

Multilingual Text Analysis of On-line Discussions for Early Warning

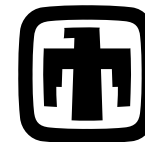
A Perspective from Sandia's "Networks Grand Challenge", ngc.sandia.gov

California Council on Science and Technology
Council Meeting on "Big Data"

Philip Kegelmeyer, wpk@sandia.gov, csmr.ca.sandia.gov/~wpk



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



**Sandia
National
Laboratories**

October 13, 2011



(Advance) Summary



- **What?**

Continuous capture and analysis of aggregated, voluntarily public text.

- **Why?**

Predict: will heated on-line discussion set something physical on fire?

- **Why Big Data?**

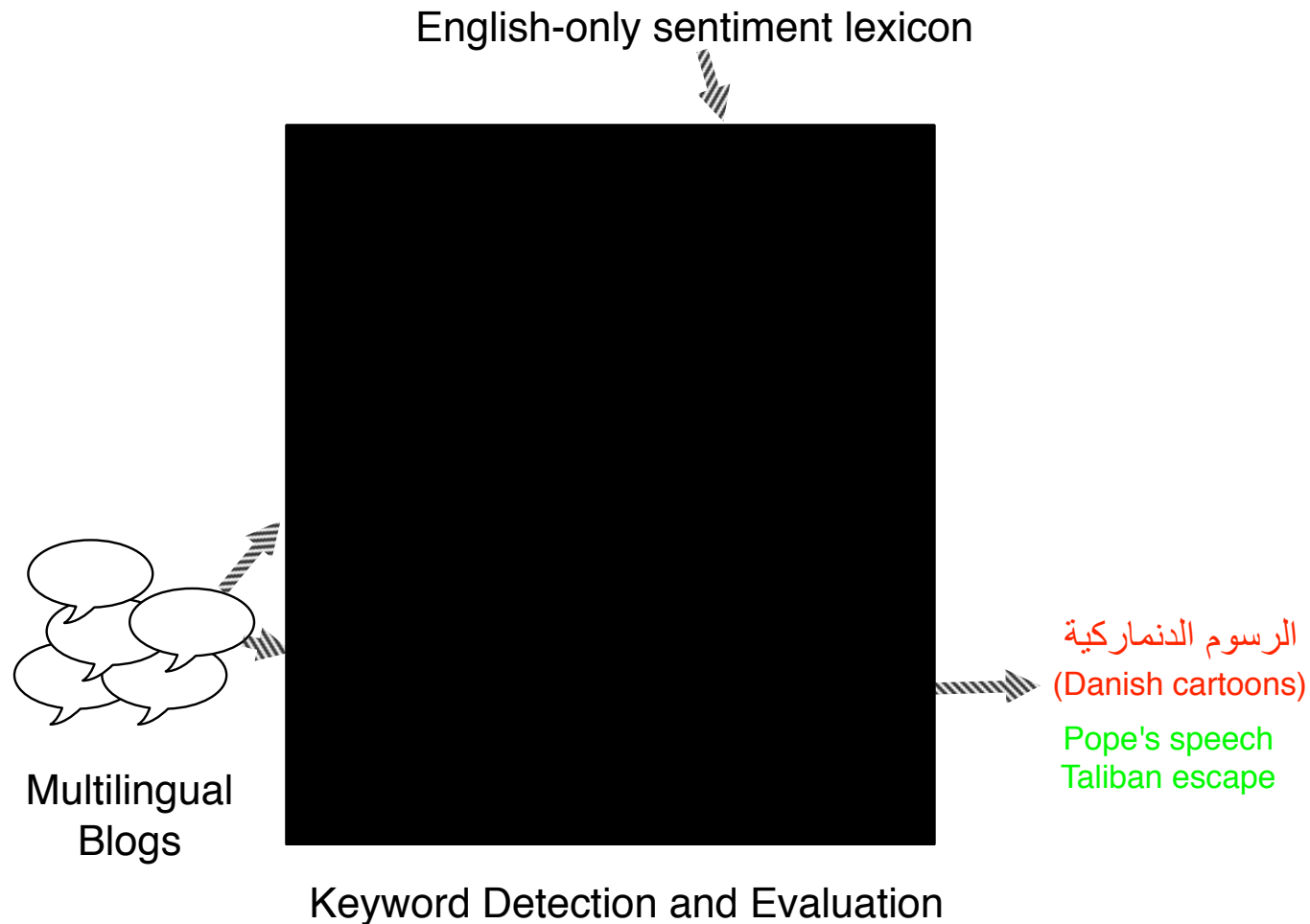
- *Not* just “the biggest net catches the most fish”.
- Techniques require big data in a very precise, mathematical way.

- (Some of the) **Policy Issues**

- Privacy, and our historical digital contrail.
- Anything that might chill speech is worrisome.
- Concerns about over-interpretation.

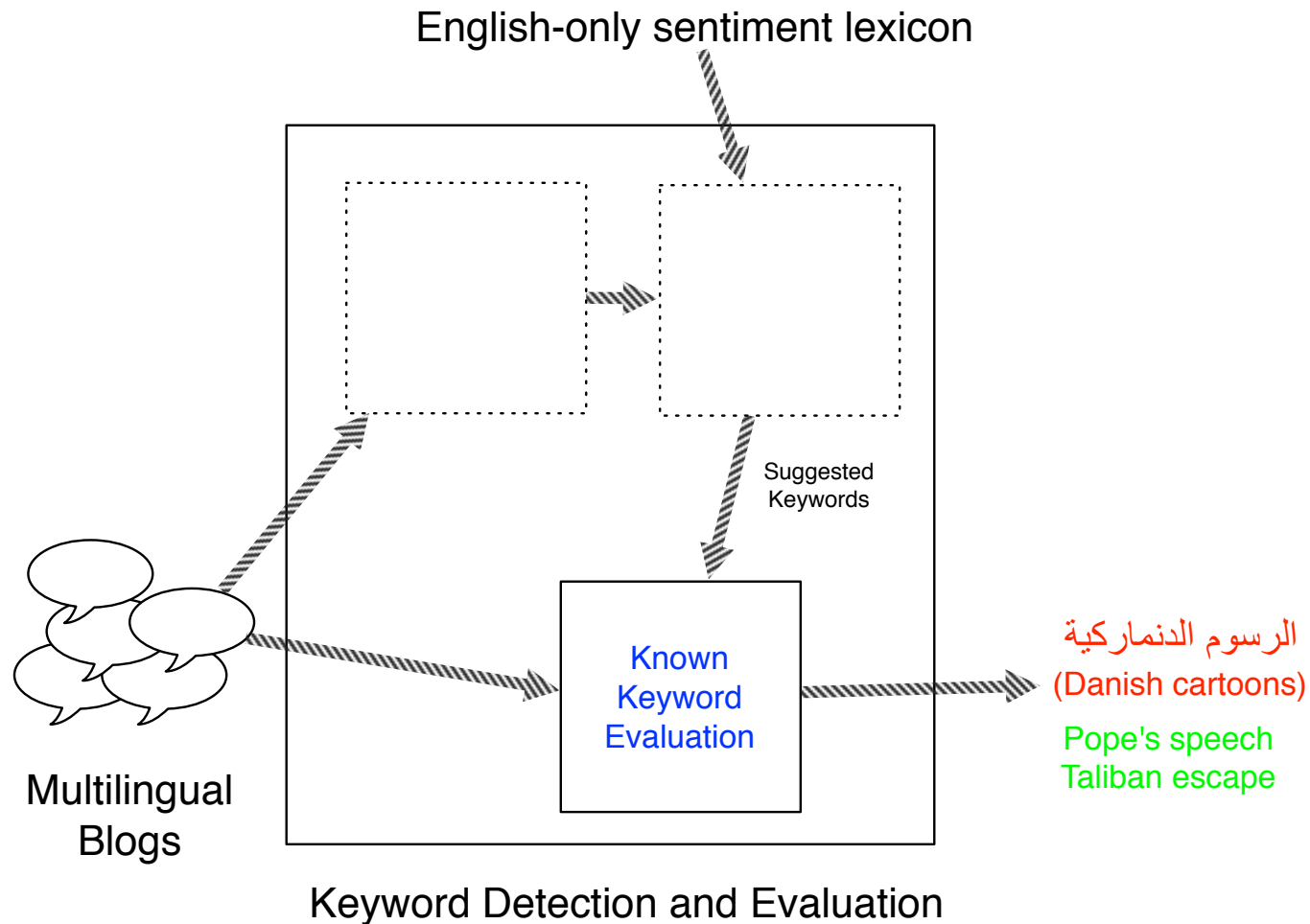


The Early Warning Black Box





Evaluating Known Keywords

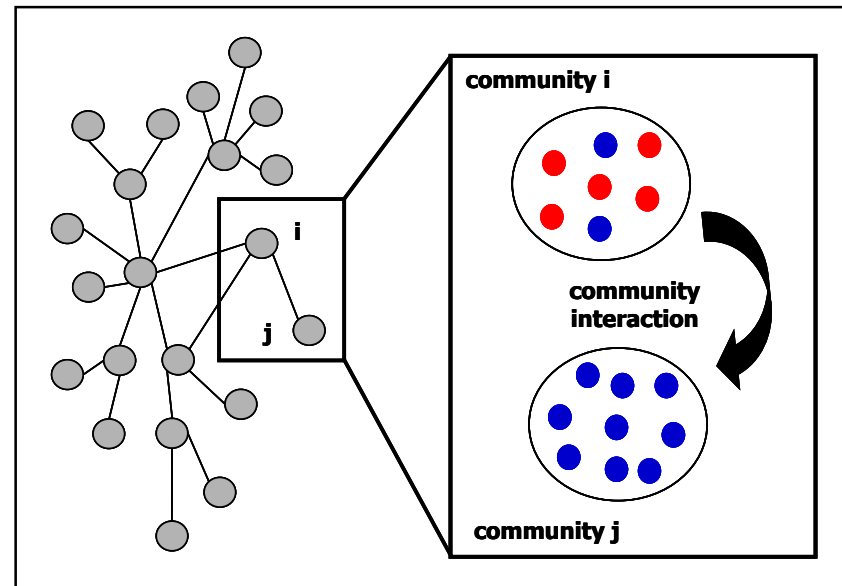
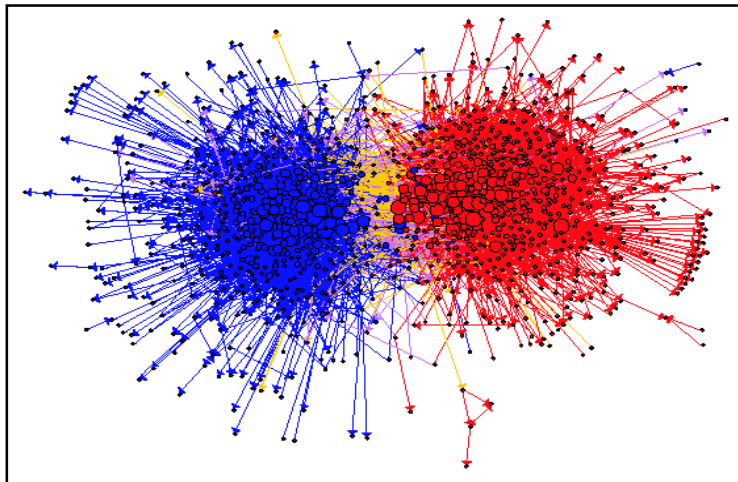




Multi-scale social dynamics model

A broad range of social dynamics phenomena can be usefully represented within a multi-scale modeling framework:

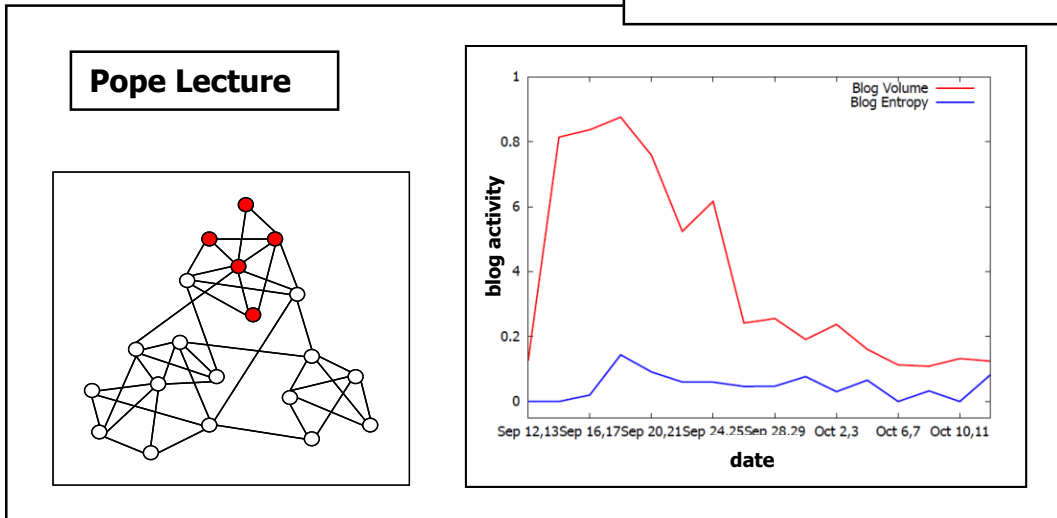
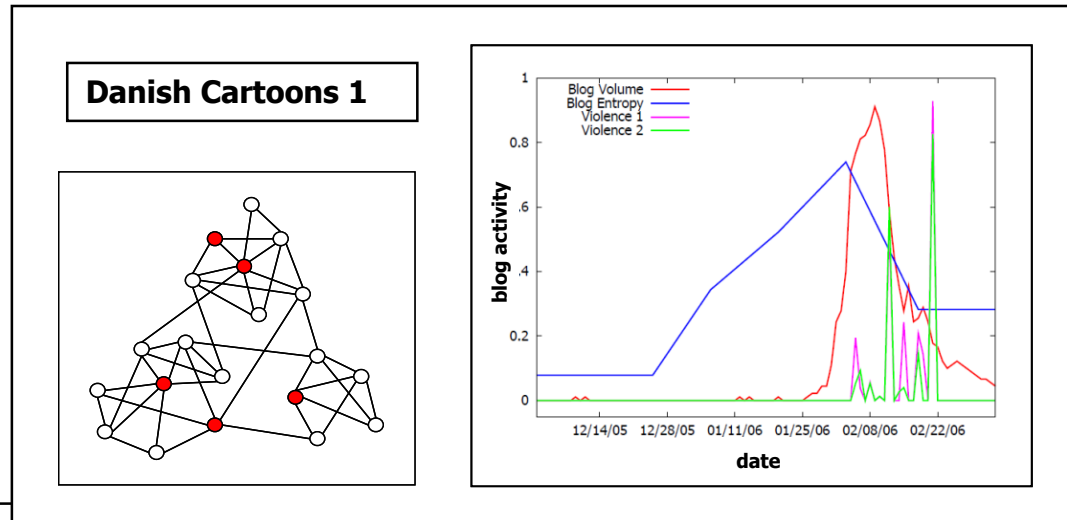
- micro-scale – behavior of individuals;
- meso-scale – interactions *within* social network communities;
- macro-scale – interactions *between* communities.





Entropy as an early indicator ...

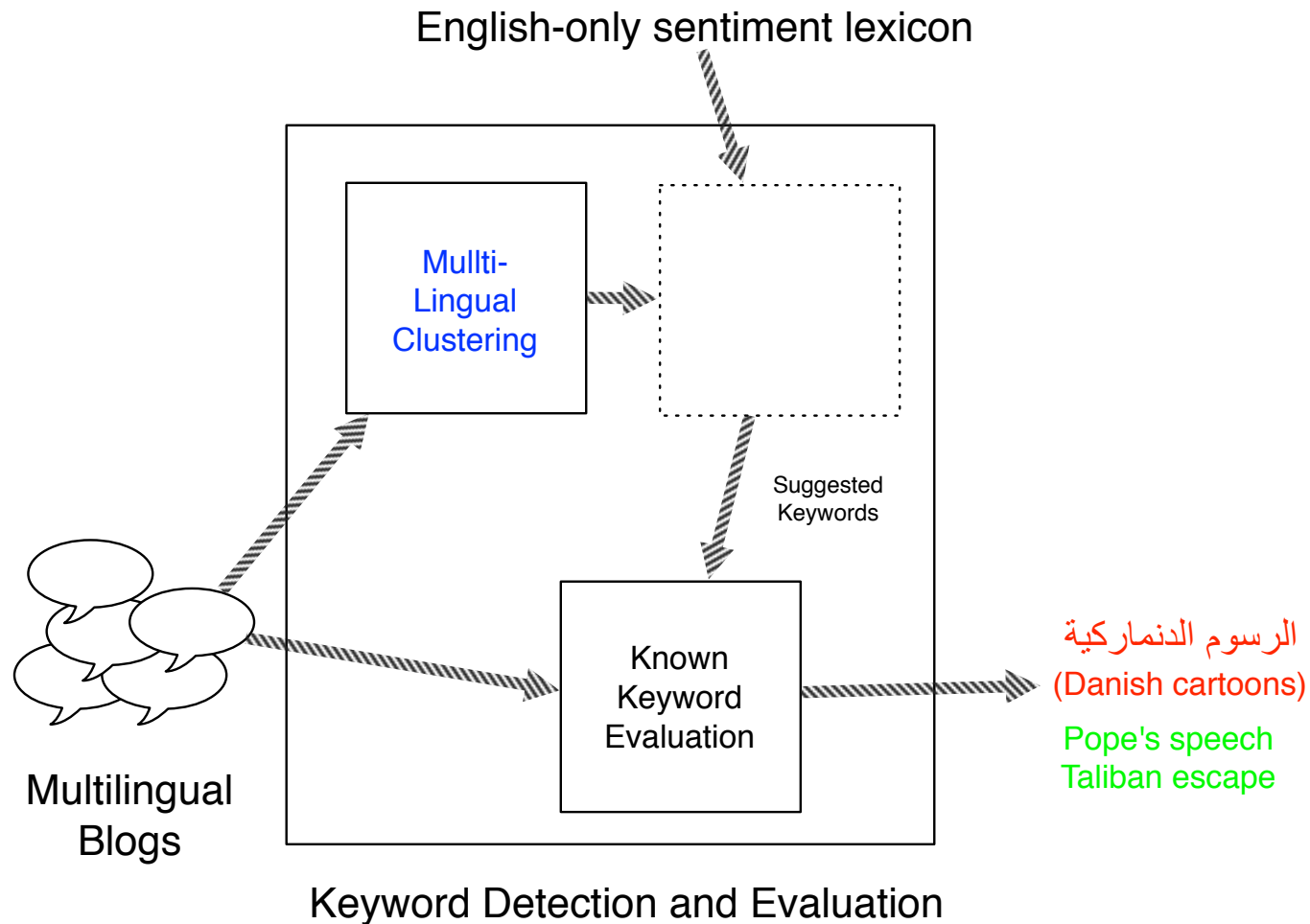
Blog post dispersion across communities is an useful early indicator of large mobilizations.



Predictive Analysis	
<u>Metric</u>	<u>Predictive?</u>
post entropy	yes (p<0.002)
post volume	no
lexicon intrinsics	no



Multi-lingual Text Clustering

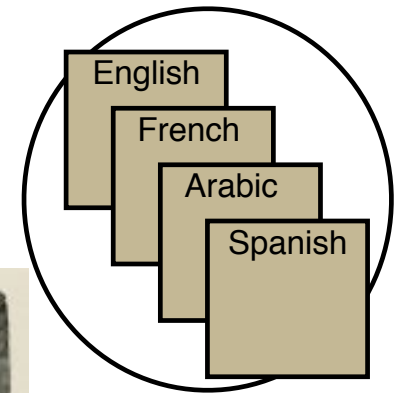


SNL has developed multilingual text analysis to link threats across multiple languages

- “Translate” new documents into a language-independent concept space, which is useful for:
 - Translation triage (i.e., translate documents in clusters of interest)
 - Ideological classification (e.g., hostile to U.S.)
 - Multilingual sentiment analysis

Sandia’s database: 54 languages: >99% coverage of web

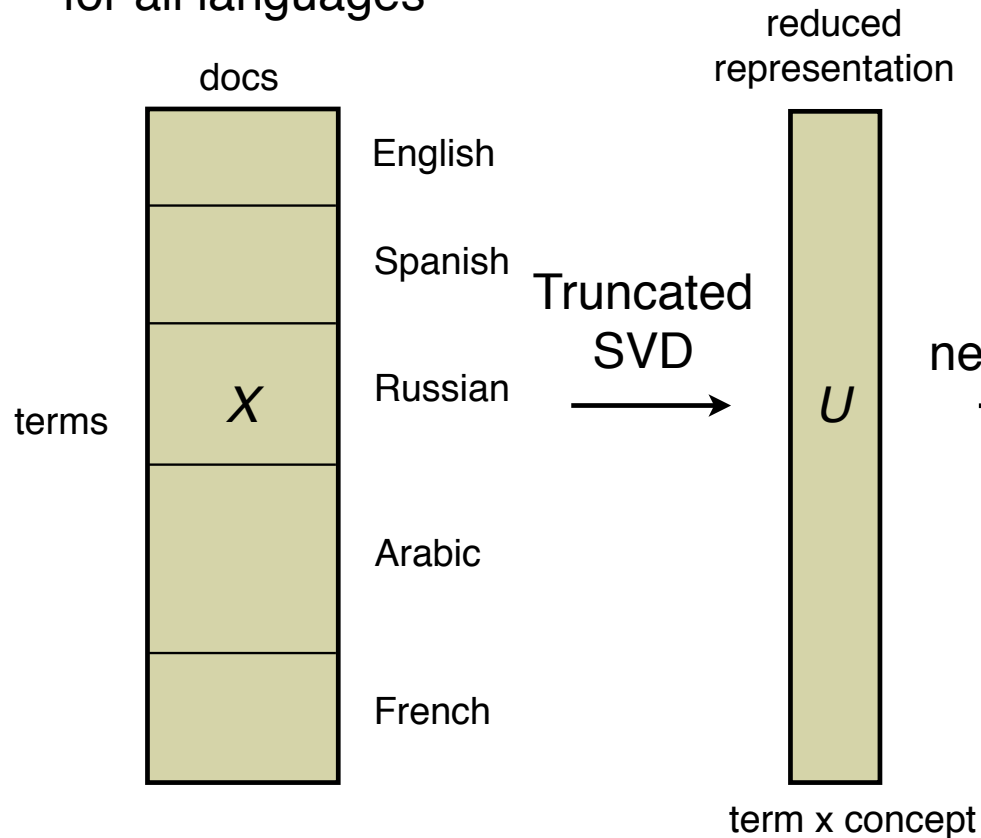
Afrikaans	Estonian	Norwegian
Albanian	Finnish	Persian (Farsi)
Amharic	French	Polish
Arabic	German	Portuguese
Aramaic	Greek (New Testament)	Romani
Armenian Eastern	Greek (Modern)	Romanian
Armenian Western	Hebrew (Old Testament)	Russian
Basque	Hebrew (Modern)	Scots Gaelic
Breton	Hungarian	Spanish
Chamorro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



The Rosetta Stone

Multilingual Latent Semantic Analysis

Term-by-doc matrix
for all languages

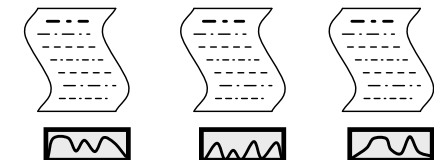


“Translate” new documents
into a small number of
language-independent features

Project
new documents

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

Document feature
vector

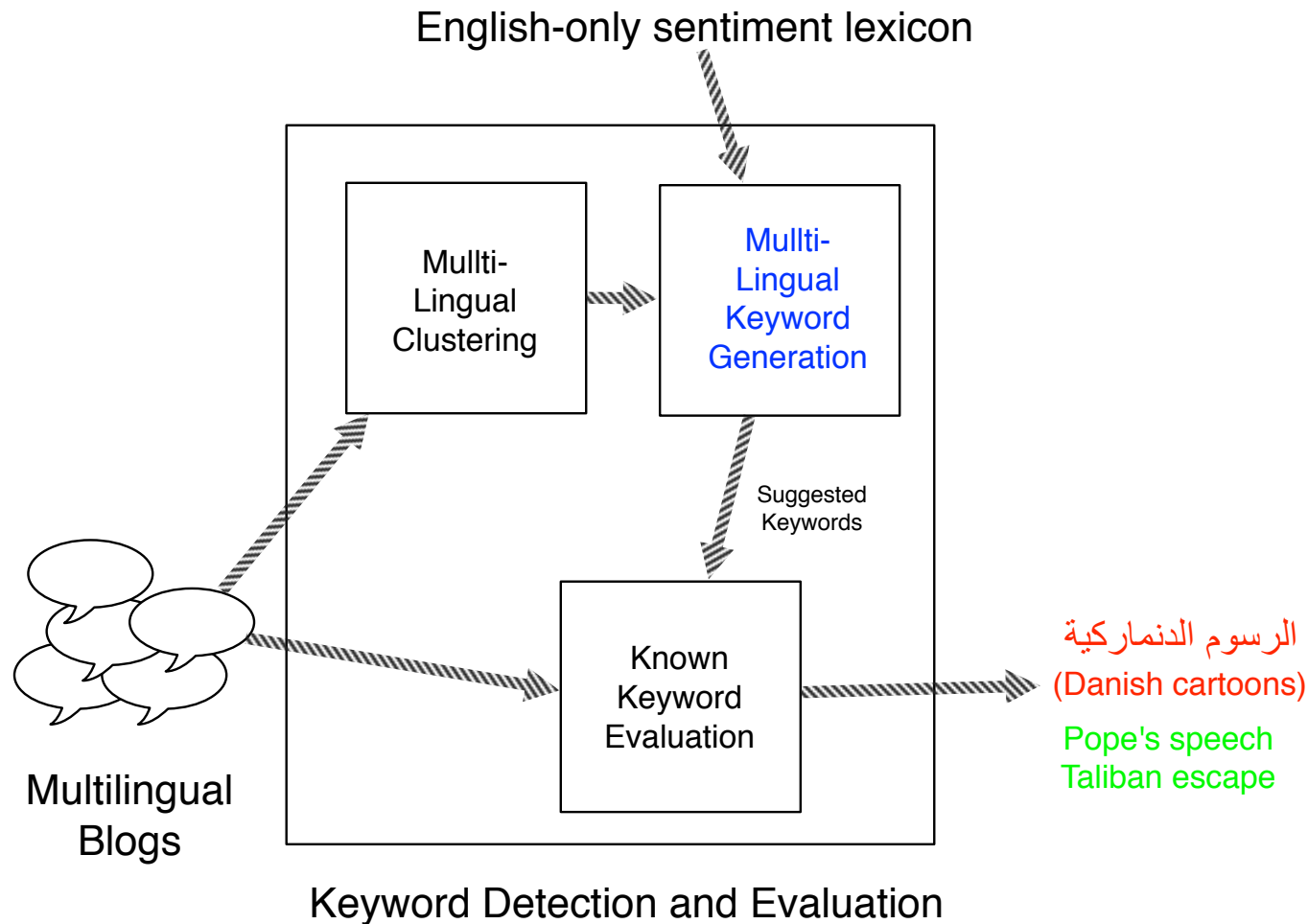


Applications

- cross-language retrieval
- pairwise similarities for clustering
- machine learning applications

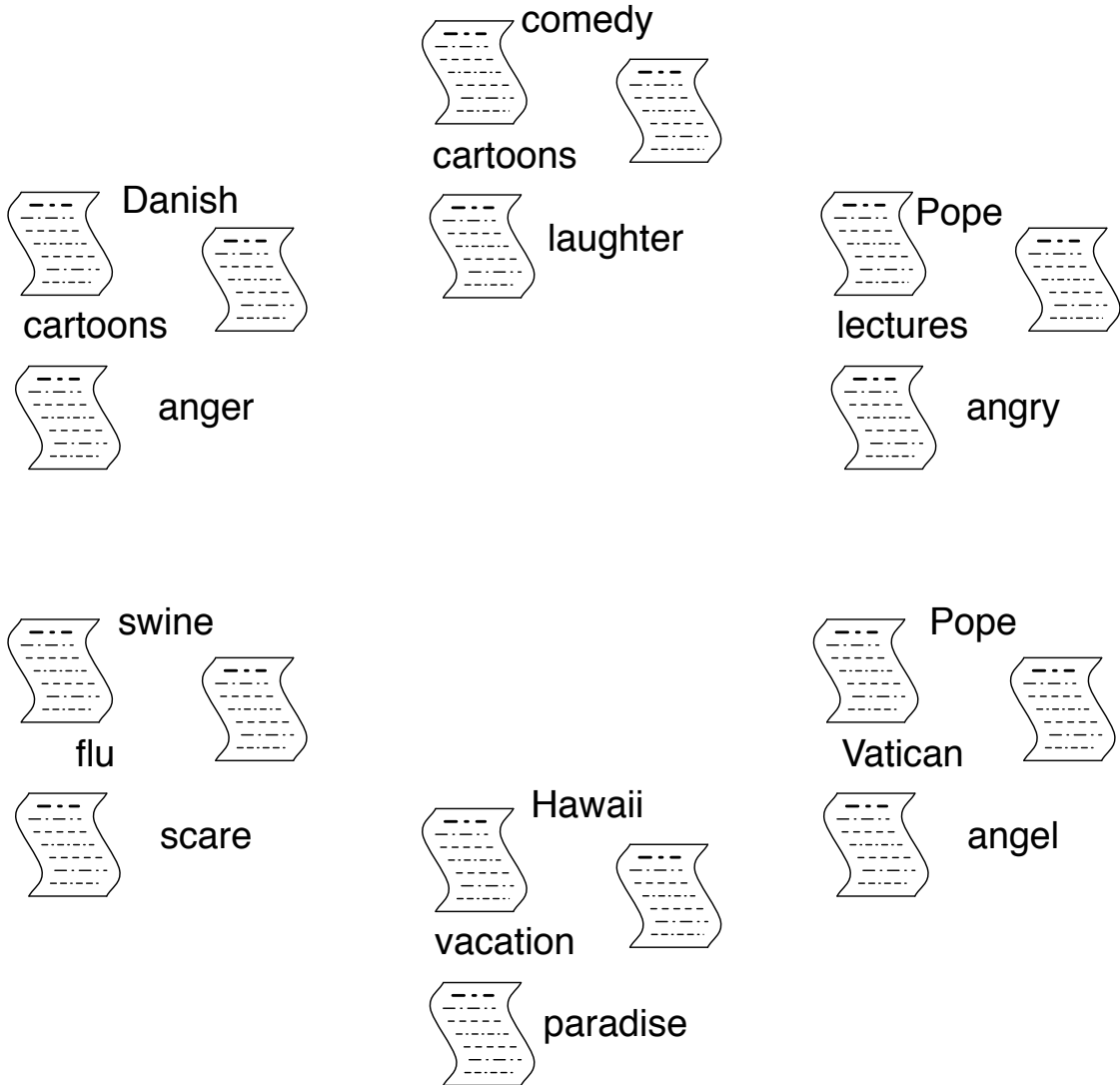


Multi-lingual Sentiment Analysis





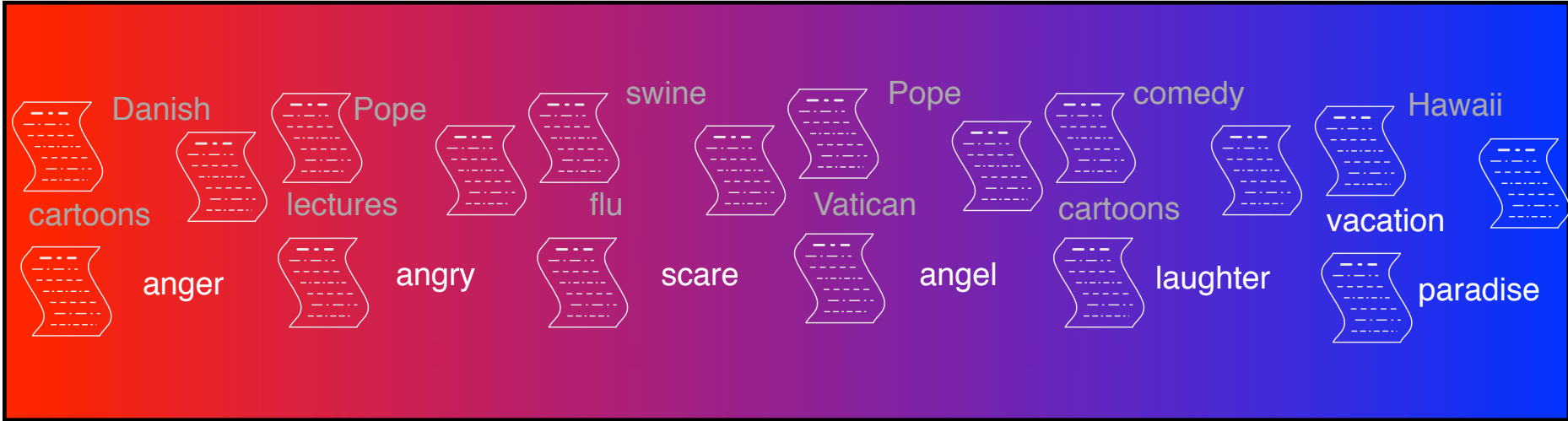
Conceptual Example





Conceptual Example

1) Sort the documents according to sentiment



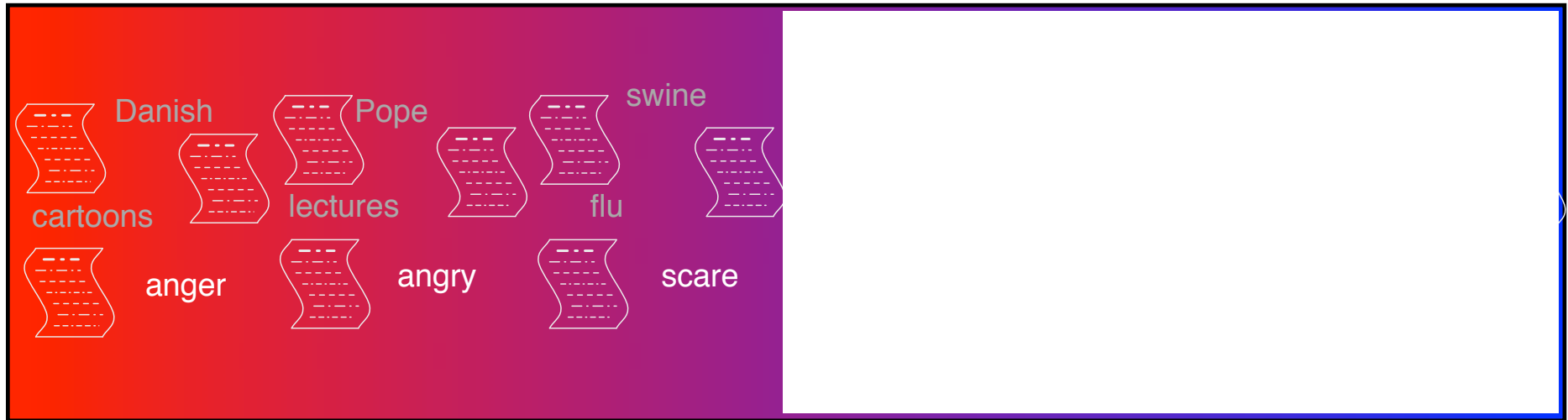
Unpleasant

Pleasant



Conceptual Example

2) Keep only the highly emotional documents



Unpleasant

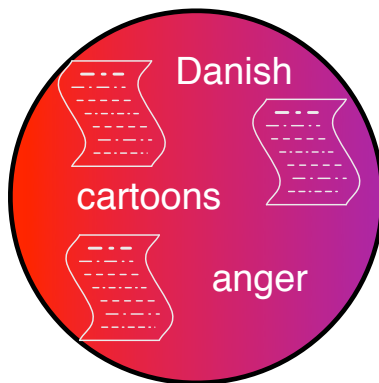
Ambivalent

Pleasant

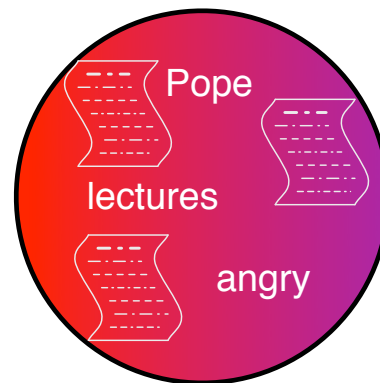


Conceptual Example

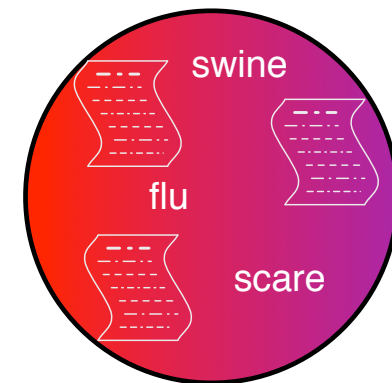
- 3) Cluster by topic using multilingual document clustering
- 4) Find unique keywords that describe each cluster



Danish, cartoons



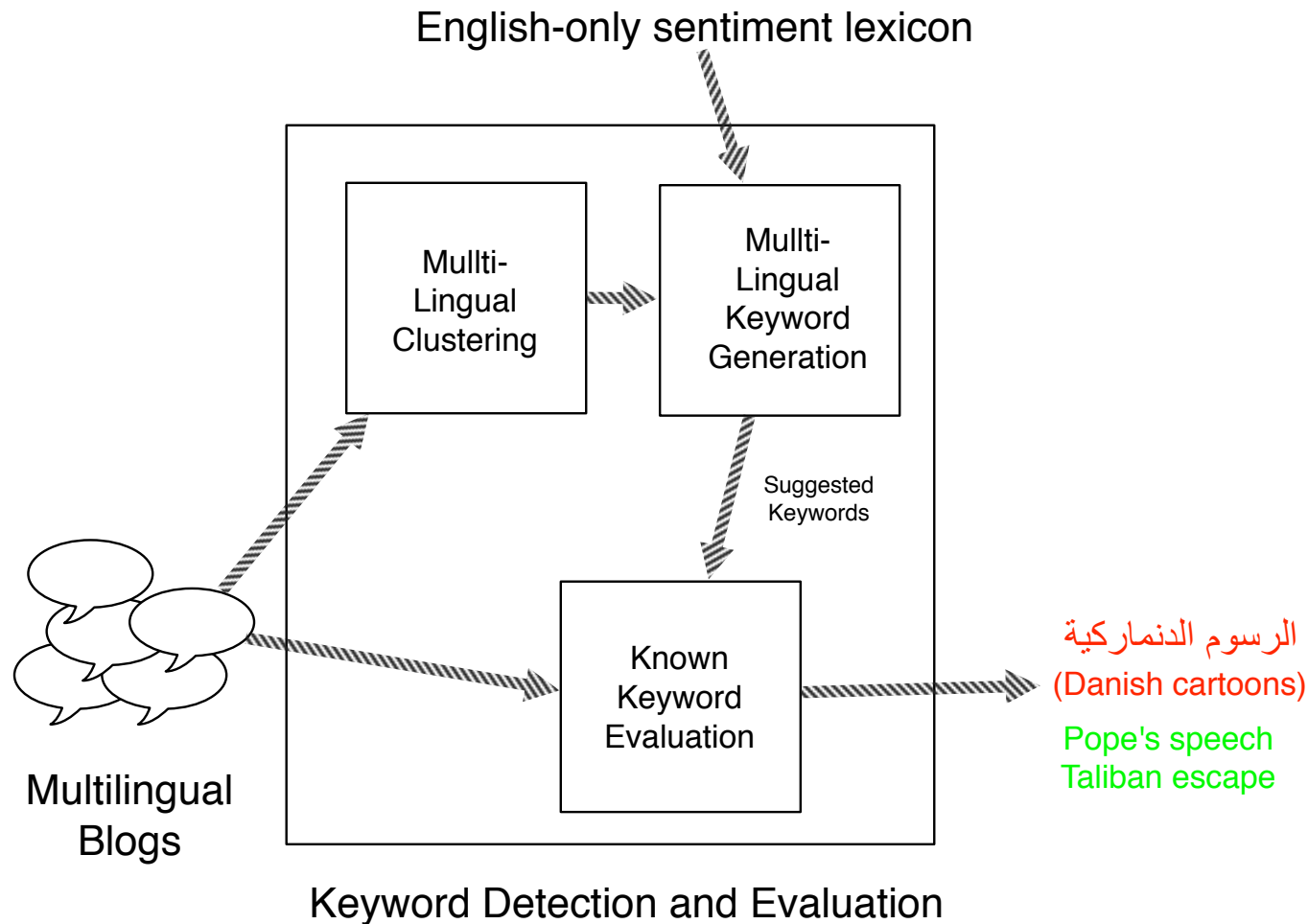
Pope, lectures



swine, flu



The Black Box, Unpacked . . .





Status and Outcomes



- **Status**

Not deployed operationally for nightly updates.

- **Challenges**

- Practical: pairwise similarity calculations, ugliness of web data.
- Validation: customer anecdotes, clustering validation, keyword replication, sentiment prediction studies, including peer-reviewed publication. (See ngc.sandia.gov.)

- **Example analyses**

Bali-bomber execution in early November 2008; correctly predicted that outrage was **not** self-sustaining.

Israel/Gaza conflict in early January 2009; correctly predicted that hacking discussion **was** self-sustaining, then actual hacking occurs.



Summary (Reprise)



- **Why Big Data?**
 - Quantitative effect on community modeling
 - More languages helps, not hurts, multi-lingual translation.
 - More text improves keyword extraction.
- (Some of the) **Policy Issues**
 - Privacy, and our digital contrails:
data captured for one purpose is inevitably used for others.
 - Anything that might chill speech is worrisome.
 - Concerns about over-interpretation:
these methods cannot predict what, when, or, especially, who.
 - Methods inherently useful only for *public*, linked discourse:
useless on private data.



For More Information ...



- NGC Final Report, all publications, all contact info: ngc.sandia.gov

Sandia National Laboratories: NGC Home

http://wwwd.sandia.gov/ngc/index.html

Sandia National Laboratories: NGC Home

Employee Locator | Index | Site Map

Sandia National Laboratories

About Sandia | Mission Areas | Newsroom | Careers | Doing Business | Education | Contact Us

NGC Home

Publications

Contacts

Network Discovery, Characterization, and Prediction Grand Challenge (NGC)

The Challenge: Adversarial Networks Impacting National Security

Networks engaged in weapons proliferation, terrorism, cyber attacks, clandestine resale of dual-use imports, arms and drug smuggling, and other illicit activities are major threats to national security. These adversarial networks in turn rely on legitimate and illegitimate secondary networks for financial, supply chain, communication, recruiting, and fund-raising activities. Complexity, dynamism, resilience and adaptability make adversarial networks extremely difficult to identify and disrupt. Often the only way an individual may be detected is through the networks they use, and the arrest of an individual may not remove the underlying threat if the networks remain intact. In short, our real adversaries are networks.

The Solution: Research and Develop Analysis Capabilities

Our goal, then, is to research and develop analysis capabilities that address adversarial networks. The full title of the project, "Network Discovery, Characterization, and Prediction," conveys the scope and challenges involved. The discovery of adversarial networks is immensely difficult in its own right. A network may only reveal itself by the union of its parts. Individual relationships and activities may appear completely benign in isolation. Data relevant to network discovery may come from communications, financial transactions, human intelligence reports, shipment records, cyber events or many other sources. It may be geographically or temporally dispersed. Thus, very large and heterogeneous data collections must be analyzed collectively to detect networks. The characterization of networks requires methods for identifying likely relationships that are not captured in the data. The structure of a network conveys information about its purpose and the roles of its component individuals, organizations and activities. It can reveal command and control structure and critical components. Structure can also suggest likely evolution and intent, allowing prediction of the possible future shapes of the network.

In sum, we are creating at Sandia, in support of the nation, the unique capability to answer currently unanswerable questions.

NGC Project Goals

- Build upon considerable existing Sandia capabilities in scalable computing and advanced analysis algorithms
- Understand and elicit the needs of the intelligence community
- Do basic research on uncertainty in the intelligence domain
- Research and evaluate novel analysis algorithms
- Implement that research to address those needs to create a flexible, interactive capability for intelligence analysis on large datasets

- Predictions from communities: Rich Colbaugh, rcolbau@sandia.gov, 505 284-4116
- Multilingual text analysis: Brett Bader, bwbader@sandia.gov, 505 845-0514
- PI, and sentiment analysis: Philip Kegelmeyer, wpk@sandia.gov, 925 294-3016