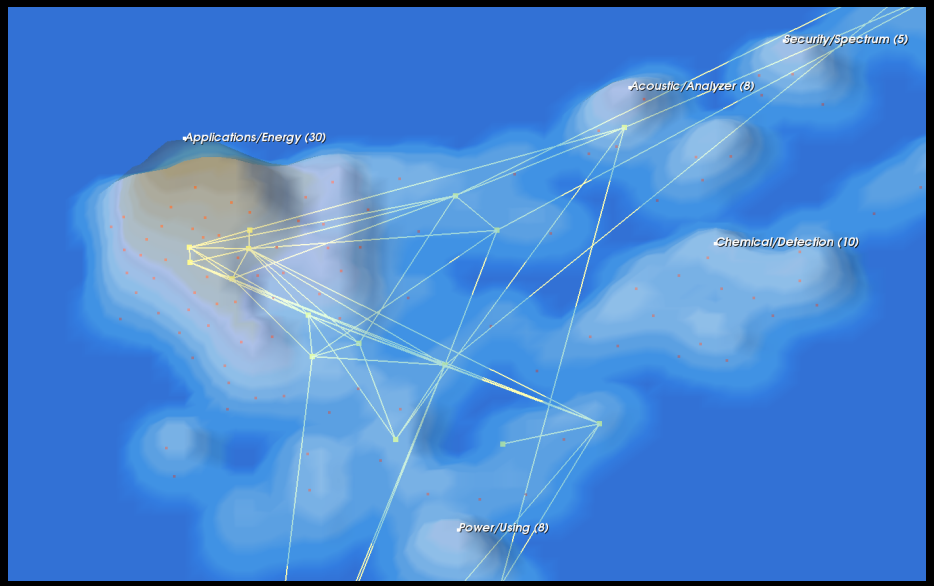
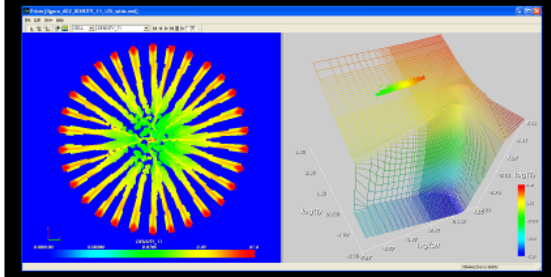
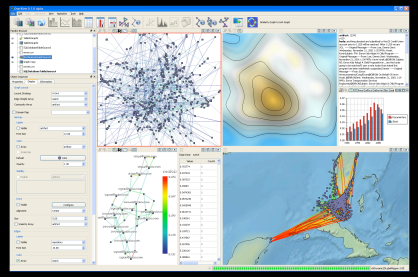
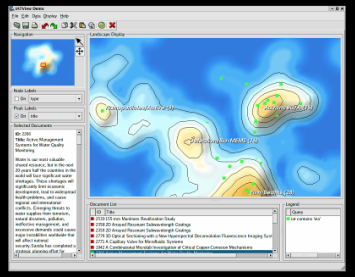
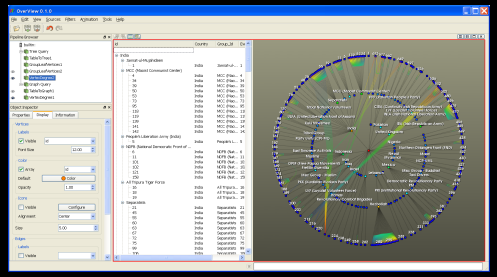
Scalable Informatics











David H. Rogers

Data Analysis and Visualization Department 1424

Sandia National Laboratories

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000

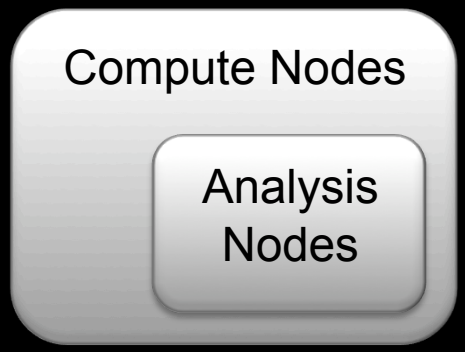
Sandia National Laboratories

Warmup Slides

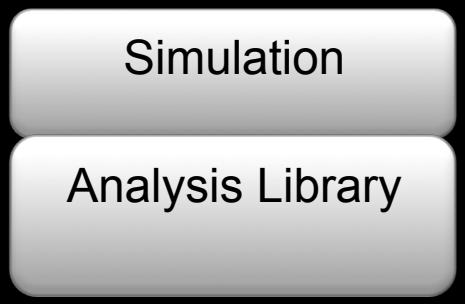
- Analysis reaches from disk to user, through algorithms
- “20% of the system devoted to Resiliency”
  - Question: How much of the system will you devote to Results?
  - 30% tax on a timestep? (example to follow)
  - Should analysis services migrate to other areas (OS, etc.)
- Metrics
  - “Let’s do away with FLOPS”
    - Discoveries/joule, Insights/Joule
    - Patents/joule
    - Nobels/joule
    - Bad guys/joule
- The ‘write results’ paradigm is dead
  - Should services be supported at a lower level
- Informatics is the next big challenge – for everyone

Lessons from Zia Design Team

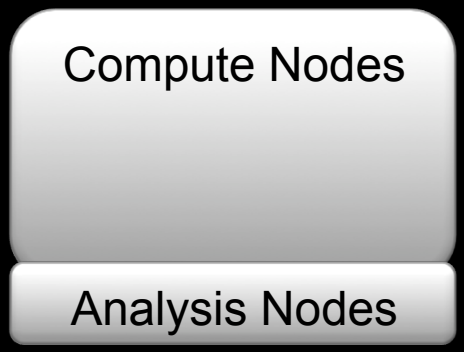
- Analysis should be included in system design
- For extreme systems, budget and FLOPS-colored glasses bias us towards analysis on the platform
- A policy-driven analysis resource is a partial solution
  - Enables In-situ visualization (coupled, or system level resource)
  - Enables a flexible analysis resource, because any set of nodes will do
- A variety of use cases are supported by analysis-on-platform



Interactive Queue on a set of nodes

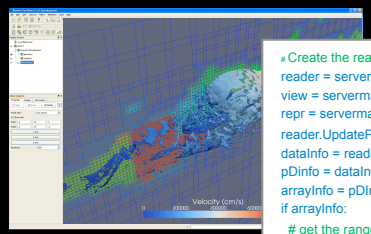


Simulation coupled to general Analysis Library



Adaptive Service Node Set (write at network speed, not disk speed)

Petascale and Beyond *Requires* In-situ Analysis



```
# Create the reader and set the filename.
reader = servermanager.sources.Reader(FileNames=path)
view = servermanager.CreateRenderView()
repr = servermanager.CreateRepresentation(reader, view)
reader.UpdatePipeline()
dataInfo = reader.GetDataInformation()
pDinfo = dataInfo.GetPointDataInformation()
arrayInfo = pDInfo.GetArrayInformation("displacement9")
if arrayInfo:
  # get the range for the magnitude of displacement9
```

ParaView script saved, used in input deck

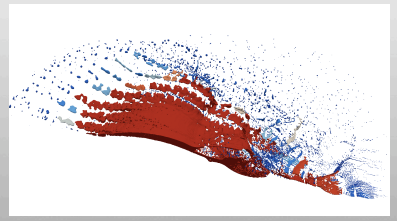
Simulation
(CTH, Presto, S3D, etc.)

Analysis Services (ParaView inside)

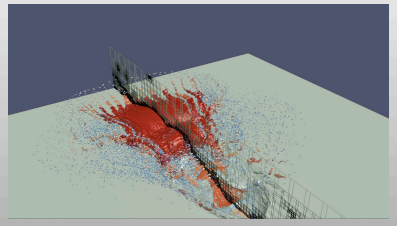
- Service coupled to simulation
- Service as part of platform

Analysis Products (Output)

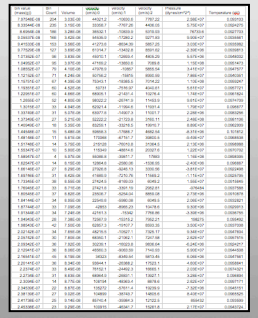
Polygonal and Field Data



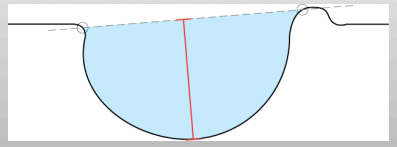
Rendered Images



Fragment Statistics

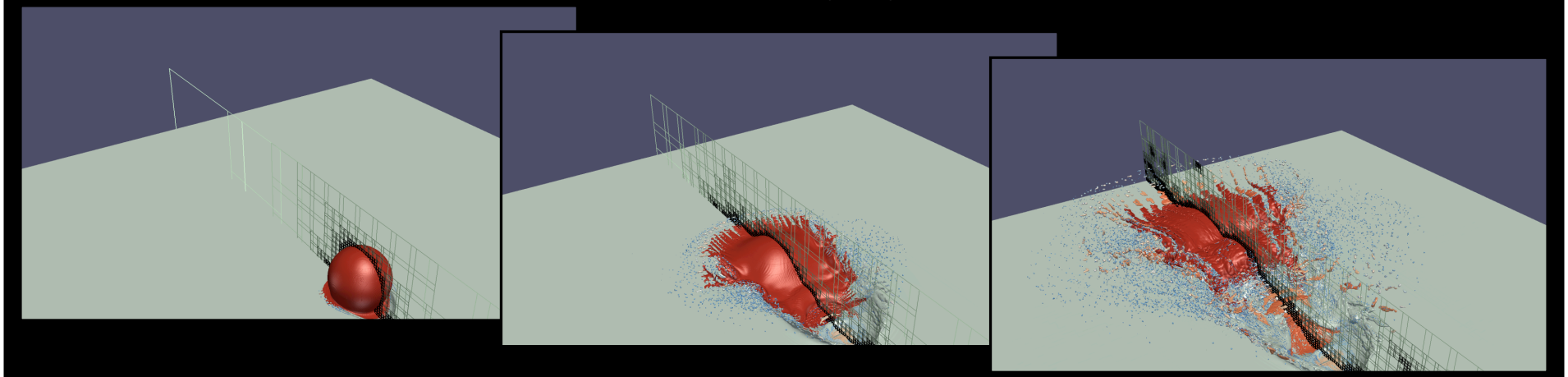


Crater Statistics

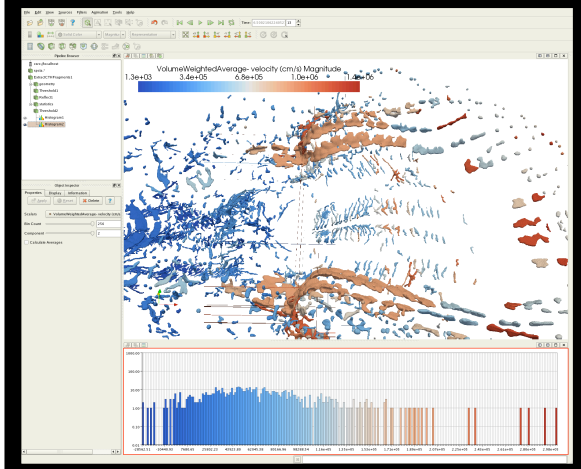


In-situ Fragment Analysis Demonstrates Value of Coupled Analysis

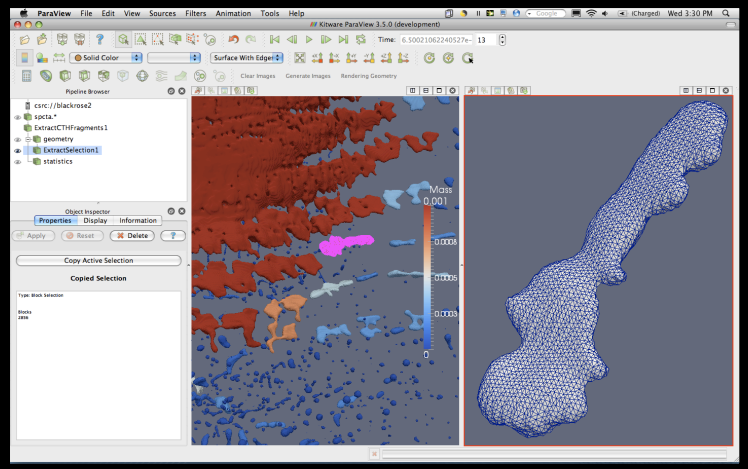
Frames from a CTH simulation, showing fragment data calculated In-situ



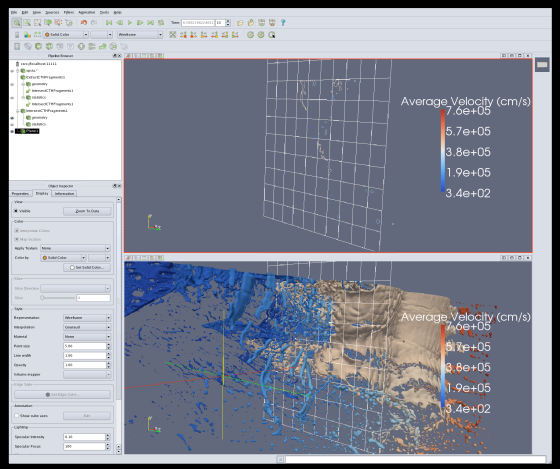
Fragment Drill-down info.



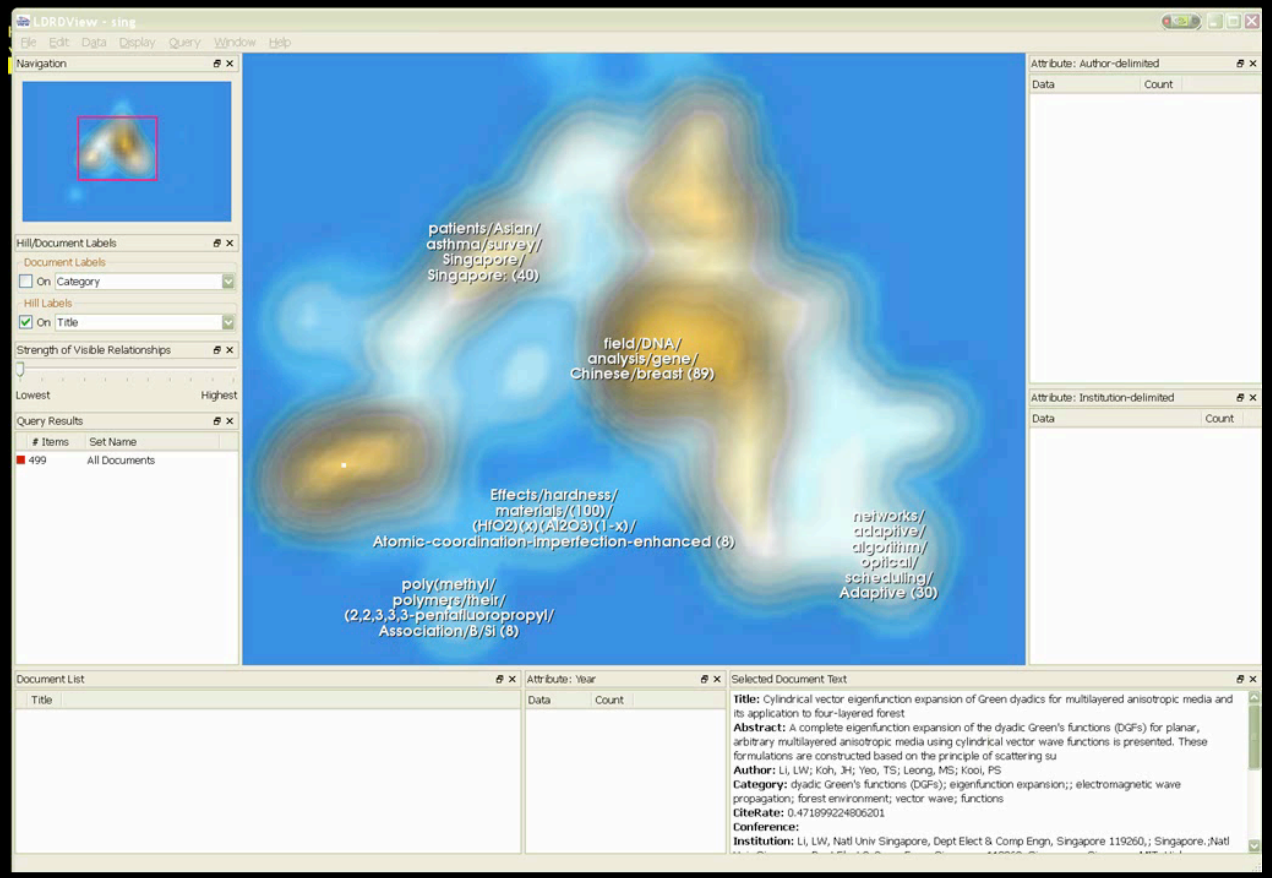
Coupled re-meshing (CUBIT)



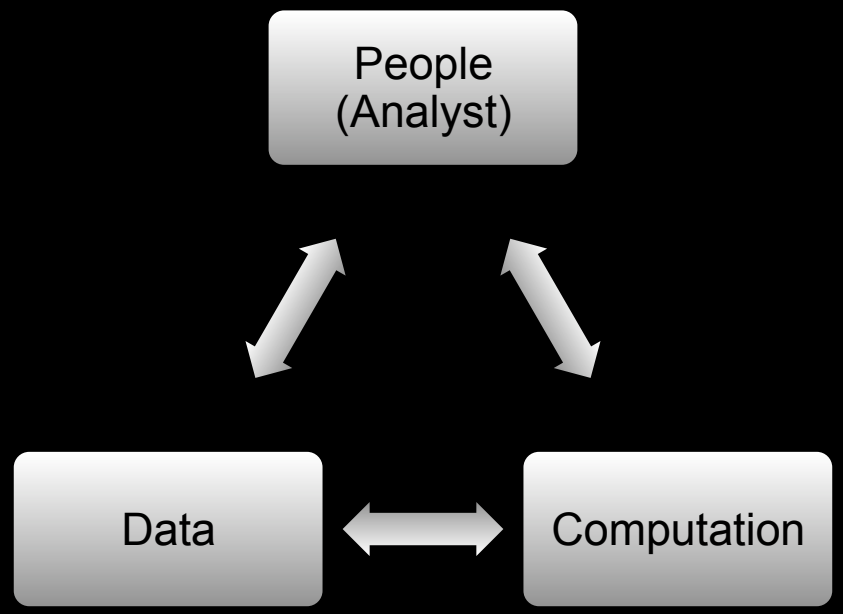
Fragment-based operations



It’s Informatics All the Way Down



Interactivity is Central to Discovery

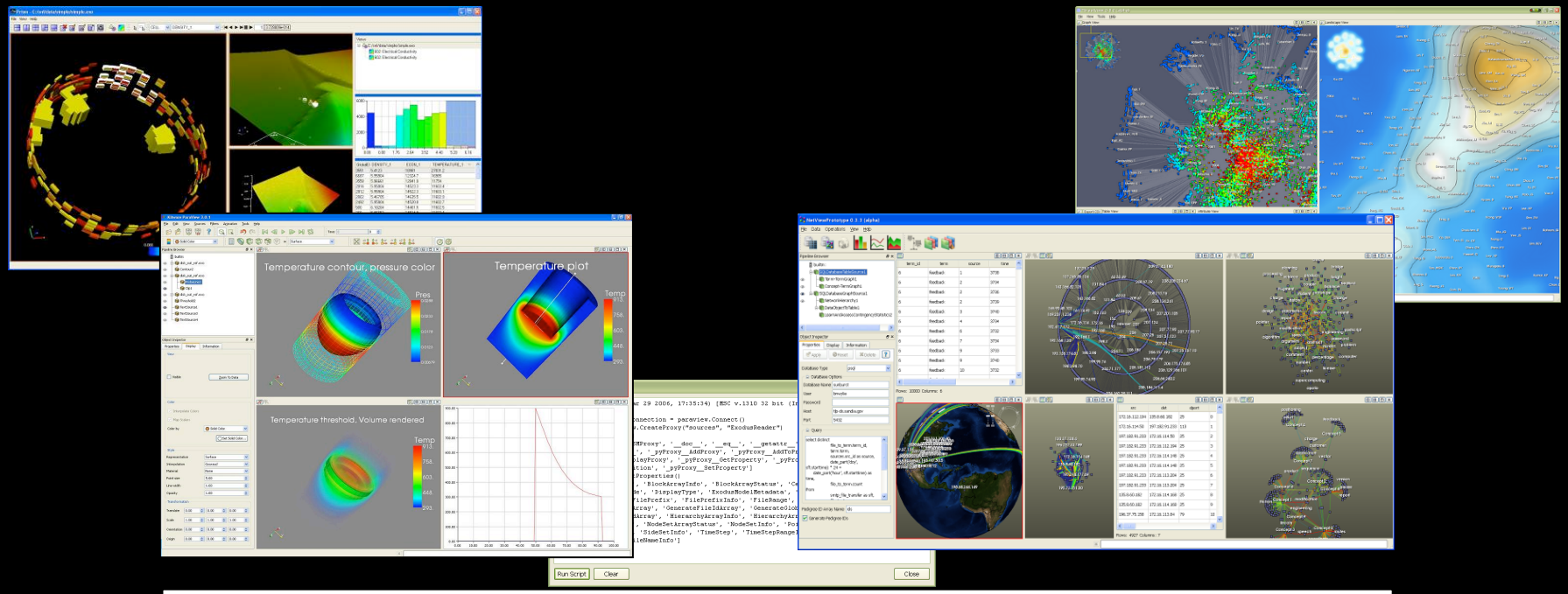


- Interactivity encompasses all interactions with user:
  - Queries, including updates to UI elements
  - Results of Adjustments to Algorithms
  - Selection (even in abstract spaces)
  - Creation of new views

… regardless of size/location of data or resource

- Insight can come from many places
  - SQL Query
  - Algorithm (social network)
  - Visualization of results (graph layout)

Family of Scalable Analysis Tools, Built on Scalable Technology



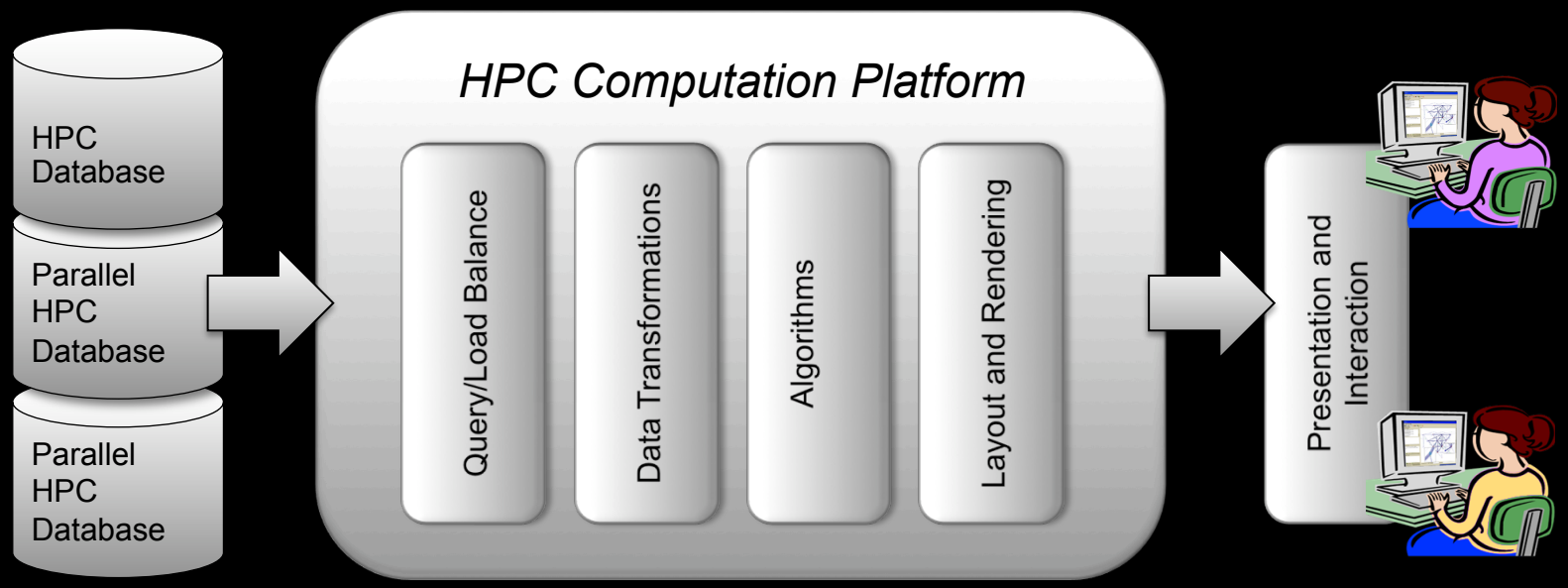
Parallel Server/Client Library

TITAN Toolkit

VTK (Visualization Toolkit)

For Big Data, We Could Use HPC Anywhere

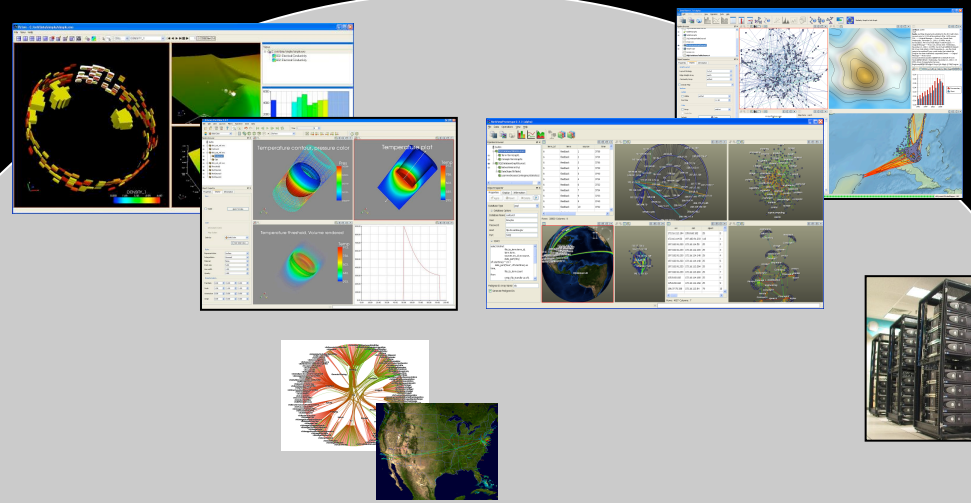
- Need Scalable Software that exploits available hardware
  - Runs on everything from laptop to remote cluster/SC
  - Runs algorithms on appropriate hardware, when available
  - Provides basic functionality, even if it's slow on large data
  - Flexible support for integration of advanced HPC algorithms



Informatics Technical Strategy

- Goal: Develop Scalable Informatics for large data problems
  - Support wide range of research
  - Develop Sandia expertise
  - Adopt solutions when necessary
- Principles
  - Users must evaluate technologies on actual problems
    - There are plenty of real problems
  - End-user tools must be flexible, to promote integration of research
  - Use Rapid Prototyping to deliver research and potential capabilities
    - Keep the stuff that works
  - Research is a primary focus
- Results
  - Titan project: Informatics additions to current scalable toolkit (VTK)
    - OverView application, providing basic informatics capability
  - Promote research projects that can stand alone
  - Projects provide components for integration into Titan
  - Capabilities can then be expressed in tools

Combine Scalable Analysis Framework, Scalable Algorithms and Ingestion



Titan

Remote Data

DB Drivers

File Readers

Streaming Data

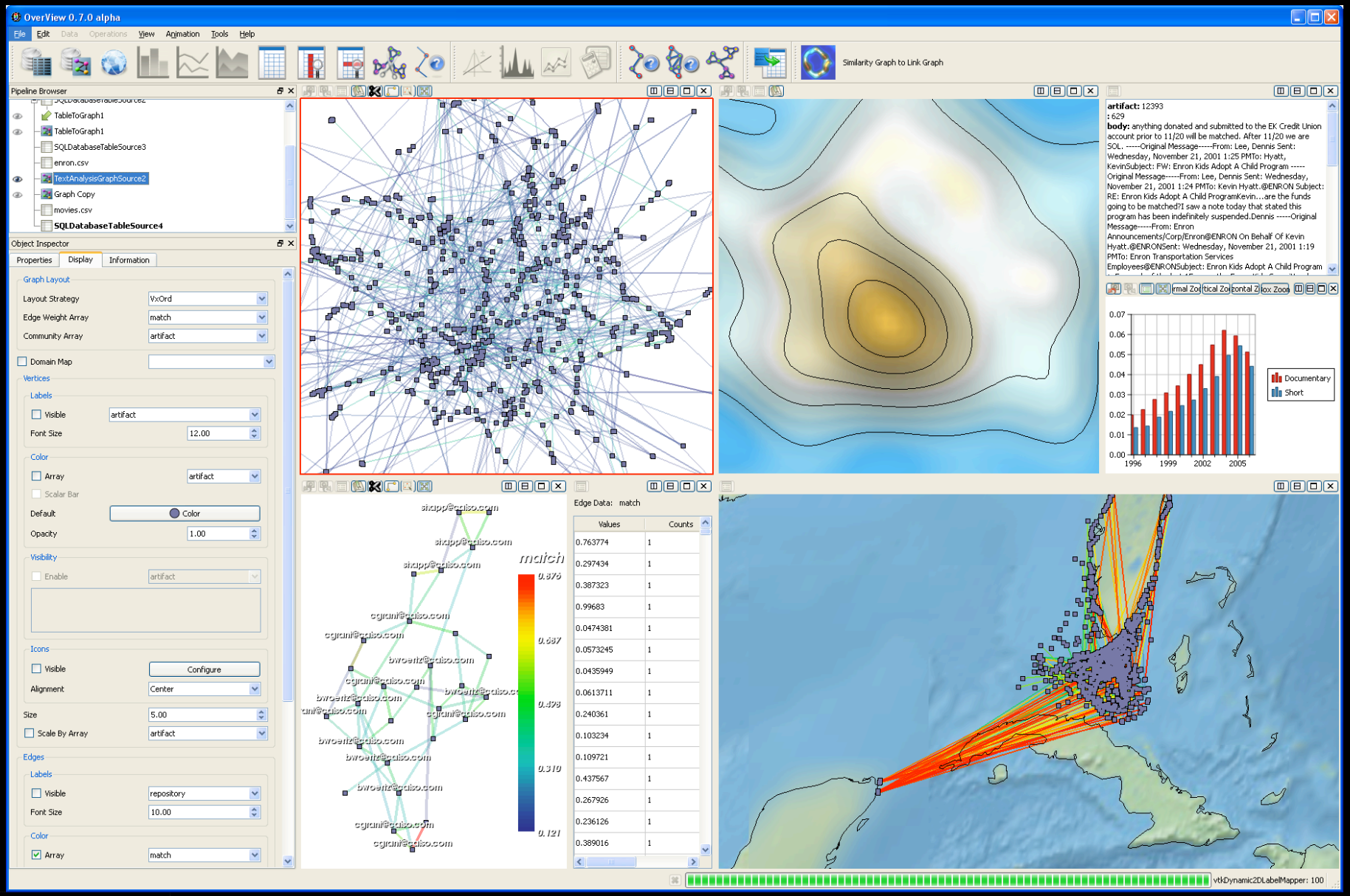
Matlab

MTGL

PBGL

Trilinos

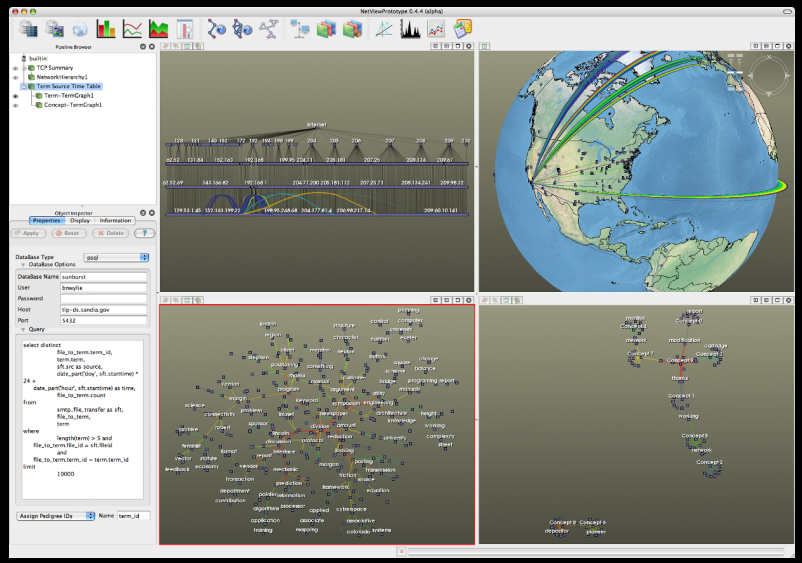
OverView: End-user Open Source tool based on scalable framework

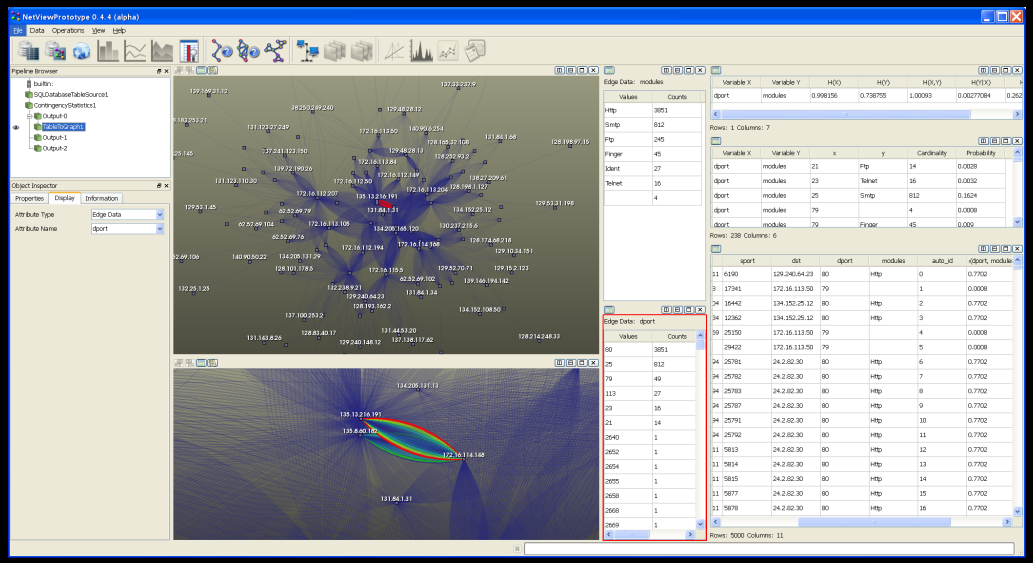


Sandia National Laboratories

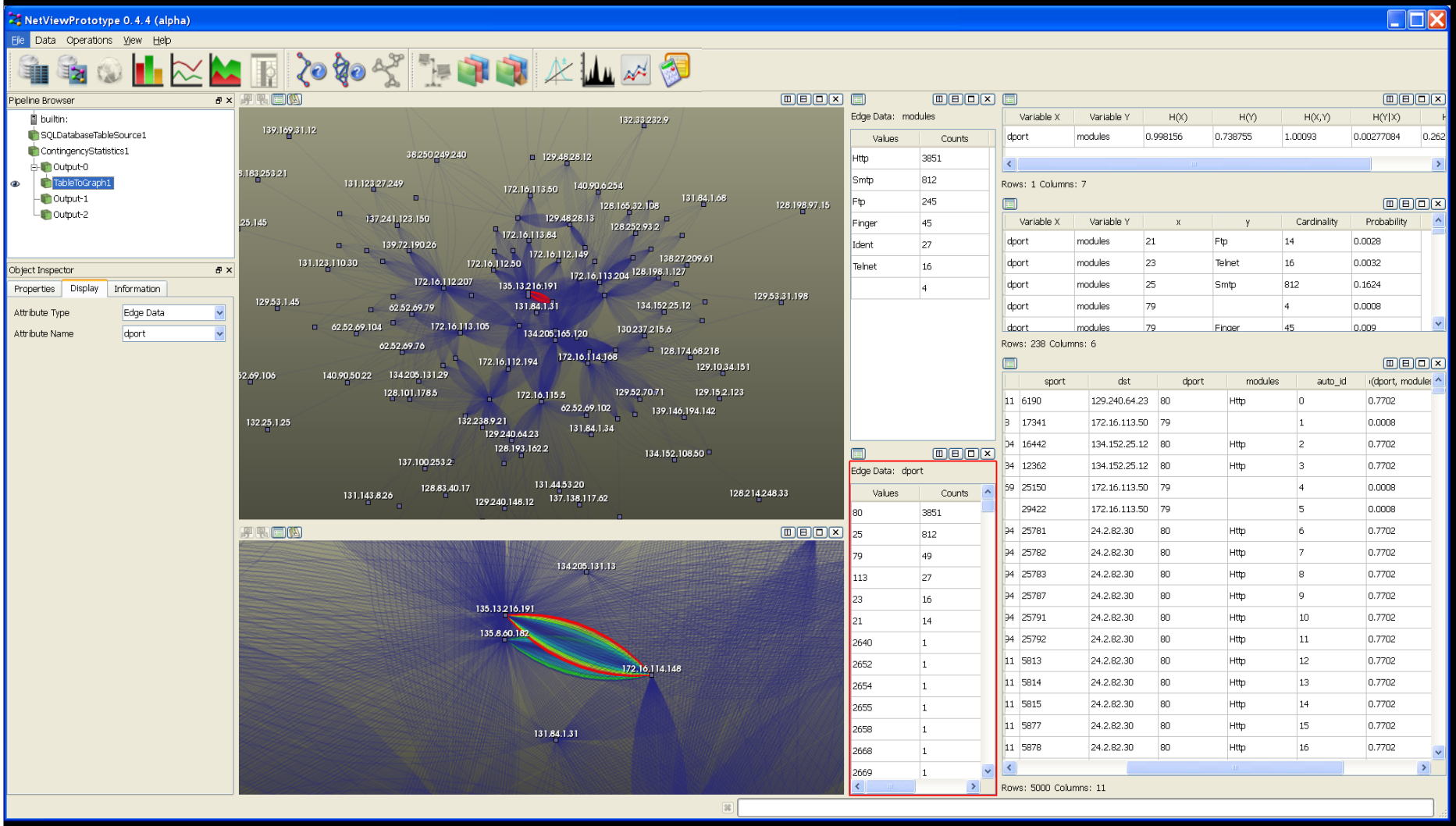
Network Grand Challenge: Research into Scalable Algorithms

- "The Network is the Adversary"
  - Social networks, cyberdata, geovisualization
- Project explicitly combines efforts in Algorithms, Databases, Visualization and tool delivery
  - Getting algorithms into users' hands and evaluating results
  - Iterating on successful components
- Strong desire to create software that can be productized



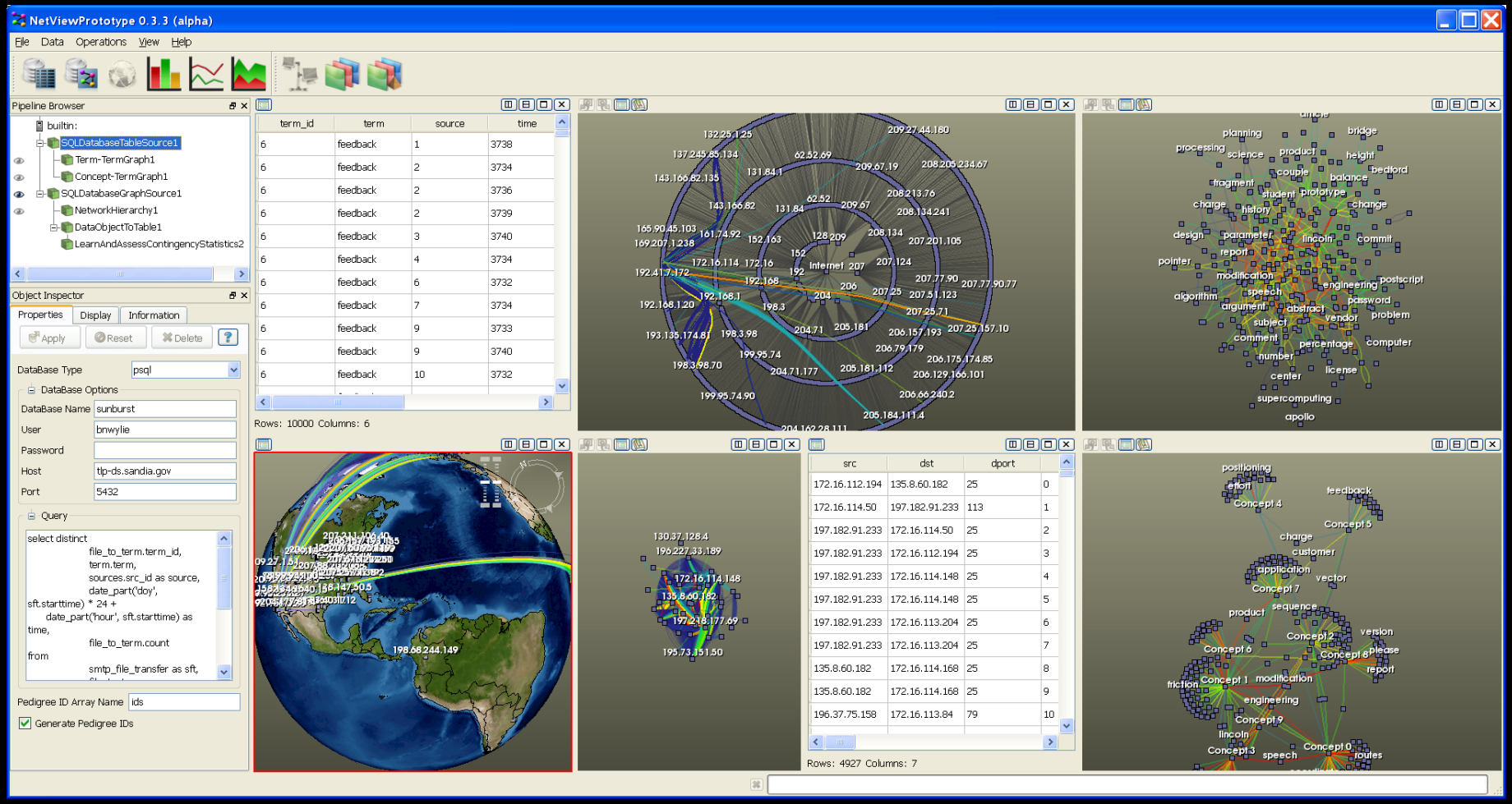


Find Me Some Anomolies, Please



- Of Particular Interest are those things that the analyst is not looking for
  - Rapid development lowers cost, makes ‘throwaways’ attractive
    - Leads to innovation

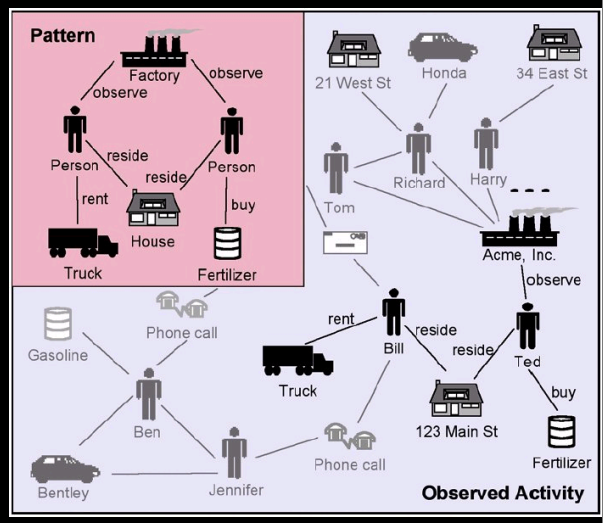
Network Analysis Plugins

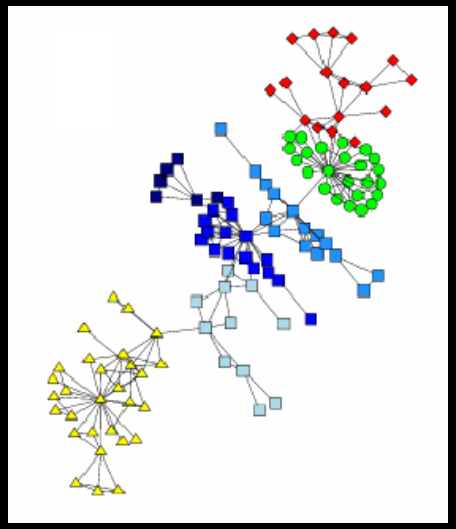


- Common Elements: Graphs, Semantic Analysis

Scalable Graph Algorithms Provide Crucial Capabilities

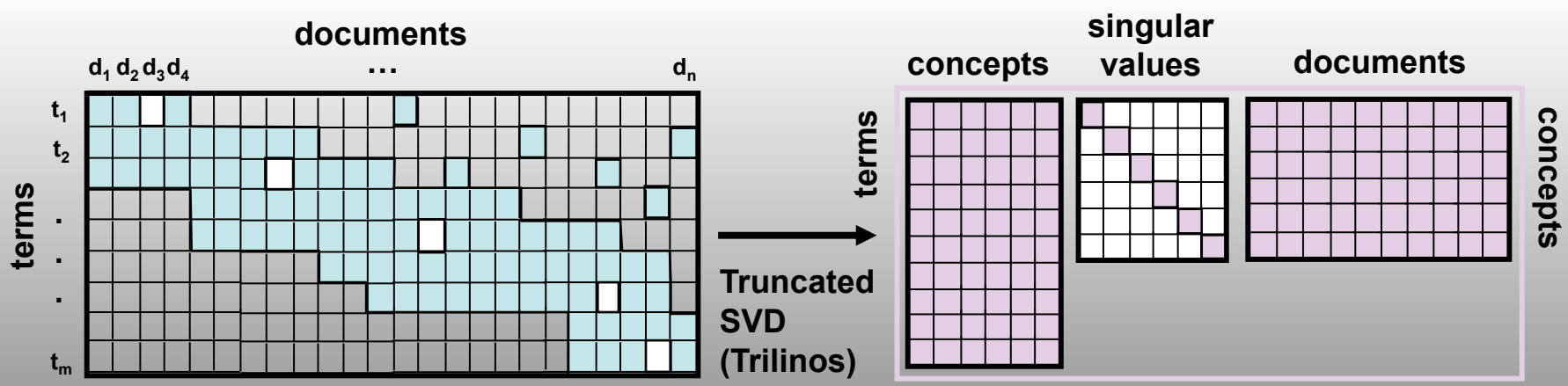
- Graphs are a unifying abstraction for much work in informatics
  - Model entities and relationships
  - Communication, control, financing, etc.
- Scalability and high performance are critical
- Only DOE discrete math research department
  - Long history of research in graphs and HPC
  - Historical focus on scientific computing applications, but current emphasis on graph models of information
- Chaco, Zoltan and MTGL open source graph toolkits
- Recent focus on alternative computer architectures
  - Handful of XMT processors can compute short paths as fast as 32,000 processors of BlueGene/ Light



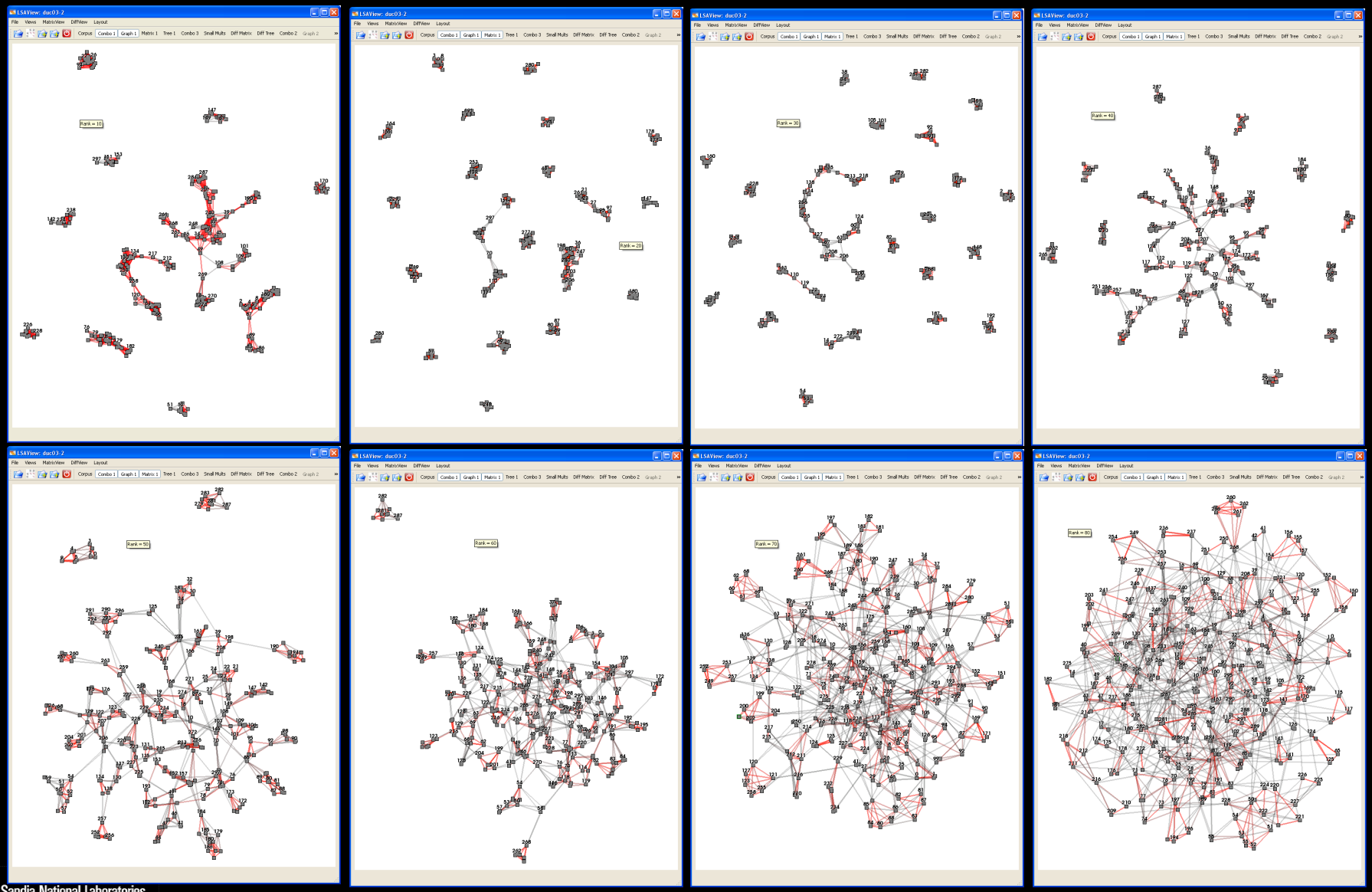


Latent Semantic Analysis

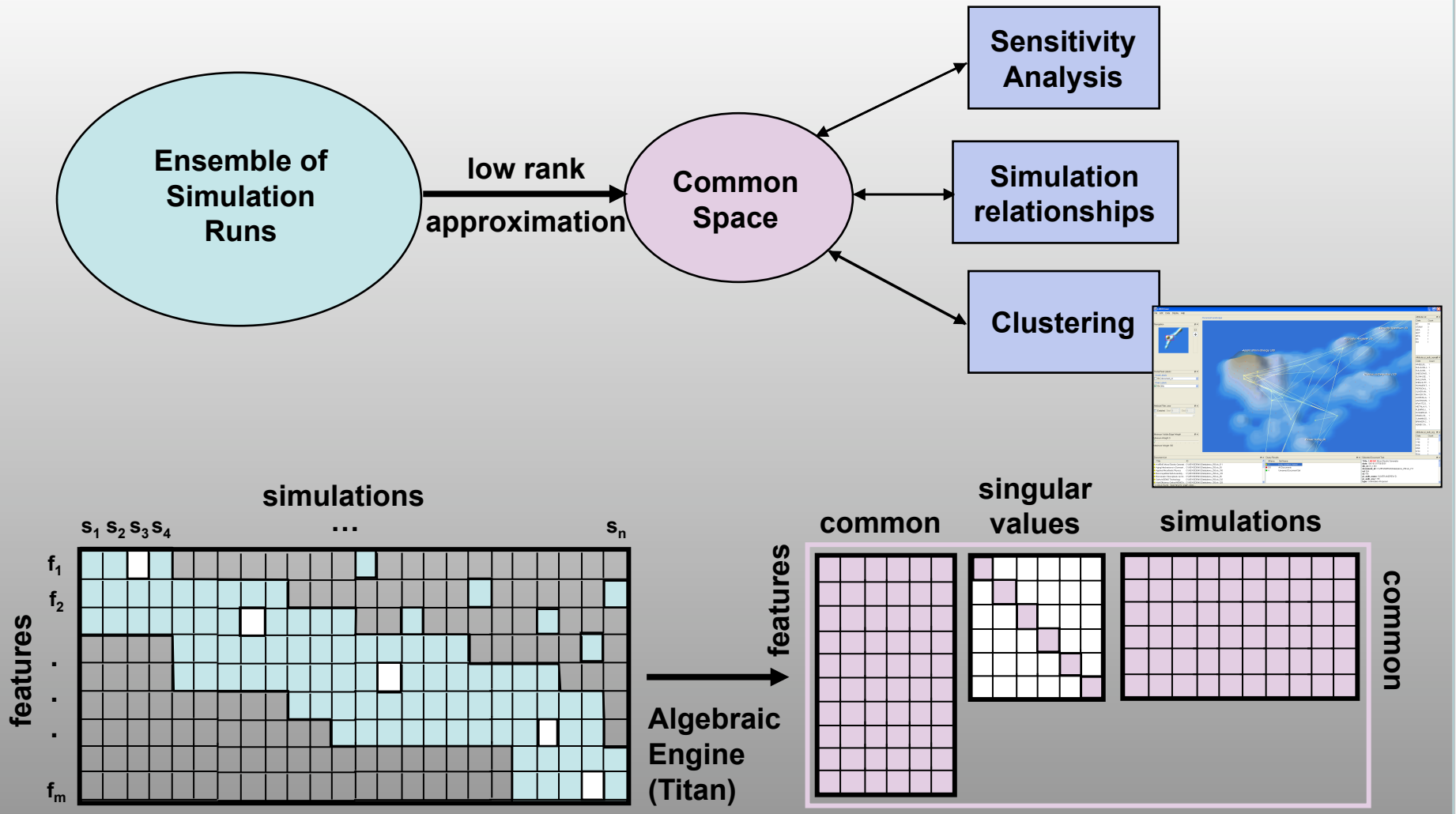
- "Scalable Ingestion"
  - Help me understand a large corpus of inputs (documents)
    - Compare documents in concept space (clustering)
    - Relationships between terms
    - Conceptual Queries
  - Implemented within the scalable framework
    - Utilizes Trilinos within Titan
- Tightly coupled viz project provides tools for evaluating algorithm



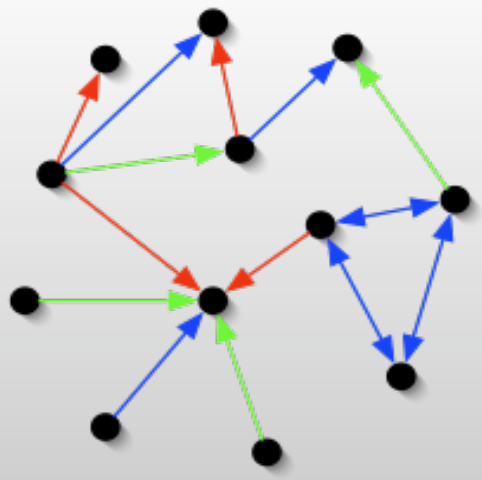
Which model to use?



ParaSpace Applies LSA to n-dimensional features in any results

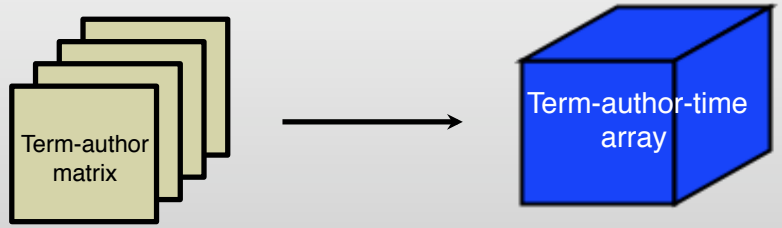


Algebraic algorithsm show great promise in addresssing informatics challenges

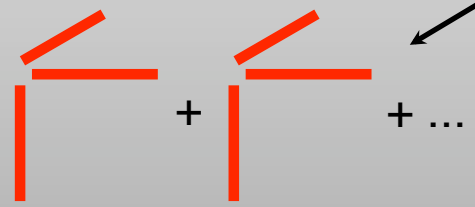


- Semantic Graphs - Edges have labels
- Examples
  - Email communications
  - Network traffic

"Multidimensional Data Mining"
Build a 3-way array such that there is a term-author matrix for each time period.



PARAFAC (Parallel Factors)



Third and higher dimensions offer more explanatory power: uncovers new latent information and reveals subtle relationships

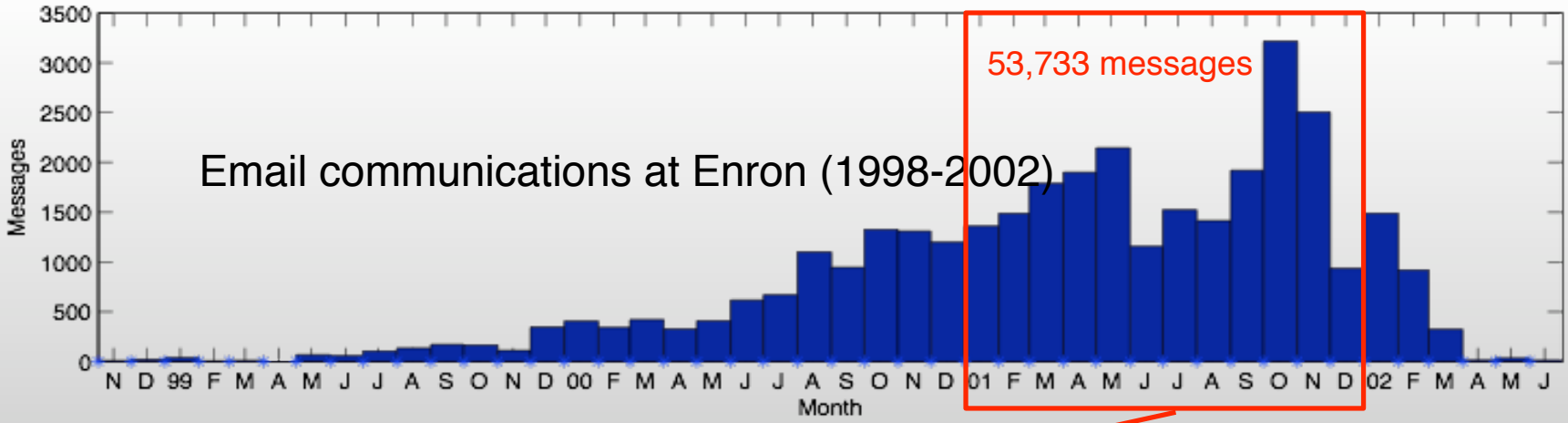
What can we learn from these multidimensional arrays?

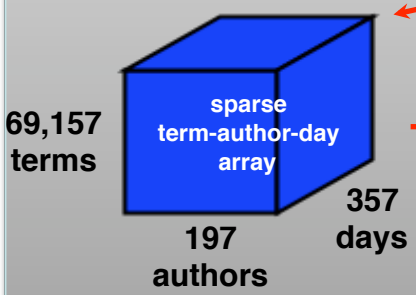
What are the major discussion topics?

Who are the major participants?

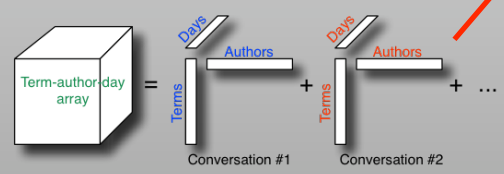
When are they taking place?

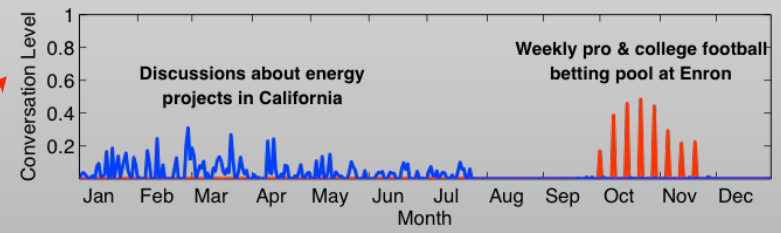
New techniques applied to the Enron email corpus reveal unexpected activities



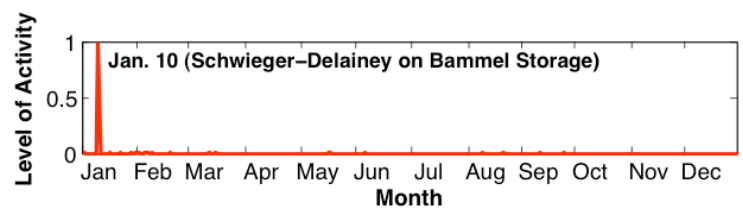


~5 Billion Cells
1,770,233 nonzeros

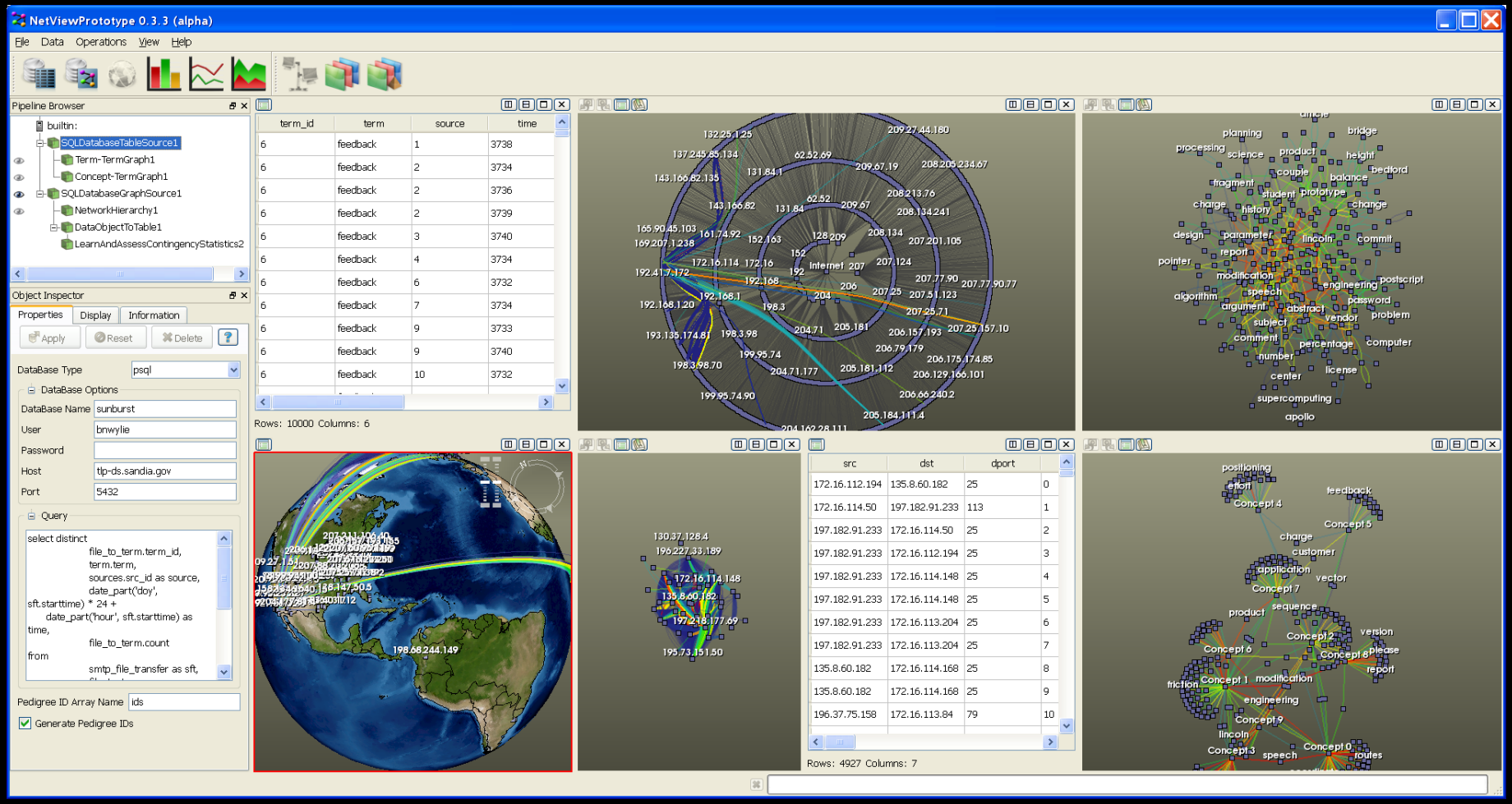




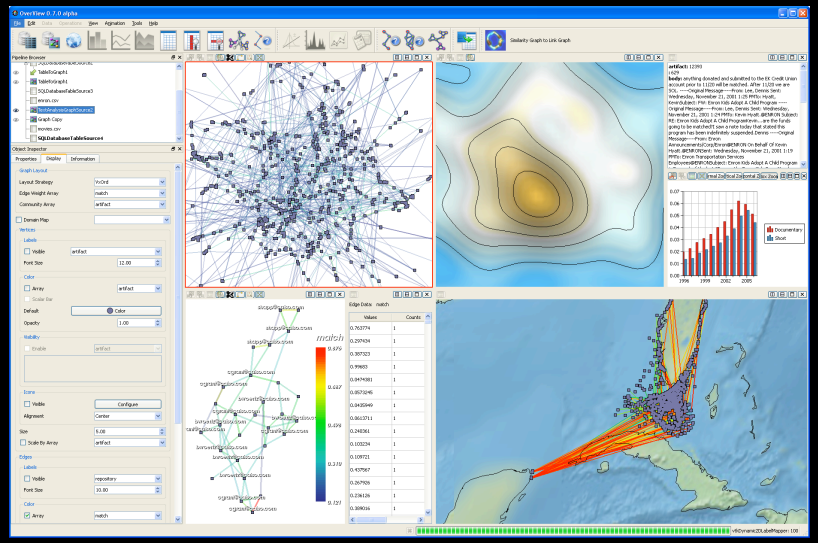
- Four-way tensor exposed conversation confirming bank fraud related to the natural gas reserves in the Bammel Storage field (Texas)—"The Enron whistle-blower who wasn't" by G. Farrell, **USA Today**, Oct. 11, 2007



Network Analysis with Plugins



Titan Toolkit Integrates Scalable Algorithms into an End-User Tool

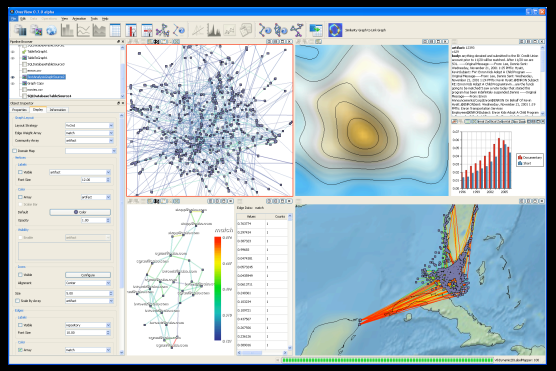


Distributed Memory Machine (RedStorm)	Massively Multithreaded Machine (XMT)	HPC Database (Netezza)	…

• Distributed Sparse Matrix (Trilinos)
• Some graph algorithms
• Statistical analysis
	• Graph Algorithms
	  • Shortest Path
	  • Isomorh. queries
		• Data Queries (SQL)
		• FPGA-assisted writes
		  • compression
		• FPGA-assisted extraction
		• FPGA-whatever	

Status

- Titan components are available in VTK now
- OverView 1.0a available
  - Basic capabilities in a number of areas
- Scalable algorithms available on a per-project basis
  - Trilinos available open source
  - MTGL available open source



Application: http://www.sandia.gov/OverView
Tim Shead, tshead@sandia.gov

Toolkit: http://www.sandia.gov/Titan
Brian Wylie, bnwylie@sandia.gov

Thank you

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000