



# **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.

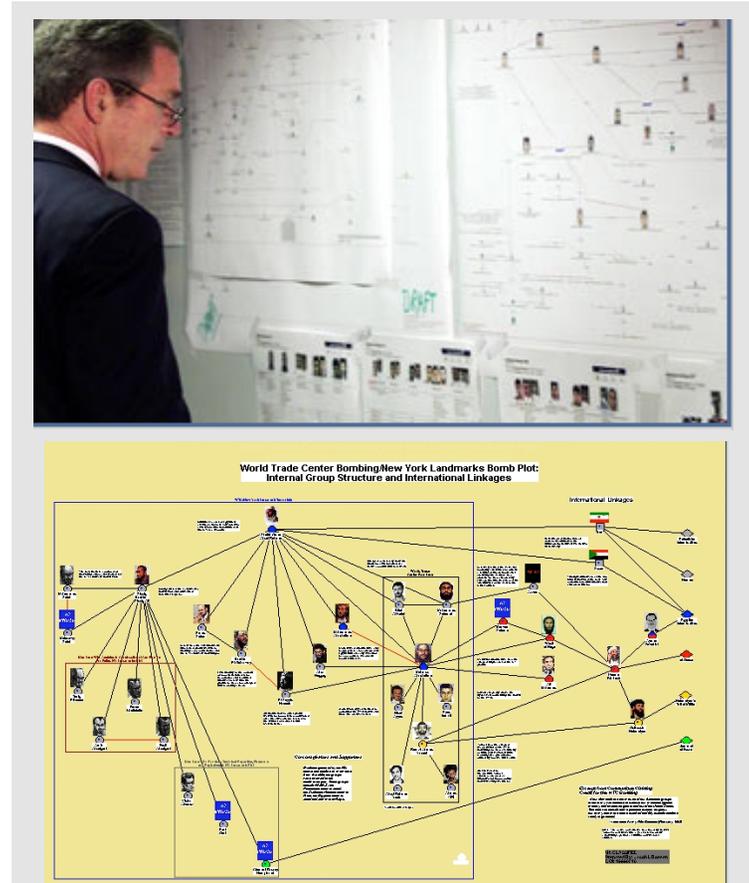


# **What is the Networks Grand Challenge?**



# “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 10<sup>13</sup> data).
  - Batch-processing unacceptable (analysts need answers within seconds).



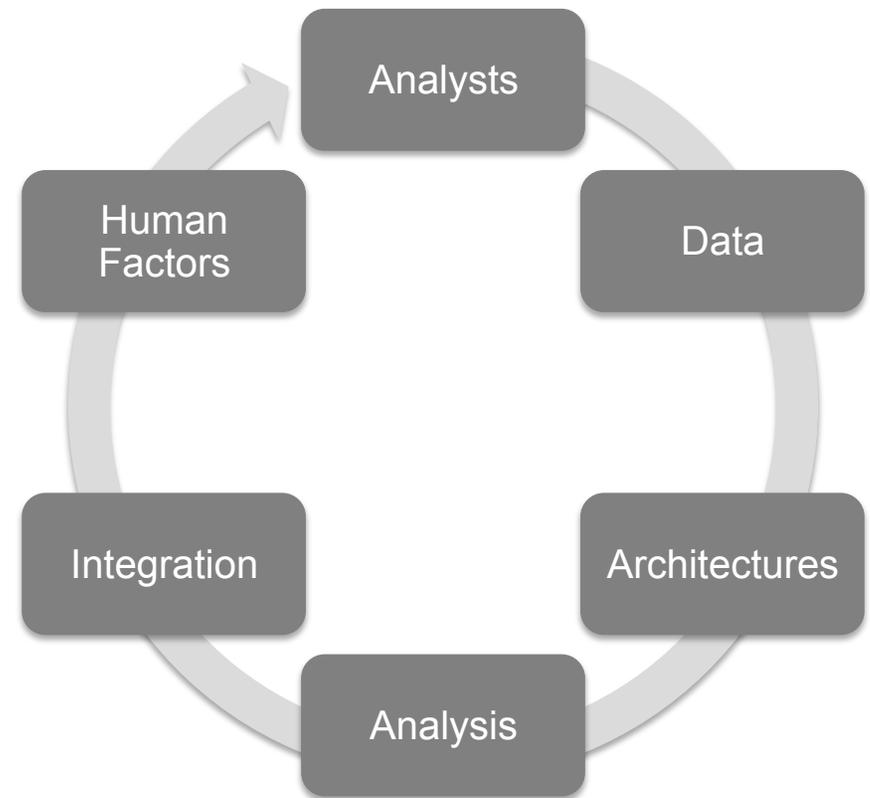
President Bush looks over a chart depicting Osama bin Laden's financial network; an excerpt from the chart (created using one of the leading tools, i2 Analyst's Notebook)

....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

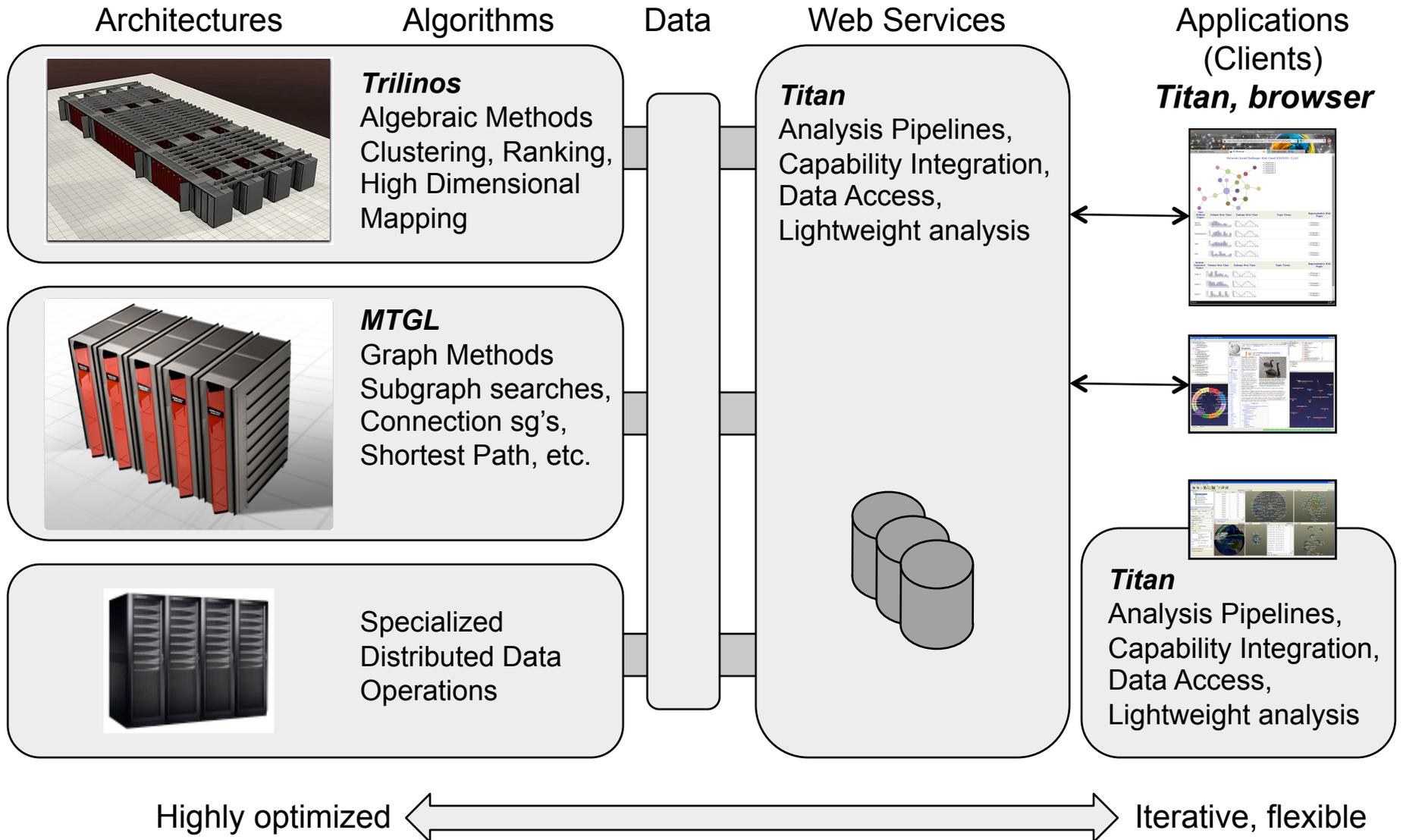
- **Data**
  - Provides data sets to the project, creates and supports the data infrastructure required by the project.
- **Architectures**
  - Develops software to efficiently integrate specialized hardware devices required by the research teams
- **Analysis (Discovery and Forecasting Teams)**
  - Develops 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.





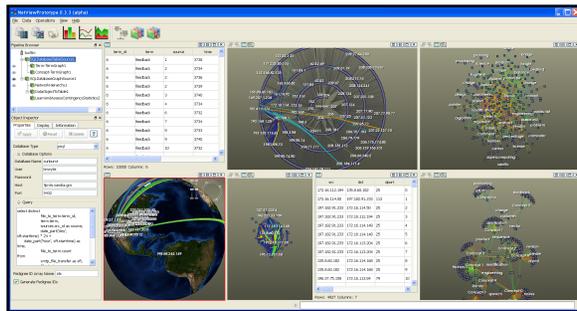
# NGC System Diagram

*“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*





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



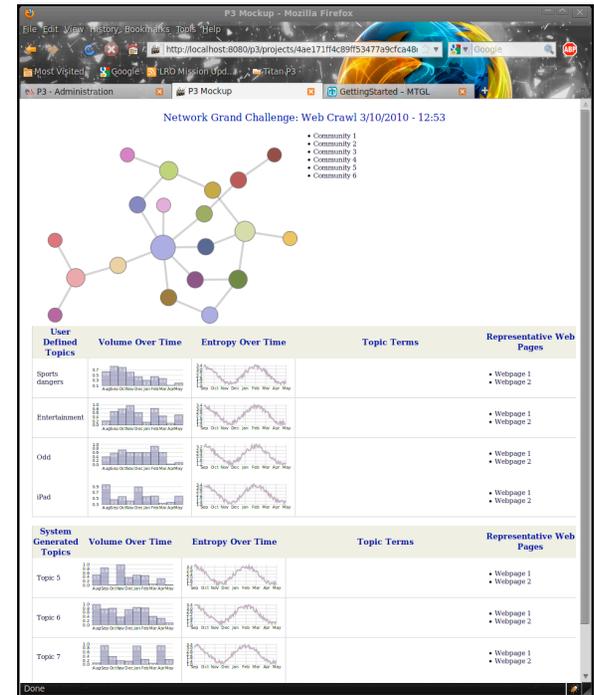
## PI: Cyber Application

- 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?**

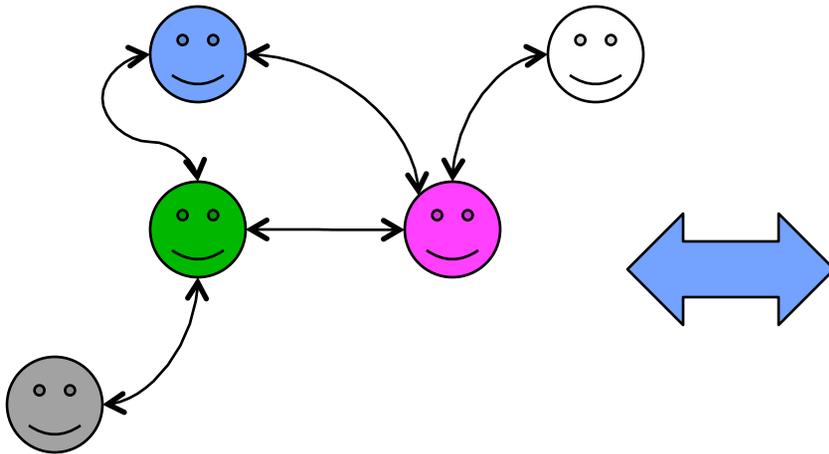


# **NGC: Unifying Data Abstractions**



# 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*

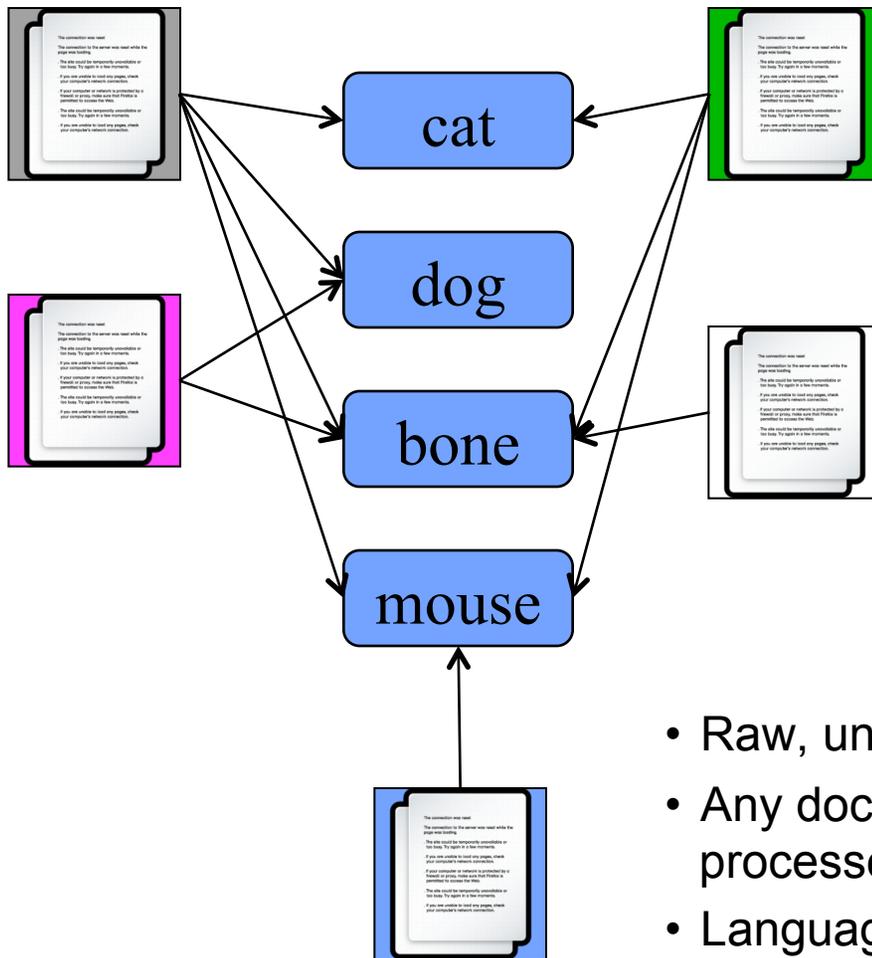


	Grey	Green	Blue	Pink	White
Grey		1			
Green	1		1	1	
Blue		1		1	
Pink		1	1		1
White				1	

- 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

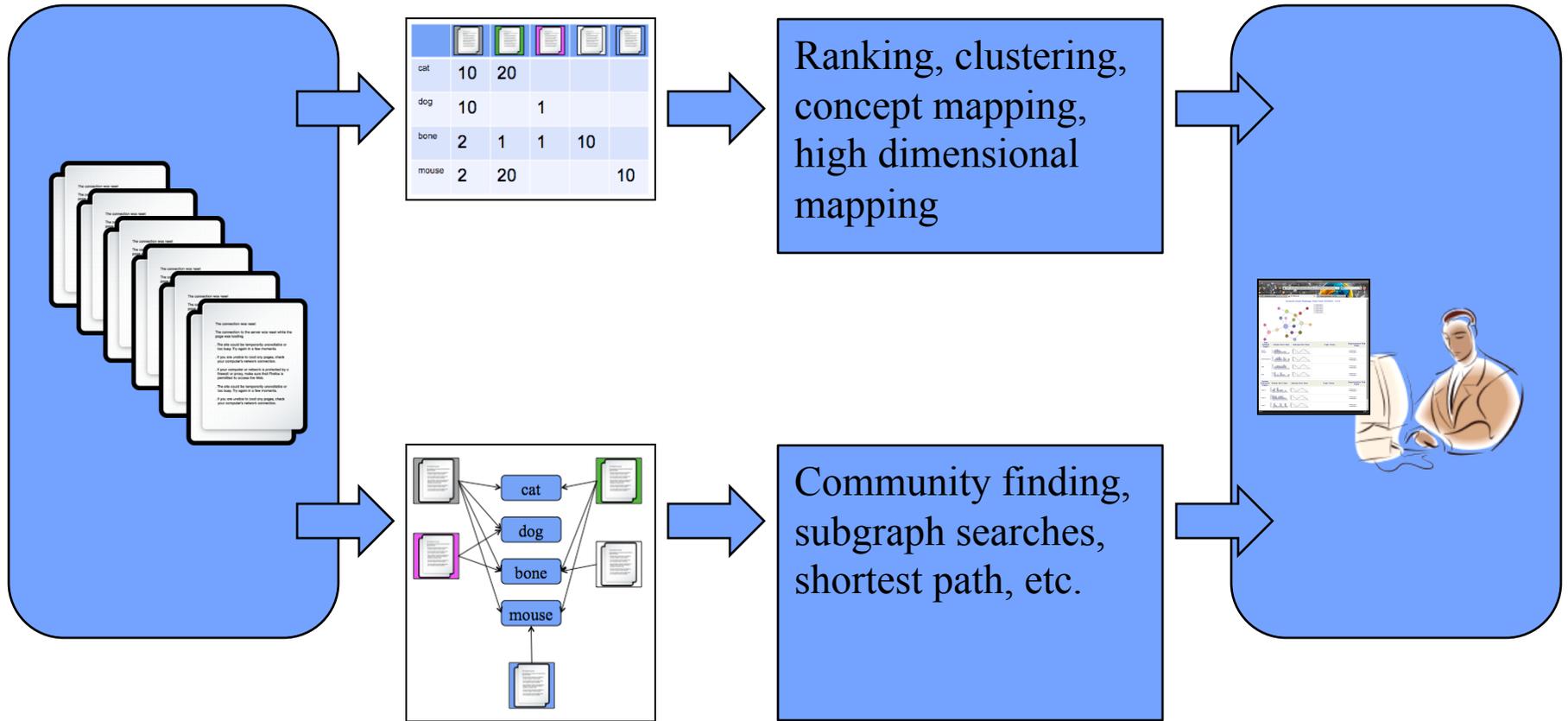


cat	10	20			
dog	10		1		
bone	2	1	1	10	
mouse	2	20			10

- 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**

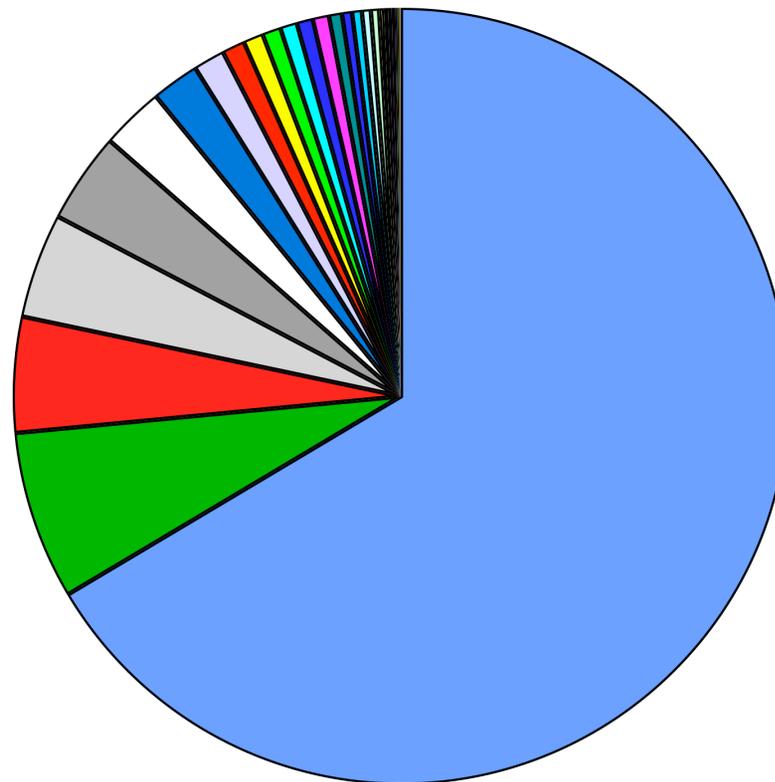


# **NGC research thrust: Multilingual Text Analysis**

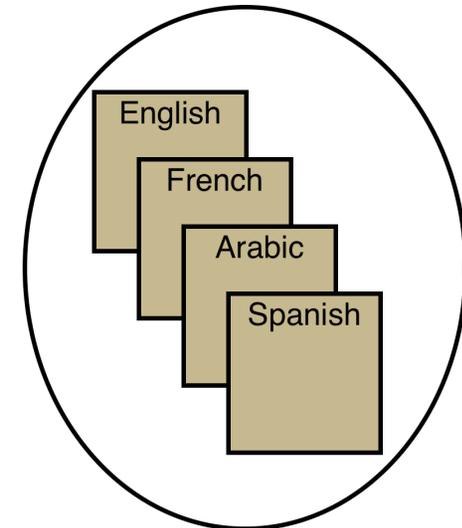
# Cross-language Information Retrieval (CLIR)

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

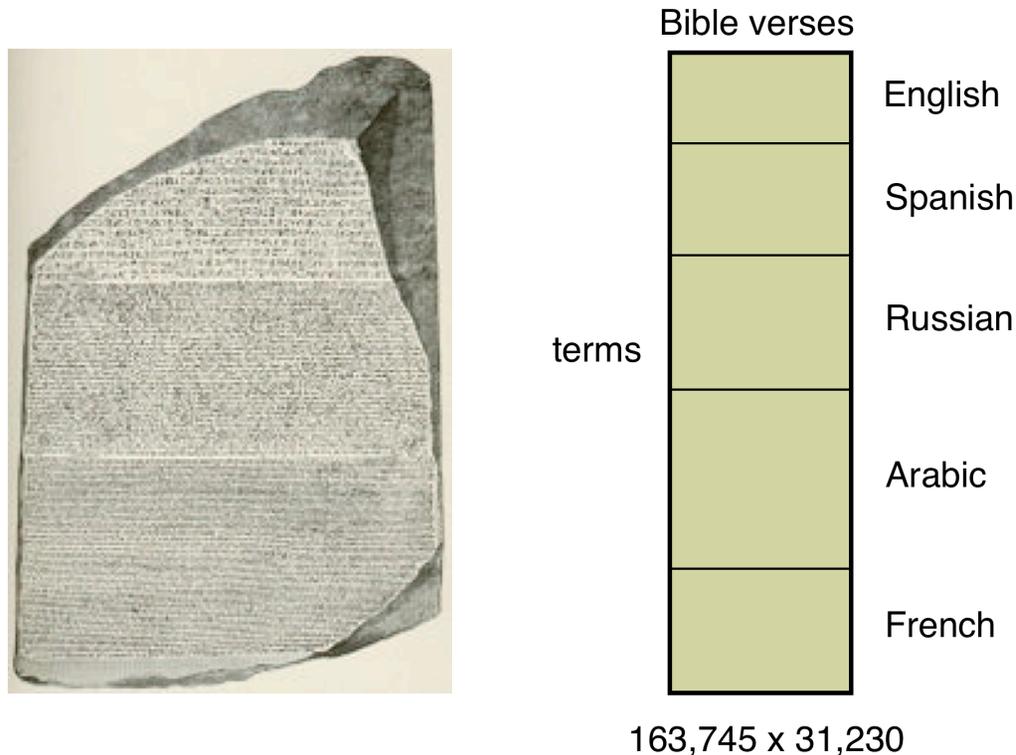
Sandia's database: 54 languages: 99.76 % coverage of web

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

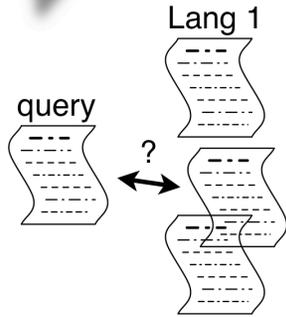


# Verification and Validation

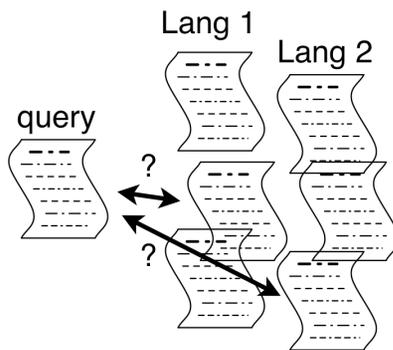
---

- 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

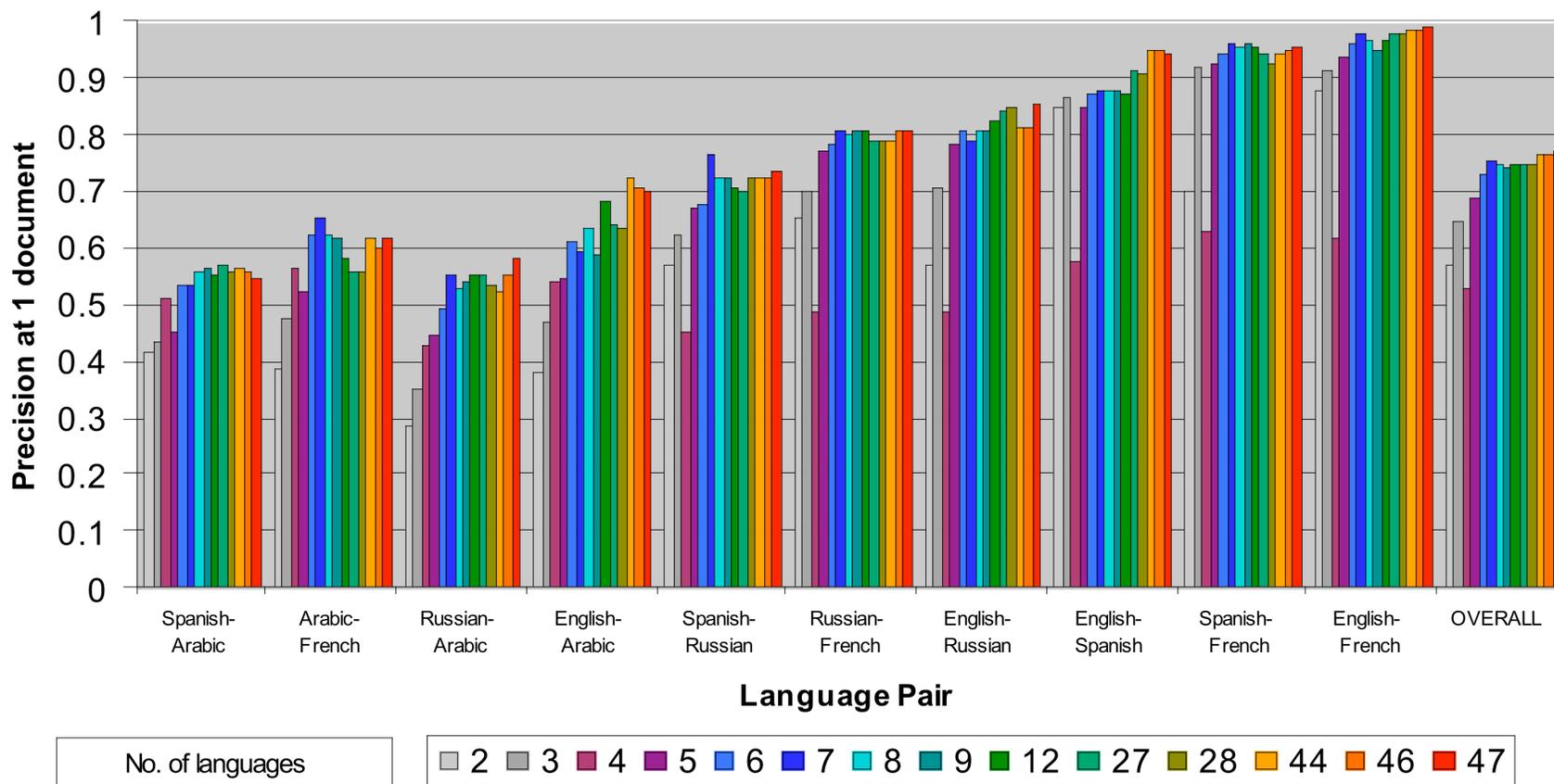


- **Average multilingual precision at 5 (or n) documents (MP5)**
  - The average percentage of the top 5 documents that are translations of the query document
  - Calculated as an average for all queries & all languages
  - Essentially, MP5 measures success in multilingual clustering
- Standard measures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

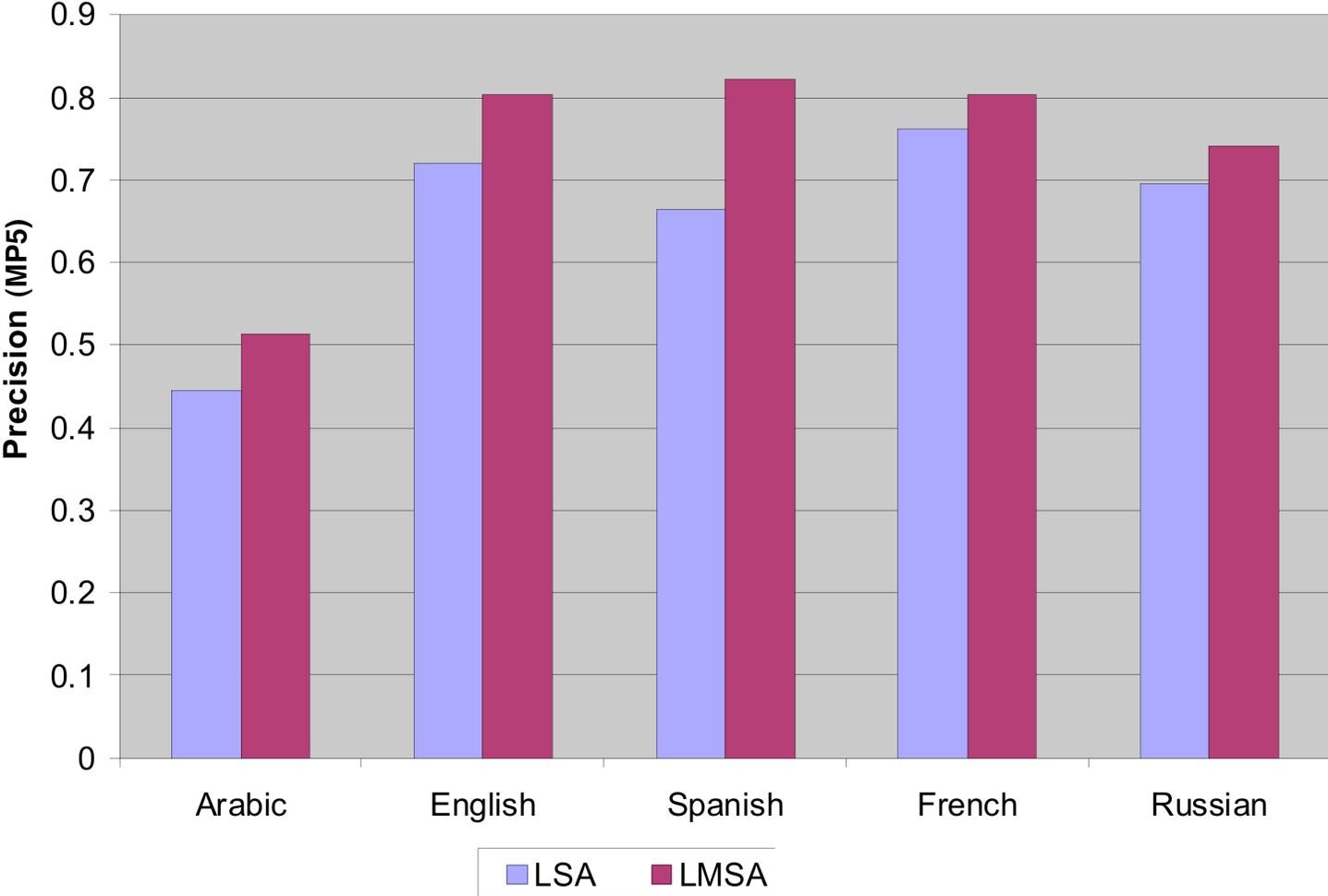
- Latent Morpho-Semantic Analysis, Term Aligned

- 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



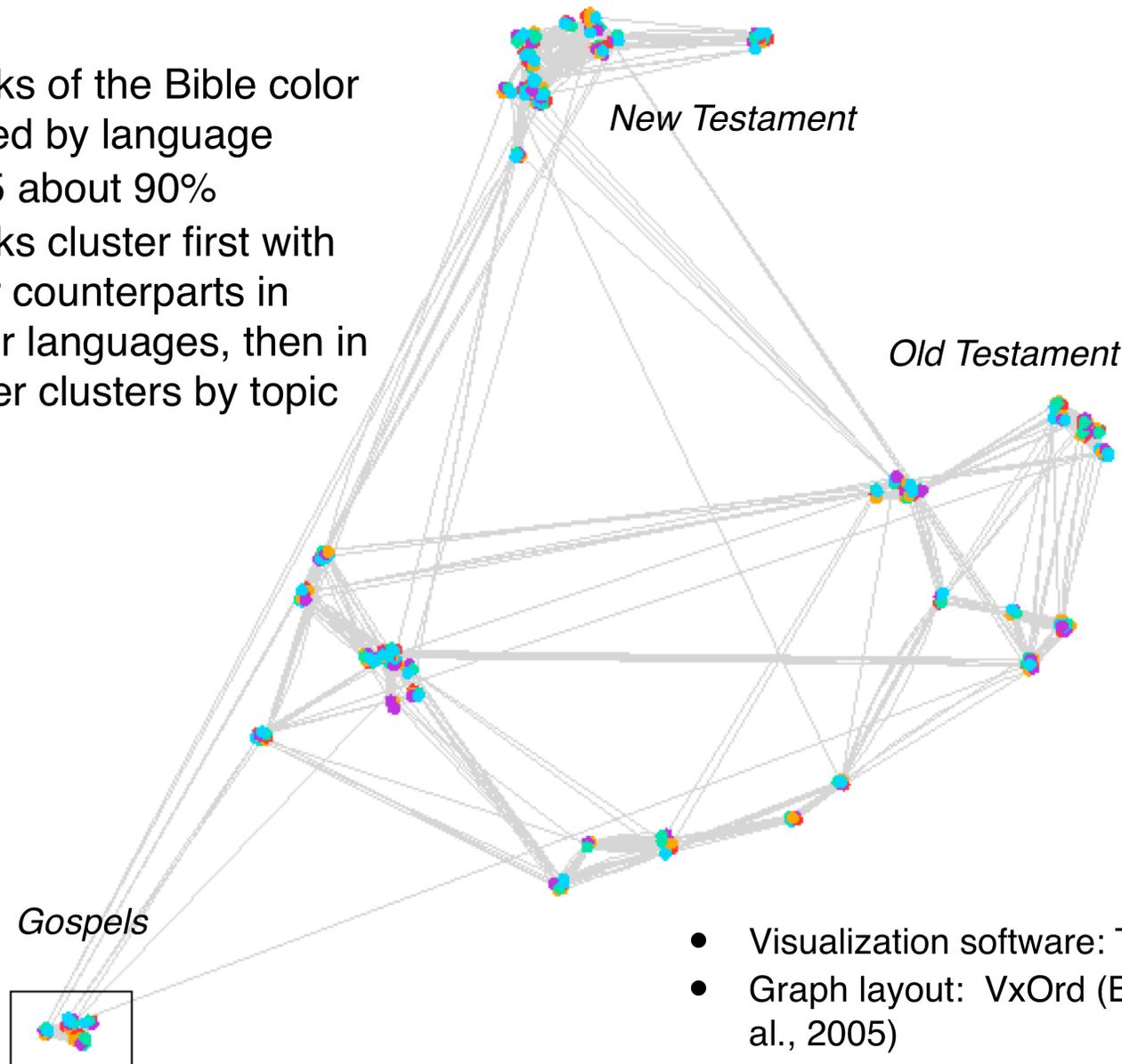
# Comparison by Language



Statistically significant improvements  
at  $p < 0.001$

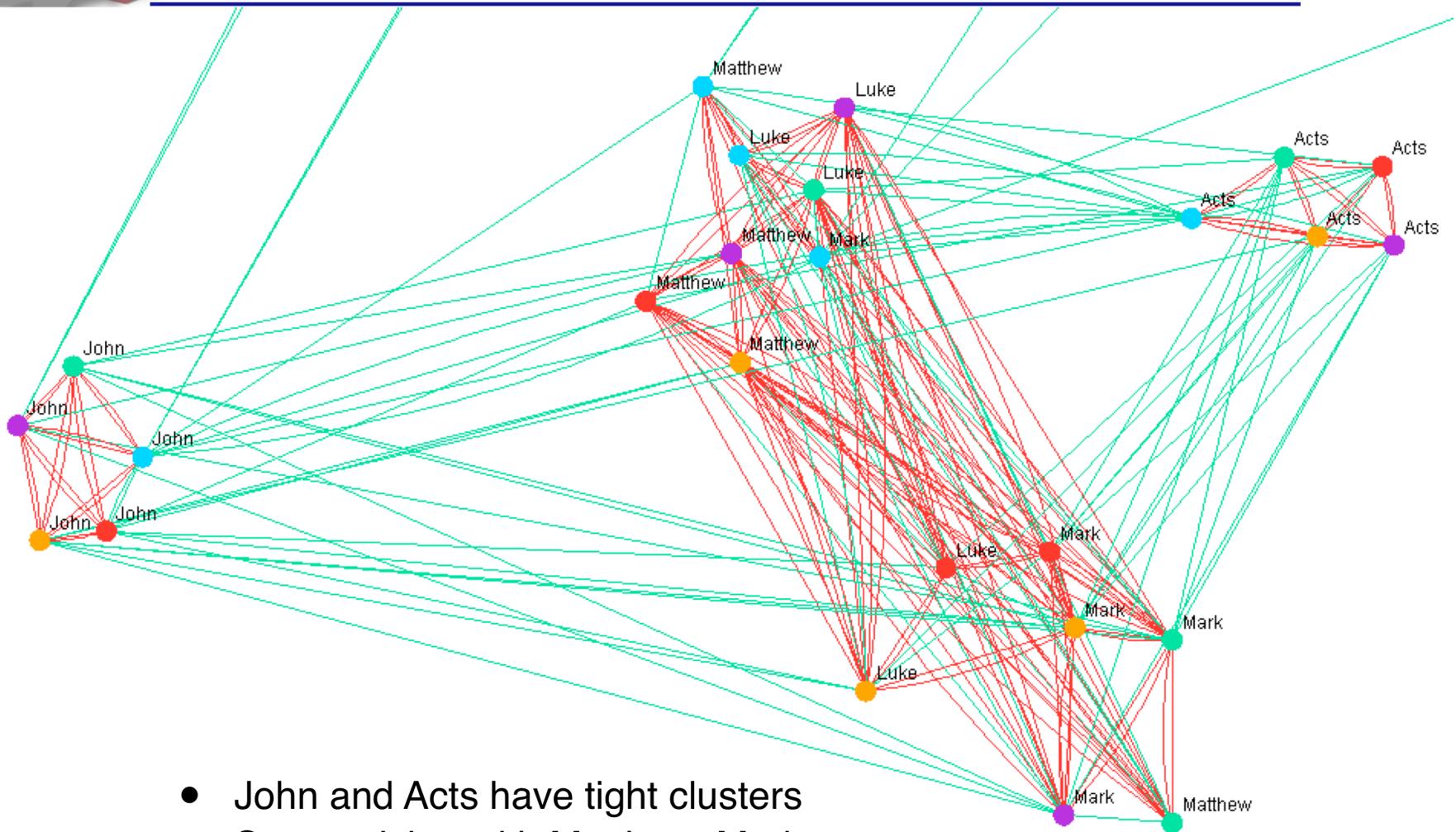
# Bible Clustering with LMSATA

- 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)

# 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

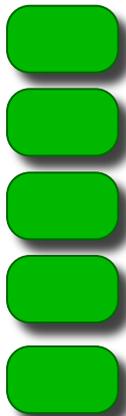
## Network Scalable Service Interface (Nessie)

- 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

Compute Nodes  
(Trilinos Code)



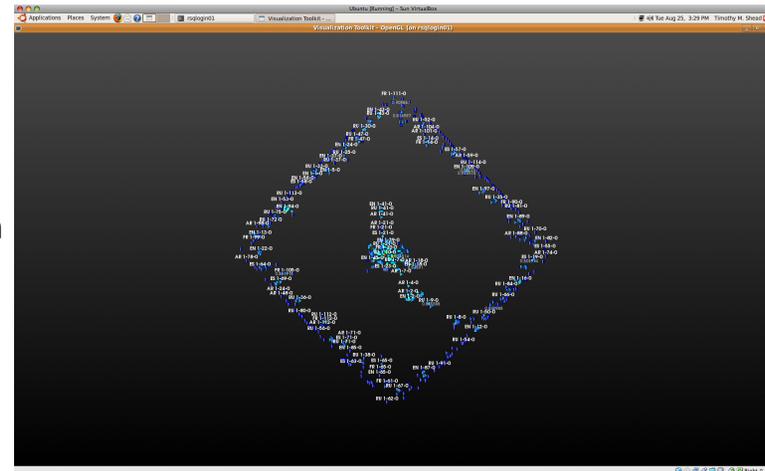
Similarity  
Matrix



Service Nodes  
(Visualization Service)



Titan  
Visualization



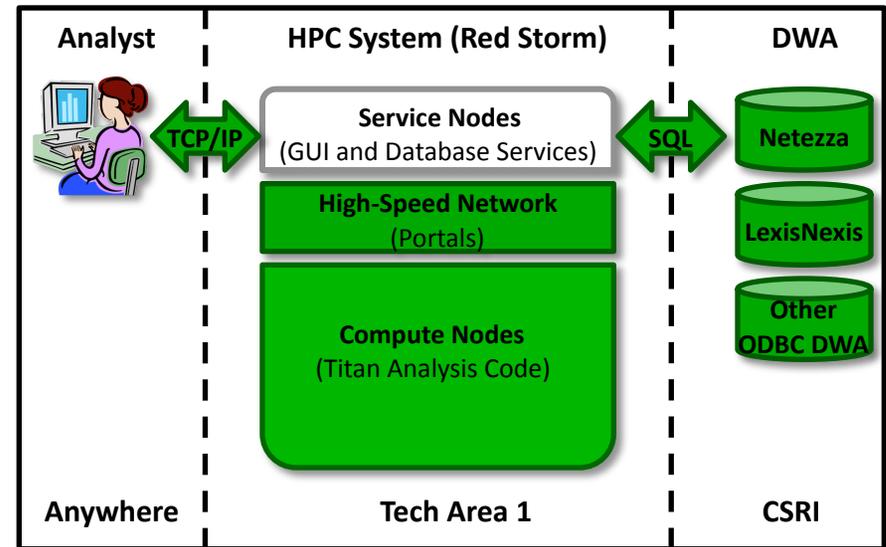


# SQL Service

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

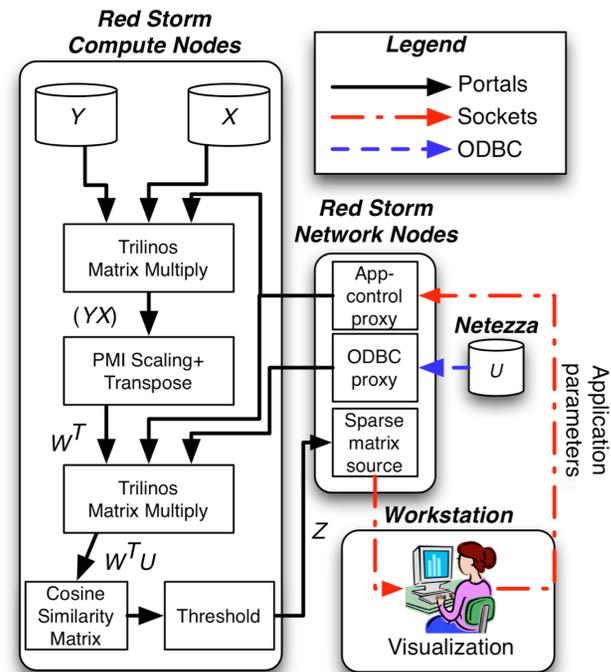
- Tokenization support on Netezza (goal is to count unique words)
- Developed a custom UTF-8 words splitter for SPU (snippet processing unit)
- Allows parallel tokenization and counting at storage device

*\* 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

Titan.P2 0.8.0 (alpha) - C:/datasets/InfoVis/EAB\_Docs

File Options Help

Document Clusters (View 1)

- agents/teams/formation
  - AgentNetworks.pdf
  - LearningMultiAgent.pdf
  - NetworkDynamicTeam.pdf
  - StrategyLearning.pdf
- function
  - Awk.pdf
  - CProgrammingLanguage.pdf
  - CrashRecovery.pdf
  - TheHideousName.pdf
- population/model/microdata
  - DatabaseConfidentialData.pdf
  - DisclosureRiskMicrodata.pdf
  - RUConfidentialityMap.pdf
- function/classification/disclosure classification/training/function
  - EmsembleBites.pdf
  - ParallelEmsemble.pdf
  - SMOTE.pdf
- classification/function/training
  - lemma/model/function
- classification/training/attributes
  - LesionsMammograms.pdf
  - LesionsMammograms2.pdf
- function/sparse/movie
  - transmission/contact/publisher
  - function/uncertainty/verification
    - Stat\_V\_Engineering.pdf
    - TestingTheUntestable.pdf
    - VV\_CompEngPhysics.pdf

Corpus Maps (View 3)

Cluster View Graph View

Document Text (View 5)

Find out more about navigating Wikipedia and finding information. Try Beta Log in / create account

## Falsifiability

From Wikipedia, the free encyclopedia

This article **may contain original research or unverified claims.** Please improve the article by adding references. See the talk page for details. (April 2009)

**Falsifiability** (or **refutability**) is the logical possibility that an assertion can be shown false by an observation or a physical experiment. That something is "falsifiable" does not mean it is false; rather, that if it is false, then this can be shown by observation or experiment. Falsifiability is an important concept in science and the philosophy of science. The term "testability" is related but more specific; it means that an assertion can be falsified through experimentation alone.

The term was made popular by Karl Popper. Popper asserted that a hypothesis, proposition, or theory is scientific only if it is falsifiable.

For example, "all men are mortal" is unfalsifiable, since no finite amount of observation could ever demonstrate its falsehood; that one or more men can live forever. "All men are immortal," by contrast, is falsifiable, by the presentation of just one dead man. Not all statements that are falsifiable in principle are falsifiable in practice. For example, "it will be raining here in one million years" is theoretically falsifiable, but not practical.

Popper stressed that unfalsifiable statements are still very important for science and are often contained in scientific theories as unfalsifiable consequences. For example, while "all men are mortal" is unfalsifiable, it is still contained as a consequence of the falsifiable theory that "every man dies before he reaches the age of 150 years". Similarly, the ancient metaphysical idea of the existence of atoms has led to corresponding falsifiable modern theories. Popper invented the notion of **metaphysical research programs** to name such ideas.

In contrast to **Positivism**, which held that statements are senseless if they cannot be verified or falsified, Popper denied that falsifiability somehow makes scientific theories special. According to Popper, falsifiability is merely a special case of the much more general notion of **criticizability**, even though he admitted that falsification is one of the most effective methods by which theories can be criticized.

Are all swans white? The classical view of the philosophy of science is that it is the goal of science to "prove" such hypotheses or induce them from observational data. This seems hardly possible, since it would require us to infer a general rule from a number of individual cases, which is logically inadmissible. However, if we find one single black swan, logic allows us to conclude that the statement that all swans are white is false. Falsificationism thus strives for questioning, for falsification, of hypotheses instead of proving them.

Contents [hide]

- Naive falsification
  - Two types of statements: observational and categorical
  - Inductive categorical inference
    - Deductive falsification
- Falsificationism
- The criterion of demarcation
  - Verificationism
  - Use in courts of law
- Criticism
  - Contemporary philosophers
  - Kuhn and Lakatos
  - Feyerabend
  - Sokal and Bricmont
- Examples
  - Economics
  - Ethics
  - Evolution
  - Historicism
  - Logic and mathematics
  - Philosophy

Entities (View 2)

- ORGANIZATION
  - Walter F. Bischof (1)
  - Walter F. Bischof (1)
  - Wang (2)
  - Warren (4)
  - Warren Avenue Los Angles (1)
  - Watkin (1)
  - Weber (2)
  - Wei Chen Dept. (2)
  - Weinberger (1)
  - Weintraub (1)
  - Weiss (1)
  - Wellesley (1)
  - Werle (3)
  - WestMeade Drive Chesterfield (1)
  - Westchester Avenue White Plains (1)
  - Westendorpand Osterhaus (1)
- PERSON
  - Keller-McNulty
  - P. Kegelmeyer
  - Gaston
  - Weinberger
- MISC
- LOCATION

Hotlist (View 7)

- Keller-McNulty
- P. Kegelmeyer
- Gaston
- Weinberger

Hotlist Map (View 4)



# 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

The screenshot displays the Titan P2 0.8.0 (alpha) interface with several key components:

- Document Text (View 1):** Shows a Wikipedia article titled "Falsifiability" with a warning that it contains original research or unverified claims. The article text discusses the concept of falsifiability in science, mentioning Karl Popper and the term's origin.
- Clustering (View 3):** A network graph showing relationships between various documents and concepts, with nodes and edges representing connections.
- Entity list (View 2):** A list of entities extracted from the documents, including ORGANIZATION, PERSON, and LOCATION, with counts for each.
- 'Soap opera' hotlist (View 7):** A list of entities identified as 'hot' or significant, including names like Keller-McNulty, P. Kegeles, and Gaslon.
- Corpus Map (View 4):** A network graph showing relationships between documents and concepts, similar to the clustering view but with a different layout.

Entity list

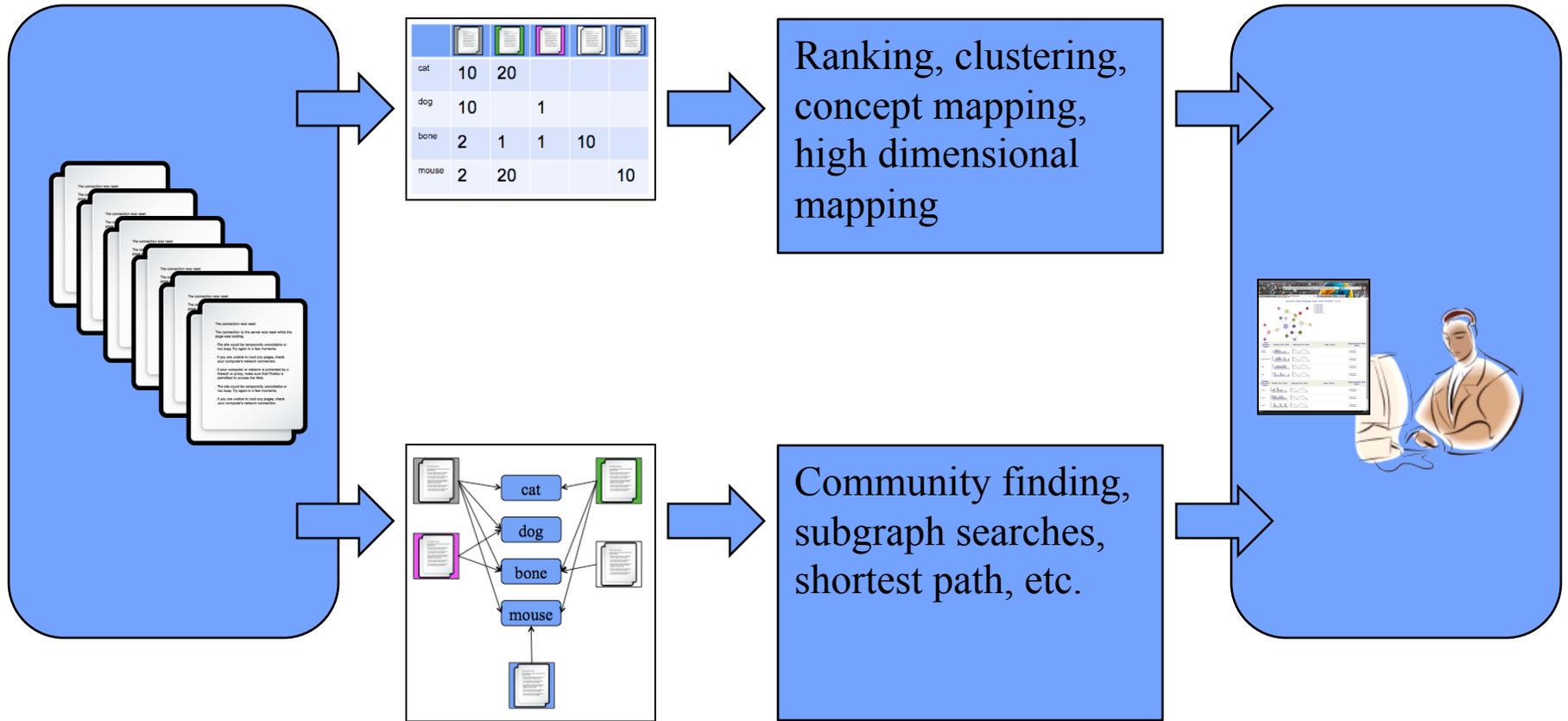
'Soap opera'

Corpus query network

Clustering



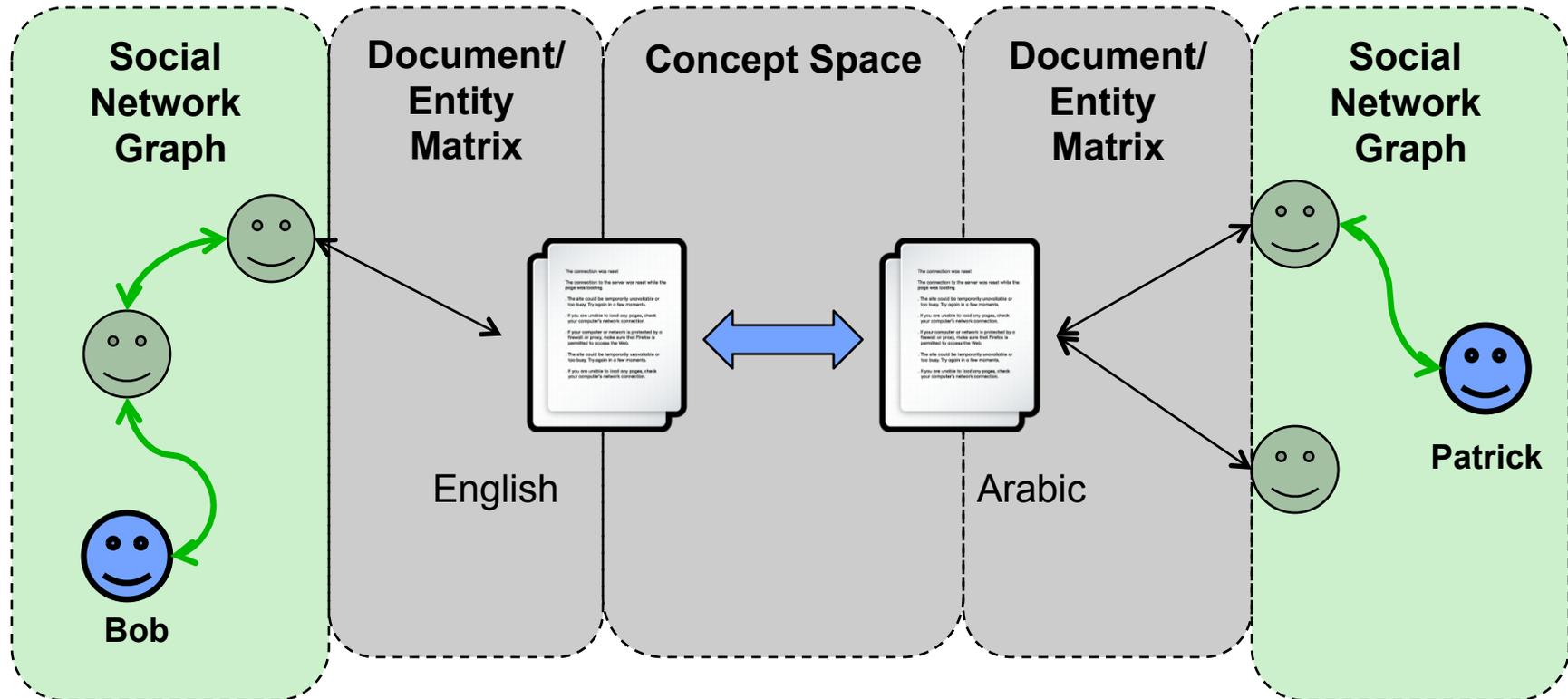
# 'Connecting the Dots' Benefits from a Range of Approaches



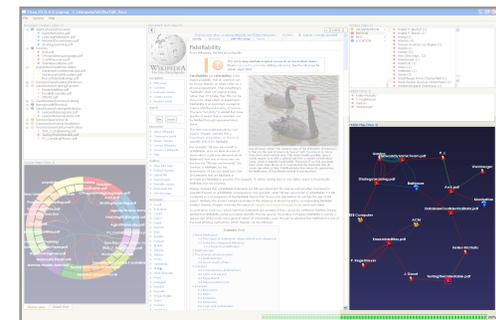
- **Supports rich relationship-centered analysis**
- **Combines large, heterogeneous data corpora**
- **Different abstractions support different analytics**



# 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

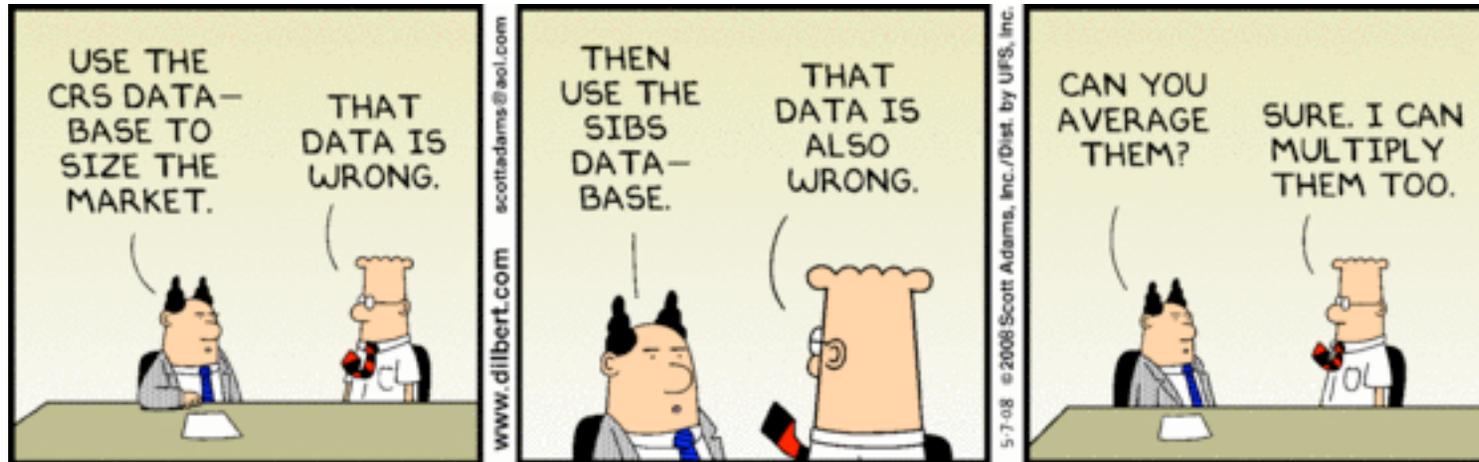


PII Prototype, emphasizing the combined graph and algebraic methods view



# 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?**