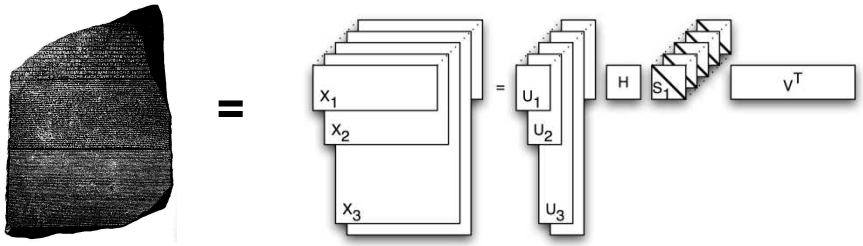
SAND2007-6648P

High-Level Visualization of Multilingual Text from the Internet: Using PARAFAC2 for Scalable Multilingual Document Clustering



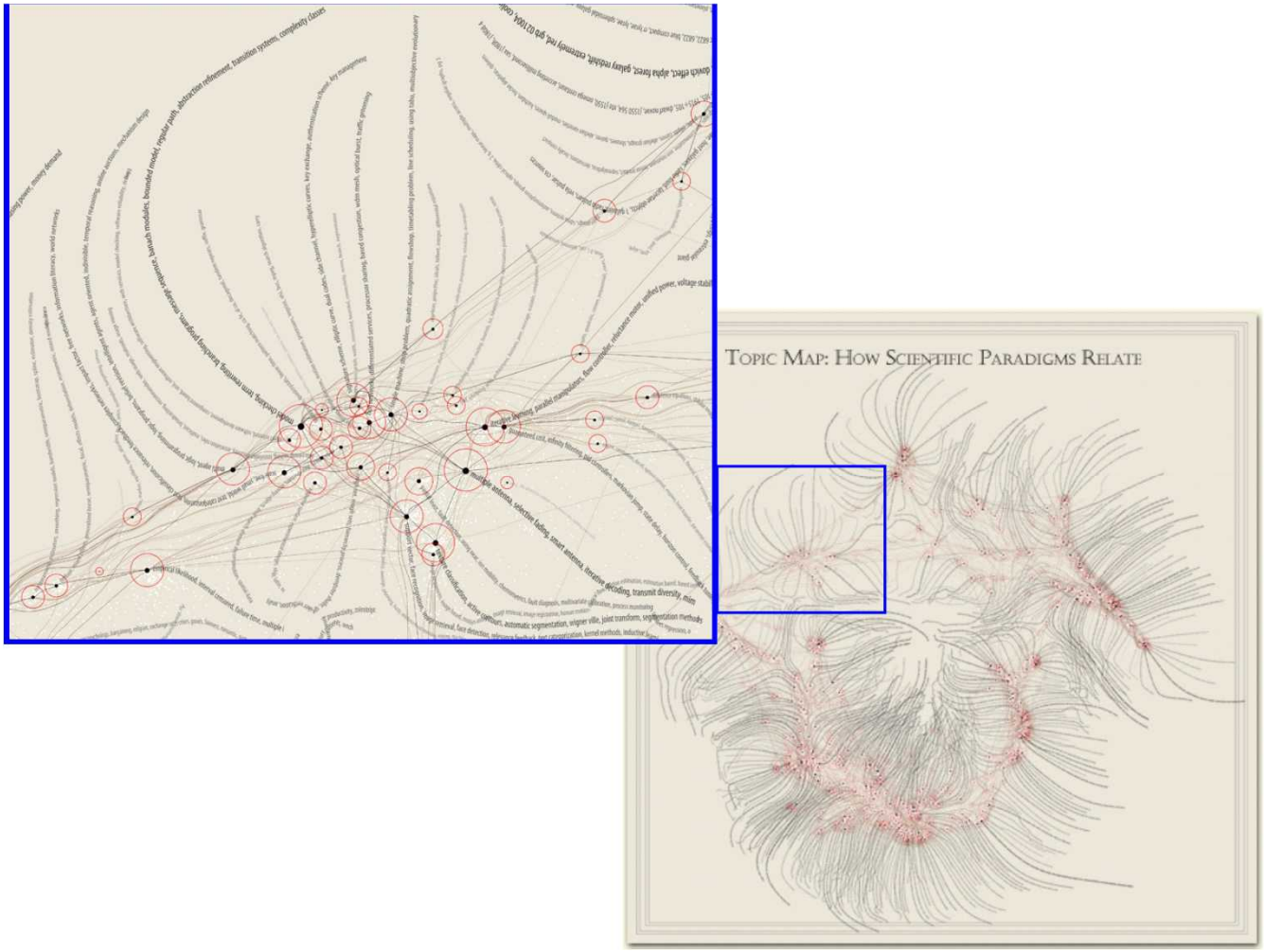
August 22, 2007

Peter Chew
Sandia National Laboratories

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract DE-AC04-94AL85000.

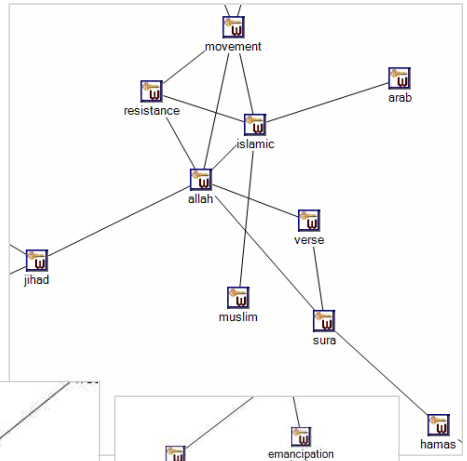
Sandia National Laboratories

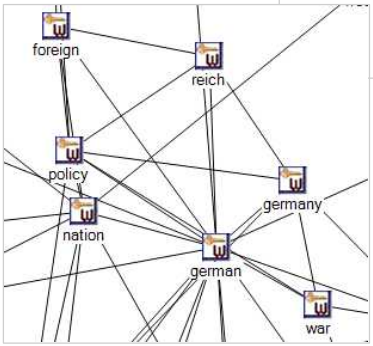
‘Identification of Threats’: the vision

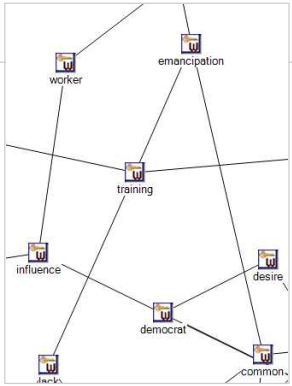


Sandia National Laboratories

Uses







words

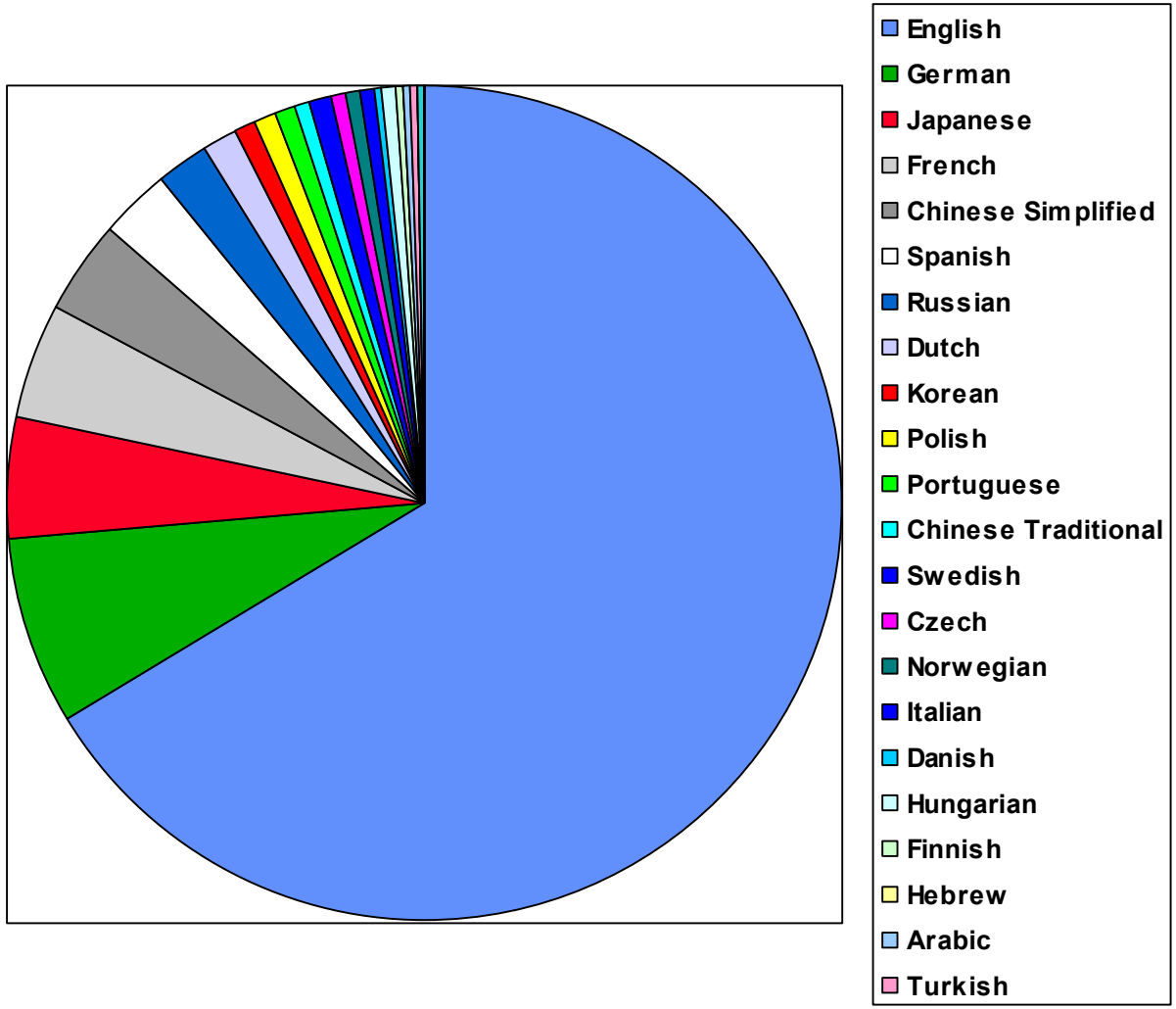
ideas

movements

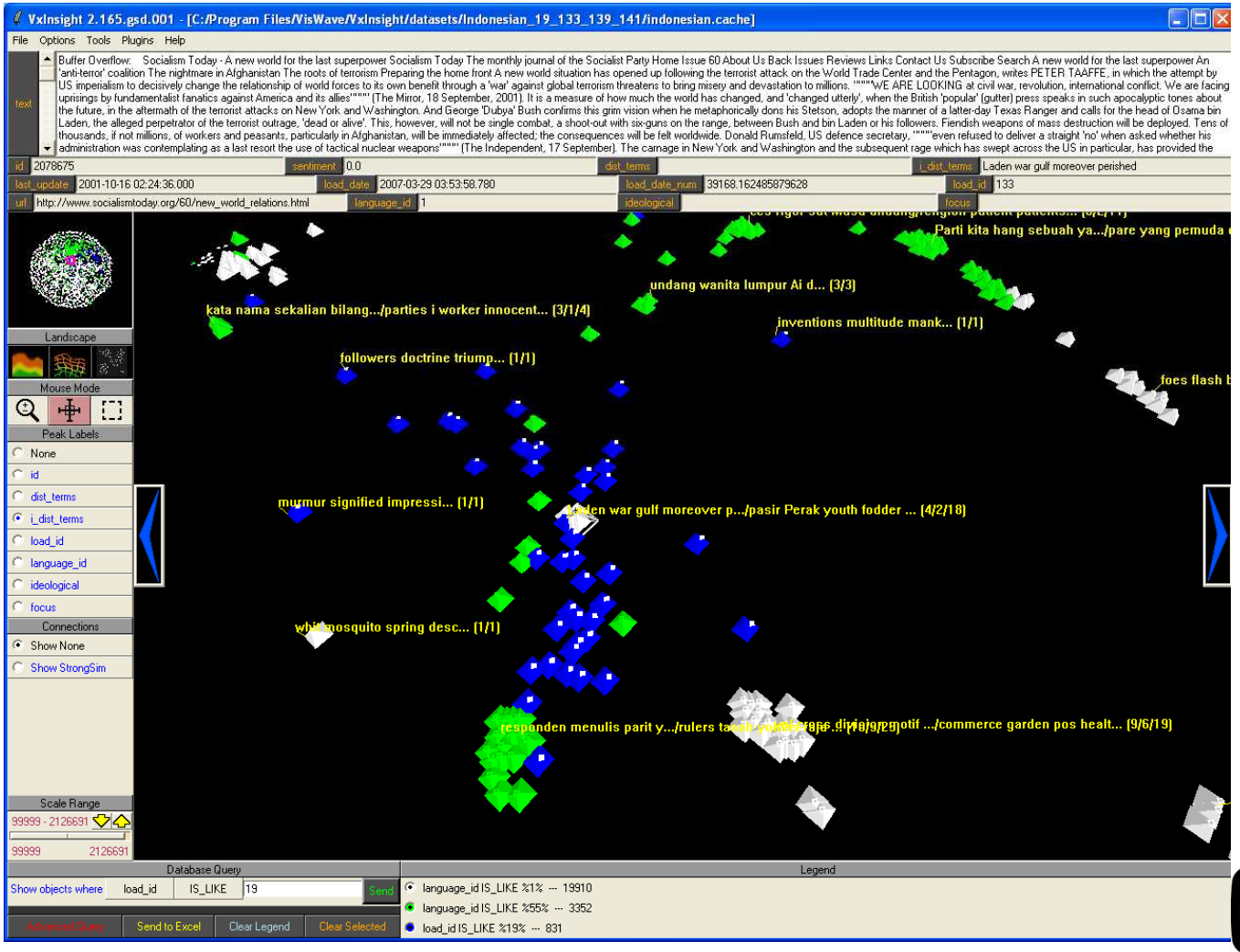
- Early detection of significant socio-political shifts
- Tracking changes over time in public interest in certain topics

Sandia National Laboratories

Languages of the internet

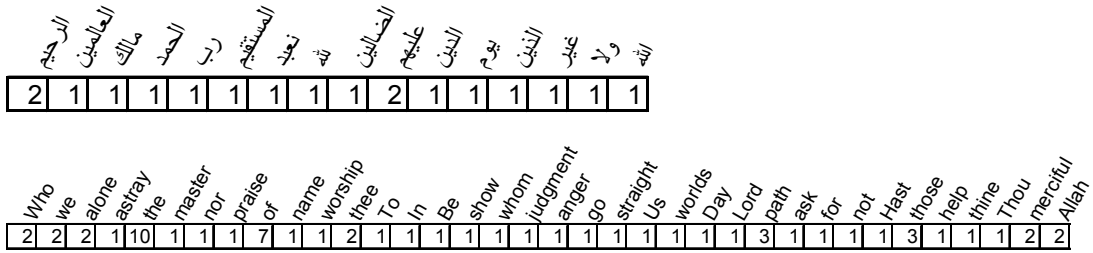


Multilingual document clustering



Sandia National Laboratories

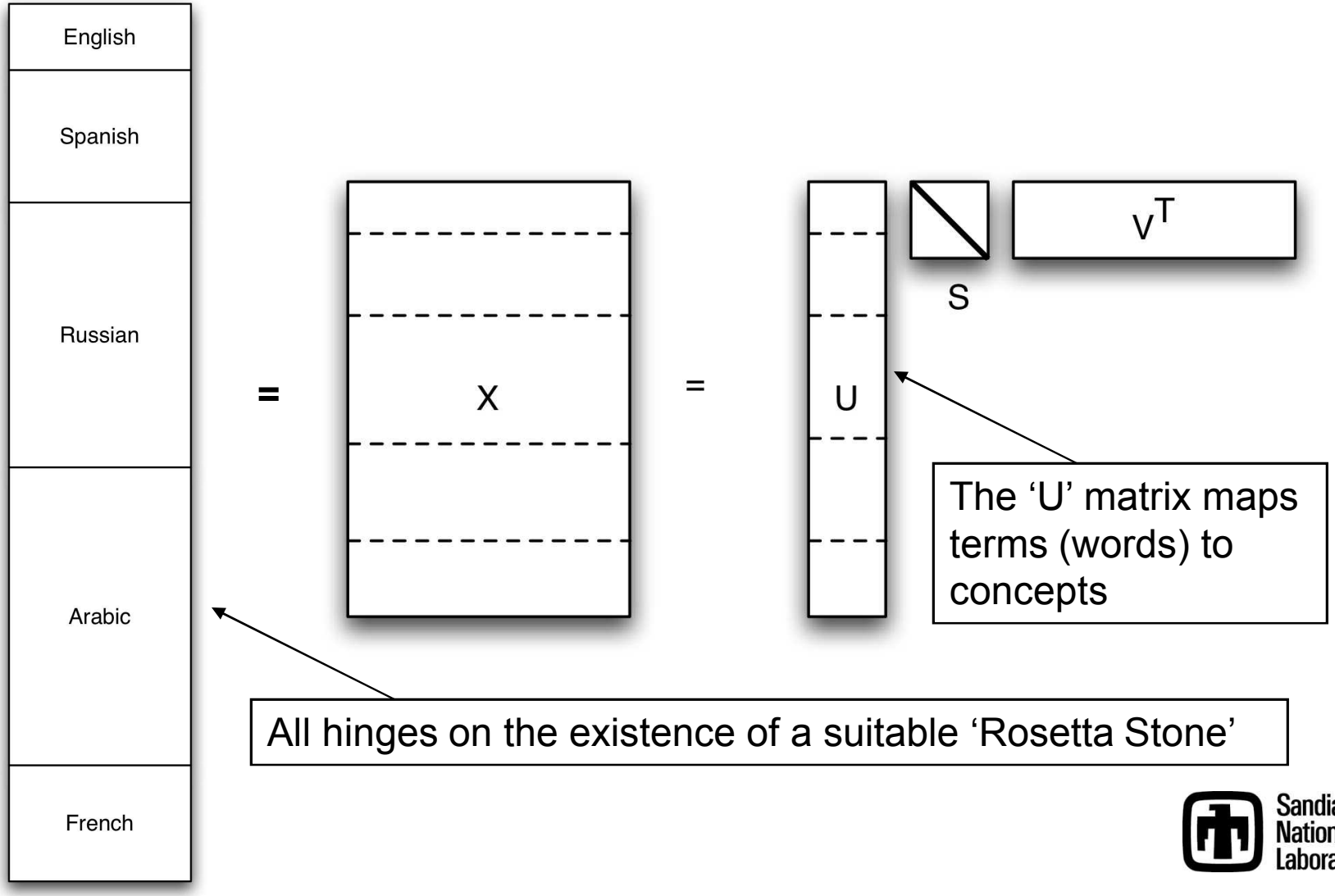
The vector space model



- documents can be represented by vectors
- cosine between vectors is a measure of document similarity

Sandia National Laboratories

The standard multilingual LSA model



Sandia National Laboratories

The Bible as our Rosetta Stone

- Resnik, Olsen & Diab (1999) discuss the properties which make the Bible an ideal parallel corpus in many ways
  - Translations in > 2,400 languages and rising
  - Great care taken over translations
  - Respectably large compared to other corpora
  - Covers many modern genres
  - Mostly public-domain
  - Electronically available (we have 54 languages)
  - Alignable (31,226 mini-documents)
  - Covers up to 85% of modern vocabulary

Sandia National Laboratories

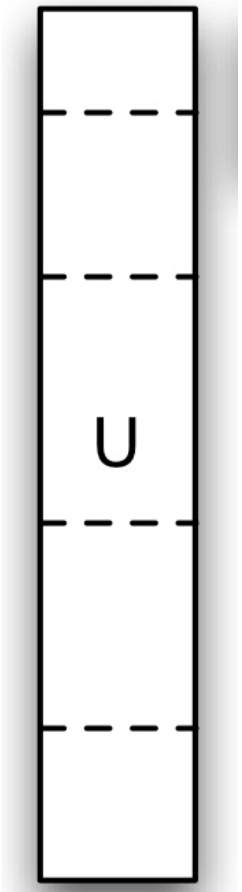
Languages in our implementation

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)	
Basque	Hebrew (Modern)	Gaelic
Breton		Spanish
Chamorro		Swahili
	Italian	Swedish
(Traditional)	Japanese	Tagalog
Croatian	Korean	Thai
Czech	Latin	Turkish
Danish	Latvian	Ukrainian
Dutch	Lithuanian	Vietnamese
English	Manx Gaelic	Wolof
Esperanto	Maori	Xhosa

About 99.76% coverage of documents on the internet

Sandia National Laboratories

The vector space model in LSA (1)



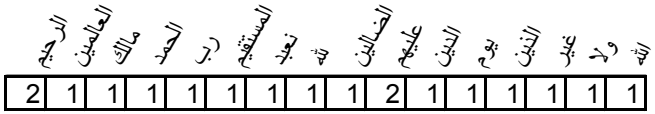
new document
(with word counts)

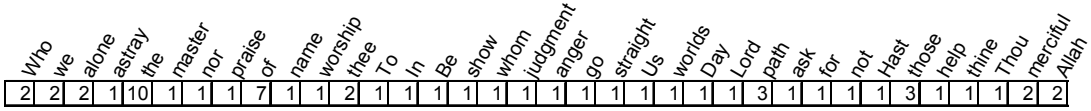
الله	ولا	غير	الذين	يوم	الدين	عليهم	الضالين	لا	نعبد	المستقيم	ربي	الحمد	مالك	العالمين	الرحيم
1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	2

=

dimension 1	0.0138
dimension 2	0.0106
dimension 3	0.0034
dimension 4	0.0044
dimension 5	-0.0009
dimension 6	0.0041
dimension 7	0.0102
dimension 8	0.0002
dimension 9	0.0052
dimension 10	0.0083
dimension 11	-0.0010
dimension 12	-0.0116
dimension 13	-0.0100
dimension 14	0.0023
dimension 15	-0.0052
dimension 16	0.0110
dimension 17	0.0030
dimension 18	0.0050
dimension 19	0.0056
dimension 20	0.0161

The vector space model in LSA (2)





dimension 1	0.1375
dimension 2	0.1052
dimension 3	0.0341
dimension 4	0.0441
dimension 5	-0.0087
dimension 6	0.0410
dimension 7	0.1011
dimension 8	0.0020
dimension 9	0.0518
dimension 10	0.0822
dimension 11	-0.0101
dimension 12	-0.1154
dimension 13	-0.0990
dimension 14	0.0228
dimension 15	-0.0520
dimension 16	0.1096
dimension 17	0.0294
dimension 18	0.0495
dimension 19	0.0553
dimension 20	0.1598

dimension 1	0.1947
dimension 2	0.1819
dimension 3	0.0202
dimension 4	0.0832
dimension 5	0.1250
dimension 6	-0.1281
dimension 7	0.0553
dimension 8	0.0747
dimension 9	-0.0004
dimension 10	-0.1386
dimension 11	0.0233
dimension 12	-0.4058
dimension 13	0.0796
dimension 14	-0.0573
dimension 15	0.1223
dimension 16	0.0507
dimension 17	-0.0201
dimension 18	0.0562
dimension 19	-0.0613
dimension 20	0.0970

With LSA, the cosine now allows us to compare documents across languages

Sandia National Laboratories

Validation of cross-language clustering

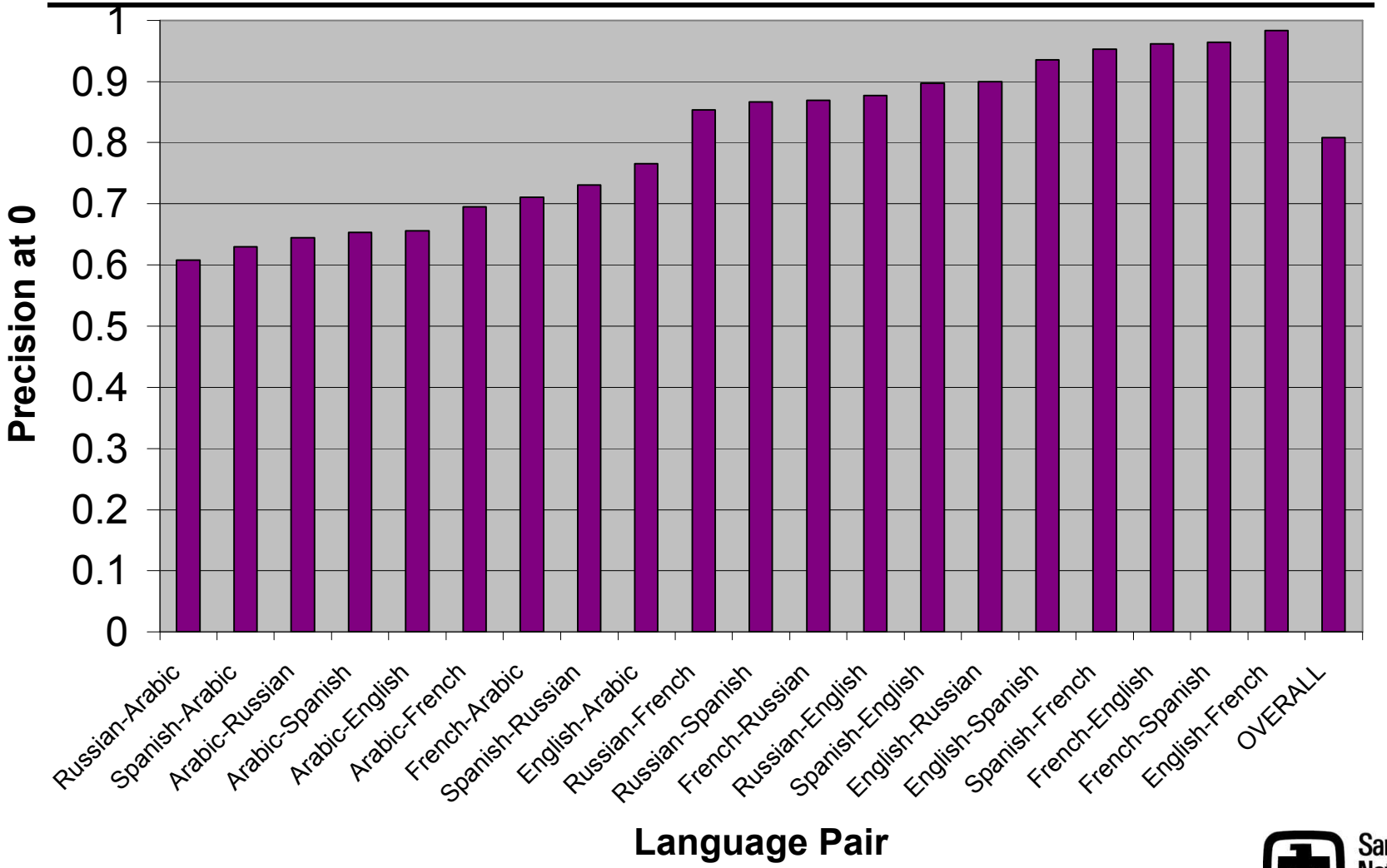
- **Test corpus: the Quran**
  - 114 chapters (suras), each in 5 languages

			Language of retrieved results																			
			English				Spanish				Russian				Arabic				French			
			1	2	…	114	1	2	…	114	1	2	…	114	1	2	…	114	1	2	…	114
Language of query	English	1	✓				✓				✓				✓				✓			
		2		✓				✓						✓		✓				✓		
		…			✓				✓				✓				✓				✓	
		114				✓				✓				✓				✓				✓
	Spanish	1	✓				✓				✓				✓				✓			
		2		✓				✓				✓				✓				✓		
		…			✓				✓				✓				✓				✓	
		114				✓				✓				✓				✓				✓
	Russian	1	✓				✓				✓				✓				✓			
		2		✓				✓				✓				✓				✓		
		…			✓				✓				✓				✓				✓	
		114				✓				✓				✓				✓				✓
	Arabic	1	✓				✓				✓				✓				✓			
		2		✓				✓				✓				✓				✓		
		…			✓				✓				✓				✓				✓	
		114				✓				✓				✓				✓				✓
	French	1	✓				✓				✓				✓				✓			
		2		✓				✓				✓				✓				✓		
		…			✓				✓				✓				✓				✓	
		114				✓				✓				✓				✓				✓

- ‘Language-specific’ vs. ‘multilingual’ precision

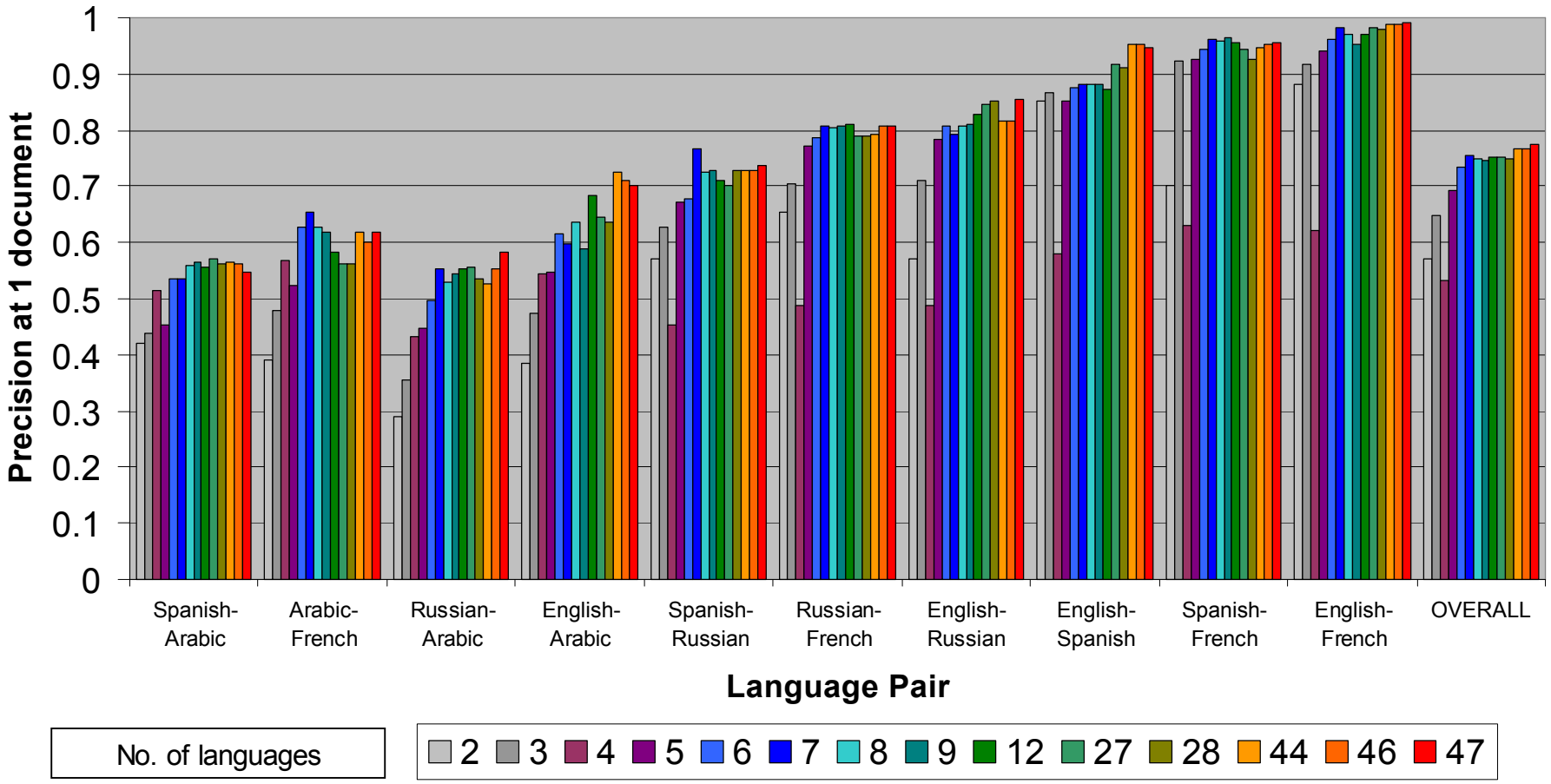
Sandia National Laboratories

Results with LSA: language-specific



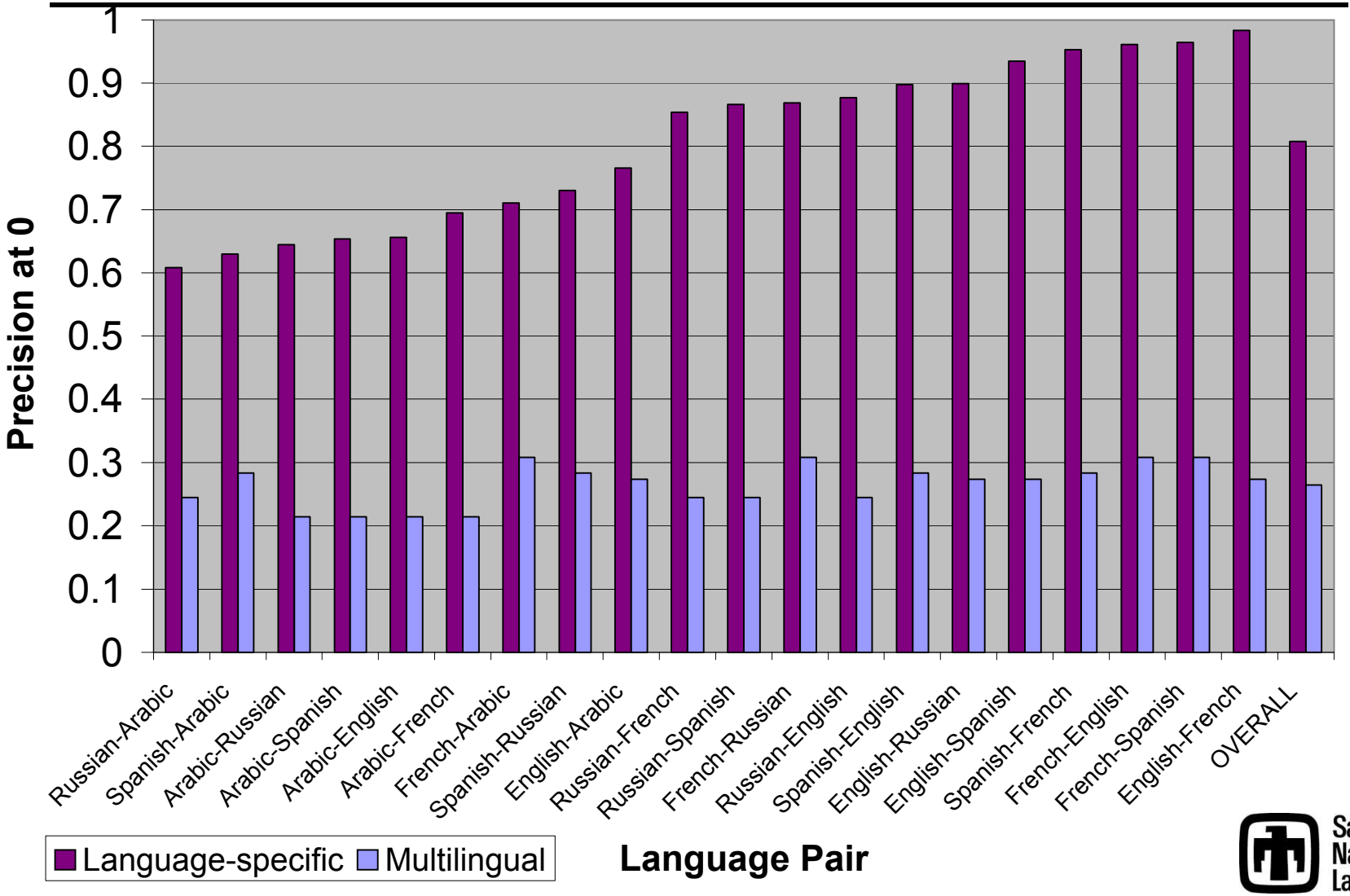
Sandia National Laboratories

‘Massive linguistic parallelism’ helps



Sandia National Laboratories

Results with LSA: multilingual



Sandia National Laboratories

Why multilingual precision can be lower

Ranking	Language of Retrieved Document	Relevant?
1	English	✓
2	English	✘
3	English	✘
4	English	✘
5	English	✘
6	French	✓
7	Spanish	✓
8	Arabic	✓
9	Russian	✓

With LSA, documents do not cluster language-independently

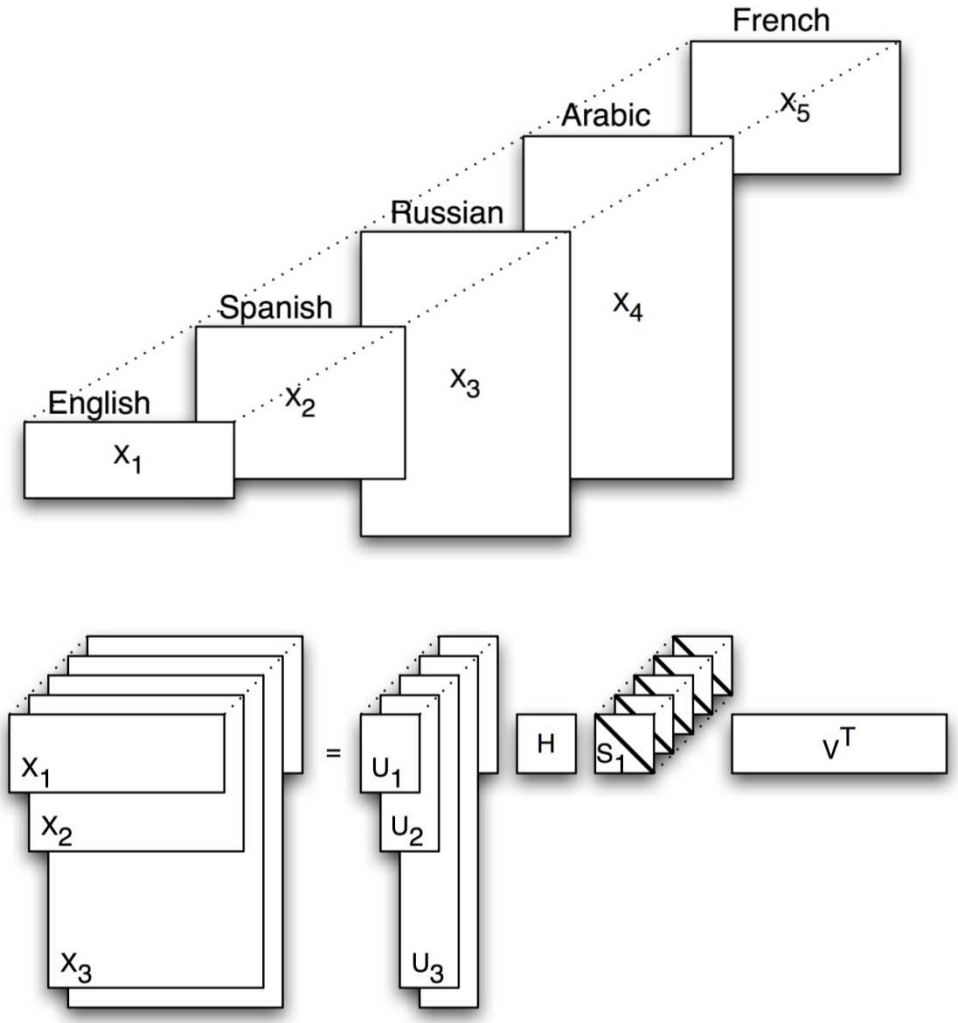
Types and tokens by language

Different languages have different statistics

Language	Types	Tokens
English	12,335	789,744
Russian	47,226	560,524
Spanish	28,456	704,004
Arabic	55,300	440,435
French	20,428	812,947

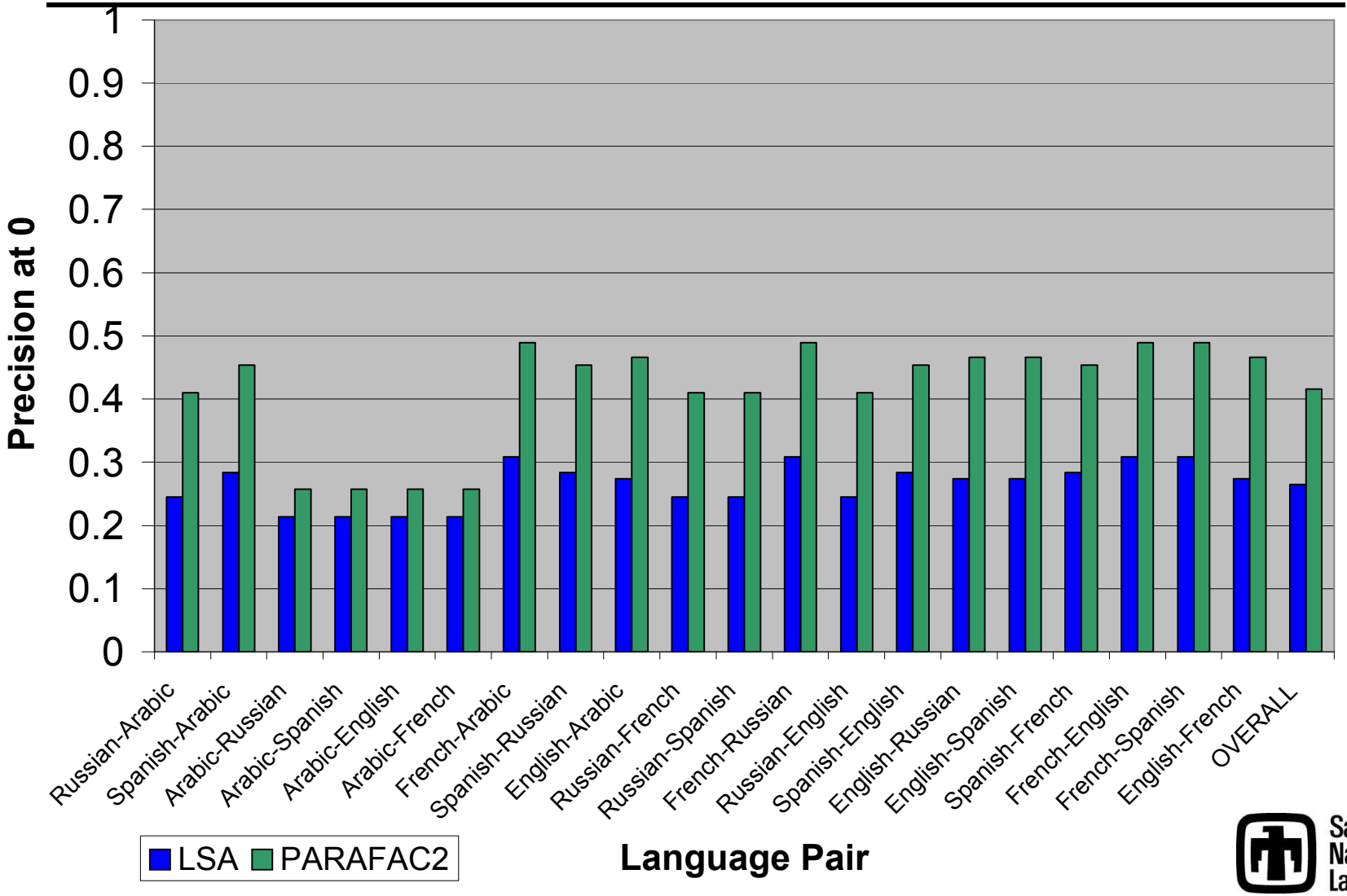
Analytic vs. synthetic languages

The PARAFAC2 model



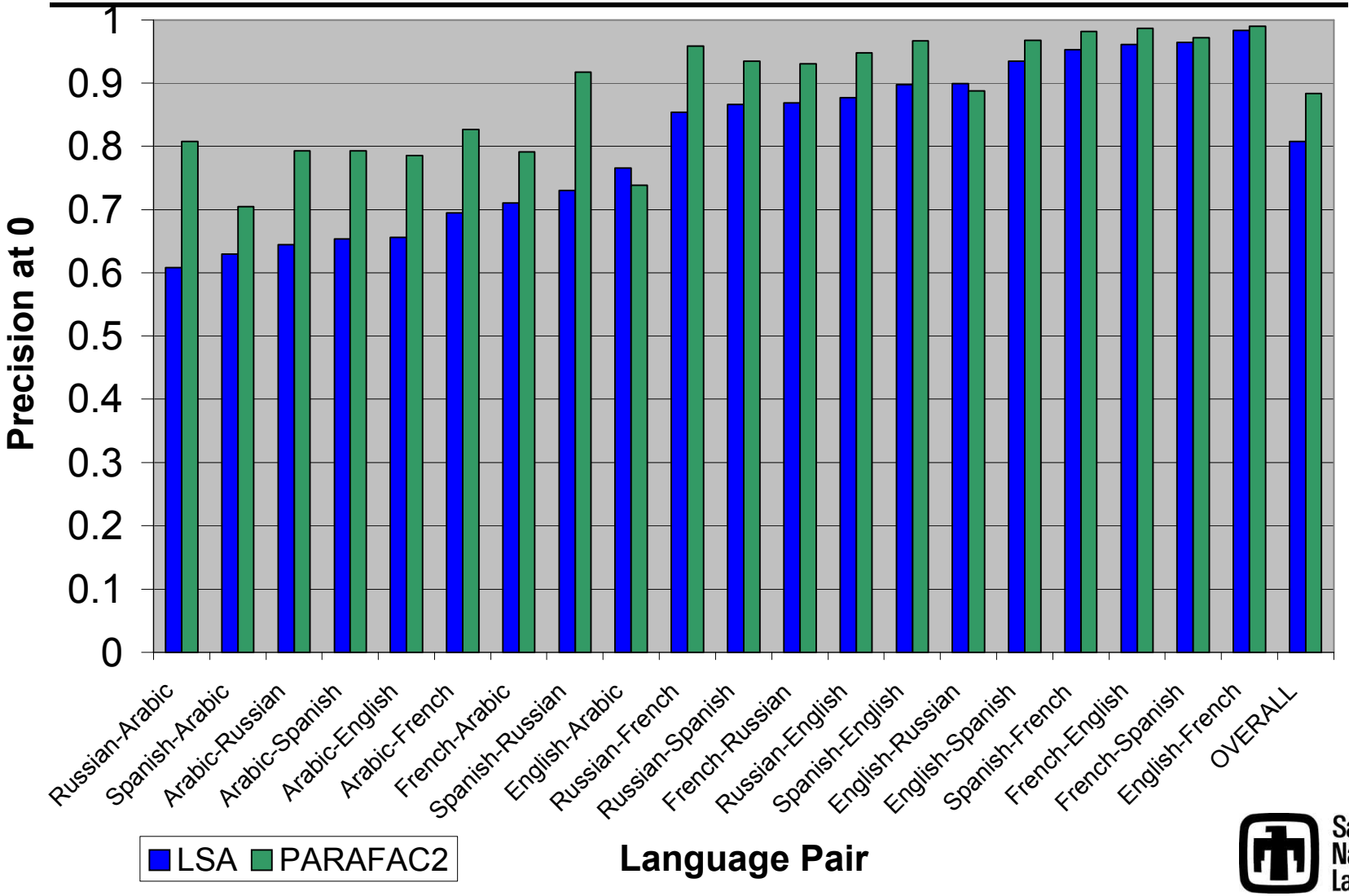
Sandia National Laboratories

Results with PARAFAC2: multilingual



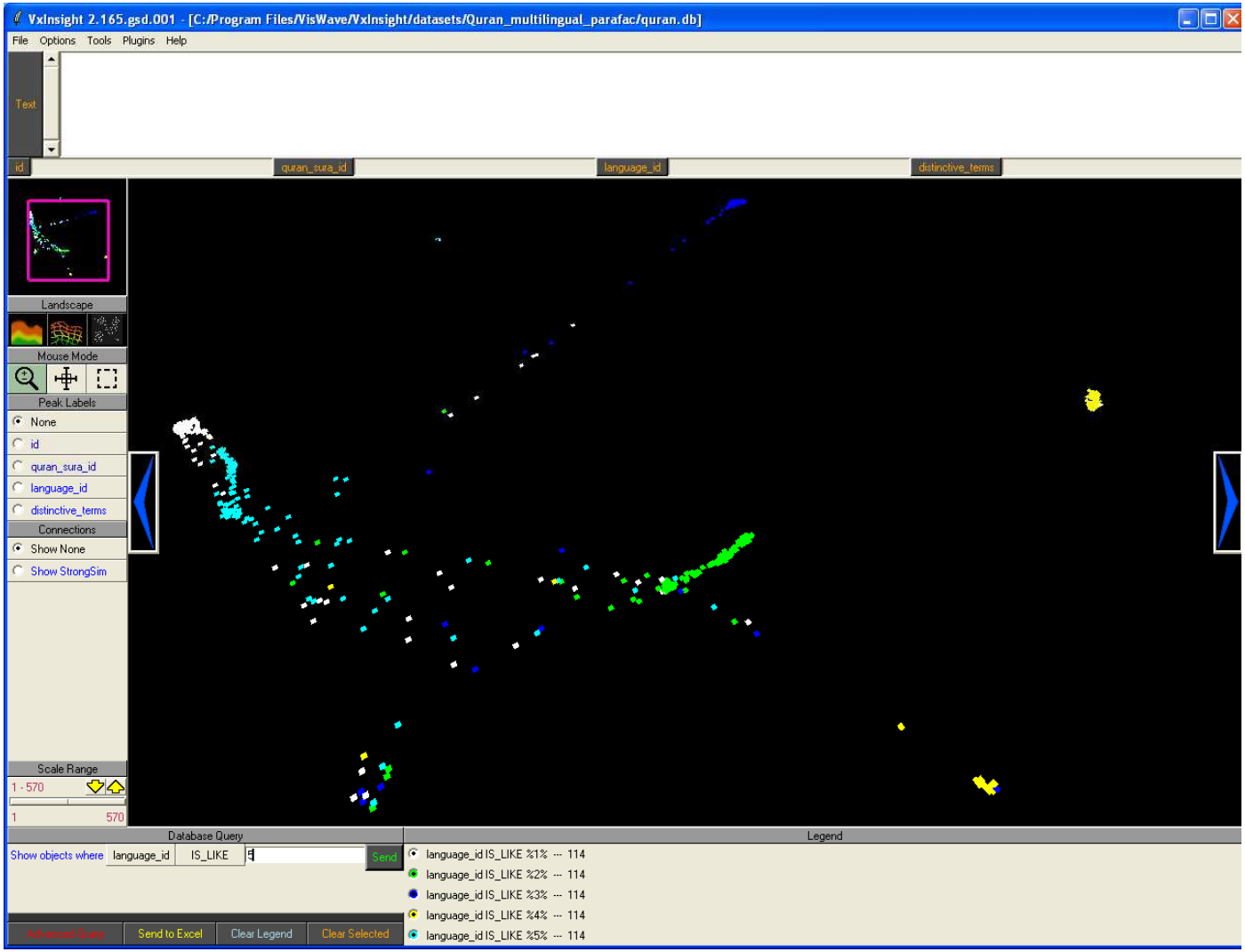
Sandia National Laboratories

Results with PARAFAC2: language-specific



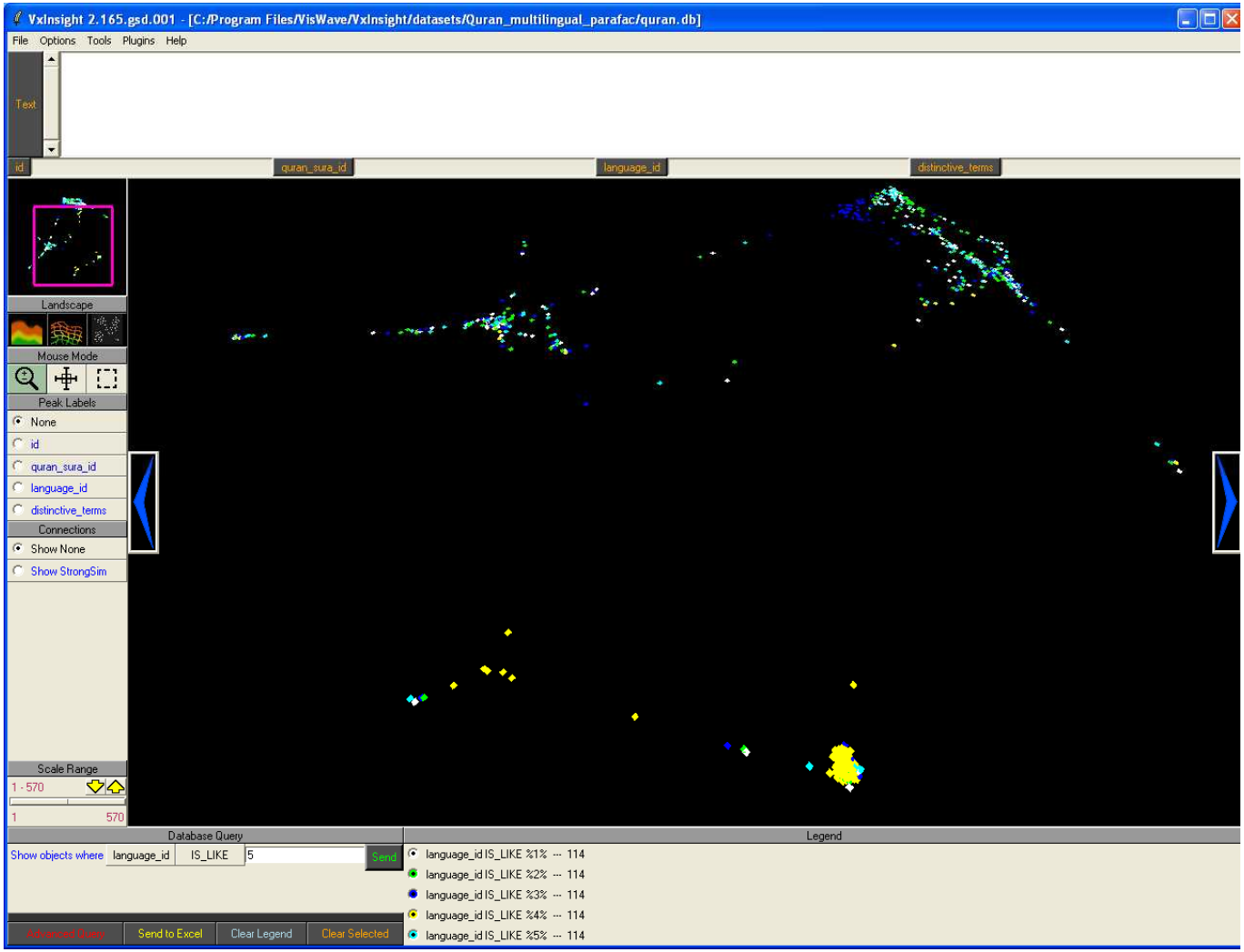
Sandia National Laboratories

Multilingual clustering with LSA



Sandia National Laboratories

Multilingual clustering with PARAFAC2

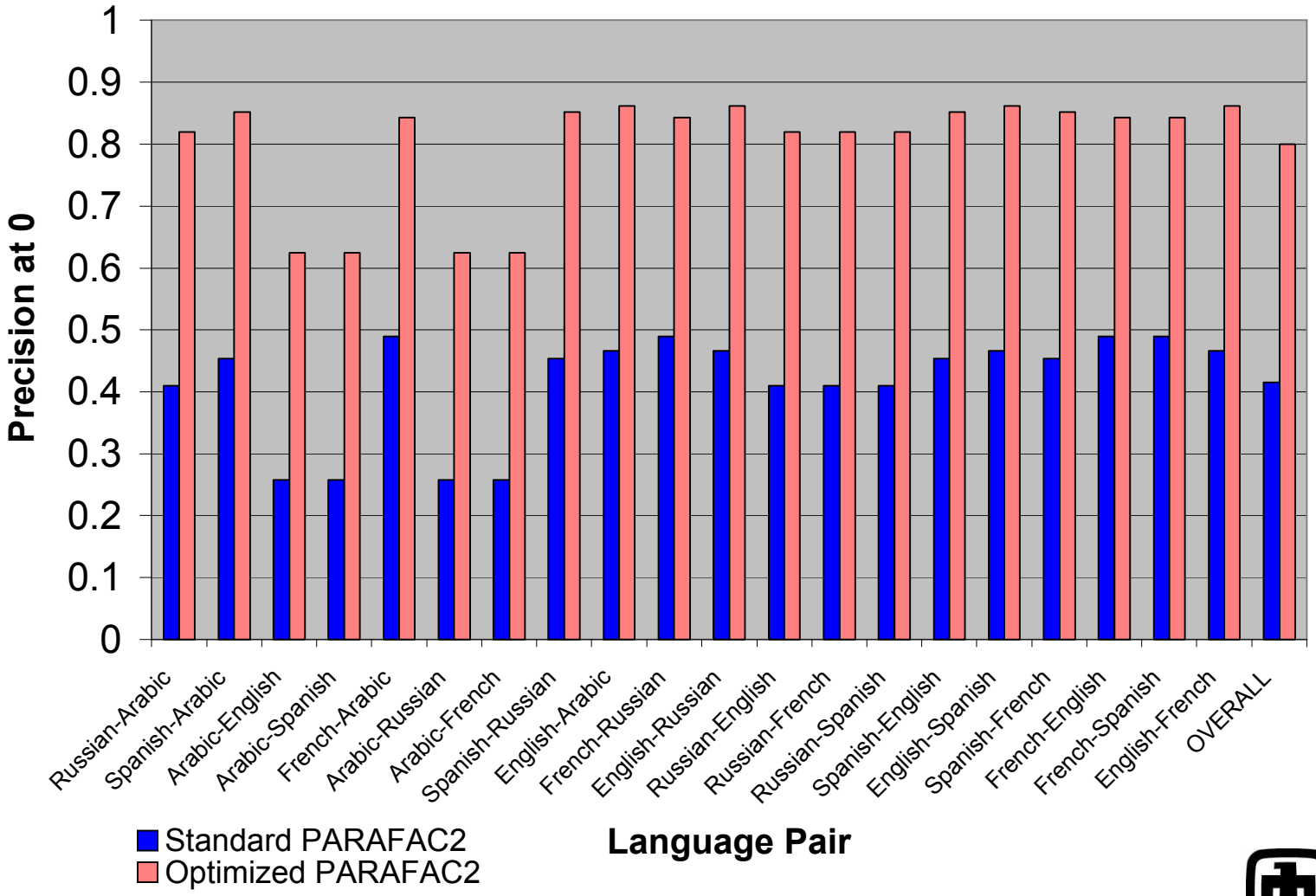


Sandia National Laboratories

Optimization

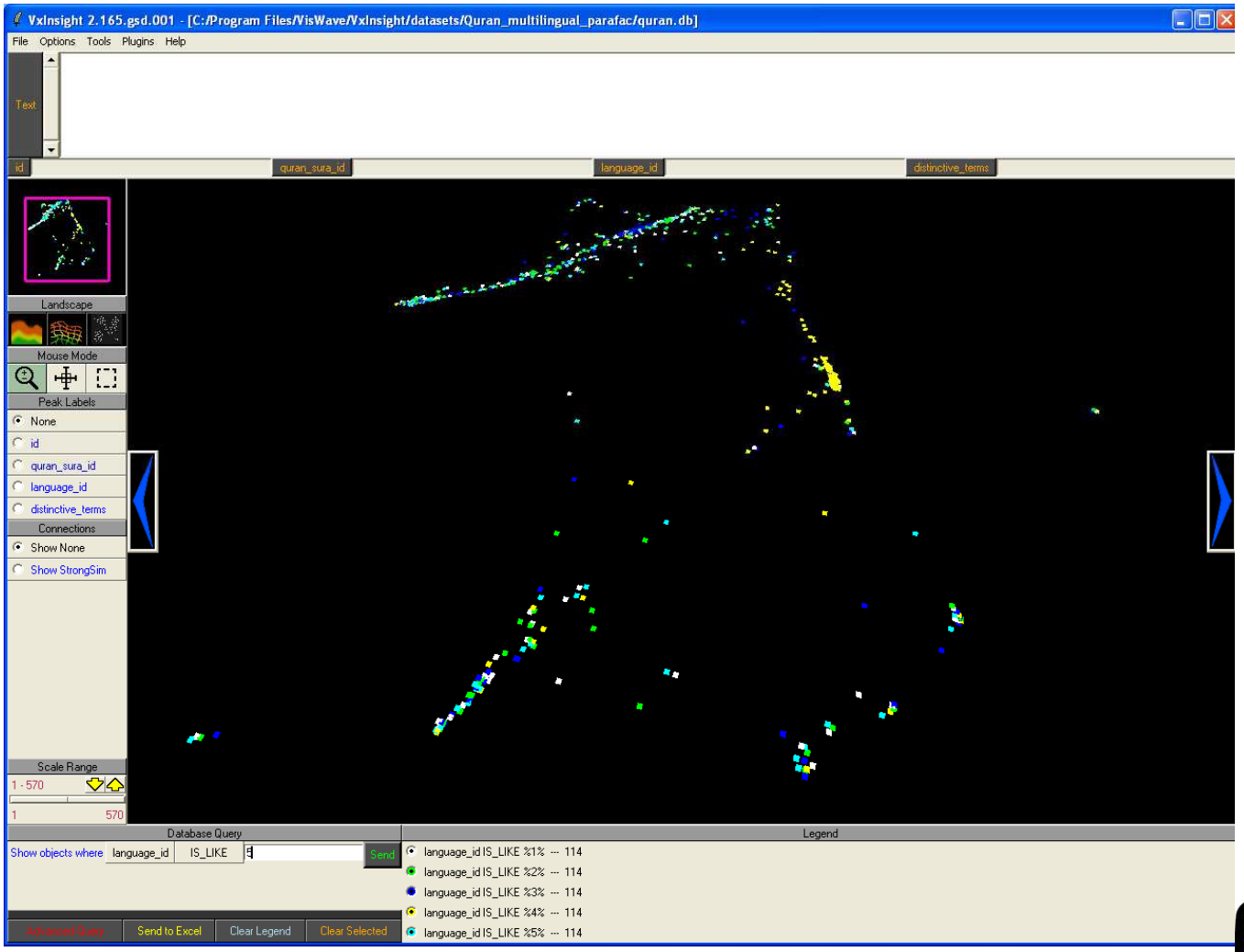
- Separate computations of entropy for training and test corpora
- Raise the global weight to some power
- Elimination of high-entropy terms
- Raise the global weight to some power which varies by language

Results with optimized PARAFAC2



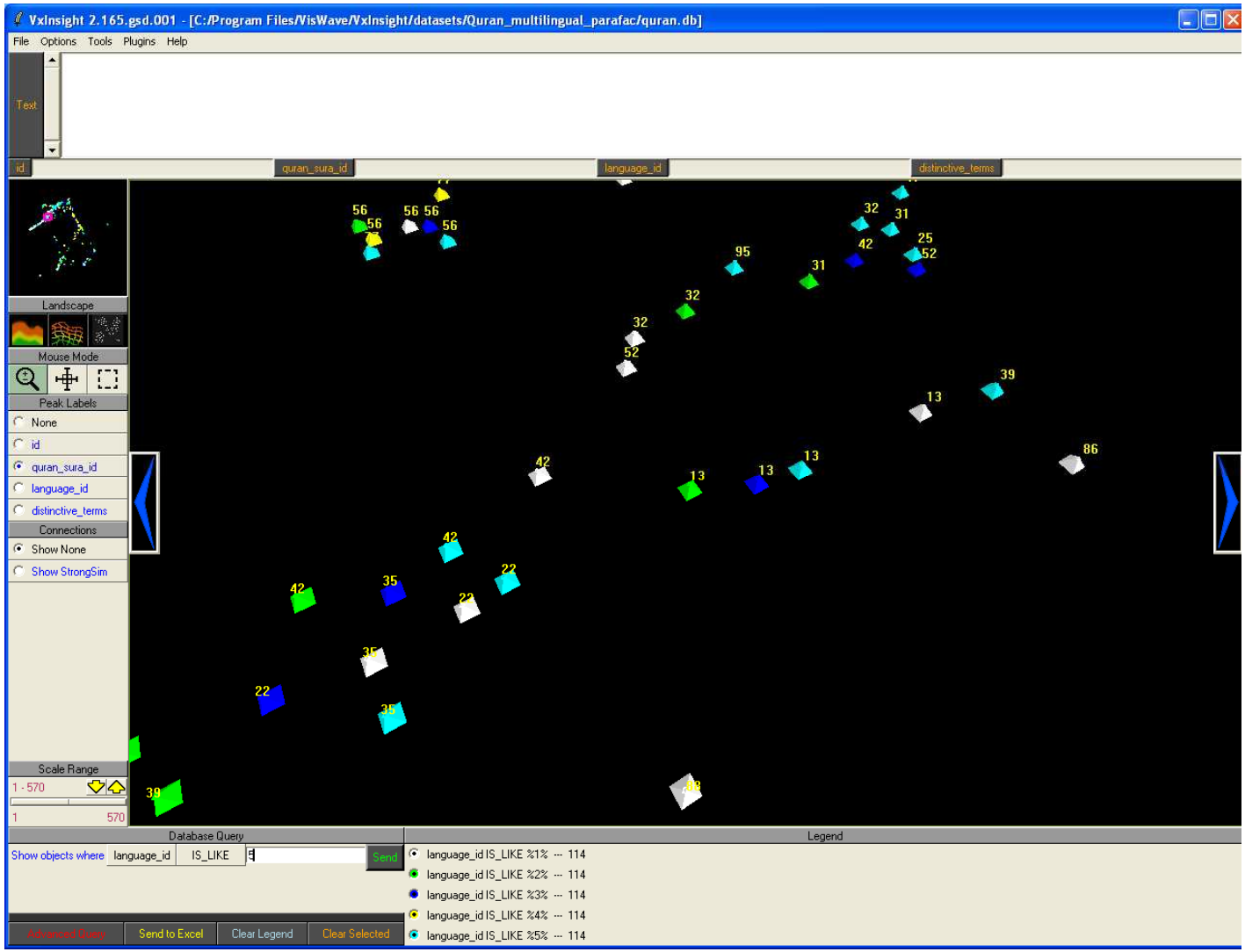
Sandia National Laboratories

Clustering with optimized PARAFAC2 (1)



Sandia National Laboratories

Clustering with optimized PARAFAC2 (2)

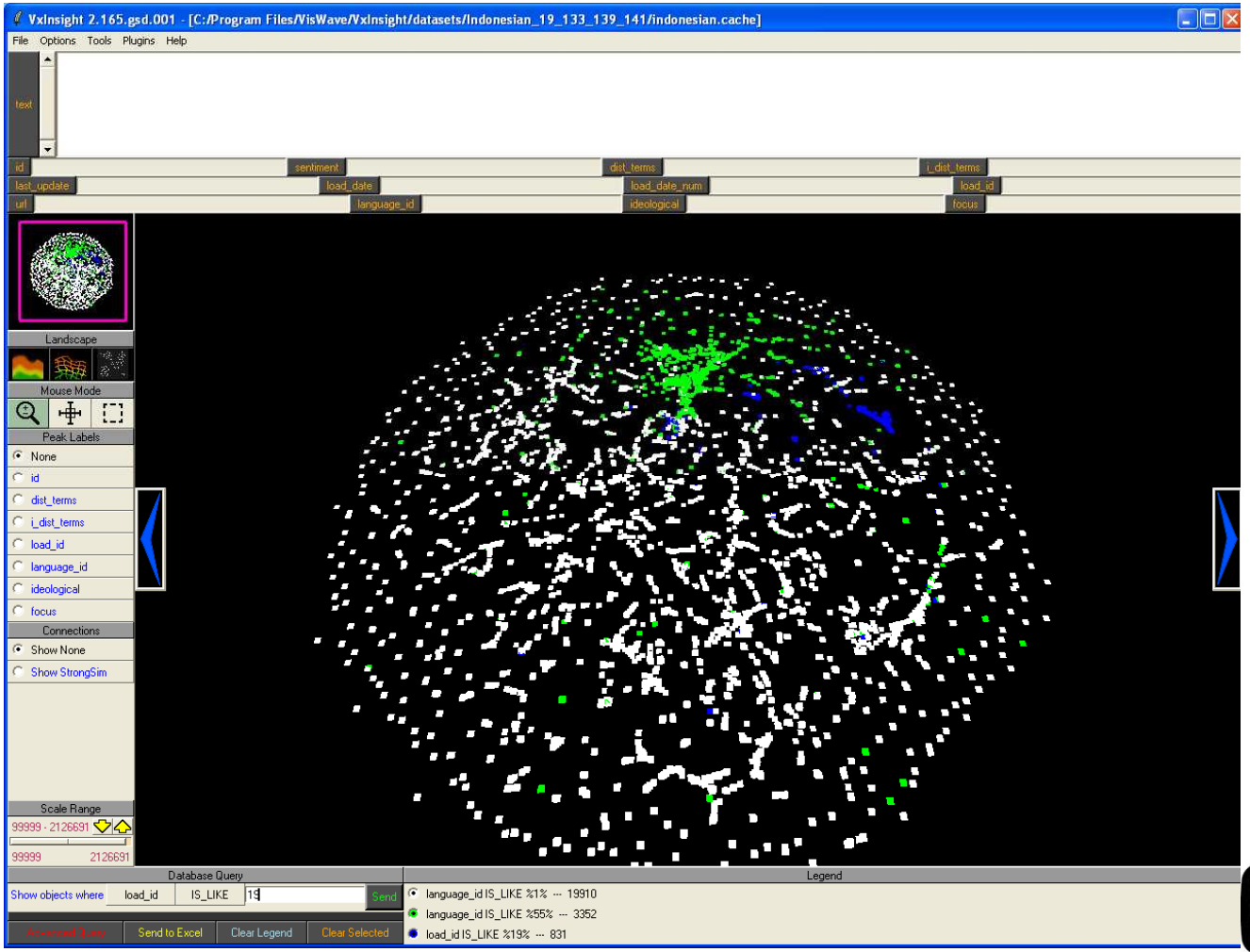


Sandia National Laboratories

Conclusions

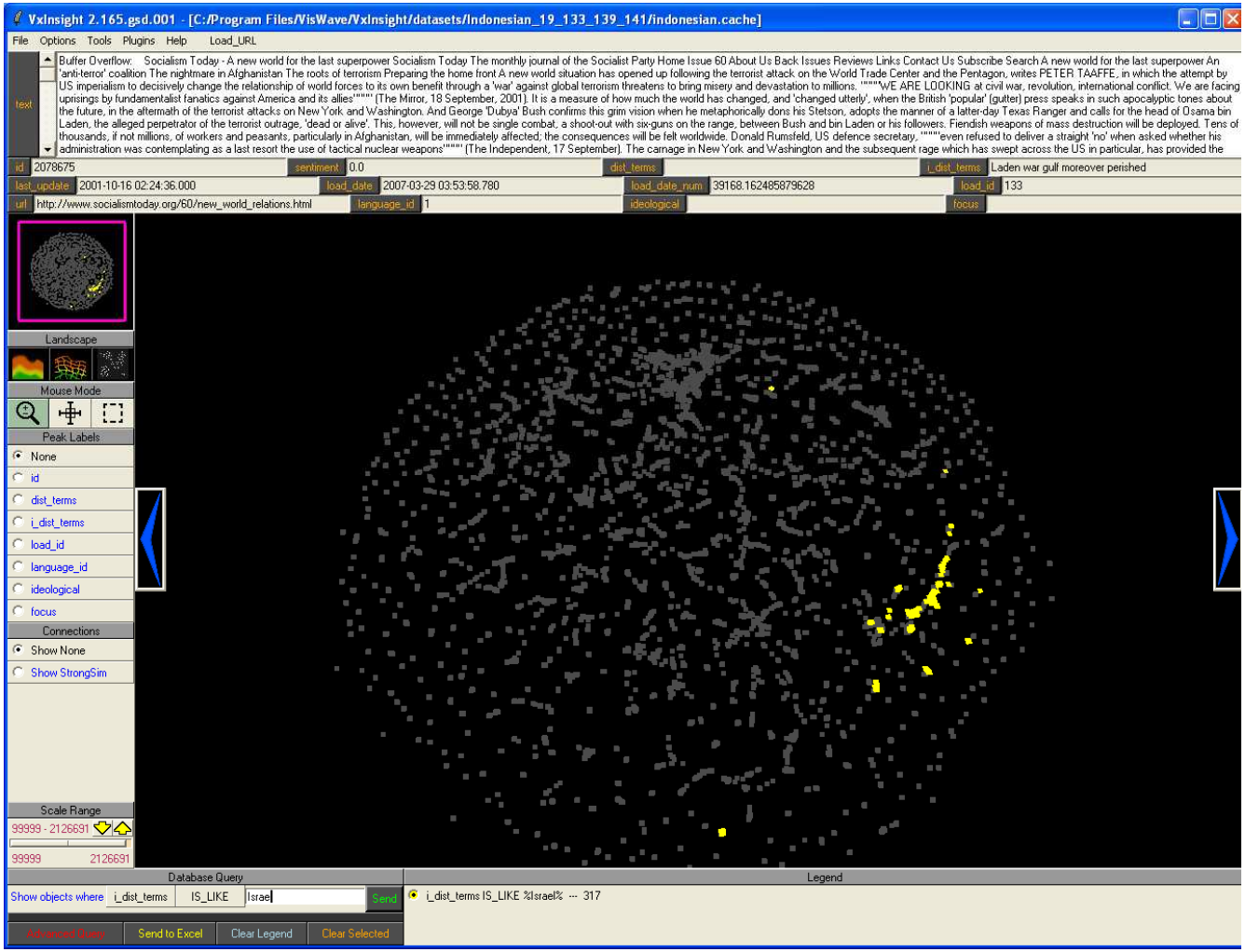
- As a method for multilingual document clustering, PARAFAC2 is clearly superior to LSA
- PARAFAC2 also 'beats LSA at its own game'; information retrieval is improved even when the target languages are known a priori
- Improved empirical results are (we believe) due to the fact that PARAFAC2 better models the underlying linguistic reality
- Disadvantage: implementations of PARAFAC2 are currently not as scalable as those of LSA

Application: visualization of the WWW



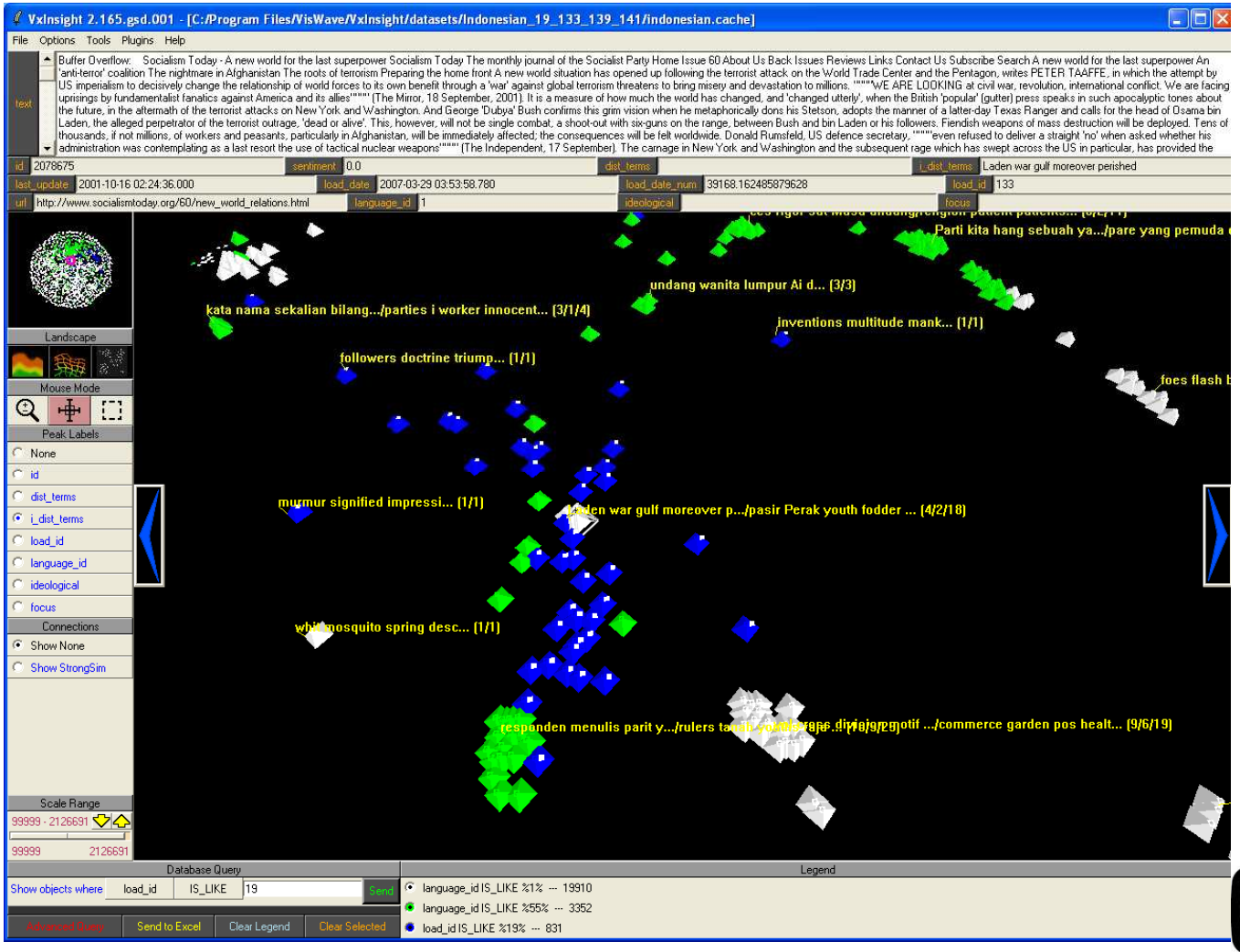
Sandia National Laboratories

Where do topics of interest fit into the overall space?



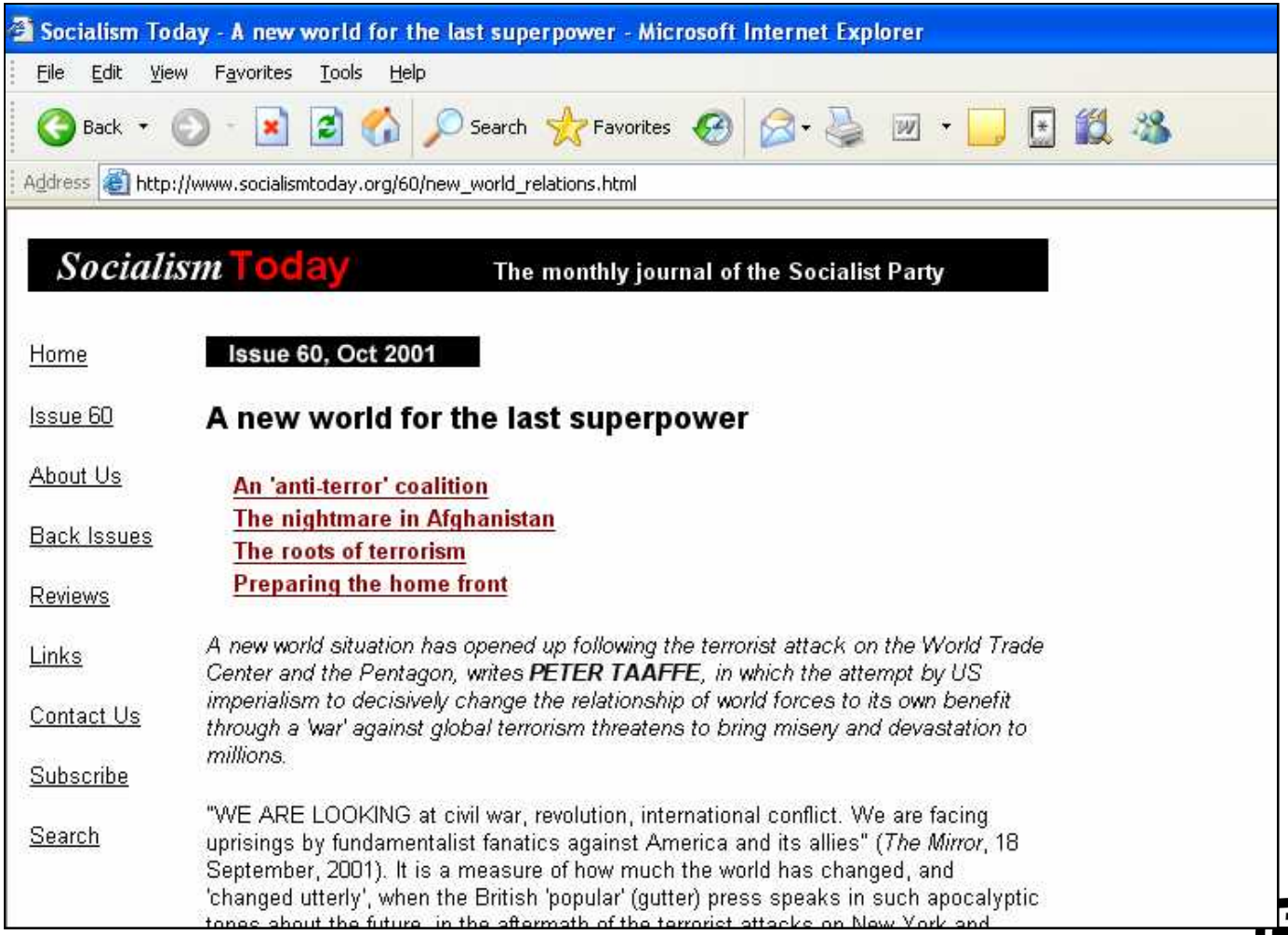
Sandia National Laboratories

Cluster detail



Sandia National Laboratories

Navigation to documents of interest



Sandia National Laboratories

Challenges and future directions: scalability

From tens of thousands of documents…

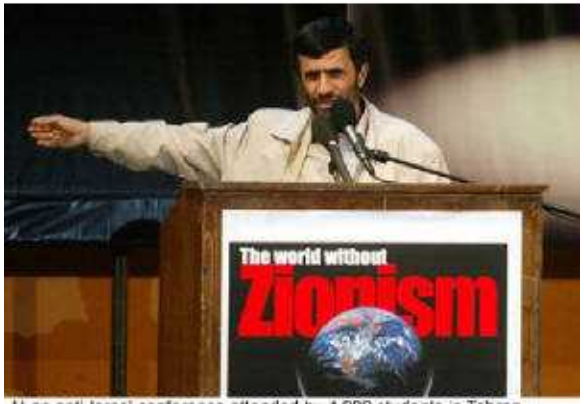
… to hundreds of thousands, or millions, of documents

Sandia National Laboratories

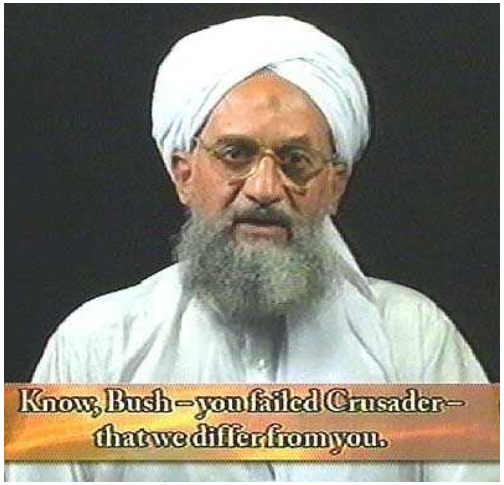
Challenges and future directions: ‘ideological spectroscopy’

Sandia National Laboratories

What is an ‘ideology’?



At an anti-Israel conference attended by 4,000 students in Tehran, titled "The World Without Zionism," President Mahmoud Ahmadinejad of Iran said "Israel must be wiped off the map." (AP Photo)



Sandia National Laboratories

Questions we would like to answer

- Can we say something about an author's beliefs based only on what he/she writes?
- If so, can we automate the process?
- Can we extend the approach so that documents in multiple languages can be evaluated?
- Can we aggregate the results in a meaningful way for millions of documents?
- Can we use the results to detect subtle shifts in opinions through time and space?

Words *do* say something about ideology

- Examples from Middle East politics (Pipes, 2005)
  - 'Jerusalem' or 'al-Quds'?
  - 'Security fence' or 'separation wall'?
  - 'Judea and Samaria' or 'The West Bank'?
  - 'Legally disputed' or 'occupied' territories?
  - The 'cycle of violence'
- An example from closer to home (Lakoff, 1996)
  - 'Tax relief'

Ideological characterization

الرحيم	العالمين	مالك	الحمد	رب	المستقيم	نعبد	اله	الضالين	عليهم	الدين	يوم	الذين	غير	ولا	الله
2	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1

dimension 1	0.1375
dimension 2	0.1052
dimension 3	0.0341
dimension 4	0.0441
dimension 5	-0.0087
dimension 6	0.0410
dimension 7	0.1011
dimension 8	0.0020
dimension 9	0.0518
dimension 10	0.0822
dimension 11	-0.0101
dimension 12	-0.1154
dimension 13	-0.0990
dimension 14	0.0228
dimension 15	-0.0520
dimension 16	0.1096
dimension 17	0.0294
dimension 18	0.0495
dimension 19	0.0553
dimension 20	0.1598

Is there some area of the n-dimensional conceptual space which tends to be inhabited by 'ideological' documents?

…if so, we should be able to predict which documents are 'ideological' simply by statistical analysis of the text

Perhaps with that area of the conceptual space, we can distinguish between sub-areas for different types of ideology?

Sandia National Laboratories

Selected Publications

- Chew, Verzi, Bauer and McClain. 2006. Evaluation of the Bible as a Resource for Cross-Language Information Retrieval. *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*, 2006, 68–74.

- Chew and Abdelali. 2007. Benefits of the 'massively parallel Rosetta Stone': cross-language information retrieval with over 30 languages, *Proceedings of the Association for Computational Linguistics conference*, 2007.

- Chew, Bader, Kolda and Abdelali. 2007. Cross-language information retrieval using PARAFAC2. Proceedings of KDD 2007.

Sandia National Laboratories

DISCUSSION
and
QUESTIONS

SANDIA POINT OF CONTACT:
Peter Chew (pchew@sandia.gov)

Sandia National Laboratories