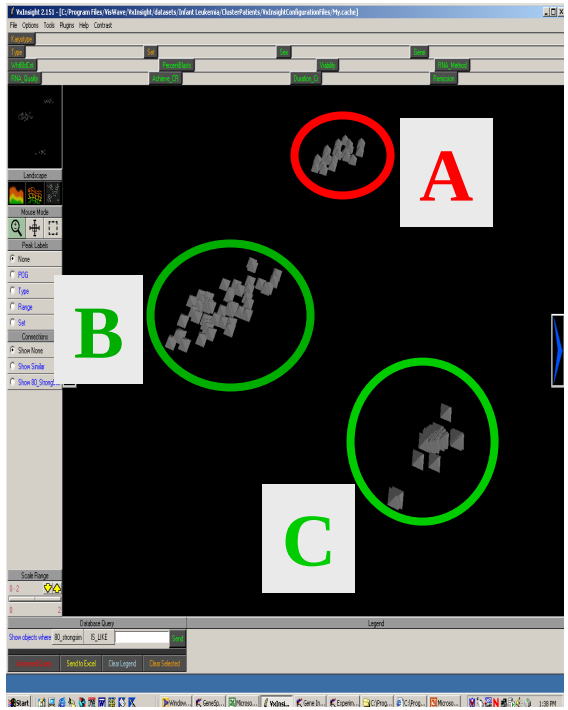
Topo Map of *C. elegans* genes

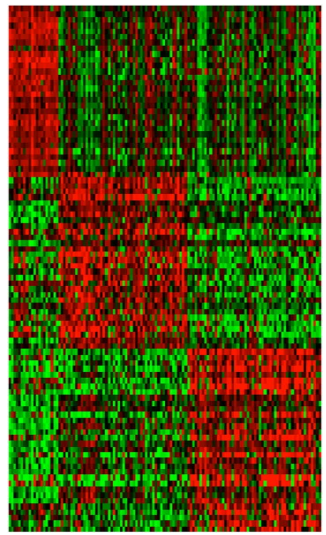
(Stuart Kim lab)
Science (**293**) 2087-2092, 14 Sept. 2001

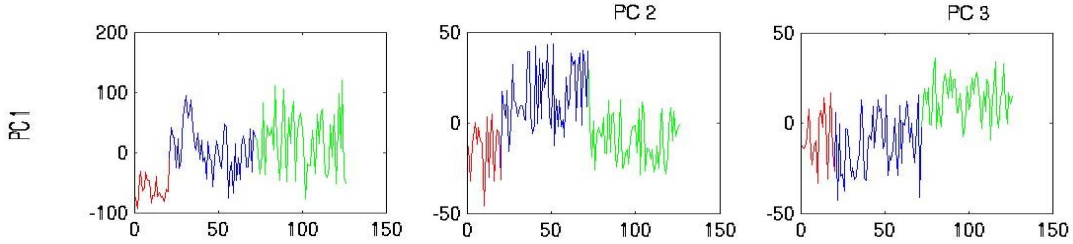


Infant Leukemia – New Results

- Vx finds structure in the expression patterns across the patients (**A**,**B**,**C**)
- All of our prediction methods improve when we condition on these patient subgroups
- What is going on biologically in these groups? (back to our fundamental question)

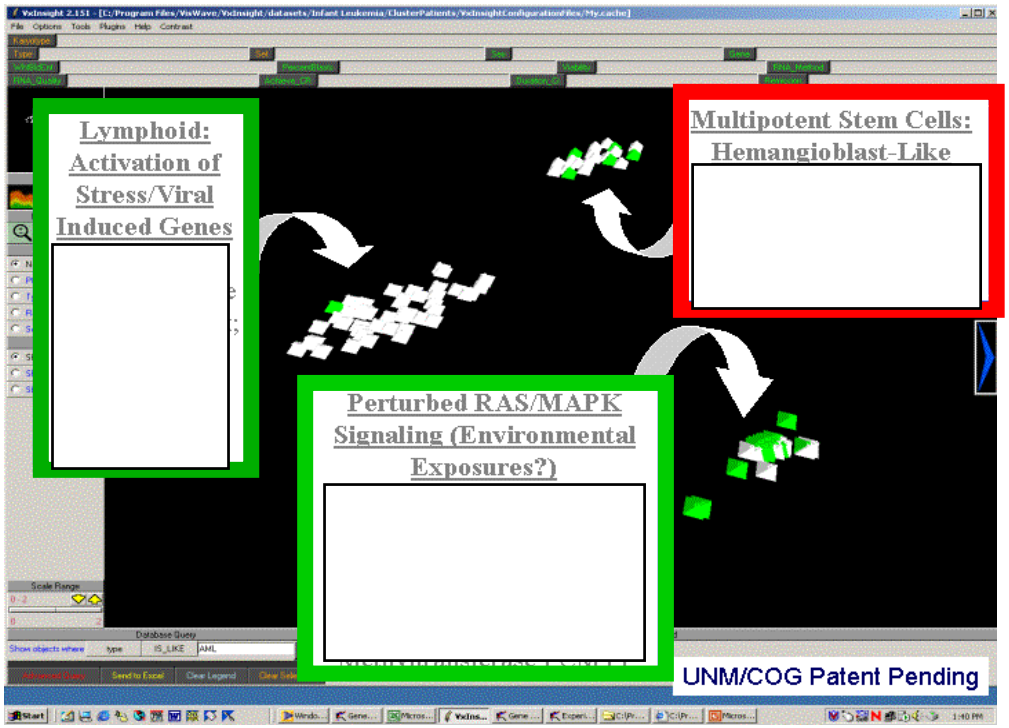






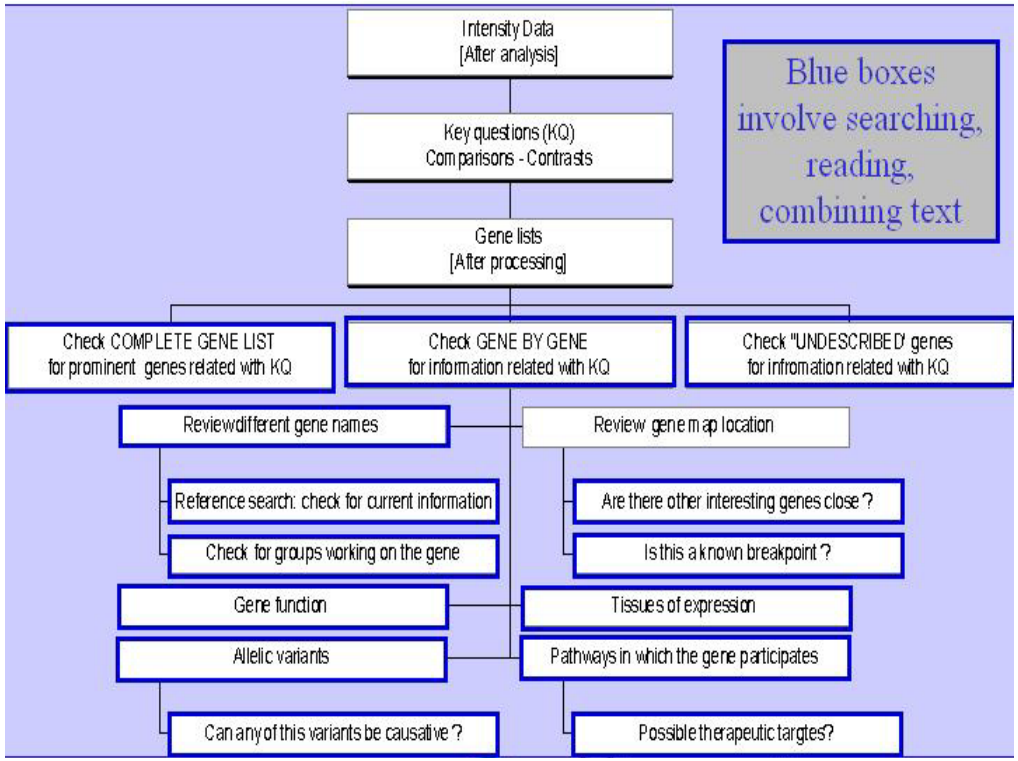
What the Genes Reveal

- **A** is a new and apparently very significant ***stem cell*** *leukemia*
- **B** is mostly ALL and, possibly, ***virally induced***
- **C** is, possibly, *induced by* ***environmental exposure***



Process to Enable Discovery

- Cluster by genes
- Cluster by arrays
- Contrast between clusters of patients to generate differentiating gene lists
- Used these gene lists as entries into the online databases and back to the original literature
- Read, read, and read some more.

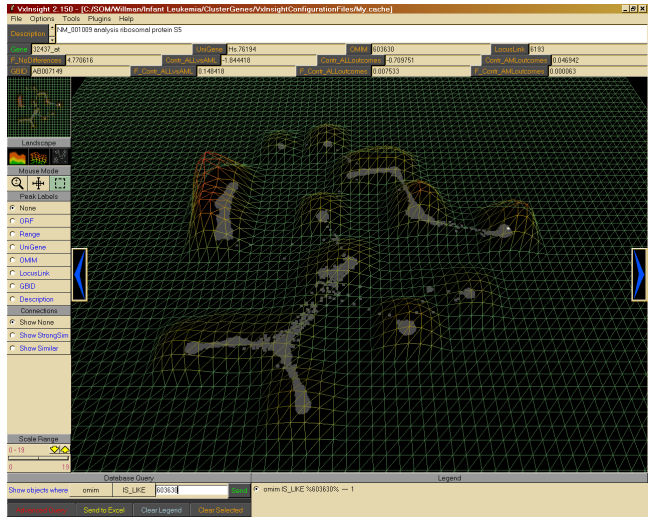


Multi-database knowledge mining

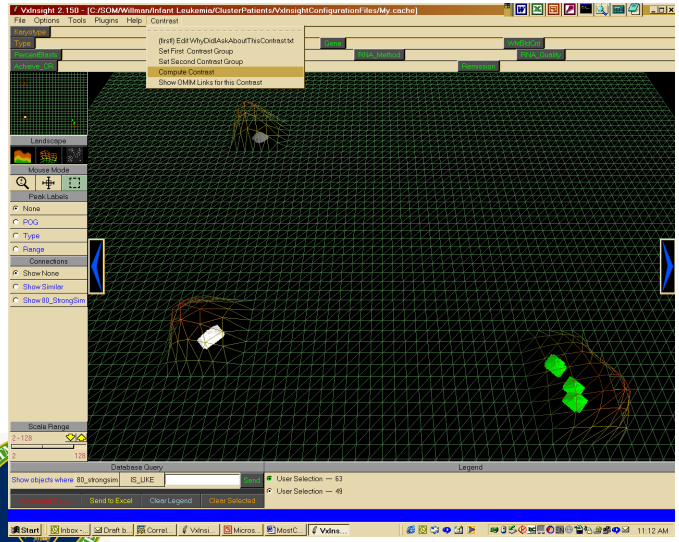
142 patients

12,625 genes

Cluster ***by genes***



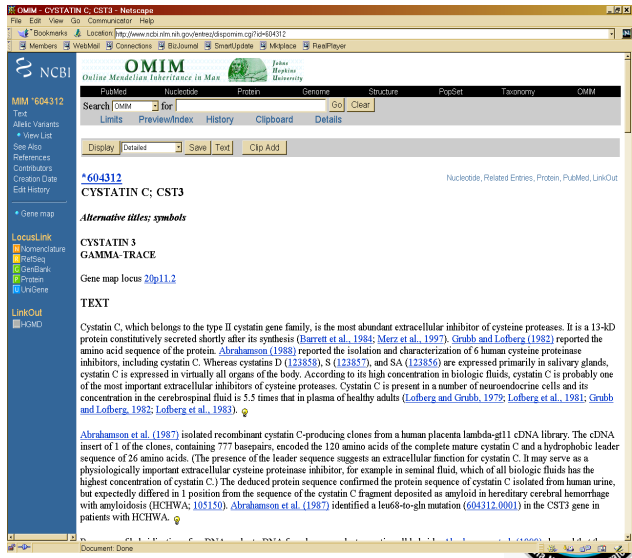
Cluster ***by arrays***



ANOVA & Contrast

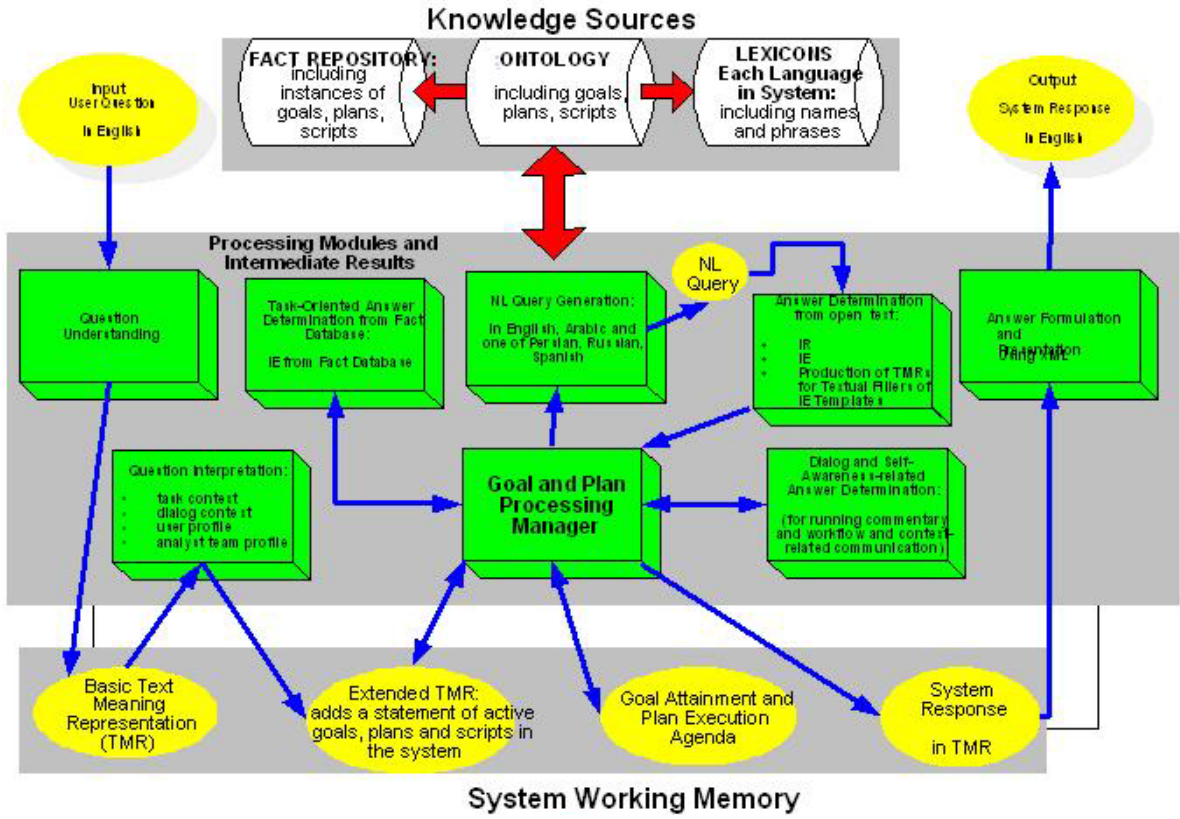
Contrast is Group1 - Group2

ORF	Contrast	F	URL
	1.55	149.72	
	1.40	149.38	Get OMIM details
	2.23	138.32	Get OMIM details
	1.87	135.59	Get OMIM details
	1.60	133.46	Get OMIM details
	1.24	126.09	Get OMIM details
	1.03	120.06	Get OMIM details
	1.08	119.77	
	2.21	118.18	Get OMIM details
	1.36	117.75	Get OMIM details
	1.52	115.49	
at	2.96	112.13	Get OMIM details
	1.19	108.09	Get OMIM details
	1.02	107.84	Get OMIM details



GLEE++

- Proposing something more intelligent for a *much* larger text corpora.
- Want deeper understanding, and biologically useful reasoning.



Some Issues

- Scale – need to analyze and visualize extreme scale (>1M nodes) networks
- Data – extraction, merging of different types and formats, relationship extraction, etc., to generate knowledge
- Understanding from NLP, deep semantics
- Validity

Recent Publications

Boyack, K. W. (2004). Mapping knowledge domains: Characterizing PNAS. *Proceedings of the National Academy of Sciences*, 101(S1), 5192-5199.

Boyack, K. W., & Börner, K. (2003). Indicator-assisted evaluation and funding of research: Visualizing the influence of grants on the number and quality of research papers. *Journal of the American Society for Information Science and Technology* 54(5), 447.

Börner, K., Chen, C., & Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology* 37, 179-255.

Werner-Washburne, M., Wylie, B., Boyack, K., Fuge, E., Galbraith, J., Fleharty, M., Weber, J., & Davidson, G.S. (2002). Concurrent analysis of multiple genome-scale datasets. *Genome Research* 12(10), 1564-1573.

Boyack, K. W., Wylie, B. N., & Davidson, G. S. (2002). Domain visualization using VxInsight for science and technology management. *Journal of the American Society for Information Science and Technology,* 53(9), 764-774.

Davidson, G. S., Wylie, B. N., & Boyack, K. W. (2001). Cluster stability and the use of noise in interpretation of clustering. *Proc. IEEE Information Visualization 2001*, 23-30.

Kim, S. K., Lund, J., Kiraly, M., Duke, K., Jiang, M., Stuart, J. M., Eizinger, A., Wylie, B. N., & Davidson, G. S. (2001). A Gene Expression Map for Caenorhabditis elegans. *Science, 293*, 2087-2092.

Boyack, K.W., Wylie, B.N., Davidson, G.S. & Johnson, D.K., *Analysis of patent databases using VxInsight*. Presented at New Paradigms in Information Visualization and Manipulation 2000, McLean, VA, Nov. 10, 2000.

Beck, D.F., Boyack, K.W., Bray, O.H. & Siemens, W.D., “Landscapes, Games, and Maps for Technology Planning,” *Chemtech* 29(6), 8-16, 1999.