SAND2010-4204P

# Data Analysis in the Networks Grand Challenge LDRD

***Brett Bader***
Senior Member of Technical Staff

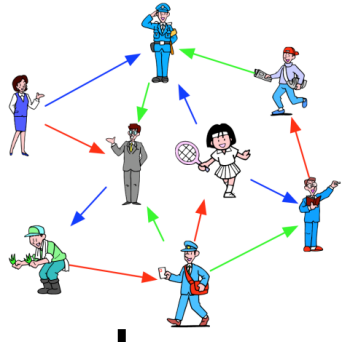
June 24, 2010

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.

# Robust data analysis requires appropriate data abstractions and algorithms

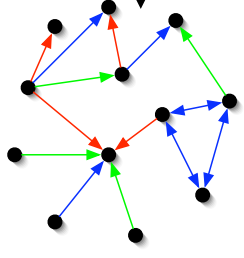
**Sandia uses *semantic graphs* and *tensors* as unifying data abstractions**

- Supports rich relationship-centered analysis
- Combines large, heterogeneous data corpora
- Different abstractions support different analytics

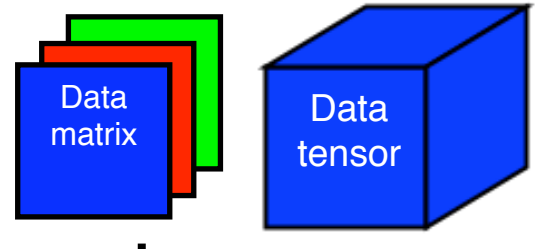


Real world networks
- Social networks
- Cyber traffic
- etc.

Raw, unstructured text
- Newspaper articles
- Web documents
- etc.



Semantic Graph

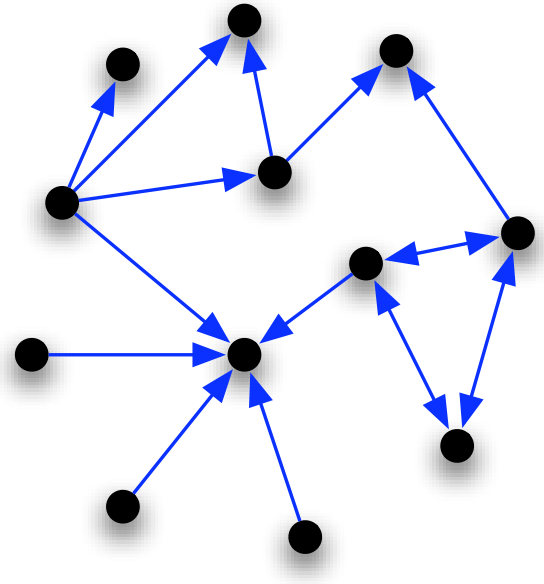



Matrices & Tensors

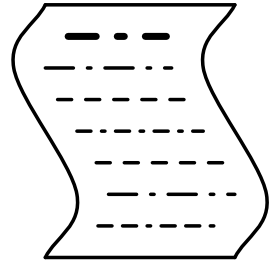
- Graph algorithms, discrete math
- Short paths, connection subgraphs, subgraph isomorphism

- Linear and multilinear algebra, statistics/probability
- Ranking, clustering

# Traditional Analysis



or



Data matrix

Best rank-*k* matrix filters out noise and captures "latent" information, which improves certain data mining tasks

Truncated SVD

Examples:

- Latent Semantic Analysis
- Text Analysis (LSI)
- Web search (HITS)
- Clustering

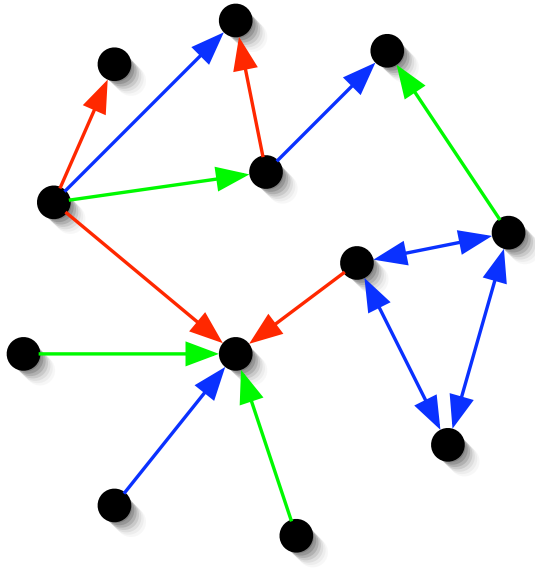
$U_k$ $\Sigma_k$ $V_k^T$

$$A_k = U_k \Sigma_k V_k^T = \sum_{i=1}^{k} \sigma_i u_i v_i^T$$

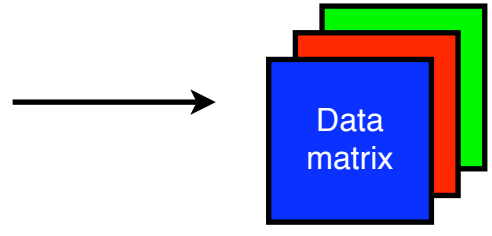
But there may be more useful information in the data!

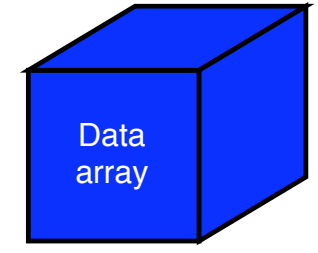
Sandia National Laboratories

# New Paradigm: "Multidimensional Data Mining"

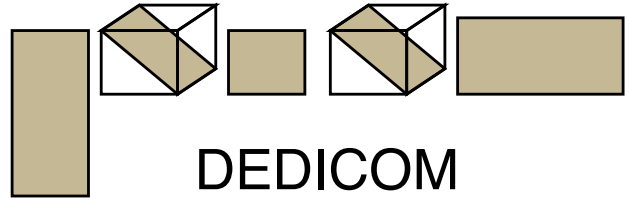


Build a "data array" such that there is a data matrix for each link type.

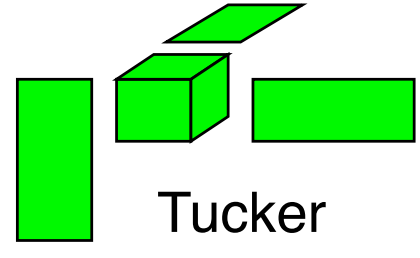


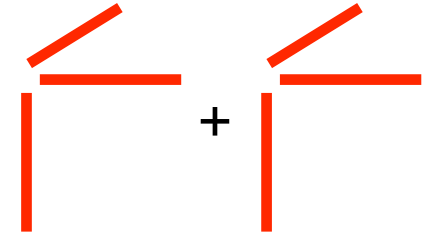

Multilinear algebra



DEDICOM



Tucker




PARAFAC

Third dimension offers more explanatory power: uncovers new latent information and reveals subtle relationships

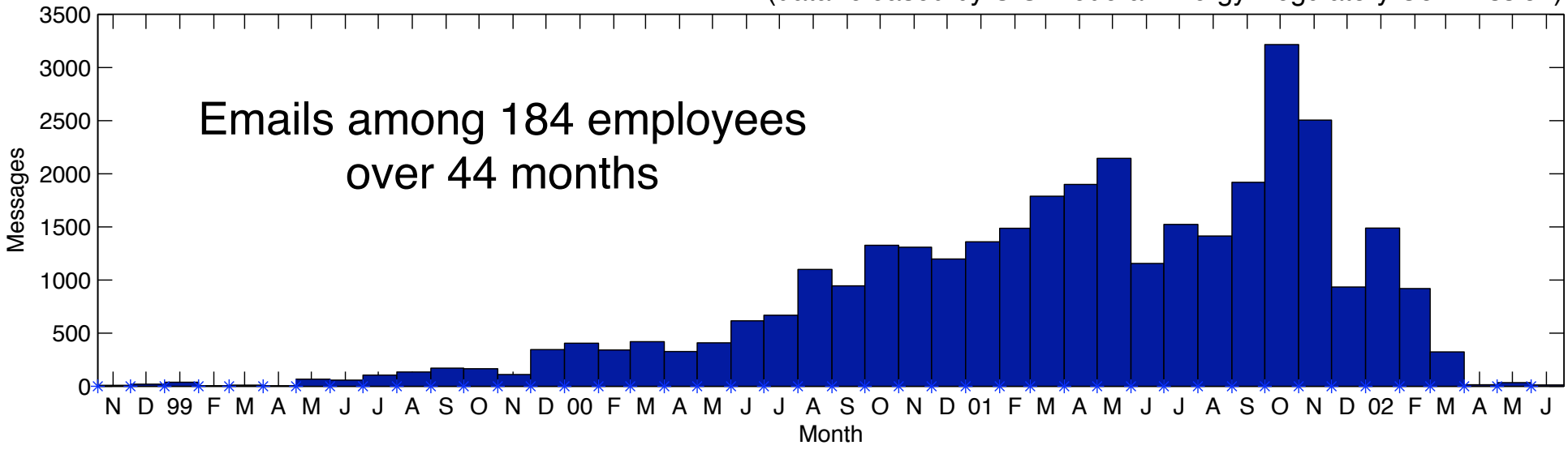
Unique data mining capability developed at Sandia

Sandia National Laboratories

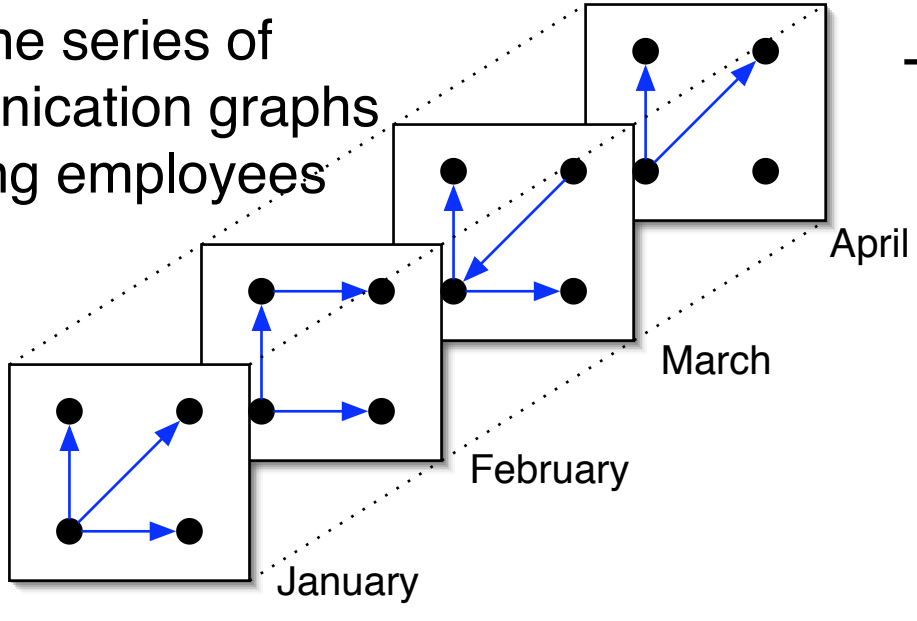
# Case Study:
# Pattern Analysis in Email Networks

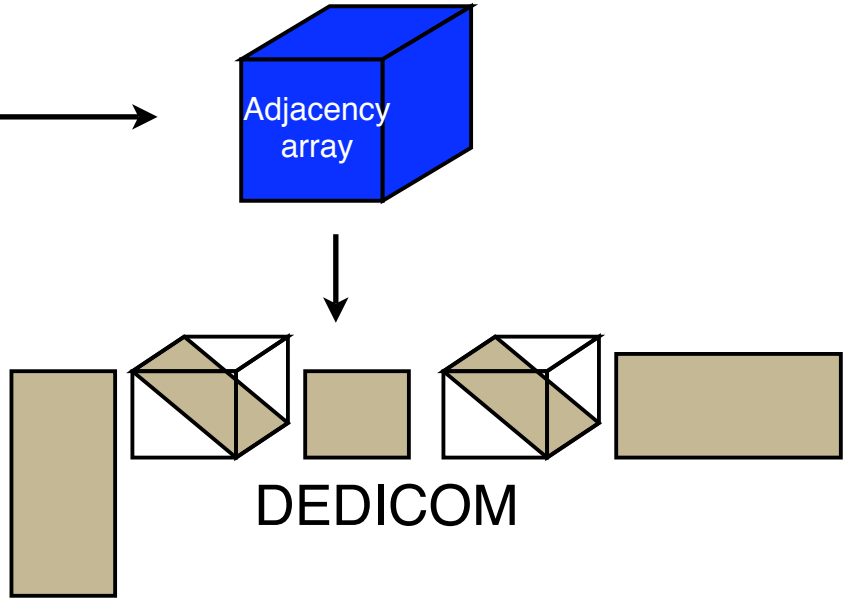
Email communications at Enron (1998-2002)

(data released by U.S. Federal Energy Regulatory Commission)



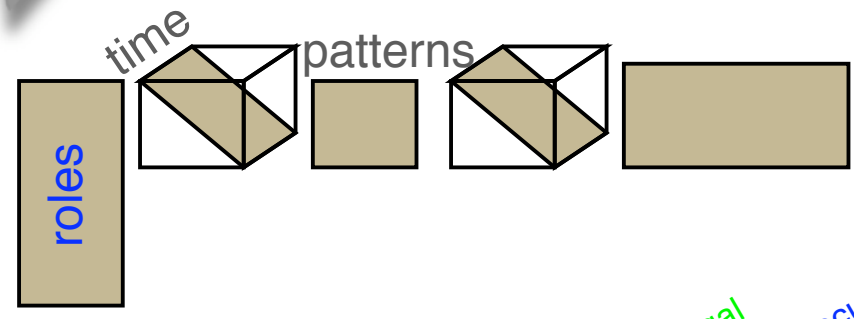

Time series of communication graphs among employees



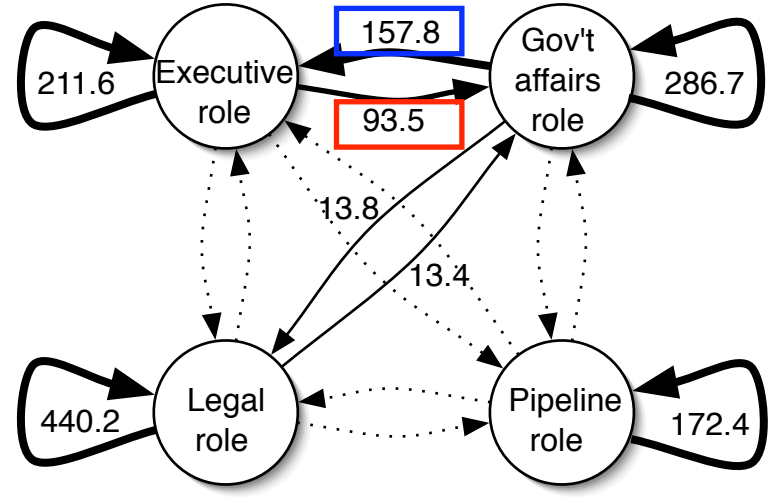

Sandia National Laboratories

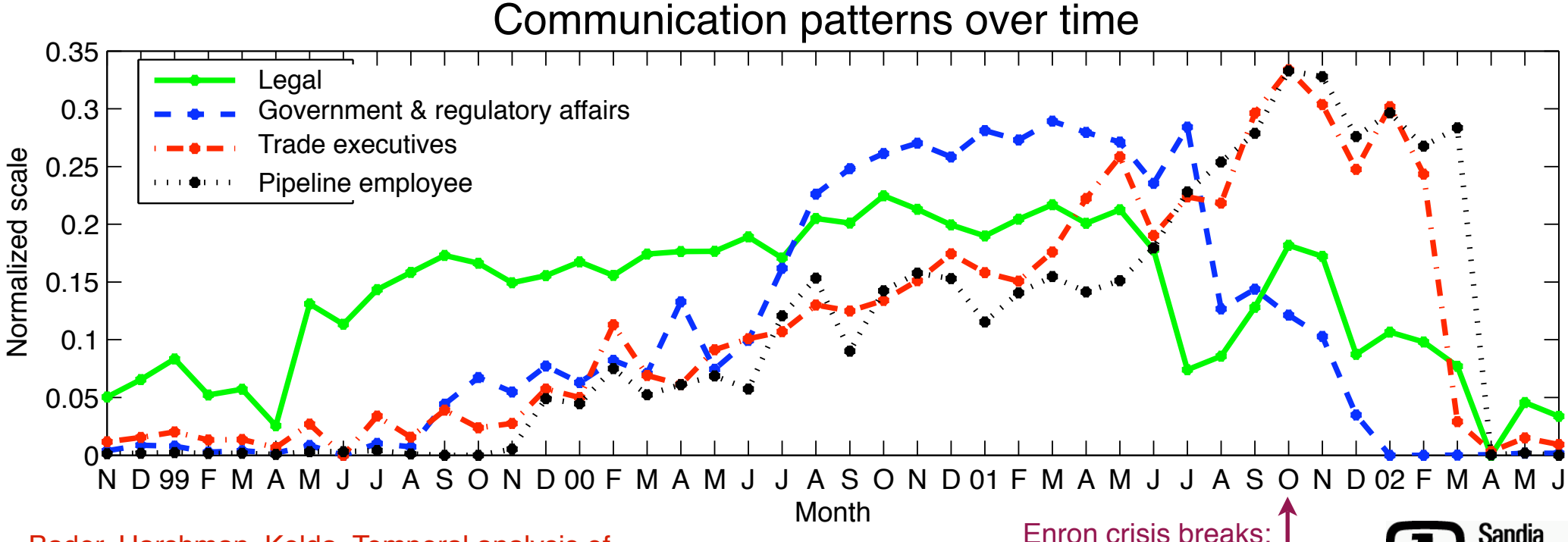
# Analysis shows employee roles and communication patterns




| | Legal | Executive (gov't affairs) | Executive (trade) | Pipeline |
|---|---|---|---|---|
| L. Kitchen - President, Enron Online | **0.11** | -0.09 | **0.53** | 0.00 |

Soft clustering



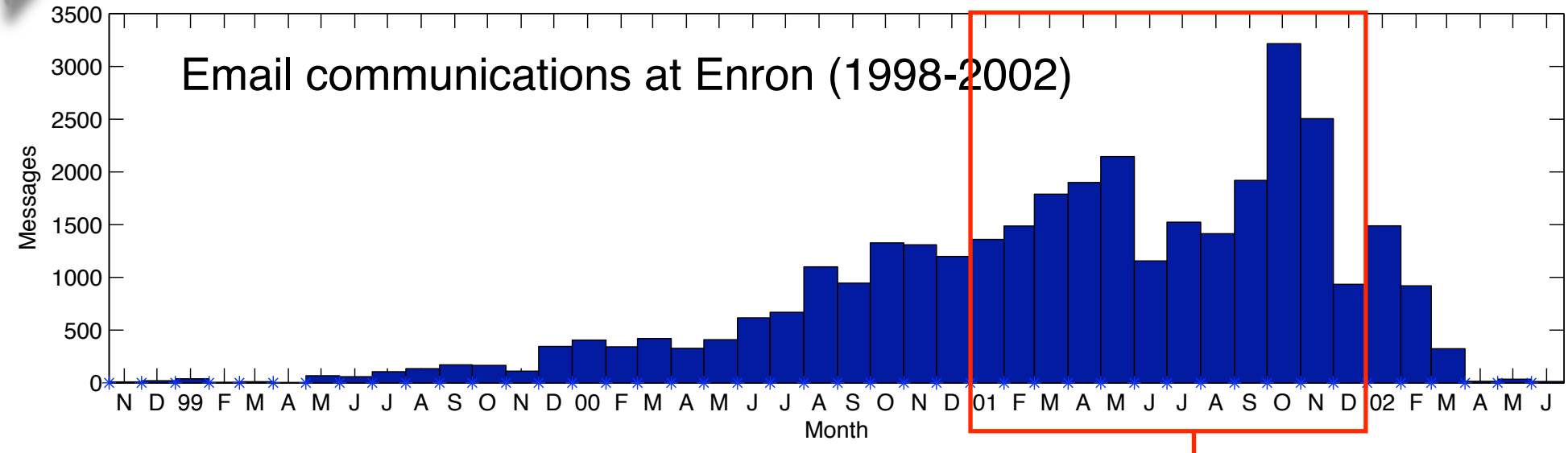




Bader, Harshman, Kolda, Temporal analysis of semantic graphs using ASALSAN, in ICDM 2008.

Enron crisis breaks; investigation begins

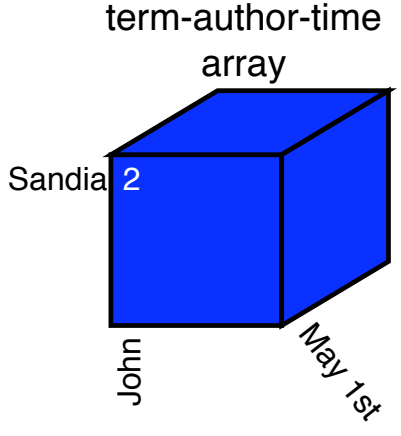
Sandia National Laboratories

# Case Study:
# Discussion Tracking in Email




53,733 messages
from 184 employees

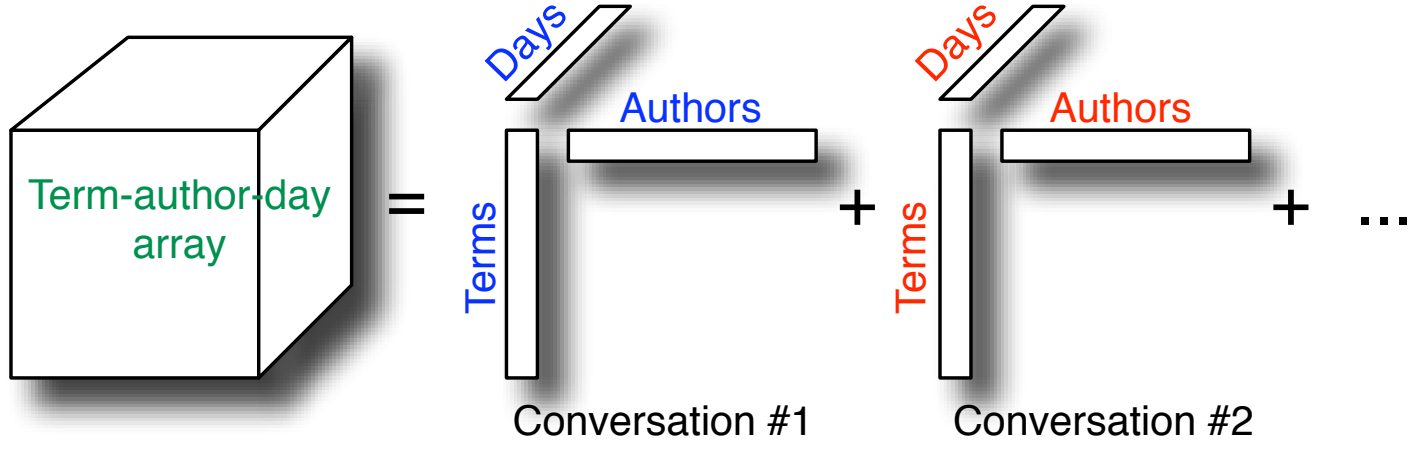
- Situational awareness
- What can we learn from these email conversations?
  - What are the major topics of conversations?
  - Who are the major participants?
  - When are they taking place?

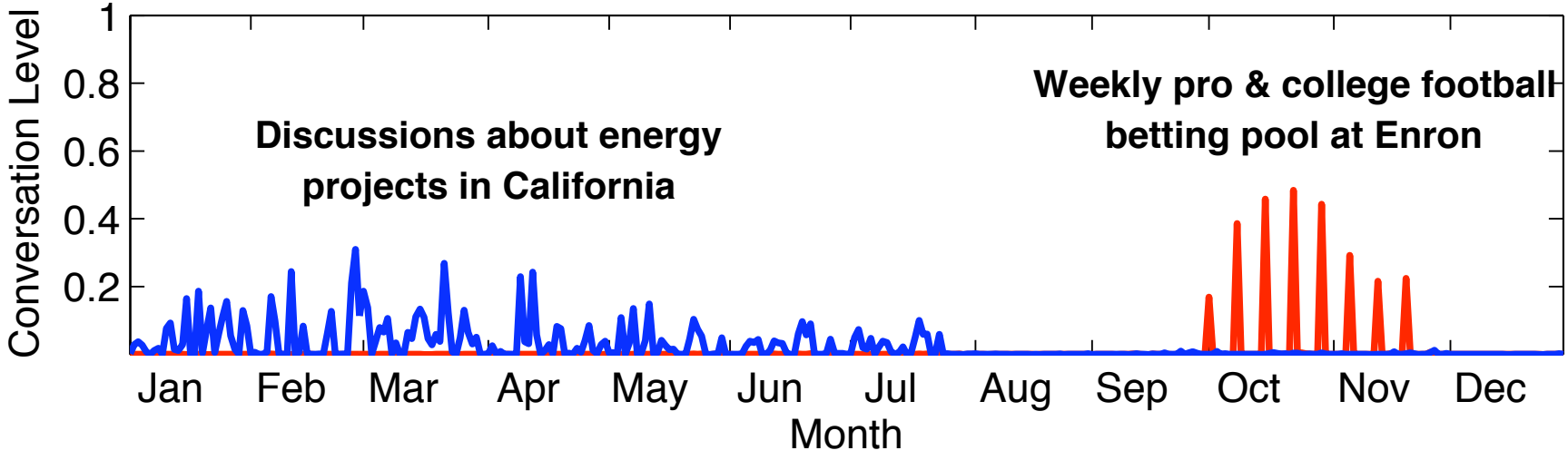



Bader, Berry, Browne, Discussion tracking in Enron email using PARAFAC, in Survey of Text Mining II, 2008.

Sandia National Laboratories

# Tensor analysis finds unusual activity by associating terms with people over time




Key terms: California, power, utilities, energy, utility, governor, market

games, week, missed, picked, prize, wins, scored, upsets

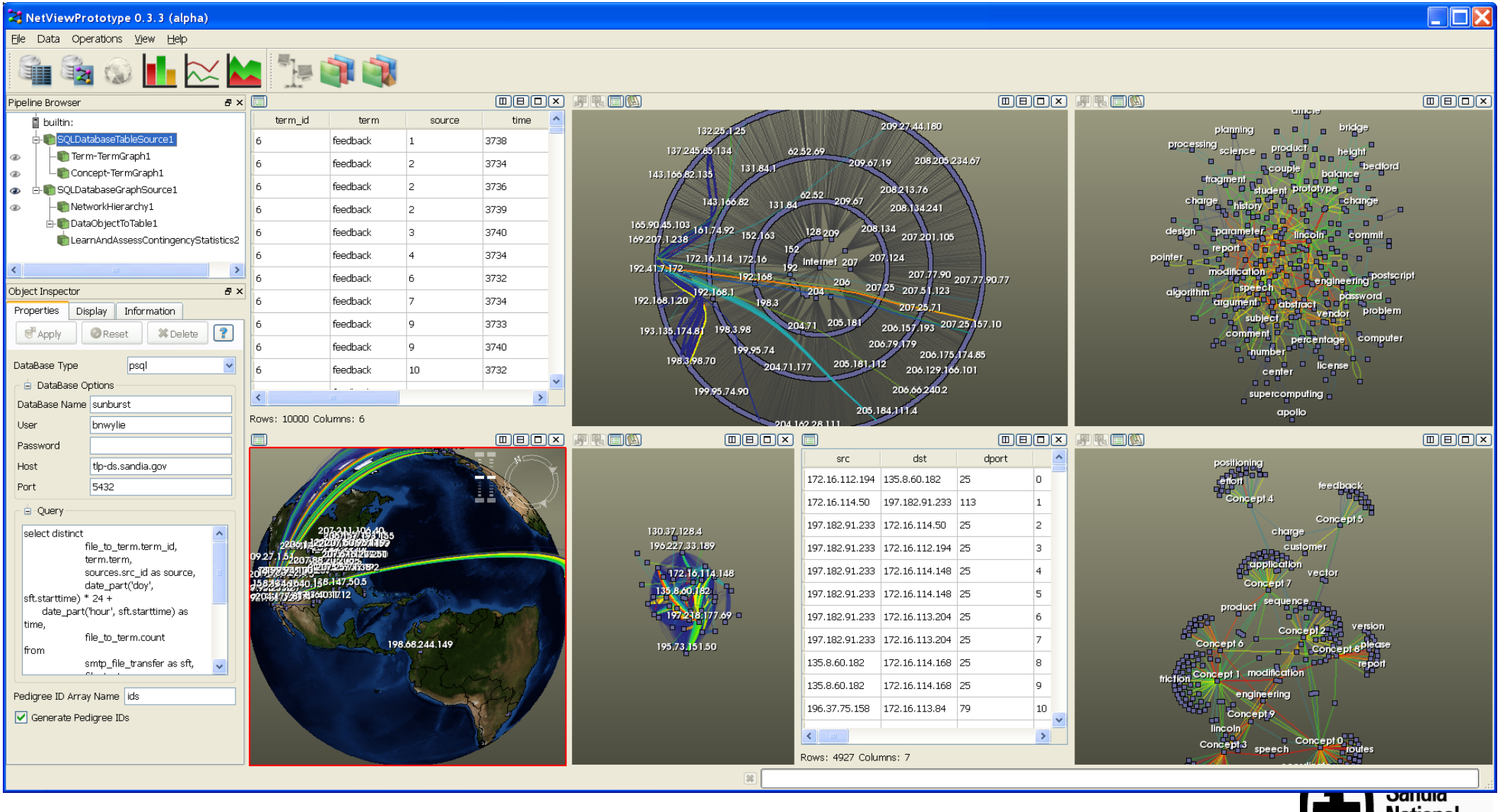
Key authors: J. Steffes, S. Kean, J. Dasovich, R. Shapiro, P. Allen, ...

A. Pace, L. Campbell, C. Dean

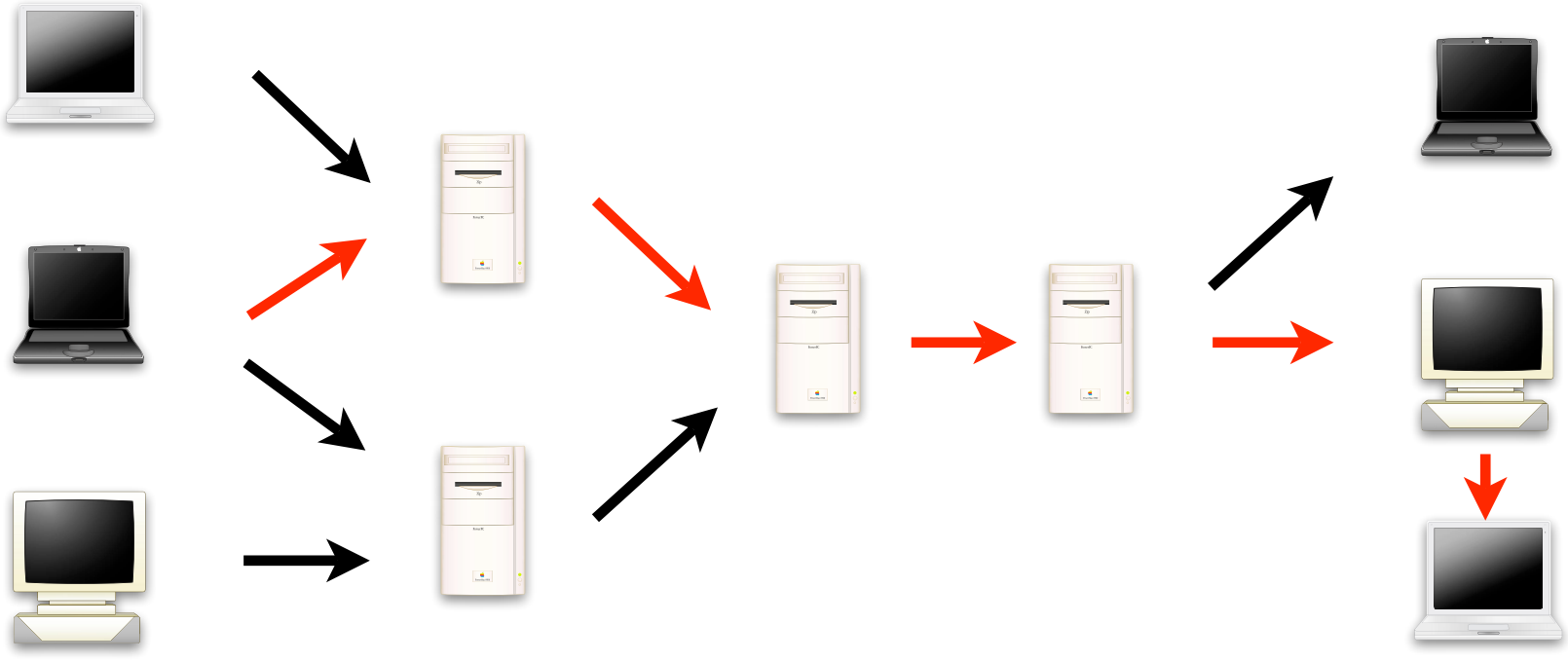
Sandia National Laboratories

# Analysis tools for deep packet analysis

- Prototype from the Networks Grand Challenge LDRD
  - Network traffic analysis
  - Content analysis

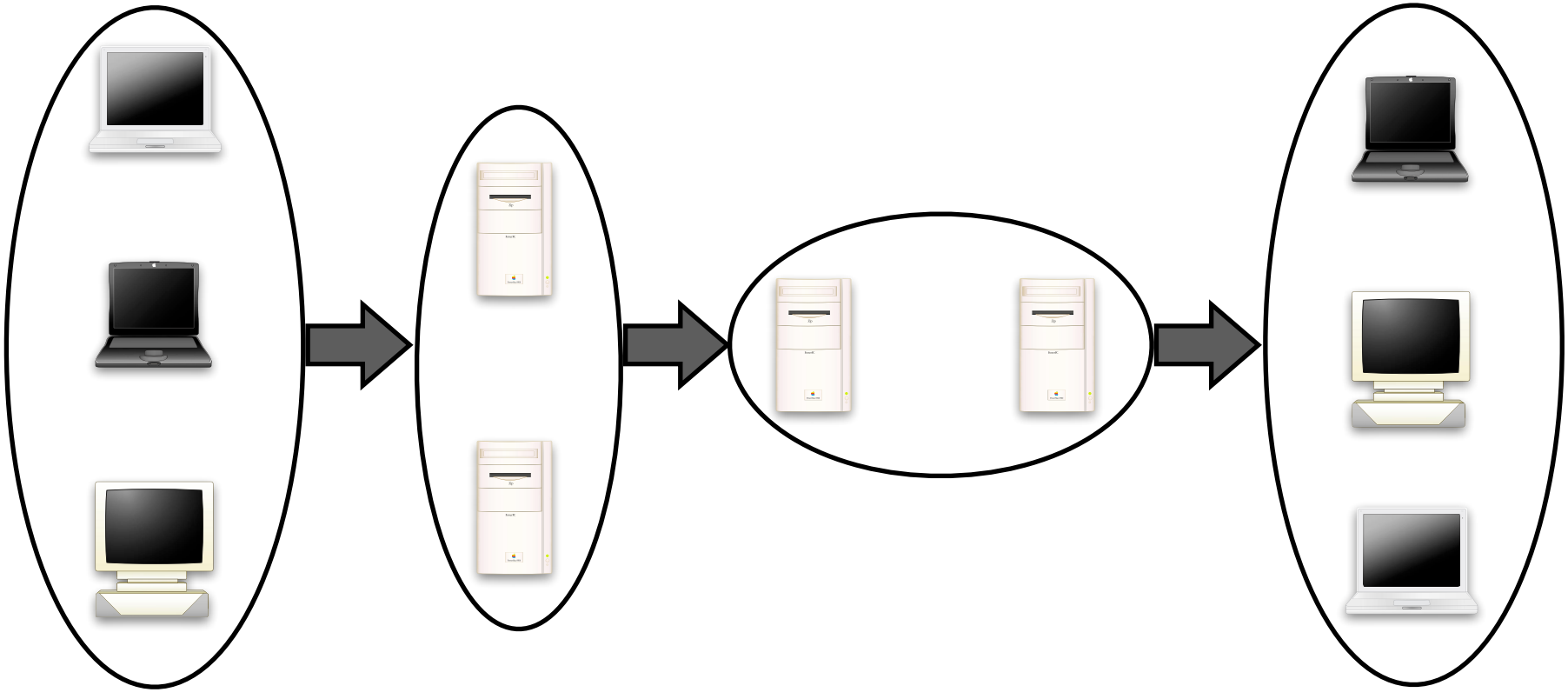


# Trace Route Analysis



- Analysis requested by Sandia's cyber analysts
- UNIX "`traceroute`" gives paths through network
  - multiple IPs per path
  - each path is directional
- Would like to know structure of network and general traffic patterns

# Trace Route Analysis



- Cluster of similar IPs based on connectivity
  - Hard or soft clustering
- Directionality of traffic between clusters
- Analysis may be used to identify “choke points”

# Sandia's statistical techniques have demonstrated unique ability to identify patterns and anomalies

## *VAST 2009 Challenge*

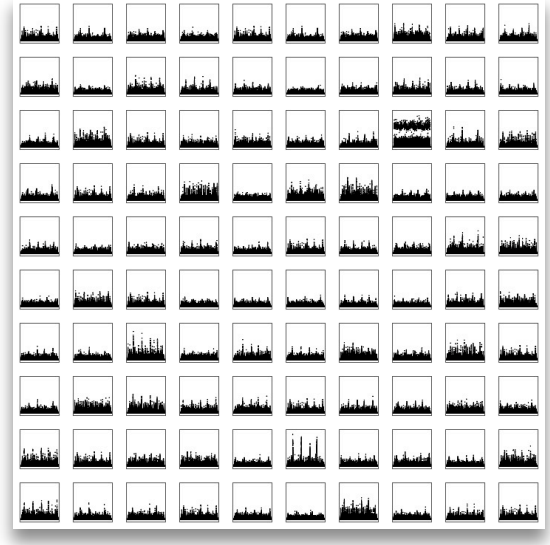
**Goal:** Identify an insider threat in a cyber environment.
**Data:** Header information from 115,414 network events over 1 month.
**Approach:** Use a probabilistic semantic analysis to identify unusual patterns.

| Source IP | Access Date/Time | Destination IP | Socket | Req Size | Resp Size |
|---|---|---|---|---|---|
| 37.170.100.38 | 01/01/08 09:40 AM | 37.170.100.200 | 80 | 7063 | 49591 |
| 37.170.100.38 | 01/01/08 09:43 AM | 37.157.76.124 | 80 | 5171 | 434285 |
| . . . | . . . | . . . | . . . | . . . | . . . |

Latent Dirichlet Allocation(LDA)

- Probabilistic version of a popular technique called LSA
- Statistical framework
- Soft clustering

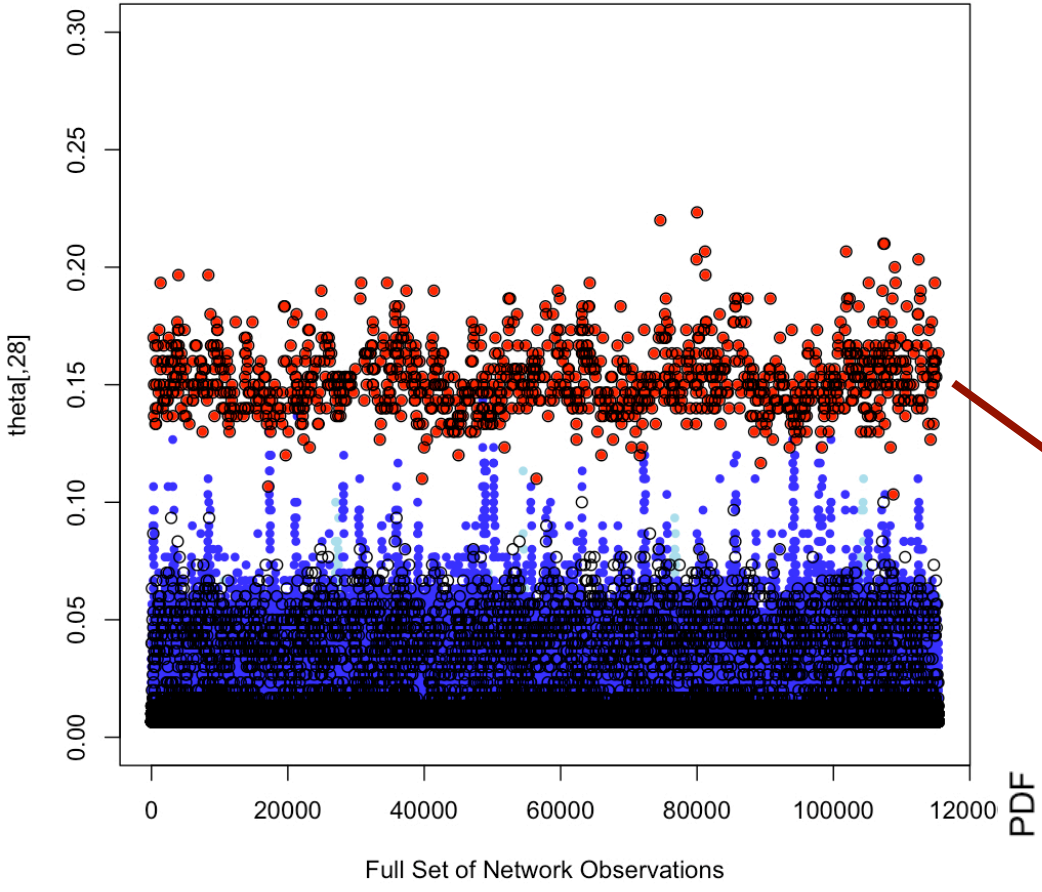
Network events clustered into 100 groups with non-exclusive membership



(Dave Robinson, 1415)

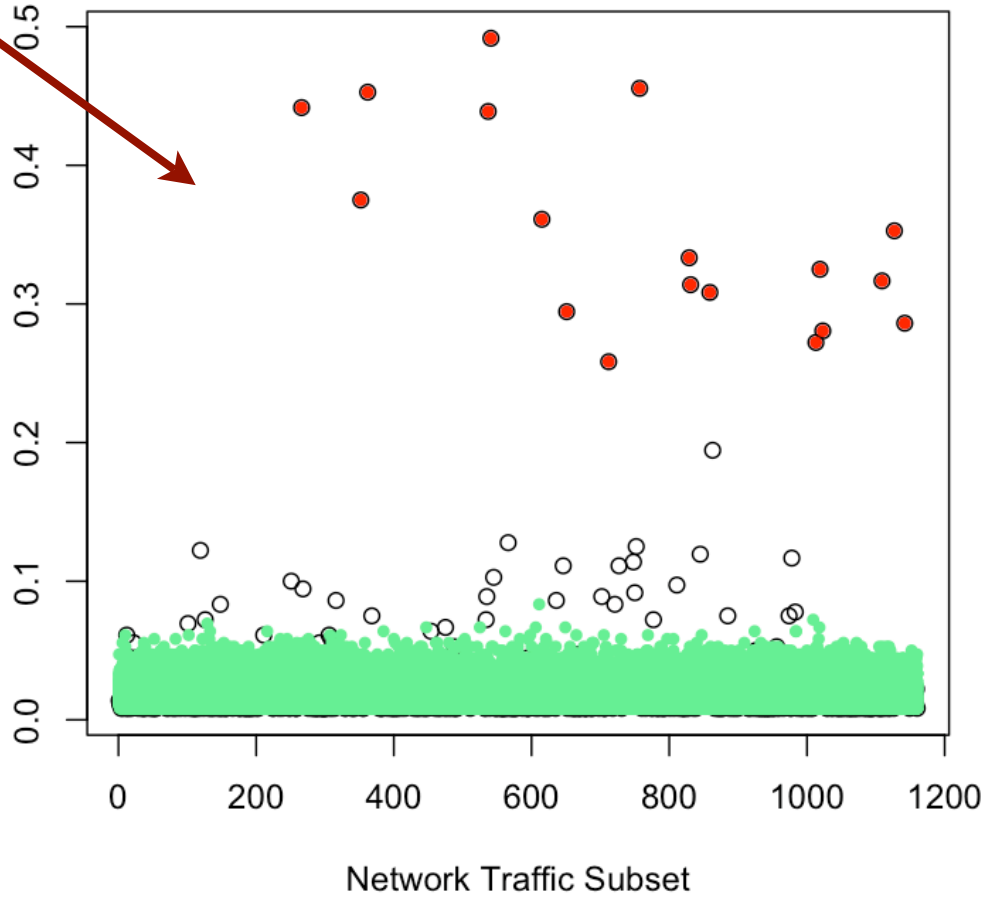
# VAST Contest Results

## LDA - First Pass




- 1st pass LDA: 1160 suspect events
- 2nd pass LDA: 18 events (the contest answer)

## LDA - Second Pass

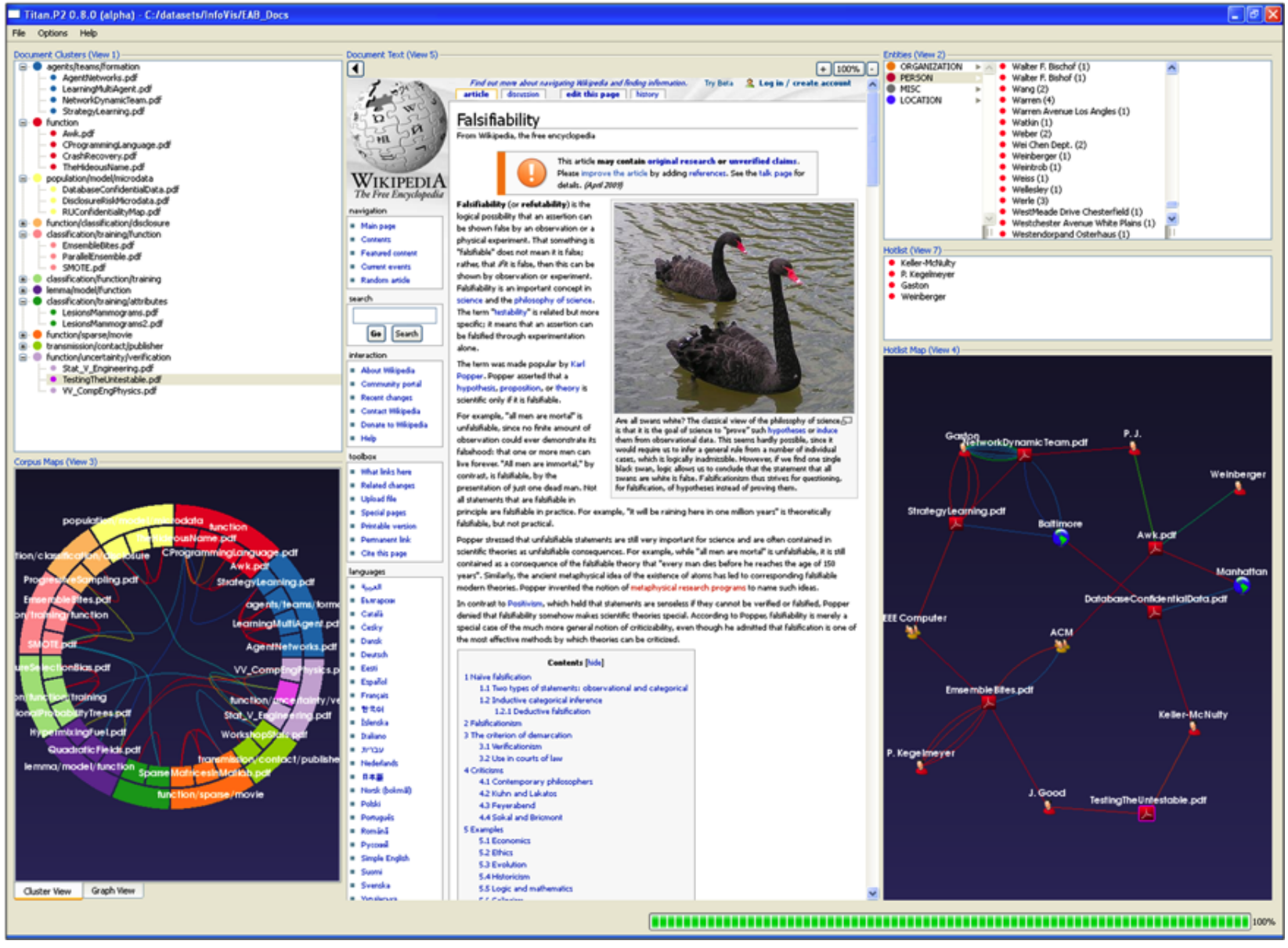



- By comparison, cyber analysts at Sandia narrowed it down to 80 events in 90 minutes
- Decision support tool to flag events that are not normal

(Dave Robinson, 1415)

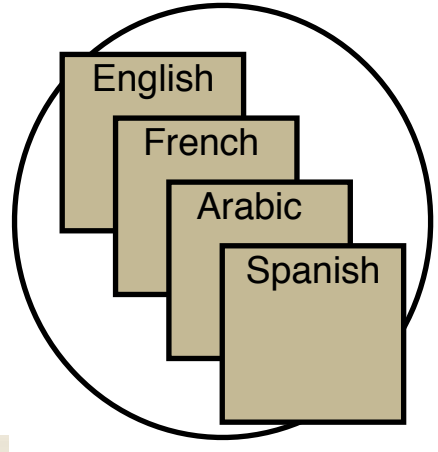
# Analysis tools for document/text analysis

- Prototype from the Networks Grand Challenge LDRD
  - Document clustering
  - Entity subgraphs



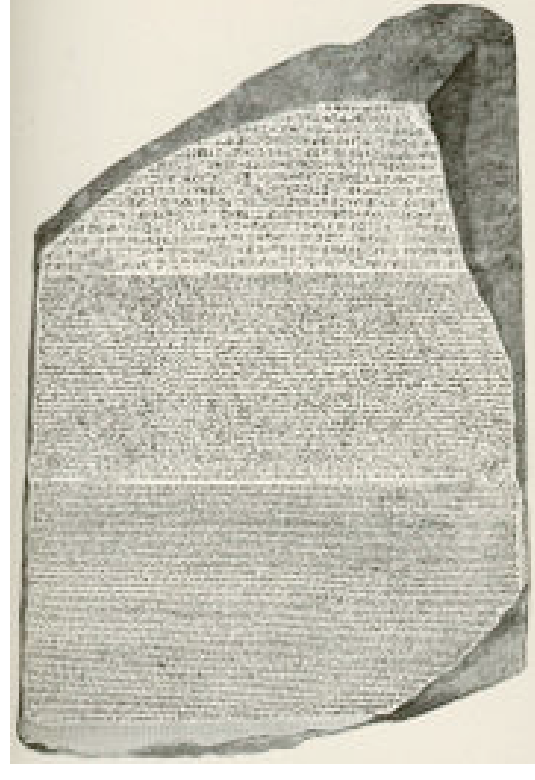
# SNL has developed multilingual text analysis to link threats across multiple languages

- "Translate" new documents into a language-independent concept space, which is useful for:
  - Translation triage (e.g., find all documents related to bomb making)
  - Multilingual sentiment analysis
  - Ideological classification (e.g., hostile to U.S.)



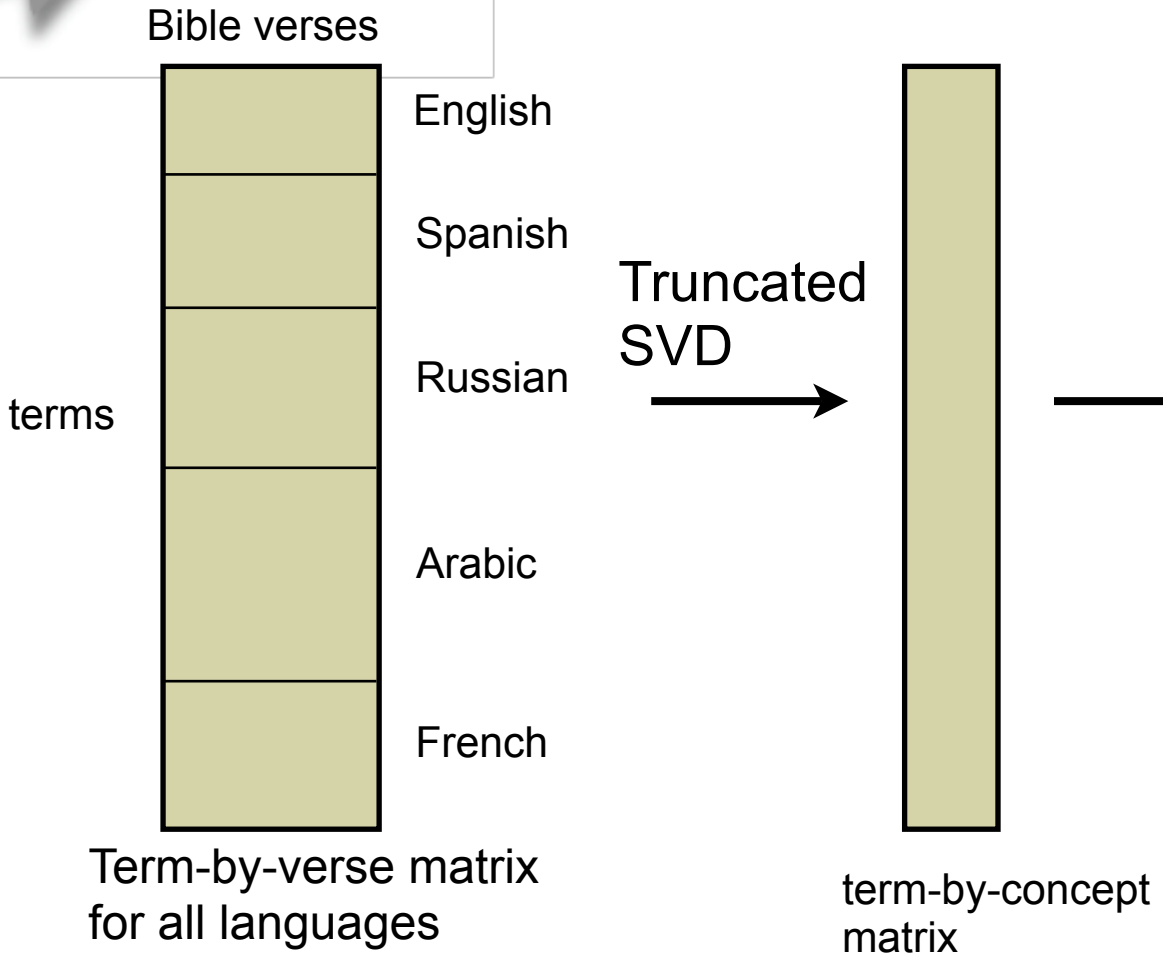

Sandia's database: 54 languages: 99.76 % coverage of web

| | | |
|---|---|---|
| Afrikaans | Estonian | Norwegian |
| Albanian | Finnish | Persian (Farsi) |
| Amharic | French | Polish |
| Arabic | German | Portuguese |
| Aramaic | Greek (New Testament) | Romani |
| Armenian Eastern | Greek (Modern) | Romanian |
| Armenian Western | Hebrew (Old Testament) | Russian |
| Basque | Hebrew (Modern) | Scots Gaelic |
| Breton | Hungarian | Spanish |
| Chamorro | Indonesian | Swahili |
| Chinese (Simplified) | Italian | Swedish |
| Chinese (Traditional) | Japanese | Tagalog |
| Croatian | Korean | Thai |
| Czech | Latin | Turkish |
| Danish | Latvian | Ukrainian |
| Dutch | Lithuanian | Vietnamese |
| English | Manx Gaelic | Wolof |
| Esperanto | Maori | Xhosa |

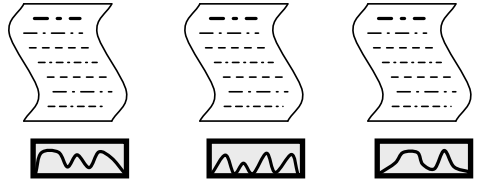


Sandia National Laboratories

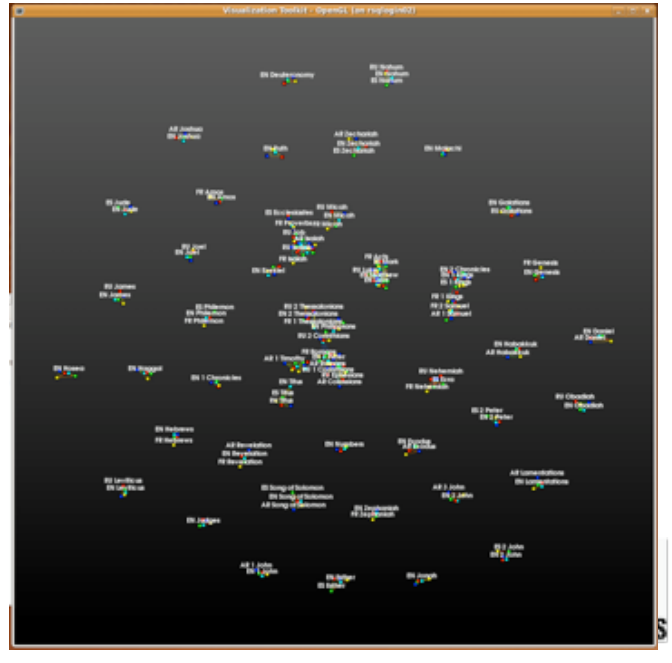
# Multilingual Latent Semantic Analysis




"Translate" new documents into a small number of language-independent features



Books of the Bible clustered by topic regardless of language



Selected applications:

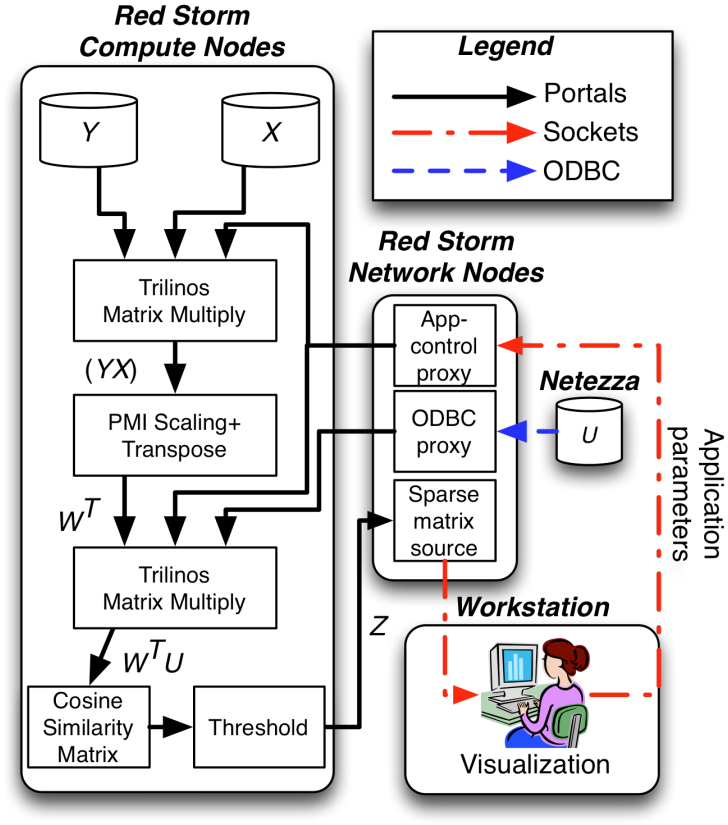
- Clustered 100k documents of European Parliament (working toward 1M documents on Red Storm)
- Identified web documents with a hostile ideology

# Architectural Challenges

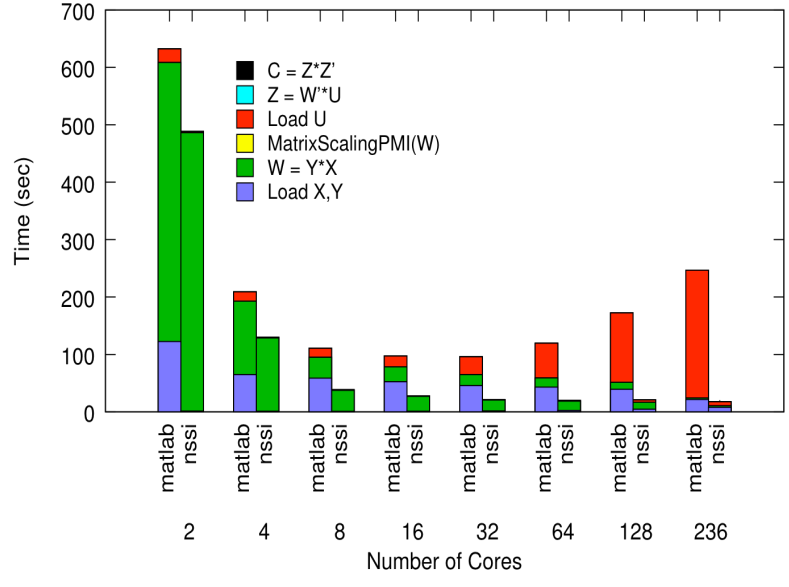
Exploiting specialized architectures

- Red Storm for numerics
- Clusters/Workstations for vis and interactive control
- Data Warehouse Appliances for database functionality

***Integrating these systems for interactive jobs has never been done***




Performance Results: Bible Dataset

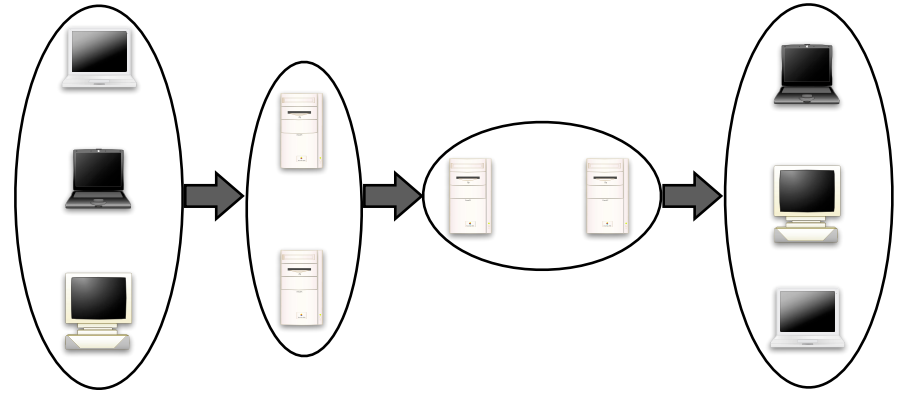



Sandia National Laboratories

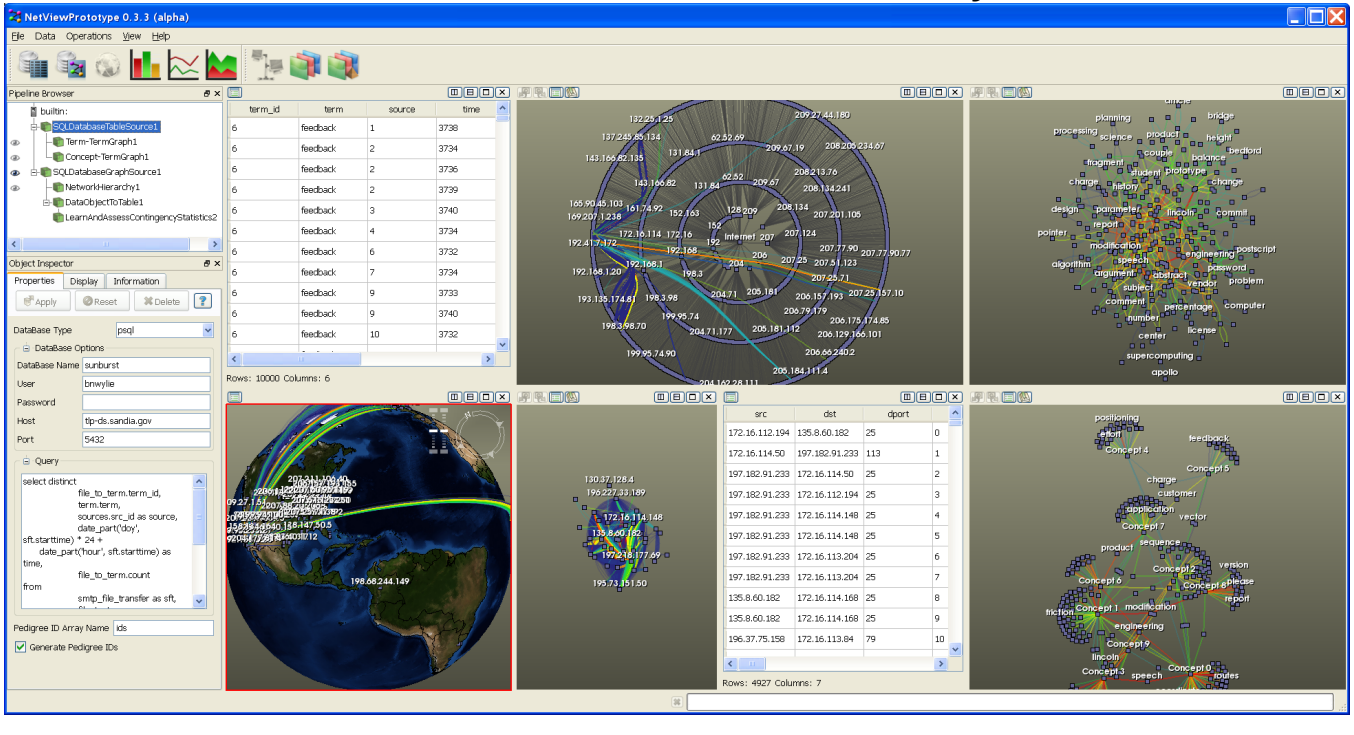
# Related Analysis Projects

- Multilingual document analysis and classification
- Uncovering plots buried in text (scenario discovery)
- IP address characterization (trace route analysis)
- Network traffic analysis (cyber, phone)
- Cyber data exfiltration analysis
- Link prediction
- Higher-order web link analysis

*Clustering nodes by their activity*



*Network traffic and content analysis*




Sandia National Laboratories