Sandia's Network Grand Challenge

Salishan 2010

April 27, 2010

David H. Rogers

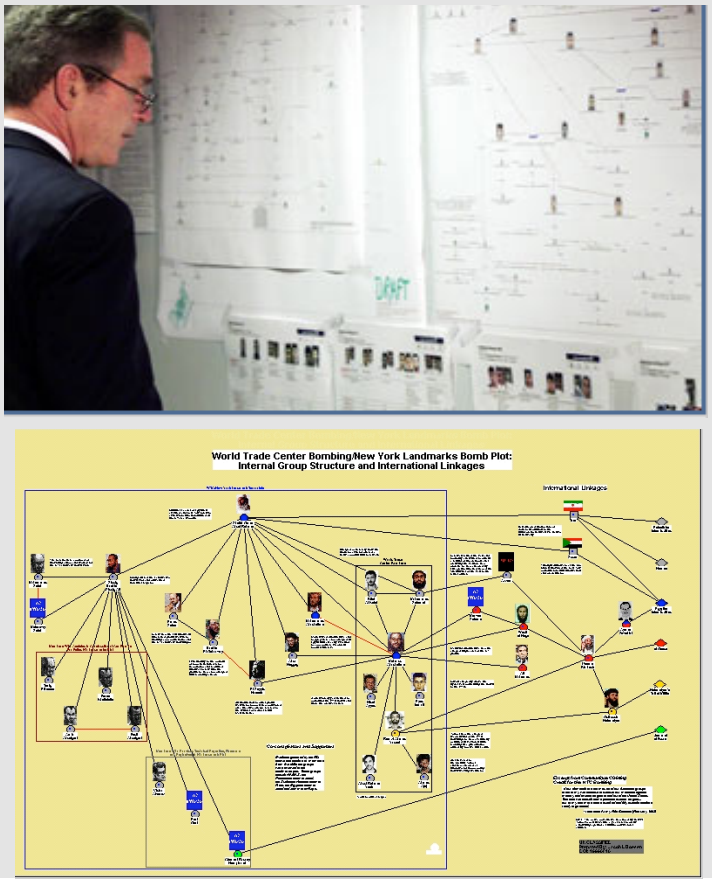
Manager, Dept 1424
Data Analysis and Visualization Department
Sandia National Labs

Sandia National Laboratories is a multi-program laboratory operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin company, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

What is the Networks Grand Challenge?

"our real adversaries are networks…"

- **Many national security threats come from loose, dynamic networks of people & organizations.**
  - **Facilitated by networks of finance, shipment, recruiting, smuggling, etc.**
  - **E.g., terrorism, proliferation, cyber, drug trafficking.**
- **Apprehending individuals or preempting events doesn't remove the threat.**
  - **Need means to discover and defeat the network.**
- **Individual tidbits of data look benign.**
  - **Only recognizable in larger context of related activities**
- **R&D gap in issues around scale and automation:**
  - **Lacking scalable methods for processing very large network graphs.**
  - **Need to find very faint signatures (e.g., 1013 bytes within 1013 data).**
  - **Batch-processing unacceptable (analysts need answers within seconds).**

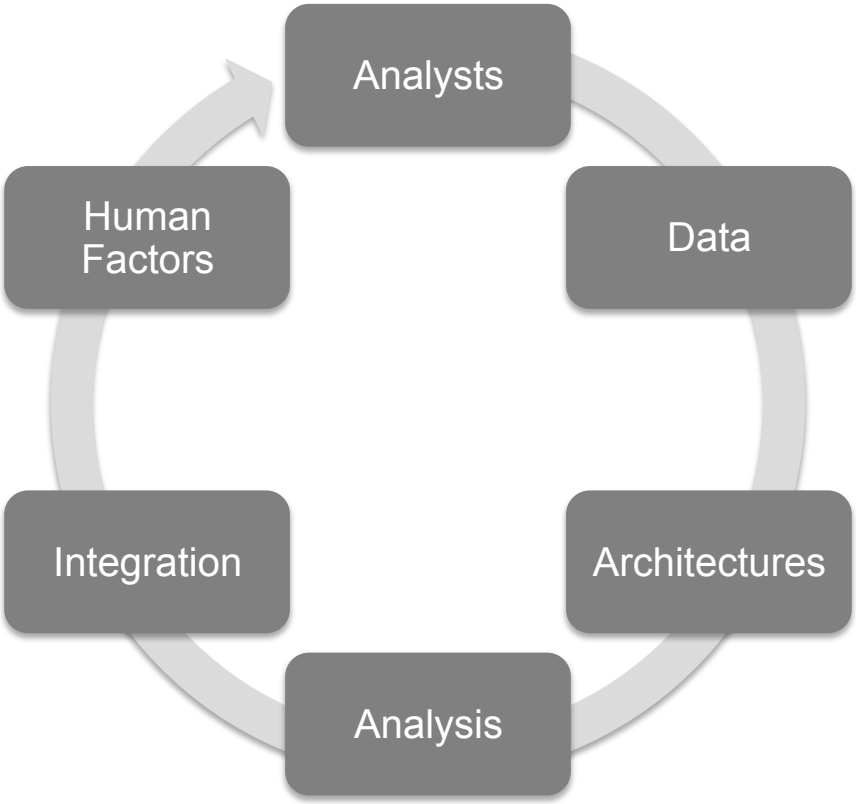


President Bush looks over a chart depicting Osama bin Laden's financial network; an excerpt from the chart (created using one of the leading tools, i2 Analyst's Notebook)

….and so, SNL is conducting R&D to yield a radical improvement on analytical methods and tools.

NGC Gathers Expertise from across the Labs, working on an End-to-end Approach

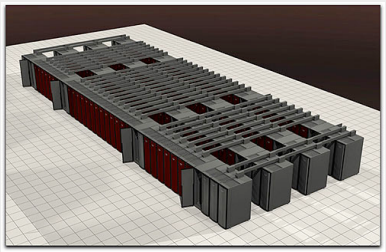
- **Data**
  - Provides data sets to the project, creates and supports the data infrastructure required by the project.
- **Architectures**
  - Develops software to efficiently integrate specialized hardware devices required by the research teams
- **Analysis (Discovery and Forecasting Teams)**
  - Develops capabilities relevant to the needle-in-a-haystack kinds of problems that concern intelligence analysts.
- **Integration**
  - Integrates NGC technologies and techniques into tools that are usable by analysts
- **Human Factors**
  - Responsible for all elicitation and knowledge representations, as well as software evaluation and assessment of technology impact on analyst performance.
  - Team of social and computer scientists interested in the relationship between human beings and computer technologies.



NGC System Diagram

"This project seeks to bring these two strengths – a solid reputation for excellence in computing, and our niche expertise in specific classes of intelligence analysis – to bear on a thorny problem: developing advanced informatics capabilities that are both usable and useful to analysts who are drowning in data." NGC project proposal

Architectures | Algorithms | Data | Web Services | Applications (Clients) ***Titan, browser***

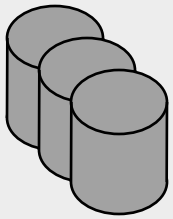


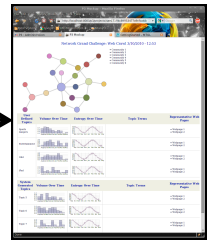
Trilinos
Algebraic Methods
Clustering, Ranking,
High Dimensional
Mapping

MTGL
Graph Methods
Subgraph searches,
Connection sg's,
Shortest Path, etc.

Specialized
Distributed Data
Operations

Titan
Analysis Pipelines,
Capability Integration,
Data Access,
Lightweight analysis

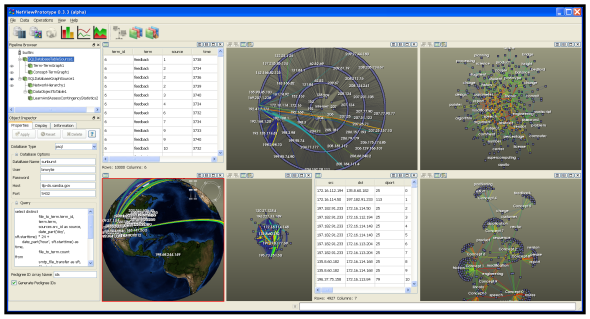




Titan
Analysis Pipelines,
Capability Integration,
Data Access,
Lightweight analysis

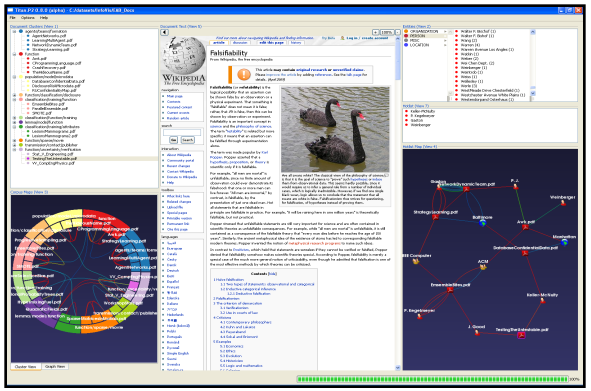
Highly optimized ⟷ Iterative, flexible

NGC's Commitment to Prototypes Promotes End-to-end Integration



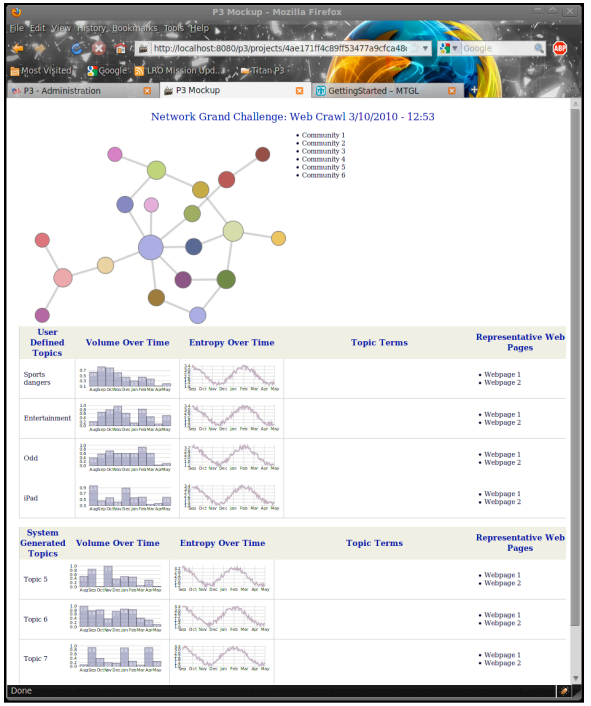
PI: Cyber Application

- Capability integration and demonstration



PII: Document Analysis

- Targeted development with analysts
- Iterative development
- Currently under consideration for funding



PIII: Web 'prediction'

- New approach: web services architecture, lightweight application in browser

There are Specific Reasons to Use HPC

- Iterative questioning in *Analyst Time*
- "Firehose, Stopwatch, or Dump truck" problems
  - Data Constraints
  - Time Constraints
  - Complexity of the query

We are Mauled by Data

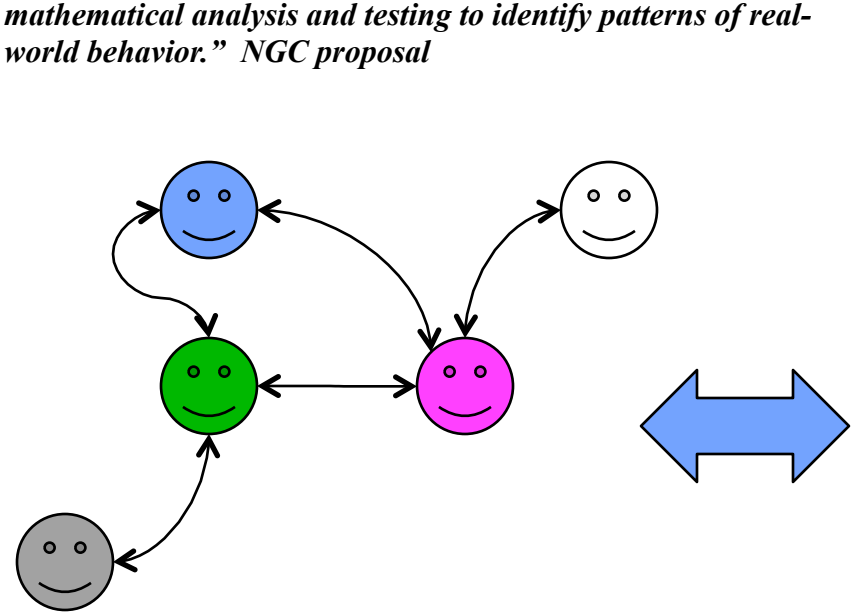
- 'Data Tsunami', etc. doesn't really capture it
- Need
  - Connect the dots
  - Find the needle in the haystack (needle in a pile of needles)
  - Find 'interesting events/connections
    - Signature detection relies on knowing what you're looking for
  - Find what I'm not looking for

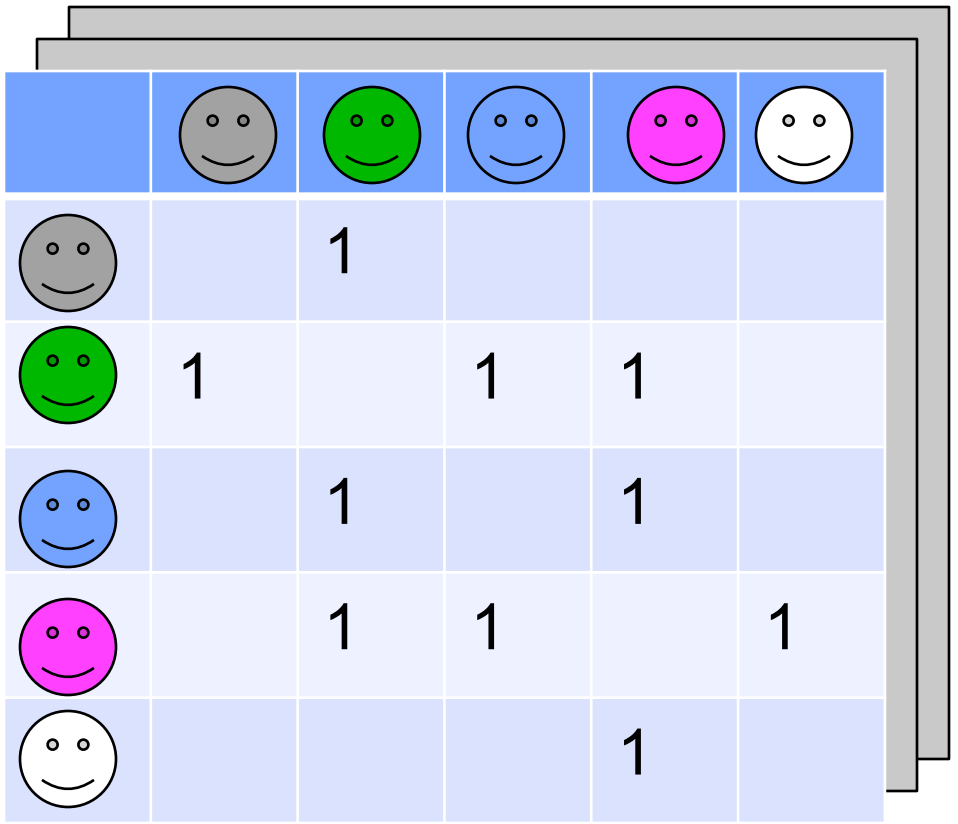
Can we develop methods that extract interesting behavior, without knowing what we're looking for?

NGC: Unifying Data Abstractions

Semantic Graphs and Tensors are Sandia's Unifying Data Structures

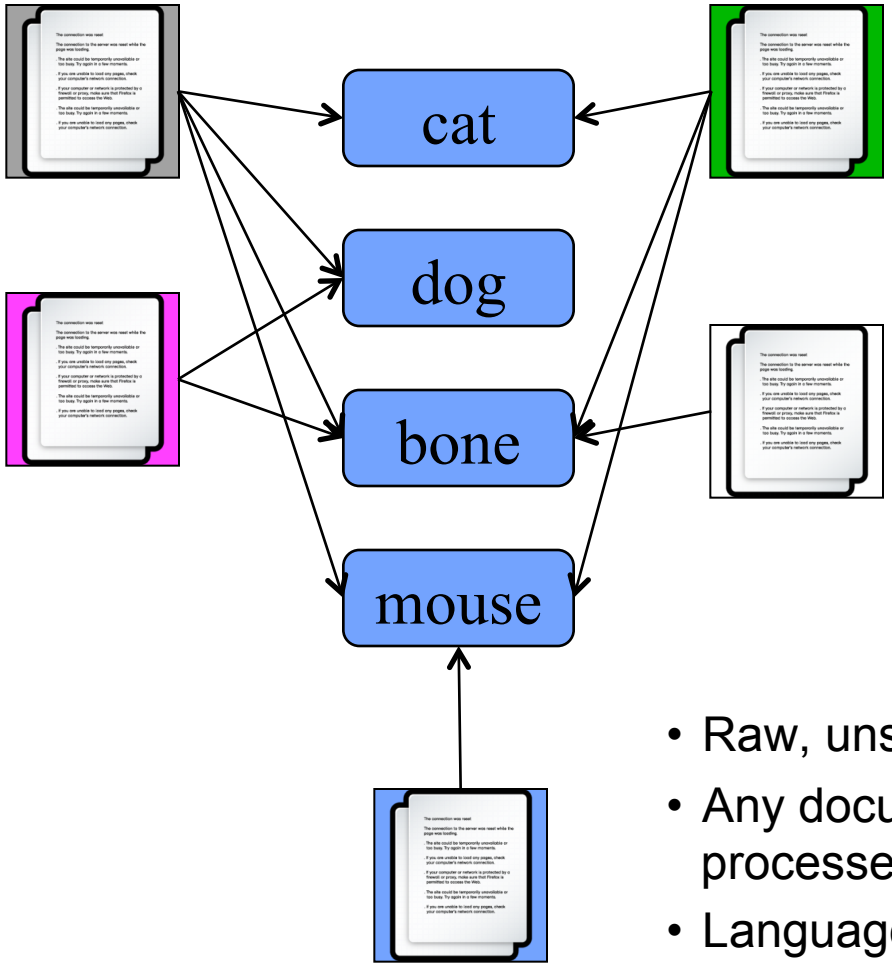
"The central hypothesis of this project is the idea that network structures extracted from large datasets can be subjected to mathematical analysis and testing to identify patterns of real-world behavior." NGC proposal





- Networks can be of many types
  - Social network
  - Cyber traffic
  - Communications
- Graph and matrix/tensor are equivalent representations
  - Extends to multiple dimensions
- Data can be easily transformed
- Both graph and algebraic algorithms can be run on the same data
  - appropriate architectures

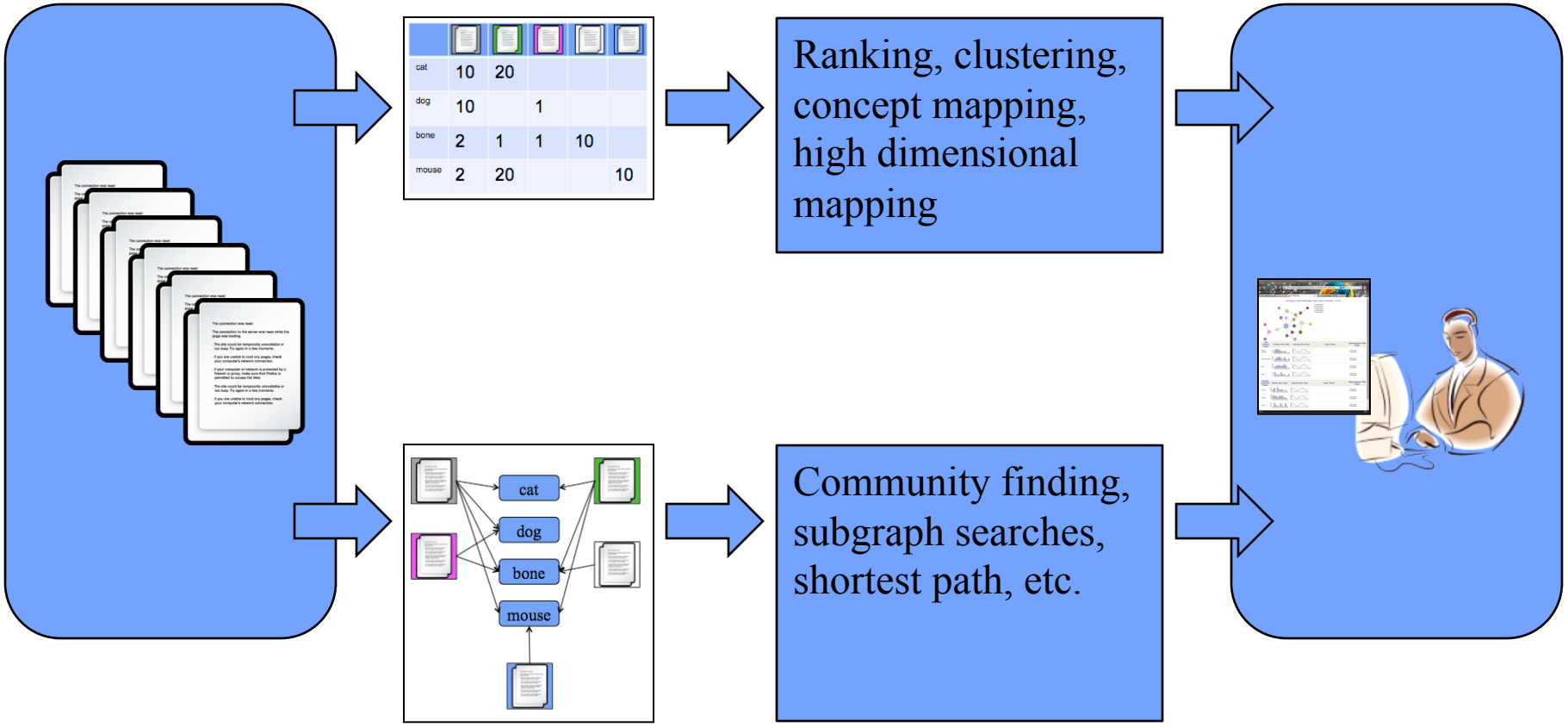
Text is Easily Transformed into Graph and Matrix Data Structures



					
cat	10	20			
dog	10		1		
bone	2	1	1	10	
mouse	2	20			10

- Raw, unstructured text input
- Any document, in any language, can be processed into these data structures
- Language expertise is embedded in techniques
  - No language expertise required to use them

‘Connecting the Dots’ Benefits from a Range of Approaches



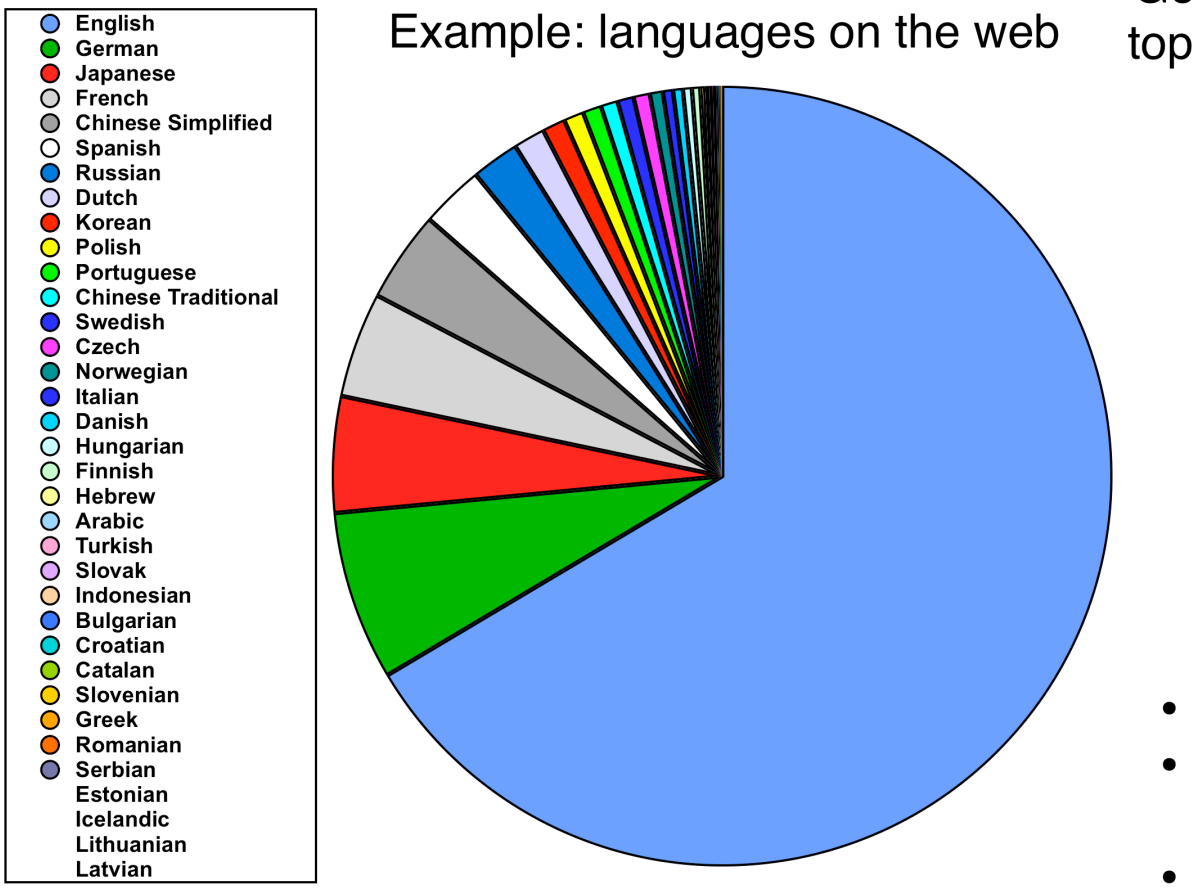
- Supports rich relationship-centered analysis
- Combines large, heterogeneous data corpora
- Different abstractions support different analytics

NGC research thrust:
Multilingual Text Analysis

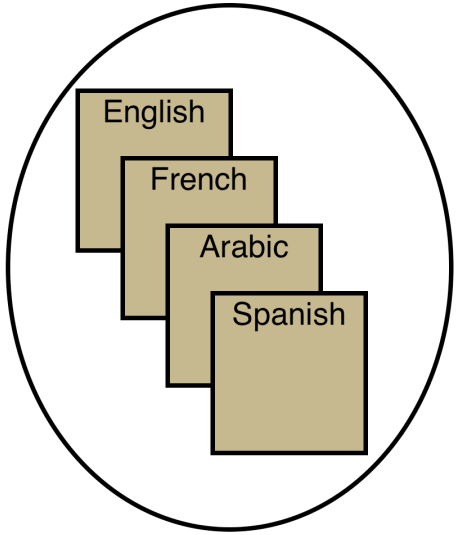
Cross-language Information Retrieval (CLIR)

Documents could be in any language

Example: languages on the web



Goal: Cluster documents by topic regardless of language



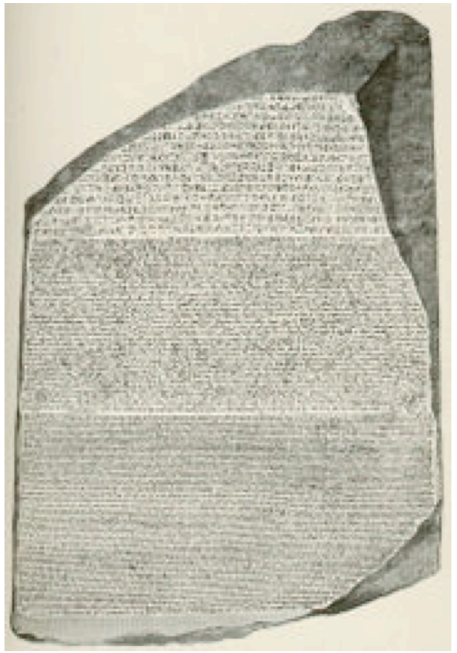
- Translation triage
- Multilingual sentiment analysis
- Ideological classification

Bible as a 'Rosetta Stone'

- The Bible has been translated carefully and widely
  - 451 complete & 2479 partial translations
- Verse aligned

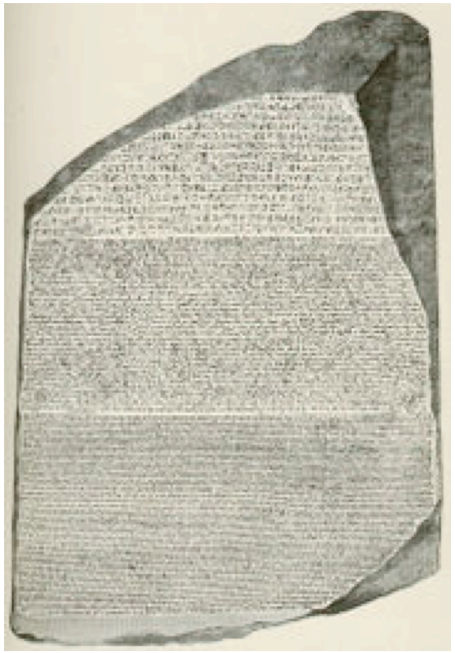
Sandia's database: 54 languages: 99.76 % coverage of web

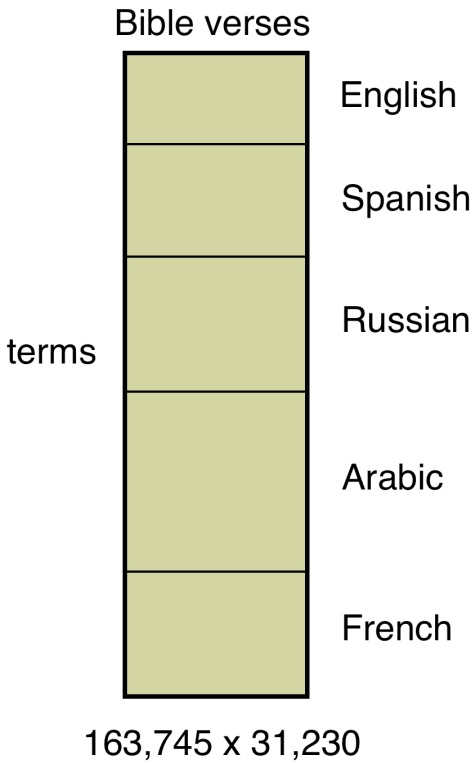
Afrikaans	Estonian	Norwegian
Albanian	Finnish	Persian (Farsi)
Amharic	French	Polish
Arabic	German	Portuguese
Aramaic	Greek (New Testament)	Romani
Armenian Eastern	Greek (Modern)	Romanian
Armenian Western	Hebrew (Old Testament)	Russian
Basque	Hebrew (Modern)	Scots Gaelic
Breton	Hungarian	Spanish
Chamorro	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



Term-Doc Matrix

Term-by-verse matrix
for all languages





Look for co-occurrence of terms in the same verses and across languages to capture latent concepts

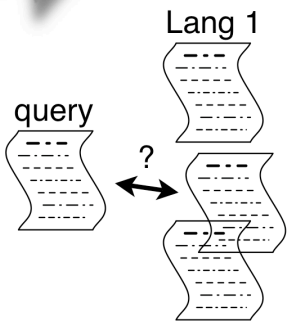
- Approach is not new: pairs of languages in Latent Semantic Analysis (LSA)
  - English and French (Landauer & Littman, 1990)
  - English and Greek (Young, 1994)
- *Multi-parallel* corpus is new

Verification and Validation

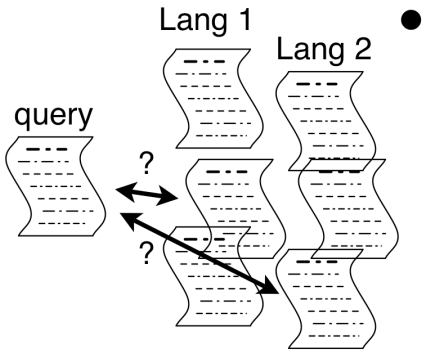
- Bible as training set
- Quran as test set

- Quran is translated into many languages, just like the Bible
  - Multi-parallel corpus
  - Ground truth
  - 114 suras (or chapters)
  - More variation across translations => harder IR task

Performance Metrics



- Average precision at 1 document (P1)
  - Equals the percentage of times the translation of the query ranked highest
  - Essentially, P1 measures success in retrieving documents when the source and target languages are specified



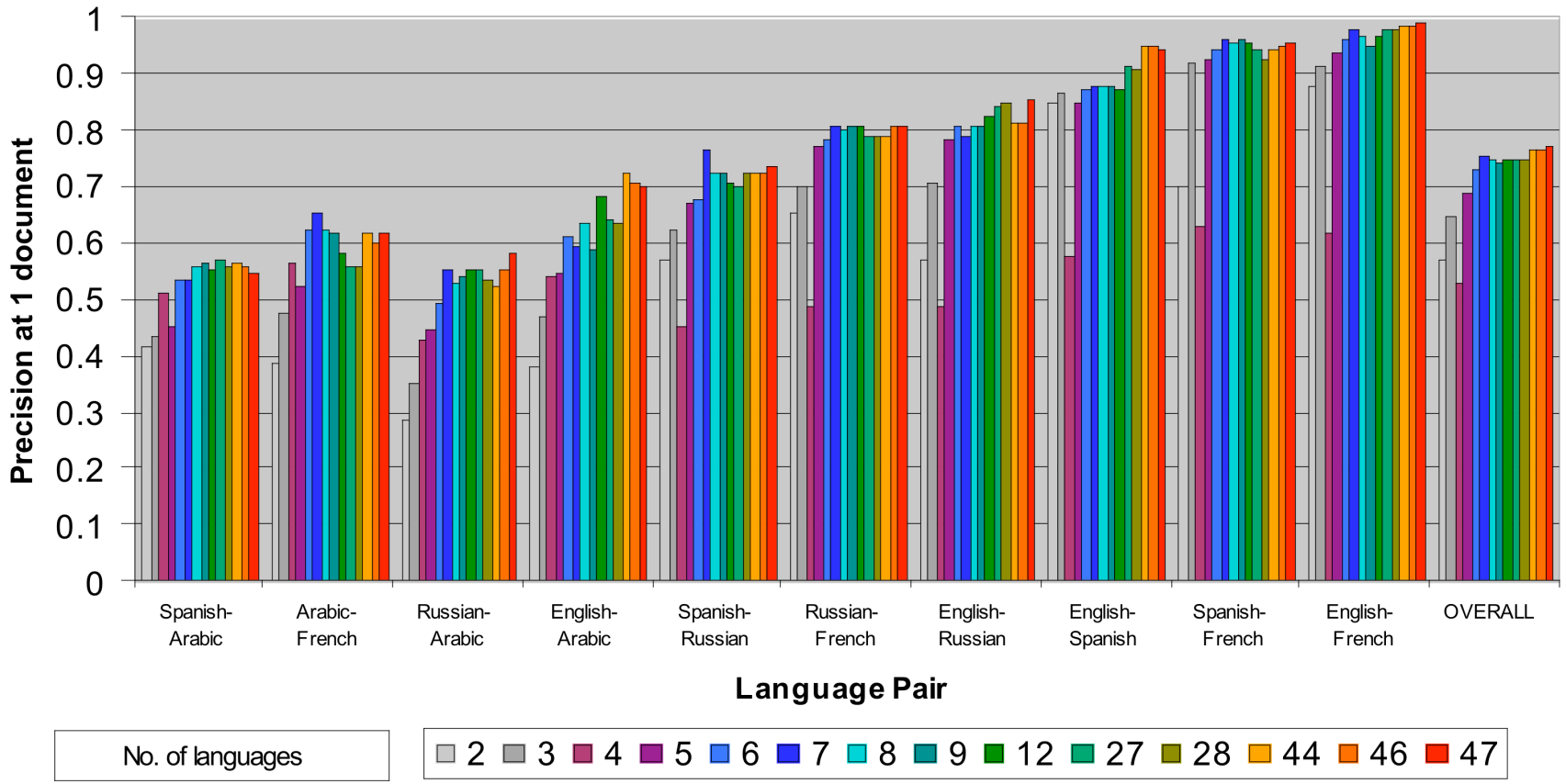
- Average multilingual precision at 5 (or n) documents (MP5)
  - The average percentage of the top 5 documents that are translations of the query document
  - Calculated as an average for all queries & all languages
  - Essentially, MP5 measures success in multilingual clustering

- Standard measures from information retrieval but adapted for multiple languages
- Striving for 90% MP5

More languages = Better results

(Chew and Abdelali, 2007)

LSA with 300 concept vectors



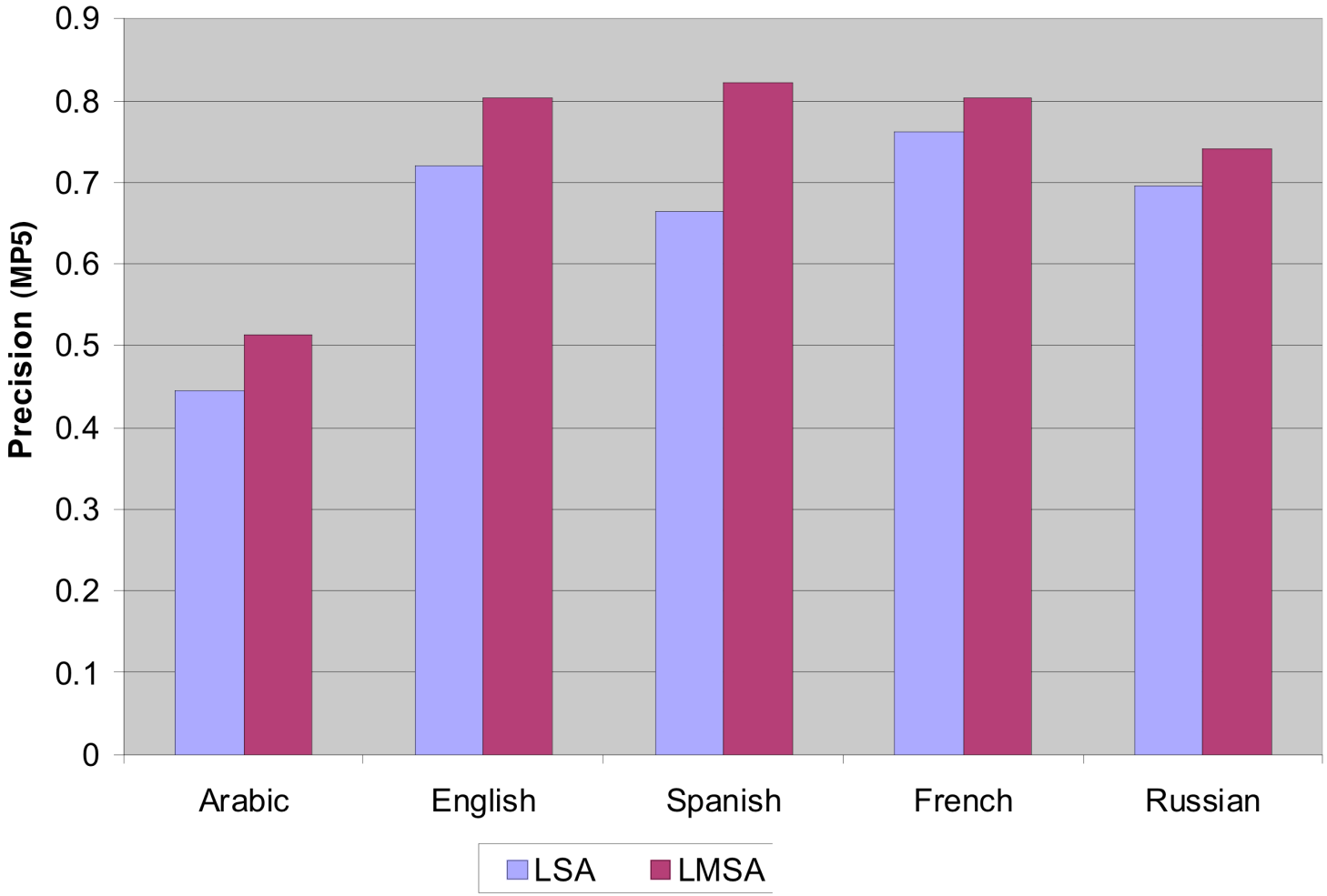
Sandia National Laboratories

Sandia's improvements on LSA

- **LMSA**
  - Latent Morpho-Semantic Analysis
    - Using morphemes, instead of terms
    - Two methods
      - Tokenization, based on mutual information of character n-grams
      - Unsupervised learning based on Minimum Description Length
    - Results in denser matrices, better results
- **LMSATA**
  - Latent Morpho-Semantic Analysis, Term Aligned
    - Statistical methods to determine highly related terms across languages
    - Collects terms that are likely to be the same across language

- Both methods are independent of human intervention
- Both methods improve the P1 and MP5 metrics

Comparison by Language

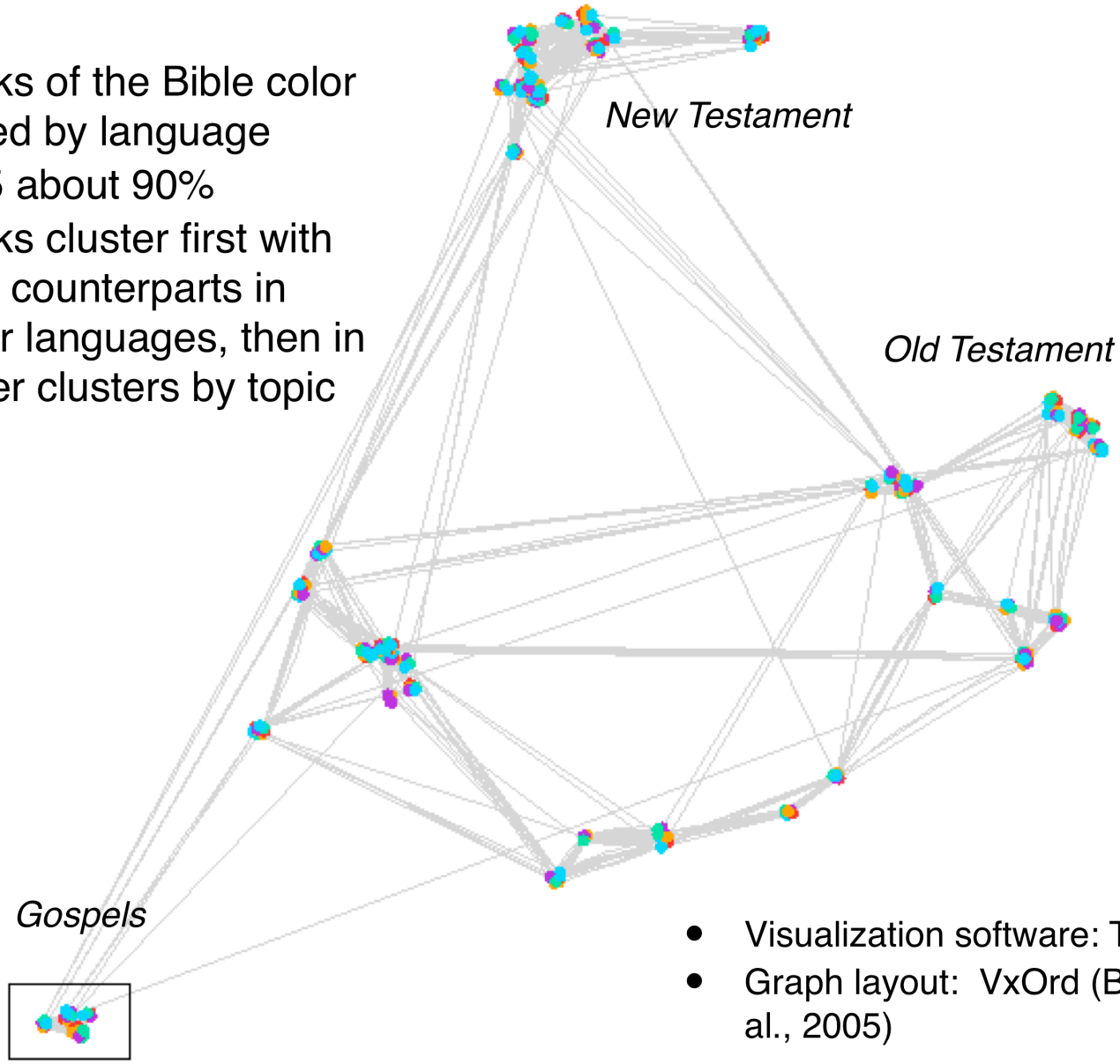


Statistically significant improvements
at $p < 0.001$

Sandia National Laboratories

Bible Clustering with LMSATA

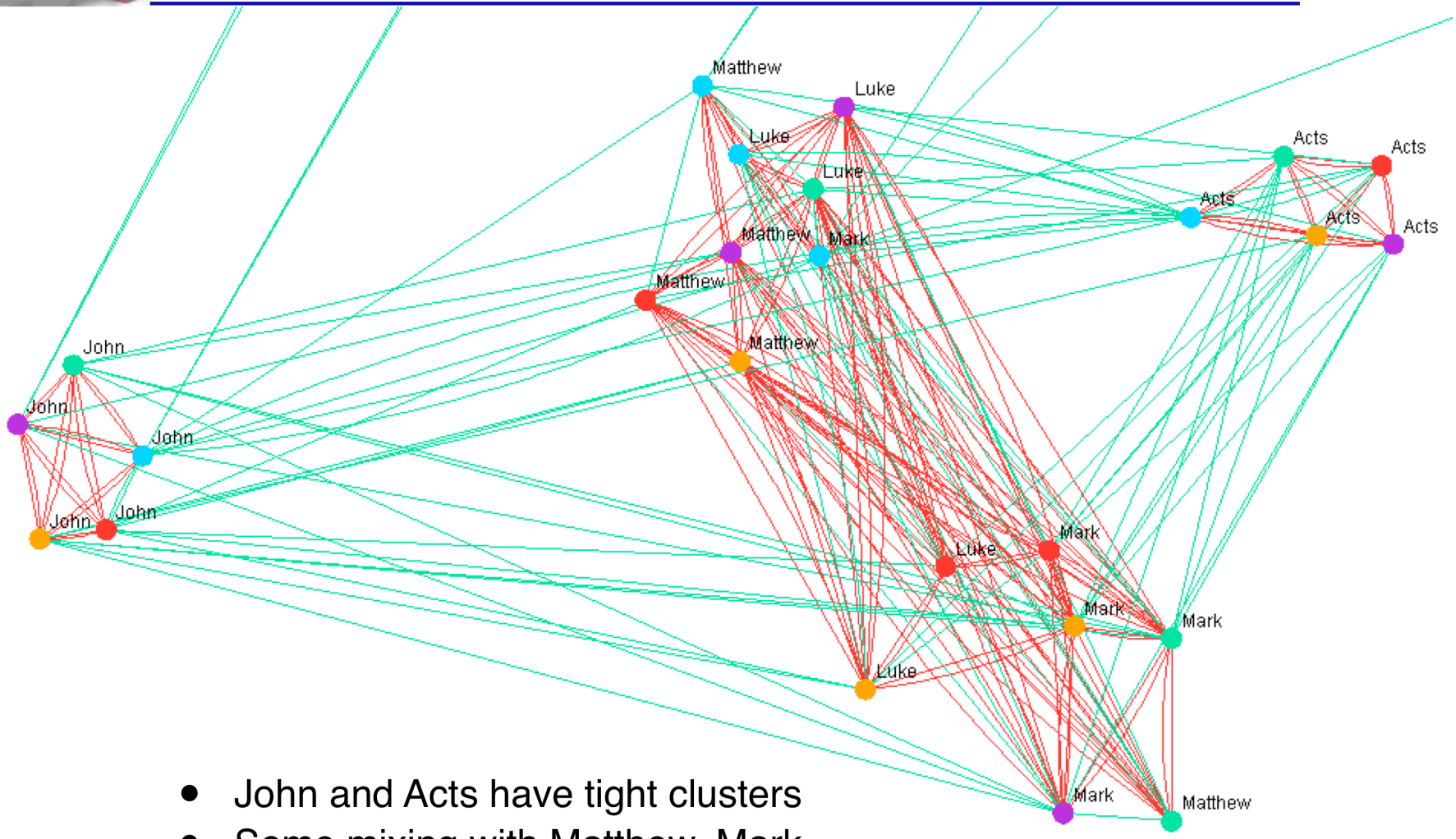
- Books of the Bible color coded by language
- MP5 about 90%
- Books cluster first with their counterparts in other languages, then in larger clusters by topic



- Visualization software: Tamale 1.2
- Graph layout: VxOrd (Boyack et al., 2005)

Sandia National Laboratories

Clustering Close-up



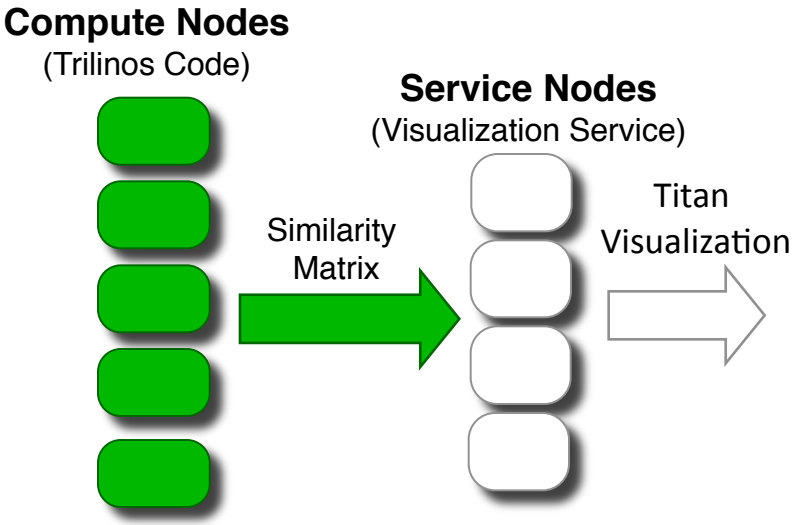
- John and Acts have tight clusters
- Some mixing with Matthew, Mark, Luke (synoptic gospels - share a similar perspective)

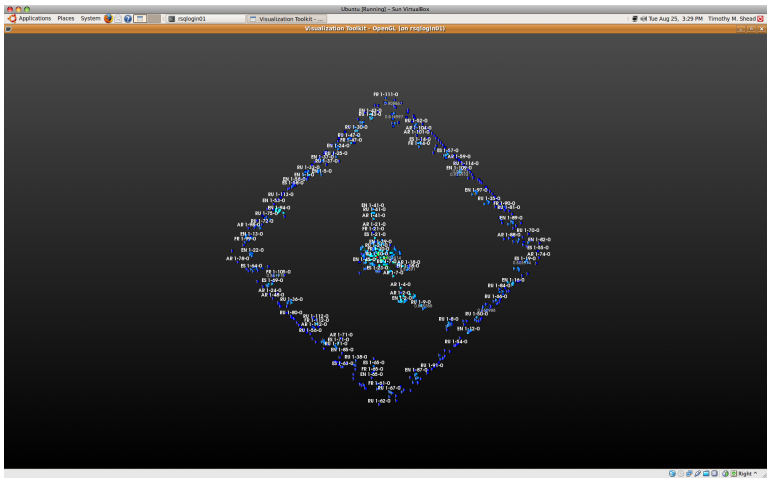
Bridging Architectures with Data Services

Network Scalable Service Interface (Nessie)

- Developed for the Lightweight File Systems Project
- Framework for HPC client/server services
- Designed for scalability and bulk data movement
- Portals and InfiniBand Implementations

Visualization Service



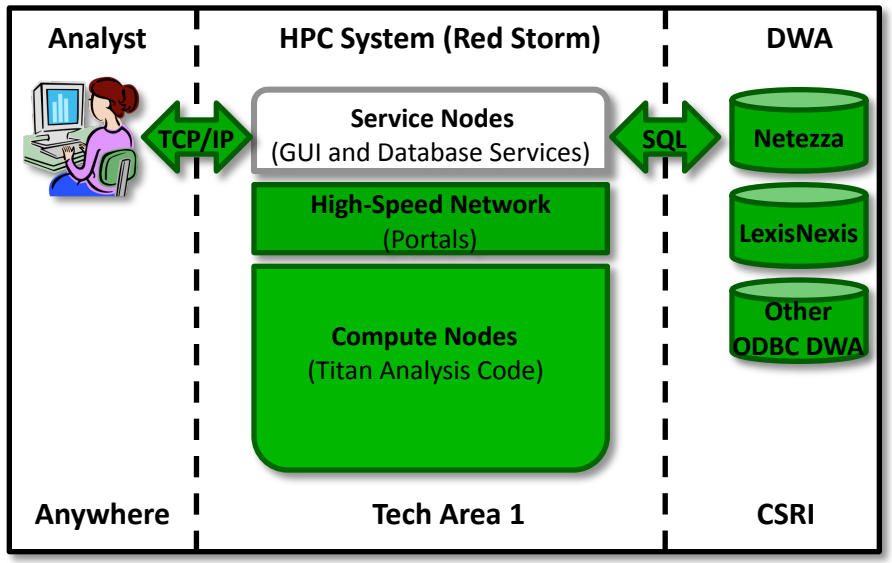


SQL Service

Enables Remote Access to Data Warehouse Appliances (DWA)

SQL Service*

- Provides "bridge" between parallel apps and external DWA
- Runs on Red Storm network nodes
- Titan applications communicate with service through Portals
- External resources (Netezza) communicate through standard interfaces (e.g. ODBC over TCP/IP)



The SQL service enables an HPC application to access a remote DWA

Additional Modifications for Multilingual

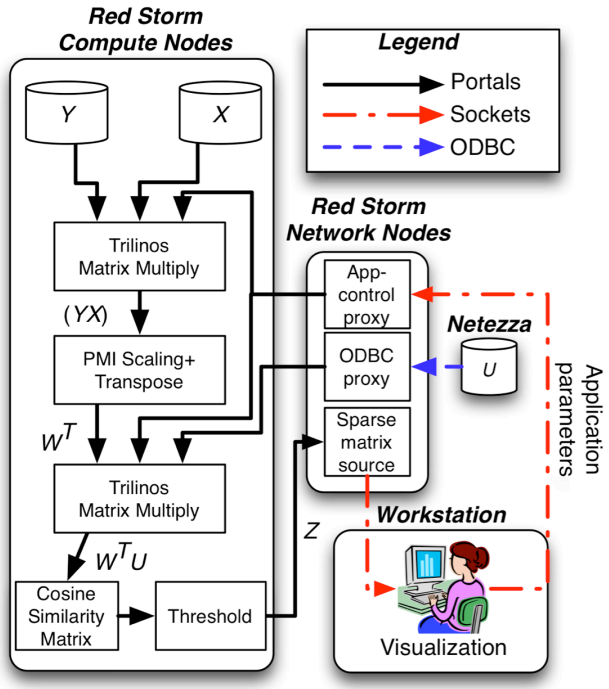
- Tokenization support on Netezza (goal is to count unique words)
- Developed a custom UTF-8 words splitter for SPU (snippet processing unit)
- Allows parallel tokenization and counting at storage device

** Results of SQL access from parallel statistics code presented at CUG'2009.*

Architectural Challenges

Exploiting specialized architectures

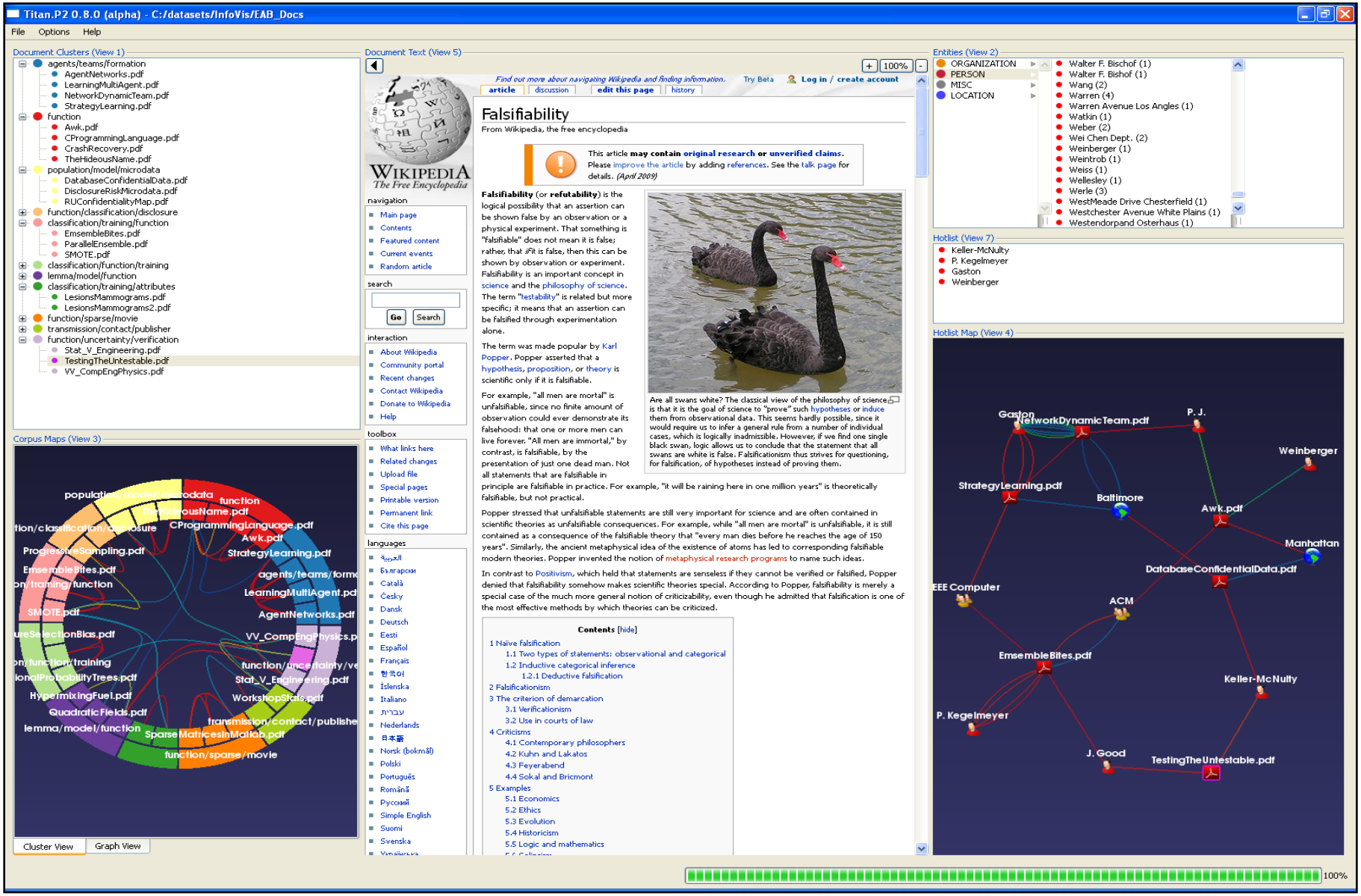
- Red Storm for numerics
- Clusters/Workstations for vis and interactive control
- Data Warehouse Appliances for database functionality



Integrating these systems for interactive jobs has never been done

Bringing it all together, in tools

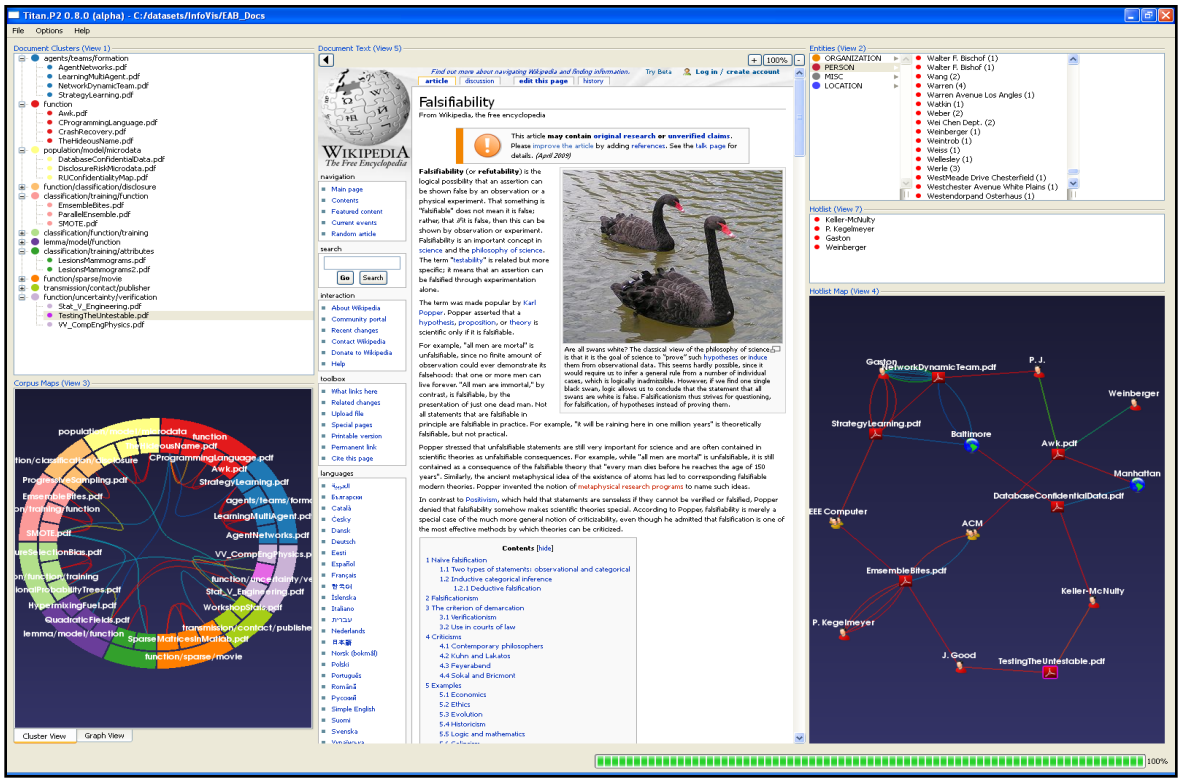
PII: Integrating Algebraic and Graph Methods



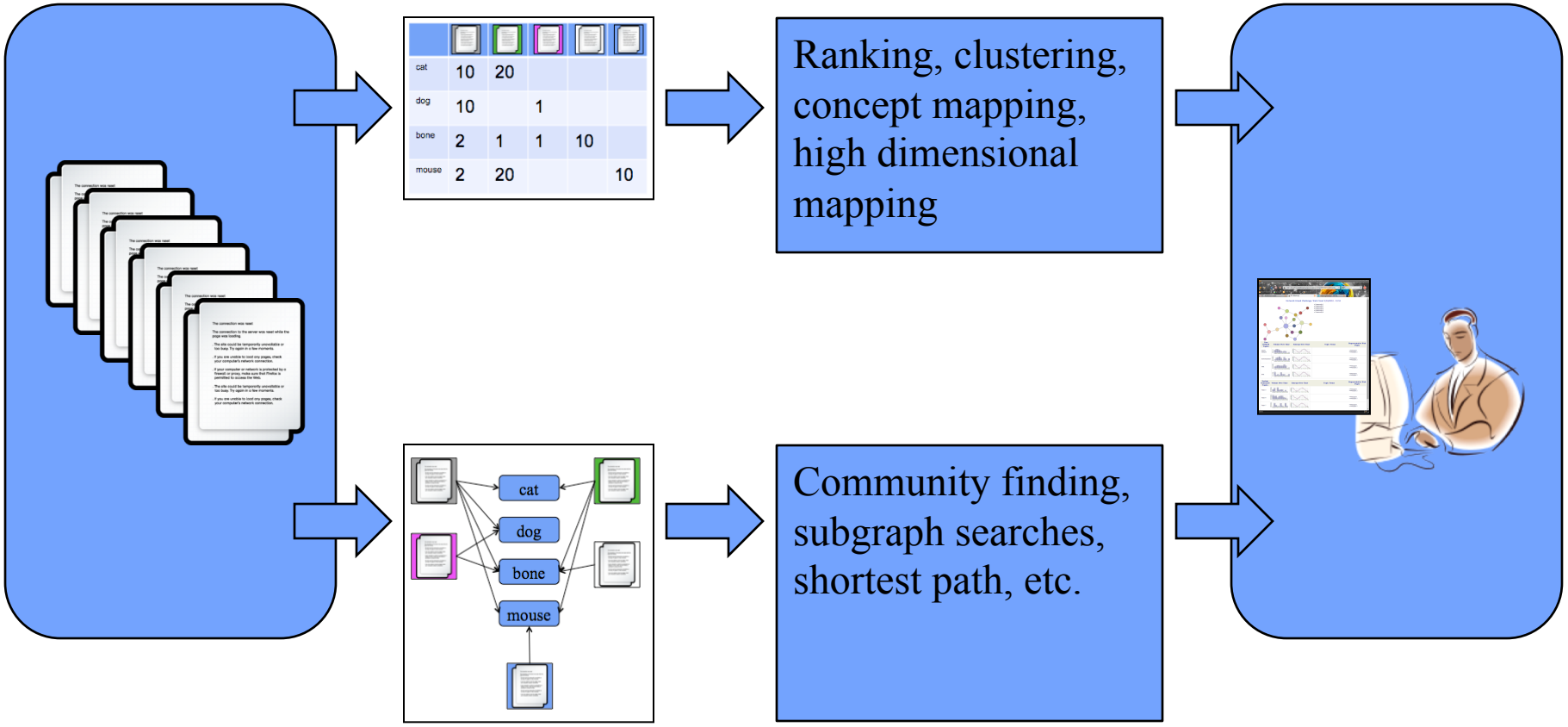
PII: Integrating Algebraic and Graph Methods

- Problem: a small dumptruck of data
- Entity extraction, clustering, entity linking, ‘soap opera’ hotlist, and corpus network/ concept query
- Query is a set of named entities
  - Can be imported; has been implemented w/Palintir

Document Display

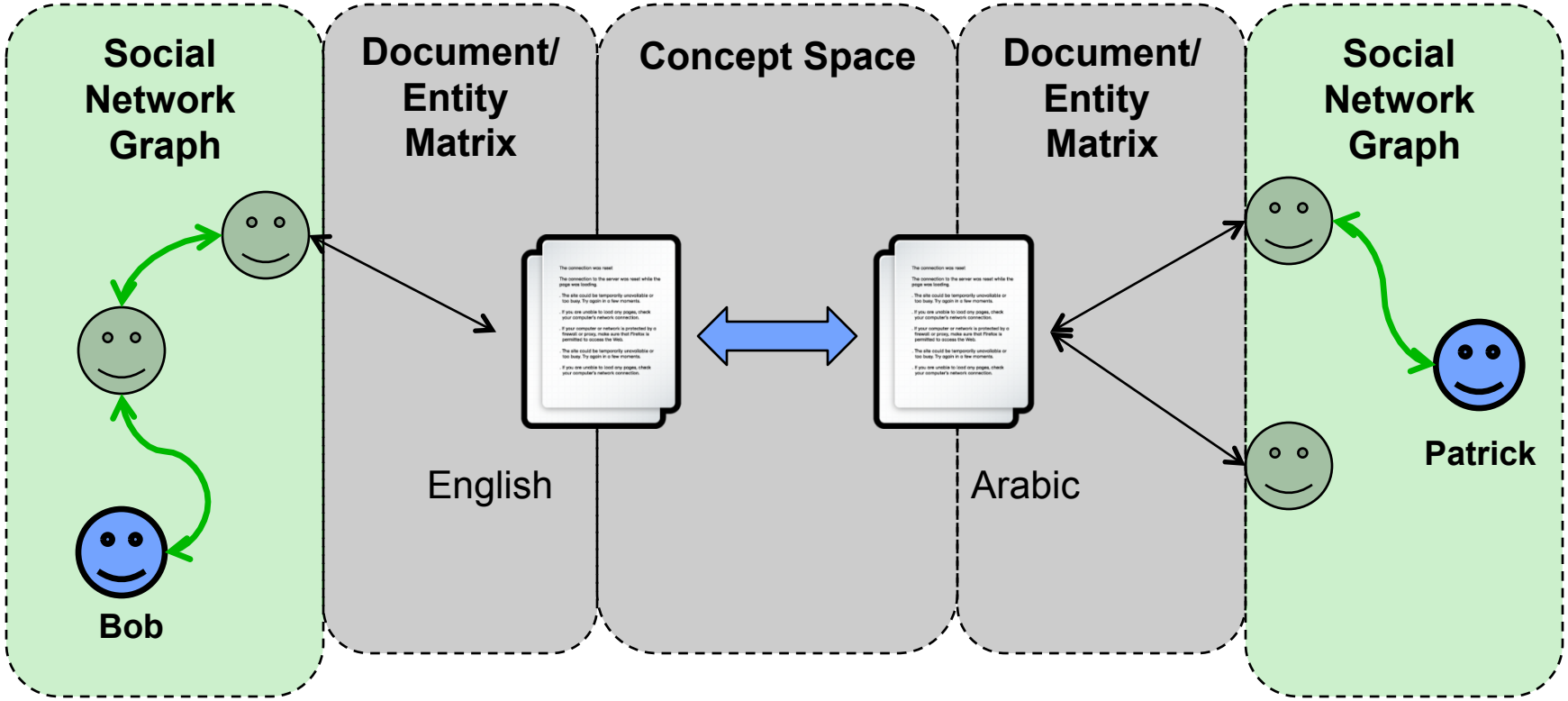


‘Connecting the Dots’ Benefits from a Range of Approaches

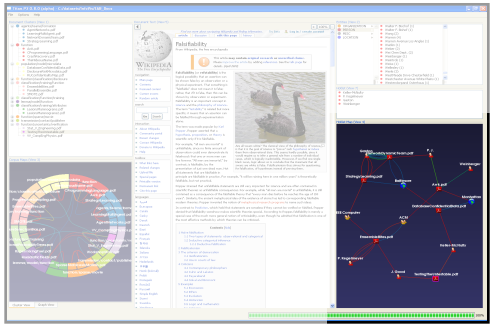


- Supports rich relationship-centered analysis
- Combines large, heterogeneous data corpora
- Different abstractions support different analytics

Linked Graph and Algebraic Methods Provide Rich Analysis Capability



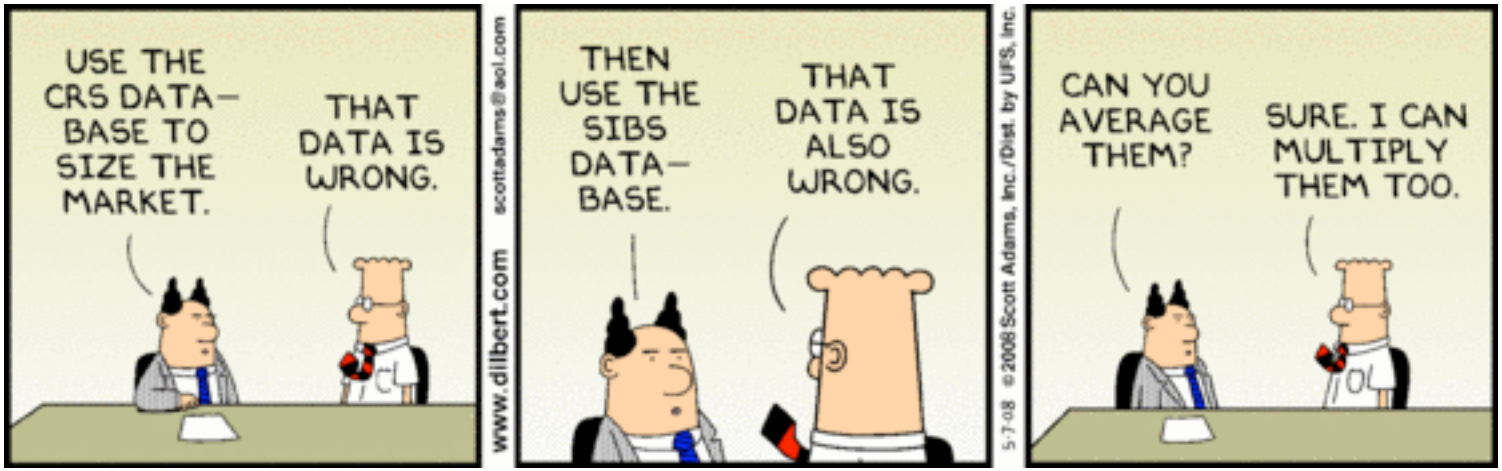
- Entities can be linked through social networks *and* concept space
- Provides rich connection data in a clear visual representation
- Promotes new paths of investigation



PII Prototype, emphasizing the combined graph and algebraic methods view

An important next step

Make sure your users understand the math …



- We find that users need to understand the underlying algorithms – everything from analysis to layout (visualization)
  - Prevents incorrect conclusions from the data
- Analysis V&V needs study
  - How do we quantify the accuracy of the conclusions that are made?
- Close connection to analysts (Human Factors Team) is critical to ensure relevance
  - An untrusted algorithm is useless (literally won't get used)
  - A trusted algorithm should be understood
- This data-algorithm-visualization-analyst cycle is crucial
  - Subject of another talk

Questions?