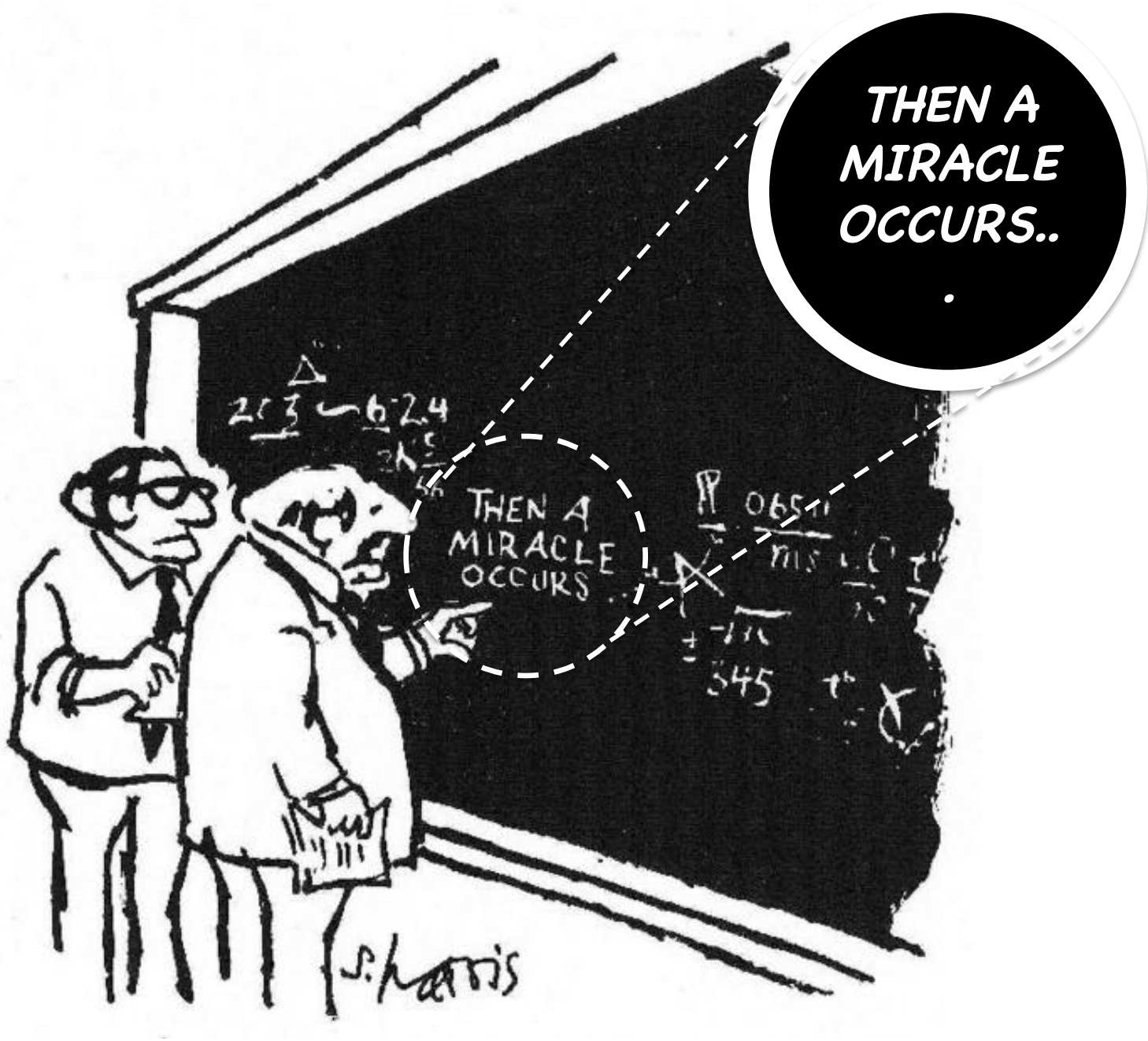
SAND2011-1593P

The Bigger the Data, the Harder it Falls: Current Problems in Large Data Analysis

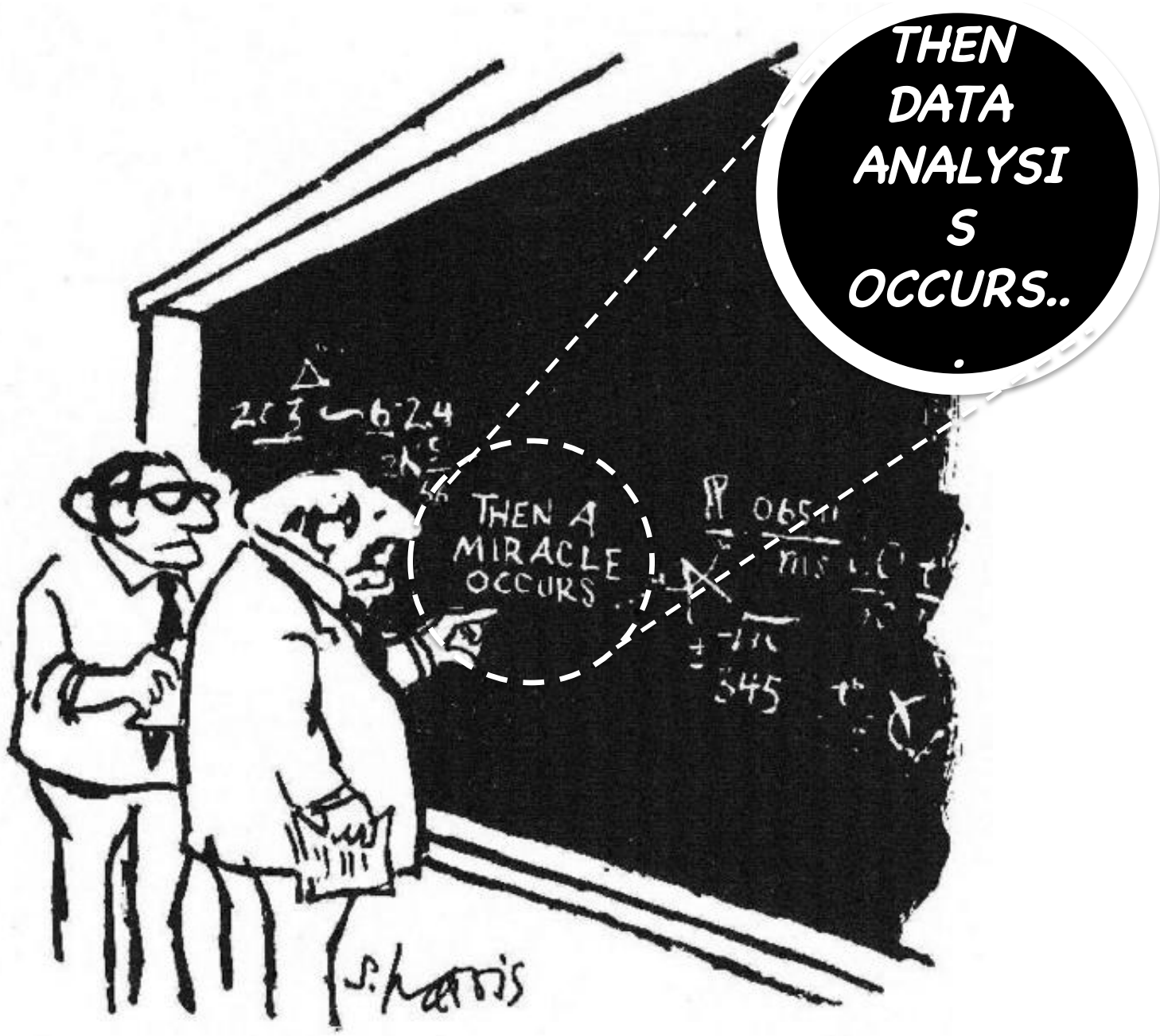
David H. Rogers

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.

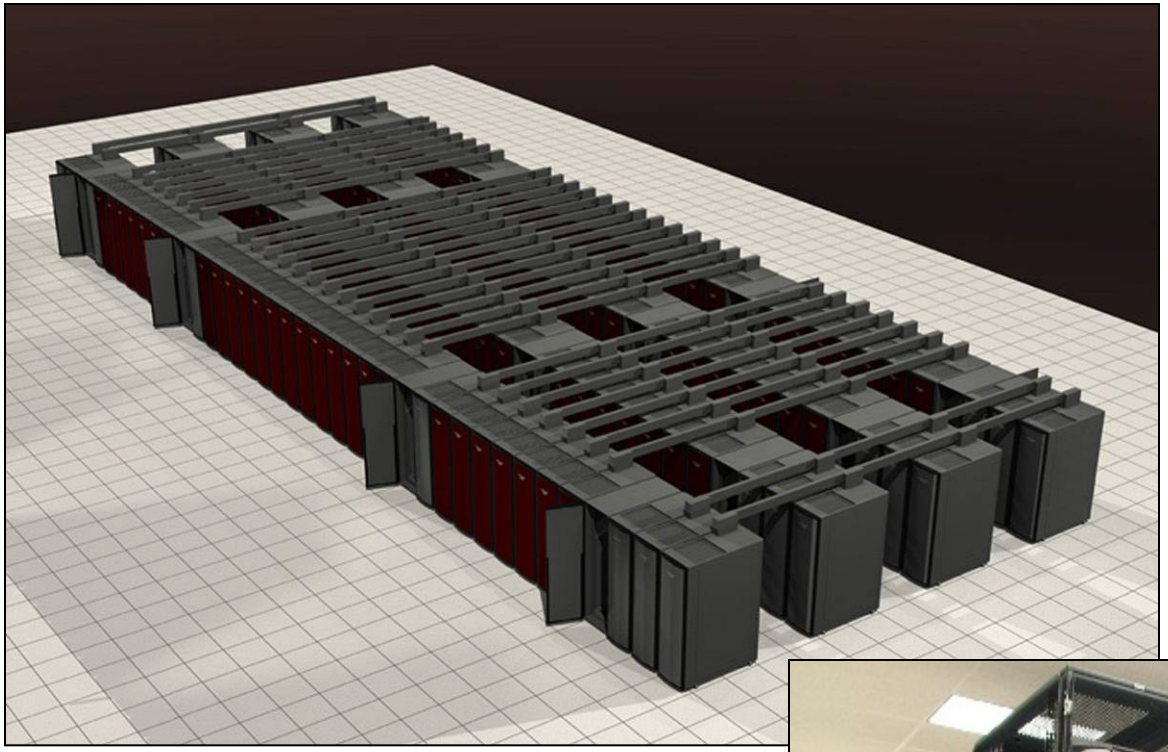


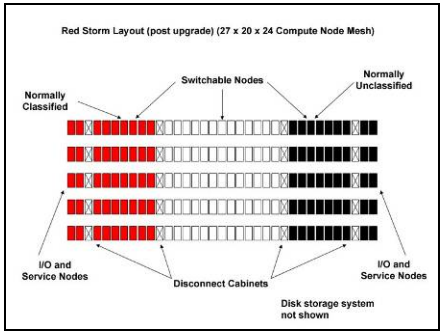
"I think you should be more explicit here in step two."



"I think you should be more explicit here in step two."

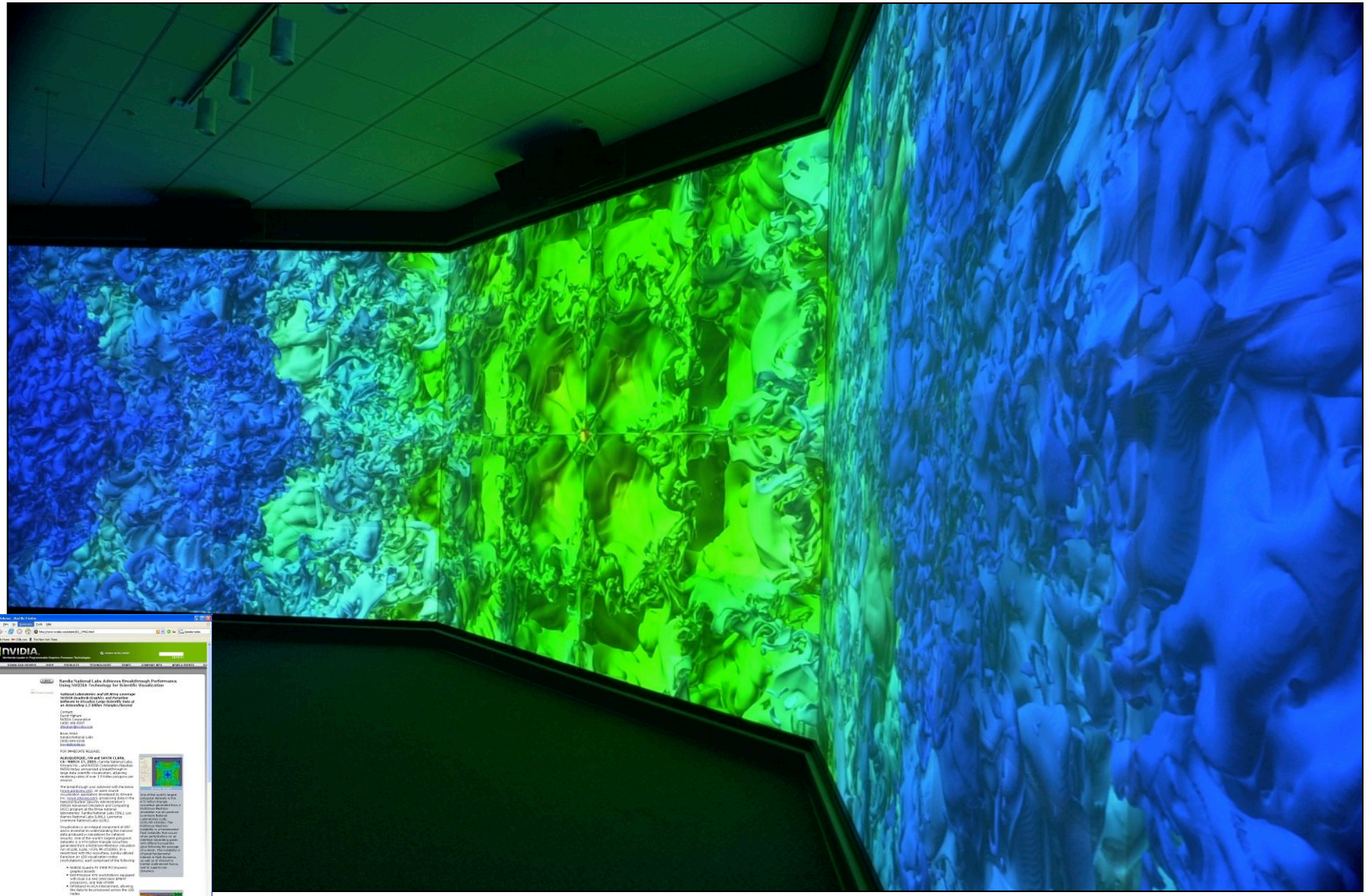
Interactive Visualization of Extremely Large Data





National Laboratories

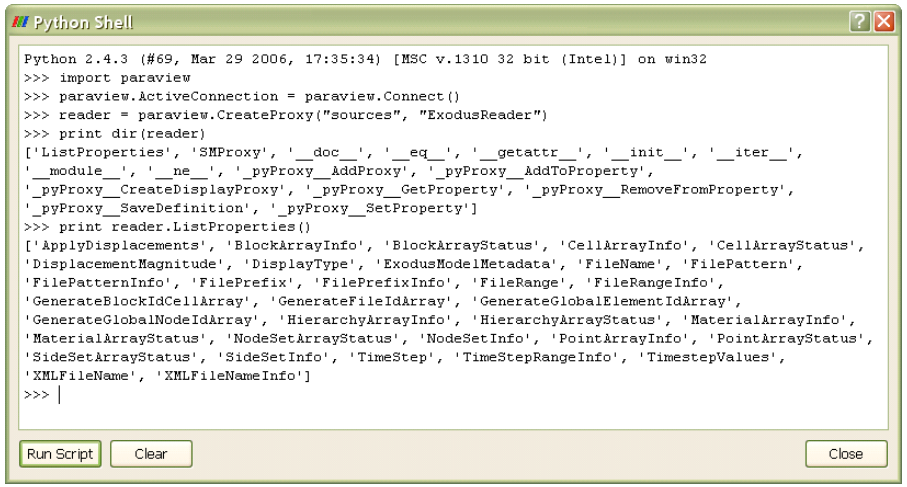
Interactive Visualization of Extremely Large Data

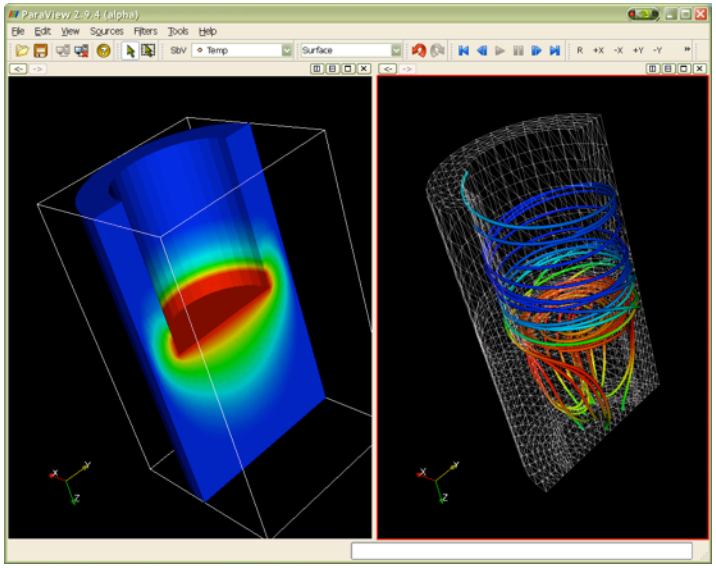


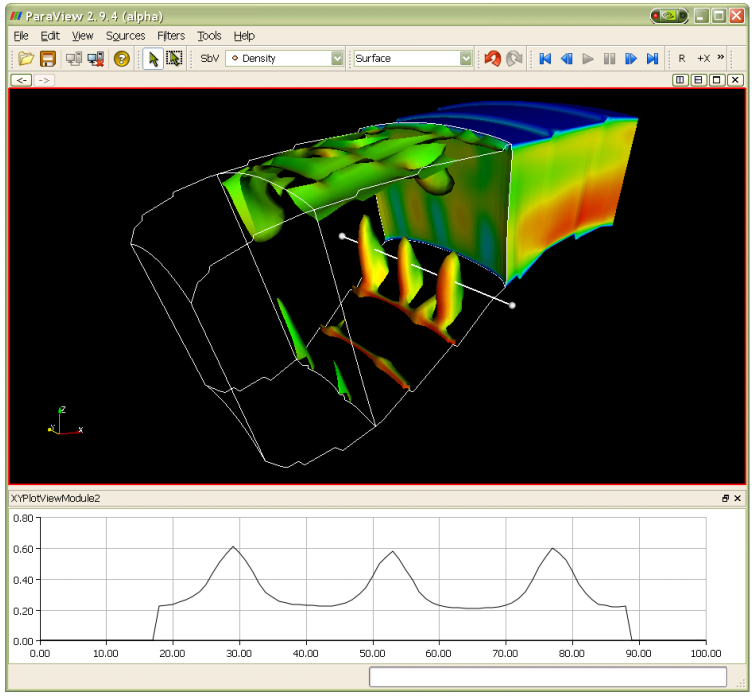
March, 2005: 473 million triangle mesh, 1.5 Billion polygons/sec.

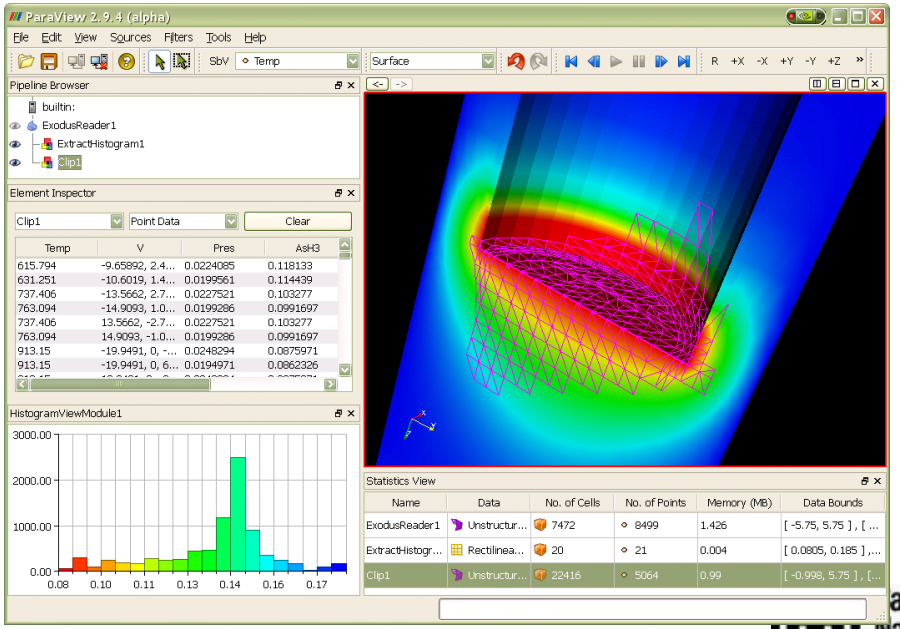
Sandia National Laboratories

Advancing Scalable Scientific Visualization with ParaView 3.0



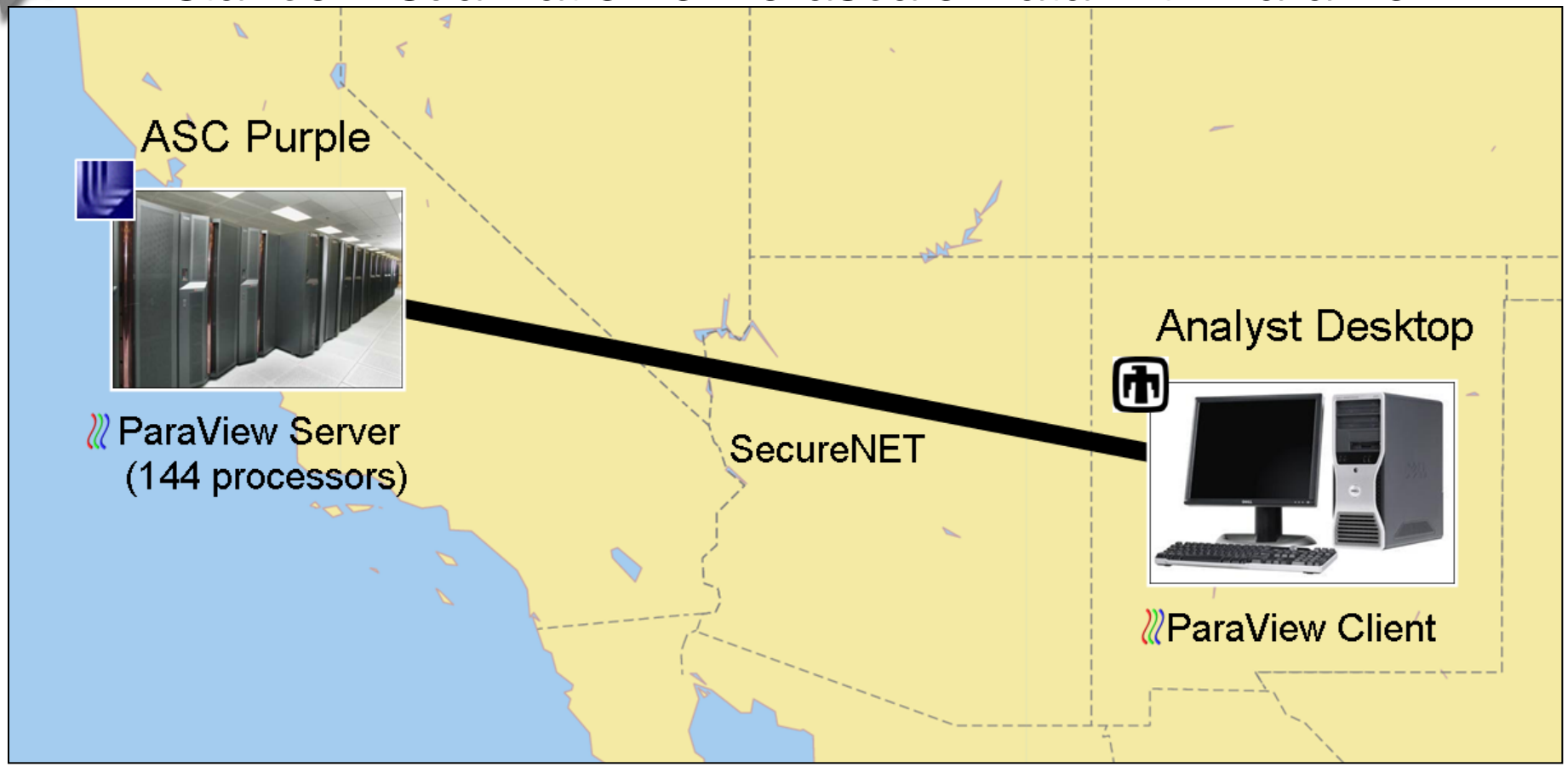






andia National Laboratories

Distance Visualization of Terascale Data with ParaView



- Out-of-the-box solution (despite complexity of hardware)
- ParaView 2x faster than other solutions (27 million unstructured cells)
- ParaView only workable solution for large data (274 million unstructured cells)

Sandia National Laboratories

Principles Resulting from Large Scientific Data Analysis

Scalability

- Run the same software on everything from laptop to cluster
  - Data size shouldn't matter to the analyst
- Algorithms, data structures and visualization must adapt to data size
- Client (user-side) software can connect to appropriate remote hardware

- Tools must be cross-platform
  - Linux, Windows, Mac

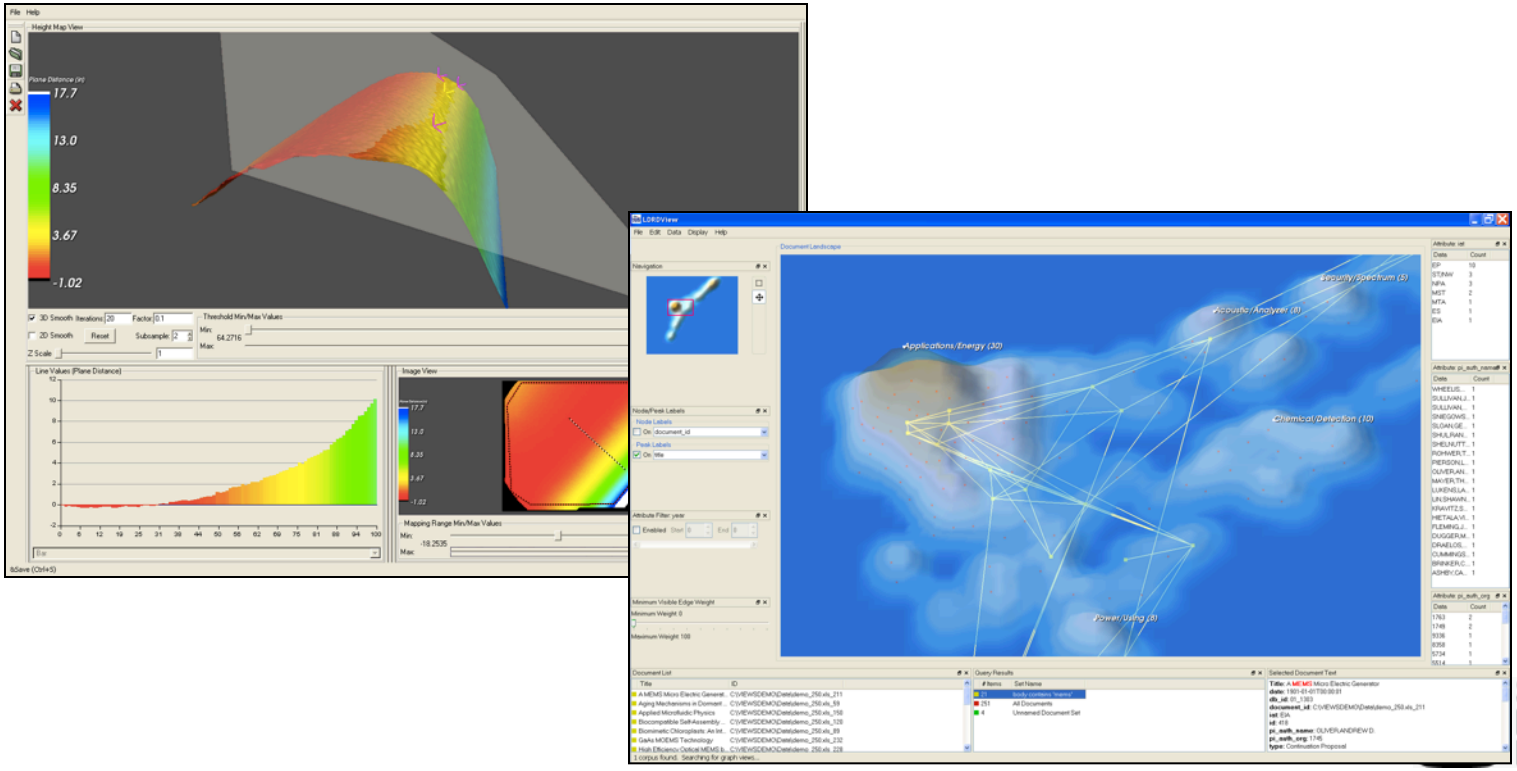
- Responsive innovation leads to success
  - Balance between research, long term planning and immediate problems

Sandia National Laboratories

Targeted Tool Development

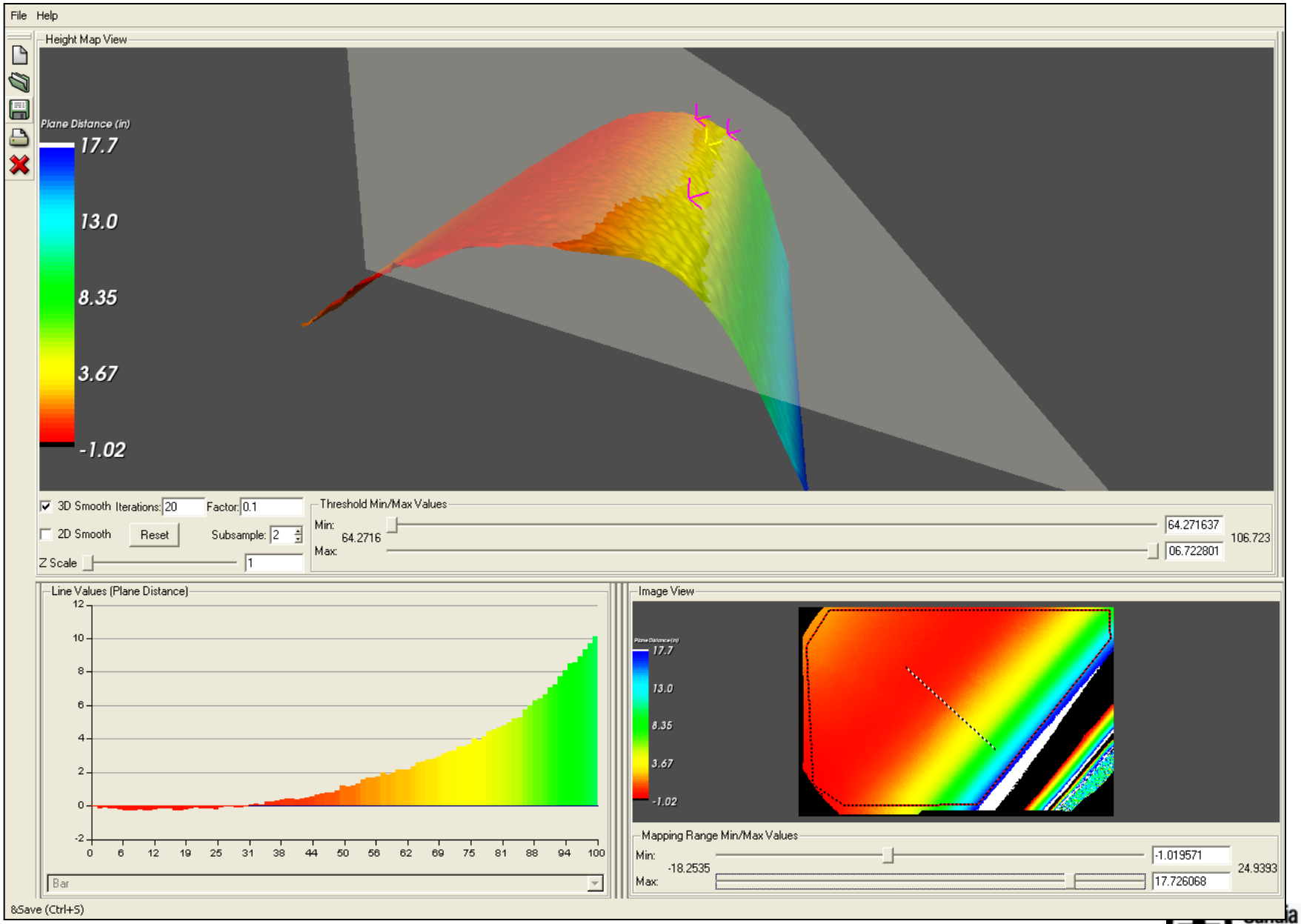
Rapid Prototyping, with flexible technologies (Qt, VTK, etc.)

- Promotes interctive, cross-platform tool development
- Attacks the ‘real problems’
- Addresses customer concerns up front
- Promotes creative, collaborative developement



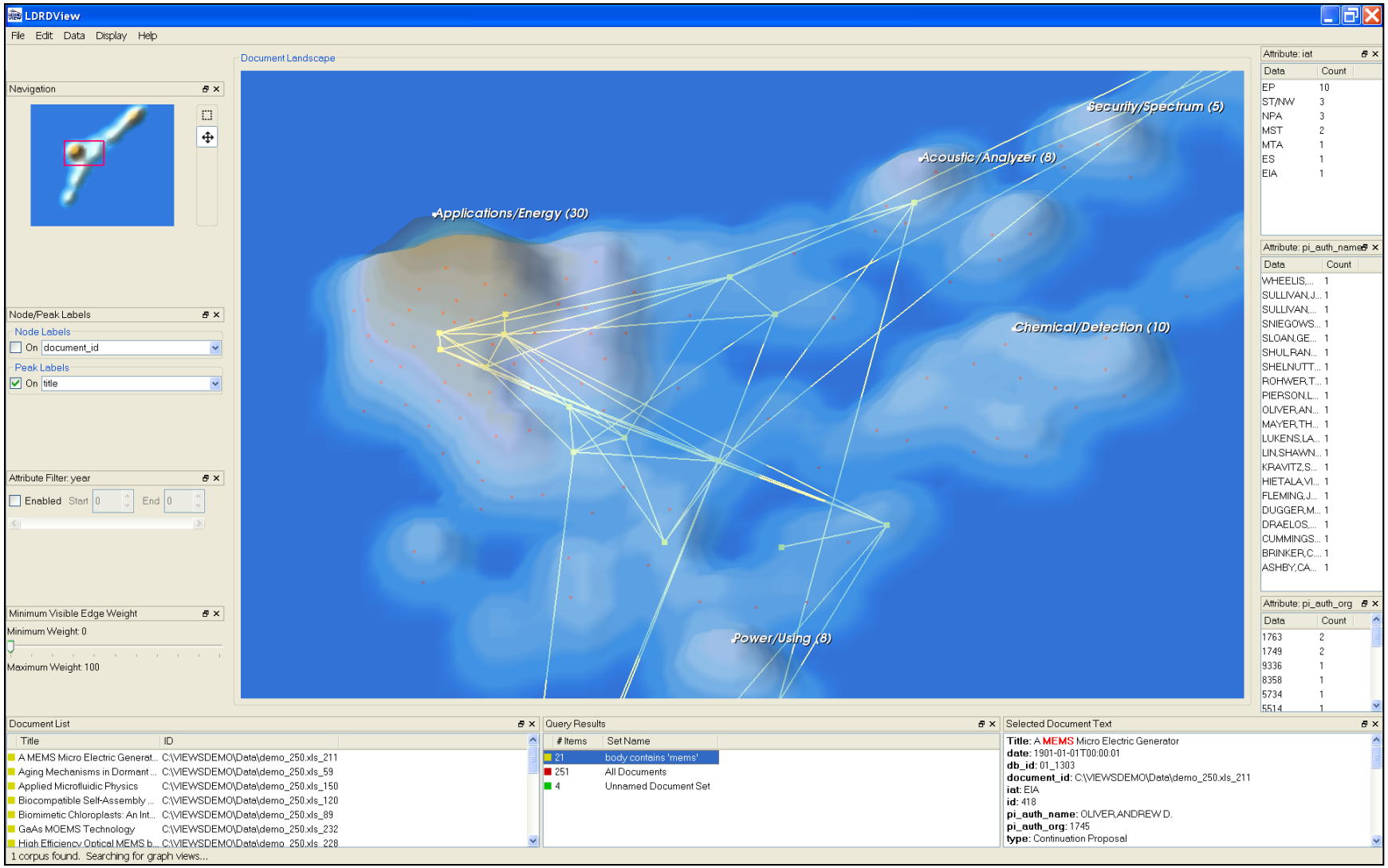
Sandia National Laboratories

Analysis of NASA Telemetry Data (Leading Edge of Orbiter Wings)



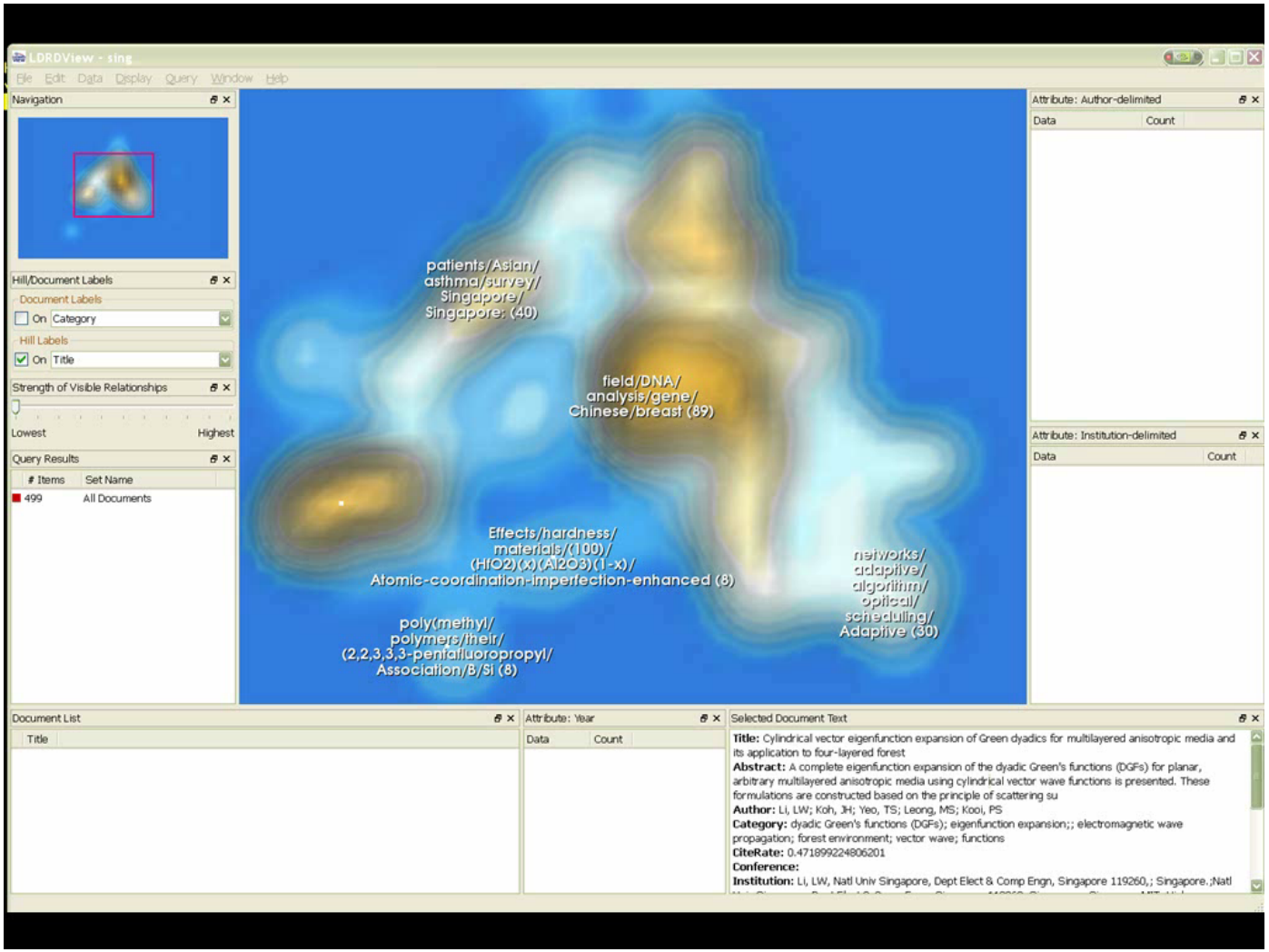
Sandia National Laboratories

Accelerating Insight into Complex Data with LDRDView



Sandia National Laboratories

Accelerating Insight into Complex Data with LDRDView



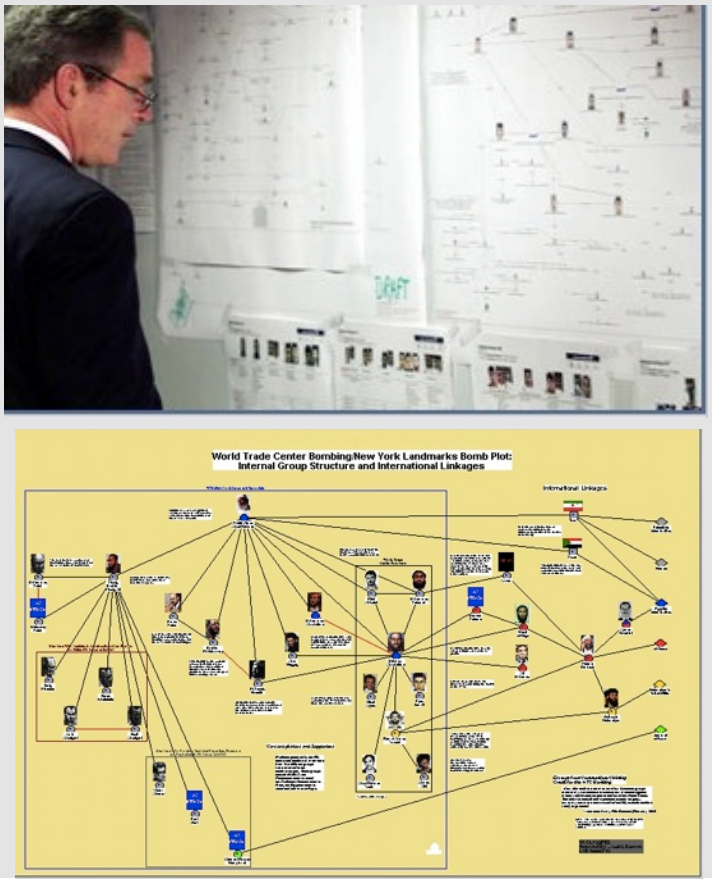
Sandia National Laboratories

The Networks Grand Challenge?

Sandia National Laboratories

"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).**



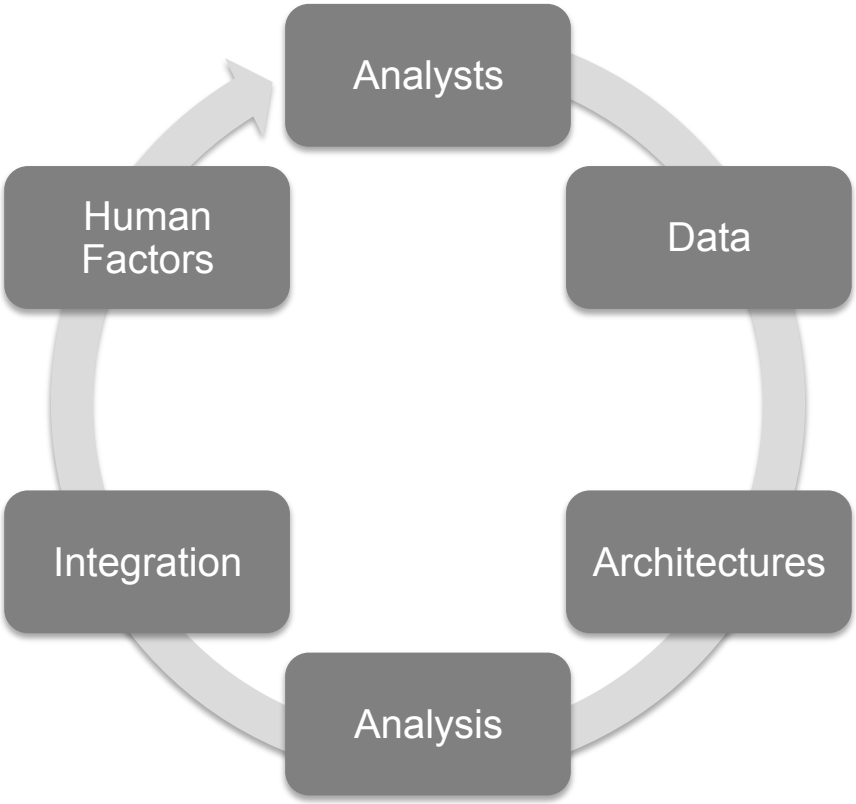
President Bush looks over a chart depicting Osama bin Laden's financial network; an excerpt from the chart (created using one of the leading tools, i2 Analyst's Notebook)

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

Sandia National Laboratories

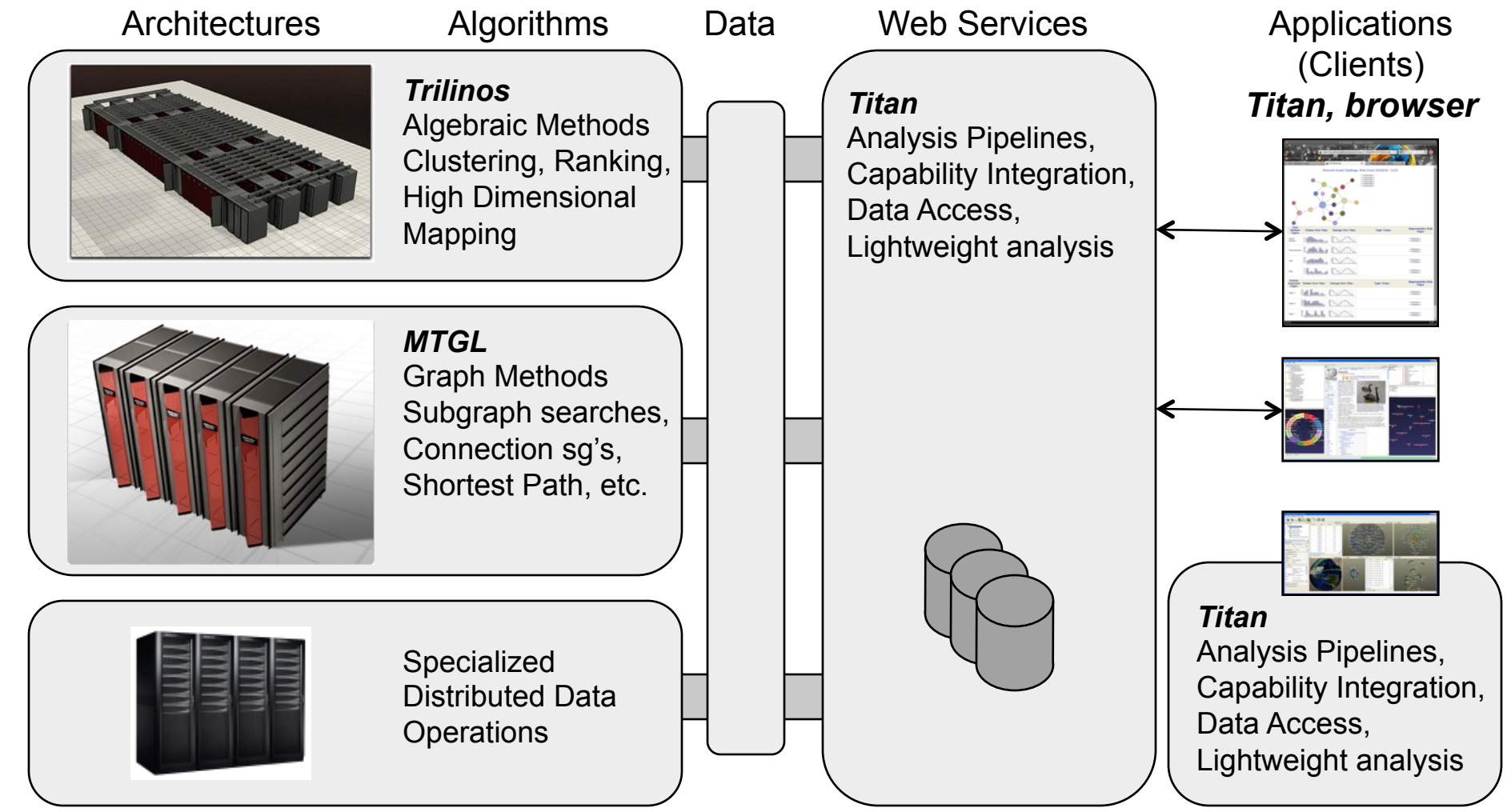
NGC Gathers Expertise from across the Labs, working on an End-to-end Approach

- **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.

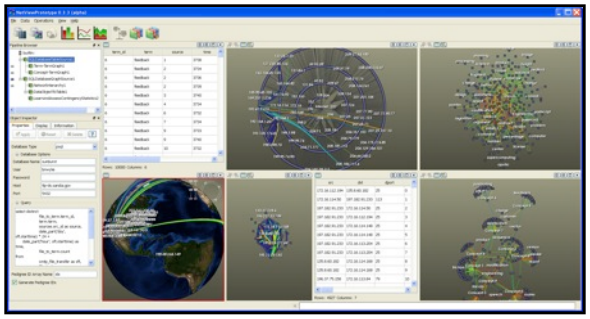


NGC System Diagram

"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

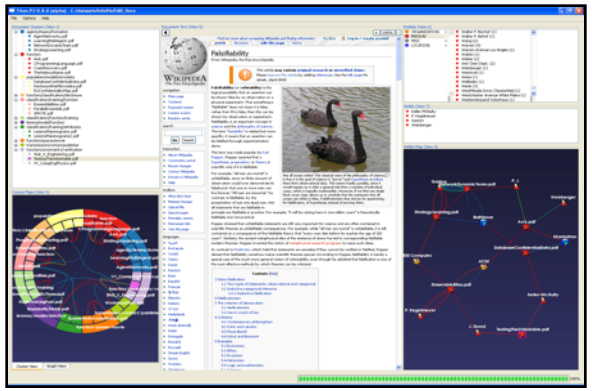


NGC's Commitment to Prototypes Promotes End-to-end Integration



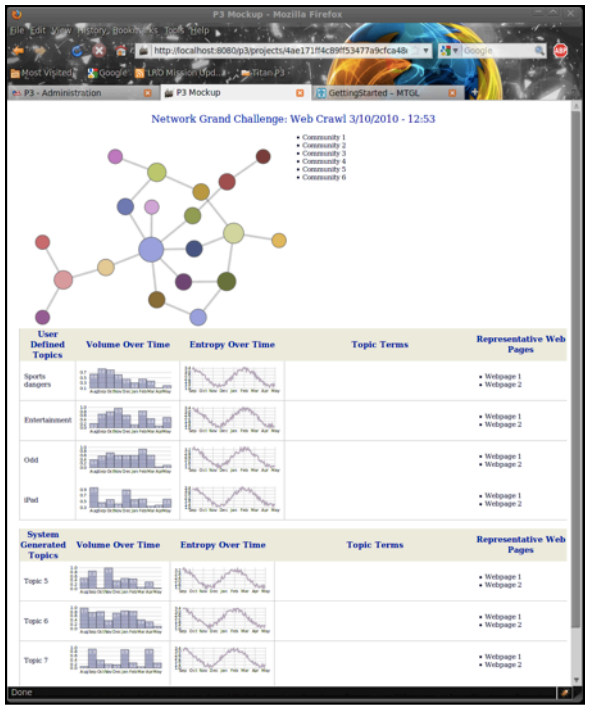
PI: Cyber Application

- 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

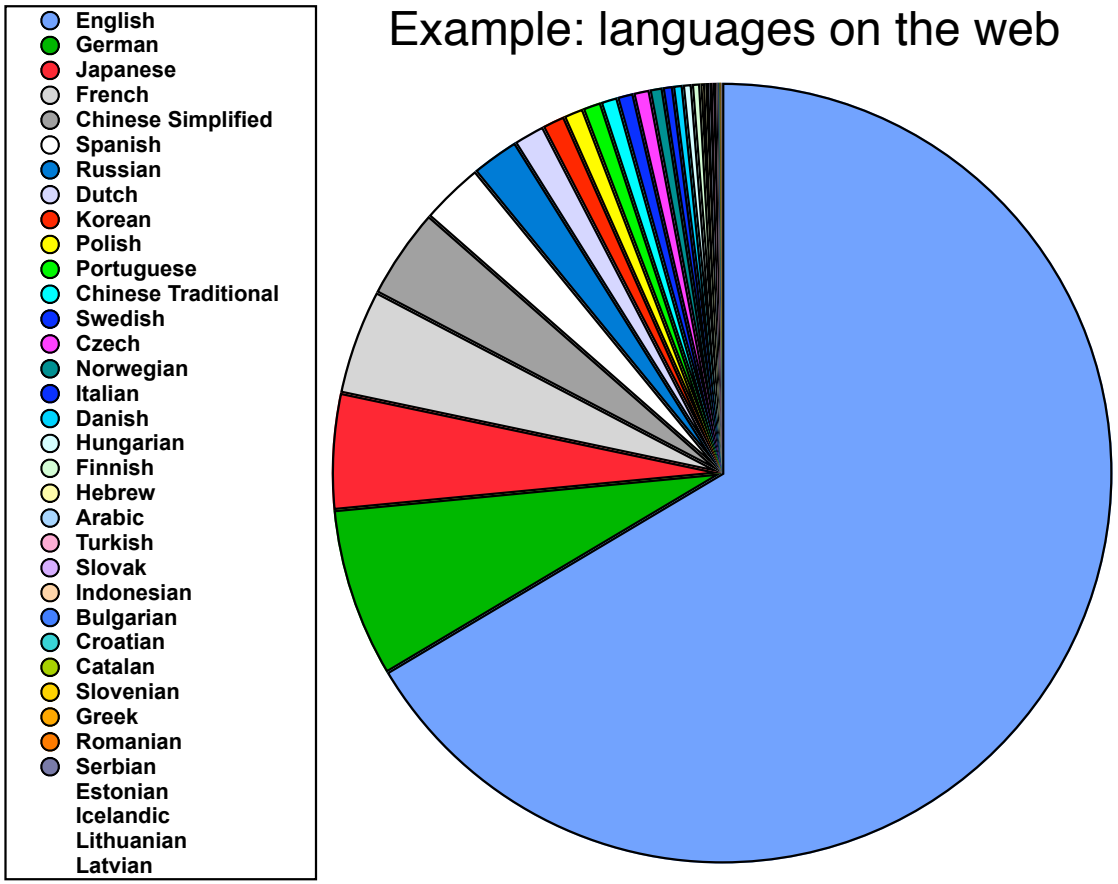
NGC research thrust: Multilingual Text Analysis

Sandia National Laboratories

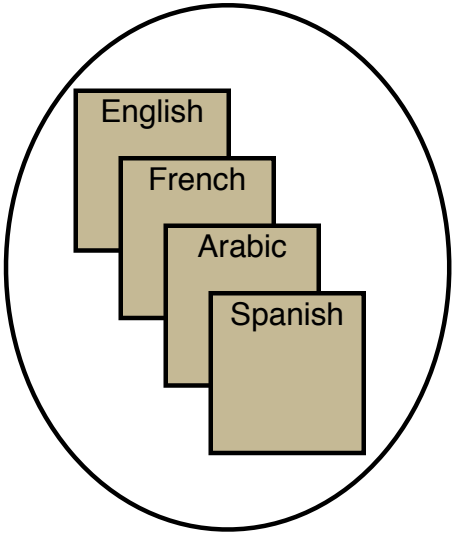
Cross-language Information Retrieval (CLIR)

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

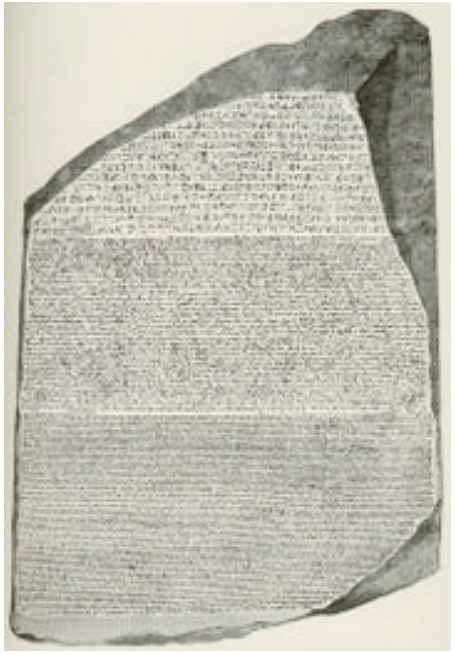
Sandia National Laboratories

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

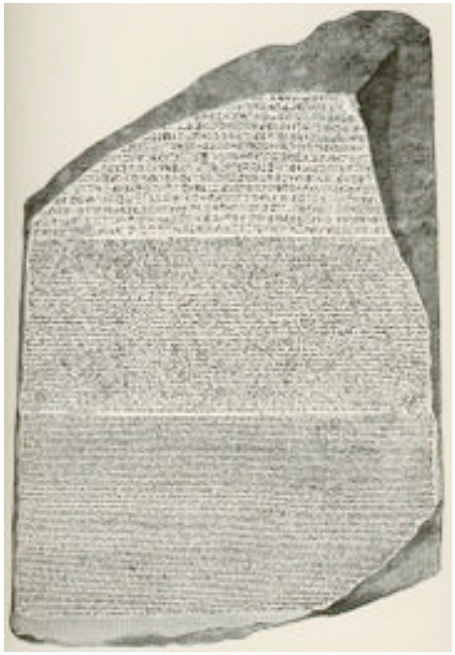
Afrikaans	Estonian	Norwegian
Albanian	Finnish	Persian (Farsi)
Amharic	French	Polish
Arabic	German	Portuguese
Aramaic	Greek (New Testament)	Romani
Armenian Eastern	Greek (Modern)	Romanian
Armenian Western	Hebrew (Old Testament)	Russian
Basque	Hebrew (Modern)	Scots Gaelic
Breton	Hungarian	Spanish
Chamorro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

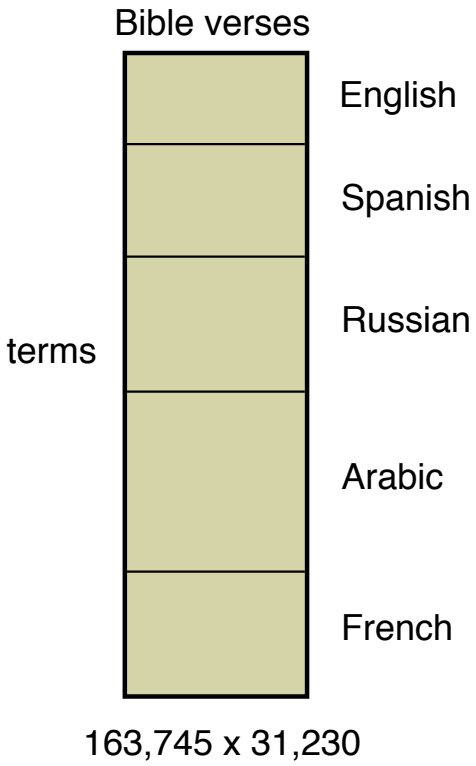


Sandia National Laboratories

Term-Doc Matrix

Term-by-verse matrix
for all languages





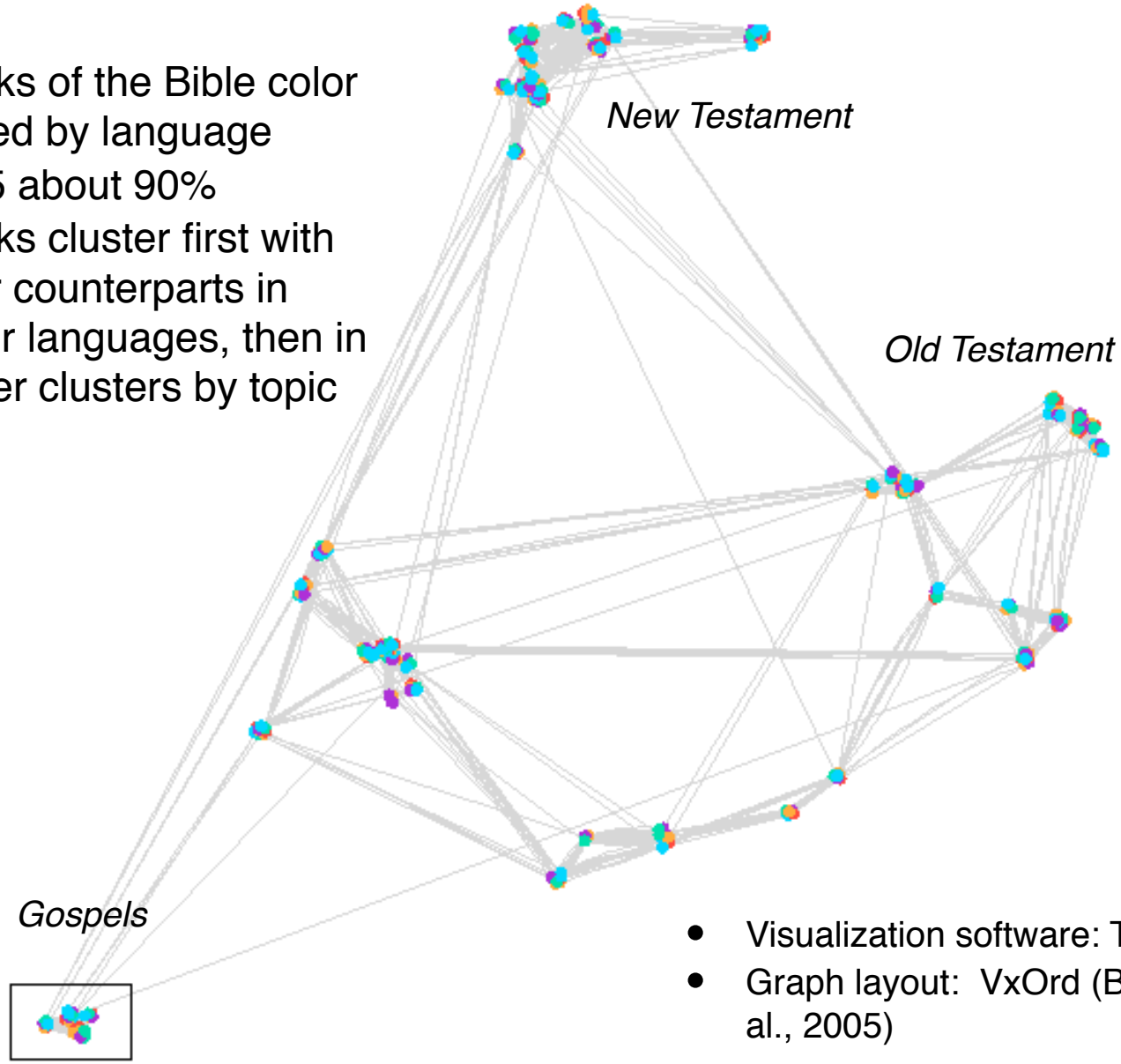
Look for co-occurrence of terms in the same verses and across languages to capture latent concepts

- 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

Sandia National Laboratories

Bible Clustering with LMSATA

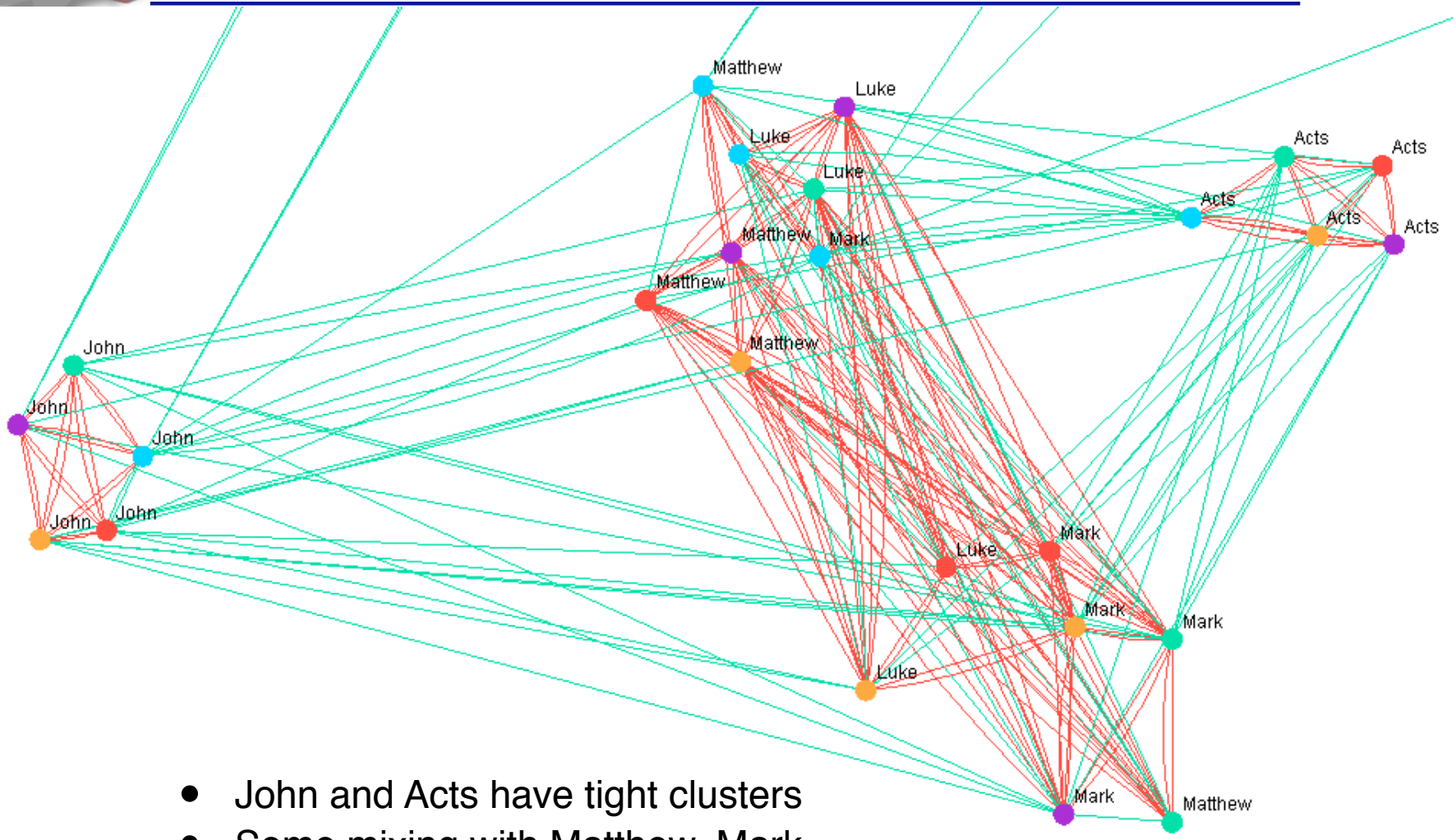
- Books of the Bible color coded by language
- MP5 about 90%
- Books cluster first with their counterparts in other languages, then in larger clusters by topic



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

Sandia National Laboratories

Clustering Close-up



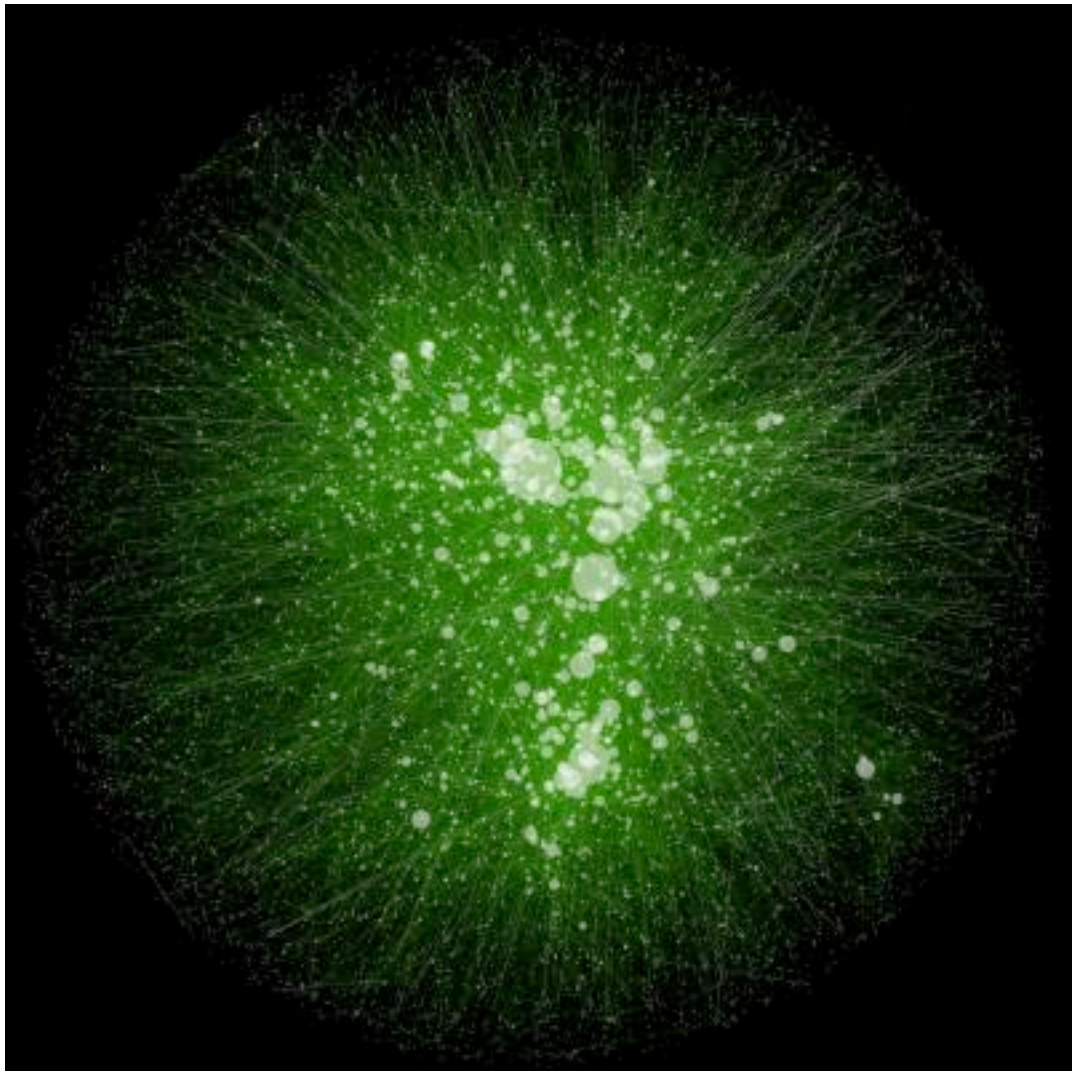
- John and Acts have tight clusters
- Some mixing with Matthew, Mark, Luke (synoptic gospels - share a similar perspective)

Sandia National Laboratories

All (Large) Data Analysis is Highly, Immediately Constrained

- Raw data must be processed
  - Example: database schema
- Analysis Algorithms op on specific data types
  - Graphs, tensors, images
- Advanced architectures support certain ops
  - XMT: multithreaded (graphs)
  - Netezza: hardware execution of SQL
  - Distributed memory: large tensor operations
- Analysis results viewed, queried many ways
  - Layout/vis introduces bias
- Human in the loop is a known unknown
  - *Especially in non-scientific arenas*

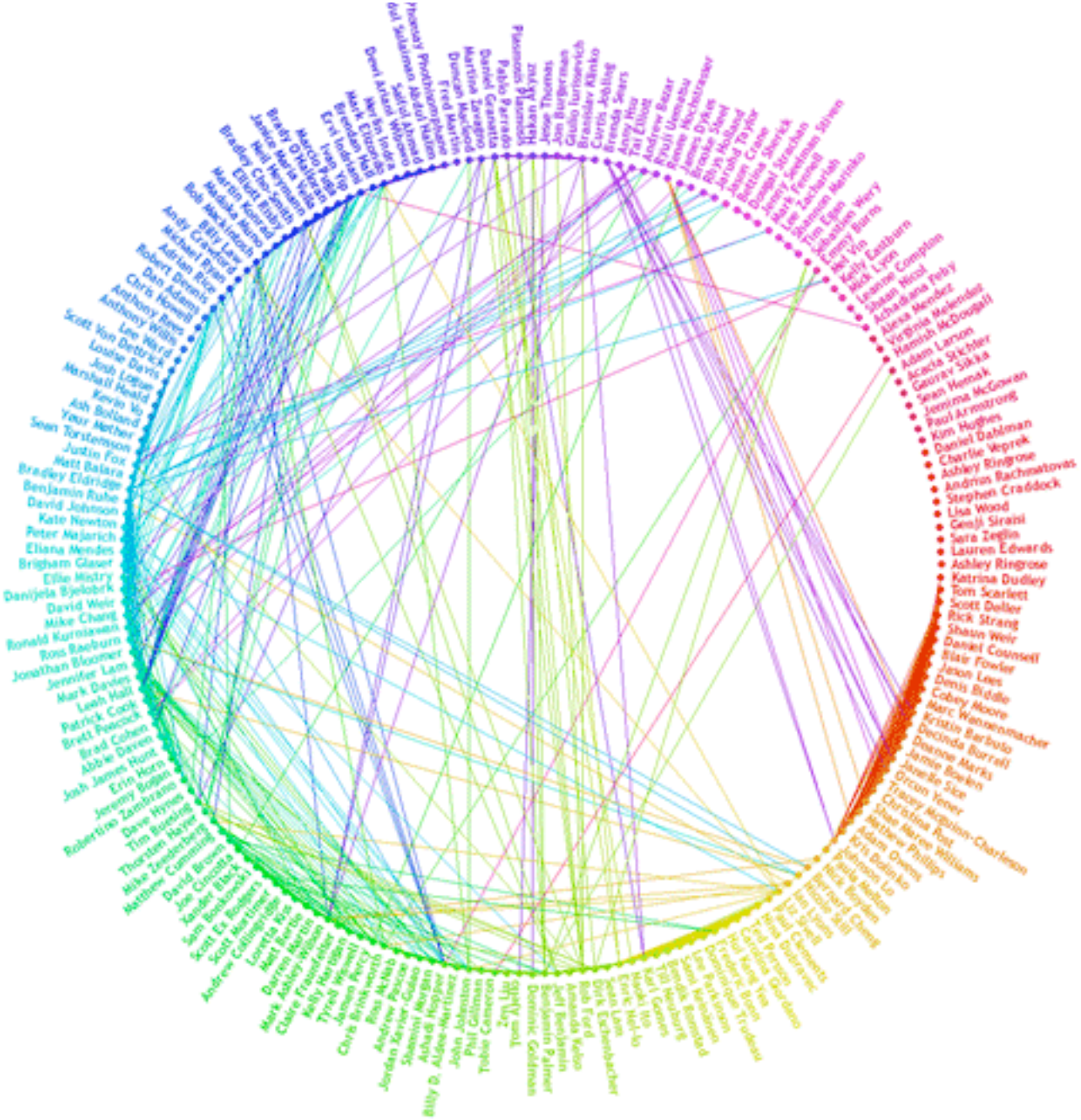
The Hairball Principle



This image, created by Trey Ideker using Cytoscape, depicts thousands of known molecular and genetic interactions occurring inside the human body.

http://www.physorg.com/news/2010-10-center-human-function.html

Sandia National Laboratories



http://www.bannerblog.com.au/news/2008/08visualizing_your_life_the_non_new_age_way.php

Sandia National Laboratories

Mapping of intensity of Facebook friends between pairs of cities, with the connections drawn as great arcs.

After a few minutes of rendering, the new plot appeared, and I was a bit taken aback by what I saw. The blob had turned into a surprisingly detailed map of the world. Not only were continents visible, certain international borders were apparent as well. What really struck me, though, was knowing that the lines didn't represent coasts or rivers or political borders, but real human relationships. Each line might represent a friendship made while travelling, a family member abroad, or an old college friend pulled away by the various forces of life.

Paul Butler, in a blogpost: **http://www.facebook.com/notes/facebook-engineering/visualizing-friendships/469716398919**

Sandia National Laboratories

UX notes

- Real data is messy. *Very* messy
- Interesting data is hopeless to view in its entirety
  - This requires interactivity across CS
    - Algorithms, architectures, data movement, vis
- Understanding data is a process, not a product
- Customers must understand the algorithms, so they do not infer unsupported information from the presentation

Elbow-to-elbow collaboration is essential to understanding large data – iterative development is essential.

The process and the data are messy – the software CANNOT be

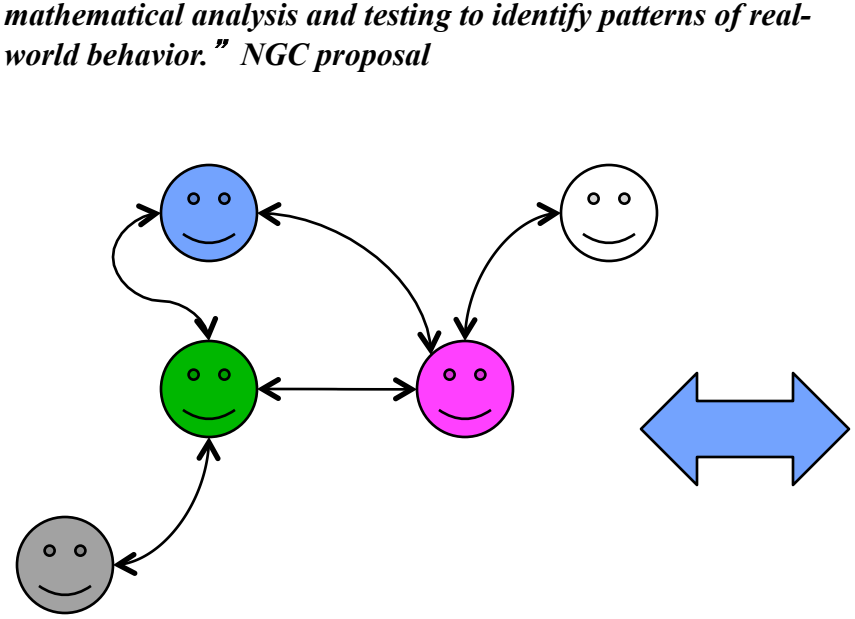
Sandia National Laboratories

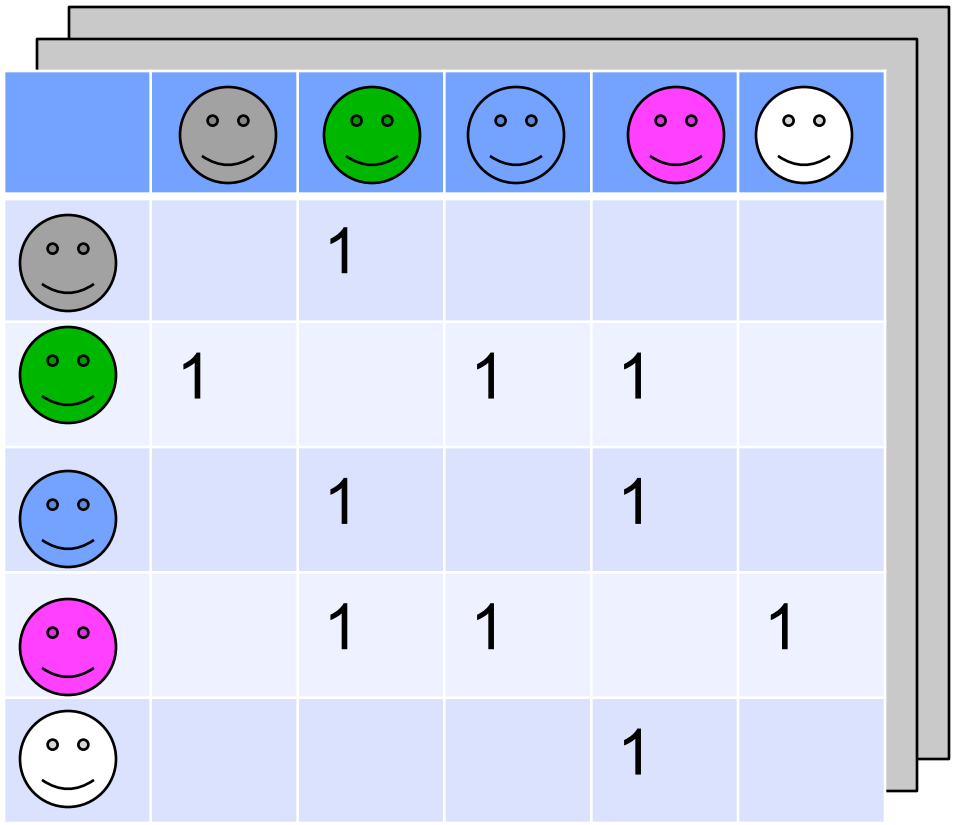
Unifying Data Abstractions

Sandia National Laboratories

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

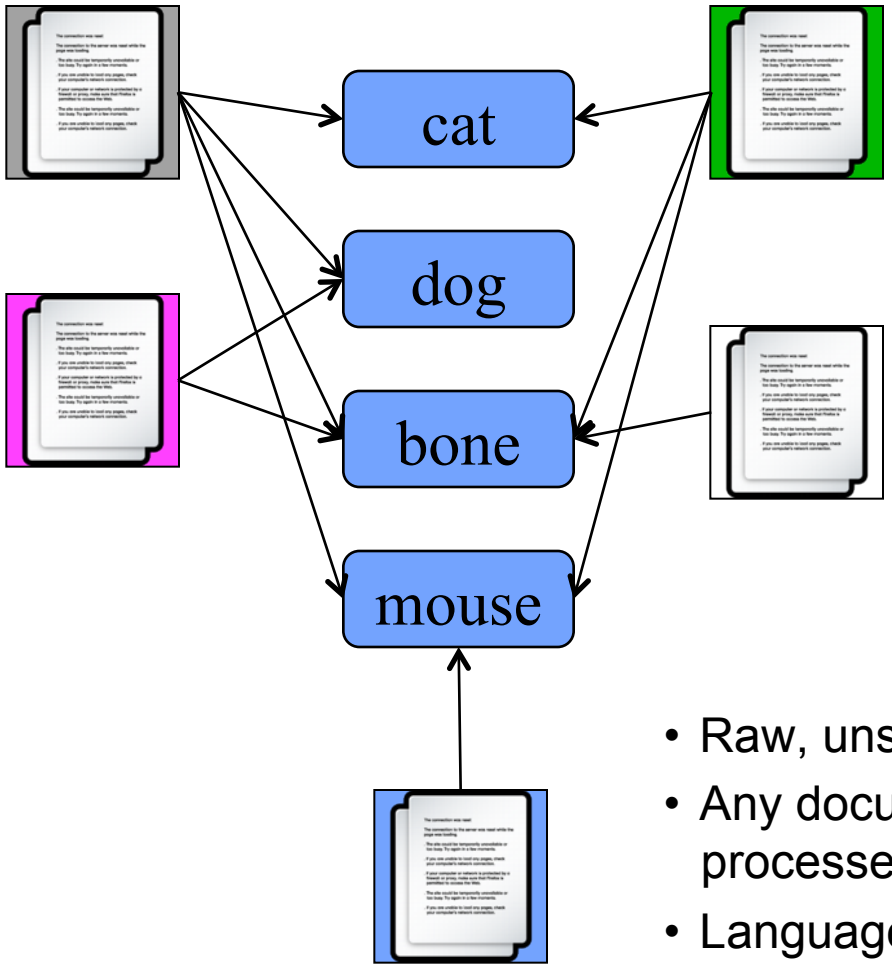




- Networks can be of many types
  - Social network
  - Cyber traffic
  - Communications
- Graph and matrix/tensor are equivalent representations
  - Extends to multiple dimensions
- Data can be easily transformed
- Both graph and algebraic algorithms can be run on the same data
  - appropriate architectures

Sandia National Laboratories

Text is Easily Transformed into Graph and Matrix Data Structures

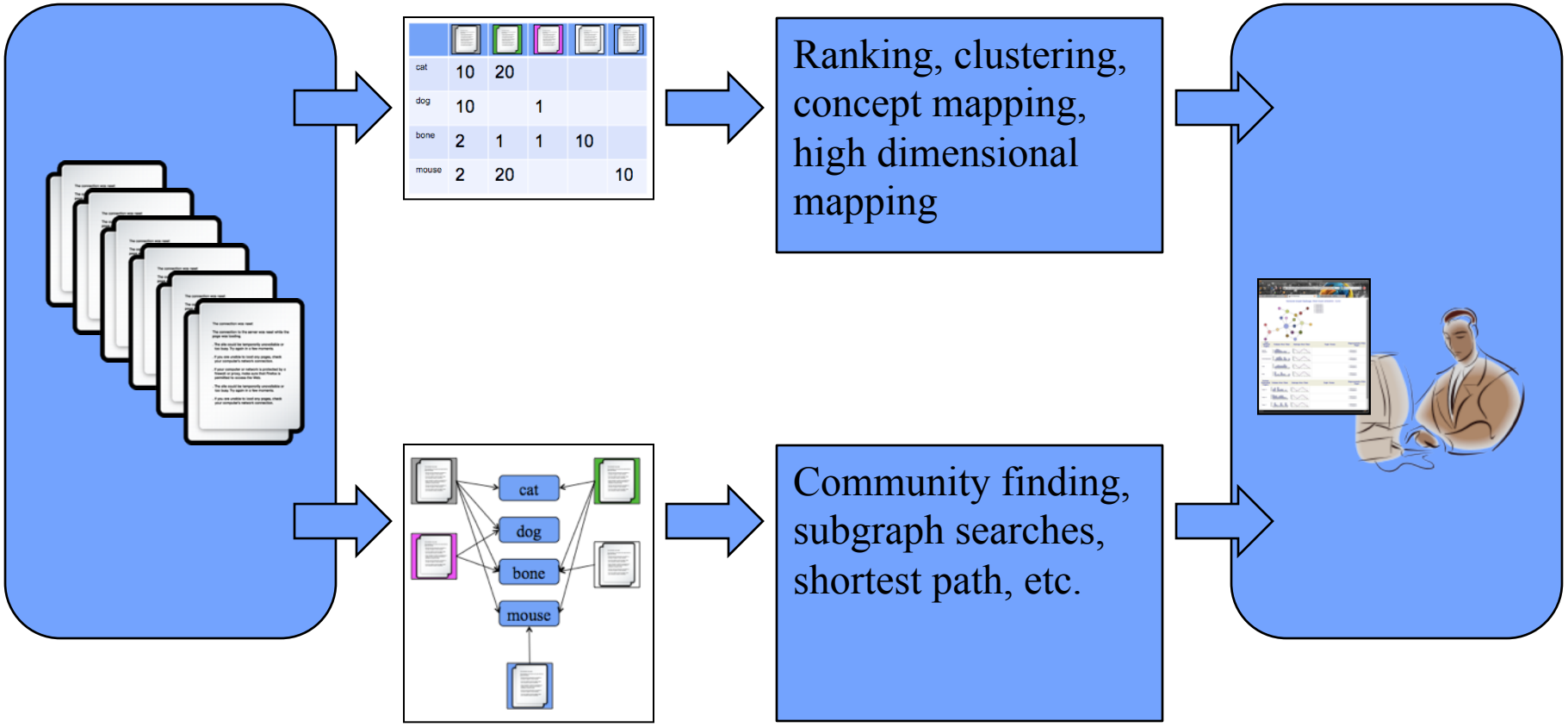


cat	10	20			
dog	10		1		
bone	2	1	1	10	
mouse	2	20			10

- 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

Sandia National Laboratories

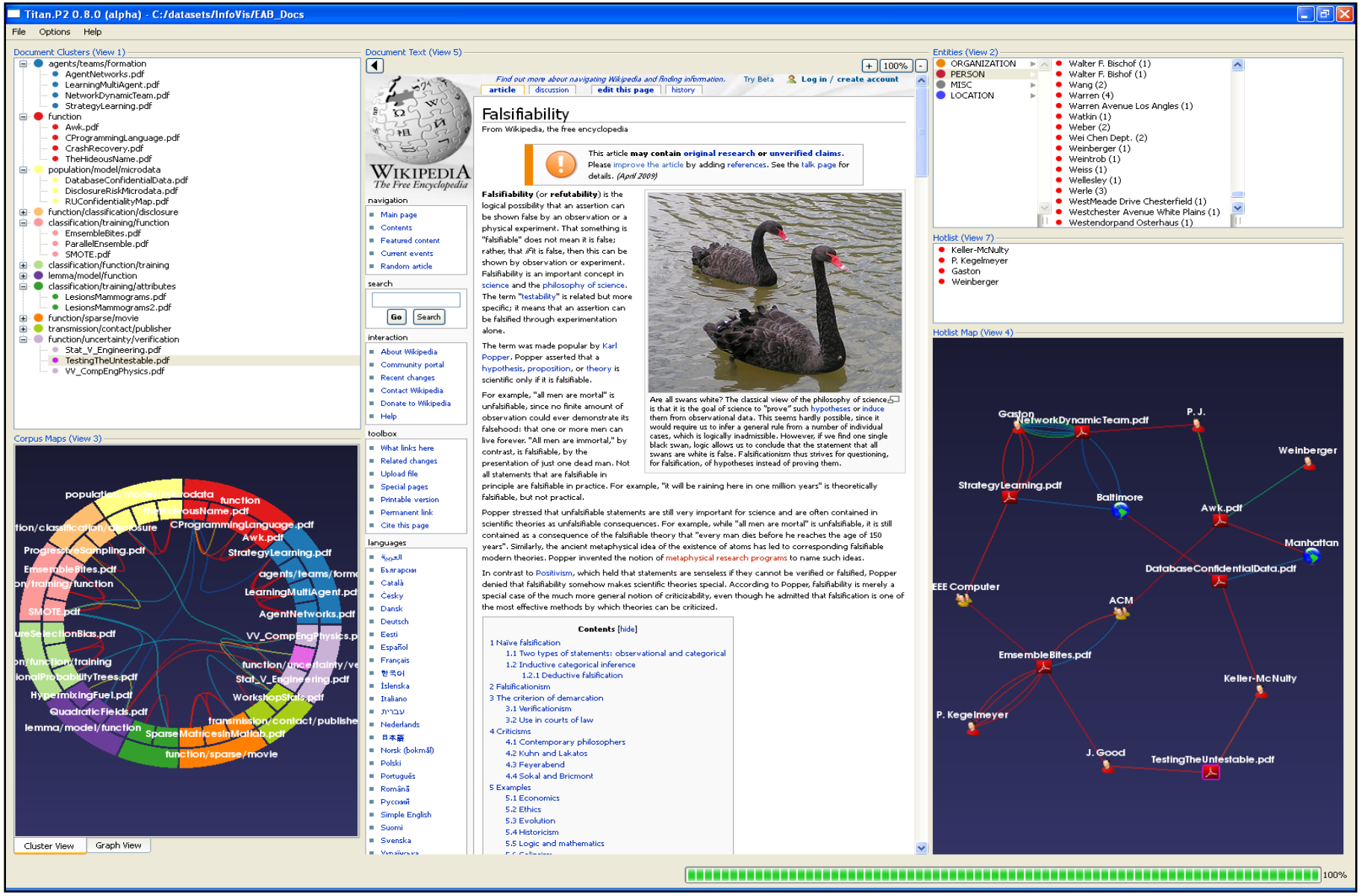
‘Connecting the Dots’ Benefits from a Range of Approaches



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

Sandia National Laboratories

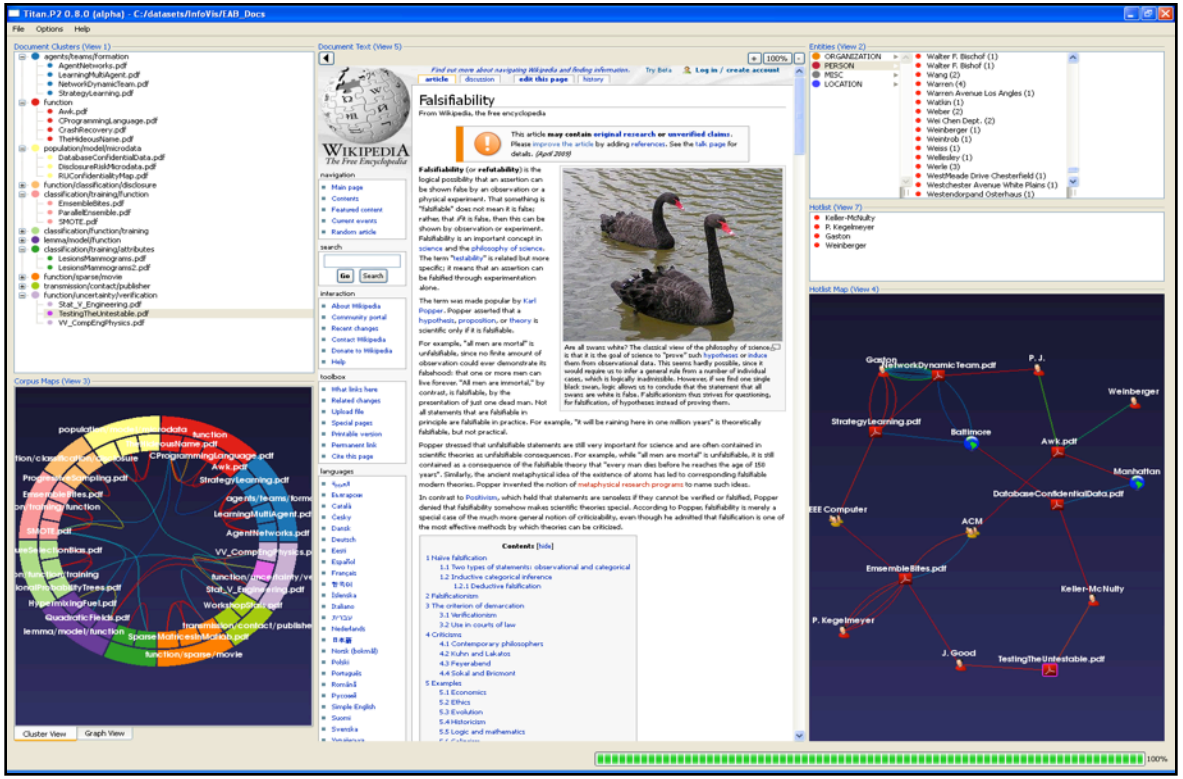
PII: Integrating Algebraic and Graph Methods



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



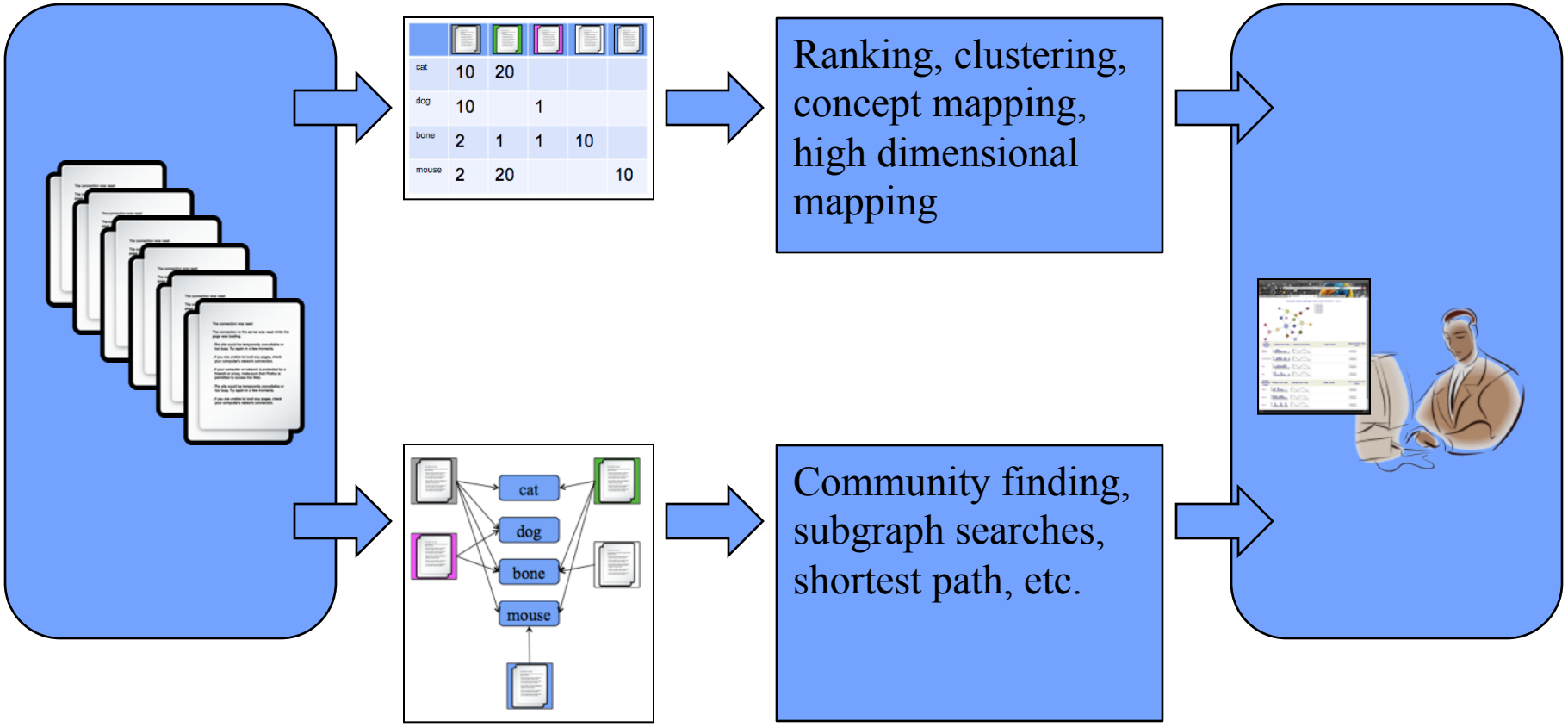
Clustering

Entity list

‘Soap opera’

Corpus query network

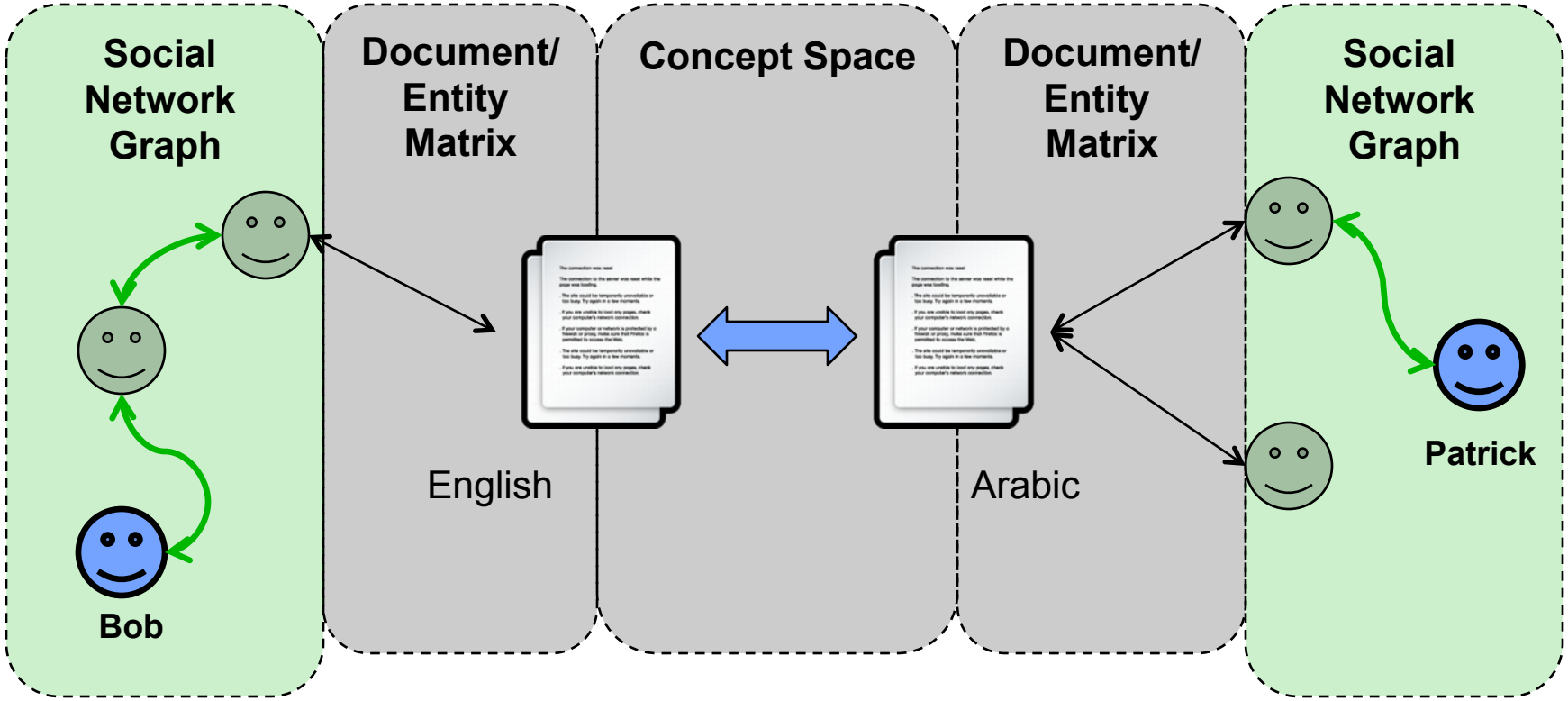
‘Connecting the Dots’ Benefits from a Range of Approaches



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

Sandia National Laboratories

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

UX notes

- Complexity arises from algorithms and data structures, as well as raw data
- Complexity comes from the human as well
  - Cognitive models, new intuitions, and *personal bias*

Sandia National Laboratories

P3 - Administration | P3 Prototype 3 | P3 Prototype 3 | Hex-to-RGB Color ... | Gregorian | Protovis - Colors | SRN Cactus Datab...

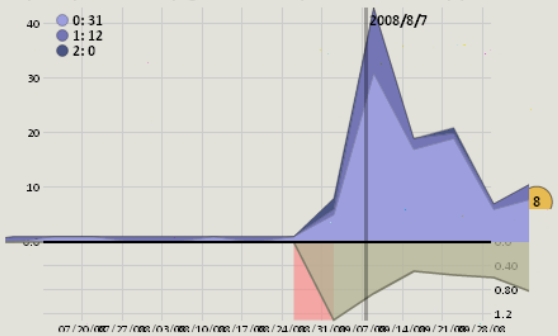
http://localhost/p3/projects/998c89927ef0ba802eabb34406b42bad42dd3980/reports/

CCHD | CSU Information | SEEMS Web Mail | SNL Directory | SNL Index | Windows Marketplace | Other bookmarks

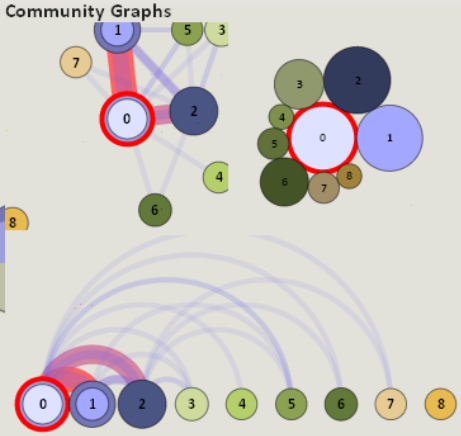
Network Grand Challenge: Web Crawl <Known Topics>

Topic Analysis | Data Statistics

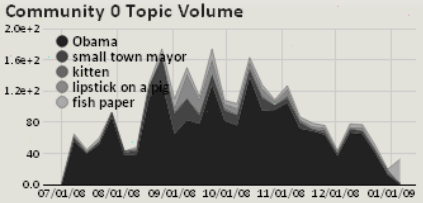
Topic 'lipstick on a pig' Community Volume / Entropy



Community Graphs



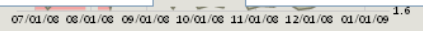
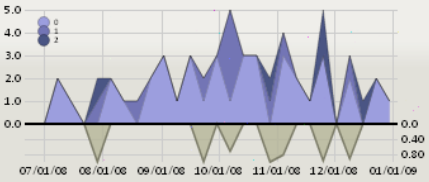
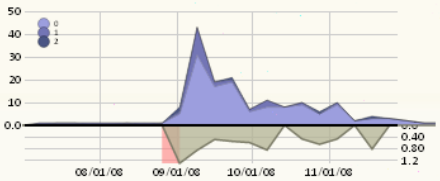
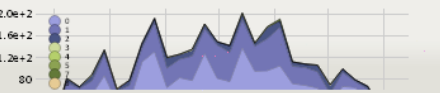
Community 0 Topic Volume

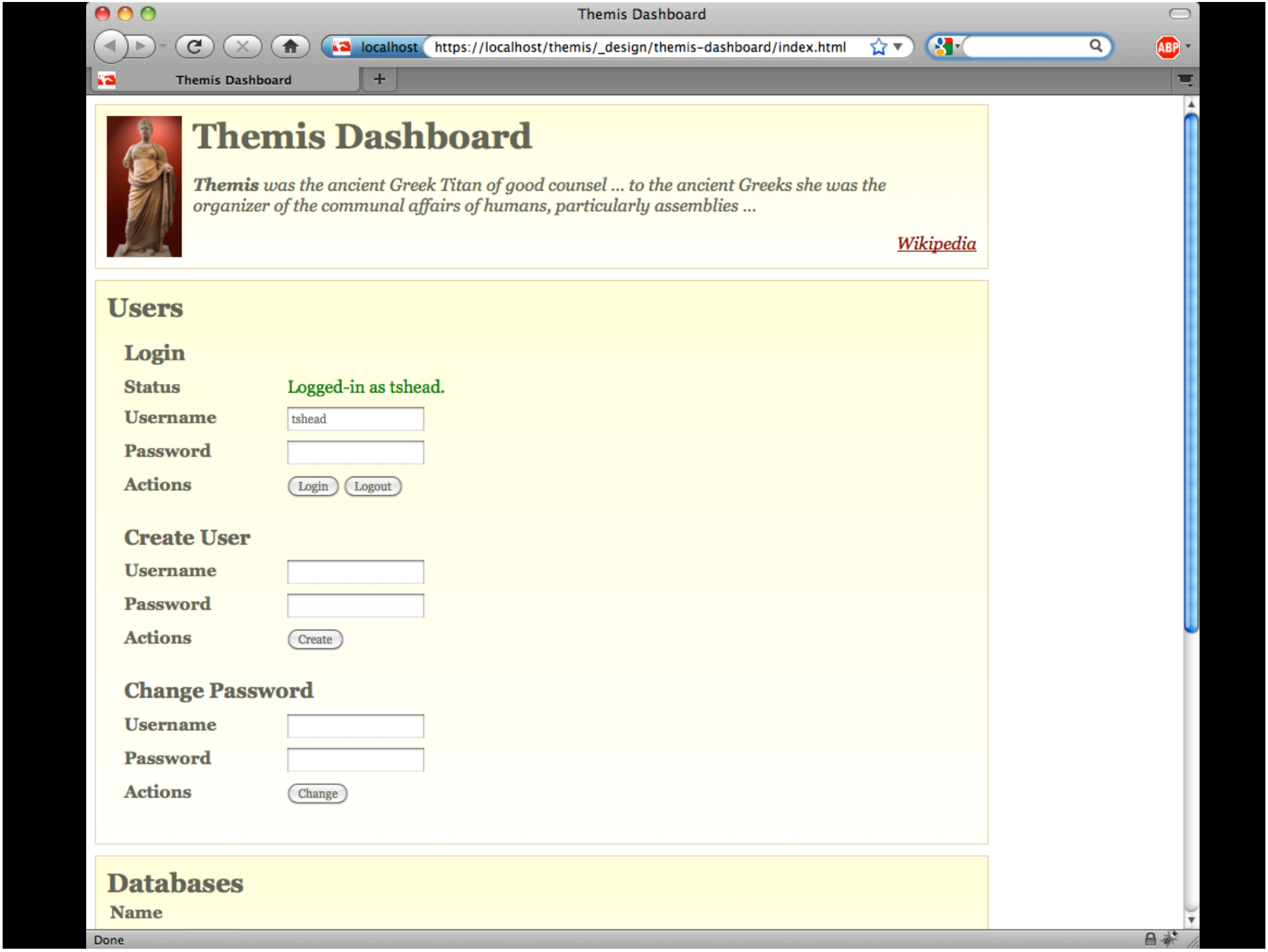


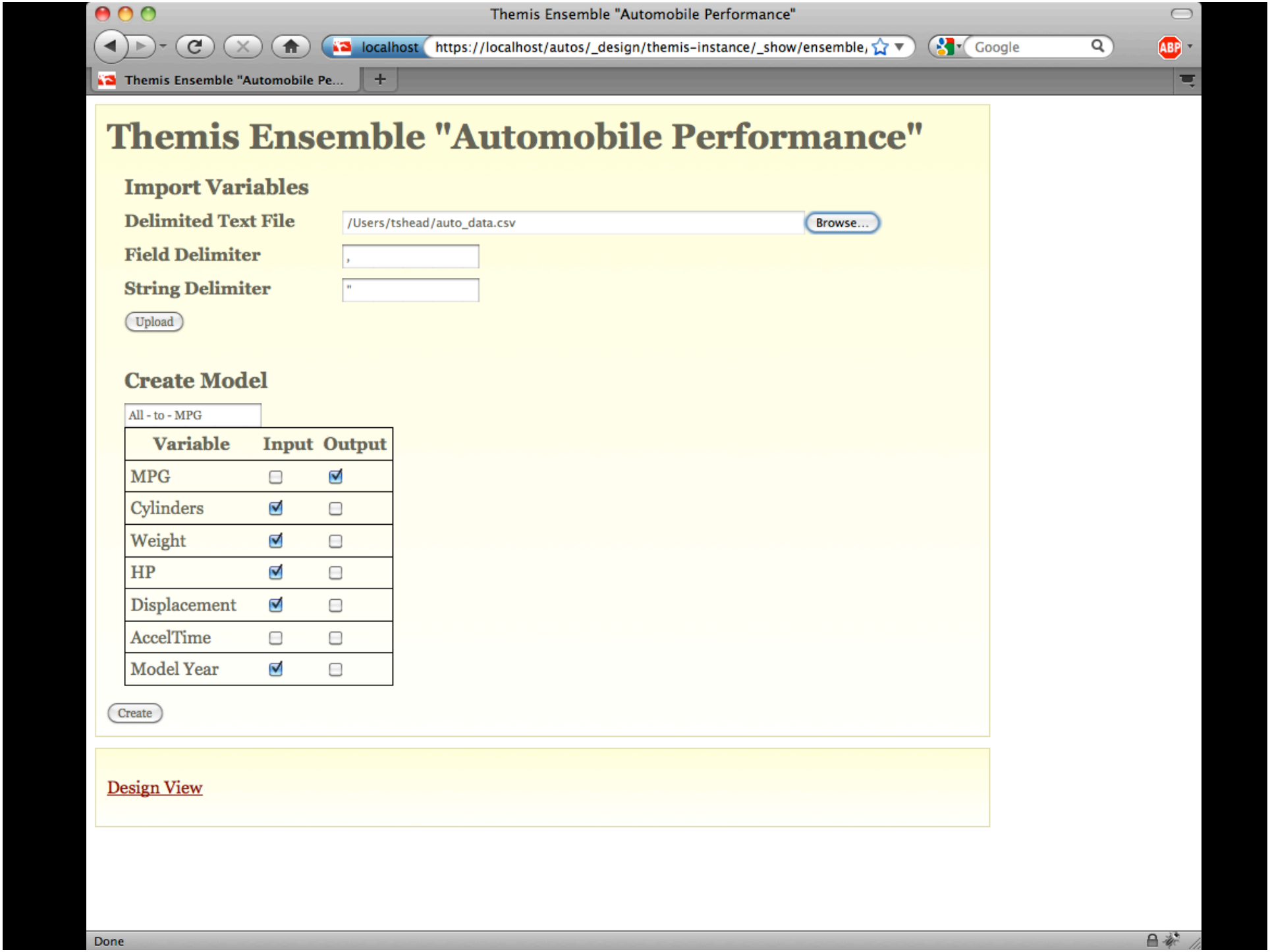
Community 0 Hostnames

politicalticker.blogs.cnn.com
attackerman.firedoglake.com
blog.reidreport.com
blogs.abcnews.com

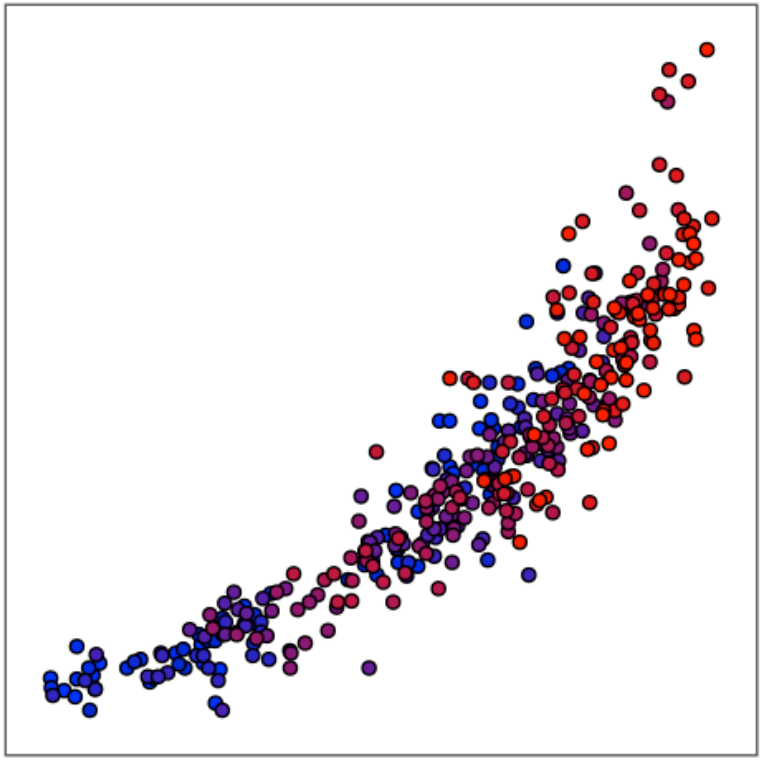
Topic Name: | Term List: | Add Topic

		http://firedoglake.com/2008/07/12/blue-america-welcomes-back-mark-begich-sen-ak
Name: kitten Terms: kittens		http://rawdawgb.blogspot.com/2008/07/i-dont-wanna-see-you-draws-jones.html http://www.samsedershow.com/node/3392 http://rawdawgb.blogspot.com/2008/07/lone-lakota-heading-east.html http://rawdawgb.blogspot.com/2008/07/top-10-reasons-4-working-for-self.html http://rawdawgb.blogspot.com/2008/07/pomegranate-puy.html http://rawdawgb.blogspot.com/2008/07/plum-foolish.html http://riverdaughter.wordpress.com/2008/07/13/pumabull-moose
Name: lipstick on a pig Terms: lipstick, on, a, pig		http://mrssatan.blogspot.com/2008/07/obama-worships-satan.html http://blogs.suntimes.com/sweet/2008/07/sarkozy_gushes_over_obama_kind.html http://politicalticker.blogs.cnn.com/2008/08/29/how-mccain-chose-palin http://asecondhandconjecture.com/index.php/2008/08/30/sarah-palin-kuwait-gallery http://newsbusters.org/blogs/warner-todd-huston/2008/08/31/stooping-lowest-level-dem http://thecaucus.blogs.nytimes.com/2008/08/31/mccain-donations-rolling-in-after-palin-p http://www.floppingaces.net/2008/09/01/sarah-palin-the-female-teddy-roosevelt
Name: Obama Terms: obama		http://attackerman.firedoglake.com/2008/07/02/hitchensboarding http://riverdaughter.wordpress.com/2008/07/02 http://www.time-blog.com/swampland/2008/07/christian_conservatives_unitin.html

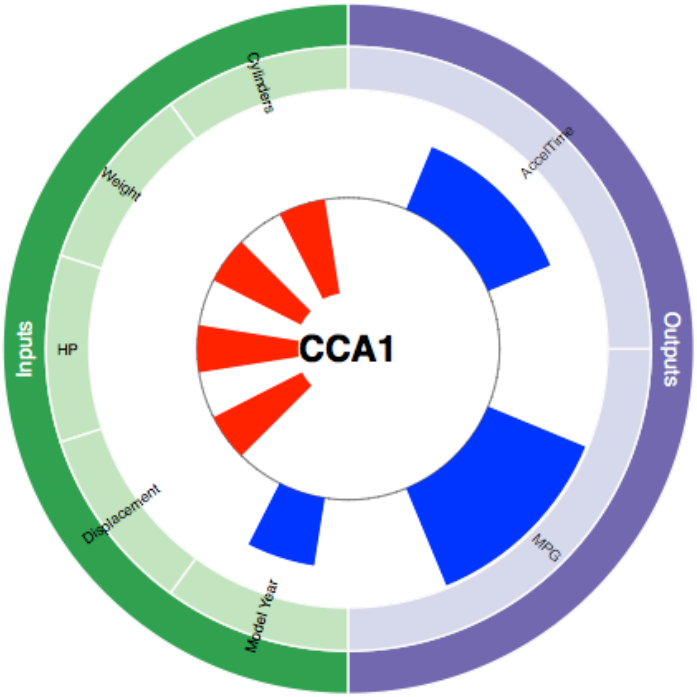

Themis Dashboard
localhost
https://localhost/themis/_design/themis-dashboard/index.html
Themis Dashboard
Themis Dashboard
Themis was the ancient Greek Titan of good counsel ... to the ancient Greeks she was the organizer of the communal affairs of humans, particularly assemblies ...
Wikipedia
Users
Login
Status
Logged-in as tshead.
Username
tshead
Password
Actions
Login
Logout
Create User
Username
Password
Actions
Create
Change Password
Username
Password
Actions
Change
Databases
Name
Done


Themis Ensemble "Automobile Performance"
localhost
https://localhost/autos/_design/themis-instance/_show/ensemble,
Google
Themis Ensemble "Automobile Pe...
Themis Ensemble "Automobile Performance"
Import Variables
Delimited Text File
/Users/tshead/auto_data.csv
Browse...
Field Delimiter
,
String Delimiter
"
Upload
Create Model
All - to - MPG
Variable
Input
Output
MPG
Cylinders
Weight
HP
Displacement
AccelTime
Model Year
Create
Design View
Done

Themis Model "All - to - all"



Component	R^2	P
CCA1	0.910	0.089
CCA2	0.698	0.512



Simulation ID	Cylinders	Weight	HP	Displacement	Model Year	MPG	AccelTime
0	8.00	3.50e+3	130	307	70.0	18.0	12.0
1	8.00	3.69e+3	165	350	70.0	15.0	11.5
2	8.00	3.44e+3	150	318	70.0	18.0	11.0
3	8.00	3.43e+3	150	304	70.0	16.0	12.0

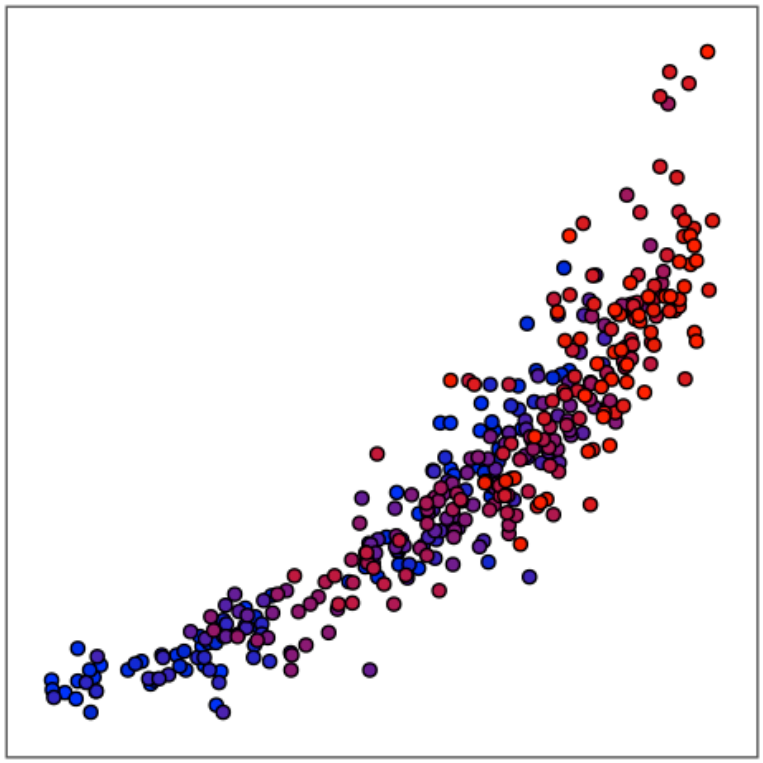
Themis Model "All – to – all"

localhost https://localhost/autos/_design/themis-instance/_show/model/feC

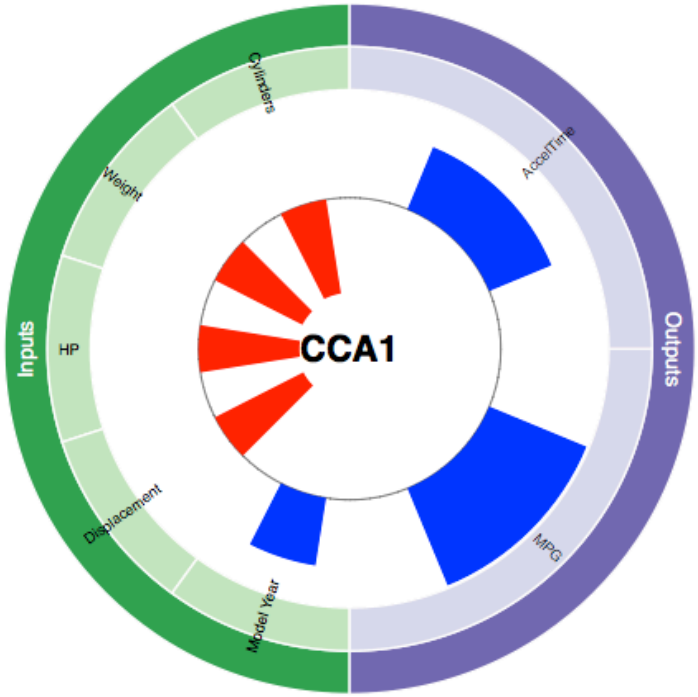
Google

Themis Model "All – to – all"

Themis Model "All - to - all"



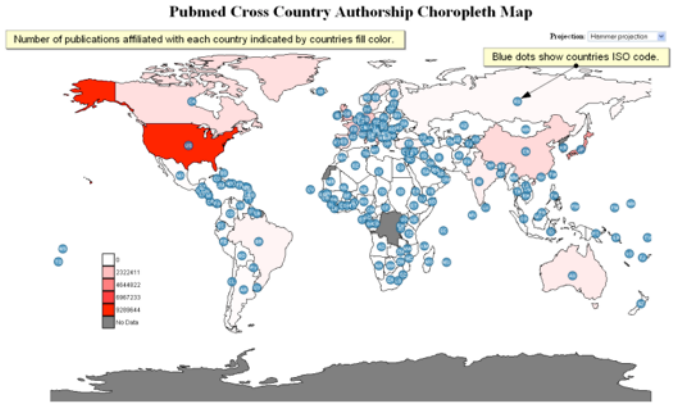
Component	R^2	P
CCA1	0.910	0.089
CCA2	0.698	0.512

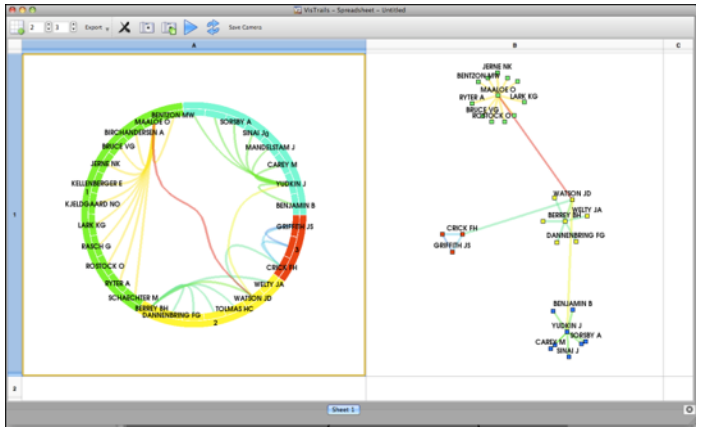


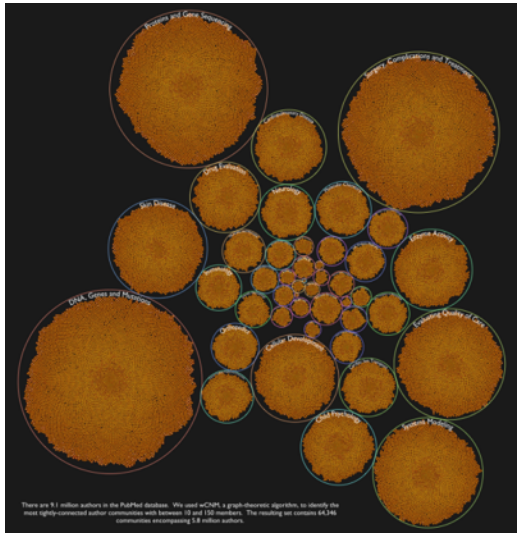
Simulation ID	Cylinders	Weight	HP	Displacement	Model Year	MPG	AccelTime
0	8.00	3.50e+3	130	307	70.0	18.0	12.0
1	8.00	3.69e+3	165	350	70.0	15.0	11.5
2	8.00	3.44e+3	150	318	70.0	18.0	11.0
3	8.00	3.43e+3	150	304	70.0	16.0	12.0

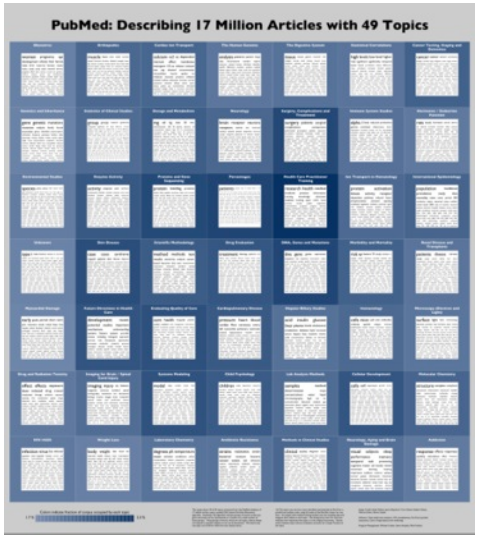
Done

We have successfully clustered and visualized 17 million records.



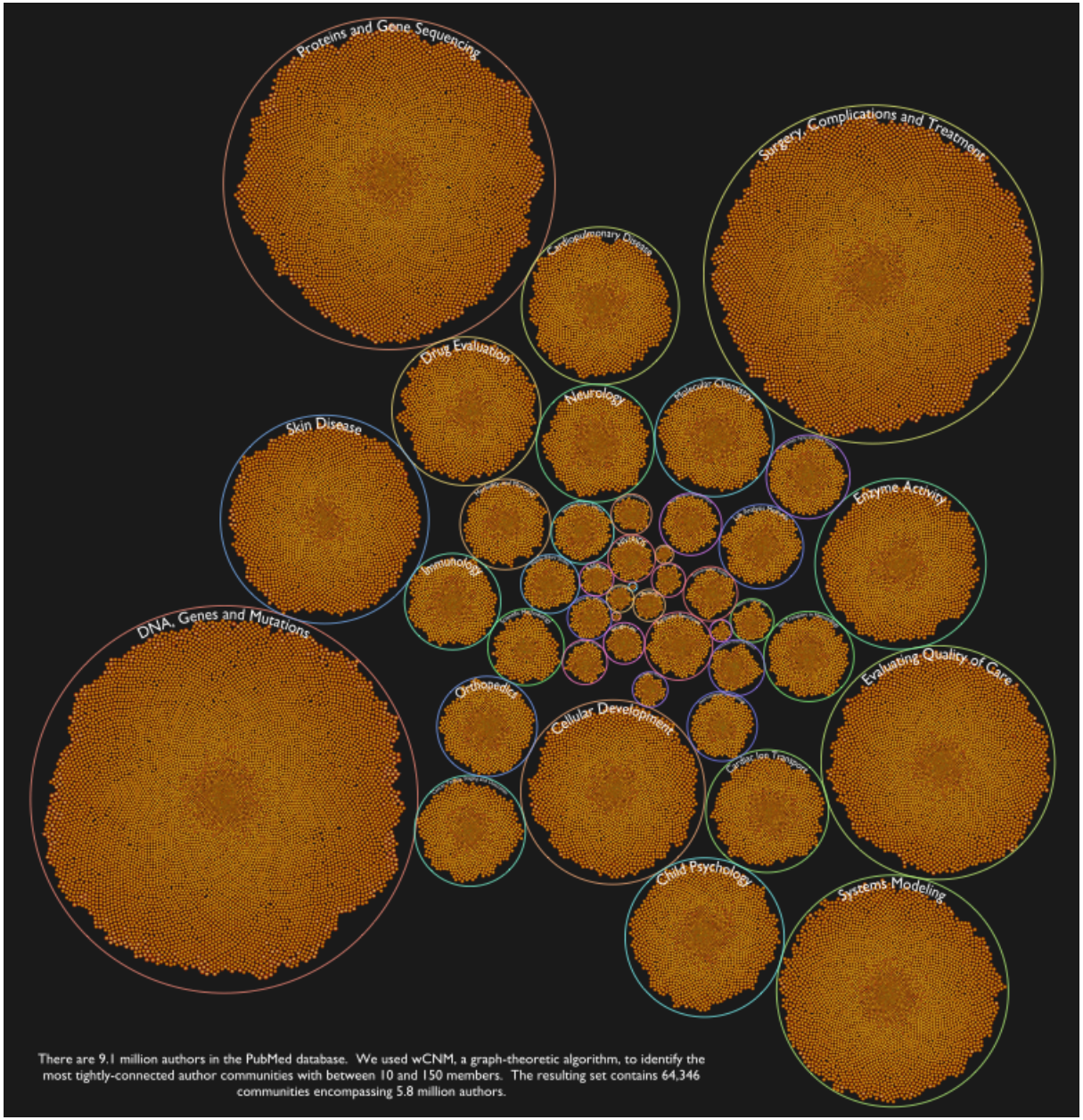






Sandia National Laboratories

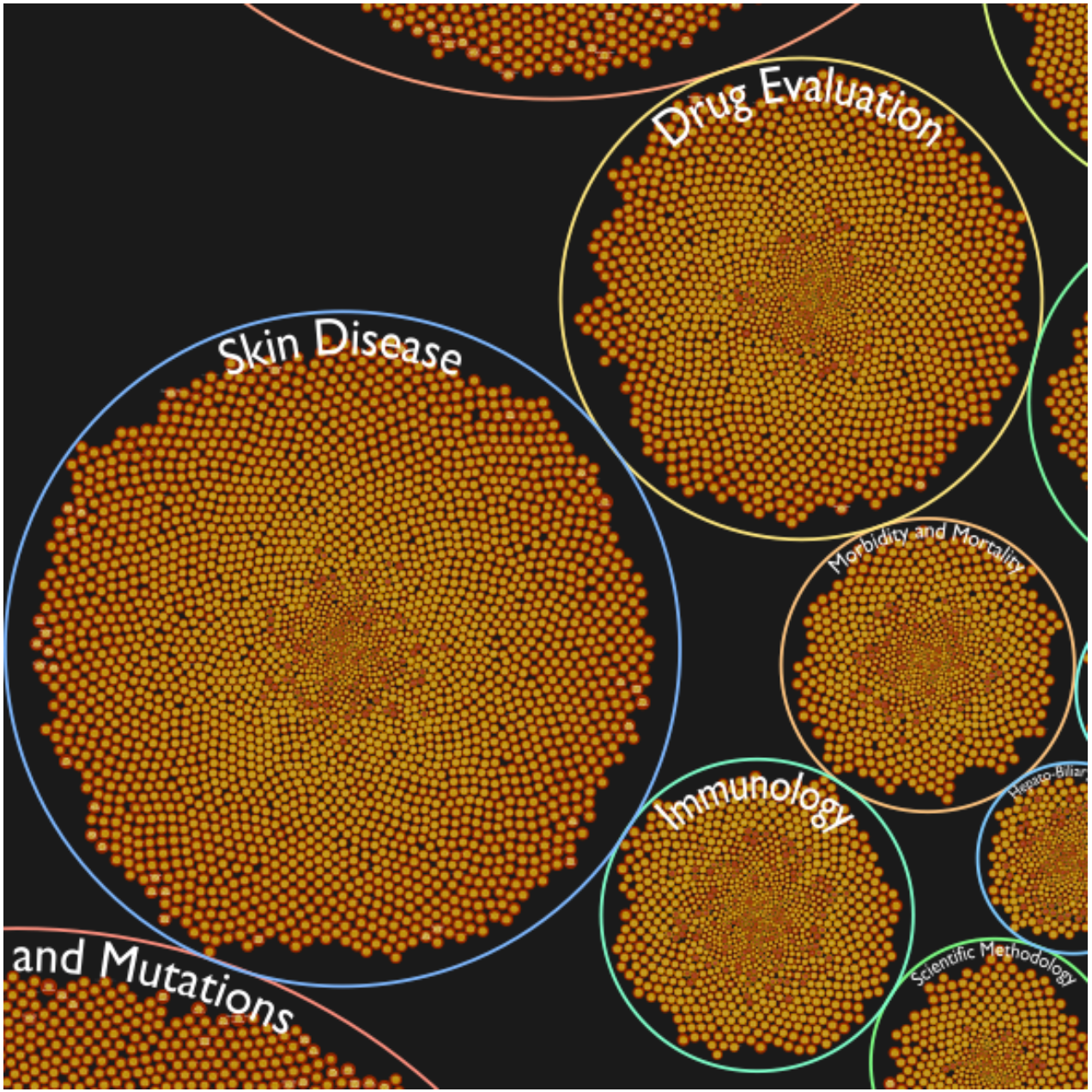
Author Communities Visualized



- Each blob is a topic area
- Each tiny orange dot is a single community
- 44 topics, 64,314 communities
  - 10-150 authors in each community

Sandia National Laboratories

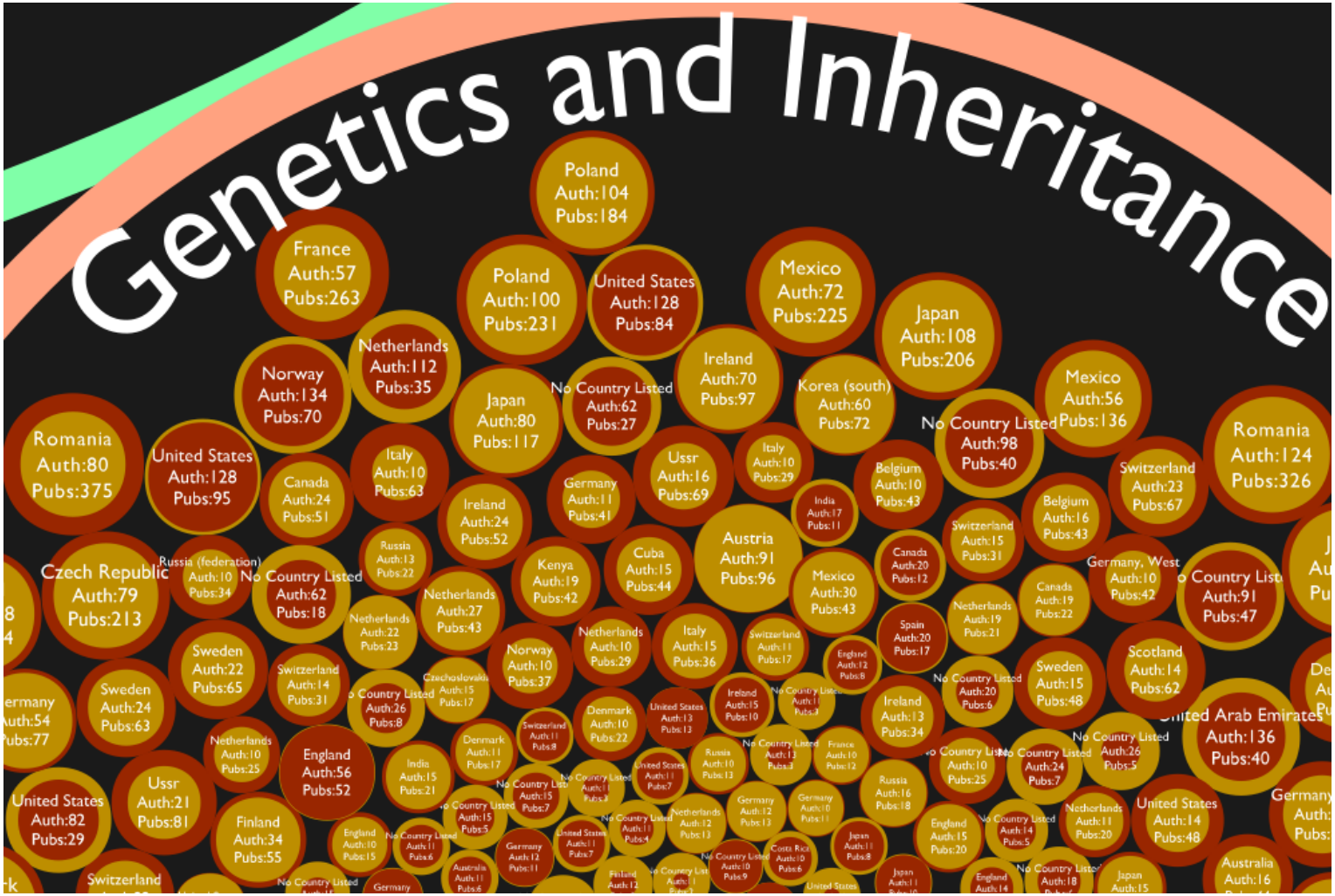
Author Communities: Zoom 1



- Topic areas contain between 5 and 9800 communities
- Each community is sized according to its author/article count

Sandia National Laboratories

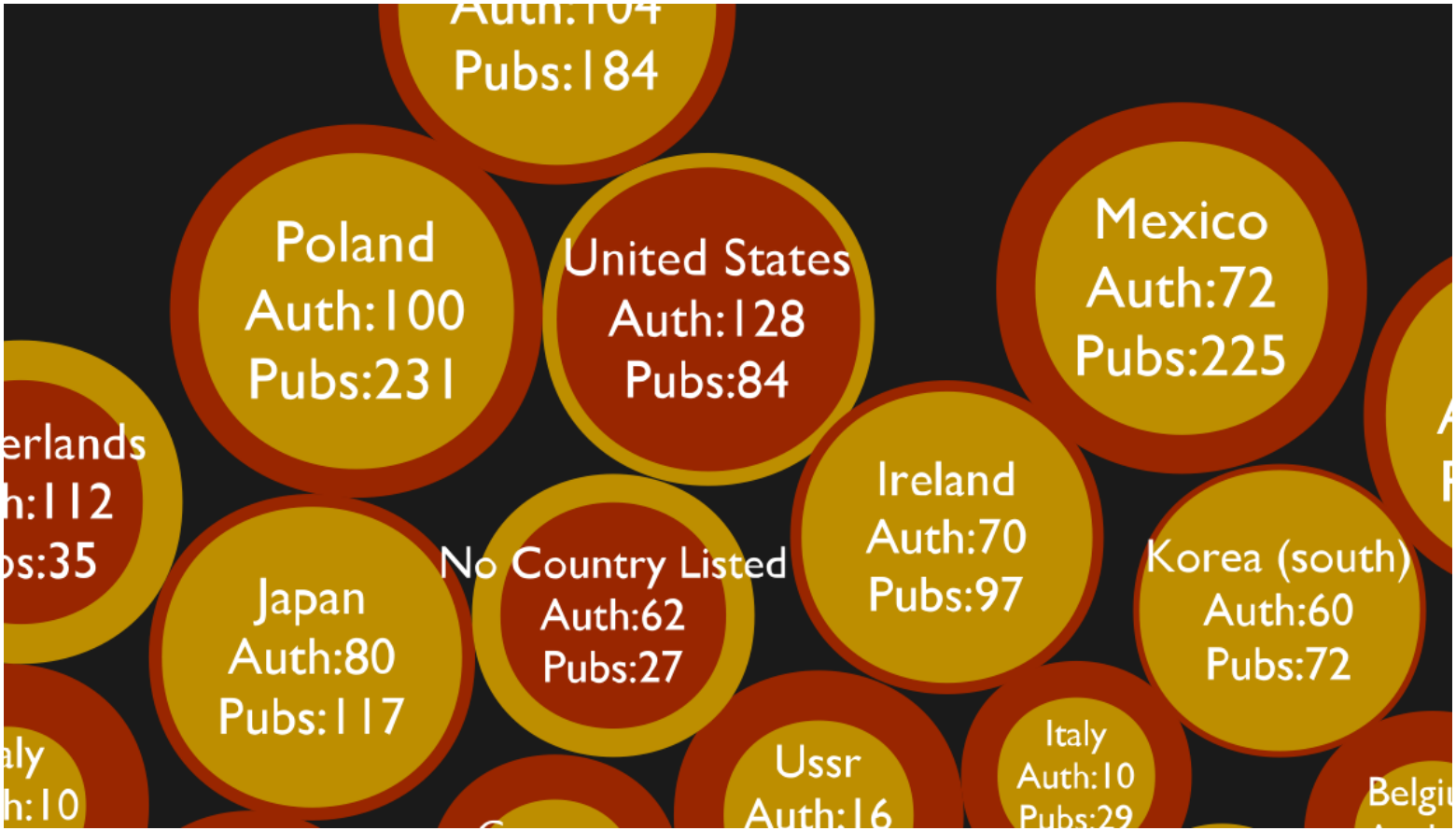
Author Communities: Zoom 2



- Finer-grained topics (and diagrams) are possible
- Hierarchical layout possible with advanced clustering algorithms

Sandia National Laboratories

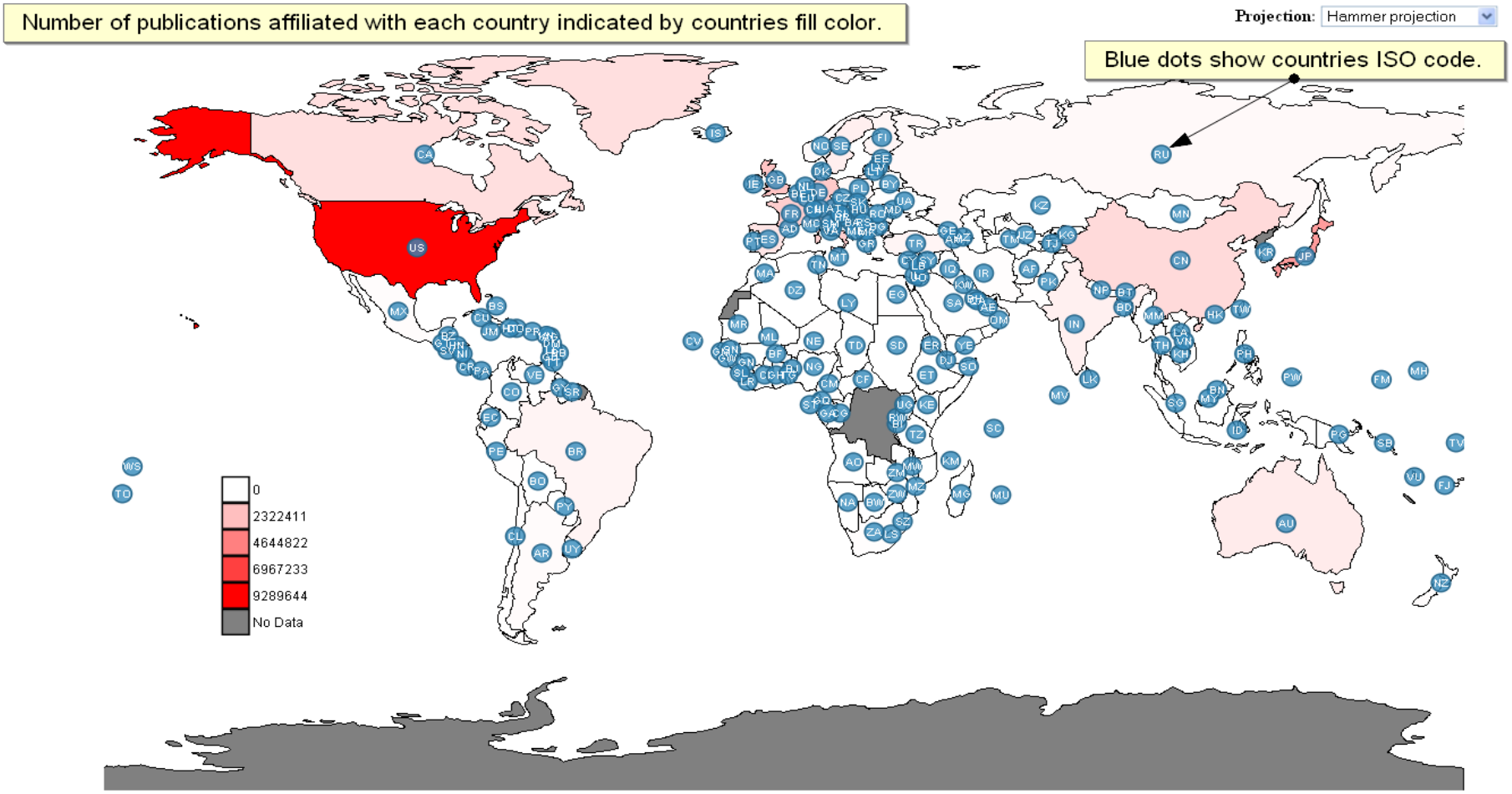
Author Communities: Zoom 3



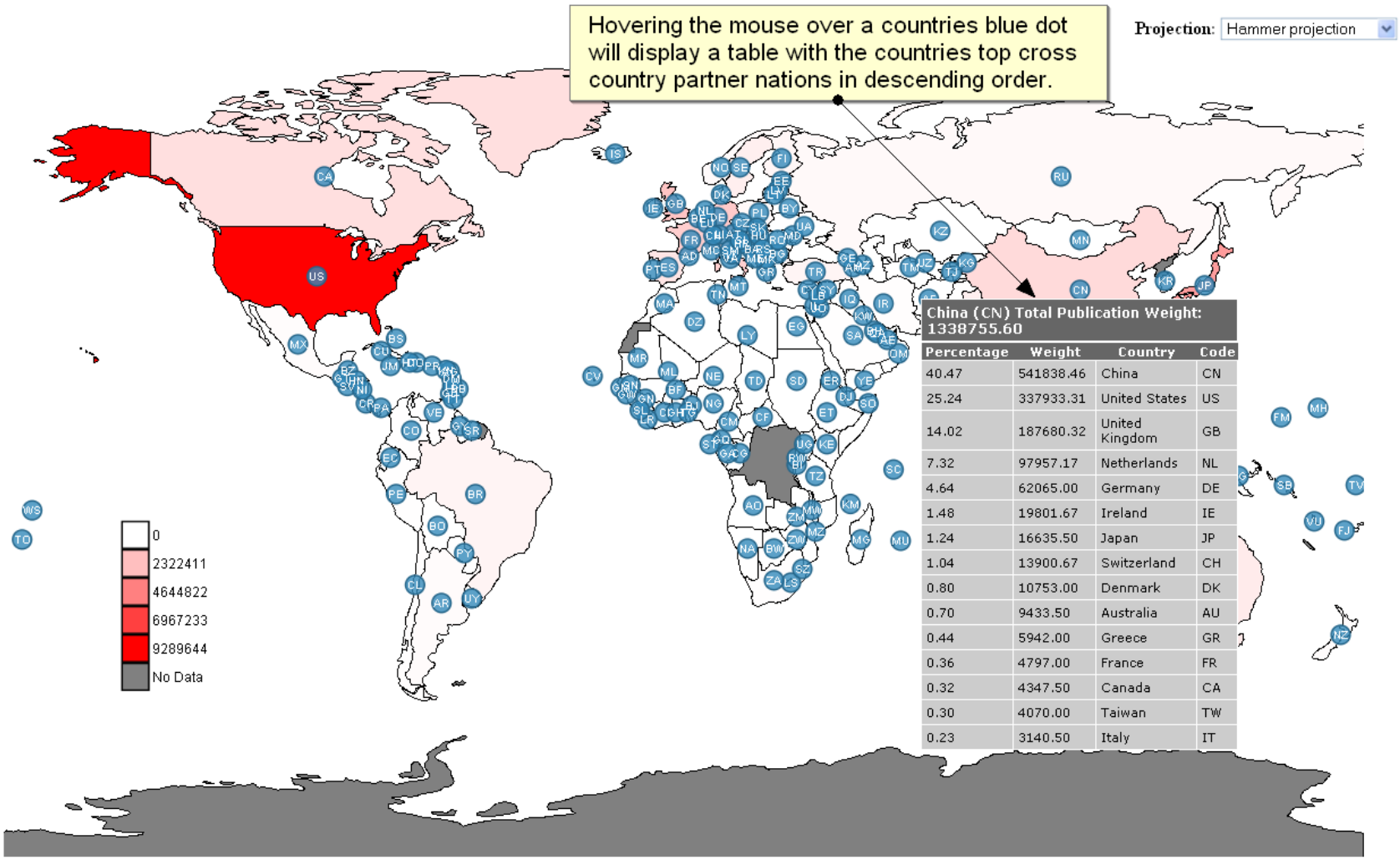
- Community glyphs indicate relative proportion of authors to articles
- Red = articles, yellow = authors
- Red outside yellow means more articles than authors (and vice versa)
- Labels show country where most papers are published

Sandia National Laboratories

Pubmed Cross Country Authorship Choropleth Map



Pubmed Cross Country Authorship Choropleth Map

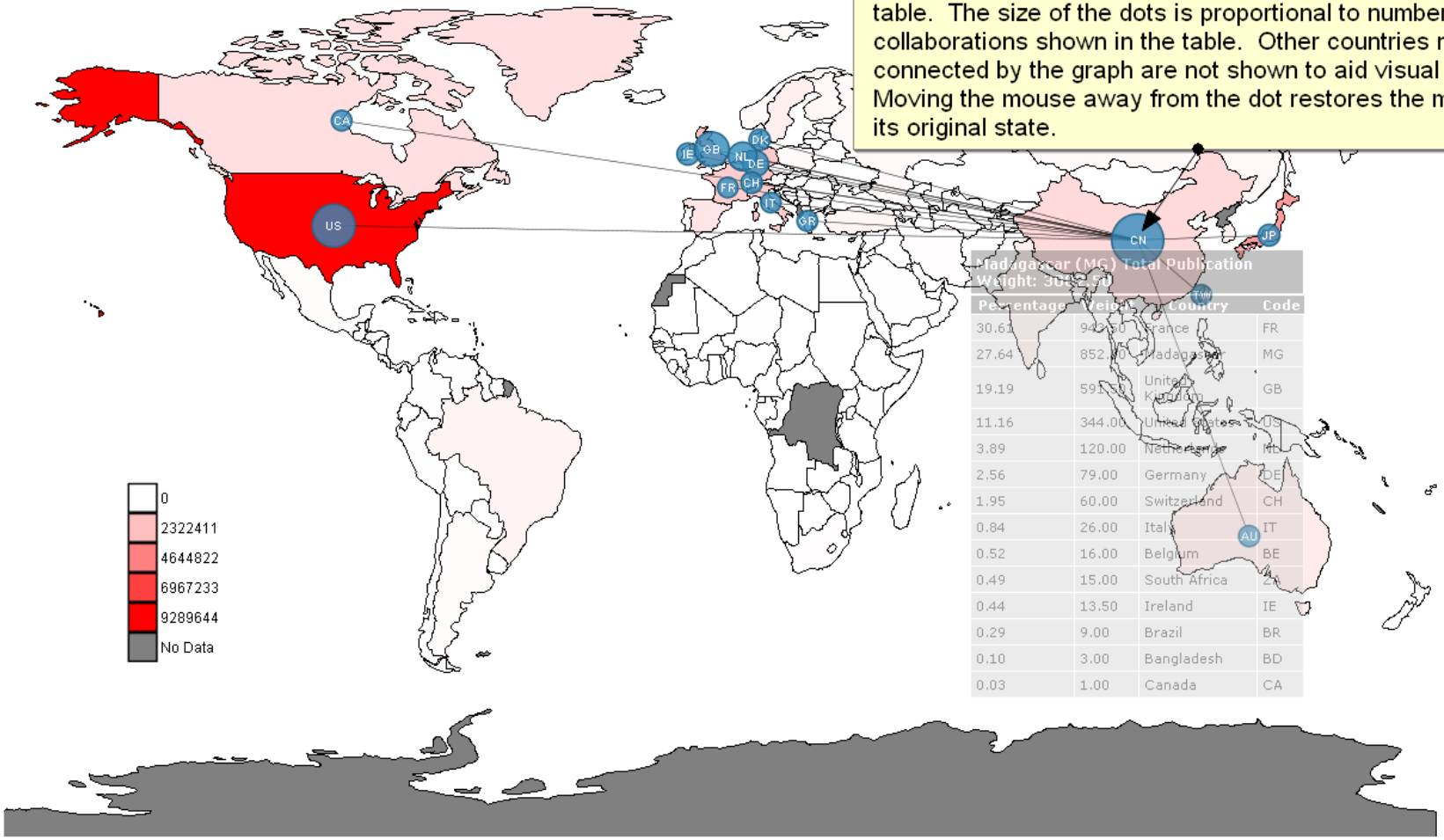


China (CN) Total Publication Weight: 1338755.60

Percentage	Weight	Country	Code
40.47	541838.46	China	CN
25.24	337933.31	United States	US
14.02	187680.32	United Kingdom	GB
7.32	97957.17	Netherlands	NL
4.64	62065.00	Germany	DE
1.48	19801.67	Ireland	IE
1.24	16635.50	Japan	JP
1.04	13900.67	Switzerland	CH
0.80	10753.00	Denmark	DK
0.70	9433.50	Australia	AU
0.44	5942.00	Greece	GR
0.36	4797.00	France	FR
0.32	4347.50	Canada	CA
0.30	4070.00	Taiwan	TW
0.23	3140.50	Italy	IT

Pubmed Cross Country Authorship Choropleth Map

Clicking the mouse on a countries blue dot will produce a graph from the information contained in the displayed table. The size of the dots is proportional to number of collaborations shown in the table. Other countries not connected by the graph are not shown to aid visual clarity. Moving the mouse away from the dot restores the map to its original state.



Madagascar (MG) Total Publication

Percentage	Weight	Country	Code
30.61	944.00	France	FR
27.64	852.00	Madagascar	MG
19.19	591.00	United Kingdom	GB
11.16	344.00	United States	US
3.89	120.00	Netherlands	NL
2.56	79.00	Germany	DE
1.95	60.00	Switzerland	CH
0.84	26.00	Italy	IT
0.52	16.00	Belgium	BE
0.49	15.00	South Africa	ZA
0.44	13.50	Ireland	IE
0.29	9.00	Brazil	BR
0.10	3.00	Bangladesh	BD
0.03	1.00	Canada	CA

All (Large) Data Analysis is Highly, Immediately Constrained

- Raw data must be processed
  - Example: database schema
- Analysis Algorithms op on specific data types
  - Graphs, tensors, images
- Advanced architectures support certain ops
  - XMT: multithreaded (graphs)
  - Netezza: hardware execution of SQL
  - Distributed memory: large tensor operations
- Analysis results viewed, queried many ways
  - Layout/vis introduces bias
- Human in the loop is a known unknown
  - *Especially in non-scientific arenas*

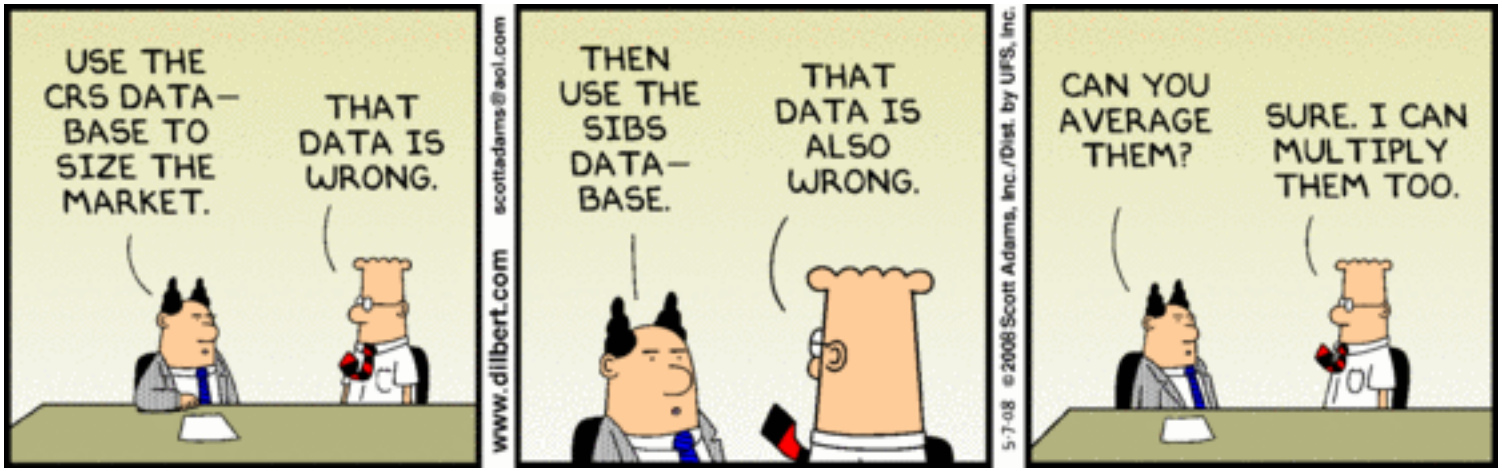
Analysis Path Forward

- Tight coupling among architectures, algorithms and applications
  - Customers must understand algorithms
- *Everything must be interactive*
  - But we must make task-specific tradeoffs
- Good software foundations are crucial to collaboration in the field
  - http://titan.sandia.gov
- It takes interactive research to find things that are valuable enough to throw away
  - Finding viable products is worse
- Good news: this is so difficult and important that visualization, UI, and design are recognized as core to success

Sandia National Laboratories

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

Sandia National Laboratories

Questions?

Sandia National Laboratories