

High-Level Visualization of Multilingual Text from the Internet: Using Linguistic Methods to Identify Topics of Interest

Seminar at the
Physical Science Laboratory, NMSU

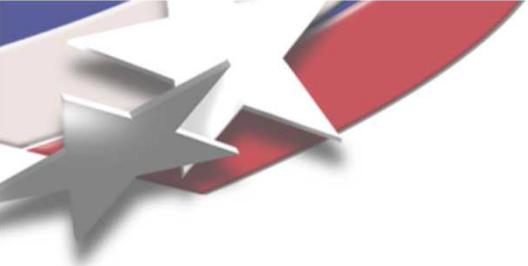
May 22, 2007

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.



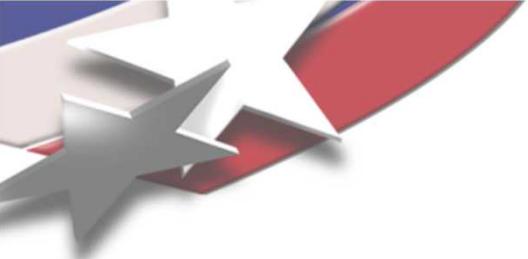


Vision

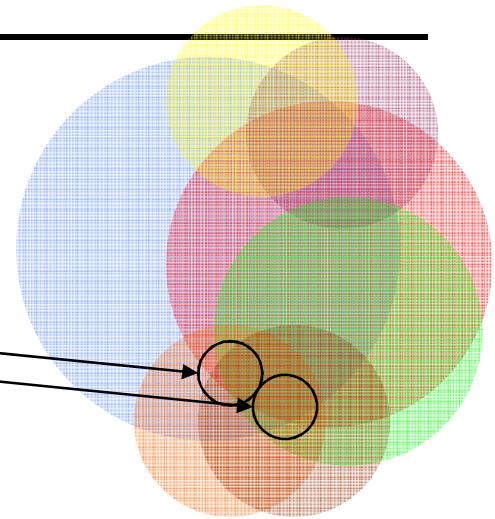
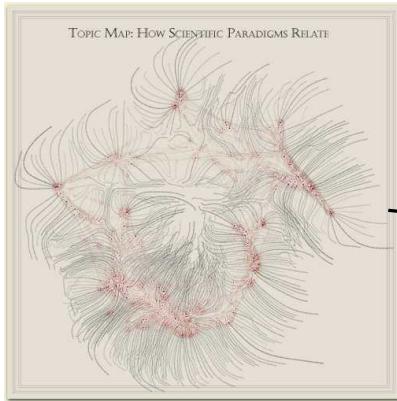
Imagine that you can create a representation, using a graphical or spatial layout, of all blogs and similar pages on the internet.

- What will it look like?
- What will it contain?
- How will it be used?
- What will people be able to learn from it?
- Can it be predictive?

- How do we make it?



Recent work at Sandia



ISI database

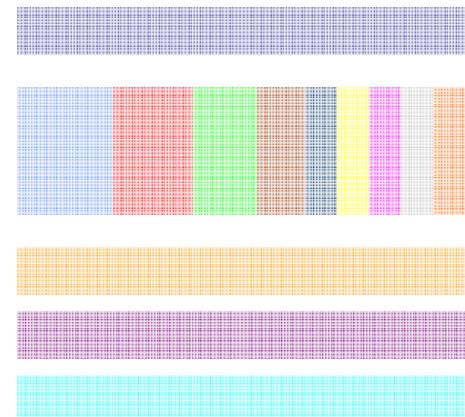
One year – 1M papers

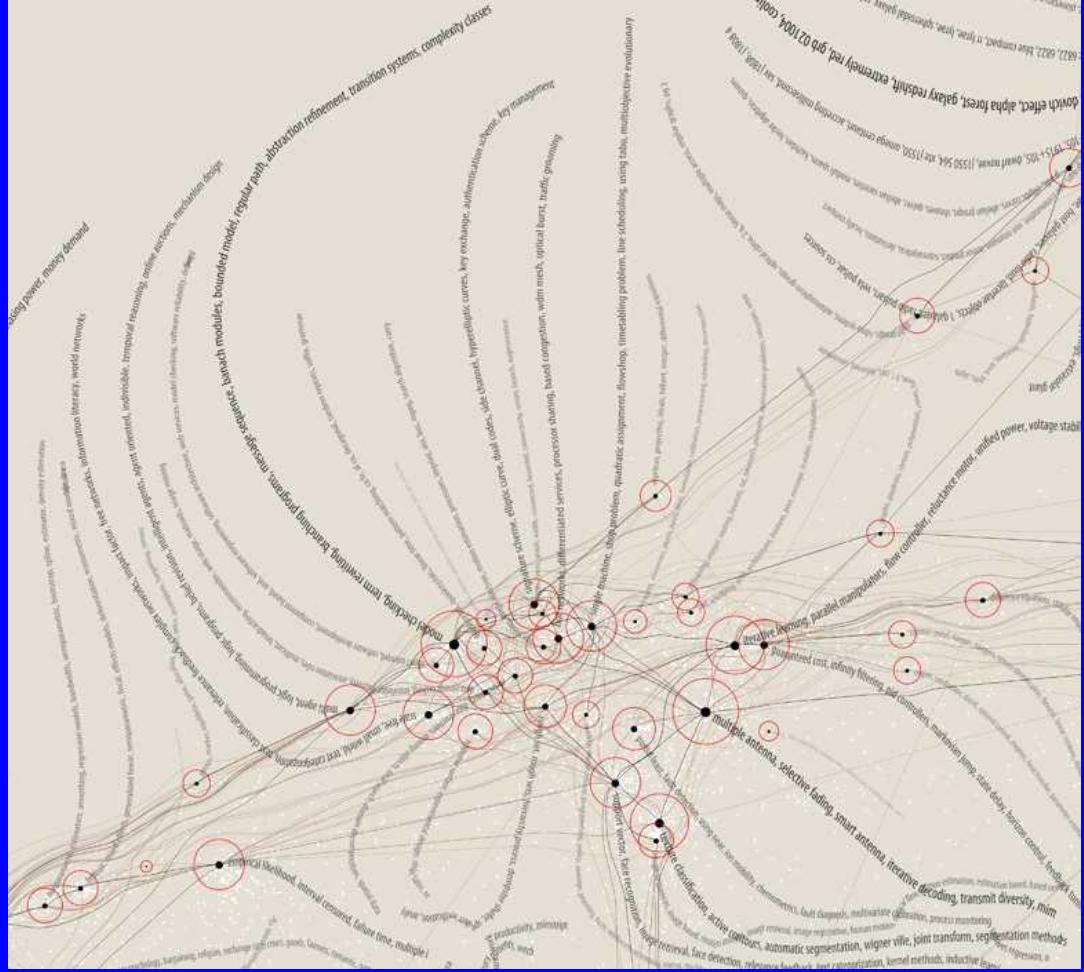
Relatively clean, fielded data

Matching refs to previous papers

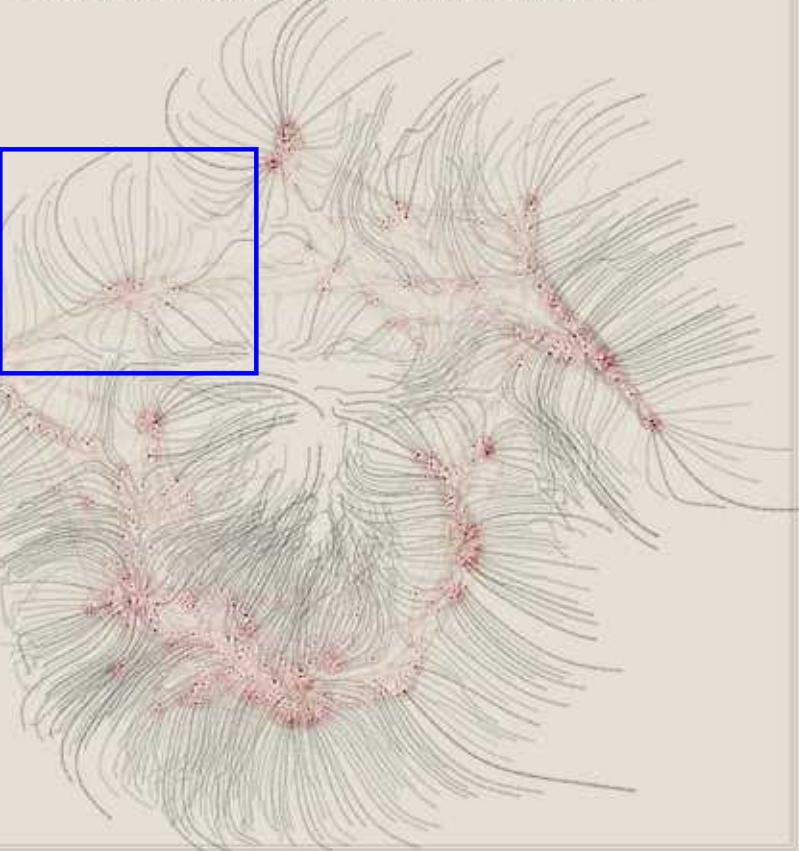
Million node layout

- 100k clusters, very precise topic focus





TOPIC MAP: HOW SCIENTIFIC PARADIGMS RELATE



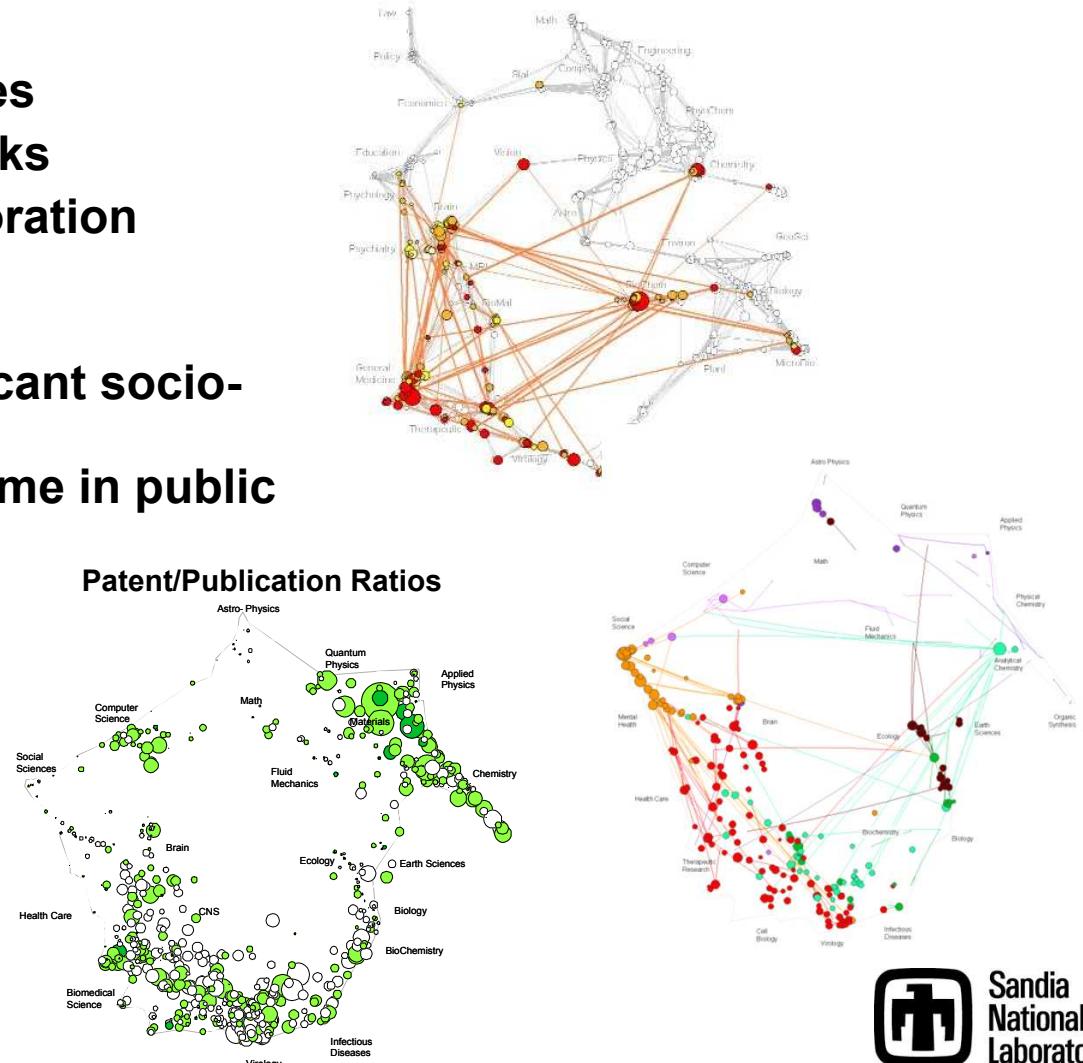
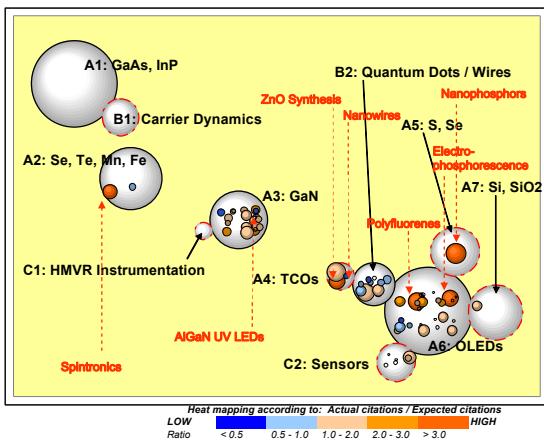
Uses

CURRENT

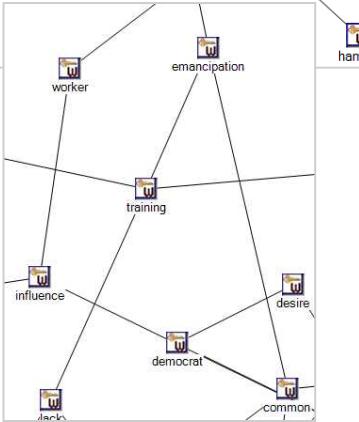
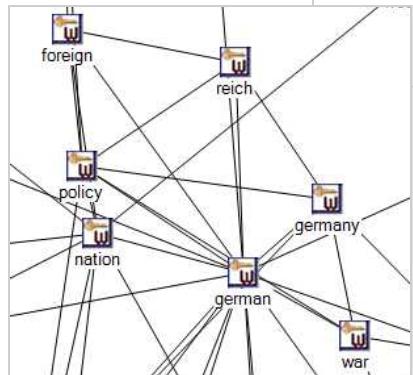
- Agency, corporate profiles
- National strength networks
- Opportunities for collaboration

FUTURE?

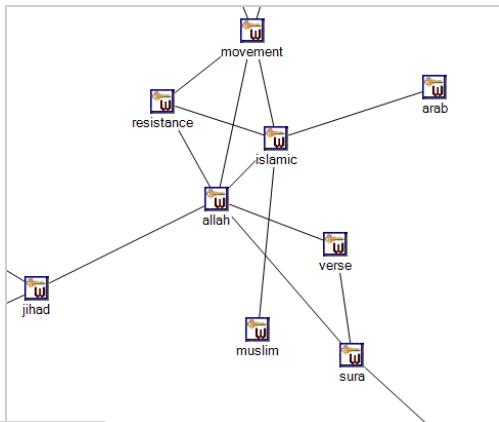
- Early detection of significant socio-political shifts
- Tracking changes over time in public interest in certain topics



The internet as a marketplace of ideas



words



ideas

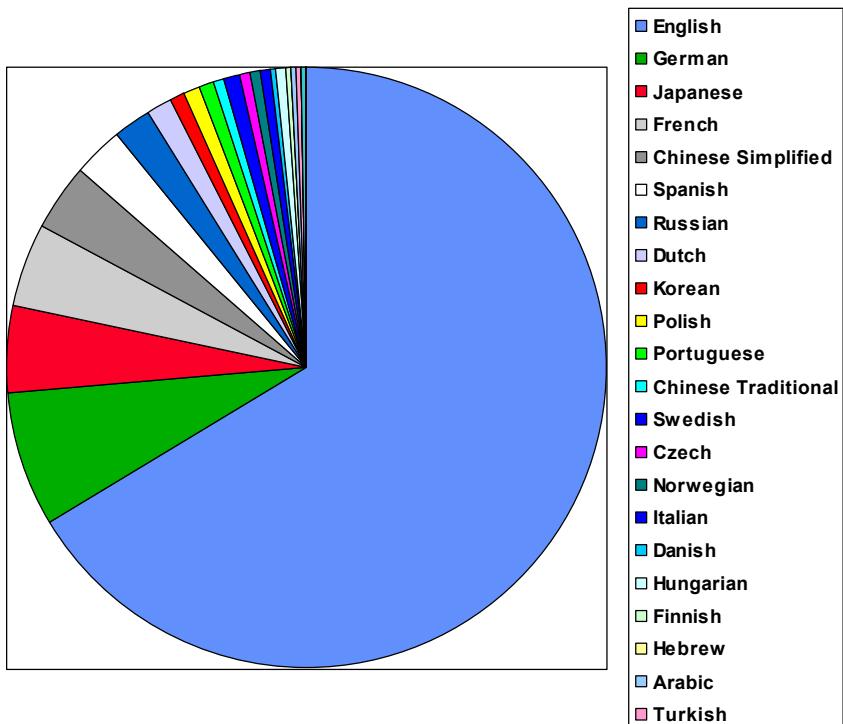


movements

Multilingual information retrieval

(identification of similar documents in different languages)

Languages of the internet



Our framework:

- Identifies similar documents across language boundaries
- Can handle 81 languages
- Covers about 99.75% of internet content (based on chart at left)
- Is easily extensible to new languages
- Identifies translations of documents with up to 99% precision
- Enables users to find documents of interest in an unknown language

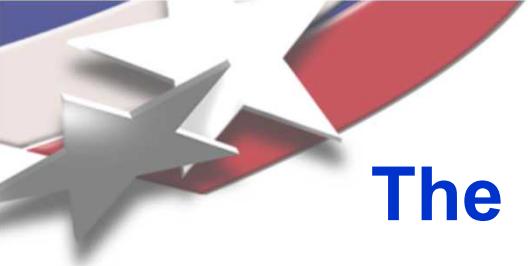


Approaches to cross-language information retrieval

- Translate the query
 - Efficacy is constrained by quality of machine translation
- Train algorithm on parallel corpora
 - Translations should:
 - Be available in target languages
 - Be reliable
 - Be sufficiently large in size
 - Cover target subject domain
 - Be free of undue copyright restrictions
 - Be electronically available
 - Be alignable

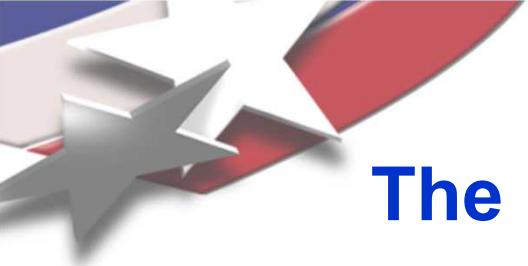


```
health of animals act ===> loi
health of the economy ===> santé
heavy burden ===> lourd fardeau
heavy equipment ===> matériel lourd
heritage day ===> jour du patrimoine
higher premiums ===> cotisation
highest bidder ===> plus offrant
historic document ===> document
historic event ===> événement
house management committee ===>
house of commons standing commit
```



The Bible as a ‘Rosetta Stone’ (1)

- Resnik, Olsen & Diab (1999) showed that the Bible fulfills all of these criteria and is surprisingly suitable as a parallel corpus
 - Translations in > 2,400 languages and rising
 - Great care taken over translations
 - Respectably large compared to other corpora
 - Covers many modern genres
 - Covers up to 85% of modern vocabulary
 - Generally free of copyright restrictions
 - Electronically available
 - Alignable



The Bible as a 'Rosetta Stone' (2)

Language	Text	Words	%
EN	In the beginning God created the heavens and the earth.	10	23.8
RU	В начале сотворил Бог небо и землю.	7	16.7
ES	EN el principio crió Dios los cielos y la tierra.	10	23.8
AR	في البدء خلق الله السموات والارض.	6	14.3
FR	Au commencement Dieu créa les cieux et la terre.	9	21.4
TOTAL		42	100.0

Parallel documents (verses)	31,226
Languages with translations	2,426
Electronically-available parallel translations	80+
Coverage of internet content	99.75%



Potential language coverage - detail

A STATISTICAL SUMMARY OF LANGUAGES WITH THE SCRIPTURES

A summary, by geographical area and type of publication, of the number of different languages and dialects in which publication of at least one book of the Bible had been registered as of December 31, 2005.

Continent/Region	Portions	Testaments	Bibles	Bibles, DC*	Total
Africa	223	301	159	(29)	683
Asia	218	244	131	(28)	593
Australia/New Zealand/ Pacific Islands	148	234	38	(9)	420
Europe	114	36	61	(47)	211
North America	39	30	7	(0)	76
Caribbean Islands / Central America / Mexico/South America	118	270	29	(9)	417
Constructed Languages	2	0	1	(0)	3
TOTALS	862	1,115	426	(122)	2,403

* This column is a sub-section of the Bibles column – for example, there is a translation of the Deuterocanon for 47 of the 61 languages of Europe in which the Bible has been translated.

[A few corrections were made to our language databases and are reflected in this statistical summary]

Per <http://www.biblesociety.org/latestnews/latest341-slr2005stats.html>

The 'Unbound Bible'

The Unbound Bible - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address http://www.unboundbible.com/index.cfm?method=downloads.showDownloadMain

The Unbound Bible

[English] [Русский] [한국어]
[Translate] [Nederlands] [Français] [עברית] [English]

Bible Search Bible Study Tools Know God Downloads Translate

Bible in a Year Search Settings FAQ Links Contact

Download Unicode Fonts and Bibles

Unicode Bibles

Afrikaans 1953 † Download

Afrikaans 1953 †
Albanian †
Amharic NT
Arabic: Smith & Van Dyke †
Aramaic NT: Peshitta †
Armenian (Eastern): (Genesis, Exodus, Gospels)
Armenian (Western): NT (Dwight/Riggs, 1853)
Basque (Navarro-Labourdin): NT
Breton Gospels
Chamorro (Psalms, Gospels, Acts)
Chinese: NCV (Traditional) †
Chinese: Union (Simplified) †
Chinese: NCV (Simplified) †
Chinese: Union (Traditional) †
Croatian 2.0 †
Czech BKR †
Czech CEP †
Czech KMS †
Czech NKB †
Danish †
Dutch Staten Vertaling †
English: King James Version 2.0
English: American Standard Version
English: Basic English Bible
English: Darby Version
English: Douay-Rheims 2.0
English: Webster's Bible
English: Weymouth NT
English: World English Bible
English: Young's Literal Translation

mir Rybant has written a free software program that works with Bibles in the Unbound Bible format. You can look up a passage,

UniBible - A Multilingual Bible Reader for PalmOS

ilingual Bible-reader program for the PalmOS platform. It uses Unicode to display Bible texts in various languages, including:

els and accents!
Hebrew, Syriac
s
load for Free!

Unicode Fonts

ut requires registration)

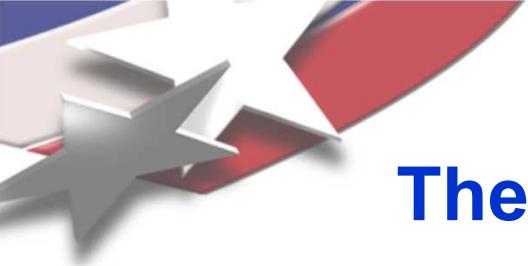
GB18030 Support Package

85 translations (some partial) in
51 languages, in common format

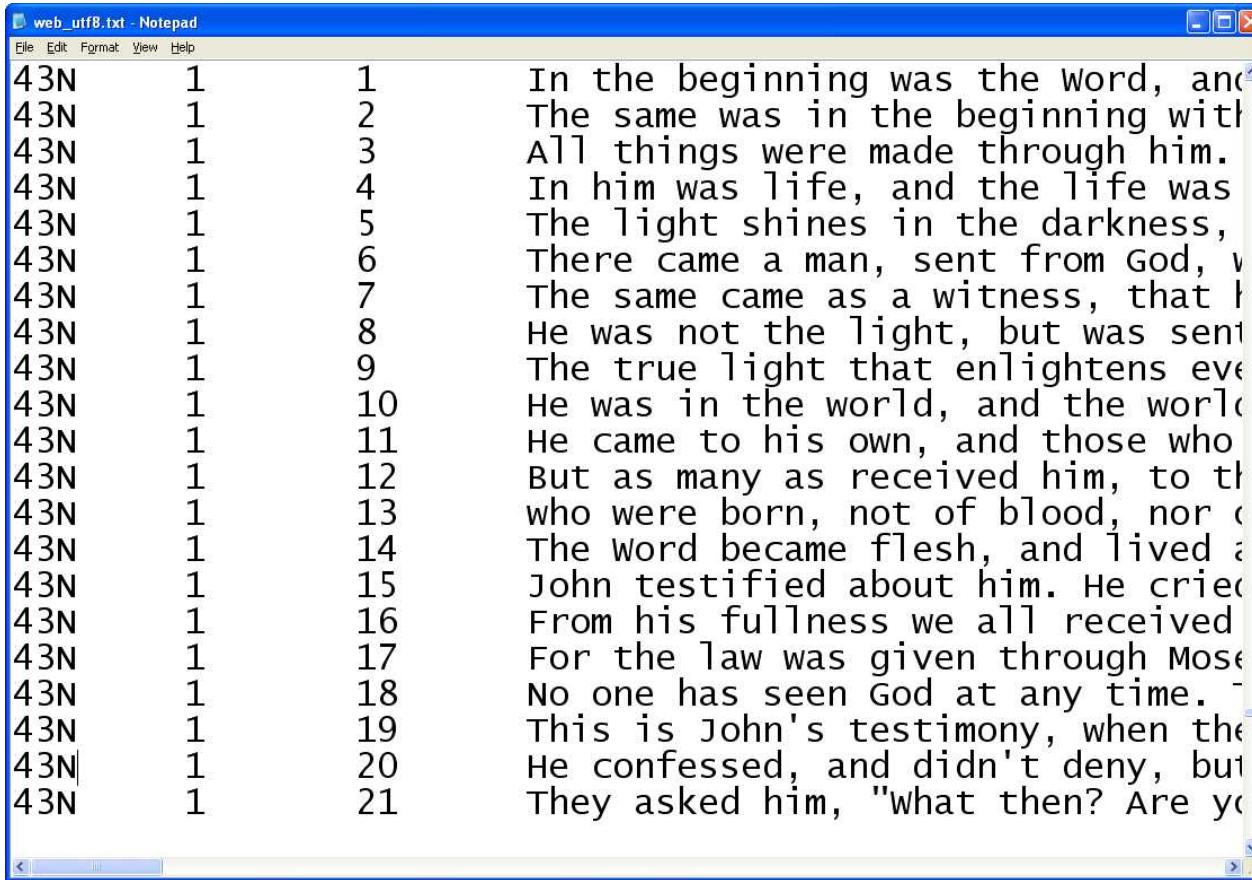
Internet



Sandia
National
Laboratories



The ‘Unbound Bible’ – a sample



43N 1 1 In the beginning was the Word, and the Word was with God, and the Word was God.
43N 1 2 The same was in the beginning with God, and the Word was God.
43N 1 3 All things were made through him; and without him was not anything made that was made.
43N 1 4 In him was life, and the life was the light of men.
43N 1 5 The light shines in the darkness, and the darkness did not comprehend it.
43N 1 6 There came a man, sent from God, whose name was John.
43N 1 7 The same came as a witness, to bear witness of that light.
43N 1 8 He was not the light, but was sent to bear witness of the light.
43N 1 9 The true light that enlightens everyone, was coming into the world.
43N 1 10 He was in the world, and the world did not know him.
43N 1 11 He came to his own, and those who were his own did not receive him.
43N 1 12 But as many as received him, to them he gave the right to become children of God, even to those who were born, not of blood, nor of the will of man, nor of the will of God.
43N 1 13 The Word became flesh, and lived among us, and we have seen his glory, the glory as of the only begotten of the Father, full of grace and truth.
43N 1 14 John testified about him. He cried out, saying, "This was my beloved Son, in whom I am well pleased."
43N 1 15 From his fullness we all received, and grace upon grace.
43N 1 16 For the law was given through Moses; but grace and truth came through Jesus Christ.
43N 1 17 No one has seen God at any time. If you have seen me, you have seen God; and he is in me, and I am in him.
43N 1 18 This is John's testimony, when the Jews sent priests and Levites from Jerusalem to ask him, "Who are you?"
43N 1 19 He confessed, and didn't deny, but said, "I am not the Christ."
43N 1 20 They asked him, "What then? Are you Elijah?" He said, "I am not." They said, "Are you the Prophet?" He answered, "I am not."
43N 1 21 They said to him, "Who are you? Let us know, so that we may give an answer to those who sent us. What do you say about yourself?"



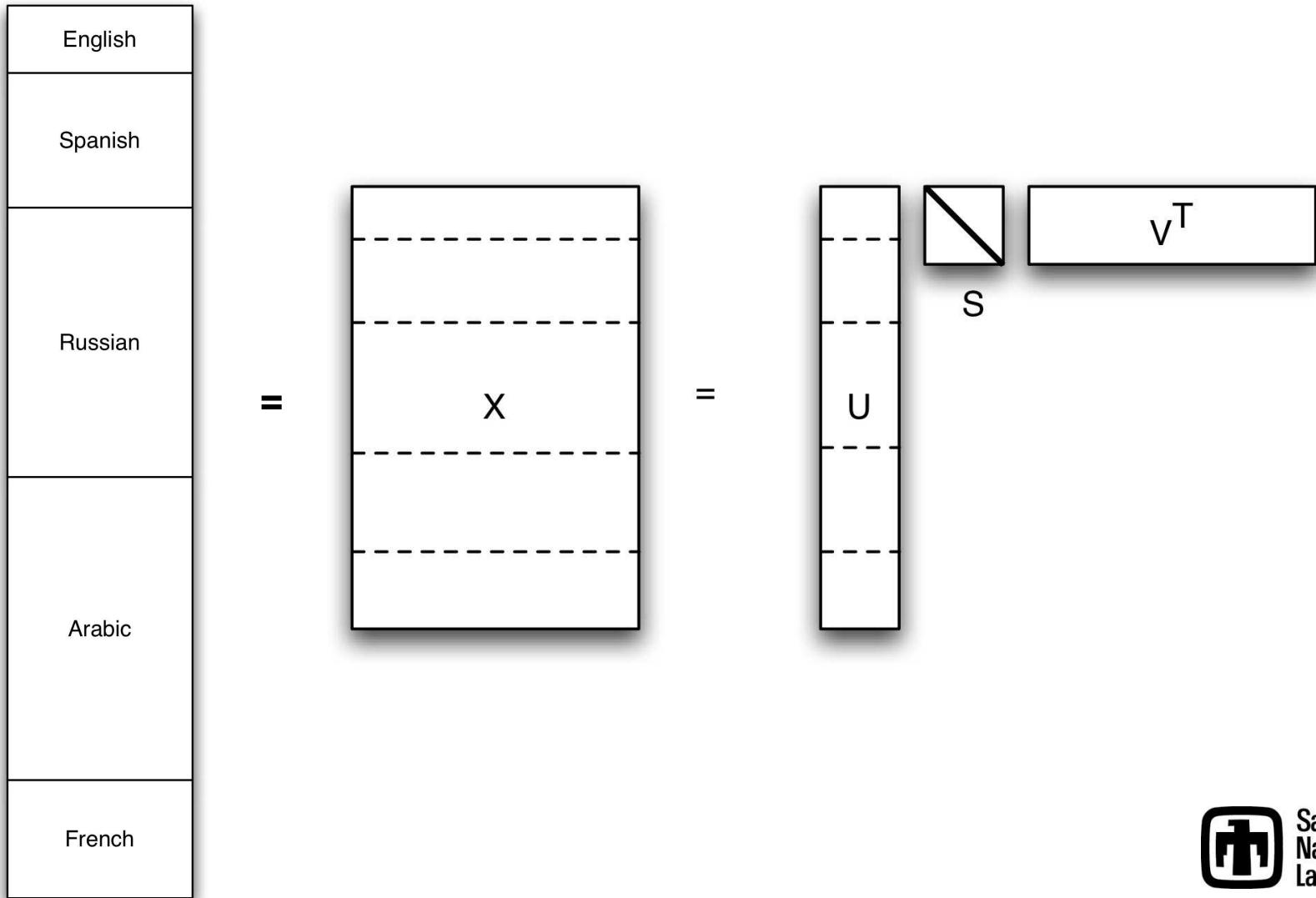
Languages in our implementation

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

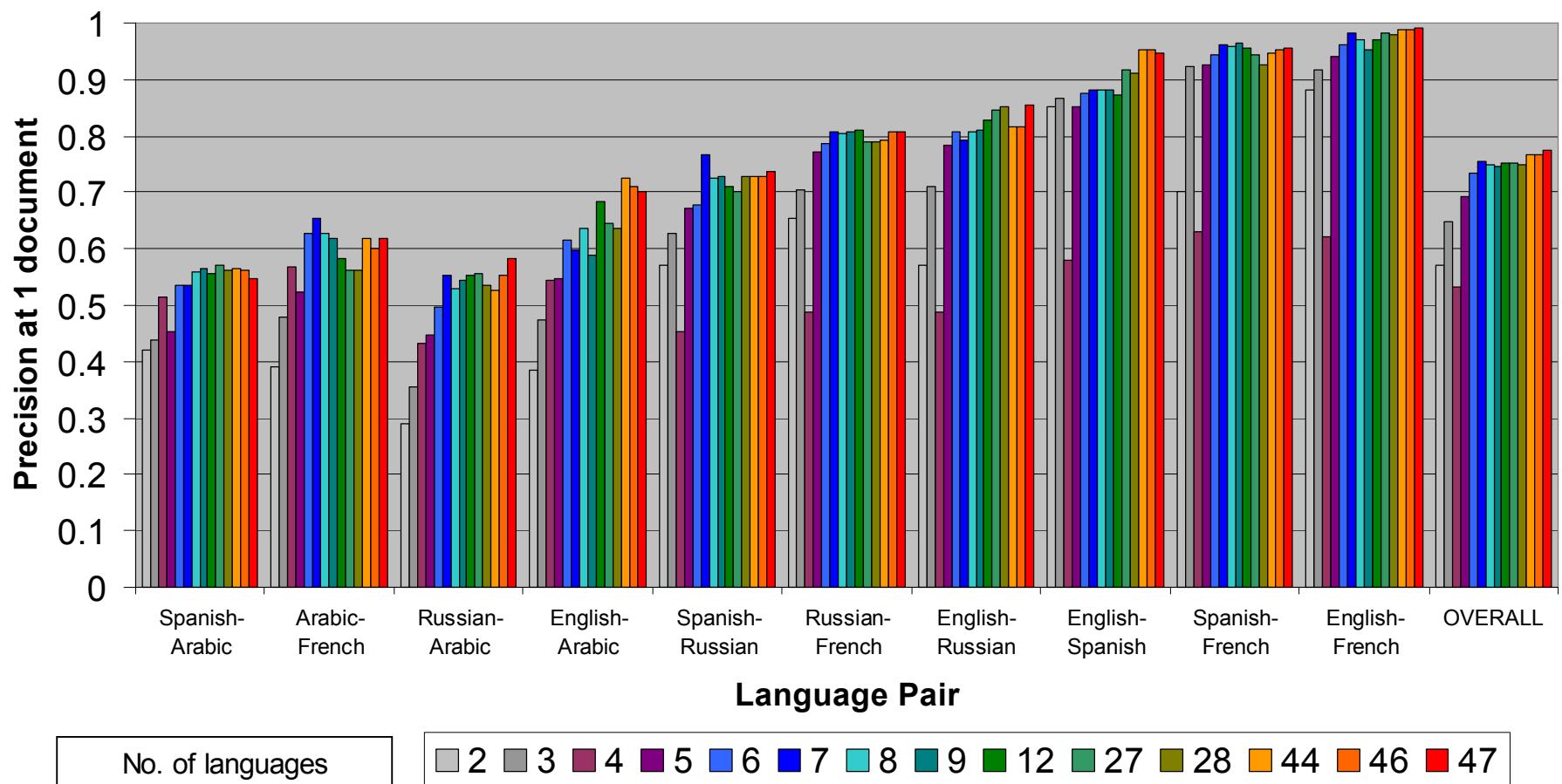


Sandia
National
Laboratories

The standard multilingual LSA model

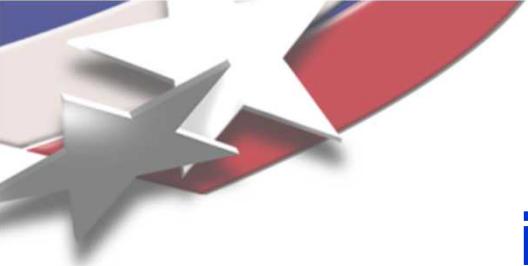


‘Massive linguistic parallelism’ helps



No. of languages

2 3 4 5 6 7 8 9 12 27 28 44 46 47



Technical challenges in multilingual clustering

- With MLSA, documents do not cluster language-independently

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



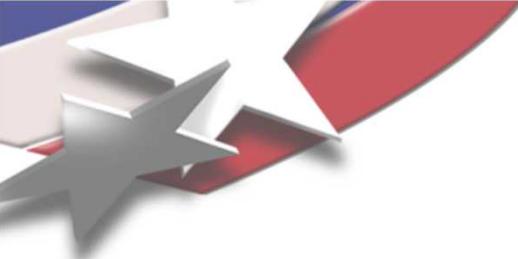


Why standard LSA has drawbacks

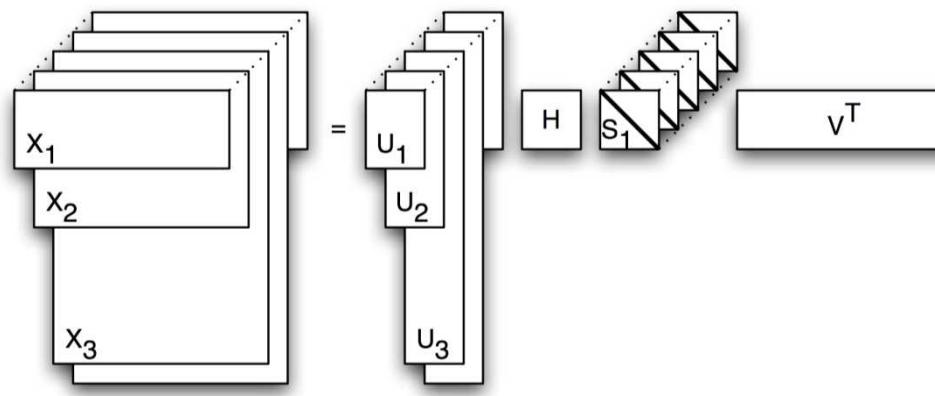
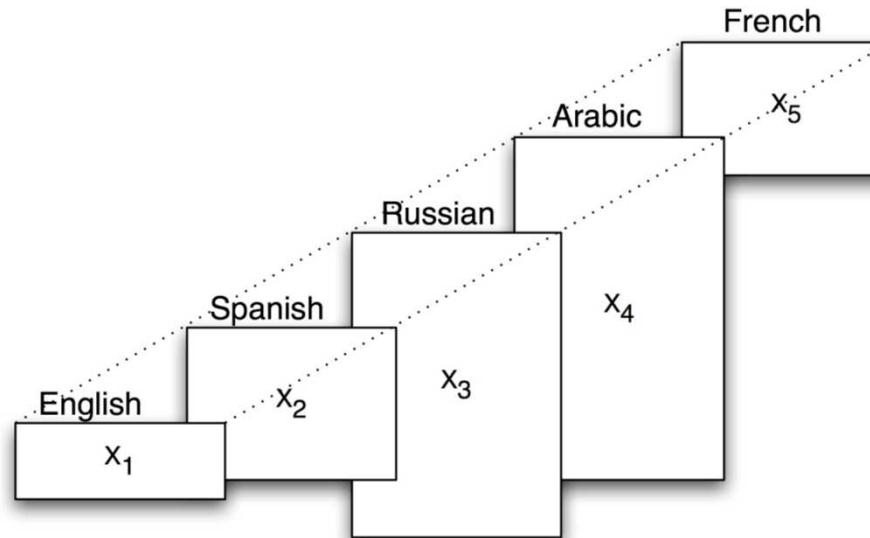
Different languages have different statistics

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





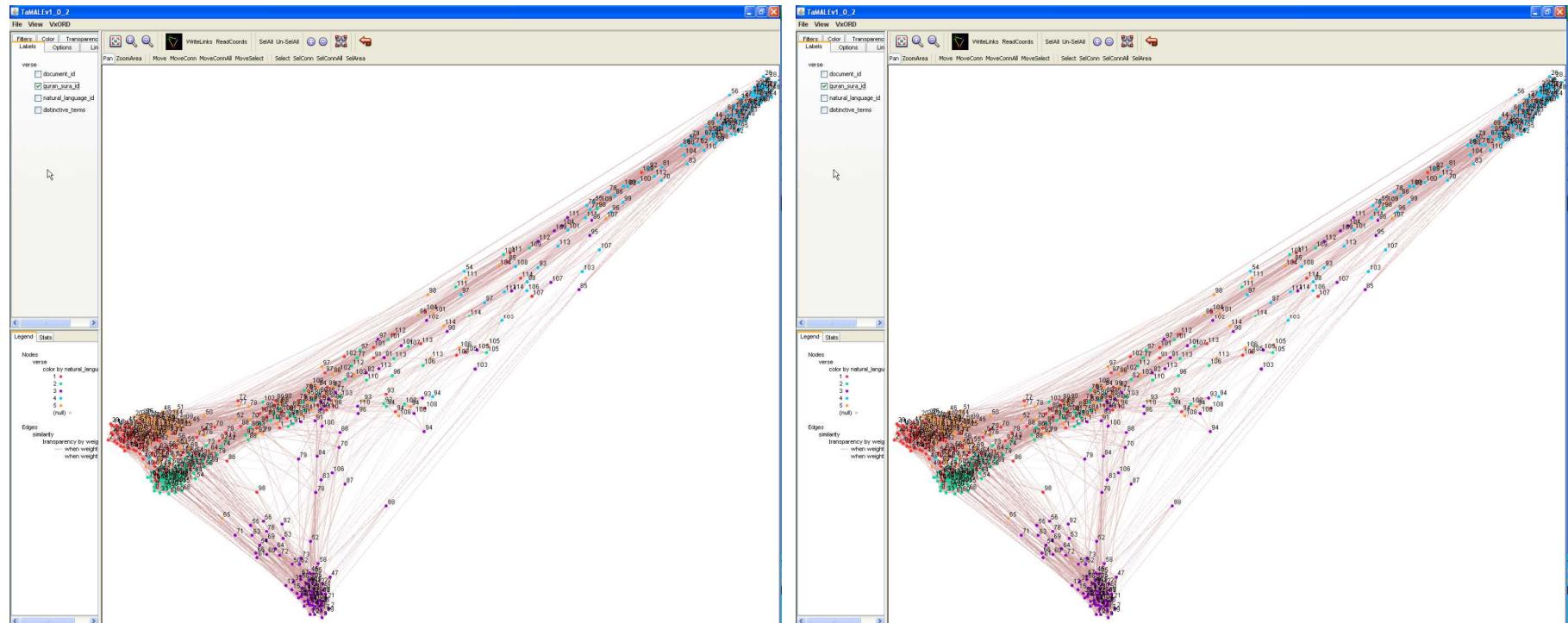
Comparison of LSA and PARAFAC2 (1)

Parameters for common-basis comparison

- 5 parallel languages in training data
- 70 dimensions
- Log-entropy weighting scheme

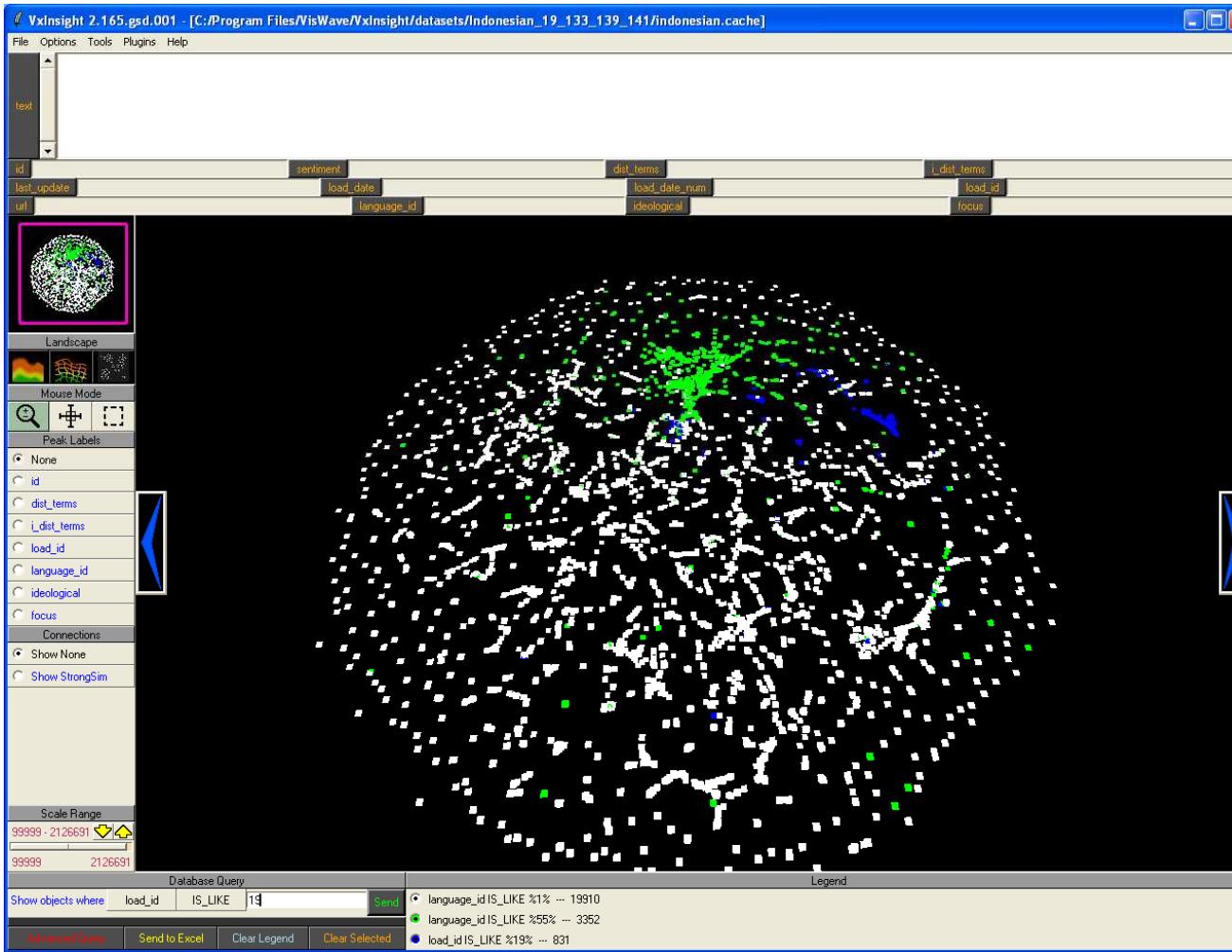
Common-basis comparison	LSA	PARAFAC2
Multilingual precision at 0	0.3473	0.5722
Multilingual precision at 5 docs	0.3330	0.5568
Precision at 0	0.7190	0.8079
Precision at 1 doc	0.6418	0.7509
Best achieved (>= 5 training languages)		
Multilingual precision at 0	0.3473	0.6504
Precision at 0	0.8819	0.9342

Comparison of LSA and PARAFAC2 (2)



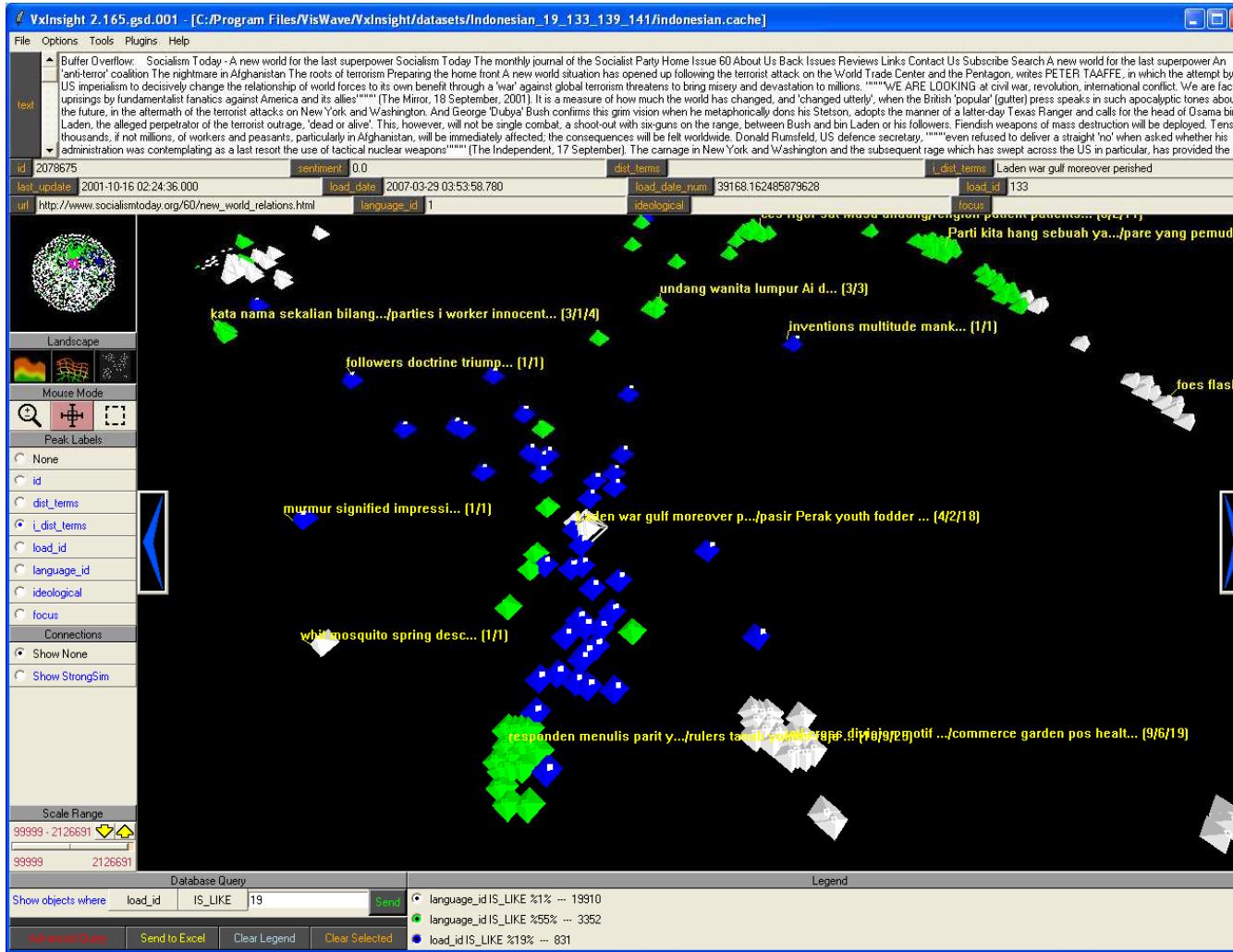


Multilingual visualization



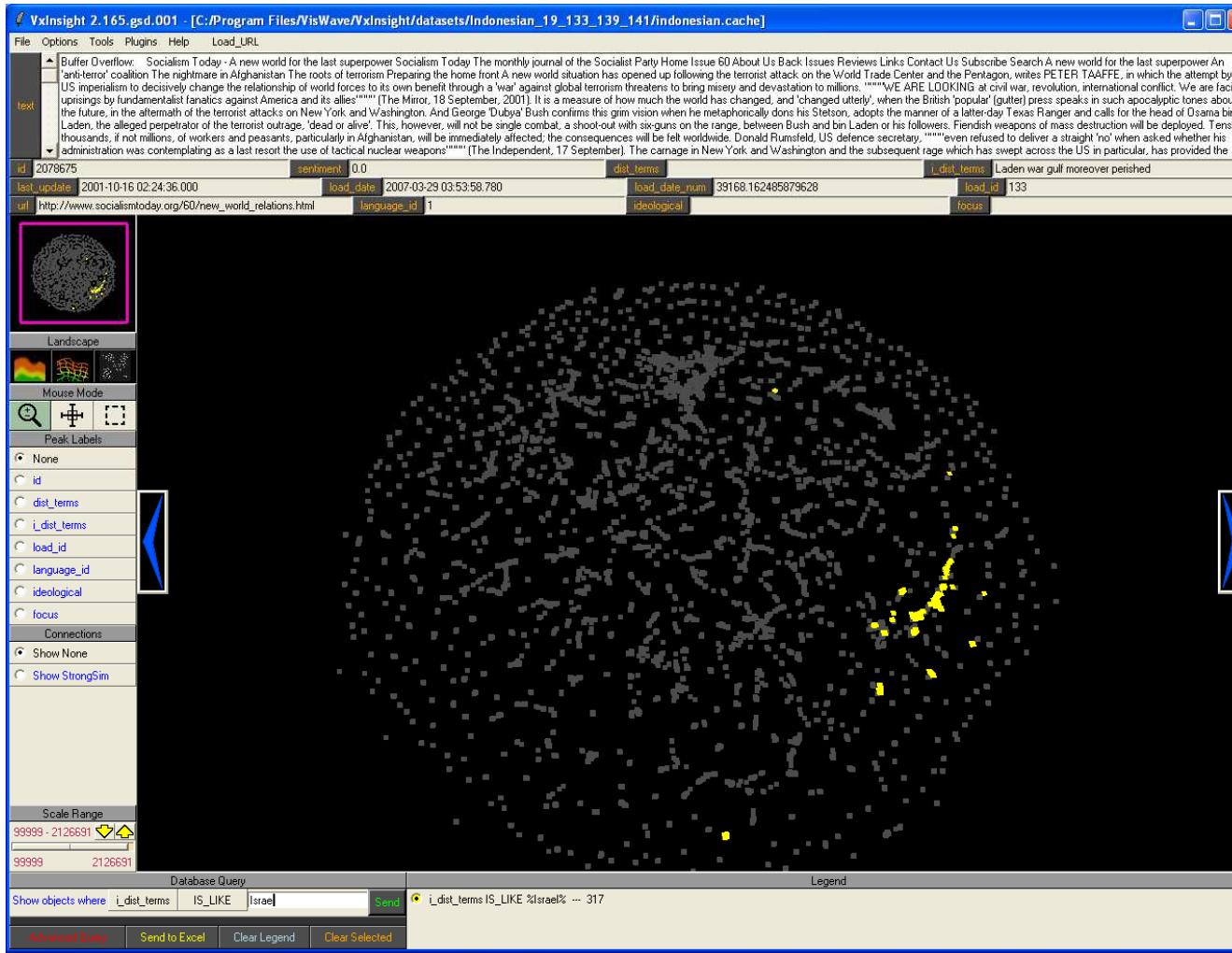
Sandia
National
Laboratories

Cluster detail



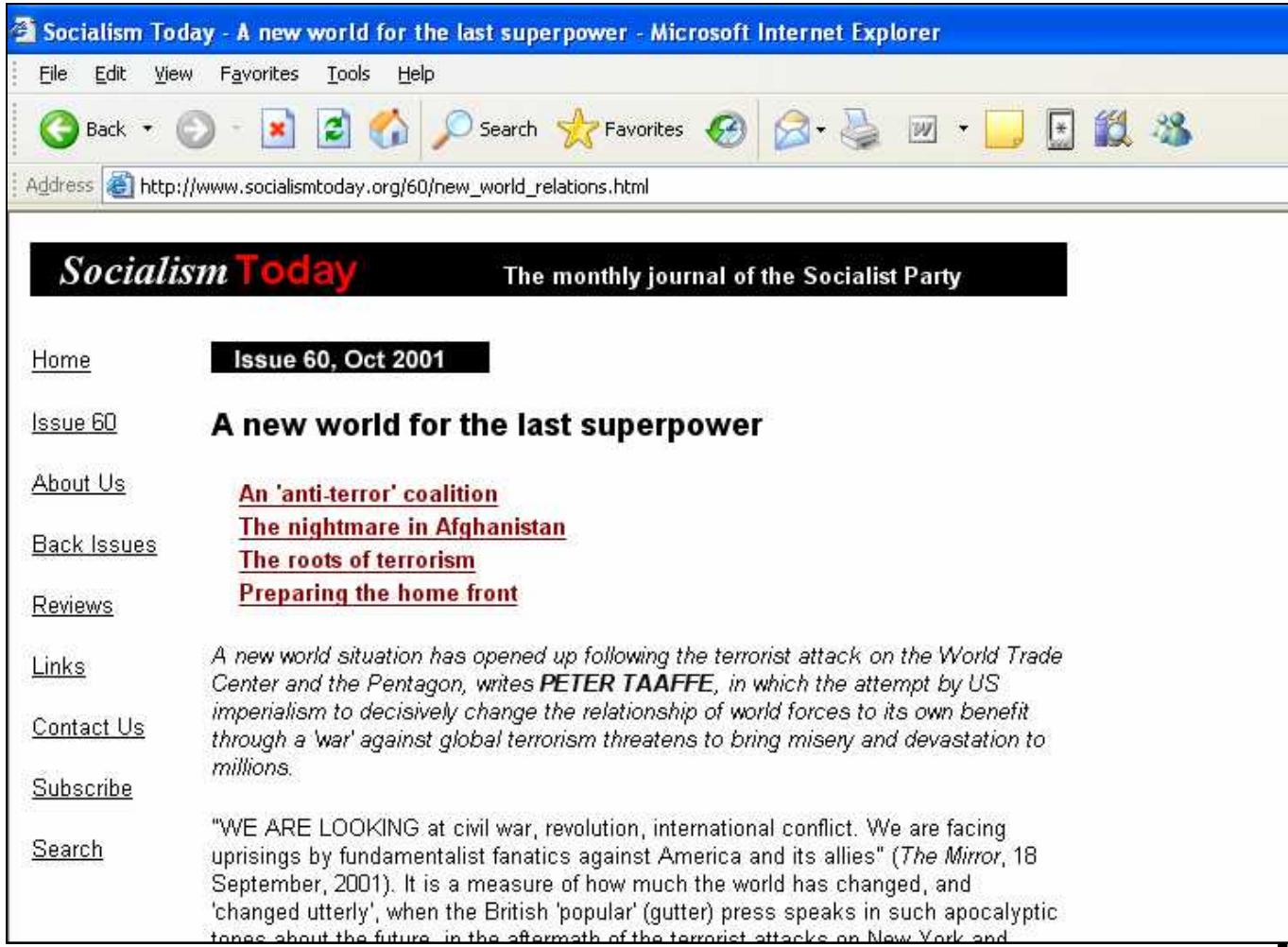
Sandia
National
Laboratories

Where do topics of interest fit into the overall space?





Navigation to documents of interest



Socialism Today - A new world for the last superpower - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back

Address http://www.socialismtoday.org/60/new_world_relations.html

Socialism Today The monthly journal of the Socialist Party

[Home](#) **Issue 60, Oct 2001**

[Issue 60](#) **A new world for the last superpower**

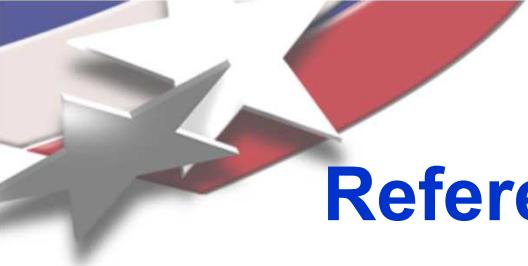
[About Us](#) [An 'anti-terror' coalition](#)
[Back Issues](#) [The nightmare in Afghanistan](#)
[Reviews](#) [The roots of terrorism](#)
[Links](#) [Preparing the home front](#)
[Contact Us](#)
[Subscribe](#)
[Search](#)

A new world situation has opened up following the terrorist attack on the World Trade Center and the Pentagon, writes PETER TAAFFE, in which the attempt by US imperialism to decisively change the relationship of world forces to its own benefit through a 'war' against global terrorism threatens to bring misery and devastation to millions.

"WE ARE LOOKING at civil war, revolution, international conflict. We are facing uprisings by fundamentalist fanatics against America and its allies" (The Mirror, 18 September, 2001). It is a measure of how much the world has changed, and 'changed utterly', when the British 'popular' (gutter) press speaks in such apocalyptic tones about the future in the aftermath of the terrorist attacks on New York and



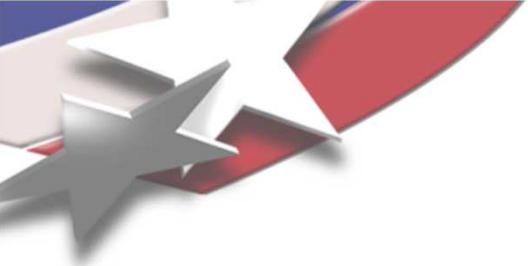
Sandia
National
Laboratories



References and contact information

- 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, forthcoming. 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. Forthcoming. Cross-language information retrieval using PARAFAC2. Submitted to KDD 2007.
- Chew and Abdelali. Forthcoming. The Effects of Language Relatedness and Morphology on Multilingual Information Retrieval: A Case Study With Semitic Languages

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



DISCUSSION and QUESTIONS

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